Iroki: automatic customization and visualization of phylogenetic trees

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

[1] Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, United States of America

[2] School of Marine Science and Policy, University of Delaware, Newark, DE, United States of America

ABSTRACT

Phylogenetic trees are an important analytical tool for evaluating community diversity and evolutionary history. In the case of microorganisms, the decreasing cost of sequencing has 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 create complex trees that can be challenging to interpret. 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. Iroki, a user-friendly web interface for tree visualization, addresses these issues by providing automatic customization of large trees based on metadata contained in tab-separated text files. Iroki's utility for exploring biological and ecological trends in sequencing data was demonstrated through a variety of microbial ecology applications in which trees with hundreds to thousands of leaf nodes were customized according to extensive collections of metadata. The Iroki web application and documentation are available at https://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.

Submitted 2 October 2019
Accepted 17 January 2020
Published 26 February 2020

Corresponding author
K. Eric Wommack,
wommack@udel.edu

Academic editor
Thiago Venancio

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.8584**

Subjects Bioinformatics, Computational Biology, Ecology, Microbiology, Virology

Keywords Bioinformatics, Environmental microbiology, Phylogeny, Sequence analysis, Metagenomics, Microbiome, Viral ecology, Microbial ecology, Data visualization, Software

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, declining sequencing cost 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, Rosen & Hershberg, 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 & Zmasek, 2009*), Dendroscope (*Huson et al., 2007*), FigTree (*Rambaut, 2006*), TreeGraph2 (*Stöver & Müller, 2010*), Treevolution (*Santamaría & 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 application programming interface (API) or command line software (e.g., APE (*Paradis, Claude & Strimmer, 2004*), Bio::Phylo (*Vos et al., 2011*), Bio.Phylo (*Talevich et al., 2012*), ColorTree (*Chen & Lercher, 2009*), ETE (*Huerta-Cepas, Serra & Bork, 2016*), GraPhlAn (*Asnicar et al., 2015*), JPhyloIO (*Stöver, Wiechers & Müller, 2016*), phytools (*Revell, 2012*), treeman (*Bennett, Sutton & Turvey, 2017*)). While these packages are powerful, they require substantial computing expertise, which can be an impediment for some scientists. 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 & Bork, 2016*), PhyD3 (*Kreft et al., 2017*), Phylemon (*Sánchez et al., 2011*), PhyloBot (*Hanson-Smith & Johnson, 2016*), Phylo.io (*Robinson, Dylus & Dessimoz, 2016*)), but often have complex user interfaces or complicated file formats to enable complex annotations. Iroki strikes a balance between flexibility and usability by combining visualization of trees in a clean, user-friendly web interface with powerful automatic customization based on simple, tab-separated text (mapping) files. Given its focus on automatic customization and a core set of key features, Iroki's user interface can remain lean and easy-to-learn while still enabling complex customizations. In addition to specifying simple color gradients directly in the mapping file, Iroki also provides a dedicated module allowing the user to generate custom gradients to embed their data into color space, enhancing visualization. Iroki stays responsive even when customizing large trees, and it does not require an account or uploading potentially sensitive data to an external service.

Here, Iroki was used to customize large trees containing hundreds to thousands of leaf nodes according to extensive collections of metadata. These applications demonstrated the utility of Iroki for distilling biological and ecological insights from microbial community sequence data. The particular use cases included examinations of phage-host interactions, relative abundance of populations across sample types, and comparisons of viral community composition across environmental gradients.

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 and documentation are

available at the following web address: https://www.iroki.net, or through 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: 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 tab-separated text files (mapping files), and the color gradient generator, which creates mapping files to use in the tree viewer based on quantitative data (such as counts) from a tab-separated text file similar to the classic-style OTU tables exported from a JSON or hdf5 format biom file (*McDonald et al., 2012*).

Tree viewer

Iroki uses JavaScript and Scalable Vector Graphics (SVG, an XML-based markup language for representing vector graphics) for rendering trees. The Document Object Model (DOM) and SVG elements are manipulated with the `D3.js` library (*Bostock, Ogievetsky & Heer, 2011*). Rectangular, circular, and radial tree layouts are provided in the Iroki web application. Rectangular and circular layouts are generated using D3's cluster layout API (`d3.cluster`). For radial layouts, Algorithm 1 from *Bachmaier, Brandes & Schlieper (2005)* was implemented in JavaScript. In addition to the SVG tree viewer, Iroki also includes an HTML5 Canvas viewer with a reduced set of features capable of displaying huge trees with millions of leaf nodes (Supplementary Materials Sec. 4).

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 leaf labels 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 so that if all descendant nodes moving towards the inner parts of the tree have the same style, then quick identification of clades sharing the same metadata is possible. Aspects of the tree that can be automatically styled using the mapping file include branches, leaf labels, leaf dots, bar charts, and arcs.

Inner node labels may represent support values (e.g., bootstrap results) or other comments that describe the inner nodes. If inner labels are numeric, then inner nodes can be decorated with filled and unfilled circles that allow quick identification of branches with high support. The semantics of support labels are key to proper tree representations (*Czech, Huerta-Cepas & Stamatakis, 2017*). As Iroki currently does not implement tree rerooting, Iroki handles these specifics implicitly rather than giving the option to map inner node labels to branches or to the nodes themselves.

While Iroki is focused mainly on automatic customization via mapping files, some interactive features are included such as node selection and the ability to modify labels after a tree has been submitted. Finally, various aspects of the tree can be adjusted directly through Iroki's user interface.

Color gradient generator

Iroki's color gradient generator accepts tab-separated text files (similar to the classic-style count tables exported by VIROME (*Wommack et al., 2012*) or QIIME 1 (*Caporaso et al., 2010*)) and converts the numerical data (e.g., counts/abundances) into a color gradient. Several single-, two-, and multi-color gradients are provided including cubehelix (*Green, 2011*) and those from ColorBrewer (*Brewer, Harrower & University, 2013*).

Iroki reads numerical data from tab-separated text files. Similar to the mapping file for the tree viewer, the first column should match leaf names in the tree, and the remaining columns describe whatever aspect of the data is of interest to the researcher (e.g., counts or abundance). In a dataset with M observations and N variables, the input file will then have $M+1$ rows (the first row is the header) and $N+1$ columns (the first column specifies observation names). From this data, Iroki can generate color gradients in a variety of ways.

Observation means

A color gradient is generated based on the mean value of each observation across all variables. In this case, an observation i would be represented as $\mu_i = \sum_{j=1}^{N} c_{ij}$, where c_{ij} is the value of observation (row) i for variable (column) j.

Observation "evenness"

A color gradient is generated based on the "evenness" of observation i across all N variables. Then, each observation i is represented by Pielou's evenness index (*Pielou, 1966*) calculated across all variables: $E_i = H_i / H_{\max}$, where H_i is the Shannon entropy for observation i with respect to the N variables specified in the input file, and H_{max} is the maximum theoretical value of H_i. In this case, H_{max} occurs when observation i has equal values c_{ij} across all N variables. Thus, we calculate Pielou's evenness index for an observation i as

$$E_i = \frac{-\sum_{j=1}^{N} p_{ij} \log_2 p_{ij}}{\log_2 N},$$

where N is the number of variables and p_{ij} is the proportion of observation i in variable j (i.e., $c_{ij} / \sum_{j=1}^{N} c_{ij}$).

In this way, the user can map observations with high evenness (i.e., an observation with approximately the same value for each variable) to one side of the color gradient and observations with low evenness (i.e., an observation with high values in a few variables and low values in most others) to the other side of the gradient for easy identification.

