TITLE:

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

AUTHORS:

Peter F. Chuckran[1]

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

Abstract:

Genomic traits, such as genome size, GC content, codon usage, and amino acid content, shed insight into the evolutionary processes of bacteria and selective forces behind microbial community composition. Nutrient limitation has been shown to reduce bacterial genome size and influence nucleotide composition, yet little research has been conducted in the soil environment, 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 the United States from the National Ecological Observation Network (NEON) and observed the distribution of these traits across numerous environmental gradients. We found that genomic trait averages were most strongly related to pH, which we suggest results in both physiological constraints on growth as well as affects availability of nutrients in soil. Low pH soils had higher carbon to nitrogen ratios (C:N) and tended to have communities with larger genomes and lower GC-content, potentially a result of increased physiological stress and increased metabolic diversity. Conversely, smaller genomes with high GC content were associated with high pH and low soil carbon to nitrogen ratios, indicating potential resource driven selection against carbon-rich AT base pairs. We found that this relationship of nutrient conservation also applied to amino acid stoichiometry, where bacteria in soils with C:N ratios tended to code for amino acids with lower C:N. Together, these relationships point towards fundamental mechanisms which 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 which reduce the cost of reproduction, such as low GC content and smaller genomes [3]. Since the AT base pair has a carbon to nitrogen ratio (C:N) of 10:7 (1.42), whereas a GC base pair has a C:N of 9:8 (1.13), the AT base pair is more advantageous in nitrogen-limited environments. 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 [4] 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 between marine metagenomes, but surprisingly negatively correlated between metagenomes collected from soils [5]. Since the growth of soil bacteria is thought to be more limited by carbon than nitrogen [6, 7], we hypothesized that the distribution of genomic traits in soil bacteria might exhibit unique patterns reflecting carbon limitation: specifically, higher GC content and smaller genomes when carbon availability is low.

To better understand the relationship between genomic traits of soil bacteria and edaphic characteristics, we analyzed 398 metagenomes collected and sequenced by the US-based National Ecological Observation Network (NEON) [8] across a broad geographic scale (Fig. 1A) and analyzed genomic traits alongside a range of environmental and soil properties. For each metagenome, QC filtered reads were assembled into contigs and annotated. Bacterial contigs were then used to generate a community-level estimate of bacterial GC content, codon frequency, and amino acid content. Using the chemical formulas for each amino acid, we calculated the total relative carbon and nitrogen content of amino acids for each metagenome

(see Methods). Average genome size of a community was estimated from the number of single copy genes recovered from the QC filtered metagenomic reads. Since this estimate uses all QC filtered reads and may therefore be biased by the presence of large fungal genomes, we calculated an additional estimate of bacterial genome size using 16S rRNA gene datasets produced from the same soil samples (sequenced separately from the metagenomes). By aligning community taxonomic data against a database of genome sizes from high-quality isolates, we were able to calculate a community weighted mean of genome size for the bacteria in a community. Genomic traits were then paired with NEON environmental data (soil chemistry, climate, etc.), as well as select meteorological data accessed from GRIDMET [9].

We hypothesized that microbial communities in carbon limited 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). Most of the communities with small genome size and high GC content were located in the mountain-west. We further 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). The C:N of the sum of all coded amino acids was positively correlated with soil $C_{extr}$:$N_{extr}$ ($R^2$=0.24, $p < 0.001$, Fig. 1D), and closely tracked metagenome GC content ($R^2$=0.51, $p < 0.001$, Fig. 1E)—reflecting the resource alignment between the stoichiometry of nucleic acids in codons and their corresponding amino acids [10]. 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. 2A). 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. 2B&C).

These results are in line with our original hypothesis wherein lower C:N lead to 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 [11, 12] and pH [13]. 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 which explain the most variance in GC content and 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 changes in the average genome size of a community as determined by single copy genes could be biased by large fungal genomes at low pH, we additionally found a relationship between pH and genome size when genome size was determined by aligning 16S gene sequences with a database of isolates of known size.

