BMC Medical Genomics

RESEARCH

Open Access

Benchmarking Illumina and Oxford Nanopore Technologies (ONT) sequencing platforms for whole genome sequencing of bacterial genomes and use in clinical microbiology

Srinithi Purushothaman[1], Tim Roloff[1], Adrian Egli[1†] and Helena MB Seth-Smith[1*†]

Abstract

Background In microbial diagnostics, whole-genome sequencing (WGS) is used to address key questions such as species identification, presence of antimicrobial resistance genes (ARGs), virulence genes, and outbreak detection. The choice of sequencing technology is crucial to ensure high-quality data, cost-effectiveness, and efficient reporting times. We aimed to compare Illumina (short-read) and ONT (long-read) sequencing methods for WGS on different bacterial species for base accuracy and reliable taxonomic and ARG identification.

Materials and methods We used clinical isolates of ESKAPE pathogens ($n=12$) and ATCC strains ($n=8$) of varying %G + C. Illumina sequencing was performed on MiSeq (PE150) and ONT sequencing using GridION with R9.4.1 and R10.4.1 flowcells. Base-calling was performed using Guppy, Dorado, and Rerio software. We performed *de novo* assembly with Unicycler for Illumina and Flye for ONT, and two types of hybrid assemblies, Unicycler and Polypolish. We annotated genomes with Bakta and assessed the quality (QUAST, GTDB-Tk). We identified ARGs (AMRFinderPlus) and plasmids (MOB-suite). We mapped reads and called SNPs using Minimap2, Pilon, vcftools, and Snippy (Illumina). Core genome MLST analysis was conducted with Ridom Seqsphere+.

Results We observed that Illumina sequencing provided consistently high-quality reads (median Q-score 35), whereas for ONT R10.4.1, SUP model showed higher median quality (median Q-score 15.3) compared to R9.4.1 (median Q-score 13.9, SUP model). We observed that Illumina-based assemblies generated fewer genes annotated as disrupted; for ONT assemblies, the base-caller affects assembly annotation accuracy, with High accuracy (HAC) and Super accuracy (SUP) base-calling models perform better than FAST model. ONT assemblies resolved rRNA operons better than Illumina assemblies. Sequencing errors were determined by SNP calling, and varied widely by species, with ONT often generating more sequencing errors compared to Illumina. Hybrid assemblies combine accuracy and completeness effectively. Taxonomic identification and ARG detection were reliable across all methods.

†Adrian Egli and Helena MB Seth-Smith contributed equally to this work.

*Correspondence:
Helena MB Seth-Smith
hsethsmith@imm.uzh.ch

Full list of author information is available at the end of the article

Conclusion Combining Illumina and ONT technologies yielded optimal bacterial genome sequencing results, leveraging the high accuracy of short reads and improved contiguity of ONT long reads. The HAC and SUP ONT models with Dorado notably enhance genome assembly annotation and resolution of complex regions, although species-specific issues, likely due to repeat regions and base modifications, remain challenging even in SUP model with Dorado. Hybrid approaches currently offer the most comprehensive and accurate genome assemblies for clinical microbiology. For reliable cgMLST even using the most recent ONT methods, resolution must be assessed on a species-by-species basis.

Keywords Whole genome sequencing, Oxford Nanopore Technologies, Illumina

Introduction

Bacterial typing and antimicrobial susceptibility testing (AST) are important services offered by clinical microbiology laboratories for infection control and prevention, and to inform treatment strategies. Whole genome sequencing (WGS) offers detailed pathogen profiling and can be used for genome-based species identification, high-resolution typing, including cluster identification, as well as prediction of antimicrobial resistance genes (ARG) and virulence genes [1–3]. WGS data can thus be used for diagnostics and, importantly, also for surveillance programs [4, 5]. To facilitate the accreditation of clinical laboratories and ensure consistent data interpretation and sharing, standardization of WGS workflows is critical [6, 7]. A clear understanding of the advantages and limitations of different sequencing technologies is essential for their reliable integration into clinical and public health microbiology [8, 9].

Short-read, Illumina platform-generated data has been used extensively for WGS of bacterial isolates in settings spanning from high- to limited-resource [10, 11]. Illumina platforms offer workflows with turnaround times (TAT) of 2–5 days (12,13), high throughput, base call accuracy of 99.99% [12, 13], and read lengths of 75–300 bp [14]. However, Illumina data has lower resolution for repetitive regions, structural variant identification, and thus limitations in generating complete, circularized bacterial genome and plasmid assemblies [15–18].

The advent of long-read sequencing technologies over the last decade is redefining the microbial genomics landscape and promising to shed light on the uncharted regions of bacterial genomes [19]. Long-read sequencing technologies generate read lengths in kilobases, often enabling complete genome assembly [13]. Of particular interest in clinical microbiology labs is Oxford Nanopore Technologies (ONT). ONT offers rapid TAT for library preparation and sequencing (range here as for Illumina), real-time analysis, lower instrument costs, ease of implementation in resource-limited settings (e.g., handheld MinION sequencer), and the ability to generate circularized bacterial chromosome and plasmid assemblies [20–24]. However, ONT data has a lower base quality of the sequenced reads due to a high error rate compared to Illumina [14, 25].

Over the years, ONT has been dynamically modifying pore chemistries and developing better base-calling algorithms to generate high-quality reads and increase read accuracy. The speed of these developments and the subsequent rapid depreciation of previous hardware pose an issue for clinical microbiology laboratories. Frequent validation is thus required for the bioinformatic software, base-callers, and base-calling models, to ensure reproducibility and standardization. Although several studies [25–30] have evaluated long- and short-read sequencing platforms for bacterial WGS-based typing, these investigations have focused on specific Gram-negative or Gram-positive species, and comparisons across different ONT base callers and models remain limited.

Our benchmarking study aimed to analyze a range of genomically diverse bacteria, including clinically relevant pathogens of the ESKAPE group (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter spp*.). We compared sequencing data from routine Illumina workflows, ONT chemistries using the R9.4.1 and R10.4.1 flowcells, and ONT base-callers, namely Guppy, Dorado, and Rerio, with different base-calling models: FAST, High accuracy (HAC), and Super accuracy (SUP). We assessed read quality, assembly quality, taxonomic identification, cluster identification, and ARG prediction. Along with the Illumina-only and ONT-only *de novo* assemblies, we also compared hybrid assemblies using the data. With this benchmarking study, we aim to address whether the improved ONT data can generate stand-alone assemblies on a par with those from Illumina data for pathogen profiling, and to investigate the potential paradigm shift in the use of short-to-long-read sequencing platforms.

Methods

Sample selection

We selected American Type Culture Collection (ATCC) bacterial isolates ($n = 8$) with %G + C content from 30% to 66%, and clinical ESKAPE ($n = 12$) isolates. Table 1 lists the bacterial isolates used.

Two *A. baumannii* routine clinical isolates with the same resistance pattern were used to demonstrate the

Table 1 Bacterial isolates used for the study. Clinical isolates were labeled as "s" for sensitive and "r" for resistant. For example, *acibau_r* or *acibau_nr* refers to *A. baumannii* with a specific resistance pattern. The resistant and sensitive ESKAPE isolates were identified with phenotypic antimicrobial sensitivity testing, carried out with routine culture-based diagnostics

Bacterial isolates	Isolate information	Gram status	%G + C	Short name
Acinetobacter baumannii	Carbapenem resistant, bla_{OXA-23}	Negative	39.0	acibau_nr
Acinetobacter baumannii	Carbapenem resistant, bla_{OXA-23}	Negative	38.8	acibau_r
Acinetobacter pittii	Carbapenem resistant, $bla_{OXA-500}$	Negative	38.7	acipit
Burkholderia stabilis	ATCCBAA-67	Negative	66.4	bursta
Campylobacter jejuni	ATCC700819	Negative	30.6	camjej
Enterococcus faecalis	ATCC29212	Positive	37.5	entfael
Enterococcus faecium	Vancomycin sensitive	Positive	37.7	entfae_s
Enterococcus faecium	Vancomycin resistant, *vanA*	Positive	37.7	entfae_r
Escherichia coli	ATCC25922	Negative	50.4	esccol
Escherichia coli	Non-Extended spectrum beta lactamase (ESBL) producing, sensitive	Negative	50.6	esccol_s
Escherichia coli	ESBL producing, $bla_{CTX-M-15}$	Negative	50.8	esccol_r
Klebsiella pneumoniae	Carbapenem sensitive	Negative	57.3	klepne_s
Klebsiella quasipneumoniae	ATCC700603	Negative	57.7	klequa
Pseudomonas aeruginosa	ATCC27853	Negative	66.1	pseaer
Pseudomonas aeruginosa	Carbapenem sensitive	Negative	66.4	pseaer_s
Pseudomonas aeruginosa	Carbapenem resistant, bla_{VIM-2}	Negative	66	pseaer_r
Staphylococcus aureus	Oxacillin sensitive	Positive	32.8	staaur_s
Staphylococcus aureus	Oxacillin resistant, *mecA*	Positive	32.9	staaur_r
Staphylococcus aureus	ATCC25923	Positive	32.9	staaur
Streptococcus pyogenes	ATCC19615	Positive	38.5	strpyo

reliability of calling the $bla_{\text{OXA-23}}$ allele across different genomes.