Observation projection

Data reduction can be a powerful method for extracting meaningful trends in large, high-dimensional data sets. Given that microbiome or other studies in microbial ecology can have hundreds of samples and a rich set of metadata associated with those samples, data reduction often proves useful. Thus, Iroki provides a method to project the data into a single dimension and then map that projection onto a color gradient. For data reduction, Iroki conducts a principal components analysis (PCA) calculated via the singular value decomposition (SVD) using the LALOLib scientific computing library for JavaScript (*Lauer, 2017*). Briefly, performing singular value decomposition on the centered

(and optionally scaled) matrix X, with observations as rows and variables as columns, the following decomposition is obtained: $X = USV^T$, where the columns of US are the principal component scores, S is the diagonal matrix of singular values, and the columns of V are the principal axes. To illustrate as much variance as possible in a single dimension, the first principal coordinate is mapped onto the chosen color gradient.

RESULTS AND DISCUSSION

Bacteriophage proteomes, taxonomy, and host phyla

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 & Thurber, 2009*). Due to their importance, there is a growing interest in connecting viruses with their hosts through the analysis of metagenome data. As such, researchers have used a variety of computational techniques to predict viral-host interactions including CRISPR-spacers (*Roux et al., 2016*; *Coutinho et al., 2017*; *Nishimura et al., 2017a*) and tRNA matches (*Bellas, Anesio & Barker, 2015*; *Roux et al., 2016*; *Coutinho et al., 2017*; *Nishimura et al., 2017a*), sequence homology (*Roux et al., 2016*; *Coutinho et al., 2017*; *Nishimura et al., 2017a*), abundance correlation (*Coutinho et al., 2017*), and oligonucleotide profiles (*Roux et al., 2015*; *Roux et al., 2016*; *Munson-McGee et al., 2018*).

We used Iroki to examine phage-host interactions at the taxonomic scale by constructing a tree based on proteomic content (*Rohwer & Edwards, 2002*) from a subset of viral genomes from the Virus-Host DB (*Mihara et al., 2016*) using ViPTree (*Nishimura et al., 2017b*) (Fig. 1; Supplementary Materials Sec. 1). A proteomic tree clusters phage based on relationships between the collection of protein-encoding genes encoded within their genomes (*Rohwer & Edwards, 2002*; *Nelson, 2004*; *Wommack et al., 2015*). Specifically, ViPTree bases its clustering on normalized tBLASTx scores between genomes following the method of *Mizuno et al. (2013)*.

Tree branches were colored by host phyla and virus family was indicated by a ring surrounding the tree using Iroki's bar plot options (Fig. 1; Supplementary Materials Sec. 1). As shown by the branch coloring, host phyla mapped well onto the proteomic tree (i.e., large clusters of viruses that are similar in their proteomic content often infect the same host phylum). Firmicutes-infecting phage (represented by blue branches of the tree in Fig. 1) are confined almost exclusively to a large cluster in the top-left quadrant of the tree. This large cluster of mostly Firmicutes-infecting viruses can be further partitioned according to virus family, with a distinct group of myoviruses clustering separately from the other clades which include mostly siphoviruses. The Actinobacteriophage (pink) also cluster near each other with most viruses being confined to a few clusters at the bottom of the tree. The tight clustering of the Actinobacteriophage phage is likely explained by the fact that many of the viruses infect a limited number of hosts including Propionibacterium and *Mycobacterium smegmatis* from the SEA-PHAGES program (https://seaphages.org) (*Pope et al., 2011*). In contrast, the Proteobacteria-infecting viruses (green) are clustered in a few locations across the tree, with each cluster showing high levels of local proteomic similarity.

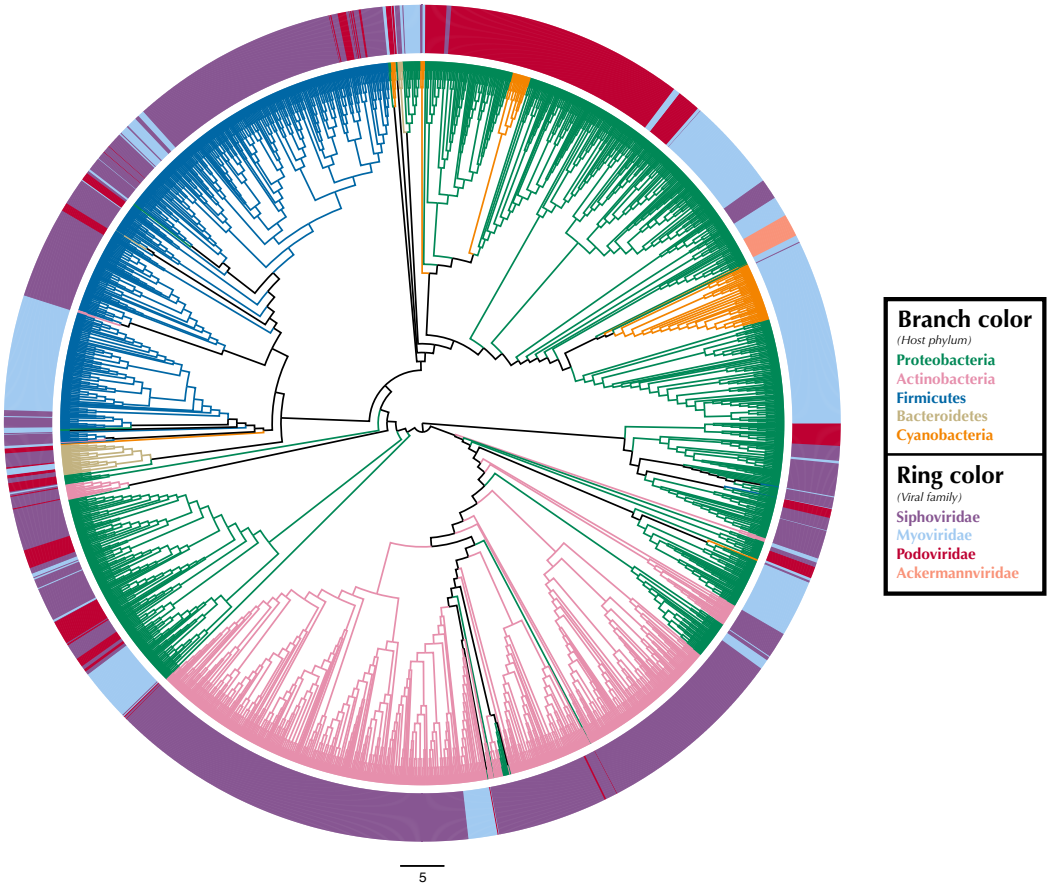

Figure 1 **Proteomic cladogram of viruses from Virus-Host DB.** Proteomic cladogram of viruses infecting Actinobacteria, Bacteroidetes, Cyanobacteria, Firmicutes, and Proteobacteria. Branches are colored by host phylum. Outer ring colors represent virus taxonomic family. Virus-host data is from the Virus-Host DB (*Mihara et al., 2016*).

