Original Article

Standardized Phylogenetic Evolutionary Analysis based on an Alignment Free Sequence Comparison Technique

Kshatrapal Singh[1], Ashish Kumar[2], Manoj Kumar Gupta[3]
[1]Department of CSE, Krishna Engineering College, Ghaziabad, India
[2]Department of CSE, I.T.S Engineering College, Greater Noida, India
[3]Department of CSE, Shri Mata Vaishno Devi University, Katra (J&K), India
Email: mekpsingh1@gmail.com
DOI: 10.47750/pnr.2022.13.S01.74

Abstract

Phylogenetic analysis explains how species evolved throughout time. Multiple sequence alignments of proteins or genomes can be used to infer a phylogenetic tree. Understanding evolution requires phylogenetics, or reconstructing organisms' evolutionary relationships. The alignment of complete gene sequences of higher eukaryotes, as well as the construction of phylogenetic trees on the basis of these alignments, is a computationally costly and ambitious undertaking. To get around these restrictions, we compared the genomes of the Brassicales clade using alignment-free technique. A Chaos Representation (CR) can be evaluated for each nucleotide sequence, which depicts each nucleotide as a point in a square specified by the four nucleotides as nodes. As a result, every CR is a distinct fingerprint of the basic sequence. Each grid square in the CRs represents the presence of oligonucleotides of a given size in the sequence if the CRs are categorized by grid lines (Frequency Chaos Representation, FCR). To build phylogenetic trees of Brassicales species, we used distance metrics between FCRs.

Keywords: Phylogenomics, Chaos Representation, Genome-scale data, DNA sequencing, NJ trees.

1. INTRODUCTION

The term "phylogenomics" was initially used in the context of genome function prediction for genome-scale data, and then in the ambience of phylogenetic conclusion shortly after. The advancements in DNA analyzing methodology over the last two decades have given rise to the discipline of phylogenomics [1-3]. It incorporates multiple fields of research at the intersection of molecular as well as evolutionary biology, with two main objectives: (i) inferring phylogenetic relationships among taxa and gaining insights into molecular advancement mechanisms; and (ii) using multi-species phylogenetic analysis to conclude putative functions for protein or DNA sequences.

This unplanned 'noise' effects the inference of backbone nodes, possibly directing to weakly carried phylogenetic trees, due to the fewer number of phylogenetically informative characteristics accessible in one or a few genomes. Using significantly higher volumes of sequence data, this challenge can be solved satisfactorily. Traditional sequencing datasets are orders of magnitude bigger than modern phylogenomic techniques, which use hundreds to thousands of sites from across the gene [4-5]. As a result, the scale of these datasets minimizes the impact of stochastic error as well as data availability as a binding factor.

High-throughput sequence method (also known as next-generation sequence method) has produced massive amounts of genome-scale data. Next-generation sequencing methods depart from the Sanger approach for which they permit for largely parallel DNA sequencing, allowing for exceptionally high throughput from several samples at the same time at a far lower cost. Large number of DNA nucleotides could be analyzed in parallel, giving orders of enormity additional value and reducing the requirement for fragment-cloning techniques employed in Sanger sequencing [6-8]. Recent advancements in NGS technology, as well as the rapid development of bioinformatics tools, have made it possible for research groups of some amount to create massive quantity of genomic sequence for organisms of consideration.

Phylogenetic studies are important in most domains of biology, but their reproducibility is often poor, with a prediction of 60% published phylogenetic comparison being 'lost to science' as a lack of data and techniques. Because the used analytical software, software versions, software settings, as well as OS versions can be difficult or not possible to unearth or revive, produced phylogenetic research can be tough or impossible to replicate or expand [9-12]. The tracking of the input and modification of information needed to get an output, known as data provenance, is a critical aspect of reproducibility. A large number of

recommendations as well as guidelines have been offered to encourage best practices in phylogenomics and bioinformatics in terms of reproducibility and data management, and several tools for assuring provenance of both data and methodologies have been evolved [13-16].

2. Proposed Methodology

With Octave, the algorithm for calculating CRs and FCRs was implemented. CR coordinates were created as lists in original text and drawn in the Scalable Vector Graphics programme for graphical presentations (SVG). FCRs were calculated for each k in 1,..., 8, based on CR position values. Calculations for distance were also implemented.