DNA preparation

We cultured the ATCC isolates on Columbia agar with 5% sheep blood and the ESKAPE isolates on LB agar plates. We extracted DNA from bacterial cultures using the QIAamp DNA Mini kit (Qiagen) with the Gram-positive protocol using 20 mg/mL lysozyme, according to the manufacturer's protocol. We used DNA from a single extraction in all experiments for each isolate, with the exception of the ATCC isolates on Illumina (see below). We used the Qubit 4™ Fluorometer with 1x dsDNA HS Assay Kit™ (Thermo Fisher Scientific) to quantify the DNA concentration.

Genome sequencing

We sequenced all ESKAPE isolates on the Illumina MiSeq platform with PE150, following QIAseq FX (Qiagen) library preparation according to diagnostic routine protocols (ISO norm 15189/17025). Illumina data for the ATCC isolate libraries prepared with QIAseq FX were obtained from a previous study [31] under ENA number PRJEB31421. For ONT sequencing, the rapid barcoding kit (RBK004 and RBK114.24) was used for library preparation and sequenced with both R9.4.1 and R10.4.1 flowcells in a GridION sequencer, multiplexing twelve samples per flowcell. We used the same DNA for library preparation for both flowcell chemistries. We sequenced with the 400-bps speed and 4 kHz model for the default 72 h runtime, on flowcells each with > 1000 available pores. The demultiplexing was done using the inbuilt MinKNOW software (v22.10.7). We further sequenced the eight ATCC isolates using the inbuilt 5 kHz Dorado-SUP model provided within the MinKNOW software (v24.06.15).

Base-calling for ONT

We generated the base-calling for the Fast5 files from R9.4.1 and R10.4.1 in parallel using the three different command line base-callers Guppy (v6.4.6), Dorado (v0.3.4), and Rerio (sup@4.0.1). For Guppy and Dorado, three base-calling models (FAST, HAC - High accuracy, and SUP - Super accuracy) were used; for Rerio, only the SUP model is available. An overview of these different base-calling options and programs is shown in Fig. 1. We converted the Fast5 files into Pod5 files for Dorado base-calling using pod5 tools (v0.1.5) [32]. We set the minimum base quality score cutoff to nine for all base-calling models. The sequencing was done using the 4 kHz model. We also compared the 5 kHz Dorado-SUP model for ATCC isolates. We used the fastq files obtained after the base-calling for downstream data analysis. Base calling was performed on the high-performance computing (HPC) cluster provided by ScienceCluster, University of Zurich. The cluster is equipped with NVIDIA A100 and

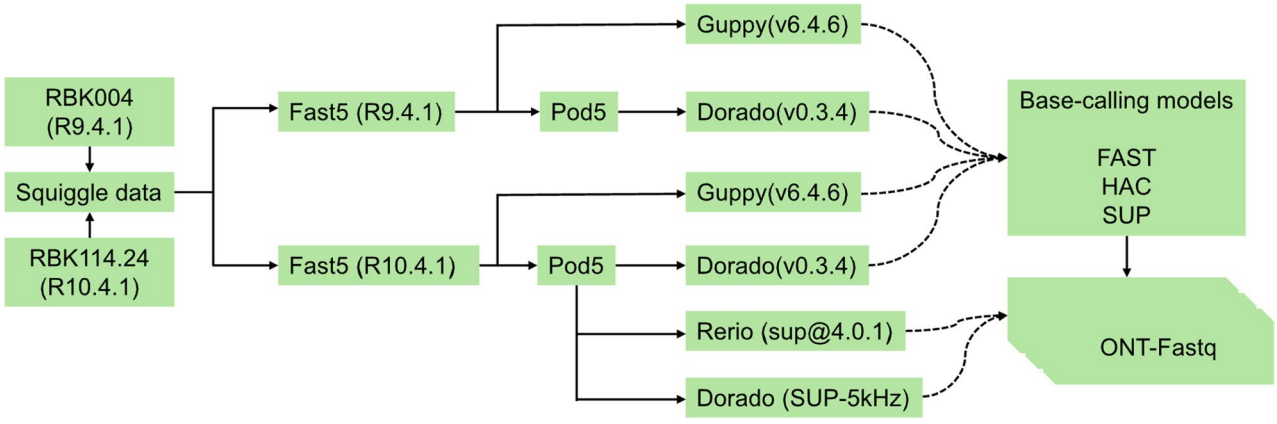

Fig. 1 Overview of different base-calling options and programs for ONT sequencing. All base callers were used in command-line model, except Dorado (SUP-5 kHz); FAST, HAC (High accuracy), and SUP (Super accurate)

V100 GPUs, supporting multi-GPU execution through CUDA (v12.0.0). A total of five GPUs were utilized for this analysis.

Illumina read QC and assembly

We performed sequencing read quality control for the Illumina data using the FastQC program (v0.11.9) [33]. The raw reads were subjected to adapter removal using Trimmomatic (v0.39) [34]. After the adapter removal, we assembled the reads using Unicycler (v0.5.0) [35]. We calculated mean genome read depth using the formula as follows: C = L*N/G; where C - Genome read depth; L - Read length; N - Total number of reads, and G - Genome size, and achieved over 30x mean read depth (our clinically validated minimum) for all samples.

ONT read QC and assembly

We filtered reads with Filtlong (v0.2.0) [36], using a minimum read length of 200 bp (--min_length 200) and retaining the top 90% of bases by quality (keep_percent 90). The sequencing read quality for the ONT data was calculated using NanoPlot (v1.40) [37]. Over 30x mean read depth was obtained for all samples, except a few assemblies generated with FAST model. We converted the fastq files to fasta using Seqtk (1.3-r106) [38] for Medaka polishing. We *de novo* assembled the filtered reads using Flye (v2.9.1-b1780) [39] and obtained the assembly summary from the Flye log. The assembled FASTA obtained from Flye was polished with Medaka (v1.7.2) [40] using reads generated with the respective base-calling models. Medaka polishing was not performed for the genomes obtained using Rerio base-calling, as no medaka polishing model is available for this respective base-caller. We utilized the Medaka (v2.0.0) [40] polishing model for bacterial methylation for the data generated using R10.4.1 flowcell chemistry for both the 4 and 5 kHz models.

Hybrid assemblies

We generated hybrid assemblies based on Illumina-first assemblies using Unicycler (v0.5.0) [35] with Illumina-only assemblies polished with long-reads generated from all possible base-callers and models, hereafter referred to as long read polished (lrp). Next, we generated ONT-first hybrid assemblies using Polypolish(v0.5.0) [41], using medaka polished ONT-only assemblies, polished using Illumina reads, hereafter referred to as short read polished (srp). For the 5 kHz sequencing, we have not performed the hybrid assemblies, as the aim is to compare whether ONT alone reads base called with the bacterial base modification genome-aware base calling model, outperforms the hybrid assemblies.

Annotation

We annotated the assembled genomes using Bakta (v1.10.3) with database version (v5.1) [42] and obtained the total number of rRNA, coding sequence (CDS), and pseudogenes from the Bakta report. Bakta relies on the Rfam database (v14.10) and Infernal models for detection and annotation of rRNA genes. Bakta identifies and annotates all three major rRNA genes, 16 S, 23 S, and 5 S. Bakta also identifies pseudogenes as frameshift mutations (insertions and deletions not including triplets) along with premature stop codons and loss of original start codons. We calculated the mean CDS length using SeqKit (v2.4.0) [43].

Reference genomes

Table 2 lists the reference genomes for the ATCC isolates. We used the lrp assemblies generated with the Rerio-SUP model as reference genomes for the ESKAPE isolates, as the Rerio-SUP model accounted for bacterial methylation.

Table 2 ATCC isolates and reference genomes used in analysis

ATCC isolates	ATCC ID	Refseq Accession number
Burkholderia stabilis	ATCCBAA-67	NZ_CP016442.1, NZ_CP016443.1, NZ_CP016444.1
Campylobacter jejuni	ATCC700819	NC_002163.1
Enterococcus faecalis	ATCC29212	NZ_CP008816.1, NZ_CP008815.1, NZ_CP008814.1
Escherichia coli	ATCC25922	CP009072.1
Klebsiella quasipneumoniae	ATCC700603	NZ_CP014696.2, NZ_CP014697.2, NZ_CP014698.2
Pseudomonas aeruginosa	ATCC27853	CP015117.1
Staphylococcus aureus	ATCC25923	NZ_CP009361.1, NZ_CP009362.1
Streptococcus pyogenes	ATCC19615	NZ_CP008926.1

Assembly quality

We assessed the assembly quality using QUAST (v5.0.2) [44]. The assembly graphs were visualized using Bandage (v0.8.1) [45]. The taxonomy identification was carried out on the assembled genome using the Genome Taxonomy Database Tool kit (GTDB-Tk) (v2.2.6) with database release (v214) [46]. We performed ARG prediction using NCBI AMRFinderPlus (v3.12) [47] on the assembled genomes, using gene length coverage and sequence identity of 95% as cutoff scores. We used MOB-suite (v3.1.9) [48] to classify assembled contigs with the identified ARG as either of chromosome or plasmid origin.

