Enterotypes of the human gut microbiome

Critique

Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier,
Denis Le Paslier, Takuji Yamada, Daniel R. Mende, *et al.*

December 6, 2012

Metagenomics offers a powerful way for viewing the microbial communities bypassing the need to isolate and culture individual species. This paper describes an approach to identification and analysis of enterotypes present in human gut published in 2011. The study based on data of tract microbiome of 39 individuals from six geographical regions (Japanese, American, Danish, French, Italian, and Spanish). Some of data was already sequenced with illumina and pyrosequncing technology. Other samples were sequenced for this study using Sanger technology. The authors claim that comparative analysis using the procedures was not biased by data-set origin. However the comparison touched only 454 and Sanger sequencing datasets and didn't touch the illumina technology. The authors also recognize the age bias in the dataset, since samples in enterotype 1 enriched with Japanese individuals was younger than the rest of the dataset.

Besides, they do not consider nutritional habits and way of life (smoking/non smoking person) which might also impact the observations.

All the data was aligned against human genome assembly *hg18* using BLAT alignment tool. Possible human DNA sequences were identified with a very low alignment threshold to maximize true positives and minimize false negatives (pslFilter -minMatch=50 from the BLAT package), and were removed. Therefore they might have discarded lots of bacterial data, which should affect the calculated statistics.

The next step was classification of the selected data to different kinds of bacterial genera. They mapped the sequenced metagenomic reads to 1,511

reference microbal genomes. To consistently estimate the functional composition of the samples they also predicted genes in the sequenced data. Orthologous group abundance patterns of the predicted genes agree with observations made in previous studies, for example, histidine kinases make up the largest group.

The formula for abundance calculation is provided, which uses the notion of overlap between two protein sequences. However it is not absolutely clear how they calculate the overlap. Since first they align the sequences with BLASTP, it could be the number of matches. In any case this issue might impact the method sensitivity.

The goal of the study was to divide the samples into clusters. They used abundance of the assigned orthologous groups as the metric for clustering, the distance between samples was calculated as the Jensen-Shannon divergence, which is based on Kullback-Leibler divergence which is known to be a native approach to estimation of the difference between probability distributions.

The multidimensional cluster analysis and principal component analysis revealed that there are three distinct clusters which are actually very good represented on the provided figures.

Authors tried to analyze the biomarkers, such as ratio between body mass index and firmicutes presented in samples, however it does not reveal a correlation. They used Spearman pairwise correlations methods with the Benjamini-Hochberg procedure for multiple testing correction, which also seems to be a natural approach.

To sum up the paper discusses a pipeline to identify and analyze enterotypes present in a set of samples in very clear and sufficiently detailed manner.