W554–W559 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 3 May 2017*
doi: 10.1093/nar/gkx351

CAFE: aCcelerated Alignment-FrEe sequence analysis

Yang Young Lu[1], Kujin Tang[1], Jie Ren[1], Jed A. Fuhrman[2], Michael S. Waterman[1,3] and Fengzhu Sun[1,3,*]

[1]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, CA 90089, USA, [2]Department of Biological Sciences and Wrigley Institute for Environmental Studies, University of Southern California, Los Angeles, CA 90089, USA and [3]Centre for Computational Systems Biology, School of Mathematical Sciences, Fudan University, 200433 Shanghai, China

Received February 27, 2017; Revised April 8, 2017; Editorial Decision April 16, 2017; Accepted April 20, 2017

ABSTRACT

Alignment-free genome and metagenome comparisons are increasingly important with the development of next generation sequencing (NGS) technologies. Recently developed state-of-the-art *k*-mer based alignment-free dissimilarity measures including *CVTree*, d_2^* and d_2^S are more computationally expensive than measures based solely on the *k*-mer frequencies. Here, we report a standalone software, aCcelerated Alignment-FrEe sequence analysis (CAFE), for efficient calculation of 28 alignment-free dissimilarity measures. CAFE allows for both assembled genome sequences and unassembled NGS shotgun reads as input, and wraps the output in a standard PHYLIP format. In downstream analyses, CAFE can also be used to visualize the pairwise dissimilarity measures, including dendrograms, heatmap, principal coordinate analysis and network display. CAFE serves as a general *k*-mer based alignment-free analysis platform for studying the relationships among genomes and metagenomes, and is freely available at https://github.com/younglululu/CAFE.

INTRODUCTION

Sequence comparison is widely used to study the relationship among molecular sequences. The dominant tools for sequence comparison are alignment-based methods, including global (1) and local (2) sequence alignments. With the advent of alignment-based tools such as BLAST (3) and sequence databases such as RefSeq (4), alignment-based methods are widely used in a broad range of applications. Despite their extensive applications, alignment-based methods are not appropriate in some situations. First, gene regulatory regions are generally not highly conserved making alignment-based approaches difficult to identify related regulatory regions that are bound by similar transcription factors (5). Second, next generation sequencing (NGS) technologies generate large amounts of short reads and it is challenging to assemble them for both genomic and metagenomic studies. Without long assembled contigs across many samples, it is challenging for alignment-based methods to compare genomes and metagenomes (6,7). Third, viruses are more likely to infect bacterial hosts having similar word pattern usage (8,9), and thus, the hosts of viruses can potentially be inferred based on their word pattern usages. However, alignment based methods are usually not applicable for studying virus-host infectious associations.

Alignment-free sequence comparison methods serve as attractive alternatives for studying the relationships among sequences when alignment based methods are not appropriate or too time consuming to be implemented in practice (10,11). Several types of alignment free approaches are available including those based on the counts of *k*-mers, longest common subsequences, shortest absent patterns, etc. that have recently been reviewed in a special issue of Briefing in Bioinformatics (12). Here we concentrate on alignment-free statistics using *k*-mer counts. These approaches project each sequence into *k*-mer (or equivalently *k*-tuple, *k*-gram) counts feature space, where sequence information is transformed into numerical values such as *k*-mer frequency. We do not consider dissimilarity measures using spaced *k*-mers due to the added computational complexity counting spaced *k*-mers. The recently developed statistics d_2^* and d_2^S have been shown to perform well theoretically (13) as well as in many applications including the comparison of gene regulatory regions (11), whole genome sequences (14), metagenomes (7) and virus-bacteria host infectious associations (8). Despite their excellent performance in many applications, the original implementation of these statistics are relatively slow due to the requirement of calculating the expected *k*-mer counts and thus limits their usage.

