Title

Reply to: Microbial dark matter could add uncertainties to metagenomic trait estimations

Authors

Piton, Gabin

Allison, Steven D

Bahram, Mohammad

et al.

Publication Date

2024-05-13

DOI

10.1038/s41564-024-01688-9

Reply to: Microbial dark matter could add uncertainties to metagenomics trait estimations

Gabin Piton[1,2,*], Steven D Allison[1,3], Mohammad Bahram[4,5], Falk Hildebrand[6,7], Jennifer BH Martiny[3], Kathleen K Treseder[3], Adam C Martiny[1,3]

1. Department of Earth System Science, University of California, Irvine, California, USA
2. Eco&Sols, INRAE-IRD-CIRAD-SupAgro, University Montpellier, Montpellier, France
3. Department of Ecology and Evolutionary Biology, University of California, Irvine, California, USA
4. Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden
5. Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia
6. Gut Microbes & Health, Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk NR4 7UA, UK
7. Digital Biology, Earlham Institute, Norwich Research Park, Norwich, Norfolk NR4 7UA, UK.

* Corresponding author: gabin.piton@inrae.fr

In a recent paper, we used soil metagenomes from global biomes to assess life history strategies across soil bacterial communities[1]. We compiled 102 traits in a multi-table coinertia analysis (MCOA) to characterize the life history strategies of soil bacteria. Although we only used annotated reads for calculation of most traits, the full metagenomes were used to quantify bacterial average genome size (AGS) and rRNA operon copy numbers. In their Matter Arising, Osburn and co-workers point out a potential bias in our method for estimating bacterial-AGS is biased. We acknowledge that estimating bacterial-AGS for an environment as complex and unexplored as soil is non-trivial. Although we agree that our approach has a bias towards bacterial-AGS overestimation, we show that the alternative calculation proposed by Osburn et al. appears to have an even stronger

underestimation bias. Still, neither approach led to a different conclusion for our proposed life history strategies, demonstrating the robustness of our results.

In their comment, Osburn et al. argue that bacterial-AGS should be calculated only with metagenomic reads assigned to bacteria, as opposed to using all metagenomic reads. While the proportion of metagenomic reads classified as eukaryotic is usually small (< 2%), systematic covariance with ecosystem variables such as pH could influence calculations. Indeed, as soil eukaryotes (mainly fungi in soil) usually have larger genomes[2,3], using the full metagenome can overestimate bacterial-AGS proportionally to the relative abundance of eukaryotic sequences in the metagenome.

However, using only annotated sequences is also problematic. Fifty to 80% of metagenomic reads from soils are typically not identified with current databases (i.e. the functional dark matter[4]) and most of these are probably of bacterial origin[4]. Indeed, MicrobeCensus, the tool used to estimate AGS with metagenomes[4] relies on the principle that the AGS in a community is inversely proportional to the relative abundance of single copy core genes (SCCG), ie. the ratio (SCCG base pair / Total base pair). Well-characterized and evolutionarily conserved SCCGs (that is, genes that are easy to identify based on homology) can be over-represented in the pool of annotated sequences. Conversely, genes outside the core pangenome are more likely to be unknown in reference databases (because they are absent from most genomes). Thus, relying solely on assigned metagenomic reads will likely lead to an overestimation of SCCGs, while the total number bacterial reads is underestimated, leading overall to an AGS underestimation[5]. That bias should be also considered along with the bias of the full-metagenome calculation that we used in our original study (Figure 1).

In our study, we found a strong negative relationship between soil pH and bacterial-AGS using the full metagenome method. Osburn et al. illustrate the consequence of the bias due to the full-metagenome method for this relationship. Using their estimate, they showed a weakened relationship between pH and genome size and that the proportion of eukaryotic sequence increased at low pH.

They thus argued that this pattern is not a real ecological pattern, but is instead 'an artefact of ecosystems with acidic soils having larger proportions of non-bacterial DNA'.

