VTAM: A robust pipeline for validating metabarcoding data using controls

Aitor González, Vincent Dubut, Emmanuel Corse, Reda Mekdad, Thomas Dechatre, Ulysse Castet, Raphaël Hebert, Emese Meglécz

Computational and Structural Biotechnology Journal 21 (2023) 1151–1156

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

VTAM: A robust pipeline for validating metabarcoding data using controls

Aitor González [a,*], Vincent Dubut [b,c], Emmanuel Corse [d,e], Reda Mekdad [a,b,1], Thomas Dechatre [a,b], Ulysse Castet [a,b], Raphaël Hebert [a,b], Emese Meglécz [b,*]

[a] Aix Marseille Univ, INSERM, TAGC, Marseille, France
[b] Aix Marseille Univ, Avignon Université, CNRS, IRD, IMBE, Marseille, France
[c] ADeNeko, Saint-Girons, France
[d] Centre Universitaire de Formation et de Recherche de Mayotte (CUFR), Dembeni, Mayotte, France
[e] MARBEC, CNRS, Ifremer, IRD, University of Montpellier, Montpellier, France

ARTICLE INFO

Article history:
Received 12 October 2022
Received in revised form 25 January 2023
Accepted 25 January 2023
Available online 27 January 2023

Keywords:
Metabarcoding
Filtering parametrization
Negative controls
Mock samples
Replicates

ABSTRACT

To obtain accurate estimates for biodiversity and ecological studies, metabarcoding studies should be carefully designed to minimize both false positive (FP) and false negative (FN) occurrences. Internal controls (mock samples and negative controls), replicates, and overlapping markers allow controlling metabarcoding errors but current metabarcoding software packages do not explicitly integrate these additional experimental data to optimize filtering. We have developed the metabarcoding analysis software VTAM, which uses explicitly these elements of the experimental design to find optimal parameter settings that minimize FP and FN occurrences. VTAM showed similar sensitivity, but a higher precision compared to two other pipelines using three datasets and two different markers (COI, 16S). The stringent filtering procedure implemented in VTAM aims to produce robust metabarcoding data to obtain accurate ecological estimates and represents an important step towards a non-arbitrary and standardized validation of metabarcoding data for conducting ecological studies. VTAM is implemented in Python and available from: https://github.com/aitgon/vtam. The VTAM benchmark code is available from: https://github.com/aitgon/vtam_benchmark.

1. Introduction

The analysis of environmental DNA (eDNA) using metabarcoding has become a powerful approach to studying biodiversity [33,36]. DNA metabarcoding offers a cost-efficient, (often) non-invasive, and highly sensitive approach for assessing biological diversity from diverse environmental sources including water, soil, bulk samples, sediment, or feces [15,2,24]. However, DNA metabarcoding is prone to a series of now well-documented methodological pitfalls that pave the way from fieldwork to the desktop, passing by the benchtop, and render metabarcoding data especially prone to false negatives (FN), and false positives (FN) [1,43]. If not correctly addressed, FN and FP can hinder robustness, repeatability, and comparability between ecological studies [13,39,8]. Many authors are therefore in search of non-arbitrary and adequate metabarcoding data filtering strategies to standardize biodiversity analyses [11,29,4]. In particular, accurate and exhaustive curation of FP and ensuring repeatability by using technical replicates have proven essential for producing accurate ecological estimates [8,25]. These bioinformatic curation steps involve a series of good practices in study design, notably, the systematic use of mock community samples, negative controls, and technical replicates [29,40,5]. However, the use of mock community and negative control samples in bioinformatics pipelines was often limited to post-hoc analyses by the user to verify the quality expectations (e.g. [23,30,42]).

To date, several original denoising or clustering algorithms are available such as Swarm [26], Unoise [18], Deblur [3], and the widely used DADA2 [9]. These denoising and clustering algorithms are generally efficient for filtering most PCR and sequencing errors, but they cannot account for inter-sample contamination, tag-jump, and chimeras and therefore must be combined with other tools. In such

* Corresponding authors.
E-mail addresses: aitor.gonzalez@univ-amu.fr (A. González), emese.meglecz@imbe.fr (E. Meglécz).
[1] Present address: Institut Curie, LSMP, CurieCoreTech, Paris, France

https://doi.org/10.1016/j.csbj.2023.01.034

integrated workflows (e. g. QIIME2, [6]; Mothur, [34]; PEMA, [42]; SLIM, [17]) several tools are used to denoise data, but control samples (mocks or negatives) are not used explicitly for conducting the filtering parametrization and obtaining robust and accurate results. Although the integration of control samples (mock and negative) is included in the metabaR [44] and Begum [41] programs, none of these programs use an explicit filtering parametrization based on control samples to minimize FP and FN. Instead, they provide graphs for helping users to establish filtering parameters. Yet, control samples have been shown extremely useful for determining data filtering thresholds and discarding false positives from metabarcoding data [11]. Moreover, mock community samples can be used to ensure the comparability of data between distinct high-throughput sequencing runs [12,40]. In these latter cases, negative and mock community samples were processed explicitly to set optimal parameters for validating metabarcoding data and standardizing biodiversity analyses across samples and High Throughput Sequencing (HTS) runs.

