Efficient computation of Faith's phylogenetic diversity with applications in characterizing microbiomes

George W Armstrong, Kalen Cantrell, Shi Huang, et al.

Genome Res. published online September 3, 2021
Access the most recent version at doi:10.1101/gr.275777.121

Efficient computation of Faith's phylogenetic diversity with applications in characterizing microbiomes

George Armstrong[1,2,3], Kalen Cantrell[2], Shi Huang[1,2], Daniel McDonald[1], Niina Haiminen[4], Anna Paola Carrieri[5], Qiyun Zhu[6,7], Antonio Gonzalez[1], Imran McGrath[2,8], Kristen L. Beck[9], Daniel Hakim[1,3], Aki S. Havulinna[10,11], Guillaume Méric[12,13], Teemu Niiranen[10,14,15], Leo Lahti[16], Veikko Salomaa[10], Mohit Jain[2,17,18], Michael Inouye[12,19], Austin D. Swafford[2], Ho-Cheol Kim[9], Laxmi Parida[4], Yoshiki Vázquez-Baeza[2], Rob Knight[1,2,20,21,#]

[1]Department of Pediatrics, School of Medicine, University of California, San Diego, California, USA; [2]Center for Microbiome Innovation, Jacobs School of Engineering, University of California San Diego, La Jolla, California, USA; [3]Bioinformatics and Systems Biology Program, University of California, San Diego, California, USA; [4]IBM T. J. Watson Research Center, Yorktown Heights, New York, USA; [5]IBM Research Europe, The Hartree Centre, Warrington, United Kingdom; [6]School of Life Sciences, Arizona State University, Tempe, Arizona, USA; [7]Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, Arizona, USA; [8]Division of Biological Sciences, University of California San Diego, La Jolla, California, USA; [9]IBM Almaden Research Center, San Jose, California, USA; [10]Department of Public Health and Welfare, Finnish Institute for Health and Welfare, Helsinki, Finland; [11]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland; [12]Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia; [13]Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia; [14]Department of Internal Medicine, University of Turku, Turku, Finland; [15]Division of Medicine, Turku University Hospital, Finland; [16]Department of Computing, University of Turku, Turku, Finland; [17]Department of Medicine, University of California, San Diego, California, USA; [18]Department of Pharmacology, University of California, San Diego, California, USA; [19]Department of Public Health and Primary Care, Cambridge University, Cambridge, UK; [20]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA; [21]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA;

Corresponding Author: robknight@ucsd.edu

Abstract

The number of publicly available microbiome samples is continually growing. As dataset size increases, bottlenecks arise in standard analytical pipelines. Faith's phylogenetic diversity is a highly utilized phylogenetic alpha diversity metric that has thus far failed to effectively scale to trees with millions of vertices. Stacked Faith's Phylogenetic Diversity (SFPhD) enables calculation of this widely adopted diversity metric at a much larger scale by implementing a computationally efficient algorithm. The algorithm reduces the amount of computational resources required, resulting in more accessible software with a reduced carbon footprint, as compared to previous approaches. The new algorithm produces identical results to the previous method. We further demonstrate that the phylogenetic aspect of Faith's PD provides increased power in detecting diversity differences between younger and older populations in the FINRISK study's metagenomic data.

Introduction

In microbiome research, particular attention is given to evaluating the diversity of microbes within samples (McDonald et al. 2018a; The Human Microbiome Project Consortium 2012; Thompson et al. 2017). Alpha diversity (within sample diversity) represents a family of summary statistics that can summarize the breadth of diversity present in an environment. More recently, many examples have been reported on the associations between various host or environmental factors and alpha diversity of microbiomes, including country and diet in human guts (McDonald et al. 2018a), disease status in humans and canines (Gevers et al. 2014; Vázquez-Baeza et al. 2016), the pH (Lauber et al. 2009), salinity (Thompson et al. 2017) and temperature (Zhou et al. 2016) of soils, among many others (Youngblut et al. 2019; Jeffery et al. 2016). A popular metric that accounts for the phylogenetic relatedness of the community members, Faith's Phylogenetic

Diversity (Faith's PD) (Faith 1992), has been noted to be more sensitive in distinguishing disease factors in the human digestive system, relative to other alpha diversity indices (Scherson and Faith 2018; Youngblut et al. 2021).

Modern DNA sequencing instruments have enabled microbiome studies at the scale of tens of thousands of samples, which presents a computational challenge for metrics that rely on a phylogeny, such as Faith's PD. This metric is computed by summing the branch lengths (edge weights) of the phylogeny that exclusively represents the sequences contained in a biological sample. The amount of memory and number of necessary operations needed to calculate Faith's PD depends on the number of edges in the phylogenetic tree, as well as the number of samples in the underlying data table.

In today's increasingly large and sparse datasets and meta-analyses, these phylogenetic trees and tables can exceed 100,000s of samples and millions of tree tips (McDonald et al. 2018b). Recent advances have enabled efficient computation of the UniFrac metric for beta diversity. UniFrac is also a metric computed over phylogenetic trees (Lozupone and Knight 2005) and is mathematically related to Faith's PD (Faith et al. 2009). Specifically, Striped UniFrac (McDonald et al. 2018b) improves upon previous UniFrac implementations (Hamady et al. 2010) by using space- and time-efficient tree data structures (Cordova and Navarro 2016) and reducing the number of vectors required to store intermediate scores in the tree.

