

Optimized use of Oxford Nanopore flowcells for hybrid assemblies

Samuel Lipworth^{1,*}, Hayleah Pickford¹, Nicholas Sanderson^{1,2}, Kevin K. Chau¹, James Kavanagh¹, Leanne Barker¹, Alison Vaughan^{1,2}, Jeremy Swann^{1,3}, Monique Andersson⁴, Katie Jeffery⁴, Marcus Morgan⁴, Timothy E. A. Peto^{1,2}, Derrick W. Crook^{1,2,4}, Nicole Stoesser^{1,4†} and A. Sarah Walker^{1,2†}

Abstract

Hybrid assemblies are highly valuable for studies of *Enterobacteriaceae* due to their ability to fully resolve the structure of mobile genetic elements, such as plasmids, which are involved in the carriage of clinically important genes (e.g. those involved in antimicrobial resistance/virulence). The widespread application of this technique is currently primarily limited by cost. Recent data have suggested that non-inferior, and even superior, hybrid assemblies can be produced using a fraction of the total output from a multiplexed nanopore [Oxford Nanopore Technologies (ONT)] flowcell run. In this study we sought to determine the optimal minimal running time for flowcells when acquiring reads for hybrid assembly. We then evaluated whether the ONT wash kit might allow users to exploit shorter running times by sequencing multiple libraries per flowcell. After 24 h of sequencing, most chromosomes and plasmids had circularized and there was no benefit associated with longer running times. Quality was similar at 12 h, suggesting that shorter running times are likely to be acceptable for certain applications (e.g. plasmid genomics). The ONT wash kit was highly effective in removing DNA between libraries. Contamination between libraries did not appear to affect subsequent hybrid assemblies, even when the same barcodes were used successively on a single flowcell. Utilizing shorter run times in combination with between-library nuclease washes allows at least 36 *Enterobacteriaceae* isolates to be sequenced per flowcell, significantly reducing the per-isolate sequencing cost. Ultimately this will facilitate large-scale studies utilizing hybrid assembly, advancing our understanding of the genomics of key human pathogens.

DATA SUMMARY

(1) Raw sequencing data are available via the National Center for Biotechnology Information (NCBI) under project accession number PRJNA604975. Sample accession numbers are provided in Table S1 (available in the online version of this article).

(2) Assemblies are available via Figshare <https://doi.org/10.6084/m9.figshare.11816532>.

INTRODUCTION

Ideally, a single sequencing technology would provide both highly accurate and structurally complete genomes. The rapid acceleration in whole-genome sequencing over the past decade has been driven primarily by short-read technologies (e.g. Illumina). The 100–300 bp reads generated are generally highly accurate and low cost, and the tools for their analysis are now relatively mature. However, the inability to resolve long genomic repeats using short reads is a significant limiting factor. In *Enterobacteriaceae*, clinically important genes, such as those involved in antimicrobial resistance (AMR)

Received 16 March 2020; Accepted 25 September 2020; Published 11 November 2020

Author affiliations: ¹Modernising Medical Microbiology, Nuffield Department of Medicine, University of Oxford, UK; ²NIHR Oxford Biomedical Research Centre, Oxford, UK; ³NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with Public Health England, Oxford, UK; ⁴Department of Clinical Microbiology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK.

***Correspondence:** Samuel Lipworth, samuel.lipworth@ndm.ox.ac.uk

Keywords: hybrid assembly; Nanopore sequencing; *Enterobacteriaceae*; long-read assembly; bacterial genomics.

Abbreviations: AMR, antimicrobial resistance; Indel, insertion and/or deletion; IQR, interquartile range; ONT, Oxford Nanopore Technologies; SNP, single-nucleotide polymorphism.

†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. Two supplementary tables and nine supplementary figures are available with the online version of this article.

000453 © 2020 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.

and virulence, are commonly carried on plasmids and other mobile genetic elements (MGEs) [1]. It is generally impossible to delineate the structure of these using short-read data alone [2].

Long-read sequencing platforms such as Oxford Nanopore Technologies (ONT) or Pacific Biosciences (PacBio) can produce reads that are thousands or tens of thousands (and even hundreds of thousands) of bases long. This greatly aids *de novo* assembly because these reads span long genomic repeats. Particularly in the case of ONT, however, longer reads are still currently associated with a higher error rate, which may be problematic for some applications (e.g. transmission inference). Improvements in laboratory and bioinformatic methods to enable sequencing using only long-reads are emerging at a rapid pace. Significant limitations remain, however, and there has been little evaluation using real-world data [3]. Hybrid approaches combine the low error rate of Illumina reads with the structural resolution of ONT/PacBio, maximizing the strengths of both technologies [4], and the widely used Unicycler tool [5] offers an automated and easy-to-use pipeline for this. Large-scale studies utilizing hybrid assemblies would likely provide valuable new insights into the biology of MGEs in *Enterobacteriaceae*; however, the significant associated cost currently limits the widespread application of this technique.

Recent research has suggested that random subsampling of ONT reads can improve hybrid assemblies [6], raising the possibility that significantly shorter sequencing times may be suitable where long reads are being created for the purpose of hybrid assembly. Producing sufficient reads to complete hybrid assemblies for one library of isolate extracts may only require a small proportion of the potential useful sequencing time of a flowcell. In theory, therefore, it should be possible to sequence multiple libraries on each flowcell, thereby reducing the per-isolate cost. The major obstacle to this is the need to eliminate contamination between libraries sequenced sequentially on the same flowcell. ONT has recently released a version 3 wash-kit with the addition of nuclease. The company quotes between-library contamination as being around 0.1% [7]; however, to our knowledge, this has not been independently verified.

