Metagenomic analysis of the Southern Brazilian Atlantic Forest soil using next-generation sequencing technologies

[1]Janynne Palheta, [2]Michelle Zibetti Tadra-Sfeir, [2]Emanuel Maltempi de Souza, [2]Fábio de Oliveira Pedrosa, [1,3]Helisson Faoro

[1]Laboratory of Bioinformatics, Professional and Technological Education Sector, Universidade Federal do Paraná, [2]Department of Biochemistry and Molecular Biology, Universidade Federal do Paraná, [3]Laboratory of Gene Expression Regulation, Carlos Chagas Institute, Fiocruz-PR

Metagenomics allows the direct access to the DNA of the environmental bacterial communities without cultivation. Appling the Next Generation Sequencing technology (NGS) to environmental DNA has been provide precise information about the species that are present in a specific environment (microbiota) and the genes that these microorganisms are caring (microbiome). In this work, we used the MiSeq and Ion Proton platforms to sequence the total DNA and the 16S rRNA gene from soil samples of the Southern Brazilian Atlantic Forest. The first group was formed by samples MA02, MA05 and MA07, collect in the winter of 2004 at 900, 653 and 32 meters of altitude. The second group was formed by samples MAF1, MAF2 and MAF3, collect at the same site of the first group in the summer of 2007. The total DNA from all six samples were purified using MoBio Power Soil kit and the 16S rRNA gene was amplified using universal primers customized with the Illumina adaptors sequence. The analysis of the resulting data, using QIIME package, revealed the presence of 33 bacterial phyla with the predominance of the Acidobacteria phylum (49.4%) and Proteobacteria (24.6%). The less abundant bacterial phyla were Chloroflexi (2.5%), Nitrospirae (2.2%) and Actinobacteria (1.9%). There was no alteration in the dominant or in the less represented phyla with the time and season. The total DNA was also sequenced on the MiSeq and Ion Proton platforms yielding 3.6 Gbp, 2 Gbp and 4 Gbp for samples MAF1, MAF2 and MAF3, respectively. The functional analysis on the MG-RAST server, using predicted protein sequences, based on COG groups showed that 41% of the reads were related to general metabolism followed by cellular process and signalization (22%). Based on the subsystems of MG-RAST, 11.43% of the reads were related to metabolism of carbohydrates and 8.55% to amino acids and derivatives. The reads obtained from total DNA sequencing were also submitted, separated by each platform and using a hybrid strategy, to the *de novo* assembling process through MegaHIT and CLC genomic workbench packages. The CLC assembler and MiSeq platform obtained the best results, measured by the number of contigs above 1,000 bp and the length of the larger contig: 16,426/35,630, 4,840/5,841 and 2,248/6,699 for samples MAF1, MAF2 and MAF3, respectively. All these data reflects the vast diversity of the soil, which make difficult to assemble large genomic regions without a large sequencing coverage, even using a hybrid sequencing strategy.