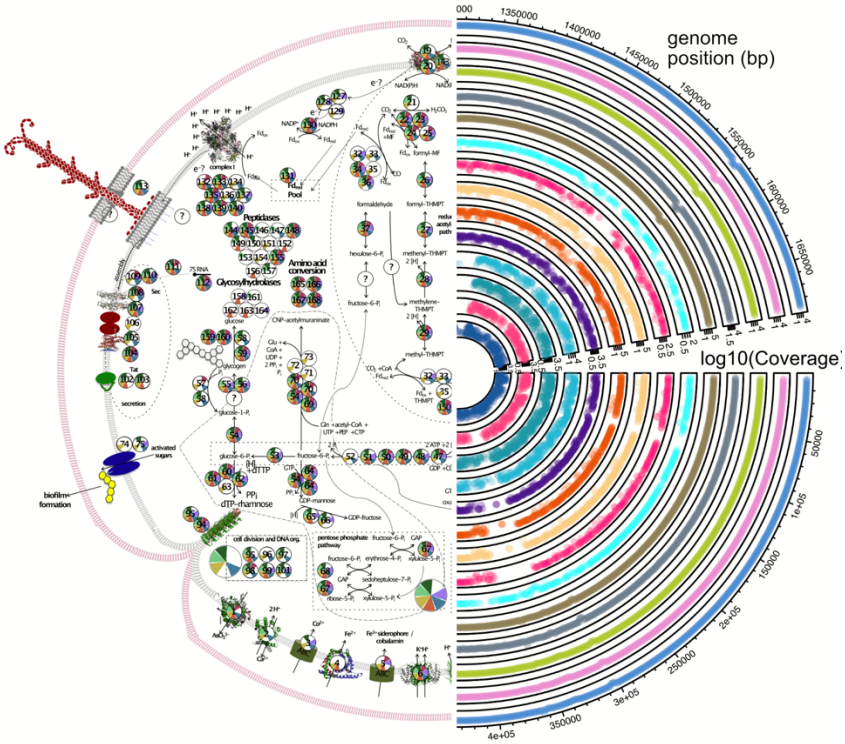

Archaeal genome fluidity in the deep biosphere

DISSERTATION

Zur Erlangung des Doktorgrades der Naturwissenschaften (Dr. rer.nat.)

Der Fakultät für Chemie der

Universität Duisburg-Essen

vorgelegt von

Till Leander Valentin Bornemann

aus Starnberg

2022

Image on title page: Genome fluidity of Ca. Altiarchaea.
Combination of Figure III.3.5 showing conserved Ca. Altiarchaea core metabolism (left) and Figure IV.2.3.1 showing Ca. Altiarchaea genome fluidity based on mapping on complete Altiarchaea genome (right). Please see respective complete figures for more details on the figures in theire respective captions as well as in the figure legends.

Das Promotionsgesuch wurde eingereicht am:	31.05.2022
Die Verteidigung wurde abgelegt am:	10.11.2022

Diese Arbeit wurde angeleitet von:		Prof. Dr. Alexander J. Probst
Prüfungsausschuss:	Vorsitzender:	Prof. Dr. Sebastian Schlücker
1.	Gutachter:	Prof. Dr. Alexander J. Probst
2.	Gutachterin:	Prof. Dr. Bettina Siebers
3.	Gutachterin:	Prof. Dr. Cara Magnabosco

Unterschrift: Till L.V. Bornemann

Archaeal genome fluidity in the deep biosphere

DISSERTATION

For archieving the doctoral degree of natural sciences (Dr. rer. Nat.)

At the faculty of Chemistry at the University of Duisburg-Essen

By

Till Leander Valentin Bornemann

From Starnberg

2022

Key words:
genome-resolved metagenomics, archaea, genome curation, software, altiarchaeota, high-CO_2, pangenomics, extreme environments, microbial ecology, subsurface, *in silico* growth estimations

The dissertation was performed under the auspices of

Prof. Dr. Alexander J. Probst (PhD supervisor)

Prof. Dr. Kai-Uwe Hinrichs (1st mentor)

Dr. Julius Lipp (2nd mentor)

At the Fakultät für Chemie under fullfillment of all guidelines according to the Promotionsordnung of the faculty

DOI: 10.17185/duepublico/77135
URN: urn:nbn:de:hbz:465-20230329-141906-1

“The cell is basically a historical document, and gaining the capacity to read it (by sequencing of genes) cannot but drastically alter the way we look at all biology.” Woese, 1987

“The powers of a man’s mind are directly proportional to the quantity of coffee he drinks” Sir James MacKintosh

Acknowledgements

I would like to take this opportunity to first and foremost thank Alexander J. Probst, who gave me the opportunity to do my PhD thesis on this ambitious and exciting interdisciplinary project, requiring me to challenge myself to combine and extend my previous research experiences (Pulse-Chase proteomics, lipidomics) along with many further new fields such as genome-resolved metagenomics to investigate the lifestyle of pretty weird putatively symbiotic little buggers (i.e, CPR). I don't think I could've asked for a more enthusiastic, supportive and relatable 'Le Boss' (©PAFG), through all the ups and downs the project has had (*e.g.*, switching ecosystems, corona,…), causing the actual presented work in this thesis to be different from the original project. I truly hope that we can bring the original project to a satisfying conclusion in the up-coming months. I'd like to thank Prof. Dr. Bettina Siebers for agreeing to be the second reviewer of my thesis, despite me being somewhat late in asking her. I'd additionally like to thank Kai-Uwe Hinrichs and Julius Lipp from the MARUM in Bremen, who are helping me explore the lipid diversity and their biosynthetic origin and are always there to talk to. I also need to thank the Ministerium für Kultur und Wissenschaft and the eSym project in particular for funding my PhD thesis.

I have a lot of amazing collaborators that I had to pleasure to interact with, sadly too many to list them all. In particular, I would like to thank Sarahi Garcia and Sari Pleura, originally from Uppsala University, for collaborating on sampling and sequencing the lake Alinen Mustayärvi in Finland, as well as Kirsten Küsel's lab from Jena (special shout-outs to Patricia, Will, Haiko, Falco, Naren and of course Kirsten Küsel), who PAFG and I sampled the Hainich CZE with.

When it comes to sampling however, I think that Perla Abigail Figueroa-Gonzalez (or PAFG) takes the cake as the person I need to thank the most as we endured through so many sampling campaigns together, she's always someone to talk to (over the project or otherwise), and just a great person. Sophie Simon deserves an additional mention since she managed to recover a complete *Ca.* Altiarchaeum genome and thus made a lot of the general discussion of this thesis possible. I also wholeheartedly need to thank the entire group, both alumni and current members, as I cannot really imagine a more fun atmosphere to work in (to quote Cedric Laczny from the LCSB in Luxembourg: "I wonder how you guys get any work done"). It is

pretty amazing how the GAME group has developed from my start with just Alex and Petra Wahl (n=3) to currently being 22 members.

I would also like to thank my family and friends for their support during my PhD thesis and helping me switch off my brain from work by *e.g.*, going to festivals ('the RAR group'), driving drunkenly at Mario Cart (Martin & Juri), playing computer games (Mark) and MTG ("biero info magic"). Last but not least, I would like to thank the proofreaders of my thesis, namely: André Rodriguez-Soarez and Sophie Simon.

List of publications

Till LV Bornemann has authored 3 first author publications and 18 publications in total, with a H index of 5 (based on google scholar profile). The individual publications are listed below, categorized into published, accepted (but not yet published), submitted/in revision and in preparation.

1. **Bornemann, T. L.V.**, Adam, P. S., and Probst, A. J. (2022) Reconstruction of archaeal genomes from short-read metagenomes. **Accepted in *Methods of Molecular Biology* 2522**, as book chapter 33.

2. **Bornemann, T.L.V.**, Esser, S.P., Stach, T.L., Burg, T., and Probst, A.J. (2022) uBin – a manual refining tool for metagenomic bins designed for educational purposes. 2020.07.15.204776. **Submitted to *Environmental Microbiology*.**

3. **Bornemann, T.L.V.,** Adam, P.S., Turzynski, V., Schreiber, U., Figueroa-Gonzalez, P.A., Rahlff, J., et al. (2022) Genetic diversity in terrestrial subsurface ecosystems impacted by geological degassing. ***Nature Communications* 13**: 284.

4. Vinklárek, I.S., **Bornemann, T.L.V.,** Lokstein, H., Hofmann, E., Alster, J., and Pšenčík, J. (2018) Temperature Dependence of Chlorophyll Triplet Quenching in Two Photosynthetic Light-Harvesting Complexes from Higher Plants and Dinoflagellates. ***Journal of Physical Chemistry B* 122**: 8834–8845. **From Master thesis.**

5. Schwank, K., **Bornemann, T.L.V.,** Dombrowski, N., Spang, A., Banfield, J.F., and Probst, A.J. (2019) An archaeal symbiont-host association from the deep terrestrial subsurface. ***ISME Journal* 13**: 2135–2139.

6. Figueroa-Gonzalez, P.A., **Bornemann, T.L.V.,** Adam, P.S., Plewka, J., Révész, F., von Hagen, C.A., et al. (2020) Saccharibacteria as Organic Carbon Sinks in Hydrocarbon-Fueled Communities. ***Frontiers in Microbiology* 11**: 587782.

7. Nuy, J., **Bornemann, T.,** Beisser, D., Probst, A., and Boenigk, J. (2020) Comparing Community Measures in Lake Microbial Ecology: Metagenomes and Metatranscriptomes and Amplicons, oh my!. **In Revision.** Preprint at https://europepmc.org/article/ppr/ppr167968**.**

8. Rahlff, J., **Bornemann, T.L.V.,** Lopatina, A., Severinov, K., and Probst, A.J. (2022) Host-Associated Phages Disperse across the Extraterrestrial Analogue Antarctica. ***Applied and Environmental Microbiology*** **88**: 10.

9. Pinto, O.H.B., **Bornemann, T.L.V.**, Oliveira, R.S., Frederico, T.D., Quirino, B.F., Probst, A.J., et al. (2022) Plume Layer Influences the Amazon Reef Sponge Microbiome Primary Producers. ***Frontiers in Marine Science*** **9**: 867234.

10. P. S. Adam, **T. L.V. Bornemann**, and A. J. Probst (2022) Progress and challenges in studying the ecophysiology of Archaea. **Accepted in *Methods of Molecular Biology* 2522, as book chapter 32.**

11. Rahlff, J., Turzynski, V., Esser, S.P., Monsees, I., **Bornemann, T.L.V.**, Figueroa-Gonzalez, P.A., et al. (2021) Lytic archaeal viruses infect abundant primary producers in Earth's crust. ***Nature Communications*** **12**: 4642.

12. Schulze-Makuch, D., Lipus, D., Arens, F.L., Baqué, M., **Bornemann, T.L.V.,** de Vera, J.-P., et al. (2021) Microbial Hotspots in Lithic Microhabitats Inferred from DNA Fractionation and Metagenomics in the Atacama Desert. ***Microorganisms*** **9**: 1038.

13. Yang, S., Liebner, S., Walz, J., Knoblauch, C., **Bornemann, T.L.V.,** Probst, A.J., et al. (2021) Effects of a long-term anoxic warming scenario on microbial community structure and functional potential of permafrost-affected soil. ***Permafrost and Periglacial Processes*** **32**: 641–656.

14. Chaudhari, N.M., Overholt, W.A., Figueroa-Gonzalez, P.A., Taubert, M., **Bornemann, T.L.V.,** Probst, A.J., et al. (2021) The economical lifestyle of CPR bacteria in groundwater allows little preference for environmental drivers. ***Environmental Microbiome*** **16**: 24.

15. Hwang, Y., Schulze-Makuch, D., Arens, F.L., Saenz, J.S., Adam, P.S., Sager, C., **et al.** (2021) Leave no stone unturned: individually adapted xerotolerant Thaumarchaeota sheltered below the boulders of the Atacama Desert hyperarid core. ***Microbiome*** **9**: 234.

16. Overholt, W.A., Trumbore, S., Xu, X., **Bornemann, T.L.V.,** Probst, A.J., Krüger, M., et al. (2021) Rates of primary production in groundwater rival those in oligotrophic marine systems. **In press at *Nature Geoscience***. Preprint at https://www.biorxiv.org/content/10.1101/2021.10.13.464073v1.

17. Adam, P.S., Kolyfetis, G.E., **Bornemann, T.L.V.**, Vorgias, C.E., and Probst, A.J. (2021) Genomic remnants of ancestral hydrogen and methane metabolism in Archaea drive anaerobic carbon cycling. 2021.08.02.454722. **Revised for Science Advances.** Preprint at https://www.biorxiv.org/content/10.1101/2021.08.02.454722v1.

18. Sarah P. Esser, Janina Rahlff, Weishu Zhao, Michael Predl, Panagiotis S. Adam, Katrin Schwank, **et al.** CRISPR-mediated resistance to archaeal episymbionts. ***In revision for Nature.***

CODL metagenomics team: Exploring functional diversity across continental and marine subsurface environments via metagenomics. ***In preparation***.

Manan Shah, Daniela Beisser, **Till L.V. Bornemann**, Julia Nuy, Martin Hahn, Alexander J. Probst, and Jens Boenigk Metagenome-assembled genomes reveal the European freshwater microbiome and provide indications for nutrient shortage or grazing pressure as driving forces for genome reduction. ***In preparation***.

Till L.V. Bornemann, Raphael Gasper, Kerstin Pieper, Claudia Buechel, and Eckhard Hofmann Crystal structure of 9-Cis Fucoxanthin. ***In preparation, for Acta Cryst E***. **From Master thesis.**

Conference Proceedings

Contributions at conferences included three poster presentations (1x International Society for Microbial Ecology in 2018; 1x Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM); 1x Transporter Colloquium 2016) and two oral presentations (1x VAAM; 1x Neujahrskolloqium).

Table of Contents

Table of Contents

I. Abstract

Archaea have long been thought to reside solely in extreme environments like hot springs, volcanos, or salt lakes and this is reflected in the majority of their available isolates. However, culture-independent methods such as metagenomics have shown Archaea to be ubiquitous in the environment, though they typically are the minority compared to Bacteria with the exception of extreme environments. The terrestrial subsurface is one of the most important environments these techniques have made accessible, as it is poorly explored and yet hosts approximately 25 % of organisms on Earth. In this environment, characterized by both extremely low energy yields and limited dispersal, the extent of horizontal gene transfer influencing evolutionary adaptation as well as the growth parameters facilitating evolution are virtually unknown.

In this thesis, we aimed to recover high quality archaeal genomes of uncultivated. Altiarchaeaota to investigate how they adapt to their deep terrestrial subsurface habitats. These Archaea dominate their moderate temperature environments, and thus identifying their adaptations, allowing them to gain an edge, is of particular interest. To accomplish this, we developed a workflow to recover archaeal genomes from metagenomes. One step frequently neglected in the recovery of genomes from metagenomes is the genome curation, due to it both being a manual task and there being a limited amount of available software. Hence, we developed the genome curation software uBin to fill this gap and enable easy, GUI-based curation of genomes. Bin curation using uBin improved the quality of 78.9 % of genomes of the CAMI dataset. Finally, we metagenomically characterized the CO_2-geyser with the highest water fountain in the world, the Geyser Andernach, which was dominated by *Ca.* Altiarchaeum GA. We binned and curated hundreds of MAGs from this and other deep terrestrial subsurface sites. To estimate the growth potential of microbes in the deep terrestrial subsurface, we compared the Geyser Andernach ecosystem to these 16 other sites. Their sampling depth ranged from near-surface caves to samples up to 3 km in depth. We identified a trend of organisms being able to replicate faster the deeper their habitant but having less replication forks at the time of sampling, a possible adaptation to oscillating nutrient availabilities. Additionally, we compared newly binned with available *Ca.* Altiarchaea genomes and identified an extreme conservation of genetic content between *Ca.* Altiarchaea of the clade Alti-1. These genomes clustered biogeographically by continent, indicating plate tectonics as

a possible route for dispersal. Their genetic repertoire showed a strong conservation of the core metabolism but differed in their peripheral genes, such as peptidases, with some showing signs of being horizontally transferred from the bacterial domain. I further substantiated these findings by using the complete *Ca.* Altiarchaeum GA genome recovered from the environment as a reference to identify sequence sections of genetic variability between populations of *Ca.* Altiarchaea. This analysis was congruent with prior biogeographic results and indicated that there is a lot more genome diversity in *Ca.* Altiarchaea than previously estimated. Some of these regions of genetic variability are likely caused by horizontal gene transfer, as evidenced by the presence of transposase genes. Thus, we conclude, that horizontal gene transfer may act to mitigate the otherwise very slow evolution within this phylum.

In summary, this thesis provides a valuable workflow for the recovery of archaeal genomes from metagenomes, along with the new, easy to use genome curation tool uBin to ensure their quality, and gives valuable insights into the genetic diversity of one of the few dominant archaea in moderate temperature deep biosphere environments.

II. General Introduction

1. Metagenomics and the controversy of naming uncultivated microbes

1.1 The established order: Naming of only cultivated microbes

Leeuwenhoek's discovery of 'little animalcules' (Lane, 2015), using a single-lens microscope, moving swiftly around in water in 1674 dawned the research field of microbiology, though most of his discoveries forgotten till their rediscovery in the 19th century (Lane, 2015). In 1838 and 1839, respectively, the cell theory was formulated, *i.e.*, that plants and animals are made up of small compartments (Mazzarello, 1999). This was followed by many further discoveries like Louis Pasteur's proof that microorganisms caused perpetual fever and Osteomyelitis (Louis Pasteur, 2017) that microbes replicated (Ligon, 2002).

In the early 20th century, life was still classified in two domains of life: Prokaryotes and Eukaryotes, based on their very distinct morphological properties (Woese and Fox, 1977). Microorganisms could only be differentiated based on their morphology and type culture characteristics (Sommerlund, 2006). No official guidelines on how to name microbes existed, though microbiologists tried to use the Botanical Code of Nomenclature as a proxy. This proved problematic as no type cultures were allowed within the Botanical code. Thus, the International Code of Nomenclature of Prokaryotes (ICNP) was established in 1936, stating some guiding and still valid principles (https://www.the-icsp.org/bacterial-code):

1. Names should be stable and the first name published is retained
2. Names should be unambiguous, assured via type cultures as references.
3. Names should only be given where needed, *i.e.*, to classify newly described organisms.

In 1977, Carl Woese and George Fox challenged this paradigm by proposing a third domain they called "archaebacteria" (later on simply called "Archaea"), which they had uncovered by investigating phylogenetic relationships of 16S ribosomal RNA (rRNA) and 18S rRNA sequences (Woese and Fox, 1977). rRNA genes are universally present, frequently in multiple copies per genome (Větrovský and Baldrian, 2013), across the tree of life, with the 16S rRNA gene being present in all Bacteria and Archaea as well as in plastids of Eukaryotes (Woese, 1987). Multiple rRNA gene copies were recently shown to increase genome stability in Bacteria (Fleurier *et al.*, 2022). It was proposed as an ideal phylogenetic marker due to its

sequence containing both highly conserved (Fox and Woese, 1975) as well as hypervariable (Woese and Fox, 1977) regions allowing examination of a wide range of phylogenetic difference ranges.

The adaptation of the 16S rRNA gene as a marker gene to characterize phylogenetic relationships started a movement away from purely morphological taxonomic categorization towards genotype-based taxonomic determination (Sommerlund, 2006). This was facilitated through newly developed techniques for the sequencing of DNA like the sequencing with chain-terminating inhibitors (Sanger, Nicklen and Coulson, 1977). In this technique, chain inhibiting nucleotide analogues were successively added to Polymerase Chain Reaction (PCR) reactions, and the resulting fragment patterns were then analyzed using gelelectrophoresis, with the band patterns revealing the original amplified sequence (Sanger, Nicklen and Coulson, 1977). To this day, this so-called Sanger sequencing is used to determine the sequence of PCR products, though only the sequence of one DNA molecule can be determined in each sequencing run. Eventually, high-throughput approaches that allowed parallel sequencing of many different DNA molecules, *e.g.*, derived from entire microbial communities, were developed, like pyrosequencing (Ronaghi *et al.*, 1996) in 1996 and to Solexa (which was later acquired and rebranded as Illumina) sequencing in 1997. These cultivation-independent new technologies soon revealed a much larger microbial diversity than previously estimated based on cultivation-dependent methods, with estimates of microbial species ranging widely from 2.2-4.3 million (Rosselló-Móra and Whitman, 2019) up to one trillion (10^{12}) species (Locey and Lennon, 2016), with actually cultured and validly described species according to the ICNP lagging extremely far behind with only 22187 prokaryotes (April 2022, https://lpsn.dsmz.de/text/numbers).

1.2 Metagenomics

In 1996, the first shotgun metagenomes of environmental samples, *i.e.*, the sequencing of random DNA fragments extracted from environmental samples without prior amplification, were reported (Stein *et al.*, 1996) and later coined ‘metagenomics’ (Handelsman *et al.*, 1998). While early sequencing techniques only sequenced a single side of these DNA fragments (single read datasets), newer techniques sequence between 150-250 bp of both edges of these DNA fragments (paired-end datasets).

These DNA fragments could then later be combined into their sequences of origin. To estimate the similarity between reads, they were decomposed into k-mers of various lengths. Two assembly methods exist: Overlap-Layout-Consensus (OLS) assemblies, which identify overlaps between reads and condense the information in a graph before making contiguous sequences (contigs) out of them, and De-Brujin-Graph (DBG) assemblers, which deconstruct reads into k-mers of varying sizes and use those to reconstruct graphs and finally assemble contiguous sequences from the graph. DBG can also use paired-end information to merge contigs, that are connected by sets of read pairs into scaffolds, filling the unknown sequence composition between the contigs with undefined nucleotides (typically designated with Ns; Figure II.1.3.1A; Please see (Li *et al.*, 2012) for an in-depth comparison of the assembly methods).

Compared to amplicon sequencing of environmental samples, metagenomics required more input DNA for library construction and was more expensive as the required sequencing depth, *i.e.*, the total length of sequenced DNA for a metagenomic sample, was much higher due as much more genetic diversity was covered (the entire genomes of organisms instead of just a specific hypervariable region of the 16S rRNA gene for these organisms). This also caused amplicon datasets to cover more of the biodiversity of the ecosystems at a comparable sequencing depth. But metagenomics also had some distinct advantages over amplicon sequencing: 1) There were no amplification biases (exceptions are mini-metagenomes or Single-amplified genomes, see section III.1.4.8), 2) they contained information about the metabolic potential of the community, 3) Organisms that were underrepresented or undetected in amplicon studies due to unusual rRNA sequences escaping amplification could be detected, 4) the metagenomes were a resource for biomarker discovery (Segata *et al.*, 2011), and could help in identifying novel metabolic pathways (Figueroa *et al.*, 2018).

One further advantage became apparent in 2004, when the first genomes were reconstructed from metagenomes of acid main drainage (Tyson *et al.*, 2004) and the Saragosso Sea (Venter *et al.*, 2004). This was the first time that genomes were recovered from the environment, without a need for prior isolation followed by cultivation. This revolutionary approach (which will be described in detail in the following section II.1.3), granted access to the vast biological diversity already indicated by prior amplicon studies (see section II.1.1) and allowed substantial expansions to the tree of life, using more robust phylogenies utilizing many marker genes (Hug *et al.*, 2016).

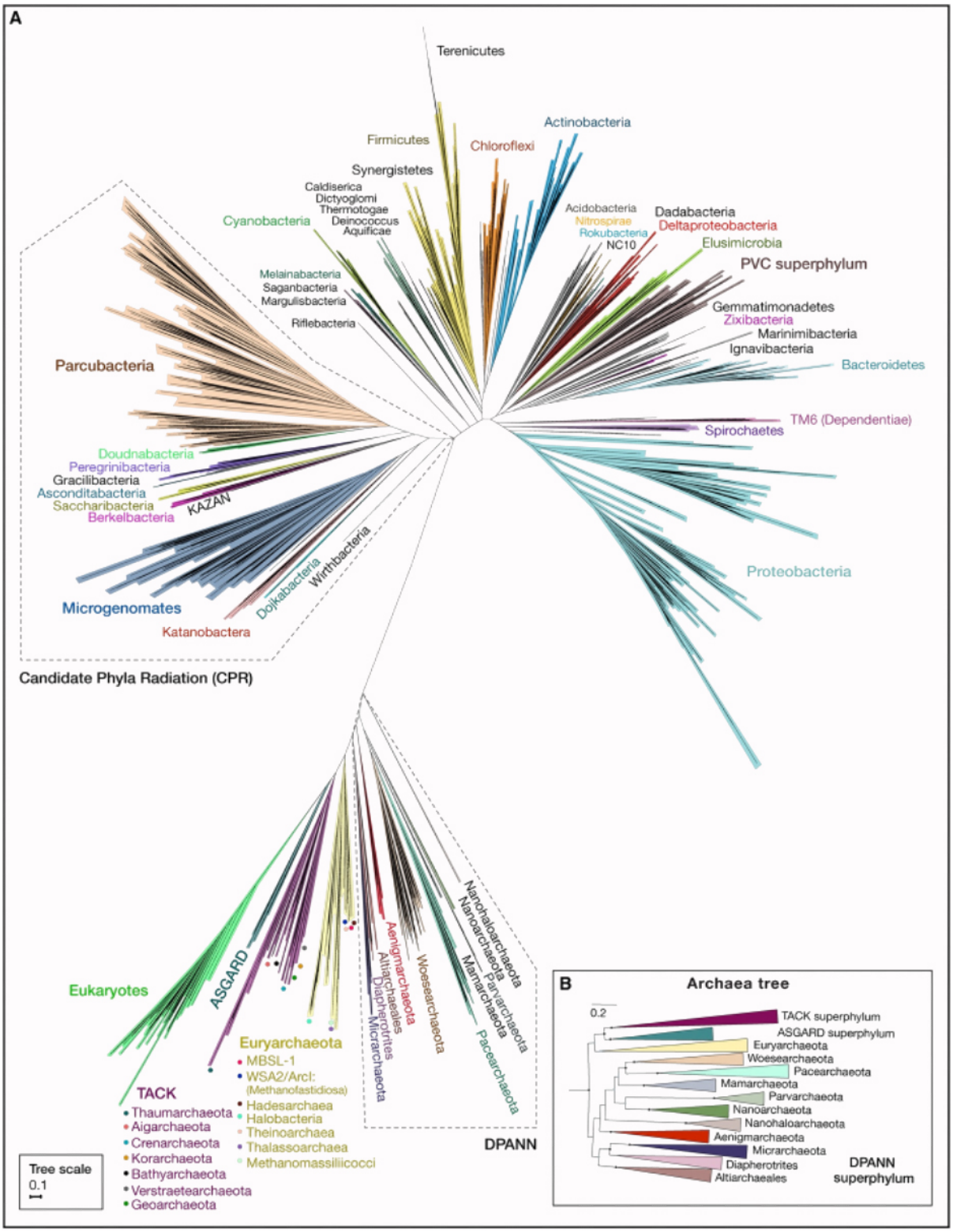

Figure II.1.2.1: The current 2-domain view of the tree of life with archaeal and bacterial domains. The displayed phylogenetic tree from (Castelle and Banfield, 2018), modified by removing the beige background, displays **A)** the current 2-domain view of the tree of life, with Eukarya branching off within the Archaea, forming a sister clade with the Asgardarchaea and **B)** A tree of the Archaeal domain, using bacteria to root the tree. Both displayed phylogenomic trees were generated based on a supermatrix of

14 conserved ribosomal proteins processed according to (Hug *et al.*, 2016) and calculated using RAxML with the PROTCATLG option (Stamatakis, 2014).

Most of the recent phylogenomic reconstructions (Spang *et al.*, 2015; Williams *et al.*, 2020) and references therein, Figure II.1.2.1) indicate that life can be divided into two domains, the Bacteria and the Archaea, with Eukaryotes having emerged from within the Archaea, with the Heimdallarchaeota being the best currently available candidate for a sister clade (Williams *et al.*, 2020).

1.3 Reconstruction of genomes from metagenomes

The initial genomes from metagenomes were reconstructed based on a combination of Guanin (G) and Cytosin (C) content, *i.e.*, the percentage of G and C nucleotides of the total sequence length per contig, and contig abundance (Tyson *et al.*, 2004), calculated by aligning ('mapping') the reads of the sample to the assembled contigs. Follow-up studies investigated the use of higher order k-mers as oligonucleotides had previously been shown to better capture species specifity than GC content (Sandberg *et al.*, 2003) and could show that 4-mers were the best compromise between specifity and runtime (Dick *et al.*, 2009). The use of 4-mers has since then become the community standard for the binning of metagenomes (Dick *et al.*, 2009; Brown *et al.*, 2015; Wu, Simmons and Singer, 2016; Graham, Heidelberg and Tully, 2017). This approach alone was still insufficient for delineating closely related organisms like species and strains, due to shared genetic patterns, and consequently frequently clustered them together (Anantharaman *et al.*, 2016). In the context of genome comparisons, species borders are defined as having between 85-95% Average Nucleotide Identity (ANI) in their head-to-head comparison (strains have >95% ANI) (Jain *et al.*, 2018).

Sharon *et al.* proposed another metric to also delineate closely related organisms: Differential coverage (Sharon *et al.*, 2013). In this extension to the original binning by coverage (Tyson *et al.*, 2004), the contig abundances are also calculated via mapping. But this process is repeated using reads of related samples, *e.g.*, from the same ecosystem, generating a set of abundances for each scaffold. Sequences belonging to the same organism are expected to share a similar abundance pattern over the sample series, making it possible to, *e.g.*, distinguish sequences belonging to different strains of the same species, provided they do not share similar abundances (Sharon *et al.*, 2013). A combination of differential coverage and 4-mer frequency-

based binning is the current standard approach to recover metagenome-assembled genomes (MAGs) from metagenomes [(Sieber *et al.*, 2018); Figure II.1.3.1B] sometimes supplemented with marker gene searches to identify clusters of marker genes within the 4-mer frequency & differential coverage space that can be used as seeds for the binning algorithms (Y.-W. Wu *et al.*, 2014). The differences are more pronounced in the used algorithms to define bins, *e.g.*, Maximum-likelyhood (Y.-W. Wu *et al.*, 2014; Wu, Simmons and Singer, 2016), dimensionality reduction followed by clustering, graph-based clustering (Kang *et al.*, 2019), modified k-medoid clustering (Kang *et al.*, 2015) and whether the binning is manual (Ultsch, 2005; Laczny *et al.*, 2015) or automatized (Wu, Simmons and Singer, 2016; Uritskiy, DiRuggiero and Taylor, 2018; Kang *et al.*, 2019). Some algorithms also use either additional metrics or completely unique metrics like assembly Graph information (Mallawaarachchi, Wickramarachchi and Lin, 2020), pre-binning alignment-based assignment of sequences to domains (Miller *et al.*, 2019) or codon usage (Yu *et al.*, 2018). Since no binner has been identified so far that performs consistently good across all types of samples, the state-of-the-art approach is to use various binners with a variety of different algorithms and metrics and aggregate these binning results into a dereplicated set of bins using DAS tool (Sieber *et al.*, 2018). DAS tool uses ubiquitous single copy marker genes for Archaea and Bacteria to assign scores to each candidate MAG based on their estimated completeness and contamination and selects the best representative of each MAG across the various input binning results based on this score, using N50 and then total genome length as tiebreaker metrics (Sieber *et al.*, 2018). The N50 of a genome is the minimum sequence length needed to cover 50% of a genome with scaffolds of this value's size or larger (Sieber *et al.*, 2018). These binning methods target prokaryotes specifically and need to be adjusted when used for eukaryotic genome binning (see section III.1.1 for more details).

The recent advent of long-read metagenomics [Pacific Biosciences (PacBIO) and Oxford Nanopore Technologies (ONT)], has caused a further revolution of the binning process, as the old adage of "a good assembly is the better binning" comes true in these technologies (Chen *et al.*, 2020; Moss, Maghini and Bhatt, 2020). Long-reads can cover otherwise problematic regions in genomes, like repeats or non-protein coding regions, and can thus assemble over these regions (Chen *et al.*, 2020). This makes long-read assemblies or hybrid short- and long-read assemblies frequently much better in quality, and can even result in recovering complete circularized genomes as single sequences directly from the environment (see section IV.2 for an example of this), a rare occurrence [62 complete prokaryotic genomes

in Sept. 2019 (Chen *et al.*, 2020), four of those generated using PacBIO] prior to the development of long reads, thereby simplifying the binning process massively.

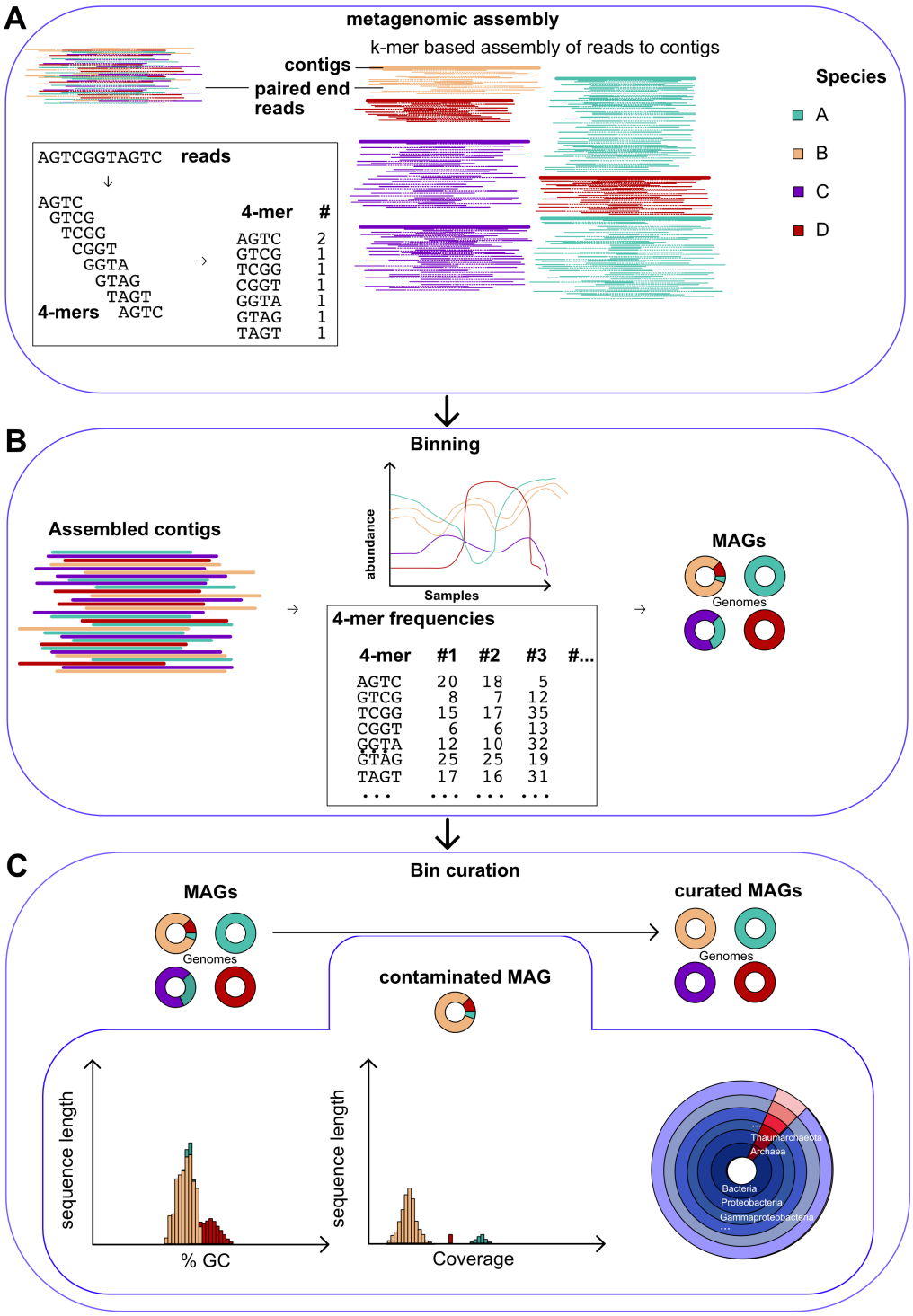

Figure II.1.3.1: Reconstruction of genomes from metagenomic reads. A: Assembly of metagenomic reads into contigs/scaffolds, with an illustration of the decomposition of a sequence (*e.g.*, reads) into k-mers. For read assembly with additional scaffolding, the default used k-mer sizes used are 21, 33 and 55 (*e.g.*, in metaSPAdes (Bankevich *et al.*, 2012)) while assembling into contigs without scaffolding employs more k-mers up to 99 (*e.g.*, in megahit (Li *et al.*, 2015)). **B:** A mixture of contigs/scaffolds is

clustered into Metagenome-Assembled Genomes (MAGs) based on shared 4-mer patterns and/or co-abundance information. The resulting bins are frequently still incomplete and have contaminating sequences. **C:** Bins can be curated of contaminating reads by removing abnormalities in GC content, coverage and taxonomy.

1.4 Knowledge gap: Why is genome curation necessary?

The complexity of most environmental metagenomes compared to metagenomes from pure cultures frequently causes errors, both in the assembly as well as the binning processes (Chen *et al.*, 2020). See Chen et al. for an extensive list of assembly error types as well as best practices on how to spot and potentially fix them (Chen *et al.*, 2020). The two types of errors encountered during binning are: 1) not binning sequences and 2) binning foreign ("contaminant") sequences to a bin. The former mainly has implications for inferences one can draw from MAGs: Unless a MAG is circularized and can thus be regarded as complete (Chen *et al.*, 2020), inferences about the absence of specific genes on the MAG are difficult to validate as their absence may just be a result of binning errors or they may have simply not assembled. The latter, however, has large implications that far exceed the individual analyses of the ecosystems they are binned from. Genomes are made public to the community, usually via upload to NCBI or similar repositories, and are thus incorporated into public databases. These public databases are then used by other researchers across the globe for further analyses, causing contaminations to propagate.

Contaminant sequences (as well as the amount of missing sequence information in a MAG), can be estimated using either universal or taxon-specific single copy marker genes, *i.e.*, genes that occur ubiquitously (within the taxon if taxon-specific markers are used) and are only present as a single copy, as each of these genes should be present exactly once per complete genome. Taxon-specific marker gene sets are more accurate than universal marker gene sets as more markers can be utilized since they only need to be universal within the given but also require prior correct phylogenetic placement of the genome to select the correct set (Parks *et al.*, 2015). CheckM (Parks *et al.*, 2015) is the community standard to perform taxon-specific contamination and completeness estimation. Universal marker genes on the other hand do not require prior phylogenetic placement and can thus be identified on entire assemblies, making them ideal for, *e.g.*, bin set aggregation (Sieber *et al.*, 2018). It should be noted that the used single copy marker genes are frequently not evenly distributed across the target genome and that thus comparatively small portions of the genome can be extremely enriched in marker genes. Examples of frequently used markers that share an operon are ribosomal proteins

(Cerretti *et al.*, 1983) and ribosomal rRNA sequences (Espejo and Plaza, 2018). This can cause completeness (and contamination) estimates to be biased for some genome regions (Chen *et al.*, 2020). While marker-based approaches are the standard approach to estimate contamination in genomes, a lot of approaches using other methods and even combinations thereof (Cornet *et al.*, 2018; Lupo *et al.*, 2021; Orakov *et al.*, 2021) have been proposed [we refer to Cornet and Baurain for a comparison between 17 of these approaches (Cornet and Baurain, 2022)].

Recent studies (Ballenghien, Faivre and Galtier, 2017; Shaiber and Eren, 2019) have identified contamination to be a significant issue in public databases, identifying metagenome-assembled Genomes (MAGS) as particularly problematic (Shaiber and Eren, 2019), due to their composite nature and the increasing number of submitted MAGs, with even studies supplying many thousands of MAGs at once becoming ever more frequent (Parks *et al.*, 2017; Almeida *et al.*, 2019; Pasolli *et al.*, 2019). This makes the quality control of MAGs even more important. And yet, there are no uniform guidelines about MAG quality that need to be followed for them to be published (Robert M. Bowers *et al.*, 2017), and consequently the quality criteria used in a publication are under the purview of the authors, reviewers and editors of respective journals alone, making the quality of published MAGs very heterogeneous indeed. Unified MAG quality standards have been proposed (Robert M. Bowers *et al.*, 2017) but are not yet widely applied. This may, in part, be due to analyses having widely different requirements in terms of MAG quality and thus enforce different priorities in MAG quality. Examples of this divide in required quality are analyses of both minimal growth rates (*e.g.*, growthpred (Vieira-Silva and Rocha, 2010)) and replication indices (*e.g.*, iRep (Brown *et al.*, 2016)) requiring fairly complete genomes to give accurate estimates, while lower completeness genomes could be included in general metabolic potential characterizations of community members to not remove large proportions of the communities.

1.5 How to curate genomes

Genome curation has been suggested to be a mandatory analysis step prior to the submission of the genome to public databases but is not yet widely applied (Robert M. Bowers *et al.*, 2017). But while established techniques to identify contamination exist (section II.1.4), the tools to curate genomes of these contaminations are sparse. All available tools like ggKbase (Wrighton *et al.*, 2012), Anvi'o (Eren *et al.*, 2015) and gbtools (Seah and Gruber-Vodicka, 2015) use a

combination of GC content, coverage and taxonomy to visually identify outlier scaffolds in MAGs (or define clusters of sequences into a MAG) (see Figure II.1.3.1C). Genome curation is mainly used to remove contaminations but can also be used to expand the genomes by recruiting additional scaffolds sharing GC, coverage and taxonomy patterns with the target genome sequences. While these tools use the same basic metrics to curate genomes, they differ widely in their MAG curation interfaces, export capabilities, additional features and the required preprocessing steps in order to use the software. They also differ in their availability, with *e.g.*, ggKbase being a feature of the titular database, requiring metagenomes to be deposited there before use (Wrighton *et al.*, 2012).

It must be noted that the gold standard for genome curation is still the completion of the sequence, *i.e.*, recovering the complete circular chromosome of a prokaryote. Only if this is the case a genome can be regarded as complete as well as free of contamination. However, this is frequently not possible due to the fragmentation level of the assemblies of environmental metagenomes, though long-read technologies have begun to change that (for more on this topic, I refer to section II.1.3). I refer to (Chen *et al.*, 2020) for a guide on curating genomes to completion.

1.6 Initiatives to solve the problem of the 'uncultivated majority'

Only a minority of genomes currently available in public databases are available as type strains in culture collection and are thus named according to the ICNP. Indeed, large proportions of the tree of life (Figure II.1.2.1) are entirely uncharted in terms of validly described species. For many groups of organisms with exacting growth requirements like obligate symbionts, it is highly unlikely that they will ever be available as type strains in culture collections (Whitman, 2015). To be able to describe and classify organisms (frequently via a combination of gene phylogenetic and morphological evidence (Whitman, 2015)) for which no type strains exist, the *Candidatus* Category was implemented (Murray and Schleifer, 1994; Murray and Stackebrandt, 1995; Stackebrandt *et al.*, 2002), which acts as a prefix to a proposed genus name with an optional epithet. *Candidatus* taxonomies have no priority, *i.e.*, if a type strain for this taxonomy is found, then the cultivator has the right to propose a new name according to the ICNP guidelines and the *Candidatus* name is discontinued (Whitman, 2015). *Candidatus* names are not subject to the ICNP and do not refer to a specific taxonomic rank, *i.e.*, there are *Candidatus* Phyla and Genera alike. *Candidatus* names are problematic as they violate two of

the founding principles of the ICNP: Names should be 1) stable and 2) unambiguous (https://www.the-icsp.org/bacterial-code).

Whitman proposed in 2015 that genome sequences could act as alternative replacements for type strains (and should accompany type strain submissions wherever possible) as they 1) allow accurate phylogenomic placement of the organism, 2) the entire diversity can be captured with this approach (in contrast to a strong bias towards easily cultivable organisms) and 3) The storage of the sequence information is much cheaper in maintenance than culture collection maintenance (current culture collections would need to expand their size at least 100-fold to encompass the entire existing diversity). However, only genome sequences from pure cultures (or single cells) should be valid type material, and needed to comply with quality restrictions in regards to genome size and fragmentation, making MAGs unsuitable due to their inherent ambiguity (Whitman, 2015). This proposal, as well as a follow-up proposal (Whitman, Sutcliffe and Rossello-Mora, 2019) to retroactively grant priority to *Candidatus* names should genome sequences become valid type material (and they otherwise comply with the ICNP rules), was ultimately rejected (Sutcliffe *et al.*, 2020).

An alternative to the amendment of the existing code was proposed by (Murray *et al.*, 2020): The establishment of a new nomenclature. This motion was put into action with the SeqCode (The International Code of Nomenclature of Prokaryotes Described from Sequence Data) initiative. The SeqCode aims to establish a platform for researchers to submit species for which only genomic data, including MAGs and SAGs, is available, and follows the ICNP rules regarding priority (though sequences are considered valid type material). It establishes rules for genome quality and data submission guidelines (Hedlund *et al.*, 2022). As this initiative is not a proposal for the adaptation of the ICNP, it does not directly rely on the acceptance by its committee. Instead, it relies on the community accepting it and propagating its use. As of this writing (May 2022), the SeqCode paper (Hedlund *et al.*, 2022) is expected to be published within the next few months (personal communication, Alexander J. Probst) and the SeqCode submission platform (https://seqco.de/) can be used by the public. It remains to be seen whether the SeqCode initiative will alleviate some of the issues being caused by the ‘uncultivated majority’.

2. Genome characteristics of Archaea in the subsurface

2.1 Beyond Photosynthesis: Carbon fixation in the deep biosphere

Traditionally, life was thought to reside solely on the surface of the land and the ocean, fueled by sunlight facilitating the fixation of carbon dioxide (CO_2) to organic carbon in the Calvin-Benson-Bassham (CBB) Reductive Pentose Phosphate Cycle, with organic carbon degrading back into H_2O and CO_2, and thus closing the biochemical cycle (Gold, 1992). Corliss *et al.* discovered 1979 microbial communities using sulfur oxidation instead of sunlight as the energy source for biomass generation in the deep-sea Galápagos Rift, providing a first example of sunlight-independent carbon fixation (Corliss *et al.*, 1979). This gave rise to the notion that microbial life might be far more widespread, postulated to extend all the way up to 5-10km in the subsurface (Gold, 1992). The nutrients were assumed to be supplied by the Earth itself, either via thermophilic vents in the oceans, fluid migrations from the crust or the rocks themselves, causing an imbalance that could be utilized to generate energy chemically (Gold, 1992). Many later studies corroborated this finding, with some examples being microbial communities living off hydrogen-releasing silicate (Telling *et al.*, 2015) and FeS_2 oxidation (Boyd *et al.*, 2014). This sparked the search for the mechanisms by which organisms used these energy sources to fixate CO_2, with up to now six additional carbon fixation pathways discovered (Berg, 2011; Steffens *et al.*, 2021).

Today, we know that the pathway an organism uses is mainly tied to the availability of oxygen and the alkalinity of the surroundings, with secondary criteria being metal co-factor availability and temperature (Berg, 2011). Aerobic carbon fixation pathways generally use NADPH (compared to more electro-negative molecules like ferredoxin in anaerobic pathways) as their reducing agent for CO_2 fixation and thus require more additional energy in Form of ATP to reduce carbon (Berg, 2011).