Additionally, the usefulness of techniques like Faith's PD and UniFrac remains underexplored for metagenomics sequencing. Recent molecular protocol optimizations, such as SHOGUN (Hillmann et al. 2018), have enabled the metagenomic characterization of large human cohorts (Borodulin et al. 2015; Kaplan et al. 2019; Salosensaari et al. 2021). In this context, the applicability of Faith's PD has largely been limited by the technical difficulties associated with constructing phylogenies from metagenomic features (Zhu et al. 2019). Efforts

like the Web of Life (WoL) (Zhu et al. 2019) and Genome Taxonomy Database (GTDB) (Parks et al. 2018, 2020) are now addressing this issue by providing a phylogenomic tree as part of their database releases that can be used for phylogeny-informed analysis.

Motivated by these advances in algorithms and resources for analyzing phylogenies, phylogenomic trees, and sparse data, we developed a new algorithm and implementation, Stacked Faith's Phylogenetic Diversity (SFPhD), for rapidly computing Faith's PD. Additionally, we aim to demonstrate concrete benefits of phylogeny-informed analysis in metagenomic studies where this metric is less frequently used.

Results

SFPhD is a new implementation for calculating Faith's PD. The key advances of SFPhD are using a sparse matrix representation, an efficient tree structure, and partial aggregation of metric constituents. Our BSD-licensed implementation of this algorithm is available in the `unifrac` package (via PyPI and bioconda (Grüning et al. 2018)), which has 57,007 total conda downloads and 40,434 conda downloads since the introduction of SFPhD, as of the time of writing (August 28, 2021). The package produces a C/C++ shared library with Python bindings and is additionally linkable by any programming language (https://github.com/biocore/unifrac). Additionally, by investigating the previously documented relationship between age and bacterial richness of the gut microbiome (de la Cuesta-Zuluaga et al. 2019), we demonstrate that accounting for phylogeny in metagenomic data can increase the statistical power for detecting group differences (**Supplemental Code**).

Stacked Faith's PD provides a faster and memory-efficient implementation over the previous state-of-the-art algorithm.

SFPhD uses the structure of microbiome data along with other practical considerations to achieve decreased time and memory requirements. An example feature table is shown in **Fig. 1A**, with a corresponding phylogenetic tree in **Fig. 1B**. Note that for a given tree $\mathcal{T}$ Faith's PD can be expressed as

$$PD_i = \sum_{j \in \mathcal{T}} I_{ij} \times \text{branchLen}_j(\mathcal{T})$$

where PD_i is Faith's PD for sample i, I_{ij} indicates if sample i has any features that descend from node j, and $\text{branchLen}_j(\mathcal{T})$ indicates the length of the branch to node j in the tree $\mathcal{T}$.

The previous state-of-the-art reference implementation (scikit-bio, http://scikit-bio.org/) computes Faith's PD for a batch of samples by first fully computing I_{ij}. I_{ij} is computed by traversing the entire phylogenetic tree in a post-order traversal, where the children of a node must be visited before the node itself can be visited (the nodes in **Fig. 1B** are labeled in the order of a post-order traversal). During the traversal, when a given node I_{ij} is visited, all j are set by determining the features present in all children of node j. Subsequently, the $I_{ij} \times \text{branchLen}_j(\mathcal{T})$ for all branches is calculated. The results are obtained by summing over the branches for each sample (**Fig. 1C**). However, this approach tends to use much more space than is needed.

Microbiome data are known to be sparse (Martino et al. 2019; Kumar et al. 2018; Morton et al. 2017), i.e., of the entries in a data table, many are likely to be zero. This issue is exacerbated in large datasets, where many microbes are only observed in a handful of samples. In an extreme case, such a table (McDonald et al. 2018b) with 113,721 samples rarefied at 500 sequences per sample, has only 0.0126% non-zero entries. Sparse representations have been used previously for storing microbiome data (McDonald et al. 2012a), and have been applied for accelerating microbiome analyses (McDonald et al. 2018b), but they have not been previously applied to Faith's PD. We identified that a major downfall of the state-of-the-art implementation

in scikit-bio is that it uses a full, dense table to represent all of I_{ij} in memory at once. A key advancement of our approach is to use a sparse matrix implementation for storing information on the taxa present for each sample and feature. Sparse matrices save space by only retaining information about positions in the matrix that have non-zero values (e.g., only the grey values in **Fig. 1A** and information about their positions are retained by a sparse matrix).

