Computational and Structural Biotechnology Journal 18 (2020) 1342–1351

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

SEQU-INTO: Early detection of impurities, contamination and off-targets (ICOs) in long read/MinION sequencing

Markus Joppich [a,1], Margaryta Olenchuk [a,1], Julia M. Mayer [a,1], Quirin Emslander [b], Luisa F. Jimenez-Soto [c], Ralf Zimmer [a,*]

[a] LFE Bioinformatics, Department of Informatics, Ludwig-Maximilians-Universität München, 80333 München, Germany
[b] Physics of Synthetic Biological Systems, Physics Department, Technische Universität München, 85748 Garching, Germany
[c] Walther Straub Institute for Pharmacology and Toxicology, Ludwig-Maximilians-Universität München, Goethestrasse 33, 80336 München, Germany

ARTICLE INFO

Article history:
Received 13 January 2020
Received in revised form 4 May 2020
Accepted 9 May 2020
Available online 23 May 2020

Keywords:
Bioinformatics
User-friendly
Contamination
Nanopore Sequencing

ABSTRACT

The MinION sequencer by Oxford Nanopore Technologies turns DNA and RNA sequencing into a routine task in biology laboratories or in field research. For downstream analysis it is required to have a sufficient amount of target reads. Especially prokaryotic or bacteriophagic sequencing samples can contain a significant amount of off-target sequences in the processed sample, stemming from human DNA/RNA contamination, insufficient rRNA depletion, or remaining DNA/RNA from other organisms (e.g. host organism from bacteriophage cultivation). Such impurity, contamination and off-targets (ICOs) block read capacity, requiring to sequence deeper. In comparison to second-generation sequencing, MinION sequencing allows to reuse its chip after a (partial) run. This allows further usage of the same chip with more sample, even after adjusting the library preparation to reduce ICOs. The earlier a sample's ICOs are detected, the better the sequencing chip can be conserved for future use. Here we present *sequ-into*, a low-resource and user-friendly cross-platform tool to detect ICO sequences from a predefined ICO database in samples early during a MinION sequencing run. The data provided by sequ-into empowers the user to quickly take action to preserve sample material and chip capacity. sequ-into is available from https://github.com/mjoppich/sequ-into

1. Introduction

Long-read sequencing is rapidly evolving as a common practice in molecular biology. In 2018 more than 130 articles mentioning *MinION* or 280 articles mentioning *PacBio* have been published. Great advances have been made in terms of feasibility, cost, throughput, and read-length, now delivering single bacterial reads of more than one million base-pairs in length [1]. Oxford Nanopore (MinION) sequencing is becoming more and more popular with diverse applications like plant pathogen identification [2], virology [3], or botany [4]. One of its major advantages is portability, allowing in-the-field sequencing, e.g. screening for pathogens [5] or new species under arctic conditions [6] - and even on the International Space Station [7].

One of the most important requirements of successful sequencing is the sample purity, whether in the lab or out in the field. However, samples containing off-target reads are still common [8,9]. A reduced number of target sequences complicates correct, high-quality downstream analysis of sequencing data. Low number of target reads may, for instance, effect transcriptomic analyses (e.g. differential expression), or reduce the evidence for specific splice isoforms. On a genomic scale, it has been reported that (public) genome assemblies contain sequences highly likely originating from contamination[9].

Particularly with MinION sequencing, the sequencing time is not fixed: a run can be aborted at any time or new material can be added for sequencing. Thus, the general success criterion of a sequencing experiment might not be the total yield of (on-target) sequences, but instead the detection (or absence) of certain target sequences. An interactive analysis of the sequenced reads can be of help to decide whether a sequencing run can be successfully concluded, or should be aborted because it will not yield the necessary data of the intended target in the required quality.

* Corresponding author.
E-mail addresses: joppich@bio.ifi.lmu.de (M. Joppich), zimmer@bio.ifi.lmu.de, ralf.zimmer@ifi.lmu.de (R. Zimmer).
[1] Contributed equally.

https://doi.org/10.1016/j.csbj.2020.05.014
2001-0370/© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

Several tools have been developed since the public introduction of the MinION sequencer in 2012. Among these are NanoOK [10], RUBRIC [11], What's in my Pot (WIMP) [12] and npAnalysis [13].

Each of these serves a particular problem. NanoOK is a toolkit to assess descriptive statistics from MinION sequencing runs. With RUBRIC, selective sequencing can be performed by ejecting unwanted sequences from the pore, requiring a complex dual-computer setup. WIMP is built into Metrichor, which requires a paid subscription. Finally, npAnalysis provides a streaming server for Nanopore Sequencing reads, which is capable of detecting sequenced organisms, similar to WIMP. While this allows an online analysis, the setup and usage are rather sophisticated.

During the preparation of DNA or RNA for sequencing, several steps, including enzymatic reactions, can hamper the quality of the samples, e.g., inefficient rRNA depletion. In metagenomics, success is determined by the choice of correct and efficient primers [14]. In both cases, the detection of off-target sequences or specific organisms could be done directly while sequencing, right after the first actual reads of the sample are available. The sooner ICO sequences in the sample are detected, the more chip capacity can be rescued for further use.

Here we describe the applicability of sequ-into for online detection of sample ICOs during the sequencing run, or also after the sequencing has been concluded. sequ-into provides an online, descriptive overview of the sequenced reads, cross-platform compatibility (Windows, MacOS, Linux) and an easy installation combined with a graphical user-interface on a typical laptop computer. Using state-of-the-art long-read alignments, sequ-into can be of great help for on-/off-target analysis when performing laboratory protocol optimization, enabling a rapid assessment of sequenced reads. It has the capability to add genomes of interest which can be specifically targeted during analysis. By providing a descriptive overview of the sequencing run and its alignment to the selected on/off-target references, sequ-into allows an easily comprehensible and sharable analysis of a sequencing run. Such a setup reflects many real-world scenarios, including the more widespread in-the-field-usage of the MinION device.

2. Material & methods

2.1. Sequencing data

The biological samples were prepared as described in the supplementary information. The sequencing time and yield has been different per sample and is summarized in Table 2. Additional external phage DNA reads have been downloaded from EMBL EBI under accession id `PRJEB8318` (Jain et al. [15]).

2.2. Read extraction

Only basecalled FAST5 reads can be used for read extraction. Thus, the read extraction script (`extract_fast5.py`) of sequ-into relies on the live-basecalling functionality of MinKNOW. It can extract basecalled sequences from one or more (e.g. demultiplexed reads) locations containing FAST5 or FASTQ files. If sequ-into performs the read extraction from a folder containing FAST5 files, in addition to the reads (in FASTQ-format), an additional file containing the read-name and its creation time is produced. This allows further analysis of the off-target rate over sequencing time.

2.3. Software

sequ-into is implemented using Electron, a framework for creating native applications with web technologies like JavaScript, HTML, and CSS [16]. The user interface is developed in Typescript [17]/React[18] based on MaterialUI[19]. The read extraction from FAST5 folders is performed using the above described `extract_fast5.py` script. Reads are aligned to given references in python using the python wrapper for Minimap2, *mappy*, an aligner specialized for long-read alignments[20]. The alignment and the generation of all statistics, figures and the HTML-report is coordinated using our python server (startAlignmentServer.py, included in sequ-into). This server allows to increment existing results and thereby an online processing of the input reads. sequ-into uses the Windows Subsystem for Linux to ensure compatibility on Microsoft Windows. All required python dependencies can be installed using *pip* [21].

2.4. Benchmark

sequ-into internally uses mappy, the python wrapper for minimap2 [20]. Hence, the mapping accuracy of sequ-into is mainly determined by the minimap2 performance on the MinKNOW basecalled reads..

We perform two benchmarks: one on simulated *E. coli* and bacteriophage *Escherichia phage ADB-2* reads in a mixture, and another benchmark on a metagenomics dataset.

The reads have been simulated using NanoSim [22] version 2.1.0 with the Nanopore R9 1D profile provided by the NanoSim authors. For the metagenomics datasets the sequencing data from Edwards et al. [23] and accession PRJEB30868 (ERX3139117-ERX3139119) have been used. These are sequences generated from MinION sequencing of the ZymoBIOMICS Microbial Community Standard from Zymo Research [24].

