Open access • Posted Content • DOI:10.1101/2021.01.18.427069

A genomic perspective on genome size distribution across Earth's microbiomes reveals a tendency to gene loss — Source link

Alejandro Rodríguez-Gijón, Julia K. Nuy, Maliheh Mehrshad, Moritz Buck ...+3 more authors

Institutions: Stockholm University, Swedish University of Agricultural Sciences, Joint Genome Institute

Published on: 18 Jan 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Bacterial genome size, Genome size, Genome and Microbiome

Title: A genomic perspective on genome size distribution across Earth's microbiomes reveals a tendency to gene loss

Running title: Archaea and Bacteria genome size distribution

Authors: Alejandro Rodríguez-Gijón #[a] (alejandro.gijon@su.se), Julia K. Nuy #[a] (julia.nuy@su.se), Maliheh Mehrshad [b] (maliheh.mehrshad@slu.se), Moritz Buck [b] (moritz.buck@slu.se), Frederik Schulz [c] (fschulz@lbl.gov), Tanja Woyke [c] (twoyke@lbl.gov), Sarahi L. Garcia [a*] (sarahi.garcia@su.se).

[a] Department of Ecology, Environment, and Plant Sciences, Science for Life Laboratory, Stockholm University, 10691 Stockholm, Sweden
[b] Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Lennart Hjelms väg 9, 75651 Uppsala, Sweden
[c] DOE Joint Genome Institute, Berkeley, CA 94720, USA

#Equal contributions
*Corresponding author

Article type: Review article

Abstract

Our view of genome size distribution in Bacteria and Archaea has remained skewed as the data used to paint its picture has been dominated by genomes of microorganisms that can be cultivated under laboratory settings. However, the continuous effort to catalogue the genetic make-up of Earth's microbiomes, specifically propelled by recent extensive work on uncultivated microorganisms, provides a unique opportunity to revise our perspective on genome size distribution. Genome size is largely a function of the expansion and contraction, by gain or loss of DNA elements. While genome expansion provides microorganisms the capability to acquire a wide repertoire of ecological functions, genome reduction increases the fitness of the microorganisms to very specific niches. Capitalizing on a recently released large catalog of tens of thousands of metagenome-assembled genomes, we here provide a comprehensive overview of genome size distributions, suggesting that the known phylogenetic diversity of environmental microorganisms possess significantly smaller genomes (aquatic bacteria average 3.1 Mb, host-associated bacterial genomes average 3.0 Mb, and terrestrial bacteria average 3.8 Mb) than the collection of laboratory isolated microorganisms (average 4.4 Mb). Moreover, the variation in genome sizes across different types of environments reflects the different ecological and evolutionary strategies used by microorganisms to thrive in their native environment. Finally, the fact that genome sizes in Bacteria and Archaea remain relatively small might be a reflection of the constraints imposed by selection and an overall dominance of gene loss as a survival strategy.

Introduction

Genomes are dynamic databases that encode the machinery behind evolution and adaptation of living organisms to environmental settings. In brief, a genome encompasses all genetic material present in one organism and includes both its genes and its non-coding DNA. Genome size is largely a function of expansion and contraction by gain or loss of DNA fragments. The genomes of extant organisms are the result of a long evolutionary history. In eukaryotes, an organism's complexity is not directly proportional to its genome size which can have variations over 64,000-fold (1, 2). However, the genome size ranges in Bacteria and Archaea are smaller and the genomes are information-rich (3), and known to range from 100 kb to 16 Mb (4, 5). While subject to genetic drift bacterial and eukaryotic genomes evolve in opposite directions. Bacteria exhibit a mutational bias that deletes superfluous sequences, whereas Eukaryotes are biased toward large insertions (6). In Bacteria and Archaea, evolutionary studies have revealed extremely rapid and highly variable flux of genes (7) with evolutionary forces acting on individual genes (8). On one hand, mechanisms for genome expansion are promoting the gain of new functions through horizontal gene transfer, *de novo* gene birth, and gene duplications (9, 10). On the other hand, it is known that the primary driving forces for genome reduction are metabolic and spatial economy and cell multiplication speed (11, 12).

Our overall view of the diversity, distribution, and genome characteristics of Bacteria and Archaea have remained biased for most of the microbial ecology history. These biases stem chiefly from the "great plate count anomaly" because for more than a decade, the genomes that were sequenced were primarily from laboratory isolates (13, 14). More than two decades have passed since the first bacterial genomes were completely sequenced (15, 16). In the first decade of genome sequencing about 300 bacterial genomes and two metagenomic projects with assembled genomes were published (17). Since then, rapid advances in metagenome sequencing and data analyses have enabled large-scale cataloging of bacterial and archaeal genomes from a wide range of environments (18-20). With all sequencing efforts the representation of number of genomes from phylogenetically diverse groups of Bacteria and Archaea has greatly increased. The Genome Taxonomy Database (GTDB) include 194,600 genomes, with 31,910 of those being species representatives and 8,792 of those species representatives are based on published named species (21). Genome catalogs such as Genomes from Earth's Microbiomes (GEMs) contain ~52,500 genomes all of them being metagenome-assembled genomes. Using these novel resources, it is now possible to obtain an updated view of microbial genome characteristics, diversity, and distribution of microbes in the environment.

Genome size and its evolution has been studied by many researchers who each focused on different taxonomic lineages or different ecological or evolutionary backgrounds (8, 12, 22-25). As microbial researchers, how do we define what is a small genome or a big genome? Perhaps, researchers working on model organisms such as *Escherichia coli* with a genome size of ~5 Mb (26), would define 'big' or 'small' very differently to researchers working on *Prochlorococcus* with a genome size of ~2 Mb (27), soil-dwelling *Minicystis rosea* with a genome size of 16 Mb (5) or bacterial endosymbionts of insects that may have genomes merely larger than 100 kb (4). The recently published expanded database of environmental bacterial and archaeal genomes (18) allows us to revisit and acquire a more complete understanding of genome size distribution across different environments in higher resolution. In this review, we provide an overview of the evolutionary and ecological drivers behind the different genome sizes of Bacteria and Archaea. Moreover, we offer an overview of the distribution of genome sizes of all known bacterial and archaeal phyla across different environments. We found that while there are phyla with consistently smaller genome sizes (< 2 Mb), such as Caldisericota, Aenigmarchaeota, Micrarchaeota, Nanohaloarchaeota, and Ianarchaeota, 78.4% of bacterial and archaeal genomes recovered through genome-resolved metagenomics represent estimated genome sizes below 4 Mb.

Extant genome size distribution in the environment

The current state of environmental sequencing, assembly, and binning technologies allows us to review and renew our view of bacterial and archaeal genome size distribution on Earth (18). To minimize representation biases (28), from the ~52,500 genomes we included one representative per mOTU, defined by 95% average nucleotide identity (ANI), from the GEMs environmental MAGs resulting in ~15,000 MAGs (Figure 1A). We complemented these data by adding ~8,000 species cluster representatives from >90% complete genomes of isolates from GTDB (Figure 1). GEMs reported that MAGs in the same species than isolate genomes were consistent in size (average estimated genome length per OTU MAGs = -0.17 + 1.01 average estimated genome length per OTU isolates, r=0.95) (18).While this suggests that there is not a big bias in metagenome assembly and binning it is important to keep in mind that MAG assembly might discriminate against ribosomal RNAs, transfer RNAs, mobile element functions and genes of unknown function (29).

