# Datenanalyse - Laborprotokoll (Gruppe 1b)

Jan Aspan, Felix Bachmann, Niklas Grundner, Julian Resch, Sebastian Sarria Suarez

28. April 2025

## Table of contents

# 1 Abstract

Oxford Nanopore sequencing enables direct, long-read analysis of complex metagenomic samples. In this study, we systematically compared two data processing workflows for taxonomic classification, starting from DNA samples of known origin: an oral swab, a room monitoring sample, a single cultured microorganism, and the ZymoBIOMICS HMW DNA standard (ZymoResearch Cat.# D6322). We evaluated a minimal workflow with basic quality control and no filtering against an extended workflow including expanded quality control and additional manual filtering. Assemblies were performed using `Raven`, `MetaFlye`, and `Canu`, the taxonomic classification via `Kraken2` and visualization using `Krona`. Our results demonstrate that comprehensive quality control significantly improves assembly accuracy and taxonomic resolution. Among assemblers, `Raven` consistently yielded the most representative results across all sample types. We recommend a workflow combining thorough quality filtering, `Raven` assembly, and `Kraken2` classification for reliable and efficient nanopore-based metagenomic analysis of diverse sample origins.

# 2 Background

## 2.1 Oxford Nanopore Sequencing

Nanopore sequencing (often referred to as Oxford Nanopore Sequencing or ONT sequencing) is a single-molecule technology in which individual strands of DNA or RNA are pulled through a biological nanopore channel in an electrically insulating membrane. An applied voltage (~180 mV) generates a constant ionic base current; as soon as a nucleic acid segment passes through the narrowest part of the pore, the combination of about five consecutive bases located there changes the electrical resistance. These characteristic current fluctuations (‘squiggles’) are digitised at a kHz sampling rate and then translated into nucleotide sequences using deep neuronal-based base callers (e.g. guppy flip-flop RNN or Dorado). An upstream motor protein - a modified helicase/polymerase - regulates the translocation speed (70 - 450 bp s-$^{1}$ with current R9.4 chemistry) and removes double strands. The method works purely electrically, does not require fluorescent dyes or PCR steps and therefore enables extremely long reads (record-verified > 2 Mbp), real-time data flow and the immediate detection of epigenetic base modifications and native RNA. Today’s commonly used por generations (R9.4, R10.3) achieve raw read error rates of 5 - 10 % and consensus accuracies > 99.9 % after polishing.[1][2]

### Metrics for performance evaluation

Current benchmarks essentially use four key figures, partly already defined in the Wick paper, now formalised:[3][4]

1. read accuracy: median sequence identity of individual reads against a reference, usually as Phred-Q (-10 log  error) or “modal accuracy”.
2. consensus accuracy: Identity of a consensus generated from many reads; important for assemblies/variant calling.
3. Failure rate: Percentage of raw streams that cannot be decoded by the base caller.
4. throughput / speed: Bases per second on defined hardware (GPU/CPU). In addition, homopolymer errors, modified-base detection F1 score (for Remora/Dorado) and computational cost (watt-hours per Gb) are shown in more recent work.

## 2.2 Metagenom

A metagenome is the entirety of all genetic sequences obtained directly from a complex sample (such as soil, seawater or the human gut). Instead of first isolating individual microorganisms in pure culture, the entire DNA or RNA mixture is extracted and sequenced; this provides an overview of all members of the community and their genetic potential in a single step.[5]

Metagenomes are being analysed because they answer questions that can hardly or not at all be solved with classical culture methods:

• Uncovering hidden biodiversity: It is estimated that more than 90 % of the microbial world cannot (or can only with difficulty) be cultivated; metagenomic approaches open up this ‘dark matter’ of the microbiome.[6]

• Understanding ecological functions: By analysing the enzymes encoded in the metagenome, it is possible to reconstruct which metabolic pathways take place in a habitat and how microbes influence global cycles (carbon, nitrogen, etc.).[7]

- Medical diagnostics and microbiome research: In the human body, metagenomics helps to detect pathogens or resistance genes without prior knowledge and to uncover connections between dysbioses and diseases.[8]

- Biotechnological treasure hunt: Novel enzymes and bioactive substances originate from environmental metagenomes, which can be used in industry, environmental technology or pharmaceuticals.[9]

- Environmental and process monitoring: Sequence-based monitoring of wastewater, fermenters or drinking water systems enables the early detection of contamination or process disruptions.[10]

