Sensitivity of A Frequency-Based Alignment-Free Approach For Phylogeny Reconstruction.

Yana Hrytsenko (✉ yana_hrytsenko@uri.edu)

University of Rhode Island

Noah M. Daniels

University of Rhode Island

Rachel S. Schwartz

University of Rhode Island

Research Article

Keywords: alignment-free methods, phylogeny reconstruction, k-mers

Posted Date: February 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1174825/v2>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Phylogenies enrich our understanding of how genes, genomes, and species evolve. Traditionally, alignment-based methods are used to construct phylogenies from genetic sequence data; however, this process can be time-consuming when analyzing the large amounts of genomic data available today. Additionally, these analyses face challenges due to differences in genome structure, synteny, and the need to identify similarities in the face of repeated substitutions resulting in loss of phylogenetic information contained in the sequence. Alignment Free (AF) approaches using k-mers (short subsequences) can be an efficient alternative due to their indifference to positional rearrangements in a sequence. However, these approaches may be sensitive to k-mer length and the distance between samples.

Results: In this paper, we analyzed the sensitivity of an AF approach based on k-mer frequencies to these challenges using cosine and Euclidean distance metrics for both assembled genomes and unassembled sequencing reads. Quantification of the sensitivity of this AF approach for phylogeny reconstruction to branch length and k-mer length provides a better understanding of the necessary parameter ranges for accurate phylogeny reconstruction. Our results show that a frequency-based AF approach can result in accurate phylogeny reconstruction when using whole genomes, but not stochastically sequenced reads, so long as longer k-mers are used.

Conclusions: In this study, we have shown an AF approach for phylogeny reconstruction is robust in analyzing assembled genome data for a range of numbers of substitutions using longer k-mers. Using simulated reads randomly selected from the genome by the Illumina sequencer had a detrimental effect on phylogeny estimation. Additionally, filtering out infrequent k-mers improved the computational efficiency of the method while preserving the accuracy of the results thus suggesting the feasibility of using only a subset of data to improve computational efficiency in cases where large sets of genome-scale data are analyzed.

Background

Phylogenies are essential in evolutionary biology to better understand evolution from genes to traits. Currently, phylogenetic analysis typically begins with DNA or RNA sequencing, often using high throughput sequencing (HTS) technology. As HTS has become more efficient and affordable, the number of species for which sequence data are easily accessible has increased greatly, which has provided many opportunities for studies of new organisms¹. Commonly, phylogenies of organisms are estimated using these data in alignment-based methods². However, challenges in this work range from identifying orthologs to handling sequencing errors to accurately aligning loci³⁻⁵.

Genome characteristics such as gene conservation, point mutations, chromosomal mutations, structural changes, and copy number variation⁶, can pose challenges for a traditional alignment-based approach for phylogeny reconstruction^{7,8}. Specifically, phylogeny reconstruction is complicated in cases where

genomes do not share the same set of genes, or if sequences have very low similarity and thus are not alignable to each other⁹. Additionally, alignment-based phylogeny reconstruction is affected by rearrangements and inversion events, and different sequence lengths among different species¹⁰. A high number of insertions and deletions in the sequences also negatively influence the accuracy of the alignment¹¹.

Sequence alignment can also be complex and error-prone due to the erroneous positioning of gaps during reconstruction^{12,13,14}. Some aligners substantially underestimate insertion rates and overestimate deletion rates and thus produce alignments with inaccurate indels¹⁵. These errors affect the construction of the phylogenetic tree that is built from DNA or RNA sequences^{16,9}.

Alignment free methods

Alignment-free (AF) approaches are less susceptible to these challenges and have been proposed as an effective alternative for phylogeny reconstruction¹⁷. AF methods do not take into account the exact position of the nucleotides¹⁸ and thus are more resistant to the effects of structural variants such as inversions, translocations, chromosome fusions, chromosome fissions, and reciprocal translocations¹⁹. AF methods are especially successful in genome-based phylogenetic studies because they are capable of processing large numbers of genome-size sequences²⁰. Additionally, AF approaches do not require genome assembly and alignment steps, which are complex and computationally expensive^{21,22}. The scalability of AF methods allows processing large sets of genome data more efficiently than the traditional alignment-based algorithms.

