skDER: microbial genome dereplication approaches for comparative and metagenomic applications

Rauf Salamzade[1,2], Lindsay R. Kalan[1,3]

[1]Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA
[2]Microbiology Doctoral Training Program, University of Wisconsin-Madison, Madison, WI, USA
[3]M.G. DeGroote Institute for Infectious Disease Research, David Braley Centre for Antibiotic Discovery, Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada

Abstract

skDER (https://github.com/raufs/skDER) combines recent advances to efficiently estimate average nucleotide identity (ANI) between thousands of microbial genomes by skani[1] with two low-memory methods for genomic dereplication. The first method implements a dynamic algorithm to determine a concise set of representative genomes. This approach is well-suited for selecting reference genomes to align metagenomic reads onto for tracking strain presence across related microbiome samples. This is because fewer representative genomes should alleviate the concern that reads belonging to the same strain get falsely partitioned across closely related genomes. The other method, which uses a greedy approach, is better suited for use in comparative genomics, where users might be overwhelmed with the high number of genomes available for certain taxa and aim to reduce redundancy and, therefore, computational requirements for downstream analytics. This method selects a larger number of representative genomes to comprehensively sample the pangenome space for the taxon of interest. To further aid usage for comparative genomics studies, skDER also features an option to automatically download genomes classified as a particular species or genus in the Genome Taxonomy Database[2–4] and we provide precomputed representative genomes for commonly studied bacterial taxa[5].

Statement of Need

The boundaries of bacterial species are commonly defined using ANI-based thresholds [6–8]. For delineation of more resolute clusters of strains or lineages within species and the selection of individual genomes to serve as representatives, the use of phylogenetic [9], alignment [10,11], and k-mer based [12] approaches have been used as alternatives to ANI-based methods. While these methods have demonstrated scalability to species featuring thousands of genomes, they typically involve multi-step workflows, such as preliminary construction of core genome sequence alignments or phylogenies. In contrast, an ANI-cutoff based approach is more direct to use and simply takes genomic assemblies as input.

For broad comparative genomic investigations of a species or genus, highly sequenced lineages can skew evolutionary statistics [13], especially if such lineages are associated with a single environment, such as hospitals [14]. Thus, for the purposes of removing such biases or simply reducing redundancy to save on computational costs for downstream analyses, genome dereplication can be extremely useful. Methods that have used ANI to perform dereplication of an input set of genomes - to select representatives - have largely been designed for metagenomic applications. A common strategy of these methods [2,15–17] has been to first perform a preliminary clustering using a more efficient though less precise ANI calculator to bin related genomes together. Afterwards, a secondary, more stringent, clustering is performed using a high precision but low efficiency approach, such as FastANI [7]. While this heuristic works well when applied for selecting representative genomes from a set of diverse metagenomic genome assemblies (MAGs), it is not designed to handle high-redundancy datasets with thousands of genomes belonging to the same species or genus. Hence, we developed skDER, which implements methods for high-resolution selection of representative genomes for both comparative genomic and metagenomic investigations based on ANI estimates computed using skani[1].

Methods

A dynamic dereplication method to select distinct reference genomes for metagenomic investigations

In the first dereplication approach, referred to as the “dynamic” algorithm, an empty set, redundant_genomes, is first initialized. Next, pairwise genome-to-genome similarity information from skani is assessed and a genome specific score is used to determine the preferable genome in the pair to use as a reference. The score for a genome is computed as the product of its assembly N50 and the number of genomes that are regarded as highly similar to it based on user-adjustable ANI and aligned fraction (AF) cutoffs. Assuming the genomes are similar to each other at the required ANI and AF cutoffs and the AF for one genome is not substantially larger than the other, the genome with the lower score, less ideal to serve as a representative, is added to the set of redundant_genomes. This dereplication approach also relies on a third filter, the difference in the AF, which is non-symmetric, between a pair of genomes. If this difference exceeds a certain threshold, the ANI between the genomes meets the ANI cutoff, and the AF for at least one genome exceeds the AF cutoff, then the smaller genome, with the higher AF, will by default be added to the redundant_genomes. The reason being because the smaller genome is largely contained within the larger genome and should not serve as a representative. Once all pairs of genomes are assessed, genomes which are absent in the redundant_genomes set are reported as representatives. It is important to note that this approach approximates selecting a single representative genome for each coarse transitive cluster of related genomes. Therefore, some non-representative genomes might exhibit greater distances to the nearest representative genome than the requested thresholds if they are indirectly connected to the representative genome.

