

Harmonizing bioinformatics procedures in microbiome amplicon high-throughput sequencing

Dubois Benjamin, Gelhay Vanessa, Muhovski Yordan and Deboide Frédéric
Department of Life Sciences, Bioengineering Unit, Walloon Agricultural Research Centre (CRA-W), Gembloux, Belgium

Introduction

The recent fast development of high-throughput sequencing (HTS) technologies and associated procedures allowed, among others, to study the microbiome composition of complex samples from very contrasting environments using amplicon HTS. Bioinformatics workflows have been developed to deal with the huge amount of generated data, and meta-analysis tools have been created to analyze microbiome data in the context of tens of thousands of other samples from other studies. However, sequencing data must be generated and processed in the same way so that the results from different studies can be appropriately compared.

Hence, the major steps of a traditional bioinformatics workflow dedicated to microbiome amplicon HTS are presented here, together with the main available options, pros/cons and recommendations for each of them.

Methods

Scientific literature has been reviewed to identify, for each bioinformatics step, the main options available and the trends in the selected options. Attention has been focused on benchmarking studies, to highlight the most relevant procedures and the context in which their use is recommended. These best practices also emerge from results and observations gathered from our different microbiome research projects.

Results

1. Selection of targets. When studying microbiome composition, the most commonly targeted regions are the 16S rRNA gene (bacteria and archaea), the internal transcribed spacer (ITS), and the 18S and 28S rRNA genes (fungi). The selected portion of these regions may have a strong impact on taxonomic assessments, just like the PCR primers used (De Filippis 2017; Soriano-Lerma 2020; Abellan-Schneyder 2021; Fadeev 2021). Using several targets in parallel and evaluating the relevance of candidate targets with mock samples is highly advised. Combining the high accuracy of short-read sequencing to the improved taxonomic resolution of long-read sequencing is also promising.

2. Read trimming. Low-quality bases at the edges of the reads must be trimmed to avoid failure when joining forward and reverse reads. Several studies have shown the positive influence of trimming on the result quality (Mohsen 2019; Abellan-Schneyder 2021). It is thus recommended to include a read-trimming step in the bioinformatics workflow and to define it on a case-by-case basis.

3. Read processing strategy. To avoid redundancy and to correct errors, sequencing reads used to be clustered into operational taxonomic units (OTUs) where sequences displaying a high degree of similarity (often 97%) were grouped together. New methods have recently been developed to resolve amplicon sequence variants (ASVs) using error models to remove amplification and sequencing errors. ASVs allow differentiation between sequence variants differing by only one nucleotide, thus improving the taxonomic resolution, and they are reusable and comparable across studies. There is now a strong scientific consensus to move towards ASVs (Callahan 2017; Almeida 2018; Joos 2020; Prodan 2020; Straub 2020).

