An integrated pipeline for large-scale phylogenetic characterization of genomes and metagenomes

Asnicar F(1)[†] , Beghini F(1), Bolzan M(1), Manara S(1), Pasolli E(1), Mirarab S(2,3), Huttenhower C(4,5), Segata N(1)

(1) Centre for Integrative Biology, University of Trento, Italy
(2) Center for Microbiome Innovation, University of California San Diego, La Jolla, California, USA
(3) Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, California, USA
(4) Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, USA
(5) The Broad Institute of MIT and Harvard, Cambridge, USA

[†] Email: f.asnicar@unitn.it

Motivation
Modern sequencing technologies are extensively applied in metagenomics studies on human and environmental microbiomes, and are generating an increasingly large amount of genomic information from thousands of microbial species. However, these data often consist of fragmentary genomes and contigs missing taxonomic labels, and phylogenetic profiling is in many cases the only possible approach for contextualizing them within characterized microbial genomes. This is especially true for the so-called "microbial dark matter", which contains the genomes of organisms without relevant similarity to any previously labeled microorganism. An additional current challenge in microbial genomics is the efficient phylogenetic characterization of thousands of whole microbial genomes. This applies also to pathogen transmission inference, microbial population genomics, and evolutionary studies. Despite a number of methods developed to tackle some of these challenges, at present, there is no comprehensive, integrated, and automatic approach able to phylogenetically model large sets of (meta)genomic data. Current phylogenetic pipelines like AMPHORA2 (Wu et al., 2012), PhyloSift (Darling et al., 2014), and PhyloPhlAn 1.0 (Segata et al., 2013), cannot accurately reconstruct strain-level phylogenies, do not scale to large datasets, and can only be used for building holistic phylogenies.
Methods
We introduce PhyloPhlAn 2.0, a new integrated pipeline for accurate automatic inference of strain-level phylogenies, fast and scalable reconstruction of evolutionary trees of life, and taxonomic placement of (draft) genomes and single contigs from isolate sequencing and metagenomic assemblies. PhyloPhlAn 2.0 is partially based on the previously developed PhyloPhlAn 1.0 package and automatizes the steps for a phylogenetic analysis, from the identification and extraction of the marker genes identified in the input genomes, to multiple-sequence alignment and phylogeny

inference. The pipeline is fully customizable through a configuration file and several input parameters. Customization options include preferences for the software to be used for marker mapping, multiple-sequence alignment, and phylogeny reconstruction, and the thresholds used for the newly implemented methods for several cleaning, trimming, and quality-control steps. PhyloPhlAn 2.0 has been designed to perform either a concatenation-based approach or a supertree pipeline based on summary methods. Moreover, it precomputes marker genes for all species and higher-level clades, without the need to identify phylogenetically relevant loci for each organism of interest. PhyloPhlAn 2.0 allows also to easily and automatically integrate the user's input genomes with thousands of genomes available in public databases.

Results

PhyloPhlAn 2.0 is available at https://bitbucket.org/nsegata/phylophlan under the "dev" branch, and provides two databases of universal markers for prokaryotes: the 400 PhyloPhlAn markers (Segata et al., 2013) and the 136 genes from AMPHORA2 (Wu et al., 2012). We successfully applied it to 4088 high-quality reference genomes from NCBI to build an updated bacterial tree of life. We also used it for estimating several strain-level phylogenies and to phylogenetically place metagenomic assemblies into the newly reconstructed bacterial tree of life. For the metagenomic application, many of the reconstructed genomes that could not confidently be assigned to any known species with at least 95% genome similarity, were accurately phylogenetically placed into the tree of life. For the study of mother-to-infant vertically inherited microbes, we used PhyloPhlAn 2.0 to analyze at strain-level resolution the transmission events of unknown microbial genomes reconstructed from metagenomes (see Figure). Overall, we proved that PhyloPhlAn 2.0 is a reliable and customizable approach to tackle a wide range of phylogenetic tasks and partially solve the problem of the microbial dark matter in the field of microbiome research.

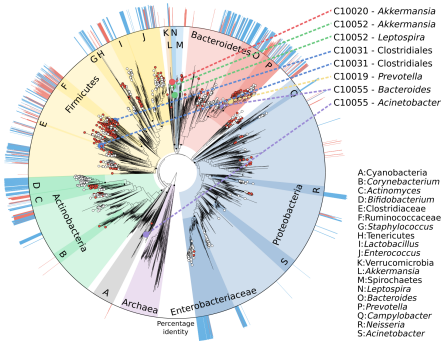

Figure 1: Phylogenetic placement of 1,132 metagenomically-reconstructed genomes and mother-to-infant transmission of taxonomically uncharacterized strains. We used PhyloPhlAn to place the 1,132 genomes reconstructed from 216 metagenomes (using metaSPAdes (Nurk et al., 2017) and binned with MetaBAT2 (Kang et al., 2015)) on the microbial 'tree of life' (Ciccarelli et al., 2006; Segata et al., 2013), which encompasses more than 4000 species with available reference genomes. Leaf nodes without circles refer to reference genomes from known species; white circles indicate reconstructed genomes that are close (>95% identity) to a known species, and red circles show reconstructed genomes that cannot be assigned (<95% identity) to known species. Phyla, families, and genera of interest are highlighted in the phylogeny, and the eight events of mother-to-infant transmission of strains from yet-to-be-described species are called out on the top right. The external ring of the phylogeny reports the percent identity of each leaf node against the closest genomes from known species (values below 95% are shown in red).