nature microbiology

Matters arising

https://doi.org/10.1038/s41564-024-01687-w

Microbial dark matter could add uncertainties to metagenomic trait estimations

Received: 24 October 2023

Accepted: 25 March 2024

Published online: 13 May 2024

Check for updates

Ernest D. Osburn[1,2], Steven G. McBride[3] & Michael S. Strickland[2]

ARISING FROM: G. Piton et al. *Nat. Microbiol.* https://doi.org/10.1038/s41564-023-01465-0 (2023).

Estimating life history traits of bacterial communities from metagenomes is an increasingly common practice in microbial ecology research[1–4]. Contributing to this area of inquiry, Piton et al.[5] use a global dataset of soil metagenomes to describe life history trait dimensions of soil bacteria. Our analysis of the same dataset suggests that the genome size analysis of Piton et al. may underestimate potential biases introduced by varying proportions of non-bacterial DNA among different ecosystem types. We demonstrate how this potential bias could influence the relationships of this trait with environmental variables, thus altering the interpretation of the identified trait dimensions.

Some of the life history traits of Piton et al., for example, average genome size (AGS), were estimated using the full metagenomes from each sample, with the reasoning that the metagenomes mostly represent bacteria as <2% of reads were annotated as eukaryotic[5]. We agree that only a small proportion of the sequences can be annotated as eukaryotic, but this does not imply that eukaryotic sequences are negligible or that the proportion of non-bacterial DNA is constant among ecosystems. As is the case in many (if not all) soil metagenome studies, the majority of sequences in this dataset cannot be classified at all (Extended Data Fig. 1a), and only 20–50% of the sequences from the metagenomes could be identified as bacterial[6] (Fig. 1a). It is likely that much of this unclassified DNA with unknown function, that is, microbial 'dark matter'[7], is, in fact, bacterial in origin, but some unknown proportion will also be eukaryotic or viral DNA. The presence of non-bacterial DNA is perhaps a minor concern if it is present in similar proportions across samples. This assumption may be valid in studies within a particular ecosystem but may be problematic in a global multi-ecosystem study such as this one. Indeed, we found systematic variation in the percentage of reads classified as bacteria versus eukaryotic, as well as the percentage of unclassified reads, among different ecosystems with different soil pH (Fig. 1 and Extended Data Fig. 1a). Specifically, low-pH soils (for example, forest soils) had higher lower proportions of reads classified as bacterial and higher proportions of reads classified as eukaryotic (Fig. 1 and Extended Data Fig. 1b). This is not surprising, as acidic soils are known to host larger biomass of microbial eukaryotes, for example, fungi[8]. The greater relative abundance of eukaryotic reads in the acidic soils would also explain the greater proportions of unclassified reads in those soils given the disproportionately poor annotation of eukaryotes in metagenomic analyses[9–11]. Overall, while we do not know the exact proportion of non-bacterial DNA in the metagenomes, we suspect that it is higher than has been reported by the authors and, critically, that the non-bacterial proportion varies among ecosystem types. It should be noted, however, that our method of classifying reads likely has some bias, for example, against particular bacterial taxa, life history groups, genome regions and/or ecosystem types, although the degree of bias is difficult to assess. The taxonomic annotation method appears to be generally effective, however, as 76% of the genome equivalents identified in the full metagenomes were recovered in the metagenomic reads identified as bacterial (hereafter, 'bacterial' metagenomes) on average (Extended Data Fig. 1c).

To investigate potential confounding influences of non-bacterial DNA, we repeated some of the authors' analyses using both the full metagenomes and the bacterial metagenomes. Ideally, the results from the full metagenomes would be confirmed by analysing the fraction of the metagenomes that can confidently be identified as bacterial. We focus on AGS, as this was the strongest contributor to the authors' first trait dimension, which, in turn, was strongly related to soil pH[5]. In the full metagenomes, AGS ranged from ~5 to 10 Mb and were strongly negatively correlated with pH (Fig. 2a). These AGS values would be very large for community averages and are all larger than the previously reported 3.7 Mb average for terrestrial bacteria[12]. By contrast, the putative bacterial metagenomes had smaller AGS ranging from ~2.5 to 4.2 Mb and were only very weakly correlated with soil pH (Fig. 2b). This discrepancy is not likely due to bias against large genomes in the bacterial metagenomes given that well-characterized isolates that are better represented in sequence databases tend to have larger genomes[12]. A more likely explanation is that the full