Furthermore, we compared the genome size distribution of all environmental MAGs versus that of their taxonomic relatives from cultivated isolates, as derived from the GTDB. The genomes from bacterial isolates have the average genome size of 4.4 Mb. When comparing this genome size distribution from isolates with that of the environmental MAGs, the first striking observation is that environmental MAGs have significantly lower genome sizes (t-test Bacteria p>2e-16

Archaea p>2e-16). Environmental aquatic bacteria average 3.1 Mb, host-associated metagenome-assembled bacterial genomes average 3.0 Mb, and terrestrial bacteria average 3.8 Mb (Figure 1A). A reason for the difference in genome size between isolates and environmental microorganisms might be the tendency to sample different types of microorganisms with culture dependent and independent methods (30). For example, it is known that current cultivation techniques with rich media bias cultivation towards copiotroph microorganisms (31). Moreover, microorganisms in nature do not live in isolation but instead have coevolved with other microorganisms and might have specific requirements that are hard to meet in batch-culture standard-media isolation techniques (32). Other reasons for biases in cultivation include slow growth of microorganisms (33), host dependence (34), dormancy (35), and microorganisms with very limited metabolic capacity (36) among others. Innovations to culturing the uncultured microbial majority might help breach this genome size gap in the future (37).

Placing bacterial and archaeal genome sizes in the context of a phylogenetic tree (Figure 2 and 3) shows that the distribution of representative genomes and its sizes vary widely not only between different phyla but widely within different phyla. To this view, we want to bring into the discussion the biphasic model of evolution (25). In this model it is discussed that genome evolution occurs in two phases. One phase involves gene gains that occur in bursts and are associated with the emergence of novel microbial groups. The other phase involves gene loss that occurs gradually. In the extant phylogenetic tree of Bacteria and Archaea it is noticeable how closely related species are shaped by the different genetic processes that influence genome size (Figure 2 and 3).

Genetic processes that shape genome size

The variability in genome size that we observe across different microbial taxa is the result of the reached equilibrium between gains and losses of genetic information (Figure 4). The evolutionary events that drive these changes are diverse. Some lineages follow a highly mutational mode of evolution (38) while other lineages have recombination as a stronger evolutionary force (39-42). The acquisition of new genetic information and metabolic capacities is often accompanied by the expansion of gene families. *In silico* studies indicate that the acquisition of new genes could have a vital role in adaptation (43). Moreover, a strong correlation has been observed between genome size with gene family expansions and length of non-coding sequences in complex cyanobacteria (44). The most important evolutionary events involved in genome expansion processes are *de novo* gene birth, the duplication of genes, and Lateral/Horizontal Gene Transfer (LGT/HGT). *De novo* gene birth is the process by which new genes emerge from non-genic DNA sequences (45). However, most of the known examples of this process are found in eukaryotes. Furthermore, comparative genomics of some bacterial taxonomic lineages has suggested that HGT is more relevant on the expansion of bacterial metabolic networks than gene duplications (46). HGT can foster the acquisition of new functions while duplications relate to a higher gene dosage (47). However, phylogenomic analysis of other lineages such as Nitrososphaerales (Thermoproteota) indicate a predominant role of gene duplications over HGT (48). These examples highlight how HGT and duplications aid microorganisms into adaptation to their niches. For example, different Archaea have shown modifications in their metabolic potential through these genome expansion processes (49-51). In a nutshell, genome expansion processes can influence gene dosage, acquisition of new ecological capacities and adaptation in both, Archaea and Bacteria.

Conversely, genome reduction fosters the development of more compact genomes (Figure 4). There are three main processes involved in gene loss: genetic drift, pseudogenization and streamlining. Genetic drift describes stochastic changes on the gene repertoire variants. Mutations which are biased towards deletions over time promote genome reduction (8). Genetic drift is more pronounced in species that have a small effective population size such as host-associated endosymbiotic microorganisms. As an example, endosymbiotic lineages of Gammaproteobacteria such as in *Buchnera aphidicola* have lost ample genes that have already reached stasis (52). When a gene loses its original function, it is often turned into a pseudogene (53). A pseudogene is a derived form of regular genes that might present a different function or turn obsolete. Comparative analyses of archaeal genomes show that up to 8.6% of their genomes are constituted by pseudogenes which usually present at least one inactivating mutation (54). Moreover, pseudogenization has been suggested to be a special type of gene loss when adaptation to new ecological niches is needed. In the *Roseobacter* lineage (class Alphaproteobacteria), this process was correlated to switches in resource recovery, energy conservation, stress tolerance and different metabolic pathways (55). Finally, streamlining is the process of gene loss through selection and it is mainly observed in free-living microorganisms with high effective population sizes. Streamlining creates a series of distinct patterns, such as increase in nutritional connectivity between individuals, reduction of genome size, lower GC content and higher coding gene density (12). Aquatic microorganisms have been used as exemplary cases of streamlining in which many have gone through community adaptive selections and gene loss (56). In fact, their gene loss goes so far that these free-living aquatic microorganisms depend on community associations and thus thrive in functional cohorts (57). The renewed view of genome sizes and characteristics confirms that genomes from aquatic microorganisms have a higher coding density compared to those from other ecosystems (Figure 1).

Both genome expansion and reduction have a vital role in the evolution of microorganisms. However, these two processes are not in perfect equilibrium. While genome expansion might allow cells to become highly flexible in terms of developmental capacities and physiological performance, gene loss allows cells to become highly successful in particular niches (44). Moreover, gene loss dominates in the evolutionary history of Bacteria and Archaea (25). For example, *in silico* studies of 34 bacterial genera and one archaeal genus show that the rate of gene gains is three times lower than that of gene loss (7). In the same study, highly dynamic genomes were found presenting these evolutionary events 25 times more often than the most stable genomes. This tendency that gene loss is more prevalent than gene gain has also been described in short term in host-associated *Pseudomonas aeruginosa* (Gammaproteobacteria) (58). There, clinical isolates of *P. aeruginosa* indicate gene loss rates six times greater than gene acquisition during the first year of a chronical infection as an adaptive strategy to avoid the host's immune response. Even evolutionary reconstructions of the Last Common Ancestor of Archaea show that genomes of early Archaea were more complex and thus gene loss played likely a critical role in their evolution (59, 60). In summary, these examples illustrate the synergy of both evolutionary processes, with genome expansion providing microorganisms the capability to acquire a wide repertoire of ecological adaptations and genome reduction increasing the fitness of the microorganisms to very specific niches (Figure 4).