The following algorithm was used to construct chaos representations of nucleotide sequences. Each vertex of a 1 × 1 square is identified with a nucleotide. We situate A to the upper left, T to the upper right, C to the lower left, as well as G to the lower right node, in agreement with other results. The geometric centre of the square at position is the beginning point (0.5, 0.5). After that, the nucleotide sequences are shown in order. A value is drawn on half the distance among the initial value (0.5, 0.5) as well as the node in agreement to this nucleotide for the first nucleotide. Following that, for each subsequent nucleotide, a value is drawn at the halfway point among the previously drawn value and the nucleotide's node.

The following equations can be used to express the algorithm:

$$CR0=(0.5,0.5) \quad (1)$$

$$CRi=\begin{cases} CR_{i-1}+0.5\,.\,\left(CR_{i-1}+(0,0)\right) if\ Seq_i='A' \\ CR_{i-1}+0.5\,.\,\left(CR_{i-1}+(1,0)\right) if\ Seq_i='T' \\ CR_{i-1}+0.5\,.\,\left(CR_{i-1}+(0,1)\right) if\ Seq_i='C' \\ CR_{i-1}+0.5\,.\,\left(CR_{i-1}+(1,1)\right) if\ Seq_i='G' \end{cases} \quad (2)$$

Each sequence generates a different plot. Every one sub-square of the plot repeats the general pattern of points. Furthermore, all plot on the basis of a subsequence of the entire sequence looks the same. As a result, similar sequences produce similar CR charts. The FCR is calculated by adding the frequencies of points within every sub-square. As a result, each FCR indicates the total number of oligonucleotides in the sequence. The binary square categorized to a 4 × 4 grid for dinucleotides (k = 2), an 8 × 8 grid for trinucleotides (k = 3), and a 2k × 2k grid in general.

If the lengths of the nucleotide sequences vary, the resulting FCRs will have different overall frequencies. Each FCR was standardized to overcome the sequence length bias. If the FCR is shown as a 2k × 2k matrix, the X = (x) 2k× 2k matrix is translated into an expected FCR as follows:

$$\bar{X} = \frac{4^k}{\sum_{i=1}^{k}\sum_{j=1}^{k} x_{i,j}}\ X \quad (3)$$

We evaluated pair-wise distances among the FCRs to identify the evolutionary relationship between the examined species [17-20]. In general, any distance that can be applied to matrices can be employed. The Hamming distance, the Euclidean distance, the Image distance explained in, and the Pearson distance have all been described as distances for comparing FCRs. The Euclidean distance was chosen since it fared the leading in a contrast of several distance algorithms. The Pythagorean equation can be used to compute the Euclidean distance among 2 points in 2-dimensional space, which is given as the size of the line segment among these 2 locations. The distance between two FCRs can be calculated using this technique. The Euclidean distance among 2 standardized FCRs X = (x) 2k× 2k and Y = (y) 2k× 2k is given by following equation:

$$ED\ (\bar{X}, \bar{Y}) = \frac{\sqrt{2^k}}{4^k}\ \sqrt{\sum_{i=1}^{2^k}\sum_{j=1}^{2^k}(x_{i,j}-y_{i,j})^2} \quad (4)$$

3. Generating Phylogenetic Trees

Pair-wise distance patterns were constructed for every k using the Euclidean distance approach while specified in Equation 4 to give the phylogenetic trees. The distance patterns were processed using the Mega 5.2 software with Neighbor Joining (NJ) and Fitch Margoliash methods. Random sampling with replacement was used to produce 200 datasets for each FCR. For each k, 200 phylogenetic trees were reconstructed using these re-sampled FCRs. The consense tool was used to summarise each dataset's trees into consensus trees. The branch lengths were computed using the Fitch–Margoliash technique once the topographies of the consensus trees were specified [21-24]. For the NJ trees, a bootstrapped tree with the identical topography

like the consensus tree was chosen, and the bootstrap rates projected onto it. The bootstrap amount indicates the proportion of internal branches that have the identical partition like the consensus tree [25-29].

