# Review history

## First round of review

### Reviewer 1

The article by Passarelli-Araujo et al., makes an interesting observation; that is, that species with smaller pangenomes tend to have a longer ANI gap to their closest relative in the genome database relative to species with larger pangenomes. This is an important topic as we (i.e., the scientific community) are trying to understand the pattern of microbial genome diversity in nature revealed by sequencing and appropriate for Genome Biology. The authors seem to attribute this finding to the former organisms having less flexible genomes and being less prone to horizontal gene transfer. My major concern with the manuscript is on the interpretation; the approach/method seems robust. The extent of gene content similarity (and thus, openness of the pangenome) correlates with ANI (e.g., more closely related genomes tend to have smaller gene content differences, on average). Therefore, more clonal groups typically have more closed pangenomes and one has to normalize for this, i.e., the genomic diversity within species (or the cluster of genomes used). No normalization was attempted; thus, the smaller pangenome is likely the direct effect of smaller diversity within the species rather than caused by (or itself causing) the ANI distance to the closest relative or anything else. Related to this, it is equally likely, in my view at least, that the pattern the authors see is due to longer time since last diversity purging event (or longer time since the emergence of novel diversity that created/invaded a new ecological niche, or even that we have not yet sequenced the close relatives), and the genomes have not evolved much yet since that event to acquire gene content difference. In other words, it is not the pangenome that drives (or is driven by) the (extent of the) ANI gap but the nature of the selective sweeps in the corresponding cases. Indeed, most of the species with small pangenomes mentioned (e.g., M. tuberculosis) are restricted-lifestyle organisms (usually specialized pathogens that can leave long time periods as spores or resting cells), which is consistent with an idea that they have uniquely invaded a (new?) ecological niche. In summary, it seems likely that (unique/specialized) ecology drives the extent of the ANI gap, rather than the size of the pangenome, and the pangenome size may have just happened to correlate to the extent of the ANI gap in some cases where the genomes of the species have not diverged enough since the diversity selective sweep yet for some reason. These reasons could be the recent invasion of a niche, "stopping the evolution clock" relative to their close relatives due to a spore/resting cell stage, or no need/advantage for horizontal transfer within that niche; the latter is the one favored/implied by the authors, but I don't think it is the only one. I think the authors should discuss these alternative scenarios. I don't necessarily see these scenarios are incompatible with the scenario favored by the authors, but they should be discussed for a more complete picture to emerge, in my view. Importantly, the code can not be found where the authors say it is: https://github.com/Passarelli-bio/Data_Bacterial_Discontinuity. Also, the authors ran many programs (mentioned in their methods) but they do not mention parameters used. Please add these for the work to be reproducible.

Minor/specific points.

From the Abstract:
Line 13, page 3. I would define what delta is or remove it. Is it a proxy of ANI distance or something else?

line 22-23, page 3. "Additionally, through a machine learning approach, we detected key features that impact genetic discontinuity prediction." Such as? Again, I would suggest the authors to mention some specifics or remove such vague sentences or terms (e.g., for "delta" above).

line 47-48, page 3. The first systematic and large-scale reports on the existence of discrete sequence clusters corresponding to species was by Konstantinidis (e.g. Konstantinidis and DeLong, ISME 2008; Caro-Quintero and Konstantinidis EMI 2012). I think you also want to check and probably cite the recent paper by his team, which seems highly relevant (Rodriguez-R et al, mBio 2024).

line 45, page 3. Were these genomes really removed from further analysis or just not added to "a community"? Because if removed, this could affect the extend of the ANI gap, right?

Line 24-25, page 5. Bacillus cereus and the very closely related Bacillus anthracis is a good example of what I am talking about above. The ANI distance between them is substantial, 3-4 ANI units, and the pangenome of B. anthracis is tiny compared to that B. cereus (a close relative, sharing about 95% ANI) but this is due to the very small intra-species diversity of B. anthracis not that they two closely related organisms are that different in genome structure or frequency of HGT. They are very closely related/similar organisms.

Line 55-56, page 5. What is an orthogroup? One needs this definition to be able to follow this part. Is it othologous groups of genes? And if so, does this really represent a metric of genome size (e.g., higher number of shared orthogroups would correlate with higher genome size) or (or in addition) to having relatives among the ~260 species or not? I hope the authors can explain better the term, and what biological property it really represents/reflects. The methods were not much more useful with this respect. The remaining of the methods were clear, I felt.

Line 45, page 6. Here and elsewhere: When you write "genetic identity", do you mean sequence identity/relatedness or gene-content sharing? Please clarify/define.