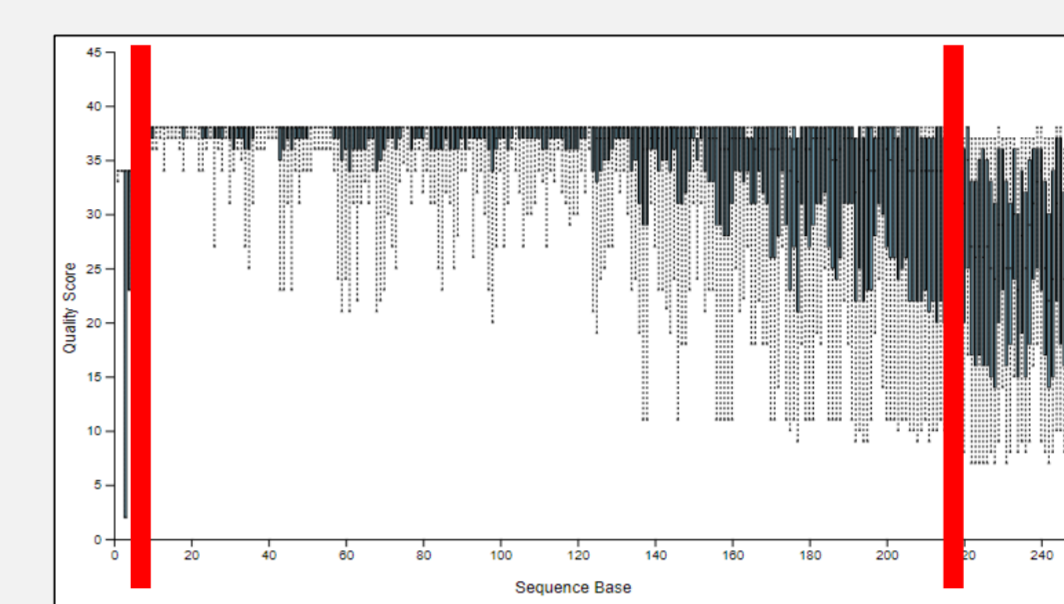
4. Reference database. Different reference databases are available to assess the composition of bacterial (SILVA, Greengenes, RDP, GTDB, NCBI RefSeq) and fungal (UNITE, Warcup, SILVA, RDP, NCBI RefSeq) communities. It is crucial to consider the database characteristics when selecting one to work with, as it can have a strong impact on the conclusions drawn (Xue 2019; Abellan-Schneyder 2021; Robeson 2021; Ramakodi 2022). For example, NCBI-RefSeq shows very good accuracy but limited taxonomic coverage, whereas GTDB should be preferred to analyze long-read sequencing data as it shows longer reference sequences. Reference databases can also undergo further curation/processing that might have an impact on the analysis outcomes (Dubois 2022).

5. Taxonomic classifier. The classifier is the algorithm used to compare sequencing reads to reference sequences and assign taxonomy. A recent study benchmarked the main classifiers according to the strategy they use and showed that the sequence-composition method implemented in the QIIME2 platform and the widely used BLAST alignment algorithm offered the best prediction accuracies (Hleap 2021). In addition, appropriate classifier parametrization is crucial, ideally using mock samples. The choice of the classifier is strongly linked to the reference database characteristics and some combinations should be preferred or avoided.

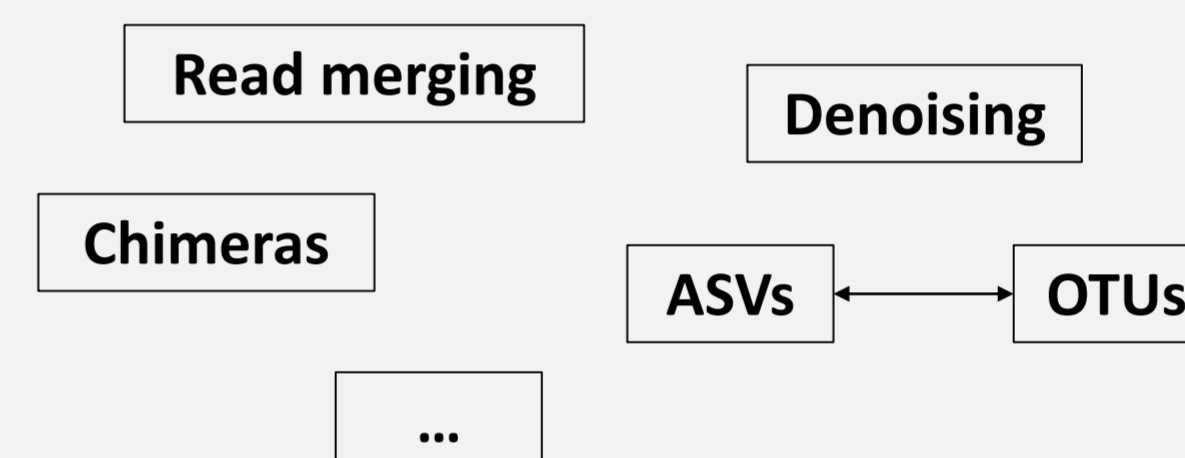
1. Selection of target(s)



