RESEARCH

A high-throughput multiplexing and selection strategy to complete bacterial genomes

Sergio Arredondo-Alonso [1,2,*], Anna K. Pöntinen [1], François Cléon [3], Rebecca A. Gladstone [1], Anita C. Schürch [4], Pål J. Johnsen [3], Ørjan Samuelsen [3,5] and Jukka Corander [1,2,6]

[1]Department of Biostatistics, University of Oslo, 0317, Oslo, Norway; [2]Parasites and Microbes, Wellcome Sanger Institute, Cambridgeshire CB10 1RQ, UK; [3]Department of Pharmacy, Faculty of Health Sciences, UiT The Arctic University of Norway, 9037, Tromsø, Norway; [4]Department of Medical Microbiology, UMC Utrecht, 3584 CX, Utrecht, the Netherlands; [5]Norwegian National Advisory Unit on Detection of Antimicrobial Resistance, Department of Microbiology and Infection Control, University Hospital of North Norway, 9038, Tromsø, Norway and [6]Department of Mathematics and Statistics, Helsinki Institute of Information Technology (HIIT), FI-00014 University of Helsinki, 02130, Espoo, Helsinki, Finland

*Correspondence address. Sergio Arredondo-Alonso. P.O.Box 1122 Blindern 0317 Oslo, Norway.
e-mail: s.a.alonso@medisin.uio.no https://orcid.org/0000-0002-4445-2442

Abstract

Background: Bacterial whole-genome sequencing based on short-read technologies often results in a draft assembly formed by contiguous sequences. The introduction of long-read sequencing technologies permits those contiguous sequences to be unambiguously bridged into complete genomes. However, the elevated costs associated with long-read sequencing frequently limit the number of bacterial isolates that can be long-read sequenced. Here we evaluated the recently released 96 barcoding kit from Oxford Nanopore Technologies (ONT) to generate complete genomes on a high-throughput basis. In addition, we propose an isolate selection strategy that optimizes a representative selection of isolates for long-read sequencing considering as input large-scale bacterial collections. **Results:** Despite an uneven distribution of long reads per barcode, near-complete chromosomal sequences (assembly contiguity = 0.89) were generated for 96 *Escherichia coli* isolates with associated short-read sequencing data. The assembly contiguity of the plasmid replicons was even higher (0.98), which indicated the suitability of the multiplexing strategy for studies focused on resolving plasmid sequences. We benchmarked hybrid and ONT-only assemblies and showed that the combination of ONT sequencing data with short-read sequencing data is still highly desirable (i) to perform an unbiased selection of isolates for long-read sequencing, (ii) to achieve an optimal genome accuracy and completeness, and (iii) to include small plasmids underrepresented in the ONT library. **Conclusions:** The proposed long-read isolate selection ensures the completion of bacterial genomes that span the genome diversity inherent in large collections of bacterial isolates. We show the potential of using this multiplexing approach to close bacterial genomes on a high-throughput basis.

Introduction

Whole-genome sequencing (WGS) of bacterial isolates has dramatically increased its routine presence in clinical genomics and epidemiological investigations [1]. The possibility of using short-read technologies, affordable and generally accurate in terms of their sequencing reads, has permitted tracking the presence of

particular sequencing types, assessing the presence of single-point mutations, or identifying antimicrobial resistance (AMR) genes in collections of thousands of bacterial isolates [2–4]. These applications are fundamental during outbreak detections and investigations, for which WGS has overcome some limitations associated with classical epidemiological techniques [2, 5, 6]. However, the read length associated with these short-read technologies (from 150 to 300 bp) cannot unambiguously span the presence of repeat elements such as insertion sequences (ISs). This results in a fragmented assembly, typically formed by contiguous sequences (contigs) of unknown order. In particular, determining the genome context (chromosome, plasmid, phage) of genes typically surrounded by IS elements (e.g., AMR genes) is challenging in draft assemblies [7, 8].

