Protein k-mer analyses for assembly- and alignment-free sequence analysis

This manuscript (permalink) was automatically generated from bluegenes/2021-paper-protein-kmers@b33a771 on February 23, 2022.

Authors

- **N. Tessa Pierce-Ward**
  0000-0002-2942-5331 · bluegenes · saltyscientist
  Department of Population Health and Reproduction, University of California, Davis · Funded by NSF 1711984, NSF 2018911

- **C. Titus Brown**
  0000-0001-6001-2677 · ctb · ctitusbrown
  Department of Population Health and Reproduction, University of California, Davis · Funded by Moore Foundation GBMF4551

Abstract

Background

As the scale of genomic sequencing continues to grow, alignment-free methods for estimating sequence similarity have become critical for conducting tasks ranging from taxonomic classification to phylogenetic analysis on large-scale datasets [1,2]. The majority of alignment-free methods rely upon exact matching of k-mers: subsequences of length k, that can be counted and compared across datasets, with or without use of subsampling methods such as MinHash [] and derivates such as FracMinHash [3]. As k-mer based methods rely on exact sequence matches, they can suffer from limited sensitivity when comparing highly polymorphic sequences or classifying organisms from groups that are not well represented in reference databases.

Current best practices methods can still only categorize a fraction of the metagenomic and metatranscriptomic data, especially for understudied and/or diverse habitats (xx% recovery for soil, xx% recovery ocean metagenomes, etc). Even well-studied environments such as human gut can produce significant uncharacterized metagenome content. “For example, a reference-based approach failed to map 35% of reads in the iHMP study on inflammatory bowel disease (Supp. Data. of (Franzosa et al., 2019)), omitting them from any further analysis. These reads may belong to unknown microbes, phage or viruses, plasmids, or accessory elements of known microbes, all of which can play a role in disease.[from RO1]”. This phenomenon is not restricted to metagenome samples. Alignment-based estimates can fail at larger evolutionary distances and even rRNA amplicon surveys may underestimate bacterial diversity [4].

To increase sensitivity of alignment-free methods, modified k-mer approaches have been introduced, including spaced seeds /split k-mers, which accommodate polymorphic sites in highly similar genomes (CITE). For larger evolutionary distances, protein-based comparisons have long been the gold-standard approach for taxonomic and functional annotation, as protein sequence is more conserved than the underlying DNA sequence [5,6]. As microbial and viral genomes are gene-dense, [MinHash-based] alignment-free comparisons of translated protein sequence have been shown to increase sensitivity for taxonomic classification and genome discovery [7,8]. Here, we demonstrate the utility of protein k-mer comparisons for phylogenomic reconstruction and taxonomic classification at larger evolutionary distances. We use FracMinhash subsampling to facilitate conducting these comparisons at scale [3].

FracMinHash is a MinHash variant for selecting and hashing a set of representative k-mers from a sequence dataset [3]. Unlike traditional MinHash, FracMinHash sketches scale with the size of the dataset, meaning each sketch is comprised of the chosen proportion of k-mers in the input dataset, rather than a chosen number of k-mers. Downsampling sequencing datasets in this way enables estimation of containment, which has been shown to permit more accurate estimation of genomic distance, particularly for genomes of very different lengths [9,10]. Streaming containment estimates have been shown to facilitate genome discovery and correlate with Mash Distance, a proxy for Average Nucleotide Identity (ANI) [8,11].

Standardized genomic measures of relatedness such as ANI and its protein counterpart, Average Amino Acid Identity (AAI) have shown lasting utility for genome relatedness and phylogenomic analysis. Traditional ANI and AAI describe the sequence similarity of all orthologous genes, either in nucleotide or protein space, respectively. Both been shown to be robust measure of overall pairwise genome relatedness even for highly incomplete datasets, such as those comprised of only ~4% of the genome or 100 genes [12,13]. ANI has emerged as the most widely-accepted method for estimating pairwise similarity of microbial genomes and delimiting species boundaries [14]. Recent research appears to confirm 95% ANI species threshold for prokaryotic species, although there is some debate as to the universality of this threshold [15,16,17]. AAI thresholds have been proposed for higher taxonomic ranks, <45%, 45-65% and 65-95% for family, genus, and species [13,18]. While traditional alignment-based estimation of ANI and AAI are computationally intensive, sketching-based estimates and sketching-facilitated estimates have permitted ANI calculations at the scale of whole-databases [1,8,15].

Rahman Hera et. al (2022) [19] introduced accurate nucleotide sequence distance estimation from FracMinHash containment estimates, while accounting for the non-independence of mutated k-mers [20]. Here, we extend that distance estimation to protein k-mers and demonstrate distance estimation across related genomes using the GTDB taxonomy. Furthermore, FracMinHash containment estimates work well for genome pairs of varying lengths and for compositional analysis of metagenome samples. Taken together, these properties enable robust assembly and alignment-free pairwise relatedness estimation that can be used on sequences separated by a wide range of evolutionary distances. Here, we demonstrate that the utility of FracMinHash protein containment, both used directly and a an approximation of ANI and AAI, for taxonomic classification and phylogenomic reconstruction for species across the tree of life.

