Alignment-free distance estimation across the tree of life

This manuscript (permalink) was automatically generated from bluegenes/2021-ani-paper@18931fe on April 14, 2021.

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

Sequence similarity estimation is critical for genome analyses ranging from taxonomic classification to phylogenetic reconstruction. Here, we introduce an alignment-free k-mer based method for quickly and accurately estimating pairwise sequence similarity, including Average Nucleotide Identity (ANI) and Average Amino Acid Identity (AAI). Because this method is assembly-independent and sequence-agnostic, it can be applied to both DNA and protein sequences across the tree of life. We demonstrate the utility of this method with similarity comparisons and protein taxonomic classification for benchmarking sets from the genome-based prokaryotic taxonomy (GTDB). We then extend the analysis to highly divergent and incomplete datasets as well as alternate domains of life.

Keywords (3-10)

Overall Genome Relatedness Index (OGRI), Average Nucleotide Identity (ANI), Average Amino Acid Identity (AAI), pairwise evolutionary distance, Jaccard Index, Containment Index, MinHash, Scaled MinHash, k-mer

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. 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 [3].

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 [4,5]. 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 [6,7]. Here, we demonstrate the utility of protein k-mer comparisons for phylogenomic reconstruction and taxonomic classification at larger evolutionary distances and across both gene-rich and [gene-sparse] sequences. We use Scaled Minhash subsampling to facilitate conducting these comparisons at scale [8].

Scaled Minhash is a MinHash variant for selecting and hashing a set of representative k-mers from a sequence dataset [8]. Unlike traditional MinHash, Scaled MinHash 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) [7,11].

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 [1,7,15].

[Pierce-Ward et al., 2021 (tbd technical paper)] showed that Scaled MinHash containment estimates can be used to approximate both ANI (nucleotide k-mers) and Average Amino Acid Identity (AAI; protein k-mers), while accounting for the non-independence of mutated k-mers [19]. Furthermore, Scaled MinHash 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 Scaled MinHash 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

Accurate distance estimation from Maximum Containment

(Correlation between Scaled MinHash Containment and ANI)

Containment searches enable similiarity estimation, especially between genomes of different lengths. Max containment normalizes the shared content by the smaller of the two genomes