Full-size DOI: 10.7717/peerj.8584/fig-1

Homology and similarity-based methods have previously been shown to be effective in predicting a phage's host (*Edwards et al., 2016*), perhaps because viruses that infect similar hosts are likely to have more similar genomes (*Villarroel et al., 2016*). Given this and the fact that the proteomic tree clusters viruses based on shared sequence content using homology and multiple sequence alignments (*Rohwer & Edwards, 2002*), it is unsurprising that viruses infecting hosts from the same phylum often cluster near each other on the proteomic tree. In fact, previous studies have used proteomic distance (*Nishimura et al., 2017a*) and other measures of genomic similarity (*Villarroel et al., 2016*) to transfer host annotations from viruses with known hosts to metagenome assembled viral genomes with unknown hosts. In contrast, virus taxonomy is primarily based on multiple phenotypic criteria including virion morphology, host range, and pathogenicity, rather than on genome sequence similarity (*Simmonds, 2015*; *Simmonds et al., 2017*). One study found that for prokaryotic viruses, members of the same taxonomic family (as defined by phenotypic criteria) were divergent and often not detectably homologous in genomic analysis (*Aiewsakun et al.,*

2018). In particular, multiple viral families in the order Caudovirales were interspersed in their dendrograms. Similar results can be seen in Fig. 1, in which several Caudovirales viral families are intermixed in clusters throughout the tree.

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 (GI) tracts of cattle and other ruminants. STEC-contaminated beef and STEC cells 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 sequenced on the Illumina MiSeq platform (*Fadrosh et al., 2014*). The study found that the structure of hide bacterial communities differed between STEC-positive and STEC-negative samples.

To illustrate Iroki's utility for exploring changes in the relative abundance of taxa in conjunction with metadata categories, a subset of cattle hide bacterial operational taxonomic units (OTUs) were selected from the aforementioned study (Supplementary Materials Sec. 2). A Mann–Whitney U test comparing OTU abundance between STEC-positive and STEC-negative samples was performed. Cluster representative sequences from any OTU with a p-value < 0.2 (selected to limit the number of OTUs on the tree and to demonstrate Iroki's features by coloring branches based on test significance) from the Mann–Whitney U test were selected and aligned against SILVA's non-redundant, small subunit ribosomal RNA reference database (SILVA Ref NR) (*Quast et al., 2012*) and an approximate-maximum likelihood tree inferred using SILVA's online Alignment, Classification and Tree (ACT) service (https://www.arb-silva.de/aligner/) (*Pruesse, Glöckner & Peplies, 2012*). Iroki was then used to display various aspects of the data set (Fig. 2; Supplementary Materials Sec. 2). Branches of the tree were colored based on the p-value of the Mann-Whitney U test examining change in relative abundance with STEC contamination (dark green: $p \leq 0.05$, light green: $0.05 < p \leq 0.10$, and gray: $p > 0.10$). Additionally, bar charts representing the log of relative abundance of each OTU (inner bars) and the abundance ratio (outer bars) of OTUs in samples positive and negative for STEC are shown. The color gradient for the inner bar series was generated using Iroki's color gradient generator. Finally, leaf labels show the order and family of the OTU and are colored by predicted OTU phylum using one of the color palettes included in Iroki.

Decorating the tree in this way allows the user to explore the data and look for high-level trends. For example, Firmicutes dominates the tree (e.g., Bacillales, Lactobacillales, Clostridiales). Members of Clostridiales are at low-to-medium relative abundance compared to other OTUs on the tree. Some Clostridiales OTUs (e.g., a majority of the Ruminococcaceae) tend to be at higher abundance in STEC-positive samples, whereas other Clostridiales OTUs, namely those classified as Lachnospiraceae, tend to be at lower abundance in STEC-positive samples. Previous studies have also identified significant positive associations between STEC shedding and Clostridiales OTU abundance in general

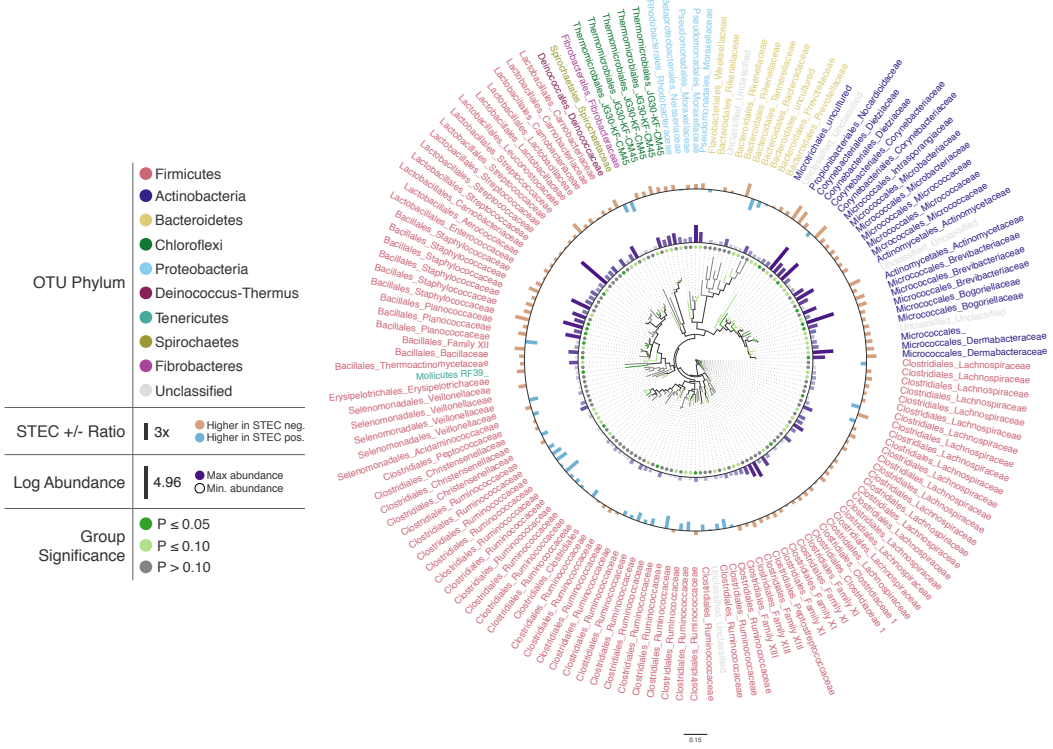

Figure 2 Changes in OTU abundance in two sample groups. Approximate-maximum likelihood tree of cattle hide SSU rRNA OTUs that showed differences in relative abundance between STEC-positive and STEC-negative samples. Branch and leaf dot coloring represents the p-value of a Mann–Whitney U test (dark green: $p \leq 0.05$, light green: $0.05 < p \leq 0.1$, gray: $p > 0.1$) for changes in OTU abundance between STEC-positive samples and STEC-negative samples. Inner bar heights represent log transformed OTU abundance, and outer bars represent the abundance ratio between STEC positive and STEC negative samples (blue bars for higher abundance in STEC positive samples and brown bars for OTUs with higher abundance in STEC negative samples). Taxa labels show the predicted order and family of the OTU and are colored by the predicted phylum using the Paul Tol Muted color palette included with Iroki.

Full-size DOI: 10.7717/peerj.8584/fig-2

(*Zhao et al., 2013*) and Ruminococcus OTUs abundance more specifically (*Zaheer et al., 2017*). In contrast, other studies have found certain Ruminococcus OTUs associated with shedding cattle and other Ruminococcus OTUs associated with non-shedding individuals (*Xu et al., 2014*). Apparent contradictions may be explained by the fact that the various studies were examining the bacterial microbiome associated with different locations on the cow (e.g., GI tract, recto-anal junction, hide). In fact, significant spatial heterogeneity in community composition exists even among different sites along the gastrointestinal tract (*Mao et al., 2015*). Other potential explanations include methodological differences, or that variation associated with STEC presence may be better explained by using more granular groupings than taxa and OTUs (e.g., amplicon sequence variants) (*Callahan, McMurdie & Holmes, 2017*).