**Reviewer 2**

In this study, Passarelli-Araujo and colleagues tested to what extent bacterial species form discrete entities rather than a continuum. This is a central question to understand how to identify bacterial species and to understand how they are formed. The authors' results support previous findings that bacterial species are typically separated by clear breaks. They further find a strong link between species discontinuity and the pan-genome openness. Overall, this is a very interesting study and a great contribution to the "bacterial species problem".

Some proponents of the idea that bacterial species are an artifact often argue that these patterns of discontinuity are the result of sampling issues. I don't think this is the case, since previous studies have convincingly shown that this is unlikely based on metagenomic data. However, I think that the manuscript would benefit from discussing this point or mentioning previous works that have shown this.

The authors are reporting a correlation between genomic fluidity and pangenome saturation (Fig. 3). Isn't it tautological? These two metrics are expected to be strongly connected.

The classification of bacterial lifestyles is a complex task. The authors have analyzed the lifestyles using the features of the pan-genome and this is something I am not very familiar with. I think that the authors should give more details on how this was done and what is the basis of it. The lifestyles reported in Table 1 appear rather approximative since free-living and opportunistic pathogens can often be found in the same species. For instance, E. coli can be pathogenic, but it has been classified

as "free-living" in Table 1. Nevertheless, the link between lifestyle and genomic discontinuity appears well supported.

The link between pan-genome openness, lifestyle and species discontinuity is very interesting but it seems that the authors are not explicitly proposing or discussing a mechanism to explain this pattern. I think that if would be interesting to discuss their results in light of other works. In particular, the connection between discontinuity and pan-genome openness hints heavily towards the role of DNA transfers in the core genome and in the accessory genome, since they would maintain genetic cohesiveness and pan-genome openness, respectively. Related to this point, I believe that this recent article would be relevant to discuss: Diop et al. Genome Biology 2022.

The approach to detect discontinuity between bacterial genomes is interesting, but I can't help wondering if this could be further improved by taking into consideration the abundance of genomes. Some clades might appear continuous due to a single genome, and expressing delta by the number of genomes could further amplify the breaks if some lineages appear continuous due to the presence of a single of very few genomes. Of course, this further opens the issue of sampling bias and metagenomic data might be less biased for such an analysis. Just a thought!

**Reviewer 3**

Passarelli-Araujo and colleagues have submitted a well-conceived study of genetic discontinuities in bacteria and introduce a simple but meaningful metric that quantifies the magnitude of discontinuities. This represents an important advance in the field of microbial genomics and the study of bacterial speciation.

That said, I do have several concerns regarding the motivation of the methodology and question whether the authors have sufficiently explored how sensitive this metric is to their specific choices. These are detailed below.

- Are the methods appropriate to the aims of the study, are they well described, and are necessary controls included? If not, please specify what is required.

Some significant clarifications and improvements to the methodology are necessary.

1. While I think I understand what differentiates the detection of discontinuities in ANI estimates from the approach the authors take, I am not confident that someone unfamiliar with the specific discourse around ANI in the field of microbial genomics has enough context. For a specific example, in the last paragraph of the Background section, why is it important to quantify the "magnitude" of these breaks, and what exactly does "magnitude" mean in this context? To be clear, I think that this is quite an elegant measure, it just needs more motivation in the text.

2. The choice of thresholds used in the analysis should be explored further. Specifically, the definition of delta as the maximum genetic rate of change above 94%. While a justification is given, I think it is important to understand just how sensitive the results are to changes to the 94% threshold, AND the inclusion or exclusion of genomes through some form of jackknifing.

3. How sensitive is the measurement of discontinuity to choosing different bait nodes for each species analyzed? How was the actual bait node chosen?

4. The ML methods need to be better-motivated. Specifically, why were the 31 different metrics chosen as candidate predictors? The authors describe these in a general way on page 5, lines 32-46 and again on page 9 lines 35-42, but it's still not clear to me exactly what all these metrics were. Table S1 was referenced but not present in my copy of the manuscript. This might provide some clarification?

5. The quantile regression relating delta to pangenome openness should be described in the methods section.

- Are the conclusions adequately supported by the data shown? If not, please explain

I believe the conclusions are largely supported by the data shown, but it is hard for me to evaluate this fully without more careful sensitivity analyses.

- Are sufficient details provided to allow replication and comparison with related analyses that may have been performed? If not, please specify what is required.

I thank the authors for making their code available on github with sufficient documentation. Given this, I don't have concerns with replicability.

- Does the work represent a significant advance over previously published studies?

Yes, in that quantitative measurements of the genetic discontinuity within species has been lacking and a huge gap in the field. This paper represents a significant step forward in exploring this important feature of microbial species and populations.

- Is the paper of broad interest to others in the field, or of outstanding interest to a broad audience of biologists?

Yes, I think that this work represents a really key advance in our ability to quantify genetic discontinuities that has potential applications beyond microbial genomics (e.g., quantifying genomic discontinuities along animal hybrid zones). However, there are a couple of areas where a broader audience could use some clarification.

