

Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction

Nicholas D. Sanderson^{1,2,*}, †, Natalia Kapel^{1,2} †, Gillian Rodger^{1,2}, Hermione Webster^{1,2}, Samuel Lipworth², Teresa L. Street^{1,2}, Timothy Peto^{1,2}, Derrick Crook^{1,2} and Nicole Stoesser^{1,2,3,*}

Abstract

Complete, accurate, cost-effective, and high-throughput reconstruction of bacterial genomes for large-scale genomic epidemiological studies is currently only possible with hybrid assembly, combining long- (typically using nanopore sequencing) and short-read (Illumina) datasets. Being able to use nanopore-only data would be a significant advance. Oxford Nanopore Technologies (ONT) have recently released a new flowcell (R10.4) and chemistry (Kit12), which reportedly generate per-read accuracies rivaling those of Illumina data. To evaluate this, we sequenced DNA extracts from four commonly studied bacterial pathogens, namely *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa* and *Staphylococcus aureus*, using Illumina and ONT's R9.4.1/Kit10, R10.3/Kit12, R10.4/Kit12 flowcells/chemistries. We compared raw read accuracy and assembly accuracy for each modality, considering the impact of different nanopore basecalling models, commonly used assemblers, sequencing depth, and the use of duplex versus simplex reads. 'Super accuracy' (sup) basecalled R10.4 reads - in particular duplex reads - have high per-read accuracies and could be used to robustly reconstruct bacterial genomes without the use of Illumina data. However, the per-run yield of duplex reads generated in our hands with standard sequencing protocols was low (typically <10%), with substantial implications for cost and throughput if relying on nanopore data only to enable bacterial genome reconstruction. In addition, recovery of small plasmids with the best-performing long-read assembler (Flye) was inconsistent. R10.4/Kit12 combined with sup basecalling holds promise as a singular sequencing technology in the reconstruction of commonly studied bacterial genomes, but hybrid assembly (Illumina+R9.4.1 hac) currently remains the highest throughput, most robust, and cost-effective approach to fully reconstruct these bacterial genomes.

DATA SUMMARY

Nanopore fast5 and fastq data are available at the NCBI/ENA under project accession: PRJEB51164 (available directly at the following link: <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB51164>).

Received 29 June 2022; Accepted 12 October 2022; Published 10 January 2023

Author affiliations: ¹NIHR OxfordBiomedical Research Centre, University of Oxford, Oxford, UK; ²Nuffield Department of Medicine, University of Oxford, Oxford, UK; ³NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, UK.

***Correspondence:** Nicholas D. Sanderson, nicholas.sanderson@ndm.ox.ac.uk; Nicole Stoesser, nicole.stoesser@ndm.ox.ac.uk

Keywords: Genome sequencing; hybrid assembly; long-read assembly.

Abbreviations: bp, base pair; CDS, coding sequence; DNA, deoxyribonucleic acid; Gb, Giga base pairs; GC-content, guanine-cytosine content; hac, High accuracy; INDEL, Insertion or deletion; IQR, Interquartile range; kb, kilo base pairs; Mb, Mega base pairs; MRSA252, Methicillin-resistant *Staphylococcus aureus* strain 252; ONT, Oxford Nanopore Technologies; Q20+, Quality higher than 20 PHRED score; QC, Quality control; SNP, Single nucleotide polymorphism; sup, Super accuracy.

Nanopore fast5 and fastq data are available at the NCBI/ENA under project accession: PRJEB51164.

Assemblies have been made available at: https://figshare.com/projects/q20_comparison_genome_assemblies/146097

Code and analysis outputs are available at: https://gitlab.com/ModernisingMedicalMicrobiology/assembly_comparison_analysis/-/tree/main (tagged version v0.5.5).

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary figures and seven supplementary tables are available with the online version of this article.

000910 © 2023 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

Impact Statement

Our understanding of microbes has been greatly enhanced by the capacity to evaluate their genetic make-up using a technology known as whole genome sequencing. Sequencers represent microbial genomes as stretches of shorter sequence known as 'reads', which are then assembled using computational algorithms. Different types of sequencing approach have advantages and disadvantages with respect to the accuracy and length of the reads they generate; this in turn affects how reliably genomes can be assembled.

Currently, to completely reconstruct bacterial genomes in a high-throughput and cost-effective manner, researchers tend to use two different types of sequencing data, namely Illumina (short-read) and nanopore (long-read) data. Illumina data are highly accurate; nanopore data are much longer, and this combination facilitates accurate and complete bacterial genomes in a so-called 'hybrid assembly'. However, new developments in nanopore sequencing have reportedly greatly improved the accuracy of nanopore data, hinting at the possibility of requiring only a single sequencing approach for bacterial genomics.

Here we evaluate these improvements in nanopore sequencing in the reconstruction of four bacterial reference strains, where the true sequence is already known. We show that although these improvements are extremely promising, for high-throughput, low-cost complete reconstruction of bacterial genomes hybrid assembly currently remains the optimal approach.

INTRODUCTION

Bacterial whole genome sequencing has become a prominent tool in the biological sciences, with wide-ranging applications from epidemiology to diagnostics [1]. Important considerations include sequencing throughput, read length (which facilitates complete reconstruction of bacterial chromosomes and plasmids), read accuracy, accessibility and cost. Historically, short-read Illumina sequencing has been the leading high-throughput, high-accuracy technology, but is limited in its capacity to completely reconstruct genomes, particularly in the presence of repetitive sequences. Nanopore sequencing (Oxford Nanopore Technologies [ONT]) has become one of the most widely adopted long-read sequencing approaches, enabled by affordable, small-footprint sequencing platforms, but has been limited to some extent by its accuracy. Combining short- and long-read sets from both technologies in the form of hybrid assembly has facilitated cost-effective, highly accurate and scalable genome reconstruction for large bacterial isolate collections [2, 3], such as by multiplexing 96 *E. coli* isolates on a single nanopore flowcell [3]. For nanopore sequencing, developments in multiplexing, rapid library preparation and flow cell reuse after washing have streamlined this process [4].

ONT have undertaken iterative development of their sequencing flowcells and chemistries, releasing the R10.3 (FLO-MIN111) flowcells for consumers in January 2020 and the Kit12 (Q20+) chemistry and R10.4 flowcell (FLO-MIN112) in their store in late 2021. The proposed advantages of the R10.4/Kit12 system include: (i) a new motor to facilitate more controlled passage of the nucleic acid template through the sequencing pore thereby avoiding template slippage; (ii) 'duplex' read sequencing - where the forward and reverse strand of a single nucleic acid molecule are sequenced in succession to improve accuracy; and (iii) an optimized pore with a longer pore head to better resolve homopolymers.

These new developments however come with some potential disadvantages. Sequencing yields for the R10.3 flowcells were lower than those using R9.4.1 flowcells (thought to be due to the slower passage of template through pores) [5]. The use of R10 flowcells also currently requires a ligation-based library preparation, which results in longer sequencing turnaround times when compared with rapid transposase-based library preparation kits which can be used with R9.4.1 flowcells. Ligation-based preparations may also miss the capture and sequencing of small plasmids [6]. The reported improvements in per-read accuracy with R10/Kit12 are also potentially dependent on the use of super accuracy (sup) basecalling models; however, on the same computing infrastructure sup basecalling takes 2–8× longer than the previous typical approach using high accuracy (hac) basecalling models, which may preclude 'on-machine' basecalling in real-time [7].

Sequencing accuracy can be characterized using several different metrics, including: (i) raw read accuracy (the accuracy achieved when reading a single nucleic acid fragment once) and (ii) assembly accuracy (the capacity to accurately reconstruct complete genomes in terms of structure, sequence identity and coding sequence content). We therefore set out to compare data and assemblies generated by R9.4.1/Kit10 and R10/Kit12 nanopore flowcells/chemistries, comparing these with Illumina-only sequence data and hybrid assembly, and investigating the impact of sup versus hac basecalling and metrics for duplex sequencing reads. We undertook this comparison for four reference bacterial strains reflecting different species, genome sizes, %GC content, plasmid content and plasmid sizes. We also evaluated the impact of sequencing depth on the capacity to reconstruct the reference bacterial genomes, and whether flowcell washing would still enable flow cell reuse with the new flowcells and chemistry.

