In-silico enrichment of bacterial plasmids by chromosomal depletion using nanopore adaptive sampling

Jens-Uwe Ulrich[1,2] ✉, **Lennard Epping**[3], **Tanja Pilz**[3], **Birgit Walther**[4], **Kerstin Stingl**[5], **Torsten Semmler**[3], **and Bernhard Y. Renard**[1,] ✉

[1]Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam, Germany
[2]Department of Mathematics and Computer Science, Free University of Berlin, 14195 Berlin, Germany
[3]Genome Sequencing Unit (MF2), Robert Koch Institute, 13353 Berlin, Germany
[4]Advanced Light and Electron Microscopy, Robert Koch Institute, 13353 Berlin, Germany
[5]National Reference Laboratory for Campylobacter, Department of Biological Safety, German Federal Institute for Risk Assessment (BfR), 12277 Berlin, Germany

Bacterial plasmids play a major role in the spread of antibiotic resistance genes. However, their characterization via DNA sequencing suffers from the small proportion of plasmid DNA in those samples. Although sample preparation methods can enrich the proportion of plasmid DNA before sequencing, these methods are expensive and laborious, and they might introduce a bias by enriching only for specific plasmid DNA sequences. Nanopore adaptive sampling could overcome these issues by rejecting uninteresting DNA molecules during the sequencing process. In this study, we assess the application of adaptive sampling for the enrichment of low abundant plasmids in known bacterial isolates using two different adaptive sampling tools. We further inspect potential cost savings for laboratories by utilizing adaptive sampling on reused or expired flow cells. We show that a significant enrichment can be achieved even on reused or expired flow cells. By applying adaptive sampling, we also improve the quality of *de novo* plasmid assemblies and reduce the sequencing time. Our experiments indicate that the choice of the read classification method and read lengths used for the classification influence the enrichment factor.

adaptive sampling | readuntil | nanopore sequencing | plasmid | bacteria | enrichment

Correspondence: ***jens-uwe.ulrich@hpi.de, bernhard.renard@hpi.de***

Introduction

Infectious diseases caused by bacterial pathogens have lost their threat to people living in high-income countries due to the discovery of antibiotic drugs within the last 70 years. However, adaptation processes within bacteria cause these drugs to lose their effectiveness in treating infectious diseases. The emergence of such antimicrobial resistance (AMR) already poses a significant threat to public health, with an estimation of 4.95 million deaths associated with bacterial AMR in 2019 (1) and will even worsen with around 10 million expected deaths per year by 2050 (2, 3).

Besides vertically passing antimicrobial resistance genes (ARG) to their offspring, bacteria are also capable of transferring ARGs across the bacterial population by horizontal gene transfer. This process is mediated via mobile genetic elements (MGE), such as plasmids, which are epichromosomal DNA elements unique to bacteria (4, 5). Plasmids are a major driver in the spread of ARGs in bacterial populations (6) and have recently been found to accelerate bacterial evolution by enhancing the adaptation of the bacterial chromosome (7).

Classifying plasmid types is crucial to understanding antibiotic resistance transmission between bacteria. Several recent studies have shown the benefit of whole genome sequencing for classifying plasmid types (8, 9). In particular, the emergence of long-read sequencing by Oxford Nanopore Technologies (ONT) promises improvements for outbreak investigations due to its lower capital investment and shorter turnaround times (10, 11). However, these methods suffer from the small proportion of plasmid DNA within the sequenced samples, primarily if bacteria can not be cultivated in the lab (12). Therefore, a large proportion of the plasmids in such samples are probably missed, or the sequencing depth is insufficient to assemble them correctly (13). Thus, additional sample preparation steps are required to isolate or enrich plasmids before DNA sequencing, but they are too expensive and laborious for applications in clinical diagnostic settings.

While nanopore sequencing has shown to accurately reconstruct plasmids (14), the technology offers a feature called adapative sampling (AS) that has the potential to improve plasmid classification. First described in 2016 by Loose et al., nanopore adaptive sampling has been increasingly used for *in-silico* target enrichment within the last two years. Here, DNA molecules can be rejected from individual nanopores if the corresponding sequence is not interesting for downstream analysis. Pulling out unwanted DNA frees the nanopore for the following molecule to be sequenced and reduces the time spent sequencing uninteresting DNA fragments. Different tools implement adaptive sampling (16–18), using dynamic time warping, read mapping, or k-mer-based strategies, all performing rejection decisions by analyzing the first approximately 500 base pairs of each read. The potential enrichment reached by using adaptive sampling was already shown, and even mathematical models that predict the enrichment factor were recently described by some groups (17, 19, 20). In one study, Marquet et al. could enrich the microbiome in human vaginal samples by depleting host DNA. Further, Kipp et al.

used adaptive sampling to enrich bacterial pathogens in tick samples, while Viehweger et al. even enriched single ARGs in human samples. However, no study has been conducted that shows the potential enrichment of plasmid sequences in bacterial isolates by depleting their chromosomal DNA sequences.

In the present study, we evaluate the performance of adaptive sampling to enrich plasmid sequences in five different bacterial species using two different adaptive sampling tools. Since laboratories try to decrease costs and increase sustainability, it is not unusual to reuse nanopore flow cells or sequence samples on expired flow cells. Hence, we also investigate whether enrichment of plasmids can be achieved when adaptive sampling is applied to such nanopore flow cells. By depleting the chromsomal reads, we show that an *in-silico* enrichment of 1.1-1.8x can be achieved with adaptive sampling even with reused or expired nanopore flow cells. We further show that adaptive sampling improves the quality of *de novo* plasmid assemblies through *in-silico* enrichment of low abundant plasmid sequences. Here, the plasmid assemblies after 2 hours of sequencing with adaptive sampling are as accurate as those without adaptive sampling after 3 hours of sequencing. Finally, we demonstrate that the read classification performance of the chosen adaptive sampling tool and read length used for classification both influence the potential enrichment of plasmids in species with conserved regions on both chromosome and plasmid.

Methods

Culture and DNA extraction

Campylobacter strains were streaked on Columbia Blood agar (Oxoid, Thermofisher Scientific, USA) and incubated at 42 °C under microaerobic atmosphere. *Enterobacter, Salmonella* and *Klebsiella* strains used in this study were streaked out on Luria Bertani (LB) plate and incubated over night at 37 °C. DNA extraction for *Campylobacter jejuni* (NCBI Refseq NZ_CP043763.1) was done using the MagAttract HMW Genomic Extraction Kit (Qiagen). For *Salmonella enterica*, *Campylobacter coli* (23), *Klebsiella pneumoniae* (NCBI BioSample SAMN30667246) and *Enterobacter hormaechei* (NCBI BioSample SAMN30667239) DNA was extracted using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instruction. The total amount of DNA was quantified using a Qubit fluorometer (Thermo Fisher Scientific) and frozen at -80ºC until further analysis.

Library preparation and sequencing

Sample preparation was performed according to the manufactures' instructions without any optional steps or size selection using the Rapid Barcoding Kit SQK-RBK004. The barcoded samples (pools of either two, three or four samples) were pooled in an equimolar manner and finally the DNA was sequenced on an Oxford Nanopore MinION (Oxford, UK) using FLO-MIN106D(R9.4.1) flow cells. All sequencing experiments were started via ONT's MinKNOW control software (version 4.5.0).

In-silico enrichment via adpative sampling

We performed five sequencing runs using MinKNOW software v4.5.0 on an Nvidia Jetson AGX Xavier (32GB) for at least 24 hours. In all five experiments, we compared adaptive sampling with standard sequencing by dividing the flow cells into two parts: Adaptive sampling was performed on the first 256 channels, and standard sequencing was performed on channels 257 to 512. We used a new flow cell with 1153 active pores for the first run (ReadBouncer1) and sequenced two multiplexed *Campylobacter* isolates. For the second run (ReadBouncer2), we used an expired flow cell with only 636 active pores for sequencing and multiplexed three bacterial isolates (*Enterobacter, Salmonella, Klebsiella*) for sequencing. The third run (MinKNOW1) used the same *Campylobacter* samples as the first, but we performed sequencing on an expired flow cell with only 557 active pores. For the fourth run (MinKNOW2), we used the identical three bacterial isolates as for the second run and performed sequencing on an expired flow cell with only 718 active pores after the initial flow cell check. Finally, the fifth run was performed on a flow cell with 725 active pores that was reused after a former SARS-CoV-2 amplicon sequencing experiment. For this run we multiplexed four bacterial isolates (*Campylobacter jejuni, Enterobacter, Salmonella, Klebsiella*) with the same setup as for the other four experiments. The reused flow cell was washed after the SARS-CoV-2 amplicon experiment with the EXP-WSH004 wash kit (ONT) and all raw reads from the bacterial sequencing run were mapped against the SARS-CoV2 reference sequence (NC_045512) with `minimap2` (24) to check for potential cross contamination from the first run of that flow cell. However, we could not detect any SARS-CoV2-related read in the final output, which indicates no cross contamination after flow cell washing.

