MetacodeR: An R package for manipulation and heat tree visualization of community taxonomic data from metabarcoding

Zachary S. L. Foster[1], Thomas J. Sharpton[2,3,4], Niklaus J. Grünwald[1,4,5*]

[1] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, 97331, USA
[2] Department of Microbiology, Oregon State University, Corvallis, OR, 97331, USA
[3] Department of Statistics, Oregon State University, Corvallis, OR, 97331, USA
[4] Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR, 97331, USA
[4] Horticultural Crops Research Laboratory, USDA-ARS, Corvallis, OR, 97330, USA
* Corresponding author: nik.grunwald@ars.usda.gov

Abstract

Community-level data, the type generated by an increasing number of metabarcoding studies, is often graphed as stacked bar charts or pie graphs; these graph types do not convey the hierarchical structure of taxonomic classifications and are limited by the use of color for categories. We developed *MetacodeR*, an R package for easily parsing, manipulating, and plotting hierarchical data. To accomplish this, *MetacodeR* provides a function to parse most text-based formats that contain taxonomic classifications, taxon names, taxon IDs, or sequence IDs. This parsed data can then be subset, sampled, and ordered using a set of intuitive functions that take into account the hierarchical nature of the data. Finally, an extremely flexible plotting function allows for the quantitative representation of up to 4 arbitrary statistics simultaneously in a tree format by mapping statistics to color and size of tree nodes and edges. *MetacodeR* also allows exploration of barcode primer bias by integrating functions to run digital PCR. *MetacodeR* has been designed for data from metabarcoding research, but can easily be applied to any data that has a hierarchical component such as gene ontology, gene expression data, or geographic location data. Our package complements currently available tools for community analysis and is provided open source with extensive online user manuals.

1 Introduction

Metabarcoding is revolutionizing our understanding of complex ecosystems by circumventing limits of traditional microbial diversity assessment including the need for culturability, the effects of cryptic diversity, and the reliance on expert identification. Metabarcoding is a technique for determining community composition that typically involves extracting environmental DNA, amplifying a gene shared by the group of interest using PCR, sequencing the amplicons, and comparing the sequences to reference databases (Cristescu, 2014). It has been used to explore communities inhabiting diverse environments, including oceans (De Vargas et al., 2015), plants (Coleman-Derr et al., 2016), animals (Yu et al., 2012), humans (Consortium et al., 2012), and soil (Gilbert, Jansson, & Knight, 2014).

The complex community data produced by metabarcoding is challenging conventional graphing techniques. Most often, bar charts, stacked bar charts, or pie graphs are employed that use color to represent a small number of taxa at the same rank (e.g. phylum, class, etc). This reliance on color for categorical information

limits the number of taxa that can be effectively displayed, so most publications to only show results at a coarse taxonomic rank (e.g class) or for only the most abundant taxa. Furthermore, these graphing techniques do not convey the hierarchical nature of taxonomic classifications, potentially obscuring patterns in unexplored taxonomic ranks that might be more important biologically.

Here, we introduce the R package *MetacodeR* that is specifically designed to address some of these problems in metabarcoding-based community ecology, focusing on parsing and manipulation of hierarchical data and community visualization. *MetacodeR* provides a visualization that we call heat trees which quantitatively depicts statistics associated with taxa, such as abundance, using the color and size of nodes and edges in a taxonomic tree. This is similar to some visulizations produced by MetaPhlAn (Segata et al., 2012). These heat trees are useful for evaluating taxonomic coverage, barcode bias, or displaying differences in taxon abundance between communities. *MetacodeR* also provides a means of extracting and parsing taxonomy information from text-based formats (e.g. reference database FASTA headers) and an intuitive set of functions for subsetting, sampling, and rearranging taxonomic data. *MetacodeR* also allows exploration of barcode primer bias by integrating digital PCR. All this functionality is made intuitive and user-friendly while still allowing extensive customization and flexibility. *MetacodeR* can be applied to any data that can be organized hierarchically such as gene ontology, geographic location, or even the nesting of HTML tags. *MetacodeR* is an open source project available on CRAN and has a comprehensive manual with examples.

2 Design and Implementation

The R package *MetacodeR* provides a set of novel tools designed to parse, manipulate, and visualize community diversity data in a tree format using any taxonomic classification (Figure 1). Figure 1 illustrates the ease of use and flexibility of MetacodeR. It shows an example analysis extracting taxonomy from the 16S RDP training set for mothur (Schloss et al., 2009), filtering and sampling the data by both taxon and sequence characteristics, running digital PCR, and graphing the proportion of sequences amplified for each taxon. Table 1 provides an overview of the core functions available in *MetacodeR*.