1. The authors switch from describing boundaries between species to discontinuities between populations in the same paragraph, and I think clarifying this distinction would be helpful for most readers.
2. At the beginning of page 3 with the discussion of G. vaginalis, I think it would be helpful to provide a little more detail on why the detect

One language comment:

1. I would recommend scanning more thoroughly for typos. For instance, on page 8, line 34, I believe the sentence should read, "All single-copy genes were aligned with Mafft..."

**Authors' response to reviewers**

**Reviewer #1**

The article by Passarelli-Araujo et al., makes an interesting observation; that is, that species with smaller pangenomes tend to have a longer ANI gap to their closest relative in the genome database relative to species with larger pangenomes. This is an important topic as we (i.e., the scientific community) are trying to understand the pattern of microbial genome diversity in nature revealed by sequencing and appropriate for Genome Biology. The authors seem to attribute this finding to the former organisms having less flexible genomes and being less prone to horizontal gene transfer.

R: Thank you for your positive feedback. We appreciate the opportunity to answer the concerns you raised in detail.

My major concern with the manuscript is on the interpretation; the approach/method seems robust. The extent of gene content similarity (and thus, openness of the pangenome) correlates with ANI (e.g., more closely related genomes tend to have smaller gene content differences, on average). Therefore, more clonal groups typically have more closed pangenomes and one has to normalize for this, i.e., the genomic diversity within species (or the cluster of genomes used). No normalization was attempted;

R: It is certainly the case that more clonal groups typically have more closed pangenomes. The correlation between pangenome openness and bacterial lifestyle is well supported (McInerney, McNally, and O'Connell 2017; Ceres, Stanhope, and Gröhn 2022; Puente-Sánchez et al. 2023). This is the main reason we used pangenome openness as a proxy of bacterial lifestyle.

Regarding the normalization, pangenome saturation (α) and fluidity (ϕ) are metrics used to describe genomic diversity within species, each from a different perspective. The pangenome saturation measures how completely the gene repertoire of a species has been captured, reflecting whether additional sequencing will likely discover many new genes (low saturation, open pangenome) or few new genes (high saturation, closed pangenome).

In contrast, fluidity quantifies the variability or turnover in gene composition among different genomes within the same species, with high fluidity indicating frequent genetic changes like gene loss or horizontal gene transfer. Together, these metrics provide insights into intra-species diversity, and how it has come about.

We could think about using the information about clonal complexes for each species in the study, but it would require the knowledge of the population structure of each species - which poses significant challenges, particularly in standardizing data across different species, including some that are not fully characterized despite numerous available genomes. Therefore, we relied on pangenome openness and fluidity to infer both direct and indirect aspects of diversity within species.

Thus, the smaller pangenome is likely the direct effect of smaller diversity within the species rather than caused by (or itself causing) the ANI distance to the closest relative or anything else.

R: Based on this and other comments from the reviewer, we believe that it may be helpful to clarify some terms such as pangenome size, pangenome saturation, and genomic fluidity:

- **Pangenome Size** refers to the total number of non-redundant gene families identified across all genomes within a species.
- **Pangenome Openness** describes the extent to which the pangenome of a species is expected to expand as additional genomes are sequenced.
- **Genomic Fluidity** measures the variability in gene content across different genomes within a species, reflecting how gene composition changes from one strain to another.

In the manuscript, we do *not* assert that the ANI distance to the closest relative results in smaller pangenomes. Instead, we claim that pangenome openness (not pangenome size) correlates with estimates of genetic discontinuity (abrupt breaks in ANI distribution), but we do not wish to make any strong claims as to causation.

Related to this, it is equally likely, in my view at least, that the pattern the authors see is due to longer time since last diversity purging event (or longer time since the emergence of novel diversity that created/invaded a new ecological niche, or even that we have not yet sequenced the close relatives), and the genomes have not evolved much yet since that event to acquire gene content difference.

R: We acknowledge that our initial interpretation may have implied a direct causative relationship between pangenome size and horizontal gene transfer (HGT) frequency without sufficient consideration of these underlying ecological and evolutionary dynamics. We have corrected this by proposing other scenarios in the discussion section.

In other words, it is not the pangenome that drives (or is driven by) the (extent of the) ANI gap but the nature of the selective sweeps in the corresponding cases. Indeed, most of the species with small pangenomes mentioned (e.g., M. tuberculosis) are restricted-lifestyle organisms (usually specialized pathogens that can leave long time periods as spores or resting cells), which is consistent with an idea that they have uniquely invaded a (new?) ecological niche.