On the first two flow cells, we performed adaptive sampling with ReadBouncer `ReadBouncer` (18) using the chromosomal references of the bacterial isolates as depletion target. Here, a k-mer size of 15, chunk length of 250 bp, fragment size of 200,000 bp and an expected error rate of 5% was used as parameters for the read classification. For flow cells three and four, `MinKNOW`'s adaptive sampling feature was used, which is based on the `Readfish` (17) scripts and uses `minimap2` (24) for mapping of reads prefixes against a given reference sequence set for read classification. For these experiments, we built a `minimap2` index file including the chromosomal reference sequences, which we used as depletion targets for adaptive sampling. Read prefixes classified as "chromosomal" were rejected from the pore, and decisions were written to log files by both tools, `ReadBouncer` and `MinKNOW`. In all experiments, `ReadBouncer` and `MinKNOW` used `guppy` GPU basecaller (fast model, v6.0.6. Oxford Nanopore Technologies) for real-time base calling of the raw signal data received from the device after at most 0.4 seconds of sequencing.

In order to investigate whether adaptive sampling could

Reference	Flow cell ID	Mean Length	Median Length	Std. Deviation
Campylobacter jejuni	ReadBouncer1	17,227.41	11,616	17,285.88
	MinKNOW1	14,208.21	9,060	15,421.39
Campylobacter coli	ReadBouncer1	4,670.54	1,313	7,473.24
	MinKNOW1	5,951.48	2,950	7,256.60
Salmonella enterica	ReadBouncer2	9,600.79	6,853	9,082.91
	MinKNOW2	8,833.55	6,126	8,648.58
Enterobacter hormaechei	ReadBouncer2	9,195.27	6,377	8,849.36
	MinKNOW2	8,457.37	6,071	8,190.38
Klebsiella pneumoniae	ReadBouncer2	9,516.86	6,690	9,185.76
	MinKNOW2	8,887.41	5,930	9,017.54

Table 1. Overview of read length metrics of the five bacterial isolates. The metrics were computed from reads sequenced on the control side of the flow cells where no adaptive sampling was applied.

be utilized on reused nanopore flow cells, we performed a fifth sequencing experiment with `ReadBouncer` as an adaptive sampling tool. We used the same `ReadBouncer` parameters and base caller settings as in the first two experiments, except that raw signal data were received from the device after at most 0.6 seconds of sequencing.

Data Analysis

After the sequencing runs were finished, we base called and demultiplexed all raw data with `Guppy` GPU basecaller (super accuracy model, v6.0.6. Oxford Nanopore Technologies). `Guppy` trimmed barcodes and adapter sequences from the resulting nanopore reads during that process. Afterwards, we computed read length metrics (see Table 1) and species abundances from the control regions (channels 257 to 512) using the sequencing_summary files provided with the `MinKNOW` and `Guppy` output directories. In the next step, we mapped all demultiplexed and base-called reads with a minimum length of 1000 nucleotides against the reference genomes of the five bacterial strains using `minimap2 v2.19` (24) with parameter `-x map-ont`. Based on the mapping results, we could assign each mapped read to either the bacterial chromosome or plasmid(s) of one of the bacterial isolates. We further separated the reads by their species of origin and partitioned them to comprise the cumulative data from the beginning of each experiment up to 24 hours, separated by 30 minutes of sequencing, which resulted in 48 individual timepoint data sets. With this information, we calculated for each experiment time point t the plasmid enrichment by yield for each bacterial strain,

$$Enrichment_{yield}(t) = \frac{yield_{AS}(plasmid, t)}{yield_{CTRL}(plasmid, t)} \quad \textbf{(1)}$$

where $yield_{AS}(plasmid, t)$ is the number of sequenced plasmid bases of a strain from the adaptive sampling region at time point t and $yield_{CTRL}(plasmid, t)$ is the number of sequenced plasmid bases of a strain from the control region (without adaptive sampling) at time point t.

We further compared the enrichment by yield of plasmid bases with the enrichment by the number of reads,

$$Enrichment_{reads}(t) = \frac{reads_{AS}(plasmid, t)}{reads_{CTRL}(plasmid, t)} \quad \textbf{(2)}$$

and the enrichment by the mean depth of coverage of the plasmid reference sequences.

$$Enrichment_{depth}(t) = \frac{depth_{AS}(plasmid, t)}{depth_{CTRL}(plasmid, t)} \quad \textbf{(3)}$$

According to the definitions above, $reads_{AS}(plasmid, t)$ and $reads_{CTRL}(plasmid, t)$ represent the number of reads from the adaptive sampling (AS) or control region (CTRL) that map to a plasmid of a given bacterial strain at experiment time point t. Further, $depth_{AS}(plasmid, t)$ denotes the mean sequencing depth of plasmids from a strain using mapping data from the adaptive sampling region at time point t and $depth_{CTRL}(plasmid, t)$ is the mean sequencing depth of plasmids on the control region at time point t. Here, we used `samtools coverage` (25) to calculate the mean depth of coverage of every species' plasmid reference at each time point for the control and adaptive sampling regions.

Since we expect plasmid sequences in our use case scenario to be usually unknown, we also did a *de novo* assembly of the demultiplexed fastq files that we cumulatively partitioned at the 48 individual timepoints. Thus, we would get an assembly out of the reads every 30 minutes of the experiment. This helps us to estimate the time required to obtain high-quality plasmid assemblies. Therefore, we assembled demultiplexed and quality filtered ($Q \geq 10$) nanopore reads with a minimum length of 1000 bases using `Flye/metaFlye` assembler (v2.9, parameter "–meta") (26, 27). Then, we polished the obtained `metaFlye` assemblies with one round of `Medaka consensus` (v.1.6.0, default parameters, model r941_min_sup_g507, Oxford Nanopore technologies) using the same nanopore read set. We assessed the quality of the final assemblies with `Quast` (v5.0.2) (28) and combined reported metrics like NGA50 and mean depth of coverage for each timepoint.

