Peer Community In
Genomics

A quick alternative method for resolving bacterial taxonomy using short identical DNA sequences in genomes or metagenomes

B. Jesse Shapiro based on reviews by Gavin Douglas and 1 anonymous reviewer

Open Access

A recommendation of:
M Briand, M Bouzid, G Hunault, M Legeay, M Fischer-Le Saux, M Barret. **A rapid and simple method for assessing and representing genome sequence relatedness** *(2020), bioRxiv, 569640, ver. 5 peer-reviewed and recommended by Peer Community In Genomics.* 10.1101/569640

Published: 24 Sept 2020

Submitted: 07 November 2019, Recommended: 15 September 2020
Cite this recommendation as:
B. Jesse Shapiro (2020) The bacterial species problem can be summarized as follows: bacteria recombine too little, and yet too much. *Peer Community in Genomics, 100001.* 10.24072/pci.genomics.100001

The bacterial species problem can be summarized as follows: bacteria recombine too little, and yet too much (Shapiro 2019).

Too little in the sense that recombination is not obligately coupled with reproduction, as in sexual eukaryotes. So the Biological Species Concept (BSC) of reproductive isolation does not strictly apply to clonally reproducing organisms like bacteria. Too much in the sense that genetic exchange can occur promiscuously across species (or even Domains), potentially obscuring species boundaries.
In parallel to such theoretical considerations, several research groups have taken more pragmatic approaches to defining bacterial species based on sequence similarity cutoffs, such as genome-wide average nucleotide identity (ANI). At a cutoff above 95% ANI, genomes are considered to come from the same species. While this cutoff may appear arbitrary, a discontinuity around 95% in the distribution of ANI values has been argued to provide a 'natural' cutoff (Jain et al. 2018). This discontinuity has been criticized as being an artefact of various biases in genome databases (Murray, Gao, and Wu 2020), but appears to be a general feature of relatively unbiased metagenome-assembled genomes as well (Olm et al. 2020). The 95% cutoff has been suggested to represent a barrier to homologous recombination (Olm et al. 2020), although clusters of genetic exchange consistent with BSC-like species are observed at much finer identity cutoffs (Shapiro 2019; Arevalo et al. 2019).

Although 95% ANI is the most widely used genomic standard for species delimitation, it is by no means the only plausible approach. In particular, tracts of identical DNA provide evidence for recent genetic exchange, which in turn helps define BSC-like clusters of genomes (Arevalo et al. 2019). In this spirit, Briand et al. (2020) introduce a genome-clustering method based on the number of shared identical DNA sequences of length *k* (or *k*-mers). Using a test dataset of *Pseudomona*s genomes, they find that 95% ANI corresponds to approximately 50% of shared 15-mers. Applying this cutoff yields 350 *Pseudomonas*species, whereas the current taxonomy only includes 207 recognized species. To determine whether splitting the genus into a greater number of species is at all useful, they compare their new classification scheme to the traditional one in terms of the ability to taxonomically classify metagenomic sequencing reads from three *Pseudomonas*-rich environments. In all cases, the new scheme (termed K-IS for "Kinship relationships Identification with Shared *k*-mers") yielded a higher number of classified reads, with an average improvement of 1.4-fold. This is important because increasing the number of genome sequences in a reference database – without consistent taxonomic annotation of these genomes – paradoxically leads to fewer classified metagenomic reads. Thus a rapid, automated taxonomy such as the one proposed here offers an opportunity to more fully harness the information from metagenomes.

KI-S is also fast to run, so it is feasible to test several values of k and quickly visualize the clustering using an interactive, zoomable circle-packing display (that resembles a cross-section of densely packed, three-dimensional dendrogram). This interface allows the rapid flagging of misidentified species, or understudied species with few sequenced representatives as targets for future study. Hopefully these initial *Pseudomonas* results will inspire future studies to apply the method to additional taxa, and to further characterize the relationship between ANI and shared identical *k*-mers. Ultimately, I hope that such investigations will resolve the issue of whether or not there is a 'natural' discontinuity for bacterial species, and what evolutionary forces maintain this cutoff.

References

Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF (2019) A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. Cell, 178, 820-834.e14. https://doi.org/10.1016/j.cell.2019.06.033

Briand M, Bouzid M, Hunault G, Legeay M, Saux MF-L, Barret M (2020) A rapid and simple method for assessing and representing genome sequence relatedness. bioRxiv, 569640, ver. 5 peer-reveiwed and recommended by PCI Genomics. https://doi.org/10.1101/569640

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications, 9, 5114. https://doi.org/10.1038/s41467-018-07641-9

Murray CS, Gao Y, Wu M (2020) There is no evidence of a universal genetic boundary among microbial species. bioRxiv, 2020.07.27.223511. https://doi.org/10.1101/2020.07.27.223511

Olm MR, Crits-Christoph A, Diamond S, Lavy A, Carnevali PBM, Banfield JF (2020) Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. mSystems, 5. https://doi.org/10.1128/mSystems.00731-19

Shapiro BJ (2019) What Microbial Population Genomics Has Taught Us About Speciation. In: Population Genomics: Microorganisms Population Genomics. (eds Polz MF, Rajora OP), pp. 31–47. Springer International Publishing, Cham. https://doi.org/10.1007/13836*2018*10

Reviewed by [Gavin Douglas](), 2020-08-25 13:11

The authors have addressed all of my comments - thanks.

