Optimized bioinformatics pipeline for fungal microbiota analyses using combined databases

Camargo-Penna, PH; Soares, RC; Braz, ASK; Paulino, LC

Centro de Ciências Naturais e Humanas (CCNH), Universidade Federal do ABC (UFABC), Santo André-SP, Brazil

pennaphc@gmail.com

Next Generation Sequencing (NGS) has been used to study microbial communities from various sources, changing our vision about microbial diversity. The large amount of data generated by NGS requires bioinformatics techniques. Bacterial microbiota has been more studied in comparison to Fungi. Therefore bioinformatics parameters and strategies are less established for Fungi and the available bioinformatics tools are time consuming. Moreover, currently available databases for Fungi taxonomy are less informative than for Bacteria. So metagenomics analysis may not reveal a complete panorama of fungal communities. The aim of this study was to implement a pipeline for fungal OTU assignment that allows a larger and more precise panorama of fungal communities in less time and consumption of computational power. Fungal ITS1 rDNA from 48 skin samples were amplified and sequenced using Illumina MiSeq platform. Sequences were analyzed in a 3 step pipeline: (1) taxonomic assignment using BLAST implemented in QIIME package against reference sequences from UNITE fungal database; (2) clustering of ITS1 sequence reads that did not match to any sequence from UNITE database, using CD-HIT; (3) online BLAST analysis of representative sequences of each cluster against NCBI database. Approximately 1.4 million fungal sequences were obtained after size and quality filtering. Taxonomic assignment was not possible for near 40% of the sequences in step 1. Sequences that were not assigned were then subjected to step 2. CD-HIT clustering allowed a 10-fold reduction of sequences to be analyzed by BLAST, decreasing computational time, against NCBI database at step 3. BLAST comparisons against NCBI allowed identification of organisms that were not in the UNITE database, including uncharacterized organisms and potential contaminants. By using a combined reference database strategy, this pipeline reduces computational time and also permits a broad and accurate panorama of fungal communities. This approach could be employed for other organism sequence databases, allowing an optimized taxonomic assignment for metagenomic analysis.

Key-words: metagenomics, taxonomic assignment, pipeline optimization, Fungi, next generation sequencing.

Financial support: L'Oréal Research & Innovation