Unlocking the Potential of MinION: A Hands-On Genomic Sequencing Workshop

Hands-on: Targeted metagenomic analysis based on MinION sequence data

This section provides a step-by-step guide to analyze 16S rDNA targeted metagenomic data using MinION sequence data. The workflow covers the entire process, from the initial quality control of raw sequencing reads to the visualization of bacterial composition and diversity within and between the samples.

This Hands-on guide covers:

- Quality control of full 16S rDNA sequence data
- Read clustering and polishing
- Taxonomy assignment of clusters
- Diversity analysis of bacterial community
- Visualization of bacterial composition.

1. Dataset Overview

For this tutorial, the bacteriome of *Aedes aegypti* well be analyzed. *Aedes aegypti* is a well-known vector of viruses such as yellow fever, dengue, chikungunya, and Zika. The dataset will be used to investigates the variation in the bacterial community across developmental stages of the mosquito—from larvae to adult (male and female). We'll work with 12 samples divided into four groups: larvae (4), adult males (4), and adult females (4).

2. Open the terminal session and change the working directory

To begin, launch a terminal session and navigate to your working directory

```
cd Desktop/AmpliconAnalysis
```

3. Activate the CONDA working environment

Before starting the analysis, activate the CONDA environment that contains the necessary tools for the analysis pipeline.

```
# Check available environments
conda info --envs

# Activate GeomeTools environment
conda activate AmpliconTools
```

4. Explore the sequence data

Before proceeding with the analysis, it is crucial to assess the quality of the raw sequence data. For this tutorial, we assume the data has already been demultiplexed and trimmed of barcodes and adapters. The raw reads of each sample are stored in individual FASTQ files.

4.1.Unzip FASTQ files

Unzip all FASTQ files in the Data/ folder:

```
#Change directory to Data folder
cd Data

#Unzip all Fastq files (* refers to all files)
gunzip *.fastq.gz
```

4.2.Assess Data Quality with NanoPlot

NanoPlot allows us to quickly assess the quality of the raw Fastq files. It generates reports on the number of reads, their quality, and the mean read length.

```
#Run Nanoplot for all the data together
NanoPlot --fastq *.fastq -o Nanoplot_raw --no_static
```

Output: The output is a folder named Nanoplot_raw containing general reports in both TXT and HTML formats, along with various plots that visualize the data quality.

4.3.Count Reads Per Sample

The script **ReadCount.sh** allows to calculate the total number of reads per FASTQ file located in the current directory.

```
#Run ReadCount.sh script (the dot “.” refers to the current directory)
ReadCount.sh .
```

Output: A summary of the read counts will be displayed in the terminal and saved in a text file (reads_count.txt).

5. Filter raw sequencing data

To improve the quality and the accuracy of the analysis, it's important to filter out the raw reads based on average quality score and length, focusing on reads that are between 1200–1600 bp, which is the expected length range for 16S rDNA.

5.1.Filter reads

The script **ProcessInput.sh** helps to filter the raw sequencing data based on quality scores and read length for all the samples. The script is based on NanoFilt with minQscore of 10 and target length between 1200bp and 1600bp.

The command needs as an input one argument:

- Path to raw FASTQ files: Data/
- Minimum quality score: 10
- Minimum read length in base pairs: 1200
- Minimum read length in base pairs: 1600

```
#Go back to the main working directory (Desktop/AmpliconAnalysis)
cd ..

# Run the ProcessInput.sh script with specified quality score and length
ProcessInput.sh Data/ 10 1200 1600
```

Output: The new filtered FASTQ file containing only high-quality reads with the target size are saved in Results/Processed_fq.

5.2.Post-Filtering Quality Check

After filtering, it's important to re-assess the quality of your data to ensure the filtering process was successfully performed.

```
#change directory to Results/processed_fq
cd Results/Processed_fq

#Run NanoPlot
NanoPlot --fastq *.fastq -o Nanoplot_filt --no_static

#run read_count command
ReadCount.sh .
```

Output: The output of NanoPolt in a folder named Nanoplot_Filt. Output of ReadCount.sh, summary of the read counts will be displayed in the terminal and saved in a text file (reads_count.txt).

