Title

EMPress Enables Tree-Guided, Interactive, and Exploratory Analyses of Multi-omic Data Sets.

Permalink

https://escholarship.org/uc/item/8fw922f0

Journal

mSystems, 6(2)

ISSN

2379-5077

Authors

Cantrell, Kalen
Fedarko, Marcus W
Rahman, Gibraan
et al.

Publication Date

2021-03-01

DOI

10.1128/msystems.01216-20

Peer reviewed

AMERICAN SOCIETY FOR MICROBIOLOGY mSystems®

RESEARCH ARTICLE
Applied and Environmental Science

EMPress Enables Tree-Guided, Interactive, and Exploratory Analyses of Multi-omic Data Sets

Kalen Cantrell,[a,b] Marcus W. Fedarko,[a,b] Gibraan Rahman,[c,d] Daniel McDonald,[d] Yimeng Yang,[b] Thant Zaw,[b] Antonio Gonzalez,[d] Stefan Janssen,[e] Mehrbod Estaki,[d] Niina Haiminen,[f] Kristen L. Beck,[g] Qiyun Zhu,[d]* Erfan Sayyari,[b,h] James T. Morton,[i] George Armstrong,[b,c,d] Anupriya Tripathi,[d] Julia M. Gauglitz,[n] Clarisse Marotz,[d,j] Nathaniel L. Matteson,[l] Cameron Martino,[b,c,d] Jon G. Sanders,[o] Anna Paola Carrieri,[m] Se Jin Song,[b] Austin D. Swafford,[b] Pieter C. Dorrestein,[b,n] Kristian G. Andersen,[l] Laxmi Parida,[f] Ho-Cheol Kim,[g] Yoshiki Vázquez-Baeza,[b] Rob Knight[a,b,d,k]

[a]Department of Computer Science, Jacobs School of Engineering, University of California, San Diego, California, USA
[b]Center for Microbiome Innovation, Jacobs School of Engineering, University of California, San Diego, California, USA
[c]Bioinformatics and Systems Biology Program, University of California, San Diego, California, USA
[d]Department of Pediatrics, School of Medicine, University of California, San Diego, California, USA
[e]Algorithmic Bioinformatics, Justus Liebig University, Giessen, Germany
[f]IBM T. J. Watson Research Center, Yorktown Heights, New York, USA
[g]IBM Almaden Research Center, San Jose, California, USA
[h]Department of Electrical and Computer Engineering, University of California, San Diego, California, USA
[i]Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, New York, USA
[j]Scripps Institution of Oceanography, University of California, San Diego, California, USA
[k]Department of Bioengineering, University of California, San Diego, California, USA
[l]Scripps Research Institute, San Diego, California, USA
[m]IBM Research, The Hartree Centre, Daresbury, United Kingdom
[n]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, California, USA
[o]Cornell Institute for Host-Microbe Interaction and Disease, Cornell University, Ithaca, New York, USA

Kalen Cantrell and Marcus W. Fedarko contributed equally to this work. Author order was determined on the basis of project seniority.

ABSTRACT Standard workflows for analyzing microbiomes often include the creation and curation of phylogenetic trees. Here we present EMPress, an interactive web tool for visualizing trees in the context of microbiome, metabolome, and other community data scalable to trees with well over 500,000 nodes. EMPress provides novel functionality—including ordination integration and animations—alongside many standard tree visualization features and thus simplifies exploratory analyses of many forms of ‘omic data.

IMPORTANCE Phylogenetic trees are integral data structures for the analysis of microbial communities. Recent work has also shown the utility of trees constructed from certain metabolomic data sets, further highlighting their importance in microbiome research. The ever-growing scale of modern microbiome surveys has led to numerous challenges in visualizing these data. In this paper we used five diverse data sets to showcase the versatility and scalability of EMPress, an interactive web visualization tool. EMPress addresses the growing need for exploratory analysis tools that can accommodate large, complex multi-omic data sets.

KEYWORDS bioinformatics, microbial ecology

The increased availability of sequencing technologies and automation of molecular methods have enabled studies of unprecedented scale (1), prompting the creation of tools better suited to store, analyze (2), and visualize (3) studies of this magnitude. Many of these tools, including for example UniFrac (4), phylofactor (5), PhILR (6), and Gneiss (7), make use of tree structures in some way: often these structures are phylogenetic trees

Citation Cantrell K, Fedarko MW, Rahman G, McDonald D, Yang Y, Zaw T, Gonzalez A, Janssen S, Estaki M, Haiminen N, Beck KL, Zhu Q, Sayyari E, Morton JT, Armstrong G, Tripathi A, Gauglitz JM, Marotz C, Matteson NL, Martino C, Sanders JG, Carrieri AP, Song SJ, Swafford AD, Dorrestein PC, Andersen KG, Parida L, Kim H-C, Vázquez-Baeza Y, Knight R. 2021. EMPress enables tree-guided, interactive, and exploratory analyses of multi-omic data sets. mSystems 6:e01216-20. https://doi.org/10.1128/mSystems.01216-20.

Editor Laura A. Hug, University of Waterloo

Address correspondence to Rob Knight, robknight@ucsd.edu.

* Present address: Qiyun Zhu, Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, Arizona, USA.

Received 23 November 2020
Accepted 21 February 2021
Published 16 March 2021

detailing the evolutionary relationships among features in a data set, although this category also includes general dendrograms that organize features in a hierarchical structure (e.g., clustering of mass spectra) (8). The challenge of enabling fully interactive analyses stems from the disconnect between tools that focus on features (for example, microbial relative abundances) and tools that focus on samples (for example, alpha diversity distributions). In addition, few tools can interactively integrate multiple representations of the data side-by-side (9) while scaling to display large data sets. We view this as a key unresolved challenge for the field: to contextualize community-level patterns (groupings of samples) together with feature-level structure, i.e., which features lead to the groupings explained in a given sample set.

