ARPHA Conference Abstracts 4: e65027
doi: 10.3897/aca.4.e65027

Conference Abstract

OTU picking on large datasets : comparing methods on a diversity of situations

Jean-Marc Frigerio[‡,§], Ester Eckert[|], Emre Keskin[¶], Frédéric Mahé[#], Michael T Monaghan[¤], Matteo Montagna[«], Franck Salin[‡,§], Paul Schmidt Yáñez[»], Douglas W Yu[^,˅], Diego Fontaneto[|], Alain Franc[‡,§]

‡ INRAE BioGeCo, Cestas, France
§ INRIA Pleiade, Talence, France
| CNR-IRSA, Verbania, Italy
¶ Evolutionary Genetics Laboratory (eGL), Ankara University Agricultural Faculty, Ankara, Turkey
# CIRAD BGPI, Montpellier, France
¤ Leibnitz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
« University of Milan, Milan, Italy
» Leibnitz Institute of freshwater Ecology and inland Fisheries, Berlin, Germany
^ University of East Anglia, School of Biological Sciences, Norwich, United Kingdom
˅ State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

Corresponding author: Alain Franc (alain.franc@inrae.fr)

Received: 25 Feb 2021 | Published: 04 Mar 2021

Citation: Frigerio J-M, Eckert E, Keskin E, Mahé F, Monaghan MT, Montagna M, Salin F, Schmidt Yáñez P, Yu DW, Fontaneto D, Franc A (2021) OTU picking on large datasets: comparing methods on a diversity of situations. ARPHA Conference Abstracts 4: e65027. https://doi.org/10.3897/aca.4.e65027

Abstract

De novo OTU picking from large metabarcoding read datasets is at the same time a current and a complex task, and several methods coexist to perform it. We present here the outcome of a collective project developed within Working Group on « Data Analysis and Storage » in DNAqua.net. Our aim has been to organize a thorough comparison of OTU composition according to some selected methods called by the wrappers, in a diversity of situations. This has been done by disposing of a set of different datasets, and a set of different methods, applying each method on each dataset, and comparing the results. We have deliberately chosen to work with cleaned datasets only, and not to include cleaning in the process.

We have worked with a set of about 60 different datasets, some environmental, some as mock communities, produced by six teams, in different countries (D, F, I, T, UK), each with

specific markers for different organisms. All datasets have been cleaned beforehand by the team proposing it. We have installed four different tools for building OTUs by unsupervised clustering : Swarm (Mahé et al. 2015), Vsearch (Rognes et al. 2016) with the same receipe for all datasets, usearch (Edgar 2010) with a unique command, the same for all datasets, and yapotu, which computes pairwise Smith-Waterman distances between all reads of a given dataset, and then clusters them with graph based techniques. Yapotu approach is expected to be the most accurate one, as there are no heuristics in the calculations.

We have harmonized common input/output format for the four methods, to make comparisons. Here is a summary of the indicators selected for comparing results.

We have first computed basic indicators per sample and method, like the number of OTU, the number of singletons, the number of OTUs with ten reads or more (after dereplication), and the fraction of reads that have been allocated to an OTU. The four methods displayed a great variety of counts, with highest number of OTUs and singletons for Swarm, then slighltly equivalent figures (but a smaller number of singletons) for yapotu, and significantly smaller counts for Vsearch and Usearch. However, the counts for the number of OTUs with 10 reads or more are much more convergent between the four methods.

We have then compared rank-size curves, which have been computed for all pairs (sample by method). Here again, yapotu and swarm results are very similar, whereas Vsearch and Usearch sometimes are close to the former pattern, sometimes very different (I attach a figure?)

We then have computed 10 different diversity indices, like OTU richness, Shannon, Chao, eveness. Here again, results provided by Swarm and Yapotu are very similar, with very strong correlations between indices over all samples by method, whereas the correlations with Vsearch and Usearch is very poor.

Finally, we have computed all contingency tables (in a sparse format) between all pairs of methods (hence, 6 pairs) for all samples, which accurately describe whether OTUs composition are similar or dissimilar between methods. We have observed that swarm OTUs are systematically nested within yapotu OTUs, and most often, there is a one to one correspondence between a Swarm and a Yapotu OTU.

As a conclusion, we show that

- Swarm and yapotu yield very similar results including for fine details, the only diffrence being a larger number of singletons provided by Swarm ;

- This shows that Swarm OTUs are very close to OTUs built by single linkage Clustering on Smith-Waterman pairwise distances, and consolidates these approaches.

- Very often, both Vsearch and Usearch diverge from those convergent results, but not always, and it is not easy to understand when and why. Some further investigations are needed therefore.

All datasets will be publicly available for further benchmarking of a wider set of methods and datasets.

Keywords

metabarcding, OTU picking, benchmarking

Presenting author

Alain Franc

Presented at

1st DNAQUA International Conference (March 9-11, 2021)

References

- Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26 (19): 2460-2461.
- Mahé F, Rognes T, Quince C, Vargas C, Dunthorn M (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3: e1420.
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4: e2584.