

Surprising biology
in uncharted
microbial sequences

F. A. Bastiaan von Meijenfeldt

Surprising biology in uncharted microbial sequences

Frederik Alexander Bastiaan von Meijenfeldt

Author: F. A. Bastiaan von Meijenfeldt

Cover layout and artwork: Verali von Meijenfeldt

Printing: Ridderprint, ridderprint.nl

Layout and design: Garcella Dings, persoonlijkproefschrift.nl

ISBN: 978-94-6483-259-4

Copyright 2023 © F. A. Bastiaan von Meijenfeldt

The Netherlands. All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author.

Surprising biology in uncharted microbial sequences

Verrassende biologie in onverkende microbiële sequenties
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 23 augustus 2023 des middags te 2.15 uur

door

Frederik Alexander Bastiaan von Meijenfeldt

geboren op 30 maart 1989
te Nieuwegein

Promotoren:

Prof. dr. B. Snel

Prof. dr. P. Hogeweg

Prof. dr. B.E. Dutilh

Beoordelingscommissie:

Prof. dr. M.H. Medema

Prof. dr. ir. C.M.J. Pieterse

Prof. dr. F.E. Rodriguez-Valera

Dr. A.C. Schürch

Prof. dr. A. Spang

Contents

	Summary	6
	Samenvatting	7
Chapter 1.	Introduction	9
Chapter 2.	Bridging the membrane lipid divide: bacteria of the FCB group superphylum have the potential to synthesize archaeal ether lipids	29
Chapter 3.	Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT	101
Chapter 4.	Integration of taxonomic signals from MAGs and contigs improves read annotation and taxonomic profiling of metagenomes	135
Chapter 5.	A social niche breadth score reveals niche range strategies of generalists and specialists	163
Chapter 6.	Structural colour in the bacterial domain: the ecogenomics of an optical phenotype	217
Chapter 7.	Discussion	247
Appendices	References	259
	Acknowledgements	284
	Curriculum vitae	286
	List of publications	287

Summary

The microbiome, all microorganisms in their natural habitat, can be collected on filters and its DNA extracted. Sequencing of this metagenome allows us to observe microorganisms not under a microscope or in the lab but via their DNA and in their natural habitat, revealing who they are and what they can do. Uncharted microbial sequences from the environment hold potential to change our view of the tree of life, by uncovering a vast microbial diversity that is not easily cultivated in the lab.

Novel microorganisms with surprising biological traits await discovery in newly sequenced metagenomes. In addition, environmental sequences published by others are available in public databases, waiting to be further explored. The chapters of this thesis describe the search for unexpected biology in environmental sequences, both in newly sequenced datasets and in those already published. Reanalysis of already published datasets allows for comparisons between microbiomes from many different habitats, and for classical ecological questions to be addressed on the global scale. The chapters also describe the design of algorithms and tools to taxonomically annotate sequences from previously unknown microorganisms that are common in underexplored habitats and still surface in those more intensely studied like human-associated or marine habitats, and to use these annotations for an accurate and comprehensive view of the microbiome.

Surprising biology is found deep down in the Black Sea, where elusive bacteria include archaeal-type lipids in their cell membrane, challenging a once-thought fundamental divide across the tree of life in the molecular composition of the membrane. Surprising biology is also found in the discovery of structural colour—the striking display of changing colours depending on the angle of observation—in the bacterial domain and in many different microbial habitats, including curiously the deep ocean where no light penetrates. By reanalysis of thousands of environmental sequencing studies across a wide range of habitats and geographical locations, the social niche breadth of microorganisms—the range of communities in which each lives—is quantified and ecological and genomic correlates of microbial specialism versus generalism are revealed.

Together, the chapters paint a picture of a microbial world that is still largely left uncharted. With the development of new algorithms and tools, and the (re)analysis of large-scale data, this thesis uncovers a fraction of that vast microbial unknown.

Samenvatting

Het microbioom, alle micro-organismen in hun natuurlijke habitat, kan op filters worden verzameld en vervolgens kan het DNA ervan worden geëxtraheerd. Sequenten van dit metageenoom stelt ons in staat om micro-organismen te observeren niet onder een microscoop of in het laboratorium maar via hun DNA en in hun natuurlijke habitat, wat onthult wie ze zijn en wat ze kunnen doen. Nog niet in kaart gebrachte microbiële sequenties uit de omgeving kunnen ons beeld van de boom van het leven veranderen, door een enorme microbiële diversiteit bloot te leggen die niet gemakkelijk in het laboratorium kan worden gekweekt.

Nieuwe micro-organismen met verrassende biologische eigenschappen kunnen worden ontdekt in nieuw gesequencete metagenomen. Daarnaast zijn sequenties uit de omgeving die door anderen zijn gepubliceerd beschikbaar in openbare databases, in afwachting van verdere verkenning. De hoofdstukken van dit proefschrift beschrijven de zoektocht naar onverwachte biologie in sequenties uit de omgeving, zowel in nieuw gesequencete als in reeds gepubliceerde datasets. Heranalyse van reeds gepubliceerde datasets maakt het mogelijk om microbiomen van veel verschillende habitats met elkaar te vergelijken, en om klassieke ecologische vragen op globale schaal te behandelen. De hoofdstukken beschrijven ook het ontwerp van algoritmes en tools om sequenties van voorheen onbekende micro-organismen taxonomisch te annoteren, die veel voorkomen in nog niet goed onderzochte habitats en die nog steeds opduiken in intensiever bestudeerde habitats zoals die in en op het menselijk lichaam of in de zee, en om deze annotaties te gebruiken voor een nauwkeurig en volledig beeld van het microbioom.

Verrassende biologie wordt diep in de Zwarte Zee gevonden, waar moeilijk te bereiken bacteriën lipiden van het archaeale type in hun celmembraan inbouwen, waardoor een ooit veronderstelde fundamentele scheiding in de moleculaire samenstelling van het membraan dwars door de boom van het leven wordt betwist. Verrassende biologie wordt ook gevonden in de ontdekking van structurele kleur—het opvallend vertoon van veranderende kleuren afhankelijk van de waarnemingshoek—in het bacteriële domein en in veel verschillende microbiële habitats, waaronder vreemd genoeg de diepe oceaan waar geen licht doordringt. Door heranalyse van duizenden sequencing studies uit de omgeving van een grote variëteit aan habitats en geografische locaties wordt de sociale niche breedte van micro-organismen—de verscheidenheid aan gemeenschappen waarin elk leeft—gekwantificeerd en worden ecologische en genomische correlaten van microbiële specialisme versus generalisme onthuld.

Samen schetsen de hoofdstukken een beeld van een microbiële wereld die nog grotendeels onbekend is. Met de ontwikkeling van nieuwe algoritmes en tools, en de (her)analyse van grootschalige data, brengt dit proefschrift een fractie van dat enorme microbiële onbekende aan het licht.





Chapter 1

Introduction

Most life is invisible

Too small to be seen without microscope, bacterial and archaeal microorganisms inhabit every corner of the biosphere. Diverse microbial communities live in soils and waters, on and inside animals and plants, and deep in the subsurface. A single gram of soil may contain over a billion prokaryotes, and the surface waters of the open ocean harbour half a million per millilitre (1). Our own body houses as many bacterial as human cells (2), and thousands of distinct prokaryotic species crowd the rhizosphere of a single plant (3).

Although tiny, microorganisms profoundly change the world in which they live. For example, the oxygenation of Earth's atmosphere 2.4 billion years ago is attributed to the rise of the Cyanobacteria (4,5), may have caused global glaciation (6,7), and has steered the course of evolution by making the highly reactive and toxic O₂ molecule common and available for use in metabolic and biosynthetic pathways (8–10). Atmospheric nitrogen is incorporated into biological compounds via nitrogen fixation by particular prokaryotes, an exclusive trait exploited via symbiosis by a few plants and insects (11–14). Microorganisms are directly involved in food production via fermentation (15), and degrade pollutants in heavily contaminated groundwater (16). The microbial communities that reside in humans are fundamental to health, digesting what we eat and supplying essential nutrients, deterring foreign invaders or making us sick (17–19). They direct behaviour and development of the brain (20). An intricate dependence on microorganisms is vital to all complex multicellular life (21,22). Microorganisms themselves depend on other microorganisms as well. Competitive and cooperative interactions shape microbial communities (23–26), and exchange of genetic material via horizontal gene transfer (HGT) promotes adaptation (27–31).

Yet, despite their ubiquity, and their ecological and evolutionary significance, most microorganisms have just been discovered in the past two decades. Only recently are we able to unbiasedly and comprehensively analyse the microbiome—all microorganisms in their natural habitat, and did we get to know them. Microbiome research was enabled by the development of cultivation-independent environmental sequencing techniques (32–35). The new view that environmental DNA sequences offer has meant a paradigm shift for microbiology.

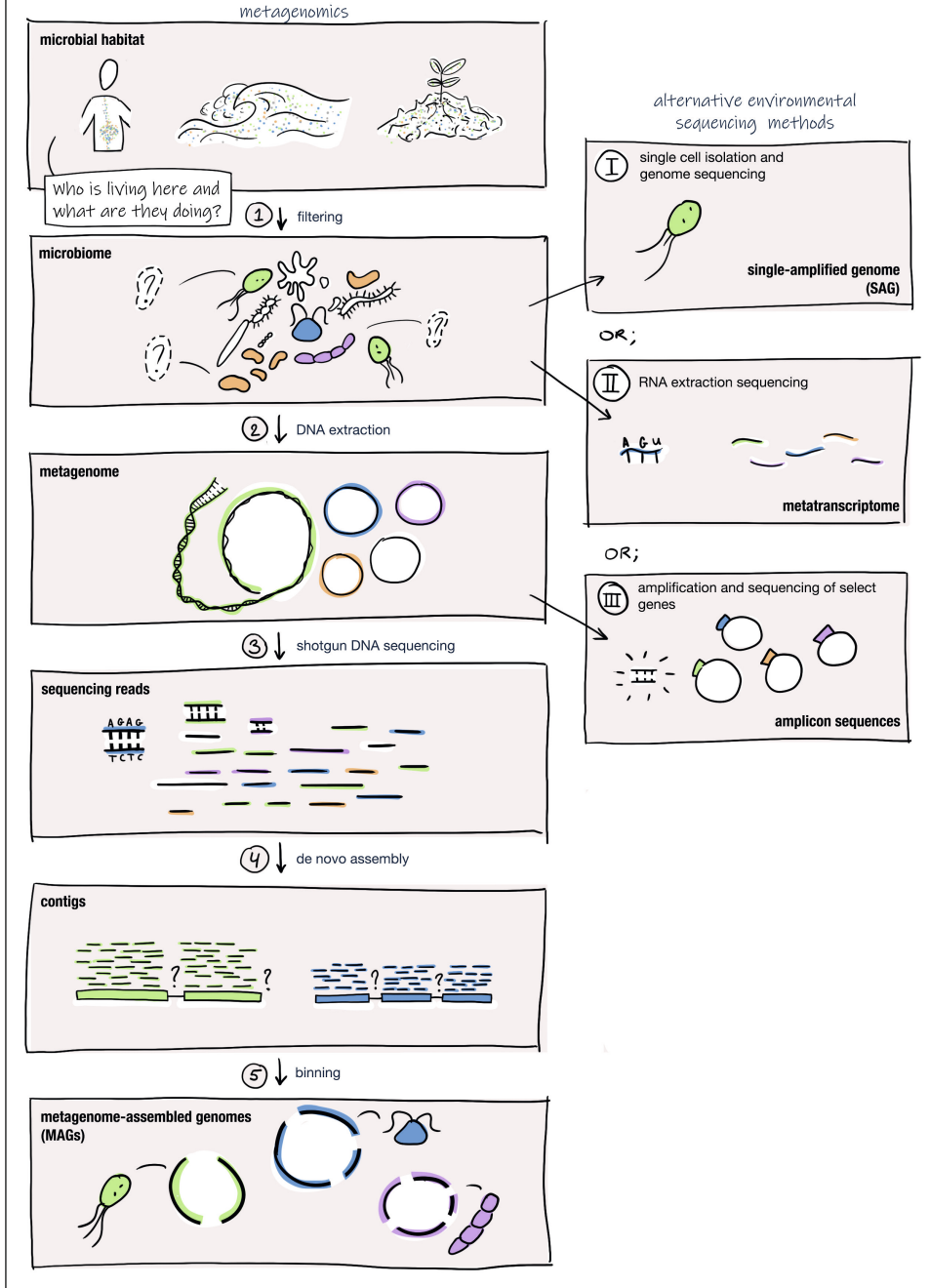
Most life is unseeable in the lab—but can be uncovered by metagenomics

Since the late nineteenth and for most of the twentieth century, microorganisms were studied in the lab. Through pure and enrichment cultures, our view of the microbial world was one of isolation. A microorganism taken from its natural habitat had to

be domesticated to ensure high clonal population numbers for further observation. Once in culture, it could be examined with a wide variety of lab-based techniques. Although very powerful and still an important tool in microbiology today, isolation limits analyses to those microorganisms that grow in the lab, while it was realised that many do not (36,37).

Direct sequencing of genetic information from the environment (**Box 1**) established the extent of the cultivation bias. It let us see microorganisms not under a microscope or in the lab but via their DNA and in their natural habitat, revealing who they are and what they can do. Over the past few decades, environmental sequencing has altered our view of the tree of life, by uncovering the vast microbial diversity that the cultivation bias left unseen. The majority of bacteria and archaea are currently uncultured, and only in the human microbiome do cultured genera dominate (38). ~97% of all cultured species fall within one of four phyla (Bacteroidetes, Proteobacteria, Firmicutes, and Actinobacteria) (39), even though at least 99 prokaryotic phyla have now been proposed (40) and sequence databases may contain over a thousand (41). Of the 27 proposed archaeal phyla, just six are cultured (42).

Box 1 | Environmental sequencing



Amplicon sequencing and metagenomics. All DNA of a microbiome is its metagenome. After a filtering step ① that selects microorganisms of a certain size and concentrates cells, the metagenome can be extracted ② and sequenced. Only a fraction is sequenced by classical methods that target specific genes via amplification (iii), like the prokaryotic gene that encodes 16S ribosomal RNA (rRNA) (43,44) and that is widely used as a taxonomic marker (45–47). While giving a general taxonomic overview of the community, amplicon sequencing of the 16S rRNA gene ignores non-targeted DNA sequences and the functional potential that they encode. More comprehensive than the reductionist gene-centric approach, DNA can be sequenced at random in an untargeted fashion ③, giving an overview of the complete metagenome (48). This ‘whole metagenome shotgun sequencing’, or metagenomics in short, provides our most unbiased view of the microbial world. No metagenome has been sequenced in its entirety yet—sequencing more DNA, referred to as deeper sequencing, will still return new information.

Assembly and genome-resolved metagenomics. The sequencers that are used in most current metagenomics studies use the Illumina sequencing technology (49) and do not sequence full-length DNA molecules but instead generate millions or even billions of randomly sequenced pieces of DNA with a maximum length of a few hundred base pairs. If multiple (almost) identical DNA molecules are present in the metagenome, as is often the case for microorganisms that clonally replicate, the random nature of the sequencing process ensures that most parts of the genome are covered by multiple short sequencing reads, and the sequence of the composite DNA molecule can be reconstructed with de novo sequence assembly ④. Due to sequencing errors, repeats, strain-diversity, and low-covered regions, the resulting contigs often do not span the entire genome. Contigs that come from the same microorganism can be binned ⑤, generating metagenome-assembled genomes (MAGs) (50) (see ‘**Pillar three: the age of computation**’). Quality of MAGs can be assessed with single-copy marker genes (51), and usually only medium- to high-quality MAGs (52) are further investigated, for example by phylogenomic placement in the tree of life and identifying metabolic pathways.

Typically, not all sequencing reads are assembled into contigs and not all contigs are binned into MAGs. The quality of assembled contigs and MAGs is related to sequencing depth, with the most abundant microorganisms having the highest coverage. Low-abundant microorganisms can thus be uncovered by deeper sequencing of the metagenome. New experimental protocols and sequencing technologies can also improve assembled contigs and MAGs, like proximity-based ligation of DNA with Hi-C and long-read single-molecule sequencing (see ‘**Discussion**’).

Reconstructing the original genomes that the short shotgun metagenomics sequencing reads derived from is like solving a puzzle, and the metagenome can be analysed at different steps in this process. For example, predicting function can be done at the level of sequencing reads, assembled contigs, and MAGs. The MAGs give a precise picture, being closest to the original genomes. Metabolic and biosynthetic pathways that the MAGs encode can be investigated on a genomic-wide scale. With the MAGs taxonomy and function can be linked: “who is doing what?”. However not all sequencing reads can typically be binned, and thus the MAGs provide a limited picture of the microbiome. The most comprehensive picture is that of the sequencing reads, representing all sequence data, but they miss the genome context that confidently links taxonomy to function. Contigs hold the middle ground.

Amplicons and metagenomics show different microbial abundances. The community composition that 16S rRNA gene amplicon sequences show differs from that shown by metagenomics. Detection by amplicon sequencing is dependent on primer specificity, the polymerase chain reaction (PCR) generates artifacts and biases (53,54), and different microorganisms have a different number of copies of the 16S rRNA gene (55–57). In addition, metagenomics reflects the complete genome sequence and large genomes are thus overrepresented in the sequencing reads compared to small genomes with similar cell abundance (58). Although metagenomics provides a less biased view of the microbiome than amplicon sequences, biases also arise in filtering, DNA extraction, library preparation, and sequencing (54,59–62).

Metatranscriptomics. Sequencers sequence DNA. However, RNA in the microbiome can be reverse transcribed into complementary DNA (cDNA) and sequenced (ii). Where the metagenome shows the genetic potential of the microbiome, the metatranscriptome—preferably enriched for messenger RNA—depicts its gene expression at that time (63). New sequencing techniques allow for the direct sequencing of RNA without reverse transcription (see ‘Discussion’).

Single-cell sequencing. In an alternative approach to environmental genomics, an individual cell can be isolated from the environment and its genome sequenced (64,65) (i). Where MAGs are ‘composite’ genomes based on a population of (almost) identical genomes, in single-cell sequencing the genome is amplified via whole-genome amplification to get enough copies for shotgun sequencing and genome assembly. Single-amplified genomes (SAGs) can be reconstructed from rare microorganisms in highly diverse microbial communities like soils, that are difficult to capture with metagenomics. SAGs are generally smaller and less complete than MAGs (66).

In addition to uncovering uncultured microorganisms, environmental sequences provide an in situ view of the microbiome. No microorganism lives in isolation. Interactions within complex microbial communities, and between microorganisms and their host and the environment, can be observed in the metagenome (25,67–71).

Environmental amplicon sequencing studies in the 1990s paved the way for metagenomics. The current state of the metagenomics research field—the motivations and innovations that advanced it, and its recent breakthroughs—is seen to take shape in three unrelated articles published at its dawn in 2004 (refs. 32,34,35). I will discuss these articles in the light of today below. They foreshadowed what was about to come, and represent three pillars of this thesis: (i) the motivation to circumvent the cultivation bias and discover novel biology, (ii) the generation and (re)analysis of large-scale data, and (iii) the development of algorithms and tools to interpret that data.

Pillar one: a less biased view of the microbial world results in discovery of the vast unknown

In 2004, Jo Handelsman published an article reviewing metagenomics (32). Pioneering the biosynthetic potential of soil microbiomes, she had coined the term

‘metagenome’ 6 years prior (72), giving name to the then-emerging technique. The review article started with a history of the realisation that most microorganisms are uncultured (**Box 2**), which was by that time well-established (36,37,73,74). It showed a phylogenetic tree based on 16S rRNA gene sequences, containing both established phyla with cultured members, and numerous candidate phyla that were only known from environmental sequences. Handelsman concluded that metagenomics would uncover those vast microbial unknowns beyond their 16S rRNA gene sequence, providing ecological insights by elucidating function.

Box 2 | the cultivation bias

The cultivation bias exists in part because certain habitats are more readily explored than others, or attract more attention. For example, the microbial habitat perhaps most distant to us is the deep subsurface, and contains an immense number of microorganisms. Dark, under high pressure, and low in energy and nutrients, the deep biosphere extends kilometres below the surface down to the upper temperature limit of life at around 120 °C (75,76). Metatranscriptomics reveals that microorganisms in the deep biosphere are active (77), but generation times may be months to more than 100 years (78). Due to its vast extent, it is likely one of the four largest reservoirs of prokaryotes on Earth, with oceans, soils, and upper marine sediments (1,79), although quantitative estimates are understandably debated (80–82). Culturing efforts have been directed to habitats that are easier to reach, and to those microorganisms with practical potential, for example in medicine or biotechnology.

However, abundant microorganisms in familiar habitats also escape cultivation. Substrate or growth conditions can be difficult to identify or replicate in the lab, cells may be dormant and not readily resuscitated, microorganisms can depend on others via mutualism or syntrophy, or slow-growing microorganisms are outcompeted by faster growing ones (39). For example, one of the most abundant organisms in the waters of the open ocean, the extremely oligotrophic *Pelagibacter*, was only cultivated in 2002 by supplying its media with very low nutrient concentrations (83). The first microorganism of the closely related LD12 clade that is highly abundant in fresh-water environments was cultivated in 2018 (ref. 84).

The MAGs generated by genome-resolved metagenomics can provide guidance for culture attempts (85). Microorganisms of the deep biosphere discussed above are difficult to culture once they do reach the lab, and metabolic predictions based on MAGs have been used to get a pure culture of thermophilic spirochetes collected in a 2 kilometre deep aquifer in Siberia (86).

Cultivation of the first microorganism of the Asgard superphylum, a close sister clade to eukaryotes and thus important for understanding eukaryogenesis, was reported in 2020, after 12 years of enrichment since sampling from a marine sediment core (87)—the authors unaware during most of that time of its evolutionary significance. This archaeon (*Candidatus* Prometheoarchaeum syntrophicum) and a second cultured Asgard archaeon (*Candidatus* Lokiarchaeum ossiferum) (88) both live together with syntrophic partners, and surprisingly have surface protrusions and constrictions that are compatible with an earlier proposed hypothesis for eukaryogenesis, the ‘inside-out’ model (89). Asgard archaea are not restricted to extreme environments and they are ubiquitous. They had been missed thus far because they are not abundant, do not cause disease, and have slow generation times (90). ‘*Ca.* P. syntrophicum’ divides every 14–25 days, and ‘*Ca.* L. ossiferum’ every 7–14 days.

Metagenomics has shaken up our understanding of the tree of life

Over the years, metagenomics has been strongly driven by the identification and characterisation of previously unknown taxa, including a wide range of newly proposed very deep taxonomic clades. Genome-resolved metagenomics (**Box 1**) allows for the identification of a microorganism and for insights into its functional potential via its genome sequence (91). MAGs have brought a wealth of surprises. For example, the bacterial ‘Candidate Phyla Radiation’ (CPR) was discovered in 2015 and may represent 15% of bacterial diversity with at least 65 proposed phyla (92,93), although that number is possibly inflated (40). CPR genomes are small and lack common biosynthetic pathways—for example for nucleotides and amino acids, suggesting a symbiotic or parasitic lifestyle. Because their 16S rRNA genes are so different from those of other taxa, they evade detection in 16S rRNA gene surveys with common primers (92). The DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaea) superphylum represents a large part of archaeal diversity and also has small genomes indicative of a symbiotic or parasitic lifestyle (64,94–96). Various CPR and DPANN genomes elusively lack homologs of known membrane lipid biosynthesis genes (97,98), and they likely receive membrane lipids from host cells (99,100). The Asgard superphylum, the closest known prokaryotic sister clade to eukaryotes that firmly places eukaryotes within the archaeal domain, was uncovered in 2015. Its first described member was identified via 16S rRNA gene surveys in marine sediment near a hydrothermal vent field at over 3 kilometres water depth, and uncovered via genome-resolved metagenomics (101). Other Asgard MAGs were subsequently found in a wide variety of natural environments (102,103).

These and other recently discovered MAGs have forced us to re-evaluate our view of the tree of life—the tree spanning a much larger diversity than thought with new deep branches at unexpected places and surprising biological traits (97,104,105). Their discovery has led to a substantial revision of prokaryotic taxonomy based on genome phylogeny (40).

Metagenomes contain uncharted sequences with different levels of unknownness

An enormous part of microbial sequence space is thus only reachable with metagenomics. From the perspective of a single metagenome, this means that a large fraction of the DNA is often still uncharted. Sequencing reads may derive from microorganisms of clades that have never been seen before, for example from a novel genus or family, or that are only known from 16S rRNA gene surveys. In addition, microbial genomes are highly flexible, with large variations in gene content even between strains of the same species, as reflected in strain-specific ‘accessory’ genes (106,107), and parts of a newly discovered microbial genome may be uncharted even if it belongs to a previously encountered species. Conversely, microbial genomes of unknown clades can contain sequences that are very similar to sequences that

are present in known microorganisms due to evolutionary conservation, HGT, or shared low-complexity regions. Some metagenomic sequencing reads may thus be very similar to known sequences, whereas others are much more diverged. Some sequencing reads do not have detectable homologs at all (108).

Over 80% of species in soils of Central Park in New York City are unknown (109). This assessment was based on the 16S rRNA gene that is much more widely sequenced than other genes (**Box 1**), and uncharted genetic diversity is thus even higher—some of the species that are known from 16S rRNA genes do not have other parts of their genome sequenced. Of just the dominant bacterial species in soils, that are ubiquitous and abundant in soils worldwide, 42% have no known relatives even within the same taxonomic family (110), again based on the 16S rRNA gene and thus underestimating uncharted genetic diversity. An analysis of 339 metagenomes from diverse environments showed that 40% of predicted proteins did not have known homologs with more than 50% sequence identity (111).

Assembly and binning illuminate the identity of uncharted sequences

This uncharted genetic diversity poses a problem when assigning taxonomy (“who is living here?”) to sequencing reads in a metagenome, as this is done by sequence similarity to known sequences in reference databases. When a short sequencing read of a few hundred base pairs only has distantly-related known homologs or no detectable homologs at all, it may represent an unknown microbial clade, or an uncharted region of a genome in an already known clade. Conversely, when a sequencing read does have closely-related known homologs, it may represent a microorganism from the known clade, a sequence that is evolutionary conserved across clades, or a sequence that was transferred via HGT from a known microorganism into an unknown microbial clade.

Assembled contigs and MAGs contain much more taxonomic information than the individual sequencing reads. A region of DNA with low or no sequence similarity to known sequences may reside on a contig or MAG with genes that are more closely related to sequences in a reference database, and a region of DNA that is shared by many microorganisms may be surrounded by sequences with more specific association. The identity of uncharted sequences can thus be more confidently assigned to longer than to shorter sequences.

If a MAG contains ‘marker’ genes, they can be used to infer taxonomy. Select marker genes are conserved in single-copy across clades and have been inherited primarily via vertical descent—with few events of HGT (112). Phylogenetically informative marker genes that are universal for prokaryotes encode primarily ribosomal proteins and RNA polymerase domains (51), and concatenated they allow for confident phylogenomic placement of a MAG in a tree with known genomes (16), even if

the rest of its genes are unknown. Marker genes that are not universal but specific for a certain clade are also used for phylogenomic placement (98). Phylogenomic trees based on concatenated marker genes are the gold standard for taxonomy assignment, and were used to place the previously discussed MAGs in the tree of life and assign novel taxonomy. However, the required multi-sequence alignment and tree inference are computationally demanding processes, and manual interpretation can be laborious, especially with many genomes.

Not all environments are equal

The fraction of uncharted sequences and their level of unknownness differs per microbiome, certain habitats being more vigorously explored or readily characterised than others. Human-associated microbiomes, especially from the gut, are relatively well-characterised, as are marine microbiomes (see '**Pillar two: the age of data**'). Aquatic environments and soils are relatively underexplored. Although soils have been researched extensively, their very high taxa richness compared to other habitats makes their characterisation with metagenomics a challenge (113,114). High taxa richness leads to low relative abundance of many microorganisms in the sequenced metagenome. Because the quality of assembled contigs and MAGs depends on sequencing depth of a microbial genome (**Box 1**), soil microorganisms are difficult to uncover with genome-resolved metagenomics (115). Microbial communities that thrive in conditions where no other survive, the so-called extreme environments, have low taxa richness, and they are largely known (116,117).

Microbiomes from habitats that are thoroughly studied still contain surprises. The bacteriophage crAssphage is present in $\frac{3}{4}$ of human guts, represents 1.68% of the metagenomic sequencing reads, and was only uncovered in 2014 (ref. 118). Close relatives of crAssphage are found in Old-World and New-World primates, indicating that this bacteriophage has been with us for a long time (119). In 2019, 92,143 bacterial MAGs were reconstructed from human gut metagenomes (120), increasing the then-known phylogenetic diversity of the gut microbiome with 281%.

Uncharted sequences can only be uncovered by exhaustive sampling of all microbial habitats, and their discovery still represents an important motivation for the field today.

Pillar two: the age of data

In 2004, the metagenomes of hundreds of litres of surface water from the Sargasso Sea were reported, after size-based selection for microorganisms on filters (34). It represented the first large-scale metagenomic study of complex microbial communities. The size of the dataset, assembled contigs with a combined length of over a billion base-pairs, was unheard of at the time. The 1.2 million previously unknown gene sequences that were recovered from the Sargasso Sea almost doubled

the size of public protein databases, leading to sequence similarity searches primarily returning proteins with unknown function for a while (121).

Next-generation sequencing technologies have accelerated data output

Thousands of metagenomic studies have been conducted since then, and public sequence databases have grown explosively (122). The advance of a new generation of sequencing technologies was an important contribution. Sequencing of the Sargasso Sea metagenomes was done with automated cloning-based Sanger sequencing (123), the same technology that was used to elucidate the first bacterial and human genome sequences (124,125). Next-generation sequencing (NGS) technologies, like 454 sequencing (126), SOLiD sequencing (127), and the now dominant Illumina sequencing (49), generate a massive amount of sequencing reads at high speed and low cost (128,129). As sequencing costs dropped and more scientists got access to metagenomics, a wide range of well-studied and thus-far unexplored habitats got sequenced ever more deeply.

Human-associated, marine, and other microbiomes

Early in line to sequence were human-associated metagenomes. The Human Microbiome Project (130,131) found that even though taxonomic composition within body sites differs between humans, metabolic potential is largely similar (132). The Metagenomics of the Human Intestinal Tract (MetaHIT) consortium first assembled 3,3 million microbial gene sequences from gut metagenomes (133), tripling that number in a subsequent study (134). Gut microbial communities separated in three distinct clusters, which suggested a limited number of balanced states for the healthy human gut microbiome (135), although the existence of these ‘enterotypes’ was later questioned (136). Human-associated metagenomes have been and are being extensively sequenced, in relation to disease (137,138), and in cohort-studies that follow human subjects and their microbiome during their lives and provide insights into the connection between microorganisms and host-associated factors like diet. Thus far, host-associated factors seem to influence microbiome composition to a limited extent (139,140).

After the first Sargasso Sea study, the Global Ocean Sampling expedition followed, the ‘Sorcerer II’ sailing boat circumnavigating the world to characterise marine microbiomes (141,142). The sailing boat ‘Tara’ of the Tara Oceans expedition subsequently set sail in 2009, generating more and more deeply sequenced metagenomes, and dared polar regions (143). The Tara Oceans expedition uncovered over 40 million gene sequences in 234 samples from epipelagic and mesopelagic waters, four times more than then-known from the human gut (144). This dataset showed that gene families that are ubiquitous in both the ocean and the human gut represent at least 73% of the sequenced ocean metagenomes, revealing a large common functional core between these very different habitats (144).

Chapter 1

Numerous human-associated and marine metagenomes have been sequenced over the years, and various other host-associated and free-living microbiomes, including the cow rumen (145), the coral holobiont (146), honey bee colonies (147), the gills of shipworms (148), mangrove forests (149), deep sea hydrothermal plumes (150), the floating fern *Azolla* (151), aquifer sediments and groundwater (93), and topsoil from around the globe (114), and many more. The MGnify database (152), that gathers environmental sequencing studies and is hosted by the European Bioinformatics Institute (EMBL-EBL), contains at the moment I write this (February 2023) metagenomic samples from 142 different annotated biomes, both natural and human-made like activated sludge from wastewater.

All data are equal

Upon publication, sequences become available in public repositories. Sequencing reads, assembled contigs, MAGs, genes, and translated proteins, should be deposited in databases such as the European Nucleotide Archive (ENA) (153), GenBank (154), or the DNA Data Bank of Japan (DDBJ) (155), together with associated metadata. The NCBI non-redundant protein database (nr) (156) contains all unique protein sequences that have been deposited with taxonomic annotation—as supplied by the dataset submitters and potentially wrong (157–161), by gathering (translated) protein sequences from curated and non-curated databases like the Protein Data Bank (162), RefSeq (163), and GenBank.

All databases are free to explore, to compare newly discovered sequences or to search for biology that was not detected by the original authors. Metagenomics generates so much data in single studies that often only a fraction is thoroughly investigated at first, the rest deposited unaddressed (108). For example, the Sargasso Sea assemblies were reanalysed by others to investigate picoeukaryotes, that are so small that they ended up in the size filter fraction with prokaryotes (164). A habitat of interest can be more deeply analysed, and microbiomes from different habitats can be compared. Early comparisons between microbiomes from distinct environments revealed habit-specific functions that allowed for the separation of environments based on metabolic composition (165,166), and it was later shown that microbial communities can be used as biomarker, by reflecting biochemical conditions (167), the health status of the host (168–170), or metabolites in the microbiome (171,172).

Reanalysis of sequencing data from different studies

The microbial composition of various habitats can be assessed based on the metagenomic sequences, and even more habitats have been taxonomically characterised with amplicon sequencing of the 16S rRNA gene, for example by the Earth Microbiome Project (173). This allows for a change of view from the microbiome to microbial taxa: “where are they living?”

Databases that aggregate sequences and reanalyse them in a uniform way are particularly useful for large-scale comparisons across studies. The previously discussed MGnify database contains taxonomic and functional information that is generated by running standardised pipelines on raw metagenomic, metatranscriptomic, and amplicon data (152). This reanalysis removes at least one source of noise between different sequencing studies—the various ways in which sequences can be annotated, for example with different bioinformatic tools and different reference databases. The MGnify database also generates new assemblies and collects existing ones. In addition, metadata annotations are readily accessible. However, even though metadata are important for the interpretation of microbiomes and standards have been developed to improve data sharing (174), metadata are still often ambiguous, redundant, or wrong (175,176).

When a novel microorganism or genetic trait is discovered, databases like MGnify can be searched to expose it in other places and habitats.

Reanalysis of genomes from different studies

Genome-resolved metagenomics studies can generate a massive number of high-quality genomes, in newly sequenced datasets or by reanalysis of earlier published data. For example, 283 MAGs were binned from the MetaHIT data (177), 913 from the cow rumen (145), 469 from the chicken caecum (178), 1,529 from thawing permafrost (179), and 2,540 from aquifer sediments and groundwater (93). 957 MAGs were binned from Tara Oceans data (180), followed by another 2,631 based on a larger set of samples (181). Data from 75 different studies were used to bin 92,143 MAGs from the human gut (120), and a study that specifically looked in different habitats besides the human gut reconstructed 7,903 (ref. 105). In 2021, then-known prokaryotic phylogenetic diversity was expanded by 44% by the binning of 52,515 MAGs from diverse habitats (182).

At the dawn of the genome sequencing age, concerns were raised that genome sequencing was directed to ‘important’ model organisms, but should instead represent a large phylogenetic diversity (183,184). MAGs (and SAGs, see **Box 1**) from across the tree of life now fill genome databases. The PathoSystems Resource Integration Center (PATRIC) database, now part of the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), collects genome sequences from cultured isolates, MAGs, and SAGs, and uses standardised pipelines for uniform functional annotation (185). In addition, metadata annotations are readily accessible.

When a novel genetic trait is discovered, genomic databases can be searched for the trait to uncover it throughout the tree of life. For example, high codon-usage bias in highly expressed genes like those coding for ribosomal proteins is associated with fast maximal growth rates (186). When maximal growth rate was predicted for

Chapter 1

217,074 genomes from cultures, MAGs, and SAGs, fast-growing microorganisms were shown to be overrepresented in culture collections (187), highlighting one of the reasons for the cultivation bias (**Box 2**).

Very large datasets are part of the foundation of metagenomics today. Data generation is ongoing, and their (re)analysis drives discovery. New algorithms and tools to handle and interpret that data are continuously developed.

Pillar three: the age of computation

In 2004, the in-depth analysis of a metagenome from a biofilm was published (35). The pink biofilm was growing in extremely acidic conditions (pH 0.83) amongst toxic metals in a mine in California. The low complexity of the community, just three bacterial and three archaeal species according to 16S rRNA gene sequencing, allowed for the separation of the assembled contigs into ‘bins’. This was the first time that multiple MAGs were reconstructed from a metagenome.

The contigs were binned based on features that were shared because they derived from the same microorganism—DNA G+C content and sequencing read depth. G+C content is relative stable across a microbial genome (188) so contigs derived from the same microorganisms have similar G+C content. Because the contigs derive from the same number of genome copies in the metagenome, they are also covered by a similar depth of sequencing reads. The different species in the pink biofilm had distinct G+C content and read depth from each other, which allowed for their separation. Because G+C content and read depth are intrinsic properties of the genome, binning is a *de novo* method that allows for recovery of MAGs from unknown microorganisms.

Binning has matured

MAGs are now binned from complex microbial communities in an automated fashion, thanks to development of new algorithms and tools. Current *de novo* binning methods still use DNA sequence composition and read depth. Separation of contigs based on G+C content and read depth however is not very discriminative—multiple microorganisms in the microbiome may have similar G+C content and abundance. For example, the ‘low G+C content 3x coverage’ bin from the pink biofilm consisted of two strains that were further separated based on sequence similarity to a known genome that was closely related to one of the strains. G+C content has made way for DNA signatures with higher information content and discriminatory power, like tetra-nucleotide frequency (TNF) (189–192), or *k*-mer frequency of other lengths (193). Separation based on read depth can be greatly improved if multiple samples are considered (194–196)—contigs from the same microorganisms have similar read depth across samples. Where early binning methods involved manual steps and custom software (192,195), fully automated methods that combine *k*-mer frequency and differential read depth across samples are now commonly used (193,197–200).

The development of new algorithms and automated binning tools has, together with deeper sequencing, advanced binning in complex microbial communities and of low-abundant microorganisms. However, short contigs (<2,500 base pairs) are still difficult to bin, even from abundant microorganisms, because k -mer frequency and read depth can be highly variable for short contigs compared to long contigs on which these signals are evened out.

Quality estimates of MAGs

Even with modern binning tools, quality of reconstructed MAGs varies considerably. Some MAGs represent (almost) complete genomes, whereas others are only partially complete or contain contigs from multiple genomes. MAG completeness and contamination are estimated by the presence of single-copy marker genes. A commonly used tool that automates quality estimation is CheckM (51). After phylogenomic placement in a tree with known genomes, the MAG is searched for clade-specific sets of single-copy marker genes. These genes are present in single-copy in most known genomes from the same clade, and a related novel genome is similarly expected to have all of them once. Clade-specific marker genes are more numerous than universal marker genes and therefore allow for more robust quality estimation. ‘Completeness’ is reported as the fraction of expected marker gene sets that are found, and ‘contamination’ as the fraction of the marker gene sets that are present as multiple copies suggesting that the MAG contains contigs from multiple organisms. It is common procedure to estimate MAG quality with CheckM, and usually only medium- to high-quality MAGs (completeness $\geq 50\%$, contamination $< 10\%$; see ref. 52) are reported.

Algorithms and tools for the age of metagenomic data

The advance of NGS technologies required advances in data analysis and handling. Assembling the enormous amount of short sequencing reads with a traditional read overlap approach is computationally demanding, and modern assemblers now use de Bruijn graphs of k -mers in sequencing reads (201–203). Metagenome assemblers are designed to reduce chimeras as sequencing reads from different microorganisms can be spuriously assembled together (204), and some focus specifically on time- and cost-efficiency (205). The fragmented nature of metagenomic assemblies poses challenges for conventional gene prediction, and the metagenomic gene caller MetaProdigal identifies translation initiation sites and alternate genetic codes based on a diverse set of training genomes (206).

Mapping short sequencing reads to contigs is commonly done with aligners that use the Burrows-Wheeler Transform (BWT) (207,208), which are very fast and memory-efficient but do not allow for a high number of mismatches. For distant homology searches, algorithms that are more computationally intensive but allow for sequence variation are required, like the DIAMOND (translated) protein aligner, that is slower

than BWT mappers but is still 20,000 times faster (in translated protein mode) than a conventional BLASTX search (209).

Benchmarking tools that deal with uncharted sequences

Testing the performance of metagenomic tools is challenging because they deal with uncharted sequences, and we cannot readily confirm that their results, for example the contigs that they assemble or the taxonomic annotation that they assign, are correct. Simulated datasets have been used for benchmarking (210,211), and especially the Critical Assessment of Metagenome Interpretation (CAMI) challenges have formalised benchmarking methods. Benchmarking methods widely differ between publications that introduce new tools. The first CAMI challenge simulated metagenomes from novel microorganisms and tested performance of commonly used assemblers, taxonomic profilers, and binners (212). Datasets from the second CAMI challenge were more complex, and more tools competed (213). One of the benefits of the CAMI challenges is that different tools are run on the same datasets, allowing for a fair comparison between the tools.

In addition to the development of new algorithms and tools that deal with metagenomic data, new statistical methods are invented, for example for the inherent compositional nature of environmental sequencing datasets (214). Computation, in different forms, is part of the foundation of metagenomics today, representing both necessity and opportunity.

Synopsis of this thesis

As explained above, environmental sequencing is transforming microbiology. Still early in the age of massive data, metagenomic studies are often explorative, discovery-driven rather than focused on testing hypotheses. After all, who knows what can be found where nobody has looked before? New habitats are explored and MAGs of microorganisms with unexpected traits challenge current biological teachings. Data are generated at an incredible pace, and surprising results come up faster than they can be tested in the lab. Algorithms and tools to interpret metagenomes and handle the data are still being invented. In the meantime, all sequences—sequencing reads, assembled contigs, MAGs, genes, translated proteins—become freely available upon publication in public databases. Usually only a fraction is analysed in the original publication, waiting to be further explored by others.

The chapters of this thesis fall within this novel microbiological tradition that revolves around environmental sequences, and its three pillars: (i) the motivation to circumvent the cultivation bias and discover novel biology, (ii) the generation and (re)analysis of large-scale data, and (iii) the development of algorithms and tools to interpret that data. The chapters describe the generation of data and their

interpretation (or sometimes of only a fraction), the discovery of unexpected biology, the development of novel algorithms and tools that deal with what we do not know, and the large-scale exploration of data published by others. In some chapters I collaborated with experimentalists, where lab and computer work complement each other, but most of the research in this thesis was done *in silico*. I analysed shotgun metagenomes, amplicon sequences, shotgun metatranscriptomes, and genomes. In all chapters, the focus is on prokaryotes.

In **Chapter 2**, I use genome-resolved metagenomics to reconstruct MAGs of bacteria and archaea living in the water column of the Black Sea. Deep down, at 2 kilometres water depth, I find an elusive group of bacteria whose genomes contain an unexpected set of genes, challenging a once-thought fundamental divide across the tree of life in the molecular composition of the cell membrane. The lab being so alien to an organism that is taken from its natural habitat close to the bottom of the Black Sea, culture attempts understandably failed—even if guided by the MAGs, preventing direct analysis of their membranes for confirmation. However, measurements of lipid molecules in the environment support these unexpected findings. Chapter 2 illustrates how genome-resolved metagenomics advances our understanding of the microbial world by uncovering novel traits in thus far uncultured organisms. It is one of many discovery-driven metagenomics studies of today. The tools and concepts of assembly and binning discussed in Chapter 2 will return in other chapters.

The CAT pack software suite is introduced in **Chapter 3**, tools for the taxonomic annotation of contigs (Contig Annotation Tool, CAT) and MAGs (Bin Annotation Tool, BAT). The CAT pack excels when annotating uncharted sequences, that dominate microbiomes from underexplored habitats and still surface in those well-studied. Where most studies thus far assign taxonomy to contigs based on the best hit in a reference database, CAT integrates taxonomic signals from all predicted proteins on the sequence, thereby greatly increasing annotation precision of unknown sequences. The same algorithm is applied by BAT, that automates MAG annotation as a first explorative step before phylogenomic placement. A large part of tool development is benchmarking, and Chapter 3 deals with the question how to test performance of a taxonomic classifier that is meant to annotate unknown organisms. Furthermore, it shows that current bioinformatic tools depend on other tools and on publicly available reference databases.

Chapter 4 turns to the taxonomic annotation of metagenomic sequencing reads and the reconstruction of taxonomic profiles, the most complete view of a microbiome. A new member is added to the CAT pack, Read Annotation Tool (RAT). Taxonomic profiles have previously been reconstructed by directly mapping individual sequencing reads or even shorter k -mers to a reference database, and while accounting for a large part of the data these profiles contain many false positives because of the

Chapter 1

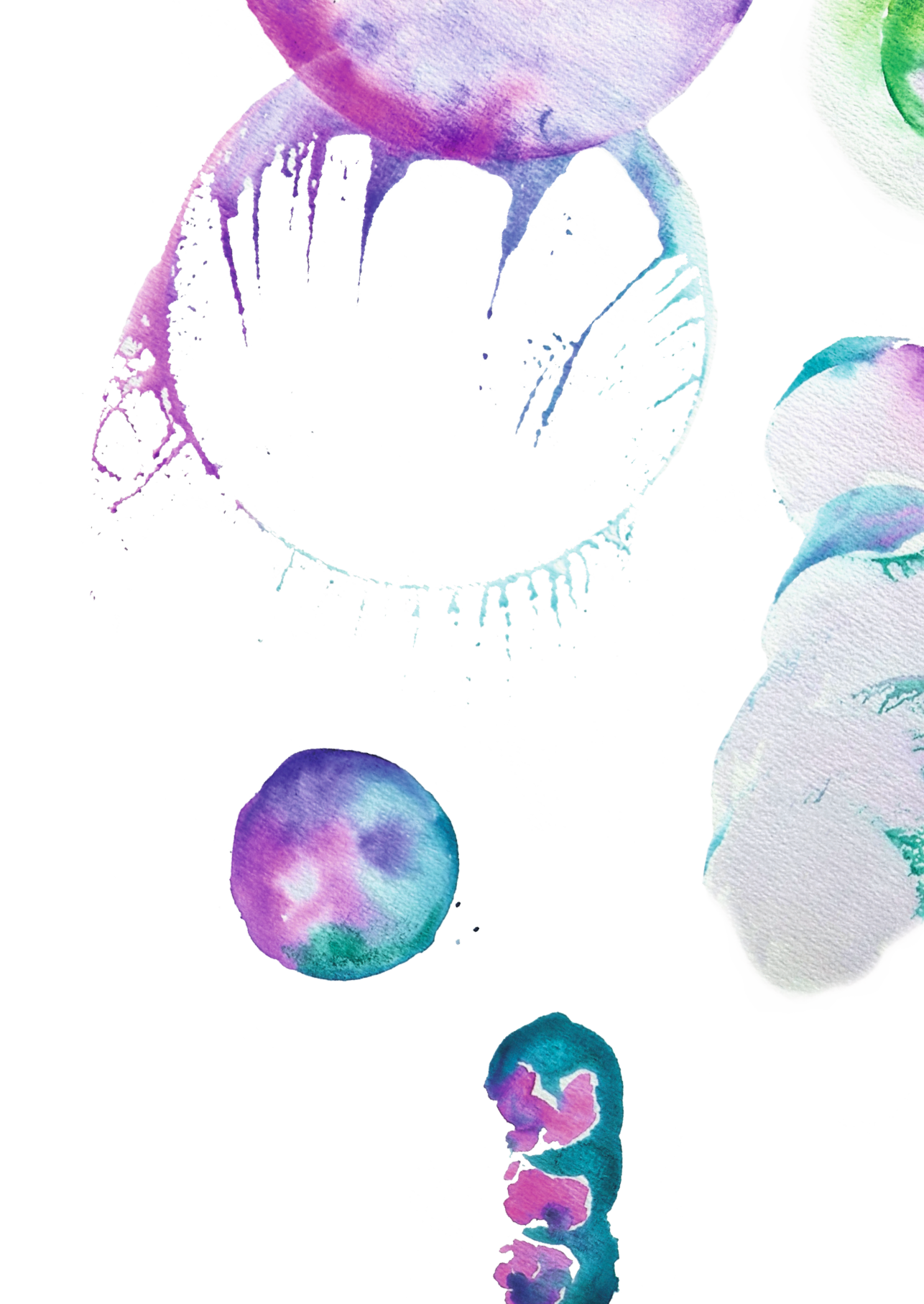
limited taxonomic information contained in short sequencing reads. RAT instead uses the much more robust taxonomic annotation of MAGs and contigs that BAT and CAT provide to assign the most reliable annotation to each read, and only resorts to direct individual read mapping when sequencing reads cannot be associated with a MAG or contig. The result is a taxonomic profile that is both comprehensive—containing as much of the data as possible, and accurate. With RAT, I show that taxa richness in microbiomes is likely overestimated by conventional taxonomic profilers. In addition, a new reference database is added to the CAT pack, allowing for assignments to the Genome Taxonomy Database—the previously discussed revised taxonomy based on genome phylogeny, in addition to the more traditional NCBI taxonomy database. Chapter 4 unites genome resolved-metagenomics and read-based analysis methods.

In **Chapter 5**, I explore over 22 thousand taxonomic profiles across a wide range of habitats and geographical locations, that are based on environmental sequencing projects published by others. Chapter 5 contains a map that shows how all microbiomes on Earth relate to each other. With this dataset, I develop a score that quantifies the niche breadth of a microorganism, the range of conditions in which it lives. Rather than defining microbial habitats a priori with often incomplete metadata, I define the ‘social niche’ of a microorganism as the communities it is found in, building upon observations that the microbiome itself can be a sensitive biomarker. Next, I quantify the range of communities in which a microorganism is found. Using this ‘social niche breadth’ score, I address the distinction between specialists and generalists, and search for associated ecological and evolutionary strategies. While specialists are often assumed to dominate local communities, the data show they are outcompeted by generalists, and a surprising habitat-dependent relation between genome size and niche breadth turns up. Combined sequencing data from vastly different study designs and even experiment types is inherently noisy. Chapter 5 shows that relevant biological signals can be extracted nonetheless.

In **Chapter 6**, structural colour in bacteria is investigated. Structural colour is the striking display of changing colours depending on the angle of observation, as seen in iridescent butterfly wings, and can be a colony phenotype of bacteria that emerges from cell organisation. Only a couple bacteria are known to produce structural colour, and little is known about the genes involved. In Chapter 6, lab work and genomics are combined to identify genes involved in structural colour. A structural colour classifier is built, that can predict the phenotype based on genome sequence. I classify all publicly available bacterial genomes (including MAGs and SAGs), and show that structural colour is more widespread than previously thought. In addition, I use the classifier to identify potential structural colour across a wide range of habitats, targeting metagenomic assemblies. I find structural colour common in the environment, and curiously present in the deep ocean where no light penetrates.

Chapter 6 is an example of a fascinating microbial world that is still largely left unexplored.

Together, these chapters paint a picture of current research on environmental sequences. Finally, in the concluding **Discussion** I will look beyond today's research and attempt a projection of the future. This will be my two cents on where the field is headed.





Chapter 2

Bridging the membrane lipid divide: bacteria of the FCB group superphylum have the potential to synthesize archaeal ether lipids

Laura Villanueva*, F. A. Bastiaan von Meijenfeldt*, Alexander B. Westbye, Subhash Yadav, Ellen C. Hopmans, Bas E. Dutilh[#], and Jaap S. Sinninghe Damsté[#]

* Laura Villanueva and F. A. Bastiaan von Meijenfeldt contributed equally to this work.

[#] Bas E. Dutilh and Jaap S. Sinninghe Damsté jointly supervised this work.

The ISME Journal **15**, 168–182 (2021)

Abstract

Archaea synthesize membranes of isoprenoid lipids that are ether-linked to glycerol-1-phosphate (G1P), while Bacteria/Eukarya produce membranes consisting of fatty acids ester-bound to glycerol-3-phosphate (G3P). This dichotomy in membrane lipid composition (i.e. the 'lipid divide') is believed to have arisen after the Last Universal Common Ancestor (LUCA). A leading hypothesis is that LUCA possessed a heterochiral 'mixed archaeal/bacterial membrane'. However, no natural microbial representatives supporting this scenario have been shown to exist today. Here, we demonstrate that bacteria of the Fibrobacteres–Chlorobi–Bacteroidetes (FCB) group superphylum encode a putative archaeal pathway for ether-bound isoprenoid membrane lipids in addition to the bacterial fatty acid membrane pathway. Key genes were expressed in the environment and their recombinant expression in *Escherichia coli* resulted in the formation of a 'mixed archaeal/bacterial membrane'. Genomic evidence and biochemical assays suggest that the archaeal-like lipids of members of the FCB group could possess either a G1P or G3P stereochemistry. Our results support the existence of 'mixed membranes' in natural environments and their stability over a long period in evolutionary history, thereby bridging a once-thought fundamental divide in biology.

Introduction

Lipid membranes are essential for all cellular life forms to preserve the integrity and individuality of cells, as well as having a direct influence in the maintenance of energy metabolism. Lipid membranes are also key in differentiating the domains of life. Bacteria and eukaryotes have membranes formed by fatty acids linked to glycerol-3-phosphate (G3P) via ester bonds, while archaea have membranes made of isoprenoid alkyl chains linked by ether linkages to glycerol-1-phosphate (G1P), leading to an opposite stereochemistry of the glycerol phosphate backbone (215). This segregation in lipid membrane composition, or 'lipid divide', has been hypothesized to have appeared early in the evolution of microbial life from the Last Universal Common Ancestor (LUCA), but the nature of the lipid membrane of LUCA and its subsequent differentiation in Bacteria and Archaea remain unknown. Some studies have proposed that a non-cellular LUCA lacked a lipid membrane (216–218). A recent study suggested that the membrane of LUCA was formed by fatty acids and isoprenoids without the glycerol phosphate backbone as a requirement to have a lower membrane permeability that could sustain a proton gradient (219). The most parsimonious hypothesis may be that LUCA had a heterochiral lipid membrane composed of both G1P and G3P together with fatty acids and isoprenoids (220,221), which later diversified into archaeal and bacterial membranes resulting in the 'lipid divide'. It was originally proposed that this differentiation may have been driven by heterochiral membrane instability (222,223), but heterochiral membranes are in fact stable (224) and a recent study has proven that they are, in some cases, more robust to environmental stresses (225).

Another critical issue in the 'lipid divide' is the membrane lipid composition of eukaryotes. Multiple lines of evidence indicate that eukaryogenesis encompassed an endosymbiosis event of a bacterial cell into an archaeal host (102,226–230). Thus, the bacterial-like composition of contemporary eukaryotic membranes implies that an early eukaryote had its archaeal-like membrane replaced by a bacterial-like one, possibly through a 'mixed membrane' intermediate containing both the archaeal membrane lipids with ether-linked isoprenoids to G1P and the bacterial ones with ester-linked fatty acids to G3P. This would imply that bacterial-like membrane molecules arose twice, in bacteria and in eukaryotes. The competing syntrophic hypothesis of eukaryogenesis (231–233) proposes that the host of the mitochondrial endosymbiont was a bacterium, avoiding the need for a transitional step from an archaeal to eukaryotic membrane. Some eukaryogenesis models also suggest that the membrane transition was facilitated by intensive bacterial lipid transfer from the endomembrane system (234). Because most models for the origin of eukaryotes require a membrane transitional step similar to the one expected in the 'mixed membrane' scenario for LUCA, it is striking that no remnants or natural microbial

representatives with a heterochiral ‘mixed membrane’ have yet been described that would support the viability of such a scenario.

The concept of the ‘lipid divide’ has been challenged by the identification of traits thought to be characteristic of archaeal membrane lipids in bacteria and vice versa, mostly restricted to specific taxonomic groups. First, some bacteria (and eukaryotes) produce ether-linked lipids (235–241). Second, membrane-spanning lipids, another trait thought to be specific for the archaea, also occur in bacteria (240–244). However, no bacteria are known to produce membrane lipids based on isoprenoidal side chains or possessing the ‘archaeal’ G1P stereochemistry. Third, some studies have reported that some archaea produce phospholipid fatty acids (e.g. ref. 245) but this may be due to contamination of the growth media (246). Nevertheless, an almost complete biosynthetic pathway for fatty acid synthesis is encoded by many archaeal genomes (247). Last, two uncultured archaeal groups, the Euryarchaeota ‘Marine Group II’ (currently known as ‘*Candidatus* Poseidoniales ord. nov.’ (248)) and *Candidatus* Lokiarchaeota of the Asgard superphylum, contain archaeal lipid biosynthesis genes alongside bacterial-like fatty acid and ester-bond formation genes, but seem to lack the capacity to synthesize the G1P backbone via glycerol-1-P-dehydrogenase (G1PDH), while they do have the genetic ability to produce G3P (249). This observation is exciting as the Asgard archaea are currently considered as the closest descendants of the archaeal ancestor leading to eukaryotes (101). However, a recent phylogenomic study of the orthologs of bacterial lipid genes present in Asgard archaea does not provide compelling support for an origin of eukaryotic lipids via an archaeal host cell (250). Moreover, there is no further evidence to date that the presence of these genes in those two archaeal groups actually leads to the synthesis of ‘mixed membranes’, i.e. membranes containing both ‘bacterial’ fatty acid ester-linked lipids and ‘archaeal’ isoprenoid ether-linked lipids, either heterochiral (containing G1P and G3P) or homochiral. Taken together, these observations expose a ‘lipid divide’ that is not as clear-cut as originally thought. Nonetheless, no natural microbial representatives have ever been reported to synthesize ‘mixed membranes’.

Here, we present the discovery of living bacteria of the phylum *Candidatus* Cloacimonetes of the Fibrobacteres–Chlorobi–Bacteroidetes (FCB) group superphylum, which are highly abundant in the deep anoxic waters of the Black Sea, harbouring a putative ‘mixed archaeal/bacterial membrane’. We observed that the metagenome-assembled genomes (MAGs) of this bacterial phylum contain the genes of the canonical bacterial fatty acid biosynthetic pathway but also, unexpectedly, homologs of key enzymes for archaeal membrane lipid biosynthesis. We validated the presence of these protein-coding genes both in silico and experimentally, and observed that they were expressed in the environment. Expression of these genes in *Escherichia coli* leads to the formation of a membrane containing ether-linked isoprenoid phospholipids. We hypothesize that the ‘archaeal’ membrane lipids of *Ca.*

Cloacimonetes have the G1P stereochemistry, awaiting validation based on isolation of these elusive bacteria and analysis of their membrane lipids. Database searches revealed the presence of the key archaeal membrane lipid biosynthetic genes not only in other *Ca. Cloacimonetes* genomes, but also in other genomes of the FCB group superphylum and related candidate phyla, indicating that the ability to produce ‘mixed membranes’ might be widespread in the tree of life.

Results and discussion

An abundant bacterium in the anoxic water column of the Black Sea

The microbial diversity in the water column of the Black Sea, a basin whose euxinic waters may resemble the ancient oceans (251), was determined by 16S rRNA gene amplicon sequencing (**Supplementary Tables 1 and 2**). Remarkably, a group of bacteria attributed to *Ca. Cloacimonetes* was very abundant (i.e. representing 5–20% of the total bacterial plus archaeal 16S rRNA gene reads) in the euxinic waters between 500 and 2,000 m depth (**Fig. 1a** and **Supplementary Table 2**). Catalyzed reporter deposition Fluorescence In Situ Hybridization (CARD-FISH) using a specific probe targeting *Ca. Cloacimonetes* confirmed their presence in the deep Black Sea waters (i.e. 2,000 m depth) and identified the cell morphology of the *Ca. Cloacimonetes* cells as oval to small rods (**Supplementary Fig. 1**).

A genome-centric metagenomics approach was subsequently undertaken to shed light on the physiology of *Ca. Cloacimonetes*; four draft genomes of this group with substantial to near completeness and low contamination were assembled (see ‘**Supplementary Information**’, **Supplementary Table 3**). The high abundance of these MAGs in the deep waters matched the *Ca. Cloacimonetes* 16S rRNA gene abundance profile (**Fig. 1b**). Phylogenomic analysis based on 43 concatenated core genes confirmed their taxonomic position (**Fig. 1c,d** and **Supplementary Fig. 2**). Based on their genomic content, the predicted metabolism of the *Ca. Cloacimonetes* found in the Black Sea is compatible with polysaccharide hydrolysis and fermentation of sugars and amino acids under an anaerobic or microaerophilic lifestyle (see ‘**Supplementary Information**’). Based on this genomic information, six different growth media supplemented with cellulose, cellobiose, microbial cell lysate, and amino acid mix were tested to obtain an enrichment but cultivation of *Ca. Cloacimonetes* was not achieved, possibly because culture conditions did not mimic the high hydrostatic pressure present in the deep Black Sea water column (see ‘**Supplementary Information**’ for details).

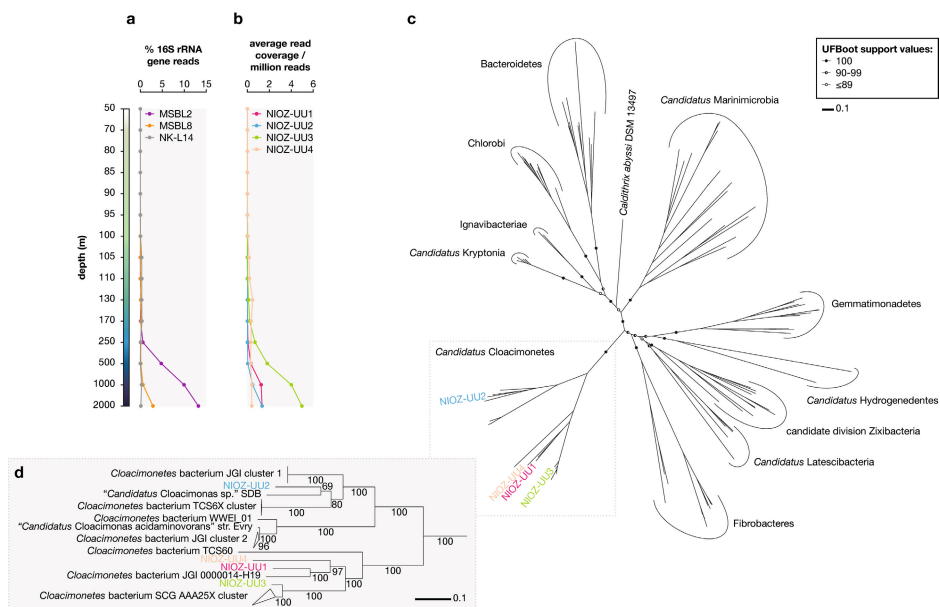


Fig. 1 | Distribution within the Black Sea water column and phylogeny of *Ca. Cloacimonetes*. **a**, Percentage of 16S rRNA gene reads attributed to different *Ca. Cloacimonetes* groups during BS2013. **b**, Estimated abundance of the four *Ca. Cloacimonetes* MAGs. **c**, Maximum likelihood phylogenetic tree of the FCB group super-phylum based on 43 concatenated core genes. Circles along branches indicate ultrafast bootstrap approximation support values, with only values for deepest nodes shown. **d**, Zoom in on the *Ca. Cloacimonetes* phylogeny. Numbers along branches indicate ultrafast bootstrap approximation support values. Scale bars in **c** and **d** represent mean number of substitutions per site.

An archaeal membrane lipid biosynthetic pathway encoded by the *Ca. Cloacimonetes* genomes

The analysis of the MAGs revealed unexpected features in the lipid biosynthetic pathways of *Ca. Cloacimonetes*. Genes encoding the canonical bacterial fatty acid biosynthetic pathway were detected, including the gene (*gps*) coding for glycerol-3-phosphate dehydrogenase (catalysing the formation of G3P), the acyltransferases responsible for the esterification of fatty acids and G3P, as well as genes coding for enzymes involved in downstream reactions (252) (see ‘**Supplementary Information**’). Hence, *Ca. Cloacimonetes* harbours the genes enabling the formation of a normal bacterial membrane (i.e. based on G3P-esterified fatty acids). Strikingly, however, putative key genes of the archaeal lipid biosynthetic pathway (**Supplementary Fig. 3**) were also detected in all four *Ca. Cloacimonetes* MAGs. The gene encoding GGGP synthase is co-localized (i.e. encoded in close proximity) with the gene encoding DGGGP synthase (**Supplementary Information**). Homology searches show that the two genes are found in other Bacteria and Archaea in the Black Sea water column as well (**Supplementary Figs. 4 and 5**). These two presumed archaeal enzymes together mediate the formation of the two ether bonds between isoprenoid

alcohols and G1P resulting in the production of unsaturated archaeol (215). Distant homologs of both GGGP and DGGGP synthase genes have previously been reported in bacterial genomes, however, their presence has never been associated with the production of 'archaeal' lipids. Whereas GGGP synthase activity has been confirmed in only a few bacteria (253), DGGGP synthase belongs to a large superfamily of UbiA prenyltransferases with several different potential functions (247) and bacterial homologs are assumed to have a different function. Moreover, there has never been an indication for co-expression of the two genes in bacteria. We tested and rejected the possibility that the co-localized GGGP and DGGGP synthase homologs in the *Ca. Cloacimonetes* MAGs could have been introduced by a methodological error, both by *in silico* analyses and by experimentally amplifying and resequencing one of the scaffolds containing these genes from the original Black Sea water ('**Supplementary Information**', **Supplementary Figs. 6** and **7**). The presence of *Ca. Cloacimonetes* GGGP and DGGGP synthase gene transcripts in the Black Sea water column confirmed that they were also expressed in the environment ('**Supplementary Information**', **Supplementary Figs. 8–10**). This resulted in the hypothesis that this bacterium uses the GGGP and DGGGP synthases actively and thus is capable of synthesizing archaeal-like membrane lipids.

Next, we set out to verify that the detected genes coding for putative GGGP and DGGGP synthase have the predicted activity leading to archaeal-like membrane lipids. Previous studies have determined that GGGP synthase homologs of the phylum Bacteroidetes have *in vitro* GGGP synthase activity with G1P, like the archaeal GGGP synthase (253), rather than the heptaprenyl synthase activity of the bacterial PcrB orthologs detected in *Bacillus subtilis* (254). To biochemically verify the predicted enzymatic activity in *Ca. Cloacimonetes*, we recombinantly produced its GGGP synthase protein in *E. coli*. The purified protein was found to catalyse formation of GGGP from GGPP in an enzymatic assay (**Fig. 2a**, '**Supplementary Information**', **Supplementary Figs. 11–13**), as expected in the archaeal membrane lipid biosynthetic pathway (215). Then, we tested if the *Ca. Cloacimonetes* GGGP and DGGGP synthases could support the formation of a 'mixed membrane' in a bacterial cell, by co-expressing them in an *E. coli* optimized for production of GGPP and G1P, the likely substrates of the two enzymes. Cells producing both GGGP and DGGGP synthase contained significant amounts of phosphatidylglycerol (PG) archaeol (i.e. archaeol, also known as diphytanyl glycerol diether, with PG as a polar head group) with eight double bonds (**Fig. 2b**, '**Supplementary Information**', **Supplementary Fig. 14**), the expected intermediate in the biosynthesis of archaeal membrane lipids in the absence of a specific geranylgeranyl reductase in *E. coli* (255,256). Hence, these experiments provide strong evidence for potential synthesis of ether-linked isoprenoid phospholipids by this group of bacteria.

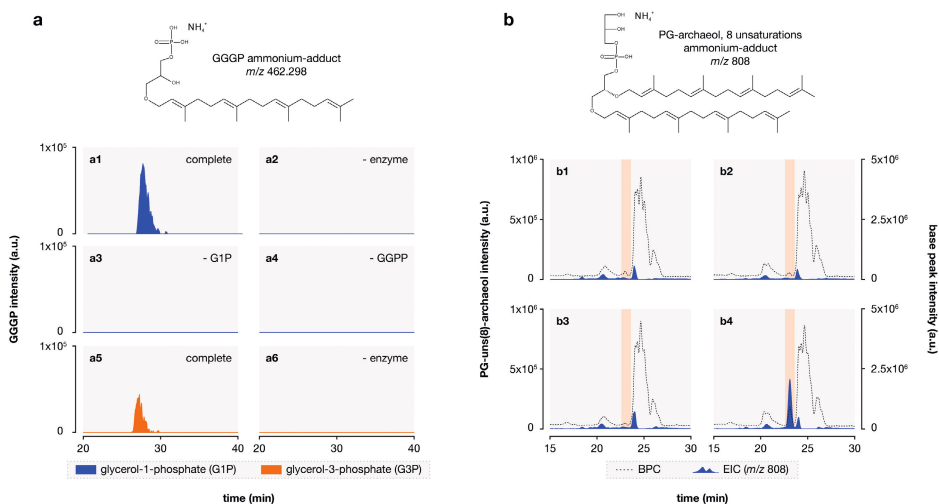


Fig. 2 | Biochemical verification of *Ca. Cloacimonetes* GGGP and DGGGP synthase activities. **a**, Enzyme assay of purified GGGP synthase with G1P (**a1–a4**) or G3P (**a5** and **a6**). Extracted ion chromatogram, within 3 ppm mass accuracy, of [GGGP + NH₄]⁺ (m/z 462.298) of complete enzymatic assay (**a1** and **a5**) or control assays lacking enzyme (**a2** and **a6**), glycerol phosphate (**a3**) or geranylgeranyl-diphosphate, GGPP (**a4**). **b**, Production of the lipid PG-archaeol with eight double bonds or unsaturations (PG-unsat(8)-archaeol) in *E. coli*. Extracted ion chromatogram (± 0.5 , mass units, mu) of [PG-unsat(8)-archaeol + H]⁺ (m/z 808) (blue filled area, left axis) or base peak chromatogram (dotted line, right axis) of optimized *E. coli* containing empty-vector (**b1**) or vector encoding GGGP synthase (**b2**), DGGGP synthase (**b3**), or both GGGP and DGGGP synthases (**b4**). Orange boxes span the retention time of PG-unsat(8)-archaeol.

Additional lipid biosynthetic genes in the *Ca. Cloacimonetes* MAGs

In addition to genes for GGGP and DGGGP synthase, other genes required for the synthesis of isoprenoidal archaeal lipids (**Supplementary Fig. 3**) were also detected in the *Ca. Cloacimonetes* MAGs. They contain the genes for a complete bacterial isoprenoid MEP/DOXP pathway, genes coding for acetyl-CoA C-acetyltransferase and hydroxymethylglutaryl-CoA synthase of the Mevalonate pathway, and two polyprenyl synthases (see ‘**Supplementary Information**’, **Supplementary Fig. 15a**), which support the existence of the biosynthetic pathway leading to GGPP, one of the substrates of GGGP synthase. Finally, the MAGs also encode geranylgeranyl reductases that are closely related to homologs found in Euryarchaeota (see ‘**Supplementary Information**’, **Supplementary Fig. 15b**). Thus, in addition to ether bond formation (mediated by the putative GGGP and DGGGP synthases), these bacteria have the capacity to synthesize isoprenoid chains (via isoprenoid biosynthetic pathways and the presence of polyprenyl synthases) and saturate them (via the putative geranylgeranyl reductases), characteristics that are fully in line with archaeal lipid membrane biosynthesis (215).

Only two genes of the known lipid biosynthetic pathway in archaea were not detected in the *Ca. Cloacimonetes* MAGs. One is the CDP-archaeol synthase (CarS)

forming the activated CDP-archaeol before the addition of the polar head groups in combination with other specific archaeal enzymes (222). However, it has been recently demonstrated that the bacterial CDP diacylglycerol synthase can replace the function of the archaeal CarS to generate CDP-archaeol (257). Subsequently, substrate promiscuity allows the bacterial phosphatidylglycerophosphate synthetase together with the phosphatidylglycerophosphatase to recognize CDP-archaeol and synthesize archaetidylglycerol (257) as in the archaeal lipid biosynthetic pathway. All these bacterial enzymes are encoded by the four MAGs, which together with the biosynthetic genes mentioned above further support the formation of archaeal-like membrane lipids.

The second gene that is not detected in the four MAGs is the gene (*egsA*) coding for the enzyme enabling G1P biosynthesis (i.e. G1PDH). Its bacterial homolog (*araM*), occurring in a few bacteria (258), is also absent. This seems enigmatic at first sight, as it would suggest that the presumed archaeal membrane lipids synthesized by *Ca. Cloacimonetes* do not have G1P as a glycerol phosphate backbone. Rather, they could have the G3P stereochemistry as promiscuity of GGGP synthase for G3P has been observed by others (225,259,260), as well as in our enzyme assay (Fig. 2a). However, it was recently demonstrated that an alternative pathway for the formation of G1P must exist within bacteria (225). Notably, a genetically engineered bacterial strain of *E. coli*, whose genome contained archaeal lipid biosynthesis genes, formed archaeal membrane lipids with the G1P stereochemistry even when the *AraM* coding gene was not included in the genetic construct. The genome did not contain *egsA* either. This shows that *araM* is not required for G1P synthesis in *E. coli*. This study also highlighted that in the presence of G3P as substrate, there is still a very high stereoselectivity towards G1P for the formation of archaeal membrane lipids. While we do not know how *E. coli* produces G1P, it is striking that the only known archaeal lipid biosynthesis gene that is missing in the *Ca. Cloacimonetes* MAGs is nonessential for the formation of archaeal-like membrane lipids.

In summary, the four *Ca. Cloacimonetes* MAGs encode a complete set of genes for the synthesis of ether-linked isoprenoid phospholipids. When archaeal GGGP and DGGGP synthase coding genes were previously expressed in *E. coli*, it formed a heterochiral membrane with G3P-bacterial lipids and G1P-archaeal lipids (225). Given the strong stereoselectivity of the archaeal membrane lipid pathway for G1P and the observation that within bacteria an unknown G1P synthesis pathway exists, we hypothesize that *Ca. Cloacimonetes* also forms ether-linked isoprenoid phospholipids with the G1P stereochemistry, which should be confirmed with the isolation and lipid analysis of members of this phylum.

The presence of the above-mentioned genes encoded by the *Ca. Cloacimonetes* MAGs are compatible with the formation of archaeol (diether) lipid membranes. While

many archaeal groups have been seen to synthesize also or exclusively tetraether membrane lipids (i.e. glycerol dibiphytanyl glycerol tetraethers, GDGTs; ref. 261), the genes required for their synthesis (i.e. potential GDGT synthase) remain unknown, thus we are unable to determine if the *Ca. Cloacimonetes* MAGs could synthesize tetraether membrane lipids.

Estimating the *Ca. Cloacimonetes* contribution to the ‘archaeal’ membrane lipid pool in the Black Sea

Next, we assessed whether the observed *Ca. Cloacimonetes* abundances could significantly contribute to the ‘archaeal’ membrane lipids of living cells. In order to do so, archaeal IPLs (i.e. archaeol and GDGTs) were quantified in the Black Sea deep water column. IPLs are relatively easily hydrolysed once the cell dies and therefore considered as biomarkers of living biomass (262).

We estimated the absolute abundance of archaeal IPLs, which showed an increase from 2.5 to 25 ng L⁻¹ from 500 to 2,000 m depth (**Fig. 3**). The archaeal population in the euxinic waters of the Black Sea mostly consists of Bathyarchaeota, DPANN archaea, and Euryarchaeota Thermoplasmatales, and ANME-1 (**Supplementary Table 2b**; see also ref. 263). Based on their estimated cell numbers in the water column and their expected lipid content per cell we predict a maximum archaeal membrane lipid concentration of <6.5 ng L⁻¹ (**Fig. 3, ‘Supplementary Information’**). At 1,000–2,000 m depth, this represents a striking offset between observed and expected archaeal IPL concentrations ranging at least from two- to fivefold. The mismatch between observed and expected is likely larger considering that these estimates represent an ideal case scenario (see ‘**Supplementary Information**’). The mismatch could be due to preservation of suspended IPLs in the anoxic waters of the Black Sea. Extracellular archaeal IPLs, especially those with more stable glycosidic bonds, could potentially be preserved as fossils as previously observed in deep anoxic sediments (e.g. ref. 264). However, this possibility seems unlikely, as the IPLs are detected in the water column, which is a more dynamic system than sediments. Moreover, most of the archaeal IPLs detected in the Black Sea do not have glycosidic-based polar head groups but the more labile PG, phosphatidylserine, or phosphatidylethanolamine head groups (263). Therefore, the mismatch is more likely explained by production of archaeal-like ether-linked isoprenoid membrane lipids by the highly abundant *Ca. Cloacimonetes* bacteria. It is currently not possible to determine the stereochemistry of the archaeal IPLs detected in environmental samples as the analysis would require much higher concentrations of purified archaeal lipids. Hence, although experimental confirmation of archaeal IPL synthesis by *Ca. Cloacimonetes* bacteria and the stereochemistry of these lipids await their isolation, these environmental observations provide enticing circumstantial evidence for the bacterial production of archaeal-like membrane lipids.

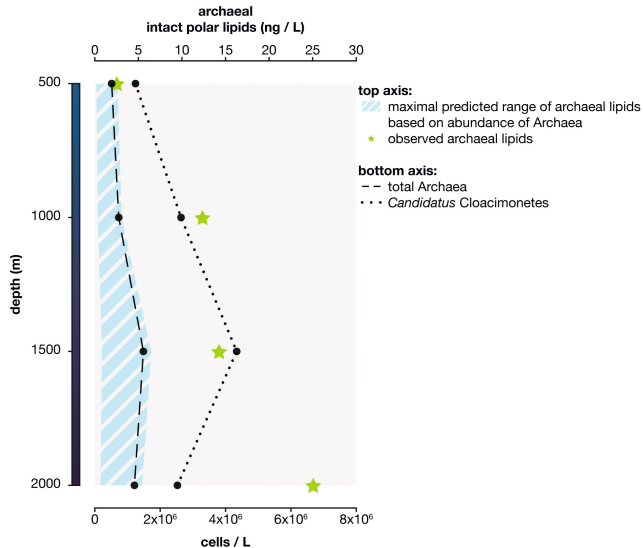


Fig. 3 | Detected archaeal intact polar lipids (IPLs) in nanograms per litre (green stars) in the Black Sea water column, and predicted archaeal IPLs (shaded area) considering different estimates of membrane lipid production and cell size for the archaeal groups found. Total archaeal cells and *Ca.* Cloacimonetes cells per litre based on qPCR and amplicon sequencing data are also indicated. All calculations are available in **Supplementary Notebook 1.**

Archaeal lipid biosynthetic genes in *Ca.* Cloacimonetes and other bacteria

Homology searches coupled to phylogenetic analyses indicated that the presence of archaeal lipid biosynthetic genes in bacterial genomes is not limited to *Ca.* Cloacimonetes from the Black Sea. Close GGGP and DGGGP synthase homologs are found together in other genomes of *Ca.* Cloacimonetes, as well as in other FCB group superphylum bacteria, related bacterial candidate phyla, and in a genome of *Candidatus* Parcubacteria of the ‘Candidate Phyla Radiation’ (104) (**Fig. 4, Supplementary Figs. 16–18, ‘Supplementary Information’**). The phylogenetic topologies of both the GGGP and DGGGP synthases in bacteria are similar with respect to both branching of bacterial groups and in sharing a close affiliation with GGGP and DGGGP synthases from the archaeal TACK group, in particular Crenarchaeota (247), suggesting that the two genes share a similar evolutionary history (**Fig. 4, Supplementary Figs. 16 and 17, ‘Supplementary Information’**). Co-localization of GGGP synthase and DGGGP synthase as observed in the four *Ca.* Cloacimonetes MAGs has previously only been seen in some Euryarchaeota, and within bacteria seems to be restricted to *Ca.* Cloacimonetes (**Supplementary Information**). Considering the basal placement of these genomes in the bacterial clade of both trees (**Fig. 4, Supplementary Figs. 16 and 17**), we hypothesize that genomic co-localization of the genes is the ancestral state.

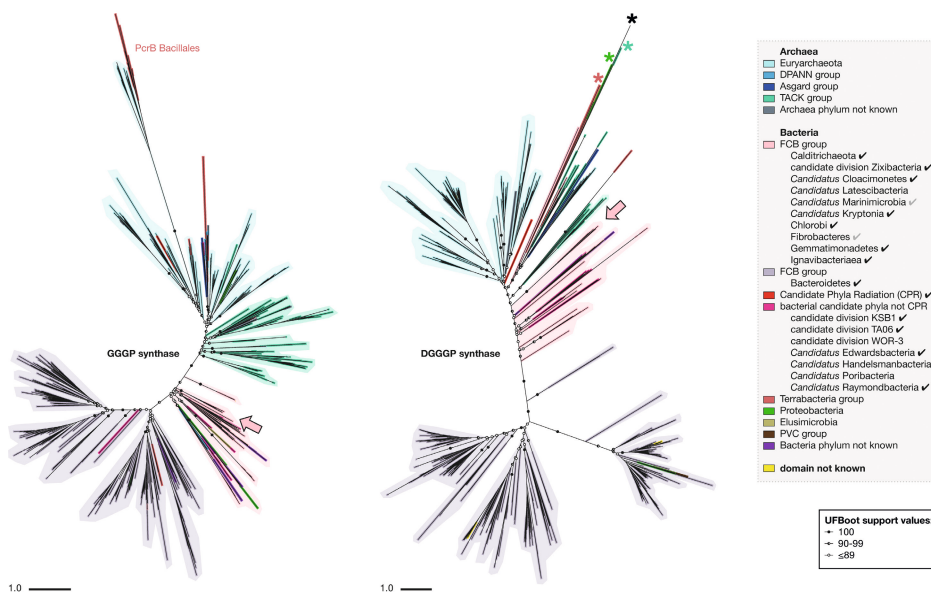


Fig. 4 | Phylogenetic trees of the GGGP and DGGGP synthase homologs detected across the tree of life. Search for homologs was performed with *Ca. Cloacimonetes* MAG sequences (arrows), in genomes from both cultures and environmental samples. Shadings illustrate the dominant group in the clade. Circles along branches indicate ultrafast bootstrap approximation support values, with only values for deepest nodes shown. See **Supplementary Figs. 6 and 7** for annotated trees with all branch support values. Scale bars represent mean number of substitutions per site. Asterisks in the DGGGP synthase tree indicate known UbiA prenyltransferases that do not have DGGGP synthase activity. A checkmark in the legend marks bacterial groups with genomes that code for both GGGP and DGGGP synthase, a grey checkmark marks groups in which both genes are found but not in the same genome. TACK includes the Thaumarchaeota and the Crenarchaeota. Archaeal or bacterial ‘phylum not known’: phylum is not known but the genome is annotated on a lower level, or sequence represents multiple groups. ‘Domain not known’: genomes for which no lineage was found on the PATRIC servers.

The extended presence and close phylogenetic associations of the GGGP and DGGGP synthase homologs in the FCB group superphylum and related candidate phyla strongly supports an origin in Bacteria before the radiation of the FCB group superphylum. This suggests that the capacity to synthesize ‘mixed membranes’ has been a trait of these bacteria for a long period in evolutionary history. Two evolutionary scenarios may explain our current observations. First, the contemporary presence of these enzymes in the FCB group superphylum bacteria could reflect an evolutionary remnant of the ‘mixed membrane’ stage after LUCA but before the diversification of Bacteria. Notably, a recent study that also evaluated the phylogeny of GGGP and DGGGP synthases using outgroup-free routing could not exclude a LUCA origin for both genes (250). Second, if the ancestor of the FCB group superphylum contained a homochiral bacterial membrane like the other contemporary bacterial groups, the phylogenetic similarities of the co-localized genes could also reflect an ancient horizontal gene transfer (HGT) from an ancestral archaeal lineage into the FCB group ancestor (see ‘**Supplementary Information**’).

Implications for the ‘lipid divide’

In contrast to the dogma of the ‘lipid divide’, bacteria of the FCB group superphylum harbour a complete archaeal-like membrane lipid biosynthetic pathway. Our data indicate that members of *Ca. Cloacimonetes* have the potential to synthesize ether-linked isoprenoid phospholipids, possibly with the G1P stereochemistry, in which case they would possess a true heterochiral ‘mixed membrane’. This is the first evidence of naturally occurring organisms with this ability. The presence of bacterial and archaeal lipid biosynthesis genes in *Ca. Cloacimonetes* strikingly resembles the ‘Marine Group II’ Euryarchaeota (currently known as ‘*Ca. Poseidoniales* ord. nov.’ (248)) (249) and some members of the Asgard superphylum, which harbour genes for the synthesis of putative bacterial membrane lipids (see ‘**Supplementary Information**’). The existence of a natural contemporary bacterial counterpart synthesizing archaeal-like membranes might provide weight to the hypothesis that these archaeal groups also synthesize ‘mixed membranes’. Hence, *Ca. Cloacimonetes*, as well as the rest of the FCB group superphylum, appear to be key in our understanding of the ‘lipid divide’. Their membranes may possibly reflect evolutionary remnants of the hypothetical ‘mixed membrane’ of LUCA, or an ancient HGT. In either case, this discovery provides further support for the existence and potential feasibility of ‘mixed membranes’ in natural environments and over a long period in evolutionary history, bridging the lipid divide.