On Earth's surface, the CBB cycle is by far the most dominant form of carbon fixation, with it's key enzyme, the RubisCO, the most abundant protein in the world (Ellis, 1979). It is responsible for the fixation of CO_2 to ribulose-1,5-biphosphate, which in subsequent steps is further reduced to 3-Phosphoglycerate (Berg, 2011). On the surface, it is commonly coupled to photosynthesis, in which light energy is used to reduce electron equivalents (NADPH) and energy (ATP) is generated.

The CBB cycle is also a fairly common carbon fixation pathway in the anoerobic (or microaerophilic) and dark subsurface, then consequently not coupled to photosynthesis. But a much wider variety of carbon fixation pathways exist there, depending on the ecological niches. One of the typical ways of organisms to produce energy is by using the Tri-Carbolic Acid (TCA) Cycle in which Acetyl-CoA and oxaloacetate are first combined into citrate, which over the seven additional reaction steps is oxidized to oxaloacetate, resulting in a net gain of producing 3 NADH, 1 FADH2 and 1 ATP per Acetyl-CoA molecule. Under anaerobic conditions, and only in some bacteria (Ramos-Vera, Berg and Fuchs, 2009; Berg *et al.*, 2010), this cycle can be reserved in the reductive TCA cycle (rTCA) and thus allows these bacteria to generate Acetyl-CoA from two CO_2 molecules. To make this possible, the enzymes performing the irreversible reactions in the standard TCA cycle are exchanged in these bacteria (Berg, 2011). In some thermophilic organisms like *Aquificae*, an additional ATP-dependent reaction to speed up the 2-oxoglutatate to isocitrate conversion is used to deplete available pool of the thermo-labile succinyl-CoA is inserted as well (Yamamoto *et al.*, 2010). Recent investigations into the TCA cycle in the *Aquificae Thermosulfidibacter takaii* (Nunoura *et al.*, 2018) showed that its citrate synthase could function bidirectionally dependent on the availability of organic versus inorganic carbon, thus making the TCA cycle natively reversible. The direction of the TCA cycle was later shown to be regulated via partial pressures of CO_2, with high pressures reversing its direction towards autotrophy (Steffens *et al.*, 2021).

The 3-Hydroxyproprionate/4-Hydroxybutyrate (3-HP/4-BH) carbon fixation pathway has so far only been identified in the archaeal phyla Thaumarchaeota and Crenarchaeota (Berg, 2011). In this aerobic carbon fixation pathway, acetyl-CoA and 2 CO_2 are converted into succinyl-CoA (Berg, 2011). While the mechanically challenging reactions of the pathway are performed by homologous enzymes in the two archaeal phyla, many of the other reaction steps are performed by paralogous enzymes, indicating that the pathway has evolved separately in Thaumarchaeota and Crenarchaeota (Liu *et al.*, 2021). The thaumarchaeal version of the pathway is more energy-efficient as it reduces 2 ATP only to 2 ADP instead of 2 AMP like the crenarchaeal version of the pathway (Liu *et al.*, 2021). This lower energy requirement is considered an adaptation for the ammonia-oxidizing thaumarchaeota compared to the hydrogen-oxidizing Crenarchaeota (Könneke *et al.*, 2014). The Crenarchaeota also harbor another related carbon fixation cycle: The dicarboxylate/4-hydroxybutylate (DC/HB) cycle. This cycle is restricted to anaerobic Crenarchaeota like *Thermoproteales* and also converts Acetyl-CoA and 2 CO_2 molecules to succinyl-CoA, albeit with different dicarboxylases (Berg,

2011) and, same as other anaerobic pathways, is cheaper in terms of ATP at the expense of using stronger reduction equivalents.

The hydroxyproprionate cycle is an oxygen-tolerant alternative, and is found most commonly in mixotrophic organisms due to its inherent ability to coassimilate fermentation products and other compounds (Berg, 2011) but can also support autotrophic growth (Meer *et al.*, 2000). The pathway has so far only been identified in *Chloroflexus aurantiacus* (Berg, 2011).

The most energy-efficient carbon fixation pathway is the Wood-Ljiungdahl pathway, in which two CO_2 molecules are fixated to acetyl-CoA. This pathway is strictly anaerobic, with the key enzyme CO-dehydrogenase/acetyl-CoA synthase being one of the most oxygen-sensitive enzymes on Earth and is used by a great diversity of prokaryotes (Berg, 2011). Oxidative stress response enzymes and consortia with oxygen-consuming microorganisms have developed to make the pathway viable in microaerophilic environments (Shima *et al.*, 2001; Drake, Gössner and Daniel, 2008). The enzymes of this pathway also have more metal co-factor requirements than other carbon fixation pathways, restricting its use further (Stupperich and Krautler, 1988). The pathway is not just used for carbon fixation but can also be used to build up, *e.g.*, electrochemical gradients (Ragsdale and Pierce, 2008), or in an oxidative manner for decarboxylation (Can, Armstrong and Ragsdale, 2014).

In the deep biosphere, the main carbon fixation pathways are the Wood-Lijungdahl pathway, the reverse TCA cycle and the CBB cycle, with the Wood-Lijungdahl pathway frequently being the dominant pathway in anaerobic ecosystems (Momper *et al.*, 2017; Smith *et al.*, 2019) while less energy-efficient but more oxygen-tolerant alternatives like CBB and rTCA cyles become more dominant in (micro-)oxygenic environments. Intolerances such as oxygen intolerance of pathways as well as the frequently being embedded into sediment also have repercussions on the dispersal of microbes in the deep biosphere.

The Wood-Ljiungdahl pathway has likely been present in the Last Universal Common Ancestor (Weiss *et al.*, 2016) and has remained present in many Archaea and Bacteria since then, though aspects of it show signs of convergent evolution, such as the methyl synthesis branch (Sousa and Martin, 2014). The 3-HP/4-BH pathway has been indicated to have evolved independently in Thaumarchaeota and Crenarchaeota (Liu *et al.*, 2021). There is also the possibility that the hydroxyproprionate cycle being present in both Archaea and Bacteria could be the result of horizontal gene transfer to Archaea (Braakman and Smith, 2012). The pathway utilizes two biotin-dependent carboxylase reactions and these are very widely distributed in

Bacteria since they are required for fatty acid synthesis (Braakman and Smith, 2012). Enzyme analyses, however, indicate independent evolution as the catalysis step performed by proprionyl-CoA synthase in the *Chloroflexus aurantiacus* is achieved by three enzymes, that are only distantly related to the bacterial proprionyl-CoA synthase, in *Sulfolobus tokodaii* (Teufel *et al.*, 2009). Hence, it remains unclear whether the distribution of the hydroxyproprionate pathway is a result of HGT or convergent evolution (Braakman and Smith, 2012).

2.2 Biogeography: Dispersal of microbes in the subsurface

Many of the common microbial dispersal routes available on the surface, such as air, rivers and oceans as well as hitchhiking on macro-organisms (Custer, Bresciani and Dini-Andreote, 2022) are not available in the deep biosphere. In the deep biosphere, most microorganisms are embedded into the sediment and thus inherently immobile (Teske *et al.*, 2015). Organisms in or adjacent to aquifer systems can however disperse through them. One further dispersal route for all microorganisms are plate tectonics, *i.e.*, the shifting of continental plates, which are responsible for the major continential reconfigurations on Earth. These shifts however, are very slow processes, frequently taking many millions of years (Maruyama *et al.*, 1997).

Biogeography aims is the study of organism distributions and aims to identify evolutionary and ecological processes shaping those distributions (Hanson *et al.*, 2012). One of the common theorems in this field is the isolation by distance theory, in which a proportional relationship between genetic and geographic distance between species is postulated (Wright, 1943). The higher dispersal of surface organisms can prevent speciation (Claramunt *et al.*, 2012). Hence, most of microbial biogeographic analyses have focused on surface organisms inhabiting extreme environments, such as hot spring inhabiting cyanobacteria (Papke *et al.*, 2003) or *Sulfolobus* (Whitaker, Grogan and Taylor, 2003) or *Comamonas testosterone* (Liu *et al.*, 2015) strains inhabiting polluted environments, which are very limited in their dispersal due to their restricted habitants.

To our knowledge, similar studies have not been performed for organisms inhabiting the deep biosphere. The only study coming close investigated the genomic conservation in the bacterium *Candidatus* Desulforudis Audaxviator (CDA) strains across Africa, North America and Eurasia, also in the context of geographic distance (Becraft *et al.*, 2021). The CDA populations did not show a distinct clustering by continent or correlation by distance and had

extremely little genomic variation between them (>= 99.5 % average nucleotide identity for all compared genomes across continents).

2.3 Genome Fluidity in the deep biosphere

Subsurface microorganisms contribute substantially to the total biomass on our planet (Magnabosco *et al.*, 2018), despite only having access to extremely low energy (D'Hondt, Rutherford and Spivack, 2002) and consequently generally having minimal metabolism, more akin to stationary phase cultures (Finkel, 2006; Røy *et al.*, 2012). Indeed, the number of cells decreases exponentially with the depth in sub-seafloor sediment, reflecting the decrease in available energy (Kallmeyer *et al.*, 2012). Analyses of depth profiles of sub-seafloor sediments have indicated that deep biosphere communities might be remnants of the most persistent near-seafloor community members instead of actively growing or adapting communities in the deep biosphere (Starnawski *et al.*, 2017; Kirkpatrick, Walsh and D'Hondt, 2019). A periodic lifestyle has also been suggested, with long periods of bacteriostasis being interrupted by short periods of active metabolism due to ephemeral nutrient pulses (Mehrshad *et al.*, 2021). This largely dormant lifestyle might also be reflected in the minor genomic differences of *Candidatus* Desulforudis audaxviator (CDA) populations from Africa and Eurasia (Becraft *et al.*, 2021). The main proposed reasons for this extreme conservation in CDA were high fidelity DNA-replication as well as repair mechanisms encoded on the genomes, coupled to a doubling time of CDA was estimated to be <1-10 years and minimal evolution was suggested since the breakup of the pangean continent between 165 and 55 Ma ago (Becraft *et al.*, 2021). However, no systematic investigation of prokaryotic growth in the subsurface have been performed yet, to be able to extrapolate these findings to other organism groups.

Since the (Starnawski *et al.*, 2017; Kirkpatrick, Walsh and D'Hondt, 2019) studies relied on the 16S rRNA marker gene to estimate community structures in the sediment column instead of entire genomes, they would only register either new species additions or strongly mutated versions of 16S rRNA genes previously detected in other depths. The potential adaptation of deep biosphere organisms via horizontal gene transfer (HGT) would, however, remain undetected. HGT, *i.e.*, the transfer of genetic material to another organism outside of proliferation, is ubiquitous and predominantly occurs in prokaryotes (but also in Eukaryotes) (Keeling and Palmer, 2008).

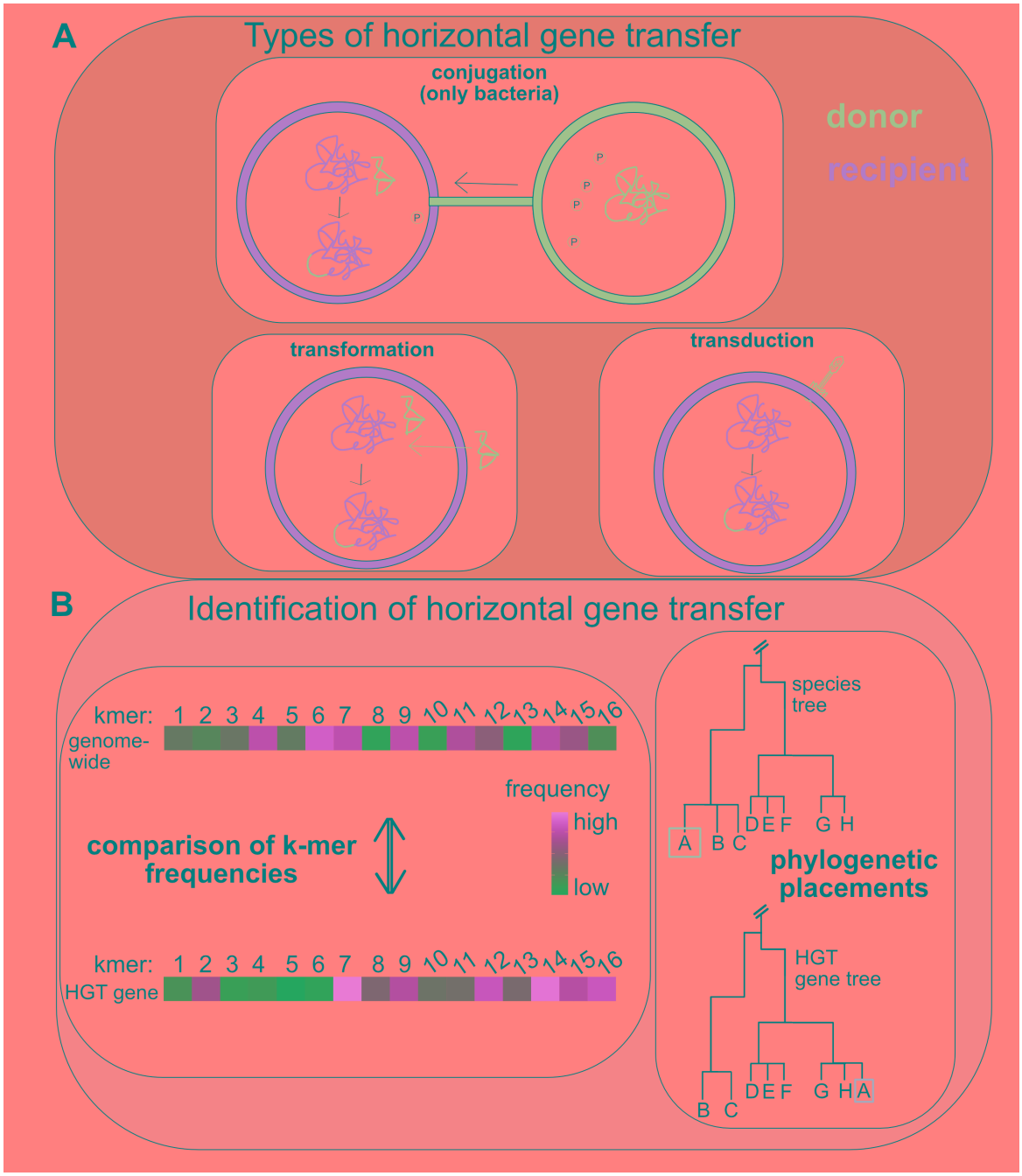

Figure II.2.3.1: Types and mechanisms of horizontal gene transfer. Only the most frequent mechanisms of horizontal gene transfer are shown (panel A), while rare variants such as Gene Transfer agents are not depicted. The phylogenetic tree and k-mer frequencies used to show techniques of identifying horizontal gene transfer (panel B) are there for illustration and do not represent real data. The heatmap was generated using ggplot2 (Wickham, 2009) and the viridis (https://cran.r-project.org/web/packages/viridis/) package supplied the color scheme.

There are a variety of HGT mechanisms, with the three main ones being conjugation (pili-mediated transfer of genetic material), transformation (uptake of external DNA) and transfection (virus-mediated uptake of genetic material) [(Soucy, Huang and Gogarten, 2015);

Figure II.2.3.1A]. While conjugation has only been described between Bacteria, transduction and transformation occur in both Bacteria and Archaea (Soucy, Huang and Gogarten, 2015). Most available bacterial genomes show evidence for HGT (Arnold, Huang and Hanage, 2022). Potential HGT candidate genes can be identified by their differences in nucleotide composition compared to the host genome (Langille, Hsiao and Brinkman, 2010). These candidates are then later validated by discrepancies between the phylogeny of the transferred gene and reference phylogenies [(Soucy, Huang and Gogarten, 2015); Figure II.2.3.1.B]. Reference phylogenies, utilizing many conserved marker genes as the basis, reflect the species ancestry. The Genome Taxonomy Database (GTDB) along with its phylogenomic placement software gtdb-tk are the currently most widely used ways to phylogenomically classify genomes and provide a reference phylogeny for bacteria with 120 marker genes and for archaea with 122 marker genes (Chaumeil *et al.*, 2020). However, such phylogenetic comparisons can only detect transfer events between distantly related organisms, making most HGT events undetectable, as successful HGT is most frequent between closely related organisms (Schaack, Gilbert and Feschotte, 2010). This higher frequency of successful HGT between closely related organisms is likely based on the likelyhood of homologuous recombination, which allows for efficient integration of closely related donor material into the recipient organism. This has been shown for Haloarchaeota, where 90% of observed HGT events were facilitated by homologuous recombination, with the frequency decreasing exponentially with the phylogenetic distance for Haloarchaeota pairs (Williams, Gogarten and Papke, 2012).

The acquisition and integration of external DNA inherently puts the host organism at a disadvantage. Thus, for horizontally transferred genetic material to persist in the recipient organism, it either needs to provide an immediate evolutionary advantage or develop one over time, being of little detriment during this development phase (Soucy, Huang and Gogarten, 2015).

(Orsi *et al.*, 2021) showed that deep sea sediment *Thalassospira* Bacteria, isolated within the sediment matrix, are more strongly affected by genetic drift, *i.e.*, fluctuations in allele frequencies with potential to become permanent fixtures in the population causing synonymous (dS; does not change protein sequence), non-synonymous (dN; changes protein sequence) and missense (introduces stop codon) mutations, than by horizontal gene transfer via homologous recombination (a special variant of transformation). This caused an accumulation of pseudogenes, though neither genome reduction nor high dN/dS ratios were

observed, possibly due to the populations not having yet experienced sufficient generations to observe large shifts due to the low energy available in the deep biosphere.

Currently, the extent to which horizontal gene transfer contributes to the adaptation of deep biosphere microbes, let alone Archaea, is completely unknown.

2.4 Altiarchaeota: The pirates of the subsurface

Archaea were originally thought to be only dominant in extreme environments, such as hot springs (Dai *et al.*, 2016), salt lakes (Gunde-Cimerman, Oren and Plemenitaš, 2005) or acidic hot springs (Ding *et al.*, 2011). The advancement of sequencing technologies revealed Archaea to be ubiquitously present, though non-extreme environments are generally dominated by Bacteria (Probst *et al.*, 2013). Notable exceptions are the Thaumarchaeota (formerly known as marine group I Archaea), who are among the most abundant prokaryotes on Earth across both marine and terrestrial environments and have the most energy-efficient aerobic carbon fixation pathway with their unique take on the 3-hydroxypoprionate/4-hydroxybutyrate pathway combined with ammonia oxidation (Könneke *et al.*, 2014). Marine group II and Marine group III Archaea (now referred to as Poseidoniales and Pontarchaea) are other phyla that can sometimes dominate ecosystems (Vetriani, Reysenbach and Doré, 1998; Iverson *et al.*, 2012; Needham and Fuhrman, 2016).

The Altiarcheaota are another archaeal phylum that can dominate in their moderate temperature environments. Altiarchaeota form almost pure-species biofilms in their cold sulfidic spring (Henneberger *et al.*, 2006) or cold geyser subsurface habitants (Probst *et al.*, 2017) and use the most energy-efficient carbon fixation pathway [the Wood-Ljiungdahl pathway; Figure II.2.4.1].

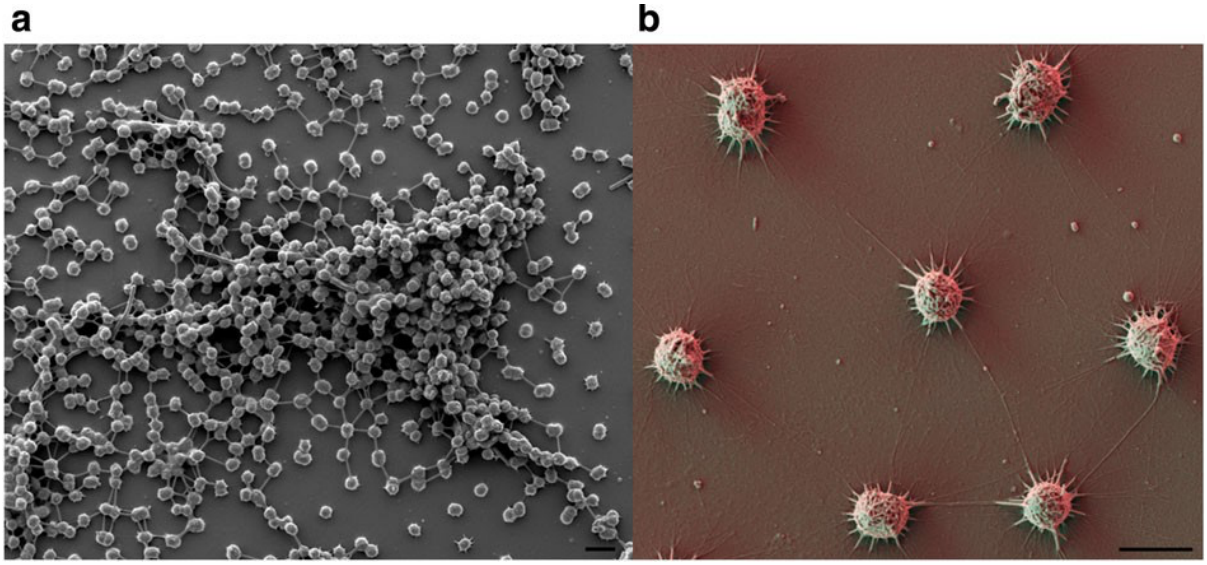

Figure II.2.4.1: Scanning electron microscopy displaying an (a) Overview and (b) Zoom of the *Ca.* Altiarchaeum SM1 biofilm from (Probst *et al.*, 2014)**.** The scale bar has 2 µm length.

Altiarchaeaota were first discovered in a cold sulfurous spring where they formed 'string of pearls'-like biofilm consortia with bacteria (Rudolph, Wanner and Huber, 2001). These pearls had a peculiar structure with the *Thiotrix* bacteria, known for being aerobic sulfide-oxidizers, forming the outer layers while the inside of the pearls was made up of Altiarcheales (Moissl, Rudolph and Huber, 2002). *Thiotrix* alone formed the strings connecting the pearls (Moissl, Rudolph and Huber, 2002). It was speculated that Thiotrix might help maintain the anaerobic environment the Altiarchaeales needed as well as supply them with sulfate, which in turn could be metabolized to sulfide by Altiarchaeales, thus closing a sulfur cycle within the pearls (Moissl, Rudolph and Huber, 2002). However, no genetic evidence supporting the use of sulfur-compounds in respiration in *Ca.* Altiarchaea was found so far (Probst *et al.*, 2014).

Altiarchaeal cells are about 0.6-0.7 µm diameter and coccoidal in shape and surrounded by around 50-100 characteristic barbwire-like cell appendages with a thrice-pronged hook at the terminal end termed *hami* with 2-3µm in length (Moissl *et al.*, 2003). These *hami* likely facilitate adhesion to both inorganic and organic surfaces (Moissl *et al.*, 2003), forming an interconnected web in the altiarchaeal biofilm and causing the characteristic 4 µm spacing between cells (Henneberger *et al.*, 2006).

In addition to inhabiting sulfidic springs, *Ca.* Altiarchaea were also found as the dominating Archaeon in the cold, high-CO_2 Crystal Geyser (Emerson *et al.*, 2016; Probst *et al.*, 2017, 2018). In this type of geyser, eruptions are caused by CO_2 over-saturation of water due

to pressure fluctuations changing the saturation limit, causing its transition into gas, causing further pressure changes in the ecosystem. This causes a chain reaction of CO_2 transitioning into gas phase and ultimately leads to the eruption of water displaced by the CO_2 gas. Gas-chromatography isotope ratio-mass spectrometry analyses of archaeal ether lipids coupled to metagenomics revealed that *Ca.* Altiarchaea fixate carbon via the strictly anaerobic reductive Acetyl-CoA (Wood-Ljiungdahl) pathway (Probst *et al.*, 2014), thus potentially benefitting from the high CO_2 concentrations in cold geyser ecosystems. Genomic analyses indicated that they use methanofuran and tetrahydromethanopterin (THMPT) as C_1 carriers. These C_1 carriers are typically used in methanogenic archaea but markers indicating methanogenesis (*e.g.*, mcr) were absent. Instead of the factor$_{420}$ -dependent THMPT dehydrogenase typically present in methanogenetic archaea, they use a NAD(P)-dependent enzyme likely acquired from methylotrophic bacteria. Indeed, no factor$_{420}$ biosynthesis genes could be identified (Probst *et al.*, 2014). Similarly, no hydrogenases supplying the WL with reduction equivalents have been identified so far, making the source of electrons for the WL pathway a mystery (Probst *et al.*, 2014).

Bird *et al.* showed in 2016, using both phylogenetic analyses based on the 16S rRNA gene and phylogenomic analyses using 10 conserved marker genes, that two clades, termed Alti-1 and Alti-2, are formed within the Altiarchaeota (Bird *et al.*, 2016). This separation was also apparent in the genomic differences between the clades: Only the Alti-1 clade forms biofilms facilitated by their *hami* cell appendages and can dominate ecosystems but seems to be restricted to sulfidic springs and cold high CO_2 geysers, while Alti-2 seem to be planktonic but are much more widespread and genetically diverse (Bird *et al.*, 2016).

3. Scope of the thesis and publication guide

The scope of this thesis was to establish a workflow for the recovery of high-quality archaeal genomes from metagenomes, including establishing a genome curation tool facilitating easy GUI-based genome curation, and using this workflow to investigate global prokaryotic replication in the deep biosphere and investigate the genetic adaptations dominating Altiarchaeota across the globe. The knowledge gaps the following publications attempt to fill are illustrated in section II.1.4 (knowledge gap: Why is genome curation necessary?), section II.1.5 (how to curate genomes), section II.2.3 (genome fluidity in the deep biosphere) and

section II.2.4 (altiarchaeota: The pirates of the subsurface). The following paragraphs will summarize the content and main findings of the three publications covered in full in section III.

3.1 Recovery of archaeal genomes from metagenomes

Metagenomics is a rapidly developing research field, with new software being released daily. Only very few steps in the metagenomics workflow have an established state-of-the art software associated with them, making the analyses of metagenomic very challenging, as researchers need to generate custom pipelines to analyze their data. The manuscript titled "**Reconstruction of archaeal genomes from short-read metagenomes**" provides our **metagenomics pipeline** for the processing of metagenomic bins, the **binning of genomes as well as their curation of contaminant sequences**, providing researchers with a ready-to-use pipeline for the recovery of genomes from metagenomes. This book chapter also contains a wealth of advice for various problems commonly encountered in the analyses of metagenomics, such as analyzing MDA-based metagenomics (single-cell metagenomes, mini-metagenomes), contamination controls and what to consider when sequencing a metagenome (*i.e.*, which technology and parameters to keep in mind).

3.2 Curation of metagenomic bins with uBin

Bin curation has been advocated to become an obligatory part in the workflow for the generation of MAGs from metagenomes but is not yet widely applied. The reason for this may be that bin curation is a manual and hence low-throughput task. Additionally, the few available software (mainly Anvi'o and ggKbase) are challenging to use for beginners, making it difficult for the practice of bin curation to disseminate through the community. Hence, we developed the **bin curation software uBin**, showcased in the manuscript titled "**uBin – a manual refining tool for genomes from metagenomes**", and tested its ability to improve bin quality on both simulated and real world data. We also explored its potential as a standalone binner for low complexity metagenomes, recovering the **first genomes from the International Space Station (ISS)** mainly consisting of human-associated microbes, and identified differences in their replication processes (iRep values), diagnosing the flight mission (1, 2 or 3) as the main factor influencing their replication.

3.3 Genomic fluidity of deep biosphere dominating Altiarchaeota

Sites of geological degassing, such as hydrothermal vents and volcanos, have long been studied in regard to how the gasses shape native communities and how organisms adapt to them. The effect of geological degassing on communities residing in the mesophilic deep biosphere, however, is rarely explored. In the manuscript titled ”**Genetic diversity in terrestrial subsurface ecosystems impacted by geological degassing**”, we explored one such system, the cold, high-CO_2 Geyser Andernach, and identified a microbial community dominated by *Ca.* Altiarchaea of clade Alti-1, one of the few known archaea able to dominate in moderate temperature environments. Biogeographic analyses of all available **Alti-1 Altiarchaea** indicated them to be **dispersed via plate tectonics**, with pangenome analyses showing a **conserved core metabolism** with **peripheral genes** showing evidence of having been **horizontally transferred from Bacteria**, possibly compensating for the otherwise extreme conservation. To evaluate the impact of geological degassing on microbial communities, we assembled a database of 895 genomes from Geyser Andernach and 16 other ecoystems, covering 3 km of depth and four continents (North America, Africa, Europe, Asia), and uncovered patterns **of faster minimal generation times but less on-going replication with depth**. Sites impacted by **geological degassing behaved similar to surface-near sites**, indicating them to be hotspots of microbial activity in the deep biosphere.

III. Publications

Overview

This quasi-cumulative dissertation comprises three articles. One manuscript (section III.3) has been published in the peer-reviewed Journal Nature Communications. Another manuscript (section III.1) has been accepted as a book chapter in the book series "Methods of Molecular Biology" and one additional manuscript (section III.2) has been submitted to the peer-reviewed journal Environmental Microbiology. The PhD student Till L. V. Bornemann has authored every manuscript printed in this thesis as first author. Supporting information as well as additional data is supplied on the supporting CD, whose contents are described in section VII.

The PhD student's contributions to the manuscripts are as follows:

1. Till L. V. Bornemann, Panagiotis S. Adam, Alexander J. Probst, 2022: Till L. V. Bornemann wrote the manuscript, incorporated revisions and verified the functionality of the given code blocks.
2. Till L.V. Bornemann, Sarah P. Esser, Tom L. Stach, Tim Burg, and Alexander J. Probst, 2022: Till L.V. Bornemann prepared the figures, prepared the supplemental material, performed statistics, wrote the code used to generate uBin input files (*i.e.*, the uBin wrapper scripts) and wrote the manuscript.
3. Till L.V. Bornemann, Panagiotis S. Adam, Victoria Turzynski *et al.*, 2022: Till L. V. Bornemann performed sampling and biochemical as well as microscopic experiments, did bioinformatic and biogeographic analyses, performed statistics, prepared figures and wrote the manuscript, submitted data to NCBI where applicable and implemented the revisions.

1. Reconstruction of archaeal genomes from short-read metagenomes

Till L. V. Bornemann, Panagiotis S. Adam, Alexander J. Probst*

Environmental Microbiology and Biotechnology, Faculty of Chemistry, University of Duisburg-Essen, Universitätsstrasse 5, 45141 Essen, Germany

*To whom the correspondence should be addressed:
Alexander J. Probst
alexander.probst@uni-due.de
+49 (201) 183-7080

Publication information:
Methods of Molecular Biology, Vol. 2522, chapter 33
Received 29.05.2021; revised 26.08.2021; second revision 05.11.2021; proofed 11.05.2021

Abstract

As the majority of biological diversity remains unexplored and uncultured, investigating it requires culture-independent approaches. Archaea in particular suffer from a multitude of issues that make their culturing problematic, from them being frequently members of the rare biosphere, to low growth rates, to them thriving under very specific and often extreme environmental and community conditions that are difficult to replicate. OMICs techniques are state of the art approaches that allow direct high throughput investigations of environmental samples at all levels from nucleic acids to proteins, lipids, and secondary metabolites. Metagenomics, as the foundation for other OMICs techniques, facilitates the identification and functional characterization of the microbial community members and can be combined with other methods to provide insights into the microbial activities, both on the RNA and protein levels. In this chapter, we provide a step-by-step workflow for the recovery of archaeal genomes from metagenomes, starting from raw short-read sequences. This workflow can be applied to recover bacterial genomes as well.

Key words

rare biosphere, genome-resolved metagenomics, short-read sequencing, genome curation, prokaryotes

1.Introduction.

Metagenomics as a technique first appeared in scientific studies of acid mine drainage biofilms and the Sargasso Sea (Venter *et al.*, 2004), analyzing the entire DNA content of samples instead of focusing on specific marker genes. Amplicon-based studies in general do not capture the entire extent of microbial diversity in environmental samples due to primer biases and introns in 16S rRNA genes (Eloe-Fadrosh, Ivanova, *et al.*, 2016). For instance, bacteria of the Candidate Phyla Radiation (CPR) often have introns and Archaea of the DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaea) radiation often escape detection using standard primer pairs (Eloe-Fadrosh, Ivanova, *et al.*, 2016; Brown *et al.*, 2015; Huber *et al.*, 2002; Baker *et al.*, 2006; Eloe-Fadrosh, Paez-Espino, *et al.*, 2016). As metagenomics is non-targeted, it enabled researchers to reveal entire clades of hitherto unknown microorganisms (Castelle *et al.*, 2013; Brown *et al.*, 2015). Shotgun sequencing of metagenomic DNA also allows the investigation of the entire biological diversity in a single experiment, including all three domains of life, DNA-based viruses as well as mobile genetic elements, provided these entities are double-stranded and DNA-based. In general, metagenomic analyses can be divided into three levels depending on the data type: Read-based metagenomics, assembly-based metagenomics, and genome-resolved metagenomics. These three levels will be briefly introduced and explained in the following paragraphs and their advantages and disadvantages are summarized in *Table* III.1.1.

Table III.1.1: Advantages (+) and disadvantages (-) of the three levels of metagenomic analyses.

Read-based metagenomics	Assembly-based metagenomics	Genome-resolved metagenomics
+ Captures the full breadth of sequenced information	+ Genes called are usually fairly complete	+ Genes called are usually fairly complete
- Short reads span only fragments of genes (problem of function prediction)	- Memory (RAM) intensive	+ linkage of function and taxonomy
- Long Nanopore reads can have frame shifts (problem of function prediction)	- Calling taxonomy only possible with marker genes	+ Proper phylogeny possible & other features like strain resolution etc.
- Time and resource intensive	- Only assembled reads are taken into account	- Only binned reads are taken into account
- Taxonomic calling is very weak	- Assembly errors	- Assembly and binning errors

Read-based metagenomics is set up with the aim of analyzing the quality-checked reads by querying them as BLAST homology searches against a single or variety of databases. Protein databases (*e.g.*, `NCBI-nr`, `UniRef100` (Suzek *et al.*, 2015)) are searched to estimate functional profiles of the metagenome and taxonomic toolkits like `SILVA` (Quast *et al.*, 2013) or `PhyloFlash` (Gruber-Vodicka, Seah and Pruesse, 2020) are used to estimate taxonomic profiles. Genome databases (*e.g.*, `RefSeq` (Pruitt *et al.*, 2011)) can also be queried to estimate genome abundances and calculate Single Nucleotide Polymorphisms (SNPs). Read-based metagenomics is less memory-intensive than other approaches, as no assembly of the reads (see below) is required but the querying against databases is computationally expensive (in CPU power/hours). Additionally, the entire metagenome is analyzed instead of only the part that was assembled or binned (compare assembly-based and genome-resolved metagenomics). On the other hand, read-based approaches mostly rely on reference databases and thus the quality of the analyses is dependent on how suitable the chosen reference databases are for the given sample. Thus, this method will perform well for sufficiently characterized ecosystems but will fail to identify novel genes or species (Teeling and Glöckner, 2012). The short read length compared to assembled sequences also makes the alignment less accurate as the

unambiguous assignment of reads to a reference is often not possible, for instance due to repetitive sequences re-occurring across proteins.

In assembly-based metagenomics, the reads are assembled into contiguous sequences, also called contigs. During sequencing, each DNA fragment is only sequenced partially, with the first 100-250 bp (forward-read) and the last 100-250 bp (reverse-read) of the fragment being sequenced, making up a read pair. If multiple of these read pairs can be aligned to the edges of two contigs, indicating their connectedness, they are combined into a single sequence named scaffold, adding in N's in the scaffold sequence between the two previous contigs to designate the unknown nucleotides linking the two contigs (scaffolding). When assembling metagenomes from environmental communities, both high and low abundance organisms can pose a challenge: The assembly of highly abundant organisms can be problematic due to the co-occurrence of strains making assembly graphs unresolvable and low abundance organisms might simply not assemble, due to a lack of available information. The assembly of highly abundant and strain diverse organisms can sometimes be improved by using small fractions of the total reads for assemblies to reduce the amount of strain interference during assembly (Hug *et al.*, 2015). Only resequencing with more sequencing depth can help resolve less abundant community members. Assemblies are more informative than pure reads, as the longer contig sequences make more accurate matching to databases possible (Breitwieser, Lu and Salzberg, 2019) and allow identification of larger genomic constructs like operons (Hu and Friedberg, 2020), viral particles (Guo *et al.*, 2021) or CRISPR arrays (Edgar, 2007). Similar to read-based metagenomics, a general overview about the taxonomic and functional composition of the community can be gleaned from assembly-based metagenomics. The assembly is the most computationally expensive (though not manual work intensive) step in the metagenomic analysis. One of the main differences to read-based approaches is the need for gene prediction. In this step, putative regions of DNA encoding for proteins are identified. This is made difficult by the presence of multiple genetic codes that alter the identification and translation of genes and can co-occur in metagenomes due to the presence of many different organisms. Thus, gene prediction needs to be performed on each scaffold individually to identify the correct genetic code and consequently the genes (Hyatt *et al.*, 2010). Additionally, eukaryotes have a very different gene structure to prokaryotes, with more complex promoter regions, regulatory signals and introns (West *et al.*, 2018). Eukaryotic gene prediction thus needs to be performed by specialized tools like `EuGene-EP` (Sallet, Gouzy and Schiex, 2019) as prediction using tools designed for prokaryotes will result in false results. To identify eukaryotic scaffolds in

metagenomes, tools like `Eukrep` (West *et al.*, 2018) can be used. See Breitwiesser et al. (2019) for a comprehensive list of read-based and assembly-based software (Breitwieser, Lu and Salzberg, 2019).

The complexity of metagenomes can be estimated from the raw sequencing data using `Nonpareil3` (Rodriguez-R *et al.*, 2018). This tool is of particular interest, as it also gives an estimate about how much of the microbial diversity is covered by the sequencing effort and predicts how much sequencing effort is needed to reach specific coverage thresholds (*e.g.*, 95%). The assembly-based analysis can be taken one step further by assigning subsets of scaffolds to genomes based on shared characteristics, *i.e.*, the "binning" of genomes in genome-resolved metagenomics. Most modern binning tools use one or both of two major characteristics: *k*-mer frequency patterns and (differential) coverage. Similar to the use of *k*-mers in metagenome assemblies, scaffolds can be decomposed into their set of *k*-mers. These *k*-mer frequency patterns can subsequently be used to cluster scaffolds belonging to the same genome based on genome-characteristic *k*-mer signatures. In read assembly, a combination of large *k*-mers is usually used with *k*'s between 21 to 128. In binning, *4*-mers (tetranucleotides) have been shown to be the most effective trade-off between signal sensitivity and runtime (Dick *et al.*, 2009). Tetranucleotide-based binning becomes problematic if multiple closely related species or strains are present in the sample as they often do not have distinguishable *4*-mer patterns. Differential coverage binning on the other hand utilizes the correlating abundance of scaffolds belonging to the same genome obtained from mapping of reads of related samples to the assembly (Albertsen *et al.*, 2013; Sharon *et al.*, 2013). This strategy makes the delineation of strains and closely related species possible, provided they have differential abundance patterns across the samples (Wu, Simmons and Singer, 2016). In some binning tools, additional characteristics like assembly graph information (`GraphBin` (Mallawaarachchi, Wickramarachchi and Lin, 2020)) or taxonomy (`Autometa` (Miller *et al.*, 2019)) are used for binning metagenome assembled genomes (MAGs). Both manual (`ESOM` (Dick *et al.*, 2009), `VizBin` (Laczny *et al.*, 2015)) as well as automatic (`MaxBin` (Wu, Simmons and Singer, 2016), `ABAWACA` (Brown *et al.*, 2015), `MetaBAT` (Kang *et al.*, 2019), `binsanity` (Graham, Heidelberg and Tully, 2017), `Automata` (Miller *et al.*, 2019), `metawrap` (Uritskiy, DiRuggiero and Taylor, 2018), `GraphBin` (Mallawaarachchi, Wickramarachchi and Lin, 2020)) binning tools exist. The success of the various binning strategies and tools is very dependent on the sample type and complexity and cannot be predicted *a priori* (Sieber *et*

al., 2018). Thus, the state of the art approach is to use a combination of binning tools followed by the aggregation of the results using `DAS Tool` (Sieber *et al.*, 2018) to extract the best representative bin set, which can also include sub-setting single bins. Even though the qualitatively best genomes are selected this way, they frequently still require further curation as both false positives (*i.e.*, contaminant scaffolds not belonging to the bin) and false negatives (*i.e.*, scaffolds belonging to the genome that were not binned) abound. The completeness of bins can be assessed using either ubiquitously present (Probst *et al.*, 2017) or branch-specific marker genes (Parks *et al.*, 2015). As these marker genes should be only present once per genome, multiple copies of marker genes can be used to approximate the degree of false positive information in the bin, *i.e.*, the degree of contamination (note that some genes can also be split, either due to an evolutionary event or due to splitting of genes in fragmented assemblies). The contamination in metagenomic bins after binning makes manual curation of the bins necessary. Different tools exist to assist this step, including `Anvi'O` (Eren *et al.*, 2021), `ggKbase` (Wrighton *et al.*, 2012), `gbtools` (Seah and Gruber-Vodicka, 2015) and `uBin` (Bornemann *et al.*, 2020). They use a combination of GC content, coverage, and taxonomy to identify contaminant sequences, but are different in the interface design, data input, scope and accessibility.

This book chapter is set out with the aim of providing the reader a guide and the requisite knowledge on reconstructing archaeal MAGs starting from raw short-read metagenomes. The method presented here can be carried out by any scientist with basic knowledge in shell programming and with access to a high-performance computing (HPC) facility. Additionally, our protocol can not only be used for reconstructing archaeal genomes from metagenomes but can also be applied for bacteria-focused research.

2.Materials

2.1 Generating sequencing reads

While this chapter focuses purely on the analysis of metagenomic data, the data generation, including sampling, DNA extraction, library preparation and sequencing, remain essential prerequisites. The sampling strategy needs to be adjusted to the ecosystem, *e.g.*, filtering water samples on filters to concentrate the cells. Particularly in ecosystems with a low expected biomass, such as subsurface communities, measuring the cell concentrations via microscopy can help determine the required amount of sample. Field blanks are also recommended for

environmental samples, particularly low biomass ecosystems (Sheik *et al.*, 2018). Optionally, DNA from intact cells can be separated from other DNA using propidium monoazide (PMA) treatment (Joo, Park and Park, 2019). DNA can be extracted with a wide variety of commercial kits designed for specific sample types (*e.g.*, biofilms, soil, water, human) or using classic chloroform-phenol-based methods. A negative control should be included to ensure detection of contaminants coming from the kit or reagents (Sheik *et al.*, 2018). If needed, the DNA can be concentrated using ethanol precipitation protocols or vaporization, depending on the volume. The genomic DNA is then fragmented and metagenomic libraries are prepared using one of many available commercial kits. The sequencing can then be performed at sequencing facilities or commercial providers. Depending on the sequencing platform, DNA fragmentation and library preparation can also be performed by the sequencing facility. For commissioning the sequencing, we recommend that 1) no amplification is performed, 2) the libraries are paired-end, *i.e.*, there is both a forward and a reverse read of one DNA strand (necessary for scaffolding in assemblies, see above) and 3) the read length is as long as possible. If 1) or 2) are not the case, alternate software needs to be utilized (see Notes sections 4.2 and 4.8).

2.2 Basic bioinformatics knowledge

Most of the analyses in this chapter are based on the UNIX command-line and consequently require basic knowledge in the UNIX shell, such as navigating and creating folders, inspecting and copying flat (*i.e.*, non-binary) files. If the user does not have such preliminary knowledge, we recommend first taking a bash tutorial course like https://linuxconfig.org/bash-scripting-tutorial-for-beginners.

2.3 Desktop computer

A desktop computer with at least 2 cores and 16 GB RAM is recommended (8 GB of RAM can work for low complexity metagenomes). Software facilitating the connection to a UNIX-based server needs to be installed. There is a wide variety of options depending on the operating system of the desktop computer. On a Windows operating system (OS), the free to use software `PuTTy` (https://www.putty.org/) can be installed and used to connect to the server. UNIX-based OS like MacOS or Linux distributions should natively be able to connect to servers using the `ssh` command. Optionally, a file sharing program like `CyberDuck` (https://cyberduck.io/) can be installed for easier transfer of files between desktop computer and Server. Furthermore, the manual binning software `esomana` (Ultsch, 2005) or `VizBin` (Laczny *et al.*, 2015) can

be installed (http://databionic-esom.sourceforge.net/index.html) and http://claczny.github.io/VizBin/). Finally, the newest release of the metagenome-assembled genome curation program `uBin` (Bornemann *et al.*, 2020) should be downloaded from https://github.com/ProbstLab/uBin/releases. Installation wrapper files are available for the various operating systems (`.exe` for Windows, `.dmg` for MacOS, `.deb` for Linux distributions).

2.4 Server requirements

The server should run a Linux distribution like `Ubuntu` (https://ubuntu.com/download). Hardware requirements depend on the size of the datasets to be analyzed. The assembly step is by far the most memory-demanding step and scales with the size and complexity of the input dataset. 500 GB of RAM are a bare minimum, but very complex and deeply sequenced metagenomes like soil ecosystems or those containing eukaryotic reads can require more than three TB of RAM. See van der Walt et al. (2017) (van der Walt *et al.*, 2017) for a benchmark about the memory requirements of various assemblers. The server should have a minimum of ten threads to be able to analyze single samples in sequence. Be aware that CPUs have a maximum amount of RAM they can use effectively, potentially making a higher number of CPUs necessary. Cloud computing platforms are an attractive alternative to purchasing and maintaining one's own server system and can be configured with the respective requirements listed above (*e.g.*, `Amazon AWS`).

2.5 Code writing conventions

This chapter contains code blocks detailing how individual analysis steps are performed. To avoid any issues with executing or understanding the code, the conventions for how different aspects of the code are formatted are described here. Any parts of the code that are enclosed in curly brackets, *e.g.*, `{int}`, require the reader to replace the brackets and its contents with the described item (in the mentioned example, an integer value). Throughout this section, the scripts that are used are assumed to be in your `PATH` variable. If this is not the case, you can either add them to your `PATH` variable using the export function or use the path to the respective script. See https://opensource.com/article/17/6/set-path-linux for an explanation of the `PATH` variable and how it can be modified, both temporarily and permanently. If the software was installed into a `conda` environment and the respective environment is active, it will automatically be in the `PATH` variable (*see Table* III.1.2).

2.6 Data and data structure

In this chapter, two formats for sequencing information will be used: the `FASTA` and the `FASTQ` format. In the `FASTA` format, each sequence entry starts with a > followed by a title or header for the sequence (*see* Subheading 2.6.1). The next line(s) then contain the sequence data and are used either for nucleotides (AGCT for individual nucleotides or N for undefined nucleotide) or amino acids (20 letters for the standard amino acids, additional letters for non-standard amino acids *e.g.*, Y for pyrrolysine, usually X or another symbol such as a question mark for undefined states).