AF methods typically analyze k-mers, which are short substrings of the genome sequence²³. AF methods can be classified broadly¹⁷ into methods based on k-mer frequencies^{24,25,26}, a number of word matches^{27,28}, lengths of common substring²⁹⁻³¹, and micro-alignments^{32-35,36}. However, various AF methods perform differently depending on the data type (mitochondrial vs. nuclear genomes), taxonomic groups (animal, bacterial, or plants), and data size (assembled genomes and unassembled sequencing reads of different coverage)¹⁷. No single method is successful across all types of data¹⁷.

Several factors must also be considered when working with AF approaches that can affect the accuracy of the phylogeny: k-mer length used in the analysis, number of substitutions per site in the genome, a metric used to calculate genome distances between sequences, and the use of assembled genomes vs. unassembled raw reads. Different AF methods use k-mers differently and thus there is no universal length of k-mer used in all AF methods³⁷. Instead, different optimal k-mer lengths have to be determined for the different methods¹⁷. Shorter k-mers are more likely to be present in a sequence (e.g. 1-mers); thus they are less informative in analyzing closely related genomes³⁸. Longer k-mers are more unique to particular species and are therefore more useful for similarity identification across species³⁹. Additionally, the problem of k-mer homoplasy should be considered when identifying k-mer length for accurate phylogeny reconstruction using AF methods. Shorter k-mers are more likely to have greater sensitivity to single

evolutionary events. However, identical k-mers could be obtained from the genome regions which are not identical in functional or evolutionary properties and thus are not homologous (k-mer homoplasy) ⁴⁰. Longer kmers are less susceptible to k-mer homoplasy. Thus, there is a tradeoff between sensitivity, which requires shorter k-mers, and reducing k-mer homoplasy, which requires longer k-mers ⁴⁰. Optimal k depends on the AF approach used ³⁷. The optimal lengths of k-mers differed for the “word count” AF approach and the “match length” approach ³⁷. K-mer length and genome size are correlated, which affects the accuracy of phylogeny reconstruction ⁴¹.

Evolutionary distance is another important parameter to consider in phylogeny reconstruction ^{42, 43, 44}. The presence of a high number of substitutions among samples causes loss of phylogenetic signal contained in the sequence (saturation) and is an important source of systematic bias ^{45, 46}. Multiple substitutions at the same site due to a high substitution rate can make individuals appear more similar than they actually are ⁴⁷. In extreme cases, the similarity between the sequences will depend entirely on the similarity in nucleotide frequencies that often does not reflect phylogenetic relationships ^{46, 48}. More specifically, the long branches present in the phylogenies may have acquired many substitutions that cause a historically random resolution of the internode. Saturation is known as a cause of erroneous outgroup placement in phylogeny reconstruction ⁴⁹. Additionally, genetic saturation introduces a long-branch attraction problem (LBA) ⁵⁰, where one lineage falsely appears to be closely related to another long-branched lineage due to the large amount of molecular change accumulated in the genomes rather than due to the actual relatedness by descent ^{51, 40}. The AF methods based on “number of word matches” and “match length” had greater accuracy of similarity identification for sequences with low divergence ⁴¹ whereas “word count” methods performed better at an intermediate level of sequence divergence ³⁷.

The metric used to calculate distances between genomes is also an important factor to consider in AF-based phylogeny reconstruction, as the distance metric affects both computational efficiency and sensitivity ⁵². Distance metrics include Minkowski (e.g., Euclidean) ⁵³, D2 (e.g., D_2^S) ⁴¹, mismatch (e.g., Jaccard) ⁴⁰, inner product (e.g, normalized vectors, cosine) ⁵⁴, and length difference ⁵⁵. Differences in the accuracy of phylogeny reconstruction depend on the metric used ³⁷. The accuracy of phylogenies based on D_2 statistics (based on the number of word matches) and kmacs (based on length of common substring) varied by the k-mer length and sequence divergence level ^{56–59, 41}.

Here, we use simulated genomes and HTS data to assess the sensitivity of AF methods based on k-mer frequencies to a range of numbers of substitutions per site in the genome using different k-mer lengths and two genome distance metrics for phylogeny reconstruction. We analyzed phylogenies with branch length (in numbers of substitutions per site) from 0.00001 to 0.1 based on differences among primates reported in the Tree of Life ^{60–62} to understand how sensitive these methods are to different sequence divergence levels. We show that AF approaches based on k-mer frequencies of common k-mers alone can generate accurate phylogenies from assembled genomes with high evolutionary distances, so long as k-

mers are sufficiently long. This work provides a better understanding of optimal parameter selection for future phylogenetic studies using frequency-based AF methods.

