

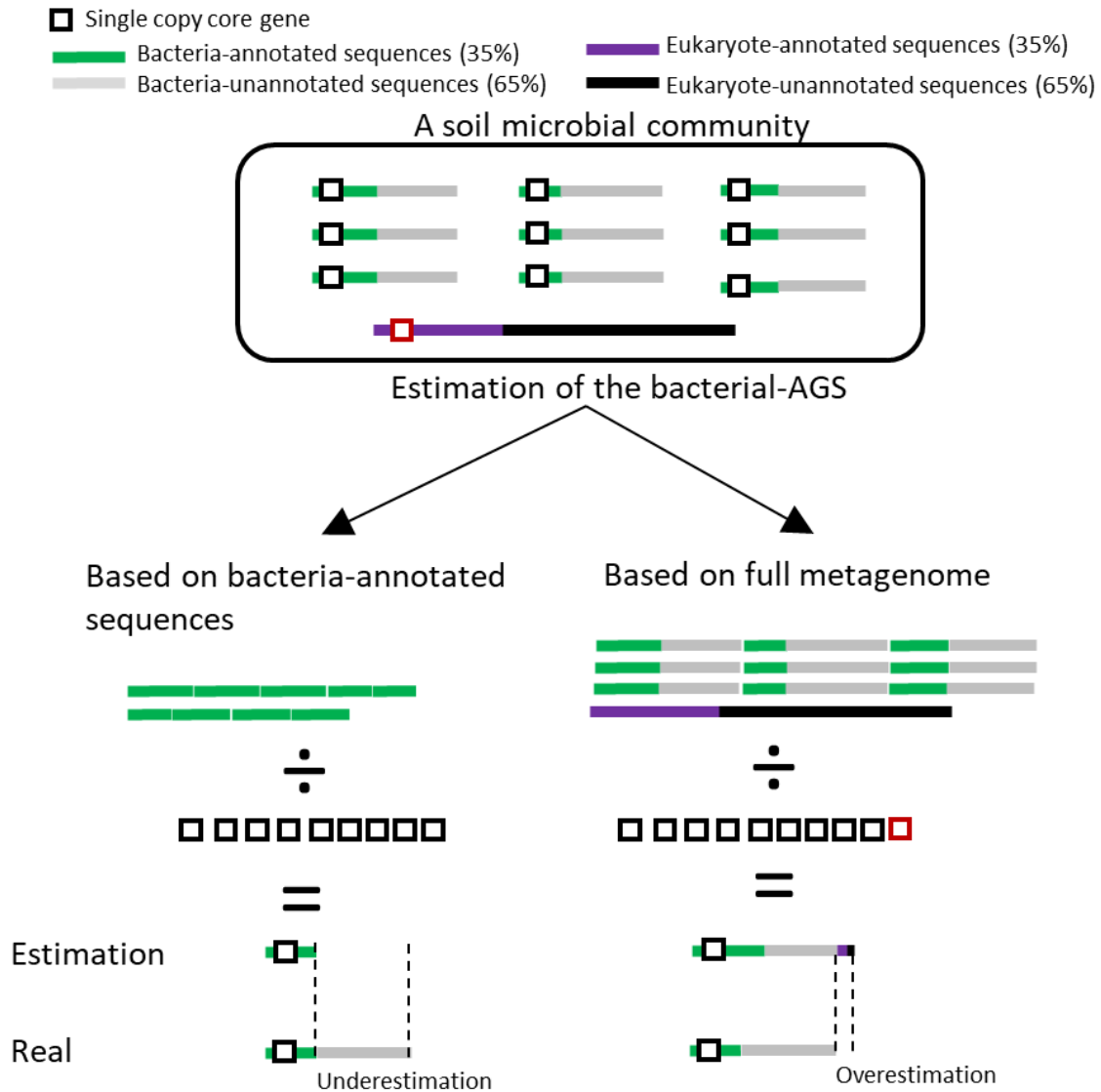
25 underestimation bias. Still, neither approach led to a different conclusion for our proposed life history
26 strategies, demonstrating the robustness of our results.

