Soil Biology and Biochemistry 178 (2023) 108935

Contents lists available at ScienceDirect

Soil Biology and Biochemistry

journal homepage: www.elsevier.com/locate/soilbio

Edaphic controls on genome size and GC content of bacteria in soil microbial communities

Peter F. Chuckran [a,d,1,*], Cody Flagg [b,2], Jeffrey Propster [a], William A. Rutherford [c,3], Ella T. Sieradzki [d], Steven J. Blazewicz [e], Bruce Hungate [a], Jennifer Pett-Ridge [e,f], Egbert Schwartz [a], Paul Dijkstra [a]

[a] Center for Ecosystem Science and Society (ECOSS) and Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA
[b] National Ecological Observation Network (NEON), Boulder, CO, USA
[c] School of Natural Resources and the Environment, University of Arizona, Tucson, AZ, USA
[d] Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA
[e] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA
[f] Life & Environmental Sciences Department, University of California Merced, Merced, CA, USA

ARTICLE INFO

Keywords:
Metagenomics
pH
Amino acid stoichiometry
C:N ratios
Genomic traits

ABSTRACT

Nutrient limitation has been shown to reduce bacterial genome size and influence nucleotide composition; however, much of this work has been conducted in marine systems and the factors which shape soil bacterial genomic traits remain largely unknown. Here we determined average genome size, GC content, codon usage, and amino acid content from 398 soil metagenomes across a broad geographic range and used machine-learning to determine the environmental parameters that most strongly explain the distribution of these traits. We found that genomic trait averages were most related to pH, which we suggest is primarily due to the correlation of pH with several environmental parameters, particularly soil carbon content. Low pH soils had higher carbon to nitrogen ratios (C:N) and tended to have communities with lower GC content and larger genomes, potentially a response to increased physiological stress and a requirement for metabolic diversity. Conversely, communities in high pH and low soil C:N had smaller genomes and higher GC content—indicating potential resource driven selection against AT base pairs, which have a higher C:N than GC base pairs. Similarly, we found that nutrient conservation also applied to amino acid stoichiometry, where bacteria in soils with low C:N ratios tended to code for amino acids with lower C:N. Together, these relationships point towards fundamental mechanisms that underpin genome size, and nucleotide and amino acid selection in soil bacteria.

In bacteria, nutrient constraints exert influence on traits such as genome size, GC content, codon frequency, and amino acid content (Batut et al., 2014; Giovannoni et al., 2014; Shenhav and Zeevi, 2020). For free-living bacteria, low nutrient concentrations often select for genomic traits that reduce the cost of reproduction, such as low GC content and smaller genomes (Giovannoni et al., 2014). A GC base pair requires more energy to produce than an AT base pair (Chen et al., 2016; Rocha and Danchin, 2002). Moreover, since the GC base pair has a carbon to nitrogen ratio (C:N) of 9:8 (1.13), whereas a AT base pair has a C:N of 10:7 (1.42), GC-rich genomes require more nitrogen, which may be disadvantageous in nitrogen-limited environments. Genomic traits are therefore not only important as fundamental metrics describing genomes, but also as indicators of evolutionary history and life strategy.

However, much of the existing and foundational literature on processes controlling genomic traits in free-living bacteria are based on the study of marine isolates (Giovannoni et al., 2005) and aspects of this framework may not cleanly transpose onto soil bacteria. For example, we had previously found that community-averaged GC content and genome size were positively correlated among marine metagenomes—a relationship attributed to N-limitation; however, GC content and

* Corresponding author. Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA.
E-mail address: pfchuckran@gmail.com (P.F. Chuckran).
[1] Current affiliation: Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA.
[2] Current affiliation: Charter Communications Inc., Greenwood Village, CO, USA.
[3] Current affiliation: Southwest Watershed Research Center, Agricultural Research Service, US Department of Agriculture, Tucson, AZ, USA.