Results

Phylogenies from simulated assembled genome data

Phylogenies built from Euclidean distances had a median RF score of 0 for all substitution rates with k-mer lengths of 9 - 16 (Figure 3a), (means = 0.44 - 1.22; sd = 1.15 - 1.57). Median RF scores of 2 - 10 were produced for phylogenies generated from distances calculated using k-mer lengths 4 - 8 for the same dataset (means = 0.94 - 9.3, sd = 1.16 - 2.52) (Supplementary material Appendix 1, Figure A1).

The cosine distance metric produced median RF scores of 0 for all substitution rates with k-mer lengths 9 - 16 (Figure 3b), (mean = 0.08 - 0.28, sd = 0.39 - 0.85). Median RF scores of 0 - 10 were produced for phylogenies built from distances calculated using k-mer lengths 4 - 8 (mean = 0.64 - 9.3, sd = 1.13 - 2.55) (Supplementary material Appendix 1, Figure A2).

Phylogenies from simulated reads at uniform coverage

Comparison of the phylogenetic trees generated from cosine distances calculated between read sequences simulated by the “sliding window” method at uniform 10X coverage to the reference tree produced a median RF score of 0 for all rates, and k-mer lengths of 10 through 23 (means = 0.06 - 0.36, sd = 0.34 - 1.18). Median RF scores of 0 - 10 were produced for phylogenies generated using k-mer lengths 4 through 9 (means = 0.64 - 9.26, sd = 1.13 - 2.64) for all rates (Supplementary material Appendix 1, Figure A3).

Analysis of simulated reads with non-uniform coverage

Comparison of phylogenetic trees generated from cosine distances calculated between simulated Illumina read sequences at 10X non-uniform coverage (as would occur in empirical sequencing data) produced median RF scores of 8 through 10 for k-mer lengths of 10 through 23 (means = 8.68 - 10, sd = 0 - 0.99) for a range of substitutions per site, 10^{-5} to 10^{-1} (Supplementary material Appendix 1, Figure A4). Comparison of phylogenetic trees generated from Euclidean distances calculated for the same dataset produced the same median RF scores of 10 for the k-mer length of 10 through 23 (mean = 10, sd = 0) (Supplementary material Appendix 1, Figure A5).

To analyze the method’s performance at higher sequencing coverage we increased the coverage of the simulated Illumina reads to 100X for 10 samples and recalculated the distances. Comparison of phylogenetic trees generated from cosine distances produced median RF scores of 8 to 10 for k-mer lengths 10 through 23 (mean = 7.6 - 10, sd = 0 - 2.27) for all rates (Supplementary material Appendix 1, Figure A6).

Filtering of infrequent k-mers and efficiency improvements

Each sample's nearest neighbor remained the same after filtering out k-mer frequencies below 10^5 . We rebuilt the dendrograms and confirmed that the clustering of samples was not affected by filtering out frequencies below this threshold value while reducing computation time. Calculating cosine distances on filtered frequencies showed to be 25 times faster than calculating on unfiltered frequencies. Calculating cosine distance between a pair of 12-mer human genome profiles took 48 seconds using TensorFlow⁶³ with a GPU-based framework nVidia Tesla K40c with 128Gb of RAM; thus, distance calculations for 25 samples would take four hours. However, calculating distance on a pair of 12-mer profiles that do not include frequencies below 10^5 takes only two seconds.

Discussion

In this work, we show that a frequency-based AF approach is suitable for analyzing data with high numbers of substitutions per site. Our AF approach was able to recover the expected phylogeny from data containing a high number of substitutions using longer k-mers and cosine and Euclidean distances. These results suggest that an AF approach may be applied in place of traditional alignment-based methods where the problem of genome saturation and LBA in phylogeny reconstruction exists⁶⁴. Identification of k-mer length ranges that are appropriate for phylogeny reconstruction from a dataset with a range of numbers of substitutions per site may serve as a recommendation for the selection of suitable parameters for frequency-based AF methods.