In this dataset, more of the OTUs had a higher average relative abundance (brown bars) in STEC-negative samples than in STEC-positive samples (blue bars). Similarly, in a study of the upper and lower gastrointestinal tract microbiome of cattle, a majority of

differentially abundant OTUs were found to be at higher abundance in animals that were not shedding *E. coli* O157:H7 (*Zaheer et al., 2017*). In contrast, another study found that over 75% of differentially expressed OTUs were at greater abundance in STEC shedding cattle (*Xu et al., 2014*).

Tara Oceans viromes

The ribonucleotide reductase (RNR) gene 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*; *Harrison et al., 2019*). As such, it can be used as a marker gene for the study of viral communities. To explore viral communities of the global ocean, we collected RNR proteins from the *Tara* Oceans viral metagenomes (viromes). The *Tara* Oceans expedition was a two-and-a-half year survey that sampled over 200 stations across the world's oceans (*Bork et al., 2015*; *Pesant et al., 2015*). Forty-four viromes were searched for RNRs (Supplementary Materials Sec. 3). Of these, three samples contained fewer than 50 RNRs and were not used in the subsequent analysis. In total, 5,470 RNR sequences across 41 samples were aligned with MAFFT (*Katoh & Standley, 2013*) and post-processed manually to ensure optimal alignment quality. Then, FastTree (*Price, Dehal & Arkin, 2010*) was used to infer a phylogeny from the alignment. Using this tree, the unweighted UniFrac distance (*Lozupone & 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 (Supplementary Materials Sec. 3). Finally, Iroki was used to generate color gradients and add bar charts to visualize the data (Fig. 3). Coloring of the dendrogram with the Viridis color palette (a dark blue, teal, green, yellow sequential color scheme) was based on a 1-dimensional projection of sample conductivity, oxygen, and latitude calculated using Iroki's color gradient generator. The color gradient generator was also used to make the color palettes used for the bar charts.

Coloring the dendrogram based on a projection of the environmental conditions of the samples results in samples with similar environmental metadata being similar in color. For example, the station 66 surface and deep chlorophyll maximum (DCM) samples are nearly identical to one another with respect to conductivity, oxygen, and latitude and have the same dark bluish branch color. In contrast, surface samples from stations 31 and 32 both have a lighter yellowish-green branch color. As the bar charts indicate, these two samples are very similar to one another with respect to the metadata (hence their similar coloring), but are rather different from the station 66 samples in branch color, reflecting the differences in metadata between the two groups.

The combination of dendrogram coloring and bar charts assists in finding trends in the data. Since the dendrogram is based on UniFrac distance between samples based on RNR OTUs, samples that cluster together on the tree have more similar viral communities, according to RNR gene allele content, than samples that are far from one another. In contrast, dendrogram branch coloring and the bar charts show environmental information about the samples themselves (conductivity, oxygen, and latitude). Combining these two

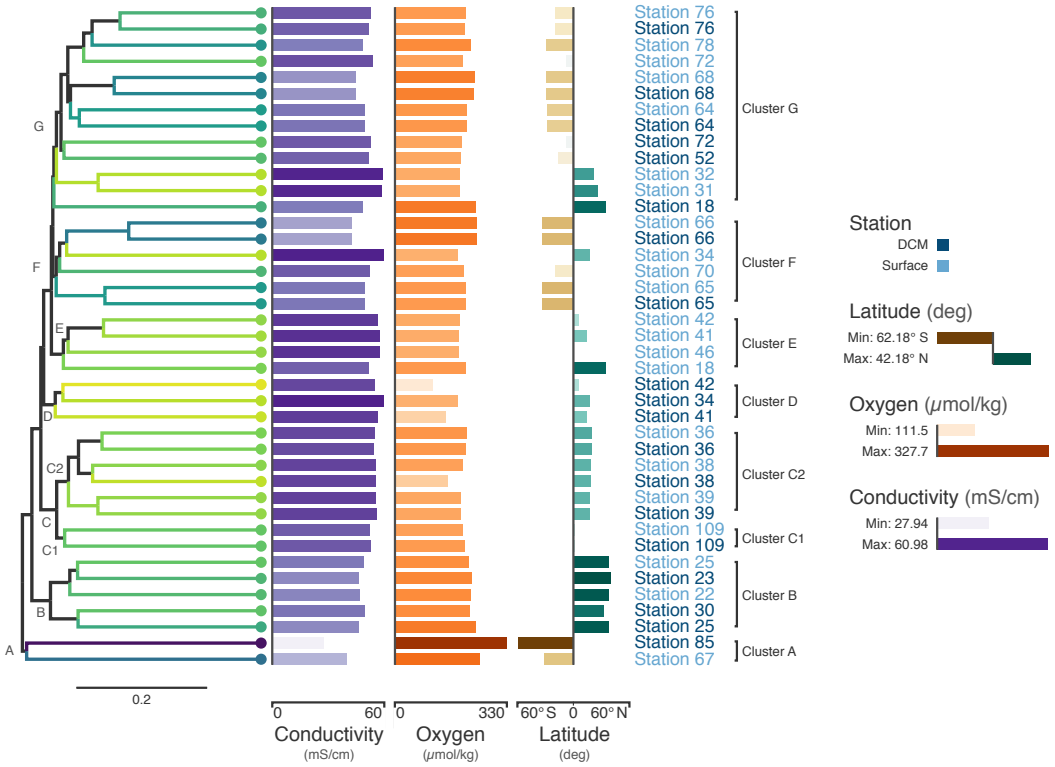

Figure 3 ***Tara* Oceans virome similarity with associated metadata.** Average-linkage hierarchical clustering of sample UniFrac distance based on RNR sequences mined from 41 *Tara* Oceans viromes. Major and sub-clusters of samples (A–G) are labeled. Branch color is based on a scaled, 1-dimensional projection of sample conductivity, oxygen, and latitude onto the Viridis color gradient. Samples that are more similar to each other in branch color represent those that are more similar to each other with respect to the environmental parameters in the ordination. The first bar series (purple) represents sample conductivity (mS/cm), the second bar series (orange) represents sample dissolved oxygen levels (μmol/kg), and the third bar series (brown/green) represents sample latitude (degrees). For the first two bar series, shorter bars with lighter colors indicate lower values, while longer bars with darker colors indicate higher values. For the third series, longer, dark brown bars indicate samples with extreme negative latitudes, whereas longer, dark blue bars indicate samples with extreme positive latitudes. Samples with intermediate latitudes are represented by shorter, light colored bars. Sample labels represent the station from which the virome was acquired and are colored by sampling depth, with light blue representing surface samples and dark blue representing samples from the deep chlorophyll maximum at that station.

Full-size DOI: 10.7717/peerj.8584/fig-3

aspects of the samples enables visualization of the relationship between the similarity of RNR-containing viral communities and the environments in which they are found.

For example, the samples in the bottom half of the tree are, in general, from northern latitudes, whereas samples towards the top tend to be from southern latitudes. In a previous study of the T4-like viral communities of Polar freshwater lakes, no significant correlation between latitude and viral community diversity was found in the Antarctic samples (*Daniel et al., 2016*). Though the Arctic lakes were not tested among themselves for significant associations between latitude and viral community richness (presumably due to the small latitudinal variation in Arctic sampling locations), Arctic and Antarctic lakes were tested against one another; however, no significant difference in viral diversity was seen with

respect to pole of origin. The Antarctic samples from the study ranged from 67.84°S to 62.64°S, whereas the *Tara* Oceans viromes used to build the tree in Fig. 3 ranged from 62.18°S to 41.18°N. The increased range of samples from the *Tara* survey may have enabled this shift in diversity to be detected. Additionally, the previous study used *g23*, the gene for major capsid protein, to survey the viral community. It is possible that a functional protein like RNR is more connected with environmental conditions than a structural protein such as the T4-like major capsid protein. RNRs reduce ribonucleotides, the rate-limiting step of DNA synthesis (*Kolberg et al., 2004*; *Ahmad et al., 2012*). There are several different types of RNR, each with specific biochemical mechanisms and nutrient requirements (*Nordlund & Reichard, 2006*). Accordingly, the type of RNR carried by a cell or virus often reflects the environmental conditions in which DNA replication occurs (*Reichard, 1993*; *Cotruvo & Stubbe, 2011*; *Sakowski et al., 2014*; *Srinivas et al., 2018*; *Harrison et al., 2019*). A survey based on RNR, then, may provide more sensitivity in detecting environmental effects on viral community structure. A significant relationship between T4-like viral communities and bacterial assemblages was found however (*Daniel et al., 2016*), and numerous other studies have reported a significant relationship between bacterial community diversity and latitude (e.g., *Ladau et al. (2013)*; *Raes et al. (2018)*). Thus, latitudinal variation in bacterial communities is likely linked to viral community variation.

