Response Letter

(Manuscript: Revisiting pangenome openness with k-mers)

Luca Parmigiani, Roland Wittler and Jens Stoye

Response to Recommender

Comment 1: The problem and approach I find very interesting, but the reviewers have a number of important questions that need to be answered first.

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

Comment 2: Justify the used definitions of open/closed genomes and explain the practical relevance of the results based on this definition;

Response: The definition of an open/closed pangenome is, in essence, historical. This definition aligns with the one utilized in the follow-up paper by Tettelin et al. [2, 3], where they propose that fitting to Heaps' law provides a superior fit compared to employing an exponential decaying function. Since then, it has been used several times.

The application of Heaps' law stems from the estimation of the pangenome size, which can be compared to a broad category of problems. Specifically, we are given a collection of objects (in this case, genomes) each containing certain items (k-mers, genes, etc.), and we want to estimate how many more unseen items are present in the object set.

Though empirical data do not fit perfectly, Heaps' law appears to serve as a strong heuristic. It helps us estimate how many more genomes we might sample, and it provides some insights into unseen genomic data. Presently, no superior function seems to be available.

In response to this feedback, we have provided a more detailed justification of the definition in our text.

Comment 3: Make the supplementary figures and tables available;

Response: See our response to Comment 2 by Reviewer #1.

Comment 4: Explain why the blue line in Figure 1 does not fit the data;

Response: See our response to Comment 3 by Reviewer #1.

Comment 5: Analyse how the practical running time depends on the number of samples;

Response: See our response to Comment 1 by Reviewer #2.

Comment 6: Analyse the distribution of alpha values in the experiments;

Response: Thank you for this suggestion. We added the distribution of the alpha values.

Comment 7: Explain why the method was compared only to Roary and Pantools.

Response: Thank you for bringing this up. See our response to Comment 3 by Reviewer #2.

Response to Reviewer #1

Comment 1: In "Revisiting pangenome openness with k-mers" the authors give a computational method and an implementation to estimate how "open" a pan-genome is, that is whether the genome of a species has many variant genes (opened) or is more constrained (closed). This is traditionally done by comparing gene content of different individual bacteria of a species, but is done here using k-mer content instead.

Although the proposed computational method seem correct, the definition of open/close pan-genome raises questions. Consequently the conclusions drawn from the experiments are affected by the flaw in the definition.

Page 4, line 153: it is stated that $0 \leq \gamma \leq 1$ and $\alpha = 1 - \gamma$ (hence $0 \leq \alpha \leq 1$ as well). Then line 158, the definition of a close genome is for $\alpha > 1$, which cannot happen by definition. A close genome would imply $\gamma < 0$, that is the number of k-mer seen would be a decreasing function of m (m = number of genomes considered). This simply cannot be observed.
Unsurprisingly, all the values reported by the proposed method (see Fig. 4) have an $\alpha < 1$ and are all declared to be open genomes. That is not an empirical conclusion based on data, but a mathematical guarantee independent of the data.

Response: Thank you very much for carefully reviewing our manuscript and providing us with professional comments and helpful suggestions. We concur that the initial mathematical definition of open and closed was not correctly defined. As you rightly pointed out, it is not feasible to obtain a closed pangenome.

In practical terms, this situation can arise where the function follows Heaps' law only for larger m_0 values (e.g., $m_0 \geq 6$), but fails to do so for smaller ones. Given that the mathematical definition of open and closed is ill-defined, throughout the paper, we have focused on discussing the "openness" of the pangenome, while using the term "closeness" only in its historical context.

Although Heaps' law may not be the ideal fit, it is widely adopted within the community. The goal of our paper was not to validate Heaps' law, but to illustrate its equal applicability for k-mers and genes, and its tendency to give comparable values.

We underscored in our revised paper the incorrectness of the definition. The use of Heaps' law is, in part, due to the lack of more suitable functions.

Comment 2: The Supplementary material does not seem to be available, even though it contains important figures.

Response: We sincerely apologize for the inconvenience. The supplementary material was available in version 1 (as can be seen from the history of the manuscript from *bioRxiv*) but was mistakenly not re-uploaded in the second version. The current version of the manuscript contains now most of the previous supplementary material as part of the main text. Additionally, a new supplementary material was produced, containing a long list of figures, which can be accessed via link directly from the paper.

Comment 3: Fig 1 page 5: the fitting of the blue line does not seem to match the data. The conclusion that $\alpha = 0.98$ for this data set is questionable. It seems like this data does not follow Heaps' law. Maybe the fact that this data does not follow Heaps' law is the signature of a closed genome?

