Comparative Evaluation of Raven, metaFlye, and Read-Based Taxonomic Classification in Nanopore Metagenomic Pipelines

Leona Valentina Vracar, Michaela Stätter, Jakob Siebenhütter,
Nicholas Wedige, Ricardo Medina, Markus Eder

April 2025

1 Abstract

Metagenome analysis is a tool for identifying organisms in complex environmental samples. The long-read NGS technology developed by Oxford Nanopore Technologies (ONT) enables long-read sequencing of DNA and RNA by using artificial nanopores. This technological advance has tremendously improved the speed and accuracy of microbial community identification. However, data processing remains a bottleneck in the identification workflow. In this study, we analyzed four samples of varying complexity: an oral swab, an environmental monitoring sample, a mock community DNA standard, and a single-organism sample. For identification of the samples, a read-based pipeline besides two assembly-based approaches using the Raven and metaFlye assemblers were compared with each other.

Initially, a library preparation was performed followed by sequencing of the samples using the ONT MinION platform. Using the obtained sequencing results, the following steps were performed: Quality control of the results followed by trimming and filtering using Porechop and NanoFilt. Assembly of the reads using Raven or metaFlye and final taxonomic classification using Kraken2 and visualization using a Krona chart. All samples were preprocessed with the same parameters to ensure consistency.

The read-based method performed best for the mock community, but was improper for the more complex samples. The other three samples were identified using an assembly-based method: Raven provided structurally accurate assemblies with fewer mismatches, while metaFlye yielded more complete assemblies with longer contigs. Based on our results, we recommend metaFlye for completeness-focused analyses and Raven for structurally precise reconstructions. Applying the different approaches, we were able to identify all the samples.

Keywords: Oxford Nanopore Technologies, MinION, Long-Read Sequencing, Read-Based Classification, Assembly-Based Analysis, Raven assembler, metaFlye assembler, Kraken2, Krona

2 Background

The MinION series by Oxford Nanopore Technologies (ONT) is a portable and affordable third-generation technology for real-time, long-read sequencing. DNA or RNA strands pass through a biological nanopore, generating characteristic changes in ionic current that are decoded into nucleotide sequences using neural network base-calling algorithms. One drawback of long-read ONT sequencing is the comparably high mean error rate on raw reads of approximately 6% relative to Next-Generation Sequencing (NGS) such as Illumina [1].

Despite the reduced accuracy of nanopore sequencing, the portability of ONT devices such as the MinION device makes them particularly suitable for in-field metagenomic studies targeting uncultivated or so far uncharacterized organisms [2]. Furthermore, the long reads produced enable more contiguous assemblies and better resolution of repetitive regions, thus showing potential for more precise taxonomic classification by spanning genomic regions that are often fragmented in short-read data [3], [4]. Still, multiple studies have shown that the reliability of MinION-based taxonomic classification is highly influenced by factors such as library preparation protocols, reference databases, and the choice of bioinformatics pipelines, all of which significantly affect the accuracy and interpretability of results [4]–[6].

This project aimed to evaluate and compare the performance of a nanopore sequencing pipeline across various microbial samples, including an oral swab, a room monitoring swab, a single cultured microorganism and a commercially available mock community ZymoBIOMICS HMW DNA standard (ZymoStandard) [7]. We tested a range of tools for quality control, read processing and assembly including MinIONQC, FastQC, Nanoplot, NanoFilt, Porechop, Raven, metaFlye and Medaka, assessing both genome assembly quality and the accuracy of taxonomic classification. While we aimed to gain a general understanding of how each step of the nanopore metagenomic pipeline influences the final results, we specifically focused on two main questions:

- (i) How does the choice of assembler (Raven vs. metaFlye) affect assembly quality and downstream taxonomic classification?
- (ii) How does direct read-based classification compare to an assembly-based approach in terms of taxonomic resolution?

3 Methods

Long-read sequencing of four high-molecular weight DNA (HMW) samples (codes A, B, C, D) was conducted according to the protocol provided by the Rapid barcoding kit (SQK-RBK114.24, Version 27.12.2024) [8] on a R10.4.1 flow cell. The run consisted of a total of eight DNA samples. The concentration was then measured by the Qubit™ dsDNA HR Assay Kit [9] and a suitable dilution was prepared. For each sample a consecutive barcode was chosen (RB9-RB12). Sequencing parameters were set to a sequencing time of 72 hours in superaccuracy mode to reduce sequencing errors.

3.1 Quality Control, Filtering and Trimming

The initial quality control (QC) of the raw fastq files was performed using FastQC to assess the number of reads generated per barcode, as well as the overall quality for each read length [10]. However, since FastQC was not intended to be used for nanopore reads, for a thorough QC, MinionQC [11] and Nanoplot were performed both before and after the trimming and filtering step. We used Porechop [12] to identify and remove sequencing adapters and chimeric reads from the raw fastq files. Porechop (v.0.2.4) first aligns a subset of reads to known adapter sequences to identify the adapters and then trim them from the end of all reads. Internal adapters i.e. chimeras are either used to split reads or to discard them if splitting would produce reads shorter than 1000 bp.

Following adapter trimming, NanoFilt [13] was used to filter reads based on quality and length, as well as to remove base-calling artifacts from the ends of reads using the parameters below:

```
NanoFilt --length 1000 --maxlength 50000 -q 12 --headcrop 50 --tailcrop 50
```

All four samples were preprocessed with the same parameters to ensure consistency in downstream analysis.

3.2 Assembly and Polishing

The filtered reads were assembled using Raven [14] [15] and metaFlye [16] [17].