R: We emphasize that our analysis focuses on pangenome openness rather than size. For instance, we note that *M. tuberculosis* exhibits a low saturation coefficient, indicative of a closed pangenome, which aligns with the restricted lifestyle which the reviewer rightly mentions. Furthermore, our manuscript discusses correlations, not causality, between these genomic features.

In summary, it seems likely that (unique/specialized) ecology drives the extent of the ANI gap, rather than the size of the pangenome, and the pangenome size may have just happened to correlate to the

extent of the ANI gap in some cases where the genomes of the species have not diverged enough since the diversity selective sweep yet for some reason.

R: A key finding in our study is the correlation between pangenome openness and genetic discontinuity (ANI gap), aligning with your observation that specialized ecology influences the extent of the ANI gap. We utilize pangenome openness as an indicator of a species' lifestyle or the uniqueness of its ecological niche. Importantly, we do not assert a correlation between ANI gap and pangenome size; rather, our focus is on pangenome openness. Our machine learning results further demonstrate that pangenome size does not significantly explain genetic discontinuity.

These reasons could be the recent invasion of a niche, "stopping the evolution clock" relative to their close relatives due to a spore/resting cell stage, or no need/advantage for horizontal transfer within that niche; the latter is the one favored/implied by the authors, but I don't think it is the only one. I think the authors should discuss these alternative scenarios. I don't necessarily see these scenarios are incompatible with the scenario favored by the authors, but they should be discussed for a more complete picture to emerge, in my view.

R: We thank the reviewer for this suggestion and in response we have included alternative scenarios in the main text.

Importantly, the code can not be found where the authors say it is: https://github.com/Passarelli-bio/Data_Bacterial_Discontinuity. Also, the authors ran many programs (mentioned in their methods) but they do not mention parameters used. Please add these for the work to be reproducible.

R: We appreciate the reviewer for pointing this out and apologize for the oversight. We have deposited the codes with parameters under a ZENODO repository (DOI: 10.5281/zenodo.13125800).

Minor/specific points. From the Abstract:

Line 13, page 3. I would define what delta is or remove it. Is it a proxy of ANI distance or

something else?

R: The sentence in the abstract defines delta as genetic discontinuity: "*we aimed to quantify the genetic discontinuity (δ) and investigate how this metric is related to ecology.*"

line 22-23, page 3. "Additionally, through a machine learning approach, we detected key features that impact genetic discontinuity prediction." Such as? Again, I would suggest the authors mention some specifics or remove such vague sentences or terms (e.g., for "delta" above).

R: Thank you for the feedback. We have updated the abstract to explicitly mention the key features affecting genetic discontinuity, namely gene conservation patterns and functional

annotations. To maintain clarity and conciseness, we refer to orthogroups-related features as gene conservation patterns in the abstract.

line 47-48, page 3. The first systematic and large-scale reports on the existence of discrete sequence clusters corresponding to species was by Konstantinidis (e.g. Konstantinidis and DeLong, ISME 2008; Caro-Quintero and Konstantinidis EMI 2012). I think you also want to check and probably cite the recent paper by his team, which seems highly relevant (Rodriguez-R et al, mBio 2024).

R: We appreciate the suggestions and added the references accordingly.

line 45, page 3. Were these genomes really removed from further analysis or just not added to "a community"? Because if removed, this could affect the extent of the ANI gap, right?

R: Regarding lines 45-46 on page 3, where we stated: "Representative genomes for each community were selected by removing edges below 95% identity in $M$". We realize this could mistakenly imply the removal of genomes. In our matrix $M$, genomes are represented as nodes, and it was the edges, not the genomes, that were removed. To prevent any confusion for future readers, we have revised the text to clarify that only edges were removed, not the genomes themselves.

Line 24-25, page 5. Bacillus cereus and the very closely related Bacillus anthracis is a good example of what I am talking about above. The ANI distance between them is substantial, 3-4 ANI units, and the pangenome of B. anthracis is tiny compared to that B. cereus (a close relative, sharing about 95% ANI) but this is due to the very small intra-species diversity of B. anthracis not that they two closely related organisms are that different in genome structure or frequency of HGT. They are very closely related/similar organisms.

R: The table below shows the key pangenome features for both species.