2.6.1 `FASTA` format with nucleotide sequences.

```
>Example_header_1
AGTCCCCCAGAGAATTTT
>Example_header_2
AGATTTTCCGAGTCCAGTTTAGAC
...
```

New lines can be used to split longer sequences over multiple lines for improved readability. The addition of characters other than new lines or sequence characters are not permitted and will cause the format to be invalid if present. We recommend avoiding the use of special characters (*e.g.*, `*&^%$.`, see https://www3.ntu.edu.sg/home/ehchua/programming/howto/Regexe.html for a list) in sequence headers as they frequently can be misinterpreted by software due to their special meaning in the context of regular expressions or can cause the sequence to be invalid for certain software. Some software even automatically replaces them with non-special characters to ensure compatibility. Spaces and tabs in particular should be avoided as most software truncate sequences at these characters. A particularly problematic case of special characters can be introduced when files are opened and saved with windows applications as alternate and in most applications invisible newline characters are in use. Those then need to be replaced by the UNIX newline characters. Multiple extensions other than `fasta` are in use. Some of those are `fa` as a shortened form for `fasta`, `faa` if the sequences present in the file are amino acids or `fna` if the sequences in the file are nucleic acid sequences.

The `FASTQ` format is an extension of the `FASTA` format that in addition to a title and the sequence contains quality information for each nucleotide position (hence `FASTQ`). This format is used to convey raw sequence information along with a quality metric indicating the

reliability of the acquired sequences. A single entry in a `FASTQ` file is composed of four lines, with the first and second line being the title and sequence, respectively. The third line always starts with a + and can contain optional further information. The fourth and final line of each entry contains a quality string based on ASCII quality symbols with one symbol per nucleotide in the nucleotide sequence (*see* Subheading 2.6.2).

2.6.2 `FASTQ` format with read sequences.

```
@{read identifier 1}
{Sequence 1}
+{optional metadata or repeat of read identifier 1}
{Quality String 1}
@{read identifier 2}
{Sequence 2}
+{optional metadata or repeat of read identifier 2}
{Quality String 2}
...
```

The quality string contains information about the confidence level of the base-calling and can be used to identify read regions with poor quality that subsequently get removed.

While read files are commonly distributed in the `FASTQ` format, the order of the reads can vary if paired-end sequencing was performed, resulting in paired reads, *i.e.*, a forward and a reverse read for each sequenced DNA fragment. Forward and reverse reads can be present as separate files for each sample, by convention containing a `R1`/`r1` or `R2`/`r2` in the name for forward reads and reverse reads, respectively. A read pair usually shares the read identifier but can optionally have an additional tab-separated identifier marking it as the forward or reverse read, like `1:N:0:4`/`2:N:0:4` (*see* Subheading 2.6.3).

2.6.3 Paired reads in separate `FASTQ` files. The `1` in `1:N:0:4` indicates the forward read file while the `2` in `2:N:0:4` indicates the reverse read file.

Forward read `FASTQ` file

```
@{read identifier 1}        1:N:0:4
{Sequence                        1}
+
{Quality String 1}
@{read identifier 2}        1:N:0:4
{Sequence                        2}
+
{Quality String 2}
...
```

Reverse read `FASTQ` file

```
@{read identifier 1}        2:N:0:4
{Sequence                        1}
+
{Quality String 1}
@{read identifier 2}        2:N:0:4
{Sequence                        2}
+
{Quality String 2}
...
```

Both files must have an identical order of reads as many programs implicitly assume the same order of reads in the files to assign mate-pairs. Deviating orders in the files would consequently generate artificial DNA fragments, making any downstream analyses invalid. Thus, comparing the read identifiers at the start and end of the paired files using the shell *`head`* and *`tail`* commands, respectively, is highly recommended. Additionally, the total number of lines per file can be determined using *`wc -l`*. This number should be identical for the forward and reverse read files. Some applications also require the forward and reverse reads to be supplied as a single file with merged forward and reverse reads. Two types of merged files exist: `shuffled` (also called `interleaved`) and `unshuffled`. In shuffled files, each forward read in the file is followed by its respective reverse read, while in unshuffled files there is no prescribed order of reads. The use of the wrong merged file format will also lead to the creation of artificial DNA fragments due to the mispairing of reads. Thus, software requirements should always be checked to see whether they require shuffled reads.

2.6.4 Shuffled `FASTQ` file. In the shuffled `FASTQ` format, the reverse read (indicated here by `2:N:0:4`) always follows its forward read mate (`1:N:0:4`). The {read identifier} string is identical between mate pairs.

```
@{read        identifier        1}                                  1:N:0:4
{Sequence                                                               1}
+
{Quality String 1}
@{read        identifier        1}                                  2:N:0:4
{Sequence                                                               2}
+
{Quality String 2}
@{read        identifier        2}                                  1:N:0:4
{Sequence                                                               3}
+
{Quality String 3}
@{read        identifier        2}                                  2:N:0:4
{Sequence                                                               4}
+
{Quality String 4}
...
```

2.7 Software

The software mentioned herein must be installed on the UNIX-server. We recommend installing the software through an open-source package management system as most software has dependencies like specific versions of other software and those dependencies can be incompatible between software. Package management systems like `miniconda` provide both a convenient way to install software as well as resolving these dependency incompatibilities, as they allow the definition of so-called environments, in which new software can be installed in isolated containers. They also do not require root permissions, making it possible for every user to install software locally. `Miniconda` can be downloaded from https://docs.conda.io/en/latest/miniconda.html. Select a Linux-based installer with `python3` as the pre-installed `python` version and your systems bit version (likely 64-bit). The installer is provided as a shell script (`.sh`) and can be executed with the *`bash`* command. Please see https://conda.io/projects/conda/en/latest/user-guide/index.html for an introduction on how to install and use `conda` to create and manage environments and install software in them.

The `uBin` wrapper scripts repository needs to be downloaded (https://github.com/ProbstLab/uBin-helperscripts). We recommend the creation of two `conda`

environments, one for each major `python` version. The `python 3.7 conda` environment can be installed using a `yaml` file provided with the `uBin` wrapper scripts. Some additional software not used within the wrapper scripts will be additionally installed:

2.7.1 Creation, loading and installation of `python3 conda` environment and software. This command will install the listed software and its dependencies. The command will create a `conda` environment called `uBin_input_generator_pyt37`. Please use the entire path to `uBin_wrapper_reqs.yaml` if you are not in the folder with the file. Prior to starting the contained software, the `conda` environment needs to be loaded using `conda activate uBin_input_generator_pyt37`.

```
conda env create -f uBin_wrapper_reqs.yaml
conda activate uBin_input_generator_pyt37

conda install -c anaconda -c bioconda -c agbiome abawaca bbtools git das_tool
megahit checkm-genome
```

Neither the required `ruby` version (2.3) nor all the `ruby` packages (`gems`) are available through `conda` and thus need to be installed separately. For this purpose, we recommend a ruby version and gem manager like `rvm` (https://rvm.io/). We provide installation instructions for the correct `ruby` version as well as the needed gems using this installer in the following section.

2.7.2 Creation of `ruby` 2.3 and installation of `ruby gem` dependencies using `rvm`. This command needs to be executed from within the downloaded `uBin-wrapper` scripts folder. Otherwise, the paths for the `nu-2.0.1.gem` and the `gemfile` file listing the required ruby dependencies need to be given to the bundle install and gem install commands, respectively.

```
rvm install 2.3
rvm use 2.3
gem install bundler -v 1.16
bundle install
gem install –local nu-2.0.1.gem
```

A `python2 conda` environment is also required to execute `MaxBin2` and `sickle`. It can be installed as follows:

2.7.3 Creation, loading and installation of the `python2 conda` environment and software. This command will install the listed software and their dependencies. Prior to starting the contained software, the `conda` environment needs to be loaded using `conda activate mg_pyt27`.

```
conda create -n mg_pyt27 python=2.7
conda activate mg_pyt27
conda install -c bioconda maxbin2 sickle-trim
```

Table III.1.2 contains all essential and optional software with a description and a source location. The source location contains information about the installation and use of the respective software. All necessary software will have been installed within the respective `conda` environment via Subheadings 2.7.1, 2.7.2 and 2.7.3 (see above). Optional software represents software that is only mentioned in this chapter without being used. We would like to note that there exists alternative software and procedures to those listed here and the field is under constant development with new software being published every day.

Table III.1.2: Software descriptions and source locations (in alphabetical order).

Software	**Description**	**Homepage**
Required software		
Programming languages		
Bourne Again SHell (bash)	GNU shell	https://www.gnu.org/software/bash/
perl	Perl programming language	https://www.perl.org/get.html
R	R programming environment	https://cran.r-project.org/
ruby	Ruby programming language	https://www.ruby-lang.org/de/downloads/
Python3 v3.7 Python2 v2.7	Python programming language. There are two incompatible releases (python2 and python3). python3 is still being actively developed but both are dependencies of other software	https://www.python.org/downloads/

Standalone software		
ABAWACA v1.07	Binning tool	https://github.com/CK7/abawaca
BBTools v37.62	Collection of tools for the analysis and preprocessing of sequencing data; BBDuk subtool that is used to trim reads	https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide
BLAST	Sequence aligner for amino acid or nucleotide sequences	https://www.ncbi.nlm.nih.gov/books/NBK279690/
Bowtie2	Read mapping tool	https://github.com/BenLangmead/bowtie2
CheckM	Estimates the completeness and contamination level in prokaryotic genomes	https://github.com/Ecogenomics/CheckM/wiki
Miniconda	Software environment manager	https://docs.conda.io/en/latest/miniconda.html
DAS Tool	MAG dereplication tool	https://github.com/cmks/DAS_Tool
DIAMOND	High-throughput aligner for amino acid and nucleotide sequences	https://github.com/bbuchfink/diamond
esomana	Manual emergent self-organizing maps-based binning tool	http://databionic-esom.sourceforge.net/user.html
git	Version control system; here mainly used to retrieve data from GitHub	https://git-scm.com/downloads
MaxBin2	Binning tool; uses differential coverage and *k*-mers	https://sourceforge.net/projects/maxbin2/
Megahit	Assembler, comparatively low memory requirement	https://github.com/voutcn/megahit
Miniconda	Software environment management and installation software	https://docs.conda.io/en/latest/miniconda.html
Prodigal	Prediction of open reading frames in prokaryotes	https://github.com/hyattpd/Prodigal
pullseq	Subsetting FASTA or FASTQ files	https://github.com/bcthomas/pullseq
rvm	Ruby package and versioning	https://rvm.io/

	software	
sickle	Read sequence trimmer	https://github.com/najoshi/sickle
SPAdes	Assembler, comparatively high memory requirement; includes many specialized modes (*e.g.*, plasmidSPAdes, metaviralSPAdes, metaSPAdes) as well as special modes for types of samples (*e.g.*, --isolate or --sc)	https://github.com/ablab/SPAdes
uBin	Genomic bin curation software	https://github.com/ProbstLab/uBin
uBin-helperscripts	Assembly processing wrapper and input generator for uBin	https://github.com/ProbstLab/uBin-helperscripts
Optional software		
Automata	Automated binner	https://github.com/KwanLab/Autometa
BinSanity	Automated binner	https://github.com/edgraham/BinSanity
CONCOCT	Automated binner using differential coverage	https://github.com/BinPro/CONCOCT
Crass	Identification of CRISPR arrays from unassembled metagenomic data	https://github.com/ctSkennerton/crass
CRISPRCasFinder	Identification of CRISPR arrays and Cas proteins	https://crisprcas.i2bc.paris-saclay.fr/
DESMAN	De novo extraction of strains from metagenomes	https://github.com/chrisquince/DESMAN
DRAM	Metabolic potential prediction	https://github.com/shafferm/DRAM
dRep	Dereplication of genomes	https://drep.readthedocs.io/en/latest/
FastANI	average nucleotide identity comparison tool	https://github.com/ParBLiSS/FastANI
FastQC	Quality control analysis tool for reads	https://github.com/s-andrews/FastQC
Graphbin	Refining bins using assembly graphs	https://github.com/Vini2/GraphBin

GRiD	Replication index calculator	https://github.com/ohlab/GRiDhttps://github.com/ohlab/GRiD
gRodon	Minimal generation time estimator	https://github.com/jlw-ecoevo/gRodon
growthpred	Minimal generation time estimator	http://ftp.pasteur.fr/pub/gensoft/projects/growthpred/
GTDB-tk	Taxonomic Classification of genomes	https://github.com/Ecogenomics/GTDBTk
HiCanu	Standalone long read assembler and hybrid long-read+short-read assembler	https://github.com/marbl/canu
inStrain	SNP calling, iRep calculation, strain estimation	https://instrain.readthedocs.io/en/latest/
iRep	Replication index calculator; only works for bacteria	https://github.com/christophertbrown/iRep
MAGE	Online platform for in-depth genome annotation and analysis	https://mage.genoscope.cns.fr/microscope/home/index.php
MetaBAT	Automatic binner	https://bitbucket.org/berkeleylab/metabat/src/master/
MetaCRAST	Reference-guided CRISPR detection in raw metagenomic reads	https://github.com/molleraj/MetaCRAST
metaFly	Hybrid assembler for nanopore+illumina data	https://github.com/agofton/metaFly
metaWRAP	Automated binning pipeline	https://github.com/bxlab/metaWRAP
Microbeannotator	Metabolic potential prediction	https://github.com/cruizperez/MicrobeAnnotator
NCBI eutilities	Command line programs to access	https://www.ncbi.nlm.nih.gov/books/NBK25497/
Nonpareil3	Estimation of diversity coverage at used sequencing depth	https://github.com/lmrodriguezr/nonpareil
pgap	Official NCBI genome annotation tool	https://github.com/ncbi/pgap
PILER-CR	CRISPR repeat identification	https://www.drive5.com/pilercr/

ra2	contig error correction; recommended to be used only on genomes due to runtime	https://github.com/christophertbrown/fix_assembly_errors
Unicycler	Hybrid assembler for nanopore+illumina data	https://github.com/rrwick/Unicycler
VICTOR	Clustering and classification of viral sequences	https://ggdc.dsmz.de/victor.php
VirFinder	Identification of viral sequences	https://github.com/jessieren/VirFinder
virsorter2	Viral particle prediction tool	https://github.com/jiarong/VirSorter2
VizBin	Manual binner using nonlinear dimensionality reduction	http://claczny.github.io/VizBin/

2.8 Databases

Various databases are required for the reconstruction of archaeal genomes from metagenomes. The installation of the `BBTools` suite contains various databases, including databases for sequencing artefacts (`sequencing_artifacts.fa.gz`), sequencing adapters (`adapters.fa`) and PhiX control sequences (`phix174_ill.ref.fa.gz`) that will be used for the trimming of raw reads. Further required databases are the `FunTaxDB`, a modified version of the `Uniref100` database (Suzek *et al.*, 2015) with added taxonomic information from `Uniref100` as well as additional taxonomic information from perfect (100 % similarity) matches to `NCBI-nr` entries. This database will be used for the functional and taxonomic annotation of genes. The `FunTaxDB` (Functional Taxonomic DataBase) can be downloaded in `FASTA` format from https://uni-duisburg-essen.sciebo.de/s/pi4cuYwyZ3KJVMl. Prior to its use, it will need to be compiled into a `DIAMOND` Basic Local Alignment Search Tool (`BLAST`) (Buchfink, Xie and Huson, 2015) database. `DIAMOND BLAST` is an alternative to conventional `BLAST` that can be up to three orders of magnitude faster at a similar accuracy. The database can be compiled using the following command (Subheading 2.8.1).

2.8.1 Formatting of the `FunTaxDB` as a `DIAMOND BLAST` database. The output file is in the binary `dmnd` format and is not compatible among major `DIAMOND` versions. This requires the database to be re-compiled if the used major version of `DIAMOND` is changed. Both versions of the database, extensions `FASTA` and `dmnd`, are about 80 GB in size.

```
diamond makedb --in FunTaxDBvX.X.fasta -d FunTaxDBvX.X
```

Additionally, various marker gene databases are used to generate single copy gene tables for `uBin` (Bornemann *et al.*, 2020) and to estimate completeness and contamination using `CheckM` (Parks *et al.*, 2015) or infer the phylogeny of recovered genomes using `GTDK-tk` (Chaumeil *et al.*, 2020). Marker genes for `CheckM` can be downloaded from https://data.ace.uq.edu.au/public/CheckM_databases/, the `GTDB-tk` reference database can be downloaded from https://data.ace.uq.edu.au/public/gtdb/data/releases/release95/95.0/auxillary_files/gtdbtk_r95_data.tar.gz, and `uBin` marker gene databases are downloaded alongside the `uBin-helperscripts`.

3.Methods

3.1 Read quality control

After sequencing, two `FASTQ` files should be present per sample, one with forward reads and one with reverse reads, or one shuffled `FASTQ` file with both types of reads (see data types). All read sequences in these files will have the same length between 100-250 bp, depending on the sequencing instrument. These sequences can be contaminated by Illumina adapters, sequencing artefacts, or spike-in controls like the PhiX controls (Mukherjee *et al.*, 2015). To check for those and remove them if present, the tool `BBDuk` from the `BBTools` suite (Bushnell, 2021) is used. In addition to contaminant sequences, the inherent read quality can also vary greatly across reads and nucleotide positions. Thus, a read quality control step is necessary to ensure that only reliable sequencing data are used in downstream analyses. For this purpose, the program `sickle` (JN Fass, 2011) is used.

Both `BBDuk` and `sickle` require a shuffled, *i.e.*, interleaved, file. The conversion of individual paired read files to a merged shuffled file and the reverse can be done with two shell scripts, https://gist.github.com/3521724 (de-interleaving) and https://gist.github.com/4544979 (interleaving). Once a shuffled file has been created, `BBDuk` and sickle can be used to remove contaminants and perform quality-based trimming, respectively (Subheading 3.1.1). If only single end reads, *i.e.*, reads without a mate pair, are available, they can be supplied to `BBDuk` using the `-in` flag and setting `interleaved=f`. Sickle provides the `se` option to switch to single end read mode. In this mode, the `FASTQ` file can be supplied using the `-f` flag and the output name can be set with the `-o` flag. The deinterleaving step can be excluded if single end

reads are used and the removal (`rm`) command should be modified to not remove the file ending on `PE.fastq` as it will contain the quality checked single reads.

3.1.1 Read Quality control and contaminant removal. The command is split into actual code, as well as comments. Lines starting with '#' are comments and are ignored by programming languages. If the required contaminant databases, *i.e.*, for adapters, phiX, and artefacts, are not in the PATH variable, they can be supplied through the full path to the respective database. The command assumes that the required `bbduk.sh`, `sickle` and `fastq_deinterleave.sh` scripts are in the current directory or in the PATH variable. If that is not the case, the full path to the software needs to be used.

```
# removes illumina adapters using BBDuk, requires an interleaved input file
bbduk.sh ref={adapters.fa} k=23 mink=11 hdist=1 tbo tpe ktrim=r ftm=5
in={interleaved}.fastq out={interleaved}_trim.fastq t=8 interleaved=t

# removes illumina artefacts and phix174 control sequences
bbduk.sh ref={phix174},{Illumina.artifacts} k=31 hdist=1
in={interleaved}_trim.fastq out={interleaved}_trim_clean.fastq t=8
interleaved=t

# Quality-string-based trimming of reads
sickle pe -t sanger -c {interleaved}_trim_clean.fastq -m
{interleaved}_trim_clean.PE.fastq -s {interleaved}_trim_clean.SR.fastq

# deinterleaving the interleaved file into PE.1 and PE.2
bash fastq_deinterleave.sh < {interleaved}_trim_clean.PE.fastq
{interleaved}_trim_clean.PE.1.fastq
{interleaved}_trim_clean.PE.2.fastq

# gathering basic statistics for orphan reads and removing intermediate
# and orphan read files
echo "cleaning up..."
echo $(( $(wc -l {interleaved}_trim_clean.SR.fastq | awk '{print$1}') / 4 ))
> {interleaved}_trim_clean.SR.txt
rm {interleaved}_trim_clean.PE.fastq {interleaved}_trim_clean.SR.fastq
{interleaved}_trim_clean.fastq {interleaved}_trim.fastq
```

This command will result in three output files: one file for quality checked forward (`PE.1`) and reverse (`PE.2`) reads, respectively, and one file containing the number of orphan reads (reads without a mate) that were removed from the files due to the mate being of too low quality (`SR.txt`). High numbers of thusly removed reads can indicate a systematic error, *e.g.*, the usage of an unshuffled input file or a problem with the databases of potential contaminants used. Consequently, the file should be checked to see whether it indicates any problems. We

also recommend checking the quality of reads using `FastQC` (https://github.com/s-andrews/FastQC) before and after quality control.

3.2 Assembly

Reads can be assembled using a multitude of different assemblers. However, we focus on the two most commonly used assemblers here: `MetaSPAdes` (Nurk *et al.*, 2017) and `MEGAHIT` (Li *et al.*, 2015). We generally prefer an assembly with `MetaSPAdes`, as it performs scaffolding. The downside of using `MetaSPAdes` is its high memory requirements, exceeding a Terabyte for very large and complex metagenomes. `MEGAHIT` can be used as a less memory-hungry alternative to `MetaSPAdes` that does not perform scaffolding. The basic execution of both assemblers is straightforward, requiring only the quality-checked read files and an output file name as the input.

3.2.1 Basic execution of `MetaSPAdes` command. Either make sure that the `metaSPAdes.py` executable is in your PATH environment or use the full path to the executable in the command.

```
python3 metaSPAdes.py -1 {for-read} -2 {rev-read} -o {outputfoldername}
```

3.2.2 Basic execution of `MEGAHIT` command.

```
megahit -1 {for-read} -2 {rev-read} -o {outputfoldername}
```

We recommend setting a fixed number of threads for the program to use with the `-t` option, *e.g.*, `-t 10`, and the user can optionally also define a memory limit using the `-m` option, *e.g.*, `-m 1000` (these flags are used in both `MetaSPAdes` and `MEGAHIT`).

The assembly of metagenomes is the individual analysis step with the longest runtime, with `MetaSPAdes` runtimes up to a week for large assemblies. The commands presented here are only suitable for non-amplified and paired-end data. If only unpaired data, *i.e.*, single read data without respective mates, are available, `MEGAHIT` can be used for the assembly. In this case, the `-r` option can be used to supply the single read file. Optionally, the used assembly parameters can be tuned further for the given assembly by *e.g.*, via modifying the *k*-mers used during the assembly process or subsampling. These modifications as well as additional `SPAdes` modes for the assembly of other types of data are covered in detail in the Notes sections 4.2 - 4.5 and 4.8 - 4.9.

If the assembly using MetaSPAdes is successful, multiple output files are produced, including FASTA files of the generated contigs (contigs.fasta) and scaffolds (scaffolds.fasta), assembly graphs (assembly_graph_with_scaffolds.gfa) and a log-file detailing the progress and results of the assembly (spades.log). In MEGAHIT assemblies, contig sequences are supplied in final_contigs.fa and a log file simply called log details the assembly process. It is recommended to check the log file in both MetaSPAdes and MEGAHIT assemblies for any warnings or errors that might have occurred during the assembly.

3.2.3 Scaffold / contig renaming. Both MetaSPAdes as well as MEGAHIT have a prescribed way of naming the output scaffolds. In MetaSPAdes, scaffolds get ranked according to their length compared to the entire assembly and their actual length as well as *k*-mer coverage are used to name them. Thus, a typical MetaSPAdes scaffold header has the format NODE_1_length_2150837_cov_25.407906, indicating that this is the longest scaffold in the assembly (NODE_1), that its length is 2150837 nucleotides and that the *k*-mer coverage is 25.407906. Note that while there is some correlation between *k*-mer coverage and the actual sequence coverage, this number does not reflect actual sequencing coverage. MEGAHIT headers are also numbered throughout the assembly and contain details on their length as well as additional metadata (>k141_579 flag=1 multi=11.0000 len=3709). We recommend modifying the scaffold/contig names with the Project and Sample names as well as creating a subset of the assembly of scaffolds/contigs >=1 Kbp in length, if multiple samples are being processed (See Note 4.7 for commands to rename and subset assembly files).

3.3 Processing of metagenomic assemblies

Once the reads are assembled, further processing is required to identify open reading frames and annotate them as well as to calculate the abundance, GC content, length, consensus taxonomy, and presence/absence of archaeal single copy genes of scaffolds/contigs. The processing is done using a wrapper script supplied along with the uBin software (https://github.com/ProbstLab/uBin-helperscripts). The wrapper requires a DIAMOND BLAST-compiled version of FunTaxDB (*see* 2.8.1 for compilation instructions). Then the wrapper can be run using the following command:

3.3.1 Execution of the `uBin` wrapper. If only an unpaired read file is available, the `-sr` option can be used to supply the single read file instead of the `-r1/-r2` options.

```
bash uBin_wrapper.sh -s {assembly-fasta} -p {Project_Sample} -r1 {QC'd for-
reads} -r2 {QC'd rev-reads} -e diamond -t {threads}
```

The command is designed so that the `DIAMOND` database is located in the `/bin/SCG` subdirectory of the wrapper script folder. If the file is located at a different location, the alternate location can be specified using the flag -u {path to `dmnd` format database}. The wrapper will calculate the GC content, length, coverage and taxonomy of scaffolds as well as predict the presence of single copy genes, all of which are used in the binning or bin curation steps. The individual analysis steps performed by the wrapper as well as their output files are explained below.

3.3.2 Mapping. Read mapping to the assembly is performed using `Bowtie2` (Langmead and Salzberg, 2012) in `--sensitive` and `--no-unal` mode and results in a file in Sequence Alignment Map (`SAM`) format containing alignment information for each read, *e.g.*, to which scaffold the read aligned, at which positions it aligned, and how many mismatches there were in the alignment. The `--no-unal` flag causes only reads that aligned to a scaffold with a score meeting the threshold to be reported. The alignment location is allocated via a scoring function, with the alignment being assigned randomly if a read has multiple equal score alignment options. The mapping not only creates a file with the extension `.sam` but also a file with the extension `.sam.log`, which contains some general information about the mapping. The percentage of reads that aligned to the assembly, *i.e.*, the overall alignment rate, is an important piece of information stored in the log file. This metric provides an indication of how representative the assembly is for the sequenced proportion of the community (represented by the quality-checked reads). Low percentages, *e.g.*, <20 %, indicate that the assembly might not be very representative of the microbial community. Thus, the alignment rate should be considered when drawing conclusions from the assembled metagenome about the community. The Sequence Alignment Map (`SAM`) file is used as the basis for the calculation of the coverage of the scaffolds, *i.e.*, how often each scaffold is covered by reads. Afterwards, as the `SAM` file is large and not used for any other application, it is deleted. Additionally, the GC content and length of the scaffolds is calculated in this step. This analysis step results in four files: the general mapping information (`.sam.log`) as well as the

coverage (`.scaffold2cov.txt`), length (`.scaffold2len.txt`), and GC content (`.scaffold2gc.txt`) information for each single scaffold.

3.3.3 ORF prediction and annotation. ORFs are predicted on scaffolds with a length >=1 kbp using `Prodigal` (Hyatt *et al.*, 2010). `Prodigal` takes the input sequences to train models to predict open reading frames on the input sequences. As genes are predicted on metagenome assemblies, the genes on the containing scaffolds frequently originate from a wide variety of organisms and could consequently be encoded with alternate genetic codes. Thus, `Prodigal` in "`-p meta`" mode is used, making sure that the models to determine the best codon usage are trained on a by-scaffold basis instead of across the entire `FASTA` file. The ORF prediction results in two files, the predicted gene sequences as amino acid sequences (`genes.faa`) and as nucleic acid sequences (`genes.fna`), respectively. Following their prediction, the ORFs (in amino acid format) are annotated against `FunTaxDB` with `DIAMOND` (Buchfink, Xie and Huson, 2015) with only the best hit at an E-value cutoff of 10^{-5} being reported. The output is in the `BLAST format 6` (see http://www.metagenomics.wiki/tools/blast/blastn-output-format-6 for an explanation about the format) and consequently has the extension `.b6`. Besides quality information about the best match between the query ORF sequences and the `FunTaxDB` database like alignment identity, length of alignment, E-values and bit-scores, it will also give the functional annotation of the best `FunTaxDB` hit, as well as its taxonomic annotation. Using this output table, a consensus taxonomy for each scaffold is determined by grouping all annotated ORFs by scaffold and determining the taxonomic assignment by majority consensus, starting with the domain level taxonomy and moving down the taxonomic levels until either the species level is reached or no single assignment at that level has a majority taxonomy (>50 %). Please see Bornemann et al., 2020 (Bornemann *et al.*, 2020) for more details about how the consensus taxonomy is retrieved. These consensus taxonomies are stored in a `scaffold2tax.txt` file.

3.3.4 Collection of scaffold information. Tables with information about the scaffolds (Consensus taxonomy, GC content, coverage, length) are collected in an overview table (`overview.txt`).

3.3.5 Prediction of single copy genes per scaffold. 51 bacterial and 38 archaeal universal single copy genes (Probst *et al.*, 2017) are predicted by blasting against sequence databases of those marker genes. The results on a per-scaffold basis are collected in a comma-separated table (`SCGS.csv`), with the first column containing scaffold-ID's and 89 other

columns containing the number of hits on the scaffold for the respective bacterial (B_ prefix in column name) or archaeal (A_ prefix) single copy genes.

3.4 Binning of Metagenome-Assembled Genomes (MAGs)

Various binning tools are employed in this section to bin genomes. They do not all need to be used for the binning of every metagenome and many additional tools exist that can be used for binning. The same tool can also be applied using different parameters. After binning using various tools, the generated sets of genomic bins are combined using the Dereplication Aggregation Scoring Tool (`DAS_Tool`) (Sieber *et al.*, 2018). Beware that any binning tools relying on differential coverage can only be used to bin MAGs but no Single-cell Amplified Genomes (SAGs) or Mini-MAGs (MiMAGs) due to amplification biases (see Note 4.8).

3.4.1 Calculation of 4-mer frequencies. The binning tools `esomana` (Ultsch, 2005; Dick *et al.*, 2009) and `ABAWACA` (Brown *et al.*, 2015) require *4*-mer frequencies to be pre-calculated. This can be done with the perl script `esomWrapper.pl`, included in the https://github.com/tetramerFreqs/Binning.git GitHub repository (Dick *et al.*, 2009). The entire repository can be downloaded using:

3.4.1.1 Copying of the GitHub directory 'Binning'. This command will create a Binning directory inside the current directory.

```
git clone https://github.com/tetramerFreqs/Binning.git
```

To execute the script, the assembly `FASTA` file needs to be in its own folder. For manual binning using `esomana`, optional control genomes can also be placed within this folder so that they can be incorporated into the *4*-mer frequency calculations. They can serve as positive controls for the later training of the Emergent self-organizing map and should form coherent clusters. We use a high-GC genome (*Streptomyces griseus*, NC_010572.1) and a low-GC genome (*Escherichia coli*, NC_000913.3) to cover both extremes of the *4*-mer frequency range. The script can be executed with:

3.4.1.2 Calculation of *4*-mer frequencies. The `-min` option controls the minimal scaffold size to be used for the *4*-mer frequency calculation. The `-max` argument indicates the maximum length a scaffold fragment should have. Scaffolds are fragmented if they are longer than this parameter and the binning is later validated by checking whether the fragments of a scaffold were binned together. The location of the folder containing the esomWrapper.pl script must be either supplied with the `-script` option or can be added into the code by replacing the paths in L113 to L114 of the `esomWrapper.pl` script.

```
perl esomWrapper.pl -min {int} -max {int} -p {directory with .fasta files} -
dir {outputfoldername, optional}
```

See `perl esomWrapper.pl -h` for an explanation of the parameters. We generally use `3000/5000` and `5000/10000` for `-min/-max` respectively in `ABAWACA` and just `5000/10000` in `esomana`. All other parameters are optional. If a directory name is not supplied in the `-dir` option, an output folder called `ESOM` will be created instead. A description of the various output files is provided in *Table* III.1.3.

***Table* III.1.3: Explanation of output files of the esomWrapper.pl script.**

Extension	Description
.names	Assignment of scaffold fragments to scaffolds
.lrn	all *k*-mer frequencies of scaffold fragments
.mod.lrn	Subset of *k*-mer frequencies not containing AUG
esom.fasta	All FASTA sequences
esom.log	Log file containing information about the run of the script and recommendations for the 'rows' and 'columns' parameters used within the 'esomana' manual binning software

4-mers containing the start-codon AUG are overrepresented due to its nearly ubiquitous use and thus are not characteristic for any specific genome. Consequently, a subset of *4*-mer frequencies containing AUG-less *4*-mers is supplied as `mod.lrn` and used in the respective binning tools.

3.4.2 ABAWACA can be executed using:

3.4.2.1 Execution of the automatic binner `ABAWACA`. The `ABAWACA` binning tool uses the output files created by the `esomWrapper.pl` script.

```
abawaca {.names} {.mod.lrn} {esom.fasta} {outputfoldername}
```

`ABAWACA` needs to be executed separately for each set of `-min/-max` calculated *4*-mer frequencies. The software uses a dimensionality reduction and clustering approach to bin

genomes and is relatively fast compared to other binning tools. The final set of genomes will be in a new subdirectory called final-clusters in the output folder as a set of `FASTA` files.

3.4.3 MaxBin is a binner utilizing both differential coverage and *4*-mer frequencies to bin genomes. It can be installed via `conda` (*see Table* III.1.2). The required input is the assembly `FASTA` file along with coverage information for the assembly. If only a single sample is available, the coverage information can be supplied in the form of the `scaffold2cov.txt` table already generated during the processing during the metagenomic assembly using the `-abund` option. If multiple samples of the same ecosystem are available, the interleaved read files can be supplied with the `-reads` or `-reads_list` options. The `MaxBin` software can be executed using the command:

3.4.3.1 Execution of the automatic binner `MaxBin`. If multiple samples are available, the `-abund` option can be exchanged for the `-reads` option and interleaved read files can be supplied. They need to be supplied separately to `-reads` flags, *e.g.*, `-reads`, `-reads2`, `-read3`, or a single file enumerating the locations to the read files can be supplied to the `-reads_list` flag.

```
perl run_MaxBin.pl -c {assembly} -abund {scaff2cov.txt 1} -t {in} -markerset
{40 or 107} -o {outputname}
```

It is recommended to make a new folder for this analysis, as `MaxBin` produces a lot of files in the folder where it is executed. `MaxBin` comes with two marker sets that are used to identify seeds for the clustering algorithm. The default 107 marker genes encompassing marker set works well for conventional prokaryotic genomes, while the 40-marker gene dataset contains universal markers for bacteria and archaea that may work better for reduced genomes like Bacteria of the Candidate Phyla Radiation or Archaea of the DPANN superphylum. Thus, running `MaxBin` with both marker sets is recommended.

3.4.4 Concoct is an alternative to `MaxBin` that also uses differential coverage and *4*-mer frequencies and requires the same input files. `Concoct` is much slower compared to `MaxBin` making it unsuitable for very large assemblies (>50 GB read file size). `Concoct` requires large scaffolds to be chopped up in a similar fashion to `ABAWACA` and `esomana` and supplies the `cut_up_fasta.py` script to do. Details on usage of both the auxiliary script and `Concoct` are supplied at https://concoct.readthedocs.io/en/latest/usage.html.

3.4.5 Manual binning tools. There are multiple manual binning tools available. Manual binning as a general rule of thumb is a time-consuming process but can also produce

comparatively higher quality bins when working with low complexity samples (Sieber *et al.*, 2018). Manual binning tools are not suitable as a high-throughput method, making them less suitable for larger sample sets. Two manual binning software that are available are `esomana` (Ultsch, 2005) and `VizBin` (Laczny *et al.*, 2015). The `esomana` software can be used to manually define genomic bins based on *4*-mer frequencies that are organized in emergent self-organizing maps. It can be downloaded from http://databionic-esom.sourceforge.net/index.html and uses the `mod.lrn` file that `ABAWACA` also utilizes as input. Please see the User Manual under the provided link for an introduction to `esomana`. We recommend training ESOMs with the default `Online training`, adjusting the `start value for radius` to `50` and the number of rows & columns in map, *i.e.*, the size of the ESOM, to the values suggested in the `esom.log` file. After the binning using `esomana` has finished and a new `CLS` file containing the new bin assignments has been saved, the bins can be extracted using the previously downloaded `getClassFasta.pl` script with the command:

3.4.5.1 Extraction of Bins in `FASTA` format from `esomana` output. This script takes the assigned bins `{.cls}` as well as the `NAMES` and `esom.fasta` files from the tetramer frequency calculation as input and extracts the bin with the number defined in `{bin-nr}` as a `FASTA` file called `{bin-nr}.fasta`. To extract all bins in a single command, the `bin-nr`'s can be looped over with `seq 1 1 {max. bin-nr}`.

```
perl getClassFasta.pl -cls {CLS} -names {NAMES} -fasta esom.fasta -loyal 51 -num {bin-nr}
```

`VizBin` uses a `FASTA` file with the assembled sequences as input and performs a nonlinear dimensionality reduction to prepare the input sequences for manual binning (Laczny *et al.*, 2015). Selected bins can directly be exported as `FASTA` files.

3.4.6 Additional binning tools. Many further binning tools are available. Some of those are `GraphBin` (Mallawaarachchi, Wickramarachchi and Lin, 2020), `Autometa` (Miller *et al.*, 2019), `MetaBAT` (Kang *et al.*, 2019), `binsanity` (Graham, Heidelberg and Tully, 2017) and `metawrap` (Uritskiy, DiRuggiero and Taylor, 2018). As any arbitrary number of binning tools can later be combined using `DAS Tool` (see below), there is no technical restriction in the amount of binning tools used. There are diminishing returns though after the first few binners, as most binners use similar input data (Sieber *et al.*, 2018), *i.e.*, either just *k*-mer frequencies or also differential coverage, and thus only differ in the grouping algorithm

used, making the differences less pronounced once a certain number of binners has been applied (depending on the complexity and data structure of the data set).

3.4.7 Generation of scaffold2bin tables. Most Binning tools produce the output bins as FASTA files and thus need to be converted to scaffold2bin tables to serve as input for DAS Tool. The conversion of all FASTA files in the current folder to a scaffold2bin.txt table can be done using the script Fasta_to_Scaffolds2Bin.sh script supplied along with the DAS Tool software with the following command (no files other than the genomes ending on FASTA must be in the current directory). Use the script with the parameter help to see optional input parameters.

3.4.7.1 Generation of scaffold2bin tables for the DAS Tool input. The script assumes that the genomes are present as FASTA files in the current folder with the fasta extension. The -i and -e options can be used to set non-default folder locations and extensions, respectively. The script will produce a table containing the scaffold names in the first column and the bin names in the second column.

```
bash Fasta_to_Scaffolds2Bin.sh. > {scaf2bin.txt}
```

The script will produce a {scaf2bin.txt} table in the current folder with the FASTA file names as bin names. DAS Tool requires the bin names to be unique across all input bin sets. Thus, some {scaf2bin.txt} tables may need to be further modified to make the bin names unique. This can be done using the bash command:

3.4.7.2 Renaming bin names to make them unique.

```
sed -i "s/$/_{unique binset identifier}/" {scaf2bin.txt}
```

3.4.8 Aggregation of bin sets. Bins are aggregated using DAS Tool (Sieber *et al.*, 2018). DAS Tool first predicts ORFs using Prodigal and identifies ORFs belonging to universal single copy genes in Archaea and Bacteria, thus assessing completeness and contamination in the bins of the various bin sets. Bins across bin sets are matched and dereplicated, aggregated, and finally scored based on their completeness and contamination. All bins meeting the score threshold are finally output as a scaffold2bin table. The basic DAS Tool command can be run like this:

3.4.8.1 Execution of DAS Tool for the aggregation of individual bin sets. Beware that there must not be any spaces in the comma-separated strings given in the -i/-l options for

the bin sets (`-i`) or their labels (`-l`), respectively. Thus, bin set file names must not contain any spaces. The labels supplied in the `-l` option are:

```
DAS_tool -c {assembly.fasta} -i {binset1},{binset2},{binset3},... \
-l {binset1},{binset2},{binset3},...-t {10} -o {outputprefix}
```

Running just the `DAS_tool` script without parameters shows all the options.

3.5 Manual curation of metagenomic bins

3.5.1 Addition of bin information. The scaffold2bin information from `DAS Tool` can be added to the overview file with the command:

3.5.1.1 Addition of bin information to `overview.txt` files. The script `09_additionbincol.sh` is supplied along with the `uBin` wrapper scripts.

```
bash 09_additionbincol.sh {scaffolds2bin} {overviewfile} > {outputfilename.txt}
```

3.5.2 Transfer of files to the local machine. The newly created overview file with bin information as well as the `SCGS.csv` file created during processing and the assembly `FASTA` file must be moved to the desktop computer that has an installation of the `uBin` software. Please either use a file sharing program, *e.g.*, `Cyberduck` or `Forklift`, or the `scp` command to download the files to your local computer. The following command is a template for the usage of the `scp` command.

3.5.2.1 `scp`-based Transfer of files from remote location to the local computer. The port number needs to be supplied to the `-P` option if it differs from the default `22`.

```
scp -P {int} {username}@{IP address}:{Path to files on server} {local file destination}
```

Optionally, `{PATH to files on server}` can also be the folder containing the files. In that case, the `-r` option needs to be used.

3.5.3 Manual curation of genomes with uBin. Open the `uBin` (Bornemann *et al.*, 2020) application and click on the `Import` tab (Layout Import Tab). Navigate to the files you just transferred in the file system on the left side. First click on the ⊕ symbol next to the overview file with the added bin information and then the `SCGS.csv` file. Enter a name for the sample. This name will be used as the basename for the curated bins and we thus recommend using the `{Project_Sample}` as the name. The name will be used to differentiate between samples

in uBin and thus needs to be unique. You can choose between three import options: 1) Everything, *i.e.*, both binned and unbinned scaffolds, 2) Binned scaffolds and 3) Unbinned scaffolds. After selecting the files, naming the sample and selecting the import option, press the import button. A loading window will appear and close once the data have been imported. Afterwards, you can switch to the Samples tab (Layout Samples Tab). Click on Import/Export>Import Records>{your Sample} to select your sample for bin curation. The Samples tab consists of six plots of which all but the Single Copy Gene (SCG) plots on the right showing genome completeness and contamination for bacteria (top) and archaea (bottom), respectively, are interactive. This means that any of the plots can be used to select specific ranges of GC content, coverage, or taxonomies and all other plots will change according to the new selection. The selected ranges are shown in the top. You can select a bin to curate through Select Bin. A video tutorial is provided on GitHub showing how to curate bins (Video tutorial) along with an interface description (Interface description). Saving a bin will save the current selection as a new bin named {SampleName}_{Taxonomy}_{GC}_{Coverage}. The original bin with the scaffolds that did not get included in the curated bin will also remain available in the Select Bin tab.

3.5.4 Identification and curation of archaeal bins in uBin. The Single Copy Genes and taxonomy wheel can be used to identify archaeal bins for curation. To do this, go through the bins in Select Bin and see whether any bins are indicated to be archaeal based on their taxonomy and Single Copy Genes and curate the ones thus identified as Archaea.

3.5.5 Export of bins from uBin. Once all bins have been curated, it is recommended to go through your bins and delete all the bins not meeting your expectations with regard to completeness & contamination. You can delete bins by clicking on the red trash bin symbol in the top right. Please see Bowers et al. (2017) as a reference for common MAG quality cut-offs (Robert M. Bowers *et al.*, 2017). After you have finalized your set of curated bins, you can click on the export button in the top right of the Samples tab. A window will pop up asking you to select a target directory. We recommend to also select a FASTA format file with the assembly sequences to directly export your bins as FASTA files. After pressing the Save button, a new overview table will be created containing the curated bin assignments. If an assembly FASTA file was supplied, a folder containing the curated bins in FASTA format is also created.

3.6 Branch-specific completeness and contamination prediction

`CheckM` (Parks *et al.*, 2015) can be used to roughly assign genomes to phylogenetic branches and then calculate their completeness and contamination levels using branch-specific sets of marker genes. Since these marker genes are not the same set as used in `DAS Tool` or `uBin`, they can be regarded as an independent confirmation of the bin quality. We refer to the `CheckM` wiki for usage instructions (Quick Start workflow). While marker genes give an indication of completeness and contamination of MAGs, they still need to be considered as estimates as these marker genes only represent a small repertoire of genes of the respective organisms. Only circularized genomes can be assumed to be complete. Nevertheless, we recommend confirming circularity via read-mapping as described in (Chen *et al.*, 2020).

3.7 Filtering out low-quality genomes

Depending on the application, various MAG quality cut-offs are in use, either using just the completeness and contamination metrics or also incorporating the presence of various rRNAs and tRNAs. The cut-offs largely depend on the type of downstream analyses the genomes will be used for, *e.g.*, growth estimators require fairly complete genomes. We refer to Bowers et al. (2017) (Robert M. Bowers *et al.*, 2017) for guidelines on quality categories of MAGs as well as genomes recovered from mini-metagenomes (MiSAGs) and single-cell genomes (SAGs).

3.8 Phylogenomic placement

Genomes are taxonomically placed using the classify workflow of `GTDB-tk` (Chaumeil *et al.*, 2020). This workflow first identifies 120 bacterial and 122 archaeal marker genes, aligns them and then finally infers the phylogeny of the genomes based on their placement on the `GTDB-tk` reference tree. Prior to running the classify workflow, the path to the reference database needs to be set. `GTDK-tk` provides the environment variable `GTDBTK_DATA_PATH` for this purpose. Thus, the command to first set the variable and then run the workflow is:

3.8.1 Phylogenomic placement of genomes using `GTDK-tk`. The modification of the `GTDBTK_DATA_PATH` variable is temporary, *i.e.* needs to be repeated with every new terminal session, unless the command is added to the shell profile (*e.g.*, '`.bashrc`').

```
# Set the GTDBTK-DATA_PATH variable
export GTDBTK_DATA_PATH={path to GTDB-tk reference data folder}

# run the classify__wf workflow of gtdb-tk
gtdbtk classify_wf --genome_dir {genomes as FASTA} --out_dir {output-dir}
```

The number of CPUs can be adjusted with the `--cpu` option. Among the many output files, there are two `summary.tsv` files, summarizing the results for bacteria and archaea, respectively. These include the GTDB taxonomy for the target genome as well as its closest reference genome and the Average Nucleotide Identity (ANI) to the reference genome if closely related (ANI >=80 %). An ANI >= 95% indicates that the reference genome might belong to the same species (Jain *et al.*, 2018). The classify workflow also produces a treefile in `Newick` format that can then be visualized in software like `iTOL` (Letunic and Bork, 2016). Since phylogenomic placement of MAGs or entire lineages is a complicated issue, we include some additional suggestions in the Note 4.14.

3.8 Beyond binning

The binning of genomes is just the beginning in genome-resolved metagenomics, as it opens up the possibility of pursuing many different types of analyses. Some of those are quickly presented in this section.