Here we introduce VTAM (Validation and Taxonomic Assignment of Metabarcoding data), a variant-based filtering pipeline that deals with Amplicon Sequence Variants (ASVs), which is based on the method described in [11]. VTAM processes explicitly the negative control and mock samples to determine optimal filtering parameter settings for minimizing both false negatives (FN) and false positives (FP), therefore ensuring a balance between the two main types of errors in metabarcoding studies. Additionally, VTAM addresses other methodological pitfalls known to be associated with eDNA metabarcoding (PCR errors, chimeras, and pseudogenes; see Table 1 in Supplementary Information 1) and includes additional features that are unique or rarely included in published pipelines: (i) the explicit use of replicates to ensure repeatability; (ii) the possibility to integrating multiple overlapping markers to further avoid false negatives [12].

2. Materials and methods

2.1. Implementation

VTAM is a command-line application that runs on Linux, MacOS or Windows Subsystem for Linux (WSL) based on the method described in [11]. VTAM is implemented in Python3, using a Conda environment to ensure repeatability and easy installation of VTAM and these third-party applications: WopMars (https://github.com/aitgon/wopmars/), NCBI BLAST, Vsearch [32], Cutadapt [27]. Data is stored in an SQLite database that ensures traceability. Detailed documentation is available at https://vtam.readthedocs.io/en/latest/.

The algorithm is described in detail in Supplementary Information 1. First raw reads are merged, demultiplexed, and trimmed. Then amplicon sequence variants (ASVs; unique sequences characterized by the number of underlying reads) are computed and ASV read counts in different samples and replicates are stored in an SQLite database.

Afterward, the VTAM "filter" command addresses known pitfalls of metabarcoding. The FilterLNF (Filter Low Frequency Noise) and FilterPCRerror filters eliminate occurrences (presence of an ASV in a sample-replicate) with a low relative or absolute frequency. These filters are based on the hypothesis that low read counts are due to weak cross-sample or exogenous contamination, PCR/sequencing errors, and tag-jump. Occurrences are filtered out by the LFN filters if they have low read counts either in absolute terms (Nijk < lfn_read_count_cutoff) or compared to the total number of reads of the sample-replicate (Nijk/Njk < lfn_sample_replicate_cutoff) or compared to the total number of reads of the ASV (Nijk/Ni < lfn_variant_cutoff), where Nijk is the read count of ASV i in sample j and replicate k, Njk is the total number of reads in sample j of replicate k and Ni is the total number of reads of ASV i in the sequencing run. The FilterPCRerror discards occurrences if an ASV closely matches another ASV with a significantly higher number of reads. The cut-off proportion of FilterPCRerror and all three cut-offs of the filterLFN can be obtained in the optimize step (see below).

Two filters eliminate non-reproducible occurrences (FilterMinReplicateNumber), or whole replicates (FilterRenkonen). Chimeras are eliminated by the FilterChimera, and pseudogenes and spurious sequences by FilterCodonStop and FilterIndel.

After the initial low stringency filtering, known true positives (expected occurrences in mock samples) and false positives (unexpected occurrences in control samples) are detected. This information is used by the VTAM "optimize" command to determine optimal parameter values for the FilterFLN and FilterPCRerror steps. This command uses the frequency of unexpected variants in mock samples to suggest a limit for lfn_sample_replicate_cutoff, and the proportion of the closely matching expected and unexpected ASV read counts for FilterPCRerror. The other two cutoffs are determined by counting the FP and FN for the combination of a series of these two parameters and the users can choose the compromise between FP and FN. The "filter" command can then be run again with the optimized parameters (Fig. 1). The "pool" command is then used to produce a single ASV table for multiple overlapping markers by grouping variants identical in their overlapping regions. Finally, a taxonomic assignment based on the Lowest Taxonomic Group method [11] is performed.

2.2. Benchmarking

We have tested VTAM 0.2.0 on two HTS datasets of Cytochrome Oxydase subunit I (COI) obtained from fish and bat feces [11,23] and a 16 S dataset from shark gut content [19]. All three datasets included both negative and positive (mock samples) controls, as well as technical (PCR) replicates in their experimental design. Most samples of the fish dataset were from freshwater, but five samples contained feces from marine fishes. We then compared the results obtained from VTAM to those obtained using the DALU and OBIbaR pipelines. The DALU pipeline is based on the denoising algorithm of DADA2 1.12.1 [9], followed by LULU 0.1.0 [22]. The OBIbaR pipeline starts with filters implemented by OBITools3 3.0.1b19 [7], followed by metabaR 1.0.0 R package [44]. To ensure comparability as much as possible, DALU and OBIbaR pipelines were completed using steps equivalent to VTAM: i) replicates were merged using the equivalent to the MinReplicateNumber step of VTAM; ii) pseudogenes were eliminated by the equivalent of FilterCodonStop and FilterIndel for the COI datasets, and iii) chimeras were eliminated by the equivalent of FilterChimera. Occurrences with a low number of reads are frequently eliminated at the end of analyses, using arbitrary thresholds [25]. Hence, the DALU and OBIbaR pipelines were both completed using 0, 10, 40, and 60 as a cut-off of the minimum number of reads for all three datasets. These values were based on the suggested cut-offs optimized by VTAM (10 for fish, 40 for shark, 60 for bat). For VTAM, this filtering step is included by default and its parameter is established by the "optimize" command. Taxonomic assignments were done by VTAM's "taxassign" command for all three pipelines to ensure homogeneity. Detailed protocols are found in Supplementary Information 2 and the result of each filtering pipeline is available at https://osf.io/rtngk/.

