Aerobic bacteria and archaea tend to have larger and more versatile genomes

Nielsen DA[1], Fierer N[2], Geoghegan JL[1], Gillings MR[1], Gumerov V[3], Madin JS[4], Moore L[5], Paulsen IT[5], Reddy TBK[6], Tetu SG[5], Westoby M[1]

[1]Dept of Biological Sciences, Macquarie University Sydney NSW 2109 Australia
[2]Department of Ecology and Evolutionary Biology, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309, USA
[3]Dept of Microbiology, Ohio State University, Columbus Ohio 43210 USA
[4]Hawaii Institute of Marine Biology, University of Hawaii, Kaneohe HI 96744 USA
[5]Dept of Molecular Sciences, Macquarie University Sydney NSW 2109 Australia
[6]DOE Joint Genome Institute, 2800 Mitchell Drive Walnut Creek, CA 94598 USA

Corresponding author: Westoby M, Dept of Biological Sciences, Macquarie University Sydney NSW 2109 Australia. E-mail: mark.westoby@mq.edu.au

Decision date: 03-Jan-2021

Abstract

A recent compilation of traits across culturable species of bacteria and archaea allows relationships to be quantified between genome size and other traits and habitat. Cell morphology, size, motility, sporulation and doubling time were not strongly correlated with genome size. Aerobic species averaged ca 35% larger genomes than anaerobic, adjusted for growth temperature. Aerobes had a similar mix of gene functions compared to anaerobes of the same genome size. Shifting proportions of aerobes to anaerobes accounted for about half of previously-known differences in mean genome size between habitats. One possible factor in these results could be if effective population sizes are larger for aerobes, reducing the potential for gene loss via genetic drift. Larger genomes also confer versatility. They can transport and metabolise a wider range of substrates. More of their genome is engaged in signal detection and response, indicating they benefit from different resources at different times or under different condition. Aerobic habitats might well present opportunities and challenges that vary through time more than anaerobic habitats. The genome size trait-dimension contributes a useful quantitative descriptor for ecological strategies.

Keywords: genome size, bacteria, archaea, ecological strategies, versatility, aerobic, anaerobic

Introduction

This work emerges from a broader project to assess ecological strategies via traits. For other taxa and notably for land plants, "trait ecology" has developed as a way to position species relative to each other in a strategy space. Rather than relying on concepts such as stress-tolerance, species can be positioned along measurable trait-dimensions, including species that do not occur together (e.g. Wright et al. 2004, Díaz et al. 2016, Kunstler et al. 2016). Trait-based ecology for bacteria and archaea has been advocated and initiated by several research groups (Fierer et al. 2007, 2014, Litchman and Klausmeier 2008, Krause et al. 2014, Litchman et al. 2015, Ho et al. 2017, Wood et al. 2018, Bewick et al. 2019, Malik et al. 2020). The topic considered here is variation of genome size across bacterial and archaeal species (using consistent species definitions from Genome Taxonomy Database, Parks et al. 2018), its correlation with other traits and with habitat, what light that may shed on the forces most responsible for observed genome size variation, and how genome size variation should be interpreted as a dimension of ecological strategies. We use a recent compilation of traits that substantially expands coverage across culturable species (Madin et al. 2020). For some relationships previously reported, the effect is to strengthen quantification. We focus especially on the relationship to oxygen use, which has not previously been quantified, and its cross-correlation with other traits and with habitat.

Genome size in prokaryotes varies across two orders of magnitude (112 kb to 13 Mb, Koonin and Wolf 2008, Moran and Bennett 2014). In comparison, eukaryote genomes vary across six orders of magnitude, largely as a result of duplications and accumulation of non-coding DNA (Cavalier-Smith 2005). Prokaryote genomes consist mainly (mean 88%) of protein-coding genes, with much of the remainder devoted to structural RNAs or to regulation, and little non-coding or parasitic DNA (Kirchberger et al. 2020). Most genes are present as single copies, such that the count of different genes is closely correlated with genome size (r^2 =0.976 across 3300 species in the Madin et al compilation (Westoby et al. in review)). Prokaryote genomes are thought to be reduced in this way because of deletional bias (Mira et al. 2001), deletion mutations being more common than insertions or duplications and leading to loss of genes or function.

Some interpretations of prokaryote genome size emphasize rates of gain and loss of genes, and some emphasize adaptive benefit. Rates of gain by horizontal gene transfer (Oliveira et al. 2017) might be higher in communities where more different useful genes are available, or where reactive oxygen is present which induces the SOS response which in turn encourages horizontal transfer (Baharoglu and Mazel 2014). Another rate-based interpretation (Bobay and Ochman 2018, Kirchberger et al. 2020) invokes the idea that in species with smaller effective population sizes drift is more important relative to selection. Selection relative to drift is indexed by dN/dS the ratio of non-synonymous to synonymous mutations, and species with lower dN/dS do tend to have larger genomes (e.g. Sela et al. 2016, Bobay and Ochman 2018). Bobay and Ochman describe the mechanism as follows: "effective population size modulates the efficacy of selection, thereby affecting the number of genes that are effectively perceived as neutral (and eliminated) and the number of genes that are retained by selection".

The most-studied instances of genome reduction fall into two groups (Giovannoni et al. 2014). Host-associated species including pathogens grow in rather consistent habitats and may be able to simplify genomes for that reason. In addition hosts may often be colonized by very few individuals leading to large drift effects and loss of genes for that reason (e.g. Ochman and Moran 2001, Toft and Andersson 2010, Murray et al. 2020). The most extreme shedding of function is found in species that make their living inside eukaryote cells (McCutcheon and Moran 2012), a specialised location where the host cell can provide many resources. For the present paper we have set aside obligate intracellular species and mycoplasmas in order to see more clearly the patterns among the remaining ways of life. The second much-studied category of genome reduction is in oligotrophic oceans, where very large populations are said to be under strong selection to minimize cell complexity and resources used for replication (Giovannoni et al. 2014). However, it has also been argued for these large populations that gene loss is associated with selective sweeps within subpopulations that have small effective population size (Kirchberger et al. 2020).

Life at elevated temperatures is associated with smaller genome sizes, suggested to arise by genome reduction during adaptation to high temperatures, and also to a small extent by genes being shorter on average (Sabath et al. 2013). A natural experiment along a thermal gradient confirmed that hotter soils had smaller microbial genomes (Sorensen et al. 2019), and similarly experimental warming of coastal plankton communities reduced mean genome size (Huete-Stauffer et al. 2016), and bacteria at warmer latitudes have smaller genomes (Lear et al. 2017).