Notes

- AAI::phylogeny https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1236649/

Results

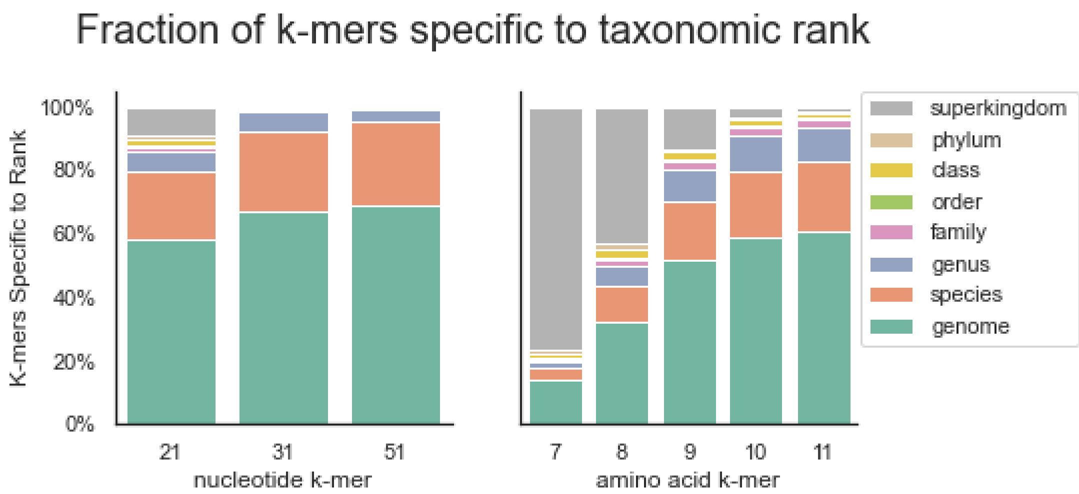
K-mer analysis methods enable similarity detection as low as a single shared k-mer between divergent genomes. As a result, exact matching of long nucleotide k-mers can be used for taxonomic classification and similarity detection between closely related genomes, including strain-level, species-level, and genus-level comparisons (often using k-mer lengths 51, 31, and 21, respectively). At larger evolutionary distances, accumulated nucleotide divergence limits the utility of exact nucleotide k-mer matching. Protein sequences, which are more conserved than their corresponding nucleotide sequences, are the gold standard for comparisons at larger evolutionary distances. Here, we evaluate the utility of amino acid k-mers for a wide range of genomic and metagenomic applications, including sequence distance estimation, taxonomic classification, and metagenome breakdown.

Amino Acid k-mer length selection

The Genome Taxonomy Database (GTDB) provides a genome-based taxonomy for bacterial and archaeal genomes [21]. We begin by assessing the prevalance of nucleotide amino acid k-mers of different k-mer lengths within genomes (/proteomes) selected for inclusion within GTDB. The most recent GTDB release, `rs202`, encompasses 258,407 genomes from 47,895 species.

To make analyses at this scale tractable, we built `sourmash` FracMinHash sketches, with a scaling factor of 1000 for nucleotide k-mers (keep ~1/1000 k-mers) and 200 for amino acid k-mers (keep ~1/200 protein k-mers) [3]. DNA FracMinHash sketches have been shown to accurately subsample genome datasets [3]. For most genomes, both genomic and protein fastas were available for download from NCBI. In remaining cases (n=36,632), genome fastas were translated into protein sequence via Prodigal [22] prior to sketching. We indexed these sketches into `sourmash` databases, which we have made available as part of the `Prepared Databases` section of the `sourmash` documentation, and archived on OSF [https://osf.io/t3fqa/] /Zenodo???.

For a range of nucleotide and amino acid k-mers lengths, we assessed the fraction of k-mers specific to each taxonomic rank. For nucleotide k-mers, we used lengths of 21, 31, and 51, which are commonly used for analyses at the genus, species, and strain level, respectively. For amino acid k-mers, we focused on k-mer lengths ranging between k=7 and k=11, which roughly correspond to nucleotide k-mer lengths 21-31. K-mers specific to a genome were only present in a single genome in the database; k-mers specific to a species were found in at least two genomes of the same species, etc. K-mers specific to a "superkingdom" were found in genomes from at least two phyla.



Fraction of k-mers specific to taxonomic rank

For the GTDB-RS202 database, the majority of nucleotide k-mers are specific to (unique at) a specific genome, species, or genus. Few k-mers are shared across superkingdoms, though these do exist at a k-mer length of 21. In contrast, all protein k-mer sizes contain a portion of k-mers that are shared across genera and above. At a protein k-mer size of 7, over 80% of k-mers are present in genomes found in more than one phylum, while at a protein k-size of 10, the number of genome-specific k-mers is closer to that observed for nucleotide k-mers. Given the difference in k-mers found across taxonomic ranks, we decided to focus on amino acid k-mer lengths 7 and 10 for our primary analyses.

This shared k-mers analysis is limited by the genomes included within GTDB. While some genera contain many thousands of genomes (e.g. 55k *Escherichia* genomes), many others are limited to a single genome or pair of genomes. Thus here we do not consider the absolute numbers of shared k-mers, but rather the proportional differences between k-mer lengths.