This study therefore evaluated whether the ONT wash-kit could enable successful reuse of flowcells to increase the number of hybrid assemblies per flowcell for isolates with existing Illumina short-read data. In doing so we investigated: (i) whether sequencing run times could be shortened without affecting assembly quality; and (ii) whether between-library contamination from reusing flowcells with the new wash kit occurs and can be mitigated. Whilst we primarily focused on hybrid assembly, we also compared hybrid to long read-only assemblies to assess whether short-read sequencing remains necessary to produce complete and accurate assemblies. Based on these evaluations, we propose a rapid and simple workflow that potentially

Impact Statement

Most existing sequencing data have been acquired from short-read platforms (e.g. Illumina). For some species of bacteria, clinically important genes, such as those involved in antibiotic resistance and/or virulence, are carried on plasmids. Whilst Illumina sequencing is highly accurate, it is generally unable to resolve complete genomic structures due to repetitive regions. Hybrid assembly uses long reads to scaffold together short-read contigs, maximizing the benefits of both technologies. A major limiting factor when using hybrid assemblies at scale is the cost of sequencing the same isolate with two different technologies. Here we show that high-quality hybrid assemblies can be created for most isolates using significantly shorter run times than are currently standard. We demonstrate that a simple washing step allows several libraries to be run on the same flowcell, facilitating the ability to take advantage of shorter running times. Adding nuclease means that contamination between libraries is minimal and has no significant effect on the quality of subsequent hybrid assemblies. This approach reduces the cost of acquiring long reads by at least 27%, paving the way for large-scale studies utilizing hybrid assemblies, which will likely significantly enhance our understanding of the genomics of important human pathogens.

reduces the consumables cost of ONT sequencing by at least 27% with no apparent impact on assembly accuracy.

METHODS

Isolate preparation, DNA extraction and sequencing

Forty-six isolates were selected for sequencing, of which 45 were cultured from bacteraemic patients presenting to Oxford University Hospitals NHS Foundation Trust between 2008–2018, and 1 was the MGH78578 *Klebsiella pneumoniae* reference (Table S1). Pure isolate cultures were stored at -80°C in 10% glycerol. Sub-cultures of isolate stocks were grown on Columbia blood agar overnight at 37°C . DNA for Illumina sequencing was extracted using the QuickGene DNA extraction kit (Autogen, MA, USA) as per the manufacturer's instructions with the addition of a mechanical lysis step (FastPrep, MP Biomedicals, CA, USA; 6 m s^{-1} for 40 s). Short-read sequencing was performed using an Illumina HiSeq 4000 instrument as described previously [6].

For Nanopore sequencing, DNA from isolates for library 1 (Table S1) was extracted using the EasyMag system (bioMerieux). A $10\ \mu\text{l}$ loop was used to inoculate $500\ \mu\text{l}$ of autoclaved phosphate-buffered saline and $100\ \mu\text{l}$ of this was transferred to the easyMAG vessel, which was then run using the manufacturer's generic short protocol and

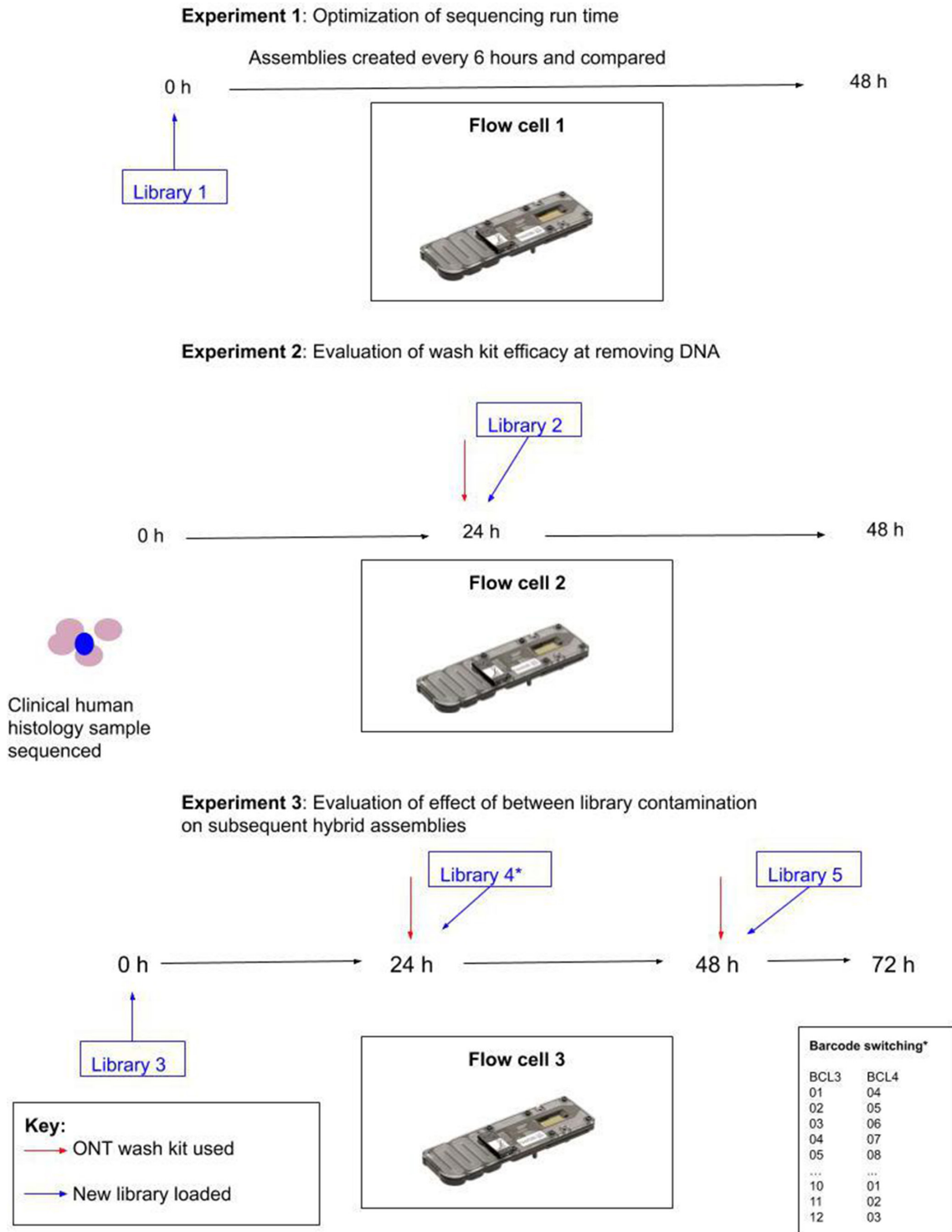


Fig. 1. Schematic representation of experiments performed, flowcells used and libraries sequenced. *The same 12 isolates were sequenced in both libraries 3 and 4, but with different barcodes, as shown in the inset table and Table 1.