Because prokaryote genome size corresponds closely with the total number of different genes, larger genome size can be interpreted as conferring increased versatility (Konstantinidis and Tiedje 2004, Guieysse and Wuertz 2012, Cobo-Simón and Tamames 2017), meaning that it is expected to reflect the range of different resources that can be transported or metabolized, together with flexibility in response to different circumstances. Species found in soils tend to larger genomes on average, and it is plausible that life in soil presents challenges and opportunities that vary through time more than life in water (Konstantinidis and Tiedje 2004, Parter et al. 2007, Cobo-Simón and Tamames 2017). Species with larger genomes also tend to be distributed across a wider range of habitats, as might perhaps be expected from their wider metabolic capacities (Cobo-Simón and Tamames 2017, Sriswasdi et al. 2017).

Our analysis focuses around the relationship of genome size to oxygen use, which has not previously been quantified. We consider also this relationship's cross-correlation with other traits and with habitat.

Methods

Data

Data for most traits were drawn from a compilation by Madin et al (2020), the traits in question being genome size, growth temperature, minimum doubling time, cell radial diameter, cell shape, oxygen use categories, number of rRNA operon copies, sporulation, motility, gram stain and source habitat. That data

paper merges 26 sources, ranging from large managed databases to tables in published papers. The merger is accomplished via an open-access scripted workflow, giving rise to tables with biological entities as rows or records, and traits as columns. After merging sources the data include multiple records within many species. Outliers were not removed automatically by Madin et al, but their underlying sources were inspected by knowledgeable microbiologists, leading to a substantial number of corrections, which are documented in the workflow. Madin et al then also provide tables where these multiple records are condensed down to a single row per species, using two alternative taxonomies. For purposes of the present report, each instance is one species in terms of the Genome Taxonomy Database GTDB (Parks et al. 2018) (https://gtdb.ecogenomic.org/ July 2019 version). When multiple records for a species were condensed down to a single row, the records were averaged (for quantitative traits) or a majority rule was applied (for categorical traits).

GTDB taxonomy was adopted because it is monophyletic, so far as can be determined from the 120 protein-coding genes used, and because it places taxonomic ranks at a consistent relative distance from the tree root. The species level condensation has been adopted as a working compromise, intended to capture ecologically meaningful variation without letting the dataset be unduly dominated by a few species with thousands of records each (e.g. *Staphylococcus aureus, Salmonella enterica, Streptococcus pneumoniae*). Madin et al (2020) focused on phenotypic traits such as cell diameter and potential rate of increase and consequently the data come largely from species that have been brought into culture. These may tend to larger genomes and faster potential growth rates and more often to be aerobic, compared to the many uncultured species (Giovannoni et al. 2014, Nayfach and Pollard 2015, Solden et al. 2016, Fierer 2017). However, the species included here do span a full range of possibilities, including strong oligotrophy, very small genome sizes and very slow potential growth rates.

For purposes of the present paper we have excluded a priori species that live inside the cells of eukaryotes, and also mycoplasmas as a group.

Traits

For genome size, records were taken only when described as “complete”, “whole genome sequence” (WGS) or “complete draft”. From the PATRIC dataset (Wattam et al. 2017), data were removed where the sequencing status was "incomplete" or "partial" or had a sequencing depth of <10 or had the words "single cell" in the isolation source.

Considering species with 10 or more records for genome size, median coefficient of variation within species was ~3% (Fig S10), and 95% of species had CV <=15%.

Attribution of species to oxygen-use types used the following rules. (1) There were records for facultative anaerobe and also for facultative aerobe. We interpret both as facultative anaerobes. In figures these are called simply facultative, for brevity. (2) Where multiple records were available within a species and they included different attributions, a majority rule was applied. (3) Where there were ties in the counts of different attributions (for example 1:1), obligate aerobes and anaerobes were given priority over unqualified aerobes and anaerobes. Microaerophiles and facultative were also given priority over aerobes, on grounds that more detailed investigation probably lay behind these attributions. (4) Where there was a tie in the count between categories that were prioritised at the same level, such as microaerophile and facultative, oxygen use was coded not available NA.

“Growth temperature” is in practice nearly always the temperature routinely adopted for culture. Engqvist (2018), the largest source, extracted data from culture collection websites. Growth temperature can be interpreted as correlated with temperatures experienced during growth in natural habitat, and with the temperature that would maximize growth rate, but is not identical to either.

Sporulation is recorded simply as present or absent. It refers always to formation of endospores. Motility is recorded mostly as present or absent, but a few times as via flagella or gliding or axial filament. For purposes of the present paper this has been condensed to present or absent. Cell shapes have been condensed to rods (bacilli) vs near-spheres (cocci and coccobacilli). Discs, filaments, flasks, stars, spirals, spindles, vibrios, fusiform, "pleomorphic", "irregular", "square" and "triangular" have been grouped as “other” and are not included in comparisons of shape reported here.

Cell diameter is measured in the radial direction. It is known to increase plastically under richer nutrient conditions, volume increasing 2-fold or more (Vadia and Levin 2015, Lever et al. 2015, Cesar and Huang 2017, Harris and Theriot 2018). The measurements brought together in Madin et al (2020) would in the great majority of cases have been measured in culture under favourable growth conditions, and can be

considered standardised for plasticity to this extent. Some sources give a single value, others give a range. For ranges, the present paper takes the midpoint.

Among phenotypic measurements of prime interest, minimum doubling times or maximum growth rates have least coverage of species. This describes the exponential growth rate as measured under favourable conditions. Some observations are taken from culture conditions, and it is uncertain how widely alternative growth conditions may have been investigated. Other observations are from studies where factors have been varied systematically, and in these cases the shortest doubling time has been taken. Undoubtedly these numbers should be treated as approximate or noisy. However, they do span from less than an hour to more than 100 hours, and they show (Table S1) the same correlation across species with rRNA16S copy number that has been widely reported (Vieira-Silva and Rocha 2010, Roller et al. 2016). We believe they do capture meaningful differences across species, though small differences or small numbers of species need to be treated with extra caution.

Habitat

Madin et al (2020) defined a system of 91 fixed environment descriptors, using a hierarchical system employing up to four individual terms for each descriptor. Here we have used the data only to aggregate species into major habitat groupings such as soil, freshwater, and host associated, similar to those used in previous studies.

Inferring clusters of orthologous groups from genomes.