Another key advance is the partial aggregation of Faith's PD (**Fig. 1D**). Note that the $I_{ij} \times \text{branchLen}_j(T)$, which we will call a metric constituent, can be added in any order, and that I_{ij} only depends on the children of node j. Thus, if node k is a child of node j, I_{ik} is no longer needed once metric constituents for node k have been computed and I_{ij} is known. As a result, we can reduce the memory used to store I_{ij} by traversing the phylogeny with a post-order traversal and freeing I_{ik} after they are no longer needed. Furthermore, we can reduce the storage needed for the metric constituents keeping a running summation of them while traversing the tree. Thus, this approach reduces the expected space complexity for storing the metrics from $O(nk)$, to $O(n \log(k))$, where n is the number of samples and k is the number of vertices in the tree.

In addition to the algorithmic improvements, we have included several practical enhancements that improve the performance of the code. The topology of the phylogenetic tree (**Fig. 1B**) is now represented as balanced-parentheses vector (**Fig. 1E**) that corresponds to additional vectors of branch lengths and node names; this structure has a lower memory footprint and a sequential memory representation which reduces the number of cache misses during a tree traversal (Cordova and Navarro 2016). Finally, the software is written using C/C++ (with Python extensions using Cython, https://cython.org/) and builds upon the foundation established by Striped UniFrac (McDonald et al. 2018b). Reuse of this library facilitated our access to a much faster Newick format parser, which reduces the overhead when reading a tree from disk. These

147 factors make for an improved expected and in-practice performance, despite the time complexity
148 and worst-case memory complexity remaining the same.

149 To demonstrate the scalability of SFPhD, we used a collection of 307,237 public and
150 anonymized private 16S rRNA V4 microbiome samples amounting to 1,264,796 phylogenetic
151 tree tips (after rarefaction at 500 sequences per sample). The samples were retrieved using the
152 redbiom command line interface (McDonald et al. 2019) which queried a cache of public and
153 anonymized private studies available in Qiita (Gonzalez et al. 2018). Amplicon sequence
154 variants (ASVs) were placed into the Greengenes (Gonzalez et al. 2018; McDonald et al. 2012b;
155 DeSantis et al. 2006) phylogeny using SEPP (Mirarab et al. 2012). Computing the full alpha
156 diversity vector took SFPhD 1 hour and 5 minutes wall-clock time and required a maximum
157 resident set size of less than 3 GB (see Methods for hardware details). In addition, we iteratively
158 measured runtime and memory consumption for increasingly large random subsets of samples
159 while fixing the size of the tree at 100,000 tips (**Fig. 2A, 2B, Table S1**). For the iteration with
160 20,000 samples, the memory usage of the reference implementation exceeded 150 GB and the
161 process ran for over 15 minutes. Contrastingly, with SFPhD, the process took 14 seconds to
162 execute and required less than 0.5 GB of memory. Additionally, using Green Algorithms
163 (Lannelongue et al. 2021), we estimated the carbon footprint of the scikit-bio reference
164 implementation on the 20,000 sample table to be 12.84 g CO_2e, whereas we estimated the carbon
165 footprint of SFPhD would be 0.04 g CO_2e in the United States, which is a 321-fold reduction in
166 impact on global warming.

167 **Phylogenetic diversity is a suitable metric to analyze stool metagenomic samples**

168 To demonstrate SFPhD's versatility and applicability to newer datasets, we reanalyzed
169 2,661 paired 16S rRNA and metagenomic data of stool samples from the FINRISK (Borodulin et
170 al. 2015, 2018; Salosensaari et al. 2021) study (n=1,563 aged 60 and older, n=1,098 aged 35 and

under). In this experiment, we select random subsets of the full sample set and compare each metric's (Observed Features and Faith's PD) ability to detect differences in mean alpha diversity distributions. For each step we randomly select N paired 16S and metagenomic samples, and then compute the difference in mean alpha diversity between samples taken from younger adults (under 35 years) and older adults (over 60 years) together with an empirical p-value. For both 16S and metagenomics, the alpha diversity of younger adults is lower than in older adults. In metagenomics, but not in 16S sequencing, Faith's PD provides improved statistical power over observed features, a phylogenetically-agnostic alternative (**Fig. 3A, B**). With 16S data, the difference between the two metrics is subtle (**Fig. 3A**). In both cases, the statistical power increases as the number of samples grows. With metagenomic data, the number of observed features shows a weaker effect compared to Faith's PD regardless of the number of samples (**Fig. 3B**). Unlike 16S datasets (5,600 features), metagenomic datasets (1,700 features) are resolution-limited by the reference databases. Whereas the nature of amplicon sequence variants (ASVs) allows for a broader feature space that can capture age-differences without the need for a phylogeny.

We investigated the difference in mean alpha diversity in metagenomic samples (**Fig. 4A**) by computing the log of the likelihood ratio of *older to younger adult* samples present for each branch in the WoL phylogenomic tree (Zhu et al. 2019). We were able to identify portions of the WoL tree responsible for the increase in phylogenetic diversity (**Fig. 4B**). From this analysis, we found that the majority of the tree is comparably represented in young and old adult samples. However, we also found two clades where *older adult* samples were more prevalent than *younger adult* samples (Clade 1 has a log likelihood ratio bounded with an 80% confidence interval of [1.20, 1.45] and Clade 2 has an 80% confidence interval of [0.55, 0.74]). Clade 1 corresponds to a majority of Lactobacillales genomes, and Clade 2 corresponds to Proteobacteria

genomes. The branches in Clade 1 primarily have a large log likelihood ratio, indicating that the features across the entire clade are more likely to be found in samples from older adults. However, the internal branches in Clade 2 additionally have low log likelihood ratios, indicating that the enrichment of features in older adults is not completely consistent across the entire clade. Lastly, although not confined to a few clades, there are several tips (e.g. *Staphylococcus aureus*, *Bavariicoccus seileri*, *Nitratireductor indicus*, and *Campylobacter ureolyticus*) in the phylogeny that are only associated with younger adults.