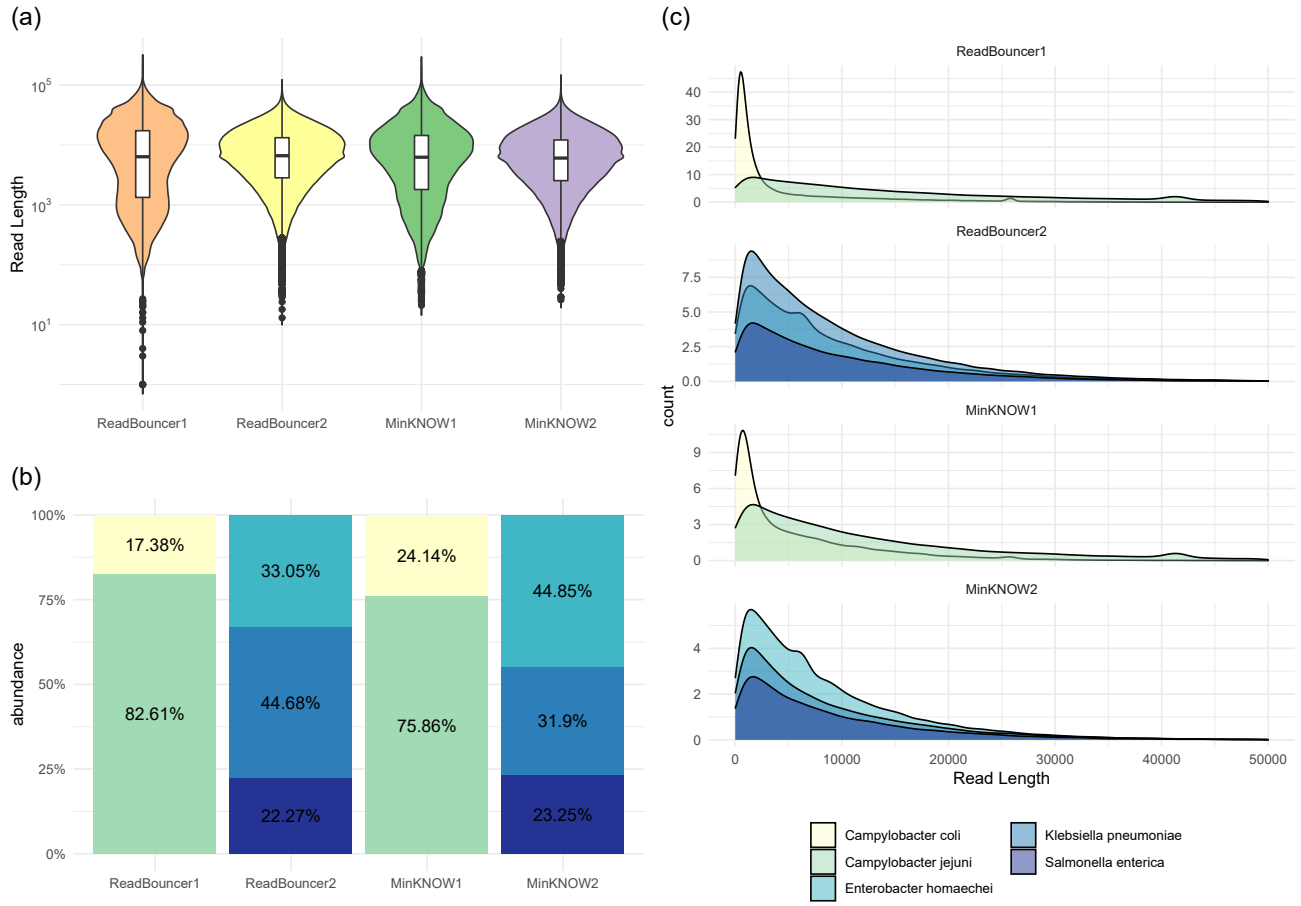

Fig. 1. Evaluation of control regions from the first four sequencing runs. (a) Violin plots (log scale) of read length distributions. Box plots for read lengths are included within the violin plots. **(b)** Species abundances for each of the four MinION runs. Similar species abundances can be observed between ReadBouncer1 and MinKNOW1 with *Campylobacter jejuni* dominating in both runs. Abundances are also similar for ReadBouncer2 and MinKNOW2 except that *Enterobacter hormaechei* is most abundant species in MinKNOW2 and *Klebsiella pneumoniae* is most abundant in ReadBouncer2. **(c)** Read length distributions for each species from control regions of the four sequencing runs.

Results

Similar read lengths, species and plasmid abundances of bacterial isolates across different experiments

In this study, we present the application of nanopore adaptive sampling on the *in-silico* enrichment of plasmids by depleting chromosomal reads during the sequencing of bacterial isolates. Therefore, we sequenced five bacterial strains - *Campylobacter jejuni*, *Campylobacter coli*, *Salmonella enterica*, *Enterobacter hormaechei* and *Klebsiella pneumoniae* - on five different flow cells, each divided into an adaptive sampling and a control region. This section will provide an overview of the generated data from the control region of the first four flow cells, with further details on plasmid abundance in every sample.

First, the quality filtered sequencing data after base calling from the control regions show a large difference in mean and median read lengths and the standard deviation for the two *Campylobacter* samples, pooled together on two of the four flow cells (see Table 1). Here, the application of the MagAttract HMW Genomic Extraction Kit (Qiagen) results in larger read lengths for the *Campylobacter jejuni* samples. We also see a difference between the read lengths of the same sample from the two *Campylobacter* sequencing runs (ReadBouncer1 and MinKNOW1), particularly for *Campylobacter coli*. However, the other three bacterial samples we pooled together show similar read lengths for both sequencing runs, with slightly shorter read lengths for samples sequenced on flow cell MinKNOW2.

In a second step, we look at the sample compositions of the four sequencing runs. In Figure 1 we see that in both sequencing runs *Campylobacter jejuni* dominates the output over *Campylobacter coli* with about $75 - 80\%$ of sequencing reads. On the other hand, we see a more even distribution of *Salmonella enterica*, *Enterobacter hormaechei* and *Klebsiella pneumoniae* in the other two sequencing runs. We further observe no considerable difference in species abundances between sequencing runs of the same samples. Additionally, we investigated the plasmid sequence abundances for each of the sequenced bacterial samples by mapping all quality filtered reads longer than or equal to 1,000 bases against the five reference genomes. Here we observe different abundances between 1 and 15% between the five species. However, the plasmid abundance is similar for every species between the sequencing runs (see Figure 3).

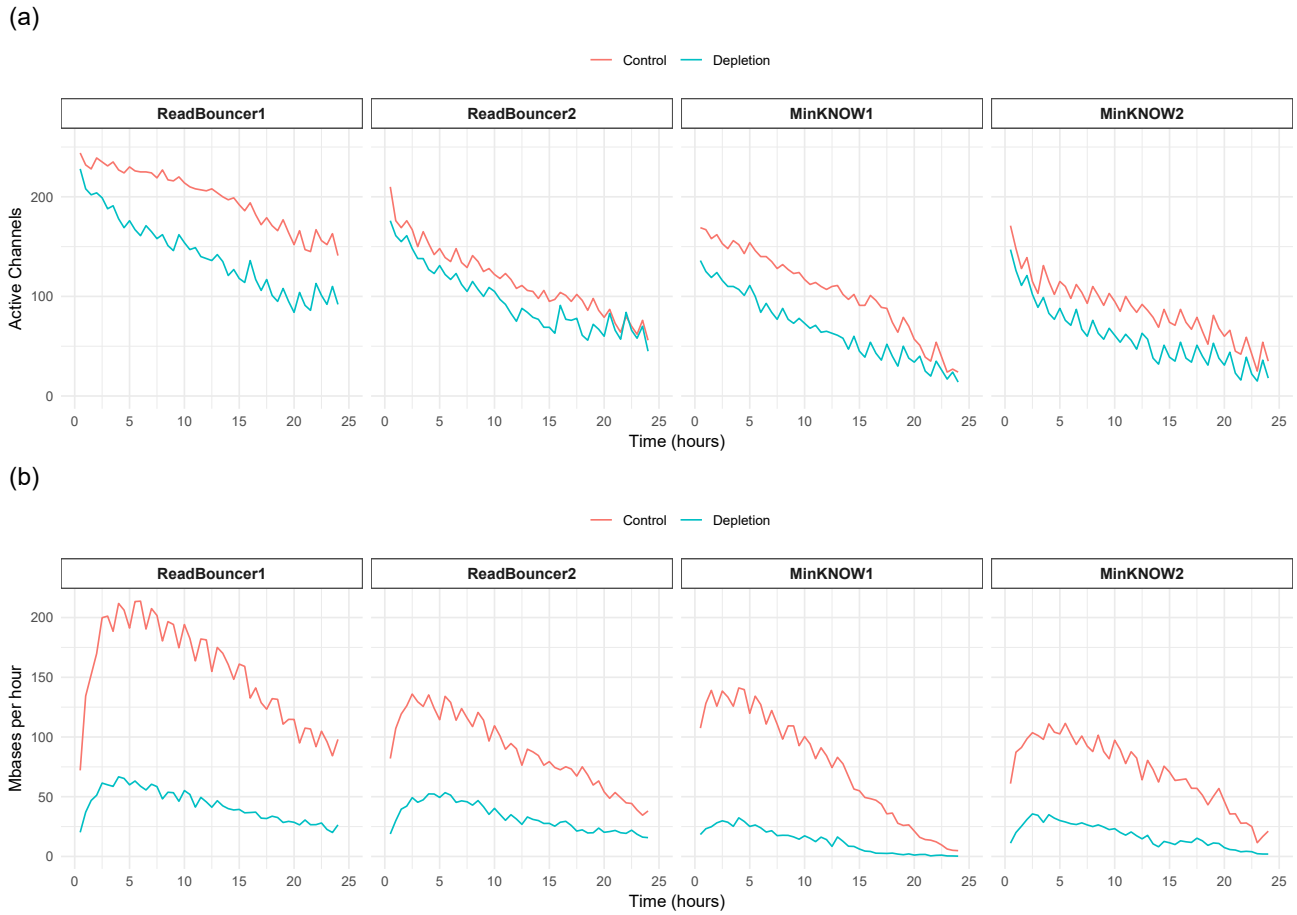

Fig. 2. Comparison of active channels and yield between control and adaptive sampling (depletion) regions in all four experiments. (a) Plots showing how the number of active channels varies with time in adaptive sampling (depletion) and control regions. There are more active sequencing channels in control regions on all four flow cells **(b)** Hourly yields from depleted channels vs control channels. Usage of adaptive sampling results in lower overall sequencing yield compared to normal sequencing.

Adaptive sampling reduces number active channels and sequencing yield

One of the major aspects of our study is that we wanted to assess whether adaptive sampling could also be applied on expired nanopore flow cells. In order to check if adaptive sampling leads to fast pore exhaustion on expired flow cells, we investigated the effect of adaptive sampling on the number of active sequencing channels and yield in sequenced Mbases per hour. Comparing the first four experiments, we consistently observe fewer active sequencing channels on flowcell regions with adaptive sampling than in control regions across all experiments. In summary, we find between 1.4 to 2.6 times more active channels in control regions than in adaptive sampling regions. However, we could not detect systematically bigger difference in active sequencing channels on expired flow cells when compared to the fresh flow cell. Our observations correspond to a higher sequencing yield in control regions than in adaptive sampling flow cell regions. Hence, our observations confirm that adaptive sampling results in less overall sequencing capacity (Payne et al.) and reduced yield from rejecting unwanted molecules (Martin et al.).

