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

This manuscript (permalink) was automatically generated from bluegenes/2022-paper-protein-kmers@576c7bf on June 10, 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

- **Taylor E. Reiter**
  0000-0002-7388-421X · taylorreiter · ReiterTaylor
  Department of Population Health and Reproduction, University of California, Davis · Funded by Grant XXXXXXXX

- **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

Advancements in sequencing over the past decades have made it feasible to investigate the vast global diversity of microbial organisms via direct sequencing of environmental samples (metagenomics). These techniques have expanded and reshaped our understanding of evolutionary relatedness across the tree of life and allowed us to move beyond organismal isolates to investigate the structure and function of microbial communities (CITE).

Metagenomic analyses rely on our ability to make sense of bulk sequencing reads by assigning taxonomic and functional groupings. However, the methods and databases used for characterization impact both the extent and accuracy of classification. As the scale of genomic sequencing continues to grow, fast and low-memory methods for estimating sequence similarity have become critical for conducting tasks ranging from taxonomic classification to phylogenetic analysis on large-scale datasets [1,2]. However, many of these methods struggle with classification specificity, with some methods reporting false positive rates as high as 25% on short read metagenomic datasets prior to thresholding [3]. At the same time, classification techniques often can suffer from limited sensitivity when comparing highly polymorphic sequences or classifying organisms from groups underrepresented in reference databases. For understudied and diverse habitats such as soil, metagenomic classification methods often only categorize a small fraction of metagenomic data, and even well-studied environments such as the human gut can produce significant uncharacterized metagenome content (CITE).

As protein sequence is more conserved than the underlying DNA sequence, protein-based comparisons have long been the gold-standard approach across larger evolutionary distances [4,5]. Protein-based metagenomics taxonomic classification approaches typically have increased sensitivity relative to nucleotide methods [6,7,8,9,10,11]. Whole-proteome relatedness indices such as Amino Acid Identity (AAI) can be used to determine whether uncharacterized sequences belong to known taxonomic groups or represent truly novel sequence. As we continue to sequence more of the biosphere, there remains a need for fast and accurate alignment-free sequence comparison tools with protein-level sensitivity.

Alignment-free methods using k-mers, short sequences of length k, can quickly compare and classify metagenomic datasets particularly when used with subsampling methods such as MinHash [1] and FracMinHash [12]. While the majority of k-mer methods utilize nucleotide k-mers, amino acid k-mers (k_{aa}-mers) have shown some promise for functional screening and annotation [11,13,14]. Here, we show that k_{aa}-mer comparisons robustly estimate Average Amino Acid Identity across large evolutionary distances, even while using FracMinHash k-mer subsampling methods. We then use FracMinHash k_{aa}mer sketches to tackle two classification challenges: taxonomic classification of assembled genomes, and compositional analysis of metagenomes. Our results suggest that k_{aa}mer sequence analysis can facilitate large-scale assembly-based and assembly-free metagenomic analyses, even when sequenced organisms are only distantly related to organisms available in reference databases.

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 has been used for taxonomic classification and similarity detection between closely related genomes (genus-level to strain-level

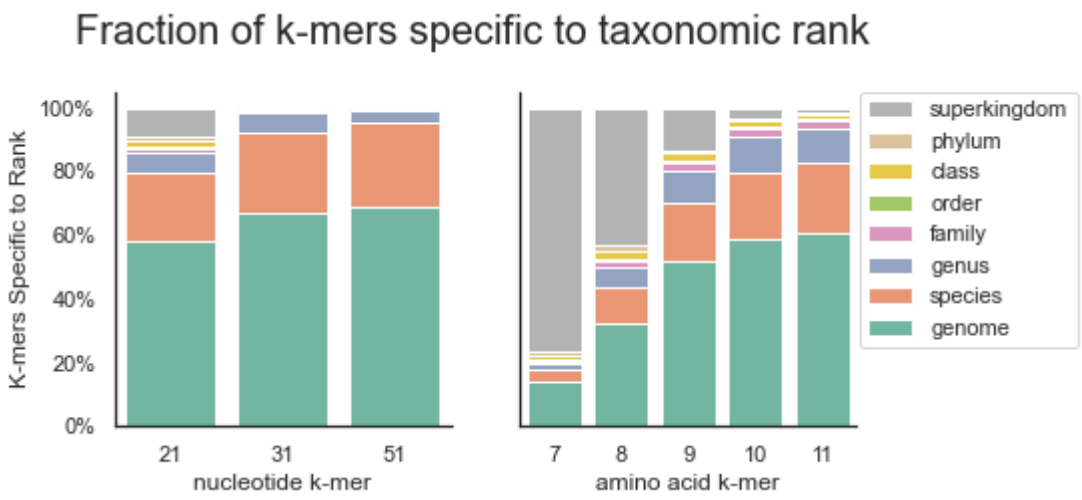
comparisons using k-mer lengths between 21-51) [1,**metapalette?**]. At larger evolutionary distances, accumulated nucleotide divergence limits the utility of exact nucleotide k-mer matching. Amino acid k-mers (kaa-mers) retain the benefits of fast, alignment-free exact k-mer matching, but with increased tolerance for evolutionary divergence. Here, we evaluate the utility of amino acid k-mers for a wide range of genomic and metagenomic applications, including sequence distance estimation and taxonomic classification.

