frontiers in Microbiology

MINI REVIEW
published: 05 January 2022
doi: 10.3389/fmicb.2021.761869

A Genomic Perspective Across Earth's Microbiomes Reveals That Genome Size in Archaea and Bacteria Is Linked to Ecosystem Type and Trophic Strategy

Alejandro Rodríguez-Gijón[1†], Julia K. Nuy[1†], Maliheh Mehrshad[2], Moritz Buck[2], Frederik Schulz[3], Tanja Woyke[3] and Sarahi L. Garcia[1]*

[1] Department of Ecology, Environment, and Plant Sciences, Science for Life Laboratory, Stockholm University, Stockholm, Sweden, [2] Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden, [3] DOE Joint Genome Institute, Berkeley, CA, United States

OPEN ACCESS

Edited by:
M. Pilar Francino, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO), Spain

Reviewed by:
Jennifer F. Biddle, University of Delaware, United States
Georg H. Reischer, Vienna University of Technology, Austria

***Correspondence:**
Sarahi L. Garcia
sarahi.garcia@su.se

[†]These authors share first authorship

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *20 August 2021*
Accepted: *15 December 2021*
Published: *05 January 2022*

Citation:
Rodríguez-Gijón A, Nuy JK, Mehrshad M, Buck M, Schulz F, Woyke T and Garcia SL (2022) A Genomic Perspective Across Earth's Microbiomes Reveals That Genome Size in Archaea and Bacteria Is Linked to Ecosystem Type and Trophic Strategy. Front. Microbiol. 12:761869. doi: 10.3389/fmicb.2021.761869

Our view of genome size in Archaea and Bacteria has remained skewed as the data has been dominated by genomes of microorganisms that have been cultivated under laboratory settings. However, the continuous effort to catalog Earth's microbiomes, specifically propelled by recent extensive work on uncultivated microorganisms, provides an opportunity to revise our perspective on genome size distribution. We present a meta-analysis that includes 26,101 representative genomes from 3 published genomic databases; metagenomic assembled genomes (MAGs) from GEMs and stratfreshDB, and isolates from GTDB. Aquatic and host-associated microbial genomes present on average the smallest estimated genome sizes (3.1 and 3.0 Mbp, respectively). These are followed by terrestrial microbial genomes (average 3.7 Mbp), and genomes from isolated microorganisms (average 4.3 Mbp). On the one hand, aquatic and host-associated ecosystems present smaller genomes sizes in genera of phyla with genome sizes above 3 Mbp. On the other hand, estimated genome size in phyla with genomes under 3 Mbp showed no difference between ecosystems. Moreover, we observed that when using 95% average nucleotide identity (ANI) as an estimator for genetic units, only 3% of MAGs cluster together with genomes from isolated microorganisms. Although there are potential methodological limitations when assembling and binning MAGs, we found that in genome clusters containing both environmental MAGs and isolate genomes, MAGs were estimated only an average 3.7% smaller than isolate genomes. Even when assembly and binning methods introduce biases, estimated genome size of MAGs and isolates are very similar. Finally, to better understand the ecological drivers of genome size, we discuss on the known and the overlooked factors that influence genome size in different ecosystems, phylogenetic groups, and trophic strategies.

Keywords: microbial ecology, genome size, bacteria, archaea, genomics

INTRODUCTION

As microbiologists, how do we define what is a small or a big genome? Perhaps, researchers working on model organisms such as *Escherichia coli* with a genome size of ∼5 Mbp (Abram et al., 2021) would define "big" or "small" differently to researchers working on soil-dwelling bacteria with a genome size of 16 Mbp (Garcia et al., 2014). On the lower genome size scale, whereas genome sizes of bacterial endosymbionts of insects may have genomes merely larger than 100 kbp (Moran and Bennett, 2014), the abundant *Prochlorococcus* range between 1.6 and 1.9 Mbp for high-light and low-light ecotypes (Berube et al., 2018). In summary, it is known that genome sizes of Archaea and Bacteria range between 100 kbp and 16 Mbp, but the genome size distribution in nature is still undefined. Therefore, the aim of this review is to provide an overview of the distribution of genome sizes in different ecosystems.

We leveraged recently published databases of archaeal and bacterial metagenome assembled genomes (MAGs) (Nayfach et al., 2020; Buck et al., 2021a) together with isolate genomes to revisit and acquire an updated understanding of the estimated genome size distribution across different ecosystems. In this review, we also discuss the ecological drivers that potentially influence genome sizes. In summary, we found that 76.3% of representative archaeal and bacterial genomes recovered through genome-resolved metagenomics present estimated genome sizes below 4 Mbp. Furthermore, all MAGs from five archaeal phyla (Micrarcheota, Ianarchaeota, Undinarchaeota, Nanohaloarchaeota, and Hadarchaeota) and two bacterial phyla (Coprothermobacterota and Dictyoglomota) were recovered exclusively from aquatic ecosystems and have genome sizes below 2 Mbp (**Figures 1A,B**).

APPROXIMATION OF GENETIC UNITS USING 95% AVERAGE NUCLEOTIDE IDENTITY AND ITS CAVEATS

Species are widely considered congruent genetic and ecological units for sexual eukaryotes (Mallet, 2008; Shapiro and Polz, 2015). However, there is no consensus regarding the concept of species for Archaea and Bacteria. Instead, 95% average nucleotide identity (ANI) has been a widely recognized as a genetic boundary to operationally estimate genetic units or "microbial species" (Konstantinidis and Tiedje, 2005; Varghese et al., 2015; Garcia et al., 2018; Jain et al., 2018). Several genomic and metagenomic studies have verified the existence of sequence discrete genetic units with 95% ANI as boundary (Olm et al., 2020; Rodriguez et al., 2021).

We used the 95% ANI boundary in published datasets (Nayfach et al., 2020; Parks et al., 2020; Buck et al., 2021a) to review and renew our view of the distribution of archaeal and bacterial estimated genome size. To minimize representation biases (Gweon et al., 2017), we included one representative per 95% ANI cluster. The approximation of these clusters will be called mOTUs (metagenomic Operational Taxonomic Units) (Buck et al., 2021b; Garcia et al., 2021). We used only one MAG per mOTU, making the estimated genome size dependent on the MAG assembly and quality. It is known that MAG assembly and binning might discriminate against ribosomal RNAs, transfer RNAs, mobile element functions and genes of unknown function (Nelson et al., 2020; Meziti et al., 2021), and also that completeness could be underestimated for streamlined genomes (Garcia et al., 2015). Despite potential biases introduced by these methodological limitations, we still offer valuable insights on genome size distribution across different environments.