Mapping and SNP error calling

We aligned the read length filtered ONT reads, and adapter-trimmed Illumina reads to the respective reference genomes according to ESKAPE or ATCC bacterial isolates using Minimap2 (v2.24) [49] in ont-mode for ONT reads and sr-mode for Illumina reads. The BAM files were subjected to sorting, duplication removal, and indexing using Samtools (v1.15.1) [50]. We performed SNP calling from the BAM files using Pilon (v1.24) [51], and vcftools (v0.1.16) [52] to process the vcf files with retaining sites with a minimum of two alleles (--min-alleles 2) and a minimum quality threshold of 1(--minQ 1). We extracted the SNPs annotated in the vcf files using vcftools (v0.1.16). Variant calling on short reads was also tested using Snippy (v4.6.0) [53] for comparison. We calculated the read depth of the called SNPs at each position using Samtools (v1.15.1) [50].

Core genome MLST (cgMLST) analysis

We analyzed all the assemblies in Ridom Seqsphere+ (v10.0.3) [54] and used cgMLST analysis with defined schemes (cgmlst.org). Statistical tests were not feasible, as each comparison involved a single species, and species-specific effects would have influenced the results.

Results

Read and base quality

The selection of diverse ATCC type strains and clinical isolates yielded a median base quality of 35.9 (IQR 33.2–36.3) with Illumina sequencing. In contrast, sequencing using ONT with R10.4.1 flowcells and base-calling in SUP model with Dorado and Guppy resulted in a median read quality score (mean Q-score) of 15.3 (IQR 14.7–15.9) and 15.4 (IQR 14.9–16.1). This was higher than the data from R9.4.1 flowcells, which were base-called in SUP model with Dorado and Guppy, producing a median Q-score of 13.9 (IQR 13.6–14.1) and 14.2 (IQR 14–14.6) (Fig. 2a). The number of bases generated with R9.4.1 flowcells was greater than that of R10.4.1 flowcells, but the R10.4.1 flowcells produced longer reads, with a higher N50 (Fig. 2b and c). The quality metrics for each sample are provided in Supplementary Table S1. ATCC isolates resequenced with Dorado 5 kHz SUP model yielded the highest median base quality of 19.5 (IQR 19.0–19.7) compared to the R10.4.1 (4 kHz) and R9.4.1 base called reads (Supplementary Table S1). We also obtained duplex reads from the R10.4.1 flowcells. While the read quality was improved with duplex reads, the number of these reads was insufficient for downstream genome assembly [29]. Therefore, these data are not included in the results.

Computational time needed for base calling

Dorado was computationally less expensive than Guppy across different flow chemistries and base calling modes. Among the different base calling models, FAST model consistently required the least time, followed by HAC, while SUP incurred the highest computational time. On the other hand, Rerio SUP model also required lower computational time compared to Dorado-SUP. Table 3 lists the base calling time across the different base callers and the total bases generated before filtering.

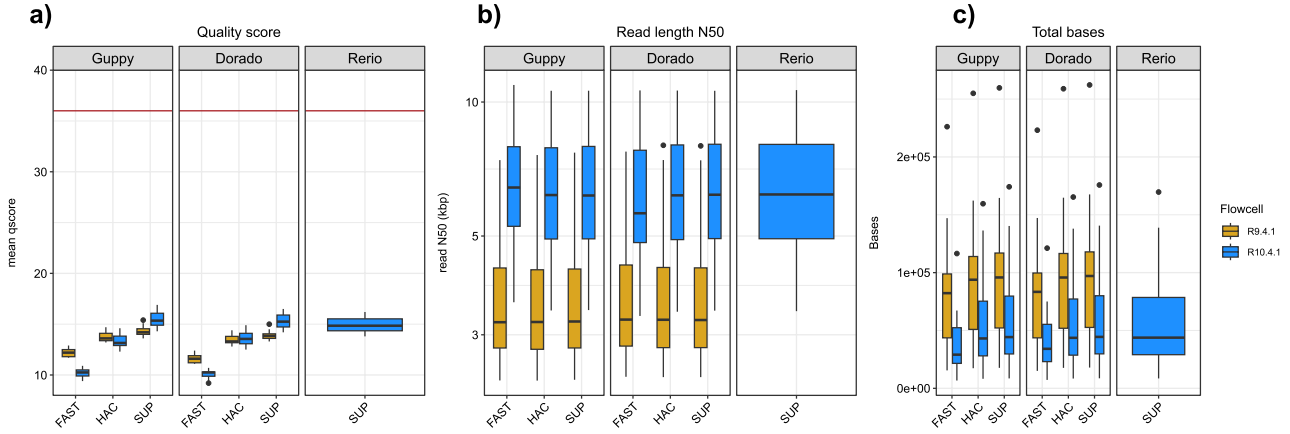

Fig. 2 ONT read quality metrics from NanoPlot for 20 sequenced isolates. **a**: mean quality score (qscore; red line represents the Illumina median quality score of 35); **b**: read length N50 in kilobase pairs, Illumina reads paired end 150 base pairs; **c**: total number of passed bases. The x-axis shows the base-caller models, and the y-axis shows the respective quality metrics. Each grid represents different ONT base-callers. The boxes represent the interquartile range, and the lines inside the boxes represent the median. The lines extending from the box (whiskers) represent the furthest data points that are not considered outliers. Gold (R9.4.1) and blue (R10.4.1) fill colors represent the different ONT flowcell chemistries. FAST; HAC (High accuracy); SUP (Super accuracy)

Table 3 The total time for base calling using different base callers with their respective models and flowcell chemistries

Base calling model	R9.4.1 (*n* = 20)		R10.4.1 (*n* = 20)		
	Guppy	Dorado	Guppy	Dorado	Rerio
FAST	Time – 2.5 h Bases – 21.7 Gbp	Time – 1.1 h Bases – 18.4 Gbp	Time – 1.2 h Bases – 11.5 Gbp	Time – 0.60 h 9.4 Gbp	-
HAC	Time – 3 h Bases – 22.3 Gbp	Time – 1.2 h Bases – 21.2 Gbp	Time – 1.5 h Bases – 13 Gbp	Time – 0.75 h Bases – 13.3 Gbp	-
SUP	Time – 3.3 h Bases – 22.4 Gbp	Time – 3.3 h Bases – 21.5 Gbp	Time – 2.2 h Bases – 13.3 Gbp	Time – 3.3 h Bases – 13.8 Gbp	Time – 1.1 h Bases – 12.3 Gbp

Assembly quality and species identification

We noted that bacterial genome assemblies generated using R10.4.1 Dorado SUP (4 and 5 kHz) and Rerio SUP base callers generally achieved higher assembly N50 values compared to those from R9.4.1 and Illumina platforms. The N50s are strongly species dependent, with the N50 of ONT and hybrid assemblies frequently reflecting longer assembly N50 lengths, with exceptions in a few isolates sequenced from R9.4.1 flow chemistries. entfae_s and acipit had shorter assembly length N50, whereas assemblies based solely on Illumina data generate shorter assembly N50 lengths (Fig. 3a). There was no difference in the observed %G + C content between the assemblies from the two sequencing platforms. The mean depth for the ONT and Illumina is also provided in Supplementary Table S2. The GTDB-Tk correctly assigned bacterial taxonomy in all de novo assembled genomes, regardless of the sequencing technology or the flowcell chemistry (Supplementary Table S3), except that bursta was not classified for assemblies generated with R10.4.1 Dorado and Guppy in FAST model, similarly in srp hybrid assemblies from R10.4.1 Dorado and Guppy in the FAST model. Additionally, the assembly from R9.4.1, HAC model data was unable to classify pseaer_s as it was incomplete.

Assemblies generated with R10.4.1 yielded more circular contigs than those from R9.4.1 or Illumina. Table 4 reports per-isolate counts of circular (C) and linear (L) contigs from Unicycler and Flye for Illumina and ONT-only assemblies. Corresponding Bandage visualization of the assembly graphs is shown in Supplementary file F1.

Annotation quality

Annotated assemblies, across all species, generated with Illumina data, exhibited relatively consistent mean CDS lengths, suggesting accurate representation of the genome content and minimal frame-shift errors (Fig. 3b). Notably, annotated assemblies from R10.4.1 Dorado SUP (4 and 5 kHz) and Rerio SUP produced mean CDS lengths that were closer to those observed in Illumina assemblies, followed by the R10.4.1 Dorado HAC model (Fig. 3b). The FAST model with Guppy and Dorado performed the worst, with a lower mean CDS length.