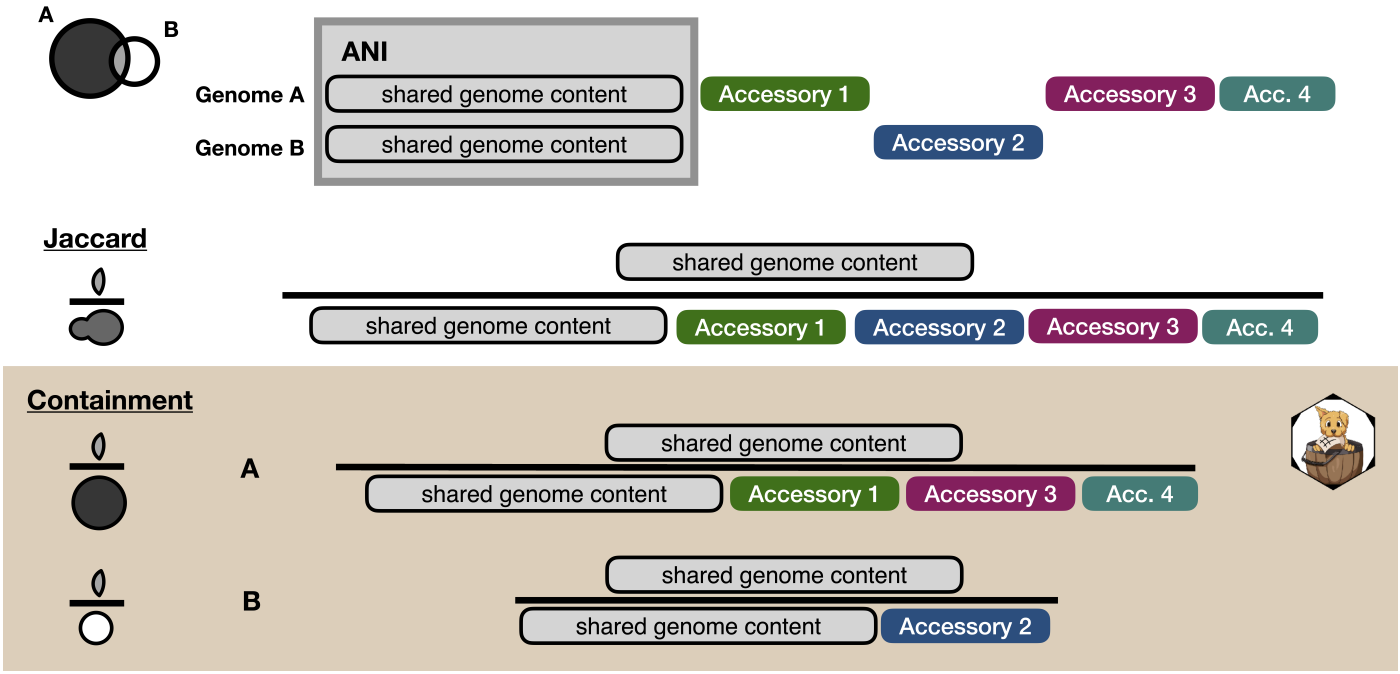

Figure 1: Max Containment to ANI and AAI. Containment calculation is guaranteed to be more similar to traditional calculation of Average Nucleotide Identity and Average Amino Acid Identity, which compared only the sections of genome that align. The shared k-mer content (containment numerator) can be thought of as the alignable sections of the genomes. The denominator of the Jaccard index is the alignable sections + the unalignable sections. The lower bound of the containment denominator will be the exact same as the numerator at 100% containment, where all k-mers are found within the comparison dataset. The upper bound will be the same as the Jaccard denominator, where all k-mers of the comparison dataset are found within the query dataset, and it is the query that contains any additional nonshared k-mers/unalignable sequence.

To assess the utility of Scaled MinHash techniques across evolutionary distance, we generated a series of “evolutionary paths” from the set of 31k representative GTDB genomes. Each evolutionary path offers six genome similarity comparisons at a range of evolutionary distances. For each genome comparison, we estimated Average Nucleotide Identity (ANI) using fastANI [15] and Average Amino Acid Identity (AAI) using compareM [**???** https://github.com/dparks1134/CompareM].