In the analysis of the whole genome data our AF approach successfully reconstructed the phylogeny from the genome with realistic substitution rates, using a range of higher k-mer lengths using cosine and Euclidean distance metrics. However, the method produced less accurate phylogeny from sequences using shorter k-mers, even from more closely related sequences. Our results of successful phylogeny reconstruction with k-mer frequencies using longer k-mers are consistent with prior work suggesting longer k-mers are required to produce accurate phylogenies^{65,66,24}. For example, another frequency-based method (FFP) successfully reconstructed the phylogeny from assembled microbial genomes using longer k-mers and Jensen–Shannon Divergence, while a method based on the number of word matches (D_2) was shown to successfully estimate phylogeny from nucleotide and amino acid sequences, using D_2 statistics family and a higher k-mer length. Here we confirm these results with alternative distance metrics.

AF approaches using HTS reads are accurate with uniform but not stochastic coverage

The AF approach based on k-mer frequencies was able to capture the correct phylogeny from reads at a uniform level of coverage. The analysis of the reads at precisely 10X coverage, where each base was sequenced exactly 10 times, produced a correct phylogeny for all datasets using longer k-mers. However, the method was not able to recover the correct phylogeny for the same dataset using a lower range of k-

mer lengths. This result mirrors the results with whole-genome data but expands to allow for the truncation of k-mers due to short reads.

In reality, a uniform level of coverage is not plausible due to the randomized nature of the sequencing process. However, analysis of simulated read data generated by the Illumina simulator at non-uniform coverage did not produce the expected phylogeny for datasets even at high coverage levels. The random nature of sequencing likely introduces stochasticity to the k-mer frequencies and negatively affects the distances between sequences⁶⁷. While low sequencing coverage and sequencing errors are known to have a negative effect on phylogeny reconstruction using AF methods based on k-mer frequencies⁴⁰, even at high read coverage, the counts of k-mers are distorted due to the randomized nature of the sequencing algorithm. Cosine and Euclidean distances were sensitive to these distortions resulting in a phylogenetic error in our analysis.

Because the accuracy of the AF approach based on k-mer frequencies is highly dependent on the precision of the read coverage, using read data would depend on correction for sequencing error and coverage depth. Both cosine and Euclidean metrics take exact counts of k-mer frequencies present in a genome. Incongruence in the frequencies due to variable read coverage and sequencing error skews the distance between two vectors of k-mer counts.

Alternatively, other non-frequency-based methods may be considered. For instance, methods based on micro-alignments^{32,34} and methods based on the number of word matches^{27,28,40}. These methods employ several techniques to correct for the effects of coverage and sequencing error. The techniques range from estimating the amount of sequencing error and coverage and including the estimates into the distance formula²⁸ to correcting tip branches and bootstrapping of the obtained phylogenetic trees⁴⁰.

Choice of distance metric

The choice of metric for calculating genome distances affected the accuracy of phylogeny reconstruction. While cosine and Euclidean distance metrics produced distances from which an accurate phylogenetic tree was recovered from genomes when analyzing reads, the Euclidean distance metric produced a less accurate phylogeny. Prior work has also found that AF tools that employ different metrics performed differently depending on the data type - assembled genomes and unassembled raw reads¹⁷. For instance, AFKS⁵², an AF tool that employs the Jaccard index metric, was more accurate at phylogeny reconstruction from assembled genomes and less accurate for phylogeny reconstruction from unassembled raw reads using the same metric. On the other hand, mash²⁷, a method that also uses the Jaccard index metric, produced accurate results for both data types.

Filtering low-frequency k-mers

Filtering out infrequent k-mers improved the computational efficiency of the method. Earlier work of^{68,69} had discussed filtering out low k-mer frequencies, in particular, singletons due to them being a result of sequencer error. Our analyses showed that infrequent k-mers constitute the majority of k-mer content in

the genome. Distance calculation using only frequent k-mers showed no negative effect on the accuracy of the results and greatly reduced the computational time of the method. We believe the majority of the information comes from a small fraction of the most frequent k-mers. These results suggest the feasibility of using only a subset of data to improve computational efficiency in cases where large sets of genome-scale data are analyzed.

Conclusion

In this study, we have shown an AF approach for phylogeny reconstruction is robust in analyzing assembled genome data for a range of numbers of substitutions using longer k-mers. Similarly, an AF approach recovered an accurate phylogeny from reads with uniform coverage from data with a high number of substitutions using higher k-mer lengths. However, using simulated reads randomly selected from the genome by the Illumina sequencer had a detrimental effect on phylogeny estimation. Thus we recommend using longer k-mers to accurately reconstruct phylogeny from assembled genome data with high numbers of substitutions. However, phylogeny reconstruction using frequency-based AF methods with read data is not recommended. Additionally, k-mer frequencies below 10^5 may be filtered out to improve the efficiency of the distance calculation between sequences while preserving the accuracy of the results.