Long-read sequencing technologies such as Oxford Nanopore Technologies (ONT) can generate genomic libraries with a mean read length of 10–30 kb [9], but with a lower associated raw read accuracy of ~97% (phred score ~15) depending on the pore chemistry [10]. These long reads can typically span repeat elements in a bacterial genome, producing a contiguous assembly consisting of single and circular contigs per replicon (chromosome and/or plasmids) [11, 12]. However, the sequence accuracy of these complete genomes can be hindered by incorrect base-calling of short homopolymer sequences, resulting in early termination of open reading frames (ORFs) in protein-coding genes [13]. This limitation can be mitigated by creating an improved consensus sequence and variant calls, reusing the ONT sequencing data with tools such as Nanopolish [11] or Medaka [14].

An attractive alternative is to combine short- and long-read technologies to generate complete and accurate genomes in a process called hybrid assembly [15]. A frequent scenario encountered by researchers is that large collections of bacterial isolates have been massively short-read sequenced and only a subset of these isolates can be further selected for long-read sequencing. Therefore, selecting isolates for long-read sequencing is a non-trivial and crucial step that can affect subsequent analyses based on the resulting complete genomes.

There are different strategies to decrease the price of completing a genome with ONT sequencing [16, 17]. Recently, Lipworth et al. [18] showed that use of wash-kits coupled with shorter sequencing times can be optimized to obtain 36 complete genomes per single flow cell, achieving a reduction in long-read sequencing costs of 27%. During the hybrid assembly process, only a fraction of the total number of long reads generated are required to bridge and span the initial short-read assembly graph and thus a low ONT coverage is sufficient to complete a genome [19, 20].

The possibility of increasing the number of multiplexed isolates per flow cell is an alternative to the use of wash-kits that reduces hands-on time and avoids contamination issues between libraries. Recently, ONT has released a new barcoding kit allowing 96 isolates to be multiplexed in the same sequencing library.

Here, we evaluated the degree of completeness and accuracy in bacterial genomes generated using the new ONT barcoding kit for 96 bacterial isolates. In addition, we provide a step-by-step computational workflow to perform an unbiased selection of isolates for long-read sequencing based on the presence/absence of genes, as previously applied in 2 large bacterial collections of isolates [21, 22].

The approach described in this study generated near-complete genomes for the majority of 96 *Escherichia coli* isolates. The computational workflow proposed can be used to maximize and complete a representative selection of isolates from large-scale bacterial collections.

Methods
Isolate selection based on existing short-read assemblies

From a large collection of short-read isolates, frequently only a subset of isolates can be completed with long-read sequencing. We propose the following approach to select a representative subset of isolates for a population sample:

1. Define M as the presence/absence matrix of orthologous genes created by pangenome tools such as Roary (Roary, RRID:SCR_018172) [23] or Panaroo (Panaroo, RRID:SCR_021090) [24]. M is a binary matrix with $s \times g$ dimensions, in which s is the total number of isolates (samples) present in the collection and g corresponds to the total number of orthologous genes predicted.

$$M_{s,g} = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,g} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,g} \\ \vdots & \vdots & \ddots & \vdots \\ m_{s,1} & m_{s,2} & \cdots & m_{s,g} \end{pmatrix}$$

2. We transform M into a Jaccard distance matrix D with $s \times s$ dimensions using the R function *parDist* (method = "fJaccard") provided in the R package parallelDist (version 0.2.4) [25, 26].

$$D_{s,s} = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,s} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,s} \\ \vdots & \vdots & \ddots & \vdots \\ d_{s,1} & d_{s,2} & \cdots & d_{s,s} \end{pmatrix}$$

For instance, the element $d_{1,2}$ can be defined as the similarity distance between the genes predicted for the first isolate and second isolate:

$$M_{1*} = \{m_{1,1}, m_{1,2}, ..., m_{1,g}\}$$

$$M_{2*} = \{m_{2,1}, m_{2,2}, ..., m_{2,g}\}$$

$$d_{1,2} = \frac{|M_{1*} \cap M_{2*}|}{|M_{1*} \cup M_{2*}|}$$

3. Next, the distance matrix D can be dimensionally reduced using the t-distributed stochastic neighbour embedding algorithm (t-sne) [27], which results in a new matrix T with only 2 dimensions. In this step, we use the R package Rtsne (version 0.15) [28] considering as default a perplexity value of 30.