Although many useful phylogenetic visualization and analysis tools are available, few focus on community analysis tasks. The current state of the art includes specialized tools like Anvi'o (10), which consolidates a large collection of methods for sequence-based analysis and visualization of metagenomic assembled-genomes, pangenomes, and proteins (among many other data types). The state of the art also includes more general-purpose tree visualization tools like PHYLOViZ (9), SigTree (11), and iTOL (12) (among many others). Although tree structures are usually stored in standard file formats like Newick, the data accompanying these trees—for example, tip-level taxonomic classifications or other metadata values—are less standardized and sometimes require the onerous creation of configuration files. Furthermore, some types of exploratory analyses are not easily possible: for example, ordination plots computed from UniFrac (4) distances (or other phylogenetically informed distances) are often used to visualize sample clustering patterns in microbiome studies. However, interpreting the patterns in these plots—and determining which features influence the separation of certain groups of samples—is not always straightforward. While biplots can improve legibility by showing information about influential features alongside samples, the phylogenetic relationships of these features are not always obvious.

To address these and other outstanding gaps in the state of the art, we introduce EMPress (https://github.com/biocore/empress), an open-source (BSD 3-clause), interactive and scalable phylogenetic tree viewer accessible as a QIIME 2 (2) plugin or as a standalone Python program. EMPress is built around the high-performance balanced parentheses tree data structure (13) and uses a hardware-accelerated WebGL-based rendering engine that allows EMPress to visualize trees with hundreds of thousands of nodes from within a laptop's web browser (see Materials and Methods). EMPress visualizations can be created solely from a tree, or users can optionally provide additional metadata files and a feature table to augment the visualization. Additionally, through integration with the widely-used EMPeror software (3), EMPress can simultaneously visualize a study's phylogenetic tree alongside an ordination plot of the samples in the data set (in what we colloquially term an "EMPire plot"). User actions in one visualization, such as selecting a group of samples in the ordination, update the other (in this case, highlighting the portions of the tree corresponding to these samples), providing context that would not be easily accessible with independent visualizations. This tight integration between displays streamlines several use-cases elaborated below that previously required manual investigation or writing custom scripts.

RESULTS

Rather than providing a programmatic interface for the procedural generation of styled phylogenetic trees (14, 15; FigTree [http://tree.bio.ed.ac.uk/software/figtree/]), EMPress provides an interactive environment to support exploratory feature- and sample-level tree-based analyses. One of the ways EMPress stands out is in its scalability in comparison to other web-based tree viewers: iTOL (12) claims trees with more than 10,000 tips to be "very large" (https://itol.embl.de/help.cgi), while EMPress readily supports trees with over hundreds of thousands of tips, as shown in Fig. 1. Many visualization customization options available in EMPeror (3), iTOL (12), and Anvi'o (10) are immediately accessible in EMPress' interface. Continuous metadata associated with the

