Iroki: automatic customization and visualization of phylogenetic trees

Ryan M. Moore[1], Amelia O. Harrison[2], Sean M. McAllister[2], and K. Eric Wommack[1]

[1]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA
[2]School of Marine Science and Policy, University of Delaware, Newark, DE, USA

Corresponding author:
K. Eric Wommack[1]

Email address: wommack@dbi.udel.edu

ABSTRACT

Phylogenetic trees are an important analytical tool for examining community diversity and the evolutionary history of species. In the case of microorganisms, decreasing sequencing costs have enabled researchers to generate ever-larger sequence datasets, which in turn have begun to fill gaps in the evolutionary history of microbial groups. However, phylogenetic analyses of these types of datasets present challenges to the extraction of meaningful trends from complex trees. Scientific inferences made by visual inspection of phylogenetic trees can be simplified and enhanced by customizing various parts of the tree. Yet, manual customization is time-consuming and error prone, and programs designed to assist in batch tree customization often require programming experience or complicated file formats for annotation. To address these issues, we developed Iroki, a user-friendly web interface for tree visualization that provides automatic customization of phylogenetic trees based on metadata and abundance information. The Iroki web app and documentation is available at `http://www.iroki.net` or through the VIROME portal (`http://virome.dbi.udel.edu`). Iroki's source code is released under the MIT license and is available at `https://github.com/mooreryan/iroki`.

INTRODUCTION

Community and population ecology studies often use phylogenetic trees as a means to assess the diversity and evolutionary history of organisms. In the case of microorganisms, the declining cost of sequencing has enabled researchers to gather ever-larger sequence datasets from unknown microbial populations within environmental samples. While large sequence datasets have begun to fill gaps in the evolutionary history of microbial groups (Simister et al., 2012; Müller et al., 2015; Lan et al., 2016; Larkin et al., 2016; Wu et al., 2016), they have also posed new analytical problems, as extracting meaningful trends from high dimensional datasets can be challenging. In particular, scientific inferences made by visual inspection of phylogenetic trees can be simplified and enhanced by customizing various parts of the tree.

Many solutions to this problem currently exist. Standalone tree visualization packages allowing manual or batch modification of trees are available (e.g., Archaeopteryx (Han and Zmasek, 2009), Dendroscope (Huson et al., 2007), FigTree (Rambaut, 2016), TreeGraph2 (Stöver and Müller, 2010), Treevolution (Santamaría and Therón, 2009)), but the process can be time consuming and error prone especially when dealing with trees containing many nodes. Some packages allow batch and programmatic customizations through the use of an API or command line software (e.g., APE (Paradis et al., 2004), Bio::Phylo (Vos et al., 2011), Bio.Phylo (Talevich et al., 2012), ColorTree (Chen and Lercher, 2009), ETE (Huerta-Cepas et al., 2016), GraPhlAn (Asnicar et al., 2015), JPhyloIO (Stöver et al., 2016), phytools (Revell, 2012), treeman (Bennett et al., 2017)). While these packages are powerful, they could prove challenging depending on the expertise of the user. Current web based tree viewers are convenient in that they do not require the installation of additional software and provide customization and management features (e.g., Evolview (He et al., 2016), IcyTree (Vaughan, 2017), iTOL (Letunic and Bork, 2016), PhyD3 (Kreft et al., 2017), Phylemon (Sánchez et al., 2011), PhyloBot (Hanson-Smith and Johnson, 2016)), but often have complex user interfaces or complicated file formats to enable complex annotations. Iroki strikes a balance between flexibility and usability by providing visualization of trees with powerful automatic customization based on simple tab-separated text files of metadata and abundance information in a clean, user-friendly web interface.

METHODS

Iroki is a web application for visualizing and automatically customizing taxonomic and phylogenetic trees with associated qualitative and quantitative metadata. Iroki is particularly well suited to projects in microbial ecology and those that deal with microbiome data, as these types of studies generally have rich sample-associated metadata and represent complex community structures. The Iroki web application is available at `http://www.iroki.net` and the VIROME portal (`http://virome.dbi.udel.edu`) (Wommack et al., 2012). Iroki's source code is released under the MIT license and is available on `GitHub` at: `https://github.com/mooreryan/iroki`.

Implementation

Iroki is built with the Ruby on Rails web application framework. The main features of Iroki are written entirely in JavaScript allowing all data processing to be done client-side. This provides the additional benefit of eliminating the need to transfer potentially private data to an online service.

Iroki consists of two main modules: the tree viewer, which also handles customization with mapping files, and the biom converter, which creates mapping files to use in the tree viewer based on abundance data from a tab-separated file (e.g., one exported from a JSON or hdf5 style biom format (McDonald et al., 2012)).

Tree viewer

Rendering trees In order to render trees, Iroki uses JavaScript and Scalable Vector Graphics (SVG, an XML-based markup language for representing vector graphics). The Document Object Model (DOM) and SVG elements are manipulated with the D3.js library (Bostock et al., 2011).

Tree layouts Iroki provides three commonly used tree layouts: rectangular, circular, and radial. Rectangular and circular layouts are generated using D3's cluster layout API (`d3.cluster()`). For radial layouts, we implemented Algorithm 1 from Bachmaier et al. (2005) in JavaScript.

Automatic customization Iroki provides the option to automatically style aspects of the tree using a tab-separated text file (mapping file). Entries in the first column of this file are matched against all leaves in the tree using either exact or substring matching. If a leaf name matches a row in the mapping file, the styling options specified by the remaining columns are applied to that node. Inner nodes are styled to match their descendant nodes if all descendants also have that same style, allowing quick identification of clades sharing the same metadata. Aspects of the tree that can be automatically styled using the mapping file include leaf label color, font, size, and name, leaf dot color and size, as well as branch width and color.

Manual customization In addition to automatic customization using a mapping file, various aspects of the tree can be manually adjusted including tree layout and rotation, branch sorting, scale bar size, and characteristics of labels (color, font, size, and alignment), dots (color, size, bootstrap support values), and branches (width and color).