Methods

To find the parameters for which frequency-based AF methods produce accurate phylogeny from sequences of different levels of divergence using k-mers, we simulated genomes for multiple related individuals and estimated phylogenies using AF methods based on k-mer frequencies. We also simulated and estimated phylogenies from simulated HTS read data to compare the accuracy of the phylogenetic trees between two types of sequence data.

1. Trees

We initially generated a balanced tree for simulations with eight individuals and equal branch lengths (Figure 1). We varied these branch lengths for simulations from 10^{-5} to 10^{-1} substitutions per site incremented by orders of magnitude. These datasets provide a testable range of sequence divergence to assess the sensitivity of alignment-free methods in recovering the topology.

2. Assembled genome data simulation.

We simulated genomes using a subset (for efficiency purposes) of the reference human genome, chromosome 19⁷⁰, as an ancestral sequence. A custom script was used to generate 5 sets of data (one per order of magnitude) with 8 genomes per set (Figure 1). We repeated the analysis 100 times per set to examine statistical significance.

3. Raw-read data simulation.

We simulated HTS read sequences on each of the simulated genomes using two methods - the Illumina read simulator in BBMap ⁷¹ and a custom “sliding window” method to generate reads containing identical coverage of every location in the genome. This allowed us to compare the sensitivity of the approach to stochastic variation in coverage. Simulated Illumina read sequences were generated for 10X and 100X coverages to compare how efficiently the method works at the two levels of coverage with no sequencing errors. We used a custom script to generate read sequences at 10X coverage using the “sliding window” method where each read was generated by sliding a window of 150 bp along the sequence with 15bp overlap. We repeated the analysis of reads at 10X coverage generated by the custom script and the reads generated at 10X coverage by the Illumina read simulator 100 times per set to examine statistical significance. Analysis of the Illumina read dataset at 100X coverage was only repeated 10 times due to high memory usage.

4. Distance calculation between simulated assembled genomes based on k-mer frequencies.

To calculate distances between genomes we first generated k-mer profiles using Jellyfish ⁷², a tool for fast, memory-efficient counting of k-mers in the DNA sequence, for k values in the range 4 through 16 for each sequence in the sets of the simulated data. Each k-mer profile contains the frequency of canonical k-mers (k-mer or its reverse complement, whichever comes first lexicographically) present in the genome. We performed genome distance calculations using two metrics, Euclidean and cosine distances. Both metrics use the exact count of k-mers present in the sequences and thus represent the extent of similarity that sequences share with higher sensitivity. The Euclidean distance is based on the inclusion of all of the k-mer frequencies through the union operation and therefore takes into account the complete k-mer profile of a sequence. Conversely, cosine distance uses the set intersection of k-mer profiles and uses less information about the sequence, but has the advantage of not being sensitive to the number of reads. Specifically, cosine distance looks at the angle between two vectors in high-dimensional space, while ignoring their length. Greater read coverage changes the magnitude of a vector but not its distance to any other vector.

We calculated the set intersection (for cosine distance) and the set union (for Euclidean distance) for each pair of k-mer profiles using custom scripts for distance calculations. Computational resources required to calculate distances on such a large set of genome data, based on the frequencies of k-mers, grow exponentially as the length of k increases. Thus, to accommodate multiple datasets with a large number of genomic-scale sequences, we computed distances between each pair of k-mer profiles using TensorFlow ⁶³ and multiple GPUs.

5. Distance calculation between simulated read sequences based on the k-mer frequencies.

To calculate distances between read sequences we built k-mer profiles using the same approach as in the genome analysis for k-mer length 4 through 16 (for reads of uniform coverage), and 10 through 23 (for reads of non-uniform coverage) for each sequence. Our preliminary results showed lower accuracy in phylogeny estimation from reads of non-uniform coverage using shorter k-mers thus we increased the k-mer length range for this dataset. To emulate the real-life analysis process we filtered out singletons, which are usually considered as possible errors produced by the sequencer in the real dataset ⁶⁸. We analyzed longer k-mers to overcome possible k-mer homoplasy in read analysis ⁴⁰. We calculated the set intersection and set union of k-mers. For reads of non-uniform 10X coverage, we calculated cosine and Euclidean distances. We saw lower accuracy in phylogeny reconstruction using Euclidean distance metric in raw reads and thus used only cosine distances for analyzing reads of uniform 10X coverage and reads of non-uniform 100X coverage. We used TensorFlow ⁶³ and multiple GPUs to accommodate large datasets of k-mer profiles which grow exponentially as k-mer length increases.