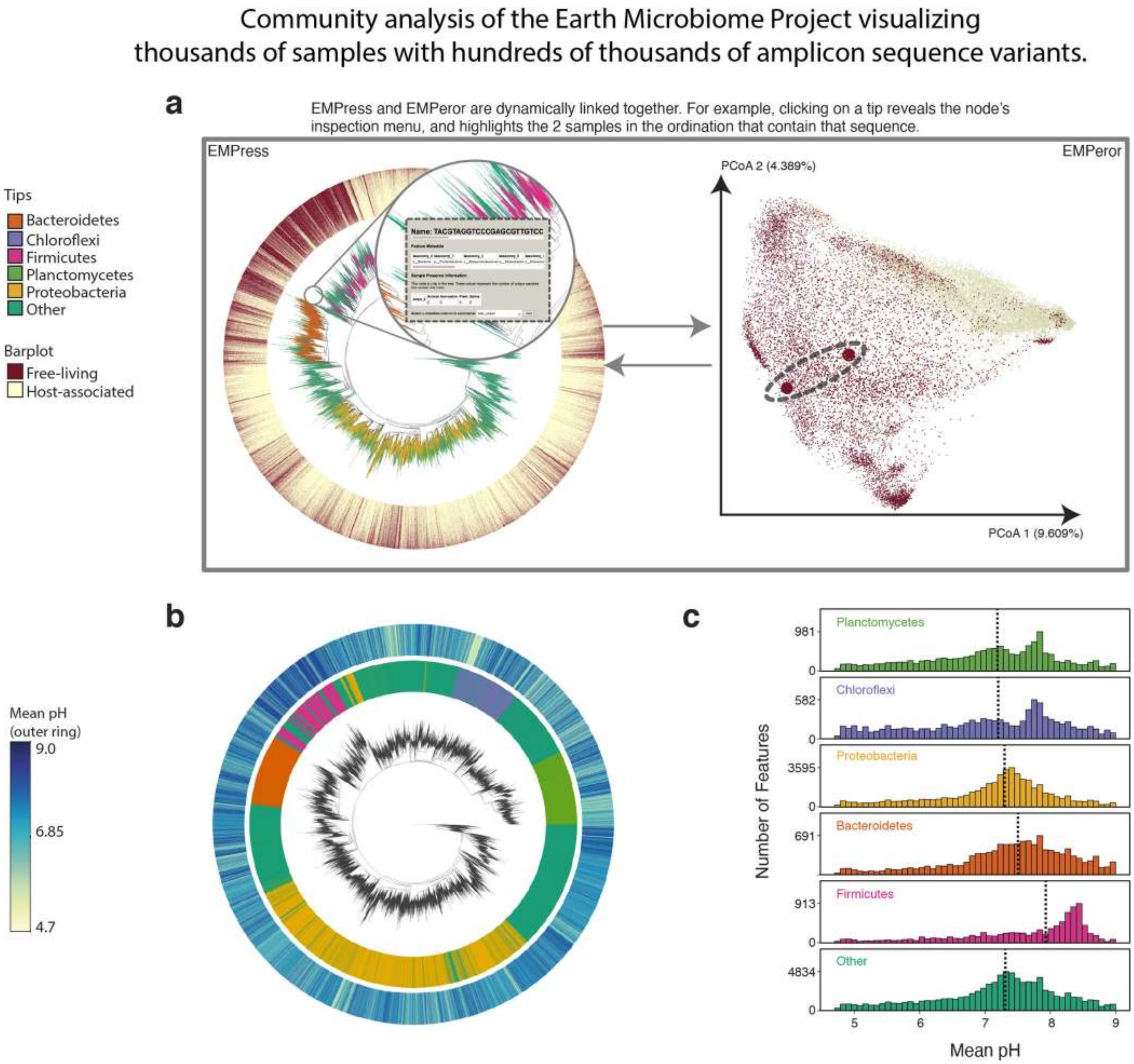

FIG 1 Earth Microbiome Project paired phylogenetic tree (including 756,377 nodes) and unweighted UniFrac ordination (including 26,035 samples). (a) Graphical depiction of EMPress' unified interface with fragment insertion tree (left) and unweighted UniFrac sample ordination (right). Tips are colored by their phylum-level taxonomic assignment; the barplot layer is a stacked barplot describing the proportions of samples containing each tip summarized by level 1 of the EMP ontology. Inset shows summarized sample information for a selected feature. The ordination highlights the two samples containing the tip selected in the tree enlarged to show their location. (b) Subset of EMP samples with pH information: the inner barplot ring shows the phylum-level taxonomic assignment, and the outer barplot ring represents the mean pH of all the samples where each tip was observed. (c) pH distributions summarized by phylum-level assignment with median pH indicated by dotted lines. Interactive figures can be accessed at https://github.com/knightlab-analyses/empress-analyses.

tips of the tree can be visualized as barplots with a color gradient and/or by mapping each value to the height of each bar. Similarly, categorical sample metadata information can be visualized using a stacked barplot showing—for each tip—the proportion of samples containing that tip stratified by category. These options are available in EMPress' user interface, based on the data provided by the user to EMPress when creating a visualization, and—providing data files are stored in an accepted format—do not require programming or the creation of configuration files.

EMPress also aids interpretation of ordination plots by optionally providing a unified interface where the tree and ordination visualizations are displayed side-by-side and "linked" through sample and feature identifiers (16). This combination of EMPress and EMPeror (3) allows for many novel exploratory data analysis tasks. For example, selecting a group of samples in the ordination highlights nodes in the tree present in those samples and vice versa (see Materials and Methods). This integration extends to biplots: clicking feature arrows in the ordination highlights the corresponding node in the tree. Lastly, EMPress supports the visualization of longitudinal studies by simultaneously showing tree nodes unique to groups of samples at each individual time point during an EMPeror animation (see Materials and Methods).

Scalability: visualizing data from the Earth Microbiome Project. Using the first data release of the Earth Microbiome Project (EMP), we demonstrate EMPress' scalability by rendering a 26,035-sample ordination and a 756,377-node tree (Fig. 1a). We also demonstrate EMPress' ability to annotate large tree visualizations with categorical "feature" (i.e., corresponding to nodes in the tree) metadata: to visualize the relative proportions of taxonomic groups at the phylum level, we use EMPress' feature metadata coloring to color tips in the tree by their phylum-level taxonomic classifications (see Materials and Methods). We extend this further by demonstrating EMPress' ability to visualize sample-level categorical metadata: we add a barplot layer showing, for each tip in the tree, the proportions of samples containing each tip summarized by level 1 of the EMP ontology ("Free-living" and "Host-associated").

The paired "EMPire plot" visualizations supported by EMPress' integration with EMPeror allow users to click on a tip in the tree (in EMPress) and view the samples that contain that feature in the ordination (in EMPeror). Clicking on an internal node in the tree functions similarly, showing all samples that contain any of the descendant tips of this node. These actions, and other functionality unique to these paired visualizations, are especially useful when analyzing data sets with outliers or mislabeled metadata. Figure 1a shows an example of this functionality in practice, in which the samples in which a tip in the tree is present are highlighted dynamically in an ordination.

EMPress' barplots can also be used to summarize environmental metadata: as a demonstration of this, Fig. 1b shows the subset of samples (4,002) with recorded pH information and a barplot layer with the mean pH where each feature was found. The barplot reveals a relatively dark section near many tips classified in the phylum *Firmicutes*; in concert with histograms showing mean pH for each phylum (Fig. 1c), we can confirm that *Firmicutes*-classified sequences are more commonly found in higher-pH environments. These and the other observations highlighted here indicate the utility of EMPress for exploratory analyses of large, complex data sets.

Versatility: visualizing diverse types of data. EMPress—both in the visualization of tree structures and in the visualization of various types of metadata alongside these structures—can be applied to many types of "'omic" data sets. To illustrate this versatility, we reanalyzed a COVID-19 metatranscriptome sequencing data set (17), a liquid chromatography-mass spectrometry (LC-MS) untargeted metabolomic food-associated data set (8), and a 16S rRNA gene sequencing oral microbiome data set (18). Despite the vastly different natures of these data sets, EMPress provides meaningful functionality for their analysis and visualization. Movie S1 in the supplemental material also shows a longitudinal exploratory analysis using EMPress and EMPeror representing a subset of SARS-CoV-2 genome data from GISAID. This paired visualization emphasizes the relationships in time and space among "community samples" and the convergence of locales in the United States with the outbreak in Italy (see Materials and Methods). The interactive nature of EMPress allows rapid visualization of strains observed in a collection of samples from different geographical locations.

Figure 2a showcases EMPress' ability to identify feature clusters that are differentially abundant in COVID-19 patients compared to community-acquired pneumonia patients and healthy controls (17). Clades showing KEGG enzyme code (EC) (19) annotations are collapsed at level two except for lyases, highlighting feature 4.1.1.20 (carboxy-lyase diaminopimelate decarboxylase) that was more abundant in COVID-19 here