Pseudogenes, being annotated as disrupted genes, were used as a proxy for incorrect base calls or indels leading to mis-annotation. Illumina-based assemblies exhibited

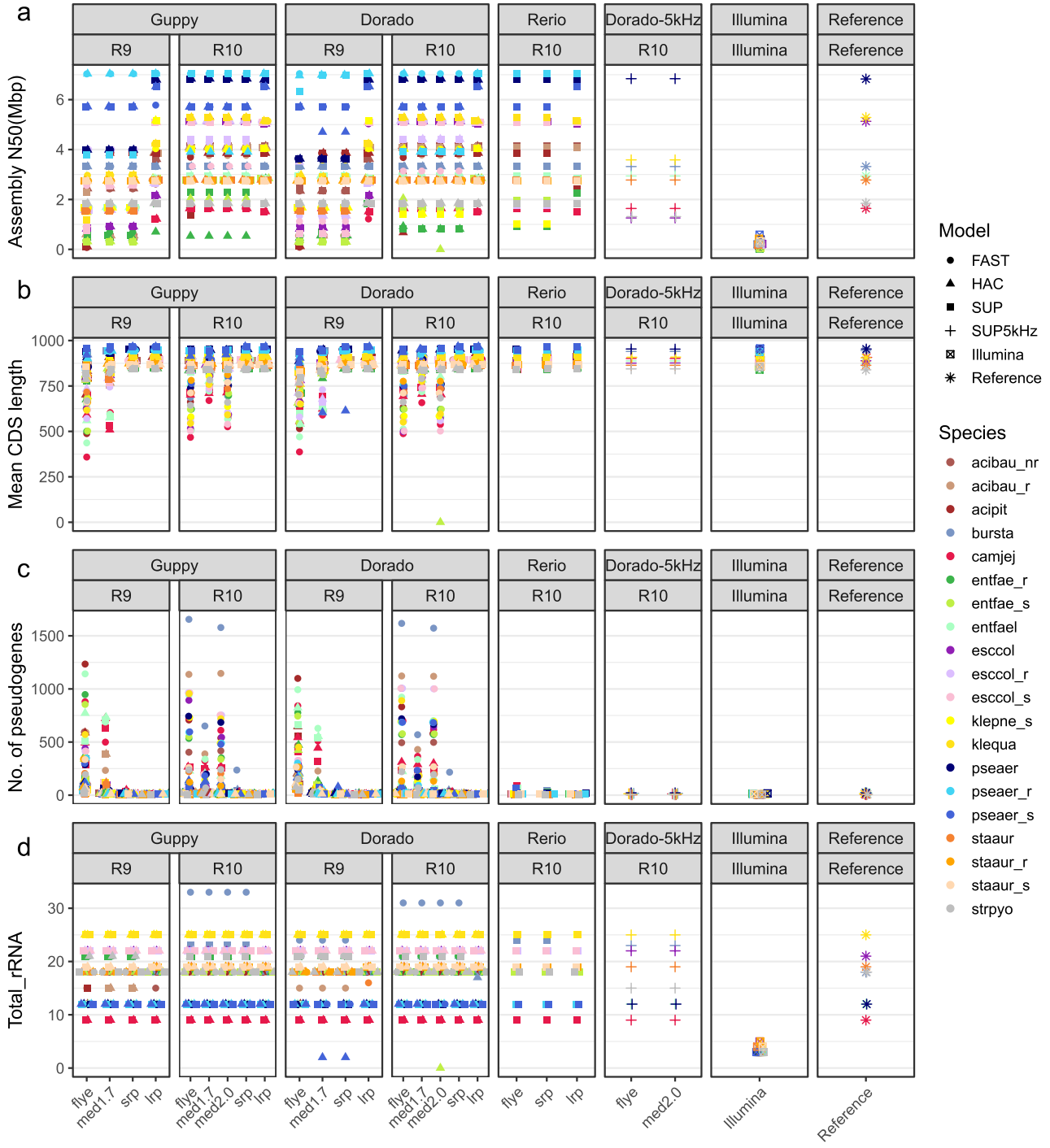

Fig. 3 **a**-**d** Assembly and annotation quality - Each dot represents a bacterial genome assembly, with colors referring to isolates. Each panel represents the base callers, and the nested panel represents the flowcell chemistry. **a** Assembly N50 length (QUAST) in millions of base pairs. **b** Mean CDS length (Bakta and SeqKit) in base pairs. **c** Number of annotated pseudogenes (Bakta). **d** Number of rRNAs annotated (Bakta). The x-axis represents the assemblies generated from the ONT (Flye, Medaka polished (v1.7.2 and v2.0.0)), srp, lrp, Illumina, and ATCC-reference genomes. The y-axis represents the respective assembly and annotation quality metrics. The reference panel represents the ATCC reference genomes

a lower number of pseudogenes compared to ONT-only assemblies. Similarly, assemblies from R10.4.1 Dorado SUP (4 and 5 kHz), Rerio SUP, and R10.4.1 Dorado HAC models also had fewer pseudogenes compared to those from other ONT flowcells and base calling models (Fig. 3c). Assemblies generated using the FAST base-calling models from both Dorado and Guppy, using data from either R9.4.1 or R10.4.1 chemistries, consistently

Table 4. Summary of circular contigs across Illumina, R9.4.1, and R10.4.1. For each species, cells show the number of circularized contigs (C) and linear contigs (L) as C/L for ONT-only assemblies generated with R9.4.1 and R10.4.1 flow cells using Guppy or Dorado basecalling models (FAST, HAC, SUP), R10.4.1 rerio SUP, and Dorado-5 kHz SUP, and for illumina assemblies. Cell shading denotes concordance with the reference: dark green = C and L match the reference column, pale green = C matches only. The lrp assemblies generated with R10 rerio SUP model are considered as the reference assemblies for the ESKAPE isolates, and the respective ATCC refseq genomes are considered as the reference for the ATCC isolates. NA indicates values that are not available

	R9.4.1						R10.4.1									
	Guppy			Dorado			Guppy			Dorado			Rerio	Dorado-5kHz		
	FAST	HAC	SUP	FAST	HAC	SUP	FAST	HAC	SUP	FAST	HAC	SUP	SUP	SUP	Illumina	Reference
Species	C/L	C/L	C/L	C/L	C/L	C/L	C/L	C/L	C/L	C/L	C/L	C/L	C/L	C/L	C/L	C/L
acibau-nr	2/12	2/16	3/16	3/11	3/13	2/12	4/3	4/2	4/2	4/2	4/2	4/2	4/2	NA	1/174	4/0
acibau-r	2/19	2/14	2/12	2/19	1/21	2/10	2/10	3/2	3/2	0/7	2/1	3/0	1/2	NA	0/181	3/0
acipit	2/2	2/1	1/2	5/0	4/1	4/0	4/1	5/1	6/0	4/3	5/2	5/3	5/2	NA	1/97	5/6
entfae-r	3/25	2/33	3/26	3/25	4/25	3/24	3/10	4/19	5/15	5/17	3/14	3/18	2/15	NA	2/320	11/23
entfae-s	4/24	5/25	3/27	5/17	4/32	2/26	4/8	3/12	4/5	5/5	3/9	2/9	2/13	NA	4/239	7/0
esccol-r	2/7	2/13	2/15	1/15	2/10	4/8	3/4	4/0	3/2	4/1	4/0	3/3	3/5	NA	4/153	5/2
esccol-s	1/20	0/16	0/18	1/13	0/15	1/19	1/2	1/0	1/0	1/6	1/1	1/0	1/0	NA	1/125	2/0
klepne-s	1/3	2/3	2/3	1/3	2/3	2/3	1/2	1/2	1/3	2/2	1/2	0/7	0/8	NA	0/94	1/2
pseaer-r	3/1	2/1	0/9	2/1	1/5	0/5	1/1	0/5	1/1	1/1	1/1	0/5	1/1	NA	0/243	1/0
pseaer-s	0/8	0/7	0/7	0/7	0/7	0/9	0/15	0/14	0/13	0/14	0/11	0/13	0/11	NA	0/220	1/0
staaur-r	1/1	1/1	1/1	1/1	1/1	1/1	2/0	2/0	2/0	2/0	2/0	2/0	2/0	NA	1/35	2/0
staaur-s	2/0	2/0	2/0	2/0	2/0	2/0	2/0	2/0	1/0	2/0	1/0	1/0	2/0	NA	0/44	2/0
bursta	5/2	6/0	5/0	6/0	6/0	5/0	7/33	6/43	7/35	8/28	7/35	6/37	7/31	6/13	0/149	3/0
camjej	2/0	2/0	2/0	2/0	2/0	2/0	2/0	2/0	2/0	2/0	2/0	2/0	2/0	2/0	0/42	1/0
entfael	0/6	0/5	0/5	2/1	2/1	2/1	2/1	2/1	2/1	2/1	2/1	2/0	2/1	2/0	0/67	3/0
esccol	2/17	3/15	3/19	3/12	3/15	4/12	4/1	2/1	4/2	3/1	3/2	4/0	2/0	1/14	3/193	1/0
klequa	2/11	1/11	1/16	2/5	1/8	1/9	3/0	3/0	3/0	3/0	3/0	3/0	2/1	2/6	0/118	3/0
pseaer	1/3	1/4	1/3	0/6	0/4	0/4	0/4	2/3	2/3	2/2	2/3	2/5	0/4	1/0	0/185	1/0
staaur	1/7	1/6	1/6	1/4	1/6	1/6	2/0	2/0	2/0	2/0	2/0	2/0	2/0	2/0	0/53	2/0
strpyo	1/0	2/0	1/0	2/0	1/0	2/0	1/0	1/0	2/0	1/0	1/0	2/0	1/0	0/2	0/43	1/0

exhibited higher pseudogene counts than those generated with the SUP models from Rerio and Dorado. In hybrid assemblies, annotation of lrp-based assemblies resulted in fewer pseudogenes than the srp-based assemblies, specifically those generated with R9.4.1 chemistry.

Ribosomal RNA (rRNA) genes are often found in multiple copies within genomes [55]. ONT-only assemblies achieved the same total numbers of annotated rRNA genes as seen in the reference genomes, whereas they often collapse into single copies in Illumina assemblies (Fig. 3d). Hybrid assemblies, integrating both long- and short-read data, showed no differences in terms of assembly length, N50, number of annotated rRNAs, and %G + C content.

