Datenanalyse - Laborprotokoll

Gruppe 1_b

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

25. April 2025

1 Abstract

From DNA samples to taxonomic classification.

Test

2 Background

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 $\log_{10}$ 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.

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]

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

Sample preparation and sequencing

Sample preparation was done according to *Rapid sequencing DNA V14 - barcoding (SQK-RBK114.24 or SQK-RBK114.96).* Any additional required steps are explained in detail in *GRUPPE_1_B_20250212_Labprotocol.* The output of the oxford nanopore basecaller is in the further steps described as "sample".

Data processing pipeline

To get one hand a meaningful result and on the other one a better insight into what effect quality control and filtering has, two slightly different workflows were defined for data analysis.
For the first one only minimal quality control and no filtering was used while for the second one further quality control and filtering was used.

Workflow 1 - minimal QC and no filtering

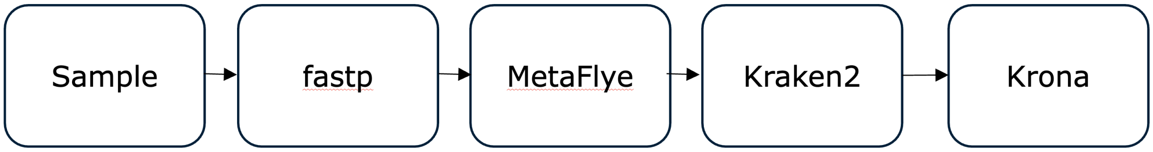


Figure 1: Workflow 1 with minimal QC and no filtering.

In this minimal workflow the received data was preprocessed with fastp which also has minimal qc functionality integrated (**fastp** is a tool for quality control and preprocessing like filtering short reads. Version: 0.24.0, default parameters, https://github.com/OpenGene/fastp).
The received result from fastp was afterwards assembled with MetaFlye (**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, Parameters: , https://github.com/fenderglass/Flye).
For classification of the reads the Kaken2 database was used and Krona to create the visuals. (**Kraken2** is a taxonomic classification database used to assign reads to taxa. Version: 2.1.3, Parameters: , https://github.com/DerrickWood/kraken2) (**Krona** is used to visualize the result as interactive html plots, Version: , Parameters: ,https://github.com/marbl/Krona/wiki)

Workflow 2 - extended QC and filtering

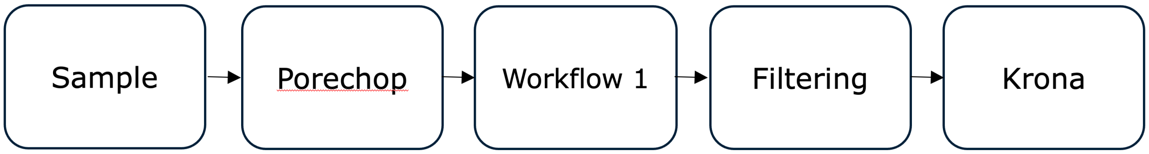


Figure 2: Workflow 2 with extended QC and filtering

The main part (preprocessing -> assemble -> classification) of workflow 2 is the same as in worklow 2, 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, Parameters: ,https://github.com/bonsai-team/Porechop_ABI)

For sample 4 Raven was used as additional assembly method (**Raven** is a de novo genome assembler optimized for long, error-prone reads, Version: 1.8.3, Parameters: ,https://github.com/lbcb-sci/raven)