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 [14]. 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 [15], 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. 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 [16] and, similarly, we found that soils with lower pH were associated with higher $C_{extr}$:$N_{extr}$ (Fig. 4), as well as larger genomes and lower GC content (Fig. 4). 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 [17] and, as discussed above, influences the pH in soil. We did find a positive relationship between mean annual precipitation and average genomes size (Supplemental Fig. 1), although the relationship was not as strong as that for pH or soil $C_{extr}$:$N_{extr}$. Previous work has shown relationships between precipitation, pH, and genome size [13], 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 [5].

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 which captures multiple environmental parameters, such as soil nutrients, precipitation patterns, and physiological stress. Soil pH is well established driver of community composition in soils [18] and our results emphasize the degree to which pH dictates belowground microbial

life and perhaps influence evolution of soil bacteria. Additionally, we found several trends which suggest that selection pressure in soil bacterial communities might reflect carbon limitation. Whereas marine communities tend to harbor small AT-rich bacteria in response to nitrogen limitation [3], we found that soils low in carbon tended to select for communities where genomes were smaller but high in GC content. These results are derived from community averages and more work must be done to uncover both the mechanisms and taxonomic level at which such changes in genomic traits occur. However, it is evident that the distribution of genomic traits in soil remains distinct from marine systems, and that physiological stress and nutrient demand are likely written into the DNA of soil bacteria.

## METHODS

Data for this project was gathered from the National Ecological Observation Network (NEON)—an observational network collecting ecological data from across the United States, funded by the US National Science Foundation. NEON maintains 81 field sites in number of distinct biomes across the US. Terrestrial sites include a central meteorological station as well as many dispersed plots from which samples are collected [19]. A full description of NEON sites can be found at https://www.neonscience.org/field-sites/about-field-sites.

*Metagenomic traits*

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

Raw reads were QC filtered using Trimmomatic v0.39 (parameters: TruSeq2-PE.fa:2:30:10 LEADING:10 TRAILING:10 SLIDINGWINDOW:10:20 MINLEN:50) [20]. We then used the program MicrobeCensus [21] to determine the average genome size of a microbial community from the QC-filtered reads. MicrobeCensus estimates the total number of single copy genes to create an estimate of average genome size from unassembled reads. Contigs were assembled from QC-filtered reads using MEGAHIT v1.2.9 (parameters: --k-list 21,29,39,59,79 --min-contig-len 400) [22] and read depth for each contig was determined using BBMap v38.87 [23]. Open reading frames (ORFs) were then identified using Prodigal v2.6.3 (parameters: -p meta) [24] and taxonomy was assigned to each ORF using Kaiju v1.7.4 using the default parameters [25].

Using the read depth for each contig and the estimated taxonomic identity for each ORF, we calculated a depth adjusted GC content for the bacterial reads in a metagenome. Similarly, we calculated the depth-adjusted amino acid content for each metagenome. Using known chemical formulas for each amino acid and the depth-adjusted amino acid content, we calculated an estimate of the total C:N of the amino acids in each metagenome. For amino acids with fourfold degenerative sites (alanine, glycine, proline, threonine, valine), we calculated the frequency of nucleotides at the third position and averaged these values across the selected amino acids. This provided a metagenome-level estimate of nucleotide frequencies for fourfold degenerative sites.

Since our estimate of average genome size for a metagenome was based on the QC filtered reads, large fungal genomes could potentially bias our estimate. To assure that trends we observed were not solely driven by the presence of large fungal genomes, we also estimated average genome size using a community-weighted mean derived from aligning bacterial 16S rRNA gene sequences against a database of genomes of known size. In August 2021 we downloaded all available 16S gene sequences and associated metadata from the Genome Taxonomy Database [26] and compiled a database consisting of high quality (>95% complete, <5% redundant) bacterial genomes. We then accessed 16S rRNA gene datasets from the NEON data portal derived from the same plots as our metagenomes. Operational taxonomic units (OTUs; clustered at 97%) identified by NEON were then aligned to our high-quality database using BLAST. The best alignment (at >99% identity) was used as the assigned taxonomy. The genome size of each assigned OTU was then adjusted for read depth and used to calculate a community-weighted mean genome size with mean coverage (read depth) as the weight factor.