Revision round #2

2020-08-03
Both reviewers appreciated the improvements made to this revised manuscripts, but suggested a few remaining revisions before the manuscript can be recommended. I trust you will be able to address these comments without too much additional work, since they mainly involve clarification and software availability.

Preprint DOI: https://doi.org/10.1101/569640

Reviewed by anonymous reviewer, 2020-07-31 13:52

Overall I believe that this manuscript has improved, and the authors are more fair in their comparison of modern methods. Other choices for benchmarking could have been chosen, but the authors have justified their tests.

My main remaining issue is on the availability of the software. While the source code is now available (after some effort, but available nonetheless), the galaxy wrapper is not. I would urge the authors to either open access to the galaxy instance linked, or put their code on the galaxy shed so that other galaxy users can most easily benefit from their work.

Reviewed by [Gavin Douglas](), 2020-07-16 17:07

The authors have addressed most of my concerns, but there are two important points I would like to see addressed to improve the manuscript.

Major comments:

First, thanks for expanding on the CLARK analysis in your response, but I still have a few concerns. Does the % classified reads that you refer to refer to just at the species level or in general? It would make sense if it corresponded to just the species level based on your explanation. In either case, a clearer discussion of this result is needed. From my understanding, this result is a proof-of-concept that clustering genomes into clusters before running taxonomic assignment can improve classification. It could be argued that this is circular because the clusters are based on shared k-mers, which is also what CLARK bases classification on: if taxa are defined based on shared k-mers then it would always be expected for a k-mer-based classification approach to classify taxa with more resolution. I think the authors should emphasize this k-mer connection between CLARK and the clustering approach and also clearly state that they can only hypothesize that a higher proportion of reads are being *correctly* classified with their workflow (and that it hasn't actually been demonstrated). Currently the authors discuss this result as showing a clear benefit to microbial ecology in general, which I think would be very misleading to readers.

Secondly, on page 10 the authors state: "Moreover, KI-S includes a friendly visualization interface that could help systematicians to curate whole genome databases.". I was able to get access to the authors' galaxy server to try out the tool thanks to their quick reply to my email and I found it straight-forward to use.

However, it wasn't clear to me whether any reader in general would be able to get an account on this server. Based on advertising the link in the manuscript I'm guessing this is true, but this should be clarified either way. If not, then users will need more documentation on how they can use the KI-S code to setup the visualization workflow themselves. I did not find it straight-forward to download the source code and it looks like the only documentation for the source code (the README.md file in the GitHub repository) is in French, which would need to be translated for an English-reading audience. Specifically, looking into this README it appears that the key circle packing visualization step is performed by the `generate_packing.pl` Perl script. Details on how to prepare the input and look at the output of this script is needed.

Minor comments:

Hierarchical clustering is mentioned later on, but it would help readers evaluate the method to know the specific details on how this clustering was performed with the custom R script when it is described in the methods.

Minor typo on Pg 4: "was first evaluate" should be "was first evaluated"

Author's reply:

Download author's reply (PDF file)

Revision round #1

2019-12-19
The manuscript has now been seen by two reviewers, who both see potential in the work but both raised concerns about precisely what the new method brings, and how it compares to other methods (e.g. FastANI). Perhaps the major contribution of the new method lies in the visualization, in which case this part should be expanded. The reviewers also have several specific comments that should be addressed in a revised manuscript.

Preprint DOI: https://doi.org/10.1101/569640

Reviewed by anonymous reviewer, 2019-11-25 18:12

In this short manuscript Briand et al describe a workflow which uses k-mer indexing software to compare bacterial genomes. This method generates a similarity measure which is comparable to ANI. They go on to use these relatedness measures to cluster genomes at various thresholds, produce a visualisation of these clusters, and test the use of these clusters in metagenomic read classification. This workflow is deployed on a galaxy server.

Overall the methods in the manuscript appear to be sound, as they are mostly based on previously published work. Though the novelty of the algorithm is limited, the implementation and pipeline, being on galaxy, may well be useful to researchers who are more comfortable with a graphical user interface than the command line. This server requires username and password to use, so I was unable to test any of this software myself. Nor was the implementation available on github (or similar), or the galaxy shed, meaning no-one else can use it. This severely limited my ability to review this aspect of the manuscript.