- Evolutionary insights: Time series or sediment samples show how microbial communities genetically adapt to climate and environmental changes.[11][12]

## 2.3 Taxonomic classification

Taxonomic classification is the computer-aided step in which sequence reads or longer contigs of a biological sample are assigned to a hierarchical category of life (i.e. species, genus, family, etc.). In practice, classification follows directly after sequencing and answers the question: 'Which organisms (or genes) are in my data set?'[13]

Read-based classification assigns each raw read directly to a reference database. This is fast and can be applied even with low coverage.[13]

Assembly-based classification first assembles reads into contigs or whole genomes; this increases accuracy and enables functional annotation, but requires higher computing power as a result.[13]

Typical tools such as Centrifuge, Kraken 2 (k-mer-based) or Kaiju (protein-level) are available on many bioinformatics servers. The choice of database and parameters influences the sensitivity and precision of the classification.[13]

Taxonomic classification is important for the following points:

- Recording biodiversity: Only taxonomic classification turns an anonymous sequence collection into an ecological 'species list' and shows which taxa are present or absent in a habitat.[14]

- Clinical diagnostics & epidemiology: Rapid detection of pathogens or resistance genes in patient samples supports therapeutic decisions and outbreak monitoring.[15]

- Food and environmental monitoring: Detection of undesirable microorganisms in production chains or waters is often only possible via metagenomic classification.[16]

- Functional interpretation: Many bioinformatics pipelines combine taxa information with gene annotation steps; this makes it possible to deduce which groups of organisms encode certain metabolic pathways or toxins.[17]

- Bias control & quality check: By comparing expected and observed taxa, contamination or gaps in reference databases can be recognised; false positives/negatives can be quantified.[13]

# 3 Methods

## 3.1 Data processing pipeline

To gain both meaningful result and insight into the effect of quality control and filtering, two slightly different workflows were defined for data analysis.
One with minimal quality control and no filtering and a second one with further quality control and filtering.

**Workflow 1 - minimal QC and no filtering**

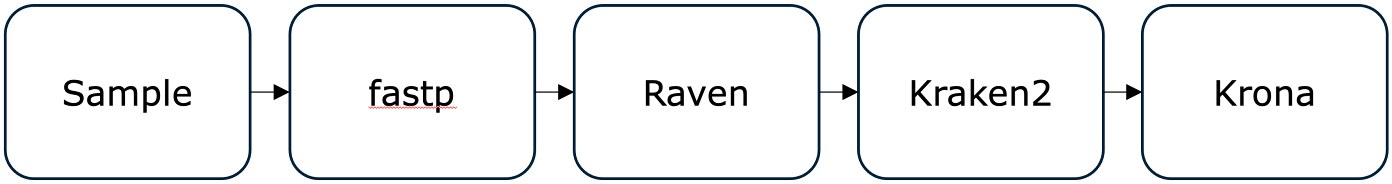


Figure 1: Workflow 1: From Sample to qc, then to assembly, classification and visualization

In this minimal workflow the received data was preprocessed with fastp which also has minimal qc functionality integrated (Version: 0.24.0, https://github.com/OpenGene/fastp).
The output from fastp was then assembled using Raven (**Raven** is a genome assembler designed for error-prone reads like oxford nanopore reads. Version: 1.8.3, https://github.com/lbcb-sci/raven).
For taxonomic classification, Kaken2 was used and Krona to create the visuals. (**Kraken2** is a taxonomic classification database used to assign reads to taxa. Version: 2.1.3, https://github.com/DerrickWood/kraken2) (**Krona** is used to visualize the result as interactive html plots, Version: 2.8.1, https://github.com/marbl/Krona/wiki)

**Workflow 2 - extended QC and filtering**

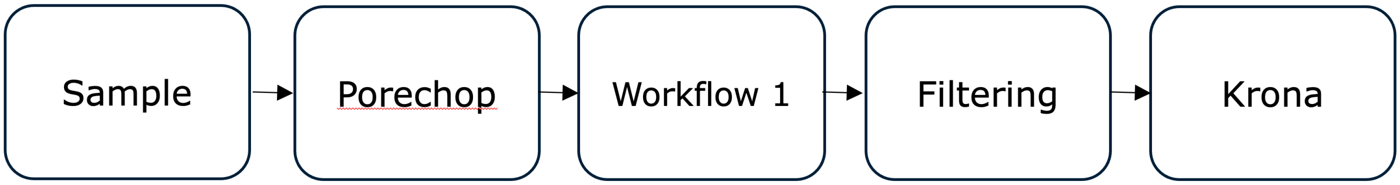


Figure 2: Workflow 2: extended qc with porechop, then fastp, assembly and classification and manual filtering before visualization

