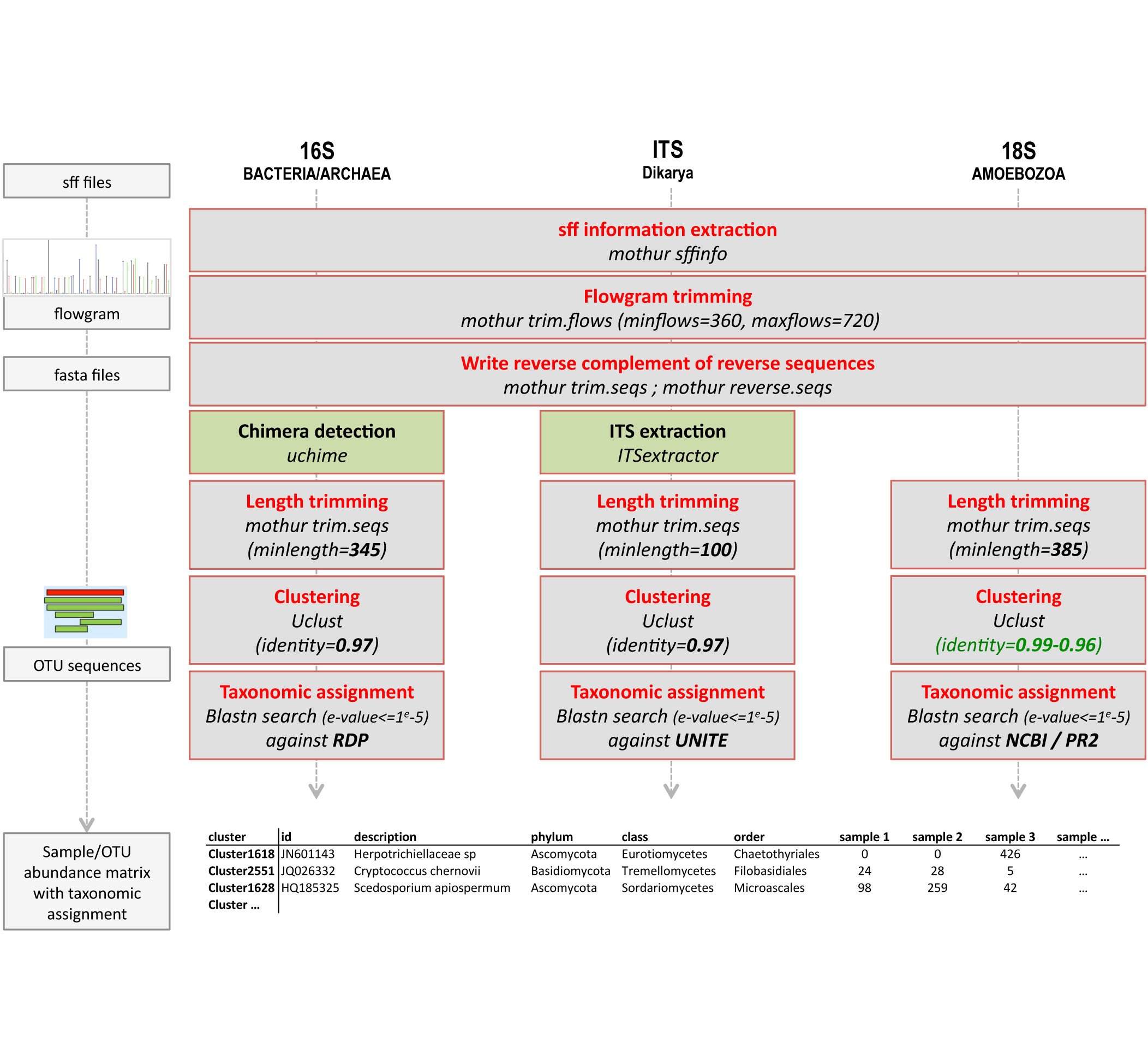
## Developing and Standardising Tools and Procedures for Assessment of Soil Biodiversity

Emilie Tisserand, Pierre Plassart, Marc Buée, Juliette Lengelle, Thibaud Payen, Claude Murat, Stéphane Uroz, Philippe Lemanceau, and Francis Martin

Soil is the reservoir of a high diversity of organisms. This biodiversity plays an important role in ecosystem functioning and nutrient cycling due to the ability of these organisms to degrade organic matter, weather minerals, and fix carbon and nitrogen. However, we are far from appreciating this telluric black box of diversity. Most of our present knowledge was established using cultivation-dependent approaches giving access to only a limited portion of this diversity. The development in the last decade of high-density 16S rRNA, 18S rRNA or ITS microarrays, monogenic and shotgun metagenomics has permitted us to greatly improve our knowledge, generating very informative views on the structure and diversity of the soil microbiome.

