NanoPack: visualizing and processing long read sequencing data

Wouter De Coster[1*], Svenn D'Hert[2], Darrin T. Schultz[3], Marc Cruts[1] and Christine Van Broeckhoven[1]

[1]Neurodegenerative Brain Diseases group, Center for Molecular Neurology, VIB & University of Antwerp, 2610 Antwerp, Belgium
[2]Bioinformatics, Neuromics Support Facility, Center for Molecular Neurology, VIB & University of Antwerp, 2610 Antwerp, Belgium
[3]Department of Biomolecular Engineering and Bioinformatics, University California Santa Cruz, Santa Cruz, CA 95064, USA

*To whom correspondence should be addressed

Wouter de Coster, MSc

Neurodegenerative Brain Diseases research group
VIB Center for Molecular Neurology
University of Antwerp – CDE, Universiteitsplein 1, 2610 Antwerp, Belgium
E-mail: wouter.decoster@molgen.vib-ua.be; Phone: +3232651951

Abstract

Summary: Here we describe NanoPack, a set of tools developed for visualization and processing of long read sequencing data from Oxford Nanopore Technologies and Pacific Biosciences.

Availability and Implementation: The NanoPack tools are written in Python3 and released under the GNU GPL3.0 Licence. The source code can be found at https://github.com/wdecoster/nanopack, together with links to separate scripts and their documentation. The scripts are compatible with Linux, Mac OS and the MS Windows 10 subsystem for linux and are available as a graphical user interface, web service at http://nanoplot.bioinf.be and command line tools.

Contact: wouter.decoster@molgen.vib-ua.be

Introduction

The dominant sequencing by synthesis technology is characterized by sequencing a fixed short read length template (50-300bp) with high accuracy (error rate <1%) (Goodwin *et al.*, 2016). In contrast, long read sequencing methods from Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) routinely achieve read lengths of 10kb, with a long tail of up to 900kb for ONT (Jain, Koren, *et al.*, 2017). These long reads come with a tradeoff of lower accuracy of about 85-95% (Jain, Tyson, *et al.*, 2017; Giordano *et al.*, 2017). It is evident that these characteristics make many existing Illumina-tailored QC tools, such as FastQC (Babraham Bioinformatics, 2010), suboptimal or unusable for long read technologies. NanoPack, a set of Python scripts for visualizing and processing long read sequencing data, was developed to partially bridge this gap.

Software description

Installation and dependencies

NanoPack and individual scripts are available through the public software repositories PyPI using pip and bioconda through conda (Dale *et al.*, 2017). The scripts build on a number of third party Python modules: matplotlib (Hunter, 2007), pysam (Li *et al.*, 2009; Heger, 2009), pandas (McKinney, 2011), numpy (Walt *et al.*, 2011), seaborn (Waskom *et al.*, 2017) and biopython (Cock *et al.*, 2009).

Scripts for statistic evaluation and visualization

NanoStat produces a comprehensive statistical data summary (Supplementary table 2). NanoPlot and NanoComp produce informative QC graphs displaying multiple aspects of sequencing data (Figure 1, Supplementary table 1) and accept input data in (compressed) fastq, bam and albacore summary files or multiple of the same. All plots and summary statistics are combined in an html report. Because long and variable read lengths may be challenging to interpret on a linear axis, there is also an option to plot the read lengths on a log scale. Plots can be produced in standard image file formats including png, jpg, pdf and svg. NanoPlot produces read length histograms, cumulative yield plots, violin plots of read length and quality over time, and bivariate plots comparing the relationship between read lengths, quality scores, reference identity and read mapping quality. Better insight in big datasets can be obtained using bivariate plots with a 2D kernel density estimation or hexagonal bins (Figure 1E, F, S3). Optional arguments include random downsampling of reads and removing all reads above a length cutoff or below a quality cutoff. Data from a multiplexed experiment in albacore summary format can be separated, resulting in plots per sample. NanoComp performs comparison across barcodes or runs of read length and quality distributions, number of reads, throughput and reference identity.

Scripts for data processing

NanoFilt and NanoLyse were developed for processing reads in streaming applications and therefore have a minimal memory footprint and can be integrated in existing pipelines prior to alignment. NanoFilt is a tool for read filtering and trimming. Filtering can be performed based on mean read quality, read length and mean GC content. Trimming can be done with a user-specified number of nucleotides from either read ends. NanoLyse is a tool for rapid removal of contaminant DNA, using the Minimap2 aligner through the mappy Python binding (Li, 2017). A typical application would be the removal of the lambda phage control DNA fragment supplied by ONT, for which the reference sequence is included in the package. However, this approach may lead to unwanted loss of reads from regions highly homologous to the lambda phage genome.

