phylogenize: a web tool to identify microbial genes underlying environment associations

Patrick H. Bradley[1] and Katherine S. Pollard[1,2,3*]

[1] *Gladstone Institute of Data Science and Biotechnology, San Francisco, CA USA*
[2] *Department of Epidemiology & Biostatistics, University of California–San Francisco, CA USA*
[3] *Chan–Zuckerberg Biohub, San Francisco, CA USA*
* *To whom correspondence should be addressed.*

Abstract

Summary: Microbes differ in prevalence across environments, but in most cases the causes remain opaque. Phylogenetic comparative methods have emerged as powerful, specific methods to identify microbial genes underlying differences in community composition. However, to apply these methods currently requires computational expertise and sequenced isolates or shotgun metagenomes, limiting their wider adoption. We present *phylogenize*, a web server that allows researchers to apply phylogenetic regression to 16S amplicon as well as shotgun sequencing data and to visualize results. Using data from the Human Microbiome Project, we show that *phylogenize* draws similar conclusions from 16S and from shotgun sequencing. Additionally, we apply *phylogenize* to 16S data from the Earth Microbiome Project, revealing both known and candidate pathways involved in plant colonization. *phylogenize* has broad applicability to the analysis of both human-associated and environmental microbiomes.

Availability: *phylogenize* is available at `https://phylogenize.org` with source code available at `https://bitbucket.org/pbradz/phylogenize`.

Contact: kpollard@gladstone.ucsf.edu

Introduction

Shotgun and amplicon sequencing have enabled previously intractable microbial communities to be characterized and compared. However, while these communities have the potential to yield clinical (Moayyedi *et al.*, 2015) and agricultural tools (Mendes *et al.*, 2011), translating microbe-to-environment correlations into gene-level mechanisms remains difficult.

Phylogenetic regression is a powerful, underutilized technique (Washburne *et al.*, 2018) that can help interpret these correlations by accounting for the confounder of common descent. Previously, we demonstrated that applying this technique to shotgun metagenomic data can identify microbial genes linked to human body sites without the high false-positive rate of standard regression (Bradley *et al.*, 2018).

Here, we present *phylogenize*, a web tool that makes this technique available to researchers without specific expertise in this area by allowing them to upload and analyze their own data. We also provide the source code of *phylogenize*, allowing more experienced users to run it locally.

In addition to shotgun metagenomic data, *phylogenize* also allows researchers to analyze abundances derived from 16S amplicon sequencing. 16S data is much less expensive to generate and already exists for many environments, allowing researchers to get more from their data.

Overview

phylogenize (Figure 1) takes the following basic inputs. First, users provide a table of taxon abundances across a set of samples. These taxa should be ASVs from DADA2 (Callahan *et al.*, 2016) or Deblur (Amir *et al.*, 2017) (for 16S data) or MIDAS species (for shotgun data). Second, users provide a table of sample annotations matching sample IDs to environments and datasets. The abundances and sample annotations can be provided separately or as a single BIOM-format (McDonald *et al.*, 2012) file.

Next, the user selects one environment out of those represented in the sample annotations. Finally, the user chooses whether to link gene presence to prevalence (the frequency a microbe is observed in the selected environment) or specificity (how specific a microbe is for the chosen environment compared to all others: see Bradley *et al.*, 2018).

phylogenize uses the fast mapper BURST (Al-Ghalith and Knights, 2017) to map sense or anti-sense ASVs to individual PATRIC genomes (Wattam *et al.*, 2014), using a cutoff of 98.5% nucleotide identity (Rodriguez-R *et al.*,

