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, 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 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`.

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 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, 2006), 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 application programming interface (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 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 and Bork, 2016), PhyD3 (Kreft et al., 2017), Phylemon (Sánchez et al., 2011), PhyloBot (Hanson-Smith and Johnson, 2016), Phylo.io (Robinson et al., 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 files. 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 et al., 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 et al. (2005) was implemented in JavaScript. In addition to the SVG based tree viewer, Iroki also includes an HTML5 Canvas based 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 leaf label color, font, size, and name, leaf dot color and size, branch width and color, as well as bar charts and arcs. In addition to automatic customization using a mapping file, 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 et al., 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 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, each 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:

$$E_i = H_i / H_{\max}, \tag{1}$$

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

$$E_i = \frac{-\sum_{j=1}^{N} p_{ij} \log_2 (p_{ij})}{\log_2 (N)}, \tag{2}$$

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) count matrix X, with observations as rows and variables as columns, the following decomposition is obtained:

$$X = USV^T, \tag{3}$$

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. In this way, the color gradient matches the first principal component, which maximizes the data variance.

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 and 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-spacer (Roux et al., 2016; Coutinho et al., 2017; Nishimura et al., 2017a) and tRNA matches (Bellas et al., 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, 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 and 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 and 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.

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 and 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. This was especially true when considering members of the Caudovirales, which make up all the phage we included in our analysis (Aiewsakun et al., 2018). Similar trends can be seen in Fig. 1, in which multiple viral families as defined by tail morphology are found in the same cluster on 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 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 et al., 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 (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 et al., 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 deferentially expressed OTUs were at greater abundance in STEC *E. coli* 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 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 (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 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 and 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 and 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)), 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, 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 et al., 2013; Finke and Suttle, 2019). 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 in 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: Supplementary Figure 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. Various example datasets from microbial ecology studies were analyzed to demonstrate Iroki's utility. In each case, Iroki simplified the processes of data exploration and presentation. 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.

ADDITIONAL INFORMATION AND DECLARATIONS

Availability of data and materials

Data used to generate figures for this manuscript are available for download on Zenodo at the following URL: https://doi.org/10.5281/zenodo.3458510.

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. This content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Acknowledgments

We would like to acknowledge Barbra D. Ferrell for editing the manuscript.

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. KEW and SWP guided the project and edited the manuscript. All authors read, edited, and approved the final manuscript.