To store the taxonomic hierarchy and associated observations (e.g. sequences) we developed a new data object class called taxmap. The taxmap class is designed to be as flexible and easily manipulated as possible. The only assumption made about the users data is that it can be represented as a set of observations assigned to a hierarchy; the hierarchy and the observations do not need to be biological. The class contains two tables in which user data is stored: a taxonomic hierarchy stored as an edge list of unique IDs and a set of observations mapped to that hierarchy (Figure 1). Users can add, remove, or reorder both columns and rows in either table using convenient functions included in the package (Table 1). For each table, there is also a list of functions stored with the class that each create a temporary column with the same name when referenced by one of the manipulation or plotting functions. These are useful for attributes that must be updated when the data is subset or otherwise modified, such as the number of observations for each taxon (see "n_obs" in Figure 1). If this kind of derived information was stored in a static column the user would have to update the column each time the data set is subset, potentially leading to mistakes if this is not done. There are many of these column-generating functions included by default, but the user can easily add their own by adding a function that takes a taxmap object. The names of columns or column-generating functions in either table of a taxmap object can be referenced as if they were independent variables in most *MetacodeR* functions in the style of the popular R packages like *ggplot2* and *dplyr*. This makes the code much easier to read and write.

2.1 Universal parsing and retrieval of taxonomic information

MetacodeR provides a way to extract taxonomic information from text-based formats so it can be manipulated within R. One of the most inefficient steps in bioinformatics can be loading and parsing data into a standardized form usable for computational analysis. Many databases have unique taxonomy formats with

Table 1: Primary functions found in *MetacodeR*.

Functions	**Description**
extract_taxonomy	Parses taxonomic data from arbitrary text and returns a *taxmap* object containing a table with rows corresponding to inputs (i.e. observations) and a table with rows corresponding to taxa.
heat_tree	Makes tree-based plots of data stored in *taxmap* objects. Color, size, and labels of tree components can be mapped to numeric data.
primersearch	Executes the EMBOSS program called primersearch on sequence data stored in a *taxmap* object. Results are parsed and added to the input *taxmap* object.
mutate_taxa *mutate_obs* *transmute_taxa* *transmute_obs*	Modify columns of taxon or observation data in *taxmap* objects. *mutate_** adds columns and *transmute_** returns only new columns.
select_taxa *select_obs*	Subset columns of taxon or observation data in *taxmap* objects.
filter_taxa *filter_obs*	Subset rows of taxon or observation data in *taxmap* objects.
arrange_taxa *arrange_obs*	Order rows of taxon or observation data in *taxmap* objects.
sample_n_taxa *sample_n_obs* *sample_frac_taxa* *sample_frac_obs*	Randomly subsample rows,of taxon or observation data in *taxmap* objects. Weights can be applied that take into account the taxonomic hierarchy and associated observations.
subtaxa *supertaxa* *observations* *roots*	Returns the indexes of rows in taxon or observation data in *taxmap* objects. Used to map taxa to related taxa and observations.