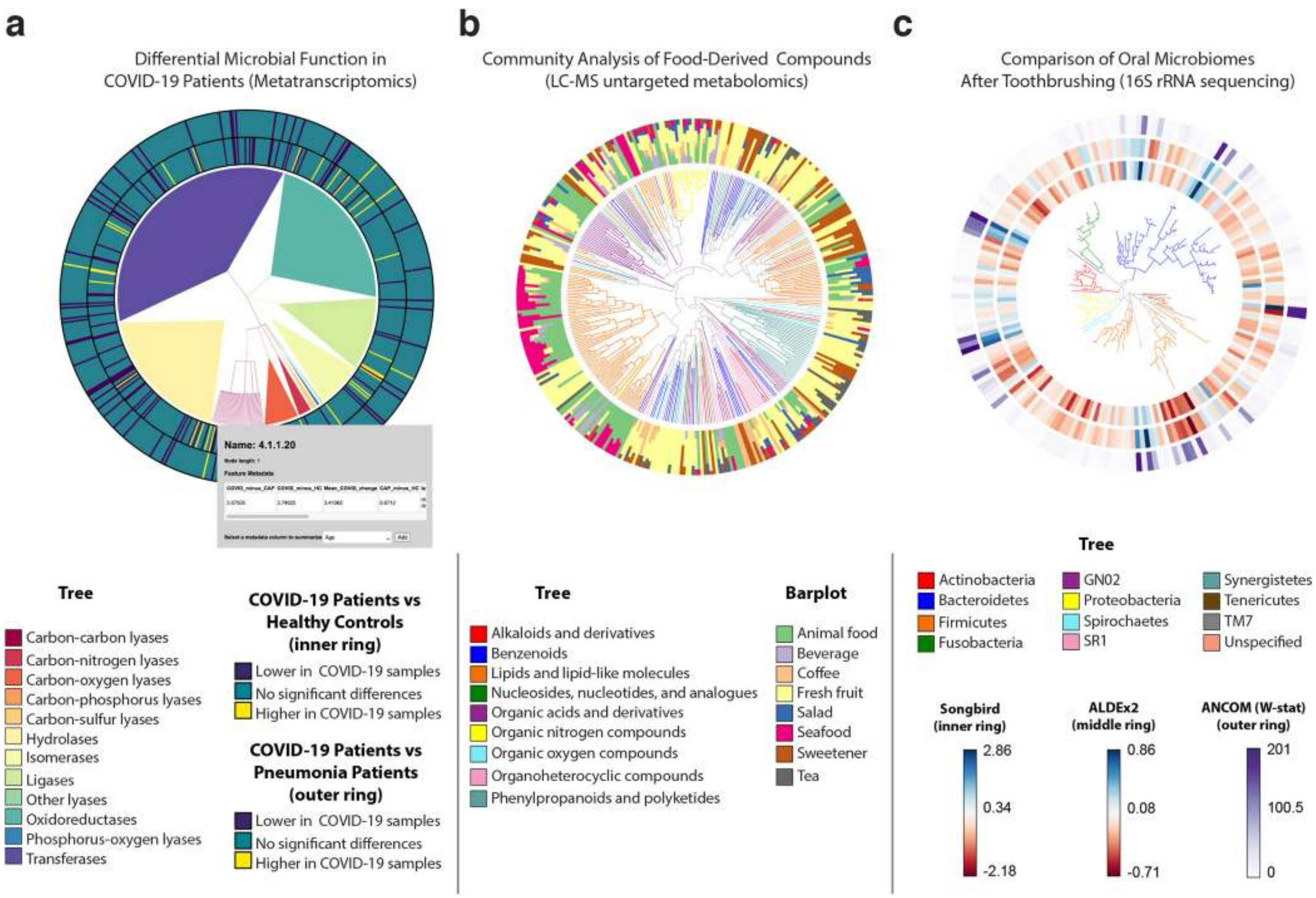

FIG 2 EMPress is a versatile exploratory analysis tool adaptable to various 'omics data types. (a) RoDEO differential abundance of microbial functions from metatranscriptomic sequencing of COVID-19 patients ($n = 8$) versus community-acquired pneumonia patients ($n = 25$) and versus healthy control subjects ($n = 20$). The tree represents the four-level hierarchy of the KEGG enzyme codes. The barplot depicts significantly differentially abundant features ($P < 0.05$) in COVID-19 patients. Clicking on a tip produces a pop-up insert tabulating the name of the feature, its hierarchical ranks, and any feature annotations. (b) Global FoodOmics Project LC-MS data. Stacked barplots indicate the proportions of samples ($n = 70$) (stratified by food) containing the tips in an LC-MS Qemistree of food-associated compounds, with tip nodes colored by their chemical superclass. (c) *De novo* tree constructed from 16S rRNA sequencing data from 32 oral microbiome samples. Samples were taken before ($n = 16$) and after ($n = 16$) subjects ($n = 10$) brushed their teeth; each barplot layer represents a different differential abundance method's measure of change between before- and after-brushing samples. The innermost layer shows estimated log-fold changes produced by Songbird, the middle layer shows effect sizes produced by ALDEx2, and the outermost layer shows the W-statistic values produced by ANCOM (see Materials and Methods). The tree is colored by tip nodes' phylum-level taxonomic classifications. Interactive figures can be accessed at https://github.com/knightlab-analyses/empress-analyses.

and in an independent metaproteomic analysis of COVID-19 respiratory microbiomes (20).

Recent developments in cheminformatics have enabled the analysis and visualization of small molecules in the context of a cladogram (8). Using a tree that links molecules by their structural relatedness, we analyzed untargeted LC-MS/MS data from 70 food samples (see Materials and Methods). With EMPress' sample metadata barplots, we can inspect the relationship between chemical annotations and food types. Figure 2b shows a tree where each tip is colored by its chemical superclass and where barplots show the proportion of samples in the study containing each compound by food type. This representation reveals a clade of lipids and lipid-like molecules that are well represented in animal food types and seafoods. In contrast, salads and fruits are broadly spread throughout the cladogram.

Lastly, in Fig. 2c, we compare three differential abundance methods in an oral microbiome data set (18) as separate barplot layers on a tree. This data set includes samples ($n = 32$) taken before and after subjects brushed their teeth (see Materials and Methods). As observed across the three differential abundance tools' outputs, all methods agree broadly on which features are particularly "differential" (for example, a group of sequences classified in the phylum *Firmicutes* in the bottom right of the tree; see Materials and Methods), although there are discrepancies due to different methods' assumptions and biases.

DISCUSSION

Future work on visualization integration. By providing an intuitive interface supporting both categorically new and established functionality, EMPress complements and extends the available range of tree visualization software. EMPress can perform community analyses across distinct "'omics" types, as demonstrated here. Moving forward, facilitating the integration of multiple orthogonal views of a data set at a more generalized framework level (for example, using QIIME 2's [2] visualization API) will be important as data sets continue to grow in complexity, size, and heterogeneity.