To compare the performance of pipelines we estimated their sensitivity (TP/(TP + FN)), and their precision (TP/(TP + FP)), where TP, FP and FN are the numbers of true positives, false positives, and false negatives, respectively [21]. Finally, the α-diversity and β-diversity estimates were calculated based on the ASV richness and the Jaccard distances respectively, and compared between the pipelines by a t-test using R [31] and the package *vegan* [28].

3. Results

3.1. Precision and sensitivity

Based on the control samples, the precision (TP/(TP+FP) and sensitivity (TP/(TP+FN)) were measured for the three pipelines (Fig. 2). Expectedly, the precision of the DALU and OBIbaR pipelines

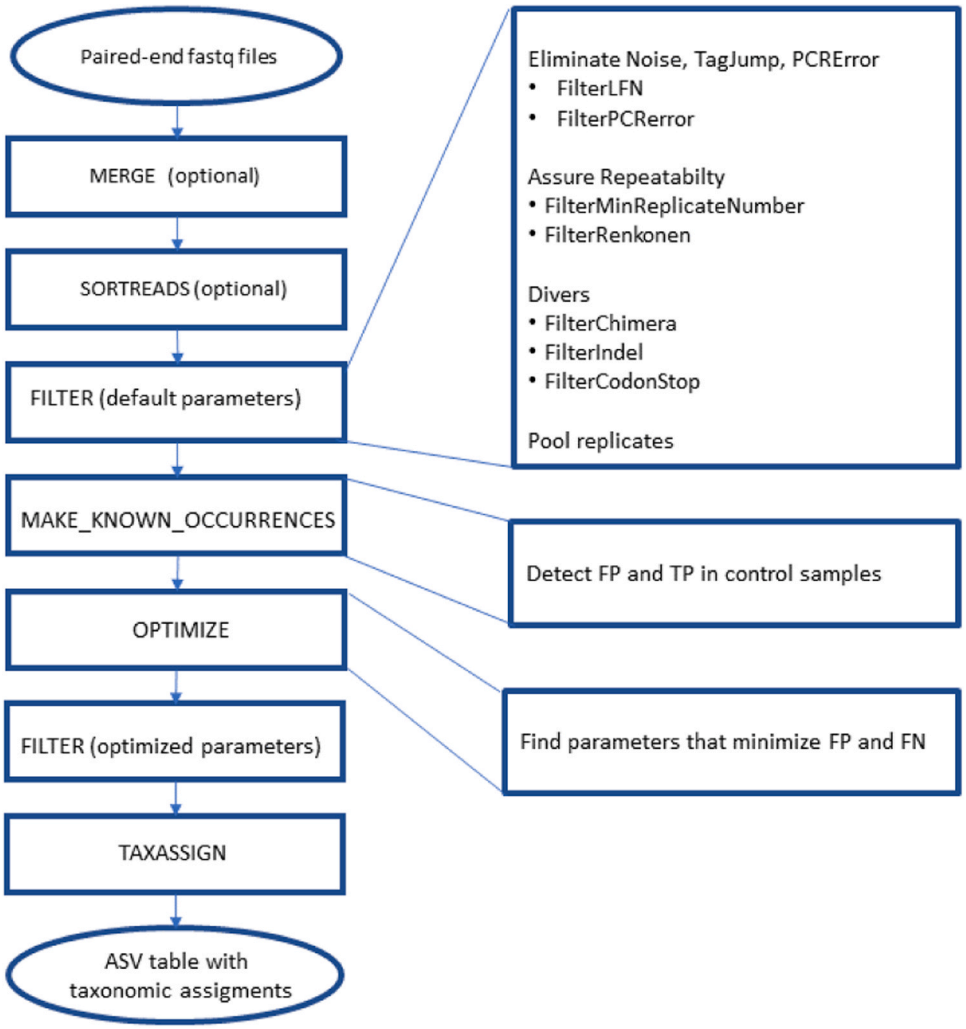

Fig. 1. Workflow of VTAM.

increased with the minimum read count cut-off. While the precision of the OBIbaR pipeline with the minimum read count cut-off 60 was comparable to the precision of VTAM for the fish and bat datasets, it decreased rapidly with the other cut-offs. Moreover, the precision of the DALU pipeline was lower than that of the VTAM even with the highest tested minimum read count cut-off 60.