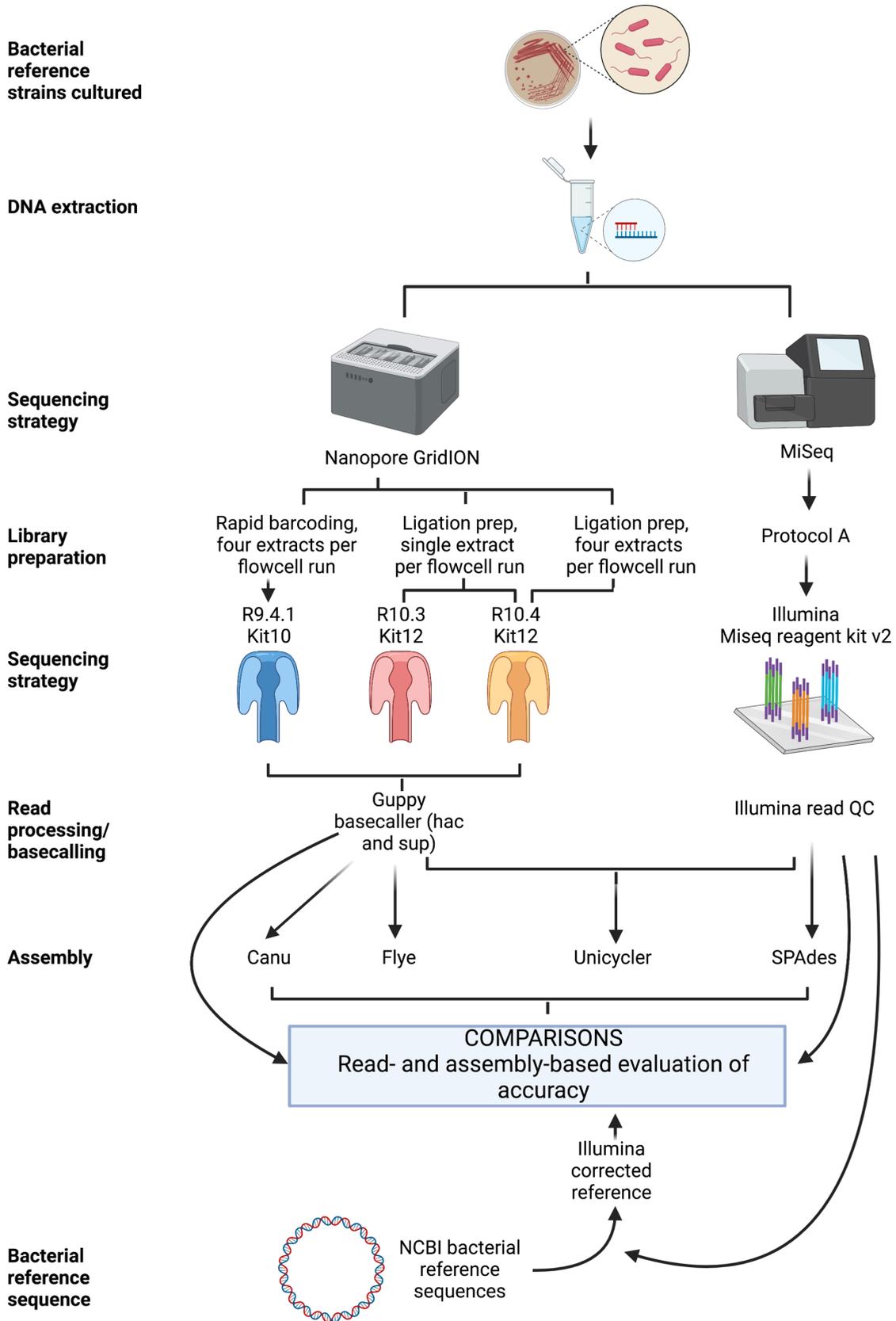


Fig. 1. Experimental workflow.

Table 1. Sequencing read statistics by sequencing modality and bacterial species. Note for R.9.4/Kit10 four isolates were plexed and the total data output is a composite of the individual outputs; for the R10.3/Kit12 and R10.4/Kit12 evaluations each isolate extract was initially run separately. The same flowcell was washed and then re-used for the R10.4 evaluation for the *S. aureus* and then *P. aeruginosa* isolates. Finally, the four DNA extracts were also multiplexed on a single R10.4/Kit12 run

Species	Sequencing modality/sub-group	Total reads	Total bases	N50	Percentage of reads with a phred score of ≥ 20
<i>E. coli</i>	Illumina	3,801,912	574,088,712	151	97.93
	R9.4 (multiplexed run)	353,317	2,364,469,570	11,705	67.1
	R9.4 (multiplexed run; sup called)	339,077	2,242,222,750	11,535	70.03
	R10.3 (single extract/run)	1,073,327	5,964,466,078	9,852	79.05
	R10.3 (single extract/run; sup called)	1,072,758	5,936,766,616	9,827	73.5
	R10.4 (single extract/run; overall)	1,174,227	6,124,985,330	10,507	66.2
	R10.4 (single extract/run; sup called)	1,167,782	6,131,556,595	10,562	79.09
	R10.4 (single extract/run; sup called and duplex reads)	52,171	229,801,689	7,274	98.21
	R10.4 (multiplexed run)	286,239	671,853,044	5,327	72.62
	R10.4 (multiplexed run; sup called and duplex reads)	6,447	10,999,797	3,403	98.06
<i>K. pneumoniae</i>	Illumina	3,202,356	483,555,756	151	97.45
	R9.4 (multiplexed run)	377,192	3,646,791,131	17,396	65.23
	R9.4 (multiplexed run; sup called)	361,657	3,458,646,526	17,157	68.59
	R10.3 (single extract/run)	789,562	7,772,922,913	19,228	77.29
	R10.3 (single extract/run; sup called)	774,119	765,899,2847	19,124	70.24
	R10.4 (single extract/run; overall)	869,853	7,481,444,246	18,612	65.83
	R10.4 (single extract/run; sup called)	865,400	7,495,921,601	18,697	79.79
	R10.4 (single extract/run; sup called and duplex reads)	54,177	452,672,411	16,484	98.62
	R10.4 (multiplexed run)	224,555	1,667,146,081	15,525	72.1
	R10.4 (multiplexed run; sup called and duplex reads)	12,114	95,832,563	15,245	98.82
<i>P. aeruginosa</i>	Illumina	5,299,866	800,279,766	151	97.25
	R9.4 (multiplexed run)	361,977	4,302,642,519	21,597	66.49
	R9.4 (multiplexed run; sup called)	351,155	4,138,688,286	21,342	71.55
	R10.3 (single extract/run)	1,024,134	8,524,041,501	17,666	81.81
	R10.3 (single extract/run; sup called)	1,017,748	8,528,041,241	17,683	76.05
	R10.4 (single extract/run; overall)	556,000	5,851,279,980	24,126	67.35
	R10.4 (single extract/run; sup called)	638,801	6,378,501,910	23,860	82.24
	R10.4 (single extract/run; sup called and duplex reads)	22,859	261,812,617	21,432	98.58
	R10.4 (multiplexed run)	208,693	1,412,016,443	14,627	73.91
	R10.4 (multiplexed run; sup called and duplex reads)	12,468	93,018,395	14,095	98.83
<i>S. aureus</i>	Illumina	9,033,160	1,364,007,160	151	98.98
	R9.4 (multiplexed run)	40,194	725,665,757	33,599	72.67
	R9.4 (multiplexed run; sup called)	39,155	699,249,807	33,066	75.51
	R10.3 (single extract/run)	1,625,258	9,724,520,340	14,338	82.06
	R10.3 (single extract/run; sup called)	1,645,001	9,819,093,990	14,337	78.84

Continued

Table 1. Continued

Species	Sequencing modality/sub-group	Total reads	Total bases	N50	Percentage of reads with a phred score of ≥ 20
	R10.4 (single extract/run; overall)	950,361	7,371,346,901	23,339	74.06
	R10.4 (single extract/run; sup called)	945,421	7,382,123,466	23,446	84.24
	R10.4 (single extract/run; sup called and duplex reads)	47,087	334,258,567	16,366	98.8
	R10.4 (multiplexed run)	80,512	287,957,484	14,301	80.04
	R10.4 (multiplexed run; sup called and duplex reads)	3,753	12,755,562	10,232	99.08

METHODS

Bacterial isolates and DNA extraction

Four reference bacterial strains were sequenced for this study, namely: *Escherichia coli* CFT073 (Genbank accession: NC_004431.1), *Klebsiella pneumoniae* MGH78578 (NC_009648.1-NC_009653.1), *Pseudomonas aeruginosa* PAO1 (NC_002516.2) and *Staphylococcus aureus* MRSA252 (NC_002952.2). Stock cultures were stored at -80°C in nutrient broth supplemented with 10% glycerol. For DNA extraction, stocks were sub-cultured on Columbia blood agar at 37°C overnight.

Long fragment DNA extraction from sub-cultured strains was performed using the Qiagen Genomic tip 100 G^{-1} kit (Qiagen). Quality and fragment length assessments were measured with the Qubit fluorometer (ThermoFisher Scientific) and TapeStation (Agilent). The same DNA extract, stored in elution buffer at 4°C was used for all sequencing experiments. DNA concentration and fragment lengths were evaluated longitudinally to ensure that there was minimal obvious degradation (Tables S1-4, Figs. S1-3).