Examples and discussion

The NanoPlot and NanoComp examples (Figure 1) are based on an ONT E.coli dataset from an ultra-long read protocol sequenced on an R9.4 MinION flow cell (Quick and Loman, 2017) generating 150,735 reads, base called using Albacore 2.0.2 and aligned to the E.coli reference genome using Minimap2 (Li, 2017). The cumulative yield (figure 1A) shows a lower efficiency when the flow cell wears out. A heat map of the physical layout of the MinION flow cell (Figure 1B) highlights more productive channels and could potentially identifying suboptimal loading conditions, such as introduction of an air bubble. The mean base call quality per 6 hours interval (Figure 1C) shows a uniform high quality in the beginning, with lower quality reads after 24h. In a bivariate plot comparing log transformed read lengths with their mean quality score (Figure 1E) the majority of reads can be identified at lengths of 10kb and quality scores of 12 by the color intensity of the hexagonal bins, with a subgroup of low quality short reads. Plotting the mean quality against the per read percent reference identity (as a proxy for accuracy) (Figure 1F) highlights a strong correlation, here with the number of reads plotted using a kernel density

estimate. Additional examples from NanoPlot can be found in the supplementary information online, including standard and log transformed histograms, optionally with the N50 metric (S1,2) and a bivariate plot comparing effective read length with aligned read length (S4), identifying reads which are only partially aligned to the reference genome.

The NanoComp plot (figure 1D) compares the log transformed read lengths of the same E.coli dataset to a K. pneumoniae (Wick *et al.*, 2017) and a human PromethION dataset (unpublished), clearly showing differences in the length profile with far longer reads in the E.coli dataset, standard read lengths in the library prep by ligation from K. pneumoniae and suboptimal read lengths from the human sample. Additional examples from NanoComp can be found in the supplementary information online, indicating that the K. pneumoniae library has both the highest yield (S5) and on average higher quality scores (S6) than both the human and E.coli dataset, but a comparable percent identity (S7) with the human dataset.

Conclusion

NanoPack is a package of efficient Python scripts for visualization and processing of long read sequencing data available on all major operating systems. Installation from the PyPI and bioconda public repositories is trivial, automatically taking care of dependencies. The plotting tools are flexible and customizable to the users need. Using a single NanoPlot or NanoComp command a full html report containing all summary statistics and plots can be prepared, and the software is easily accessible through the graphical user interface and web service, in addition to the command line scripts.

References

Babraham Bioinformatics,S.A. (2010) FastQC A Quality Control tool for High Throughput Sequence Data.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Dale,R. *et al.* (2017) Bioconda: A sustainable and comprehensive software distribution for the life sciences. *bioRxiv*, 207092.

Giordano,F. *et al.* (2017) De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci. Rep.*, **7**, 3935.

Goodwin,S. *et al.* (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.

Heger,A. (2009) pysam Github.

Hunter,J.D. (2007) Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.*, **9**, 90–95.

Jain,M., Tyson,J.R., *et al.* (2017) MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Res.*, **6**, 760.

Jain,M., Koren,S., *et al.* (2017) Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv*, 128835.

Li,H. (2017) Minimap2: fast pairwise alignment for long DNA sequences. *arXiv [q-bio.GN]*.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

McKinney,W. (2011) pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 1–9.

Quick,J. and Loman,N. (2017) Thar she blows! Ultra long read method for nanopore sequencing · Loman Labs.

Walt,S. van der *et al.* (2011) The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.*, **13**, 22–30.

Waskom,M. *et al.* (2017) seaborn.

Wick,R.R. *et al.* (2017) Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*, **3**.

