An Interactive Visualization Tool for Exploratory Data Analysis of Taxonomic and Functional Metagenomic Data

Cecilia Noecker[1] cnoecker@uw.edu
Colin McNally[1] cmcn@uw.edu
Alex Eng[1] engal@uw.edu
Will Gagne-Maynard[2] gagnemaw@uw.edu

[1]School of Genome Sciences, University of Washington

[2]School of Oceanography, University of Washington

ABSTRACT

Author Keywords
Data visualization, metagenomics, microbiome, web interfaces

INTRODUCTION
Microbial communities living in and on humans, known as the human microbiome, are complex systems associated with a range of diseases including obesity, allergies, and inflammatory bowel disease. Other microbial communities found in water and soil play important ecological roles in food webs and nutrient cycling. High-throughput DNA sequencing has enabled comprehensive profiling of these communities using a technology known as metagenomics. These analyses traditionally choose between two main sequencing techniques to focus on either 1) which microbial species are present, or 2) what genetic functions are present? However, our lab and others have developed methods to answer both of these questions from the same dataset, using either species deconvolution of gene functions [1] or prediction of functions based on taxonomic composition [2].

The resulting datasets present several challenges for visualization. First, sequencing is an inherently compositional technology: these datasets are largely limited to describing the relative abundances of species or genes across samples. These data can also be quite high-dimensional, with up to 1,000 species and 10,000 annotated genes in a gut microbiome sample. However, these dimensions are not independent: species are evolutionarily related to varying degrees, which can be approximately summarized by the taxonomic classification hierarchy of kingdoms, phyla, classes, etc., and genes can have related functions and/or structures, which can be summarized in terms of metabolic pathways and other functional categories.

Importantly, though researchers tend to analyze taxonomic and functional variation separately, gene abundances are essentially a linear combination of species abundances: The amount of a particular gene function is determined by the number of organisms belonging to species that have that gene. Shifts in these contributions may affect the overall activity of the community and its relationship with its host or the environment. We decided to develop a visualization tool to enable initial exploratory analysis of these relationships in a particular dataset, while simultaneously addressing the other inherent challenges detailed above.

RELATED WORK
Traditionally, analyses of metagenomic data focus on trends in either the taxonomic composition (who is there -- species) of a community or the functional capacity (what can they do -- genes). It is becoming increasingly feasible to process metagenomic data to obtain high-resolution information on both of these questions. 16S targeted sequencing used to obtain taxonomic data can be applied to predict functional data using tools such as PICRUSt [2]. Whole metagenome shotgun sequencing, which is traditionally used to obtain functional data, can also be used to obtain taxonomic information [3], and to deconvolve which functions belong to which species (MetaDecon [1]).

The data produced by these tools has the potential to reveal new insights about microbial community functions and interactions, but tools to visualize and explore the complex relationships between genes and species are lacking. For example, the OmicTools website (http://omictools.com/data-visualization-c372-p1.html) lists a handful of examples of currently available interactive metagenomic data visualizations, but these all focus on exploring either single communities or a comparison of communities across single attributes (such as relative abundance of species) at a time. Another visualization tool was developed specifically for displaying the relative abundances of bacterial species across multiple samples using an interactive 3D heatmap view [4]. MG-RAST [5], a server for uploading and automated analysis of metagenomic data, also generates a set of predefined interactive visualizations. These include embedded Krona plots [6], which are interactive

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

layered pie or donut charts enabling hierarchical exploration of compositional data.

METHODS

Data and data structures

Taxonomic and gene function hierarchies

Both microbial species and gene functions have classification hierarchies. The visualization displays the hierarchies through trees and uses this hierarchical classification information to generate the trees and allow users to summarize or expand different aspects of the data based on interest. These hierarchies are stored as tree-like nest objects that allow for easy traversal, classification lookup, and efficiently identifying child and parent classification relationships when updating the various parts of the visualization as users expand or collapse nodes of the two trees.

Taxonomic and functional abundances

Abundances of different taxa and functions are provided as relative abundances within each sample, with the taxonomic abundances summing to 100 and the functional abundances summing to 100. The visualization displays the taxonomic abundance data through a stacked bar plots. These data are stored as two default d3 tsv structures, with each row containing a single sample's relative abundances and each column mapping to a taxon or function, which allows for easy iteration over the samples and lookup for specific abundances when updating the taxonomic stacked bar plot or looking up a function's relative abundance in a sample for functional contributions as described below.