Amino Acid k-mers can be used to discriminate between taxa

The Genome Taxonomy Database (GTDB) provides a genome-similarity-based taxonomy for bacterial and archaeal genomes [15]. GTDB release `rs202` encompasses 258,407 genomes from 47,895 species. We begin by assessing the prevalance of nucleotide and amino acid k-mers of different k-mer lengths within assemblies included in GTDB.

To make analyses at this scale tractable, we use FracMinHash sketching to randomly subsample all available k-mers, retaining approximately 1/1000 nucleotide k-mers and 1/200 amino acid k-mers [12]. DNA FracMinHash sketches have been shown to representatively subsample genome datasets [12]. 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 [16] 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/proteomes 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. 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 (CTB: do we want to say specificity here?) at different taxonomic ranks. We choose amino acid k-mer lengths 7 and 10 for our primary analyses, and have set a default kaa-mer length of 10 within `sourmash`.

Evolutionary Paths 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 from genus to superkingdom.

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.

For DNA comparisons, each genome was sketched from the available genome downloaded from genbank. For protein comparisons, we conducted both protein comparisons and translated comparisons. In both workflows, all anchor genomes were sketched from available proteomes (either downloaded or generated via Prodigal, as above). For the direct protein assessment, comparison proteomes were also sketched from the available proteome. For these sketches, k-mer containment results are equally valid in both directions, so we report the mean containment for the comparison alongside the Jaccard Index. For the second workflow, comparison genomes were 6-frame translated to build protein kaa-mers. As 6-frame translation introduces erroneous k_{aa}-mers, we report only the containment estimate relative to the anchor proteome (CTB: perhaps note that the intuition here is that for long k, only correct k-mers will match). We term this "anchor containment", where the trusted genome is the "anchor" upon which we base the comparison. We conduct k-mer comparisons using `sourmash` FracMinHash sketches default fractional scaling: 1/1000 k-mers from DNA sketches and 1/200 k-mers for protein sketches (including 6-frame translated sketches). For amino acid k-mers, we focus on k-mer lengths k=7 and k=10, which are closest to nucleotide k-mer lengths 21 and 31. To verify results and estimate the impact of FracMinHash scaling, we also conducted all comparisons using all available k-mers (no subsampling).

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

We begin by assessing k-mer containment across the 6 comparisons (each genome compared with the anchor genome) within each of 4095 evolutionary paths. When plotted by the rank of the lowest common ancestor, the dynamic range of containment values is much larger for kaa-mer comparisons. While DNA k-mers can provide resolution at the genus level, log-transformed containment values for protein k-mers continue to decrease, providing resolution for comparisons even between genomes of different phyla. Average containment estimated from proteome sequence is very similar to anchor containment estimated from 6-frame translation of genome sequence, suggesting that either value

can be used for this type of comparison. We obtained similar results when comparing all available k-mers, suggesting that these results are not affected by FracMinHash scaling (*Supplemental Figure XX*).

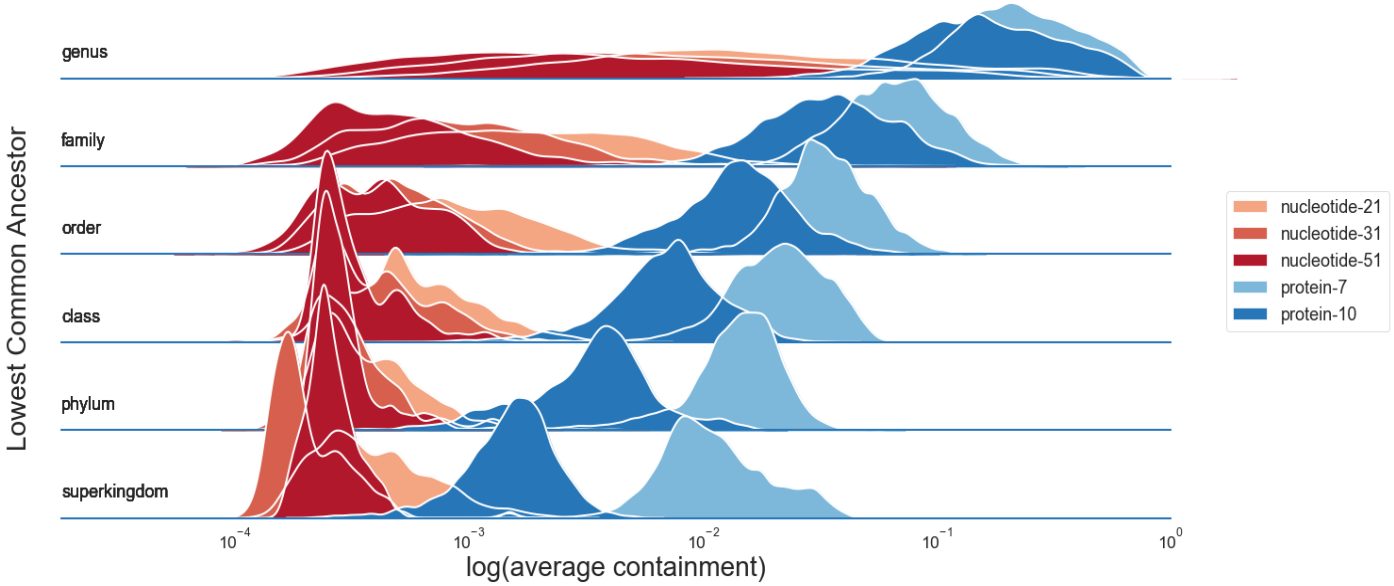


Figure 1: Protein k-mers are shared at higher taxonomic ranks Default scaled values 1000, 200