4. Next, we use the k-means algorithm [29] with T as input data and define a number of centroids L such that it corresponds to the desired number of isolates to be long-read sequenced, whereby k-means assigns each point t_i in the t-sne map cor-

responding to an isolate i to its closest centroid. For each cluster C_k, the algorithm iteratively updates the position of the centroid by computing the average Euclidean distance of each point t_i assigned to C_k to the mean value μ_k of all points assigned to the same cluster. The within-square variation V for a particular cluster C_k can be defined as:

$$V(C_k) = \sum_{t_i \in C_k} (t_i - \mu_k)^2.$$

After a random initialization of the centroid positions, the algorithm iterates between updating centroid positions and cluster memberships for all points until convergence is reached, and thus the positions of the centroids in T no longer vary or the maximum number of iterations is reached (default = 1,000 iterations).

To run the algorithm, we use the function "kmeans" provided in the R package stats (version 3.6.3). For a fixed number of centroids L, we run the k-means algorithm using 10,000 distinct initializations and select the outcome with the highest ratio between-cluster sum of squares to the total sum of squares (between_ss/tot_ss).

$$\text{tot.within\_ss} = \sum_{L=1}^{L} V(C_k)$$

$$\text{tot\_ss} = \frac{1}{N} \sum_{i,i' \in \{1,2,\dots N\}} (t_i - t_{i'})^2$$

$$\text{between\_ss} = \text{tot\_ss} - \text{tot.within\_ss}$$

N corresponds to the total number of points t_i present in T.

The ratio (between_ss/tot_ss) is considered as the percent of total variance from T explained by the chosen number of centroids (clusters). The relationship between the percent of total variance explained and the number of centroids can be considered to visually determine the ideal numbers of clusters required to capture the diversity present in the collection (elbow method).

5. For each cluster C_k, we select for long-read sequencing the isolate i with the lowest Euclidean distance with respect to its associated centroid c_k.

$$\text{long read } i = \min_{t_i \in C_k} t_i - c_k$$

6. The t-sne matrix T together with the final coordinates of the centroids c and of the isolates i selected for long-read sequencing are plotted using the R package ggplot2 (version 3.3.3).

The proposed workflow aims at capturing a comprehensive representation of the gene content variation across the lineages in the target population and is available at [30] as a Snakemake pipeline [31] that only requires (i) a presence/absence matrix in the same format as created by Roary/Panaroo and (ii) the desired number of long-read isolates.

Collection of Illumina short-read assemblies

To showcase the proposed long-read selection, we considered the Norwegian *Escherichia coli* collection of 3,254 isolates caus-

ing bloodstream infections recently described by Gladstone et al. [32]. The DNA extraction of this collection was performed using the DNeasy 96 Blood and Tissue kit (Qiagen, Hilden, Germany), the isolates were short-read sequenced with the Illumina HiSeq platform, and short-read contigs were created using VelvetOptimiser (version 2.2.5) [33] and Velvet (version 1.2.10) [34]. Furthermore, an assembly improvement step was applied to the assembly with the best N50 and contigs were scaffolded using SSPACE version 2.0 (SSPACE, RRID:SCR_005056) [35] and sequence gaps filled using GapFiller (version 1.11) [36]. These short-read contigs will be later considered to compare the accuracy of the hybrid and ONT-only assemblies (section Genome accuracy and completeness).

The presence/absence of orthologous genes defined by Panaroo (version 1.0.2) [24] and PopPUNK lineages (version 2.0.2) [37] associated with the isolates were also extracted from Gladstone et al. [32] and considered as input for the long-read selection process. As an example, we fixed the number of centroids to 96 and considered the between_ss/tot_ss ratio to estimate whether the maximum number of isolates that can be multiplexed in an ONT sequencer would suffice to capture the genomic diversity of this particular collection.