In the frame of the EcoFINDERS project, the impact of the type and intensity of land management was tested across Europe in different long-term observatories using amplicon-based metagenomics. The objectives were to decipher the diversity and structure of a wide range of organisms including archaea, bacteria, fungi (Glomeromycota and Dikarya), protists, and nematodes. This approach was a scientific challenge because our present knowledge of the diversity and associated molecular markers of the range of organisms are not at the same level. Here, we present the tools to investigate soil biodiversity and we discuss the major results highlighting the diversity of archaea, bacteria, fungi, amoebozoa, flagellates and nematodes in European soils.

For most of the organisms considered ribosomal genes have been used as marker genes: i) 16S rRNA for archea and bacteria and 18S rRNA for the others. The internal transcribed spacer region (region ITS1) was used to decipher the fungal diversity, focusing on Dikarya sub-kingdom (Basidiomycota and Ascomycota). The same metagenomic DNA samples were used to assess to the diversity of all the soil organisms. After amplification, the sequencing of these marker genes generated a data set containing from 245,867 sequences for the amobozoa to 3 millions of sequences for the bacteria. The next step was to analyse this large diversity of sequences. To date, several bioinformatic pipelines have been developed such as RDP (ribosomal data project, Cole et al., 2009), Mothur (Schloss et al., 2009), QIIME (Caporaso et al., 2010) and UPARSE (Edgar, 2013), and most of them have been optimized to analyse the bacterial diversity. As one important task of the project was to analyse the sequences generated for each type of organism in a similar way, different pipelines have been tested and one composite pipeline - named pipeline #1 - described in Figure 1, was at last established to analyse all the sequences



**Figure 1: Description of Pipeline #1.** This figure presents the different steps from the sequences extraction from the raw data to the taxonomic assignation using the convenient database.

Bioinformatic analysis of the sequences was performed and several critical points were highlighted:

1) the raw data generated need a flowgram trimming step,

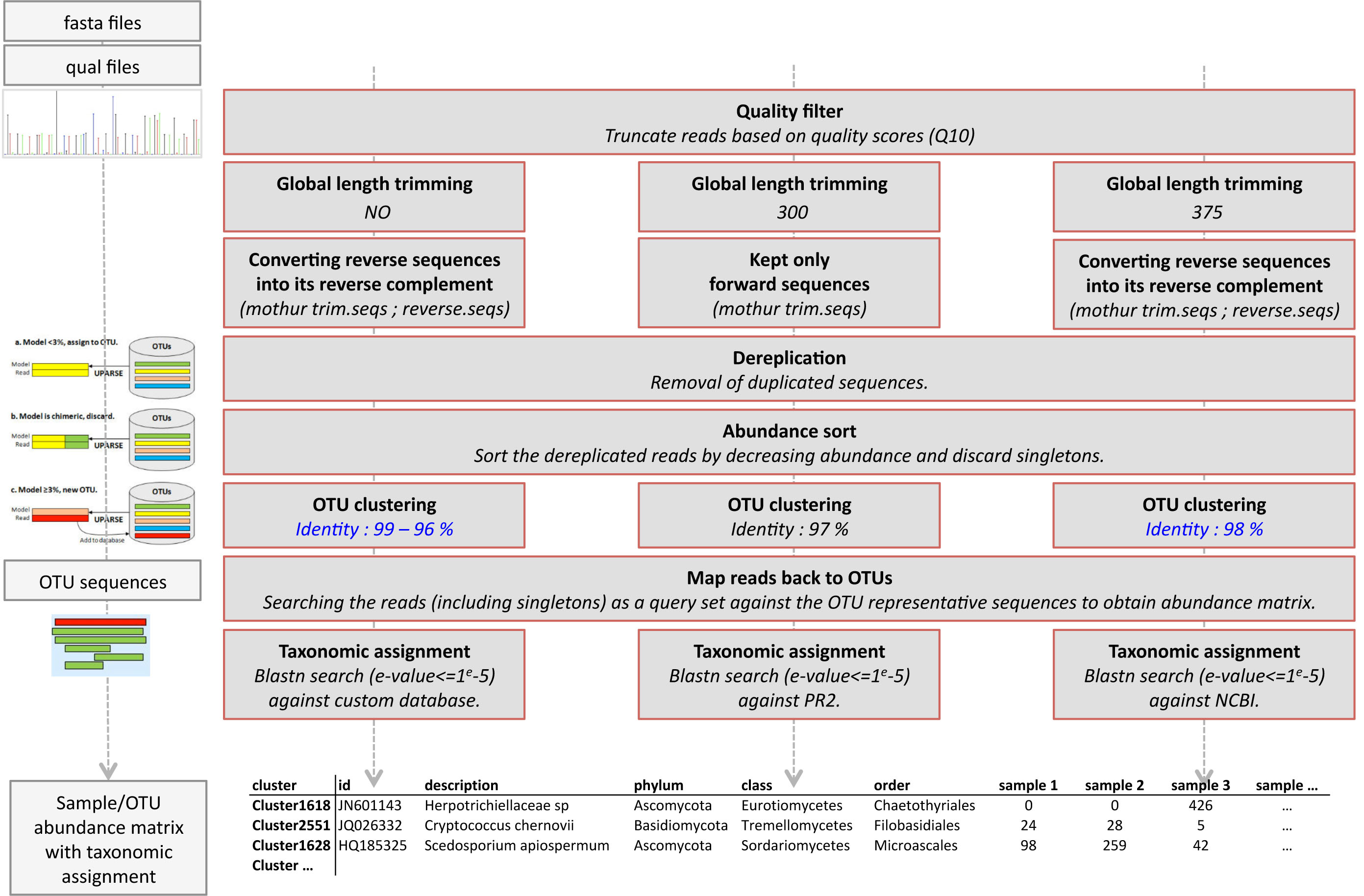
2) ITS analysis needs a specific ITS extraction step, which is not necessary for the ribosomal genes,