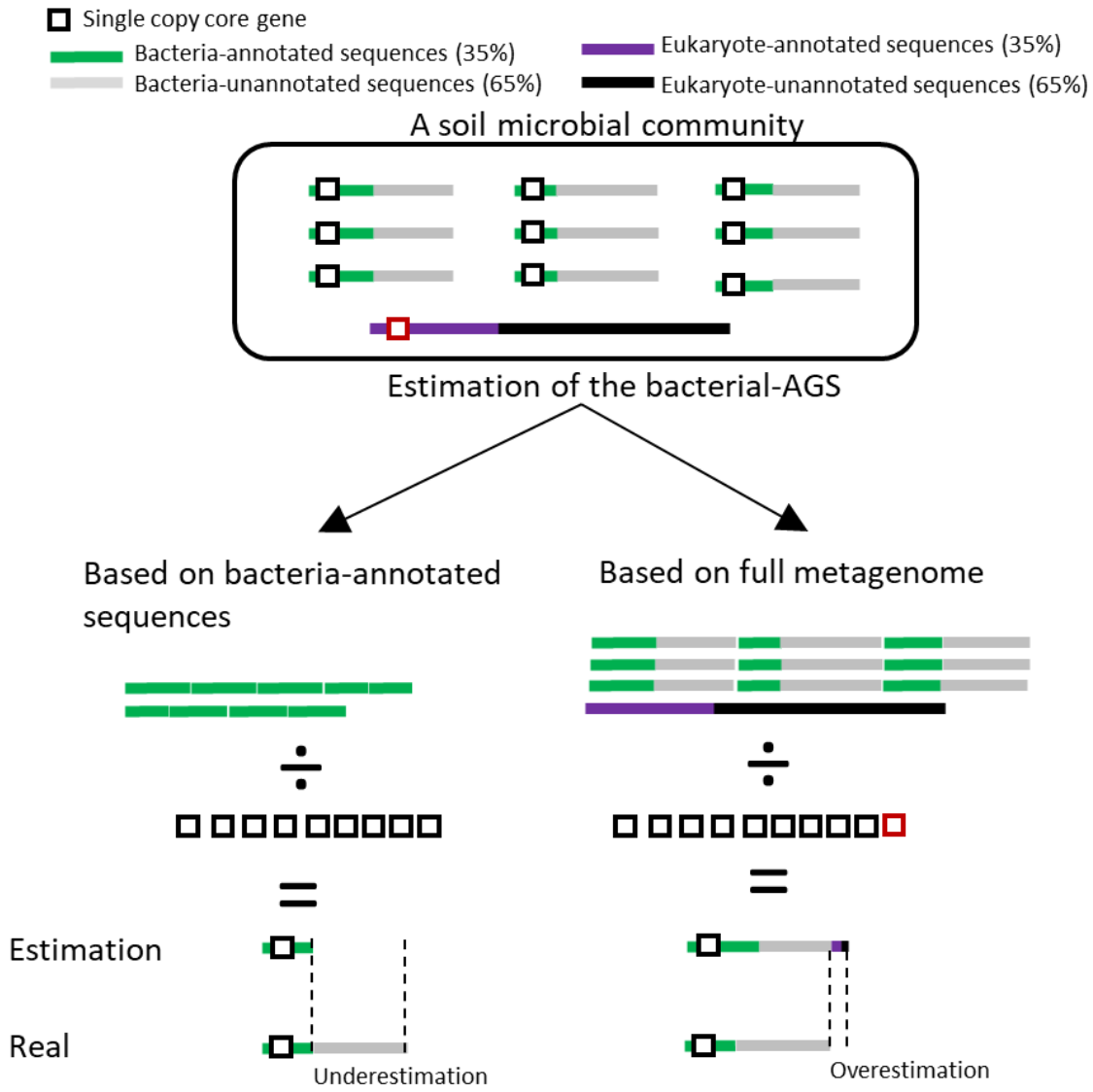


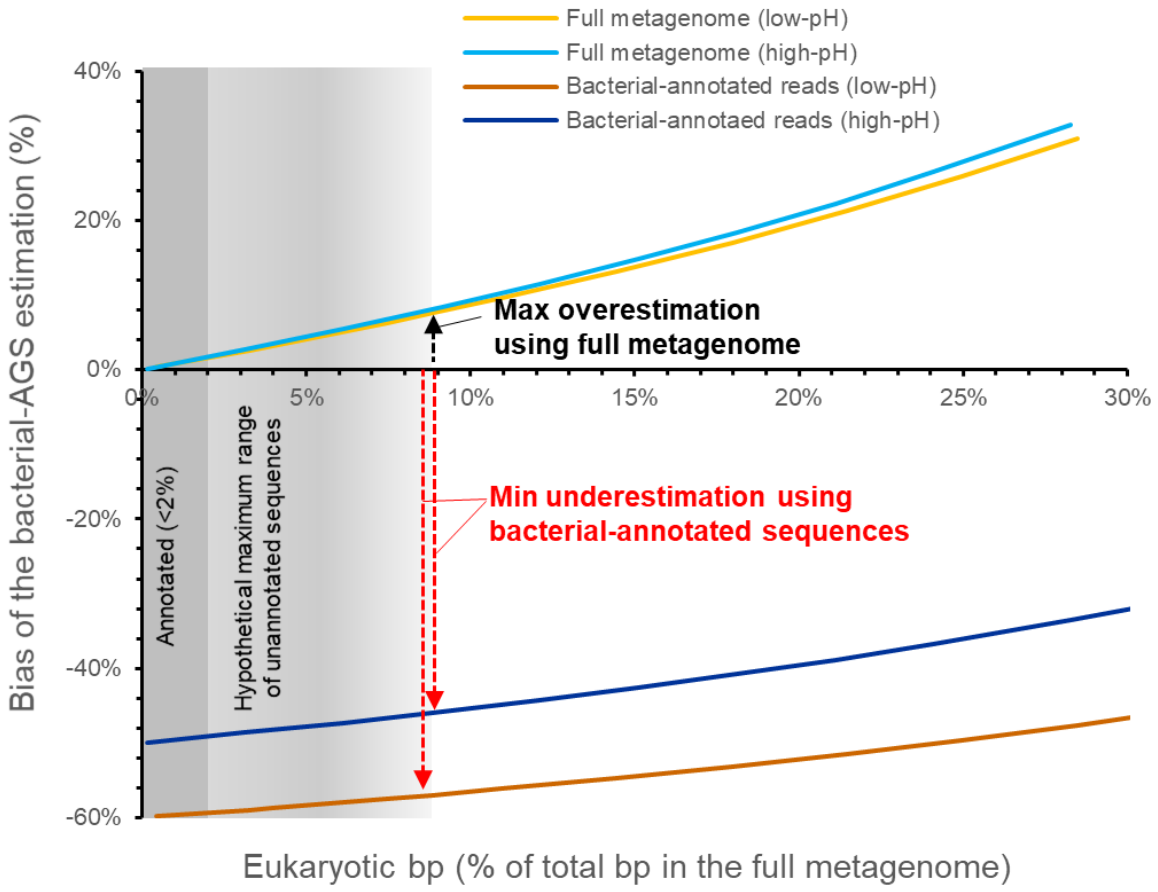
Figure 1. Average genome size (AVG) calculation using metagenome. Simplified representation of the process of bacterial-AGS estimation with MicrobeCensus if only bacteria-annotated sequences are used versus the full metagenome. The values of 35% and 65% for annotated and unannotated sequences respectively were chosen for this illustration as they were the average percentages in our original study. One Eukaryotic genome out of 10 genomes is represented for clarity, but eukaryotes

represent only about 1% of the genomes in a metagenome, leading to lower positive bias than represented in the figure.

The difference in the AGS-pH relationship observed with the two methods might also be explained by the bias of their estimation using bacterial-annotated metagenomes. Indeed, unidentified sequences increase from 60 to 70% at low pH (Extended Data Figure 1 of their Matters Arising). Such an increase of the proportion of unannotated reads in metagenomes of low-pH soil very likely includes a dominant portion of bacterial sequences. Disregarded in their calculation, this bacterial dark matter would accentuate the underestimation of their approach, which would weaken a real negative relationship with pH. Thus, several mechanisms can influence the quantification of average genome size, but we currently lack direct measurements to fully validate estimates of this key trait across natural environments.