Nanopore sequencing

The experimental workflow is shown in Fig. 1. For the experiment using the R9.4.1 (FLO-MIN106) flowcell (denoted as R.9.4 throughout), ONT sequencing libraries were prepared by multiplexing DNA extracts from all four isolates using the Rapid Barcoding Sequencing (SQK-RBK004) kit according to the manufacturer's protocol; sequencing was undertaken on a GridION for 48 h.

For the experiments using the R10.3 (FLO-MIN111) and R10.4 (FLO-MIN112) flowcells, ONT sequencing libraries were prepared from DNA extracts using the Q20 +Early Access Kit (SQK-Q20EA) ligation-based protocol. During adapter ligation and clean-up the long fragment buffer was used to enrich for DNA fragments >3 kb. Each DNA extract was sequenced on a single flowcell. After sequencing the *S. aureus* MRSA252 library, the R10.4 (FLO-MIN112) flowcell was washed with the flowcell wash kit

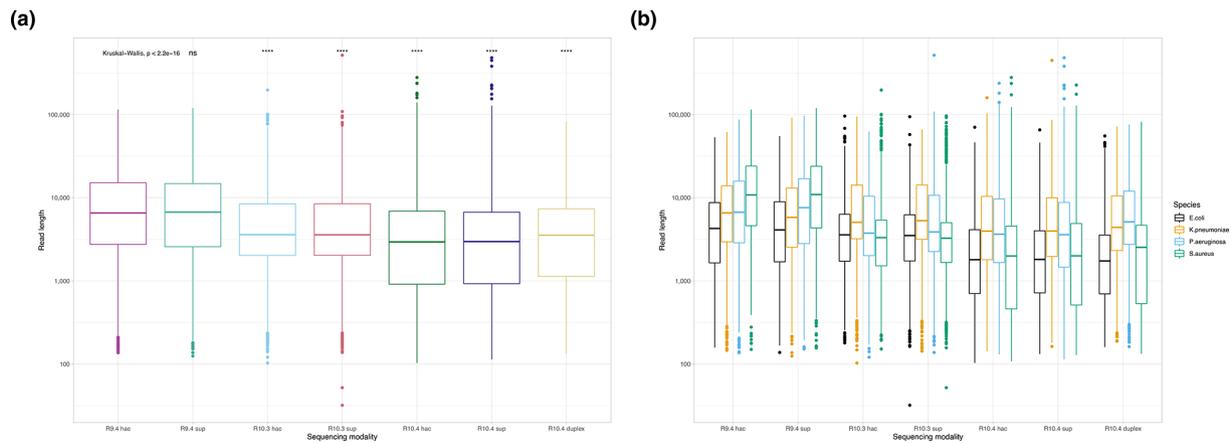


Fig. 2. Read length distributions by (a) modality and (b) by modality and species. Boxplots reflect median (central line) and IQR (box hinges) values, whiskers the smallest and largest values $1.5 \times \text{IQR}$, and dots the outlying points beyond these ranges. Note the y-axis is a log-scale. Median differences in read length were significant across the whole dataset (Kruskal-Wallis test, $P < 0.001$); other significance values represent comparisons with the median read length for R9.4 hac as the reference category (two-sample Wilcoxon test, 'ns' - not significant, '****' - $P < 0.001$). (a) Modality (b) Modality and species.

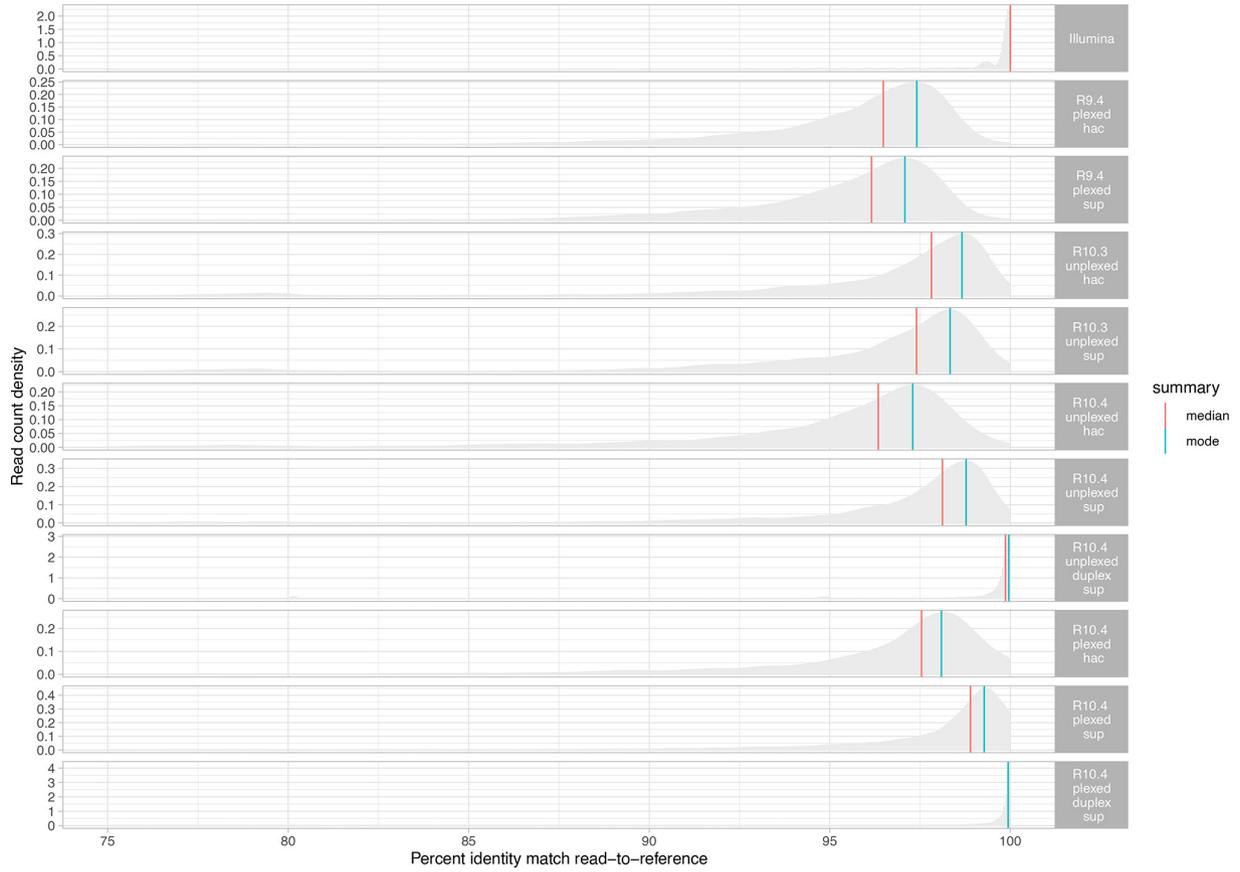


Fig. 3. Median and modal raw read accuracy (% identity when reads are mapped to the Illumina-corrected reference) for each of the major nanopore sequencing modalities, flowcells/kit and basecalling combinations. Reads matching to the reference with <75% identity have been excluded. Complete details summarising all accuracies across all modality, flowcell/kit and basecalling combinations, and stratified by species are represented in Supplementary Table S6.

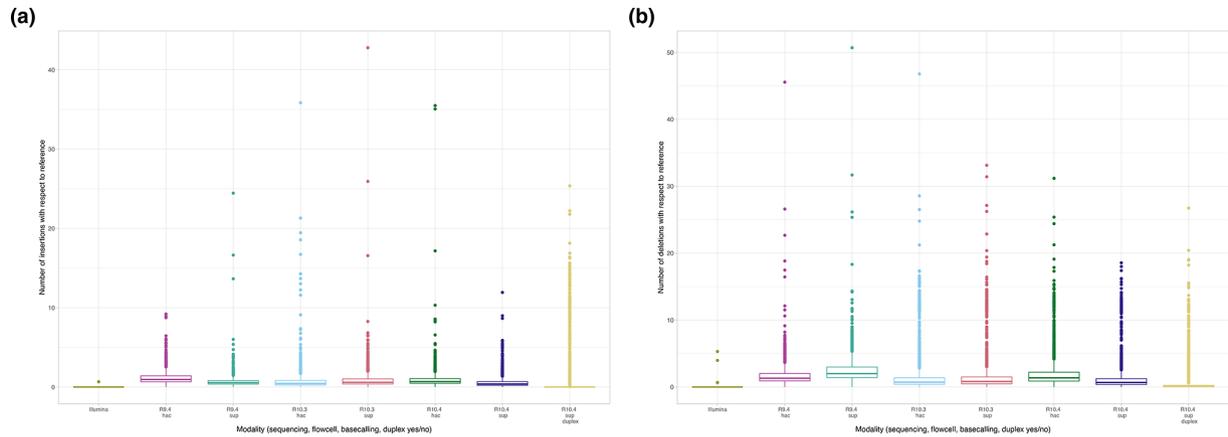


Fig. 4. Number of insertions (panel A) and deletions (panel B) amongst reads mapped to the Illumina-corrected reference for all sequencing modalities. (a) Insertions (b) Deletions.