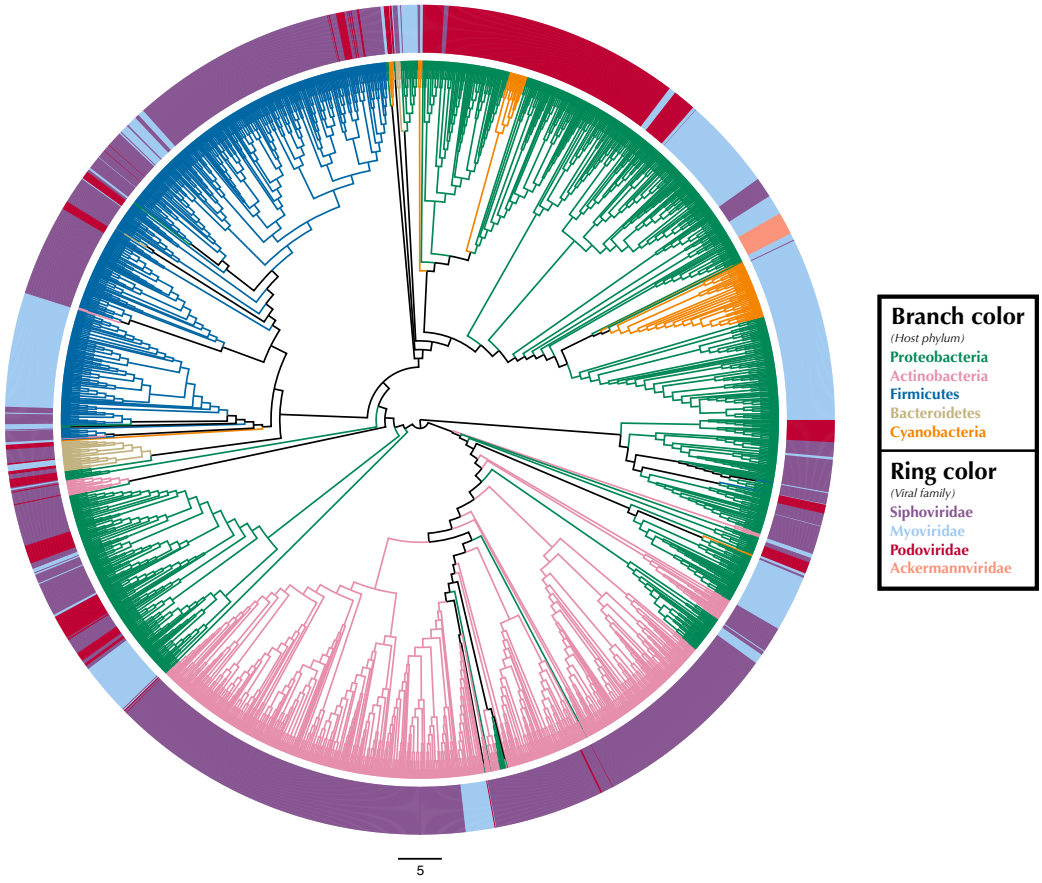

Figure 1. Proteomic cladogram of viruses from Virus-Host DB. Proteomic cladogram of viruses infecting Actinobacteria, Bacteroidetes, Cyanobacteria, Firmicutes, and Proteobacteria from the Virus-Host DB (Mihara et al., 2016). Branches are colored by host phylum. Outer ring colors represent virus taxonomic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

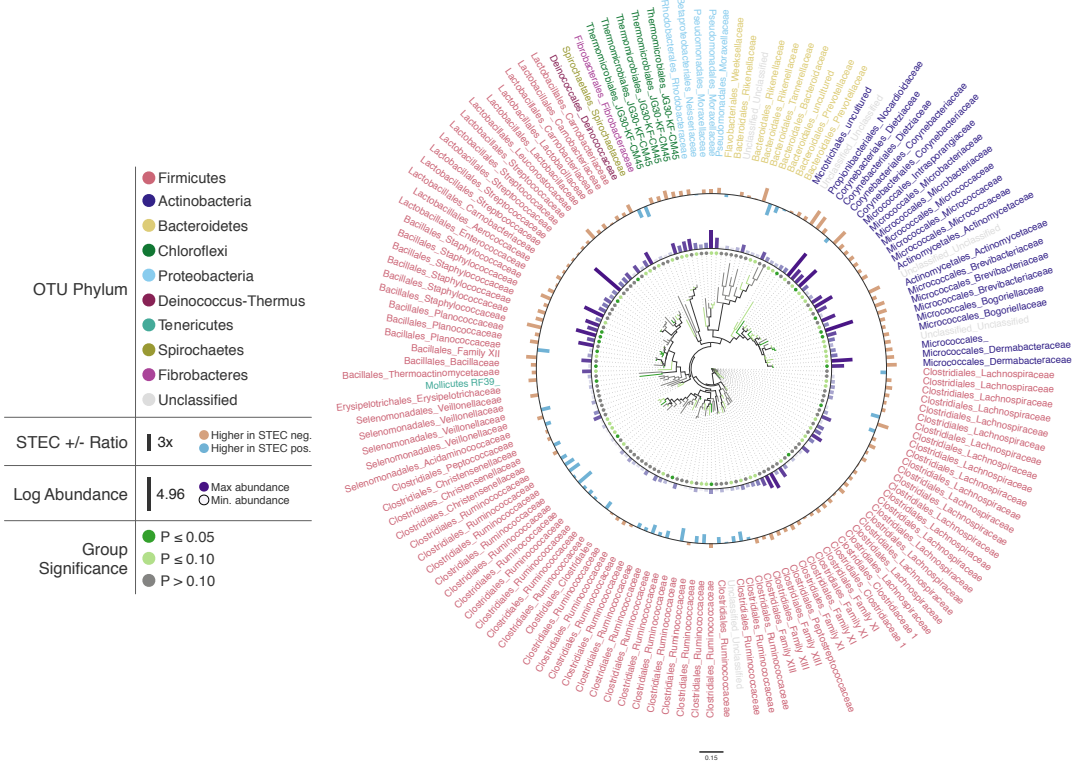

Figure 2. Changes in OTU abundance in two sample groups. Approximate-maximum likelihood tree of hide SSU rRNA OTUs that showed differences in relative abundance between STEC positive and STEC negative cattle hide samples. Branch and leaf dot coloring represents the p-value of a Mann-Whitney U test (dark green: $p \leq 0.05$, light green: $0.05 < p \leq 0.1$, gray: $p > 0.1$) testing 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.