Biom converter

Overview

Iroki's biom converter accepts tab-separated biom files such as those exported by VIROME (Wommack et al., 2012) or QIIME (Caporaso et al., 2010) and converts the count data into a color space.

Input format Iroki reads continuous data from tab-separated biom files (classic format tables). These files can be created automatically from JSON or hdf5 formatted biom files using the `biom convert` utility or manually in a spreadsheet program such as Microsoft Excel or any text editor.

Color space Iroki has two options for color space: CIELCh, which is a cylindrical representation of the CIELab color space (a color space that attempts to better approximate human vision), and hue saturation lightness (HSL) (Aisch, 2018).

Data reduction Data reduction can be a powerful method for extracting meaningful trends in high dimensional data sets. Given that microbiome or other studies in microbial ecology can have hundreds of samples, Iroki provides the option to perform data reduction on the data in the biom file. Iroki can use a specified number of principle components to convert into color space rather than the full data set. For example, using Iroki's geometry coloring mode (see below), many samples can be simplified by first projecting the data into three dimensions. For data reduction, we implement principal components analysis (PCA) calculated via the singular value decomposition using the LALOLib scientific computing library for JavaScript (Lauer, 2017). Briefly, performing singular value decomposition on the centered (and optionally scaled) count matrix X, we obtain the following decomposition

$$X = USV^T, \tag{1}$$

where the columns of US are the principal components, S is the diagonal matrix of singular values, and the columns of V are the principal directions.

Data characteristics

Iroki uses several characteristics to project data into color space.

Sample evenness One component is evenness across samples for each node in the biom file. For this we use Pielou's evenness index (Pielou, 1966):

$$J' = H'/H'_{max}, \tag{2}$$

where H' is the Shannon entropy, and H'_{max} is the maximum value of H'. In our case, H'_{max} occurs when the leaf node has equal counts across all samples in the biom file. Putting it together, we calculate Pielou's index as

$$J' = \frac{-\sum_{n=1}^{N} p_i \log_2(p_i)}{\log_2(N)}, \tag{3}$$

where N is the number of samples and p_i is the relative abundance of the node in the i-th sample.

Mean abundance The abundance of a node across samples is calculated by taking the mean abundance across all samples or the mean abundance across non-zero count samples depending on the option selected by the user.

Sample weight For each leaf node, Iroki calculates a value for each sample to represent the overall ”weight” of each sample for that node, according to the following formula:

$$w_{ij} = \frac{c_{ij}}{c_i\text{max}}, \tag{4}$$

where c_{ij} is the count for the i-th leaf node in the j-th sample and $c_i$max is the max count for the i-th leaf node. In this way, a leaf node that is at high abundance in Sample A but low abundance in Sample B would have a higher weight for sample A and a lower weight for Sample B. Note that Iroki does not normalize counts/abundance values (for example, to account for differences in sequencing effort) before calculating the sample weight. If such normalizations are required, they should be performed prior to analysis with Iroki.

Mapping data to color space

Iroki uses the three data characteristics discussed above in various combinations to project leaves into color space based on sample count data.

Geometry mode Iroki's geometry mode uses all three data characteristics to project count data into the color space (Fig. S1). In this mode, sample weight is mapped to hue, sample evenness is mapped to chroma, and mean abundance is mapped to lightness.

In order to map sample weights to hue, we first map each sample (or principal component if data reduction is used) to an angle θ. Then, for each leaf node, a polygon is drawn within the confines of the unit circle (as the sample weights are bound from zero to 1) using the following formula:

$$P_{ij} = (w_{ij} \cos\theta_j, w_{ij} \sin\theta_j)\,, \tag{5}$$

where P_{ij} is the point for the i-th leaf and the j-th sample, w_{ij} is the sample weight as calculated in Eq. 4, and θ is the angle for the j-th sample (see Fig. S1 for an example).

Next, we find the centroid of the whole polygon by dividing the entire shape into its constituent triangles, finding the centroid of those triangles, and then taking the weighted average of those centroids. Thus, the angle from the origin to the centroid of the polygon determines the hue angle.

Sample evenness is mapped to chroma and mean abundance is mapped to lightness. The user has the option to set whether high values for evenness and abundance represent high or low values for chroma and lightness respectively.

Geometry mode works best with three or fewer samples (e.g., treatment vs. control, or high, medium, low) or in cases where samples are grouped by similarity beforehand since the order of the samples in the biom file affects the perception of their similarity (e.g., S2 is closer in color to S1 or S3 than to S5). For example, it may be beneficial to group samples in the biom file by a certain metadata type or by a pre-computed similarity measure. If either of these cases do not hold, data reduction should be performed or one of Iroki's other coloring modes should be used.

Other modes Iroki gives the user the option to map data characteristics into color space in various ways. After selecting one of many built in color gradients (single-, two-, or multi-color gradients including cubehelix (Green, 2011) and those from ColorBrewer (Brewer et al., 2013) are available), Iroki can map data characteristics directly onto the chosen gradient. For example, evenness across samples can be examined separately from abundance by choosing to map the sample evenness data characteristic onto a blue-green color scale. Iroki also can place leaf nodes on a color gradient based on the first component of a PCA.

RESULTS AND DISCUSSION

Global diversity of bacteriophage

Viruses are the most abundant biological entities on Earth, providing an enormous reservoir of genetic diversity, driving evolution of their hosts, influencing composition of microbial communities, and affecting global biogeochemical cycles (Suttle, 2007; Rohwer and Thurber, 2009). The viral taxonomic system developed by the International Committee on Taxonomy of Viruses (ICTV) is based on a suite of physical characteristics of the virion rather than on genome sequences. Noting this limitation, the phage proteomic tree was created to provide a genome-based taxonomic system for bacteriophage classification (Rohwer and Edwards, 2002). The phage proteomic tree was recently updated to include hundreds of new phage genomes with taxonomy and host information from the Phage SEED reference database (Overbeek et al., 2014), as well as long assembled contigs from viral shotgun metagenomes (viromes) collected from the Chesapeake Bay (SERC sample) (Wommack et al., 2015) and the Mediterranean Sea (Mizuno et al., 2013).