FracMinHash kaa-mer containment accurately estimates sequence similarity

Pairwise Overall Genome Relatedness indices (OGRI's) such as Average Nucleotide Identity (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 have been shown to be robust measures of overall pairwise genome relatedness even for highly incomplete datasets, such as those comprised of only ~4% of the genome or 100 genes [17,18]. These measures have utility for phylogenomic comparisons and taxonomic classification, but alignment-based analyses not tractable for large-scale comparisons.

As first noted by [19] and extended for MinHash [1], DNA k-mer comparisons can be used to estimate the sequence similarity between two genomes without alignment. These methods have primarily assumed a simple mutational model of random nucleotide substitution in order to estimate sequence identity from the Jaccard Index between the two samples. While this strategy permits ANI estimation at the scale of whole-databases, they work best with high quality genomes with high similarity (>90% ANI) [1,11,20]. Compared with Jaccard, the Containment Index permits more accurate estimation of genomic distance, particularly for genomes of very different lengths [11,21,22]. As described in [23], using the Fractional Containment Index $C_{\text{frac}}(A, B)$, estimated using a k-mer size of k, we can obtain a point estimate of sequence identity between two genomes as follows:

$$ANI = C_{\text{frac}}(A, B)^{1/k}$$

Note that this method assumes a simple mutational model of random substitution and estimates sequence divergence solely using the fraction of shared and divergent k-mers between the two

FracMinhash sketches. When FracMinhash sketches are instead generated with amino acid k_{aa}-mers of length $k\ aa$, the corresponding equation can be used to generate alignment-free AAI estimates.

$$AAI = C_{\text{frac}}(A, B)^{1/k\ aa}$$

To account for the variance of FracMinHash subsampling, we also derive confidence intervals around this point estimate (see Methods for details).

To assess the accuracy and utility of k_{aa}-mer Amino Acid Identity estimation, we compared our results with alignment-based methods that leverage three different algorithms: EzAAIb (BLASTp), EzAAIm (MMSeqs2), and CompareM(DIAMOND). As BLAST-based alignment remains the gold-standard method, we compare all AAI values to the BLAST AAI values. As above, we utilize anchor containment for comparisons involving 6-frame translated sketches.

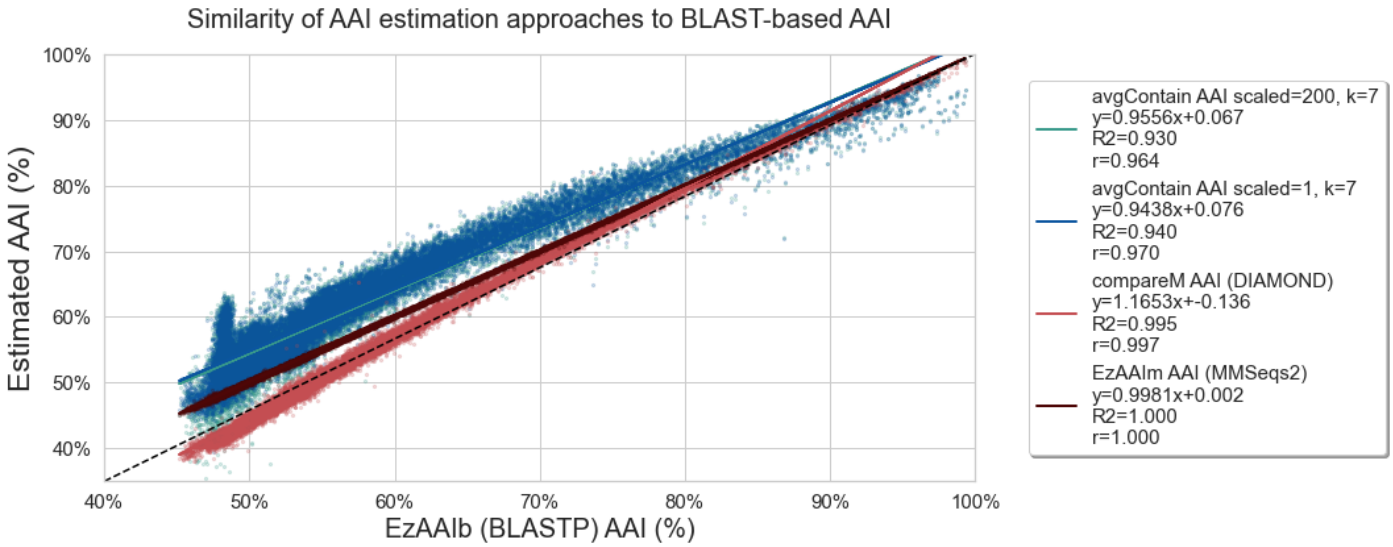


Figure 2: FracMinHash AAI vs BLASTP alignment based AAI (CTB: suggest putting parameters in caption rather than in figure legend)

Protein k-mer containment and AAI can be used for taxonomic classification

Given that protein k-mers facilitate similarity estimation across these larger evolutionary distances, we next assess the utility of protein k-mers for taxonomic assignment, both for metagenome breakdown/classification and for assembled genomes.

