An optimal data-analysis pipeline to determine a potential microbiome shift in COPD patients

Birgit Rijvers, Jeanine Gastreich & Eefje Schrauwen

Avans University of Applied Sciences, School of Life Sciences and Technology, Breda, the Netherlands

Abstract

Chronic obstructive pulmonary disease (COPD) is a prevalent respiratory condition characterized by airflow obstruction and inflammation, often leading to acute exacerbations. While azithromycin shows promise in reducing exacerbations, concerns regarding antibiotic resistance and side effects arise with prolonged use. Understanding the impact of discontinuing azithromycin treatment on exacerbations and patient well-being requires assessing changes in the COPD-associated microbiome. We aimed to identify an optimal pipeline for analyzing 16S rRNA gene sequencing data by comparing five pipelines: NanoCLUST, NanoRTax, BugSeq, Emu, and Kraken2/Bracken. We evaluated their accuracy using a Zymo mock sample with known bacterial DNA composition and a sputum sample spiked with *Allobacillus* and *Imtechella*. NanoCLUST demonstrated robust performance for taxonomic classification on a genus level. Subsequently, we utilized NanoCLUST to analyze sputum samples from a single patient collected at months 0, 3, 6, and 9. Rstudio facilitated data visualization through heatmaps, enabling comparison of microbiome compositions across patient groups. This study provides insights into COPD-associated microbiome changes and highlights NanoCLUST as a reliable pipeline for taxonomic analysis of 16S rRNA gene sequencing data. The visualization approach enhances interpretation and comparison of microbiome data in COPD research, advancing our understanding of the effects of discontinuing azithromycin treatment in COPD patients.

Introduction

Chronic obstructive pulmonary disease (COPD) is a prevalent respiratory condition that significantly impacts global healthcare systems, resulting in high mortality rates and substantial costs. This disease is distinguished by the obstruction of airflow and inflammation occurring within the respiratory tract. While smoking remains the primary cause, other genetic and environmental factors have also been implicated.[1][2][3][4] This progressive condition often manifests as acute exacerbations, worsening COPD symptoms, including dyspnea and coughing. Often, patients suffering from acute exacerbation must be hospitalized.[4][5] The primary objective of COPD treatment is to minimize exacerbations. Macrolide antibiotics have proven to be effective in reducing exacerbation frequency. Azithromycin, belonging to the macrolide class, offers multiple benefits due to its anti-inflammatory, antibacterial and immunomodulatory properties.[6] It is commonly employed in the treatment of chronic inflammatory airway diseases and helps regulate epithelial cell interactions in the airways. Studies have demonstrated that prolonged administration of azithromycin and other antibiotics can effectively decrease exacerbations in COPD patients. [7] However, extended antibiotic use carries the risk of antibiotic resistance development, posing a significant concern.[8] Additionally, the use of azithromycin may give rise to potential side effects ranging from common symptoms like abdominal pain, nausea, vomiting, and diarrhea to more serious complications like hearing loss, tinnitus and long QT syndrome.[9]

To understand the various host-microbe interactions happening in the human body, precise characterization of the human microbiome is a critical first step. The predominant approach employed to achieve this goal involves sequencing the 16S ribosomal RNA (rRNA) gene of bacteria. This gene, consisting of approximately 1500 base pairs (bp), plays a vital role in bacterial functioning and encompasses nine hypervariable regions dispersed throughout an otherwise highly conserved sequence.[10][11] Notably, Oxford Nanopore Technologies (ONT) have gained significant attention in recent research as a means to sequence the complete 16S rRNA gene, serving various purposes such as antimicrobial susceptibility testing and species identification within the human skin and gut microbiome. Furthermore, the utilization of 16S rRNA gene sequencing has substantially contributed to our understanding of the lung microbiome in both healthy individuals and those with diseases.[12] Most scientific studies of the lung microbiome that are based on 16S rRNA gene sequencing identify microbes to the genus level, but often also mention phyla. Occasionally, species can be identified.[13][14][15]

The analysis of Next-Generation Sequencing (NGS) data poses a significant computational challenge due to the complexity and noise inherent in the produced data. To address this challenge, numerous tools and pipelines have been developed. Pipelines, in particular, offer comprehensive data analysis, taking raw data as input and generating (visualized) results as output. In recent years, novel tools and methods have emerged to enhance the accuracy and efficiency of NGS data analysis

In this study, we compare and evaluate several pipelines for the analysis of 16S rRNA gene sequencing data. The pipelines discussed in this section include BugSeq, NanoCLUST, NanoRTax, Emu and Kraken 2/Bracken. BugSeq is a pipeline specifically designed for the analysis of raw basecalled Nanopore reads, packaged with Nextflow. Although the tool is free, it is only available online and not open source, with limited flexibility in modifying the settings of the tools within the pipeline. The output of BugSeq includes QC summary, krona plots, and relative abundance data. BugSeq adopts a hybrid approach that combines long-read sequencing data with reference-based and de novo assembly methods.[16] NanoCLUST, on the other hand, is a pipeline developed specifically for the analysis of amplicon-based full-length 16S rRNA gene Nanopore sequencing reads. It is implemented in Nextflow and can be run on Linux if Docker or conda is installed.[17] NanoRTax, developed by the same authors as NanoCLUST and released in 2022, is also a Nextflow-based pipeline designed for taxonomic classification of bacteria based on 16S rRNA gene sequencing data. It offers state-of-the-art read classification tools, quality control, and real-time analysis capabilities. NanoRTax is free, open-source, and intended to run on a laptop or desktop computer using Docker or Conda on Linux.[18] Emu is a tool that utilizes an expectation-maximization (EM) algorithm for taxonomic abundance estimation from long 16S rRNA gene sequencing reads. The Emu pipeline consists of two stages: read alignment to a reference database, followed by the EM algorithm to refine the relative abundance of detected species. The output provides an estimation of the microbial community composition for each sample. Emu is compatible with Linux and utilizes conda.[19] Lastly, Kraken 2 is a tool specifically developed for taxonomic classification of DNA sequence reads. It employs a probabilistic hash table method to map minimizers to lowest common ancestors of reads, which enables accurate classification of millions of reads within minutes, achieving high speed and accuracy.[20] Bracken (Bayesian Reestimation of Abundance with KrakEN) is a statistically robust methodology that effectively determines the species abundance within DNA sequences obtained from a metagenomics sample. It functions as a supplementary software to Kraken 2. While Kraken assigns reads to various levels in the taxonomic hierarchy, Bracken enables the estimation of abundance at a specific level utilizing these classifications.[21]

To understand the impact of stopping azithromycin treatment on exacerbations and patient wellbeing, the main objective of the Vasco da Gama study is to assess the potential changes in the microbiome composition of COPD patients after discontinuation of azithromycin treatment using 16S rRNA gene sequencing. Specifically, in this paper we aim to identify the most suitable method for analyzing 16S rRNA gene sequencing data of COPD patients. Additionally, we intend to find a visualization tool that enables comparison of microbiome data from consecutive sputum samples across different patient groups.

