Targeted decontamination of sequencing data with CLEAN

Marie Lataretu[1,2,*], Sebastian Krautwurst[2], Adrian Viehweger[3], Christian Brandt[4], Martin Hölzer[1]

[1] Genome Competence Center (MF1), Methodology and Research Infrastructure, Robert Koch Institute, Berlin, Germany
[2] RNA Bioinformatics and High-Throughput analysis, University of Jena, Jena, Germany
[3] Institute of Medical Microbiology and Virology, University Hospital Leipzig, Leipzig, Germany
[4] Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany

* corresponding author

Marie Lataretu: lataretum@rki.de

Sebastian Krautwurst: sebastian.krautwurst@uni-jena.de

Adrian Viehweger: adrian.viehweger@medizin.uni-leipzig.de

Christian Brandt: christian.brandt@med.uni-jena.de

Martin Hölzer: hoelzerm@rki.de

Abstract

Background

Many biological and medical questions are answered based on the analysis of sequence data. However, we can find contaminations, artificial spike-ins, and overrepresented rRNA sequences in various read collections and assemblies; complicating data analysis and making interpretation difficult. In particular, spike-ins used as controls, such as those known from Illumina (PhiX phage) or Nanopore data (DNA CS lambda phage, yeast enolase ENO2), are often not considered as contaminants and also not appropriately removed during bioinformatics analyses.

Findings

To address this, we developed CLEAN, a pipeline to remove unwanted sequence data from both long and short read sequencing techniques from a wide range of use cases. While focusing on Illumina and Nanopore data and removing of their technology-specific control sequences, the pipeline can also be used for everyday tasks, such as host decontamination of metagenomic reads and assemblies, or the removal of rRNA from RNA-Seq data. The results are the purified sequences and the sequences identified as contaminated with statistics summarized in an HTML report.

Conclusions

The decontaminated output files can be used directly in subsequent analyses, resulting in faster computations and improved results. Although decontamination is a task that seems mundane, many contaminants are routinely overlooked, cleaned by steps that are not fully reproducible or difficult to trace by the user. CLEAN will facilitate reproducible, platform-independent data analysis in genomics and transcriptomics and is freely available at https://github.com/hoelzer/clean under a BSD3 license.

Keywords

sequencing, decontamination, pipeline, Nextflow

Background

The high-throughput sequencing of DNA and RNA has become a standard approach in molecular biology. Next-Generation Sequencing (NGS), predominantly provided by Illumina, is capable of generating high-quality data from DNA and cDNA with low costs and error rates. The relatively short reads (50-300 nt) produced by NGS are, amongst other topics, used for the reconstruction of genomes, identifying SNPs, or characterizing differentially expressed genes. One technological limitation of NGS, the short read length, was overcome in recent years with the development of long-read sequencing technologies (Third-Generation Sequencing; TGS). In particular, Oxford Nanopore Technologies (ONT) provides a small, affordable, and mobile device that can generate reads of unprecedented length from DNA, cDNA, and also native RNA (Hu et al. 2021; Quick et al. 2016). Next to Illumina, the technology was also widely used to sequence SARS-CoV-2 samples during the COVID-19 pandemic (Brandt et al. 2021). The longer reads are used to significantly improve assembly contiguity (Nurk et al. 2022), the taxonomic classification of metagenomic samples (Overholt et al. 2020), or help to characterize alternative splicing in more

depth (Naftaly et al. 2021) while technological advances continue to push error rates more closely towards the level of short-read data (Sereika et al. 2022).

Since NGS and TGS (or simply “sequencing technologies”) are widely used, quality control of the raw sequencing data is becoming increasingly important. Most bioinformatics tools and pipelines identify and trim low-quality bases and remove remaining adapter sequences. However, one crucial step is often overlooked and still poses a challenge for sequencing technologies (Nieuwenhuis et al. 2020): the identification of DNA and/or RNA contamination where material from two or more sources is accidentally mixed or is simply a natural component of the sample, for example originating from cell line preparation (Chrisman et al. 2022) or in metagenomic samples. When contamination happens after sample collection or shipping, the preparation of the sequencing library involving multiple steps in the lab is another possible source (Porter et al. 2021). Apart from such unwanted contaminations, short and well-described control sequences are frequently spiked into sequencing runs to function as calibrations for basecalling and monitor the sequencing run's quality. Most commonly known is the PhiX phage genome, frequently used as a control in Illumina experiments. PhiX sequences were already shown to be large-scale contaminations in microbial isolate genomes because the reads were not cleaned before assembly and publication of genomes in public databases (Mukherjee et al. 2015). Also, we found that the positive control in ONT DNA sequencing (known as DCS), a 3.6 kb standard amplicon mapping the 3' end of the Lambda phage genome, is wrongly labeled as *E. coli* or *Klebsiella quasipneumoniae* subsp*. similipneumoniae* plasmid in the NCBI GenBank (CP077071.1, CP092122.1), see Supplemental Figure 1. For ONT native RNA sequencing, a yeast ENO2 Enolase II transcript of strain S288C, YHR174W, functions as a positive control. Spike-in steps are usually optional; however, the information if a spike-in was used, often does not reach the user working with raw reads.

Besides the decontamination of such manually introduced control sequences and other accidentally introduced but known contaminations, other use cases exist, where specific biological sequences should be removed, that can be a natural part of a sample or are still remaining after experimental steps. One prominent example is the removal of ribosomal or mitochondrial RNA from Illumina RNA-Seq samples before read-count normalization and differential gene expression estimation (Wolf 2013; Zhao et al. 2018; Raz et al. 2011). Even if rRNA depletion kits are frequently used to reduce the amount of rRNA before sequencing, rRNA can still be present in a sample. This applies in particular to non-model species where no optimized kit exists (Hölzer et al. 2019). Another example is the removal of host sequences, for example, in human gut microbiome sequencing data (Almeida et al. 2019), which is becoming increasingly important with the advent of metagenomic and metatranscriptomic sequencing.