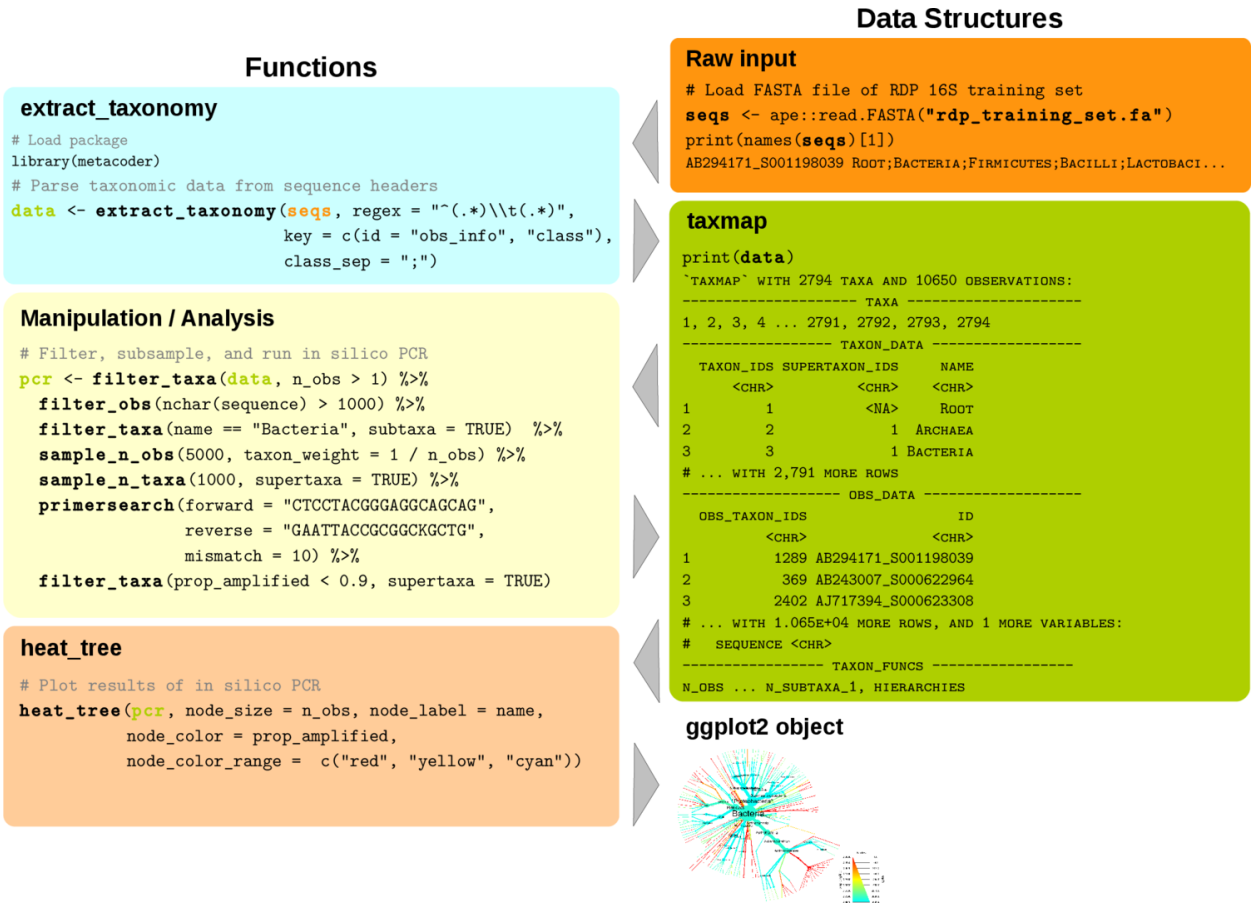

Figure 1: ***MetacodeR* has an intuitive and easy to use syntax.** The code in this example analysis parses the taxonomic data from the RDP (Maidak et al., 1996) 16S training set, filters and subsamples the data by sequence and taxon characteristics, conducts digital PCR, and displays the results as a heat tree. All functions in bold are from in the *MetacodeR* package.

differing types of taxonomic information. The structure and nomenclature of the taxonomy used can be unique to the database or reference another database such as GenBank (Benson et al., 2013). Rather than creating a parser for each data format, *MetacodeR* provides a single function to parse any format definable by regular expressions that contains taxonomic information (Figure 1). This makes it easier to use multiple data sources with the same downstream analysis (Figure 3).

The `extract_taxonomy` function can parse hierarchical classifications or retrieve classifications from online databases using taxon names, taxon IDs, or Genbank sequence IDs. The user supplies a regular expression with capture groups (parentheses) and a corresponding key to define what parts of the input can provide classification information. The `extract_taxonomy` function has been used successfully to parse several major database formats including Genbank (Benson et al., 2013), UNITE (Kõljalg et al., 2013), Protist Ribosomal Reference Database (PR2) (Guillou et al., 2012), Greengenes (DeSantis et al., 2006), Silva (Quast et al., 2013), and, as illustrated in figure 1, the Ribosomal Database Project (RDP) (Maidak et al., 1996). Examples for each database are provided in the user manuals (Foster & Grünwald, 2016).

2.2 Intuitive manipulation of taxonomic data

MetcodeR makes it easy to subset and sample large data sets composed of thousands of observations (e.g. sequences) assigned to thousands of taxa, while taking into account hierarchical relationships. This allows for exploration and analysis of manageable subsets of a large data set. Taxonomies are inherently hierarchical, making them difficult to subset and sample intuitively compared with typical tabular data. In addition to the taxonomy itself, there is usually also a set of things assigned to taxa in the taxonomy, which we refer to as "observations". Subsetting either the taxonomy or the associated observations, depending on the goal, might require subsetting both to keep them in sync. For example, if a set of taxa are removed or left out of a random subsample, should the subtaxa and associated observations also be removed, left as is, or reassigned to a supertaxon? If observations are removed, should the taxa they were assigned to also be removed?