Sensitivity was high (0.95–1)Agorize for all datasets and pipelines, and it was independent of the minimum read count cut-off for most comparisons. However, for the bat dataset with the DALU pipeline, sensitivity was the highest with the lowest minimum read count cut-off (0) but decreased with increasing minimum read count cut-off, suggesting a trade-off between sensitivity and precision. Overall, the filtering parametrization based on control samples (including but not limited to the read count cut-off) implemented by VTAM allowed us to achieve high precision and sensitivity, outperforming the other pipelines.

Furthermore, we reviewed the ASVs found in the five marine samples from the fish dataset. The ASVs validated by DALU and OBIbaR (using 10 as read count cut-off) contained false positives likely to originate from tag-jump (8 for OBIbaR 41 for DALU), since they were variants also found in freshwater samples and from taxa that cannot be encountered in marine environments. Based on these marine samples VTAM revealed far higher precision (1.00) than OBIbaR (0.66) or DALU (0.19).

3.2. The effect of the filtering procedure on biodiversity estimates

The ASV richness was significantly higher in DALU and OBIbaR compared to VTAM, in the bat dataset, for all read count cut-offs, in the fish dataset using 0 and 10 cut-offs and in the shark dataset using 0, 10, and 40 cut-offs (Fig. 3, Supplementary Fig. S1, Supplementary Table S1). On the contrary, between-sample dissimilarities (related to β-diversity) were significantly lower when estimated from data obtained from DALU and OBIbaR compared to VTAM in all comparisons (Fig. 3 and Supplementary Fig. S1). This indicates that VTAM filters out more variants per sample and implies the ASV data obtained from VTAM has a more discriminating power than the data obtained from the other pipelines (Fig. 3).

4. Discussion

4.1. VTAM addresses most technical pitfalls and determines non-arbitrary thresholds

The biggest challenge in filtering metabarcoding data is to consider the trade-off between false positive (FP) and false negative (FN) occurrences. All filtering procedures are a compromise between eliminating spurious sequences and losing true signals [16,25,37]. It is therefore critical to find optimal filtering parameters to achieve a