Materials and methods

Oceanographic sampling

All cruises in the Black Sea western gyre were performed with the R/V Pelagia. Suspended particulate matter (SPM) from 15 depths across the water column (50–2,000 m) was collected at sampling station 2 (N42°53.8', E30°40.7', 2,107 m depth) during the Phoxy cruise 64PE371 (BS2013) on 9–10 June 2013 (**Supplementary Table 1**). At the same station, SPM was also collected from 1,000, 1,500, and 1,980 m depth during the NESSC cruise 64PE408 (BS2016) on 31 January–2 February 2016, and from 2,000 m depth during the 64PE444 cruise (BS2018) on 17 August 2018. SPM from four depths (500, 1,000, 1,500, and 2,000 m) was collected at sampling station 4 (N42°46.9', E29°21.1', 2,100 m depth) during the 64PE418 cruise (BS2017) on 27 March–5 April 2017. For the BS2013 and BS2017 cruises, SPM was collected with McLane WTS-LV in situ pumps (McLane Laboratories Inc., Falmouth) on pre-combusted glass fibre filters with 142 mm diameter and 0.7 and 0.3 μm pore size, respectively. For the BS2016 and BS28018 cruises, SPM was collected on 0.22 μm Sterivex cartridge filters (Millipore). In all cases, all samples were stored at $-80\text{ }^{\circ}\text{C}$ until nucleic acid or lipid extraction (only for the glass fibre filters) was performed. Water samples were collected during the BS2018 cruise to attempt enrichment cultures and for visualization of the cell morphology as specified in the '**Supplementary Information**'.

Lipid analysis environmental samples

Total lipids were extracted from freeze-dried glass fibre filters as described in ref. 263. The Bligh and Dyer lipid extracts are expected to contain archaeal intact polar lipids (IPLs), which are composed of the core lipid (CL) attached to one or two polar head groups (265). The Bligh and Dyer extracts were both analysed directly for the presence of total archaeal CL, and also after acid hydrolysis to remove the polar head groups and quantify both the archaeal CLs and the IPL-derived CLs. Subtracting CLs from CL + IPL-derived CLs allows for determination of the IPL-derived CLs linked to living archaeal biomass. Acid hydrolysis was performed in nitrogen-dried Bligh and Dyer extracts (266). Extracts were analysed by UHPLC–atmospheric pressure chemical ionization MS for archaeol and GDGTs, according to Hopmans et al. (267) with some modifications. Briefly, analysis was performed on an Agilent 1260 UHPLC coupled to a 6130 quadrupole MSD in selected ion monitoring mode. Separation was achieved on two UHPLC silica columns (BEH HILIC columns, $2.1 \times 150\text{ mm}$, $1.7\text{ }\mu\text{m}$; Waters) in series, fitted with a $2.1 \times 5\text{ mm}$ pre-column of the same material (Waters) and maintained at $30\text{ }^{\circ}\text{C}$. Archaeol and GDGTs were eluted isocratically for 10 min with 10% B, followed by a linear gradient to 18% B in 20 min, then a linear gradient to 100% B in 20 min, where A is hexane and B is hexane:isopropanol (9:1). Flow rate was 0.2 ml min^{-1} . Total run time was 61 min with a 20 min re-equilibration. Source settings were identical to Schouten et al. (268). Typical injection volume was

10 μl of a 1 mg ml^{-1} solution. The m/z values of the protonated molecules of archaeol and isoprenoid GDGTs were monitored. Archaeol and GDGTs were quantified by adding a C_{46} GTGT internal standard by using an archaeol:GDGT-0 standard (1:1) to correct for response differences between archaeol and GDGTs (269).

Nucleic acid extraction and 16S rRNA gene amplicon sequencing

DNA and RNA were extracted from sections of the glass fibre filters (1/8 filter from 50 to 130 m depth and 1/4 from 170 to 2,000 m depth) or from the Sterivex filter cartridge with the RNA PowerSoil[®] Total Isolation Kit plus the DNA elution accessory (Mo Bio Laboratories, Carlsbad, CA). RNA extracts were treated with DNase and reverse-transcribed to cDNA using random nonamers as described previously (270). The 16S rRNA gene amplicon sequencing and analysis was performed as described previously (271) (see **Supplementary Table 2** and '**Supplementary Information**' for details). Taxonomy of the reads was assigned based on BLAST (272) and the SILVA database version 123 (ref. 273).

Metagenome sequencing and assembly

Unamplified DNA extracts from the 15 SPM samples of BS2013 were used to prepare TruSeq nano libraries which were further sequenced with Illumina MiSeq (5 samples multiplexed per lane) at Utrecht Sequencing Facility, generating 45 million 2×251 bp paired-end reads. Quality control was performed with FastQC v0.11.3 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and reads with uncalled bases and remaining TruSeq adapters were removed with FLEXBAR v2.5 (ref. 274), keeping the longer side of the read with the '--ae any' flag. All reads were cross-assembled with SPAdes v3.8.0 in '--meta' mode (204), with read error correction turned on. BWA-MEM v0.7.12-r1039 (ref. 275) was used to map the forward and reverse reads from individual samples to the cross-assembled scaffolds.

Scaffold binning and assessment of MAG quality

Scaffolds were binned into draft genome sequences based on coverage profile across samples and tetra-nucleotide frequency with MetaBAT v0.32.4 (ref. 197). The '--super-specific' preset was used to minimize contamination. To increase sensitivity without losing specificity, MetaBAT was run with ensemble binning, which aims to combine highly similar bins ('-B' and '--pB' were set to 20% and 50%, respectively). Quality of the MAGs was assessed based on absence and presence of lineage-specific marker gene sets after genome placement in a reference tree with CheckM v1.0.7 (ref. 51) in '--lineage_wf' mode.

Genome annotation and abundance estimation

MAGs were annotated with Prokka v1.11 (ref. 276) in '--metagenome' mode. GenBank output files generated by Prokka were also annotated with the Rapid Annotation using Subsystem Technology (RAST) pipeline v2.0 (ref. 277). The geranylgeranylglyceryl

phosphate (GGGP) synthase of NIOZ–UU2 is only partially predicted by Prokka as it bridges a scaffold boundary (see ‘**Supplementary Information**’). For the homology searches and tree constructions detailed below a longer protein was reconstructed by concatenating the predicted partial protein sequence with the last part of the translated linked scaffold.

MAG abundance was estimated from the shotgun metagenomics data by generating depth files per sample for all the scaffolds with SAMtools v.1.3.1 (ref. 278), using the mpileup utility with flags ‘-aa’ and ‘-A’ (count orphans) set. Average read coverage per nucleobase of a scaffold was calculated by dividing the sum of depth of all positions by the length of the scaffold with N’s removed. Coverage of a MAG was calculated likewise after concatenating the depth files of the scaffolds in that MAG. Average read coverage per nucleobase was normalized across samples by dividing by the total number of reads after quality control in the sample times 1,000,000. Normalized data is only shown in **Fig. 1**.

Manual cleaning of the MAGs

The four *Ca. Cloacimonetes* MAGs were cleaned by plotting coverage across samples for all the scaffolds in the MAG, and manually removing those scaffolds that did not clearly have the same coverage profile as the majority (see ‘**Supplementary Information**’ for details). NIOZ–UU3 appeared clean based on its coverage profile so none of its scaffolds were removed. Because the removed scaffolds do not contain marker genes, completeness and contamination estimates did not change (see **Supplementary Table 3** for extended CheckM results of the four MAGs). Genome annotations and abundance estimations of MAGs were newly generated as described above after cleaning.

Placement of MAGs in the FCB group superphylum tree

To further check the phylogenetic affiliation of the four MAGs in comparison with close relatives, Bacteroidetes/Chlorobi group genomes were downloaded from RefSeq (163), and other FCB group genomes from GenBank (279), both on February 8, 2017 (**Supplementary Table 4**). For Ignavibacteriae and Chlorobi all the RefSeq representative genomes were downloaded, and for Bacteroidetes only the first ten representative genomes. We only included genomes that were estimated to be less than 10% contaminated by CheckM in ‘--lineage_wf’ mode, and that contained at least 4 out of 43 phylogenetically informative marker genes in single copy that CheckM uses for bin placement (51) (**Supplementary Table 4**). We realigned the 43 marker genes individually with Clustal Omega v1.2.3 (ref. 280). The genes were concatenated, gaps included if a gene was not present, and identical sequences collapsed (**Supplementary Table 4**). A maximum likelihood tree was inferred with IQ-TREE v1.6.3 (ref. 281). Model selection of nuclear models was performed with ModelFinder (282) and the best-fit model (LG+R8) chosen according to Bayesian

Information Criterion. Branch support was based on 1,000 ultrafast bootstraps (283). The tree was visualized in iTOL (284).

We identified 16S rRNA genes in the FCB group genomes included in the tree with the CheckM *ssu_finder* utility. 16S rRNA gene sequences were found in 21 *Ca. Cloacimonetes* genomes, including one of the MAGs (NIOZ-UU1) (**Supplementary Table 4**). We aligned these sequences with the *Ca. Cloacimonetes* 16S rRNA gene amplicon sequences using MAFFT v7.394 (ref. 285), sliced out the amplicon region from the genome sequences, and removed identical sequences (**Supplementary Table 4**).

Homology searches of GGGP synthase and (S)-2,3-di-O-geranylgeranylglyceryl phosphate (DGGGP) synthase in the Black Sea water column

Predicted GGGP and DGGGP synthases from the four MAGs were queried with TBLASTN v2.6.0+ (refs. 272,286) (e -value $< 1e-5$ and query coverage $\geq 70\%$) against all scaffolds in the assembly. For the scaffolds with significant hits (61 for GGGP synthase and 43 for DGGGP synthase), we extracted the aligned part of the subject sequence of the best hit based on e -value on each scaffold. In addition to the queries we included a set of known GGGP and DGGGP synthases based on biochemical evidence and phylogenetic analyses (287), as well as other non-DGGGP prenyltransferases as ‘outgroups’ (see ‘**Supplementary Information**’). We also queried the GGGP and DGGGP synthases from the four MAGs against the NCBI nonredundant protein database (nr) (156) with BLASTP (272), and added the four best hits for each gene, all of which were found in *Ca. Cloacimonetes*. The protein sequences were aligned with MAFFT v7.394 (ref. 285), using a maximum number of 1,000 iterative refinements and local pair alignment (L-INS-i). The sequence alignments were trimmed with trimAl v1.4.rev22 (ref. 288) in ‘gappyout’ mode, and identical sequences were collapsed. Final alignment lengths were 171 and 200 positions for GGGP synthase and DGGGP synthase, respectively. Maximum likelihood trees were inferred with IQ-TREE. Model selection of nuclear models was performed with ModelFinder and the best-fit model (LG+R7 for GGGP synthase and LG+F+R7 for DGGGP synthase) chosen according to Bayesian Information Criterion. Branch support was based on 1,000 ultrafast bootstraps. Trees were visualized in iTOL.

Taxonomy was assigned to hits based on taxonomic classification of the scaffolds on which they were found with Contig Annotation Tool v5.1.2 (ref. 289). Database files were constructed on March 4, 2020. We used Prodigal v2.6.3 (ref. 290) for gene prediction and DIAMOND v0.9.21 (ref. 209) for protein alignment to nr. The f parameter was set to 0.3 to allow for speculative classifications. One scaffold had multiple classifications, and we chose the lowest classification that reached a majority vote based on bit-score support in this case.

Homology searches of GGGP synthase and DGGGP synthase across the tree of life

Predicted GGGP and DGGGP synthases from the four MAGs were queried with BLASTP v2.6.0+ (refs. 272,286) against 110,421 annotated genomes available in the PATRIC genome database (291) on November 29, 2017. BLASTP was run per genome (e -value $< 1e-10$ and query coverage $\geq 70\%$) with a fixed database size of 20,000,000 to make e -values comparable across genomes. For GGGP synthase, we collected all hits and included the entire protein in the analysis. For DGGGP synthase, we only included hits that were annotated as ‘Digeranylgeranyl glyceryl phosphate synthase (EC 2.5.1.42)’, and (‘similar to’) ‘(S)-2,3-di-O-geranylgeranyl glyceryl phosphate synthase’, and hypothetical proteins with $\geq 90\%$ query coverage, to exclude other prenyltransferases from the superfamily. Again, entire proteins were included in the analysis. We added the queries, and the set of extra sequences (see above): known GGGP and DGGGP synthases, outgroups, and four best hits for each gene from nr. Identical sequences were collapsed.

Alignment and tree inference were performed with MAFFT, trimAl, and IQ-TREE as described above. Final alignment lengths were 246 and 266 positions for GGGP synthase and DGGGP synthase, respectively. The best-fit models were LG+R10 for GGGP synthase and LG+F+R10 for DGGGP synthase. For both trees, the consensus tree had a higher likelihood than the maximum likelihood tree found. Major clade separations were comparable between the maximum likelihood and consensus tree for both genes. Consensus trees were visualized in iTOL.

Co-localization of GGGP and DGGGP synthase in FCB group superphylum genomes

Predicted GGGP and DGGGP synthases from the four *Ca. Cloacimonetes* MAGs were queried with TBLASTN (e -value $< 1e-5$ and query coverage $\geq 70\%$) against all downloaded GenBank/RefSeq FCB group superphylum genomes (see above) to identify homologs. If hits were located on the same scaffold, the minimum base pair distance between GGGP and DGGGP synthase homologs was considered as a measure of co-localization.

Amplification, sequencing, gene expression, and quantification of specific genes in NIOZ-UU3

To experimentally assess the assembly accuracy of NIOZ-UU3, primers were designed to amplify and sequence the region of the scaffold that contains the genes predicted to code for GGGP synthase, DGGGP synthase, polyprenyl synthase, and the bacterial marker gene predicted by CheckM (helicase PriA; see ‘**Supplementary Information**’) from the BS2013 2,000 m depth sample. PCR reaction mixture was the following (final concentration): Q-solution (PCR additive) 1 \times ; PCR buffer 1 \times ; bovine serum albumin (200 $\mu\text{g ml}^{-1}$); dNTPs (40 μM); primers (0.4 pmol); MgCl_2

(1.5 mM); 2.5 U Taq polymerase (Qiagen, Valencia, CA, USA). PCR conditions were the following: 95 °C, 5 min; 35 × (95 °C, 15 s; 62 °C (melting temperature verified by gradient PCR), 30 s; 72 °C, 1 min per kilobase); final extension 72 °C, 5 min. PCR product was gel purified (QIAquick gel purification kit, Qiagen), cloned in the TOPO-TA cloning® kit from Invitrogen (Carlsbad, CA, USA), and sequenced for verification. In addition, to determine the presence and expression of the GGGP synthase, DGGGP synthase, and polyprenyl synthase coding genes, PCRs targeting a fragment of the genes (see primers in ‘**Supplementary Information**’) were tested with DNA and cDNA from the SPM samples recovered at 1,000 and 2,000 m depth from the Black Sea 2013 campaign as a template with the same PCR master mix as described above but with half amount of dNTPs and Taq Polymerase (melting temperature 62 °C verified by gradient PCR). Moreover, quantitative PCRs (qPCR) with the same specific primers targeting the GGGP and the DGGGP synthase coding genes were performed on DNA and cDNA extracts of the Black Sea 2013 campaign from 50 to 2,000 m depth (15 samples). qPCR master mix was the same as used for amplifications above with the addition of EvaGreen fluorescent nucleic acid dye (0.625 nM final concentration). PCR analyses were performed on a Bio-Rad CFX96 Real-Time System/C1000 Thermal cycler equipped with CFX Manager Software. Standard curves were generated by dilution of the verified amplicon of the GGGP and DGGGP synthase coding gene fragment obtained with DNA extracts of the SPM at 2,000 m depth as described above.

Construction of GGGPS and DGGGPS expression plasmids

Plasmids for in vivo lipid production were constructed by PCR amplifying the GGGP synthase ORF from pLVA01 and the plasmid pRSFDuet-1 (NovaGen) using the primers GGGPS-RSFDuet-F/R and RSFDuet-GGGPS-F/R, respectively, and further recombined (292) (**Supplementary Information**). A non-synonymous mutation (Met to Val, based on NIOZ-UU3) was corrected by site-directed mutagenesis using the primers GGGPS-SDM-F/R, resulting in pABW1. The DGGGP synthase ORF was amplified using the primers DGGGPS-F/R and cloned into pRSFDuet-1 using NcoI and BamHI to construct pABW2. pABW3 (encoding both GGGPS and DGGGPS) was constructed by re-amplifying DGGGPS from pABW2 and cloning the amplicon as a NcoI-BamHI fragment into pABW1. An N-terminally 6His-tagged GGGP synthase overproduction plasmid (pABW4) was constructed by recombining (292) the GGGP synthase ORF from pABW1 and pET-28a(+) (NovaGen) amplified using the primers GGGPS-ET28-F/R and ET28-GGGPS-F/R, respectively. All inserts were verified to encode correct proteins by sequencing.

Recombinant production, purification, and in vitro enzyme assay of GGGP synthase

Purification and enzymatic assay of the NIOZ-UU3 GGGP synthase was based on the method of Jain et al. (293). ‘*E. coli* BL21(DE3)’ harbouring pABW4 was cultured

in 250 ml Lysogeny broth (LB) medium (37 °C, 200 rpm) and induced with 0.5 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) at 0.6 OD_{600nm} for 4 h. Cells were centrifuged (3,500 rcf) and frozen. Subsequent steps were performed at room temperature, with samples and buffers kept on ice. Thawed cells were resuspended in ~5 ml lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 20 mM imidazole, 1 mg ml⁻¹ lysozyme), and a protease inhibitor (cOmplete™, EDTA- free; Roche, Basel) and sonicated to facilitate cell lysis. The lysate was cleared by centrifugation (20,000 rcf, 10 min), glycerol was added (10% final), and sample loaded onto a gravity column containing 1.5 ml of nickel-nitrilotriacetic acid agarose beads (Qiagen, Venlo, NL) pre-equilibrated with protein buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 20 mM imidazole, 10% glycerol). Beads were washed with ~30 column volumes of protein buffer, and initially eluted by a titration of imidazole with GGGP synthase protein eluting at 200 mM. Subsequent elutions were performed with 250 mM imidazole. The purity of the GGGP synthase protein was verified using 12% TGX™ precast gels (Bio-Rad), stained with Bio-Safe™ Coomassie stain (Bio-Rad). The concentration of the purified protein was calculated based on the absorbance at 280 nm using a predicted extinction coefficient of 9002 M⁻¹ cm⁻¹ (ExpASy ProtParam, averaged Cys reduced, and cysteine form).

Enzymatic activity of the purified *Ca. Cloacimonetes* GGGP synthase was determined in an end-point assay using 0.1 μ M GGGP synthase, 10 mM GIP or G3P and 100 μ M geranylgeranyl pyrophosphate in a reaction buffer consisting of 50 mM Tris-HCl (pH 7.5), 10 mM MgCl, and carryover amounts of imidazole (1 mM) and glycerol (0.5%). Reactions were incubated for 2 h at 37 °C in glass vials, extracted twice with 300 μ L *n*-butanol (water saturated), pooled and stored at -20 °C. Analysis of the GGGP synthase enzyme assays were performed using Ultra High Pressure Liquid Chromatography – High Resolution Mass Spectrometry (UHPLC–HRMS) based on Sturt et al. (265) with some modifications as detailed below. Pooled butanol extracts were evaporated under a stream of nitrogen, redissolved in 50 μ L methanol:dichloromethane (1:1) and filtered (0.45 μ m, regenerated cellulose). Analysis was performed using an Agilent 1290 Infinity I UHPLC, equipped with thermostatted auto-injector and column oven, coupled to a Q Exactive Orbitrap MS with Ion Max source with heated electrospray ionization (HESI) probe (Thermo Fisher Scientific, Waltham, MA). Injection volume was 1 μ l (out of 50 μ l). Separation was achieved on a YMC-Triart Diol-HILIC column (250 \times 2.0 mm, 1.9 μ m particles, pore size 12 nm; YMC Co., Ltd, Kyoto, Japan) maintained at 30 °C. The following elution program was used with a flow rate of 0.2 ml min⁻¹: 100% A for 5 min, followed by a linear gradient to 66% A: 34% B in 20 min, maintained for 15 min, followed by a linear gradient to 40% A:60% B in 15 min, followed by a linear gradient to 30% A:70% B in 10 min, where A = hexane/2-propanol/formic acid/14.8 M NH_{3aq} (79:20:0.12:0.04 [volume in volume in volume in volume, v/v/v/v]) and B = 2-propanol/water/formic acid/14.8 M NH_{3aq} (88:10:0.12:0.04 [v/v/v/v]). HESI settings were as follows: sheath

gas (N₂) pressure 35 (arbitrary units), auxiliary gas (N₂) pressure 10 (arbitrary units), auxiliary gas (N₂) T 50 °C, sweep gas (N₂) pressure 10 (arbitrary units), spray voltage 4.0 kV (positive ion ESI), capillary temperature 275 °C, S-Lens 70 V. Lipids were analysed with a mass range of m/z 350–2,000 (resolving power 70,000) followed by data dependent MS² (resolving power 17,500), in which the ten most abundant masses in the mass spectrum (with the exclusion of isotope peaks) were fragmented successively (stepped normalized collision energy 15, 22.5, 30; isolation window 1.0 m/z). An inclusion list was used with a mass tolerance of 3 ppm, targeting the ammoniated molecule [C₄₆H₇₉O₈P + NH₄]⁺ of GGGP at m/z 462.2979. The Q Exactive was calibrated within a mass accuracy range of 1 ppm using the Thermo Scientific Pierce LTQ Velos ESI Positive Ion Calibration Solution (containing a mixture of caffeine, MRFA, Ultramark 1621, and *n*-butylamine in an acetonitrilemethanol-acetic solution). Identification of GGGP was aided by the analysis of 1-O-octadecyl-2-hydroxy-*sn*-glycero-3-phosphate (C₁₈-LPA; Avanti Polar Lipids, Inc. Alabama, USA), a structural analog of GGGP, which has a C₁₈ carbon chain attached to the glycerol backbone instead of the geranylgeranyl carbon chain present in GGGP. C₁₈-LPA showed similar chromatographic and mass spectral behaviour to GGGP.

Co-expression of GGGP and DGGGP synthases in *E. coli*

'*E. coli* C43(DE3)' (294) harbouring geranylgeranyl-diphosphate (GGPP) synthase (*crtE*) and G1PDH (*araM*) on plasmid pMS148 (225) was used for expression of the GGGP and DGGGP synthases (encoded on plasmids pABW1, -2 and -3). Cells growing exponentially in LB medium were diluted into magnesium-supplemented terrific broth medium (295) (Mg-TB; 1.2% tryptone, 2.4% yeast extract, 0.4% glycerol, 2 mM MgSO₄, 0.23% KH₂PO₄, and 1.25% K₂HPO₄) to 0.01 OD_{600nm} induced with 0.4 mM IPTG and incubated at 37 °C (200 rpm) for 16 h. Cells (10 ml culture normalized to 1.9 OD_{600nm}) were harvested by centrifugation (4,000 rcf, 10 min) and washed twice with 0.85% NaCl, lyophilized, and then stored at –80 °C. Analysis for production of archaeal-like lipids in *E. coli* was performed by extracting IPLs by a modified Bligh and Dyer extraction (296) that was analysed according to ref. 242 with some modifications: lyophilized cells were extracted 3× with BDE solvent mixture (2:1:0.8) methanol:dichloromethane (DCM):potassium phosphate buffer (50 mM, pH 7) aided by sonication and centrifugation. The extracts were pooled and solvent ratios adjusted to 1:1:0.9, vigorously mixed, centrifuged, and the lower DCM phase transferred. The upper fraction was re-extracted twice with DCM, and the pooled extract was evaporated under a stream of nitrogen and stored dry at –20 °C until analysis. For analysis, samples were dissolved in 200 µl hexane:isopropanol:H₂O (718:271:10) and filtered (0.45 µm) regenerated cellulose. Analysis was performed on an Agilent 1200 series LC (Agilent, San Jose, CA), equipped with thermostatted auto-injector and column oven, coupled to a Thermo LTQ XL linear ion trap with Ion Max source with electrospray ionization (ESI) probe (Thermo Scientific, Waltham, MA). Separation was achieved on a YMC-Pack-Diol-120-NP column (250 × 2.1 µm,

5 μm particles; YMC Co., Ltd, Japan) maintained at 30 °C. Elution program and ESI settings are described in ref. 242. The lipid extract was analysed by an MS routine where a positive ion scan (m/z 400–2,000) was followed by a data dependent MS² experiment where the base peak of the mass spectrum was fragmented (normalized collision energy (NCE) 25, isolation width 5.0, activation Q 0.175). This was followed by a data dependent MS³ experiment where the base peak of the MS² spectrum was fragmented under identical fragmentation conditions. This process was repeated on the 2nd–4th most abundant ions of the initial mass spectrum.

Data availability

The 16S rRNA gene amplicon reads (raw data) have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject ID PRJNA423140, PRJNA649254–57. The *Ca. Cloacimonetes* MAGs are deposited in IMG under the following IMG accession IDs: 134200 (NIOZ-UU1), 134201 (NIOZ-UU2), 134202 (NIOZ-UU3), 151202 (NIOZ-UU4). The 15 metagenomes (raw data), assembly, and MAGs generated in this study are available under BioProject ID PRJNA649215.

Acknowledgements

We thank Melvin Siliakus for providing several of the plasmid constructs and for useful suggestions to improve the manuscript. We are also thankful to Julian Vosseberg, John van Dam, Anja Spang, and Jan de Leeuw for suggestions and constructive discussions. We acknowledge the Utrecht Sequencing Facility (USF), which is partially subsidized by the Hubrecht Institute, Utrecht University, and UMC Utrecht, for the sequencing data and service. We thank Jan Kees van Amerongen, Elda Panoto, Maartje Brouwer, Michele Grego, and Michel Koenen for providing technical support. We acknowledge the crew and scientists of the R/V Pelagia cruises 64PE371 (chief scientist Gert-Jan Reichart) and 64PE408 (chief scientist Marcel van der Meer). J.S.S.D. received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n° 694569—MICROLIPIDS). L.V. and J.S.S.D. receive funding from the Soehngen Institute for Anaerobic Microbiology (SIAM) through a Gravitation Grant (024.002.002) from the Dutch Ministry of Education, Culture and Science (OCW). B.E.D. is supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004. F.A.B.v.M. and B.E.D. are supported by ERC Consolidator grant 865694.

Contributions

L.V., B.E.D., and J.S.S.D. conceived the study. L.V., F.A.B.v.M., and B.E.D. analysed environmental sequencing data and performed phylogenetic analyses. A.B.W.

performed studies of the recombinant proteins. S.Y. performed CARD-FISH and enrichment experiments. A.B.W. and E.C.H. analysed enzymatic assays and cells using LC-MS. L.V., F.A.B.v.M., and J.S.S.D. wrote, and all authors edited and approved the manuscript.

Supplementary Information

Supplementary Materials and methods, Results and discussion

Sampling and sample overview

Four research cruises were performed with the R/V Pelagia to the Black Sea as specified in **Supplementary Table 1**. Suspended particulate matter (SPM) from 15 depths across the water column (50–2,000 m) was collected at sampling station 2 (N42°53.8', E30°40.7', 2,107 m depth) during the Phoxy cruise 64PE371 (BS2013) on 9–10 June 2013. DNA was extracted from the BS2013 SPM 0.7 µm pore size glass fibre filters and used to estimate microbial diversity with 16S rRNA gene amplicon sequencing and quantitative PCR (qPCR) as specified below, as well as metagenomic sequencing (see **Supplementary Table 1**). Archaeal intact polar lipid data of those SPM filters have been previously reported by Sollai et al. (263). At station 2, water was also collected from 1,000, 1,500, and 1,980 m depth during the NESSC cruise 64PE408 (BS2016) on 31 January–2 February 2016 on Sterivex filter cartridges (Millipore). DNA extracted from the BS2016 samples was used for 16S rRNA gene amplicon sequencing only (**Supplementary Table 2a**). Water was also collected from 2,000 m depth during the 64PE444 cruise (BS2018) on 17 August 2018, which was used for attempting enrichment cultures and also to extract DNA which was used for 16S rRNA gene amplicon sequencing and qPCR estimations. Lastly, SPM from 4 depths (500, 1,000, 1,500, and 2,000 m) was collected at sampling station 4 (N42°46.9', E29°21.1', 2,100 m depth) during the 64PE418 cruise (BS2017) on 27 March–5 April 2017. DNA was extracted from the BS2017 samples and used for 16S rRNA gene amplicon sequencing. Except for the BS2017 samples (**Supplementary Table 1**), total lipids were also extracted to quantify archaeal intact polar lipids as specified below.

Diversity estimates by 16S rRNA gene amplicon sequencing and quantification

Microbial diversity was estimated by 16S rRNA gene amplicon sequencing of the SPM collected during the four cruises (**Supplementary Table 1**). In the case of the BS2013 samples, 16S rRNA gene amplicon sequencing was performed initially with 454 GS FLX sequencing as described in Moore et al. (271) and Besseling et al. (297) (**Supplementary Table 2a**) and later repeated with Illumina MiSeq 2 × 300 bp as described in van Grinsven et al. (298) to increase sequence resolution and allow for comparison with the 16S rRNA gene sequencing results of later campaigns.

In the BS2013 samples, 16S rRNA gene reads attributed to *Ca. Cloacimonetes* ranged from 5 to 16% of the total bacterial plus archaeal 16S rRNA gene reads between 500 and 2,000 m depth (**Supplementary Table 2**). Most sequences belonged to the *Ca. Cloacimonetes* MSBL2 group (13%), while the sequences closely related to *Ca. Cloacimonetes* group MSBL8 represented 3% of the total 16S rRNA gene reads at

2,000 m depth (Fig. 1a, Supplementary Table 2a). For the BS2013 SPM samples, the relative fraction of reads attributed to *Ca. Cloacimonetes* was similar for the 454 GS FLX sequencing and Illumina MiSeq runs (Supplementary Table 2b) while the relative fraction of archaeal 16S rRNA gene reads was lower in the Illumina MiSeq libraries. Moreover, the relative fraction of *Ca. Cloacimonetes* reads was similar in the SPM of BS2013 and BS2016 with percentages ranging from 6 to 14% of total reads between 1,000 and 2,000 m depth (Supplementary Table 2), while the percentage of *Ca. Cloacimonetes* reads in station 4 of BS2017 was slightly higher (from 7 to 20% from 500 to 2,000 depth). qPCR using primers 515F/806RB was also performed as described in ref. 298 with DNA extracted from SPM from 500, 1,000 and 2,000 m depth obtained during BS2013, SPM from 1,000 and 2,000 m depth collected during BS2018, as well as all SPM samples of station 4 of BS2017. Total 16S rRNA gene abundance estimates were similar from 500 to 2,000 m depth in both station 2 in BS2013, BS2018 and in station 4 in BS2017 with average values of 4×10^7 16S rRNA gene copies per litre.

Microscopic detection and characterization of *Ca. Cloacimonetes* at 2,000 m depth

Water samples were also collected at 2,000 m depth during the BS2018 sampling as specified above (see Supplementary Table 1). Water was collected with Niskin bottles under N_2 overpressure directly into glass pressure bottles acid-washed, autoclaved and also overpressurised with N_2 to minimize exposure to oxygen and keep anoxic conditions. This water was used on board to start enrichments as described in the following section. The presence of *Ca. Cloacimonetes* was assessed in both the original samples and in the enrichments by Catalyzed reporter deposition Fluorescence In Situ Hybridization (CARD-FISH) with the specific HRP-labelled probe Cloa1 5'-GGT TGT GCC CCT TCG GGG G-3', which was designed based on the 16S rRNA gene fragment sequence obtained from the MAG NIOZ-UU1. The CARD-FISH protocol was performed as described earlier (ref. 299 and http://www.environmental-microbiology.de/pdf_files/CARDFISH_2march2013.pdf). To avoid cell loss during cell wall permeabilization, filters were dipped in low-gelling-point agarose (0.2% [wt/vol] in MQ water, dried face up on glass slides at 30 °C, and subsequently dehydrated in 96% (vol/vol) ethanol for 1 min. To inhibit endogenous peroxidases water samples from the Black Sea at 2,000 m depth were treated overnight with 0.1% H_2O_2 at room temperature for 2 min. For cell wall permeabilization, filters were incubated in a lysozyme solution (10 mg ml⁻¹ in 0.05 M EDTA, 0.1 M Tris-HCl [pH 7.5]) at 37 °C for at least 30 min. The sections were washed with MQ water, dehydrated with 96% ethanol, dried at room temperature, and subsequently stored in petri dishes at -20°C until further processing. The optimal stringency for the Cloa1 probe consisted of 55% formamide. A volume of about 30 ml of water collected from the Black Sea at 2,000 m depth as described above was filtered on a 0.22 µm 24 mm diameter polycarbonate filter. A piece of about 1 cm² of filter paper was cut and used for CARD-FISH analysis with the Cloa1, EUB388 (general bacteria), and Arc915 (for

archaea) HRP-labelled probes (**Supplementary Fig. 1**). Visualization was performed on an Axio Imager.M2 Microscope system (Zeiss).

Double staining with EUB338-Alexa488 and Cloa1-Alexa555 (in yellow, **Supplementary Fig. 1b**) showed the presence of *Ca. Cloacimonetes* cells among the total bacteria. Likewise, the double staining with EUB338-Alexa488 and Arch915-Alexa555 (in yellow, **Supplementary Fig. 1e**) showed the presence of archaeal cells among the total bacteria, indicating that the archaeal cells were a minority among the total cells and of smaller size than the *Ca. Cloacimonetes* cells. *Ca. Cloacimonetes* cells were oval to small rods in shape. The oval shaped cells were 0.8–0.9 μm in diameter whereas small rods were 0.8–9 μm wide and 2–3 μm long.

Genome-centric metagenomics of the Black Sea water column

Assembly and binning of the metagenomes from the SPM samples collected during the BS2013 cruise generated 181 MAGs. NIOZ-UU1–4 fell within the *Ca. Cloacimonetes* phylum (**Fig. 1c**, **Supplementary Fig. 2**). These four MAGs are estimated to be substantial to near complete with low to no detectable contamination (**Supplementary Table 3**). Due to difficulties in metagenome assembly of the 16S rRNA gene, only NIOZ-UU1 contained a copy. In support of our shotgun metagenomics and bioinformatic assembly pipeline, we built a maximum likelihood phylogeny of the representative 16S rRNA gene sequences of the amplicon sequencing analysis assigned to the phylum *Ca. Cloacimonetes*, 16S rRNA gene sequences of closely related species obtained from ARB-SILVA (273), the 16S rRNA gene sequence obtained from NIOZ-UU1, and the 16S rRNA gene sequences found in the *Ca. Cloacimonetes* genome placed in **Fig. 1c** (see also **Supplementary Fig. 2a**, **Supplementary Table 4**). The tree was inferred with MEGA6 (ref. 300) with the generalised time reversible model and gamma distribution, employing 1,000 bootstraps. The comparison of this tree with the concatenated marker gene phylogeny suggests that NIOZ-UU1, 3 and 4 are closely related to MSBL2 group *Ca. Cloacimonetes* genomes (**Supplementary Fig. 2**). Mapping shotgun metagenomic sequencing reads back to the cross-assembly revealed that NIOZ-UU3 (MSBL2) was the most abundant *Ca. Cloacimonetes* MAG at 2,000 m depth and showed comparable abundance patterns between MSBL8 and NIOZ-UU2 (**Fig. 1**).

Predicted metabolism of the Black Sea *Ca. Cloacimonetes* MAGs

Various enzymes related to anaerobic lifestyles were detected in NIOZ-UU1 and NIOZ-UU3 (see **Supplementary Tables 5–8** for functional annotation of NIOZ-UU1 to 4) including ribonucleoside triphosphate reductase, ferredoxin oxidoreductases, and radical S-adenosylmethionine-dependent proteins, which indicated that *Ca. Cloacimonetes* is well adapted to the permanent anoxic conditions of the Black Sea. However, presence of genes related to microaerophilic growth (e.g. superoxide reductase (EC 1.15.1.2), ruberythrin and thioredoxin reductase (EC 1.8.1.9)) in

NIOZ-UU1 and NIOZ-UU2 (**Supplementary Table 6**) also indicated that the *Ca. Cloacimonetes* found in our samples may also thrive in the suboxic zones of the Black Sea. α -amylase (EC 3.2.1.1) and β -glycosyl hydrolase, which are involved in the hydrolysis of polysaccharides (cellulose and starch), were detected in NIOZ-UU1 and NIOZ-UU3, supporting their capabilities to obtain energy by hydrolysing polysaccharides. Furthermore, annotation of NIOZ-UU1 and NIOZ-UU3 also confirmed the presence of genes responsible for the production of ethanol (alcohol dehydrogenase; EC 1.1.1.1).

NIOZ-UU MAGs also harbour genomic indications of a syntrophic lifestyle and of propionate oxidation. Their genomes harbour NiFe dependent hydrogenases, which are known to couple the oxidation of reduced ferredoxin to the production of H₂ during carbohydrate and protein fermentation. NIOZ-UU1 and NIOZ-UU3 also harbour various ferredoxin oxidoreductases, which are primarily involved in amino acid fermentation. Evidence for syntrophic propionate oxidation has previously been observed for *Ca. Cloacimonetes* '*Candidatus Syntrophosphaera thermopropionivorans*', and propionate oxidation was speculated to be performed via methylmalonyl CoA but with absence of the complete methylmalonyl CoA pathway genes (301). In the case of the NIOZ-UU MAGs, evidence for this process are the presence of genes coding for the Propionyl-CoA carboxylase β chain (EC 6.4.1.3) and the Malonyl CoA-acyl carrier protein transacylase (EC 2.3.1.39). Genes involved in glycolysis were also present in NIOZ-UU1 and NIOZ-UU3. However, the complete absence of the electron transfer chain involved in anaerobic respiration indicates that these *Ca. Cloacimonetes* likely obtain their energy through the hydrolysis of polysaccharides, glycolytic pathway and ultimately the fermentation of sugars and amino acids.

The various genes detected in NIOZ-UU1 and NIOZ-UU3 indicated that these *Ca. Cloacimonetes* might be actively involved in the degradation of polysaccharides and amino acids sinking from the upper oxic zones of the Black Sea water column. Furthermore, a complete set of the heterodisulphide reductase (HDR) system was identified in those MAGs. The HDR system serves as an elemental sulphur oxidation enzyme in the cytoplasmic space of bacteria and archaea (302). The HDR system catalyses the reversible reduction of the disulphide bond (R-S-S-R) coupled with energy conservation (303). The presence of this system in the *Ca. Cloacimonetes* MAGs is expected to be associated with energy generation and conservation by sulphide oxidation in the deeper water column of the Black Sea, where the sulphide concentration can be up to 425 μ M (304).

Moreover, a set of genes required for assembly of type IV pili (i.e. Type IV prepilin-like) is encoded in the NIOZ-UU MAGs. Type IV pili can be involved in motility, adherence, DNA uptake, and carrying electric current during direct interspecies

electron transfer (DIET) (305), suggesting that the *Ca. Cloacimonetes* detected in the Black Sea may be capable of DIET. Previous studies have also indicated that *Ca. Cloacimonetes* can be involved in cellulose or sugar degradation (ref. 306, among others), which is further supported by the presence of Glucan endo-1,3- β -glucosidase A1 and Glucosidase YgjK coding genes in NIOZ-UU3. Based on the predicted metabolism of the *Ca. Cloacimonetes* present in the Black Sea water column, we designed a culture enrichment strategy as specified below.

Attempts of enrichment and isolation of *Ca. Cloacimonetes*

We followed different strategies for the enrichment of *Ca. Cloacimonetes* using different growth media by considering the in situ environmental parameters of the deep water column of the Black Sea and by extracting information from the four *Ca. Cloacimonetes* MAGs. Six different media were prepared for the selective enrichments. Growth medium 1 (GM1) contained (g L^{-1} , pH 7.0) cellulose (2.0); tryptone (2.0); yeast extract (1.0); $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$ (1.0), NaCl (20.0), $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$ (3.6), $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ (4.3), KCl (0.5), $\text{Na}_2\text{S} \cdot 9\text{H}_2\text{O}$ (10 mg). Growth medium 2 was made by 100 times dilution of GM1, while growth medium 3 contained (g L^{-1} , pH 7.0) cellobiose (0.1); KCl (0.55); Na_2SO_4 (2.34); yeast extract (1.0); $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$ (2.38), NaHCO_3 solution (1 ml of 10% w/v); NaCl (20.0), $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$ (8.8), ferric citrate solution (5 ml of 0.1% w/v) and sterilized (i.e. autoclaved) Black Sea water (994 ml). Growth medium 4 (GM4) was made by amending GM2 with 10 ml L^{-1} of autoclaved cell lysates. Cell lysates were raised from Black Sea water enriched with GM1. GM5 was made by amending a basal medium containing (L^{-1}) Na_2HPO_4 (5.51 g); KH_2PO_4 (3.4 g); $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ (0.2 g); $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$ (0.06 g); $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ (0.5 mg); vitamin solution (2 ml); and trace elements (1 ml) with different amino acids (lysine (4.56 g); leucine (13 mg); isoleucine (13 mg); valine (17 mg); threonine (19 mg); methionine (14 mg); proline (11 mg); arginine (17 mg); histidine (20 mg); phenylalanine (16 mg); cysteine (12 mg); and tryptophan (4 mg). Vitamins and trace elements were prepared as described earlier (307,308). The amino acids and the cofactors were freshly prepared, filter sterilized, and added to the medium after being autoclaved. The headspace was flushed with ultrapure nitrogen. The enrichments were maintained in 120 ml serum vials containing 50 ml of the medium. These anaerobic enrichments were kept at 10 °C and 28 °C in order to enrich for psychrophilic and mesophilic bacterial members, respectively. Subsamples of 50 μl of the enrichments were pipetted on 0.22 μm 24 mm diameter polycarbonate filters and CARD-FISH was performed as described above. Filters were mounted on microscopic slides with mounting medium containing DAPI and analysed on an Olympus microscope with 100 \times magnification. Only the GM4 enrichment showed positive signals with the Cloa1 probe (data not shown). After 10 days of incubation, the enrichments were streaked on the agar media mentioned above and grown in anoxic conditions in anaerobic jars supplemented with Anaerocult® (VWR, The Netherlands). For the isolation and purification of *Ca. Cloacimonetes* from the enrichments, samples were streaked on solid media (1.8%

agar). None of the obtained cultures, however, showed a close affiliation with *Ca. Cloacimonetes* based on 16S rRNA gene sequence analysis.

Archaeal lipid biosynthetic genes in *Ca. Cloacimonetes* and beyond

All four *Ca. Cloacimonetes* MAGs contained homologs of the archaeal lipid biosynthetic pathway (see ‘Materials and methods’; **Supplementary Fig. 3** for details), including a homolog of the geranylgeranylgeranyl phosphate (GGGP) synthase, co-localized with an homolog of (S)-2,3-di-O-geranylgeranylgeranyl phosphate (DGGGP) synthase, the two enzymes that mediate the formation of the first and second ether bond in archaeal membrane lipids, respectively (**Supplementary Table 9, Supplementary Fig. 3**). Homology searches show that the two genes are found in other Bacteria and Archaea in the Black Sea water column as well (**Supplementary Figs. 4 and 5, Supplementary Table 10**). Moreover, an extended search in publicly available genomes from other environments and across the tree of life shows that the two genes occur widespread in the FCB group superphylum and related candidate phyla (**Fig. 4, Supplementary Tables 11 and 12**) although only co-localized in *Ca. Cloacimonetes* (**Supplementary Table 13**). Homologs of the two genes were found in representatives from *Ca. Cloacimonetes*, Bacteroidetes, Chlorobi, Calditrichaeota, ‘candidate division Zixibacteria’, *Candidatus* Kryptonina, Gemmatimonadetes, Ignavibacteriae, *Candidatus* Edwardsbacteria, *Candidatus* Raymondbacteria, ‘candidate division KSBI’ and ‘candidate division TA06’, and in one genome of *Ca. Parcubacteria* of the ‘Candidate Phyla Radiation’ (104). The two genes are found in *Candidatus* Marinimicrobia and Fibrobacteres as well, but not together in single genomes. Genomes with only one of the genes span an even larger part of the tree of life (**Supplementary Tables 11 and 12**).

Validation of the metagenomic assembly and co-localization of GGGP and DGGGP synthase in silico and experimentally

We thoroughly tested and rejected the possibility that the co-localized bacterial GGGP and archaeal DGGGP synthase coding genes in the *Ca. Cloacimonetes* MAGs could have been introduced by a methodological error. First, chimeras could have been produced during cross-assembly, i.e. sequencing reads that are not derived from the same species in the original sample could have been merged into scaffolds. Second, scaffolds could have been binned erroneously into MAGs, e.g. because of relaxed binning parameters or because binning signals (coverage across samples and TNF) are very similar between species. To address the first hypothesis, we plotted coverage along the full length of the scaffolds of interest in their deepest covered samples, and found the scaffolds evenly covered (**Supplementary Fig. 6**), the remaining small-scale peaks in part reflecting biases introduced during the TruSeq nano library preparation. Importantly, the GGGP and DGGGP synthase ORFs are co-localized on a single assembled scaffold in NIOZ-UU1, NIOZ-UU3, and NIOZ-UU4, and read coverage across them is continuous (**Supplementary**

Figs. 6a and 6c–d). Whereas GGGP and DGGGP synthase coding genes are not directly co-localized on the same scaffold in NIOZ–UU2, one read-pair bridges the two scaffolds and an alignment between the two scaffolds and GGGP synthases in the three other MAGs shows the two genes to be concatenated as well in NIOZ–UU2 (data not shown).

Moreover, the scaffold of interest in NIOZ–UU3 carries both the archaeal lipid biosynthesis genes and a bacterial (node ID: UID2495) marker gene, PriA (primosomal protein N') from the TIGR00595 family, which is present single-copy in the MAG, and again, connected to the lipid genes through even coverage (**Supplementary Fig. 6c**). We, therefore, conclude that the scaffolds are not chimeras. To address the second hypothesis of erroneous binning, we compared the average read coverage profiles across samples for the scaffolds of interest. Importantly, for the four MAGs, the scaffolds of interest show comparable coverage profiles across samples as the full MAGs (**Supplementary Fig. 7**). Moreover, each of the four MAGs has a unique coverage profile (data not shown). Thus, we have no reason to suspect that the scaffolds of interest are archaeal sequences that were erroneously binned with *Ca. Cloacimonetes* scaffolds. Furthermore, subsequent database searches identified the presence of both the GGGP and DGGGP synthase encoding genes in a range of other FCB group and related candidate phyla genomes as well, further supporting the existence of a bacteria synthesizing both bacterial and archaeal membrane lipids (**Supplementary Tables 11 and 12**).

We additionally verified our binning-analysis experimentally by PCR and sequencing: DNA extracted from the 1,000 and 2,000 m depth SPM obtained during BS2013 were PCR-amplified using primers designed to span the GGGP and DGGGP synthases, polyprenyl synthase and the marker gene PriA in the specific scaffold of NIOZ–UU3 (**Supplementary Table 9**). The resultant amplicon only from the 2,000 m depth SPM sample was cloned and sequenced, confirming our assembly (data not shown).

Quantification of GGGP and DGGGP synthase coding gene and gene expression

In order to evaluate if the *Ca. Cloacimonetes* 'archaeal'-like membrane lipid biosynthetic pathway is functional, we determined the transcriptional activity of the GGGP, DGGGP, and polyprenyl synthase genes by a RT-PCR approach. The positive expression of those genes was confirmed in the SPM 1,000 and 2,000 m depth samples of BS2013 (**Supplementary Fig. 8**). No amplification was detected when a negative control of the reverse transcription reaction (RNA extract without reverse transcriptase) was included as DNA template of the PCR reaction (**Supplementary Fig. 8**). Amplified fragments were further sequenced to confirm the products (data not shown).

In addition, we performed a qPCR approach to quantify the NIOZ-UU3 GGGP and DGGGP coding genes both at the DNA and RNA level and through the vertical profile of SPM from 50 to 2,000 m depth of the BS2013 campaign (15 depths). The *Ca. Cloacimonetes* GGGP synthase coding gene was detected from 250 m depth downwards with a maximum of 1.7×10^7 copies per litre at 2,000 m depth (qPCR efficiency = 80%, $R^2 = 0.997$; **Supplementary Fig. 9**). Gene transcripts of the *Ca. Cloacimonetes* GGGP synthase coding gene were also detected from 250 m depth downwards with a minimum value at this depth (2.7×10^2 copies L^{-1} ; detection limit qPCR assay estimated to be 15 copies L^{-1}), which increased with depth to a maximum of 1.8×10^4 copies L^{-1} at 2,000 m depth. These results point to an increasing number of *Ca. Cloacimonetes* GGGP synthase genes with depth, in agreement with the estimation of *Ca. Cloacimonetes* 16S rRNA copies L^{-1} based on 16S rRNA gene amplicon sequencing and the 16S rRNA gene qPCR assay as estimated above. Gene expression was detectable but low. It is likely that gene expression of this gene may be affected by the sampling procedures we are applying, as the SPM samples do not get fixed until they are retrieved on the deck of the ship, after they have gone through decompression from deep waters (i.e. 2,000 m depth) to surface in a short period of time (approximately 30 min). These factors have been seen to significantly affect the gene expression profile of deep sea samples (309). Similarly, the *Ca. Cloacimonetes* DGGGP synthase abundance and gene expression was evaluated by qPCR, with qPCR efficiency too low for an accurate quantification. However, we estimated the *Ca. Cloacimonetes* DGGGP synthase gene to be detectable from 250 m depth downwards and with an increasing abundance reaching maximum values at 2,000 m depth (**Supplementary Fig. 10a**). *Ca. Cloacimonetes* DGGGP synthase gene transcripts were also detected from 250 m downwards, which was supported by a positive qPCR signal and correct melting behaviour in the melting curve (**Supplementary Fig. 10b**).

Biochemical verification of the encoded archaeal-like lipid biosynthesis proteins

In order to confirm the enzymatic function of the putative GGGP synthases annotated in the four *Ca. Cloacimonetes* MAGs, the amplified GGGP synthase open reading frame (ORF) from NIOZ-UU3 (**Supplementary Table 14**) was expressed from a T7-promoter in plasmid pABW4 in '*E. coli* BL21(DE3)', and the 6His tagged protein was purified by Ni-NTA affinity chromatography (see '**Materials and methods**', **Supplementary Table 15**). Purity was verified using 12% TGX™ precast gels (Bio-Rad), stained with Bio-Safe™ Coomassie stain (Bio-Rad). The purified recombinant *Ca. Cloacimonetes* GGGP synthase (**Supplementary Fig. 11**) used in the enzymatic assay below was identified based on the predicted size. Note that three larger bands were also present (**Supplementary Fig. 11b**) and tentatively identified as multimers of the purified GGGP synthase, which is in line with the observations of Peterhoff et al. (253), who reported that all group I GGGP synthases are dimers, while group II GGGP synthases enzymes are either dimers or hexamers.

The enzymatic activity of the protein was tested in an assay with geranylgeranyl-diphosphate (GGPP, 20 carbons) and either G1P or G3P as substrates to further test the stereo-selectivity of the enzyme, performed in duplicate. Samples were analysed by UPLC-HRMS (see ‘**Materials and methods**’), and GGGP formation (detected as GGGP ammonium adduct [GGGP + NH₄]⁺ (**Fig. 2a, Supplementary Fig. 12**) and GGGP-sodium adduct [GGGP + Na]⁺, not shown) was observed both in the presence of G1P and G3P (**Fig. 2a, Supplementary Fig. 12**), consistent with previous studies of archaeal GGGP synthases (225,259,260). The identification of GGGP was confirmed by MS² fragmentation analysis of the GGGP produced (**Supplementary Fig. 13a**) and furthermore had a retention time and MS² fragmentation pattern consistent with the structural analogue C₁₈-lyso phosphatidylglycerolphosphate (1-O-octadecyl-2-hydroxy-*sn*-glycero-3-phosphate; Avanti Polar Lipids, Cat no. 857228). No GGGP formation was detected in the absence of enzyme, GGPP or glycerol-phosphate. For GGGP-positive samples, some variation in the amount of GGGP was observed and the use of G1P appeared to result in increased GGGP levels compared to G3P (**Supplementary Fig. 12**), consistent with the reported preference of archaeal GGGP synthases for G1P over G3P (225,253,259,260).

To test whether the GGGP and DGGGP synthases encoded by the *Ca. Cloacimonetes* MAGs could support the formation of archaeal-like lipids in a bacterium, we co-expressed NIOZ-*UU3* GGGP and DGGGP synthases (encoded on plasmids pABW1, -2 and -3) in ‘*E. coli* C43(DE3)’, a strain optimized for membrane protein production (294) that contained plasmid pMS148 (ref. 310). Plasmid pMS148 encodes GGPP synthase (CrtE) and G1PDH (AraM), enzymes that produce G1P and GGPP, respectively, the likely biosynthetic substrate for the *Ca. Cloacimonetes* GGGP synthase. For analysis, lipids were extracted from cells and analysed by Ultra High Pressure Liquid Chromatography – High Resolution Mass Spectrometry (UHPLC-HRMS) for the formation of archaeal lipid intermediates (see ‘**Materials and methods**’).

Cells expressing both GGGP and DGGGP synthases produced significant amounts of phosphatidylglycerol archaeol with 8 double bonds or unsaturations (**Fig. 2b, Supplementary Fig. 14**; PG-unsat(8)-archaeol, i.e. an octaunsaturated (8 double bonds or unsaturations) archaeol, *n*-2,3-diphytanyl-glycerol diether with isoprenoid chains of 20 carbons, with a phosphatidylglycerol head group, also known as phosphatidylglycerol digeranylgeranyl-glyceryl phosphate), the expected intermediate in the biosynthesis of archaeal membrane lipids in the absence of a specific geranylgeranyl reductase in *E. coli* (255,256). The identity of this compound was verified using MS² fragmentation analysis and the interpretation of the fragmentation spectrum of PG-unsat(8)-archaeol was based on Yoshinaga et al. (311) (**Supplementary Fig. 13b**). No PG-unsat(8)-archaeol was detected in cells lacking

either GGPP or DGGPP synthase, or in cells lacking the upstream enzymes GGPP synthase and GIPDH (encoded on pMS148) (**Supplementary Fig. 14**).

Other lipid biosynthetic genes in *Ca. Cloacimonetes*

We subsequently focused on genes of the lipid biosynthesis pathways other than GGPP and DGGPP synthase encoding genes present in the four Black Sea MAGs affiliated with *Ca. Cloacimonetes*. The gene coding for the glycerol-3-phosphate dehydrogenase (G3PDH; *gps* gene) catalysing the formation of G3P was detected in three of the MAGs (NIOZ-UU1, 3, and 4; **Supplementary Table 16**). In addition to genes for GGPP and DGGPP synthase, other genes required for the synthesis of isoprenoidal archaeal lipids were also detected in the *Ca. Cloacimonetes* MAGs, including the genes for a complete bacterial isoprenoid MEP/DOXP pathway (**Supplementary Table 17**), genes coding for acetyl-CoA C-acetyltransferase and hydroxymethylglutaryl-CoA synthase of the Mevalonate pathway (see **Supplementary Table 17, Supplementary Fig. 3**), and two polyprenyl synthases (**Supplementary Table 17**). In addition, several genes of the bacterial fatty acid pathway were detected in the MAGs (**Supplementary Table 18**), including the acyltransferases responsible for the esterification of G3P, as well as genes coding for enzymes involved in downstream reactions (**Supplementary Table 16, Supplementary Fig. 3**). The complete annotation of NIOZ-UU1 to 4 is found in **Supplementary Tables 5–8**.

A phylogenetic analysis of the polyprenyl synthases indicated the presence of close relatives in other *Ca. Cloacimonetes* genomes, and they were classified as either short-chain or geranylgeranyl-diphosphate (GGPP) synthases and medium-chain length prenyltransferases based on their sequence (253) (**Supplementary Fig. 15a**). Finally, the four *Ca. Cloacimonetes* MAGs also included a putative digeranylgeranyl-glycerophospholipid reductase (identified based on homology with experimentally verified geranyl reductase from ref. 310) that is also closely related to homologs of the *Ca. Cloacimonetes* and FCB group superphylum genomes. These homologs are closely related to putative digeranylgeranyl-glycerophospholipid reductases of archaeal genomes of the Euryarchaeota (**Supplementary Fig. 15b**). Both these two maximum likelihood phylogenetic analyses were performed with PHYML v3.0 (ref. 312) using the model indicated by ProtTest v2.4 (ref. 313) (LG model plus gamma distribution and invariant site, LG+G+I). Sequences were aligned with MUSCLE (314). Alignments were trimmed with Gblocks v0.91b (ref. 315) using relaxed parameters and manually curated.

Potential sources of archaeal membrane lipids in the Black Sea deep water column

Archaeal lipid diversity and abundance was estimated in the SPM samples collected from station 4 (500, 1,000, 1,500 and 2,000 m depth) during BS2017. The diversity of IPL-derived CLs detected in station 4 in 2017 was similar to that observed in

station 2 in 2013 (see ref. 263) with predominance of GDGT-1 and 2 as well as archaeol (**Supplementary Table 19**). The total archaeal IPLs abundance increased from 2.5 to 25 ng per litre from 500 to 2,000 m depth (**Supplementary Table 19**). We next assessed whether the archaeal IPLs observed in the water column could be attributed to the living archaeal cells. All calculations are available in **Supplementary Notebook 1**. Absolute abundances of total 16S rRNA gene copies per litre were determined by qPCR as described above (**Supplementary Table 2**) in the samples from which IPLs were measured. Cells per litre for a given taxon at each depth was estimated by multiplying the fraction of total 16S rRNA gene amplicon sequencing reads attributed to that taxon with the total 16S rRNA gene copies per litre estimated by qPCR and dividing by the expected 16S rRNA gene copy number in the genome (**Supplementary Notebook 1**). For the archaeal groups we assumed one 16S rRNA gene per genome, and for *Ca. Cloacimonetes* two, as the genome of “*Ca. Cloacimonas acidaminovorans*” str. Evry’ contains two copies (<https://rrndb.umms.med.umich.edu/>, ref. 56). We estimated the IPLs abundances that the observed archaeal cells could theoretically produce at each depth based on the estimated abundances of archaeal cells per group, their estimated cell size based on literature, and different models for lipid production per cell surface area (**Supplementary Notebook 1**). We took uncertainty in measurements for both lipid production and cell size estimates into account by including the most extreme cases reported in literature, arriving at a range for the amount of IPLs that the archaeal population could produce. The maximum of this range represents an ideal case scenario, where all archaea in the water column are at maximum known size and produce the maximum reported number of membrane lipid molecules.

Lipid abundance estimates were based on the proposed 0.86–1.85 femtograms (fg) of archaeal lipids per cell for rods sized 0.5–0.9 μm length \times 0.2 μm width (316), 1 fg per cell for 0.8 \times 0.5 μm rods (317), and 0.25 fg per cell for 0.5 \times 0.15 μm rods (318). Cell size estimates for the different archaeal groups were: Thermoplasmatales (rods 0.5–3 μm length \times 0.2–0.5 μm width (319)), ANME-1 (cells within aggregates 1.2 μm length \times 0.3–0.4 μm width (320)), *Candidatus* Bathyarchaeota (spherical cells 0.4–0.5 μm size (321)), and DPANN ‘*Candidatus* Woesearchaeota’ (spherical cells between 400 and 500 nm diameter inferred based on the diameter of the DPANN ‘*Nanoarchaeum equitans*’, 400 nm (322) and the 500 nm diameter of the DPANN ‘ARMAN’ Nanoarchaea (323)). For the remaining archaeal cells (referred to as “archaea, others” in **Supplementary Table 2**) we took a range between 0.25 and 5 femtograms of archaeal lipids per cell.

We arrive at a predicted IPLs concentration of <6.5 ng per litre based on the archaeal population (**Supplementary Notebook 1**). This represents a striking offset at 1,000, 1,500 and 2,000 m depth ranging from two- to fivefold between observed archaeal IPLs and predicted IPLs. This means that even the most ideal situation, where all

archaea in the water column are of maximum known size and produce the maximum reported number of membrane lipid molecules, cannot explain the observed amount of archaeal IPLs at deeper depths. An explanation for this mismatch could be that suspended IPLs are preserved for a long time in the anoxic waters of the Black Sea. In deep anoxic sediments, preservation as fossils of archaeal IPLs with stable glycosidic bonds has been observed (e.g. ref. 264). However, this seems an unlikely explanation as the IPLs reported here are found in the water column, which is expected to have a much higher degradation rate than sediments. Moreover, most of the archaeal IPLs detected in the Black Sea do not contain the stable glycosidic-based polar head groups but rather the more labile headgroups phosphatidylglycerol, phosphatidylserine, or phosphatidylethanolamine (263). We, therefore, argue that the mismatch provides evidence for the production of archaeal-like ether-linked isoprenoid membrane lipids by the highly abundant *Ca. Cloacimonetes* bacteria at these depths, which are 2–4 times more abundant than the total archaeal population.

Archaeal lipid biosynthetic genes in *Ca. Cloacimonetes* and other bacteria

The GGGP synthases of the Black Sea *Ca. Cloacimonetes* MAGs have close homologs in other recently released *Ca. Cloacimonetes* genome sequences and in members of the FCB group superphylum and related candidate phyla (**Fig. 4, Supplementary Fig. 16, Supplementary Tables 11 and 12**). Our phylogeny shows a clear separation between group I and group II GGGP synthases in line with earlier findings (247), with the *Ca. Cloacimonetes* sequences falling within group II (**Fig. 4, Supplementary Fig. 16**). Crenarchaeota contain the closest archaeal relatives to the detected GGGP synthases (**Fig. 4, Supplementary Fig. 16**), however the phylogeny of this enzyme is inconclusive regarding the origin of the GGGP synthase homologs in the FCB group superphylum.

The extended presence and close phylogenetic associations of the GGGP synthase homologs in this superphylum and related candidate phyla strongly supports the presence of this enzyme in Bacteria at least before radiation of the FCB group. Thus, this GGGP synthase could be a ‘remnant’ of the ‘mixed membrane’ stage after LUCA and before the diversification of Bacteria. Alternatively, the GGGP synthase could have been transferred after the bacterial membrane origin from Archaea to the ancestor of the FCB group and related bacterial phyla, in line with earlier suggestions (247).

Like GGGP synthase, the DGGGP synthase genes detected in the four *Ca. Cloacimonetes* MAGs (**Supplementary Table 9**) also have close homologs in other *Ca. Cloacimonetes*, FCB group superphylum, and recently released candidate phyla genomes (**Fig. 4, Supplementary Fig. 17; Supplementary Tables 11 and 12**). The DGGGP synthases of *Ca. Cloacimonetes* are closely related to those of TACK group genomes, in particular Crenarchaeota and *Candidatus* Korarchaeota (**Fig. 4,**

Supplementary Fig. 17). The topology of the phylogeny is similar to that of GGGP synthase with respect to sharing the TACK group as sister clade and branching of bacterial clades, suggesting that the two genes share a similar evolutionary history.

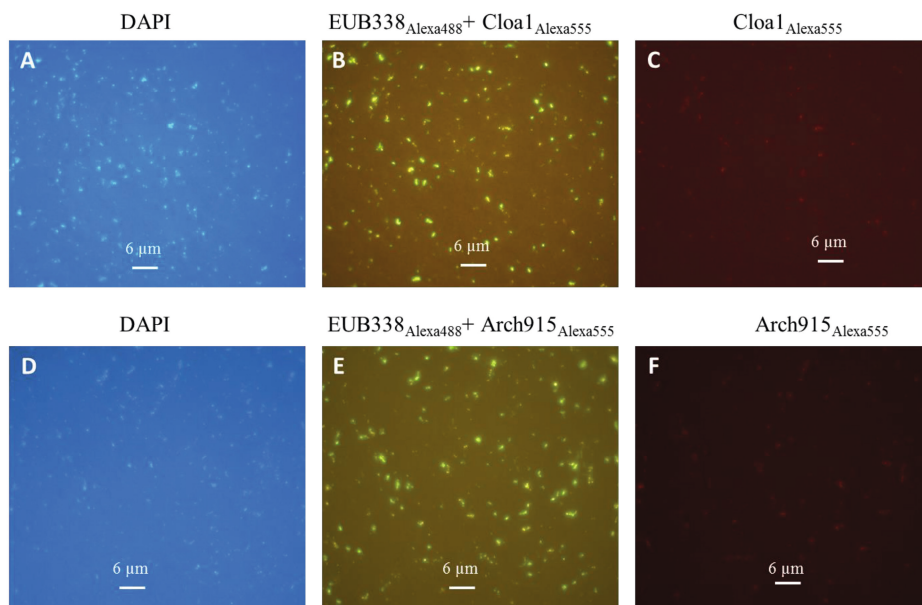
Interestingly, many genomes from the class Flavobacteriia within the phylum Bacteroidetes contain the gene in double copy (**Supplementary Table 11**), suggestive of a duplication within the phylum. Sequence identity between the Bacteroidetes cluster and other bacterial/archaeal sequences is low (**Supplementary Fig. 18**), which might be an indication for a divergent activity of the gene in the phylum. DGGGP synthases belong to a superfamily of UbiA prenyltransferases (324) including prenyltransferases of the ChlG/BchG (i.e. chlorophyll *a* synthase ChlG from '*Synechosystis* sp. strain PCC6803', accession number BAA10281) and UbiA/COQ2 prenyltransferases for the biosynthesis of ubiquinone (AAK40480.1 ubiA-1 *Sulfolobus solfataricus*; CAA96321.1 COQ2 *Saccharomyces cerevisiae*; AAC43134.1 4-hydroxybenzoate octaprenyltransferase UbiA *Escherichia coli*), therefore we included some of these sequences to possibly elucidate the LUCA or HGT origin of the DGGGP synthase gene. However, the placement of these 'outgroups' close to the base of the archaeal tree does not unequivocally solve the question of the origin of the gene; the phylogenetic signal in DGGGP synthase is insufficient to identify the archaeal species tree, wherever a root is placed.

Screening of membrane lipid biosynthetic genes in the Asgard archaea

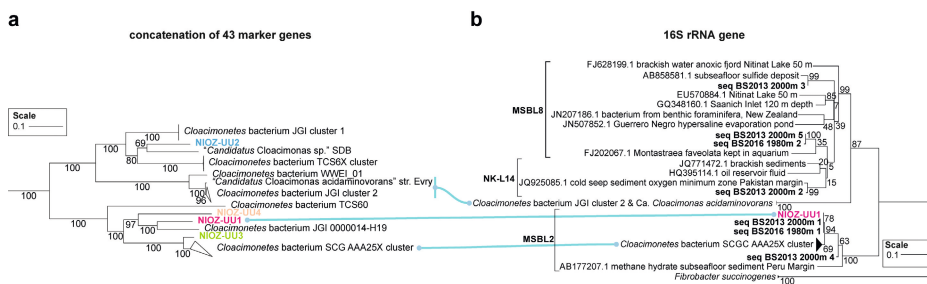
Previous studies have reported that two uncultured archaeal groups, the Euryarchaeota 'Marine Group II' and *Ca.* Lokiarchaeota of the Asgard superphylum, contain archaeal lipid biosynthesis genes alongside bacterial-like fatty acid and ester-bond formation genes, but apparently they lack the gene coding for glycerol 1-P-dehydrogenase (G1PDH), suggesting that they are not able to synthesize the typical archaeal-like lipids with G1P stereochemistry (249). This observation is interesting as these genomes also have the genes required to produce G3P, making it possible that they synthesize archaeal-like lipids with G3P stereochemistry. The study by Caforio et al. (225) suggests that G1P could be synthesized in the absence of the typical G1PDH, thus archaea of the 'Marine Group II' Euryarchaeota and some of the Asgard archaea might still synthesize G1P archaeal-like lipids. They might also synthesize fatty acid-based G3P bacterial-like lipids as they have some of the genes for fatty acid synthesis, G3P synthesis and acyltransferases for the esterification of the fatty acids to the G3P backbone. We further screened currently available Asgard archaea MAGs for the presence of the genes of the archaeal and bacterial lipid biosynthetic pathways (**Supplementary Table 20**). In contrast to the *Ca.* Lokiarchaeota MAG CR4, Asgard archaea MAGs of the *Candidatus* Heimdallarchaeota, *Candidatus* Odinarchaeota, and *Candidatus* Thorarchaeota do harbour a G1PDH coding gene homolog. In addition, several of the *Ca.* Heimdallarchaeota MAGs (**Supplementary Table 20**) also harbour a homolog of

the PlsY acyltransferase, and both the *Ca.* Heimdallarchaeota MAGS LC2 and LC3 (**Supplementary Table 20**) also have homologs of the bacterial PlsC acyltransferase as observed for the *Ca.* Lokiarchaeota CR4 MAG. The lack of G1PDH in *Ca.* Lokiarchaeota and potentially of G1P archaeal-like lipids in Asgard archaea is very appealing for supporting the eukaryogenesis scenario with eukaryotes originating from within the Asgard archaea (227). However, the presence of G1PDH in other Asgard archaea as seen here weakens this hypothesis. Also, the absence of a gene from incomplete MAGs reconstructed from environmental samples is not a warranty that the gene is actually missing from the genome. However, we also confirmed the absence of the G1PDH homolog (i.e. BLASTP with the G1PDH of '*Ca.* Odinarchaeota archaeon LCB_4' as query sequence) in the complete genome of an archaeon of the *Ca.* Lokiarchaeota that has been recently obtained from an enrichment culture (87). We therefore consider this further support that *Ca.* Lokiarchaeota either do not synthesize G1-based archaeal-like membrane lipids or they use a novel and unexpected alternative pathway to synthesize G1P as observed in the study of Caforio et al. (225) for the bacterium *E. coli*. Regardless of this genetic evidence, there is still no experimental evidence of the formation of a 'mixed membrane', neither in the Asgard archaea nor in the 'Marine Group II' Euryarchaeota.

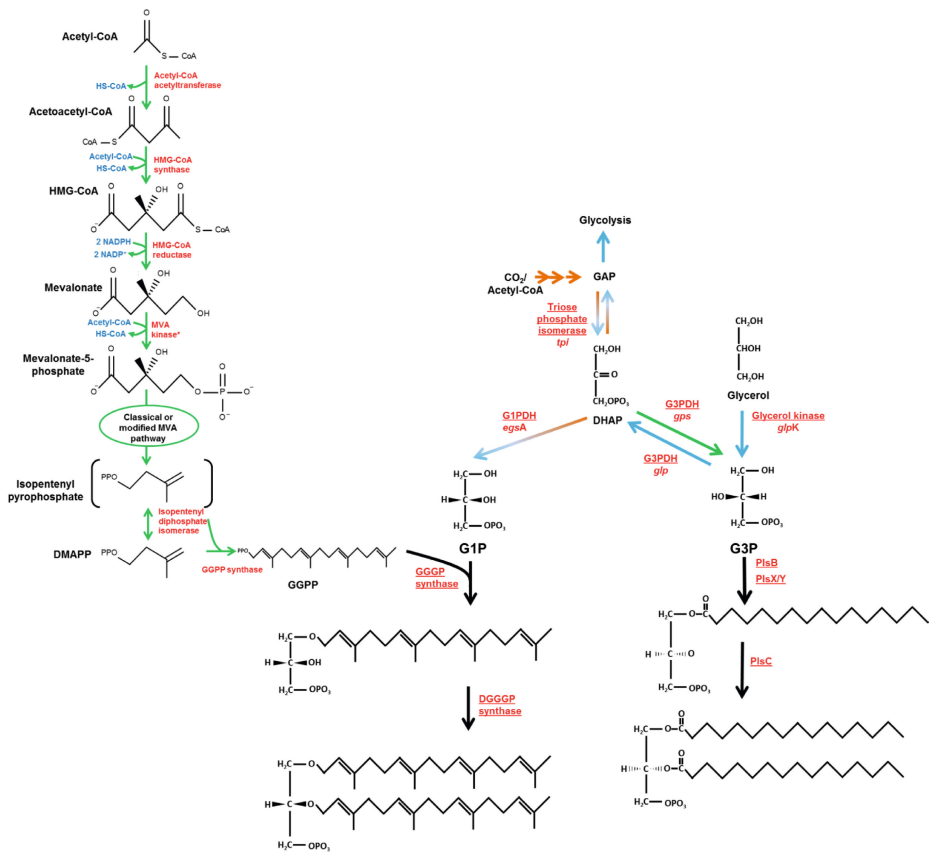
Supplementary Figures



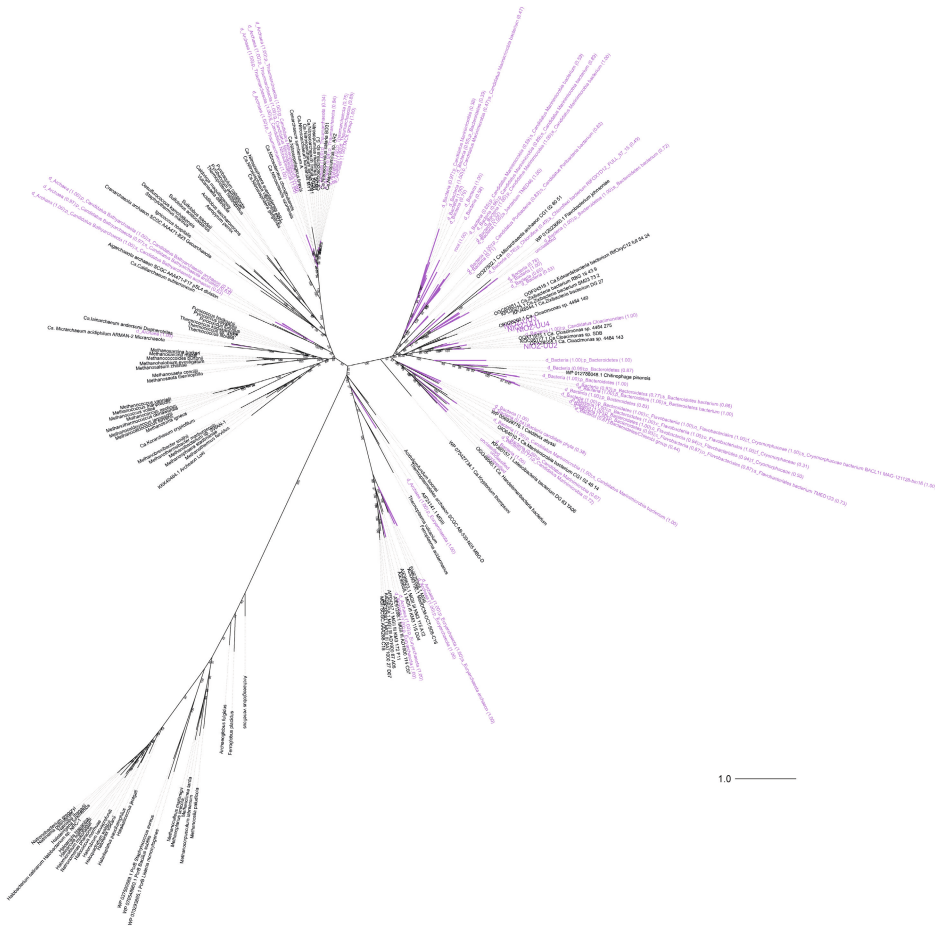
Supplementary Fig. 1 | Epifluorescence microscope images of *Ca. Cloacimonetes* cells hybridized by Catalyzed reporter deposition Fluorescence In Situ Hybridization (CARD-FISH) with either DAPI (a,d) or with the specific HRP-labelled probes Cloa1 for *Ca. Cloacimonetes* (b,c), EUB388 for general bacteria (b,e), and Arch915 for archaea (e,f) of Black Sea water collected at 2,000 m depth. Images were obtained with $\times 63$ magnification.



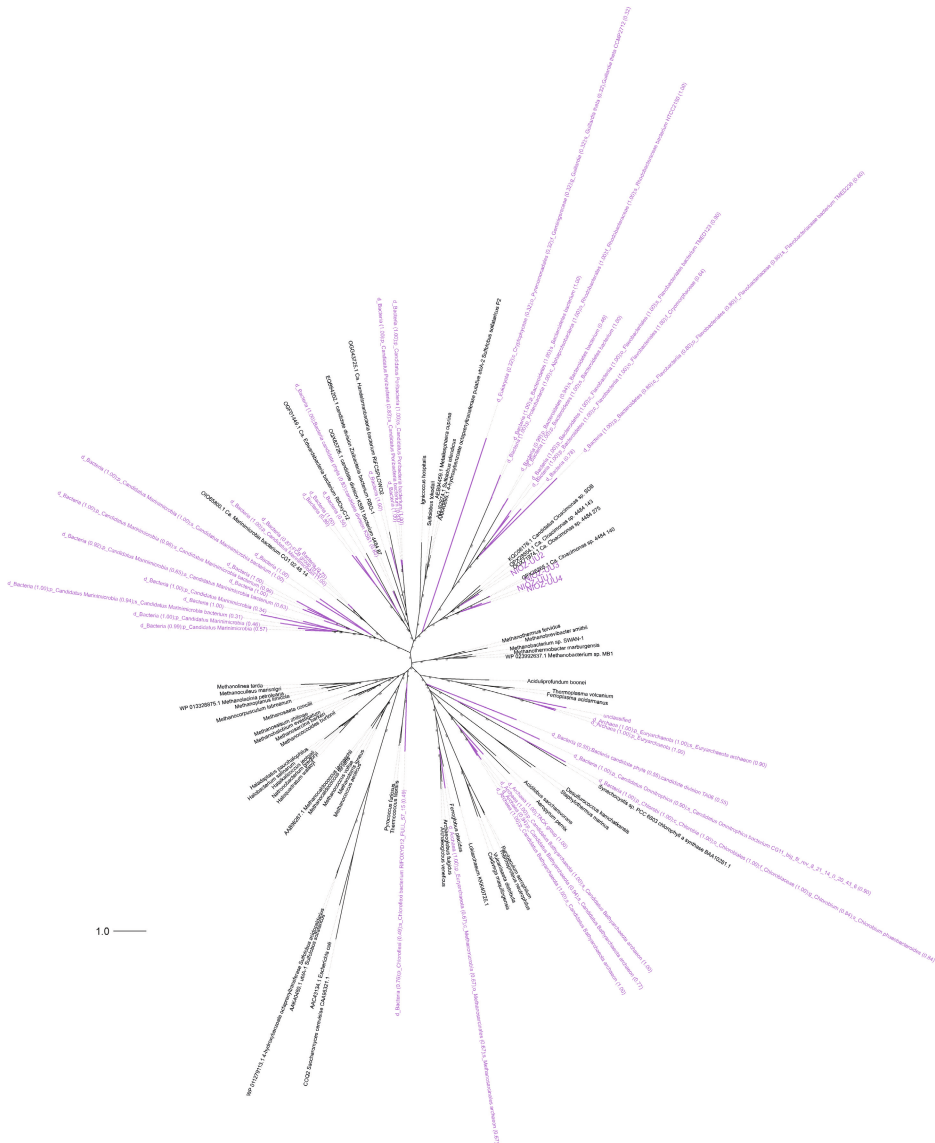
Supplementary Fig. 2 | Phylogenetic affiliation of the *Ca. Cloacimonetes* MAGs. a, Maximum likelihood phylogenetic tree based on 43 concatenated marker genes indicating the position of the *Ca. Cloacimonetes* MAGs with respect to other *Ca. Cloacimonetes* genome sequences available in databases. See ‘Materials and methods’ for details. b, 16S rRNA gene phylogenetic tree inferred by using the maximum likelihood method based on the generalised time reversible model, applying the neighbour-joining method with a discrete gamma distribution plus invariable sites. The analysis involved a total of 1,661 positions. Evolutionary analyses were conducted in MEGA6 (ref. 300). Correspondence of the sequences between the two trees is indicated with blue lines. 16S rRNA sequences from NIOZ-UU1, seq_BS2013_2000m_1, and seq_BS2016_1980m_1 are identical. For a description of the clustered sequences, see **Supplementary Table 4. Scale bars in **a** and **b** indicate mean number of amino acid and nucleotide substitutions per site, respectively.**



Supplementary Fig. 3 | Scheme of the archaeal membrane lipid biosynthetic pathway listing the main genes and enzymes involved.

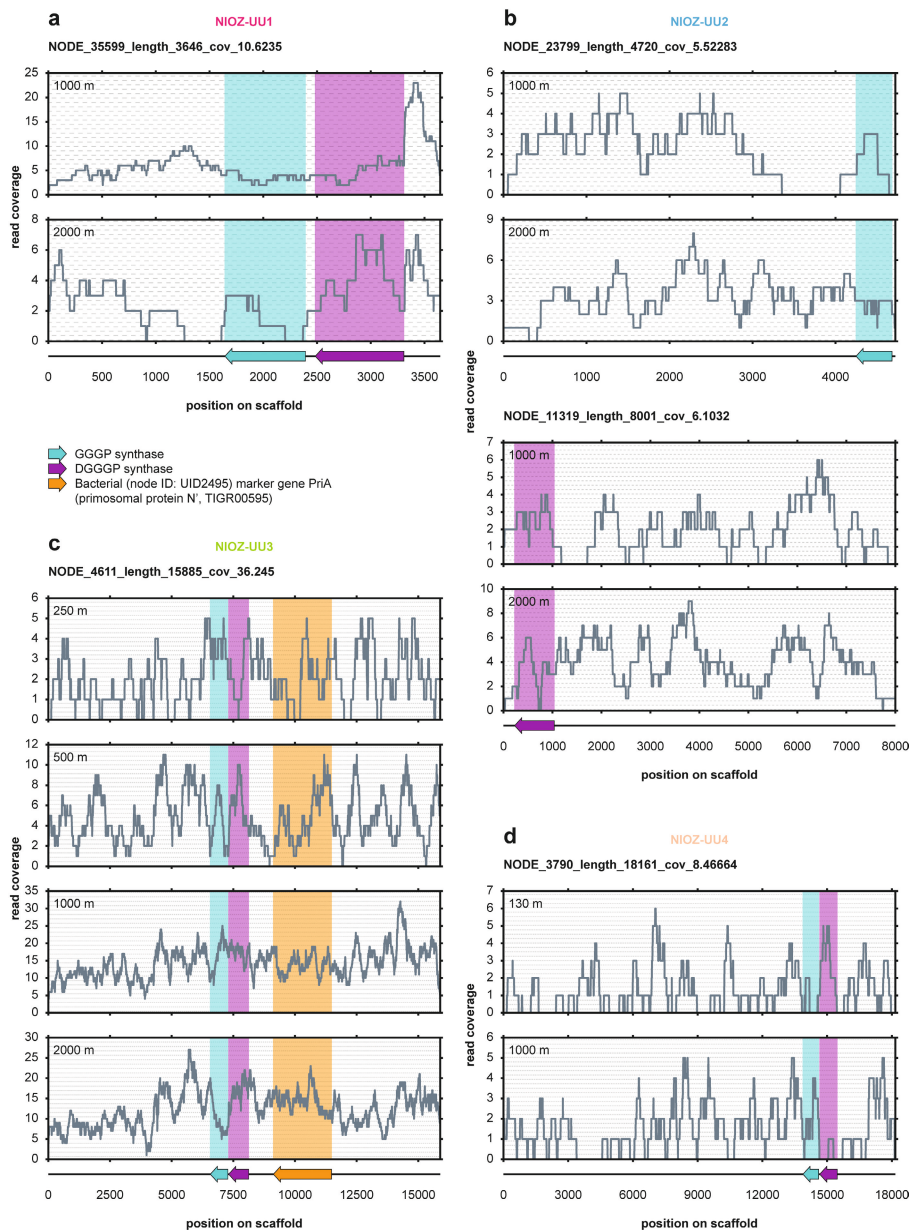


Supplementary Fig. 4 | Maximum likelihood tree of GGGP synthases found in the Black Sea assembly (purple) and reference sequences (black). *Ca.* Cloacimonetes reference sequences are based on a BLASTP search against nr. Branch support is based on 1,000 ultrafast bootstraps. Scale bar represents mean number of substitutions per site. Taxonomic classification of the sequences from the assembly are based on taxonomic classification of the scaffolds on which they were found with Contig Annotation Tool (CAT) (289). Only classifications at official taxonomic ranks are shown, unless intermediate ranks are informative (e.g. ‘candidate division TA06’). Numbers between brackets indicate the fraction of bit-score support for that classification (see ref. 289). Note that taxonomic classifications that have low bit-score support and/or are based on few ORFs (see **Supplementary Table 10** for full CAT results) are speculative.