Certain clusters have been marked on the tree for further analysis. Cluster A (Station 85 DCM, Station 67 surface) contains the samples with the most divergent RNR-containing viral populations (Fig. 3) according to the dendrogram. Station 85 DCM is also the sample with the lowest conductivity, highest dissolved oxygen, and most southerly latitude, suggesting that the divergent conditions of the sample with respect to the other included samples could be influencing the divergent RNR-containing viral population. Clusters B and C also offer a good point of comparison (Fig. 3). In addition to the similarity of their RNR-containing viral populations, samples in cluster B have highly similar conductivity, oxygen, and latitude (as shown by their highly similar branch color and bar charts), suggesting a close connection between sample composition and viral population. Cluster C is separate from cluster B on the dendrogram, implying their RNR-containing viral populations are less similar. The sample metadata between the two clusters is less similar as well, with Cluster B having on average a lower conductivity and higher dissolved oxygen content than samples from cluster C.

Connections between viral community composition and environment have been seen before. Salinity, which can be estimated from measurements of electrical conductivity (*Pawlowicz, 2012*; *Pawlowicz, 2019*), has been shown to affect viral-host interactions. In a viral-host system of halovirus SNJ1 with its host, *Natrinema* sp. J7-2, viral adsorption rates and lytic/lysogenic rates were measured at varying salt concentrations. Adsorption and lytic rate were found to increase with salt concentration, whereas the lysogenic rate decreased (*Mei et al., 2015*). In a system of tropical coastal lagoons, salinity was found to be one of the main factors positively affecting viral abundance (*Junger et al., 2018*). Viral community structure has also been associated with shifts in salinity in various environments (*Bettarel et al., 2011*; *Emerson et al., 2013*; *Winter, Matthews & Suttle, 2013*; *Finke & Suttle, 2019*).

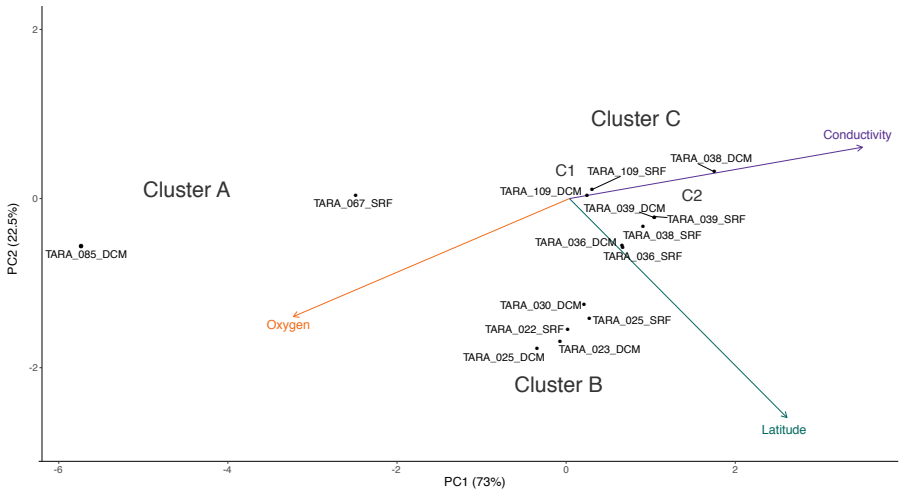

Figure 4 PCA biplot of *Tara* Oceans virome clusters A, B, and C. Principal components analysis biplot of *Tara* Oceans viromes based on sample oxygen, conductivity, and latitude. Ordination was done on all viromes, but only those from clusters A, B, and C are shown here for clarity.

Full-size DOI: 10.7717/peerj.8584/fig-4

These shifts likely effect a change in the host communities, which is reflected in the shifts in viral communities.

Cluster C can be further divided into two clusters, C1 and C2. While the samples in C1 are closer to those in C2 than to those in cluster B in terms of their RNR-carrying viral populations, the samples in C1 are more similar to the samples in cluster B with respect to their metadata projection. The similar branch coloring between samples in clusters B and C1, despite their large differences in latitude, occurs because more of the variation the first principal component (the principal component on which the Viridis coloring is based) is explained by conductivity and oxygen than by latitude (Fig. 4; full ordination: Fig. S1). More striking examples can be found elsewhere in the tree. For example, station 66 surface, station 66 DCM, and station 34 surface cluster together on the dendrogram based on viral community similarity (cluster F), but the conductivity, oxygen, and latitude values for sample 34 surface are quite different from the station 66 samples. Thus, while these three metadata categories were significantly correlated with sample UniFrac distance, other factors also play a role in shaping the viral communities. Overall, using Iroki to add color and bar charts based on environmental metadata to the dendrogram based on RNR-carrying viral community structure helps visualize that high-level viral community structure can be influenced by the environmental parameters of the sample in which they originate.

CONCLUSIONS

Iroki is a web application for fast, automatic customization and visualization of large phylogenetic trees based on user specified, tab-delimited configuration files with categorical and numeric metadata. Through the use of simple configuration files, Iroki provides a 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. While Iroki includes many key features, future work is planned to increase its utility. There is no mechanism within Iroki to handle rerooting trees. As such, users must use an external program to reroot their tree before viewing it in Iroki. Customizing the tree is mainly handled by modifying the mapping file, however, Iroki could be made more interactive by allowing the user to edit certain aspects of the tree "by-hand" without having to reupload a new mapping file. Currently, Iroki allows editing leaf labels after a tree is submitted. More interactive features, such as editing label and branch styles, are planned for a future release. Finally, bringing the full feature set of Iroki's SVG viewer to the Canvas viewer will allow users to visualize and customize huge trees quickly and easily.

Various example datasets from microbial ecology studies were analyzed to demonstrate Iroki's utility. Iroki simplified the processes of data exploration and presentation by facilitating the mapping of various aspects of the data directly on the tree. 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.

ACKNOWLEDGEMENTS

We would like to acknowledge Barbra D. Ferrell for editing the manuscript, and the reviewers for their constructive feedback. This content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

ADDITIONAL INFORMATION AND DECLARATIONS

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, the National Science Foundation Advances in Biological Informatics program (award number DBI-1356374), the National Science Foundation Grant No. 1736030, the Established Program to Stimulate Competitive Research (award number OIA-1736030) from the Office of Integrated Activities, and a Doctoral Fellowship provided by University of Delaware in conjunction with the Unidel Foundation. Computational infrastructure support by the University of Delaware Center for Bioinformatics and Computational Biology Core Facility was made possible through funding from the Delaware Biotechnology Institute, and the Delaware INBRE program with a grant from the National Institute of General Medical Sciences (NIGMS P20 GM103446) from the National Institutes of Health and the State of Delaware. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
USDA Agriculture and Food Research Initiative: 2012-68003-30155.
National Science Foundation Advances in Biological Informatics program: DBI-1356374.