Raven is a fast and memory-efficient de-novo genome assembler built for handling high-error-rate data such as the long-reads produced by Nanopore sequencing. Raven assemblies were generated both using the default settings as well as with stricter parameters to improve overlap detection in complex metagenomic samples. A larger k-mer size of 50 (default: 15) was chosen to make overlaps more specific. The minimum relative k-mer frequency was set to 0.003, meaning a k-mer had to appear in at least 0.3% of reads to be considered. This reduces potential false overlaps caused by sequencing errors but may also exclude rare variants or low-abundance organisms. Since the focus of our analysis was not the detection of rare species, this trade-off was acceptable. Finally, the minimum unitig size was lowered from 9,999 to 500 bp to retain shorter, potentially valid sequences from fragmented or low-coverage genomes.

```
raven --kmer-len 50 --polishing-rounds 2 --min-unitig-size 500 -f 0.003 -t 10
```

MetaFlye is specifically designed for metagenomic samples and uses repeat graphs to better reconstruct genomes in complex microbial communities [17]. It was run with the `--nano-hq` parameter which adjusts the assembly to recent ONT data basecalled with Guppy5+ SUP, and with `--meta` to use metaFlye when calling the Flye assembler.

```
flye --nano-hq --meta
```

After assembling with both tools, we polished the resulting contigs with Medaka [18], a neural network consensus polishing tool for ONT reads, to improve accuracy and correct base-level sequencing errors. To assess and compare assembly quality, we used MetaQUAST [19], a reference-based evaluation tool that calculates a range of assembly metrics, including genome fraction, number of contigs, misassemblies, and duplication ratio. MetaQUAST was applied to both unpolished and polished assemblies, and was also used to directly compare the performance of the two assemblers across all samples.

3.3 Taxonomic Classification and Visualization

To determine the taxonomic composition of the samples, we used Kraken2 with the Standard PlusPF database which includes a comprehensive set of microbial genomes. Kraken2 works by assigning DNA sequences to taxa using exact k-mer matches against a reference database. It constructs a hash table of k-mers (default: k = 35) from reference genomes and uses the lowest common ancestor of matching k-mers to classify reads (or contigs if using assembly output).

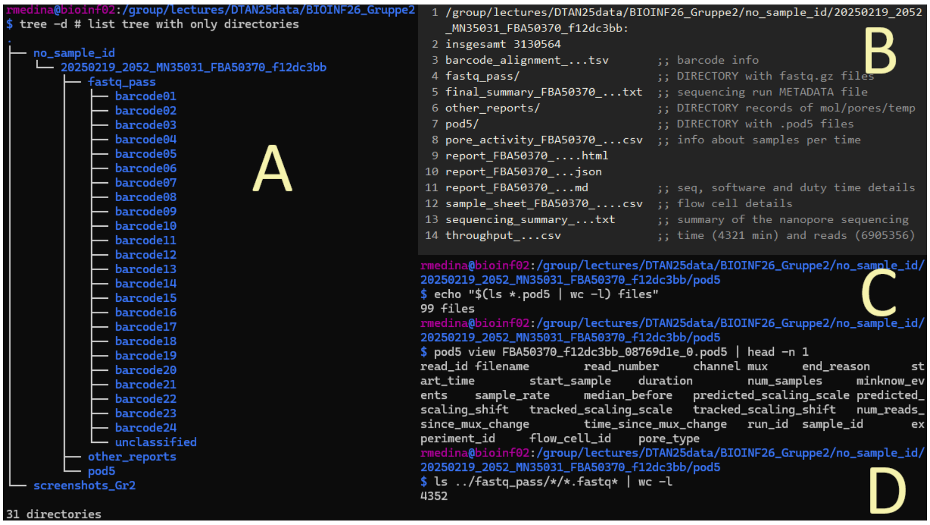
Classification was performed both on the quality-filtered reads (after Porechop and Nanofilt) and on the unpolished assemblies produced by Raven and metaFlye.

For visualization, we used the Galaxy implementation of KrakenTools [20] to convert Kraken2 report files into Krona-compatible text format. The tool (Galaxy version 1.2+galaxy2) was run with the `--no-intermediate-ranks` setting which excludes non-standard taxonomic levels from the output file to produce a cleaner, more interpretable Krona chart. Interactive pie charts of the taxonomic groups were generated with the visualization tool Krona [21] in Galaxy (v2.7.1+galaxy0) using the processed KrakenTools output.

4 Results

4.1 Data Exploration and Quality Control

An overview of the file structure and key data types generated during sequencing and basecalling is shown in Figure 1.



```
rmedina@bioinf02:/group/lectures/DTAN25data/BIOINF26_Gruppe2
$ tree -d # list tree with only directories
.
├── no_sample_id
│   └── 20250219_2052_MN35031_FBA50370_f12dc3bb
│       ├── fastq_pass
│       │   ├── barcode01
│       │   ├── barcode02
│       │   ├── barcode03
│       │   ├── barcode04
│       │   ├── barcode05
│       │   ├── barcode06
│       │   ├── barcode07
│       │   ├── barcode08
│       │   ├── barcode09
│       │   ├── barcode10
│       │   ├── barcode11
│       │   ├── barcode12
│       │   ├── barcode13
│       │   ├── barcode14
│       │   ├── barcode15
│       │   ├── barcode16
│       │   ├── barcode17
│       │   ├── barcode18
│       │   ├── barcode19
│       │   ├── barcode20
│       │   ├── barcode21
│       │   ├── barcode22
│       │   ├── barcode23
│       │   ├── barcode24
│       │   └── unclassified
│       ├── other_reports
│       └── pod5
└── screenshots_Gr2

31 directories
```

```
/group/lectures/DTAN25data/BIOINF26_Gruppe2/no_sample_id/20250219_2052
_MN35031_FBA50370_f12dc3bb:
insgesamt 3130564
barcode_alignment_...tsv       ;; barcode info
fastq_pass/                    ;; DIRECTORY with fastq.gz files
final_summary_FBA50370_...txt  ;; sequencing run METADATA file
other_reports/                 ;; DIRECTORY records of mol/pores/temp
pod5/                          ;; DIRECTORY with .pod5 files
pore_activity_FBA50370_...csv  ;; info about samples per time
report_FBA50370_....html
report_FBA50370_...json
report_FBA50370_...md          ;; seq, software and duty time details
sample_sheet_FBA50370_....csv  ;; flow cell details
sequencing_summary_...txt      ;; summary of the nanopore sequencing
throughput_...csv              ;; time (4321 min) and reads (6905356)
```

```
rmedina@bioinf02:/group/lectures/DTAN25data/BIOINF26_Gruppe2/no_sample_id/
20250219_2052_MN35031_FBA50370_f12dc3bb/pod5
$ echo "$(ls *.pod5 | wc -l) files"
99 files
rmedina@bioinf02:/group/lectures/DTAN25data/BIOINF26_Gruppe2/no_sample_id/
20250219_2052_MN35031_FBA50370_f12dc3bb/pod5
$ pod5 view FBA50370_f12dc3bb_08769d1e_0.pod5 | head -n 1
read_id filename        read_number     channel mux     end_reason      st
art_time        start_sample    duration        num_samples     minknow_ev
ents    sample_rate     median_before   predicted_scaling_scale predicted_
scaling_shift   tracked_scaling_scale   tracked_scaling_shift   num_reads_
since_mux_change        time_since_mux_change   run_id  sample_id       ex
periment_id     flow_cell_id    pore_type
rmedina@bioinf02:/group/lectures/DTAN25data/BIOINF26_Gruppe2/no_sample_id/
20250219_2052_MN35031_FBA50370_f12dc3bb/pod5
$ ls ../fastq_pass/*/*.fastq* | wc -l
4352
```