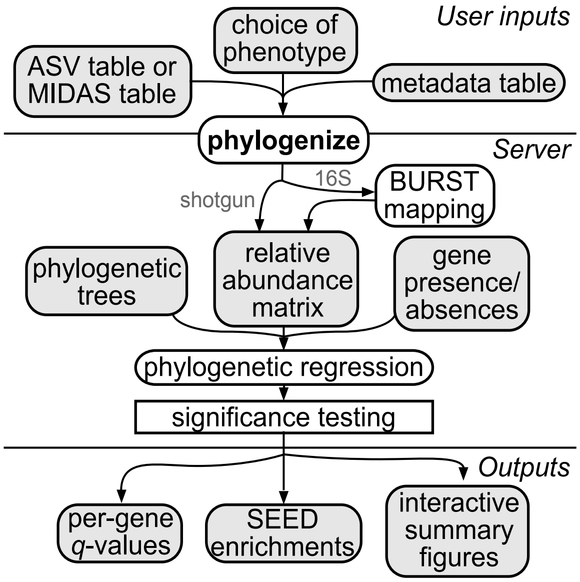

Figure 1: Schematic showing *phylogenize* pipeline. Dark gray indicates user-provided data or options; light gray indicates data included in *phylogenize*.

2018), then matches these genomes to MIDAS species (which are clusters of PATRIC genomes). Reads for sequences mapping to the same species are summed within samples.

The web front-end for *phylogenize* is written in Python using the Flask framework with a Beanstalk-based queueing system. For each job, *phylogenize* uses RMarkdown (Allaire *et al.*, 2018) and knitr (Xie, 2014) to generate an HTML report. This report includes interactive trees showing the phenotype's phylogenetic distribution, heatmaps of significantly positively-associated genes, and tables showing which SEED subsystems (Overbeek *et al.*, 2005) were significantly enriched at a 25% FDR. *phylogenize* also provides tab-delimited files containing the calculated phenotype, p-values and effect sizes for all FIGfams tested, and protein annotations for the significant, positively-associated hits.

Example Applications

Human Microbiome Project comparison: We first used *phylogenize* to associate gene presence-absence with microbial prevalence in the gut. To do so, we used 454 16S amplicon sequencing data from the Human Microbiome Project (HMP) (Human Microbiome Project Consortium, 2012). 6,577 samples from 192 individuals across 16 sites were downloaded from the Sequence Read Archive and denoised with DADA2 (Callahan *et al.*, 2016). Reads were combined for all samples from the same individual and site.

Previously, we performed a similar analysis using HMP's shotgun sequencing data (Bradley *et al.*, 2018), which we use here as a benchmark. Despite differences in read depth and technology, species prevalence estimates obtained by mapping 16S ASVs to MIDAS genomes were similar to those from shotgun sequencing (r = 0.6), and the effect sizes calculated for genes associated with gut prevalence were also broadly similar ($0.339 \leq r \leq 0.601$, Figure S1). When we compared the significantly-associated genes, we also observed shared pathway enrichments, including for genes in the SEED subsystems ”Sporulation gene orphans” in Firmicutes ($q_{\text{shotgun}} = 2.7 \times 10^{-22}$, $q_{16S} = 0.019$), and ”Type III, Type IV, Type VI, ESAT secretion systems” in Proteobacteria ($q_{\text{shotgun}} = 1.69 \times 10^{-11}$, $q_{16S} = 2.23 \times 10^{-6}$).

Earth Microbiome Project: The Earth Microbiome Project (EMP) (Thompson *et al.*, 2017) comprises 16S data sampled across many biomes and habitats. Using the balanced subset of 2,000 samples processed using Deblur (Amir *et al.*, 2017), we calculated a specificity score for being plant-associated, as opposed to being animal-associated or free-living. *phylogenize* identified genes enriched in processes known to be relevant to a plant-associated lifestyle, such as nitrogen fixation (Mylona *et al.*, 1995), the metabolism of opines (metabolites whose biosynthesis in plants is induced by parasitic *Agrobacterium* species (Schell *et al.*, 1979)), and xylose metabolism (xylose is a plant cell wall component: Liu *et al.*, 2015).

Conclusion

Phylogenetic regression offers a computational way to identify genes potentially involved in site colonization, even for clinically or ecologically important microbes that are poorly characterized and/or experimentally intractable. Previously, applying this method to microbiome data required specialized computational expertise and either shotgun metagenomics data (Bradley *et al.*, 2018) or a large collection of sequenced isolates (Levy *et al.*, 2018). By making it significantly easier to analyze either 16S or shotgun data with phylogenetic regression, *phylogenize* expands the toolkit for researchers studying microbial communities.