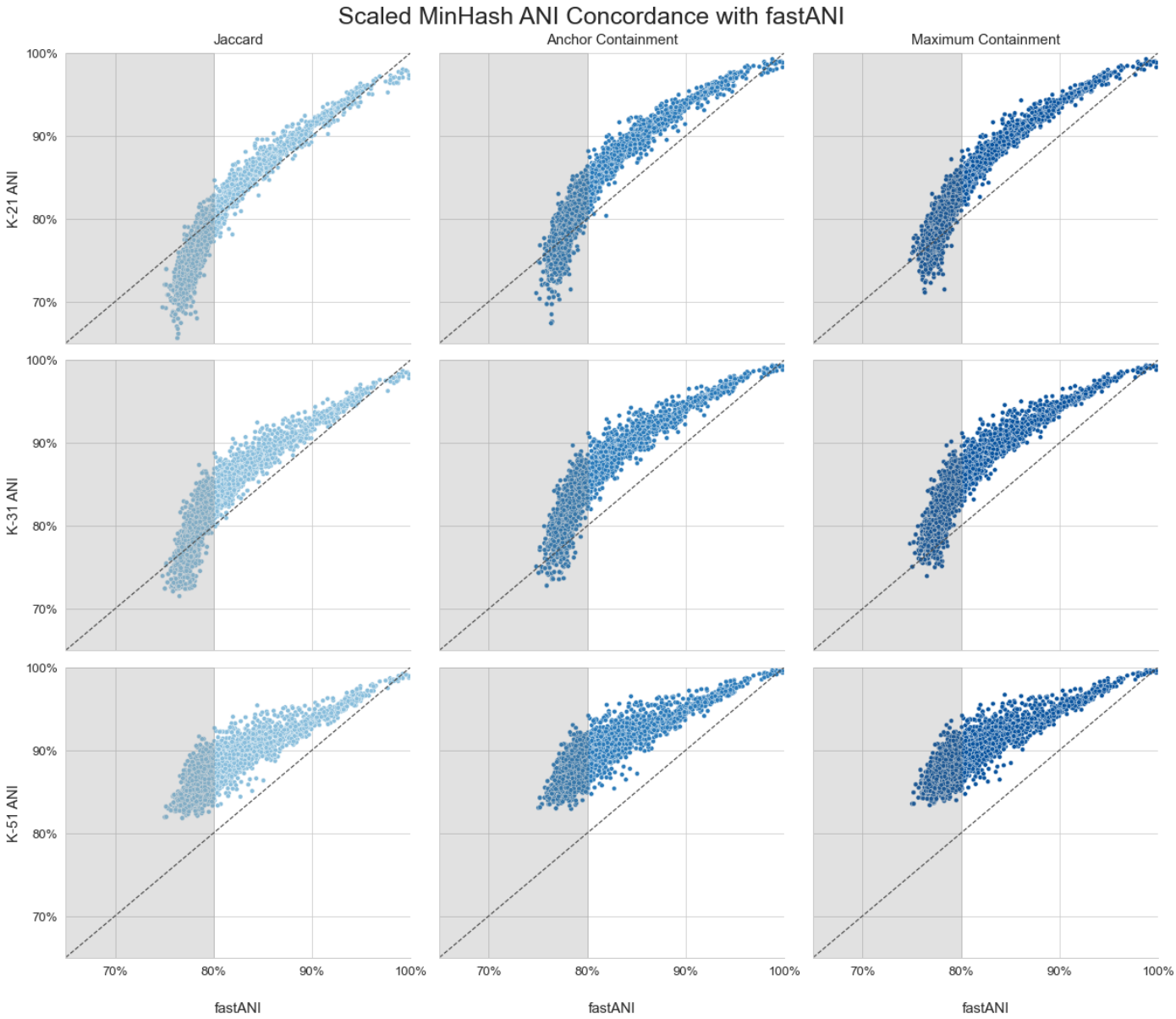

Figure 2: Scaled MinHash ANI vs FastANI GTDB Evolpaths Dataset

Similarity detection and clustering at increased evolutionary distances

Protein k-mers! <compare heatmap w/ max containment for subset of gtdb data?>

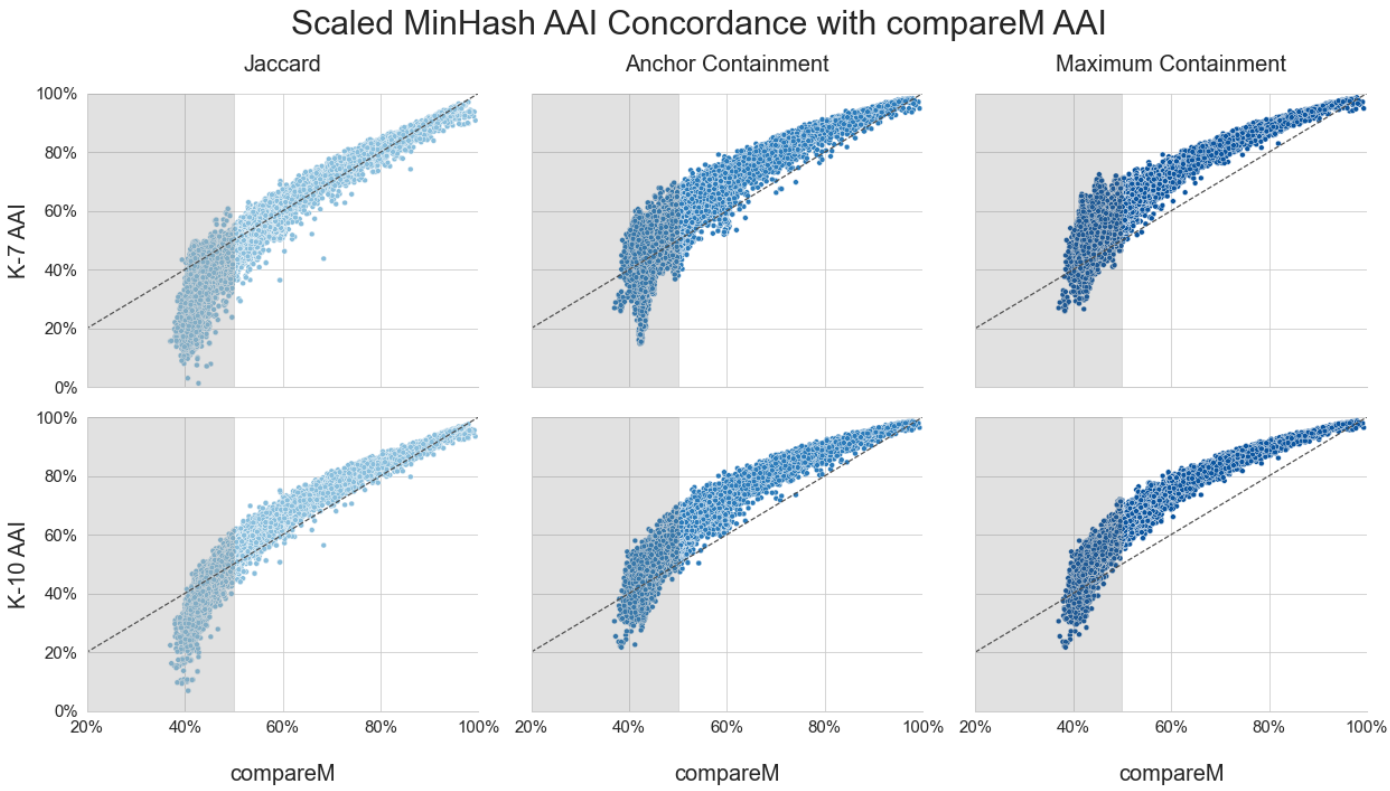

Figure 3: Scaled MinHash AAI vs CompareM GTDB Evolpaths dataset

(DNA vs Protein) - *(just containment, no ANI/AAI) - include dayhoff or just protein?*

K-mer analysis methods enable similarity detection as low as a single shared k-mer between divergent genomes. As a result, exact matching long nucleotide k-mers can be used for taxonomic classification between closely related genomes, including at the strain, species, and genus level (k-mer lengths 51, 31, and 21, respectively). At larger evolutionary distances, accumulated nucleotide divergence limits the utility of exact nucleotide k-mer matching.

As protein sequences are more conserved than their coding nucleotide sequences, exact matching of protein k-mers can detect sequence similarity in spite of synonymous nucleotide substitutions that may have accumulated across the two sequences.

Exact matching of k-mers has long been deemed a shortcoming for k-mer based analyses, limiting similarity detection power across larger evolutionary distances. However, protein k-mers (and k-mers leveraging reduced protein alphabets)

Protein sequences are more conserved than their underlying DNA sequences. Whole-proteome MinHash sketches are more similar than whole-genome DNA sketches, enabling us to find protein-level similarity across divergent genomes.