2.5. Riboseq library

Ribosomal RNA contamination is detected by aligning the reads against a library of microbial ribosomal RNA sequences: the riboseq library. Using the list of available bacterial genomes from EMBL-EBI [25], for each species one representative strain (the first in list) is downloaded. This ensures that the additionally needed disk space is small. From those genomes, all sequences either directly annotated as *rRNA* or where the product (or any other description) is annotated as *ribosomal RNA*, are extracted. Currently included eukarya are *Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Danio rerio*. For these species, ribosomal RNA sequences provided by Rfam [26] in RNA families *Bacterial small subunit ribosomal RNA (RF00177)* and *Eukaryotic small subunit ribosomal RNA (RF01960)* are available. Some eukaryotic rRNA sequences are also contained in RF00177, which thus is included as well. Finally, ribosomal RNA sequences for 1485 species are accessible from within sequ-into, searchable via a text input with auto completion.

3. Results & discussion

3.1. Software

sequ-into is available as a cross-platform compatible software providing a mean of interaction (e.g. file dialog) known to users know from everyday computer usage (Fig. A.2). It has been designed such that the user can perform an off-target analysis using an easy-to-follow workflow. Each step for this analysis is supported by a brief graphical and written description. Providing a GUI makes the application accessible to most scientists [27,28].

While we anticipate the online use of sequ-into (while sequencing), it can also be used to analyse reads after the sequencing run has already been concluded (post-sequencing). To start the off-target read detection, the user can choose the current sequencing folder (online or post-sequencing detection) or regular FASTQ files (post-sequencing detection). sequ-into is designed to work with both FASTQ files, and FAST5 files. In the latter case (FAST5, real-time base calling), it will extract the first thousand reads in FASTQ format for further analysis (or all, if demanded), taking advantage of the live-basecalling functionality of MinKNOW. In our data, using the MinKNOW live-basecalling introduces an average delay of 48 s in sequence availability, with a maximal delay of 100 s.

The input reads are aligned against given reference sequences, e.g. genomes or rRNA sequences. By default, an *Escherichia coli k12 MG1655* genome is included in the distribution. In addition, sequences from the riboseq library (ribosomal RNAs from over 1400 organisms, see Materials & Methods) can be selected here. The user may also upload/use custom genomes in FASTA format, allowing to prepare custom ICO libraries. Any given reference may be defined as either an *on-* or *off-target* sequence, depending on whether it is the intended target sequence or an ICO.

Mappy, the python wrapper for Minimap2 [20] is used to align reads. This has several advantages: first it eases installation because sequ-into only depends on python tools, which are all installable via the standard python package installer *pip*[21]. Second, no intermediate SAM/BAM-files are written to disk. Moreover, no I/O bandwidth is taken from the actual sequencing and basecalling process. In addition, it would be risky to use BAM files to store alignments of ultra-long (genomic) reads due to the CIGAR size limit in the bam-format.

Aligning the selected reads against the references assesses the off-target rate (in terms of target versus off-target sequences). For example: bacterial RNA is intended to be sequenced and the user defines the bacterial rRNA as off-target reference. All aligned reads then originate from the off-target sequence, hence stem from ICOs. The off-target rate is then "% aligned reads". Contrarily, in case the user specifies the transcriptome or genome of the intended species, all aligned reads are considered as on-target reads. The on-target rate is then "% aligned reads".

Analysing transcriptomic reads may incur extra complexity, particularly in eukaryotes, due to their intron/exon structure. Counting the aligned bases of eukaryotic transcriptomes requires the handling of intronic gaps. The user can select to ignore aligned fragments with CIGAR code *N* in the calculations, e.g. because the reference contains intronic regions not present in the sequenced sample. sequ-into uses an online algorithm for calculating the required statistics and alignments. Upon starting sequ-into, the alignment server is ready to start or update an analysis by first loading any existing results, updating these results with the statistics of the newly processed reads, and, finally, saving the combined result for the next iteration.

As final output, sequ-into provides an overview of the performed alignment via the on- and off-target rate, the fraction of aligned bases and an analysis of the on-target rate over time (Fig. 1). In addition to the descriptive overview, sequ-into also shows a notification if the samples contain more than 30% off-target sequences (Fig. 1).

Due to varying read lengths, the number of (un-)aligned bases are considered. On the base-level, two measures are useful: the length of the alignment on the reference (alignment bases) and the length of the matching bases in the alignment (aligned bases). While the first measure is important to determine how well the reference is covered, the latter also gives an estimate of the alignment quality (regarding substitutions). Explanations and a description of how to interpret the reported descriptive values help the users to understand the values, also decreasing chances of misunderstandings.

If reads are extracted from FAST5 files and more than 1,000 reads are available, sequ-into provides a plot shows a binned (bin-size 1,000 reads) histogram of the alignment ratio of the reads (e.g. in the first 5,000 reads). This analysis is of particular interest for mixed samples, e.g. phage DNA sequencing with left-over phage host DNA, because changes in the off-target rate have been observed.

The read length distribution of the reads and the results of the alignment analysis are shown in a result summary, supported by pie charts of aligned reads and bases, and histograms of read length distributions. Besides the output in the sequ-into app (Fig. 1), where this overview is displayed, sequ-into saves the created plots together with an HTML report, which can easily be shared among colleagues.

In order to save computational resources, sequ-into uses an online and incremental algorithm. Before the alignment and read extraction, existing results are loaded. Only new reads are extracted and further processed. Alignment counts are updated incrementally and the descriptive statistics are updated and stored for the next analysis round. Thus, sequ-into runs on laptop computers, matching the portability of the MinION sequencer. The analysis of 1,000 reads with suspected *E. coli* contamination took 12 s including read extraction from FAST5 files on a Microsoft Windows 10 laptop with an Intel i7-7820HQ CPU and 32 GB RAM. Even on a more mainstream and (computationally) less powerful Microsoft Surface Book with 16 GB RAM and a 128 GB SSD, the sample was analyzed in less than 10 s. For detecting ribosomal off-target sequences in 34,782 *Helicobacter pylori* transcriptomic reads (Run 11), less than 10 s are needed (without read extraction) on the Windows 10 laptop. Neither sequ-into nor the live basecalling caused a bottleneck in this analysis. It can thus be used directly side-by-side with the MinION sequencer, either in the field or the lab.

3.2. Benchmark

Two benchmarks have been performed to assess the accuracy and correctness of sequ-into. The results are shown in Table 1.

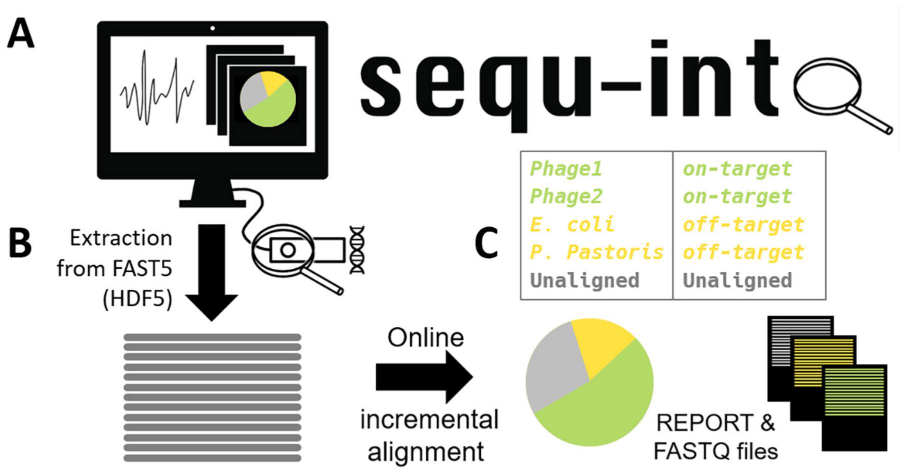
For the simulated dataset the number of simulated *E. coli* reads was 50.000 and 25.000 for *E. phage ADB-2*. Of all reads aligned by sequ-into only 6 reads fail to align. Otherwise, sequ-into, respectively the underlying minimap2 mapper, performs perfect.