Simulation of metagenome composition can help in perceiving what conditions would underlie different biases in bacterial-AGS quantification. We simulated how eukaryotic sequences might bias the estimation of bacterial-AGS (Supplementary Note 1). This simulation (Figure 2) shows that the bacterial-AGS estimated with bacteria-annotated sequences would be accurate and the one using the full-metagenome fully artificial if 65% of the sequences were eukaryotic in low-pH soil (93.5% of the unannotated sequences). However, given that eukaryotic sequences represent less than 2% of annotated sequences, we might also expect only a small fraction of eukaryotic sequences in the unannotated base pairs[4]. Indeed, assuming an extreme range of 4 to 9% eukaryotic base pairs (2% annotated and 2-7% unannotated, Supplementary Note 2) would lead to an overestimation of bacterial-AGS by +3.2 to +7.5% in low pH soil using the full metagenome method (Real value=7.4-7.8Mb, Estimated value=8Mb, Figure 2). In contrast, using only bacteria-annotated sequences would lead to an underestimate of genome size between -57 and -59% (Estimated value=3.2Mb, Figure 2). In such case, it thus appears that the assumptions associated with Osburn et al. method would strongly

bias the results. Finally, the observed negative trend between pH and bacterial-AGS would be an artifact if the metagenome of low-pH soil would be composed of ~30% more eukaryotic base pairs



than metagenome of high-pH soil, whereas only 0.3% more was observed (Figure 1 of the Matters Arising). Thus, we conclude that there is support for the negative link between bacterial genome size and soil pH. Supporting this relationship, Wang et al.[6] recently found the same negative relationship using the same metagenome dataset. However, they calculated bacterial-AGS using the putative genome size of taxa from the Genome Taxonomy Database.