Uncultured phage contigs from the Chesapeake Bay and Mediterranean viromes make up a large portion of all phage sequences shown on the tree, and are widely distributed among known phage. In general, viruses in the same family claded together, e.g., branch coloring highlights large groups of closely related Siphoviridae and Myoviridae. This label-coloring scheme also shows that in general, viruses infecting hosts within the same phylum are phylogenetically similar. For example, though there are multiple large groups of Siphoviridae spread across the tree, these clades are distinguished by the fact that they generally infect host species within different phylums (e.g., Siphoviridae infecting Actinobacteria clade away from Siphoviridae infecting Firmicutes or Proteobacteria).

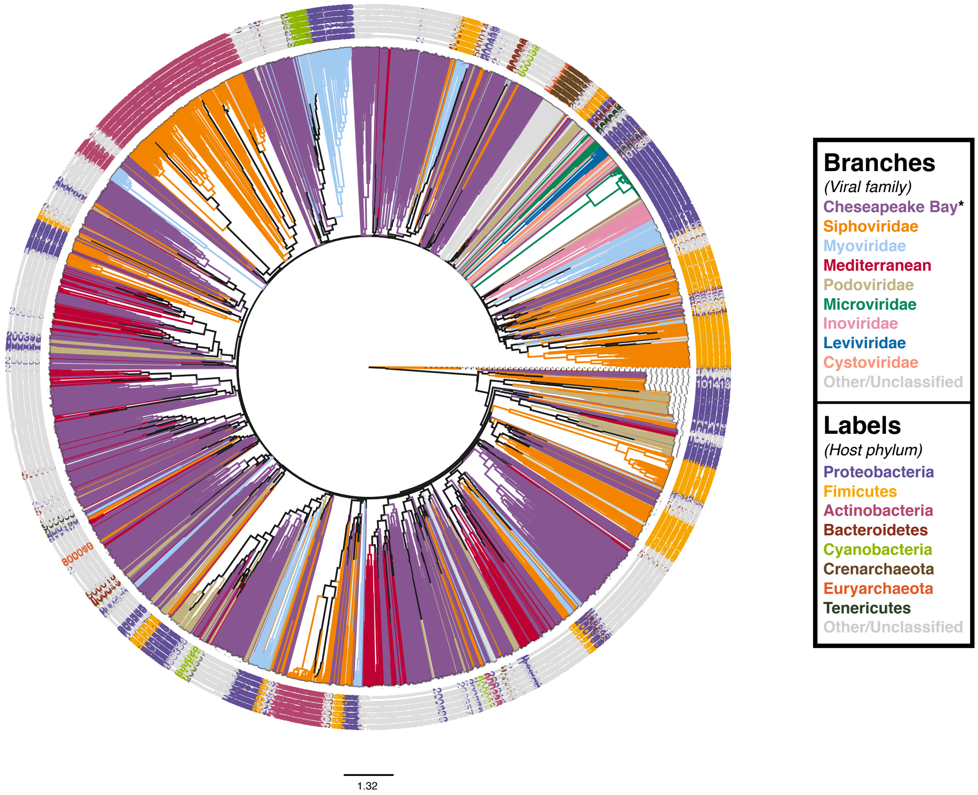

Figure 1. **Comparing phage and their host phyla.** All phage genomes from Phage SEED with assembled virome contigs from the Chesapeake Bay and Mediterranean Sea. Branches are colored according to viral family or sampling location in the case of virome contigs. Leaf node labels are colored accorting to host phylum of the phage.

Bacterial community diversity and prevalence of *E. coli* in beef cattle

Shiga toxin-producing *Escherichia coli* (STEC) are dangerous human pathogens that colonize the lower gastrointestinal tracts of cattle and other ruminants. STEC-contaminated beef and STEC shed in the feces of these animals are major sources of foodborne illness (Hancock et al., 1994; Caprioli et al., 2005). To identify possible interactions between STEC populations and the commensal cattle microbiome, a recent study examined the diversity of the bacterial community associated with beef cattle hide (Chopyk et al., 2016). Hide samples were collected over twelve weeks and SSU rRNA amplicon libraries were constructed and analyzed by Illumina sequencing (Fadrosh et al., 2014). The study indicated that the community structure of hide bacterial communities was altered when the hides were positive for STEC contamination.

Iroki was used to visualize changes in the relative abundance of each cattle hide bacterial OTU according to the presence or absence of STEC. A Mann-Whitney U test comparing OTU abundance between STEC positive and STEC negative samples was performed. Cluster representatives of any bacterial OTU with a p-value < 0.2 were aligned with SILVA Ref NR using SILVA's online ACT service (`https://www.arb-silva.de/aligner/`). This p-value was chosen to illustrate Iroki's ability to combine continuous data with categorical data (in this case, abundance and significance of a statistical test). Branches of the tree were colored based on whether there was a significant change in relative abundance with STEC contamination (pink: $p < 0.05$, gray: $p \leq 0.05$). We used Iroki to generate a color gradient based on the log transformed abundance of each OTU in samples with and without STEC using the geometry mode of the biom converter (Fig. 2). Most OTUs on the tree did not have a significant difference in abundance between STEC positive and STEC negative samples. Also, the color gradient makes it clear that most OTUs are at low abundance with only a few highly abundant OTUs (i.e., most leaf dots are dark and only a few are light), and that the highly abundant OTUs were spread across the tree in many different phylogenetic groups. In this way, Iroki enables an end-user to visualize trends that would have been hidden in an undecorated tree.

