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

This manuscript (permalink) was automatically generated from bluegenes/2022-paper-protein-kmers@34aa18a on March 11, 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]".

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. 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 [8,9]. Streaming containment estimates have been shown to facilitate genome discovery and correlate with Mash Distance, a proxy for Average Nucleotide Identity (ANI) [7,10].

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 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 [11,12]. ANI has emerged as the most widely-accepted method for estimating pairwise similarity of microbial genomes and delimiting species boundaries [13]. Recent research appears to confirm 95% ANI species threshold for prokaryotic species, although there is some debate as to the universality of this threshold [14,15,16]. AAI thresholds have been proposed for higher taxonomic ranks, <45%, 45-65% and 65-95% for family, genus, and species [12,17]. 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 [1,7,14]. Hera et. al (2022) [18] introduced accurate nucleotide sequence distance estimation from FracMinHash containment estimates, while accounting for the non-independence of mutated k-mers [19].

Here, we show that protein FracMinHash sketches can be used to find similarity across larger evolutionary distances than nucleotide k-mers. We demonstrate that FracMinHash Containment estimates can robustly estimate Average Amino Acid Identity across a range of evolutionary distances. We then use FracMinHash comparison methods to tackle two classification challenges: taxonomic classification of assembled genomes, and compositional analysis of metagenomes. Taken together, these results suggest that protein FracMinHash analyses can be used for metagenome sequence analysis, and may be particularly useful when sequenced organisms are more distantly related to organisms available in reference databases.

Results

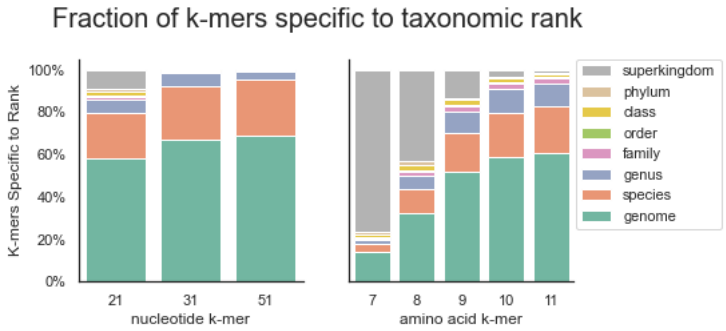
K-mer analysis methods enable similarity detection as low as a single shared k-mer between divergent genomes. As a result, exact matching of long nucleotide k-mers can be used for taxonomic classification and similarity detection between closely related genomes, including strain-level, species-level, 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 [20]. We begin by assessing the prevalance of nucleotide and amino acid k-mers of different k-mer lengths within assemblies 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 [21] 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. 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.