Depletion of chromosomal reads massively increases plasmid abundance

In our first four experiments, we investigate the potential enrichment of plasmid sequences in bacterial samples by depletion of the chromosomal reads using adaptive sampling. In Figure 3, we see that after 24 hours of sequencing, adaptive sampling massively increases the abundance of plasmid base pairs (bp) in the final output. For instance, we could increase the abundance of *Campylobacter jejuni* plasmid bases from 3.89% to 99.88% when depleting chromosomal reads with `MinKNOW`. We further observe that plasmid abundances are much higher when using `MinKNOW` instead of `ReadBouncer` for adaptive sampling, suggesting that `ReadBouncer` either had a much lower read classification sensitivity or took too long to reject unwanted chromosomal DNA molecules. Thus, we investigated the lengths of rejected reads in the final output ($end_reason =' data_service_unblock_mux_change'$) of the experiments by drawing histograms comparing lengths of reads rejected by `MinKNOW` with lengths of reads rejected by `ReadBouncer` (see Figure 6). In the histogram, we see that reads rejected by `ReadBouncer` are much longer than those rejected by `MinKNOW`, with an average length of 848 bp compared to 520 bp. This confirmed our

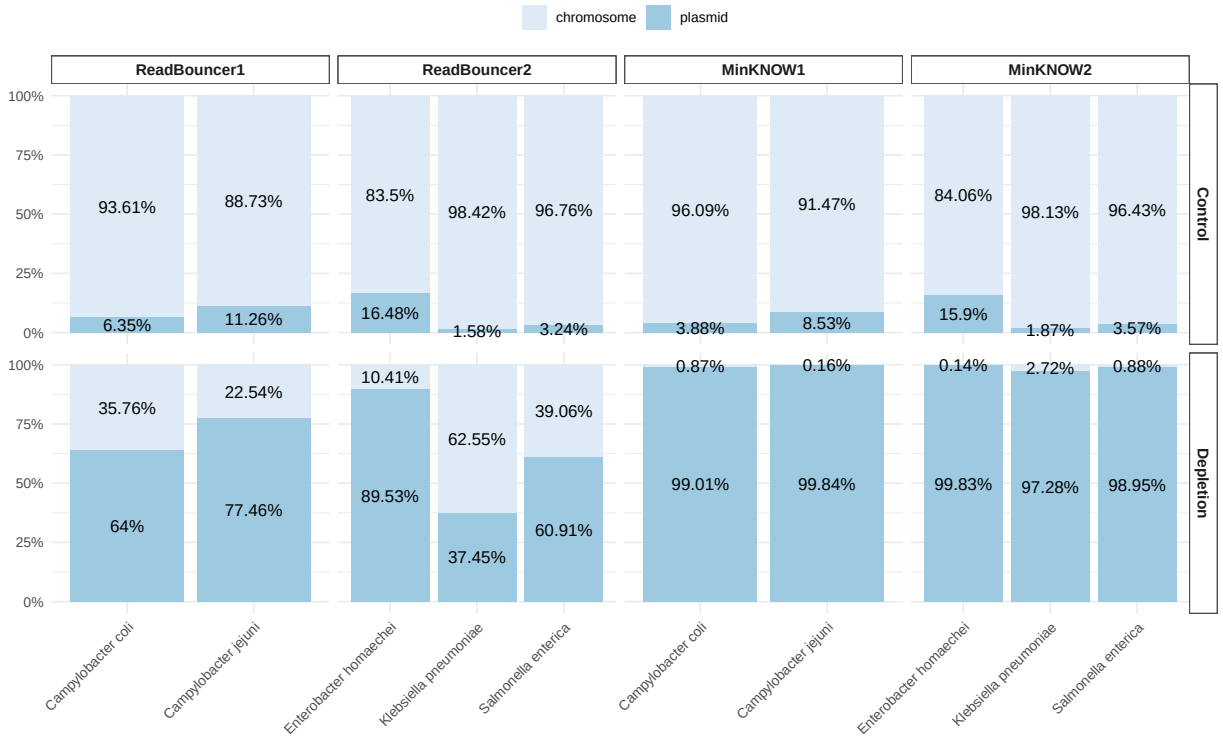

Fig. 3. Comparison of plasmid abundances in five bacterial samples. Adaptive sampling with MinKNOW was used on flowcells MinKNOW1 and MinKNOW2 and ReadBouncer was used as adaptive sampling tool on flowcells ReadBouncer1 and ReadBouncer2. For all experiments, plasmid abundances for each sample were measured after 24 hours of sequencing for control regions and adaptive sampling regions (Depletion). Plasmid abundances are highest when using MinKNOW for depletion of chromosomal nanopore reads.

assumption that `ReadBouncer` rejects reads too late during the adaptive sampling process resulting in a higher abundance of unwanted chromosome reads in the final output.

We also examined read classification metrics by combining information from read-until log files produced by both tools with mapping results of reads from the base called FASTQ output files (Table 2). Here, we see that `MinKNOW` shows a better sensitivity, while `ReadBouncer` has a much better specificity. Thus, `MinKNOW` is better at correctly rejecting unwanted chromosomal reads but also falsely rejects much more plasmid reads than `ReadBouncer`. Together with the longer rejected reads, this explains the difference in plasmid abundance between experiments using the two tools.

Consistent enrichment of plasmids by yield, read number and mean depth of coverage

We calculated the plasmid enrichment for the five bacterial species for the four experiments in 30-minute intervals. According to equation 1 the enrichment is the ratio of cumulated plasmid bases from the adaptive sampling region and the control region at time point t. Figure 4 illustrates that we obtain an enrichment of plasmid bases at any time point during the experiments for all but one sample. For the *Klebsiella pneumoniae* sample of flow cell MinKNOW2, we observe an enrichment $2.5 - 3.0$ hours after starting the experiment. At all other time points, the number of plasmid bases from adaptive sampling is lower than that from the control channels. Thus, we failed to obtain an enrichment of *Klebsiella pneumoniae* plasmids in that experiment where we used `MinKNOW` for the depletion of chromosomal reads. For all other samples, we observe an enrichment of 1.1x to 1.8x after 24 hours, which translates to $10 - 80\%$ more plasmid data when using adaptive sampling on flow cells with reduced active pores.

When comparing plasmid enrichment by yield, number of reads and mean depth of coverage, we see little differences between them for all sequenced samples. Independent on the tool used for adaptive sampling, we observe a peak enrichment of plasmids after $3 - 6$ hours for all samples. After reaching the peak, the enrichment decreases over time and stabilizes towards the end of the experiments. Although the number of active sequencing channels is higher for ReadBouncer1 than for MinKNOW1, we only see a slight difference in the enrichment of the *Campylobacter* strains when comparing `ReadBouncer` and `MinKNOW`. On the other hand, we see a higher enrichment of plasmids for *Salmonella enterica*, *Enterobacter hormaechei*

