Bioinformatics Advance Access published September 8, 2011

TopiaryExplorer: Visualizing large phylogenetic trees with environmental metadata

Meg Pirrung[1], Ryan Kennedy[2], J. Gregory Caporaso[3], Jesse Stombaugh[3], Doug Wendel[3] and Rob Knight[3,4,*]

[1]Computational Bioscience Program, University of Colorado Denver, Aurora, CO 80045.

[2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 19104.

[3]Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, CO 80309.

[4]Howard Hughes Medical Institute, Boulder, CO, 80309.

Associate Editor: Dr. Jeffrey Barrett

ABSTRACT

Motivation: Microbial community profiling is a highly active area of research, but tools that facilitate visualization of phylogenetic trees and associated environmental data have not kept up with the increasing quantity of data generated in these studies.

Results: TopiaryExplorer supports the visualization of very large phylogenetic trees, including features such as the automated coloring of branches by environmental data, manipulation of trees, and incorporation of per-tip metadata (e.g., taxonomic labels).

Availability: http://topiaryexplorer.sourceforge.net.

1 INTRODUCTION

Microbial community profiling using marker genes such as the 16S rRNA greatly expanded our knowledge of the diversity of the microbial world (Tringe and Hugenholtz, 2008). Phylogenetic trees are key to our understanding of the microbial world (Pace, 1997), and provide an important view into microbial data (Ludwig, et al., 2004). Software such as QIIME (Caporaso, et al., 2010) has kept pace with the increasing rate of sequence acquisition to allow statistical analysis of these data, but tools to visualize and manipulate phylogenetic trees and corresponding metadata (Huson, et al., 2007; Letunic and Bork, 2011; www.phylosoft.org/archaeopteryx; tree.bio.ed.ac.uk/software/figtree), which worked well on data sets that were characteristic a few years ago, are less suitable for datasets containing millions of sequences.

A key question in microbial ecology is which portions of a phylogenetic reference tree are differentially represented in specific groups of samples. To address this question, users should be able to load trees with thousands of tips, assign taxonomic labels to the tips, and color the branches based on data about each sample.

Here we present TopiaryExplorer, a software package that facilitates visual exploration of large phylogenetic trees, including information about each sample and each tip. This integration of what is often called "sequence metadata" is crucial to understanding how sequences (and their source organisms) are distributed across environments, and the processes underlying the observed patterns. TopiaryExplorer additionally allows display and revision of the taxonomy (including multiple taxonomies for the same tree, facilitating taxonomic comparisons), and integration with databases that contain sample/tree information (an example database is provided to assist users in creating their own: see below). It also provides key user interface improvements including: the ability to dynamically collapse or expand the whole tree using several different tree layout algorithms, allowing rapid visual exploration of which lineages are shared among or unique to specific subsets of environments; the ability to spawn new windows for investigation of specific subtrees and to view multiple trees at the same time; control over labels and layout features critical for producing publication-quality graphics; the ability to export results in any of several graphical and machine-readable formats for further analysis; and the ability to handle datasets of hundreds of thousands of tips, which can easily be created from larger datasets by OTU picking with UCLUST (Edgar, 2010) or related tools. TopiaryExplorer metadata is provided as tab-separated text, and trees are provided as Newick-formatted strings. These standard file formats allow data generated with different tools to be easily imported into TopiaryExplorer.

TopiaryExplorer is written in Java using Processing, which allows for rapid tree visualization and PDF export using OpenGL. The tree layout rendering algorithms in TopiaryExplorer were adapted from PyCogent (Knight, et al., 2007). Several strategies were applied to efficiently visualize large trees and associated metadata, including caching node lookups rather than running multiple lookups of the same node, and using sparse table representations for storing and accessing metadata.