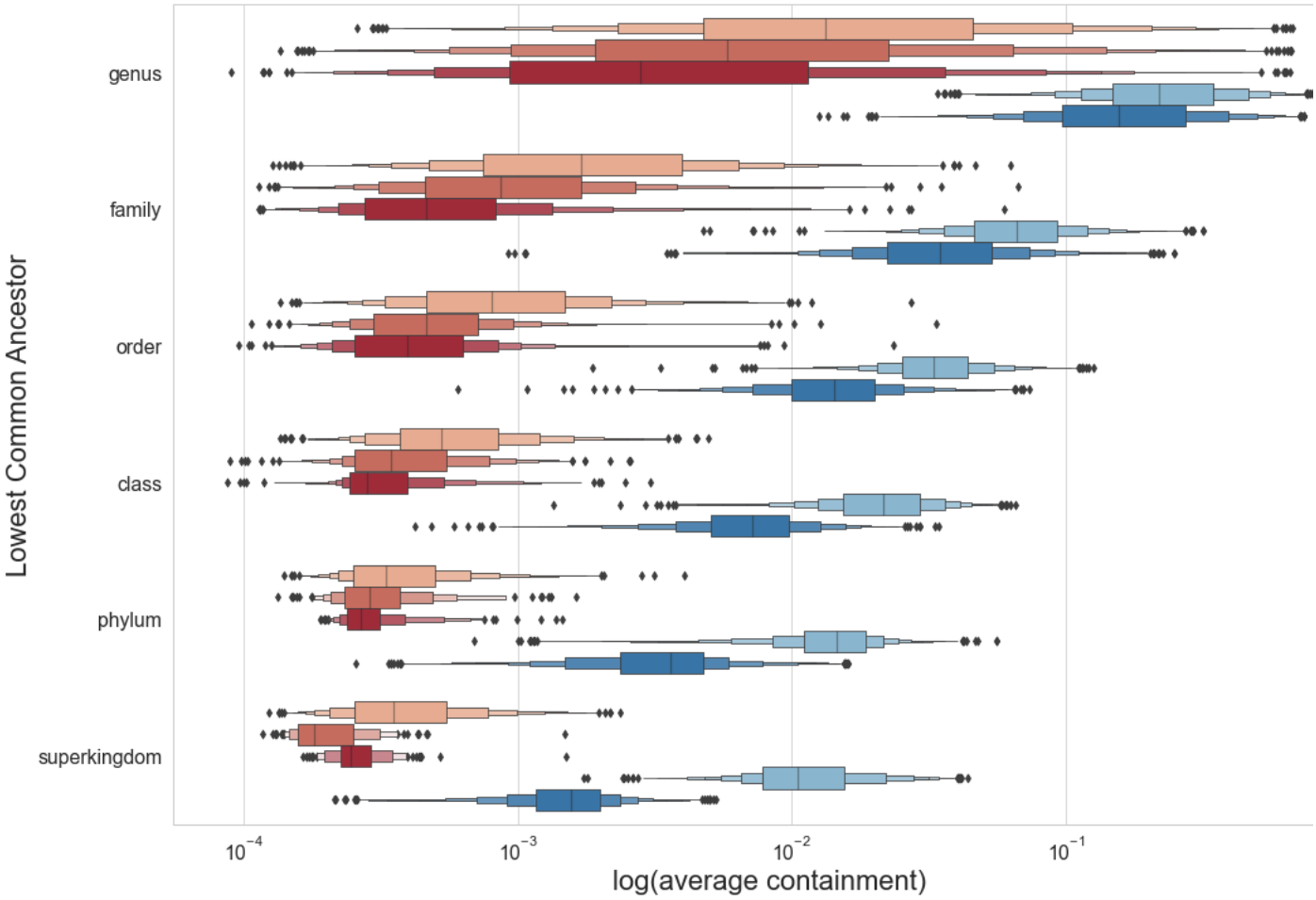
For each genus with at least two species clusters in GTDB, one representative genome was randomly selected as an “anchor” genome. Then, one additional genome was selected from the GTDB representative genomes matching the anchor’s taxonomy at each higher taxonomic rank. This “evolutionary path” consists 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.

For DNA comparisons, each genome was sketched from the available genome downloaded from genbank. For protein comparisons, we conducted two comparison workflows. In both workflows, all anchor genomes were sketched from available proteomes (either downloaded or generated via Prodigal, as above). For the first workflow, 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 FracMinHash sketches. As 6-frame translation introduces erroneous k-mers, we report only the containment estimate relative to the anchor proteome. We term this “anchor containment”, where the trusted genome is the “anchor” upon which we base the comparison. We conduct k-mer comparisons using 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). To verify results and estimate the impact of FracMinHash scaling, we also conduct comparisons using all available k-mers.

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

We begin by assessing k-mer statistics 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 protein k-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*).



To do: - Containment instead of Jaccard, so can easily compare protein sketches vs 6-frame translated sketches

Questions: - Also display jaccard for protein sketch comparisons? - Is there a better way to visualize this? - a pair of heatmaps?

Distance estimation from FracMinHash sketch comparisons

DNA k-mer Jaccard has previously been used to estimate of the Average Nucleotide identity between genomes [cite Ondov Mash, Koslicki k-mer paper, Criscuolo?]. Recently, equations have been developed for FracMinHash Containment that estimate ANI while accounting for the nonindependence of mutated k-mers [18]. These equations assume a simple mutational model and estimate distance solely based on k-mer statistics. Here we apply FracMinHash distance estimation to protein k-mer comparisons to obtain an alignment-free estimate of Amino Acid Identity [18]. As above, we utilize anchor containment for comparisons involving 6-frame translated sketches.

To assess whether k-mer methods can be used to approximate AAI, 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 the BLAST AAI values.

