Jan 16, 2025 Version 1

DBP Metabarcoding Pipeline (for metabarcoding data using nanopore) V.1

DOI
https://dx.doi.org/10.17504/protocols.io.j8nlk94rxv5r/v1

Muhammad Danie Al Malik[1], Ni Kadek Dita Cahyani[2], Aji Wahyu Anggoro[3]

[1]Diponegoro Biodiversity Project (DBP), Universitas Diponegoro, Indonesia;
[2]Biology Department, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia;
[3]Yayasan Konservasi Alam Nusantara, Jakarta, Indonesia

Muhammad Danie Al Malik
Diponegoro Biodiversity Project (DBP)

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.j8nlk94rxv5r/v1

Protocol Citation: Muhammad Danie Al Malik, Ni Kadek Dita Cahyani, Aji Wahyu Anggoro 2025. DBP Metabarcoding Pipeline (for metabarcoding data using nanopore) . **protocols.io** https://dx.doi.org/10.17504/protocols.io.j8nlk94rxv5r/v1

Protocol status: Working
We use this protocol and it's working

Created: November 24, 2024

Last Modified: January 16, 2025

Protocol Integer ID: 112701

Keywords: MinION, Nanopore, Metabarcoding, Bioinformatic, eDNA, raw fastq data from nanopore base, analysis of environmental dna, using nanopore, environmental dna, taxonomic assignment, nanopore base, dbp metabarcoding pipeline, operational taxonomic unit, metabarcoding data, artifact detection with chimera, taxon table format, taxon table format suitable for assignment, dna, raw fastq data, nanofilt, phyloseq object, metabarcoding purpose, quality filtering with nanofilt, using blastn, pipeline

Disclaimer

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.The content in this pipeline is provided for informational and sharing purposes only. Any actions taken based on this information are at your own risk. You agree that the Company, its authors, contributors, administrators, or anyone associated with protocols.io cannot be held responsible for your use of the information contained in or linked to this protocol.

Abstract

This pipeline was developed to process MinION nanopore data output for metabarcoding purposes. The process begins with raw FASTQ data from nanopore base-calling output and includes several steps: quality filtering with NanoFilt, primer removal and length trimming with Cutadapt, artifact detection with Chimera, and building an Operational Taxonomic Unit (OTU) table with VSEARCH. Taxonomic assignment is then performed using Blastn. Additionally, this pipeline integrates an R script to build a taxon table format suitable for assignment in a phyloseq object. All processes were tested and run from the terminal on a Linux system using the Ubuntu distribution. This pipeline ensures a streamlined workflow, facilitating the analysis of environmental DNA (eDNA) data.

Before start

This pipeline must execute on Terminal via Linux (the test was run on Ubuntu 22.04.5 via Windows Sub-system Linux or WSL).

Acquire Pipeline Data

1 The pipeline requires fastq.gz (format data) from MinION nanopore sequencing output. This protocol's example data comes from the Flongle MinION nanopore of the R10.1 flow cell. Please visit Malik et al. (2024) for details on the data information.

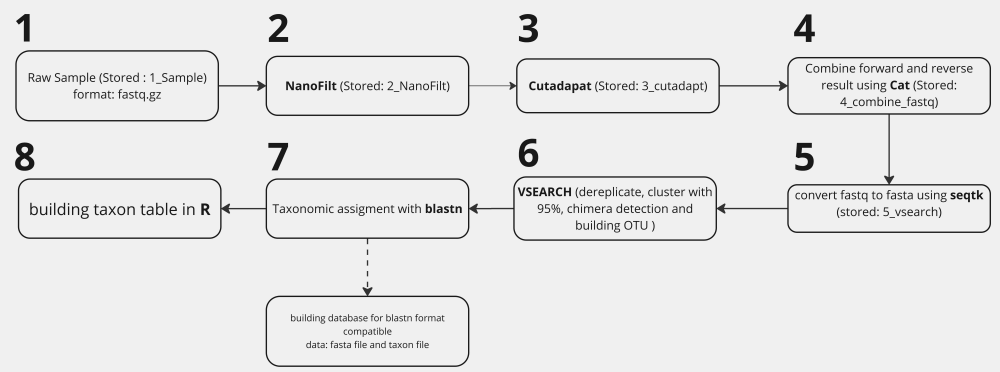

Figure 1. This is the flow chat from the pipeline

This is the explanation about each process on this pipeline:

1. **Raw Sample (Stored: 1_Sample)**: The journey begins with a raw sample in the fastq.gz format.
2. **NanoFilt (Stored: 2_NanoFilt)**: The raw sample undergoes filtering on quality and/or read length, which is tailored for nanopore sequencing data.
3. **Cutadapt (Stored: 3_cutadapt)**: The filtered data then goes through Cutadapt, a tool designed to trim adapter/primer sequences from sequencing reads.
4. **Combine forward and reverse result using Cat (Stored: 4_combine_fastq)**: Forward and reverse reads are combined using the cat command.
5. **Convert fastq to fasta using seqtk (Stored: 5_vsearch)**: The combined fastq file is then converted to the fasta format using seqtk.
6. **VSEARCH**: The fasta file is processed with VSEARCH for dereplication, clustering (with 95% similarity), chimera detection, and building OTUs (Operational Taxonomic Units).
7. **Taxonomic assignment with blastn**: The OTUs are assigned taxonomic labels using blastn.
8. **Building taxon table in R**: Finally, the taxonomic assignments are utilized to construct a taxon table in R.

1.1 Prepare the sample:

- Ensure that your sample is in fastq.gz format. Examples fastq.gz files from this protocol could be downloaded through this [**link**]

- Place the sample in a folder named "1_Sample".

1.2 Prepare the Database:
- Obtain a database in fasta (sequence list) and txt (taxon names) formats. An example of the required format can be seen at this [**link**].
- Rename the fasta file to "database.fasta" and the txt file to "database.txt".

1.3 Download the Syntax Script to run the pipeline:
- Download the syntax script in .sh format from this [**link**].

1.4 Download the R Script:
- Download the R script to build a taxon table from this [**link**].

1.5 Download the Syntax Script to Install Software (Optional):
- Download a syntax script in .sh format to compile software installation via the terminal. This script allows you to install all necessary software with a single command. You can download it from this [**link**].

1.6

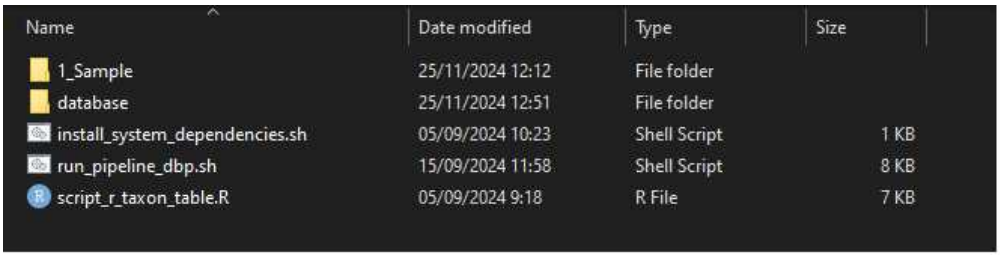

Figure 2. This is an example list of files required to run this pipeline.

Set up Pipeline

2 Installing several software need for this pipeline

- NanoFilt (**wdecoster/nanofilt: Filtering and trimming of long read sequencing data**)
- Cutadapt (**Cutadapt — Cutadapt 0.1 documentation**)
- seqtk (**lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats**)
- VSEARCH (**torognes/vsearch: Versatile open-source tool for microbiome analysis**)
- blastn (**BLAST+ executables — BLASTHelp documentation**)
- R (**https://www.r-project.org/**)

2.1 We have created a .sh file named "install_system_dependencies.sh" to install all the necessary software listed above [link].

Processing Pipeline

3 All the processing pipelines conducted from Linux (for demo using Ubuntu distro)

3.1 Open the terminal on Ubuntu and navigate to the folder that contains the dependency files.

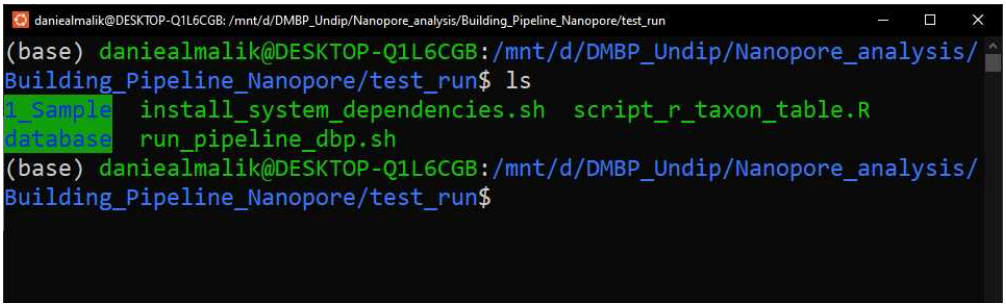

Figure 3. This example demonstrates that the dependency files are in the "test_run" folder.