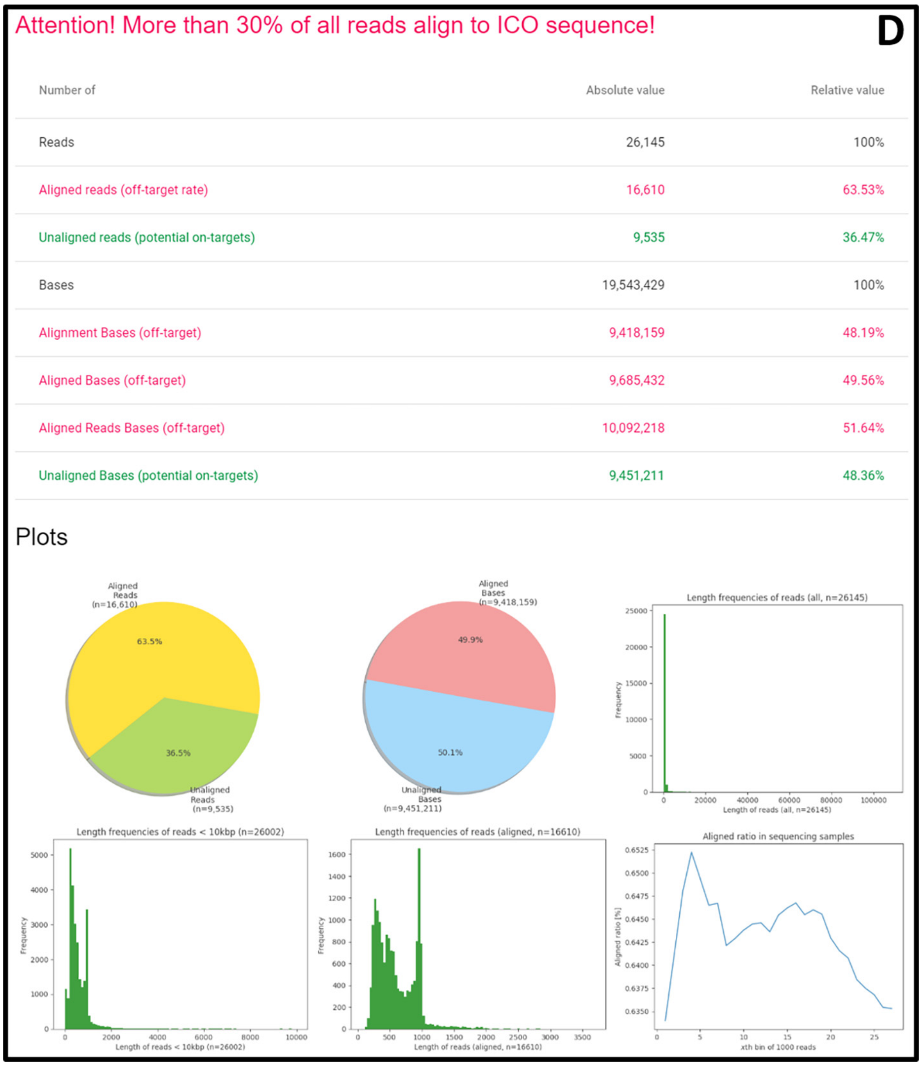
More interesting is the metagenomics benchmark. Here, the number of aligned reads deviates slightly from the number of expected reads and the expected fraction, respectively. For both yeast species, the fraction of identified reads matches the expected ratio well. *Staphylococcus aureus* is more prevalent than expected, but only by 4% - which is the maximal deviation observed in the benchmark. Interesting are the high read counts for *Escherichia coli* and *Salmonella enterica* (Table 1). This is because 14,582 reads align to both genomes. sequ-into does not try to untangle this, but reports these multi-mapping instead (Fig. 2). Assigning half of the reads to each organism, the align fraction resembles the expected fraction well enough. The fact, that the expected fraction already differs from theory is known and is based on the fact, that the DNA extraction and library prep may induce a bias. This is also reported by the manufacturer.

Given the low deviations from the expected fractions in the metagenomics sample, sequ-into/minimap2 performs considerably well. This is supported by the simulated reads, where all but 6 reads are assigned correctly. This is no surprise, since the accuracy of sequ-into is strongly determined by its underlying mapper, minimap2, which achieves an alignment rate of 98% and more for long reads [20].

3.3. Use-cases

In order to demonstrate that sequ-into supports lab experiments, three use-cases are presented. The first one demonstrates





Attention! More than 30% of all reads align to ICO sequence!

Number of	Absolute value	Relative value
Reads	26,145	100%
Aligned reads (off-target rate)	16,610	63.53%
Unaligned reads (potential on-targets)	9,535	36.47%
Bases	19,543,429	100%
Alignment Bases (off-target)	9,418,159	48.19%
Aligned Bases (off-target)	9,685,432	49.56%
Aligned Reads Bases (off-target)	10,092,218	51.64%
Unaligned Bases (potential on-targets)	9,451,211	48.36%

Plots

Fig. 1. A-C: The steps from (raw) sequencing data to output from sequ-into. D: Example run of sequ-into: result for a *Helicobacter pylori* RNA-seq example. Here, 63.5% of all reads originate from *H. pylori* ribosomal RNA (off-target reference).

how sequ-into helps in the sequencing of genomic samples with high off-target susceptibility. Here the protocol for extracting phage DNA and depleting *E.coli* host DNA was improved in only a few rounds of the rapid prototyping cycle (Fig. A.1).

The second use-case analyses a transcriptome sequencing project (post-sequencing analysis) and shows how sequ-into could have helped to improve sample quality by detecting a high ribosomal RNA content in the first experiment. This post analysis led to a

Table 1
Summary of the benchmarking results for simulated and metagenomics data. The number of total simulated reads is 75, 000. In the metagenomics dataset, the number of total aligned reads (149,742) is determined by the 177, 599 reads in the dataset and the 27, 857 unaligned reads.

Organism	Reads	Reads (after change)	Fraction Measured	Theoretic Fraction	Expected Fraction	Difference Theory	Difference Expected
NanoSim *E. coli*	49, 996		66.66%	66.6%	66.6%	-	-
NanoSim *E. phage ADB-2*	24, 998		33.33%	33.33%	33.3%	-	-
Meta: *Pseudomonas aeruginosa*	15, 096		10.08%	12.00%	12.50%	−1.92%	−2.42%
Meta: *Escherichia coli*	27, 698	20, 407	13.63%	12.00%	12.50%	1.63%	1.13%
Meta: *Salmonella enterica*	26, 106	18, 815	12.56%	12.00%	12.50%	0.56%	0.06%
Meta: *Lactobacillus fermentum*	15, 904		10.62%	12.00%	12.30%	−1.38%	−1.68%
Meta: *Enterococcus faecalis*	18, 624		12.44%	12.00%	9.50%	0.44%	2.94%
Meta: *Staphylococcus aureus*	24, 036		16.05%	12.00%	12.00%	4.05%	4.05%
Meta: *Listeria monocytogenes*	21, 653		14.46%	12.00%	12.50%	2.46%	1.96%
Meta: *Bacillus subtilis*	19, 901		13.29%	12.00%	14.70%	1.29%	−1.41%
Meta: *Saccheromyces cerevisiae*	3, 560		2.38%	2.00%	2.08%	0.38%	0.30%
Meta: *Cryptococcus neoformans*	2, 960		1.98%	2.00%	1.56%	−0.02%	0.42%

Table 2
Summary of the sequencing runs analysed by sequ-into. The number of reads refers to the number of basecalled reads.