https://doi.org/10.1016/j.soilbio.2022.108935
Received 13 September 2022; Received in revised form 19 December 2022; Accepted 26 December 2022
Available online 27 December 2022

genome size were negatively correlated between metagenomes collected from soils (Chuckran et al., 2022). Since the growth of soil bacteria is thought to be more limited by carbon than nitrogen (Hobbie and Hobbie, 2013; Soong et al., 2020), we hypothesized that the distribution of genomic traits in soil bacteria might exhibit unique patterns reflecting carbon limitation. Specifically, because the GC base pair (1.13) has a lower C:N than the AT base pair (1.42), we predicted communities would exhibit higher GC content and smaller genomes when carbon availability is low. Carbon limitation has been indicated as a potential driver of high GC content in microbial models (Hellweger et al., 2018), and previous studies have shown that microbial communities from bulk soil (often carbon-poor) tend to have smaller genomes and higher GC content than those in the rhizosphere (Chen et al., 2021). However, the relationship between soil carbon and the genomic traits of soil bacteria has not been properly assessed and represents a fundamental unknown in our understanding of belowground microbial life.

To better understand the relationship between genomic traits of soil bacteria and environmental characteristics, we analyzed 398 metagenomes collected and sequenced by the US-based National Ecological Observation Network (NEON) (National Ecological Observatory Network (NEON), 2021) across a broad geographic scale and analyzed community-averaged genomic traits alongside a range of environmental and edaphic properties (see Supplemental Methods). Genome size was estimated from metagenomic raw reads (using single-copy genes), bacterial contigs, and by aligning amplicon datasets to genomes of known size. Although each of these methods may contain biases, the source of error is different for each method, and agreement between estimates deepens evidence for observed trends. We hypothesized that microbial communities in low carbon environments would exhibit smaller genome sizes, higher GC content, and amino acid composition with a lower carbon to nitrogen ratio (C:N). To test this, we assessed the relationship between extractable soil C:N (C_{extr}:N_{extr}) and genomic traits across all NEON sites (Fig. 1A). We found a negative correlation between GC content and C_{extr}:N_{extr} ($p < 0.001$; Fig. 1B), a trend which extended to the average GC content of the 16S rRNA gene (Supplemental Fig. 2D). We also found a positive correlation between C_{extr}:N_{extr} and estimates of average genome size derived from multiple estimates of genome size ($p < 0.001$; Fig. 1C). Further, we found that average genome size and GC content calculated from bacterial contigs were negatively correlated ($p < 0.01$, Fig. 1D). This relationship is unique in comparison to free-living marine bacteria, where smaller genomes tend to have lower GC content (Giovannoni et al., 2014).

The C:N of the sum of all predicted amino acids was positively correlated with soil C_{extr}:N_{extr} ($R^2 = 0.24$, $p < 0.001$, Fig. 2A), and closely tracked metagenome GC content ($R^2 = 0.52$, $p < 0.001$, Fig. 2B), reflecting the resource alignment between the stoichiometry of nucleic acids in codons and their corresponding amino acids (Bragg and Hyder, 2004). We found that synonymous codon usage skewed towards codons with a higher GC content, perhaps best represented by the strong preference for guanine and cytosine at fourfold degenerate sites ($p < 0.01$; Fig. 2C). This suggests that GC content is not solely a consequence of amino acid composition, and that there is additional nucleotide bias within synonymous substitutions. Notably, we also observed a higher abundance of cytosine than guanine and fourfold degenerate sites (Fig. 2C), which could be a consequence of the lower metabolic cost of cytosine production (Chen et al., 2016). Preferential selection for codons with higher AT content was most pronounced where soil C:N was high. For each amino acid, codons with higher GC were more often negatively correlated with soil C_{extr}: N_{extr} compared to codons with lower GC, which more often were positively correlated with soil C_{extr}:N_{extr} (Fig. 2D&E).