We used the selection procedure described above on the *E. coli* collection of 3,254 isolates for the following 2 purposes: (i) to showcase the percent of variance (ratio between_ss/tot_ss) recovered by the pipeline with an arbitrary number of isolates (n = 96) and (ii) to select a large number of isolates (1,085) that are planned to be sequenced with ONT in future.

Of these 1,085 selected isolates, we explored the recently released ONT SQK-NBD110-96 barcoding kit to sequence 96 of these bacterial isolates. This run was crucial to define whether the rest of the isolates (n = 989) could be completed using this multiplexing kit or whether a different strategy would be preferable.

DNA isolation for ONT libraries

In total, 96 *E. coli* isolates, originally from clinical samples of human bloodstream infections, were grown on MacConkey agar No. 3 (Oxoid Ltd., Thermo Fisher Scientific Inc., Waltham, MA, USA) at 37°C, and individual colonies were picked for overnight growth in LB (Miller) broth (BD, Franklin Lakes, NJ, USA) at 37°C in 700 rpm shaking. High-molecular-weight (HMW) genomic DNA from cell pellets of 1.6 mL overnight cultures was extracted using MagAttract® HMW DNA Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions and using a final elution volume of 100 μL. DNA concentration and integrity were verified using NanoDrop One spectrophotometer (Thermo Scientific) and the Qubit dsDNA HS assay kit (Thermo Fisher Scientific) on a CLARIOstar microplate reader (BMG Labtech, Ortenberg, Germany).

ONT library preparation

Prior to library preparation, the samples were adjusted to 400 ng. The ONT library was prepared using SQK-NBD110-96 barcoding kit, and 40 fmol was loaded onto flow cells. Sequencing was run for 72 hours on GridION using FLO-MIN106 flow cells and MinKNOW v.20.10.6 software. Base-calling was conducted with the high-accuracy base-calling model and demultiplexing was formed by Guppy (version 4.2.3).

Hybrid assemblies

Porechop (version 0.2.4) [38] was used to trim and remove ONT adapters with default parameters. Filtlong (version 0.2.0) [39] was used to filter the ONT reads considering a minimum length of 1 kb (–min_length 1000), a weight given to the mean quality score of 20 (–mean_q_weight 20), retaining 90% of the total number of ONT reads (–keep_percent 90) from a maximum 40× coverage (–target_bases). Unicycler (version 0.4.7) [19] was run using the normal mode to perform a hybrid assembly with the Illumina trimmed reads and ONT reads retained after Filtlong.

We extracted the number of segments (contigs), links (edges), N50, and size of the components present at the resulting hybrid assembly graph. For each component, we defined its contiguity as follows:

$$\text{Contiguity} = \frac{\text{N50 Component}}{\text{Component Size}}$$

Mlplasmids (version 1.0.0) [40] was used with the *E. coli* model to confirm the origin (plasmid- or chromosome-derived) from the longest segment of each component.

Illumina and ONT depth per replicon

Illumina reads were mapped against the genome assembled by Unicycler, using Bowtie2 (version 2.4.2) [41] with the argument –very-sensitive-local, with a minimum and maximum fragment length of 0 (-I 0) and 2,000, respectively (-X 2000). The ONT reads were mapped against Unicycler assemblies using bwa mem (version 0.7.17) [42] and indicating the read type (-x) as ont2d. Samtools (version 1.12) [43] was used with the commands sort, index, and depth to process the alignments and retrieve the number of reads covering an individual nucleotide.

For each component present in the graph file given by Unicycler, we considered its longest contig and computed the mean number of reads. To normalize the contig depth with respect to its chromosome, we divided the mean depth of the contig against the mean depth of the longest contig assembled by Unicycler (longest chromosome segment).