Abridged GTDB Benchmarking Dataset

To rigorously assess the utility of protein k-mers for comparisons at an array of evolutionary distances, we selected a subset of GTDB genomes that would allow standardized comparisons across taxonomic ranks and overcome the database-inclusion limitations mentioned above.

For each genus with at least two species clusters in GTDB, one representative genome was randomly selected as an "anchor" genome. Then, one additional genome was selected from the GTDB representative genomes matching the anchor's taxonomy at each higher taxonomic rank. This "evolutionary path" consists of seven genomes: an anchor genome, a genome matching anchor taxonomy down to the genus level, one matching anchor taxonomy to the family level, one matching to the order level, and so on. This creates a gradient of similarity, where comparisons to the anchor genome range from genus-level to superkingdom-level.

Path selection using the representative genomes in GTDB rs202 resulted in 4095 paths comprised of 9213 unique genomes (8790 Bacteria, 333 Archaea). These paths include genome comparisons across 40 phyla (36 Bacteria, 4 Archaea), covering roughly a quarter of the 169 phyla (149 Bacteria, 20 Archaea) in GTDB release rs202. While paths are limited to taxonomies with at least two GTDB representative genomes for each taxonomic rank, these paths provide a rich resource for comparisons at increasing evolutionary distances.

Protein k-mers facilitate alignment-free comparisons at increased evolutionary distances

We begin by assessing standard k-mer comparisons across the 6 comparisons (each genome compared with the anchor genome) within each of 4095 evolutionary paths. We estimate Jaccard Index (number of k-mers shared between two samples divided by the total number of k-mers across both samples) from FracMinHash sketches. When plotted by the rank of the lowest common ancestor, the dynamic range of Jaccard values is clearly much larger for protein k-mer comparisons. While DNA k-mers can provide resolution at the genus level, log-transformed jaccard values for protein k-mers continue to decrease, providing resolution for comparisons even between genome in different phyla. We obtained similar results when comparing all dataset k-mers, suggesting FracMinHash sketching does not greatly impact these results (*Supplemental Figure XX*).

Protein k-mers are shared at higher taxonomic ranks Default scaled values 1000, 200

Protein k-mers are shared at higher taxonomic ranks Default scaled values 1000, 200

Distance estimation from FracMinHash sketch comparisons

Jaccard and Containment of DNA k-mers can be transformed into an estimate of the Average Nucleotide identity between genomes [cite Ondov Mash, Koslicki k-mer paper, koslicki scaled mh paper]. Recently, equations have been developed for FracMinHash that account for the nonindependence of mutated k-mers [19]. Here we apply the FracMinHash distance estimation to protein k-mer comparisons to obtain an alignment-free estimate of Amino Acid Identity [19]. In addition to k-mer based FracMinHash AAI, we also conducted alignment-based AAI methods for each comparison. *We focus on AAI programs that can be run via the command line, and include CompareM (DIAMOND), EzAAIm (MMSeqs2), and EzAAIb (BLAST), each of which use a different aligment algorithm, DIAMOND, MMSeqs2, and BLAST respectively. As BLAST-based is the gold-standard method, we compare all AAI values the BLAST AAI values.* Note that FracMinHash sketches enable estimation of the Containment Index in addition to the more commonly used Jaccard Index. Unlike Jaccard comparisons, which estimate the similarity between sets, Containment estimates are relative to each individual set. When one set is highly trusted, such as a reference genome or proteome, the containment relative to that set may be most informative. When both proteomes are equally trusted, the directional containment can be averaged, as done for BLAST-based AAI(CITE?), which often differ depending on mapping direction. FracMinHash AAI values produced by Jaccard and Containment (here, average containment) methods are very similar.