Discussion

By accounting for the relationship between features in a dataset, Faith's PD can mitigate issues with sparsity and heterogeneity common to modern 'omics' datasets. Although this metric was first introduced 30 years ago, the underlying algorithm for computing this metric had largely remained unchanged. In this paper we demonstrated that our novel algorithm, SFPhD, performed efficiently on datasets with hundreds of thousands of samples and millions of tree tips, producing identical results to those of previous algorithms for computing this metric while producing a speedup of up to 64× and requiring as little as 0.21% of the memory in our benchmarks.

An important aspect of SFPhD's underlying algorithm is substituting calculation of the full presence/absence table over the phylogeny, for a tree traversal that partially aggregates diversity values and frees presence/absence information when no longer needed. The result is a high-performance implementation that demonstrates improved scaling with the number of samples in the input dataset. Much of the engineering work here was facilitated by the balanced parenthesis tree implementation provided in the UniFrac package (McDonald et al. 2018b). Therefore, we believe that increasing the availability of efficient and flexible data structures for phylogenetic analyses is likely to accelerate and facilitate the development of novel analytical

methods. In a broader sense, this is similar to the impact of NumPy's (McDonald et al. 2018b; Harris et al. 2020) N-dimensional array in image processing, machine learning, neuroscience, and other fields.

In addition, in a stool metagenomic study Faith's PD demonstrates increased statistical power compared to observed features for differentiating younger from older subjects based on their microbial communities. In this context, we show that Faith's PD consistently provided increased statistical power for determining age-based differences in the shotgun metagenomic sequencing data. While this metric was originally developed to analyze data with vastly different statistical and biological properties, its use here demonstrates the versatility and applicability behind measuring diversity using a tree. Furthermore, enabling efficient Faith's PD computation on microbiome datasets is of particular importance when examining the impact of COVID-19 on gut health (Kim et al. 2021).

Although we show the utility of SFPhD in large and complex microbiome studies, the underlying implementation is not tied to a particular molecular technology. Thus, this implementation will be relevant to fields outside of microbiology, such as conservation prioritization, which inspired the original version of Faith's PD (Faith 1992) and where it continues to be applied (Rosauer et al. 2017). We also envision our implementation will be applicable in fields like nutrition and metabolomics research, that only recently began adopting trees for analytical tasks (Tripathi et al. 2021; Johnson et al. 2019).

Methods

Construction of benchmarking tables

Data for the benchmarking in this study were subsampled from a BIOM table of 113,721 and 761,003 ASVs, which is composed of studies aggregated from several large sources of publicly available microbiome data in Qiita (Amir et al. 2017; Gonzalez et al. 2018). This data table was

produced as previously described (McDonald et al. 2018b). The data was subset by uniformly randomly sampling the desired number of ASVs and samples from the table. Ten different tables were created for each number of samples and ASVs. The published insertion tree (McDonald et al. 2018b) was collapsed to only contain sequences that were selected to be included in the given subsampled table.

The table with 307,237 public and anonymized private 16S rRNA V4 microbiome samples and 1,264,796 phylogenetic tree tips was also prepared as previously described (McDonald et al. 2018b), but included samples with private sequencing data from Qiita.

Benchmarking time and memory estimates

The SFPhD implementation available in the Python package unifrac v0.10.0 was used. The reference implementation uses the Faith's PD implementation from scikit-bio v0.5.4.

All methods were run single-threaded on shared compute nodes that were not running other compute tasks. The nodes all had Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz processors. A job was terminated if it exceeded 6 hours of wall time or 250 GB of memory (system max). Space was tracked using GNU Time. Time for both implementations was tracked with a Python wrapper script. The time needed to parse data is not included in the scikit-bio timings, but is included in the SFPhD timings, due to the lack of access to this information in the unifrac interface. This is acceptable given that it results in a conservative estimate of the speedup with SFPhD.

Carbon footprint estimation

The Green Algorithms interface (Lannelongue et al. 2021) was used to estimate the Carbon Dioxide equivalent (CO_2e) of the benchmarked methods. The Intel(R) Xeon(R) CPU E5-2640 v3 CPUs used in benchmarking have a Thermal Design Power (TDP) per core of the 11.25 TDP / core.

FINRISK processing

The 16S rRNA data were demultiplexed, quality filtered, and denoised with deblur (Amir et al. 2017). The Greengenes (McDonald et al. 2012b) 13.8 with a clustering level of 99% was used as the reference phylogeny for open-reference feature picking with SEPP (Mirarab et al.2012). ASVs with a total frequency fewer than 10 were discarded, and the table was then rarefied to a sampling depth of 1000 reads/sample. The resulting table and insertion tree were used for calculation of Faith's PD.