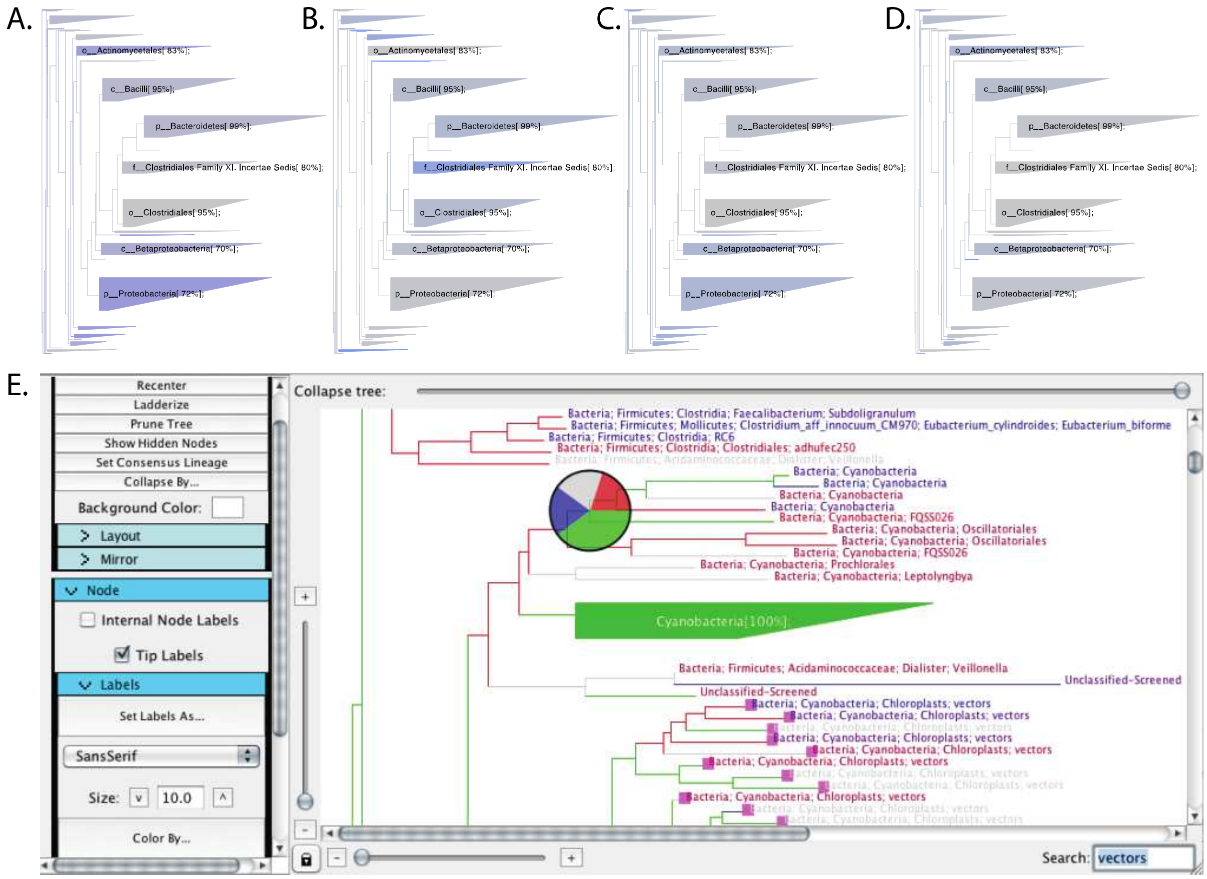
2 RESULTS AND CONCLUSIONS

To illustrate the utility of TopiaryExplorer we applied it to a microbial survey of the hands and keyboards of three individuals (Fierer, et al., 2010). This study illustrated how individuals can be matched to objects they have touched through the use of "microbial fingerprints". It has previously been difficult to identify the specific taxonomic differences between the hand or keyboard microbial communities from different individuals.

[*]To whom correspondence should be addressed.

Downloaded from bioinformatics.oxfordjournals.org at NERL on October 6, 2011

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.


A.
o__Actinomycetales[83%];
c__Bacilli[95%];
p__Bacteroidetes[99%];
f__Clostridiales Family XI. Incertae Sedis[80%];
o__Clostridiales[95%];
c__Betaproteobacteria[70%];
p__Proteobacteria[72%];
B.
o__Actinomycetales[83%];
c__Bacilli[95%];
p__Bacteroidetes[99%];
f__Clostridiales Family XI. Incertae Sedis[80%];
o__Clostridiales[95%];
c__Betaproteobacteria[70%];
p__Proteobacteria[72%];
C.
o__Actinomycetales[83%];
c__Bacilli[95%];
p__Bacteroidetes[99%];
f__Clostridiales Family XI. Incertae Sedis[80%];
o__Clostridiales[95%];
c__Betaproteobacteria[70%];
p__Proteobacteria[72%];
D.
o__Actinomycetales[83%];
c__Bacilli[95%];
p__Bacteroidetes[99%];
f__Clostridiales Family XI. Incertae Sedis[80%];
o__Clostridiales[95%];
c__Betaproteobacteria[70%];
p__Proteobacteria[72%];
E.
Recenter
Ladderize
Prune Tree
Show Hidden Nodes
Set Consensus Lineage
Collapse By...
Background Color:
> Layout
> Mirror
v Node
Internal Node Labels
Tip Labels
v Labels
Set Labels As...
Color By...
Collapse tree:
Bacteria; Firmicutes; Clostridia; Faecalibacterium; Subdoligranulum
Bacteria; Firmicutes; Mollicutes; Clostridium_aff_innocuum_CM970; Eubacterium_cylindroides; Eubacterium_biforme
Bacteria; Firmicutes; Clostridia; RC6
Bacteria; Firmicutes; Clostridia; Clostridiales; adhufec250
Bacteria; Firmicutes; Acidaminococcaceae; Dialister; Veillonella
Bacteria; Cyanobacteria
Bacteria; Cyanobacteria
Bacteria; Cyanobacteria
Bacteria; Cyanobacteria
Bacteria; Cyanobacteria; FQSS026
Bacteria; Cyanobacteria; Oscillatoriales
Bacteria; Cyanobacteria; Oscillatoriales
Bacteria; Cyanobacteria; FQSS026
Bacteria; Cyanobacteria; Prochlorales
Bacteria; Cyanobacteria; Leptolyngbya
Cyanobacteria[100%];
Bacteria; Firmicutes; Acidaminococcaceae; Dialister; Veillonella
Unclassified-Screened
Unclassified-Screened
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors
Bacteria; Cyanobacteria; Chloroplasts; vectors