Table 2. Number of unique contigs by sequencing modality and bacterial species. Using the complete data available (i.e. no subsampling). The first number in each cell represents the number of contigs assembled and matching to the Illumina-corrected reference using dnadiff, the total number of contigs assembled is shown in curved brackets, and the proportion of the reference chromosomal contig covered in square brackets. For *E. coli*, *P. aeruginosa*, *S. aureus* the total number of expected contigs is one, for *K. pneumoniae* one chromosome +5 plasmids. Orange shading shows absent contigs, and/or incomplete assembly (>1 contig matching to reference), and/or extra contigs not matching to reference. Green shaded cells denote complete singular contigs which reflect the reference DNA content at 100+/-1%. "-" denotes no relevant contig assembled

Sequencing modality	Assembler	Plexed Y/N	<i>E. coli</i> chromosome [1] 5231428 bp 51% GC	<i>P. aeruginosa</i> chromosome [1] 6264404 bp 66.2%GC	<i>S. aureus</i> chromosome [1] 2902619 bp 32.8%GC	<i>K. pneumoniae</i> chromosome [1] 5315120 bp 57.0%GC	<i>K. pneumoniae</i> , pKPN3 plasmid [1] 175879 bp 51.7%GC	<i>K. pneumoniae</i> , pKPN4 plasmid [1] 107576 bp 53.4%GC	<i>K. pneumoniae</i> , pKPN5 plasmid [1] 88582 bp 53.8%GC	<i>K. pneumoniae</i> , pKPN6 plasmid [1] 4259 bp 41.4%GC	<i>K. pneumoniae</i> , pKPN7 plasmid [1] 3478 bp 45.7%GC
Illumina	SPAdes	Y	226 (326) [98.64%]	115 (152) [99.72%]	86 (130) [98.42%]	117 (312) [98.9%]	41 [79.7%]	45 [100.72%]	22 [82.49%]	1 [102.82%]	1 [103.45%]
R9.4.1 hac + Illumina	Unicycler	Y	1 [1] [100.09%]	1 [1] [100.49%]	1 [1] [100.69%]	1 [7] [100.1%]	1 [100%]	1 [100%]	1 [100%]	2 [100%]	1 [96.15%]
R9.4.1 hac	Canu	Y	1 [1] [100.59%]	1 [1] [101.29%]	1 [1] [103.18%]	1 (24) [100.72%]	1 [131.55%]	1 [172.51%]	1 [100%]	2 [133.5%]	1 [100%]
R9.4.1 sup	Canu	Y	1 [1] [100.65%]	1 [1] [101.18%]	1 [1] [102.43%]	1 [7] [100.96%]	1 [129.41%]	1 [100%]	1 [125.07%]	1 [100%]	1 [100%]
R10.3 hac	Canu	N	1 [6] [100.49%]	1 [2] [101.19%]	1 [3] [102.78%]	1 [8] [100.72%]	1 [100%]	1 [100%]	1 [100%]	1 [100%]	1 [100%]
R10.3 sup	Canu	N	1 [2] [100.6%]	1 [3] [101.13%]	1 [3] [102.85]	1 [8] [100.74%]	1 [100%]	1 [138.13%]	1 [143.94%]	1 [192.58%]	1 [100%]
R10.4 hac	Canu	N	1 [6] [100.62%]	1 [1] [101.06%]	1 [2] [103.6%]	1 [8] [100.8%]	1 [127.89%]	1 [141.81%]	1 [147.6%]	1 [100%]	1 [112.85%]
R10.4 sup	Canu	N	1 [3] [100.61%]	1 [2] [100.1%]	1 [2] [102.33%]	1 [7] [100%]	1 [100.94%]	1 [100%]	1 [145.43%]	1 [100%]	1 [100%]
R10.4 sup duplex	Canu	N	4 (42) [100.12%]	1 [14] [101.5%]	1 (29) [101.72%]	1 (41) [100.1%]	1 [100%]	3 [130.98%]	1 [149.19%]	1 [100%]	1 [100%]
R10.4 hac	Canu	Y	1 [2] [100.42%]	1 [1] [100.87%]	1 [1] [102.01%]	1 [7] [100.82%]	1 [124.01%]	1 [100%]	1 [142.89%]	1 [100%]	1 [100%]
R10.4 sup	Canu	Y	1 [1] [100.48%]	1 [1] [100.17%]	1 [1] [102.29%]	1 [12] [100.61%]	1 [117.3%]	2 [134.06%]	1 [137.68%]	1 [100%]	1 [100%]
R10.4 sup duplex	Canu	Y	-*	23 (25) [99.2%]	-*	15 (25) [99.2%]	1 [80.39%]	2 [97.73%]	1 [112.86%]	1 [94.01%]	1 [100%]
R9.4.1 hac	Flye	Y	1 [1] [100.09%]	1 [1] [100.14%]	1 [1] [100.7%]	1 [4] [100.11%]	1 [75.7%]	1 [103.68%]	1 [100%]	-	-
R9.4.1 sup	Flye	Y	1 [1] [100.09%]	1 [1] [100.47%]	1 [1] [100%]	1 [5] [100.1%]	1 [100%]	1 [99.99%]	1 [100%]	-	-
R10.3 hac	Flye	N	1 [1] [100.10%]	1 [1] [100.44%]	1 [1] [100.69%]	1 [4] [100.1%]	1 [95.83%]	1 [73.13%]	1 [100%]	-	-
R10.3 sup	Flye	N	1 [1] [100.09%]	1 [1] [100.44%]	1 [1] [100.69%]	1 [4] [100.1%]	1 [100%]	1 [100%]	1 [100%]	-	-
R10.4 hac	Flye	N	1 [1] [100.10%]	1 [1] [100.11%]	1 [1] [100.69%]	1 [4] [100.1%]	1 [100%]	1 [100%]	1 [100%]	-	-
R10.4 sup	Flye	N	1 [1] [100.09%]	1 [1] [100.4%]	1 [1] [100.69%]	1 [5] [100.1%]	1 [100%]	1 [98.92%]	1 [99.99%]	-	-
R10.4 sup duplex	Flye	N	1 [3] [100.10%]	1 [1] [100.8%]	1 [2] [100.69%]	1 [7] [100.21%]	1 [94.27%]	1 [102.93%]	1 [101.54%]	-	1 [100%]
R10.4 hac	Flye	Y	1 [1] [100.10%]	1 [1] [100.16%]	1 [1] [100.69%]	1 [5] [100.1%]	1 [100%]	1 [100%]	1 [100%]	-	-
R10.4 sup	Flye	Y	1 [1] [100.00%]	1 [1] [100.48%]	1 [1] [100.69%]	1 [5] [100.11%]	1 [100%]	1 [100%]	1 [100%]	-	-
R10.4 sup duplex	Flye	Y	37(38) [8.47%]	1 [5] [100.48%]	25(25) [83.64%]	1 [4] [100.4%]	1 [84.71%]	1 [105.27%]	1 [100%]	-	-

*Insufficient read depth for canu to assemble using default settings.