In order to explore shifts in genome composition we developed a separate dataset from annotated genomes in RefSeq (Tatusova et al. 2016). Protein coding sequences were extracted from bacteria (n=1582) and archaea (n=146). Protein sequences were then analysed with RPS-BLAST+ against NCBI's Conserved Domain Database (CDD, downloaded July 2018) (Marchler-Bauer et al. 2011). Clusters of orthologous groups (COGs) (Galperin et al. 2014) were inferred using CDD2COG.pl v0.1 (Leimbach 2016). Hits for each query protein were filtered for the lowest e-value, in other words each gene was attributed to a single best-fit COG. COGs and Pfams are generally thought to be consistent since the underlying assignment method ensures even remote sequence similarity is detected across conserved residues (Chen et al. 2013). This

dataset was cross-referred to our main dataset for purpose of attributing aerobic vs anaerobic, using NCBI Taxon IDs which are associated to particular genome records in RefSeq and nested into species in GTDB.

Genes involved with signalling

We developed a further dataset from the MISTdb database (https://mistdb.com/) (Ulrich and Zhulin 2010, Gumerov et al. 2020), which provides more detailed information on genes involved with signalling than does COG T "signal transduction mechanisms". Data were downloaded in November 2018. Here we use the counts for total genes involved with signalling and for the two major categories of one-component and two-component systems. These data were cross-referred to our main dataset for purpose of attributing aerobic vs anaerobic, using NCBI Taxon IDs which are associated to particular genome records in MIST and nested into species in GTDB. Counts were averaged within GTDB species, where there was more than one record.

Data analysis

Statistics presented describe the variation in genome size across species and its main correlates. Phylogenetic generalised least squares is mathematically equivalent to taking phylogenetic divergences, rather than present-day species, as the cases under study (Blomberg et al. 2012). That is the way we interpret it here. It can also be thought of as a special sort of partial correlation, corresponding to a causal model where phylogeny is fitted first as a predictor and only residuals are able to be attributed to other present-day traits or to habitat (Westoby 1999, Uyeda et al. 2018). In reality past history and present-day selection are not mutually exclusive as causes.

Genome size, cell diameter and minimum doubling time were $\log_{10}$-scaled for analysis to reduce skew. Phylogenetic generalised least squares used phylolm (Tung Ho and Ané 2014); v2.6.1 was installed from https://github.com/lamho86/phylolm. The phylogenetic tree adopted corresponded to GTDB taxonomy with seven levels (superkingdom, phylum, class, order, family, genus, species), star phylogeny at each node and unit branch lengths. GTDB has standardized the depth of these seven taxonomic levels relative to the tree root.

Results

There was little difference in genome size between species recorded as aerobic and obligate aerobic, nor between anaerobic and obligate anaerobic (Fig 1a). Accordingly for relationships reported in this main text, oxygen use is condensed to a binary contrast between aerobic including obligate, and anaerobic including obligate. Not included are facultative species, which spanned a range of genome size covering both aerobes and anaerobes, and microaerophiles (further comment in Supp Mat) which tended to small genomes (Fig 1a).

Correlates of genome size

The two strongest correlates of genome size were aerobic vs anaerobic oxygen tolerances (Pearson r = 0.44), and growth temperature (r = -0.43) (Fig 1b, partitioning of correlated variation in Fig S1, models in Tables S2 and S3). Quantitative relationships are best appreciated through percentage responses of genome size to different predictors (columns 4 and 9 of Table 1). Again the largest responses are to aerobic vs anaerobic (54%) and to growth temperature, where the slope of -0.0186 per ^{o}C translates to more than 50% decline in genome size across 30 ^{o}C.

The different correlates of genome size were themselves cross-correlated to some extent. The difference in genome size between aerobes and anaerobes at a given growth temperature and within bacteria or archaea averaged 35% (Table S3, Fig 1b), not as much as the 54% when taken as a raw difference across all species (Table 1, column 4). Response coefficients after partialling for superkingdom, oxygen use and growth temperature were smaller than the raw response coefficients, but mostly still different from zero (column 6 compared to column 3 in Table 1, associated confidence intervals in column 7).

Relations to phylogeny

Aerobic vs anaerobic metabolism and its association with genome size is fairly conservative down the phylogeny. Of 447 families where metabolism was known for two or more representatives, 242 were entirely aerobic or anaerobic, as also were some orders, classes and phyla. A further 37 large families (10 or more representatives in the data) were at least 90% one or the other. Nevertheless, the tendency of aerobes to have larger genomes recurred consistently within the great majority of clades (Fig 2a for major phyla, and Fig S4 for subclades within particular phyla). Genome size differences between aerobes and anaerobes at each phylogenetic divergence (across each node in the phylogenetic tree) averaged 9% (Table 1, phylogenetic generalised least squares slope and its equivalent percentage effect). They were rather consistent in direction, with 95% confidence intervals on the slope translating to 5.6% and 11.3%. Divergences at different depths in the phylogenetic tree accumulated to produce the 53% mean difference between all present-day aerobes and anaerobes. The 9% average difference at individual divergences translates into about 270 genes, for a 3Mb genome size and assuming about 1000 genes per Mb (e.g. Xu et al. 2006). Given differences this substantial, it is not surprising to see phylogenetic niche conservatism

whereby aerobic habitats have often been colonised from clades that are already aerobic. A trait-habitat combination can be both phylogenetically conservative and also adaptive in the present day, these are not mutually exclusive interpretations.

Relations to habitat

Genome size is known to differ between different major habitats (e.g. Konstantinidis and Tiedje 2004, Parter et al. 2007, Cobo-Simón and Tamames 2017), tending to be highest in species from soils and lowest from thermal, oral and gut habitats, with marine and freshwater environments intermediate. However, differences between aerobes and anaerobes were clear within each of these habitat-types (Fig 2b). About half of the variation across habitat means (red dots in Fig 2b, which show a pattern very similar to that reported by Cobo-Simon and Tamames) can be understood as resulting from shifting mixtures of aerobes with anaerobes across habitat (sum of squares SS across habitats after aerobe vs anaerobe 12.47, compared to SS across habitats 25.49, Table S4).

Genome composition

The extra genes in aerobes did not in general appear to be contributed disproportionately from particular categories or activities. Rather, they followed established scaling rules that relate particular gene categories to genome size (van Nimwegen 2003, 2006, Konstantinidis and Tiedje 2004, Koonin and Wolf 2008, Molina and van Nimwegen 2009). Genes involved with signalling (Fig 3a) have been reported to scale with log-log slope ~2, meaning that they contribute an increasingly large fraction of the total as genome size rises. However, an initial slope of 2 must necessarily shallow to a slope of 1 at some stage, otherwise the number of signalling genes would exceed the total number of genes in the genome. A slope in the order of 2 was indeed observed among genomes up to about 3-4 Mb (Fig 3a), spanning the whole range of archaeal but only part of the range of bacterial genome sizes. Above 3-4 Mb the relationship in bacteria approximated a slope of 1, as signalling genes reached some maximum percentage of the genome. The reference line of slope 1 in Fig 3a is at a height that corresponds to ca 10% of genes involved with signalling, assuming an average 1000 genes per Mb. For purposes of the aerobe-anaerobe comparison, the important point is that while aerobes did indeed tend to have more genes involved with signalling as well as larger genomes, nevertheless at a given genome size aerobes had broadly similar numbers of signalling genes compared to anaerobes (Fig 3a). Similar results apply for genes grouped into COG-categories (clusters of orthologous groups (Galperin et al. 2015)). For example, genes involved with transport and metabolism of all sorts (amino-acids, carbohydrates, inorganic ions, coenzymes and lipids) were more numerous in aerobes, but only in proportion to genome size. These COGs contributed a similar proportion of the genome in large-genome anaerobes compared to aerobes of the same genome size (Fig 3b; other COG-categories are described in Figs S4-S8). Broadly a similar pattern applied throughout: each category of genes made up a similar fraction of the genome in aerobes compared to anaerobes with equally large genomes.