ARG detection

All ARGs presented in Table 1 were successfully identified in assemblies of ESKAPE isolates, regardless of the sequencing platforms used. However, assemblies generated using the FAST model from ONT showed incomplete annotation of the *vanA* operon in entfae_r or entirely missed ARGs for certain bacterial species, notably esccol_r. Additionally, when using ONT R9.4.1 chemistry with Dorado and Guppy base calling in High accuracy (HAC) and Super accuracy (SUP) models, the *vanA* operon in entfae_r was detected with only 81% gene length coverage.

To investigate the chromosomal or plasmidic location of the ARGs, MOB-suite analysis of assembled genomes was correlated. The identified ARGs are located on the chromosomes, except for the *vanA* gene, which was classified as plasmid-encoded. Of note, the majority of the ARGs from lrp assemblies were not on contigs assigned specifically to chromosome or plasmid categories. The ARG results are shared in Supplementary Table S4.

Sequencing error calling

The number of sequencing errors was assessed using variant calling/SNP analysis, using ATCC ($n = 8$) isolates, which have high-quality reference genomes available in the RefSeq database. Overall, Illumina reads mapped to the reference sequences generated fewer sequencing errors than ONT reads, with more errors seen in R10.4.1 than R9.4.1 reads (Fig. 4).

We observed that in ONT data (Dorado-SUP, R10.4.1), more sequencing errors were observed in camjej(~ 10,000), followed by pseaer (~ 466), esccol (~ 140), and entfael (~ 73), compared to Illumina. Pilon identified sequencing errors with at least 30x read depth. A species-specific effect is observed between the variant calling for different base-callers and models, which could be attributed to the varying %G + C content, presence of modified bases (miscalled methylated sites), or highly repetitive regions in the genome.

For ATCC staaur, fewer sequencing errors were detected in ONT reads from Dorado-SUP and Rerio-SUP (R10.4.1 and R9.4.1 flowcells), with one error in ONT reads compared to two in Illumina. For klequa, no errors were detected in ONT data from R10.4.1, Rerio (4 kHz), and Dorado SUP 4 kHz ($n = 0$ or $n = 1$). *S. pyogenes* had seven sequencing errors in ONT data (R10.4.1 Dorado 4 kHz and Rerio SUP) and six errors in Illumina data. For bursta, Illumina reads had more errors than ONT data. No difference was observed in sequencing errors between the 4 kHz and 5 kHz models of Dorado SUP in R10.4.1. The complete number of sequencing errors detected and the alignment percentage to the respective

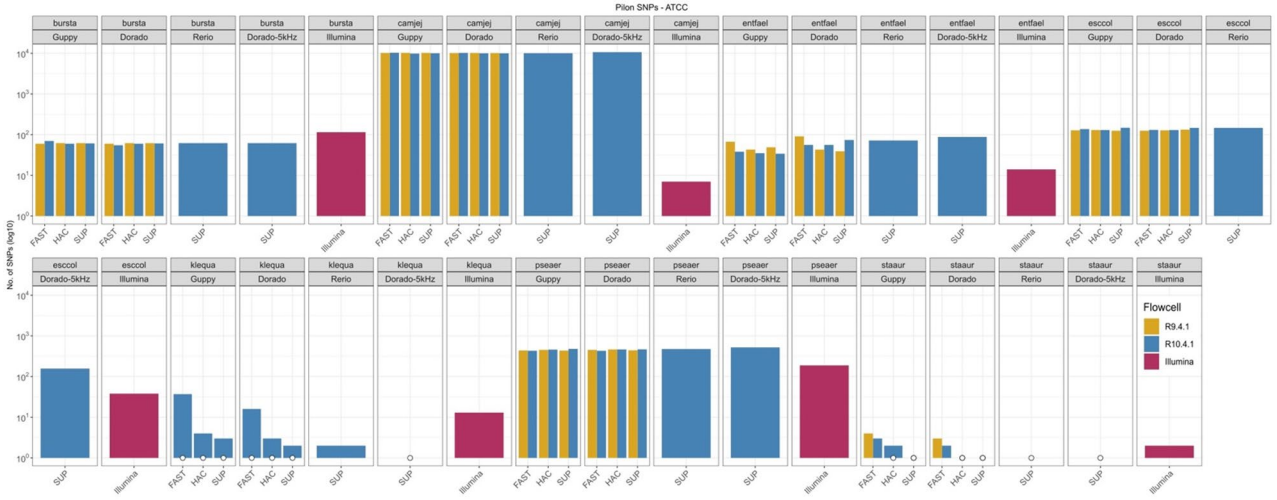

Fig. 4 Sequencing errors detected on the read level - The x-axis represents the different models (FAST, HAC, and SUP) used for the base-callers (Dorado, Guppy, and Rerio) along with Illumina. The y-axis represents the number of errors detected by Pilon on the log 10 scale. The bar color represents the flowcells (Gold - R9.4.1; Blue - R10.4.1; Red - Illumina). bursta - *B. stabilis;* entfael - *E. faecalis;* esccol - *E. coli;* klequa - *K. quasipneumoniae;* pseaer - *P. aeruginosa*; staaur - *S. aureus;* strpyo - *S. pyogenes.* The circle indicates zero SNPs called

reference genomes are provided in Supplementary Table S5 for ATCC and ESKAPE isolates.

Similarly, the number of indels was higher in the ONT data compared to Illumina, with the exception of klequa and staaur. The Snippy comparison is also provided in Supplementary Table S5.

Core genome analysis of assemblies

Multilocus sequence typing (MLST) is commonly used as a low-discrimination typing scheme, and sequence types (STs) were assigned where schemes were available. Across isolates with established MLST schemes (*A. baumannii*, C. *jejuni*, *E. faecium*, *E. faecalis*, *E. coli*, *K. pneumoniae*, *K. quasipneumoniae*, *P. aeruginosa*, *and S. aureus*), we observed that not all the assemblies have an assigned ST. This includes assemblies from R10.4.1 FAST-model data, but also R9.4.1 HAC and SUP. The absence of assigned ST was even seen in some cases among HAC/SUP assemblies with good target percentage (≥ 90%), with data from both R9.4.1 and R10.4.1.

For the entfae_s isolate, discrepancies were noted: assemblies from R9.4.1 basecalled with Guppy FAST and SUP, and srp assemblies generated from R9.4.1 and base called with Guppy and Dorado, were assigned ST1478, whereas the remaining assemblies were classified as ST117. In camjej, the expected ST43 (ATCC reference) was recovered from the Illumina, lrp, and srp assemblies, whereas R10.4.1 assemblies basecalled with Dorado SUP (4 kHz and 5 kHz) were classified as ST50, consistent with the elevated allelic divergence observed for camjej (Supplementary S2).

An important readout in transmission and outbreak analysis is cgMLST. This approach allows for high-resolution comparison of bacterial isolates by analyzing the allelic differences across a defined set of core genes shared within a species. This is often a genome assembly-based comparison. Where schemes were available, we compared our assemblies using cgMLST. Assemblies from Illumina and hybrid approaches (lrp and srp) exhibit lower allelic distances from their respective reference genomes, indicating higher similarity to the reference genomes at the core genome level compared to ONT data (Fig. 5). ONT-only assemblies generated with R10.4.1 Dorado SUP and Rerio SUP for staaur, klequa and klepne_s also met the defined cluster thresholds. However, for other species, although some of the ONT-only assemblies from Dorado, Guppy, and Rerio (HAC and SUP models) achieved good target percentages above 90% (a quality control cut off), their allelic distances exceeded the cgMLST cutoff, meaning that the isolates would not be identified in the same cluster in an outbreak context (Supplementary file S2). In contrast, ONT-only assemblies using the FAST model for Dorado and Guppy displayed lower than 90% good target percentages and higher allelic distances from the reference. The ATCC isolates sequenced with the Dorado SUP-5 kHz model showed lower allelic distances from their respective ATCC references and were on par with Illumina assemblies, except for camjej, which still displayed 348 allele differences from the reference genome.

Discussion

In this study, we evaluated the short-read Illumina and long-read ONT technologies, as well as basecalling software versions, for bacterial whole-genome sequencing using eight ATCC strains and twelve ESKAPE

