TITLE:

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

AUTHORS:

Peter F. Chuckran[1,4*]

Cody Flagg[2]

Jeffrey Propster[1]

William A. Rutherford[3]

Ella Sieradzki[4]

Steven J. Blazewicz[5]

Bruce Hungate[1]

Jennifer Pett-Ridge[5,6]

Egbert Schwartz[1]

Paul Dijkstra[1]

[1] Center for Ecosystem Science and Society (ECOSS) and Department of Biological Sciences, Northern Arizona University, Flagstaff, Arizona, USA

[2] National Ecological Observation Network (NEON), Boulder, Colorado, USA

[3] School of Natural Resources and the Environment, University of Arizona, Tucson, Arizona, USA

[4] Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA

[5] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California, USA

[6] Life & Environmental Sciences Department, University of California Merced, Merced, California, USA

[*] Current affiliation

Competing interest statement:

The authors have no competing interests to disclose.

ABSTRACT:

Bacteria in soil microbial communities are crucial to terrestrial ecosystem function, yet our understanding of the fundamental characteristics of their genomes, such as GC content and genome size, is not complete. Much of our understanding of the mechanisms which shape their genomic traits is derived from other systems or from isolated bacteria. 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 which environmental parameters 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, especially 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, smaller genomes with high GC content were associated with high pH and low soil C:N, indicating potential resource driven selection against AT base pairs. As soil bacteria tend to be more carbon limited, smaller genomes with higher GC content may reduce the cost of reproduction in carbon-limited soils. 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 nucleotide and amino acid selection in soil bacterial communities.

MAIN TEXT:

In bacteria, nutrient constraints exert influence on traits such as genome size, GC content, codon frequency, and amino acid content [1–3]. 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 [3]. A GC base pair requires more energy to produce than an AT base pair [4, 5]. 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. Additionally, there is a correlation between the elemental requirements for nucleic and amino acids, such that codons with higher GC content (which have lower C:N) tend to code for amino acids with low C:N [6]. Therefore, nucleotide frequency not only controls the energy requirements for DNA synthesis, but also directly relates to the nutritional demand for amino acid synthesis [7].

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 [8] 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, the correlation between GC content and genome size was negatively correlated between metagenomes collected from soils [9]. Since the growth of soil bacteria is thought to be more limited by carbon than nitrogen [10, 11], 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 [12], 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 [13]. 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) [14] across a broad geographic scale and analyzed community-averaged genomic traits alongside a range of environmental and edaphic properties.

RESULTS AND DISCUSSION:

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); which would be predominantly driven by carbon limitation in the bacterial community. 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), and a positive correlation between C_{extr}:N_{extr} and both estimates of average 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 [3].