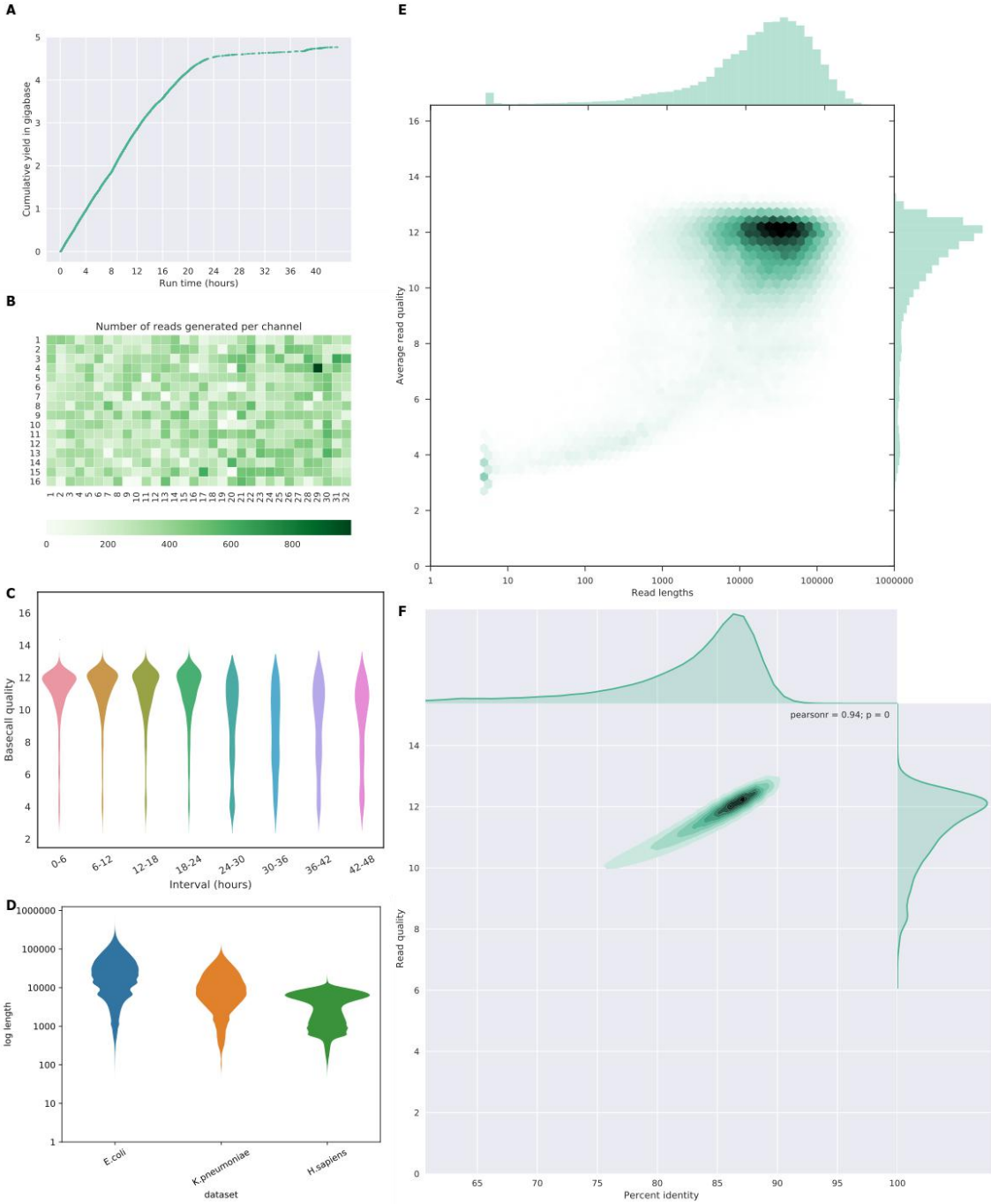


Figure 1 Examples of plots of NanoPlot and NanoComp. A Cumulative yield plot **B** Flow cell activity heatmap showing number of reads per channel. **C** Violin plots comparing base call quality over time. **D** NanoComp plot comparing log transformed read lengths of the E.coli dataset with a K.pneumoniae and human dataset. **E** bivariate plot of log transformed read length against base call quality with hexagonal bins and marginal histograms. **F** bivariate plot of base call quality against percent identity with a kernel density estimate and marginal density plots.

Funding

The study was in part funded by the VIB (Flanders Institute for Biotechnology, Belgium), the University of Antwerp and the Flanders Agency for Innovation and Entrepreneurship (VLAIO). W.D.C. is a recipient of a PhD fellowship from VLAIO. D.T.S. is supported by NSF DGE 1339067.

Acknowledgments

The authors acknowledge Mick Watson for contributing the mean GC content filtering to NanoFilt and Andreas Sjödin for maintaining the bioconda build recipes. The authors are also thankful to the many users who provided helpful suggestions and feature requests for these scripts.

Supplementary information

Supplementary table 1: NanoPlot plots

Plot	input data					Options
	Fastq	Fastq rich	Fastq minimal	Bam	Albacore Summary	
Histogram of read length	x	x	x	x	x	N50
Histogram of log transformed read length	x	x	x	x	x	N50
Bivariate plot of length against base call quality	x	x		x	x	log transformation
Heatmap of reads per channel		x			x	
Cumulative yield plot		x	x		x	
Violin plot of read length over time		x	x		x	
Violin plot of base call quality over time		x			x	
Bivariate plot of aligned read length against sequenced read length				x		
Bivariate plot of percent reference identity against read length				x		log transformation
Bivariate plot of percent reference identity against base call quality				x		
Bivariate plot of mapping quality against read length				x		log transformation
Bivariate plot of mapping quality against basecall quality				x		

Supplementary table 2: NanoStat example output

General statistics

Number of reads:	150,735
Total bases:	4,769,579,795
Median read length:	18,796.00
Mean read length:	31,642.15
Readlength N50:	61,411

Top 5 read lengths and their average base call quality score

Length: 763,488bp	Q: 11.64
Length: 760,365bp	Q: 11.46
Length: 744,361bp	Q: 10.38
Length: 656,649bp	Q: 12.13
Length: 628,176bp	Q: 11.96