EXTANT GENOME SIZE DISTRIBUTION IN THE ENVIRONMENT

In this review, we have included ∼64,500 environmental MAGs available via two recently published datasets, stratfreshDB and GEMs. StratfreshDB offers ∼12,000 MAGs (>40% completeness) from 41 stratified lakes and ponds assembled with Megahit (v1.1.13) and binned with Metabat (v2.12.1) (Buck et al., 2021a). GEMs offers ∼52,000 MAGs (>50% completeness) from > 10,000 metagenomes collected from diverse habitats on Earth (Nayfach et al., 2020). GEMs dataset was assembled using metaSPAdes and binned with Metabat (v0.32.5). After dereplication using fastANI and mOTUlizer (Jain et al., 2018; Buck et al., 2021b), our meta-analysis includes 17 834 mOTUs, with one representative MAG each (completeness > 50% and contamination < 5%, assessed with CheckM v1.1.3) (Parks et al., 2015). We complemented these environmental MAGs by adding 8 267 representative genomes (>90% complete) of isolates from GTDB (Parks et al., 2020; **Figure 1A**). These genomes are marked in the GTDB databases (release 95) to originate from culture collections. After clustering at 95% ANI threshold, 540 mOTUs contained representatives from both environmental MAGs and isolate genomes (**Figure 1C**). Previous surveys based on 16S rRNA have found that the uncultured microbial fraction could constitute up to 81% of the total microbial diversity on Earth (Lloyd et al., 2018). However, it is known that 16S rRNA underestimate prokaryotic diversity (Rodriguez et al., 2018). Overall, our review shows that 3% of the reconstructed environmental mOTUs are represented among cultured microbes.

Furthermore, using completeness estimates from CheckM, we compared the estimated genome size distribution of all MAGs vs. genomes from isolates. The estimated genome size was calculated by dividing the MAG's assembly size by CheckM completeness (ranging from 0 to 1). Representative genomes from isolates have an average genome size of 4.3 Mbp which is significantly larger than that of MAGs (*t*-test $p < 0.0001$), both when comparing Archaea and Bacteria combined and separately. To compare estimated genome sizes between MAGs, ecosystem type was used according to the GEMs database. Although the ecosystem classification presented here is coarse and might contain countless niches, it still allowed us to see trends for genome sizes. Estimated genome sizes of aquatic MAGs have an average of 3.1 Mbp, host-associated MAGs average 3.0 Mbp, and terrestrial MAGs average 3.7 Mbp (**Figure 1A**). For the 540 mOTUs that contained both environmental MAGs and isolate