4. Results and Discussion

Table 1 shows information about various genome files obtained from NCBI database, every in FASTA format. The nuclear and mitochondrial genomes, as well as any contamination from other species' DNA, are included in the whole genome assemblies made accessible by sequencing centers. The subscriptions of mitochondrial genomes as well as tainting DNA to the FCRs, however, are minimal given the lengths of the all over genome datasets. As a result, the FCRs of all over genome data can be examined same in order to FCRs of nuclear genome.

Table 1: The species that were employed in our study

Species	**Accession Numbers**
Arabidopsis lyrata	GL348613–GL348407
Arabidopsis thaliana	NC_003074–NC_003076, NC_001284
Brassica rapa	AENI01000001–AENI01051658, NC_016325
Capsella rubella	134834574
Carica papaya	DS981420–DS984526
Citrus clementina	295550349
Citrus sinensis	319231331
Eucalyptus camaldulensis	DF096775–DF127446
Eucalyptus grandis	691297852
Eutrema halophilum	243117811
Eutrema parvulum	CM001587–CM001293
Gossypium raimondii	763818933
Theobroma cacao	FR720257–FR725448
Vitis vinifera	FN597025–FN597047, NC_012119
Brassica oleracea	NC_008285
Limnanthes alba	8582959
Raphanus raphanistrum	104536170
Raphanus sativus	97973538
Brassica carinata	NC_016320
Brassica juncea	NC_016223
Lotus japonicus	NC_016783
Millettia pinnata	NC_016772
Ricinus communis	NC_015131

The Euclidean distance method was combined with the NJ or Fitch Margoliash tree regeneration methods to generate phylogenetic trees on the basis of all over genome, mitochondrial genome, as well as EST data. We shown trees of FCRs constructed with k = 3 (trinucleotides, 64 data values) and k = 8 (oligonucleotides, 64 data values) to reveal the effect of oligonucleotide lengths (octanucleotides, 65,536 data values).

We looked for nearly connected plant species like that all over genome assemblies were present with the aim to examine the phylogenetic organizing of B. rapa in an all-over genome environment. The genomes of 13 distinct Malvidae species have been compared and analyzed: A. lyrata, A. thaliana, B. rapa subsp. pekinensis, Capsella rubella, Carica papaya, C. clementina, C. sinensis (sweet orange), Eucalyptus camaldulensis (Murray red gum), Eucalyptus grandis (Flooded gum), Eutrem (cacao plant). Furthermore, the Vitis vinifera (grape vine) genome was selected as an out category for rooting the plants. There isn't a species tree that includes all of these critters. We generated trees of these species on the basis of alignment for the actin CAP, Arp2, and Arp3 proteins' concatenated protein sequences for comparison (Figures 1). The main difference between the NJ and ML trees is whether the two Citrus species are grouped as a separate clade (NJ, Fig 1A) or like a sister clade as regards Malvales (ML, Fig 1B).