MetacodeR provides functions to intuitively and efficiently subset complex hierarchical data sets using a cohesive set of functions inspired by the popular *dplyr* data-manipulation philosophy. *Dplyr* is an R package for providing a conceptually consistent set of operations for manipulating tabular information (Wickham & Francois, 2016). Whereas *dplyr* functions each act on a single table, *MetacodeR*'s analogous functions act on both the taxon and oberservation tables in a taxmap object. For each major *dplyr* function there are two analogous *MetacodeR* functions: one that manipulates the taxon table and one that manipulates the observations table. The functions take into account the relationship between the two tables and can modify both depending on parameterization, allowing for operations on taxa to affect their corresponding observations and vice versa. They also take into account the hierarchical nature of the taxon table. For example, the *MetacodeR* functions `filter_taxa` and `filter_obs` are based on the *dplyr* function `filter` and are used to remove rows in the taxon and observation tables corresponding to some criterion. Unlike simply applying a filter to these tables directly, these functions allow the subtaxa, supertaxa, and/or observations of taxa passing the filter to be preserved or discarded, making it easy to subset the data in diverse ways. There are also functions for ordering rows (`arrange_taxa`, `arrange_obs`), subsetting columns (`select_taxa`, `select_obs`), and adding columns (`mutate_taxa`, `mutate_obs`).

MetacodeR also provides functions for random sampling of taxa and corresponding observations. The function `taxonomic_sample` is used to randomly sub-sample items such that all taxa of one or more given ranks have some specified number of observations representing them. Taxa with too few sequences are excluded and taxa with too many are randomly subsampled. Furthermore, whole taxa can be sampled based on the number of sub-taxa they have. Alternatively, there are *dplyr* analogues called `sample_n_taxa` and `sample_n_obs`, which can sample some number of taxa or observations. In both functions, weights can be assigned to taxa or observations, influencing how likely each is to be sampled. For example, the probability of sampling a given observation can be determined by a taxon characteristic, such as the number of observations assigned to that taxon, or it could be determined by an observation characteristic, like sequence length. Similar to the `filter_*` functions, there are parameters controlling whether selected taxa's subtaxa, supertaxa, or observations are included or not in the sample. Figure 1 provides an example of these functions in the manipulation/analysis box.

2.3 Heat tree plotting of taxonomic data

Visualizing the massive data sets being generated by modern sequencing of complex ecosystems is typically done using traditional stacked barcharts or pie graphs, but these ignore the hierarchical nature of taxonomic classifications and their reliance on colors for categories limits the number of taxa that can be displayed at once. Generic trees can convey a taxonomic hierarchy, but displaying how statistics are distributed throughout the tree, including internal taxa, is difficult. *MetacodeR* provides a function that plots up to 4 statistics on a tree with quantitative legends by automatically mapping any set of numbers to the color and width of nodes and edges. The size and content of edge and node labels can also be mapped to custom values. These publication-quality graphs provide a method for visualizing community data that is richer than is currently possible with stacked bar charts. Although there are other R packages that can plot variables

on trees, like phyloseq, these have been designed for phylogenetic rather than taxonomic trees and therefore variables are only plotted on the tips of the tree, not internal nodes.

The function `heat_tree` creates a tree utilizing color and size to display taxon statistics (e.g., sequence abundance) for many taxa and ranks in one intuitive graph (Figure 1, see heat tree box for code and *ggplot* box for heat tree example). Taxa are represented as nodes and both color and size are used to represent any statistic associated with taxa, such as abundance. Although the `heat_tree` function has many options to customize the appearance of the graph, it is designed to minimize the amount of user-defined parameters necessary to make an effective visualization. The size range of graph elements is optimized for each graph to minimize overlap and maximize size range. Raw statistics are automatically translated to size and color and a legend is added that displays the relationship. Unlike most other plotting functions in R, the plot looks the same regardless of output size, allowing the graph to be saved at any size or used in a complex, composite figures such as figure 3 without changing parameters. These characteristics allow `heat_tree` to be used effectively in pipelines and with minimal parametrization since a small set of parameters displays diverse taxonomy data. The output of the `heat_tree` function is a *ggplot2* object, making it compatible with many existing R tools.

3 Applications

3.1 Heat trees allow visualization of community diversity data

We developed heat trees to allow visualization of community data in a taxonomic context by mapping any statistic to the color or size of tree components. Here, we reanalyzed data set 5 from the TARA oceans eukaryotic plankton diversity study to visualize the similarity between OTUs observed in the data set and their closest match to a sequence in a reference database (De Vargas et al., 2015). The TARA ocean expedition analysed DNA extracted from ocean water throughout the world. Even though a custom reference database was made using curated 18S sequences spanning all known eukaryotic diversity, many of the OTUs observed had no close match. Figure 2 shows a heat tree that illustrates the proportion of OTUs that were well characterized in each taxon (at least 90% identical to a reference sequence). Color indicates the percentage of OTUs that are somewhat similar to the closest reference sequence. Node width indicates the number of OTUs assigned to each taxon and edge width is mapped to the number of reads. Taxa with ambiguous names and those with less than 200 reads have been filtered out for clarity. This figure illustrates one of the principal advantages of heat trees, as it reveals many clades in the tree that contain only red lineages, which indicate that the entire taxonomic group is poorly represented in the reference sequence database. Of particular interest are those clades with predominantly red lineages that also have relatively large nodes, such as Harpacticoida. These represent taxonomic groups that were found to have high amounts of diversity in the oceans, but for which we have a paucity of genomic information. Investigators interested in improving the genomic resolution of the biosphere can thus use these approaches to rapidly assess which taxa should be prioritized for focused investigations. Note that a large portion of the taxa shown in red, yellow or orange have many OTUs with a poor match to the reference taxonomic hierarchy.