Metagenome breakdown using protein k-mers

As developed in Irber et al., 2022 [12], minimum set cover of nucleotide k-mers can be used to find the set of genomes that cover all known k-mers in a metagenome. This approach, implemented in `sourmash gather`, works by using k-mer containment relative to reference genomes ("anchor containment", as above) and "assigning" metagenome k-mers iteratively to the reference genome with highest containment. Anchor containment is then re-estimated using the remaining unassigned query k-mers until all known k-mers have been assigned. This step provides us with an ordered list of reference genomes, each of which represent a non-overlapping portion of the metagenome. The taxonomy of these matched reference genomes thus represents the closest match for each of these non-overlapping portions of the metagenome. In addition to reporting these exact matches, we can aggregate these taxonomic assignments of these matches to obtain taxonomic summarization at each rank.

Here, we assess the utility of protein k-mers for this application using the same metagenome samples described in Irber et al., 2022 [12]. As metagenome samples are unassembled, we use the 6-frame translation approach described above to obtain protein k-mers for comparison. No modification to the min-set-cov approach is required, as it already relies upon anchor containment to the reference genomes.

add figure: genome-grist mg breakdown, nucl k-mers, prot k-mers, nucl mapping

do we need an additional metagenome w/more divergent genomes, to show advantage of protein methods?

Robust taxonomic classification from protein k-mers

We use a similar approach for taxonomic classification of assembled genomes from protein k-mer containment. We apply the same minimum set cover approach to find the set of reference genomes that cover all known k-mers in our sample (in this case, a genome itself rather than a metagenome). If the most contained reference genome is sufficiently similar (passes default or user-defined threshold) to our query, we can annotate the query with taxonomic information from this reference genome. If not, we can use the genome-based lowest common ancestor approach to classify the query genome to the taxonomic rank where it contains sufficient similarity to matched reference genome sequence.

We select two sets of genomes: first, a set of 1000 genomes from the MGNify project ("MGNify-1000"), which are assembled from human gut and likely to be well-represented in existing databases. We next choose a set of 885 microbial ("Delmont-885"; 820 *Bacteria*, 65 *Archaea*) metagenome-assembled genomes (MAGs) assembled from TARA Oceans metagenomes [24]. As the marine environment is understudied relative to human gut, these genomes are more challenging for classifiers as they are less likely to have close relatives available in reference databases.

To assess the utility of protein k-mers for genome classification, we conduct this classification using three k-mer approaches: direct nucleotide k-mers, 6-frame translated protein k-mers, and direct protein k-mers from prodigal-translated proteomes. Where reference taxnonomic lineages were available (MGNify-1000), we compared our results directly to these annotations. With experimental genomes where no reference taxonomic lineage is available, we assessed our annotation relative to `gtdb-tk` classification [25].

Dataset	Exact Match	Higher Rank	Unclassified (sourmash)	Unclassified (GTDB-Tk)
MGNify-1000	95.7%	4.3%	N/A	N/A
Delmont-885	73.5%	26.5%	1 (0.1%)	15 (1.7%)

to do: - discuss/utilize AAI threshold at all?

Notes

Include Jaccard –> AAI results anywhere? - FracMinHash AAI values produced by Jaccard and Containment (here, average containment) methods are very similar.

Discussion

Protein sequences are more conserved than their underlying DNA sequence, allowing protein k-mer comparisons to match across larger evolutionary distances. Protein sequence matching is also less impacted by sequencing errors due to codon degeneracy.

FracMinHash kaa-mer sketches support whole-proteome analysis at scale

Our results show that amino acid k-mers can be used for global proteome analyses, including distance estimation (AAI) and taxonomic classification. For most microbial proteomes, systematic subsampling via FracMinHash maintains accuracy while enabling much faster comparisons.

Shorter amino acid k-mers (kaa = 7) can be matched even across phyla, suggesting they may be most useful for comparisons at larger evolutionary distances. These k-mers also have the potential to work well for functional analyses. Longer amino acid k-mers (kaa=10+) are more discriminatory and may be best for taxonomic classification.

kaa-mer Amino Acid Identity estimation

Kaa-mer estimation of Amino Acid Identity (AAI) correlates well with mapping-based AAI approaches, while requiring far less memory and computation. Different mapping approaches vary slightly in the AAI reported for a given pair of genomes, suggesting that comparisons are best made between values produced by the same method.