A summary of this study including links to the used pipelines and tools can be found on our GitHub page.[1]

Materials & Methods

Phenotypic determination

Upon collection of sputum from a patient, investigation of the bacterial composition was conducted using both Gram staining and culturing techniques.

A high quality flake of sputum was washed with physiological saline in a sterile petri dish. This mixture was inoculated onto a blood agar and chocolate agar plate and incubated at 36 degrees Celsius for 48 hours. Additionally, a Gram stain was prepared from the same mixture. On day 1 and 2 each of the plates was investigated for the presence of *Staphylococcus aureus*, *Staphylococcus pneumoniae*, β-hemolytic streptococci, *Moraxella catharralis*, *Haemophilus influenzae*, *Enterobacter*, non-fermentors and over colonization by other bacteria.

Datasets

To assess the performance of the different pipelines two different types of samples were used. A clinical sputum sample was prepared for analysis by adding ZymoBIOMICS™ Spike-in Control I (High Microbial Load, D6320. Zymo Research Corp. located in Irvine, CA, United States). This spike-in control contained equal cell numbers of *Imtechella halotolerans* and *Allobacillus halotolerans*. Additionally, the analysis included ZymoBIOMICS™ Microbial Community DNA standard (D6305), which provided a DNA composition with known precision (Table 1). FASTQ files of both samples were obtained using ONT MinION sequencing.

Table 1: Theoretical Composition ZymoBIOMICS™ Microbial Community DNA Standard based on Genomic DNA

Micro-organism	Percentage amount
Listeria monocytogenes	12%
Pseudomonas aeruginosa	12%
Bacillus subtilis	12%
Escherichia coli	12%
Salmonella enterica	12%
Lactobacillus fermentum	12%
Enterococcus faecalis	12%
Staphylococcus aureus	12%
Saccharomyces cerevisiae	2%
Cryptococcus neofromans	2%

[1] https://github.com/BirgitRijvers/COPD-Microbiome-Shift-Analysis.git

Quality control and filtering

Nanopore signal basecalling was carried out using the high-accuracy model of Guppy (v6.0.7) pipeline (Oxford Nanopore Technologies – ONT, Oxford, UK). The quality and quantity of reads form the Zymo mock sample and the sputum sample 5004 were assessed using NanoStat (v1.6.0). Subsequently, Filtlong (v0.2.1) was utilized to create subsets of 10% and 90% of reads from both the Zymo mock sample and the clinical sputum sample. The parameter--keep_percent 10 was employed to retain the top 10% of reads, while --keep_percent 90 was used to discard the lowest-quality 10% of reads.

Data analysis

In order to identify the most effective bioinformatic approach for analyzing the relative abundance of the microbiome at the genus level, various pipelines and tools were employed.

BugSeq The FASTQ files containing the subsets of reads were imported into the BugSeq (v2023-05-08) 16S sequencing analysis pipeline along with the NCBI nt Metagenomic database. The pipeline was configured with the following settings: platform Nanopore, Device & Chemistry MinION/GridION/Flonge -R9.4.1 Guppy 5/6 Super accuracy Basecalling, and lower respiratory sample type. The default DNA settings were applied for the sequenced material.

The pipelines NanoCLUST, NanoRTax and the tool Emu were executed using a Linux terminal Ubuntu Desktop 20.04, equipped with 16 cores and 64 GB RAM as the computational environment. Initially, miniconda3 (v23.5.0) was installed to manage the software dependencies. The workflow management system Nextflow (v22.10.6) was installed with Miniconda3 Linux 64-bit to facilitate the utilization of NanoCLUST and NanoRTax.

NanoCLUST The NanoCLUST software (V1.0dev) was installed following the instructions provided on the corresponding GitHub page[2]. As a containerization platform Docker was utilized. Additionally, the configuration file nextflow.config in the NanoCLUST directory was modified by making changes to line 104 as follows:

```
withName: consensus_classification { container
= 'ncbi/blast:latest'
```

To further improve the of NanoCLUST, it was necessary to modify the get_abundances.py file located in the templates directory of NanoCLUST. Specifically, changes were made to line 22 and onwards. The original code snippet:

```
Try:
        name =
json.loads(complete_tax)[0][tax_level_tag]
    except:
        name = str(int(tax_id))
    return
json.loads(complete_tax)[0][tax_level_tag]
```

[2] https://github.com/genomicsITER/NanoCLUST

was replaced by the following code:

```
path =
'http://api.unipept.ugent.be/api/v1/taxonomy.j
son?input[]=' + str(int(tax_id)) +
'&extra=true&names=true'
    complete_tax = requests.get(path).text

    # Check if the list returned by
json.loads() is not empty
    tax_list = json.loads(complete_tax)
    if len(tax_list) > 0:
        name = tax_list[0][tax_level_tag]
    else:
        name = str(int(tax_id))

    return name
```

The NanoCLUST analysis utilizes the blastn algorithm, with the NCBI RefSeq database and the NCBI taxonomy database as reference databases.

NanoRTax For NanoRTax, the installation was performed following instructions on the corresponding GitHub page[3]. As a containerization platform Conda was utilized. The NCBI blast database, Kraken2 RDP database and the Centrifuge database were used in the analysis.

Prior to utilizing the tools Emu and Kraken2/Bracken, a quality control and filtering step was implemented using the fastp (v0.23.4) software. The following settings were applied during the fastp process:--qualified_quality_phred 8,--length_required 1400, and –length_limit 1700.

Emu Emu (v3.4.5) was installed using Conda, following the guidelines specified on the GitLab page[4]. To analyze the data, the provided Emu database, which is a combination of rrnDB v5.6 and NCBI 16S RefSeq data from 17 September 2020, was employed.

Kraken2/Bracken The analysis using Kraken2 was conducted on the EU servers of the Galaxy platform. The FASTQ files were uploaded to Galaxy for further analysis. Kraken2 (v2.1.1+galaxy1) was employed with the parameter for single reads, and the specific Kraken2 standard database (created 2021-05-17) was utilized, and the -report option was enabled to create a report as input for Bracken. To re-estimate the abundance at the taxonomic level based on the Kraken output, Bracken (v2.8+galaxy0) was employed. The analysis was performed using the parameters for the Kmer distribution level Standard (2021-05-17) and the taxonomy class at genus level.

Visualization

Following the retrieval of relative abundances from different pipelines and tools, the data was visualized using R (v4.3.0) and RStudio (v2023.03.1). Barplots depicting relative abundances higher than 0.1% per pipeline/tool were generated using R packages ggplot2 (v3.4.2), RColorBrewer (v1.1-3), tidyr (v1.3.0) and phyloseq (v1.44.0). Additionally, a barplot visualizing the count of all uniquely identified genera identified per pipeline/tool was generated using the ggplot2 package.