A greedy dereplication method to sufficiently sample the pangenome space for a taxon and simplify comparative genomic studies

The greedy algorithm based approach begins by initializing a set called representative_genomes after which pairwise genome-to-genome similarity information from skani is assessed. For each genome, a list is constructed of alternate genomes deemed similar to it using user-adjustable ANI and AF cutoffs. Similar to the dynamic algorithm-based approach, a score indicating the value of a genome to serve as a representative is computed. Genomes are sorted in descending order by this score. The first genome, with the greatest score, is automatically selected as a representative and appended to the representative_genomes set. The direct list of alternate genomes regarded as redundant to this genome are noted and kept track of in another set, nonrepresentative_genomes, disqualifying them from serving as representatives. As the full list of genomes is traversed linearly, they are appended to the set of representative_genomes if they have not previously been added to the set of nonrepresentative_genomes. Similar to the first representative genome, genomes subsequently selected as representatives also have their respective set of genomes deemed as redundant to them appended to the nonrepresentative_genomes set.

Comparison of skDER to other ANI-based approaches for representative genome selection across *Enterococcus*

We compared the performance of skDER (v1.0.7) for representative genome selection across the diverse genus of *Enterococcus*[18] to two other ANI-based dereplication programs, dRep (v 3.2.2) and galah (v0.3.1), which are primarily designed for selecting reference genomes for mapping metagenomic readsets. Both the dynamic and greedy based approaches implemented in skDER as well as dRep and galah were independently applied to a set of 5,291 genomes classified as *Enterococcus* or *Enterococcus*-like (e.g. Enterococcus_A, Enterococcus_B, etc.) in GTDB release 207[3] (**Table 1**). FastANI[7] was requested for secondary clustering in dRep. For all four methods, dereplication was performed using an ANI-threshold of 99% with otherwise default parameters. galah and dRep selected 472 and 463 representative genomes in 466 and >1000 minutes when using 30 threads, respectively. When similarly provided 30 threads, the dynamic algorithm of skDER selected a comparable 459 genomes as representatives in only 40 minutes. All non-representative genomes exhibited $\geq$ 98.27 ANI to their nearest representative genome, with 90.4% exhibiting $\geq$ 99% ANI to a representative. The greedy algorithm ran for approximately the same time, 34 minutes, and selected 807 representative genomes with the default alignment fraction cutoff of 90%. When the alignment fraction cutoff was relaxed to 25%, the greedy algorithm selected only 412 representative genomes in roughly 35 minutes. All non-representative genomes identified by the greedy-based approach exhibited $\geq$ 99% ANI to their nearest representative genome. The faster runtime of skDER relative to dRep and galah is largely due to skani's improved efficiency in computing pairwise ANI relative to FastANI.

All methods for dereplication selected at least one genome representative for each individual *Enterococcus* species in GTDB. One of the key differences between the four methods was the number of representative genomes selected for the two most highly sequenced species in the

genus, *E. faecium* and *E. faecalis*. Therefore, we further investigated the coverage of the *E. faecalis* pangenome by representative genomes selected using skDER in comparison to those selected by dRep and galah. Panaroo[19] was used to group genes into ortholog groups for all genomes regarded as belonging to *E. faecalis* by GTDB. Importantly, Prokka [20] was used for gene calling and predicted coding sequences on the edge of scaffolds were disregarded to avoid overinflating the pangenome with partial CDS features due to assembly fragmentation [21]. Of the total 1,902 *E. faecalis* genomes considered, dRep and galah had selected a similar number of representatives, 101 and 99, respectively. These representative genomes featured approximately 60.5% to 61.8% of the total number of distinct genes in the *E. faecalis* pangenome. The dynamic dereplication approach within skDER achieved a similar level of pangenome coverage, finding 10,263 (57.6%) of 17,829 distinct genes, but through selection of only 66 representative genomes. This concise selection approach thus makes it well-suited for metagenomic applications through avoiding over-selection of multiple reference genomes for the same strain. In contrast, the greedy approach in skDER with default parameters led to the greatest number of distinct representative genomes being selected, 205, however achieved the best recovery of the full known *E. faecalis* pangenome, featuring 81.6% of the total distinct genes. This showcases the greedy dereplication approach's benefit for simplifying comparative genomics analyses through reducing the number of genomes needed to adequately sample the taxa-wide pangenome.