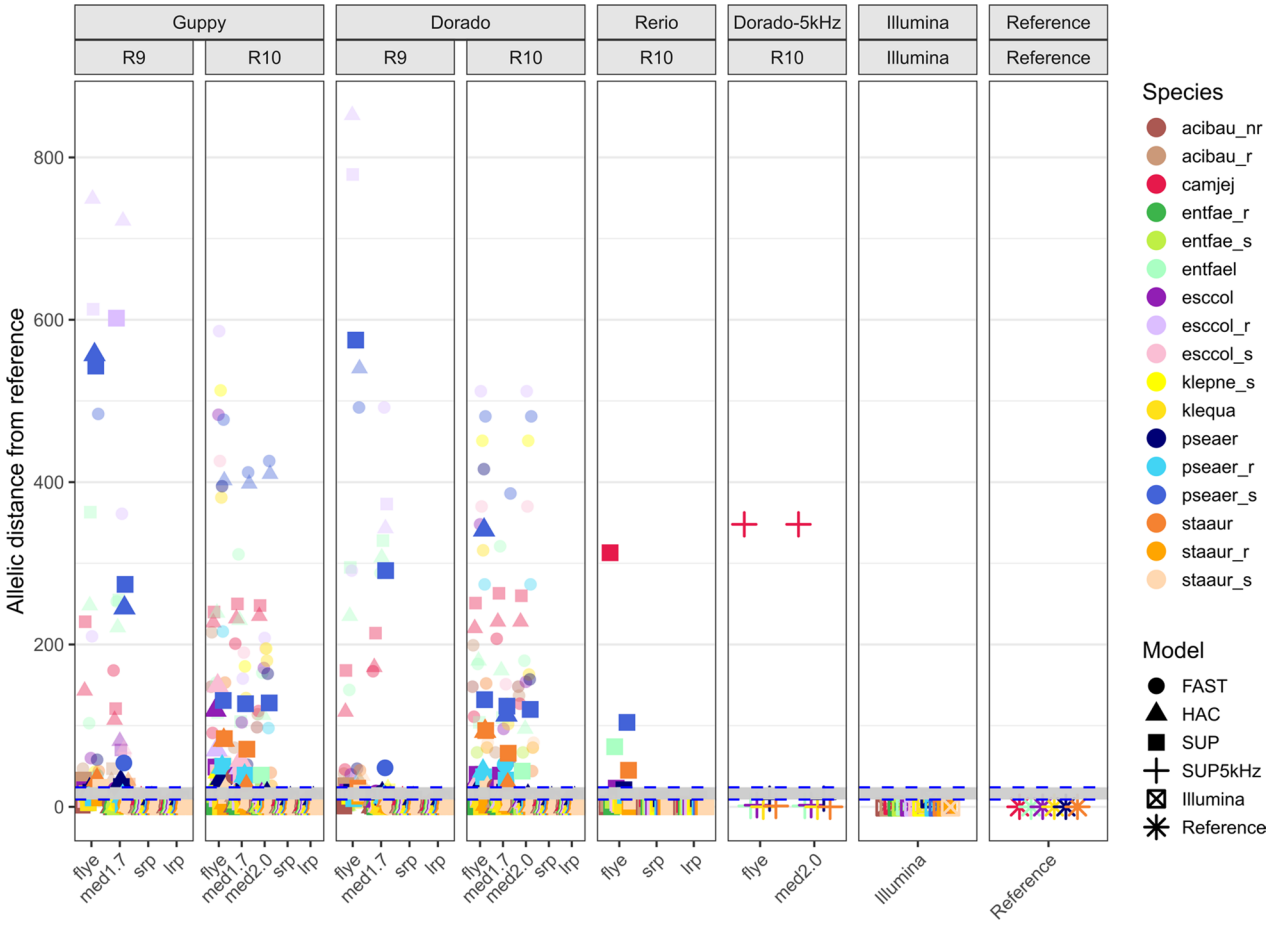

Fig. 5 Sequencing errors defined by cgMLST allelic distance from reference - Each point represents a bacterial genome assembly; point color indicates the species, and point shape indicates the basecalling model. Panels are arranged by basecaller, with nested facets showing the corresponding flowcell chemistry. Along the x-axis, assemblies are grouped by assembly strategy: ONT-only assemblies (Flye, Medaka-polished v1.7.2 and v2.0.0), hybrid assemblies (srp and lrp), Illumina-only assemblies, and the ATCC reference genomes. The y-axis shows the cgMLST allelic distance to the reference genome as calculated with Ridom SeqSphere+. Assemblies with ≥ 90% good target loci are plotted as larger, opaque points, whereas assemblies with < 90% good targets are shown as semi-transparent points. The two blue dashed horizontal lines indicate the minimum and maximum cgMLST allelic distance cut-offs (9 (minimum) to 24 (maximum)), and the grey-shaded band between them marks the accepted cgMLST range in which assemblies are considered to cluster with the reference

clinical isolates. We performed relevant analyses for clinical microbiology, including genome assembly quality, antimicrobial resistance identification, and taxonomic identification.

Illumina data consistently achieved much higher median quality scores compared to both ONT flow cell versions. The ONT results were also below the expected Q20, potentially influenced by library preparation methods [56]. The higher total base yields produced by the R9.4.1 chemistry was likely due to shorter read lengths being preferentially sequenced, compared to the longer reads of R10.4.1. (Fig. 2b). Longer reads typically require more time to translocate through nanopores than shorter reads. Consequently, ONT platforms preferentially sequence shorter reads, resulting in an overall higher data yield for the R9.4.1 chemistry [57, 58].

The observed differences in computational time (Table 3) across FAST, HAC, and SUP models mirror the guidance from Oxford Nanopore [59], that increased model complexity trades speed for accuracy. Consistently, Dorado outperformed Guppy in time-to-result across chemistries and models, indicating more efficient GPU utilization. However, SUP remained the most computationally demanding for both callers. Given that our runs used five GPUs, this configuration may be impractical in resource-limited settings, where the added hardware requirements and infrastructure costs are prohibitive. In such contexts, prioritizing the FAST or HAC model might offer a more feasible balance between turnaround time and hardware burden.

We observed that genome assemblies generated using ONT R10.4.1 chemistry and the SUP basecalling models from Rerio and Dorado (4 and 5 kHz)

produced notably higher N50 values compared to R9.4.1 and Illumina assemblies. This increase in N50 indicates greater genome contiguity, which is particularly beneficial for resolving complex genomic structures, repetitive elements, and large structural variants. Similar findings have been reported by previous studies using long-read technologies, highlighting the advantage of ONT sequencing for improved assembly continuity and completeness [60–62]. These findings highlight the advantage of ONT-only assemblies in generating more contiguous genomes relative to Illumina-only assemblies. This is also reflected in a higher number of circular assemblies generated with R10.4.1 compared to R9.4.1 and Illumina (Table 4).

Pseudogene counts and CDS mean length varied by sequencing technology and base-calling model. Illumina assemblies consistently had the fewest pseudogenes, reflecting fewer sequencing errors and more accurate coding sequence annotations. Conversely, assemblies from FAST base-calling showed increased pseudogene counts due to lower raw read accuracy (63), causing frameshift mutations, indels, and misplaced stop codons. Thus, selecting high-accuracy base-calling models (SUP or HAC) is crucial for precise genome annotation and comparative genomic studies, such as pan-genome analysis and Genome-wide association studies, which rely on core gene annotations.

ONT-alone assemblies and hybrid assemblies were able to resolve individual rRNA operons compared to Illumina-alone assemblies, in which they collapsed and assembled. This highlights the importance of long-read sequences provided by ONT in resolving repetitive genomic regions compared to short-read Illumina. These results suggest that ONT-alone assemblies R10.4.1 chemistry, specifically base-called on Dorado (5 kHz) in the SUP model, can yield accurate annotations.

Taxonomic identification of all assembled bacterial genomes was accurate, regardless of the sequencing technology used. GTDB-Tk assigned all bacterial genomes to the correct species. This robustness could be attributed to the reliance on a conserved set of single-copy marker genes detected via amino-acid Hidden Markov Models (HMMs), which tolerate indels and %G + C-extreme underrepresentation, and from Average Nucleotide Identity (ANI) comparisons focused on those conserved regions. By enforcing minimal marker-gene completeness thresholds, GTDB-Tk ensures sufficient phylogenetic signals for accurate taxonomic assignment [63–65].

The consistent detection of ARGs across the sequencing platforms highlights their reliability and effectiveness in identifying resistance determinants among ESKAPE pathogens. However, the limitations observed with ONT's FAST model assemblies emphasize the importance of selecting appropriate sequencing and base calling strategies to accurately identify ARG alleles. Sequencing errors, particularly in ONT data, can lead to miscalling ARG alleles, which, particularly in the cases of some beta-lactamase genes, could have clinical implications.

SNP calling is a critical process in bacterial WGS and comparison that relies on high-accuracy reads at a good read depth and is confounded by sequencing errors. Single-nucleotide changes can provide key insights during epidemiological investigations and outbreak analyses. We used SNP calling as a proxy for sequencing error analysis. A pronounced species-specific effect was evident in bursta, entfael, esccol, camjej, and pseaer, which displayed greater divergence between platforms, with ONT detecting up to four orders of magnitude more sequencing errors depending on the base-caller model. By contrast, staaur, strpyo, and klequa remained almost invariant, reflecting their relatively homogeneous genomes and lower repeat content. These results underscore that genomic features such as the presence of methylated motifs, %G + C skew, and repeat density might lead to mapping artifacts [66, 67].

An important test of the generated assemblies was to evaluate the allelic distances from reference genomes using cgMLST. Our results demonstrate that Illumina and hybrid assemblies consistently yielded lower allelic distances from their reference genomes and clustered closely, within the accepted cgMLST thresholds. Interestingly, ONT-only assemblies produced using base caller models R10.4.1 Dorado SUP 4 kHz and Rerio SUP 4 kHz also showed comparable performance to Illumina for specific organisms, notably staaur, klequa, and klepne_s. However, our analysis revealed limitations for other species (camjej, entfael, esccol, and pseaer) when employing ONT-only assemblies with Dorado, Guppy, and Rerio using High accuracy (HAC) and Super accuracy (SUP) models. Despite some of these assemblies having good target percentages of alleles identified above 90%, their allelic distances frequently surpassed the cgMLST cutoffs. Especially poor was the performance with camjej. This observation suggests species-specific variability in the assembly quality, potentially linked to genome complexity and issues in sequencing native modified bases. Furthermore, ONT-only assemblies employing FAST models for Dorado and Guppy performed notably worse, demonstrating both suboptimal target percentages (< 90%) and increased allelic distances from references. This underscores the limited suitability of FAST models for applications requiring high precision, such as cgMLST-based epidemiological investigations. As a consequence, ONT sequencing should only be used with great caution in outbreak investigations, and with

model- and species-specific validation. The missed ST assignments in some assemblies typed with Ridom SeqSphere + likely reflect misassemblies at one or more MLST target loci, yielding sequences that fail exact allele-matching criteria.