The shotgun metagenomic data were trimmed and quality filtered using Atropos (Didion et al.2017). They were aligned to the WoL database using SHOGUN pipeline (v1.0.8) with a Bowtie 2 alignment option. A table was generated from the alignments using the OGU workflow (Zhu et al. 2021). OGUs with a total frequency fewer than 10 were discarded, and the table was then rarefied to a sampling depth of 1000 reads/sample. The WoL phylogenomic tree (Zhu et al. 2021, 2019) was used for Faith's PD.

Both tables were filtered to include only samples from individuals 35 and younger (younger criteria) or 60 and older (older criteria).

Power estimation for mean difference in alpha diversity

For a given N (shown on horizontal axis in **Fig. 3a,b**), the FINRISK processed samples matching the younger/older criteria were sampled to this depth. On the subsampled data, the difference in mean alpha diversity between younger and older adults $\bar{d}$, was computed. A null distribution, $\hat{D}$, was generated by repeating 1000 repetitions of shuffling the age category associated with an alpha diversity and recomputing the difference of mean alpha diversity between the groups. The p-value was computed by finding the percentile of $\bar{d}$ in $\hat{D}$.

This test procedure was repeated for 1000 repetitions. The power for N is estimated as the proportion of tests found significant at $\alpha = 0.05$.

Older-younger log likelihood-ratio calculation

The WoL tree (Zhu et al. 2019) was pruned and filtered to only include the OGUs (Zhu et al. 2021) belonging to the FINRISK samples with age <= 35 and >= 60. For each node $t \in T$ in the tree,

$$logLikelihoodRatio_t = \log\left(\frac{|Samples_{older}(Descendants(t))|}{|Samples_{younger}(Descendants(t))|}\right) - \log\left(\frac{|Samples_{older}(T)|}{|Samples_{younger}(T)|}\right)$$

where $Descendants(t)$ is the set of descendants of t in T, and for a set of nodes N, $Samples_{group}(N)$ is the set of samples that contain any features in N.

Phylogenetic Visualization

Tree was visualized using EMPress (Cantrell et al. 2021). A node in the tree was considered old if its $age_{log} > 0$ and young if its $age_{log} < 0$.

Software Availability

The data used for benchmarking Faith's PD timing and memory usage are available as per the Striped UniFrac paper (McDonald et al. 2018b). The code for the benchmarking is available on GitHub (https://github.com/biocore/faiths-pd-benchmarking). The data and code needed for benchmarking the FINRISK metagenomics data are also available on GitHub. The SFPhD code is available in the unifrac Python package (https://github.com/biocore/unifrac). All of the software is also available in the **Supplemental Code**.

Competing Interests

We declare that we have no competing interests.

Acknowledgements

This work was supported in part by IBM Research AI through the AI Horizons Network, the Center for Microbiome Innovation at UC San Diego, the Academy of Finland grant 321351 and the Emil Aaltonen Foundation (to T.N.), the National Institutes of Health grant R01ES027595 (to M.J.), the Academy of Finland grants 321356 and 335525 (A.S.H), the Academy of Finland grant 295741 (L.L.). MI was supported by the Munz Chair of Cardiovascular Prediction and Prevention. VS was supported by the Finnish Foundation for Cardiovascular Research.

Figure Legends

Figure 1. Partially aggregating branch lengths reduces the space complexity of the algorithm. a) Faith's PD calculation depends on the representation of features present in samples. In the table, the letters (R, O, B, K) represent samples and the numbers (0, 1, 2, 4, 6, 9, 10) represent features. A 1 in an entry indicates the presence of a feature in the sample. SFPhD uses sparse table data structures, which reduce memory by only keeping track of the non-zero values in a matrix (highlighted in gray). b) A mock reference phylogenetic tree is shown, with the features from (a) as tips. Labels for the samples from (a) are located next to tips that they contain. The nodes are labeled by their order in a post-order traversal of the tree. c) Graphic depiction of the reference implementation's calculation of Faith's PD by first aggregating the presence/absence information for each branch in the tree, followed by multiplication by the branch lengths to get the metric constituents, and finally a sum over the entire branch ×metric constituent table. d) Graphic representation of the execution of SFPhD. On the left, the stack of presence/absence information is shown at three points during the algorithm's execution (i, ii, ii). Each of these times shows the stack immediately before memory is freed. On the right, the state of the partially aggregated phylogenetic diversity (PD) is shown after each node is added to the

stack. Each row represents the vector after a step in the algorithm. In practice, there is only one such vector. e) The balanced parentheses representation for the phylogenetic tree from (b).

Figure 2. SFPhD outperforms the reference implementation in terms of runtime and memory usage. a) Runtime in seconds for computing Faith's PD on datasets with thousands of samples and 100,000 tips in the phylogeny. Data is independently sub-sampled from a collection of 113,721 public samples in Qiita (Zhu et al. 2019; Gonzalez et al. 2018) as previously processed (McDonald, Vázquez-Baeza, et al. 2018). Mean of n=10 repetitions with 95% CI error bars. b) Memory usage for the same experiment as in (a). For both a and b jobs were terminated if they exceeded 250 GB of memory.