3.8.1 Extending and completing genomes. During the curation of genomes, the main objective is to remove contaminating sequences. The removal of contaminating sequences reduces the possibility for false positive predictions caused by foreign sequences in genomes and is essential to be able to interpret the genomic data. But these genomes are frequently fragmented and very incomplete. Just like the process of curating genomes by removing contaminating sequences, they can also be improved in completeness by assigning additional scaffolds whose actual assignment needs to be carefully checked (including GC-content, coverage, etc.). Additionally, scaffolds can be extended via mapping reads and further assembly to produce more complete and finally possibly circularized genomes (we recommend the software `Geneious` for this). This process is very time- and labor-intensive as it requires many cycles of mapping, manual inspection, and sequence editing. Only circularized genomes can allow reaching conclusions about the absence of genes and thus the completeness or patchiness of certain pathways (setting aside proteins of unknown function). Consequently, completing genomes might be an option for key organisms in datasets, provided the initial bins are already of high quality. We refer to Chen et al. (2020) for a guide on how to go about the extending and merging of scaffolds and the completion of genomes (Chen *et al.*, 2020).

3.8.2 ORF prediction on genomes. `Prodigal` in normal mode is usually the first option used to predict ORFs from genomes, as most organisms share a genetic code.

`Prodigal` in normal mode uses the entire genome to train its prediction models and thus is generally more accurate than Prodigal in meta mode. By default, Prodigal in normal mode uses the universal genetic code 11 but also tries genetic code 4 (Mycoplasma/Spiroplasma), if the average training gene length is too low. Other codes can be used (see genetic codes for all available genetic codes), if set manually. As per its documentation, Prodigal does not distinguish between genetic codes 4 and 25 (in the CPR lineages Absconditabacteria and Gracilibacteria), thus the latter has to be set manually if the user is aware of such a taxonomic affiliation.

3.8.3 Dereplication of Genomes. If multiple metagenomic samples are analyzed from the same ecosystem, it is likely that many of the recovered genomes are recovered multiple times. This makes a dereplication step necessary to get a representative set of genomes for this ecosystem and select the best representative for each genome cluster. For this purpose, `dRep` (Olm *et al.*, 2017) has been developed which scores genomes based on `CheckM`-derived completeness, contamination and strain heterogeneity measures as well as the total genome size and its N50 value [a metric describing the fragmentation degree of the genome (Earl *et al.*, 2011)]. It then clusters the genomes based on `Mash` (Ondov *et al.*, 2016) followed by `gANI` clustering and then uses the score to select the best representative for each genome cluster (Olm *et al.*, 2017).

3.8.4 Prediction of minimal generation time. Minimal generation times can be calculated based on the difference in codon usage bias between housekeeping genes like ribosomal proteins and the other genes in the genome. These calculations are implemented in the `growthpred` (Vieira-Silva and Rocha, 2010) software as well as the `R` (R Core Team, 2008) package `gRodon` (Weissman, Hou and Fuhrman, 2020). Please note, that this calculation provides only an estimate of the minimum generation that is theoretically possible and does not provide evidence that this is the case in the ecosystem or during sampling. At the moment, calculation of replication indices using Peak to Trough ratio methods, like `iRep` (Brown *et al.*, 2016) is impossible for Archaea, due to a series of peculiarities in their replication, such as the existence of multiple origins of replication (Zatopek, Gardner and Kelman, 2018).

3.8.5 Prediction of metabolic potential. One of the main advantages of genome-based analyses compared to either assembly- or read-based analyses is that specific ecosystem functions can be assigned to community members based on predicting the functions encoded

in the genes of the respective genomes. For this purpose, manual BLAST analysis against various databases from NCBI (Uniprot100, ncbi-nr, Uniref) or against the FunTaxDB (included in uBin wrappers, see above) can be performed. However, many more sophisticated metabolic potential analyses suites have been developed that either query the genomes using Hidden-Markov Models (HMMs) of target genes or the ORFs against various databases. Some of these suites are DRAM (Shaffer *et al.*, 2020), MicrobeAnnotator (Ruiz-Perez, Conrad and Konstantinidis, 2021), KOALA (Kanehisa, Sato and Morishima, 2016) and PGAP (Tatusova *et al.*, 2016). For a very detailed analysis of individual genomes, MAGE by *Genoscope* (Vallenet *et al.*, 2006) can also be employed.

3.8.6 Virus-host matching. Viruses are the most abundant entities on Earth (Pan *et al.*, 2017). Consequently, investigating the virome of a respective sample is often necessary to obtain a complete picture of the interactions in a microbial community. Many tools exist to identify viral particles, including Virsorter2 (Guo *et al.*, 2021) and Virfinder (Ren *et al.*, 2017). They can also be clustered to identify viral clusters using Victor (Meier-Kolthoff and Göker, 2017). A virus-specific HMM-database, the VogDB (http://vogdb.org/download) is available to annotate viral genes. The CRISPR-Cas system is one of the mechanisms by which cellular organisms defend against foreign DNA and RNA like viruses and works by incorporating snippets of this foreign DNA ("spacers") into CRISPR arrays in the genome to later recognize re-occurring invasion by these particles and defend against it (Lillestøl *et al.*, 2006; Makarova *et al.*, 2006). Thus, by matching these spacers to protospacers, *i.e.*, sequences matching to the spacers on viral genomes, potential hosts of these viruses can be determined and infection histories can be reconstructed (Shmakov, Wolf, *et al.*, 2020). CRISPR arrays can be identified in assembled sequences with PILER-CR (Edgar, 2007), Crass (Skennerton, Imelfort and Tyson, 2013), with CRISPRCasFinder (Couvin *et al.*, 2018) additionally identifying Cas proteins, or directly from reads using Crass (Skennerton, Imelfort and Tyson, 2013). CRISPR Spacers can be extracted from the reads using MetaCRAST (Moller and Liang, 2017) using previously identified repeat sequences. As an example for a detailed study on novel archaeal viruses and their infection histories, *see* (Rahlff *et al.*, 2021).

3.8.7 Single-Nucleotide-Polymorphism (SNP) and strain analysis. Genomes recovered from metagenomes are so-called population genomes as they do not reflect the genotype of individual organisms but rather the average sequence across an entire population of the

respective organism (Olm *et al.*, 2021). To resolve the intra-population differences in a population genome, SNP analyses can be utilized. Various tools have been developed for this purpose, including `inStrain` (Olm *et al.*, 2021) and `STRONG` (Quince *et al.*, 2021).

4.Notes

4.1 Read quality control for new sequencing technology

When using a new sequencing technology, it should be checked whether their adapter sequences are in the respective adapter sequence file. For instance, adapter sequences can be found on the Illumina homepage (Illumina Adapters).

4.2 Assembling single-end metagenomes

While recent Illumina sequencing projects almost exclusively use paired-end sequencing, many of the metagenomes available on SRA are still single-end. In single-end sequencing, only one side of each DNA fragment is sequenced and thus no paired-end information is available. Consequently, no scaffolding can be performed. Thus, assemblers like `MEGAHIT` are used instead of `MetaSPAdes`.

4.3 Assembly of pure cultures

If your dataset is expected to contain only a single organism, the original `SPAdes` algorithm (instead of `MetaSPAdes` for mixed communities) should be used. This algorithm will generally result in a better assembly for pure culture genomes, as the assumption that only a single organism is present in the sample makes a simplification of the assembly procedure possible (Bankevich *et al.*, 2012) as well as allow for the use of more sets of *k*-mers during assembly. Thus, we highly recommend using a dedicated pure culture metagenome assembler like `SPAdes` for this type of metagenome. `SPAdes` also has an `--isolate` option that is recommended for high coverage prokaryote/viral isolate as well as multi-cell organism data. Apart from these de-novo assembly strategies, reference assemblies guided by genomes of closely related organisms can sometimes also bolster genome assembly, *e.g.*, see `AMOScmp` (Pop *et al.*, 2004).

4.4 Communities with extremely highly abundant organisms

Some metagenomic samples like biofilms are enriched in specific organisms and individual species sometimes constitute 80% of the community (Probst *et al.*, 2014). These population

genomes often contain multiple strains that break the assembly, due to ambiguous paths in assembly graphs. In these cases, a separate assembly strategy is applied which relies on a) assembling a subset of reads (*e.g.*, 1%) that map to an initial draft genome (often of low quality) of the dominant species and b) assembling the reads that were not mapped separately (Wrighton *et al.*, 2012). Coverage estimates on subassemblies should still use all the reads of a sample.

4.5 Modification of k-mers used during assembly

Each assembly software has its default set of *k*-mers that are used for the respective assembly. These *k*-mers can range from as low as 21-mers up to *k*-mers corresponding to the minimum read length in the sample. The modification of *k*-mers used and/or the addition of more sets of *k*-mers used can result in an improved assembly quality. Fine-tuning this parameter is a rather time-consuming process and frequently only improves the assembly marginally, if at all. Thus, it should only be attempted when trying to, *e.g.*, recover a specific genome in the highest possible quality. *See* Page et al. (2016) for an example of fine-tuning the *k*-mer parameter for a specific assembly (Page *et al.*, 2016).

4.6 Targeted assembly of plasmids and viruses

In addition to being identifiable in the assemblies, virus- and plasmid-specific assembly workflows also exist within the `SPAdes` software suite called `MetaViralSPAdes` (Antipov *et al.*, 2020) and `PlasmidSPAdes` (Antipov *et al.*, 2016), respectively. These can be used to specifically recover viruses and circular elements, respectively.

4.7 Modification of scaffold/contig names

We recommend modifying the scaffold/contig names by replacing either the `NODE` in `MetaSPAdes` scaffold names or the ‘k’ in `MEGAHIT` scaffold headers with a `{Project}_{Sample}` designation. This has two advantages: 1) When working with multiple samples, the header name lets you know to which sample this sequence belongs to and 2) it ensures that the headers are unique across samples. Most tools have an implicit requirement of sequences having unique headers but only sometimes check if this requirement is met and can thus produce erroneous results. We provide commands to modify the scaffold/contig headers for `MetaSPAdes` scaffolds (*see* 4.7.1) and `MEGAHIT` contigs (*see* 4.7.2).

4.7.1 Renaming of `MetaSPAdes` headers. This command will remove the decimals of the *k*-mer coverage in the scaffold header name, replace `NODE` with the supplied name in `{Project_Sample}` and replace any '.' or '-' occurring in the header. Note that the `-i` option in sed will cause the original file to be modified instead of the output being printed to stdout.

```
sed -i "s/\.[0-9]\+$//g;s/>NODE_/>{Project_Sample}_/g;s/\./_/g;s/-/_/g" {fasta}
```

4.7.2 Renaming of `MEGAHIT` headers. This command will replace `NODE` with the supplied name in `{Project_Sample}` and replace any '.' or '-' occurring in the header.

```
sed -i "s/>k/>{Project_Sample}_/g;s/\./_/g;s/-/_/g" {fasta}
```

We also recommend using only scaffolds of at least 1 Kbp in length for gene prediction. Thus, a command creating a subset of the assembly file containing only sequences >=1 Kbp of is provided (*see* 4.7.3).

4.7.3 Making a subset >=1000 bp of the assembly. The `pullseq` (https://github.com/bcthomas/pullseq) application is required. It should either be in the `PATH` variable or the full path to the application needs to be used.

```
pullseq -i {fasta} -m 1000 > {Project_Sample}_min1000.fasta
```

4.8 Amplified metagenomes

Multiple Displacement Amplification (MDA) amplifies DNA using non-specific primers. This method can provide sufficient DNA for library preparation even from single cells and makes it possible to analyze genomes from individual organisms instead of population genomes recovered from conventional metagenomes (Woyke *et al.*, 2010). This process comes with similar drawbacks as amplicon-based metagenomics, *i.e.*, sequences are amplified unevenly and thus the observed abundance no longer reflects the abundance in the community (Rinke *et al.*, 2013). It also affects the assembly strategy as it can no longer be assumed that *k*-mers belonging to the same DNA sequence share an abundance (Rinke *et al.*, 2013). Thus, alternate assembly modes need to be used. `SPAdes` has a separate `--sc` mode for the analysis of single cell metagenomes and mini-metagenomes (Bankevich *et al.*, 2012) that takes uneven coverage into account. Please note, that coverage information cannot then be used to bin draft genomes from the dataset due to a distortion in sequence profiles.

4.9 Long-read data

Recent advances in nanopore sequencing have enabled assembling complete circularized bacterial genomes directly from the reads without any need for the binning of genomes (Moss, Maghini and Bhatt, 2020). While the base-calling in Nanopore sequencing is very inaccurate compared to conventional Illumina short-read sequencing (Noakes *et al.*, 2019), the extremely long reads of up to Mbps in length significantly improve the assembly of short reads as problematic regions in short-read sequencing like conserved repeats can be resolved (Moss, Maghini and Bhatt, 2020). However, Nanopore and other long-read sequencing methods require specialized DNA extraction protocols, as long continuous DNA strands are necessary (Moss, Maghini and Bhatt, 2020). Long-read sequencing, either PacBio or Oxford Nanopore, is currently mainly used in conjunction with short-read data to support short-read assemblies in hybrid approaches (*i.e.*, scaffolding). While long-reads can also be added to the assembly in (`meta-`)`SPAdes`, many specific hybrid assemblers exist (*e.g.*, `Unicycler` (Wick *et al.*, 2017), `MetaFlye` (Kolmogorov *et al.*, 2020) or `HiCanu` (Nurk *et al.*, 2020)), with no specific hybrid assembler currently as the community standard (as of the writing of this chapter).

4.10 Contamination controls

Environmental metagenomics is frequently hindered by low biomass and consequently extremely vulnerable to contamination. This contamination can be introduced during all steps of the wet lab component of the metagenomic analysis, *i.e.*, sampling, DNA extraction, library preparation, and sequencing. Thus, including negative controls into the analyses where possible is recommended, *e.g.*, blank extractions or field blanks. There are lists of bacterial genera that are frequent contaminants in molecular reagents (Sheik *et al.*, 2018) and any genomes that are taxonomically classified to be members of these genera should be investigated further by comparing the recovered genomes to publicly available representatives using Average Nucleotide Identity (ANI) using, *e.g.*, the `FastANI` tool (Jain *et al.*, 2018). Genomes showing close similarity to these putative contaminant species (>95% ANI indicating same species), should be considered for exclusion from further analyses.

4.11 Targeted binning

In some instances, the objective of metagenomic binning is to recover a specific bin in the dataset instead of all bins. To do this, the target bin can be identified using a scaffold carrying

a respective marker gene. Afterwards, automatic binning followed by `DAS Tool` dereplication of the individual bin sets is performed and the bin that carries the specific scaffold is identified and curated in, *e.g.*, `uBin`. *See* Speth et Orphan. (2018) for an example of a marker gene-aided identification of target metagenomes in followed by the targeted binning of a *Methanomassiliicoccales* genome (Speth and Orphan, 2018).

4.12 Assembly error correction

Assembly errors can occur both in the assembly of contiguous sequences and in scaffolding. These errors can be detected by mapping reads to the genomes or assemblies, using various scoring thresholds for the mapping and identifying regions on scaffolds with little to no coverage support. One tool for the correction of assembly errors is `ra2` (Brown *et al.*, 2015) (https://github.com/christophertbrown/fix_assembly_errors). This software 1) identifies putative assembly errors using stringent mapping, 2) recruits reads mapping to the region at lower stringency and 3) re-assembles the sequence region (if possible), splits apart the sequence or replaces the non-assembled area with Ns. However, this tool is designed for individual genome bins only as it is computationally expensive.

4.13 Assessment of diversity coverage

The degree to which reads map back to the assembly, indicates how representative the assembly is for the sequenced reads. However, it does not show how well the microbial diversity in the ecosystem is represented by the metagenome, *i.e.*, the sequenced reads. The software `Nonpareil3` (Rodriguez-R *et al.*, 2018) estimates how much of the microbial diversity is covered at the current sequencing effort, *i.e.*, how representative a metagenome is of the microbial diversity in the ecosystem. It will also estimate how much sequencing depth would be needed to reach certain diversity coverage thresholds, *e.g.*, 95% or 99% of the microbial diversity. We refer to the GitHub page of `Nonpareil3` (https://github.com/lmrodriguezr/nonpareil/tree/master/docs) for instructions on its use.

4.14 Systematic classification and phylogenomics

`GTDB-tk` is but one of many automated pipelines that can be used for the taxonomic classification of MAGs, with common alternatives including `Phylosift` (Darling *et al.*, 2014) and `PhyloPhlan` (Asnicar *et al.*, 2020). Differences among such tools usually concern

the (presumably vertically inherited) marker genes used to create alignment supermatrices and the algorithm choices for alignments, trimming, and phylogenetic inference in each pipeline. `GTDB-Tk` has the added advantage of using its own standardized system of prokaryotic classification (GTDB) (Parks *et al.*, 2018, 2020) that counters misclassification issues that have crept into public databases like NCBI Taxonomy over time. However, since `GTDB-Tk` uses `FastTree` (Price, Dehal and Arkin, 2010) to construct phylogenies instead of more sophisticated Maximum Likelihood and Bayesian methods (Lartillot, Lepage and Blanquart, 2009; Stamatakis, 2014; Minh *et al.*, 2020) there exists potentially a speed-accuracy tradeoff. Domain-level phylogenies, such as the placement of a novel MAG within Archaea, tend to carry multiple bias the most pervasive of which is mutational saturation manifesting as long branch attraction *e.g.*, *see* (Kapli, Yang and Telford, 2020). Automated pipelines do not afford much modularity in phylogenomic analyses and one should be aware of alternative algorithms for the different analysis steps, as well as downstream analyses that can be used to counter biases in deep phylogenies, such as recoding, desaturation, and mixture models. The reader is referred to previous work using sophisticated phylogenomic approaches on archaeal phylogenies, such as (Raymann, Brochier-Armanet and Gribaldo, 2015; Dombrowski *et al.*, 2020; Martijn *et al.*, 2020).

2. uBin – a manual refining tool for genomes from metagenomes

Till L.V. Bornemann[1], Sarah P. Esser[1], Tom L. Stach[1], Tim Burg[2], and Alexander J. Probst[1,3]

1: Institute for Environmental Microbiology and Biotechnology, Department of Chemistry, University Duisburg-Essen, Germany
2: Tim Burg, Im Acker 59, 56072 Koblenz, Germany
3: Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, Universitätsstraße 5, 45141 Essen, Germany

To whom the correspondence should be addressed:
alexander.probst@uni-due.de
phone: +49 (0) 201 183 7080
fax: +49 (0) 201 183 6603

Publication information:
Article was submitted to Environmental Microbiology
Submitted 25.05.2022

running title: Refining genomes from metagenomes with uBin

Abstract

Resolving bacterial and archaeal genomes from metagenomes has revolutionized our understanding of Earth's biomes yet producing high quality genomes from assembled fragments has been an ever-standing problem. While automated binning software and their combination produce prokaryotic bins in high throughput, their manual refinement has been slow, sometimes difficult or missing entirely facilitating error propagation in public databases. Here, we present uBin, a GUI-based, standalone bin refiner that runs on all major operating platforms and was additionally designed for educational purposes. When applied to the public CAMI dataset, refinement of bins was able to improve 78.9% of bins by decreasing their contamination. We also applied the bin refiner as a standalone binner to public metagenomes from the International Space Station and demonstrate the recovery of near-complete genomes, whose replication indices indicate active proliferation of microbes in Earth's lower orbit. In a worker variability test, the majority of first-time users (students) showed a significant improvement of their reconstructed genomes. uBin is an easy to install software for bin

refinement, binning of simple metagenomes and communication of metagenomic results to other scientists and in classrooms. The software and its helper scripts are open source and available under https://github.com/ProbstLab/uBin.

Originality-Significance Statement

Environmental genomics, foremost genome-resolved metagenomics, has substantially increased our knowledge regarding Earth's genetic diversity, ecosystem functioning and given raise to numerous biotechnological inventions. The basis for this research is the reconstruction of accurate genomes from mixed communities, which is, however, prone to errors, particularly for complex data. The software uBin that we present in this study has the potential to substantially decrease binning errors in reconstructed genomes and thus improve the accuracy of the predictive power of metagenomics and any downstream analysis including public databases.

Keywords

Genome-resolved metagenomics, genomics, genome curation, education, ISS, bacteria, archaea

The authors declare no competing interest. All data is publicly available.

Introduction

Genome-resolved metagenomics aims at recovering genomes from shotgun sequencing data of DNA of mixed populations. The genomes allow determination of the metabolic capacities of the individual community members and provide the basis for many downstream 'omics techniques like metatranscriptomics and metaproteomics. Since the percentage of closed genomes from complex ecosystems remains as low as 5.3% even with applying long-read sequencing (Singleton *et al.*, 2020), genomes generally need to be reconstructed ("binned") from metagenomes using genome-wide shared characteristics like their similar abundance pattern and *k*-mer frequencies (Teeling *et al.*, 2004; Albertsen *et al.*, 2013), which can be accomplished using a multitude of automatic and semi-automatic tools (Dick *et al.*, 2009; Alneberg *et al.*, 2014; Wu, Simmons and Singer, 2016; Sieber *et al.*, 2018; Kang *et al.*, 2019). The quality of the resulting bins, however, can vary greatly depending on metagenome complexity (*e.g.*, strain heterogeneity, microbial community characteristics, repetitive genomic regions and combinations thereof) (Sieber *et al.*, 2018). Recent studies have shown that contamination in genomes from metagenomes in public databases is a frequent occurrence (Ballenghien, Faivre and Galtier, 2017; Shaiber and Eren, 2019) and suggested genome curation as a mandatory analysis step prior to genome submission to public databases (Robert M Bowers *et al.*, 2017).

While established tools exist to determine the bin quality (Parks *et al.*, 2015; Sieber *et al.*, 2018), software to improve the bin quality are sparse. Some established tools are used for genome refinement (Wrighton *et al.*, 2012; Eren *et al.*, 2015) but have not been designed for educational purposes and are sometimes not open source (Wrighton *et al.*, 2012). Consequently, we developed uBin as an interactive graphical-user interface that is easy to install on Mac OS, Windows, and Ubuntu for usage in, *e.g.*, classrooms. uBin is inspired by ggKbase (Wrighton *et al.*, 2012) and enables the curation of genomes based on a combination of GC content, coverage and taxonomy and couples this to information on completeness and contamination for supervised binning. In addition, uBin can be directly used as a standalone software to bin genomes from low complexity samples.

Results and Discussion

The designed software called uBin is an interactive tool that enables live curation of metagenome-derived bins by first formatting the dataset using helper scripts (https://github.com/ProbstLab/uBin-helperscripts), followed by import of the full metagenome data (*Figure* III.2.1). We ensured uBin's compatibility with any binning software currently available (including DAS Tool (Sieber *et al.*, 2018)). After successful import, the predefined bins from automated binners can be selected and interactively improved using %GC-content, coverage information and taxonomy (*Figure* III.2.1). The process can be supervised using 51 bacterial and 38 archaeal single copy genes as used in DAS Tool (Sieber *et al.*, 2018). After successfully improving all bins in one sample, all data can be batch exported as tab-delimited file assigning scaffolds/contigs to bins or additionally as individual fasta files for prompt downstream analysis (*Figure* III.2.1).

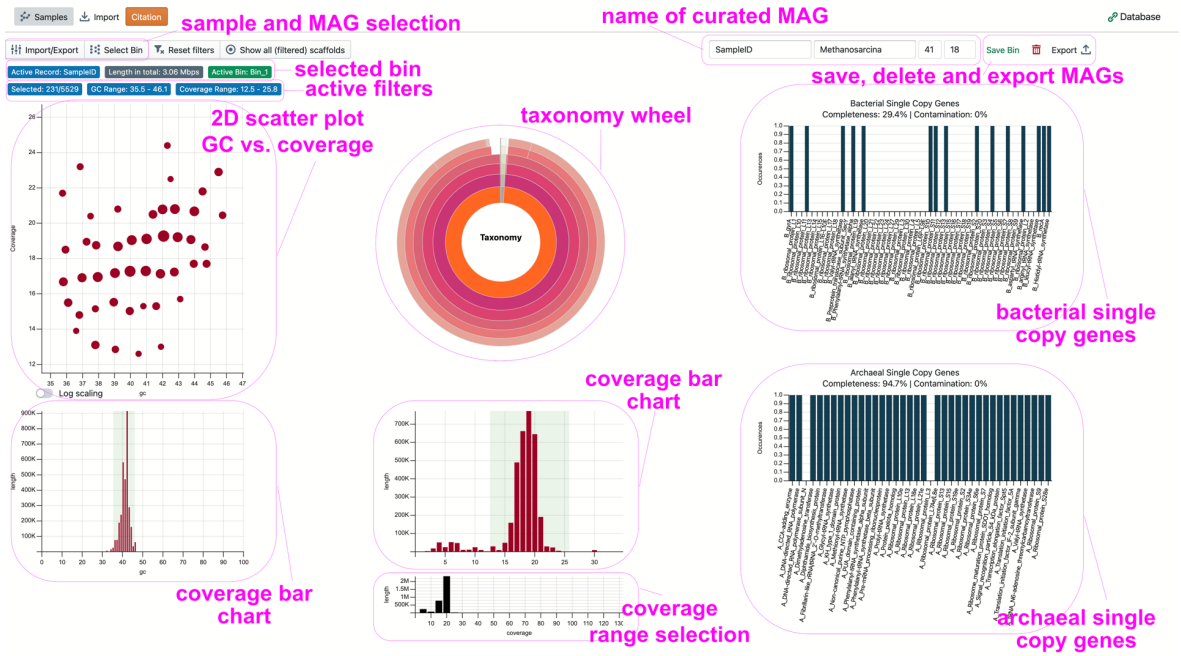

Figure III.2.1 | The bin curation interface of uBin. All plots shown in the interface are interactive, *i.e.*, selections in the scatterplot, histograms or in the taxonomy wheel modifies the availability of data in the other plots**.** Selected areas in the interface are highlighted. All selections can be reverted using the 'Reset filters' button; once bins are curated, they can be saved (or deleted). After bin curation of all bins in a dataset is completed, the curated bins can be exported in FASTA format. *Figure* S6 provides an explanation for both the bin curation interface (*i.e.*, the Samples tab) and the Import tab. A video tutorial on how to use uBin is provided at uBin_video_tutorial. The uBin software was generated using the Electron app generation toolkit (https://www.electronjs.org/).

We first tested the performance of uBin on simulated datasets with varying complexity. For this purpose we used the data of the Critical Assessment of Metagenome Interpretation (CAMI) challenge and applied four automated binners and aggregation via DAS Tool to generate bins (Sieber *et al.*, 2018). The quality of the bins before (after DAS Tool) and after curation with uBin was compared to the correct assignment based on the CAMI dataset (*Table* S1). uBin curated bins showed a highly significant quality improvement in medium ($p < 10^{-4}$) and high complexity datasets ($p < 10^{-5}$, *Figure* III.2.2A), while no significant difference could be detected for the low complexity dataset ($p > 0.70$ / 0.65).

The bin quality of the low complexity dataset was significantly higher than the bin quality in medium (0.197 higher F-score, $p < 10^{-6}$) and high complexity (0.078 higher F-score, $p < 0.10$) datasets (ANOVA coupled to TukeyHSD, $p < 2x10^{-6}$) after application of DAS Tool and prior to refinement with uBin [(Sieber *et al.*, 2018); *Figure* S1A]. Subsequent to curation with uBin the differences between these datasets were much less pronounced (ANOVA, $p < 0.01$, *Figure* S1B), leading to the conclusion that mainly high-complexity datasets can greatly benefit from manual curation.

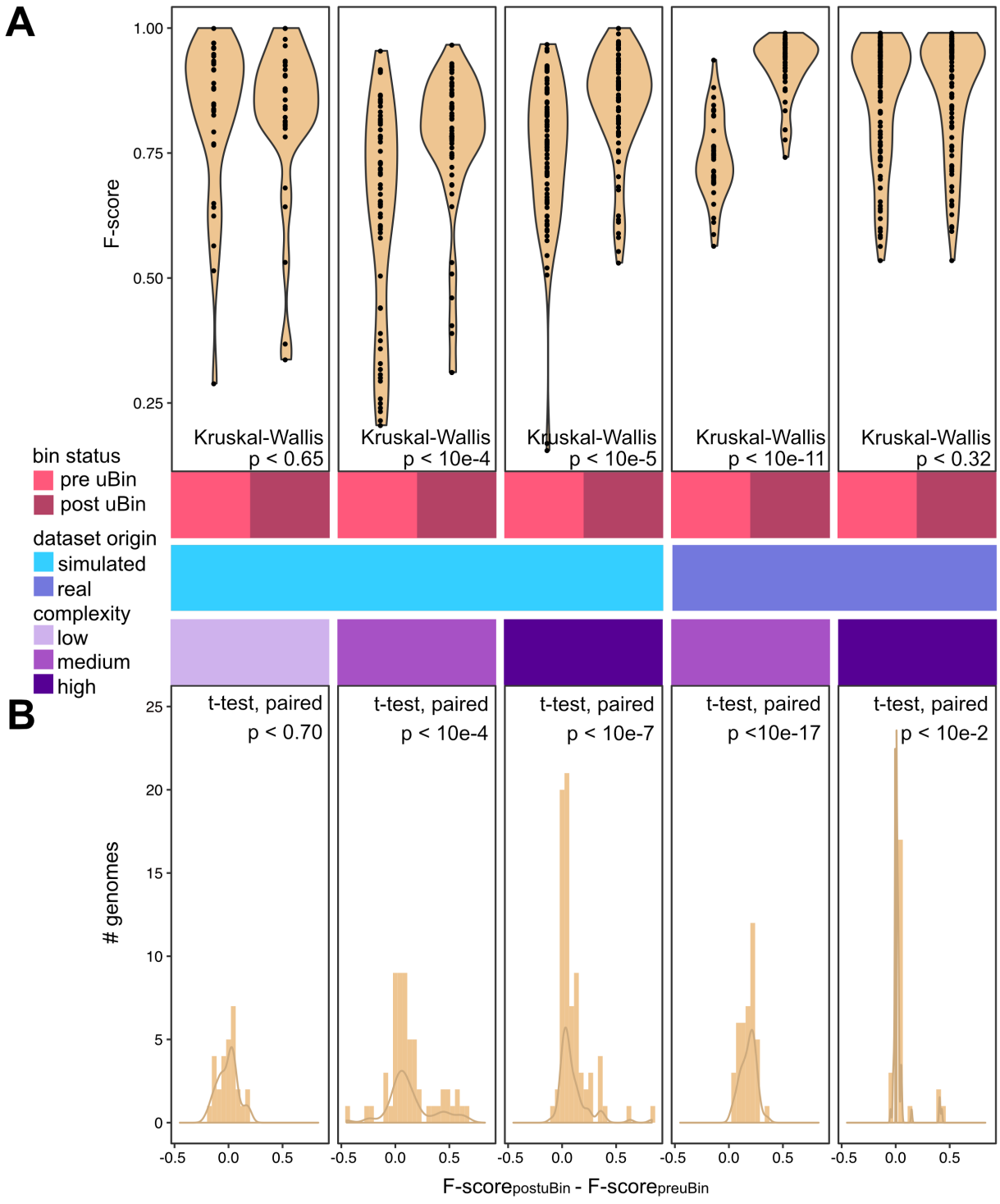

Figure III.2.2 | Performance of uBin on simulated and real datasets with varying degrees of complexity. A: Violin plots of the F-score (mean between recall and precision) of genomes prior to uBin curation (pre uBin) and after uBin curation (post uBin) across simulated low, medium and high complexity datasets of the CAMI challenge as well as real world metagenomic datasets of medium (Tomsk) and high (SulCav AS07-7) complexity. CheckM completeness and redundancy estimates were used to calculate F-scores for bins from environmental metagenomes where the true assignment of scaffolds was unknown. Unpaired Kruskal-Wallis p-values are depicted. **B:** Histograms of the F-score differences for each bin prior to and post uBin curation and their density distribution. Paired Welch t-

test p-values are shown. F-score calculation, statistics, as well as violin and histogram plot generation was done in R (R Core Team, 2008) and using ggplot2 (Wickham, 2009). The panels were assembled, and legends, statistics and titles were added in Affinity Designer v1.10.4.

Secondly, we applied uBin to subsurface metagenomes from the Tomsk aquifer (Kadnikov *et al.*, 2018) and the Acquasanta Therme (AS) cave system (Hamilton *et al.*, 2015). We used Nonpareil3 (Rodriguez-R *et al.*, 2018) to estimate their complexity, indicating that they were less diverse than the low complexity CAMI dataset (*Figure* S3A). A comparison of the number of individual *rpS3* sequences recovered from the assemblies indicated that these environmental metagenomes have a much larger number of organisms than the low complexity CAMI dataset and placed Tomsk and AS metagenomes solidly between low-mid and mid-high CAMI challenge datasets (*Table* S4). We estimated F-scores based on the independent genome quality estimation tool CheckM (Parks *et al.*, 2015) to assess MAGs from Tomsk and AS samples before and after curation with uBin. Comparison of F-scores using paired t-tests revealed that uBin curation improved the MAG quality significantly for both datasets (*Figure* III.2.2B).

We further tested uBin's capability as a standalone binner of low-complexity samples only using the metrics of %GC content, coverage, and taxonomy profile and the supervision via single copy genes. We chose the indoor environment of the International Space station (ISS) due to its low complexity, both based on rpS3 gene number and nonpareil3 diversity estimates, and the availability of public metagenomes (see *Table* S4). The results were compared to those of Emergent-Self-Organizing Maps (ESOMs), one of the semi-automated binners available for metagenomics, because this software still involves manual definition of bins by the user and can thus lead to highly accurate bins (Dick *et al.*, 2009). uBin outperformed ESOM-based binning when used as a standalone tool and also when uBin was used for curation of the bins generated via ESOMs (*Figure* III.2.3A, see Supplementary Material for details). Using uBin, we successfully reconstructed 53 genomes with at least 94 percent completeness (*Figure* III.2.3B) and only 6% or less contamination (Table S2). Their phylogenetic placement agreed in 98% of the genomes with the taxonomic classification provided by uBin (*Table* S3). The reconstruction of these 53 genomes represents an important step for space science since these are the first environmental genomes reconstructed from the ISS or associated transport flights. To investigate if the genomes are actively replicated under these conditions, we were able to calculate the *in situ* replication measure iRep (Brown *et al.*, 2016) for 43 out of 53 genomes. Across all sampling sites, the replication rates of the recovered population genomes

varied from 1.20 to 2.55, which implies an active metabolism. For instance, the lowest iRep value, which was calculated for *Methylobacterium aquaticum*, indicated that on average 20% of its sampled population was undergoing genome replication. While closely related organisms often had similar replication measures (*Figure S4*), the main discriminatory factor for varying replication indices was the sample origin, *i.e.*, the space flight from which the samples were retrieved (*Figure* III.2.3C). This result indicates community-wide shifts in replication between the different flights. The dataset also enabled the answer to a long-standing question of indoor microbiology relating to how external DNA influences the measurements of iRep values in metagenomics. Samples of the third sampled ISS flight were analyzed using both regular metagenomics as well as metagenomics following propidium monoazide (PMA) treatment, which removes external DNA fragments and enables DNA sequencing of cells with intact membranes. When comparing the iRep values of the paired samples (n=7 per group), no significant difference could be observed (paired t- and Wilcoxon-tests, *Figure* III.2.3D), although the variance of the iRep values increased tremendously after PMA treatment. Equivalence testing confirmed that there are no differences between these two sample types ($p < 0.01$). We suggest that PMA-treatment can improve the accuracy of iRep measures of environmental samples and recommend its usage where appropriate.

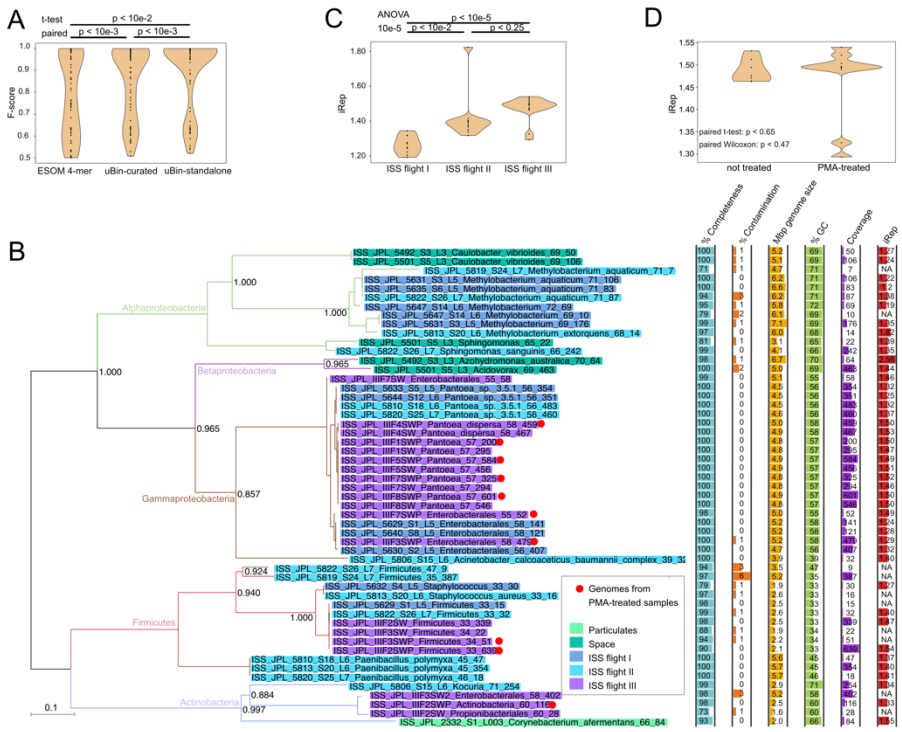

Figure III.2.3 | Reconstruction of genomes from the ISS, scoring of their curation and their phylogeny. A: Comparison of genome statistics after ESOM 4-mer binning, after uBin curation of ESOM bins, and after standalone binning using uBin. p-values correspond to paired Welch t-tests. **B:** Phylogenetic reconstruction based on the concatenation of 16 ribosomal proteins of 53 genomes from ISS metagenomes when using uBin as standalone binner. Branch colors indicate phyla assignments with coloring of leaves on tree displaying the sampling origin of the genomes. Genomes from PMA-treated samples (see main text) are highlighted with a red circle. The bargraphs on the right panel display completeness, contamination, genome size, GC content, coverage (relative abundance based on read-mapping) and the *in situ* replication measure (iRep (Brown *et al.*, 2016)). **C:** Replication index dependency on flight of origin and significance testing thereof using ANOVA followed by TukeyHSD. **D:** Effect of PMA-treatment for removal of extracellular DNA on iRep of genomes from PMA-treated samples having increased iRep variance but no significant differences in iRep value based on paired Wilcoxon and paired t-tests (n=7 per group). Genomes were paired based on sample ID as well as using their shared uBin-taxonomy and GC content. Plots presented in panels A, C and D were generated in R (R Core Team, 2008) with ggplot2 (Wickham, 2009), with statistic annotation and axis labels being added in Affinity designer during Figure assembly. The phylogenomic tree presented in panel B was generated using Fasttree 2.1.8 (Price, Dehal and Arkin, 2009) from a supermatrix of 16 ribosomal proteins and visualized in Dendroscope 3.7.2 (Huson *et al.*, 2007).

The herein presented uBin software is designed for improvement of bins and as a standalone binner for simple metagenomes with few species. It is independent of the operating system (available for Windows, MacOS, Linux) and GUI-based so that a wide audience of non-bioinformaticians can make use of it. The initial data processing (as general metagenomic data processing) necessitates bioinformatics knowledge, but respective easy-to-use wrapper scripts are provided along with the software. Thus, uBin is ideally used by bioinformaticians

to communicate metagenomic data to non-bioinformatics peers and to students in classrooms (for worker variability between students please see *Figure* S5; please see *Figure* S7 for survey results of students after using uBin in classrooms). The curated genomes can be further explored for metabolic analyses with, *e.g.*, MAGE (Vallenet *et al.*, 2009) or KEGG mapper (Kanehisa and Sato, 2020). Consequently, uBin represents an important software link between automated binners along with the widely-used software DAS Tool and downstream analyses including genome refinement to completion (Moss, Maghini and Bhatt, 2020).

Acknowledgments

This study was funded by the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen ("Nachwuchsgruppe Dr. Alexander Probst"). We thank the students who tested and worked with uBin over the last two years in classrooms. We thank Christine Sun for her contribution to the script for calculating consensus taxonomy of scaffolds and Kasthuri Venkateswaran for input regarding sampling locations of the ISS samples. We acknowledge the students who participated in testing worker variability of uBin. We thank Ken L. Dreger for the administration and maintenance of our servers.

Supplemental Information

Supplemental information is supplied on the accompanying CD (see section VII Content of supporting CD for more information).

3. Genetic diversity in terrestrial subsurface ecosystems impacted by geological degassing

Till L.V. Bornemann[1], Panagiotis S. Adam[1], Victoria Turzynski[1], Ulrich Schreiber[2], Perla Abigail Figueroa-Gonzalez[1], Janina Rahlff[1#], Daniel Köster[3], Torsten C. Schmidt[3], Ralf Schunk[4], Bernhard Krauthausen[5], and Alexander J. Probst[1,6*]

1: Environmental Microbiology and Biotechnology, Faculty of Chemistry, University Duisburg-Essen, Germany
2: Department of Geology, University Duisburg-Essen, Germany
3: Instrumental Analytical Chemistry and Centre for Water and Environmental Research (ZWU), University of Duisburg-Essen, Germany
4: Geyser-Center, Andernach, Germany
5: Institute of Applied Geosciences, Karlsruhe Institute of Technology, Germany
6: Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, Universitätsstraße 5, 45141, Essen, Germany
[#]Present address: Centre for Ecology and Evolution in Microbial Model Systems (EEMiS), Department of Biology and Environmental Science, Linneaus University, Kalmar, Sweden

*To whom the correspondence should be addressed:
Alexander J. Probst
Environmental Microbiology and Biotechnology, Faculty of Chemistry, University of Duisburg-Essen
alexander.probst@uni-due.de

Publication information:
Nat Commun (2022) **13**: 284. https://doi.org/10.1038/s41467-021-27783-7
Received 25.03.2020; accepted 02.12.2021; published 12.01.2022
Link: https://www.nature.com/articles/s41467-021-27783-7

Abstract

Earth's mantle releases 38.7 ± 2.9 Tg/yr CO_2 along with other reduced and oxidized gases to the atmosphere shaping microbial metabolism at volcanic sites across the globe, yet little is known about its impact on microbial life under non-thermal conditions. Here, we perform comparative metagenomics coupled to geochemical measurements of deep subsurface fluids from a cold-water geyser driven by mantle degassing. Key organisms belonging to uncultivated *Candidatus* Altiarchaeum show a global biogeographic pattern and site-specific adaptations shaped by gene loss and inter-kingdom horizontal gene transfer. Comparison of the geyser community to 16 other publicly available deep subsurface sites demonstrate a conservation of chemolithoautotrophic metabolism across sites. In silico replication measures suggest a linear relationship of bacterial replication with ecosystems depth with the exception of impacted sites, which show near surface characteristics. Our results suggest that subsurface ecosystems affected by geological degassing are hotspots for microbial life in the deep biosphere.

Introduction

The continental subsurface is a huge reservoir for life, hosting about 60% of all microorganisms on Earth (Magnabosco *et al.*, 2018; Flemming and Wuertz, 2019). Carbon, nitrogen and sulfur turnover by these microorganisms have a vast contribution to all biogeochemical cycles on the planet (Falkowski, Fenchel and Delong, 2008). In addition to the great number of microorganisms, subsurface ecosystems can accommodate a large diversity of different bacteria and archaea (Castelle *et al.*, 2015; Anantharaman *et al.*, 2016; Probst *et al.*, 2018), with even single ecosystems containing representatives of almost all known bacterial phyla (Anantharaman *et al.*, 2016). Subsurface ecosystems are categorized as either detrital or productive, depending on whether buried organic carbon or inorganic carbon are the main carbon sources of the community (Stevens, 1997). Since no light is available as an energy source in the deep biosphere, alternative electron donors to water like hydrogen (H_2) or sulfide (H_2S) are used to fuel mostly anaerobic carbon fixation pathways such as the Wood-Ljungdahl pathway (Stevens, 1997). Subsurface lithoautotrophic microbial communities (Stevens and McKinley, 2000) have been reported for many terrestrial ecosystems including the Fennoscandian Shield (Nyyssönen *et al.*, 2014), the Columbia River Basalt (Stevens and McKinley, 2000), the Witwatersrand Basin (Lau *et al.*, 2016), and subsurface fluids discharged by Crystal Geyser (Probst *et al.*, 2017). While these subsurface ecosystems are usually

dominated by bacteria, one exception are archaea belonging to the Alti-1 clade of the *Ca.* Altiarchaeota (Probst *et al.*, 2014, 2018; Hernsdorf *et al.*, 2017) Alti-1 form biofilms using their characteristic nano-grappling hooks (hami) (Moissl *et al.*, 2005; Bird *et al.*, 2016). The other clade, Alti-2, are more widespread and diverse but found at lower abundances in their ecosystems (Bird *et al.*, 2016). *Ca.* Altiarchaeota live autotrophically using the Wood-Ljungdahl carbon fixation pathway (Wood, 1991), which was the most dominant carbon fixation pathway prior to the evolution of photosynthesis (Gutiérrez-Preciado *et al.*, 2018; Adam, Borrel and Gribaldo, 2019).

Chemolithoautotrophic life in subsurface ecosystems necessitates the presence of adequate electron donors like hydrogen, hydrogen sulfide or methane. One source of such gases can be Earth's mantle, which also releases 38.7 ± 2.9 Tg/yr of oxidized carbon (Aiuppa *et al.*, 2019), mainly in form of carbon dioxide (CO_2), into the crust and the atmosphere (Bräuer *et al.*, 2013; Werner *et al.*, 2019). This process, also termed mantle degassing, is the transition of volatiles from the mantle (supercritical) to the subcritical zone of the upper crust fueled by lower pressure of volatiles near the surface compared to the mantle (Zhang, 2014). Modern Earth has few areas with active mantle degassing, which are usually restricted to terrestrial volcanoes, subduction zones or hydrothermal vents in oceans (Caracausi and Paternoster, 2015; Loreto *et al.*, 2015; Fullerton *et al.*, 2019; Gilfillan *et al.*, 2019; Lee *et al.*, 2019). At hydrothermal vents, chemolithoautotrophs initiate the microbial trophic network and proliferate at high rates leading to high microbial cell numbers(Magnabosco *et al.*, 2018; Adam, Borrel and Gribaldo, 2019; Aiuppa *et al.*, 2019). While volcanic sites and hydrothermal vent fields have been studied fairly thoroughly regarding both their microbial community composition and activity(Hedrick *et al.*, 1992; Schrenk, Holden and Baross, 2008; Ding *et al.*, 2017; Tu *et al.*, 2017; Galambos *et al.*, 2019), little is known about deep subsurface ecosystems with low temperatures (283-293 K) and still impacted by gases released from the mantle.