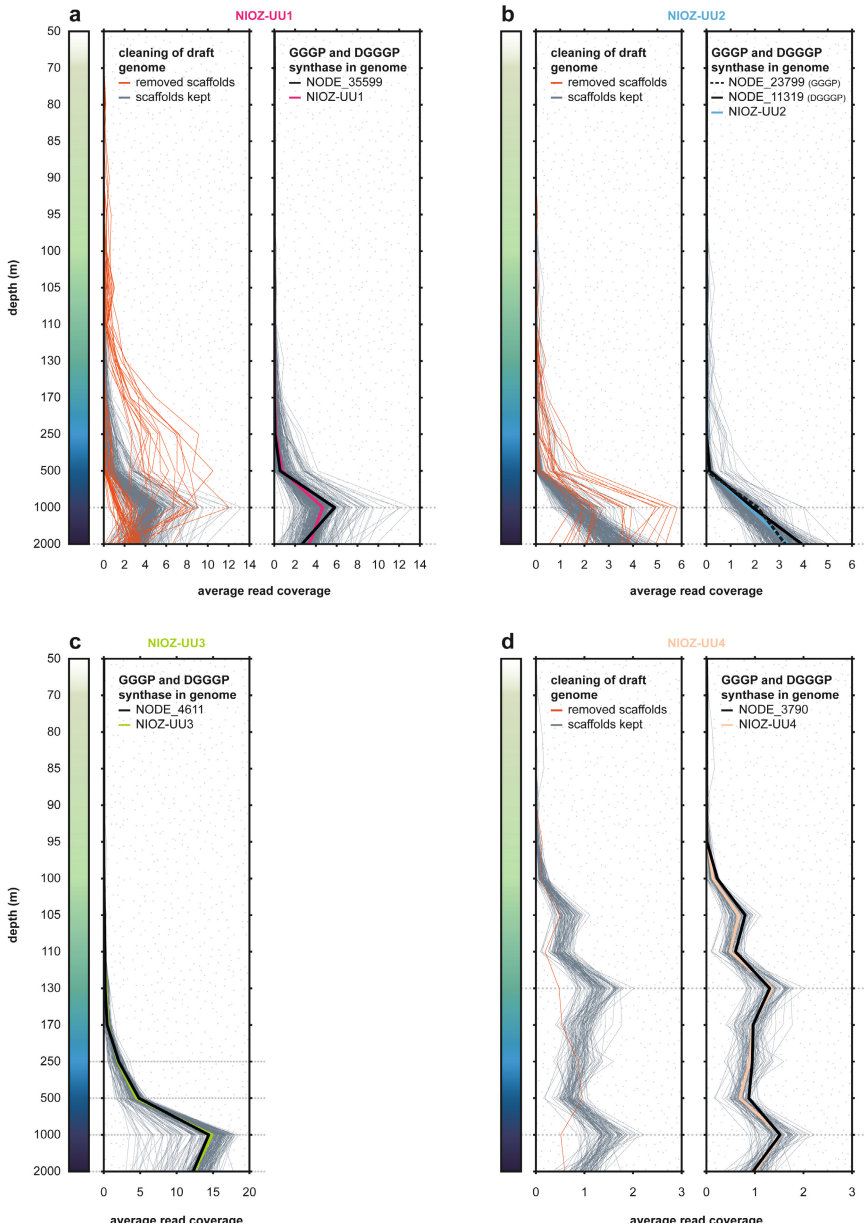


Supplementary Fig. 5 | Maximum likelihood tree of DGGGP synthases found in the Black Sea assembly (purple) and reference sequences (black). *Ca.* Cloacimonetes reference sequences are based on a BLASTP search against nr. Branch support is based on 1,000 ultrafast bootstraps. Scale bar represents mean number of substitutions per site. Taxonomic classification of the sequences from the assembly are based on taxonomic classification of the scaffolds on which they were found with Contig Annotation Tool (CAT) (289). Only classifications at official taxonomic ranks are shown, unless intermediate ranks are informative (e.g. ‘TACK group’ if there is no lower rank classification). Numbers between brackets indicate the fraction of bit-score support for that classification (see ref. 289). Note that taxonomic classifications that have low bit-score support and/or are based on few ORFs (see **Supplementary Table 10** for full CAT results) are speculative. The eukaryotic scaffold classification is tentative because it is ultimately based on a single ORF out of 30 that was classified as *Guillardia theta* (see **Supplementary Table 10**).

Chapter 2

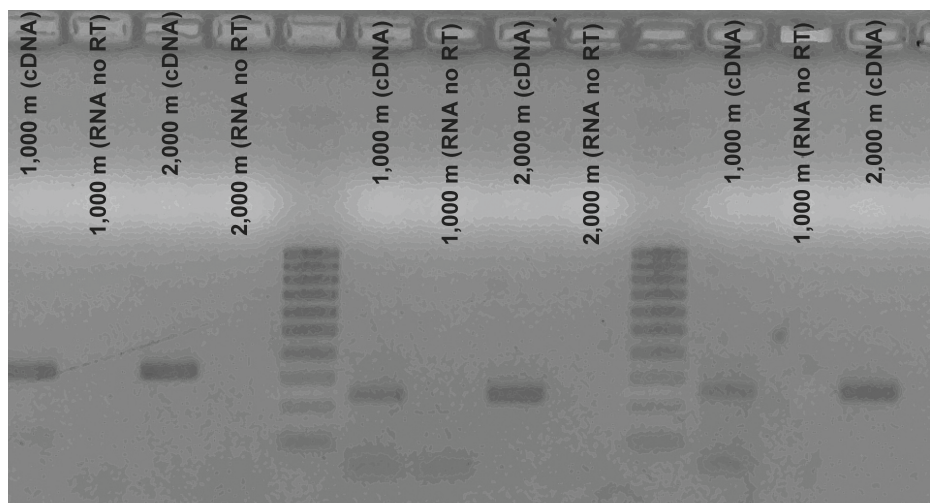


Supplementary Fig. 6 | Coverage across *Ca. Cloacimonetes* scaffolds containing GGGP and DGGGP synthase. a–d, Panels showing the position of the genes on the scaffolds, together with read coverage per nucleobase, as generated with SAMtools mpileup (278). Coverage is only shown for samples in which average read coverage is considerable (see horizontal dashed lines in **Supplementary Fig. 7**). **b,** In NIOZ–UU2 the genes are located on different scaffolds. **c,** The location of a bacterial marker gene on the same scaffold as the two archaeal homologs in NIOZ–UU3 is shown.

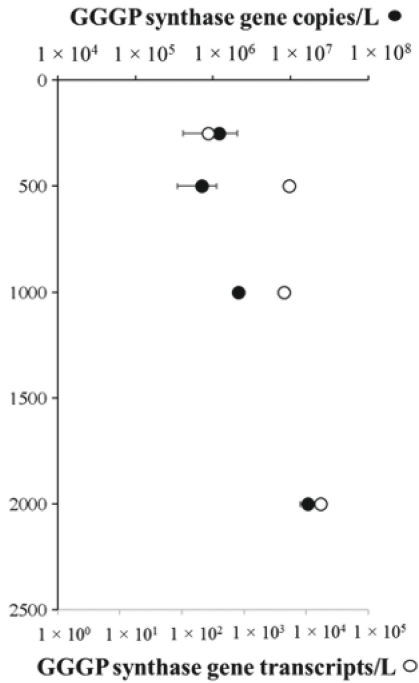


Supplementary Fig. 7 | Average read coverage (depth of mapped reads per nucleobase) for all the scaffolds in the four MAGs. Left panels in **a,b,d** show removed scaffolds due to manual cleaning. Since the coverage profile of NIOZ-UU3 looks clean (**c**), no scaffolds were removed. The right panel in **a-d** shows average read coverage of the scaffolds in the cleaned MAGs, with the scaffolds that contain GGGP and DGGGP synthase genes highlighted (2 in NIOZ-UU2). Coloured lines depict average read coverage of the cleaned MAGs, see ‘**Materials and methods**’ for details. Horizontal dashed lines indicate depths at which read coverage across the scaffolds containing GGGP and DGGGP synthase is plotted in **Supplementary Fig. 6**.

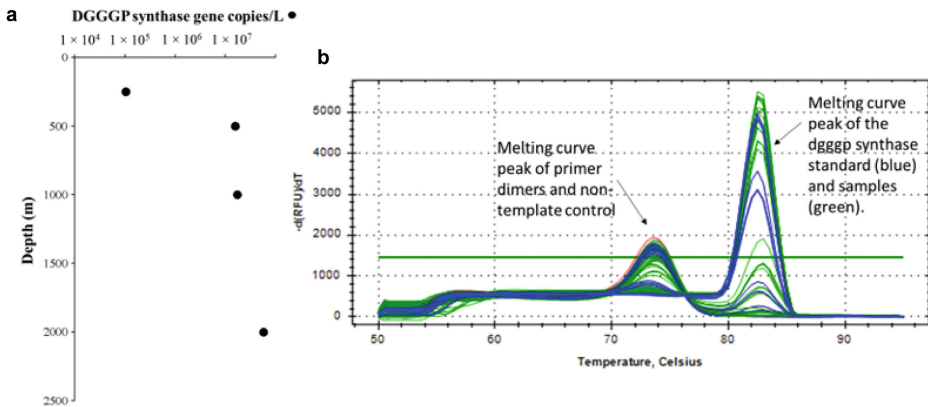
2



Supplementary Fig. 8 | Assessment of gene expression of the putative GGGP, DGGGP and polyprenyl transferases of NIOZ-UU3 in the 1,000 and 2,000 m depth Black Sea samples. Band indicates positive amplification with the primers listed in **Supplementary Table 17**. no RT: indicates negative control of RNA extract with absence of reverse transcriptase. cDNA: PCR using complementary DNA generated by reverse transcription. Amplicons were further sequenced for verification (data not shown).

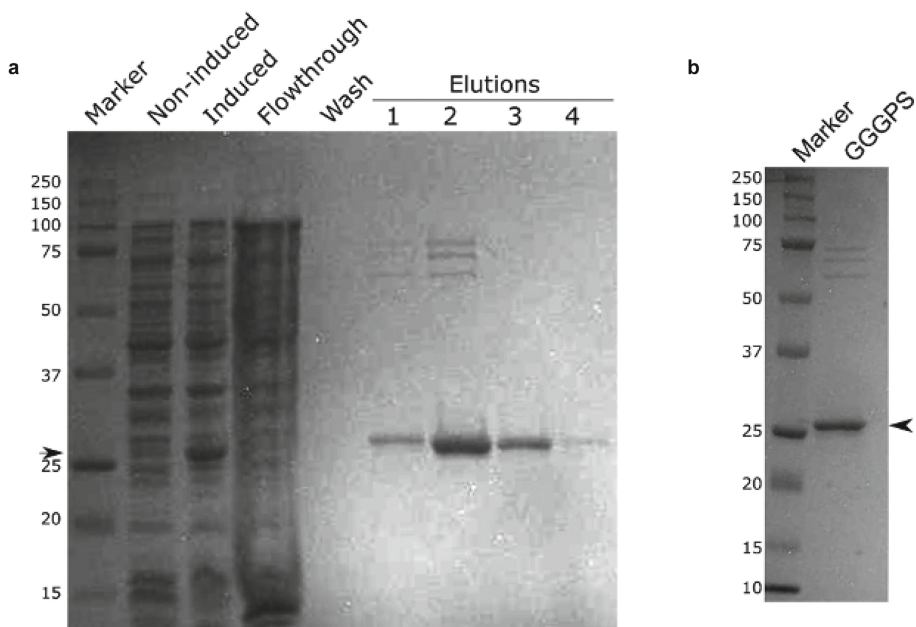


Supplementary Fig. 9 | *Ca. Cloacimonetes* GGGP synthase gene (black circles) and gene transcripts (white circles) copies per litre estimated by qPCR in the BS2013 campaign SPM samples.

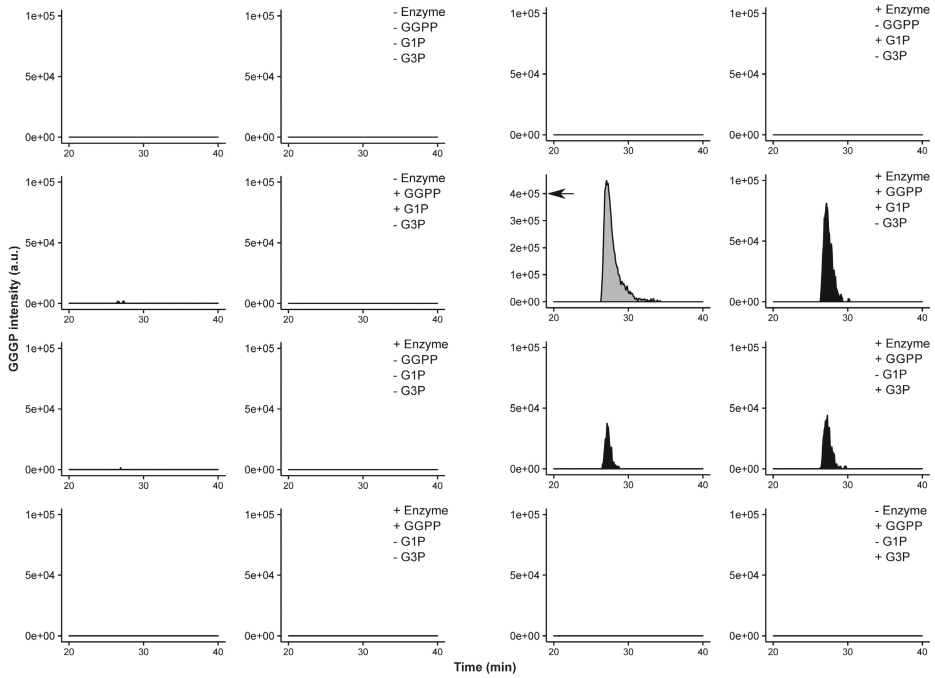


Supplementary Fig. 10 | *Ca. Cloacimonetes* DGGGP synthase coding gene and gene expression quantification. **a**, *Ca. Cloacimonetes* DGGGP synthase gene (black circles) copies per litre estimated by qPCR in the BS2013 campaign SPM samples. Note that the abundance is not accurate as explained in the text. **b**, Melting curve of the *Ca. Cloacimonetes* DGGGP synthase gene and gene transcript quantification, indicating that the melting curve behaviour of the standard (PCR amplicon of the *Ca. Cloacimonetes* DGGGP synthase gene fragment) is identical to that found in gene and gene transcript estimations in SPM extracts from 250 m downwards.

2

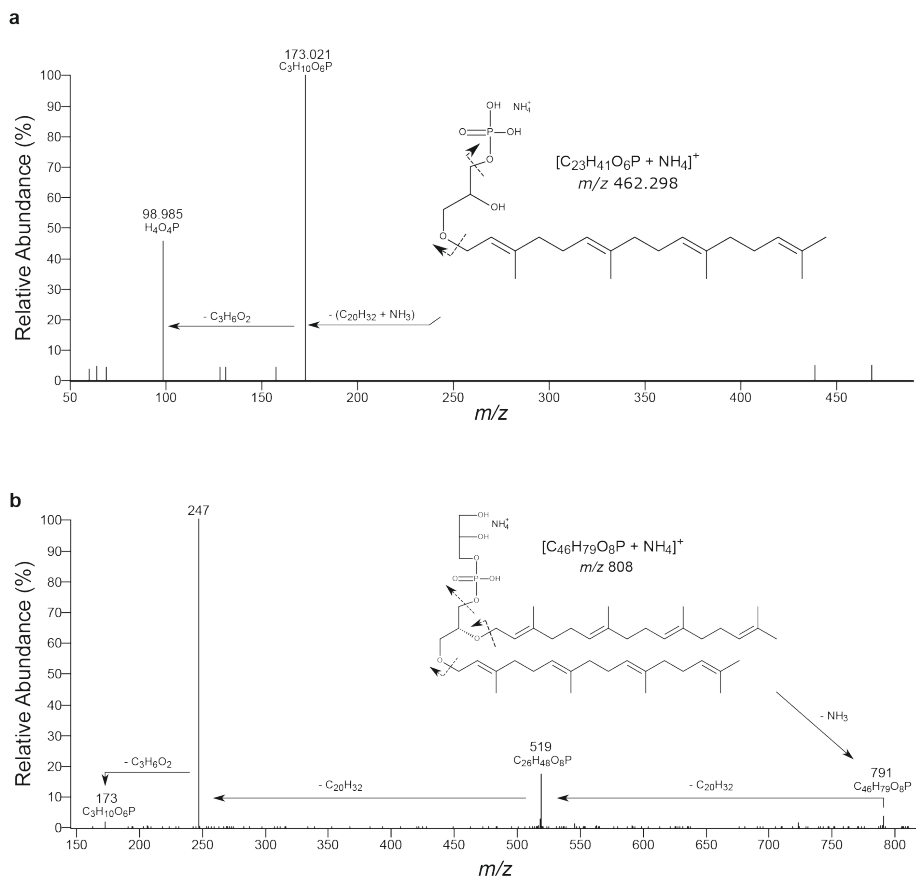


Supplementary Fig. 11 | SDS-PAGE gel images of the purified recombinant *Ca. Cloacimonetes* GGGP synthase. a, Gel indicating the protein extract in the different purification phases. Elution 2 was used in the enzymatic assay. **b,** Re-run of the protein extract obtained in Elution 2 with less material. Arrow points to predicted size (see text for further details).

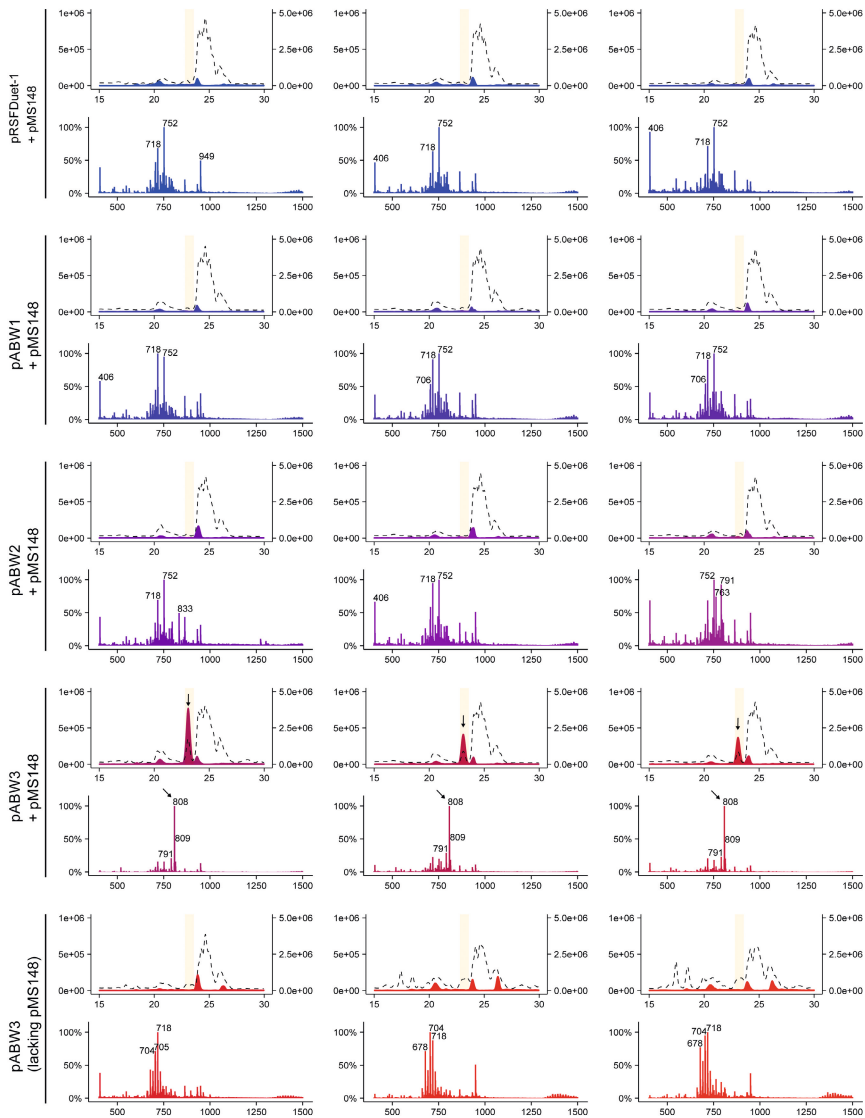


Supplementary Fig. 12 | GGGP production in vitro. Extracted ion chromatogram within 3 parts per million (ppm) mass accuracy of $[GGGP + NH_4]^+$ (m/z 462.298) showing the retention time in minutes vs. the GGGP intensity in arbitrary units, a.u. Results with inclusion or absence of purified enzyme, geranylgeranyl-diphosphate, glycerol-1-phosphate (G1P) or glycerol-3-phosphate (G3P) in reaction assay as indicated. Two replicate enzyme assays (performed different days) are indicated. Note that one sample (grey fill) is plotted with a 5-fold increased maximum y-scale value (arrow), to allow visual comparison to the other samples.

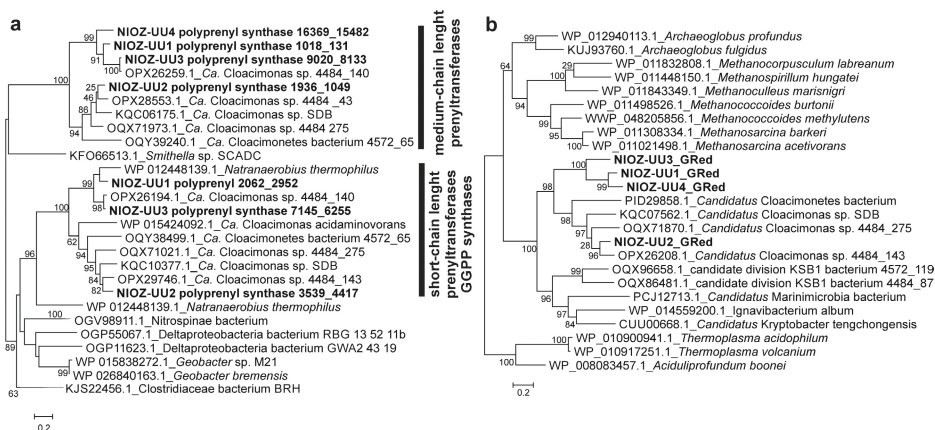
2



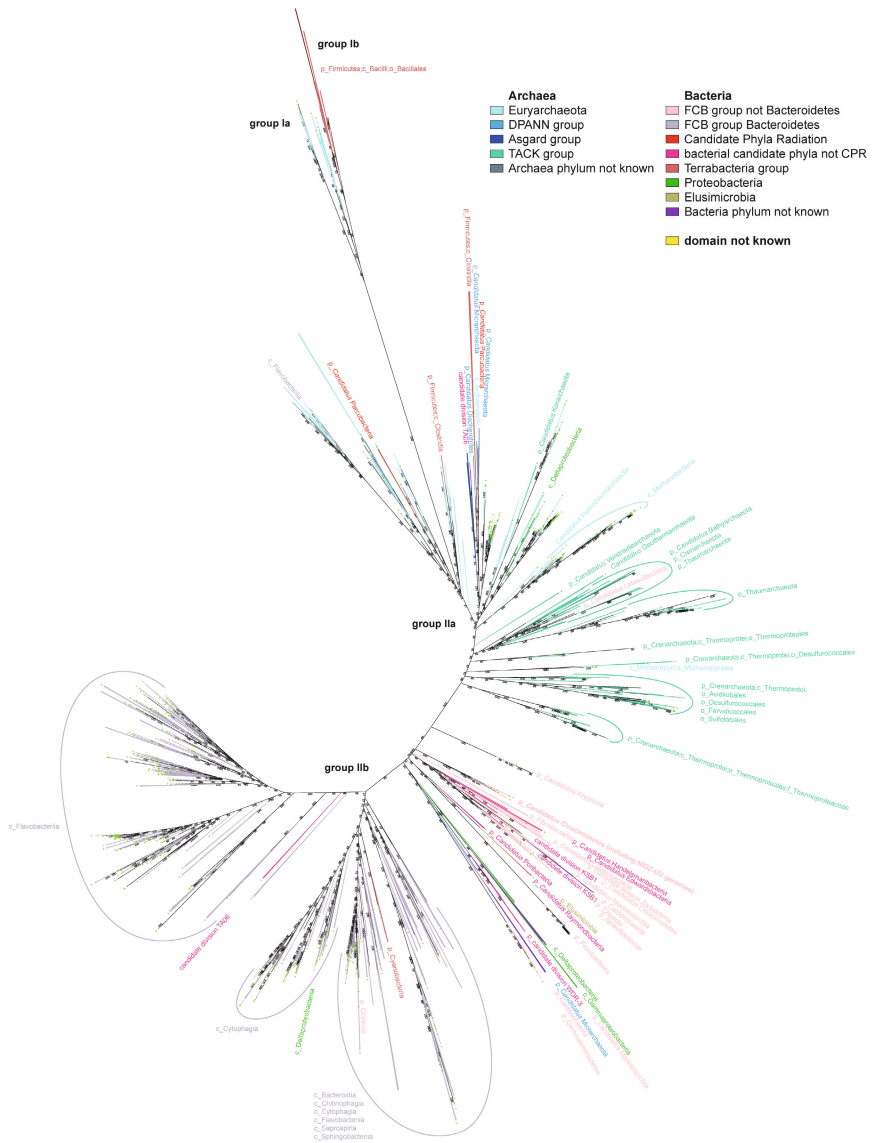
Supplementary Fig. 13 | Representative MS² fragmentation spectra of the GGGP produced in vitro and the phosphatidylglycerol-archaeol with 8 double bonds produced by the recombinant *E. coli* strain. **a**, Geranylgeranylgeranyl phosphate (GGGP) produced in vitro, obtained by HRMS using a quadrupole orbitrap hybrid MS. The MS² fragmentation spectrum showed a loss of the geranylgeranyl carbon chain ($-C_{20}H_{32} + NH_3$) generating a fragment at m/z 173.021 representing the phosphatidylglycerol ($C_3H_{10}O_6P$). Subsequent loss of the glycerol moiety ($-C_3H_6O_2$) results in a fragment at m/z 98.985 representing the remaining phosphatidic acid moiety (PO_4H_4). **b**, Phosphatidylglycerol-archaeol with 8 double bonds or unsaturations (PG-unsat(8)-archaeol) produced by recombinant '*E. coli* C43(DE3)' obtained by ion trap MS, as described in '(Supplementary) Materials and methods'. The fragmentation mass spectrum of PG-unsat(8)-archaeol shows an initial loss of NH_3 resulting in formation of the $[M + H]^+$ at m/z 791. Subsequent losses of the 2 geranylgeranyl moieties ($2 \times -C_{20}H_{32}$) results in the formation of fragments at m/z 519 and 247, with the latter representing the glycerol backbone with phosphatidylglycerol headgroup. Loss of a glycerol moiety results in the formation of the fragment at m/z 173 representing phosphatidylglycerol.



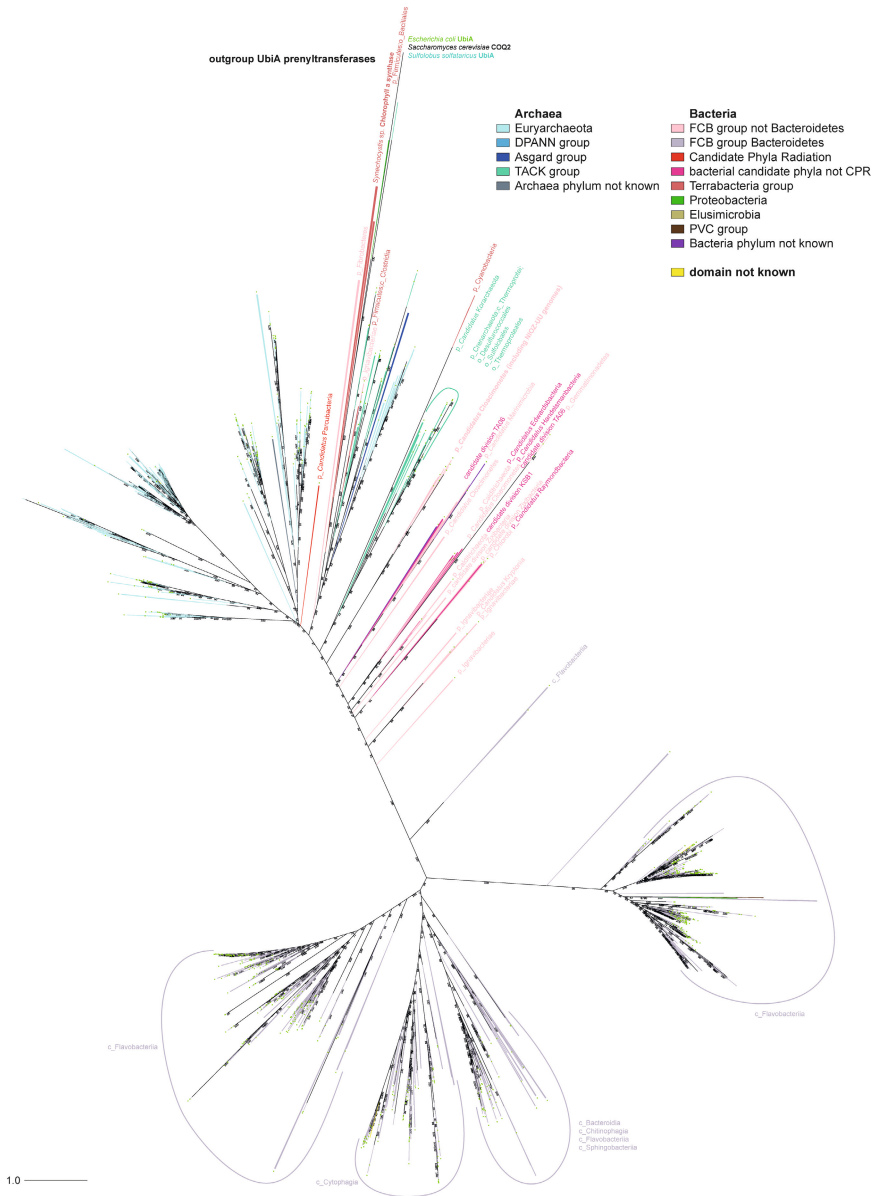
Supplementary Fig. 14 | PG-unsat(8)-archaeol production in recombinant *E. coli* C43(DE3). Upper panel of each subfigure shows the extracted ion chromatogram (± 0.5 mass units) of $[\text{PG-unsat(8)-archaeol} + \text{H}]^+$ (m/z 808; filled coloured area, left y-axis [$\text{PG-unsat(8)-archaeol} + \text{H}]^+$, in arbitrary units, a.u.), and the base peak intensity in a.u. (right y-axis, dotted line) vs. the retention time in minutes (x-axis). Orange box highlights the retention time where PG-unsat(8)-archaeol is detected (in the positive samples) and used for summed MS¹ analysis. The lower panel shows the average MS¹ spectrum for the region spanning the PG-unsat(8)-archaeol retention time (22.8 to 23.3 min) with the three most intense m/z indicated (x-axis, m/z ; y-axis, relative abundance in percentage). The plasmids harboured by *E. coli* C43(DE3) (see **Supplementary Table 15**) are indicated and arrows indicate formation of PG-unsat(8)-archaeol. The analysis results of three biological replicates are shown.



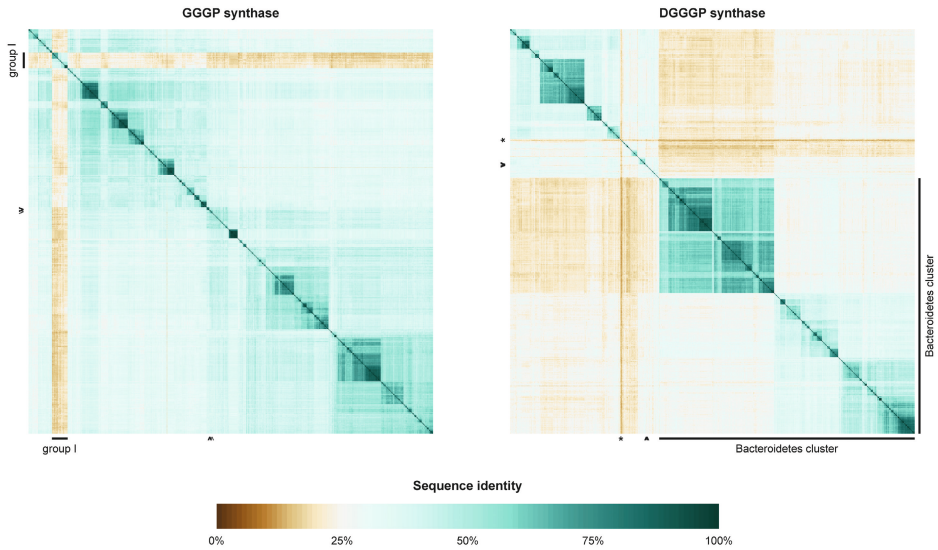
Supplementary Fig. 15 | Phylogeny of the polyprenyl transferases and digeranylgeranylglycerophospholipid reductases detected in the *Ca. Cloacimonetes* MAGs. a, The two putative polyprenyl transferases, and **b**, putative digeranylgeranylglycerophospholipid reductases and closest relatives. Scale bar represents mean number of substitutions per site. Branch support was calculated with the approximate likelihood ratio test (aLRT) and values $\geq 50\%$ are indicated on the branches.



Supplementary Fig. 16 | Consensus tree of GGGP synthases across the tree of life. The tree is based on 1,068 unique sequences representing 1,389 genomes from cultures and environmental samples. Bacterial clades are labelled as a certain group if a large fraction of sequences in the clade shows a consistent annotation. Interesting or aberrant placements are labelled as well, as are close archaeal sister groups to the bacterial part of the tree. Groups sensu Peterhoff et al. (253). Branch support is based on 1,000 ultrafast bootstraps. Scale bar represents mean number of substitutions per site. Archaeal or bacterial 'phylum not known': phylum is not known but the genome is annotated on a lower level, or sequence represents multiple groups. 'Domain not known': genomes for which no lineage was found on the PATRIC servers. Green dots: DGGGP synthase is also found in the same genome, or in at least one of the genomes if the branch represents multiple genomes.



Supplementary Fig. 17 | Consensus tree of DGGGP synthases across the tree of life. The tree is based on 1,258 unique sequences representing 1,385 genomes from cultures and environmental samples. Bacterial clades are labelled as a certain group if a large fraction of sequences in the clade shows a consistent annotation. Interesting or aberrant placements are labelled as well, as are close archaeal sister groups to the bacterial part of the tree. Branch support is based on 1,000 ultrafast bootstraps. Scale bar represents mean number of substitutions per site. Archaeal or bacterial ‘phylum not known’: phylum is not known but the genome is annotated on a lower level, or sequence represents multiple groups. ‘Domain not known’: genomes for which no lineage was found on the PATRIC servers. Green dots: GGGP synthase is also found in the same genome, or in at least one of the genomes if the branch represents multiple genomes.



Supplementary Fig. 18 | Pairwise sequence identity between protein sequences used in trees for GGGP synthase and DGGGP synthase. Sequence identity was calculated excluding regions where the beginning or end of one of the sequences consisted of gaps and excluding gap-gap alignments. Sequences are ordered according to placement in the tree (**Supplementary Figs. 16 and 17**), starting from the upper left branch. GGGP synthase sequences falling within Group I sensu Peterhoff et al. (253) are marked, as are DGGGP synthase sequences falling within the Bacteroidetes cluster. Asterisk indicates the position of outgroup UbiA prenyltransferases. Carets indicate the position of the enzymes from the four *Ca. Cloacimonetes* MAGs described in this study.

2

Supplementary Tables

Supplementary Tables 1–20 (captions below) are available from Zenodo at <https://doi.org/10.5281/zenodo.8090260>.

Supplementary Table 1 | Overview of the research cruises and the analyses performed with the samples collected as specified in ‘Materials and methods’.

Supplementary Table 2 | 16S rRNA amplicon sequencing counts. **a**, Heat map (red, higher; green, lower) of the percentage of 16S rRNA gene reads affiliated to the *Ca. Cloacimonetes* obtained by amplicon sequencing in the BS2013 (50 to 2,000 m) and the BS2016 campaign (1,000 to 1,980 m). 16S rRNA gene amplicon data generated with 454 GS FLX sequencing as specified in ‘**Materials and methods**’. **b**, Percentage of 16S rRNA gene reads affiliated to the *Ca. Cloacimonetes* and Archaea obtained by amplicon sequencing in the BS2013 campaign and sequenced with 454 GS FLX and Illumina MiSeq 3 × 200 bp (methods in table). Samples collected in station 4 BS2017 and in station 2 BS2018 and sequenced with Illumina MiSeq 2 × 200 bp are also specified. More details can be found in ‘**Supplementary Materials and methods, Results and discussion**’.

Supplementary Table 3 | Extended CheckM output of the four *Ca. Cloacimonetes* MAGs after cleaning.

Supplementary Table 4 | Information about FCB group genomes included in Fig. 1c,d, Supplementary Fig. 2, and Supplementary Table 13. The genome GCA_000402135.1 annotated as *Candidatus* Marinimicrobia contains marker gene sequences very similar to GCA_000384735.1, which is annotated as *Candidatus* Latescibacteria. Together they cluster with *Ca. Latescibacteria* genomes, albeit with low bootstrap support. The two genomes are enigmatic, we believe the *Ca. Marinimicrobia* genome is wrongly labelled, and we include it with *Ca. Latescibacteria* in Fig. 1c. na: not available.

Supplementary Table 5 | Functional annotation of NIOZ–UU1 based on the Rapid Annotation using Subsystem Technology (RAST) pipeline v2.0.

Supplementary Table 6 | Functional annotation of NIOZ–UU2 based on the Rapid Annotation using Subsystem Technology (RAST) pipeline v2.0.

Supplementary Table 7 | Functional annotation of NIOZ–UU3 based on the Rapid Annotation using Subsystem Technology (RAST) pipeline v2.0.

Supplementary Table 8 | Functional annotation of NIOZ–UU4 based on the Rapid Annotation using Subsystem Technology (RAST) pipeline v2.0.

Supplementary Table 9 | Location and predicted function of genes related to archaeal lipid biosynthesis in the MAGs NIOZ–UU1 to 4 based on the Rapid Annotation using Subsystem Technology (RAST) pipeline v2.0. Note that the putative GGGP synthase is listed here as ‘homolog of geranylgeranylgeranyl glyceryl phosphate synthase’, and DGGGP synthase as ‘(S)-2,3-di-O-geranylgeranylgeranyl glyceryl phosphate synthase’. Putative GGGP and DGGGP synthases are also concatenated in NIOZ–UU2 as the putative GGGP synthase bridges two scaffolds as specified in ‘Supplementary Materials and methods, Results and discussion’.

Supplementary Table 10 | Scaffolds with significant hits to *Ca. Cloacimonetes* GGGP synthase or DGGGP synthase in the Black Sea assembly via TBLASTN. Predicted proteins are the Prodigal predictions that were generated by CAT from the same region as the hit. Note that trees (Supplementary Figs. 4 and 5) are based on the aligned part of the subject sequence and not on the full protein as predicted by Prodigal. Taxonomic classification is based on CAT. A single scaffold (NODE_19451_length_5423_cov_10.7351) has multiple classifications. Taxonomic classifications that have low bit-score support and/or are based on few ORFs are speculative. The eukaryotic classification of NODE_3056_length_20971_cov_7.58568 is tentative because it is based on 6/30 ORFs, of which only a single ORF was classified as *Guillardia theta* (data not shown).

Supplementary Table 11 | Screening of GGGP synthase and DGGGP synthase genes in 110,421 genomes deposited in the PATRIC database, based on a BLASTP search of the genes from the 4 *Ca. Cloacimonetes* MAGs. X indicates absence, numbers indicate the copy number of the gene in the genome if it is found. ‘not known’ in the lineage indicates that the genome has no annotation on the specific level, ‘no lineage found’ denotes genomes that were uploaded on the PATRIC server but were not present in the supplied lineage file.

Supplementary Table 12 | Summary of Supplementary Table 11, indicating how many genomes of the respective phylum contain GGGP synthase, DGGGP synthase, and both. ‘not known’ in the lineage indicates that the genome has no annotation on the phylum level, ‘no lineage found’ denotes genomes that were uploaded on the PATRIC server but were not present in the supplied lineage file. ‘Candidate division TA06’ and ‘candidate division KSBI’ don’t have a rank assigned, thus they fall within the ‘Bacteria;p_not known’ category. They are added below the tables.

Supplementary Table 13 | Screening of GGGP synthase and DGGGP synthase and co-localization of the two genes in the DNA sequences of FCB group genomes downloaded from GenBank and RefSeq (Supplementary Table 4), based on a TBLASTN search of the genes from the 4 *Ca. Cloacimonetes* MAGs. X indicates absence, GGGPs and DGGGPs indicate presence of the respective gene.

Supplementary Table 14 | Primers used in this study.

Supplementary Table 15 | Strains and plasmids used in this study.

Chapter 2

Supplementary Table 16 | Overview of the presence of genes of the glycerophospholipid pathway (find functions, pathways by KEGG orthology (KO) terms in the Integrated Microbial Genomes (IMG) system) in the four MAGs. Numbers in the table indicate lack (0) or specific number of copies. Enzymes discussed in the text are highlighted in grey.

Supplementary Table 17 | Overview of the presence of genes of the terpenoid backbone biosynthetic pathway (find functions, pathways by KEGG orthology (KO) terms in the Integrated Microbial Genomes (IMG) system) in the four MAGs. Numbers in the table indicate lack (0) or specific number of copies. Enzymes of the MEP/DOXP pathway and those discussed in the text are highlighted in grey.

Supplementary Table 18 | Overview of the presence of genes of the bacterial fatty acid pathway (find functions, pathways by KEGG orthology (KO) terms in the Integrated Microbial Genomes (IMG) system) in the four MAGs. Numbers in the table indicate lack (0) or specific number of copies. Enzymes discussed in the text are highlighted in grey.

Supplementary Table 19 | Determination of lipid abundances (in nanograms per litre of filtered seawater) in suspended particulate matter collected from 500 to 2,000 m depth in station 4 during the BS2017 cruise (Supplementary Table 1), as core lipids (CLs) detected in the Bligh and Dyer lipid extracts (BDE) as well as the CLs + IPL-derived CLs detected in the BDE extract after acid hydrolysis (i.e. H⁺ BDE), and those CLs IPL derived (i.e. IPL derived).

Supplementary Table 20 | Overview of the presence of genes of the glycerophospholipid pathway (find functions, pathways by KEGG orthology (KO) terms in the Integrated Microbial Genomes (IMG) system) of Asgard archaea MAGs currently available (September 2019). Numbers in the table indicate lack (0) or specific number of copies. Enzymes discussed in the text are highlighted in grey.

Supplementary Notebook 1

```
[1]: #/usr/bin/env python3

import matplotlib.pyplot as plt
import numpy as np
```

```
[2]: from IPython.display import set_matplotlib_formats
set_matplotlib_formats('png', 'pdf')
```

This notebook describes the data and calculations that demonstrate that there is a striking offset between the amount of archaeal membrane lipids (intact polar lipids or IPLs) observed in the Black Sea water column and the amount of IPLs that the archaeal population can theoretically produce in the most ideal situation. We take uncertainty in measurements for both membrane lipid production and size estimates into account by taking the most extreme cases reported in literature and plotting the expected range of IPLs abundances given observed archaeal abundances. The most ideal situation, where all Archaea in the water column are at maximum known size and produce the maximum reported number of membrane lipids (i.e. have just come out of growth phase), cannot explain the observed amount of IPLs at deeper depths.

1. Functions

The following formula is used to throughout this notebook to calculate cell surface area, A , based on its radius r and length l :

$$A = 2\pi rh + 4\pi r^2$$

with

$$h = l - 2r$$

Thus, cells are modelled as spheres if $l = 2r$ (coccus shaped), and as hemisphere-capped cylinders if $l > 2r$ (rod shaped).

Chapter 2

```
[3]: def calc_surface_area(r, l):  
      h = l - 2 * r  
  
      return 2 * np.pi * r * h + 4 * np.pi * r ** 2
```

2. Data

In this section we define and show the data on which subsequent calculations are based.

2.1. Measured data

All measured data are a list of length 4, where each element is a different depth in the water column.

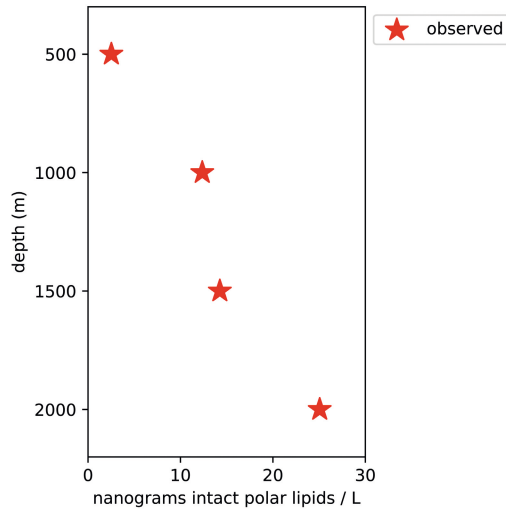
```
[4]: depths = [500, 1000, 1500, 2000] # In m.
```

2.1.1. Archaeal lipid data

The observed archaeal lipid data is from station 4 during the BS2017 campaign and is described in **Supplementary Table 18**. We plot it to show what it looks like.

```
[5]: observed_IPLs = [2.52431557301726,  
                    12.3592656904217,  
                    14.2605958190464,  
                    25.0652634330246] # In ng / L.  
  
# Plot the data.  
fig = plt.figure(figsize=(3, 5))  
ax = plt.subplot2grid((1, 1), (0, 0))  
  
ax.scatter(observed_IPLs,  
          depths,  
          marker='*',  
          color='red',  
          s=200,  
          label='observed')  
  
# Styling.  
ax.set_xlim([0, 30])  
ax.set_xlabel('nanograms intact polar lipids / L')  
ax.set_ylim([2200, 3000])  
ax.set_yticks(depths)  
ax.set_ylabel('depth (m)')  
ax.legend(loc='upper left', bbox_to_anchor=(1, 1))  
plt.suptitle('Figure 1: observed archaeal membrane lipids')  
  
plt.show()  
plt.close()
```

Fig. 1: observed archaeal membrane lipids



2.1.2. 16S rRNA gene read data

The 16S rRNA gene read data is from station 4 during the BS2017 campaign and is described in **Supplementary Table 2b**.

```
[6]: total_16S = [33483253.5212389,
                 35714158.946356,
                 43608346.1585536,
                 26792197.1985571] # In copies / L based on qPCR.

archaeal_taxa = ['Thermoplasmatales',
                 'ANME-1b',
                 'MCG+C3', # Part of the phylum Candidatus Bathyarchaeota.
                 'DHVE-6', # Part of the DPANN superphylum.
                 'Archaea, others']

# Fractions of total reads that are attributable to Archaea or
# Candidatus Cloacimonetes.
```

Chapter 2

```
fractions = dict()
fractions['Candidatus Cloacimonetes'] = [0.0742909208685631,
                                           0.147748016535324,
                                           0.199093934608142,
                                           0.188686561035215]
fractions['Thermoplasmatales'] = [0.00221381463039215,
                                   0.00104816453435197,
                                   0.00134980791195053,
                                   0.00262319733310745]
fractions['ANME-1b'] = [0.000145926687856,
                        0.00465589190165,
                        0.0126674280968,
                        0.0157199859607]
fractions['MCG+C3'] = [0.00554165495100603,
                       0.0076276542705795,
                       0.0098257272295019,
                       0.00787755533844039]
fractions['DHVE-6'] = [0.00358054263373126,
                       0.00348603036144377,
                       0.00469427124034476,
                       0.0124533684865375]
fractions['Archaea, others'] = [0.00391510625953947,
                                0.00369801869423497,
                                0.00554131669117553,
                                0.00657199837971533]
fractions['Total Archaea'] = [sum(fractions[taxon][i] for
                                  taxon in archaeal_taxa) for
                              i, depth in enumerate(depths)]
```

We convert 16S rRNA gene read count to number of cells per liter. The following formula is used for a certain taxonomic group at a given depth, with C in cells L^{-1} :

$$C_{taxon} = \frac{f_{taxon}T}{n_{taxon}}$$

where f is the fraction of total reads attributable to the taxonomic group at that depth, T the total 16S count in copies L^{-1} at that depth, and n the number of 16S copies per genome. We assume one 16S copy per genome for Archaea and two 16S copies per genome for *Candidatus Cloacimonetes* as “*Candidatus Cloacimonas acidaminovorans*” str. Evry’ has two. We plot the predicted cell counts to show what it looks like.

```
[7]: SSU_copy_number = {taxon: 1 for
      taxon in archaeal_taxa + ['Total Archaea']}
SSU_copy_number['Candidatus Cloacimonetes'] = 2

cells_per_L = dict()
for taxon in fractions:
    cells_per_L[taxon] = [fractions[taxon][i] * total_16S[i] /
                        SSU_copy_number[taxon] for
                        i, depth in enumerate(depths)]

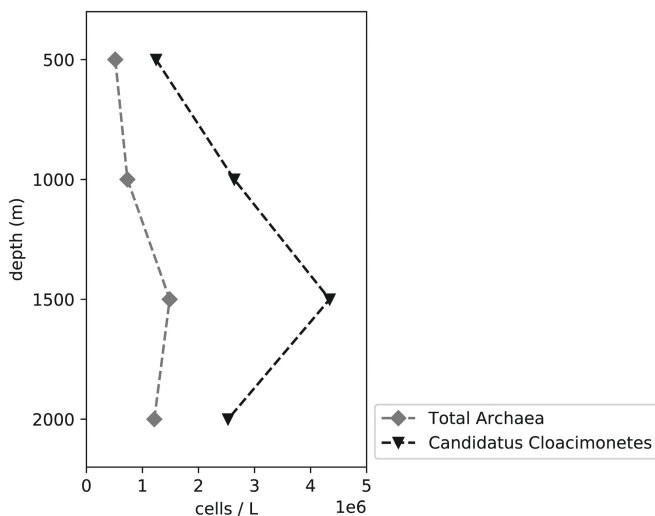
# Plot the data. Only plot total archaeal abundances.
fig = plt.figure(figsize=(3, 5))
ax = plt.subplot2grid((1, 1), (0, 0))

ax.plot(cells_per_L['Total Archaea'],
        depths,
        marker='D',
        color='grey',
        linestyle='--',
        label='Total Archaea')
ax.plot(cells_per_L['Candidatus Cloacimonetes'],
        depths,
        marker='v',
        color='black',
        linestyle='--',
        label='Candidatus Cloacimonetes')

# Styling.
ax.set_xlim([0, 5000000])
ax.ticklabel_format(axis='x', style='sci', scilimits=(-2, 2))
ax.set_xlabel('cells / L')
ax.set_ylim([2200, 300])
ax.set_yticks(depths)
ax.set_ylabel('depth (m)')
ax.legend(loc='lower left', bbox_to_anchor=(1, 0))
plt.suptitle('Figure 2: observed cell counts')

plt.show()
plt.close()
```

Fig. 2: observed cell counts



The main question this notebook addresses is: do the Archaea in the water column (**Fig. 2**) support the observed amount of archaeal IPLs (**Fig. 1**)?

2.2. Literature estimates for membrane lipid abundances in archaeal cells

There are three estimates for membrane lipid abundances in archaeal cells:

1. Sinninghe Damsté et al. (317) estimate the number of Crenarchaeota cells based on measured GDGT lipids. Cells are assumed rod-shaped with 0.8 μm length and a radius of 0.25 μm . Membrane lipid abundance is estimated 1.0 fg per cell.
2. Schouten et al. (318) estimate the expected amount of IPL-GDGT based on Thaumarchaeota abundance. Cells are assumed rod-shaped with 0.5 μm length and a radius of 0.075 μm . Membrane lipid abundance is estimated 0.25 fg per cell.
3. Elling et al. (316) report lipid production by marine ammonia-oxidizing Archaea in a growth experiment. Cells are assumed rod-shaped with 0.5-0.9 μm length and a radius of 0.1 μm . Membrane lipid abundance estimates range from 0.86 fg per cell for small cells in the early growth phase to 1.85 fg per cell for large cells right after growth phase. This represents an increased production of ~20% per cell surface area right after growth. Stationary phase production is lower again at 0.92 fg per cell.

We construct the dictionary 'estimates' that contains per study both the estimated membrane lipids per cell and their estimated surfaces. Multiple estimates indicate a range. We print the dictionary to show what it looks like.

```
[8]: estimates = dict()
      (estimates['Sinninghe Damsté'], estimates['Schouten'],
       estimates['Elling']) = dict(), dict(), dict()

      # Amount of lipids per cell are given in femtogram, radius and length in
      # micrometre.
      estimates['Sinninghe Damsté']['lipids per cell'] = [1.0]
      estimates['Sinninghe Damsté']['surface area'] = [calc_surface_area(0.25,
                                                                    0.8)]

      estimates['Schouten']['lipids per cell'] = [0.25]
      estimates['Schouten']['surface area'] = [calc_surface_area(0.075, 0.5)]

      estimates['Elling']['lipids per cell'] = [0.86, 1.85]
      estimates['Elling']['surface area'] = [calc_surface_area(0.1, 1) for
                                             1 in [0.5, 0.9]]

      # Print. Limit precision for the display of floats to 5 decimal points.
      for i, study in enumerate(estimates):
          print(study)
          print('membrane lipids per cell (fg):\t',
                [float(f'{v:.5f}') for v in
                 estimates[study]['lipids per cell']])
          print('cell surface area (micrometre^2):\t',
                [float(f'{v:.5f}') for v in
                 estimates[study]['surface area']])
          if i < 2:
              print()
```

```
Sinninghe Damsté
membrane lipids per cell (fg):  [1.0]
cell surface area (micrometre^2):  [1.256664]

Schouten
membrane lipids per cell (fg):  [0.25]
cell surface area (micrometre^2):  [0.235662]

Elling
membrane lipids per cell (fg):  [0.86, 1.85]
cell surface area (micrometre^2):  [0.31416, 0.56549]
```

2.3. Estimates of cell surface area of different archaeal taxa based on reported size ranges in literature

Size estimates for the four encountered archaeal taxa are:

- Thermoplasmatales: rod-shaped cells with a length of 0.5-3.0 μm and radius of 0.1-0.25 μm (319).
- ANME-1: cells within aggregates with 1.2 μm length and a radius of 0.15-0.2 μm (320).
- *Candidatus* Bathyarchaeota: spherical cells with a radius of 0.2-0.25 μm (321).
- DPANN: spherical cells with a radius of 0.2-0.25 μm , based on the *Nanoarchaeum equitans* (322) and ARMAN archaea (323).

Chapter 2

We calculate the upper and lower bound of cell surface area by taking the smallest and largest combination of reported l and r . We construct the dictionary 'surface_area_range' that contains the range of surface areas for different archaeal taxa, and print it to show what it looks like.

```
[9]: surface_area_range = dict()
      # Radius and length are given in micrometre.
      surface_area_range['Thermoplasmatales'] = [calc_surface_area(r, l) for
                                                  r in [0.1, 0.25] for
                                                  l in [0.5, 3.0]]
      surface_area_range['ANME-1'] = [calc_surface_area(r, 1.2) for
                                       r in [0.15, 0.2]]
      surface_area_range['Bathyarchaeota'] = [calc_surface_area(r, 2 * r) for
                                                r in [0.2, 0.25]]
      surface_area_range['DPANN'] = [calc_surface_area(r, 2 * r) for
                                      r in [0.2, 0.25]]

      # Print. Limit precision for the display of floats to 5 decimal points.
      print('range of cell surface area (micrometre^2)\n')
      for i, taxon in enumerate(sorted(surface_area_range)):
          print('{0}:\t{1}'.format(taxon,
                                   [float(f'{v:.5f}') for v in
                                    surface_area_range[taxon]]))
```

```
range of cell surface area (micrometre^2)

ANME-1: [1.13097, 1.50796]
Bathyarchaeota: [0.50265, 0.7854]
DPANN: [0.50265, 0.7854]
Thermoplasmatales: [0.31416, 1.88496, 0.7854, 4.71239]
```

3. Calculations

Next, we calculate the theoretical range of membrane lipids that the Archaea in the water column can produce.

3.1. Predictions of membrane lipids per cell for different archaeal taxa

We predict how many lipids per cell the different archaeal taxa produce, based on membrane lipid abundance estimates by Sinninghe Damsté et al., Schouten et al., and Elling et al., and their membrane surface area relative to the surface area of the organisms investigated in the respective studies. We estimate the number of lipids per cell, L , with the formula:

$$L_{\text{taxon}} = \frac{A_{\text{taxon}} L_{\text{study}}}{A_{\text{study}}}$$

where L_{study} and A_{study} are the membrane lipid estimate and surface area of the organisms in that particular study, respectively. We calculate L_{taxon} for the upper and lower surface area bounds of that taxon, and for the upper and lower bounds of cell surface area and membrane lipid abundance estimates reported in the particular study. The upper and lower bounds of all these combinations are reported. For the ‘other Archaea’, we assume a cell membrane lipid abundance between 0.25 and 5 femtogram per cell. We print the dictionary ‘lipids_per_cell’ to show what it looks like.

```
[10]: def predict_lipids_per_cell(surface_areas, estimates, study):
        """surface areas = list of cell surface areas.
           estimates = the estimates dictionary.
           study = either Sinnighe Damsté, Schouten, or Elling.
           """
        lipids = list()

        for surface_area in surface_areas:
            lipids.append(surface_area *
                          min(estimates[study]['lipids per cell']) /
                          min(estimates[study]['surface area']))
            lipids.append(surface_area *
                          max(estimates[study]['lipids per cell']) /
                          max(estimates[study]['surface area']))

        return (min(lipids), max(lipids))
```

```
[11]: # Amount of lipids per cell are given in femtogram.
lipids_per_cell = dict()
for study in estimates:
    lipids_per_cell[study] = dict()
    lipids_per_cell[study]['ANME-1b'] = predict_lipids_per_cell(
        surface_area_range['ANME-1'],
        estimates,
        study)
    lipids_per_cell[study]['DHVE-6'] = predict_lipids_per_cell(
        surface_area_range['DPANN'],
        estimates,
        study)
    lipids_per_cell[study]['MCG+C3'] = predict_lipids_per_cell(
        surface_area_range['Bathyarchaeota'],
```

```

    estimates,
    study)
lipids_per_cell[study]['Thermoplasmatales'] = predict_lipids_per_cell(
    surface_area_range['Thermoplasmatales'],
    estimates,
    study)
lipids_per_cell[study]['Archaea, others'] = (0.25, 5.0)

# Print. Limit precision for the display of floats to 5 decimal points.
print('minimum and maximum membrane lipid abundance estimates (fg / cell)\n')
for i, study in enumerate(lipids_per_cell):
    print(study)
    for taxon in lipids_per_cell[study]:
        print('{0}:\t{1}'.format(taxon,
                                tuple([float(f'{v:.5f}') for v in
                                       lipids_per_cell[study][taxon])))
    if i < 2:
        print()

```

```

minimum and maximum membrane lipid abundance estimates (fg / cell)

```

```

Sinninghe Damsté
ANME-1b:      (0.9, 1.2)
DHVE-6: (0.4, 0.625)
MCG+C3: (0.4, 0.625)
Thermoplasmatales: (0.25, 3.75)
Archaea, others: (0.25, 5.0)

Schouten
ANME-1b:      (1.2, 1.6)
DHVE-6: (0.53333, 0.83333)
MCG+C3: (0.53333, 0.83333)
Thermoplasmatales: (0.33333, 5.0)
Archaea, others: (0.25, 5.0)

Elling
ANME-1b:      (3.096, 4.93333)
DHVE-6: (1.376, 2.56944)
MCG+C3: (1.376, 2.56944)
Thermoplasmatales: (0.86, 15.41667)
Archaea, others: (0.25, 5.0)

```

3.2. Calculate the range of amount of lipid molecules produced by the Archaea in the water column

The amount of lipid molecules produced by an archaeal taxon in the water column, M is calculated as:

$$M_{\text{taxon}} = C_{\text{taxon}} L_{\text{taxon}}$$

We calculate minimum and maximum values based on the possible ranges of estimates, and sum the values for all archaeal taxa.

```
[12]: predictions_low = dict()
      predictions_high = dict()

      for study in lipids_per_cell:
          predictions_low[study] = [0 for depth in depths]
          predictions_high[study] = [0 for depth in depths]

          for i, depth in enumerate(depths):
              for taxon in archaeal_taxa:
                  predictions_low[study][i] += (
                      cells_per_L[taxon][i] *
                      min(lipids_per_cell[study][taxon]) /
                      1000000) # Convert fg to ng.
                  predictions_high[study][i] += (
                      cells_per_L[taxon][i] *
                      max(lipids_per_cell[study][taxon]) /
                      1000000) # Convert fg to ng.
```

The expected ranges of IPLs the archaeal population in the water column can theoretically support differs per study and looks like this:

```
[13]: # Plot the data.
      fig = plt.figure(figsize=(3, 5))
      ax = plt.subplot2grid((1, 1), (0, 0))

      ax.fill_betweenx(depths,
                      predictions_low['Elling'],
                      predictions_high['Elling'],
                      color='red',
                      alpha=0.25,
                      linewidth=0,
                      label='Elling et al.')
```

```
      ax.fill_betweenx(depths,
                      predictions_low['Schouten'],
                      predictions_high['Schouten'],
                      color='blue',
                      alpha=0.25,
                      linewidth=0,
                      label='Schouten et al.')
```

```
      ax.fill_betweenx(depths,
                      predictions_low['Sinninghe Damsté'],
                      predictions_high['Sinninghe Damsté'],
```

```

        color='green',
        alpha=0.25,
        linewidth=0,
        label='Sinninghe Damsté et al.')
```

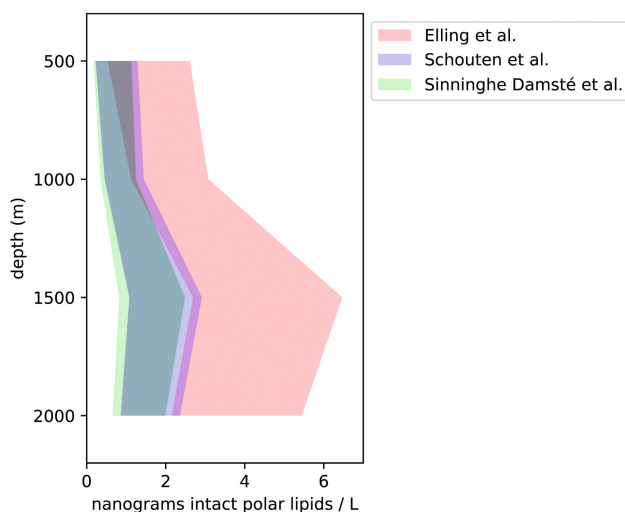
Styling.

```

ax.set_xlim([0, 7])
ax.set_xlabel('nanograms intact polar lipids / L')
ax.set_ylim([2200, 300])
ax.set_yticks(depths)
ax.set_ylabel('depth (m)')
ax.legend(loc='upper left', bbox_to_anchor=(1, 1))
plt.suptitle('Figure 3: expected archaeal membrane lipids')

plt.show()
plt.close()
```

Fig. 3: expected archaeal membrane lipids



The theoretical maximum of expected IPLs in the water column, i.e. the maximum predictions based on Elling et al., represents an ideal situation:

- All Archaea in the water column should be the maximum size reported in literature.
- All Archaea in the water column should produce the maximum reported amount of lipids per cell surface area. In Elling et al. this only happens right after growth phase, in the early stationary phase.

Both of these conditions are unlikely in the deep waters of the Black Sea. Nevertheless, even if we assume both, there is still an offset between expected and observed IPLs in the water column, which can be seen if we plot expected and observed together.

4. Plot expected versus observed amount of archaeal lipids in the water column, together with abundances of Archaea and *Candidatus Cloacimonetes*

```
14]: # Plot the data.
fig = plt.figure(figsize=(3.5, 7))
ax1 = plt.subplot2grid((1, 1), (0, 0))

# Plot abundances of Archaea and Candidatus Cloacimonetes.
ax1.plot(cells_per_L['Total Archaea'],
         depths,
         marker='D',
         color='grey',
         linestyle='--',
         label='Total Archaea')
ax1.plot(cells_per_L['Candidatus Cloacimonetes'],
         depths,
         marker='v',
         color='black',
         linestyle='--',
         label='Candidatus Cloacimonetes')

# Styling.
ax1.set_xlim([0, 8e6])
ax1.ticklabel_format(axis='x', style='sci', scilimits=(-2, 2))
ax1.set_xlabel('cells / L')
ax1.set_ylim([2200, 300])
ax1.set_yticks(depths)
ax1.set_ylabel('depth (m)')
ax1.legend(loc='lower left', bbox_to_anchor=(1, 0))

ax2 = ax1.twinx()

# Plot predicted and observed membrane lipids.
```

```

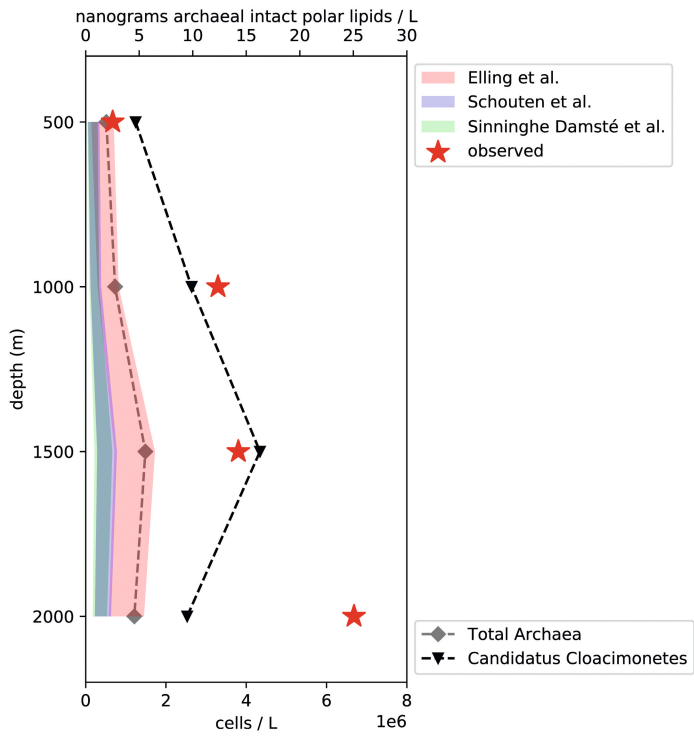
        predictions_high['Elling'],
        color='red',
        alpha=0.25,
        linewidth=0,
        label='Elling et al.')
ax2.fill_between(depths,
                 predictions_low['Schouten'],
                 predictions_high['Schouten'],
                 color='blue',
                 alpha=0.25,
                 linewidth=0,
                 label='Schouten et al.')
ax2.fill_between(depths,
                 predictions_low['Sinninghe Damsté'],
                 predictions_high['Sinninghe Damsté'],
                 color='green',
                 alpha=0.25,
                 linewidth=0,
                 label='Sinninghe Damsté et al.')
ax2.scatter(observed_IPLs,
            depths,
            marker='*',
            color='red',
            s=200,
            label='observed')

# Styling.
ax2.set_xlim([0, 30])
ax2.set_xlabel('nanograms archaeal intact polar lipids / L')
ax2.legend(loc='upper left', bbox_to_anchor=(1, 1))
plt.suptitle('Figure 4: observed versus expected archaeal membrane lipids')

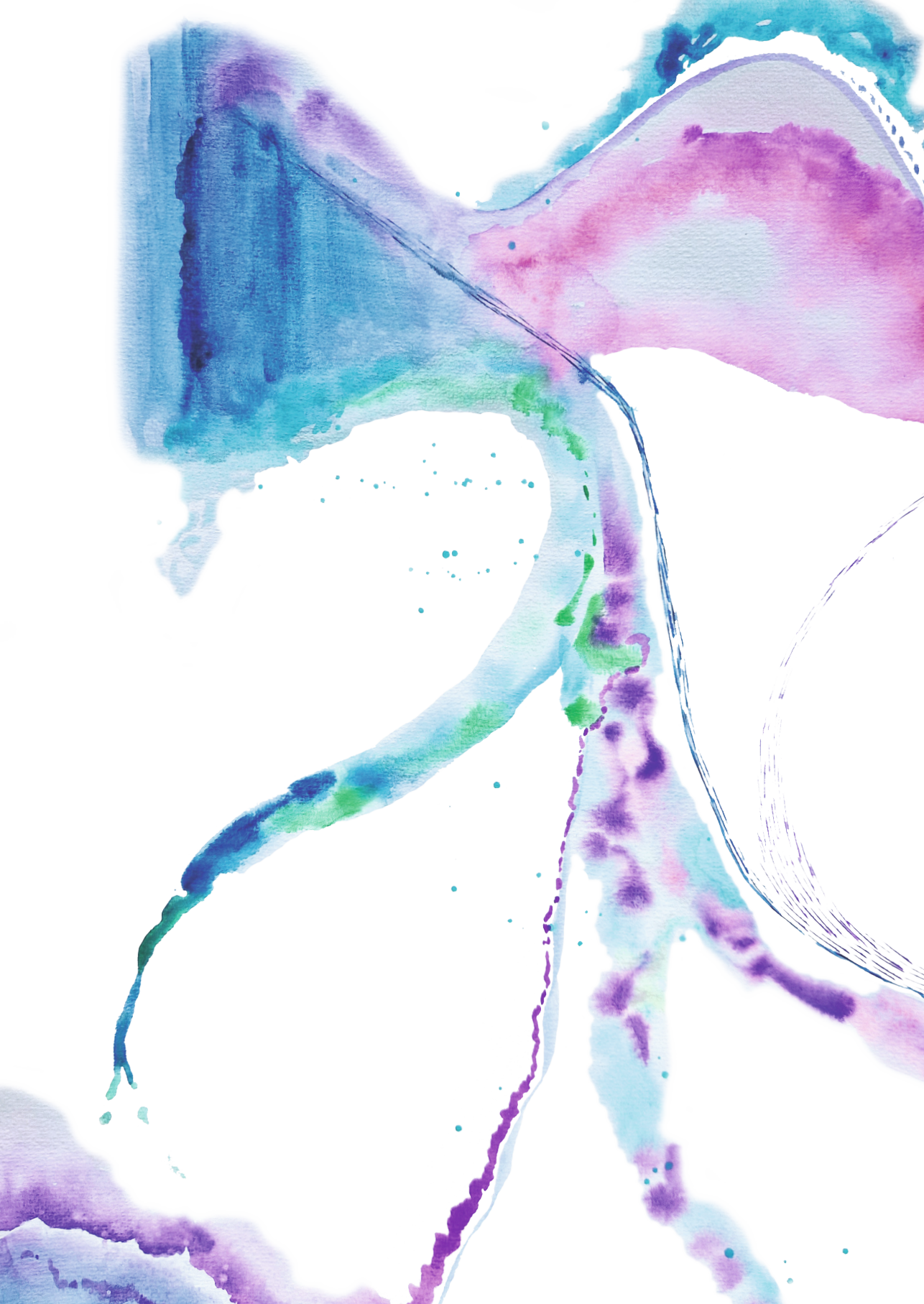
plt.show()
plt.close()

```


Fig. 4: observed versus expected archaeal membrane lipids



Thus, at 1,000m, 1,500m, and 2,000m there is a clear offset between observed IPLs and expected IPLs based on archaeal abundances in the water column.





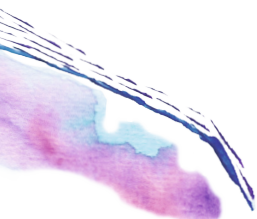
Chapter 3

Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT

F. A. Bastiaan von Meijenfeldt*, Ksenia Arkhipova*, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh

* F. A. Bastiaan von Meijenfeldt and Ksenia Arkhipova contributed equally to this work

Genome Biology **20**, 217 (2019)



Abstract

Current-day metagenomics analyses increasingly involve de novo taxonomic classification of long DNA sequences and metagenome-assembled genomes. Here, we show that the conventional best-hit approach often leads to classifications that are too specific, especially when the sequences represent novel deep lineages. We present a classification method that integrates multiple signals to classify sequences (Contig Annotation Tool, CAT) and metagenome-assembled genomes (Bin Annotation Tool, BAT). Classifications are automatically made at low taxonomic ranks if closely related organisms are present in the reference database and at higher ranks otherwise. The result is a high classification precision even for sequences from considerably unknown organisms.

Background

Metagenomics, the direct sequencing of DNA from microbial communities in natural environments, has revolutionized the field of microbiology by unearthing a vast microbial sequence space in our biosphere, much of which remains unexplored (97,108,111). With increases in DNA sequencing throughput, metagenomics has moved from analysis of individual reads to sequence assembly, where increases in sequencing depth have enabled de novo assembly of high-quality contiguous sequences (contigs), sometimes many kilobases in length (204). In addition, current state-of-the-art encompasses binning of these contigs into high-quality draft genomes, or metagenome-assembled genomes (MAGs) (92,93,105,145). The advance from short reads to contigs and MAGs allows the metagenomics field to answer its classical questions (32), “who is there?” and “what are they doing?” in a unified manner: “who is doing what?”, as both function and taxonomy can be confidently linked to the same genomic entity. Because assembly and binning can be done de novo, these questions can be applied to organisms that have never been seen before, and the discovery of entirely novel phyla is common still (93).

Several efficient tools for taxonomic classification of short-read sequences have been developed over the years, reflecting the read-based focus of the time. Most tools consider each read as an independent observation, whose taxonomic origin can be estimated by identifying best-hit matches in a reference database, either on read, k -mer, or translated protein level (see ref. 325 for an overview). Widely used programs such as Kraken (326) (k -mer based), CLARK (327) (discriminative k -mer based), and Kaiju (328) (protein-based) can process hundreds of thousands of sequencing reads per second. Without compromising accuracy, still faster approaches use mixture modelling of k -mer profiles, as implemented in FOCUS (329). Sometimes a Last Common Ancestor (LCA) algorithm is applied to allow for multiple hits with similar scores as the best hit (e.g. Kraken, MEGAN (330)).

Similar approaches are often applied to contigs, with classification often based on the best hit to a reference database. Although fast, the best-hit approach can lead to spurious specificity in classifications, for example when a genomic region is highly conserved or recently acquired via horizontal gene transfer (HGT) from a distantly related organism. As we will show below, the problem is particularly grave when the query contigs are very divergent from the sequences in the database, i.e. they are distantly related to known organisms. Whereas specificity (correctly classified/total classified) can be increased when only classifications at higher taxonomic ranks are considered, this approach is not desirable as taxonomic resolution is unnecessarily lost for query contigs that are closely related to known organisms.

Depending on their length, contigs may contain multiple open reading frames (ORFs), each of which contains a taxonomic signal. Integrating these signals should enable a more robust classification of the entire contig, yet surprisingly few tools exist that integrate distributed signals for contig classification. The viral-specific pipeline MetaVir2 (ref. 331) assesses the classification of up to five ORFs encoded on a contig. Recently, the MEGAN long-read algorithm was introduced (332), which allows users to taxonomically classify long sequences such as those generated by Oxford Nanopore Technologies or Pacific Biosciences sequencers. The algorithm works by partitioning the sequence into intervals based on the location of hits of a LAST (333) search.

In contrast, for taxonomic classification of MAGs, it is common to include information from multiple ORFs. Since the classification of complete genomes by using phylogenetic trees of multiple marker genes is well-established (334), MAG classification has followed these best practices. Some steps in the process can be automated, including initial placement in a low-resolution backbone tree by CheckM (51), specific marker gene identification, and backbone tree taxon selection by phyloSkeleton (335), and many tools are available for protein alignment, trimming, tree building, and display. However, interpretation of the resulting phylogeny remains a critical manual step, making this approach for genomic taxonomy a laborious task that does not scale well with the increasing number of MAGs being generated (see e.g. ref. 145).

Here we present Contig Annotation Tool (CAT) and Bin Annotation Tool (BAT), two taxonomic classifiers whose underlying ORF-based algorithm is specifically designed to provide robust taxonomic classification of long sequences and MAGs. Both tools exploit commonly used tools for ORF calling and homology searches. They require minimal user input and can be applied in an automated manner, yet all aspects are flexible and can be tuned to user preferences.

Benchmarking classification of sequences from novel taxa

Taxonomic classifiers are often benchmarked by testing them on sequences from novel taxa, i.e. that are not (yet) in the reference database (e.g. as in the CAMI challenge (212), and refs. 326,327,329). Alternatively, unknown query sequences can be simulated by using a “leave-one-out” approach, where the genome that is being queried is removed from the database (e.g. refs. 328,332). However, due to taxonomic biases in database composition, other strains from the same species, or other species from the same genus, may still be present. Thus, the leave-one-out approach does not reflect the level of sequence unknownness that is often encountered in real metagenomes, where the query sequences may be only distantly related to the ones in the reference database. A benchmark better suited to address this novelty is a “leave-entire-taxa-out” approach also known as clade exclusion, where all related sequences belonging to a certain taxonomic rank are removed from the database (e.g. refs. 326,336,337).

Here, we rigorously assess the performance of taxonomic classification tools by developing an extensive database reduction approach at different taxonomic ranks, where novel species, genera, and families are simulated by removing all the sequences of entire taxa from the database. In a second benchmark, we classified the high-complexity CAMI dataset (212). We show that the algorithm of CAT and BAT allows for the correct classification of organisms from known and unknown taxa and outperforms existing methods, especially for sequences that are highly unknown (i.e. with no close relatives in the database). Third, we used BAT in a real-world challenge to classify a large, recently published set of 913 MAGs from the cow rumen (145) that represent a wide range of novelty at all taxonomic ranks, and whose published taxonomic classifications involved extensive phylogenetic analyses.

Results and discussion

To test the performance of our newly developed taxonomic classification tools CAT and BAT, we thoroughly tested them in three independent benchmarks: (i) A clade exclusion experiment with increasing levels of sequence unknownness, (ii) the high-complexity gold standard CAMI assembly, and (iii) a recently published set of MAGs where the BAT classifications are compared to the published taxonomic classifications.

Contig classification with CAT

Benchmark 1: Classification of increasingly unknown sequences

We used CAT (Fig. 1) to classify ten simulated contig sets in the context of four reference databases with different levels of simulated unknownness, representing query sequences from (A) known strains, (B) novel species, (C) novel genera, and (D) novel families (see ‘Methods’). To assess the effect of the two key user parameters, r (hits included within *range* of top hits) and f (minimum *fraction* classification support), on precision, fraction of classified sequences, sensitivity, and taxonomic rank of classification, we ran CAT with a wide range of possible parameter values against all four reference databases (Fig. 2). This parameter sweep revealed a trade-off between the classification precision on the one hand and the taxonomic resolution and the fraction of classified sequences on the other hand. This general trend can be understood by considering that classifications at a low taxonomic rank (i.e. close to the species rank, high taxonomic resolution) will inevitably be increasingly imprecise, especially if closely related organisms are absent from the reference database. This might be resolved by classifying sequences at a higher taxonomic rank, but this leads to increased numbers of sequences not being classified or classified at trivially informative taxonomic ranks such as “cellular organisms” or “root”.

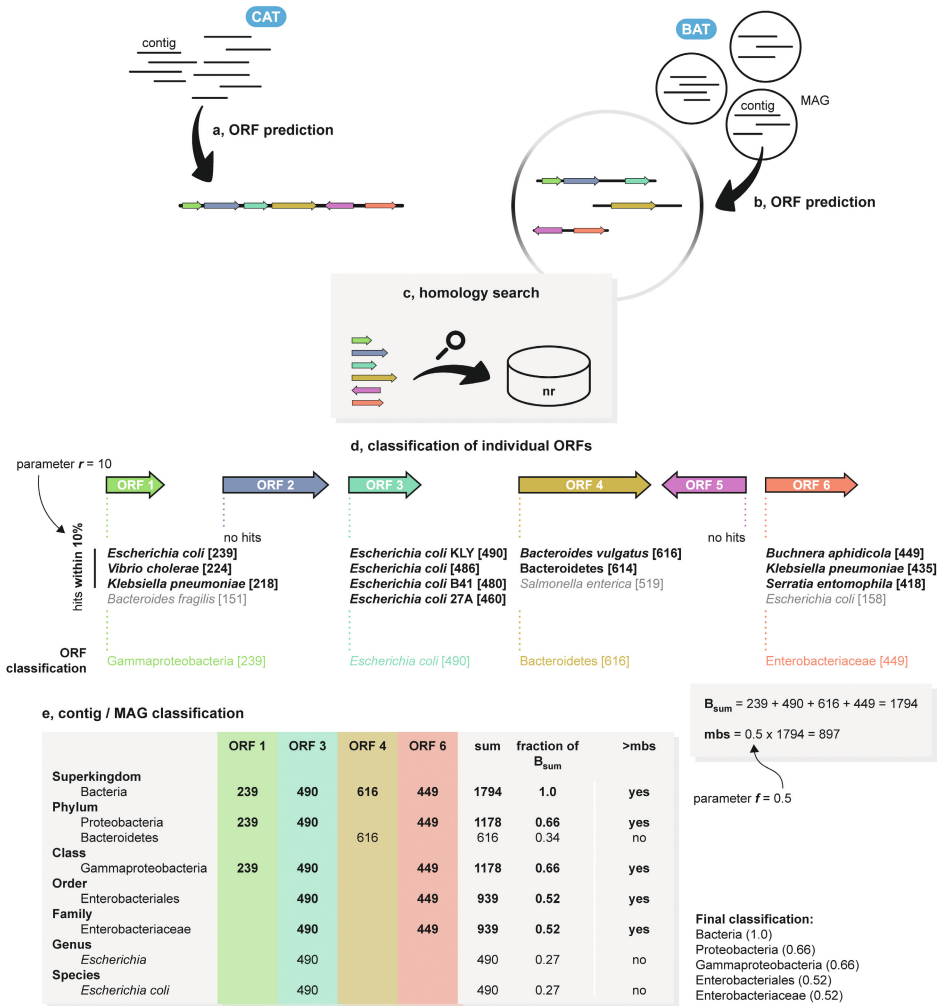


Fig. 1 | Contig and MAG classification with CAT and BAT. **a,b**, Step 1: ORF prediction with Prodigal. CAT analyses all ORFs on a contig, BAT analyses all ORFs in a MAG. **c**, Step 2: predicted ORFs are queries with DIAMOND to the NCBI non-redundant protein database (nr). **d**, Step 3: ORFs are individually classified based on the LCA of all hits falling within a certain range of the top hit (parameter r), and the top-hit bit-score is assigned to the classification. Bit-scores of hits are depicted within brackets. Hits in grey are not included in final annotation of the ORF. Parameter f defines minimal bit-score support (mbs). **e**, Step 4: contig or MAG classification is based on a voting approach of all classified ORFs, by summing all bit-scores from ORFs supporting a certain classification. The contig or MAG is classified as the lowest classification reaching mbs. The example illustrates the benefit of including multiple ORFs when classifying contigs or MAGs; a best-hit approach might have selected *Bacteroides vulgatus* or Bacteroidetes if an LCA algorithm was applied as its classification, as this part has the highest score to proteins in the database in a local alignment-based homology search. In the example, only six taxonomic ranks are shown for brevity; in reality, CAT and BAT will interpret the entire taxonomic lineage.

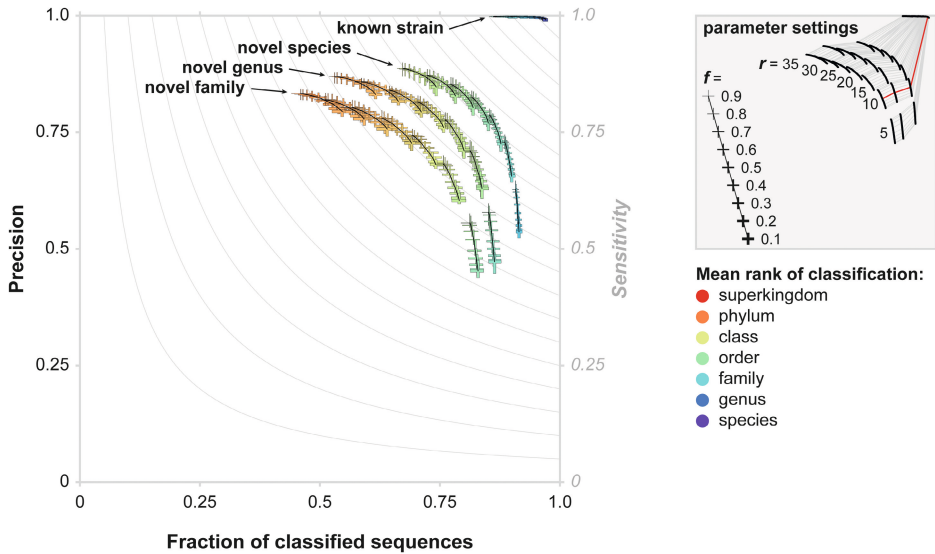


Fig. 2 | Classification performance of CAT for different levels of unknownness across a range of parameter settings. Thickness of markers indicates values of the f parameter; runs with similar r parameter values are connected with black lines. Markers indicate maximum and minimum values out of ten benchmarking datasets, bars cross at the means. Colour coding indicates the mean taxonomic rank of classification averaged across the then benchmarking datasets (minimum and maximum values not shown for brevity). Gray lines in the plot depict sensitivity, which is defined as the fraction of classified sequences times precision. Runs with equal parameter settings are connected in the parameter settings figure, showing that CAT achieves a high precision regardless of unknownness of the query sequence, by classifying sequences that are more unknown at higher taxonomic ranks. Default parameter combination ($r = 10, f = 0.5$) is shown in red.

The r parameter, which governs the divergence of included hits for each ORF, has the largest effect. As increasing r includes homologs from increasingly divergent taxonomic groups, their LCA is pushed back and classifications at low taxonomic ranks are lost, resulting in fewer classified sequences and classifications at lower taxonomic resolution (i.e. at higher taxonomic ranks), but with higher precision. The f parameter, which governs the minimum bit-score support required for classifying a sequence, has a smaller effect. Decreasing f results in classifications that are based on evidence from fewer ORFs, leading to more tentative classifications at lower taxonomic ranks. As a result, more sequences are classified at lower taxonomic ranks, albeit with a lower precision.