National Science Foundation: 1736030.
The Established Program to Stimulate Competitive Research: OIA-1736030.
University of Delaware.
UNIDEL foundation.
Delaware Biotechnology Institute.
Delaware INBRE program.
National Institute of General Medical Sciences: NIGMS P20 GM103446.
National Institutes of Health.
The State of Delaware.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Ryan M. Moore conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Amelia O. Harrison, Sean M. McAllister, Shawn W. Polson and K. Eric Wommack conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The Iroki web app is available at https://www.iroki.net. The Iroki source code is available under the MIT license at https://github.com/mooreryan/iroki.

Data and scripts used to generate the figures are available on Zenodo: Ryan M. Moore, Amelia O. Harrison, Sean M. McAllister, Shawn W. Polson, & K. Eric Wommack. (2019). Manuscript data for "Iroki: automatic customization and visualization of phylogenetic trees" [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3458510.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.8584#supplemental-information.

REFERENCES

Ahmad MF, Kaushal PS, Wan Q, Wijerathna SR, An X, Huang M, Dealwis CG. 2012. Role of arginine 293 and glutamine 288 in communication between catalytic and allosteric sites in yeast ribonucleotide reductase. *Journal of Molecular Biology* **419(5)**:315–329 DOI 10.1016/j.jmb.2012.03.014.

Aiewsakun P, Adriaenssens EM, Lavigne R, Kropinski AM, Simmonds P. 2018. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. *The Journal of General Virology* **99(9)**:1331–1343 DOI 10.1099/jgv.0.001110.

Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**:e1029 DOI 10.7717/peerj.1029.

Bachmaier C, Brandes U, Schlieper B. 2005. Drawing phylogenetic trees. (Extended abstract). In: Deng X, Du D-Z, eds. *ISAAC: 16th international symposium on algorithms and computation, vol. 3827*. Springer, 1110–1121 DOI 10.1007/11602613.

Bellas CM, Anesio AM, Barker G. 2015. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Frontiers in Microbiology* **6(JUL)**:656 DOI 10.3389/fmicb.2015.00656.

Bennett DJ, Sutton MD, Turvey ST. 2017. treeman: an R package for efficient and intuitive manipulation of phylogenetic trees. *BMC Research Notes* **10(1)**:30 DOI 10.1186/s13104-016-2340-8.

Bettarel Y, Bouvier T, Bouvier C, Carré C, Desnues A, Domaizon I, Jacquet S, Robin A, Sime-Ngando T. 2011. Ecological traits of planktonic viruses and prokaryotes along a full-salinity gradient. *FEMS Microbiology Ecology* **76(2)**:360–372 DOI 10.1111/j.1574-6941.2011.01054.x.

Bork P, Bowler C, De Vargas C, Gorsky G, Karsenti E, Wincker P. 2015. Tara oceans studies plankton at planetary scale. *Science* **348(6237)**:873 DOI 10.1126/science.aac5605.

Bostock M, Ogievetsky V, Heer J. 2011. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics* **17(12)**:2301–2309 DOI 10.1109/TVCG.2011.185.

Brewer C, Harrower M, University TPS. 2013. ColorBrewer2. *Available at http://colorbrewer2.org*.

Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11**:2639–2643 DOI 10.1038/ismej.2017.119.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pea AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7(5)**:335–336 DOI 10.1038/nmeth.f.303.

Caprioli A, Morabito S, Brugre H, Oswald E. 2005. Enterohaemorrhagic Escherichia coli: emerging issues on virulence and modes of transmission. *Veterinary Research* **36(3)**:289–311 DOI 10.1051/vetres:2005002.

Chen W-H, Lercher MJ. 2009. ColorTree: a batch customization tool for phylogenic trees. *BMC Research Notes* **2(1)**:155 DOI 10.1186/1756-0500-2-155.

Chopyk J, Moore RM, DiSpirito Z, Stromberg ZR, Lewis GL, Renter DG, Cernicchiaro N, Moxley RA, Wommack KE. 2016. Presence of pathogenic Escherichia coli is correlated with bacterial community diversity and composition on pre-harvest cattle hides. *Microbiome* **4(1)**:9 DOI 10.1186/s40168-016-0155-4.

Cotruvo JA, Stubbe J. 2011. Class I Ribonucleotide reductases: metallocofactor assembly and repair in vitro and in vivo. *Annual Review of Biochemistry* **80(1)**:733–767 DOI 10.1146/annurev-biochem-061408-095817.

Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, Dutilh BE, Thompson FL. 2017. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications* **8(May)**:1–12 DOI 10.1038/ncomms15955.

Czech L, Huerta-Cepas J, Stamatakis A. 2017. A critical review on the use of support values in tree viewers and bioinformatics toolkits. *Molecular Biology and Evolution* **34(6)**:1535–1542 DOI 10.1093/molbev/msx055.

Daniel ADC, Pedrós-Alió C, Pearce DA, Alcamí A. 2016. Composition and interactions among bacterial, microeukaryotic, and T4-like viral assemblages in lakes from both polar zones. *Frontiers in Microbiology* **7**:337–337 DOI 10.3389/fmicb.2016.00337.

Dwivedi B, Xue B, Lundin D, Edwards RA, Breitbart M. 2013. A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evolutionary Biology* **13(1)**:33 DOI 10.1186/1471-2148-13-33.

Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2016. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews* **40(2)**:258–272 DOI 10.1093/femsre/fuv048.

Emerson JB, Thomas BC, Andrade K, Heidelberg KB, Banfield JF. 2013. New approaches indicate constant viral diversity despite shifts in assemblage structure in an Australian Hypersaline Lake. *Applied and Environmental Microbiology* **79(21)**:6755–6764 DOI 10.1128/AEM.01946-13.

Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, Ravel J. 2014. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* **2(1)**:6 DOI 10.1186/2049-2618-2-6.

Finke JF, Suttle CA. 2019. The environment and cyanophage diversity: insights from environmental sequencing of DNA polymerase. *Frontiers in Microbiology* **10**:167 DOI 10.3389/fmicb.2019.00167.

Green DA. 2011. A colour scheme for the display of astronomical intensity images. *Bulletin of the Astronomical Society of India* **39(2)**:289–295 DOI 10.1109/CVPR.2015.7298594.

Han MV, Zmasek CM. 2009. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10(1)**:356 DOI 10.1186/1471-2105-10-356.

Hancock DD, Besser TE, Kinsel ML, Tarr PI, Rice DH, Paros MG. 1994. The prevalence of Escherichia coli O157.H7 in dairy and beef cattle in Washington State. *Epidemiology and Infection* **113(2)**:199–207 DOI 10.1017/S0950268800051633.

Hanson-Smith V, 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 DOI 10.1371/journal.pcbi.1004976.

Harrison AO, Moore RM, Polson SW, Wommack KE. 2019. Reannotation of the ribonucleotide reductase in a cyanophage reveals life history strategies within the virioplankton. *Frontiers in Microbiology* **10**:134 DOI 10.3389/fmicb.2019.00134.

He Z, Zhang H, Gao S, Lercher MJ, Chen WH, Hu S. 2016. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Research* **44(W1)**:W236–W241 DOI 10.1093/nar/gkw370.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution* **33(6)**:1635–1638 DOI 10.1093/molbev/msw046.

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. 2007. Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8(1)**:460 DOI 10.1186/1471-2105-8-460.