In the past, several tools were developed for the fast classification of sequence data and thus also applicable for decontamination. One approach involves the taxonomic classification of all reads followed by removing unwanted sequences. Tools implementing such an approach are Kraken2/Kraken software suite (Wood et al. 2019 and Lu et al. 2022), Clark (Ounit et al. 2015) and Kaiju (Menzel et al. 2016). HoCoRT (Rumbavicius et al. 2022) offers a wrapper around well-known mapping and classification tools. SourceTracker (Knights et al. 2011), microDecon (McKnight et al. 2019) and Decontam (Davis et al. 2018) follow the metagenomics approach by analyzing the composition of the sample and finding unexpected proportions of contamination

taxa. The latter focus on short-read data, while other tools focus on the ONT DNA spike-in, nanolyse (De Coster et al. 2018), or on cloned, exogenous cDNA removal from NGS data, cDNA-detector (Qi et al. 2021). However, and although decontamination of already known species is in many cases a rather easy task with potentially huge benefits, many studies still lack appropriate decontamination of their sequenced samples. One reason for that might be that the output files of many pipelines cannot be directly used for downstream steps such as assembly or annotation and additional formatting of the files and extraction of the results are needed. As a direct result, we can find contamination omnipresent in genomic resources (Steinegger and Salzberg 2020). In particular, with the rise of TGS data, specialized methods are also needed for the fast decontamination of long reads.

Mapping reads to a reference genome for decontamination can be a general step while working with sequencing data. Therefore, we developed CLEAN (https://github.com/hoelzer/clean) as an easy-to-use all-in-one decontamination pipeline for short reads, long reads, and any FASTA-formatted sequence file. While initially developed for the decontamination of Illumina and Nanopore positive spike-in controls and host sequences in metagenomic samples, we extended the functionality of the pipeline to clean against any provided reference sequence(s). Also, we implemented the removal of rRNA from Illumina RNA-Seq samples in a faster and easier way than current state-of-the-art software (Kopylova, Noé, and Touzet 2012). Furthermore, CLEAN includes a convenient QC report and outputs the intermediate mapping files, which can be used for further investigation. Thus, CLEAN can be easily downloaded, installed, and executed with a single command on a local laptop, a high-performance cluster, or the cloud. We especially focused on well-structured output files and formats so that the decontaminated data files can be directly used in further downstream analyses such as assembly or annotation, thus allowing direct integration of CLEAN in other workflows. We believe that by providing an easy-to-use, expandable and reproducible pipeline, the decontamination of all kinds of sequencing data in molecular biology studies and genomic resources will increase.

Findings

Implementation

We implemented the pipeline in the workflow manager Nextflow v21.04.0 or higher (Di Tommaso et al. 2017). Every step is encapsulated in a software container (Docker (Boettiger 2015) or Singularity) or virtual environment (Conda (Grüning et al. 2018)). The modular structure allows updating of the containers and environments periodically. The user can deploy the software directly from GitHub. CLEAN can be easily installed - only Nextflow and one of Docker, Singularity, or Conda must be installed. We offer configurations for local execution, LSF and SLURM workload managers, and a simple cloud execution.

Workflow

CLEAN's input can be single- and paired-end Illumina FASTQ files, ONT FASTQ read files, and FASTA files, see Figure 1. The input is the only required parameter. The user can optionally add a FASTA file for a custom contamination reference. We provide common host genomes, e.g., *Homo sapiens*, *Mus musculus*, and *Escherichia coli*, spike-in sequences for Illumina, direct RNA ONT and DNA ONT sequencing, and an rRNA contamination reference (derived from SortMeRNA (Kopylova et al. 2012) (https://github.com/biocore/sortmerna/tree/master/data/rRNA_databases).

CLEAN concatenates all specified contaminations, e.g., to clean reads of the host and the spike-in in one step. Each input file (FASTQ and/or FASTA) is mapped against the contamination reference with minimap2 v2.18 (Li 2018). For Illumina, we also offer a kmer-based option with bbduk (https://sourceforge.net/projects/bbmap/). After the mapping, we separate mapped from unmapped reads or contigs by the primary alignment with SAMtools (Li 2018; Danecek et al. 2021). For ONT data and the DSC control, we provide the parameter --dcs_strict: only reads that map to the DCS and cover at least one of the artificial DCS ends are considered as contamination. By that, we avoid removing similar phage DNA that is actually part of, e.g., a metagenomics sample. If the user sets the parameter --min_clip, mapped reads are filtered by the total length (sum of both ends) of the soft-clipped positions. If --min_clip >= 1, the total number is considered, else the fraction of soft-clipped positions to the read length. The user can optionally specify FASTA files with --keep. Input reads are separately mapped to this reference. If a read maps to the "keep"-reference but was classified as contamination before, CLEAN moves the read to the set of clean reads. Thus, the user can reduce false negatives. This can help in particular when working with closely related species or metagenomic samples. CLEAN creates for the input files as well as the clean and contamination files quality reports with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Illumina reads), NanoPlot (De Coster et al. 2018) (ONT reads), or QUAST (FASTA files). MulitQC (Ewels et al. 2016) summarizes all quality reports and mapping statistics in an HTML report. Besides the MultiQC summary report and the de- and contaminated reads or FASTA files for direct downstream usage, CLEAN also emits the indexed mapping files in BAM format and the indexed contamination reference. If necessary, the user can further examine the results in a genome browser such as IGV (Robinson et al. 2011).

