Bioinformatics, 2025, **41(3)**, btaf089
https://doi.org/10.1093/bioinformatics/btaf089
Advance Access Publication Date: 20 March 2025
Applications Note

OXFORD

Sequence analysis

NanoASV: a snakemake workflow for reproducible field-based Nanopore full-length 16S metabarcoding amplicon data analysis

Arthur Cousson[1,*], Frédéric Mahé[2,3], Ulysse Guyet[4], Damase Razafimahafaly[5], Laetitia Bernard[1]

[1]Eco&Sols, University of Montpellier, IRD, INRAe, CIRAD, Inst Agro, Montpellier F-34060, France
[2]CIRAD, UMR PHIM, F-34398 Montpellier, France
[3]PHIM, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France
[4]Metabolic Genomics, Genoscope, Institut de Biologie François Jacob, CEA, CNRS, Université d'Evry, Université Paris Saclay, Evry 91000, France
[5]Laboratoire des Radio-Isotopes, BP 3383, Route d'Andraisoro, Antananarivo 101, Madagascar

*Corresponding author. Eco&Sols, University of Montpellier, IRD, INRAe, CIRAD, Inst Agro, 2 Pl Viala, Montpellier F-34060, France. E-mail: NanoASV@proton.me.

Associate Editor: Can Alkan

Abstract

Summary: NanoASV is a conda environment and snakemake-based workflow using state-of-the-art bioinformatics software to process full-length SSU rRNA (16S/18S) amplicons acquired with Oxford Nanopore Sequencing technology. Its strength lies in reproducibility, portability, and the possibility to run offline, allowing in-field analysis. It can be installed on the Nanopore MK1C sequencing device and process data locally.

Availability and implementation: Source code and documentation are freely available at https://github.com/ImagoXV/NanoASV and Zenodo archive at https://doi.org/10.5281/zenodo.14730742.

1 Introduction

Oxford Nanopore Technologies (ONTs) offer an inexpensive and mobile solution for acquiring third-generation high-throughput sequencing data. However, their commercial computational solution, Epi2Me, suffers from several technical limitations (e.g. no taxonomy curation, unknown sequences grouping, and non-universal data format). Additionally, its dependency on internet connectivity and the transmission of large data volumes, particularly during the taxonomic affiliation step, make it challenging to use in remote areas with unreliable internet access. Only a few solutions addressing these challenges were published (Santos *et al.* 2020, Rodríguez-Pérez *et al.* 2021, Zorz *et al.* 2023).

Among these, NanoCLUST (Rodríguez-Pérez *et al.* 2021) is a reliable and efficient operational taxonomical unit-based analysis pipeline that offers species-level resolution. NanoCLUST does not seem to be maintained anymore, and the taxonomic recovery by SituSeq (Zorz *et al.* 2023) is hindered by a high prevalence of unknown classifications (see Supplementary Information S1). To overcome these limitations, we present NanoASV, a reference-based ASV workflow specifically designed for locally processing full-length SSU rRNA (16S/18S) metabarcoding sequencing data.

NanoASV is inspired by Nygaard *et al.*'s ASV-based pipeline (Nygaard *et al.* 2020) but aims to be more portable through encapsulation and more precise when processing unknown sequences.