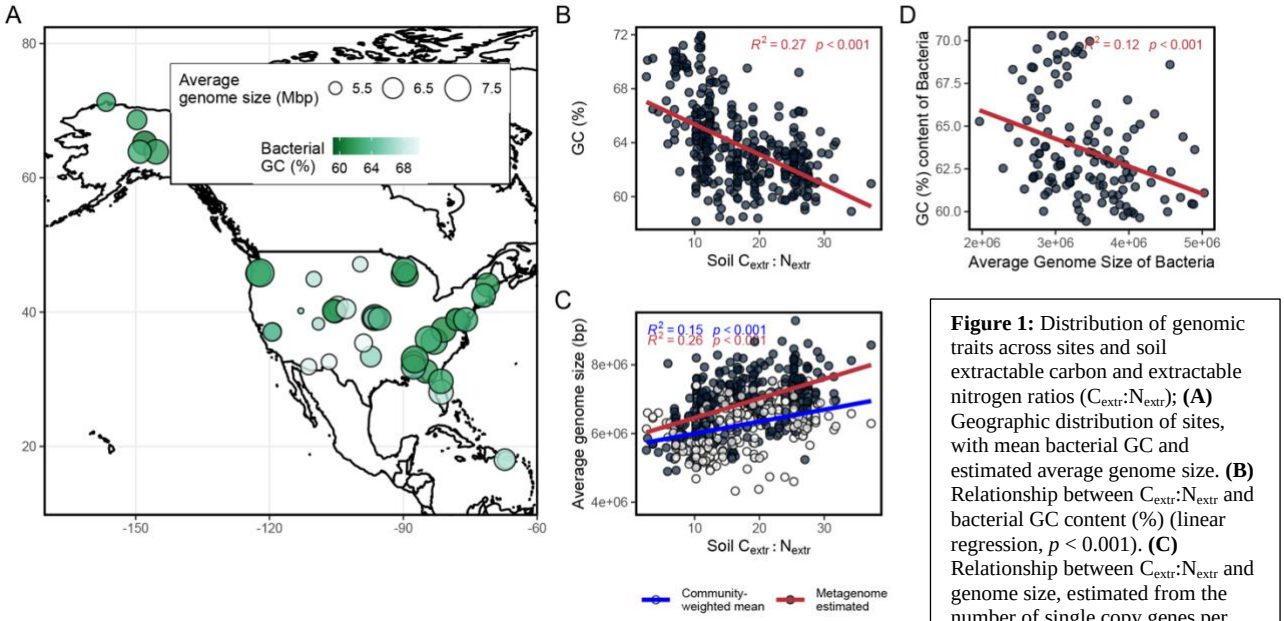

Figure 1: Distribution of genomic traits across sites and soil extractable carbon and extractable nitrogen ratios (C_{extr}:N_{extr}); **(A)** Geographic distribution of sites, with mean bacterial GC and estimated average genome size. **(B)** Relationship between C_{extr}:N_{extr} and bacterial GC content (%) (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 (dark circles, linear regression, $p < 0.001$) and from 16S rRNA gene datasets and genome size estimated from isolates (open circles, linear regression, $p < 0.001$). **(D)** The average GC content of all bacterial contigs in a metagenome vs average genome size estimated from bacterial contigs.

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=0.51, $p < 0.001$, Fig.

2B), reflecting the resource alignment between the stoichiometry of nucleic acids in codons and their corresponding amino acids [6]. We found that synonymous codon usage skewed towards codons with a higher GC content, perhaps best represented by the strong preference for guanine