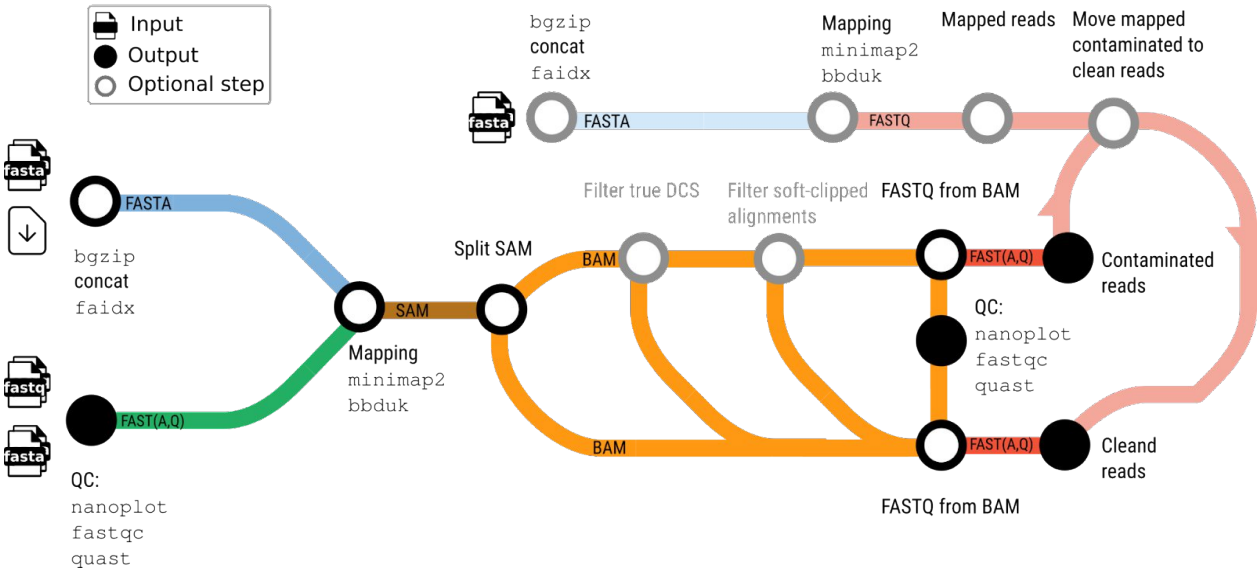


Figure 1. Schematic overview of the CLEAN workflow. Gray/blurred elements are optional and depend on the user input. The pipeline can search multiple FASTA or FASTQ inputs against a user-defined set of reference sequences (potential contamination). CLEAN automatically combines different user-defined FASTA reference sequences, built-in spike-in controls, and downloadable host species into one mapping index for decontamination. The user can also specify FASTA files comprising sequences that should explicitly not be counted as contamination. The output is finally filtered to provide well-formatted FASTA or FASTQ files for direct downstream analyses. The icons and diagram components that make up the schematic view were originally designed by James A. Fellow Yates & nf-core under a CCO license (public domain).

External resources

The user can define a contamination reference or choose from included ones. These are the currently provided host genomes in CLEAN version v1.0.0-alpha: *Homo sapiens* (Ensembl release 99), *Mus musculus* (Ensembl release 99), *Gallus gallus* (Ensembl release 99), *Escherichia coli* (Ensembl release 45), *Chlorocebus sabeus* (NCBI GCF_000409795.2), and *Columba livia* (NCBI GCF_000337935.1). A genome is only downloaded once on-demand and can be reused. The list of automatically downloadable references can be easily extended upon request or by experienced users. However, the user can also always provide additional reference FASTAs via a parameter. As an rRNA reference, we provide the rRNA database from SortMeRNA, a tool commonly used to filter rRNA from metatranscriptomic data. The database contains representative rRNA sequences from the Rfam and SILVA databases (see https://github.com/biocore/sortmerna/blob/master/data/rRNA_databases/README.txt). Spike-in sequences for direct RNA and DNA ONT sequencing are taken from Guppy, the basecaller developed by ONT: yeast enolase ENO2/YHR174W of 1.2 kb and a Lambda Phage amplicon of 3.6 kb. By further investigating the latter, we found another resource at the ONT community for the DCS sequence (https://assets.ctfassets.net/hkzaxo8a05x5/2IX56YmF5ug0kAQYoAg2Uk/159523e326b1b791e3b842c4791420a6/DNA_CS.txt). This 3560 nt long sequence is a substring of the Guppy sequence (3587 nt), where the first 27 nucleotides are duplicated at the start, see Supplemental Figure 2 and 3. The first 65 nt (Guppy 92 nt) and the last 48 nt seem to be artificial as they show no hits in a BLAST search against the NCBI nucleotide collection (nr/nt).

Results & Discussion

In the following, we will show the application of CLEAN for three common use cases: 1) removal of DNA attributed to *Chlorocebus* species (Green Monkey cell line) contamination from hybrid-assembled *Chlamydiifrater* samples improves assembly quality, 2) decontamination of the yeast enolase control in Nanopore native RNA-Seq data of a Coronavirus sequencing run, and finally, 3) fast removal of rRNA from an Illumina RNA-Seq data set.