Top 5 average base call quality scores and their read lengths

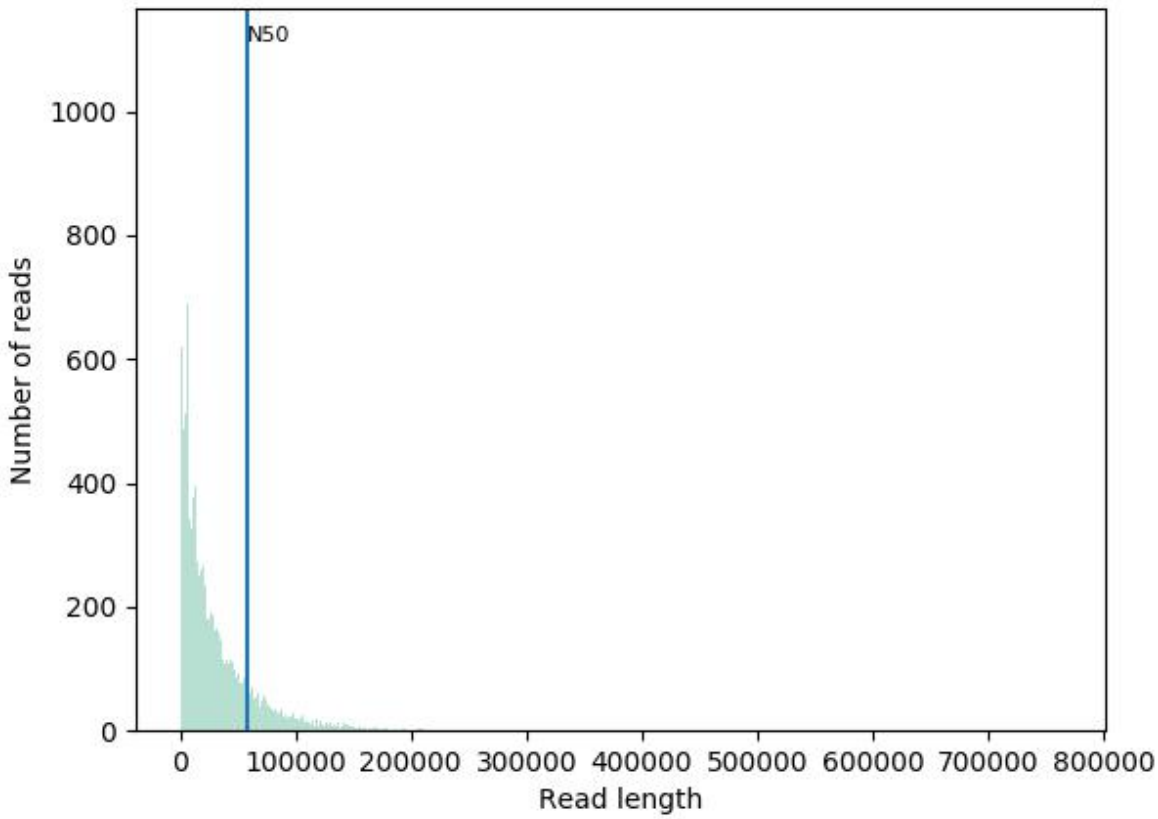
Length: 323bp	Q: 16.57
Length: 29,090bp	Q: 14.1
Length: 8,297bp	Q: 14.05
Length: 1,233bp	Q: 13.94
Length: 112,093bp	Q: 13.92

Number of reads and fraction above quality cutoffs:

Q5:	143,721 (95.35%)
Q10:	114,990 (76.29%)
Q15:	1 (0.00%)
Q20:	0 (0.00%)