In line with our original hypothesis, lower soil C:N was associated with communities averaging smaller genomes, higher GC, and lower C:N of amino acids. However, genomic traits in soil microbial communities may also be driven by other environmental factors, such as temperature (Sabath et al., 2013; Sorensen et al., 2019) and pH (Gravuer and Eskelinen, 2017). To assess the relationships between genomic traits and other environmental drivers, we used a machine learning, random-forest model approach to determine the environmental variables that explain the most variance in GC content and predicted average genome size. With this model, we assessed the importance of over 100 environmental factors and geographic range in shaping genomic features.

Random forest models indicated that GC content and the average genome size of a community were most strongly related to soil pH (Fig. 3A), where soils with low pH fostered communities with low GC content (Fig. 3B) and larger average genome size (Fig. 3C). Although, large fungal genomes tended to be more present in low pH soils (Supplemental Fig. 1A) and correlated with larger community-average genome size (Supplemental Fig. 1B), similar relationships were observed when genome size was predicted from single copy genes detected in bacterial contigs (Supplemental Results & Discussion; Supplemental Fig. 1C) as well as 16S rRNA gene taxonomy (Supplemental

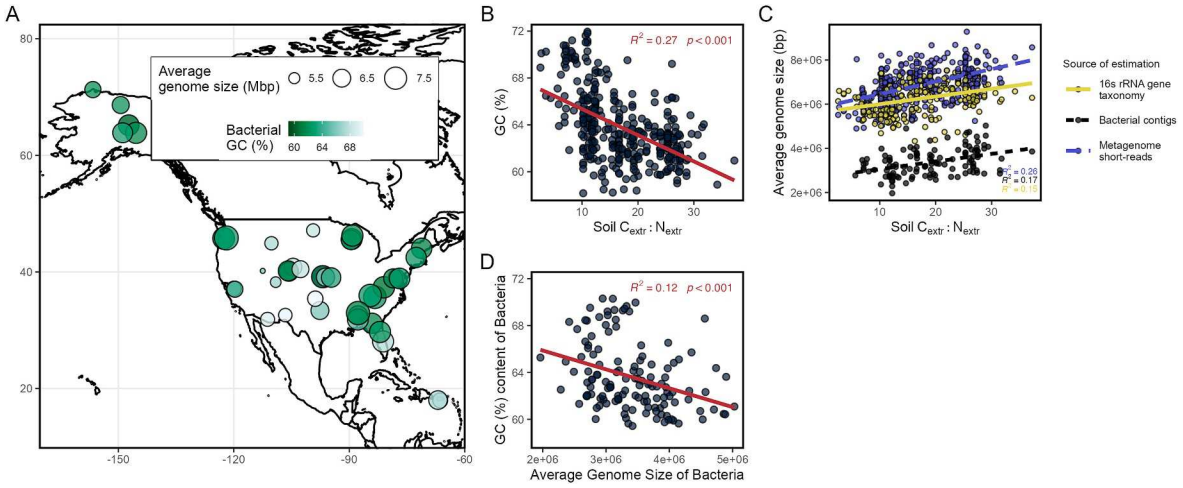

Fig. 1. Distribution of genomic traits across sites and soil extractable carbon and extractable nitrogen ratios (C_{extr}:N_{extr}); **(A)** Geographic distribution of sites, with mean bacterial GC and estimated average genome size. **(B)** Relationship between C_{extr}:N_{extr} and bacterial GC content (%) (linear regression, $p < 0.001$). **(C)** Relationship between C_{extr}:N_{extr} and genome size, estimated from the number of single copy genes per metagenome (blue, linear regression, $p < 0.001$), from 16S rRNA gene datasets and genome size estimated from isolates (yellow circles, linear regression, $p < 0.001$), and from single copy genes in the assembled bacterial contigs (black, $p < 0.01$). **(D)** The average GC content of all bacterial contigs in a metagenome vs average genome size estimated from bacterial contigs. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

