We thank the reviewers for their time and feedback to our manuscript. Please see below point by point responses, where line numbers refer to the clean version of the manuscript. A version of the manuscript with substantial changes highlighted has also been uploaded.

Reviewer #1:

1.1: Eisenhofer and colleagues present a short manuscript that further informs an ongoing debate in the literature: what percentage of soil metagenomic DNA is eukaryotic, how does that influence estimates of average prokaryotic genome size, and how does that further influence observed relationship between the average prokaryotic genome size in a community and the soil pH of that community? They apply a program they've developed called "singlem microbial_fraction" to estimate the microbial (and therefore also the eukaryotic) fractions of soil metagenomes to answer this question.

Overall the manuscript is well-written, well-presented, and the analysis methodology is largely sound and builds upon existing literature and their software that I have used myself before. Their finding that eukaryotic DNA can represent ~30% of soil metagenomic DNA is certainly surprising: this is a higher number than many people would have guessed.

I have several comments that arise from my reading of the manuscript:

We thank the reviewer for their enthusiasm and accurate reading of the manuscript. The finding that ~30% of metagenomic reads are non-microbial in soil metagenomes is somewhat surprising, but since the SMF method is the first that has been able to interrogate such questions, it has not been previously possible to assess metagenomes in this way.

1.2: 1. The authors should avoid saying that they are "settling" the debate. Rather, they should use more collegial language, such as "further informs this debate", "builds upon this debate", etc. I don't think their approach is so assumption-less that it completely settles this debate in all contexts.

While contributing substantially to the debate in the literature, we agree with the reviewer that our method is not without assumptions and therefore the discussion cannot be said to be fully "settled". However, in both instances the original text stated that the manuscript "helps settle" the debate, which is already a weaker version than that quoted by the reviewer. Nevertheless, we have modified the wording in the two instances:

Line 23: ***Here, we add to ~~help resolve~~ a recent debate***

Lines 52-54: ***We recently developed a tool that can account for such biases in AGS estimations[8], and we use it here to address the central uncertainty in ~~help settle~~ this debate.***

Lines 111-113: ***The application of our newly developed tool contributes substantially to ~~helps to settle~~ an ongoing scientific discussion that only a few months ago seemed unresolvable, and highlights the complexity and dynamism of the metagenomics research field.***

1.3: 2. An advantage of their approach is it does not rely on any estimates or references for eukaryotic genomes, instead only estimating the prokaryotic fraction and then assuming what is left over is the eukaryotic fraction. This is a good assumption, although it does not account for e.g. viruses - I think the authors should point this out (but can admit that viral DNA is likely to be fairly limited). It is also what is relevant for the question at hand where they are actually trying to estimate the prokaryotic fraction for average genome size estimates.

Both reviewers are right that our method does not distinguish between the different sources of non-microbial DNA in metagenomes, and that it is likely a combination of eukaryotic and viral DNA. We used the word eukaryotic as a shorthand for the non-prokaryotic fraction as this is likely the dominant source in soil samples, and because this is the way both the Piton et. al. and Osburn et. al. articles referred to it.

In response to the reviewers' valid concerns, we have added a paragraph (lines 80-84) describing that viral DNA may also contribute metagenomic reads alongside eukaryotic cells, and now refer to the "non-microbial" fraction throughout the remainder of the manuscript rather than the less precise "eukaryotic" fraction as previous.

1.4: 3. A challenge with the approach though is that it *does* rely on being able to infer the genome size of a detected taxon based on its species representative presence in GTDB. This is the main limitation of the approach, especially in soils, where this number may be high.

The authors delve into this assumption in their SingleM paper (https://www.biorxiv.org/content/10.1101/2024.05.16.594470v1.full.pdf), but not in this work. But because it is a core assumption, it is important to explain this assumption in this work as well.