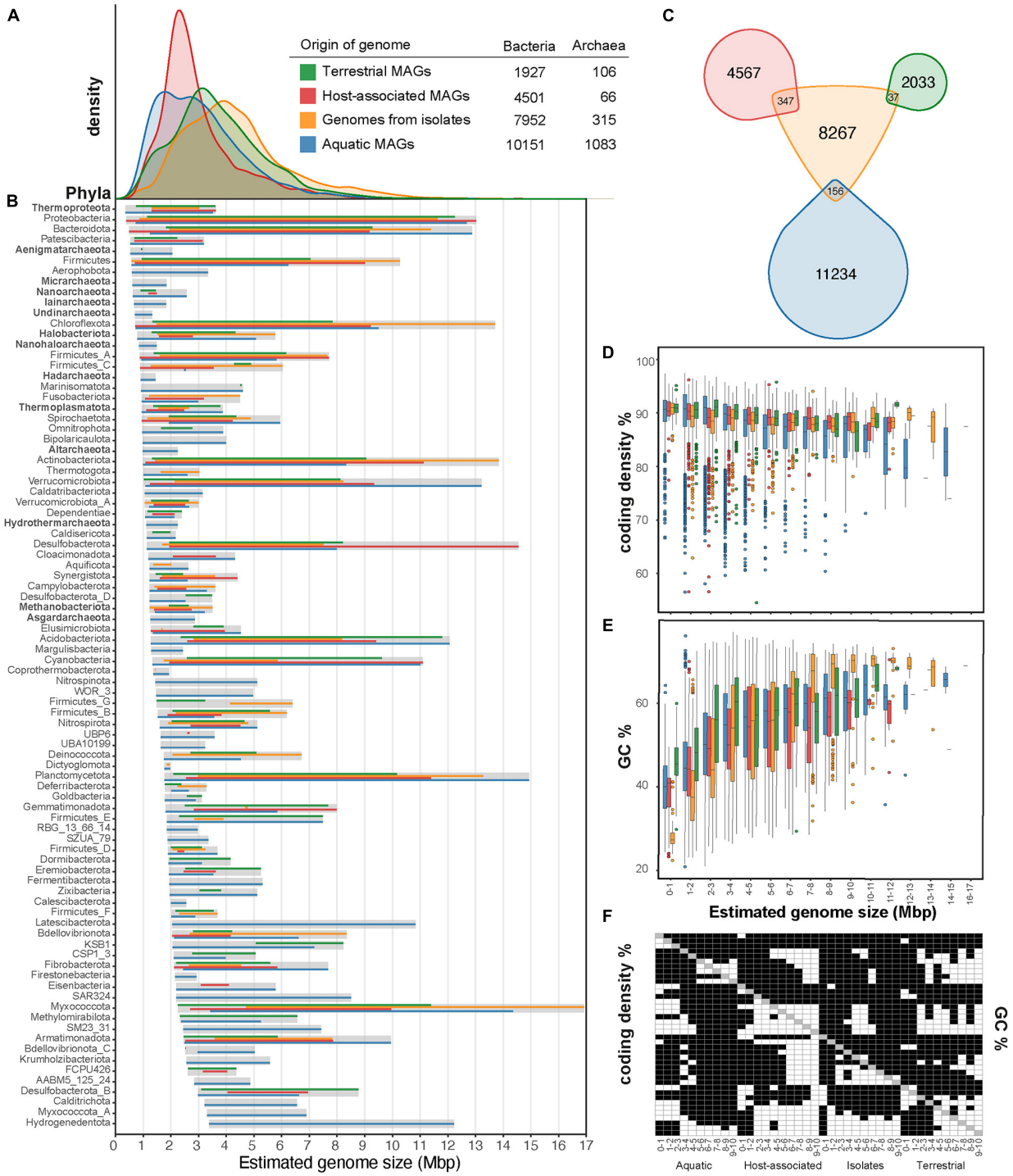

FIGURE 1 | Overview of the genome size distribution across Earth's microbiomes. Genome size distribution of Archaea and Bacteria **(A)** from different environmental sources and across different archaeal and bacterial phyla **(B)** are shown for a total of 26,101 representative genomes. Isolate genomes were gathered from GTDB (release95) and environmental MAGs were gathered from GEMs (Nayfach et al., 2020) and stratfreshDB (Buck et al., 2021a). We use one representative genome per mOTU (defined by 95% ANI) from the union of GEMs catalog and stratfreshDB in the plots. From the GTDB database, we selected one representative isolate genome per species cluster that was circumscribed based on the ANI (≥95%) and alignment fraction [(AF) > 65%] between genomes (Parks et al., 2020). To construct the figures, we plotted the min-max estimated genome sizes, which were calculated based on the genome assembly size and completeness estimation provided. Venn diagram of the intersection between the representative environmental MAGs and the representative isolate genomes **(C)**. The intersection was calculated using FastANI (Jain et al., 2018) and was determined with a threshold of 95%. The coding density **(D)** and GC content (%) **(E)** are shown for the archaeal and bacterial MAGs across different ecosystem categories and isolates. Pair-wise *t*-test was performed in all variables of **(D,E)** and shown in **(F)**, where white is significant ($p < 0.05$) and black is not significant ($p > 0.05$). In **(B)**, we only included phyla with more than five genomes.

genomes (**Figure 1C**), we found that MAGs were estimated on average 3.7% smaller than isolate genomes (**Supplementary Figure 1**). In other words, even when assembly and binning methods introduce biases, estimated genome size of MAGs and isolates are very similar. Overall, this suggests that the bias in metagenome assembly and binning would not account for the genome size difference observed between all isolate representatives and ecosystem MAGs, neither for the differences among ecosystem MAGs.

A reason for the difference in genome size between isolates and genomes reconstructed from metagenomes might be related to the fact that traditional isolation techniques select for rare microorganisms (Shade et al., 2012) and do not capture the entire ecosystem's diversity (**Figure 1C**). For example, it is known that classical cultivation techniques with rich media bias the cultivation toward copiotrophic and fast-growing microorganisms (Swan et al., 2013). Cultivation biases our view of nature because it selects against slow growing microorganisms (Imachi et al., 2020), host dependency (Cross et al., 2019), and dormancy (Hoehler and Jorgensen, 2013) among others. In nature there are many microorganisms with very limited metabolic capacity (Figueroa-Gonzalez et al., 2020) that is linked with dependencies and smaller genomes sizes (Morris et al., 2012). Microorganisms in nature have coevolved with other microorganisms and might have specific requirements that are difficult to mimic in batch-culture standard-media isolation techniques (Garcia, 2016). Although there have been many advances on cultivation techniques (Dedysh, 2011; Carini et al., 2013; Henson et al., 2016; Imachi et al., 2020), more innovations to culture the uncultivated microbial majority (Lewis et al., 2020) will enable us to bring more natural abundant representatives to culture.

Placing archaeal and bacterial genome sizes in phylogenetic trees using GTDB-tk (**Figures 2A,B**) shows that the distribution of representative genomes and their estimated sizes varies widely between different phyla and within phyla. MAGs assigned to eight phyla in the domain Archaea were reconstructed exclusively from aquatic ecosystems, whereas eight other archaea phyla were reconstructed from multiple ecosystems. There was no significant difference between the genome sizes of aquatic archaea phyla or those from non-specific ecosystem (**Figure 2C**). However, estimated genome sizes in bacterial phyla were significantly larger than those in archaeal phyla. Moreover, genera from phyla with genome sizes below 3 Mbp, such as Halobacteriota, Thermoproteota, and Patescibacteria, do not show genome size variation in different ecosystems (**Figures 2D,E,I**). Nevertheless, genera from these smaller genome sizes phyla are significantly smaller than genera with more genome size variation in any ecosystem category (**Figures 2K–N**). For phyla with genome sizes above 3 Mbp, the genome sizes in aquatic or host-associated genera are significantly smaller than those in terrestrial or non-specific ecosystems (**Figures 2F–J**). We observe that while the microorganisms' ecosystem can certainly be linked to genome size, phyla where genome sizes are mostly below 3 Mbp show no variation in estimated genome size across ecosystems.

Clustering microorganisms together by the three ecosystem categories is not optimal since each contains innumerable niches. In each niche, there will be different selective pressures on the genome size. An example is clearly shown in a study (Nayfach and Pollard, 2015) in which it is observed that Archaea and Bacteria sampled from different parts of the human body have differences in genome sizes. Low metadata resolution and clustering of all genomes into three main ecosystem types might be a reason why we see a range of genome sizes in the genera of different ecosystems (**Figure 2**). With more precise metadata and higher sampling resolution of microhabitats, it might be possible to identify the ecological drivers of genome sizes in the different niches in fine-scale.