As a user increases r and f , this will increasingly result in high-rank classifications that are correct but ultimately uninformative. When low values of r and f are chosen, the classifications will be more specific (i.e. at a lower taxonomic rank) but more speculative (i.e. precision goes down). Based on the parameter sweep described above, we set the default values for CAT contig classification to $r = 10$ and $f = 0.5$ (red line in the legend of Fig. 2). Note that this value of $f = 0.5$ results in at most one classification, since $>50\%$ of the bit-score supports that classification.

Comparison to state-of-the-art taxonomic classifiers

We compared classification by CAT in this first benchmark to (i) the recently published LAST+MEGAN-LR algorithm (332), (ii) the widely used Kaiju algorithm (328), and (iii) a conventional best-hit approach with DIAMOND (209). Kaiju, designed for short-read classification, uses a best-hit approach with an LCA algorithm if equally good top-hits are found. Its underlying algorithm allows for the classification of long sequences as well and has recently been used as such (332,338,339). Final Kaiju classification is based on the hit with the maximum exact match (MEM), or on the highest scoring match allowing for mismatches (Greedy).

When classifying simulated contigs against the full reference database (known strains), all programs showed a similar precision and fraction of classified sequences (**Fig. 3a**). The mean taxonomic rank of classification is slightly higher for CAT and LAST+MEGAN-LR than for the other approaches (**Supplementary Table 1**), reflecting the conservative LCA-based classification strategies of the former two. DIAMOND best-hit does not use an LCA algorithm, and Kaiju only in cases where multiple hits have identical scores, and thus, they classify contigs according to the taxonomic rank of their match in the reference database.

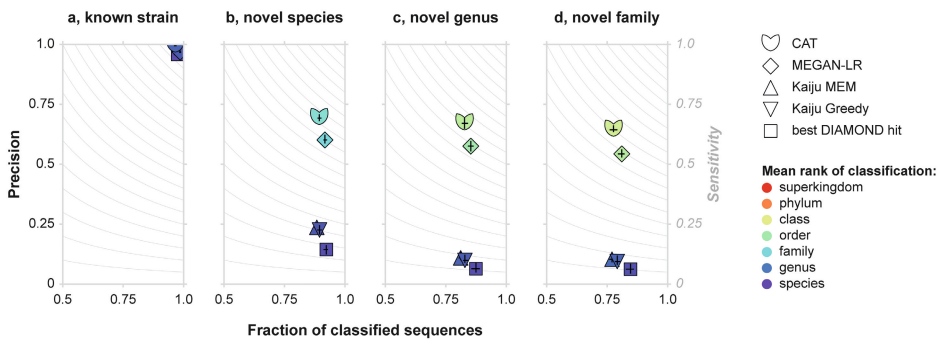


Fig. 3 | Classification performance of CAT, LAST+MEGAN-LR, Kaiju, and DIAMOND best-hit for different levels of unknownness. a, Classification of known sequences. b–d, Classification of simulated novel taxa for different levels of divergence from reference databases. Black bars indicate maximum and minimum values out of ten benchmarking datasets, bars cross at the means. Colour coding indicates the mean taxonomic rank of classification averaged across the then benchmarking datasets (minimum and maximum values not shown for brevity).

When novel species, genera, and families were simulated by removing related sequences from the database, precision declined rapidly for DIAMOND best-hit and Kaiju (**Fig. 3b–d**). The classifications called by these approaches are often too specific, because in databases where closely related sequences are absent, the singular best hit may still match a sequence that is annotated at a low taxonomic rank, although this annotation cannot match that of the query. This spurious specificity can be seen in the mean rank of classification, which stays close to the species rank, even when sequences from the same species, genus, or family were removed from the database

(Fig. 3b–d, Supplementary Table 1). CAT and LAST+MEGAN-LR clearly perform better in the face of such uncharted sequences. With default parameter settings, CAT has higher precision and sensitivity than MEGAN-LR and classifications are made at slightly higher taxonomic ranks.

Precision for CAT and LAST+MEGAN-LR increases when the sequence contains more ORFs with a DIAMOND hit to the database, whereas this is not the case for DIAMOND best-hit and Kaiju (Supplementary Fig. 1). Algorithms that integrate multiple taxonomic signals are thus well suited for taxonomic classification of long metagenomic sequences and MAGs (see below), but even the majority of contigs in our benchmarking sets that contained a single ORF are still classified correctly (Supplementary Fig. 1).

Sequences are classified correctly and automatically at the appropriate taxonomic rank

As a solution to the spurious specificity of the best-hit approach described above, classifications are sometimes assigned to a higher taxonomic rank such as genus, family, or even phylum. However, applying a rank cut-off may unnecessarily sacrifice taxonomic resolution in cases where the query sequences do have close relatives in the reference database and classification at a low taxonomic rank would be justified. Supplementary Fig. 2 shows that application of a rank cut-off to the best-hit classifications (e.g. reporting all classifications at the genus or phylum rank) does not solve the problem of spurious specificity as effectively as CAT does. CAT classifications have a higher precision than a best-hit cut-off on a rank comparable to its mean rank. For example, when novel families are simulated, the mean rank of classification for CAT is between order and class, and precision is much higher than best-hit classifications on those ranks, with a similar fraction of classified sequences (Supplementary Fig. 2d). Importantly, CAT has the highest precision on a per rank basis of any of the tested tools (Supplementary Fig. 3, Supplementary Table 2). This shows that CAT approach of integrating multiple taxonomic signals across a sequence leads to better classifications.

As shown in Fig. 2, the ORF-based voting algorithm ensures a high precision regardless of the level of unknownness of the query sequences, i.e. whether closely related sequences are present in the reference database or not. In some circumstances, taxonomic resolution is traded for precision: when classifying sequences that are more distantly related to the sequences in the reference database, hits will have weaker bit-scores and match sequences that are taxonomically more diverse. As a result of these conflicting signals, the algorithm automatically increases the taxonomic rank when classifying more divergent query sequences. Thus, no rank cut-off is needed for precise classifications, regardless of the composition of the metagenome.

Benchmark 2: Comparison to CAMI tools

Our second benchmark consisted of classifying the high-complexity gold standard assembly of the CAMI challenge (212). Classifying the CAMI dataset has two benefits. First, it allows us to compare CAT to any of the taxonomic classifiers tested in the CAMI challenge (referred to as “taxonomic bidders” in ref. 212). Second, CAMI simulated novel organisms, making it a complementary benchmarking approach as compared to the database reduction method in our first benchmark.

Since novel sequences are simulated, it is crucial that search databases are used that do not contain the simulated sequences. For this reason, an old copy of RefSeq (dated January 30th, 2015) was supplied during the CAMI challenge. Here, we also ran CAT with that old RefSeq reference database for a fair comparison against the other tools. However, one of the advantages of CAT and BAT is that they can be run with very large protein databases and hence have a larger search space for taxonomic classification beyond RefSeq. Thus, we also ran CAT with the nr databases from a similar date (January 23, 2015) as a reference. The nr database is the default option for CAT and BAT runs.

CAT performance measures on the high-complexity gold standard contig set (**Supplementary Table 3**) are plotted in **Supplementary Fig. 4** and can be compared to Supplementary Figure 18 and Supplementary Figure 19 in ref. 212. Average precision increases sharply if 99% of the data are considered (i.e. removal of taxa summing up to less than 1% of the total assembly length) as opposed to 100%. This is also true for most of the tools tested in the CAMI challenge. The reason for this observation is that precision in the CAMI challenge is measured on a ‘per bin’ basis, and erroneous classifications of single contigs thus weigh very heavily in this benchmark. If classifications that are seen in only a single or few contigs (i.e. are supported by short sequence length overall) are excluded, CAT showed very high average precision at all taxonomic ranks down to the genus level (**Supplementary Fig. 4**). Accuracy and average recall were high for higher ranks and decreased towards the species level. Misclassification was very low, with misclassification rates of up to 11% only at the lowest taxonomic ranks. Notably, CAT results with nr as a reference database (**Supplementary Fig. 4b**) were better than with RefSeq as reference (**Supplementary Fig. 4a**) for any of the measures. Average precision stayed above 90% down to the genus level if nr was used as a reference, higher than what is achieved by any of the tools tested in the CAMI challenge (see below). This highlights the benefit of using a large reference database for taxonomic classification.

We compared CAT to the other tools tested in the CAMI challenge by downloading their performance measures from the CAMI GitHub (**Supplementary Fig. 5**). The CAMI tools fall within two categories: One set of tools (taxator-tk 1.4pre1e, taxator-tk 1.3.0e, PhyloPythiaS+ mg c400, MEGAN 6.4.9) had low misclassification but

also low average recall and accuracy. The other set (PhyloPythiaS+ c400, Kraken 0.10.6-unreleased, Kraken 0.10.5) had high recall and accuracy, but very high misclassification rates towards species level. In contrast, CAT managed a medium (when using RefSeq as reference database) to high (when using nr as reference database) average recall and accuracy, with a very low misclassification rate. The misclassification rate was lower than that of the CAMI tools, with the exception of taxator-tk (both versions), which classified very few sequences in general. CAT scored among the highest average precision with 99% of the data. Thus, CAT has a high average precision and combines the high average recall and accuracy of the second set of tools with the low misclassification of the first.

The ORF-based algorithm is fast and has a very low memory requirement

CAT is about two times faster than LAST+MEGAN-LR (Fig. 4a) and outperforms all other programs tested in our first benchmark in terms of memory usage (Fig. 4b). The slowest and most memory intensive step is the DIAMOND search for homologs in the vast nr database, which due to the flexible nature of our implementation can be optimized for a specific use case (see **Supplementary Table 4**) or replaced by any protein aligner of a user's choice, as can the search database.

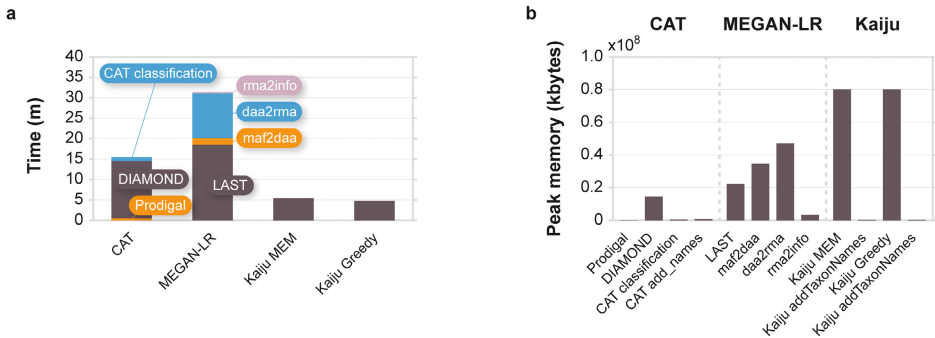


Fig. 4 | Computer resource usage by CAT, LAST+MEGAN-LR, and Kaiju. a, Run-time and **b**, peak memory usage. In **a**, classification by CAT and Kaiju includes adding taxonomic names to the classification; in **b**, these steps are depicted separately.

We classified the CAMI high complexity dataset with recent versions of the tools tested in our first and second benchmarks. This analysis showed that CAT is faster than MEGAN6, LAST+MEGAN-LR, and taxator-tk and has a memory footprint that is similar to or lower than any of the tested tools (**Supplementary Table 4**).

MAG classification with BAT

Benchmark 3: Classification of 913 metagenome-assembled genome bins (MAGs)

Next, we set out to apply the algorithm to MAGs, i.e. draft genomes that can be generated from metagenomes by assembly and binning. Since the typical pipeline

to generate MAGs is reference database independent, they can be distantly related to known organisms. As benchmark set, we picked 913 recently published MAGs from the cow rumen (145) that represented a wide range of novelty at different taxonomic ranks (**Supplementary Fig. 6a**). The published classifications were based on the placement of the MAGs in a backbone tree and subsequent refinement, a slow process that includes various manual steps and visual screening (145). At the time of our study, the MAGs were not yet included in the reference database, providing an ideal test case for our automated classification tool BAT.

The 913 MAGs were previously assessed to be $\geq 80\%$ complete and have $\leq 10\%$ contamination and contain between 541 and 5,378 ORFs each (**Supplementary Fig. 6b**). We ran BAT with default parameter settings for MAGs classification ($r = 5$, $f = 0.3$). The low r value ensures that individual ORFs are annotated to an LCA with a relatively low taxonomic rank, as hits within 5% of the highest bit-score are considered. The low f value reports taxonomic classifications that are supported by at least 30% of the bit-score evidence. While this could be considered a speculative call when contigs with relatively few encoded ORFs are annotated, the much higher number of ORFs in MAGs means that even classifications with relatively low f values are backed by a high number of ORFs and precision is thus expected to be high (**Supplementary Fig. 1**). We scored the consistency between BAT and the published classifications (**Fig. 5a**), dividing consistent classifications into three groups: (i) BAT can be more conservative than the published classification, i.e. BAT classifies the MAG to an ancestor of the published classification; (ii) classifications can be equal; and (iii) BAT can be more specific. Alternatively, BAT can classify a MAG inconsistently, i.e. in a different taxonomic lineage than the original publication. As shown in **Fig. 5a**, 885 of 913 MAGs (97%) were classified consistently with the original publication. If parameter f is relaxed, mean rank of classification for the MAGs increases (**Fig. 5b**). Importantly, decreasing the value of f has little effect on inconsistency rate. Thus, changing this parameter will mainly lead to a change in the rank of classification, while the taxonomic lineage will remain unchanged. Finally, classifying these MAGs with two MAG classification tools that are still under development, lastTaxa (<https://gitlab.com/jfroula/lasttaxa>) and GTDB-Tk (<https://github.com/Ecogenomics/GTDBTk>), yielded very similar results (**Supplementary Table 5**).

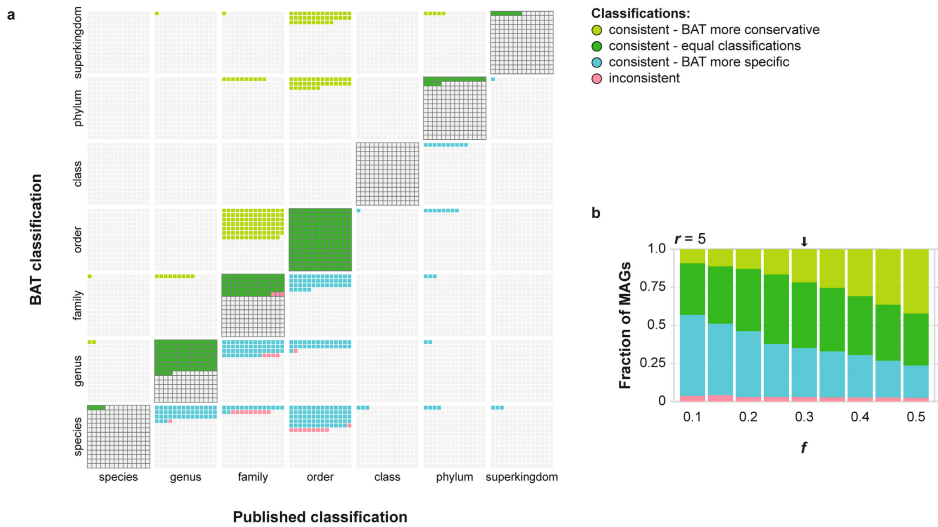
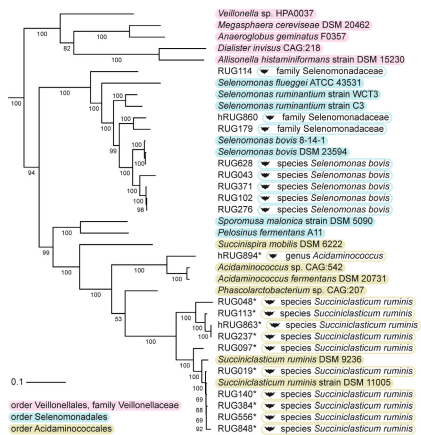


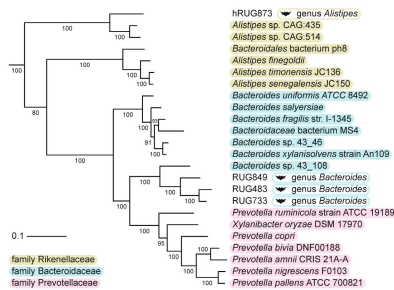
Fig. 5 | Classification of 913 MAGs with BAT. a, Consistency between BAT classifications and published classifications with default parameter settings ($r = 5, f = 0.3$). **b**, The mean rank of classification can be increased by increasing f . Arrow indicates BAT results for its default parameter settings.

To assess the taxonomy of the 28 inconsistently classified MAGs (at $r = 5, f = 0.3$), we placed them in a phylogenomic tree with closely related genomes and observed their closest relatives, the published classifications, and the BAT classifications. As shown in **Fig. 6**, BAT classified all 28 inconsistently classified MAGs more precisely and at a higher taxonomic resolution than the published classifications. Note that this may be due to these closely related reference genomes being new additions to the database since the research was performed. Together, these results highlight the benefit of using BAT for the rapid, automated, and high-resolution taxonomic classification of novel microbial lineages at a range of unknownness.

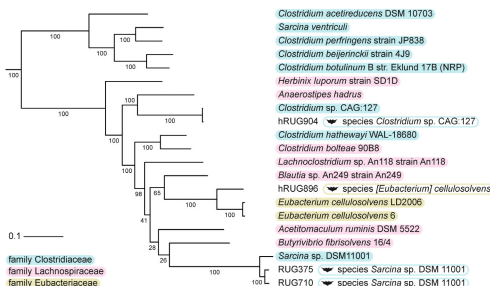
a, family Veillonellaceae / order Selenomonadales (*)



b, family Prevotellaceae



d, family Lachnospiraceae



c, genus Succinatimonas

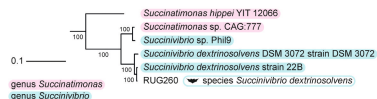


Fig. 6 | Tree placement of the 28 inconsistently classified MAGs that were assigned to five different taxa according to the original classifications (a–d). Headers of subfigures refer to the published classifications. In a, MAGs published as Selenomonadales are marked with an asterisk. Taxonomic classification of reference genomes is indicated in shades boxes. BAT classifications of MAGs are indicated in open boxes.

Conclusions

Metagenomics continues to reveal novel microorganisms in all environments in the biosphere, whose genome sequences can be reconstructed with high accuracy by using high-throughput DNA sequencing and modern sequence assembly and binning tools. Taxonomically classifying these uncharted sequences remains challenging, partly because the vast natural biodiversity remains highly underrepresented in even the largest reference databases, partly because existing classifiers are built to classify short sequencing reads, and partly because interpreting trees is manual work.

We presented CAT and BAT, a set of tools that exploits DIAMOND homology searches of individual ORFs called by Prodigal, LCA annotation, and a user-definable weighting to classify long contigs and metagenome-assembled genomes (MAGs). As we have shown, these query sequences contain a wealth of information that allows their accurate taxonomic classification at appropriate taxonomic ranks, i.e. at a low rank when closely related organisms are present in the database, and at a high rank

when the sequences are divergent or highly novel. We have shown that the low precision of conventional best-hit approaches when classifying novel taxa can be overcome by a voting algorithm based on classifications of multiple ORFs. Elegantly, sequences from organisms that are distantly related to those in the reference database are automatically classified at a higher taxonomic rank than known strains. ORFs on divergent sequences will hit a wider variety of different taxa both on the individual ORF level and between ORFs. Such conflict of classifications is automatically resolved by the algorithm by providing a more conservative classification, so no taxonomic cut-off rank for classification needs to be pre-defined. In metagenomes containing both known and unknown sequences, the algorithm vastly outperforms best-hit approaches and a range of state-of-the-art taxonomic classifiers in precision.

CAT and BAT supplement a modern metagenomics workflow in various ways. For example, CAT can be used after metagenome assembly to confidently classify all contigs. Since contigs are longer sequences and thus contain more information than individual reads, we expect that classification of the original reads in terms of classified contigs results in better profiling estimates than those based on the reads alone. Indeed, a comparison in ref. 212 between taxonomic bidders and dedicated taxonomic profilers (whose output is an abundance profile but not classification of individual sequences) showed that on average bidders estimated taxon abundance more accurately than profilers. With increases in contig lengths due to advances in assembly algorithms and more deeply sequenced metagenomes, as well as increasingly available long-read metagenomic sequencing datasets, CAT classifications will become even more precise in the future. Moreover, BAT will rapidly provide taxonomic classifications of MAGs without requiring a full phylogenomics pipeline and subsequently visual inspection of the tree. CAT classifications of individual contigs within MAGs can be used to identify taxonomic outliers, and flag those as possible contamination. As most binning tools do not incorporate taxonomic signals (e.g. refs. 193,197), CAT classification can be considered as independent evidence and might be used to decide on the inclusion of specific contigs in a MAG.

BAT provides a robust and rapid classification of MAGs in a single operation, but is not a replacement for high-confidence phylogenomic tree construction based on marker gene superalignments which remains the gold standard (334). However, BAT queries the full NCBI non-redundant reference database (nr) and the taxonomic context is thus much bigger than any phylogenomic tree that depends on completely sequenced genomes. For example, the backbone tree of CheckM currently includes only 5,656 genomes (51). BAT classification is fully automated and can be run on a set of MAGs with minimal user input, allowing MAG classification to be scaled up considerably as we showed here for over 900 MAGs that were classified consistently with the original publication in almost all cases. Notably, in all inconsistent cases,

Chapter 3

we identified genomes that were more closely related to the BAT classification than to the published (manual) classification.

As long as sequence space is incompletely explored and reference databases represent a biased view of the tree of life (97,108), algorithms designed to address the abundant uncharted microbial sequences will be needed to make sense of the microbial world. Decreasing sequencing costs and improvement of alignment and binning algorithms have moved metagenomics from the analysis of short reads towards contigs and MAGs, improving our understanding of microbial ecosystems to a genomic resolution. As these data will only increase in the coming years, we presented a robust solution to their specific challenges that we expect will play an important role in future metagenomics workflows.

Methods

Explanation of the algorithm

Both CAT and BAT take high-quality long DNA sequences in FASTA format as input (**Fig. 1**), such as assembled contigs or corrected long Oxford Nanopore Technologies or Pacific Biosciences reads (340,341). First, ORFs are predicted with Prodigal (290) in metagenome mode, using default parameter settings (genetic code 11) (**Fig. 1a, b**). Predicted proteins can also be independently supplied to CAT/BAT in case a user prefers a different gene caller than Prodigal.

Next, protein translations of the predicted ORFs are queried against the National Center for Biotechnology Information (NCBI) non-redundant protein database (nr) (156) using DIAMOND (209) blastp (*e*-value cut-off of 0.001, BLOSUM62 alignment matrix, reporting alignments within 50% range of top hit bit-score) (**Fig. 1c**). The nr database is currently the largest sequence database where all sequences are assigned to clades in NCBI Taxonomy (342). A separate BLAST tabular output file can also be supplied together with the predicted protein file, in which case CAT/BAT starts directly with classification.

Taxonomic classification of the query sequences is then carried out based on a voting approach that considers all ORFs on a query with hits to the reference database. Here, the main difference between CAT and BAT is that CAT considers ORFs on a single contig, whereas BAT considers ORFs on all contigs belonging to a MAG. CAT and BAT also have slightly different default parameter values (see below).

First, the algorithm infers the taxonomic affiliation of individual ORFs based on the top DIAMOND hits (**Fig. 1d**). To account for similarly high-scoring hits in potentially different clades, hits within a user-defined range of the top hit bit-score to that ORF are considered and the ORF is assigned to the LCA of their lineages (parameter *r* for *range*, by default hits with bit-scores within 10% or 5% range of the top hit bit-score are included, *r* = 10 for CAT and *r* = 5 for BAT, respectively). By adjusting parameter *r*, the user can tune how conservative CAT is in the classification of individual ORFs. For example, increasing *r* results in more divergent hits being included that together are likely to have a deeper LCA, thus leading to a more conservative ORF classification at a higher taxonomic rank. In contrast, decreasing *r* leads to a more specific classification since fewer and more similar hits will be included, likely with a narrower taxonomic range. This accounts for conserved or HGT-prone genes that are highly similar in diverse taxa by assigning them a high-rank classification. The top hit bit-score for each ORF is registered for the subsequent voting process (**Fig. 1d**).

Next, the query contig or MAG is evaluated by summing the bit-scores for each taxon identified among the classifications of all ORFs, as well as their ancestral lineages up to the taxonomy root (**Fig. 1e**). The query contig or MAG is then assigned to a taxon, if the total bit-score evidence for that taxon exceeds a cut-off value (mbs, minimal bit-score support), which is calculated as a fraction (parameter f for *fraction*) of the sum of the bit-scores of all ORFs ($mbs = f \times B_{\text{sum}}$, by default $f = 0.5$ for CAT and $f = 0.3$ for BAT). For example, if parameter f is set to 0.5, this means that a contig is assigned to a taxon if the majority of the sum of the bit-scores of all ORFs supports that classification ($mbs = 0.5 \times B_{\text{sum}}$). This is done at multiple taxonomic ranks including phylum, class, order, family, genus, and species. The algorithm stops at the taxonomic rank where the total bit-score supporting the classification drops below the minimal bit-score support value, so CAT/BAT automatically finds the lowest rank taxonomic classification that is still reliable (**Fig. 1e**). Note that with CAT default values ($f = 0.5$), only one classification is given per sequence, and there can be no conflicting classifications at different ranks (e.g. a species-level classification conflicting with a genus-level classification). When $f < 0.5$ is set by the user, multiple lineages at a given taxonomic rank may exceed the threshold, and all will be written to the output file. A user can decide on the appropriate (rank of) classification based on support values that represent the fraction of summed bit-score that supports the classification. While these support values are indicative of the prediction precision (**Supplementary Fig. 7a**), in contrast to the total bit-score alone (**Supplementary Fig. 7b**), it should be noted that they cannot be interpreted as statistical probabilities.

Output files

For each query contig or MAG, the full taxonomic lineage of the lowest-rank supported classification is written to the output file, together with support values per rank (i.e. the fraction of B_{sum} that is represented by the taxon). In addition, the number of ORFs found on the contig or MAG and the number of ORFs on which the classification is based are written to the output file. An extra output file containing information about individual ORFs is also generated, including classifications of ORFs and an explanation for any ORF that is not classified. We advise the user caution when interpreting the classifications of short contigs that are based on relatively few ORFs as they will be less robust than the classifications of long contigs or MAGs (**Supplementary Fig. 1**).

Helper programs

The CAT/BAT package comes bundled with three helper utilities, “prepare”, “add_names”, and “summarise”. “Prepare” only needs to be run once. It downloads all the needed files including NCBI taxonomy files and the nr database. It constructs a DIAMOND database from nr and generates the files needed for subsequent CAT and BAT runs. Because the first protein accession in nr not always represents the LCA of all protein accessions in the entry, “prepare” corrects for this in the protein

accession to taxonomy id mapping file (prot.accession2taxid). After running CAT/BAT, “add_names” will add taxonomic names to the output files, either of the full lineage or of official taxonomic ranks alone (superkingdom, phylum, class, order, family, genus, species). “Summarise” generates summary statistics based on a named classification file. For contig classification, it reports the total length of the contigs that are classified to each taxon. For MAG classification, it reports the number of MAGs per taxon.

Generation of contigs for clade exclusion benchmarking datasets

To test the performance of the algorithm in a first benchmark, we artificially generated contigs from known genome sequences in the RefSeq database (343) (**Supplementary Table 6**). We randomly downloaded one genome per taxonomic order from bacterial RefSeq on July 7, 2017 (163 orders in total) and cut the genomes into at most 65 non-overlapping contigs, generating a set of ~10,500 contigs with known taxonomic affiliation. Contig lengths were based on the length distribution of eight assembled real metagenomes deposited in the Sequence Read Archive (SRA) (344) (assembly with metaSPAdes v3.10.1 (ref. 204) after quality filtering with BBDuk that is included with BBTools v36.64 (<https://sourceforge.net/projects/bbmap/>); see **Supplementary Table 6**), with a minimum length of 300 nucleotides. This was done ten times to construct ten different benchmarking datasets sampled from 163 different genomes, each from a different taxonomic order.

Viruses remain vastly under-sampled, and the sequences in the database remain a small fraction of the total viral sequence space (345). Moreover, the hierarchy of the viral taxonomy is not as deeply structured as the taxonomy of cellular organisms (346). Based on these considerations, we did not explicitly assess the performance of our tool on viral sequences. However, we expect that classification of viruses will be readily possible when closely related viruses are present in the reference database.

Reference databases with increasing levels of unknownness

The benchmarking datasets generated above are derived from genomes whose sequences are also present in the reference database, corresponding to the perhaps unlikely scenario where the query sequences in the metagenome are identical to known strains in the database. To benchmark our tools in the context of discovering sequences from novel taxa, we next generated novel reference databases with increasing levels of unknownness by removing specific taxonomic groups from nr. In addition to the original nr database (known strains), three derived databases were constructed to reflect the situation of discovering novel species, genera, and families. This was done by removing all proteins that are only present in the same species, genus, or family as any of the 163 genomes in the benchmarking dataset. To do this, either we removed the sequences from the database itself, or if a protein was identical in sequence to a protein in another clade, we changed the protein

accession to taxonomy id mapping file to exclude the query taxon. In contrast to many other taxonomic classification tools, all the programs that we compared (CAT, DIAMOND best-hit, LAST+MEGAN-LR, and Kaiju) allowed such custom files to be used. The three reduced databases and associated mapping files thus reflect what nr would have looked like if the species, genus, or family of the genomes present in the benchmarking dataset were never seen before. This was done independently for each of the ten different benchmarking datasets, resulting in a total of 30 new reference databases to rigorously test the performance of our sequence classification tools in the face of uncharted microbial sequences. Simulating unknownness like this provides a better benchmark for classification of unknown sequences than a leave-one-out approach where only the query genome is removed from the reference database (e.g. refs. 328,332), because close relatives of the query may still be present in the latter case.

Programs, parameters, and dependencies

Nr database and taxonomy files were downloaded on November 23, 2017. Prodigal v2.6.3 (ref. 290) was used to identify ORFs on the simulated contigs. DIAMOND v0.9.14 (ref. 209) was used to align the encoded proteins to the reference databases for CAT and for the DIAMOND best-hit approach. Kaiju v1.6.2 (ref. 328) was run both in MEM and Greedy mode with SEG low complexity filter enabled. The number of mismatches allowed in Greedy mode was set to 5. For LAST+MEGAN-LR, LAST v914 (ref. 333) was used to map sequences to the databases with a score penalty of 15 for frameshifts, as suggested in ref. 332. Scripts in the MEGAN v6.11.7 (ref. 332) tools directory were used to convert LAST output to a classification file. The maf2daa tool was used to convert LAST output to a .daa alignment file. The daa2rma tool was used to apply the long-read algorithm. ‘--minSupportPercent’ was set to 0 and the LCA algorithm to longReads, and the longReads filter was applied. ‘--topPercent’ was set to 10 and ‘--lcaCoveragePercent’ to 80 (MEGAN-LR defaults). The rma2info tool was used to convert the generated .rma file to a classification file. When a reduced database was queried, the appropriate protein accession to taxonomy id mapping file was supplied via its respective setting (see the section ‘**Reference databases with increasing levels of unknownness**’ above).

Scoring of contig classification performance

For contig classification, we scored (i) the fraction of classified contigs, (ii) sensitivity, (iii) precision, and (iv) mean and median rank of classification (**Supplementary Fig. 8**). Classifications were compared at the taxonomic ranks of species, genus, family, order, class, phylum, and superkingdom. In those cases where $f < 0.5$ and multiple classifications reached the mbs threshold, we chose the lowest classification that reached a majority vote (i.e. as if $f = 0.5$) for calculating the four performance measures i–iv. This means CAT classifications were more conservative in those (rare) cases. Contigs with a classification higher than the superkingdom rank (e.g.

“cellular organisms” or “root”) were considered unclassified, as these classifications are trivially informative in our benchmark. For all tools, a classification was considered correct if it was a subset of the true taxonomic lineage, regardless of rank of classification. If a classification was consistent with the true taxonomic lineage but classified too specifically (e.g. at the species rank whereas the query is a novel family), it was considered incorrect. For classifications that are shown per rank, only that part of the lineage that is too specific is considered incorrect.

The mean and median taxonomic rank of classification were calculated for all classified contigs, where the ranks species-phyllum were given the integer values 0–6, respectively. Even though the true distance between taxonomic ranks may vary (40), calculating mean taxonomic rank in this fashion does serve as a proxy to show that classifications are called at higher taxonomic ranks “on average” under certain parameter conditions or e.g. with higher divergence of the query sequence from the reference database. Sensitivity and precision were scored as (correctly classified/total number of contigs) and (correctly classified/total number of classified contigs), respectively. Thus, all performance measures are a property of the whole contig set and not of single taxonomic classifications as with some measures in the CAMI challenge benchmark further on. Wherever error bars are shown, they represent the maximum and minimum values out of the ten benchmark datasets.

CAMI high-complexity gold standard benchmark

In a second benchmark, we downloaded the high-complexity gold standard assembly together with the taxonomy files and NCBI RefSeq database (dated January 30, 2015) that was supplied with the CAMI challenge (212). We ran CAT on the assembly with RefSeq and nr (dated January 23, 2015) as reference databases. Importantly, both databases did not contain any of the query sequences yet.

We scored performance in exactly the same way as in the CAMI challenge, which allows us to compare the results of CAT to any of the taxonomic classifiers tested (“taxonomic binners”). In short, all four measures (accuracy, misclassification, average precision, average recall) are a function of the number of classified base pairs and not of classified contigs as in the benchmark above. If a tool classifies a sequence on a taxonomic rank that is not present in the gold standard, it is not taken into account. Thus, there is no penalty for classifications that are too specific. Accuracy is (number of correctly classified base pairs/total number of base pairs), misclassification (number of incorrectly classified base pairs/total number of base pairs), and both are thus a property of the whole assembly. Precision is a measure of the purity of a predicted taxonomic bin (i.e. all sequences from a single predicted taxon) with (number of correctly assigned base pairs/total assigned base pairs). Average precision is the mean precision of all predicted taxonomic bins and is thus very sensitive to misclassified small bins. Therefore in ref. 212 in addition to precision

measures of the full data, small bins summing up to 1% of the data are excluded and precision is recalculated. We did the same. Recall is a measure of the completeness of a real taxon bin (i.e. all sequences from a single query taxon), with (number of correctly assigned base pairs/real number of base pairs). Average recall is mean recall for all real taxon bins.

For a comparison with all taxonomic classifiers tested in the CAMI challenge, we downloaded the summaries from https://github.com/CAMI-challenge/firstchallenge_evaluation/tree/master/binning/tables/plot/supervised/summary_high.csv and https://github.com/CAMI-challenge/firstchallenge_evaluation/tree/master/binning/tables/plot/supervised/summary99_high.csv.

MAG classification

For a third benchmark, 913 high-quality draft genome bins (MAGs) (completeness \geq 80%, contamination \leq 10%) from the cow rumen generated with both conventional metagenomics as well as Hi-C binning methods (145) were downloaded from the DataShare of the University of Edinburgh (<https://datashare.is.ed.ac.uk/handle/10283/3009>). Taxonomic classification of the MAGs was downloaded from the supplementary data that accompanies the paper and manually corrected if the names did not match our taxonomy files (**Supplementary Table 5**). To save disk space on the alignment file being generated, we ran BAT on batches of 25 genomes each. Akin to the contig classification case in the first benchmark, we only considered classifications by BAT at official taxonomic ranks and chose the majority classification in those cases where BAT gave more than one classification for a MAG (i.e. as if $f = 0.5$ for that MAG) resulting in more conservative classifications.

To manually assess the 28 MAGs whose classification was inconsistent with the published classifications, we created a phylogenomic tree of those bins together with closely related genomes that were downloaded from PATRIC (291) on January 16, 2018. CheckM v1.0.7 (ref. 51) was used to extract 43 phylogenetically informative marker genes that were realigned with Clustal Omega v1.2.3 (ref. 280). We concatenated the alignments to create a superalignment and included gaps if a protein was absent. We constructed a maximum likelihood tree with IQ-TREE v1.6.3 (ref. 281), with ModelFinder (282) set to fit nuclear models (best-fit model LG+R7 based on Bayesian Information Criterion), including 1,000 ultrafast bootstraps (283). Per clade, rooted subtrees were visualized in iTOL (284).

We classified the MAGs with 2 MAG classification tools that are still under development, lastTaxa (<https://gitlab.com/jfroula/lasttaxa>) and GTDB-Tk v0.2.2 (<https://github.com/ECogenomics/GTDBTk>). LastTaxa predicts ORFs with Prodigal and searches the nr database with LAST, after which classification is based on the majority classification of individual ORFs. LastTaxa was run on the same nr dataset

as BAT, and they can thus be directly compared. GTDB-Tk first identifies marker genes and places the MAG in a reference genome tree based on these marker genes (see also ref. 40). GTDB-Tk was run with the classify workflow with release 86 of the GTDB-tk reference database. This database was constructed after the publication of ref. 145. The results of these comparisons can be found in **Supplementary Table 5**.

Usage of computer resources

Run time and peak memory usage were estimated with the Linux `/usr/bin/time` utility. Elapsed wall clock time and maximum resident set size were scored for runs of CAT, MEGAN-LR, and Kaiju, classifying contig set #1 (10,533 contigs, see **Supplementary Table 6**) with the nr reference database. All tools were run with default parameter settings. Runs were performed on a machine with an Intel Xeon Gold 6136 Processor, 128 GB of memory, 24 cores, and 48 threads. Whenever one of the programs allowed for the deployment of multiple threads, all were used.

We estimated run time and peak memory usage for CAT, MEGAN-LR, Kaiju, and recent versions of the CAMI tools on the CAMI high-complexity dataset, with the NCBI RefSeq database that was supplied with the CAMI challenge as a reference. PhyloPythiaS+ was excluded because it needs a custom database that cannot be constructed based on RefSeq. The CAMI tools were run as suggested in their respective manuals and/or as done in the CAMI challenge (see **Supplementary Table 4**). MEGAN was run on a single metagenomic read file (out of 5 in the challenge); all the other tools were run on the gold standard assembly (42,038 contigs). Runs were performed on a machine with an Intel Xeon E5-2667 v3 Processor, 512 GB of memory, and 16 cores/threads. Whenever one of the programs allowed for the deployment of multiple threads, all were used.

CAT and BAT have been tried and tested on 128 GB machines.

Availability of data and materials

CAT and BAT are available under the MIT License at GitHub: <https://github.com/dutilh/CAT>. A version of the source code used in this manuscript is deposited on Zenodo at <https://doi.org/10.5281/zenodo.3403695>. All benchmarking datasets and reference databases with increasing levels of unknownness are available from the authors upon request. The contigs in the first benchmark are based on Bacterial RefSeq (343) and the databases based on nr (156). The second benchmark is the CAMI benchmark (212), and the third is from Stewart et al. (145).

Acknowledgements

We thank Jan Kees van Amerongen for his technical support and Johannes Dröge for his advice on taxator-tk.

Funding

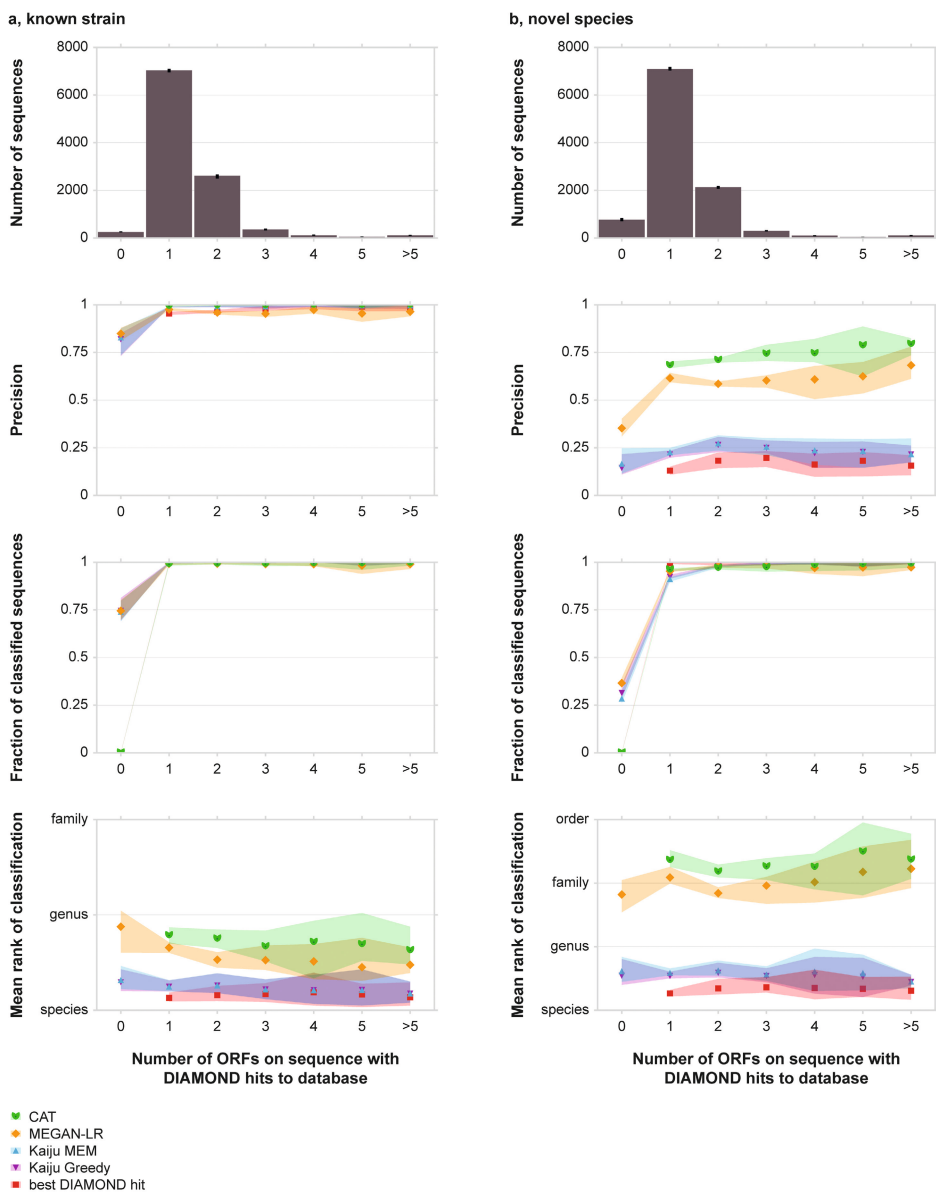
This work was supported by the Netherlands Organization for Scientific Research (Vidi grant 864.14.004) to B.E.D. and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (Science Without Borders program) to D.D.C. and F.H.C.

Contributions

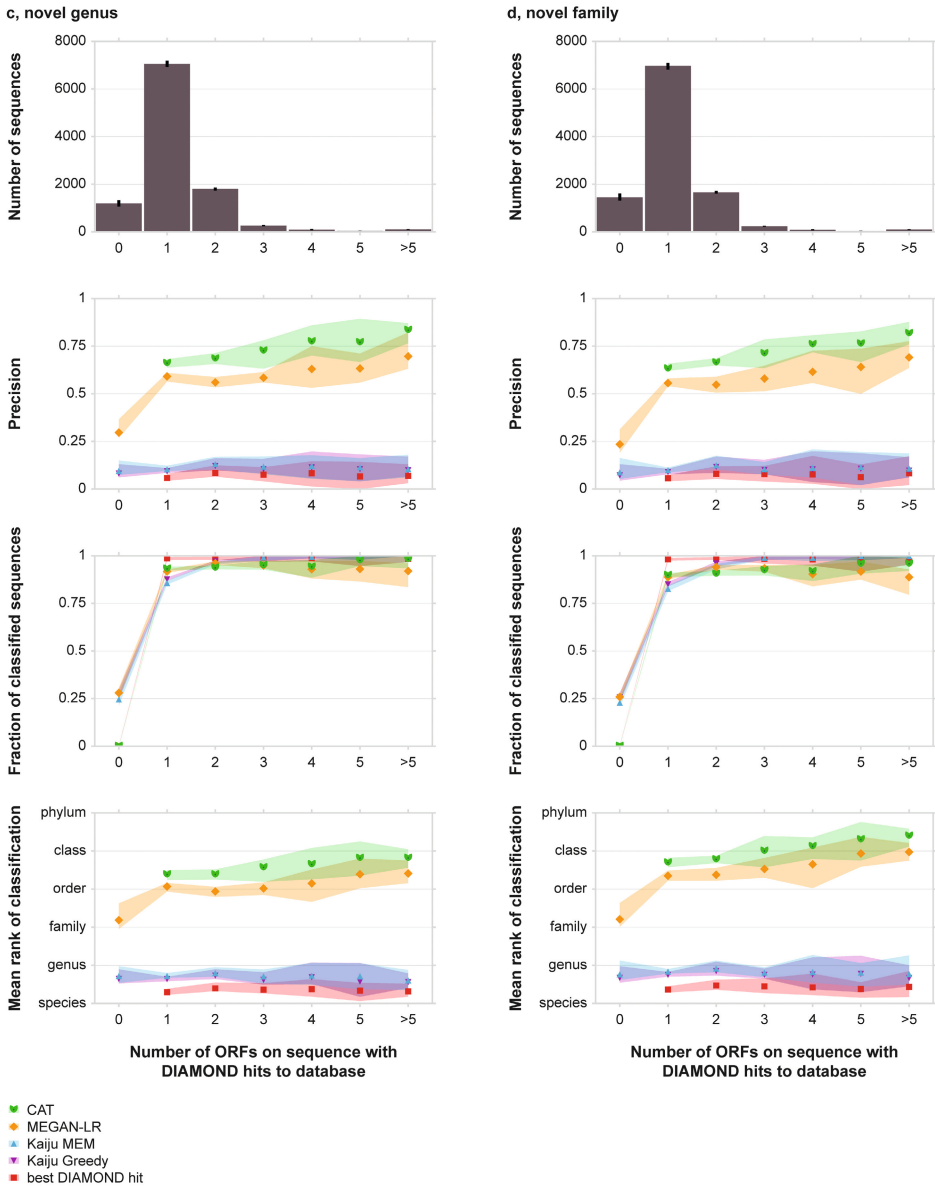
B.E.D. conceived the study. F.A.B.v.M., D.D.C., and K.A. wrote the code. F.A.B.v.M. and F.H.C. analysed the data. F.A.B.v.M. and B.E.D. wrote the paper. All authors read and approved the final manuscript.

Supplementary Information

Supplementary Figures

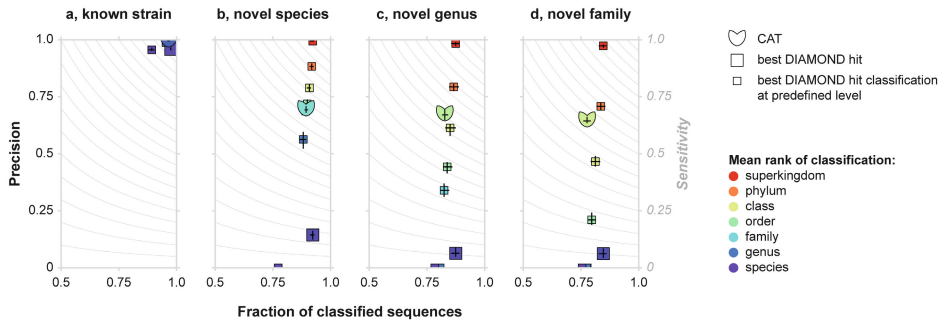


Supplementary Fig. 1 | Caption on next page.



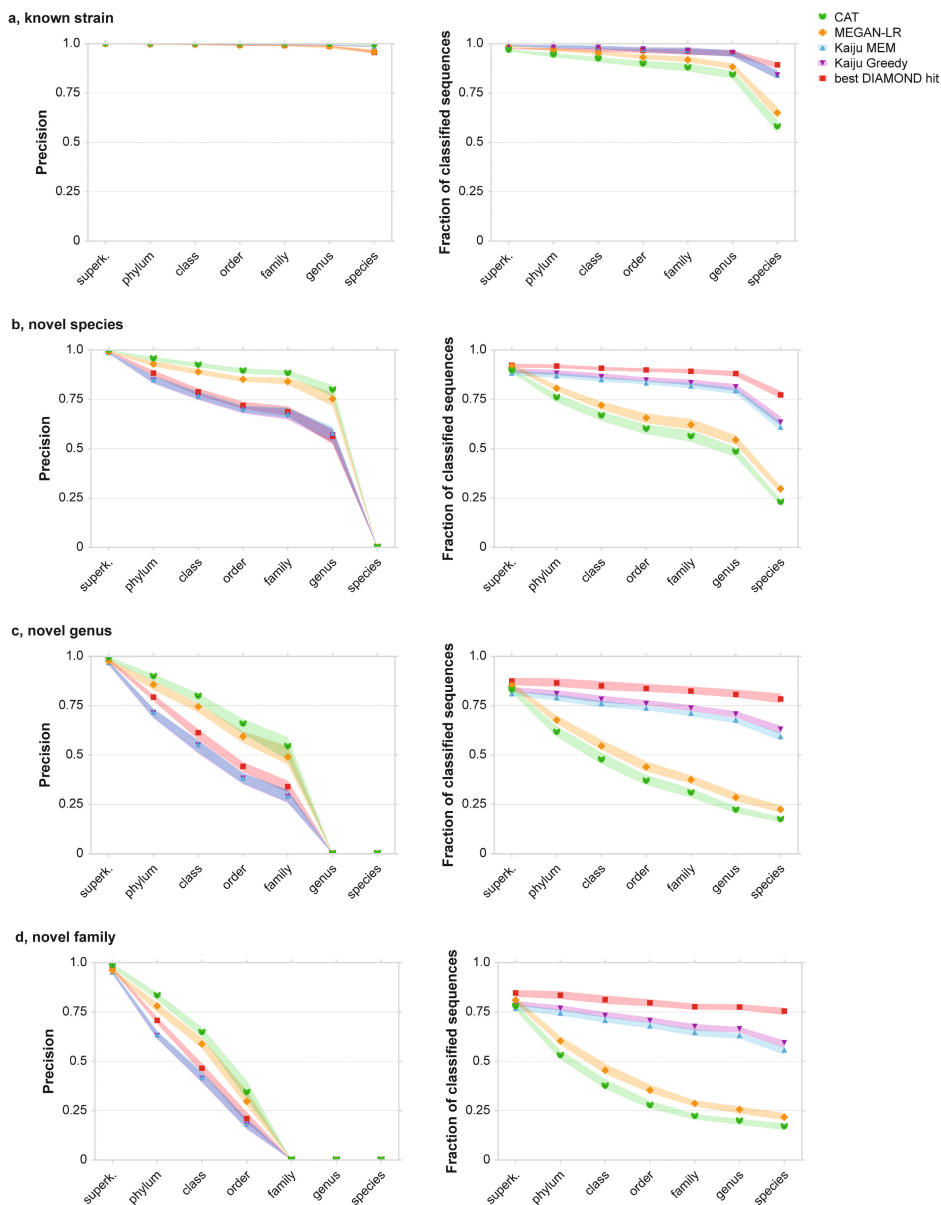
Supplementary Fig. 1 | Classification performance of CAT, LAST+MEGAN-LR, Kaiju, and DIAMOND best-hit with sequences binned according to the number of ORFs on the sequence with DIAMOND hits to the database. a, Classification of known sequences, b–d, classification of simulated novel taxa for different levels of divergence from reference databases. Sequences can fall in the 0 bin for three reasons: no ORFs are recognised on the sequence, ORFs are predicted but they do not have any hits to the database, or the ORF does have hits but its accession number cannot be found in the NCBI taxonomy files. In those cases where 1 or more ORFs on a sequence have DIAMOND hits, the fraction of classified sequences by CAT and DIAMOND best-hit can only be lower than 1 when some classifications are made at trivially informative taxonomic ranks such as “cellular organisms” or “root”. Black error bars (in top figures) and shaded areas indicate maximum and minimum values out of ten benchmarking datasets.

Contig Annotation Tool and Bin Annotation Tool



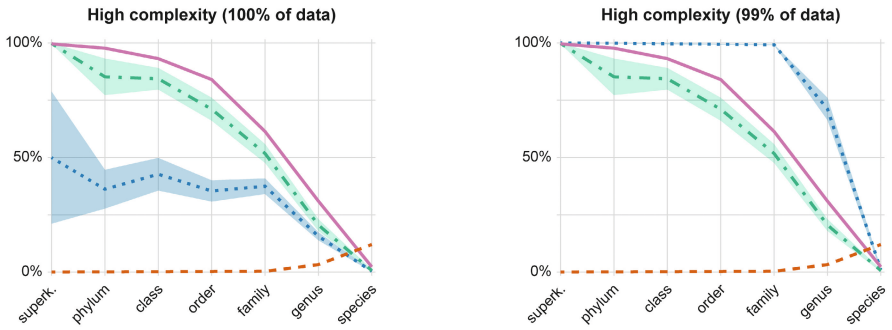
Supplementary Fig. 2 | Classification performance of CAT, DIAMOND best-hit, and DIAMOND best-hit with different taxonomic rank cut-offs. **a**, Classification of known sequences, **b–d**, classification of simulated novel taxa for different levels of divergence from reference databases. Black bars indicate maximum and minimum values out of ten benchmarking datasets, bars cross at the means. Colour coding indicates the mean taxonomic rank of classification averaged across the then benchmarking datasets (minimum and maximum values not shown for brevity).

Chapter 3

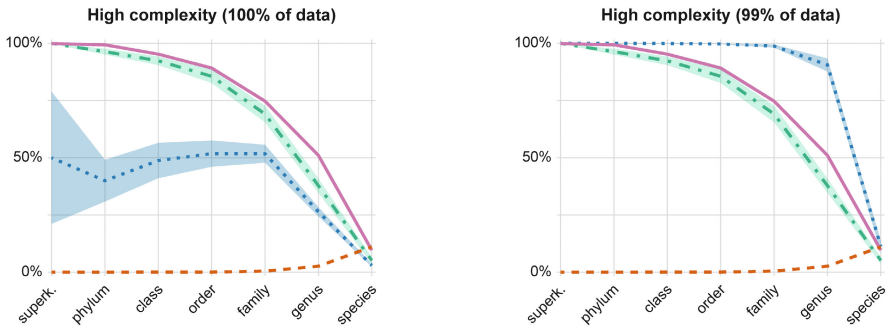


Supplementary Fig. 3 | Classification performance of CAT, LAST+MEGAN-LR, Kaiju, and DIAMOND best-hit for different levels of unknownness across taxonomic ranks. a, Classification of known sequences, **b–d,** classification of simulated novel taxa for different levels of divergence from reference databases. Shaded areas show maximum and minimum values across the ten benchmarking datasets.

a, RefSeq as reference database



b, nr as reference database

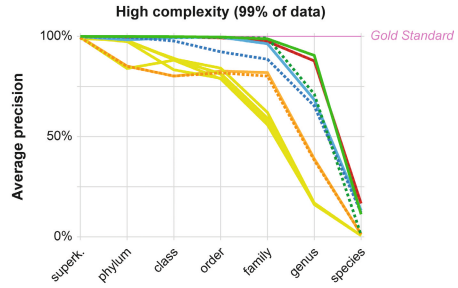
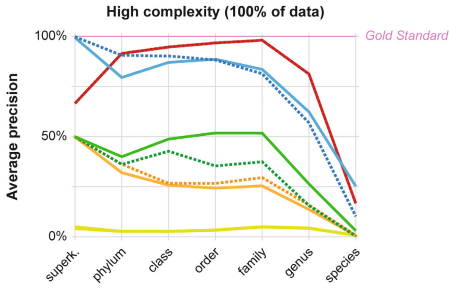


Metric:
 — accuracy
 - - misclassification
 ··· average precision
 - · - average recall

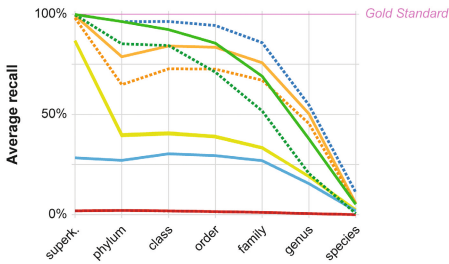
Supplementary Fig. 4 | Classification performance of CAT on the CAMI high-complexity gold standard assembly, with a, RefSeq as reference database, and b, nr as a reference database. Plots on the left show average precision if all taxonomic annotations are included, plots on the right if shortest classifications representing less than 1% of the total assembly length are excluded. Accuracy, misclassification, and average recall are the same in both plots. Note that measures are the same as those calculated in the CAMI challenge, and thus precision here reflects something different from the precision we used earlier in the clade exclusion experiments. Shaded areas show the Standard Error of the Mean (SEM) for average precision and average recall. **a** can be compared to Supplementary Figure 18 and **b** to Supplementary Figure 19 in the CAMI paper (212).

Chapter 3

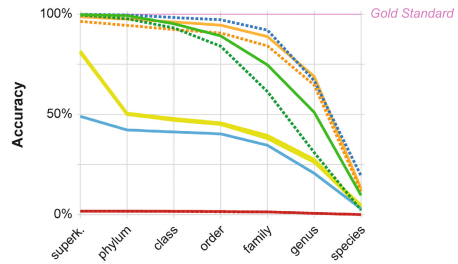
a



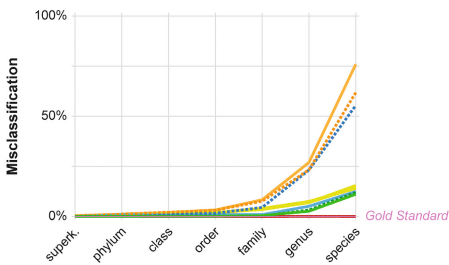
b



c

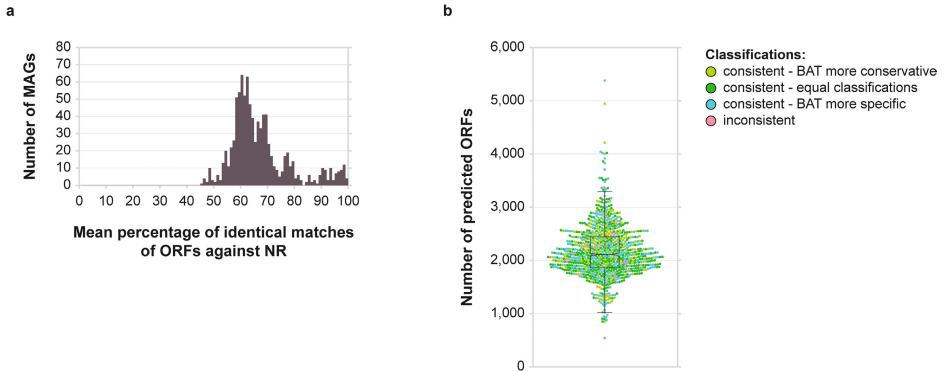


d

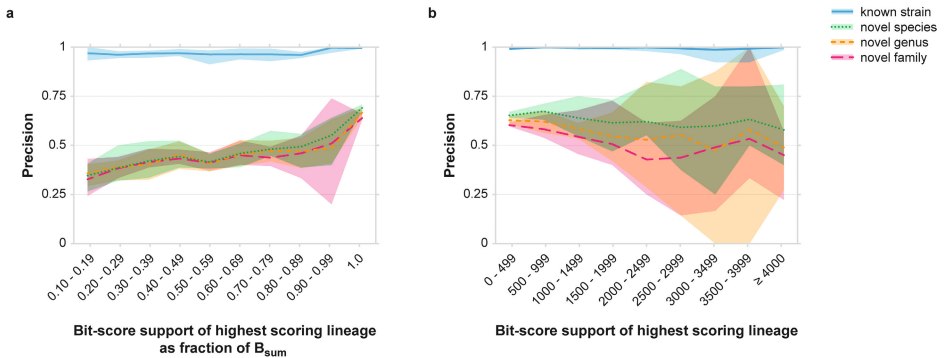


- Taxonomic binner:**
- CAT (nr)
 - CAT (RefSeq)
 - taxator-tk 1.4pre1e
 - taxator-tk 1.3.0e
 - PhylPythiaS+ mg c400
 - PhylPythiaS+ c400
 - Kraken 0.10.6-unreleased
 - Kraken 0.10.5
 - MEGAN 6.4.9 (5x)

Supplementary Fig. 5 | Comparison of CAT with RefSeq and nr as reference database against the taxonomic classifiers tested in the CAMI challenge, for a, average precision, b, average recall, c, accuracy, and d, misclassification. The left plot in a shows average precision if all taxonomic annotations are included, the right plot if shortest classifications representing less than 1% of the total assembly length are excluded.

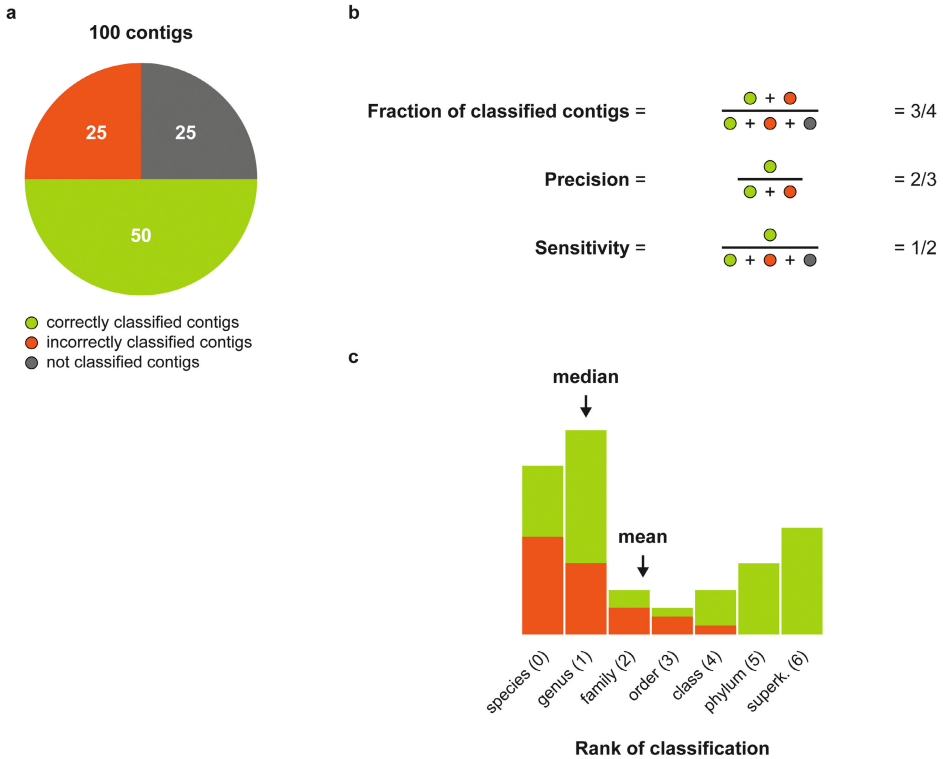


Supplementary Fig. 6 | Predicted ORFs on 913 MAGs. **a**, Average percentage of identical matches with the best DIAMOND hit in the nr database for all predicted ORFs in a MAG. The wide distribution shows that the MAGs represent a wide range of novelty, i.e. most MAGs are organisms that are not present in the nr database yet. **b**, Swarmplot showing the number of predicted ORFs per MAG. MAGs are coloured as in Fig. 5 ($r = 5, f = 0.3$).



Supplementary Fig. 7 | Classification performance of CAT, binned per a, fraction of summed bit-score support and b, total bit-score support. CAT was run with $f = 0.1$ and only the values of the lowest classification of the chosen lineage are shown. If multiple lineages have a score higher than 0.1, the majority classification is chosen (i.e. the sequence is classified as if $f = 0.5$). Shaded areas indicate maximum and minimum values out of ten benchmarking datasets. Since only values of the lowest classification are shown per lineage, and taxon classifications higher up the lineage have a higher chance of being correct and also higher support values, the relation between fraction of summed bit-score support and precision is even more pronounced for all taxa in a lineage.

Chapter 3



Supplementary Fig. 8 | Measuring performance for contig classification. **a**, Example contig set. Classifications above superkingdom rank (e.g. “cellular organisms” or “root”) are considered not classified. Half of the total classifications is contained within the true taxonomic lineage and is thus scored as correct, and a quarter is not. If a classification is in the correct lineage but too specific, it is considered incorrect. **b**, Measures of performance. Precision is a measure for how trustworthy a classification is, sensitivity for how much of the total data is correctly classified. Sensitivity is fraction of classified contigs \times precision. **c**, Mean and median taxonomic rank of classification are calculated for all classified contigs (75 in the example), where the ranks species-phyllum are given the integer values 0–6, respectively, allowing a mean to be calculated.

Supplementary Tables

Supplementary Tables 1–6 (captions below) are available from Zenodo at <https://doi.org/10.5281/zenodo.8090260>.

Supplementary Table 1 | Performance measure results for the tested taxonomic classifiers on ten benchmarking contig sets. For CAT the entire parameter sweep is included.

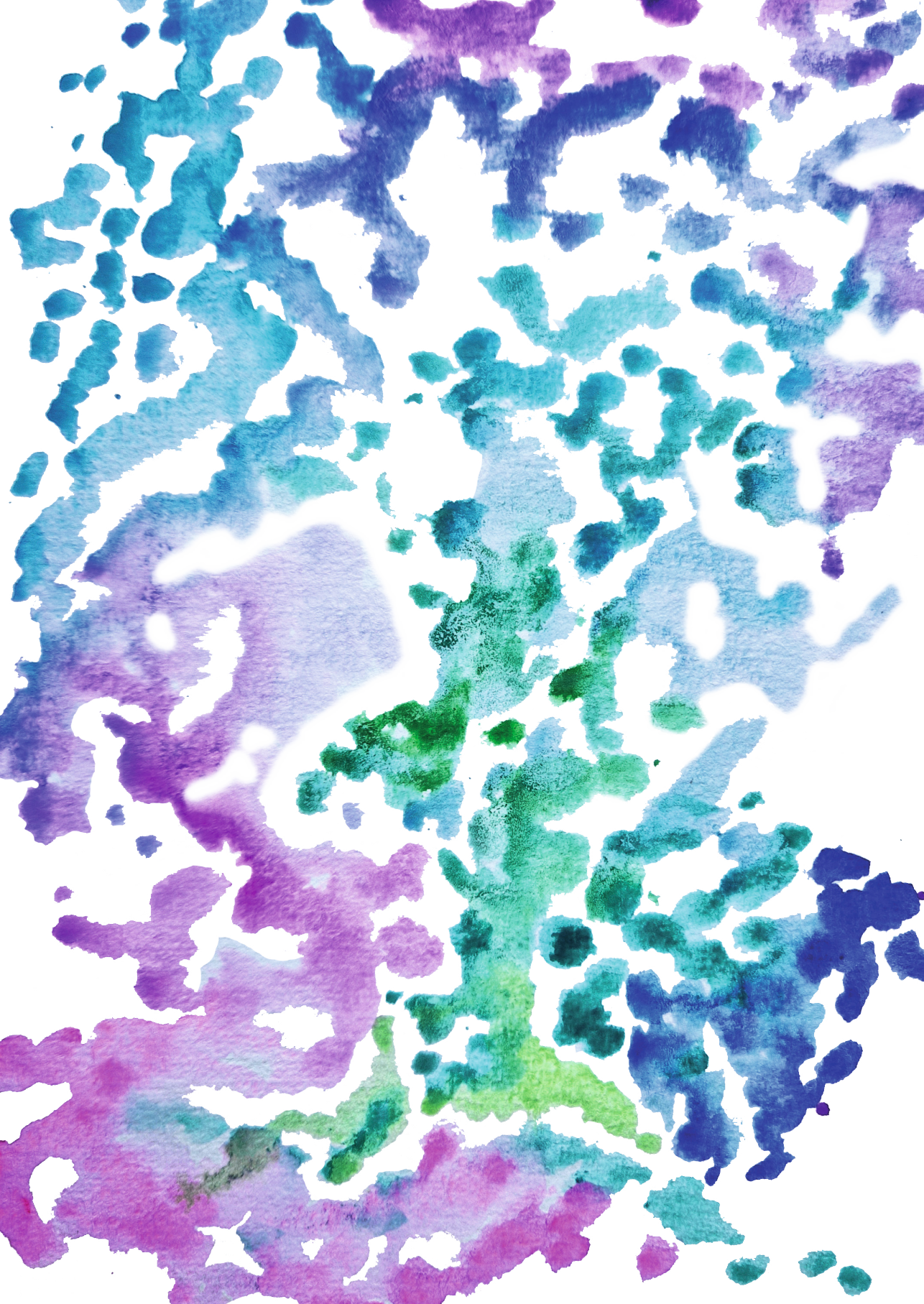
Supplementary Table 2 | Performance measure results per rank for the tested taxonomic classifiers on ten benchmarking contig sets. Only CAT results with default parameter settings are shown.

Supplementary Table 3 | Performance measure results of CAT runs with default parameter settings ($r = 10, f = 0.5$) on the CAMI high-complexity gold standard assembly. Note that measures are the same as those calculated in the CAMI challenge, and thus precision here reflects something different from the precision we used in the clade exclusion experiments.

Supplementary Table 4 | Run time and peak memory usage for CAT, MEGAN-LR, Kaiju, and recent versions of the CAMI tools on the CAMI high-complexity dataset. The NCBI RefSeq database was used as a reference, either on the DNA or protein level. CAT was run with default parameter settings, and with parameter settings that reduce run time of the DIAMOND alignment step substantially. The CAMI tools were run as suggested in their respective manuals, and/or as done in the CAMI challenge. As in the CAMI challenge, MEGAN was run on metagenomic reads and not on the gold standard assembly. We ran MEGAN on a single read file out of 5 in the challenge. Run time and peak memory usage are given for the discrete steps of each pipeline, colour fills indicate total and maximum for the entire pipeline, respectively.

Supplementary Table 5 | MAG classification by BAT, lastTaxa, and GTDB-Tk. Published classifications of the MAGs were manually adjusted to match our taxonomy files. BAT classifications are shown for $r = 5, f = 0.1$. With these parameter settings a MAG can have multiple classifications. The single BAT majority classification is given as well (i.e. the classifications of $f = 0.5$). Numbers in parentheses are the fraction of summed bit-score that supports the taxon. For lastTaxa the best guess classification is shown. The percentage of proteins with database hits assigned to the taxon is shown in parentheses when they were given. GTDB-Tk classifications are based on release 86 of the GTDB-Tk reference database, which was constructed after publication of the queried MAGs.

Supplementary Table 6 | Ten benchmarking contig sets were generated from genomes deposited in bacterial RefSeq. Lengths were based on the length distribution of eight assembled real metagenomes deposited in SRA (libraries SRR2922420, ERR1198954, ERR315808, ERR315819, SRR3666246, ERR594326, ERR599045, SRR3732372). Reads were quality filtered with BBduk (BBTools v36.64), and assembled with metaSPAdes v3.10.1. Contigs had a minimum length of 300 nucleotides. RefSeq id, length, start and stop coordinate in the genome, and taxonomic classifications of contigs are shown. Datasets are available from the authors upon request.





Chapter 4

Integration of taxonomic signals from MAGs and contigs improves read annotation and taxonomic profiling of metagenomes

Ernestina Hauptfeld, Nikolaos Pappas, Sandra van Iwaarden, Basten L. Snoek, Andrea Aldas-Vargas, Bas E. Dutilh, and F. A. Bastiaan von Meijenfildt

Under review

Abstract

Metagenomic analysis typically includes read-based taxonomic profiling, assembly, and binning of metagenome-assembled genomes (MAGs). Here we integrate these steps in Read Annotation Tool (RAT), which uses robust taxonomic signals from MAGs and contigs to enhance read annotation. RAT reconstructs taxonomic profiles with high precision and sensitivity, outperforming other state-of-the-art tools. In high-diversity groundwater samples, RAT annotates a large fraction of the metagenomic reads, calling novel taxa at the appropriate, sometimes high taxonomic ranks. Thus, RAT integrative profiling provides an accurate and comprehensive view of the microbiome. CAT/BAT/RAT is available at <https://github.com/MGXlab/CAT>. The CAT pack now also supports GTDB annotations.

Background

Metagenomic shotgun sequencing provides a single platform for exploring both the composition and the functional potential of diverse microbial communities (16,114,133,144,347). While functional profiling maximizes the usage of the shotgun data, taxonomic profiling of metagenomes may involve mapping reads to a reference database containing specific marker genes (152,348–351), in which case only a portion of the data is used and function can only be coupled to taxonomy for those reads that contain the marker gene. Alternatively, taxonomy can be assigned to as much of the data as possible by querying reads to a full reference database (328,352–354). Metagenomic profilers carry out direct homology searches in DNA (353), protein (328), or *k*-mer space (329,352,355), and the resulting taxonomic profiles have been used in large scale studies to characterize microbial communities of the oceans (144), the global topsoil (114), and to estimate the niche range of every known microbial taxon (356).

Taxonomic profiles that are based on direct queries of individual reads to full reference databases give a comprehensive view of a microbiome, but often contain spurious annotations. Assigning taxonomy based on homology searches is challenging, particularly for relatively short reads: (i) some genomic regions are highly conserved across taxa, making it difficult to discriminate between them; (ii) microbes have high rates of horizontal gene transfer (27,357), so the best hit in the reference database might be an unrelated taxon; (iii) environmental microbiomes may contain many novel taxa without close representatives in the reference database, resulting in possible annotation to e.g. a genus or species when the organism only shares the same order (212,289); (iv) known taxa may contain novel genomic regions, resulting in no annotation of reads covering that region or annotation to a more distant relative, and (v) reference databases contain mis-annotated sequences (161). These challenges are especially pronounced when directly comparing individual reads. With the exception of data from recent long-read sequencing platforms (358), reads are short sequences that contain limited taxonomic information, leading to reads derived from a single strain potentially being assigned to several different taxa. Thus, while comprehensive, taxonomic profiles based on read annotations are inherently noisy with spurious annotations, and often inaccurate (212).

Over the past decade, best practices in shotgun metagenomics have been established, including reference database-independent (*de novo*) assembly (204,205) and binning of metagenome-assembled genomes (MAGs) (199,200). The resulting contiguous sequences (contigs) and especially MAGs allow for accurate detection of novel taxa. Contigs and MAGs are significantly longer than the original short sequencing reads, the additional data allowing for more reliable taxonomic annotation, either by multiple homology searches (289,359) or phylogenetic placement (360). Long

sequence length mitigates the errors in annotation discussed earlier because multiple taxonomic signals can be integrated (confidence in annotation: MAGs > contigs > reads). However, even though taxonomic annotation is more accurate for longer sequences, they often represent only part of the metagenomic data and therefore provide an incomplete picture of the microbiome (data explained: reads > contigs > MAGs). As de novo assembly and binning depends on sufficient coverage of the genome sequence, it may be expected that especially rare microorganisms will be missed when MAGs or contigs are assessed. For a robust taxonomic profile that also includes rare microorganisms, an annotation protocol that integrates both taxonomic information from long sequences where available and short reads where not may thus be desirable.

Here, we present Read Annotation Tool (RAT), an annotation pipeline for metagenomic sequencing reads that integrates accurate annotation of contigs and MAGs derived from de novo assembly and binning, and direct homology searches of the remaining unassembled reads. RAT estimates taxonomic profiles by associating reads to longer sequences when possible and assigning taxonomy according to the most reliable taxonomic signal it can find (MAGs > contigs > reads). Contigs and MAGs are taxonomically annotated with the previously published tools CAT and BAT (289), which provide robust annotation based on open reading frame (ORF) prediction and comparisons to a protein database (209,290,361). We show that, by integrating taxonomic signals from MAGs, contigs, and reads, RAT provides more accurate read annotations and taxonomic profiles than other state-of-the-art tools, and accurately characterizes groundwater microbiomes with many novel taxa.

Results and discussion

Natural microbial communities consist of many different microorganisms that can be identified and characterized by sequencing their DNA with shotgun metagenomics. To get an accurate overview of all microorganisms and their relative abundances in a sample, the most comprehensive approach is to obtain reliable taxonomic annotation for as many of the sequencing reads as possible. While contigs and MAGs can be more reliably annotated than individual reads, in most metagenomic datasets not all reads are assembled into contigs and not all contigs are binned into MAGs (**Fig. 1**). To address this trade-off between annotation accuracy and the fraction of data that can be explained in a metagenome, we developed Read Annotation Tool (RAT) (**Fig. 1b,c**).

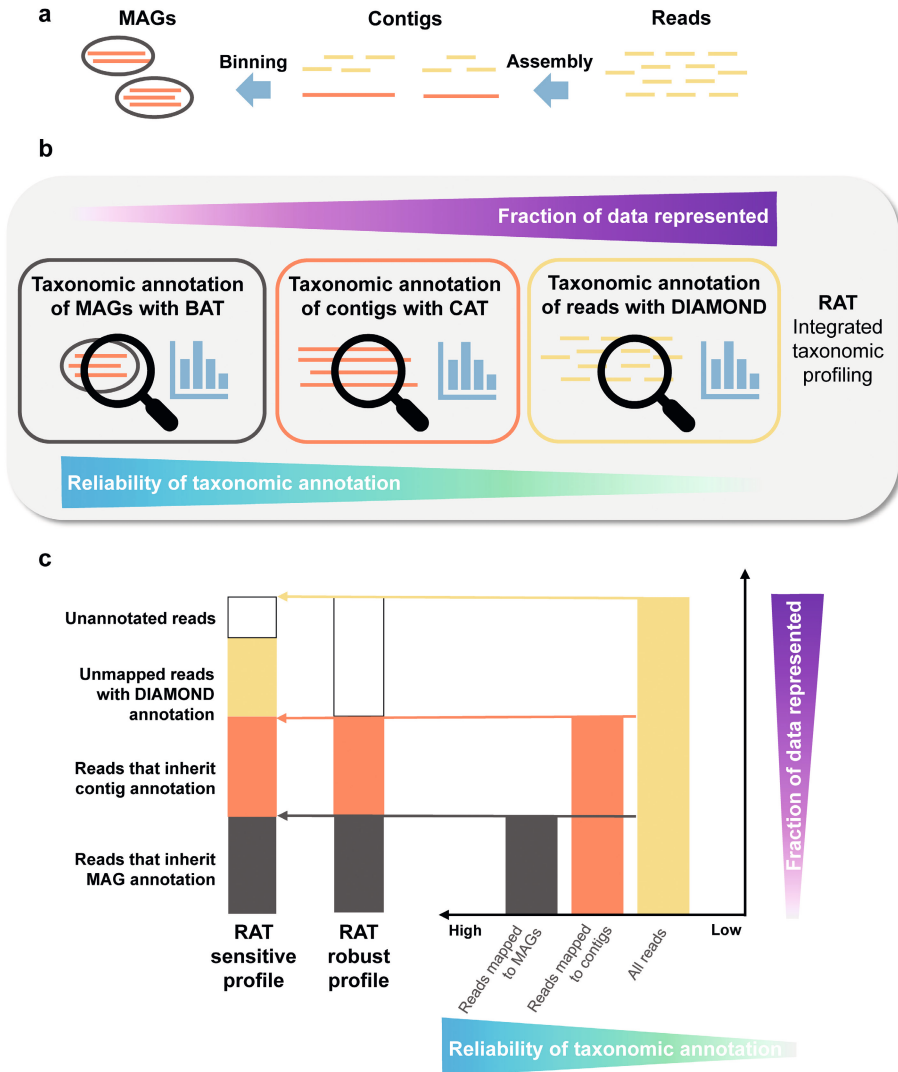


Fig. 1 | The RAT workflow. **a**, Overview of a standard state-of-the-art metagenomics pipeline. **b**, Overview of the RAT workflow: reads are mapped to contigs which are binned into MAGs or unbinned. MAGs and contigs are taxonomically annotated using BAT/CAT. Unmapped reads and unclassified contigs are annotated using DIAMOND. **c**, Left: composition of an integrated taxonomic profile as reconstructed by RAT sensitive and RAT robust. Right: schematic bar plot showing the fraction of the metagenome that can be annotated as reads, contigs, and MAGs.

RAT annotates contigs and MAGs with the previously published tools Contig Annotation Tool (CAT) and Bin Annotation Tool (BAT), respectively. CAT and BAT query predicted ORFs on these longer sequences to a protein reference database with DIAMOND blastp (209), and assign taxonomy based on the combined taxonomic signal (289). Default options for the reference database include the non-redundant protein database (nr) (156) and in the latest update the non-redundant set of proteins in the Genome Taxonomy Database (GTDB) (362), and alternatively any protein database with taxonomic annotations can be supplied by the user. Next, individual reads are mapped to the contigs and each read inherits the taxonomic annotation with the highest confidence. Finally, the remaining sequences (reads that do not map to a contig and contigs that cannot be annotated by CAT) are annotated individually by querying them to the protein database with DIAMOND blastx (209). Thus, by assigning reads to the taxonomic annotation with the highest confidence, RAT reconstructs a comprehensive taxonomic profile with high accuracy (**Fig. 1c**, **Supplementary Fig. 1**). The final step in which sequences are individually queried to the protein database is optional, and depending on whether this step is included, we distinguish two RAT modes: in ‘robust’ mode, RAT only uses the most reliable read annotations, which are based on MAGs and contigs with ORFs. In ‘sensitive’ mode, RAT also uses the read and contig annotations with DIAMOND blastx, which also include more tentative annotations while representing more of the data.

We evaluated the performance of RAT for read annotation, and how well the final taxonomic profile represents the microbial community. First, to address the trade-off between the annotation accuracy and the fraction of reads that can be annotated by the different steps in RAT, we used simulated data from the Critical Assessment of Metagenome Interpretation (CAMI) challenge (213). Second, we used the same dataset to compare taxonomic profiles predicted by RAT to those predicted by other commonly used state-of-the-art profilers. Third, we assessed the performance of RAT and the other profilers on real metagenomes. To this end, we analysed samples from three groundwater monitoring wells, a relatively unexplored high-diversity environment that contains many novel taxa (363).

Including taxonomic signals from MAGs and contigs improves read annotation

To evaluate how the integration of different taxonomic signals influences the annotation of individual reads, we annotated simulated metagenomic datasets from the second CAMI challenge (213) with RAT. The CAMI challenge simulated well-characterized microbiomes of the mouse gut. The 10 datasets contained between 97–225 species, and included raw reads, gold standard assemblies (the best possible assembly of the sequencing reads in a sample), and genome sequences of these species. In our benchmarks, we used the gold standard assemblies as contig input.

We compared five different methods for read annotation: (i) we annotated all reads directly with DIAMOND blastx, without mapping them to contigs or MAGs, (ii) We ran RAT sensitive using only contig annotations, and direct read annotations for reads that do not map to contigs, (iii) We ran RAT sensitive also using MAG annotations with the genome sequences included in the dataset ('CAMI genomes') as input, (iv) We ran RAT sensitive with MAG annotations using MAGs binned by MetaBAT 2 (ref. 200), and (v) we ran RAT robust using only contig and MetaBAT 2 MAG annotations but no direct read annotations (**Fig. 2**). Results were assessed at six taxonomic ranks (phylum, class, order, family, genus, and species) and we scored whether a read was correctly or incorrectly annotated, or unclassified (**Fig. 2**).

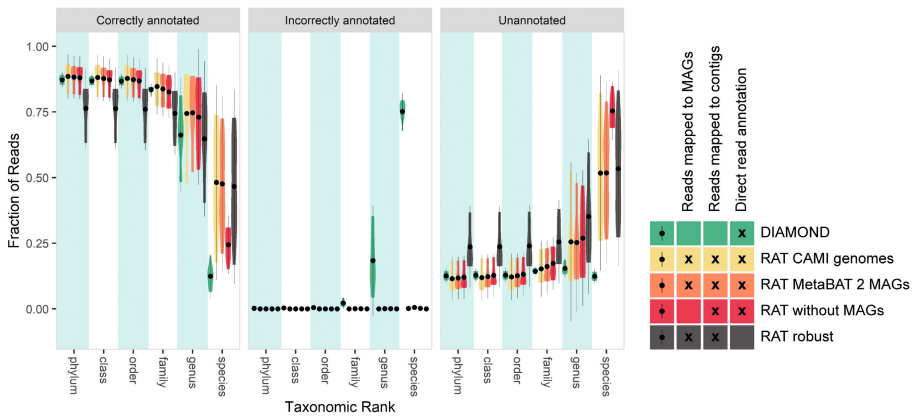


Fig. 2 | Outcome of incorporating different taxonomic signals into read annotations. ‘DIAMOND’ refers to using only direct read annotation. ‘RAT CAMI genomes’ refers to a RAT sensitive run using the genomes that were provided by the CAMI challenge as MAG input. ‘RAT MetaBAT 2 MAGs’ refers to a RAT sensitive run with contigs binned by MetaBAT 2. ‘RAT without MAGs’ refers to a RAT sensitive run without MAG input. ‘RAT robust’ refers to a RAT robust run, using only read annotation via mapping to MetaBAT 2 MAGs and contigs, but no direct read annotation.