In that other work, they write:

"In more complex communities such as those found in soils and many animal guts, the genome sizes of individual lineages may also be uncertain."
"For SMF to be inaccurate, many genome size estimates have to be incorrect, and either systematically overestimated or systematically underestimated"

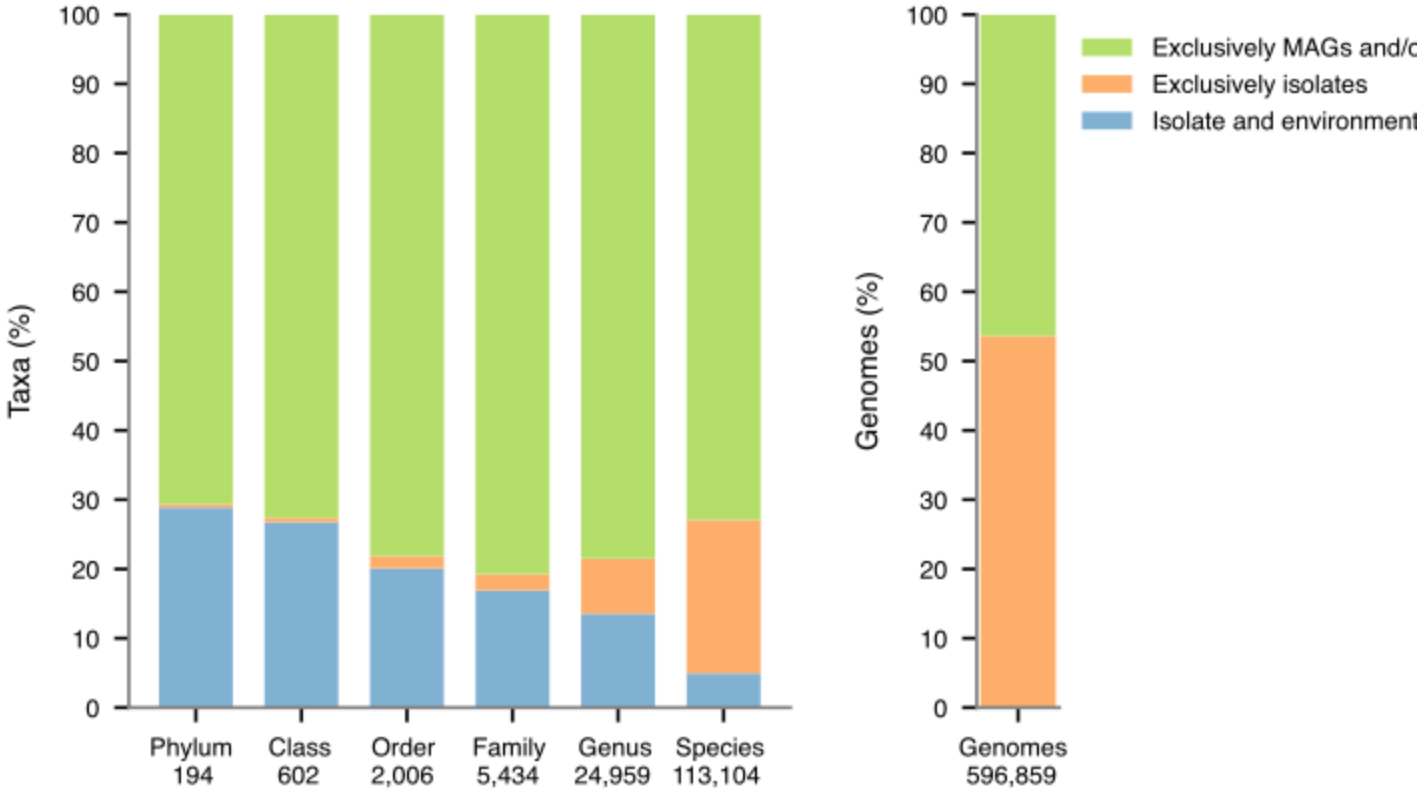
However, there is are in fact good reasons to guess that genome sizes of unrepresented taxa might be systematically different from those that are. For example, uncultivated microbes often have smaller genome sizes (which in part can explain why they are uncultivated). Simultaneously, MAGs, especially from soil, may be missing accessory gene content and thus be underestimates of true genome size.

The degree of each of these confounders is uncertain, but they certainly could play some role, and for that the authors should afford some discussion thereof of this caveat of their work.