CAFE significantly speeds up the calculation of recently developed measures based on background adjusted *k*-mer counts, such as *CVTree* (15), d_2^* (13) and d_2^S (13), with

*To whom correspondence should be addressed. Tel: +1 213 740 2413; Fax: +1 213 740 8631; Email: fsun@usc.edu
Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

reduced memory requirement. In addition, CAFE integrates 10 conventional measures based on *k*-mer counts such as *Chebyshev* (*Ch*), *Euclidean* (*Eu*), *Manhattan* (*Ma*), d_2 dissimilarity (16), Jensen-Shannon divergence (*JS*) (17), feature frequency profiles (*FFP*) (18) and *Co-phylog* (19). CAFE also offers 15 measures based on presence/absence of *k*-mers, such as *Jaccard* and *Hamming* distances. We further demonstrate the value of alignment-free dissimilarity measures using CAFE on real datasets, ranging from primate, vertebrate and microbial genomic sequences, to metagenomic sequence reads.

MATERIALS AND METHODS

Workflows

CAFE works with sequence data, both assembled genomic sequences and unassembled shotgun sequence reads from NGS technologies and counts *k*-mers by JELLYFISH (20), a fast and memory-efficient *k*-mer counting tool. JELLYFISH produces compressed databases containing all *k*-mer counts given the query sequences in parallel. CAFE subsequently loads the databases and generates necessary transformed information with respect to various dissimilarity measures. For example, measures based on presence/absence of *k*-mers binarize *k*-mer counts into presence/absence indicators. Most conventional measures normalize *k*-mer counts into the *k*-mer frequencies. Besides, expected *k*-mer counts are involved in recently developed measures based on background adjusted *k*-mer counts, such as *CVTree*, d_2^* and d_2^S. In such cases, the Markov models for the sequences are assumed as the underlying generative models, with the parameters estimated from the sequence data accordingly. The Markov order can be either manually set or automatically chosen using the Bayesian information criterion (BIC) (21).

The resulting pairwise dissimilarities among the sequences form a symmetric matrix. CAFE can directly output the dissimilarity matrix in a standard PHYLIP format. Alternatively, CAFE provides four types of built-in downstream visualized analyses, including clustering the sequences into dendrograms using the UPGMA algorithm, heatmap visualization of the matrix, projecting the matrix to a 2D space using principal coordinate analysis (PCoA) and network display. A graphical illustration of CAFE workflow is shown in Figure 1.

Graphical user interface

The CAFE user interface consists of four major tools—data selection toolbar, dissimilarity setting toolbar, image toolbar and visualized analyses. The data selection toolbar enables users to browse and add/delete genome sequences or NGS shotgun reads of the file extension '.fasta', '.fa' or '.fna'. The selected files are shown in the input data list. The data selection toolbar also supports loading pre-computed results in a standard PHYLIP format.

The dissimilarity setting toolbar determines the choice of dissimilarity measures as well as the involved parameter configuration, including the *k*-mer length, the order of potential Markov model, the cutoff of the minimum *k*-mer occurrences and whether to consider the reverse complement of each *k*-mer, a common practice in dealing with NGS shotgun reads. When a certain parameter is unnecessary for particular dissimilarity measures, the corresponding configuration is disabled. In the cases of *CVTree*, d_2^* and d_2^S, usually the proper order of Markov model remains unclear to the user. A simple yet time-consuming way is to set '-1', which will infer the order automatically using the BIC (21).

After the 'Run' button is pressed, the CAFE workflow starts, and consolidated dissimilarity results are saved in a standard PHYLIP format, together with the run-time information trackable from the console. Meanwhile, four types of built-in analyses are provided in tabbed windows, including dendrograms, heatmap, PCoA and network display.

The view of the visualized analyses can be adjusted by using the 'zoom-in' and 'zoom-out' buttons located in the image toolbar. CAFE also supports downloading the visualized results for publication. To access this function, users can either use the 'save' button in the image toolbar or right-click on the figure directly. A screenshot of the CAFE user interface is shown in Figure 2.

Design

CAFE is designed for extensibility and reusability, following the software engineering paradigm. For example, users can specify a threshold to filter out *k*-mers whose counts are below the threshold. In this case, the Iterator hides the details of filtering, wrapping the enumeration of qualified *k*-mer counts or frequencies uniformly. Also, some dissimilarity measures do not need the expected *k*-mer counts. Hence the Proxy provides the calculation of expected *k*-mer counts as a service on demand. Moreover, the dissimilarity measures are encapsulated in Strategy, enabling users to integrate customized dissimilarity measures into CAFE easily as plug-in.