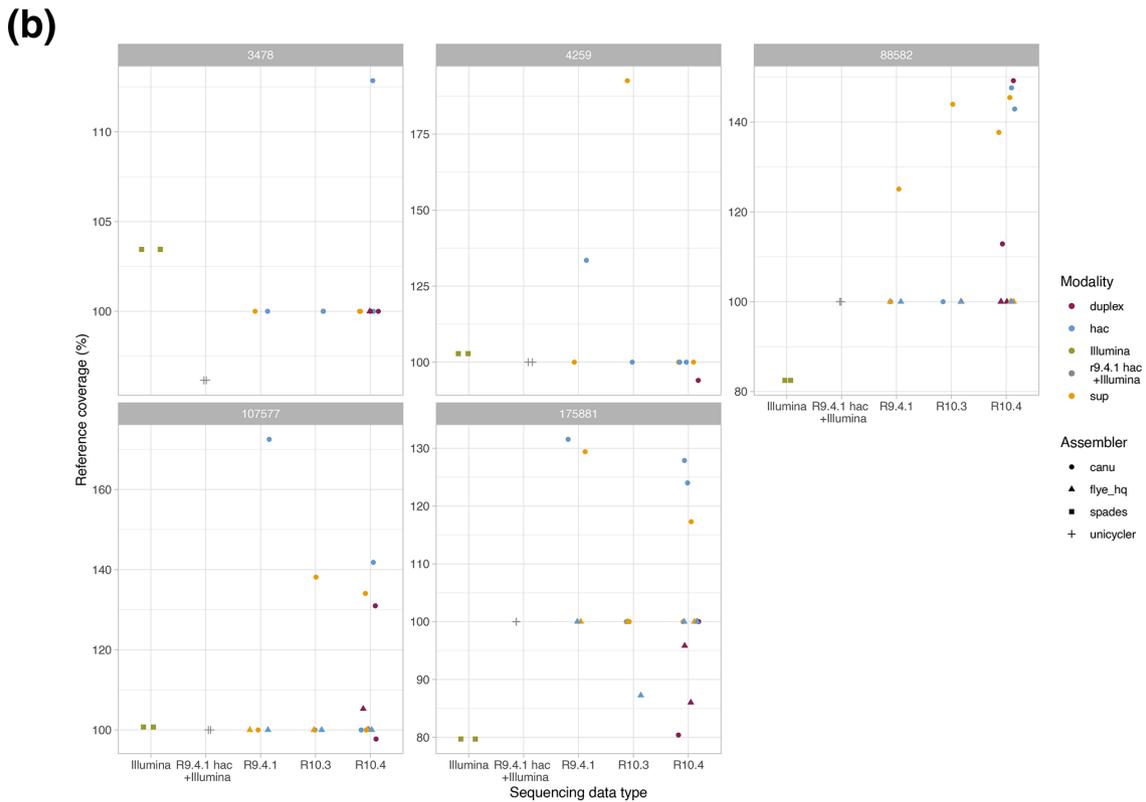
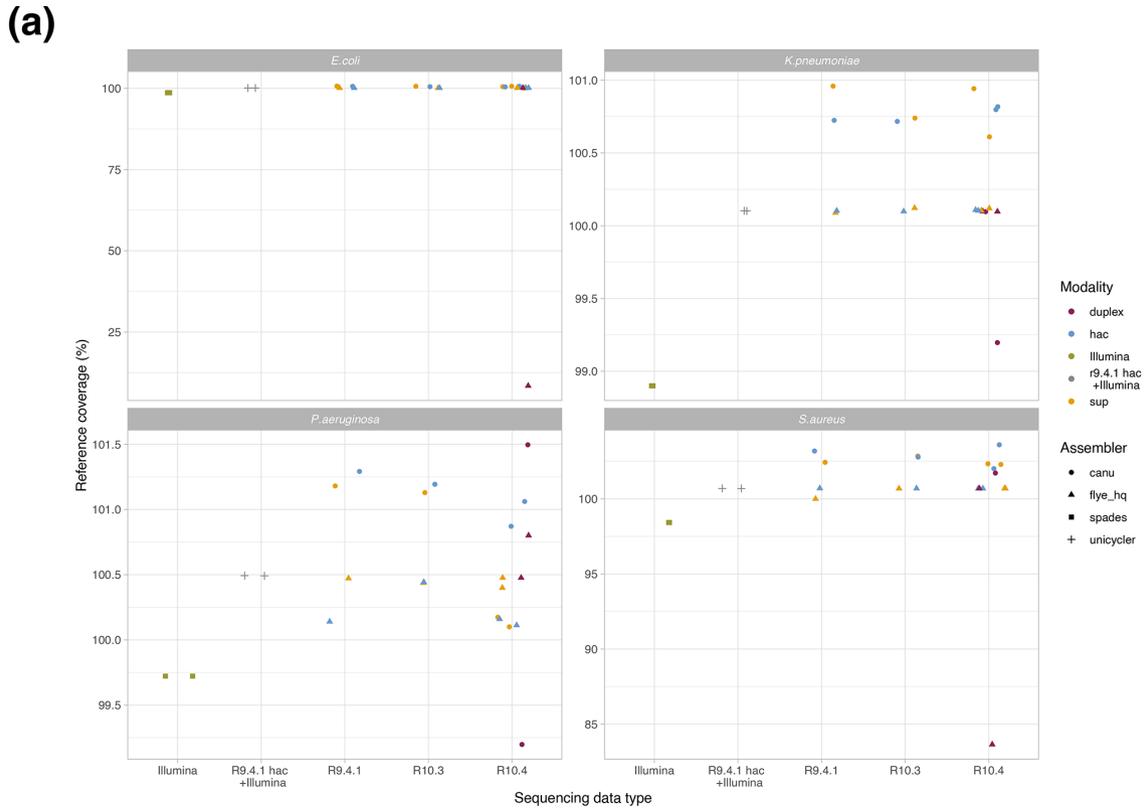


Fig. 5. Assembly reference coverage percentage (%) by sequencing modality, assembler and species. Panel A represents the data for chromosomes and panel B evaluations for the five plasmids known to occur in the *K. pneumoniae* reference strain (labelled by their lengths in bp). Data shown for complete data only (i.e. no sub-sampling performed). (a) Chromosomes (b) Plasmids.



Fig. 6. Assembly accuracy by sequencing modality, assembly strategy and species. Accuracy evaluated on the basis of contig comparisons to Illumina-corrected references using dnadiff, for (a) Indels, and (b) SNPs. NB - SPAdes was only used on Illumina data, and Unicycler hybrid assembly was only performed on R9.4.1+Illumina data. For R10.4, data presented are those from unplexed runs. Dashed black vertical line indicates a threshold of 1 error/100 kb. (a) Indel errors (b) Single nucleotide-level errors.

(EXP-WSH004) according to the manufacturer's protocol, before reusing the flowcell to sequence the *P. aeruginosa* PAO1 library. For the R10.3 experiments, sequencing was undertaken on a GridION for 48 h; for the unplexed R10.4 experiments sequencing times were terminated prematurely. The flowcell usage strategy and pore counts for each flowcell prior to use are summarised in Table S5.

Finally, in a separate experiment, the four DNA extracts were also multiplexed on the R10.4 (FLO-MIN112) flowcell using the Native Barcoding Kit (SQK-NBD112.24); sequencing was undertaken on a GridION for 48 h.

Illumina sequencing

DNA extracts for all isolates were also sequenced on the Illumina MiSeq, as part of two runs plexing 3 bacterial extracts each. Libraries were constructed following the Illumina DNA Prep protocol, according to the manufacturer's instructions (including standard normalization for libraries ['Protocol A']). Library DNA concentrations were quantified by Qubit fluorometry and size distributions of libraries determined using the TapeStation, as above. Sequencing was performed using the MiSeq Reagent Micro Kit v2, generating 150 bp paired-end reads.

Data processing and bioinformatic methods

R10.4 duplex read pairs were identified and prepared for basecalling using ONT's duplex tools (<https://pypi.org/project/duplex-tools/>; v 0.2.9). R9.4, R10.3, and R10.4 raw nanopore reads were hac basecalled with Guppy (ONT) versions 5.0.12+eb1a981 (dna_r9.4.1_450bps_hac.cfg), 5.0.13+bbad529 (res_dna_r103_q20ea_crf_v034.cfg), and 5.0.16+b9fcd7b (dna_r10.4_e8.1_hac.cfg) respectively, as recommended by ONT. R9.4, R10.3, R10.4 (all reads) and R10.4 duplex raw nanopore reads were also basecalled using sup models dna_r9.4.1_e8.1_sup.cfg, dna_r10.3_450bps_sup.cfg, dna_r10.4_e8.1_sup.cfg. Basecalled read summary statistics were generated with SeqKit (v2.2.0) stats using '-T' and '-all' flags [8].

Nanopore reads were subsampled using Rasusa (0.6.1) [9] to depths of 10, 20, 30, 40, 50, and 100 average coverage. Nanopore reads were assembled with Canu (version 2.2, using maxInputCoverage=100 and otherwise default parameters) [10], or Flye (using the --meta and --nano-hq parameters and otherwise defaults, version 2.9-b1768) [11], both of which are commonly used long-read only assemblers that have been shown to optimize long-read only assembly quality [12]. We also explored the impact of polishing nanopore assemblies with one, two and three rounds of Medaka (1.6.0; default settings; <https://github.com/nanoporetech/medaka>).

Subsampled nanopore reads were combined with Illumina reads for hybrid assembly using Unicycler (version 0.4.8, default parameters; includes read polishing with Racon as part of its default workflow) [13]. The SPAdes (version 3.15.3) [14] assemblies generated from Illumina data as part of the Unicycler pipeline were used as the Illumina-only assemblies for comparative evaluations.

Given the previous discrepancies observed between multiple resequenced assemblies for *E. coli* CFT073 and *K. pneumoniae* MGH78578 [15], and the genetic and phenotypic differences observed in different laboratory sub-culture stocks of *P. aeruginosa* PAO1 [16, 17], we generated an Illumina-corrected reference sequence to use as the 'gold standard' comparator for this evaluation. Reference genomes for *E. coli* CFT073 (Genbank accession: AE014075.1), *K. pneumoniae* MGH78578 (CP000647.1), *P. aeruginosa* PAO1 (NC_002516.2), *S. aureus* MRSA252 (NC_002952.2) and the respective Illumina datasets generated for this study were used as inputs for the SNIPPY pipeline (version 4.6.0) (<https://github.com/tseemann/snippy>); output consensus fasta files represented the new Illumina-corrected reference sequences used in this study.