Validating visual observations and aiding reproducibility. It is important to note that making visual observations using EMPress does not eliminate the need for providing statistical support. For example, various layout and branch length options available in EMPress can drastically affect the perceived size of a clade or taxonomic group. We therefore recommend that users remain careful of the need to validate their claims, in order to ensure that conclusions drawn are not solely visual artifacts. In the context of microbial ecology studies, tools like Phylofactor, Gneiss, and PhILR can complement observations made with EMPress well.

Similarly, although the various exploratory (*post hoc*) analyses shown in this paper are simplified by EMPress, they are not a substitute for sound hypothesis-driven science and should not be presented as such (21). Exploratory analyses, when documented transparently, can be useful to the scientific community (21); our hope is that, by providing a tool that simplifies these analyses of complex data, EMPress can fulfill a legitimate scientific need. One way we believe EMPress helps fulfill this need is through the inherent shareability of its outputs (see Materials and Methods). This shareability simplifies the process of reproducing a visualization in EMPress, as well as the interactive exploration of alternative representations of a data set. As an example of this, we have provided QZV files for Fig. 1 and 2 on GitHub (see Materials and Methods), and we encourage readers to reproduce these figures for themselves. We acknowledge that "reproducibility" of this kind is limited to the files the user provides when creating an EMPress visualization—and since the same data and same methods are used in the reproduction as in the original visualization, this therefore does not necessarily add evidence that conclusions derived from the visualization are completely correct (22). However, we contend that it can still be beneficial for readers and authors alike, and we hope that by simplifying the sharing of EMPress visualizations we can encourage researchers to share their visualizations and make clear the exploratory nature of their work.

Tree shearing in the context of community data. In order to aid in the analysis of community data, EMPress—when a feature table is provided—by default shears the tree to include only tips that are found in the feature table. This preprocessing step visually emphasizes samples spanning only a small number of tips within larger reference trees. Since this option may not always be desired—if, for example, the focus of an analysis is to compare novel diversity to a reference database—this option can be disabled from the command line.

MATERIALS AND METHODS

All EMPress plots in this paper were visualized and stylized in Safari (14.0) or Google Chrome (85.0.4183.121) using a MacBook Pro (15-inch, 2017) with a 2.9-GHz Quad-Core Intel Core i7 7820QM, 16 GB of RAM, a Radeon Pro 560 4 GB, and Intel HD Graphics 630 1,536-MB graphics processor. Data, analyses, and steps to reproduce the figures in this paper can be found at https://github.com/knightlab-analyses/empress-analyses. Due to file size restrictions on GitHub, some files are downloaded by executing the Jupyter notebook associated with each figure.

Earth Microbiome Project. The EMP release 1 (1) table, tree, and metadata were used to generate the visualization (ftp://ftp.microbio.me/emp/release1). The original feature table was subset to remove sterile water blanks and mock community samples. The table contains the taxonomic assignments used to annotate tips. For ease of visualization, only the top 5 most abundant phylum annotations in the data set were kept while microbial features annotated with any other phylum (or unspecified) were categorized as "Other." A distance matrix was generated by computing the unweighted UniFrac distances between samples (4, 23) that was then used to generate the principal-coordinate plot.

A subset of the feature table was generated by extracting all samples in the middle 90% of the pH range (4.7 to 9) to remove outliers. Samples without a valid pH value were removed. For each remaining feature, the number of samples ($\log_{10}$) in which this feature occurred and the mean pH of those samples were calculated and saved as a feature metadata file passed into EMPress. Calculations were performed using NumPy v1.18.1 (24) and Pandas v0.25.3 (25, 26). Distributions of pH were plotted using matplotlib v3.1.3 (27) and seaborn v0.10.0 (28).

COVID-19 metatranscriptome data set. The COVID-19 bronchoalveolar lavage fluid metatranscriptome sequencing data (17) consists of COVID-19 ($n = 8$), community-acquired pneumonia ($n = 25$), and healthy control ($n = 20$) samples. The tree for this data set corresponds to the KEGG enzyme code (EC) (19) hierarchy. Sequencing reads were processed and annotated with EC feature labels using PRROMenade (29, 30) with a database of bacterial and viral protein domains from the IBM Functional Genomics Platform (31), as previously reported (30). Differential abundance per feature was determined by performing a Kolmogorov-Smirnov test on average RoDEO-processed (30, 32) values per sample. A cutoff ($P < 0.05$) was applied to focus the visualization on the features that are significantly more abundant or less abundant in COVID-19 patients than in healthy controls and/or community-acquired pneumonia samples.

Global FoodOmics data set. The untargeted metabolomics data set was generated using a quadrupole time of flight (QTOF) mass spectrometer in positive ionization mode (Bruker). The samples presented in this data set were processed using Qemistree version 2020.1.1+14.g1b4edb4 running in QIIME 2 version 2019.7. For ease of interpretation, the data set was subset to keep features with a superclass assignment and keep samples with a *common meal type* classification. The tip barplots show the proportion of samples where each small molecule is present summarized by *meal type*.

Differential abundance comparison of oral microbiomes. The oral microbiome 16S rRNA sequencing data used in reference 18 was revisualized for Fig. 2c. This data set comprises $n = 32$ samples total, taken before and after subjects brushed their teeth. Some paired samples were taken more than once from the same subject; in total, these 32 samples were contributed by 10 unique subjects.

The sequences in this data set were processed (in July 2018) using Deblur v1.0.4 (33) through q2-deblur in QIIME 2 2018.6 (2). Taxonomic classifications were assigned (in August 2018) using q2-feature-classifier's (34) classify-sklearn method (35), using the Greengenes reference database (36), also in QIIME 2 2018.6. The table provided lacks provenance information due to not being stored as a QIIME 2 artifact, but since its features are a subset of those in the sequences file—and since the lowest number of samples that a feature within it is present in is 6—it was likely filtered at some point. To construct a rooted tree from the sequences in this data set (in September 2020), we used QIIME 2 2019.10's qiime phylogeny align-to-tree-mafft-fasttree pipeline (37–39).