RESULTS

Application to primate and vertebrate genomic sequences

We compared various alignment-free dissimilarity measures using CAFE on three real genomic datasets. We first investigated the evolutionary relationship of 21 primates whose complete genome sequences are available in the NCBI database (22). For each dissimilarity measure, the calculated pairwise dissimilarity measures are directly compared against the corresponding evolutionary distances calculated by Ape (An R package) (23) as the benchmark, in terms of Spearman correlations. Comparison using Pearson correlations between the estimated alignment-free dissimilarity and the evolutionary distances, and normalized Robinson-Foulds distance (24) between the clustering tree using UPGMA and the standard phylogenetic tree are also available in the supplementary material. Similarly, we investigated the evolutionary relationship of 28 vertebrate species and compared the alignment-free dissimilarity measures with the pairwise evolutionary distances given in (25). Finally, we combined the two datasets to see how the alignment-free dissimilarity measures relate to evolutionary distances

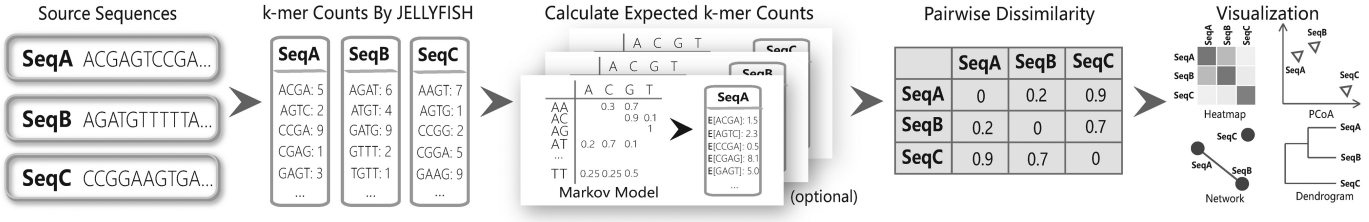

Figure 1. The workflow of CAFE. The JELLYFISH software parses the input sequence files (in Fasta format), counts *k*-mers and saves compressed information into separate databases. CAFE subsequently loads the databases and constructs a symmetric dissimilarity matrix among the inputs. CAFE also integrates four types of visualized downstream analysis, including dendrograms, heatmap, principal coordinate analysis (PCoA) and network display.