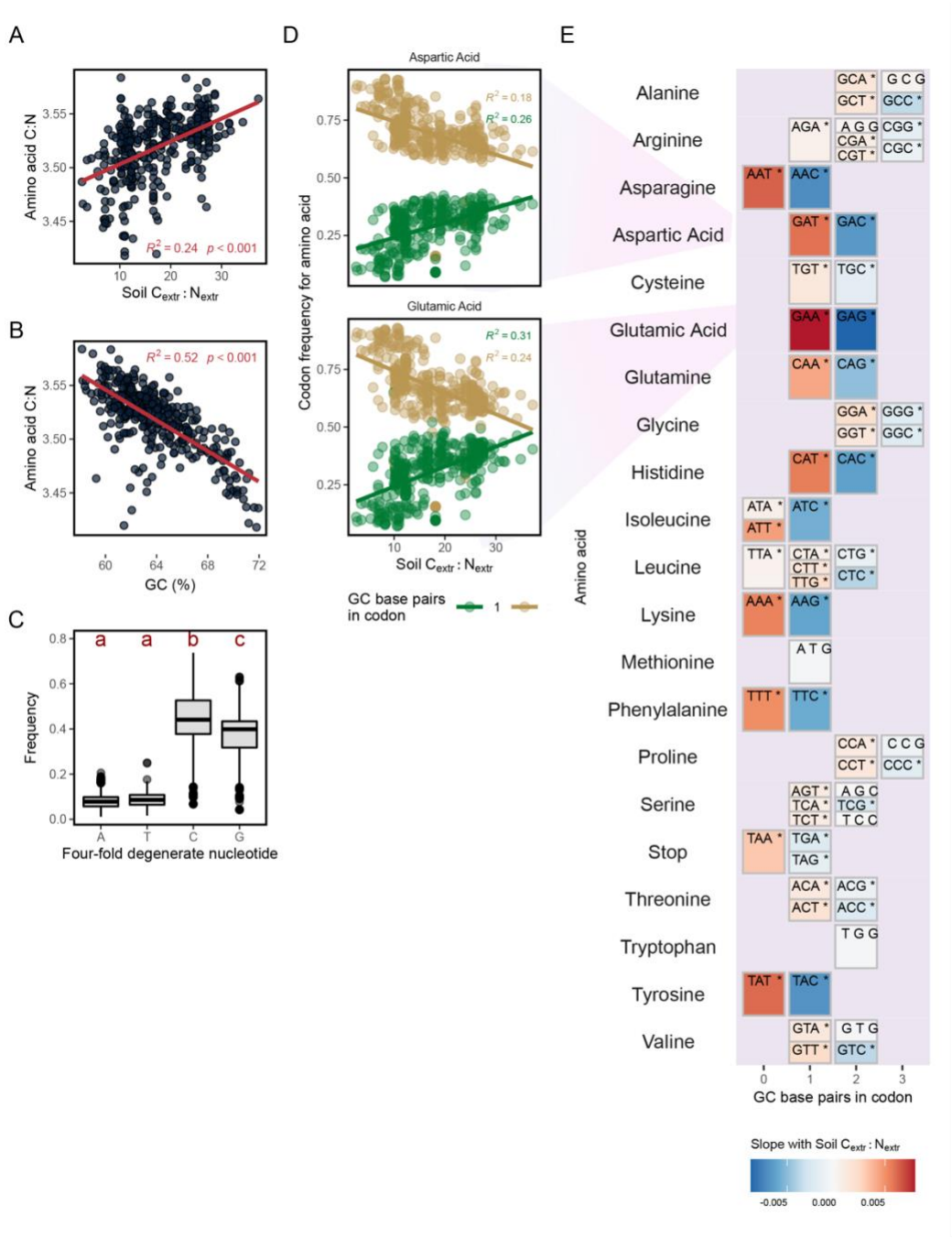

Figure 2: Averaged nucleotide, codon, and amino acid composition of bacteria in metagenomes **(A)** Relationship between C_{extr}:N_{extr} and bacterial amino acid C:N ratios averaged per metagenome (linear regression, $p < 0.001$). **(B)** Relationship between bacterial GC and pooled amino acid C:N 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.

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.

Notably, we also observed a higher abundance of cytosine than guanine and fourfold degenerate sites (Fig. 2C) that could be a consequence of the lower metabolic cost of cytosine production [4]. 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 [15, 16] and pH [17]. 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). 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). However, 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. 2), indicating that the observed trends are also being

driven by changes in bacterial genome size (as opposed to solely changes in the abundance of non-bacterial reads).