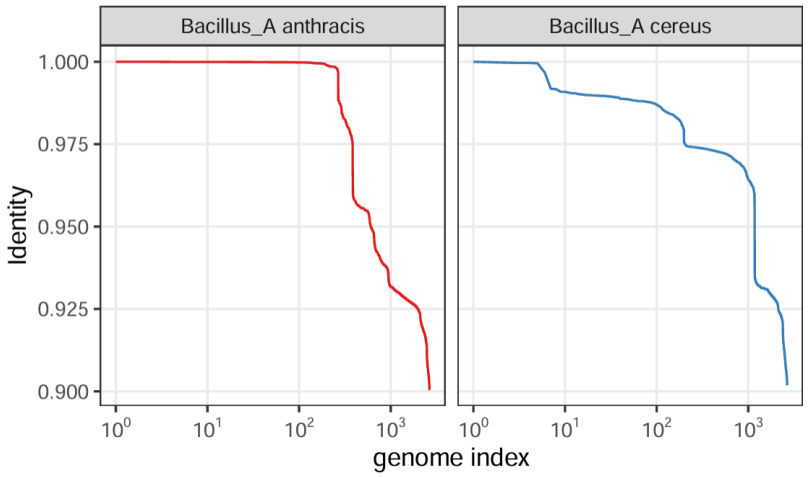
| **Species** | **# Samples** | **Pangenome size** | **Pan. Saturation (α)** | **Fluidity (ϕ )** |
|---|---|---|---|---|
| *B. anthracis* | 500 | 36,400 | 0.6618117 | 0.17 |
| *B. cereus* | 500 | 49,130 | 0.6404846 | 0.18 |

As noted, the pangenome size of *B. anthracis* is substantially smaller than that of *B. cereus*; and it is correct that this difference primarily reflects the reduced intra-species diversity of *B. anthracis*. However, we are not focusing on pangenome size alone, but two other metrics that inform intra-species diversity: pangenome saturation and genomic fluidity.

It is crucial to highlight that the apparent substantial ANI distance between these two species can be misleading if considered in isolation. This does not necessarily imply a significant difference in their genome structures or HGT frequency. Although the pangenome size differs

between species, the pangenome saturation and genomic fluidity are very similar. These two metrics, as shown in our analysis, support similar levels of genomic stability and HGT frequency, which corroborates their close genetic relationship, as also stated by the reviewer.

Further, the identity plots provided for both species illustrate how the genetic identity within each species varies less dramatically in *B. anthracis* compared to *B. cereus*, highlighting the small intra-species diversity of *B. anthracis* compared to *B. cereus*.




Line 55-56, page 5. What is an orthogroup? One needs this definition to be able to follow this part. Is it orthologous groups of genes? And if so, does this really represent a metric of genome size (e.g., higher proportion of shared orthogroups would correlate with higher genome size) or (or in addition) to having relatives among the ~260 species or not? I hope the authors can explain better the term, and what biological property it really represents/reflects. The methods were not much more useful with this respect. The remaining methods were clear, I felt.

R: We are grateful for this constructive feedback regarding the clarity of most of the methods. An orthogroup refers to a collection of genes across multiple species that are identified as orthologs, meaning they are derived from a single ancestral gene in the last common ancestor of the species being studied. The proportion of shared orthogroups between species does not directly represent genome size but rather indicates the degree of shared evolutionary history and functional traits. A higher proportion of shared orthogroups typically suggests closer evolutionary relationships or more extensive shared ecological roles, rather than directly correlating with genome size. We revised the manuscript to add more information about orthogroups in the methods section to make the text clearer for a broader audience.

**Reviewer #2**

In this study, Passarelli-Araujo and colleagues tested to what extent bacterial species form discrete entities rather than a continuum. This is a central question to understand how to identify bacterial species and to understand how they are formed. The authors' results support previous findings that bacterial species are typically separated by clear breaks. They further find a strong link between species discontinuity and the pan-genome openness. Overall, this is a very interesting study and a great contribution to the "bacterial species problem".

R: We greatly appreciate the reviewer's support.

Some proponents of the idea that bacterial species are an artifact often argue that these patterns of discontinuity are the result of sampling issues. I don't think this is the case, since previous studies have convincingly shown that this is unlikely based on metagenomic data. However, I think that the manuscript would benefit from discussing this point or mentioning previous works that have shown this.

R: We have now added additional references that have been exploring bacterial discontinuity from metagenomic data (Caro-Quintero and Konstantinidis 2012).

The authors are reporting a correlation between genomic fluidity and pangenome saturation (Fig. 3). Isn't it tautological? These two metrics are expected to be strongly connected.

R: Thank you for your observation on the correlation between genomic fluidity and pangenome saturation. While these metrics are related, their correlation is not tautological; rather, it elucidates distinct yet complementary aspects of genomic diversity. Genomic fluidity measures the turnover of genes within a species, indicating gene content variability across strains, whereas pangenome saturation assesses how comprehensively the gene pool is sampled as more genomes are sequenced. Assessing the correlation between these two metrics helps understand how gene variability within strains (fluidity) contributes to the overall scope of the gene pool (saturation). Thus, while the correlation might appear tautological, it is quantified to determine if the observed pattern diverges from expected trends.