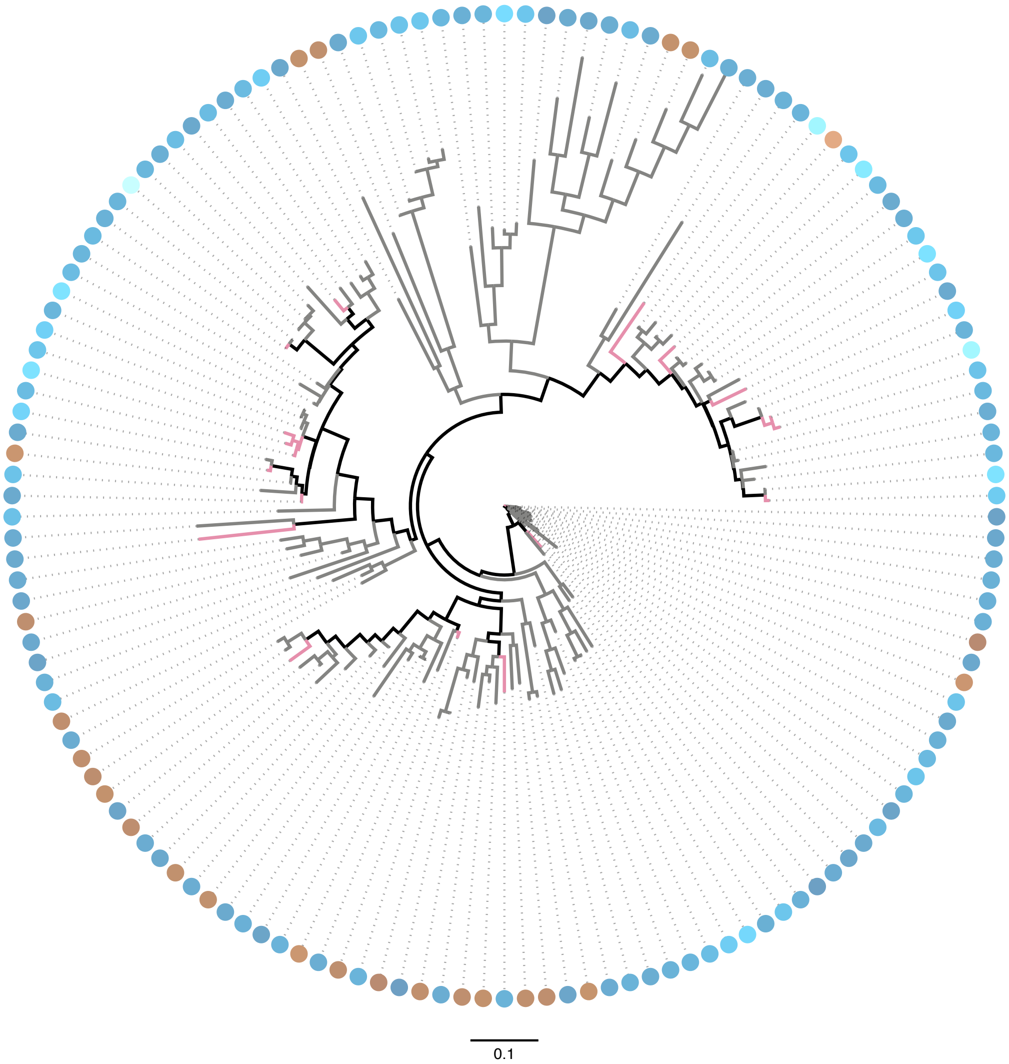

Figure 2. **Changes in OTU abundance in two sample groups.** Approximate-maximum likelihood tree of hide OTUs that showed differences in relative abundance between STEC positive and STEC negative cattle hide samples. Branches show significance based on coloring by the *p*-value of a Mann-Whitney U test ($p < 0.05$ – pink, $p \geq 0.05$ – gray) examining changes in abundance between samples positive for STEC and samples negative for STEC. Label color on a brown-blue color gradient highlights OTU occurrence based on the abundance ratio between STEC positive samples (brown) and STEC negative samples (blue). Dot lightness represents overall abundance with lighter nodes being more abundant than darker nodes.

Tara Oceans viromes

Ribonucleotide reductase (RNR) is common within viral genomes (Dwivedi et al., 2013) and RNR polymorphism is predictive of certain biological and ecological features of viral populations (Sakowski et al., 2014). As such, it is an often used marker gene for the study of viral communities. To explore viral communities of the global ocean, we mined the *Tara* Oceans viromes for RNR genes. The *Tara*

Oceans expedition was a two-and-a-half year survey that began in March 2012 and sampled over 200 stations across the world oceans (Bork et al., 2015; Pesant et al., 2015). Any virome containing less than 50 RNR sequences was removed leaving a total of 5,680 RNR genes across 41 samples. These sequences were aligned with MAFFT (Katoh and Standley, 2013) and post-processed manually to ensure optimal alignment quality. Then, FastTree (Price et al., 2010) was used to infer a phylogeny from the alignment. Using this tree, the unweighted UniFrac distance (Lozupone and Knight, 2005) between samples was calculated using QIIME (Caporaso et al., 2010). A tree was generated from this distance matrix in R using average-linkage hierarchical clustering. Additionally, Mantel tests identified that conductivity, oxygen, and latitude were significantly correlated ($p < 0.05$) with the UniFrac distance between samples. Finally, Iroki's biom converter was used to generate a cubehelix color palette (Green, 2011) based on a scaled, 1-dimensional projection of the conductivity, oxygen, and latitude values for each sample (Fig. 3).

Samples that are more similar to one another with respect to the three environmental parameters will appear more similar in color (e.g., TARA 031 SRF and TARA 032 SRF are both lilac and therefore more similar to each other than to TARA 085 DCM, which is black). Samples that cluster together with respect to UniFrac distance based on RNR protein presence/absence are often very similar in color, indicating that they have a similar pattern among the three environmental parameters. In this way, we can see that the high-level viral community structure is influenced by the environmental parameters of the sample in which they originate. Additionally, it supports the use of RNR as a marker gene for studying viral diversity, as the structure of the viral communities as inferred by this gene appear to recapitulate environmental parameters.