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

34 However, using only annotated sequences is also problematic. Fifty to 80% of metagenomic
35 reads from soils are typically not identified with current databases (i.e. the functional dark matter⁴)
36 and most of these are probably of bacterial origin⁴. Indeed, MicrobeCensus, the tool used to estimate
37 AGS with metagenomes⁴ relies on the principle that the AGS in a community is inversely
38 proportional to the relative abundance of single copy core genes (SCCG), ie. the ratio (SCCG base
39 pair / Total base pair). Well-characterized and evolutionarily conserved SCCGs (that is, genes that are
40 easy to identify based on homology) can be over-represented in the pool of annotated sequences.
41 Conversely, genes outside the core pangenome are more likely to be unknown in reference databases
42 (because they are absent from most genomes). Thus, relying solely on assigned metagenomic reads
43 will likely lead to an overestimation of SCCGs, while the total number bacterial reads is
44 underestimated, leading overall to an AGS underestimation⁵. That bias should be also considered
45 along with the bias of the full-metagenome calculation that we used in our original study (Figure 1).

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

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



52

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

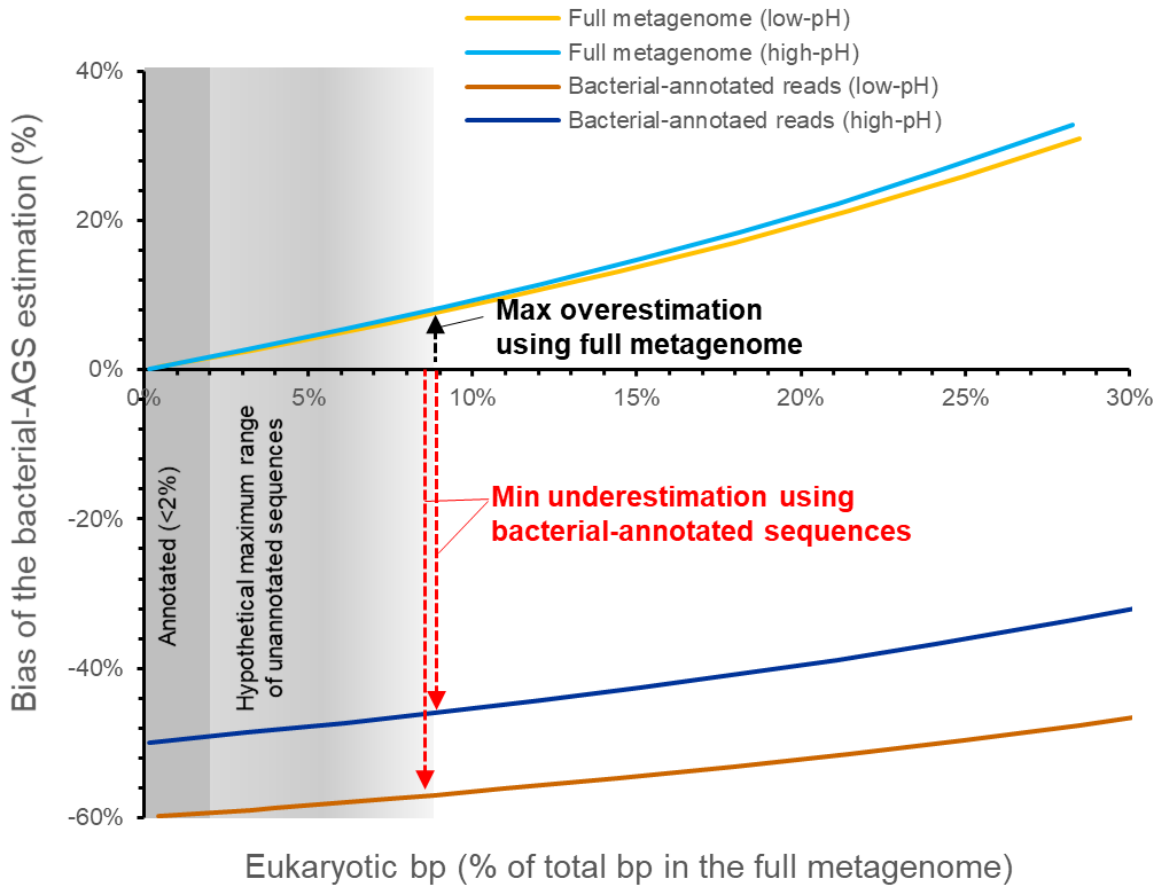
58 *represent only about 1% of the genomes in a metagenome, leading to lower positive bias than*
59 *represented in the figure.*