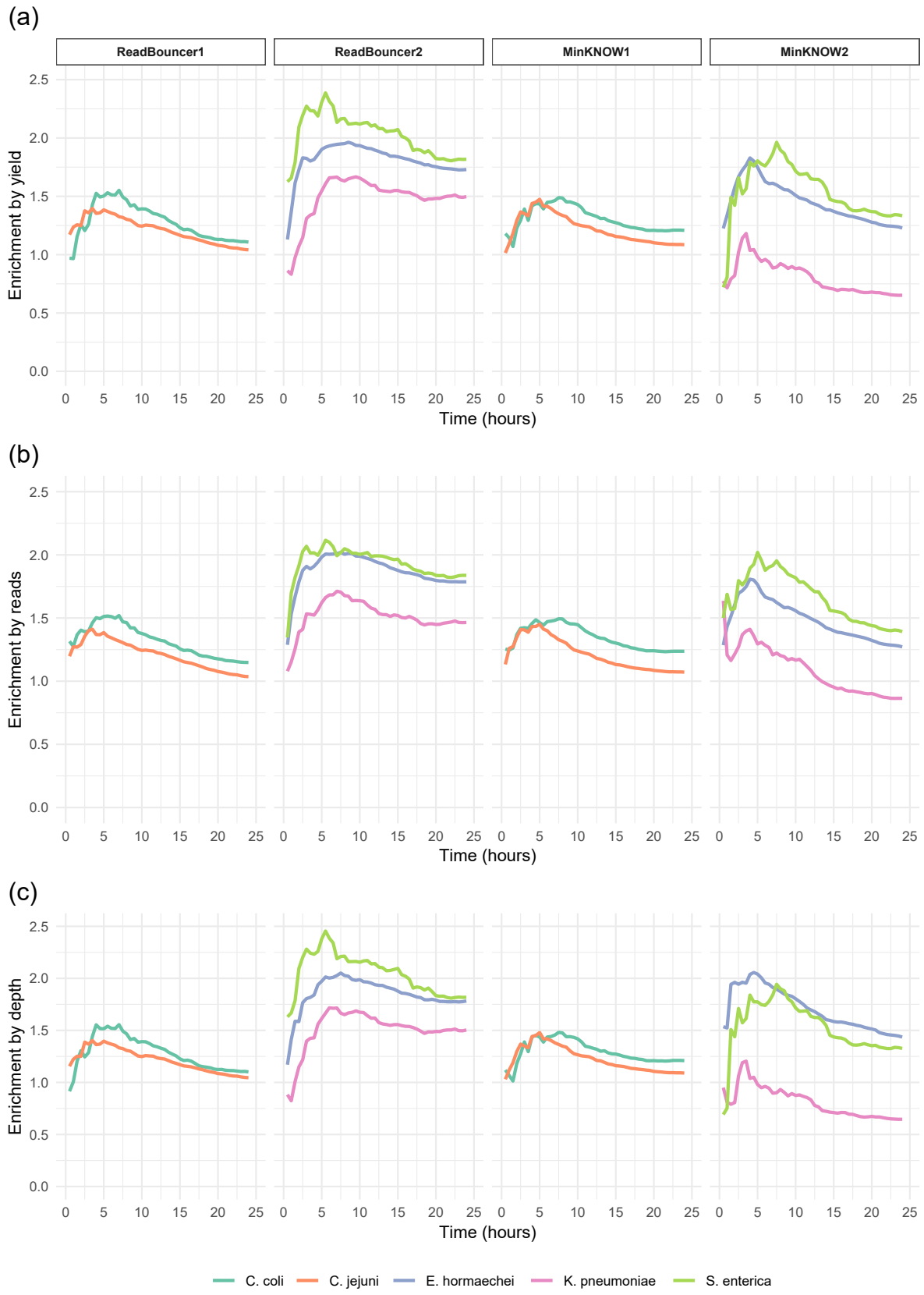

Fig. 4. Comparison of enrichment in five bacterial samples. (a) Enrichment by number of sequenced plasmid bases for all five bacterial strains across the four sequencing runs. **(b)** Enrichment by number of plasmid reads for the five bacterial strains across all four sequencing runs. **(c)** Enrichment by mean depth of coverage of plasmid references for the five bacterial strains across the four sequencing runs. All strains but the *Klebsiella pneumoniae* sample, where MinKNOW was used for adaptive sampling, show a slight enrichment.

	Accuracy	Precision	Recall	Specificity	F1-Score	Balanced Accuracy
All						
MinKNOW	**95.88**	98.50	**96.73**	90.12	**97.61**	93.43
ReadBouncer	93.48	**99.58**	93.33	**95.23**	96.35	**94.28**
Salmonella enterica						
MinKNOW	**96.87**	99.95	**96.84**	98.22	**98.37**	**97.53**
ReadBouncer	93.74	**99.97**	93.60	**99.08**	96.68	96.34
Enterobacter hormaechei						
MinKNOW	**95.21**	97.38	**96.86**	86.93	**97.12**	91.89
ReadBouncer	94.05	**99.18**	93.75	**95.72**	96.39	**94.73**
Klebsiella pneumoniae						
MinKNOW	**95.79**	99.16	**96.52**	61.02	**97.82**	78.77
ReadBouncer	92.83	**99.74**	92.96	**85.74**	96.23	**89.35**

Table 2. Comparison between MinKNOW and ReadBouncer on various read classification evaluation metrics. Nanopore reads from flow cells ReadBouncer2 and MinKNNOW2 were mapped against the three bacterial references using minimap2. Mapping results were matched with ReadUntil log files, which include for each read the unblock decisions made by the two adaptive sampling tools (`MinKNOW` and `ReadBouncer`). From this data, evaluation metrics for the read classification could be calculated for all and for each species separately. Overall, classification with `MinKNOW` shows higher accuracy, recall and F1-score, while `ReadBouncer` has better precision, specificity and balanced accuracy. `ReadBouncer` in particular shows much higher specificty and balanced accuracy when classifying plasmid reads of *Enterobacter hormaechei* and *Klebsiella pneumoniae*. Here, `ReadBouncer`'s classification approach outperforms `MinKNOW`'s mapping approach when regions with high sequence similarity are located on both target and depletion reference sequences.

and *Klebsiella pneumoniae* in the `ReadBouncer` adaptive sampling experiment. In particular, we reached an enrichment of 1.5x for *Klebsiella pneumoniae* plasmids even after 24 hours of sequencing.

Read classification performance impacts plasmid enrichment

As described in the subsection above, we observe a large difference in the enrichment of plasmids between the three bacterial strains sequenced on flow cells MinKNOW2 and ReadBouncer2. While some differences might be explained by the differing overall sequencing yield over time between control and adaptive sampling regions (see Table 2), it does not fully explain why adaptive sampling with `MinKNOW` results in plasmid depletion for *Klebsiella pneumoniae* after 5 hours of sequencing. Thus, we investigated the read classification performance of the two adaptive sampling tools by mapping all unfiltered reads against the three reference sequences of *Salmonella enterica*, *Enterobacter hormaechei* and *Klebsiella pneumoniae* with `minimap2` and combined the results with data from the log files of `ReadBouncer` and `MinKNOW`, which store unblock decisions for each read from the adaptive sampling regions. Based on the mapping data, we calculated the number of *true positives* (unblocked chromosome reads), *false positives* (unblocked plasmid reads), *true negatives* (plasmid reads not unblocked) and *false negatives* (chromosome reads not unblocked) for both adaptive sampling tools. Furthermore, we calculated overall read classification evaluation metrics from these four values and for each of the three bacterial isolates separately (see Table 2).

In general, `MinKNOW` shows higher accuracy, recall and F1-score, suggesting that it unblocks more chromosome reads correctly than `ReadBouncer`, which is also one of the explanations for the difference in final plasmid abundance between the two tools (see Figure 3). On the other hand, `ReadBouncer` shows better precision, specificity and balanced accuracy, suggesting that the probability of falsely unblocking plasmid reads is much lower with `ReadBouncer` than with `MinKNOW`. In particular, we observe a massive difference between the two tools when classifying reads from *Klebsiella pneumoniae*, where `ReadBouncer` shows a specificity of 85.74 and `MinKNOW` only 61.02.

To further examine whether the difference in specificity is due to the classification approach of the two tools used for classification, we extracted all plasmid reads of *Salmonella enterica*, *Enterobacter hormaechei* and *Klebsiella pneumoniae* that were falsely unblocked by `MinKNOW` and mapped them to the corresponding bacterial chromosome references with `minimap2`. Using `samtools depth`, we first identified four regions (between 829 and 2,101 bp long) on the *Klebsiella pneumoniae* chromosome with read depth ≥ 10, and 2 such regions on the *Enterobacter hormaechei* chromosome. These findings reveal regions of high identity between plasmid and chromosome for both bacterial strains, which explains the decreased classification specificity for `MinKNOW` on these isolates. In order to test whether `ReadBouncer` would also unblock plasmid reads matching these chromosomal regions, we used 250 bp prefixes of the unblocked plasmid reads and classified them with `ReadBouncer classify` using the same parameter values for k-mer size (15), fragment size (200,000) and expected error rate (5%) as in the adaptive sampling experiments performed in this study. We found that for *Klebsiella pneumoniae* only 321 out of 1,125 falsely unblocked plasmid reads would also be classified as chromosomal DNA by `ReadBouncer`. For the sequence of *Enterobacter hormaechei*, only 4 out of 2,647 falsely

	Salmonella enterica			*Enterobacter hormaechei*			*Klebsiella pneumoniae*		
Adaptive Sampling (channels 1-256)									
Time (hours)	1	2	3	1	2	3	1	2	3
Plasmid reads	73	190	321	853	2,404	4,219	77	187	318
NGA50	93,838	93,830	93,828	292,652	292,567	292,570	102,925	154,202	154,190
Ref. avg. coverage depth	9	23	40	12	37	65	3	8	16
Ref. coverage $\geq$ 5x(%)	98.17	100	100	89.55	99.84	99.84	22.91	93.53	97.53
Ref. coverage $\geq$ 10x(%)	50.39	100	100	36.98	99.81	99.84	0.0	34.10	89.10
Mismatches per 100kb	188	132	132	91	25	24	577	29	5
Indels per 100kb	676	658	655	130	25	24	722	78	21
Control (channels 257-512)									
Time (hours)	1	2	3	1	2	3	1	2	3
Plasmid reads	43	98	150	606	1,429	2,357	124	289	492
NGA50	91,692	93,851	93,828	109,532	292,998	292,565	103,578	153,446	154,178
Ref. avg. coverage depth	5	11	17	9	22	36	4	8	14
Ref. coverage $\geq$ 5x(%)	61.97	100	100	69.66	96.40	99.84	44.47	79.01	97.3
Ref. coverage $\geq$ 10x(%)	6.19	75.92	100	17.54	77.40	98.73	3.17	35.73	78.62
Mismatches per 100kb	263	145	129	190	35	25	161	177	45
Indels per 100kb	876	685	656	276	54	26	290	198	38

Table 3. Plasmid Assembly statistics of adaptive sampling and control region at three different time points of sequencing for three different species from sequencing run ReadBouncer2. All quality filtered reads from the adaptive sampling and control regions were separately assembled for *Salmonella enterica*, *Enterobacter hormaechei* and *Klebsiella pneumoniae* using `metaFlye` assembler. At all time points, assembly statistics provided by `Quast` show better results for plasmid assemblies in adaptive sampling regions. For *Salmonella enterica* and *Enterobacter hormaechei*, plasmid assemblies with adaptive sampling after two hours are even better than plasmid assemblies from the control region after three hours.

unblocked plasmid reads are classified as chromosomal reads by `ReadBouncer`, confirming a higher specificity of `ReadBouncer`'s classification approach when regions with high identity are located on both target and depletion reference sequences.

