ARPHA Conference Abstracts 4: e64659
doi: 10.3897/aca.4.e64659

Conference Abstract

VTAM: A robust pipeline for validating metabarcoding data using optimized parameters based on internal controls

Emese Meglécz‡, Vincent Dubut‡, Emmanuel Corse§, Aitor González|

‡ Aix Marseille Univ, Avignon Université, CNRS, IRD, IMBE, Marseille, France
§ Centre Universitaire de Mayotte/ MARBEC, CNRS, Ifremer, IRD, University of Montpellier, Dembeni, France
| Aix Marseille Univ, INSERM, TAGC, Turing Center for Living Systems, Marseille, France

Corresponding author: Emese Meglécz (emese.meglecz@imbe.fr)

Received: 19 Feb 2021 | Published: 04 Mar 2021

Citation: Meglécz E, Dubut V, Corse E, González A (2021) VTAM: A robust pipeline for validating metabarcoding data using optimized parameters based on internal controls. ARPHA Conference Abstracts 4: e64659. https://doi.org/10.3897/aca.4.e64659

Abstract

Metabarcoding has become a powerful approach to study biodiversity from environmental samples but it is still prone to some pitfalls. Several papers have called for good practice in study design, data production and analyses to ensure repeatability and comparability between studies. Notably, the importance of mock community samples, negative controls, and replicates is frequently highlighted (Alberdi et al. 2018, O'Rourke et al. 2020). However, their use in bioinformatics pipelines is often limited to *post hoc* verification of expectations by the user. Indeed, one of the biggest challenges in metabarcoding analyses is to take into account the trade-off between false positive (FP) and false negative (FN) occurrences. We thus developed the VTAM (Validation and Taxonomic Assignation of Metabarcoding data) pipeline, which is the first tool to use explicitly the negative control and mock samples to find optimal parameters to minimize false positive and negative occurrences. In addition, VTAM addresses all known technical error types including tag-jumps, repeatability among replicates, and also it is able to integrate more than one overlapping markers to further minimize false negative occurrences.

In order to evaluate VTAM, we compared it with two other pipelines: a pipeline based on DADA2 (Callahan et al. 2016) and LULU (Frøslev et al. 2017), and a pipeline based on

OBITools3 (Boyer et al. 2016) and metabaR (Zinger et al. 2020). Two datasets from fish and bat diet studies were analysed with the three different pipelines. Based on mock and negative samples, we demonstrate that VTAM showed the best precision for mock samples in both datasets, while specificity in negative controls were comparable among the three pipelines (Fig. 1).

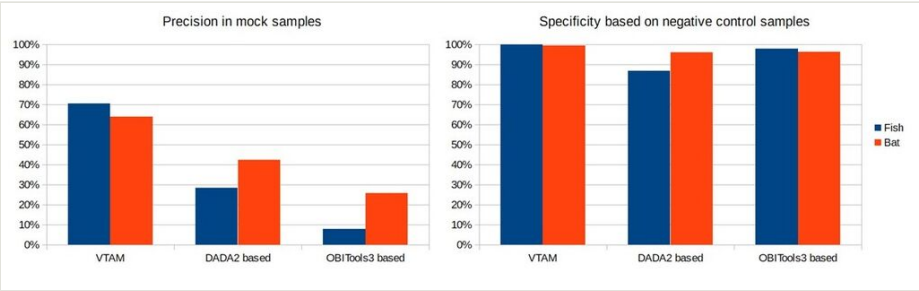

Figure 1. doi

Precision (True positives / (True positives + False positives)) and Specificity (True negatives / (True negative + False positives)) of three pipelines, based on mock samples and negative controls, respectively.

VTAM therefore constitutes a complete pipeline to filter and validate metabarcoding data, from raw FASTQ data to Amplicon Sequence Variant tables with taxonomic assignments. Our pipeline aggregates a series of features rarely grouped in a single pipeline and performs a non-arbitrary parameter optimization based on internal control samples to generate conservative but informative metabarcoding datasets. We believe VTAM provides a very valuable tool for the validation of metabarcoding data, which is essential for conducting robust analyses of biodiversity.

Keywords

metabarcoding, mock sample, negative control, replicates, taxonomic assignation, false positives, false negatives

Presenting author

Emese Meglécz

Presented at

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

References

- Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution 9 (1): 134-147. https://doi.org/10.1111/2041-210X.12849
- Boyer F, Mercier C, Bonin A, Bras YL, Taberlet P, Coissac E (2016) obitools: a unix-inspired software package for DNA metabarcoding. Molecular Ecology Resources 16 (1): 176-182. https://doi.org/10.1111/1755-0998.12428
- Callahan B, McMurdie P, Rosen M, Han A, Johnson A, Holmes S (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13 (7): 581-583. https://doi.org/10.1038/nmeth.3869
- Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications 8 (1): 1-11. https://doi.org/10.1038/s41467-017-01312-x
- O'Rourke D, Bokulich N, Jusino M, MacManes M, Foster J (2020) A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. Ecology and Evolution 10 (18): 9721-9739. https://doi.org/10.1002/ece3.6594
- Zinger L, Lionnet C, Benoiston A, Donald J, Mercier C, Boyer F (2020) metabaR : an R package for the evaluation and improvement of DNA metabarcoding data quality. bioRxiv https://doi.org/10.1101/2020.08.28.271817