IMPACT OF ECOSYSTEM AND TROPHIC STRATEGY ON GENOME SIZE

Terrestrial ecosystems harbor immense microbial diversity (Delgado-Baquerizo et al., 2018). Yet, the most up-to-date data compilation provided here shows only 2033 MAGs from terrestrial ecosystems (**Figure 1C**) with an average genome size of 3.7 Mb (**Figure 1A**). The sub-ecosystems considered in this view are soil and deep subsurface, among others (**Supplementary Figure 2**). While the terrestrial microorganism's genome size is the biggest of the three ecosystem categories in this review, they are smaller than expected based on previous metagenomic predictions, which placed the genome size of soil bacteria at 4.74 Mb (Raes et al., 2007). Trends for larger genome sizes in soil have been hypothesized to be related to scarcity and high diversity of nutrients, and a fluctuating environment combined with little penalty for the slow growth rate (Konstantinidis and Tiedje, 2004; Cobo-Simon and Tamames, 2017; Chen et al., 2021). Although terrestrial environments are physically structured, they are generally characterized by two to three orders of magnitude greater variations (in temperature and currents) than marine environments (Steele et al., 2019). *In silico* studies predict that large genome sizes could result from higher environmental variability (Bentkowski et al., 2015). A recent example showed that isolates of terrestrial Cyanobacteria have genomes on the larger size scale (6.0–8.0. Mb) that are enriched in genes involved in regulatory, transport and motility functions (Chen et al., 2021). These functional categories enable them to thrive in a fluctuating environment with high nutrient diversity. Despite these general trends showing larger genome sizes in terrestrial ecosystems, it is worth noting that the diversity captured in the GEMs survey is probably a small fraction of the total terrestrial microbial diversity. It is, for example, also known that streamlined microorganisms such as Patescibacteria (Ortiz et al., 2021) and "*Candidatus* Udaeobacter copiosus" (Verrucomicrobiota) are abundant in soils (Brewer et al., 2016). Therefore, we predict that the view on genome size distribution and microbial diversity in terrestrial ecosystems will become more complete with more sequencing, assembly, binning and novel isolation efforts.

In host-associated microbiomes, genetic drift, deletion biases and low populations sizes drive the reduction of genomes (Li et al., 2021). In these environments, the differing levels of intimacy with their host can influence the evolution of the genome of microorganisms. For example, within the

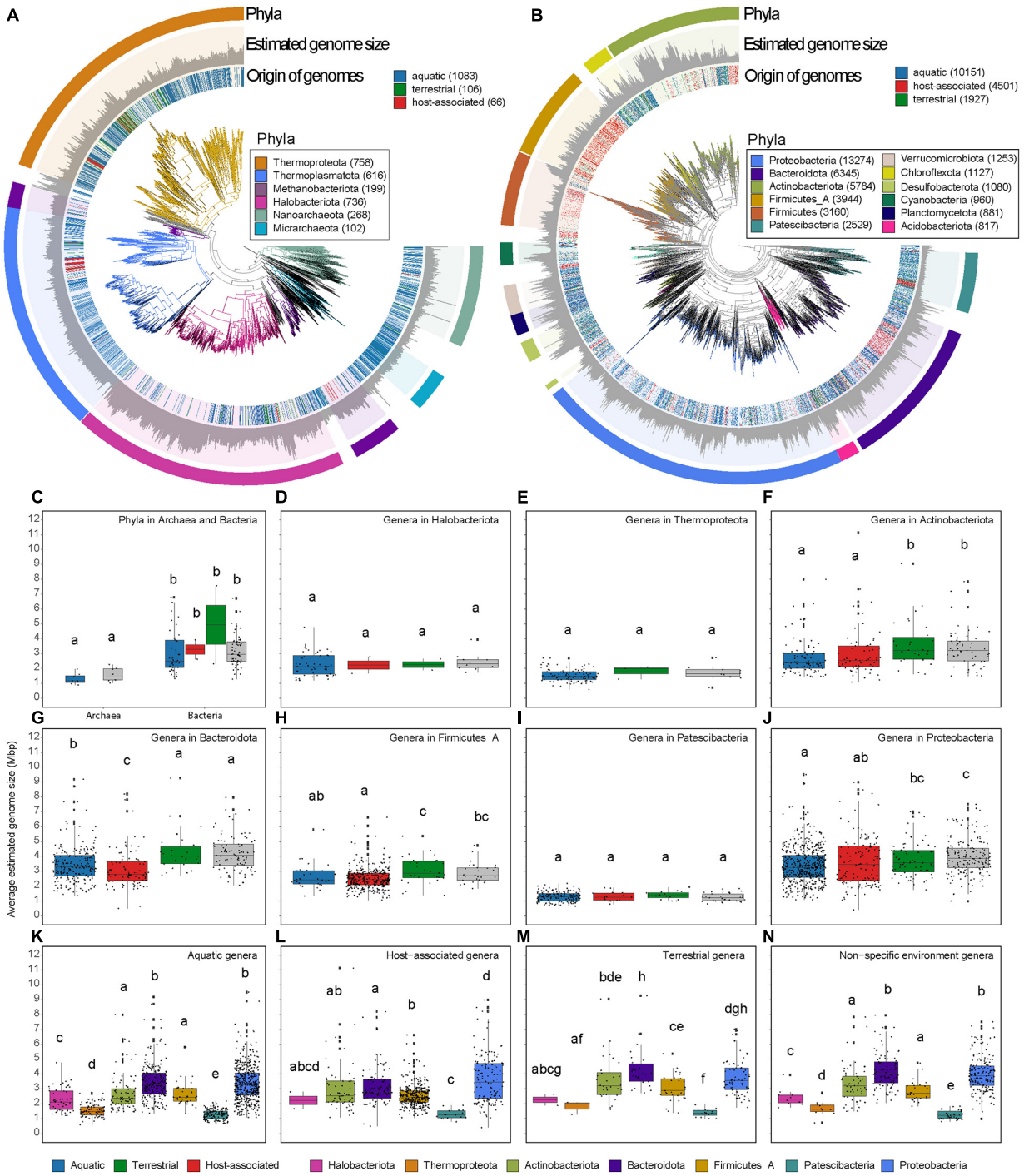

FIGURE 2 | Phylogenetic trees of archaeal **(A)** and bacterial **(B)** representative genomes show variation in genome size between and within phyla. The trees were constructed using GTDB-tk (v 1.5.0) using *de novo* workflow using aligned concatenated set of 122 and 120 single copy marker proteins for Archaea and Bacteria, respectively (Chaumeil et al., 2020). Moreover, in this mode, GTDB-tk adds 1,672 and 30,238 backbone genomes for Archaea and Bacteria, respectively. Tree is visualized in anvi'o (Eren et al., 2021). Estimated genome size is presented in scale from 0 to 6 Mpb or 0 to14 Mbp for Archaea or Bacteria. In the tree, the origin of the environmental genomes is labeled: aquatic, terrestrial and host-associated (same MAGs as **Figure 1**). Highlighted phyla with more representative genomes are color-coded. Boxplots show the average estimated genome size per phyla within archaeal and bacterial **(C)** domains. The average estimated size per genus within Halobacteriota **(D)**, Thermoproteota **(E)**, Actinobacteriota **(F)**, Bacteroidota **(G)**, Firmicutes A **(H)**, Patescibacteria **(I)**, Proteobacteria **(J)**. The presence of phyla and genera is colored in gray if they contain MAGs from different ecosystem category (non-specific ecosystem). The average estimated size per genus extracted from aquatic ecosystems **(K)**, host-associated ecosystems **(L)**, terrestrial ecosystems **(M)**, or non-specific ecosystems **(N)**. Letters in boxplot panels are the result of non-parametric tests, Wilcoxon **(C)** or Kruskal-Wallis **(D–N)**. Different letters show significant differences $p < 0.05$ (all statistical test with multiple testing were corrected with Benjamini-Hochberg).