[1]Department of Plant and Soil Sciences, University of Kentucky, Lexington, KY, USA. [2]Department of Soil and Water Systems, University of Idaho, Moscow, ID, USA. [3]Department of Biology, Radford University, Radford, VA, USA. e-mail: e.osburn@uky.edu

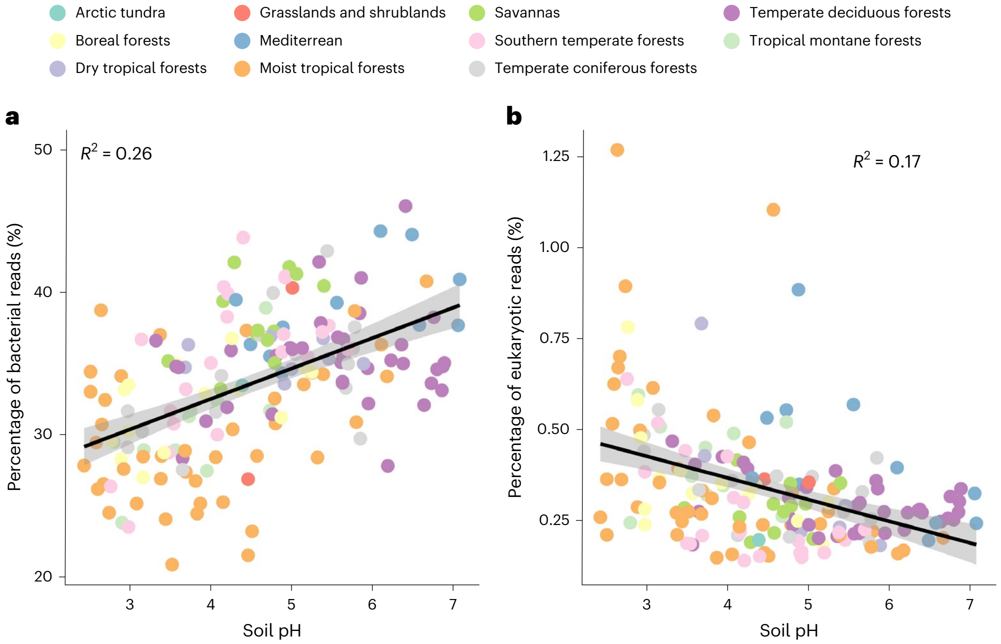


Fig. 1 | Evidence of varying proportions of non-bacterial DNA in soil metagenomes from different ecosystems. a, The percentage of sequence reads identified as bacterial as a function of soil pH. **b**, The percentage sequence reads identified as eukaryotic as a function of soil pH. Classification of sequence reads was performed with kraken2 with the Refseq genomes for bacteria, archaea, viruses, fungi and protists as reference databases. On all panels, R^2 values and best-fit lines are from linear regression, and for all models, $P < 0.001$.