Figure 3. Phylogenetic diversity provides increased statistical power to differentiate age groups in shotgun metagenomics but not in 16S rRNA sequencing. a) Statistical power to differentiate young adults from old adults in two alpha diversity metrics at different sample sizes using 16S rRNA sequencing in the FINRISK cohort. b) Same as (a) but for shallow shotgun metagenomic sequencing.

Figure 4. Phylogenetic tree colored by age-group log of the likelihood ratio of older to younger adults per node. a) Distribution of Faith's PD by age group on the full dataset. b) Web of Life (WoL) Phylogenetic tree with branches colored by the log of likelihood ratio of old adults compared to young adults in descendants of the branch, for the FINRISK dataset. The inner circle is colored by the log of likelihood ratio of older adults compared to younger adults in the tips of the tree. The outer circle is colored by the phylum of the taxon represented by each tree tip. Red ellipses mark two clades enriched for samples from older individuals.

References

Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2** (2). https://doi.org/10.1128/mSystems.00191-16.

Borodulin K, Tolonen H, Jousilahti P, Jula A, Juolevi A, Koskinen S, Kuulasmaa K, Laatikainen T, Mannisto S, Peltonen M et al. 2018. Cohort Profile: The National FINRISK Study. *Int J Epidemiol.* **47** (3): 696-696i. doi: 10.1093/ije/dyx239.

Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T, Männistö S, Salomaa V, Sundvall J, Puska P. 2015. Forty-Year Trends in Cardiovascular Risk Factors in Finland. *Eur J Public Health.* **25** (3): 539–46. doi: 10.1093/eurpub/cku174.

Cantrell K, Fedarko MW, Rahman G, McDonald D, Yang Y, Zaw T, Gonzalez A, et al. 2021. EMPress Enables Tree-Guided, Interactive, and Exploratory Analyses of Multi-Omic Data Sets. *mSystems* **6** (2). doi: 10.1128/mSystems.01216-20.

Cordova J, and Navarro G. 2016. Simple and Efficient Fully-Functional Succinct Trees. *Theor*. doi: 10.1016/j.tcs.2016.04.031.

de la Cuesta-Zuluaga J, Kelley ST, Chen Y, Escobar JS,Mueller NT, Ley RE, McDonald D, Huang S, Swafford AD, Knight R, Thackray VG. 2019. Age- and Sex-Dependent Patterns of Gut Microbial Diversity in Human Adults. *mSystems* **4** (4). doi: 10.1128/mSystems.00261-19.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol*. **72** (7): 5069–72. doi: 10.1128/AEM.03006-05.

Didion JP, Martin M, Collins FS. 2017. Atropos: Specific, Sensitive, and Speedy Trimming of Sequencing Reads. *PeerJ*. **5**: e3720. doi: 10.7717/peerj.3720.

Faith DP. 1992. Conservation Evaluation and Phylogenetic Diversity. *Biol Conserv* 61 (1): 1–10.

Faith DP, Lozupone CA, Nipperess D, and Knight R. 2009. The Cladistic Basis for the Phylogenetic Diversity (PD) Measure Links Evolutionary Features to Environmental Gradients and Supports Broad Applications of Microbial Ecology's 'Phylogenetic Beta Diversity' Framework. *Int J Mol Sci* **10** (11). doi: 10.3390/ijms10114723.

Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M et al. 2014. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host Microbe* **15** (3): 382–92. doi: 10.1016/j.chom.2014.02.005.

Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB et al. 2018. Qiita: Rapid, Web-Enabled Microbiome Meta-Analysis. *Nat Methods* **15** (10): 796–98. doi: 10.1038/s41592-018-0141-9.

Grüning B, The Bioconda Team, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J. 2018. Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences. *Nat Methods*. doi: 10.1038/s41592-018-0046-7.

Hamady M, Lozupone C, Knight R. 2010. Fast UniFrac: Facilitating High-Throughput Phylogenetic Analyses of Microbial Communities Including Analysis of Pyrosequencing and PhyloChip Data. *ISME J* **4** (1): 17–27. doi: 10.1038/ismej.2009.97.

Harris CR., Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ et al. 2020. Array Programming with NumPy. *Nature* **585** (7825): 357–62. doi: 10.1038/s41586-020-2649-2.

Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q,Gohl DM, Beckman KB, Knight R, Knights D. 2018. Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems* **3** (6). doi: 10.1128/mSystems.00069-18.

Jeffery IB., Lynch DB, O'Toole PW. 2016. Composition and Temporal Stability of the Gut Microbiota in Older Persons." *ISME J* **10** (1): 170–82.

Johnson AJ., Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, Kim AD, Shmagel AK, Syed AN, Personalized Microbiome Class Students et al. 2019. Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host Microbe* **25** (6): 789–802.e5.