Unblocked read lengths can influence plasmid enrichment

We showed that read classification performance impacts plasmid enrichment if regions with high sequence similarity are located on both the bacterial chromosome and one of the plasmids. Since read lengths influence the read classification performance and rejected read lengths were different depending on the adaptive sampling tool we used in the first four experimental runs, we examined whether decreasing rejected read lengths with `ReadBouncer` affects plasmid enrichment. Therefore, we re-sequenced four of our five samples, for which we had material, with `ReadBouncer` as an adaptive sampling tool. Using the same experimental setup as in the other two `ReadBouncer` runs, we only changed the $break_reads_after_seconds$ parameter within the sequencing control software to 0.6 seconds. This resulted in a less skewed read length distribution with a mean length of rejected reads of 690 bp (see Figure 5 (e)). Although we improve accuracy and recall, we see a drop in read classification specificity from 85.74 to 70.45 for *Klebsiella pneumoniae* when using shorter read prefixes. Further, the enrichment decreased by yield from 1.5 to 1.1 for *Klebsiella pneumoniae* plasmid base pairs in comparison to the first plasmid enrichment experiment with `ReadBouncer`. However, we also see a drop in the enrichment factor of $0.4 - 0.5$ for *Salmonella enterica* and *Enterobacter hormaechei* between the first and the second experiment with `ReadBouncer`. Since we do not see a large difference in classification performance for these two bacterial strains between the two experiments, we cannot estimate how significant the influence of the classification performance on enrichment by yield is.

Adaptive sampling improves plasmid assembly quality

In order to evaluate the effect of adaptive sampling on the assembly of low abundant plasmids, we took the reads available at one-hour intervals from the adaptive sampling and control regions of the sequencing run ReadBouncer2. We did not include the *Campylobacter* strains from sequencing run ReadBouncer1 because the coverage of plasmids from sequencing only two isolates on a fresh flow cell was extremely high which led to similar assembly statistics from control and adaptive sampling regions. We separately assembled all quality filtered and demultiplexed reads with a minimum length of 1,000 bp for *Salmonella enterica*, *Enterobacter hormaechei* and *Klebsiella pneumoniae* using `metaFlye` assembler (27). After one round of polishing with `Medaka consensus`, we measured quality metrics for the final assemblies using `Quast` (28). Table 3 shows results for the first three hours into sequencing the three bacterial isolates from sequencing run ReadBouncer2. At all time points, the final plasmid assemblies from the adaptive sampling regions show higher accuracies than those

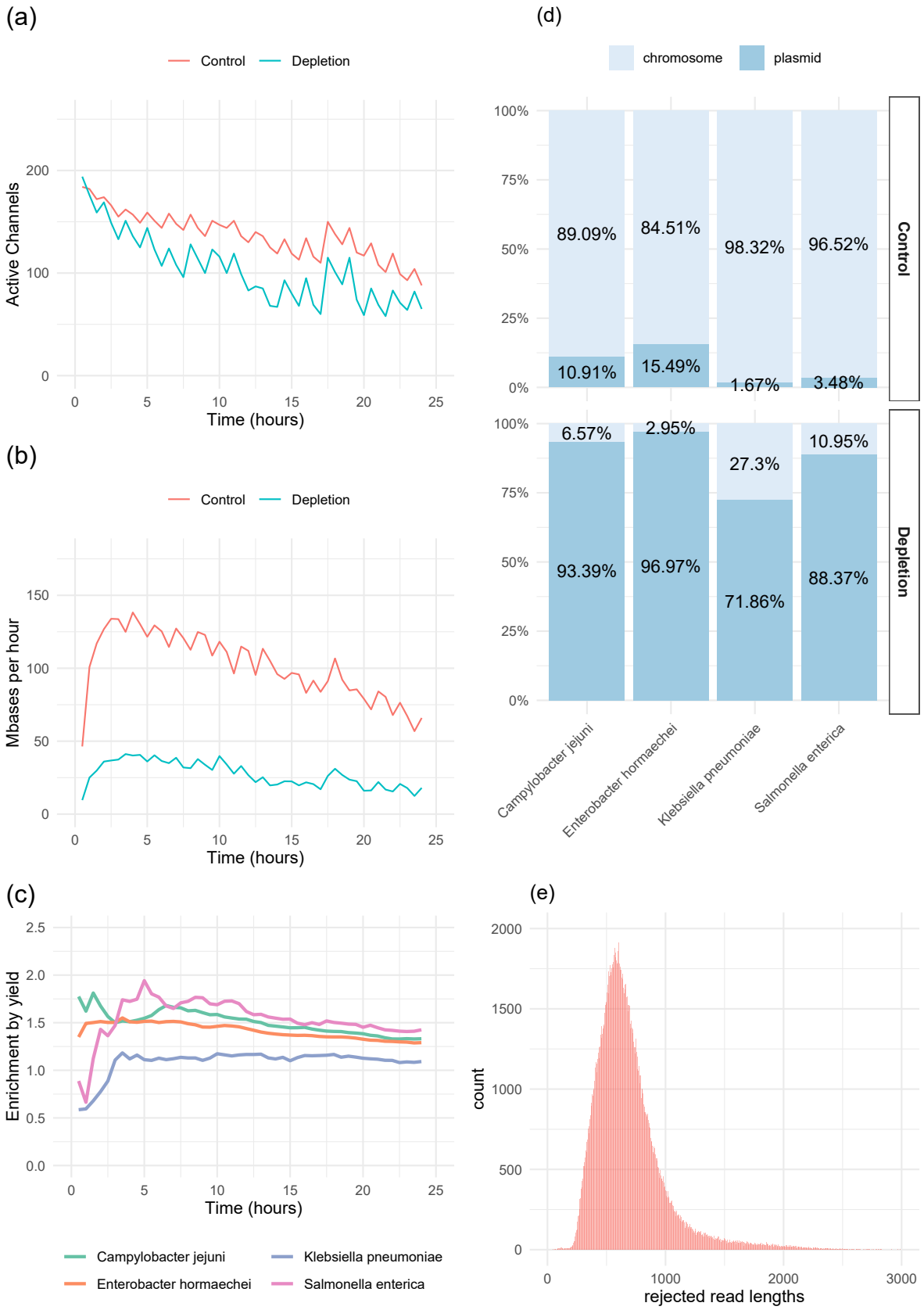

Fig. 5. Overview of results from third adaptive sampling run with ReadBouncer. **(a)** Comparison of the number of active sequencing channels over time for adaptive sampling and control reagion. **(b)** Hourly comparison of yield between adaptive sampling and control region. **(c)** Enrichment factor by yield over time for each of the four sequenced bacterial strains. **(d)** Comparison of plasmid and chromosomal sequence abundance between adaptive sampling and control region for each sequenced bacterial strain. **(e)** Read length histogram of rejected reads.

	Accuracy	Precision	Recall	Specificity	F1-Score	Balanced Accuracy
All	95.77	99.50	95.76	95.81	97.60	95.79
Salmonella enterica	96.33	99.99	96.25	99.62	98.08	97.93
Enterobacter hormaechei	96.17	99.54	95.98	97.35	97.73	96.66
Klebsiella pneumoniae	95.22	99.34	95.75	70.45	97.51	83.10

Table 4. Read classification metrics from third ReadBouncer adaptive sampling run. Nanopore reads from fifth flowcell run using ReadBouncer for adaptive sampling were mapped against the three bacterial references using minimap2. Mapping results were matched with ReadUntil log files, which include for each read the unblock decisions made by ReadBouncer. From this data, evaluation metrics for the read classification could be calculated for all and for each species separately.

from the control region. For *Salmonella enterica* and *Enterobacter hormaechei*, plasmid assemblies with adaptive sampling after two hours show even better results than plasmid assemblies from the control region after three hours. For *Klebsiella pneumoniae* the plasmid assembly from the adaptive sampling region after two hours of sequencing is as least as good as the assembly from the control region after three hours.