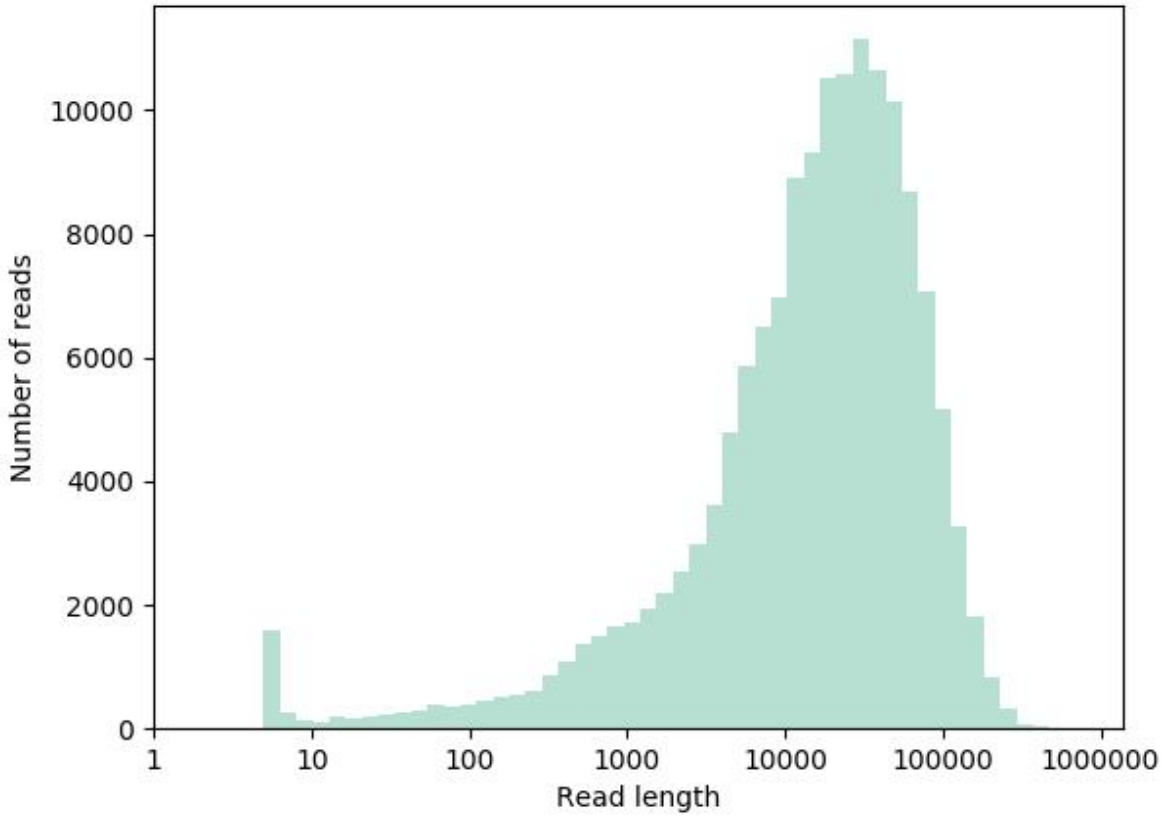
Data produced using 506 active channels.

Supplementary Figure S1: Histogram of read lengths



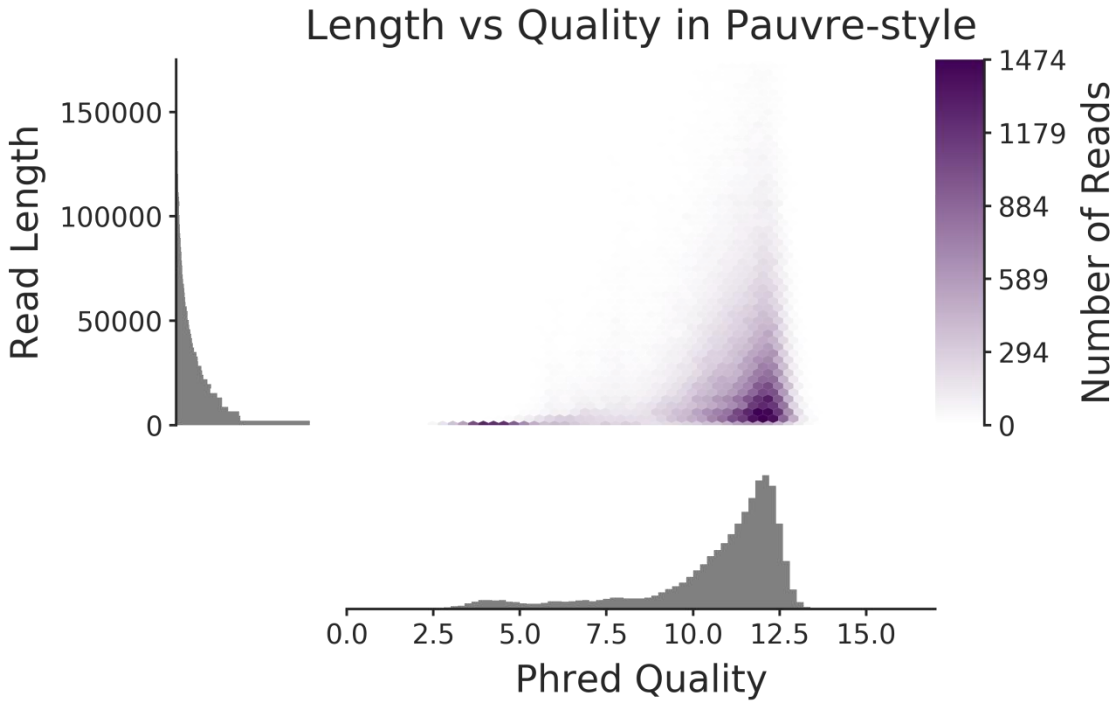
NanoPlot histogram of read lengths in bins of 100bp, showing the read length N50 value.