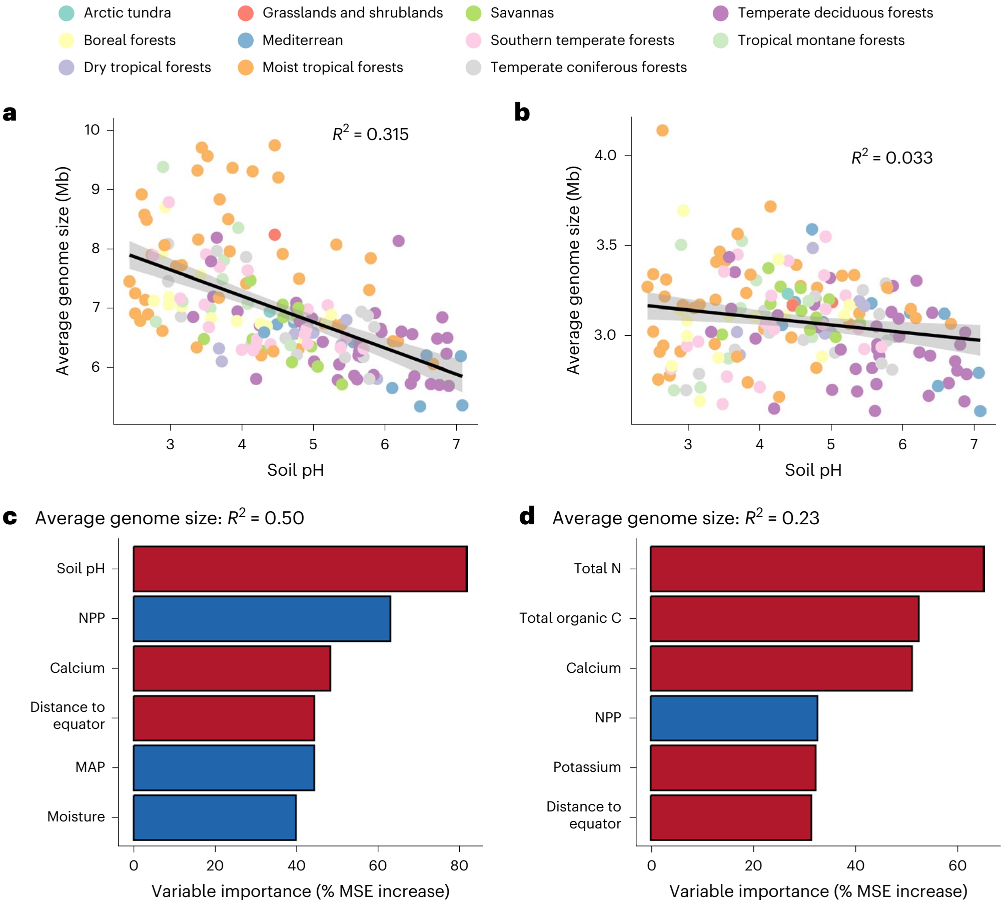


Fig. 2 | Evidence that varying proportions of non-bacterial DNA influences estimates of bacterial community life history and relationships with environmental variables. a,**b** Community-averaged genome sizes for each metagenome in the full metagenomes (**a**) and the putative bacterial metagenomes (**b**) as a function of soil pH. For **a** and **b**, R^2 values and best-fit lines are from linear regression. **c**,**d**, The most important environmental predictors (from random forest regression) for AGSs in the full metagenomes (**c**) and the bacterial metagenomes (**d**). For **c** and **d**, blue bars indicate positive effects, while red bars indicate negative effects. For **c** and **d**, variable importance was quantified by determining the increase in model error (mean square error (MSE)) after randomly shuffling each candidate predictor across the dataset. NPP, net primary productivity; MAP, mean annual precipitation.

metagenomes are contaminated with non-bacterial sequences to varying degrees, which has been previously shown to inflate calculated AGS[13]. Supporting this explanation, we found that AGS in the full metagenomes were strongly negatively correlated with the percentage of reads classified as bacteria (Extended Data Fig. 2a) and positively correlated with the ratio of eukaryotic to bacterial reads (Extended Data Fig. 2b). Therefore, we suggest that the strong association of very large AGS with acidic soils in the full metagenomes is likely an artefact of ecosystems with acidic soils having larger proportions of non-bacterial DNA.

We were also interested in determining whether biases from non-bacterial DNA influenced the identification of environmental drivers of AGS. Not surprisingly, for the full metagenomes, soil pH was identified as the dominant driver of AGS (Fig. 2c). This, again, is likely an artefact of acidic soils simply having larger proportions of non-bacterial DNA. By contrast, in the putative bacterial metagenomes, soil C and N content emerge as the most important drivers (Fig. 2d). Other studies have also observed reduced bacterial AGS and weakened relationships with soil pH when analysing full versus bacterial metagenomes[1,4]. Our conclusion from these results is that inferences regarding bacterial AGS derived from full metagenomes can be biased by varying proportions of non-bacterial DNA among different ecosystem types. By contrast, other bacterial genomic traits (for example, 16S ribosomal RNA gene copy number, GC content) only showed minor differences between the full and bacterial metagenomes (Extended Data Fig. 3), probably because those calculations are not dependent upon the total number of sequences present. While AGS was only one of many community traits used in the analysis of Piton et al., it was a key trait in delineating the trait dimensions in their analysis—our results suggest that this conclusion should be reconsidered.