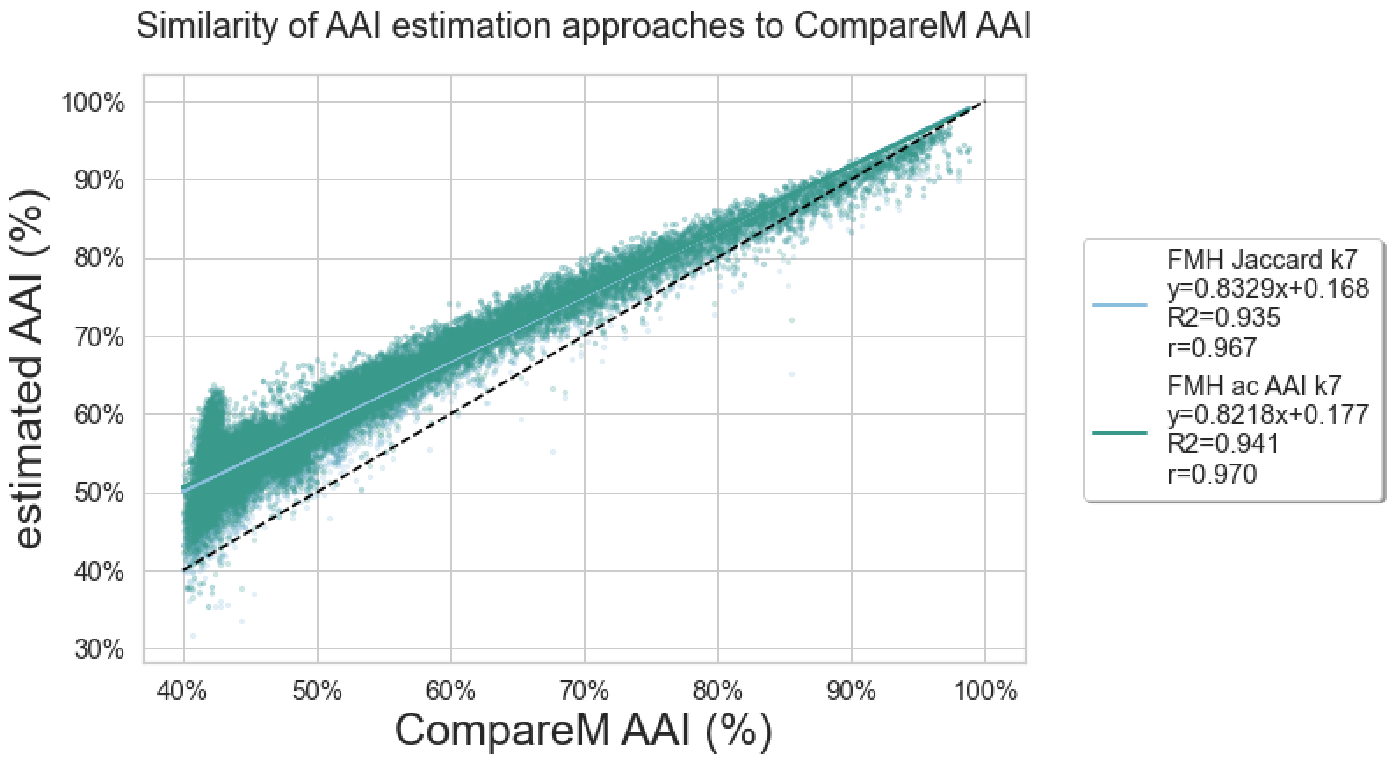


Figure 1: FracMinHash AAI vs CompareM Scaled 200

FracMinHash Containment enables comparison directly from DNA sequence

For protein k-mer comparisons to be useful, any DNA queries must be translated into protein sequence. Often this limits amino acid comparisons to assembly-based workflows, as assemblies can be reliably translated into predicted Open Reading Frames (ORFs).

Here we can utilize direct 6-frame translation, which is assembly-free but does not attempt to find the correct open reading frame. Assuming a single open reading frame, only 1/6th of the k-mers generated by 6-frame translation will belong to the correct ORFs. These erroneous k-mers greatly impact the Jaccard Index when comparing samples. However, the containment index, which enables k-mer comparisons between datasets of different sizes, here provides an added benefit: by using only the FracMinHash containment estimate relative to reference proteomes, we can obtain accurate Amino Acid Identity estimates directly from DNA sequence. We term this “anchor” containment, where the trusted genome is the “anchor” upon which we base the comparison.

We have generalized this type of comparison as the “maximum containment”, which is the containment relative to the smaller set of k-mers. Since 6-frame translation should always yield excess k-mers relative to genomes of similar size max containment should always provide AAI comparisons of 6-frame translated k-mers to reference proteomes.

Note that comparing two 6-frame translated datasets is not recommended, as there is no mechanism to exclude erroneous k-mers introduced during translation.

figure: AAI from translated nucleotide –> reference protein

Robust Taxonomic classification from protein k-mer containment

Anchor containment can also be used to enable robust taxonomic classification from either assembled proteomes or 6-frame translated DNA queries.

With experimental genomes where no reference taxonomic lineage is available, we assessed our annotation relative to `gtdb-tk` classification [23].

Taxonomic utilities are implemented in the `sourmash taxonomy` module.

Dataset	Exact Match	Higher Rank	Unclassified (sourmash)	Unclassified (GTDB-Tk)
MGNify-1000	95.7%	4.3%	N/A	N/A
Delmont-886	73.5%	26.5%	1 (0.1%)	15 (1.7%)

Metagenome breakdown using protein k-mers

Anchor containment also enables metagenome breakdown via min-set-cov.

Metagenome sequences can also be translated in 6-frames, and then the anchor containment can be assessed relative to proteomes in a reference database.

Discussion

K-mer based estimation of sequence identity has been limited to nucleotide sequences of similar size with high sequence identity (>80%),outside of which MinHash Jaccard is less well correlated with sequence identity [1,15].

Shared k-mers

K-mers shared at such a high level may be indicative of true shared biological sequence, contamination, or k-mer homoplasy: the presence of k-mers that are identical by chance rather than evolutionary descent.](images/gtdb-rs202.lca_f_aggregated_kmers.png){#fig:gtdb-kmers height=2in}

The differences observed between nucleotide and amino acid k-mers, as well as across different k-mer lengths suggests that these different k-mer sizes may provide resolution at different taxonomic ranks. The exact characterization here is of course impacted by which are genomes included in the database, but we are confident that the 258k genomes included within GTDB provide a good testing ground for this assessment.

By leveraging the Containment Index of Scaled MinHash sketches with both nucleotide and protein k-mers, we can extend accurate k-mer sequence identity to sequences of different sizes and to >50% Amino Acid Identity.

Cricuolo [24] (suggests w/ appropriate correction, nucl MinHash Jaccard can be used up to >65% ANI??)

