Comparative analyses of 454 pyrosequencing fungal ITS sequences processing methods

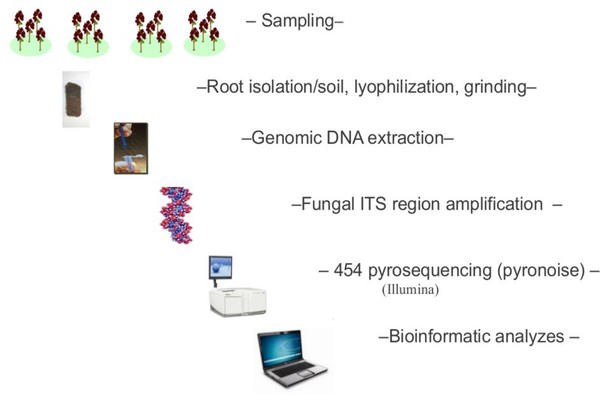
Juliette Lengellé(1), Marc BUEE(1), Claude Murat(1), Emmanuelle Morin(1), Francis MARTIN(1)

(1) INRA de Nancy, UMR INRA/UHP 1136, Interactions Arbres/Microorganismes, 54280 Champenoux, France

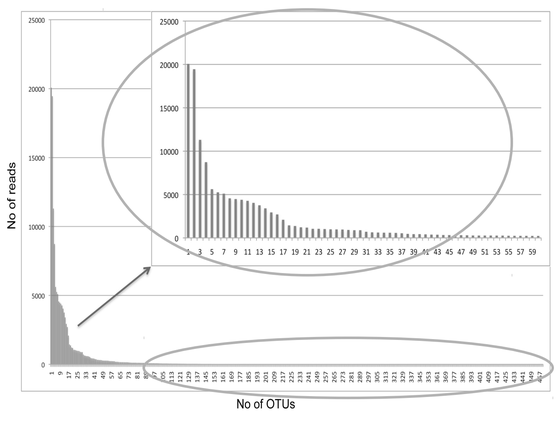
Introduction

The ecosystem management requests knowledge of biodiversity, spatial and temporal dynamics of diverse communities, their roles and their potential interactions in the environment. The new sequencing technologies (e.g. 454 pyrosequencing), allow now to obtain a detailed taxonomic composition of microbial communities. However, inherent sequencing errors of these technologies lead to artefactual OTUs (Operation Taxon Units), in particular singletons which represent around 50% of OTUs. These errors induce bias in OTU's identification (and OTU's clustering) and can induce an overestimation of the real diversity. That's why, it's very important to identify and eliminate sequencing errors before taxonomic assignation step. Different programs, which solve these problems, are available to study bacterial biodiversity using gene coding ribosomal DNA 16S. However, rare programs are currently available to study fungal biodiversity using ribosomal intergenic spacer sequence (ITS). The aim of this study was to associate different bioinformatic programs used to clean 454 pyrosequencing sequences to reduce artefactual OTU's redundancy without modifying the supposed fungal diversity of the studied environment.