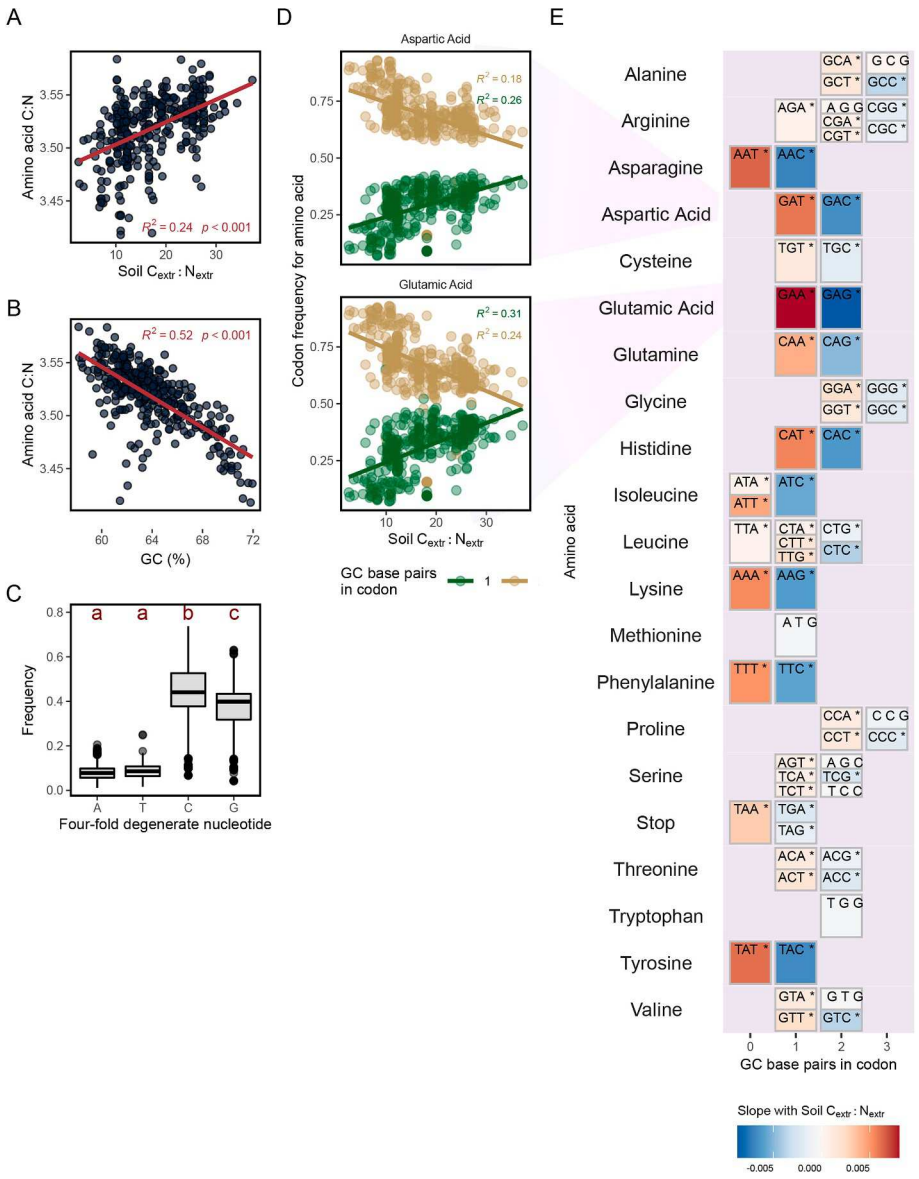

Fig. 2. Averaged nucleotide, codon, and amino acid composition of bacteria in metagenomes **(A)** Relationship between C_{extr}:N_{extr} and bacterial amino acid C:N ratios averaged per metagenome (linear regression, $p < 0.001$). **(B)** Relationship between bacterial GC and pooled amino acid C:N ratios (linear regression, $p < 0.001$). **(C)** The distribution of nucleotides at the third position in fourfold degenerate codons across all metagenomes, with letters corresponding to groups identified via Tukey's post-hoc test ($p < 0.01$). **(D)** The relationship between codon frequency and soil C_{extr}:N_{extr} shown for aspartic acid and glutamic acid, with number of GC base pairs in each codon being indicated by color (linear regression, $p < 0.001$). **(E)** The relationship between codon frequency and soil C_{extr}:N_{extr} shown for each codon, with color indicating the slope of the relationship and an asterisk indicating significance (linear regression, $p < 0.05$). Codons are arranged left to right in increasing number of GC base pairs. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Fig. 2). All three of these estimates may be subject to biases; however, the source of bias of each varies, and the co-occurrence of these patterns provides strong support that the observed changes in average genome size are driven by changes in bacterial genome size.

There are several reasons why pH might be correlated with this set of genomic characteristics in soil bacteria. Soil pH represents the intersection of numerous environmental vectors and, accordingly, we hypothesize that there are several mechanisms underpinning the relationship between pH and genomic traits. First, both low and high pH can cause physiological stress in bacteria, with acidic soils having been shown to be associated with a greater number of repair mechanisms, such as chaperones (Malik et al., 2018). This might preferentially select for bacteria with a greater investment in stress alleviation and maintenance, and