Case study I: Removal of cell cultivation contamination from Nanopore- and Illumina-sequenced *Chlamydiaceae*

The polished assemblies based on cleaned reads reveal 1.19 Mb circular genomes and the plasmids 6 kb for each of the four *Chlamydiifrater* isolates. Without prior decontamination of the cell line DNA, contigs belonging to *Chlorocebus* species can be found in the final assemblies. Using an older version of Unicycler, running the assemblies without a CLEAN step of the raw read data also yields more fragmented final assembly results, likely due to the inflated complexity of the initial short-read graph. However, this issue was resolved by using a newer version of Unicycler, but still, contigs belonging to the used cell line could be found. Thus, decontamination of DNA belonging to a host cell line can 1) improve the general assembly process and 2) results in a much cleaner assembly.

Case study II: Yeast enolase is a highly abundant spike-in control in Nanopore native RNA-Seq data

Nanopore sequencing is currently the only technology that allows the sequencing of native RNA strands without a cDNA intermediate (Ergin, Kherad, and Alagoz 2022). This ‘direct RNA’ protocol includes the addition of a calibration strand (amplified RNA sequences of the *S. cerevisiae*

Enolase 2 mRNA, GenBank, NP_012044.1) as a spike-in positive control. Depending on the concentration of sample input RNA, this spike-in can represent a substantial fraction of the sequenced reads. In our study of direct RNA sequencing of Human Coronavirus genomes (Viehweger et al. 2019) these sequences made up 15.8% and 10.2% of the two samples, respectively. Due to algorithmic advances, re-basecalling the raw data with version 4.0.11 of the Guppy basecaller (RNA models are unchanged since then) yields more reads and a higher fraction of spike-in reads (31.4% and 31.0%, see Figure 2). Guppy does not filter these with default parameters but has an optional parameter (--calib_detect) to enable detection and filtering calibration strand reads. However, we found that this functionality does not adequately detect spike-in reads: 35.4% and 19.8% of spike-in reads were still present using this parameter. Applying CLEAN to this dataset removes all calibration strand reads (see Figure 2).

Generally, if a positive control is not needed for the experiment, we suggest skipping the addition of this spike-in. This can increase the yield of desired RNA reads by freeing up throughput capacity. For all direct RNA read data with added spike-in, we propose using CLEAN to remove these sequences reliably and quickly before downstream analyses are performed.

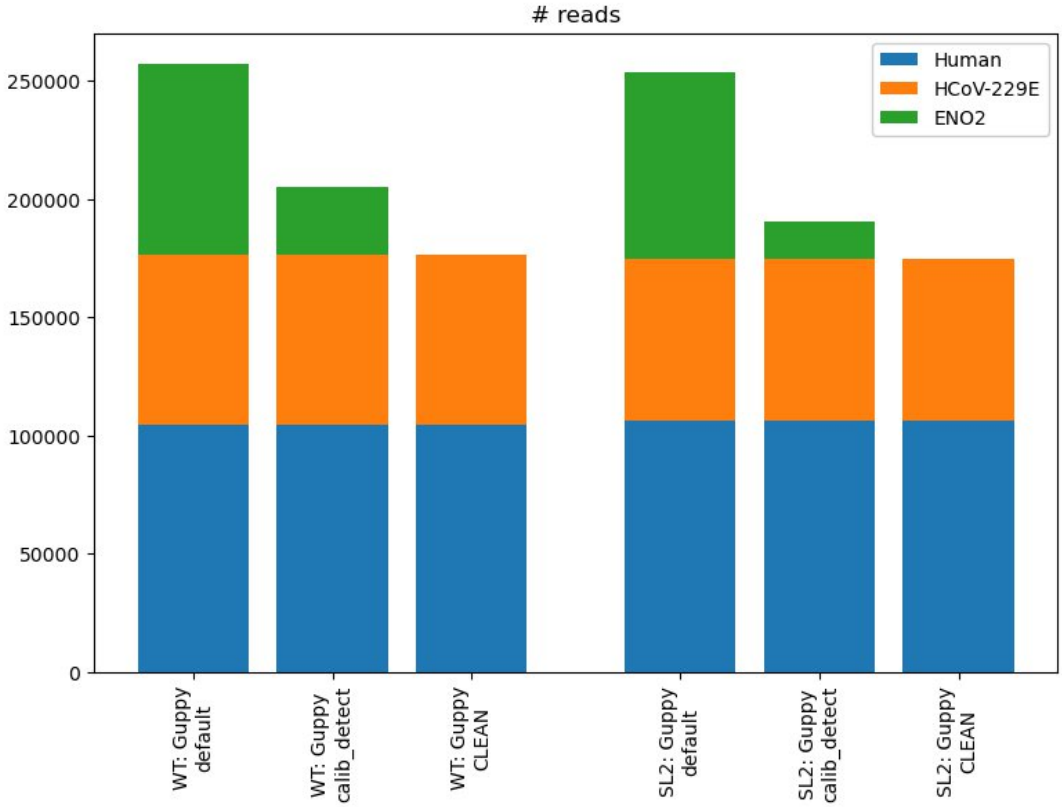


Figure 2. Number of reads mapping to the human genome, HCoV-229E or *S. cerevisiae* Enolase 2 (from bottom to top) for two HCoV-229E samples WT (left) and SL2 (right) after Guppy (default parameters), Guppy with --calib_detect or after CLEAN usage. Only CLEAN is able to remove all reads deriving from the dRNA control sequence. WT - wild type sample, SL2 - sample with different RNA secondary structure.

Case study III: Speeding up an everyday task in transcriptomics – removal of rRNA from Illumina RNA-Seq data

CLEAN performs equally compared to SortMeRNA in terms of selectivity: 99.99 % of simulated non-rRNA reads are detected as non-rRNA reads with CLEAN; SortMeRNA achieves slightly

less with 99.94 %. Regarding sensitivity, CLEAN performs at the same level as SortMeRNA with a maximum difference of 6.17 % at Set 2, see Table 1.