Table 1. Comparison of SNPs and indels detected by DNADiff between hybrid assemblies of the same isolates sequenced in libraries 3 and 4

Isolate name	Barcode library 4	Barcode library 3	SNPs	Indels	% identity	Species	MLST
blc-44	1	10	9	64	99.99	<i>E. coli</i>	420
blc-45	2	11	27	1	99.99	<i>K. pneumoniae</i>	490
blc-46	3	12	36	2	99.98	<i>E. coli</i>	372
blc-47	4	1	0	0	100	<i>K. pneumoniae</i>	490
blc-48	5	2	0	0	100	<i>K. pneumoniae</i>	490
blc-49	6	3	0	0	99.99	<i>K. pneumoniae</i>	45
blc-50	7	4	0	0	100	<i>E. coli</i>	127
blc-51	8	5	0	0	100	<i>K. pneumoniae</i>	15
blc-52	9	6	0	1	99.99	<i>E. coli</i>	428
blc-53	10	7	0	0	100	<i>E. coli</i>	127
blc-54	11	8	1	5	99.99	<i>E. coli</i>	88
blc-55	12	9	3	1	99.99	<i>K. pneumoniae</i>	490

a final elution volume of 25 µl. For all other extractions for nanopore sequencing, the Qiagen Genomic tip 100/G kit (Qiagen) was used according to the manufacturer's protocol. DNA concentration was quantified using the Qubit 2.0 instrument (Life Technologies).

DNA extracts were multiplexed as 10 (library 1) or 12 (all other libraries) samples per flowcell using the ONT Rapid Barcoding kit (SQK-RBK004) according to the manufacturer's protocol. Sequencing was performed for 48 (library 1) and 24 h for all other libraries on a GridION using version FLO-MIN106 R9.4 flowcells. Flowcells were washed using the ONT Flowcell Wash kit (EXP-WSH003) and bias voltages were adjusted between runs according to the manufacturer's recommendations. One isolate on library 1 was excluded from all further analysis because the long- and short-read assemblies produced a different species identification, strongly suggesting a laboratory error.

Read pre-processing and assembly

We compared several filtering and demultiplexing approaches, particularly to try to reclaim 'unclassified reads' that might be important when using shorter sequencing times. Overall, using Guppy v3.1.5 (<https://community.nanoporetech.com>) for base-calling and demultiplexing followed by Deepbinner [8] (v0.2.0) to reassign reads binned as 'unclassified' by Guppy produced the most complete assemblies. We therefore adopted this approach for the rest of the analysis (see supplementary methods). The quality of ONT reads was assessed by kmer identity compared to Illumina reads using Filtlong [9]. Unicycler v0.4.8-beta was used to create hybrid assemblies utilizing both the long- and short-read data. We assessed both Unicycler's bold and normal '--mode' options (supplementary methods), and elected to use the bold mode results for analysis due to the

fact that they produced more complete assemblies and a structurally accurate assembly of the MGH78578 reference.

Long read-only assembly was performed using Flye (version 2.6) with the '--plasmids' option [10]. All assembly graphs were visualized using Bandage [11], which was also used to perform BLASTN searches. Isolates ($n=3$) with $<5\times$ estimated genome coverage were excluded from the long-read vs hybrid assembly comparison. All computation was performed on the Oxford University Biomedical Research Computing cluster, with eight threads used for each assembly. Deepbinner was run on a cluster of NVIDIA GeForce GTX 1080 Ti GPUs.

Assembly comparison

We compared assemblies created under different conditions using various different metrics.

- Completeness – the number of plasmids/chromosomes in each assembly marked as being circular by Unicycler.
- ALE – assembly likelihood estimator that estimates the likelihood of hybrid assemblies created using the same Illumina short-read sequencing data [12]. Short reads were mapped to hybrid assemblies using minimap2 [13].
- DNADiff – whole-genome alignment with calculations of gSNP and gIndel differences between assemblies [14]. gSNPs and gIndels represent high-confidence single-nucleotide polymorphisms (SNPs) and indels bounded between at least 20 exact nucleotide matches on both sides.

The relationship between the number of long-read bases and completeness (assessed as all structures marked as being circular by Unicycler) was estimated using a Wilcoxon rank sum test in R version 3.6. Minimap2 was used to map contigs from long-read to hybrid assemblies. ML plasmids [15] was used as a further arbitrator of the chromosomal/

plasmid origin of sequences. Human reads were detected using Centrifuge [16] as part of the Crumpit [17] pipeline. Simulations of shorter sequencing times were performed by selecting reads from fastq files produced between the beginning of the run and the simulated endpoint using a Python script (available at <https://github.com/samlipworth/ONT-wash-hybrid>).

Phases of laboratory evaluation

Three laboratory phases were performed (Fig. 1).

- (1) **Optimization of flowcell run time** (flowcell 1, library 1)
 - a. Ten isolates (one excluded from analysis, see above) sequenced for 48 h with comparison of assemblies created every 6 h (i.e. the first assembly used the first 6 h of data and the second the first 12 h etc.).
- (2) **Quantification of between-library contamination after using ONT wash kit** (flowcell 2, library 2)
 - a. Assessed by washing and then reusing a flowcell that had been used to sequence a clinical pathology sample for 24 h for an unrelated project. As library 2 contained 12 pure culture bacterial samples, no human DNA should have been detected if the wash kit was completely effective.
- (3) **Evaluation of the effect of between library contamination on subsequent hybrid assemblies** (flowcell 3, libraries 3–5)
 - a. Assessed by first sequencing 12 isolates for 24 h (library 3) and then washing the flowcell and resequencing the same 12 isolates with all barcodes switched (library 4, Table 1). We subsequently compared hybrid assemblies created using long-read data from libraries 3 and 4.
 - b. We then washed flowcell 3 for a second time and sequenced 12 different isolates for a further 24 h. We checked for between-library contamination by blasting contigs (BLASTN) from short-read to hybrid assemblies.

RESULTS

Optimization of sequencing run time

We ran the first flowcell with library 1 for 48 h (multiplexing 10 isolates of which 1 was excluded from analysis due to laboratory error). For the nine evaluable isolates, read length peaked at an N50 of 8774 base pairs (bp) after 7 h and subsequently decreased to a minimum of 7094 bp at 42 h. Median read quality score peaked at 5 h (74, IQR 49–85) and reduced

to a minimum of 56 (IQR 27–75) at 38 h. The rate of bases called for each barcode over time was very unequal (Fig. S1); at 24 h there was a median output of 447 Mb per barcode (range 131–863 Mb) and at 48 h there was a median of 552 Mb per barcode (range 158–1085 Mb).