*Analysis with environmental characteristics*

From the NEON data portal and meteorological data, we gathered 205 environmental parameters associated with each site and date of collection. A full list of the NEON data products used in this analysis can be found in Supplemental Table 1. Data were accessed between January and April 2021. Where possible, edaphic characteristics were paired with metagenomes from the same soil sample. Otherwise, metagenomes were paired with data from plot or site level averages. Microbial PLFA biomarkers were used to calculate fungal to bacterial biomass ratios [27, 28]. Standard precipitation and standard precipitation evapotranspiration indices (SPI and SPEI) were gathered from GRIDMET [9] using Google Earth Engine [29]. Regression analysis was used to determine the relationship between soil carbon and genomic traits. All models were constructed in R (v 3.6.1) [30].

Random forest models were constructed using the Tidymodels [31] and ranger [32] packages in R, with the objective of determining the environmental parameters which most strongly influenced genomic traits. A non-parametric, machine learning/random forest regression model approach was selected given its proven ability to handle non-linear, complex interactions in space and time across multiple predictor variables [33]. Prior to tuning the random forest hyperparameters, predictors which did not provide data for at least two-thirds of metagenomes were dropped, and the remaining variables were scaled and centered prior to analysis. The full 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.

Upon publication, all code and corresponding data generated through this analysis will be made publicly available through Github.

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, 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. 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.
5. Chuckran PF, Hungate B, Schwartz E, Dijkstra P. Soil, ocean, hot spring, and host-associated environments reveal unique selection 1 pressures on genomic features of bacteria in microbial communities 2 3 AUTHORS: 4. *bioRxiv* 2021; 2021.04.05.438506.
6. 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.
7. 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.
8. National Ecological Observatory Network (NEON). Soil microbe metagenome sequences (DP1.10107.001). 2021.
9. Abatzoglou JT. Development of gridded surface meteorological data for ecological applications and modelling. *Int J Climatol* 2013; **33**: 121–131.

10. Bragg JG, Hyder CL. Nitrogen versus carbon use in prokaryotic genomes and proteomes. *Proc R Soc B Biol Sci* 2004; **271**: 374–377.

11. 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.

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

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

14. 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.

15. 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.

16. 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.

17. Simonsen AK. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *ISME J 2021* 2021; 1–12.

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

19. 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.

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

21. 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.

22. 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.

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

24. 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.

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

26. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, 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* 2021.

27. 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.

28. 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.

29. 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.

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

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

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

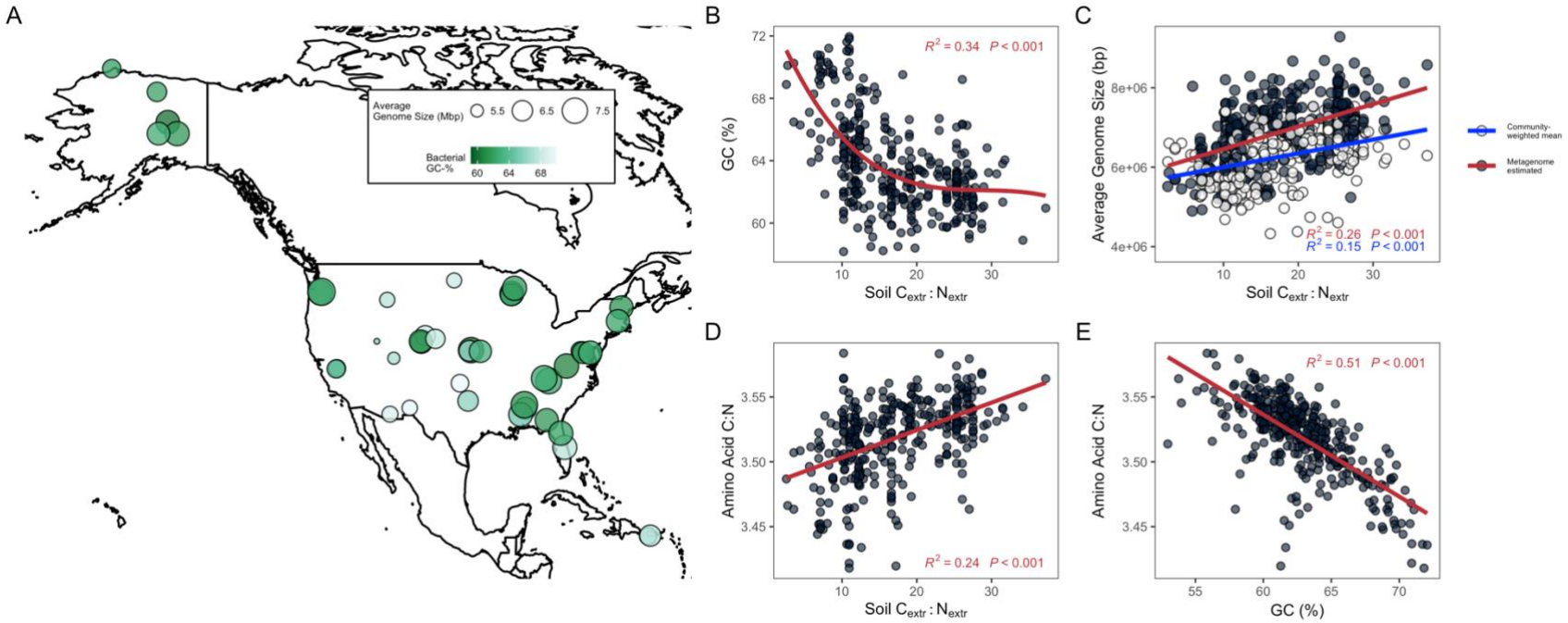
33. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.