Discussion

When interpreting these results it should be borne in mind that the data come from species that have been brought into culture. Not-yet-cultured species may tend to have smaller genomes and slower maximum growth rates and may be more likely to be anaerobic (see Methods). Is it possible that not-yet-cultured aerobes tend to smaller genomes than not-yet-cultured anaerobes, reversing the relationship between traits reported here for cultured species? Since oxygen tolerance data are not available for not-yet-cultured species, this cannot yet be known. But we do not know of reasons to expect relationships between traits to be structured differently among cultured compared to not-yet-cultured species.

The correlation patterns reported here were largely similar between bacteria and archaea (e.g. Figs 1 and 3 and figures in Supplementary Materials). Most likely this was because they arise from basics of cell size and growth and from the fact that genome size reflects mainly number of different genes, with these basics applying across both bacteria and archaea.

The tendency toward smaller genomes in warmer habitats (Fig 1b) has been noted previously (Lear et al. 2017, Sauer and Wang 2019). The tendency to larger genomes in aerobes has not previously been quantified, although a dataset used by Wang et al (2006) shows mean genome size 4.17 Mb for 37 aerobes vs 3.03 Mb for 39 anaerobes. Nor has oxygen use featured in discussions of genome size (e.g. Konstantinidis and Tiedje 2004, Koonin and Wolf 2008, Guieysse and Wuertz 2012, Bentkowski et al. 2015, Cobo-Simón and Tamames 2017). However, larger genomes in aerobes do reflect the known structure of metabolic networks. Oxic metabolic pathways are built on top of a core of anoxic pathways (Raymond and Segrè 2006, Sousa et al. 2016). Raymond and Segrè (2006) ran simulations where metabolic networks could self-extend from a core, provided that metabolites arose from pre-existing pathways, and that the additional pathways were known to be possible from the Kyoto Encyclopedia of Genes and Genomes. Where molecular oxygen was provided, it increased the total size of the network by about 50% compared to anoxic. This increase consisted largely of new pathways, with a smaller contribution from alternative enzymes for previously existing pathways. About half of the added pathways used molecular oxygen directly, and the other half were made possible by emergence of new reactants. The 50% increase in that simulation is interesting to compare with the differences observed here: 54% gross between all aerobes and all anaerobes (Table 1), and 35% within each of bacteria and archaea and at a given optimal growth temperature (Fig 1, Table S3).

What can be said about factors or situations favouring larger vs smaller genomes? Considering first ideas that invoke rates of gain or loss of genes, it is possible that some anaerobic habitats might tend to have smaller effective population sizes, leading to faster rates of gene loss, and possible also that some anaerobic habitats supply new genes by horizontal transfer at a slower rate. We do not have new data to

test these ideas, beyond the negative correlations between genome size and dN/dS described by Sela et al (2016) and Bobay and Ochman (2018).

Considering second ideas that invoke adaptation, one possibility is that the extra ATP (up to 19-fold more) made available by using oxygen as electron acceptor might make it possible to exploit a wider range of substrates, including those that are expensive to transport or metabolize. This possible reason for larger genomes in aerobes has not been noted elsewhere, to our knowledge.

An idea modelled by Bentkowski et al (2016) is that larger genomes are favoured by high mortality. Since mortality rates balance rates of multiplication over evolutionary time, high mortality can be thought of as fast-turnover populations. However, available estimates for fast potential rates of increase (short minimum doubling time) were only very weakly correlated with genome size ($r^2 \sim 0.01$, Table S1; Fig S2b), as previously reported (Vieira-Silva and Rocha 2010). To the extent evidence is available, it does not support the idea that population turnover is a major influence on genome size across bacteria and archaea as a whole.

A prevalent interpretation of larger genomes is that they should be favoured where resources are variable or diverse or recalcitrant (Konstantinidis and Tiedje 2004, Madsen 2008). There are a number of reasons for thinking this credible. First, it corresponds with the actual functions of the extra genes, in giving access to a wider variety of resources or under a wider variety of conditions. Second, the observed pattern of genome size across major habitats (Fig 2, and Parter et al. 2007, Cobo-Simón and Tamames 2017) can plausibly be interpreted as arising from thermal and host-associated habitats being less variable, aquatic habitats intermediate and soils most variable. Third, taxa with larger genomes also tend to be distributed across a wider range of habitats (Cobo-Simón and Tamames 2017, Sriswasdi et al. 2017). Wider habitat breadth and coping with variability through time can both be thought of as aspects of niche breadth or of generalism vs specialism. Fourth, there is a quantitative model showing larger genomes selected by variability (Bentkowski et al. 2015). In this model an environment index moves to and fro over time, more vigorous movement representing higher variability. Different resource-uptake genes confer most benefit at different points along this index. Consequently when the model is run with higher variability a larger number of different genes are favoured and genome size is greater. Fifth, species that sporulate tend to have larger genomes, by amounts in the order of 35% or 1-2 Mb (Table 1, Fig S2A). This is considerably beyond the number of genes directly involved in sporulation (Galperin et al. 2012). Sporulating life histories are likely to be associated with environmental challenges that vary through time. Sixth, as genome size increases so also does the relative contribution of genes involved with signal detection and response (e.g. Fig 3a). This indicates that changes over time in gene expression and in which resources are being exploited are important.

These interpretations might apply to the tendency of aerobes to have larger genomes that has been quantified here, if aerobic habitats tend to offer more variable resource conditions over time, for example by shifting between aerobic and anaerobic conditions, as would be common at many microsites within

soils. Alternatively or additionally, the extra ATP from using oxygen as electron acceptor might make it possible to exploit a wider range of resources.