Run ID	Sequence Type	Duration of Sequencing Run	Number of reads	Off-target rate (%)
1	CP & EP Phage DNA	3 : 42h	65, 964	42.11
2	Kit Phage DNA	2 : 00h	14, 750	7.66
3	DNAseI Phage DNA	2 : 00h	20, 756	2.15
11	*H. pylori* RNA	6 : 00h	26, 145	63.52
12	*H. pylori* RNA	5 : 50h	15, 332	57.21
13	*H. pylori* RNA	3 : 00h	22, 940	55.35
21	*H. pylori* RNA	5 : 00h	24, 540	65.86
1117	*E. coli* phage genome	59 : 46h	108, 026	9.37
1118	*E. coli* phage genome	59 : 40h	103, 384	10.01
1121	*E. coli* phage genome	59 : 40h	49, 720	9.47

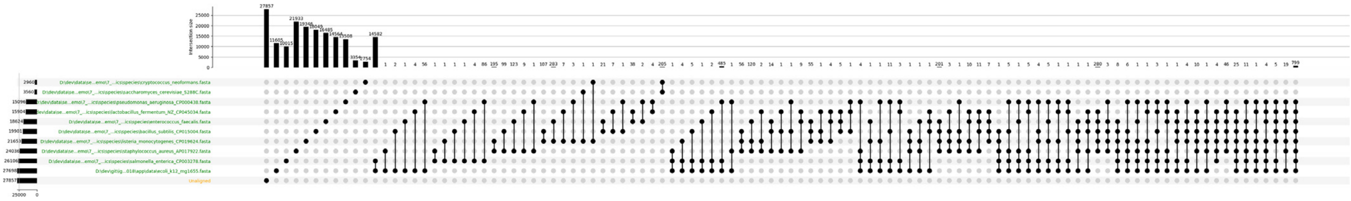


Fig. 2. Upset plot showing the number of aligned reads per expected genome of the metagenomics sample (Zymo Research Mock Community). It can be seen that most reads only map to a single organism. Noticeable are the 14, 852 reads which are shared by *S. enterica* and *E. coli*.

better ribosomal RNA depletion before the sequencing of further samples.

In the third use-case an external, publicly available dataset is re-analysed regarding its off-target rate.

The sequencing details are given in Table 2.

3.3.1. Case I: DNA purification

In this use-case a DNA sequencing analysis targeting phage DNA was performed. The practical problem is to determine levels of host (*E. coli*) DNA contamination after phage isolation for faster evaluation of extraction protocols. For later applicability, host DNA levels must be as low as possible.

MinION sequencing was used to assess the purity of the extracted DNA. In three rounds the purification protocol was improved (Table 2). The initial run (run 1) used only chloroform-phenol extraction and ethanol precipitation for DNA extraction and contains a peak *E. coli* off-target rate of 80%. For run 2, a standard phage isolation kit was used for DNA extraction, leading to off-target rates of below 20%. Final adjustments performed for run 3, with DNAseI incubation, led to an off-target rate of below 5% (run 3, details in A.1).

Using sequ-into, the experimenters are able to analyse their data conveniently. This allows to employ rapid-prototyping of the laboratory protocol, as the results are available directly after sequencing (Fig. A.1). In case of unwanted/bad results, a new strategy can be tested without first having to wait for the bioinformaticians to finish the analysis. The report function of sequ-into allows experimenters to easily share the report to discuss the results within a team.

3.3.2. Case II: RNA sequencing

The second use-case is from an *Helicobacter pylori* transcriptomic sequencing project (Table 2, method Supplement A.2). Common RNA purification techniques like poly-A-tail selection do not work in bacteria. Only ribosomal RNA depletion kits may get rid of rRNA using enzymatic reactions, making rRNA depletion particularly important for transcriptomic sequencing. Considering that rRNA can make up more than 85% of a cell's RNA [29], while not giving any information about the transcriptional regulation. After applying an enzymatic rRNA depletion on the input library, the initial rRNA content, in the experiments performed, was between 58–65% per sample, considering either the first 10% of the total sequencing time (data not shown) or the first 1,000 sequenced reads (Fig. A.3). Using sequ-into after data collection, it was possible to determine how well the library preparation and also the rRNA depletion worked. In the presented cases, the sequencing yielded enough support of a high rRNA content after only 10 min. With this result in mind, further measures could then be taken to deplete the rRNA to the desired level more efficiently (99% purity has been reported [30]). Knowing the rRNA fraction of a sequencing sample as soon as possible saves valuable sequencing time (and costs).

3.4. Case III: Analysing the off-target ratio over time

sequ-into has been developed in the context of phage genome sequencing with a focus on assessing sample (im-)purity. Besides the descriptive final overview, we also wanted to check whether the content of the host organism (*E. coli*) remains constant during sequencing. For this reason, sequ-into analyses the off-target rate for every set of 1,000 reads.

In the off-target rate plot of the phage DNA sequencing (Fig. A.4, *E. coli* genome as reference) we observed a high fraction of reads originating from *E. coli* at the beginning, getting fewer towards the end. Such an effect was not observed in the transcriptomic data (Fig. A.3).

Not knowing whether this observation is special to our data only (e.g. library preparation), we also analysed the FAST5 raw data from a public dataset (accession id PRJEB8318, runs 1117, 1118 and 1121) [15]. In that experiment, an *E. coli str. K12 substr. ER2738* is used as host organism, which is closely related to the *E. coli str. K-12 substr. MG1655* genome available from sequ-into and used here. In this data we made similar observations regarding the contamination over the number of sequenced reads (Fig. 3). It can be seen that for all three runs the off-target rate decreases. The fraction of reads originating from *E. coli* are at about 10% at the beginning, rising up to 20%, before getting lower. Again, similar to the phage DNA sequencing in use-case 1 (Fig. A.4), the read buckets at the start of the sequencing run align better to *E. coli* than those afterwards. Nonetheless, even such data can be successfully analysed by sequ-into. In concordance with the results from use-case 1, we show that even with this surprising behaviour of the sequenced reads, the first few thousand reads are an useful estimate of the overall off-target rate.

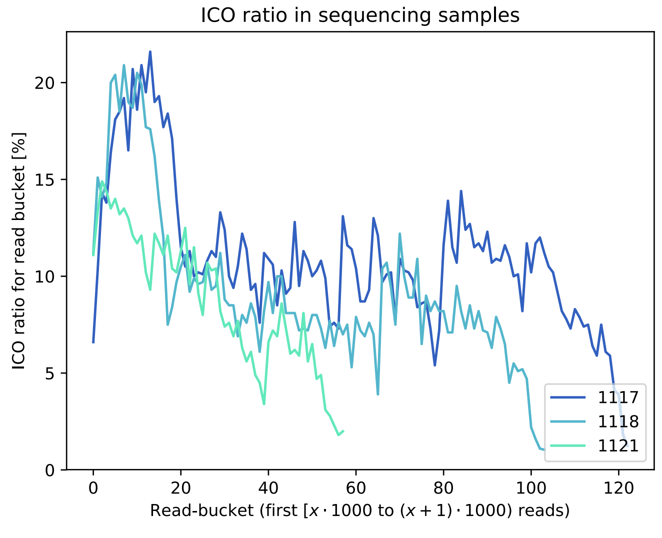


Fig. 3. Fraction of off-target reads for the external phage DNA dataset (accession id PRJEB8318, runs 1117, 1118 and 1121). The reads were binned into buckets of 1000 reads with respect to their sequencing order over time. For the buckets of the 3 runs the respective percentage of ICOs is shown.

3.5. Read extraction

For sequ-into only the first 1,000 reads of an experiment are extracted in the real-time mode by default - the user can change this. In our analysis we have considered two frequent scenarios: the detection of ICOs in either transcriptomic reads (RNAseq) or genomic reads (DNAseq). With the genomic/phage samples we observed that occasionally off-target (*E. coli*) reads are slightly more frequent in the first few thousands reads (Fig. A.4 and Fig. 3) but then remain constant throughout the sequencing runs (Fig. 1). Thus, already the first (few) 1,000 reads provide a useful estimate of the overall off-target rate or its upper-bound. Analysing more reads is not necessary for fast decision making, yet possible with sequ-into.

4. Conclusion

sequ-into offers a cross-platform, graphical-user-interface and uses state-of-the-art long-read alignment software such that everyone can perform an on-/off-target analysis, even during the sequencing run (use-case 1).