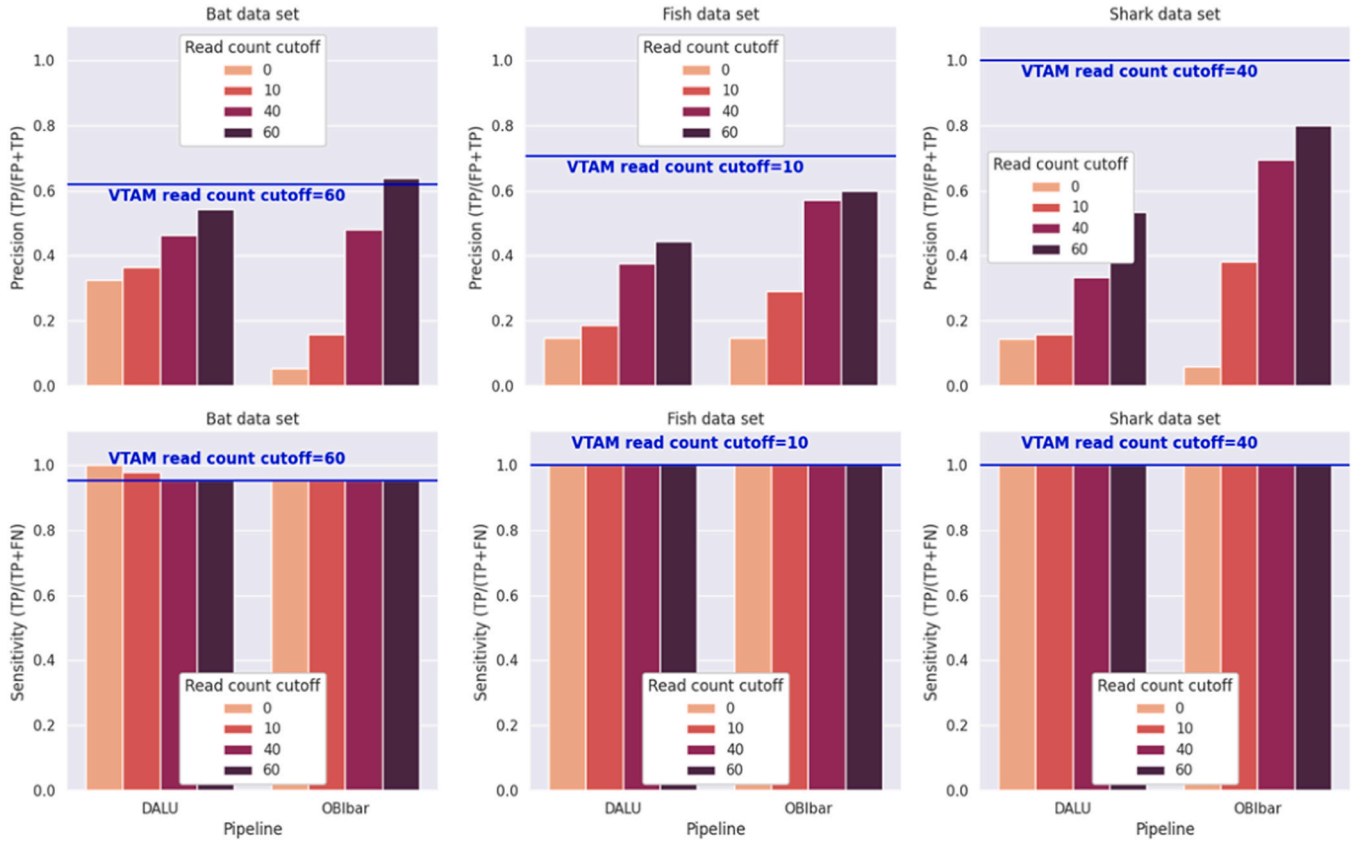

Fig. 2. Precision and sensitivity of the three pipelines based on control samples (mock and negative control). The horizontal blue lines give the precision and sensitivity of the VTAM software with the given cut-off. The read count cut-off was optimized by VTAM (60 for bat, 10 for fish and 40 for shark). For the DALU and OBIbaR pipelines, 4 cut-offs were used based on the optimized VTAM cut-offs and arbitrary values (0, 10, 40, 60).

balance between false positives and negatives. We have compared the performance of VTAM to two different pipelines using three different datasets and two different markers. Each pipeline is composed of a denoising step (DADA2 and OBITools) enhanced by a series of further steps to decrease false positives and thus increase precision. The precision of VTAM estimated both from control samples or a subset of marine environmental samples is the highest among the pipelines, and its sensitivity is also high and similar to the sensitivity of the other pipelines. To achieve a similar sensitivity and precision to VTAM, the original denoising algorithms (DADA2, OBITools) had to be completed by further filtering steps including a series of arbitrary cut-offs of minimum read count to find the best performance. On the contrary, VTAM integrates several filters to control for known artifacts of metabarcoding dealing with false

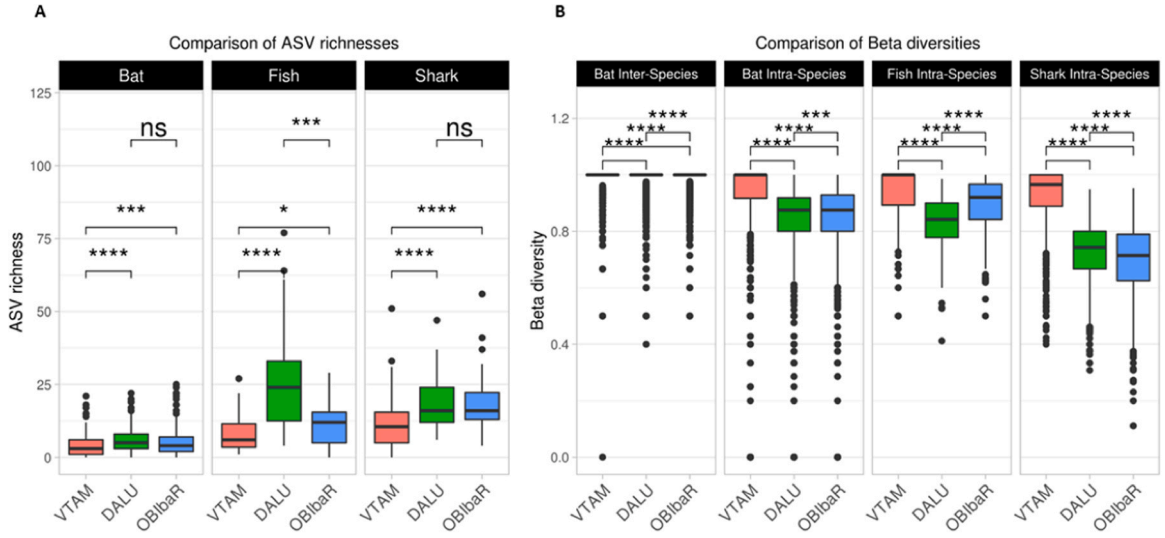

Fig. 3. ASV richness (A) and beta-diversity (B) of the three metabarcoding pipelines calculated for the three datasets (bat, fish and shark). The diversity estimates were calculated based on real samples (without control samples). The bat dataset is the result of diet analyses of different species. Therefore, beta diversity was calculated separately within and between bat species. The read count cut-off was 60 for the bat, 10 for the fish, and 40 for shark datasets as optimized by VTAM. **** P < 0.0001, *** P < 0.001, ** P < 0.01, * P < 0.05.

positives (sequencing and PCR errors, tag-jump, chimeras, pseudogenes) and repeatability (PCR heterogeneity). Furthermore, the real strength of VTAM is to include the explicit and non-arbitrary optimization of parameter combinations, while the parametrization must be done separately and more subjectively for the other pipelines. The underlying hypothesis of the parameter optimization based on control samples is that the best parameter setting that produces clean control samples can also be applied to real samples and reduces FP and FN in the whole dataset. If all species are present in an equimolar concentration in the mock and amplify well by PCR, the parameter setting suggested by the optimize command can be very stringent, and rare or badly amplified species can be eliminated from real samples. This calls for the careful construction of mock communities to include species representative of the diversity of the targeted group. Tag-jump [35] and cross-sample contamination are rarely explicitly considered in current metabarcoding pipelines. However, failing to filter out these artifacts is likely to inflate false positive occurrences, which in turn skew β-diversity estimates by inflating inter-sample similarities. This is the pattern we observed in our comparison: the DALU and OBIbaR pipelines produced a significantly higher ASV richness per sample than VTAM, but dissimilarities between samples were lower. Moreover, several false positives due to tag-jump were not filtered out by DALU and OBIbaR in real (marine) samples.