If we understand the reviewer correctly, the assertion that SMF "*relies on being able to infer the genome size of a detected taxon based on its species representative presence in GTDB*" is not strictly true. SMF does rely on the genome size of the species when that species is detected as being present in the metagenome, but this does not mean that only lineages that are detected at the species level are used in the SMF calculation. Instead, when coverage is assigned to a taxon level higher than species (e.g. to a family), then the average genome size of the genera in that family is used as the estimate for the family. In this way, the microbial fraction can be estimated even if no species are assigned at the species level in the SingleM community profile.

Taking this comment alongside R1.6 below, it appears the reviewer has misunderstood the SMF algorithm. In response, we have expanded the brief description of the SMF algorithm in the introduction of the manuscript (lines 56-70).

The reviewer suggests that the algorithm may overestimate the genome sizes of uncultivated species since the reference is largely based upon genomes from cultivated lineages. However, the assertion that the database is largely based on cultivated lineages is incorrect. Being based upon the GTDB, the majority of genomes in the SingleM reference database are MAGs or SAGs. From the GTDB R220 statistics page (https://gtdb.ecogenomic.org/stats/r220):



The reviewer also comments that "*MAGs, especially from soil, may be missing accessory gene content and thus be underestimates of true genome size.*" While potentially true to an extent, the SMF algorithm corrects for genome incompleteness by scaling reference genome sizes according to their completeness (and contamination) as estimated by CheckM2 (Chklovski et al. 2023)**. Relying upon the accuracy of the widely cited CheckM2 algorithm, SMF accounts for this potential loss of accessory gene content.**

We hope that we have convinced the reviewer that the method is robust in the absence of species-level representatives in the GTDB. While we agree that there is always some level of methodological uncertainty, we feel that discussing such a complex topic in this concise article would detract from its core message. We delve into these intricacies in the SMF paper (reference #8; https://www.biorxiv.org/content/10.1101/2024.05.16.594470v1), and have further referenced it in response to 1.5 below.

1.5: 4. I think it is prudent that the authors report the fraction of SingleM OTUs that do in fact have species representatives in GTDB, and are thereof informing their estimate. This is the main piece of analysis missing from the work.

We have reported this fraction on lines 92-94 of the clean version of the updated manuscript. We also added that SMF is robust to novel species by referencing the main SMF manuscript, related to the previous point (1.4) above.

1.6: 5. If the authors wanted to extend their analysis significantly, they could see if they could construct a model for genome size estimation of new microbes based on phylogenetic conservation of the trait (e.g. try to guess the genome size of a new species based on the genome sizes in its genus), which could further inform their estimates in communities with few represented taxa. However, this is just an idea for the authors, and is not necessary.

The reviewer's suggestion here is certainly a good one, in fact we already do this. In cases where coverage is assigned to the genus level but no further, genome sizes are estimated as the mean of the genome sizes of the species in that genus. Similar methods are used at higher taxonomic levels e.g. genome size at the family level is estimated as the mean of its genus level genome sizes.

Further details of this can be found in the methods section of the manuscript describing SMF (Eisenhofer, Alberdi, and Woodcroft 2024) **starting line 141, and copied below:**

Estimation of genome size for SMF calculation

To estimate the true $genome\_size$ of each species, accounting for imperfections which arise from genome recovery efforts, we take the genome size of the GTDB species representative and adjust it based on its completeness and contamination as predicted by CheckM2 (Chklovski et al. 2023)**:**

$$genome\_size\_adjusted = \frac{genome\_size}{completeness \times (1 + contamination)}$$

Community profiles generated by SingleM also contain entries from lineages not resolved to the species level e.g. coverage assigned to the genus or family level. For these cases we use genome sizes estimated from lower taxonomic ranks. For genera:

$$genus\_genome\_size = mean(species\_genome\_sizes\_adjusted)$$

Where $species\_genome\_sizes\_adjusted$ is the set of adjusted genome sizes of species within the genus. For family, order, class, phylum and domain levels we average the average genome sizes of the taxons immediately below them. For families:

$$family\_genome\_size = mean(genus\_genome\_sizes)$$

Where $genus\_genome\_sizes$ is the set of genome sizes of genera belonging to that family. Orders, classes, phyla and domains are calculated similarly based on the mean genome sizes of their families, orders, classes and phyla respectively.