Figure 1: **File structure and content**. (**A**) General folder structure without files. (**B**) Key directories and files within the raw data, including fastq, pod5 and documentation reports. (**C**) Folder pod5 containing a total of 99 pod5 files generated using guppy 7.6.7, which can be visualized with the command `pod5 view`. (**D**) 4352 fastq.gz files distributed across barcode-specific folders, with approximately 433 files per folder to enable parallelization and multiplexing.

4.1.1 Flow Cell Quality Control

The MinIONQC channel heatmap revealed typical variation in sequencing yield across the flow cells, with some channels contributing less data than others. The right panel in Figure 2 (Q≥7) confirmed that many of the higher yielding channels also produced high-quality reads, indicating that while the overall flow cell output was somewhat unbalanced, a sufficient number of active pores contributed high-quality data suitable for downstream analysis.

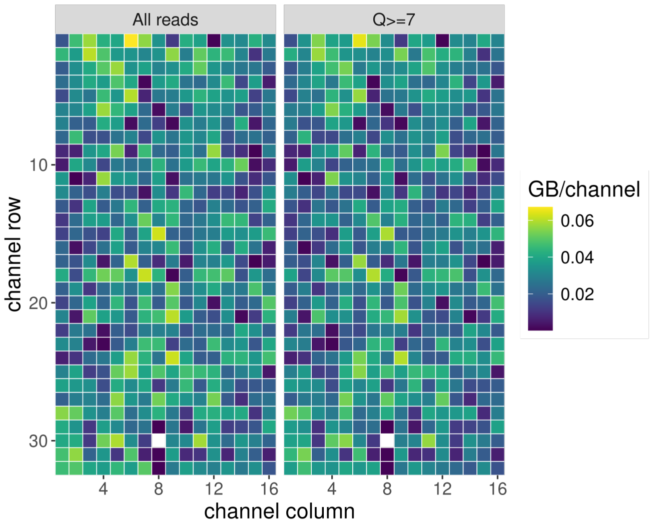


Figure 2: **Per-channel sequencing yield and high-quality output (Q≥7) accross the flow cell as visualized by MinIONQC**. Channel performance was within expected variation for a standard nanopore run.

Initial quality control, performed on the raw basecalled reads, revealed a high proportion of short, low-quality reads across all samples. This was especially apparent in BC09, where the median read length was only 700 bp and the median quality score was below Q13.

4.1.2 Filtering and Trimming

The effects of adapter trimming and chimera splitting across all samples are summarized in Table 1

Sample Barcode	**Total Reads**	**Trimmed at start** (%)	**Trimmed at end** (%)	**Reads Split** (%)
BC09	1,065,437	1,009,339 (94.7%)	0%	41,804 (3.9%)
BC10	729,889	710,757 (97.4%)	0%	4,825 (0.7%)
BC11	1,217,908	1,167,741 (95.9%)	0%	46,966 (3.9%)
BC12	667,557	647,810 (97.0%)	0%	6,533 (1.0%)

Table 1: **Summary of adapter trimming and chimera splitting using Porechop (v0.2.4).**

Porechop removed adapters from the start of over 94% of reads in all samples, with no adapters detected at the read ends. A small number of chimeric reads, ranging from 0.7% to 3.9%, were also identified and split.

Using Nanofilt, we filtered reads below the mean quality score Q12 as well as short reads (<1000 bp) and ultra-long reads (>50kb). Additionally, 50 bases were trimmed from both the 5' and 3' ends of each read to remove low-quality regions.

Ultra long-reads (>50bp) were removed as they can cause compatibility issues with taxonomic classification tools, especially with short-read programs such as kraken2 [22]. Reads shorter than 1000 bp were also excluded with the aim of focusing the assembly on more informative fragments and reducing noise from short sequences.

Sample	**Raw Reads**	**Post-Filter Reads**	**Filtered Reads**	**% Filtered**	**Read N50**	**Median Length**	**Median Quality**
	(10^3)	(10^3)	(10^3)		(bp)	(bp)	
BC09	1 065.4	351.7	713.7	66.95%	3 504 → 4 248	700 → 2 406	12.9 → 19.7
BC10	729.9	360.8	369.1	50.57%	5 604 → 5 871	1 501 → 2 783	14.9 → 20.2
BC11	1 217.9	575.8	642.1	52.70%	4 651 → 5 099	1 292 → 2 766	14.4 → 19.1
BC12	667.6	366.8	300.8	45.06%	7 604 → 7 861	1 902 → 3 591	14.9 → 19.6

Table 2: **Summary of raw vs. filtered Nanopore reads, extracted from the NanoStats.txt report generated by NanoPlot.** Read counts in thousands; lengths in bp.