To empirically estimate the optimum run time we compared hybrid assemblies produced at cumulative six-hourly intervals during the 48 h over which library 1 was sequenced. Maximum circularity was achieved by 24 h, by which point 6/9 assemblies (24/27 contigs) had fully circularized (Fig. 2); notably there was no further benefit gained from an additional 24 h of sequencing. By 24 h, 17/18 plasmids had circularized; 1 was composed of a single contig but not marked as circular by unicycler (which was also the case at 48 h). Comparison of the assembly of the reference strain (MGH78578 – barcode 1) at 12 h (the only time point at which it completely circularized) to the published sequence revealed the correct number of plasmids ($n=5$) and a high degree of genetic similarity (1 unaligned base, 99.97% average identity, 64 gSNPs and 31 gIndels).

The three assembly graphs of the three non-complete assemblies (barcodes 02, 04, 05; isolates blc-23, blc-24, blc-25) at 24 h were relatively simple (Fig. S2). There was no relationship between the number of long-read bases and the probability of hybrid assembly completion at 24 h ($P=0.17$), reinforcing the likely futility of longer sequencing times. We also compared the assemblies created at different time points using the ALE tool, which revealed a similar pattern of results: in two cases (bc04 and bc09, isolates blc-24 and blc-29), a more likely assembly vs that at 48 h was obtained after 24 h (Fig. 3). For the rest of this study we therefore elected to stop all sequencing runs at 24 h.

Evaluation of wash kit efficacy at removing human DNA

We first attempted to use the ONT wash kit on a flowcell that had previously been used to sequence a human clinical pathology sample for 24 h. This first 24 h of sequencing yielded 2059966 reads, of which 2028024 (98.4%) were binned as human by centrifuge. The flowcell was then washed and reloaded with library 2 (bacterial isolates only), which was sequenced for 24 h. After demultiplexing, 818091 reads (3942 Mb) were obtained, of which 147 (0.02%) were binned by centrifuge as being of human origin. The number of human reads was within the range of human reads called by centrifuge for all other flowcells (which had not sequenced any human DNA, Table S2), suggesting that this number is compatible with background noise from the kit-ome/false-positive binning.

Using this 24-hour-old recycled flowcell to sequence library 2, we acquired complete assemblies for 10/12 genomes. Barcode 1 failed, returning only 3.8×10^6 bases of data and barcode 10 yielded an assembly of four contigs comprising a chromosome and three plasmids, but one of the plasmids was marked as incomplete by Unicycler.

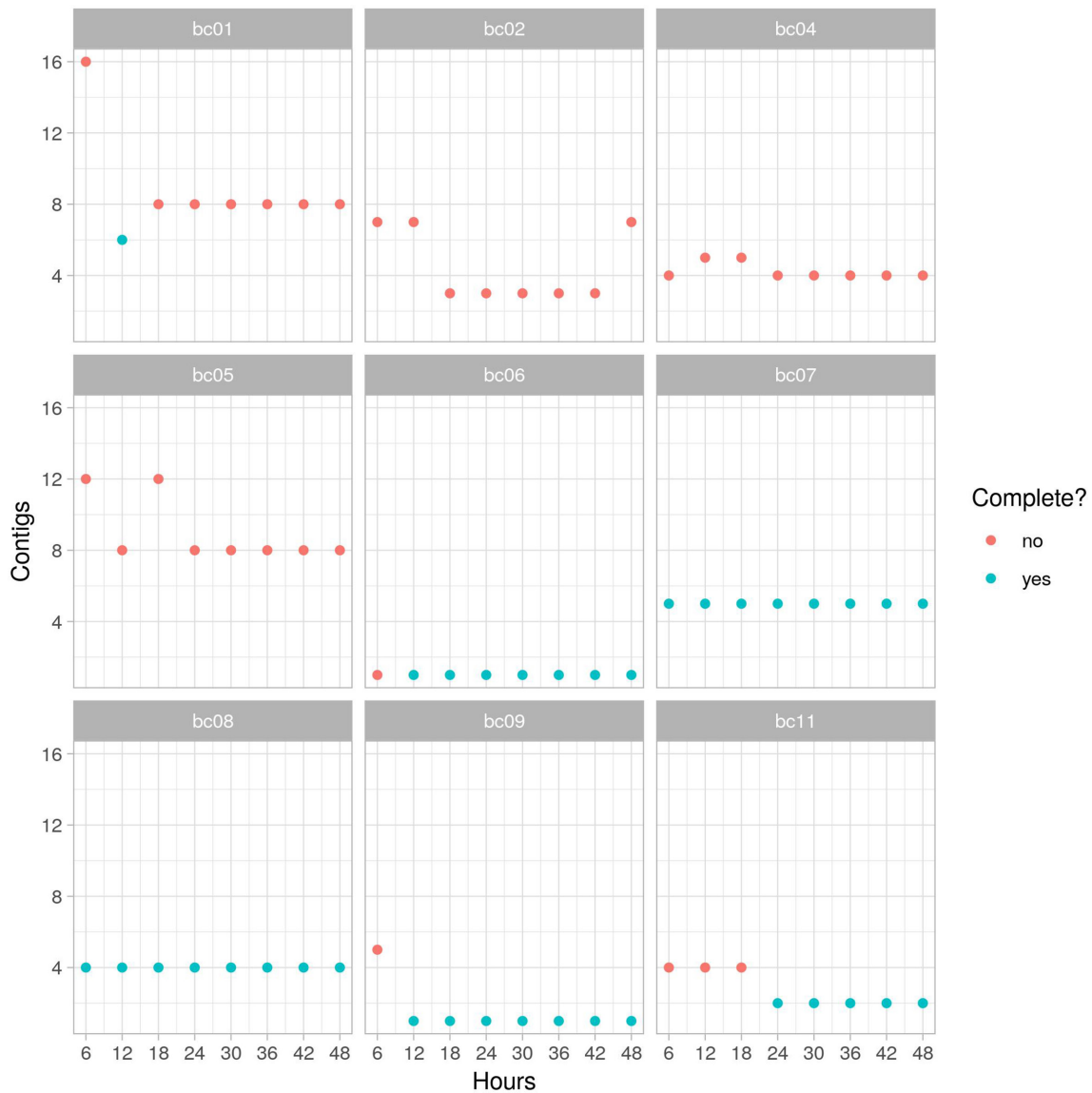


Fig. 2. Number of contigs generated by Unicycler for the nine included barcoded (bc) samples in library 1 over time (one isolate was excluded, see the Methods section). Complete assemblies (where the chromosome and all plasmids are formed of single circularized contigs) are shown in blue.

