Deep sequencing of 16S rRNA gene to efficiently assess bacterial richness and the rare biosphere in soils

Sébastien Terrat, Samuel Dequiedt, Dipankar Bachar, Richard Christen, Christophe Mougel, Mélanie M. Lelievre, Pierre-Alain Maron, Pierre Plassart, Patrick Wincker, C. Cruaud, et al.

Authors and affiliation
S Terrat, S Dequiedt, D Bachar, R Christen, C Mougel, M Lelievre, PA Maron, P Plassart, P Wincker, C Cruaud, C Jolivet, D Arrouays, A Bispo, P Lemanceau, and L Ranjard

TITLE

Deep sequencing of 16S rRNA gene to efficiently assess bacterial richness and the rare biosphere in soils.

Text (500 words)

Soil is one of the most important reservoirs of microbiological diversity on our planet and, above all, one of the last bastions of such biodiversity. Since two decades, numerous molecular tools have been developed to assess accurately this huge diversity directly from soil DNA. In this context, 16S rRNA gene is widely used to study microbial community taxonomic diversity, as it can be easily amplified by PCR, and large databases are available relating obtained sequences to bacterial phylogenies. Moreover, the recent development of high-throughput sequencing technology allows the scientific community to assess 'MetaTaxogenomic' studies by getting hundreds of thousands of ribosomal rRNA gene sequences from a single metagenomic DNA. The impressive power of these emerging tools calls now for a more thorough examination of microbial diversity in terms of richness (efficient detection of major and minor populations) and evenness. To date, only a small number of the estimates have been based on a sufficient number of sequences to provide clear information.

Here, we used pyrosequencing to obtain between 100,000 and more than 600,000 raw sequences for each of the seven different soils. Studied soils originate from the soil library of the French Soil Quality Monitoring Network (RMQS), which contains more than 2,200 soils sampled across France and were chosen because of their contrasting pedoclimatic and land-use characteristics, known to strongly influence microbial abundance and diversity. To define the number of reads needed to obtain a good taxonomical inventory, the number of sequences are randomly re-sampled in subsets between 10,000 and 600,000 for each soil.

Whatever the number of sequences used to analyze microbial diversity and composition, the discrimination between soils is similar at the *phylum* level. Meanwhile, for soils with close physico-chemical characteristics, this differentiation is a bit more difficult with a low number of sequences (due to the high variability of groups with abundance lower than 1% of the detected community). Also, irrespective of the soil tested, results show that a high number of high-quality sequences are needed to obtain a good snapshot of the total community composition and richness at lower levels (e.g. *genus* level). Taxonomy-independent analyses highlight that the number of detected operational taxonomic units (or OTUs) is increasing with the number of sequences, without ever reaching a clear plateau. Many of these novel OTUs, composed of a low number of reads correspond to low-abundance organisms. These rare organisms can be considered as members of the 'rare biosphere,' when closely related to a known sequence, but since they are usually sorted into "unclassified" or "other" categories, assigning them to novel lineages requires careful examination.

If the aim of the project is to compare different environments, 10,000 high-quality sequences are sufficient. But, to compare efficiently different soils in terms of microbial diversity and composition, a higher number of sequences are necessary to obtain a good snapshot of microbial community composition and representativeness. In fact, with few sequences, rare OTUs may be detected or not only by chance due to random sampling.