Humongous trees

Because browser performance degrades when rendering many thousands of SVG elements, to render giant trees without collapsing nodes, Iroki has a HTML5 Canvas, straight-to-png tree viewer with the ability to display trees of up to 1,000,000 leaf nodes in under 30 seconds on our test machine (MacBook Pro, 2.66 GHz Intel Core i7, 8GB RAM, with Opera version 54.0.2952.64) (Fig. S2). To illustrate the utility of this feature, we downloaded the large GreenGenes (DeSantis et al., 2006) ”two-study filter” tree (331,550 sequences) from the FastTree website and used Iroki to visualize the tree (Fig. 4). Briefly, full length SSU rRNA sequences from GreenGenes were collected and clustered based on divergence from an ancestor in a minimum-evolution tree. Sequences were kept only if they were in a cluster with sequences from two different studies or if they were named isolates. Sequences were then aligned with NAST and the phylogeny inferred with FastTree (Price et al., 2010) (see www.microbesonline.org/fasttree for full details).

CONCLUSIONS

Iroki is a web application for fast, automatic customization and visualization of large phylogenetic trees based on user specified configuration files with categorical metadata or biom-style files of continuous metadata. Various example datasets from microbial ecology studies were analyzed to demonstrate Iroki's utility. In each case, Iroki simplified the processes of data exploration, data presentation, and hypothesis testing. Though these examples focused specifically on applications in microbial ecology, Iroki is applicable to any problem space with hierarchical data that can be represented in the Newick tree format. Iroki provides a simple and convenient way to rapidly visualize and customize trees, especially in cases where the tree in question is too large to annotate manually or in studies with many trees to annotate.

ACKNOWLEDGMENTS

We would like to acknowledge Jessica M. Chopyk and Daniel J. Nasko for their work on the phage proteomic tree, and Barbra D. Ferrell and Shawn W. Polson for editing the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Availability of data and materials

Data and code used to generate figures are available on GitHub at https://github.com/mooreryan/iroki_manuscript_data.

Funding

This project was supported by the Agriculture and Food Research Initiative grant no. 2012-68003-30155 from the USDA National Institute of Food and Agriculture and the National Science Foundation Advances in Biological Informatics program (award number DBI_1356374). Computational infrastructure support by the University of Delaware Center for Bioinformatics and Computational Biology Core Facility was made possible through funding from Delaware INBRE (NIH P20 GM103446) and the Delaware Biotechnology Institute.

Competing interests

The authors declare that they have no competing interests.

Author contributions

RMM and SMM conceived the project. RMM wrote the manuscript and implemented Iroki with assistance from AOH. All authors read, edited, and approved the final manuscript.

REFERENCES

Aisch, G. (2011-2018). chroma.js. `https://github.com/gka/chroma.js`. Accessed: 2018-08-16.

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, 3:e1029.

Bachmaier, C., Brandes, U., and Schlieper, B. (2005). Drawing phylogenetic trees. (Extended abstract). *Algorithms and computation. 16th international symposium, ISAAC 2005, Sanya, Hainan, China, December 19–21, 2005. Proceedings*, pages 1110–1121.

Bennett, D. J., Sutton, M. D., and Turvey, S. T. (2017). treeman: an R package for efficient and intuitive manipulation of phylogenetic trees. *BMC Research Notes*, 10(1):30.

Bork, P., Bowler, C., De Vargas, C., Gorsky, G., Karsenti, E., and Wincker, P. (2015). Tara Oceans studies plankton at Planetary scale. *Science*, 348(6237):873.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

Brewer, C., Harrower, M., and University, T. P. S. (2013). ColorBrewer2.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336.

Caprioli, A., Morabito, S., Brugère, H., and Oswald, E. (2005). Enterohaemorrhagic Escherichia coli: emerging issues on virulence and modes of transmission. *Vet. Res.*, 36(3):289–311.

Chen, W.-H. and Lercher, M. J. (2009). ColorTree: a batch customization tool for phylogenic trees. *BMC Research Notes*, 2(1):155.

Chopyk, J., Moore, R. M., DiSpirito, Z., Stromberg, Z. R., Lewis, G. L., Renter, D. G., Cernicchiaro, N., Moxley, R. A., and Wommack, K. E. (2016). Presence of pathogenic Escherichia coli is correlated with bacterial community diversity and composition on pre-harvest cattle hides. *Microbiome*, 4(1):9.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072.

Dwivedi, B., Xue, B., Lundin, D., Edwards, R. A., and Breitbart, M. (2013). A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evolutionary Biology*, 13(1):33.

Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M., and Ravel, J. (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, 2(1):6.

Green, D. A. (2011). A colour scheme for the display of astronomical intensity images. *Bulletin of the Astronomical Society of India*, 39(2):289–295.

Han, M. V. and Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10:356.

Hancock, D. D., Besser, T. E., Kinsel, M. L., Tarr, P. I., Rice, D. H., and Paros, M. G. (1994). The prevalence of Escherichia coli O157.H7 in dairy and beef cattle in Washington State. *Epidemiology and Infection*, 113(2):199–207.

Hanson-Smith, V. and Johnson, A. (2016). PhyloBot: A Web Portal for Automated Phylogenetics, Ancestral Sequence Reconstruction, and Exploration of Mutational Trajectories. *PLoS Computational Biology*, 12(7):1–10.

He, Z., Zhang, H., Gao, S., Lercher, M. J., Chen, W. H., and Hu, S. (2016). Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic acids research*, 44(W1):W236–W241.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6):1635–1638.

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, 8:1–6.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kreft, L., Botzki, A., Coppens, F., Vandepoele, K., and Van Bel, M. (2017). PhyD3: A phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*, 33(18):2946–2947.

Lan, Y., Rosen, G., and Hershberg, R. (2016). Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome*, 4(1):18.

Larkin, A. a., Blinebry, S. K., Howes, C., Lin, Y., Loftus, S. E., Schmaus, C. A., Zinser, E. R., and Johnson, Z. I. (2016). Niche partitioning and biogeography of high light adapted Prochlorococcus across taxonomic ranks in the North Pacific. *The ISME journal*, pages 1–13.

Lauer, F. (2017). MLweb: A toolkit for machine learning on the web. *Neurocomputing*, 282:74–77.

Letunic, I. and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, 44(W1):W242–W245.