There is one other method that can function at large scale: MIGA webserver (paper has 11,000 pairwise comparisons) – and this method uses hAAI (heuristic AAI), only doing complete AAI if/wehn the hAAI cannot be estimated or is >= 90% (“close to saturation”). MiGA “applies a hierarchical approach: hAAI, AAI, then ANI” to identify the best match genome/proteome. AAI thresholds have been proposed for higher taxonomic ranks, <45%, 45-65% and 65-95% for family, genus, and species [18,26].

diffs vs fastaai: - whole proteome - taxonomy-agnostic -

Here we could envision doing this with protein k-mers –> doing a quick high-scaled proteome search to find the right family, then doing a more detailed DNA/genome analysis.

Containment enables assembly-free protein comparisons

For protein k-mer comparisons to be useful, any DNA queries must be translated into protein sequence. This can limit amino acid comparisons to assembly-based workflows, as assemblies can be reliably translated into predicted Open Reading Frames (ORFs). By using only the k-mer containment estimate relative to reference proteomes, we can obtain accurate Amino Acid Identity estimates directly from DNA sequence. In this way, we can use the more permissive nature of protein analyes for assembly-free genome and metagenome assignment.

Taxonomic Assignment is database-dependent

(but protein helps with sensitivity + min-set-cov helps with specificity) discuss in relation to: Kaiju, CAT/BAT, MEGAN-prot, MMSeqs taxonomy (+ probably move some of this to intro)

K-mer based taxonomic assignment relies upon matching k-mers found in previously sequenced reference proteomes. While this approach will always be database-dependent and improved by

presence of closely-related proteomes in the database, protein-based matching allows for classificaiton at larger evolutionary distances. While protein matching increases the sensitivity by matching across synonymous substitutions in the DNA sequence

classification LCA approaches often suffer from sensitivity/specificity trade-offs Here, the use of `sourmash gather` minimum set cover approach assigns each protein k-mer to its most likely/parsimonious match based on presence of other proteome k-mers present in the query genome/metagenome.

distinguishing features this vs kaiju: - min-set-cov –> low false pos - fracminhash –> faster, smaller databases (though might need to increase scaled value) -

Limitations

- Note that comparing two 6-frame translated datasets is not recommended, as there is no mechanism to exclude erroneous k-mers introduced during translation.
- by using proteins, we're only looking at coding regions/ignoring noncoding (same as with all protein analyses)

Future directions and utility

- functional comparisons/all the k-mers
- abundance comparisons with cosine, f_unique_weighted, etc
- clustering at protein level
- While eukaryotic datasets are out of scope of this paper, these methods should work fine (except repeats/indels) – discuss a bit and/or do some analyses?

Using genomes from the Genome Taxonomy Database (GTDB) we develop showcase amino acid k-mer distributions across phyla and demonstrate taxonomic classification using sketch containment.

Protein search has long been used for comparisons conducted at increased evolutionary distances. By using amino acid FracMinHash sketches, we can make global protein similarity assessment tractable for the current scale of sequencing.

Main points: - Protein k-mers approximate proteomes - FracMinHash sketches can be used for global proteome comparisons - Longer protein k-mers can be used for taxonomic classification and metagenome breakdown - maximum containment enables these things directly from DNA sequence - For many classification methods, an increase in sensitivity is accompanied by a concomitant decrease in specificity, yielding large number of false positives, often even on mock communities [3,27].

We expect sourmash protein k-mer analyses to be especially useful for species with few representatives in published databases.

For many mapping-based AAI approaches, it is important to report both the percent identity of matched regions and the fraction of the genomes that were mapped. This prevents believing artificially high similarity values when only small fractions of the genomes overlap. In contrast, containment-based AAI by necessity considers all of the sequence of at least one of the two genomes, as the containment measure is the matched k-mers divided by the total k-mers in the query genome. Since containment is directional, when both proteomes are equally trusted (e.g. neither set of protein k-mers is being 6-frame translated from genome sequence), then the average containment considers the entire set of protein sequence from both proteomes. While the AAI value is based on this measure, it may be useful to also consider/report the percent containment of each proteome alongside the AAI value, as this describes the percent of each proteome that matched.

While several studies have proposed utilization of more complex evolutionary models, the simple mutational model accurately estimates nucleotide similarity when compared with mapping-based estimates [23].

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.

While here we have focused on demonstrating utility of amino acid k-mers for global proteome similarity, the correlation of unique amino acid 10-mers with genes should enable gene-level analyses, if sketching with sufficient k-mers.

Methods

Large-scale k-mer comparisons with FracMinHash sketches

FracMinHash sketching, as implemented in sourmash [12,28,29], is a MinHash variant that uses a scaling factor to subsample the unique k-mers in the dataset to the chosen fraction (1/`scaled`). As k-mers are randomized prior to systematic subsampling, FracMinHash sketches are representative subsets that can be used for comparisons across datasets sketched with consistent k-mer lengths and scaling factors.

While FracMinHash sketches can be used to estimate both the Jaccard Index [1] and Containment Index [21], containment has been shown to permit more accurate estimation of genomic distance when genomes or datasets differ in size [21,22,30,31]. We focus here on the utility of containment comparisons for similarity estimation. Containment comparisons are directional: the containment of genome A in sample B is the interection of k-mers in A and B divided by the k-mers in genome A (and vice versa). Thus, two containment values can be estimated for a given pairwise comparison. The choice of which containment value to use (or whether to average the two values) depends on the particular comparison. FracMinHash containment has been shown to be an unbiased estimator of the true containment index, as long as the sketches contain sufficient k-mers for comparison or utilize a high-quality estimation of the true cardinality of the dataset [12,23].