3.2 (Optional step) Install any necessary software needs from "install_system_dependencies.sh".

```
chmod +x install_system_dependencies.sh
./install_system_dependencies.sh
```

- "chmod +x" is a command used to add the execute permission to a file, allowing it to be run as a program; essentially, it makes the file executable.
- "./" is a command to execute a file.

3.3 Execute the sh file "run_pipeline_dbp.sh"

```
chmod +x run_pipeline_dbp.sh
./run_pipeline_dbp.sh
```

4 The output file on the folder will be like this:

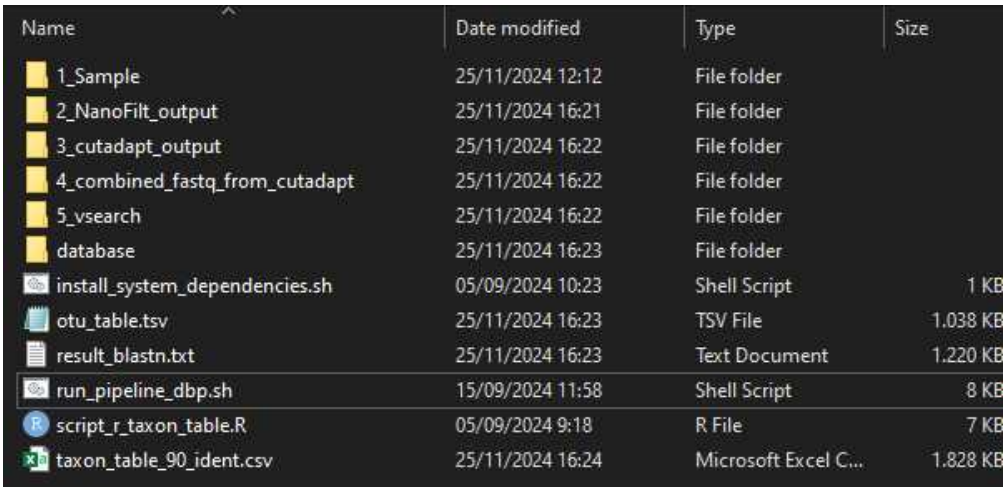

Figure 4. The output file on the folder will be like this.

Table 1. The explanation about output folders and files

	Output	Explanation
	2_NanoFilt_output	This folder contain files output from NanoFilt process with fastq format
	3_cutadapt_output	This folder contain files output from cutadapt process (trim primer forward and reverse) with fastq format
	4_combine_fastq_from_cutadapt	This folder contain files output from cutadapt process (combine output forward and reverse primer) with fastq format
	5_vsearch	This folder contain files from vsearch output with fasta format
	otu_table.tsv	This file contain otu table from each samples
	result_blastn.txt	This file contain result of blastn process
	taxon_table_90_ident.csv	This file contains taxon table after filtering with minimum of 90% identity

Editing Parameter

5 We can edit the parameter based on our necessary analysis by editing "run_pipeline_dbp.sh" through Notepad++. Please check the syntax line that could be edited for each process.

- NanoFilt parameter

Quality threshold (line 13)
min_length (line 16)
max_length (line 19)

- Cutadapt parameter

Primer list from Forward and Reverse (line 43-44)
ERROR_RATE (line45)
MIN_LENGTH (line 46)
MAX_LENGTH (line 47)

- VSEARCH parameter

clustering and building otu table from minimum identity "--id" (line 192 and line 202)

- blastn parameter

blastn parameter (line 208)

- R script to build taxon table "script_r_taxon_table.R"

Protocol references

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*, 1-9.

De Coster, W., D'hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, *34*(15), 2666-2669.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), 299-314.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), 10-12.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584.