After filtering, all samples reduced in read count (ranging from 45-67%) but notably improved in median length, N50 and median read quality. For example, in the sample with the lowest quality BC09, filtering removed approximately 67% of reads and lead to a >3-fold increase in mean read length (700 to 2,406 bp), suggesting that the raw data contained many short, low quality fragments that would have negatively affected the downstream analysis.

4.2 Taxonomic Classification and Visualization

During sequencing, we did not initially know which barcode corresponded to which sample. Read-based classification on quality-filtered reads with Kraken2 proved to be most accurate in detecting the taxonomic composition of the known ZymoStandard. Based on the dominant taxa and their ecological presence, we identified the samples as follows:

- **BC09 (Oral Swab)**: Read-based classification revealed a predominance of *Homo sapiens* (64%) and oral commensals such as *Streptococcus mitis* (6%) and *Haemophilus haemolyticus* (3%). These species are typical of the human oral microbiome, suggesting that BC09 corresponds to the oral swab.
- **BC10 (Cultured Organism)**: The sample was overwhelmingly classified as *Staphylococcus epidermidis* (91%) and 8% was assigned as belonging to the phylum *Actinomycetota* with *Micrococcus luteus* (7%) being the most abundant species. Although the expected organism (*Streptomyces noursei*) was not detected in the raw-read classification, it belongs to the phylum *Actinomycetota* class *Actinomycetes*, which also includes *Micrococcus luteus*. We therefore infer that BC10 contains the cultured microorganism, possibly misclassified due to sequence similarity or low coverage. This is discussed further in the assembly-based analysis. As for *Staphylococcus epidermidis*, it is a bacterium present in the human skin microbiota and therefore most likely originated from sample contamination.
- **BC11 (Room Monitoring Sample)**: This sample displayed the most diverse and complex community, with a wide distribution across multiple genera including *Prevotella*, *Rothia*, *Fusobacterium*, and environmental

taxa. The presence of numerous low-abundance species as well as homo sapiens DNA (21%) suggest a mixed environmental origin, consistent with a swab from a room environment.

- **BC12 (ZymoStandard)**: This sample closely matched the known composition of the ZymoStandard, correctly identifying all expected organisms with approximately correct relative abundances. Notably, *Saccharomyces cerevisiae* was detected at 2%, and the remaining seven bacterial taxa were evenly represented, confirming the identity of BC12 as the mock community standard.

4.3 Assembly and Polishing

4.3.1 Raven

Based on MetaQUAST reports, the overal assembly quality and complexity varied considerably across the four samples (BC09-BC12):

Sample BC11, identified as the room monitoring swab, yielded the most fragmented assemblies with the highest number of contigs (841 from the default assembly, 912 with custom settings) as well as the largest total assembly length (~51–56 Mbp). These findings are consistent with its complex and diverse microbial content.

In contrast, sample BC10, identified as the single cultured organism, produced the smallest and most compact assemblies (~5.5 Mbp total length) with the lowest number of contigs (24 from the default assembly, 21 custom settings).

For BC12, all organisms from the Zymo Standard were recovered in the assembly except for *Saccharomyces cerevisiae*, the only eukaryotic member, which was present at just 2% relative abundance in the standard. The likely reasons for this include insufficient sequencing depth for the fungal genome and inherent challenges in assembling larger, more repetitive eukaryotic genomes. The assembly included unexpected alignments to three *Shigella* species, which are not part of the standard. This is most likely due to taxonomic misclassification as *Shigella* species are closely related to *Escherichia coli* which is in the Zymo Standard.

4.3.2 Raven: Standard vs. Custom Settings

For BC12 (Zymo Standard) the custom Raven settings clearly outperformed the standard configuration, generating a cleaner, more contiguous assembly (20 vs. 44 contigs) with fewer misassemblies, fewer errors, and lower fragmentation (see Figure 3).

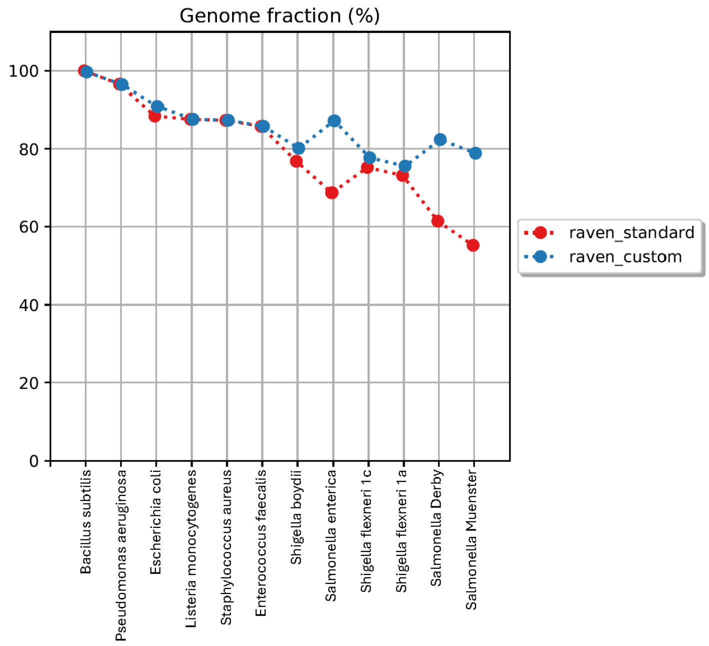


Figure 3: **Genome fraction BC12**: Genome fraction improved with custom parameters in the ZymoStandard, particularly for assembled *Salmonella* strains and *E. coli*.