2. Read trimming



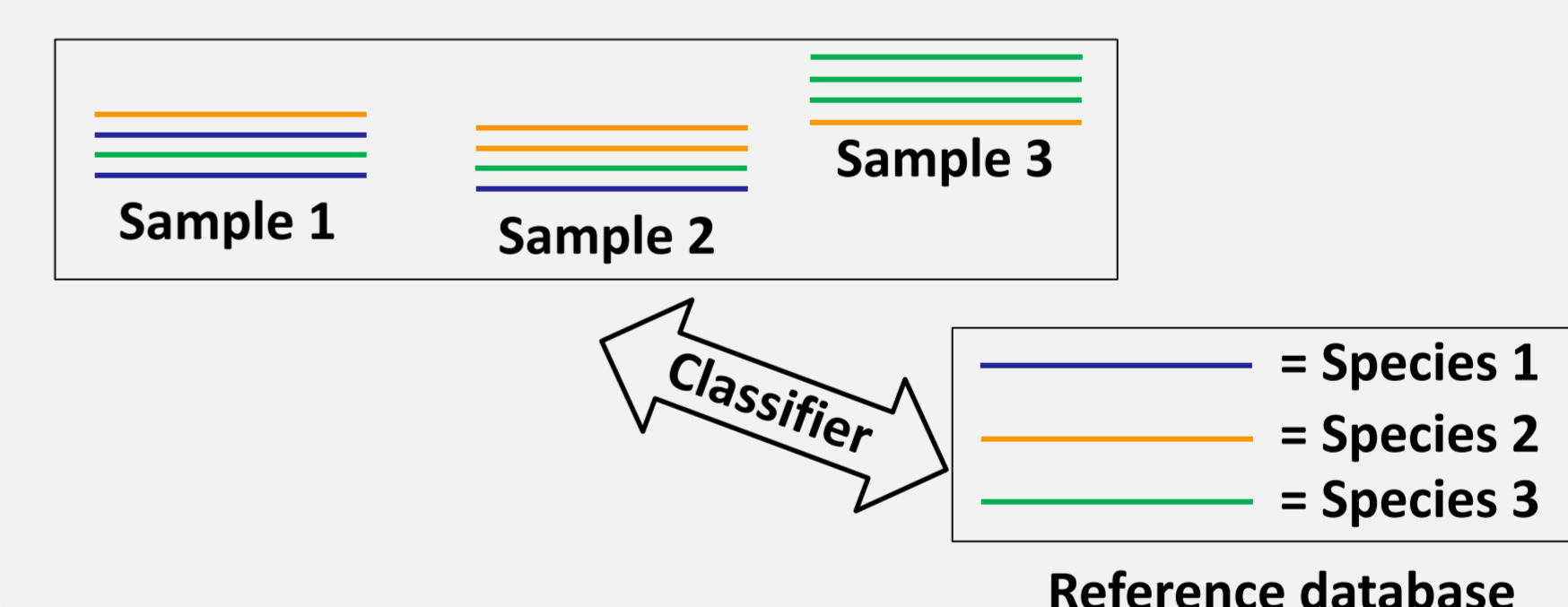
3. Read processing strategy



4. Reference database



5. Taxonomic classifier



BIOINFORMATICS WORKFLOW

Conclusion

The main steps constituting a traditional bioinformatics workflow have been identified, together with the major trends and recommendations. Two main lessons can be learned from them: (i) understand the features of the tools used and evaluate their suitability according to the study specifications (e.g. with reference material). (ii) Adopt practices favoring data exportability and re-usability. All together, these harmonizing efforts will help increase the range of microbiome studies by allowing the comparison of sequencing data from different works in a relevant way, based on results produced with identical procedures.

Literature cited

- Abellan-Schneyder 2021, MSphere.
- Almeida 2018, GigaScience.
- Callahan 2017, ISME J.
- De Filippis 2017, Appl Environ Microbiol.
- Dubois 2022, BMC Genomic Data.
- Fadeev 2021, Front Microbiol.
- Hleap 2021, Mol Ecol Resour.
- Joos 2020, BMC Genomics.
- Mohsen 2019, BMC Bioinformatics.
- Prodan 2020, PLoS ONE.
- Ramakodi 2022, Biotechnol Lett.
- Robeson 2021, PLoS Comput Biol.
- Soriano-Lerma 2020, Sci Rep.
- Straub 2020, Front Microbiol.
- Xue 2019, Biol Fert Soils.