It can detect large fractions of ribosomal RNA early in a sequencing experiment. If applied early in the sequencing project, sequ-into can show the high rRNA content, and thereby help to avoid a significant loss of reads to ICOs (use case 2). Additionally, users have easy access to our riboseq library, with ribosomal RNA sequences for more than 1,000 species, from within sequ-into.

Using sequ-into we investigated the (im-)purity of several phage DNA sequencing runs (use-case 3). We observed that the sequenced reads stem more frequently from the off-target (*E. coli*) at the beginning of a sequencing run, than towards the end. Still these results show that the first few thousand reads provide a useful estimate of the overall off-target rate in all evaluated cases.

From within sequ-into, mappy, the python wrapper for minimap2, aligns the reads to the references. For easy sharing of the results, and for later reference, sequ-into creates an HTML report for each analysis. In our use-cases, we observe that a few 1,000 reads are already sufficient to obtain a useful estimate of the off-target rate, allowing a fast availability of the results in less than a minute, even on a typical laptop computer. sequ-into supports the idea of fast protocol optimization at very low cost, by everyone and at any place.

Program and Data Availability

sequ-into is available from GitHubhttps://github.com/mjoppich/sequ-into with demo data. Documentation is available onlinehttps://sequ-into.readthedocs.io/en/latest/. sequ-into has been tested on Windows 10 Build 18363 with Ubuntu 18.04 LTS Windows Subsystem for Linux app. sequ-into has also been tested on Mac OS X 10.15 and Xubuntu 18.04.2 LTS.

Funding

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) via SFB 1123/2/Z2 (MJ/RZ) and DFG JI 221/1-1 (LFJS).

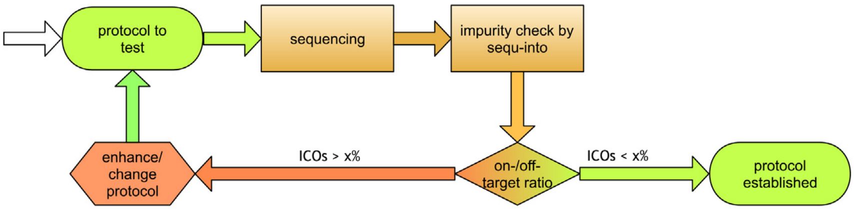


Fig. A.1. The rapid prototyping cycle performed to establish the new protocol. The quality-control step was performed using sequ-into.

CRediT authorship contribution statement

Markus Joppich: Software, Conceptualization, Methodology, Visualization, Validation, Data curation, Supervision, Writing - original draft, Writing - review & editing. **Margaryta Olenchuk:** Software, Visualization, Data curation, Methodology, Writing - original draft, Writing - review & editing. **Julia Mayer:** Software, Visualization, Data curation, Methodology, Writing - original draft, Writing - review & editing. **Quirin Emslander:** Investigation, Resources, Conceptualization, Validation, Writing - review & editing. **Luisa F. Jimenez-Soto:** Investigation, Resources, Writing - review & editing. **Ralf Zimmer:** Conceptualization, Resources, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

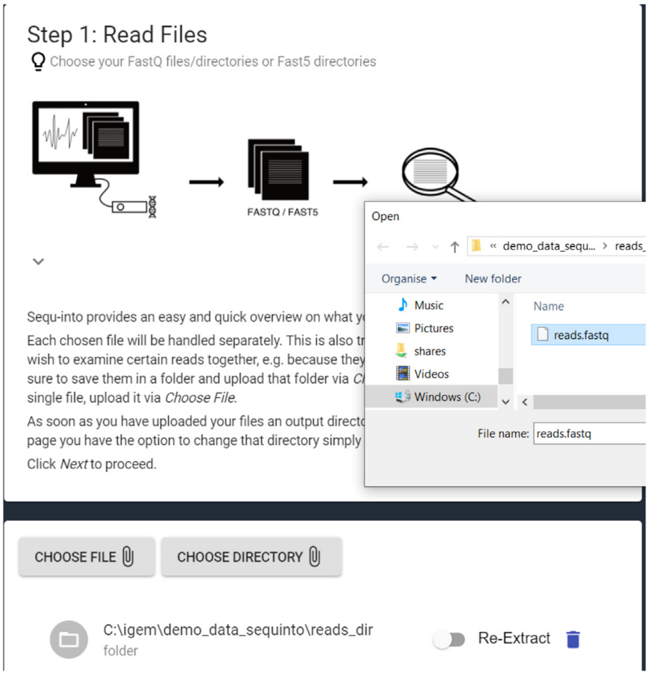
Acknowledgments

We would like to thank the group of Prof. Simmel (TU Munich) for providing the access to the ONT MinION device and hosting the iGEM 2018 Munich team.

Appendix A. Supplement

A.1. Isolation of Phage DNA

The Phage-DNA was at first isolated by chloroform-phenol extraction later with a Phage-DNA isolation kit (Cat. 46800, 46850, Norgen, Canada). For the chloroform-phenol extraction 200μL of phage stock (10^6-10^{10}pfu/ml) were mixed with 200μL of Roti$^{®}$-Phenol/Chloroform/Isoamyl alcohol(pH 7.5–8.0) in a 5PRIME Phase Lock Gel^{TM}tube (Quantabio, USA), to enable a better phase separation. The tubes were gently inverted and centrifuged for 5 min at $16,000 \times g$ at room temperature. Afterwards, 400μL of pure chloroform (Roth, Germany) were added to the upper phase of the tube, inverted and centrifuged as previously. The supernatant was transferred to a separate Eppendorf tube and 20μL of 3 M of sodium acetate were added. The DNA was precipitated with 1 mL of $-80\,^{\circ}$C cold pure ethanol (Roth, Germany) at 1 h at $-80\,^{\circ}$C. Afterwards, the DNA the sample was centrifuged at $16,000 \times g$ for 15 min at room temperature. The supernatant was discarded and 1 mL of $-20\,^{\circ}$C cold 70% (v/v) ethanol (Roth, Germany) was added. The DNA was pelleted at $16,000 \times g$ for 5 min at $4\,^{\circ}$C and the supernatant was discarded again. The sample was stored at room temperate for approximately 15 min to evaporate the remaining ethanol. The pellet was dissolved with 30μL of nuclease-free water (Thermo Fischer Scientific, USA) (run1). To achieve DNA isolation based on the Phage Isolation Kit (Norgen, Canada) user manufacturing protocol was followed (run2). To eliminate host DNA, which was extracted by chloroform phenol extraction, DNaseI (NEB, USA) was applied to the purified DNA for 45 min (run 3).


Step 1: Read Files

Choose your FastQ files/directories or Fast5 directories

Sequ-into provides an easy and quick overview on what y
Each chosen file will be handled separately. This is also tr
wish to examine certain reads together, e.g. because they
sure to save them in a folder and upload that folder via Ch
single file, upload it via *Choose File*.
As soon as you have uploaded your files an output directo
page you have the option to change that directory simply
Click *Next* to proceed.

File name: reads.fastq

C:\igem\demo_data_sequinto\reads_dir
folder

Fig. A.2. sequ-into provides explanations in its graphical user-interface leading through each step of the off-target analysis (here: selecting correct input).

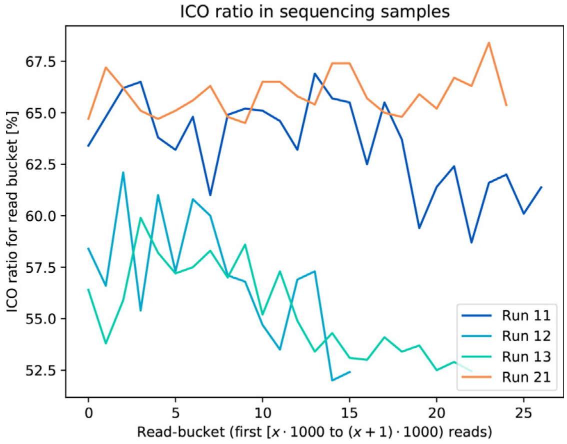


Fig. A.3. Off-target rate for every 1,000 reads in *Helicobacter pylori* transcriptome sequencing (with rRNA as off-target). It can be seen that the ribosomal RNA content is conserved over read buckets. The first 1,000 reads already give an estimate for the overall off-target rate $(+/-5\%)$.