Previous studies have analyzed the influence of mantle degassing via volcanic mofettes, *i.e.*, CO_2 seeps below 373 K, on near surface biomes, particularly soil microbial communities (Frerichs *et al.*, 2013; Mehlhorn *et al.*, 2014; Beulig *et al.*, 2015, 2016). (Mehlhorn *et al.*, 2014) showed that gases from the mantle can alter the availability of different heavy metals including metalloid arsenic and predicted impacts on microbial communities. Beulig and co-workers reported an increase in dark carbon fixation and found evidence that the CO_2 from the degassing is indeed incorporated into biomass based on IR-GC/MS measurements of fatty acid methyl-esters and DNA Stable-Isotope Probing experiments of microcosms fed with ^{13}C-labelled CO_2

(Beulig *et al.*, 2015, 2016). Along with fermentation processes, the pathways for the turnover of organic carbon were similar in both systems, while the microbial diversity of soils in mofettes was lower compared to controls. Carbon and sulfate respiration were enriched during degassing, while aerobic respiration declined (Beulig *et al.*, 2016), and acetogenesis was suggested to play a major role in these systems (Beulig *et al.*, 2015). However, these studies were limited to the upper 50-cm of Earth's critical zone, and the influence of mantle degassing on mesophilic microbial communities in the deep subsurface including their metabolic capacity and activity has not been investigated so far.

The cold-water (291 K) Geyser Andernach is located in the Rhine Valley near Koblenz in western Germany and is driven by gases discharged from the mantle (Bräuer *et al.*, 2013). Since 2001 the geyser has had an intact tubing, thus tapping into a unique ecosystem. Once released by a mechanical shutter, the gases from the mantle (mainly CO_2) permeating the groundwater cause the eruption of cold subsurface fluids sourced from an uniform aquifer system. Thus, Geyser Andernach is an ideal ecosystem to investigate how mantle degassing shapes mesophilic microbial life in the subsurface.

Here, we used a combination of long-term geochemical characterization coupled to genome-resolved metagenomics to investigate the geyser's microbial community. To analyze how mantle degassing impacts mesophilic microbial communities, we set the bacterial replication index values, minimal generation times and microbial metabolism abundances in Geyser Andernach into relation to 16 other deep continental subsurface ecosystems across the globe. We identified a pattern of decreasing replication indices but shorter minimal generation times with increasing depth. Sites impacted by mantle degassing showed similar replication indices and generation times as near-surface sites, rendering them hotspots for microbial activity in the subsurface. Comparative genomics applied to a key player at sites impacted by geological degassing (*Ca.* Altiarchaeum sp.), revealed that the slow evolutionary rate present in this phylum might be counteracted by horizontal gene transfer (HGT) and gene loss events in this organism group.

Results

Geyser Andernach provides access to a stable ecosystem impacted by mantle degassing

Geyser Andernach was drilled to a depth of 351 m in 1903 tapping into a shale-hosted aquifer with quartz veins. Its eruptions are driven by mantle degassing and can be controlled via mechanical shutters (a diagram of the plumbing system is provided in *Supplementary Figure*

1). Geochemical measurements averaged over 14 years have demonstrated that the subsurface fluids provide a constant environment (*Supplementary Table* 1). The gaseous and ionic composition of the geyser showed the predominance of CO_2 in the system and previously reported traces of hydrogen and hydrogen sulfide (Bräuer *et al.*, 2013). Prominent electron donors and acceptors were determined to be hydrogen and ferric iron as well as sulfate, respectively. To investigate the microbial community in subsurface fluids impacted by mantle degassing, we sampled two eruptions of Geyser Andernach, and collected the planktonic fraction of microorganisms onto three individual 0.1-µm filters. Metagenomic sequencing of the community resulted in ~7 billion bp per sample (5% SD), covering about 80% of the microbial diversity as estimated by Nonpareil3 [(Rodriguez-R *et al.*, 2018); *Supplementary Figure* 2]. Reads were assembled into 921,520 scaffolds on average (20% SD, for further statistics please see *Supplementary Table* 2). Approximately 75% of the reads (2.6% SD) mapped back the assembly providing evidence that the reconstructed metagenome is representative for the planktonic community at the time of sampling. The community composition based on ribosomal protein S3 (*rpS3*) sequences assembled from the metagenome displayed a fairly restricted diversity consisting of 52 organisms, which spanned twelve phyla (*Figure* III.3.1). The core community was composed of 15 organisms detected via *rpS3* across all three metagenomes (*Figure* III.3.1), and they accounted for 42.8% (1.3% SD) of the total relative abundance of the community. For 20 of these 52 microorganisms, we reconstructed high quality genomes with at least 70% estimated completeness (and less than 10% estimated contamination, details in *Supplementary Data* 1). The most abundant species recruited 42.8% (1.3% SD) of the metagenomic reads and belonged to the *Ca.* phylum Altiarchaeota (Probst *et al.*, 2018) (in the following denoted as *Ca.* Altiarchaeum GA) and specifically grouped within the Alti-1 (Bird *et al.*, 2016) clade. The second most abundant organism was classified as Caldiserica, which were originally known to inhabit hot springs (Mori *et al.*, 2009) but were recently also detected in subsurface ecosystems populated by mesophiles (Probst *et al.*, 2017, 2018).

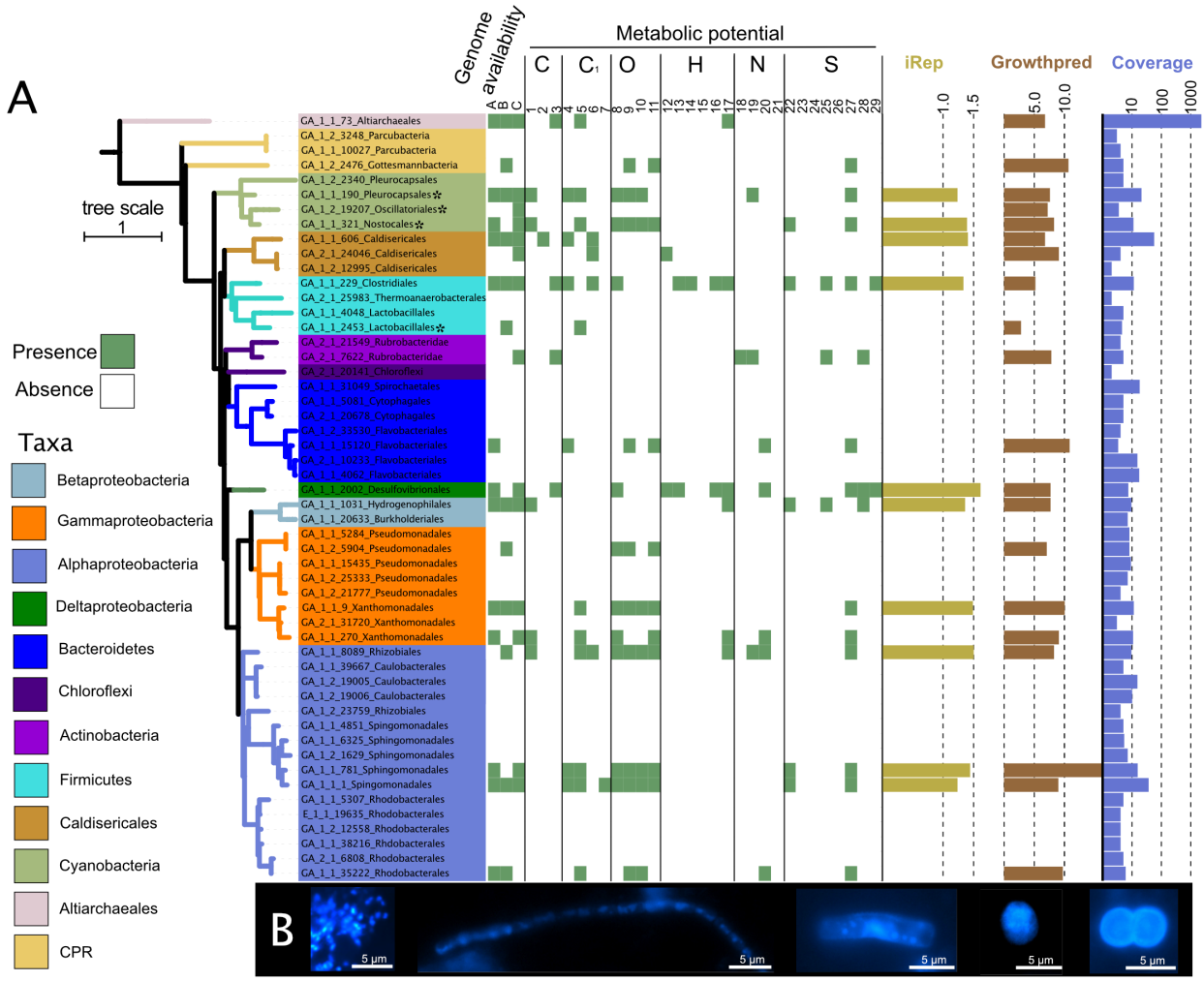

Figure III.3.1 | Metagenomic and microscopic characterization of the community in subsurface fluids discharged by Geyser Andernach. A: RpS3-based phylogenetic diversity of the organisms in the Geyser Andernach. Centroid rpS3 sequences (after clustering at 99% similarity using cdhit) were used for the calculation of the phylogenetic tree using IQTree. Colors of the different branches signify different phyla. Matching recovered draft genomes in each sample (A, B and C for samples GA_E1-1, GA_E1-2 and GA_E2-1 respectively), *i.e.*, genomes binned from these samples, are provided as green boxes (otherwise left white). The presence of marker genes based on a marker gene search using HMMs on these genomes for specific chemolithoautotrophic pathways is shown as green boxes (otherwise left white). C signifies carbon fixation with 1) CBB, 2) rTCA and 3) WL, C_1 for C_1-metabolism with 4) carbon monoxide oxidation, 5) Formaldehyde oxidation and 6) methanol oxidation, O for oxygen metabolism with 7) cytochrome c bd, 8) cytochrome c bo, 9) cytochrome c caa_3 and 10) cytochrome cbb_3, H for hydrogen metabolism with 12) FeFe-Hydrogenases type A, 13) NiFe-Hydrogenases type 3b, 14) NiFe-Hydrogenases type 3c, 15) Nife-Hydrogenases, 16) NiFe-Hydrogenases type 4 and 17) NiFe-Hydrogenases type 1, N for nitrogen metabolism with 18) Nitrate reduction, 19) Nitric oxide reduction, 20) nitrite reduction and 21) nitrous oxide reduction, S for sulfur metabolism with 22) sulfide Oxidation, 23) sulfite reduction with dsr, 24) sulfite reduction with asr, 25) sulfur oxidation with dsr, 26) sulfur oxidation with sor, 27) sulfur oxidation with sdo, 28) sulfate reduction via APS with sat and 29) Thiosulfate disproportionation. Olive bars show the average iRep value of the respective bacterial population, brown bars show the maximal growth rate of the representative genome as estimated by growthpred and blue bars show the average log10-scaled coverage. **B:** Morphologies of microorganisms as determined via DAPI staining and fluorescence microscopy (scale bars = 5 µm) are shown. The

morphologies were documented in two sampling campaigns (June 2016 and February 2018 with three and two samples in technical duplicates, respectively).

We verified that bacteria in this community were replicating at the time of sampling using in situ replication index values. Replication index values are calculated from the difference of sequencing coverage between the origin of replication and terminus of replication. Proliferating organisms replicate their genomes with multiple replication forks starting at the replication origin and thus contributing more to sequencing reads. In our study, these index values ranged between 1.4-1.5, indicating that 40-50 % of those microbial populations, whose iRep values were calculated, underwent genome replication at the time of sampling. Microscopic cell counts of organisms from the subsurface fluids ranged from 2.7 x 10^6 to 4.2 x 10^6 (average 3.5 x 10^6) cells ml^{-1} (*Supplementary Figure* 3) and displayed various morphologies ranging from cocci and rods to filamentous-shaped microorganisms (*Figure* III.3.1). Importantly, we also observed clusters of small cocci, which are similar to previously reported biofilm structures of *Ca.* Altiarchaeota (Probst *et al.*, 2014) and whose presence was confirmed by metagenomic results. We estimated the total amount of erupting carbon (CO_2 and hydrogen carbonate (HCO_3^-)) to be 6,270 kg per year, while the microbial cells account for approximately 111.5 g of carbon, suggesting that about 0.0018% of carbon degassing from the mantle is fixed in this ecosystem.

Replication index values and maximal growth rates across multiple deep continental subsurface ecosystems

To investigate if mantle degassing has an impact on microbial replication in the continental subsurface, we used in situ replication index values (iRep) of bacterial genomes and maximal growth rate estimates of bacterial and archaeal genomes. We first investigated if iRep can be used as a measure of replication by comparing groundwater fluids to sediments, because microbes in sediments are known to be more active (Kairesalo *et al.*, 1995). Indeed, iRep suggested a significantly higher replication of microbes in sediments than groundwater (p-value < 10^{-3}). Replication measures from Geyser Andernach were then compared with those from other public datasets from deep subsurface environments of varying depth (overview of samples and ecosystems is provided in *Supplementary Table* 4). The sampling depth varied from 0 m below ground (cave systems) to 3140 m depth. We reconstructed genomes of previously unbinned metagenomes resulting in 560 newly assembled and classified

prokaryotes (*Supplementary Data* 1) representing 415 different organisms after dereplication. Combined with genomes and iRep results from previous studies (Anantharaman *et al.*, 2016; Hernsdorf *et al.*, 2017; Probst *et al.*, 2018), we leveraged in situ replication measures for 895 bacteria (*Supplementary Data* 2) spanning the vast majority of all known bacterial phyla (see *Supplementary Data* 5). The average iRep value of bacteria of the individual ecosystems correlated negatively and highly significantly with sample depth across all individual iRep values (Pearson's test, p-value < 10^{-8}) and across median per sampled ecosystem (p-value < 0.0007, *Figure* III.3.2, *Supplementary Table* 5). In other words, the deeper the origin of the retrieved sample, the lower the genome replication measure.

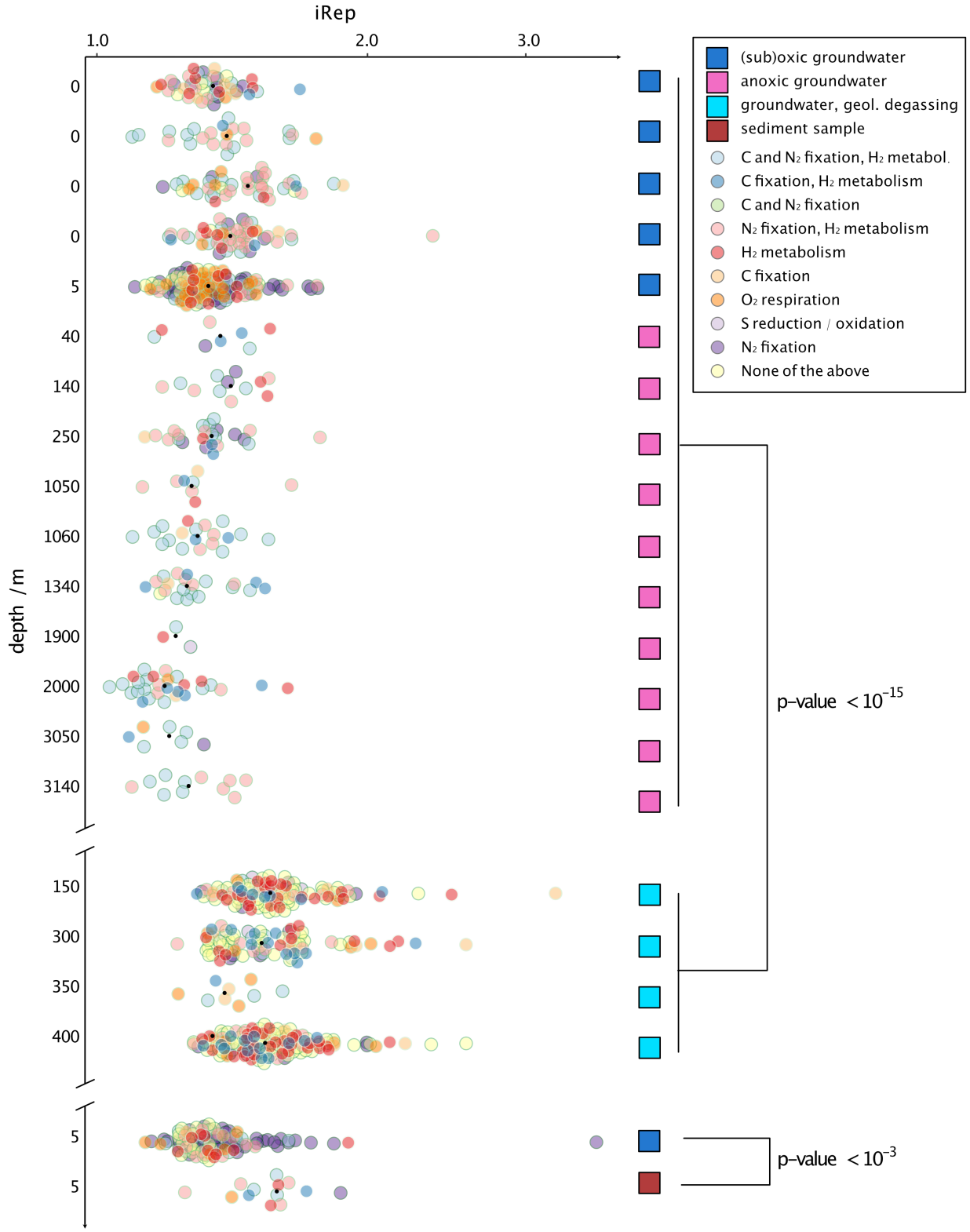

Figure III.3.2 | In situ bacterial replication rates across subsurface ecosystems ordered by ecosystem depth. The figure depicts a beeswarm plot of iRep values of genomes (x-axis) across ecosystems (y-axis) with genomes colored according to their predicted metabolic potential and the black dot representing the median iRep value (individual iRep values in *Supplementary Data* 2). C represents carbon, N_2 nitrogen, H_2 hydrogen, O_2 oxygen and S sulfur. Colored squares depict the sample type.

Samples impacted by geological degassing and a sediment sample along with the respective aquifer sample are plotted separately. The top y-axis shows the sampling depth of the different ecosystems (*Supplementary Table* 5). In total, 895 genomes were used for this analysis with >=70% completeness and <=10% contamination based on 51 bacterial and 38 archaeal single copy genes. The order of samples is given in *Supplementary Table* 5. P-values are derived from two-sided student's t-tests. The exact p-values from top to bottom are p < 2.2x10^16 (minimal value in R) and p=0.0003934, respectively.

In particular, organisms with the capacity of carbon fixation (cor=-0.47), of sulfur oxidation (cor=-0.46) or of metabolizing hydrogen (cor=-0.45) contributed to this observation (correlations are summarized in *Supplementary Table* 6). Samples impacted by high CO_2 concentrations, either solely from mantle degassing (this study) or from both mantle degassing and thermal activity(Probst *et al.*, 2018), were outliers in this correlation analysis. In fact, iRep measures of bacteria in these samples were significantly higher than iRep measures of other subsurface samples (p-value < 10^{-15}) and nearly reached values of samples that are close to Earth's surface (*Figure* III.3.2). When excluding these samples from the correlation analysis with depth, the respective correlation coefficient decreased from -0.20 to -0.28 (p-value < 10^{-8}). We also tested how the availability of oxygen influences genome replication measures of bacteria in the continental subsurface. iRep values were on average 0.09 higher for bacteria in oxygenic samples (p-value < 10^{-8}) meaning that about 9% more of the bacteria were undergoing genome replication.

While iRep values indicated that there is less ongoing replication in deeper regions of the subsurface, they do not allow any inference about the speed at which organisms are replicating. Thus, we also calculated maximal possible growth rates, *i.e.*, minimal generation times, based on the codon usage bias between constitutionally expressed ribosomal proteins and the rest of the genes per genome using growthpred (Vieira-Silva and Rocha, 2010). Correlation analyses of these maximal growth rates with the sampling depth revealed that the maximally possible replication speed increases, *i.e.*, shorter doubling times, with increasing depth (p < 0.0011, cor = -0.143, *Supplementary Figure* 4).

Conserved chemolithoautotrophic metabolism of subsurface microbial communities

Since bacterial replication is predicted to differ between sites impacted by mantle degassing and reference sets, we investigated if the general metabolism for carbon, nitrogen and sulfur turnover of entire communities is adapted to high-CO_2 subsurface environments. We searched for key enzymes for metabolic pathways across our entire metagenomic assemblies (*Supplementary Table* 2) and used the abundance of scaffolds that carried a key enzyme as relative abundance measure of the respective metabolism (*Figure* III.3.3, *Supplementary Figure* 5, *Supplementary Figure* 6). The core metabolism remained relatively stable across all tested ecosystems. We performed both Student's t-tests and Kruskal-Wallis-tests along with equivalence testing to determine whether there was a significant difference between high-CO_2 and non-high-CO_2 metabolisms and could only detect a significant difference in the nitrite reduction metabolism (Kruskal-Wallis group comparison, p-value = 6×10^{-4}, details on tests in *Supplementary Table* 7). Consequently, and in congruence with previous studies investigating the metabolic diversity in a subseafloor aquifer (Tully *et al.*, 2017), little difference exists in the metabolic potential between regular subsurface microbial communities and those at sites impacted by mantle degassing, although the indigenous organisms at these sites appear to have higher replication index values.

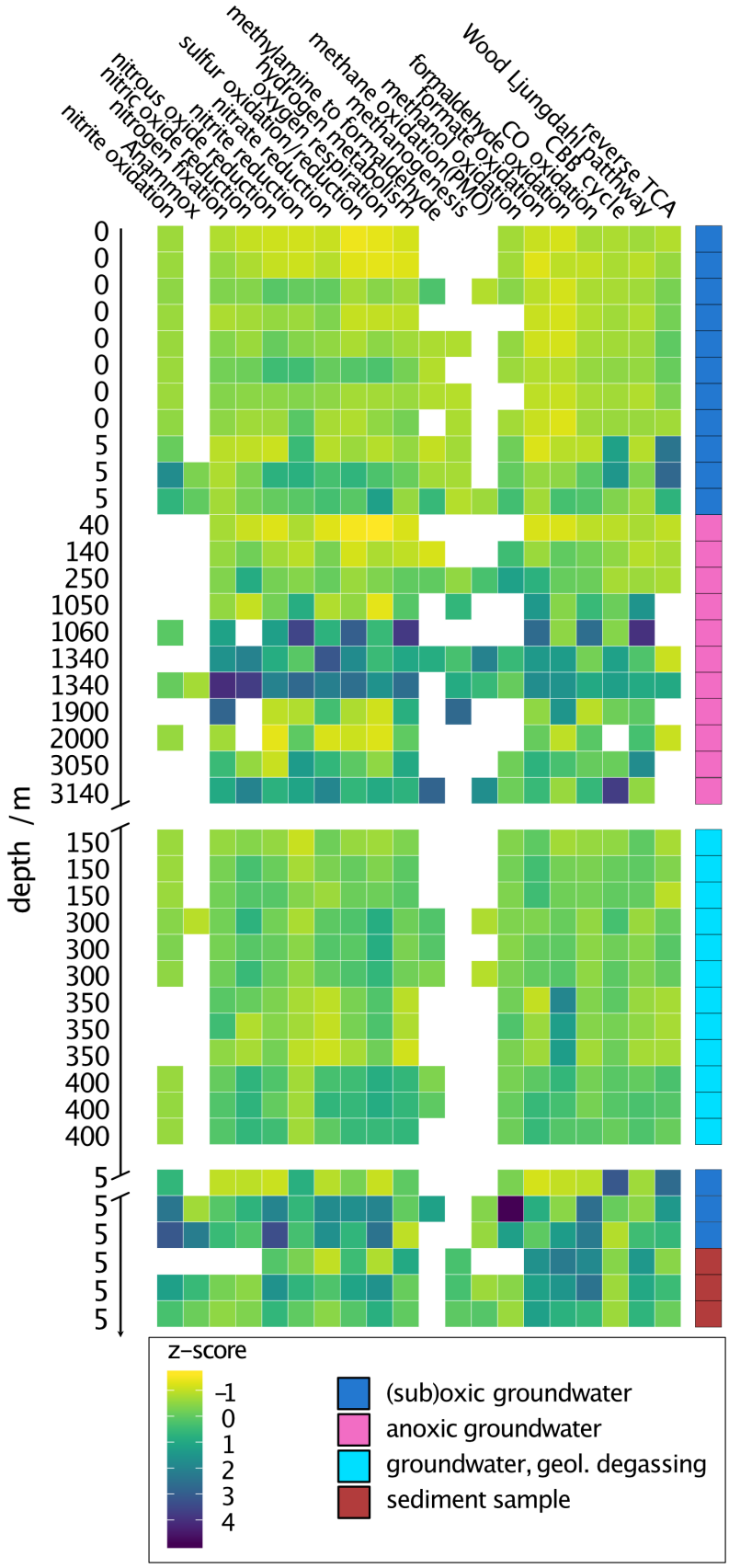
nitrite oxidation
Anammox
nitrogen fixation
nitric oxide reduction
nitrous oxide reduction
nitrite reduction
nitrate reduction
sulfur oxidation/reduction
oxygen respiration
hydrogen metabolism
methylamine to formaldehyde
methane oxidation
methanogenesis
methanol oxidation(PMO)
formate oxidation
formaldehyde oxidation
CO oxidation
CBB cycle
reverse TCA
Wood Ljungdahl pathway
depth /m
0
0
0
0
0
0
0
0
5
5
5
40
140
250
1050
1060
1340
1340
1900
2000
3050
3140
150
150
150
300
300
300
350
350
350
400
400
400
5
5
5
5
5
5
z-score
−1
0
1
2
3
4
(sub)oxic groundwater
anoxic groundwater
groundwater, geol. degassing
sediment sample

Figure III.3.3 | Chemolithoautotrophic metabolic potential across ecosystems. The heatmap shows the read-normalized abundance of chemolithoautotrophic pathways, Z-score scaled for the respective metabolisms. Colored squares on the right depict the sample type. If multiple biological replicates of samples were available, up to three were depicted. Sample order is according to *Supplementary Table* 5. *Supplementary Figure* 5 and *Supplementary Figure* 6 display the Z-scaled number of hits (*Supplementary Figure* 5) or normalized abundance (*Supplementary Figure* 6) of the individual genes aggregated into their pathways in this figure.

Biogeography and functional adaptations of deep subsurface Altiarchaeota

Key organisms in continental subsurface ecosystems impacted by geological degassing belong to the *Ca.* phylum Altiarchaeota due to their high abundance. *Ca.* Altiarchaeota can currently be divided into two clusters, Alti-1 and Alti-2, with the latter having a broader metabolic variability than Alti-1 (Bird *et al.*, 2016). In the following, we are going to refer to Alti-1 Altiarchaeota as *Ca.* Altiarchaea. However, organisms of the *Ca.* Altiarchaea are those that can dominate entire ecosystems, as shown for multiple sites across the globe (Probst *et al.*, 2014, 2018; Hernsdorf *et al.*, 2017). Nearly all of the ecosystems dominated by *Ca.* Altiarchaea have all been reported to have high CO_2 partial pressure or great amounts of carbonate deposits (Probst *et al.*, 2013). The average nucleotide (ANI) and amino acid (AAI) identity of all so-far recovered *Ca.* Altiarchaea genomes indicated that they belong to the same genus (*Supplementary Figure* 7), although 16S ribosomal RNA gene similarity suggested the same species. When correlating the genomic differences based on ANI to the geographical distance between sampling sites of the *Ca.* Altiarchaea genomes, a highly significant negative correlation (*Pearson*, cor = -0.77, $p = 9 \times 10^{-4}$) could be observed, indicating that a greater distance led to greater dissimilarity (*Supplementary Figure* 7). We challenged this observation by using robust phylogenetic analyses based on a supermatrix of 30 ribosomal proteins and found that *Ca.* Altiarchaea cluster based on geographical sampling site going all the way to continent scale (*Figure* III.3.4C, *Supplementary Figure* 8). However, we did not observe any biogeographic pattern for *Ca.* Altiarchaeota of the Alti-2 clade, which mainly occur in ocean sediments (Bird *et al.*, 2016). Based on Hidden Markov Model (HMM) profiles of key chemolithoautotrophic genes of Alti-2 and Alti-1 genomes, some of which we newly reconstructed from public datasets, we identified substantial differences particularly in the hydrogen metabolism (*Figure* III.3.4B, details on *Ca.* Altiarchaeota genomes in *Supplementary Table* 3). However, Alti-2 showed a significantly smaller minimal generation time than Alti-1 (U-test $p < 0.0024$; *Supplementary Figure* 9).

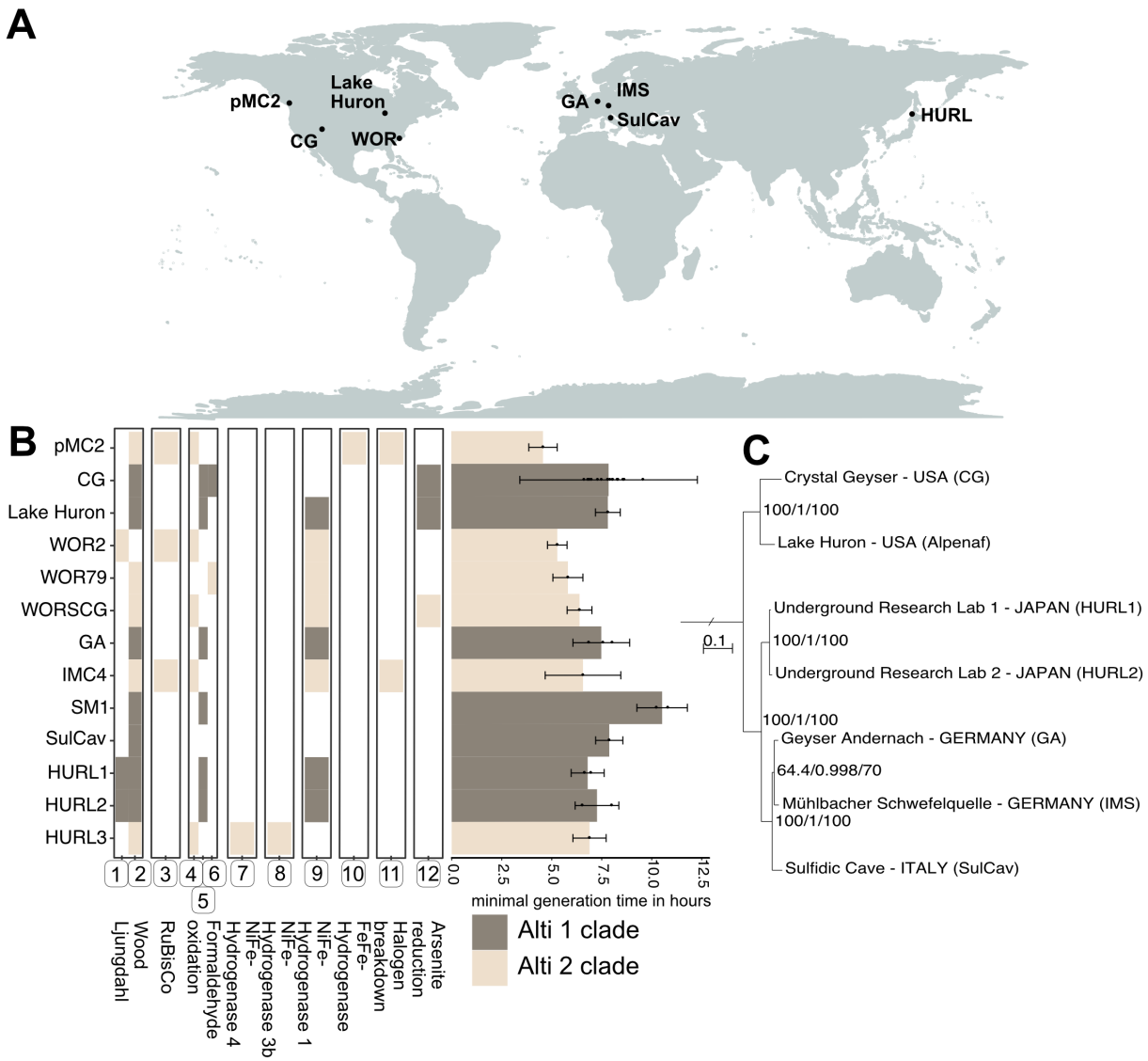

Figure III.3.4 | Geographical distribution and chemolithoautotrophic potential of *Ca.* Altiarchaeota. A: Global map with locations from which *Ca.* Altiarchaeota genomes were recovered. **B:** Metabolic potential of *Ca.* Altiarchaeota genomes. Genomes belonging to the Alti-1 clade are highlighted in dark grey, Alti-2 genomes in beige. If multiple genomes from a specific site were available, they were all used to identify the metabolic potential. The bar chart shows averaged growthpred-predicted minimal generation times across all genotypes recovered from a specific genome, with error bars denoting the averaged standard deviations (growthpred returns both an average minimal generation time and a standard deviation for this value). Additionally, the mean minimal generation time for each genome is indicated by black dots. The circled numbers below the heatmap depict the genes identified as markers and stand for: 1) codhC, 2) codhD, 3) rubisco form III, 4) fae, 5) fmtf, 6) mtmc, 7) NiFe-Hydrogenase group 4, 8) NiFe-Hydrogenase group 3b, 9) NiFe-Hydrogenase group 1, 10) FeFe-Hydrogenase, 11) hdh, 12) ars. **C:** Phylogeny of Alti-1 genotypes based on 30 universal ribosomal proteins (5136 aa positions, IQTree JTTDCMut+F+G4) and using the Alti-2 genome IMC4 as the outgroup. Branch supports correspond to ultrafast bootstraps (Hoang *et al.*, 2018) (1000 replicates), the SH-aLRT test (Guindon *et al.*, 2010) (1000 replicates), and the approximate Bayes test (Anisimova *et al.*, 2011) respectively (tree with outgroup in *Supplementary Figure* 8). Details on Altiarchaeales genomes in *Supplementary Table* 3.

Since *Ca.* Altiarchaea showed a strict biogeographic pattern, we further investigated their differences in metabolic capacities in depth using a genome model published previously (Probst *et al.*, 2014) (*Figure* III.3.5). We identified that all *Ca.* Altiarchaea share a central NAD(P)H-based Wood-Ljungdahl pathway for carbon fixation and carbon monoxide utilization. The main difference of *Ca.* Altiarchaeum GA to the reference genome *Ca.* Altiarchaeum hamiconexum (Probst *et al.*, 2014) was the presence of genes for a NiFe hydrogenase (*Figure* III.3.4B), which seems to be a specific adaptation to hydrogen containing gases from the mantle. Indeed, we identified that this NiFe hydrogenase existed in multiple other *Ca.* Altiarchaea and was lost in *Ca.* Altiarchaeum hamiconexum from IMS. The phylogenetic relatedness revealed that NiFe-hydrogenases of Alti-1 were sister to those of Alti-2 suggesting a conservation of this key enzyme in their last common ancestor (tree is provided in *Supplementary Data* 7). Other genes affected by gene loss across *Ca.* Altiarchaea encoded for proteins, which function as mechanosensitive channels, desulfoferredoxin, polysaccharide biosynthesis enzymes and some peptidases and glycosylhydrolases (*Supplementary Data* 8-15). By contrast, rubyerythrine and multiple peptidases spanning the families C44 (precursor of amidophosphoribosyltransferase), M06 (metalloendopeptidases), and C01b (endo- and exo-peptidases) were horizontally acquired by *Ca.* Altiarchaea species, mostly from the bacterial domain (*Supplementary Data* 16-19).

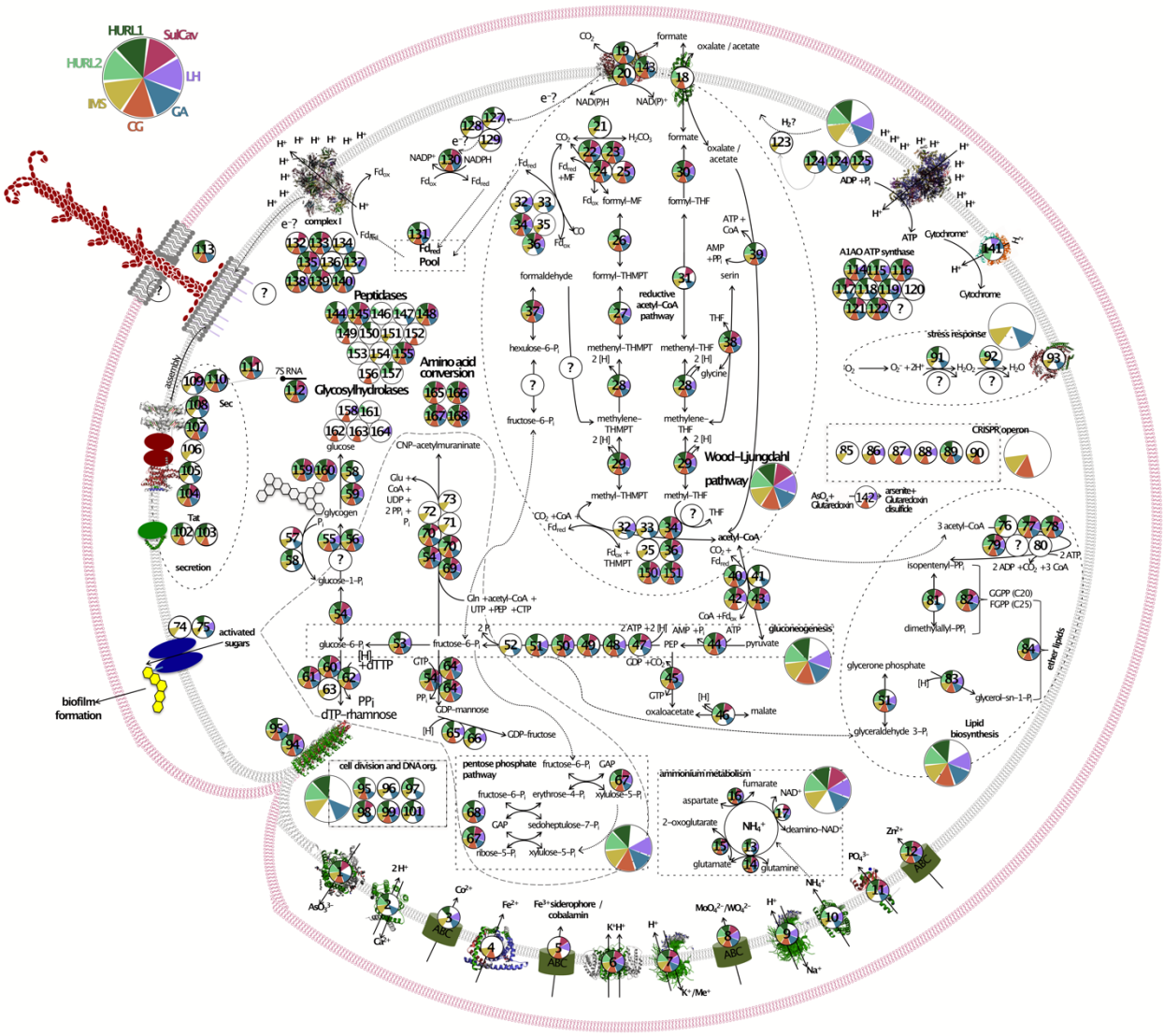

Figure III.3.5 | Metabolic capacities of *Ca.* Altiarchaeum pangenome. Previously identified genes in *Ca.* Altiarchaeum hamiconexum IMS (Probst *et al.*, 2014) were used as the basis to query the other genomes of known Altiarchaea clade members (see *Figure* III.3.4 for all members used in this analysis). To expand the predictable metabolic capacity of the genomes, METABOLIC (Zhou *et al.*, 2019) was used to annotate genes, which mainly resulted in peptidases and glycosylhydrolases. If multiple genomes copies per site were available, they were all used to query for the respective genes. All gene functions are listed in *Supplementary Data* 3.

This indicates an extreme degree of biogeographic provincialism across Earth. The small genetic divergence of *Ca.* Altiarchaea organisms in their core genome combined with their previously determined constant cell division (Probst *et al.*, 2014) implies a very slow evolutionary rate of these organisms. However, gene loss and horizontal gene transfer in *Ca.* Altiarchaea suggests a compensation for these slow evolutionary rates potentially providing a substantial advantage over other organisms in deep subsurface environments.

Discussion

Modelling of current cell counts estimate the amount of prokaryotic microorganisms in the continental subsurface to 2 to 6 x 10^{29} (Magnabosco *et al.*, 2018) which amounts to 60% of the prokaryotic life on our planet (Flemming and Wuertz, 2019). The diversity of microorganisms declines with sampling depth in the continental subsurface (Magnabosco *et al.*, 2018). Our metagenome assemblies showed the same trend in diversity change (based on the *rpS3* marker gene, cor=-0.40, p-value = 0.021, *Supplementary Figure* 10). This indicates that they are representative of general subsurface microbial communities and were consequently used to establish a genome database to calculate genome replication index values and minimal generation times across various subsurface ecosystems. These metrics revealed an apparent contradiction, with both replication index values and minimal generation times decreasing, thus indicating that organisms in the deep biosphere can replicate faster though they replicated less at the time of sampling. Prior studies (Starnawski *et al.*, 2017; Kirkpatrick, Walsh and D'Hondt, 2019) observed a reduction in microbial load with marine sediment depth and age, indicating that communities in older sediments were probably formed by members of surface communities that have a higher degree of persistence compared to others. Thus, subsurface communities would not be formed by actively replicating organisms but instead be shaped by the differing mortality of surface community members (Starnawski *et al.*, 2017; Kirkpatrick, Walsh and D'Hondt, 2019). The upper ten centimeters of sediment were found to be an exception showing active proliferation (Lloyd *et al.*, 2020). Although we analyzed many different ecosystems, our data do not allow drawing conclusions about the impact of mortality shaping subsurface microbial communities as they originate from different geologic formations. However, our observed decrease in replication measures with sampling depth does agree with these prior observations of a reduction of microbial load with depth and indicate that replication is occurring, albeit with less replication forks in the subsurface compared to near-surface ecosystems. On the other hand, the genome structures indicated a faster ability to replicate for organisms in the deep subsurface. This faster possible generation times with depth can be explained by the strategy employed by subsurface microorganisms recently termed as 'halt and catch fire' (Mehrshad *et al.*, 2021). This strategy refers to an adaptation to nutrient-poor environments like the deep subsurface, where organisms need to adapt to utilize short bursts of available nutrients and thus replicate fast during times when nutrients are available. Sites impacted by geological degassing showed a similar pattern compared to surface samples,

both in terms of replication index values and minimal generation time estimates. This could be caused by the unique geology of sites impacted by geological and thermal degassing. In these fracture-controlled aquifers, which are characterized by solid rock formation-embedded channels, flows can reach up to multiple magnitudes greater speeds than flows in comparable sediment-hosted aquifers. Thus, the availability of reduced mantle gases like H_2 and H_2S as microbial electron donors highlight the absence of nutrient bursts and the presence of a continuous nutrient flow similar to biomes on Earth's surface.

At Geyser Andernach, *Ca.* Altiarchaeota of the Alti-1 clade reach high cell densities in the CO_2 subsurface ecosystem and represent the main primary producers similar to the other high-CO_2 aquifer system Crystal Geyser, which additionally harbors a tremendous amount of bacterial diversity but also taps into three different aquifer ecosystems (Probst *et al.*, 2017, 2018). The predicted higher minimal generation time for the Alti-1 clade compared to their sister clade Alti-2 is likely caused by their higher costs of living. In contrast to their sister clade, *Ca.* Altiarchaea (Alti-1) live in biofilms, likely granting them increased survivability against a multitude of biotic and abiotic factors (See Olsen 2015 for a review on biofilm resistance (Olsen, 2015)). But this increased resistance also comes with a cost of requiring the synthesis of hundreds of their characteristic cell surface appendages called hami (Moissl *et al.*, 2005; Probst *et al.*, 2014) as well as other materials making up the extracellular polymeric substances (EPS) matrix. Additionally, *Ca.* Altiarchaea all need to assimilate CO_2 via the Wood-Ljungdahl-Pathway instead of also supplementing their carbon compounds by taking up organic carbon compounds as only gases can freely penetrate the biofilms. Thus, their proliferation would presumably be much more expensive than for their planktonic sister clade. This leads to the hypothesis that not replication speed but energy requirements limit *Ca.* Altiarchaea proliferation, making an optimization of the codon code to increase replication speed unnecessary.

The abovementioned hypothesis regarding replication speed of *Ca.* Altiarchaea would also align well with their strict biogeography. The clustering by continent of origin (North America, Europe, Asia), also reproducible in ANI and AAI (*Supplementary Figure* 7), indicate a strict provincialism. As dispersal via the surface is unlikely due to the high oxygen sensitivity of *Ca.* Altiarchaea (Probst *et al.*, 2014), plate tectonics could have been a viable alternative dispersal route providing ample opportunities for the common ancestor to distribute to North America and Europe. Plate tectonics have recently been implicated as the potential dispersal route for *Ca.* Desulforudis audaxviator to Africa, North America and Eurasia between 55 to

165 Myr (Becraft *et al.*, 2021). The dispersal of *Ca.* Altiarchaea could have occurred within the Phanerozoic, starting with the early Devonian (~400 Myr), when the continental margins Laurentia and Baltica, which form today's North America and Europe, respectively, collided to form Laurasia(Cocks and Torsvik, 2005; Torsvik *et al.*, 2012). Japan, on the other hand, has not been in contact with those margins since the break-up of Rodinia 750-600 Myr ago (Maruyama *et al.*, 1997), thus making dispersal to Japan during the Phanerozoic unlikely. As European and Japanese *Ca.* Altiarchaea are indicated to have a common ancestor, one possible route of dispersal from Europe to Japan could be across the Siberian plate through China in the early Mesozoic and then transferal to Japan during the plate processes, which uplifted the Japanese islands from the sea 25 Myr ago. Future studies are necessary to recover *Ca.* Altiarchaea genomes from Asia to further underpin this hypothesis of dispersal, since current public datasets from this continent are substantially underrepresented in databases.

The strict biogeography of the *Ca.* Altiarchaea is reflected by the conserved core metabolism, with most pathways being present in every *Ca.* Altiarchaea genome and indicate a slow evolving genus. However, observed putative gene loss and gene transfer events in investigated *Ca.* Altiarchaea populations indicate a compensatory strategy to counteract the slow evolutionary rate. This observed gene loss and transfer might be exuberated by the exclusive living in biofilms, which have generally been known as hotspots of horizontal gene transfer (HGT) for Bacteria (Hausner and Wuertz, 1999). The genes in *Ca.* Altiarchaea acquired via HGT are mainly from the bacterial domain, an evolutionary process frequently occurring in nature (Nelson-Sathi *et al.*, 2015). This HGT likely took place in the subsurface due to the immobility of *Ca.* Altiarchaea mediated by the anchoring of cells via their hami. Consequently, our analyses provide evidence that subsurface ecosystems impacted by geological degassing can be hotspots of microbial life and of increased evolutionary rates bolstered by lateral gene transfer across domains.