Discussion

Recent studies have demonstrated the utility of adaptive sampling for the enrichment of underrepresented sequences in various applications, such as host depletion in human vaginal samples or antibiotic resistance gene enrichment in metagenomics samples. In this study, we examine the potential of adaptive sampling for the enrichment of low-abundant plasmid sequences by depletion of chromosomal sequences in bacterial isolate samples. We demonstrate the possibility of using even older or expired flow cells with a reduced number of active sequencing pores for the *in-silico* enrichment via adaptive sampling. We further show that different adaptive sampling tools, namely `ReadBouncer` and ONT's `MinKNOW` sequencing control software, consistently enrich for low abundant sequences. The enrichment by yield, the most critical value for researchers, lies for all but one sample in our experiments between 1.1x and 1.8x after 24 hours of sequencing on an ONT MinION sequencing device. We also demonstrated that the difference between enrichment by yield, number of reads and mean depth of coverage is negligible in all our samples. High-quality assemblies of plasmids are possible within two hours of sequencing with adaptive sampling and show even better results than plasmid assemblies after three hours of standard nanopore sequencing. These results reflect the benefit of adaptive sampling in assembling low abundant plasmid sequences. Since we sequenced three bacterial isolates on only half a reused flow cell, we reason that up to 20 bacterial isolates can be sequenced on a flow cell with adaptive sampling for plasmid enrichment.

Our experiments showed that expired or reused flow cells with decreased number of active pores could be used in combination with adaptive sampling. Previous studies demonstrated that the number of active sequencing pores decreases faster when using adaptive sampling. Although we show the same trend in our study, we do not see a negative impact on the enrichment of target sequences. Thus, we encourage researchers to use expired or reused flow cells in adaptive sampling experiments where overall sequencing yield is not crucial for the experiment's success. This leads to more sustainable lab experiments and cost savings for core facilities and larger research institutions.

Our results exhibit that the depletion of chromosomal sequences with adaptive sampling increases the abundance of plasmid sequences in the final output. Furthermore, a consistent enrichment of plasmid sequences with regard to the number of base pairs, reads and depth of coverage was shown by using adaptive sampling. Independent of the size of the sequencing libraries, we could see a plasmid enrichment of 1.1x to 1.8x after 24 hours of sequencing. However, in one experiment, we recognized even depletion of plasmid sequences of *Klebsiella pneumoniae* after 24 hours when ONT's `MinKNOW` was used as an adaptive sampling tool. Our investigations reveal that regions with high sequence identity located both on the chromosome and the plasmid lead to lots of false read rejections, which result in a depletion of the targeted plasmid sequences. Comparing both adaptive sampling tools examined in this study, `ReadBouncer` handles such problematic regions better than ONT's `MinKNOW` by having a higher read classification specificity. However, the length of reads at the moment of classification also influences the classification specificity, suggesting that changing the corresponding parameters in `MinKNOW` could decrease the number of falsely rejected reads and thus improve the enrichment of plasmids.

Both adaptive sampling tools used in this study need known reference sequences for the depletion of the chromosomal reads. If the bacterial species in the given sample are unknown, a more extensive reference database of all potential bacterial chromosome references must be used to enrich plasmids successfully. Alternatively, researchers could also do a targeted enrichment of the plasmids by using plasmid databases such as PLSDB (29) and deplete all reads that do not match the database. However, this approach risks missing unknown plasmids not covered by the database. Using specific plasmid markers, like the origin of replication, for correctly classifying unknown plasmids is also tricky in an adaptive sampling experiment. Here, the specific markers would need to be located on the first 1000 bp of the read to prevent false rejection of plasmid reads. These limitations reinforce the need for improved classification algorithms that can even classify reads from unknown plasmids based on the raw nanopore signals.

Overall, our results illustrate the application of adaptive

sampling on expired nanopore flow cells for *in-silico* enrichment of bacterial plasmids. Given an enrichment of 50% of plasmid base pairs, as shown in this study, it is likely that expensive and time-consuming wet lab methods for plasmid enrichment can be omitted, which leads to cost savings and permits the sequencing of more plasmids from pathogenic bacteria.

We envisage several applications for the *in-silico* enrichment of plasmids in the near future. One possibility is the surveillance of plasmid outbreaks in hospital settings. Here, clinicians are interested in the trasmission of specific antibiotic resistance genes harboring plasmids from one bacterial species to another. Such community transmissions can indicate the selection pressure on bacteria caused by antibiotic pharmaceuticals and help decide on the corresponding drugs' future usage.

Another possible application of adaptive sampling is the improvement of known bacterial assemblies. In this study, we demonstrated the improved time-to-assembly of plasmids by depleting the known bacterial chromosomes. We plan to develop a pipeline for the real-time *de novo* assembly of bacterial isolates in the future. Using adaptive sampling, we could reject reads that cover assembled regions with a minimum depth of coverage, enriching for unseen or assembled regions with low sequencing depth. In such a way, we could complement the dynamic resequencing framework BOSS-RUNS (30) with a dynamic *de novo* adaptive sampling framework. We believe that this could improve the quality of bacterial and plasmid assemblies as well as metagenomics assemblies of unknown bacterial species.

Conclusions

Using nanopore adaptive sampling, we demonstrated an *in-silico* enrichment of low abundant plasmid sequences in known bacterial isolates by depleting chromosomal sequences. We discovered that enrichment could also be achieved with reused or expired flow cells, which in combination with adaptive sampling shows the potential for cost savings in clinical and laboratory sequencing of bacterial pathogens. The combination or flow cell re-usage, which can save approximately 500 Euro per flow cell, and omittance of time-consuming wet lab enrichment of plasmids make the application of plasmid sequencing more attractive for routine clinical diagnostics.

We also observed that conserved regions between depletion and enrichment targets pose a challenge for adaptive sampling tools that use mapping algorithms for read classification and that the read lengths used for the classification influence the level of enrichment in such scenarios. By performing targeted enrichment of low abundant plasmids, we significantly improved the quality of *de novo* assemblies of the plasmids and reduced the sequencing time needed to achieve this quality. We conclude that nanopore adaptive sampling will be useful for many bacterial plasmid sequencing studies.

Availability of data and materials

Scripts and software used in the analysis are available in the github repository https://github.com/JensUweUlrich/PlasmidEnrichmentScripts, as described in the methods.

The sequence datasets generated and analysed during the current study are available in the NCBI BioProject repository under accession number PRJNA862336.

Funding

This work was funded by the German Federal Ministry of Education and Research (BMBF) in the Computational Life Science program (Live-DREAM, 031L0175B) and has been supported by a grant from the BMBF/German Center for Infection Research (TI 06.904 - FP2019 to BYR). LE was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-number 425959793.

Conflict of Interest

JUU and BYR have filed a patent application on selective nanopore sequencing approaches.

ACKNOWLEDGEMENTS

The authors thank The authors thank Martin Hölzer, Matthew Huska and Aleksandar Radonic (RKI) for valuable discussions and comments on the usage nanopore adaptive sampling. We thank the Genome Sequencing Unit at Robert Koch Institute for sequencing of the bacterial strains.