For , e.g. Pseudomonas, XX% of k-mers are shared within the chosen/published genomes within species. For all published genomes within the genus, a median of xx% of k-mers are shared between genomes of one species and genomes of the a different species in the same genus.

rankinfo ... at ksize of 10... -xx% of DNA k-mers are shared within-species -yy% of protein k-mers are shared within-species - zz% of DNA k-mers are shared within-genus ... etc == median or mean containment at rank? containent = % of a genome's k-mers that are shared – do using ALL of gtdb, BUT, start with just a single set of genomes.. e.g. Pseudomonas? == similar to "shared k-mers" paper [20]

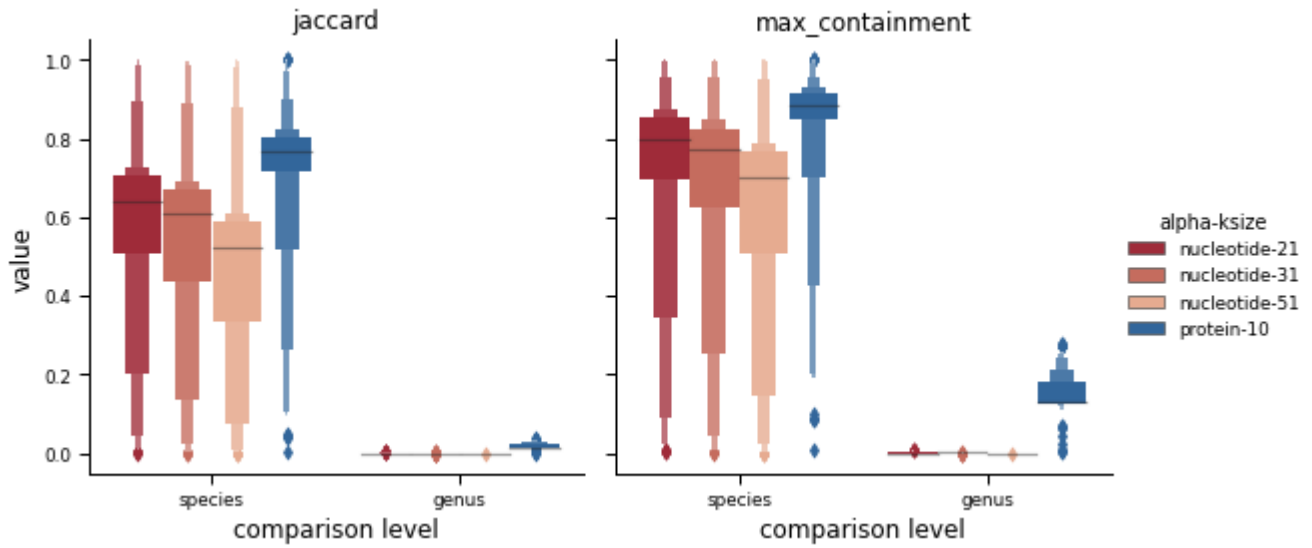

Figure 4: Protein k-mer containment facilitates genus-level comparisons 10k pseudomonas genome sequences, median containment at each alphabet