3) the trimmed sequences need to be adjusted in a length adapted to the marker gene considered,

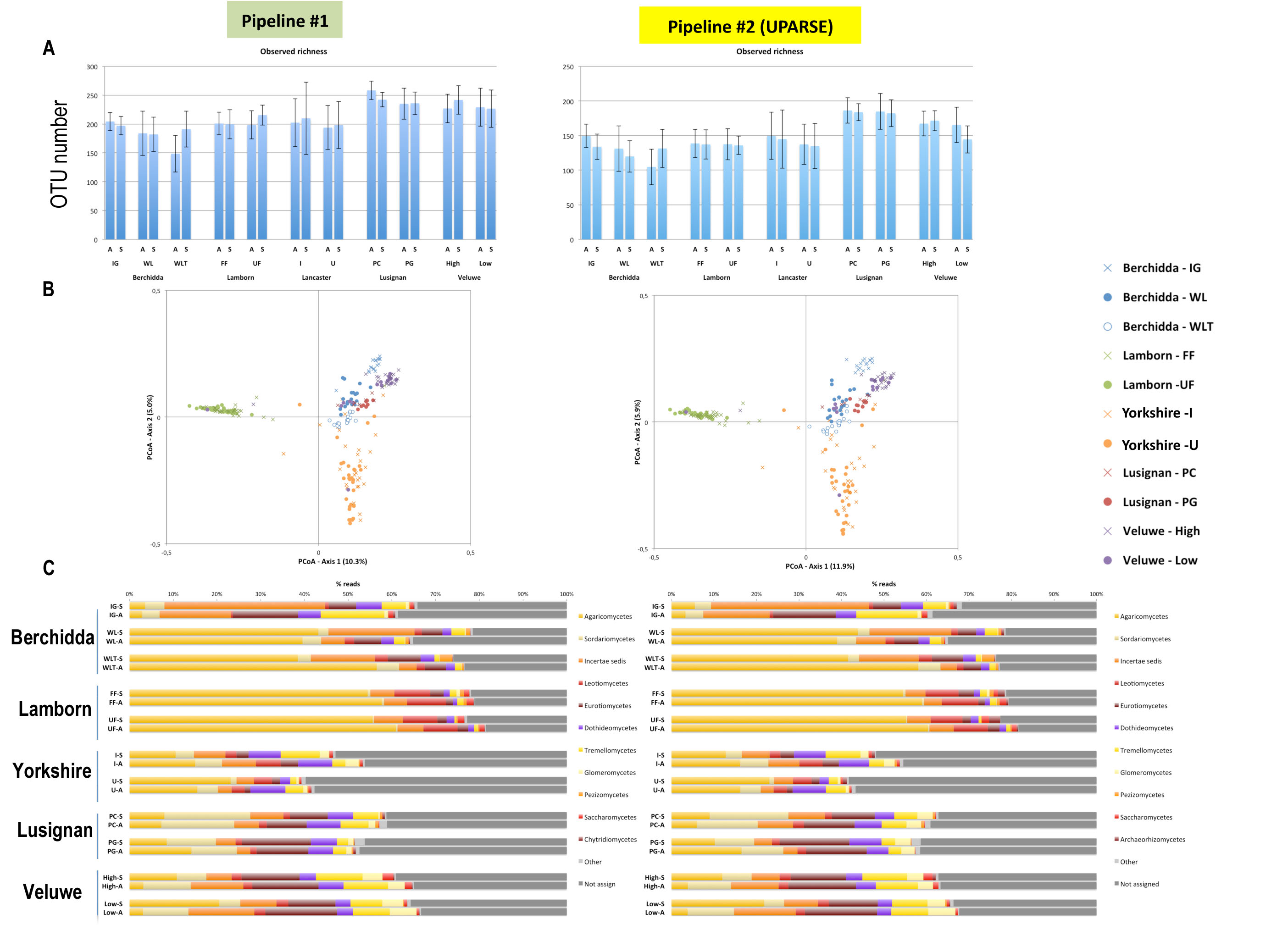
4) the clustering step must be adapted in relation with the organism considered and the inter-specific polymorphism of targeted barcodes,