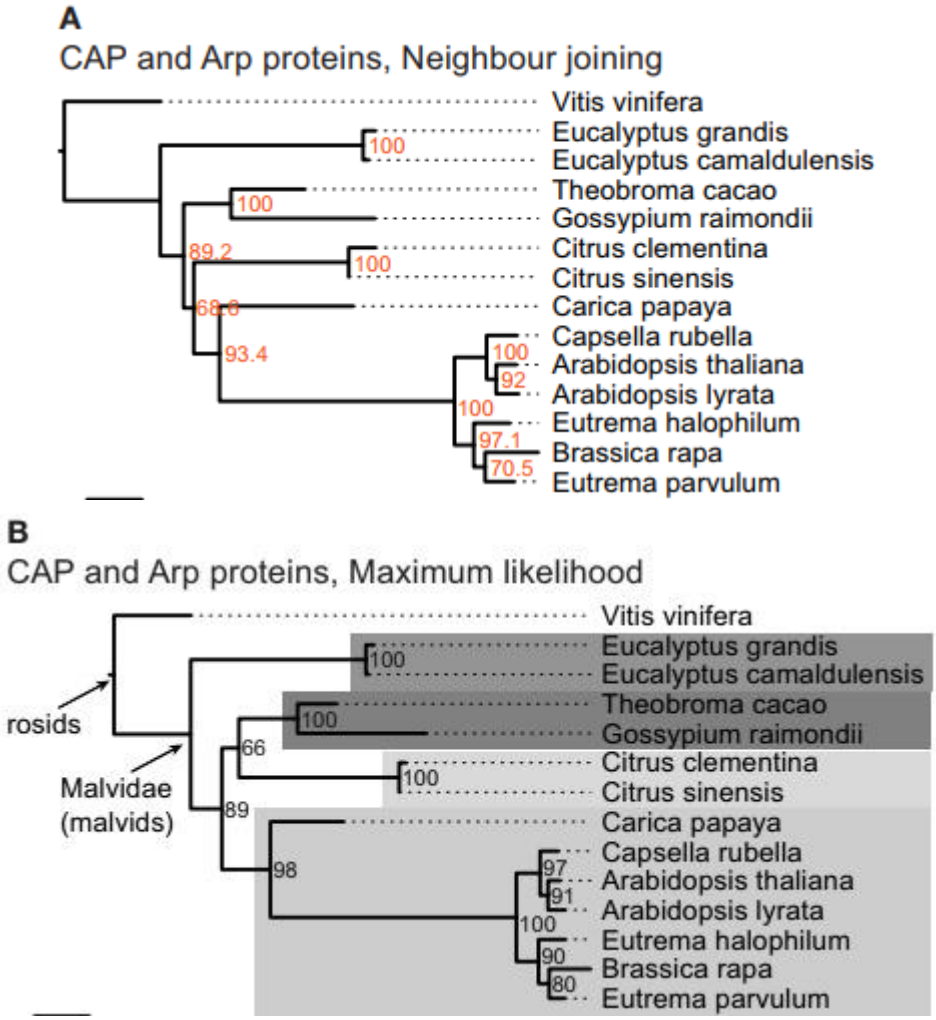

Figure 1: The trees in (A,B) on the basis of multiple sequence analysis of the CAP and Arp2/3 protein sequences

All trees agree that the Malvales, Sapindales, as well as Brassicales belong to single clade, also that the Myrtales belong to sister clade, with C. papaya representing the maximum contrasting as far as examined Brassicales species as well as C. rubella representing the nearest Arabidopsis species relation.

The phylogenetic trees of the FCRs that generate vary depending on the data and methodologies used (Figures 2 and 3). In (Figure 2), we reconstructed a tree using the Euclidean distance as well as the Fitch– Margoliash method, but with a resolution of k = 3, and a tree using a new approach for tree reconstruction, the NJ approach (Figure 3). Excluding the Eucalyptus species, which are more assigned as a sister category to E. halophilum or foundation of the Brassicales. So it is far from their location as specified to the reference tree, the trees generally coincide with the reference tree. T. cacao in Fig 2 and C. papaya in Fig 3 are both in incorrect places.

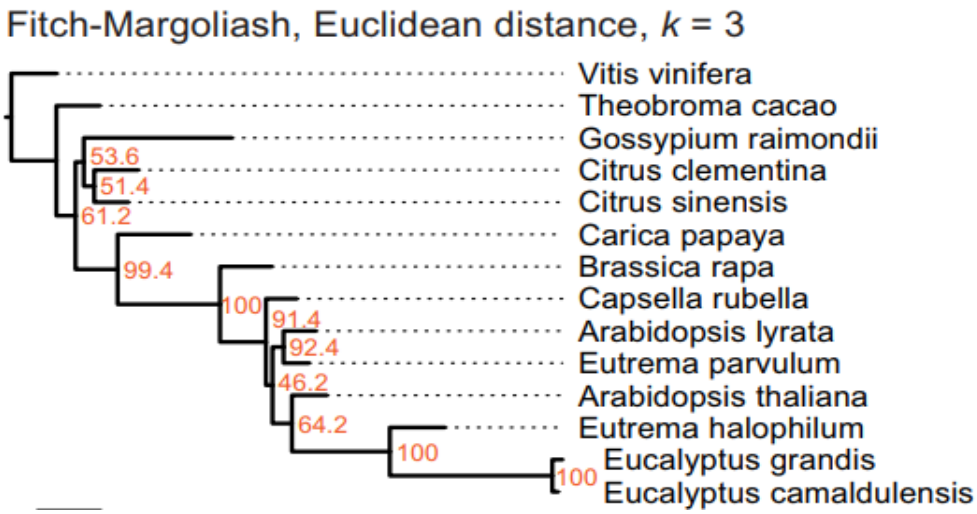

Figure 2: Fitch Margoliash algorithm (k=3)

