bioRxiv preprint doi: https://doi.org/10.1101/425231; this version posted May 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

phylogenize: correcting for phylogeny reveals genes associated with microbial distributions

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.*

May 30, 2019

Abstract

Summary: Phylogenetic comparative methods are powerful but presently under-utilized ways to identify microbial genes underlying differences in community composition. These methods help to identify functionally important genes because they test for associations beyond those expected when related microbes occupy similar environments. We present *phylogenize*, a pipeline with web, QIIME2, and R interfaces that allows researchers to perform phylogenetic regression on 16S amplicon and shotgun sequencing data and to visualize results. *phylogenize* applies broadly to both host-associated and environmental microbiomes. Using Human Microbiome Project and Earth Microbiome Project data, we show that *phylogenize* draws similar conclusions from 16S versus shotgun sequencing and reveals both known and candidate pathways associated with host colonization.
Availability: *phylogenize* is available at https://phylogenize.org and https://bitbucket.org/pbradz/phylogenize.
Contact: kpollard@gladstone.ucsf.edu

Introduction

Shotgun and amplicon sequencing allow previously intractable microbial communities to be characterized and compared, but translating these comparisons into gene-level mechanisms remains difficult. Researchers typically correlate microbial gene abundance with environments using metagenomes, either from shotgun sequencing (Nayfach and Pollard, 2016) or imputed from amplicon sequences (Langille *et al.*, 2013; Aßhauer *et al.*, 2015). However, related microbes tend to both share genes and occupy similar environments, causing confounding. Phylogenetic methods can correct for such confounding in metagenomics data (Bradley *et al.*, 2018), but are currently implemented only in command-line, computationally intensive software.

We developed *phylogenize*, a pipeline allowing researchers without specific expertise in phylogenetic regression to analyze their own data via the web, an R package (R Core Team, 2017), or the popular microbiome workflow tool QIIME2 (Bolyen *et al.*, 2018). An important innovation specific to *phylogenize* is that input data can be shotgun metagenomes or 16S amplicon data, the latter being lower-cost and available for more environments. Using these taxonomic profiles and sample environments (i.e., sources), the tool returns genes associated with differences in community composition across environments.

Overview

Users provide *phylogenize* with taxon abundances and sample annotations, in tabular or BIOM (McDonald *et al.*, 2012) format. Shotgun data should be mapped to MIDAS species (Nayfach *et al.*, 2016); amplicon data should be denoised to amplicon sequence variants (ASVs) with DADA2 or Deblur. *phylogenize* uses BURST (Al-Ghalith and Knights, 2017) to map ASVs to MIDAS species via individual PATRIC genomes (Wattam *et al.*, 2014), using a default cutoff of 98.5% nucleotide identity (Rodriguez-R *et al.*, 2018) and summing reads mapping to the same species. Taxa are linked to genes using MIDAS and PATRIC, and then gene presence is tested for association with one of two phenotypes: prevalence (frequency microbes are observed) or specificity (enrichment of microbes relative to other environments; see Bradley *et al.*, 2018).

phylogenize is an R package with a QIIME2 wrapper written in Python and a web front-end written in Python with the Flask framework (Ronacher, 2018) and a Beanstalk-based queueing system (Rarick, 2014). *phylogenize* reports include interactive trees showing the phenotype's phylogenetic distribution, heatmaps of significantly positively-associated genes, tables showing which SEED subsystems (Overbeek *et al.*, 2005) are significantly enriched, and links to tab-delimited files containing complete results.

Example Applications

Human Microbiome Project

The Human Microbiome Project (HMP; Human Microbiome Project Consortium, 2012) collected both 16S amplicon and shotgun sequences from 16 body sites on 192 individuals. Shotgun data processing was previously described (Bradley *et al.*, 2018). 6,577 amplicon samples were downloaded from the NCBI SRA and denoised with DADA2 (Callahan *et al.*, 2016), combining reads from the same individual and site. We ran *phylogenize* on both data types to identify genes whose presence is associated with prevalence in the gut. Despite differing read depth and sequencing technology (454 versus Illumina), effect sizes for genes associated with gut prevalence were similar for amplicon and shotgun ($0.339 \leq r \leq 0.601$) and similar pathways were enriched (Figure 1A).