Lozupone, C. and Knight, R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12):8228–8235.

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R., and Caporaso, J. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, 1(1):7.

Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., and Ghai, R. (2013). Expanding the marine virosphere using metagenomics. *PLoS genetics*, 9(12):e1003987.

Müller, A. L., Kjeldsen, K. U., Rattei, T., Pester, M., and Loy, A. (2015). Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi)sulfite reductases. *The ISME journal*, 9(5):1152–65.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., and Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42(D1):206–214.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., Dimier, C., Searson, S., Coordinators, T. O. C., Acinas, S. G., Bork, P., Boss, E., Bowler, C., De Vargas, C., Follows, M., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Karp-Boss, L., Karsenti, E., Krzic, U., Not, F., Ogata, H., Pesant, S., Raes, J., Reynaud, E. G., Sardet, C., Sieracki, M., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Velayoudon, D., Weissenbach, J., and Wincker, P. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, 2:150023.

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13(C):131–144.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3).

Rambaut, A. (2006-2016). FigTree. http://tree.bio.ed.ac.uk/software/figtree/. Accessed: 2016-07-21.

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223.

Rohwer, F. and Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of bacteriology*, 184(16):4529–35.

Rohwer, F. and Thurber, R. V. (2009). Viruses manipulate the marine environment. *Nature*, 459(7244):207–212.

Sakowski, E. G., Munsell, E. V., Hyatt, M., Kress, W., Williamson, S. J., Nasko, D. J., Polson, S. W., and Wommack, K. E. (2014). Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 111(44):15786–91.

Sánchez, R., Serra, F., Tárraga, J., Medina, I., Carbonell, J., Pulido, L., De María, A., Capella-Gutíerrez, S., Huerta-Cepas, J., Gabaldón, T., Dopazo, J., and Dopazo, H. (2011). Phylemon 2.0: A suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Research*, 39(SUPPL. 2):470–474.

Santamaría, R. and Therón, R. (2009). Treevolution: Visual analysis of phylogenetic trees. *Bioinformatics*, 25(15):1970–1971.

Simister, R. L., Deines, P., Botté, E. S., Webster, N. S., and Taylor, M. W. (2012). Sponge-specific clusters revisited: A comprehensive phylogeny of sponge-associated microorganisms. *Environmental Microbiology*, 14(2):517–524.

Stöver, B. C. and Müller, K. F. (2010). TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics*, 11:7.

Stöver, B. C., Wiechers, S., and Müller, K. F. (2016). JPhyloIO — A Java library for event-based reading and writing of different alignment and tree formats through one common interface Aims and concept Event based document reading Writing events using data adapters. (2009):48149.

Suttle, C. A. (2007). Marine viruses — major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801–812.

Talevich, E., Invergo, B. M., Cock, P. J., and Chapman, B. A. (2012). Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, 13(1).

Vaughan, T. G. (2017). IcyTree: Rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, 33(15):2392–2394.

Vos, R. A., Caravas, J., Hartmann, K., Jensen, M. A., and Miller, C. (2011). BIO :: Phylo-phyloinformatic analysis using perl.

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S., and Nasko, D. J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences*, 6(3):421–433.

Wommack, K. E., Nasko, D. J., Chopyk, J., and Sakowski, E. G. (2015). Counts and sequences, observations that continue to change our understanding of viruses in nature. *Journal of Microbiology*, 53(3):181–192.

Wu, Z., Yang, L., Ren, X., He, G., Zhang, J., Yang, J., Qian, Z., Dong, J., Sun, L., Zhu, Y., Du, J., Yang, F., Zhang, S., and Jin, Q. (2016). Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. *The ISME Journal*, 10(3):609–620.