On the real-data sample, CLEAN runs about 1.7 fold faster than SortMeRNA, see Figure 3. Results vary slightly with <0.014 % divergence.

sensitivity	CLEAN	SortMeRNA
Set 1	96.29	99.92
Set 2	93.68	99.85
Set 3	99.21	99.88
Set 4	96.50	98.76
Set 5	96.30	99.95
Set 6	99.29	99.91

Table 1. Sensitivity comparison of CLEAN and SortMeRNA v4.3.4 for six simulated datasets consisting of 1 Mio Illumina rRNA reads each. CLEAN has a slightly decreased sensitivity than SortMeRNA; however, it is much faster (Figure 3).

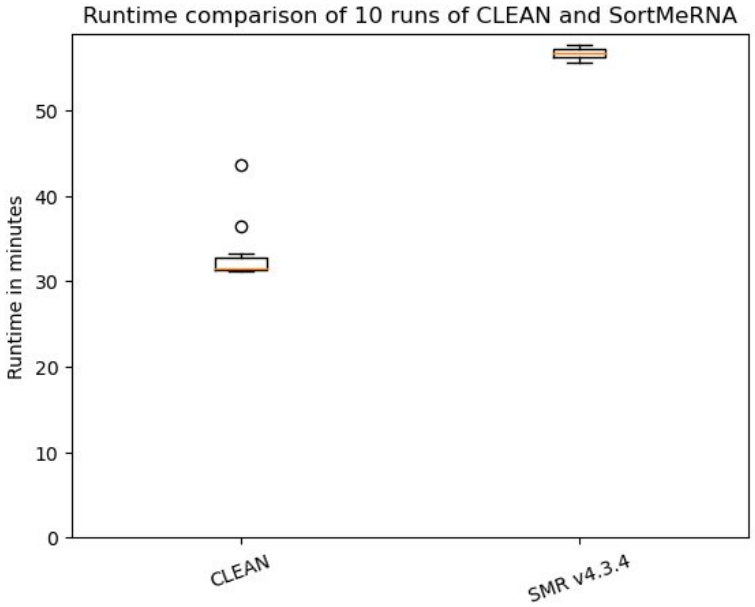


Figure 3. Comparison of the runtime for ten repeated runs of CLEAN and SortMeRNA v4.3.4. Both tools were executed on a Linux server (CPU: Opteron 6376, 64 x 2,1 GHz, RAM: 768 GB) with 30 threads. Time was measured with the Linux time command.

Methods

Test data sets and computations

Case study I: Removal of cell cultivation contamination from Nanopore- and Illumina-sequenced *Chlamydiaceae*

We obtained Nanopore (FAST5) and Illumina (FASTQ) data of two recently defined *Chlamydiifrater volucris* isolates, 15-2067_O50 (SAMEA6565319) and 15-2067_O99 (SAMEA6565320) (Vorimore et al. 2021) and re-basecalled the Nanopore raw signal data with Guppy (v6.0.0 and SUP accuracy model). In addition, we obtained data for two more unpublished isolates (15-2067_O09 and 15-2067_O77), probably also belonging to the species *Chlamydiifrater volucris*, which were cultivated on a cell line derived from *Chlorocebus sabeus* (Green monkey). DNA was extracted and sequenced with Oxford Nanopore and Illumina by colleagues at ANSES, France (Fabien Vorimore) and as described for the already published *Chlamydiifrater* strains (Vorimore et al. 2021). We used CLEAN to decontaminate all reads against DCS (--control dcs, for Nanopore) and phix (--control phix, for Illumina), *Chlorocebus sabeus* (--host csa) and the mitochondrial genome of *Chlorocebus pygerythrus* (--own NC_009747.1). Unfortunately, it is not known which species of Chlorocebus was exactly used for the construction of this cell line (Vorimore et al. 2021). Thus, we decided to use the complete chromosomal and mitochondrial genome of *C. sabeus* and add the mtDNA of *C. pygerythrus* (no chromosomal sequences are available) to increase our chances for proper decontamination (CLEAN seamlessly allows the usage of multiple references). During our analyses, we also discovered that the mitochondrial DNA of *C. pygerythrus* provides an even better matching than the mtDNA of *C. sabeus*. After decontamination, we length-filtered the ONT reads with filtlong (v0.2.0, parameters: --target_bases 1.2 * 200000000) (https://github.com/rrwick/Filtlong) and quality-trimmed Illumina reads with fastp (v0.20.1, parameters: -5 -3 -W 4 -M 20 -l 15 -x -n 5 -z 6) (Chen et al. 2018). Finally, we *de novo* assembled the cleaned and filtered short- and long-reads with Unicycler (v0.5.0, default parameters) (Wick et al. 2017) followed by independently mapping the Illumina short reads with BWA (v0.7.17) (Li 2013) to the respective resulting Unicycler assembly and subsequent polishing the assembly with polypolish (v2.2.0) (Wick and Holt 2022).

Case study II: Coronavirus native RNA sequencing with Nanopore

Virus generation, RNA isolation, sample preparation, and sequencing are detailed in (Viehweger et al. 2019). Briefly, Huh7 cells were infected with recombinant HCoV-229E variants, yielding two samples in cell culture (WT and SL2). Total RNA of these was isolated, and 1 µg of RNA in 9 µL was carried into the library preparation with the Oxford Nanopore direct RNA-Seq (DRS) protocol (SQK-RNA001). Sequencing ran for 48h on an R9.4 flow cell on a MinION device.

For this study, the raw data was basecalled with Guppy (version 4.0.11), once with and once without the --calib_detect parameter. Assignment of reads to either HCoV-229E, *S. cerevisiae* Enolase 2, or human was done by mapping to a combined reference of all three with minimap2 (version 2.17, parameters: -ax splice -k14).