Kaplan RC., Wang Z, Usyk M, Sotres-Alvarez D, Daviglus ML, Schneiderman N, Talavera GA, Gellman MD, Thyagarajan B, Moon JY et al. 2019. Gut Microbiome Composition in the Hispanic Community Health Study/Study of Latinos Is Shaped by Geographic Relocation, Environmental Factors, and Obesity. *Genome Biology* **20**. doi: 10.1186/s13059-019-1831-z.

Kim HN , Joo EJ, Lee CW, Ahn KS, Kim HL, Park DI, Park SK. 2021. Reversion of Gut Microbiota during the Recovery Phase in Patients with Asymptomatic or Mild COVID-19: Longitudinal Study. *Microorganisms* **9** (6): 1237. doi: 10.3390/microorganisms9061237.

Kumar MS, Slud EV, Okrah K, Hicks SC, HannenhalliS, Bravo HC. 2018. Analysis and Correction of Compositional Bias in Sparse Sequencing Count Data. *BMC Genomics* **19** (1): 799.

Lannelongue L, Grealey J, Inouye M. 2021. Green Algorithms: Quantifying the Carbon Footprint of Computation. *Adv Sci* **2100707**. doi: 10.1002/advs.202100707.

Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Appl Environ*

Microbiol. doi: 10.1128/aem.00335-09.

Lozupone C, Knight R. 2005. UniFrac: A New Phylogenetic Method for Comparing Microbial Communities. *ApplEnviron Microbiol* **71** (12): 8228–35.

Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, Zengler K. 2019. A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* **4** (1). doi: 10.1128/mSystems.00016-19.

McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F et al. 2012a. The Biological Observation Matrix (BIOM) Format or: How I Learned to Stop Worrying and Love the Ome-Ome. *GigaScience* **1** (1): 7. doi: 10.1186/2047-217X-1-7.

McDonald D , Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y et al. 2018a. American Gut: An Open Platform for Citizen Science Microbiome Research. *mSystems* **3** (3). doi: 10.1128/mSystems.00031-18.

McDonald D, Kaehler B, Gonzalez A, DeReus J, Ackermann G, Marotz C, Huttley G, Knight R. 2019. Redbiom: A Rapid Sample Discovery and Feature Characterization System. *mSystems* **4** (4). doi: 10.1128/mSystems.00215-19.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012b. An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea. *ISME J* **6** (3): 610–18.

McDonald D, Vázquez-Baeza Y, Koslicki D, McClelland J, Reeve N, Xu Z, Gonzalez A, Knight R. 2018b. Striped UniFrac: Enabling Microbiome Analysis at Unprecedented Scale. *Nat Methods* **15** (11): 847–48.

Mirarab S, Nguyen N, Warnow T. 2012. SEPP: SATé-Enabled Phylogenetic Placement. *Pac Symp*, 247–58. doi: 10.1142/9789814366496_0024.

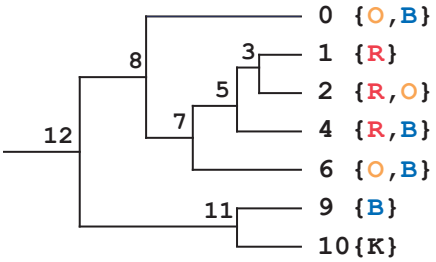
450 Morton JT, Toran L, Edlund A, Metcalf JL, Lauber C, Knight R. 2017. Uncovering the
451 Horseshoe Effect in Microbial Analyses. *mSystems* **2** (1). doi:10.1128/mSystems.00166-16.
452 Parks DH, Chuvochina M, Chaumeil PA, Rinke C,Mussig AJ, Hugenholtz P. 2020. A Complete
453 Domain-to-Species Taxonomy for Bacteria and Archaea. *Nat Biotechnology* **38** (9): 1079–
454 86.
455 Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P.
456 2018. A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially
457 Revises the Tree of Life. *Nat Biotechnology* **36** (10): 996–1004.
458 Rosauer DF , Pollock LJ, Linke S, Jetz W. 2017. Phylogenetically Informed Spatial Planning Is
459 Required to Conserve the Mammalian Tree of Life. *Proc R Soc. B* **284** (20170627).
460 doi:10.1098/rspb.2017.0627.
461 Salosensaari A , Laitinen V, Havulinna AS, Meric G, Cheng S, Perola M, Valsta L, Alfthan G,
462 Inouye M, Watrous JD et al. 2021. Taxonomic Signatures of Cause-Specific Mortality Risk
463 in Human Gut Microbiome. *Nat Communications* **12** (1): 2671.
464 Scherson RA, Faith DP. 2018. In *Phylogenetic Diversity: Applications and Challenges in*
465 *Biodiversity Science*. Springer International Publishing, New York, New York.
466 The Human Microbiome Project Consortium. 2012. Structure, Function and Diversity of the
467 Healthy Human Microbiome. *Nature*. doi:10.1038/nature11234.
468 Thompson LR , Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A,
469 Gibbons SM, Ackerman G et al. 2017. A Communal Catalogue Reveals Earth's Multiscale
470 Microbial Diversity. *Nature* **551** (7681): 457–63.
471 Tripathi A , Vázquez-Baeza Y, Gauglitz JM, Wang M, Dührkop K, Nothias-Esposito M,
472 Acharya DD, Ernst M, van der Hooft JJJ, Zhu Q et al. 2021. Chemically Informed Analyses
473 of Metabolomics Mass Spectrometry Data with Qemistree. *Nat Chem Biol* **17** (2): 146–51.