The classification of bacterial lifestyles is a complex task. The authors have analyzed the lifestyles using the features of the pan-genome and this is something I am not very familiar with. I think that the authors should give more details on how this was done and what is the basis of it. The lifestyles reported in Table 1 appear rather approximative since free-living and opportunistic pathogens can often be found in the same species. For instance, E. coli can be

pathogenic, but it has been classified as "free-living" in Table 1. Nevertheless, the link between lifestyle and genomic discontinuity appears well supported.

R: We chose to classify lifestyles based on predominant ecological behaviors sourced from literature, despite known pathogenic capabilities in some strains. We have added a footnote in Table 1 to clarify for the reader the reasons and bases for those classifications, as demonstrated by *E. coli*, that exhibit dual behaviors (e.g serotype O157 regarded as pathogenic). This update will clarify our approach in correlating genomic features with ecological roles and ensure our findings' robustness and transparency.

The link between pan-genome openness, lifestyle and species discontinuity is very interesting but it seems that the authors are not explicitly proposing or discussing a mechanism to explain this pattern. I think that it would be interesting to discuss their results in light of other works. In particular, the connection between discontinuity and pan-genome openness hints heavily towards the role of DNA transfers in the core genome and in the accessory genome, since they would maintain genetic cohesiveness and pan-genome openness, respectively. Related to this point, I believe that this recent article would be relevant to discuss: Diop et al. Genome Biology 2022.

R: We are grateful for the suggestion. We have added this reference to the manuscript.

The approach to detect discontinuity between bacterial genomes is interesting, but I can't help wondering if this could be further improved by taking into consideration the abundance of genomes. Some clades might appear continuous due to a single genome, and expressing delta by the number of genomes could further amplify the breaks if some lineages appear continuous due to the presence of a single or very few genomes. Of course, this further opens the issue of sampling bias and metagenomic data might be less biased for such an analysis. Just a thought!

R: The about considering genome abundance to refine the detection of genetic discontinuity is well taken. In our study, we have strived to address potential biases associated with genome representation by setting a threshold for inclusion in the analysis. As outlined in our methods section, only communities with more than 40 genomes were considered. This criterion ensures that our pangenome estimates (besides discontinuity) are based on a sufficiently robust sample size.

**Reviewer #3**

Passarelli-Araujo and colleagues have submitted a well-conceived study of genetic discontinuities in bacteria and introduce a simple but meaningful metric that quantifies the magnitude of discontinuities. This represents an important advance in the field of microbial

genomics and the study of bacterial speciation. That said, I do have several concerns regarding the motivation of the methodology and question whether the authors have sufficiently explored how sensitive this metric is to their specific choices. These are detailed below.

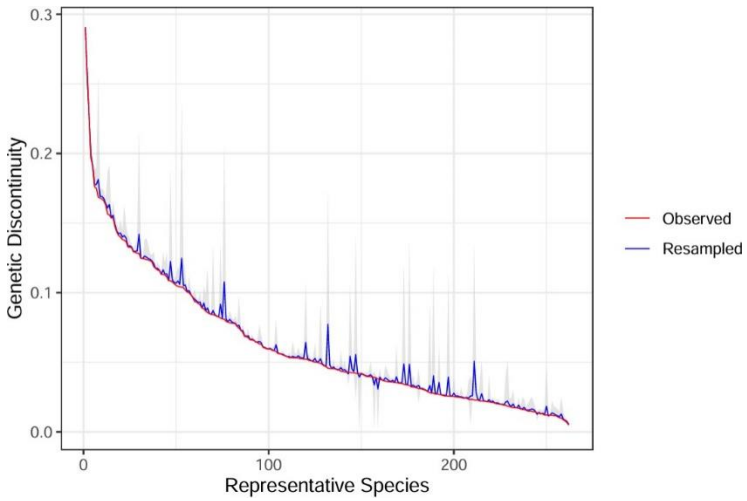
- Are the methods appropriate to the aims of the study, are they well described, and are necessary controls included? If not, please specify what is required.

Some significant clarifications and improvements to the methodology are necessary.While I think I understand what differentiates the detection of discontinuities in ANI estimates from the approach the authors take, I am not confident that someone unfamiliar with the specific discourse around ANI in the field of microbial genomics has enough context. For a specific example, in the last paragraph of the Background section, why is it important to quantify the "magnitude" of these breaks, and what exactly does "magnitude" mean in this context? To be clear, I think that this is quite an elegant measure, it just needs more motivation in the text.

R: Thank you for your feedback. Based on this and the comments of other reviewers we have recognized the need for a clearer explanation of "magnitude" in the context of ANI discontinuities, especially for those unfamiliar with microbial genomics. In our study, "magnitude" quantifies the extent of genetic divergence between microbial genomes, highlighting significant speciation events or ecological adaptations. To ensure clarity and accessibility, we have expanded the Background section to better explain ANI and the importance of measuring these discontinuities, thus helping readers appreciate the significance of our findings and the elegance of the measure.