60

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

70 Simulation of metagenome composition can help in perceiving what conditions would
71 underlie different biases in bacterial-AGS quantification. We simulated how eukaryotic sequences
72 might bias the estimation of bacterial-AGS (Supplementary Note 1). This simulation (Figure 2) shows
73 that the bacterial-AGS estimated with bacteria-annotated sequences would be accurate and the one
74 using the full-metagenome fully artificial if 65% of the sequences were eukaryotic in low-pH soil
75 (93.5% of the unannotated sequences). However, given that eukaryotic sequences represent less than
76 2% of annotated sequences, we might also expect only a small fraction of eukaryotic sequences in the
77 unannotated base pairs⁴. Indeed, assuming an extreme range of 4 to 9% eukaryotic base pairs (2%
78 annotated and 2-7% unannotated, Supplementary Note 2) would lead to an overestimation of
79 bacterial-AGS by +3.2 to +7.5% in low pH soil using the full metagenome method (Real value=7.4-
80 7.8Mb, Estimated value=8Mb, Figure 2). In contrast, using only bacteria-annotated sequences would
81 lead to an underestimate of genome size between -57 and -59% (Estimated value=3.2Mb, Figure 2).
82 In such case, it thus appears that the assumptions associated with Osburn et al. method would strongly

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



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

90

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

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

97 Osburn et al. also argue that their estimate fits into the range of soil bacteria based on results from soil
98 bacteria metagenome-assembled-genomes (MAG)⁷. However, this range is also likely biased as
99 MAGs are more easily obtained for small genomes⁸. Moreover, Madin et al.³ report from ~2,000 soil
100 isolates a range of genome sizes (median, minimum and maximum values of 6.41, 1.26 and 16 Mb,
101 respectively) that covers both of our estimates' ranges (median, minimum and maximum values of
102 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
103 Osburn et al.²).

104 Finally, we quantified the degree to which our conclusions are affected by the different
105 methodologies for estimating genome size. Using the bacterial-AGS from both methods, we found
106 that the two trait dimensions used to characterize life history strategies¹ remained very much the same
107 (Supplementary Figure 1 and 2). Both AGS estimates were associated with both dimensions;
108 however, the estimation from bacterial-annotated sequences suggested a weaker association with the
109 first dimension and a tighter one with the second dimension than predicted based on the full-
110 metagenome method alone. In other words, estimation from bacteria-annotated sequences also
111 supports that bacterial AGS captures an extension of metabolic capacities (dimension 1) and
112 emphasizes that the AGS becomes especially large when this extension is oriented towards nutrient
113 recycling capacities (dimension 2)—a profile that we associate with the competitor strategy¹.

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

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

127

128 **References**

- 129 1. Piton, G. *et al.* Life history strategies of soil bacterial communities across global terrestrial
130 biomes. *Nature microbiology* 1–10 (2023).
- 131 2. Madin, J. S. *et al.* A synthesis of bacterial and archaeal phenotypic trait data. *Scientific Data* **7**,
132 1–8 (2020).
- 133 3. Mohanta, T. K. & Bae, H. The diversity of fungal genome. *Biological procedures online* **17**, 1–9
134 (2015).
- 135 4. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global metagenomics.
136 *Nature* 1–9 (2023).
- 137 5. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative
138 metagenomics and sheds light on the functional ecology of the human microbiome. *Genome*
139 *biology* **16**, 1–18 (2015).
- 140 6. Wang, C. *et al.* Bacterial genome size and gene functional diversity negatively correlate with
141 taxonomic diversity along a pH gradient. *Nature Communications* **14**, 7437 (2023).
- 142 7. Rodriguez-Gijón, A. *et al.* A genomic perspective across Earth’s microbiomes reveals that genome
143 size in Archaea and Bacteria is linked to ecosystem type and trophic strategy. *Frontiers in*
144 *Microbiology* **12**, 761869 (2022).

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

148

149 **Author contributions**

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

152 **Competing interests**

153 The authors declare no competing interests.

154 **Additional information**

155 **Supplementary information** The online version contains supplementary material available at

156 <https://doi.org/10.1038/s41564-024-01688-9>.

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

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