5) Convenient databases (which is not yet available for all organisms) need to be used for a better taxonomic affiliation.

Due to the continual development of bioinformatics software and programs, a comparative analysis was performed using the pipeline #1 and a adaptation of the recent pipeline ‘UPARSE’ developed by Edgar (2013), on the same data set randomly subsampled to conserve only 500 sequences per sample. The critical steps of this second analytical pipeline are presented in Figure 2. A detailed comparison of the fungal ITS sequences revealed that pipeline #1 gave a higher richness (based on the number of OTU (operational taxonomic unit)) than pipeline #22 based on UPARSE (Figure 3A), including a larger proportion of singletons (single read for one OTUs). However, the same trends of richness were observed for the two pipelines. Notably, taxonomic affiliation gave similar results. The multivariate analyses (Figure 3B) performed using the two pipelines gave very similar patterns. Finally, the distribution of the sequences (relative abundance) was very similar between the two pipelines (Figure 3C).



**Figure 2: Description of pipeline #2 based on UPARSE.**



**Figure 3: Analysis of the fungal communities occurring across the different LTOs using pipeline #1 (left part) and pipeline #2 (UPARSE) (right part).** For this analysis, the dataset was subsampled to conserve 500 sequences for each sample. A) Richness analysis based on the number of OTUs. A: autumn sampling; S: spring sampling; I: improved management; U: non-improved management. B) Multivariate analysis performed on the relative distribution of the OTUs in taxonomic groups. Geographical location and management type are presented in the legend. C) Relative distribution of the sequences at the class level. Berchidda (IG: improved grassland, WL: wooded land; WLT: forest land), Lamborn (FF: fertilised forest; UF: non fertilised forest), Yorkshire (I: improved grassland; U: non-improved grassland), Lusignan (PC: permanent culture; PG: permanent grassland), Veluwe (High: long term abandoned grassland; Low: short term abandoned grassland).

Altogether, these analyses revealed that it is possible to analyse different marker genes (16S rRNA, 18S rRNA and ITS) with the same bioinformatics pipeline. However, adaptations are required to fit with i) the type of molecular marker, ii) the size and size homogeneity of marker genes and iii) the threshold applied to cluster the barcode sequences in OTUs. This approach permitted us to generate the first comprehensive view of the diversity and structure of the archaea, bacteria, fungi, amoebozoa, flagellates and nematodes in the same soil samples obtained at the different long-term observatories across Europe. The comparative analysis performed using two different pipelines provides a significant contribution in the understanding of the limitations and robustness of the bioinformatics methods. Notably, such comparison showed that richness data need to be interpreted with caution as they are strongly dependent on the pipeline used. In contrast, similar patterns of distribution of the main taxa (community structure) were obtained with the pipeline #1 and #2.

We still have some way to travel on our journey to fully appreciate the diversity and structure of the soil biota. Multidisciplinary studies linking metagenomic survey and characterization of soil organisms are needed to fully achieve this.

**References**

Caporaso et al., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 7(5):335-336.

Cole et al., 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37 (Database issue), D141eD145.

Edgar, R.C. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads, Nature Methods

Schloss et al. 2009. Introducing mothur: open-source, platformindependent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75:7537–7541.

*Further information Contact: uroz@nancy.inra.fr*