Reviewer #2:

2.1: Shotgun metagenomics is commonly used to study environmental microbial communities to infer microbial composition, metabolic potentials, and functional traits of prokaryotic microorganisms. Bioinformatics applications developed for analyzing prokaryotic microbial communities can be biased by the presence of variable amounts of eukaryotic sequences in the metagenomes. In this study, Eisenhofer et al employed a novel method (SingleM Microbial Fraction, SMF) that claims to be able to estimate the prokaryotic fraction of a metagenome accurately using a taxon-specific abundance aware community average genome size (AGS) estimation. Using this method, the authors found that the proportions of eukaryotic sequences in global soil metagenomes can be considerably high, which could lead to an overestimation of AGS using the conventional method, which calculates AGS as the ratio of total reads to the number of reads recruited by detected marker genes. When this bias was corrected using SMF, a stronger correlation between AGS and soil pH could be observed, suggesting this improved method could enhance the reliability of microbial genomic trait estimation.

Although this manuscript is timely and resolves a critical debate, several issues should be addressed before it is accepted for publication, as listed below.

We thank the reviewer for their accurate reading of the manuscript.

2.2: 1. MAGs in GTDB are usually incomplete and sometimes with contaminations, so are these issues addressed by the SMF method? If yes, how?

Yes, these issues are accounted for in the SMF algorithm. In particular, the adjusted genome sizes used by SMF are estimated from GTDB genome sizes, using a straightforward interpretation of completeness and contamination estimates produced by CheckM2 (We have added a brief description of this process to the paragraph on lines 56-70). The specific procedure for doing so is described in the methods section of the SMF manuscript (Eisenhofer, Alberdi, and Woodcroft 2024) **and copied below:**

To estimate the true $genome\_size$ of each species, accounting for imperfections which arise from genome recovery efforts, we take the genome size of the GTDB species representative and adjust it based on its completeness and contamination as predicted by CheckM2 (Chklovski et al. 2023)**:**

$$genome\_size\_adjusted = \frac{genome\_size}{completeness \times (1 + contamination)}$$

2.3: 2. It should be better explained how SMF corrects the bias of variable eukaryotic sequences. Although I can find some of the explanations in the singleM documentation, as

the core of method improvement, I believe it should be well described in the Methods section of the manuscript.

The core SMF methodology is described in detail as part of the manuscript dedicated to it (Eisenhofer, Alberdi, and Woodcroft 2024)**. However, we have expanded the section which briefly described the SMF (and SingleM) methodology so that it now provides a fuller conceptual context for the reader (lines 56-70).**

2.4: 3. If I understand correctly, based on the singleM documentation, only the microbial fraction is calculated, while the source of the remaining non-microbial fraction can be eukaryotes, viruses, or plasmids. In this case, the estimation of eukaryotic fractions (Figure 1A) is not accurate either.

As per R1.3 above, we have now modified the wording of the entire manuscript to use "non-microbial" rather than "eukaryotic", including in the legend of Figure 1.

We note that while eukaryotic and viral DNA is considered non-microbial, plasmids may contribute genes necessary for the function of the host bacteria or archaea, and so we consider plasmid DNA to be microbial for the purposes of SMF. We have clarified this in documentation of SingleM microbial_fraction (https://wwood.github.io/singlem/tools/microbial_fraction).

Cited literature:

Chklovski, Alex, Donovan H. Parks, Ben J. Woodcroft, and Gene W. Tyson. 2023. "CheckM2: A Rapid, Scalable and Accurate Tool for Assessing Microbial Genome Quality Using Machine Learning." *Nature Methods* 20 (8): 1203–12.

Eisenhofer, Raphael, Antton Alberdi, and Ben J. Woodcroft. 2024. "Large-Scale Estimation of Bacterial and Archaeal DNA Prevalence in Metagenomes Reveals Biome-Specific Patterns." *bioRxiv*. https://doi.org/10.1101/2024.05.16.594470.