Title

ghost-tree: creating hybrid-gene phylogenetic trees for diversity analyses

Authors

Jennifer Fouquier[1], Jai Ram Rideout[2], Evan Bolyen[2], John Chase[2], Arron Shiffer[2,3], Daniel McDonald[4], Rob Knight[5], J Gregory Caporaso[2,3] and Scott T. Kelley[1,4*]

[1] Graduate Program in Bioinformatics and Medical Informatics, San Diego State University, San Diego, CA, USA.
[2] Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, AZ, USA.
[3] Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA.
[4] Department of Computer Science and BioFrontiers Institute, University of Colorado at Boulder, CO, USA.
[5] Department of Pediatrics, and Department of Computer Science and Engineering, University of California San Diego, San Diego, CA USA.
[6] Department of Biology, San Diego State University, San Diego, CA, USA.

Corresponding Author

Scott T. Kelley
Department of Biology
5500 Campanile Drive
San Diego State University
San Diego, CA, USA, 92182-4614
Phone: +1 619 206 8014
Email: skelley@mail.sdsu.edu

Abstract

ghost-tree is a bioinformatics tool that integrates sequence data from two genetic markers into a single phylogenetic tree that can be used for diversity analyses. Our approach uses one genetic marker whose sequences can be aligned across organisms spanning divergent taxonomic groups (e.g., fungal families) as a "foundation" phylogeny. A second, more rapidly evolving genetic marker is then used to build "extension" phylogenies for more closely related organisms (e.g., fungal species or strains) that are then grafted on to the foundation tree by mapping taxonomic names. We apply ghost-tree to graft fungal extension phylogenies derived from ITS sequences onto a foundation phylogeny derived from fungal 18S sequences. The result is a phylogenetic tree, compatible with the commonly used UNITE fungal database, that supports phylogenetic diversity analysis (e.g., UniFrac) of fungal communities profiled using ITS markers.

Availability: ghost-tree is pip-installable. All source code, documentation, and test code are available under the BSD license at https://github.com/JTFouquier/ghost-tree.

Introduction

In recent years, next-generation sequencing and bioinformatics methods have rapidly expanded our knowledge of the microbial world by enabling marker gene surveys of bacterial, archaeal and eukaryotic microbial communities. Phylogenetic diversity metrics such as *Phylogenetic Diversity* (PD) (Faith, 1992) and *UniFrac* (Lozupone and Knight, 2005) have improved resolution of community differences relative to their non-phylogenetic analogs that were mostly developed for studying communities of macro-organisms (e.g., Chao1 and Bray-Curtis dissimilarity). An ideal marker gene has highly conserved regions that facilitate development of PCR primers and multiple sequence alignment, and highly variable regions that support detailed taxonomic resolution.

For some taxonomic groups, such as the fungi, an ideal marker gene does not exist. The small subunit ribosomal RNA (SSU) is highly conserved in the fungi, and therefore sequences from different species are too similar for detailed taxonomic assignment. The Internal Transcribed Spacer (ITS), a non-coding region found between the ribosomal genes, has thus become popular for achieving high-resolution taxonomic profiles of fungal communities. This marker "gene" is so non-conserved, however, that no meaningful alignment is possible across distant fungal lineages, thus making phylogenetic diversity analyses impossible.

Here we present ghost-tree, an open-source bioinformatics software tool for creating phylogenetic trees using multiple genetic loci. Sequences from a more conserved locus are aligned across distant taxonomic lineages to build a foundation tree, and sequences from a less conserved locus are aligned for many groups of closely related taxa to create extension trees that are then grafted onto the foundation tree. We apply "ghost-tree" to build foundation trees from the fungal 18S rRNA and extension trees from fungal ITS, to create a single "ghost-tree" that can be used in phylogenetic diversity analyses of fungal communities. There are thus two outcomes of this project: a tree for use with popular community analysis tools such as QIIME, and a software package for developing phylogenetic trees for other sets of marker genes.

Methods

ghost-tree takes as input (1) the *Foundation Alignment* (for example, the Silva 18S alignment) where sequences are annotated with taxonomy; (2) the *Extension Sequence Collection* (for example, unaligned ITS sequences from the UNITE database); and (3) a *taxonomy map*, which contains taxonomic annotations of the sequences in (2). The Foundation Alignment is filtered to remove highly gapped and high entropy positions, and FastTree is used to build a phylogenetic tree from the resulting filtered alignment. This is the *Foundation Tree*. In parallel, the *Extension Sequence Collection* is clustered with SUMACLUST http://metabarcoding.org/sumatra resulting in an *operational taxonomic unit (OTU) map* that groups sequences into OTUs by percent identity. For each *Extension Sequence OTU* a consensus taxonomy is determined and OTUs with the same consensus taxonomy are further grouped into a single OTU. The sequences in each OTU are then aligned, and FastTree is applied to build an *Extension Tree*. The OTU's consensus taxonomy is associated with the root of the *Extension Tree*. The taxa at the root of the extension trees are then used to graft the extension tree on to the tip in the foundation tree with the same taxonomy, resulting in the *Ghost Tree*. We applied ghost-tree to build a phylogenetic tree from Silva (Ver. SSU 119.1) 18S sequences (the foundation) (Pruesse *et al*., 2007) and UNITE (Ver. 12_11_otus) ITS sequences (Kõljalg *et al*., 2013). This tree is available in the GitHub repository, and this ghost-tree workflow is illustrated in Supplementary Figure S1.

ghost-tree is hosted on GitHub under the BSD open source software license. It is implemented in Python, using scikit-bio and Click, and adheres to the PEP8 Python style guide. ghost-tree is subject to continuous integration testing using Travis CI which, on

each pull request, runs unit tests with nose, monitors code style using flake8, and monitors test coverage with Coveralls.

