Its all fun guys: a comparison of bioinformatic pipelines for metabarcoding plant and soil fungal communities
Charlie Pauvert, Marc Buee, Valerie Laval, Véronique Edel-Hermann, Laure Fauchery, Angelique Gautier, Isabelle Lesur, Jessica Vallance, Corinne Vacher

To cite this version:

HAL Id: hal-01938707
https://hal.science/hal-01938707
Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
**ITS ALL FUN GUYS**

**A COMPARISON OF BIOINFORMATIC PIPELINES FOR METABARCODING PLANT AND SOIL FUNGAL COMMUNITIES**

**Charlie Pauvret**, **Marc Buée**, **Valérie Lalav**, **Véronique Edel-Hermann**, **Laure Fauchery**, **Angélique Gautier**, **Isabelle Lesur**, **Jessica Vallance**, **Corinne Vacher**

---

**Introduction**

Plant and soil fungal communities influence plant fitness and ecosystem functioning.

These fungi are classically studied by metabarcoding approaches targeting the ribosomal internal transcribed spacer (ITS).

**Methods**

DNA were pooled and sequenced with Illumina MiSeq.

DNA were independently amplified and sequenced with Sanger.

**Figure 1** Taxonomic composition of the artificial fungal community. The nodes of the tree are the ITS1 sequences (n = 175) of the fungal strains (n = 189) that constitute the artificial or mock community.

**Figure 2** Overview of the bioinformatic pipelines compared in this study.

- **Pipeline comparison**
  - **Sensitivity**: Percentage of recovered mock strains
  - **Precision**: Percentage of OTU/ASV matching mock strains
  - **Compositional similarity**: Bray-Curtis similarity between the community recovered and the mock community

**Results**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Highest Sensitivity</th>
<th>Highest Precision</th>
<th>Highest Similarity</th>
<th>Best Trade-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly</td>
<td>QUALITY_R1</td>
<td>FASTQJOIN_150</td>
<td>QUALITY_R1</td>
<td>No</td>
</tr>
<tr>
<td>Extraction</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>VSEARCH</td>
<td>VSEARCH</td>
<td>DADA2</td>
<td>DADA2</td>
<td>DADA2</td>
</tr>
<tr>
<td>Chimera</td>
<td>Retained</td>
<td>Removed</td>
<td>Retained</td>
<td>Removed</td>
</tr>
<tr>
<td>Filtering</td>
<td>All</td>
<td>10</td>
<td>All</td>
<td>All</td>
</tr>
</tbody>
</table>

Table 1 Pipelines with the best performances for each three criteria and the pipeline with the best trade-off. Pipelines are labeled below the table and the steps are numbered as in Fig 2.

**Figure 3** Total number of OTUs (or ASVs) retrieved by the top four pipelines (Table 1). The black horizontal line indicates the expected richness.

**Figure 4** Values of precision and compositional similarity to the mock fungal community, for all 288 bioinformatic pipelines. Only the top four pipelines are labeled.

**Conclusion**

The Si5 pipeline used single forward (R1) sequences with DADA2 and no filter other than the removal of low-quality and chimeric sequences.

**Acknowledgments**

Attendance at this congress was possible thanks to the award for the best presentation of the Pathobiome 2018 congress. We thank Matthieu Barret, Martial Briand, Lucas Auer, Gregory Gambetta, Guilherme Martins, Frédéric Barraquand and Tania Fort for useful comments and discussions. We also thank the INRA MEM metaprogramme (Meta-Omics of Microbial Ecosystems) for financial and scientific support. The mock community sequencing was funded by the INRA MEM MetaBAR project (PI: MB) and the bioinformatic analyses were performed as part of the INRA MEM Learn-biocontrol project (PI: CV). Additional funding was received from the LABEX COTE (ANR-10-LABX-45) and the LABEX CEBA (ANR-10-LABX-25-01). CP’s PhD grant was funded by the INRA and Bordeaux Sciences Agro (BSA). We thank the Genotoul sequencing facility (Get-PlaGe) for sequencing the mock community and the Genotoul bioinformatics facility (Bioinfo Genotoul) for providing computing and storage resources.

---

**Objectives**

1. Make and sequence a mock community of 189 Dikarya strains commonly found in agricultural and forest soils and in plant tissues.

2. Compare the ability of 288 combinations of bioinformatic softwares and parameters to recover the fungal strains present in this mock community, in the expected proportions.

---

1. BIOGECO, INRA, Univ. Bordeaux, 33615 Pessac, France
2. INRA, Université de Lorraine, UMR IAM 1136, Laboratoire d’Excellence ARBBRE, Centre INRA-Lorraine, 54280
3. Agroécologie, AgroSup Dijon, CNRS, INRA, Univ. Bourgogne Franche-Comté, 21000 Dijon, France
4. école Polytechnique, PSL Paris-Saclay, France
5. HelaVenture, 33700 Mérignac, France
6. SAVE, Bordeaux Sciences Agro, INRA, ISYV, Univ. Bordeaux, 33882 Villenave d’Ornon, France