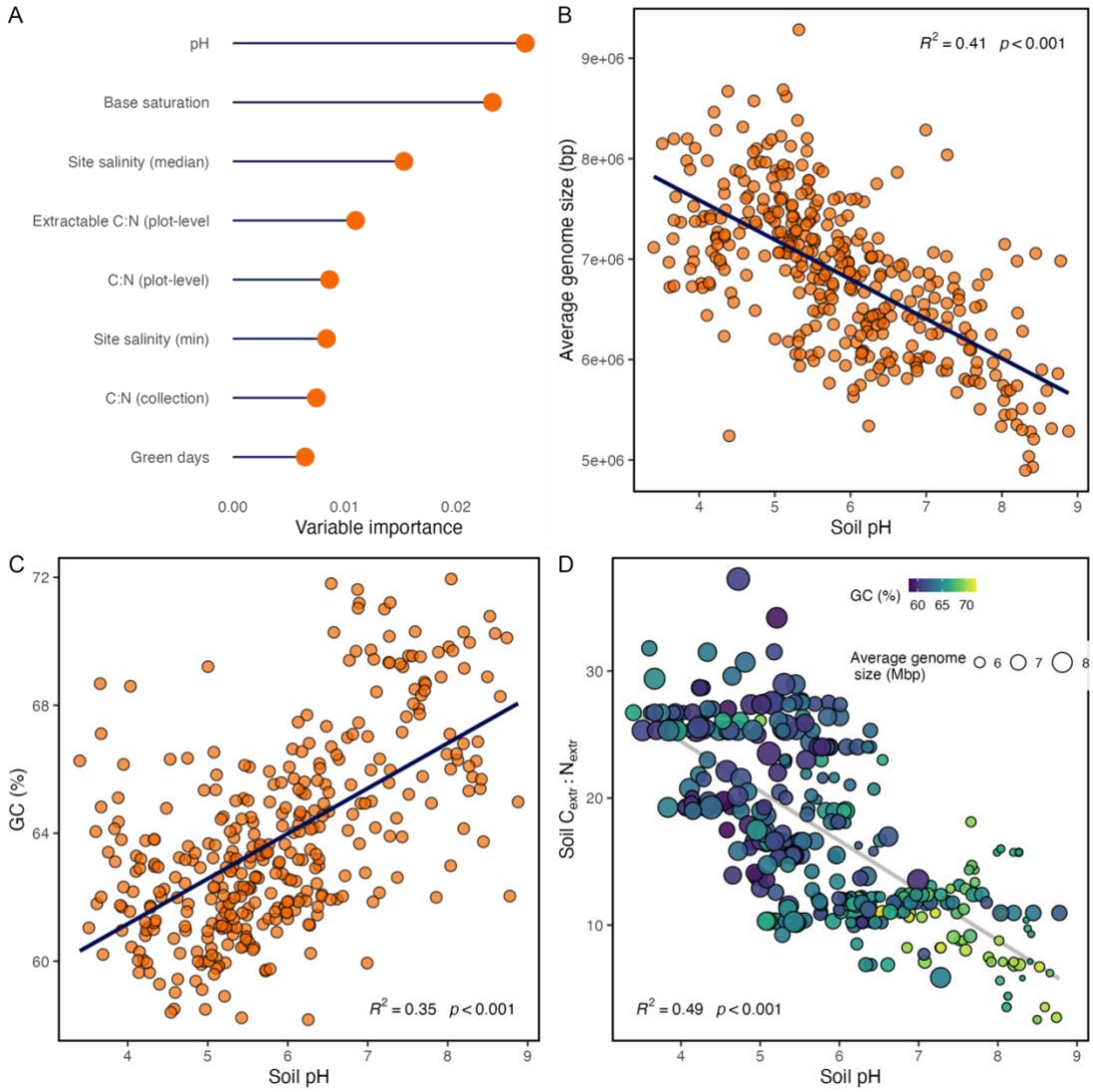

Figure 3: Results from the random forest model and relationships between soil pH and genomic traits; **(A)** Variable importance plot for the top 8 environmental parameters predicting GC content from the random forest model (RMSE = 0.017; R^2 = 0.66). **(B)** The relationship between soil pH and GC content (%) 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.

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, low pH causes physiological stress in soil bacteria and is often associated with a greater number of repair mechanisms, such as chaperones [18]. This might preferentially select for bacteria with a greater investment in stress alleviation

and maintenance, and thus larger genomes. 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 [19], 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 [18]. 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 [20] 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 [18], which could explain why we found that more alkaline soils had higher C_{extr}:N_{extr}, larger genomes, and lower GC content than soils with a neutral pH. 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 [21] and, as discussed above, influences the pH in soil. Previous work has shown relationships between precipitation, pH, and genome size [17], 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 [9]. Although we did find a positive relationship between mean annual precipitation and average genomes size (Supplemental Fig. 3), the relationship was not as strong as that for pH or soil C_{extr}:N_{extr}.

Our work demonstrates that soil pH determines the broad-scale distribution of genomic traits between soil bacterial communities, which we suggest can be attributed to pH being a

metric that captures multiple environmental parameters, such as soil nutrients and precipitation patterns, as well as physiological stress. Soil pH is well established driver of community composition in soils [22] and our results emphasize the degree to which pH dictates belowground microbial life and perhaps influences the evolution of soil bacteria. Additionally, 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.

173 METHODS

174 Data for this project was gathered from the National Ecological Observation Network (NEON)—
175 an observational network collecting ecological data from across the United States, funded by the
176 US National Science Foundation. NEON maintains 81 field sites in a number of distinct biomes
177 across the US. Terrestrial sites include a central meteorological station as well as many dispersed
178 plots from which samples are collected [23]. A full description of NEON sites can be found at
179 https://www.neonscience.org/field-sites/about-field-sites. Data products used in this analysis can
180 be found in Supplemental Table 1. A full list of data used in this analysis, as well as associated
181 descriptions of this data, can be found in Supplemental Tables 2 and 3, and metagenome
182 assembly statistics in Supplemental Table 4.