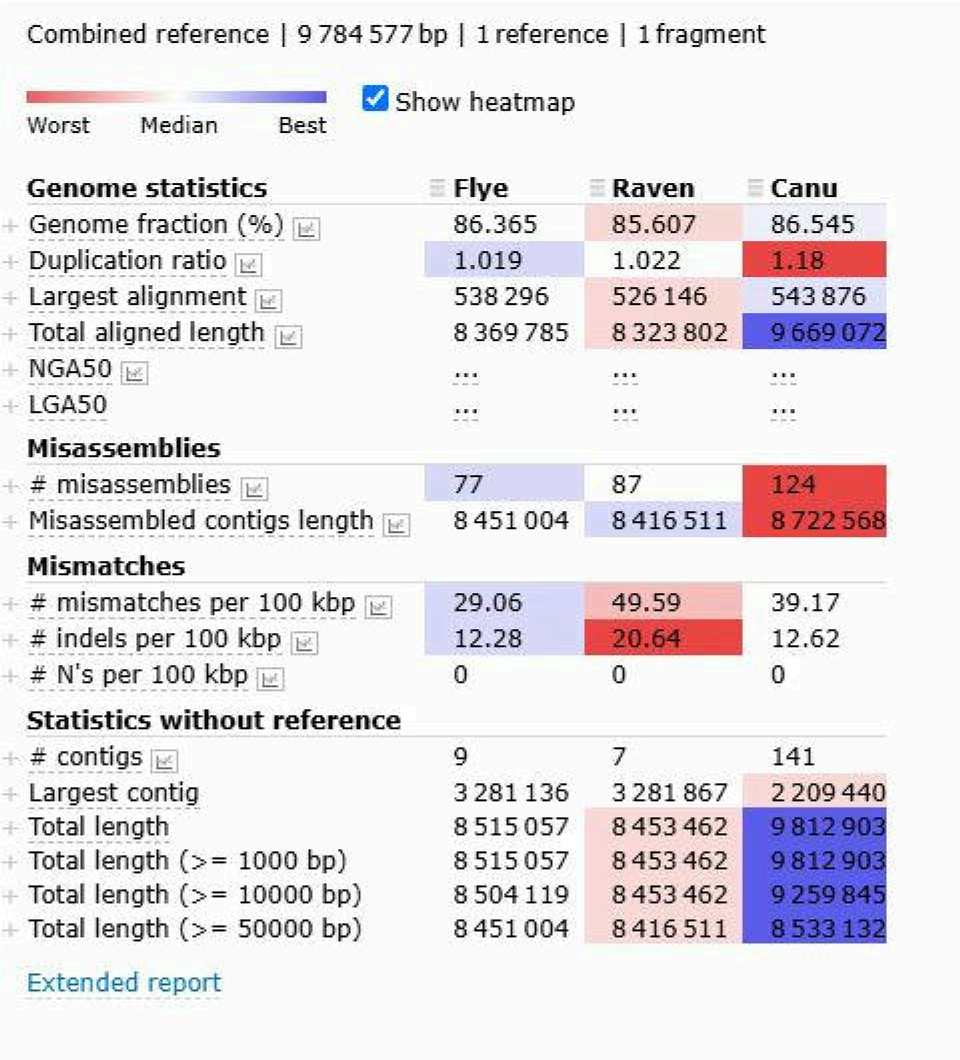
4 Results

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 1.



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

Worst Median Best

☑ Show heatmap

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 3: Assembly statistics comparison generated by MetaQUAST.

Taxonomic classification of all assemblies was performed with `Kraken2`. The resulting taxonomic profiles were visualized using a Krona chart (Figure 2).

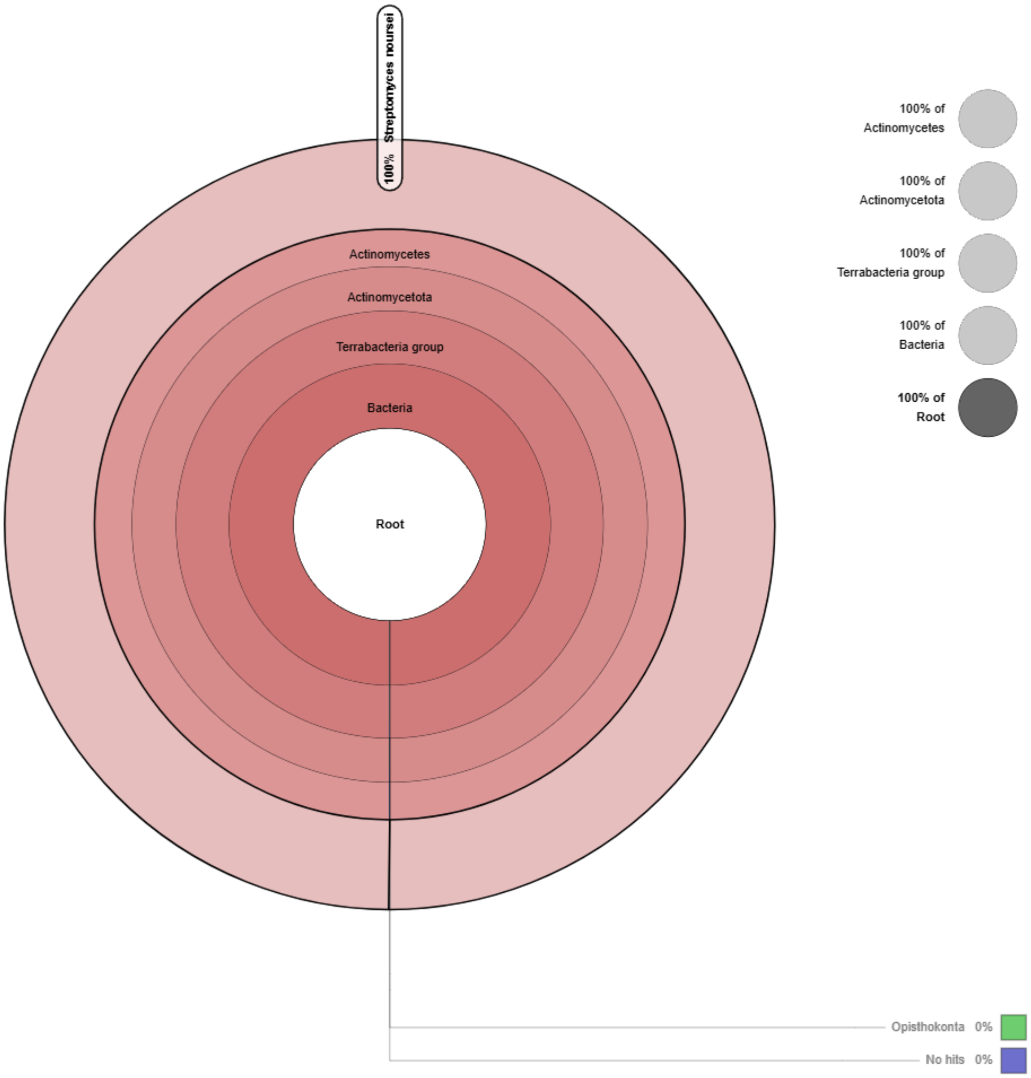


Figure 4: Krona chart showing taxonomic composition.

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 3.

```
![Krona chart showing taxonomic composition of Sample 4.](figures/krona-4.png)
{#fig-krona3 width="600"}
```

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

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

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)

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]