Direct annotation with DIAMOND blastx resulted in low accuracy at low taxonomic ranks with a high fraction of mis-annotated reads (**Fig. 2**), revealing spurious annotations when mapping short sequences to a reference database. Accuracy is particularly low on species rank, where only $12.4 \pm 4\%$ (mean \pm standard deviation) of the reads were correctly annotated by DIAMOND. Despite using DIAMOND with the same reference database, RAT runs reduced mis-annotations and improved the fraction of correctly annotated reads at all taxonomic ranks, highlighting the value of integrating information from taxonomically annotated MAGs and contigs (**Fig. 2**).

When only taxonomic signals from contigs are integrated, the fraction of correctly annotated and unclassified reads increases compared to direct annotation with DIAMOND blastx, while the fraction of incorrectly annotated reads drops to

0.1–1%. This indicates that many previously mis- or unannotated reads are correctly annotated if they map to contigs. The fact that many of the reads that were mis-annotated using DIAMOND blastx are unclassified in RAT sensitive shows that these reads are mapped to contigs that cannot be annotated on lower taxonomic ranks by CAT. This indicates that the contigs have hits to multiple different taxa in the database, in which case CAT chooses a higher rank taxon as the most robust taxonomic annotation (289), precluding annotation on lower taxonomic ranks.

When taxonomic signals from both contigs and MAGs are integrated, the fraction of correctly annotated reads increases while the fraction of unclassified reads decreases compared to annotating without MAGs. In the CAMI mouse gut dataset, using the CAMI genomes as MAG input and binning the contigs with MetaBAT 2 gave very similar results, indicating that current binning tools accurately group contigs from the same species together. Without using DIAMOND blastx to annotate the remaining unmapped reads and unclassified contigs (RAT robust), the fraction of annotated reads drops, while the fraction of unclassified reads increases. This effect is likely more pronounced in some real biological datasets, where higher taxon diversity makes it more difficult to assemble reads into contigs than in the simulated CAMI samples (for which a gold standard assembly is supplied), and which thus contain more unmapped reads (see below, **Supplementary Figs. 3 and 4**).

Concluding, using the taxonomic signals from contigs and MAGs for read annotation leads to more reliable annotations than using direct querying of individual reads.

Including information from contigs and MAGs improves accuracy of taxonomic profiling

Metagenomics is used to analyse high complexity microbial communities, including many different taxa with orders of magnitude of difference in their abundances. Taxonomic profilers aim to chart the community composition by listing all taxa in a sample and estimate their relative abundance. A good taxonomic profile contains as many members of the microbial community as possible, while avoiding taxa that are not present in the sample. In practice, this often leads to a compromise between sensitivity (finding all taxa that are present and maybe some false positives) and precision (avoiding taxa that are not present and maybe some false negatives). To assess how including contig and MAG annotations affects the reconstruction of taxonomic profiles, we used four metrics (sensitivity, precision, L1 distance, and weighted UniFrac distance) to compare the CAMI ground truth taxonomic profiles to those reconstructed by RAT and four state-of-the-art taxonomic profilers that carry out annotations by direct read mapping (**Fig. 3**). Centrifuge classifies microbial DNA sequences, Kaiju annotates sequences in protein space, Kraken2 annotates DNA sequences using exact k -mer matches, and Bracken uses Kraken2 annotations for a Bayesian reestimation of the abundances of taxa in the sample. As direct read

annotations can be inaccurate, we limited the amount of noise by only counting taxa that were detected in the profile with a minimum relative abundance of 0.001%.

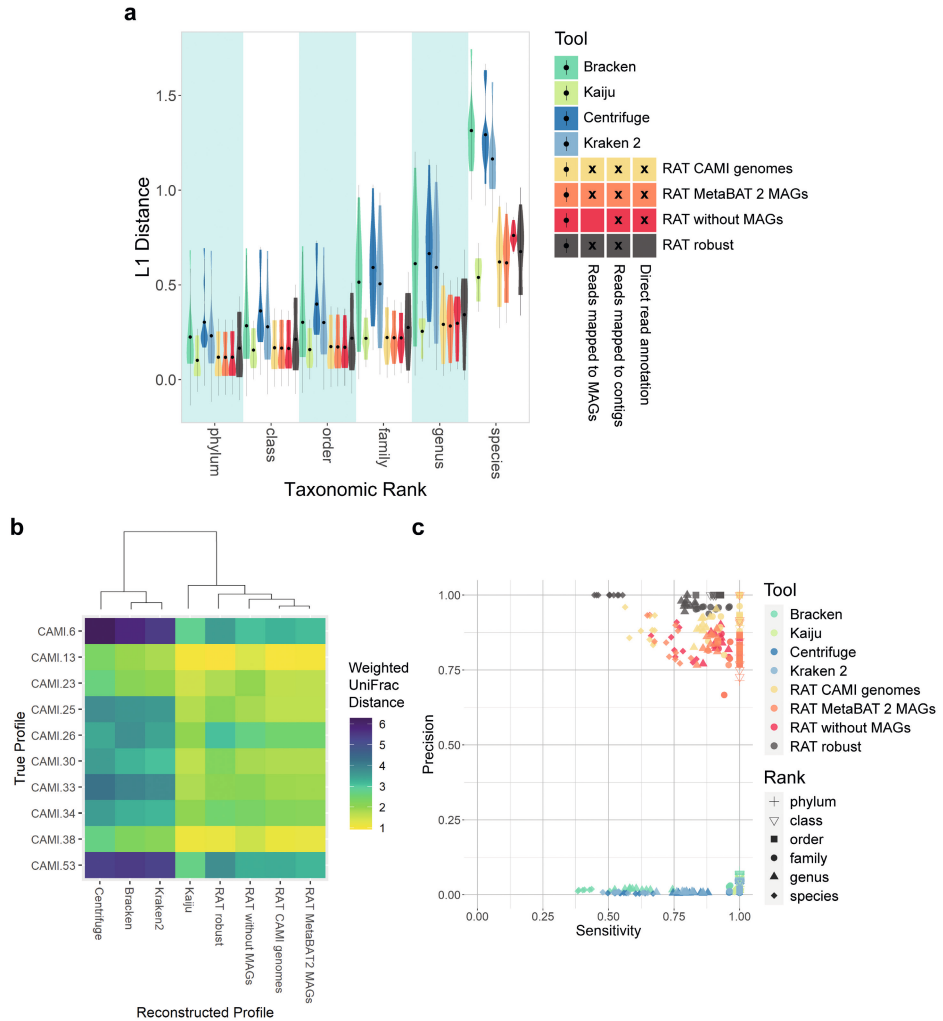


Fig. 3 | Similarities between true profiles and profiles reconstructed by different tools of the CAMI mouse gut dataset. We only counted taxa as detected if their relative abundance was at least 0.001%. **a**, L1 values between profiles reconstructed by RAT/other tools and the true profiles. An L1 value of 0 means that two profiles are identical (thus lower is better). **b**, Heatmap of weighted UniFrac distances between reconstructed and true profiles (a shorter distance is better). **c**, Sensitivity vs. precision of the different tools. Different shapes signify the sensitivity/precision on different taxonomic ranks, different colours indicate tools (high precision + sensitivity is better).

In line with our first benchmark, incorporating taxonomic signals from MAGs led to more accurate profiles than using only taxonomic signals from contigs, as seen in the L1 distance and in the weighted UniFrac distance. RAT sensitive slightly

outperformed RAT robust (**Fig. 3**) in L1 distance and sensitivity, indicating that including direct read annotation leads to reconstructed profiles that are more similar to the ground truth profile than when the step is not included. Taxonomic profiles reconstructed by RAT consistently had lower L1 distances to the ground truth profiles than profiles reconstructed by Bracken, Centrifuge, and Kraken2 (**Fig. 3a**). In comparison to taxonomic profiles reconstructed by Kaiju, RAT runs had slightly higher L1 distances on genus and species rank. Taxonomic profiles reconstructed by RAT had lower weighted UniFrac distances to the ground truth profiles than Bracken, Centrifuge, and Kraken2 (**Fig. 3b**) while Kaiju performed similarly.

RAT had a higher precision on all taxonomic ranks than the other evaluated tools. This means that RAT had fewer falsely detected taxa, in line with earlier observations of high precision of CAT and BAT annotations (289). RAT robust maintained >0.94 precision on all taxonomic ranks, even when detected taxa were not limited by a minimum relative abundance cut-off (**Supplementary Fig. 2**). Thus, like CAT and BAT on which its annotations are based, RAT robust tends to avoid spurious annotations at deeper taxonomic ranks in cases where conflicting taxonomic signals arise. For RAT sensitive, precision remained higher than that of the other evaluated tools across taxonomic ranks, albeit was lower than that of RAT robust. The minimum relative abundance cut-off greatly improved precision of RAT sensitive (cf. **Fig 3c** and **Supplementary Fig. 2**). Spurious annotations are introduced wherever short sequencing reads are directly annotated, in the direct annotation step of RAT and in the other evaluated tools. However, because of the prioritization of taxonomic signals in RAT, a smaller fraction of total reads is annotated directly, leading to fewer spurious annotations in the first place. By setting an abundance cut-off (e.g. 0.001% of reads as in this benchmark), RAT can profit from the high sensitivity of the DIAMOND blastx step (finding taxa that might not be detected using just contig and MAG annotations) while further minimizing the number of falsely detected taxa (by excluding spurious annotations that have very low abundance).

RAT's overall high precision can be explained by the integrated taxonomic profiling approach, which improves annotations in most of the challenges discussed above. Reads that map to conserved or horizontally transferred regions, or map to novel genomic regions of a known taxon are likely to get the correct annotation with RAT, because the surrounding (regions of the) genome is considered in the annotation via the contig and/or MAG. Reads belonging to novel taxa within known clades are also more likely to get correctly annotated, as when the reads are assembled into contigs or MAGs, RAT may annotate on a higher, appropriate taxonomic rank instead of on the lower taxonomic rank of closely related taxa. The difference in precision between the different approaches shows that reads that are not annotated by being associated with a contig or MAG, are far more likely to get falsely annotated. RAT's approach

reduces the number of falsely detected taxa from 200–4,000 by the other evaluated tools to between 0 (RAT robust) and 38 (RAT sensitive).

All evaluated tools showed high sensitivity from phylum down to family rank, detecting most of the taxa that were present in the ground truth profiles (**Fig. 3c**). This is consistent with increased barriers to horizontal gene transfer at higher taxonomic ranks (364). On the genus and species ranks, RAT and Kaiju outperformed Bracken, Centrifuge, and Kraken2, while Kaiju showed higher sensitivity. RAT's high performance on the CAMI datasets is in part due to the fact that a large fraction of the reads map back to contigs ($81.6 \pm 6.6\%$) and MAGs ($75.8 \pm 6.4\%$). These numbers are often lower in metagenomic datasets from other environments (see below). This leads to most reads being annotated in the most reliable MAG and contig annotation steps and few reads being annotated directly with DIAMOND, reducing the probability of spurious annotations (**Supplementary Figs. 3 and 4**).

Usage, runtime and memory requirements

Next, we compared the runtime and memory requirement of RAT with the other tools (**Table 1**). We did not take assembly and binning steps into account, since RAT does not assemble and bin metagenomes but rather takes assembled contigs and associated MAGs (of advised medium to high quality, see below) as input from the user. Although assembly and contig binning can take hours or days to run, they are a common procedure in many metagenomics studies, as they provide valuable genomic context information to short sequencing reads with relatively little risk of generating chimeras (365).

Table 1 | Runtime and memory usage for RAT and four other tools. Tools were run on two simulated datasets (smp6: 33,098,456 reads, smp13: 33,184,772 reads) from the CAMI 2 challenge. RAT robust: annotation based only on contigs and MAGs, and RAT sensitive: annotation also based on DIAMOND annotation of unmapped reads. Kraken2 and bracken are run together. All runs were performed with 16 parallel CPUs.

		Smp6	Smp13
Runtime (in minutes)	<i>RAT robust</i>	16	16
	<i>RAT sensitive</i>	106	139
	<i>Centrifuge</i>	11	11
	<i>Kaiju</i>	16	15
	<i>Kraken2/Bracken</i>	2	2
Maximum memory usage (in GB)	<i>RAT robust</i>	43	48
	<i>RAT sensitive</i>	84.1	115.4
	<i>Centrifuge</i>	240.4	240.6
	<i>Kaiju</i>	127.4	127.4
	<i>Kraken2/Bracken</i>	55.6	55.4

Kraken2 was the fastest tool (01m49s), RAT sensitive was the slowest (02h05m10s), and all other tools including RAT robust performed the jobs in 16 minutes or less. In terms of memory usage, all tools can be run on a 256Gb server. RAT sensitive had a higher memory footprint than Kraken2, but lower than Kaiju and Centrifuge. RAT sensitive varied in RAM and runtime between the two samples because it loads different amounts of unclassified reads and contigs into memory depending on the sample.

The expanded CAT pack facilitates the detection and annotation of unknown microorganisms

The simulated data provided by the CAMI challenge differs from real biological datasets. In the CAMI data, Illumina sequencing experiments were simulated of relatively low-diverse microbiomes containing genomes of known species. Annotations are facilitated by the fact that on average >80% of the reads mapped back to a MAG or contig from a gold-standard assembly, while in biological datasets, this percentage can be much lower (**Supplementary Figs. 3 and 4**). In addition, particularly in microbiomes from under-studied environments, unknown lineages are often detected that are only distantly related to known taxa in reference databases. Awaiting taxonomic classification of these microorganisms, a higher-rank taxonomic annotation of the sequence at e.g. family or phylum rank may be appropriate in these cases.

RAT provides a framework for assessing these “unknowns”. Because reads are classified via CAT and BAT, annotations are made at the appropriate taxonomic rank. CAT and BAT assign individual ORFs to the last common ancestor of all hits that have a similar bit-score to the best hit, and annotate the contig or MAG using a bit-score based voting scheme that selects the taxon at which a certain fraction (in the RAT workflow the majority) of the ORF assignments agree (289). Novel sequences have many distinct hits and are thus only annotated at a high taxonomic rank, reflecting their unknownness. MAGs that only receive a high taxonomic rank annotation by BAT may be further investigated with phylogenomic software for strain-level resolution. Since the quality of RAT results is highly dependent on the quality of the input data, we recommend using high-quality assemblies, and only including MAGs with low contamination (e.g. <10% contamination according to CheckM (51)). Contaminated MAGs can be mis-annotated or annotated at a high trivial taxonomic rank in which case a contig annotation is more reliable. MAG completeness is less relevant for RAT, as MAGs with low completeness typically still include more than one contig from the same microorganism, creating a stronger taxonomic signal than present on the individual contigs.

To challenge RAT with real datasets, we selected relatively unexplored groundwater samples taken 12–64 m below surface level from three different monitoring wells in

a Dutch agricultural area, which we previously found had high microbial diversity and contain many novel taxa (363). We performed a metagenomic analysis including quality control, assembly (204), and binning (199,200,366), which produced 514 MAGs. We supplied the reads, 2,770,251 contigs, and 423 medium- to high-quality MAGs (completeness $\geq 50\%$, contamination $< 10\%$; see ref. 52) to RAT to reconstruct taxonomic profiles of the groundwater samples, using nr as a reference database. In addition, the medium- to high-quality MAGs were dereplicated (367), and the resulting 195 representative MAGs were placed in a phylogenetic tree showing their relationships and abundance across samples (**Supplementary Fig. 5**).

RAT annotated $22.0 \pm 8.7\%$ (mean \pm standard deviation) of reads by mapping them to MAGs, much less than in the simulated CAMI mouse gut datasets (see **Supplementary Figs. 3 and 4**), reflecting the high complexity of the groundwater samples. RAT classified $20.9 \pm 3.2\%$ of reads via unbinned contigs annotated by CAT, and $0.35 \pm 0.23\%$ via contigs annotated by DIAMOND. Finally, DIAMOND blastx annotated an additional $23.0 \pm 3.3\%$ of the reads. These unmapped reads represent sequences with low coverage that could not be assembled into contigs, and based on the results with simulated data above, we expect represent more spurious results.

The taxonomic profile reconstructed by RAT sensitive showed that most reads belonged to unclassified Bacteria, including of the phyla Chloroflexi and Deltaproteobacteria (**Fig. 4a**). Chloroflexi bacteria utilize a variety of electron acceptors including oxidized nitrogen or sulphur compounds. Comparison of the 18 reconstructed taxonomic profiles showed that Sample 23-2 contained relatively many Chloroflexi reads, while the Deltaproteobacteria were rare. Although many of the microorganisms in this sample could only be classified on high taxonomic ranks, 22 MAGs from these phyla represented 31.1% of the reads in the sample (see **Supplementary Fig. 5**).

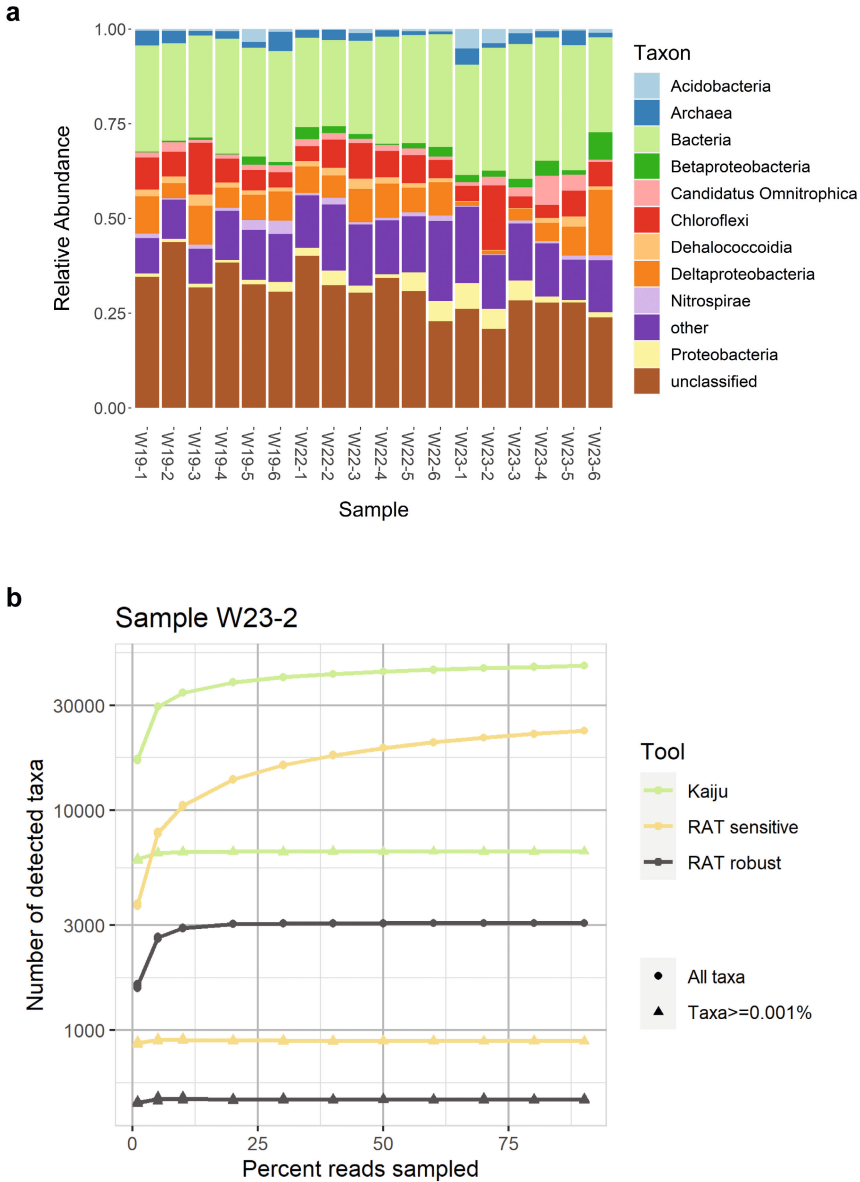


Fig. 4 | Taxonomic profiling of groundwater metagenomes. a, Microbial profiles of groundwater samples on taxonomic rank class as reconstructed by RAT sensitive. **b,** Rarefaction curves of the number of taxa detected in profiles in sample W23-2 by RAT robust, RAT sensitive, and Kaiju. Triangles indicate the number of taxa detected in profiles when a minimum abundance is required to consider an organism as detected. Circles indicate the number of taxa detected without a cut-off.

Next, we compared the taxonomic profiles of the groundwater metagenomes as predicted by RAT and Kaiju, as it was the best performing other tool in the previous benchmark. Both tools classified two thirds of the data (RAT: $68.9 \pm 5.8\%$ of reads, Kaiju: $63.8 \pm 5.6\%$ of reads, **Supplementary Table 2**). However, RAT classified these reads into roughly 20% of the taxa that Kaiju predicted (**Fig. 4b**). Bearing in mind the high precision of RAT (**Fig. 3**), we propose that the taxa predicted by RAT are a more parsimonious interpretation of the metagenomic data than those predicted by Kaiju. To visualize the potential overestimation of taxa due to spurious annotations, we made rarefaction curves for RAT sensitive, RAT robust, and Kaiju. Without a minimum relative abundance cut-off, rarefaction curves for RAT sensitive and Kaiju did not level off. This reflects the spurious annotations of individual reads and indicates that, without a cut-off, deeper sequencing of the same sample would lead to higher predicted richness. This pattern was also observed in simulated data containing a known number of 110 species (**Supplementary Fig. 6**). The rarefaction curve of RAT robust levelled off in the groundwater data, indicating robustness towards falsely detected taxa of the RAT robust workflow. With a minimum relative abundance cut-off of 0.001%, all rarefaction curves levelled off, although the different tools predicted different taxa richness. Kaiju estimated a much higher richness than RAT in both robust and sensitive mode (**Fig. 4c**). Combined with the RAT results on simulated data where RAT robust underestimated richness while RAT sensitive included some false positives (**Supplementary Fig. 6**), this shows that: (i) RAT robust is the best-suited RAT workflow in experiments where reliability is crucial, but it will likely not detect all taxa present while (ii) RAT sensitive will detect more taxa at the risk of including a few of them spuriously.

RAT estimates sequence abundance as opposed to taxonomic abundance (58). This means that RAT reports the abundance of a taxon as fraction of total DNA in the sample, rather than as genome copies which can for example be estimated by querying marker genes (348–350). The resulting relative abundance profile is skewed towards microorganisms with larger genomes, since they provide more DNA to the sequencing machine and thus contribute more reads than organisms with small genomes. To convert sequence abundance to genome copies, relative abundances have to be normalized by genome length, which is often unknown and can vary widely even between strains of the same species (368). For novel microorganisms, genome sizes of closely related species might not be available. For these reasons, RAT by default does not convert sequence abundance to taxonomic abundance. However, the CAT pack provides a table with weighted mean genome sizes for most known bacteria and archaea at all taxonomic ranks based on genomes deposited in the BV-BRC database (369) which allows users to do this conversion if they wish.

Conclusion

In this study, we presented the Read Annotation Tool (RAT), a new tool to strengthen the CAT pack metagenome analysis suite. We showed how annotating each read by using the best available taxonomic information (based on MAGs, contigs, or direct read mapping) leads to fewer falsely detected taxa and improves accuracy of taxonomic profiles. RAT is flexible to future improvements in sequencing technologies, as well as in assembly and binning software, as they are run by the user before the mapping and classification steps. RAT will be useful in the exploration and understanding of metagenomic datasets by robust classification of a majority of sequencing reads, even in unexplored environments that are rich in novel microorganisms.

Methods

RAT workflow

Read Annotation Tool (RAT) provides individual metagenomic sequencing reads with the most reliable taxonomic annotations, and uses these results to reconstruct an accurate taxonomic profile of the microbiome. RAT requires an input of sequencing reads, de novo assembled contigs or scaffolds, and optionally affiliated MAGs. We advise to filter the MAGs based on quality and only supply MAGs that have low contamination (<10%). Completeness of MAGs is less critical, as multiple contigs of the same organism carry a stronger taxonomic signal than individual contigs even if a part of the genome is not binned. These different DNA sequences are queried against a protein database for taxonomic annotations. Next, taxonomic annotations of individual reads are based on the associated data type with the highest confidence of annotation (MAGs > unbinned contigs > unassembled reads). RAT can be run in two different modes: sensitive (complete workflow, see below), and robust (skips step 3, only evidence from MAGs or contigs is used). The complete workflow of RAT consists of five steps:

1. RAT maps the reads back to the assembled contigs using BWA-MEM (275). Reads mapping to each contig are extracted with SAMtools (278), including only primary mappings and excluding low-quality primary mappings (default: Phred quality score of 2, which can be changed by the user). In case of multiple mappings with equal Phred scores, one of the mappings is assigned at random.
2. RAT performs taxonomic annotation of the contigs and MAGs with the previously published tools CAT and BAT (289), respectively. CAT and BAT annotate contigs and MAGs by predicting open reading frames (ORFs) with Prodigal (290) and comparing these with DIAMOND blastp to the non-redundant protein database of NCBI (nr) (156) or the non-redundant set of proteins in GTDB (362), both of which can be downloaded and prepared by running ‘CAT download’ and ‘CAT prepare’. MAGs consist of binned contigs and therefore a contig in a MAG gets assigned both a BAT and a CAT annotation that may not be identical. As a MAG contains more taxonomic signals than a contig, RAT will prioritize the MAG annotation. In most metagenomic datasets, not all contigs are binned, and not all contigs can be annotated with CAT (289). By default, RAT runs CAT with standard settings, and BAT with an *f* parameter value of 0.5 to prevent multiple annotations per MAG (see ref. 289 for details). Currently, *f* values < 0.5 are not supported by RAT.
3. Contigs that are not classified and reads that could not be mapped to any contig in step 1 are now classified simultaneously by comparing them to the protein database using DIAMOND blastx (209), and assigning the taxon of the last common ancestor of the organisms found within a certain range of the top hit

- (default: hits within 10% of the top-hit bit-score, which can be changed by the user), similar to the *r* parameter in CAT (289). Thus, these direct mappings do not involve ORF predictions as in step 2.
- Each individual read is classified according to the taxonomic signal with the highest confidence, in the following order: (i) If the read is mapped to a contig that is binned, the MAG annotation is assigned to it. (ii) If the read is mapped to an unbinned contig, the contig annotation is assigned to it. (iii) If a read is mapped to an unbinned contig that could not be annotated with CAT, or not mapped at all, the direct annotation is assigned to it (see step 3). (iv) Reads that do not have any taxonomic annotation are binned in an ‘unclassified’ category.
 - RAT calculates the abundance of a taxon by summing the total number of reads assigned to it, and normalizes abundances by dividing by the total number of sequenced reads in the sample. This final table constitutes the taxonomic profile. The relative abundances are sequence abundance (fraction of sequenced DNA), as opposed to taxonomic abundance (genome copies) (58). A user may convert fraction of sequenced DNA to an estimate of genome copies by normalizing by genome size. The CAT pack provides a table with weighted mean genome sizes for most known bacteria and archaea at all taxonomic ranks based on a genomes deposited in the BV-BRC (previously PATRIC) database (369) which allows a user to do this conversion.

RAT is written in Python 3.8.3 and available on GitHub at: <https://github.com/MGXlab/CAT>. We have tested RAT in the following configuration: BWA v0.7.17-r1188, SAMtools v1.10, Prodigal v2.6.3, DIAMOND v2.0.5.

Benchmarking on simulated datasets

To evaluate RAT’s performance as read classifier and taxonomic profiler, we used datasets generated for the second Critical Assessment of Metagenome Interpretation (CAMI) challenge (213). We used 10 randomly selected samples of the mouse gut benchmark dataset (samples 6, 13, 23, 25, 26, 30, 33, 34, 38, and 53), which contain between 97–225 bacterial species each. We taxonomically annotated the reads with RAT and four other commonly used profilers: Bracken, Centrifuge, Kraken2, and Kaiju. All tools included in this benchmark also report relative abundance as sequence abundance and a comparison to RAT is thus fair (58).

For each read, we assessed at six taxonomic ranks (phylum, class, order, family, genus, and species) whether it was correctly or incorrectly annotated, or unclassified. To evaluate the taxonomic profiles, we used the same measures used in the original CAMI challenge (212): the L1 and weighted UniFrac distances between the true and inferred profiles, as well as the precision and sensitivity of detected taxa. We only counted taxa as detected if they had been assigned at least 0.001% of the reads in the taxonomic profile and applied the same cut-off for all tools. L1 and weighted

UniFrac are pairwise similarity measures between taxonomic profiles. L1 ranges from 0 (profiles are identical) to 2 (profiles do not share any taxa) according to the formula:

$$L1 = \sum_{i=1}^n |p1_i - p2_i|$$

where i is the i th out of n total taxa in the union of the two profiles, and $p1_i$ and $p2_i$ are its relative abundances in the profiles that are being compared (212). L1 is calculated at each taxonomic rank, contrary to weighted UniFrac distance. The weighted UniFrac distance incorporates both the relative phylogenetic or taxonomic relatedness between taxa and their abundance. We calculated weighted UniFrac distances using EMDUniFrac (370) using the taxonomy as measure for relatedness with a distance of 1 between taxonomic ranks. Precision and sensitivity are defined as in ref. 212 and only depend on the binary detection of each organism and not on their abundance. They were calculated using the following formulas:

$$precision = \frac{TP}{TP + FP}$$

$$sensitivity = \frac{TP}{TP + FN}$$

where TP (true positives) is the number of taxa that are correctly detected, FP (false positives) is the number of taxa that are incorrectly detected, and FN (false negatives) is the number of taxa that are not detected but are present in the dataset and thus should have been detected.

We ran Bracken v2.6.1 (ref. 371), Kraken2 v2.1.2 (ref. 352), Centrifuge v1.0.4 (ref. 353), and Kaiju v1.8.2 (ref. 328) using default settings. We ran all tools using the nr/nt database from the 8th of January 2019 that were provided with the CAMI challenge.

Rarefaction curves

Rarefaction curves were calculated for RAT sensitive, RAT robust, and Kaiju results. We randomly sampled 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100% of all reads ten times and counted the number of taxa detected in these subsets.

Biological datasets

To demonstrate the performance of the RAT workflow on real-world data, we sequenced metagenomes from groundwater, a relatively unexplored biome (93).

18 samples were collected from three groundwater monitoring wells in an agricultural area in the Netherlands. The same samples were used in an earlier study where they were analysed with 16S rRNA amplicon sequencing (363). Each well was sampled at six discrete depths between 12 and 64 m below the surface. We filtered 5–7 L of groundwater through 0.2µm filters and extracted DNA from the filters using the DNeasy PowerSoil Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. DNA quality (average molecular size) was checked with 1% (w/v) agarose gels stained with 1x SYBR® Safe (Invitrogen, Grand Island, NY) and quantified using the dsDNA HS Assay kit for Qubit fluorometer (Invitrogen). Whole metagenome shotgun sequencing was performed on the DNA by Novogene in Hong Kong on the Illumina MiSeq Platform, generating 35,289,790–58,902,006 paired-end sequencing reads of 2 × 251 bp per sample (**Supplementary Table 1**).

For quality-control, assembly, and binning, we used the ATLAS pipeline v2.4 (ref. 372). ATLAS uses BBTools (<https://sourceforge.net/projects/bbmap/>) to remove PCR duplicates and adapters and to trim the reads, assembles the reads using SPAdes v3.13.1 in metagenomic mode (204), and bins the contigs using MetaBAT 2 v2.14 (ref. 200) and MaxBin 2 v2.2.7 (ref. 199), after which DASTool v1.1.2 (ref. 366) is used to optimize MAGs resulting from the two binning approaches. We used MAGs of medium to high quality (>50% completeness, <10% contamination (52)) based on CheckM estimates in '--lineage_wf' mode (51). We ran RAT on multiple samples at a time using GNU parallel v20210622 (ref. 373) with the nr database downloaded on the 4th of March 2020. We ran Kaiju on the reads using default settings with a database containing NCBI nr for bacteria, archaea, viruses, fungi and microbial eukaryotes from the 24th of February 2021.

The N50 of the assembled contigs was between 1,435–3,012 nt per sample, the L50 was between 15,196–39,823 nt. Out of the 2,770,251 total contigs that were generated from the 18 samples, CAT annotated 2,411,810. All 423 medium- to high-quality MAGs were annotated at superkingdom rank or lower by BAT.

To further assess the diversity of groundwater organisms represented by the MAGs, we dereplicated all medium- to high-quality MAGs with dRep using default settings (367). We performed a phylogenetic analysis of the dereplicated MAGs based on the CheckM alignment of 43 universal marker genes that are used for phylogenetic placement (51). A phylogenetic tree was inferred with IQ-TREE v2.1.2 (ref. 281), ModelFinder (282), and UFBoot (283), using the model LG+R10 chosen according to BIC. The resulting tree was visualized with iTOL (374). The tree was rooted between the archaeal and bacterial MAGs based on their BAT classification (by RAT).

Plotting

All figures were made using R v4.1.3 and RStudio v1.1.456. The packages used for plotting were ggplot2 (375), tidyverse (376), reshape2 (377), ggalluvial (378), dplyr (<https://dplyr.tidyverse.org>), tidyr (<https://tidyr.tidyverse.org>), RColorBrewer (see <http://colorbrewer2.org>), Hmisc (<https://hbiostat.org/R/Hmisc/>), vegan (379), ape (380), and gridExtra (<http://CRAN.R-project.org/package=gridExtra>).

Availability of data and materials

RAT is available on GitHub at <https://github.com/MGXlab/CAT>. The scripts used in the downstream analyses are available on GitHub at https://github.com/thauptfeld/RAT_paper. The raw sequencing reads used for the biological analyses are available at SRA under BioProject ID PRJNA947390. Data from the CAMI challenge is available at <https://data.cami-challenge.org/participate>.

Acknowledgements

We thank Nora B. Sutton and her group for providing us with groundwater metagenomes to use for this manuscript. We thank Jan Kees van Amerongen for critical technical support. We thank the members of TBB at the University of Utrecht for their valuable input on the text.

Funding

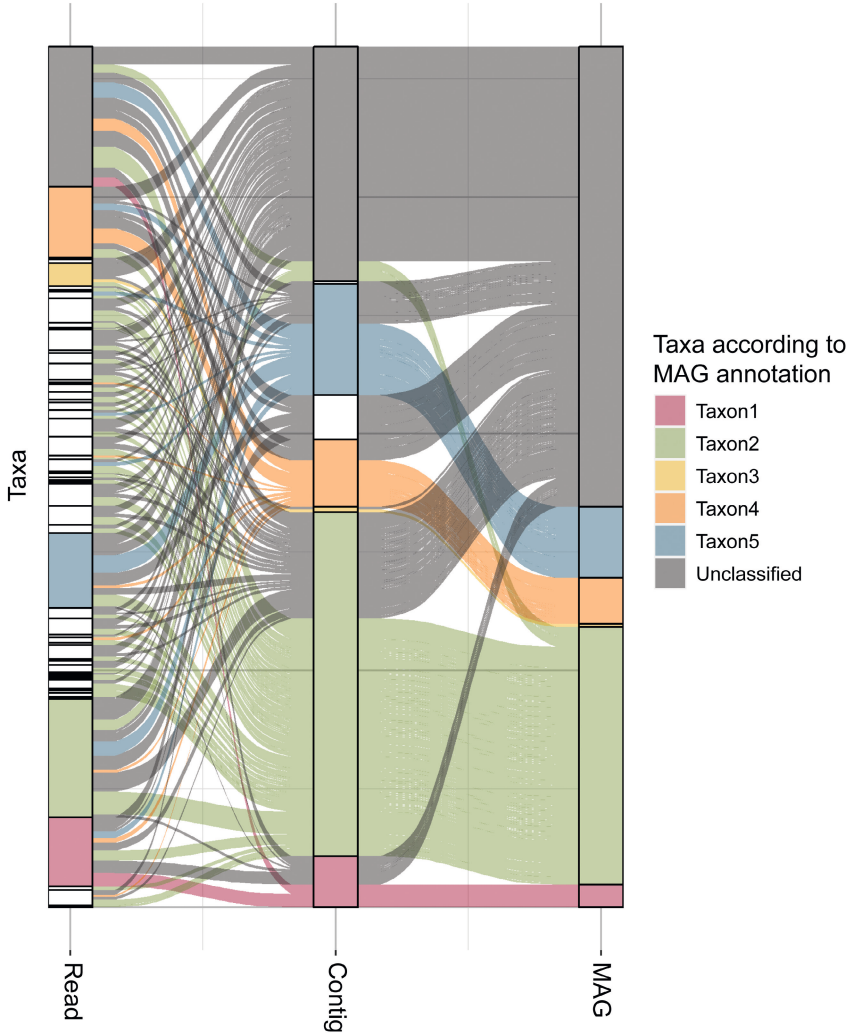
This work was supported by the European Research Council (Consolidator Grant 865694: DiversiPHI to B.E.D.), the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy (EXC 2051; Project-ID 390713860 to B.E.D.) and the Alexander von Humboldt Foundation in the context of an Alexander von Humboldt Professorship founded by the German Federal Ministry of Education and Research (to B.E.D.).

Contributions

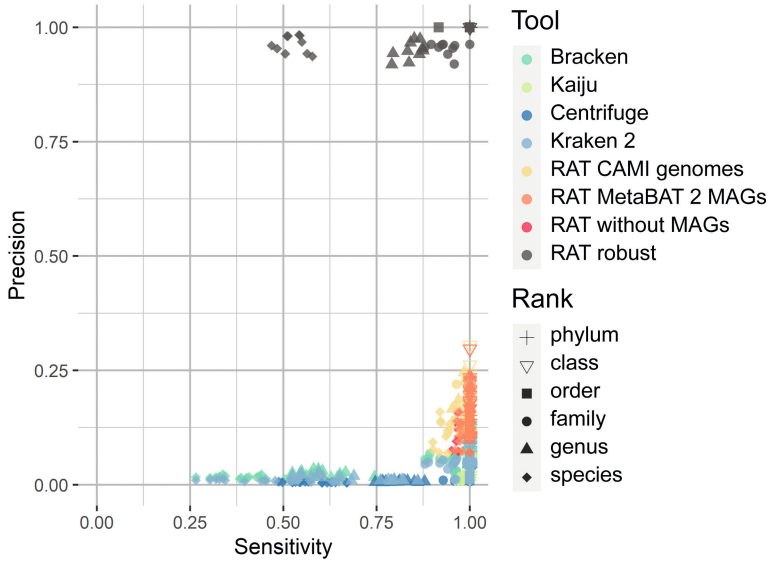
E.H. developed RAT, integrated it into the CAT pack, performed the CAMI benchmarks, and wrote the manuscript. N.P. integrated the GTDB option into CAT prepare. S.v.I. and B.L.S. performed the analysis on biological data. A.A.V. sampled the groundwater, extracted DNA and sent the groundwater samples for sequencing. B.E.D. and F.A.B.v.M. supervised the research and co-wrote the manuscript.

Supplementary Information

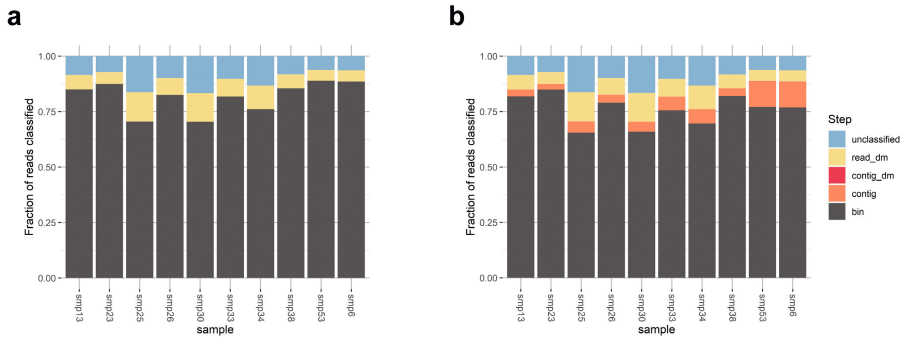
Supplementary Figures



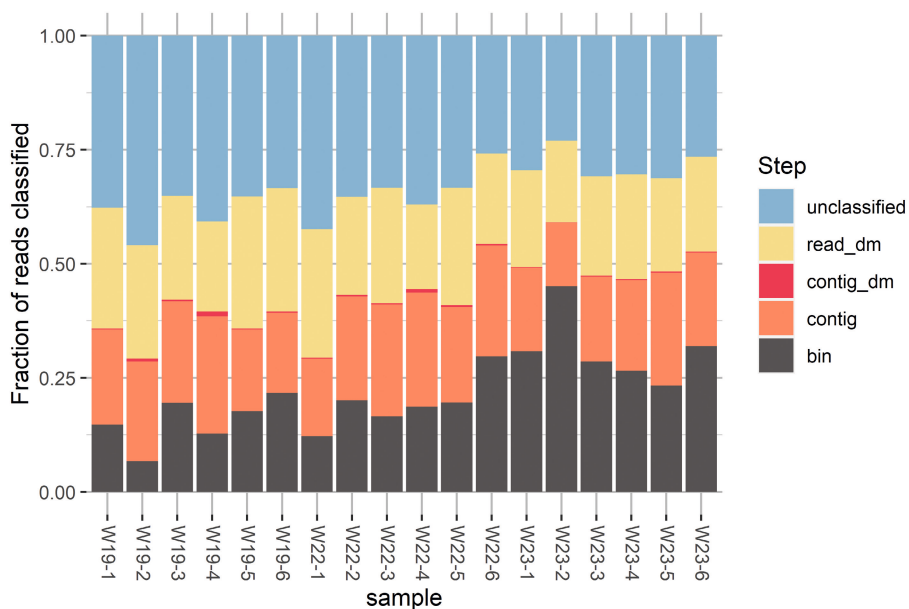
Supplementary Fig. 1 | Schematic depiction of noise reduction by using reliable taxonomic signals. Each column segment represents a taxon, each column represents an annotation step. In the read annotation step, many taxa are detected, the profile is noisy. At contig level, the number of detected taxa is much lower. Reads that were previously unannotated or had a spurious annotation now get annotated to one of eight main taxa. However, many reads that had an annotation on read level do not get an annotation on contig level, either because they don't map to any contigs, or because they map to a contig without annotation. At MAG level, there are even fewer detected taxa, and more unclassified data.



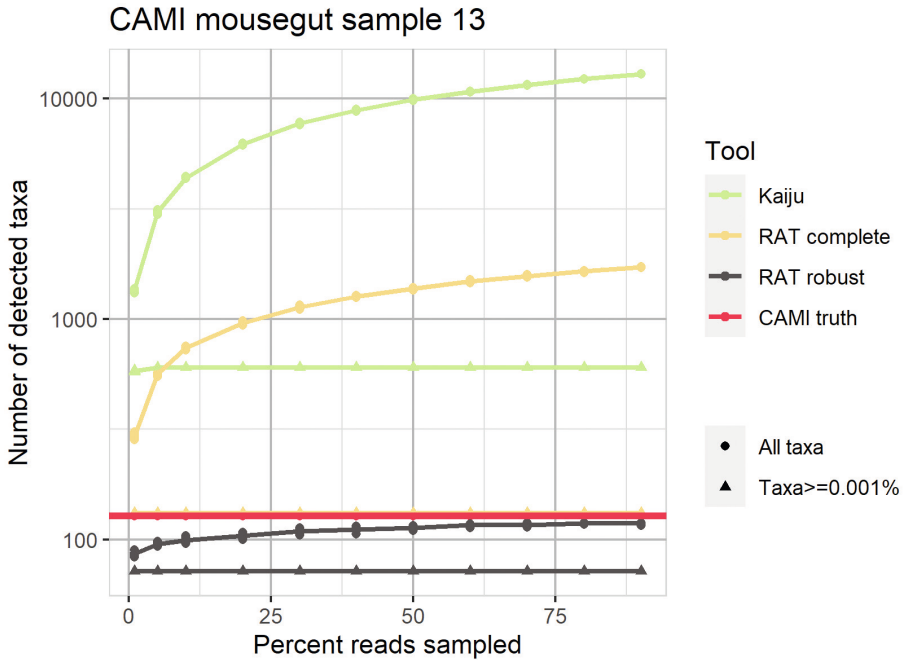
Supplementary Fig. 2 | Sensitivity versus precision of the different tools without an abundance cut-off to include taxa. Different shapes signify the sensitivity/precision on different ranks, different colours indicate tools (high precision + sensitivity is better).



Supplementary Fig. 3 | Fraction of read annotations in the simulated CAMI dataset and the taxonomic signal they originate from. 'bin' refers to BAT annotation of a MAG, 'contig' refers to CAT annotation of a contig, 'contig_dm' and 'read_dm' refer to DIAMOND blastx direct annotations of contigs/reads. **a**, Read annotations per taxonomic signal using CAMI genomes as MAG input. **b**, Read annotations per taxonomic signal using MAGs binned by MetaBAT 2.



Supplementary Fig. 4 | Fraction of read annotations in the biological groundwater dataset and the taxonomic signal they originate from. 'bin' refers to BAT annotation of a MAG, 'contig' refers to CAT annotation of a contig, 'contig_dm' and 'read_dm' refer to DIAMOND blastx direct annotations of contigs/reads.



Supplementary Fig. 6 | Number of taxa detected in one of the simulated CAMI datasets by RAT robust, RAT sensitive, and Kaiju. ‘CAMI truth’ refers to the actual number of taxa present in the sample. Triangles indicate the number of taxa detected in profiles when a minimum abundance is required to consider an organism as detected. Circles indicate the number of taxa detected without a cut-off.

Supplementary Tables

Supplementary Table 1 | Number of reads sequenced in the groundwater samples.

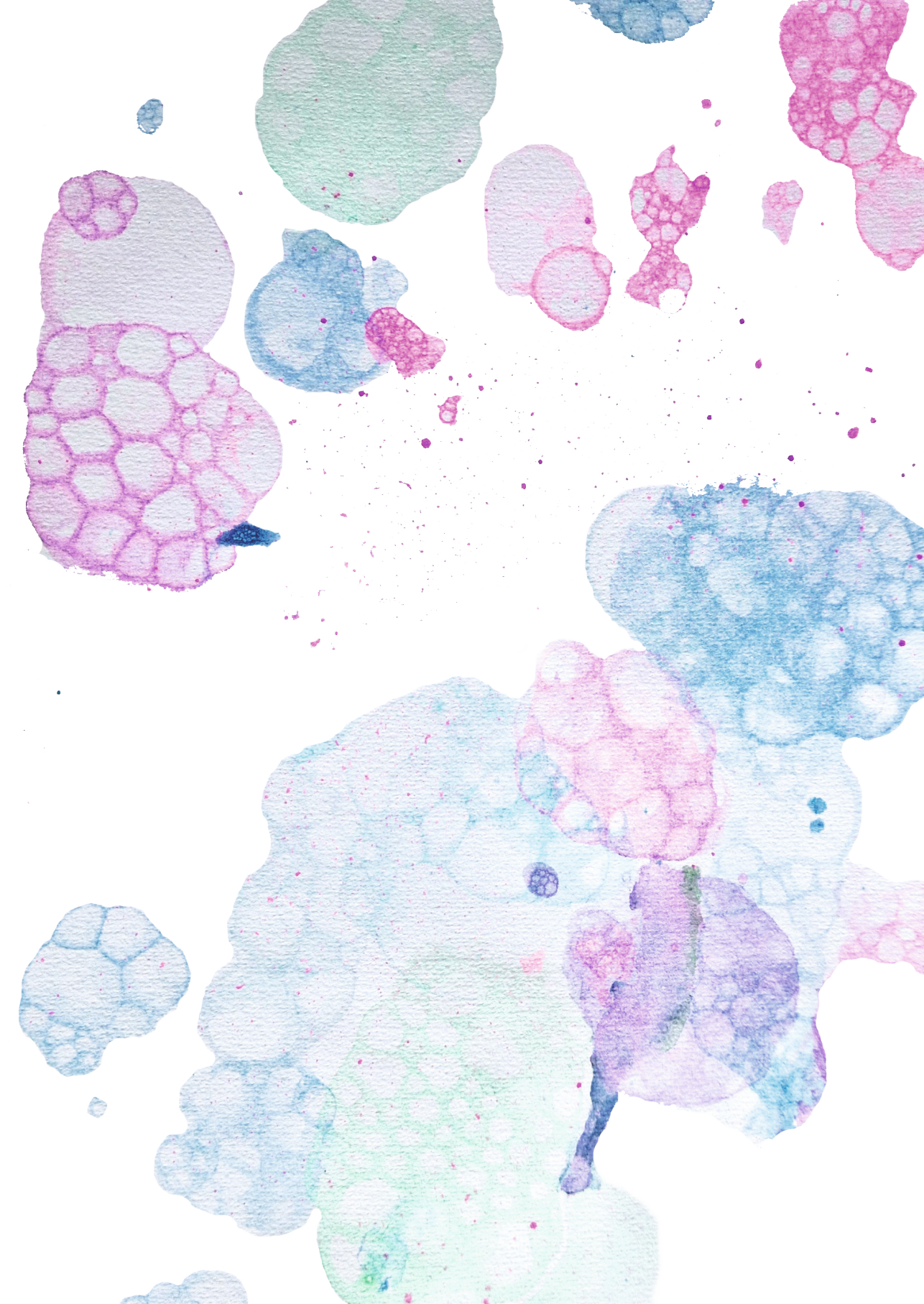
Sample	Reads sequenced
W19-1	36,438,396
W19-2	35,289,790
W19-3	42,407,036
W19-4	42,875,403
W19-5	38,591,389
W19-6	39,998,587
W22-1	42,251,696
W22-2	40,904,059
W22-3	43,615,855
W22-4	41,004,335
W22-5	39,219,955

**Supplementary Table 1 | Number of reads sequenced in the groundwater samples.
(continued)**

Sample	Reads sequenced
W22-6	58,902,006
W23-1	41,964,387
W23-2	40,896,174
W23-3	42,448,523
W23-4	41,463,768
W23-5	47,496,123
W23-6	46,181,232

Supplementary Table 2 | Fraction of reads with a superkingdom annotation by Kaiju and RAT in the groundwater samples.

Sample	RAT	Kaiju
W19-1	0.654212	0.646401
W19-2	0.562158	0.510839
W19-3	0.681671	0.642909
W19-4	0.611402	0.494389
W19-5	0.673378	0.666271
W19-6	0.693093	0.677197
W22-1	0.598599	0.614838
W22-2	0.67577	0.643249
W22-3	0.695687	0.668228
W22-4	0.657298	0.571904
W22-5	0.690264	0.646235
W22-6	0.77066	0.696525
W23-1	0.738719	0.676679
W23-2	0.791681	0.665519
W23-3	0.715761	0.631563
W23-4	0.719839	0.668849
W23-5	0.716869	0.667231
W23-6	0.760786	0.698729





Chapter 5

A social niche breadth score reveals niche range strategies of generalists and specialists

F. A. Bastiaan von Meijenfeldt, Paulien Hogeweg, and Bas E. Dutilh

Nature Ecology & Evolution 7, 768–781 (2023)

Abstract

Generalists can survive in many environments, whereas specialists are restricted to a single environment. Although a classical concept in ecology, niche breadth has remained challenging to quantify for microorganisms because it depends on an objective definition of the environment. Here, by defining the environment of a microorganism as the community it resides in, we integrated information from over 22,000 environmental sequencing samples to derive a quantitative measure of the niche, which we call social niche breadth. At the level of genera, we explored niche range strategies throughout the prokaryotic tree of life. We found that social generalists include opportunists that stochastically dominate local communities, whereas social specialists are stable but low in abundance. Social generalists have a more diverse and open pan-genome than social specialists, but we found no global correlation between social niche breadth and genome size. Instead, we observed two distinct evolutionary strategies, whereby specialists have relatively small genomes in habitats with low local diversity, but relatively large genomes in habitats with high local diversity. Together, our analysis shines data-driven light on microbial niche range strategies.

Introduction

Culture-independent sequencing studies have greatly expanded our understanding of the microbial world. They uprooted the tree of life (97,104), revolutionized our view of the human microbiome and virome (119,135) and advanced our comprehension of early evolution (101,381). By using standardized protocols across large numbers of samples (132,134,144,173), classical ecological questions can now be addressed on the global scale. A quintessential question is that of ecological niche breadth (382,383)—the range of conditions under which an organism can live. Although the distinction between specialists and generalists is a fundamental property of life and its evolution, general mechanisms that determine niche breadth are poorly understood (384) and quantification has proven challenging (385).

Microbial niche breadth has been measured for specific aspects of the environment (for example, temperature (386,387), pH (388) and nutrient dependence (389,390)). Niche breadth definitions that assess the full n -dimensional niche space (391) have been based on occurrence in environmental samples. Rather than the theoretical fundamental niche, microbial occurrence represents its empirical realized niche. Because of complex interactions within microbial communities, the realized niche can be both smaller (for example, due to competition (392)) or larger (for example, due to metabolic dependencies (393)). Previous studies defined organisms that are present in many samples or predefined habitats as generalists, and rare organisms as specialists (394–397). Based on this definition, Sriswasdi et al. (398) suggested an important evolutionary role for generalist species in maintaining taxonomic diversity, with generalists having higher speciation rates and persistence advantages over specialists. Others defined the niche breadth of an organism by the uniformity of its distribution across habitats (399), suggesting that community assembly of specialists is driven by deterministic processes, whereas for generalists neutral processes are more important (400,401). Notwithstanding these intriguing results, niche breadth studies based on occurrence in microbiomes have been sensitive to biases due to habitat definition and sample selection.

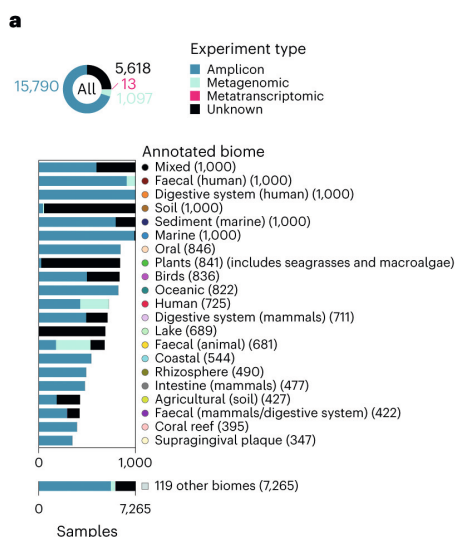
Microbiomes are sensitive biomarkers capable of detecting geochemical gradients (167), host health status (168–170) and metabolites in a given niche (171,172). We thus reason that the vast collection of tens of thousands of environmental sequencing datasets that are available in the public domain (152) could be used to implement an unbiased, data-driven and comprehensive niche breadth definition, based on community similarity between samples where microbial taxa occur. As such, we treat community composition as a proxy for the realized niche of a microorganism that reflects both the abiotic environment and the microbial interactions within. Similar reasoning has been used to quantify the niche range of eukaryotes without the use of external habitat definitions (402). In this view, organisms that occur in

compositionally similar samples are social specialists, as their niche is restricted to the same local neighbours, and organisms that occur in compositionally dissimilar samples are social generalists, as they are more flexible in their interaction partners. Using community similarity as a substitute for ecological range, we developed a social niche breadth (SNB) score that allowed us to quantify the social niche range for taxa at all taxonomic ranks and assess strategies for specialization and niche range expansion throughout the prokaryotic tree of life.

Results

SNB captures global heterogeneity in microbial communities

To compare the niche breadth of microbial taxa, we devised and extensively benchmarked (**Supplementary Information**) an SNB score that exploits the abundantly available meta-omics datasets derived from diverse environments around the world (**Fig. 1** and **Supplementary Data 1** and **2**). These microbiomes are taxonomically annotated with the same MGnify pipeline (152), which allows for a comparison of vastly different environments, studies and experiment types (**Supplementary Information**). First, we assessed the biome annotations of these datasets, as provided by the dataset submitters. The annotations highlighted the main drivers of microbiome composition (**Fig. 1d** and **'Supplementary Information'**), including salinity (t-distributed stochastic neighbour embedding dimension 1 (t-SNE 1)) and host association (t-SNE 2) (173,403–405). The 22,518 samples covered a total of 140 annotated biomes that differed markedly in within-sample (α) and between-sample (β) diversity. Annotated biomes with high mean α diversity, such as soils, had low β diversity (**Fig. 1f**), implying a relatively stable core community across these high-diversity habitats.



Niche range strategies of generalists and specialists

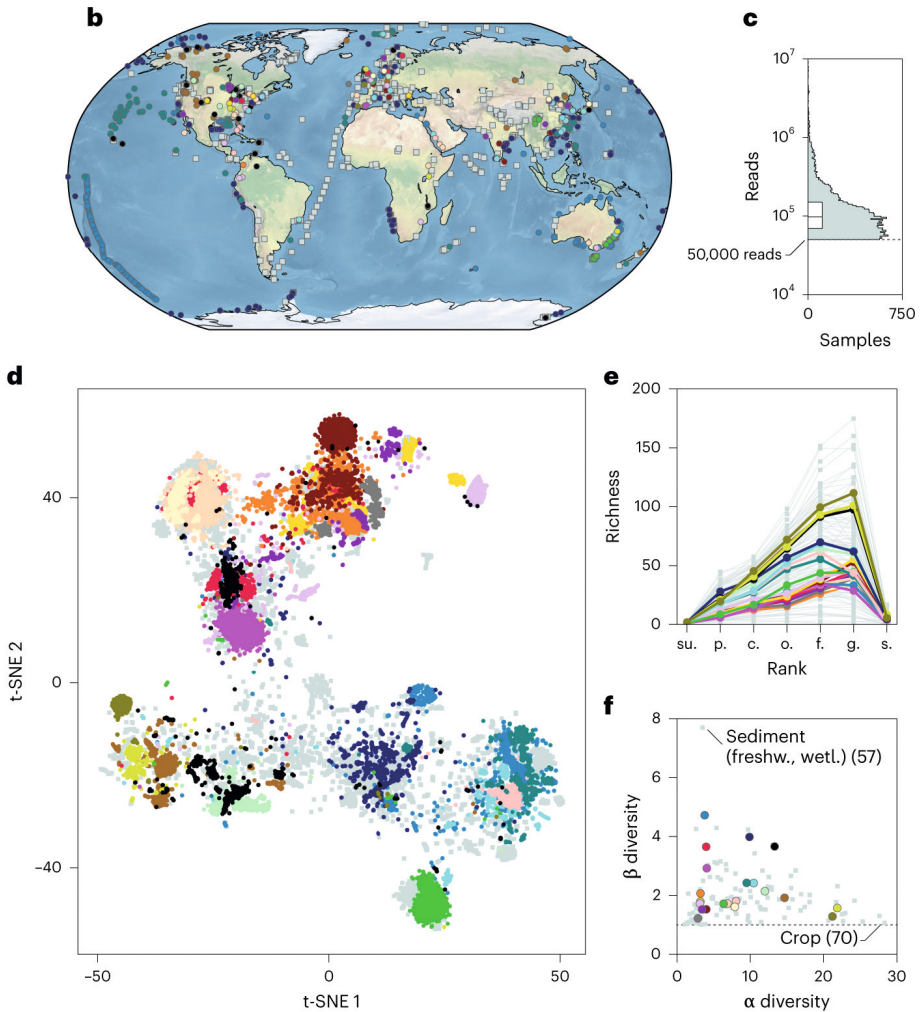


Fig. 1 | A diverse and global microbial dataset. **a**, Samples were received from vastly different annotated biomes and study designs. The numbers in parentheses indicate the number of samples within the annotated biome. Annotated biomes with fewer than 347 samples have been grouped as other. For a hierarchical tree of all annotated biomes, see **Supplementary Fig. 1b**. **b**, Geographical distribution of the samples. **c**, Total number of taxonomically annotated reads per sample ($n = 22,518$ samples). The box plot shows the interquartile range and median. No samples with fewer than 50,000 reads were selected. **d**, Samples from similar annotated biomes cluster together based on taxonomic profile in a t-SNE visualization (perplexity = 500), with the same ecological dissimilarity measure used as for SNB (namely, the Spearman's rank correlation coefficient ($0.5 - (\rho/2)$) of known taxa at taxonomic rank order). For a PCoA visualization of the same data and the positions of all 140 annotated biomes on the PCoA, see **Supplementary Figs. 2 and 3**, respectively. Most samples from the plants biome were derived from seagrasses and macroalgae from kelp forests. **e**, Taxa richness differs per annotated biome and taxonomic rank. The low number of annotated species is a consequence of a relatively unexplored biosphere. su., superkingdom; p., phylum; c., class; o., order; f., family; g., genus; s., species. **f**, Annotated biomes with high mean α diversity have low β diversity, whereas both low and high β diversity is found among annotated biomes with low mean α diversity. freshw., freshwater; wetl., wetlands.

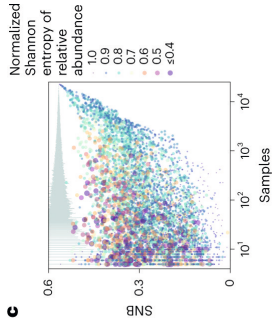
Most samples from the same annotated biome are relatively similar, as reflected by a low β diversity. Nevertheless, annotated biome definitions are arbitrarily delineated and may be subject to human error. For example, the plants biome includes both freshwater plants (406–408) and seagrasses (409), as well as macroalgae from kelp forests (410) (**Supplementary Data 1**). Also, it is difficult to quantify the degree of similarity between categorical biomes in a biologically meaningful way. We used the observation that microbiomes are biomarkers (167–172) and developed SNB, which captures the compositional heterogeneity of samples for which a taxon is found to quantify niche breadth.

We assume that the small subunit rRNA gene that is queried is a proxy for the genetic content of a taxon that defines its traits. Specific traits exist at all taxonomic ranks and determine their occurrence across microbiomes (411). Since the taxonomic annotations are based on a reference taxonomy and the biosphere is relatively unexplored, sometimes high-ranking taxa do not have low-ranking annotations like species (**Fig. 1e**). We considered that members of a taxon are alive and growing if the taxon represented a relative abundance of at least 1/10,000 of the prokaryotic reads in a sample, and thus ignored the possibility of migration from other sources and the potential for dead organic matter contributing DNA to the sequencing results (412). Next, we defined SNB as the mean pairwise dissimilarity between these microbiomes. After benchmarking 150 different ecological dissimilarity measures for their ability to separate the annotated biomes, we chose mean pairwise dissimilarity based on the inverse Spearman's rank correlation of known taxa at taxonomic rank order to quantify SNB (**Supplementary Information**). Thus, taxa with a low SNB score are found in samples with very similar microbial composition (social specialists) and taxa with a high SNB score are found in dissimilar samples (social generalists). Our approach accounts for database biases, as some environments are much more frequently sampled than others (**Fig. 1a** and '**Supplementary Information**'). Indeed, taxa that are detected in the same number of samples or annotated biomes may have very different SNB scores (**Fig. 2a–c**). Different from studies that investigate the co-occurrence of taxa across samples (25,413), SNB quantifies the range of communities that a taxon can occur in. SNB treats each sample as a local niche and infers that taxa that occur across highly differing communities are social generalists, while taxa that occur in similar communities are social specialists. Since SNB is calculated when a taxon is present over a detection limit of 1/10,000 reads, the relative abundance and associated variability of a taxon's distribution are observables and can be associated with niche range rather than part of the definition, as in ref. 399.

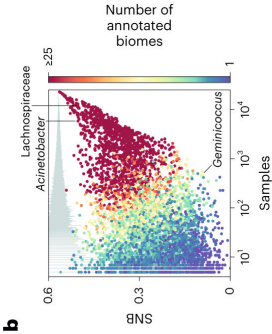
SNB throughout the prokaryotic tree of life

To investigate the distribution of social generalists and specialists throughout the prokaryotic tree of life, we calculated SNB for taxa at all ranks (**Fig. 2d** and **Supplementary Data 3**). For the vast majority of taxa, the SNB score is lower than expected based on random permutations (**Fig. 2a–c**), indicating that all microorganisms are social specialists to some extent because they occur in a non-random subset of all samples. Exceptions to this rule include the high-ranking superkingdom Bacteria and phylum Proteobacteria, which are widespread, occurring in 22,295 and 22,211 of the 22,518 samples, respectively. While there is a clear positive correlation between SNB and the number of samples in which a taxon occurs (**Fig. 2a–c**), very rare taxa such as *Aminobacter* (five samples) and *Methanimicrococcus* (28 samples) still have a high SNB (SNB = 0.56 and SNB = 0.51, respectively). Alternatively, some taxa that are found in many samples have a relatively low SNB because these samples are very similar in composition (for example, *Phyllobacterium* (226 samples; SNB = 0.03) and *Geminicoccus* (473 samples; SNB = 0.09)).

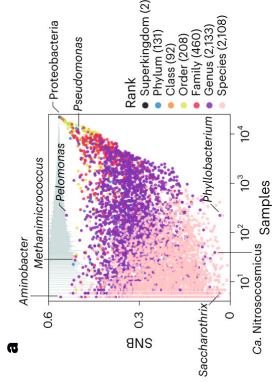
The distribution of SNB scores differs per taxonomic rank. High-ranking taxa tend to have higher SNB scores than low-ranking taxa (**Fig. 2d,e**), which intrinsically occur in a subset of the samples of their parent taxa. High-ranking taxa can have high SNB scores either because they contain subtaxa that are specialists in different communities or because the subtaxa are also generalists. To compare taxa at different ranks, we calculated a rank-specific modified z score (**Fig. 2d** and **Supplementary Data 4**), where positive z scores indicate that the SNB of the taxon is higher than the median for its rank and the taxon is thus relatively generalist and negative z scores indicate that it is relatively specialist. For example, the family Flavobacteriaceae and the genus *Prevotella* are social generalists (with z scores of 2.02 and 0.61, respectively), but their subtaxa are relatively specialized for their rank (median z score of genera in Flavobacteriaceae = -0.53 ; median z score of species in *Prevotella* = -0.71). The family Lactobacillaceae on the other hand is generalist (z score = 0.46) and its genera are also generalists (median z score = 1.73). In addition, high-ranking taxa with high SNB scores often have more subtaxa than high-ranking taxa with low SNB scores (**Supplementary Fig. 16**). This suggests that the diversity of taxa, as currently represented by taxonomy, reflects their ecological range well. The four best-studied phyla, Proteobacteria, Firmicutes, Bacteroidetes and Actinobacteria, which together cover 97% of cultured prokaryotic species (39), are dominant across a wide range of environments (**Supplementary Fig. 17**) and have a higher SNB than others (**Fig. 2d**).



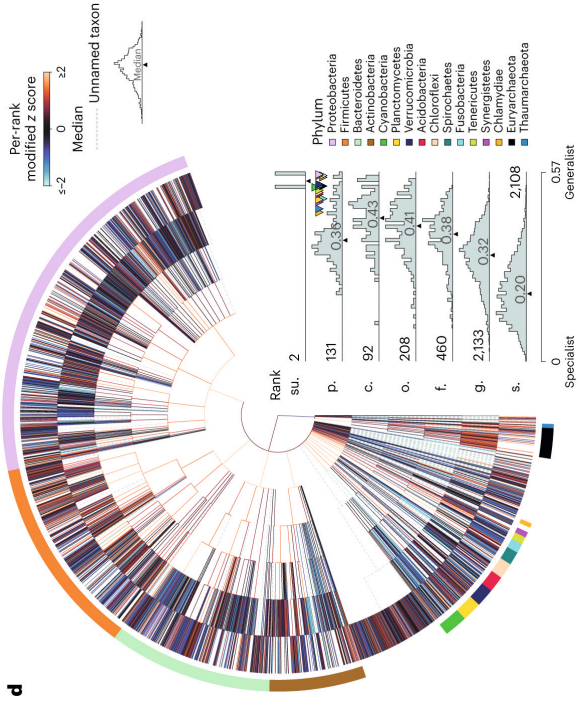
c



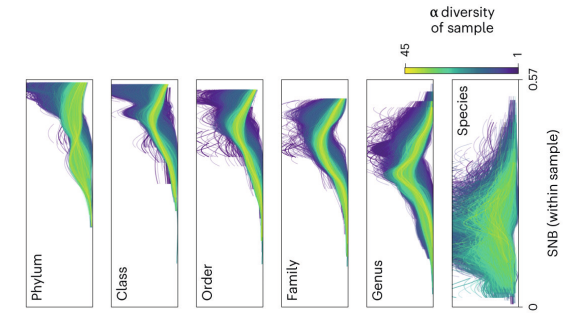
b



a



d



e

Fig. 2 | SNB throughout the prokaryotic tree of life and across samples. a–c, Relationship between SNB and the number of samples in which a taxon is found, coloured by rank (a), number of annotated biomes (b) and normalized Shannon entropy of relative abundance (c). The grey bars on top show the range of SNB scores of imaginary taxa that were present in 100 randomly picked subsets of samples of the specific size. The locations of some outlier taxa are indicated. Numbers within brackets indicate the number of taxa measured at each rank. The normalized Shannon entropy of relative abundance across samples is represented as shading for each taxon in c. Both colour coding and the size of the markers represent the Shannon entropy. Note that higher entropy is indicated with smaller markers. Relative abundance across samples was more constant for social specialists than for social generalists. Ca., *Candidatus*. d, SNB of taxa throughout the prokaryotic tree of life. SNB scores are standardized per rank based on the median absolute deviation (modified z scores), with low z scores representing taxa that are relatively specialist for their rank and high z scores representing taxa that are relatively generalist. The distributions of SNBs at different taxonomic ranks are shown as histograms, for which the numbers on the distributions show the number of taxa at that rank. The most diverse phyla are colour coded. e, Distribution of SNBs within samples at different taxonomic ranks. The α diversity of a sample was calculated on the rank order.

There are many phyla that have low SNB scores and contain few classes, orders and families compared with the dominant ones described above. Subtaxa of these low-scoring phyla are thus under-represented at the class, order and family ranks and we observe that the distribution of SNB scores is more skewed towards social specialism at the phylum rank (median SNB = 0.36) than at these lower ranks (median SNB = 0.38–0.43; see **Fig. 2d**). Many phyla with the *Candidatus* status have a low SNB compared with validly described phyla (**Supplementary Fig. 17**). The connection between the *Candidatus* status and low SNB may reflect a discovery bias of these phyla where widespread lineages tend to be discovered and described sooner than rare ones, although some candidate phyla are widespread (**Supplementary Fig. 17**). Candidate phyla may require specific growth conditions, which can be reflected in relatively stable specialized microbial communities, consistent with their low SNB. In addition, several candidate phyla, including the bacterial ‘Candidate Phyla Radiation’ and DPANN archaea, may consist of obligate symbionts of specific hosts (97). Whereas it was recently shown that consortia of obligate symbionts can grow on a wider range of carbon sources than their individual members and thus expand their metabolic niche (393), the individual microorganisms in these consortia are social specialists as they require specific partners in their local communities.

Taxa with high and low z scores are dispersed throughout the prokaryotic tree of life (**Fig. 2d**), indicating that social specialization and niche range expansion happened independently numerous times in evolution. Phyla with relatively specialized genera include Proteobacteria (median z score = -0.07), Bacteroidetes (median z score = -0.26), Actinobacteria (median z score = -0.17), Cyanobacteria (median z score = -0.72), Planctomycetes (median z score = -0.37), Acidobacteria (median z score = -0.47) and Chloroflexi (median z score = -0.17), whereas Firmicutes, Tenericutes and Euryarchaeota have genera that are relatively generalist (median z scores of 0.43, 1.18 and 1.06, respectively). Taxa with relatively low SNB for their ranks include known specialists such as the genus *Christensenella* (398) (z

score = -1.01), but also the family Pelagibacteraceae (z score = -1.94) and genus *Prochlorococcus* (z score = -1.25), which hold some of the most abundant organisms on Earth (414,415). These taxa, known for their highly streamlined genomes (416), are found in aquatic samples with a uniform microbial composition (**Supplementary Fig. 5b**) and thus have a low SNB. While the family Pelagibacteraceae contains both marine and freshwater representatives (in the SAR11 and LD12 clades, respectively (417)), in our dataset it is found primarily in marine samples (**Supplementary Fig. 5b**). This highlights that future sampling of even more habitats, combined with more sensitive detection methods, could change or refine SNB scores for some taxa. The genus *Roseobacter*, whose members are considered marine metabolic generalists with large genomes and a versatile metabolism (418,419), is found in more diverse samples (**Supplementary Fig. 5b**) and has an SNB closer to the median of all genera (z score = -0.30). At the generalist end of the spectrum are taxa that are ubiquitously present in our dataset (**Fig. 2a,b**), such as the genera *Acinetobacter* (z score = 2.30) and *Pseudomonas* (z score = 2.33 ; however, this genus may be ubiquitous in part because it is a common contaminant of DNA extraction kits (420)) (**Supplementary Fig. 5b**). The family Lachnospiraceae (z score = 1.74 ; found in over half of all samples; $n = 11,887$) and its genera (median z score = 0.79)—obligate anaerobes that were previously regarded as habitat specialists (398)—also have a high SNB for their ranks, highlighting the heterogeneity of the communities in which they are found.

Generalists dominate, whereas specialists are stable but scarce

Next, we set out to find patterns in SNB. We focused our analysis on genera because they balance a high taxonomic resolution with a good representation in the dataset (**Fig. 1e**) and show a broad range of SNB values (**Fig. 2d,e**), allowing for a comprehensive investigation of niche range strategies.

It has been suggested that generalists, being Jacks of all trades, can be masters of none (382), while specialists are adapted to become dominant within their habitats under stable conditions (421). The niche range may thus reflect a trade-off, where specialists gain local dominance at the expense of ecological versatility. Alternatively, computational models of microbial metabolism have suggested that metabolically flexible generalists have faster growth rates than specialists (422). We correlated SNB with local abundance and found that social generalists are dominant in most annotated biomes, as indicated by a consistent positive correlation across samples, with exceptions including marine host organisms such as corals, seagrasses and macroalgae (**Fig. 3a**). SNB positively correlates with abundance within samples, meaning that social generalists locally outcompete their more specialist neighbours, disputing the expected trade-off mentioned above. While these are general results based on correlations across samples, an exception is *Prochlorococcus*, which has a low SNB but a high local abundance (mean relative abundance = 0.63%). This genus is among the top 10% of the most abundant genera (**Supplementary Data 3**)

and in the majority of its samples belongs to the top 20% of genera in terms of local abundance. Local dominance of habitat generalists has previously been observed in specific environmental settings such as highly dynamic sandy ecosystems (423). Some soil microorganisms are both abundant and ubiquitous (424) and only ~500 dominant phylotypes (that is, 2%) represent >40% of soil bacterial communities (110). Our results show that these observations reflect a general pattern wherein generalists are dominant.

Whereas samples are typically dominated by social generalists, we find that the relative abundance of generalists is more variable across samples than that of specialists, whose abundance is relatively stable. This is evident when comparing the niche range of organisms that locally co-occur within samples, where social generalists have a higher variability of relative abundance than social specialists (**Fig. 3a**). It is also evident for taxa throughout the prokaryotic tree of life, where social specialists have an even relative abundance while taxa with a high variability of relative abundance are social generalists (**Fig. 2c**). Even *Prochlorococcus*, while having a high local dominance for a social specialist, still has an even abundance across samples (normalized Shannon entropy of relative abundance = 0.86).

Our data counter the classic Jack of all trades argument, which suggests that specialists should have a local fitness advantage at the expense of ecological versatility. We explored possible explanations for the local dominance of generalists over specialists and their relatively variable abundance. First, although generalist genera contained more species than specialist genera in the total dataset (**Supplementary Fig. 18a**), we did not find evidence that they also contain a higher number of species within samples (**Supplementary Fig. 18a**), but note that only a small fraction of genera could be taxonomically classified on the species rank (**Supplementary Fig. 18b**). Alternatively, SNB may reflect the classical distinction between r strategists and K strategists (425). Social specialists have a low but constant abundance near carrying capacity (K selected) and some (but not all) social generalists are opportunistic taxa that reach high relative abundance when circumstance permits (r selected). To test this hypothesis, we compared the SNBs of microorganisms with their predicted maximal growth rates based on the EGGGO database (187) (**Fig. 3a**) and confirmed that, within samples, social generalists have shorter doubling times than social specialists. These results support the idea that generalist genera include more opportunistic growers than specialist genera.

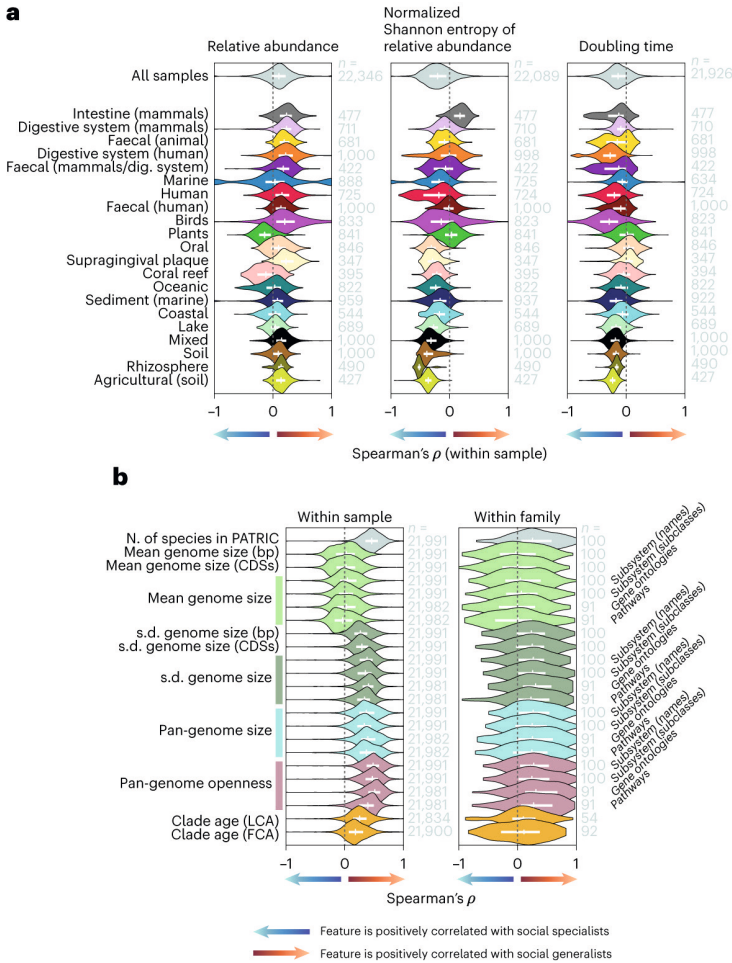


Fig. 3 | Ecological and genomic features correlated with SNB. **a**, Spearman's rank correlation coefficient (ρ) within samples between SNB and features related to local dominance on the rank genus. The violin plots depict the distribution of ρ across all samples or those from the annotated biomes with the most samples. Annotated biomes are arranged according to mean α diversity. Positive values indicate that the feature is positively correlated with social generalists and negative values indicate that the feature is positively correlated with social specialists. Doubling time estimates are from the EGGO database. dig., digestive. **b**, ρ between SNB and genomic features in the PATRIC database on the rank genus within samples and within families. The violin plots depict the distribution of ρ across all samples and across all families with at least five genera. Genome sizes are given in base pairs (bp) and numbers of coding sequences (CDS). Genomic measures with annotations to the right are in numbers of unique functions for that specific functional universe. Genome size estimates for a genus are based on the genome size of its species, which is defined as the mean size of all strains for base pair and coding sequence measures, and as the majority set of functions of all strains for the functional universe measures. Pan-genome openness is the total pan-genome size divided by the mean genome size. Correlations with time to the last common ancestor (LCA) and first common ancestor (FCA) are based on the TimeTree database. Numbers to the right of violins show sample sizes. Lines within violin plots show the interquartile range and median. Supporting data are available in **Supplementary Data 6** and **7**. N., number; s.d., standard deviation.

SNB reflects genomic heterogeneity

Next, we used our dataset to assess the suggestion that social generalists have large genomes that encode many functions, reflecting a versatile metabolism that allows them to colonize diverse habitats (396,398). For example, bacteria that are found in a diverse range of habitats encode more extracellular proteins than bacteria that are restricted to few habitats (397), and habitats with temporal variation may select for larger genomes (426). In contrast, specialization may be associated with a reduction in genome size due to a loss of unnecessary genes (as has been observed in members of the phylum Planctomycetes transitioning from soil to freshwater habitats (427)) or genome streamlining (428) (which is common in oligotrophic marine waters (429,430)). Genomic versatility of high-ranking taxa, reflected in a large pan-genome (106,107), may either result from small yet diverse genomes in individual subtaxa (open pan-genome) or genomically versatile yet functionally similar strains (closed pan-genome). We set out to identify genomic features associated with SNB using publicly accessible genome sequences from the Pathosystems Resource Integration Center (PATRIC) database (185) (**Supplementary Data 5**). These features include the mean genome size of all species in the genus, the variation in these genome sizes, the pan-genome size (that is, the total number of functions present in all genomes) and the pan-genome openness (calculated as the pan-genome size divided by the mean genome size). The PATRIC database contains genome sequences for 1,704 of the 2,133 genera that we investigated in our global microbiome dataset (**Supplementary Data 3**). Although these genomes probably belong to different strains or species than those observed in MGnify, we decrease the inconsistencies in our analysis by assessing their genomic features at the genus rank.

We compared genera within samples (**Fig. 2e**) for an ecological view and within their taxonomic families for an evolutionary view. Both perspectives gave qualitatively similar results (**Fig. 3b**), indicating that genomic signatures of SNB (see below) are generalizable across habitat and phylogeny. Although the number of samples is larger than the number of families, the correlation between genomic features and SNB is more consistent within samples than within families, possibly suggesting that ecology is a stronger driver of (pan)-genome evolution than phylogenetic history (431).

When comparing taxa across all samples, we found no consistent correlation between SNB and genome size, whether measured in the number of nucleotides, genes or unique functions (**Fig. 3b**). We did, however, observe that the genomes in generalist genera are more variable in size than the genomes in specialist genera, as seen in their standard deviation. Moreover, the pan-genomes of generalist genera contain more functions, in line with theoretical models that suggest that the ability to migrate to new niches is associated with pan-genome size (368). The pan-genome size of microorganisms may be positively associated with effective population size

(432), which may be larger for social generalists. The same study found that rapidly growing microorganisms have large effective population sizes, in line with our earlier discussed observation that opportunistic growers are generalists. Finally, the pan-genomes of social generalists are more open than those of social specialists. These results did not depend on the higher number of species in generalist genera (**Supplementary Fig. 19**).

In conclusion, species in specialist genera are genomically more similar than species in generalist genera, which is reflected in more similar genome sizes and less variation in functions, and an associated smaller and more closed pan-genome than generalist genera. We hypothesize that the observed genomic flexibility allows members of generalist genera to rapidly acquire the genes needed to thrive in a given local environment (433,434), with their higher growth rate potential allowing them to outgrow specialists. The correspondence between genomic heterogeneity and the heterogeneity of communities (SNB) confirms the strong association between ecological and genomic diversification. To further explore this diversification we correlated SNB with clade age based on the TimeTree database (435). It was previously suggested that habitat generalist species are evolutionarily younger than habitat specialist species (398). Our data do not allow analysis of these trends at the species rank, but at the genus rank we found that social specialists are younger than generalists, as indicated by a consistent positive correlation between SNB and clade age (**Fig. 3b**). Together, these results support a model of continuous diversification whereby old generalist clades that share a diverse pan-genome may invade new niches, leading to the emergence of specialized subtaxa.

Two contrasting genomic niche range strategies

As discussed above, SNBs of genera were not consistently associated with mean genome size. However, we did observe a habitat-dependent relationship between genome size and social niche range (**Fig. 4a**). We found two contrasting strategies that broadly depended on the α diversity of the samples. In samples with low local diversity (α diversity ≤ 11 ; **Fig. 4b**), including most animal-associated and saline habitats, there was a mean positive correlation between genome size and SNB. In contrast, in samples with high local diversity (α diversity > 11 ; **Fig. 4b**), including most free-living non-saline habitats and the rhizosphere, the correlation was often negative. Because databases contain a majority of samples from animal-associated and marine habitats with relatively low α diversity (**Fig. 4c**), and because genome size estimates are often based on cultivated microorganisms that differ markedly from environmentally derived genomes (436), previous suggestions of a positive correlation between genome size and niche range may have overlooked the contrasting correlations.

In habitats with low α diversity, social generalists tend to have large genomes that may encode the functions needed to utilize many different resources, whereas social specialists (low-diversity specialists) have the smallest genomes known (**Fig. 4d,e**). The coding density—a signature of genomic streamlining (428)—is significantly lower in low-diversity specialists than in social generalists ($P < 0.003$; one-tailed t -test; measured as the number of coding sequences per base pair, with generalists defined as $\text{SNB} > 0.35$ (697 genera) and low-diversity specialists defined as present in samples with a mean α diversity ≤ 11 (552 genera)). This suggests that genome streamlining is not the common route to genome reduction in low-diversity specialists. Instead, their small genomes could reflect specialization to habitat-specific metabolites and the loss of genes through drift (427). In addition, cooperating metabolic specialists could supplement each other's nutrient requirements (25) or they could depend on the co-occurring generalists.