Reusing flowcells for similar isolates

Given the low contamination observed, we next sought to reuse a flowcell to sequence closely related *Enterobacteriaceae* using a single set of barcodes. After 24h, library 3 produced 8/12 fully complete assemblies, following which we resequenced the same isolates after changing the barcodes used, as shown in Table 1, and washing the flowcell between runs. Starting channel availability decreased by about 28% (~1400 to ~1000 at the beginning of library 3 vs 4, respectively; Fig. S3).

Completeness was identical for 11/12 isolates between the runs. There was, however, a major discrepancy in one sample, where a ~868 kb region was called as chromosomal in library

3 and a circularized super-plasmid-like component in library 4 (Fig. S4). As expected, ML plasmids [15] predicted with high confidence (97% probability) that this contig was of chromosomal origin. Interestingly, this error was fixed after filtering with FilTlong, suggesting that it may have arisen from low-quality reads.

For 6/12 isolates (blc-46, blc-48, blc-50, blc-51, blc-53, blc-55,) the ALE score suggested a better assembly for library 3. Comparing hybrid assemblies of the same isolates between libraries (i.e. library 3 vs library 4) using the DNAdiff tool revealed near-identical assemblies (identity $\geq 99.98\%$), and low numbers of SNPs and indels in all instances (Table 1). The worst performing assembly (blc-46) contained the major

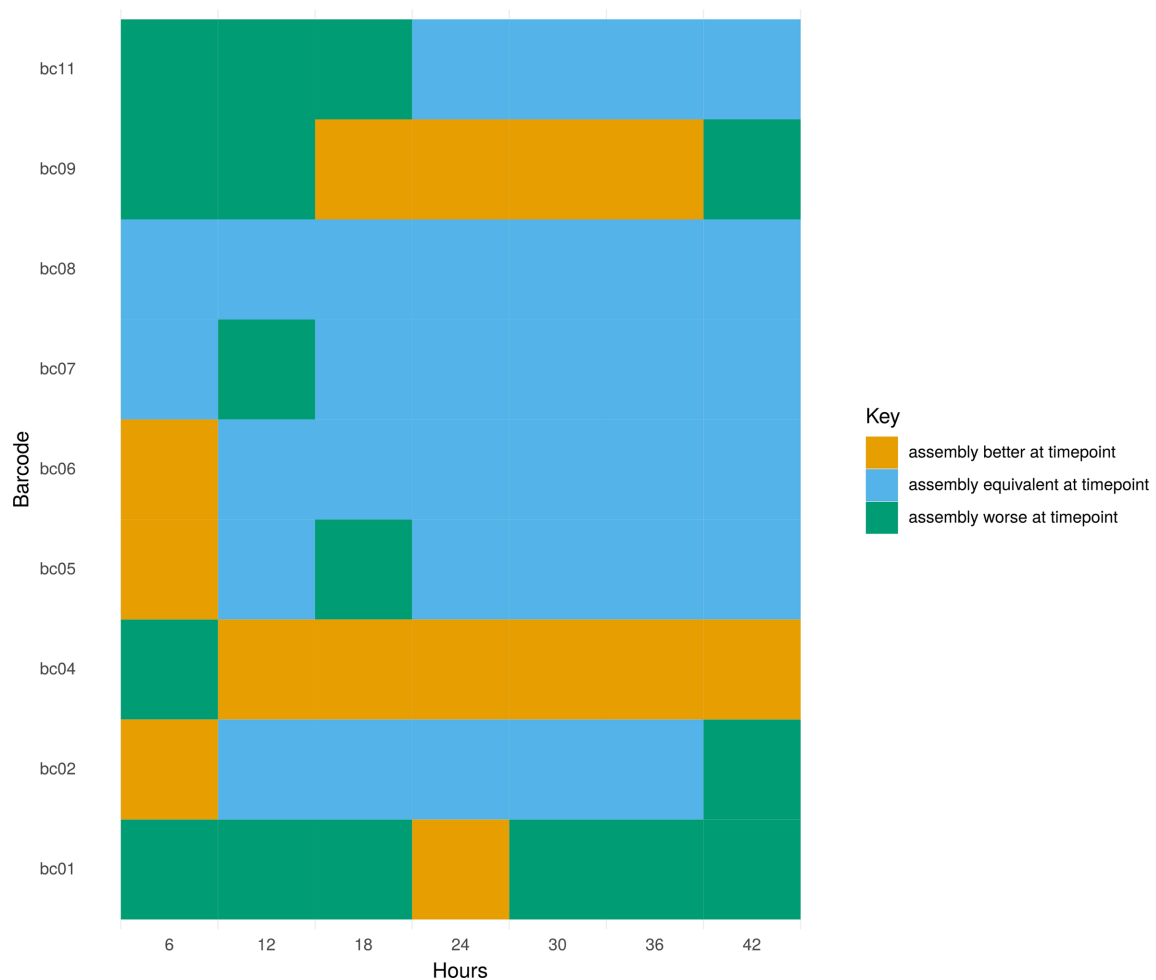


Fig. 3. Assembly likelihoods were calculated with the assembly likelihood estimator (ALE) by mapping Illumina reads to hybrid assemblies. Likelihoods were calculated for assemblies of each barcoded isolate (y-axis) created at 6 h intervals up to 48 h (x-axis). A likelihood difference of 0 (blue) implies that the assembly at time t is equally as likely as that at $t=48$ h. A positive likelihood difference (orange) implies that the assembly at time t was better than at 48 h and a negative likelihood difference (green) implies a worse assembly at time t vs 48 h.

structural disagreement discussed above, which likely caused the slightly higher number of SNPs between assemblies in this sample. We subsequently reloaded the same flowcell that had been used to sequence libraries 3 and 4 with library 5 (different isolates) and generated a further 7/12 completed assemblies.

A possible explanation for differences in assemblies between runs might be that, as demonstrated above, read length and quality deteriorate markedly over the course of a single run cycle, which may introduce false artificial variation. However, use of the ONT wash kit on flowcell 3 between libraries 3 and 4 restored median read quality scores almost to their original values [before=70 (IQR 10–84), after=69 (IQR 25–82)]. From 18–24 h of sequencing library 3, median read length was 1204 (IQR 169–4429). After using the wash kit, the next 6 h of sequencing of library 4 yielded median read length of 1439 (IQR 292–4554) (Fig. S5). A similar effect appeared to occur

after the flowcell was washed and reloaded with library 5, although quality scores and read lengths decayed more quickly on the third run. However, the results from library 5 were not directly comparable because they comprised sequencing data from different extractions and isolates.