Taxonomic contributions to gene functions

The contribution of each taxon to each function (at the most granular taxonomic and functional classification levels) is given as the relative contritubtion to the function provided by a taxon in a given sample. Thus, all contributions to a given function in a given sample will sum to 100. Relative contributions are then converted to relative functional abundances by multiplying the contribution by the relative abundance of the function within the sample. These data are then stored in a set of nested maps, keyed by sample, then taxon, then function, which allows the efficient lookup for specific contribution data displayed and linked between various parts of the visualization.

Currently displayed taxa and functions

The current leaf taxa and functions are kept synchronized between all parts of the visualization through two arrays, one for taxa and one for functions. These arrays are also used to provide a shared vertical ordering of taxa and functions within the trees, bipartite graph, and stacked bar plots.

Currently displayed contributions

To speed up transitions made when changing taxonomic or functional classification levels, all currently displayed contribution data is stored in a set of nested maps similar to the overall contribution data. However, this displayed contribution data structure instead contains only entries for the displayed taxa and the displayed functions. These summarized contributions are defined as, for a given non-leaf taxon and non-leaf function, the sum of all contributions of all corresponding descendent taxa to all corresponding descendent functions.

This data structure then allows access to the summarized contributions for all displayed taxa and functions without having to perform these summations on every update. When a single taxon or function is expanded or collapsed in the visualization, all non-involved summarized taxa or fucntions can be kept constant. Collapsing a taxon or function is made more efficient by determining the new displayed contribution as the sum of contributions from all nodes that were displayed but are removed due to the collapse. Node expansion is still inefficient, such that for each new non-leaf child node, the relevant contribution sums must be calculated from the data structure containing all of the lowest-level contribution data.

Color mappings

Colors for taxa and functions are defined through global color maps that are passed to each part of visualization to maintain consistent color keys.

Visualization components

Hierarchy trees

The trees are generated from the taxonomic and gene function hierarchies using the default d3 object. Interior node labels are hidden by default to reduce clutter, but ancestor node labels can be displayed on demand through mouse-over.

Bipartite graph

The bipartite graph is generated based on the contribution data and currently displayed taxa and functions. Each edge corresponds to a non-zero contribution relationship between taxon and function, found by iterating through the data structure of displayed contributions.

Stacked bar plots

The taxonomic relative abundances stacked bar plot is generated from the corresponding data. Each rectangle is defined separately for every displayed taxon and sample, which allows easy subselection for highlighting. The functional relative abundances stacked bar plot is generated not from the functional relative abudance data, but instead from the currently displayed contribution data. Each contribution in each sample is generated as a separate rectangle, with rectangles belonging to the same

function but separate contributions colored the same. The rectangles for the same function are stacked vertically adjacent within a sample in an order consistent with the taxonomic vertical orderings. This setup allows for easy highlighting through the selection of either all rectangles belonging to a single function, which will appear as a single large highlighted rectangle for the entire function, or all rectangles contributed by a single taxon, which will highlight sub-rectangles across all samples visualizing the contribution of a given taxon to all functions across all samples.

Highlight linking

Each part of the visualization provides a separate highlight capability defined by a selected taxon or function. Each of these highlight functions are then called through a global method, which is passed to each individual component of the visualization upon initialization. A given component can then highlight itself and all other components upon mouse-overs within that component corresponding to specific taxa or functions.

RESULTS

Description of the Visualization Tool

We plan to enable users to upload their own dataset for exploration, but currently our visualization shows only a sample dataset from [7], describing the taxonomic and predicted functional composition of the cecal microbiota of mice at 2 days and 6 weeks following treatment with the antibiotic cefoperazone, along with age-matched controls (29 samples in all).

As shown in Figure 1, our visualization consists of traditional stacked bar plots showing the relative abundances of taxa and functions in each of these samples, beneath a "control panel" that enables the user to switch between hierarchical levels of detail and query the contributions of particular taxa to particular functions across samples. The central bipartite graph shows presence-absence links between taxa and functions – whether a particular taxon possesses a specific function. This is not especially informative at high levels, but we believe it provides useful information for domain scientists searching for taxa contributing to very specific functions of interest such as antibiotic resistance or synthesis of particular metabolites.

There are two primary modes of interaction. First, the trees in the upper left and upper right show the taxonomic and functional hierarchies, respectively, and clicking on a node expands or collapses the tree to show more or less detail. It also results in a re-drawing of the bipartite graph and the stacked bar plots to reflect the newly expanded or collapsed data.