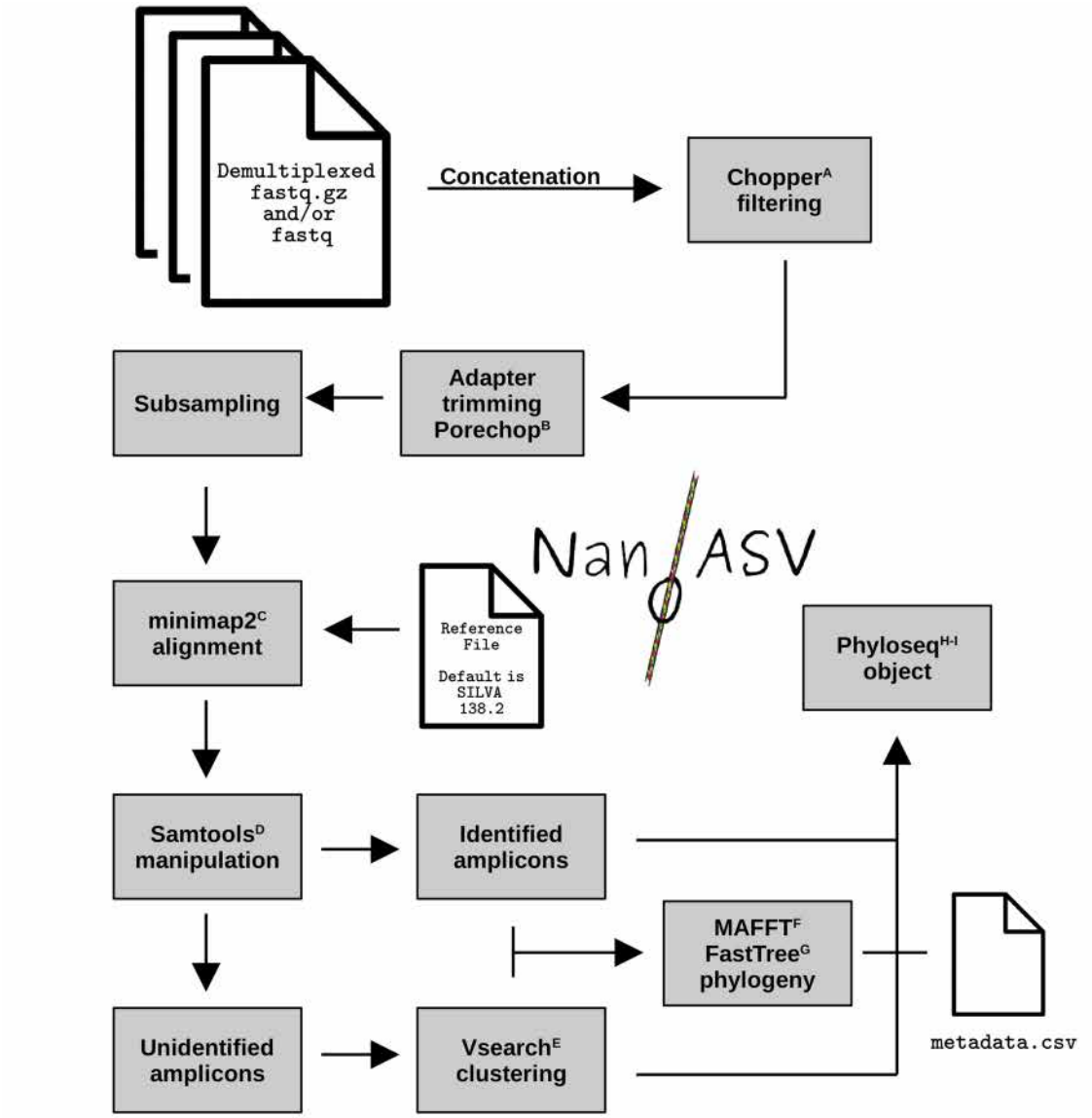
2 Description

NanoASV uses a conda environment to ensure portability and reproducibility. By default, NanoASV accepts demultiplexed fastq files organized in barcode0 to barcodeXX sub-folders (e.g. "fastq_pass" if basecalled by the sequencing device), a CSV metadata file that is used to build a phyloseq object (*.Rdata* output file), and a CSV assignment file. NanoASV can also directly process unorganized fastq files (if they are found in the specified *--dir* directory). In that context, if no metadata file is provided, NanoASV will generate a dummy one.

The workflow (summarized in Fig. 1) processes both compressed and uncompressed fastq files, which are typically 4000 sequences long. It groups sequences by sample, using barcode identifiers. Subsequently, the sequences are filtered using Chopper (De Coster and Rademakers 2023) with specific parameters (*--quality 8*, *--minlength 1400 --maxlength 1700* by default) to remove low-quality reads. Porechop (https://github.com/rrwick/Porechop—Ryan Wick) is employed to identify and remove sequencing adapters from the filtered sequences.