Our analyses show the pitfalls of inferring bacterial community traits from full metagenomes that have varying domain-level taxonomic composition and/or varying proportions of unclassified dark matter DNA. Our findings have broad relevance beyond their implications for Piton et al., as inference of bacterial community traits from full soil metagenomes is currently commonplace[1–5]. While our alternative methods are also imperfect, our results do provide evidence that the non-bacterial proportion of metagenomes can bias estimates of bacterial community life history traits and their relationships with environmental variables. We expect that the biases present metagenomic analyses will become easier to identify and account for over time as reference databases improve. In the meantime, we suggest that metagenomic analyses should 'stress-test' their results by applying multiple analytical approaches to their datasets. An additional stress-test approach to complement the method we used here would be the removal of annotated eukaryotic reads or contigs before analysis[4]. Observed patterns can also be validated by analysing additional datasets or by performing experimental manipulations to robustly establish relationships between environmental change and metagenomic traits. By following these recommendations, influences of confounding factors and subsequent misinterpretations of data can be minimized.

Methods

Raw metagenomic sequence reads were downloaded from National Center for Biotechnology Information accession number PRJEB18701. Raw reads were quality filtered using trimmomatic[14] and taxonomy assigned to the quality-filtered reads using kraken2[6] with the Refseq genomes for bacteria, archaea, viruses, protists and fungi as reference databases. We then generated the putative bacterial metagenomes by extracting the sequence reads identified as bacterial in origin (taxid '2'). For both the full metagenomes and the bacterial metagenomes, we quantified the number of genome equivalents and the AGS using MicrobeCensus (version 1.1.1)[13]. Full details of our methods can be found in the Supplementary Information.

References

1. Chuckran, P. F. et al. Edaphic controls on genome size and GC content of bacteria in soil microbial communities. *Soil Biol. Biochem.* **178**, 108935 (2023).
2. Chen, Y., Neilson, J. W., Kushwaha, P., Maier, R. M. & Barberán, A. Life-history strategies of soil microbial communities in an arid ecosystem. *ISME J* **15**, 649–657 (2021).
3. Fierer, N., Barberán, A. & Laughlin, D. C. Seeing the forest for the genes: using metagenomics to infer the aggregated traits of microbial communities. *Front. Microbiol.* **5**, 614 (2014).
4. Wang, C. et al. Bacterial genome size and gene functional diversity negatively correlate with taxonomic diversity along a pH gradient. *Nat. Commun.* **14**, 7437 (2023).
5. Piton, G. et al. Life history strategies of soil bacterial communities across global terrestrial biomes. *Nat. Microbiol.*, https://doi.org/10.1038/s41564-023-01465-0 (2023).
6. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
7. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
8. Strickland, M. S. & Rousk, J. Considering fungal:bacterial dominance in soils – methods, controls, and ecosystem implications. *Soil Biol. Biochem.* **42**, 1385–1395 (2010).
9. Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48 (2020).
10. Pronk, L. J. U. & Medema, M. H. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microb. Genom.* **8**, mgen000823 (2022).
11. Lind, A. L. & Pollard, K. S. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome* **9**, 58 (2021).
12. Rodríguez-Gijón, A. et al. A genomic perspective across earth's microbiomes reveals that genome size in archaea and bacteria is linked to ecosystem type and trophic strategy. *Front. Microbiol.* **12**, 761869 (2022).
13. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* **16**, 51 (2015).
14. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

Author contributions

E.D.O and S.G.M. conceived the study. E.D.O conducted the bioinformatic and statistical analyses. M.S.S. supervised the project. All authors contributed to the writing and editing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