Acknowledgements

The authors acknowledge members of the Pollard, Hernandez, Wall, and Turnbaugh labs for helpful discussions.

Funding

Funding for this research was provided by the National Science Foundation [DMS-1069303 and DMS-1563159]; Gordon & Betty Moore Foundation [#3300]; and the Gladstone Institutes.

Conflict of Interest: none declared.

References

Al-Ghalith, G. and Knights, D. (2017). BURST enables optimal exhaustive DNA alignment for big data. DOI: `doi.org/10.5281/zenodo.806850`

Allaire, J. J. *et al.* (2018). *rmarkdown: Dynamic Documents for R.*

Amir, A. *et al.* (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, **2**(2), e00191–16.

Bradley, P. H. *et al.* (2018). Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLOS Comput. Biol.*, **14**(8), e1006242.

Callahan, B. J. *et al.* (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods*, **13**(7), 581–583.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, **486**(7402), 207–14.

Levy, A. *et al.* (2018). Genomic features of bacterial adaptation to plants. *Nat. Genet.*, **50**(1), 138–150.

Liu, Y. *et al.* (2015). Molecular mechanisms of xylose utilization by Pseudomonas fluorescens: overlapping genetic responses to xylose, xylulose, ribose and mannitol. *Mol. Microbiol.*, **98**(3).

McDonald, D. *et al.* (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, **1**(1), 7.

Mendes, R. *et al.* (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science*, **332**(6033), 1097–1100.

Moayyedi, P. *et al.* (2015). Fecal microbiota transplantation induces remission in patients with active ulcerative colitis in a randomized controlled trial. *Gastroenterology*, **149**(1), 102–109.e6.

Mylona, P. *et al.* (1995). Symbiotic Nitrogen Fixation. *Plant Cell*, **7**(7), 869–885.

Overbeek, R. *et al.* (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**(17), 5691–5702.

Rodriguez-R, L. M. *et al.* (2018). How much do rRNA gene surveys underestimate extant bacterial diversity? *Appl. Env. Microbiol.*, **84**(6), AEM.00014–18.

Schell, J. *et al.* (1979). Interactions and DNA transfer between Agrobacterium tumefaciens, the Ti-plasmid and the plant host. *Proc. R. Soc. Lond. B. Biol. Sci.*, **204**(1155), 251–66.

Thompson, L. R. *et al.* (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, **551**(7681), 457.

Washburne, A. D. *et al.* (2018). Methods for phylogenetic analysis of microbiome data. *Nat. Microbiol.*, **3**(6), 652–661.

Wattam, A. R. *et al.* (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**(Database issue), D581–91.

Xie, Y. (2014). *knitr: a comprehensive tool for reproducible research in R.*

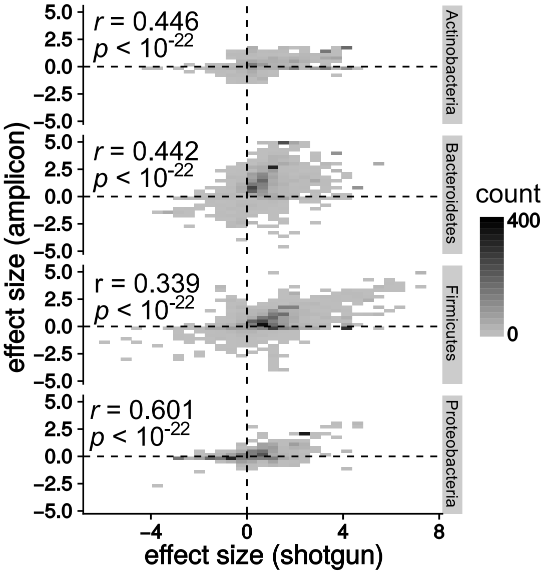

Figure S1: ***phylogenize* makes similar inferences from 16S and shotgun data.** On the x-axis are effect sizes of genes associated with gut prevalence using shotgun data from HMP; the y-axis has effect sizes derived from 454 16S data. Only genes significant at $q \leq 0.05$ in at least one dataset are shown.