Here, we utilize Scaled MinHash sketches with Containment to overcome size differences between sequences being compared.

To accurately estimate sequence identity from sequence files of different sizes(genomes, metagenomes, etc), we employ Scaled Minhash sketches, which enables estimation of the

Containment Index.

A number of methods have used discriminatory k-mer analysis for taxonomic classification. However, most rely upon first developing a reference of discriminatory k-mers, e.g. k-mers unique to / diagnostic of a taxonomic group. Instead, sourmash gather leverages the Containment Index to find the reference match that shares the largest number of k-mers with the query sequence.

At k=21 (dna) and k=7 (protein), many k -mers are shared across taxonomic groups. Unlike many k-mer based classifiers, we do not need to explicitly characterize the discriminatory k-mers for each taxonomic group. The Containment Index uses all matched k-mers between the query and each reference, finding the % of each reference genome present in the query. Gather then selects the most covered (highest percent contained) reference genome, thus utilizing the combination of shared and discriminatory k-mers to find the most parsimonious match. After finding the best match, all matched k-mers are removed for the query in order to repeat the analysis to find the next most parsimonious genome match.

While this method is still dependent on a good set of reference genomes, updating the set of references with new data does not require recalculation of discriminatory k=mer sets…

** discussion of k-mer size **

- Scaled Minhash distance estimation is robust to completeness (unlike standard minhash https://drep.readthedocs.io/en/latest/choosing_parameters.html#importance-of-genome-completeness)

containment is imp: Assembly methods can exclude up to XX% of data.

Unlike Jaccard comparisons, which estimate the similarity between sets, containment estimates are relative to each individual set. When one set is highly trusted, such as a reference genome or proteome, the containment relative to that set may be most informative. In these cases, we can consider the trusted genome as an “anchor” upon which we are basing our comparison, and the containment relative to this set as “anchor containment.”

Conclusions

Containment-based pairwise distance estimation via Scaled Minhash enables accurate assembly-free and alignment-free phylogenomic reconstruction and taxonomic classification across a wide range of evolutionary distances.

Methods

Scaled MinHash Sketching with Sourmash

As implemented in sourmash [25,26,27], Scaled MinHash is a MinHash variant that uses a scaling factor to subsample the unique k-mers in the dataset to the chosen proportion (1/ `scaled`). As k-mers are randomized prior to systematic subsampling, Scaled MinHash sketches are representative subsets that can be used for comparisons, as long as the k-mer size and chosen scaled value remain consistent. Unlike traditional MinHash sketches, Scaled MinHash sketches enable similarity estimation with containment, which permits more accurate estimation of genomic distance when genomes or datasets differ in size [9,10].

Sourmash v4.x supports sketching from either nucleotide or protein input sequence. All genome sequences were sketched with sourmash v4.0 using the `sourmash sketch dna` command, k-mer sizes of 21,31,51, a scaling factor of 1000. Sourmash also supports 6-frame translation of nucleotide sequence to amino acid sequence. To assess the utility of these translated sketches, genome sequences were also sketched with the `sourmash sketch translate` command at protein k-sizes (*kaa-mer sizes?*) of 7-12 and a scaling factor of 100. All proteome sequences were sketched with sourmash v4.0 using the `sourmash sketch protein` command at protein k-sizes (*kaa-mer sizes?*) of 7-12 and a scaling factor of 100. Where higher scaling factors were evaluated, these original sketches were downsampled using the sourmash `downsample` method prior to conducting sequence similarity comparisons.

Sequence Identity Estimation from Scaled MinHash

(very DRAFTy)

Sourmash contains standard implementations of Jaccard Index [1] and Containment Index [9] set comparisons.

Estimating Sequence Similarity from Jaccard For a comparison between two genomes (genomeA, genomeB), the Jaccard Index represents the k-mers shared between the two genomes (sketch intersection) divided by the k-mers present in both sketches (sketch union). Thus the Jaccard Index represents the percent of shared k-mers relative to all k-mers across both genomes (intersection/genomeA+genomeB). MinHash Sketch Jaccard has been shown to correlate well with ANI at high sequence identities (>=90% sequence identity) [1]; (>=80% sequence identity [15].

Mash Distance from Scaled MinHash Jaccard

TBD

Estimating Sequence Similarity from Containment As the Jaccard Index utilizes the union of all k-mers in a dataset, it is greatly affected by differences in dataset size [28]. The Containment Index instead represents the percent of a genome found in the comparison genome. Containment is directional: while the number of shared k-mers is fixed for a pairwise comparison, the Containment of each dataset will depend on the unique k-mers found in that particular dataset. Containment for genomeA will be (intersection/genomeA), while Containment for genomeB will be (intersection/genomeB).

Alignment-based ANI represents the sequence similarity of the alignable fraction of two genomes. In this way, ANI only compares the shared sequences, and discounts/ignores all other sequence present in either genome. Bidirectional containment comparisons use the same numerator (shared k-mers), but may contain different numbers of non-shared k-mers in the denominator.

In cases where both genomes are high-quality and highly complete, we can most closely approximate ANI by using the maximum value between the bidirectional containment values: that is, using the comparison that represents the shared sequence over the genome with the smallest number of non-shared k-mers.