Environmental impact on genome size in different taxonomic lineages

The most up-to-date view of genome sizes on Earth provided here shows that the distribution of genomes from terrestrial environments average at size of 3.8 Mb (Figure 1). The sub-ecosystems considered in this view are soil, and deep subsurface among others (Figure 5). Terrestrial microorganism's genome size is larger than what is commonly found in aquatic and host-associated ecosystems. However, it is smaller than expected based on previous metagenomic predictions which placed the genome size of soil bacteria at 4.74 Mb (61). Trends of larger genome sizes in soil have been hypothesized to be related to scarcity and diversity of nutrients, fluctuating environment combined with little penalty for the slow growth rate (23, 62, 63). In fact, although terrestrial or soil environments are physically structured, they are generally characterized by two to three orders of magnitude greater variations (in temperature and currents) than marine environments (64). *In silico* studies predict that large genome sizes could be the result of higher environmental variability (65). A recent example showed that isolates of terrestrial Cyanobacteria have larger genomes (6.0-8.0. Mb), as compared to their freshwater counterparts (4.0-6.0 Mb) and their relatives originating from the marine environment (1.5-2.5 Mb) (62). The general theory of genome expansion states that the genetic repertoire increases to allow microorganisms to gain adaptive capacities to face perturbations and survive in variable environments. Despite these general trends showing larger genome sizes in terrestrial environments, it is worth noting that streamlined microorganisms such Patescibacteria (Fig 1B) as '*Candidatus* Udaeobacter copiosus' (Verrucomicrobiota) are abundant in soils (66).

Some of the most numerically abundant and streamlined microorganisms known to date, such as Pelagibacter (class Alphaproteobacteria) (12), marine methylotrophs (class Gammaproteobacteria) (67), *Prochlorococcus* (phylum Cyanobacteria) (27) Thermoproteota (68) and Patescibacteria (69) are commonly found in aquatic niches. This is well reflected in the MAG data, illustrating that genomes from aquatic sites are among the smallest (Figure 1). Aquatic environments are less physically structured than soils. However, there is some vertical structure in physicochemical parameters connected to depth variables such as light penetration, temperature, oxygen, and nutrient gradients, as well as microscale spatial structure due to the presence of heterogeneous particles. Aquatic structures are drivers of the genetic repertoire of aquatic microorganisms. For instance, metagenomic sequencing reported the increase of genome sizes for Bacteria and Archaea with increasing depths (70). While several hypotheses have been proposed as drivers of such evolutionary trends, nutrient limitation might be one of the central factors determining genomic properties (71) (Table 1). Temperature might be as important, for example, a study based on twenty-one Thermoproteota and Euryarchaeota fosmids (Euryarchaetoa is now reclassified into Methanobacteriota, Halobacteriota and Nanohaloarchaeota) showed high rates of gene gains through HGT to adapt to cold and nutrient-depleted marine environments (72). Moreover, aquatic hyperthermophilic microbes show reduced genomes compared to those of microorganisms adapted to very cold environments (73-75) supporting this negative relation between temperature and genome size. One last driver we want to point out in aquatic environments is light which 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 Mb), and a low-light-adapted ecotype with slightly bigger genome size (average 1.9 Mb) (76).

In host-associated microbiomes, microorganisms are shaped in their ecological and evolutionary history by the differing levels of intimacy they might have with their host. For example, within the Chlamydiaceae family there are lineages that have evolved intracellular associations with eukaryotes (77, 78). Recent metagenomic studies uncovered that extensive metabolic capabilities were present in the common ancestor of environmental Chlamydiia (class) and subsequently lost in Chlamydiaceae (79). 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 (endosymbiotic, ectobiotic or epibiotic). Generally, microorganisms associated with Arthropoda (52), humans (80) and other mammals show smaller genomes whereas protist- and plant-associated bacteria present bigger genomes (81) (Figure 5). *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 (82). Within the Chloroflexota, genomes associated with plants or algae range between 4.75 and 7.5 Mb, and genomes associated with Arthropoda range between 0.75 and 1.75 Mb (Figure S1). Although microorganisms that are host-associated are widely known for their reduced genomes, the characteristics of metagenomic host-associated bacterial genomes show lower coding density than streamlined genomes in aquatic environments in the genome sizes ranged 1 – 4 Mb (Figure 1F). However, at size range below 1 Mb the MAGs and available genomes of endosymbionts are often reduced and at same time have high coding density of ~91% (Figure 1F) (83).

Table 1

Chemical, physical or biological variable influencing genome size		**Taxa**	**References**
Temperature	*Literature review indicates a negative correlation between genome size and temperature.*		
	Comparative genomic of genomes of hyperthermophilic microorganisms shows average genome sizes of about 2.3 Mb with very active horizontal gene transfer (HGT) mechanisms	*Thermus thermophilus* (phylum Deinococcota) *Thermus spp.*	(73, 74)
	Metagenomics suggest that gene gains would have played an important role in adaptation to low temperature and oligotrophic deep marine environments	Thermoproteota and Euryarchaeota (phyla)	(72)
	Comparative genomics of isolates in one genus indicate larger genomes in colder environments.	*Janthinobacterium* spp. (class Gammaproteobacteria)	(75)
	Environmental samples indicate that hypersaline environments could increase gene gain via HGT, whereas thermal environments decrease it.	Halobacteria and Thermoproteia (class)	(84)
Nutrients	*When talking about nutrients, diversity and quantity of nutrients are two factors that drive ecology and evolution. Some literature present conflicting results on how these two dimensions of nutrient influence genome sizes.*		
	Metagenomics indicate dominance of reduced genomes in the Baikal Lake. Small genomes are thought to reflect the extremely oligotrophic conditions.	Actinobacteria, Bacteroidetes, Cyanobacteria Verrucomicrobia and Thermoproteota	(85)
	Online databases indicate that larger	70 closely related bacterial	(23)

	genome-sized species may dominate environments where resources are scarce but diverse.	genomes	
	Phylogenomics of isolates show gene loss in functions like resource scavenging and energy acquisition when adapting to nutrient-rich environments in algae and corals.	Roseobacter spp. (class Alphaproteobacteria)	(55)
	Oceanic metagenomic data show positive correlation between nutrient concentration and genome size.	Different bacteria phyla	(86)
	Metagenomics indicates small genomes in mesopelagic environments are the result of adaptation to energy scarcity.	Some Thermoproteota (phylum)	(68)
	Whole-genome shotgun sequencing indicated that deep oligotrophic marine environments are dominated by large genomes with high GC content.	Lactobacillales (phylum Firmicutes)	(87)
	Oceanic metagenomic samples suggest that deeper areas with more nitrate and phosphate as nutrients are dominated by large genomes and high GC content.	Bacteria (SAR11, *Prochlorococcus* spp., *Roseobacter* spp., etc.,) and Archaea (Thermoproteota and Euryarchaeota)	(70)
Light	*In oxygenic phototrophs there is negative correlation between light irradiance and the genome size.*		
	Genomes of cultures and single cells show high-light-adapted ecotypes with smaller genome sizes and low-light-adapted ecotypes with bigger genomes.	*Prochlorococcus* spp. (phylum Cyanobacteria)	(27, 76, 88)
Particles	*Microorganisms with particle associated lifestyle tend to have larger genome sizes.*		
	Comparison of metagenomes in coastal ecosystems show larger genome sizes for particle associated microorganisms than free-living.	Metagenomic data	(89)
	Particle associated microbes have larger genome sizes than free-living bacteria.	Cyanobacteria and Bacteroidetes	(86)
Host-association	*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 (endosymbiotic, ectobiotic or epibiotic)*		
	In silico studies indicate massive genome expansions in plant-associated bacteria.	Alphaproteobacteria (class)	(82)
	Isolates from sugarcane (*Saccharum* sp.) rhizosphere and endophytic roots and stalks show 26 individual genomes of associated bacteria whose genomes ranged from 3.9 to 7.5 Mbp.	Diverse bacterial taxa (Burkholderiaceae, Rhizobiaceae, Caulobacteraceae, Xanthomonadaceae, etc.)	(90)
	Genomic comparison of 3837 bacterial genomes identified thousands of plant-associated gene clusters and found genomes of plant associated microorganisms tended to be larger	Diverse bacterial taxa	(81)
	Intense genome reduction in isolates of microbes associated with aphids (Arthropoda).	*Buchnera aphidicola* (class Gammaproteobacteria)	(52)