Effect of washing on flowcell output

Nanopore has informally reported that the wash kit ‘at least doubles’ the output of a flowcell ‘in experiments where throughput is limited by the increase in pores in the ‘recovering’/‘unavailable’ state’ [7]. To independently investigate this, we compared the output of flowcell 1 (library 1) with the combined output of flowcell 3 (libraries 3, 4 and 5; Fig. 1). The output of flowcell 1 was 4.40 Gbp at 24 h; a further 24 h of sequencing yielded only an additional 1.01 Gbp (total 5.41 Gbp at 48 h). The first library to be sequenced on flowcell 3 (library 3) yielded 6.68 Gbp after 24 h. After washing and

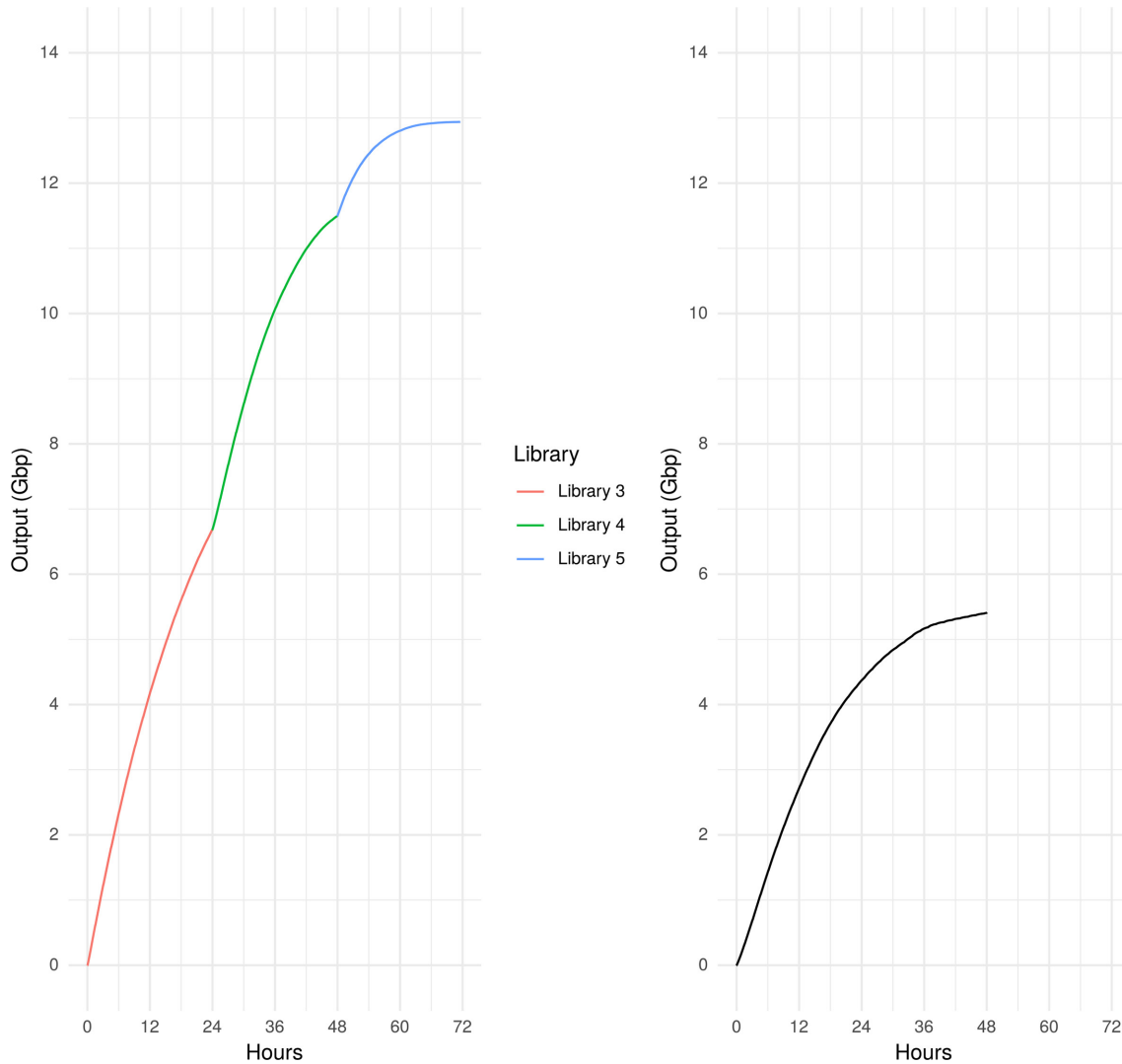


Fig. 4. Output over time for flowcell 3 (libraries 3/4/5, left) and flowcell 1 (library 1, right). The ONT wash kit was used between libraries 3/4/5 on flowcell 3, whereas flowcell 1 was run for 48 h with no washing steps.

reloading with library 4, a further 4.82 Gbp was acquired in the subsequent 24 h period (a total of 11.5 Gbp over the 48 h period; Fig. 4). Whilst flowcell 3 was better performing in the first 24 h and therefore not all the observed improvement can be attributed to washing, in this limited evaluation it did seem to greatly improve output.

Evaluating sequencing run times using all data

Combining data from libraries 1, 2, 3 and 5 (i.e. excluding library 4 as the isolates were the same as in library 3), we simulated shorter sequencing times by assembling reads produced 3, 6 and 12 h after the start of the run. At 24 h, 30/45 isolates had complete assemblies (chromosome and all plasmids circular) compared to 29/45 at 12 h, 24/45 at 6 h and 21/45 at 3 h. The number of incomplete plasmids was similar across all time points (8/150 at 24 h, 6/150 at 12 h, 9/150 at 6 h and 9/150 at 3 h).

Comparison with long-read assembly

Finally, we compared hybrid assemblies to those generated using only long reads to assess whether generating Illumina reads is still likely to be necessary for future studies. Overall, long read-only assemblies had a high average identity to the reference hybrid assemblies (Table 2). When created using data demultiplexed by Guppy alone, however, most of the long read-only assemblies contained contigs that did not map to the hybrid assemblies (median number 3, range 0–15, median length 3780 bp, range 545–19197 bp, median coverage 24, range 4–995) (Fig. S6). This was true both for libraries sequenced on new flowcells and those that had been reused after washing, but not for library 1. Using *BLASTN* (in *Bandage*) we were able to identify that some of these likely represented between barcode contamination from isolates sequenced in the same library (Fig. S7).