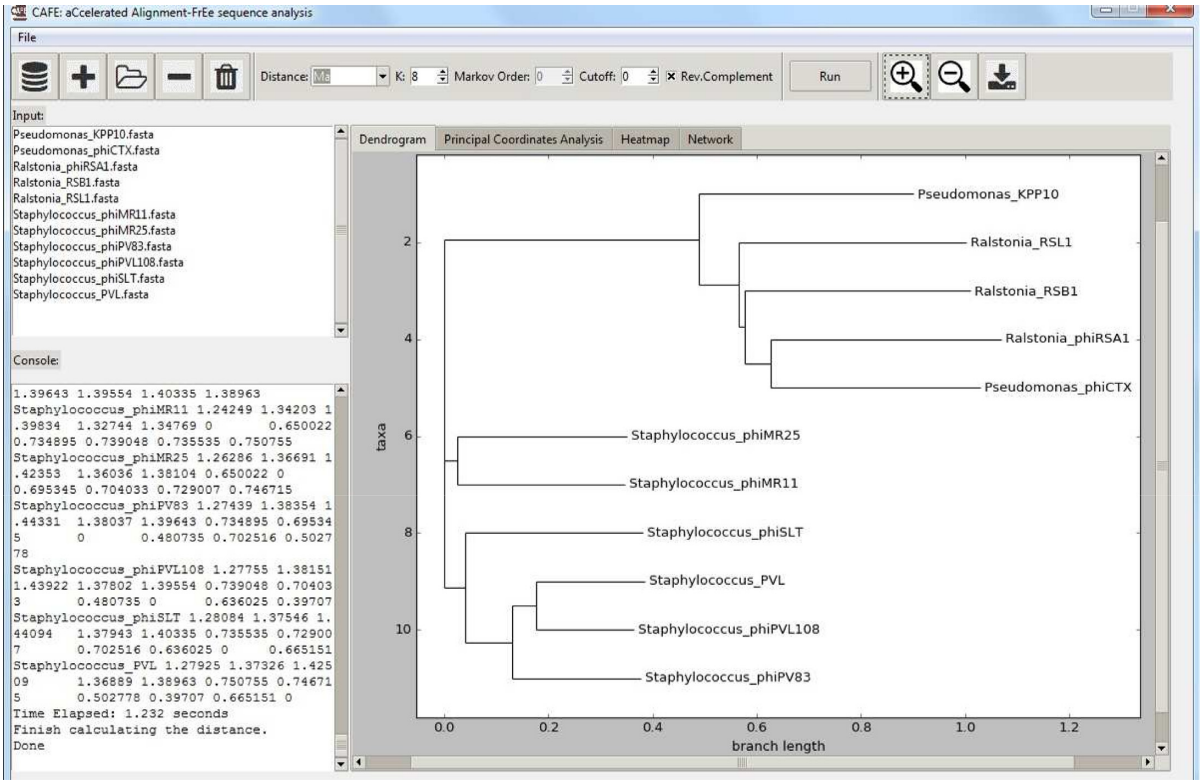

Figure 2. Screenshot of CAFE user interface based on a toy example. The user interface layout divides into six parts in terms of functionality: (i) data selection toolbar (top left), (ii) dissimilarity setting toolbar (top middle), (iii) image toolbar (top right), (iv) input data list (middle left), (v) run-time information console (bottom left) and (vi) visualized analyses (bottom right).

calculated based on maximum likelihood approach from a large number of genomic regions.

The comparison involves three dissimilarity measures based on background adjusted *k*-mer counts including *CVTree*, d_2^* and d_2^S, 10 conventional measures based on *k*-mer counts, including *Canberra*, *Ch*, *Cosine*, *Co-phylog*, d_2, *Eu*, *FFP*, *JS*, *Ma* and *Pearson*, and 15 measures based on presence/absence of *k*-mers including *Anderberg*, *Antidice*, *Dice*, *Gower*, *Hamman*, *Hamming*, *Jaccard*, *Kulczynski*, *Matching*, *Ochiai*, *Phi*, *Russel*, *Sneath*, *Tanimoto* and *Yule*. We used $k = 14$ as in (14). The results are illustrated in Figure 3. The Markov order 12 is used for d_2^*, d_2^S and *JS* as most of the sequences have estimated order 12 based on BIC (21). Consistent with previous studies, the background adjusted dissimilarity measures outperform markedly the non-background adjusted measures.

We then evaluate the computational speed of CAFE compared to the original implementation for d_2^* in (14). We calculate the dissimilarity using d_2^* measure (d_2^* and d_2^S share highly similar formulation) on random pairs of genome sequences. As shown in Figure 4, CAFE achieves 24.0× speedup with 55.3% peak memory consumption on average.

Application to microbial genomic sequences

We applied CAFE to analyze 27 *E. coli* and *Shigella* genomes dataset as in (26). These genomes can be assigned to 6 *E. coli* reference (ECOR) groups: A, B1, B2, D, E and S. We investigated how well various alignment-free dissimilarity measures can identify these groups. For each dissimilarity measure, we used UPGMA to cluster the samples based on the calculated pairwise dissimilarity matrix. The Markov