In reference 18, three differential abundance tools were run on this data set, using the "brushing_event" metadata field (indicating before/after toothbrushing status) as the sole field across which to identify differentially abundant features. The three differential abundance tools used in reference 18 and visualized in Fig. 2c are Songbird (18), ALDEx2 (40), and ANCOM (41). Songbird's column of feature differentials (describing the estimated log-fold changes of each feature between the "after" and "before" brushing states) is shown as the innermost barplot layer in Fig. 2c; ALDEx2's per-feature effect size is shown as the middle layer; and ANCOM's per-feature W-statistic is shown as the outermost layer. For both Songbird and ALDEx2 results, higher values indicate association with before-brushing samples (i.e., features that decreased most from toothbrushing, for example, secondary metabolizers present on the outer layers of dental plaque biofilms such as *Haemophilus*) while lower values indicate association with after-brushing samples (i.e., features that decreased least from toothbrushing, for example, primary metabolizers such as *Actinomyces* that are rooted at the base of the biofilm) (18). ANCOM's W-statistic corresponds to the number of log-ratio hypothesis tests in which a given feature was found to be differentially abundant between before- and after-brushing samples (41) (https://forum.qiime2.org/t/1844/10). Since Songbird and ALDEx2's results include directionality between before and after brushing, they are shown in Fig. 2c with a "diverging" color map; ANCOM's W-statistic does not include this information and is therefore shown with a "sequential" color map (see Fig. S1 in the supplemental material).

We note that the Songbird results from reference 18 did not include differentials for nine features in the data set; this may have been due to software bugs or other unknown factors in the data analysis, since (although Songbird does filter out features present in less than a given number of samples) these absent features are present in the same number of samples as other features which were included in the Songbird differentials. For the sake of simplicity here, and since the purpose of this subfigure is primarily to demonstrate the utility of EMPress in the context of existing data, we simply reused the data from reference 18, filtering these nine features out of the data set before constructing and visualizing the tree.

Animated analysis of SARS-CoV-2. The GISAID (42) SARS-CoV-2 genome alignment and genome metadata were obtained on 21 September 2020. Sequences were converted to DNA and subset to the set of sequences associated with Italy, Madrid, King County, San Diego, Brooklyn, Queens, and Manhattan. Highly gapped and high-entropy positions in the alignment were filtered using q2-alignment (2020.6; default parameters). A tree was estimated using FastTree (38) (v2.1.10 compiled with double precision support; default options except -fastest) and subsequently rooted using midpoint rooting as implemented by q2-phylogeny (2020.6; default parameters).

Separately, a sliding window procedure was developed to assess the observed SARS-CoV-2 genomes within a given time period within a geographic location. To do so, the metadata were partitioned into the respective locations (note that the three New York boroughs were treated as New York) and ordered by the genome date information. A sliding window width of 7 days was used, and a sample was retained only if five or more strains were observed within a window. These windows were then aggregated into a BIOM table (43) with the GISAID strain identifier on one axis and a "community sample" identifier on the other. Unweighted UniFrac (q2-diversity 2020.6 [2, 23]) was then computed over these samples followed by a principal-coordinate analysis. The tree and ordination were visualized with a development version of EMPress (version 0.3.0-dev), and therefore, the visualization shown in the video may look slightly different from more recent versions of EMPress.

Implementation details. EMPress is implemented as a QIIME 2 plugin (or standalone Python program, usable outside QIIME 2) capable of generating HTML documents with a self-contained visualization user interface. The code-base is composed of a Python component and a JavaScript component. The Python code-base is responsible for data validation, preprocessing, filtering, and formatting. User interaction, rendering, and figure generation are all handled by the JavaScript code-base. In both cases, we rely on the balanced parentheses data structure (13) to rapidly operate on the tree structures.

EMPress' Python code-base currently uses NumPy (24), SciPy (44), Pandas (25, 26), Click (https://palletsprojects.com/p/click/), Jinja2 (https://jinja.palletsprojects.com/), scikit-bio (http://scikit-bio.org), the BIOM format (43), iow (https://github.com/wasade/improved-octo-waddle) (13), and EMPeror (3). The JavaScript code-base uses Chroma.js (https://gka.github.io/chroma.js/), FileSaver.js (https://github.com/eligrey/FileSaver.js/), glMatrix (http://glmatrix.net/), jQuery (https://jquery.com/), Require.js (https://requirejs.org/), Spectrum (https://bgrins.github.io/spectrum/), and Underscore.js (https://underscorejs.org/). For testing and linting, EMPress' Python code-base uses flake8 (https://flake8.pycqa.org/en/latest/) and nose (https://nose.readthedocs.io/), and EMPress' JavaScript code-base uses QUnit (https://qunitjs.com/), qunit-puppeteer (https://github.com/davidtaylorhq/qunit-puppeteer/), jshint (https://jshint.com/about/), and Prettier (https://prettier.io/).

As of writing, EMPress supports drawing trees using three standard layout algorithms ("rectangular," "circular," and "unrooted"), coloring the tree using sample and feature metadata, collapsing clades based on common metadata values, adding tip-aligned barplots for sample and feature metadata, and summarizing tips' presence within sample groups using interactive node selections. This is in addition to integration with EMPeror, described further below, as well as various other visualization options.

EMPress' "unrooted" layout algorithm is translated from code from Gneiss (7), which was in turn adapted from PyCogent (45), and is an implementation of the equal-angle algorithm described in reference 46. EMPress' "rectangular" and "circular" layout algorithms are adapted from code from TopiaryExplorer (47) and resemble the rooted tree drawing algorithms described in reference 46. EMPress also includes the ability to reorder sibling clades in the rectangular and circular layouts by the number of tips contained within each clade; this functionality was inspired by iTOL's (12) "leaf sorting" option and uses tree traversal code adapted from scikit-bio (http://scikit-bio.org).