1. The choice of thresholds used in the analysis should be explored further. Specifically, the definition of delta as the maximum genetic rate of change above 94%. While a justification is given, I think it is important to understand just how sensitive the results are to changes to the 94% threshold, AND the inclusion or exclusion of genomes through some form of jackknifing.

R: We acknowledge the need for a sensitivity analysis regarding our discontinuity estimates. In response, we have conducted a sensitivity analysis using jackknife resampling, excluding 20% of genomes randomly across 100 replicates. The results, showing averages and confidence intervals, will be included in the supplementary material.




While the jackknife averages sometimes exceed the observed metrics, there is strong correspondence, and all observed metrics fall within the confidence intervals.

Regarding the 94% threshold for delta, we chose this value given the existing practice in species delineation practices (Konstantinidis and Tiedje 2005; Jain et al. 2018). The 94%-96% ANI range is widely accepted in microbial genomics for defining species boundaries, supported by literature indicating that genomes from different species with ANI values between 90% and 95% are rare (Rodriguez-R et al. 2024). Figure 2b in our manuscript illustrates this concept effectively. However we recognize that this is based on historical practice in this field and, to address concerns, we have provided a more detailed justification for selecting the 94% threshold, ensuring our approach's transparency for readers unfamiliar with these conventions.

2. How sensitive is the measurement of discontinuity to choosing different bait nodes for each species analyzed? How was the actual bait node chosen?

R: Page 8, lines 14-19 describes the process of selecting representative genomes for each community: "*Representative genomes of each species were chosen based on the following criteria: (i) prior designation as representative in both GTDB and RefSeq databases, (ii) representative at least in GTDB, or (iii) fewest scaffolds if the previous criteria were not met. Ties were resolved by random selection. This yielded a subset of 261 representative genomes (t) used for subsequent analysis*".

Page 8, lines 43-44 we highlighted that each representative genome served as a bait: "*we employed an egocentric-based approach using representative genomes as "baits*" within the network $M$"

Regarding the sensitivity analysis, we utilized type strains as reference points. Type strains are defined as well-characterized isolates used as references for defining species in bacteriology. The List of Prokaryotic names with Standing in Nomenclature (LPSN) and the Genome Taxonomy Database (GTDB) provide lists of representative genomes for several bacterial

species. As stated in the main text, we only chose random (yet well assembled) representative genomes in case no information was available for a given community/species in our network.

3. The ML methods need to be better-motivated. Specifically, why were the 31 different metrics chosen as candidate predictors? The authors describe these in a general way on page 5, lines 32-46 and again on page 9 lines 35-42, but it's still not clear to me exactly what all these metrics were. Table S1 was referenced but not present in my copy of the manuscript. This might provide some clarification?

R: We appreciate the reviewer for highlighting that it could raise some questions from a reader. We apologize for the oversight - the relevant features are summarized in Table S2.

4. The quantile regression relating delta to pangenome openness should be described in the methods section.

R: Many thanks for pointing this out. The quantile regression is now described in the methods section.

*Are the conclusions adequately supported by the data shown? If not, please explain*

I believe the conclusions are largely supported by the data shown, but it is hard for me to evaluate this fully without more careful sensitivity analyses.

- *Are sufficient details provided to allow replication and comparison with related analyses that may have been performed? If not, please specify what is required*.

I thank the authors for making their code available on github with sufficient documentation. Given this, I don't have concerns with replicability.

- *Does the work represent a significant advance over previously published studies?*

Yes, in that quantitative measurements of the genetic discontinuity within species has been lacking and a huge gap in the field. This paper represents a significant step forward in exploring this important feature of microbial species and populations.

- *Is the paper of broad interest to others in the field, or of outstanding interest to a broad audience of biologists?*

Yes, I think that this work represents a really key advance in our ability to quantify genetic discontinuities that has potential applications beyond microbial genomics (e.g., quantifying genomic discontinuities along animal hybrid zones). However, there are a couple of areas where a broader audience could use some clarification.

1. The authors switch from describing boundaries between species to discontinuities between populations in the same paragraph, and I think clarifying this distinction would be helpful for most readers.

Ultimately, species boundaries represent discontinuities between distinct populations. We anticipate that early stages of speciation will manifest as such discontinuities. However, this paper focuses on the characteristics of established species rather than the processes of their formation. We have updated the manuscript to clarify this point.

2. At the beginning of page 3 with the discussion of G. vaginalis, I think it would be helpful to provide a little more detail on why the detect

R: We have now addressed these concerns in the main text. One language comment:

3. I would recommend scanning more thoroughly for typos. For instance, on page 8, line 34, I believe the sentence should read, "All single-copy genes were aligned with Mafft..."

R: Many thanks for such close reading and noticing that typo. We have changed the text accordingly. We have read over the text during response to reviewers and hope to have removed typos.