	In vitro cultures and metagenomic datasets indicate reduced genome sizes in microbes associated with humans and other mammmals	*Salmonella enterica* (class Gammaproteobacteria) Patescibacteria (phylum)	(80, 91)
	Environmental samples indicate that symbionts and epibionts of other microbes present highly reduced genomes.	Bacteria of the CPR clade (such as *Vampirococcus lugosii*) and Archaea of the DPANN	(92, 93)
Viruses	Marine isolates support the "Cryptic Escape Theory". In here small cell size is a strategy to minimize viral predation. This article also finds a correlation between genome size and cell size.	Different bacteria lineages (Cyanobacteria, Proteobacteria, Actinobacteria, among others)	(94)

Conclusion

Since the sequencing of the first isolate bacterial genomes in 1995, profound improvements in both sequencing technologies and bioinformatic analysis tools have accelerated our access to the genetic make-up of the uncultivated majority. This allowed us for the first time to provide a more global view of the distribution of bacterial and archaeal genomes from a wide array of microbiomes on Earth. In this review, we offer an overview where genomes obtained from environmental samples show to be smaller than those obtained from laboratory isolates. This is not because isolates and MAGs from the same species differed in size but because cultivation methods bias the sampling of nature towards obtaining copiotrophs, fast growers, and more metabolically independent microorganisms. Moreover, we find the distribution of genome sizes across the phylogenetic tree of Bacteria and Archaea reflects that genome evolution occurs in a gene gain phase and gene loss phase, as the biphasic model theory suggests. Finally, we review the ecological and evolutionary effectors causing the varying sizes of genomes in different environments. Soils might have the microorganisms with the bigger genome sizes due to higher environmental variability. Genomes in aquatic environments might be shaped by vertical stratification in nutrients, particles, and light penetration. Host-associations might shape genomes differentially based on the kind of relationship between the microorganisms and the host. We expect that as the microbial ecology field keeps moving forward, we get a deeper resolution on physicochemical, spatial, and biological drivers of bacterial and archaeal genome sizes.

Figures

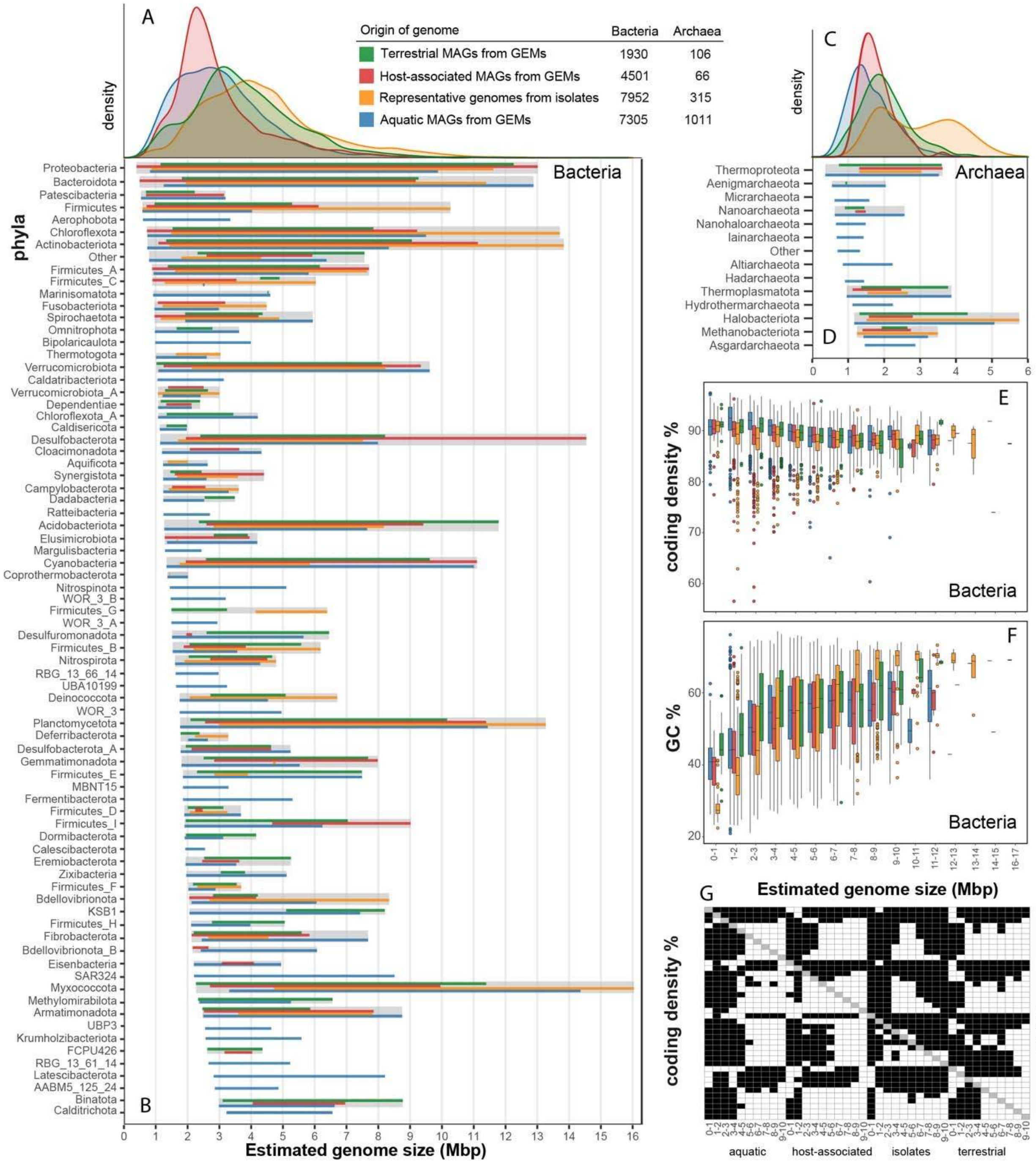

Figure 1. Overview of the genome size distribution across Earth's microbiomes. Genome size distribution of Bacteria [A] and Archaea [C] from different environmental sources and across different bacterial [B] and archaeal [D] phyla is shown for a total of 23,186 genomes. The coding density [E] and GC content (%) [F] is shown for the bacterial MAGs across different environments and isolates. Pair-wise t-test was performed in all variables of panel E and F and if the pair-wise comparison was significant ($p < 0.05$) it is shown in panel G in black. The figure was constructed in R (95) using representative isolate genomes from GTDB database as well as MAGs (metagenome assembled genomes) from GEMs catalog. The GEMs genomes available

were clustered into mOTUs (metagenomic operational taxonomic unit) at the threshold of the operational definition of species (95% ANI). To eliminate over-representation biases for some mOTUs, we used only one representative genome per mOTU from the GEMs catalog in the plots. We addressed the same bias for the GTDB database by selecting the representative isolate genome per species cluster that were circumscribed based on the ANI (>=95%) and alignment fraction ((AF) >65%) between genomes (21). To construct the figures, we plotted the estimated genome sizes which was calculated based on the genome assembly size and completeness estimation provided. In panel B, 'other' includes 45 phyla all with less than 5 genomes. For a complete list of bacterial phyla please see Figure S2. In panel D, 'other' includes 2 phyla all with less than 5 genomes. For a complete list of archaeal phyla please see Figure 3.