183

184 *Metagenomic traits*

185 From the NEON data portal (https://www.neonscience.org/data), we accessed the metadata for
186 all available metagenomes as of January 2021. We selected 398 metagenomes which
187 demonstrated the greatest read depth while maximizing the number of collection sites (43 total).
188 In January of 2021, selected metagenomes were downloaded from the NEON data portal to the
189 high-performance computing cluster at Northern Arizona University.

190 Raw reads were QC filtered using Trimmomatic v0.39 (parameters: TruSeq2-
191 PE.fa:2:30:10 LEADING:10 TRAILING:10 SLIDINGWINDOW:10:20 MINLEN:50) [24]. We
192 then used the program MicrobeCensus [25] to determine the average genome size of a microbial
193 community from the QC-filtered reads. MicrobeCensus estimates the total number of single copy
194 genes to create an estimate of average genome size from unassembled reads. Contigs were
195 assembled from QC-filtered reads using MEGAHIT v1.2.9 (parameters: --k-list 21,29,39,59,79 --

196 min-contig-len 400) [26] and read depth for each contig was determined using BBMap v38.87
197 (bbmap.sh default parameters) [27]. Open reading frames (ORFs) were then identified using
198 Prodigal v2.6.3 (parameters: -p meta) [28] and taxonomy was assigned to each ORF using Kaiju
199 v1.7.4 using the default parameters against a NCBI *nr* database (constructed with: kaiju-makedb
200 -s nr_euk) [29]. HMMER was used to annotate genes against the pfam database [30].
201 Using the read depth for each contig and the estimated taxonomic identity for each ORF,
202 we calculated a depth-adjusted GC content for the bacterial contigs in a metagenome. Similarly,
203 we calculated the depth-adjusted amino acid content for each metagenome. Using known
204 chemical formulas for each amino acid and the depth-adjusted amino acid content, we calculated
205 an estimate of the total C:N ratio of the amino acid pool in each metagenome. For amino acids
206 with fourfold degenerative sites (alanine, glycine, proline, threonine, valine), we calculated the
207 frequency of nucleotides at the third position and averaged these values across the selected
208 amino acids. This provided a metagenome-level estimate of nucleotide frequencies for fourfold
209 degenerative sites.
210 Since our estimates of average genome size for a metagenome was based on the QC
211 filtered reads, non-bacterial reads could potentially bias these estimates. To assess whether the
212 trends we observed were driven by changes in bacterial average genome size, we generated
213 another estimate of genome size based solely on the assembled bacterial contigs. We searched
214 our bacterial pfam gene annotations for 139 single copy genes (based on Campbell et al. 2013
215 [31]) and adjusted for coverage and read depth. We then used the mean number of single copy
216 genes as an estimate of the number of bacterial genomes represented in our metagenomic
217 assembly. We divided the total number of base pairs of our bacterial contigs (adjusted for read
218 depth) by the estimated number of bacterial genomes to generate an estimate of bacterial average

219 genome size. Metagenomes with fewer than 200 million assembled bacterial base pairs where
220 discarded—which provided a similar sample size to MicrobeCensus with default parameters.
221 Specific commands for metagenomic assembly, mapping, and annotation, as well as
222 custom scripts for genetic data processing, can be found at the following Github repository page:
223 https://github.com/PChuckran/Soil_genomic_traits/tree/main/code/upstream_bioinformatics.
224
225 *Traits from 16S rRNA genes*
226 We also estimated average genome size using a community-weighted mean derived from
227 aligning bacterial 16S rRNA gene sequences against a database of genomes of known size. 16S
228 rRNA gene amplicon datasets were downloaded from the NEON data portal. We selected
229 samples from the same sites and collection times as the metagenomes used in this study. To build
230 a reference database, in August 2021 we downloaded all available bacterial 16S gene sequences
231 and associated metadata from the Genome Taxonomy Database (release 202) [32] and created a
232 BLAST [33] database (parameters: makeblastdb -parse_seqids -dbtype nucl) consisting of high
233 quality (>95% complete, <5% redundant) bacterial genomes. Operational taxonomic units
234 (OTUs; clustered at 97%) identified by NEON were then aligned to our high-quality database
235 using BLAST [33] (parameters: blastn -max_target_seqs 1). The best alignment (at >99%
236 identity) was used as the assigned taxonomy. The genome size of each assigned OTU was then
237 adjusted for read depth and used to calculate a community-weighted mean genome size with
238 mean coverage (read depth) as the weight factor. This provided us with a secondary estimate of
239 genome size based only on the bacterial community. Although genome size for a given taxon can
240 vary, we believe this approach still detects general trends in genome size with changes in
241 community composition. Importantly, the source of error in this estimate is different from the