thus larger genomes. Since our analysis does not include as many alkaline as acidic soils, we are unable to assess the impact of high pH on traits; however, alkaline soils are highly prevalent worldwide and the distribution of traits in these soils deserves further study. Second, low pH is often associated with the accumulation of soil organic carbon (SOC). Since soil pH is largely driven by the balance between precipitation and evapotranspiration (Slessarev et al., 2016), low pH often coincides with greater precipitation excess and primary production. Higher biomass inputs into acidic soils, combined with a reduction in the decomposition rate due to low pH, results in the build-up of SOC (Malik et al., 2018). The accumulation of SOC not only alleviates carbon limitation—which may reduce GC content—but also potentially favors larger genomes with increased metabolic diversity. It has been suggested that the requirement for increased metabolic diversity might explain why soil bacterial genomes tend to have large genomes (Barberán et al., 2014) and, similarly, we found that soils with lower pH were associated with higher C_{extr}:N_{extr} as well as larger genomes and lower GC content (Fig. 3D). High pH in soil can also result in the build-up of microbial biomass and increases in SOC (Malik et al., 2018). However, due to the limited range of our alkaline soil data, we cannot determine the impact of that effect here. Third, genomic traits in soil bacteria may relate to other forms of stress coinciding with pH. Aridity has been shown to drive streamlining in certain soil bacteria (Simonsen, 2022) and, as discussed above, influences the pH in soil. Previous work has shown relationships between precipitation, pH, and genome size (Gravuer and Eskelinen, 2017), and in a previous analysis we found that soil metagenomes collected from both hot and cold deserts often had smaller genomes and greater GC content than soils collected in more mesic systems (Chuckran et al., 2022). We found that mean annual precipitation and average