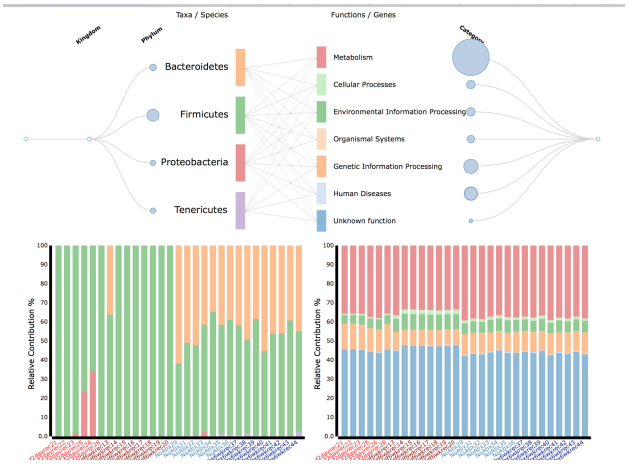

Figure 1. Initial display of our visualization of the example dataset.

The second and most important interaction is the unified mouseover, in which all items linked to the selection are highlighted to show their relationships. Mousing over a node on the taxa tree or a particular taxon in the stacked bar highlights and colors the linked function edges, its associated bars in the taxa stacked bar graph, and the contributions of that taxa to the various functions shown in the function bar graph (see figure 2). Similarly, mousing over an edge in the bipartite graph highlights the associated taxon in the taxa stacked bar and its contributions to the specific linked function node in the function stacked bar, and mousing over a node in the function graph or a bar in the function stacked bar highlights the entire function and linked taxa. The stacked bar mouseovers also display the associated relative abundance quantities for the selected item.

User Experiences

Users familiar with these datasets commented that this tool is fun and intuitive for exploration. They understood the interpretation of each part of the tool quickly and tended to click around the tree hierarchies rather rapidly, which caused the data transitions to lag a bit. As expected, they tended to identify and dig into specific functions of interest, drilling down into the hierarchy to access one element and closing all the other categories. They were able to quickly navigate back and forth between the stacked bar plots and the control panel, identifying interesting variation across samples in the bar plots and then going back up to the hierarchy to access more detail about the identified component.

Users unfamiliar with metagenomic datasets were substantially slower to interact with our visualization and tended to perform fewer interactions. This is not unexpected as they are not the primary intended audience. These users were frequently confused by the expanding and collapsing of the hierarchies, as well as the specific relationships between the different visual elements.

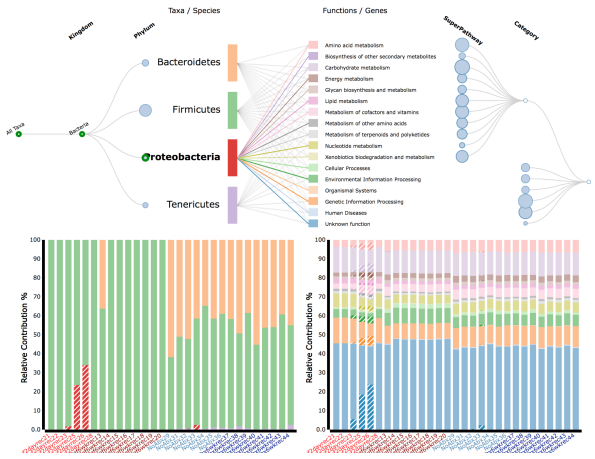

Figure 2. Mousing over the Proteobacteria node highlights the relative abundances of Proteobacteria across samples and its associated contributions to the functions it possesses.

They were less likely to explore using the bipartite graph as it looked to them as though most everything is connected and did not have specific biological questions or hypotheses in mind. However, they still frequently made observations about interesting sample variation and taxa-function links, including the differences seen in the antibiotic treatment samples and the general consistency of functions across samples.

DISCUSSION

We have developed a visualization tool that addresses a major gap in what is currently available to microbial community researchers. There is currently a gap in the way researchers study the distribution of taxa and the distribution of genetic functions present in samples and the interdependence between the two. Our tool provides a powerful first step in the process to unify these methods of research and accelerate the pace of discovery in microbiome research. Our tool has received positive feedback from preliminary user tests, and has potential to be a useful tool for data exploration. However there are some shortcomings of the tool in its current state. These include limitations to the ability of the user to focus on examining specific taxa and functions, compromises in data detail made due to computational demands, and the current lack of support for user-uploaded data. These problems are explained in more detail in the Future Directions section and the plans for how to address them are detailed.

FUTURE WORK

The future goals for this visualization project break down into three broad goals:

1. Give the user greater control over the visualization and ability to drill down.
2. Allow visualizations of larger and more detailed datasets.
3. Allow the user to upload their own data.