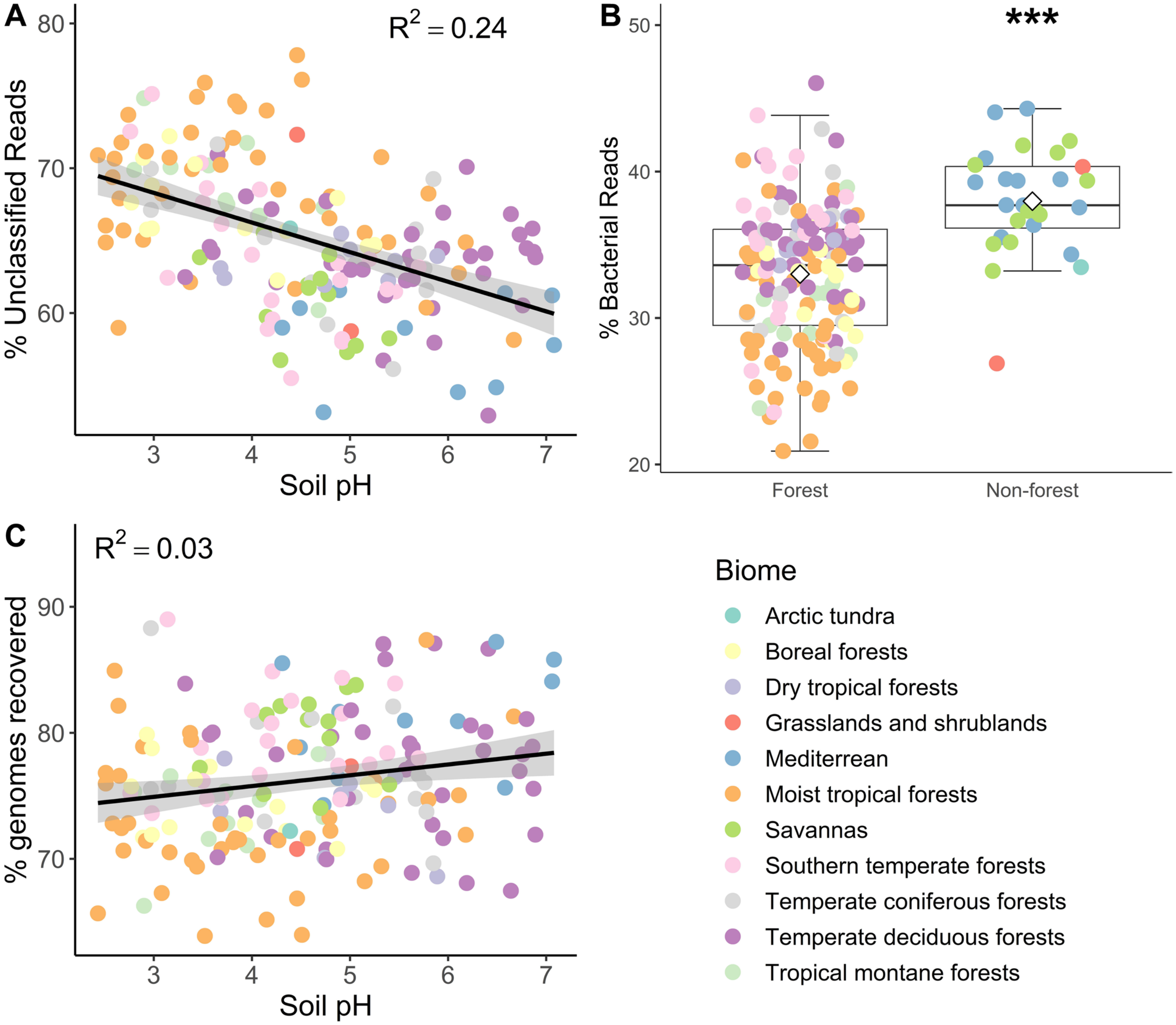
Extended data is available for this paper at https://doi.org/10.1038/s41564-024-01687-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-024-01687-w.