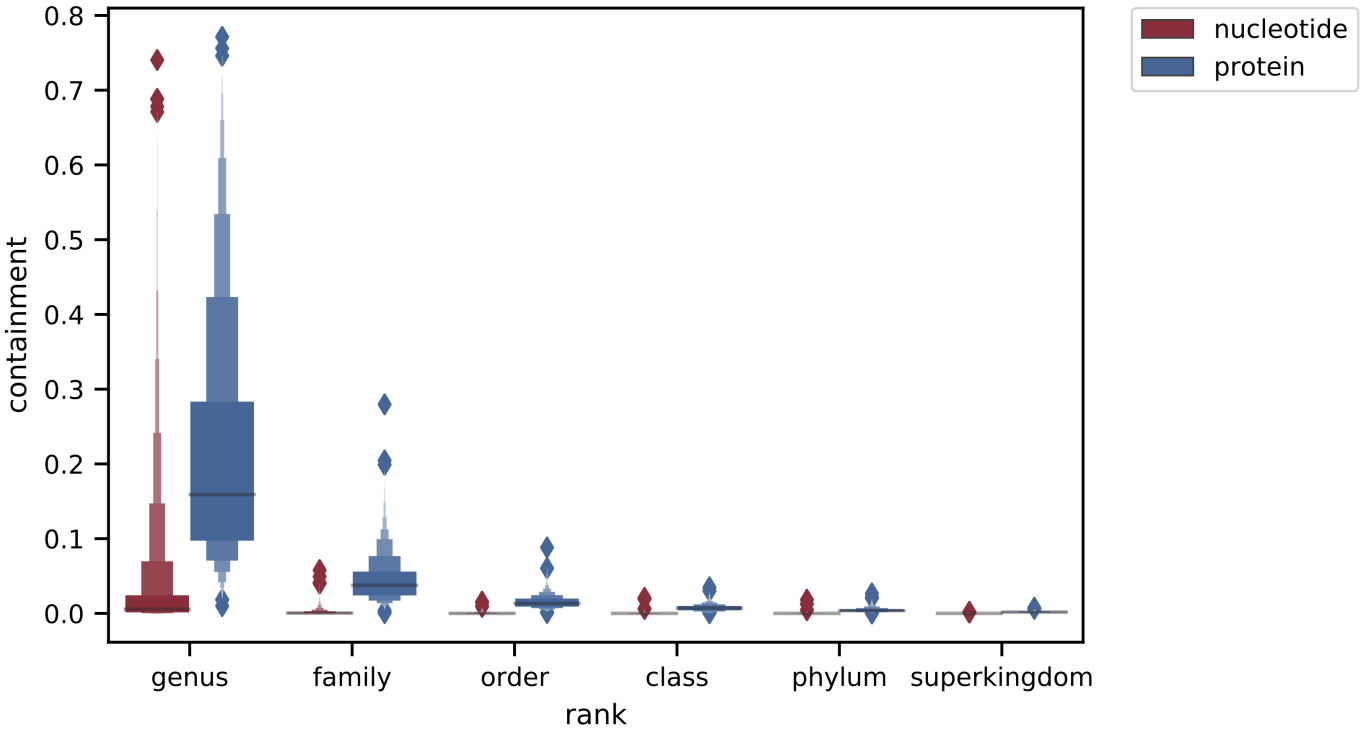

Figure 5: Protein k-mers facilitate comparisons across species This currently uses the evolutionary paths dataset. Perhaps better to demonstrate with a different test set – say, just the species, genus family level, using something like Pseudomonas that has a lot of published genomes. Also show jaccard to emphasize how it gets progressively worse when you start comparing genomes that are different sizes? Or separate figure for this...?