Table 2. DNAdiff comparisons between Flye long read-only assemblies and hybrid assemblies

Library	Isolate name	Barcode	SNPs	Indels	% reference bases aligned	% identity
1	blc-22	bc01	3385	14679	99.95	99.55
1	blc-23	bc02	251	9953	100	99.76
1	blc-24	bc04	3202	15949	99.98	99.58
1	blc-25	bc05	3387	12262	100	99.53
1	blc-26	bc06	3082	11494	99.96	99.53
1	blc-27	bc07	3203	11764	99.98	99.56
1	blc-28	bc08	157	12793	99.97	99.61
1	blc-29	bc09	3309	12036	100	99.39
1	blc-31	bc11	4070	13861	99.98	99.51
2	blc-33	bc02	3179	15572	99.97	99.48
2	blc-34	bc03	4942	16498	100	99.34
2	blc-35	bc04	5488	16499	99.9	99.36
2	blc-36	bc05	5458	16842	100	99.35
2	blc-37	bc06	5442	16499	99.98	99.36
2	blc-38	bc07	5460	16559	100	99.36
2	blc-39	bc08	4311	20783	97.82	98.45
2	blc-40	bc09	5428	16882	99.83	99.36
2	blc-41	bc10	3064	15975	99.93	99.37
2	blc-42	bc11	2960	15133	100	99.44
2	blc-43	bc12	5434	16748	100	99.35
3	blc-47	bc01	4482	19204	99.98	99.31
3	blc-48	bc02	4352	19444	99.99	99.3
3	blc-49	bc03	3983	18206	99.98	99.3
3	blc-50	bc04	2615	17537	100	99.4
3	blc-51	bc05	3728	19560	100	99.36
3	blc-52	bc06	2645	16326	100	99.4
3	blc-53	bc07	2645	17567	100	99.4
3	blc-48	bc08	2580	17145	99.95	99.39
3	blc-49	bc09	4326	19080	99.95	99.31
3	blc-50	bc10	2607	15458	100	99.4
3	blc-51	bc11	4406	19273	99.99	99.31
3	blc-52	bc12	2555	17336	100	99.4
5	blc-56	bc01	4746	19881	99.89	98.96
5	blc-57	bc02	4174	20993	99.46	98.56
5	blc-58	bc03	4876	20019	99.92	99.01
5	blc-59	bc04	4649	21014	94.41	97.93
5	blc-61	bc06	3115	14943	100	99.45
5	blc-62	bc07	4709	15872	100	99.33

Continued

Table 2. Continued

Library	Isolate name	Barcode	SNPs	Indels	% reference bases aligned	% identity
5	blc-63	bc08	4680	15201	99.98	99.34
5	blc-64	bc09	3051	15973	100	99.39
5	blc-65	bc10	4212	18101	99.36	99.22
5	blc-66	bc12	4986	21189	98	98.61

To try to correct this, we created further assemblies using only reads where both Deepbinner and Guppy agreed on the barcode assignment. Whilst this greatly improved the assemblies and most (but not all) spurious contigs were removed (Fig. S6), structural differences compared to the hybrid references remained in several assemblies (Fig. S8). We hypothesized that this might be an issue with rapid barcoding, but saw the same signal in data multiplexed with the native barcoding kit in a recent study (Fig. S9) [6].

DISCUSSION

In this study we have demonstrated that, for the purposes of creating ONT reads from pure isolates for hybrid assembly, there is unlikely to be benefit in extending sequencing runs beyond 24 h; indeed, for most assemblies, 12 h is likely to be sufficient. We have also shown that after utilizing the ONT flowcell wash kit, between-library contamination is minimal, and is unlikely to have an important effect on subsequent hybrid assemblies. This appears to be true even when the same barcodes are used for successive libraries. Despite significantly shortened run times and reusing flowcells, we were able to completely assemble the vast majority of plasmids. This marks a significant milestone for ONT sequencing for the purposes of hybrid assembly and unlocks the potential for large-scale studies of plasmid epidemiology in the near future. We have additionally independently replicated the very limited data available on ONT's website, showing that flowcell washing can dramatically increase output, a finding relevant to all investigators performing Nanopore sequencing.

Previous studies have demonstrated successful completion of 12 genomes on a single flowcell using molecular barcoding; here we have demonstrated that this can be increased to at least 22. Based on ONT's quoted figures of £540 (GBP) per flowcell (when buying 24) and £86.70 (£520/6) for library preparation and barcoding, the current per-sample cost for the long-read sequencing component is approximately £52.20 (i.e. £540 for flowcell+£520/6 for library preparation/12). In the most conservative interpretation of this study, we have shown an approximately 27% per-sample reduction in cost to £38.20 [i.e. £540 for flowcell+3×(£520/6) for library preparation+3×(£80/6) for the wash kit/22], assuming downstream analysis demanded complete circularization of all contigs. We envisage that for most current use cases, however, particularly plasmid genomics, the standard of data produced in the majority of

our assemblies would be sufficient to answer the biological questions posed. Even with ultra-short run times of 12 h (1/6th of the total run time that is currently standard in our laboratory and others) we were able to circularize the vast majority of plasmids (and most chromosomes). Combining washing steps with a recently published chemistry-free approach to demultiplexing may provide additional savings and would seem worth exploring in future work [18]. Whilst we had access to a large university computer cluster, this will not be the case for all investigators, though we estimate that if one were to use cloud computing (e.g. Amazon AWS), the additional cost of this is small (e.g. 2 h on a 1.4× large=£0.31) [19].

If, as seems plausible from our data, 12 h is a viable run time for most research questions and we assume a useful period of 72 h per flowcell, then long-read sequencing costs would be further reduced to approximately £15.80 per isolate (i.e. £540 for a flowcell+6×£520/6 for library prep+£80 for a wash kit/72 isolates; a ~70% reduction on current costs). Investigators would, of course, also have to factor in the cost of Illumina sequencing (around £38/isolate at our institution). The calculations above might be limited by the effect of repeated washing of the flowcell and deterioration of pores over time; however, even in library 5 in our study (which used a 48 h-old flowcell that had been washed twice), 8/12 chromosomes and 33/36 plasmids were complete at 12 h. We envisage that after stopping runs at 12 or 24 h, investigators would be able to carefully select the few isolates that require further sequencing and avoid wasting valuable pore time where complete assemblies have already been acquired. We would caution, however, that, in this study, increasing run times did not usually lead to improved assemblies. This is consistent with recent data from a different study in our laboratory that demonstrated that in some cases random subsampling of reads can even improve assemblies [6]. The reason for this is unclear and requires further investigation but could plausibly be a result of differences between real sequencing data (as used in this study and a previous one [6]) and simulated data, which are often used extensively in the development of assembly tools. A better understanding of these differences and how they affect the metrics that assemblers use might lead to improvements in these tools.