It is important to note that different ideas about processes and factors influencing genome size need not be mutually exclusive. For example, it would be possible for a genome size difference between aerobes and anaerobes to arise both from positive selection for a wider range of metabolic capacities in aerobic situations, and from a tendency to lose genes faster with smaller effective population sizes in anaerobic habitats.

Obviously, aerobes differ from anaerobes in specific aspects of energy metabolism. But also, they have genomes larger by a thousand or more genes on average. This is consistent with an interpretation that aerobes are substantially shifted from anaerobes along an ecological strategy spectrum expressing versatility.

Data statement

Trait data analysed here are mostly drawn from a data paper by Madin et al (2020). That paper describes a merger of 26 data sources, with open-source code for the merger at GitHub (https://github.com/bacteria-archaea-traits/bacteria-archaea-traits/releases/tag/v1.0.0). The dataset used here is the product condensed to one row per species (using Genome Taxonomy Database https://gtdb.ecogenomic.org/) and is available at Figshare https://doi.org/10.6084/m9.figshare.c.4843290. Data on signal transduction proteins are drawn from the MIST database (Gumerov et al. 2020).

Acknowledgments

We thank David Warton, Will Cornwell, Rob Willows, Phil Hugenholtz, Tim Ghaly, Andrew Bissett, Elena Litchman and Jennifer Martiny for help and advice. Funds have been contributed from Macquarie U Species Spectrum Research Centre, from Macquarie U Biomolecular Discovery and Design Research Centre, and from Australian Research Council fellowships to MW and ITP.

Author contributions

Following CRT taxonomy (https://casrai.org/credit/) author contributions were as follows. Conceptualization: Westoby; data curation: Madin, Nielsen; formal analysis: Madin, Nielsen, Westoby; investigation: all authors; software: Madin, Nielsen; project administration: Westoby; resources: Fierer, Gumerov, Reddy; visualization: Madin, Nielsen, Westoby; writing draft: Westoby; writing review: all authors.

Competing interests

The authors declare no competing financial interests.

References

Baharoglu, Z. and Mazel, D. 2014. SOS, the formidable strategy of bacteria against aggressions. - FEMS Microbiol. Rev. 38: 1126–1145.

Bentkowski, P. et al. 2015. A Model of Genome Size Evolution for Prokaryotes in Stable and Fluctuating Environments. - Genome Biol. Evol. 7: 2344–2351.

Bentkowski, P. et al. 2016. The effect of extrinsic mortality on genome size evolution in prokaryotes. - ISME J. 11: 1011–1018.

Bewick, S. et al. 2019. Trait-based analysis of the human skin microbiome. - Microbiome 7: 1–15.

Blomberg, S. P. et al. 2012. Independent contrasts and PGLS regression estimators are equivalent. - Syst. Biol.: syr118.

Bobay, L.-M. and Ochman, H. 2018. Factors driving effective population size and pan-genome evolution in bacteria. - BMC Evol. Biol. 18: 1–12.

Cavalier-Smith, T. 2005. Economy, Speed and Size Matter: Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion. - Ann. Bot. 95: 147–175.

Cesar, S. and Huang, K. C. 2017. Thinking big: the tunability of bacterial cell size. - FEMS Microbiol. Rev. 41: 672–678.

Chen, I.-M. A. et al. 2013. Improving Microbial Genome Annotations in an Integrated Database Context. - PLOS ONE 8: e54859.

Cobo-Simón, M. and Tamames, J. 2017. Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. - BMC Genomics 18: 499.

Díaz, S. et al. 2016. The global spectrum of plant form and function. - Nature 529: 167–171.

Engqvist, M. K. M. 2018. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. - BMC Microbiol. 18: 177.

Fierer, N. 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. - Nat. Rev. Microbiol. 15: 579.

Fierer, N. et al. 2007. Toward an Ecological Classification of Soil Bacteria. - Ecology 88: 1354–1364.

Fierer, N. et al. 2014. Seeing the forest for the genes: using metagenomics to infer the aggregated traits of microbial communities. - Front. Microbiol. 5: 614. doi: 10.3389/fmicb.2014.00614.

Galperin, M. Y. et al. 2012. Genomic determinants of sporulation in Bacilli and Clostridia: towards the minimal set of sporulation-specific genes. - Environ. Microbiol. 14: 2870–2890.

Galperin, M. Y. et al. 2014. Expanded microbial genome coverage and improved protein family annotation in the COG database. - Nucleic Acids Res.: gku1223.

Galperin, M. Y. et al. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. - Nucleic Acids Res. 43: D261–D269.

Giovannoni, S. J. et al. 2014. Implications of streamlining theory for microbial ecology. - ISME J. 8: 1553–1565.

Guieysse, B. and Wuertz, S. 2012. Metabolically versatile large-genome prokaryotes. - Curr. Opin. Biotechnol. 23: 467–473.

Gumerov, V. M. et al. 2020. MiST 3.0: an updated microbial signal transduction database with an emphasis on chemosensory systems. - Nucleic Acids Res. 48: D459–D464.

Harris, L. K. and Theriot, J. A. 2018. Surface Area to Volume Ratio: A Natural Variable for Bacterial Morphogenesis. - Trends Microbiol. 26: 815–832.

Ho, A. et al. 2017. Revisiting life strategy concepts in environmental microbial ecology. - FEMS Microbiol. Ecol. in press.

Huete-Stauffer, T. M. et al. 2016. Experimental Warming Decreases the Average Size and Nucleic Acid Content of Marine Bacterial Communities. - Front. Microbiol. in press.

Kirchberger, P. C. et al. 2020. The Ingenuity of Bacterial Genomes. - Annu. Rev. Microbiol. 74: 815–834.

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. 101: 3160–3165.

Koonin, E. V. and Wolf, Y. I. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. - Nucleic Acids Res. 36: 6688–6719.

Krause, S. et al. 2014. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. - Front. Microbiol. in press.

Kunstler, G. et al. 2016. Plant functional traits have globally consistent effects on competition. - Nature 529: 204–207.

Lear, G. et al. 2017. Following Rapoport's Rule: the geographic range and genome size of bacterial taxa decline at warmer latitudes. - Environ. Microbiol. 19: 3152–3162.

Leimbach, A. 2016. bac-genomics-scripts: bovine E. coli mastitis comparative genomics edition. - Zenodo in press.

Lever, M. A. et al. 2015. Life under extreme energy limitation: a synthesis of laboratory- and field-based investigations. - FEMS Microbiol. Rev. 39: 688–728.

Litchman, E. and Klausmeier, C. A. 2008. Trait-Based Community Ecology of Phytoplankton. - Annu. Rev. Ecol. Evol. Syst. 39: 615–639.

Litchman, E. et al. 2015. Microbial resource utilization traits and trade-offs: implications for community structure, functioning, and biogeochemical impacts at present and in the future. - Front. Microbiol. in press.