A.2. RNA isolation and preparation of H. pylori transcriptome

Bacteria were grown in liquid culture complemented with Cholesterol as previous published [31] until an OD550 of 0.6. A final pellet containing approx. $3.6 \cdot 10^8$ bacteria were frozen in $-70\,^\circ$C before RNA extraction. For RNA extraction the RNA extraction kit from Qiagen was used and their protocol followed. To eliminate rRNA, samples were digested with the Terminator 5'-Phosphate-Dependent Exonuclease from Illumina® (Cat. Nr. TER51020) before first-strand DNA was created. After evaluation of RNA quality with Bio-Rad's Experion Electrophoresis System, protocols were followed as recommended by Oxford Nanopore Technologies. Library preparation has been performed using the SQK-LSK208 kit for 2D sequencing (R9.4 chemistry). The sequencing has been performed using a FLO-MIN106 flowcell.

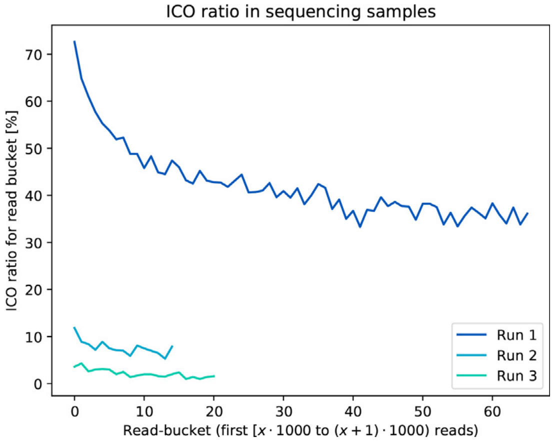


Fig. A.4. Off-target rate for every 1,000 reads in phage genome sequencing (with *E. coli* as off-target). It can be seen that the off-target reads are decreasing for later read buckets. The first ten buckets (10,000 reads) seem to be an estimate for the upper bound of the off-target rate.

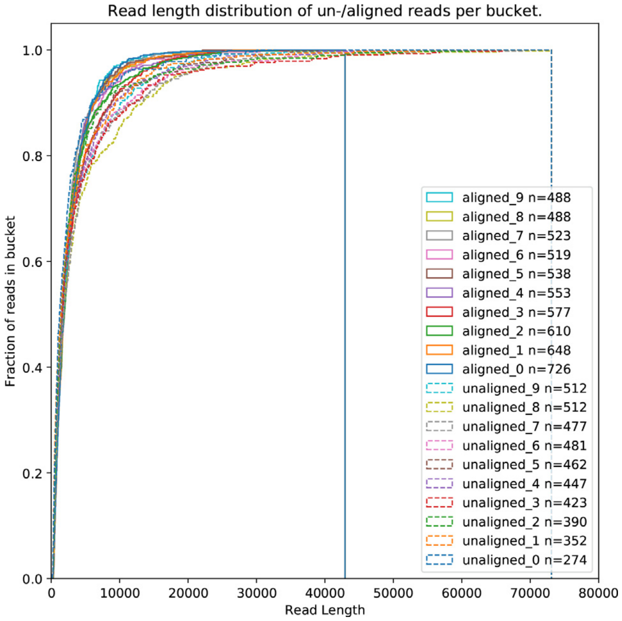


Fig. A.5. The length distribution of aligned (*E. coli* off-target) and unaligned reads from the phage DNA run 1. It can be seen that both the aligned and unaligned read length distributions do not differ much, with the unaligned reads having a tendency of being longer.

A.3. DNA/RNA sequencing & basecalling

The collected DNA and RNA samples have been sequenced using an Oxford Nanopore MinION sequencer. The sequencing time and yield has been different per sample and is summarized in Table 2. The *Phage DNA* reads have been basecalled using MinKNOW-Live-Basecalling 1.14.1. The *H. pylori* RNA reads have been basecalled using Albacore version 1.2.2.

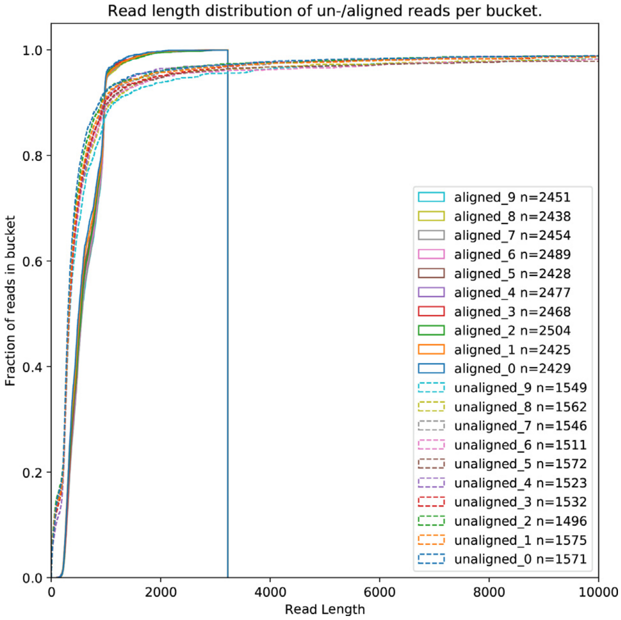


Fig. A.6. The length distribution of aligned (*H. pylori* ribosomal RNA/off-target) and unaligned reads from the *H. pylori* transciptomic sequencing (runs 11,12,13,21). The x-axis has been cut at 10,000*bp* for better readability. Each bin (0 to 9) contains 4,000 reads (1,000 of each experiment).

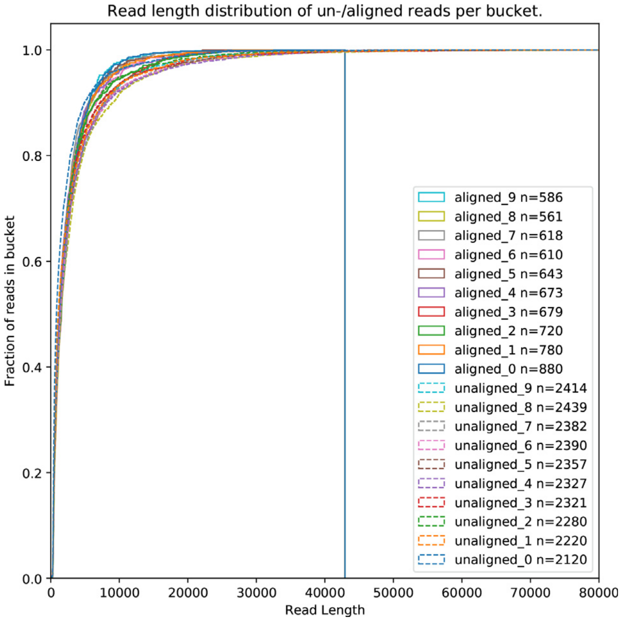


Fig. A.7. The length distribution of aligned (*E. coli* off-target) and unaligned reads from the combined phage DNA runs (runs 1,2,3). The x-axis has been cut at 80,000*bp* for better readability. Each bin (0 to 9) contains 3,000 reads (1,000 of each experiment).

A.4. Usability

Fig. A.1 and Fig. A.2.

A.5. ICO rates

Fig. A.3 and Fig. A.4.