Response: Thank you for pointing this out. We acknowledge that the curve for *Yersinia pestis* does not strictly follow Heaps' law across all points but adheres to the power-law distribution predominantly at the tail (with $m_0 = 6, R^2 \approx 0.99$), which is a common phenomenon in power-law distributions, as noted in the literature [1]. In response to this comment, we have included an image for each tool to illustrate how the fit improves when the fitting is done starting at larger points, reinforcing that the data follows Heaps' law (Figures S1-S4).

It is worth mentioning that this behavior is not unique to the k-mer approach. We observed that also *Rhodopseudomonas palustris* does not fit Heaps' law but only for the gene-based approaches. Unfortunately the dataset consists only of 8 genomes, which limits the possibility of starting to fit later.

In light of this comment, we have reevaluated our choice of using *Streptococcus pneumoniae* and *Y. pestis* to illustrate the concept of different types of openness, since it raises more questions than it helps to explain. We decided instead for *Helicobacter pylori* and *Campylobacter jejuni*. These pangenomes were selected because they have almost identical genome sizes and similar numbers of genes, and they comprise the same number of genomes. Interestingly, *H. pylori* shows two to three times more items (both k-mers and genes) than *C. jejuni*, reflecting a different amount of richness in the sequence.

Comment 4: The use of GMP to compute f_{tot} is not well justified. The ratio $(n-i)^{m}/n^{m}$ (where ^{m} is the falling factorial as in the text) probably doesn't need an infinite precision library. It is the product of the ratios $(n-i-j)/(n-j)$ for $0 \leq j < m$. These ratios and their product can most likely be stored in double floats without significant loss in precision (and is likely cheaper to compute).

Response: Thank you for this helpful suggestion. Our implementation using a double was able to report the correct value for most cases but it failed when the number of genomes in the pangenome increases (it started reporting wrong values around 2000 genomes). Based on this comment, we changed the implementation to compute the pangenome growth in log space, storing the value $\log((n-i)^m/n^m)$. The new implementation was updated in the Gitlab repository and the dependency of GMP was removed.

Comment 5: There is no timing or memory usage information given for the bacterial experiments.

Response: Thank you for this suggestion. We added in the manuscript the time and memory usage and we discussed them.

Comment 6: GNU should be capitalized (it is an accronym)

Response: Thank you for carefully checking our manuscript. Since the dependecy from GMP was removed we also removed the related sentence.

Comment 7: Page 9, line 274: "making k-mers more suitable" is ambiguous. k-mers are more suitable for bacterial genomes or eukariotic genomes?

Response: Thank you for the question. The k-mers in this case are more suitable for eukaryotic genomes since we are losing a huge amount of potentially relevant variants present in non-coding DNA regions or other non-annotated portions of the DNA. We made it more clear in the text.

Response to Reviewer #2

Comment 1: Parmigiani et al. used k-mers to estimate pan-genome openness. It's a nice idea, but also challenging work. I have some small questions:
Based on the different numbers of samples (10, 20, 50, 100), what is the running time of this algorithm?

Response: Thank you very much for carefully reviewing our manuscript and providing us with professional comments and helpful suggestions to improve the manuscript. We have now evaluated our algorithm on 10, 20, 50, 100, 1000, 2000, 4000, and 8000 *Escherichia coli* samples, and reported the running time and maximum space usage in the revised version of the paper (Table 3). We separately detail the measurements for the two stages of our tool: i) generation of the histogram h, and ii) calculation of the pangenome growth using this histogram. Despite the quadratic complexity of the second step, we show that, given the histogram h of 8000 *E. coli* samples, the pangenome growth can be calculated in under a second, requiring minimal memory (less than 5MB).

Comment 2: The author tested twelve bacterial species, how many strains were tested for each species?

Response: Thank you for this suggestion. In the Supplementary material we now have a table showing the number of genomes/strains for each species and reference it in the paper.

Comment 3: In addition to Roary and Pantools, should it be compared with other software?

Response: Roary and Pantools were used for the following reasons: Roary is one of the highest cited pangenomics tool that performs the openness analysis, while Pantools is well documented, easy to install and can perform directly the openness analysis, too. We also considered other possible tools and provided reasons for not including them in our study. These varies between not performing directly the openness analysis, to being too slow to run of our dataset. Moreover, we decided to additionally compare to BPGA as its speed and capability to estimate the openness made it a good candidate. We have now elaborated on our choice of tools in the manuscript (Section "Related work").

Comment 4: The author compares the sensitivity of different k-mers to the pan-genome openness estimation. Compared with other software, what is the distribution of α values for the twelve different species under different k-mers?

Response: Thank you for the suggestion. In Figure 4(diagonal), we added histograms showing the distribution of α values for both k-mers and gene-based approaches.

Response to Reviewer #3

Comment 1: The mathematical equation were well explained and it was articulated. I endorse this and say it should be accepted.

Thank you

Response: Thank you for the supportive feedback.

References

[1] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, Nov. 2009.

[2] H. Tettelin, V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, I. M. y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pangenome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955, 2005.

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