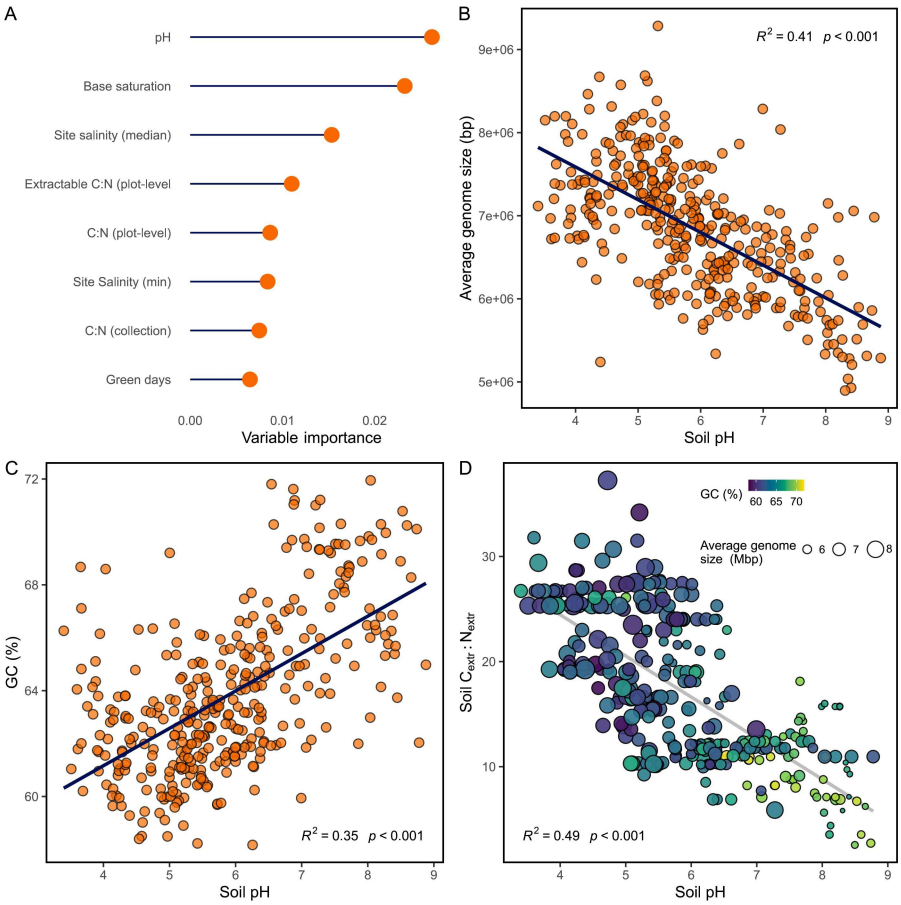

Fig. 3. Results from the random forest model and relationships between soil pH and genomic traits; **(A)** Variable importance plot for the top 8 environmental parameters predicting GC content from the random forest model (RMSE = 0.017; R^2 = 0.66). **(B)** The relationship between soil pH and GC content (%) of bacterial contigs, with a linear relationship as selected by the random forest model ($p < 0.01$). **(C)** The relationship between soil pH and average genome size (derived from metagenomes) with a linear relationship as selected by the random forest model (blue; $p < 0.01$). **(D)** The relationship between soil pH and soil C_{extr}:N_{extr} with points colored by bacterial GC content and point size corresponding to average genome size. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

genomes size were related (Supplemental Fig. 3); however, the relationship was not as strong as with pH or soil C_{extr}:N_{extr}.

The underlying forces which drive these trends is likely a mix of changes in community composition and selection for traits within a taxon across an environmental gradient. Genome size estimated from 16S rRNA gene taxonomy demonstrated trends similar to those estimated in metagenomes, suggesting that community assemblage is a predominant force driving relationships between average genome size and edaphic characteristics (Fig. 1C). Genome size is a phylogenetically conserved trait in bacteria (Martinez-Gutierrez and Aylward, 2022), varies greatly in soil microbes, and is known to relate to life strategies. For example, *Candidatus* Udaeobacter copiosus is a streamlined soil bacterium with a genome of only 2.81 Mbp, largely at the cost of metabolic diversity (Brewer et al., 2017). Streptomyces species are known to be ubiquitous in soils and have genomes longer than 8 Mbp and are known to have in increased investment in regulation and nutrient transport (Konstantinidis and Tiedje, 2004). Interestingly, average metagenomic and 16S rRNA gene GC content were not related to GC content predicted from gene taxonomy (Supplemental Discussion; Supplemental Fig. 4), indicating that nucleotide frequency may be a less phylogenetically conserved trait than genome size in soil bacteria.