The room monitoring swab, sample BC11, showed slightly increased contiguity with custom settings (fewer contigs, longer total length), but also lead to a slight increase in misassemblies (634 vs. 649). Overall, custom Raven settings led to slightly better genome recovery across most taxa in all our samples.

4.3.3 Metaflye: Comparison of Pre- and Post-Polishing Assemblies

The results obtained from all four samples were highly consistent with regard to assembly quality before and after polishing. Sample BC09 is presented in Figure 4 as a representative example.

Polishing with Medaka resulted in only minimal changes in the number of contigs, contig sizes, and total assembly length. While the number of mismatches per 100,000 bases slightly increased post-polishing, small indel errors were marginally reduced. No substantial structural differences were observed between the polished and unpolished assemblies.

Overall, these findings indicate that Medaka slightly enhances assembly accuracy by correcting smaller errors, while maintaining structural consistency. The results before and after polishing are thus consistent and can be considered reliable.

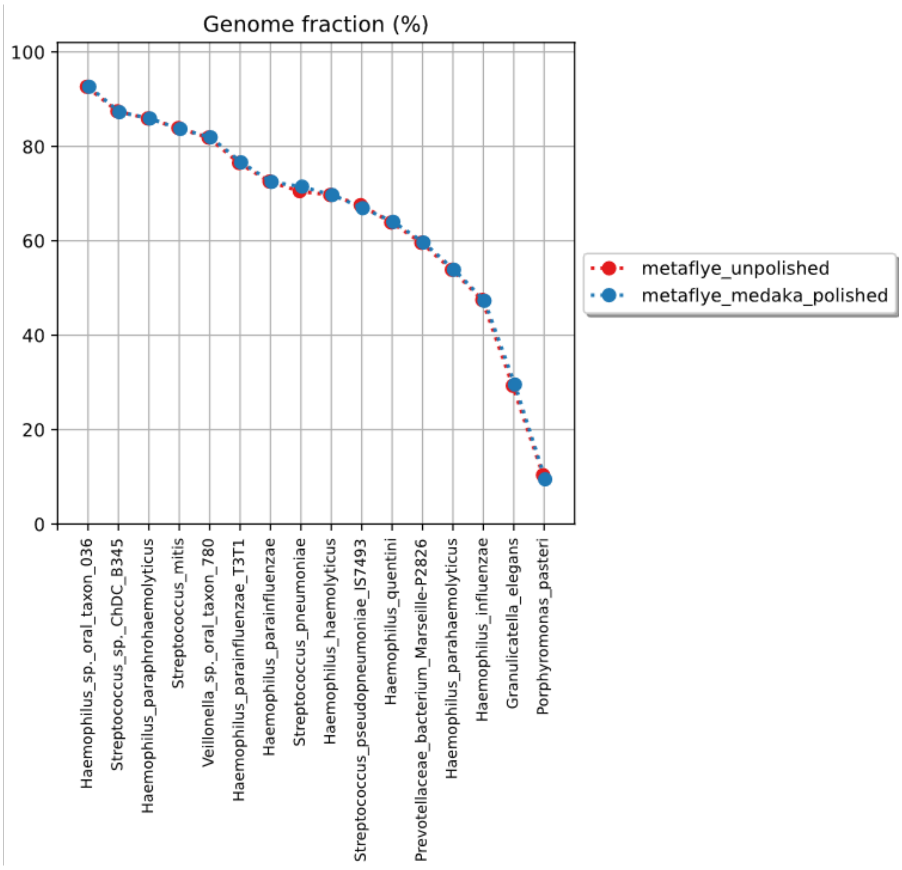


Figure 4: **Genome fraction BC09**: No major differences in genome fraction between polished and unpolished assemblies. BC09 is shown as a representative example; similar results were observed for all samples.

4.3.4 Metaflye: Visualization of assembly-graphs using Bandage

Metaflye generates several assembly-graph files as additional output. We visualized these assembly graphs using the tool Bandage (Bioinformatics Application for Navigating De novo Assembly Graphs Easily)[23], allowing us to explore connectivity, coverage and contig lengths for each assembly. Node-size is proportional to contig length, while edges represent sequence overlaps, meaning complex, branched graphs hint to a sample containing high strain-level variation or high numbers of repetitive regions. In a ”clean” or homogeneous sample we expect to see self-connecting edges, i.e. fully assembled circular genomes, while we expect more ”tangled” graphs from samples containing more varied bacterial strains. For visualization, all nodes were filtered, so only nodes with a minimum coverage depth of 20x are shown.

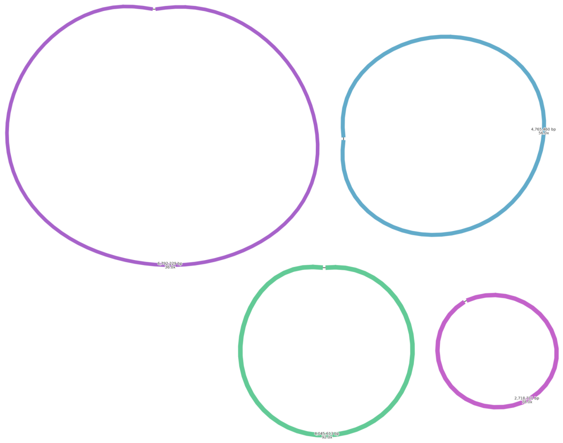


Figure 5: **Assembly graph for BC12**: As expected for sample BC12, we can observe properly assembled, circular genomes, lacking any branched structures. As this sample is the Zymo Standard, one can anticipate high strain-level homogeneity, leading to a clean assembly.