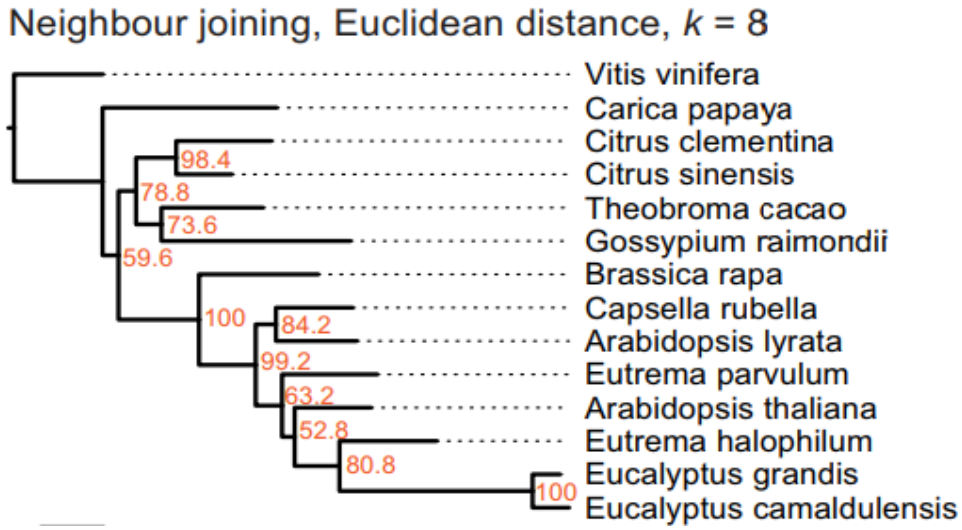

Figure 3: Neighbour joining algorithm (k=8)

Analysis of Mitochondrial Genome

Close neighbors of B. rapa were selected for this study from NCBI database as sequenced mitochondria. A. thaliana, Brasicaa carinata, Brasicaa juncea, B. napus, B. oleracea, B. rapa, Lotus japonicus, Millettia pinata, as well as Ricinus communis were among nine species for which mitochondrial gene sequences were present (Table 1). As an outgroup, the mitochondrial genome of V. vinifera employed. The trees on the basis of the FCRs of the mitochondrial genes were quite comparable for the two distinct approaches, in contrast to the analysis of the other datasets (Figures 4 and 5). The topology of the sub-branches comprising the 5 nearly associated Brasicaa species, in particular, is identical, as evidenced by strong bootstrap amounts. While the Brassicales sub-family tree's structure is clearly defined, the Fabales L. japonicus as well as M. pinnata, as well as the Malpighiales R. communis, all belong to the Fabales [30-34]. The trees based upon Euclidean distance along more-resolution FCRs (k = 8) show the identical well-backed topography collecting the Fabales together, regardless of the tree regeneration approach chosen. It is consistent with the findings from the full genome, which show that using high-resolution FCRs produces more plausible trees and that using the Euclidean approach to calculate distances is preferable.

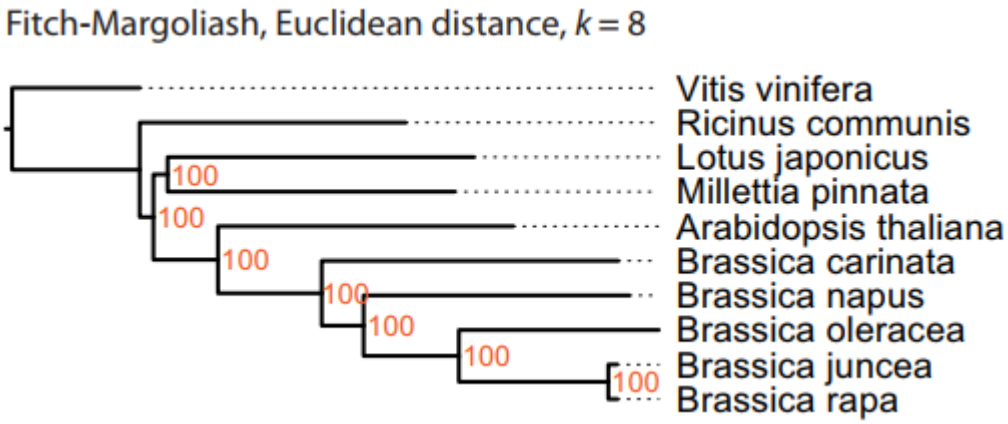

Figure 4: Fitch Margoliash algorithm (k=8)