The main steps of workflow 2 (preprocessing -> assemble -> classification) are the same as in worklow 1, however in this workflow porechop was used before fastp to filter adapters. (**Porechop** is used for the removal of sequencing adapters and improving the accuracy of downstream analyses. Version: 0.5.1, https://github.com/bonsai-team/Porechop_ABI).
After classification the results were manually filtered using this command: `awk -F'\t' '\$2 \> 5' report.txt \> filtered_report.txt`

As not every assembler works with every sample, MetaFlye and Canu were used as alternative for some samples (**MetaFlye** is a de novo assembler based on Flye and optimized for metagenomic data and is ideally for oxford nanopore reads. Version: 2.1.3, https://github.com/fenderglass/Flye, **Canu** is specifically designed for long-read data from Oxford Nanopore, Version: 2.3, Parameters: `canu -p canu_assembly -d genomeSize=10m maxThreads=8 useGrid=false -nanopore` , https://github.com/marbl/canu).

MetaQuast was used to compare the quality of metagenome asselblies from different asselblers, specially for sample 4. (Version: 5.2.0, https://github.com/ablab/quast)

# 4 Results

## 4.1 Sample 1

Sample 1 taxonomy classification with Kraken 2 was performed with different approaches to the sequence assembly. For each approach polishing and filtering steps with Porechop (adapter trimming), Fastp (discard nucleotides <10 phred score) and a manual awk filtering step were performed. After the preparation steps assembly was performed with Metaflye, Raven and a classification with no assembly at all. Since Raven assemblies are the most representative assemblies for our samples the results of Raven and the classification without an assembly are shown here (the Metaflye assembly data shows very similar results, but Raven results are the most representative for the four given samples and the results of sample 1 with no assembly are, unexpectedly, also very similar and therefore shown here).

Quality control with `FastQC` revealed 1460282 sequences (GC 42%), with low base quality in short reads. After adapter and quality filtering using `Porechop` and `Fastp`, 1384890 high-quality reads (GC 42%) remained.

Krona output of workflow with Raven assembly of sample 1:

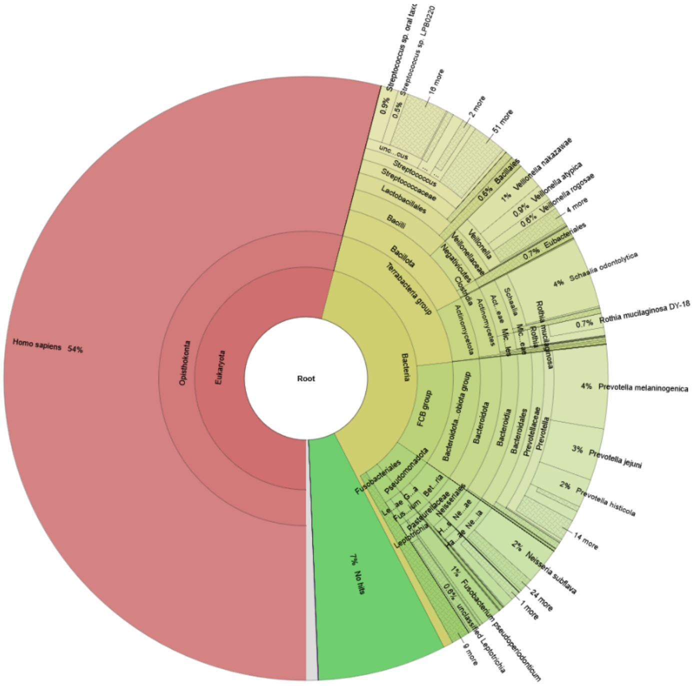


Figure 3: Krona chart showing taxonomic composition of sample 1 assembled via Raven

Krona output of workflow with no assembly of sample 1:

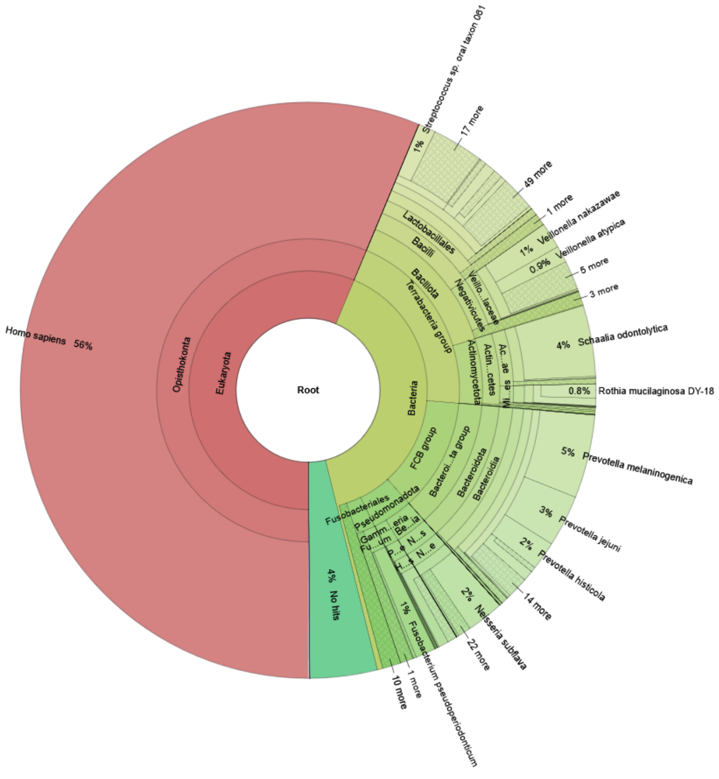


Figure 4: Krona chart showing taxonomic composition of sample 1 alternative workflow (no assembly)

Table 1: Comparison of Top 8 relativ abundand organisms from Krona classificiation output

| Taxon (Top 8) | Relative abundance (Raven) [%] | Relative abundance (no assembly) [%] |
|---|---|---|
| Homo sapiens | 54 | 56 |
| Prevotella melaninogenica | 4 | 5 |
| Schaalia odontolytica | 4 | 4 |
| Prevotella jejuni | 3 | 3 |
| Prevotella histicola | 2 | 2 |
| Neisseria subflava | 2 | 2 |
| Fusobacterium pseudoperiodonticum | 1 | 1 |
| Veillonella nakazawae | 1 | 1 |

## 4.2 Sample 2

Sample 2 analysis was performed with the same polishing and filtering steps as sample 1 (Porechop (adapter trimming) Fastp (discard nucleotides <10 phred score) and a manual awk filtering. Here too an assembly with Metaflye and Raven was performed. The results showed very similar numbers but Metaflye outputs are not shown in this report as Raven stays the most representative assembler for our samples.

Quality control with `FastQC` revealed 281153 sequences (GC 41%), with low base quality in short reads. After adapter and quality filtering using `Porechop` and `Fastp`, 254349 high-quality reads (GC 41%) remained.

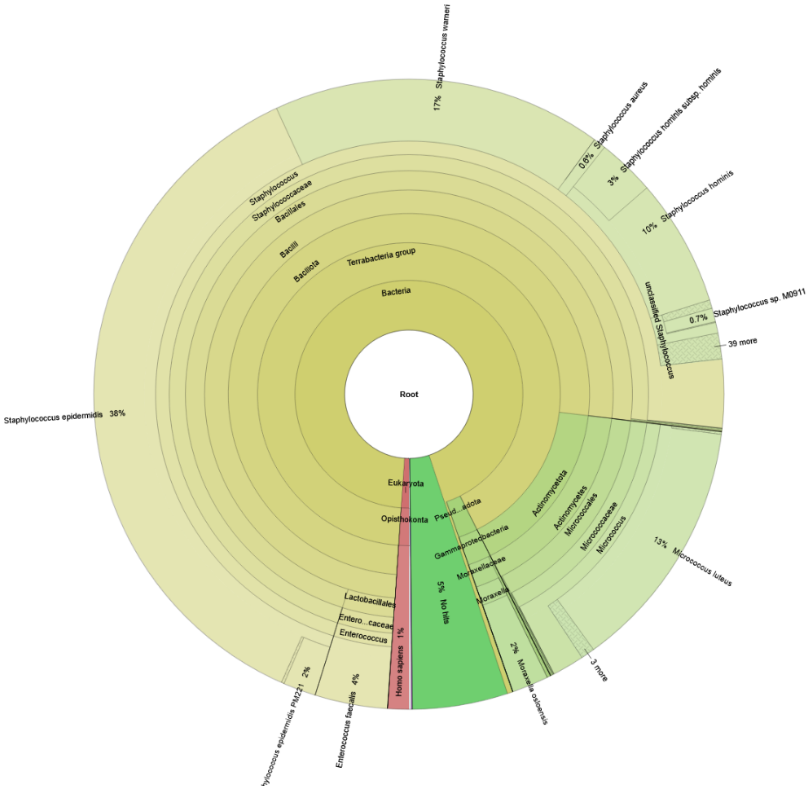


Figure 5: Krona chart showing taxonomic composition of sample 2 assembled via Raven