6. Phylogeny estimation

To analyze the accuracy of our approach we examined the grouping of samples based on the distances among them. We built phylogenetic trees using the Neighbour-Joining algorithm for each dataset ⁷³. We measured the similarity between the generated tree and the reference tree by calculating the Robinson-Foulds (RF) distance ⁷⁴.

7. Adding efficiency to calculations.

We hypothesized that the majority of k-mers are infrequent. Approximately 99% of 12-mer frequencies in a human genome at 30X coverage are $< 10^5$ (Figure 2) and will not contribute to the dot product in the cosine distance calculation. Thus, removing these k-mers from the calculation should reduce computation time without impacting the cosine distances. To better understand the distribution of k-mer frequencies, we built histograms. Frequencies were normalized by the level of genome coverage. In the analysis of complete genomes, we filtered out k-mers with frequencies below thresholds ranging from 10^0 - 10^7 at different orders of magnitude, recalculated distances for each set of filtered k-mers, and examined the closest sample to each sample. This ensured an appropriate comparison between the tree generated by analysis of the full dataset and the tree from the filtered data.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All the code used in this work is available in the GitHub repository

https://github.com/YanaHrytsenko/kmer_phylogeny_sensitivity and can be used freely for reproducibility purposes.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by a grant to R. Schwartz from the National Science Foundation (DBI-1942273) and by the USDA National Institute of Food and Agriculture, Hatch project accession no. 1017848.

Authors' contributions

YH implemented the code, performed all the experiments and wrote the manuscript. ND and RS guided the study. All authors contributed to conceiving the study, provided contributions to the written manuscript, and read and approved the final manuscript.

Acknowledgments

We thank members of the Schwartz lab for valuable comments and suggestions. The High-Performance Research Computing (HPC) facility at the University of Rhode Island is acknowledged for providing the computational resources for the analyses in this manuscript. Additionally, we thank Kevin Bryan, the HPC systems manager, for invaluable help with software installation on the cluster and GPU operations management.

References

1. Kulski, J. K. Next-Generation Sequencing — An Overview of the History, Tools, and 'Omic' Applications. *Next Generation Sequencing - Advances, Applications and Challenges* (2016) doi:10.5772/61964.
2. Phillips, A., Janies, D. & Wheeler, W. Multiple Sequence Alignment in Phylogenetic Analysis. *Molecular Phylogenetics and Evolution* vol. 16 317–330 (2000).
3. Tekaia, F. Inferring Orthologs: Open Questions and Perspectives. *Genomics Insights* **9**, 17–28 (2016).
4. Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 10950 (2018).
5. Armstrong, J., Fiddes, I. T., Diekhans, M. & Paten, B. Whole-Genome Alignment and Comparative Annotation. *Annu Rev Anim Biosci* **7**, 41–64 (2019).

6. Loewe, L. & Hill, W. G. The population genetics of mutations: good, bad and indifferent. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 1153–1167 (2010).
7. Som, A. Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.* **16**, 536–548 (2015).
8. Ogden, T. H. & Rosenberg, M. S. Multiple sequence alignment accuracy and phylogenetic inference. *Syst. Biol.* **55**, 314–328 (2006).
9. Dessimoz, C. & Gil, M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol.* **11**, R37 (2010).
10. Comin, M. & Verzotto, D. Alignment-free phylogeny of whole genomes using underlying subwords. *Algorithms Mol. Biol.* **7**, 34 (2012).
11. Nuin, P. A. S., Wang, Z. & Tillier, E. R. M. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* **7**, 471 (2006).
12. Lees, J. A. *et al.* Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Res* **3**, 33 (2018).
13. Ezawa, K. Characterization of multiple sequence alignment errors using complete-likelihood score and position-shift map. *BMC Bioinformatics* **17**, 133 (2016).
14. Ashkenazy, H., Cohen, O., Pupko, T. & Huchon, D. Indel reliability in indel-based phylogenetic inference. *Genome Biol. Evol.* **6**, 3199–3209 (2014).
15. Westesson, O., Lunter, G., Paten, B. & Holmes, I. Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. *PLoS One* **7**, e34572 (2012).
16. Dwivedi, B. & Gadagkar, S. R. Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evol. Biol.* **9**, 211 (2009).
17. Zielezinski, A. *et al.* Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **20**, 144 (2019).
18. Ren, J. *et al.* Alignment-Free Sequence Analysis and Applications. *Annu Rev Biomed Data Sci* **1**, 93–114 (2018).
19. Vinga, S. & Almeida, J. Alignment-free sequence comparison-a review. *Bioinformatics* **19**, 513–523 (2003).
20. Chan, C. X. & Ragan, M. A. Next-generation phylogenomics. *Biol. Direct* **8**, 3 (2013).

21. Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. Comparative genome assembly. *Brief. Bioinform.* **5**, 237–248 (2004).
22. Dewey, C. N. Whole-Genome Alignment. *Methods Mol. Biol.* **1910**, 121–147 (2019).
23. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* **29**, 987–991 (2011).
24. Jun, S.-R., Sims, G. E., Wu, G. A. & Kim, S.-H. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 133–138 (2010).
25. Sims, G. E. & Kim, S.-H. Whole-genome phylogeny of *Escherichia coli*/Shigella group by feature frequency profiles (FFPs). *Proc. Natl. Acad. Sci. U. S. A.* **108**, 8329–8334 (2011).
26. Choi, J. & Kim, S.-H. A genome Tree of Life for the Fungi kingdom. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9391–9396 (2017).
27. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
28. Sarmashghi, S., Bohmann, K., P Gilbert, M. T., Bafna, V. & Mirarab, S. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biol.* **20**, 34 (2019).
29. Leimeister, C.-A. & Morgenstern, B. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics* **30**, 2000–2008 (2014).
30. Haubold, B., Pfaffelhuber, P., Domazet-Loso, M. & Wiehe, T. Estimating mutation distances from unaligned genomes. *J. Comput. Biol.* **16**, 1487–1500 (2009).
31. Haubold, B. Alignment-free phylogenetics and population genetics. *Brief. Bioinform.* **15**, 407–418 (2014).
32. Haubold, B., Klötzl, F. & Pfaffelhuber, P. andi: fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics* **31**, 1169–1175 (2015).
33. Klötzl, F. & Haubold, B. Phylonium: fast estimation of evolutionary distances from large samples of similar genomes. *Bioinformatics* **36**, 2040–2046 (2020).
34. Yi, H. & Jin, L. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Res.* **41**, e75 (2013).
35. Dencker, T. *et al.* 'Multi-SpaM': a maximum-likelihood approach to phylogeny reconstruction using multiple spaced-word matches and quartet trees. *NAR Genom Bioinform* **2**, lqz013 (2020).

36. Leimeister, C.-A., Sohrabi-Jahromi, S. & Morgenstern, B. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics* **33**, 971–979 (2017).
37. Bernard, G., Chan, C. X. & Ragan, M. A. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Sci. Rep.* **6**, 28970 (2016).
38. Bernard, G. *et al.* Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Brief. Bioinform.* **20**, 426–435 (2019).
39. Greenfield, P. & Roehm, U. Answering biological questions by querying k-mer databases. *Concurrency and Computation: Practice and Experience* vol. 25 497–509 (2013).
40. Fan, H., Ives, A. R., Surget-Groba, Y. & Cannon, C. H. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* **16**, 522 (2015).
41. Chan, C. X., Bernard, G., Poirion, O., Hogan, J. M. & Ragan, M. A. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci. Rep.* **4**, 6504 (2014).
42. Varki, A. & Altheide, T. K. Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res.* **15**, 1746–1758 (2005).
43. Wolfe, K. H., Li, W. H. & Sharp, P. M. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 9054–9058 (1987).
44. Bielawski, J. P., Dunn, K. A. & Yang, Z. Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156**, 1299–1308 (2000).
45. Xia, X., Xie, Z., Salemi, M., Chen, L. & Wang, Y. An index of substitution saturation and its application. *Mol. Phylogenet. Evol.* **26**, 1–7 (2003).
46. Dávalos, L. M. & Perkins, S. L. Saturation and base composition bias explain phylogenomic conflict in *Plasmodium*. *Genomics* **91**, 433–442 (2008).
47. Philippe, H. *et al.* Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* **9**, e1000602 (2011).
48. Xia, X. *Data Analysis in Molecular Biology and Evolution*. (Springer Science & Business Media, 2007).
49. Peer, Y. V. de, Van de Peer, Y., Frickey, T., Taylor, J. S. & Meyer, A. Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. *Gene* vol. 295 205–211 (2002).
50. Bawono, P. & Heringa, J. Phylogenetic Analyses. *Comprehensive Biomedical Physics* 93–110 (2014) doi:10.1016/b978-0-444-53632-7.01108-4.