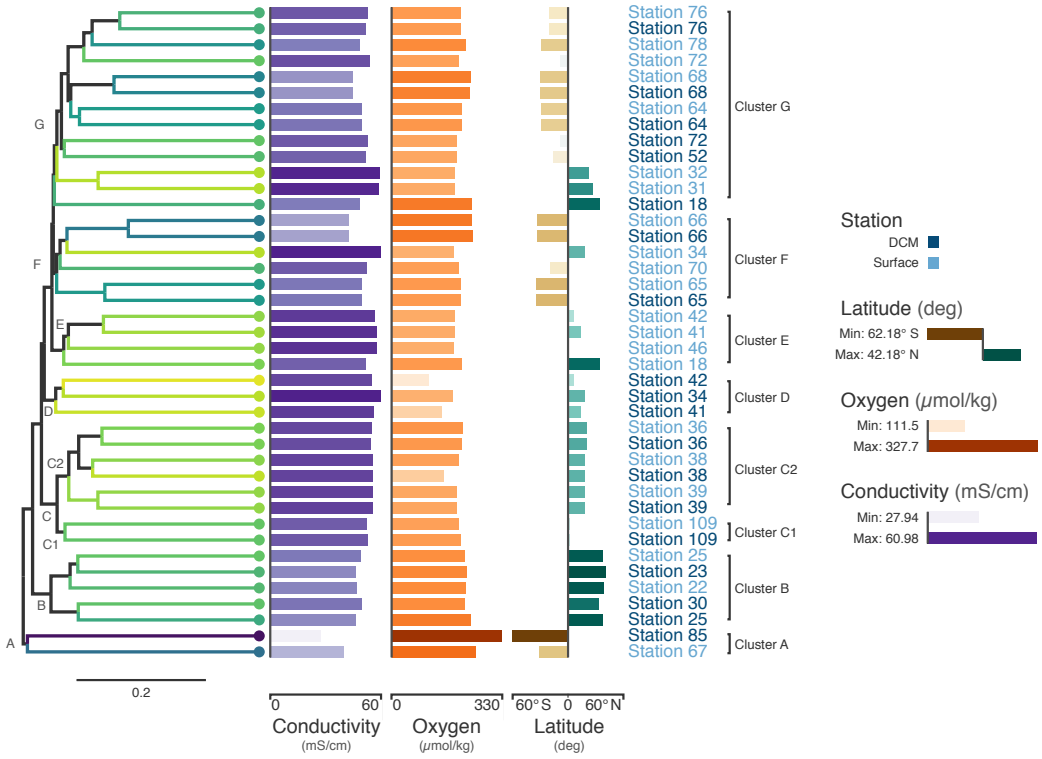

Figure 3. ***Tara*** **Oceans virome similarity with associated metadata.** Average-linkage hierarchical clustering of sample UniFrac distance based on RNR sequences mined from 41 *Tara* Oceans viromes. Major and sub-clusters of samples (A-G) are labeled. Branch color is based on a scaled, 1-dimensional projection of sample conductivity, oxygen, and latitude onto the cubehelix 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.