474 Vázquez-Baeza Y, Hyde ER, Suchodolski JS, Knight R. 2016. Dog and Human Inflammatory
475 Bowel Disease Rely on Overlapping yet Distinct Dysbiosis Networks. *Nat Microbiol*. doi:
476 10.1038/nmicrobiol.2016.177.
477 Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M,
478 Hidalgo G, Baldassano RN Anokhin AP et al. 2012. Human Gut Microbiome Viewed across
479 Age and Geography. *Nature* **486** (7402): 222–27.
480 Youngblut ND, de la Cuesta-Zuluaga J, Ley RE. 2021. Incorporating Genome-Based Phylogeny
481 and Functional Similarity into Diversity Assessments Helps to Resolve a Global Collection
482 of Human Gut Metagenomes. *bioRxiv*. doi: 10.1101/2020.07.16.207845.
483 Youngblut ND, Reischer GH, Walters W, Schuster N, Walzer C, Stalder G, Ley RE, Farnleitner
484 AH. 2019. Host Diet and Evolutionary History Explain Different Aspects of Gut
485 Microbiome Diversity among Vertebrate Clades. *Nat Commun* **10** (1): 2200.
486 Zhou J, Deng Y, Shen L, Wen C, Yan Q, Ning D, Qin Y, Xue K, Wu L, He Z et al. 2016.
487 Temperature Mediates Continental-Scale Diversity of Microbes in Forest Soils. *Nat*
488 *Commun* **7** (July): 12083.
489 Zhu Q, Huang Q, Gonzalez A, McGrath I, McDonald D, Haiminen N, Armstrong G, Vazquez-
490 Baeza Y, Yu J, Kuczynski J et al. 2021. OGUs Enable Effective, Phylogeny-Aware Analysis
491 of Even Shallow Metagenome Community Structures. *bioRxiv*. doi:
492 10.1101/2021.04.04.438427.
493 Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA,
494 Kopylova E, McDonald D et al. 2019. Phylogenomics of 10,575 genomes reveals
495 evolutionary proximity between domains Bacteria and Archaea. *Nat Commun* **10** (1): 5477.

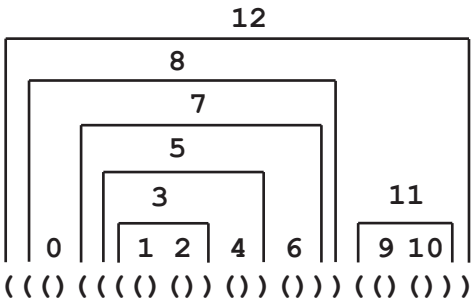
A Feature Table

	0	1	2	4	6	9	10
R	0	1	1	1	0	0	0
O	1	0	1	0	1	0	0
B	1	0	0	1	1	1	0
K	0	0	0	0	0	0	1

B Phylogenetic Tree



E Balanced Parentheses



C Reference Implementation

	Branch	R	O	B	K	Length	R*L	O*L	B*L	K*L
Tips	0	0	1	1	0	0.7	0	0.7	0.7	0
	1	1	0	0	0	0.2	0.2	0	0	0
	2	1	1	0	0	0.2	0.2	0.2	0	0
	4	1	0	1	0	0.3	0.3	0	0.3	0
	6	0	1	1	0	0.5	0	0.5	0.5	0
	9	0	0	1	0	0.3	0	0	0.3	0
	10	0	0	0	1	0.3	0	0	0	0.3
Internal Nodes	3	1	1	0	0	0.1	0.1	0.1	0	0
	5	1	1	1	0	0.2	0.2	0.2	0.2	0
	7	1	1	1	0	0.2	0.2	0.2	0.2	0
	8	1	1	1	0	0.3	0.3	0.3	0.3	0
	11	0	0	1	1	0.7	0	0	0.7	0.7
	12	1	1	1	1	0.3	0.3	0.3	0.3	0.3

Final Faith's PD = sum = 1.8 2.5 3.5 1.3

D Stacked Faith's PD Implementation

	Branch	R	O	B	K	Length	Partially Aggregated PD			
	0	0	1	1	0	0.7	0.0	0.7	0.7	0.0
i.	1	1	0	0	0	0.2	0.2	0.7	0.7	0.0
	2	1	1	0	0	0.2	0.4	0.9	0.7	0.0
	0	0	1	1	0	0.7				
ii.	3	1	1	0	0	0.1	0.5	1.0	0.7	0.0
	4	1	0	1	0	0.3	0.8	1.0	1.0	0.0
	0	0	1	1	0	0.7				
iii.	5	1	1	1	0	0.2	1.0	1.2	1.2	0.0
	6	0	1	1	0	0.5	1.0	1.7	1.7	0.0

Final Faith's PD = 1.8 2.5 3.5 1.3