[3] https://github.com/genomicsITER/NanoRTax

[4] https://gitlab.com/treangenlab/emu

Sputum microbiome analysis patient

Upon selecting the most suitable pipeline for analyzing the relative abundance of sputum, the sputum samples collected from a patient at four different timepoints (Table 2) were utilized to investigate potential shifts in the microbiome. Nanopore signal basecalling was performed using the high-accuracy model of Guppy (v6.1.5) pipeline. The quality and quantity of reads obtained from the four sputum samples were assessed using NanoStat (v1.6.0.) Subsequently, NanoCLUST analysis was conducted using the blastn algorithm, utilizing the NCBI Refseq database and the NCBI taxonomy database as a reference database for analyzing the sputum samples. The resulting data was visualized in a heatmap using the ggplot2 and phyloseq packages in Rstudio (v4.3.0).

Table 2: Sputum samples collected for Microbiome analysis

Sample name	Date collected
M0	03-06-2021
M3	14-09-2021
M6	14-12-2021
M9	15-03-2022

Results

Phenotypic determination

Culturing of the sputum samples collected unveiled the presence of bacteria belonging to the *Streptococcus* and *Haemophilus* genera within the one sputum sample used for testing the performance of the different classification methods. The four sputum samples collected from one patient at different time points showed that the sputum contained *Streptococcus* bacteria. *Streptococcus* was also present in het sputum collected in month 3, together with *Haemophilus* bacteria. The sample from month 6 contained Gram-positive cocci, as well as the sample collected at month 9. This sample additionally contained *Haemophilus* bacteria (Table 3).

Table 3: Cultured bacteria in sputum samples

Sample	Cultured bacteria
Sputum used for testing pipelines	*Streptococcus*, *Haemophilus*
Patient sputum M0	*Streptococcus*
Patient sputum M3	*Streptococcus*, *Haemophilus*
Patient sputum M6	Gram-positive cocci
Patient sputum M9	*Haemophilus*, Gram-positive cocci

Quality control of 16S rRNA amplicon sequencing reads

In order to assess the quality of the 16S rRNA amplicon sequencing reads obtained from the sputum and Zymo mock reads, several basic quality measurements were conducted on the four subsets used for testing the various pipelines (Table 4).

Table 4: Basic quality measurements sequenced Sputum and Zymo mock samples with selection of top 10% and 90% reads based on quality

	Sputum 10%	Sputum 90%	Zymo Mock 10%	Zymo Mock 90%
Mean read length	1,661.0	1,580.9	1,645.8	1,560.9
Mean read quality	14.3	12.0	13.1	11.0
Amount of reads	84,905	802,845	11,238	106,645
Median read length	1,610.0	1,604.0	1,608.0	1,596.0
Median read quality	14.7	12.7	13.6	11.7

The mean read length was found to be around the expected length of the 16s rRNA gene. For the sputum samples the mean read lengths were 1,661.0 (10% reads) and 1,580.9 (90% reads), while for the Zymo mock samples, they were 1,645.8 (10% reads) and 1,560.9 (90% reads).

The read quality was measured in terms of Phred score. In the sputum samples, the mean read quality scores were 14.3 (10% reads) and 12.0 (90% reads), whereas in the Zymo mock samples, they were 13.1 (10% reads) and 11.0 (90% reads).

Median read lengths and qualities were also assessed. In sputum samples, median read lengths ranged from 1,610 (10% reads) to 1,604 (90% reads), and in Zymo mock samples, they ranged from 1,608 (10% reads) to 1,596 (90% reads). The corresponding median read quality scores were 14.7 (10% reads) and 12.7 (90% reads) for sputum samples, and 13.6 (10% reads) and 11.7 (90% reads) for Zymo mock samples.For the sputum samples, the number of reads ranged from 84,905 (10% reads) to 802,845 (90% reads), while for the Zymo mock samples, the range was 11,238 (10% reads) to 106,645 (90% reads).

Quality measurements were also performed on the data obtained from the sequencing of four sputum samples from the same patient, collected at different moments (Table 5). Similar to the previous analysis, the mean read length was consistent with the expected length of the 16S rRNA gene, ranging from 1,462.0 to 1,443.5 across the different time points. The mean read quality scores for these sputum samples ranged from 11.2 to 11.3, indicating a relatively consistent sequencing read quality throughout the time course. The total number of reads obtained for each time point was also determined, the values varied between 508,008 and 844,936.

Additionally, the median read length and median read quality were determined for the sputum samples at each time point. The median read lengths ranged from 1,600 to 1,611, while the median read quality scores ranged from 12.1 to 11.9 across the different time points.

Table 5: Basic quality measurements sequenced Sputum samples from 1 patient

	M0	M3	M6	M9
Mean read length	1,462.0	1,346.5	1,443.5	1,403.0
Mean read quality	11.3	11.2	11.3	11.2
Amount of reads	563,555	844,936	508,008	655,185
Median read length	1,600.0	1,595.0	1,603.0	1,611.0
Median read quality	12.1	11.9	12.0	11.9

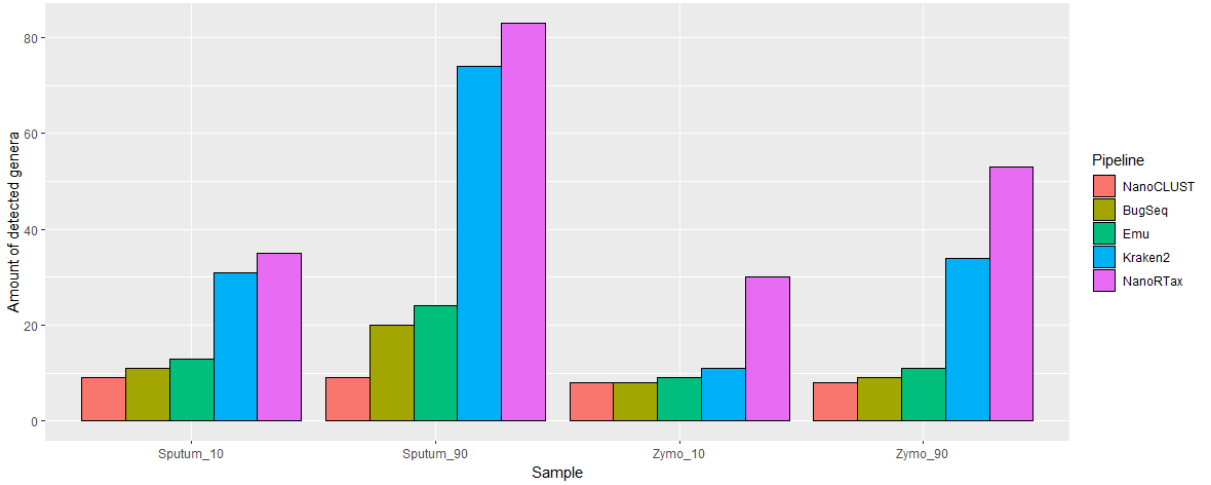

Figure 1: Barplot of amount of unique genera detected by each pipeline in each of the four subsets

Comparison of the detection of unique genera by each pipeline

All five different classifiers were employed to analyze the two subsets of Zymo mock reads and sputum sample reads. The amount of unique genera detected by each of the pipelines in the four different samples was quantified, no cut-off value was used for this analysis. (Figure 1).