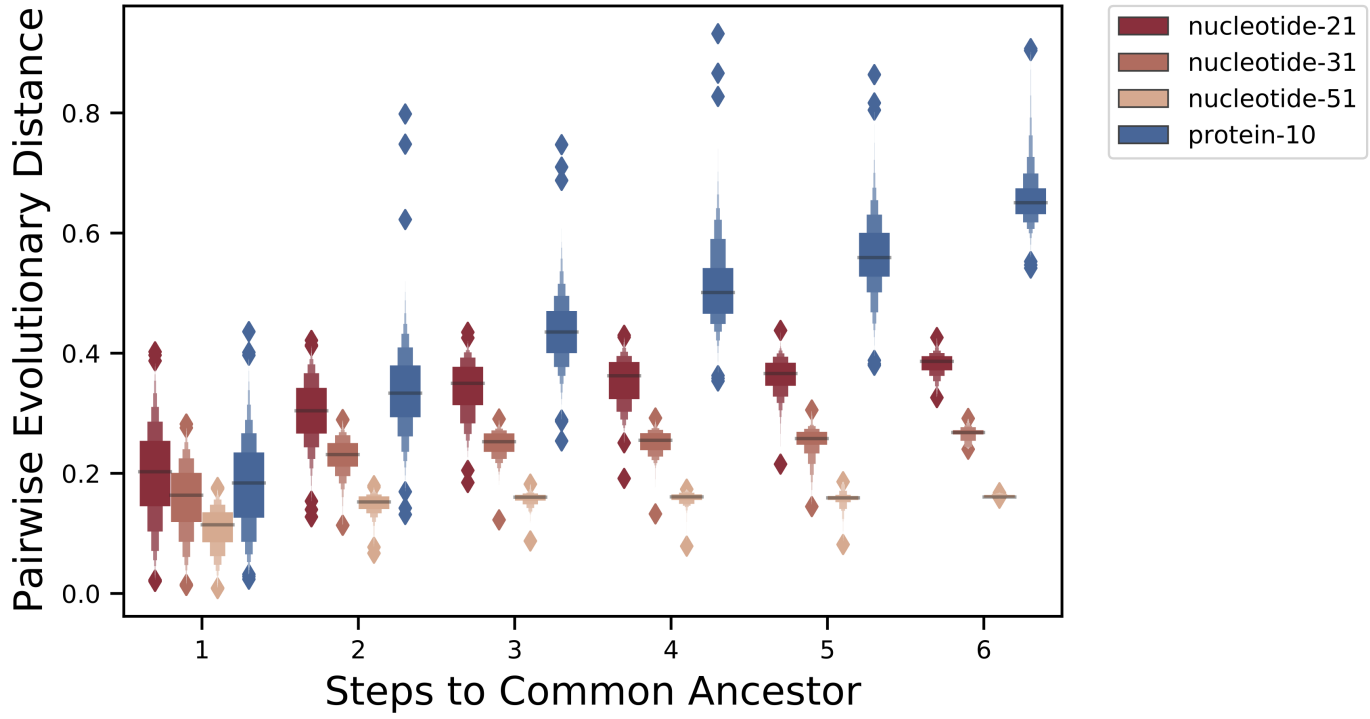

Figure 6: Containment-based ANI, AAI estimates, evolpaths

Alignment-free phylogeny recapitulates core-genome phylogeny

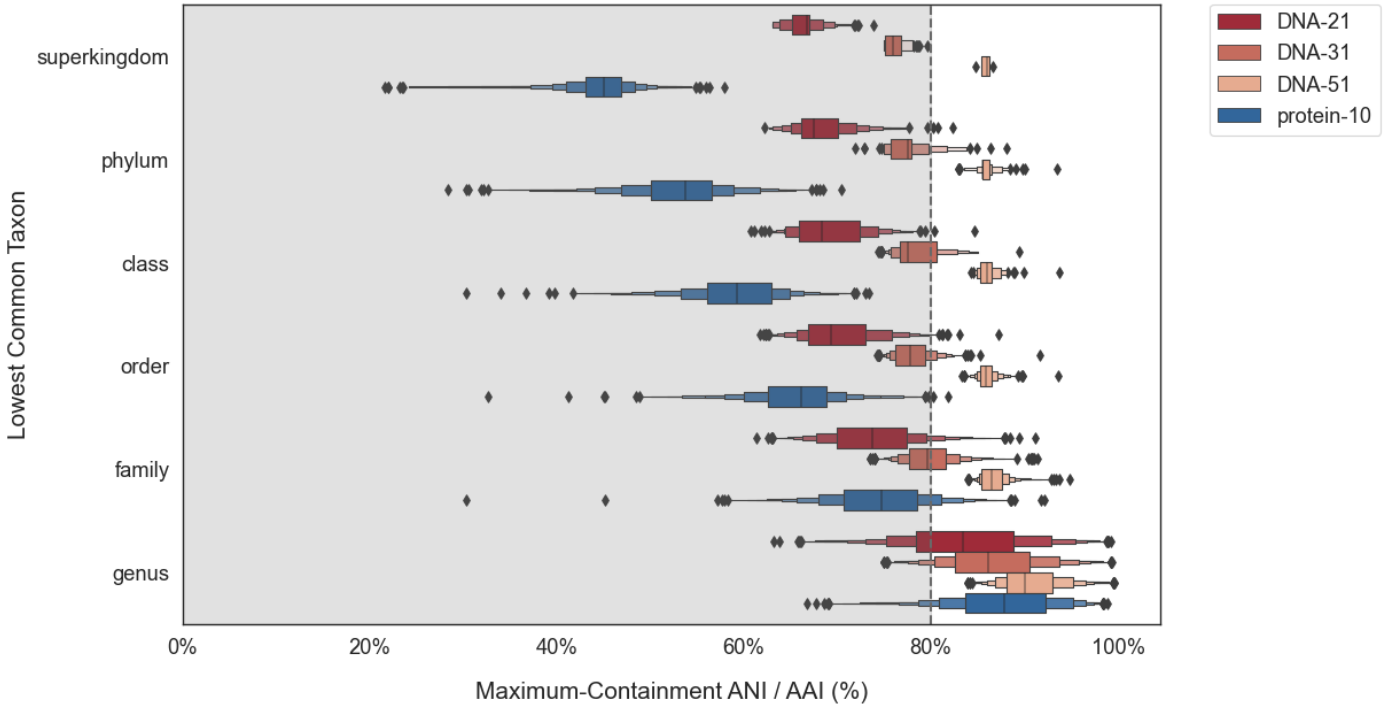

Figure 7: K-mer Based Sequence Identity by Lowest Common Taxon GTDB Evolpaths dataset

Robust taxonomic classification using Sourmash Gather

Ref databases are incomplete (may not have good representation of sp. of interest). Query genomes /mags/ metagenomes are incomplete.

Benchmarking Taxonomic Classification

First, we benchmarked protein-based gather classification using the high quality, highly complete reference genomes within the GTDB representative genome set. For each genus, we randomly selected one reference genome for inclusion in the benchmarking reference database (n=9428). For each genus with at least two species clusters, we randomly selected a second species within that genus for the test set of genomes (n=3911). Thus, each test genome shares genus-level taxonomy with one (and only one) genome in the reference database. Since we know that no test genome shares species-level taxonomy with the reference database, we used the lowest/least common ancestor approach described above to report taxonomic classifications at the genus level. To assess the impact of 6-frame translation of nucleotide sequence on classification accuracy, we compared classification accuracy between published proteome queries and (6-frame) translated genome queries. Using the same reference database, we selected an environmental dataset [**???**] to assess the impact of genome completeness on taxonomic classification.

We compared gather-LCA classification to `GTDB-Tk`, a tool ...

We also generated nucleotide and protein Scaled MinHash reference databases for all GTDB representative genomes (release 95, n=31,910).

Classification is robust to incompleteness of query genome and reference database

Classification is robust to query genome contamination

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].

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 [21] (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)

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 [8,22,23], 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 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].

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 [24]. 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 [8], 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 [19] presented a method to estimate the mutation rate between MinHash sketches while accounting for the non-independence of mutated k-mers. Using [25], 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 [26]...

GTDB “Evolutionary Paths” Dataset

The Genome Taxonomy Database (GTDB) provides a genome-based taxonomy for bacterial and archaeal genomes [27]. To assess the utility of Scaled MinHash techniques across evolutionary distance, we generated a series of “evolutionary paths” from the set of 31k representative GTDB genomes. For each genus with at least two species clusters, one representative genome was randomly selected as a path “anchor” genome. To build the path, one additional genome was selected from the representative genomes matching the anchor’s taxonomy at each higher taxonomic rank. Each path

thus consisted 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 release 95 resulted in 2957 paths comprised of 6690 unique genomes (6543 Bacteria, 237 Archaea). These paths include genome comparisons across 33 phyla (29 Bacteria, 4 Archaea), covering roughly a quarter of the 129 phyla (111 Bacteria, 18 Archaea) in GTDB release 95. 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.