Finally, we used CLEAN on the basecalled DRS reads with calibration strand detection and compared the results to the manual assignment. All commands and the plotting script are available from the supplement.

Case study III: rRNA removal from bulk RNA-Seq Illumina data

We tested and compared CLEAN's functionality to remove ribosomal RNA in terms of sensitivity and selectivity against SortMeRNA (v4.3.4) (Kopylova, Noé, and Touzet 2012). All seven simulated datasets were downloaded from (Kopylova, Noé, and Touzet 2012). Briefly, here 1 million single-end rRNA Illumina reads with a read length of 100 bp were simulated with different identities with respect to the SILVA database, or origin from truncated sections of the bacteria phylogenetic tree. One of the seven simulated samples contains non-rRNA reads to test for selectivity. We converted the provided FASTA files into FASTQ files with seqtk (v1.3-r106, https://github.com/lh3/seqtk).

To compare runtime performance, we chose a non-simulated Illumina RNA-Seq sample (GEO Accession GSM3431091) from a bat transcriptome study (Hölzer et al. 2019). For total RNA obtained from a bat (*Myotis daubentonii*) cell line, cDNA libraries were prepared utilizing the Illumina Ribo-Zero rRNA Removal Kit for human/mouse/rat. We used CLEAN with the --rrna parameter and SortMeRNA to remove rRNA reads from the sample.

We run each tool ten times with 30 threads to compare runtime differences measured with Linux's time command on a Linux server (CPU: Opteron 6376, 64 x 2,1 GHz, RAM: 768 GB).

Conclusion

We developed CLEAN to easily screen any nucleotide sequences against reference sequences to identify and remove potential contamination. Therefore, common tasks are the removal of positive controls added during library preparation, host contamination, or ribosomal RNAs. Decontamination with CLEAN can be easily pre-connected to the actual analysis as the output needs no further processing or reformatting. The pipeline uses alignment-based approaches for short- and long-reads that subsequently also allow for inspection of the reads aligned to a potential contamination reference in more detail. Furthermore, CLEAN provides quality control reports for more insights. CLEAN is freely available at https://github.com/hoelzer/clean and can be easily installed and executed using Nextflow.

Limitations

CLEAN cannot be used for the removal of unexpected contaminations. For such a task, DecontaMiner, a tool to remove contaminating sequences of unmapped reads (Sangiovanni et al. 2019), or QC-Blind, a tool for quality control and contamination screening without a reference genome (Xi et al. 2019) can be used. Other tools try to find unexpected compositions in metagenomics samples to identify contaminations (McKnight et al. 2019), (Davis et al. 2018). With CLEAN we did also not focus on the detection of cross-contamination where other tools such as ART-DeCo (Fiévet et al. 2019) can be used. Furthermore, CLEAN should not be used where tools with higher sensitivity are available, e.g., SortMeRNA for rRNA annotation and Kraken2 for taxonomic classification.

Availability of Supporting Source Code and Requirements

- Project name: CLEAN
- Project home page: https://github.com/hoelzer/clean
- Operating system(s): Platform independent due to workflow managment system and container usage

- Programming language: Nextflow, Bash
- Other requirements: Nextflow v21.04.0 or higher (compatible on POSIX systems and Windows via WSL; requires Bash 3.2 or higher, Java 11 up to 18), Conda or Singularity or Docker
- License: BSD3

Data Availability

The user manual is available on GitHub. All supporting analysis scripts are available in OSF (https://osf.io/CUXEM/, DOI 10.17605/OSF.IO/CUXEM). Data used in this work are available in public databases:

Study case I: SRA BioSample IDs SAMEA6565319 (15-2067_O50), SAMEA6565320 (15-2067_O99) and https://osf.io/DKRB5/ (15-2067_O09 and 15-2067_O77)

Study case II: https://osf.io/UP7B4/, DOI 10.17605/OSF.IO/UP7B4

Study case III: SRA BioSample ID SAMN10246232

Declarations

Competing interests

CB, AV, and MH hold shares of nanozoo GmbH.

Funding

This work was supported by the European Centre for Disease Prevention and Control (grant number 2021/008 ECD.12222 to ML). The computational experiments were also tested on resources of the Friedrich Schiller University Jena supported in part by DFG grants INST 275/334-1 FUGG and INST 275/363-1 FUGG.

Authors' contributions

MH provided conceptualization, initial design, and a first implementation. ML optimized the pipeline code, realized the final implementation, conducted the experiments, and created the figures. SK performed the benchmark for the Coronavirus dRNA-Seq experiment and provided corresponding results and methods. CB and AV provided the initial backbone code structure for the workflow. MH and ML wrote the first draft of the manuscript. All authors actively participated in the writing and final editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

We thank Fabien Vorimore from ANSES, France for sequencing of the two *Chlamydiifrater* strains and providing the raw data for our benchmark. We thank Stephan Fuchs from RKI, Germany for fruitful discussions.

372 # References

373 Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor,
374 Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. 2019. "A New Genomic Blueprint
375 of the Human Gut Microbiota." Nature 568 (7753): 499–504.

376 Boettiger, Carl. 2015. "An Introduction to Docker for Reproducible Research."
377 https://doi.org/10.1145/2723872.2723882.

378 Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. "Fastp: An Ultra-Fast All-in-One
379 FASTQ Preprocessor." Bioinformatics 34 (17): i884–90.

380 Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O.
381 Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." GigaScience
382 10 (2). https://doi.org/10.1093/gigascience/giab008.

