Response Letter

(Manuscript: Revisiting pangenome openness with k-mers)

Luca Parmigiani, Roland Wittler and Jens Stoye

Response to Recommender

Comment 1: The authors did a very good job incorporating most comments of the riviewers. Unfortunately, the main comment has not yet been resolved sufficiently. Although I do believe that the added text has improved the paper, I agree with the reviewer that more work needs to be done.

Response: We express our sincere gratitude to the Recommender and Reviewers for their professional comments and helpful suggestions. We have carefully read the comments and revised the manuscript accordingly.

Comment 2: For example, in the introduction you state that "one of the most outstanding discoveries at the time was that some species possess an open pangenome and others a closed pangenome" but then you use a model in which it is impossible for a pangenome to be closed.

The other remaining problem is that you test your method only on open pangenomes. Of course this makes sense because it cannot work for closed pangenomes, but then this weakness of the method should at least be stated clearly.

The reviewer gives a very nice suggestion for how to resolve these issues. I recommend to follow this suggestion if possible.

Response: In our model, we already account for closed pangenomes, since we fit $K_2m^{-\alpha}$ on f_{new}, which allows for both open ($\alpha < 1$) and closed ($\alpha > 1$) outcomes. As you rightly observed, this approach would not apply if we had fitted K_1m^{γ} on f_{tot}. We acknowledge that our original mathematical definition led to a contradiction. We have updated the definition according to Reviewer #1's suggestion.

Additionally, we ran Pangrowth on two new datasets, one composed of a closed virus pangenome and the second of a marginally closed bacterial pangenome.

Response to Reviewer #1

Comment 1: The manuscript "Revisiting pangenome openness with k-mers" describes a method to estimate how "open" a pan-genome is, that is, to estimate whether the number of novel sequences expected as more genomes are sequence will keep growing to infinity or is bounded. The method proposed uses k-mers rather than say genes or open reading frames, and is shown to have good correlation with previous methods. Although the computational method to efficiently compute the openness (the parameter α), the mathematical background is still flawed, and the extra text added to the manuscript since the first revision does not properly address that issue.

Major issues

1) Adding that "the traditional concept of open and closed pangenome may be mathematically flawed" (line 171, page 5), does not properly address the issue. It is well understood that models are only approximation of the phenomena they represent, nevertheless these models need to be at least intrinsically coherent to draw any conclusion. Presenting a mathematically flawed model with an admission that the model is flawed is not a way to resolve the issue or design software methods. I would agree that the description in Tettelin et al., on which this manuscript is based, is also (very) confusing. Regardless, the method in this manuscript should be correct. As it is described currently, the condition $\alpha > 1$ implies that f_{tot} is decreasing with a limit of 0. This cannot model the size of a union of sets of elements as the union is necessarily increasing in size. As such, the distinction between open and closed genomes is vacuous as no pan-genome can satisfy the close definition. Maybe surprisingly, it is only the presentation of the model that needs fixing, the method itself seems correct. As I understand it, the definitions are as follow. It is the growth of the number of elements (be it k-mers, genes, etc.) that follows a power law. The number of elements is a power law plus a possible constant.

(In the following, all the additive constants are named C, even though they might not be all equal. There actual values are not important for the exposition.) That is:

- For open genomes: $f_o(m) = C + K_1 m^\gamma$, with $\gamma > 0$, ie, a constant plus an increasing power law. The derivate is $f_0'(m) = \gamma K_1 m^{\gamma-1}$, and it is positive for all m. $f_o(m)$ grows to infinity as m grows
- For close genomes, $f_c'(m) = C - K_1 m^\gamma$, with $\gamma < 0$, ie, a constant minus a decreasing power law. The derivative is $f_c'(m) = -\gamma K_1 m^{\gamma-1}$, and it is also positive for all m. $f_c(m)$ grows to C as m grows.

Both derivate have the form $K_2 m^{-\alpha}$ with $K_2 > 0$ and $\alpha = 1 - \gamma$. And the value of the exponent α(ie, < 1 or > 1) determines the openness of the pangenome. Conversely, define

$$f_{tot}(m) = \int_{m_0}^{m} K_2 x^{-\alpha} dx = C + K_2 m^{1-\alpha}/(1-\alpha)$$

Then, if $\alpha < 1$, f_{tot} has the same form as f_o(i.e., a constant plus an increasing power law), and if $\alpha > 1$, f_{tot} has the same form as fc (a constant minus a decreasing power law) Note that the constant C in f_{tot} depends on the starting point m_0, which properly models what is actually done in practice in this method (e.g., line 264, page 10).

Response: Thank you very much for carefully reviewing our manuscript. This is a great suggestion, that we have not seen in other research on pangenome openness. We changed the original definition with the help of the one you provided. In practice, none of the results were affected, since all the results were based on the fitting of $K_2 m^{-\alpha}$ on f_{new} (except for Figure 1, where we fit on f_{tot}, but this does not change any result in the manuscript).

We believe this is a significant contribution to the paper, worthy of co-authorship if you would agree. Alternatively, with your permission, we would be glad to express our thanks to you by name in the acknowledgments.

Comment 2: 2) There are no examples of closed genomes in the evaluation. There should be.

Response: In our analysis, we did not deliberately exclude closed pangenomes, however, we did not find any that matched the criteria for being "closed" as defined by Tettelin et al. (2008) [1].

We now included a comparison of Pangrowth with two new datasets: a closed virus pangenome and a marginally closed bacterial pangenome. We compared these results with their respective publications. In both cases, Pangrowth's findings were consistent with the values of α reported there.

Furthermore, we have added a new subsection titled "Closeness" to address the inclusion of the new datasets and to outline the challenges encountered in identifying mathematically closed pangenomes. The identification was challenging due to several reasons: first, the use of different fitting methods, combined with the lack of a goodness-of-fit measure, can yield varying conclusions. Second, the decision to include or exclude certain genomes from the dataset can lead to reclassification, complicating definitive categorization of pangenomes. Lastly, some species were identified as having closed pangenomes due to their highly similar genomes, even when reported with α values marginally below one.

Response to Reviewer #2

Comment 1: I support this article Accept

Response: Thank you for your support.

References

[1] H. Tettelin, D. Riley, C. Cattuto, and D. Medini. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5):472–477, 2008.