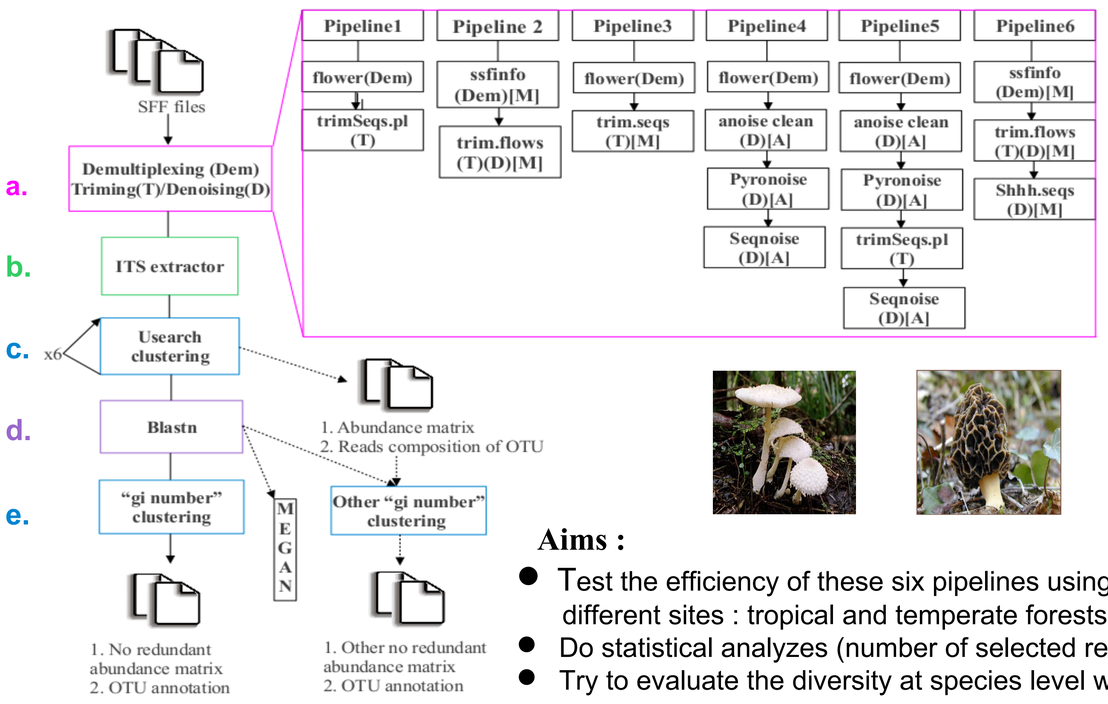
Methods for environmental studies



Usual OTU distribution example



Methods



Rk : - *Mothur* (Schloss, 2009) and *AmpliconNoise* (Quince, 2011) programs dedicated for DNAr 16S analyzes
- Pipeline 1 used as an indicator of diversity overestimation (Dickie, 2010)
- Pipeline 6 (*Mothur)* equivalent of Pipeline 4 (*AmpliconNoise*)
- Add two steps of clustering in all the pipelines to limit the redundancy effect

a. Association of different sequence processing programs (six different pipelines) :
- sff file demultiplexing programs (Dem): *sffinfo , flower*
- trimming (T)/ denoising (D) programs : *Mothur [M], AmpliconNoise [A], trimSeq.pl*

b. Extraction of ITS1 region with *Fungal ITSextractor* program (Nilson *et al.*, 2010)
Selection of ITS1 regions according to their length (≥ 100 bp)

c. Six successive clusterings of ITS1 regions with more than 97% of similarity with Usearch (iddef = 2)
Creation of abundance matrix (number of sequences by cluster by samples)

d. OTU Taxonomic assignation by similarity using *BlastN* program (e-value cut off $1e^{-05}$) against NCBI nt database curated

e. OTU clustering by gi number

Aims :

- Test the efficiency of these six pipelines using three different data sets of around 100,000 reads (ITS sequences), obtained from environmental DNA of different sites : tropical and temperate forests and truffle orchards inoculated with *Tuber melanosporum* (most abundant species and internal standard).
- Do statistical analyzes (number of selected reads, number of OTU, number of singletons, mean number of reads by OTU, mean length of OTU).
- Try to evaluate the diversity at species level with *MEGAN*.