The CRs and FCRs are calculated using an algorithm with linear time complexity O(L), where L is the nucleotide sequence length. The calculation of the CRs and FCRs for complete genomes took roughly 190 seconds, and 110 seconds for mitochondrial genomes. The distance matrix created for every species versus other species determines how long the approach takes to generate the phylogenetic trees. The time complexity of this calculation is O(4ks2), as well as the space complexity is O(s2), In which s represents count of species and k being the oligonucleotide size [35-39]. The phylogenetic tree reconstruction took 48 seconds for k = 8 and all over genome datavalues (k = 6: 10 s, k = 3: 4 s), and 36 seconds for k = 8 as well as mitochondrial genome datavalues (k = 6: 9 s, k = 3: 2 s).

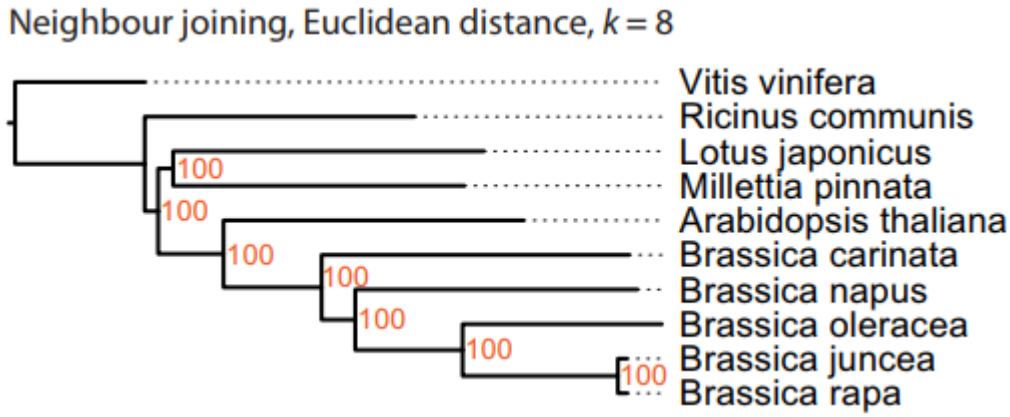

Figure 5: Neighbour joining algorithm (k=8)

5. Conclusion

Phylogenetic trees of various species redrawn using amino acid or nucleotide sequence values, morphological comparisons, or a combination of above methods. Whereas mostly sequence-centric comparisons based on single genomes, merged sequences, which can include entire transcriptomes, are becoming more popular (phylogenomics). We aimed to use alignment-free sequence data to rebuild the phylogeny of a few Brassicales species. We used CRs, which are scale-independent depictions for genomic sequences, as our approach. CRs cannot be directly compared since they are distinctive fingerprints of the relevant sequences. As a result, we constructed FCRs at various resolutions in order to reconstruct phylogenetic trees. The Euclidean (a geometric distance) was utilised to calculate the distances between FCRs, as well as trees were redrawn using the Fitch–Margoliash as well as NJ algorithms.

We analyzed two categories of nucleotide sequence, nuclear genome sequence and mitochondrial genome sequence, due to their differences. The GC content and codon use patterns of nuclear as well as mitochondrial genomes have been found to differ. The mitochondrial and entire genomes of the Brassicales species studied differ significantly in size. The existence as well as frequency of the corresponding oligonucleotides, and hence the size of the examined sequence, naturally affect FCRs. Finding the appropriate stability among sequence size and FCR resolution, that indicates amount of value accessible for tree calculation and too the key determinant for calculating time, is crucial for a decent outcome. To rule out the possibility that the sizes of concatenated sequences affect the phylogenetic tree regenerations of Brassicales species at higher FCR resolution, we

constructed trees that included total-length sequences as well as certain determined subsets.