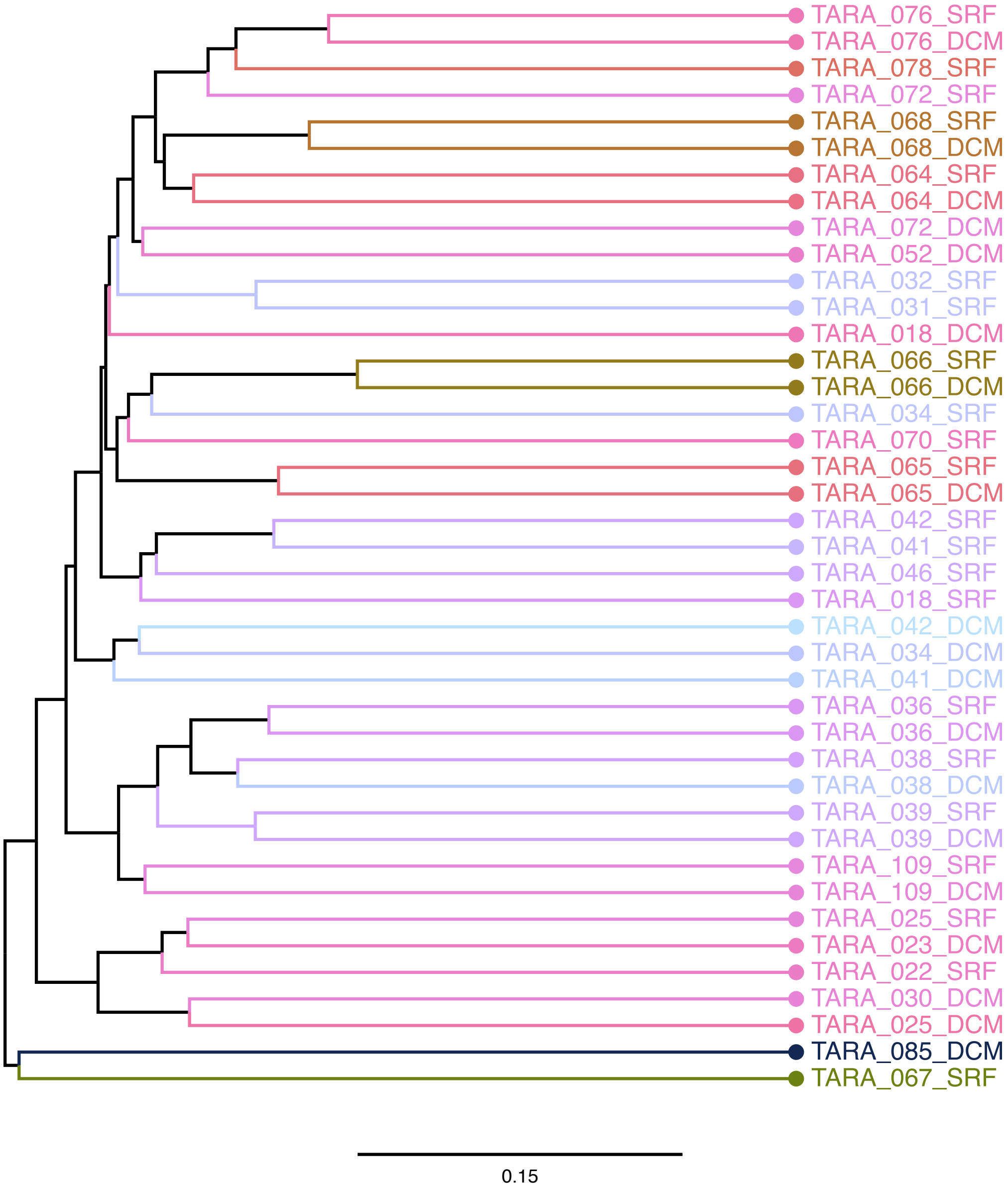

Figure 3. ***Tara* viromes.** Average-linkage hierarchical clustering based on RNR sequences mined from 41 *Tara* viromes. Color is based on a scaled, 1-dimensional projection of sample conductivity, oxygen, and latitude onto the cubehelix color gradient. Samples that are more similar to each other in color represent those that are more similar to each other with respect to the environmental parameters in the ordination.

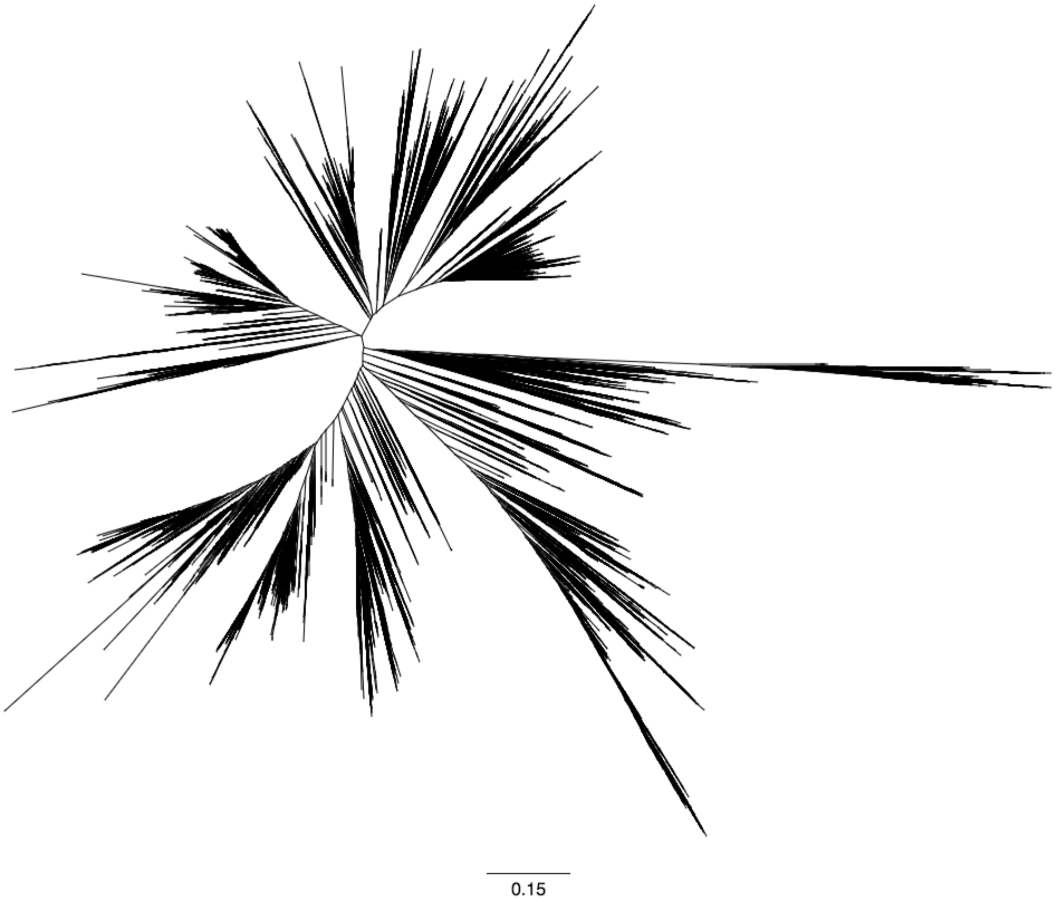

Figure 4. GreenGenes SSU rRNA tree. A collection of 331,550 full length SSU rRNA sequences from GreenGenes rendered with Iroki's Canvas tree viewer.