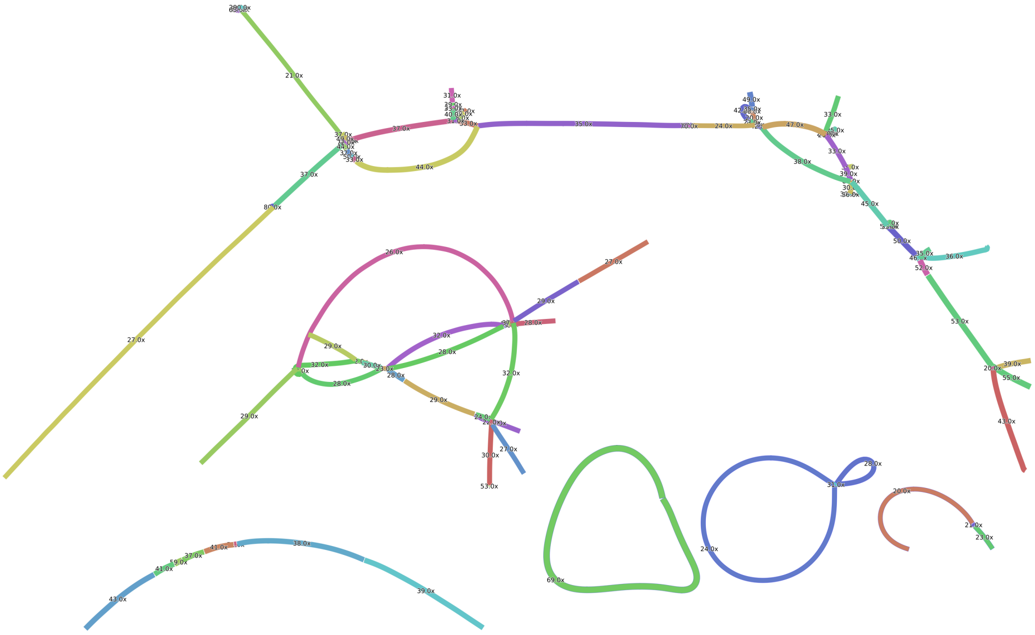


Figure 6: **Assembly graph for BC11**: In comparison, the assembly graph of sample BC11 shows highly branched, tangled sub-graphs. With this sample being the room monitoring sample, we would indeed expect rather high strain variety, leading to a less accurate assembly on the one hand, and more branching structures due to genomic variability on the other.

4.4 Assembly vs. Read-Based Classification

As mentioned previously, read-based classification using Kraken2 successfully identified all eight species present in the Zymo Standard, with approximately correct relative abundances (see Figure 7, Krona chart based on read-level classification).

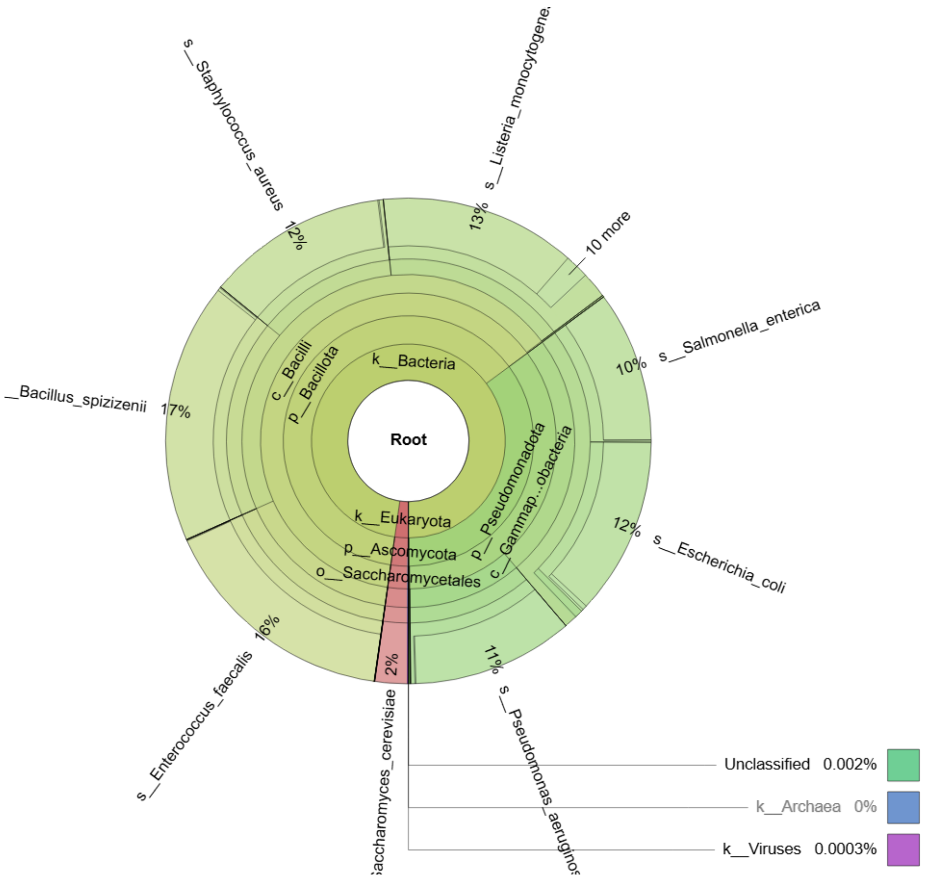


Figure 7: **Krona chart of Kraken2 read-based classification for BC12 (raw reads).**

When comparing the classification of assembled contigs, Kraken2 revealed notable differences between the two assemblers. *Saccharomyces cerevisiae* was heavily overrepresented in both cases, comprising 25% of the classified contigs in the Raven assembly and as much as 94% in the metaFlye assembly. Notably, this eukaryotic genome was not reported at all in the MetaQUAST results of neither of the assemblers. Discrepancies between the species detected by MetaQUAST assembly analysis and those classified via Kraken2 stem from the differing underlying methodologies: MetaQUAST relies on alignment against selected reference genomes, while Kraken2 performs k-mer-based taxonomic classification using a broader reference database.