Table 2: Top 8 relativ abundand organisms from Krona classificiation output

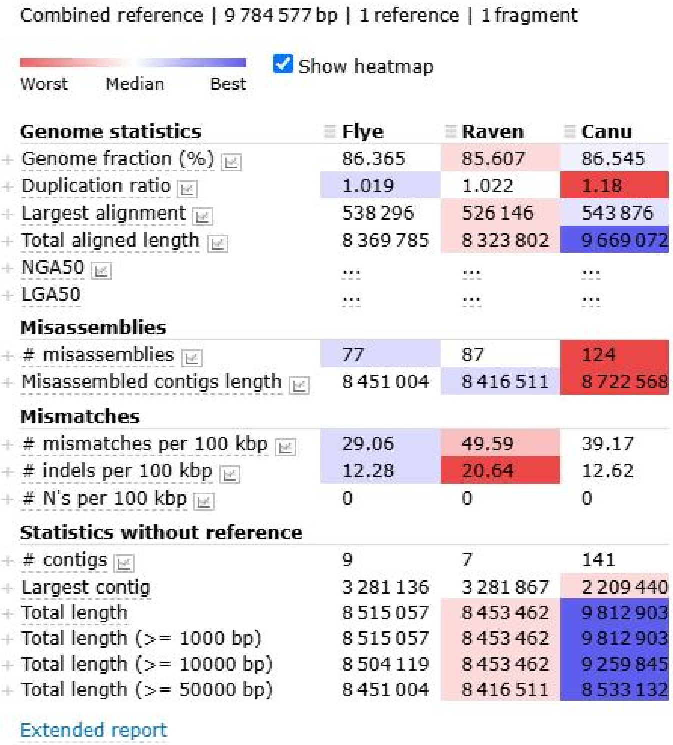
| Taxon (Top 8) | Relative abundance [%] |
|---|---|
| Staphylococcus epidermidis | 38 |
| Staphylococcus warneri | 17 |
| Micrococcus luteus | 13 |
| Staphylococcus hominis | 10 |
| Enterococcus faecalis | 4 |
| Staphylococcus hominis subsp. hominis | 3 |
| Moraxella osloensis | 2 |

## 4.3 Sample 3

Sample 3 was processed using `Metaflye`, `Raven`, and `Canu` for assembly. Quality control with `FastQC` revealed 672,323 sequences (GC 69%), with low base quality in short reads. After adapter and quality filtering using `Porechop` and `Fastp`, 481,272 high-quality reads (GC 72.8%) remained, with a mean length of 1,833 bp and a duplication rate below 7.4%.

Assembly with `Metaflye` produced contigs ranging from 2,610 bp to 3.28 Mbp, including one circular contig. `Canu` assembled 22 contigs (total length ~8.7 Mbp, NG50: 816 kb), but 2,664 sequences remained unassembled. `Raven` generated a single assembly file, which was compared together with the other assemblies using `MetaQUAST`.

MetaQUAST analysis showed similar genome fractions (86.4%) for all three assemblers. `Metaflye` and `Raven` produced 9 contigs each, while `Canu` had 141. The largest contigs were over 3.2 Mbp for `Metaflye` and `Raven`, and 2.2 Mbp for `Canu`. Summary statistics are presented in Figure 6.