Chlamydiaceae family, some lineages have evolved intracellular associations with eukaryotes (Toft and Andersson, 2010; Collingro et al., 2020). These intracellular Chlamydiaceae have lost many genes that were likely present in their common ancestor that lived in the environment (Dharamshi et al., 2020). Moreover, host-associated bacterial genomes show a variation in size depending on the type of host (plant, animal, etc.) and the type of association they have with the host, such as endosymbiotic, ectobiotic, or epibiotic (**Supplementary Table 1**). Generally, microorganisms associated with Arthropoda (Tamas et al., 2002), humans (McLean et al., 2020) and other mammals show smaller genomes sizes, whereas protist- and plant-associated bacteria present larger genomes (Levy et al., 2017; **Supplementary Figure 2**). In fact, *in silico* studies of Alphaproteobacteria show massive genome expansions diversifying plant-associated Rhizobiales and extreme gene losses in the ancestor of the intracellular lineages Rickettsia, Wolbachia, Bartonella, and Brucella that are animal- and human-associated (Boussau et al., 2004). Although host-associated microorganisms are widely known for their reduced genomes, the characteristics of host-associated MAGs show coding densities of ∼91% for genomes below 2 Mbp (**Figure 1D**).

Small genomes exhibit either strong dependency on other community members or have specific nutrient requirements. Two diverging views on genome reduction have emerged to explain mechanisms of gene loss. On the one hand, genetic drift is more pronounced in species that have a small effective population size, such as host-associated endosymbiotic microorganisms. These microorganisms might thrive because hosts provide energy or nutrients. On the other hand, streamlining is the process of gene loss through selection and it is mainly observed in free-living microorganisms with high effective population sizes (Giovannoni et al., 2014). Some of the most numerically abundant and streamlined microorganisms known to date, such as Pelagibacter (class Alphaproteobacteria) (Giovannoni et al., 2014), *Prochlorococcus* (phylum Cyanobacteria) (Rocap et al., 2003) Thermoproteota (Aylward and Santoro, 2020) and Patescibacteria (Tian et al., 2020), are commonly found in aquatic niches. Paradoxically, even though these microorganisms are free-living, their small genomes increase their nutritional connectivity to other individuals (Giovannoni et al., 2014). Free-living aquatic microorganisms have been used as exemplary streamlining cases in which many have gone through community adaptive selections and gene loss (Morris et al., 2012). Genome reduction can be so intense that microorganisms lose the capacity to biosynthesize essential metabolites and, thus, become auxotrophs. To overcome required nutritional needs, microorganisms thrive in functional cohorts (Mondav et al., 2020). As opposed to prototrophic lifestyle, the auxotrophic lifestyle is reflected by smaller genome sizes (Grote et al., 2012; Garcia et al., 2015; Brewer et al., 2016; Kang et al., 2017; **Supplementary Table 1**). An opportunity for future studies includes research on auxotrophy prevalences across the entire spectrum of metabolites (amino acids, nucleotides, fatty acids, vitamins, etc.) in different microbial communities and how those auxotrophies are linked with genome size.

In this review, the largest fraction of MAGs is recovered from aquatic environments. The two main sub-ecosystems in our survey are freshwater with MAGs estimated average genome size of 3.2 Mbp significantly different ($p < 0.0001$) from marine genome size distribution with average estimated genome size of 2.9 Mbp. When comparing freshwater and marine environments, the most obvious difference is salinity followed by nutrient concentration. Further exploring the impact of differing levels of salinity on genome size is an interesting research prospect. Additionally, we compared the union of representative freshwater MAGs from both databases (StratfreshDB and GEMs) (**Supplementary Figure 3**). The difference of mean estimated genome size between the representatives from freshwater GEMs and StratfreshDB is 0.52 Mbp. However, this is because each database captures genetic units that were not found in the other database.

In general, aquatic environments are vertically structured by gradients of light penetration, temperature, oxygen, and nutrient (**Supplementary Table 1**). Moreover, microorganisms might experience a microscale spatial and nutrient structure due to the presence of heterogeneous particles. These aquatic structures are drivers of the genetic repertoire of aquatic microorganisms. Metagenomic sequencing reported the increase of genome sizes for Archaea and Bacteria with increasing depths (Mende et al., 2017). Temperature may be as important; for example, a study based on twenty-one Thermoproteota and Euryarchaeota fosmids (Euryarchaetoa currently reclassified into Methanobacteriota, Halobacteriota, and Nanohaloarchaeota) showed high rates of gene gains through HGT to adapt to cold and deep marine environments (Brochier-Armanet et al., 2011). It has been observed that light is a relevant driver of genome size in aquatic environments as it decreases with depth. Photosynthetic bacteria such as *Prochlorococcus* spp. are well-differentiated into a high-light adapted ecotype with smaller genome sizes (average 1.6 Mbp) and a low-light-adapted ecotype with a slightly bigger genome size (average 1.9 Mbp) (Berube et al., 2018; **Supplementary Table 1**). Limitation of nutrients such as nitrogen (Elser et al., 2007) might also be one of the central factors determining genomic properties (Grzymski and Dussaq, 2012). Nitrogen fixation is a complex process that requires a great amount of genes (Franche et al., 2008) and most nitrogen-fixing marine cyanobacteria have the largest genomes in its phylum (Bergman et al., 2013).