Still, more work must be done to assess at what phylogenetic level these traits emerge and how they connect to life-strategies and frameworks commonly used by soil microbial ecologists—such as the Yield-Acquisition-Stress (YAS) (Malik et al., 2020). In this analysis, we found that low pH was associated with larger genomes in resource-abundant systems, suggesting that large genomes maybe be a stress (S) related trait, whereas high GC content might be more closely associated with resource limitation and an acquisition (A) life strategy. Acquisition life strategists seemed to also be associated with a dimension of stress through water limitation but exhibited traits opposite to pH stress tolerators. It may be that water, as a necessary element for life, activates trait relationships that are more like those that arise from resource limitation as opposed to long-term chemical stress, such as acidity.

Our work demonstrates that the broad-scale distribution of genomic traits in soil bacterial communities is correlated with soil pH, which we suggest can be attributed to pH being a metric that captures multiple parameters, such as soil nutrients and precipitation patterns, as well as physiological stress. We found several trends that suggest that selection pressure in soil bacterial communities might reflect carbon limitation, for example, the negative relationship between genome size and GC content. The overall influence of C:N on nucleotide selection is similar to what has been observed in marine systems (i.e. low C:N selects for higher GC content, and higher C:N selects for lower GC content); however, the reduction in genome size with high GC content and low soil carbon is distinct, and suggests that carbon limitation is driving the distribution of these traits. These results are derived from community averages and more work must be done to uncover both the mechanisms and taxonomic level where the observed changes in genomic traits occur. However, it is evident that the distribution of genomic traits in soil is related to edaphic properties and deserves further study.

Funding

This work was supported by funding from the USDA National Institute of Food and Agriculture Foundational Program (award #2017-67019-26396). Support for SB, ES, JP, PD, and BH was provided by the U.S. Department of Energy, Office of Biological and Environmental Research, Genomic Science Program LLNL 'Microbes Persist' Soil Microbiome Scientific Focus Area (award #SCW1632). Work conducted at LLNL was conducted under the auspices of the US Department of

Energy under Contract DE-AC52-07NA27344. Funding agencies did not play a role in study design; the collection, analysis, and interpretation of data; or writing of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and code are published on github with corresponding links in manuscript

Acknowledgements

We would like to thank Anita Antoninka, Michaela Hayer, Alicia Purcell, Junhui Li, Megan Foley, Raina Fitzpatrick, Bram Stone, Victoria Monsaint-Queeney, and Carl Roybal for their intellectual contributions to this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.soilbio.2022.108935.

References

Barberán, A., Ramirez, K.S., Leff, J.W., Bradford, M.A., Wall, D.H., Fierer, N., 2014. Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. Ecology Letters 17, 794–802. https://doi.org/10.1111/ELE.12282.

Batut, B., Knibbe, C., Marais, G., Daubin, V., 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. Nature Reviews Microbiology 12, 841–850. https://doi.org/10.1038/nrmicro3331.

Bragg, J.G., Hyder, C.L., 2004. Nitrogen versus carbon use in prokaryotic genomes and proteomes. Proceedings of the Royal Society B: Biological Sciences 271, 374–377. https://doi.org/10.1098/rsbl.2004.0193.

Brewer, T.E., Handley, K.M., Carini, P., Gilbert, J.A., Fierer, N., 2017. Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus. ' Nature Microbiology 2, 16198. https://doi.org/10.1038/nmicrobiol.2016.198.

Chen, W.-H., Lu, G., Bork, P., Hu, S., Lercher, M.J., 2016. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. Nature Communications 7, 11334. https://doi.org/10.1038/ncomms11334.

Chen, Y., Neilson, J.W., Kushwaha, P., Maier, R.M., Barberán, A., 2021. Life-history strategies of soil microbial communities in an arid ecosystem. The ISME Journal 15, 649–657. https://doi.org/10.1038/s41396-020-00803-y.

Chuckran, P.F., Hungate, B.A., Schwartz, E., Dijkstra, P., 2022. Variation in genomic traits of microbial communities among ecosystems. FEMS Microbes 2, 20. https://doi.org/10.1093/FEMSMC/XTAB020.