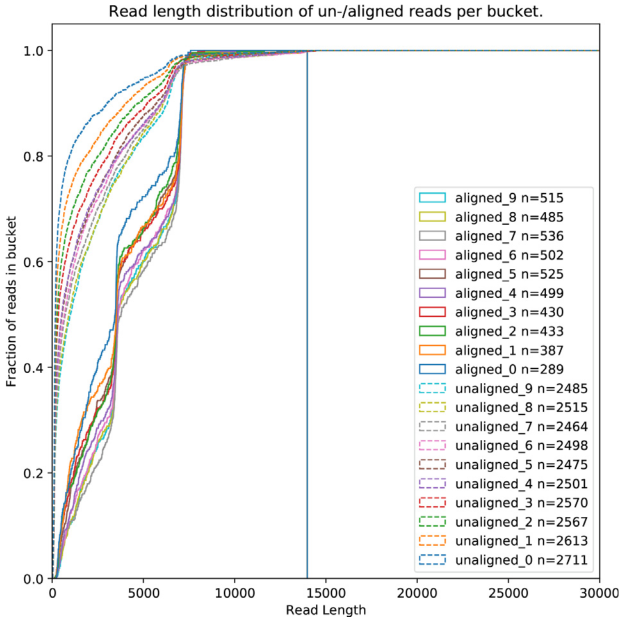


Fig. A.8. The length distribution of aligned (*E. coli* off-target) and unaligned reads from the combined external phage DNA runs (PRJEB8318). The x-axis has been cut at 30,000*bp* for better readability. Each bin (0 to 9) contains 3,000 reads (1,000 of each experiment).

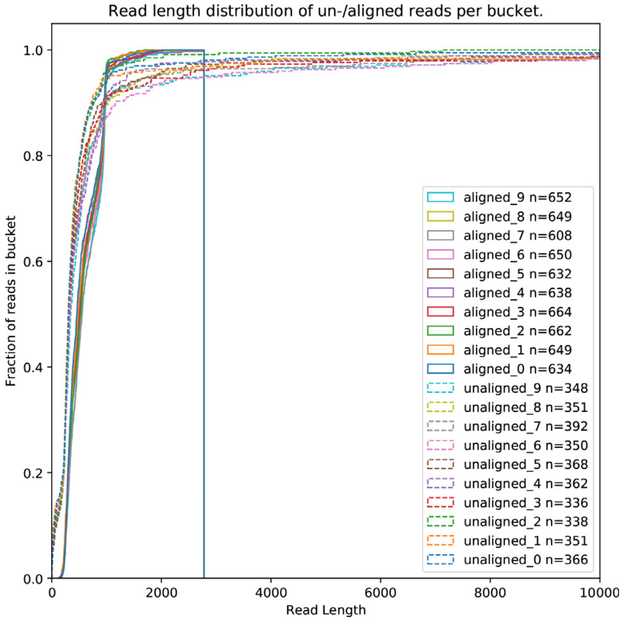


Fig. A.9. The length distribution of aligned (*H. pylori* ribosomal RNA/off-target) and unaligned reads from the *H. pylori* transciptomic sequencing (run R11). The x-axis has been cut at 10,000*bp* for better readability. Each bin (0 to 9) contains 1,000 reads.

A.6. Length distributions

In use-case 3 (Section 3.4), it was noted that there were more *E. coli* reads at the start of the sequencing run than at the end. A possible explanation was that the sequenced read lengths change over time, e.g. that shorter reads are sequenced first. However, an analysis of the read length distributions reveals that the read lengths of the off-target-*E. coli* reads (aligned) does not differ within the (from all experiments combined) first ten 1,000-read-bins, neither for our sequencing runs (Fig. A.5 (phage), Fig. A.6 (*H. pylori*)) nor for the public data (Fig. A.8). Additionally, the

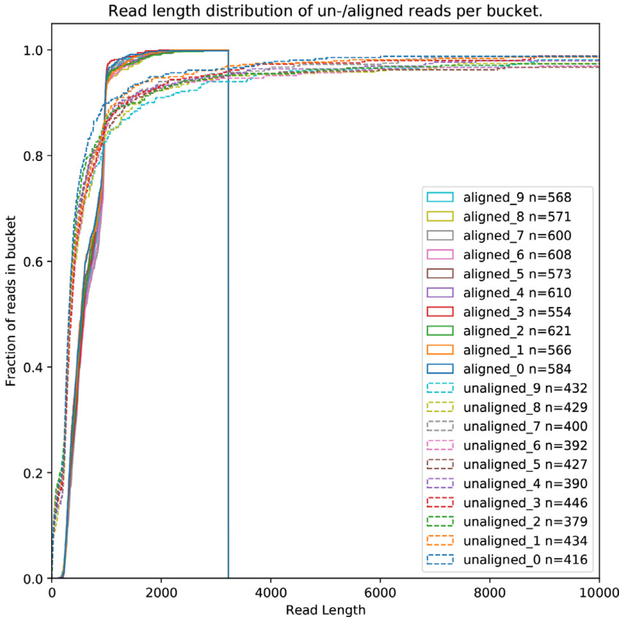


Fig. A.10. The length distribution of aligned (*H. pylori* ribosomal RNA/off-target) and unaligned reads from the *H. pylori* transciptomic sequencing (run R12). The x-axis has been cut at 10,000*bp* for better readability. Each bin (0 to 9) contains 1,000 reads.

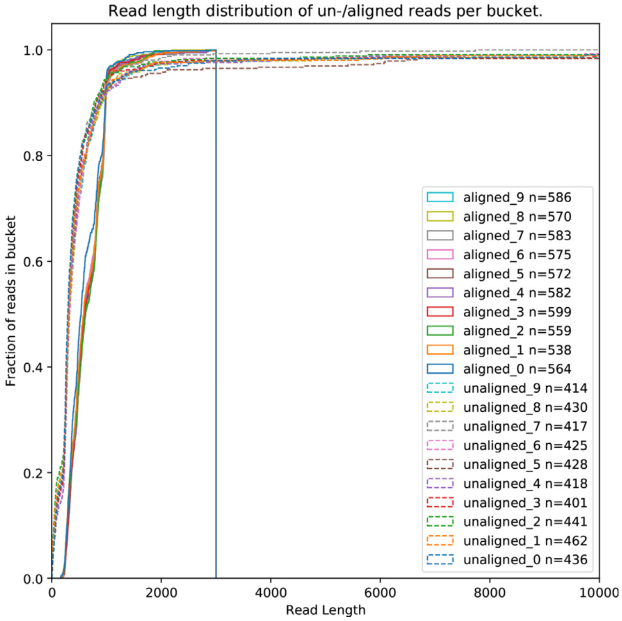


Fig. A.11. The length distribution of aligned (*H. pylori* ribosomal RNA/off-target) and unaligned reads from the *H. pylori* transciptomic sequencing (run R13). The x-axis has been cut at 10,000*bp* for better readability. Each bin (0 to 9) contains 1,000 reads.

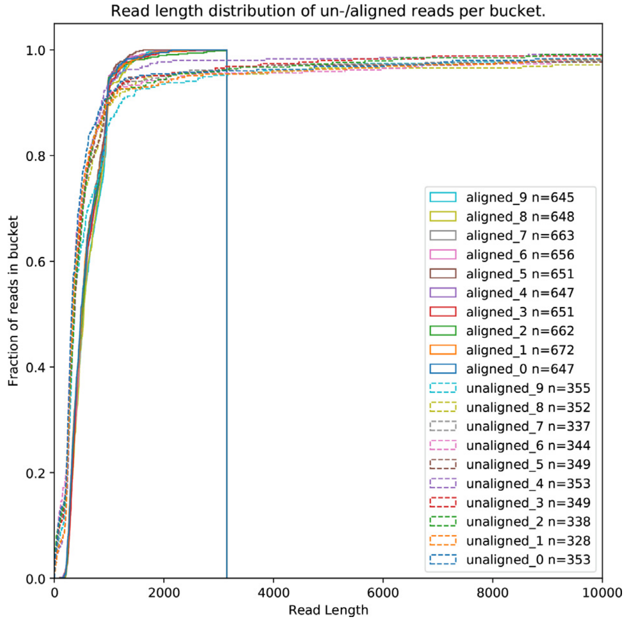


Fig. A.12. The length distribution of aligned (*H. pylori* ribosomal RNA/off-target) and unaligned reads from the *H. pylori* transciptomic sequencing (run R21). The x-axis has been cut at $10,000bp$ for better readability. Each bin $(0$ to $9)$ contains $1,000$ reads.

aligned and unaligned reads are more similar to each other, than to a specific bin. This can also be observed in the public sequencing data (Fig. A.8). Interestingly, the same observation can be made for the cDNA transcriptomic reads from *H. pylori* (Fig. A.6). Thus, the fragment size of the off-target reads does not explain the observed bias. The reason why we see such a bias for *E. coli* in the phage DNA sequencing samples remains unclear and needs further investigation.