Using the Dorado SUP-5 kHz model provided within the MinKNOW software (v24.06.15), the sequenced ATCC isolates displayed lower allelic distances, comparable to Illumina assemblies, across most species, except for camjej. Based on these results, ONT-only assemblies generated using the Dorado (5 kHz) model can be considered on par with Illumina in terms of cgMLST cluster detection. However, care should be taken for species like camjej, which consistently showed inaccurate MLST assignments and higher allelic distances using ONT data. This may be due to its low %G + C content and the presence of complex methylation motifs that are known to reduce ONT sequencing accuracy, as discussed in the benchmarking efforts [68].

Interestingly, several studies have also reported strain-specific erroneous clustering with the Dorado SUP 5 kHz model in several species, like *Listeria monocytogenes* and *Burkholderia pseudomallei* [69, 70]. A multicenter study led by Dabernig-Heinz et al. [71] also demonstrated similar strain-specific erroneous clustering in *E. faecium*, *K. pneumoniae*, *S. aureus*, and *L. monocytogenes* species. The authors also discussed PCR preamplification of input DNA for ONT sequencing as an error mitigation strategy to overcome the faulty cgMLST clustering arising from methylated sites in the bacterial genomes. Similarly, PCR-based library preparation for ONT sequencing and masking of problematic genomic regions in *Corynebacterium diphtheriae* strains and Vancomycin-resistant Enterococci strains has been proven to match Illumina results in cgMLST clustering [72]. Another ONT error mitigation is to mask the modified bases in the bacterial genomes through the polishing of the assemblies using the cgMLST polisher tool provided within the Ridom SeqSphere + software suite. The strategy is to identify and mask ambiguous base calls at methylation-related error hotspots using allele- aware masking derived from raw- signal patterns to refine allele assignments and eliminate systematic miscalls. Studies in diverse multi-drug-resistant nosocomial important Gram-negative and Gram-positive bacterial isolates and %G ± C-rich *B. pseudomallei* isolates [70, 73] highlight that targeted masking or polishing of ONT errors via the ONT-cgMLST-Polisher is critical to reach Illumina-comparable resolution in cgMLST clustering.

A limitation of the study is the limited number of bacterial isolates that were used to validate the Dorado SUP 5 kHz performance across species, and the study did not compare the latest cgMLST ONT polisher made available within Ridom Seqsphere+. Nonetheless, the study underscores the importance of leveraging the synergy between long- and short-read sequencing platforms and highlights the development of ONT across different stages from R9.4.1, 4 kHz R10.4.1 Dorado, followed by Rerio to 5 kHz Dorado SUP, and emphasizes the need to further improve and validate ONT's bacterial base-calling models to account for modified bases across different bacterial species.

Conclusion

We systematically and comprehensively evaluated bacterial WGS across two platforms, Illumina and ONT, comparing multiple flowcell chemistries, base callers, and hybrid assembly approaches. ONT data, particularly those generated with R10.4.1 flowcells and the SUP models in Dorado and Rerio, yielded highly contiguous genome assemblies and resolved repetitive regions more effectively. In contrast, Illumina consistently provided more accurate base calls, albeit with more fragmented assemblies. Hybrid assemblies leverage the strengths of both technologies, producing comprehensive assemblies alongside high-level accuracy. Nevertheless, challenges persist in addressing base modifications in certain bacterial strains, and although ONT chemistries and base-calling algorithms continue to improve, there are still species-specific sequencing errors that affect especially for outbreak investigations. The complementary use of long and short reads remains the most reliable strategy for achieving both complete and highly accurate bacterial genome assemblies.

Abbreviations

ANI	Average Nucleotide Identity
ARG	Antimicrobial resistance gene
ATCC	American Type Culture Collection
CDS	Coding DNA sequence
cgMLST	Core genome Multilocus Sequence Typing
ENA	European Nucleotide Archive
ESBL	Extended-spectrum beta-lactamase
ESKAPE	*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa,* and *Enterobacter spp*
FAST	FAST (ONT base-calling model)
GPU	Graphics Processing Unit
GTDB-Tk	Genome Taxonomy Database Toolkit
HAC	High accuracy (ONT base calling model)
HMMs	Hidden Markov Models
HPC	High Performance Computing Cluster
IQR	Interquartile Range
ONT	Oxford Nanopore Technologies
MLST	Multilocus Sequence Typing
PE	Paired end
PCR	Polymerase Chain Reaction
QC	Quality control
RBK	Rapid Barcoding Kit (ONT)
rRNA	Ribosomal RNA
SNP	Single Nucleotide Polymorphism
SUP	Super accuracy (ONT base calling model)
TAT	Turnaround time
VCF	Variant Call Format
WGS	Whole Genome Sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-025-02305-2.

Supplementary Material 1 - Read quality metrics

Supplementary Material 2 - Assembly quality metrics

Supplementary Material 3 - GTDB Taxa

Supplementary Material 4 - ARG

Supplementary file F1 - Bandage visualizations

Supplementary Material 5 - Sequencing errors

Acknowledgements

We would like to thank Dr. Vladimira Hinic and PD Dr. Stefano Mancini from the Institute of Medical Microbiology, University of Zurich for kindly providing the clinical ESKAPE isolates. We would also like to thank the High-Performance Cluster support provided by the Science IT team from the University of Zurich for performing the bioinformatics analysis. We thank Dr. Marco Meola for the helpful discussion and advice.

Authors' contributions

Conceptualization and study design plan: SP, HSS, TR, and AE. Sequencing, bioinformatics analysis, and visualization: SP and HSS. Writing of the manuscript: SP. Providing critical feedback on the manuscript: HSS, TR, and AE.

Funding

We thank the Swiss National Science Foundation (reference 310030_192515) for supporting the research of AE.

Data availability

The raw fastq files from the ONT and the Illumina have been deposited in the European Nucleotide Archive (ENA) with accession ID: PRJEB90249 and will be made public upon acceptance of the manuscript.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Institute of Medical Microbiology, University of Zurich, Gloriastrasse 28, Zurich 8006, Switzerland