Junger PC, Amado AM, Paranhos R, Cabral AS, Jacques SMS, Farjalla VF. 2018. Salinity drives the virioplankton abundance but not production in tropical coastal lagoons. *Microbial Ecology* **75(1)**:52–63 DOI 10.1007/s00248-017-1038-3.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30(4)**:772–780 DOI 10.1093/molbev/mst010.

Kolberg M, Strand KR, Graff P, Kristoffer Andersson K. 2004. Structure, function, and mechanism of ribonucleotide reductases. *Biochimica et Biophysica Acta (BBA)—Proteins and Proteomics* **1699(1)**:1–34 DOI 10.1016/j.bbapap.2004.02.007.

Kreft L, Botzki A, Coppens F, Vandepoele K, Van Bel M. 2017. PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics* **33(18)**:2946–2947 DOI 10.1093/bioinformatics/btx324.

Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J, Koeppel AF, Green JL, Pollard KS. 2013. Global marine bacterial diversity peaks at high latitudes in winter. *The Isme Journal* **7**:1669–1677 DOI 10.1038/ismej.2013.37.

Lan Y, Rosen G, 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 DOI 10.1186/s40168-016-0162-5.

Larkin AA, Blinebry SK, Howes C, Lin Y, Loftus SE, Schmaus CA, Zinser ER, Johnson ZI. 2016. Niche partitioning and biogeography of high light adapted Prochlorococcus across taxonomic ranks in the North Pacific. *The ISME Journal* **10**:1555–1567 DOI 10.1038/ismej.2015.244.

Lauer F. 2017. MLweb: a toolkit for machine learning on the web. *Neurocomputing* **282**:74–77 DOI 10.1016/j.neucom.2017.11.069.

Letunic I, 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 DOI 10.1093/nar/gkw290.

Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71(12)**:8228–8235 DOI 10.1128/AEM.71.12.8228-8235.2005.

Mao S, Zhang M, Liu J, Zhu W. 2015. Characterising the bacterial microbiota across the gastrointestinal tracts of dairy cattle: membership and potential function. *Scientific Reports* **5**:16116 DOI 10.1038/srep16116.

McDonald D, Clemente JC, Kuczynski J, Rideout J, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, 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 DOI 10.1186/2047-217X-1-7.

Mei Y, He C, Huang Y, Liu Y, Zhang Z, Chen X, Shen P. 2015. Salinity regulation of the interaction of halovirus SNJ1 with its host and alteration of the halovirus replication strategy to adapt to the variable ecosystem. *PLOS ONE* **10**(**4**):e0123874 DOI 10.1371/journal.pone.0123874.

Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P, Goto S, Ogata H. 2016. Linking virus genomes with host taxonomy. *Viruses* **8**(**3**):66–66 DOI 10.3390/v8030066.

Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. *PLOS Genetics* **9**(**12**):1–13 DOI 10.1371/journal.pgen.1003987.

Munson-McGee JH, Peng S, Dewerff S, Stepanauskas R, Whitaker RJ, Weitz JS, Young MJ. 2018. A virus or more in (nearly) every cell: ubiquitous networks of virus—host interactions in extreme environments. *The ISME Journal* **12**(**7**):1706–1714 DOI 10.1038/s41396-018-0071-7.

Müller AL, Kjeldsen KU, Rattei T, Pester M, Loy A. 2015. Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi)sulfite reductases. *The ISME Journal* **9**(**5**):1152–1165 DOI 10.1038/ismej.2014.208.

Nelson D. 2004. Phage taxonomy: we agree to disagree. *Journal of Bacteriology* **186**(**21**):7029–7031 DOI 10.1128/JB.186.21.7029-7031.2004.

Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, Blanc-Mathieu R, Yamamoto K, Hingamp P, Sako Y, Sullivan MB, Goto S, Ogata H, Yoshida T, Viral E, Shed G, Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, Blanc-Mathieu R, Yamamoto K, Hingamp P, Sako Y, Sullivan MB, Goto S, Ogata H, Yoshida T. 2017a. Environmental viral genomes shed new light on virus-host interactions in the ocean. *mSphere* **2**(**2**):e00359-16 DOI 10.1128/mSphere.00359-16.

Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. 2017b. ViPTree: the viral proteomic tree server. *Bioinformatics* **33**(**15**):2379–2380 DOI 10.1093/bioinformatics/btx157.

Nordlund P, Reichard P. 2006. Ribonucleotide reductases. *Annual Review of Biochemistry* **75**(**1**):681–706 DOI 10.1146/annurev.biochem.75.103004.142443.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**(**2**):289–290 DOI 10.1093/bioinformatics/btg412.

Pawlowicz R. 2012. The electrical conductivity of seawater at high temperatures and salinities. *Desalination* **300**:32–39 DOI 10.1016/j.desal.2012.06.001.

Pawlowicz R. 2019. Electrical properties of sea water: theory and applications. In: Cochran JK, Bokuniewicz HJ, Yager PL, eds. *Encyclopedia of ocean sciences*. Third Edition. Oxford: Academic Press, 71–80.

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 TOC, Acinas

SG, 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 EG, Sardet C, Sieracki M, Speich S, Stemmann L, Sullivan MB, Sunagawa S, Velayoudon D, Weissenbach J, Wincker P. 2015. Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data* **2**:150023 DOI 10.1038/sdata.2015.23.

Pielou EC. 1966. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology* **13(C)**:131–144 DOI 10.1016/0022-5193(66)90013-0.

Pope WH, Jacobs-Sera D, Russell DA, Peebles CL, Al-Atrache Z, Alcoser TA, Alexander LM, Alfano MB, Alford ST, Amy NE, Anderson MD, Anderson AG, Ang AAS, Ares Jr M, Barber AJ, Barker LP, Barrett JM, Barshop WD, Bauerle CM, Bayles IM, Belfield KL, Best AA, Borjon Jr A, Bowman CA, Boyer CA, Bradley KW, Bradley VA, Broadway LN, Budwal K, Busby KN, Campbell IW, Campbell AM, Carey A, Caruso SM, Chew RD, Cockburn CL, Cohen LB, Corajod JM, Cresawn SG, Davis KR, Deng L, Denver DR, Dixon BR, Ekram S, Elgin SCR, Engelsen AE, English BEV, Erb ML, Estrada C, Filliger LZ, Findley AM, Forbes L, Forsyth MH, Fox TM, Fritz MJ, Garcia R, George ZD, Georges AE, Gissendanner CR, Goff S, Goldstein R, Gordon KC, Green RD, Guerra SL, Guiney-Olsen KR, Guiza BG, Haghighat L, Hagopian GV, Harmon CJ, Harmson JS, Hartzog GA, Harvey SE, He S, He KJ, Healy KE, Higinbotham ER, Hildebrandt EN, Ho JH, Hogan GM, Hohenstein VG, Holz NA, Huang VJ, Hufford EL, Hynes PM, Jackson AS, Jansen EC, Jarvik J, Jasinto PG, Jordan TC, Kasza T, Katelyn MA, Kelsey JS, Kerrigan LA, Khaw D, Kim J, Knutter JZ, Ko C-C, Larkin GV, Laroche JR, Latif A, Leuba KD, Leuba SI, Lewis LO, Loesser-Casey KE, Long CA, Lopez AJ, Lowery N, Lu TQ, Mac V, Masters IR, McCloud JJ, McDonough MJ, Medenbach AJ, Menon A, Miller R, Morgan BK, Ng PC, Nguyen E, Nguyen KT, Nguyen ET, Nicholson KM, Parnell LA, Peirce CE, Perz AM, Peterson LJ, Pferdehirt RE, Philip SV, Pogliano K, Pogliano J, Polley T, Puopolo EJ, Rabinowitz HS, Resiss MJ, Rhyan CN, Robinson YM, Rodriguez LL, Rose AC, Rubin JD, Ruby JA, Saha MS, Sandoz JW, Savitskaya J, Schipper DJ, Schnitzler CE, Schott AR, Segal JB, Shaffer CD, Sheldon KE, Shepard EM, Shepardson JW, Shroff MK, Simmons JM, Simms EF, Simpson BM, Sinclair KM, Sjoholm RL, Slette IJ, Spaulding BC, Straub CL, Stukey J, Sughrue T, Tang T-Y, Tatyana LM, Taylor SB, Taylor BJ, Temple LM, Thompson JV, Tokarz MP, Trapani SE, Troum AP, Tsay J, Tubbs AT, Walton JM, Wang DH, Wang H, Warner JR, Weisser EG, Wendler SC, Weston-Hafer KA, Whelan HM, Williamson KE, Willis AN, Wirtshafter HS, Wong TW, Wu P, Yang YJ, Yee BC, Zaidins DA, Zhang B, Zúniga MY, Hendrix RW, Hatfull GF. 2011. Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. *PLOS ONE* **6(1)**:e16329 DOI 10.1371/journal.pone.0016329.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5(3)**:e9490 DOI 10.1371/journal.pone.0009490.