242 metagenome-derived estimate (i.e. unknown intraspecific variations as opposed to the presence
243 of non-bacterial genomes). We also used the GC content of the 16S rRNA gene itself to generate
244 an additional estimate of GC content. Similar trends of genomic traits with edaphic
245 characteristics would provide independent support that these relationships are not due to errors in
246 the methods used for estimation.

247 *Analysis with environmental characteristics*

248 From the NEON data portal and meteorological data, we gathered 205 environmental parameters
249 associated with each site and date of collection. Data were accessed between January and April
250 2021. Where possible, edaphic characteristics were paired with metagenomes from the same soil
251 sample. Otherwise, metagenomes were paired with data from plot or site level averages.
252 Microbial PLFA biomarkers were used to calculate fungal to bacterial biomass ratios [34, 35].
253 Standard precipitation and standard precipitation evapotranspiration indices (SPI and SPEI) were
254 gathered from GRIDMET [36] using Google Earth Engine [37]. Linear regression analysis was
255 used to determine the relationship between soil carbon and genomic traits. All models were
256 constructed in R (v 3.6.1) [38].

257 Random forest models were constructed using the Tidymodels [39] and ranger [40]
258 packages in R, with the objective of determining the environmental parameters which most
259 strongly influenced genomic traits. A non-parametric, machine learning/random forest regression
260 model approach was selected given its proven ability to handle non-linear, complex interactions
261 in space and time across multiple predictor variables [41]. Prior to tuning the random forest
262 hyperparameters, predictors which did not provide data for at least two-thirds of metagenomes
263 were dropped, and the remaining variables were scaled and centered prior to analysis. The full
264 dataset contained several environmental measurements which were either very similar or

functionally identical metrics collected at multiple levels (e.g., site, plot, and soil core). Redundant predictors were removed by first running a random forest which included all available predictors, which were then ranked by variable importance. We calculated Pearson's correlation coefficients between all predictors and then removed those which were highly correlated ($R^2 > 0.8$) with a predictor which explained more variability. We then ran a random forest model which included only remaining predictors.

The strongest predictor of genomic traits, as indicated by random forest models, was then fit to the full dataset. It is important to note that the random sampling associated with this approach means that the best-fit relationship indicated by the random forest, although potentially more accurate as to the true nature of the relationship, may not represent the best-fit for the entire dataset. The code for custom bioinformatics analysis, downstream statistical analysis, and refined data is publicly available at https://github.com/PChuckran/Soil_genomic_traits.

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.

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.

REFERENCES:

1. Shenhav L, Zeevi D. Resource conservation manifests in the genetic code. *Science (80-)* 2020; **370**: 683–687.
2. Batut B, Knibbe C, Marais G, Daubin V. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat Rev Microbiol* 2014; **12**: 841–850.
3. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. *ISME J* 2014; **8**: 1553–1565.
4. Chen W-H, Lu G, Bork P, Hu S, Lercher MJ. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat Commun* 2016; **7**: 11334.
5. Rocha C, Danchin A. Base composition bias might result from competition for metabolic resources. *TRENDS Genet* 2002; **18**: 291–294.
6. Bragg JG, Hyder CL. Nitrogen versus carbon use in prokaryotic genomes and proteomes. *Proc R Soc B Biol Sci* 2004; **271**: 374–377.
7. 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**: 1367–1373.
8. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, et al. Genetics: Genome streamlining in a cosmopolitan oceanic bacterium. *Science (80-)* 2005; **309**: 1242–1245.
9. Chuckran PF, Hungate BA, Schwartz E, Dijkstra P. Variation in genomic traits of microbial communities among ecosystems. *FEMS Microbes* 2022; **2**: 20.
10. Soong JL, Fuchslueger L, Marañon-Jimenez S, Torn MS, Janssens IA, Penuelas J, et al. Microbial carbon limitation: The need for integrating microorganisms into our

understanding of ecosystem carbon cycling. *Glob Chang Biol* 2020; **26**: 1953–1961.