Among the pipelines evaluated, NanoCLUST was the only one that consistently detected the same number of genera in both the sputum subsets and the Zymo mock subsets. Specifically, NanoCLUST identified 9 unique genera in the sputum subsets and 8 in the Zymo mock subsets. BugSeq displayed different results for each of the four samples, detecting 11 genera in the 10% sputum subset, 20 genera in the 90% sputum subset, and 8 and 9 genera respectively for the Zymo mock 10% and 90% subsets.

Emu also demonstrated varying results depending on the dataset and subset analyzed. Emu classified the reads in the sputum 10% subset into 13 different genera and identified 24 genera in the sputum 90% subset. In the Zymo mock dataset, Emu identified 9 unique genera in the 10% subset but 11 in the 90% subset.

The pipeline with Kraken2 and Bracken exhibited differential performance depending on the dataset and subset analyzed. In the sputum subset with 10% of the total reads, Kraken2 and Bracken detected 31 unique genera. When the subset percentage was increased to 90% of the reads, the amount of detected genera significantly increased to 74. Similarly, in the Zymo mock dataset, Kraken2/Bracken initially detected 11 unique genera in the 10% subset, but this number increased to 34 when analyzing the 90% subset.

In the sputum dataset, NanoRTax identified 35 unique genera in the 10% subset, while in the 90% subset, it detected 83 different genera. These findings of NanoRTax were further evident in the Zymo mock dataset, where it detected 30 unique genera in the 10% subset and 53 genera in the 90% subset.

Comparative analysis of genus-level taxonomic assignment in Zymo mock community reads subsets

The relative abundance data produced by the pipelines was used to compare the classification performances of each pipeline on the two Zymo mock subsets. If a genus had a relative abundance lower than 0.1%, it was excluded and the corresponding percentage was allocated to the 'other' category (Figure 2).

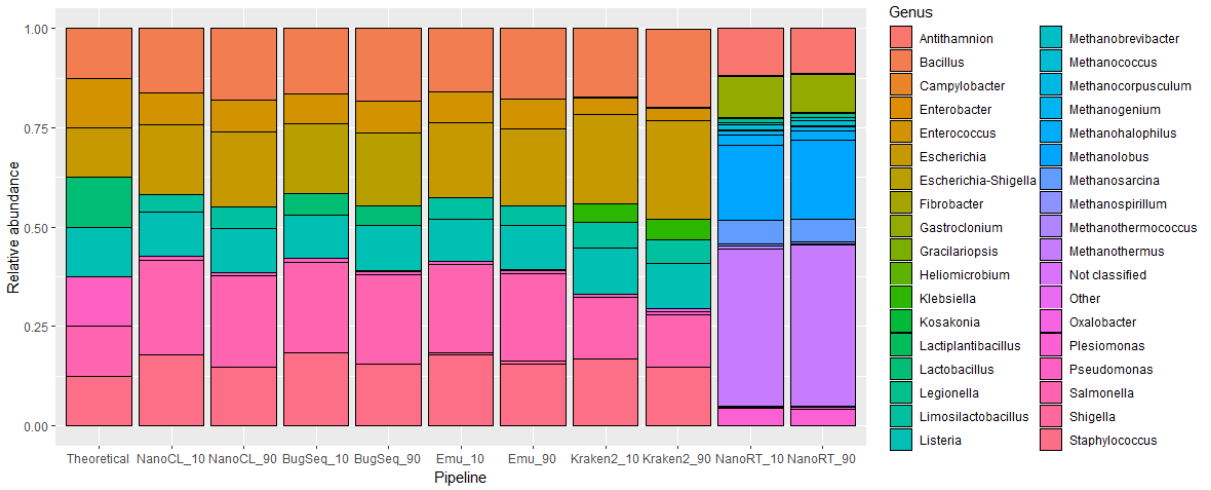

Figure 2: Stacked barplot of genera with more than 0,1% relative abundance distribution in Zymo mock sample subsets across pipelines, with a bar representing the theoretical composition of the Zymo mock sample. Relative abundances of <0.1% allocated to 'other'.

Except for NanoRTax, all pipelines successfully identified the presence of several genera that were part of the theoretical distribution, namely *Bacillus*, *Enterococcus*, *Listeria*, *Pseudomonas*, *Salmonella*, and *Staphylococcus*. These genera were not the only genera present in the Zymo mock sample.

NanoCLUST detected all genera except *Lactobacillus*, but the pipeline identified *Limosilactobacillus* in the subsets which is a genus splitted from *Lactobacillus*. Emu and Kraken2 combined with Bracken also detected *Limosilactobacillus* in the reads but not *Lactobacillus*. BugSeq was the only pipeline that did identify *Lactobacillus* in both subsets.

Although BugSeq did not specifically detect the *Escherichia* genus, it classified over 15% of the reads as *Escherichia-Shigella*. It is worth noting that BugSeq was the only pipeline that has left a small percentage (0.1%) of the reads unclassified.

Emu and Kraken2/Bracken identified some genera in both subsets that were not present in the Zymo mock sample according to the manufacturer's information. Specifically, Emu identified 0.6% of the reads in the 10% subset and 0.9% in the 90% subset as belonging to the *Shigella* genus. Kraken2 and Bracken, identified 4.5% of the reads in the 10% subset and 5.1% in the 90% subset as *Klebsiella*.

NanoRTax did not detect any of the eight genera present in the Zymo mock sample. Instead, it identified 19 other genera with a relative abundance higher than 0.1%. According to the results of NanoRTax the Zymo mock reads mostly (40%) consist of bacterial DNA that belongs to the *Methanothermus* genus.

Comparative analysis of genus-level taxonomic assignment of sputum sample reads subsets

In order to assess the classification performance of each pipeline on the two subsets of sputum samples, the relative abundance data was employed. If a genus had a relative abundance lower than 0.1%, it was excluded and the corresponding percentage was allocated to the 'other' category. (Figure 3).

The sputum sample was spiked with microbial cells from the *Imtechella* and *Allobacillus* genera. NanoCLUST accurately identified *Imtechella* with relative abundances of 2.4% and 2.1% in the 10% and 90% subsets, respectively. It also detected *Allobacillus* with relative abundances of 33.2% and 37.2% in the corresponding subsets. Similarly, the Emu pipeline successfully detected both spike-ins with relative abundances of more than 0.1% in both subsets (34.9% and 37.0% for *Allobacillus*, 2.2% and 2.1% for *Imtechella*). BugSeq did not detect the spike ins at the genus leve but left a significant portion of reads unqualified, accounting for 26.8% in the 10% subset and 33.8% in the 90% subset.