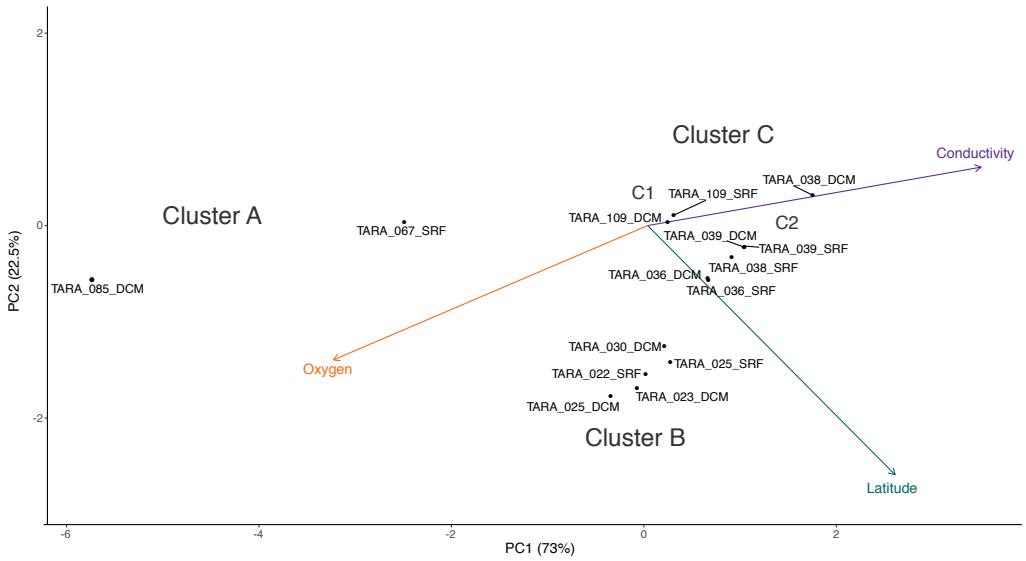

Figure 4. PCA biplot of *Tara* Oceans virome clusters A, B, and C. Principal components analysis biplot of *Tara* Oceans viromes based on sample oxygen, conductivity, and latitude. Ordination was done on all viromes, but only those from clusters A, B, and C are shown here for clarity.

REFERENCES

Ahmad, M. F., Kaushal, P. S., Wan, Q., Wijerathna, S. R., An, X., Huang, M., and Dealwis, C. G. (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.

Aiewsakun, P., Adriaenssens, E. M., Lavigne, R., Kropinski, A. M., and 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.

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). In Deng, X. and Du, D.-Z., editors, *ISAAC: 16th International Symposium on Algorithms and Computation*, volume 3827 of *Lecture Notes in Computer Science*, pages 1110–1121. Springer.

Bellas, C. M., Anesio, A. M., and Barker, G. (2015). Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Frontiers in Microbiology*, 6(JUL):656.

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.

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

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.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The Isme Journal*, 11:2639.

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. *Veterinary Research*, 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.

Cotruvo, J. A. and Stubbe, J. (2011). Class I Ribonucleotide Reductases: Metallocofactor Assembly and Repair In Vitro and In Vivo. *Annual Review of Biochemistry*, 80(1):733–767.

Coutinho, F. H., Silveira, C. B., Gregoracci, G. B., Thompson, C. C., Edwards, R. A., Brussaard, C. P. D., Dutilh, B. E., and Thompson, F. L. (2017). Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications*, 8(May):1–12.

Daniel, A. d. C., Pedrós-Alió, C., Pearce, D. A., and 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.

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.

Edwards, R. A., McNair, K., Faust, K., Raes, J., and Dutilh, B. E. (2016). Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews*, 40(2):258–272.

Emerson, J. B., Thomas, B. C., Andrade, K., Heidelberg, K. B., and Banfield, J. F. (2013). New Approaches Indicate Constant Viral Diversity despite Shifts in Assemblage Structure in an Australian Hypersaline Lake. *Applied and Environmental Microbiology*, 79(21):6755.

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.

Finke, J. F. and Suttle, C. A. (2019). The Environment and Cyanophage Diversity: Insights From Environmental Sequencing of DNA Polymerase. *Frontiers in Microbiology*, 10:167.

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(1):356.

Hancock, D. D., Besser, T. E., Kinsel, M. L., Tarr, P. I .and 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.

Harrison, A. O., Moore, R. M., Polson, S. W., and Wommack, K. E. (2019). Reannotation of the Ribonucleotide Reductase in a Cyanophage Reveals Life History Strategies Within the Virioplankton. *Frontiers in Microbiology*, 10:134.

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):460.

Junger, P. C., Amado, A. M., Paranhos, R., Cabral, A. S., Jacques, S. M. S., and Farjalla, V. F. (2018). Salinity Drives the Virioplankton Abundance but Not Production in Tropical Coastal Lagoons. *Microbial Ecology*, 75(1):52–63.

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.

Kolberg, M., Strand, K. R., Graff, P., and Kristoffer Andersson, K. (2004). Structure, function, and mechanism of ribonucleotide reductases. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1699(1):1–34.

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.

Ladau, J., Sharpton, T. J., Finucane, M. M., Jospin, G., Kembel, S. W., O'Dwyer, J., Koeppel, A. F., Green, J. L., and Pollard, K. S. (2013). Global marine bacterial diversity peaks at high latitudes in winter. *The Isme Journal*, 7:1669.

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*, 10:1555–1567.

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.

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

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.

Mei, Y., He, C., Huang, Y., Liu, Y., Zhang, Z., Chen, X., and 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.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., and Ogata, H. (2016). Linking Virus Genomes with Host Taxonomy. *Viruses*, 8(3):66–66.

Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., and Ghai, R. (2013). Expanding the marine virosphere using metagenomics. *PLoS Genetics*, 9(12):1–13.

Munson-McGee, J. H., Peng, S., Dewerff, S., Stepanauskas, R., Whitaker, R. J., Weitz, J. S., and Young, M. J. (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.

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):1152–1165.

Nelson, D. (2004). Phage taxonomy: we agree to disagree. *Journal of bacteriology*, 186(21):7029–7031.

Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., Blanc-Mathieu, R., Yamamoto, K., Hingamp, P., Sako, Y., Sullivan, M. B., 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, M. B., Goto, S., Ogata, H., and Yoshida, T. (2017a). Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. *mSphere*, 2(2).

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

Nordlund, P. and Reichard, P. (2006). Ribonucleotide Reductases. *Annual Review of Biochemistry*, 75(1):681–706.

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

Pawlowicz, R. (2012). The electrical conductivity of seawater at high temperatures and salinities. *Desalination*, 300:32–39.

Pawlowicz, R. (2019). Electrical Properties of Sea Water: Theory and Applications. In Cochran, J. K., Bokuniewicz, H. J., and Yager, P. L., editors, *Encyclopedia of Ocean Sciences (Third Edition)*, pages 71–80. Academic Press, Oxford.

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.

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

Pope, W. H., Jacobs-Sera, D., Russell, D. A., Peebles, C. L., Al-Atrache, Z., Alcoser, T. A., Alexander, L. M., Alfano, M. B., Alford, S. T., Amy, N. E., Anderson, M. D., Anderson, A. G., Ang, A. A. S., Ares, Jr., M., Barber, A. J., Barker, L. P., Barrett, J. M., Barshop, W. D., Bauerle, C. M., Bayles, I. M., Belfield, K. L., Best, A. A., Borjon, Jr., A., Bowman, C. A., Boyer, C. A., Bradley, K. W., Bradley, V. A., Broadway, L. N., Budwal, K., Busby, K. N., Campbell, I. W., Campbell, A. M., Carey, A., Caruso, S. M., Chew, R. D., Cockburn, C. L., Cohen, L. B., Corajod, J. M., Cresawn, S. G., Davis, K. R., Deng, L., Denver, D. R., Dixon, B. R., Ekram, S., Elgin, S. C. R., Engelsen, A. E., English, B. E. V., Erb, M. L., Estrada, C., Filliger, L. Z., Findley, A. M., Forbes, L., Forsyth, M. H., Fox, T. M., Fritz, M. J., Garcia, R., George, Z. D., Georges, A. E., Gissendanner, C. R., Goff, S., Goldstein, R., Gordon, K. C., Green, R. D., Guerra, S. L., Guiney-Olsen, K. R., Guiza, B. G., Haghighat, L., Hagopian, G. V., Harmon, C. J., Harmson, J. S., Hartzog, G. A., Harvey, S. E., He, S., He, K. J., Healy, K. E., Higinbotham, E. R., Hildebrandt, E. N., Ho, J. H., Hogan, G. M., Hohenstein, V. G., Holz, N. A., Huang, V. J., Hufford, E. L., Hynes, P. M., Jackson, A. S., Jansen, E. C., Jarvik, J., Jasinto, P. G., Jordan, T. C., Kasza, T., Katelyn, M. A., Kelsey, J. S., Kerrigan, L. A., Khaw, D., Kim, J., Knutter, J. Z., Ko, C.-C., Larkin, G. V., Laroche, J. R., Latif, A., Leuba, K. D., Leuba, S. I., Lewis, L. O., Loesser-Casey, K. E., Long, C. A., Lopez, A. J., Lowery, N., Lu, T. Q., Mac, V., Masters, I. R., McCloud, J. J., McDonough, M. J., Medenbach, A. J., Menon, A., Miller, R., Morgan, B. K., Ng, P. C.,