3.2 Taxonomic information from multiple sources can be conveniently compared

Metabarcode studies often rely on techniques or data that may introduce bias into an investigation. For example, the specific set of PCR primers used to amplify genomic DNA and taxonomic annotation database can both have an effect on the study results. A quick and inexpensive way to estimate biases caused by primers is to use digital PCR, which simulates PCR success using alignments between reference sequences and primers. MetacodeR can be used to explore different databases or primer combinations to assess these effects since it supplies functions to parse divserse data sources, conduct digital PCR, and plot the results. Figure 3 shows a series of heat tree comparisons that were produced by using a common 16S rRNA metabarcoding

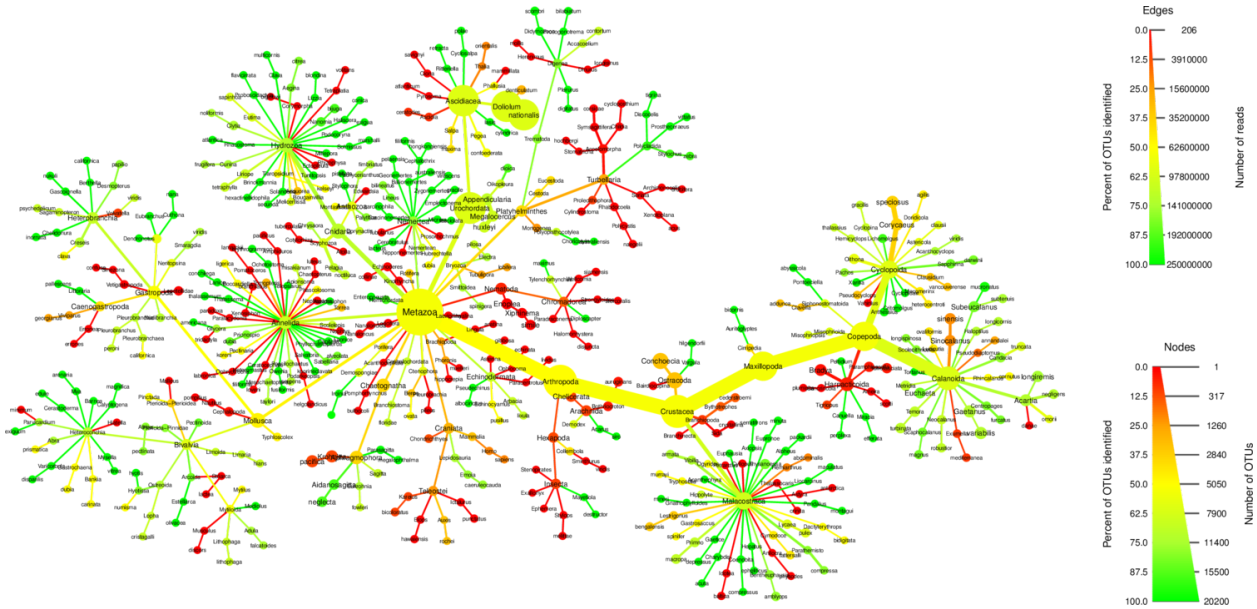

Figure 2: **Heat trees display data in a taxonomic context**. Most graph components, such as the size and color of text, nodes, and edges, can be automatically mapped to arbitrary numbers, allowing for a quantitative representation of multiple statistics simultaneously. This graph depicts the uncertainty of metazoan OTU classifications from the TARA global oceans survey (De Vargas et al., 2015). Each node represents a taxon used to classify OTUs and the edges determine where it fits in the overall taxonomic hierarchy. Node diameter is proportional to the number of OTUs classified as that taxon and edge width is proportional to the number of reads. Color represents the percent of OTUs assigned to each taxon that are somewhat similar to their closest reference sequence (>90% sequence identity).

primer set (Caporaso et al., 2012) and digital PCR against the full-length 16S sequences found in three taxonomic annotation databases: Greengenes (DeSantis et al., 2006), RDP (Maidak et al., 1996), and SILVA (Quast et al., 2013). These heat trees reveal subsets of the full taxonomies for these three databases that poorly amplify by digital PCR using the selected primers. As a result, they indicate which lineages within each of the taxonomies may be challenging to detect in a metabarcoding study that uses these primers. Importantly, all primer sets are likely to produce due to differential taxonomic structures and database sequences, but investigators that are interested in specific lineages may elect to use this approach in conjunction with various primer sets to identify those that maximize the likelihood of discovery. Additionally, these heat maps do not indicate whether one database is necessarily preferable over another, as they differ in the structure of their taxonomies, as well as the number and phylogenetic diversity of their reference sequences. For example, most of the bacterial clades that do not amplify well in the SILVA lineages are unnamed lineages that are not found in the other databases, indicating that they warrant further exploration.