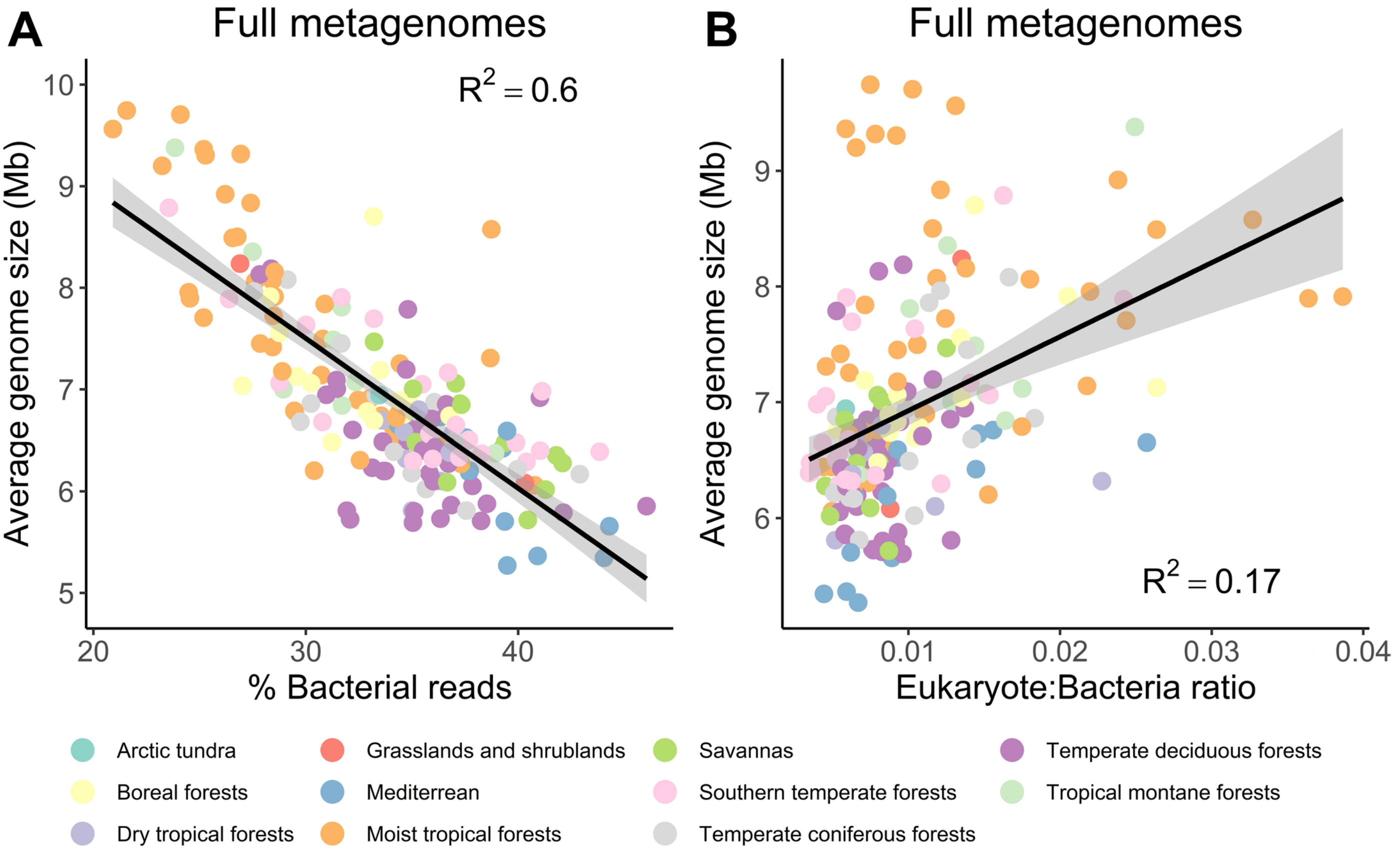
Correspondence and requests for materials should be addressed to Ernest D. Osburn.

Peer review information *Nature Microbiology* thanks Kate Buckeridge and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