Supplementary Figure S2: Histogram of log transformed read lengths



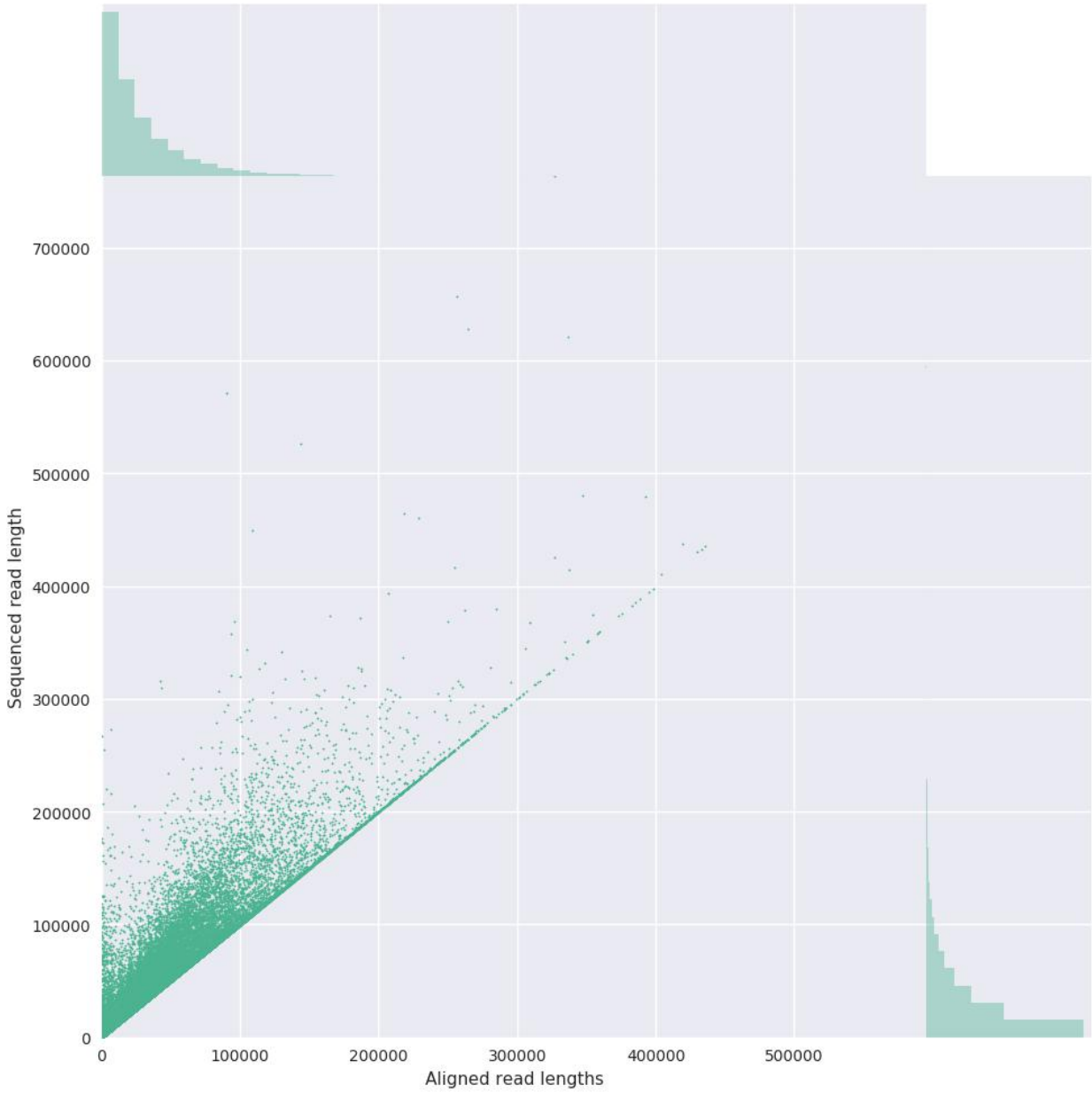
NanoPlot histogram of log transformed read lengths