51. Bergsten, J. A review of long-branch attraction. *Cladistics* vol. 21 163–193 (2005).
52. Luczak, B. B., James, B. T. & Girgis, H. Z. A survey and evaluations of histogram-based statistics in alignment-free sequence comparison. *Brief. Bioinform.* **20**, 1222–1237 (2019).
53. Schwende, I. & Pham, T. D. Pattern recognition and probabilistic measures in alignment-free sequence analysis. *Brief. Bioinform.* **15**, 354–368 (2014).
54. Murray, K. D., Webers, C., Ong, C. S., Borevitz, J. & Warthmann, N. kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLOS Computational Biology* vol. 13 e1005727 (2017).
55. Ulitsky, I., Burstein, D., Tuller, T. & Chor, B. The average common substring approach to phylogenomic reconstruction. *J. Comput. Biol.* **13**, 336–350 (2006).
56. Song, K. *et al.* New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief. Bioinform.* **15**, 343–353 (2014).
57. Torney, D. C., Burks, C., Davison, D. & Sirotkin, K. M. Computation of d 2: A Measure of Sequence Dissimilarity. *Computers and DNA* 109–125 (2018) doi:10.4324/9780429501463-11.
58. Wan, L., Reinert, G., Sun, F. & Waterman, M. S. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *J. Comput. Biol.* **17**, 1467–1490 (2010).
59. Reinert, G., Chew, D., Sun, F. & Waterman, M. S. Alignment-free sequence comparison (I): statistics and power. *J. Comput. Biol.* **16**, 1615–1634 (2009).
60. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
61. Feng, S. *et al.* Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252–257 (2020).
62. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
63. Martín Abadi *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015).
64. Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G. & Philippe, H. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.* **54**, 743–757 (2005).
65. Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2677–2682 (2009).

66. Wang, Y., Liu, L., Chen, L., Chen, T. & Sun, F. Comparison of metatranscriptomic samples based on k-tuple frequencies. *PLoS One* **9**, e84348 (2014).
67. Song, K. *et al.* Alignment-free sequence comparison based on next-generation sequencing reads. *J. Comput. Biol.* **20**, 64–79 (2013).
68. Melsted, P. & Pritchard, J. K. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* **12**, 333 (2011).
69. Roy, R. S., Bhattacharya, D. & Schliep, A. Turtle: Identifying frequent k -mers with cache-efficient algorithms. *Bioinformatics* vol. 30 1950–1957 (2014).
70. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
71. Bushnell, B. *BBMap: a fast, accurate, splice-aware aligner*. <https://www.osti.gov/biblio/1241166> (2014).
72. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
73. Pavlopoulos, G. A., Soldatos, T. G., Barbosa-Silva, A. & Schneider, R. A reference guide for tree analysis and visualization. *BioData Min.* **3**, 1 (2010).
74. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences* vol. 53 131–147 (1981).

Figures

Figure 1

A balanced reference tree of 8 samples.

Figure 2

Histogram of 12-mer frequencies per single human genome at 30X coverage

Figure 3

*Example violin plots of the distribution of Robinson-Foulds scores calculated for 100 trees generated from simulated genomic data when compared to the reference tree for a range of substitution per site of 10^{-5} to 10^{-1} incremented by orders of magnitude. The median RF score was 0 for all values. Each sample is an assembled genome. **a)** Trees were generated from Euclidean distances calculated using the k-mer length of 12. K-mer lengths 10 -16 produced comparable plots. Each sample is an assembled genome. **b)** Trees were generated from cosine distances calculated using the k-mer length of 9. K-mer lengths 10 -16 produced comparable plots.*

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SuppMatFigA1A6.pdf](#)