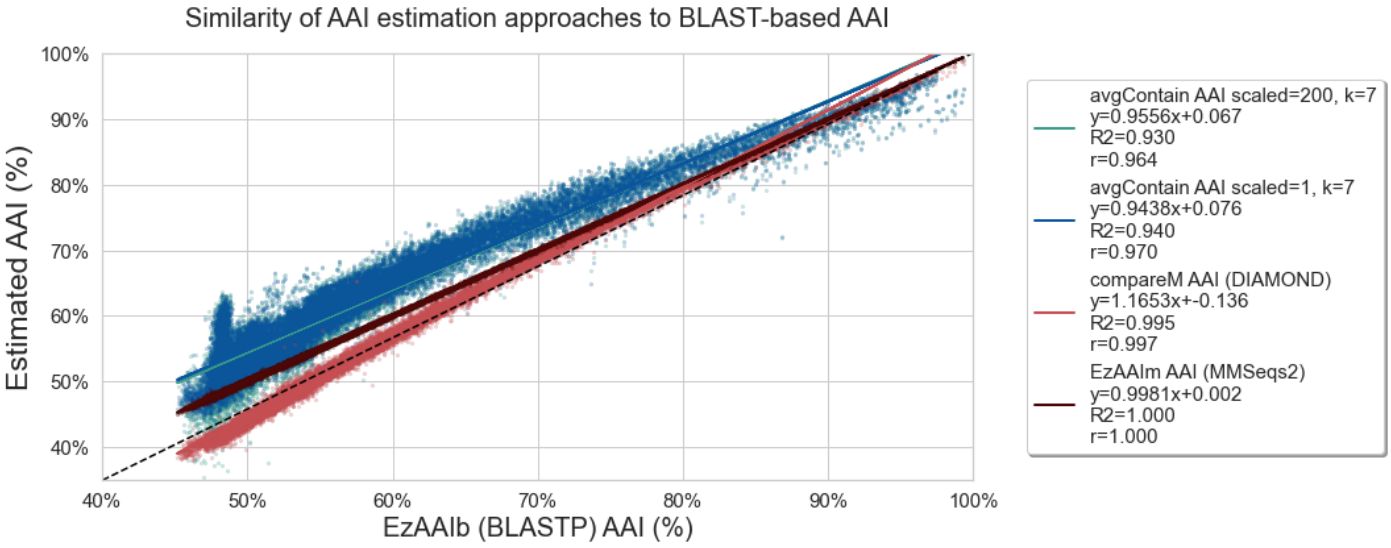


Figure 1: FracMinHash AAI vs BLASTP alignment based AAI

To do: - Rerun BLAST-AAI with default thresholds (current was lower identity thresholds) and re-plot - finish and add results from 6-frame translated sketches (anchor containment): same plot? or adjacent plot if needed.

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 [3], minimum set cover of nucleotide k-mers can be used to find the set of genomes that cover all known k-mers in your 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 [3]. 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 [22]. 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 [23].

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

just some disjoint notes, at the moment :)

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 [1,14].

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 FracMinHash 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 FracMinHash 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 FracMinhash sketches, which enables estimation of the Containment Index.

At k=21 (dna) and k=7 (protein), many k -mers are shared across taxonomic groups.

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 **

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

Containment enables comparison directly from DNA sequence

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.”

Maximum Containment

For both 6-frame translation and metagenome breakdown comparisons, the most informative containment value will be relative to the smaller set of k-mers (typically reference proteomes), rather than relative to all metagenome k-mers or all 6-frame translated genome or metagenome k-mers. As such, we have implemented “maximum containment,” a shorthand method to always select the greater of the two containment values for AAI estimation. 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 AAI with this method, so we suggest using containment relative to the more trusted sample if known, or using average containment AAI or jaccard AAI when comparing two genomes of approximately equal quality.

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

For the 6-frame translated comparisons, only ~1/6th of the k-mers generated by 6-frame translation will belong to true ORFs.

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). 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. However, these k-mers only impact the containment index in one direction (relative to the set with erroneous k-mers) 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.* 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). 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 only the FracMinHash containment estimate relative to reference proteomes, we can obtain accurate Amino Acid Identity estimates directly from DNA sequence.

evolpath containment values –driven by the k-mer frequency patterns observed in gtdb db...

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

Large-scale k-mer comparisons with FracMinHash sketches

FracMinHash sketching, as implemented in sourmash [3,25,26], 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 [8], containment has been shown to permit more accurate estimation of genomic distance when genomes or datasets differ in size [8,9,27,28]. 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 [3,18]. *As of `v4.x`, `sourmash` sketches store a Hyper-Log-Log estimate of dataset cardinality, calculated during sketching. Use of this estimate ensures that `sourmash` FracMinHash containment results will be unbiased estimates of the true containment, even for very small genomes (e.g. viruses) or large scaling factors (e.g. keep 1/1e6 k-mers).*

Sourmash v4.x supports sketching from either nucleotide or protein input sequence. 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. 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 200. 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.

In select cases, we also conduct 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).

Estimating Average Amino Acid Identity

discuss HLL / bias factor?

MinHash Sketch Jaccard has been shown to correlate well with ANI at high sequence identities (>=90% sequence identity) [1]. Recently, Blanca et al, 2021 [19] 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 [18] 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. If we instead generate amino acid k-mer sketches, we can apply the same equations to 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. The equation for sequence similarity estimation (ANI or AAI) from FracMinHash Containment is reproduced here for completeness (see [18] for details).

to do: ADD EQUATION

Sequence distance estimation (ANI, AAI) is implemented in `sourmash` as of `v4.4` ([**NTPW?**]: check version). The distance estimation equations can be found in the `distance_utils.py` file.