3.3 Heat trees can be used for pairwise comparisons of communities across treatments with differential abundance

A major challenge in metabarcoding studies is visually determining how specific sub-sets of samples vary in their taxonomic composition. Unlike most other graphing software in R, *MetacodeR* produces graphs that look the same at any output size or aspect ratio, allowing heat trees to be easily integrated into larger composite figures without changing the code for individual subplots. Using color to depict the difference in read or OTU abundance between two treatments can result in particularly effective visualizations, especially when the presence of color is made dependent on a statistical test. To examine more than two treatments at

once, a matrix of these kind of heat trees can be combined with with a labeled "guide" tree. Figure 4 shows application of this idea to human microbiome data showing pairwise differences between body sites. Coloring indicates significant differences between the median proportion of reads for samples from different body sites as determined using a Wilcox Rank-Sum test followed by a Benjamini-Hochberg (FDR) correction for multiple testing. The intensity of the color is relative to the log-10 ratio of difference in median proportions. Brown taxa indicate an enrichment in body sites listed on the top of the graph and green is the opposite. While the original study (Consortium et al., 2012) showed abundance plots our visualization provides the taxonomic context. For example, *Haemophilus*, *Streptococcus*, and *Prevotella* spp. are enriched in saliva (brown) relative to stool where *Bacteroides* is enriched (green). We also see that in the Lachnospiraceae clade several genera shown in both green and brown taxa are differentially abundant. These observations are consistent with known differences in the human-associated microbiome across body sites, but the heat-tree uniquely provides an integrated view of how all levels of a taxonomy vary for all pairs of body sites.

4 Availability and Future Directions

The R package *MetacodeR* is an open-source project licensed under GNU General Public License version 3. Stable releases of *MetacodeR* are available on CRAN while recent improvements can be downloaded from github (https://github.com/grunwaldlab/metacoder). A manual with documentation and examples is provided (Foster & Grünwald, 2016).

We are currently continuing development of *MetacodeR*. We welcome contributions and feedback (preferably as GitHub issues) from the community. We are considering making *MetcodeR* functions and classes compatible with those from other bioinformatic R packages such as *phyloseq*, *ape*, *seqinr*, and *taxize*. We might integrate more options for digital PCR and barcode gap analysis, perhaps using ecoPCR or the R packages *PrimerMiner* and *Spider*. We are also considering adding alternate visualizations.

5 Acknowledgments

This work was supported in part by funds from USDA ARS CRIS Project 2027-22000-039-00 and the USDA ARS Floriculture Nursery Research Initiative. The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the United States Department of Agriculture or the Agricultural Research Service of any product or service to the exclusion of others that may be suitable.

6 Author Contributions

Conceived and designed the experiments: ZSLF, NJG, TJS. Performed the experiments:ZSLF. Analyzed the data: ZSLF. Contributed reagents/materials/analysis tools: ZSLF, NJG. Wrote the paper: ZSLF, NJG, TJS. Designed, developed scripts: ZSLF.

References

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2013). GenBank. *Nucleic acids research*, *41*(D1), D36–D42.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal*, *6*(8), 1621–1624.

Coleman-Derr, D., Desgarennes, D., Fonseca-Garcia, C., Gross, S., Clingenpeel, S., Woyke, T., et al. (2016). Plant compartment and biogeography affect microbiome composition in cultivated and native Agave species. *New Phytologist*, *209*(2), 798–811.

Consortium, H. M. P., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214.

Cristescu, M. E.(2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in ecology & evolution*, *29*(10), 566–571.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, *72*(7), 5069–5072.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605.

Foster, Z. S. L., & Grünwald, N. J.(2016). *MetacodeR user documentation.*

Gilbert, J. A., Jansson, J. K., & Knight, R.(2014). The Earth Microbiome project: successes and aspirations. *BMC biology*, *12*(1), 1.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., et al. (2012). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic acids research*, gks1160.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F., Bahram, M., et al.(2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular ecology*, *22*(21), 5271–5277.

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., & Woese, C. R. (1996). The ribosomal database project (RDP). *Nucleic acids research*, *24*(1), 82–85.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, *41*(D1), D590–D596.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, *75*(23), 7537–7541.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C.(2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, *9*(8), 811–814.