Although the genomes of social generalists are large compared with those of low-diversity specialists (**Fig. 4d**), the genomes in samples with low α diversity are still moderate in size compared with the large genomes in samples with high α diversity (**Fig. 4e**). Samples with high α diversity contain many specialists (**Fig. 2e**) and their high richness may be a driver of specialization through competitive exclusion, as has been suggested to explain a negative correlation between niche breadth and local diversity in eukaryotes (437). There is negative correlation between genome size and social niche range in samples of high α diversity, where social specialists (high-diversity specialists) have larger genomes than co-occurring social generalists (**Fig. 4b**). The largest known prokaryotic genomes belong to high-diversity specialists (**Fig. 4d**); for example, the genus *Polyangium*, with a mean genome size of 12.7 megabases and an SNB of 0.21 (z score = -1.36). Selection may favour large genomes in habitats with diverse but scarce nutrient availability where slow growth is no disadvantage, such as in soils (438,439). Moreover, in contrast with the earlier-mentioned cooperative taxa, microorganisms in competitive consortia carry many metabolic functions (25). High-diversity specialists may thus reflect a competitive metabolism. Social generalists in these habitats may use metabolites that are irregularly available and rapidly depleted, consistent with their opportunistic nature and variable occurrence. Alternatively, they could exploit metabolic byproducts generated by metabolic specialists (390). Regardless of the mechanism, it appears that adaptation of high-diversity specialists to their habitats by genome expansion decreases their competitiveness in differing communities.

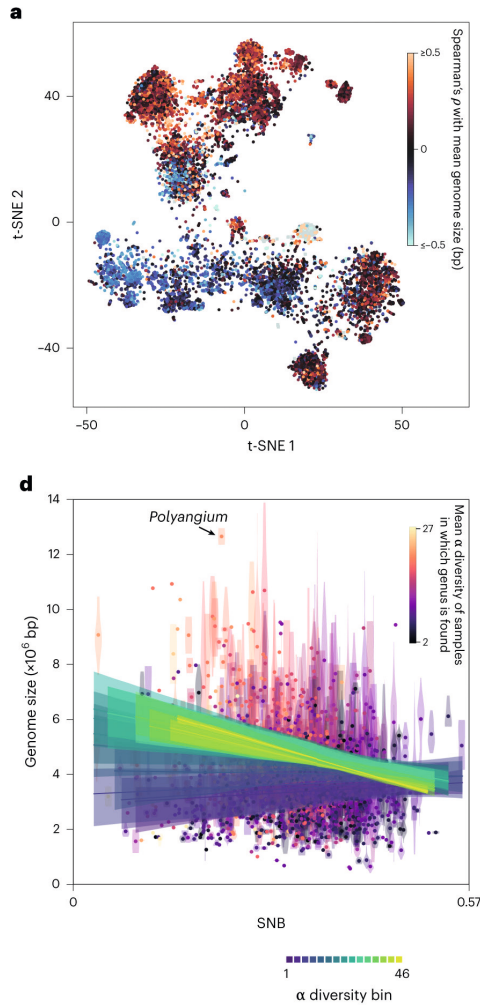
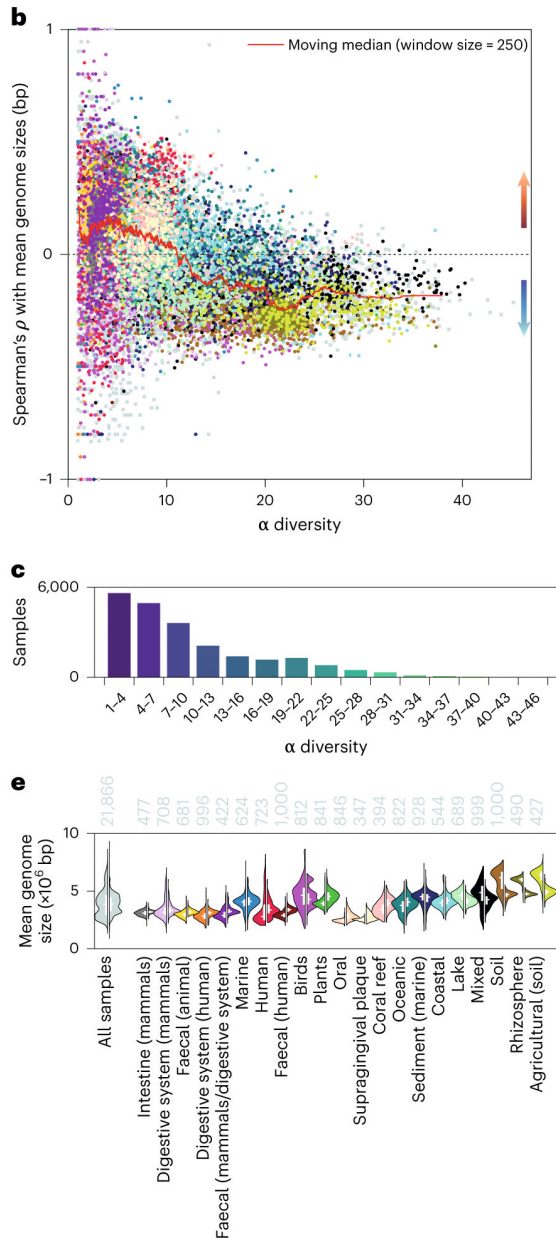


Fig. 4 | Contrasting genomic niche range strategies. **a**, Spearman's rank correlation coefficient (ρ) per sample between SNB and mean genome size on the rank genus plotted on the t-SNE (**Fig. 1d**). Positive values indicate an association with social generalists and negative values indicate an association with social specialists. **b**, ρ as a function of the α diversity of the sample. The colour coding represents annotated biomes (see **e**). **c**, Number of samples per bin of α diversity. **d**, SNB versus genome size on the rank genus. The violin plots show the distribution of genome sizes of species within a genus and the dots represent mean values. The lines depict the mean of linear regression lines between SNB and mean genome size of all samples in a specific bin of α diversity. The shaded areas show the interquartile range of the regression lines. **e**, Violin plots depicting the distribution of mean genome size of the top 25% social specialist taxa (left) and top 25% social generalist taxa (right) within a sample across all samples or those from the annotated biomes with the most samples. The annotated biomes are arranged according to mean α diversity. The numbers at the top of the violin plots show the sample size. The lines within the violin plots show the interquartile range and median. Supporting data relating to **a** and **b** are available in **Supplementary Data 6**.

Niche range strategies of generalists and specialists



5

The pan-genomes of social specialists and generalists

To further characterize generalist and specialist microorganisms, we explored differences in genomic content by dividing all genera into two groups based on SNB (**Fig. 5a**) and performing gene set enrichment analysis (GSEA) (440) on the genus-level pan-genomes. We performed two GSEAs: one with all genera comparing social specialists with social generalists (**Fig. 5b** and **Supplementary Data 8**); and one comparing low-diversity specialists with high-diversity specialists (**Fig. 5c** and **Supplementary Data 9**). The functions enriched in social generalists (false discovery rate (FDR) < 0.1; **Fig. 5b**) included associations with genome fluidity, such as (pro-)phages and plasmid-related functions, highlighting the mechanisms by which they keep an open pan-genome. Other generalist functions reflected an investment in species–species interactions, observation and response to a fluctuating environment. Of the 33 generalist-enriched functions, 13 were related to metabolism, including functions associated with secondary metabolites, such as coenzyme F₄₂₀ (refs. 396,441). We also found quorum sensing and biofilm formation, adhesion, locomotion via the flagellum, and functions concerning the cell envelope and transport across it (S-layers, protein secretion systems and siderophores) to be generalist enriched. Finally, pathogenicity islands could point to opportunistic interactions with eukaryotic host organisms.

Fewer genomic functions were enriched in social specialists than in social generalists (**Fig. 5b**). Specialist-enriched functions include energy-related processes and some specifically related to Cyanobacteria, such as heterocyste formation, which is involved in nitrogen fixation in one evolutionary lineage of the phylum (442). Most functions that were enriched in specialists also occurred in the pan-genomes of many generalist genera (**Fig. 5b**), suggesting that the smaller pan-genome size of social specialists does not involve consistent loss of functions. The absence of widespread specialist functions highlights that there are many ways to be a specialist. There is not a single type of social specialist, but instead many different specialists exist, each with a functional arsenal that fits its niche.

Comparing low-diversity specialists with high-diversity specialists (**Fig. 5c**), we observed several specific metabolic adaptations to these different types of habitat. Half of the 60 enriched functions in the GSEA were related to metabolism. High-diversity specialists have more enriched functions than low-diversity specialists. For example, functions associated with stationary phase, dormancy and persistence are enriched in high-diversity specialists, consistent with slow growth and persistence in soil. Moreover, functions related to lipid metabolism (for example, steroids and hopanoids, (unsaturated) fatty acids, sphingolipids and phospholipids) are enriched in high-diversity specialists. Low-diversity specialists, like social generalists, also contain some functions associated with genome fluidity (for example, transposable elements, (pro-)phages and plasmid-related functions), suggesting that their genomes, although small in size, may still be in flux.

Niche range strategies of generalists and specialists

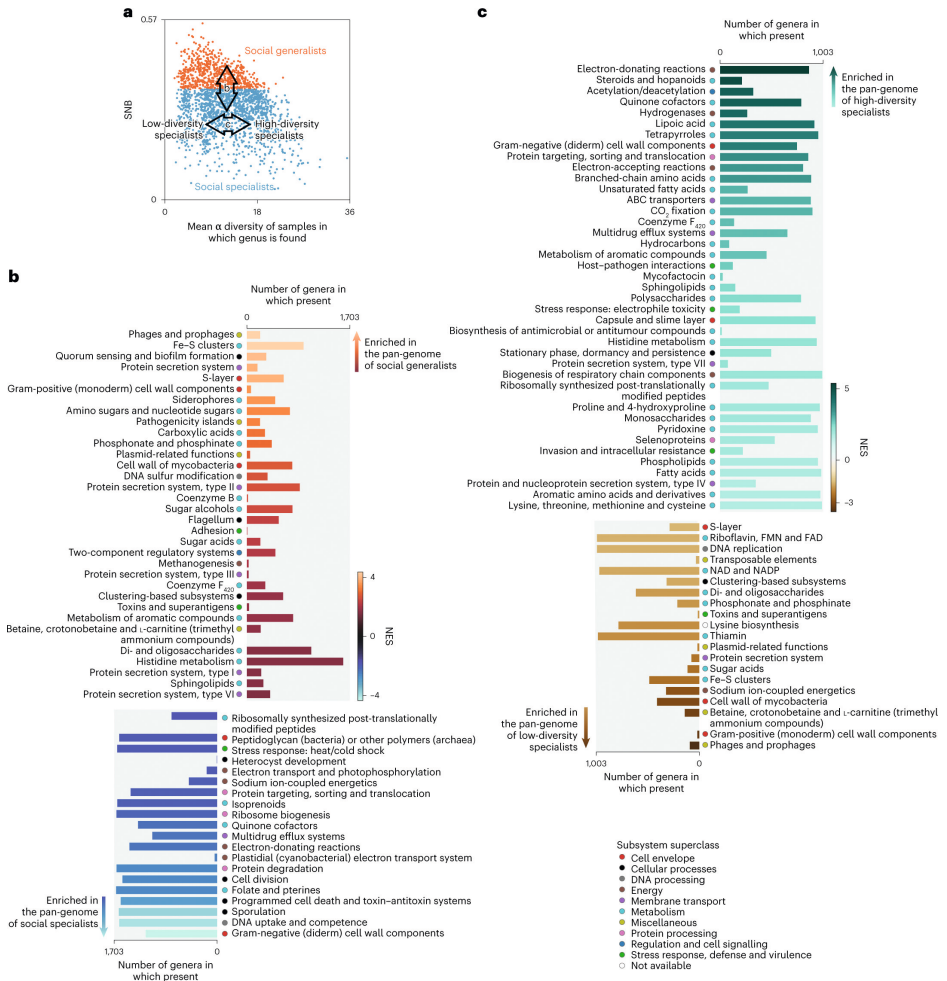


Fig. 5 | Functional characterization of generalists and low- and high-diversity specialists. **a**, Division of genera in social specialists (SNB < 0.35) and social generalists (other). Arrows represent the direction of the GSEA in panels **b** and **c**. **b**, GSEA on the pan-genome of all genera sorted by SNB, on the level of subsystem subclasses. **c**, GSEA on the pan-genome of specialist genera sorted by mean α diversity of the samples in which they are found, on the level of subsystem subclasses. Enriched functions (FDR < 0.1) in **b** and **c** are sorted according to normalized enrichment score (NES). The size of the bars indicates the total number of genera in the dataset having the function. The most enriched functions are in the upper and lower parts of the panels. Full GSEA information, including analyses of other functional universes (subsystem names, gene ontologies and pathways), is available in **Supplementary Data 8 and 9**.

Discussion

We present an SNB score for microbial taxa that is based on the community similarity of the samples in which they occur. Integrating information from over 22,000 samples, SNB represents a global and comprehensive view on niche range throughout the prokaryotic tree of life. With continued and ever-deeper

Chapter 5

sequencing efforts and associated expansion of public databases, the environmental and taxonomic resolution of our picture of the microbial world increases, as does our understanding of the processes shaping microbial niche breadth. In contrast with earlier suggestions, we found that most habitats are dominated by generalists. Specialists occur at low but stable abundances.

Generalist genera are older than specialist genera and have large and open pan-genomes with which they have adapted to different habitats. Individual genome size and SNB are differentially related depending on the diversity of the habitat, with social generalists having larger genomes than social specialists in low-diversity habitats and smaller genomes than social specialists in high-diversity habitats. High-diversity specialists may need a large genetic repertoire as they are continually exposed to many different interaction partners and possibly high environmental variability at small spatial scales. Low-diversity specialists have decreased genome sizes due to the loss of unnecessary functions. Large genomes may thus reflect increased environmental versatility in two different settings. In habitats with low local diversity, generalists are relatively versatile, as they can survive in a range of different communities. In habitats with high local diversity, specialists are relatively versatile, allowing them to persist in their local complex community. Since social generalists and social specialists are dispersed throughout the prokaryotic tree of life, these genomic adaptations have repeatedly occurred and represent fundamental eco-evolutionary processes.

Methods

Sample selection

We downloaded taxonomic profiles deposited in the MGnify microbiome resource (152) on 20 August 2019. MGnify contains taxonomic profiles based on studies that amplify taxonomic marker gene regions (amplicons), shotgun metagenomics studies and shotgun metatranscriptomics studies. We selected taxonomic profiles that were constructed with the pipeline 4.1 version of MGnify and based on the small subunit ribosomal RNA gene, contained at least 50,000 taxonomically annotated reads at the rank superkingdom and had <10% of those reads classified as eukaryotic. We randomly picked one taxonomic profile per sample in cases where there were multiple. To balance the large over-representation of several environments in the database (for example, human gut, soil and ocean), at most 1,000 samples were randomly selected per annotated biome. The 22,518 selected samples (**Supplementary Data 1**) spanned 140 different annotated biomes across a wide geographical range and consisted of amplicon, metagenomic, metatranscriptomic and unknown experiment types (**Supplementary Fig. 1**).

We removed eukaryotic classifications including those classified as mitochondria and chloroplast from the taxonomic profiles, as well as those not classified at the taxonomic rank superkingdom. When relative abundances were used, they were calculated as the number of reads assigned to a taxon divided by the total number of prokaryotic reads, unless otherwise stated in the section ‘**Ecological dissimilarity measures**’.

Ecological dissimilarity measures

We calculated ecological dissimilarity between all sample pairs based on their taxonomic profiles (compositional dissimilarity) at different taxonomic ranks using ten commonly used ecological measures: Aitchison distance; Bray–Curtis dissimilarity; Sørensen–Dice coefficient; Jaccard distance; weighted Jaccard distance; Kendall’s τ_b coefficient; Pearson correlation coefficient; Spearman’s rank correlation coefficient; unweighted UniFrac distance; and weighted UniFrac distance. Some are true distance or dissimilarity measures, whereas others can be readily converted to a scale from 0–1, with 0 being compositionally similar. The three correlation measures were converted to dissimilarity with the formula $0.5 - (\text{coefficient}/2)$ and we used $1 - \text{Sørensen–Dice coefficient}$.

Taxa that were represented by fewer than five reads in the sample were removed before dissimilarity calculations. This was done per rank; therefore, the total number of included reads for a sample could differ depending on the rank considered. To ensure that the pairwise calculations were based on the deepest attainable resolution, we decided on a low absolute read cut-off as opposed to a relative abundance cut-off.

Chapter 5

For each pairwise calculation, we only included taxa that were present in the union of the two samples, thus avoiding the vast scarcity (that is, the presence of zeros in the abundance matrix) often associated with microbiome studies. This scarcity is especially likely because our study compares many different habitats. Those taxa that were only present in one of the samples were given an abundance of zero in the other for all ecological dissimilarity measures except the Aitchison distance, which cannot handle zeros. For the Aitchison distance, a pseudocount was added. This pseudocount differed per pair of samples and was based on the lowest relative abundance that could be reached by an undetected taxon (namely, one read in the sample with the highest number of taxonomically annotated reads). We defined N_1 as the sum of reads represented by the taxa in sample 1 and N_2 as the sum of reads represented by the taxa in sample 2, with $N_1 \geq N_2$. A pseudocount of one read was added to all taxa in sample 1 and a pseudocount of $1/N_1 \times N_2$ reads was added in sample 2.

We calculated the ecological dissimilarity measures at all ranks up to phylum with three different methods for dealing with unknowns in the data. For the UniFrac distances, we used a different method (see below). For approach (i), we considered any taxon on the specific rank. If there was no classification at that rank but the taxon contained lower rank classifications, the first classified rank below was used. If there was no classification at the specific rank and no lower rank classification, we used the first classified rank above. For approach (ii), we exclusively considered taxa that were classified at the specific rank. Taxa that were classified at lower or higher ranks alone were removed. For approach (iii), we treated taxa that were not classified at the specific rank but did have lower rank classifications the same as in approach (i). If taxa had no classification at the specific rank or at a lower rank, we used the first classified rank above, unless the taxon was present in both samples. In this case, the taxon was removed. The rationale is that for these taxa it is unknown whether they are the same or different for the rank of interest.

UniFrac distance takes relatedness between taxa into account. We used distance across the taxonomic tree as a measure for relatedness, with the distance between successive ranks defined as 1. We used the EMDUniFrac implementation (370), which is suited for samples with many unknowns because it allows for the placement of taxa at different ranks in the tree. UniFrac distances were calculated at the ranks species, family and class. For taxa that had no classification at the specific rank but did have a lower rank classification, we used an artificial classification based on the first classified rank below, ensuring uniqueness of the taxon and appropriate distance to the root. Taxa that did not have a classification at the specific rank or a lower rank were placed at the first classified rank above in the tree.

For the ecological dissimilarity calculations that took the number of reads into account, the numbers of reads per taxon were converted to relative abundance values by dividing by the sum of reads represented by the taxa in the sample (for example, for Bray–Curtis dissimilarity calculations and the addition of pseudocounts before Aitchison distance calculations). As explained above, the taxa considered in a sample may have differed per method of dealing with unknowns, and so may the relative abundance of a taxon.

Because the taxonomic profiles contain many unknowns at lower ranks, pairwise comparisons are sometimes based on only few taxa. For each rank and method of dealing with unknowns, samples were removed that did not contain any taxon at that rank. If the pairwise comparison was based on one taxon, we set the dissimilarity to 0. We removed samples from the correlation measures whose correlation coefficient with itself could not be calculated.

Permutational multivariate analysis of variance

Permutational multivariate analysis of variance pseudo F statistics were calculated for all ecological dissimilarity measures with the scikit-bio version 0.5.5 implementation (<http://scikit-bio.org/>). As predefined groups, we used either the annotated biomes of the samples or their experiment types. P values were based on 99 random permutations and we calculated the coefficient of determination (R^2) with the formula:

$$R^2 = \frac{F}{F + \frac{N - G}{G - 1}}$$

where F is the pseudo F statistic, N is the sample size and G is the number of groups.

Diversity measures

Diversity measures were calculated for the subset of samples belonging to an annotated biome and in which a taxon was found. If a subset contained fewer than three samples, it was excluded from analysis. Taxa were removed whose relative abundance was less than 1/10,000. We used approach (ii) to deal with unknowns, as explained in the section ‘**Ecological dissimilarity measures**’.

Zeroth-order α diversity (that is, richness—the mean number of taxa found in a set of samples) was calculated for all ranks. Zeroth-, first- and second-order α diversity (${}^qD_\alpha$) and β diversity (${}^qD_\beta$) were calculated on the taxonomic rank order and based on relative abundances, with ${}^qD_\beta$ defined as the total effective number of taxa (${}^qD_\gamma$) divided by ${}^qD_\alpha$. ${}^qD_\gamma$ was calculated based on the summed relative abundance of the

Chapter 5

individual samples. For the first- and second-order diversity measures, two samples were excluded that did not contain any classification at order rank after the relative abundance threshold.

When the terms α and β diversity are used, we refer to first-order diversity measures on the rank order unless otherwise stated. For a more in-depth discussion of these diversity measures, see ref. 443. The Shannon entropy and Gini-Simpson index, which were used for diversity calculations, were calculated with the scikit-bio version 0.5.5 implementation.

Local dominance and Shannon entropy across samples

For each taxon, we calculated local dominance and Shannon entropy. Local dominance was defined as the mean relative abundance across all samples in which the taxon was found. Shannon entropy (base e) was used as a measure for the randomness of its relative abundances across these samples (N) and was normalized by dividing by $\ln[N]$.

SNB definition

SNB was defined as the mean of the pairwise dissimilarity between the samples in which a taxon was found, n :

$$\text{SNB} = \frac{\sum_{i=1}^n \sum_{j=1}^n d_{ij}, i \neq j}{n^2 - n}$$

with the dissimilarity d_{ij} based on the Spearman's rank correlation coefficient ($0.5 - (\rho/2)$) on the rank order with method (ii) for dealing with unknowns (see the section 'Ecological dissimilarity measures'). A taxon was considered present in a sample if it had a relative abundance of at least 1×10^{-4} . A taxon with a low score was thus found in samples with similar taxonomic profiles (social specialists) and a taxon with a high score was found in dissimilar samples (social generalists). Taxa that were present in fewer than five samples were removed from analyses unless otherwise stated.

To benchmark SNB, we also calculated SNB with different detection thresholds of 1×10^{-3} and 1×10^{-5} . SNB was moreover calculated for imaginary taxa (iSNB) that were present in all samples from a given annotated biome, in half of the samples from a given annotated biome (100 random permutations per annotated biome) and in all samples from pairs of annotated biomes. In addition, iSNB was calculated for randomly picked sets of samples of equal size to the encountered taxa (100 random permutations per sample size). Lastly, we calculated SNB for real taxa only based on the marine and human hierarchical subsets of the samples.

Selection of genomes

We downloaded all genomes from the PATRIC genome database (185) that had a quality marked as good and were not plasmids on 14 November 2019. We only included genomes for which we had a valid taxonomy ID in our NCBI taxonomy (342) files that were downloaded on the same date. PATRIC contains identical genomes with different identifiers. We identified replicate genomes based on concatenated DNA sequences and concatenated sorted DNA sequences and removed all but one. In cases where identical genomes had different taxonomic annotations, all were discarded. Completeness and contamination estimates were generated with CheckM version 1.0.7 (ref. 51) in the lineage-specific workflow. We excluded genomes for which $\text{completeness} - 5 \times \text{contamination} < 70$. The final selection consisted of 225,101 prokaryotic genomes representing 34,304 species, from both cultures and environmental sequencing projects (**Supplementary Data 5**).

Inferences about (pan)-genome size and genomic functions (see below) and number of subtaxa were made at all taxonomic ranks and reconstructions at higher ranks were based on lower-rank taxa in the PATRIC database that were not always present in the MGnify dataset.

Genome size estimates and GC content

Genome size (in number of base pairs and number of coding sequences) and GC content were obtained from the metadata in the PATRIC database. We corrected genome size estimates by taking completeness and contamination into account, via multiplication with a scaling factor s :

$$s = \frac{100}{\text{completeness} + \text{contamination}}$$

For each measure including the number of coding sequences per million base pairs, we reconstructed species by averaging the values of its genomes. For higher ranks, mean values were calculated for all species belonging to the taxon. Some high-ranking taxa contain many low-ranking taxa from the same taxonomic group, such as genus or family. To correct for this over-representation and possible skew towards the values of these taxa, we also calculated mean values by averaging over the taxonomy at all ranks (taxonomy-corrected values). For example, the size of a family is the mean of the sizes of its genera and the size of a genus is the mean of the sizes of its species. These values were calculated for the ranks family and higher and are available in **Supplementary Data 5**.

Genome functions

Functional profiles of the genomes were created based on the PATRIC annotations of coding sequences for three functional universes: subsystems, gene ontologies and pathways. For gene ontologies, the profiles were based on the exact terms found in the annotation files, whereas for subsystems we made different profiles for the name and subclass level of the hierarchy. Genomes with ≤ 20 unique functions were discarded from further analyses for subsystem names, gene ontologies and pathways. For the included list of genomes in each analysis, see **Supplementary Data 5**.

Functional genome size was defined as the number of unique functions present in a species. A function was considered present in a species if at least 50% of the genomes with this species annotation contained it in the PATRIC database. Mean functional genome sizes were calculated for all taxa, as well as the standard deviation. Pan-genomes were defined at all ranks as the total set of unique functions present in the genomes of a taxon.

Pan-genomes can be open or closed, meaning that they can be more or less susceptible to changes in gene content (106,107). We devised a score that represents pan-genome openness for all ranks higher than species. Pan-genome openness was defined as the total pan-genome size divided by the mean pan-genome size of a species. Because taxa with many subtaxa tend to have large pan-genomes, we also calculated pan-genome features for a random subset of three daughter species (1,000 random permutations per taxon) to correct for this effect of taxonomy. This measure thus reflects how many functions are on average added to the pan-genome by including two more species. Permuted measures were calculated for all taxa with at least three species.

GSEA of pan-genomes

To detect functions that were significantly enriched in social specialists and generalists, we deployed GSEA (440) based on the pan-genomes of genera. We performed a GSEA on all genera sorted by SNB to compare specialists with generalists, and on specialist genera ($SNB < 0.35$) sorted by α diversity to compare low-diversity specialists with high-diversity specialists. We used the classical Kolmogorov–Smirnov statistic for the enrichment score ($p = 0$). Enrichment score normalizations and P values were based on 100,000 random permutations of the gene set. Multiple hypothesis correction was carried out via the FDR as suggested in ref. 440. GSEA computations were done with a modified version of the algorithm.py script from GSEAPy version 0.7.3.

Growth rate and clade age estimates

We downloaded the maximal growth rate predictions of RefSeq genomes from the EGGO database (187) and defined species- and genus-rank maximal growth rates as the mean of their genomes.

Clade ages were based on the TimeTree database (435). Times to the first and last common ancestor were extracted from the species rank phylogenies of bacteria and archaea using ete3 version 3.1.1 (ref. 444).

Software packages used for calculations and visualizations

Calculations were done with the Python 3 standard library, NumPy (445) and the SciPy library (446) unless otherwise stated. Visualizations were done with Python 3 and Matplotlib (447) in JupyterLab (<https://jupyter.org/>), with the use of NumPy, pandas (448) and seaborn (<https://seaborn.pydata.org/>). Principal coordinates analysis (PCoA) was performed with the scikit-bio version 0.5.5 implementation. t-SNE was performed with scikit-learn version 0.21.3 (ref. 449). Samples were drawn on the world map with Cartopy version 0.17.0 (<http://scitools.org.uk/cartopy/>) using Natural Earth data (<https://www.naturalearthdata.com>). The taxonomic tree was visualized with iTOL (374) and the hierarchical tree of annotated biomes was visualized with ete3 version 3.1.1 (ref. 444).

Data availability

All of the data analysed during this study are included in this article and **Supplementary Data 1–9** or available in public repositories. The selected samples from the MGnify resource (<https://www.ebi.ac.uk/metagenomics/>) are described in **Supplementary Data 1**. The cleaned taxonomic profiles based on these data are available in **Supplementary Data 2**. The selected genomes from the PATRIC database (<https://www.bv-brc.org/>) are described in **Supplementary Data 5**. Measures derived from the PATRIC genomes and the EGGO (<https://github.com/jlw-ecoevo/eggo>) and TimeTree (<https://timetree.org/>) databases are available in **Supplementary Data 3**.

Code availability

All of the code used for this manuscript is available from Zenodo at <https://doi.org/10.5281/zenodo.7651594>. A stand-alone script to calculate SNB is available from https://github.com/MGXlab/social_niche_breadth.

Acknowledgements

We thank J. K. van Amerongen for technical support. This work is funded/supported by the European Research Council (Consolidator Grant 865694: DiversiPHI to

Chapter 5

B.E.D.), the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy (EXC 2051; Project-ID 390713860 to B.E.D.) and the Alexander von Humboldt Foundation in the context of an Alexander von Humboldt Professorship founded by the German Federal Ministry of Education and Research (to B.E.D.).

Funding

Open access funding provided by Friedrich-Schiller-Universität Jena.

Contributions

All authors conceived of and designed the study and analysed and interpreted the results. F.A.B.v.M. carried out the experiments and wrote the manuscript, with substantial contributions to the writing from P.H. and B.E.D.

Supplementary Information

Supplementary Results and discussion

A cross-biome dataset

We compiled a diverse set of 22,518 environmental sequencing samples from 592 studies, spanning 140 annotated biomes across a wide geographical range based on the MGnify resource (152) (**Supplementary Fig. 1, Fig. 1, Supplementary Data 1**, see ‘**Methods**’ for selection criteria). MGnify uses standardised pipelines to process environmental sequencing datasets, allowing for the comparison of samples across a wide range of different environments, studies, and experiment types. We only included taxonomic profiles that were constructed with the 4.1 pipeline version of MGnify and based on the small subunit (SSU) rRNA gene. Because these taxonomic classifications are all based on queries with the same rRNA gene models (152) that can be found in both targeted amplicon and shotgun studies, we included sequencing samples from amplicon, metagenomic, metatranscriptomic, and even the elusive ‘unknown’ experiment types.

Samples from similar annotated biomes clustered together based on microbial composition, despite the samples coming from vastly different locations and study designs including experiment type (**Fig. 1d, Supplementary Figs. 2a and 3**). Samples were mostly separated by association to a vertebrate host versus free-living habitats, and saline versus non-saline habitats (173,403–405) (**Supplementary Figs. 4a and 2b**). An exception are fish, whose foregut and intestinal microbiomes were more similar to microbiomes from aquatic habitats than to those in other vertebrate guts (**Supplementary Fig. 3**). Within free-living habitats, saline samples differed from non-saline samples including soils. Aquatic sediments resembled their saline or non-saline provenance. In line with earlier findings (404), invertebrate-associated samples clustered together with free-living samples and not with vertebrate-associated samples. Invertebrates in our dataset include sponges, molluscs, Cnidaria, and Echinodermata whose internal microbiomes are in direct or semi-direct contact with the surrounding environment, and marine arthropods like *Calanus finmarchicus* (**Supplementary Data 1**).

Host-associated samples typically had lower taxa richness and α diversity than free-living samples (**Supplementary Figs. 4b,c, 2c–e, and 5a**). Rhizospheres have been shown to resemble soils in terms of richness (173) and we also observed this (**Fig. 1e**). Notably, annotated biomes with a high mean α diversity had a low β diversity, while annotated biomes with low mean α diversity had either low or high β diversity (**Fig. 1f**).

While taxa richness increases towards low taxonomic ranks, richness in the microbiomes was lower at the species rank and in some cases also at the genus rank than at higher ranks (**Fig. 1e**). This anomaly reflects the still low classification rate of the organisms in natural environments at the species and genus ranks. In addition, low rank classifications more easily fall below our detection limit of 1/10,000 than a higher rank classification whose abundance is the sum of its lower ranks.

Quantifying microbial social niche breadth

To quantify the range of habitats in which a microbial taxon is found we formulated a social niche breadth score that is data-driven and independent of human-defined biome annotations. To do this, we calculated the dissimilarity between taxonomic profiles of pairs of samples (see below), and defined the social niche breadth (SNB) of a taxon as the mean pairwise dissimilarity between the microbial communities of the samples where it is found. Thus, taxa that always occur in samples with similar microbial composition have a low SNB (social specialists), whereas taxa that occur in dissimilar samples have a high SNB (social generalists).

Benchmarking microbiome dissimilarity measures

To arrive at a quantitative niche breadth definition that optimally reflects annotated biomes as recognised by the research community, we benchmarked 150 different microbiome dissimilarity measures for their correspondence with the biome annotations of the underlying datasets. Many dissimilarity measures have been proposed in ecological literature, each with their own merit. For example, the Aitchison distance (450) is relevant for microbiomes because it takes the inherent compositionality of sequencing data into account (214), while the UniFrac distance considers phylogenetic (or taxonomic) information (451). Measures that take relative abundances into account (weighted measures) better reflect quantitative relationships between taxa than unweighted measures, but put less emphasis on low abundant taxa that might be instrumental for ecosystem functioning (452). Other factors of importance include the taxonomic rank of comparison, as well as the method for handling unknowns (sequences that cannot be classified).

We calculated ten ecological dissimilarity measures between all 253,518,903 sample pairs at six taxonomic ranks and with four different methods for dealing with unknowns (see ‘**Methods**’), totalling 150 different measures (**Supplementary Fig. 6a–d**). The dissimilarity measures are based on: Aitchison distance; Bray–Curtis dissimilarity; Sørensen–Dice coefficient; Jaccard distance; weighted Jaccard distance; Kendall’s τ_b coefficient; Pearson correlation coefficient; Spearman’s rank correlation coefficient; unweighted UniFrac distance; and weighted UniFrac distance. Most measures are true distance or dissimilarity measures whereas the correlation and Sørensen–Dice coefficients were converted to a scale from 0 to 1, with 0 being compositionally similar (see ‘**Methods**’).

Since the MGnify taxonomic annotation pipeline depends on sequence similarity to a reference database and environmental sequencing studies contain both known and unknown taxa (47), many reads in a sample are not classified at all ranks (superkingdom, phylum, class, order, family, genus, species). Furthermore, taxonomy is incomplete, with lower rank classifications sometimes present but intermediate ones missing. For example, taxonomy of the phylum Cyanobacteria is debated (453) and currently only one order has a taxonomic annotation at rank class in NCBI taxonomy. Likewise, the genus *Methyloceanibacter* of the order Rhizobiales does not have a family annotation yet (454). For these reasons, we calculated the ecological dissimilarity measures with four different methods for dealing with unknowns (see ‘Methods’). Approach (i) substitutes unknowns at the considered rank with known lower or higher rank classifications. Whereas this approach has the benefit that all reads are considered, it potentially clusters different unknown taxa under the same higher rank umbrella, or may divide a would-be single taxon into multiple lower rank groups. Approach (ii) only compares reads that are annotated at the considered rank. This allows for a robust taxonomic comparison when samples contain unknowns, at the expense of using all reads. Approach (iii) is similar to approach (i) but removes unknowns that do not have lower rank classifications if they are shared between the two samples. The rationale is that for these taxa it is unknown if they are the same or different for the rank of interest. Finally, for the UniFrac distances we allow for placement of unknown taxa at different ranks in the tree (approach (iv)), with lower rank classifications artificially placed at the considered rank, and higher rank classifications kept. This allows for all reads to be considered, and has similar trade-offs compared to approach (i). In addition, samples with many unknowns that only have higher rank classifications can have an artificially short pairwise distance.

We scored how well the 150 pairwise dissimilarity measures represented the annotated biomes using PERMANOVA, and found that these groups are best represented by a dissimilarity measure based on an inverted Spearman’s rank correlation coefficient ($0.5 - (\rho/2)$) at the taxonomic order rank while ignoring unknowns (**Supplementary Fig. 6a–d**). We thus use the Spearman’s rank-based microbiome dissimilarity score to quantify SNB, while noting that another choice would not qualitatively affect our results, as social niche breadth scores based on six alternative ecological dissimilarity measures spanning the four different methods of dealing with unknowns showed a high correlation with the one we selected (**Supplementary Fig. 6e**). In addition, we investigated the robustness of our results to our choice for the mean pairwise dissimilarity by comparing it to the median and third quartile, and found high correlations ($\rho = 0.977$ ($P = 0.000$) and $\rho = 0.967$ ($P = 0.000$), respectively; **Supplementary Fig. 7**). In agreement with our premise that social niche breadth should reflect the co-occurrence of a taxon with other taxa, our SNB score is strongly negatively correlated with the fraction of shared taxa between samples ($\rho = -0.878$ ($P = 0.000$); **Supplementary Fig. 8**).

Social niche breadth robustly reflects community heterogeneity

Robustness to sampling bias

To investigate the robustness of the SNB score to sampling bias, we further calculated social niche breadth for imaginary taxa (iSNB) that occur in all samples of an annotated biome. This showed a strong association between iSNB and the β diversity of an annotated biome (**Supplementary Figs. 9a and 4d**). Thus, taxa that are ubiquitously present in a very heterogeneous annotated biome would have a high SNB, while taxa that are ubiquitous in an annotated biome with a low heterogeneity would have a low SNB. Importantly, iSNB does not depend on the number of available samples for a given annotated biome (**Supplementary Fig. 9b**). Random subsets of samples from single annotated biomes showed low variation in iSNB (**Supplementary Figs. 9c and 4e**), implying robustness of SNB to sporadically missed presence of a taxon in a sample, but standard deviation increased when the number of samples becomes very low (**Supplementary Figs. 9c and 4e**). For this reason, we use caution when interpreting SNB of rare taxa, i.e. that are present in only a few samples, and exclude taxa that are present in less than 5 samples from our analyses. iSNB calculated for imaginary taxa that occur across two annotated biomes revealed low iSNB for highly similar annotated biomes like different human oral sites (**Supplementary Fig. 10**, top left corner). In addition, even though presence in a single annotated biome with low β diversity results in a low iSNB, presence in two different annotated biomes with low β diversity still results in a high iSNB if they are very different from each other (**Supplementary Fig. 10**).

For real microbial taxa, we observed a striking independence of SNB on the number of samples in which a taxon is found (**Fig. 2a–c**). Some moderately ubiquitous taxa are exclusively present in similar samples (low SNB), whereas many uncommon taxa are present in very different samples (high SNB). That some specialist taxa are still quite ubiquitous can partly be explained by the overrepresentation of some environments in our microbiome dataset, even though we selected a maximum of 1,000 samples per annotated biome (**Supplementary Fig. 1b**). Further, we observed widely different SNB for taxa encountered in the same number of annotated biomes (**Fig. 2b**), pointing to differences in community dissimilarity between annotated biomes and heterogeneity within annotated biomes as discussed above. Nonetheless, many taxa that are found in only a few samples have a low SNB (**Supplementary Fig. 11a**), indicating that the samples where they are found are similar in composition, and suggesting that rare taxa are often social specialists. The most cosmopolitan taxa are all social generalists (high SNB) (**Fig. 2a–c**), since they are present in many dissimilar samples.

We also calculated SNB based only on the subsets of samples belonging to all human and marine annotated biomes, and found good correlations with the SNB scores based on all samples ($\rho = 0.546$ ($P = 0.000$) and $\rho = 0.662$ ($P = 0.000$), respectively;

Supplementary Fig. 12), implying that the sampling of annotated biomes does not strongly affect our calculated SNB values and suggesting that our general results would be qualitatively similar if different habitats were sampled.

Robustness to detection limit

The detection limit of taxa in environmental sequencing datasets is an important parameter that could influence SNB, as higher detection thresholds obscure our view of rare taxa (452) and decrease the number of samples and habitats in which taxa are found. To assess this effect, we calculated SNB with a ten-fold higher (1×10^{-3}) and ten-fold lower (1×10^{-5}) detection threshold than used for our main results (**Supplementary Fig. 13**), and observed shifts in SNB as expected; overall, taxa become more specialist with a higher and more generalist with a lower detection threshold. Importantly, the list of taxa ranked by SNB was consistent (**Supplementary Fig. 13c,g**), especially if uncommon taxa were excluded (**Supplementary Fig. 13d,h**). In addition, exclusion of taxa that have very low relative abundance across samples does not change the distribution of SNB (**Supplementary Fig. 11b**).

Robustness to experiment type

SNB is based on the presence of taxa in sequencing samples (22,518 in total) that are coming from different experiment types: amplicon (15,790 samples), metagenomic (1,097 samples), metatranscriptomic (13 samples), and 'unknown' (5,618 samples). The standardised taxonomic pipeline of MGnify allows for a comparison of these different experiment types which maximises the number of habitats on which the SNB score is based. The taxonomic classifications that we use are all based on the same SSU rRNA gene models that are queried in the samples irrespective of experiment type. We moreover selected analyses with at least 50,000 taxonomically annotated reads, ensuring that the targeted genes are there.

To further investigate the effect of including these different experiment types on SNB and our results, we first performed a PERMANOVA analysis with experiment type as the predefined groups, which showed very low R^2 values (**Supplementary Fig. 6a–d**), implying a low impact of experiment type on the ecological clustering. To confirm that SNB does not depend considerably on the selection of experiment type, we calculated SNB for all taxa only based on the subsets of samples from the different experiment types (**Supplementary Data 3**). The taxa that were present in at least 5 samples in these subsets had similar SNB scores to their original SNB scores that are based on all samples, according to their rank order distribution: amplicon, 4,540 taxa, $\rho = 0.950$ ($P = 0.000$); metagenomic, 1,373 taxa, $\rho = 0.604$ ($P = 0.000$); metatranscriptomic, 530 taxa, $\rho = 0.677$ ($P = 0.000$); and 'unknown', 2,663 taxa, $\rho = 0.784$ ($P = 0.000$). Thus, taxa that are specialists or generalists in all samples also tend to be specialists or generalists in the experiment type subsets, respectively, justifying our decision to include these different data types in this global analysis.

Next, we investigated whether we would have obtained qualitatively different conclusions if we would have used only samples from one experiment type (**Supplementary Fig. 14**). Importantly, our observations that generalist genera dominate local communities and have shorter doubling times than specialist genera are consistent across experiment type. Only the metagenomic samples do not show a higher variability of relative abundance across samples for social generalists than for social specialists but instead the opposite correlation, which may be a habitat specific observation as most metagenomic samples are from animal-associated environments (**Supplementary Fig. 15**). The observations that social generalists have more diverse genomes than social specialists (measured in the standard deviation of their genome size) and a larger and more open pan genome are also consistently observed when basing our analyses on samples from single experiment types. In addition, clade age of social generalists is also older than that of social specialists across experiment type, with the exception of the FCA clade age in the metagenomic and metatranscriptomic subsets.

The most important observation that qualitatively differs when using only samples from a single experiment type as opposed to using all samples combined is the correlation between SNB and genome size. When using all samples we found no consistent relation between SNB and genome size. In contrast, the amplicon and metagenomic experiment types show a small positive correlation (i.e. social generalists have larger genomes than social specialists), whereas the metatranscriptomic and 'unknown' experiment types show a small negative correlation (i.e. social specialists have larger genomes than social generalists). This discrepancy between experiment types can be largely attributed to differences in sampled habitats. The amplicon and metagenomic datasets contain relatively many animal-associated low α diversity samples, where social generalists tend to have larger genomes than social specialists (**Supplementary Fig. 15**). In contrast, the 'unknown' dataset has relatively many free-living high α diversity samples, where social specialists tend to have larger genomes than social generalists (**Supplementary Fig. 15**). This is consistent with the habitat specific relation between SNB and genome size that we observed earlier. However, most low α diversity samples in the 'unknown' dataset do not show a positive correlation between SNB and genome size (**Supplementary Fig. 15**), in which they differ from the dataset based on all samples, and the metagenomic and amplicon samples. For example, the relatively low α diversity plant-associated samples of the 'unknown' dataset do not show a consistent correlation between SNB and genome size. This illustrates that the α diversity of a sample is only a proxy for habitat type and does not fully represent a taxon's niche. Lastly, all metatranscriptomic samples show a negative correlation between SNB and genome size, regardless of α diversity (**Supplementary Fig. 15**). Although the low number of low α diversity samples in this subset prevents strong conclusions, comparing RNA-based samples (representing the

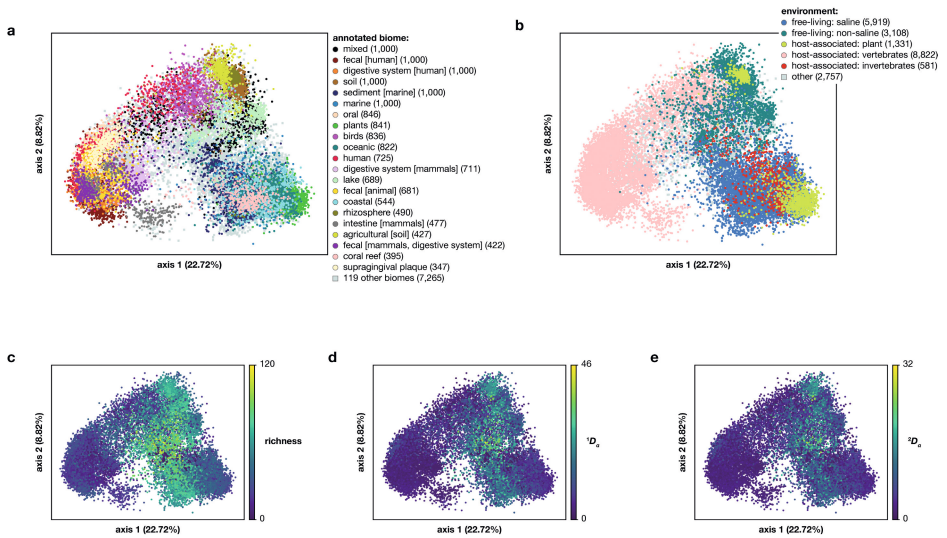
Niche range strategies of generalists and specialists

active biomass) to DNA-based samples (representing all biomass) will be interesting for future research.

In conclusion, incorporation of different experiment types for the calculation of SNB does not fundamentally affect our general conclusions.

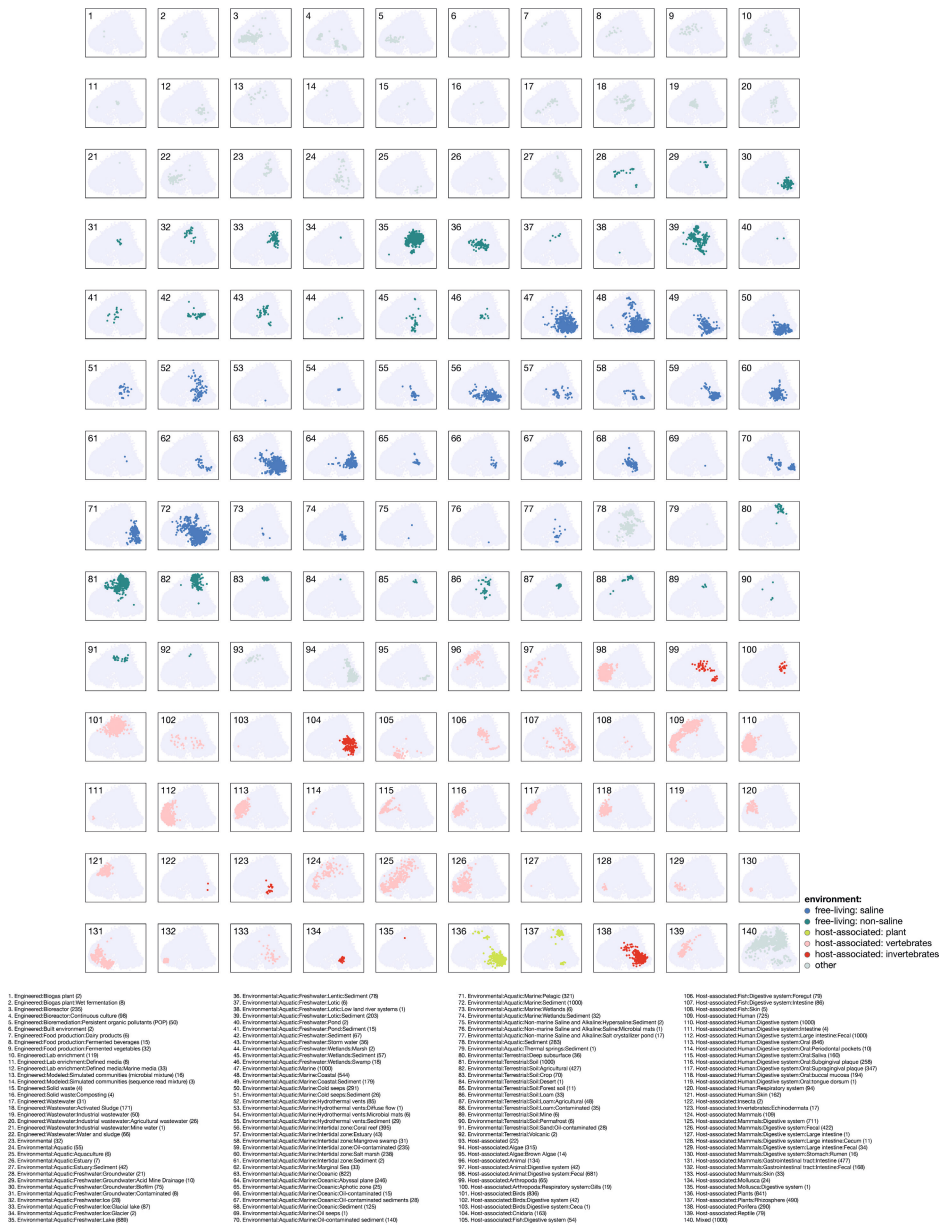
Together, the results presented in this section show that the SNB score is robust to the specific community dissimilarity measure, sampling bias, detection threshold, and experiment type, suggesting that our results represent a meaningful quantification of microbial niche range.

Niche range strategies of generalists and specialists



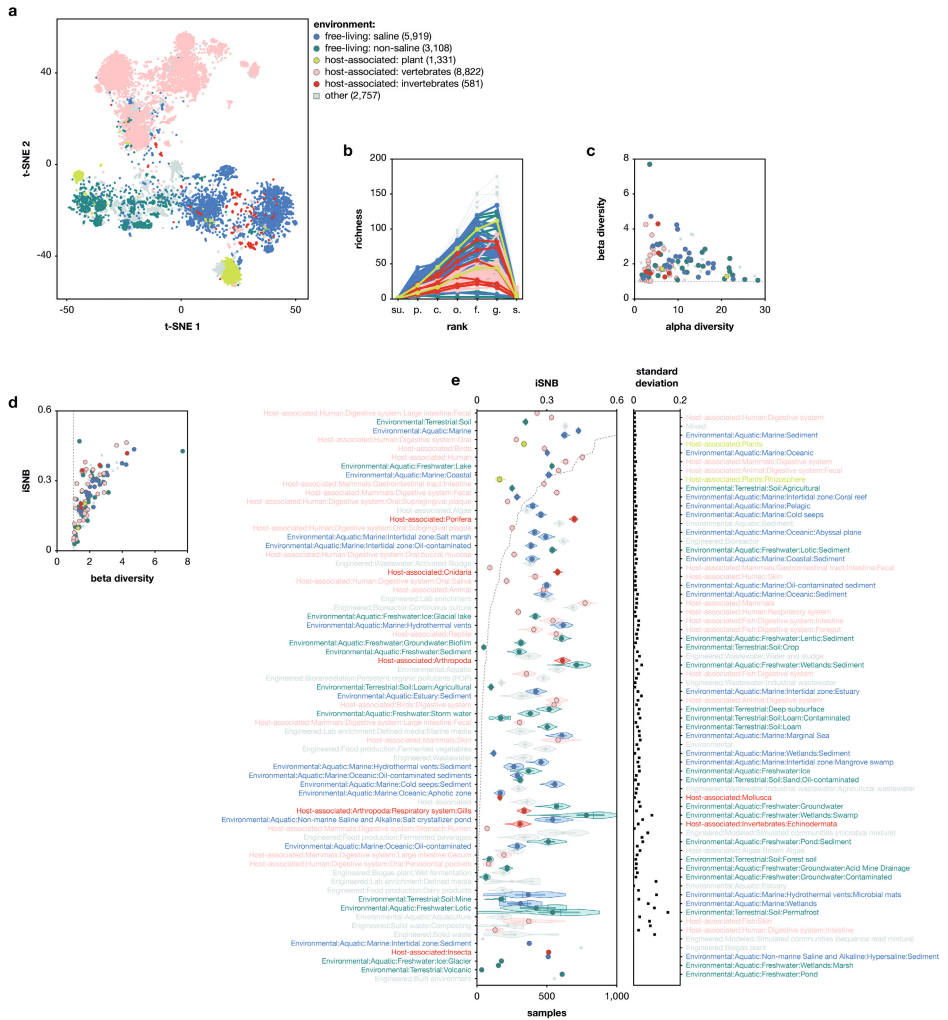
Supplementary Fig. 2 | PCoA visualisation of 22,518 microbiomes. **a**, Samples from similar annotated biomes cluster together based on taxonomic profile, with the same ecological dissimilarity measure used as for SNB, namely Spearman's rank correlation coefficient ($0.5 - (\rho/2)$) of known taxa at rank order. **b**, Samples are separated by host-association and salinity. Invertebrate-associated communities cluster together with free-living communities and not with vertebrate-associated communities. See **Supplementary Fig. 3** for the division of annotated biomes in free-living and host-associated. See **Supplementary Fig. 4a** for a t-SNE visualisation of the same data. **c-e**, α diversity of samples on the rank order for three different diversity measures, zeroth order diversity (richness) (**c**), first order diversity ($e^{\text{Shannon index}}$) (**d**), and second order diversity (inverse Simpson index) (**e**).

Chapter 5



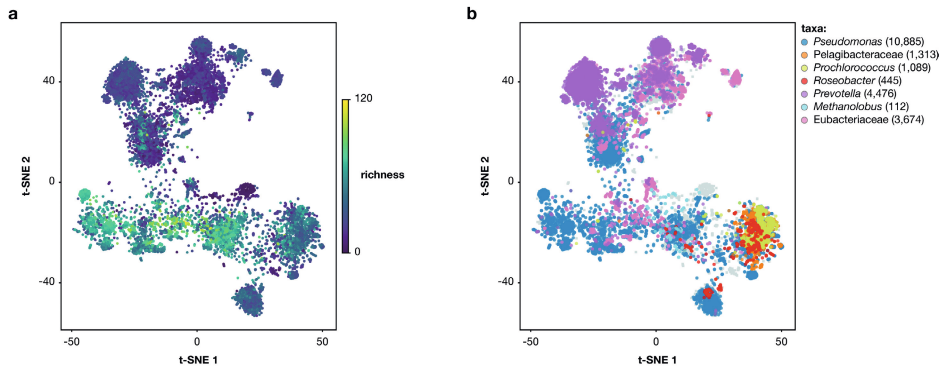
Supplementary Fig. 3 | Location of all annotated biomes on the PCoA. Annotated biomes are coloured according to their environment type, free-living or host-associated. The 'Animal' biomes contain samples from frog, iguana, cats, dog, pigeon, fish, rat, cow, pig, mouse, poultry, and mammalia-associated habitat (Supplementary Data 1), and are thus included in the vertebrate environment type.

Niche range strategies of generalists and specialists



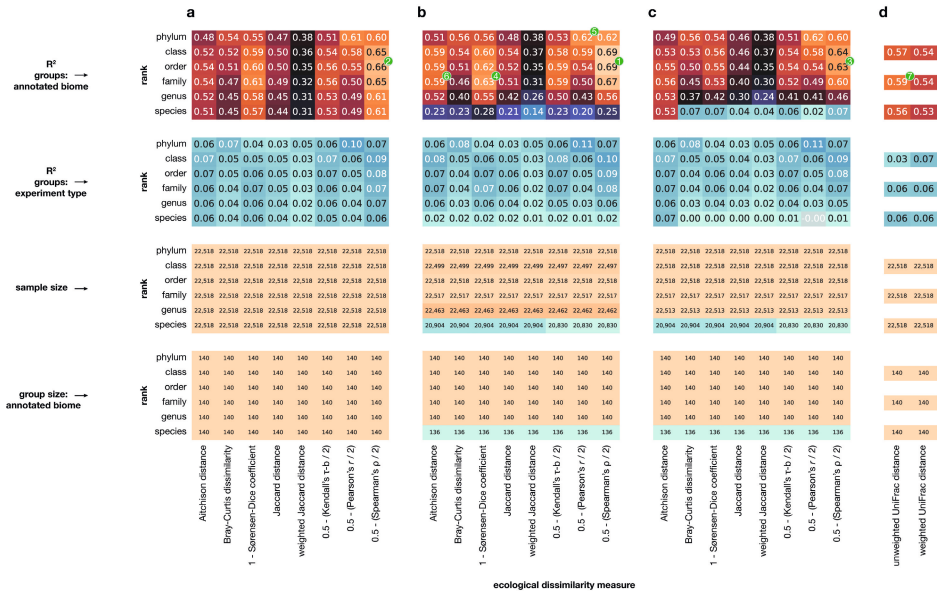
Supplementary Fig. 4 | Remake of some figures colour-coded according to environment type. a, Idem to Fig. 1d. b, Idem to Fig. 1e. c, Idem to Fig. 1f. d, Idem to Supplementary Fig. 9a. e, Idem to Supplementary Fig. 9c. Figures are identical except for the colour-coding.

Chapter 5

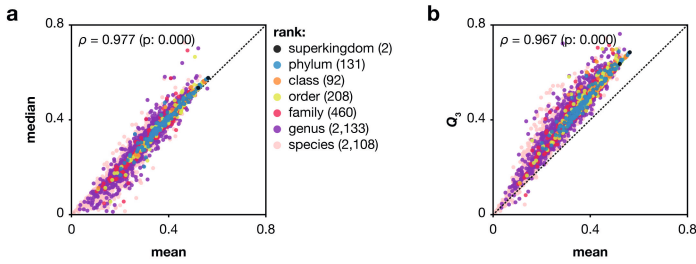


Supplementary Fig. 5 | Richness of samples and presence of some taxa visualised on the t-SNE of Fig. 1d. a, Zeroth order α diversity (richness) on the rank order of samples. b, Presence of some microbial taxa.

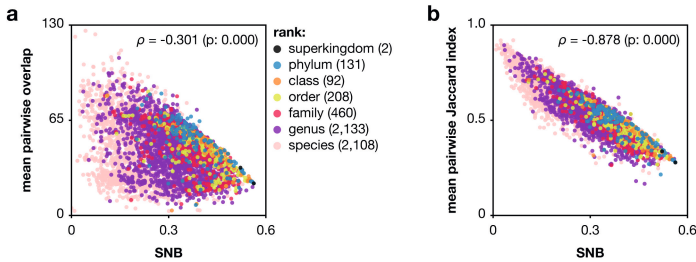
Niche range strategies of generalists and specialists



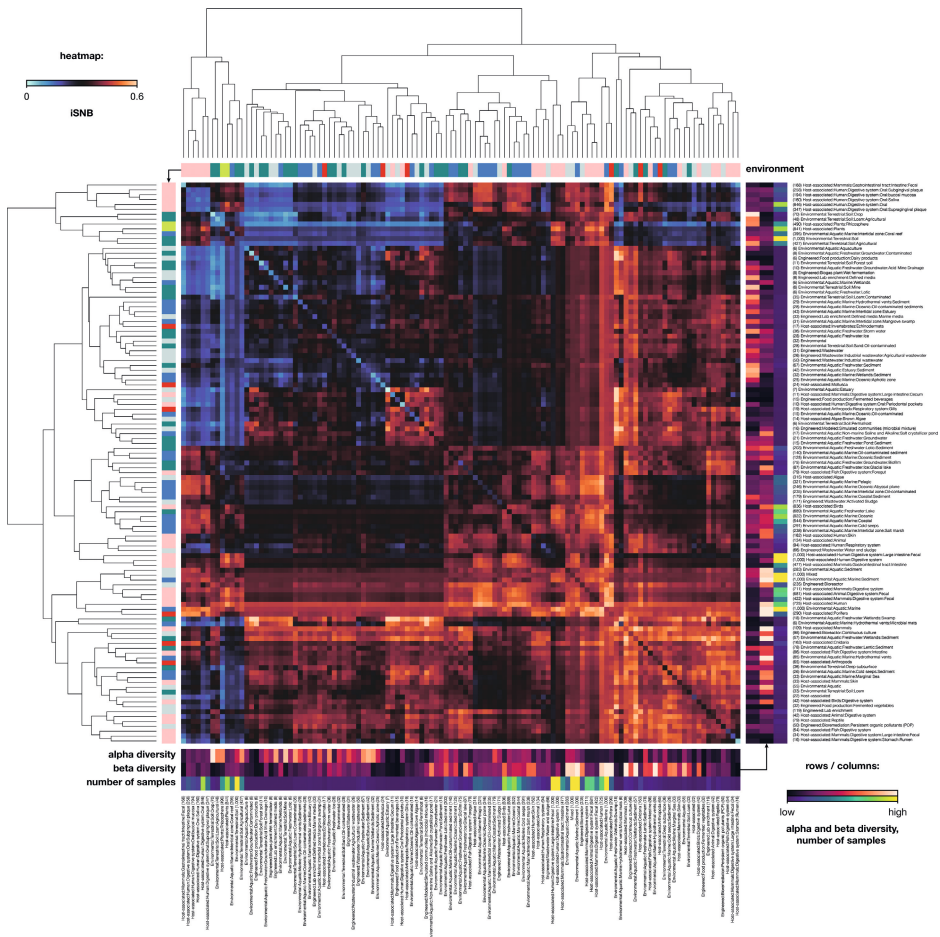
Supplementary Fig. 6 | Selection of ecological dissimilarity measure. PERMANOVA results with annotated biomes or experiment type (amplicon, metagenomic, metatranscriptomic, or 'unknown') as predefined groups (see **Supplementary Fig. 1**). Dissimilarity between any two samples was calculated for ten different dissimilarity measures and three different methods for dealing with unknowns. See **'Methods'** for a description of these methods. **a**, Approach i. **b**, Approach ii. **c**, Approach iii. **d**, UniFrac distances. The sample size and group size for the annotated biomes analyses are indicated which can be smaller than the total number of samples or annotated biomes in the dataset, respectively, if samples do not contain enough taxa for a pairwise comparison. The group size for the experiment type analyses was four. **e**, Correlations between niche breadth of taxa calculated with 7 different ecological dissimilarity measures (green badges in panels **a-d**) at different taxonomic ranks. Values in panels **a-e** that are not significant ($P > 0.05$) are coloured grey.



Supplementary Fig. 7 | SNB is robust against the choice for mean pairwise distance between the samples containing a taxon. a, SNB of taxa calculated with mean versus median pairwise distance. **b**, SNB of taxa calculated with mean versus the 75th percentile pairwise distance. Numbers within brackets behind ranks indicate number of taxa. Text in the top of the panels are Spearman’s rank correlation coefficient and associated *P* value calculated for all taxa.

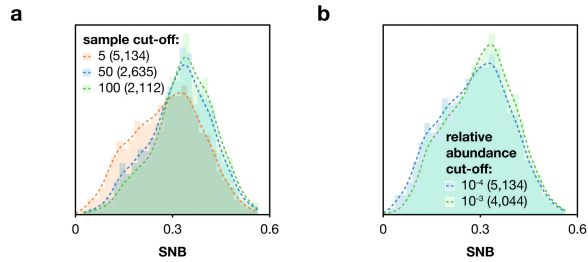


Supplementary Fig. 8 | SNB correlates with the fraction of overlapping taxa between the samples containing a taxon. a, SNB of taxa versus mean absolute pairwise overlap. **b**, SNB of taxa versus mean relative pairwise overlap. Overlap was calculated on the taxonomic rank order. Numbers within brackets behind ranks indicate number of taxa. Text in the top of the panels show Spearman’s rank correlation coefficient and associated *P* value calculated for all taxa.

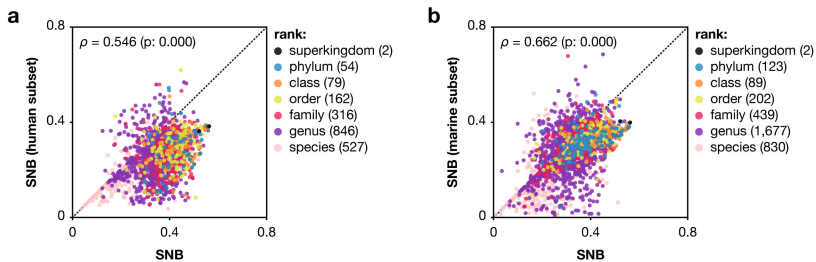


Supplementary Fig. 10 | Niche breadth of hypothetical taxa (iSNB) that are present in all samples of all combinations of two annotated biomes. Hierarchical clustering of the heatmap is based on Euclidean distance and the UPGMA algorithm. α and β diversity and number of samples are indicated with colour-coding along the axes. Number of samples are also indicated in the labels within brackets. The environment type colour-coding corresponds to **Supplementary Fig. 3.**

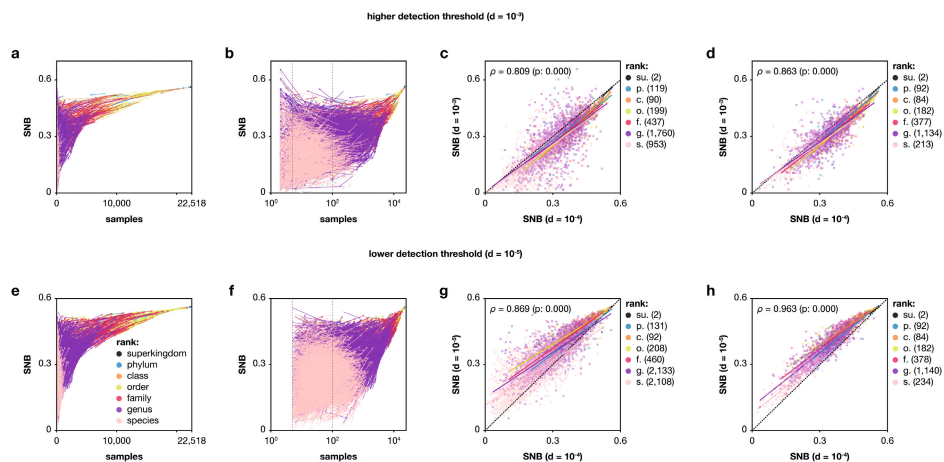
Niche range strategies of generalists and specialists



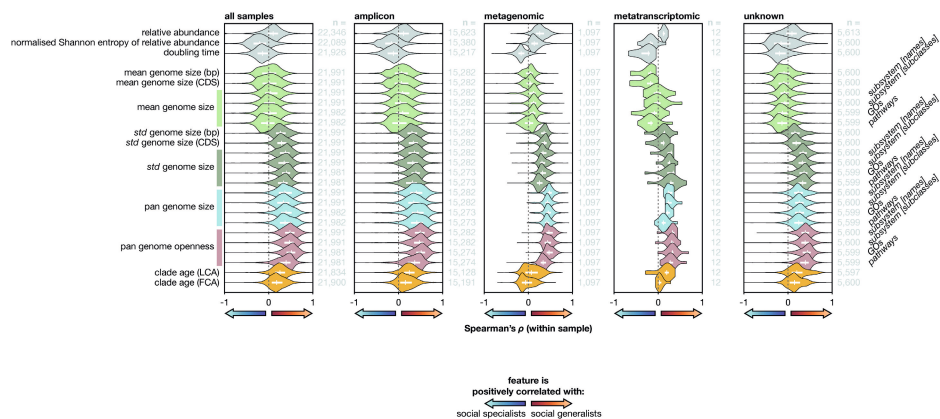
Supplementary Fig. 11 | Response of the overall distribution of SNB scores to different parameter cut-offs in our pipeline. a, The minimum number of samples in which a taxon must be found. **b**, The minimum relative abundance that must be reached by a taxon in at least 1 sample. Numbers within brackets indicate number of taxa.



Supplementary Fig. 12 | SNB is relatively invariant to environmental scale. SNB based on all samples versus SNB based on the hierarchical subsets of **a**, human annotated biomes and **b**, marine annotated biomes. See **Supplementary Fig. 1b** for which annotated biomes are included in the ‘Human’ and ‘Marine’ hierarchical subsets. Numbers within brackets behind ranks indicate number of taxa. Text in the top of the panels are Spearman’s rank correlation coefficient and associated P value calculated for all taxa.

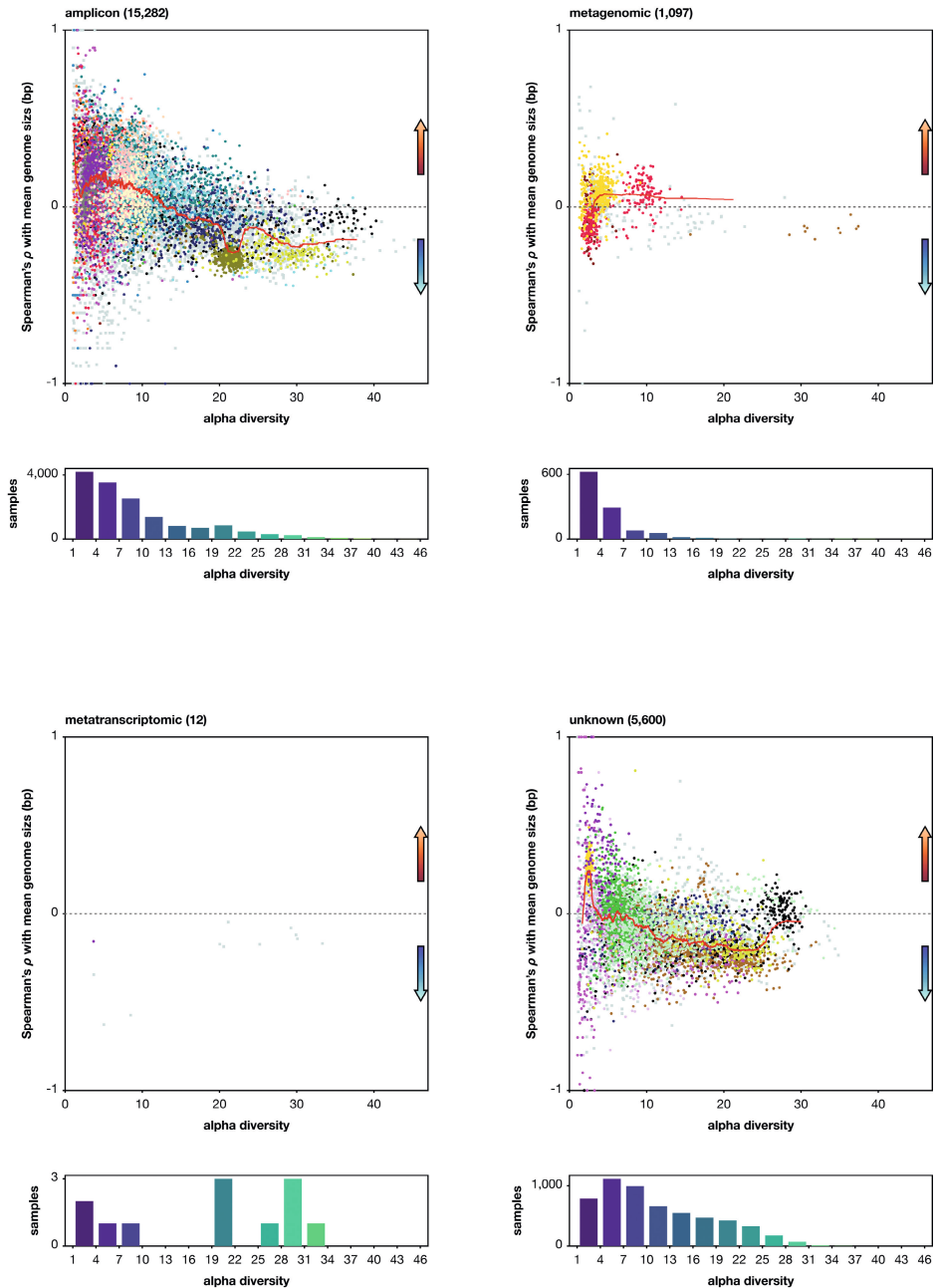


Supplementary Fig. 13 | Effect of the detection threshold on SNB. a,e, Arrows show the change in number of samples and SNB due to a higher and lower detection threshold. **b,f,** The same figure but with a logarithmic x-axis. Dashed lines indicate a presence in 5 samples which is the default cut-off for taxa to be included in most analyses in this study, and a presence in 100 samples which is the cut-off for panels **d** and **h**. **c,g,** SNB with default detection threshold versus SNB with a higher or lower detection threshold. Coloured lines are linear regression lines for different taxonomic ranks. Taxa that had a presence in less than 5 samples with a higher detection threshold (see panel **b**) are included. **d,h,** The same figure but with taxa that are present in less than 100 samples removed (dashed lines in panels **b** and **f**). Numbers within brackets behind ranks indicate number of taxa. Text in the top of panels **c,d,g,h** are Spearman's rank correlation coefficient and associated P value calculated for all taxa.



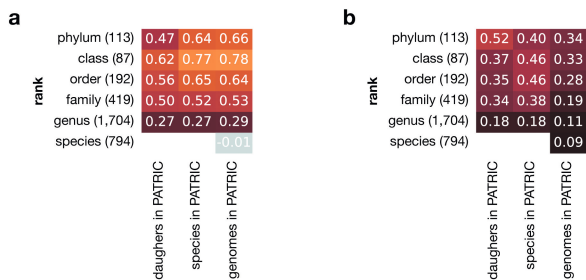
Supplementary Fig. 14 | Ecological and genomic features correlated with SNB if only samples from a single experiment type are used. Spearman's rank correlation coefficient (ρ) within samples between SNB and features related to local dominance and genomic features on the rank genus. The first panel is a repetition of **Fig. 3**, the other panels are what these results would have looked like if SNB and the shown correlations are only based on samples from that specific experiment type. Numbers to the right of violins show sample size, lines within violins show interquartile range and median. Source data of the figure are available in **Supplementary Data 6**.

Niche range strategies of generalists and specialists

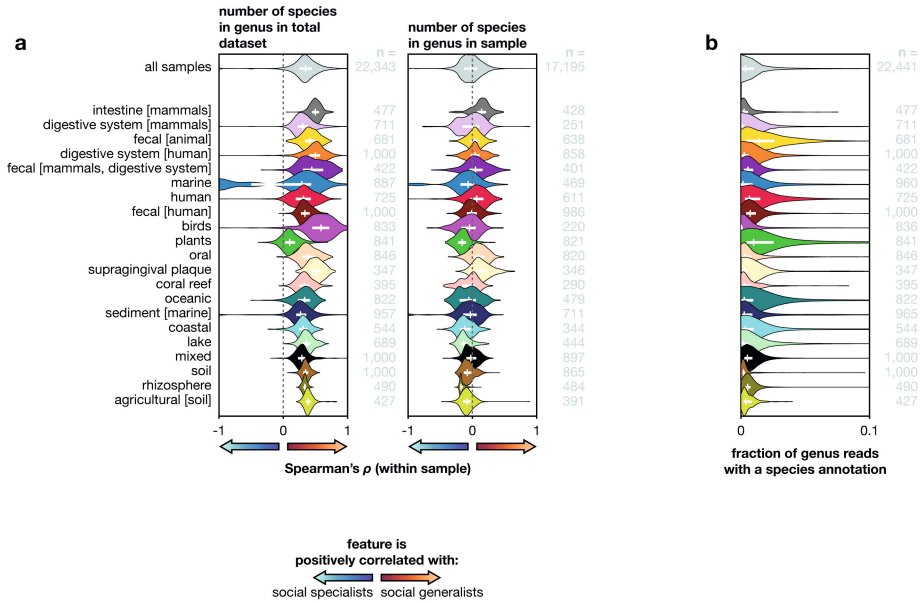


Supplementary Fig. 15 | Contrasting genomic niche range strategies if only samples from a single experiment type are used. The panels depict what Fig. 4b,c would have looked like if SNB is only based on samples from the specific experiment type. Numbers within brackets show the number of samples that are included in that experiment type, which can be lower than their total number of samples because a correlation coefficient can not be calculated when the number of genera in a sample is ≤ 2 , and samples with no classifications at order rank (two in total) do not have an α diversity definition. Source data of the figure are available in **Supplementary Data 6**.

Chapter 5

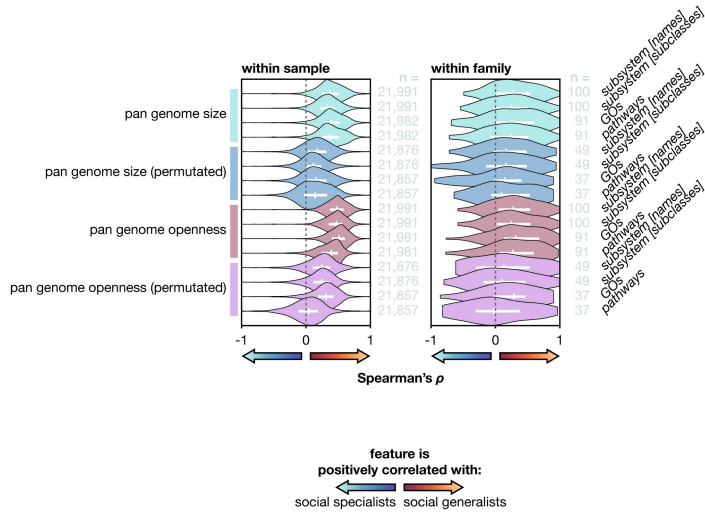


Supplementary Fig. 16 | Correlation between SNB of a taxon and its number of subtaxa in the PATRIC database. Daughters are the number of taxa one rank below the current rank. **a**, Spearman's rank correlation coefficient, and **b**, Pearson correlation coefficient. The correlation with number of genomes in PATRIC on the species rank **a** was not significant ($P > 0.05$) and is coloured grey. Numbers within brackets indicate number of taxa.



Supplementary Fig. 18 | Number of species in genera in the MGnify dataset. **a**, Spearman’s rank correlation coefficient (ρ) within samples between SNB and features related to the number of species in a genus. The number of species in the total dataset are all species that have an absolute abundance of at least 5 reads in one of the samples. The number of species in a sample are all species that have an absolute abundance of at least 5 reads in the sample. When at least 5 reads that map to a genus in a sample are not accounted for by its species, an extra ‘unknown’ species was added. In both plots, when a genus did not have any species annotation its number of species was set to 1. Violins depict the distribution of ρ across all samples or those from the annotated biomes with the most samples. Source data are available in **Supplementary Data 6**. **b**, Fraction of genus reads that have a species annotation within the sample. Numbers to the right of violins show sample size, lines within violins in panel **a** and **b** show interquartile range and median.

Niche range strategies of generalists and specialists



Supplementary Fig. 19 | Correlations between SNB and pan-genome size and pan-genome openness do not depend on a higher number of species in generalists. Spearman's rank correlation coefficient (ρ) between SNB and pan-genomic features on the rank genus within samples and within families. Violins depict the distribution of ρ across all communities or across all families with at least 5 genera. Measures are in number of unique functions for the functional universe on the right. Measures with '(permutated)' in the name are based on the mean of 1,000 randomly picked subsets of 3 species from the genus and thus correct for the high number of species in some genera. Genera with less than 3 species were excluded from these analyses. The non-permutated violins are identical to the violins in **Fig. 3**. Genome size estimates for a genus are based on the genome size of its species, which is defined as the majority set of functions of all strains for the functional universe measures. Pan-genome openness is total pan-genome size divided by mean genome size. Numbers to the right of violins show sample size, lines within violins show inter-quartile range and median. Source data of the figure are available in **Supplementary Data 6** and **7**.

Supplementary Data

Supplementary Data 1–9 (captions below) are available from Zenodo at <https://doi.org/10.5281/zenodo.8090260>.

Supplementary Data 1 | Selection of taxonomic analyses from the MGnify resource. Taxonomic analyses are associated with runs, samples and studies. Metadata of each data type are shown. At most, 1,000 samples were selected per annotated biome. For other selection criteria, see ‘**Methods**’. Zeroth-, first- and second-order α diversity measures are calculated on the taxonomic rank order.

Supplementary Data 2 | Taxonomic profiles of MGnify analyses. The numbers in the header indicate the total numbers of prokaryotic reads in the analysis. High-ranking taxa include the reads annotated to lower-ranking taxa and reads that could not be annotated to a lower rank due to a relatively unexplored biosphere.

Supplementary Data 3 | SNB for taxa throughout the prokaryotic tree of life and other features. The database from which the feature was derived is indicated as the first word in square brackets. If no database is indicated, the feature was derived from the MGnify data. Features with corrected in the name refer to our completeness and contamination correction (see ‘**Methods**’). Genome size estimates and GC content features were calculated as the mean of all daughter species (species in the name) and taxonomy corrected (no species in the name) (see ‘**Methods**’). Pan-genome features were calculated for all daughter species ([all] in the name) and for a random subset of three daughter species (1,000 random permutations per taxon) ([3] in the name).

Supplementary Data 4 | Per-rank modified zscore. The modified z score represents the number of median absolute deviations from the median for that rank divided by the constant scale factor 1.4826 to make the interpretation of the scores comparable to conventional z scores based on the standard deviation from the mean. A positive z score indicates that the SNB of the taxon is higher than the median for its rank and the taxon is thus relatively social generalist. A negative z score indicates that it is relatively social specialist.

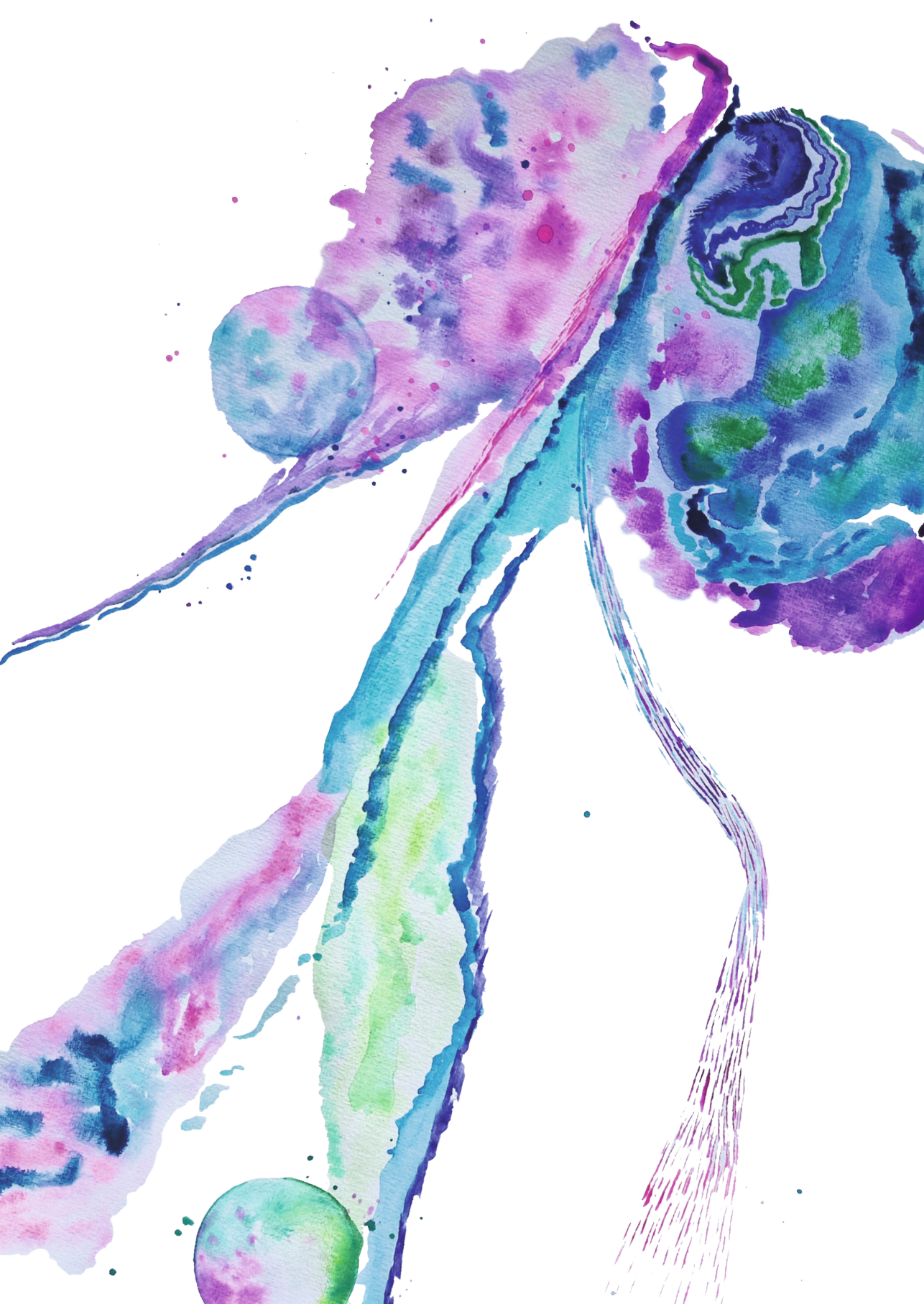
Supplementary Data 5 | PATRIC genomes and associated data. Features starting with genome. were sourced from the PATRIC database.

Supplementary Data 6 | Supporting data of figures plotting ρ within samples between SNB and features. The first value is the number of genera that the correlation is based on; the second value is ρ ; and the third value is the significance.

Supplementary Data 7 | Supporting data of figures that plot ρ within families between SNB and features. The first value is the number of genera that the correlation is based on; the second value is ρ ; and the third value is the significance.