Results

Pipeline's results comparison for truffle orchards data set

A

	Pipeline1 (1T)	Pipeline 2 (1D + 1 T)	Pipeline 3 (1 T)	Pipeline 4 (3 D)	Pipeline 5 (3 D + 1 T)	Pipeline 6 (2D + 1 T)
Percentage of selected reads	42.9%	35.2%	21.7%	40.2%	38.7%	7.0%
Number of OTUs	4954	4005	2352	5434	4770	2268
Percentage of singlet OTU	49.0%	47.4%	46.7%	64.4%	61.8%	51.8%
Mean number of reads by OTU	18	18	19	3	3	6
Mean length of OTU (bp)	184	187	154	170	179	203
Number of species	1054	958	668	868	834	514

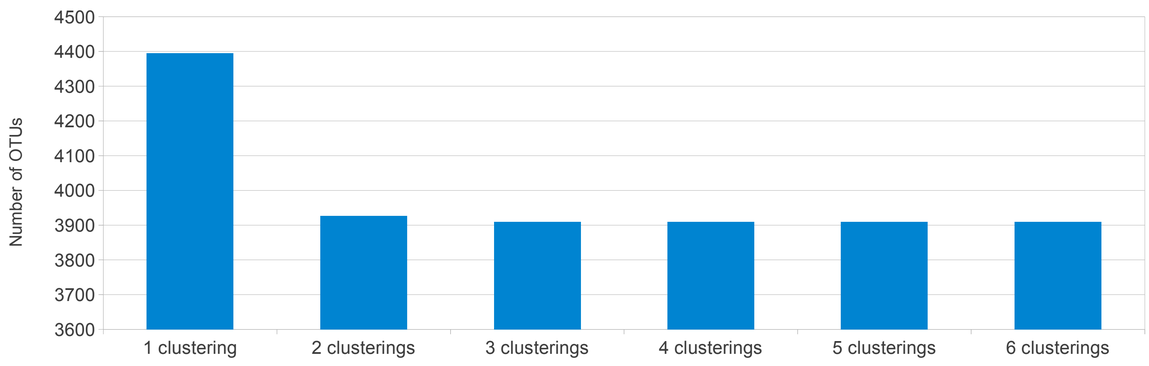
- Results for pipelines with several steps of denoising (Pipelines 4, 5 and 6) :
  => Important reduction of the reads number supporting OTUs (A and B): relative abundance
  => Important singleton number > 50% (A)
  => *Tuber melanosporum* (control) ranked in the 12th place for pipeline 4 and pipeline 5
- Results for the three first pipelines:
  => *Tuber melanosporum* ranked in the first place supported by more 10, 000 reads (B)
  => Reduction of singleton number < 50% (A)
  => For pipeline 3 only, shorter consensus sequences, small number of selected reads