Giovannoni, S.J., Cameron Thrash, J., Temperton, B., 2014. Implications of streamlining theory for microbial ecology. The ISME Journal 8, 1553–1565. https://doi.org/10.1038/ismej.2014.60.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., Rappé, M.S., Short, J.M., Carrington, J. C., Mathur, E.J., 2005. Genetics: genome streamlining in a cosmopolitan oceanic bacterium. Science 309, 1242–1245. https://doi.org/10.1126/science.1114057.

Gravuer, K., Eskelinen, A., 2017. Nutrient and rainfall additions shift phylogenetically estimated traits of soil microbial communities. Frontiers in Microbiology 8, 1271. https://doi.org/10.3389/fmicb.2017.01271.

Hellweger, F.L., Huang, Y., Luo, H., 2018. Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model. The ISME Journal 12, 1180–1187. https://doi.org/10.1038/s41396-017-0023-7.

Hobbie, J.E., Hobbie, E.A., 2013. Microbes in nature are limited by carbon and energy: the starving-survival lifestyle in soil and consequences for estimating microbial rates. Frontiers in Microbiology 4, 324. https://doi.org/10.3389/fmicb.2013.00324.

Konstantinidis, K.T., Tiedje, J.M., 2004. Trends between Gene Content and Genome Size in Prokaryotic Species with Larger Genomes, vol. 10, pp. 3160–3165.

Malik, A.A., Martiny, J.B.H., Brodie, E.L., Martiny, A.C., Treseder, K.K., Allison, S.D., 2020. Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. The ISME Journal 14, 1–9. https://doi.org/10.1038/s41396-019-0510-0.

Malik, A.A., Puissant, J., Buckeridge, K.M., Goodall, T., Jehmlich, N., Chowdhury, S., Gweon, H.S., Peyton, J.M., Mason, K.E., van Agtmaal, M., Blaud, A., Clark, I.M., Whitaker, J., Pywell, R.F., Ostle, N., Gleixner, G., Griffiths, R.I., 2018. Land use driven change in soil pH affects microbial carbon cycling processes. Nature Communications 9. https://doi.org/10.1038/s41467-018-05980-1.

Martinez-Gutierrez, C.A., Aylward, F.O., 2022. Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. PLoS Genetics 18, e1010220. https://doi.org/10.1371/JOURNAL.PGEN.1010220.

National Ecological Observatory Network (NEON), 2021. Soil Microbe Metagenome Sequences (DP1.10107.001). https://doi.org/10.48443/FZZJ-G053.

Rocha, C., Danchin, A., 2002. Base composition bias might result from competition for metabolic resources. TRENDS in genetics 18, 291–294.

Sabath, N., Ferrada, E., Barve, A., Wagner, A., 2013. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. Genome Biology and Evolution 5, 966–977. https://doi.org/10.1093/gbe/evt050.

Shenhav, L., Zeevi, D., 2020. Resource conservation manifests in the genetic code. Science 370, 683–687. https://doi.org/10.1126/science.aaz9642.

Simonsen, A.K., 2022. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. The ISME Journal 16, 423–434. https://doi.org/10.1038/s41396-021-01082-x.

Slessarev, E.W., Lin, Y., Bingham, N.L., Johnson, J.E., Dai, Y., Schimel, J.P., Chadwick, O. A., 2016. Water balance creates a threshold in soil pH at the global scale. Nature 540, 567–569. https://doi.org/10.1038/nature20139.

Soong, J.L., Fuchslueger, L., Marañon-Jimenez, S., Torn, M.S., Janssens, I.A., Penuelas, J., Richter, A., 2020. Microbial carbon limitation: the need for integrating microorganisms into our understanding of ecosystem carbon cycling. Global Change Biology 26, 1953–1961. https://doi.org/10.1111/gcb.14962.

Sorensen, J.W., Dunivin, T.K., Tobin, T.C., Shade, A., 2019. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. Nature Microbiology. https://doi.org/10.1038/s41564-018-0276-6.