Methods

Geological setting. The cold-water Geyser Andernach is located 2 km downstream of Andernach (Rhine kilometer 615) on a 0.21 km^2 peninsula called Namedyer Werth in the Middle Rhine valley. Driven by magmatic CO_2, the geyser erupts regularly and intermittently approx. every two hours, when the groundwater filling the well is saturated with CO_2 and a reinforced chain reaction (domino effect) concludes in a gas/water-eruption up to >60 meters

in height (Schunk, 2012), lasting for 15-20 minutes. The well (drilling Ø 750/312/216 mm; casing/screens Ø 150 mm) was drilled in 2001 and is the third borehole (after 1903 and 1955) on this peninsula. The drilling taps 14 m of Quaternary fluvial deposits and continues then until its total depth of 351.5 m in lower Devonian formation called 'Hunsrück Schiefer s.l.' (shale) (Krauthausen, Deuster and Lang, 2007). A diagram of the plumbing system of the Geyser Andernach is provided in *Supplementary Figure* 1.

The small peninsula is part of the Pleistocene terrace which is covered by a thin sandy layer of fluvial Holocene deposits. Only at the NE margin of the peninsula the terrace is bare of deposits. The thickness of the Quaternary layer varies from 14 m (drilling 2001) to 20.75 m (drilling 1903) (Altfeld, 1913) and 24.2 m (drilling 1955) in the vicinity of the cold-water geyser. Beneath the Quaternary deposits follow lower Devonian rock formations of low metamorphic shale, such as clayish shale and intercalated minor layers of quarzitic sandstones; the thickness of these series is up to 5000 m.

The peninsula is located in the Middle Rhine Valley, which is a part of the European Cenozoic Rift System (Dèzes, Schmid and Ziegler, 2004). This rift system runs between the cities Bingen and Bonn in SE-NW-direction and crosses the Variscian complex of the Rhenish massif. Located at the SE edge of the lower Middle Rhine Valley, Geyser Andernach is situated on the intersection of two major fault structures: about one km to the NW the Variscian Siegen thrust fault running SW-NE crosses the Rhine Valley and can be traced for over 100 km from the Eifel area to the Westerwald. This fault shows a vertical displacement of several thousand meters, which occurred during the Variscian orogenesis, thus bringing rocks of the middle Siegenian stage in lateral contact with lower Emsian stage (Meyer and Stets, 2000). About two km to the SE the lower Middle Rhine valley is morphologically separated from the adjacent intraplate Tertiary Neuwied basin by an approx. 100 m vertical displacement caused by the SW-NE trending Andernach fault.

The Andernach fault and the Siegen thrust fault were in post-Variscian time intersected and 200-300 m displaced by a SE-NW trending dextral strike-slip fault (Meyer and Striem, 1983; Schreiber and Rotsch, 1998). The fault is supposed in the river Rhine bed and covered by Quaternary deposits. The horizontal movement was probably combined with shear strain and cataclastic rocks in the vicinity of the fault. This fault is the cause for pathways of mantle gases to reach the subsurface aquifers and ultimately the atmosphere.

Starting in Tertiary, a mantle plume under the Eifel area caused an uplift of the Rhenish massif during the last two million years and is the driving force for the volcanic activity in the Quaternary Eifel area since 700 k years (Ritter, 2007).

The mantle plume is the basic requirement for the rise of magma under and into the crust, whereby magmatic gases are released.

Sampling and geochemical measurements. The mesophilic and CO_2-driven Geyser Andernach (50.448588°N, 7.375355°E) in western Germany was sampled on 21 February in 2018 by collection of erupting water in sterile, DNA-free containers and subsequent filtration onto 0.1 µm pore size filters of 142 mm diameter (Merck Millipore, JVWP14225) and storage on dry ice / 193 K until DNA extraction. Water samples were collected during eruption of the geyser and analyzed biochemically as well as microscopically (see Suppl. Material for details). In total, two sequential eruptions were sampled, resulting in two filter samples for the first eruption and one filter for the second eruption. The upper 83 m of the geyser well have a casing and are sealed with cement so that no water can enter the well from the sides. The residual length of the geyser borehole (83-351.5 m) is intermittently covered by bridge-slotted screens which allow entry of CO_2-saturated water into the geyser well (*Supplementary Figure* 1). Each eruption flushes the tubing system (cylindric shape, 7.5 cm radius, 351.5 m length, approximate volume 6.2 m^3) with 6-7 m^3 water and an additional eruption was performed prior to the sampled eruptions to rid the tubing system of any stagnant water. The metagenomes recovered from both eruptions show identical community compositions and consequently, the sampled communities should be representative of subsurface communities and not contamination from the tubing system.

Metagenomic sequencing and processing. DNA was extracted from three individual 0.1 µm bulk water filtration filter membranes using the DNeasy PowerMax Soil DNA extraction Kit (Qiagen, JVWP14225) according to the manufacturer's instructions and further concentrated using ethanol precipitation with glycogen as the carrier. The samples were sequenced as part the Census of Deep Life phase 13 sequencing grant using Illumina NextSeq (paired end, 150 bps each). The three samples were processed individually as follows: Quality control of raw reads was performed using BBduk (Bushnell, https://sourceforge.net/projects/bbtools/) and Sickle (JN Fass, 2011). The metagenomic coverage and sequence diversity of metagenomes was estimated using Nonpareil3 (Rodriguez-R *et al.*, 2018) using k-mers of size 20. Reads were assembled into contigs and scaffolded using metaSPAdes 3.11 (Nurk *et al.*, 2017). For the sample IMS-BF, a sub-assembly of reads not mapping to the available *Ca.* Altiarchaeum

SM1 genome (GCA_000821205.1) was performed to improve assembly quality and this sub-assembly was used for the binning of additional genomes. Open reading frames were predicted for scaffolds larger than 1kbp using Prodigal (Hyatt *et al.*, 2010) in meta mode and annotated using DIAMOND blast (Buchfink, Xie and Huson, 2015) against UniRef100 (state Dec. 2017) (Suzek *et al.*, 2007), which contained the NCBI taxonomic information of the respective protein sequences. Taxonomy of each scaffold was predicted by considering the taxonomic rank of each protein on the scaffold on each taxonomic level and choosing the lowest taxonomic rank when more than 50% of the protein taxonomies agree. Reads were mapped to scaffolds using Bowtie2 (Langmead and Salzberg, 2012) and the average scaffold coverage was estimated along with scaffolds' length and GC content.

Binning of GA samples. Abawaca (Brown *et al.*, 2016), MaxBin2 (Wu, Simmons and Singer, 2016), tetranucleotide-based Emergent Self-Organizing Maps (ESOM (Dick *et al.*, 2009)), and CONCOCT (Alneberg *et al.*, 2014) were used to identify metagenome assembled genomes and DAS Tool with standard parameters was used to aggregate the results (Sieber *et al.*, 2018). (See supplementary methods for a detailed listing of the parameters used). Binning of publicly available datasets was carried out using a combination of MaxBin2, Abawaca and tetranucleotide ESOM, if possible. Bins were refined using GC content, coverage and taxonomy and their completeness and contamination was accessed by a set of 51 bacterial and 38 archaeal single copy genes as described previously (Probst *et al.*, 2017, 2018). Only bins with >= 70% estimated completeness and <= 10% estimated contamination were used for downstream analysis. For each sample, genomes were dereplicated using dRep (Olm *et al.*, 2017).

Ribosomal protein S3 (rpS3) analysis. Genes annotated as ribosomal protein S3 were extracted and assigned to genomes where possible based on shared GC, coverage, and taxonomy. *rpS3* coverage was determined based on the scaffold coverage (see above) containing the ribosomal protein. Ribosomal protein sequences were clustered using MUltiple Sequence Comparison by Log-Expectation (MUSCLE) (Edgar, 2004), trimmed using BMGE 1.0 (Criscuolo and Gribaldo, 2010) with the BLOSUM62 scoring matrix and aligned using IQ-TREE (Nguyen *et al.*, 2015) multicore 1.3.11.1 with -m TEST -bb (Hoang *et al.*, 2018) 1000 and -alrt (Guindon *et al.*, 2010) 1000 options. The tree was visualized along with other genomic data using the iToL platform version 5.5 (Letunic and Bork, 2016).

Identification of potential contaminant genomes. The GTDB-Tk (Chaumeil *et al.*, 2020) classify_wf workflow with default parameters was used to place the recovered genomes from

the Geyser Andernach in relation to a reference dataset. If a close relative genome was identified in this approach, we calculated the ANI between the reference and the newly recovered genome. The only genome showing a similarity >= 80 % ANI to the reference dataset was GA_180221_E-1-2_metaspades_Carnobacterium_36_4 (96.42 % ANI to *Carnobacterium alterfunditum* GCF_000744115.1) and was thus identified as a potential contaminant and excluded from further analyses.

Determination of bacterial in situ replication index. Reads were mapped onto concatenated genomes per sampling site using Bowtie2 with the reorder flag (Langmead and Salzberg, 2012) and the index of replication (iRep (Brown *et al.*, 2016)) was calculated, allowing for 2% mismatches relative to the read length (3 mismatches for 150 bp). The calculation of in situ replication index values is based on the assumption that organisms, that are actively proliferating, replicate their genome starting at the origin of replication and ending at the terminus of replication. Replicating organisms can thus have already replicated the parts of their genome close to the origin of replication but have not yet completed replicating sequences close to the terminus of replication. This can result in higher relative coverage of the sequence close to the origin of replication compared to the terminus of replication. Multiple simultaneous replication processes can exuberate this difference further. The in situ index of replication (iRep) estimates the number of replication processes based on this coverage difference but only works in Bacteria as Archaea can have multiple origins of replication (Z. Wu *et al.*, 2014) and thus the iRep signal is distorted and cannot be applied in a comparative manner. If multiple samples were available for one ecosystem, all iRep values for one genome were calculated and averaged to ensure comparability with other samples.

Prediction of maximal growth rates. Growthpred (Vieira-Silva and Rocha, 2010) values were calculated on prodigal-predicted genome gene sets in nucleotide format with the -t parameter and otherwise default options. Growthrate estimators like Growthpred utilize differences in codon usage between genes which are continuously expressed like housekeeping genes (by default growthpred uses ribosomal proteins) and the rest of the gene pool to predict how optimized the genome is for a faster replication. In contrast to iRep, growthpred does predict the actual fastest rate at which a genome can replicate.

Metabolic potential predictions. A set of Hidden Markov-Models (HMM) with respective score thresholds for chemolithoautotrophic key enzymes (Anantharaman *et al.*, 2016) was used to predict the metabolic potential of recovered genomes and overall in entire assemblies (See suppl. Material for more detailed information).

Biogeographical analysis. The R package sp (Pebesma and Bivand, 2005) was used to calculate the geographical elliptical distance between two sampling sites (based on longitude/latitude), in which putative genomes of the *Ca.* Altiarchaeales subclade Alti-1 were identified. The average nucleotide identities (ANI) between all available putative genomes of the *Ca.* Altiarchaeales subclade Alti-1 were calculated using the ANI calculator (Rodriguez-R and Konstantinidis, 2016) with default parameters. Correlations between geographical distance and ANI were done using Pearson's r (R Core Team, 2008).

Genome comparison of *Ca.* Altiarchaeota. Genes of all *Ca.* Altiarchaeota genomes were blasted against each other (E-value: 10^{-5}) and matches were filtered to matches with the similarity [(AlignmentLength x Identity) / QueryLength] thresholds of >= 40 %, 50 %, 60%, 70% or 80% . Cytoscape 3.7.2 (Shannon *et al.*, 2003) was used to visualize the networks at the respective similarity thresholds.

Metabolic network of Ca. Altiarchaea (Alti-1). The annotated genes from (Probst *et al.*, 2014) were used as the basis to identify homologues in other Alti-1 genomes using an E-value of 10^{-5} as the cutoff. If multiple versions of a genome were available, their results were concatenated. Additionally, genomes were annotated using METABOLIC (Zhou *et al.*, 2019), mainly incorporating annotations for glycosyl hydrolases, peptidases and aminotransferases.

Phylogenomic analysis of *Ca.* Altiarchaeota. Amino acid sequences and annotations for Alti-1 ORFs plus one Alti-2 serving as outgroup were predicted using Prokka 1.14.0 (Seemann, 2014) with options: --kingdom archaea --metagenome --compliant). The resulting protein datasets were searched with HMMER 3.2.1 (Eddy, 2011) for homologs of 30 universal ribosomal proteins using the v4 HMM profiles from Phylosift (Darling *et al.*, 2014). A 10^{-4} cutoff was applied, and the resulting datasets were curated manually to remove distant homologs and multiple copies in each genome, as well as to fuse contiguous fragmented genes. Individual genes were aligned with MUSCLE v3.8.31 (Edgar, 2004) and trimmed with BMGE (Criscuolo and Gribaldo, 2010) under the BLOSUM30 matrix. The genes were then concatenated into a supermatrix of 5156 aa positions. The phylogeny was reconstructed in IQTree 1.6.11 (Nguyen *et al.*, 2015) under the JTTDCMut+F+G4 model as selected by ModelFinder (Kalyaanamoorthy *et al.*, 2017).

Tracking of gene loss and gene transfer events in *Ca.* Altiarchaea. To identify genes that were lost in multiple *Ca.* Altiarchaea or identify genes that were acquired by individual *Ca.* Altiarchaea through HGT, we selected genes only present in one or two *Ca.* Altiarchaea genomes (*Figure* III.3.5) for phylogenetic analyses. The selected genes were used as BLASTp

queries (E-value: 10^{-5}) against a reference database of bacterial and archaeal genomes, retaining up to 2,000 hits per search. The database is a concatenation of bacterial and archaeal genomes in NCBI Genome database (accessed 2019.06.01), dereplicated using rpS3 amino acid sequence clustering with CD-Hit at 99% identity followed by dRep at 95% ANI to get a single representative genome per species. This resulted in a databank of 25,226 bacterial and 1,808 archaeal genomes. Taxonomic information and functional annotation (when available for genomes with protein datasets) were used directly from NCBI. If no protein dataset was available, the translated ORFs were predicted with Prodigal. Genes were aligned with MUSCLE, trimmed using BMGE with the BLOSUM30 matrix and their phylogeny was reconstructed using IQTree2.0-rc2 with the -m MFP, -bb 1000 and -alrt 1000 options.

Community-wide analyses. Genes were predicted on assemblies with scaffolds longer than 1 kbp and chemolithoautotrophic key enzymes were predicted as described above. The abundance of the genes was estimated using the coverage of the encoding scaffolds after adjustment to unequal sequencing depths by normalization using the total bps per library. If a pathway was represented by multiple key enzymes, the enzyme with the highest frequency of hits was selected. Abundances of individual key enzymes were summed to provide the total relative abundance of each pathway in the respective samples. Likewise, diversity within each assembly was estimated based on *rpS3* diversity and relative abundance of the respective scaffolds.

Estimations of annual total erupted carbon and intracellular erupted carbon. The annual total erupted carbon was calculated based on the available CO_2, HCO_{3-} and cell concentrations, the eruption volume (*Supplementary Table 1*), the average estimate of the intracellular carbon amount from (Kallmeyer *et al.*, 2012) of 14 fg $cell^{-1}$ and the number of eruptions during tourist season (roughly 1 April – 31 October ~ 210 days). See the Supplementary Material for the calculations.

Statistical analysis. Statistical analyses were performed in the R programming environment (R Core Team, 2008). These included paired and independent *t*-tests, Pearson correlations, analysis of variance (ANOVA), TukeyHSD significance tests (Haynes, 2013), the Shannon-Wiener index (Shannon, 1948) and equivalence testing using TOSTER (Lakens, Scheel and Isager, 2018). As the upper and lower equivalence boundaries for equivalence testing of two groups, we used the effect size the CO_2-poor sample group had a 33% power to detect as recommended previously (Simonsohn, 2015). Results were visualized using ggplot2 (Wickham, 2009).

Methods for DAPI staining, cell counting, geochemical measurements are provided in the Supplementary Methods.

Data availability

Raw sequencing data and MAGs from Geyser Andernach have been deposited at SRA and Genbank, respectively, and are available under the BioProject PRJNA627655. MAGs binned from additional ecosystems have been deposited at Genbank in the BioProject PRJNA767587. Individual BioSample ID's of all MAGs are listed in *Supplementary Data* 1 and individual SRA accession codes are listed in *Supplementary Table 4*.

Acknowledgements

This study was funded by the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (Nachwuchsgruppe Dr. Alexander Probst). The Geyser Andernach metagenomes were sequenced within the Census of Deep Life Sequencing call 2017, phase 13 project Microbial metabolism in a deep subsurface, shale-hosted aquifer of the Volcanic Eifel (central Europe): A comparative analysis of two cold, high-CO_2 geysers. We thank Hubert Müller for technical assistance, Sabrina Eisfeld for laboratory maintenance, Ken Dreger for server administration and maintenance and Karen L Lloyd for scientific discussions.

Author contributions

TLVB performed main bioinformatics analysis. PSA performed phylogenomics. VT and AJP performed microscopy. US, RS and BK performed geological analyses and geological data interpretation. TLVB and PAFG analyzed genomes. TLVB, JR and AJP took samples. DK and TCS performed geochemical analyses. AJP conceptualized the study. TLVB and AJP wrote the manuscript with revisions from all co-authors.

Competing interests

The authors declare no competing interests.

Supplementary material

The supplementary material is available under https://www.nature.com/articles/s41467-021-27783-7 as well as on the accompanying CD (see section VII Content of supporting CD for more information).

IV. General Discussion

1. The need for high quality genomes from metagenomes

1.1 Considerations for established workflows to generate genomes from metagenomes

The development of next generation sequencing (NGS) has revolutionized our understanding of the biological diversity in the environment, opening up the vast majority of organisms for research. Due to the novelty [first genome-resolved metagenomes in 2004; (Tyson *et al.*, 2004; Venter *et al.*, 2004)] and extreme speed of development, the workflows and software used to analyze metagenomes are in a steady flux, with no specific approach being the community standard. Indeed, the approach already varies a lot depending on the type of metagenomic analyses performed: read-based, assembly-based or genome-resolved metagenomics (see section III.1 for more information on these types of metagenomic analyses).

Within book chapter 33 in "Methods in Molecular Biology, vol. 2522" (see section III.1) , we aimed to provide a workflow to retrieve archaeal genomes from metagenomic raw reads. Many steps of the given workflow can be accomplished with alternate software, with only a few steps of the workflow having an actual community standard in terms of used software. Among those few established steps is the assembly process, in which metaSPAdes (Nurk *et al.*, 2017) are the commonly accepted state of the art approach, as well as the prokaryotic ORF annotation using prodigal (Hyatt *et al.*, 2010). For most other steps, many different software solutions exist, though, generally, the principles of what those do remains the same, starting with quality trimming of the reads, their assembly, ORF prediction and annotation, abundance estimation via mapping. Binning has an even greater variety of different tools and approaches [see (Sieber *et al.*, 2018) for an overview of a small collection of binning tools], with none being consistently the best, hence requiring the use of many different types of binners and aggregating the results to get the best results (Sieber *et al.*, 2018). Even the combination of various binners, while being the state of the art, is not yet widely applied, as it complicates the workflow, making it unattractive for studies not primarily focused on genome-resolved metagenomics.

Bin curation on the other hand is not commonly applied, likely due to a missing automated software that facilitates this step. This makes it the bottleneck in the otherwise high-throughput and automatable workflow. Instead, various manual bin curation tools are

sometimes used. Once bins have been produced, CheckM (Parks *et al.*, 2015) and GTDB-tk (Chaumeil *et al.*, 2020) are the community standard for their quality assessment and classification, respectively. Other than that, follow-up analyses are very diverse, depending on research question, ranging from strain delineation, metabolic prediction, growth estimates, detailed phylogenomic analyses and virus-host analyses to name just a few.

Since the field is so broad, both in terms of available software and types of possible analyses, the supplied workflow should just be seen as an option to recover genomes from metagenomes and shows the combination of tools we commonly use for this task. It should be noted that while the book chapter (section III.1) named for recovering archaeal genomes specifically, there is no reason why it cannot be used for bacteria as well. Indeed, the only difference between recovering archaeal and bacterial genomes are the single copy genes used to estimate completeness and contamination[and the possibility of there being alternatively coded bacterial genomes in the community such as the Gracilibacteria (Wrighton *et al.*, 2012)].

1.2 Why uBin is needed in the bin curation landscape

Prior to the development of uBin, various other bin curation tools were already available. So one might ask, why an additional tool is needed when other tools are already available. In general, until options to reliably automate genome curation are developed, researchers will need to continue the tedious curating process manually. Despite being suggested as an obligatory part in genome-resolved metagenomics studies (Chen *et al.*, 2020), bin curation is still not used everywhere. Hence, it is extremely important to make these tools as easy to use as possible and thus give researchers as much of an incentive to curate their genomes. We believe that existing tools all have a much higher barrier of entry, either requiring prior upload of metagenomic data, then also using their entire server structure (Wrighton *et al.*, 2012), being just a subroutine of a much larger software suite (Eren *et al.*, 2015) or being mainly a reporting/visualization tool (Seah and Gruber-Vodicka, 2015; Vollmers *et al.*, 2022). Hence, we tried to make uBin as easily usable and accessible as possible.

uBin itself is inspired by ggKbase. We aimed at improving the ggKbase curation interface (ggKbase curation inferface) by improving upon the individual selection plots, *e.g.*, also allowing selection of target taxonomies instead of purely excluding taxonomies, and also by including a combined GC and coverage scatterplot to find outliers in both metrics. The uBin software is available on all commonly used platforms as an open-source desktop application

(in contrast to ggKbase which is not easily accessible due to requiring user account generation, upload of genomes and being browser-based). We hope that its ease of use will make bin curation much more accessible to a wider audience and thus in the long run improve upon the genome quality in individual research but also in databases. We reiterate it is very subjective, which bin curation software one likes to use, and there consequently is no real "best" bin curation software.

Note that the uBin software was not used for bin curation in section III as the binning and bin curation performed for this section predates even its alpha versions. Bin curation in this section was performed in Tableau, as it makes approximation of the later uBin interface (GC and coverage bar charts, scatterplot, single copy gene plots) and selection within those possible, albeit without a taxonomy wheel to easily filter taxonomy (deselection had to be done by ticking off specific taxonomic names). While functional, this approach left much to be desired, as large datasets were very slow in loading and the handling of the software interface, designed to easily plot and manipulate economics data, had an entry barrier. Thus, curating these (and other) genomic datasets were the main reason why we felt that the uBin software was needed.

1.3 Implications of decreasing growth in the deep biosphere on genome fluidity

The established metagenomic processing, binning and curation workflow described in sections III.1 and sections III.2 allowed us to reconstruct a database of subsurface genomes, spanning a depth gradient of 0-3 km, and allowed us to verify whether some of the trends observed in individual sites, such as the microbial load decreasing with depth (Magnabosco *et al.*, 2018), or functional against taxonomic conservation, hold true across continents and a large depth span.

In this work, a trend of decreasing ongoing growth (*i.e.*, lower replication indices) with faster replication speed (*i.e.*, minimal generation times) with depth was detected (Figure III.3.2), and was attributed to cells at greater depths being exposed to more short nutrient bursts to which they should react fast. For small populations, it is conceivable that there might be a slight dependency between the number of detected replication forks, approximated by iRep values, and the minimal generation time, approximated by growthpred values, as the time spent

with detectable replication forks would be only as long as the minimal generation time. However, since metagenomics generally looks at entire populations of many thousands of organisms, this is unlikely to have much of an effect as enough of a sample size is measured to accurately represent the average replication forks happening.

The lower replication happening in the deep subsurface might have severe implications on deep biosphere communities and their genetic adaptation to their environment as well as their arms race against viral predation as genome replication as a source of adaptation happens much less frequently than in the critical zone. Hence, faster ways of adaptation, such as horizontal gene transfer (Figure II.3.1A) might be much more important in these environments, particularly if genetic exchange is easily possible such as in aquifer systems (in contrast to, *e.g.*, microbes encased in stone).

In this thesis, I specifically looked at the genetic diversity in *Ca.* Altiarchaea of clade Alti-1, which are indicated to have been distributed via plate tectonics (Figure III.3.4A,C), though more Alti-1 populations are needed to accurately map the entire dispersal route, and specifically the timing of the allopatric speciation events. Both the core metabolism as well as the overall genomes of Alti-1 *Ca.* Altiarchaea are very conserved (Figure III.3.5; Figure III.3.4B), indicating very little genetic divergence, though their auxiliary metabolism shows signs of adaptation, with some of these adaptations being caused by horizontal gene transfer from Bacteria. Since all *Ca.* Altiarchaea identified so far inhabit either sulfidic springs & caves or cold geysers, this may restrict their potential diversity, though biogeographic distribution seems to outweigh ecosystem type, as the two Alti-1 genomes from cold geysers (Crystal Geyser, Utah and Geyser Andernach, Germany) do not form a cluster (Figure III.3.4C). Additionally, their lifestyle as biofilms might be sufficient to negate the need for individual genome adaptation to stressors, as biofilms are known to be very beneficial as a stress defense (Davies, 2003). On the other hand, biofilms are known to be hotspots of horizontal gene transfer (Madsen *et al.*, 2012) and thus might make it easier for beneficial transfers to disseminate throughout the community.

The following sections will utilize a complete *Ca.* Altiarchaeum GA genome recovered from the Geyser Andernach by Sophie Simon (Group for Aquatic Microbial Ecology, University-Duisburg-Essen) to relate *Ca.* Altiarchaeum *in situ* replication indices to the overall community as well as take an in-depth look into areas of genetic divergence between *Ca.* Altiarchaeum genomes.

2. Genome characteristics and fluidity as revealed by complete *Ca.* Altiarchaeum GA genome

2.1 *Ca.* Altiarchaeum GA have non-canonical origins of replication

Ca. Altiarchaea of the clade Alti-1 were estimated to have a minimal generation time of ~7.5 h based on the differences in the codon usage between ribosomal proteins, *i.e.*, house-keeping genes, and the residual gene repertoire (see section III.3). Other techniques, available for bacterial draft genomes, such as the calculation of replication indices (Brown *et al.*, 2016), cannot be used for most Archaea since they can have multiple origins of replication (Z. Wu *et al.*, 2014). The calculation of bacterial replication indices relies on arranging the scaffolds in a draft genome so that the difference in sequence abundance between the origin of replication (where replication starts) and the terminus of replication (where replication ends) can be calculated and thus identify the percentage of the population of actively replicating. Due to the possible multiple origins of replication in Archaea, it is not possible to arrange the genome from origin to terminus of replication with any certainty if the genome is fragmented as is typical for draft genomes (Zhang and Zhang, 2005). Hence, complete genomes, *i.e.*, circular genomes, of Archaea are needed to be able to calculate replication indices. But these complete genomes are very difficult to obtain using short read metagenomics. Difficulties arise from strain heterogeneity, repetitive genome regions (repeats longer than read length), highly similar genomic regions across populations (*e.g.*, transposons or CRISPR arrays, gene duplications), non-protein coding regions having an atypical k-mer frequency (*e.g.*, 16S rRNA genes). This assemblies from short reads to be rather fragmented, making it very rare to recover complete genomes. Long-read metagenomics like ONT or PacBio are much better suited to generate complete genomes as their up to many kbp-sized reads can assemble across problematic regions like repeats (Chen *et al.*, 2020). However, current long-read metagenomes still have some disadvantages over short-read metagenomes, such as higher error rates, being more expensive and generally having lesser sequencing depth [see (Amarasinghe *et al.*, 2020) for a review on the advantages and disadvantages of long-read metagenomics].

Using ONT, Sophie A Simon reconstructed a complete genome of *Ca.* Altiarchaaeum from Geyser Andernach (sampled on 30.11.2021, 40L on 0.2 µm pore size filter) with a total

length of 1687308 bp, using metaFlye (Kolmogorov *et al.*, 2020) assembly, inspection of altiarchaeal scaffolds via uBin and identification of single circular scaffold of altiarchaeal origin, *i.e.*, a complete *Ca.* Altiarchaeum GA genome, followed by read mapping on the *Ca.* Altiarchaeum GA genome using Minimap2 (Li, 2018) and finally reassembly of the mapped reads using Trycycler [(Wick *et al.*, 2021); Supplemental file sectionIV_FileS1 contains the genome in FASTA format, please see section VII Content of supporting CD for more information on Supplemental files]. CheckM (Parks *et al.*, 2015) assigned it 86.1% completeness, 4.46% contamination and 16.67% strain heterogeneity using the Euryarchaeota-specific marker set (there is no marker set specifically for *Ca.* Altiarchaea). Since this genome is circularized and hence presumed complete (excluding possible extrachromosomal elements (Wang *et al.*, 2015)), the low completeness indicates that some of the markers used for this estimation are either missing or obfuscated by *e.g.*, frameshifts that can occur in ONT sequencing (Hackl *et al.*, 2021). Thus, based on this genome, CheckM consistently underestimates *Ca.* Altiarchaea completeness. Multiple sequence alignments of duplicated ("contaminating") marker sequences indicated that some of these contaminations could be the result of gene duplications due to their high sequence similarity and close proximity on the genome. This is also indicated by the strain heterogeneity value, which measures the average amino acid identity between duplicated sequences to identify closely related contaminations (Parks *et al.*, 2015). Examples of gene duplications are the ribosomal protein S19e and NAD+ synthetase. The ribosomal protein S19e is of particular interest as it is regarded as an universal single copy gene and used as such in both DAS tool (Sieber *et al.*, 2018) and uBin (Bornemann *et al.*, 2020), consequently overestimating the contamination level in *Ca.* Altiarchaea genomes if both are assembled and binned into the same genome. The predicted rpS19e loci are 43 kbp distant from each other. This fact excludes that ORF prediction might have split the single rpS19e gene into two, resulting in a false dual annotation of this gene as the loci should in that case be next to each other. All genomes of *Ca.* Altiarchaeum GA recovered in section III.3 have only either zero (GA E1-1) or one copy of the gene (GA E1-2; GA E2-1). This might indicate that the nearly identical genes got assembled together due to their shared k-mer patterns and provide one example why short-read metagenomics was not able to recover a closed *Ca.* Altiarchaeum genome. Bird *et al.* also reported in 2016 that a gene encoding for a putative phenylalanine tRNA synthetase beta subunit, which is regarded as an universal marker gene for Archaea [(Probst *et al.*, 2017); and used as such in uBin], is split in Altiarchaea (Bird

et al., 2016). However, CheckM does not use this marker gene in its Euryarchaea marker set and thus their apparent contamination value remains unaffected.

Origins of replications are usually located at positions of asymmetry in the nucleotide composition across a genome (Lobry, 1996). In contrast to Bacteria, Archaea can have multiple origins of replication, with, *e.g.*, *Sulfolobus solfataricus* P2 having three origins of replication (Zhang and Zhang, 2005). These are expected to have resulted from the uptake of extrachromosomal elements (Samson *et al.*, 2013). Many techniques like the GC skew (Lobry, 1996) and others listed in Zhang and Zhang (2005) were developed to identify the origin of replication in Bacteria as well as Archaea, with the Z-curve method (Zhang and Zhang, 2005) being one of the latest approaches. In this method, the sequence characteristics (amine versus ketone (MK), purine versus pyrimidine bases (RY), strong versus weak hydrogen bonds (SW)) are cumulatively summed up across the genome sequence and plotted in a 3D line graph. Each of these characteristics can also be plotted separately on the Y-axis in 2D line graphs across the genome sequence (X-Axis; Figure IV.2.4.1). While most Z-curve plots of Archaea (and Bacteria) show clear extremes at origins of replication (Zhang and Zhang, 2005), the *Ca.* Altiarchaeum GA genome shows no clear maxima (Figure IV.2.4.1). Ambiguous results from the Z-curve method have been reported previously for a miniority of archaeal genomes (Zhang and Zhang, 2005). Examples of these are *Nanoarchaeum equitans* and *Archaeoglobus fulgidus* (Zhang and Zhang, 2005).

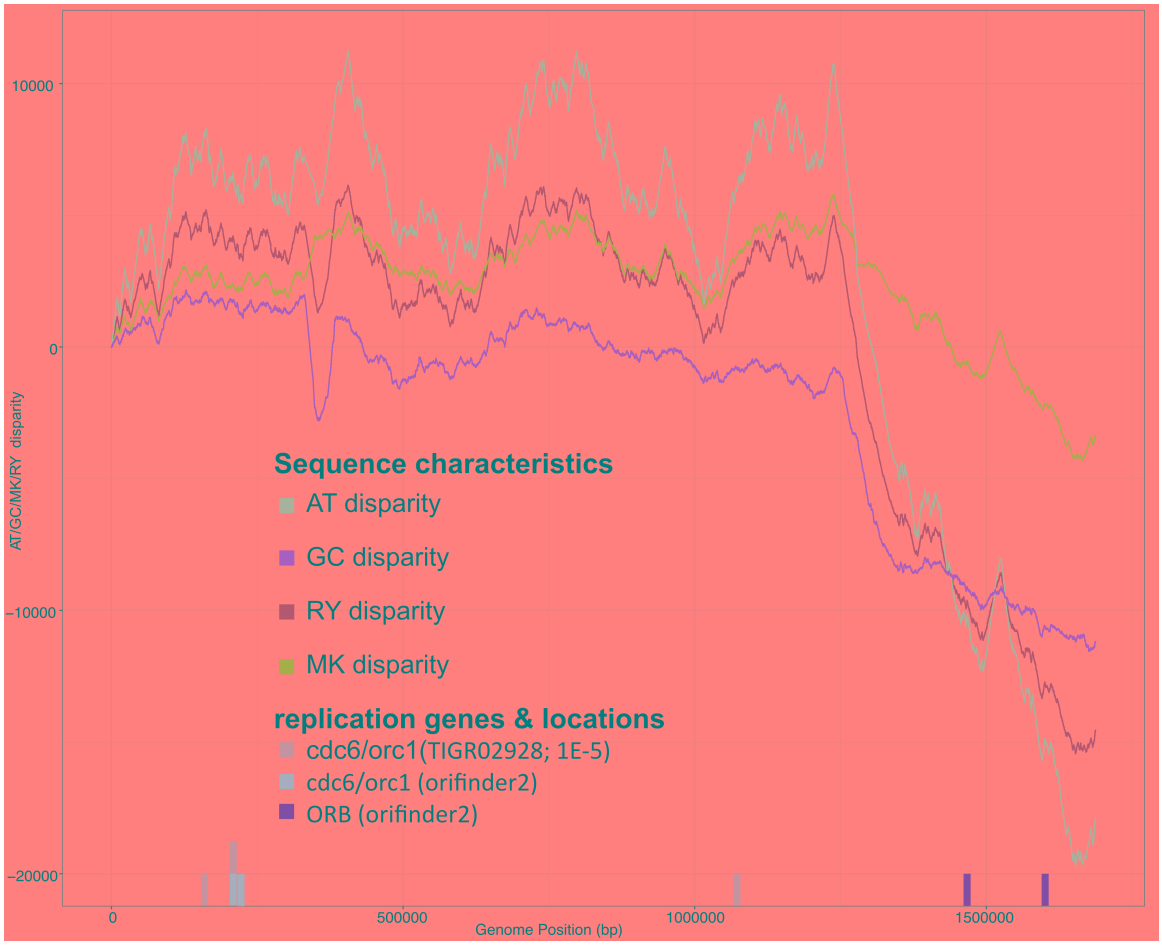

Figure IV.2.4.1: Identification of putative origins of replication and replication initiator genes in the complete *Ca.* Altiarchaeum GA genome. Sequence characteristics were calculated with an in-house shell script and plotted with ggplot2 (Wickham, 2009) in R (R Core Team, 2008). Origin recognition boxes (ORB) and cdc6/orc1 initator protein sequences were identified either with orifinder2 (Luo, Zhang and Gao, 2014) or hmmsearch (Eddy, 2011) using the publicly available TIGR02928 hidden-markov model at an E-value threshold of 1E-5. The x, y and z coordinates per nucleotide position used to generate the figures are given in supplementary file sectionIV_FileS2.

A second indicator for the location of origins of replication are origin recognition boxes (ORB) and cdc6/orc1 initiator protein sequences, which are typically adjacent to each other (Ausiannikava and Allers, 2017). Both of these are necessary for the start of replication in Archaea, with the Cdc6 initiator protein binding to the ORB sequence and thus initiating the replication process (Ausiannikava and Allers, 2017). In the *Ca.* Altiarchaeum GA genome, two ORB regions and 4 *cdc6* genes were identified but *cdc6* and ORB regions were not co-localized (Figure IV.2.4.1). While archaeal genomes can have multiple *cdc6* homologs and are suggested to use their differing affinity for different ORB sequences and their respective origins of replication to modulate the replication process (Coker *et al.*, 2009), they are almost always co-localized [though the *cdc6* gene is not always essential (Coker *et al.*, 2009)]. One exception is

the *Sulfolobus solfataricus* P2 genome, where the *cdc6* gene has a distance of about 80 kbp to its closest origin (Lundgren *et al.*, 2004), which is still much closer than the distances observed for the *Ca.* Altiarchaeum GA genome. It has been suggested that origins of replication with distant *cdc6* genes could be replicating with a different mechanism (Zhang and Zhang, 2005). In the case of *Sulfolobus islandicus*, the Cdc6 protein is non-essential as *S. islandicus* additionally encodes for the Whip protein, which can also bind ORB regions and thus replace functions (Samson *et al.*, 2013). No *whip* homolog could be identified in the *Ca.* Altiarchaeum GA genome via DIAMOND BLAST (Buchfink, Xie and Huson, 2015) against FunTaxDB [(Bornemann *et al.*, 2020); E-value: 1E-5].

Both of the ORB regions identified in the *Ca.* Altiarchaeum GA genome have a preceding DEDD_Tnp_IS110 domain-containing protein, a relatively uncharacterized type of transposase (Garcia, Cavanaugh and Kacar, 2021), which could indicate that the ORB has been transferred via a mobile genetic element (see sectionIV_FileS3 for a table with blast results for each ORF along with its position on the genome). The second ORB region (1598752-1599190 bp) also contains many more adjacent transposase genes while the first ORB region (1466766-1467456 bp) contains mostly uncharacterized proteins instead. Thus, both detected ORB regions might be the result of HGT and may not be actively used origins of replication.

In summary, based on sequence information, no origins of replication can be confidentially identified based on sequence information in the *Ca.* Altiarchaeum GA genome, as no initiator proteins and ORB regions are co-localized (and all ORB regions are adjacent to transposons). Possible reasons could be non-canonical ORB regions or Cdc6 proteins.

2.2 *Ca.* Altiarchaeum GA genome coverage indicates active replication

Provided that large proportions of the *Ca.* Altiarchaeum GA population are replicating, the coverage around the origin of replication(s) should be at the coverage maxima across the genome (also assuming that the same origins of replication are used across the population for replication), as this part of the genome is replicated first, with the terminus of replication being replicated last. This characteristic is employed in the iRep algorithm extensively used in section III.3 to estimate replication indices across the subsurface (see Figure III.3.2). However, the iRep algorithm only works with genomes having a single origin of replication, *i.e.*, in Bacteria

and a few Archaea, as it relies on rearranging the fragments of a draft genome in a coverage-descending order, thus generating a slope from the origin of replication to the terminus of replication (Brown *et al.*, 2016).

When applied to the *Ca.* Altiarchaeum GA genome, and using various available metagenomic datasets from GA, a single maximum is clearly visible in all datasets, peaking at around 1.5Mbp for samples from 2018 (GA E1-1, GA E1-2, GA E2-1) and peaking at 1.3Mbp for samples from 2019 (GA2019_1, GA2019_2). Figure IV.2.2.1.A shows the coverage distribution exemplified for GA E1-1.

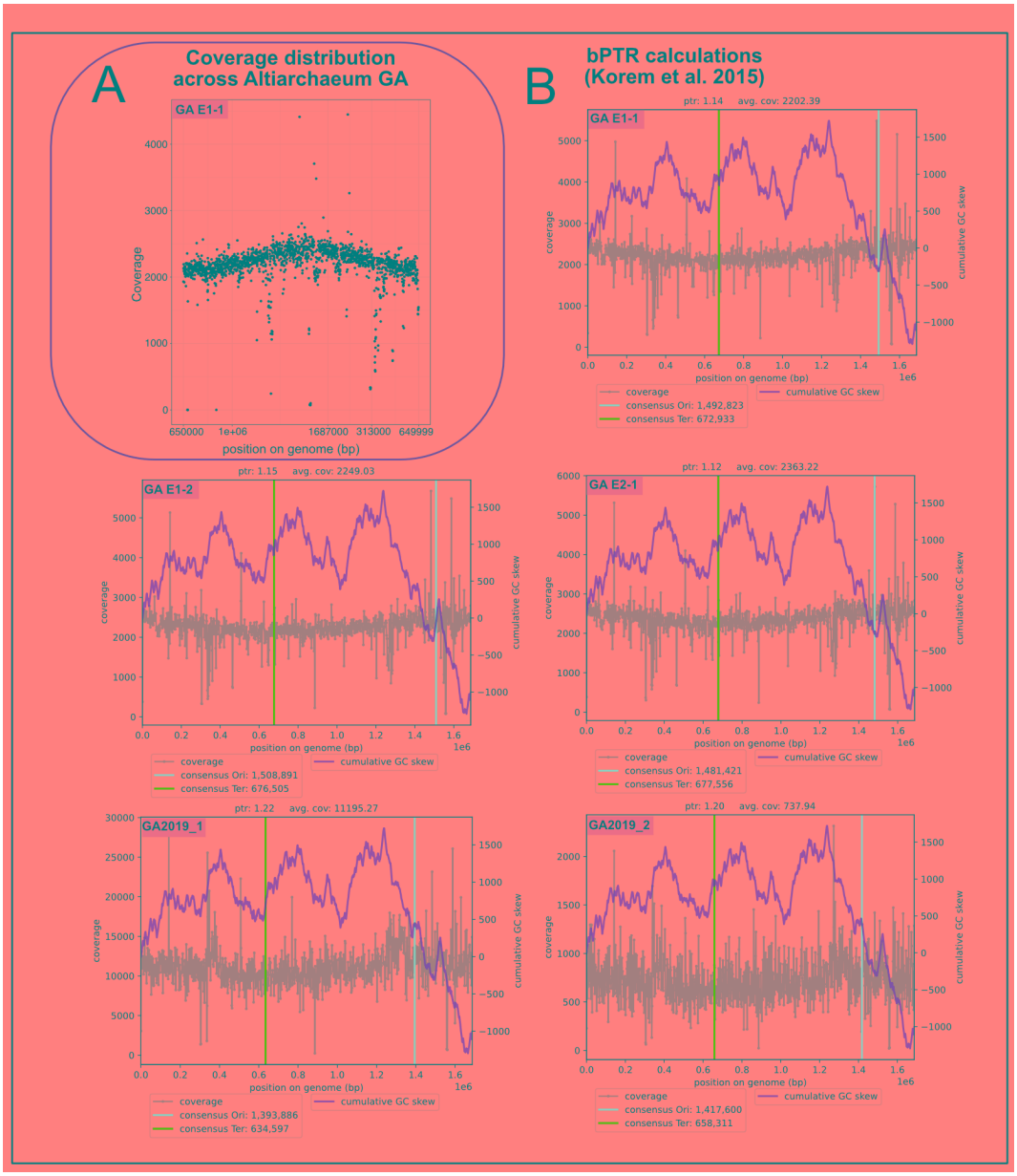

Figure IV.2.2.1: Peak-to through ratio calculations for *Ca.* Altiarchaeum GA genomes. A: Coverage distribution after mapping using Bowtie2 (Langmead and Salzberg, 2012) of short reads to *Ca.* Altiarchaeum GA genome, using the metagenomic dataset GA E1-1 as an example. The genome

was rearranged to have the maximum in the middle of the sequence. **B:** bPTR calculation plots as described in (Korem *et al.*, 2015) and implemented as a script in (Brown *et al.*, 2016) for all five metagenomic datasets available for the Geyser Andernach. The origin of replication and terminus was determined using the coverage mode as GC skew was an unreliable measure for this genome (see Figure IV.2.1.1).

No clear coverage minimum could be identified, with coverage profile plateauing in the sequence area of 535kbp – 835 kbp (Figure IV.2.2.1.A**)**. The coverage profile, however, allowed identification of a clear maximum, which should correspond to the origin of replication (and a single origin of replication is the requirement for Peak-to-Through-Ratio methods). Hence, the bPTR method (Korem *et al.*, 2015), made available as a script (Brown *et al.*, 2016), was used to estimate a putative replication index. The coverage mode was used to identify the origin of replication (Figure IV.2.2.1.B, GC skew-based origin localization gave similar results, data not shown). Estimated growth index values ranged between 1.12-1.22 (*i.e.*, 12-22 % of the population were replicating at the time of sampling) and were thus far below the ~1.5 iRep value estimated for the rest of the community, and indeed lower than any replication measures calculated during section III.3. Provided that the bPTR method is comparable to these iRep values (shown in (Brown *et al.*, 2016), though only on single genome and with bPTR having consistently slightly lower values), this indicates that a far smaller percentage of the *Ca.* Altiarchaeum GA population is replicating than for the rest of the community in Geyser Andernach. One explanation for this finding could be the presence of other Altiarchaea strains obfuscating the coverage profile, causing misleading bPTR results. Another more biologically relevant explanation was already postulated to explain the slower generation times of the Alti-1 clade compared to its planktonic sister clade Alti-2 (section III.3B), this could also be caused by the higher energetic costs of living in a biofilm, *i.e.*, the necessity of EPS production and living autotrophically (see section III.3), which make replication a costly process. Additionally, since Biofilms are very heterogeneous in terms of microbial activity, due to only a small compartment of the biofilm being exposed to nutrients (Ren *et al.*, 2018; Flemming and Wuertz, 2019), the bPTR value might simply reflect the ratio of dormant to proliferating cells in the biofilm as each inactive or dormant cell would shift the detected value towards one. Simulation studies have also indicated that in pore spaces, slower growing bacterial biofilms can have an advantage over fast-growing biofilms as clogging up the pores and consequently interrupting the nutrient flow can be disadvantageous (Coyte *et al.*, 2017). It is currently unknown where *Ca.* Altiarchaeum GA resides in the Geyser Andernach ecosystem (please see

Supplementary Figure 1 of section III.3 for a schematic of the subsurface environment of the Geyser Andernach). The majority of water in the Geyser Andernach subsurface likely passes through quartz veins, which are probably too large to clog. But water is also expected to move through smaller passageways and those could indeed benefit from slower growing biofilms to avoid clogging. Depending on where Ca. Altiarchaum GA populations reside in the ecosystem, they could thus benefit from slower growth.

In any case, the slow replication values align well with the observed conservation of Alti-1 genetic content across ecosystems and continents (see Figure III.3.5), as mutations incurred during replication are the traditional way of genetic diversification. Horizontal Gene Transfer as an independent, and in biofilms highly effective, way to introduce genetic diversification might then be doubly important.

2.3 Mapping on complete *Ca.* Altiarchaeum GA genome reveals large genome rearrangements

To identify how much of the complete *Ca.* Altiarchaeum GA genome is contained in the short read datasets of all ecosystems with populations of the Alti-1 clade (Figure III.3.4A) and get a more general understanding on similar *Ca.* Altiarchaeum GA is to other Altiarchaeota populations outside of the Central metabolism (Figure III.3.5) and marker genes used to reconstruct phylogenies (Figure III.3.4C), metagenomic reads of all ecosystems were mapped to the *Ca.* Altiarchaeum GA genome (Figure IV.2.3.1).