6. Reads clustering and polishing

Once you have high-quality reads, the next step is to cluster them. To this end we will use a modified pipeline based on NanoClust[1] for clustering and polishing. The script NB_ReadClust.sh automates the various steps of this process.
The command needs as an input one argument:

- Path to processed FASTQ files: Results/processed_fq/

```
#Return to the main working directory
cd ~/Desktop/AmpliconAnalysis

## run NB_ReadClust.sh command indicating the path to filtered fastq
files (≈ 11min)
NB_ReadClust.sh Results/Processed_fq
```

Output: The outputs of **NB_ReadClust.sh** command will be saved in the folder Results/, which includes various intermediate files and folder for each step of the analysis:

- **0_cluster_reads/**: K-mer frequencies and clustering results.
- **1_split_cluster/**: Reads separated by cluster.
- **2_draft_polish/**: Polished sequences for each cluster.
- **all_cluster_seq.fa**: A FASTA file containing polished representative sequences for each cluster.

clusters_count_tab.txt: Tab delimited table showing the read count per sample for each cluster.

7. Assign taxonomy

After the clustering/polishing and the read count for each cluster, the next step is taxonomic assignment for the representative sequences of each cluster. We will use QIIME 2 with the Silva database for this step. The taxonomy assignment of the selected representative of each cluster can be performed using QIIM2 algorithm based on silva database. To this end, the command **AssignTaxa.sh** will be used using all_cluster_seq.fa file which contains the cluster sequences. The command needs as an input three arguments:

- Fasta file: all_cluster_seq.fa
- Path to Silva database sequences: SILVA_db/ silva-138-99-seqs.qza
- Path to Silva database taxa: SILVA_db/ silva-138-99-tax.qza

[1] Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. Bioinformatics. 2021 Jul 12;37(11):1600-1601.

```
#Run AssignTaxa command (≈ 12min)
AssignTaxa.sh Results/all_cluster_seq.fa SILVA_db/silva-138-99-seqs.qza
SILVA_db/silva-138-99-tax.qza
```

Output: The output is a tab delimited table (TSV format) (taxonomy.tsv). The file contains the cluster ID, and the closest taxonomy identified, formatted as: domain;phylum;class;order;family;genus;species.

Additionally, the AssignTaxa.sh script will generate a folder called final_output/, which contains the following files essential for downstream analysis using MetaXplore:

- **taxonomy.tsv:** the taxonomic classification table.
- **clusters_count_tab.txt:** the read counts per cluster for each sample.

8. Run MetaXplore application

Once the taxonomy assignment and data processing are complete, you can use MetaXplore to perform the downstream analysis. MetaXplore[2] is a user-friendly web application designed to explore alpha diversity, beta diversity, relative abundance, and visualize the results in various graph formats.

To run MetaXplore, start the app on the system using the following command:

```
#Run a local MetaXplore app
R -e "shiny::runApp('~/Desktop/AmpliconAnalysis/MetaXplore', port=5277)"
```

MetaXplore requires the following input files, which are generated in previous steps:

- **clusters_count_tab.txt:** located on the folder (Results/final_output/), this file contains the read counts per cluster for each sample
- **taxonomy.txt:** located on the folder (Results/final_output/), this file contains the taxonomic classification of each cluster
- **Metadata:** located on the folder (AmpliconAnalaysis/), this file contains information about the samples (Developmental stage and Sex)

Once the application is running, open a web browser and navigate to http://localhost:5277 to access the MetaXplore interface and begin your analysis.

[2] Bel Mokhtar N, Asimakis E, Galiatsatos I, Maurady A, Stathopoulou P, Tsiamis G. Development of MetaXplore: An Interactive Tool for Targeted Metagenomic Analysis. Curr Issues Mol Biol. 2024 May 15;46(5):4803-4814

For remote access, you can use MetaXplore by visiting the following link: http://metaxplore.eu/

The source data for MetaXplore is available on GitHub Repository: https://github.com/nbel15/MetaXplore.