By default, fastq files are subsampled to 50 000 sequences per barcode, but this can be changed with the option

Received: 8 July 2024; **Revised:** 18 February 2025; **Editorial Decision:** 21 February 2025; **Accepted:** 18 March 2025



A) De Coster, Wouter, and Rosa Rademakers. 2023. "NanoPack2: Population-Scale Evaluation of Long-Read Sequencing Data." Edited by Can Alkan. Bioinformatics 39 (5): btad311. https://doi.org/10.1093/bioinformatics/btad311.
B) https://github.com/rrwick/Porechop – Ryan Wick
C) Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." Edited by Inanc Birol. Bioinformatics 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
D) Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." GigaScience 10 (2): giab008. https://doi.org/10.1093/gigascience/giab008.
E) Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. "VSEARCH: A Versatile Open Source Tool for Metagenomics." PeerJ 4 (October):e2584. https://doi.org/10.7717/peerj.2584.
F) Katoh, K., and D. M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." Molecular Biology and Evolution 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.
G) Price, M. N., P. S. Dehal, and A. P. Arkin. 2009. "FastTree: Computing Large Minimum Evolution Trees with Profiles Instead of a Distance Matrix." Molecular Biology and Evolution 26 (7): 1641–50. https://doi.org/10.1093/molbev/msp077.
H) McMurdie, Paul J., and Susan Holmes. 2013. "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." Edited by Michael Watson. PLoS ONE 8 (4): e61217. https://doi.org/10.1371/journal.pone.0061217.
I) Paradis, Emmanuel, and Klaus Schliep. 2019. "Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R." Edited by Russell Schwartz. Bioinformatics 35 (3): 526–28. https://doi.org/10.1093/bioinformatics/bty633.

Figure 1. NanoASV workflow chart.

--subsampling. These subsampled sequences are then aligned with minimap2 (Li 2018) (*--model map-ont* by default) against a user-provided reference fasta file (see README file for reference requirements), or by default against SILVA138.2 (Quast *et al.* 2013). Unmapped reads, secondary alignments, and supplementary alignments are removed using samtools (*samtools view -F 4 -F 256 -F 272 -F 2048 -F 2024*) (Danecek *et al.* 2021) to only keep correctly mapped reads. Assignment data and statistics are written to flat .csv files. Mapping quality threshold (MapQ) can be adjusted with *--samtools-qual* (default 0, see Supplementary Information S3).

One of the key innovations of NanoASV lies in its treatment of unknown sequences (see Supplementary Information S2). As an example, with Epi2Me or SituSeq (Zorz *et al.* 2023), unassigned sequences are typically assigned the label "Unknown" without extracting any additional information. This approach is

problematic because it does not differentiate between singletons or rare erroneous sequences and abundant potentially meaningful DNA sequences within the dataset. To address this challenge, we have implemented a vsearch-based clustering step (Rognes *et al.* 2016) specifically for unassigned sequences. To account for the high error rate associated with ONT data, vsearch is set to use a low clustering similarity threshold (*--id 0.7* by default). Unknown clusters with a total abundance lower than 5 are excluded from the final results.

Reference and unknown consensus sequences are pooled and aligned with MAFFT (Katoh and Standley 2013). A 16S-based phylogenetic tree is computed with FastTree (Price *et al.* 2009).

Abundance tables of assigned and unassigned clusters, taxonomy table, and 16S-based phylogeny are used to produce a phyloseq object. NanoASV expects to find a metadata.csv file in the same directory as the demultiplexed barcodes (location controlled by the option *--metadata*). Using *ape* and *phyloseq* R packages (McMurdie and Holmes 2013, Paradis and Schliep 2019), NanoASV outputs a readily usable phyloseq object containing comprehensive information for both assigned and unknown clusters, as well as an abundance and taxonomy table (csv), and a phylogenetic tree. Sequences assigned to Eukaryota, Chloroplast, and Mitochondria are removed by default from the phyloseq object. This can be modified with the *--no-r-cleaning* option.

Benchmarking NanoASV against the SituSeq and Nygaard *et al.* pipelines showed that it is faster and more memory-efficient while recovering similar alpha diversity trends and taxonomic profiles (see Supplementary Information S1).