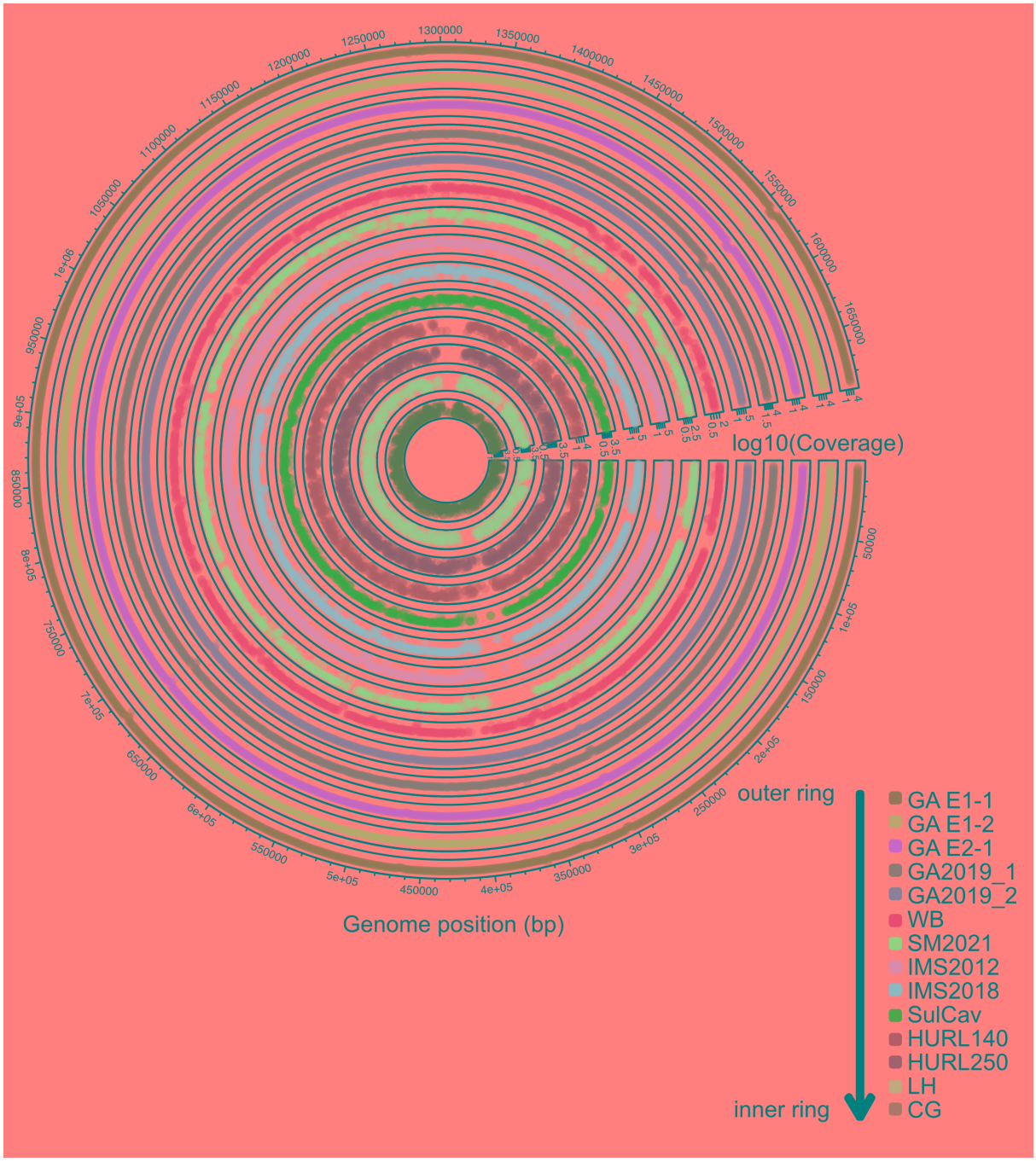

Figure IV.2.3.1: Genomic variability of Altiarchaeota clade Alti-1 Genomes revealed by mapping. Reads of metagenomic datasets were mapped on the complete genome of *Ca.* Altiarchaeum GA, recovered from a sampling campaign in November 2021. Names largely correspond to the names used in section III.3, with the following notable exceptions: GA2019_1, GA2019_2 and IMS2018 are new metagenomic datasets not used in the main manuscript of section III.3 (GA_2019_1 is only used in Figure S12 of the Supplementary Material in this section) and represent additional time points for the respective ecosystems Mühlbacher Schwefelquelle (IMS) and Geyser Andernach (GA), respectively. WB and SM2021 represent two additional ecosystems containing *Ca.* Altiarchaea, namely the cold geyser Wallender Born (WB), also located in the Volcanic Eiffel in Germany, and the Sippenauer Moor (SM2021) adjacent to the IMS site in Regensburg, Germany. For Crystal Geyser (CG), one metagenomic dataset with the highest coverage of Altiarchaeota (CG_2014-23_combo_of_CG_2014-06-09_8_20_14_all_150) was taken as the representative. A similar approach was taken for WB, where

the sample WB02no17 of a time series had the highest *Ca.* Altiarchaea coverage. The ecosystems are as follows: Geyser Andernach in the Volcanic Eiffel in Germany (GA), Wallender Born in the Volcanic Eiffel in Germany (WB), Aquasanta Terme sulfidic cave in Italy (SulCav), Lake Huron on the Canadian/USA border (LH), Crystal Geyser in Utah, USA (CG), the Honorobe Underground Research Station in Japan (HURL), the Sippenauer Moor in Regensburg, Germany (SM2021) and the Islinger Muehlbach in Regensburg, Germany (IMS). The metagenomic read datasets were mapped onto the Altiarchaeum GA genome with Bowtie2 in sensitive mode (Langmead and Salzberg, 2012). The coverage per nucleotide position was calculated with an in-house script and loaded into R (R Core Team, 2008) for data analyses. In R, medians coverages were calculated for 1000 bp window sizes to reduce the impact of coverage discrepancies, with each point in the figure corresponding to the median coverage for one 1000 bp window. Median coverages of 0, *i.e.*, non-covered regions in the respective metagenomic dataset, were not plotted. Data was log10-transformed for plotting and plotting was performed using the circlize R package (Gu *et al.*, 2014). The raw coverage data for each ecosystem, prior aggregation of 1000 bp window steps by median, is supplied in sectionIV_FileS4.

In general, the mapping onto the *Ca.* Altiarchaeum GA genome shows congruent results with the phylogenomic analyses (Figure III.3.4C). The genome coverage of *Ca.* Altiarchaeum GA decreases along the phylogenetic (and geographic) distance, causing progressively more 1000-bp windows with zero median coverage (see sectionIV_FileS5 for a list of the number of empty windows per ecosystem). This causes European sites (outer rings: 1-5 Geyser Andernach in the Volcanic Eiffel, Germany; 6 Wallender Born in the Volcanic Eiffel, Germany; Sippenauer Moor in Regensburg, Germany; 8-9 Islinger Muehlbach in Regensburg, Germany; 10 Aquasanta therme, Italy) to have the greatest similarity to the genome from Geyser Andernach, followed by Asian sites (outer rings: 11-12 HURL140 and HURL250 from the Honorobe Underground Research Facility, Japan). The American sites (13-14 Lake Huron, USA-Canadian Border, and the Crystal Geyser, Utah, USA) have the least coverage of the *Ca.* Altiarchaeum GA genome. This allows for the conclusion, that the clustering by continent observed via phylogenomics is likely also transferrable on a whole genome-basis, as was already indicated by ANI comparisons (Supplementary Figure 7 in section III.3). However, some areas show abnormalities and differences between *Ca.* Altiarchaea populations and will be analyzed in more detail below.

The mapping of Geyser Andernach reads from 2018 and 2019 (Figure IV.2.3.1, outer rings 1-3 and 4-5, respectively) show complete coverage of the *Ca.* Altiarchaeota genome and no noticeable differences incurred during the intermittent year. The only area where the coverage of all GA metagenomes drops to close to zero (and actually reaching zero for GA2019_1 in a single 1000-bp window) can be found in the genomic region 678-682 kbp and contains (based on DIAMOND BLAST at 1E-5 E-value threshold against FunTaxDB; identical to the approach used in section III.3) two (extremely short with ~200 bp) putative Signal

recognition particle (SRP) GTPases, a Phage shock protein A PspA/IM30, a TFR region domain-containing protein and a putative Rhombioid protease GluP. The specific role of GluP is unclear (Rather, 2013) and SRP GTPases act as molecular switches (Shan, Schmid and Zhang, 2009), though the recovered SRP GTPases are much shorter than known functional SRP GTPases (Bange, Wild and Sinning, 2007). The phage shock protein (Psp) system is conserved in many Bacteria and Archaea (Popp *et al.*, 2021) and, contrary to its name, does not only get expressed as a response to filamentous phage infection but also many other types of stressors (Darwin, 2005). It modulates the response to infractions upon cell membrane integrity (Popp *et al.*, 2021). Particularly the Psp system gene might give *Ca.* Altiarchaeum GA enough of a direct benefit to make this HGT viable and thus also transfer the auxiliary genes into the genome.

Some additional areas with coverage abnormalities visible in the mapping of GA short reads to the *Ca.* Altiarchaeum GA genome (Figure IV.2.3.1, outer rings 1-5) are the ranges 301-306 kbp, 333-340 kbp, 461-464 kbp and 1556-1562 kbp. 1556-1562 kbp contains, in addition to unclassified proteins, a CRISPR *cas5* protein and a Type I-B CRISPR associated protein *cas7*. Clustered Regularly Interspaced Palindromic Repeats (CRISPR) systems are a type of adaptive immunity many prokaryotes employ to defend against alien DNA, relying on first assimilating the foreign DNA and then utilizing that known sequence to identify and defend against sequential invasions by this DNA (Nidhi *et al.*, 2021). While this system is primarily regarded as a defense against phages (Nidhi *et al.*, 2021), both self-targeting (Wimmer and Beisel, 2020) as well as targeting of symbionts has been reported for *Ca.* Altiarchaeota Alti-1 in the Crystal Geyser ecosystem [(Esser *et al.*, 2022), in revision for Nature]. For *Ca.* Altiarchaeum GA specifically, no viruses targeting could be identified in prior analyses based on metagenomes from 2018 (Rahlff *et al.*, 2021). Directly adjacent to this sequence region (1548-1551 kbp; 1551-1552 kbp), two conserved CRISPR arrays are located (identified via PILERC-CR v 1.06 (Edgar, 2007)), sharing the same repeat sequence (GTTTCCATACTACATAGTGCGATTTAAAC; see sectionIV_FileS6 for the PILER-CR output file). The drop in coverage for the *cas* protein region indicates that it is only present in a fraction of the *Ca.* Altiarchaeota populations in 2018 and 2019, either via loss of *cas* proteins or a new acquisition in a subpopulation (Shmakov, Utkina, *et al.*, 2020). It is generally rather unlikely that a mobile genetic element with exactly the needed *cas* proteins was integrated into the genome right next to the CRISPR arrays needing them as most genome integration mechanisms (transfection, transposons) integrate at random positions (Soucy, Huang and

Gogarten, 2015). Hence, I postulate that the section containing the *cas* proteins was lost in the majority of *Ca.* Altiarchaeum GA. We cannot detect a further decrease in the metagenomes of 2018 and 2019 as the ratio between the sequence area containing the Cas proteins and the rest of the genome remains constant, but this may simply be caused by the process being too slow. The putative loss of *cas* proteins does not align well to the conservation of the adjacent CRISPR arrays, though CRISPR systems do not necessarily need to have adjacent *cas* proteins (Shmakov, Utkina, *et al.*, 2020).

The region between 461-464 kbp does not contain any proteins while 302-306 kbp is likely a transposon, containing multiple transposase genes, unclassified proteins as well as two copies of an UPF0020 domain-containing protein that is involved in the methylation of a guanosine nucleotide in tRNAs and is thus likely also a recent addition to the *Ca.* Altiarchaeum genome.

The coverage of reads from other sites (WB, SM, IMS, SulCav, HURL) mapped to the *Ca.* Altiarchaeum GA also reveals some further large-scale differences. These are two big gaps in coverage between 330-384 kbp and 1495-1537 kbp. The first of these is not present in WB while the second gap is much smaller than in the other ecosystems. This indicates that these sequence regions may be unique to the Volcanic Eiffel area where both WB and GA are located, and cover a wide variety of gene types. It should be noted that while these are the two large differences to the *Ca.* Altiarchaeum GA, a lot of other 1000 bp windows interspersed between the otherwise covered regions are not covered by reads, revealing more minute differences to *Ca.* Altiarchaeum GA (the size of points in Figure IV.2.3.1 makes them easily visible but causes a lot of overlap between adjacent points, masking the absence of individual points in otherwise well covered regions; this is particularly problematic for inner rings due to increased overlap). The coverage profile of the Japanese HURL site shows an additional gap not visible in European *Ca.* Altiarchaea (GA, WB, SM, IMS and SulCav) at 1278-1353 kbp, containing relatively few open reading frames (n=17), with only three classified proteins that interact with DNA (A peptidase u62 modulator of DNA gyrase, a DNA primase *dnaG* as well as a subtilisin serine protease).

Overall, the mapping of short read datasets of various *Ca.* Altiarchaeum containing metagenomes to the complete *Ca.* Altiarchaeum genome revealed that differences between *Ca.* Altiarchaea populations are much larger than previously estimated based on phylogenomic analyses (Figure III.3.4C), ANI analyses (Supplementary Figure 7 of section III.3) as well as core metabolism analyses (Figure III.3.5) and indicates that there is much more to explore

about the genomic fluidity within the *Ca.* Altiarchaea and the extent to which horizontal gene transfer might contribute to their individual adaptations in their environments.

3. Future perspectives for exploring genomic fluidity in Altiarchaea

Genome-resolved metagenomics has revolutionized our understanding of Earth's environment, vastly extending the biodiversity available for research, and thus generating entire new fields of research [*e.g.*, genome-informed cultivation (Liu *et al.*, 2022), genome-informed microscopy (Rahlff *et al.*, 2021), phylogenomics (Hug *et al.*, 2016)]. It also has revolutionized other OMIC's techniques which rely on reference data, such as proteomics and transcriptomics.

However, this is likely only the beginning of the genome-based revolution, with new technologies, particularly long-read technologies, extending the available research fields even further. The continued development of ONT in particular, *e.g.*, by improving kit chemistry and flow cell architecture, enhances sequencing quality as well as evaluation software. Long-read sequencing technologies have already begun to warp the field of genome-resolved metagenomics. By contrast, environmental short-read metagenomes are very fragmented, rendering assembly, binning and bin curation dauting tasks. In long-read metagenomics, the much more intact assembly already simplifies the binning (and consequently curation) process immensely. Depending on how far this technology progresses, it might even make traditional binning and bin curation redundant, if (near-)complete genomes can consistently be assembled directly from the reads (which was the case for the *Ca.* Altiarchaeum GA genome investigated in detail in section IV.2. Both ONT as well as Pacific Biosciences can also identify base modifications and thus could open up the research field of environmental epigenetics, which has so far mainly been explored in humans and cultured organisms through specialized technologies, as an additional dimension. Indeed, other polymers such as proteins can also be sequenced via ONT, though the 20 different possible amino acids (with additional dimensionality via Post-Translational Modifications) make their sequencing much more complex.

ONT-based detection of epigenetic elements in particular might be very interesting in multiple regards to the research questions tackled within this thesis. First, as of today, the only epigenetic elements known in Archaea are chromatin protein modifications, which have been identified in *Sulfolobus (see* (Blum and Payne, 2019) for a review on epigenetics in Archaea).

Thus, more large-scale surveys into possible DNA modifications might be able to either provide more broad evidence for there being no DNA modification systems, such as Adenine methylation, in Archaea or disproof this current view. Second, adenine methylation in Bacteria has been shown to affect replication initiation, making it relevant for the replication component of this thesis, and also DNA repair (Willbanks *et al.*, 2016), which could affect genome variability and fluidity (though rather in Bacteria than Archaea). One further open question that might be explorable using model systems, might be whether hereditary epigenetics, i.e., the transfer of epigenetic information to one's offspring, plays a role in prokaryotes. If it was identified to play a role, then the epigenetic element in horizontal gene transfer might be another layer of research.

A related research topic could also be tRNA modifications, which can also be detected via ONT (Thomas *et al.*, 2021). The mapping survey done in section IV.2.3.3 putatively identified a transposon in the 302-306 kbp region of the *Ca.* Altiarchaeum GA genome. This transposon contained two copies of a protein involved in guanine methylation in tRNAs. While tRNA guanine methylation in Archaea is a known mechanism (Armengaud *et al.*, 2004), this could provide the opportunity to investigate how the introduction of new methylation capabilities might influence the organism. As a first step, this would require an in-depth analysis of the other tRNA methylation capabilities in *Ca.* Altiarchaeum GA to make sure that this type of tRNA methylation is a newly acquired function. In the best case scenario, i.e., the transferred genes are active based on methylation patterns of tRNA's and methylation events can be unambiguously assigned to their activity due to no homologs being present in the rest of the genome, this might actually show an example of a HGT event influencing not just the genome of the recipient but also their epigenome.

Some more immediate follow-up analyses could be a more in-depth biogeographic analysis of *Ca.* Altiarchaea, using Ancestral Sequence Reconstruction to try to test whether the proposed time points of divergence, e.g., the splitting up of the Pangean supercontinent, are reflected in their sequence divergence. Since a complete Ca. Altiarchaeum GA genome is now available, one could identify potential horizontal gene transfers across the whole genome and see whether specific horizontal gene transfers, *e.g.*, from specific bacterial phyla, are prevalent, and thus reconstruct a horizontal gene transfer interaction map. Single Nucleotide Polymorphism analyses to detect strain variants of *Ca.* Altiarchaeum GA over the years could also be very interesting to see whether shifts in their ratios have occurred from 2018 to 2021.

The effect of genetic drift, i.e., SNP allele frequency fluctuations across a population with potential to become permanent, on *Ca.* Altiarchaeum GA could be explored via identification of synonymous, non-synonymous and missense SNPs, to evaluate the effect of normal replicative evolution on *Ca.* Altiarchaea. (Orsi *et al.*, 2021) could act as a guide for this type of analyses.

In the complete *Ca.* Altiarchaeum GA genome, two maintained CRISPR arrays were detected but the adjacent Cas proteins are indicated have been lost in a large proportion of the community. This abnormality also needs more in-depth investigation due to their being conflicting signals regarding the CRISPR system activity. On the one hand, the CRISPR arrays are maintained, i.e., the repeats do not show any SNPs, indicating that it might be active. On the other hand, adjacent Cas proteins have most likely been lost in most of the population, which should in principle signal that the CRISPR system is not active. Thus, in depth CRISPR analyses akin to Esser *et. al.* could be performed (Esser *et al.*, 2022). This would hopefully also elucidate what these CRISPR systems target, as prior analyses of GA (Rahlff *et al.*, 2021) were not able to find viruses targeting *Ca.* Altiarchaeum GA and consequently the purpose of these arrays is unclear. Another potential area of interest might be the transcription machinery of Ca. Altiarchaeum GA, *e.g.*, small regulatory RNAs and transcription factors, or focusing in on the *hami*-encoding region on the genome to see whether details about its assemblage can be gleaned. One might also explore the genome with the goal of trying to cultivate *Ca.* Altiarchaeum GA (which is the ultimate goal of the BMBF-funded MultiKulti Project of the Probst lab).

These are the ideas I am currently most exited about to further explore regarding the genomic fluidity of *Ca.* Altiarchaeum specifically but also prokaryotes as a whole in the deep biosphere. The availability of a full genome really opens up the amount of available research directions immensely, with a close to unlimited amount of potential research objectives becoming available, and highlights why long-read metagenomics is so important for genome-resolved metagenomics.

V. Zusammenfassung

Archaeen wurden für lange Zeit als nur Extremumgebungen, wie Vulkane, heiße Quellen oder Salzseen, besiedelnde Organismen gehalten, was sich in der Mehrheit der verfügbaren Isolate.

widerspiegelt. Jedoch haben kultivierungsunabhängige Methoden wie Metagenomik gezeigt, dass Archaeen ubiquitär verbreitet, aber häufig deutlich weniger abundant als Bakterien sind. Der terrestrische Untergrund ist eine der bedeutendsten Umgebungen, der durch diese Methoden erschlossen wurde, da er zugleich sehr wenig erforscht und jedoch Hochrechnungen zu Folge 25 % der Organismen auf der Erde beinhaltet. In dieser Umgebung lebende Organismen sind gekennzeichnet von einem Leben nahe am energetischen Minimum sowie mit wenig Ausbreitungsmöglichkeiten. Wie sich Archaeen an diese Extrembedingungen anpassen und inwiefern horizontaler Gentransfer eine Rolle dabei spielt, ist gänzlich unbekannt.

In dieser Dissertation hatte ich das Ziel, qualitativ hochwertige archaeelle Genome von unkultivierten Altiarchaeota zu generieren, um deren Anpassung an ihre Ökosysteme untersuchen zu können. Diese Archaeen dominieren ihre moderat-temperierten Umgebungen, daher besteht besonderes Interesse, zu verstehen, wie sie sich durch ihre Adaptionen einen Vorteil verschaffen. Zunächst haben wir einen Arbeitsablauf zur Erstellung archaeeller Genome aus Metagenomen etabliert. Ein Schritt, der bei dabei oft vernachlässigt wird, ist die Kuration von Genomen, da dieser Schritt manuell und mit begrenzter Softwareunterstützung durchgeführt wird. Daher haben wir die Genomkurationssoftware uBin entwickelt. uBin ermöglicht eine einfache, GUI-basierte Kuration von Genomen. Die Verwendung von uBin verbesserte die Genomqualität von 78.9 % Genomen im öffentlich erhältlichen CAMI-Datensatz. Zuletzt haben wir den CO_2-getriebenen Kaltwassergeysir Andernach, der die höchste Wasserfontäne weltweit hat und von *Ca.* Altiarchaeum dominiert wird, metagenomisch charakterisiert und hunderte prokaryotische Genome von diesem Ökosystem und anderen Umgebungen im tiefen terrestrischen Untergrund rekonstruiert und kuriert. Um das Wachstumspotential von Bakterien im tiefen Untergrund zu beziffern, verglichen wir das Ökosystem Geysir Andernach mit 16 weiteren Ökosystemen aus dem terrestrischen Untergrund. Diese Ökosysteme überspannen einen Bereich von oberflächennahen Höhlen bis zu 3 km Tiefe. Wir identifizierten einen Zusammenhang zwischen minimaler Generationszeit und Ökosystemtiefe in Bakterien, *i.e.*, je tiefer ihr Vorkommen, desto schneller konnten sie sich teilen. Gleichzeitig replizierten sie aber weniger zum Probenahmezeitpunkt, je tiefer ihr Ökosystem. Dies deuten wir als Adaptation an im terrestrischen Untergrund herrschende schwankende Nährstoffverfügbarkeiten.

Ein Vergleich neu generierter und bereits verfügbarer *Ca.* Altiarchaea Genome zeigte eine starke Konservierung der Genome von *Ca.* Altiarchaea der Klade Alti-1. Wir deuten die

biogeographische, nach Kontinent gruppierte, Konservierung der Genome als Hinweis für eine Ausbreitung durch Plattentektonik. Das genetische Repertoire zeigte eine starke Konservierung des Kernmetabolismus und Variationen der peripheren Gene, wie Peptidasen, wovon einige womöglich aus horizontalem Gentransfer mit Bakterien abstammen. Auf diesen Analysen aufbauend nutzte ich ein neu konstruiertes zirkuläres *Ca.* Altiarchaeum GA Genom als Referenz, um Regionen von genetischer Variabilität zwischen *Ca.* Altiarchaea Populationen zu identifizieren. Die Ergebnisse stehen in Einklang mit vorherigen biogeographischen Analysen, zeigten jedoch auch, dass *Ca.* Altiarchaea deutlich mehr Genomvariationen haben als angenommen. Einige dieser Bereiche genetischer Variation wurden vermutlich von horizontalem Gentransfer verursacht, wie Transposasegene in diesen Bereichen belegen. Daraus schlossen wir, dass horizontaler Gentransfer die sonst sehr langsame Evolution in diesem Phylum mitigieren könnte.

Zusammengefasst zeigt diese Arbeit einen Arbeitsablauf für die Gewinnung von archaeellen Genomen aus Metagenomen auf, zusammen mit der neuen, leicht verwendbaren uBin Software zur Genomkuration, um die Genomqualität zu gewährleisten. Darüber hinaus gibt die Arbeit wertvolle Einblicke in die genetische Diversität von einem der wenigen dominanten Archaeen in moderat temperierten tiefen Biosphäre Umgebungen.

VI. Bibliography

REFERENCES

A

Adam, P.S., Borrel, G. and Gribaldo, S. (2019) ‘An archaeal origin of the Wood–Ljungdahl H 4 MPT branch and the emergence of bacterial methylotrophy’, *Nature Microbiology*, 4(12), pp. 2155–2163. doi:10.1038/s41564-019-0534-2.

Aiuppa, A. *et al.* (2019) ‘CO 2 flux emissions from the Earth’s most actively degassing volcanoes, 2005–2015’, *Scientific Reports*, 9(1), p. 5442. doi:10.1038/s41598-019-41901-y.

Albertsen, M. *et al.* (2013) ‘Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes’, *Nature Biotechnology*, 31(6), pp. 533–538. doi:10.1038/nbt.2579.

Almeida, A. *et al.* (2019) ‘A new genomic blueprint of the human gut microbiota’, *Nature*, 568(7753), pp. 499–504. doi:10.1038/s41586-019-0965-1.

Alneberg, J. *et al.* (2014) ‘Binning metagenomic contigs by coverage and composition’, *Nature Methods*, 11(11), pp. 1144–1146. doi:10.1038/nmeth.3103.

Altfeld, E. (1913) ‘Die physikalischen Grundlagen des intermittierenden Kohlensäuresprudels zu Namedy bei Andernach a. Rh.’ Dissertation, Universität Marburg.

Amarasinghe, S.L. *et al.* (2020) ‘Opportunities and challenges in long-read sequencing data analysis’, *Genome Biology*, 21(1), p. 30. doi:10.1186/s13059-020-1935-5.

Anantharaman, K. *et al.* (2016) ‘Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system’, *Nature Communications*, 7(1), pp. 1–11. doi:10.1038/ncomms13219.

Anisimova, M. *et al.* (2011) ‘Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes’, *Systematic Biology*, 60(5), pp. 685–699. doi:10.1093/sysbio/syr041.

Antipov, D. *et al.* (2016) ‘plasmidSPAdes: assembling plasmids from whole genome sequencing data’, *Bioinformatics*, 32(22), pp. 3380–3387. doi:10.1093/bioinformatics/btw493.

Antipov, D. *et al.* (2020) ‘MetaviralSPAdes: assembly of viruses from metagenomic data’, *Bioinformatics*, 36(14), pp. 4126–4129. doi:10.1093/bioinformatics/btaa490.

Armengaud, J. *et al.* (2004) ‘N2-Methylation of Guanosine at Position 10 in tRNA Is Catalyzed by a THUMP Domain-containing, S-Adenosylmethionine-dependent Methyltransferase, Conserved in Archaea and Eukaryota*’, *Journal of Biological Chemistry*, 279(35), pp. 37142–37152. doi:10.1074/jbc.M403845200.

Arnold, B.J., Huang, I.-T. and Hanage, W.P. (2022) ‘Horizontal gene transfer and adaptive evolution in bacteria’, *Nature Reviews Microbiology*, 20(4), pp. 206–218. doi:10.1038/s41579-021-00650-4.

Asnicar, F. *et al.* (2020) ‘Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0’, *Nature Communications*, 11(1), p. 2500. doi:10.1038/s41467-020-16366-7.

Ausiannikava, D. and Allers, T. (2017) ‘Diversity of DNA Replication in the Archaea’, *Genes*, 8(2), p. 56. doi:10.3390/genes8020056.

B

Baker, B.J. *et al.* (2006) ‘Lineages of Acidophilic Archaea Revealed by Community Genomic Analysis’, *Science*, 314(5807), pp. 1933–1935. doi:10.1126/science.1132690.

Ballenghien, M., Faivre, N. and Galtier, N. (2017) ‘Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions’, *BMC Biology*, 15(1), p. 25. doi:10.1186/s12915-017-0366-6.

Bange, G., Wild, K. and Sinning, I. (2007) ‘Protein Translocation: Checkpoint Role for SRP GTPase Activation’, *Current Biology*, 17(22), pp. R980–R982. doi:10.1016/j.cub.2007.09.041.

Bankevich, A. *et al.* (2012) 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing', *Journal of Computational Biology*, 19(5), pp. 455–477. doi:10.1089/cmb.2012.0021.

Becraft, E.D. *et al.* (2021) 'Evolutionary stasis of a deep subsurface microbial lineage', *The ISME Journal* 15(10), pp. 2830-2842. doi:10.1038/s41396-021-00965-3.

Berg, I.A. *et al.* (2010) 'Autotrophic carbon fixation in archaea', *Nature Reviews Microbiology*, 8(6), p. 447. doi:10.1038/nrmicro2365.

Berg, I.A. (2011) 'Ecological Aspects of the Distribution of Different Autotrophic CO2 Fixation Pathways', *Applied and Environmental Microbiology*, 77(6), pp. 1925–1936. doi:10.1128/AEM.02473-10.

Beulig, F. *et al.* (2015) 'Carbon flow from volcanic CO2 into soil microbial communities of a wetland mofette', *The ISME Journal*, 9(3), pp. 746–759. doi:10.1038/ismej.2014.148.

Beulig, F. *et al.* (2016) 'Altered carbon turnover processes and microbiomes in soils under long-term extremely high CO_2 exposure', *Nature Microbiology*, 1(2), pp. 1–10. doi:10.1038/nmicrobiol.2015.25.

Bird, J.T. *et al.* (2016) 'Culture Independent Genomic Comparisons Reveal Environmental Adaptations for Altiarchaeales', *Frontiers in Microbiology*, 7(1221). doi:10.3389/fmicb.2016.01221.

Blum, P. and Payne, S. (2019) 'Evidence of an Epigenetics System in Archaea', *Epigenetics Insights*, 12, p. 2516865719865280. doi:10.1177/2516865719865280.

Bornemann, T.L.V. *et al.* (2020) 'uBin – a manual refining tool for metagenomic bins designed for educational purposes', *bioRxiv*, p. 2020.07.15.204776. doi:10.1101/2020.07.15.204776.

Bowers, Robert M. *et al.* (2017) 'Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea', *Nature Biotechnology*, 35(8), pp. 725–731. doi:10.1038/nbt.3893.

Boyd, E.S. *et al.* (2014) 'Chemolithotrophic Primary Production in a Subglacial Ecosystem', *Applied and Environmental Microbiology*, 80(19), pp. 6146–6153. doi:10.1128/AEM.01956-14.

Braakman, R. and Smith, E. (2012) 'The Emergence and Early Evolution of Biological Carbon-Fixation', *PLoS Computational Biology*, 8(4), p. e1002455. doi:10.1371/journal.pcbi.1002455.

Bräuer, K. *et al.* (2013) 'Indications for the existence of different magmatic reservoirs beneath the Eifel area (Germany): A multi-isotope (C, N, He, Ne, Ar) approach', *Chemical Geology*, 356, pp. 193–208. doi:10.1016/j.chemgeo.2013.08.013.

Breitwieser, F., Lu, J. and Salzberg, S. (2019) 'A review of methods and databases for metagenomic classification and assembly.', *Briefings in Bioinformatics*, 20(4), pp. 1125–1136. doi:10.1093/bib/bbx120.

Brown, C.T. *et al.* (2015) 'Unusual biology across a group comprising more than 15% of domain Bacteria', *Nature*, 523(7559), p. 208. doi:10.1038/nature14486.

Brown, C.T. *et al.* (2016) 'Measurement of bacterial replication rates in microbial communities', *Nature Biotechnology*, 34(12), pp. 1256–1263. doi:10.1038/nbt.3704.

Buchfink, B., Xie, C. and Huson, D.H. (2015) 'Fast and sensitive protein alignment using DIAMOND', *Nature Methods*, 12(1), pp. 59–60. doi:10.1038/nmeth.3176.

Bushnell (2021) *BBmaps, https://sourceforge.net/projects/bbmap/*.

C

Can, M., Armstrong, F.A. and Ragsdale, S.W. (2014) 'Structure, Function, and Mechanism of the Nickel Metalloenzymes, CO Dehydrogenase, and Acetyl-CoA Synthase', *Chemical Reviews*, 114(8), pp. 4149–4174. doi:10.1021/cr400461p.

Caracausi, A. and Paternoster, M. (2015) 'Radiogenic helium degassing and rock fracturing: A case study of the southern Apennines active tectonic region', *Journal of Geophysical Research: Solid Earth*, 120(4), pp. 2200–2211. doi:10.1002/2014JB011462.

Castelle, C.J. *et al.* (2013) 'Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment', *Nature Communications*, 4(2120). doi:10.1038/ncomms3120.

Castelle, C.J. *et al.* (2015) 'Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling', *Current Biology*, 25(6), pp. 690–701. doi:10.1016/j.cub.2015.01.014.

Castelle, C.J. and Banfield, J.F. (2018) 'Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life', *Cell*, 172(6), pp. 1181–1197. doi:10.1016/j.cell.2018.02.016.

Cerretti, D.P. *et al.* (1983) 'The spc ribosomal protein operon of Escherichia coli: sequence and cotranscription of the ribosomal protein genes and a protein export gene.', *Nucleic Acids Research*, 11(9), pp. 2599–2616.

Chaumeil, P.-A. *et al.* (2020) 'GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database', *Bioinformatics*, 36(6), pp. 1925–1927. doi:10.1093/bioinformatics/btz848.

Chen, L.-X. *et al.* (2020) 'Accurate and complete genomes from metagenomes', *Genome Research*, 30(3), pp. 315–333. doi:10.1101/gr.258640.119.

Claramunt, S. *et al.* (2012) 'High dispersal ability inhibits speciation in a continental radiation of passerine birds', *Proceedings of the Royal Society B: Biological Sciences*, 279(1733), pp. 1567–1574. doi:10.1098/rspb.2011.1922.

Cocks, L.R.M. and Torsvik, T.H. (2005) 'Baltica from the late Precambrian to mid-Palaeozoic times: The gain and loss of a terrane's identity', *Earth-Science Reviews*, 72(1–2), pp. 39–66. doi:10.1016/j.earscirev.2005.04.001.

Coker, J.A. *et al.* (2009) 'Multiple Replication Origins of Halobacterium sp. Strain NRC-1: Properties of the Conserved orc7-Dependent oriC1', *Journal of Bacteriology*, 191(16), pp. 5253–5261. doi:10.1128/JB.00210-09.

Corliss, J.B. *et al.* (1979) 'Submarine Thermal Springs on the Galápagos Rift', *Science*, 203(4385), pp. 1073–1083. doi:10.1126/science.203.4385.1073.

Cornet, L. *et al.* (2018) 'Consensus assessment of the contamination level of publicly available cyanobacterial genomes', *PLoS ONE*, 13(7), p. e0200323. doi:10.1371/journal.pone.0200323.

Cornet, L. and Baurain, D. (2022) 'Contamination detection in genomic data: more is not enough', *Genome Biology*, 23, p. 60. doi:10.1186/s13059-022-02619-9.

Couvin, D. *et al.* (2018) 'CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins', *Nucleic Acids Research*, 46(W1), pp. W246–W251. doi:10.1093/nar/gky425.

Coyte, K.Z. *et al.* (2017) 'Microbial competition in porous environments can select against rapid biofilm growth', *Proceedings of the National Academy of Sciences*, 114(2), pp. E161–E170. doi:10.1073/pnas.1525228113.

Criscuolo, A. and Gribaldo, S. (2010) 'BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments', *BMC Evolutionary Biology*, 10, p. 210. doi:10.1186/1471-2148-10-210.

Custer, G.F., Bresciani, L. and Dini-Andreote, F. (2022) 'Ecological and Evolutionary Implications of Microbial Dispersal', *Frontiers in Microbiology*, 13(855859). Available at: https://www.frontiersin.org/article/10.3389/fmicb.2022.855859 (Accessed: 11 May 2022).

D

Dai, X. *et al.* (2016) 'Genome Sequencing of Sulfolobus sp. A20 from Costa Rica and Comparative Analyses of the Putative Pathways of Carbon, Nitrogen, and Sulfur Metabolism in Various Sulfolobus Strains', *Frontiers in Microbiology*, 7, p. 1902. doi:10.3389/fmicb.2016.01902.

Darling, A.E. *et al.* (2014) 'PhyloSift: phylogenetic analysis of genomes and metagenomes', *PeerJ*, 2, p. e243. doi:10.7717/peerj.243.

Darwin, A.J. (2005) 'The phage-shock-protein response', *Molecular Microbiology*, 57(3), pp. 621–628. doi:10.1111/j.1365-2958.2005.04694.x.

Davies, D. (2003) 'Understanding biofilm resistance to antibacterial agents', *Nature Reviews Drug Discovery*, 2(2), pp. 114–122. doi:10.1038/nrd1008.

Dèzes, P., Schmid, S.M. and Ziegler, P.A. (2004) 'Evolution of the European Cenozoic Rift System: interaction of the Alpine and Pyrenean orogens with their foreland lithosphere', *Tectonophysics*, 389(1), pp. 1–33. doi:10.1016/j.tecto.2004.06.011.

D'Hondt, S., Rutherford, S. and Spivack, A.J. (2002) 'Metabolic Activity of Subsurface Life in Deep-Sea Sediments', *Science*, 295(5562), pp. 2067–2070. doi:10.1126/science.1064878.

Dick, G.J. *et al.* (2009) 'Community-wide analysis of microbial genome sequence signatures', *Genome Biology*, 10(8), p. R85. doi:10.1186/gb-2009-10-8-r85.

Ding, J. *et al.* (2011) 'A novel acidophilic, thermophilic iron and sulfur-oxidizing archaeon isolated from a hot spring of tengchong, yunnan, China', *Brazilian Journal of Microbiology*, 42(2), pp. 514–525. doi:10.1590/S1517-83822011000200016.

Ding, J. *et al.* (2017) 'Microbial Community Structure of Deep-sea Hydrothermal Vents on the Ultraslow Spreading Southwest Indian Ridge', *Frontiers in Microbiology*, 8(1012). doi:10.3389/fmicb.2017.01012.

Dombrowski, N. *et al.* (2020) 'Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution', *Nature Communications*, 11(1), p. 3939. doi:10.1038/s41467-020-17408-w.

Drake, H.L., Gössner, A.S. and Daniel, S.L. (2008) 'Old acetogens, new light', *Annals of the New York Academy of Sciences*, 1125, pp. 100–128. doi:10.1196/annals.1419.016.

E

Earl, D. *et al.* (2011) 'Assemblathon 1: A competitive assessment of de novo short read assembly methods', *Genome Research*, 21(12), pp. 2224–2241. doi:10.1101/gr.126599.111.

Eddy, S.R. (2011) 'Accelerated Profile HMM Searches', *PLoS computational biology*, 7(10), p. e1002195. doi:10.1371/journal.pcbi.1002195.

Edgar, R.C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32(5), pp. 1792–1797. doi:10.1093/nar/gkh340.

Edgar, R.C. (2007) 'PILER-CR: Fast and accurate identification of CRISPR repeats', *BMC Bioinformatics*, 8(1), p. 18. doi:10.1186/1471-2105-8-18.

Ellis, R.J. (1979) 'The most abundant protein in the world', *Trends in Biochemical Sciences*, 4(11), pp. 241–244. doi:10.1016/0968-0004(79)90212-3.

Eloe-Fadrosh, E.A., Paez-Espino, D., *et al.* (2016) 'Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs', *Nature Communications*, 7(10476). doi:10.1038/ncomms10476.

Eloe-Fadrosh, E.A., Ivanova, N.N., *et al.* (2016) 'Metagenomics uncovers gaps in amplicon-based detection of microbial diversity', *Nature Microbiology*, 1(4), pp. 1–4. doi:10.1038/nmicrobiol.2015.32.

Emerson, J.B. *et al.* (2016) 'Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla', *Environmental Microbiology*, 18(6), pp. 1686–1703. doi:10.1111/1462-2920.12817.

Eren, A.M. *et al.* (2015) 'Anvi'o: an advanced analysis and visualization platform for 'omics data', *PeerJ*, 3, p. e1319. doi:10.7717/peerj.1319.

Eren, A.M. *et al.* (2021) ‘Community-led, integrated, reproducible multi-omics with anvi’o’, *Nature Microbiology*, 6(1), pp. 3–6. doi:10.1038/s41564-020-00834-3.

Espejo, R.T. and Plaza, N. (2018) ‘Multiple Ribosomal RNA Operons in Bacteria; Their Concerted Evolution and Potential Consequences on the Rate of Evolution of Their 16S rRNA’, *Frontiers in Microbiology*, 9(1232). Available at: https://www.frontiersin.org/article/10.3389/fmicb.2018.01232 (Accessed: 11 May 2022).

Esser, S. *et al.* (2022) ‘CRISPR-mediated resistance to archaeal episymbionts’, *In revision at Nature* [Preprint].

F

Falkowski, P.G., Fenchel, T. and Delong, E.F. (2008) ‘The microbial engines that drive Earth’s biogeochemical cycles’, *Science (New York, N.Y.)*, 320(5879), pp. 1034–1039. doi:10.1126/science.1153213.

Figueroa, I.A. *et al.* (2018) ‘Metagenomics-guided analysis of microbial chemolithoautotrophic phosphite oxidation yields evidence of a seventh natural CO_2 fixation pathway’, *Proceedings of the National Academy of Sciences*, 115(1), pp. E92–E101. doi:10.1073/pnas.1715549114.

Finkel, S.E. (2006) ‘Long-term survival during stationary phase: evolution and the GASP phenotype’, *Nature Reviews Microbiology*, 4(2), pp. 113–120. doi:10.1038/nrmicro1340.

Flemming, H.-C. and Wuertz, S. (2019) ‘Bacteria and archaea on Earth and their abundance in biofilms’, *Nature Reviews Microbiology*, 17(4), pp. 247–260. doi:10.1038/s41579-019-0158-9.

Fleurier, S. *et al.* (2022) ‘rRNA operon multiplicity as a bacterial genome stability insurance policy’, *Nucleic Acids Research*, p. gkac332. doi:10.1093/nar/gkac332.

Fox, G.E. and Woese, C.R. (1975) ‘The architecture of 5S rRNA and its relation to function’, *Journal of Molecular Evolution*, 6(1), pp. 61–76. doi:10.1007/BF01732674.

Frerichs, J. *et al.* (2013) ‘Microbial community changes at a terrestrial volcanic CO_2 vent induced by soil acidification and anaerobic microhabitats within the soil column’, *FEMS Microbiology Ecology*, 84(1), pp. 60–74. doi:10.1111/1574-6941.12040.

Fullerton, K.M. *et al.* (2019) *Plate tectonics drive deep biosphere microbial community composition*. preprint. EarthArXiv. doi:10.31223/osf.io/gyr7n.

G

Galambos, D. *et al.* (2019) ‘Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents’, *Environmental Microbiology*, 21(11), pp. 4395–4410. doi:10.1111/1462-2920.14806.

Garcia, A.K., Cavanaugh, C.M. and Kacar, B. (2021) 'The curious consistency of carbon biosignatures over billions of years of Earth-life coevolution', *The ISME Journal*, 15(8), pp. 2183–2194. doi:10.1038/s41396-021-00971-5.

Gilfillan, S.M.V. *et al.* (2019) 'Noble gases confirm plume-related mantle degassing beneath Southern Africa', *Nature Communications*, 10(1), pp. 1–7. doi:10.1038/s41467-019-12944-6.

Gold, T. (1992) 'The deep, hot biosphere.', *Proceedings of the National Academy of Sciences*, 89(13), pp. 6045–6049. doi:10.1073/pnas.89.13.6045.

Graham, E.D., Heidelberg, J.F. and Tully, B.J. (2017) 'BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation', *PeerJ*, 5(e3035). doi:10.7717/peerj.3035.

Gruber-Vodicka, H.R., Seah, B.K.B. and Pruesse, E. (2020) 'phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes', *mSystems*, 5(5). doi:10.1128/mSystems.00920-20.

Gu, Z. *et al.* (2014) 'circlize implements and enhances circular visualization in R', *Bioinformatics*, 30(19), pp. 2811–2812. doi:10.1093/bioinformatics/btu393.

Guindon, S. *et al.* (2010) 'New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0', *Systematic Biology*, 59(3), pp. 307–321. doi:10.1093/sysbio/syq010.

Gunde-Cimerman, N., Oren, A. and Plemenitaš, A. (eds) (2005) *Adaptation to Life at High Salt Concentrations in Archaea, Bacteria, and Eukarya*. Dordrecht: Springer Netherlands (Cellular Origin, Life in Extreme Habitats and Astrobiology). doi:10.1007/1-4020-3633-7.

Guo, J. *et al.* (2021) 'VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses', *Microbiome*, 9(1), p. 37. doi:10.1186/s40168-020-00990-y.

Gutiérrez-Preciado, A. *et al.* (2018) 'Functional shifts in microbial mats recapitulate early Earth metabolic transitions', *Nature ecology & evolution*, 2(11), pp. 1700–1708. doi:10.1038/s41559-018-0683-3.

H

Hackl, T. *et al.* (2021) 'proovframe: frameshift-correction for long-read (meta)genomics'. bioRxiv, p. 2021.08.23.457338. doi:10.1101/2021.08.23.457338.

Hamilton, T.L. *et al.* (2015) 'Metagenomic insights into S(0) precipitation in a terrestrial subsurface lithoautotrophic ecosystem', *Frontiers in Microbiology*, 5(756). doi:10.3389/fmicb.2014.00756.

Handelsman, J. *et al.* (1998) 'Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products', *Chemistry & Biology*, 5(10), pp. R245–R249. doi:10.1016/S1074-5521(98)90108-9.

Hanson, C.A. *et al.* (2012) 'Beyond biogeographic patterns: processes shaping the microbial landscape', *Nature Reviews. Microbiology*, 10(7), pp. 497–506. doi:10.1038/nrmicro2795.

Hausner, M. and Wuertz, S. (1999) 'High rates of conjugation in bacterial biofilms as determined by quantitative in situ analysis', *Applied and Environmental Microbiology*, 65(8), pp. 3710–3713. doi:10.1128/AEM.65.8.3710-3713.1999.

Haynes, W. (2013) 'Tukey's Test', in Dubitzky, W. et al. (eds) *Encyclopedia of Systems Biology*. New York, NY: Springer New York, pp. 2303–2304.

Hedlund, B.P. *et al.* (2022) 'SeqCode, a nomenclatural code for prokaryotes described from sequence data'. preprint at https://disc-genomics.uibk.ac.at/seqcode//files/Hedlund_et_al.pdf

Hedrick, D.B. *et al.* (1992) 'In situ microbial ecology of hydrothermal vent sediments', *FEMS Microbiology Letters*, 101(1), pp. 1–10. doi:10.1016/0378-1097(92)90691-G.

Henneberger, R. *et al.* (2006) 'New Insights into the Lifestyle of the Cold-Loving SM1 Euryarchaeon: Natural Growth as a Monospecies Biofilm in the Subsurface', *Applied and Environmental Microbiology*, 72(1), pp. 192–199. doi:10.1128/AEM.72.1.192-199.2006.