Madin, J. S. et al. 2020. A synthesis of bacterial and archaeal phenotypic trait data. - Sci. Data in press.

Madsen, E. L. 2008. Environmental Microbiology: From Genomes to Biogeochemistry. - Wiley-Blackwell.

Malik, A. A. et al. 2020. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. - ISME J. 14: 1–9.

Marchler-Bauer, A. et al. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. - Nucleic Acids Res. 39: D225–D229.

McCutcheon, J. P. and Moran, N. A. 2012. Extreme genome reduction in symbiotic bacteria. - Nat. Rev. Microbiol. 10: 13–26.

Mira, A. et al. 2001. Deletional bias and the evolution of bacterial genomes. - Trends Genet. 17: 589–596.

Molina, N. and van Nimwegen, E. 2009. Scaling laws in functional genome content across prokaryotic clades and lifestyles. - Trends Genet. 25: 243–247.

Moran, N. A. and Bennett, G. M. 2014. The Tiniest Tiny Genomes. - Annu. Rev. Microbiol. 68: 195–215.

Murray, G. G. R. et al. 2020. Genome reduction is associated with bacterial pathogenicity across different scales of temporal and ecological divergence. - bioRxiv: 2020.07.03.186684.

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.

Ochman, H. and Moran, N. A. 2001. Genes Lost and Genes Found: Evolution of Bacterial Pathogenesis and Symbiosis. - Science 292: 1096–1099.

Oliveira, P. H. et al. 2017. The chromosomal organization of horizontal gene transfer in bacteria. - Nat. Commun. 8: 841.

Parks, D. H. et al. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. - Nat. Biotechnol. 36: 996–1004.

Parter, M. et al. 2007. Environmental variability and modularity of bacterial metabolic networks. - BMC Evol. Biol. 7: 169.

Raymond, J. and Segrè, D. 2006. The Effect of Oxygen on Biochemical Networks and the Evolution of Complex Life. - Science 311: 1764–1767.

Roller, B. R. K. et al. 2016. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. - Nat. Microbiol. 1: 16160.

Sabath, N. et al. 2013. Growth Temperature and Genome Size in Bacteria Are Negatively Correlated, Suggesting Genomic Streamlining During Thermal Adaptation. - Genome Biol. Evol. 5: 966–977.

Sauer, D. B. and Wang, D.-N. 2019. Predicting the optimal growth temperatures of prokaryotes using only genome derived features. - Bioinformatics 35: 3224–3231.

Sela, I. et al. 2016. Theory of prokaryotic genome evolution. - Proc. Natl. Acad. Sci. 113: 11399–11407.

Solden, L. et al. 2016. The bright side of microbial dark matter: lessons learned from the uncultivated majority. - Curr. Opin. Microbiol. 31: 217–226.

Sorensen, J. W. et al. 2019. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. - Nat. Microbiol. 4: 55–61.

Sousa, F. L. et al. 2016. One step beyond a ribosome: The ancient anaerobic core. - Biochim. Biophys. Acta 1857: 1027–1038.

Sriswasdi, S. et al. 2017. Generalist species drive microbial dispersion and evolution. - Nat. Commun. in press.

Tatusova, T. et al. 2016. NCBI prokaryotic genome annotation pipeline. - Nucleic Acids Res. 44: 6614–6624.

Toft, C. and Andersson, S. G. E. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. - Nat. Rev. Genet. 11: 465–475.

Tung Ho, L. si and Ané, C. 2014. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. - Syst. Biol. 63: 397–408.

Ulrich, L. E. and Zhulin, I. B. 2010. The MiST2 database: a comprehensive genomics resource on microbial signal transduction. - Nucleic Acids Res. 38: D401–D407.

Uyeda, J. C. et al. 2018. Rethinking phylogenetic comparative methods. - Syst Biol 67: 1091–1109.

Vadia, S. and Levin, P. A. 2015. Growth rate and cell size: a re-examination of the growth law. - Curr. Opin. Microbiol. 24: 96–103.

van Nimwegen, E. 2003. Scaling laws in the functional content of genomes. - Trends Genet. 19: 479–484.

van Nimwegen, E. 2006. Scaling laws in the functional content of genomes: fundamental constants of evolution? - In: Koonin, E. V. et al. (eds), Power Laws, Scale-Free Networks and Genome Biology. Eurekah.com and Springer Science, pp. 236–253.

Vieira-Silva, S. and Rocha, E. P. C. 2010. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. - PLOS Genet. 6: e1000808.

Wang, H.-C. et al. 2006. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: Data quality and confounding factors. - Biochem. Biophys. Res. Commun. 342: 681–684.

Wattam, A. R. et al. 2017. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. - Nucleic Acids Res. 45: D535.

Westoby, M. 1999. Generalization in functional plant ecology: the species sampling problem, plant ecology strategies schemes, and phylogeny. - In: Pugnaire, F. and Valladares, F. (eds), Handbook of Functional Plant Ecology. Marcel Dekker, pp. 847–872.

Westoby, M. et al. in review. Cell size, genome size and maximum growth rate are near-independent dimensions of ecological variation across bacteria and archaea. in press.

Wood, J. L. et al. 2018. Competitive Traits Are More Important than Stress-Tolerance Traits in a Cadmium-Contaminated Rhizosphere: A Role for Trait Theory in Microbial Ecology. - Front. Microbiol. in press.

Wright, I. J. et al. 2004. The worldwide leaf economics spectrum. - Nature 428: 821–827.

Xu, L. et al. 2006. Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms. - Mol. Biol. Evol. 23: 1107–1108.

Figure legends

Figure 1 (a) Distribution of genome sizes (all available species, excluding intracellular parasites and symbionts) in relation to oxygen use. (b) Relationships of genome size to aerobic vs anaerobic metabolism, to bacteria vs archaea and to optimum growth temperature. Ancova model giving rise to the fitted lines is described in Table S3.