Supplementary Data 8 | GSEA of genera sorted by SNB for different functional universes. The classical Kolmogorov–Smirnov statistic was used for the enrichment score (ES) and normalized enrichment score (NES), and *P* values were based on 100,000 random permutations of the gene set. Multiple hypothesis correction was carried out via the FDR. Tab 1: subsystem names. Tab 2: subsystem subclasses. Tab 3: gene ontologies. Tab 4: pathways.

Supplementary Data 9 | GSEA of specialist genera (SNB < 0.35) sorted by the mean α diversity of the samples in which they are found for different functional universes. The classical Kolmogorov–Smirnov statistic was used for the enrichment score and NES, and *P* values were based on 100,000 random permutations of the gene set. Multiple hypothesis correction was carried out via the FDR. Tab 1: subsystem names. Tab 2: subsystem subclasses. Tab 3: gene ontologies. Tab 4: pathways.





Chapter 6

Structural colour in the bacterial domain: the ecogenomics of an optical phenotype

Aldert Zomer*, Colin J. Ingham*, F. A. Bastiaan von Meijenfeldt*, Álvaro Escobar Doncel, Gea T. van de Kerkhof, Raditijo Hamidjaja, Sanne Schouten, Lukas Schertel, Karin H. Müller, Laura Catón, Richard L. Hahnke, Henk Bolhuis, Silvia Vignolini, and Bas E. Dutilh

* Aldert Zomer, Colin J. Ingham, and F. A. Bastiaan von Meijenfeldt contributed equally to this work.

Under review

Abstract

Structural colour (SC) is an optical mechanism that results from light interacting with matter ordered on the nano- to microscale. Although SC is widespread in the tree of life, the underlying genetics and genomics are not well understood. We collected and sequenced a set of 87 structurally coloured bacterial isolates, and 30 related strains lacking SC. Optical analysis of colonies indicated that diverse bacteria from at least two different phyla (Bacteroidetes and Proteobacteria) can create 2D photonic crystals. Pan-genome-wide association approaches were used to identify genes associated with SC. The biosynthesis of uroporphyrin and pterins, as well as carbohydrate utilisation and metabolism, were found to be involved. Using this information, we constructed a classifier to predict SC directly from bacterial genome sequences, validated it by scoring 100 strains that were not involved in creating the SC classifier, and predicted that photonic structures are widely distributed within Gram-negative bacteria. Analysis of over 13 thousand assembled metagenomes predicted that SC is nearly absent from most habitats associated with multicellular organisms except macroalgae and is abundant in marine waters and surface/air interfaces. This work provides the first large-scale ecogenomics view of SC in bacteria and identifies microbial pathways and evolutionary relationships that underlie this optical phenomenon.

Introduction

Besides pigmentation, nature's colour palette includes nanostructures reflecting light at specific wavelengths and angles. This is structural colour (SC), which allows organisms to modify their optical appearance from striking displays of colour to near invisibility (455,456). SC is common within the animal kingdom (in birds, cephalopods, shellfish and other marine invertebrates, insects, fish, arachnids and a few mammals) in which it is involved in inter- and intraspecies interactions and camouflage (457–463), and in plants, where it is used for light management or signalling (464–467). Finally, there is microbial SC, thus far observed in Myxomycetes and some Flavobacteriia, and sporadically in other bacteria as well. Flavobacteriia displaying SC have been isolated from marine or littoral environments as well as brackish water and the soil (468–470).

Microbial SC may be an optical phenotype, for example playing a role in photoprotection, but it could also be a side effect of optimal cellular organisation for nutrient uptake or intermicrobial competition, such as predation. It is fundamentally a population phenotype, since it can only be achieved by colonies and not individual cells. In two Flavobacteriaceae, *Cellulophaga lytica* and *Flavobacterium* IR1, the optical structures of colonies have been determined (470,471). In both cases, these bacteria collectively form 2D photonic crystals in which aligned rod-shaped cells form a hexagonal lattice when viewed in cross-section. Outside the Flavobacteriia, several Gammaproteobacteria have an unusual reflective, metallic or iridescent appearance that may indicate SC, an example are the 'metallic' colonies of *Pseudomonas aeruginosa* (472), although the underlying optical structures have not been determined. The literature is of limited use in identifying SC, as the language used to describe bacterial colonies is often ambiguous regarding the optical mechanisms (473). For example, some *Listeria* colonies have frequently been described as iridescent (474) although it is not known if *Listeria* displays SC. In addition, it remains unclear how widely phylogenetically or ecologically distributed bacterial SC is, to what extent there is an evolutionary function, and whether this function is similar in different SC bacteria. To date, the only role for SC known is an indirect one, as highly organised colonies of *Flavobacterium* IR1 appear to predate other bacteria more effectively than disorganised colonies (475). This is in sharp contrast to bacterial pigments which are known to have important ecological roles including light harvesting and photoprotection.

Our knowledge of the genetics and genomics of SC is surprisingly limited for such a striking effect. Within eukaryotes, only in butterflies have two genes been identified controlling structural effects in wing patterning (476,477). Some of the Flavobacteriia offer an accessible genetic system to study SC (470) and 25 genes involved in a number of pathways have been identified by transposon mutagenesis in *Flavobacterium* IR1,

demonstrating that gliding motility is important, but not essential, for the formation of SC (470,478). In addition, other genes coding for tRNA modification enzymes, the stringent response, and many with no previously assigned function have been identified as relevant to SC in *Flavobacterium* IR1 (470).

Structural colour implies a formidable capacity for cells to organise. A deeper understanding of SC in bacteria should facilitate our understanding of the evolution and mechanisms behind SC. In addition, SC may form the basis of industrial processes to create sustainable colourants to replace conventional pigments. Here, we have curated a collection of bacteria, largely Gram-negatives, scored for the presence or absence of SC. SC was initially identified in colonies as a pointillistic, angle-dependent, saturated colour reflection when illuminated with white light (470,478). We excluded colonies displaying rainbow effects seen in transmission as this is a common but distinct optical effect, one likely to indicate a diffraction grating which can probably be formed by many disordered aggregates of bacteria (479). The genome sequences of SC strains and non-SC strains were used to create a computational tool that predicts SC from gene content. This predictive tool was used to identify pathways common to SC in Gram-negative bacteria, and to search metagenomics datasets to define likely SC-rich biomes and discover strains showing SC, particularly within the phylum Proteobacteria.

Results

Selection of SC and non-SC bacterial strains

To sequence and compare the genomes of non-SC and SC bacterial strains, we created a collection of bacteria showing SC by screening environmental samples on agar plates and scoring for structural colour or sourcing strains from microbial culture collections. All strains were cultivated on plates containing nigrosin (see ‘Methods’ and **Supplementary Table 1**). SC was considered to be present if punctuate, angle-dependent colour was visible upon illumination with a broad spectrum, white LED (**Fig. 1**). A full spectrum of structural colours was obtained within the collection. Strains showing SC contained a complex mix of pointillistic colour when viewed under low power microscopy (**Fig. 1**) and showed variation in colour and/or intensity when viewed from multiple angles (**Supplementary Fig. 1**). SC was confirmed by mechanical disruption of colonies, demonstrating such mixing reduced or eliminated SC. The collection was supplemented with strains from the same taxonomic order as those showing SC, but which did not show SC under any growth condition. It was notable that the most intense SC was found in Flavobacteriia, and when mechanically disrupted on an agar plate, these gliding bacteria could reform SC rapidly, over a period of 10–30 minutes. SC was generally duller of isolates outside the class Flavobacteriia, although there were exceptions within the Gammaproteobacteria, notably ‘*Marinobacter algicola* HM-28’ (**Fig. 1d**, **Supplementary Fig. 1**).

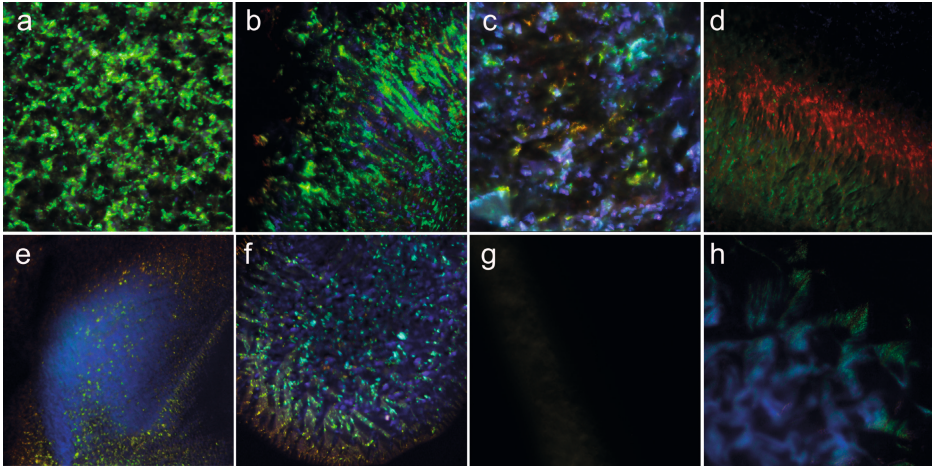


Fig. 1 | Examples of colonies from selected strains contributing genomes to the SC classifier. Each panel shows a 2×2 mm area of a colony on nigrosin containing agar, illuminated at the optimal angle. **a**, *Flavobacterium* IR1. **b**, *Cellulophaga lytica* HM-52'. **c**, *Cellulophaga fucicola* HM-74'. **d**, *Marinobacter algicola* HM-28'. **e**, *Muricauda ruestringensis* HM-37' **f**, *Tenacibaculum gallaicum* HM-45'; **g**, *Flavobacterium succinicans* DD5b', this strain does not show SC and is less reflective than the other strains in this figure that all display SC; **h**, *Virgibacillus dokdonensis* HM-38'.

Identification of genes associated with bacterial SC using a pan-GWAS approach

We hypothesised that SC may be genetically determined. To compare the genetic content of non-SC and SC bacterial strains, we selected 69 strains from our own collection (both with and without SC) and sequenced their genomes. Additionally, 48 genomes were selected from GenBank (279), from isolated strains that had previously been shown to display SC. Their characteristics are listed in **Supplementary Table 1**. These are 93 from the phylum Bacteroidetes, 23 from the phylum Proteobacteria and one from the phylum Firmicutes. A phylogenetic tree of the isolates is shown in **Fig. 2a**. Based on the pan-genome of the 117 isolates, an ortholog table of 29,850 protein coding genes was constructed using Roary (480), revealing a relevant gene set of 366 orthologs. To associate specific orthologs with structural colour, we used a pan-GWAS approach on the ortholog table using Scoary (481). We found a total of 199 orthologs to be associated with the SC phenotype (**Fig. 2, Supplementary Table 3**); 31 were detected using the Fisher exact method after Bonferroni correction, 100 using the permutation method, 79 using the phylogenetically informed pairwise comparison test in Scoary and 31 from mutagenesis (470,482). A complete list of the proteins and their presence or absence can be found in **Supplementary Table 3**. Interestingly, some of the SC-associated orthologs that are common in Bacteroidetes are shared by a member of Proteobacteria (*Marinobacter algicola* HM-30'; **Fig. 2**). We hypothesise that at least some of these genes were transferred from Bacteroidetes, although contamination of the isolate cannot be ruled out until the genome sequence is closed.



Fig. 2 | Genes identified by pan-GWAS and mutagenesis. a, Phylogenetic tree of the 16S ribosomal RNA gene, showing 117 strains included in this study (**Supplementary Table 1**). **b**, Gene presence/absence matrix of 199 proteins associated with structural colour based on both pan-GWAS and earlier knockout studies. Proteins are clustered using Ward’s method, details can be found in **Supplementary Table 3**. Proteins in green were found using mutagenesis. Proteins from the pan-GWAS analysis were clustered by STRING (see **Supplementary Fig. 2**) and their clusters are indicated using the colour legend displayed below. The importance of each gene for predicting SC in the RF model is given as vertical red bars at the top of each cluster (GINI importance score). **c**, The SC phenotype is displayed in the two columns in light blue with the first bar displaying the initial RF model input and the second column the final RF model input. The horizontal blue bars labelled ‘RF votes’ display the fraction of decision trees in the second corrected random forest model supporting classification as a strain with SC.

Functional annotation of phylum-crossing marker genes predicts key processes involved in bacterial structural colour

To predict functional associations between the genes, most of which were annotated as hypothetical proteins, we uploaded the protein sequences of the orthologs selected by the pan-GWAS approach to the STRING database, which integrates diverse sources of evidence for functional interactions between proteins (483). We detected six large clusters of proteins, which we named after the functions that were encoded by some of their members (**Supplementary Fig. 2, Fig. 2b**). These were clusters associated with pterin, porphyrin, carbohydrate, methionine, acetolactate biosynthesis, and gliding motility. The latter has previously been shown to facilitate the formation of SC in the Flavobacteriia (470,478), the other categories have not been previously associated with bacterial SC.

Genes encoding proteins that had the highest GINI importance for predicting the SC phenotype in the RF model were those linked to pterin metabolism. Pterins are widespread cofactors that have previously been shown to play a role in modulating structural colour in butterflies. They affect light scattering and selective absorption in the wing scales of pierid butterflies in which they are present as granules, which increase light reflection and amplify iridescent ultraviolet signalling (484,485). In bacteria, pteridine molecules act as enzymatic cofactors and they produce various pigments. Pteridines can act as sensors of environmental stress, and are involved in environmental transitions such as biofilm formation. Pterins have been implicated in phenotypes related to UV protection and phototaxis in Cyanobacteria (486) and accumulate in some photosynthetic bacteria when they are exposed to light (487).

The porphyrin cluster includes uroporphyrin biosynthesis proteins. Porphyrins strongly absorb light, which is then converted to energy and heat in the illuminated areas and are responsible for the colours in feathers of certain birds (488). In bacteria, accumulation of porphyrins causes photosensitivity. In a *visA* mutant of *Escherichia coli*, accumulation of protoporphyrin IX and subsequent exposure to visible light produces reactive oxygen species that are harmful (489). In addition, porphyrins have been used in artificial systems to create SC. Bacterial species observed to have structural colour frequently live in an environment characterised by regular light exposure, such as in air–water interfaces in tidal flats. Porphyrins may function as photoprotectants or light sensitive switches although any direct role in bacterial SC has not been demonstrated.

The presence of a shared Carbohydrate cluster is in agreement with observations from *Flavobacterium* IR1 that iridescence induction depends on the selected carbon source, particularly algal polysaccharides such as fucoidan and *k*-carrageenan (470,482). The organisation of the *Flavobacterium* IR1 colony is strongly regulated by cultivation on fucoidan, and transposon knockouts have implicated a specific PUL

operon (polymer utilisation locus) in mediating both the uptake and metabolism of fucoidan and linking this process to the SC displayed by the colony (482). In addition, methionine and acetolactate clusters are part of the gene set and have links to amino acid metabolism, the latter being a precursor in the synthesis of branched chain amino acids (490). A link with the stringent response may be possible, in which both acetolactate and branched chain amino acid metabolism are relevant. In support of this hypothesis, we observed that a *spoT* mutant, this gene being responsible for regulation of the stringent response, lacks structural colour in *Flavobacterium* IR1 (470).

In addition, we identified the *quiP* gene as being relevant for SC (**Fig. 2, Supplementary Table 3**), which encodes an acyl-homoserine lactone acetylase (491). This is the first indication of any involvement of quorum sensing in the formation of SC colonies. This is interesting given that the ‘metallic/iridescent’ phenotype of *Pseudomonas aeruginosa* colonies appears complicated but has been suggested both to be related to SC (473) and linked to quorum sensing (492). As we show in **Supplementary Fig. 3**, *Pseudomonas aeruginosa* does indeed appear to show SC, growing as colonies with punctuate, bright focal points of colour that are lost by mechanically disrupting the ordering of cells within a colony.

Finally, some of the strongest predictive genes had no assigned function, suggesting SC involves novel pathways.

Prediction and confirmation of structurally coloured strains using machine learning

Profile Hidden Markov Models (HMMs) were constructed of the sequence alignments of the 199 orthologs that were associated with the SC phenotype. A machine learning model was constructed using Random Forest (493) with the 117 bacterial genomes as training set, in which 2/3 of the data was used for training and 1/3 for testing the RF model. An OOB prediction error of 6.8% was observed, suggesting that the model is indeed capable of predicting SC from genome sequences. Unexpectedly, five genomes that were considered SC-negative in the past, displayed a score above 0.6 in the classifier. These included the Proteobacteria *Pseudomonas aeruginosa* and four Bacteroidetes: *Cyclobacterium marinum*, *Zobellia galactanivorans*, *Zobellia uliginosa*, and *Kriegella aquimaris*. These strains were tested for SC. In *Pseudomonas aeruginosa* punctuate iridescence was observed that was lost when the colony was mixed with a loop, suggesting SC (**Supplementary Fig. 3a–c**). The Bacteroidetes strains revealed that they did display SC. A new RF model adding these five strains to the positive set was constructed with an OOB error of 3%. This improved model is available at <https://github.com/aldertzomer/structuralcolour> and can be used to classify assembled (meta)genomes for SC.

The SC classifier was validated by scoring a collection of strains for SC that were not part of the initial classifier group (**Supplementary Table 4**) by calculating the area under the curve (AUC) of the output of the model versus their SC phenotype. The overall AUC was 0.91 ($n = 93$) for Gram-negative bacteria, 0.92 ($n = 55$) for the Bacteroidetes and 0.90 ($n = 38$) for the Proteobacteria. Taken together, this suggests good predictive accuracy for the two phyla that comprise most of the strains found that show SC.

SC throughout the bacterial domain of life

With our RF-based machine learning model, we set out to predict SC-positive strains throughout the bacterial domain, by calculating the SC score for 240,981 bacterial genome sequences downloaded from the PATRIC database (185) (**Supplementary Table 5**). We considered genomes scoring above 0.68 as likely SC-positives, and genomes scoring below 0.39 as likely SC-negative, based on the score boundaries of the non-SC and SC species in the training set of the classifier (**Fig. 3a**). The other genomes were considered putatively SC-positive.

Although the model was built using a taxonomically biased selection of organisms, we found strong support for SC in many different phyla (**Fig. 3b**). Moreover, predicted SC does not seem to be highly taxonomically biased, i.e. there are high- and low-scoring genomes in many of the taxa. This is an interesting prediction given to date, SC has only been demonstrated within the phylum Bacteroidetes. SC in the Proteobacteria is novel, and was particularly well represented in cultured isolates. It is therefore discussed in more detail below.

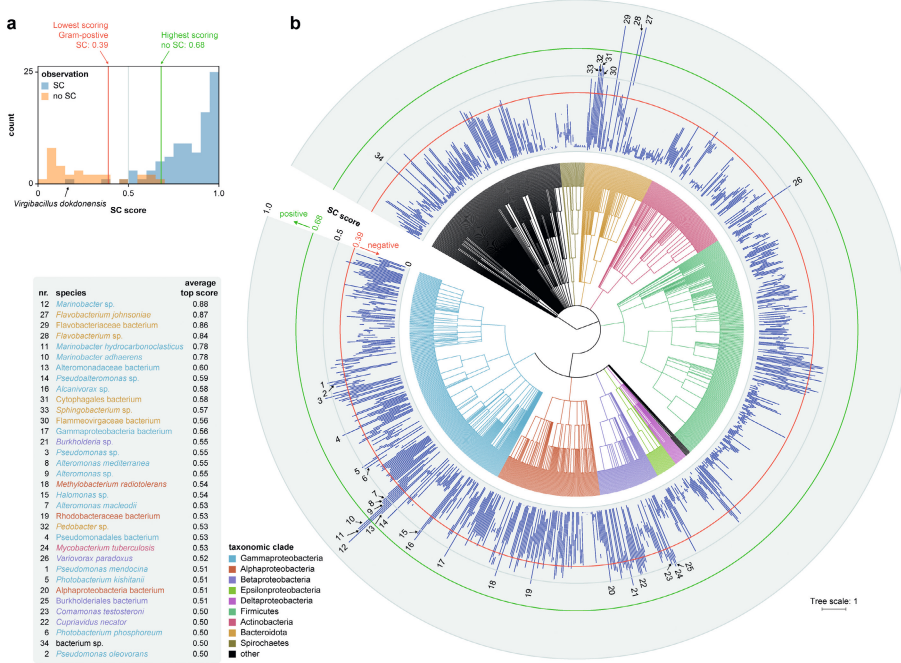


Fig. 3 | Range of structural colour scores per species as predicted using the RF model. a. Distribution of 117 SC scores in the training set. The single structurally coloured Gram-positive bacterium *Virgibacillus dokdonensis* has a low classifier score. **b.** SC was predicted for 240,981 bacterial genomes from the PATRIC database (see **Supplementary Table 5**). Species with at least 10 genomes are displayed, as incidental high scores may not be genuine and these need to be investigated experimentally. Both ends of each blue line indicate the mean of the five minimum and maximum SC scores (min = 0.03, max = 0.88) (see **Supplementary Table 6**). The tree indicates the NCBI taxonomy as annotated by PATRIC. High scoring species are indicated by numbers. Coloured rings show the relevant boundaries of the training set in panel **a** for reference. For details see **Supplementary Tables 5 and 6**.

Host-associated bacteria like members of the order Enterobacteriales, which contains the human pathogens *Klebsiella pneumoniae* and *Escherichia coli*, scored poorly within the Gammaproteobacteria (SC score < 0.3) and did not show SC in our experiments. In contrast, SC colonies were developed by members of the taxonomic orders Pseudomonadales (including the previously mentioned *Pseudomonas aeruginosa* and other pseudomonads), Oceanospirillales (*Kangiella sediminilitoris*), Xanthomonadales (*Pseudoxanthomonas* spp.) and Alteromonadales (*Marinobacter algicola*, *Microbulbifer arenaceus*) (**Fig. 3b**, **Supplementary Tables 1 and 4**, **Supplementary Figs. 1 and 3**). In addition, we note that *Alcanivorax balearicus*, a member of the Gammaproteobacteria, has been described as forming iridescent colonies (494). Generally, SC within the Gammaproteobacteria appeared weaker than of many isolates of the Flavobacteriia. This may explain the results of previous studies that reported Flavobacteriia showing SC on plates with a translucent background,

which makes weak SC harder to detect, but failed to isolate iridescent colonies from other types of bacteria (468,469). Subsequently, we optimised on diverse cultivation media what appeared to be weak SC (**Supplementary Tables 1 and 4**), leading to the observation of intense colouration, for example both '*Marinobacter algicola* HM-32' and several *Kangiella* and *Hoeflea* spp. colonies (**Fig. 1d**, **Supplementary Figs. 1 and 3d**, **Supplementary Table 4**).

Within the Alphaproteobacteria, we are aware of one prior observation of SC in an *Agrobacterium* species (order Rhizobiales), isolated from an aquatic fern that formed iridescent colonies, although the authors did not indicate whether this was due to SC (495). Our prediction based on the RF classifier, that SC may be more widespread within the Alphaproteobacteria (**Fig. 3**), was made despite the fact that no genomes from this class were used to construct the RF classifier. For example, an SC score of 0.85 was predicted for 'Rhodobacteraceae bacterium 4F10' (order Rhodobacterales) from coastal ocean surface water (496) (**Supplementary Table 5**). To validate these predictions, we screened environmental samples from littoral locations in the Zeeland province (the Netherlands), resulting in the isolation of three Alphaproteobacteria displaying weak SC, including two *Hoeflea* species (order Rhizobiales) and one *Sulfitobacter* (order Rhodobacterales) (**Supplementary Fig. 3**, **Supplementary Table 4**).

Whilst the number of predicted SC strains in the Betaproteobacteria were comparatively small in number, *Cupriavidus basilensis* appears to show SC. Glitter-like iridescence likely to be SC was observed in '*Cupriavidus basilensis* RK1' (**Supplementary Table 4**), a member of the Burkholderiales. Colonies of other Burkholderiales (*Acidovorax delafieldii*, *Janthinobacterium svalbardensis*, and *Chitinimonas korensis*) appeared to show SC (**Supplementary Table 4**). In addition, other Betaproteobacteria showing a glitter-like iridescence, *Sphingomonas pruni* and *Sphingobium herbicidovorans* (both order Sphingomonadales) also show similar phenotypes (**Supplementary Table 4**). Within the order Neisseriales, a species of *Pseudogulbenkiania* appeared to grow as SC colonies, as did the members of the order Rhodocyclales (*Dechloromonas* sp. and *Thauera aromatica*) (**Supplementary Table 4**).

Gram-positive bacteria with SC

There was only one Gram-positive genome in the SC classifier validation, from *Virgibacillus dokdonensis* which showed green/blue SC when grown on RMAR agar with high salinity (6% w/v sea salt) and at 50 °C (**Fig. 1h**). In this strain of *Virgibacillus dokdonensis* the purple/green colouration, apparently structural, was visible during vigorous motility over agar, with the colony spreading up to 5 mm/h. The original isolate of *Virgibacillus dokdonensis* (SC score = 0.18; **Fig. 3a**) was previously described as having purple tinted colonies, but this was not recognized

as SC at the time. This suggests that there may be other groups of SC bacteria that are not captured by our classifier, within the Gram-positive strains and possibly within other groups of microorganisms. We hypothesise that Gram-positive bacteria use a different genomic mechanism that is not predicted well by our classifier. *Listeria goaensis* and *Listeria monocytogenes* have both been described as producing iridescent colonies (497,498), suggesting possible SC. Also, in the validation study we found apparent SC in swarming *Paenibacillus vortex* (**Supplementary Table 4**). Representative genomes of these species have SC scores of <0.25, again suggesting SC in Gram-positive bacteria is not scored highly by the classifier developed from Gram-negative bacteria.

Optical analysis of Gammaproteobacteria shows a 2D photonic crystal arrangement that is similar to the Flavobacteriia

Most of the isolates showing SC appeared to spread over agar suggesting active surface translocation or motility. It has been established previously that gliding by Flavobacteria facilitates the formation of SC (470,478) and this is supported by the identification of gliding-related genes in this work (**Fig. 2b**). However, we did not find a universal motility mechanism that was shared by all SC strains studied. Within the Gammaproteobacteria, multiple mechanisms of motility (e.g. flagella and various pili-related) are known (499). To test the role of flagella-based motility, the formation of SC was tested in *Marinobacter subterrani* by comparing the formation of SC colonies in the WT strain with a *flaBG* knockout mutant (500). Despite the loss of flagella motility, as indicated by the mutant strain displaying a non-spreading phenotype when inoculated into 0.2% (w/v) sloppy RMAR plates (**Supplementary Fig. 4, Supplementary Table 7**), SC was not impaired when cultivated on hard RMAR plates (0.8% w/v agar). In addition, the *flaBG* strain still could spread over the hard agar RMAR plates at 0.6 cm/day, i.e. at the same rate as the WT, suggesting a flagella-independent mechanism of surface translocation. Also, other Gammaproteobacteria, *Kangiella sediminitoris* and *Hoeflea* sp., were also shown to have SC and displayed surface translocation on RMAR agar (0.8% w/v) plates at rates of up to 1 cm/day (**Supplementary Table 7**).

In order to check and validate that SC was present in the Gammaproteobacteria, the underlying optical structure was investigated. When we compare the organisation of cells in colonies of Flavobacteriia with SC with colonies of SC Gammaproteobacteria, we observe that the hexagonal packing of the cells is a universal feature across both classes (**Fig. 4, Supplementary Fig. 5**). To compare the optical response produced by these structures we measure angle-resolved reflectance spectra, using a goniometer. The goniometer illuminates the sample at a fixed incident angle (-45° with respect to the surface of the colony), and subsequently records spectra at a wide range of detection angles (**Fig. 4d**; ref. 501). The obtained spectra show intensity spots at specific wavelengths and angles that are characteristic for these types of structures

(Fig. 4, Supplementary Fig. 5; refs. 470,471,482,501). The angular dependency of these diffraction spots can be fitted to the grating equation, which reveals the period d of each colony (as explained in more detail in ref. 501). We find a lattice constant of the photonic crystal of $d = 490$ nm for ‘*Marinobacter algicola* HM-28’, $d = 395$ nm for *Flavobacterium* IR1, $d = 440$ nm for ‘*Marinobacter subterranei* JG233’ (500). In conclusion, the different colours observed by the different bacteria strains do not stem from different forms of cell packing, but from variations in the periodicity of the structure. Such variations can either be caused by differences in the interbacterial distances or in cell size (482). Additionally, ‘*Marinobacter algicola* HM-28’ shows extra diffraction spots at around 45° that are not visible in *Flavobacterium* IR1, but contribute to the observed colour appearance. This does not necessarily mean that these photonic features are not present in *Flavobacterium* IR1, but are hidden by the strong specular reflection (mirror-like surface reflection). For ‘*Marinobacter algicola* HM-28’, the specular reflection is confined to a smaller angular range than in *Flavobacterium* IR1, and the diffraction spots at 45° have a higher intensity.

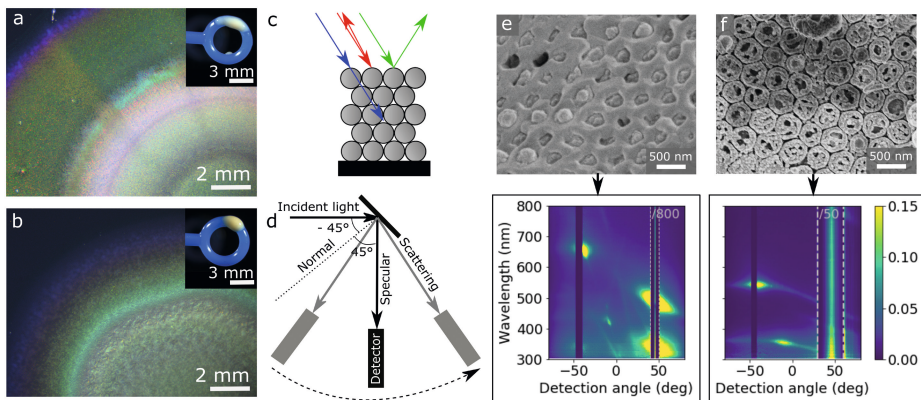


Fig. 4 | Comparison of the cell arrangement and optical properties of representative strains from Marinobacteraceae (*Marinobacter algicola* HM-28) and Flavobacteriia (*Flavobacterium* IR1). **a**, Colony of ‘*Marinobacter algicola* HM-28’ showing SC, inset shows material from the colony losing SC after mixing with an inoculation loop. **b**, As panel **a**, but showing *Flavobacterium* IR1. **c**, Schematic showing a hexagonally packed photonic crystal in cross section. When white light reaches the crystal structure, some wavelengths can pass through it (blue arrow), but others cannot (green and red arrow). **d**, Schematic of the goniometer setup used to capture the spectra shown in panels **e** and **f**. **e**, Cryogenic SEM images of cross sections of colonies of ‘*Marinobacter algicola* HM-28’ and **f**, *Flavobacterium* IR1. The corresponding angle resolved spectra for both panels **e** and **f** are given below, showing the intensity of the reflected light on a blue to yellow colour scale. The incident light angle is kept at -45° , and due to a limitation in the setup no spectra can be taken at the incident light angle. Because the mirror-like reflection (specular) around 45 degrees is far brighter than the scattered light at other angles, and can therefore not be shown on the same scale without saturating the signal, the reflected light intensity between the dashed lines has been divided by the number shown at the top. Cryo-SEM image and angle-resolved spectra for *Flavobacterium* IR1 are reproduced from ref. 501.

In addition to the different colours that are reflected by the various strains, there are also differences in the angular range over which the diffraction spots are spread (goniometer plots in **Fig. 4e,f**). This type of spreading of the reflected light over a wider angular range is typically caused by local variations in orientation of crystalline domains (471,482,501). Wherever the crystal structure appears in a tilted orientation, either because the surface of the colony is not entirely flat, or because of imperfections in the structure, light will be reflected under a slightly different angle. However, this feature is again strain-specific, and cannot be attributed as phylum-specific, as the angular range over which the light is reflected varies between both of the Gammaproteobacteria strains that were investigated here (**Fig. 4e** and **Supplementary Fig. 5c**).

Thus, we conclude that all strains, both Gammaproteobacteria and Flavobacteriia, follow the same principle mechanism of a hexagonally packed photonic crystal, but with strain-specific variations in the interplay between order and disorder. The two classes of Gram-negative bacteria form the same type of photonic structures, despite being from distinct phylogenetic groups and with different mechanisms of motility.

Identification of environments that contain bacteria that display evidence for SC

To investigate the ecological distribution of SC across microbial systems, we applied the classifier to 13,873 assembled metagenomes from 108 different biomes available in the MGnify database (152) and scored each metagenome for SC (**Fig. 5a–c**, **Supplementary Table 8**). Animal- and plant-associated microbiomes scored consistently low for SC (**Fig. 5b**). The notable exceptions were the metagenomes from macroalgae (**Fig. 5c**). This is consistent with previous studies, as bacteria showing SC have been isolated from macroalgae (469,470). SC bacteria have genes for the metabolism of algal polysaccharides and algal polysaccharides regulate SC in *Flavobacterium* IR1 (482). This is supported by the importance of carbohydrate metabolising genes identified with pan-GWAS. We find SC in aquatic and engineered biomes, which include aerobic and light-exposed habitats (**Fig. 5c**). However, SC bacteria were also found in non-illuminated biomes such as pond sediment and groundwater (**Fig. 5c**; ref. 500). Strikingly, from marine metagenomic studies we observed that many of the assembled metagenomes with the highest SC scores were from depths with limited light (**Fig. 5d,e**, **Supplementary Table 9**), i.e. below the photic zone which extends to 200 m water depth. One possible explanation for the depth profile seen is that SC is found on ‘marine snow’ which may permit assembly into highly-organised groups of cells. It has been suggested that highly organised groups of Flavobacteriia, perceived as structurally coloured colonies, have a competitive advantage against other bacteria (475). It is possible that such competition on space is occurring on marine snow and explains the high SC scores in the ocean depths. To test this hypothesis, we downloaded and assembled 62

metagenomes from sinking particulate organic matter captured at 4,000 m water depth (502) and confirmed that most had high SC scores (Fig. 5f, Supplementary Table 10).

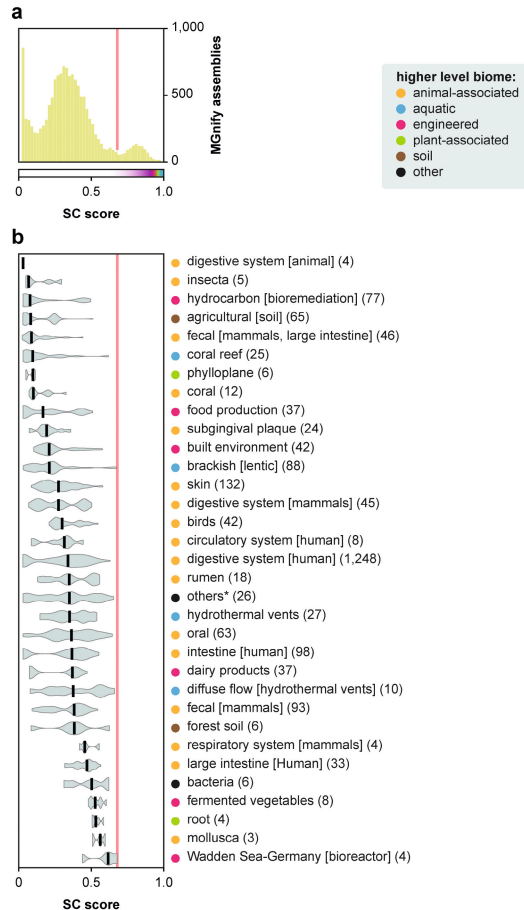
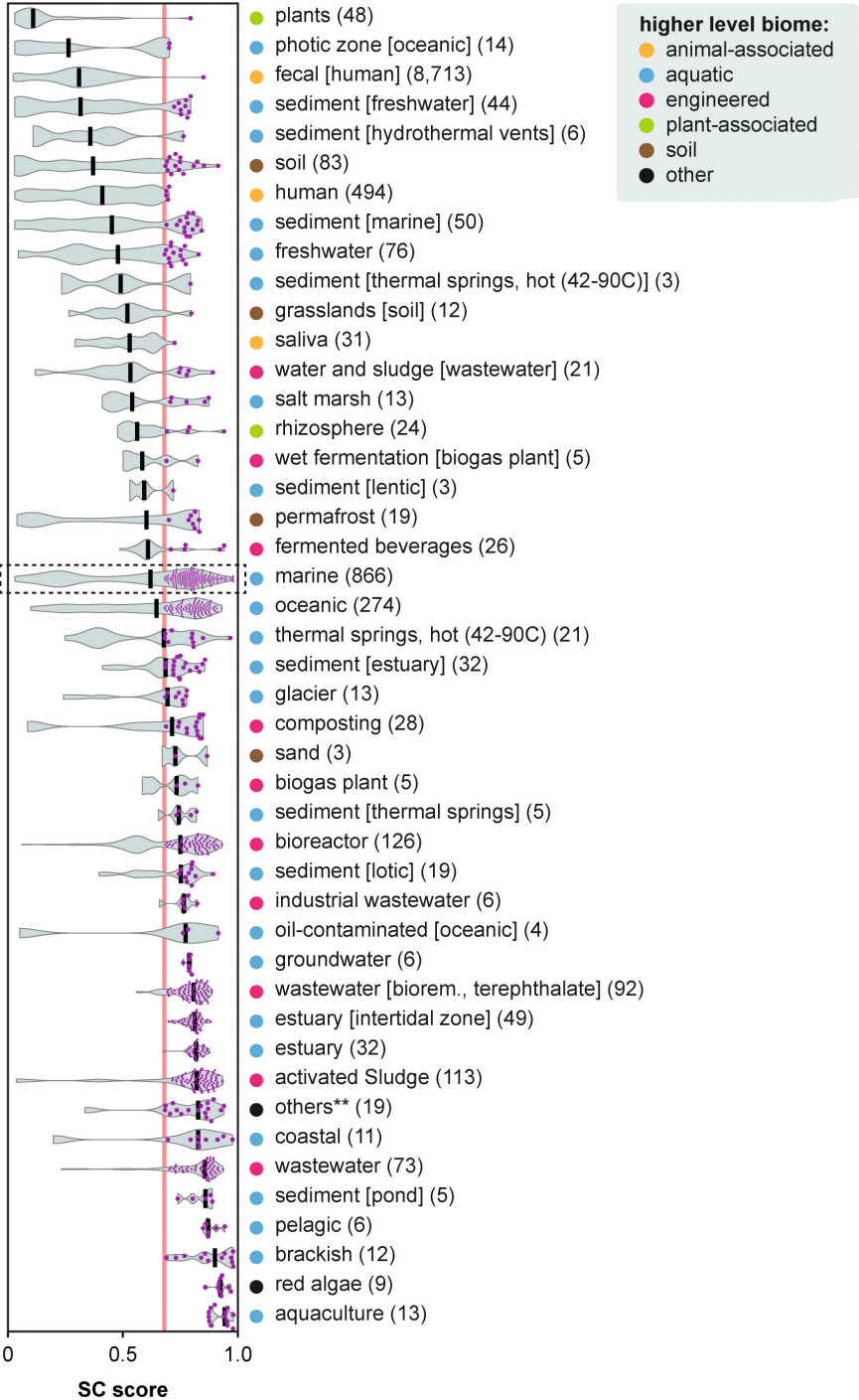
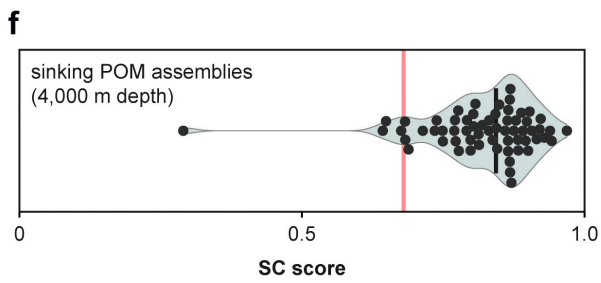
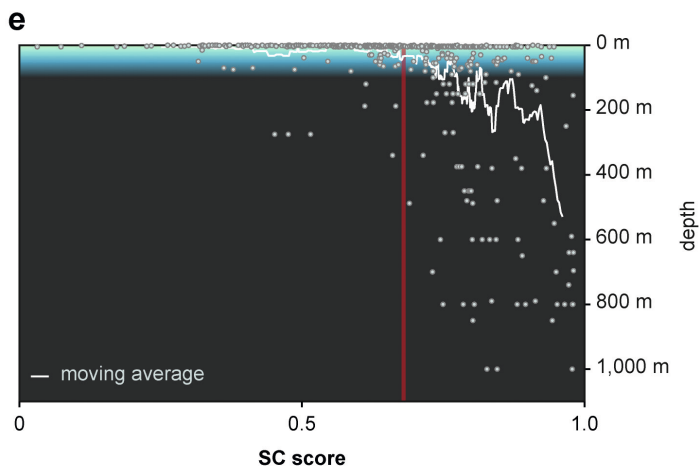
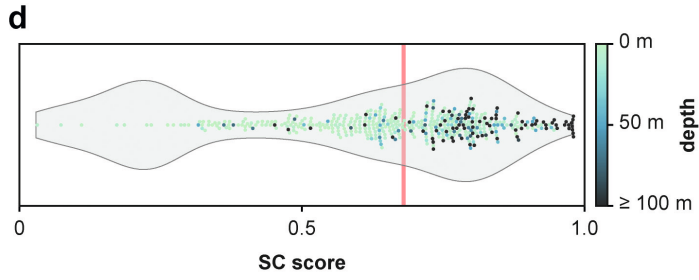


Fig. 5 | Classifier scores for assembled metagenomes from diverse environments. **a**, The distribution of SC scores of all metagenomes ($n = 13,873$) reveals a valley at 0.68 (red line), which we used as a cut-off score. **b**, The subset of biomes from which no metagenome was found scoring >0.68 . ‘Others’ (indicated with *) is the combined set of biomes with less than three associated datasets each ($n = 19$). **c**, Biomes containing at least one high scoring metagenome above the threshold of 0.68. High-scoring metagenomes are indicated with purple dots. ‘Others’ (indicated with **) is the combined set of biomes with less than three associated datasets each ($n = 13$). Black vertical lines in panels **b** and **c** show the median for each biome. **d**, Distribution of SC scores in the marine biome (see dashed rectangle in panel **c**) with associated water depth metadata. Only assemblies with specific depth information are plotted ($n = 450$), i.e. samples annotated with ‘surface’ or ‘5-160m’ depth are not shown. **e**, The same figure as panel **d** but with depth in the water column on the y-axis. The white line is the moving average with a window size of 20 data points. **f**, Distribution of SC score in 62 assembled metagenomes from sinking particulate organic matter (POM) collected at 4,000 m water depth (502).

C



Structural colour in the bacterial domain



Conclusions

Structural colour is found in many living organisms. In Eukaryotes, SC is intensively studied in the field of optics but there is little genomics. This work has studied bacteria, in which individual cells coordinate to form colonies showing SC. We created a curated collection of genomes, largely from Gram-negative bacteria, both with and without SC, and used this to generate a predictive, genome-based classifier. This work suggests that SC is present in a much more diverse group of bacteria than thought, particularly from the phylum Proteobacteria. We showed that members of the Gammaproteobacteria can form 2D photonic crystals that are essentially the same as previously determined for the Flavobacteriia. The classifier was validated with the genomes from additional SC-positive and -negative strains. We identified clusters of genes involved in pterin, porphyrin, carbohydrate, methionine and acetolactate metabolism that appeared to be common signifiers. The pterin/pteridine associated cluster was particularly interesting as the most predictive set of genes and because SC can be enhanced by pterin-related pigments in some eukaryotes (484,503). In addition, genes associated with gliding motility were associated with SC in Flavobacteriia. Gammaproteobacteria also show SC but do not require the flagella to organise (**Supplementary Fig. 4**). The classifier was applied to metagenomics datasets. SC was predicted to be common in aquatic and engineered biomes, but rarely associated with multicellular organisms with the notable exception of macroalgae. This supports previous observations on the ecological distribution and metabolic properties of bacteria showing SC (468–470,482). Interestingly, bacteria capable of SC were predicted to be common in the deep ocean which, taken with an apparent illumination-independent role of the cell organisation of SC *Flavobacterium* IR1 underlying structural colour in intermicrobial competition, suggests SC may be a side effect of colony organisation (475). In addition, our screen of the bacterial domain predicted SC in the Alpha- and Betaproteobacteria which we confirmed by isolating SC strains from these taxonomic classes. Our goniometry experiments clearly demonstrated the existence of a two-dimensional photonic crystalline colony structure in Gammaproteobacteria, similar to that already demonstrated in the Flavobacteriia. Finally, our study suggests the existence of SC within the Gram-positive bacteria.

This is the first large-scale, genomic-based analysis of SC in any organism. Bacterial colonies with SC are living nanostructures that manipulate light in intricate ways, and we have identified several molecular pathways that are linked to the process of generating them. The identified genetic signature of SC may also contribute to answering the question whether SC is selective as an optical phenotype, or a side effect of structural organisation of the colonies for a different reason. The pathways identified in our work may lead to an understanding of the function of SC in bacteria and the evolutionary relationships and processes that have created this radiant population phenotype.

Materials and methods

Strains and culturing

Unless stated otherwise, environmental samples from soil or freshwater sources, including *Flavobacterium* IR1, were cultivated on ASWBC or ASWBLow agar plates (470) at 20 °C, except that the 1% (w/v) KCl in the original formulation was replaced with the same amount of sea salts (Sel Marin, Portugal). Samples from marine or littoral origin were cultivated on RMAR plates (10 g/l peptone from animal sources, 4 g/l yeast extract from Sigma, NL, 30 g/l, sea salts as above, 200 mg/l nigrosin, 10 g/l agar (Life Technologies, NL), unless stated otherwise. The strains isolated in this study and optimal growth conditions are listed in **Supplementary Table 1**. Screening for new structurally coloured isolates was performed on the same media, plating dilutions of the source material and incubating from 3 to 10 days at 20 °C. Microcolonies showing angle-dependent colouration, suggestive of SC, were isolated by toothpick for further analysis.

Testing and curation of structural colour

Strains were tested under multiple growth conditions on agar before scoring for SC. Isolates were considered as showing SC if colonies showed angle-dependent, glitter-like colouration on plates with a dark background when illuminated from the side with a broad spectrum white LED and/or direct sunlight. In addition, if this colouration was disrupted by mixing the colony with an inoculation loop then this was considered confirmatory. Strains were scored as negative if they failed these criteria under all conditions tested. The nutrient agar formulations used were based around RMAR medium under aerobic conditions, varying sea salt from 0 to 6% (w/v), cultivation temperature from 20 to 45 °C, with or without peptone, and using agar concentrations of 0.8, 1.5 and 2.5% (w/v). Most tests were conducted on plates containing nigrosin, but this was omitted if the dye appeared to inhibit growth. In addition, strains obtained from the DSMZ were cultivated on the recommended medium supplemented with 200 mg/l nigrosin to give optical contrast.

Microscopy and motility testing

Surface motility was judged by direct observation of the colony edge by microscopy and observing spread over several days, with visualisation of SC when necessary using side illumination with a 50W white LED lamp, all as previously described (470). Swimming was tested in 0.2% (w/v) sloppy RMAR agar (504).

Genome sequencing

DNA was isolated using the Qiagen UltraClean Microbial DNA isolation kit (Qiagen, Venlo, NL). DNA libraries were prepared with the Illumina Nextera kit according to manufacturer's instructions and sequenced using NextSeq sequencing with 150 base pairs reads (Illumina, San Diego, CA, USA). Reads-quality-check and adapter

trimming was performed with Trim Galore v0.4.4 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The genomes were assembled with SPAdes v3.10.1 (ref. 505), and contigs smaller than 200 base pairs and with a *k*-mer coverage lower than 10 were removed. Genome quality was assessed with CheckM v1.1.2 in ‘--lineage_wf’ mode for completeness (>95%) and contamination (<5%) (51). Reads were submitted to the SRA under accession PRJEB47515.

Pan-GWAS approach and phylogeny

Genes on the genomes were predicted using Prokka v1.13 (ref. 276), followed by orthology clustering using Roary v3.12.0 (ref. 480) using a 20% amino acid identity cut-off and a relaxed MCL granularity parameter ‘-iv 1.3’. Orthologs from Roary were associated with the structural colour phenotype using Scoary (481) using default settings including permutation testing (1,000×). Orthologs were considered significantly associated with SC when the Fisher Exact algorithm in Scoary showed a Benjamini Hochberg adjusted $P < 0.05$, or when the contrasting pair algorithm in Scoary showed $P < 0.05$, or when the permutation testing $P < 0.001$. In all cases only orthologs with an odds ratio >1 were selected. Functional enrichment analysis was performed using STRING (483) and genes were manually assigned to clusters; “gliding”, “carbohydrate”, “pterin”, “porphyrin”, “acetolactate” and “methionine” based on gene functions and interactions between proteins. Phylogenetic trees were constructed using RaxML v8.2.4 (ref. 506), using the BINCAT model for gene presence absence and the GTR-gamma model for full length 16S genes, and visualised using iTOL (507).

Identification of structural colour related genes using transposon mutagenesis

Genes inferred to be involved in SC were identified by transposon mutagenesis of *Flavobacterium* IR1 (ref. 470). This dataset included those previously identified (470) and 6 additional genes identified in a subsequent round of screening (**Supplementary Table 2**).

Machine learning model construction

All proteins from the 117 isolates assigned to the orthologs selected from the pan-GWAS approaches and the transposon mutagenesis data were extracted from the respective genomes using roary-query_pan_genome (480) and aligned per orthologous group using MAFFT v7.407 (ref. 285) with default settings. Hidden Markov Models were constructed using HMMer v3.1b2 (<http://hmmer.org/>) using default settings. HMM profiles were aligned against all proteins with an *e*-value cut-off of $1e-30$. Presence absence data of the HMM profiles was used as input for randomForest v4.6-14 (ref. 508) with 5,000 trees and by making use of the `sampsize=(c(20,20))` option to handle class imbalances to generate a prediction model. A script in bash was constructed that automates the HMM profile searches

and predicts SC using randomForest. The script and the associated random forest model is available on GitHub (<https://github.com/aldertzomer/structuralcolour>). An online version of the prediction method is available on <http://klif.uu.nl/structuralcolourweb/>.

Public genome and metagenome classification

All bacterial genomes with a valid species annotation (240,981 in total) were downloaded from the PATRIC genome database (185) on 14 November 2019. Assembled metagenomes (13,873 unique assemblies) and associated metadata were obtained from MGnify (152) on 3 July 2019. In addition, the metagenomes from sinking particulate organic matter (502) were downloaded from SRA and individually assembled with MEGAHIT v1.2.9 (ref. 205). Proteins were predicted in all genomes and metagenomes using Prodigal v.2.60 (ref. 290) using either default settings or the metagenome settings, respectively.

Optical analysis of structural colour

Angle-dependent spectra were taken using a custom-built goniometer setup (501). On this setup, the sample was mounted on a rotating stage so that the angle of incident light could be varied. The incident light from a Ocean Optics HPX-2000 xenon lamp was collimated, with a spot size of 5 mm diameter. Light reflected or scattered from the sample was then collected by an optical fibre connected to a spectrometer (AvaSpec-HS2048, Avantes). This optical fibre was mounted on a rotating arm so that the angle of detection could be varied. At the detection angle which equals the negative of the incident angle, the detection arm blocked the incident light so that no signal could be collected. All the spectra reported here were normalised against a white diffuser (labsphere SRS-99-010).

Electron microscopy

Cryo-SEM was performed on a FEI Verios 460 scanning electron microscope at the Cambridge Advanced Imaging Centre (University of Cambridge). A piece of agar (approximately 0.5×0.2 cm) with bacteria was cut and placed on a small piece of filter paper, which was then placed in a shuttle well containing colloidal graphite paste to hold the agar slice in place during freezing and fracturing and to provide conductivity. Care was taken to not cover the upper part of the sample with graphite paste so as to not contaminate the fracture plane. Next, the shuttle containing the sample was plunge-frozen in liquid ethane and subsequently transferred to a cryo-transfer system (Quorum PP3010T) that was cooled down to approximately -140 °C. The samples were then fractured with a blade, sublimated at -90 °C and sputter-coated with platinum at 10mA for 60 seconds. The images were taken at 2.00 keV acceleration voltage with 6.3 pA probe current using the EDT detector in field-free mode for low magnification images and the TLD detector in immersion-mode for high magnification images.

Data availability

Genomic sequence data is available under accession PRJEB56913. The aligned sequence data, HMMs and scripts for classifying sequence data for structural colour are available at DOI 10.5281/zenodo.7859454. A web based interface to the classifier is available at <http://klif.uu.nl/structuralcolourweb/>.

Acknowledgements

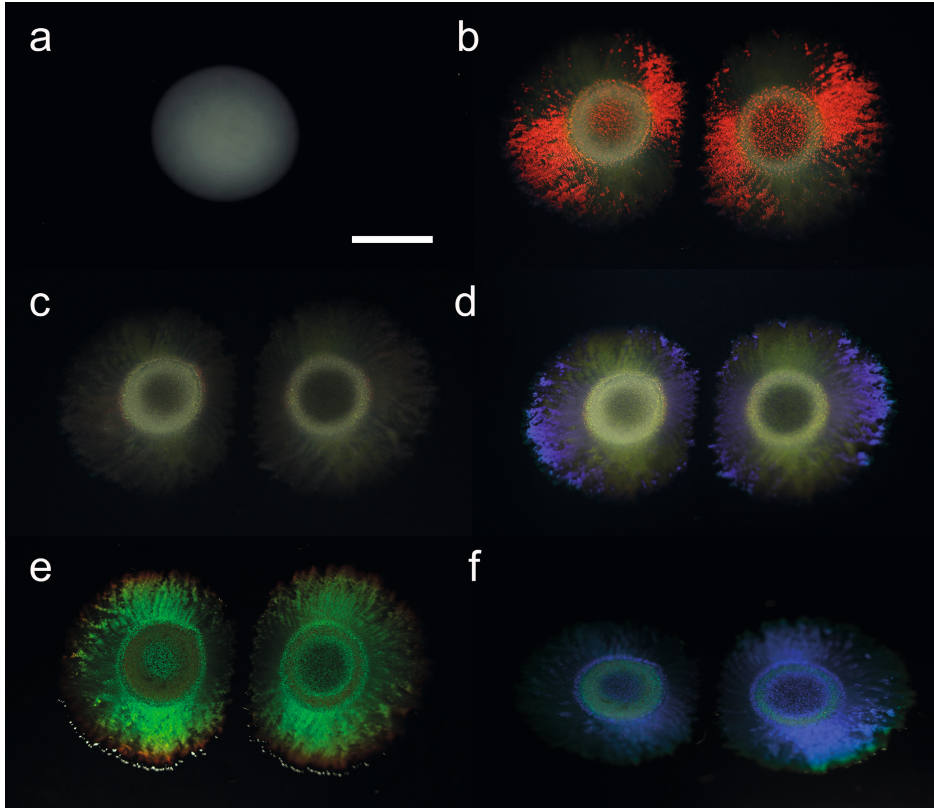
Thanks to Jens Harder, Anne Olsen, J-P Bernadet, Eric Duchaud, Francois Thomas, Gurvan Michel, Jeffrey Gralnick and Lars Jelsbak for strains and information on phenotyping. Thanks to the MBA Lab, Plymouth (UK) and CNRS Roscoff (FR) for access to sampling facilities via the ASSEMBLE Plus program of the EU and the iLAB, Utrecht for access to lab facilities (Hoekmine).

Funding

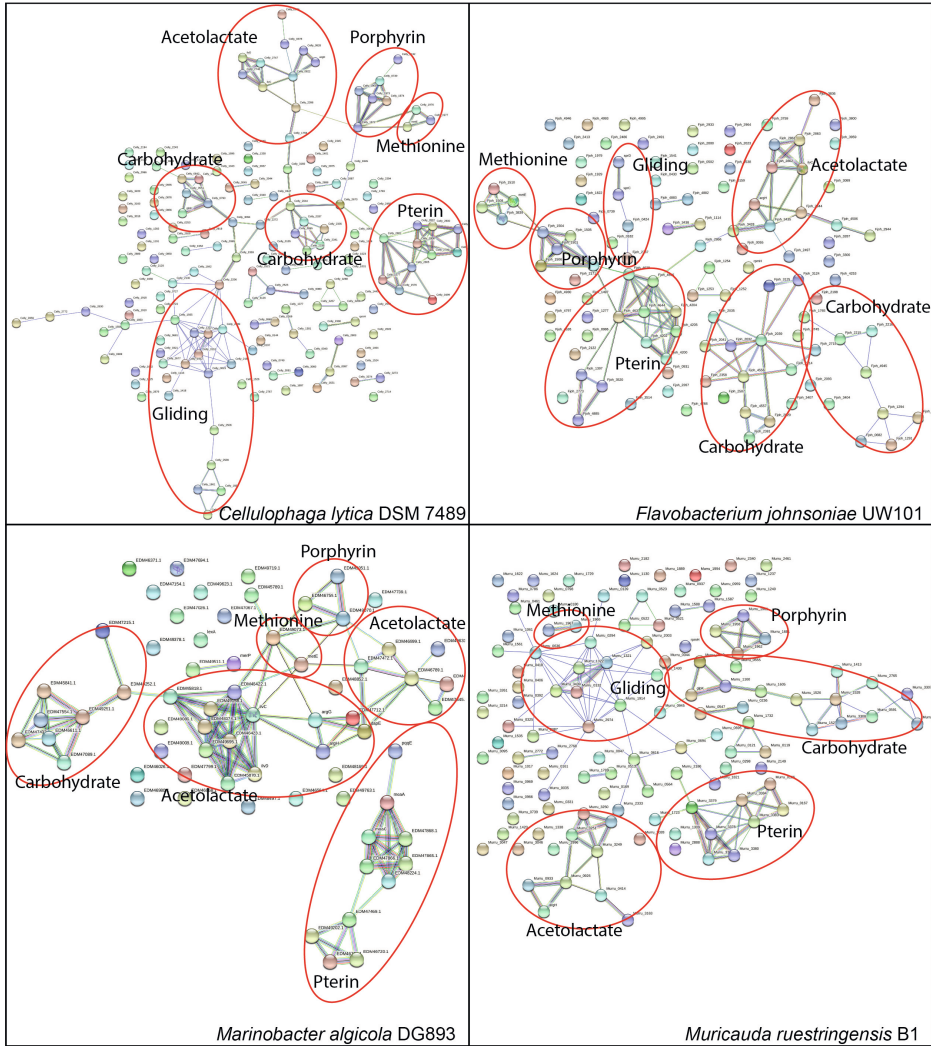
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860125 (C.J.I., A.E.D., S.V.), ZonMW Enabling Technologies Hotels grant 40-43500-98-4102/435004516 (H.B.), BBSRC UK iCASE fellowship 2110570 (L.C.). the EU's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreements No. 722842 (ITN Plant-inspired and Surfaces-PlaMatSu) (G.v.d.K.), the Swiss National Science Foundation under project P2ZHP2_183998 (L.S.), the Isaac Newton Trust (grant SNSF3) (L.S.), the Swiss national Science Foundation SNSF 40B1-0_198708 (L.S.), the European Research Council (ERC) Consolidator grant 865694: DiversiPHI (B.E.D.), the ERC Consolidator grant 101001637: BiTe (S.V., L.C.) and the BBSRC grant BB/V00364X/1 (S.V., L.C.), the Alexander von Humboldt Foundation in the context of an Alexander von Humboldt-Professorship founded by German Federal Ministry of Education and Research (B.E.D.), and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2051 – Project-ID 390713860.

Supplementary Information

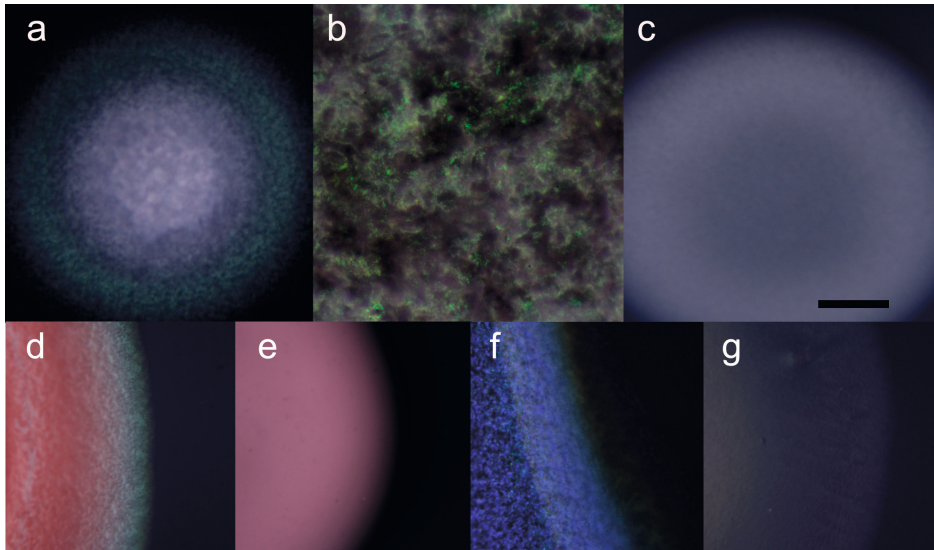
Supplementary Figures



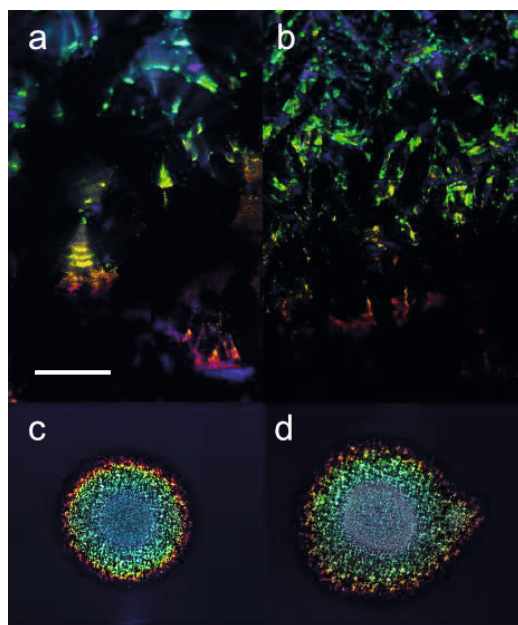
Supplementary Fig. 1 | Images of colonies ‘*Marinobacter algicola* HM-28’ showing photonic properties. **a**, Colony of this bacterium mixed to show pigmented colouration with the mechanically disrupted structural colour. **b–f**, Images of the same colonies, illuminated by white light, taken from different angles to show the range of structural colours. Scale bar indicates 1 cm for all panels.



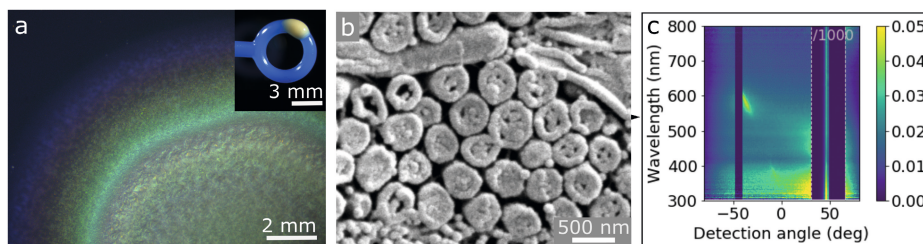
Supplementary Fig. 2 | STRING analysis showing functional clusters. STRING image of *Cellulophaga lytica*, *Flavobacterium johnsoniae*, *Marinobacter algicola*, and *Muricauda ruestringensis*, showing the six functional clusters: pterin, porphyrin, carbohydrate, methionine, acetolactate biosynthesis, and gliding motility.



Supplementary Fig. 3 | Examples of structural colour in bacteria from strains used in validation. **a**, Pinpoint colony of *Pseudomonas aeruginosa* cultured on TSA agar (0.8% w/v) plates at 30 °C showing green colouration at the edges when illuminated from the side. **b**, Microscopy image from the edge of the colony from the previous panel, showing pointillistic, saturated motes of green. **c**, Colony from panel **a**, mixed and redeposited on agar, showing loss of green colouration by mechanical disruption. **d,e**, Images of a colony of '*Hoeflea alexandrii* HHA1' grown on RMAR agar (0.8% w/v) plates at 30 °C. Panel **d** shows the intact colony with green SC and red pigmentation. Panel **e** shows that, after mixing the colony, the SC but not the pigmentation is lost. **f,g**, Images of a colony of *Microbulbifer* sp. grown on RMAR agar (0.8% w/v) plates at 30 °C. Panel **f** shows an intact colony, panel **g** shows a mixed colony. Scale bar indicates 100 μm for panels **a,c-g**, 20 μm for panel **b**. All plates contained 0.05% nigrosin to give contrast and were imaged by microscopy ($\times 40$ magnification) with illumination from the side using a 50W white LED.



Supplementary Fig. 4 | The effect of a flagellar motility gene disruption on structural colour in *Marinobacter subterrani* JG233. Panels **a** and **c** are the WT strain, panels **b** and **d** a *flaBG* knockout that is not capable of flagellar motility. Panels **a** and **b** are taken by low power microscopy with side illumination. Panels **c** and **d** show colonies after 5 days incubation. All images show growth on RMAR plates (0.8% w/v agar) at 29 °C. The scale bar indicates 0.4 mm for panels **a** and **b**, and 12 mm for panels **c** and **d**.



Supplementary Fig. 5 | Cell ordering and optical response of a colony of *Marinobacter subterrani* JG233. **a**, Colony of *Marinobacter subterrani* JG233 showing SC, inset shows material from the colony losing SC after mixing with an inoculation loop. **b**, Cryo-SEM image of *Marinobacter subterrani* JG233 in cross section. **c**, Angle-dependent spectra showing the optical response of the colony. The intensity of the reflected light is given on a blue to yellow heat map, with yellow high intensity and blue low intensity. The incident light angle is kept at -45° , and due to a limitation in the setup no spectra can be recorded at the incident light angle. Because the mirror-like reflection (specular) around 45° is far brighter than the scattered light at other angles, and can therefore not be shown on the same scale without saturating the signal, the reflected light intensity between the dashed lines has been divided by the number shown at the top.

Supplementary Tables

Supplementary Tables 1–10 (captions below) are available from Zenodo at <https://doi.org/10.5281/zenodo.8090260>.

Supplementary Table 1 | 117 bacterial genome sequences were used to create the random forest classifier for structural colour. The table lists their taxonomic affiliation, structural colour phenotype (our observations), accession number for the genome sequence, source of the strain (DSMZ, the German Strain Collection of the Leibniz Institute, this work, or as cited), and the cultivation media. All strains were cultivated on agar plates under aerobic conditions with media: ASWB, Artificial Seawater Agar Black; RMAR, Rich Marine Agar; LA, Luria Agar; DSMZ, as recommended for the strain in the culture collection (<https://www.dsmz.de/collection/catalogue/microorganisms/catalogue>).

Supplementary Table 2 | Proteins implicated in structural colour in *Flavobacterium* IRI were identified by transposon mutagenesis (this study and refs) with NCBI annotation. The table lists the mutant identifier, independent isolate identifiers, NCBI annotation of the protein function, NCBI accession identifier, and the associated citation. This dataset was used to create the structural colour classifier.

Supplementary Table 3 | Quantitative details of Fig. 2.

Supplementary Table 4 | Strains used in validation study. The table lists the species and strain name, taxonomic classification (class and phylum), Gram staining, source and habitat of the strains (either as part of this work from the DSMZ strain collection or otherwise obtained as cited), predicted SC score, and the SC phenotype (YES for structural colour under at least one growth condition on agar plates and NO if structural colour was not observed under any condition).

Supplementary Table 5 | The RF-based machine learning model was used to assign a SC score to 240,981 sequences downloaded from the PATRIC database. The table contains the PATRIC identifier, SC score, the complete taxonomic lineage, completeness and contamination information according to CheckM and PATRIC, and the status of the sequence (complete, plasmid, or whole genome shotgun).

Supplementary Table 6 | Range of RF-based SC scores of species containing at least ten genomes, as shown by the blue lines in Fig. 3b. The listed values are the mean of the five lowest and the five highest SC scores.

Supplementary Table 7 | Proteobacterial strains tested for swimming in sloppy agar plates. The table lists the species and strain name, the agar percentage of the plates, culture medium (Rich Marine Agar; LA, Luria Agar; DSMZ), and the colony expansion rate in millimetres per

Chapter 6

day. Spreading can only occur beyond a few mm if the bacteria use flagella motility to swim through the sloppy agar.

Supplementary Table 8 | List of 13,873 assembled metagenomes from the MGnify database.

The table lists the assembly file, identifier of the MGnify analysis and the assembly, the biome of the sample, and the SC score.

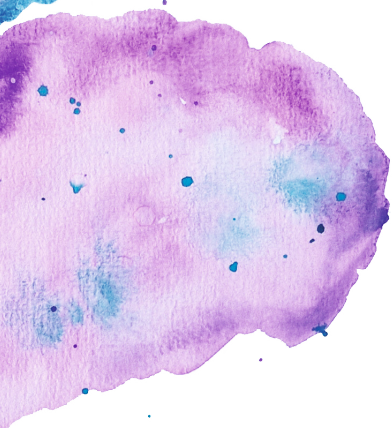
Supplementary Table 9 | List of 866 MGnify metagenomes from the root:Environmental:Aquatic:Marine biome lineage.

Columns are identical to **Supplementary Table 8**, plus the depth in the water column, number of contigs of at least 500 bp, and number of predicted proteins on all contigs.

Supplementary Table 10 | List of 62 metagenomes from sinking particulate organic matter (marine snow) that were assembled with MEGAHIT.

The table lists the run identifier, collection date, water depth, number of contigs of at least 500 bp, number of predicted proteins on all contigs, and the SC score.





Chapter 7

Discussion



In the Introduction of this thesis, I placed my research in light of what came before, describing the three pillars it builds upon: (i) the motivation to circumvent the cultivation bias and discover novel biology, (ii) the generation and (re)analysis of large-scale data, and (iii) the development of algorithms and tools to interpret that data. Here, turning to the future, I will project how these pillars will change in the coming years.

Pillar one: what is left to discover?

The chapters of this thesis describe the exploration of DNA that is sequenced directly from the environment, and the uncharted biology that these sequences represent. A compelling question is how much novelty is still left to discover. The moment I write this (February 2023), the MGnify database houses environmental sequencing data from 491 different annotated biomes. The database contains 356,309 amplicon, 33,827 metagenomic, and 2,205 metatranscriptomic analyses, and 28,873 assemblies. The BV-BRC (previously PATRIC) database contains 651,054 genomes from 72,766 prokaryotic species. There is no doubt that these numbers will grow vastly (122,152). Will there be diminishing returns in terms of biological discovery?

There may be over 1×10^{12} prokaryotic species on Earth (509). The vast expanse of undiscovered species can represent different levels of taxonomic novelty. The deeper branches of the tree of life may by now have been uncovered, in which case taxonomic novelty will only be found in close relatives to known microorganisms. Alternatively, deep branches may still have evaded detection, and superphyla (92) or even domains of life (510) are waiting to be discovered. The only way to find out is to look for them. Uncharted sequences hide in two places: in microbiomes that have not been sequenced before, and in low abundance below the detection limit of those that have.

Uncharted sequences hide in unexplored microbiomes and at low abundance

Remote habitats hold promise for the discovery of novel taxa, like the deep subsurface or the bottom of the ocean. As metagenomic sequencing will become standard procedure in environmental surveys, the microbiomes of remote environments that can be reached will be uncovered, such as those from a subsurface coal seam (511) or a subglacial Antarctic lake (512). Habitats closer to home are also underexplored, like a local pond, or host-associated habitats that do not have immediate economic value and have thus largely been ignored. The microbiomes of 101 marine fish species were recently taxonomically characterised with 16S rRNA gene amplicon sequencing (513), and future metagenomic sequencing will reveal their functional potential, along with microorganisms that escape detection with 16S rRNA gene primers. Even habitats that are extensively studied have not been taxonomically exhausted.

An analysis of human gut microbiomes from nearly 4,000 individuals from the Netherlands, Belgium, the United Kingdom, and the United States, found a shared core of only 17 genera, out of 664 in total (140), based on 16S rRNA gene amplicon sequencing. It was estimated that another 40,739 individuals had to be sampled to capture all genera present in Western guts.

The majority of microorganisms exists in low abundance in microbiomes, and they may represent a microbial seed bank for when conditions are favourable and store ecological potential (452). Uncharted sequences hide in low abundance, and will be uncovered by deeper sequencing. After the data in Chapter 2 were published, the same samples from the Black Sea water column were sequenced again, much deeper this time. 45 million reads were sequenced for Chapter 2, for this second round we sequenced over 4 billion. 90 medium- to high-quality MAGs were binned from the original dataset, of which four are discussed in this thesis (others are discussed elsewhere (70)). With the new dataset I binned 5,070. Many of the new MAGs reside in the low-abundant tail of the rank-abundance curve.

An alternative to deeper sequencing is sampling the habitat multiple times. A microorganism that is low-abundant now may be abundant at another time (452). In light of Chapter 5, sampling a habitat multiple times will uncover social generalists, with likely relatively large genomes in habitats with low local diversity, and relatively small genomes in habitats with high local diversity. Social specialists are uncovered by deeper sequencing.

Novel biology hides in known sequences

Surprising biology also awaits discovery in already published data. In Chapter 2, the 'archaeal-type' lipid biosynthesis genes of *Candidatus* Cloacimonetes were found in the Black Sea, but also showed up in *Ca.* Cloacimonetes genomes in other habitats and throughout the FCB group superphylum. The significance of the presence of these genes was not recognised when the other genomes were sequenced. Nobody may have looked for them. Lipid biosynthesis genes are rarely screened across the tree of life or in the environment (249,514).

Many genes involved in lipid biosynthetic pathways have just recently been discovered, in part because lipids constitute diverse molecules that cannot be uniformly built from smaller building blocks like polymeric macromolecules such as nucleic acids and proteins. Examples are the recent discoveries of genes responsible for the synthesis of membrane spanning lipids in archaea (515) and in bacteria (516), and for the synthesis of ring structures in archaeal membrane spanning lipids that are used as indicators of past sea surface temperature when found in the sediment (517). Now that their function is established, these genes can be searched in known sequences for potentially surprising taxonomic associations

and presence in unexpected habitats, as was done in Chapter 6 with genes involved in structural colour.

These recently discovered lipid biosynthesis genes and the identification of structural colour genes in Chapter 6 highlight that phenotype prediction based on the genome sequence is still in its infancy for many traits. Structural colour is a complex phenotype involving various genes and different genes in different microorganisms, that were identified in Chapter 6 by combining specialised lab techniques with genomics in a pan-genome-wide association study. The genomic basis of other traits is still unknown, especially for recently discovered microorganisms. Genes with unknown function are numerous in MAGs.

We have spent the last two decades uncovering “who is there”, the challenge will now be to uncover “what they are doing” beyond the pathways that are known. In addition, genetic potential encoded on the genome does not equal expression in the environment, and what a microorganism does may differ in time and place. The focus of discovery will advance from identifying surprising branches of the tree of life to elucidating condition-dependent phenotype of microorganisms from known branches in their habitat. This will be accomplished by sequencing metagenomes deeper together with the collection of other datatypes that are directly observed in the microbiome, like metatranscriptomes and meta-metabolomes. In short, more data are coming.

Pillar two: future data and uncovering phenotype

When new microorganisms are uncovered with metagenomics, we can confidently find their place in the tree of life. The debate on how to formally name MAGs that do not have cultured representatives is ongoing (518), but we fundamentally know “who they are” via phylogenomics. The second question, “what are they doing?”, is more difficult to answer. The genome sequence encodes the potential to express functions. A large number of genes in MAGs have unknown function or are only distantly related to homologs with known function. In addition, under which conditions genes are expressed and result in a phenotype cannot be answered based on the genome sequence alone. Finally, plasmids and other mobile genetic elements that may confer important functions are usually not binned with MAGs of their host because sequence composition and read depth, signals that are used for binning, are often different for plasmids and chromosomes.

Are we left again with the daunting task of persuading the uncultured majority into culture (39)? Luckily, we can also get insights into what microorganisms are doing in the natural habitat, via metagenomics and via other methods that directly interrogate the microbiome.

Metagenomics provides insights into what microorganisms are doing

Microbial characteristics that cannot be seen in the genome sequence can be found in the metagenome. Co-occurrence and abundance correlations reveal interactions between microorganisms (24,25), and will likely aid in identifying hosts for the CPR bacteria and DPANN archaea discussed in the Introduction (99), together with evidence of horizontal gene transfer between host and symbiont (98). Time-series data are still rare today but will improve understanding of microbial community dynamics. Co-occurrence and abundance correlations will become more informative as more metagenomes are sequenced and deeper sequencing will uncover low-abundant microorganisms.

An exciting new development is the estimation of in situ growth rates from the metagenome. In contrast to maximal growth rate estimates that were discussed in the Introduction and in Chapter 5 and that are based on codon-usage bias, in situ growth rate is estimated from uneven sequencing read depth across the genome. DNA replication of a circular chromosome is bidirectional from the origin of replication towards the terminus, and a growing population has more DNA copies of regions near the origin of replication because these regions are duplicated earlier than regions near the terminus. The difference in read depth between the origin of replication and the terminus in a metagenome reflects growth rate (519). This reasoning has been applied to MAGs for which the origin of replication is unknown by sorting contigs according to read depth and calculating overall slope (520,521). Although estimated and observed in situ growth rates still have low correlation (522), it is likely that deeper sequencing will reduce noise that affects slope calculations and thereby improve predictive power.

Finally, new experimental protocols and sequencing technologies will advance metagenomics. Proximity-based ligation of DNA with Hi-C (523) before metagenomic sequencing uncovers sequences that are physically close within cells. Third-generation single-molecule sequencing technologies (see **'Pillar three: future computation'**) can identify modified bases in DNA and RNA and so report epigenetics of the metagenome. Both Hi-C and epigenetic signals have already been used to bin plasmids and other genetic elements with the chromosome of the host (30,524).

The microbiome can be observed with other meta-omics approaches

Together with metagenomics, other methods that directly interrogate the microbiome are referred to as 'meta-omics' approaches, and they represent different intermediates on the path from genotype to phenotype.

Metatranscriptomics shows all genes that are expressed in the microbiome (see **Box 1** in the Introduction), and shotgun mass spectrometry-based metaproteomics shows

the proteins that are present. A discrepancy between functional potential (as seen in the metagenome) and expressed functions (as seen in the metatranscriptome and metaproteome) is observed in the human gut (525,526), and for example gene expression may change in relation to disease of the host when the microbial community does not (527). Metagenomics combined with metatranscriptomics established that microorganisms were actively involved in hydrocarbon degradation in the Deep water Horizon oil spill (528), and metagenomics combined with metaproteomics has elucidated uptake of organic matter throughout the marine water column (529). Permafrost in different stages of thawing has been characterised with metagenomics, metatranscriptomics, and metaproteomics, revealing that high rates of methane production are related to methanogenesis in thermokarst bogs (530).

Untargeted meta-metabolomics shows the metabolites and small molecules in the microbiome, that are the product of enzymatic reactions, and that mediate interactions between microorganisms, as well as many other small molecules in the environment. Recently, hundreds of microbiomes from a wide range of habitats collected by the Earth Microbiome Project (EMP) were characterised with metagenomics and meta-metabolomics, and samples from different annotated biomes could be separated based on microbial taxa—in line with the findings of Chapter 5, and based on metabolites (531). Meta-metabolomics differs from the earlier discussed meta-omics approaches in that it does not depict genes or their translation products. Metabolites are still largely anonymous molecules, and can thus not readily be assigned to specific microorganisms yet as can be done with genes, RNA, and proteins. To some extent, the meta-metabolome can be mechanistically predicted from the metagenome with genome-scale metabolic models, linking metabolic potential in the genome to the metabolic environment (171).

Methodologically related to metaproteomics and meta-metabolomics is the emerging field of untargeted environmental lipidomics (532,533)—a field still young enough to not have been consistently named ‘meta’lipidomics yet. Metalipidomics allows for the discovery of uncharted lipid molecules (the larger molecules in the metabolome) in the environment. 930 metalipidomes from the Atlantic and Pacific Oceans were recently analysed, revealing a strong correlation between fatty-acid saturation and temperature in surface waters (534).

Integrating different datatypes at the global scale

Although other meta-omics approaches are not new, they have yet to reach the same large-scale adoption as metagenomics. As amplicon sequencing preceded metagenomics, metagenomics is now ahead of the other meta-omics approaches. Part of the reason is that raw data from metaproteomics, meta-metabolomics, and metalipidomics is not as readily generated and interpretable as DNA

sequences. However, large-scale meta-metabolomics studies like that of the EMP foreshadow what will come. Time-series that are analysed with metagenomics, metatranscriptomics, metaproteomics, and meta-metabolomics are already available (535).

In Chapter 2, I used RNA transcripts to confirm gene expression and molecules from the environment combined with DNA abundance estimates to confirm that the archaeal-type lipids were produced by *Ca. Cloacimonetes*. This shows that integration of different datatypes improves biological comprehension. A future in which habitats across the world are characterised with metagenomics and other meta-omics approaches, and in which these data are freely available, will facilitate discovery. For example, presence of metabolites or lipids of interest for which the genes involved are unknown can be linked to gene expression in the microbiome, identifying possible candidates. Gene expression or protein abundance in relation to environmental parameters across samples will reveal under which conditions pathways are expressed, and effects can be observed in the meta-metabolome. The integration of meta-omics and other biological data will help in providing a mechanistic understanding of what shapes microbial communities, including cross-feeding and other metabolic dependencies. In Chapter 5, I showed that the microbial community itself can define the niche of a microorganism, and that different ecological and evolutionary strategies are associated with wide and narrow niche ranges. Future meta-omics studies will further a functional understanding of these findings.

At the base of all meta-omics approaches is metagenomics. Ultimately, the metagenome encodes everything that happens in the microbiome, and it is thus crucial for understanding the metatranscriptome, metaproteome, meta-metabolome, and metalipidome. As habitats around the world will be observed with different meta-omics approaches, the metagenome will continue to be sequenced along. Tools and algorithms will be instrumental to its interpretation.

Pillar three: future computation

Algorithms and tools for metagenomic interpretation evolve and change. Two trends will shape their future: the increasingly large scale of metagenomic datasets, and the move from short-read next-generation sequencing to long-read single-molecule third-generation sequencing technologies.

Larger datasets require new approaches to metagenomic interpretation

Sequencing projects generate ever more sequencing reads due to declining sequencing costs and availability of sequencing machines with higher throughput, and methods for metagenomic interpretation adapt to this. Adaptation to larger datasets is visible

in this thesis. In Chapter 2, I assembled the sequencing reads from the 15 Black Sea water column samples together into a single cross-assembly. The benefit of cross-assembly compared to assembling sequencing reads from individual samples is that low abundant microorganisms are more likely to be assembled into contigs if they are present in multiple samples—the combined pool contains more of their DNA. The risk is that chimeras are generated, as sequencing reads from different samples that do not originate from the same microbial strain can be spuriously assembled together into contigs, and I had to address the possibility of chimeras to confirm that archaeal-type lipid biosynthesis genes were present in *Ca. Cloacimonetes* genomes. The metagenomes of 18 groundwater samples in Chapter 4 were sequenced a couple of years later and were too large and complex for cross-assembly. Assembly is computer memory intensive, especially for highly diverse metagenomes, and the sequencing reads could not be cross-assembled even on our largest server with 1.5 terabytes memory. We therefore assembled the sequencing reads per sample individually.

Assembly of sequencing reads from individual samples eliminates the risk of between-sample chimeras, but may recover (almost) identical DNA sequences from the same microbial strain multiple times—one from each sample where the strain is found. To alleviate interpretation and reduce computational costs, for example of phylogenomic placement of many highly similar MAGs, the 423 medium to high quality groundwater MAGs were compared based on sequence similarity with dRep (367) and 195 groups of nearly identical genomes were identified. One representative genome per group was placed in a phylogenomic tree. This redundancy would have been reduced in the assembly step with cross-assembly.

The development of assemblers that require less memory is ongoing (205), and as the size of metagenomic datasets increases some of the algorithms and tools that are used today will become unfeasible because of impractical memory requirement or runtime, and will be replaced by others.

Chapter 4 also showed another trend in the field: the development of pipelines that automate common procedures. In Chapter 2 I ran every tool myself, in Chapter 4 the ATLAS pipeline was used for the steps of genome-resolved metagenomics from sequencing reads to MAGs (372). Pipelines make metagenomics more accessible and reduce time spent on repetitive tasks.

Sifting out noise in taxonomic annotations

The CAT pack software suite that was introduced in Chapter 3 uses nr as its default reference database that contains (translated) protein sequences from curated and non-curated sources. Annotations in nr are provided by the dataset submitters, that are prone to error for both function (536) and taxonomy. For example, eukaryotic DNA can be misclassified as prokaryotic because of unidentified contamination

in sequencing labs (157) or because eukaryotic contigs are binned in prokaryotic MAGs (158), and conversely prokaryotic sequences are sometimes misclassified as eukaryotic (159). Microbial genome sequences have been unknowingly contaminated with PhiX, a common control in Illumina sequencing (160). Taxonomic annotations in nr can be a best guess and incorrect. A screen of the nr database from 2018 identified over two million misclassified proteins (161).