Supplementary Figure S3: Quality versus length plot in ‘pauvre’ style



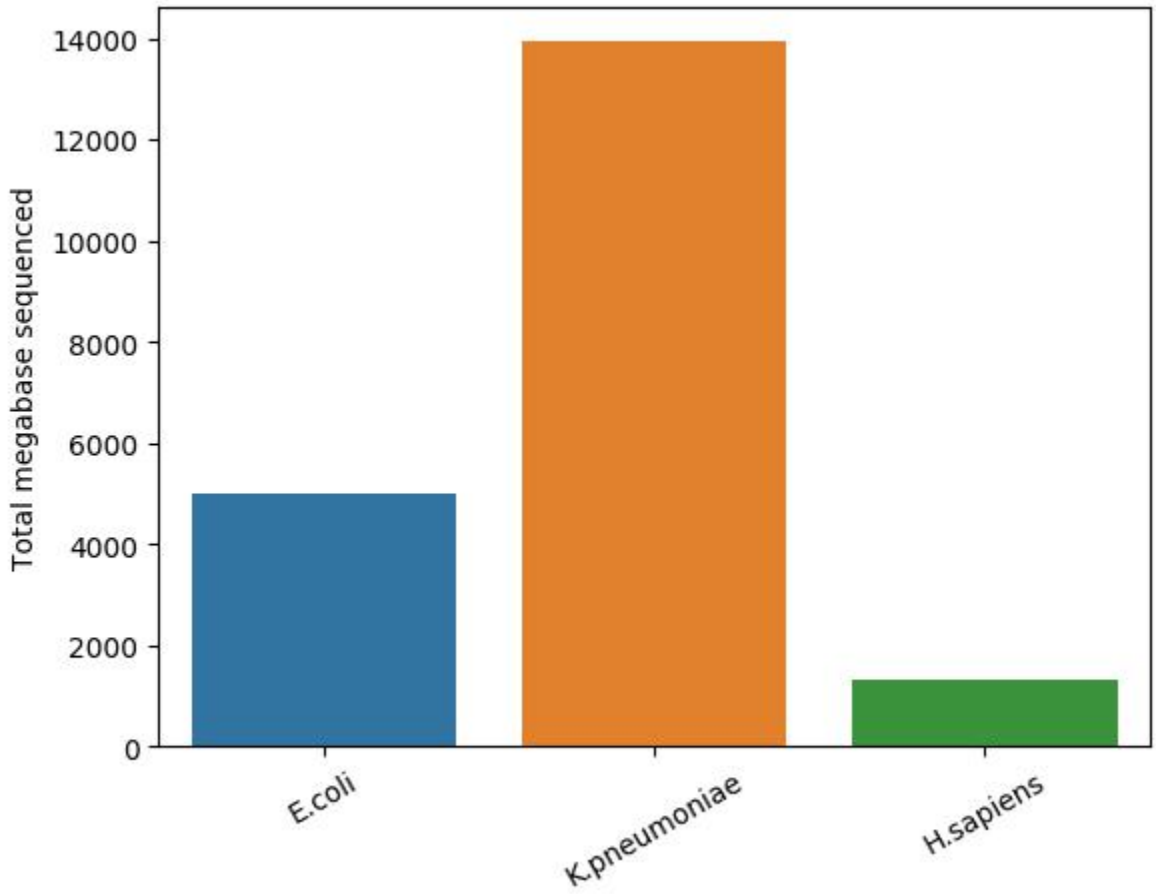
NanoPlot bivariate plot of per read reads lengths against mean basecall quality score in pauvre style.

Supplementary Figure S4: Bivariate plot of sequenced read lengths against aligned read length



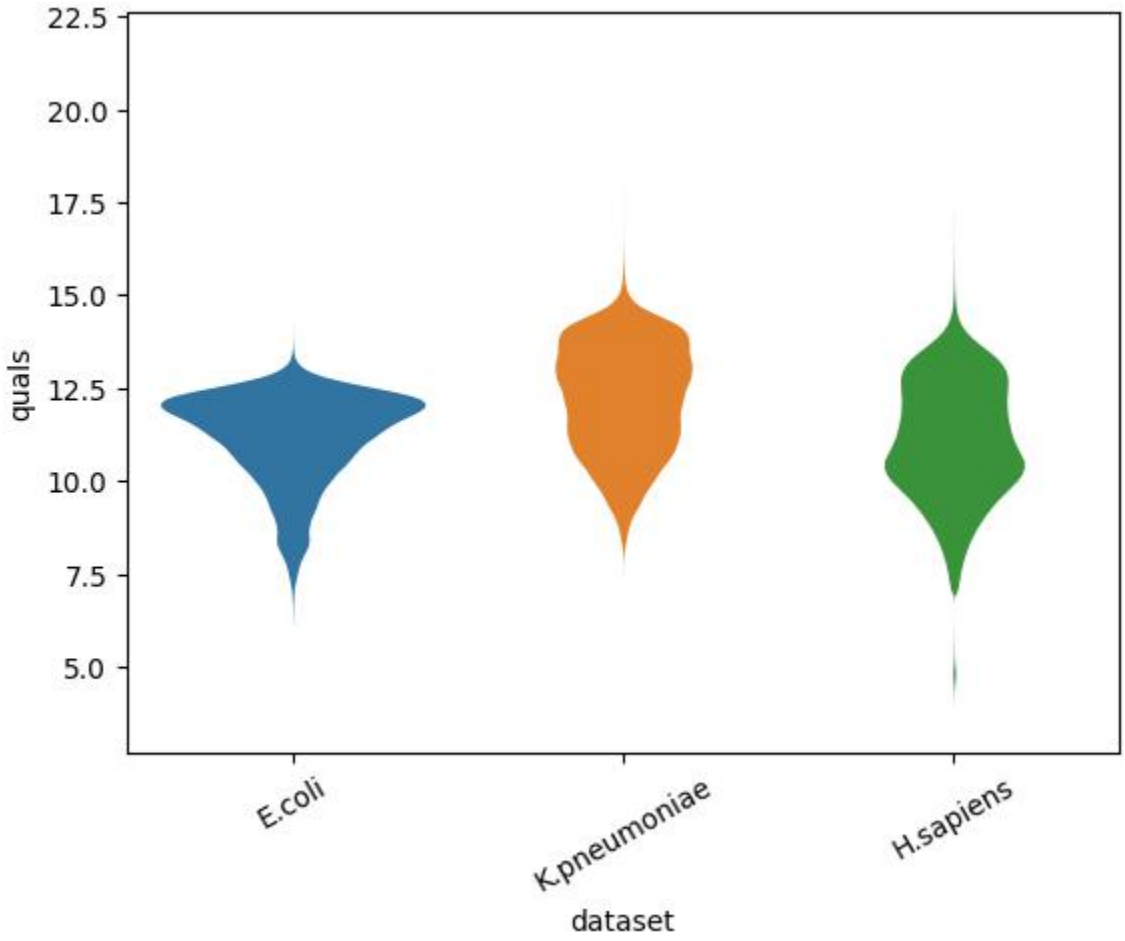
NanoPlot bivariate plot of sequenced reads lengths against aligned read lengths, showing reads not aligning end to end to the reference genome.