In cases where one genome is more trusted (high quality and highly complete), Containment may be best calculated relative to the trusted genome. This use case also allows us to estimate sequence identity from larger sequence collections, such as metagenomes. By definition, metagenomes contain k-mers from many organisms. We can take advantage of directional Containment by calculating the Containment Index of Reference genomes that share many k-mers with the Metagenome. We have already shown the utility of Containment for metagenome classification [25], but now we can report estimated average sequence identity between the matching sequence regions and the reference genome.

Estimating Sequence Identity from Scaled MinHash

TBD

Blanca et al, 2021 [20] presented a method to estimate the mutation rate between MinHash sketches while accounting for the non-independence of mutated k-mers. Using [29], we estimate Sequence Identity from Scaled MinHash Containment.

Estimating sequence similarity from Scaled MinHash requires a good estimate of the number of unique k-mers in the sketched sequencing dataset [30]…

Scaled MinHash Distance Correlates with Standard Methods

FastANI v1.32 ([15]; run with default parameters) was used to obtain Average Nucleotide Identity between the anchor genome and each additional genome in its evolutionary path. FastANI is targeted at ANI values between 80%-100%, so only values in this range are considered "trusted" and used in **assessing the correlation between Scaled MinHash estimates and FastANI._(TBD)_**

CompareM v0.1.2 ([31]; run with `--sensitive` parameter for DIAMOND mapping) was used to obtain Average Amino Acid Identity between the anchor proteome and each additional proteome in its evolutionary path. CompareM reports the mean and standard deviation of AAI, as well as the fraction of orthologous genes upon which this estimate is based. Briefly, CompareM calls genes for each genome or proteome using PRODIGAL [6] and conducts reciprocal best-hit mapping via DIAMOND [22]. By default, CompareM requires at least 30% percent sequence identity and 70% percent alignment length to identify orthologous genes. As DIAMOND alignment-based homology identification may correlate less well with BLAST-based homology under 60% sequence identity [32/], **we also ran compareM with a percent sequence identity threshold of 60% to obtain a set of high-confidence orthologous genes for AAI estimation. We report correlation between Scaled MinHash AAI estimation and each of these compareM parameter sets in XX *(TBD)***. *CompareM was also used to obtain AAI values directly from each genome, using PRODIGAL to translate sequences prior to gene calling. These results [were not significantly different from proteome-based AAI estimation??] (Supplemental XX).*

Taxonomic Classification with Sourmash `Gather` and `Taxonomy`

To take advantage of the increased evolutionary distance comparisons offered by protein k-mers, we apply compositional analysis with sourmash gather [25] to protein sequences (amino acid input and 6-frame translation from nucleotides). Sourmash gather is conducted in two parts: First (preselection), gather searches the query against all reference genomes, building all genomes with matches into a smaller, in-memory database for use in step 2. Second (decomposition), gather does iterative best-containment decomposition, where query k-mers are iteratively assigned to the reference genome with best containment match. In this way, gather reports the minimal list of reference genomes that contain all of the k-mers that matched any reference in the database.

For reference matches with high sequence identity (ANI) to the query, we classify the query sequence as a member of the reference taxonomic group, as in [25]. **However, when ANI between the query and the top reference match exceeds the taxonomic rank threshold (e.g. species default 95%), we use a least/lowest common ancestor (LCA) approach to report likely taxonomy at a higher**

taxonomic rank ***(TBD)***. Briefly, as gather reports non-overlapping genome matches, we can sum the k-mer matches for all genomes with shared taxonomies at the next higher taxonomic rank to report the best query containment at that rank. As this gather-LCA approach first uniquely assigns k-mers to their best reference genome, it bypasses the impact of increasing database size on taxonomic assignment observed for other LCA-based k-mer classification approaches [33].

Workflows and Computing Resources

Reproducible workflows associated with this paper are available at XX (gh link + doi for release), with datasets available at OSF (XX). All workflows were executed using snakemake >= 5.26 [34)] on the FARM cluster at UC Davis, using practices outlined in [35].

Supplemental

Protein k-mers facilitate alignment-free comparisons at increased evolutionary distances

Protein k-mers are shared at higher taxonomic ranks: ALL KMERS

Protein k-mers are shared at higher taxonomic ranks: ALL KMERS

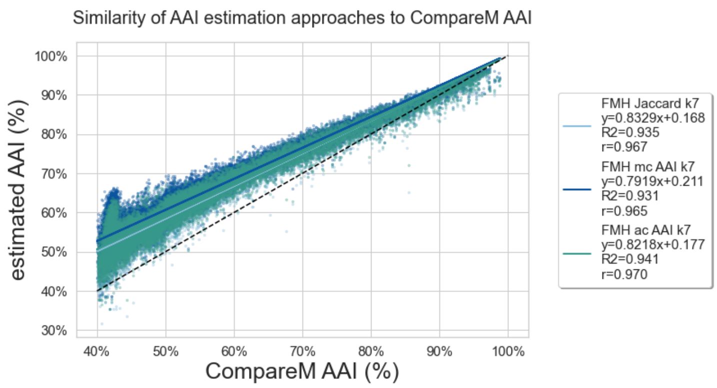


Figure 2: FracMinHash AAI vs CompareM Scaled 1

References

1. **Mash: fast genome and metagenome distance estimation using MinHash**
Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, Adam M Phillippy
Genome Biology (2016-12) https://doi.org/gfx74q
DOI: 10.1186/s13059-016-0997-x · PMID: 27323842 · PMCID: PMC4915045