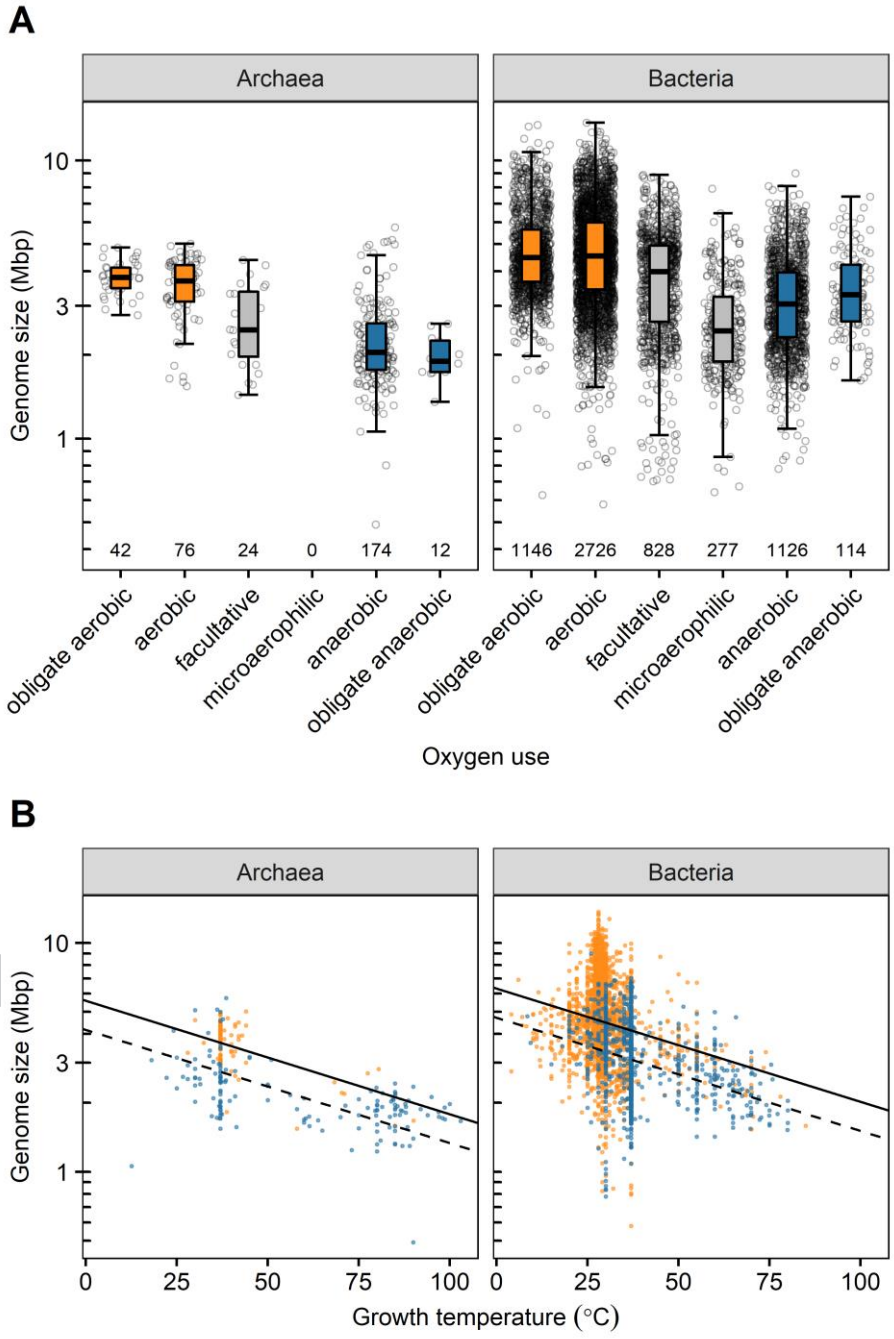

Fig 2 (a) Aerobic vs anaerobic genome sizes within the best-represented phyla. Taxonomy follows the genome taxonomy database (Parks et al. 2018), where clades are monophyletic as best as can be determined from the 120 genes used to build the tree, and where taxonomic ranks are normalized to a consistent range of relative depths in the tree. (b) Distribution of genome sizes of cultured species in relation to major habitats from which they come. Red dots show means for each habitat across the data compilation reported here. The boxplots separate contributions from aerobes and anaerobes.