Diversity and quantity of nutrients might be two understudied factors that drive ecology and genome size evolution. A recent example shows that polysaccharide xylan triggers microcolonies, whereas monosaccharide xylose promotes solitary growth in Caulobacter (D'Souza et al., 2021). This is a striking example of how nutrient complexity can foster diverse niches for well-studied cells such as Caulobacter with genome size 4 Mbp. To fully understand the link between genome size and nutritional requirements of diverse environmental microorganisms, we need to systematically explore the ∼90% of molecules/metabolites still unknown (Wienhausen et al., 2017; Hawkes et al., 2018; Patriarca et al., 2020). The wide nutrient complexity in the environment might prompt microorganisms to shape their genome. Their genomic content and metabolic potential defines whether they

are capable to feed on the available nutrients, forcing them to develop dependencies with other community members in order to acquire energy and metabolic precursors. Metagenomics combined with metabolomics will provide an understanding of the link between genome size evolution of microorganisms and their nutritional and trophic strategy.

CONCLUSION

This review offers a broad overview of genome size distribution across three different ecosystem categories, showing that MAGs recovered from aquatic and host-associated ecosystems present smaller estimated genome sizes than those recovered from terrestrial ecosystems. Moreover, genomes obtained from environmental samples present a smaller estimated genome size than obtained by cultivation approaches. We find that the distribution of genome sizes across the phylogenetic tree of Archaea and Bacteria can be linked to the ecosystem type from which the microorganisms' genomes have been extracted (aquatic, host-associated or terrestrial). Finally, we review the ecological factors that may cause the varying sizes of genomes in different ecosystems. In comparison with the aquatic and host-associated ecosystems, terrestrial ecosystems might harbor microorganisms with bigger estimated genome sizes mainly due to higher fluctuations in this ecosystem. Host-associations might shape genomes sizes differentially based on the type of host and level of intimacy between the microorganisms and the host. Genomes in aquatic ecosystems might be shaped by vertical stratification of abiotic factors such as nutrient distribution, light penetration, and temperature. Moreover, different trophic strategies such as auxotrophies might be connected to smaller genome sizes. We expect that as the microbial ecology field keeps moving forward with sequencing, bioinformatics, chemical analysis, and novel cultivation techniques, we will get a deeper resolution on physicochemical, metabolic, spatial, and biological drivers of archaeal and bacterial genome sizes.

AUTHOR CONTRIBUTIONS

SG, AR-G, and JN conceptualized the literature and data review idea. JN, MB, and FS gathered the data. AR-G and JN performed data analysis. SG, AR-G, and MM drafted the first manuscript. All authors did literature searches, contributed to the writing, and editing of the manuscript.

FUNDING

This work was supported by SciLifeLab and Kungl. Vetenskapsakademiens stiftelser grant CR2019-0060. The computations and data handling were enabled by resources in the project SNIC 2021/6-99 and SNIC 2021/5-133 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. The work conducted by the U.S. Department of Energy Joint Genome Institute, an Office of Science User Facility, made use of the National Energy Research Scientific Computing Center and was supported under Contract No. DE-AC02-05CH11231.

ACKNOWLEDGMENTS

We are grateful to John Paul Balmonte, Sergio Tusso, and Alexander Probst for helpful discussions. AR-G thanks Fede Berckx for technical advice.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.761869/full#supplementary-material

REFERENCES

Abram, K., Udaondo, Z., Bleker, C., Wanchai, V., Wassenaar, T. M., Robeson, M. S. II., et al. (2021). Mash-based analyses of *Escherichia coli* genomes reveal 14 distinct phylogroups. *Commun. Biol.* 4:117. doi: 10.1038/s42003-020-01626-5

Aylward, F. O., and Santoro, A. E. (2020). Heterotrophic *Thaumarchaea* with small genomes are widespread in the dark ocean. *mSystems* 5, e00415–e00420. doi: 10.1128/mSystems.00415-20

Bentkowski, P., Van Oosterhout, C., and Mock, T. (2015). A model of genome size evolution for prokaryotes in stable and fluctuating environments. *Genome Biol. Evol.* 7, 2344–2351. doi: 10.1093/gbe/evv148

Bergman, B., Sandh, G., Lin, S., Larsson, J., and Carpenter, E. J. (2013). *Trichodesmium*–a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol. Rev.* 37, 286–302. doi: 10.1111/j.1574-6976.2012.00352.x

Berube, P. M., Biller, S. J., Hackl, T., Hogle, S. L., Satinsky, B. M., Becker, J. W., et al. (2018). Single cell genomes of *Prochlorococcus*. *Synechococcus*, and sympatric microbes from diverse marine environments. *Sci. Data* 5:180154. doi: 10.1038/sdata.2018.154

Boussau, B., Karlberg, E. O., Frank, A. C., Legault, B. A., and Andersson, S. G. (2004). Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9722–9727. doi: 10.1073/pnas.0400975101

Brewer, T. E., Handley, K. M., Carini, P., Gilbert, J. A., and Fierer, N. (2016). Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus'. *Nat. Microbiol.* 2:16198. doi: 10.1038/nmicrobiol.2016.198

Brochier-Armanet, C., Deschamps, P., Lopez-Garcia, P., Zivanovic, Y., Rodriguez-Valera, F., and Moreira, D. (2011). Complete-fosmid and fosmid-end sequences reveal frequent horizontal gene transfers in marine uncultured planktonic archaea. *ISME J.* 5, 1291–1302. doi: 10.1038/ismej.2011.16

Buck, M., Garcia, S. L., Fernandez, L., Martin, G., Martinez-Rodriguez, G. A., Saarenheimo, J., et al. (2021a). Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater lakes and ponds. *Sci. Data* 8:131. doi: 10.1038/s41597-021-00910-1

Buck, M., Mehrshad, M., and Bertilsson, S. (2021b). mOTUpan: a robust Bayesian approach to leverage metagenome assembled genomes for core-genome estimation. *bioRxiv* [Preprint]. doi: 10.1101/2021.06.25.449606

Carini, P., Steindler, L., Beszteri, S., and Giovannoni, S. J. (2013). Nutrient requirements for growth of the extreme oligotroph 'Candidatus Pelagibacter ubique' HTCC1062 on a defined medium. *ISME J.* 7, 592–602. doi: 10.1038/Ismej.2012.122

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927. doi: 10.1093/bioinformatics/btz848

Chen, M. Y., Teng, W. K., Zhao, L., Hu, C. X., Zhou, Y. K., Han, B. P., et al. (2021). Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. *ISME J.* 15, 211–227. doi: 10.1038/s41396-020-00775-z

Cobo-Simon, M., and Tamames, J. (2017). Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. *BMC Genomics* 18:499. doi: 10.1186/s12864-017-3888-y