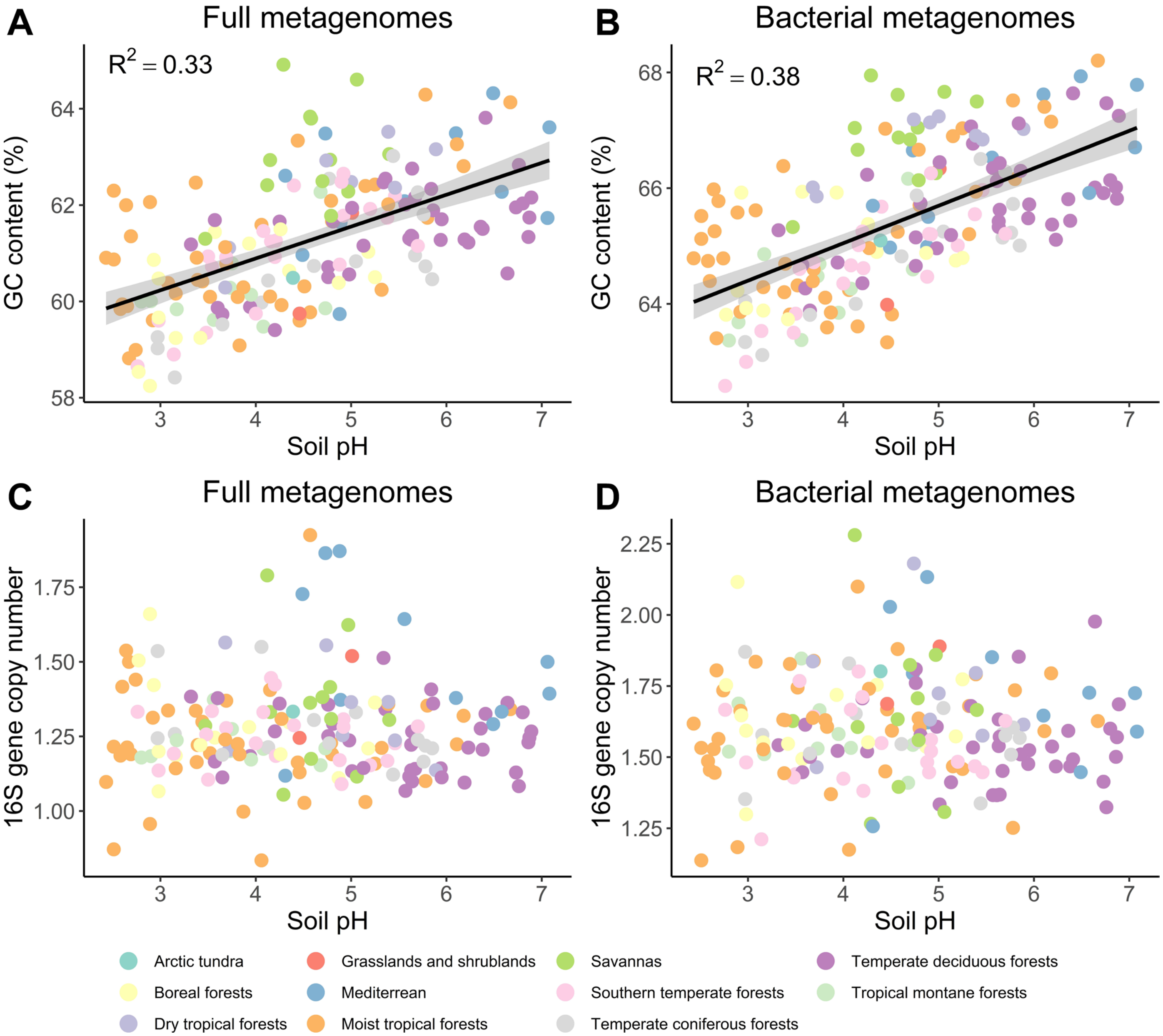
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Percent of metagenomic reads not classified to any taxon (**a**), percent of reads classified as bacteria between forested and non-forested ecosystems (**b**), and percentage of genome equivalents recovered in the 'bacterial' metagenomes that were found in the full metagenomes (**c**). Classification was done using kraken2 with RefSeq genomes for bacteria, archaea, viruses, fungi, and protists as reference databases. Genomes within the metagenomes were quantified by determining mean coverage of 30 single-copy genes using MicrobeCensus. The R^2 value and line of best fit in (**a**) and (**c**) are from linear regression. Asterisks in (**b**) indicate significantly higher percentage of reads classified as bacterial in non-forested environments ($p < 0.001$).



Extended Data Fig. 2 | Average genome size in the full metagenomes as a function of the percentage of reads classified as bacteria (**a**) and the ratio of eukaryotic to bacterial reads (**b**). R^2 values and lines of best fit are from linear regression (both $p < 0.001$). Average genome sizes were determined using MicrobeCensus.



Extended Data Fig. 3 | Metagenome GC content in the full (**a**) and bacterial (**b**) metagenomes and average 16S rRNA gene copy number in the full (**c**) and bacterial (**d**) metagenomes as a function of soil pH. R^2 values and best-fit lines on (**a**) and (**b**) are from linear regression (both $p < 0.001$). Regression models for 16S gene copy number in (**c**) and (**d**) were not significant (both $p > 0.05$).