Supplementary Figure S5: NanoComp comparison of throughput in megabases



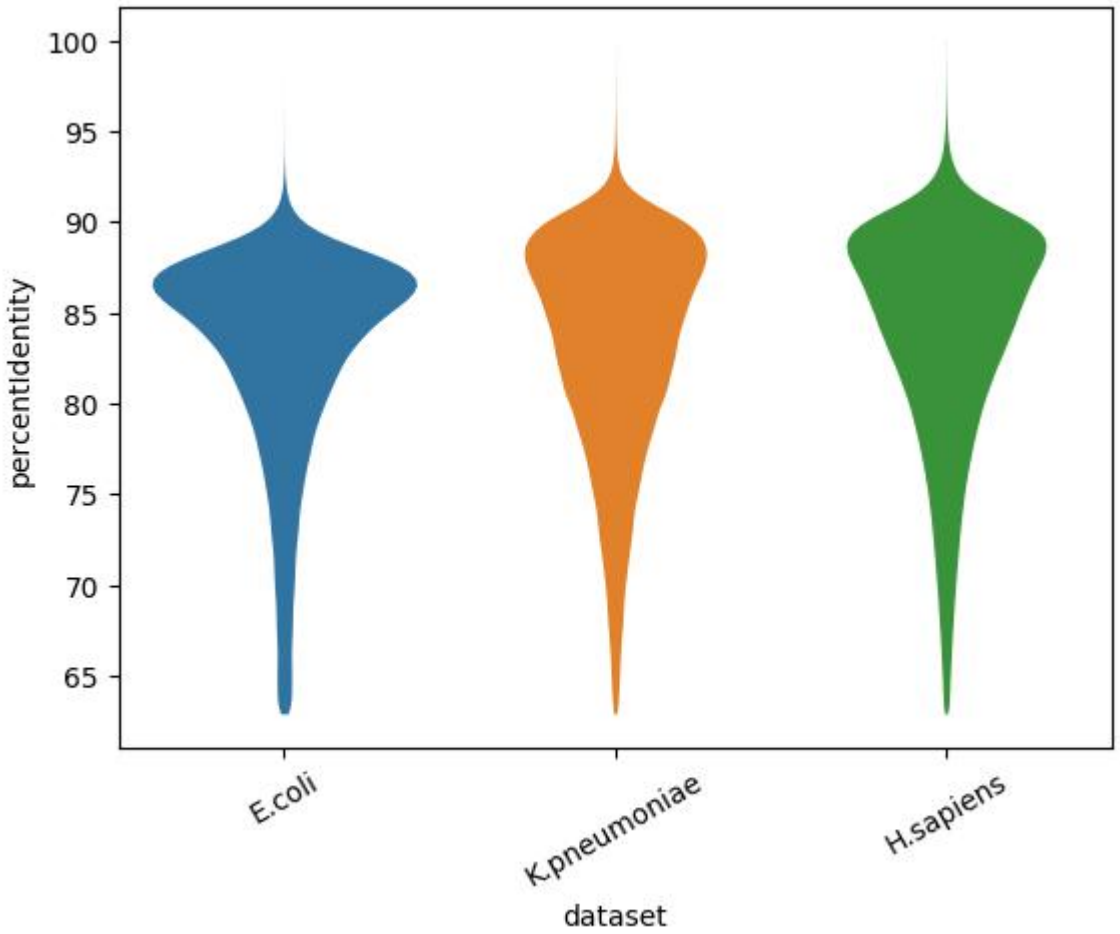
NanoComp bar chart of throughput in megabases per dataset.

Supplementary Figure S6: NanoComp comparison of quality distribution using violin plots



NanoComp violin plot of per read average read quality per dataset.

Supplementary Figure S7: NanoComp comparison of percent reference identities



NanoComp violin plot of per read percent reference identity (edit distance scaled to aligned read length) compared to the reference genome.