Despite overestimating *S. cerevisiae*, the Raven assembly maintained a broader and more balanced bacterial profile, with key taxa like *P. aeruginosa*, *E. coli*, and *S. aureus* still detectable. In contrast, metaFlye largely failed to recover the expected diversity, overwhelmingly reporting *S. cerevisiae* and underrepresenting all other species.

Organism	**ZymoStandard (%)**	**Raw Reads (%)**	**Raven (%)**	**metaFlye (%)**
Pseudomonas aeruginosa	14	11	5	0.4
Escherichia coli	14	12	10	0.9
Salmonella enterica	14	10	40	0.9
Enterococcus faecalis	14	16	5	0.4
Staphylococcus aureus	14	12	5	2
Listeria monocytogenes	14	13	5	0.4
Bacillus subtilis	14	17	5	0.4
Saccharomyces cerevisiae	2	2	25	94
Other species	0	7	0	0.6
Unclassified	0	0.002	0	0

Table 3: **Relative abundances (%) of species in the ZymoStandard as detected by Kraken2 in raw reads, Raven assembly, and metaFlye assembly for sample BC12.**

While read-based classification was the most accurate approach for classifying the Zymo Standard (BC12), the read-based classification of BC10, identified as the single cultured organism, showed to be far less accurate. As discussed in the sample-identification section, read-based classification of BC10 classified 91% as *Staphylococcus epidermidis* (91%) and 8% to the phylum *Actinomycetota* with *Micrococcus luteus* (7%) being the most abundant species. Kraken2 classification of the Raven assembly identified 60% as *Staphylococcus epidermidis* and 40% *Micrococcus luteus* while classifying the metaFlye assembly revealed 82% *Staphylococcus epidermidis*, 2% *Staphylococcus hominis*, 5% *Staphylococcus aureus*, and 11% *Micrococcus luteus*.

Interestingly, MetaQUAST assembly evaluation revealed that both Raven and metaFlye assembled contigs corresponding only to two species—*Micrococcus aloeverae* and *Micrococcus luteus*—both belonging to the same class (*Actinomycetes*) as the expected cultured organism *Streptomyces noursei*. This suggests that discrepancies in assembly and classification results, such as the absence of the most abundant read-classified species from the assembled contigs, may come from misclassifications at the read level and solely relying on read-based classification is not recommended.

5 Discussion

5.1 Assembly vs. Read-Based Classification

In the case of the Zymo Standard, read-based classification outperformed contig-based approaches, recovering all expected species at approximately correct relative abundances. This suggests that direct classification of raw reads can provide a more accurate snapshot of community composition, especially when assemblies are incomplete or biased. However, this does not mean assembly-based classification should be dismissed. When key organisms are not successfully assembled, their sequences may be absent from downstream analyses, making assembly a potential bottleneck and unintentional filtering step. Still, assemblies can offer important insights into genome structure, variation, and functional potential—information that read-based methods alone cannot provide. Thus, the choice between read- and assembly-based classification should consider the study's goals, data quality, and the complexity of the sample.

5.2 Assembly Quality Assessment: Raven vs. metaFlye

The comparison between Raven and metaFlye shows that both assemblers have distinct strengths. MetaFlye tends to produce assemblies with a slightly higher genome fraction and longer contigs, which may indicate more complete assemblies. However, this comes at the cost of increased misassemblies, which can compromise structural accuracy. Raven, on the other hand, delivers cleaner assemblies with significantly fewer misassemblies, although its genome fraction and total contig length are somewhat lower in some samples. The mismatch rates were relatively similar, with no consistent advantage for either assembler. Overall, the choice between the two may depend on the specific application — metaFlye may be preferred when completeness is prioritized, whereas Raven may be more suitable for structurally accurate assemblies.

6 Limitations and Recommendations

Integrating multiple methodologies in ONT data analysis is essential to balance the strengths and limitations of individual tools, fostering a more reliable consensus [22]. Additionally, parameter optimization is crucial, as default settings may not always yield optimal results [24]. Adjusting these configurations to align with dataset-specific characteristics enhances sequencing accuracy, data integrity, and overall analytical precision [25].

When examining limitations and recommendations in ONT sequencing, filtering and trimming emerge as critical factors. Although both processes are widely recognized as essential, the optimal methodology remains uncertain. In this regard, Portik *et al*. [22] highlight that reads shorter than 2 kb or longer than 50 kb may present compatibility issues, particularly with short-read classification tools like Kraken2. While our study follows these recommendations, alternative strategies deserve further consideration. For instance, Lee *et al*. [25] argue that the Phred Quality Score cutoff in NanoFilt may inadvertently exclude high-fidelity sequences. Their proposed window-based trimming method, which retains valuable genomic segments, has demonstrated improvements in alignment precision and genome assembly reliability through their tool, Prowler.

While this method offers potential benefits in recovering ultra-long read fragments and refining ONT data analysis, it also introduces additional complexities, particularly regarding the trade-off between error rate and assembly contiguity. Furthermore, Lee *et al*. investigated whether Prowler-trimmed reads enhanced genome polishing compared to untrimmed reads. Their findings revealed no significant improvement, suggesting that trimming may not be necessary when the primary goal is final genome refinement rather than optimizing raw sequence alignment. These results highlight the ongoing lack of consensus in ONT data analysis, reinforcing the need for continued advancements to improve sequencing accuracy and data integrity.

A Software Versions and Databases

- **MinIONQC** v1.4.1
- **FastQC** v0.11.9
- **NanoFilt** v2.8.0
- **Porechop** v0.2.4
- **Raven** v1.8.3
- **metaFlye** v2.9.4-b1799
- **Medaka** v1.11
- **MetaQUAST** QUAST v5.2.0 (MetaQUAST mode)
- **Kraken2** v2.1.3 (database: Standard PlusPF, updated 2023-10)
- **KrakenTools** (commit: latest as of March 2025)
- **Krona** v2.8
- **Bandage** v0.9.0

References

[1] C. Delahaye and J. Nicolas, “Sequencing dna with nanopores: Troubles and biases,” *PLOS ONE*, vol. 16, no. 10, e0257521, 2021. DOI: 10.1371/journal.pone.0257521.