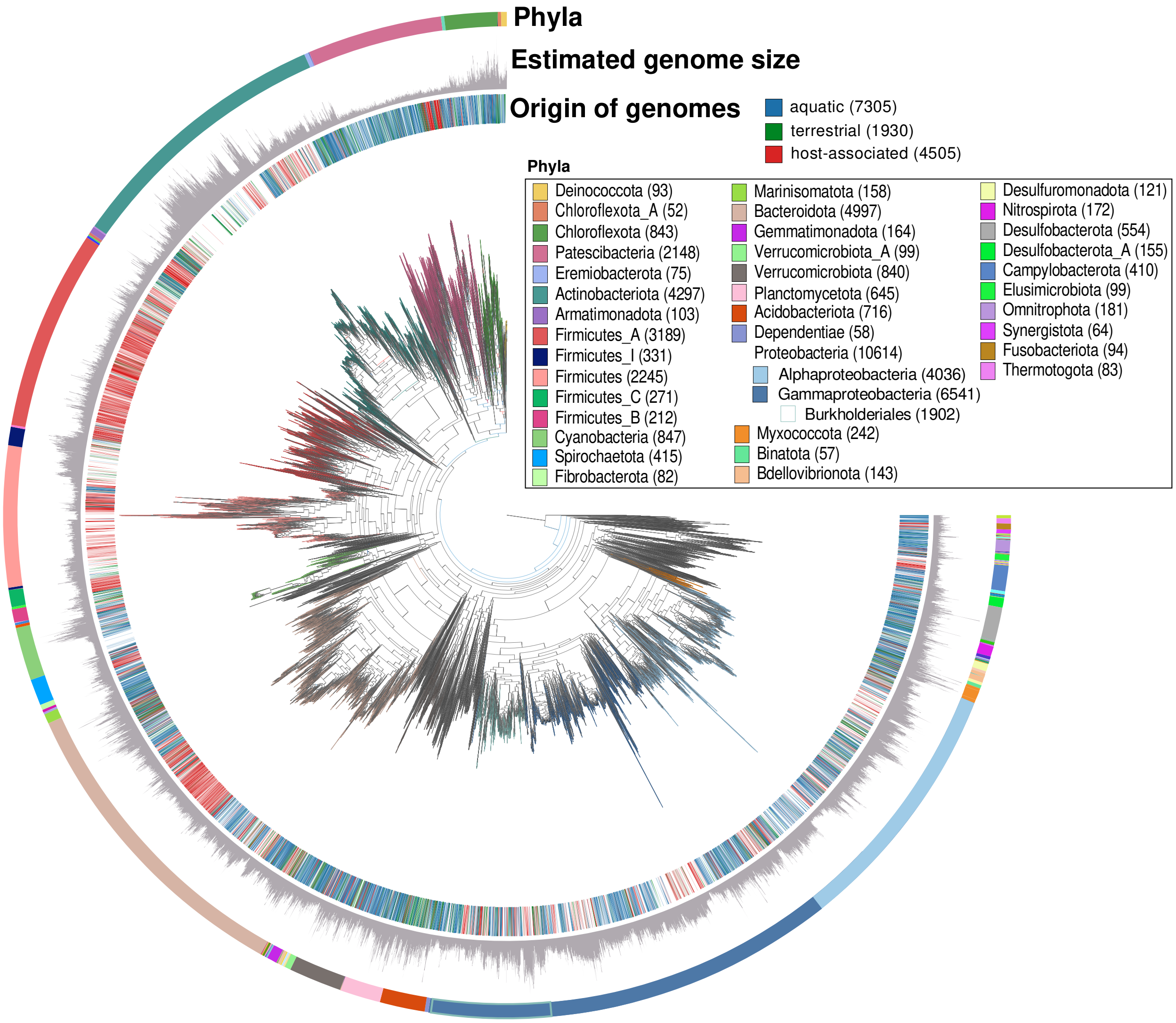

Figure 2. Phylogenetic tree of bacterial representative genomes shows variation in genome size between and within phyla. Tree was constructed using GTDB-tk and aligned concatenated set of 120 single copy marker proteins for Bacteria (96). Estimated genome size shows distribution of larger and smaller genomes sizes are non-monophyletic. The tree shows origin of the genomes: aquatic, terrestrial and host-associated genomes are MAGs from GEMs database. The backbone genomes were added by GTDB-tk and it consists of their representative genomes. Estimated genome size scale is from 0 Mb to 14 Mb. Phyla are color-coded and legend includes the phyla with most representatives. Phyla with less than 50 genomes are not included in the legend. For full legend please refer to Figure S2. Burkholderiales is the Order with most genomes.