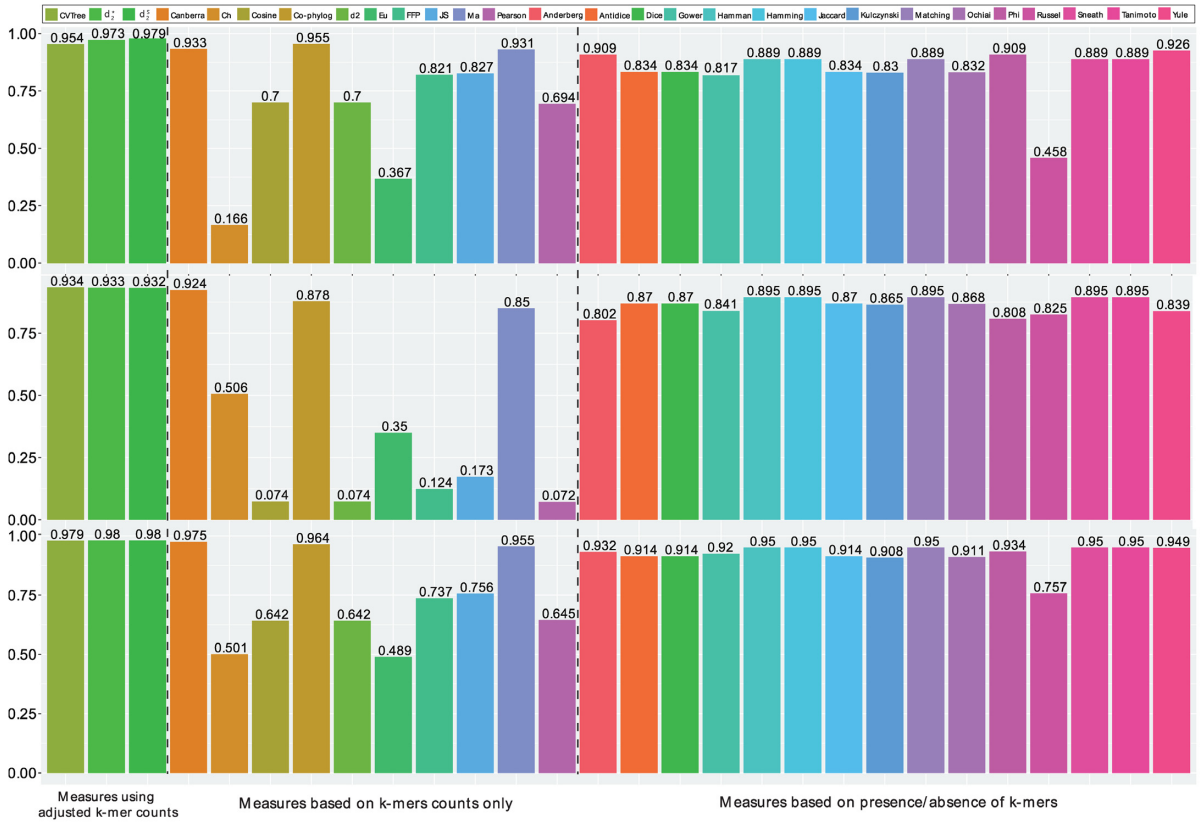

Figure 3. The Spearman correlation of various dissimilarity measures with the evolutionary distances using maximum likelihood approach across many genomic regions based on 21 primate species (top), 28 vertebrate species (middle) and the combination of both (bottom).

Sequence Model	Original Implementation		CAFE		
	Wall time	Peak memory	Wall time	Speedup	Peak memory
order=0	0:42′32″	64.0G	0:6′09″	6.9x	31.1G
order=1	1:44′18″	64.0G	0:6′13″	16.8x	31.1G
order=2	2:11′32″	64.0G	0:6′12″	21.2x	31.1G
order=3	2:34′28″	62.4G	0:5′05″	30.4x	24.8G
order=4	2:34′11″	62.3G	0:6′10″	25.0x	31.1G
order=5	3:24′43″	64.0G	0:5′08″	39.9x	24.8G
order=6	2:53′08″	63.9G	0:5′14″	33.1x	24.8G
order=7	2:40′04″	64.0G	0:6′29″	24.7x	31.1G
order=8	2:33′19″	64.0G	0:6′08″	25.0x	48.1G
order=9	2:37′50″	64.2G	0:6′19″	25.0x	48.2G
order=10	2:22′18″	64.7G	0:5′15″	27.1x	48.5G
order=11	2:05′55″	60.4G	0:6′29″	19.4x	49.6G
order=12	1:53′40″	74.6G	0:6′39″	17.1x	37.0G

Figure 4. Wall time, peak memory usage and speedup ratio comparison between CAFE and the original implementation to calculate d_2^* dissimilarity between a pair of genomes for $k = 14$.

order 1 is used for d_2^* and d_2^S as most of the sequences have estimated order 1 based on BIC (21).

We used $k = 14$ for the comparison. The comparison involves three dissimilarity measures based on background adjusted k-mer counts including *CVTree*, d_2^* and d_2^S, and the results are illustrated in Figure 5. The results using the other 10 conventional measures based on 14-mer counts as well as 15 measures based on presence/absence of 14-mers, are given in the supplementary material. Consistent with previous studies, for d_2^S, each ECOR is monophyletic except A and B2. The normalized Robinson-Foulds distances (24) between the estimated clustering tree and the standard phylogenetic tree are available in the Supplementary Data.