2. **Kraken: ultrafast metagenomic sequence classification using exact alignments**
Derrick E Wood, Steven L Salzberg
Genome Biology (2014) https://doi.org/gfkndk
DOI: 10.1186/gb-2014-15-3-r46 · PMID: 24580807 · PMCID: PMC4053813

3. **Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers**
Luiz Irber, Phillip T Brooks, Taylor Reiter, NTessa Pierce-Ward, Mahmudur Rahman Hera, David Koslicki, CTitus Brown
Bioinformatics (2022-01-12) https://doi.org/gn34zt
DOI: 10.1101/2022.01.11.475838

4. **How Much Do rRNA Gene Surveys Underestimate Extant Bacterial Diversity?**
Luis M Rodriguez-R, Juan C Castro, Nikos C Kyrpides, James R Cole, James M Tiedje, Konstantinos T Konstantinidis
Applied and Environmental Microbiology (2018-03-15) https://doi.org/ghtrdq
DOI: 10.1128/aem.00014-18 · PMID: 29305502 · PMCID: PMC5835724

5. **Basic local alignment search tool.**
SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman
Journal of molecular biology (1990-10-05) https://www.ncbi.nlm.nih.gov/pubmed/2231712
DOI: 10.1016/s0022-2836(05)80360-2 · PMID: 2231712

6. **Fast and sensitive protein alignment using DIAMOND**
Benjamin Buchfink, Chao Xie, Daniel H Huson
Nature Methods (2015-01) https://doi.org/gftzcs
DOI: 10.1038/nmeth.3176 · PMID: 25402007

7. **Fast and sensitive taxonomic classification for metagenomics with Kaiju**
Peter Menzel, Kim Lee Ng, Anders Krogh
Nature Communications (2016-09) https://doi.org/f8h4b6
DOI: 10.1038/ncomms11257 · PMID: 27071849 · PMCID: PMC4833860

8. **Mash Screen: high-throughput sequence containment estimation for genome discovery**
Brian D Ondov, Gabriel J Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B Buck, Adam M Phillippy
Genome Biology (2019-12) https://doi.org/ghtqmb
DOI: 10.1186/s13059-019-1841-x · PMID: 31690338 · PMCID: PMC6833257

9. **Improving MinHash via the containment index with applications to metagenomic analysis**
David Koslicki, Hooman Zabeti
Applied Mathematics and Computation (2019-08) https://doi.org/ghtqrv
DOI: 10.1016/j.amc.2019.02.018

10. **Dashing: fast and accurate genomic distances with HyperLogLog**

Daniel N Baker, Ben Langmead
Genome Biology (2019-12) https://doi.org/ggkmjc
DOI: 10.1186/s13059-019-1875-0 · PMID: 31801633 · PMCID: PMC6892282

11. **Metalign: efficient alignment-based metagenomic profiling via containment min hash**
Nathan LaPierre, Mohammed Alser, Eleazar Eskin, David Koslicki, Serghei Mangul
Genome Biology (2020-12) https://doi.org/ghtqrz
DOI: 10.1186/s13059-020-02159-0 · PMID: 32912225 · PMCID: PMC7488264

12. **Toward a More Robust Assessment of IntraspeciesDiversity, Using Fewer GeneticMarkers**
Konstantinos T Konstantinidis, Alban Ramette, James M Tiedje
Applied and Environmental Microbiology (2006-11) https://doi.org/dcmw9q
DOI: 10.1128/aem.01398-06 · PMID: 16980418 · PMCID: PMC1636164

13. **Uncultivated microbes in need of their own taxonomy**
Konstantinos T Konstantinidis, Ramon Rosselló-Móra, Rudolf Amann
The ISME Journal (2017-11) https://doi.org/gbprgw
DOI: 10.1038/ismej.2017.113 · PMID: 28731467 · PMCID: PMC5649169

14. **Shifting the genomic gold standard for the prokaryotic species definition**
Michael Richter, Ramon Rosselló-Móra
Proceedings of the National Academy of Sciences (2009-11-10) https://doi.org/dvchzz
DOI: 10.1073/pnas.0906412106 · PMID: 19855009 · PMCID: PMC2776425

15. **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries**
Chirag Jain, Luis M Rodriguez-R, Adam M Phillippy, Konstantinos T Konstantinidis, Srinivas Aluru
Nature Communications (2018-12) https://doi.org/gfknmg
DOI: 10.1038/s41467-018-07641-9 · PMID: 30504855 · PMCID: PMC6269478

16. **Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries**
Matthew R Olm, Alexander Crits-Christoph, Spencer Diamond, Adi Lavy, Paula B Matheus Carnevali, Jillian F Banfield
mSystems (2020-02-11) https://doi.org/ggwqh6
DOI: 10.1128/msystems.00731-19 · PMID: 31937678 · PMCID: PMC6967389