FIGURES:

**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 (%). **(C)** Relationship between $C_{extr}$:$N_{extr}$ and genome size, estimated from the number of single copy genes per metagenome (black circles) and from 16S rRNA gene datasets and genome size estimated from isolates (open circles)**. (D)** Relationship between $C_{extr}$:$N_{extr}$ and bacterial amino acid C:N ratios averaged per metagenome. **(E)** Relationship between bacterial GC-% and pooled amino acid C:N ratios**.**

335


A
Average Genome Size (Mbp)
5.5
6.5
7.5
Bacterial GC-%
60
64
68
B
$R^2$ = 0.34 $P$ < 0.001
GC (%)
Soil $C_{extr}$ : $N_{extr}$
C
Average Genome Size (bp)
$R^2$ = 0.26 $P$ < 0.001
$R^2$ = 0.15 $P$ < 0.001
Community-weighted mean
Metagenome estimated
Soil $C_{extr}$ : $N_{extr}$
D
Amino Acid C:N
$R^2$ = 0.24 $P$ < 0.001
Soil $C_{extr}$ : $N_{extr}$
E
Amino Acid C:N
$R^2$ = 0.51 $P$ < 0.001
GC (%)

336

**Figure 2:**

Nucleotide bias in codons derived from open reading frames of bacteria in soil metagenomes; **(A)** 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. **(B)** 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. **(C)** 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 ($p < 0.05$). Codons are arranged left to right in increasing number of GC base pairs.