The CAT pack is robust against misclassifications because the taxonomic signals of all ORFs on a sequence are considered, at the cost of annotating uncharted sequences too high in the taxonomy. When an ORF is annotated to an incorrect clade, conflicting taxonomic signals between ORFs will result in a classification at a high taxonomic rank. As metagenomic sequencing projects have scaled up, misclassifications in nr may become more prevalent, especially if proteins are annotated with a best-hit approach which propagates database errors. In my experience, the CAT pack is providing more conservative annotations with more recent databases, likely due to an increase in misclassifications.

A solution is to use a curated reference database, as was introduced in Chapter 4 with the non-redundant set of proteins from GTDB. An alternative approach is to identify and correct misclassifications (161,537). This will be the next big update to the CAT pack.

Metagenomics embraces long sequencing reads

Long in the making, third-generation sequencing technologies promise a revolution (538). Single-molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio) works by direct observation of DNA polymerase that is attached to a well (539), and nanopore sequencing by Oxford Nanopore Technologies (ONT) detects individual bases as they pass through a pore (540). Both SMRT and nanopore sequencing generate long sequencing reads (>10 kb) from a single molecule. Both technologies allow for the detection of modified bases, and RNA can be sequenced directly via nanopores without reverse transcription (541). The ONT MinION sequencer is so small that it fits in a pocket and has been used for real-time sequencing in the field (542). Initial high error rates in long-read sequencing have declined, and PacBio now promises similar accuracy to Illumina sequences with its HiFi sequencing.

Long-read sequencing holds the promise of bridging regions that are difficult to assemble into contigs with short sequencing reads, like repeats within genomes and regions that are shared between genomes of different microorganisms in the community (543). Long sequencing reads with relatively high error rates pose a problem for complex metagenomes, because the sequencing depth required to assemble accurate sequences is not met for low-abundant microorganisms. While sequencing quality and throughput of third-generation sequencing increase,

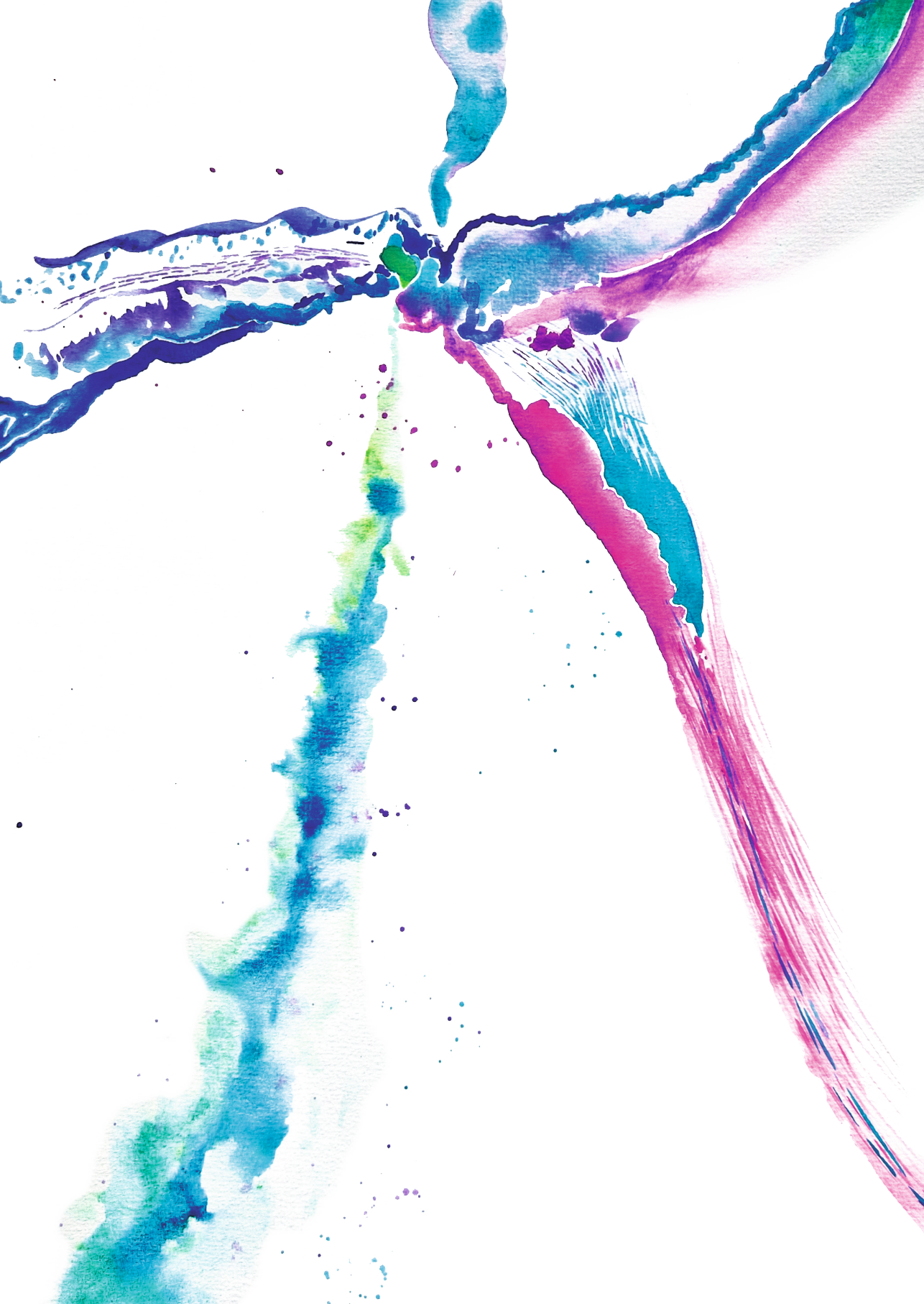
Chapter 7

combining shallow long-read sequencing with deep short-read sequencing to improve accuracy is a compromise (544). 1,083 high-quality MAGs including 57 complete circular genomes have been reconstructed by combining long-read nanopore sequences with short-read Illumina data from wastewater treatment plants (545).

Throughput, read length, sequencing quality, and cost are the defining factors for large-scale adoption, and as the technologies advance, long sequencing reads will become more prevalent. Algorithms and tools will have to adapt—numerous tools that I used in this thesis will be obsolete for high-quality long-read data, like Burrows-Wheeler Transform aligners, and maybe eventually even de Bruijn graph assemblers and bidders. Other tools will fill their place. A study in 2020 already counted 354 long-read analysis tools (358). More will surely follow.

Conclusion of this thesis

This thesis paints a picture of a microbial world that is still large left uncharted. With the development of new tools and algorithms, and the (re)analysis of large-scale data, I have uncovered a fraction of that vast microbial unknown.



Appendices

References
Acknowledgements
Curriculum vitae
List of publications

References

1. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA* **95**, 6578–6583 (1998).
2. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biol* **14**, e1002533 (2016).
3. Mendes, R., Garbeva, P. & Raaijmakers, J. M. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiol Rev* **37**, 634–663 (2013).
4. Olson, J. M. Photosynthesis in the Archean era. *Photosynth Res* **88**, 109–197 (2005).
5. Allen, J. F. & Martin, W. Out of thin air. *Nature* **445**, 610–612 (2007).
6. Kopp, R. E., Kirschvink, J. L., Hilburn, I. A. & Nash, C. Z. The Paleoproterozoic snowball Earth: A climate disaster triggered by the evolution of oxygenic photosynthesis. *Proc Natl Acad Sci USA* **102**, 11131–11136 (2005).
7. Warke, M. R. *et al.* The Great Oxidation Event preceded a Paleoproterozoic “snowball Earth”. *Proc Natl Acad Sci USA* **117**, 13314–13320 (2020).
8. Raymond, J. & Segrè, D. The Effect of Oxygen on Biochemical Networks and the Evolution of Complex Life. *Science* **311**, 1764–1767 (2006).
9. Knoll, A. H. & Nowak, M. A. The timetable of evolution. *Sci Adv* **3**, e1603076 (2017).
10. Ren, M. *et al.* Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. *ISME J* **13**, 2150–2161 (2019).
11. Desbrosses, G. J. & Stougaard, J. Root Nodulation: A Paradigm for How Plant-Microbe Symbiosis Influences Host Developmental Pathways. *Cell Host Microbe* **10**, 348–358 (2011).
12. Santi, C., Bogusz, D. & Franche, C. Biological nitrogen fixation in non-legume plants. *Ann Bot* **111**, 743–767 (2013).
13. Benemann, J. R. Nitrogen Fixation in Termites. *Science* **181**, 164–165 (1973).
14. Breznak, J. A., Brill, W. J., Mertins, J. W. & Coppel, H. C. Nitrogen Fixation in Termites. *Nature* **244**, 577–580 (1973).
15. Sharma, R., Garg, P., Kumar, P., Bhatia, S. K. & Kulshrestha, S. Microbial Fermentation and Its Role in Quality Improvement of Fermented Foods. *Ferment* **6**, 106 (2020).
16. Hauptfeld, E. *et al.* A metagenomic portrait of the microbial community responsible for two decades of bioremediation of poly-contaminated groundwater. *Water Res* **221**, 118767 (2022).
17. V., L. S. & Oluf, P. The Human Intestinal Microbiome in Health and Disease. *New Engl J Med* **375**, 2369–2379 (2016).
18. Oliphant, K. & Allen-Vercoe, E. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. *Microbiome* **7**, 91 (2019).
19. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat Med* **24**, 392–400 (2018).
20. Morais, L. H., Schreiber, H. L. & Mazmanian, S. K. The gut microbiota–brain axis in behaviour and brain disorders. *Nat Rev Microbiol* **19**, 241–255 (2021).
21. McFall-Ngai, M. *et al.* Animals in a bacterial world, a new imperative for the life sciences. *Proc Natl Acad Sci USA* **110**, 3229–3236 (2013).
22. Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T. & Singh, B. K. Plant–microbiome interactions: from community assembly to plant health. *Nat Rev Microbiol* **18**, 607–621 (2020).
23. Freilich, S. *et al.* Competitive and cooperative metabolic interactions in bacterial communities. *Nat Commun* **2**, 589 (2011).
24. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat Rev Microbiol* **10**, 538–50 (2012).
25. Machado, D. *et al.* Polarization of microbial communities between competitive and cooperative metabolism. *Nat Ecol Evol* **5**, 195–203 (2021).

26. Palmer, J. D. & Foster, K. R. Bacterial species rarely work together. *Science* **376**, 581–582 (2022).
27. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
28. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**, 472–482 (2015).
29. Vos, M., Hesselman, M. C., Beek, T. A. te, van Passel, M. W. J. & Eyre-Walker, A. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends Microbiol* **23**, 598–605 (2015).
30. Yaffe, E. & Relman, D. A. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat Microbiol* **5**, 343–353 (2019).
31. Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2015).
32. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**, 669–685 (2004).
33. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99**, 14250–14255 (2002).
34. Venter, J. C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
35. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
36. Staley, J. T. & Konopka, A. Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annu Rev Microbiol* **39**, 321–346 (1985).
37. Stahl, D. A., Lane, D. J., Olsen, G. J. & Pace, N. R. Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences. *Appl Environ Microb* **49**, 1379–1384 (1983).
38. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *Msystems* **3**, e00055-18 (2018).
39. Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z. & Ettema, T. J. G. Innovations to culturing the uncultured microbial majority. *Nat Rev Microbiol* **19**, 225–240 (2020).
40. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996–1004 (2018).
41. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12**, 635–645 (2014).
42. Baker, B. J. *et al.* Diversity, ecology and evolution of Archaea. *Nat Microbiol* **5**, 887–900 (2020).
43. Woese, C. R. Bacterial evolution. *Microbiol Rev* **51**, 221–71 (1987).
44. Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* **82**, 6955–6959 (1985).
45. Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**, D141–D145 (2009).
46. DeSantis, T. Z. *et al.* Greengenes, a chimeric-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microb* **72**, 5069–72 (2006).
47. Schloss, P. D., Girard, R. A., Martin, T., Edwards, J. & Thrash, J. C. Status of the Archaeal and Bacterial Census: an Update. *Mbio* **7**, e00201-16 (2016).
48. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**, 833–844 (2017).
49. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
50. Hugerth, L. W. *et al.* Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol* **16**, 279 (2015).

Appendices

51. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043–1055 (2015).
52. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725–731 (2017).
53. Quail, M. A. *et al.* Optimal enzymes for amplifying sequencing libraries. *Nat Methods* **9**, 10–11 (2012).
54. van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: Tone down the bias. *Exp Cell Res* **322**, 12–20 (2014).
55. Farrelly, V., Rainey, F. A. & Stackebrandt, E. Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl Environ Microb* **61**, 2798–801 (1995).
56. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. K. & Schmidt, T. M. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res* **43**, D593–8 (2014).
57. Louca, S., Doebeli, M. & Parfrey, L. W. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* **6**, 41 (2018).
58. Sun, Z. *et al.* Challenges in benchmarking metagenomic profilers. *Nat Methods* **18**, 618–626 (2021).
59. Guo, F. & Zhang, T. Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. *Appl Microbiol Biotechnol* **97**, 4607–4616 (2013).
60. Sato, M. P. *et al.* Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *Dna Res Int J Rapid Publ Reports Genes Genomes* **26**, 391–398 (2019).
61. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol* **14**, R51 (2013).
62. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform* **3**, lqab019 (2021).
63. Bashiardes, S., Zilberman-Schapira, G. & Elinav, E. Use of Metatranscriptomics in Microbiome Research. *Bioinform Biology Insights* **10**, 19–25 (2016).
64. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
65. Zheng, W. *et al.* High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome. *Science* **376**, eabm1483 (2022).
66. Alneberg, J. *et al.* Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* **6**, 173 (2018).
67. Speth, D. R., Zandt, M. H. in 't, Guerrero-Cruz, S., Dutilh, B. E. & Jetten, M. S. M. Genome-based microbial ecology of anammox granules in a full-scale wastewater treatment system. *Nat Commun* **7**, 11172 (2016).
68. Zaneveld, J. R., McMinds, R. & Thurber, R. V. Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nat Microbiol* **2**, 17121 (2017).
69. Zmora, N., Suez, J. & Elinav, E. You are what you eat: diet, health and the gut microbiota. *Nat Rev Gastroenterol* **16**, 35–56 (2019).
70. van Vliet, D. M. *et al.* The bacterial sulfur cycle in expanding dysoxic and euxinic marine waters. *Environ Microbiol* **23**, 2834–2857 (2021).
71. Coutinho, F. H. *et al.* Ecogenomics and metabolic potential of the South Atlantic Ocean microbiome. *Sci Total Environ* **765**, 142758 (2021).
72. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**, R245–R249 (1998).
73. Hugenholtz, P., Goebel, B. M. & Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180**, 4765–74 (1998).

74. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3**, reviews0003.1 (2002).
75. Heuer, V. B. *et al.* Temperature limits to deep seafloor life in the Nankai Trough subduction zone. *Science* **370**, 1230–1234 (2020).
76. Beulig, F. *et al.* Rapid metabolism fosters microbial survival in the deep, hot seafloor biosphere. *Nat Commun* **13**, 312 (2022).
77. Orsi, W. D., Edgcomb, V. P., Christman, G. D. & Biddle, J. F. Gene expression in the deep biosphere. *Nature* **499**, 205–208 (2013).
78. Trembath-Reichert, E. *et al.* Methyl-compound use and slow growth characterize microbial life in 2-km-deep seafloor coal and shale beds. *Proc Natl Acad Sci USA* **114**, E9206–E9215 (2017).
79. Flemming, H.-C. & Wuertz, S. Bacteria and archaea on Earth and their abundance in biofilms. *Nat Rev Microbiol* **17**, 247–260 (2019).
80. Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C. & D'Hondt, S. Global distribution of microbial abundance and biomass in seafloor sediment. *Proc Natl Acad Sci USA* **109**, 16213–6 (2012).
81. Magnabosco, C. *et al.* The biomass and biodiversity of the continental subsurface. *Nat Geosci* **11**, 707–717 (2018).
82. McMahon, S. & Parnell, J. Weighing the deep continental biosphere. *FEMS Microbiol Ecol* **87**, 113–120 (2014).
83. Rappé, M. S., Connon, S. A., Vergin, K. L. & Giovannoni, S. J. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**, 630–633 (2002).
84. Henson, M. W., Lanclos, V. C., Faircloth, B. C. & Thrash, J. C. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. *ISME J* **12**, 1846–1860 (2018).
85. Gutleben, J. *et al.* The multi-omics promise in context: from sequence to microbial isolate. *Crit Rev Microbiol* **44**, 212–229 (2018).
86. Karnachuk, O. V. *et al.* Targeted isolation based on metagenome-assembled genomes reveals a phylogenetically distinct group of thermophilic spirochetes from deep biosphere. *Environ Microbiol* **23**, 3585–3598 (2021).
87. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).
88. Rodrigues-Oliveira, T. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* **613**, 332–339 (2023).
89. Baum, D. A. & Baum, B. An inside-out origin for the eukaryotic cell. *BMC Biol* **12**, 76 (2014).
90. Löwe, J. Mysterious Asgard archaea microbes reveal their inner secrets. *Nature* **613**, 246–248 (2023).
91. Sharon, I. & Banfield, J. F. Genomes from Metagenomics. *Science* **342**, 1057–1058 (2013).
92. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–11 (2015).
93. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**, 13219 (2016).
94. Castelle, C. J. *et al.* Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. *Curr Biol* **25**, 690–701 (2015).
95. Williams, T. A. *et al.* Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc Natl Acad Sci USA* **114**, E4602–E4611 (2017).
96. Sakai, H. D. *et al.* Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses. *Proc Natl Acad Sci USA* **119**, e2115449119 (2022).
97. Castelle, C. J. & Banfield, J. F. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).
98. Dombrowski, N. *et al.* Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat Commun* **11**, 3939 (2020).

Appendices

99. He, C. *et al.* Genome-resolved metagenomics reveals site-specific diversity of epismymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat Microbiol* **6**, 354–365 (2021).
100. Jahn, U., Summons, R., Sturt, H., Grosjean, E. & Huber, H. Composition of the lipids of Nanoarchaeum equitans and their origin from its host Ignicoccus sp. strain KIN4/I. *Arch Microbiol* **182**, 404–413 (2004).
101. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–9 (2015).
102. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
103. Liu, Y. *et al.* Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).
104. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).
105. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).
106. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”. *Proc Natl Acad Sci USA* **102**, 13950–13955 (2005).
107. Lefébure, T., Bitar, P. D. P., Suzuki, H. & Stanhope, M. J. Evolutionary Dynamics of Complete Campylobacter Pan-Genomes and the Bacterial Species Concept. *Genome Biol Evol* **2**, 646–655 (2010).
108. Dutilh, B. E. Metagenomic ventures into outer sequence space. *Bacteriophage* **4**, e979664 (2014).
109. Ramirez, K. S. *et al.* Biogeographic patterns in below-ground diversity in New York City’s Central Park are similar to those observed globally. *Proc Royal Soc B Biological Sci* **281**, 20141988 (2014).
110. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).
111. Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P. & Baptiste, E. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biol Evol* **10**, 707–715 (2018).
112. Ciccarelli, F. D. *et al.* Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science* **311**, 1283–1287 (2006).
113. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* **15**, 579–590 (2017).
114. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).
115. Libis, V. *et al.* Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. *Nat Commun* **10**, 3848 (2019).
116. Schmid, A. K., Allers, T. & DiRuggiero, J. SnapShot: Microbial Extremophiles. *Cell* **180**, 818–818.e1 (2020).
117. Shu, W.-S. & Huang, L.-N. Microbial diversity in extreme environments. *Nat Rev Microbiol* **20**, 219–235 (2022).
118. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* **5**, 4498 (2014).
119. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol* **4**, 1727–1736 (2019).
120. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
121. Galperin, M. Y. Metagenomics: from acid mine to shining sea. *Environ Microbiol* **6**, 543–545 (2004).
122. Katz, K. *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res* **50**, D387–D390 (2022).
123. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**, 5463–5467 (1977).

124. Fleischmann, R. D. *et al.* Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
125. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
126. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
127. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**, 1051–1063 (2008).
128. Schloss, J. A. How to get genomes at one ten-thousandth the cost. *Nat Biotechnol* **26**, 1113–5 (2008).
129. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet* **30**, 418–426 (2014).
130. Methé, B. A. *et al.* A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
131. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
132. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
133. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
134. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**, 834–841 (2014).
135. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
136. Knights, D. *et al.* Rethinking “Enterotypes”. *Cell Host Microbe* **16**, 433–437 (2014).
137. Vatanen, T. *et al.* The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).
138. Gigliucci, F. *et al.* Metagenomic Characterization of the Human Intestinal Microbiota in Fecal Samples from STEC-Infected Patients. *Front Cell Infect Mi* **8**, 25 (2018).
139. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–9 (2016).
140. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
141. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLOS Biol* **5**, e77 (2007).
142. Yooseph, S. *et al.* Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**, 60–66 (2010).
143. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol* **18**, 428–445 (2020).
144. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
145. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun* **9**, 870 (2018).
146. Littman, R., Willis, B. L. & Bourne, D. G. Metagenomic analysis of the coral holobiont during a natural bleaching event on the Great Barrier Reef. *Env Microbiol Rep* **3**, 651–660 (2011).
147. Cox-Foster, D. L. *et al.* A Metagenomic Survey of Microbes in Honey Bee Colony Collapse Disorder. *Science* **318**, 283–287 (2007).
148. Brito, T. L. *et al.* The gill-associated microbiome is the main source of wood plant polysaccharide hydrolases and secondary metabolite gene clusters in the mangrove shipworm *Neoteredo reynei*. *PLOS One* **13**, e0200437 (2018).
149. Nóbrega, M. S. *et al.* Mangrove microbiome reveals importance of sulfur metabolism in tropical coastal waters. *Sci Total Environ* **813**, 151889 (2022).
150. Anantharaman, K., Breier, J. A. & Dick, G. J. Metagenomic resolution of microbial functions in deep-sea hydrothermal plumes across the Eastern Lau Spreading Center. *ISME J* **10**, 225–239 (2016).

Appendices

151. Dijkhuizen, L. W. *et al.* Is there foul play in the leaf pocket? The metagenome of floating fern *Azolla* reveals endophytes that do not fix N₂ but may denitrify. *New Phytologist* **217**, 453–466 (2018).
152. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* **48**, D570–D578 (2020).
153. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res* **39**, D28–D31 (2011).
154. Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res* **48**, D84–D86 (2019).
155. Okido, T. *et al.* DNA Data Bank of Japan (DDBJ) update report 2021. *Nucleic Acids Res* **50**, D102–D105 (2021).
156. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **47**, D23–D28 (2018).
157. Breitwieser, F. P., Pertea, M., Zimin, A. V. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res* **29**, 954–960 (2019).
158. Arkhipova, I. R. Metagenome Proteins and Database Contamination. *Mosphere* **5**, e00854-20 (2020).
159. Merchant, S., Wood, D. E. & Salzberg, S. L. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* **2**, e675 (2014).
160. Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand Genomic Sci* **10**, 18 (2015).
161. Bagheri, H., Severin, A. & Rajan, H. Detecting and correcting misclassified sequences in the large-scale public databases. *Bioinformatics* **36**, 4699–4705 (2020).
162. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
163. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61–D65 (2007).
164. Piganeau, G. & Moreau, H. Screening the Sargasso Sea metagenome for data to investigate genome evolution in *Ostreococcus* (Prasinophyceae, Chlorophyta). *Gene* **406**, 184–190 (2007).
165. Tringe, S. G. *et al.* Comparative Metagenomics of Microbial Communities. *Science* **308**, 554–557 (2005).
166. Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).
167. Smith, M. B. *et al.* Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors. *Mbio* **6**, e00326-15 (2015).
168. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**, 766 (2014).
169. Gupta, V. K. *et al.* A predictive index for health status using species-level gut microbiome profiling. *Nat Commun* **11**, 4635 (2020).
170. Zolti, A., Green, S. J., Sela, N., Hadar, Y. & Minz, D. The microbiome as a biosensor: functional profiles elucidate hidden stress in hosts. *Microbiome* **8**, 71 (2020).
171. Garza, D. R., van Verk, M. C., Huynen, M. A. & Dutilh, B. E. Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat Microbiol* **3**, 456–460 (2018).
172. Fahimipour, A. K. & Gross, T. Mapping the bacterial metabolic niche space. *Nat Commun* **11**, 4887 (2020).
173. Thompson, L. R. *et al.* A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
174. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
175. Vangay, P. *et al.* Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative’s Workshop and Follow-On Activities. *Msystems* **6**, e01194-20 (2021).
176. Leite, M. F. A., Broek, S. W. E. B. van den & Kuramae, E. E. Current Challenges and Pitfalls in Soil Metagenomics. *Microorg* **10**, 1900 (2022).

177. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32**, 822–828 (2014).
178. Glendinning, L., Stewart, R. D., Pallen, M. J., Watson, K. A. & Watson, M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. *Genome Biol* **21**, 34 (2020).
179. Woodcroft, B. J. *et al.* Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
180. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* **3**, 804–813 (2018).
181. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).
182. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat Biotechnol* **39**, 499–509 (2021).
183. Woese, C. R. A manifesto for microbial genomics. *Curr Biol* **8**, R780–R783 (1998).
184. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060 (2009).
185. Davis, J. J. *et al.* The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res* **48**, D606–D612 (2019).
186. Vieira-Silva, S. & Rocha, E. P. C. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLOS Genet* **6**, e1000808 (2010).
187. Weissman, J. L., Hou, S. & Fuhrman, J. A. Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proc Natl Acad Sci USA* **118**, e2016810118 (2021).
188. Kariin, S. & Burge, C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**, 283–290 (1995).
189. Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases. *Genome Res* **13**, 145–158 (2003).
190. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**, 938–947 (2004).
191. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**, R85–R85 (2009).
192. Iverson, V. *et al.* Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science* **335**, 587–590 (2012).
193. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144–1146 (2014).
194. Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**, 111–120 (2013).
195. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533–538 (2013).
196. Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, e603 (2014).
197. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
198. Lin, H.-H. & Liao, Y.-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* **6**, 24175 (2016).
199. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
200. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

Appendices

201. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821–829 (2008).
202. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**, 455–477 (2012).
203. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. Why are de Bruijn graphs useful for genome assembly? *Nat Biotechnol* **29**, 987–991 (2011).
204. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824–834 (2017).
205. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–6 (2014).
206. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230 (2012).
207. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
208. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
209. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
210. Mavromatis, K. *et al.* Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* **4**, 495–500 (2007).
211. Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* **178**, 779–794 (2019).
212. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat Methods* **14**, 1063–1071 (2017).
213. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat Methods* **19**, 429–440 (2022).
214. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* **8**, 2224 (2017).
215. Koga, Y. & Morii, H. Biosynthesis of Ether-Type Polar Lipids in Archaea and Evolutionary Considerations. *Microbiol Mol Biol R* **71**, 97–120 (2007).
216. Koga, Y., Kyuragi, T., Nishihara, M. & Sone, N. Did Archaeal and Bacterial Cells Arise Independently from Noncellular Precursors? A Hypothesis Stating That the Advent of Membrane Phospholipid with Enantiomeric Glycerophosphate Backbones Caused the Separation of the Two Lines of Descent. *J Mol Evol* **47**, 631–631 (1998).
217. Martin, W. & Russell, M. J. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philosophical Transactions Royal Soc Lond Ser B Biological Sci* **358**, 59–85 (2003).
218. Koonin, E. V. & Martin, W. On the origin of genomes and cells within inorganic compartments. *Trends Genet* **21**, 647–654 (2005).
219. Sojo, V., Pomiankowski, A. & Lane, N. A Bioenergetic Basis for Membrane Divergence in Archaea and Bacteria. *PLOS Biol* **12**, e1001926 (2014).
220. Lombard, J., López-García, P. & Moreira, D. The early evolution of lipid membranes and the three domains of life. *Nat Rev Microbiol* **10**, 507–515 (2012).
221. Peretó, J., López-García, P. & Moreira, D. Ancestral lipid biosynthesis and early membrane evolution. *Trends Biochem Sci* **29**, 469–477 (2004).

222. Koga, Y. Early Evolution of Membrane Lipids: How did the Lipid Divide Occur? *J Mol Evol* **72**, 274–282 (2011).
223. Wächtershäuser, G. From pre-cells to Eukarya—a tale of two lipids. *Mol Microbiol* **47**, 13–22 (2003).
224. Shimada, H. & Yamagishi, A. Stability of Heterochiral Hybrid Membrane Made of Bacterial sn-G3P Lipids and Archaeal sn-G1P Lipids. *Biochemistry* **50**, 4114–4120 (2011).
225. Caforio, A. *et al.* Converting *Escherichia coli* into an archaeobacterium with a hybrid heterochiral membrane. *Proc Natl Acad Sci USA* **115**, 3704–3709 (2018).
226. López-García, P. & Moreira, D. Open Questions on the Origin of Eukaryotes. *Trends Ecol Evol* **30**, 697–708 (2015).
227. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nat Rev Microbiol* **15**, 711–723 (2017).
228. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).
229. Archibald, J. M. The eocyte hypothesis and the origin of eukaryotic cells. *Proc Natl Acad Sci USA* **105**, 20049–20050 (2008).
230. Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. The archaeobacterial origin of eukaryotes. *Proc Natl Acad Sci USA* **105**, 20356–20361 (2008).
231. Moreira, D. & López-García, P. Symbiosis Between Methanogenic Archaea and δ -Proteobacteria as the Origin of Eukaryotes: The Syntrophic Hypothesis. *J Mol Evol* **47**, 517–530 (1998).
232. López-García, P. & Moreira, D. Selective forces for the origin of the eukaryotic nucleus. *Bioessays* **28**, 525–533 (2006).
233. López-García, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655–667 (2020).
234. Gould, S. B., Garg, S. G. & Martin, W. F. Bacterial Vesicle Secretion and the Evolutionary Origin of the Eukaryotic Endomembrane System. *Trends Microbiol* **24**, 525–534 (2016).
235. Langworthy, T. A., Holzer, G., Zeikus, J. G. & Tornabene, T. G. Iso- and Anteiso-Branched Glycerol Diethers of the Thermophilic Anaerobe *Thermodesulfotobacterium commune*. *Syst Appl Microbiol* **4**, 1–17 (1983).
236. Huber, R. *et al.* *Aquifex pyrophilus* gen. nov. sp. nov., Represents a Novel Group of Marine Hyperthermophilic Hydrogen-Oxidizing Bacteria. *Syst Appl Microbiol* **15**, 340–351 (1992).
237. Huber, R. *et al.* Formation of Ammonium from Nitrate During Chemolithoautotrophic Growth of the Extremely Thermophilic Bacterium *Ammonifex degensii* gen. nov. sp. nov. *Syst Appl Microbiol* **19**, 40–49 (1996).
238. Rosa, M. D. *et al.* A new 15,16-dimethyl-30-glyceroyloxytriacontanoic acid from lipids of *Thermotoga maritima*. *J Chem Soc Chem Commun* **0**, 1300 (1988).
239. Sinninghe Damsté, J. S. *et al.* Linearly concatenated cyclobutane lipids form a dense bacterial membrane. *Nature* **419**, 708–712 (2002).
240. Sinninghe Damsté, J. S. *et al.* Structural characterization of diabolic acid-based tetraester, tetraether and mixed ether/ester, membrane-spanning lipids of bacteria from the order Thermotogales. *Arch Microbiol* **188**, 629–641 (2007).
241. Weijers, J. W. H. *et al.* Membrane lipids of mesophilic anaerobic bacteria thriving in peats have typical archaeal traits. *Environ Microbiol* **8**, 648–657 (2006).
242. Sinninghe Damsté, J. S. *et al.* 13,16-Dimethyl Octacosanedioic Acid (iso-Diabolic Acid), a Common Membrane-Spanning Lipid of Acidobacteria Subdivisions I and 3. *Appl Environ Microb* **77**, 4147–4154 (2011).
243. Sinninghe Damsté, J. S. *et al.* An overview of the occurrence of ether- and ester-linked iso-diabolic acid membrane lipids in microbial cultures of the Acidobacteria: Implications for brGDGT paleoproxies for temperature and pH. *Org Geochem* **124**, 63–76 (2018).

244. Sinninghe Damsté, J. S. *et al.* A mixed ladderane/n-alkyl glycerol diether membrane lipid in an anaerobic ammonium-oxidizing bacterium. *Chem Commun* **0**, 2590–2591 (2004).
245. Gattinger, A., Schloter, M. & Munch, J. C. Phospholipid etherlipid and phospholipid fatty acid fingerprints in selected euryarchaeotal monocultures for taxonomic profiling. *FEMS Microbiol Lett* **213**, 133–139 (2002).
246. Dibrova, D. V., Galperin, M. Y. & Mulki-djanian, A. Y. Phylogenomic reconstruction of archaeal fatty acid metabolism. *Environ Microbiol* **16**, 907–918 (2014).
247. Lombard, J., López-García, P. & Moreira, D. Phylogenomic Investigation of Phospholipid Synthesis in Archaea. *Archaea* **2012**, 630910 (2012).
248. Rinke, C. *et al.* A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.). *ISME J* **13**, 663–675 (2018).
249. Villanueva, L., Schouten, S. & Sinninghe Damsté, J. S. Phylogenomic analysis of lipid biosynthetic genes of Archaea shed light on the ‘lipid divide’. *Environ Microbiol* **19**, 54–69 (2017).
250. Coleman, G. A., Pancost, R. D. & Williams, T. A. Investigating the Origins of Membrane Phospholipid Biosynthesis Genes Using Outgroup-Free Rooting. *Genome Biol Evol* **11**, 883–898 (2019).
251. Lyons, T. W. Ironing Out Ocean Chemistry at the Dawn of Animal Life. *Science* **321**, 923–924 (2008).
252. Parsons, J. B. & Rock, C. O. Bacterial lipids: Metabolism and membrane homeostasis. *Prog Lipid Res* **52**, 249–276 (2013).
253. Peterhoff, D. *et al.* A comprehensive analysis of the geranylgeranyl glyceryl phosphate synthase enzyme family identifies novel members and reveals mechanisms of substrate specificity and quaternary structure organization. *Mol Microbiol* **92**, 885–899 (2014).
254. Guldan, H., Matysik, F.-M., Bocola, M., Sterner, R. & Babinger, P. Functional Assignment of an Enzyme that Catalyzes the Synthesis of an Archaea-Type Ether Lipid in Bacteria. *Angewandte Chemie Int Ed* **50**, 8188–8191 (2011).
255. Yokoi, T., Isobe, K., Yoshimura, T. & Hemmi, H. Archaeal Phospholipid Biosynthetic Pathway Reconstructed in *Escherichia coli*. *Archaea* **2012**, 438931 (2012).
256. Isobe, K. *et al.* Geranylgeranyl Reductase and Ferredoxin from *Methanosarcina acetivorans* Are Required for the Synthesis of Fully Reduced Archaeal Membrane Lipid in *Escherichia coli* Cells. *J Bacteriol* **196**, 417–423 (2014).
257. Caforio, A. *et al.* Formation of the ether lipids archaeetidylglycerol and archaeetidylethanolamine in *Escherichia coli*. *Biochem J* **470**, 343–355 (2015).
258. Guldan, H., Sterner, R. & Babinger, P. Identification and Characterization of a Bacterial Glycerol-1-phosphate Dehydrogenase: Ni²⁺-Dependent AraM from *Bacillus subtilis*. *Biochemistry* **47**, 7376–7384 (2008).
259. Chen, A., Zhang, D. & Poulter, C. D. (S)-geranylgeranyl glyceryl phosphate synthase. Purification and characterization of the first pathway-specific enzyme in archaeobacterial membrane lipid biosynthesis. *J Biological Chem* **268**, 21701–5 (1993).
260. Zhang, D. & Poulter, C. D. Biosynthesis of Archaeobacterial lipids in *Halobacterium halobium* and *Methanobacterium thermoautotrophicum*. *J Org Chem* **58**, 3919–3922 (1993).
261. Schouten, S., Hopmans, E. C. & Sinninghe Damsté, J. S. The organic geochemistry of glycerol dialkyl glycerol tetraether lipids: A review. *Org Geochem* **54**, 19–61 (2013).
262. Harvey, H. R., Fallon, R. D. & Patton, J. S. The effect of organic matter and oxygen on the degradation of bacterial membrane lipids in marine sediments. *Geochim Cosmochim Acta* **50**, 795–804 (1986).

263. Sollai, M., Villanueva, L., Hopmans, E. C., Reichart, G.-J. & Sinninghe Damsté, J. S. A combined lipidomic and 16S rRNA gene amplicon sequencing approach reveals archaeal sources of intact polar lipids in the stratified Black Sea water column. *Geobiology* **17**, 91–109 (2019).
264. Schouten, S., Middelburg, J. J., Hopmans, E. C. & Sinninghe Damsté, J. S. Fossilization and degradation of intact polar lipids in deep subsurface sediments: A theoretical approach. *Geochim Cosmochim Acta* **74**, 3806–3814 (2010).
265. Sturt, H. F., Summons, R. E., Smith, K., Elvert, M. & Hinrichs, K.-U. Intact polar membrane lipids in prokaryotes and sediments deciphered by high-performance liquid chromatography/electrospray ionization multistage mass spectrometry—new biomarkers for biogeochemistry and microbial ecology. *Rapid Commun Mass Sp* **18**, 617–628 (2004).
266. Buckles, L. K., Villanueva, L., Weijers, J. W. H., Verschuren, D. & Sinninghe Damsté, J. S. Linking isoprenoidal GDGT membrane lipid distributions with gene abundances of ammonia-oxidizing Thaumarchaeota and uncultured crenarchaeotal groups in the water column of a tropical lake (Lake Challa, East Africa). *Environ Microbiol* **15**, 2445–2462 (2013).
267. Hopmans, E. C., Schouten, S. & Sinninghe Damsté, J. S. The effect of improved chromatography on GDGT-based palaeoproxies. *Org Geochem* **93**, 1–6 (2016).
268. Schouten, S., Hugué, C., Hopmans, E. C., Kienhuis, M. V. M. & Sinninghe Damsté, J. S. Analytical Methodology for TEX₈₆ Paleothermometry by High-Performance Liquid Chromatography/Atmospheric Pressure Chemical Ionization-Mass Spectrometry. *Anal Chem* **79**, 2940–2944 (2007).
269. Hugué, C. *et al.* An improved method to determine the absolute abundance of glycerol dibiphytanyl glycerol tetraether lipids. *Org Geochem* **37**, 1036–1041 (2006).
270. Holmes, D. E., Nevin, K. P. & Lovley, D. R. In Situ Expression of *nifD* in Geobacteraceae in Subsurface Sediments. *Appl Environ Microb* **70**, 7251–7259 (2004).
271. Moore, E. K. *et al.* Abundant Trimethylornithine Lipids and Specific Gene Sequences Are Indicative of Planctomycete Importance at the Oxidic/Anoxic Interface in Sphagnum-Dominated Northern Wetlands. *Appl Environ Microb* **81**, 6333–6344 (2015).
272. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
273. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590–D596 (2013).
274. Dodt, M., Roehr, J., Ahmed, R. & Dieterich, C. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology* **1**, 895–905 (2012).
275. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv.org* <https://doi.org/10.48550/arXiv.1303.3997> (2013).
276. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
277. Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**, 75 (2008).
278. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
279. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res* **45**, D37–D42 (2017).
280. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).
281. Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).
282. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Haeseler, A. von & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587–589 (2017).
283. Hoang, D. T., Chernomor, O., Haeseler, A. von, Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518–522 (2018).

Appendices

284. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**, W242–W245 (2016).
285. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).
286. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
287. Villanueva, L., Sinninghe Damsté, J. S. & Schouten, S. A re-evaluation of the archaeal membrane lipid biosynthetic pathway. *Nat Rev Microbiol* **12**, 438–448 (2014).
288. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
289. Meijenfeldt, F. A. B. von, Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol* **20**, 217 (2019).
290. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
291. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res* **45**, D535–D542 (2017).
292. Li, C. *et al.* FastCloning: a highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. *BMC Biotechnol* **11**, 92 (2011).
293. Jain, S. *et al.* Identification of CDP-Archaeol Synthase, a Missing Link of Ether Lipid Biosynthesis in Archaea. *Chem Biol* **21**, 1392–1401 (2014).
294. Miroux, B. & Walker, J. E. Over-production of Proteins in *Escherichia coli*: Mutant Hosts that Allow Synthesis of some Membrane Proteins and Globular Proteins at High Levels. *J Mol Biol* **260**, 289–298 (1996).
295. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Express Purif* **41**, 207–234 (2005).
296. Rütters, H., Sass, H., Cypionka, H. & Rulíkötter, J. Phospholipid analysis as a tool to study complex microbial communities in marine sediments. *J Microbiol Meth* **48**, 149–160 (2002).
297. Besseling, M. A., Hopmans, E. C., Boschman, R. C., Sinninghe Damsté, J. S. & Villanueva, L. Benthic archaea as potential sources of tetraether membrane lipids in sediments across an oxygen minimum zone. *Biogeosciences* **15**, 4047–4064 (2018).
298. Grinsven, S. van *et al.* Methane oxidation in anoxic lake water stimulated by nitrate and sulfate addition. *Environ Microbiol* **22**, 766–782 (2019).
299. Salcher, M. M., Pernthaler, J. & Posch, T. Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria “that rule the waves” (LD12). *ISME J* **5**, 1242–52 (2011).
300. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**, 2725–9 (2013).
301. Dyksma, S. & Gallert, C. Candidatus Syntrophosphaera thermopropionivorans: a novel player in syntrophic propionate oxidation during anaerobic digestion. *Env Microbiol Rep* **11**, 558–570 (2019).
302. Wang, R. *et al.* Sulfur Oxidation in the Acidophilic Autotrophic Acidithiobacillus spp. *Front Microbiol* **9**, 3290 (2018).
303. Wagner, T., Koch, J., Ermler, U. & Shima, S. Methanogenic heterodisulfide reductase (HdrABC-MvhAGD) uses two noncubane [4Fe-4S] clusters for reduction. *Science* **357**, 699–703 (2017).
304. Volkov, I. I. & Neretin, L. N. Hydrogen Sulfide in the Black Sea. In Kostianoy, A.G. & Kosarev, A.N. (eds) *The Handbook of Environmental Chemistry: The Black Sea Environment* (Springer, 2008).
305. Reguera, G. *et al.* Extracellular electron transfer via microbial nanowires. *Nature* **435**, 1098–1101 (2005).
306. Stolze, Y. *et al.* Identification and genome reconstruction of abundant distinct taxa in microbiomes from one thermophilic and three mesophilic production-scale biogas plants. *Biotechnol Biofuels* **9**, 156 (2016).

307. Biebl, H. & Pfennig, N. Growth yields of green sulfur bacteria in mixed cultures with sulfur and sulfate reducing bacteria. *Arch Microbiol* **117**, 9–16 (1978).
308. Jannasch, H. W., Wirsen, C. O. & Molyneux, S. J. Chemoautotrophic sulfur-oxidizing bacteria from the Black Sea. *Deep Sea Res Part Oceanogr Res Pap* **38**, S1105–S1120 (1991).
309. Edgcomb, V. P. *et al.* Comparison of Niskin vs. in situ approaches for analysis of gene expression in deep Mediterranean Sea water samples. *Deep Sea Res Part II Top Stud Oceanogr* **129**, 213–222 (2016).
310. Murakami, M. *et al.* Geranylgeranyl reductase involved in the biosynthesis of archaeal membrane lipids in the hyperthermophilic archaeon *Archaeoglobus fulgidus*. *FEBS J* **274**, 805–814 (2007).
311. Yoshinaga, M. Y. *et al.* Systematic fragmentation patterns of archaeal intact polar lipids by high-performance liquid chromatography/electrospray ionization ion-trap mass spectrometry. *Rapid Commun Mass Sp* **25**, 3563–3574 (2011).
312. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biol* **59**, 307–321 (2010).
313. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
314. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
315. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol Biol Evol* **17**, 540–552 (2000).
316. Elling, F. J. *et al.* Effects of growth phase on the membrane lipid composition of the thaumarchaeon *Nitrosopumilus maritimus* and their implications for archaeal lipid distributions in the marine environment. *Geochim Cosmochim Acta* **141**, 579–597 (2014).
317. Sinninghe Damsté, J. S. *et al.* Distribution of Membrane Lipids of Planktonic Crenarchaeota in the Arabian Sea. *Appl Environ Microb* **68**, 2997–3002 (2002).
318. Schouten, S. *et al.* Intact polar and core glycerol dibiphytanyl glycerol tetraether lipids in the Arabian Sea oxygen minimum zone: I. Selective preservation and degradation in the water column and consequences for the TEX₈₆. *Geochim Cosmochim Acta* **98**, 228–243 (2012).
319. Ghuneim, L.-A. J., Jones, D. L., Golyshin, P. N. & Golyshina, O. V. Nano-Sized and Filterable Bacteria and Archaea: Biodiversity and Function. *Front Microbiol* **9**, 1971 (2018).
320. Holler, T. *et al.* Thermophilic anaerobic oxidation of methane by marine microbial consortia. *ISME J* **5**, 1946–56 (2011).
321. Kubo, K. *et al.* Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J* **6**, 1949–65 (2012).
322. Jahn, U. *et al.* Nanoarchaeum equitans and Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea. *J Bacteriol* **190**, 1743–50 (2007).
323. Golyshina, O. V. *et al.* ‘ARMAN’ archaea depend on association with euryarchaeal host in culture and in situ. *Nat Commun* **8**, 60 (2017).
324. Liang, P.-H., Ko, T.-P. & Wang, A. H.-J. Structure, mechanism and function of prenyltransferases. *Eur J Biochem* **269**, 3339–3354 (2002).
325. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* **3**, 31 (2017).
326. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46 (2014).
327. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**, 236 (2015).

Appendices

328. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* **7**, 11257 (2016).
329. Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E. & Edwards, R. A. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* **2**, e425 (2014).
330. Huson, D. H. *et al.* MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Comput Biol* **12**, e1004957 (2016).
331. Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**, 76 (2014).
332. Huson, D. H. *et al.* MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biol Direct* **13**, 6 (2018).
333. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res* **21**, 487–493 (2011).
334. Dutilh, B. E. *et al.* Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics* **23**, 815–824 (2007).
335. Guy, L. phyloSkeleton: taxon selection, data retrieval and marker identification for phylogenomics. *Bioinformatics* **33**, 1230–1232 (2017).
336. Gregor, I., Dröge, J., Schirmer, M., Quince, C. & McHardy, A. C. PhyloPythiaS+ : a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* **4**, e1603 (2016).
337. Dröge, J., Gregor, I. & McHardy, A. C. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics* **31**, 817–824 (2015).
338. Xia, H. *et al.* Comparative Metagenomic Profiling of Viromes Associated with Four Common Mosquito Species in China. *Virologica Sinica* **33**, 59–66 (2018).
339. Young, J. M., Skvortsov, T., Arkhipova, K. & Allen, C. C. R. Draft Genome Sequence of the Predatory Marine Bacterium *Halobacteriovorax* sp. Strain JY17. *Genome Announcements* **6**, e01416–17 (2018).
340. Bao, E. & Lan, L. HALC: High throughput algorithm for long read error correction. *BMC Bioinformatics* **18**, 204 (2017).
341. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722–736 (2017).
342. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res* **40**, D136–D143 (2012).
343. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).
344. Leinonen, R., Sugawara, H., Shumway, M. & Collaboration, I. N. S. D. The Sequence Read Archive. *Nucleic Acids Res* **39**, D19–D21 (2011).
345. Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* **2**, 63–77 (2012).
346. King, A. M. Q. *et al.* Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018). *Arch Virol* **163**, 2601–2631 (2018).
347. Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).
348. Truong, D. T. *et al.* MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**, 902–903 (2015).
349. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* **10**, 1014 (2019).
350. Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* **12**, S4 (2011).

351. Nguyen, N., Mirarab, S., Liu, B., Pop, M. & Warnow, T. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics* **30**, 3548–3555 (2014).
352. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257 (2019).
353. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* **26**, 1721–1729 (2016).
354. Silva, G. G. Z., Green, K. T., Dutilh, B. E. & Edwards, R. A. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics* **32**, 354–361 (2016).
355. Koslicki, D. & Falush, D. MetaPalette: a k-mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation. *Msystems* **1**, e00020-16 (2016).
356. Meijenfeldt, F. A. B. von, Hogeweg, P. & Dutilh, B. E. A social niche breadth score reveals niche range strategies of generalists and specialists. *Nat Ecol Evol* **7**, 768–781 (2023).
357. Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. *Annu Rev Microbiol.* **55**, 709–742 (2001).
358. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**, 30 (2020).
359. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res* **17**, 377–386 (2007).
360. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
361. Meyer, F. *et al.* Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. *Nat Protoc* **16**, 1785–1801 (2021).
362. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* **50**, D785–D794 (2021).
363. Aldas-Vargas, A. *et al.* Selective pressure on microbial communities in a drinking water aquifer – Geochemical parameters vs. micropollutants. *Environ Pollut* **299**, 118807 (2022).
364. Popa, O. & Dagan, T. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol* **14**, 615–623 (2011).
365. Mineeva, O., Rojas-Carulla, M., Ley, R. E., Schölkopf, B. & Youngblut, N. D. DeepMASeD: Evaluating the quality of metagenomic assemblies. *Bioinformatics* **36**, 3011–3017 (2020).
366. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
367. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**, 2864–2868 (2017).
368. McInerney, J. O., McNally, A. & O’Connell, M. J. Why prokaryotes have pangenomes. *Nat Microbiol* **2**, 17040 (2017).
369. Gillespie, J. J. *et al.* PATRIC: the Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species. *Infect Immun* **79**, 4286–4298 (2011).
370. McClelland, J. & Koslicki, D. EMDUniFrac: exact linear time computation of the UniFrac metric and identification of differentially abundant organisms. *J Math Biol* **77**, 935–949 (2018).
371. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput Sci* **3**, e104 (2017).
372. Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M. & McCue, L. A. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics* **21**, 257 (2020).
373. Tange, O. GNU Parallel - The Command-Line Power Tool. *login: The USENIX Magazine* **36**, 42–47 (2011).

Appendices

374. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256–W259 (2019).
375. Wickham, H. *ggplot2, Elegant Graphics for Data Analysis* (Springer, 2016).
376. Wickham, H. *et al.* Welcome to the Tidyverse. *J Open Source Softw* **4**, 1686 (2019).
377. Wickham, H. Reshaping Data with the reshape Package. *J Stat Softw* **21**, 1–20 (2007).
378. Brunson, J. C. *ggalluvial: Layered Grammar for Alluvial Plots. J open source Softw* **5**, 2017 (2020).
379. Dixon, P. VEGAN, a package of R functions for community ecology. *J Veg Sci* **14**, 927–930 (2003).
380. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018).
381. Villanueva, L. *et al.* Bridging the membrane lipid divide: bacteria of the FCB group superphylum have the potential to synthesize archaeal ether lipids. *ISME J* **15**, 168–182 (2021).
382. Futuyma, D. J. & Moreno, G. The Evolution of Ecological Specialization. *Annu Rev Ecol Syst* **19**, 207–233 (1988).
383. Carscadden, K. A. *et al.* Niche Breadth: Causes and Consequences for Ecology, Evolution, and Conservation. *Q Rev Biology* **95**, 179–214 (2020).
384. Sexton, J. P., Montiel, J., Shay, J. E., Stephens, M. R. & Slatyer, R. A. Evolution of Ecological Niche Breadth. *Annu Rev Ecol Syst* **48**, 183–206 (2017).
385. Muller, E. E. L. Determining Microbial Niche Breadth in the Environment for Better Ecosystem Fate Predictions. *Msystems* **4**, e00080-19 (2019).
386. Bennett, A. F. & Lenski, R. E. Evolutionary Adaptation to Temperature II. Thermal Niches of Experimental Lines of *Escherichia coli*. *Evolution* **47**, 1 (1993).
387. Sauer, D. B., Karpowich, N. K., Song, J. M. & Wang, D.-N. Rapid Bioinformatic Identification of Thermostabilizing Mutations. *Biophys J* **109**, 1420–1428 (2015).
388. Kuang, J.-L. *et al.* Contemporary environmental variation determines microbial diversity patterns in acid mine drainage. *ISME J* **7**, 1038–1050 (2013).
389. Kits, K. D. *et al.* Kinetic analysis of a complete nitrifier reveals an oligotrophic lifestyle. *Nature* **549**, 269–272 (2017).
390. Bello, M. D., Lee, H., Goyal, A. & Gore, J. Resource–diversity relationships in bacterial communities reflect the network structure of microbial metabolism. *Nat Ecol Evol* **5**, 1424–1434 (2021).
391. Hutchinson, G. E. Concluding Remarks. *Cold Spring Harbor Symposium on Quantitative Biology* 415–427 (1957).
392. Bauer, M. A., Kainz, K., Carmona-Gutierrez, D. & Madeo, F. Microbial wars: Competition in ecological niches and within the microbiome. *Microb Cell* **5**, 215–219 (2018).
393. Oña, L. *et al.* Obligate cross-feeding expands the metabolic niche of bacteria. *Nat Ecol Evol* **5**, 1224–1232 (2021).
394. Thomas, T. *et al.* Diversity, structure and convergent evolution of the global sponge microbiome. *Nat Commun* **7**, 11870 (2016).
395. Malard, L. A., Anwar, M. Z., Jacobsen, C. S. & Pearce, D. A. Biogeographical patterns in soil bacterial communities across the Arctic region. *FEMS Microbiol Ecol* **95**, (2019).
396. Cobo-Simón, M. & Tamames, J. Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. *BMC Genomics* **18**, 499 (2017).
397. Garcia-Garcera, M. & Rocha, E. P. C. Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. *Nat Commun* **11**, 758 (2020).
398. Sriswasdi, S., Yang, C. & Iwasaki, W. Generalist species drive microbial dispersion and evolution. *Nat Commun* **8**, 1162 (2017).
399. Pandit, S. N., Kolasa, J. & Cottenie, K. Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology* **90**, 2253–2262 (2009).

400. Luo, Z. *et al.* Biogeographic Patterns and Assembly Mechanisms of Bacterial Communities Differ Between Habitat Generalists and Specialists Across Elevational Gradients. *Front Microbiol* **10**, 169 (2019).
401. Liao, J. *et al.* The importance of neutral and niche processes for bacterial community assembly differs between habitat generalists and specialists. *FEMS Microbiol Ecol* **92**, fiw174 (2016).
402. Fridley, J. D., Vandermast, D. B., Kuppinger, D. M., Michael & Peet, R. K. Co-occurrence based assessment of habitat generalists and specialists: a new approach for the measurement of niche width. *J Ecol* **95**, 707–722 (2007).
403. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104**, 11436–11440 (2007).
404. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **6**, 776–788 (2008).
405. Auguet, J.-C., Barberan, A. & Casamayor, E. O. Global ecological patterns in uncultured Archaea. *ISME J* **4**, 182–190 (2010).
406. Schwob, G., Roy, M., Pozzi, A. C., Herrera-Belaroussi, A. & Fernandez, M. P. In Planta Sporulation of *Frankia* spp. as a Determinant of Alder-Symbiont Interactions. *Appl Environ Microb* **84**, e01737-18 (2018).
407. Zhao, K. *et al.* Actinobacteria associated with Chinaberry tree are diverse and show antimicrobial activity. *Sci Rep* **8**, 11103 (2018).
408. Chabbert, B. *et al.* Multimodal assessment of flax dew retting and its functional impact on fibers and natural fiber composites. *Ind Crop Prod* **148**, 112255 (2020).
409. Kaimenyi, D. K., Villiers, E. P. D., Ngoi, J., Ndiso, J. B. & Villiers, S. M. D. Microbiome of two predominant seagrass species of the Kenyan coast, *Enhalus acoroides* and *Thalassodendron ciliatum*. Preprint at *PeerJ Preprints* <https://doi.org/10.7287/peerj.preprints.27387> (2018).
410. Marzinelli, E. M. *et al.* Continental-scale variation in seaweed host-associated bacterial communities is a function of host condition, not geography: Host condition explains bacterial communities. *Environ Microbiol* **17**, 4078–4088 (2015).
411. Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: A phylogenetic perspective. *Science* **350**, aac9323 (2015).
412. Liang, C., Amelung, W., Lehmann, J. & Kästner, M. Quantitative assessment of microbial necromass contribution to soil organic matter. *Glob Change Biol* **25**, 3578–3590 (2019).
413. Faust, K. *et al.* Microbial Co-occurrence Relationships in the Human Microbiome. *PLOS Comput Biol* **8**, e1002606 (2012).
414. Morris, R. M. *et al.* SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002).
415. Chisholm, S. W. *et al.* A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334**, 340–343 (1988).
416. Luo, H., Huang, Y., Stepanauskas, R. & Tang, J. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. *Nature Microbiology* **2**, 17091 (2017).
417. Logares, R., Bråte, J., Heinrich, F., Shalchian-Tabrizi, K. & Bertilsson, S. Infrequent Transitions between Saline and Fresh Waters in One of the Most Abundant Microbial Lineages (SAR11). *Mol Biol Evol* **27**, 347–357 (2010).
418. Newton, R. J. *et al.* Genome characteristics of a generalist marine bacterial lineage. *ISME J* **4**, 784–798 (2010).
419. Christie-Oleza, J. A., Fernandez, B., Nogales, B., Bosch, R. & Armengaud, J. Proteomic insights into the lifestyle of an environmentally relevant marine bacterium. *ISME J* **6**, 124–135 (2012).
420. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12**, 87 (2014).
421. Mariadassou, M., Pichon, S. & Ebert, D. Microbial ecosystems are dominated by specialist taxa. *Ecol Lett* **18**, 974–982 (2015).

Appendices

422. Freilich, S. *et al.* Metabolic-network-driven analysis of bacterial ecological strategies. *Genome Biol* **10**, R61 (2009).
423. Chen, Y.-J. *et al.* Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *ISME J* **15**, 2986–3004 (2021).
424. Brewer, T. E., Handley, K. M., Carini, P., Gilbert, J. A. & Fierer, N. Genome reduction in an abundant and ubiquitous soil bacterium ‘Candidatus *Udaeobacter copiosus*’. *Nat Microbiol* **2**, 16198 (2016).
425. MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton University Press, 1967).
426. Bentkowski, P., Oosterhout, C. V. & Mock, T. A Model of Genome Size Evolution for Prokaryotes in Stable and Fluctuating Environments. *Genome Biol Evol* **7**, 2344–2351 (2015).
427. Andrei, A.-Ş. *et al.* Niche-directed evolution modulates genome architecture in freshwater Planctomycetes. *ISME J* **13**, 1056–1071 (2019).
428. Giovannoni, S. J., Thrash, J. C. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J* **8**, 1553–1565 (2014).
429. Giovannoni, S. J. *et al.* The small genome of an abundant coastal ocean methylotroph. *Environ Microbiol* **10**, 1771–1782 (2008).
430. Swan, B. K. *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* **110**, 11463–11468 (2013).
431. Maistrenko, O. M. *et al.* Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J* **14**, 1247–1259 (2020).
432. Bobay, L.-M. & Ochman, H. Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol Biol* **18**, 153 (2018).
433. Boucher, Y. *et al.* Local Mobile Gene Pools Rapidly Cross Species Boundaries To Create Endemicity within Global *Vibrio cholerae* Populations. *Mbio* **2**, e00335-10 (2011).
434. Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C. & Martiny, J. B. H. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* **10**, 497–506 (2012).
435. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**, 1812–1819 (2017).
436. Rodríguez-Gijón, A. *et al.* A Genomic Perspective Across Earth’s Microbiomes Reveals That Genome Size in Archaea and Bacteria Is Linked to Ecosystem Type and Trophic Strategy. *Front Microbiol* **12**, 761869 (2022).
437. Granot, I. & Belmaker, J. Niche breadth and species richness: Correlation strength, scale and mechanisms. *Global Ecol Biogeogr* **29**, 159–170 (2020).
438. Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci USA* **101**, 3160–3165 (2004).
439. Guieysse, B. & Wuertz, S. Metabolically versatile large-genome prokaryotes. *Curr Opin Biotech* **23**, 467–473 (2012).
440. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550 (2005).
441. Bashiri, G. Cofactor F₄₂₀, an emerging redox power in biosynthesis of secondary metabolites. *Biochem Soc Trans* **50**, 253–267 (2022).
442. Shih, P. M. *et al.* Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci USA* **110**, 1053–1058 (2013).
443. Tuomisto, H. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography* **33**, 2–22 (2010).
444. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33**, 1635–1638 (2016).
445. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

446. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).
447. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput Sci Eng* **9**, 90–95 (2007).
448. McKinney, W. Data Structures for Statistical Computing in Python. In *Proc 9th Python in Science Conference* 56–61 (SciPy, 2010).
449. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
450. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. & Pawłowsky-Glahn, V. Logratio Analysis and Compositional Distance. *Math Geol* **32**, 271–275 (2000).
451. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: an effective distance metric for microbial community comparison. *ISME J* **5**, 169–172 (2011).
452. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat Rev Microbiol* **13**, 217–229 (2015).
453. Garcia-Pichel, F., Zehr, J. P., Bhattacharya, D. & Pakrasi, H. B. What's in a name? The case of cyanobacteria. *J Phycol* **56**, 1–5 (2020).
454. Takeuchi, M. *et al.* Methyloceanibacter caenitepidi gen. nov., sp. nov., a facultatively methylotrophic bacterium isolated from marine sediments near a hydrothermal vent. *Int J Syst Evol Micr* **64**, 462–468 (2014).
455. Kinoshita, S., Yoshioka, S. & Miyazaki, J. Physics of structural colors. *Rep Prog Phys* **71**, 076401 (2008).
456. Parker, A. R. 515 million years of structural colour. *J Opt Pure Appl Opt* **2**, R15–R28 (2000).
457. Parker, A. R. & Martini, N. Structural colour in animals—simple to complex optics. *Opt Laser Technology* **38**, 315–322 (2006).
458. Seago, A. E., Brady, P., Vigneron, J.-P. & Schultz, T. D. Gold bugs and beyond: a review of iridescence and structural colour mechanisms in beetles (Coleoptera). *J Roy Soc Interface* **6**, S165–S184 (2009).
459. Prum, R. O., Quinn, T. & Torres, R. H. Anatomically diverse butterfly scales all produce structural colours by coherent scattering. *J Exp Biol* **209**, 748–765 (2006).
460. Parker, A. R. & Hegedus, Z. Diffractive optics in spiders. *J Opt Pure Appl Opt* **5**, S111–S116 (2003).
461. Prum, R. O. & Torres, R. Structural colouration of avian skin: convergent evolution of coherently scattering dermal collagen arrays. *J Exp Biol* **206**, 2409–2429 (2003).
462. Yoshioka, S. *et al.* Mechanism of variable structural colour in the neon tetra: quantitative evaluation of the Venetian blind model. *J Roy Soc Interface* **8**, 56–66 (2011).
463. Parker, A. R., McPhedran, R. C., McKenzie, D. R., Botten, L. C. & Nicorovici, N.-A. P. Aphrodite's iridescence. *Nature* **409**, 36–37 (2001).
464. Moyroud, E. *et al.* Disorder in convergent floral nanostructures enhances signalling to bees. *Nature* **550**, 469–474 (2017).
465. Whitney, H. M. *et al.* Floral Iridescence, Produced by Diffractive Optics, Acts As a Cue for Animal Pollinators. *Science* **323**, 130–133 (2009).
466. Jacobs, M. *et al.* Photonic multilayer structure of Begonia chloroplasts enhances photosynthetic efficiency. *Nat Plants* **2**, 16162 (2016).
467. Thomas, K. R., Kolle, M., Whitney, H. M., Glover, B. J. & Steiner, U. Function of blue iridescence in tropical understorey plants. *J Roy Soc Interface* **7**, 1699–1707 (2010).
468. Hahnke, R. L. & Harder, J. Phylogenetic diversity of Flavobacteria isolated from the North Sea on solid media. *Syst Appl Microbiol* **36**, 497–504 (2013).
469. Kientz, B., Agogué, H., Lavergne, C., Marié, P. & Rosenfeld, E. Isolation and distribution of iridescent Cellulophaga and other iridescent marine bacteria from the Charante-Maritime coast, French Atlantic. *Syst Appl Microbiol* **36**, 244–251 (2013).
470. Johansen, V. E. *et al.* Genetic manipulation of structural color in bacterial colonies. *Proc Natl Acad Sci USA* **115**, 2652–2657 (2018).

Appendices

471. Kientz, B. *et al.* A unique self-organization of bacterial sub-communities creates iridescence in *Cellulophaga lytica* colony biofilms. *Sci Rep* **6**, 19906 (2016).
472. Zierdt, C. H. Autolytic nature of iridescent lysis in *Pseudomonas aeruginosa*. *Antonie Van Leeuwenhoek* **37**, 319–337 (1971).
473. Kientz, B., Vukusic, P., Luke, S. & Rosenfeld, E. Iridescence of a Marine Bacterium and Classification of Prokaryotic Structural Colors. *Appl Environ Microb* **78**, 2092–2099 (2012).
474. Rodriguez, L. D., Fernández, G. S., Garayzabal, J. F. F. & Ferri, E. R. New methodology for the isolation of *Listeria* microorganisms from heavily contaminated environments. *Appl Environ Microb* **47**, 1188–1190 (1984).
475. Hamidjaja, R., Capoulade, J., Catón, L. & Ingham, C. J. The cell organization underlying structural colour is involved in *Flavobacterium* IR1 predation. *ISME J* **14**, 2890–2900 (2020).
476. Ficarrota, V. *et al.* A genetic switch for male UV iridescence in an incipient species pair of sulphur butterflies. *Proc Natl Acad Sci USA* **119**, e2109255118 (2022).
477. Zhang, L., Mazo-Vargas, A. & Reed, R. D. Single master regulatory gene coordinates the evolution and development of butterfly color and iridescence. *Proc Natl Acad Sci USA* **114**, 10707–10712 (2017).
478. Kientz, B. *et al.* Glitter-Like Iridescence within the Bacteroidetes Especially *Cellulophaga* spp.: Optical Properties and Correlation with Gliding Motility. *PLOS One* **7**, e52900 (2012).
479. Ponder, E. Diffraction Patterns Produced by Bacteria. *J Exp Biol* **11**, 54–57 (1934).
480. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
481. Brynildsrud, O., Bohlin, J., Scheffer, L. & El-dholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* **17**, 238 (2016).
482. Kerkhof, G. T. van de *et al.* Polysaccharide metabolism regulates structural colour in bacterial colonies. *J Roy Soc Interface* **19**, 20220181 (2022).
483. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* **49**, D605–D612 (2020).
484. Morehouse, N. I., Vukusic, P. & Rutowski, R. Pterin pigment granules are responsible for both broadband light scattering and wavelength selective absorption in the wing scales of pierid butterflies. *Proc Royal Soc B Biological Sci* **274**, 359–366 (2007).
485. Rutowski, R. L., Macedonia, J. M., Morehouse, N. & Taylor-Taft, L. Pterin pigments amplify iridescent ultraviolet signal in males of the orange sulphur butterfly, *Colias eurytheme*. *Proc Royal Soc B Biological Sci* **272**, 2329–2335 (2005).
486. Feirer, N. & Fuqua, C. Pterin function in bacteria. *Pteridines* **28**, 23–36 (2017).
487. Maclean, F. I., Forrest, H. S. & Hoare, D. S. Pteridine content of some photosynthetic bacteria. *Arch Biochem Biophys* **117**, 54–58 (1966).
488. Galván, I., Camarero, P. R., Mateo, R. & Negro, J. J. Porphyrins produce uniquely ephemeral animal colouration: a possible signal of virginity. *Sci Rep* **6**, 39210 (2016).
489. Nakahigashi, K., Nishimura, K., Miyamoto, K. & Inokuchi, H. Photosensitivity of a protoporphyrin-accumulating, light-sensitive mutant (*visA*) of *Escherichia coli* K-12. *Proc Natl Acad Sci USA* **88**, 10520–10524 (1991).
490. Franco, T. M. A. & Blanchard, J. S. Bacterial Branched-Chain Amino Acid Biosynthesis: Structures, Mechanisms, and Drugability. *Biochemistry* **56**, 5849–5865 (2017).
491. Huang, J. J., Petersen, A., Whiteley, M. & Leadbetter, J. R. Identification of QuiP, the product of gene PA1032, as the second acyl-homoserine lactone acylase of *Pseudomonas aeruginosa* PAO1. *Appl Environ Microb* **72**, 1190–7 (2006).
492. D’Argenio, D. A. *et al.* Growth phenotypes of *Pseudomonas aeruginosa* lasR mutants adapted to the airways of cystic fibrosis patients. *Mol Microbiol* **64**, 512–533 (2007).
493. Ho, T. K. Random decision forests. *Proc 3rd Int Conf Document Analysis Recognit* **1**, 278–282 vol.1 (1995).

494. Rivas, R. *et al.* *Alcanivorax balearicus* sp. nov., isolated from Lake Martel. *Int J Syst Evol Microbiol* **57**, 1331–1335 (2007).
495. Banach, A., Kuźniar, A., Mencfel, R. & Wolińska, A. The Study on the Cultivable Microbiome of the Aquatic Fern *Azolla Filiculoides* L. as New Source of Beneficial Microorganisms. *Appl Sci* **9**, 2143 (2019).
496. Datta, M. S., Sliwerska, E., Gore, J., Polz, M. F. & Cordero, O. X. Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nat Commun* **7**, 11965 (2016).
497. Doijad, S. P. *et al.* *Listeria goaensis* sp. nov. *Int J Syst Evol Microbiol* **68**, 3285–3291 (2018).
498. Thorat, S. R. *et al.* Virulence Profiling of *Listeria monocytogenes* Isolated from Different Sources. *Int J Curr Microbiol Appl Sci* **8**, 2010–2017 (2019).
499. Miyata, M. *et al.* Tree of motility – A proposed history of motility systems in the tree of life. *Genes Cells* **25**, 6–21 (2020).
500. Bonis, B. M. & Gralnick, J. A. *Marinobacter subterranei*, a genetically tractable neutrophilic Fe(II)-oxidizing strain isolated from the Soudan Iron Mine. *Front Microbiol* **6**, 719 (2015).
501. Schertel, L. *et al.* Complex photonic response reveals three-dimensional self-organization of structural coloured bacterial colonies. *J Roy Soc Interface* **17**, 20200196 (2020).
502. Boeuf, D. *et al.* Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *Proc Natl Acad Sci USA* **116**, 11824–11832 (2019).
503. Wijnen, B., Leertouwer, H. L. & Stavenga, D. G. Colors and pterin pigmentation of pierid butterfly wings. *J Insect Physiol* **53**, 1206–1217 (2007).
504. Tittsler, R. P. & Sandholzer, L. A. The Use of Semi-solid Agar for the Detection of Bacterial Motility. *J Bacteriol* **31**, 575–580 (1936).
505. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr Protoc Bioinform* **70**, e102 (2020).
506. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
507. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293–W296 (2021).
508. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
509. Lennon, J. T. & Locey, K. J. More support for Earth’s massive microbiome. *Biol Direct* **15**, 5 (2020).
510. Wu, D. *et al.* Stalking the Fourth Domain in Metagenomic Data: Searching for, Discovering, and Interpreting Novel, Deep Branches in Marker Gene Phylogenetic Trees. *PLOS One* **6**, e18011 (2011).
511. McKay, L. J. *et al.* Activity-based, genome-resolved metagenomics uncovers key populations and pathways involved in subsurface conversions of coal to methane. *ISME J* **16**, 915–926 (2022).
512. Gura, C. & Rogers, S. O. Metatranscriptomic and Metagenomic Analysis of Biological Diversity in Subglacial Lake Vostok (Antarctica). *Biology* **9**, 55 (2020).
513. Minich, J. J. *et al.* Host biology, ecology and the environment influence microbial biomass and diversity in 101 marine fish species. *Nat Commun* **13**, 6978 (2022).
514. Villanueva, L., Schouten, S. & Sinninghe Damsté, J. S. Depth-related distribution of a key gene of the tetraether lipid biosynthetic pathway in marine Thaumarchaeota. *Environ Microbiol* **17**, 3527–39 (2013).
515. Zeng, Z. *et al.* Identification of a protein responsible for the synthesis of archaeal membrane-spanning GDGT lipids. *Nat Commun* **13**, 1545 (2022).
516. Sahonero-Canavesi, D. X. *et al.* Disentangling the lipid divide: Identification of key enzymes for the biosynthesis of membrane-spanning and ether lipids in Bacteria. *Sci Adv* **8**, eabq8652 (2022).
517. Zeng, Z. *et al.* GDGT cyclization proteins identify the dominant archaeal sources of tetraether lipids in the ocean. *Proc Natl Acad Sci USA* **116**, 22505–22511 (2019).

Appendices

518. Murray, A. E. *et al.* Roadmap for naming uncultivated Archaea and Bacteria. *Nat Microbiol* **5**, 987–994 (2020).
519. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101–1106 (2015).
520. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* **34**, 1256–1263 (2016).
521. Gao, Y. & Li, H. Quantifying and comparing bacterial growth dynamics in multiple metagenomic samples. *Nat Methods* **15**, 1041–1044 (2018).
522. Long, A. M., Hou, S., Ignacio-Espinoza, J. C. & Fuhrman, J. A. Benchmarking microbial growth rate predictions from metagenomes. *ISME J* **15**, 183–195 (2021).
523. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–93 (2009).
524. Beaulaurier, J. *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* **36**, 61–69 (2018).
525. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci USA* **111**, E2329–E2338 (2014).
526. Tanca, A. *et al.* Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome* **5**, 79 (2017).
527. Schirmer, M. *et al.* Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol* **3**, 337–346 (2018).
528. Mason, O. U. *et al.* Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J* **6**, 1715–1727 (2012).
529. Bergauer, K. *et al.* Organic matter processing by microbial communities throughout the Atlantic water column as revealed by metaproteomics. *Proc Natl Acad Sci USA* **115**, E400–E408 (2018).
530. Hultman, J. *et al.* Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* **521**, 208–212 (2015).
531. Shaffer, J. P. *et al.* Standardized multi-omics of Earth’s microbiomes reveals microbial and metabolite diversity. *Nat Microbiol* **7**, 2128–2150 (2022).
532. Bale, N. J. *et al.* Lipidomics of Environmental Microbial Communities. I: Visualization of Component Distributions Using Untargeted Analysis of High-Resolution Mass Spectrometry Data. *Front Microbiol* **12**, 659302 (2021).
533. Ding, S. *et al.* Lipidomics of Environmental Microbial Communities. II: Characterization Using Molecular Networking and Information Theory. *Front Microbiol* **12**, 659315 (2021).
534. Holm, H. C. *et al.* Global ocean lipidomes show a universal relationship between temperature and lipid unsaturation. *Science* **376**, 1487–1491 (2022).
535. Herold, M. *et al.* Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat Commun* **11**, 5281 (2020).
536. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLOS Comput Biol* **5**, e1000605 (2009).
537. Steinegger, M. & Salzberg, S. L. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol* **21**, 115 (2020).
538. Tederloo, L., Albertsen, M., Anslan, S. & Callahan, B. Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. *Appl Environ Microb* **87**, e00626–21 (2021).
539. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133–138 (2009).
540. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**, 265–270 (2009).
541. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**, 201–206 (2018).
542. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).

543. Somerville, V. *et al.* Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol* **19**, 143 (2019).
544. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).
545. Singleton, C. M. *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* **12**, 2009 (2021).

Acknowledgements

Scientific discovery, I learned, is a group effort. I am grateful to the people that surrounded me during my PhD, no doubt shaping my scientific mind for good. At the Theoretical Biology and Bioinformatics group, I was lucky to be close to the people that inspire me. Berend, I admire your broad knowledge and balanced opinions, and will remember your modesty. Paulien, you teach a different kind of thinking, and talking to you is always stimulating. Most of all, thank you for your kindness. And finally, Bas—you lured me into your world of vast datasets and wild biological hypotheses. I have enjoyed every bit of it. *Thankfully*, in that world complex problems are balanced by pragmatic solutions. I thank all three of you for your patience.

I thank my coauthors for collaboration, Laura, Aldert, and Colin for ‘contributing equally’, Jaap for joint supervision with Bas of a chapter, and Tina for carrying an entire chapter on her own. I am also grateful to the students I have supervised, for their curiosity and the soft skills that they taught me. Jan Kees, you will be my benchmark for a great system administrator.

Most of the work in this thesis is reanalysis of data that was generated by others for a different aim, using open-source software. I acknowledge the immense amount of work that came before on which my thesis depends. Science would be a bleak place without unconditional sharing. I am happy that open science seems to be the future.

The Theoretical Biology and Bioinformatics group is a weird home filled with smart people and creative minds, and with honest curiosity. I am grateful to my former colleagues for creating that environment, and talks and discussions in many shapes have contributed to this thesis. I am certain that all of you have a bright future ahead.

I owe each of my friends an apology, for spending too much time in my secluded world of ‘finishing my thesis’. I hope we can make up for time lost. Here I especially want to mention the friends that were there when the finish line was close yet seemed so very far away—without them this thesis would not have been completed. Stephan and Jordy, you have always been there for laughs and serious conversation, and I am certain you always will be. Tina, thank you for being both awesome and my paranymp. Christian and Roderick, thanks for keeping calling me, now is finally the time to go kitesurfing whenever we want. Molly, it has finally happened!—no more jokes about finishing my PhD. And Ruth PG, thanks for all the tea, I am still surprised by how we are similar and different.

Veer, my sister and friend, we know each other our entire lives, and I am happy that you have also been a part of my thesis. Thank you for the art, thank you for being my paranymp, but above all for knowing what to do when I do not know myself.

Cor and Ruth, our parents, thank you for always supporting me no matter what I do—sometimes with a food delivery.

This thesis is not perfect, but it is finished. In the end, I am proud of what I achieved, and I will move on to the next project. I hope to see you there.

Curriculum vitae

Frederik Alexander Bastiaan von Meijenfeldt was born together with his sister in Nieuwegein (the Netherlands) on the 30th of March, 1989. Bastiaan grew up in Gouda. During his secondary education at the Coornhert Gymnasium he regularly visited Leiden University following lectures in Astronomy (via LAPP-TOP) and subsequently attending Pre-University College. After his secondary education he moved to Utrecht University to study Biology, and Arabic Language and Culture—obtaining his Bachelor of Science in Biology (cum laude) with a thesis on the biogeography of corbiculate bees. As a Master’s student Bastiaan investigated Silurian carbon cycle dynamics with marine organic microfossils in Utrecht and visited Duke University in Durham (North Carolina, the United States of America) to study population connectivity of deep sea hydrothermal vent limpets. Bastiaan went to the University Centre in Svalbard (UNIS) and boarded the RV Helmer Hanssen to explore the ecosystems in and under the North Pole sea ice. He obtained his Master of Science in Biological Sciences (cum laude) as ‘Marine Scientist of the Netherlands’—with a thesis on metabolic models for the origin of life supervised by Paulien Hogeweg. Bastiaan stayed in the Theoretical Biology and Bioinformatics group at Utrecht University for his PhD research supervised by Bas Dutilh, Berend Snel, and Paulien Hogeweg—the results of this research are described in this thesis.

During his PhD, Bastiaan boarded the RV Pelagia to collect viral DNA and RNA from the North Atlantic Ocean. Currently, he investigates lipids in the environment and throughout the tree of life at the Royal Netherlands Institute for Sea Research (NIOZ) on the Dutch island of Texel.



List of publications

For a recent list of publications, see <https://orcid.org/0000-0002-0037-0007>.

Hauptfeld, E., Pappas, N., van Iwaarden, S., Snoek, B. L., Aldas-Vargas, A., Dutilh, B. E. & von Meijenfeldt, F. A. B. Integration of taxonomic signals from MAGs and contigs improves read annotation and taxonomic profiling of metagenomes. *Under review*. (In this thesis: **Chapter 4**)

Zomer, A.*, Ingham, C. J.*, von Meijenfeldt, F. A. B.*, Doncel, A. E., van de Kerkhof, G. T., Hamidjaja, R., Schouten, S., Schertel, L., Müller, K. H., Catón, L., Hahnke, R. L., Bolhuis, H., Vignolini, S. & Dutilh, B. E. Structural colour in the bacterial domain: the ecogenomics of an optical phenotype. *Under review*. (In this thesis: **chapter 6**)

Cabrol, L., Capo, E., van Vliet, D. M., von Meijenfeldt, F. A. B., Bertilsson, S., Villanueva, L., Sánchez-Andrea, I., Björn, E., Bravo, A. G., Heimbürger-Boavida, L.-E. Redox gradient shapes the abundance and diversity of mercury-methylating microorganisms along the water column of the Black Sea. *In press in mSystems*.

von Meijenfeldt, F. A. B., Hogeweg, P. & Dutilh, B. E. A social niche breadth score reveals niche range strategies of generalists and specialists. *Nature Ecology & Evolution* **7**, 768–781 (2023). (In this thesis: **Chapter 5 and ref. 356**)

Sahonero-Canavesi, D. X., Siliakus, M. F., Asbun, A. A., Koenen, M., von Meijenfeldt, F. A. B., Boeren, S., Bale, N. J., Engelman, J. C., Fiege, K., van Schijndel, L. S., Sinninghe Damsté, J. S. & Villanueva, L. Disentangling the lipid divide: Identification of key enzymes for the biosynthesis of membrane-spanning and ether lipids in Bacteria. *Science Advances* **8**, eabq8652 (2022). (In this thesis: **ref. 516**)

Hauptfeld, E., Pelkmans, J., Huisman, T. T., Anocic, A., Snoek, B. L., von Meijenfeldt, F. A. B., Gerritse, J., van Leeuwen, J., Leurink, G., van Lit, A., van Uffelen, R., Koster, M. C. & Dutilh, B. E. A metagenomic portrait of the microbial community responsible for two decades of bioremediation of poly-contaminated groundwater. *Water Research* **221**, 118767 (2022). (In this thesis: **ref. 16**)

Garza, D. R., von Meijenfeldt, F. A. B., van Dijk, B., Boleij, A., Huynen, M. A. & Dutilh, B. E. Nutrition or nature: using elementary flux modes to disentangle the complex forces shaping prokaryote pan-genomes. *BMC Ecology and Evolution* **22**, 101 (2022).

Villanueva, L.*, von Meijenfeldt, F. A. B.*, Westbye, A. B., Yadav, S., Hopmans, E. C., Dutilh, B. E. & Sinninghe Damsté, J. S. Bridging the membrane lipid divide: bacteria of the FCB group superphylum have the potential to synthesize archaeal ether lipids. *The ISME Journal* **15**, 168–182 (2021). (In this thesis: **Chapter 2 and ref. 381**)

Appendices

Coutinho, F. H., von Meijenfeldt, F. A. B., Walter, J. M., Haro-Moreno, J. M., Lopéz-Pérez, M., van Verk, M. C., Thompson, C. C., Cosenza, C. A. N., Appolinario, L., Paranhos, R., Cabral, A., Dutilh, B. E. & Thompson, F. L. Ecogenomics and metabolic potential of the South Atlantic Ocean microbiome. *Science of the Total Environment* **765**, 142758 (2021). (In this thesis: ref. 71)

van Vliet, D. M., von Meijenfeldt, F. A. B., Dutilh, B. E., Villanueva, L., Sinninghe Damsté, J. S., Stams, A. J. M. & Sánchez-Andrea, I. The bacterial sulfur cycle in expanding dysoxic and euxinic marine waters. *Environmental Microbiology* **23**, 2834–2857 (2021). (In this thesis: ref. 70)

de Jonge, P. A., von Meijenfeldt, F. A. B., Costa, A. R., Nobrega, F. L., Brouns, S. J. J. & Dutilh, B. E. Adsorption Sequencing as a Rapid Method to Link Environmental Bacteriophages to Hosts. *iScience* **23**, 101439 (2020).

de Jonge, P. A., von Meijenfeldt, F. A. B., van Rooijen, L. E., Brouns, S. J. J. & Dutilh, B. E. Evolution of BACON Domain Tandem Repeats in crAssphage and Novel Gut Bacteriophage Lineages. *Viruses* **11**, 1085 (2019).

von Meijenfeldt, F. A. B.*, Arkhipova, K.*, Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology* **20**, 217 (2019). (In this thesis: Chapter 3 and ref. 289)

Plouviez, S., LaBella, A. L., Weisrock, D. W., von Meijenfeldt, F. A. B., Ball, B., Neigel, J. E. & Dover, C. L. V. Amplicon sequencing of 42 nuclear loci supports directional gene flow between South Pacific populations of a hydrothermal vent limpet. *Ecology and Evolution* **9**, 6568–6580 (2019).

Brito, T. L., Campos, A. B., von Meijenfeldt, F. A. B., Daniel, J. P., Ribeiro, G. B., Silva, G. G. Z., Wilke, D. V., de Moraes, D. T., Dutilh, B. E., Meirelles, P. M. & Trindade-Silva, A. E. The gill-associated microbiome is the main source of wood plant polysaccharide hydrolases and secondary metabolite gene clusters in the mangrove shipworm *Neoteredo reynei*. *PLOS One* **13**, e0200437 (2018). (In this thesis: ref. 148)

Gigliucci, F., von Meijenfeldt, F. A. B., Knijn, A., Michelacci, V., Scavia, G., Minelli, F., Dutilh, B. E., Ahmad, H. M., Raangs, G. C., Friedrich, A. W., Rossen, J. W. A. & Morabito, S. Metagenomic Characterization of the Human Intestinal Microbiota in Fecal Samples from STEC-Infected Patients. *Frontiers in Cellular and Infection Microbiology* **8**, 25 (2018). (In this thesis: ref. 138)

* These authors contributed equally to the work.