383 De Coster, Wouter, Svenn D'Hert, Darrin T. Schultz, Marc Cruts, and Christine Van Broeckhoven.
384 2018. "NanoPack: Visualizing and Processing Long-Read Sequencing Data." Bioinformatics 34
385 (15): 2666–69.

386 Di Tommaso, Paolo, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and
387 Cedric Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." Nature
388 Biotechnology 35 (4): 316–19.

389 Ergin, Selvi, Nasim Kherad, and Meryem Alagoz. 2022. "RNA Sequencing and Its Applications in
390 Cancer and Rare Diseases." Molecular Biology Reports 49 (3): 2325–33.

391 Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize
392 Analysis Results for Multiple Tools and Samples in a Single Report." Bioinformatics 32 (19):
393 3047–48.

394 Fiévet, Alice, Virginie Bernard, Henrique Tenreiro, Catherine Dehainault, Elodie Girard, Vivien
395 Deshaies, Philippe Hupe, et al. 2019. "ART-DeCo: Easy Tool for Detection and Characterization
396 of Cross-Contamination of DNA Samples in Diagnostic next-Generation Sequencing Analysis."
397 European Journal of Human Genetics: EJHG 27 (5): 792–800.

398 Grüning, Björn, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H.
399 Tomkins-Tinch, Renan Valieris, Johannes Köster, and Bioconda Team. 2018. "Bioconda:
400 Sustainable and Comprehensive Software Distribution for the Life Sciences." Nature Methods 15
401 (7): 475–76.

402 Hölzer, Martin, Andreas Schoen, Julia Wulle, Marcel A. Müller, Christian Drosten, Manja Marz,
403 and Friedemann Weber. 2019. "Virus- and Interferon Alpha-Induced Transcriptomes of Cells
404 from the Microbat Myotis Daubentonii." iScience 19 (September): 647–61.

405 Hu, Taishan, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. 2021. "Next-Generation Sequencing
406 Technologies: An Overview." Human Immunology 82 (11): 801–11.

407 Kopylova, Evguenia, Laurent Noé, and Hélène Touzet. 2012. "SortMeRNA: Fast and Accurate
408 Filtering of Ribosomal RNAs in Metatranscriptomic Data." Bioinformatics 28 (24): 3211–17.

409 Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-
410 MEM." https://doi.org/10.48550/ARXIV.1303.3997.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” Bioinformatics 34 (18): 3094–3100.

Lu, J., Rincon, N., Wood, D.E. et al. 2022. “Metagenome analysis using the Kraken software suite.” Nat Protoc . https://doi.org/10.1038/s41596-022-00738-y

Mukherjee, Supratim, Marcel Huntemann, Natalia Ivanova, Nikos C. Kyrpides, and Amrita Pati. 2015. “Large-Scale Contamination of Microbial Isolate Genomes by Illumina PhiX Control.” Standards in Genomic Sciences 10 (March): 18.

Porter, Ashleigh F., Joanna Cobbin, Ci-Xiu Li, John-Sebastian Eden, and Edward C. Holmes. 2021. “Metagenomic Identification of Viral Sequences in Laboratory Reagents.” Viruses 13 (11): 2122.

Qi, Meifang, Utthara Nayar, Leif S. Ludwig, Nikhil Wagle, and Esther Rheinbay. 2021. “cDNA-Detector: Detection and Removal of cDNA Contamination in DNA Sequencing Libraries.” BMC Bioinformatics 22 (1): 611.

Raz, Tal, Philipp Kapranov, Doron Lipson, Stan Letovsky, Patrice M. Milos, and John F. Thompson. 2011. “Protocol Dependence of Sequencing-Based Gene Expression Measurements.” PloS One 6 (5): e19287.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” Nature Biotechnology 29 (1): 24–26.

Rumbavicius, Ignas, Rounge, Trine B. and Rognes, Torbjorn. 2022. “HoCoRT: Host contamination removal tool” bioRxiv. doi: https://doi.org/10.1101/2022.11.18.517030

Sangiovanni, Mara, Ilaria Granata, Amarinder Singh Thind, and Mario Rosario Guarracino. 2019. “From Trash to Treasure: Detecting Unexpected Contamination in Unmapped NGS Data.” BMC Bioinformatics 20 (Suppl 4): 168.

Steinegger, Martin, and Steven L. Salzberg. 2020. “Terminating Contamination: Large-Scale Search Identifies More than 2,000,000 Contaminated Entries in GenBank.” Genome Biology 21 (1): 115.

Viehweger, Adrian, Sebastian Krautwurst, Kevin Lamkiewicz, Ramakanth Madhugiri, John Ziebuhr, Martin Hölzer, and Manja Marz. 2019. “Direct RNA Nanopore Sequencing of Full-Length Coronavirus Genomes Provides Novel Insights into Structural Variants and Enables Modification Analysis.” Genome Research 29 (9): 1545–54.

Vorimore, F., Hölzer, M., Liebler-Tenorio, E. M., Barf, L. M., Delannoy, S., Vittecoq, M., ... & Sachse, K. (2021). Evidence for the existence of a new genus Chlamydiifrater gen. nov. inside the family Chlamydiaceae with two new species isolated from flamingo (Phoenicopterus roseus): Chlamydiifrater phoenicopteri sp. nov. and Chlamydiifrater volucris sp. nov. Systematic and Applied Microbiology, 44(4), 126200.

Wick, Ryan R., and Kathryn E. Holt. 2022. “Polypolish: Short-Read Polishing of Long-Read Bacterial Genome Assemblies.” PLoS Computational Biology 18 (1): e1009802.