Hernsdorf, A.W. *et al.* (2017) 'Potential for microbial H 2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments', *The ISME Journal*, 11(8), pp. 1915–1929. doi:10.1038/ismej.2017.39.

Hoang, D.T. *et al.* (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution*, 35(2), pp. 518–522. doi:10.1093/molbev/msx281.

Hu, X. and Friedberg, I. (2020) 'Identifying Core Operons in Metagenomic Data', *bioRxiv*, p. 2019.12.20.885269. doi:10.1101/2019.12.20.885269.

Huber, H. *et al.* (2002) 'A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont', *Nature*, 417(6884), pp. 63–67. doi:10.1038/417063a.

Hug, L.A. *et al.* (2015) 'Aquifer environment selects for microbial species cohorts in sediment and groundwater', *The ISME Journal*, 9(8), pp. 1846–1856. doi:10.1038/ismej.2015.2.

Hug, L.A. *et al.* (2016) 'A new view of the tree of life', *Nature Microbiology*, 1(5), p. 16048. doi:10.1038/nmicrobiol.2016.48.

Huson, D.H. *et al.* (2007) 'Dendroscope: An interactive viewer for large phylogenetic trees', *BMC Bioinformatics*, 8, p. 460. doi:10.1186/1471-2105-8-460.

Hyatt, D. *et al.* (2010) 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics*, 11, p. 119. doi:10.1186/1471-2105-11-119.

I

Iverson, V. *et al.* (2012) 'Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota', *Science*, 335(6068), pp. 587–590. doi:10.1126/science.1212665.

J

Jain, C. *et al.* (2018) 'High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries', *Nature Communications*, 9(5114). doi:10.1038/s41467-018-07641-9.

JN Fass, N.J. (2011) *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files*. https://github.com/najoshi/sickle.

Joo, S., Park, P. and Park, S. (2019) 'Applicability of propidium monoazide (PMA) for discrimination between living and dead phytoplankton cells', *PLoS ONE*, 14(6). doi:10.1371/journal.pone.0218924.

K

Kadnikov, V.V. *et al.* (2018) 'A metagenomic window into the 2-km-deep terrestrial subsurface aquifer revealed multiple pathways of organic matter decomposition', *FEMS microbiology ecology*, 94(10). doi:10.1093/femsec/fiy152.

Kairesalo, T. *et al.* (1995) 'The role of bacteria in the nutrient exchange between sediment and water in a flow-through system', *Microbial Ecology*, 29(2), pp. 129–144. doi:10.1007/BF00167160.

Kallmeyer, J. *et al.* (2012) 'Global distribution of microbial abundance and biomass in subseafloor sediment', *Proceedings of the National Academy of Sciences of the United States of America*, 109(40), pp. 16213–16216. doi:10.1073/pnas.1203849109.

Kalyaanamoorthy, S. *et al.* (2017) 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods*, 14(6), pp. 587–589. doi:10.1038/nmeth.4285.

Kanehisa, M. and Sato, Y. (2020) 'KEGG Mapper for inferring cellular functions from protein sequences', *Protein Science: A Publication of the Protein Society*, 29(1), pp. 28–35. doi:10.1002/pro.3711.

Kanehisa, M., Sato, Y. and Morishima, K. (2016) 'BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences', *Journal of Molecular Biology*, 428(4), pp. 726–731. doi:10.1016/j.jmb.2015.11.006.

Kang, D.D. *et al.* (2015) 'MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities', *PeerJ*, 3, p. e1165. doi:10.7717/peerj.1165.

Kang, D.D. *et al.* (2019) 'MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies', *PeerJ*, 7, p. e7359. doi:10.7717/peerj.7359.

Kapli, P., Yang, Z. and Telford, M.J. (2020) 'Phylogenetic tree building in the genomic age', *Nature Reviews Genetics*, 21(7), pp. 428–444. doi:10.1038/s41576-020-0233-0.

Keeling, P.J. and Palmer, J.D. (2008) 'Horizontal gene transfer in eukaryotic evolution', *Nature Reviews Genetics*, 9(8), pp. 605–618. doi:10.1038/nrg2386.

Kirkpatrick, J.B., Walsh, E.A. and D'Hondt, S. (2019) 'Microbial Selection and Survival in Subseafloor Sediment', *Frontiers in Microbiology*, 10(956). doi:10.3389/fmicb.2019.00956.

Kolmogorov, M. *et al.* (2020) 'metaFlye: scalable long-read metagenome assembly using repeat graphs', *Nature Methods*, 17(11), pp. 1103–1110. doi:10.1038/s41592-020-00971-x.

Könneke, M. *et al.* (2014) 'Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation', *Proceedings of the National Academy of Sciences of the United States of America*, 111(22), pp. 8239–8244. doi:10.1073/pnas.1402028111.

Korem, T. *et al.* (2015) 'Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples', *Science (New York, N.Y.)*, 349(6252), pp. 1101–1106. doi:10.1126/science.aac4812.

Krauthausen, B., Deuster, J. and Lang, R. (2007) 'Die Flucht des Wassers aus der Tiefe. Der Geysir von Andernach am Rhein', in *Faszination Geologie. Die bedeutendsten Geotope Deutschlands*, pp. 110–111.

L

Laczny, C.C. *et al.* (2015) 'VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data', *Microbiome*, 3(1), p. 1. doi:10.1186/s40168-014-0066-1.

Lakens, D., Scheel, A.M. and Isager, P.M. (2018) 'Equivalence Testing for Psychological Research: A Tutorial':, *Advances in Methods and Practices in Psychological Science*, 1(2), pp. 259–269. doi:10.1177/2515245918770963.

Lane, N. (2015) 'The unseen world: reflections on Leeuwenhoek (1677) "Concerning little animals"', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1666), p. 20140344. doi:10.1098/rstb.2014.0344.

Langille, M.G.I., Hsiao, W.W.L. and Brinkman, F.S.L. (2010) 'Detecting genomic islands using bioinformatics approaches', *Nature Reviews Microbiology*, 8(5), pp. 373–382. doi:10.1038/nrmicro2350.

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. doi:10.1038/nmeth.1923.

Lartillot, N., Lepage, T. and Blanquart, S. (2009) 'PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating', *Bioinformatics*, 25(17), pp. 2286–2288. doi:10.1093/bioinformatics/btp368.

Lau, M.C.Y. *et al.* (2016) 'An oligotrophic deep-subsurface community dependent on syntrophy is dominated by sulfur-driven autotrophic denitrifiers', *Proceedings of the National Academy of Sciences*, 113(49), pp. E7927–E7936. doi:10.1073/pnas.1612244113.

Lee, H. *et al.* (2019) 'Mantle degassing along strike-slip faults in the Southeastern Korean Peninsula', *Scientific Reports*, 9(1), pp. 1–9. doi:10.1038/s41598-019-51719-3.

Letunic, I. and Bork, P. (2016) 'Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees', *Nucleic Acids Research*, 44(Web Server issue), pp. W242–W245. doi:10.1093/nar/gkw290.

Li, D. *et al.* (2015) 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics (Oxford, England)*, 31(10), pp. 1674–1676. doi:10.1093/bioinformatics/btv033.

Li, H. (2018) 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics*, 34(18), pp. 3094–3100. doi:10.1093/bioinformatics/bty191.

Li, Z. *et al.* (2012) 'Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph', *Briefings in Functional Genomics*, 11(1), pp. 25–37. doi:10.1093/bfgp/elr035.

Ligon, B.L. (2002) 'Biography: Louis Pasteur: A controversial figure in a debate on scientific ethics', *Seminars in Pediatric Infectious Diseases*, 13(2), pp. 134–141. doi:10.1053/spid.2002.125138.

Lillestøl, R. *et al.* (2006) 'A putative viral defence mechanism in archaeal cells', *Archaea*, 2(1), pp. 59–72. doi:10.1155/2006/542818.

Liu, L. *et al.* (2015) 'High correlation between genotypes and phenotypes of environmental bacteria Comamonas testosteroni strains', *BMC Genomics*, 16(1). doi:10.1186/s12864-015-1314-x.

Liu, L. *et al.* (2021) 'Convergent Evolution of a Promiscuous 3-Hydroxypropionyl-CoA Dehydratase/Crotonyl-CoA Hydratase in Crenarchaeota and Thaumarchaeota', *mSphere*, 6(1), pp. e01079-20. doi:10.1128/mSphere.01079-20.

Liu, S. *et al.* (2022) 'Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation', *Microbiome*, 10(1), p. 76. doi:10.1186/s40168-022-01272-5.

Lloyd, K.G. *et al.* (2020) 'Evidence for a Growth Zone for Deep-Subsurface Microbial Clades in Near-Surface Anoxic Sediments', *Applied and Environmental Microbiology*, 86(19). doi:10.1128/AEM.00877-20.

Lobry, J.R. (1996) 'Asymmetric substitution patterns in the two DNA strands of bacteria', *Molecular Biology and Evolution*, 13(5), pp. 660–665. doi:10.1093/oxfordjournals.molbev.a025626.

Locey, K.J. and Lennon, J.T. (2016) 'Scaling laws predict global microbial diversity', *Proceedings of the National Academy of Sciences*, 113(21), pp. 5970–5975. doi:10.1073/pnas.1521291113.

Loreto, M.F. *et al.* (2015) 'Mantle degassing on a near shore volcano, SE Tyrrhenian Sea', *Terra Nova*, 27(3), pp. 195–205. doi:10.1111/ter.12148.

Louis Pasteur (2017) *On the extension of the germ theory to the etiology of certain common diseases*. Available at: https://web.archive.org/web/20170908042333/https://ebooks.adelaide.edu.au/p/pasteur/louis/exgerm/complete.html (Accessed: 7 April 2022).

Lundgren, M. *et al.* (2004) 'Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination', *Proceedings of the National Academy of Sciences of the United States of America*, 101(18), pp. 7046–7051. doi:10.1073/pnas.0400656101.

Luo, H., Zhang, C.-T. and Gao, F. (2014) 'Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes', *Frontiers in Microbiology*, 5(482). Available at: https://www.frontiersin.org/article/10.3389/fmicb.2014.00482 (Accessed: 2 May 2022).

Lupo, V. *et al.* (2021) 'Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics', *Frontiers in Microbiology*, 12(755101). Available at: https://www.frontiersin.org/article/10.3389/fmicb.2021.755101 (Accessed: 12 April 2022).

M

Madsen, J.S. *et al.* (2012) 'The interconnection between biofilm formation and horizontal gene transfer', *FEMS immunology and medical microbiology*, 65(2), pp. 183–195. doi:10.1111/j.1574-695X.2012.00960.x.

Magnabosco, C. *et al.* (2018) 'The biomass and biodiversity of the continental subsurface', *Nature Geoscience*, 11(10), pp. 707–717. doi:10.1038/s41561-018-0221-6.

Makarova, K.S. *et al.* (2006) 'A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action', *Biology Direct*, 1, p. 7. doi:10.1186/1745-6150-1-7.

Mallawaarachchi, V., Wickramarachchi, A. and Lin, Y. (2020) 'GraphBin: refined binning of metagenomic contigs using assembly graphs', *Bioinformatics*, 36(11), pp. 3307–3313. doi:10.1093/bioinformatics/btaa180.

Martijn, J. *et al.* (2020) 'Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition', *Nature Communications*, 11(1), p. 5490. doi:10.1038/s41467-020-19200-2.

Maruyama, S. *et al.* (1997) 'Paleogeographic maps of the Japanese Islands: Plate tectonic synthesis from 750 Ma to the present', *Island Arc*, 6(1), pp. 121–142. doi:10.1111/j.1440-1738.1997.tb00043.x.

Mazzarello, P. (1999) 'A unifying concept: the history of cell theory', *Nature Cell Biology*, 1(1), pp. E13–E15. doi:10.1038/8964.

Meer, M.V.D. van der *et al.* (2000) 'Autotrophy of green non-sulphur bacteria in hot spring microbial mats: biological explanations for isotopically heavy organic carbon in the geological record.', *Environmental microbiology* [Preprint]. doi:10.1046/J.1462-2920.2000.00124.X.

Mehlhorn, J. *et al.* (2014) 'Carbon dioxide triggered metal(loid) mobilisation in a mofette', *Chemical Geology*, 382, pp. 54–66. doi:10.1016/j.chemgeo.2014.05.027.

Mehrshad, M. *et al.* (2021) 'Energy efficiency and biological interactions define the core microbiome of deep oligotrophic groundwater', *Nature Communications*, 12(1), p. 4253. doi:10.1038/s41467-021-24549-z.

Meier-Kolthoff, J.P. and Göker, M. (2017) 'VICTOR: genome-based phylogeny and classification of prokaryotic viruses', *Bioinformatics*, 33(21), pp. 3396–3404. doi:10.1093/bioinformatics/btx440.

Meyer, W. and Stets, J. (2000) 'Geologische Übersichtskarte und Profil des Mittelrheintales - 1:100000, mit Erläuterungen', in. (Geologisches Landesamt Rheinland-Pfalz) Mainz, p. 49.

Meyer, W. and Striem, H.L. (1983) 'Geological indications for young horizontal displacements in the Central Rhenish Massif', *Geological indications for young horizontal displacements in the Central Rhenish Massif*, (2), pp. 97–100.

Miller, I.J. *et al.* (2019) 'Autometa: automated extraction of microbial genomes from individual shotgun metagenomes', *Nucleic Acids Research*, 47(10), p. e57. doi:10.1093/nar/gkz148.

Minh, B.Q. *et al.* (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37(5), pp. 1530–1534. doi:10.1093/molbev/msaa015.

Moissl, C. *et al.* (2003) 'In situ growth of the novel SM1 euryarchaeon from a string-of-pearls-like microbial community in its cold biotope, its physical separation and insights into its structure and physiology', *Archives of Microbiology*, 180(3), pp. 211–217. doi:10.1007/s00203-003-0580-1.

Moissl, C. *et al.* (2005) 'The unique structure of archaeal "hami", highly complex cell appendages with nano-grappling hooks: Unique structure of archaeal "hami"', *Molecular Microbiology*, 56(2), pp. 361–370. doi:10.1111/j.1365-2958.2005.04294.x.

Moissl, C., Rudolph, C. and Huber, R. (2002) 'Natural Communities of Novel Archaea and Bacteria with a String-of-Pearls-Like Morphology: Molecular Analysis of the Bacterial Partners', *Applied and Environmental Microbiology*, 68(2), pp. 933–937. doi:10.1128/AEM.68.2.933-937.2002.

Moller, A.G. and Liang, C. (2017) 'MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes', *PeerJ*, 5, p. e3788. doi:10.7717/peerj.3788.

Momper, L. *et al.* (2017) 'Energy and carbon metabolisms in a deep terrestrial subsurface fluid microbial community', *The ISME Journal*, 11(10), pp. 2319–2333. doi:10.1038/ismej.2017.94.

Mori, K. *et al.* (2009) 'Caldisericum exile gen. nov., sp. nov., an anaerobic, thermophilic, filamentous bacterium of a novel bacterial phylum, Caldiserica phyl. nov., originally called the candidate phylum OP5, and description of Caldisericaceae fam. nov., Caldisericales ord. nov. and Caldisericia classis nov.', *International Journal of Systematic and Evolutionary Microbiology*, 59(11), pp. 2894–2898. doi:10.1099/ijs.0.010033-0.

Moss, E.L., Maghini, D.G. and Bhatt, A.S. (2020) 'Complete, closed bacterial genomes from microbiomes using nanopore sequencing', *Nature Biotechnology*, 38, pp. 1–7. doi:10.1038/s41587-020-0422-6.

Mukherjee, S. *et al.* (2015) 'Large-scale contamination of microbial isolate genomes by Illumina PhiX control', *Standards in Genomic Sciences*, 10(1), p. 18. doi:10.1186/1944-3277-10-18.

Murray, A.E. *et al.* (2020) 'Roadmap for naming uncultivated Archaea and Bacteria', *Nature Microbiology*, 5(8), pp. 987–994. doi:10.1038/s41564-020-0733-x.

Murray, R.G. and Schleifer, K.H. (1994) 'Taxonomic notes: a proposal for recording the properties of putative taxa of procaryotes', *International Journal of Systematic Bacteriology*, 44(1), pp. 174–176. doi:10.1099/00207713-44-1-174.

Murray, R.G. and Stackebrandt, E. (1995) 'Taxonomic note: implementation of the provisional status Candidatus for incompletely described procaryotes', *International Journal of Systematic Bacteriology*, 45(1), pp. 186–187. doi:10.1099/00207713-45-1-186.

N

Needham, D.M. and Fuhrman, J.A. (2016) 'Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom', *Nature Microbiology*, 1, p. 16005. doi:10.1038/nmicrobiol.2016.5.

Nelson-Sathi, S. *et al.* (2015) 'Origins of major archaeal clades correspond to gene acquisitions from bacteria', *Nature*, 517(7532), pp. 77–80. doi:10.1038/nature13805.

Nguyen, L.-T. *et al.* (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32(1), pp. 268–274. doi:10.1093/molbev/msu300.

Nidhi, S. *et al.* (2021) 'Novel CRISPR–Cas Systems: An Updated Review of the Current Achievements, Applications, and Future Research Perspectives', *International Journal of Molecular Sciences*, 22(7), p. 3327. doi:10.3390/ijms22073327.

Noakes, M.T. *et al.* (2019) 'Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage', *Nature Biotechnology*, 37(6), pp. 651–656. doi:10.1038/s41587-019-0096-0.

Nunoura, T. *et al.* (2018) 'A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile', *Science (New York, N.Y.)*, 359(6375), pp. 559–563. doi:10.1126/science.aao3407.

Nurk, S. *et al.* (2017) 'metaSPAdes: a new versatile metagenomic assembler', *Genome Research*, 27(5), pp. 824–834. doi:10.1101/gr.213959.116.

Nurk, S. *et al.* (2020) 'HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads', *Genome Research*, p. gr.263566.120. doi:10.1101/gr.263566.120.

Nyyssönen, M. *et al.* (2014) 'Taxonomically and functionally diverse microbial communities in deep crystalline rocks of the Fennoscandian shield', *The ISME journal*, 8(1), pp. 126–138. doi:10.1038/ismej.2013.125.

O

Olm, M.R. *et al.* (2017) 'dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication', *The ISME Journal*, 11(12), p. 2864. doi:10.1038/ismej.2017.126.

Olm, M.R. *et al.* (2021) 'inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains', *Nature Biotechnology*, pp. 1–10. doi:10.1038/s41587-020-00797-0.

Olsen, I. (2015) 'Biofilm-specific antibiotic tolerance and resistance', *European Journal of Clinical Microbiology & Infectious Diseases*, 34(5), pp. 877–886. doi:10.1007/s10096-015-2323-z.

Ondov, B.D. *et al.* (2016) 'Mash: fast genome and metagenome distance estimation using MinHash', *Genome Biology*, 17(1), p. 132. doi:10.1186/s13059-016-0997-x.

Orakov, A. *et al.* (2021) 'GUNC: detection of chimerism and contamination in prokaryotic genomes', *Genome Biology*, 22, p. 178. doi:10.1186/s13059-021-02393-0.

Orsi, W.D. *et al.* (2021) 'Genome Evolution in Bacteria Isolated from Million-Year-Old Subseafloor Sediment', *mBio*, 12(4), pp. e01150-21. doi:10.1128/mBio.01150-21.

P

Page, A.J. *et al.* (2016) 'Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data', *Microbial Genomics*, 2(8), p. e000083. doi:10.1099/mgen.0.000083.

Pan, D. *et al.* (2017) 'Abundance and Distribution of Microbial Cells and Viruses in an Alluvial Aquifer', *Frontiers in Microbiology*, 8. doi:10.3389/fmicb.2017.01199.

Papke, R.T. *et al.* (2003) 'Geographical isolation in hot spring cyanobacteria', *Environmental Microbiology*, 5(8), pp. 650–659. doi:10.1046/j.1462-2920.2003.00460.x.

Parks, D.H. *et al.* (2015) 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Research*, 25(7), pp. 1043–1055. doi:10.1101/gr.186072.114.

Parks, D.H. *et al.* (2017) 'Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life', *Nature Microbiology*, 2(11), pp. 1533–1542. doi:10.1038/s41564-017-0012-7.

Parks, D.H. *et al.* (2018) 'A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life', *Nature Biotechnology*, 36(10), pp. 996–1004. doi:10.1038/nbt.4229.

Parks, D.H. *et al.* (2020) 'A complete domain-to-species taxonomy for Bacteria and Archaea', *Nature Biotechnology*, 38(9), pp. 1079–1086. doi:10.1038/s41587-020-0501-8.

Pasolli, E. *et al.* (2019) 'Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle', *Cell*, 176(3), pp. 649-662.e20. doi:10.1016/j.cell.2019.01.001.

Pebesma, E. and Bivand, R. (2005) 'Classes and Methods for Spatial Data in R', *R News*, 5.

Pop, M. *et al.* (2004) 'Comparative genome assembly', *Briefings in Bioinformatics*, 5(3), pp. 237–248. doi:10.1093/bib/5.3.237.

Popp, P.F. *et al.* (2021) 'Phyletic distribution and diversification of the Phage Shock Protein stress response system in bacteria and archaea'. bioRxiv, p. 2021.02.15.431232. doi:10.1101/2021.02.15.431232.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2009) 'FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix', *Molecular Biology and Evolution*, 26(7), pp. 1641–1650. doi:10.1093/molbev/msp077.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) 'FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments', *PLOS ONE*, 5(3), p. e9490. doi:10.1371/journal.pone.0009490.

Probst, A.J. *et al.* (2013) 'Tackling the minority: sulfate-reducing bacteria in an archaea-dominated subsurface biofilm', *The ISME Journal*, 7(3), pp. 635–651. doi:10.1038/ismej.2012.133.

Probst, A.J. *et al.* (2014) 'Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface', *Nature Communications*, 5, p. 5497. doi:10.1038/ncomms6497.

Probst, A.J. *et al.* (2017) 'Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations', *Environmental Microbiology*, 19(2), pp. 459–474. doi:10.1111/1462-2920.13362.

Probst, A.J. *et al.* (2018) 'Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface', *Nature Microbiology*, 3(3), pp. 328–336. doi:10.1038/s41564-017-0098-y.

Pruitt, K. *et al.* (2011) 'NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy', *Nucleic acids research*, 40, pp. D130-5. doi:10.1093/nar/gkr1079.

Q

Quast, C. *et al.* (2013) 'The SILVA ribosomal RNA gene database project: improved data processing and web-based tools', *Nucleic Acids Research*, 41(D1), pp. D590–D596. doi:10.1093/nar/gks1219.

Quince, C. *et al.* (2021) 'STRONG: metagenomics strain resolution on assembly graphs', *Genome Biology*, 22(1), p. 214. doi:10.1186/s13059-021-02419-7.

R

R Core Team (2008) ‘R: A Language and Environment for Statistical Computing’, R Foundation for Statistical Computing, Vienna, Austria.

Ragsdale, S.W. and Pierce, E. (2008) ‘Acetogenesis and the Wood-Ljungdahl Pathway of CO2 Fixation’, *Biochimica et biophysica acta*, 1784(12), pp. 1873–1898. doi:10.1016/j.bbapap.2008.08.012.

Rahlff, J. *et al.* (2021) ‘Lytic archaeal viruses infect abundant primary producers in Earth’s crust’, *Nature Communications*, 12(1), p. 4642. doi:10.1038/s41467-021-24803-4.

Ramos-Vera, W.H., Berg, I.A. and Fuchs, G. (2009) ‘Autotrophic carbon dioxide assimilation in Thermoproteales revisited’, *Journal of Bacteriology*, 191(13), pp. 4286–4297. doi:10.1128/JB.00145-09.

Rather, P. (2013) ‘Role of rhomboid proteases in bacteria’, *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1828(12), pp. 2849–2854. doi:10.1016/j.bbamem.2013.03.012.

Raymann, K., Brochier-Armanet, C. and Gribaldo, S. (2015) ‘The two-domain tree of life is linked to a new root for the Archaea’, *Proceedings of the National Academy of Sciences*, 112(21), pp. 6670–6675. doi:10.1073/pnas.1420858112.

Ren, J. *et al.* (2017) ‘VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data’, *Microbiome*, 5(1), p. 69. doi:10.1186/s40168-017-0283-5.

Ren, Y. *et al.* (2018) ‘Emergent heterogeneous microenvironments in biofilms: substratum surface heterogeneity and bacterial adhesion force-sensing’, *FEMS Microbiology Reviews*, 42(3), pp. 259–272. doi:10.1093/femsre/fuy001.

Rinke, C. *et al.* (2013) ‘Insights into the phylogeny and coding potential of microbial dark matter’, *Nature*, 499(7459), pp. 431–437. doi:10.1038/nature12352.

Ritter, J.R.R. (2007) *The Seismic Signature of the Eifel Plume*, *springerprofessional.de*. Available at: https://www.springerprofessional.de/the-seismic-signature-of-the-eifel-plume/2715234 (Accessed: 22 January 2020).

Rodriguez-R, L.M. *et al.* (2018) ‘Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity’, *mSystems*, 3(3). doi:10.1128/mSystems.00039-18.

Rodriguez-R, L.M. and Konstantinidis, K.T. (2016) *The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes*. e1900v1. PeerJ Inc. doi:10.7287/peerj.preprints.1900v1.

Ronaghi, M. *et al.* (1996) ‘Real-time DNA sequencing using detection of pyrophosphate release’, *Analytical Biochemistry*, 242(1), pp. 84–89. doi:10.1006/abio.1996.0432.

Rosselló-Móra, R. and Whitman, W.B. (2019) ‘Dialogue on the nomenclature and classification of prokaryotes’, *Systematic and Applied Microbiology*, 42(1), pp. 5–14. doi:10.1016/j.syapm.2018.07.002.

Røy, H. *et al.* (2012) 'Aerobic Microbial Respiration in 86-Million-Year-Old Deep-Sea Red Clay', *Science*, 336(6083), pp. 922–925. doi:10.1126/science.1219424.

Rudolph, C., Wanner, G. and Huber, R. (2001) 'Natural communities of novel archaea and bacteria growing in cold sulfurous springs with a string-of-pearls-like morphology', *Applied and Environmental Microbiology*, 67(5), pp. 2336–2344. doi:10.1128/AEM.67.5.2336-2344.2001.

Ruiz-Perez, C.A., Conrad, R.E. and Konstantinidis, K.T. (2021) 'MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes', *BMC Bioinformatics*, 22(1), p. 11. doi:10.1186/s12859-020-03940-5.

S

Sallet, E., Gouzy, J. and Schiex, T. (2019) 'EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes', *Methods in Molecular Biology (Clifton, N.J.)*, 1962, pp. 97–120. doi:10.1007/978-1-4939-9173-0_6.

Samson, R.Y. *et al.* (2013) 'Specificity and Function of Archaeal DNA Replication Initiator Proteins', *Cell Reports*, 3(2), pp. 485–496. doi:10.1016/j.celrep.2013.01.002.

Sandberg, R. *et al.* (2003) 'Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content', *Gene*, 311, pp. 35–42. doi:10.1016/s0378-1119(03)00581-x.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463–5467.

Schaack, S., Gilbert, C. and Feschotte, C. (2010) 'Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution', *Trends in Ecology & Evolution*, 25(9), pp. 537–546. doi:10.1016/j.tree.2010.06.001.

Schreiber, U. and Rotsch, S. (1998) 'Cenozoic block rotation according to a conjugate shear system in central Europe — indications from palaeomagnetic measurements', *Tectonophysics*, 299(1), pp. 111–142. doi:10.1016/S0040-1951(98)00201-7.

Schrenk, M.O., Holden, J.F. and Baross, J.A. (2008) 'Magma-to-microbe networks in the context of sulfide hosted microbial ecosystems', *Washington DC American Geophysical Union Geophysical Monograph Series*, 178, pp. 233–258. doi:10.1029/178GM12.

Schunk, R. (2012) 'Der Ausbruch – ein faszinierendes Naturschauspiel', in *Naturschauspiel Geysir Andernach*, pp. 20–36.

Seah, B.K.B. and Gruber-Vodicka, H.R. (2015) 'gbtools: Interactive Visualization of Metagenome Bins in R', *Frontiers in Microbiology*, 6(1451). doi:10.3389/fmicb.2015.01451.

Seemann, T. (2014) 'Prokka: rapid prokaryotic genome annotation', *Bioinformatics (Oxford, England)*, 30(14), pp. 2068–2069. doi:10.1093/bioinformatics/btu153.

Segata, N. *et al.* (2011) 'Metagenomic biomarker discovery and explanation', *Genome Biology*, 12(6), p. R60. doi:10.1186/gb-2011-12-6-r60.

Shaffer, M. *et al.* (2020) 'DRAM for distilling microbial metabolism to automate the curation of microbiome function', *Nucleic Acids Research*, 48(16), pp. 8883–8900. doi:10.1093/nar/gkaa621.

Shaiber, A. and Eren, A.M. (2019) 'Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories', *mBio*, 10(3). doi:10.1128/mBio.00725-19.

Shan, S., Schmid, S.L. and Zhang, X. (2009) 'Signal recognition particle (SRP) and SRP receptor: A new paradigm for multi-state regulatory GTPases', *Biochemistry*, 48(29), pp. 6696–6704. doi:10.1021/bi9006989.

Shannon, C.E. (1948) 'A mathematical theory of communication', *The Bell System Technical Journal*, 27(3), pp. 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

Shannon, P. *et al.* (2003) 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Research*, 13(11), pp. 2498–2504. doi:10.1101/gr.1239303.

Sharon, I. *et al.* (2013) 'Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization', *Genome Research*, 23(1), pp. 111–120. doi:10.1101/gr.142315.112.

Sheik, C.S. *et al.* (2018) 'Identification and Removal of Contaminant Sequences From Ribosomal Gene Databases: Lessons From the Census of Deep Life', *Frontiers in Microbiology*, 9. doi:10.3389/fmicb.2018.00840.

Shima, S. *et al.* (2001) 'Characterization of a heme-dependent catalase from Methanobrevibacter arboriphilus', *Applied and Environmental Microbiology*, 67(7), pp. 3041–3045. doi:10.1128/AEM.67.7.3041-3045.2001.

Shmakov, S.A., Utkina, I., *et al.* (2020) 'CRISPR Arrays Away from cas Genes', *The CRISPR Journal*, 3(6), pp. 535–549. doi:10.1089/crispr.2020.0062.

Shmakov, S.A., Wolf, Y.I., *et al.* (2020) 'Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes', *Communications Biology*, 3(1), pp. 1–9. doi:10.1038/s42003-020-1014-1.

Sieber, C.M.K. *et al.* (2018) 'Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy', *Nature Microbiology*, 3(7), pp. 836–843. doi:10.1038/s41564-018-0171-1.

Simonsohn, U. (2015) 'Small Telescopes: Detectability and the Evaluation of Replication Results', *Psychological Science*, 26(5), pp. 559–569. doi:10.1177/0956797614567341.

Singleton, C.M. *et al.* (2020) 'Connecting structure to function with the recovery of over 1000 high-quality activated sludge metagenome-assembled genomes encoding full-length rRNA

genes using long-read sequencing', *bioRxiv*, p. 2020.05.12.088096. doi:10.1101/2020.05.12.088096.

Skennerton, C.T., Imelfort, M. and Tyson, G.W. (2013) 'Crass: identification and reconstruction of CRISPR from unassembled metagenomic data', *Nucleic Acids Research*, 41(10), p. e105. doi:10.1093/nar/gkt183.

Smith, A.R. *et al.* (2019) 'Carbon fixation and energy metabolisms of a subseafloor olivine biofilm', *The ISME Journal*, 13(7), pp. 1737–1749. doi:10.1038/s41396-019-0385-0.

Sommerlund, J. (2006) 'Classifying Microorganisms: The Multiplicity of Classifications and Research Practices in Molecular Microbial Ecology', *Social Studies of Science*, 36(6), pp. 909–928.

Soucy, S.M., Huang, J. and Gogarten, J.P. (2015) 'Horizontal gene transfer: building the web of life', *Nature Reviews Genetics*, 16(8), pp. 472–482. doi:10.1038/nrg3962.

Sousa, F.L. and Martin, W.F. (2014) 'Biochemical fossils of the ancient transition from geoenergetics to bioenergetics in prokaryotic one carbon compound metabolism', *Biochimica Et Biophysica Acta*, 1837(7), pp. 964–981. doi:10.1016/j.bbabio.2014.02.001.

Spang, A. *et al.* (2015) 'Complex archaea that bridge the gap between prokaryotes and eukaryotes', *Nature*, 521(7551), pp. 173–179. doi:10.1038/nature14447.

Speth, D.R. and Orphan, V.J. (2018) 'Metabolic marker gene mining provides insight in global mcrA diversity and, coupled with targeted genome reconstruction, sheds further light on metabolic potential of the Methanomassiliicoccales', *PeerJ*, 6, p. e5614. doi:10.7717/peerj.5614.

Stackebrandt, E. *et al.* (2002) 'Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology.', *International Journal of Systematic and Evolutionary Microbiology*, 52(3), pp. 1043–1047. doi:10.1099/00207713-52-3-1043.

Stamatakis, A. (2014) 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30(9), pp. 1312–1313. doi:10.1093/bioinformatics/btu033.

Starnawski, P. *et al.* (2017) 'Microbial community assembly and evolution in subseafloor sediment', *Proceedings of the National Academy of Sciences*, 114(11), pp. 2940–2945. doi:10.1073/pnas.1614190114.

Steffens, L. *et al.* (2021) 'High CO_2 levels drive the TCA cycle backwards towards autotrophy', *Nature*, 592(7856), pp. 784–788. doi:10.1038/s41586-021-03456-9.

Stein, J.L. *et al.* (1996) 'Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon', *Journal of Bacteriology*, 178(3), pp. 591–599. doi:10.1128/jb.178.3.591-599.1996.

Stevens, T. (1997) 'Lithoautotrophy in the subsurface', *FEMS Microbiology Reviews*, 20(3–4), pp. 327–337. doi:10.1111/j.1574-6976.1997.tb00318.x.

Stevens, T.O. and McKinley, J.P. (2000) 'Abiotic Controls on H_2 Production from Basalt−Water Reactions and Implications for Aquifer Biogeochemistry', *Environmental Science & Technology*, 34(5), pp. 826–831. doi:10.1021/es990583g.

Stupperich, E. and Krautler, B. (1988) 'Pseudo vitamin B_{12} or 5-hydroxybenzimidazolyl-cobamide are the corrinoids found in methanogenic bacteria', *Archives of microbiology* [Preprint]. Available at: https://scholar.google.com/scholar_lookup?title=Pseudo+vitamin+B12+or+5-hydroxybenzimidazolyl-cobamide+are+the+corrinoids+found+in+methanogenic+bacteria&author=Stupperich%2C+E.&publication_year=1988 (Accessed: 22 May 2022).

Sutcliffe, I.C. *et al.* (2020) 'Minutes of the International Committee on Systematics of Prokaryotes online discussion on the proposed use of gene sequences as type for naming of prokaryotes, and outcome of vote', *International Journal of Systematic and Evolutionary Microbiology*, 70(7), pp. 4416–4417. doi:10.1099/ijsem.0.004303.

Suzek, B.E. *et al.* (2007) 'UniRef: comprehensive and non-redundant UniProt reference clusters', *Bioinformatics*, 23(10), pp. 1282–1288. doi:10.1093/bioinformatics/btm098.

Suzek, B.E. *et al.* (2015) 'UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches', *Bioinformatics*, 31(6), pp. 926–932. doi:10.1093/bioinformatics/btu739.

T

Tatusova, T. *et al.* (2016) 'NCBI prokaryotic genome annotation pipeline', *Nucleic Acids Research*, 44(14), pp. 6614–6624. doi:10.1093/nar/gkw569.

Teeling, H. *et al.* (2004) 'Application of tetranucleotide frequencies for the assignment of genomic fragments', *Environmental Microbiology*, 6(9), pp. 938–947. doi:10.1111/j.1462-2920.2004.00624.x.

Teeling, H. and Glöckner, F.O. (2012) 'Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective', *Briefings in Bioinformatics*, 13(6), pp. 728–742. doi:10.1093/bib/bbs039.

Telling, J. *et al.* (2015) 'Rock comminution as a source of hydrogen for subglacial ecosystems', *Nature Geoscience*, 8(11), pp. 851–855. doi:10.1038/ngeo2533.

Teske, A. *et al.* (2015) *Deep Subsurface Microbiology*. Frontiers Media SA.

Teufel, R. *et al.* (2009) '3-hydroxypropionyl-coenzyme A dehydratase and acryloyl-coenzyme A reductase, enzymes of the autotrophic 3-hydroxypropionate/4-hydroxybutyrate cycle in the Sulfolobales', *Journal of Bacteriology*, 191(14), pp. 4572–4581. doi:10.1128/JB.00068-09.

Thomas, N.K. *et al.* (2021) 'Direct Nanopore Sequencing of Individual Full Length tRNA Strands', *ACS Nano*, 15(10), pp. 16642–16653. doi:10.1021/acsnano.1c06488.

Torsvik, T.H. *et al.* (2012) 'Phanerozoic polar wander, palaeogeography and dynamics', *Earth-Science Reviews*, 114(3), pp. 325–368. doi:10.1016/j.earscirev.2012.06.007.

Tu, T.-H. *et al.* (2017) 'Microbial Community Composition and Functional Capacity in a Terrestrial Ferruginous, Sulfate-Depleted Mud Volcano', *Frontiers in Microbiology*, 8. doi:10.3389/fmicb.2017.02137.

Tully, B.J. *et al.* (2017) 'A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer', *The ISME Journal*, 12(1), p. 1. doi:10.1038/ismej.2017.187.

Tyson, G.W. *et al.* (2004) 'Community structure and metabolism through reconstruction of microbial genomes from the environment', *Nature*, 428(6978), pp. 37–43. doi:10.1038/nature02340.

U

Ultsch, A. (2005) 'ESOM-maps: tools for clustering viszalization and classification with Emergent SOM', *Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany*, 46.

Uritskiy, G.V., DiRuggiero, J. and Taylor, J. (2018) 'MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis', *Microbiome*, 6(1), p. 158. doi:10.1186/s40168-018-0541-1.

V

Vallenet, D. *et al.* (2006) 'MaGe: a microbial genome annotation system supported by synteny results', *Nucleic Acids Research*, 34(1), pp. 53–65. doi:10.1093/nar/gkj406.

Vallenet, D. *et al.* (2009) 'MicroScope: a platform for microbial genome annotation and comparative genomics', *Database: The Journal of Biological Databases and Curation*, 2009(bap021). doi:10.1093/database/bap021.

Venter, J.C. *et al.* (2004) 'Environmental genome shotgun sequencing of the Sargasso Sea', *Science (New York, N.Y.)*, 304(5667), pp. 66–74. doi:10.1126/science.1093857.

Vetriani, C., Reysenbach, A.-L. and Doré, J. (1998) 'Recovery and phylogenetic analysis of archaeal rRNA sequences from continental shelf sediments', *FEMS Microbiology Letters*, 161(1), pp. 83–88. doi:10.1111/j.1574-6968.1998.tb12932.x.

Větrovský, T. and Baldrian, P. (2013) 'The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses', *PLoS ONE*, 8(2), p. e57923. doi:10.1371/journal.pone.0057923.

Vieira-Silva, S. and Rocha, E.P.C. (2010) 'The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics', *PLOS Genetics*, 6(1), p. e1000808. doi:10.1371/journal.pgen.1000808.

Vollmers, J. *et al.* (2022) 'How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner', *Nucleic Acids Research*, p. gkac294. doi:10.1093/nar/gkac294.

van der Walt, A.J. *et al.* (2017) 'Assembling metagenomes, one community at a time', *BMC Genomics*, 18(1), p. 521. doi:10.1186/s12864-017-3918-9.

W

Wang, H. *et al.* (2015) 'Archaeal extrachromosomal genetic elements', *Microbiology and molecular biology reviews: MMBR*, 79(1), pp. 117–152. doi:10.1128/MMBR.00042-14.

Weiss, M.C. *et al.* (2016) 'The physiology and habitat of the last universal common ancestor', *Nature Microbiology*, 1(9), pp. 1–8. doi:10.1038/nmicrobiol.2016.116.

Weissman, J.L., Hou, S. and Fuhrman, J.A. (2020) 'Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns', *bioRxiv*, p. 2020.07.25.221176. doi:10.1101/2020.07.25.221176.

Werner, C. *et al.* (2019) 'Carbon Dioxide Emissions from Subaerial Volcanic Regions: Two Decades in Review', in Orcutt, B.N., Daniel, I., and Dasgupta, R. (eds) *Deep Carbon*. 1st edn. Cambridge University Press, pp. 188–236. doi:10.1017/9781108677950.008.

West, P.T. *et al.* (2018) 'Genome-reconstruction for eukaryotes from complex natural microbial communities', *Genome Research*, 28(4), pp. 569–580. doi:10.1101/gr.228429.117.

Whitaker, R.J., Grogan, D.W. and Taylor, J.W. (2003) 'Geographic barriers isolate endemic populations of hyperthermophilic archaea', *Science (New York, N.Y.)*, 301(5635), pp. 976–978. doi:10.1126/science.1086909.

Whitman, W.B. (2015) 'Genome sequences as the type material for taxonomic descriptions of prokaryotes', *Systematic and Applied Microbiology*, 38(4), pp. 217–222. doi:10.1016/j.syapm.2015.02.003.

Whitman, W.B., Sutcliffe, I.C. and Rossello-Mora, R. (2019) 'Proposal for changes in the International Code of Nomenclature of Prokaryotes: granting priority to Candidatus names', *International Journal of Systematic and Evolutionary Microbiology*, 69(7), pp. 2174–2175. doi:10.1099/ijsem.0.003419.

Wick, R.R. *et al.* (2017) 'Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads', *PLOS Computational Biology*, 13(6), p. e1005595. doi:10.1371/journal.pcbi.1005595.

Wick, R.R. *et al.* (2021) 'Trycycler: consensus long-read assemblies for bacterial genomes', *Genome Biology*, 22(1), p. 266. doi:10.1186/s13059-021-02483-z.

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag (Use R!).

Willbanks, A. *et al.* (2016) 'The Evolution of Epigenetics: From Prokaryotes to Humans and Its Biological Consequences', *Genetics & Epigenetics*, 8, pp. 25–36. doi:10.4137/GEG.S31863.

Williams, D., Gogarten, J.P. and Papke, R.T. (2012) 'Quantifying homologous replacement of loci between haloarchaeal species', *Genome Biology and Evolution*, 4(12), pp. 1223–1244. doi:10.1093/gbe/evs098.

Williams, T.A. *et al.* (2020) 'Phylogenomics provides robust support for a two-domains tree of life', *Nature ecology & evolution*, 4(1), pp. 138–147. doi:10.1038/s41559-019-1040-x.

Wimmer, F. and Beisel, C.L. (2020) 'CRISPR-Cas Systems and the Paradox of Self-Targeting Spacers', *Frontiers in Microbiology*, 10(3078). Available at: https://www.frontiersin.org/article/10.3389/fmicb.2019.03078 (Accessed: 5 May 2022).

Woese, C.R. (1987) 'Bacterial evolution.', *Microbiological Reviews*, 51(2), pp. 221–271.

Woese, C.R. and Fox, G.E. (1977) 'Phylogenetic structure of the prokaryotic domain: The primary kingdoms', *Proceedings of the National Academy of Sciences*, 74(11), pp. 5088–5090. doi:10.1073/pnas.74.11.5088.

Wood, H.G. (1991) 'Life with CO or CO2 and H2 as a source of carbon and energy.', *The FASEB Journal*, 5(2), pp. 156–163. doi:10.1096/fasebj.5.2.1900793.

Woyke, T. *et al.* (2010) 'One Bacterial Cell, One Complete Genome', *PLoS ONE*, 5(4). doi:10.1371/journal.pone.0010314.

Wright, S. (1943) 'Isolation by Distance', *Genetics*, 28(2), pp. 114–138.

Wrighton, K.C. *et al.* (2012) 'Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla', *Science (New York, N.Y.)*, 337(6102), pp. 1661–1665. doi:10.1126/science.1224041.

Wu, Y.-W. *et al.* (2014) 'MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm', *Microbiome*, 2(1), p. 26. doi:10.1186/2049-2618-2-26.

Wu, Y.-W., Simmons, B.A. and Singer, S.W. (2016) 'MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets', *Bioinformatics*, 32(4), pp. 605–607. doi:10.1093/bioinformatics/btv638.

Wu, Z. *et al.* (2014) 'DNA replication origins in archaea', *Frontiers in Microbiology*, 5(179). doi:10.3389/fmicb.2014.00179.

Y

Yamamoto, M. *et al.* (2010) 'Carboxylation reaction catalyzed by 2-oxoglutarate:ferredoxin oxidoreductases from Hydrogenobacter thermophilus', *Extremophiles: Life Under Extreme Conditions*, 14(1), pp. 79–85. doi:10.1007/s00792-009-0289-4.

Yu, G. *et al.* (2018) 'BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage', *Bioinformatics (Oxford, England)*, 34(24), pp. 4172–4179. doi:10.1093/bioinformatics/bty519.

Z

Zatopek, K.M., Gardner, A.F. and Kelman, Z. (2018) 'Archaeal DNA replication and repair: new genetic, biophysical and molecular tools for discovering and characterizing enzymes, pathways and mechanisms', *FEMS Microbiology Reviews*, 42(4), pp. 477–488. doi:10.1093/femsre/fuy017.

Zhang, R. and Zhang, C.-T. (2005) 'Identification of replication origins in archaeal genomes based on the Z-curve method', *Archaea*, 1(5), pp. 335–346.

Zhang, Y. (2014) 'Degassing History of Earth', in *Treatise on Geochemistry*. Elsevier, pp. 37–69. doi:10.1016/B978-0-08-095975-7.01302-4.

Zhou, Z. *et al.* (2019) 'METABOLIC: A scalable high-throughput metabolic and biogeochemical functional trait profiler based on microbial genomes', *bioRxiv*, p. 761643. doi:10.1101/761643.

VII. Content of supporting CD

The accompanying CD contains the supplemental data for the publications presented in sections III and IV. The three folders with the respective supplemental Material on the CD are:

1. sectionIII.2_uBin
2. sectionIII.3_genetic_diversity
3. sectionIV.2_archaeal_diversity

Supplemental files are supplied in a separate subfolder while the main Supplementary Information document is directly supplied in PDF format in the respective section folder. Supplemental files were named to match the name listed in the main and supplemental manuscript if the journal had given them an alphanumeric identifier. Section III.1 does not contain any supplemental data and is thus not listed on the CD. For each folder, the supplemental files are additionally explained in a File descriptions document in the main subfolder. Due to their size, the not yet publicly available metagenomes used in section IV are not supplied on the CD. They will be published soon but in the interim will be supplied upon

request. For archiving purposes, these folders are all archived for posterity in our in-house servers in /ICEAGE/Archive/Theses/TLVB_PhDthesis. Requests for the supplemental files, not yet public metagenomes and in-house scripts used in section IV and can be send to till.bornemann@uni-due.de.

VIII. Eidesstattliche Erklärung

Ich bestätige hiermit an Eides statt, dass die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe des Literaturzitats gekennzeichnet.

Essen, 31.05.2022

Till L.V. Bornemann