I also had a number of serious issues with the presentation of the work: 1) The comparison with PYANI is not really appropriate. The authors used Simka, which by my understanding is a k-mer indexing package, so is unsurprisingly orders of magnitude faster than nucleotide alignment with mummer and blast. A more modern comparison would be with either other k-mer indexes, or sketch based approaches such as fastANI. These approaches have been around for a number of years, and are the standard now used in this field. 2) There is not enough description of the methods, and code is also needed. Describing briefly how components work, what they do and why parameters values were chosen all need to be added. I was not able to find information on simka without following references, and this is an integral part of the method. The difference in how simka and other potential methods work needs to be explained, and why this is expected to lead to large differences in computation time. Likewise, the section on metagenomic read sets needs further description. (What is Clark and how does it work? Why does adding further classifications in helps classify more reads?) 3) The results also lack context. It was difficult to understand what problems were being solved by the presented method, and how much of the method is new compared to e.g. fastANI. The first section of the results was mostly a technical methods description, finding similar results on k-mer size as has been previously reported. How did the original Clark database and the newly assigned genome sequences differ in classification, and why exactly did this change the number of reads that could be assigned? How does this relate to the broader issue of misclassification and missing identifiers in RefSeq, which has been noted previously (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1554-6)? More explicit example use cases could be added. In figure 3, describing how to read the figure and e.g. identify misclassiciations would be useful. 4) More care needs to be taken with some of the species and genus name terminology. Particularly, the words 'strain' and 'clique' kept appearing without definition. How do these terms relate to species and genus level differences? 5) Why was the Pseudomonas dataset used? What was the original species classifications breakdown, and how (quantitatively) did this compare with the reclassification? Is this one example sufficient? Other fast distance estimators have been run on all of RefSeq.

Reviewed by Gavin Douglas, 2019-11-22 19:11

Briand et al. describe a new approach for computing inter-genome relatedness based on the percentage of shared kmers. The main motivation for this project is that the computation time for computing many relatedness metrics, like the average nucleotide identity, can be prohibitive for many pairwise genome comparisons. I think this tool could be valuable for the field, especially after addressing a few issues which I think currently make the benefits of the tool difficult to evaluate (see below). In particular, I think the authors' tool could be great for running quality control on taxonomy assignments in genome databases. This quality control can be run using an interactive approach for visualizing genome relatedness that the authors have implemented, which I think could be used for quickly spotting problematic taxonomic assignments.

Major comments

I think it would be important to clarify how the results of KI-S clustering in practice differ from other similar tools. One tool in particular is Mash, which was published in 2016. This tool can be used to rapidly compute distances between genomes after performing dimension reduction based on kmer counts. The motivation for Mash was to speed up calculations of inter-genome (and sequences in general) distances. In the Mash paper the authors describe their approach as comparable to ANI while being much faster and so I think it would be important to directly compare to Mash in terms of both the compute time and results. If the authors don't agree that Mash is a comparable tool then this should be explained.

A related issue is that it currently is not clear whether using the % of shared kmers results in comparable genome clusters to existing approaches like average nucleotide identity. It seems like this would likely be the case, but I think this is important for the authors to clearly describe either way so that users can better evaluate the tool.

I also do not agree that they have shown evidence that their approach can be used to improve the taxonomic classification of metagenomics samples. This analysis focused on the percentage of classified reads, which cannot alone be used to evaluate how well a taxonomic classification performed.

