Metagenomic microbial community profiling using unique cladespecific marker genes (i.e. MetaPhlAn)

Dreycey Albin (da39)

February 19, 2019

Scientific question

The scientific for this paper is focused on *making a program, termed MetaPhlAn, that can both identify and utilize marker genes to measure relative abundance in a metagenomic sample.* Because of the importance of understanding the composition of a metagenomic sample to study disorders and the environment, many have undertaken different techniques to study the relative abundance of a metagenomic sample. Namely, laboratory techniques that use 16S ribosomal RNA have been implemented, which do fairly well at characterizing the relative abundance of a sample with little information. There have also been other computational methods made to accomplish this goal, however most of these techniques contain costly time complexities.

Hypothesis tested

The hypothesis tested in this research is that *MetaPhlAn is able to uniquely identify multiple species within a metagenomic sample.* Their method for testing this hypothesis consists of comparing their program to other programs for estimating relative abundance, and running their program on simulated and real, not yet defined, metagenomic samples.

Methods

The methods section for this paper is quite extensive, taking up the later half of the paper. There are three inputs that go into MetaPhlAn to create the marker gene database: (1) The raw nucleotide sequences; (2) the CDS calls; and lastly, (3) the taxonomic classification of the genomes. Per species/genome, all of the CDS sequences are clustered together, then the representative seeds from the clustering, for all of the genomes, are clustered together. This is repeatedly performed, while climbing up the taxonomic tree, to identify signature sequences for all of the taxonomic clades. As the genes are clustered and made sure to be repeated within the species genomes, these are defined as "core gene". Next these core genes are further tested for uniqueness, thus ridding of all multicopy genes (even like that of 16S rRNA- which is a multicopy gene). This portion of the process is performed by the MetaPhlAn team, so the user does not have to worry about this portion of the process. After the gene marker catalog has been created, different metagenomic samples can be mapped to the marker genes. This part of the process essentially counts the number of hits to a specific marker gene and then assigns a number to that taxonomic position. Thereafter, a relative abundance is assigned based on the number of counts to a specific species/genus.

Results

The major result in the paper is the comparison of MetaPhlAn with the other available programs for measuring relative abundance for a metagenomic sample. It was there that they were able to show that MetaPhlAn ran nearly 100x faster than the nearest competing program, PhyloPythiaS, which only is able to compute genus level abundance levels. In addition to the speed differences, the authors convincingly showed that the ranked error distribution was lower for MetaPhlAn than that of the of the other programs. I personally think a great way to think about this could be the integral of the MetaPhlAn distribution, over the entire error rank domain, and that it was smaller for MetaPhlAn than that of the other competing programs. In addition, the authors of the paper tried MetaPhlAn on several different metagenomic samples: (1) Vaginal samples, (2) Soil samples, and (3) gut microbiota samples. The MetaPhlAn results for the vaginal samples were compared to 16S rRNA results, and it showed that there was significant overlap between abundance levels for both of the results. For the gut microbiota sample, it was shown that the abundance level recapitulate what is expected to be in the human gut, and it also showed that there are specific microbial communities within the gut.

Key implications of the results

Overall, the key implication for the research discussed in this paper is an improved pipeline for retrieving marker sequences for a database of species. The subsequent taxonomic clustering also delivers a unique approach to classify the metagenomic samples at higher taxinomic levels. The significance of their approach is a major improvement for the time, as 16S rRNA was the most commonly used method at the time the paper was published.