Assembled contigs from nanopore, Illumina, and hybrid assemblies were compared against the Illumina-corrected reference sequences using DNAdiff version 1.3(18).

Assembled contigs from nanopore, Illumina, and hybrid assemblies as well as the Illumina-corrected reference sequences were annotated with Prokka (version 1.14.6) [18], using the corresponding reference GenBank files to ascertain reference proteins using the '--proteins' flag.

Translated amino acid sequences for Prokka annotations in the different test assemblies (Canu, Flye [long-read only], Unicycler [hybrid long-/short-read], SPAdes [short-read only]) and Illumina-corrected reference sequences were compared using the script AAcompare.py in the workflow provided (see below for the repository link). This looked for exact amino acid sequence matches (i.e. 100% identity along 100% of the translated protein) between the Illumina-corrected reference and assembled contigs to determine how intact assembled coding sequences were for each assembly method.

Per-read error rates were calculated by mapping the raw reads to the Illumina corrected reference sequences using minimap2 (version 2.22-r1101) [19]. The percent identity was calculated from the query distance (NM tag) divided by the query length, multiplied by 100, using the bamreadstats.py script provided in the gitlab repository (link below).

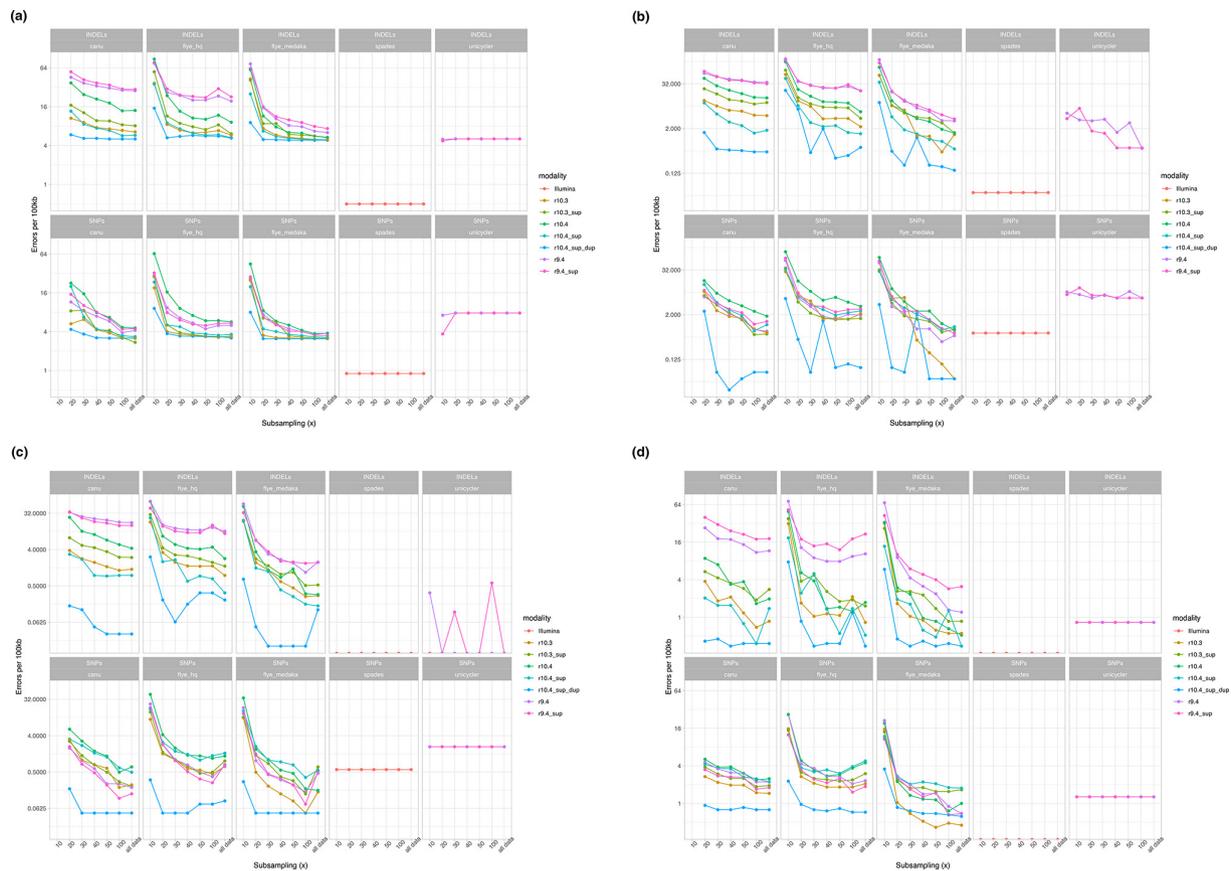


Fig. 7. Impact of subsampling of long-read datasets on assembly accuracy. Presented here by species for Indels (top panels), and SNPs (lower panels). For ease of representation, only data for Flye assemblies polished with one round of Medaka are shown, as the effects of additional polishing was shown to be marginal for most modalities (Fig. S6, Table S7). Data for 10x long-read coverage is omitted for Canu assemblies as this coverage was considered too low for default settings and was unlikely to improve results. (a) *E. coli* (b) *K. pneumoniae* (chromosome only) (c) *P. aeruginosa* (d) *S. aureus*.

A workflow for this analysis has been written using nextflow [19] and is available on gitlab (https://gitlab.com/ModernisingMedicalMicrobiology/assembly_comparison, tagged version v0.5.5). Outputs from the analyses are also available in this separate repository (https://gitlab.com/ModernisingMedicalMicrobiology/assembly_comparison_analysis).

Statistical analysis

The Kruskal-Wallis test was used to evaluate any statistically significant difference in median read lengths across all sequencing modalities and the Mann-Whitney-Wilcoxon (two-sample Wilcoxon) test used to evaluate statistical significance of pairwise differences in median read lengths for each sequencing modality when compared with R9.4/Kit 10 median read lengths.

Data visualization

Figures and plots for this manuscript were generated using the ggplot2 and patchwork packages in R (v3.6.2), and Biorender (www.biorender.com).

RESULTS

Sequencing yield and read length distributions

The total data yield after 48 h of sequencing from the R9.4 flowcell was 11.0 Gb (four isolate extracts multiplexed on one sequencing run), compared with 4.0 Gb for the R10.4 multiplexed run (Table 1, Fig. S4, available in the online version of this article). For the individual R10.3 flowcells a median of 8.2 Gb/flowcell (IQR: 7.3–8.8 Gb) were generated by 48 h of sequencing, and 6.7 Gb/flowcell (IQR: 6.6–7.4 Gb) for the R10.4 flowcells respectively by 20–30 h of sequencing (Table 1, Fig. S4). In total, 32.2 Gb of data were generated for the extracts from the Illumina runs (Table 1).