The first goal is to give the user greater power to drill down and search through the data. In the current iteration of our tool the abundances on the stacked bar chart of the displayed taxa and functions always sum to 100%, which makes it difficult to observe trends across samples in taxa or functions that are rare or at a lower hierarchical level and comprise less than 1% of the abundance. One component of the future solution to this problem is to implement selection options that will let the user select a specific hierarchical group or set of groups and hide all others. This selection would be made using the tree in the control panel and would result in all un-selected groups being hidden on the tree and removed from the bar graph. The y-axis of the bar graph would re-scale as appropriate to show the groups of interest. The selected groups can then be further subdivided by hierarchy, or hidden groups can be added back to the display using the selection mechanism on the tree. Sometimes the user will want to still be viewing multiple groups, but the nature of stacked bar charts makes fair comparisons difficult. Thus we intend to implement a feature to change the sorting of the displayed taxonomic or functional groups as a way to alter the view and move groups to the bottom or top of the bar charts where they can be viewed with less bias. A final component of our plans to improve usefulness is to allow the user to search through taxonomic and functional groups to more easily find specific data of interest to them. The search will automatically expand the trees and/or subset to the data of interest.

Many of the current challenges in drilling down on the data in our tool will be addressed via the above plans, but another major limiting step is that the current version of our tool is displaying a summarized version of the dataset rather than the raw data. The lowest level of taxa and function in metagenomics samples are Operational Taxonomic Units (OTUs, which roughly correspond to species), and KEGG Orthologs (KOs, which roughly correspond to genes). Our tool currently displays data where the OTUs have been summed together at the Genus level, and the KOs have been summed at the SubPathway level. This was done because the full version of the data proved to be too large and difficult to work with in javascript running in a client's web-browser, but it is a major limitation that the user is not able to drill down to visualize the lowest level of the data. Our planned future solution is to rewrite the database portion of our code to run on a server so that large, complete datasets can be displayed. The primary bottleneck in storing and processing the data is the handling of the complete set of data including all raw OTUs and KOs, but the user never needs to or should want to view this entire set of data. Thus the raw data and the unfiltered data cube will be stored on the server, and given the users desired set of taxonomic and functional groups to visualize, the server side code will compute sums of contributions and abundances over the raw data and send the summarized data to the user. The client-side code will store this summarized data in memory to allow rapid highlighting, sorting, and details on demand,

but the less frequent tasks of changing the displayed data will then rely on the server computation.

The last primary future goal of this project is to allow end-users to use it to visualize their own metagenomic sequencing data. Accomplishing this will require several steps. In the current state our tool reads in a set of different data files we prepared from the source data. Allowing the user to upload data that is in the same format would not be difficult, but several processing steps were performed on our data that the user would need to replicate, the most significant of which is the inference of relationships between taxa and functions. For the data used here this was done using PICRUSt [2], but other tools could also be used. The optimal solution is to set up a service wherein the user can upload their raw metagenomic data, which will then be processed using our lab's proprietary pipeline, and inference of taxonomic and functional data and the interdependence between the two will be calculated using PICRUSt or a similar program. The data will then become available to the user to view using our tool.

We believe that the sum of these future goals for the project will elevate this visualization tool to be truly useful to the microbiome research community.

ACKNOWLEDGMENTS

We thank Professor Jeff Heer for his instruction in class and his feedback on the initial project proposal. We also thank the TA's of CSE 512, Domonik Moritz and Jeffrey Snyder for their assistance and input. We would also like to acknowledge Professor Elhanan Borenstein for his input on this project.

REFERENCES

[1] R. Carr, S. S. Shen-Orr, and E. Borenstein, "Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution," *PLoS Comput Biol*, vol. 9, no. 10, p. e1003292, Oct. 2013.

[2] M. G. I. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko, and C. Huttenhower, "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences," *Nat. Biotechnol.*, vol. 31, no. 9, pp. 814–821, Sep. 2013.

[3] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes," *Nat. Methods*, vol. 9, no. 8, pp. 811–814, Aug. 2012.

[4] Pacific Symposium on Biocomputing and R. Altman, *Biocomputing 2011 Proceedings of the Pacific Symposium*. New Jersey: World Scientific, 2010.

[5] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. Edwards, "The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC Bioinformatics*, vol. 9, no. 1, p. 386, 2008.

[6] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a Web browser," *BMC Bioinformatics*, vol. 12, no. 1, p. 385, 2011.

[7] C. M. Theriot, M. J. Koenigsknecht, P. E. Carlson Jr, G. E. Hatton, A. M. Nelson, B. Li, G. B. Huffnagle, J. Z. Li, and V. B. Young, "Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to Clostridium difficile infection," *Nat. Commun.*, vol. 5, Jan. 2014.