Linking microbial community composition and potential functional roles using shotgun metagenomic libraries

Laura Rabelo Leite [1] , Julliane Dutra Medeiros [1,2], Francislon Silva de Oliveira [1,2], Victor Satler Pylro [1], Sara Cuadros Orellana [1] , Guilherme Corrêa de Oliveira [3], Gabriel da Rocha Fernandes [1]

Centro de Pesquisas Rene Rachou – Fiocruz Minas [1], Universidade Federal de Minas Gerais [2] and Vale Institute of Technology – Biodiversity and Biotechnology [3]

Metagenomics involves the study of genetic material recovered directly from environmental sample, allowing microbiologists to analyze not cultivable organisms. Shotgun metagenomic reads can be taxonomically or functionally classified. Classification methods can be divided in three strategies: (a) sequence similarity methods, which use the results of a sequence similarity search against a database of a reference set of sequences, (b) sequence composition methods, which are based on characteristics of their nucleotide composition, (c) marker-based methods which identify species based on the occurrence of specific marker sequences. However, the accurate identification of microorganisms at the species level remains extremely challenging. Here, we proposed a pipeline to analyze, in house, shotgun metagenomic reads (classifying them taxonomically and functionally), resulting in a relational database to link these information. Raw reads were subject to quality filtering using Trimmomatic. A taxonomic profile was obtained by comparison of all reads against the NT database, and the Nucleotide BLAST output file was used to calculate a taxonomic classification based on the lowest common ancestor method. Contigs larger than 500bp were obtained with SPADES assembler. Coding sequences were predicted with MetaGeneMark. Proteins were extracted using an in-house Perl script and classified into UniRef Enriched KEGG Orthology using DIAMOND. The KO counts were normalized according to the length of the target gene. Reads were then mapped back onto the contigs, using Bowtie2, to determine the contribution of each taxon to the environment gene pool. Our pipeline allows the inference of direct relations between taxonomic and functional data in high-throughput libraries. Moreover, is possible compare the relative abundance of organisms and metabolic pathways between environments and look for taxa that carry out functions of interest. Our pipeline was evaluated against a mock community metagenome, downloaded from Human Microbiome Project (SRR172903), and compared to two approaches: MetaPhlAn2 (based on marker genes) and GSM (based on k-mer composition). The sensitivity evaluation suggested that our pipeline was sufficient to identify microbial strains with ≥0.2× coverage, and 63% of selected genus should be detected based on the positive predictive value. Still, GSM showed better specificity results (51%), due to low false positive rate, highlighting the indicator that needs to be improved in our pipeline. The proposed pipeline was also used to build the first catalog of the biotechnological potential of microorganisms in brazilian copper mines, identifying about 3,800 potential commercial applications related to CUEs (Commercially Useful Enzymes).