NTP: working here

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 [29] alignment. For both, we utilized EzAAI default parameters: 40% coverage threshold, 40% sequence identity threshold. CompareM v0.1.2 ([30]; 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 [21]. 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 [31/], **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 [32] 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 [32]. **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].

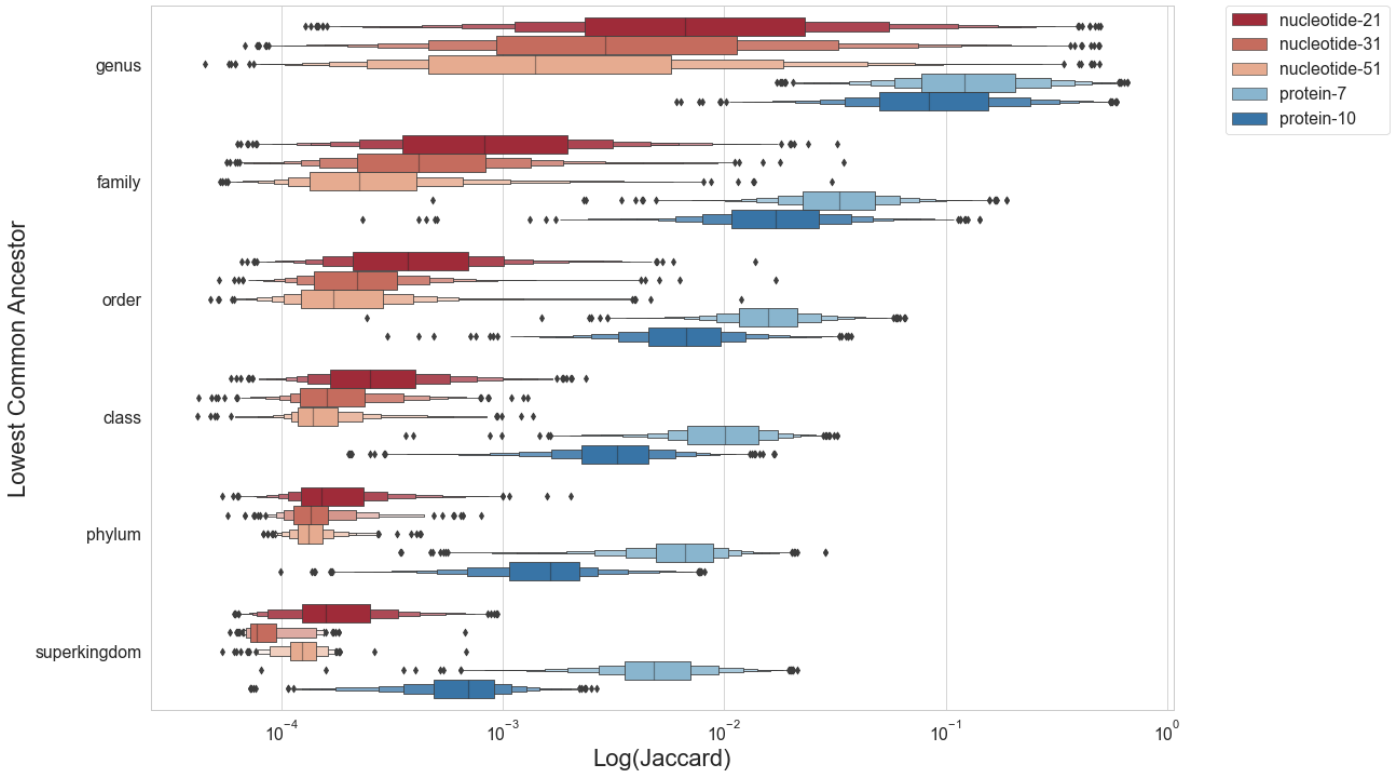
Taxonomic utilities are implemented in the `sourmash taxonomy` module.

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



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

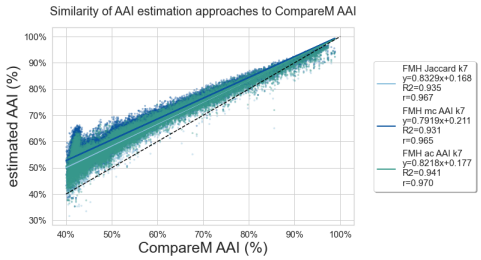


Figure 2: FracMinHash AAI vs CompareM Scaled 1

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. **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. **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

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 (2019-12) https://doi.org/ghtqmb
DOI: 10.1186/s13059-019-1841-x · PMID: 31690338 · PMCID: PMC6833257

8. **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

9. **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

10. **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

11. **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

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

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

20. **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

21. **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

22. **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

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. **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

26. **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

27. **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

28. **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

29. **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

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

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

32. **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/

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