Walters, W., Hyde, E. R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., et al.(2016). Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems*, *1*(1), e00009–15.

Wickham, H., & Francois, R.(2016). *dplyr: A Grammar of Data Manipulation.* (R package version 0.5.0)

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., et al.(2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), 613–623.

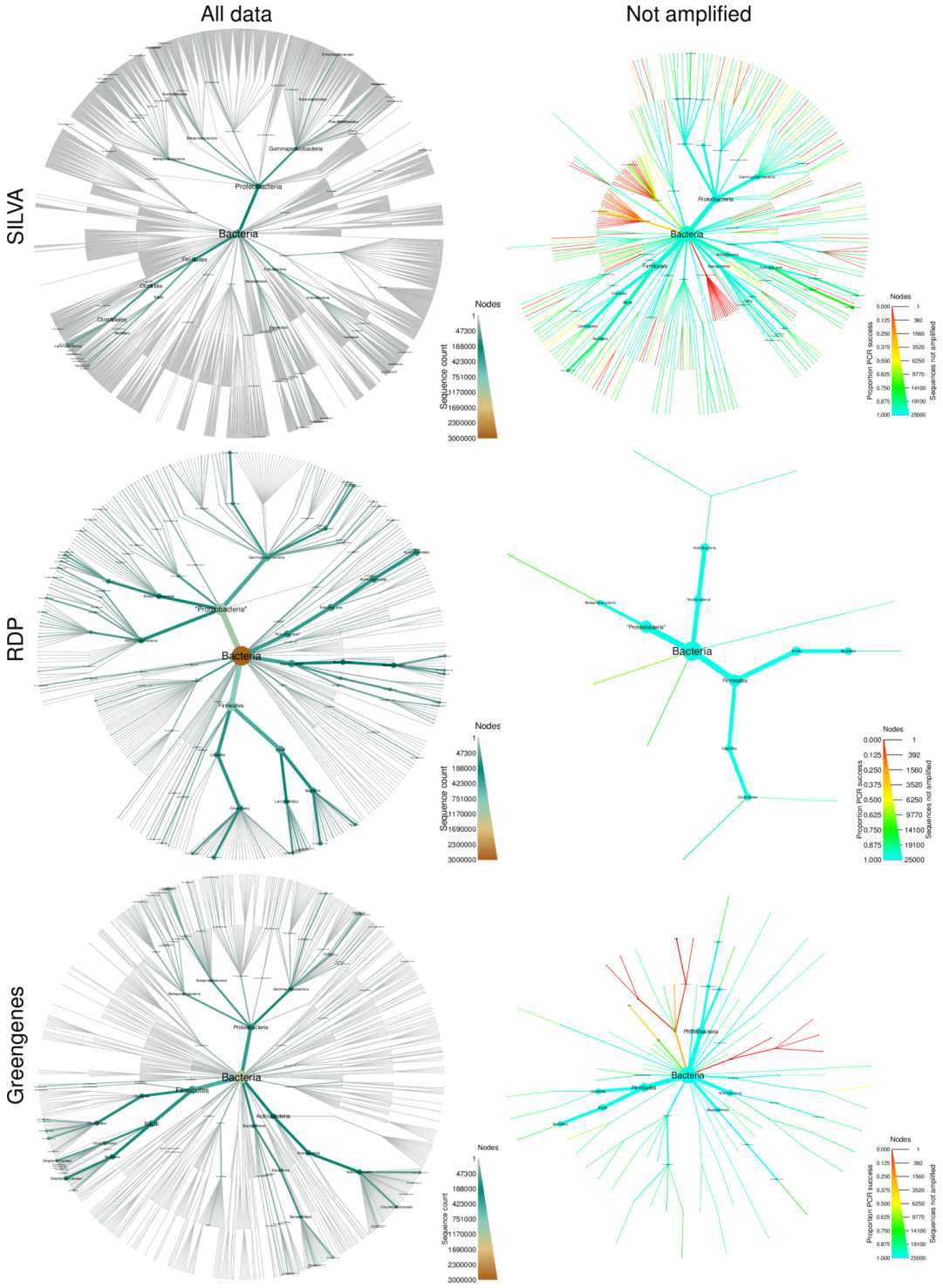

Figure 3: **Taxonomic information from diverse sources can be visualized and manipulated in the same way.** Flexible parsing and subsetting facilitates exploration of diverse data sources and primer combinations. Comparison of in digital PCR results using three different databases: SILVA (Top), RDP (Middle), and Greengenes (Bottom). Left plots display abundance of all bacterial 16S sequences. Right plots display all taxa with subtaxa not entirely amplified by digital PCR using universal 16S primers (Walters et al., 2016) and node color and size display the proportion and number of sequences not amplified respectively.