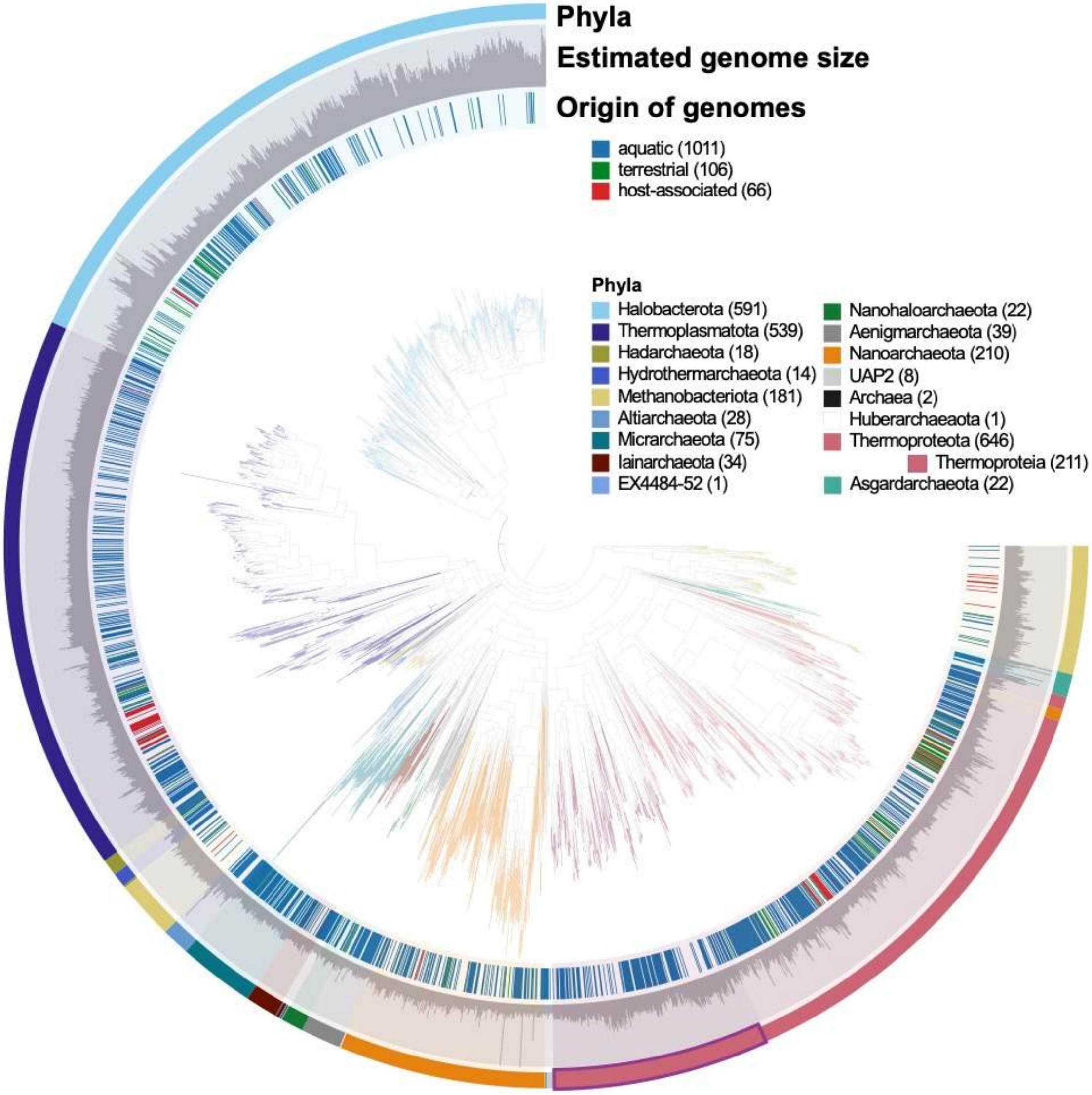

Figure 3. Phylogenetic tree of archaeal representative genomes shows variation in genome size between and within phyla. Tree was constructed using GTDB-tk and aligned concatenated set of 122 single copy marker proteins for Archaea (96). Estimated genome size shows distribution of larger and smaller genomes sizes are non-monophyletic. The tree shows origin of the genomes: aquatic, terrestrial and host-associated genomes are MAGs from GEMs database. The backbone genomes were added by GTDB-tk and it consists of their representative genomes. Estimated genome size scale is from 0 Mb to 6 Mb. Phyla are color-coded.

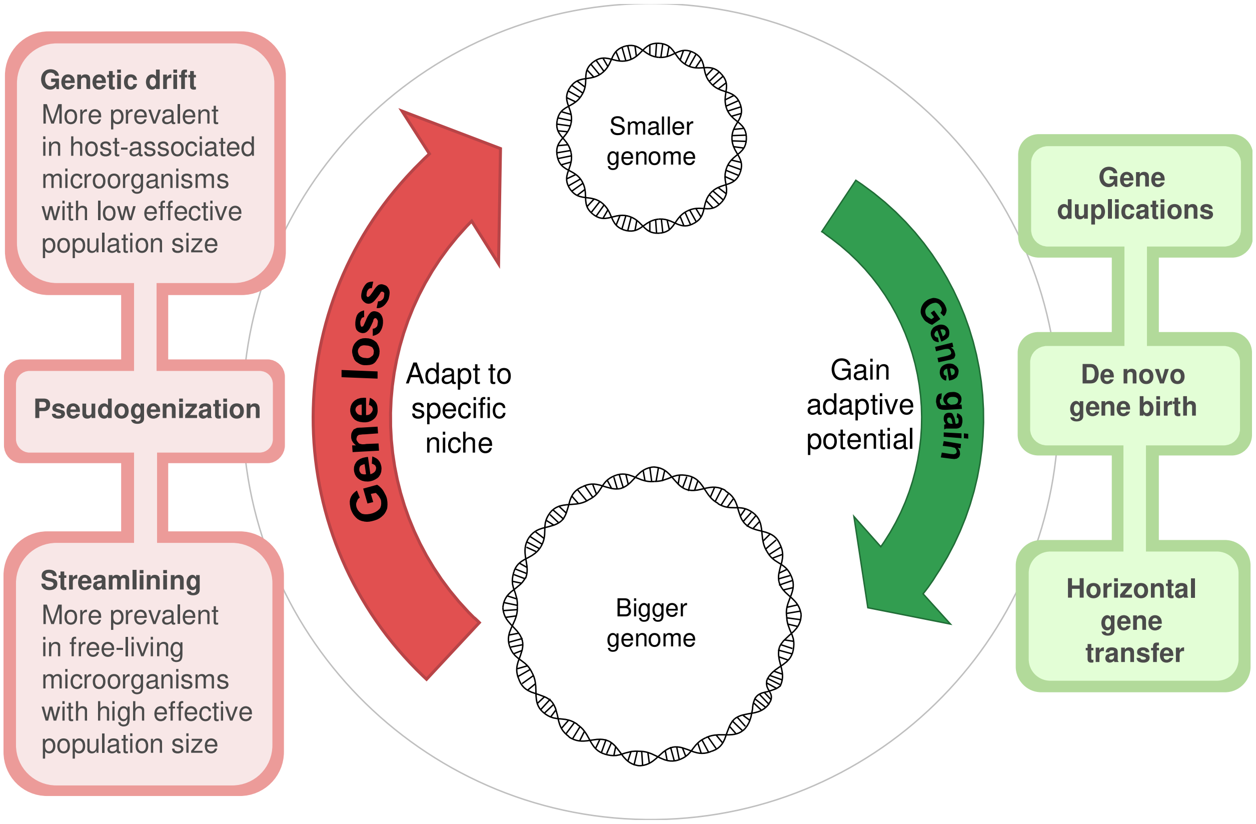

Figure 4. Conceptual figure of the evolutionary forces driving the expansion and reduction of genome sizes. Gene loss is represented with a bigger arrow because it dominates the evolutionary history we know based on extant microorganisms.

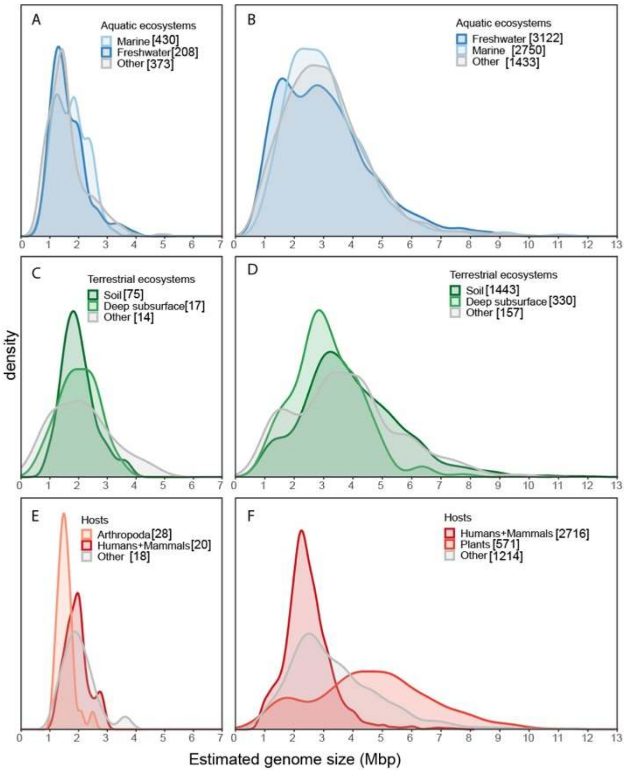

Figure 5. Genome size distribution in different sub-categories of environments. [A] Aquatic archaeal genomes, [B] aquatic bacterial genomes, [C] terrestrial archaeal genomes, [D] terrestrial bacterial genomes, [E] hos-associated archaeal genomes and [F] host-associated bacterial genomes. Inside the parenthesis is stated the number of MAGs per sub-environment.