Sourmash supports sketching from either nucleotide or protein input sequence, to generate either nucleotide or protein FracMinHash sketches. We generated nucleotide and protein sketches directly from genome and protome files, respectively. All genome sequences were sketched with sourmash v4.2.1 using the `sourmash sketch dna` command, k-mer sizes of 21,31,51, a scaling factor of 1000. All proteome sequences were sketched with sourmash v4.2.1 using the `sourmash sketch protein` command at protein k-sizes (*kaa-mer sizes?*) of 7-12 and a scaling factor of 200. Sourmash also supports 6-frame translation of nucleotide sequence to amino acid sketches. 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 200.

In select cases, we also conducted comparisons using all available k-mers, rather than using FracMinHash sketch subsampling. While `sourmash` sketching is not optimized for this use case, we can generate these complete k-mer sketches using the same `sourmash` commands with a scaling factor of 1 (`scaled`=1).

Anchor Containment analysis for protein comparisons directly from DNA sequence

For protein k-mer comparisons to be useful, any DNA queries must be translated into protein sequence. This typically limits amino acid comparisons to assembly-based workflows, as assemblies can be reliably translated into predicted Open Reading Frames (ORFs). With k-mer methods, 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 true ORFs. The remaining erroneous k-mers greatly impact the Jaccard Index (set similarity) when comparing samples, but only impact the containment index in one direction (relative to the 6-frame translated set). The containment estimate relative to reference proteomes will be an accurate comparison directly from DNA sequence. We term this "anchor" containment, where the trusted genome is the "anchor" upon which we base the comparison. Since 6-frame translation should always yield excess k-mers relative to genomes of similar size, this desired containment will generally be the larger of the two containment values (maximum containment).

To facilitate these comparisons within `sourmash`, we have implemented "maximum containment," a shorthand method to select the greater of the two containment values. The maximum containment method may also provide advantages for genomes with potential contamination, as containment will always be relative to the smaller, and presumably less contaminated, genome. However, highly incomplete genomes may overestimate similarity with this method, so we suggest using containment relative to the more trusted sample if known, or considering both containment values when comparing two genomes of approximately equal quality. Note that comparing two 6-frame translated datasets is not recommended, as there is no mechanism to exclude erroneous k-mers introduced during translation.

Estimating Average Amino Acid Identity

MinHash Sketch Jaccard has been shown to correlate well with ANI at high sequence identities (>=90% sequence identity) [1]. Recently, Blanca et al, 2021 [32] presented a method to increase the accuracy of sequence similarity estimation from MinHash Jaccard by recognizing that k-mers generated from mutated sequence are not independent. Hera et al, 2022 [23] extended this approach to estimate sequence identity from FracMinHash Containment estimates. Each of these methods assumes a simple mutational model, with equal substitution probability for each nucleotide, and then estimates sequence identity based on k-mer comparisons. Here, we note that there is nothing unique to nucleotide sequence included in these equations. By applying the same equations to comparisons between amino acid k-mer sketches, we can estimate average Amino Acid Identity (AAI) between proteomes. For this application, we maintain the assumption of a simple mutational model of equal substitution probability at each position, but recognize that it now applies to any amino acid, rather than any nucleotide.

Using the Fractional Containment Index...

$$\text{scale} := \frac{|\mathbf{FRAC}_s(A) \cap \mathbf{FRAC}_s(B)|}{|\mathbf{FRAC}_s(A)| \left(1 - (1-s)^{|A|}\right)} \mathbb{1}_{|\mathbf{FRAC}_s(A)|>0}$$

to do: ADD EQUATION

The equation for sequence similarity estimation (ANI or AAI) from FracMinHash Containment is reproduced here for completeness (see [23] for details).

See [23] for these and other analytical details.

Immplementation of ANI and AAI estimation

We provide an implementation of Fractional Containment to average sequence identity (ANI/AAI) in the software package `sourmash`, which is implemented in Python and Rust and developed under the BSD license [28,29]. ANI and AAI values can be reported from sequence comparisons The distance estimation equations can be found in the `distance_utils.py` file and ANI/AAI values can be reported from a variety of `sourmash` comparison and search commands as of version `4.4`. sourmash is available at github.com/sourmash-bio/sourmash. The results in this paper were generated with sourmash v4.4.1.

FracMinHash Amino Acid Identity Correlates with Alignment-based Methods

To assess whether k-mer methods can be used to approximate AAI, we ran generated alignment AAI values for each pairwise comparison using methods that leverage different mapping algorithms: EzAAIb (BLASTp), EzAAIm (MMSeqs2), and CompareM (DIAMOND). As BLAST-based alignment remains the gold-standard method, we compare all AAI values the BLAST AAI values.