Combined reference | 9 784 577 bp | 1 reference | 1 fragment

| Genome statistics | Flye | Raven | Canu |
|---|---|---|---|
| Genome fraction (%) | 86.365 | 85.607 | 86.545 |
| Duplication ratio | 1.019 | 1.022 | 1.18 |
| Largest alignment | 538 296 | 526 146 | 543 876 |
| Total aligned length | 8 369 785 | 8 323 802 | 9 669 072 |
| NGA50 | ... | ... | ... |
| LGA50 | ... | ... | ... |
| **Misassemblies** | | | |
| # misassemblies | 77 | 87 | 124 |
| Misassembled contigs length | 8 451 004 | 8 416 511 | 8 722 568 |
| **Mismatches** | | | |
| # mismatches per 100 kbp | 29.06 | 49.59 | 39.17 |
| # indels per 100 kbp | 12.28 | 20.64 | 12.62 |
| # N's per 100 kbp | 0 | 0 | 0 |
| **Statistics without reference** | | | |
| # contigs | 9 | 7 | 141 |
| Largest contig | 3 281 136 | 3 281 867 | 2 209 440 |
| Total length | 8 515 057 | 8 453 462 | 9 812 903 |
| Total length (>= 1000 bp) | 8 515 057 | 8 453 462 | 9 812 903 |
| Total length (>= 10000 bp) | 8 504 119 | 8 453 462 | 9 259 845 |
| Total length (>= 50000 bp) | 8 451 004 | 8 416 511 | 8 533 132 |

Extended report

Figure 6: Assembly statistics comparison generated by MetaQUAST.

Taxonomic classification of all assemblies was performed with `Kraken2`. The resulting taxonomic profile, S.noursei, was visualized using a Krona chart (Figure 7).

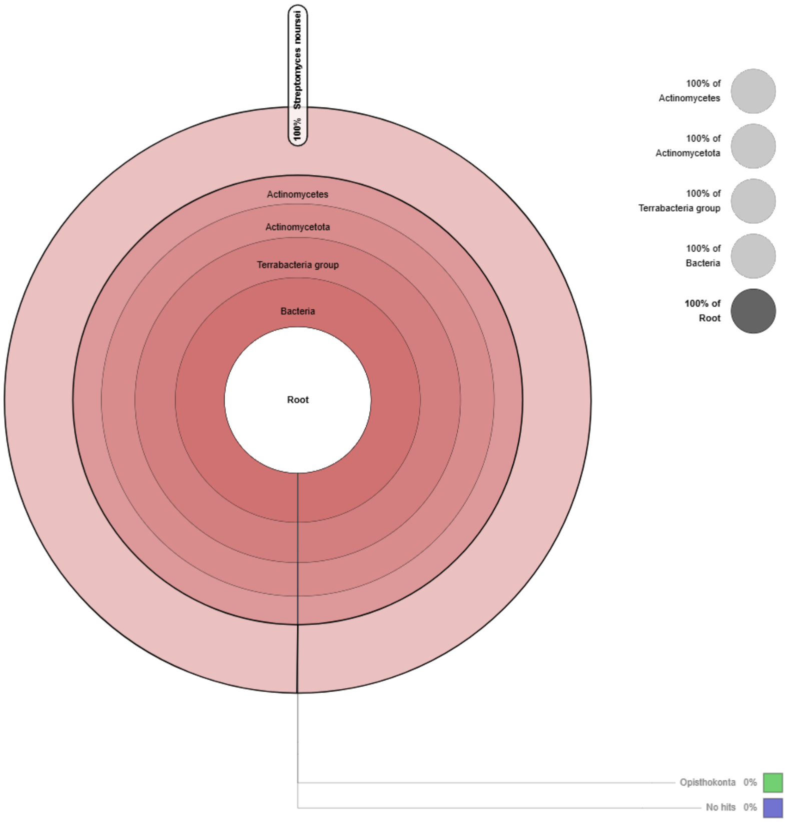


Figure 7: Krona chart showing taxonomic composition.

## 4.4 Sample 4

Sample 4 was successfully assembled using `Raven`, while assemblies with `Metaflye` and `Canu` did not yield usable results. Quality control after adapter and quality trimming indicated that the dataset contained a sufficient number of high-quality reads for assembly.

The assembly with `Raven` resulted in multiple high-coverage contigs, with total assembly size and contig statistics indicating a comprehensive reconstruction of the sample's genomic content. Several of the resulting contigs were identified as circular, suggesting the presence of plasmids or complete genomes. The taxonomic composition is visualized in the Krona chart in Figure 8.