Ideally, one would want to use a unique set of barcodes for each library run on a single flowcell. At present, however, there are only 12 barcodes available in ONT's rapid barcoding

kit, which has a substantially easier and less time-consuming protocol compared to the Native Barcoding kit (for which 24 barcodes are available). Nevertheless, the number of SNPs and indels between alignments of the isolates sequenced in different libraries on the same flowcell (using the same set of barcodes but reassigned to different isolates) was similar to that seen when comparing Illumina/ONT and Illumina/PacBio assemblies of a single isolate [6]. Our assembly of the MGH78578 reference diverged by a similar number of SNPs compared to the published sequence and that in a recent study [6]. To our knowledge, there are limited data available on the variation produced by successive cycles of culturing, DNA extraction and sequencing of the same isolate using ONT technology and further investigation of this using reference sequences seems warranted. Based on our data, using the same barcodes for consecutive libraries on the same flowcell is likely to be acceptable when generating long reads for hybrid assembly.

Multiplexed ONT sequencing holds the promise of allowing complete and accurate genomes to be obtained from a single platform. Our results suggest that both *in silico* demultiplexing and laboratory kits need to improve before this is a reliable alternative to hybrid sequencing. Such development will be critical to ensuring the viability of ONT sequencing, particularly in routine clinical settings in the future. R10 nanopore chemistry may enable improvements to the results we report here and potentially dispense with the requirement for Illumina data, offering complete and accurate genomes on a single platform, although this will require independent validation. It has previously been hypothesized that the bimodal distribution observed in quality scores of reads (as calculated by their kmer identity to Illumina reads) delineates 'good' from 'junk' reads [9]. We speculate that in fact ONT reads with low identity to Illumina reads represent cross-barcode contamination. It is possible that between-library contamination also contributed to low-quality reads in libraries 2/4/5, although the distribution of read quality scores in these was comparable to that seen on brand-new flowcells (Fig. S5). The long-read assembly problem is somewhat improved by consensus demultiplexing using two tools, but this is resource-intensive, increases reads binned as 'unclassified' and is still not completely reliable. Hybrid assemblies are much less vulnerable to cross-barcode contamination, which appears to be effectively removed by Unicycler's process of mapping long reads to the short-read assembly. Whilst reasonably high-quality long read-only assemblies can be achieved by running a single isolate per flowcell with subsequent polishing steps, the cost of this would currently be significantly higher than that for hybrid sequencing.

Different demultiplexing, filtering and assembly parameters can produce different assemblies from the same input data. Whilst our assembly of the MGH757878 reference was very similar to the published sequence, further benchmarking of the effect of using different parameters is required, although this is beyond the scope of this project. We only included a single *K. pneumoniae* reference strain, meaning that the ground truth for most assemblies we performed

was unknown, although notably, in our first library the overall structures did not change with an additional 24 h of sequencing. An additional limitation is that we used a different extraction method for library 1 compared to all other libraries; however, the similar results obtained also demonstrate that fully automated DNA extraction could be deployed to facilitate high-throughput hybrid sequencing workflows.

In conclusion, we have demonstrated that high-quality hybrid assemblies can be generated with much shorter sequencing times than are currently standard. The new ONT wash kit appears to be highly effective, even to the point where reuse of the same barcodes on a flowcell seems acceptable when acquiring long reads for hybrid assemblies. Reusing flowcells for multiple libraries produces substantial potential per-isolate cost reductions. Ultimately the opportunity to take advantage of this and conduct large-scale studies incorporating hybrid assembly is likely to help better inform future efforts to tackle some of the most important human pathogens.

Funding information

The research was supported by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (HPRU 2012–10041) at the University of Oxford in partnership with Public Health England (PHE) and by Oxford NIHR Biomedical Research Centre. T. E. A. P., A. S. W. and D. W. C. are NIHR Senior Investigators. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The report presents independent research funded by NIHR. The views expressed in this publication are those of the authors and not necessarily those of the NHS, NIHR, the Department of Health or Public Health England. K. K. C. is funded by the Medical Research Foundation. S. L. is supported by a Medical Research Council Clinical Research Training Fellowship.

Acknowledgements

The authors thank Anthony Brown for his assistance performing DNA extractions.

Author contributions

Conceptualization: S. L., N. S. t., A. S. W.; methodology: S. L., N. S.; software: J. S., N. Sa.; formal analysis: S. L.; investigation: S. L.; resources: A. S. W., D. C., T. P., M. A., M. M.; data curation: H. P., K. C., L. B., D. G., S. L., J. K., M. M.; writing - original draft preparation: S. L.; writing - review and editing: All authors; supervision: D. C., T. P., A. S. W., N. S.; funding: S. L., N. S., A. S. W., D. C., T. P.

Conflicts of interest

The authors declare that there are no conflicts of interest

References

1. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B *et al*. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J Antimicrob Chemother* 2018;73:1121–1137.
2. George S, Pankhurst L, Hubbard A, Votintseva A, Stoesser N *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.
3. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res* 2019;8:2138.

4. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* 2017;3:e000132.
5. 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.
6. 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.
7. Flow Cell Wash Kit: Nanopore Store. Oxford Nanopore Technologies. <https://store.nanoporetech.com/flow-cell-wash-kit-r9.html> (accessed 27 January 2020).
8. Wick RR, Judd LM, Holt KE. Deepbinner: demultiplexing barcoded Oxford nanopore reads with deep convolutional neural networks. *PLoS Comput Biol* 2018;14:e1006583.
9. Wick R. Filtlong documentation.
10. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;37:540–546.
11. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics* 2015;31:3350–3352.
12. Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* 2013;29:435–443.
13. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.
14. 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.
15. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J et al. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb Genom* 2018;4.
16. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016;26:1721–1729.
17. Sanderson ND, Street TL, Foster D, Swann J, Atkins BL et al. Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. *BMC Genomics* 2018;19:714.
18. Diltthey AT, Meyer SA, Kaasch AJ. Ultrplexing: increasing the efficiency of long-read sequencing for hybrid assembly with k-mer-based multiplexing. *Genome Biol* 2020;21:68.
19. Amazon EC2 On-Demand Pricing. <https://aws.amazon.com/ec2/pricing/on-demand/> (accessed 8 December 2020).

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

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

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