☐ Internal Node Labels

☑ Tip Labels

SansSerif

Size: 10.0

Search: vectors

Figure 1. Coverage of samples over tree of 16S rRNA OTUs observed in this study. Wedges summarize groups of tips and colored on a grey to blue scale based on the abundance of tips in that wedge which are represented in the given sample type ((A) M3 (subject 3) keyboard, (B) M2 (subject 2) fingertips, (C) M3 fingertips, (D) M9 (subject 9) fingertips). (E) TopiaryExplorer interface: branches are colored by individual (M2, M3, M9), tips labels are colored by source (specific keyboard key or fingertips), and the pie chart summarizes the representation of individuals in the clade. Pink boxes highlight tips matching the search term.

We applied TopiaryExplorer to specifically address this question and find that between-community taxonomic differences are present and easily discernable (Figure 1A-D). The trees visually represent the results of the original study, that M3's keyboard is more similar to M3's fingertips than the fingertips of either M2 or M9, but additionally allows us to immediately determine which taxonomic groups are differentially represented in the different sample types. For example, M3's fingertips and keyboard contain Proteobacteria in higher abundance than M2 or M9. Figure 1E represents a screenshot of the TopiaryExplorer interface.

These results show how visual inspection of phylogenetic trees with environmental data can facilitate interpretation of the relative differences between microbial communities. TopiaryExplorer fills a necessary gap in tools for the comparison of microbial communities. The tree of life is being rapidly filled by large-scale projects such as Genomic Encyclopedia of Bacteria and Archaea (GEBA), the Human Microbiome Project and the Earth Microbiome Project, and annotated with emerging standards such as Minimum Information about any Sequence (MIxS). The ability to determine what lineages are novel in a new dataset, and what lineages distinguish among samples associated with clinical or environmental parameters, will be crucial for understanding the ecology and evolution of the microbes that comprise the vast majority of life on earth.

ACKNOWLEDGEMENTS

We thank Cathy Lozupone, Jessica Metcalf, Laura Wegener Parfrey, Justin Kuczynski, and Jose Clemente for feedback on this manuscript and TopiaryExplorer. *Funding*: This work was supported in part by the Bill and Melinda Gates Foundation, the National Institutes of Health (DK78669, HG4872 HG4866, 3T15LM009451-03S1), and the Howard Hughes Medical Institute.

REFERENCES

Caporaso, J.G., *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data, *Nat Methods*, **7**, 335-336.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics*, **26**, 2460-2461.

Fierer, N., *et al.* (2010) Forensic identification using skin bacterial communities, *Proc Natl Acad Sci U S A*, **107**, 6477-6481.

Huson, D.H., *et al.* (2007) Dendroscope: An interactive viewer for large phylogenetic trees, *BMC Bioinformatics*, **8**, 460.

Downloaded from bioinformatics.oxfordjournals.org at NERL on October 6, 2011

Knight, R., *et al.* (2007) PyCogent: a toolkit for making sense from sequence, *Genome Biol*, **8**, R171.

Letunic, I. and Bork, P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy, *Nucleic Acids Res*.

Ludwig, W., *et al.* (2004) ARB: a software environment for sequence data, *Nucleic Acids Res*, **32**, 1363-1371.

Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere, *Science*, **276**, 734-740.

Tringe, S.G. and Hugenholtz, P. (2008) A renaissance for the pioneering 16S rRNA gene, *Curr Opin Microbiol*, **11**, 442-446.

Downloaded from bioinformatics.oxfordjournals.org at NERL on October 6, 2011