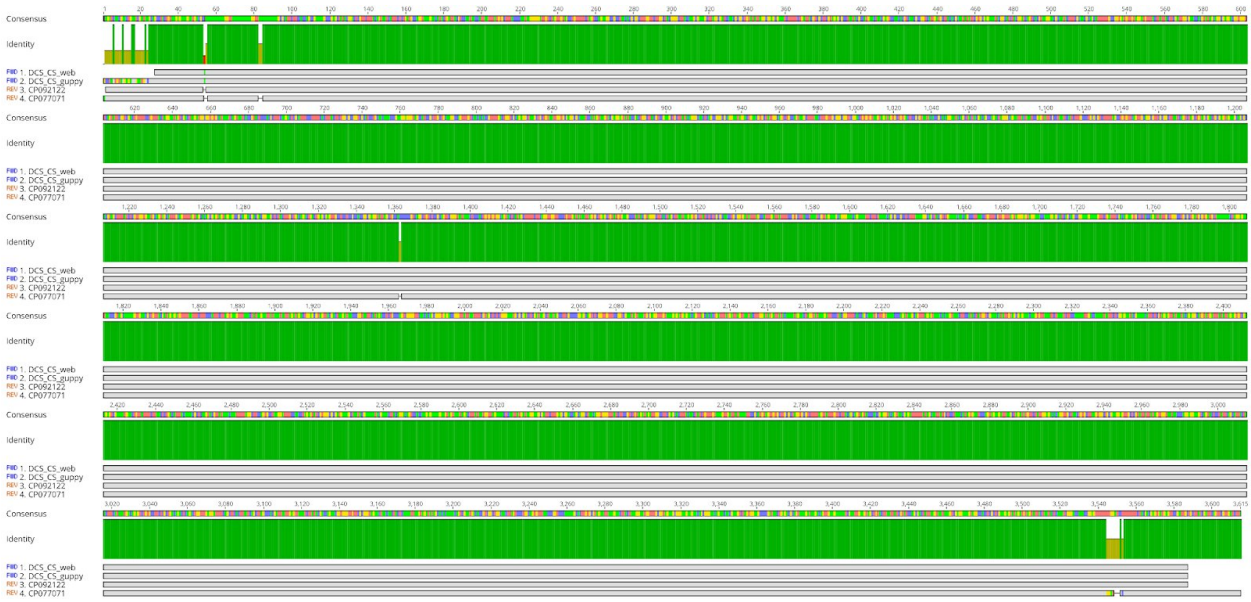
Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017. “Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads.” PLoS Computational Biology 13 (6): e1005595.

Wolf, Jochen B. W. 2013. “Principles of Transcriptome Analysis and Gene Expression Quantification: An RNA-Seq Tutorial.” Molecular Ecology Resources 13 (4): 559–72.

Xi, Wang, Yan Gao, Zhangyu Cheng, Chaoyun Chen, Maozhen Han, Pengshuo Yang, Guangzhou Xiong, and Kang Ning. 2019. “Using QC-Blind for Quality Control and Contamination Screening of Bacteria DNA Sequencing Data Without Reference Genome.” Frontiers in Microbiology 10 (July): 1560.

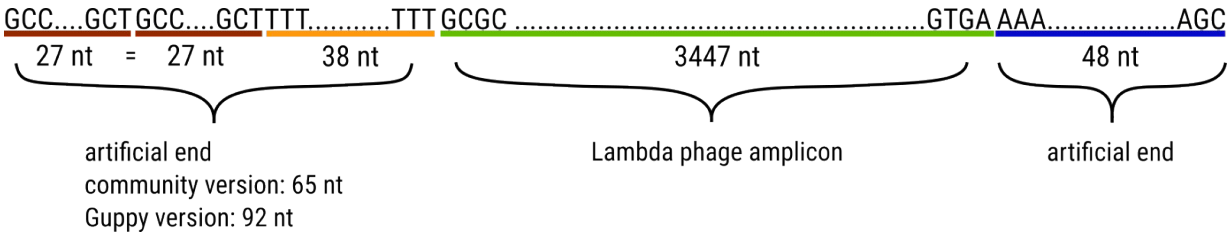
Zhao, Shanrong, Ying Zhang, Ramya Gamini, Baohong Zhang, and David von Schack. 2018. “Evaluation of Two Main RNA-Seq Approaches for Gene Quantification in Clinical RNA Sequencing: polyA+ Selection versus rRNA Depletion.” Scientific Reports 8 (1): 4781.

Supplement



Supplement Figure 1. Geneious Prime (v2021.2.2, Geneious alignment, default parameters, https://www.geneious.com) alignment of E. coli (CP077071.1), Klebsiella quasipneumoniae subsp. similipneumoniae plasmids (CP092122.1) and DCS control sequences from Guppy (DCS_CS_guppy) and the ONT community (DCS_CS_web, https://assets.ctfassets.net/hkzaxo8a05x5/2IX56YmF5ug0kAQYoAg2Uk/159523e326b1b791e3b842c4791420a6/DNA_CS.txt). The high similarity suggests that both plasmids are contaminations and falsely classified as plasmids.

Supplement Figure 2. Geneious Prime (v2021.2.2, Geneious alignment, default parameters, https://www.geneious.com) alignment of DCS control sequences from Guppy (DCS_CS_guppy) and the ONT community (DCS_CS_web, https://assets.ctfassets.net/hkzaxo8a05x5/2IX56YmF5ug0kAQYoAg2Uk/159523e326b1b791e3b842c4791420a6/DNA_CS.txt). Sequences are identical except for the first 27 nt in the Guppy version, which are duplicated subsequently.



Supplement Figure 3. Schematic illustration of the DCS control sequence: artificial ends frame a part of the Lambda phage genome. Available sequences (ONT community and Guppy installation) differ by a duplication of the first 27 nucleotides.