The use of technical replicates is an important tool to minimize both false positives and false negatives [2,20]. False positives can be strongly reduced by accepting an occurrence (presence of a variant in a sample) only if the variant is present in at least a certain number of replicates. This strategy is strongly advised to reduce experimental stochasticity to validate ASV occurrences. Furthermore, removing highly dissimilar replicates from other replicates of the same sample (renkonen filter) further reduces the effect of experimental stochasticity [14]. Although these repeatability filters are implemented in VTAM and metabaR, they are absent from all other existing tools.

4.2. Not only bioinformatics: the importance of the benchtop

Adequate use of VTAM implies a thorough experimental design and reAgorizelies on negative and positive controls. In particular, standardized mock samples are critical. While commercial mock samples are available for some taxonomic groups (bacteria and fungi), the metabarcoders that study metazoa or plants must not neglect the construction of adequate mock samples to standardize their analyses (see: [12,30]). Apart from finding adequate parameters for data analyses, mock samples are also useful in ensuring that sufficient sequencing coverage has been achieved. Furthermore, in large-scale studies involving several distinct high-throughput sequencing runs, the systematic use of identical mock samples as standards for parametrization of the filtering should minimize the effect of random fluctuations and make samples comparable between runs [12,40].

Finally, the use of multiple primer sets presents an invaluable means for limiting false negatives [12,37]. It improves the detection of taxa and haplotypes by mitigating the loss of poorly amplified taxa or haplotypes by a single primer pair.

4.3. Implications in metabarcoding studies

The choice of metabarcoding data filtering and validation strategies is critical for obtaining robust data and accurate ecological estimates [16,37,43]. Only robust and standardized filtering are expected to produce adequate data for conducting thorough metaphylogeographic studies (e.g. [38]), ecological studies (e.g. [40]), and ecosystem monitoring (e.g. [10]). In this perspective, a precise curation of false positives (e.g. sequencing and PCR errors, chimeras, internal and external contamination) and ensuring repeatability by using technical replicates have both proven essential for producing accurate biodiversity estimates [8,25]. The stringent filtering procedure implemented in VTAM aims to produce robust metabarcoding data for the estimation of accurate ecological estimates and represents an important step towards the standardization of the validation of metabarcoding data for conducting ecological studies.

5. Conclusion

VTAM is a comprehensive pipeline that provides tables of validated ASVs with taxonomic assignments from raw input FASTQ files. VTAM addresses many technical issues listed in the metabarcoding literature for validating metabarcoding data by including features rarely considered in most metabarcoding pipelines: (i) parameter optimization based on control samples, (ii) explicit handling of replicates and (iii) multiple overlapping markers. Its precision is higher than other frequently used bioinformatic pipelines. Specifically, the filtering procedure of VTAM explicitly uses control samples and technical replicates to provide an accurate curation of false positives. The optimization procedure based on control samples provides an objective data filtering strategy to standardize biodiversity analyses. While the stringent filtering of VTAM decreases (within-sample) α-diversity and increases (between-samples) β-diversity, it has proven to provide accurate estimates to conduct robust ecological studies (e.g. [40]). We, therefore, believe VTAM represents an innovative and so-expected tool for the robust validation of metabarcoding data and for conducting ecological analyses.

Funding

This work is a contribution to the European project SEAMoBB, funded by ERA-Net Mar-TERA and managed by ANR (number ANR_17_MART-0001_01).

CRediT authorship contribution statement

Aitor González: Software, Validation, Visualization, Supervision, Writing - review & editing. **Vincent Dubut:** Conceptualization, Methodology, Writing - original draft. **Emmanuel Corse:** Conceptualization, Methodology. **Reda Mekdad:** Software. **Thomas Dechatre:** Software. **Ulysse Castet:** Software. **Raphaël Hebert:** Software. **Emese Meglécz:** Conceptualization, Methodology, Writing - original draft, Validation, Writing - review & editing, Supervision.

Acknowledgments

We thank Diane Zarzoso-Lacoste and Samanta Ortuño-Miquel for valuable comments on the operational use of VTAM, Luc Giffon and Lionel Spinelli for the development of Wopmars, and Kurt Villsen for English editing. The Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high-performance computing resources.

Supplementary Information 1

Detailed algorithm of the VTAM pipeline.

Supplementary Information 2

A detailed protocol of the benchmarking.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.01.034.