Scaled MinHash Sequence Identity 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 ([28]; 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 [5] and conducts reciprocal best-hit mapping via DIAMOND [29]. 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 [30], **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

To take advantage of the increased evolutionary distance comparisons offered by protein k-mers, we apply compositional analysis with sourmash gather [8] 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 [8]. **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 [31].

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 [32] on the FARM cluster at UC Davis, using practices outlined in [33].

To Do

k-size selection for optimal comparisons / distance estimation

- num shared k-mers at different ksizes
- e.g. k=7 much more common – share far more k-mers. I assumed this would hurt, rather than help classification. Check!
- do rankinfo on each database??

because kmer size matters –> conversion to AAI is useful!? conversion to AAI does two things: accounts for k-mer length, ...

Classification

0. fix thumper (refactor branch) –> working + tests
1. implement “leave one xx clade out classification check” –> instead of just ignoring exact matches, ignore any matches in same species/genus/family
2. prelim figure for tara classification vs GTDB-Tk vs BAT

- classification of incomplete genomes
- for “contaminated” genomes .. can we randomly add contigs from diff species, see the impact? Like classification still works until xx% contaminated with something present in our database?

3. ksize diffs for classification? k=7 vs k=10 vs k=11?

- time, sensitivity, specificity

4. virus testing?

benchmarking :: Leave one out classification

leave one clade out version? see CAT/BAT paper)

[protein vs dna]

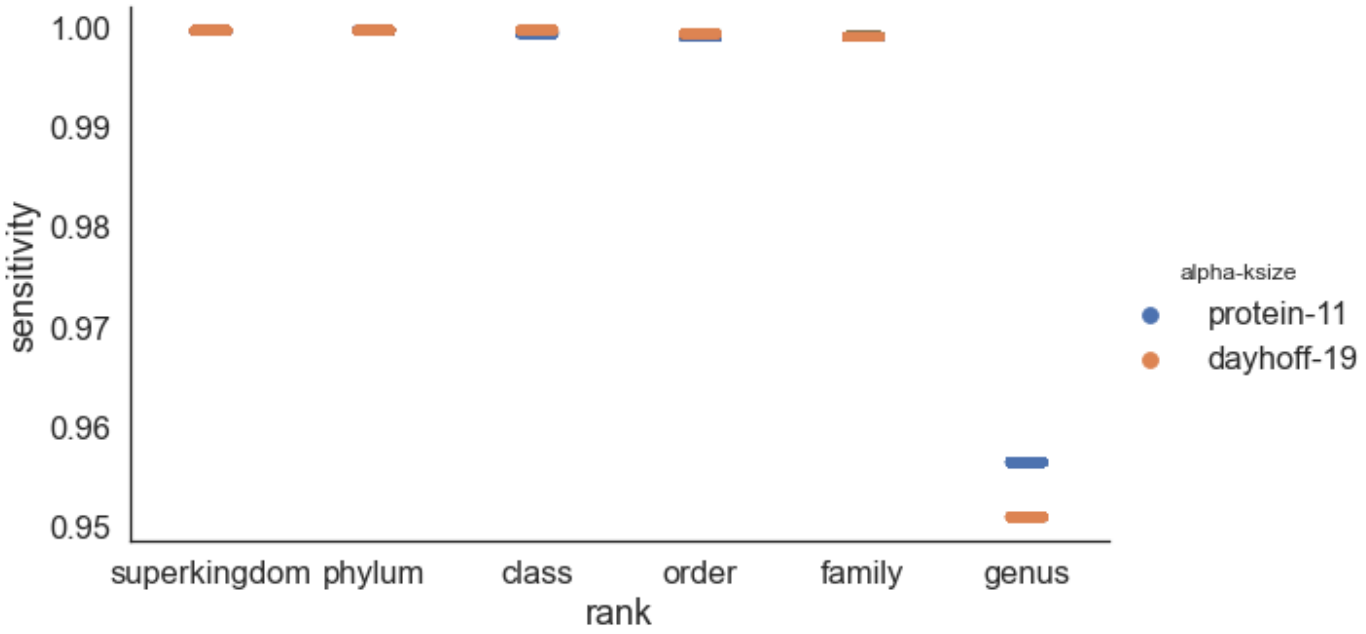

Figure 8: Protein classification sensitivity

include 6-frame translation works well for database search (sensitivity/specificity of Prodigal-translated vs 6-frame translated)

CAT/BAT paper [34] ("cat" = contig annotation, "bat" = bin annotation)

main point: more k-mers are shared = more k-mers available for matching

Classification of incomplete and contaminated genomes

virus classification

euk classification?? Too much.

median AAI across GTDB?

alphabet and k-size selection for optimal distance estimation

- num shared k-mers at different ksizes
- e.g. k=7 much more common – do rankinfo on each database!