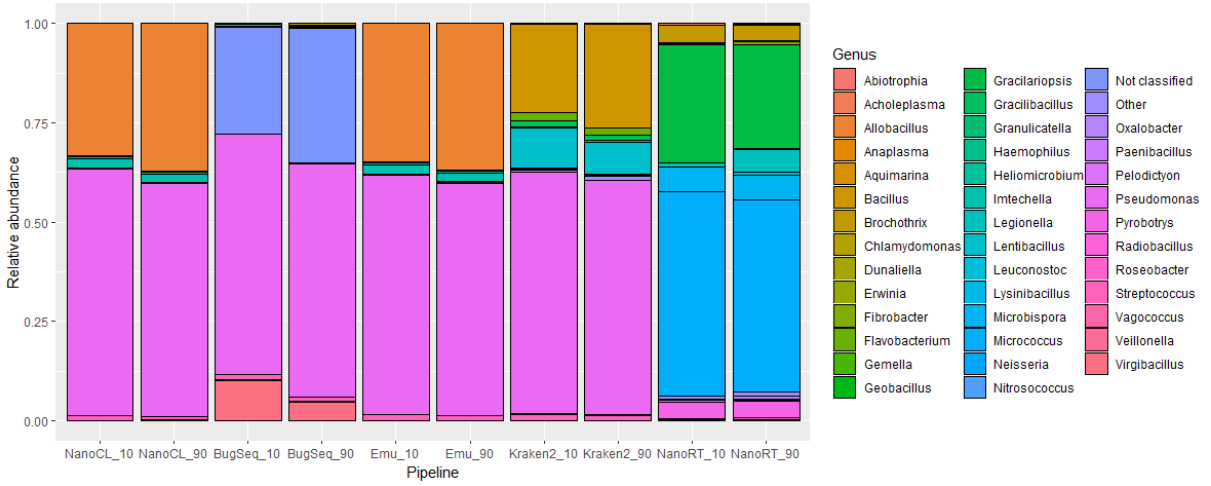

Figure 3: Stacked barplot of genera with more than 0,1% relative abundance distribution in sputum sample subsets across pipelines. Relative abundances of <0.1% allocated to 'other'.

Consistent detection of *Haemophilus*, *Neisseria*, *Streptococcus*, and *Veillonella* genera was achieved by all pipelines, except for NanoRTax. These genera were identified in both subsets by the pipelines at small relative abundances ranging from 0.2% to 1.4%. These findings align partially with the results obtained from culturing, as both Streptococcus and Haemophilus were identified.

The *Pseudomonas* genus was detected by all pipelines except NanoRTax, with a relative abundance of more than 58%. *Granulicatella* reads were identified by NanoCLUST, BugSeq, and Emu pipelines, albeit at small percentages ranging from 0.18% to 2.4% relative abundance.

The Kraken2 and Bracken pipeline classified reads into seven genera that remained undetected by any other pipeline. The NanoRTax pipeline identified 20 genera with relative abundances of more than 0.1% that were not detected by the other pipelines.

Computational performances

In order to compare the computational performances of the three command-line pipelines the data of the analysis of the sputum sample subset with 10% of the best reads was used (Table 6).

NanoCLUST processed the dataset in 13 minutes and 56 seconds, utilizing 1 hour and 42 minutes of computational resources. NanoRTax, demonstrated faster performance with a runtime of 1 minute and 8 seconds. The CPU hours for NanoRTax were described as "a few seconds," by the pipeline implying that its resource consumption was minimal.

Emu exhibited the longest runtime among the three command-line pipelines, requiring 3 hours, 50 minutes, and 43 seconds to complete the classification process. The corresponding CPU hours for Emu were 44 minutes and 33 seconds.

The installation complexity of the pipelines was also considered, Emu stood out as the most efficient in terms of installation, largely attributed to its clear and well-documented installation process. In the case of NanoRTax, the installation itself was not particularly complicated. However, it required the prior installation of Conda and Nextflow, which added an additional step before running the pipeline. NanoCLUST also required prior installation of Nextflow and Docker but encountered challenges due to errors in the scripts. As a result, considerable time was spent troubleshooting and resolving these issues before successfully setting up the pipeline.

BugSeq was utilized on its dedicated online platform, which, does not provide specific computational performance metrics as part of the output. The platform operates using a queuing system, meaning the time it took from data upload to result retrieval can vary significantly depending on the platform's current workload. The BugSeq website provided a user-friendly interface that facilitated easy navigation and usage and has the opportunity to easily let the user visualize the taxonomic classification results.

Kraken 2 and Bracken were used within the Galaxy platform, similar to the BugSeq platform, the runtime of these tools within the Galaxy environment is also influenced by the current server workload. An advantage of the Galaxy platform was its user-friendly nature, enabling convenient tool execution with some prior knowledge.

*Table 6: Computational performances of the three command-line pipelines, * indicates a difficult installation process, ** indicates an average installation process and *** indicates an easy installation process*

	NanoCLUST	NanoRTax	Emu
Run time (HH:MM:SS)	00:13:56	00:01:08	03:50:43
CPU time (HH:MM:SS)	01:42:00	'a few seconds'	00:44:33
Installation	*	**	***

Visualizing temporal dynamics in the sputum microbiome of one patient

Over a period of 9 months a sputum sample was taken from a patient once every 3 months and sequenced, the relative abundance of genera in the 16s rRNA gene reads was analyzed with NanoCLUST (Figure 4). No cut-off value for the relative abundance was used during this analysis.

The relative abundance of *Haemophilus* genera exhibited a decline over the course of the monitoring period. It started at 54% and gradually decreased to 42% by the final month. On the other hand, the relative abundance of *Streptococcus* genus displayed dynamic fluctuations throughout the entire monitoring period. It began at 42% and experienced an increase to 76% within the first 3 months. It remained at this level until month 6 before dropping significantly to 41% in month 9.

The percentage of *Veillonella* reads initially decreased in the first 3 months, but then showed an increase from month 3 to month 6, followed by another increase afterwards. Similarly, the percentage of *Gemella* reads increased during the first 3 months, but decreased from month 3 to 6, only to increase again in the last 3 months of the monitoring period. In contrast, the percentage of *Granulicatella* reads demonstrated a consistent increase throughout the entire monitoring period. It started at 1.1% in month 0 and gradually reached 6.6% by the final month. *Lautropia* showed an increase in relative abundance over a period of 6 months. Although it was initially absent in month 0, its relative abundance increased to 0.4% in the first 3 months. It further increased to 0.5% at month 6 but was absent again in month 9.

Each sputum sample contained reads that were assigned to genera that were not detected in any other sample. The sample taken at month 0 contained *Campylobacter*, *Dialister* and *Parvimonas* while those genera were not detected in the other samples. The sample taken 3 months later is the only sample that contained some reads that could not be classified, and also was the only one were *Tenuibacillus* was detected. Reads from the sputum sample from month 6 uniquely contained *Cardiobacterium*, *Fusobacterium*, Kingella and Rothia. The sputum collected at the last month of the monitoring period contained reads assigned to *Bacillus*, *Priestia* and *Virgibacillus* exclusively.