Earth Microbiome Project

The Earth Microbiome Project (EMP) (Thompson *et al.*, 2017) comprises 16S data from many biomes and habitats. Using the balanced subset of 2,000 samples processed using Deblur (Amir *et al.*,

bioRxiv preprint doi: https://doi.org/10.1101/425231; this version posted May 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

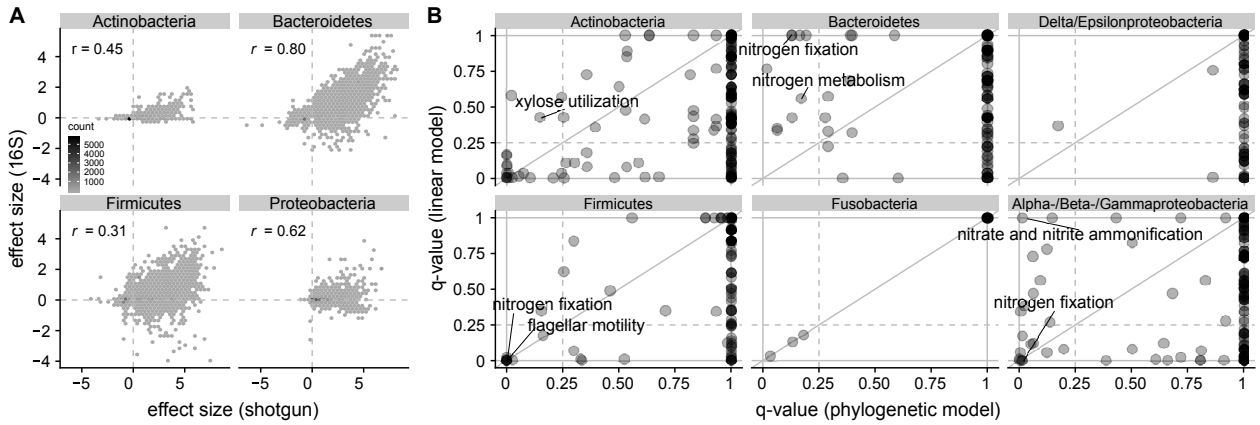

Figure 1: A. Effect sizes from HMP shotgun (x-axis) versus 16S amplicon (y-axis) data are correlated. Genes with $q < 0.05$ in one or both analyses shown with their Pearson correlation. Examples of SEED subsystems enriched for positively-associated genes with both data types include "Sporulation gene 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 ($q_{\text{shotgun}} = 1.69 \times 10^{-11}$, $q_{16S} = 2.23 \times 10^{-6}$). B. SEED enrichments in EMP data using *phylogenize* (x-axis; 61 subsystems) or a linear model (y-axis; 202 subsystems). Many shared subsystems are relevant to a plant-associated lifestyle, such as nitrogen fixation (Mylona *et al.*, 1995) and the metabolism of xylose (a pentose component of plant cell walls, Liu *et al.*, 2015). Selected enrichments labeled; full list in Supplemental Table 1.

2017), we ran *phylogenize* and linear models (no phylogenetic correction) to identify genes whose presence is specific to plant rhizosphere compared to other environments. Linear models identified many more positively-associated genes (24,728 versus 7,490, $q \leq 0.05$), but these discoveries were less enriched for processes known to be linked to plant rhizospheres (Figure 1B), suggesting dilution by false positives, as previously seen in HMP shotgun data and simulations (Bradley *et al.*, 2018).

Conclusion

Many microbes of interest to clinicians, ecologists, and microbiologists are poorly characterized or experimentally intractable. By making it easier to analyze either 16S or shotgun data with more precise statistical tools, *phylogenize* expands the toolkit for identifying mechanisms driving differences in microbial community composition.

Acknowledgements

Funding from the National Science Foundation [DMS-1069303 and DMS-1563159] and Gordon & Betty Moore Foundation [#3300]. *Conflict of Interest*: none declared.

References

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

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

Aßhauer, K. P. *et al.* (2015). Tax4Fun: Predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*, **31**(17), 2882–2884.

Bolyen, E. *et al.* (2018). QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints*.

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

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.

Langille, M. G. *et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, **31**(9), 814–821.

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.

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

Nayfach, S. and Pollard, K. S. (2016). Toward Accurate and Quantitative Comparative Metagenomics. *Cell*, **166**(5), 1103–1116.

Nayfach, S. *et al.* (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.*, **26**(11), 1612–1625.

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.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rarick, K. (2014). Beanstalkd [computer software] v1.10; https://beanstalkd.github.io.

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

Ronacher, A. (2018). Flask [computer software] v1.0.2; https://www.palletsprojects.com/p/flask.

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

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