

Title	PhyloToAST: Bioinformatics tools for species-level analysis and visualization of complex microbial communities
Author	Shareef M. Dabdoub and Purnima S. Kumar
Affiliation	The Ohio State University College of Dentistry Columbus, OH, USA
Contact	dabdoub.2@osu.edu
URL	http://phylotoast.readthedocs.org/
License	MIT

Purpose: Understanding human-associated microbial ecology is essential for insight into health, as well as identifying disease states, risk factors, and etiology. The 16S ribosomal RNA gene is the most common genetic marker for taxonomic identification due to its near universal presence and static function over time, and a wide variety of tools exist for quantifying samples. The popular QIIME software was developed to gather such tools into a single, usable pipeline, but is less usable when species-level analysis is important; as is the case with highly complex oral biofilms. We have developed a new set of tools, called PhyloToAST, that wholly integrate with the QIIME pipeline to reduce primer bias, enhance species-level analysis and visualization, and greatly improve analysis speed.

Methods: All tools were developed using the Python programming language, and use QIIME output files as input, but do not directly depend on a QIIME installation. All analysis and generation of data for visualization was performed with resources provided by the [Ohio Supercomputer Center](#).

Results: Our new pipeline was applied to three oral microbiome datasets examining subgingival bacterial community composition in smokers and non-smokers, diabetics, and dental implant recipients. Compared with the QIIME-provided Amazon EC2 instance, our pipeline reduced processing time for 2 million 16S sequences from over one week to a matter of hours. Furthermore, we developed and applied algorithms to reduce single-primer bias, condense redundant taxonomic output, and automatically generate visual representations of phylogenetic quantification and comparison.

Conclusion: We have developed a set of analysis tools (PhyloToAST) targeted at enhancing taxonomic identification of microbial samples while using the QIIME software pipeline. When applied to several large samples of the oral microbiome, analysis times were drastically decreased and species-level analysis was substantially enhanced. Our tools integrate with the QIIME pipeline and are available free and open source (MIT license).

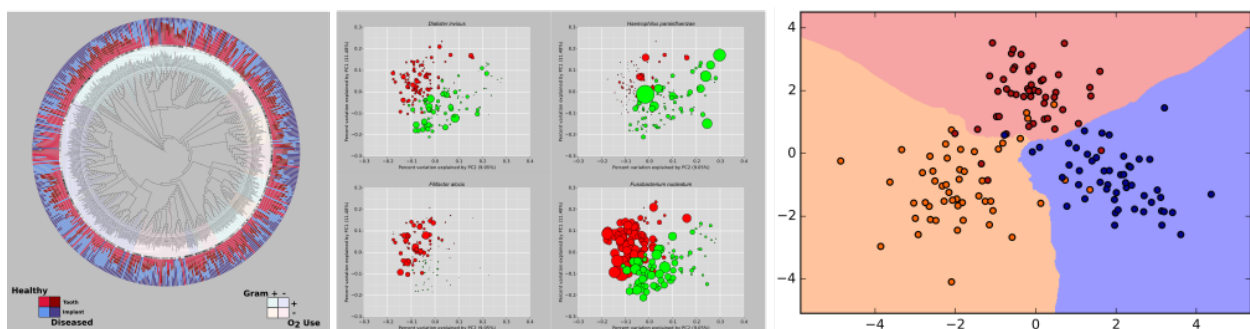


Figure 1: Examples of visualizations created with PhyloToAST. From left to right: A phylogenetic tree with stacked bar charts providing species-level abundance comparisons between four groups. Next is a series of PCoA plots, based on UNIFRAC distances, visually representing the abundance of several species between smokers and non-smokers. Finally, a k-Nearest Neighbors classification of microbial composition data. The input into the KNN algorithm was high-dimensional OTU data simplified with Linear Discriminant Analysis.