Bacterial characterization of soils for ecological assessment with high troughput metagenomic tools

Gruber CEM [a], Pietrucci D [b], Retico F [b],Biagini T [c], Bueno S [d] Chillemi G [d], Alessio Valentini [be], Alessandro Desideri [ef]

[a] Laboratory of Virology, National Institute for Infectious Disease "L. Spallanzani", Rome, 00185, Italy
[b] DIBAF, University of Tuscia , 01100,Viterbo, Italy
[c] Bioinformatics Unit, IRCCS Casa Sollievo della Sofferenza, Mendel Laboratory, Rome, 00198, Italy
[d] CINECA, Via dei Tizii 3, Rome 00166, Italy
[e] Molecular Digital Diagnostics (MDD), Viterbo, 01100, Italy
[f] Department of Biology, University of Rome Tor Vergata, Rome 00133, Italy
e-mail: cesare.gruber@inmi.it

Keywords: (metagenomics , microenvironment, bacteria, high-performance computing)

Metagenomic analysis of bacterial populations in soils is a new and emerging technique that can uncover the bacterial ecology of a soil sample through metagenomic analysis of its populations. The assessment of the bacterial composition of soil has several purposes: from basic ecological studies to precision agriculture, including also forensic applications [1,2]. However, the analysis is usually focused on the detection of a few number of bacterial species and, when genetic analysis are done, they are used to screen molecular fragments, rather than their specific nucleotide composition. The new available high-throughput sequencing technologies permit a comprehensive representation of the total bacterial genome composition of a given soil sample. As a consequence, metagenomic analysis of soils has emerged as a novel technique that can recognize the similarity and diversity within and between soil samples, and elucidate their origin. In order to provide the most accurate depiction of bacterial composition, it is necessary to use the most comprehensive and up-to-date genome databases. Managing a database of billion of reference sequences can be quite computationally demanding. Moreover, to avoid an excess of representation of reads that come from bacteria with the highest representation in soil samples, a deep sequencing is needed, and the limited amount of material available can also complicate the analysis. Approaches that involve extraction and amplification followed by a targeted sequencing rather than shotgun sequencing are usually more useful.

We recently developed a semiautomatic, parallelized, high-throughput metagenomic tool to improve metagenomics analysis through QIIME [3] pipeline, which has emerged as the new standard procedure to define the microbial diversity via ribosomal 16S and 18S gene sequencing. Moreover we format the most recent version of the Silva database (Silva 119) [4] by an homemade python scripts, and we test all the pipeline for being available on CINECA clusters and adjusted for metabarcoding approach. Parallelization has been performed in 4 different steps: 1) clustering and picking OTUs with USEARCH [5] algorithm; 2) assigning the taxonomy at each representative sequence with UCLUST [5]; 3) making a phylogenetic tree for all OTUs with MUSCLE [6] software; 4) evaluation of biodiversity with Shannon Index and Jackknife Method. Here, we applied our metagenomic pipeline to analyze 16S ribosomal RNA (rRNA) of bacterial populations, to assess ecological characterization of ten soil samples, in a blind trial experiment. Our analysis were performed on a total data set of around ten million pair-end reads. As the rarefaction curves go to saturation below the total amount of read data produced, a good representation of bacterial population can be assumed. Within-sample diversity analysis produced a taxonomy plot for each sample, providing a biological fingerprint of bacterial population from each soil. The samples cluster in five distinct groups (Figure 2) letting us suppose that samples were collected from five distinct environments. In conclusion, our work have demonstrated the feasibility of QIIME pipeline on CINECA clusters for ecological analysis of bacterial populations in soil samples. Further analysis will be done, with larger sample size and with different metagenomic data.

[1] S. Giampaoli et al, Forensic Sci Int 240 (2014) 41-47
[2] H.Y. Buse et al., FEMS Microbiol Ecol 88 (2014), 280-295
[3] J.G. Caporaso et al., Nat Methods (2010) 7 335-336.
[4] C. Quast et al., Nucleic Acids Res 41 (2013) D590-596
[5] R.C. Edgar, Bioinformatics 26 (2010), 2460-2461
[6] R.C. Edgar, Nucleic Acids Res 32 (2004) 1792-1797

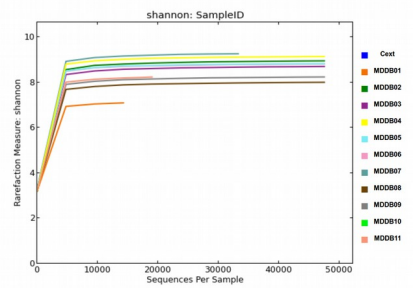

Figure 1. Rarefaction plots for each sample. The species richness is assessed with Shannon index. The most amount of species have been sampled for all samples since all curves reaches a plateau.

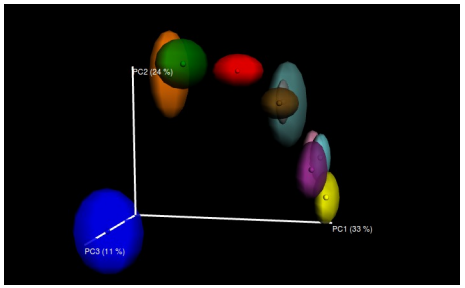

Figure 2. Principal Coordinate Analysis (PCoA) estimated with a Jackknife bootstrap method to evaluate taxomic composition diversity between samples. Samples clearly clusters into five different groups.