Bibliography

1. Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, Eve Wool, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 2022.
2. J O'Neil. Tackling a crisis for the health and wealth of nations. *World Health Organization*, 2014.
3. Jim O'Neill. Tackling drug-resistant infections globally: final report and recommendations. 2016.
4. Sally R Partridge, Stephen M Kwong, Neville Firth, and Slade O Jensen. Mobile genetic elements associated with antimicrobial resistance. *Clinical microbiology reviews*, 31(4): e00088–17, 2018.
5. OS Gonçalves, JN Rosa, KC França, JT Bossé, MF Santana, Paul Richard Langford, DMS Bazzolli, et al. Mobile genetic elements drive antimicrobial resistance gene spread in pasteurellaceae species. *Frontiers in microbiology*, 12:773284–773284, 2021.
6. Alessandra Carattoli. Plasmids and the spread of resistance. *International journal of medical microbiology*, 303(6-7):298–304, 2013.
7. Jerónimo Rodríguez-Beltrán, Javier DelaFuente, Ricardo Leon-Sampedro, R Craig MacLean, and Alvaro San Millan. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, 19(6):347–359, 2021.
8. Alex Orlek, Nicole Stoesser, Muna F Anjum, Michel Doumith, Matthew J Ellington, Tim Peto, Derrick Crook, Neil Woodford, A Sarah Walker, Hang Phan, et al. Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Frontiers in microbiology*, 8:182, 2017.
9. Laura Hidalgo, Mark de Been, Malbert RC Rogers, Anita C Schürch, Jelle Scharringa, Anneke van der Zee, Marc JM Bonten, and Ad C Fluit. Sequence-based epidemiology of an oxa-48 plasmid during a hospital outbreak. *Antimicrobial Agents and Chemotherapy*, 63(12):e01204–19, 2019.
10. Tonya L Taylor, Jeremy D Volkening, Eric DeJesus, Mustafa Simmons, Kiril M Dimitrov, Glenn E Tillman, David L Suarez, and Claudio L Afonso. Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Scientific reports*, 9(1):1–11, 2019.
11. Joep JJM Stohr, Jaco J Verweij, Anton GM Buiting, John WA Rossen, and Jan AJW Kluytmans. Within-patient plasmid dynamics in klebsiella pneumoniae during an outbreak of a carbapenemase-producing klebsiella pneumoniae. *PLoS One*, 15(5):e0233313, 2020.
12. Aya Brown Kav, Itai Benhar, and Itzhak Mizrahi. A method for purifying high quality and high yield plasmid dna for metagenomic and deep sequencing approaches. *Journal of microbiological methods*, 95(2):272–279, 2013.
13. Michael DJ Lynch and Josh D Neufeld. Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*, 13(4):217–229, 2015.
14. Ryan R Wick, Louise M Judd, Kelly L Wyres, and Kathryn E Holt. Recovery of small plasmid sequences via oxford nanopore sequencing. *Microbial genomics*, 7(8), 2021.

15. Matthew Loose, Sunir Malla, and Michael Stout. Real-time selective sequencing using nanopore technology. *Nature methods*, 13(9):751–754, 2016.
16. Sam Kovaka, Yunfan Fan, Bohan Ni, Winston Timp, and Michael C Schatz. Targeted nanopore sequencing by real-time mapping of raw electrical signal with uncalled. *Nature Biotechnology*, 39(4):431–441, 2021.
17. Alexander Payne, Nadine Holmes, Thomas Clarke, Rory Munro, Bisrat J Debebe, and Matthew Loose. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature biotechnology*, 39(4):442–450, 2021.
18. Jens-Uwe Ulrich, Ahmad Lutfi, Kilian Rutzen, and Bernhard Y Renard. ReadBouncer: precise and scalable adaptive sampling for nanopore sequencing. *Bioinformatics*, 38 (Supplement_1):i153–i160, 06 2022.
19. Samuel Martin, Darren Heavens, Yuxuan Lan, Samuel Horsfield, Matthew D Clark, and Richard M Leggett. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome Biology*, 23(1):1–27, 2022.
20. Adrian Viehweger, Mike Marquet, Martin Hölzer, Nadine Dietze, Mathias Pletz, and Christian Brandt. Adaptive nanopore sequencing on miniature flow cell detects extensive antimicrobial resistence. *bioRxiv*, 2021.
21. Mike Marquet, Janine Zöllkau, Jana Pastuschek, Adrian Viehweger, Ekkehard Schleußner, Oliwia Makarewicz, Mathias W Pletz, Ralf Ehricht, and Christian Brandt. Evaluation of microbiome enrichment and host dna depletion in human vaginal samples using oxford nanopore's adaptive sequencing. *Scientific reports*, 12(1):1–10, 2022.
22. Evan J Kipp, Laramie L Lindsey, Benedict S Khoo, Christopher Faulk, Jonathan D Oliver, and Peter A Larsen. Enabling metagenomic surveillance for bacterial tick-borne pathogens using nanopore sequencing with adaptive sampling. *BioRxiv*, 2021.
23. Nicole Tegtmeyer, Delara Soltan Esmaeili, Irshad Sharafutdinov, Jakob Knorr, Michael Naumann, Thomas Alter, and Steffen Backert. Importance of cortactin for efficient epithelial nf-κb activation by helicobacter pylori, salmonella enterica and pseudomonas aeruginosa, but not campylobacter spp. *European Journal of Microbiology and Immunology*, 11(4):95–103, 2022.
24. Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18): 3094–3100, 2018.
25. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.
26. Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, 37(5):540–546, 2019.
27. Mikhail Kolmogorov, Derek M Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy PL Smith, et al. metaflye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17(11):1103–1110, 2020.
28. Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.
29. Georges P Schmartz, Anna Hartung, Pascal Hirsch, Fabian Kern, Tobias Fehlmann, Rolf Müller, and Andreas Keller. Plsdb: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Research*, 50(D1):D273–D278, 2022.
30. Nicola De Maio, Charlotte Manser, Rory Munro, Ewan Birney, Matthew Loose, and Nick Goldman. Boss-runs: a flexible and practical dynamic read sampling framework for nanopore sequencing. *bioRxiv*, 2020.

Supplementary Data

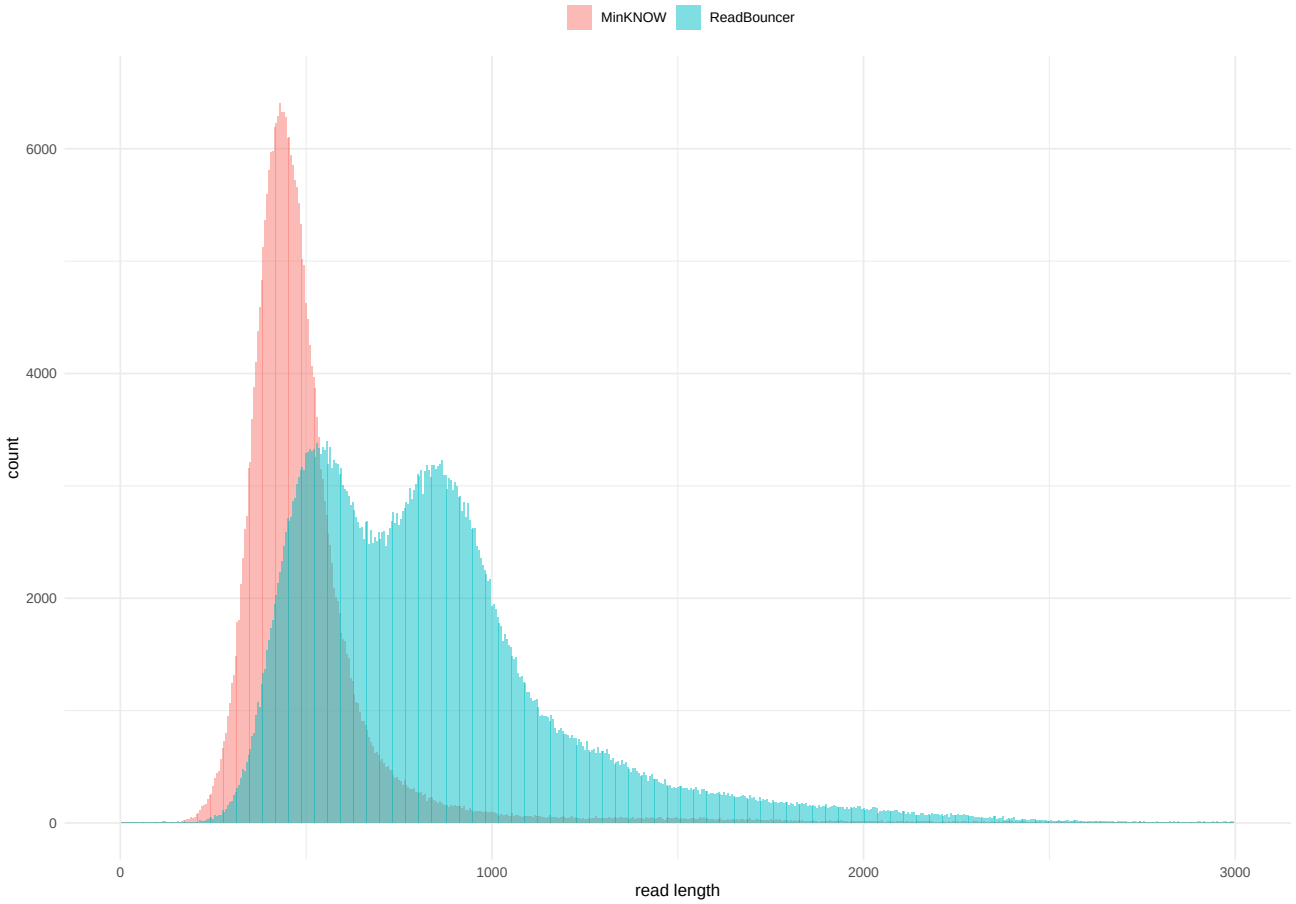

Fig. 6. Read length histogram of rejected reads by MinKNOW and ReadBouncer. Adaptive sampling with MinKNOW leads to rejected read lengths of about 450 to 500 nucleotides while lots of rejected reads are even longer than 1,000 nucleotides when using ReadBouncer