Collingro, A., Kostlbacher, S., and Horn, M. (2020). *Chlamydiae* in the Environment. *Trends Microbiol.* 28, 877–888. doi: 10.1016/j.tim.2020.05.020

Cross, K. L., Campbell, J. H., Balachandran, M., Campbell, A. G., Cooper, S. J., Griffen, A., et al. (2019). Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat. Biotechnol.* 37, 1314–1321. doi: 10.1038/s41587-019-0260-6

Dedysh, S. N. (2011). Cultivating uncultured bacteria from northern wetlands: knowledge gained and remaining gaps. *Front. Microbiol.* 2:184. doi: 10.3389/fmicb.2011.00184

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-Gonzalez, A., Eldridge, D. J., Bardgett, R. D., et al. (2018). A global atlas of the dominant bacteria found in soil. *Science* 359, 320–325. doi: 10.1126/science.aap9516

Dharamshi, J. E., Tamarit, D., Eme, L., Stairs, C. W., Martijn, J., Homa, F., et al. (2020). Marine sediments illuminate chlamydiae diversity and evolution. *Curr. Biol.* 30, 1032–1048.e7. doi: 10.1016/j.cub.2020.02.016

D'Souza, G. G., Povolo, V. R., Keegstra, J. M., Stocker, R., and Ackermann, M. (2021). Nutrient complexity triggers transitions between solitary and colonial growth in bacterial populations. *ISME J.* 15, 2614–2626. doi: 10.1038/s41396-021-00953-7

Elser, J. J., Bracken, M. E., Cleland, E. E., Gruner, D. S., Harpole, W. S., Hillebrand, H., et al. (2007). Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol. Lett.* 10, 1135–1142. doi: 10.1111/j.1461-0248.2007.01113.x

Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., et al. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* 6, 3–6. doi: 10.1038/s41564-020-00834-3

Figueroa-Gonzalez, P. A., Bornemann, T. L. V., Adam, P. S., Plewka, J., Revesz, F., von Hagen, C. A., et al. (2020). *Saccharibacteria* as Organic carbon sinks in hydrocarbon-fueled communities. *Front. Microbiol.* 11:587782. doi: 10.3389/fmicb.2020.587782

Franche, C., Lindström, K., and Elmerich, C. (2008). Nitrogen-fixing bacteria associated with leguminous and non-leguminous plants. *Plant Soil* 321, 35–59. doi: 10.1007/s11104-008-9833-8

Garcia, R., Gemperlein, K., and Muller, R. (2014). *Minicystis rosea* gen. nov., sp. nov., a polyunsaturated fatty acid-rich and steroid-producing soil myxobacterium. *Int. J. Syst. Evol. Microbiol.* 64, 3733–3742. doi: 10.1099/ijs.0.068270-0

Garcia, S. L. (2016). Mixed cultures as model communities: hunting for ubiquitous microorganisms, their partners, and interactions. *Aquat. Microb. Ecol.* 77, 79–85. doi: 103354/ame01796

Garcia, S. L., Buck, M., McMahon, K. D., Grossart, H. P., Eiler, A., and Warnecke, F. (2015). Auxotrophy and intrapopulation complementary in the "interactome' of a cultivated freshwater model community. *Mol. Ecol.* 24, 4449–4459. doi: 10.1111/mec.13319

Garcia, S. L., Mehrshad, M., Buck, M., Tsuji, J. M., Neufeld, J. D., McMahon, K. D., et al. (2021). Freshwater chlorobia exhibit metabolic specialization among cosmopolitan and endemic populations. *mSystems* 6, e01196–e01220. doi: 10.1128/mSystems.01196-20

Garcia, S. L., Stevens, S. L. R., Crary, B., Martinez-Garcia, M., Stepanauskas, R., Woyke, T., et al. (2018). Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. *ISME J.* 12, 742–755. doi: 10.1038/s41396-017-0001-0

Giovannoni, S. J., Cameron Thrash, J., and Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *ISME J.* 8, 1553–1565. doi: 10.1038/ismej.2014.60

Grote, J., Thrash, J. C., Huggett, M. J., Landry, Z. C., Carini, P., Giovannoni, S. J., et al. (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *Mbio* 3, e00252–e00312. doi: 10.1128/mBio.00252-12

Grzymski, J. J., and Dussaq, A. M. (2012). The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J.* 6, 71–80. doi: 10.1038/ismej.2011.72

Gweon, H. S., Bailey, M. J., and Read, D. S. (2017). Assessment of the bimodality in the distribution of bacterial genome sizes. *ISME J.* 11, 821–824. doi: 10.1038/ismej.2016.142

Hawkes, J. A., Patriarca, C., Sjöberg, P. J. R., Tranvik, L. J., and Bergquist, J. (2018). Extreme isomeric complexity of dissolved organic matter found across aquatic environments. *Limnol. Oceanogr. Lett.* 3, 21–30. doi: 10.1002/lol2.10064

Henson, M. W., Pitre, D. M., Weckhorst, J. L., Lanclos, V. C., Webber, A. T., Thrash, J. C., et al. (2016). Artificial seawater media facilitate cultivating members of the microbial majority from the gulf of Mexico. *mSphere* 1, e28–e116. doi: 10.1128/mSphere.00028-16

Hoehler, T. M., and Jorgensen, B. B. (2013). Microbial life under extreme energy limitation. *Nat. Rev. Microbiol.* 11, 83–94. doi: 10.1038/nrmicro2939

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., et al. (2020). Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* 577, 519–525. doi: 10.1038/s41586-019-1916-6

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9:5114. doi: 10.1038/s41467-018-07641-9

Kang, I., Kim, S., Islam, M. R., and Cho, J.-C. (2017). The first complete genome sequences of the acI lineage, the most abundant freshwater Actinobacteria, obtained by whole-genome-amplification of dilution-to-extinction cultures. *Sci. Rep.* 7:42252. doi: 10.1038/sre42252

Konstantinidis, K. T., and Tiedje, J. M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 3160–3165. doi: 10.1073/pnas.0308653100

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2567–2572. doi: 10.1073/Pnas.0409727102

Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., et al. (2017). Genomic features of bacterial adaptation to plants. *Nat. Genet.* 50, 138–150. doi: 10.1038/s41588-017-0012-9