Figure 2. Biases of bacterial-AGS estimation methods. Simulated effect of increasing % of eukaryotic base pairs on the biases of the bacterial-AGS estimations using the full metagenome or only bacterial annotated reads, with full-metagenome AGS (average across bacteria and eukaryotes) set at 8Mb and 6Mb for low-pH and high-pH soils respectively (values of the original studies). See

Supplementary Note 1 for equations and assumptions used in this simulation. The gray zone represents the maximum range of eukaryotic sequences expected (Supplementary note 2).

Osburn et al. also argue that their estimate fits into the range of soil bacteria based on results from soil bacteria metagenome-assembled-genomes (MAG)[7]. However, this range is also likely biased as MAGs are more easily obtained for small genomes[8]. Moreover, Madin et al.[3] report from ~2,000 soil isolates a range of genome sizes (median, minimum and maximum values of 6.41, 1.26 and 16 Mb, respectively) that covers both of our estimates' ranges (median, minimum and maximum values of 6.8, 5.2 and 10.3 Mb, respectively, for Piton et al.[1] and 3.07, 2.58 and 4.14 Mb, respectively, for Osburn et al.[2]).

Finally, we quantified the degree to which our conclusions are affected by the different methodologies for estimating genome size. Using the bacterial-AGS from both methods, we found that the two trait dimensions used to characterize life history strategies[1] remained very much the same (Supplementary Figure 1 and 2). Both AGS estimates were associated with both dimensions; however, the estimation from bacterial-annotated sequences suggested a weaker association with the first dimension and a tighter one with the second dimension than predicted based on the full-metagenome method alone. In other words, estimation from bacteria-annotated sequences also supports that bacterial AGS captures an extension of metabolic capacities (dimension 1) and emphasizes that the AGS becomes especially large when this extension is oriented towards nutrient recycling capacities (dimension 2)—a profile that we associate with the competitor strategy[1].

This Matter Arising of Osburn et al. stress an important point: observed patterns of bacterial-AGS can be biased by co-commitment changes in the dark matter composition, and accurately quantifying these biases will remain difficult without better annotation of soil metagenomes. Methodological improvements, development of genomic databases, annotation tools and reference-free approaches will lead to better estimation of bacterial-AGS. For instance, removing eukaryotic, viral and archaeal sequences before calculating the bacterial-AGS using all the remaining sequences

(bacterial and unidentified) is one possibility. Reference free methods (eg.[4]) also represent promising approaches to investigate the taxonomic composition of the metagenomic dark matter and account for it in bacterial-AGS estimation. However, a perfect estimation of bacterial-AGS using soil metagenomes is not yet possible. We agree with Osburn et al that comparing different estimates and understanding their biases is the best strategy to investigate patterns of bacterial-AGS. Although our results appear robust to varied methodological approaches, our discussion highlights the relevance of continued research on inferring the traits and life history strategies of soil microbes.

References

1. Piton, G. *et al.* Life history strategies of soil bacterial communities across global terrestrial biomes. *Nature microbiology* 1–10 (2023).
2. Madin, J. S. *et al.* A synthesis of bacterial and archaeal phenotypic trait data. *Scientific Data* **7**, 1–8 (2020).
3. Mohanta, T. K. & Bae, H. The diversity of fungal genome. *Biological procedures online* **17**, 1–9 (2015).
4. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* 1–9 (2023).
5. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome biology* **16**, 1–18 (2015).
6. Wang, C. *et al.* Bacterial genome size and gene functional diversity negatively correlate with taxonomic diversity along a pH gradient. *Nature Communications* **14**, 7437 (2023).
7. Rodriguez-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. *Frontiers in Microbiology* **12**, 761869 (2022).

8. Royalty, T. M. & Steen, A. D. Theoretical and Simulation-Based Investigation of the Relationship between Sequencing Effort, Microbial Community Richness, and Diversity in Binning Metagenome-Assembled Genomes. *Msystems* **4**, 10–1128 (2019).

Author contributions

This reply was written by G.P. with inputs from A.C.M., F.H., S.D.A. and J.B.H.M. Metagenome composition simulations were carried out by G.P. All authors read and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-024-01688-9.

Correspondence and requests for materials should be addressed to Gabin Piton.

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