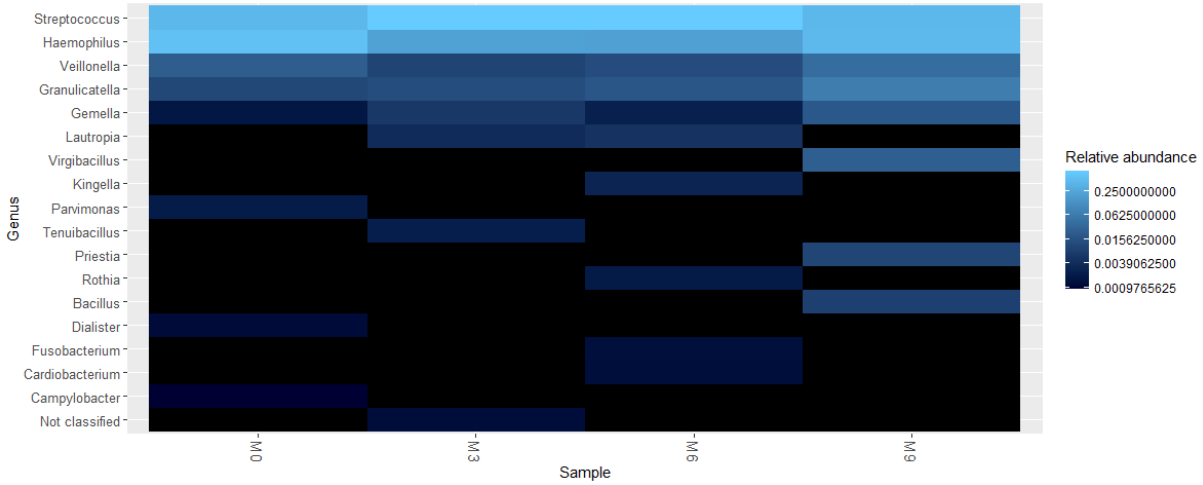

Figure 4: Temporal abundance heatmap of detected genera in sputum samples from 1 patient collected at 4 different moments

The taxonomic classification results align with the culturing results obtained from the same sputum samples. NanoCLUST identified each of the genera detected by culturing in the samples. For instance, in month 6 when only Gram-positive cocci were detected in the culture, NanoCLUST detected Streptococcus, a Gram-positive cocci. Similarly, Haemophilus, which was detected by culturing in months 3 and 9, was identified by NanoCLUST across all four months. NanoCLUST revealed the presence of several genera that were not detected by culturing the sputum samples.

Discussion

The impact of discontinuing azithromycin treatment on exacerbations and patient well-being in COPD patients has prompted investigations into the potential changes in the COPD-associated microbiome using 16S rRNA gene sequencing, considering concerns related to antibiotic resistance development and potential side effects. This paper aims to address two key objectives. Firstly, to assess the performance of three pipelines and two tools for taxonomic classification at the genus level using 16s rRNA gene reads derived from sputum samples from COPD patients. Subsets containing different percentages of the best reads of one sputum sample and one Zymo mock sample were used to evaluate the performances of NanoCLUST, NanoRTax, BugSeq, Emu and Kraken2 in combination with Bracken. The ideal pipeline or tool should accurately detect all genera in Zymo mock samples without false positives and classify spike-ins in sputum correctly. Additionally the pipeline or tool should provide ease of use.

We found that NanoCLUST was the only one of all five classifiers that correctly classified all genera in the Zymo mock subsets together with the two spike-ins in the sputum sample subsets. NanoCLUST demonstrated accurate classification by not assigning reads to any genus that was not included in the Zymo mock used for the analysis. Additionally, NanoCLUST effectively detected all genera in the sputum samples that were also identified during the culturing process. In general, NanoCLUST outperformed the other classifiers.

Several previous studies have also shown that NanoCLUST outperforms other 16s rRNA classification programs. In their research paper, the developers of the NanoCLUST pipeline highlighted that it accurately identified the expected number of species in two mock communities, while Kraken2 combined with Bracken identified a larger number of species. The authors emphasized that NanoCLUST excels in terms of identification accuracy and estimation of abundance profiles, surpassing state-of-the-art software.[17] A study conducted by A.S.G Borges *et al.* demonstrated the effective utilization of NanoCLUST in combination with MinION sequencing for species and genera identification in both pure cultures and complex samples.[22] Another study conducted by A.W.T. Lee *et al.* compared the

performance of several tools including MegaBLAST, ARGpore2, Emu, Kraken2/Bracken, and NanoCLUST. According to their findings, NanoCLUST was preferred for 16S microbial profiling due to its similar clustering results with MegaBLAST. Additionally, Kraken2/Bracken exhibited comparable clustering results, while Emu demonstrated the highest accuracy rate and F1 score.[19]

Comparison of pipeline performance in this study revealed comparable results up to genus level, although discrepancies were observed in bacterial proportions and identified genera. Several factors contribute to these differences. Primarily, the utilization of different databases across the pipelines and tools significantly impacts performance for various reasons. Databases may vary in their taxonomic coverage, with some encompassing a wider range of known genera.[23] Variations exist in the reference sequences incorporated within the databases, which are crucial for aligning and comparing the sequenced reads to enable accurate taxonomic classification. Additionally, 16S rRNA gene sequences exhibit variability within and between genera, and databases employ different clustering or sequence alignment methods to handle this variability.[23] [24]

The observed differences in the classification of *Lactobacillus* and *Limosilactobacillus* among the various pipelines and tools can be attributed to the inherent complexity of the *Lactobacillus* genus. With 2016 species (as of March 2020), *Lactobacillus* exhibits significant phenotypic, ecological and genotypic diversity. In a study by Jinshui Zeng *et al*., a reclassification of the *Lactobacillus* genus was proposed into 25 distinct genera, including *Limosilactobacillus*. [25] It is possible that certain databases have already implemented these new genera, while others may not have incorporated them yet.

NanoCLUST has a notable limitation because it requires manual script modifications for its functionality. This customization affects the output of NanoCLUST, the relative abundance tables contain taxids instead of taxonomical names for certain detected genera. If the database used by the pipeline lacks comprehensive information for certain taxonomic levels, NanoCLusT will only provide the taxid for that entry. By consulting the NCBI database, one can easily retrieve the corresponding taxonomical names and replace the taxids in the output. [17]

The inability of BugSeq to identify the spike-ins *Imtechella* and *Allobacillus* at the genus level could be attributed to the resolution of the reference database utilized. In order to ensure accurate taxonomic assignments, BugSeq prioritizes assigning bacteria to a higher taxonomic level when the genus-level information is insufficient or uncertain.[26]

The variations observed in taxonomic classification between Emu and Kraken2 can be ascribed to the algorithms employed. Emu functions as a tool that operates on the Kraken2 database, which encompasses a collection of reference sequences and associated taxonomic information. The discrepancies encountered in taxonomic classification between Emu and Kraken2 arise due to the distinct algorithms utilized by each tool and their respective utilization of the Kraken 2 taxonomic database. Emu primarily focuses on estimating abundance at a single taxonomic level, whereas Kraken 2 provides comprehensive taxonomic assignments spanning multiple levels within the taxonomic hierarchy. [19]

NanoRTax failed to identify any of the genera found in both the Zymo mock sample subsets and the spike-ins in the sputum sample. Furthermore, it inaccurately classified genera that were confirmed to be present in the sputum through culturing. Although the pipeline provided a web-app for visualizing the estimated relative abundance, the associated scripts proved to be non-functional. Apart from the paper authored by the pipeline developers, no other studies have utilized or evaluated the performance of NanoRTax.[18] The authors themselves acknowledge the necessity for additional experimental demonstrations to validate its efficacy.