References

[1] Alberdi A, Aizpurua O, Bohmann K, Gopalakrishnan S, Lynggaard C, Nielsen M, Gilbert MTP. Promises and pitfalls of using high-throughput sequencing for diet analysis. Mol Ecol Resour 2019;19(2):327–48. https://doi.org/10.1111/1755-0998.12960

[2] Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K. Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods Ecol Evol 2018;9(1):134–47. https://doi.org/10.1111/2041-210X.12849

[3] Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. Deblur rapidly resolves single-nucleotide community sequence patterns. MSystems 2017;2(2). https://doi.org/10.1128/mSystems.00191-16

[4] Antich A, Palacin C, Wangensteen OS, Turon X. To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. BMC Bioinforma 2021;22(1):177. https://doi.org/10.1186/s12859-021-04115-6

[5] Bakker MG. A fungal mock community control for amplicon sequencing experiments. Mol Ecol Resour 2018;18(3):541–56. https://doi.org/10.1111/1755-0998.12760

[6] Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. (Article). Nat Biotechnol 2019;37(8):8. https://doi.org/10.1038/s41587-019-0209-9

[7] Boyer F, Mercier C, Bonin A, Bras YL, Taberlet P, Coissac E. obitools: a unix-inspired software package for DNA metabarcoding. Mol Ecol Resour 2016;16(1):176–82. https://doi.org/10.1111/1755-0998.12428

[8] Calderón-Sanou I, Münkemüller T, Boyer F, Zinger L, Thuiller W. From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices. J Biogeogr 2020;47(1):193–206. https://doi.org/10.1111/jbi.13681

[9] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 2016;13(7):581–3. https://doi.org/10.1038/nmeth.3869

[10] Cordier T, Alonso-Sáez L, Apothéloz-Perret-Gentil L, Aylagas E, Bohan DA, Bouchez A, Chariton A, Creer S, Frühe L, Keck F, Keeley N, Laroche O, Leese F, Pochon X, Stoeck T, Pawlowski J, Lanzén A. Ecosystems monitoring powered by environmental genomics: a review of current strategies with an implementation roadmap. Mol Ecol 2021;30(13):2937–58. https://doi.org/10.1111/mec.15472

[11] Corse E, Meglécz E, Archambaud G, Ardisson M, Martin J-F, Tougard C, Chappaz R, Dubut V. A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies. Mol Ecol Resour 2017;17(6):e146–59. https://doi.org/10.1111/1755-0998.12703

[12] Corse E, Tougard C, Archambaud-Suard G, Agnèse J-F, Mandeng FDM, Bilong CFB, Duneau D, Zinger L, Chappaz R, Xu CCY, Meglécz E, Dubut V. One-locus-several-primers: a strategy to improve the taxonomic and haplotypic coverage in diet metabarcoding studies. Ecol Evol 2019;9(8):4603–20. https://doi.org/10.1002/ece3.5063

[13] Cristescu ME, Hebert PDN. Uses and misuses of environmental DNA in biodiversity science and conservation. Annu Rev Ecol, Evol, Syst 2018;49(1):209–30. https://doi.org/10.1146/annurev-ecolsys-110617-062306

[14] De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, Taberlet P. DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. Mol Ecol Resour 2014;14(2):306–23. https://doi.org/10.1111/1755-0998.12188

[15] Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, Vere N, de, Pfrender ME, Bernatchez L. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Mol Ecol 2017;26(21):5872–95. https://doi.org/10.1111/mec.14350

[16] Drake LE, Cuff JP, Young RE, Marchbank A, Chadwick EA, Symondson WOC. An assessment of minimum sequence copy thresholds for identifying and reducing the prevalence of artefacts in dietary metabarcoding data. Methods Ecol Evol 2022;13(3):694–710. https://doi.org/10.1111/2041-210X.13780

[17] Dufresne Y, Lejzerowicz F, Perret-Gentil LA, Pawlowski J, Cordier T. SLIM: a flexible web application for the reproducible processing of environmental DNA metabarcoding data. BMC Bioinforma 2019;20(1):88. https://doi.org/10.1186/s12859-019-2663-2

[18] Edgar RC. UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv 2016:081257 https://doi.org/10.1101/081257

[19] Esposito A, Sasal P, Clua É, Meglécz E, Clerissi C. Shark Provisioning Influences the Gut Microbiota of the Black-Tip Reef Shark in French Polynesia. (Article). Fishes 2022;7(6):6. https://doi.org/10.3390/fishes7060312

[20] Ficetola GF, Pansu J, Bonin A, Coissac E, Giguet-Covex C, De Barba M, Gielly L, Lopes CM, Boyer F, Pompanon F, Rayé G, Taberlet P. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Mol Ecol Resour 2015;15(3):543–56. https://doi.org/10.1111/1755-0998.12338

[21] Fletcher RH, Fletcher SW, Fletcher GS. Clinical epidemiology: the essentials (Fifth edition). Wolters Kluwer Health/Lippincott Williams & Wilkins; 2014.