[2] M. Jain, I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten, and M. Akeson, “Improved data analysis for the minion nanopore sequencer,” *Nature Methods*, vol. 12, no. 4, pp. 351–356, Apr. 2015, Epub 2015 Feb 16. DOI: 10.1038/nmeth.3290.

[3] Á. Latorre-Pérez, P. Villalba-Bermell, J. Pascual, and C. Vilanova, “Assembly methods for nanopore-based metagenomic sequencing: A comparative study,” *Scientific Reports*, vol. 10, no. 1, p. 13 588, 2020. DOI: 10.1038/s41598-020-70491-3. [Online]. Available: https://www.nature.com/articles/s41598-020-70491-3.

[4] A. Bertolo, E. Valido, and J. Stoyanov, “Optimized bacterial community characterization through full-length 16s rrna gene sequencing utilizing minion nanopore technology,” *BMC Microbiology*, vol. 24, p. 58, 2024. DOI: 10.1186/s12866-024-03208-5. [Online]. Available: https://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-024-03208-5.

[5] W.-K. Wu, C.-C. Chen, S. Panyod, *et al.*, “Optimization of fecal sample processing for microbiome study—the journey from bathroom to bench,” *Journal of the Formosan Medical Association*, vol. 118, no. 2, pp. 545–555, 2019.

[6] P. Zheng, C. Zhou, Y. Ding, *et al.*, “Nanopore sequencing technology and its applications,” *MedComm*, vol. 4, no. 4, e316, Jul. 2023. DOI: 10.1002/mco2.316.

[7] Zymo Research, *Zymobiomics hmw dna standard*, https://zymoresearch.eu/products/zymobiomics-hmw-dna-standard, Accessed: 2025-04-12, 2025.

[8] Oxford Nanopore Technologies, *Rapid barcoding rbk*$_{9}176_{v}114_{r}evq_{2}7dec2024$, https://nanoporetech.com/document/rapid-sequencing-gdna-barcoding-sqk-rbk114, Accessed: 2025-04-12, 2024.

[9] Thermofisher, *Qubit™ dsdna hs assay kit - user guide*, https://assets.thermofisher.com/TFS-Assets/LSG/manuals/Qubit_dsDNA_HS_Assay_UG.pdf, Accessed: 2025-04-12, 2025.

[10] *Fastqc*, Jun. 2015. [Online]. Available: https://qubeshub.org/resources/fastqc.

[11] R. Lanfear, M. Schalamun, D. Kainer, W. Wang, and B. Schwessinger, “Minionqc: Fast and simple quality control for minion sequencing data,” *Bioinformatics*, vol. 35, no. 3, pp. 523–525, 2019.

[12] R. Wick, *Porechop: Adapter trimmer for oxford nanopore reads*, https://github.com/rrwick/Porechop, Accessed: 2025-04-23, 2017.

[13] W. D. Coster, *Nanofilt: Filtering tool for nanopore sequencing data*, https://github.com/wdecoster/nanofilt, Accessed: 2025-04-23, 2018.

[14] R. Vaser and I. Sović, *Raven: De novo genome assembler for long reads*, https://github.com/lbcb-sci/raven, Accessed: 2025-04-23, 2020.

[15] R. Vaser and M. Šikić, "Time-and memory-efficient genome assembly with raven," *Nature Computational Science*, vol. 1, no. 5, pp. 332–336, 2021.

[16] M. Kolmogorov, *Metaflye: Long-read metagenome assembler*, https://github.com/mikolmogorov/Flye, Accessed: 2025-04-23, 2019.

[17] M. Kolmogorov, D. M. Bickhart, B. Behsaz, *et al.*, "Metaflye: Scalable long-read metagenome assembly using repeat graphs," *Nature methods*, vol. 17, no. 11, pp. 1103–1110, 2020.

[18] O. N. Technologies, *Medaka: Sequence correction provided by ont research*, https://github.com/nanoporetech/medaka, Accessed March 2025, 2023.

[19] A. Mikheenko, V. Saveliev, and A. Gurevich, "Metaquast: Evaluation of metagenome assemblies," *Bioinformatics*, vol. 32, no. 7, pp. 1088–1090, 2016.

[20] J. Lu, N. Rincon, D. E. Wood, *et al.*, "Metagenome analysis using the kraken software suite," *Nature protocols*, vol. 17, no. 12, pp. 2815–2839, 2022.

[21] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a web browser," *BMC bioinformatics*, vol. 12, pp. 1–10, 2011.

[22] D. M. Portik, C. T. Brown, and N. T. Pierce-Ward, "Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets," *BMC bioinformatics*, vol. 23, no. 1, p. 541, 2022.

[23] R. R. Wick, M. B. Schultz, J. Zobel, and K. E. Holt, *Bandage: Interactive visualization of de novo genome assemblies*, 2015. DOI: 10.1093/bioinformatics/btv383. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/20/3350/49036207/bioinformatics_31_20_3350.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btv383.

[24] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, "Canu: Scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation," *Genome Research*, vol. 27, no. 5, pp. 722–736, Mar. 2017, ISSN: 1549-5469. DOI: 10.1101/gr.215087.116. [Online]. Available: http://dx.doi.org/10.1101/gr.215087.116.

[25] S. Lee, L. T. Nguyen, B. J. Hayes, and E. M. Ross, "Prowler: A novel trimming algorithm for oxford nanopore sequence data," *Bioinformatics*, vol. 37, no. 21, P. Robinson, Ed., pp. 3936–3937, Sep. 2021, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btab630. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btab630.