Pruesse E, Glöckner FO, Peplies J. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28(14)**:1823–1829 DOI 10.1093/bioinformatics/bts252.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41(D1)**:D590–D596 DOI 10.1093/nar/gks1219.

Raes EJ, Bodrossy L, van de Kamp J, Bissett A, Waite AM. 2018. Marine bacterial richness increases towards higher latitudes in the eastern Indian Ocean. *Limnology and Oceanography Letters* **3(1)**:10–19 DOI 10.1002/lol2.10058.

Rambaut A. 2006. FigTree. *Available at http://tree.bio.ed.ac.uk/software/figtree/*.

Reichard P. 1993. From RNA to DNA, why so many ribonucleotide reductases? *Science* **260(5115)**:1773–1777 DOI 10.1126/science.8511586.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3(2)**:217–223 DOI 10.1111/j.2041-210X.2011.00169.x.

Robinson O, Dylus D, Dessimoz C. 2016. Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. *Molecular Biology and Evolution* **33(8)**:2163–2166 DOI 10.1093/molbev/msw080.

Rohwer F, Edwards R. 2002. The phage proteomic tree: a genome-based taxonomy for phage. *Journal of Bacteriology* **184(16)**:4529–4535 DOI 10.1128/JB.184.16.4529-4535.2002.

Rohwer F, Thurber RV. 2009. Viruses manipulate the marine environment. *Nature* **459(7244)**:207–212 DOI 10.1038/nature08060.

Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaqué D, Bork P, Acinas SG, Wincker P, Sullivan MB. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537(7622)**:689–693 DOI 10.1038/nature19366.

Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* **4**:1–20 DOI 10.7554/eLife.08490.001.

Sakowski EG, Munsell EV, Hyatt M, Kress W, Williamson SJ, Nasko DJ, Polson SW, Wommack KE. 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)**:15786–15791 DOI 10.1073/pnas.1401322111.

Santamaría R, Therón R. 2009. Treevolution: visual analysis of phylogenetic trees. *Bioinformatics* **25(15)**:1970–1971 DOI 10.1093/bioinformatics/btp333.

Simister RL, Deines P, Bott ES, Webster NS, Taylor MW. 2012. Sponge-specific clusters revisited: a comprehensive phylogeny of sponge-associated microorganisms. *Environmental Microbiology* **14(2)**:517–524 DOI 10.1111/j.1462-2920.2011.02664.x.

Simmonds P. 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *Journal of General Virology* **96(6)**:1193–1206 DOI 10.1099/jgv.0.000016.

Simmonds P, Adams MJ, Benk M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, Hull R, King AM, Koonin EV, Krupovic M, Kuhn JH, Lefkowitz EJ, Nibert ML, Orton R, Roossinck MJ, Sabanadzovic S, Sullivan MB, Suttle CA, Tesh RB, Van der Vlugt RA, Varsani A, Zerbini FM. 2017. Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology* **15**:161–168 DOI 10.1038/nrmicro.2016.177.

Srinivas V, Lebrette H, Lundin D, Kutin Y, Sahlin M, Lerche M, Eirich J, Branca RMM, Cox N, Sjöberg B-M, Högbom M. 2018. Metal-free ribonucleotide reduction powered by a DOPA radical in Mycoplasma pathogens. *Nature* **563(7731)**:416–420 DOI 10.1038/s41586-018-0653-6.

Stöver BC, Müller KF. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* **11**:7 DOI 10.1186/1471-2105-11-7.

Stöver BC, Wiechers S, Müller KF. 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. *Available at http://bioinfweb.info/JPhyloIO/*.

Suttle CA. 2007. Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology* **5(10)**:801–812 DOI 10.1038/nrmicro1750.

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, Dopazo H. 2011. Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Research* **39**:470–474 DOI 10.1093/nar/gkr408.

Talevich E, Invergo BM, Cock PJ, Chapman BA. 2012. Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics* **13**:209 DOI 10.1186/1471-2105-13-209.

Vaughan TG. 2017. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics* **33(15)**:2392–2394 DOI 10.1093/bioinformatics/btx155.

Villarroel J, Kleinheinz AK, Jurtz IV, Zschach H, Lund O, Nielsen M, Larsen VM, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, Nielsen M, Larsen MV. 2016. HostPhinder: a phage host prediction tool. *Viruses* **8(5)**:1–22 DOI 10.3390/v8050116.

Vos RA, Caravas J, Hartmann K, Jensen MA, Miller C. 2011. BIO::phylo-phyloinformatic analysis using perl. *BMC Bioinformatics* **12**:63 DOI 10.1186/1471-2105-12-63.

Winter C, Matthews B, Suttle CA. 2013. Effects of environmental variation and spatial distance on Bacteria, Archaea and viruses in sub-polar and arctic waters. *The Isme Journal* **7**:1507–1518 DOI 10.1038/ismej.2013.56.

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ. 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences* **6(3)**:421–433 DOI 10.4056/sigs.2945050.

Wommack KE, Nasko DJ, Chopyk J, Sakowski EG. 2015. Counts and sequences, observations that continue to change our understanding of viruses in nature. *Journal of Microbiology* **53**(**3**):181–192 DOI 10.1007/s12275-015-5068-6.

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, 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 DOI 10.1038/ismej.2015.138.

Xu Y, Dugat-Bony E, Zaheer R, Selinger L, Barbieri R, Munns K, McAllister TA, Selinger LB. 2014. Escherichia coli O157:H7 super-shedder and non-shedder feedlot steers harbour distinct fecal bacterial communities. *PLOS ONE* **9**(**5**):e98115 DOI 10.1371/journal.pone.0098115.

Zaheer R, Dugat-Bony E, Holman D, Cousteix E, Xu Y, Munns K, Selinger LJ, Barbieri R, Alexander T, McAllister TA, Selinger LB. 2017. Changes in bacterial community composition of Escherichia coli O157:H7 super-shedder cattle occur in the lower intestine. *PLOS ONE* **12**(**1**):e0170050–e0170050 DOI 10.1371/journal.pone.0170050.

Zhao L, Tyler P, Starnes J, Bratcher C, Rankins D, McCaskey T, Wang L. 2013. Correlation analysis of Shiga toxin-producing Escherichia coli shedding and faecal bacterial composition in beef cattle. *Journal of Applied Microbiology* **115**(**2**):591–603 DOI 10.1111/jam.12250.