References

1. Gregory TR. Genome Size Evolution in Animals. The Evolution of the Genome2005. p. 3-87.

2. Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. Genome Size Diversity and Its Impact on the Evolution of Land Plants. Genes (Basel). 2018;9(2).

3. Kirchberger PC, Schmidt ML, Ochman H. The Ingenuity of Bacterial Genomes. Annu Rev Microbiol. 2020;74:815-34.

4. Moran NA, Bennett GM. The tiniest tiny genomes. Annu Rev Microbiol. 2014;68:195-215.

5. Garcia R, Gemperlein K, Muller R. Minicystis rosea gen. nov., sp. nov., a polyunsaturated fatty acid-rich and steroid-producing soil myxobacterium. Int J Syst Evol Microbiol. 2014;64(Pt 11):3733-42.

6. Bobay LM, Ochman H. The Evolution of Bacterial Genome Architecture. Front Genet. 2017;8:72.

7. Puigbo P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC biology. 2014;12:66.

8. Kuo CH, Moran NA, Ochman H. The consequences of genetic drift for bacterial genome complexity. Genome Res. 2009;19(8):1450-4.

9. Gillings MR. Lateral gene transfer, bacterial genome evolution, and the Anthropocene. Annals of the New York Academy of Sciences. 2017;1389(1):20-36.

10. Han K, Li ZF, Peng R, Zhu LP, Zhou T, Wang LG, et al. Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. Scientific reports. 2013;3:2101.

11. Cavalier-Smith T. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot. 2005;95(1):147-75.

12. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014;8(8):1553-65.

13. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15(2):141-61.

14. Staley JT, Konopka A. Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. Annual Review of Microbiology. 1985;39(1):321-46.

15. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269(5223):496-512.

16. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of Mycoplasma genitalium. Science. 1995;270(5235):397-403.

17. Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, La T, et al. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct Integr Genomics. 2006;6(3):165-85.

18. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol. 2020.

19. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2(11):1533-42.

20. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun. 2016;7:13219.

21. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38(9):1079-86.

22. Batut B, Knibbe C, Marais G, Daubin V. Reductive genome evolution at both ends of the bacterial population size spectrum. Nature reviews Microbiology. 2014;12(12):841-50.

23. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A. 2004;101(9):3160-5.

24. Lynch M. Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol. 2006;60:327-49.

25. Wolf YI, Koonin EV. Genome reduction as the dominant mode of evolution. Bioessays. 2013;35(9):829-37.

26. Abram K, Udaondo Z, Bleker C, Wanchai V, Wassenaar TM, Robeson MS, et al. What can we learn from over 100,000 Escherichia coli genomes? bioRxiv. 2020.

27. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, et al. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. 2003;424(6952):1042-7.

28. Gweon HS, Bailey MJ, Read DS. Assessment of the bimodality in the distribution of bacterial genome sizes. ISME J. 2017;11(3):821-4.

29. Nelson WC, Tully BJ, Mobberley JM. Biases in genome reconstruction from metagenomic data. PeerJ. 2020;8:e10119.

30. Shade A, Hogan CS, Klimowicz AK, Linske M, McManus PS, Handelsman J. Culturing captures members of the soil rare biosphere. Environmental Microbiology. 2012;14(9):2247-52.

31. Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonzalez JM, et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc Natl Acad Sci U S A. 2013;110(28):11463-8.

32. Garcia SL. Mixed cultures as model communities: hunting for ubiquitous microorganisms, their partners, and interactions. Aquatic Microbial Ecology. 2016;77(2):79-85.

33. Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, et al. Isolation of an archaeon at the prokaryote-eukaryote interface. Nature. 2020;577(7791):519-25.

34. Cross KL, Campbell JH, Balachandran M, Campbell AG, Cooper SJ, Griffen A, et al. Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. Nat Biotechnol. 2019;37(11):1314-21.

35. Hoehler TM, Jorgensen BB. Microbial life under extreme energy limitation. Nature reviews Microbiology. 2013;11(2):83-94.

36. Figueroa-Gonzalez PA, Bornemann TLV, Adam PS, Plewka J, Révész F, von Hagen CA, et al. Saccharibacteria as Organic Carbon Sinks in Hydrocarbon-Fueled Communities. Frontiers in microbiology. 2020;11.

37. Lewis WH, Tahon G, Geesink P, Sousa DZ, Ettema TJG. Innovations to culturing the uncultured microbial majority. Nature reviews Microbiology. 2020.

38. Marais GA, Calteau A, Tenaillon O. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. Genetica. 2008;134(2):205-10.

39. Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD. A high-resolution view of genome-wide pneumococcal transformation. PLoS pathogens. 2012;8(6):e1002745.

40. de Vries J, Wackernagel W. Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. Proc Natl Acad Sci U S A. 2002;99(4):2094-9.

41. Shapiro BJ, Polz MF. Microbial Speciation. Cold Spring Harb Perspect Biol. 2015;7(10):a018143.

42. Zaremba-Niedzwiedzka K, Viklund J, Zhao WZ, Ast J, Sczyrba A, Woyke T, et al. Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. Genome Biology. 2013;14(11).

43. Sela I, Wolf YI, Koonin EV. Theory of prokaryotic genome evolution. Proc Natl Acad Sci U S A. 2016;113(41):11399-407.

44. Larsson J, Nylander JA, Bergman B. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC evolutionary biology. 2011;11:187.

45. Van Oss SB, Carvunis AR. De novo gene birth. PLoS Genet. 2019;15(5):e1008160.

46. Pal C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet. 2005;37(12):1372-5.

47. Treangen TJ, Rocha EP. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 2011;7(1):e1001284.

48. Sheridan PO, Raguideau S, Quince C, Holden J, Zhang L, Thames C, et al. Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. Nat Commun. 2020;11(1):5494.

49. Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, et al. The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol. 2002;4(4):453-61.

50. Lurie-Weinberger MN, Peeri M, Tuller T, Gophna U. Extensive Inter-Domain Lateral Gene Transfer in the Evolution of the Human Commensal Methanosphaera stadtmanae. Front Genet. 2012;3:182.

51. Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, et al. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc Natl Acad Sci U S A. 2012;109(50):20537-42.

52. Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, et al. 50 million years of genomic stasis in endosymbiotic bacteria. Science. 2002;296(5577):2376-9.

53. Andersson JO, Andersson SG. Pseudogenes, junk DNA, and the dynamics of Rickettsia genomes. Molecular biology and evolution. 2001;18(5):829-39.

54. van Passel MW, Smillie CS, Ochman H. Gene decay in archaea. Archaea. 2007;2(2):137-43.