The comparison of different pipelines and tools in this study is subject to certain limitations in terms of computational performance. NanoCLUST, NanoRTax, and Emu were utilized within a virtual machine with known specifications. BugSeq was employed on its own website, which lacks information about the computational resources employed. Kraken2 and Bracken were utilized on the Galaxy platform, making it difficult to directly compare them to the command-line execution on the virtual machine. Additionally, since Emu and Kraken2/Bracken are tools rather than pipelines, we implemented a trimming step prior to running these tools to align them more closely with the pipelines that encompass quality control and trimming steps. However, the computational resources and runtime associated with the trimming step were not factored into the overall runtime and computational resource analysis of Emu itself.

The subsets created for comparing the pipelines consisted of 10% and 90% of the total reads in the original sample. However, it is important to note that these subsets were not randomly selected reads. Only the best 10% and 90% of the total reads were included, meaning that at least the poorest 10% of reads were excluded from both subsets. As a result, the 10% subsets exhibited higher mean and median quality scores compared to the 90% subsets. Our findings indicate that there is minimal difference in the detection of genera between the two subsets of the same sample, while the estimated relative abundance does exhibit variation between the subsets. Our QC data shows that the read length is as it is to be expected with 16s rRNA gene sequencing, and the quality scores are good for Nanopore data.

Culturing results revealed that certain bacteria could not yet be identified at the genus level and were instead characterized based on their morphology and Gram stain results. This poses a challenge in accurately evaluating the performance of pipelines that were able to classify bacteria at the genus level.

The second main goal of this study is to identify a visualization strategy that facilitates the comparison of microbiome data obtained from consecutive sputum samples across different patient groups.

By implementing a visualization strategy in the form of a heatmap, researchers will be able to visually compare and analyze the dynamics of the microbial community in relation to treatment cessation and continuation. The heatmap will enable the identification of patterns, trends, and potential shifts in the abundance of specific genera over time, facilitating a deeper understanding of the impact of treatment on the lung microbiome in patients with COPD.

When assessing differences in microbiomes, alpha and beta diversity metrics are commonly used. Alpha diversity metrics measure the diversity within a sample and can be compared between groups, while beta diversity measures the diversity between samples by analyzing the dissimilarity of features and generating a distance matrix for all sample pairs. However, analyzing microbiome datasets presents challenges due to their high-dimensionality, zero-inflation and compositional nature. To tackle these challenges, researchers often employ dimension-reduction-based ordination methods such as principal coordinate analysis (PCoA) or principal component analysis (PCA) to visualize the data a represent sample distances in two or three dimensions. Statistical evaluations, such as ANOSIM, PERMANOVA, or ANCOM, are then applied to assess the biological significance of clustering.

In the context of comparing lower-level taxonomic differences in microbiome data, it is essential to employ appropriate statistical approaches that account for the large number of variables and mitigate the risk of false positives. We recommend utilizing Linear discriminant analysis of effect sizes (LEfSe), a widely adopted method specifically developed for microbiome data analysis. LEfSe incorporates a Kruskal-Wallis test, followed by subsequent Wilcoxon rank-sum tests on subgroups, to identify taxonomic differences between groups. By considering both the effect size and statistical significance, LEfSe provides a more robust and accurate assessment of taxonomic differences. This approach has been successfully applied in various microbiome studies, making it a reliable and recommended tool for comparing lower-level taxonomic differences in microbiome research. [28][29]

This research contributes to advancing our understanding of COPD pathogenesis and aids in selecting the most suitable classification strategy for 16S rRNA gene reads for microbiome studies. NanoCLUST emerges as a robust pipeline for accurate taxonomic classification, and further investigation can focus on refining its functionalities. Accurate classification of the microbiome data from sputum samples is crucial to identify and track specific bacterial genera associated with COPD and azithromycin treatment. This allows for a deeper understanding of the microbial dynamics and potential shifts in the lung microbiome that may contribute to exacerbations and disease progression. Additionally, the visualization of microbiome data obtained from consecutive sputum samples across different patient groups plays a crucial role in interpreting and comparing the results. By employing the appropriate visualization strategy, researchers can effectively analyze and identify patterns or trends in the microbiome composition over time, facilitating the investigation of the effect of ceasing azithromycin treatment in COPD patients.

Acknowledgements

We would like to express our gratitude to Amphia and Microvida for generously providing the sputum samples used in this study. Special thanks are extended to Bazante Sanders, Sander Boden, and Donny Vrins for their support and guidance during the bioinformatic analysis process. Additionally, we would like to acknowledge Liuwei Su for providing the necessary modified script to enable the successful implementation of NanoCLUST in this study. Their contributions and assistance were instrumental in the successful completion of this research.

Literature

1. Djamin, R. S. *et al*. Prevalence and abundance of selected genes conferring macrolide resistance genes in COPD patients during maintenance treatment with azithromycin. *Antibicrobial Resistance & Infection Control* **9** (2020).

2. Ditz, B. *et al*. Sputum microbiome profiling in COPD: beyond singular pathogen detection. *Thorax* **75**, 338-344 (2020).

3. Raherison, C. & Girodet, P. -. Epidemiology of COPD. *European Respiratory Review* **18**, 213-221 (2009).

4. Decramer, M., Janssens, W. & Miravitlles, M. Chronic obstructive pulmonary disease. *Lancet* **379**, 1341-1351 (2012).

5. MacLeod, M. *et al*. Chronic obstructive pulmonary disease exacerbation fundamentals: Diagnosis, treatment, prevention and disease impact. *Respirology* **26**, 532-551 (2021).

6. Yang, J. Mechanism of azithromycin in airway diseases. *J Int Med Res* **48**, 0300060520932104 (2020).

7. Albert, R. K. *et al*. Azithromycin for Prevention of Exacerbations of COPD. *N. Engl. J. Med.* **365**, 689-698 (2011).

8. Heidary, M. *et al*. Mechanism of action, resistance, synergism, and clinical implications of azithromycin. *J Clin Lab Anal* **36**, e24427 (2022).

9. Herath, S. C., Normansell, R., Maisey, S. & Poole, P. Prophylactic antibiotic therapy for chronic obstructive pulmonary disease (COPD). *Cochrane Database Syst Rev* **2018**, CD009764 (2018).

10. Santos, A., van Aerle, R., Barrientos, L. & Martinez-Urtaza, J. Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Computational and Structural Biotechnology Journal* **18**, 296-305 (2020).

11. Rozas, M., Brillet, F., Callewaert, C. & Paetzold, B. MinION™ Nanopore Sequencing of Skin Microbiome 16S and 16S-23S rRNA Gene Amplicons. *Frontiers in Cellular and Infection Microbiology* **11**, 1317 (2022).