We were able to demonstrate that FCRs are capable of phylogenetically grouping plant genes and exonomes from uniform nearly established species. This has been proved in part using the phylogeny of 26 mitochondrial genes, of which only 3 were put incorrectly using the Euclidean distance method. High resolution data yield in finer tree topologies as well as more brace for branchings, as per our research of the Brassicales clade. Trees constructed using the Pearson distance, a statistical distance metric, are slighter dependable with other constructed using Euclidean distances. The Fitch–Margoliash as well as NJ algorithms produce trees that are nearly identical. We demonstrated that the bootstrap idea for determining the support of branchings in trees, which has been well established for decades for trees on the basis of sequence alignments, can also be applied to trees on the basis of FCRs.

REFERENCES

1. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. Model Finder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589, https://doi.org/10.1038/nmeth.4285 (2017).
2. Freitas, T. A., Li, P. E., Scholz, M. B. & Chain, P. S. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. 43, e69, https://doi.org/10.1093/nar/gkv180 (2015).
3. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274, https://doi.org/10.1093/molbev/msu300 (2015).
4. Gruning, B. et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat. Methods 15, 475–476, https://doi.org/10.1038/s41592-018-0046-7 (2018).
5. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589, https://doi.org/10.1038/nmeth.4285 (2017).
6. Simon-Loriere, E. et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. Nat. 524, 102–104, https://doi.org/10.1038/nature14612 (2015).
7. Park, D. J. et al. Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. Cell 161, 1516–1526, https://doi.org/10.1016/j.cell.2015.06.007 (2015).
8. Dudas, G. et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nat. 544, 309–315, https://doi.org/10.1038/nature22040 (2017).
9. Smith, M. D. et al. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol. Biol. Evol. 32, 1342–1353, https://doi.org/10.1093/molbev/msv022 (2015).
10. Li, P. E. et al. Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. Nucleic Acids Res. 45, 67–80, https://doi.org/10.1093/nar/gkw1027 (2017).
11. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132, https://doi.org/10.1186/s13059-016-0997-x (2016).
12. Peter, J. et al. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nat. 556, 339–344, https://doi.org/10.1038/s41586-018-0030-5 (2018).
13. Dujon, B. A. & Louis, E. J. Genome Diversity and Evolution in the Budding Yeasts (Saccharomycotina). Genet. 206, 717–750, https://doi.org/10.1534/genetics.116.199216 (2017).
14. Gallone, B. et al. Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. Cell 166, 1397–1410 e1316, https://doi.org/10.1016/j.cell.2016.08.020 (2016).
15. Sulo, P. et al. The evolutionary history of Saccharomyces species inferred from completed mitochondrial genomes and revision in the 'yeast mitochondrial genetic code'. DNA Res. 24, 571–583, https://doi.org/10.1093/dnares/dsx026 (2017).
16. Carroll, M. W. et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. Nat. 524, 97–101, https://doi.org/10.1038/nature14594 (2015).
17. Lindsey, R. L. et al. Complete Genome Sequences of Two Shiga Toxin-Producing Escherichia coli Strains from Serotypes O119:H4 and O165:H25. Genome Announc 3, https://doi.org/10.1128/genomeA.01496-15 (2015).
18. Lorenz, S. C., Monday, S. R., Hoffmann, M., Fischer, M. & Kase, J. A. Plasmids from Shiga Toxin-Producing Escherichia coli Strains with Rare Enterohemolysin Gene (ehxA) Subtypes Reveal Pathogenicity Potential and Display a Novel Evolutionary Path. Appl. Env.Microbiol. 82, 6367–6377, https://doi.org/10.1128/AEM.01839-16 (2016).
19. Lorenz, S. C. et al. Complete Genome Sequences of Four Enterohemolysin-Positive (ehxA) Enterocyte Effacement-Negative Shiga Toxin-Producing Escherichia coli Strains. Genome Announc 4, https://doi.org/10.1128/genomeA.00846-16 (2016).
20. Lorenz, S. C., Gonzalez-Escalona, N., Kotewicz, M. L., Fischer, M. & Kase, J. A. Genome sequencing and comparative genomics of enterohemorrhagic Escherichia coli O145:H25 and O145:H28 reveal distinct evolutionary paths and marked variations in traitsassociated with virulence & colonization. BMC Microbiol. 17, 183, https://doi.org/10.1186/s12866-017-1094-3 (2017)
21. Depoorter, E. et al. Burkholderia: an update on taxonomy and biotechnological potential as antibiotic producers. Appl. Microbiol. Biotechnol. 100, 5215–5229, https://doi.org/10.1007/s00253-016-7520-x (2016).
22. Tuanyok, A. et al. Burkholderia humptydooensis sp. nov., a New Species Related to Burkholderia thailandensis and the Fifth Member of the Burkholderia pseudomallei Complex. Appl Environ Microbiol 83, https://doi.org/10.1128/AEM.02802-16 (2017)
23. Pettengill, E. A., Pettengill, J. B. & Binet, R. Phylogenetic Analyses of Shigella and Enteroinvasive Escherichia coli for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct GeneraDesignation. Front. Microbiol. 6, 1573, https://doi.org/10.3389/fmicb.2015.01573 (2015).
24. Hata, H. et al. Phylogenetics of family Enterobacteriaceae and proposal to reclassify Escherichia hermannii and Salmonella subterranea as Atlantibacter hermannii and Atlantibacter subterranea gen. nov., comb. nov. Microbiol. Immunol. 60, 303–311, https://doi.org/10.1111/1348-0421.12374 (2016)
25. Kurylo, C. M. et al. Genome Sequence and Analysis of Escherichia coli MRE600, a Colicinogenic, Nonmotile Strain that Lacks RNase I and the Type I Methyltransferase, EcoKI. Genome Biol. Evol. 8, 742–752, https://doi.org/10.1093/gbe/evw008 (2016).
26. Lee, T. H., Guo, H., Wang, X., Kim, C. & Paterson, A. H. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15, 162, https://doi.org/10.1186/1471-2164-15-162 (2014).
27. Faison, W. J. et al. Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human