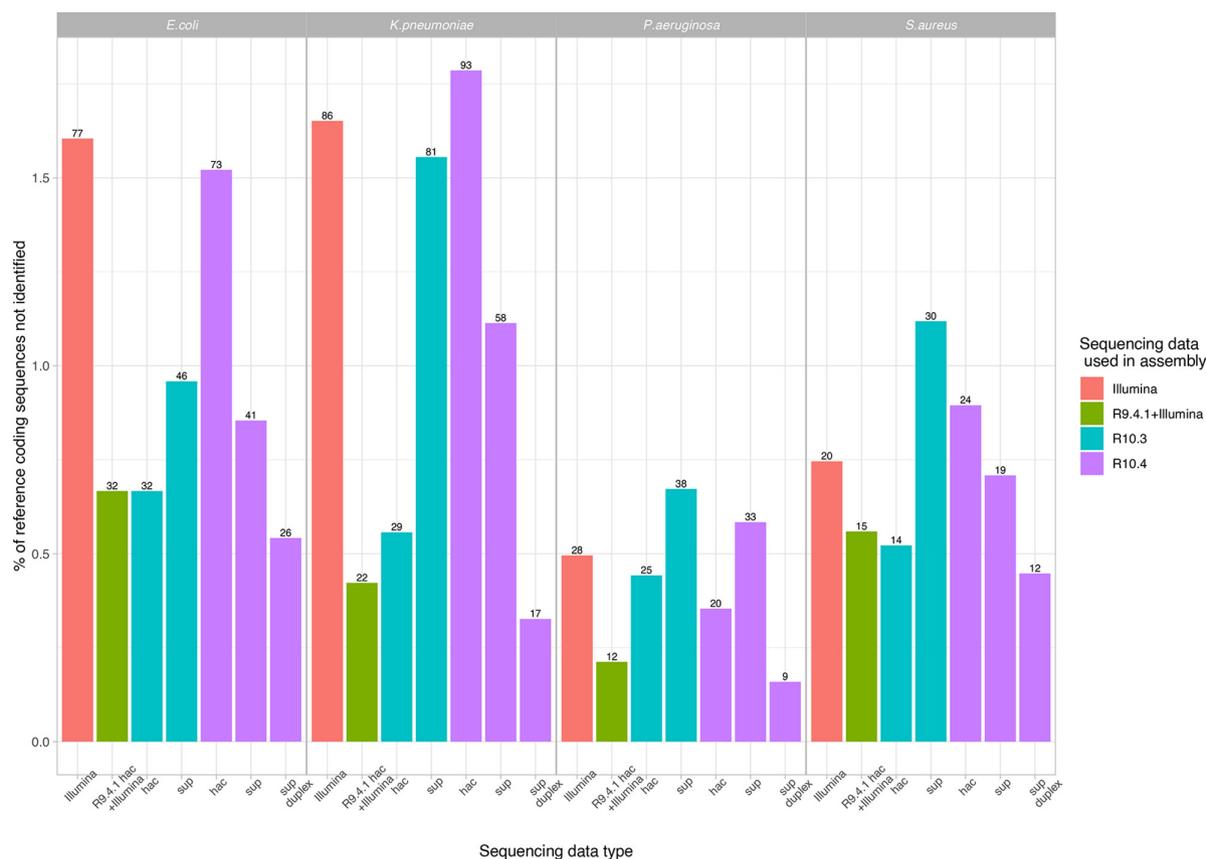


Fig. 8. Coding sequence (CDS) recovery on the basis of exact CDS (amino acid sequence) matches with respect to the Prokka-annotated Illumina-corrected reference (chromosome +all plasmids for *K. pneumoniae*). Plot shows the percentage of reference coding sequences missed by each modality. For long-read data only Flye assemblies with one round of polishing with Medaka are shown; for R10.3 and R10.4 datasets these were from non-multiplexed evaluations (i.e. only single extracts per flowcell). For Unicycler, the assembly using R.9.4 hac +Illumina data is shown. The total number of coding sequences missed by each approach is shown as a number at the top of each bar.

Read length distributions for a subsample of 1000 reads by modality and species are shown in Fig. 2; overall, across species for nanopore data the median read length was 3580 bp, the maximum read length 388620 bp and the minimum read length 77 bp. Median read lengths generated using R9.4 were longer (6273 bp versus 2930 bp for R10.4; two-sample Wilcoxon test, $P < 0.001$, comparison for hac basecalled data; Fig. 2a). N50s are represented in Table 1; median N50 across species was 19496 bp for R9.4.1 hac, 16002 bp for R10.3, 20976 bp for R10.4 (all) and 16425 bp for R10.4 duplex reads.

Duplex reads

The median proportion of duplex reads across the four unplexed, single-extract R10.4 runs was 4.5% (3.8% for *E. coli*, 6.1% for *K. pneumoniae*, 4.5% for *P. aeruginosa*, and 4.5% for *S. aureus*). For the multiplexed R10.4 run for each species these proportions were 2.3, 5.4, 6.0 and 4.7%.

Raw read accuracy by sequencing modality and species

Raw read accuracy (percent identity when mapped to the reference) for a subsample of 1000 reads by sequencing data type/process (i.e. 'sequencing modality') and species was highest (as expected) for Illumina reads (modal accuracy: 100.0%), followed by R10.4 duplex reads basecalled with the sup model (modal accuracy: 99.9%); modal accuracies for all the other approaches were $>97.0\%$ (Fig. 3). Sup basecalling improved modal accuracy for R10.4 reads, but not R10.3 or R9.4 reads; multiplexing had no impact (Fig. 3). Median and modal accuracies for each sequencing modality by species are detailed in Table S6.

In terms of insertions and deletions with respect to the reference, for long-read modalities R10.4 sup called duplex data performed best (Fig. 4a, b). The median number of insertions observed per read was 0.94, 0.45, 0.37 and 0.0 for R9.4 hac, R10.3 hac and R10.4 sup and R10.4 sup duplex respectively (two-sample Wilcoxon test for each versus R9.4 hac as the reference category; all

$P < 0.001$), and for deletions 1.31, 0.73, 0.63 and 0.10 respectively (two-sample Wilcoxon test for each versus R9.4 hac as the reference category; all $P < 0.001$).

Assembly accuracy with respect to number of expected contigs in the reference sequences and reference sequence size

We evaluated the capacity of each sequencing approach to accurately reconstruct (i) the number of known contigs present in each reference isolate, and (ii) what percentage of the Illumina-corrected reference was covered. All isolates contained single chromosomes only, except the *K. pneumoniae* reference, which contained a chromosome and five plasmids ranging in size from 3478 to 175879 bp (Table 2).

Approaches using all the data and Unicycler or Flye largely generated single chromosomal contigs, except those using R10.4 duplex reads only, particularly for multiplexed extracts, likely because these reads were insufficient to cover the whole genome (Table 2; Fig. S5A). Illumina-only assemblies generated much larger numbers of contigs as expected (Table 2). Using all the data, single *K. pneumoniae* plasmid contigs were mostly obtained using any of the long-read data and Flye, or hybrid assembly with Unicycler (Table 2, Fig. S5B). Using all the data, Flye long-read only assemblies largely all missed the two smallest plasmids (Table 2, Fig. S5B).

Sub-sampling the data to 10×, 20×, 30×, 40×, 50× or 100× depth had variable effect - for the most part single chromosomal contigs were assembled using long-reads only with $>20\times$ depth; Unicycler could mostly be used with 10× long-read depth (Fig. S5A). The same effect was seen for plasmids, except Flye struggled to reliably assemble the two largest plasmids into single contigs with lower sequencing depths (Fig. S5B). Canu assemblies failed with 10× sub-sampling, as expected given the default cut-offs.

For chromosomes, Canu long-read only assemblies tended to over-assemble structures (i.e. reference coverage $>100\%$, Fig. 5a) whilst Illumina-only assemblies under-assembled structures. Reference coverage percentage for Unicycler hybrid (R9.4+Illumina) was largely unaffected by sub-sampling the data to 10×, 20×, 30×, 40×, 50× or 100× (Fig. 5a). For plasmids, Canu assembly again largely over-assembled the structures; Unicycler hybrid (R9.4+Illumina) assembly was the only approach which consistently assembled all plasmids at near 100% reference coverage across all sub-sampling depths (Fig. 5b).

Assembly accuracy with respect to insertions, deletions and nucleotide-level mismatches

For each sequencing and assembly modality the number of indels and nucleotide-level mismatches (SNPs) were evaluated by species (Fig. 6a and b) and overall (Table S7). The impact of sub-sampling and relevance of long-read sequencing depth was also considered (Fig. 7).

Overall, SPAdes assemblies had the fewest indels (0.02 indels/100 kb), followed by Medaka-polished Flye-assembled R10.4 sup basecalled/duplex reads (0.18 indels/100 kb), Medaka-polished Flye-assembled R10.4 sup basecalled data (0.41 indels/100 kb), Medaka-polished Flye-assembled R10.3 hac basecalled data (for three rounds of polishing: 0.44 indels/100 kb) and Unicycler assemblies (0.56 indels/100 kb) (Table S7). There were apparent species-specific differences, with the *E. coli* reference proving the most challenging to assemble accurately (Fig. 6a). The improvements in the indel error rates of R9.4 or R10.4 Flye assemblies polished with two or three rounds of Medaka versus one round were negligible; however, additional rounds of polishing improved indel errors in R10.3 hac basecalled assemblies (Fig. 6a, Fig. S6, Table S7).

Similar trends were observed overall for SNPs, with the lowest error rates (0.21 SNPs/100 kb of sequence) observed for multiply-Medaka-polished Flye-assembled R10.3 hac basecalled data, or singly-Medaka-polished Flye assembled R10.4 sup basecalled/duplexed data (0.21 SNPs/100 kb of sequence) (Fig. 6b, Table S7). SNP error rates for Unicycler assemblies however were higher than for the other optimised assembly modalities (4.38 SNPs/100 kb) (Table S7). Polishing Flye assemblies with Medaka improved SNP error rates over unpolished assemblies, but there were no obvious benefits of multiple rounds of polishing (Fig. 6b, Fig. S6). Again, species-specific differences were observed, with the *E. coli* reference the most challenging to assemble (Fig. 6B).

Error rates for Unicycler assemblies were largely consistent at all long-read sequencing depths from 10× to up to strategies using all the data; error rates for long-read-only assemblies were optimised when coverage was $\geq 20\times$ (Fig. 7).