Results

To evaluate whether ghost-tree supports improved resolution in studies of fungal community analysis we compiled two real-world ITS sequence datasets: one collection of human saliva samples (Ghannoum *et al*., 2010) and one of surfaces in public restrooms (Fouquier *et al*., in review). These communities are expected to differ substantially from one another (large effect size). We next simulated 10 samples based on each of the samples from these studies using QIIME's simsam.py workflow. This generates sample replicates that are phylogenetically similar to each other. We therefore consider the grouping of each set of simulated samples to be a small effect size. We computed distances between all samples using three approaches: using binary Jaccard, a qualitative non-phylogenetic diversity metric where no tree is required; using unweighted UniFrac, a qualitative phylogenetic metric of beta diversity with a tree generated using MUSCLE and FastTree; and using unweighted UniFrac with a tree generated with ghost-tree.

Figure 1 contains principal coordinates analysis of these data based on binary Jaccard distances (Fig. 1a), UniFrac/ITS distances (Fig. 1b) and UniFrac/ghost-tree distances (Fig. 1c). UniFrac/ghost-tree resolves the small and the large effect sizes (ANOSIM R=0.38 and R=0.48, respectively), while binary Jaccard (R=0.02 and R=0.04, respectively) and unweighted UniFrac/ITS (R=0.19 and R=0.08, respectively) do not.

Summary

Widely used bacterial sequence databases such as GreenGenes (DeSantis *et al*., 2006) annotate 16S rRNA gene data, providing a reference tree and taxonomy for bacterial and archaeal community analysis. ghost-tree facilitates development of phylogentic trees that can be used in a similar way for marker genes that are less conducive to phylogenetic reconstruction at a global scale. We show that, as with bacterial community analysis, incorporating phylogeny in diversity metrics is useful for resolving differences in fungal communities. The Silva/UNITE-based ghost tree presented here integrate into a user's exiting fungal analysis pipelines and the ghost-tree software package can be used to develop phylogenetic trees for other marker gene sets that provide different taxonomic resolution, or for bridging genome trees with amplicon trees.

Acknowledgements

This work was supported in part by a grant from the Alfred P. Sloan Foundation.

Figure Legends

Figure 1. Principal Coordinates comparing samples based on (a) Binary Jaccard distances, (b) UniFrac distances where trees are computed by aligning ITS sequences using MUSCLE, and (c) UniFrac distances where trees are computed using ghost-tree. Blue points are simulated and real human saliva samples, and red points are simulated and real restroom surface samples. Plots were made using EMPeror software (Vázquez-Baeza *et al*., 2013).

Suppl. Figure S1. ghost-tree workflow diagram.

References

DeSantis,T.Z., *et al*. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microb.*, 72, 5069-5072.

Faith,D.P. (1992) Conservation Evaluation and Phylogenetic Diversity. *Biol. Conserv.*, 61, 1-10.

Ghannoum,M.A., *et al*. (2010) Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog.*, 6, e1000713.

Kõljalg,U., *et al*. (2013) Towards a unified paradigm for sequence-based identification of Fungi. *Mol. Ecol.*, 22, 5271-5277.

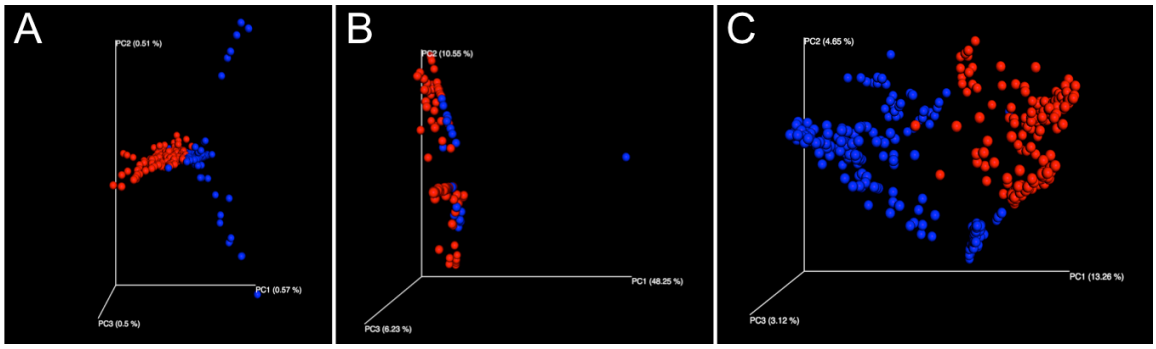
Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microb.*, 71, 8228-8235.

Pruesse,E., *et al*. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, 35, 7188-7196.

Quast,C., *et al*. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, 41, D590-596.

Vázquez-Baeza,E., *et al*. (2013) EMPeror: a tool for visualizing high-throughput microbial community data. *GigaScience.*, 2:16. doi:10.1186/2047-217X-2-16

171 **Figure 1**
172



173
174

Supplemental Figure S1