genomes. Genomics 104, 1–7, https://doi.org/10.1016/j.ygeno.2014.06.001 (2014)
28. Griffing, S. M. et al. Canonical Single Nucleotide Polymorphisms (SNPs) for High-Resolution Subtyping of Shiga-Toxin Producing Escherichia coli (STEC) O157:H7. PLoS One 10, e0131967, https://doi.org/10.1371/journal.pone.0131967 (2015).
29. Gardner, S. N., Slezak, T. & Hall, B. G. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinforma. 31, 2877–2878, https://doi.org/10.1093/bioinformatics/btv271 (2015).
30. Sahl, J. W. et al. Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data. Genome Med. 7, 52, https://doi.org/10.1186/s13073-015-0176-9 (2015).
31. Sahl, J. W. et al. NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. Microb. Genom. 2, e000074, https://doi.org/10.1099/mgen.0.000074 (2016).
32. Davis, S. et al. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Computer Sci. 1, e20 (2015).
33. Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M. & Lund, O. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS One 9, e104984, https://doi.org/10.1371/journal.pone.0104984 (2014)
34. Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B. & van Nimwegen, E. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol. Biol. Evol. 31, 1077–1088, https://doi.org/10.1093/molbev/msu088 (2014).
35. Sankarasubramanian, J., Vishnu, U. S., Gunasekaran, P. & Rajendhran, J. A genome-wide SNP-based phylogenetic analysis distinguishes different biovars of Brucella suis. Infect. Genet. Evol. 41, 213–217, https://doi.org/10.1016/j.meegid.2016.04.012 (2016).
36. Lomsadze, A., Gemayel, K., Tang, S. & Borodovsky, M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Res. 28, 1079–1089, https://doi.org/10.1101/gr.230615.117 (2018)
37. Girault, G., Blouin, Y., Vergnaud, G. & Derzelle, S. High-throughput sequencing of Bacillus anthracis in France: investigating genome diversity and population structure using whole-genome SNP discovery. BMC Genomics 15, 288, https://doi.org/10.1186/1471-2164-15-288 (2014)
38. Katz, L. S. et al. A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. Front. Microbiol. 8, 375, https://doi.org/10.3389/fmicb.2017.00375 (2017)
39. Petkau, A. et al. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. Microb. Genom. 3, e000116, https://doi.org/10.1099/mgen.0.000116 (2017)