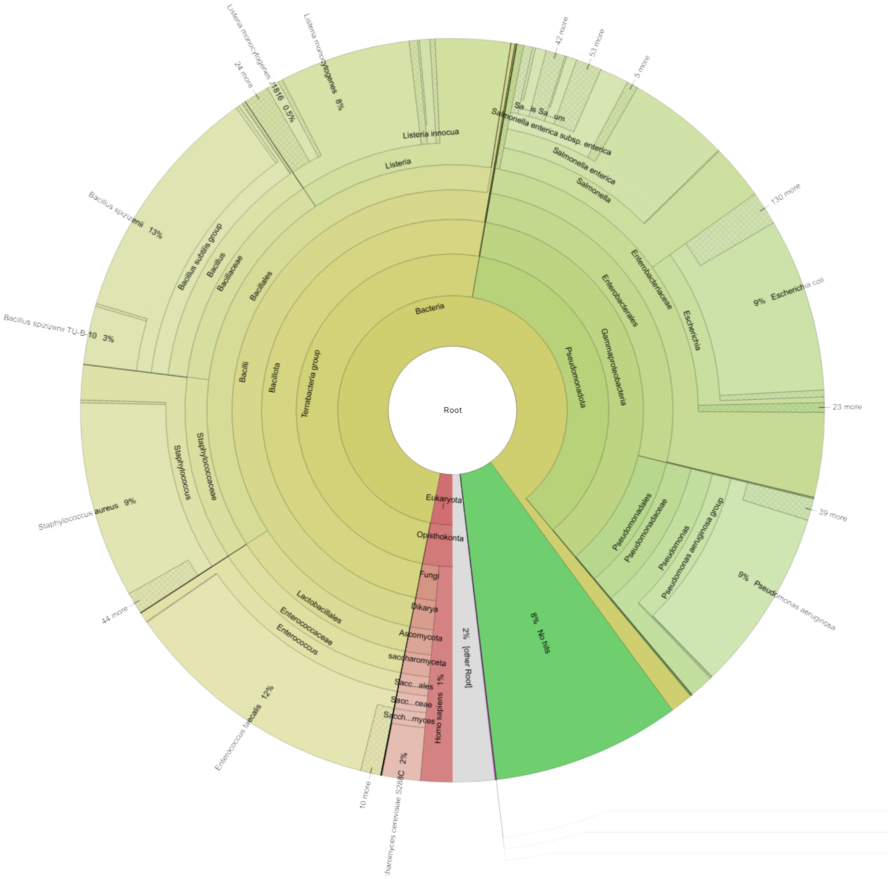


Figure 8: Krona chart showing taxonomic composition of Sample 4.

The most abundant taxa included `Bacillus subtilis` (13%), `Enterococcus faecalis` (12%), and `Salmonella enterica` (10%). The following table summarizes the most represented species detected.

Table 3: Taxons and their relative abundance in Sample 4

| Taxon | Relative abundance [%] |
|---|---|
| Bacillus subtilis | 13 |
| Enterococcus faecalis | 12 |
| Salmonella enterica | 10 |
| Escherichia coli | 9 |
| Pseudomonas aeruginosa | 9 |
| Listeria monocytogenes | 8 |
| Saccharomyces cerevisiae | 2 |

# 5 Discussion

## 5.1 Initial Assembly using metaFlye

The initial round of Assembly was performed using metaFlye. The first two samples could be identified as the oral swab sample and the room monitoring sample. The two remaining samples could not be identified as the DNA standard. Both samples only appeared to contain a single organism. A possible explanation could be the high variance in the length of the genomes in the sample (Saccharomyces cerevisiae ~ 12MB vs Escherichia coli ~ 5 MB).

## 5.2 Filtering prior to Visualization

The step Filtering prior to visualization using the command outlined in Workflow 2 was also a contributing factor in the initial misidentification of Sample 4. This step was initially performed without thorough examination of the data. As evidence of microorganisms other than Saccharomyces cerevisiae could be observed prior to filtering and not after. This resulted in a more detailed examination of the data prior to performing this step for all other following classifications.

## 5.3 Assembly using Canu

The Samples were also assembled using the Canu assembly, but for the most part the resulting data was not included in this report. As per the developers canu is mainly used for the assembly of individual genomes. https://github.com/marbl/canu Canu also resulted in the most time-consuming assemblies as well as the most computational expensive. These findings are corroborated by a study by Latorre-Pérez et al. [18] who also report the longest assembly times. But they also report canu as the best performing assembly using optimized parameters as well as multiple rounds of polishing.
As we primarily used the parameters given via the canu faq
https://canu.readthedocs.io/en/latest/faq.html#what-parameters-should-i-use-for-my-reads we do not recommend the use of canu for assembly of our samples. But further optimization and polishing could result in higher quality outputs.