**References**


Caro-Quintero, Alejandro, and Konstantinos T. Konstantinidis. 2012. “Bacterial Species May Exist, Metagenomics Reveal.” *Environmental Microbiology* 14 (2): 347–55.

Ceres, Kristina M., Michael J. Stanhope, and Yrjö T. Gröhn. 2022. “A Critical Evaluation of Pangenomics, with Reference to Its Utility in Outbreak Investigation.” *Microbial Genomics* 8

(6). https://doi.org/10.1099/mgen.0.000839.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. “High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries.” *Nature Communications* 9 (1): 5114.

Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. “Genomic Insights That Advance the Species Definition for Prokaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (7): 2567–72.

McInerney, James O., Alan McNally, and Mary J. O’Connell. 2017. “Why Prokaryotes Have Pangenomes.” *Nature Microbiology* 2 (March): 17040.

Puente-Sánchez, Fernando, Matthias Hoetzinger, Moritz Buck, and Stefan Bertilsson. 2023. "Exploring Environmental Intra-Species Diversity through Non-Redundant Pangenome Assemblies." *Molecular Ecology Resources* 23 (7): 1724–36.

Rodriguez-R, Luis M., Roth E. Conrad, Tomeu Viver, Dorian J. Feistel, Blake G. Lindner, Stephanus N. Venter, Luis H. Orellana, Rudolf Amann, Ramon Rossello-Mora, and Konstantinos T. Konstantinidis. 2024. "An ANI Gap within Bacterial Species That Advances the Definitions of Intra-Species Units." *mBio* 15 (1): e0269623.

## Second round of review

### Reviewer 1

The revised article by Passarelli-Araujo et al., addressed all my previous major and minor concerns. I felt that the authors did a great job in responding to my comments as well as those of the other reviewers and I want to thank the authors for providing additional explanations and data in their rebuttal. I have only one follow-up question/suggestion for the authors to consider. If I understand this correctly, and the authors could make this clearer in their manuscript by adding an explicit definition, the pangenome saturation metric alpha ($\alpha$) is the slope of the number of non-redundant genes in the classical pangenome curves (i.e., the top line in the graph of Figure 3B). In the specific example we were discussing in the rebuttal, the B. cereus vs B. anthracis comparison, the authors say that 500 genomes of each of these species have ~49,000 and ~36000 unique, non-redundant genes, respectively. That means each B. cereus genome has, on average, about 100 unique genes that it contributes to the pangenome curve (i.e., 49000/500= ~100) vs. 70 genes for each B. anthracis genome. So, the curve should grow/increase faster on the y-axis for B. cereus compared to B. anthracis due to this, given also that both species start at the same -more or less- point on the y-axis because they have similar average genome sizes, right? Or what am I missing? If so, then how is the alpha metric so similar between the two, around 0.62? I was anticipating a lower alpha value for B. cereus compared to B. anthracis due to the higher number of genome-specific genes. Can the authors explain this and perhaps add a discussion in the text for a more complete understanding of the results presented?

### Reviewer 2

The authors have addressed all my concerns.

### Reviewer 3

I am satisfied with the authors' revisions and particularly commend their additional sensitivity analyses and clarifications. I believe the manuscript should be accepted.

### Authors' response to reviewers

Regarding the reviewer's remarks on pangenome saturation and the growth rate of pangenome curves for Bacillus cereus and Bacillus anthracis, we confirm that pangenome saturation, represented by alpha, is indeed calculated as the slope of non-redundant gene families added with the sequencing of additional genomes. However, a key clarification is necessary regarding the reviewer's note on the gene counts: the assertion "500 genomes of each species have approximately 49,000 and 36,000 unique, non-redundant genes" is incorrect. Those numbers refer to the total pangenome size, not to unique genes. The actual counts of unique genes (or

cloud genes) are 9,050 for B. anthracis and 14,319 for B. cereus, detailed in Table S1.

Furthermore, the expectation that “B. cereus would have a lower alpha value than B. anthracis, due to a higher number of genome-specific genes”, is supported by our results. Although the difference is subtle, our data confirm this assumption: B. cereus exhibits a saturation coefficient of 0.64, which is lower than the 0.66 observed for B. anthracis.

We have chosen not to delve deeply into the mathematical estimation of pangenome saturation estimation within this manuscript to maintain the focus on estimating genetic discontinuity. The aim is to present saturation and fluidity in a manner that correlates with bacterial lifestyle and contrasts genetic discontinuities among significantly different bacteria, such as Mycobacterium tuberculosis and B. cereus. We believe this approach prevents potential confusion and keeps the manuscript streamlined, focusing on essential information for understanding the core theme of our study. In summary, all of our results presented align with the reviewer’s observations.