Other comments

- The way KI-S is mentioned in the abstract makes it seems like it is a pre-existing tool, but on page 5 it sounds like the authors developed it from scratch – this should be clarified. Also, it is unclear whether all the steps like running Simka and the custom R script are run by KI-S itself. Lastly, it would be good to state what KI-S stands for, which I may have missed.
- P3,L45 – I recommend re-wording to make the first few sentences of the Background a little clearer. In particular, it reads like specifically Bacteria vs Archaea are the taxonomic groups being delineated, rather than prokaryotic species in general
- Figure 1 – Axis labels are needed, which might be easiest to do if fewer panels were shown. In particular, it seems like K15-K20 are extremely similar so maybe a couple could be removed. It is also not clear to me from the figure legend what "the number of values by class in the subset of 934 Pseudomonas genomic comparison" refers to on the y-axis. I think this is the ANI / % shared kmers for every pairwise comparison of Pseudomonas genomes, but I think this could be clarified either way.
- When describing the % overlapping species in each peak in figure 1 – how were the cut-offs for which data points to include in each peak decided (e.g. what cut-offs of % shared kmers were used to call data points in peak 2?)
- P7 – The authors imply that using 15-mers is the best or at least equally good as higher kmer values. This decision is discussed in the discussion, but I think it would be useful to explicitly mention this decision here (esp. when contrasting the 15-mer and 20-mer comparisons for instance) – perhaps at the end of paragraph 1 of the results.
- P7,L132 – I think the paragraph starting with "Fifty percent of 15-mers is close to ANIb value of 0.95" would benefit by making it clear what the goal of these analyses were, possibly with something like this: "We next investigated what percentage of shared kmers corresponds to an ANIb value of 0.95, which is a common cut-off for delineating species".
- P7,L145-147 – I am not sure what the sentence starting with "In addition, 15-mers allows the investigation of inter and intra-specific…" refers to and I think this should be clarified. One possible way to make the authors' point clearer might be to contrast why they think this is true specifically for 15-mers and not the 10 or 20-mer distributions also shown in Fig 2.
- P7,L149 – do these run times correspond to running the jobs on a single core? It would be useful to mention the memory usage as well if that's possible.
- P8,L158-159 – "185 cliques were composed of a single genome sequences, therefore highlighting the high Pseudomonas strain diversity" – an alternative explanation would be that KS-I is incorrectly calling those genomes as individual cliques. If there are species (and strain) names for all genomes then that would be one way to evaluate whether these genomes are expected to be in different cliques or not.
- On a related note to the above it would be useful to compare the cliques identified based on KS-I compared to ANI-b – based on Fig 2 it looks like they would be extremely similar, but I do not think that is clear from the main text.

- P8,L159 – I think using estimates of Chao1 alpha-diversity to estimate the expected number of Pseudomonas clusters would only make sense if you're considering genome cliques in a single environment (and at a particular time). I do not think the numbers of singleton and doubleton genomes in NCBI can really tell you about how many more Pseudomonas genome clusters are out there in general, if only because many Pseudomonas habitats have not been sampled.
- Fig 3 – I really like the zoomable circle packing representation of the data – this seems like a great way to summarize the relationships between many genomes. It is not clear to me how novel this visualization approach is, but if the authors believe that it is novel then I would emphasize that more in the introduction and discussion.
- P8 – I am not familiar with the term "clique" – maybe "cluster" would be clearer?
- P8 – It's not clear to me why changing the taxonomic label of the Pseudomonas genomes added to the database results in a higher proportion of classified reads. Is this because the CLARK algorithm tends not to collapse taxonomy to higher ranks if reads map to genomes associated with different species? If so, that is surprising to me, but I am not sure why else there would be a different in the proportion of classified reads. It would be useful to briefly explain why the authors think this is occurring. Unless I am missing something I also do not think this would make a difference for most metagenomics taxonomic classifiers like centrifuge, kraken2, and MEGAN
- P11,L221-223 – I think these are great examples for how this tool could be used to clean up and perform quality control on taxonomy assignments in genome databases
- P11, L229 – end – As mentioned above I do not fully follow why more reads were classified with CLARK after changing the taxonomic labels of the Pseudomonas genomes, but either way I do not think this is evidence that the taxonomic classifier is actually working better as indicated in the concluding paragraph currently. I think some sort of validation would be needed to be able to state that first organizing the genomes into cliques actually improves taxonomic classification. This is difficult to do because we almost never know the right answer in microbiome datasets. However, one potential way to do this would be to create a simulate dataset enriched for Pseudomonas (ideally with metagenome-assembled genomes from the seed datasets) and then compare the relative abundances of the taxa inferred using the 3 approaches mentioned in Fig 4 to the expected relative abundances.

Example of grammatical errors

Lastly, there are numerous grammatical errors throughout the manuscript – I have made a non-exhaustive list of example errors and possible, which hopefully will be useful for the authors.

- P2,L29-30: "…datasets composed of thousand genome sequences" change to "datasets composed of thousands of genome sequences".
- P2,L31 – "kmers counts" should be "kmer counts"
- P3,L64 – "for one pair of genome sequence" should be "one pair of genome sequences"
- P4,L71 – "classifiers differ in term" should be "classifiers differ in terms"
- P4,L72-73 – "for affiliating read to a" should be "for affiliating a read to a taxonomic rank"
- P4,L81 – add "the" before "relatedness"
- P5,L100 – should be "were selected" instead of "was selected"
- P6,L116 – need to add either "the" or "a" in front of "common bean" depending on which is correct

- P7,L143 – "Fifty percent of 15-mers is close to ANIb value of 0.95" should be "Fifty percent of 15-mers are close to an ANIb value of 0.95".
- P7,L153 – "used to investigate relatedness" should be "used to investigate the relatedness"
- P10,L188 – "prohibited its used for comparing" should be "prohibit its use for comparing"
- P11,L216 - "based ANIb" should be "based on ANIb"
- P11,L219 – "Moreover, KI-S tool, provides…" should be re-written, perhaps as "Moreover, KI-S includes…"

Author's reply:

Download author's reply (PDF file)