346

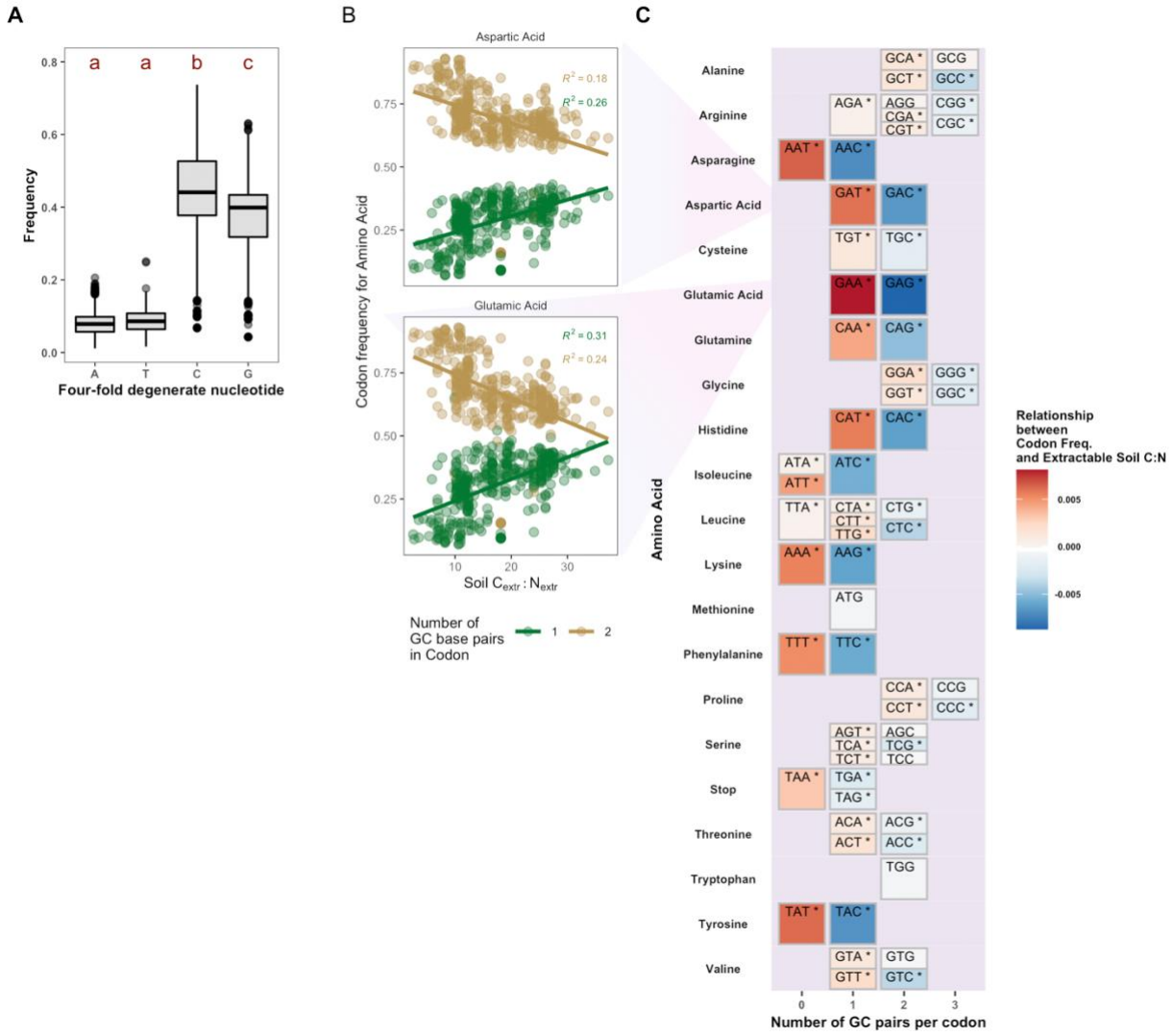

A
a a b c
Frequency
Four-fold degenerate nucleotide
B
Aspartic Acid
Glutamic Acid
Codon frequency for Amino Acid
Soil Cextr : Nextr
Number of GC base pairs in Codon
C
Amino Acid
Number of GC pairs per codon
Relationship between Codon Freq. and Extractable Soil C:N

**Figure 3:**

Results from the random forest model and relationships between soil pH and genomic traits; **(A)** Variable importance plot for the top 10 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. **(C)** The relationship between soil pH and both average genome size (derived from metagenomes) and community-weighted genome size (derived from the identification of 16S rRNA genes).