## 5.4 Recommended Workflow

We do not recommend the use of canu, as well as the use of metaFlye for the assembly of our samples. Canu because of the time intensive assembly process as well as the need for further optimization of the parameters used. MetaFlye is not recommended, as the composition of sample 4 might be the cause of the lower quality of the assembly.
The filtering step prior to visualization should only be used after verification that there is no loss of important data in the resulting plots.
We recommend the use of workflow 2 for the classification of our sample because of the ease of use as well as the lower time and computation resources necessary to perform the classification. Using this workflow

## 5.5 Final sample assignment

Sample 1: oral swab sample

Sample 2: room monitoring sample

Sample 3: single cultured microorganism

Sample 4: ZymoBIOMICS HMW DNA standard from ZymoResearch (Cat.# D6322)

# 6 References

[1] Wang, Y., Zhao, Y., Bollas, A. et al. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 39, 1348–1365 (2021)

[2] Waldherr M. Data Analysis Lab – Sequencing (Vorlesungsfolien, Master Bioinformatik, SS 2025)

[3] Pagès-Gallego, M., de Ridder, J. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. Genome Biol 24, 71 (2023)

[5] Thomas, T., Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. Microb Informatics Exp 2, 3 (2012)

[6] Vigdis Torsvik, Lise Øvreås, Microbial diversity and function in soil: from genes to ecosystems, Current Opinion in Microbiology, Volume 5, Issue 3, Pages 240-245, ISSN 1369-5274 (2002)

[9] Zhang L, Chen F, Zeng Z, Xu M, Sun F, Yang L, Bi X, Lin Y, Gao Y, Hao H, Yi W, Li M, Xie Y. Advances in Metagenomics and Its Application in Environmental Microorganisms. Front Microbiol. (2021)

[10] Prayogo FA, Budiharjo A, Kusumaningrum HP, Wijanarka W, Suprihadi A, Nurhayati N. Metagenomic applications in exploration and development of novel enzymes from nature: a review. J Genet Eng Biotechnol. (2020)

[11] Krinos, A., Bowers, R., Rohwer, R. et al. Time-series metagenomics reveals changing protistan ecology of a temperate dimictic lake. Microbiome 12, 133. (2024)

[12] Garner RE, Gregory-Eaves I, Walsh DA. Sediment Metagenomes as Time Capsules of Lake Microbiomes. mSphere. (2020)

[13] Waldherr M. Data Analysis Lab – Classification (Vorlesungsfolien, Master Bioinformatik, SS 2025)

[14] Kumar Awasthi M, Ravindran B, Sarsaiya S, Chen H, Wainaina S, Singh E, Liu T, Kumar S, Pandey A, Singh L, Zhang Z. Metagenomics for taxonomy profiling: tools and approaches. Bioengineered. (2020)

[15] Jia, X., Hu, L., Wu, M. et al. A streamlined clinical metagenomic sequencing protocol for rapid pathogen identification. Sci Rep 11, 4405 (2021)

[16] Maguire, Meghan & Kase, Julie & Roberson, Dwayne & Muruvanda, Tim & Brown, Eric & Allard, Marc & Musser, Steven & González-Escalona, Narjol. Precision long-read metagenomics sequencing for food safety by detection and assembly of Shiga toxin-producing Escherichia coli in irrigation water (2021)

[17] Maranga M, Szczerbiak P, Bezshapkin V, Gligorijevic V, Chandler C, Bonneau R, Xavier RJ, Vatanen T, Kosciolek T. Comprehensive Functional Annotation of Metagenomes and Microbial Genomes Using a Deep Learning-Based Method. mSystems. (2023)

[18] Latorre-Pérez A, Villalba-Bermell P, Pascual J, Vilanova C. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. Sci Rep. (2020)

# 7 Links

[4] Oxford Nanopore Technologies. https://nanoporetech.com/support/software/data-analysis/where-can-i-find-out-more-about-quality-scores [Accessed 24 April 2025]

[7] National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Washington (DC): National Academies Press (US); 2007. 1, Why Metagenomics? Available from: https://www.ncbi.nlm.nih.gov/books/NBK54011/ [Accessed 24 April 2025]

[8] CD Genomics the Genomics Services Company: Applications of Metagenomics in Biotechnology and Health Care. https://www.cd-genomics.com/applications-of-metagenomics-in-biotechnology-and-health-care.html? [Accessed 24 April 2025]