Comparison with other alignment-free methods (advantages, disadvantages, etc)

Alignment-based metrics are looking at the specific sequence variation of aligned regions, while k-mer based comparisons are comparing shared k-mers vs distinct k-mers. Since each nucleotide polymorphims generates mutated k-mers with an expected frequency, k-mer containment estimates can be used to accurately estimate both the Average Nucleotide Identity and Average Amino Acid Identity [1,19]

Using nucleotide k-mers This property allows for low-level homology detection at the n

Add'l thoughts, etc

```
** core vs accessory distances **
ANI/AAI == really getting at _core_
```

Leftover Text

Here, we apply k-mer based sequence identity estimation to generate taxonomic classification from the compositional results.

apply k-mer based sequence identity estimation with known taxonomic thresholds to report the most likely taxonomy for a given query genome.

While more protein k-mers are shared across genomes within the same genus (and different species), min-set-cov + LCA allows us to find/report the most similar genome.

Availability of data and materials

Competing Interests

The authors declare that they have no competing interests.

Funding

Authors' Contributions

Author	Contributions
NTP	Conceptualization; Methodology; Writing - Original Draft; Writing - Review and Editing; Visualization; Supervision; Funding Acquisition
...	...
...	...
...	...
CTB	Conceptualization; Methodology; Writing - Original Draft; Writing - Review and Editing; Visualization; Supervision; Funding Acquisition

Acknowledgements

Thank you to all the members and affiliates of the Lab for Data-Intensive Biology at UC Davis for providing valuable feedback on this manuscript. This manuscript was written using manubot [35] and is available in a GitHub repository [36].

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 (2016-06-20) 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. **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 (2018-03-01) https://doi.org/ghtrdq
   DOI: 10.1128/aem.00014-18 · PMID: 29305502 · PMCID: PMC5835724

4. **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

5. **Fast and sensitive protein alignment using DIAMOND**
   Benjamin Buchfink, Chao Xie, Daniel H Huson
   Nature Methods (2014-11-17) https://doi.org/gftzcs
   DOI: 10.1038/nmeth.3176 · PMID: 25402007

6. **Fast and sensitive taxonomic classification for metagenomics with Kaiju**
   Peter Menzel, Kim Lee Ng, Anders Krogh
   Nature Communications (2016-04-13) https://doi.org/f8h4b6
   DOI: 10.1038/ncomms11257 · PMID: 27071849 · PMCID: PMC4833860

7. **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 (2019-11-05) https://doi.org/ghtqmb
   DOI: 10.1186/s13059-019-1841-x · PMID: 31690338 · PMCID: PMC6833257

8. **Lightweight compositional analysis of metagenomes with sourmash gather**
   Luiz Irber, C. Titus Brown
   Manubot (2021-01-11) https://dib-lab.github.io/2020-paper-sourmash-gather/

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 (2019-12-04) 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 (2020-09-10) https://doi.org/ghtqrz
DOI: 10.1186/s13059-020-02159-0 · PMID: 32912225 · PMCID: PMC7488264

12. **Toward a More Robust Assessment of Intraspecies Diversity, Using Fewer Genetic Markers▿**
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 (2017-07-21) 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 (2018-11-30) 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 (2020-01-14) 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
Cold Spring Harbor Laboratory (2020-08-01) 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. **The statistics of *k*-mers from a sequence undergoing a simple mutation process without spurious matches**
Antonio Blanca, Robert S. Harris, David Koslicki, Paul Medvedev
Cold Spring Harbor Laboratory (2021-02-09) https://doi.org/fq3g
DOI: 10.1101/2021.01.15.426881

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

Jesse Shapiro
Peer Community In Genomics (2020-09-24) https://doi.org/ghvgd3
DOI: 10.24072/pci.genomics.100001

21. **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

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

23. **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

24. **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) https://doi.org/gjnvx4
DOI: 10.1093/bib/bbz083 · PMID: 31624847

25. **KoslickiLab/mutation-rate-ci-calculator**
KoslickiLab
(2021-04-14) https://github.com/KoslickiLab/mutation-rate-ci-calculator

26. https:https://github.com/dib-lab/sourmash/pull/1270

27. **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 (2020-04-27) https://doi.org/ggtbk2
DOI: 10.1038/s41587-020-0501-8 · PMID: 32341564

28. **dparks1134/CompareM**
Donovan Parks
(2021-03-23) https://github.com/dparks1134/CompareM

29. **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 (2010-03-08) https://doi.org/cktxnm
DOI: 10.1186/1471-2105-11-119 · PMID: 20211023 · PMCID: PMC2848648

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

31. **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 (2018-10-30) https://doi.org/ggc9db
DOI: 10.1186/s13059-018-1554-6 · PMID: 30373669 · PMCID: PMC6206640

32. **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

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

34. **Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT**
F. A. Bastiaan von Meijenfeldt, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, Bas E. Dutilh
Genome Biology (2019-10-22) https://doi.org/ggfm6r
DOI: 10.1186/s13059-019-1817-x · PMID: 31640809 · PMCID: PMC6805573

35. **Open collaborative writing with Manubot**
Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653

36. https://github.com/bluegenes/2021-ani-paper