Dereplication approach (major dependencies)	**Runtime in minutes***	**Representative genomes selected (% of total genomes)**	**Representative *E. faecalis* genomes selected (% of total *E. faecalis* genomes)**	**Distinct *E. faecalis* ortholog groups / genes discovered (% of *E. faecalis* pangenome across all genomes)**
dRep (MASH & FastANI)	>1,000	463 (8.8%)	101 (5.3%)	10,787 (60.5%)
galah (Dashing & FastANI)	466	472 (8.9%)	99 (5.2%)	11,025 (61.8%)
skDER – dynamic (skani)	40	459 (8.7%)	66 (3.5%)	10,263 (57.6%)
skDER – greedy; AF cutoff of 90% (skani)	34	807 (15.3%)	205 (10.8%)	14,546 (81.6%)
skDER – greedy AF cutoff of 25% (skani)	35	412 (7.8%)	73 (3.8%)	9,627 (53.9%)

Table 1: Comparison of different dereplication approaches for selecting representative genomes across the genus of *Enterococcus*. * The time to completion of methods was influenced by other operations taking place on the server at the time and variability should be expected – likely due to Input/output (I/O) performance. Heavy I/O processes on the server were thus restricted and each dereplication approach was run one at a time to get comparative values of runtime. The same number of threads (30) was used for each of the approaches.

Limitations and future directions

Notably, bacterial genomes often feature mobile genetic elements (MGEs), such as plasmids and temperate phages, that might be appropriate to disregard during dereplication for certain research objectives. Therefore, in the future, we plan to incorporate an option to use established tools for prediction of such MGEs[22–25] and mask them prior to dereplication. In addition, contamination is currently not considered, which might be particularly problematic for metagenomic assembled genomes (MAGs). We do not plan to account for contamination directly, but do recommend users assess genome contamination using tools such as CheckM[26] or charcoal [27] and remove genomes or individual contigs deemed as contaminated from consideration prior to running skDER.

Acknowledgments

This work was supported by grants from the National Institutes of Health awarded to L.R.K (NIAID U19AI142720 and NIGMS R35GM137828). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Dr. C. Titus Brown, Dr. N. Tessa Pierce-Ward, and Dr. Karthik Anantharaman for helpful discussions on the development of skDER.

References

1. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *bioRxiv* 2023.01.18.524587 (2023) doi:10.1101/2023.01.18.524587.
2. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
3. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab776.
4. Blin, K. *ncbi-genome-download: Scripts to download genomes from the NCBI FTP servers*. (Github).
5. Salamzade, R. & Kalan, L. SkDER representative genomes for select bacterial taxa. (2023) doi:10.5281/ZENODO.10041203.
6. Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome

sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57**, 81–91 (2007).

7. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 1–8 (2018).
8. Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5**, (2020).
9. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018).
10. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224–1228 (2013).
11. Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W. & Corander, J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* **47**, 5539–5549 (2019).
12. Lees, J. A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).
13. Salamzade, R. *et al.* zol & fai: large-scale targeted detection and evolutionary investigation of gene clusters. *bioRxiv* (2023) doi:10.1101/2023.06.07.544063.
14. David, S. *et al.* Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. *Nat Microbiol* **4**, 1919–1929 (2019).
15. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
16. Woodcroft, B. J. *galah: More scalable dereplication for metagenome assembled genomes*. (Github).
17. Evans, J. T. & Denef, V. J. To Dereplicate or Not To Dereplicate? *mSphere* **5**, (2020).
18. Lebreton, F. *et al.* Tracing the Enterococci from Paleozoic Origins to the Hospital. *Cell* **169**,

849-861.e13 (2017).

19. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
20. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
21. Klassen, J. L. & Currie, C. R. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* **13**, 14 (2012).
22. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* **4**, (2018).
23. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
24. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01953-y.
25. Seemann, T. *phastaf: Identify phage regions in bacterial genomes for masking purposes*. (Github).
26. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
27. *charcoal: Remove contaminated contigs from genomes using k-mers and taxonomies*. (Github).