Application to metagenomic samples

We used CAFE to analyze a mammalian gut metagenomic dataset (7) comprised of NGS short reads from 28 samples. These samples further split into 3 groups: 8 hindgut-fermenting herbivores, 13 foregut-fermenting herbivores and 7 simple-gut carnivores. We investigated how well various alignment-free dissimilarity measures can identify these groups. For each dissimilarity measure, we used UPGMA to cluster the samples based on the calculated pairwise dissimilarity matrix.

We used $k = 5$ as in (7). The comparison involves three dissimilarity measures based on background adjusted k-mer counts including *CVTree*, d_2^* and d_2^S, and the results are illustrated in Figure 6. The results based on nine conventional measures based on k-mer counts are given in the Supplementary Data. Other measures are not applicable because $k = 5$ is not large enough. The Markov order 0 is used in d_2^* and d_2^S as in (7). Consistent with previous studies, d_2^S achieves clear separations among the three groups.

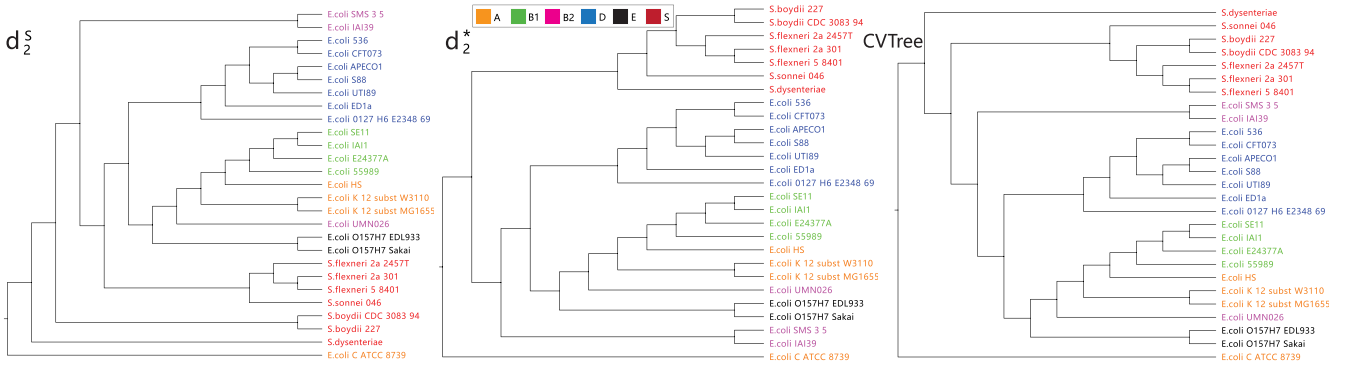

Figure 5. The clustering results of 27 *Escherichia coli* and *Shigella* genomes using measures based on background adjusted 14-mer counts: d_2^S, d_2^* and *CVTree*. The Markov order of the sequences were set at 1. The colors indicate the six different *E. Coli* reference groups.

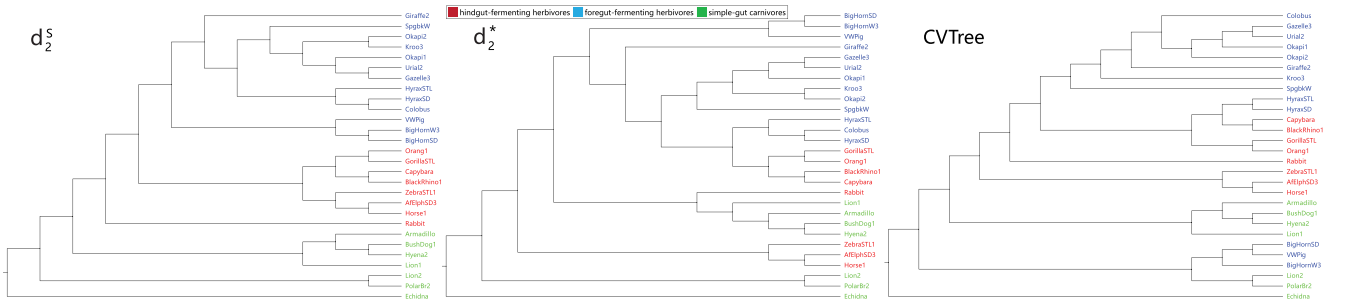

Figure 6. The clustering results of the mammalian gut samples using measures based on background adjusted *k*-mer counts: d_2^S, d_2^* and *CVTree*.