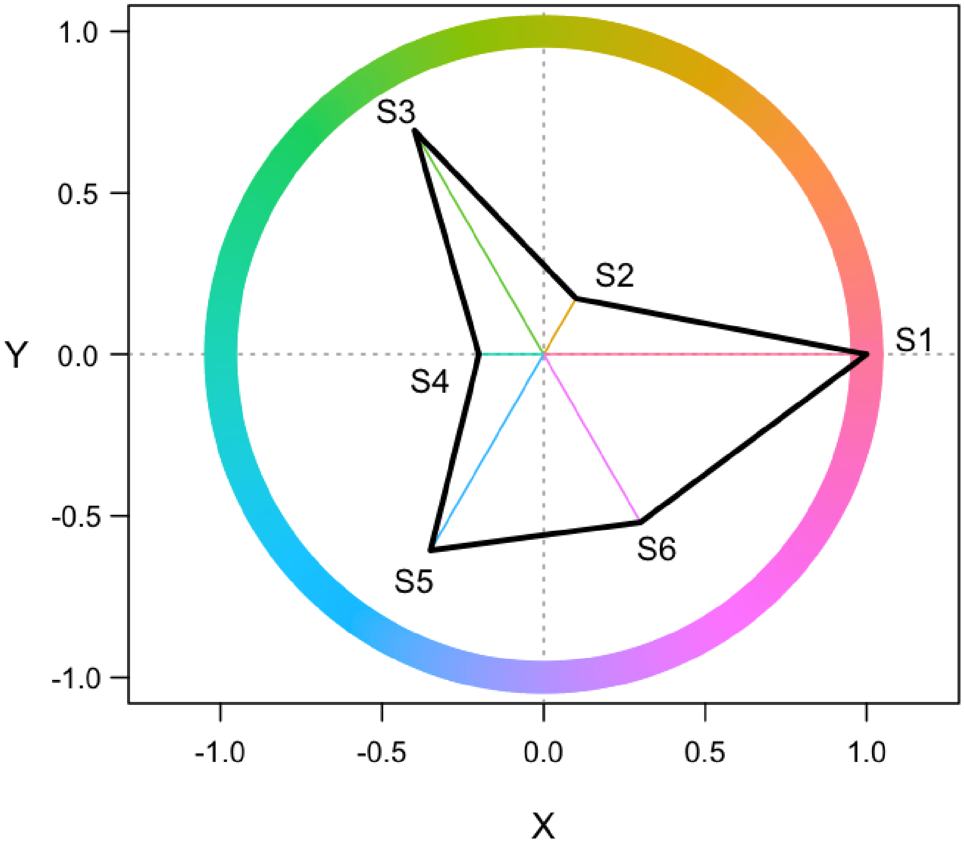

Figure S1. **Example sample weight polygon.** This is the polygon that would be produced by a leaf in a six sample biom file with the following counts, S1: 10, S2: 2, S3: 8, S4: 2, S5: 7, S6: 0. In this case, samples are evenly distributed across the circle.

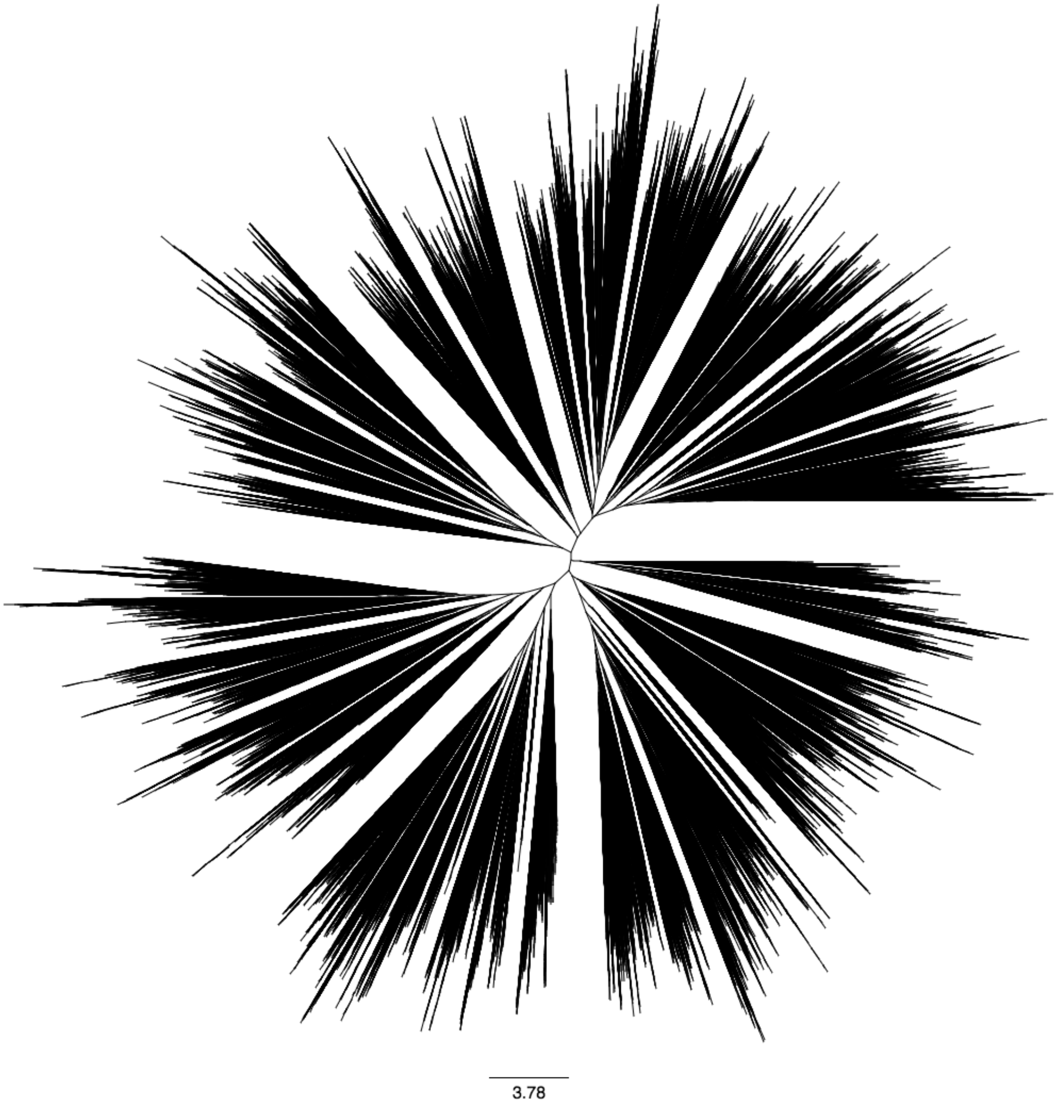

Figure S2. A very large tree. A 1,000,000 leaf tree with random branch lengths generated with `rtree` (using `runif` with default arguments for branch lengths) from the ape R package.