In order to integrate EMPress and EMPeror, we link together events triggered by each of the applications by inserting "callback" code that can be executed in one application when a given event occurs. These events notify each tool that a particular action needs to take place, and if needed, what data should be used in this context. For example, when a user selects a group of samples in EMPeror, the "select" event is triggered with a collection of sample objects. EMPress responds to this event by searching for the tips in the tree corresponding to features contained within these samples and updates the color according to the object's attributes. The subscription mechanism also enables users to select a node in EMPress to highlight the samples containing this node or one of its descendant tips in EMPeror, link biplot (48) arrows in EMPeror to nodes in the tree, highlight groups by double-clicking a category in EMPeror's color legend, and synchronize animated ordinations (49) by coloring the tree according to the current frame on screen.

Sharing EMPress visualizations. When used as a QIIME 2 (2) plugin, EMPress generates visualizations in the QZV format (which can be viewed using https://view.qiime2.org, in addition to other methods); when used outside QIIME 2, EMPress creates visualizations as a directory (containing an HTML file that can be opened to show the visualization). In either case, an EMPress visualization is easily shareable with a wide audience of users who may not have EMPress or QIIME 2 installed, for example, via uploading the visualization file(s) to GitHub or by hosting this file(s) on any other website. We note that the ability to share visualizations is not unique to EMPress and is also inherent to other QIIME 2 (2) plugins and in other tree visualization tools like iTOL (12).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, PDF file, 0.1 MB.

TABLE S1, XLSX file, 0.2 MB.

MOVIE S1, MOV file, 11.1 MB.

ACKNOWLEDGMENTS

We thank members of the Knight Lab and IBM AIHL Bioinformatics team for feedback during code reviews and presentations. We gratefully acknowledge the authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which a portion of this research is based, these authors are listed in Table S1 in the supplemental material.

This work is partly supported by IBM Research AI through the AI Horizons Network and the UC San Diego Center for Microbiome Innovation (to K.C., M.W.F., J.T.M., Q.Z., G. A., T.Z., J.G.S., A.D.S., S.J.S., Y.V.-B., and R.K.); CCF foundation no. 675191 (R.K. and P.C.D.), U19 AG063744 01 (R.K., J.M.G., and P.C.D.), CDC contract no. 75D30120C09795 (K.G.A. and R.K.), and U19 AI135995 (K.G.A.).

K.C., Q.Z., Y.Y., J.T.M., T.Z., J.G.S., and R.K. conceived the original idea for the project. K.C., M.W.F., G.R., D.M., A.G., S.J., M.E., Y.Y., E.S., J.T.M., G.A., T.Z., Q.Z., and Y.V.-B. wrote source code and/or documentation for the project. K.C., M.W.F., A.G., and Y.V.-B. wrote code to facilitate integration with EMPeror. L.P., H.-C.K., S.J.S., A.D.S., Y.V.-B., and R.K. managed the project. K.C., M.W.F., G.R., N.H., K.L.B., A.T., J.M.G., A.P.C., N.L.M., C.M., P.C.D., K.G.A., L.P., and Y.V.-B. analyzed and interpreted the data sets presented in this paper. K. C., M.W.F., G.R., D.M., N.H., K.L.B., and Y.V.-B. contributed text to Materials and Methods. All the authors contributed to the final version of the manuscript.

We declare no competing interests.

REFERENCES

1. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621.
2. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9.
3. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience 2:16. https://doi.org/10.1186/2047-217X-2-16.
4. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71:8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005.
5. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ 5:e2969. https://doi.org/10.7717/peerj.2969.
6. Silverman JD, Washburne AD, Mukherjee S, David LA. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. Elife 6:e21887. https://doi.org/10.7554/eLife.21887.
7. Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, Navas-Molina JA, Song SJ, Metcalf JL, Hyde ER, Lladser M, Dorrestein PC, Knight R. 2017. Balance trees reveal microbial niche differentiation. mSystems 2:e00162-16. https://doi.org/10.1128/mSystems.00162-16.
8. Tripathi A, Vázquez-Baeza Y, Gauglitz JM, Wang M, Dührkop K, Nothias-Esposito M, Acharya DD, Ernst M, van der Hooft JJJ, Zhu Q, McDonald D, Brejnrod AD, Gonzalez A, Handelsman J, Fleischauer M, Ludwig M, Böcker S, Nothias L-F, Knight R, Dorrestein PC. 2021. Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. Nat Chem Biol 17:146–151. https://doi.org/10.1038/s41589-020-00677-3.
9. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. 2017. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics 33:128–129. https://doi.org/10.1093/bioinformatics/btw582.
10. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319. https://doi.org/10.7717/peerj.1319.
11. Stevens JR, Jones TR, Lefevre M, Ganesan B, Weimer BC. 2017. SigTree: a microbial community analysis tool to identify and visualize significantly responsive branches in a phylogenetic tree. Comput Struct Biotechnol J 15:372–378. https://doi.org/10.1016/j.csbj.2017.06.002.
12. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.
13. Cordova J, Navarro G. 2016. Simple and efficient fully-functional succinct trees. Theor Comput Sci 656:135–145. https://doi.org/10.1016/j.tcs.2016.04.031.
14. Yu G. 2020. Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinformatics 69:e96. https://doi.org/10.1002/cpbi.96.
15. Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol 33:1635–1638. https://doi.org/10.1093/molbev/msw046.
16. Becker RA, Cleveland WS, Wilks AR. 1987. Dynamic graphics for data analysis. Stat Sci 2:355–383. https://doi.org/10.1214/ss/1177013104.
17. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, Zhou Z, Yang J, Zhong J, Yang D, Guo L, Zhang G, Li H, Xu Y, Chen M, Gao Z, Wang J, Ren L, Li M. 2020. Genomic diversity of severe acute respiratory syndrome–coronavirus 2 in patients with coronavirus disease 2019. Clin Infect Dis 71:713–720. https://doi.org/10.1093/cid/ciaa203.
18. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, Zengler K, Knight R. 2019. Establishing microbial composition measurement standards with reference frames. Nat Commun 10:2719. https://doi.org/10.1038/s41467-019-10656-5.
19. Kanehisa M. 2017. Enzyme annotation and metabolic reconstruction using KEGG. Methods Mol Biol 1611:135–145. https://doi.org/10.1007/978-1-4939-7015-5_11.
20. Maras JS, Sharma S, Bhat AR, Aggarwal R, Gupta E, Sarin SK. 2020. Multi-omics integration analysis of respiratory specimen characterizes baseline molecular determinants associated with COVID-19 diagnosis. medRxiv 2020.07.06.20147082. https://www.medrxiv.org/content/10.1101/2020.07.06.20147082v1.