EzAAI v1.12 [**doi?** 10.1007/s12275-021-1154-0] was used to run BLAST-based and MMSeqs-based Amino Acid Identity. The EzAAI workflow begins with PRODIGAL-based translation of genome sequence [5], followed by reciprocal BLAST [**doi?** 10.1016/S0022-2836(05)80360-2] or MMSeqs2 [33] alignment. For both, we utilized EzAAI default parameters: 40% coverage threshold, 40% sequence identity threshold. CompareM v0.1.2 ([34]; 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 [16]. 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 [35/], **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 FracMinHash 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 [12] 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 queries with high sequence identity (ANI) to reference matches, we classify the query sequence as a member of the reference taxonomic group, as in [12]. 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. 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 [36].

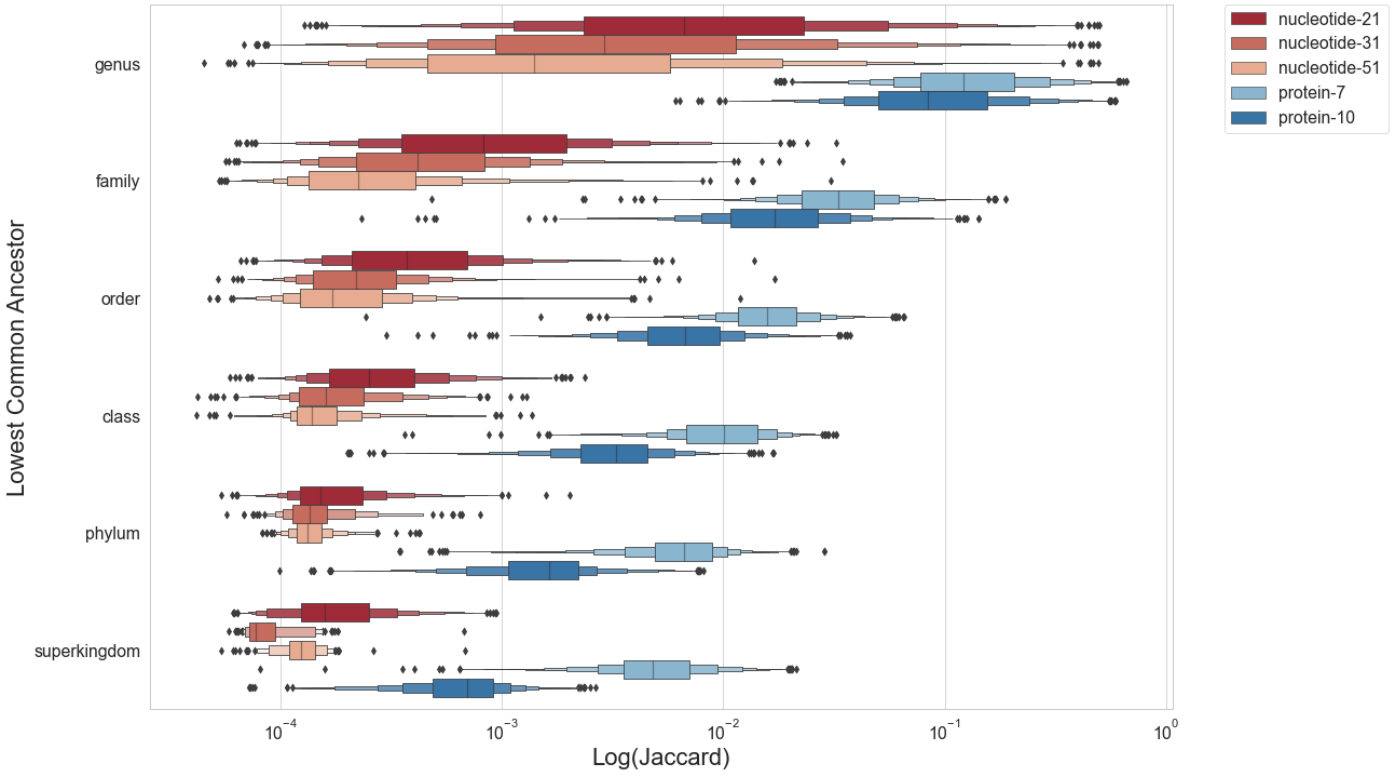
These taxonomic utilities are implemented in the `sourmash taxonomy` module, and classifications were run on gather output via `sourmash tax genome`.

Workflows and Computing Resources

Reproducible workflows associated with this paper are available at https://github.com/bluegenes/2022-protein-kmers-workflow (ADD DOI for release), with datasets available at OSF (XX). All workflows were executed using snakemake >= 5.26 [37)] on the FARM cluster at UC Davis, using practices outlined in [38].