55. Chu X, Li S, Wang S, Luo D, Luo H. Gene loss through pseudogenization contributes to the ecological diversification of a generalist Roseobacter lineage. ISME J. 2020.

56. Morris JJ, Lenski RE, Zinser ER. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. Mbio. 2012;3(2).

57. Mondav R, Bertilsson S, Buck M, Langenheder S, Lindstrom ES, Garcia SL. Streamlined and Abundant Bacterioplankton Thrive in Functional Cohorts. 2020.

58. Gabrielaite M, Johansen HK, Molin S, Nielsen FC, Marvig RL. Gene Loss and Acquisition in Lineages of Pseudomonas aeruginosa Evolving in Cystic Fibrosis Patient Airways. Mbio. 2020;11(5).

59. Csuros M, Miklos I. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Molecular biology and evolution. 2009;26(9):2087-95.

60. Wolf YI, Makarova KS, Yutin N, Koonin EV. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biology direct. 2012;7:46.

61. Raes J, Korbel J, Lercher M, von Mering C, Bork P. Prediction of effective genome size in metagenomic samples. Genome Biology. 2007;8(1):R10.

62. Chen MY, Teng WK, Zhao L, Hu CX, Zhou YK, Han BP, et al. Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. ISME J. 2020.

63. Cobo-Simon M, Tamames J. Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. BMC genomics. 2017;18(1):499.

64. Steele JH, Brink KH, Scott BE. Comparison of marine and terrestrial ecosystems: suggestions of an evolutionary perspective influenced by environmental variation. ICES Journal of Marine Science. 2019;76(1):50-9.

65. Bentkowski P, Van Oosterhout C, Mock T. A Model of Genome Size Evolution for Prokaryotes in Stable and Fluctuating Environments. Genome biology and evolution. 2015;7(8):2344-51.

66. Brewer TE, Handley KM, Carini P, Gilbert JA, Fierer N. Genome reduction in an abundant and ubiquitous soil bacterium ‘Candidatus Udaeobacter copiosus’. Nature Microbiology. 2016;2:16198.

67. Giovannoni SJ, Hayakawa DH, Tripp HJ, Stingl U, Givan SA, Cho JC, et al. The small genome of an abundant coastal ocean methylotroph. Environ Microbiol. 2008;10(7):1771-82.

68. Aylward FO, Santoro AE. Heterotrophic Thaumarchaea with Small Genomes Are Widespread in the Dark Ocean. mSystems. 2020;5(3).

69. Tian R, Ning D, He Z, Zhang P, Spencer SJ, Gao S, et al. Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. Microbiome. 2020;8(1):51.

70. Mende DR, Bryant JA, Aylward FO, Eppley JM, Nielsen T, Karl DM, et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol. 2017;2(10):1367-73.

71. Grzymski JJ, Dussaq AM. The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME J. 2012;6(1):71-80.

72. Brochier-Armanet C, Deschamps P, Lopez-Garcia P, Zivanovic Y, Rodriguez-Valera F, Moreira D. Complete-fosmid and fosmid-end sequences reveal frequent horizontal gene transfers in marine uncultured planktonic archaea. ISME J. 2011;5(8):1291-302.

73. Blesa A, Averhoff B, Berenguer J. Horizontal Gene Transfer in Thermus spp. Curr Issues Mol Biol. 2018;29:23-36.

74. Borges KM, Bergquist PL. Genomic restriction map of the extremely thermophilic bacterium Thermus thermophilus HB8. J Bacteriol. 1993;175(1):103-10.

75. Dieser M, Smith HJ, Ramaraj T, Foreman CM. Janthinobacterium CG23_2: Comparative Genome Analysis Reveals Enhanced Environmental Sensing and Transcriptional Regulation for Adaptation to Life in an Antarctic Supraglacial Stream. Microorganisms. 2019;7(10).

76. Berube PM, Biller SJ, Hackl T, Hogle SL, Satinsky BM, Becker JW, et al. Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. Sci Data. 2018;5:180154.

77. Toft C, Andersson SG. Evolutionary microbial genomics: insights into bacterial host adaptation. Nature reviews Genetics. 2010;11(7):465-75.

78. Collingro A, Kostlbacher S, Horn M. Chlamydiae in the Environment. Trends Microbiol. 2020;28(11):877-88.

79. Dharamshi JE, Tamarit D, Eme L, Stairs CW, Martijn J, Homa F, et al. Marine Sediments Illuminate Chlamydiae Diversity and Evolution. Current biology : CB. 2020;30(6):1032-48 e7.

80. McLean JS, Bor B, Kerns KA, Liu Q, To TT, Solden L, et al. Acquisition and Adaptation of Ultra-small Parasitic Reduced Genome Bacteria to Mammalian Hosts. Cell reports. 2020;32(3):107939.

81. Levy A, Salas Gonzalez I, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S, Miao J, et al. Genomic features of bacterial adaptation to plants. Nat Genet. 2017;50(1):138-50.

82. Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SG. Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc Natl Acad Sci U S A. 2004;101(26):9722-7.

83. Serra V, Gammuto L, Nitla V, Castelli M, Lanzoni O, Sassera D, et al. Morphology, ultrastructure, genomics, and phylogeny of Euplotes vanleeuwenhoeki sp. nov. and its ultra-reduced endosymbiont "Candidatus Pinguicoccus supinus" sp. nov. Scientific reports. 2020;10(1):20311.

84. Rhodes ME, Spear JR, Oren A, House CH. Differences in lateral gene transfer in hypersaline versus thermal environments. BMC evolutionary biology. 2011;11:199.

85. Cabello-Yeves PJ, Zemskaya TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov VV, et al. Genomes of Novel Microbial Lineages Assembled from the Sub-Ice Waters of Lake Baikal. Appl Environ Microbiol. 2018;84(1).

86. Allen LZ, Allen EE, Badger JH, McCrow JP, Paulsen IT, Elbourne LD, et al. Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. ISME J. 2012;6(7):1403-14.

87. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, et al. Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci U S A. 2006;103(42):15611-6.

88. Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, et al. Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci U S A. 2003;100(17):10020-5.

89. Smith MW, Zeigler Allen L, Allen AE, Herfort L, Simon HM. Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. Frontiers in microbiology. 2013;4:120.

90. de Souza RSC, Armanhi JSL, Damasceno NB, Imperial J, Arruda P. Genome Sequences of a Plant Beneficial Synthetic Bacterial Community Reveal Genetic Features for Successful Plant Colonization. Frontiers in microbiology. 2019;10:1779.

91. Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JC, Andersson DI. Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci U S A. 2005;102(34):12112-6.

92. Moreira D, Le Guyader H, Philippe H. The origin of red algae and the evolution of chloroplasts. Nature. 2000;405(6782):69-72.

93. Castelle CJ, Banfield JF. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell. 2018;172(6):1181-97.

94. Yooseph S, Nealson KH, Rusch DB, McCrow JP, Dupont CL, Kim M, et al. Genomic and functional adaptation in surface ocean planktonic prokaryotes. Nature. 2010;468(7320):60-6.

95. Team RC. R: A language and environment for statistical computing. 2020 [R Foundation for Statistical Computing , Vienna, Austria.:[Available from: http://www.R-project.org/.

96. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2020;36(6):1925-7.