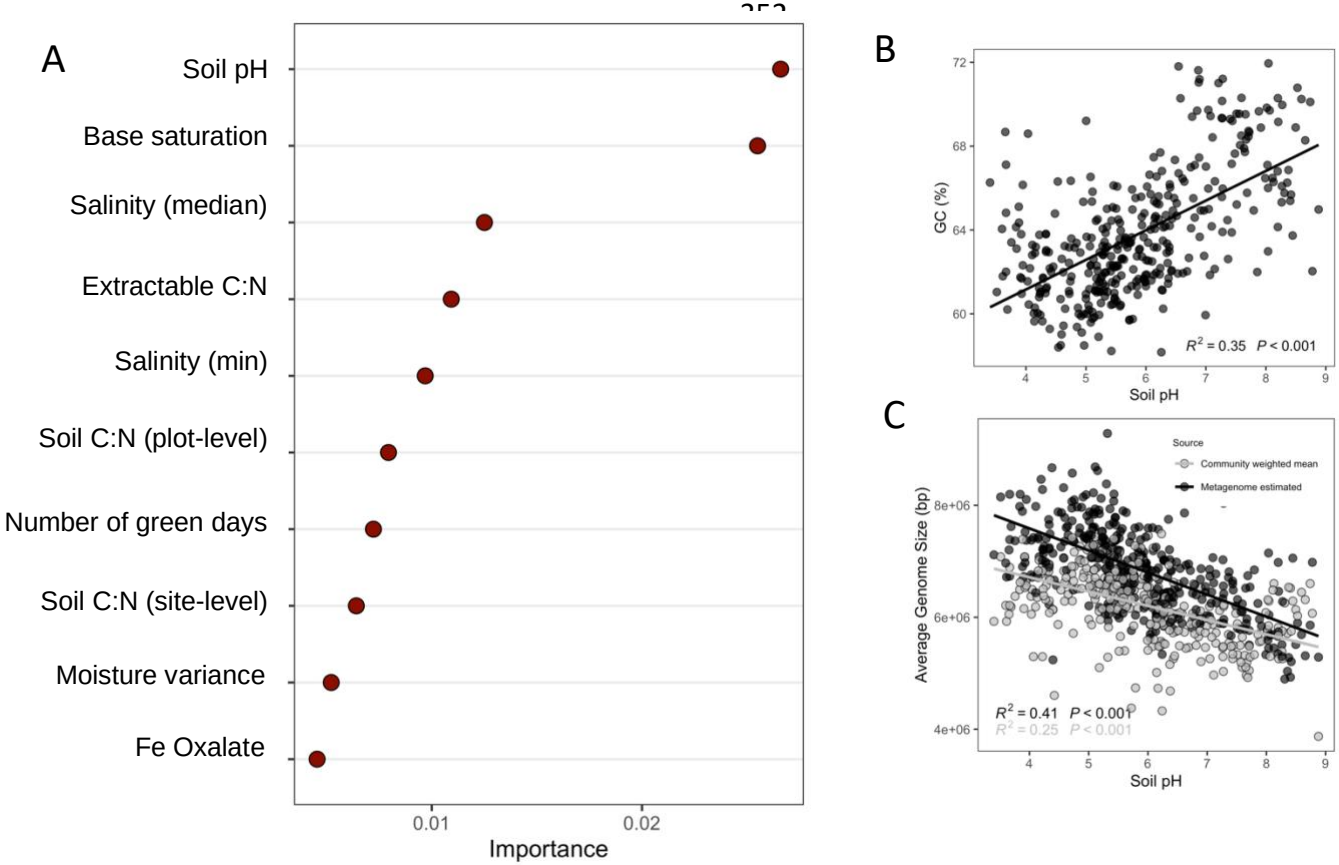
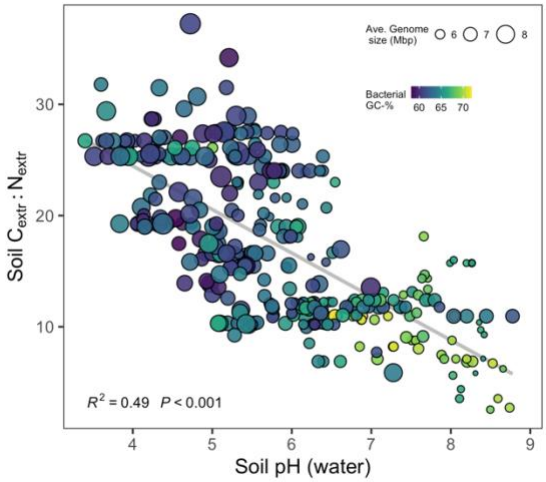


**Figure 4:**

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.