DISCUSSION

We have developed a fast and user-friendly alignment-free analyses platform, CAFE, for studying the relationships among genomes and metagenomes. With reduced memory usage, CAFE speeds up the calculation of the state-of-the-art alignment-free measures that perform well theoretically and practically. For easy usage, CAFE not only integrates 28 dissimilarity measures extensively but also integrates four types of downstream visualized analyses. CAFE will make the usage of alignment-free methods more accessible to researchers. We encourage users to contribute their own dissimilarity measures to CAFE as plug-ins.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

NSF [DMS-1518001]; OCE [1136818]; NIH [R01GM120624]; Gordon and Betty Moore Foundation Marine Microbiology Initiative [GBMF3779]. Funding for open access charge: NSF [DMS-1518001].
Conflict of interest statement. None declared.

REFERENCES

1. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
2. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
3. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
4. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**(Suppl. 1), D501–D504.
5. Leung,G. and Eisen,M.B. (2009) Identifying cis-regulatory sequences by word profile similarity. *PLoS One*, **4**, e6901.
6. Willner,D., Vega,T.R. and Rohwer,F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol.*, **11**, 1752–1766.
7. Jiang,B., Song,K., Ren,J., Deng,M., Sun,F. and Zhang,X. (2012) Comparison of metagenomic samples using sequence signatures. *BMC Genomics*, **13**, 730.
8. Ahlgren,N.A., Ren,J., Young,L.Y., Fuhrman,J.A. and Sun,F. (2017) Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.*, **45**, 39–53.
9. Roux,S., Hallam,S.J., Woyke,T. and Sullivan,M.B. (2015) Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife*, **4**, e08490.
10. Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison-a review. *Bioinformatics*, **19**, 513–523.
11. Song,K., Ren,J., Reinert,G., Deng,M., Waterman,M.S. and Sun,F. (2014) New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief. Bioinformtics*, **15**, 343–353.
12. Vinga,S. (2014) Editorial: Alignment-free methods in computational biology. *Brief. Bioinform.*, **15**, 341–342.
13. Reinert,G., Chew,D., Sun,F. and Waterman,M.S. (2009) Alignment-free sequence comparison (I): statistics and power. *J. Comput. Biol.*, **16**, 1615–1634.
14. Ren,J., Song,K., Deng,M., Reinert,G., Cannon,C.H. and Sun,F. (2016) Inference of Markovian properties of molecular sequences

from NGS data and applications to comparative genomics. *Bioinformatics*, **32**, 993–1000.
15. Qi,J., Luo,H. and Hao,B. (2004) CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res.*, **32**(**Suppl. 2**), W45–W47.
16. Blaisdell,B.E. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl. Acad. Sci.*, **83**, 5155–5159.
17. Jun,S.R., Sims,G.E., Wu,G.A. and Kim,S.H. (2010) 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.
18. Sims,G.E., Jun,S.R., Wu,G.A. and Kim,S.H. (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 2677–2682.
19. Yi,H. and Jin,L. (2013) Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Res.*, **41**, e75.
20. Marçais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.
21. Narlikar,L., Mehta,N., Galande,S. and Arjunwadkar,M. (2013) One size does not fit all: on how Markov model order dictates performance of genomic sequence analyses. *Nucleic Acids Res.*, **41**, 1416–1424.
22. Perelman,P., Johnson,W.E., Roos,C., Seuánez,H.N., Horvath,J.E., Moreira,M.A.M., Kessing,B., Pontius,J., Roelke,M., Rumpler,Y. *et al.* (2011) A molecular phylogeny of living primates. *PLoS Genet.*, **7**, e1001342.
23. Paradis,E., Claude,J. and Strimmer,K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.
24. Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.
25. Miller,W., Rosenbloom,K., Hardison,R.C., Hou,M., Taylor,J., Raney,B., Burhans,R., King,D.C., Baertsch,R., Blankenberg,D. *et al.* (2007) 28-way vertebrate alignment and conservation track in the UCSC genome browser. *Genome Res.*, **17**, 1797–1808.
26. Bernard,G., Chan,C.X. and Ragan,M.A. (2016) Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Sci. Rep.*, **6**, 28970.