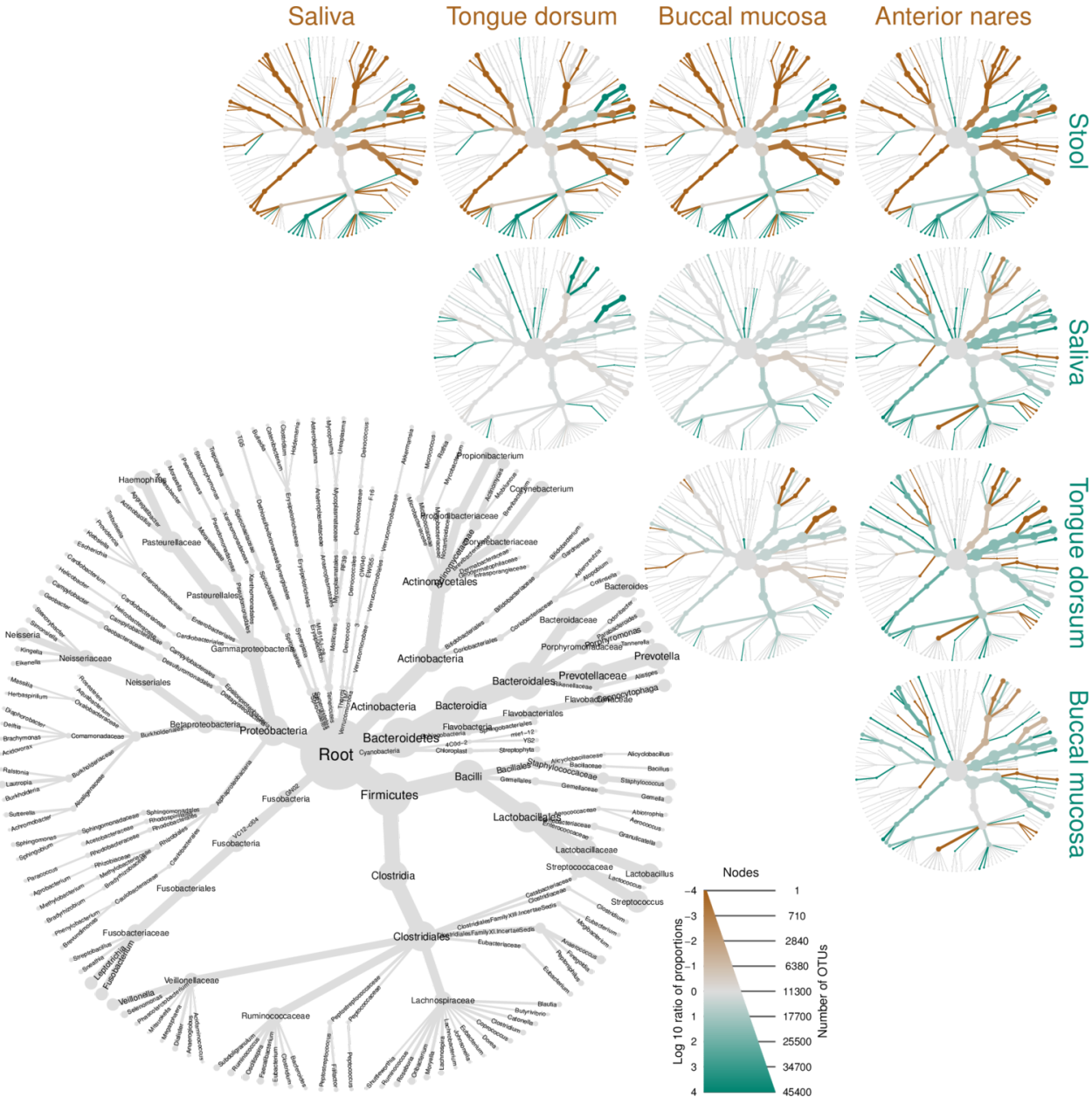

Figure 4: **Scale-independent appearance facilitates complex, composite figures**. All graph components, including text, have the same relative sizes independent of output size, unlike most graphical packages in R, making it easier to create composite figures entirely within R. This graph uses 16S metabarcoding data from the Human Microbiome Project study. The gray tree on the lower left functions as a key for the smaller unlabeled trees. The color of each taxon represents the log-10 ratio of median proportions of reads observed at each body site. Only significant differences are colored, determined using a Wilcox rank-sum test followed by a Benjamini-Hochberg (FDR) correction for multiple comparisons. For example, *Haemophilus*, *Streptococcus*, *Prevotella* are enriched in saliva (brown) relative to stool where *Bacteroides* is enriched (green).