12. Benítez-Páez, A., Portune, K. J. & Sanz, Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *Gigascience* **5**, 4 (2016).

13. Han, M. K. *et al*. Significance of the microbiome in obstructive lung disease. *Thorax* **67**, 456-463 (2012).

14. Pragman, A. A., Kim, H. B., Reilly, C. S., Wendt, C. & Isaacson, R. E. The Lung Microbiome in Moderate and Severe Chronic Obstructive Pulmonary Disease. *PLoS One* **7**, e47305 (2012).

15. Yagi, K., Huffnagle, G. B., Lukacs, N. W. & Asai, N. The Lung Microbiome during Health and Disease. *Int J Mol Sci* **22**, 10872 (2021).

16. Fan, J., Huang, S. & Chorlton, S. D. BugSeq: a highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinformatics* **22**, 160 (2021).

17. Rodríguez-Pérez, H., Ciuffreda, L. & Flores, C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics* **37**, 1600-1601 (2021).

18. Rodríguez-Pérez, H., Ciuffreda, L. & Flores, C. NanoRTax, a real-time pipeline for taxonomic and diversity analysis of nanopore 16S rRNA amplicon sequencing data. *Computational and Structural Biotechnology Journal* **20**, 5350-5354 (2022).

19. Curry, K. D. *et al*. Emu: Species-Level Microbial Community Profiling for Full-Length Nanopore 16S Reads. *bioRxiv*, 2021.05.02.442339 (2021).

20. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

21. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* **3** (2017).

22. Borges, A. S. G. *et al*. Fast Identification Method for Screening Bacteria from Faecal Samples Using Oxford Nanopore Technologies MinION Sequencing. *Curr Microbiol* **80**, 101 (2023).

23. Somervuo, P. *et al*. Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods in Ecology and Evolution* **8**, 398-407 (2017).

24. Hakovirta, J. R., Prezioso, S., Hodge, D., Pillai, S. P. & Weigel, L. M. Identification and analysis of informative single nucleotide polymorphisms in 16S rRNA gene sequences of the Bacillus cereus group. *J. Clin. Microbiol.* **54**, 2749-2756 (2016).

25. Zheng, J. *et al*. A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. *Int. J. Syst. Evol. Microbiol.* 70, 2782-2858 (2020).

26. Ceuppens, S., De Coninck, D., Bottledoorn, N., Van Nieuwerburgh, F. & Uyttendaele, M. Microbial community profiling of fresh basil and pitfalls in taxonomic assignment of enterobacterial pathogenic species based upon 16S rRNA amplicon sequencing. *Int. J. Food Microbiol.* 257, 148-156 (2017).

27. Olman, V., Mao, F., Wu, H. & Xu, Y. Parallel clustering algorithm for large data sets with applications in bioinformatics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 6, 344-352 (2008).

28. Galloway-Peña, J. & Hanson, B. Tools for analysis of the microbiome. *Dig. Dis. Sci.* 65, 674-685 (2020).

29. Cadena, A. M. *et al*. Profiling the airway in the macaque model of tuberculosis reveals variable microbial dysbiosis and alteration of community structure. *Microbiome* 6, 1-12 (2018).

Appendix

Flowchart

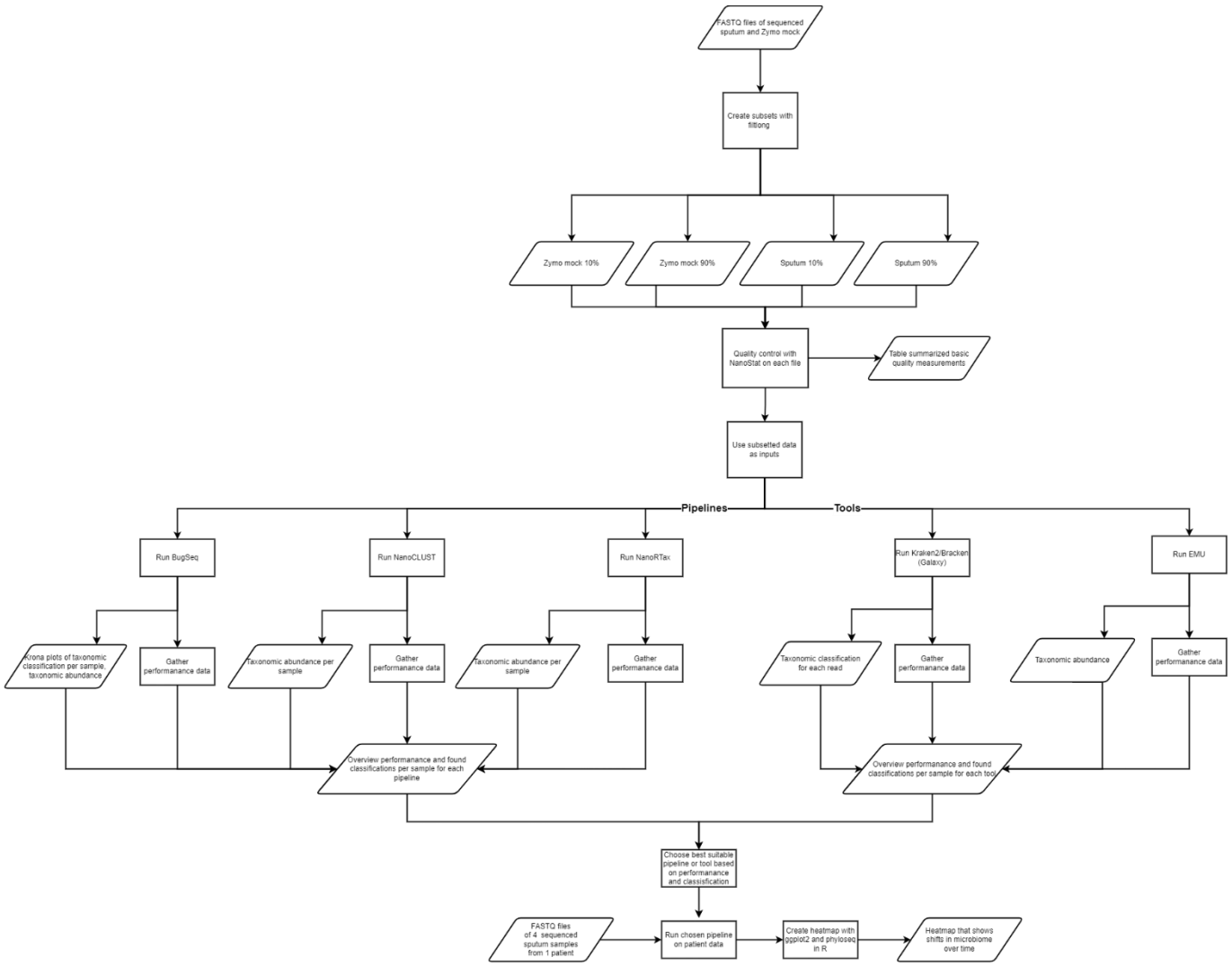

Figure 5: Flowchart of the process of testing various pipelines and using the best one on the data of 1 patient