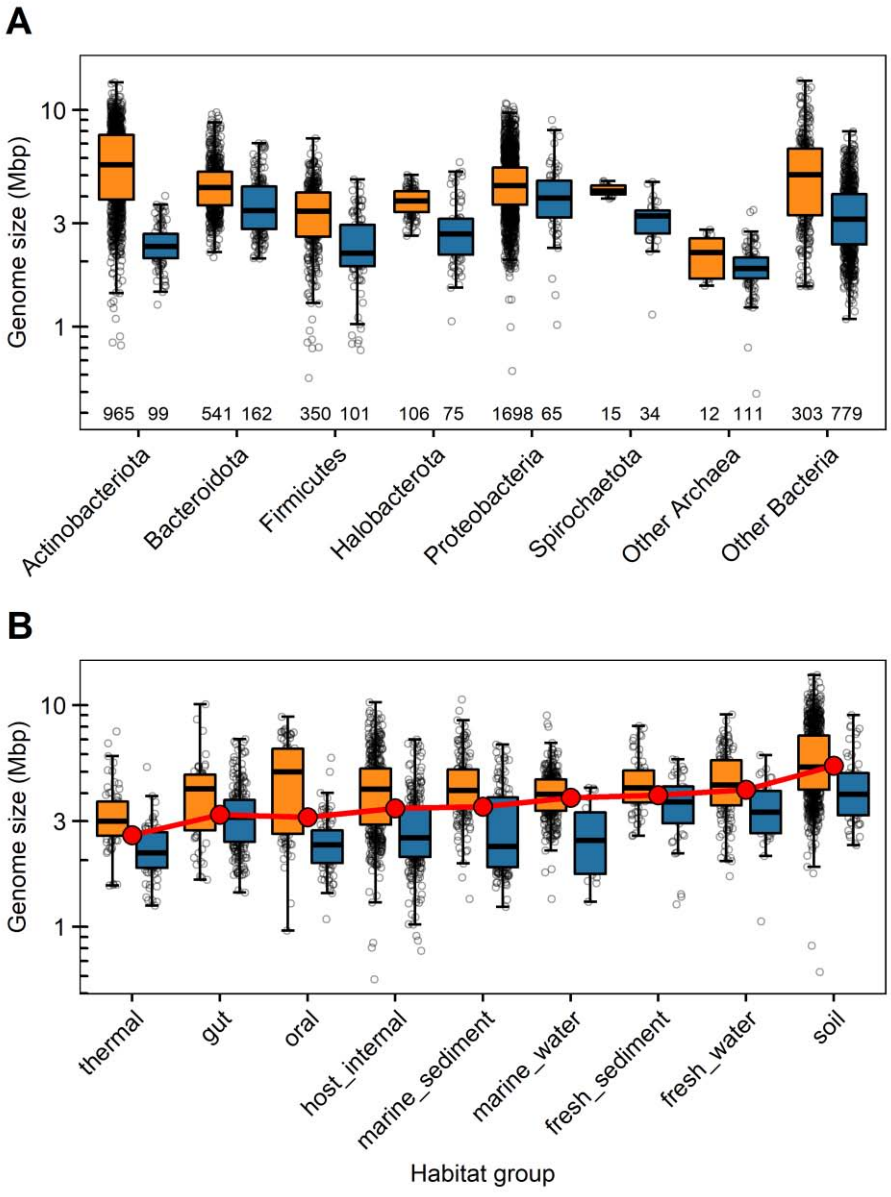

Figure 3. Scaling for particular categories of genes in relation to genome size, comparing aerobic (orange) with anaerobic (blue) species. Reference lines have slope 1 (isometry). Arrows run from the centroid for anaerobic to the centroid for aerobic species. To the extent they run parallel to the reference line, the relative contribution of a gene category to the genome is not changing. (a) Genes involved with signal reception and transduction, data from MIST database (https://mistdb.com/) (Ulrich and Zhulin 2010, Gumerov et al. 2020). Genes contributing to both one-component and two-component signalling systems are included; subsets are given in Fig S3. Reference line is set at a level that represents 100 signalling genes per Mb of total genome, or approximately 1 in 10 genes involved with signalling. (b) Numbers of genes for transport and metabolism of amino-acids, nucleotides, carbohydrates, inorganic ions, coenzymes and lipids. Genes attributed to their single best-fit COG-categories.

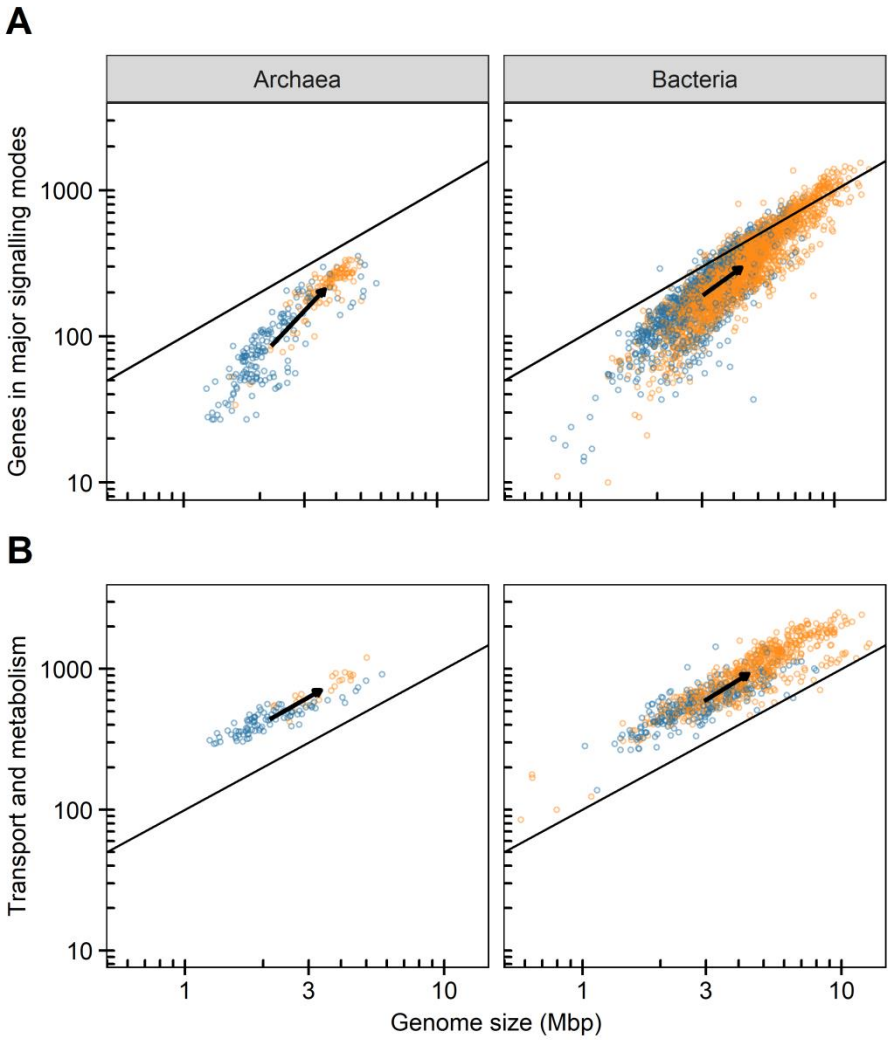

Table Legend

Table 1. Response coefficients (regression slopes) with 95% confidence intervals of $\log_{10}$ genome size in relation to other species traits, as simple OLS regression, as partials after superkingdom, growth temperature and oxygen use, and as phylogenetic generalised least squares. Percentage responses are the OLS and PGLS slopes translated to become more understandable. For binary contrasts they are percentage differences between the two groups. For cell diameter and minimum doubling time it is a percentage increase per tenfold change in the predictor. For rRNA copies and temperature it is a percentage increase per copy or per degree.. Asterisks indicate relationships where zero lies outside the 95% confidence interval. Tips indicates the number of species pairs for each trait-combination. Degrees of freedom for partial correlations are n-5.

		Ordinary least squares			Partial after superkingdom, growth temp and oxygen use		Phylogenetic generalized least squares		
trait	tips	coeff	% response	95% CI	Partial coeff	95% CI	PGLS coeff	% response	95% CI
Aerobe vs anaerobe	5520	0.187*	54	0.177, 0.198	not applic	not applic	0.0376*	9.0	0.0235, 0.0517
sporulation	4660	0.144*	39.4	0.13, 0.159	0.129*	0.117, 0.142	0.0604*	14.9	0.0436, 0.0772
motility	4800	0.0538*	13.2	0.0431, 0.0644	0.0292*	0.0182, 0.0401	0.0221*	5.2	0.0145, 0.0297
Rod vs spheroid	1490	0.14*	38.1	0.126, 0.154	0.0887*	0.074, 0.103	0.0239*	5.7	0.0126, 0.0352
Gram stain	7120	0.00873	2	-0.00111, 0.0186	0.00783	-0.00237, 0.018	0.0113	2.6	-0.00772, 0.0304
Cell diameter	3420	0.0814*	20.6	0.0551, 0.108	0.0937*	0.0681, 0.119	0.0397*	9.6	0.0242, 0.0552
Min doubling time	617	0.0294*	7	0.00531, 0.0536	-0.000829	-0.0238, 0.0221	0.0115	2.7	-0.0069, 0.0299
rRNA operon copies	2730	0.0265*	6.3	0.0236, 0.0294	0.00931*	0.00601, 0.0126	0.0322*	7.7	0.0292, 0.0352
Growth temp		-0.00816*	-1.9	-0.00859, -0.00773	not applic	not applic	-0.00368*	-0.8	-0.00418, -0.00318