Top twelve of the most abundant species for the six pipelines

B

Pipeline 1	Pipeline 2	Pipeline 3	Pipeline 4	Pipeline 5	Pipeline 6
Tuber melanosporum (19, 377 reads)	*Tuber melanosporum* (16, 825 reads)	*Tuber melanosporum* (13, 699 reads)	*Mortierella alpina* (109 reads)	*Mortierella alpina* (105 reads)	*Tuber melanosporum* (2, 107 reads)
Cryptococcus aerius (3, 757 reads)	*Cryptococcus aerius* (3, 092 reads)	*Cryptococcus aerius* (2, 849 reads)	*Mortierella elongata* (92 reads)	*Ascomycota sp.* (98 reads)	*Cryptococcus aerius* (686 reads)
Fusarium dlaminii (3, 095 reads)	*Fusarium napiforme* (2, 461 reads)	*Fusarium dlaminii* (1, 622 reads)	*Ascomycota sp.* (85 reads)	*Mortierella elongata* (81 reads)	*Mortierella alpina* (309 reads)
Mortierella sp. (2, 389 reads)	*Chaetomium sp.* (1, 146 reads)	*Mortierella sp.* (1, 107 reads)	*Cladosporium oxysporium* (65 reads)	*Mortierella sp.* (69 reads)	*Mortierella alpina* (214 reads)
Scleroderma aerolatum (1, 308 reads)	*Mortierella sp.* (1, 107 reads)	*Ascomycota sp.* (741 reads)	*Mortierella alpina* (64 reads)	*Mortierella elongata* (64 reads)	*Mortierella alpina* (143 reads)
Chaetomium sp. (1, 177 reads)	*Scleroderma aerolatum* (1, 105 reads)	*Mortierella alpina* (552 reads)	*Cryptococcus aerius* (57 reads)	*Mortierella alpina* (62 reads)	*Penicillium sp.* (118 reads)
Mortierella sp. (1, 076 reads)	*Mortierella alpina* (1, 020 reads)	*Phoma eupyrena* (550 reads)	*Phoma sp.* (55 reads)	*Phoma sp.* (58 reads)	*Phoma sp.* (118 reads)
Mortierella alpina (1, 043 reads)	*Mortierella sp.* (787 reads)	*Inocybe oblectabilis* (380 reads)	*Mortierella sp.* (54 reads)	*Fusarium dlaminii* (57 reads)	*Scleroderma aerolatum* (117 reads)
Mortierella alpina (905 reads)	*Scytalidium lignicola* (625 reads)	*Hymenogaster citrinus* (338 reads)	*Didymella pisi* (53 reads)	*Cryptococcus aerius* (57 reads)	*Ascomycota sp.* (100 reads)
Scytalidium lignicola (784 reads)	*Hymenogaster citrinus* (557 reads)	*Fusarium equiseti* (309 reads)	*Fusarium dlaminii* (49 reads)	*Ascomycete sp.* (54 reads)	*Alternaria alternata* (95 reads)
Mortierella alpina (693 reads)	*Inocybe oblectabilis* (550 reads)	*Cordyceps bassina* (293 reads)	*Ascomycete sp.* (49 reads)	*Cladosporium oxysporium* (53 reads)	*Hymenogaster citrinus* (85 reads)
Alternaria alternata (640 reads)	*Mortierella alpina* (544 reads)	*Mortierella sp.* (287 reads)	*Tuber melanosporum* (47 reads)	*Tuber melanosporum* (51 reads)	*Mortierella alpina* (85 reads)

Impact of the six successive clusterings on *Usearch* program results



	Number of gi number found for *Tuber melanosporum*	Total number of OTUs
1 clustering	3	13
2 clusterings	3	7
3 clusterings	2	6
4 clusterings	2	6
5 clusterings	2	6
6 clusterings	2	6

=> two clustering steps essential to reduce redundancy effect of 50%
=> two clustering steps essential to have 100% of unique consensus sequences

Rq : *Tuber melanosporum* represented by 2 gi number and 6 OTUs, after 6 clustering steps
=> real nucleotide variability or bad reads cleaning ?

Conclusions

Increasing number of denoising steps impacts richness values and the possibility to use read numbers as a relative abundance proxy. These steps can modify also the real order of most abundant species (e.g. *Tuber melanosporum*) and increase the number of singletons, potentially artificial. However, only one trimming step is not efficient. Indeed, whether the trimming program isn't enough stringent and the diversity is overestimate or it's too stringent and the number of selected reads is reduced and consensus sequence lengths are too short . Finally, only one step of denoising, before trimming, is recommended.

=> Pipeline 2, which have an unique step of denoising and trimming done by the same program *trim.flows* (*Mothur*) seems the best option in our studies

=> several steps of clustering improves the quality of analysis with *Usearch,* reducing redundancy effect with 100% of unique consensus sequences

References

Dickie IA. 2010. Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytologist***188**: 916–918
Nilsson RH, Bok G, Ryberg M, Kristiansson E, Hallenberg N. 2009a. A software pipeline for processing and identification of fungal ITS sequences. *Source Code for Biology and Medicine***4**: 1
Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. 2011. Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics***12**: 20–38
Schloss P, Westcott SL, Ryabin T, Hall, JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and environmental microbiology***75**: 7537–7541

Acknowledgements

This work was supported by European ECOFINDERS program, Region of Lorraine and the FEDER