11. Hobbie JE, Hobbie EA. Microbes in nature are limited by carbon and energy: the starving-survival lifestyle in soil and consequences for estimating microbial rates. *Front Microbiol* 2013; **4**: 324.

12. Hellweger FL, Huang Y, Luo H. Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model. *ISME J* 2018; **12**: 1180–1187.

13. Chen Y, Neilson JW, Kushwaha P, Maier RM, Barberán A. Life-history strategies of soil microbial communities in an arid ecosystem. *ISME J* 2021; **15**: 649–657.

14. National Ecological Observatory Network (NEON). Soil microbe metagenome sequences (DP1.10107.001). 2021.

15. Sabath N, Ferrada E, Barve A, Wagner A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol Evol* 2013; **5**: 966–977.

16. Sorensen JW, Dunivin TK, Tobin TC, Shade A. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nat Microbiol* . 2019. Nature Publishing Group. , **4**: 55–61

17. Gravuer K, Eskelinen A. Nutrient and rainfall additions shift phylogenetically estimated traits of soil microbial communities. *Front Microbiol* 2017; **8**: 1271.

18. Malik AA, Puissant J, Buckeridge KM, Goodall T, Jehmlich N, Chowdhury S, et al. Land use driven change in soil pH affects microbial carbon cycling processes. *Nat Commun* 2018; **9**.

19. Slessarev EW, Lin Y, Bingham NL, Johnson JE, Dai Y, Schimel JP, et al. Water balance

creates a threshold in soil pH at the global scale. *Nature* 2016; **540**: 567–569.

20. Barberán A, Ramirez KS, Leff JW, Bradford MA, Wall DH, Fierer N. Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecol Lett* 2014; **17**: 794–802.

21. Simonsen AK. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *ISME J* 2022; **16**: 423–434.

22. Fierer N. Embracing the unknown: Disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* 2017; **15**: 579–590.

23. Hinckley ELS, Bonan GB, Bowen GJ, Colman BP, Duffy PA, Goodale CL, et al. The soil and plant biogeochemistry sampling design for The National Ecological Observatory Network. *Ecosphere* 2016; **7**: e01234.

24. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**: 2114–2120.

25. Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 2015; **16**: 51.

26. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015; **31**: 1674–1676.

27. Bushnell B. BBTools Software Package. https://sourceforge.net/projects/bbmap/. .

28. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**: 1–11.

29. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. 2016.

30. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. *Nucleic Acids Res* 2019; **47**: D427–D432.

31. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, et al. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci U S A* 2013; **110**: 5540–5545.

32. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022; **50**: D785–D794.

33. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. *BMC Bioinformatics* 2009; **10**: 1–9.

34. Willers C, Jansen van Rensburg PJ, Claassens S. Phospholipid fatty acid profiling of microbial communities-a review of interpretations and recent applications. *J Appl Microbiol* 2015; **119**: 1207–1218.

35. Orwin KH, Dickie IA, Holdaway R, Wood JR. A comparison of the ability of PLFA and 16S rRNA gene metabarcoding to resolve soil community change and predict ecosystem functions. *Soil Biol Biochem* 2018; **117**: 27–35.

36. Abatzoglou JT. Development of gridded surface meteorological data for ecological applications and modelling. *Int J Climatol* 2013; **33**: 121–131.

37. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens Environ* 2017; **202**:

18–27.

38. Team RC. R: A language and environment for statistical computing. *R Found Stat Comput Vienna, Austria* 2018.

39. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. 2020.

40. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J Stat Softw* 2015; **77**.

41. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.