3 Installation and usage

NanoASV is easy to install. Detailed instructions can be found in the README file: (i) clone the github repository; (ii) run the installation script; and (iii) activate the conda environment and run

nanoasv --dir path/to/dir --out path/to/output [--options]

A self-test with a small dataset can be run with *nanoasv --mock*, producing a typical NanoASV analysis output. All available options are detailed on the GitHub README file.

NanoASV runs on GNU/Linux x86-64 and Aarch64 systems, as well as on the Nanopore MK1C sequencing device (Aarch64—Minion MK1C version).

4 Conclusion

NanoASV is an efficient and reliable local container-based solution to process Nanopore metabarcoding amplicon data. Its innovative approach to handling unassigned sequences allows users to discover potentially new molecular diversity. Its portability allows it to be installed directly on the Nanopore MK1C sequencing device and to process data locally, making it suited for offline usage and in-field analysis.

Acknowledgements

We are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing help, computing, and storage resources. We thank Enrique Ortega for his valuable contributions in the early stages of the project. We also appreciate Antoine Cousson, Fiona Elmaleh, and Meren for their role in software beta testing. Finally, we are grateful to the two anonymous reviewers for their insightful comments and suggestions.

Author contributions

Arthur Cousson (Conceptualization [lead], Funding acquisition [supporting], Investigation [lead], Methodology [lead], Software [lead], Validation [lead], Visualization [lead]), Frédéric Mahé (Methodology [equal], Resources [equal], Software [equal], Supervision [lead]), Ulysse Guyet (Methodology [equal], Resources [equal], Software [equal]), Damase Razafimahafaly (Resources [Supporting]), and Laetitia Bernard (Funding acquisition [lead], Project administration [equal], Supervision [equal])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

The PhD grant that allowed this production is a Contrat Doctoral Spécifique Normalien granted by the École Normale Supérieure de Paris and the French Ministère de l'Enseignement Supérieur et de la Recherche. This study was also funded by the French National Institute of Research for Development (IRD). This study was financially supported by ANR under the framework of the U2Worm project (ANR-20-CE01-0015-01) and under the Investissements d'Avenir programme with the reference ANR-10-LABX-001-01 Labex Agro and coordinated by Agropolis Fondation, under the framework of the Innov'Earth Project (convention 2101-003) and the MetaCast project (convention 2202-214).

Data availability

The benchmark data underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.14979767

References

Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008

De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* 2023;**39**:btad311. https://doi.org/10.1093/bioinformatics/btad311

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;**30**:772–80. https://doi.org/10.1093/molbev/mst010

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191

McMurdie PJ, Holmes S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;**8**:e61217. https://doi.org/10.1371/journal.pone.0061217

Nygaard AB, Tunsjø HS, Meisal R *et al.* A preliminary study on the potential of Nanopore MinION and Illumina MiSeq 16S rRNA gene sequencing to characterize building-dust microbiomes. *Sci Rep* 2020;**10**:3209. https://doi.org/10.1038/s41598-020-59771-0

Paradis E, Schliep K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 2019;**35**:526–8. https://doi.org/10.1093/bioinformatics/bty633

Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009;**26**:1641–50. https://doi.org/10.1093/molbev/msp077

Quast C, Pruesse E, Yilmaz P *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–96. https://doi.org/10.1093/nar/gks1219

Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics* 2021;**37**:1600–1. https://doi.org/10.1093/bioinformatics/btaa900

Rognes T, Flouri T, Nichols B *et al.* VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016;**4**:e2584. https://doi.org/10.7717/peerj.2584

Santos A, van Aerle R, Barrientos L *et al.* Computational methods for 16S metabarcoding studies using nanopore sequencing data. *Comput Struct Biotechnol J* 2020;**18**:296–305. https://doi.org/10.1016/j.csbj.2020.01.005

Zorz J, Li C, Chakraborty A *et al.* SituSeq: an offline protocol for rapid and remote nanopore 16S rRNA amplicon sequence analysis. *ISME Commun* 2023;**3**:33. https://doi.org/10.1038/s43705-023-00239-3

Bioinformatics, 2025, 41, 1–4
https://doi.org/10.1093/bioinformatics/btaf089
Applications Note