21. Hollenbeck JR, Wright PM. 2017. Harking, sharking, and tharking: making the case for post hoc analysis of scientific data. J Manage 43:5–18. https://doi.org/10.1177/0149206316679487.
22. Schloss PD. 2018. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. mBio 9:e00525-18. https://doi.org/10.1128/mBio.00525-18.
23. McDonald D, Vázquez-Baeza Y, Koslicki D, McClelland J, Reeve N, Xu Z, Gonzalez A, Knight R. 2018. Striped UniFrac: enabling microbiome analysis at unprecedented scale. Nat Methods 15:847–848. https://doi.org/10.1038/s41592-018-0187-8.
24. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE. 2020. Array programming with NumPy. 7825. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2.
25. McKinney W. 2010. Data structures for statistical computing in Python, p 56–61. *In* Proc 9th Python Sci Conf (SCIPY 2010).
26. Reback J, McKinney W, Van den Bossche J, Augspurger T, Cloud P, Hawkins S, Klein A, Roeschke M, Tratner J, She C, Petersen T, Ayd W, Garcia M, Schendel J, Hayden A, Jancauskas V, Saxton D, McMaster A, Battiston P, Seabold S, Hoyer S, Dong K, Overmeire W, Winkel M. 2020. pandas-dev/pandas: Pandas 1.1.2. https://zenodo.org/record/4019559#.YDaEiGhKi8k.
27. Caswell TA, Droettboom M, Lee A, Hunter J, Firing E, Stansby D, Klymak J, Hoffmann T, de Andrade ES, Varoquaux N, Hedegaard Nielsen J, Root B, Elson P, May R, Dale D, Lee J-J, Seppänen JK, McDougall D, Straw A, Hobson P, Gohlke C, Yu TS, Ma E, Vincent AF, Silvester S, Moad C, Kniazev N, Ivanov P, Ernest E, Katins J. 2020. matplotlib/matplotlib v3.1.3. https://zenodo.org/record/3633844#.YDaFSGhKi8k.
28. Waskom M, Botvinnik O, Ostblom J, Lukauskas S, Hobson P, Gelbart M, Gemperline DC, Augspurger T, Halchenko Y, Cole JB, Warmenhoven J, de Ruiter J, Pye C, Hoyer S, Vanderplas J, Villalba S, Kunter G, Quintero E, Bachant P, Martin M, Meyer K, Swain C, Miles A, Brunner T, O'Kane D, Yarkoni T, Williams ML, Evans C. 2020. mwaskom/seaborn: v0.10.0 (January 2020). https://zenodo.org/record/3629446#.YDaIQGhKi8k.
29. Utro F, Haiminen N, Siragusa E, Gardiner L-J, Seabolt E, Krishna R, Kaufman JH, Parida L. 2020. Hierarchically labeled database indexing allows scalable characterization of microbiomes. iScience 23:100988. https://doi.org/10.1016/j.isci.2020.100988.
30. Haiminen N, Utro F, Seabolt E, Parida L. Functional profiling of COVID-19 respiratory tract microbiomes. Sci Rep, in press.
31. Seabolt E, Nayar G, Krishnareddy H, Agarwal A, Beck KL, Kandogan E, Kuntomi M, Roth M, Terrizzano I, Kaufman J, Mukherjee V. 2020. IBM functional genomics platform, a cloud-based platform for studying microbial life at scale. IEEE/ACM Trans Comput Biol Bioinform https://doi.org/10.1109/TCBB.2020.3021231.
32. Haiminen N, Klaas M, Zhou Z, Utro F, Cormican P, Didion T, Jensen CS, Mason CE, Barth S, Parida L. 2014. Comparative exomics of *Phalaris* cultivars under salt stress. BMC Genomics 15:S18. https://doi.org/10.1186/1471-2164-15-S6-S18.
33. 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:e00191-16. https://doi.org/10.1128/mSystems.00191-16.
34. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6:90. https://doi.org/10.1186/s40168-018-0470-z.
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830.
36. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618. https://doi.org/10.1038/ismej.2011.139.
37. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
38. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
39. Lane D. 1991. 16S/23S rRNA sequencing, p 115–175. In Stackebrandt E, Goodfellow M (ed), Nucleic acid techniques in bacterial systematics. John Wiley and Sons, New York, NY.
40. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2:15. https://doi.org/10.1186/2049-2618-2-15.
41. Mandal S, Treuren WV, White RA, Eggesbø M, Knight R, Peddada SD. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis 26:27663. https://doi.org/10.3402/mehd.v26.27663.
42. Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 1:33–46. https://doi.org/10.1002/gch2.1018.
43. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG. 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Gigascience 1:7. https://doi.org/10.1186/2047-217X-1-7.
44. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2.
45. Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z, Lozupone C, McDonald D, Robeson M, Sammut R, Smit S, Wakefield MJ, Widmann J, Wikman S, Wilson S, Ying H, Huttley GA. 2007. PyCogent: a toolkit for making sense from sequence. Genome Biol 8:R171. https://doi.org/10.1186/gb-2007-8-8-r171.
46. Felsenstein J. 2003. Inferring phylogenies. Sinauer Associates, Sunderland, MA.
47. Pirrung M, Kennedy R, Caporaso JG, Stombaugh J, Wendel D, Knight R. 2011. TopiaryExplorer: visualizing large phylogenetic trees with environmental metadata. Bioinformatics 27:3067–3069. https://doi.org/10.1093/bioinformatics/btr517.
48. Aitchison J, Greenacre M. 2002. Biplots of compositional data. J R Stat Soc C 51:375–392. https://doi.org/10.1111/1467-9876.00275.
49. Vázquez-Baeza Y, Gonzalez A, Smarr L, McDonald D, Morton JT, Navas-Molina JA, Knight R. 2017. Bringing the dynamic microbiome to life with animations. Cell Host Microbe 21:7–10. https://doi.org/10.1016/j.chom.2016.12.009.
50. Hunter JD. 2007. Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55.