The length distributions for the experimental data are shown in Fig. A.6 for the *H. pylori* samples, in Fig. A.7 for the *E. coli* phage samples and in Fig. A.8 for the publicly available *E. coli* phage samples.

The single *H. pylori* samples are analysed in Fig. A.9 (Sample R11), A.10 (Sample R12), A.11 (Sample R13) and A.12 (Sample R21).

References

[1] Goldstein S, Beka L, Graf J, Klassen JL. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. BMC Genomics 2019;20(1):23. https://doi.org/10.1186/s12864-018-5381-7. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-5381-7.

[2] Schneider W, Wilson V, King J, Sherman D, Bronzato Badial A, Gopakumar A, Stone A. Nanopore Sequencing as a Surveillance Tool for Plant Pathogens in Plant and Insect Tissues. Plant Dis 2018;102(8):1648–52. https://doi.org/10.1094/pdis-04-17-0488-re.

[3] Magiorkinis G, Mbisa JL, Harrison I, Karamitros T, Katzourakis A, Piorkowska R. De Novo assembly of human herpes virus Type 1 (HHV-1) genome, mining of non-canonical structures and detection of novel drug-resistance mutations using short- and long-read next generation sequencing technologies. PLOS ONE 2016;11(6). https://doi.org/10.1371/journal.pone.0157600.

[4] Fleming MB, Patterson EL, Reeves PA, Richards CM, Gaines TA, Walters C. Exploring the fate of mRNA in aging seeds: protection, destruction, or slow decay?. J Exp Bot 2018;69(18):4309–21. https://doi.org/10.1093/jxb/ery215. URL http://www.ncbi.nlm.nih.gov/pubmed/29897472, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6093385.

[5] Casale Brunet S, Schuepbach T, Guex N, Iseli C, Bridge A, Kuznetsov D, et al. Towards in the field fast pathogens detection using FPGAs. Proceedings – 2018 International Conference on Field-Programmable Logic and Applications, FPL 2018. IEEE; 2018. p. 463–4. https://doi.org/10.1109/FPL.2018.00091. URL https://ieeexplore.ieee.org/document/8532495/.

[6] Goordial J, Altshuler I, Hindson K, Chan-Yam K, Marcolefas E, Whyte LG. In situ field sequencing and life detection in remote (79°26'N) Canadian high arctic permafrost ice wedge microbial communities. Front Microbiol 2017;8:2594. https://doi.org/10.3389/fmicb.2017.02594.

[7] Rainey K, First dna sequencing in space a game changer (2016).URL https://www.nasa.gov/mission_pages/station/research/news/dna_sequencing..

[8] Kryukov K, Imanishi T. Human contamination in public genome assemblies. PLoS ONE 2016;11(9). https://doi.org/10.1371/journal.pone.0162424. e0162424.

[9] Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. PeerJ 2014;2014(1). https://doi.org/10.7717/peerj.675. e675.

[10] Leggett RM, Heavens D, Caccamo M, Clark MD, Davey RP. NanoOK: Multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. Bioinformatics 2015;32(1):142–4. https://doi.org/10.1093/bioinformatics/btv540. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv540.

[11] Edwards HS, Krishnakumar R, Sinha A, Bird SW, Patel KD, Bartsch MS. Real-Time Selective Sequencing with RUBRIC: Read Until with Basecall and Reference-Informed Criteria. bioRxiv 2018:. https://doi.org/10.1101/460014. URL https://www.biorxiv.org/content/early/2018/11/02/460014460014.

[12] Juul S, Izquierdo F, Hurst A, Dai X, Wright A, Kulesha E, et al. What's in my pot? Real-time species identification on the MinION. bioRxiv 2015:. https://doi.org/10.1101/030742. arXiv:030742, URL https://www.biorxiv.org/content/10.1101/030742v1 http://biorxiv.org/content/early/2015/11/06/030742.abstract030742.

[13] Cao MD, Ganesamoorthy D, Elliott AG, Zhang H, Cooper MA, Coin LJ. Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION sequencing. GigaScience 2016;5(1):32. https://doi.org/10.1186/s13742-016-0137-2. URL https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-016-0137-2.

[14] Mancabelli L, Milani C, Lugli GA, Fontana F, Turroni F, van Sinderen D, Ventura M. The impact of primer design on amplicon-based metagenomic profiling accuracy: Detailed insights into bifidobacterial community structure. Microorganisms 2020;8(1):1–11. https://doi.org/10.3390/microorganisms8010131.

[15] Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nat Methods 2015;12(4):351–6. https://doi.org/10.1038/nmeth.3290. URL http://www.nature.com/articles/nmeth.3290.

[16] Electron — Build cross platform desktop apps with JavaScript, HTML, and CSS. (2019).URL https://electronjs.org/.

[17] TypeScript – JavaScript that scales. (2019).URL https://www.typescriptlang.org/index.html..

[18] React, React – A JavaScript library for building user interfaces (2019).URL https://reactjs.org/..

[19] The world's most popular React UI framework - Material-UI (2019).URL https://material-ui.com/.

[20] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191. URL https://academic.oup.com/bioinformatics/article/34/18/3094/4994778, http://arxiv.org/abs/1708.01492.

[21] PyPi, Python package index (2019).URL https://pypi.org..

[22] Yang C, Chu J, Warren RL, Birol I. NanoSim: Nanopore sequence read simulator based on statistical characterization. GigaScience 2017;6(4):1–6. https://doi.org/10.1093/gigascience/gix010.

[23] Edwards A, Debbonaire AR, Nicholls SM, Rassner SM, Sattler B, Cook JM, et al. In-field metagenome and 16S rRNA gene amplicon nanopore sequencing robustly characterize glacier microbiota. bioRxiv 2019:. https://doi.org/10.1101/073965. URL https://www.biorxiv.org/content/10.1101/073965v3073965.

[24] Zymo Research, ZymoBIOMICS Microbial Community Standard Catalog No. D6300 (2020) 1–8.URL https://www.bioscience.co.uk/cpl/136244/.

[25] Genomes Pages - Bacteria (2019).URL https://www.ebi.ac.uk/genomes/bacteria.html..

[26] Argasinska J, Quinones-Olvera N, Nawrocki EP, Finn RD, Bateman A, Eddy SR, Petrov AI, Kalvari I, Rivas E. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 2017;46(D1):D335–42. https://doi.org/10.1093/nar/gkx1038. URL http://academic.oup.com/nar/article/46/D1/D335/4588106.

[27] Joppich M, Zimmer R, From command-line bioinformatics to bioGUI, PeerJ (Nov 2019). doi: 10.7717/peerj.8111. URL https://peerj.com/articles/8111/.

[28] Větrovský T, Baldrian P, Morais D. SEED 2: A user-friendly platform for amplicon high-throughput sequencing data analyses. In: Berger B, editor. Bioinformatics, vol. 34. Narnia; 2018. p. 2292–4. https://doi.org/10.1093/bioinformatics/bty071. URL https://academic.oup.com/bioinformatics/article/34/13/2292/4857359.

[29] Karpinets TV, Greenwood DJ, Sams CE, Ammons JT. RNA: Protein ratio of the unicellular organism as a characteristic of phosphorous and nitrogen stoichiometry and of the cellular requirement of ribosomes for protein synthesis. BMC Biol 2006;4(1):30. https://doi.org/10.1186/1741-7007-4-30. URL http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-4-30.

[30] Petrova OE, Garcia-Alcalde F, Zampaloni C, Sauer K. Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. Sci Rep 2017;7(1):41114. https://doi.org/10.1038/srep41114. URL http://www.nature.com/articles/srep41114.

[31] Jiménez-Soto LF, Rohrer S, Jain U, Ertl C, Sewald X, Haas R. Effects of cholesterol on helicobacter pylori growth and virulence properties in vitro. Helicobacter 2012;17(2):133–9. https://doi.org/10.1111/j.1523-5378.2011.00926.x. URL http://www.ncbi.nlm.nih.gov/pubmed/22404444.