[22] Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nat Commun 2017;8(1):1–11. https://doi.org/10.1038/s41467-017-01312-x

[23] Galan M, Pons J-B, Tournayre O, Pierre É, Leuchtmann M, Pontier D, Charbonnel N. Metabarcoding for the parallel identification of several hundred predators and their prey: application to bat species diet analysis. Mol Ecol Resour 2018;18(3):474–89. https://doi.org/10.1111/1755-0998.12749

[24] Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, Hamer KC, Wilcove DS, Bruce C, Wang X, Levi T, Lott M, Yu DW. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol Lett 2013;16(10):1245–57. https://doi.org/10.1111/ele.12162

[25] Littleford-Colquhoun BL, Freeman PT, Sackett VI, Tulloss CV, McGarvey LM, Geremia C, Kartzinel TR. The precautionary principle and dietary DNA metabarcoding: commonly used abundance thresholds change ecological interpretation. Mol Ecol 2022;31(6):1615–26. https://doi.org/10.1111/mec.16352

[26] Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ 2014;2:e593https://doi.org/10.7717/peerj.593

[27] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 2011;17(1):10–2. https://doi.org/10.14806/ej.17.1.200

[28] Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. Vegan: community ecology package. R Package Version 2020;2:5–7 ⟨https://cran.r-project.org/web/packages/vegan/⟩.

[29] O'Rourke DR, Bokulich NA, Jusino MA, MacManes MD, Foster JT. A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. Ecol Evol 2020;10(18):9721–39. https://doi.org/10.1002/ece3.6594

[30] Pauvert C, Buée M, Laval V, Edel-Hermann V, Fauchery L, Gautier A, Lesur I, Vallance J, Vacher C. Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. Fungal Ecol 2019;41:23–33. https://doi.org/10.1016/j.funeco.2019.03.005

[31] R Development Core Team. (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

[32] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016;4:e2584 https://doi.org/10.7717/peerj.2584

[33] Ruppert KM, Kline RJ, Rahman MS. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. Glob Ecol Conserv 2019;17:e00547 https://doi.org/10.1016/j.gecco.2019.e00547

[34] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Horn DJV, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75(23):7537–41. https://doi.org/10.1128/AEM.01541-09

[35] Schnell IB, Bohmann K, Gilbert MTP. Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. Mol Ecol Resour 2015;15(6):1289–303. https://doi.org/10.1111/1755-0998.12402

[36] Taberlet P, Bonin A, Zinger L, Coissac É. Environmental DNA: For Biodiversity Research and Monitoring. Oxford University Press; 2018 ⟨https://academic.oup.com/book/32663⟩.

[37] Tercel MPTG, Cuff JP. The complex epistemological challenge of data curation in dietary metabarcoding: Comment on "The precautionary principle and dietary DNA metabarcoding: commonly used abundance thresholds change ecological interpretation" by Littleford-Colquhoun et al. (2022). Mol Ecol 2022;31(22):5653–9. https://doi.org/10.1111/mec.16576

[38] Turon X, Antich A, Palacín C, Præbel K, Wangensteen OS. From metabarcoding to metaphylogeography: Separating the wheat from the chaff. Ecol Appl 2020;30(2):e02036 https://doi.org/10.1002/eap.2036

[39] van der Loos LM, Nijland R. Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Mol Ecol 2021;30(13):3270–88. https://doi.org/10.1111/mec.15592

[40] Villsen K, Corse E, Meglécz E, Archambaud-Suard G, Vignes H, Ereskovsky AV, Chappaz R, Dubut V. DNA metabarcoding suggests adaptive seasonal variation of individual trophic traits in a critically endangered fish. Mol Ecol 2022;31(22):5889–908. https://doi.org/10.1111/mec.16698

[41] Yang C, Bohmann K, Wang X, Cai W, Wales N, Ding Z, Gopalakrishnan S, Yu DW. Biodiversity Soup II: a bulk-sample metabarcoding pipeline emphasizing error reduction. *BioRxiv*, 2020 07 2020;07:187666 https://doi.org/10.1101/2020.07.07.187666

[42] Zafeiropoulos H, Viet HQ, Vasileiadou K, Potirakis A, Arvanitidis C, Topalis P, Pavloudi C, Pafilis E. PEMA: a flexible pipeline for environmental DNA metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. GigaScience 2020;9(3). https://doi.org/10.1093/gigascience/giaa022

[43] Zinger L, Bonin A, Alsos IG, Bálint M, Bik H, Boyer F, Chariton AA, Creer S, Coissac E, Deagle BE, De Barba M, Dickie IA, Dumbrell AJ, Ficetola GF, Fierer N, Fumagalli L, Gilbert MTP, Jarman S, Jumpponen A, Taberlet P. DNA metabarcoding—need for robust experimental designs to draw sound ecological conclusions. Mol Ecol 2019;28(8):1857–62. https://doi.org/10.1111/mec.15060

[44] Zinger L, Lionnet C, Benoiston A-S, Donald J, Mercier C, Boyer F. metabaR: an R package for the evaluation and improvement of DNA metabarcoding data quality. Methods Ecol Evol 2021;12(4):586–92. https://doi.org/10.1111/2041-210X.13552