Nguyen, E., Nguyen, K. T., Nguyen, E. T., Nicholson, K. M., Parnell, L. A., Peirce, C. E., Perz, A. M., Peterson, L. J., Pferdehirt, R. E., Philip, S. V., Pogliano, K., Pogliano, J., Polley, T., Puopolo, E. J., Rabinowitz, H. S., Resiss, M. J., Rhyan, C. N., Robinson, Y. M., Rodriguez, L. L., Rose, A. C., Rubin, J. D., Ruby, J. A., Saha, M. S., Sandoz, J. W., Savitskaya, J., Schipper, D. J., Schnitzler, C. E., Schott, A. R., Segal, J. B., Shaffer, C. D., Sheldon, K. E., Shepard, E. M., Shepardson, J. W., Shroff, M. K., Simmons, J. M., Simms, E. F., Simpson, B. M., Sinclair, K. M., Sjoholm, R. L., Slette, I. J., Spaulding, B. C., Straub, C. L., Stukey, J., Sughrue, T., Tang, T.-Y., Tatyana, L. M., Taylor, S. B., Taylor, B. J., Temple, L. M., Thompson, J. V., Tokarz, M. P., Trapani, S. E., Troum, A. P., Tsay, J., Tubbs, A. T., Walton, J. M., Wang, D. H., Wang, H., Warner, J. R., Weisser, E. G., Wendler, S. C., Weston-Hafer, K. A., Whelan, H. M., Williamson, K. E., Willis, A. N., Wirtshafter, H. S., Wong, T. W., Wu, P., Yang, Y. j., Yee, B. C., Zaidins, D. A., Zhang, B., Zúniga, M. Y., Hendrix, R. W., and Hatfull, G. F. (2011). Expanding the Diversity of Mycobacteriophages: Insights into Genome Architecture and Evolution. *PLOS ONE*, 6(1):e16329.

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

Pruesse, E., Glöckner, F. O., and Peplies, J. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28(14):1823–1829.

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

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

Rambaut, A. (2006). FigTree.

Reichard, P. (1993). From RNA to DNA, why so many ribonucleotide reductases? *Science*, 260(5115):1773.

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

Robinson, O., Dylus, D., and Dessimoz, C. (2016). Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. *Molecular Biology and Evolution*, 33(8):2163–2166.

Rohwer, F. and Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of bacteriology*, 184(16):4529–4535.

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

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., Poulos, B. T., Solonenko, N., Lara, E., Poulain, J., Pesant, S., Kandels-Lewis, S., Dimier, C., Picheral, M., Searson, S., Cruaud, C., Alberti, A., Duarte, C. M., Gasol, J. M., Vaqué, D., Bork, P., Acinas, S. G., Wincker, P., and Sullivan, M. B. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537(7622):689–693.

Roux, S., Hallam, S. J., Woyke, T., and Sullivan, M. B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*, 4:1–20.

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):15786–15791.

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.

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

Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., Hull, R., King, A. M., Koonin, E. V., Krupovic, M., Kuhn, J. H., Lefkowitz, E. J., Nibert, M. L., Orton, R., Roossinck, M. J., Sabanadzovic, S., Sullivan, M. B.,

Suttle, C. A., Tesh, R. B., van der Vlugt, R. A., Varsani, A., and Zerbini, F. M. (2017). Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, 15:161.

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

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.

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

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*, 39:470–474.

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*, 13:209.

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

Villarroel, J., Kleinheinz, A. K., Jurtz, I. V., Zschach, H., Lund, O., Nielsen, M., Larsen, V. M., Kleinheinz, K. A., Jurtz, V. I., Zschach, H., Lund, O., Nielsen, M., and Larsen, M. V. (2016). HostPhinder: A Phage Host Prediction Tool. *Viruses*, 8(5):1–22.

Vos, R. A., Caravas, J., Hartmann, K., Jensen, M. A., and Miller, C. (2011). BIO::Phylo-phyloinformatic analysis using perl. *BMC Bioinformatics*, 12:63.

Winter, C., Matthews, B., and Suttle, C. A. (2013). Effects of environmental variation and spatial distance on Bacteria, Archaea and viruses in sub-polar and arctic waters. *The Isme Journal*, 7:1507.

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.

Xu, Y., Dugat-Bony, E., Zaheer, R., Selinger, L., Barbieri, R., Munns, K., McAllister, T. A., and Selinger, L. B. (2014). Escherichia coli O157:H7 Super-Shedder and Non-Shedder Feedlot Steers Harbour Distinct Fecal Bacterial Communities. *PLOS ONE*, 9(5):e98115.

Zaheer, R., Dugat-Bony, E., Holman, D., Cousteix, E., Xu, Y., Munns, K., Selinger, L. J., Barbieri, R., Alexander, T., McAllister, T. A., and Selinger, L. B. (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.

Zhao, L., Tyler, P., Starnes, J., Bratcher, C., Rankins, D., McCaskey, T., and 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.