Assembly accuracy with respect to coding sequence content

Coding sequence content was most accurately recovered using Flye-assembled sup basecalled R10.4 duplex data and hybrid assembly (Fig. 8; missing between 9–32 [~ 0.25 – 0.75%] of coding sequences across species). Long-read only assembly with R9.4 data missed up to 10–15% of coding sequences (data not plotted in Fig. 8). Notably, the duplex datasets from the unplexed 10.4 runs were used, as from multiplexed runs the duplex yields were insufficient to facilitate assembly in most cases (Table 2).

DISCUSSION

In this pragmatic study evaluating the impact of different nanopore sequencing flowcells and chemistries on the capacity to fully reconstruct genomes of four commonly studied bacteria, we have shown that sup basecalled R10.4/Kit12 data and sup called

duplex data have read- and assembly-level accuracies that would enable these to be effectively used for the reconstruction of bacterial genomes without requiring Illumina data to generate hybrids. However, hybrid assembly (Illumina+R9.4.1 PacBio data) remains the most robust approach in terms of contig (both chromosomes and plasmids) and CDS recovery without over-assembly, and facilitates the multiplexing of large numbers of isolates per flowcell, given that in this and at least one other study [3], $\leq 10\times$ long-read depth is required for the accurate reconstruction of chromosomes and plasmids by combining R9.4.1 and Illumina data using Unicycler. Highly accurate long-read only assembly and genome reconstructions was optimized by generating duplex reads, which in our hands made up a small proportion of the output ($<10\%$); as such, it would come at a significant cost per isolate as a result of being able to only generate data for 1–2 isolates per flowcell. Very approximate costs per genome therefore for hybrid assembly versus duplex/long-read-only assembly would be £50–70/genome versus £300–600/genome.

Although barcoding up to 96 isolates has recently been enabled for the R10.4/Kit12 combination, the data yields per flowcell (~ 4 Gb) would likely preclude viable assembly for 96 *E. coli* isolates with a typical genome size of ~ 5 Mb (would give $<8\times$ coverage). There is also a current requirement to use a ligation-based library preparation, which lengthens the processing time, and may impact on plasmid recovery [6]. We observed issues with recovering small plasmids (<5 kb) using Flye in this study although both of these small plasmids could be reliably recovered in Canu assemblies; consistent with this a previous evaluation has shown that 8–15% of small plasmids are not recovered using these long-read-only assemblers [12]. Similarly, as shown in this study and in other work [12], the basic Canu workflow ‘over-assembles’ the data, and contigs require trimming of overlaps in order to recreate accurate, single, circularized structures. We observed some apparent species-specific differences, suggesting that assemblers are more challenged in accurately reconstructing certain genomes; these differences, as well as differences related to genome length and the impact on long-read sequencing depth may be important to consider in study design.

There are currently few other published studies on the performance of R10.4/Kit12 for bacterial analyses. We found only one publication investigating its use on a mock microbial community (seven bacterial species and one fungal species) which found similar modal accuracy scores of 99% using sup basecalling, and a requirement of $40\times$ to be able to reliably assemble a bacterial genome [20]. Their hypothesis was that improved read accuracies were due to an improved ability to call homopolymers for lengths up to ten bases, which we did not investigate in this manuscript. It was unclear what proportion of reads they characterized as duplex reads.

There are several limitations of our study. We have not exhaustively investigated all possible approaches to genome assembly, but rather taken a pragmatic approach in assembling the data with several commonly used assemblers, without additional bespoke management or combination of workflows; the data are however available for other researchers to trial different approaches. We had low duplex read yields compared with those reported by ONT (up to 30–40% per flowcell); further optimization is needed to see if these can be achieved. We have investigated only a limited number of isolates and plasmids, but these represent a range of %GC and sizes, and are likely to reflect genetic content more widely in other species; we have not generated replicate datasets. Similarly, because we only investigated one isolate per species, it may be that the differences observed are not generalisable or are strain and not species-specific; this would be interesting future work. Improvements and upgrades to nanopore flowcells, chemistries and basecallers occur regularly and nanopore will be releasing the R10.4.1 flowcell and Kit14 chemistries later in 2022 which may further optimise the quality of long-read only outputs.

In summary, the combination of R10.4/Kit12 flowcells/chemistries look very promising for highly accurate, long-read only bacterial genome assembly; however, this requires superior accuracy basecalling, and is optimised by the generation of duplex reads, which currently make up only a small proportion of sequencing yield. In addition, for large-scale projects to fully reconstruct 100s-1000s of bacterial isolates, hybrid assembly, multiplexing and the use of flowcells/chemistries that support rapid barcoding are currently better suited for higher throughput and are more cost-effective per reconstructed genome. The optimal strategy in any given context will depend on the specific use case and resources available, and may evolve rapidly over short timescales.

Funding information

This study was funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (NIHR200915), a partnership between the UK Health Security Agency (UKHSA) and the University of Oxford, and was supported by the NIHR Oxford Biomedical Research Centre (BRC). The computational aspects of this research were funded from the NIHR Oxford BRC with additional support from the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NHS, NIHR, UKHSA or the Department of Health and Social Care. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. NS is an NIHR Oxford BRC Senior Research Fellow and an Oxford Martin Fellow.

Acknowledgements

We are grateful to Dr Celiq Souque and Professor Craig Maclean at the Department of Zoology, University of Oxford, for supplying the *Pseudomonas aeruginosa* PAO1 strain. We are also grateful for helpful feedback from the Twitter community following the release of this manuscript as a preprint.

Author contribution

N.K., N.Sa., D.W.C. and N.St. designed the study. N.K., G.R., T.S. and H.W. performed the laboratory experiments and sequencing. N.Sa. performed the bioinformatics analysis; S.L. provided input on bioinformatics approaches. N.St. generated the data visualisations. N.St., N.Sa. and N.K. wrote the first draft. All authors reviewed and approved the final draft.

Conflicts of interest

Oxford Nanopore Technologies supplied the R10.3 and R10.4 flowcells free of charge for this study. They were also involved in discussions regarding which data processing approaches to use to optimise basecalling and assembly outputs; however, they did not impact on the presentation of any of the results.

Ethical statement

Not applicable.

References

1. Van Goethem N, Descamps T, Devleeschauwer B, Roosens NHC, Boon NAM, *et al.* Status and potential of bacterial genomics for public health practice: a scoping review. *Implement Sci* 2019;14:79.
2. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, *et al.* Niche and local geography shape the pangenome of wastewater- and livestock-associated *Enterobacteriaceae*. *Sci Adv* 2021;7:15.
3. Arredondo-Alonso S, Pöntinen AK, Cléon F, Gladstone RA, Schürch AC, *et al.* A high-throughput multiplexing and selection strategy to complete bacterial genomes. *Gigascience* 2021;10:giab079.
4. Lipworth S, Pickford H, Sanderson N, Chau KK, Kavanagh J, *et al.* Optimized use of Oxford Nanopore flowcells for hybrid assemblies. *Microb Genom* 2020;6:11.
5. Oxford Nanopore Technologies; (n.d.). <https://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store>
6. Wick RR, Judd LM, Wyres KL, Holt KE. Recovery of small plasmid sequences via Oxford Nanopore sequencing. *Microb Genom* 2021;7.
7. Benton M. Nanopore Guppy GPU basecalling on Windows using WSL2; 2021. <https://hackmd.io/@Miles/rkYKDHPsO>
8. 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.
9. Hall MB. (n.d.) Rasusa: randomly subsample sequencing reads to a specified coverage. *JOSS*;7:3941.
10. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al.* Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* 2017;27:722–736.
11. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;37:540–546.
12. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res* 2019;8:2138.
13. 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.
14. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. *Curr Protoc Bioinformatics* 2020;70:e102.
15. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, *et al.* Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb Genom* 2019;5.
16. Klockgether J, Munder A, Neugebauer J, Davenport CF, Stanke F, *et al.* Genome diversity of *Pseudomonas aeruginosa* PAO1 laboratory strains. *J Bacteriol* 2010;192:1113–1121.
17. Chandler CE, Horspool AM, Hill PJ, Wozniak DJ, Schertzer JW, *et al.* Genomic and phenotypic diversity among ten laboratory isolates of *Pseudomonas aeruginosa* PAO1. *J Bacteriol* 2019;201:e00595-18.
18. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
19. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, *et al.* Versatile and open software for comparing large genomes. *Genome Biol* 2004;5:R12.
20. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, *et al.* Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods* 2022;19:823–826.

Five reasons to publish your next article with a Microbiology Society journal

1. When you submit to our journals, you are supporting Society activities for your community.
2. Experience a fair, transparent process and critical, constructive review.
3. If you are at a Publish and Read institution, you'll enjoy the benefits of Open Access across our journal portfolio.
4. Author feedback says our Editors are 'thorough and fair' and 'patient and caring'.
5. Increase your reach and impact and share your research more widely.

Find out more and submit your article at microbiologyresearch.org.