Received: 3 July 2025 / Accepted: 30 December 2025

Published online: 15 January 2026

References

1. Kozińska A, Seweryn P, Sitkiewicz I. A crash course in sequencing for a microbiologist. J Appl Genet. 2019;60:103–11.
2. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13:601–12.
3. Simar SR, Hanson BM, Arias CA. Techniques in bacterial strain typing: past, present, and future. Curr Opin Infect Dis. 2021;34:339–45.
4. Price V, Ngwira LG, Lewis JM, Baker KS, Peacock SJ, Jauneikaite E, et al. A systematic review of economic evaluations of whole-genome sequencing for the surveillance of bacterial pathogens. Microb Genom. 2023;9. https://doi.org/10.1099/mgen.0.000947.
5. Schadron T, van den Beld M, Mughini-Gras L, Franz E. Use of whole genome sequencing for surveillance and control of foodborne diseases: status quo and quo vadis. Front Microbiol. 2024;15:1460335.
6. Mutschler E, Roloff T, Neves A, Vangstein Aamot H, Rodriguez-Sanchez B, Ramirez M, et al. Towards unified reporting of genome sequencing results in clinical microbiology. PeerJ. 2024;12:e17673.
7. Lau KA, da Gonçalves Silva A, Theis T, Gray J, Ballard SA, Rawlinson WD. Proficiency testing for bacterial whole genome sequencing in assuring the quality of microbiology diagnostics in clinical and public health laboratories. Pathology. 2021;53:902–11.
8. Rossen JWA, Friedrich AW, Moran-Gilad J, ESCMID Study Group for Genomic and Molecular Diagnostics (ESGMD). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. Clin Microbiol Infect. 2018;24:355–60.
9. Bianconi I, Aschbacher R, Pagani E. Current uses and future perspectives of genomic technologies in clinical microbiology. Antibiotics (Basel). 2023;12:1580.
10. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, et al. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. Clin Microbiol Rev. 2017;30:1015–63.
11. Kigen C, Muraya A, Kyanya C, Kingwara L, Mmboyi O, Hamm T, et al. Enhancing capacity for National genomics surveillance of antimicrobial resistance in public health laboratories in Kenya. Microb Genom. 2023;9. https://doi.org/10.1099/mgen.0.001098.
12. Bruzek S, Vestal G, Lasher A, Lima A, Silbert S. Bacterial whole genome sequencing on the illumina iSeq 100 for clinical and public health laboratories. J Mol Diagn. 2020;22:1419–29.
13. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom. 2019;5. https://doi.org/10.1099/mgen.0.000294.
14. Bejaoui S, Nielsen SH, Rasmussen A, Coia JE, Andersen DT, Pedersen TB, et al. Comparison of illumina and Oxford nanopore sequencing data quality for clostridioides difficile genome analysis and their application for epidemiological surveillance. BMC Genomics. 2025;26:92.
15. Bachmann JA, Tedder A, Laenen B, Steige KA, Slotte T. Targeted long-read sequencing of a locus under long-term balancing selection in capsella. G3 (Bethesda). 2018;8:1327–33.
16. Luo Y, Jang JH, Balkey M, Hoffmann M. 217 closed Salmonella reference genomes using PacBio sequencing. BMC Genom Data. 2025;26:15.
17. Sierra R, Roch M, Moraz M, Prados J, Vuilleumier N, Emonet S, et al. Contributions of long-read sequencing for the detection of antimicrobial resistance. Pathogens. 2024;13:730.
18. George S, Pankhurst L, Hubbard A, Votintseva A, Stoesser N, Sheppard AE, et al. Resolving plasmid structures in Enterobacteriaceae using the minion nanopore sequencer: assessment of minion and minion/Illumina hybrid data assembly approaches. Microb Genom. 2017;3:e000118.
19. Shelenkov A, Petrova L, Mironova A, Zamyatin M, Akimkin V, Mikhaylova Y. Long-read whole genome sequencing elucidates the mechanisms of Amikacin resistance in multidrug-resistant *Klebsiella pneumoniae* isolates obtained from COVID-19 patients. Antibiot (Basel). 2022;11:1364.
20. Avershina E, Frye SA, Ali J, Taxt AM, Ahmad R. Ultrafast and cost-effective pathogen identification and resistance gene detection in a clinical setting using nanopore Flongle sequencing. Front Microbiol. 2022;13:822402.
21. Wasswa FB, Kassaza K, Nielsen K, Bazira J. MinION whole-genome sequencing in resource-limited settings: challenges and opportunities. Curr Clin Microbiol Rep. 2022;9:52–9.
22. Pronyk PM, de Alwis R, Rockett R, Basile K, Boucher YF, Pang V, et al. Advancing pathogen genomics in resource-limited settings. Cell Genom. 2023;3:100443.
23. Bayliss SC, Hunt VL, Yokoyama M, Thorpe HA, Feil EJ. The use of Oxford nanopore native barcoding for complete genome assembly. Gigascience. 2017;6:1–6.
24. Wick RR, Judd LM, Holt KE. Assembling the perfect bacterial genome using Oxford nanopore and illumina sequencing. PLoS Comput Biol. 2023;19:e1010905.
25. Lerminiaux N, Fakharuddin K, Mulvey MR, Mataseje L. Do we still need illumina sequencing data? Evaluating Oxford nanopore technologies R10.4.1 flow cells and the rapid v14 library prep kit for gram negative bacteria whole genome assemblies. Can J Microbiol. 2024;70:178–89.
26. Linde J, Brangsch H, Hölzer M, Thomas C, Elschner MC, Melzer F, et al. Comparison of Illumina and Oxford Nanopore Technology for genome analysis

of *Francisella tularensis*, *Bacillus anthracis*, and *Brucella suis*. BMC Genomics. 2023;24:258.
27. Bogaerts B, Van den Bossche A, Verhaegen B, Delbrassinne L, Mattheus W, Nouws S, et al. Closing the gap: Oxford nanopore technologies R10 sequencing allows comparable results to illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. J Clin Microbiol. 2024;62:e0157623.
28. Foster-Nyarko E, Cottingham H, Wick RR, Judd LM, Lam MMC, Wyres KL, et al. Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, Klebsiella pneumoniae. Microb Genom. 2023;9. https://doi.org/10.1099/mgen.0.000936.
29. Sanderson ND, Kapel N, Rodger G, Webster H, Lipworth S, Street TL, et al. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford nanopore flowcells and chemistries in bacterial genome reconstruction. Microb Genom. 2023;9. https://doi.org/10.1099/mgen.0.000910.
30. Sanderson ND, Hopkins KMV, Colpus M, Parker M, Lipworth S, Crook D, et al. Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. Microb Genom. 2024. https://doi.org/10.1099/mgen.0.001246.
31. Seth-Smith HMB, Bonfiglio F, Cuénod A, Reist J, Egli A, Wüthrich D. Evaluation of rapid library preparation protocols for whole genome sequencing based outbreak investigation. Front Public Health. 2019;7:241.
32. pod5-file-format. Pod5: a high performance file format for nanopore reads. https://github.com/nanoporetech/pod5-file-format.
33. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 6 June 2025.
34. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30:2114–20.
35. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13:e1005595.
36. Wick R. Filtlong: quality filtering tool for long reads. https://github.com/rrwick/Filtlong.
37. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics. 2023. https://doi.org/10.1093/bioinformatics/btad311.
38. Li H. seqtk: Toolkit for processing sequences in FASTA/Q formats. https://github.com/lh3/seqtk.
39. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6.
40. medaka. Sequence correction provided by ONT Research. https://github.com/nanoporetech/medaka.
41. Wick RR, Holt KE. Polypolish: short-read polishing of long-read bacterial genome assemblies. PLoS Comput Biol. 2022;18:e1009802.
42. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genom. 2021;7. https://doi.org/10.1099/mgen.0.000685.
43. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11:e0163962.
44. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
45. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31:3350–2.
46. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022;38:5315–6.
47. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11:12728.
48. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom. 2018;4. https://doi.org/10.1099/mgen.0.000206.
49. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
51. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963.
52. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
53. Seemann T. Snippy::scissors: rapid haploid variant calling and core genome alignment. https://github.com/tseemann/snippy.
54. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, et al. Updating benchtop sequencing performance comparison. Nat Biotechnol. 2013;31:294–6.
55. Espejo RT, Plaza N. Multiple ribosomal RNA operons in bacteria; their concerted evolution and potential consequences on the rate of evolution of their 16S rRNA. Front Microbiol. 2018;9:1232.
56. Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Comput Struct Biotechnol J. 2023;21:2352–64.
57. Purushothaman S, Meola M, Roloff T, Rooney AM, Egli A. Evaluation of DNA extraction kits for long-read shotgun metagenomics using Oxford Nanopore sequencing for rapid taxonomic and antimicrobial resistance detection. Sci Rep. 2024;14:29531.
58. De La Cerda GY, Landis JB, Eifler E, Hernandez AI, Li F-W, Zhang J, et al. Balancing read length and sequencing depth: optimizing nanopore long-read sequencing for monocots with an emphasis on the Liliales. Appl Plant Sci. 2023;11:e11524.
59. Data analysis. Oxford Nanopore Technologies. 2017. https://nanoporetech.com/document/data-analysis. Accessed 24 Oct 2025.
60. Soto-Serrano A, Li W, Panah FM, Hui Y, Atienza P, Fomenkov A, et al. Matching excellence: Oxford nanopore technologies' rise to parity with Pacific biosciences in genome reconstruction of non-model bacterium with high G+C content. Microb Genom. 2024;10. https://doi.org/10.1099/mgen.0.001316.
61. Wagner GE, Dabernig-Heinz J, Lipp M, Cabal A, Simantzik J, Kohl M, et al. Real-time nanopore Q20 + sequencing enables extremely fast and accurate core genome MLST typing and democratizes access to high-resolution bacterial pathogen surveillance. J Clin Microbiol. 2023;61:e0163122.
62. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12:733–5.
63. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics. 2019;36:1925–7.
64. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 2022;50:D785–94.
65. Mussig AJ, Chaumeil P-A, Chuvochina M, Rinke C, Parks DH, Hugenholtz P. Putative genome contamination has minimal impact on the GTDB taxonomy. Microb Genom. 2024;10. https://doi.org/10.1099/mgen.0.001256.
66. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. Gigascience. 2020. https://doi.org/10.1093/gigascience/giaa008.
67. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front Genet. 2015;6:235.
68. Wick R. ONT-only accuracy: 5 kHz and Dorado. Ryan Wick's bioinformatics blog. 2023. https://rrwick.github.io/2023/10/24/ont-only-accuracy-update.html. Accessed 6 June 2025.
69. Biggel M, Cernela N, Horlbog JA, Stephan R. Oxford nanopore's 2024 sequencing technology for Listeria monocytogenes outbreak detection and source attribution: progress and clone-specific challenges. J Clin Microbiol. 2024;62:e0108324.
70. Weigl S, Dabernig-Heinz J, Granitz F, Lipp M, Ostermann L, Harmsen D, et al. Improving nanopore sequencing-based core genome MLST for global infection control: a strategy for GC-rich pathogens like Burkholderia pseudomallei. J Clin Microbiol. 2025;63:e0156924.
71. Dabernig-Heinz J, Lohde M, Hölzer M, Cabal A, Conzemius R, Brandt C, et al. A multicenter study on accuracy and reproducibility of nanopore sequencing-based genotyping of bacterial pathogens. J Clin Microbiol. 2024;62:e0062824.
72. Neuenschwander S, Borcard L, Gempeler S, Miani MT, Casanova C, Ramette A. Evaluation of Oxford Nanopore Technologies workflows for genomic epidemiology of outbreak-associated bacterial isolates in the clinical setting. medRxiv. 2025. https://doi.org/10.1101/2025.04.25.25326448.
73. Prior K, Becker K, Brandt C, Cabal Rosel A, Dabernig-Heinz J, Kohler C, et al. Accurate and reproducible whole-genome genotyping for bacterial genomic surveillance with nanopore sequencing data. J Clin Microbiol. 2025 63:e0036925.

Publisher's note