Supplemental

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



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

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

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

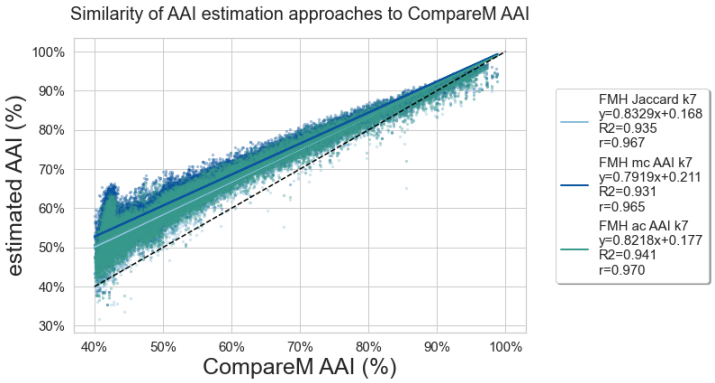


Figure 3: FracMinHash AAI vs CompareM Scaled 1

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. **Improved metagenomic analysis with Kraken 2**
Derrick E Wood, Jennifer Lu, Ben Langmead
Genome Biology (2019-12) https://doi.org/ggfk55
DOI: 10.1186/s13059-019-1891-0 · PMID: 31779668 · PMCID: PMC6883579

3. **Evaluation of taxonomic profiling methods for long-read shotgun metagenomic sequencing datasets**
Daniel M Portik, CTitus Brown, NTessa Pierce-Ward
Bioinformatics (2022-02-02) https://doi.org/hhqs
DOI: 10.1101/2022.01.31.478527

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 (2015-01) https://doi.org/gftzcs
DOI: 10.1038/nmeth.3176 · PMID: 25402007

6. **A review of methods and databases for metagenomic classification and assembly**
Florian P Breitwieser, Jennifer Lu, Steven L Salzberg
Briefings in Bioinformatics (2019-07-19) https://doi.org/gdq95k
DOI: 10.1093/bib/bbx120 · PMID: 29028872 · PMCID: PMC6781581

7. **Fast and sensitive taxonomic assignment to metagenomic contigs**
M Mirdita, M Steinegger, F Breitwieser, J Söding, E Levy Karin
Bioinformatics (2021-09-29) https://doi.org/gnnprm
DOI: 10.1093/bioinformatics/btab184 · PMID: 33734313 · PMCID: PMC8479651

8. **Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT**
FABastiaan von Meijenfeldt, Ksenia Arkhipova, Diego D Cambuy, Felipe H Coutinho, Bas E Dutilh
Genome Biology (2019-12) https://doi.org/ggfm6r
DOI: 10.1186/s13059-019-1817-x · PMID: 31640809 · PMCID: PMC6805573

9. **MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs**
Daniel H Huson, Benjamin Albrecht, Caner Bağcı, Irina Bessarab, Anna Górska, Dino Jolic, Rohan BH Williams
Biology Direct (2018-01) https://doi.org/gnnprp
DOI: 10.1186/s13062-018-0208-7 · PMID: 29678199 · PMCID: PMC5910613

10. **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

11. **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

12. **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

13. **Fast protein database as a service with kAAmer**
Maxime Déraspe, Sébastien Boisvert, François Laviolette, Paul H Roy, Jacques Corbeil
Bioinformatics (2020-04-02) https://doi.org/ggrj73
DOI: 10.1101/2020.04.01.019984

14. **Single-cell transcriptomics for the 99.9% of species without reference genomes**
Olga Borisovna Botvinnik, Venkata Naga Pranathi Vemuri, NTessa Pierce, Phoenix Aja Logan, Saba Nafees, Lekha Karanam, Kyle Joseph Travaglini, Camille Sophie Ezran, Lili Ren, Yanyi Juang, … CTitus Brown
Bioinformatics (2021-07-10) https://doi.org/gns4sg
DOI: 10.1101/2021.07.09.450799

15. **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

16. **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

17. **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

18. **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

19. **An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data**
Huan Fan, Anthony R Ives, Yann Surget-Groba, Charles H Cannon
BMC Genomics (2015-12) https://doi.org/f7s6tp
DOI: 10.1186/s12864-015-1647-5 · PMID: 26169061 · PMCID: PMC4501066

20. **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

21. **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

22. **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

23. **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

24. **Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes**
Tom O Delmont, Christopher Quince, Alon Shaiber, Özcan C Esen, Sonny TM Lee, Michael S Rappé, Sandra L McLellan, Sebastian Lücker, AMurat Eren
Nature Microbiology (2018-07) https://doi.org/gdvhp5
DOI: 10.1038/s41564-018-0176-9 · PMID: 29891866 · PMCID: PMC6792437

25. **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

26. **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

27. **Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome**
Donovan H Parks, Fabio Rigato, Patricia Vera-Wolf, Lutz Krause, Philip Hugenholtz, Gene W Tyson, David LA Wood
Frontiers in Microbiology (2021-04-20) https://doi.org/gmhdhm
DOI: 10.3389/fmicb.2021.643682 · PMID: 33959106 · PMCID: PMC8093879

28. **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

29. **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

30. **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

31. **The minimizer Jaccard estimator is biased and inconsistent***
Mahdi Belbasi, Antonio Blanca, Robert S Harris, David Koslicki, Paul Medvedev
Bioinformatics (2022-01-17) https://doi.org/gpm78w
DOI: 10.1101/2022.01.14.476226

32. **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
Bioinformatics (2021-01-17) https://doi.org/fq3g
DOI: 10.1101/2021.01.15.426881

33. **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets**
Martin Steinegger, Johannes Söding
Nature Biotechnology (2017-11) https://doi.org/ggctnw
DOI: 10.1038/nbt.3988 · PMID: 29035372

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

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

36. **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

37. **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

38. **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