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., and Ettema, T. J. G. (2020). Innovations to culturing the uncultured microbial majority. *Nat. Rev. Microbiol.* 19, 225–240. doi: 10.1038/s41579-020-00458-8

Li, L., Liu, Z., Zhou, Z., Zhang, M., Meng, D., Liu, X., et al. (2021). Comparative genomics provides insights into the genetic diversity and evolution of the DPANN superphylum. *mSystems* 6:e00602211. doi: 10.1128/mSystems.00602-21

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. Q., and Crosby, L. (2018). Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *mSystems* 3, e55–e118. doi: 10.1128/mSystems.00055-18

Mallet, J. (2008). Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 363, 2971–2986. doi: 10.1098/rstb.2008.0081

McLean, J. S., Bor, B., Kerns, K. A., Liu, Q., To, T. T., Solden, L., et al. (2020). Acquisition and adaptation of ultra-small parasitic reduced genome bacteria to *Mammalian* hosts. *Cell Rep.* 32:107939. doi: 10.1016/j.celrep.2020.107939

Mende, D. R., Bryant, J. A., Aylward, F. O., Eppley, J. M., Nielsen, T., Karl, D. M., et al. (2017). Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat. Microbiol.* 2, 1367–1373. doi: 10.1038/s41564-017-0008-3

Meziti, A., Rodriguez, R. L., Hatt, J. K., Pena-Gonzalez, A., Levy, K., and Konstantinidis, K. T. (2021). The reliability of metagenome-assembled genomes (mags) in representing natural populations: insights from comparing mags against isolate genomes derived from the same fecal sample. *Appl. Environ. Microbiol.* 87, e02593–e02620. doi: 10.1128/AEM.02593-20

Mondav, R., Bertilsson, S., Buck, M., Langenheder, S., Lindstrom, E. S., and Garcia, S. L. (2020). Streamlined and abundant bacterioplankton thrive in functional cohorts. *mSystems* 5, e00316–e00420. doi: 10.1128/mSystems.00316-20

Moran, N. A., and Bennett, G. M. (2014). The tiniest tiny genomes. *Annu. Rev. Microbiol.* 68, 195–215. doi: 10.1146/annurev-micro-091213-112901

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The black queen hypothesis: evolution of dependencies through adaptive gene loss. *Mbio* 3, e00036–e00112. doi: 10.1128/mBio.00036-12

Nayfach, S., and Pollard, K. S. (2015). Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* 16:51. doi: 10.1186/s13059-015-0611-7

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., et al. (2020). A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* 39, 499–509. doi: 10.1038/s41587-020-0718-6

Nelson, W. C., Tully, B. J., and Mobberley, J. M. (2020). Biases in genome reconstruction from metagenomic data. *PeerJ.* 8:e101191.

Olm, M. R., Crits-Christoph, A., Diamond, S., Lavy, A., Matheus Carnevali, P. B., and Banfield, J. F. (2020). Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems* 5, e00731–e00819. doi: 10.1128/mSystems.00731-19

Ortiz, M., Leung, P. M., Shelley, G., Jirapanjawat, T., Nauer, P. A., Van Goethem, M. W., et al. (2021). Multiple energy sources and metabolic strategies sustain microbial diversity in Antarctic desert soils. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2025322118. doi: 10.1073/pnas.2025322118

Parks, D. H., Chuvochina, M., Chaumeil, P. A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and *Archaea*. *Nat. Biotechnol.* 38, 1079–1086. doi: 10.1038/s41587-020-0501-8

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Patriarca, C., Sedano-Núñez, V. T., Garcia, S. L., Bergquist, J., Bertilsson, S., Sjöberg, P. J. R., et al. (2020). Character and environmental lability of cyanobacteria-derived dissolved organic matter. *Limnol. Oceanogr.* 66, 496–509. doi: 10.1002/lno.11619

Raes, J., Korbel, J., Lercher, M., von Mering, C., and Bork, P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biol.* 8:R10. doi: 10.1186/gb-2007-8-1-r10

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., et al. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424, 1042–1047. doi: 10.1038/nature01947

Rodriguez, R. L., Castro, J. C., Kyrpides, N. C., Cole, J. R., Tiedje, J. M., and Konstantinidis, K. T. (2018). How much do rRNA gene surveys underestimate extant bacterial diversity? *Appl. Environ. Microbiol.* 84, e00014–e00118. doi: 10.1128/AEM.00014-18

Rodriguez, R. L., Jain, C., Conrad, R. E., Aluru, S., and Konstantinidis, K. T. (2021). Reply to: "re-evaluating the evidence for a universal genetic boundary among microbial species". *Nat. Commun.* 12:4060. doi: 10.1038/s41467-021-24129-1

Shade, A., Hogan, C. S., Klimowicz, A. K., Linske, M., McManus, P. S., and Handelsman, J. (2012). Culturing captures members of the soil rare biosphere. *Environ. Microbiol.* 14, 2247–2252. doi: 10.1111/j.1462-2920.2012.02817.x

Shapiro, B. J., and Polz, M. F. (2015). Microbial speciation. *Cold Spring Harb. Perspect. Biol.* 7:a0181431.

Steele, J. H., Brink, K. H., and Scott, B. E. (2019). Comparison of marine and terrestrial ecosystems: suggestions of an evolutionary perspective influenced by environmental variation. *ICES J. Mar. Sci.* 76, 50–59. doi: 10.1093/icesjms/fsy149

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., Gonzalez, J. M., et al. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11463–11468. doi: 10.1073/pnas.1304246110

Tamas, I., Klasson, L., Canback, B., Naslund, A. K., Eriksson, A. S., Wernegreen, J. J., et al. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science* 296, 2376–2379.

Tian, R., Ning, D., He, Z., Zhang, P., Spencer, S. J., Gao, S., et al. (2020). Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome* 8:51. doi: 10.1186/s40168-020-00825-w

Toft, C., and Andersson, S. G. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat. Rev. Genet.* 11, 465–475. doi: 10.1038/nrg2798

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., et al. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* 43, 6761–6771. doi: 10.1093/nar/gkv657

Wienhausen, G., Noriega-Ortega, B. E., Niggemann, J., Dittmar, T., and Simon, M. (2017). The Exometabolome of two model strains of the roseobacter group: a marketplace of microbial metabolites. *Front. Microbiol.* 8:1985. doi: 10.3389/fmicb.2017.01985

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.