17. **There is no evidence of a universal genetic boundary among microbial species**
Connor S Murray, Yingnan Gao, Martin Wu
Microbiology (2020-07-27) https://doi.org/ghtrdw
DOI: 10.1101/2020.07.27.223511

18. **Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead**
Konstantinos T Konstantinidis, James M Tiedje
Current Opinion in Microbiology (2007-10) https://doi.org/b2q3jd
DOI: 10.1016/j.mib.2007.08.006 · PMID: 17923431

19. **Debiasing FracMinHash and deriving confidence intervals for mutation rates across a wide range of evolutionary distances**
Mahmudur Rahman Hera, NTessa Pierce-Ward, David Koslicki
Bioinformatics (2022-01-12) https://doi.org/gn342h
DOI: 10.1101/2022.01.11.475870

20. **The statistics of <i>k</i> -mers from a sequence undergoing a simple mutation process without spurious matches**

Antonio Blanca, Robert S Harris, David Koslicki, Paul Medvedev
Bioinformatics (2021-01-17) https://doi.org/fq3g
DOI: 10.1101/2021.01.15.426881

21. **A complete domain-to-species taxonomy for Bacteria and Archaea**
Donovan H Parks, Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J Mussig, Philip Hugenholtz
Nature Biotechnology (2020-09-01) https://doi.org/ggtbk2
DOI: 10.1038/s41587-020-0501-8 · PMID: 32341564

22. **Prodigal: prokaryotic gene recognition and translation initiation site identification**
Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, Loren J Hauser
BMC Bioinformatics (2010-12) https://doi.org/cktxnm
DOI: 10.1186/1471-2105-11-119 · PMID: 20211023 · PMCID: PMC2848648

23. **GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database**
Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, Donovan H Parks
Bioinformatics (2019-11-15) https://doi.org/ggc9dd
DOI: 10.1093/bioinformatics/btz848 · PMID: 31730192 · PMCID: PMC7703759

24. **On the transformation of MinHash-based uncorrected distances into proper evolutionary distances for phylogenetic inference**
Alexis Criscuolo
F1000Research (2020-11-10) https://doi.org/gjn4jw
DOI: 10.12688/f1000research.26930.1 · PMID: 33335719 · PMCID: PMC7713896

25. **Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers**
Luiz Irber, Phillip T Brooks, Taylor Reiter, NTessa Pierce-Ward, Mahmudur Rahman Hera, David Koslicki, CTitus Brown
Manubot (2022-01-17) https://dib-lab.github.io/2020-paper-sourmash-gather/

26. **Large-scale sequence comparisons with sourmash**
NTessa Pierce, Luiz Irber, Taylor Reiter, Phillip Brooks, CTitus Brown
F1000Research (2019-07-04) https://doi.org/gf9v84
DOI: 10.12688/f1000research.19675.1 · PMID: 31508216 · PMCID: PMC6720031

27. **sourmash: a library for MinHash sketching of DNA**
C Titus Brown, Luiz Irber
The Journal of Open Source Software (2016-09-14) https://doi.org/ghdrk5
DOI: 10.21105/joss.00027

28. **Beware the Jaccard: the choice of similarity measure is important and non-trivial in genomic colocalisation analysis**
Stefania Salvatore, Knut Dagestad Rand, Ivar Grytten, Egil Ferkingstad, Diana Domanska, Lars Holden, Marius Gheorghe, Anthony Mathelier, Ingrid Glad, Geir Kjetil Sandve
Briefings in Bioinformatics (2020-09-25) https://doi.org/gjnvx4
DOI: 10.1093/bib/bbz083 · PMID: 31624847

29. **GitHub - KoslickiLab/mutation-rate-ci-calculator: This software calculates a confidence interval for the mutation rate from a set of observed containment indices under a simple nucleotide mutation process.**
GitHub
https://github.com/KoslickiLab/mutation-rate-ci-calculator

30. **[WIP] Ertl estimators for scaled minhash by luizirber · Pull Request #1270 · sourmash-bio/sourmash**
GitHub
https://github.com/sourmash-bio/sourmash/pull/1270

31. **GitHub - dparks1134/CompareM: A toolbox for comparative genomics.**
GitHub
https://github.com/dparks1134/CompareM

32. **AAI: BLAST vs Diamond**
LM Rodriguez-R
https://rodriguez-r.com/blog/aai-blast-vs-diamond/

33. **RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification**
Daniel J Nasko, Sergey Koren, Adam M Phillippy, Todd J Treangen
Genome Biology (2018-12) https://doi.org/ggc9db
DOI: 10.1186/s13059-018-1554-6 · PMID: 30373669 · PMCID: PMC6206640

34. **Sustainable data analysis with Snakemake**
Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, … Johannes Köster
F1000Research (2021-01-18) https://doi.org/gjjkwv
DOI: 10.12688/f1000research.29032.1 · PMID: 34035898 · PMCID: PMC8114187

35. **Streamlining data-intensive biology with workflow systems**
Taylor Reiter, Phillip T Brooks†, Luiz Irber†, Shannon EK Joslin†, Charles M Reid†, Camille Scott†, CTitus Brown, NTessa Pierce-Ward
GigaScience (2021-01-13) https://doi.org/gjfk22
DOI: 10.1093/gigascience/giaa140 · PMID: 33438730 · PMCID: PMC8631065