ONT-only assemblies and long-read polishing

Flye (version 2.8.3-b1695) [44] was run with all the QC-passed ONT reads (phred score >7) available per barcode (–nano-raw), specifying the option to recover plasmids (–plasmids), indicating an expected genome size of 5 Mb (–genome-size 5m) and considering 3 polishing rounds (–iterations 3). For each component present in the assembly graph, we extracted the number of segments (contigs), links (edges), N50, and component size and defined its component contiguity. Mlplasmids (version 1.0.0) [40] was used with the *E. coli* model to confirm the origin (plasmid or chromosome derived) from the longest segment of each component.

Following the recommendations to further polish Flye assemblies [45], we used Medaka (version 1.2.5) [46] with the model "r941_min_high_g360" and considered as input all the passed ONT reads available per barcode. In addition, we performed a consecutive round of Medaka to observe whether additional rounds of polishing could improve the resulting genomes.

Genome accuracy and completeness

Quast (version 5.0.2) [47] was used to compute alignments with a minimum size of 5 kb (–min-alignment 5000) considering as reference short-read contigs assembled with Velvet (section Collection of Illumina short-read assemblies) from the same isolate against Unicycler, Flye, and Medaka-polished contigs. For each reference contig, we kept the best alignment (–ambiguity-usage one). The mean numbers of single-nucleotide polymorphisms (SNPs) and insertions and deletions (indels) per 100 kb were considered as a reference-based metric of the genome accuracy shown by the hybrid and ONT-only assemblies.

For Unicycler assemblies, we also extracted the contigs present in the file "001_best_spades_graph.gfa" corresponding to an optimal SPAdes assembly graph. Unicycler uses a range of *k*-mer sizes and computes a score between the resulting number of contigs and dead ends to choose the SPAdes graph ideal for the long-read bridge process. These SPAdes contigs were also compared using Quast against the reference contigs assembled by Velvet. This was relevant to determine the SNPs and indels per 100 kb present at the SPAdes contigs. In this manner, we could assess whether the accuracy (SNPs, indels) estimated for the hybrid assemblies was influenced by (i) differences between the short-read assemblers and/or (ii) regions of the genome from which its sequence is determined by the ONT reads, such as regions connecting dead ends in the SPAdes graph.

The ideel test [48] was used to obtain the number of early terminated ORFs in the assemblies [13]. Diamond (version 2.0.8) [49] was run with the blastp algorithm [50], specifying only a single target sequence per alignment (–max-target-seqs 1), a block size of 12 (-b12), and index chunk of 1 (-c1) against an index of the UniProt TREMBL database (retrieved in April 2021). For each hit, the ratio between query length and its own length was considered to assess the presence of interrupted ORFs (ratio <0.9). This assesses whether the length of the predicted proteins is shorter than their closest hit, most likely caused by the introduction of a stop codon. Importantly, the total number of interrupted ORFs may include true pseudogenes; however, most of these hits are considered as non-true errors introduced owing to a low sequencing accuracy. This ideel test provided a reference-free approach to assess the number of indels present in the final assemblies.

BUSCO version 5.1.2 (BUSCO, RRID:SCR_015008) [51] with the genome mode (-m genome) and the lineage dataset (-l) "enterobacterales_odb10" was used to assess the presence of 440 single-copy orthologous genes. BUSCO uses the following notation to annotate the genes as complete (length within 2 standard deviations of the group mean length), fragmented (partially recovered), and missing (totally absent). The number of BUSCO complete genes was considered as a metric to assess the completeness of each assembly.

Results

An isolate selection spanning the genome diversity inherent in a short-read collection

In large bacterial collections for which possibly thousands of isolates are short-read sequenced, the selection of a subset of isolates for long-read sequencing is crucial to obtain a representative set of complete genomes spanning the genomic diversity present in the collection.

We propose the following approach summarized in the following steps (see Methods for a formal description): (i) consider the presence/absence matrix of orthologous genes computed by established pangenome tools such as Roary or Panaroo; (ii) compute a distance matrix based on Jaccard distances; isolates are compared on the basis of their shared number of orthologous

genes; (iii) reduce the dimensionality of the distance matrix using t-sne by preserving local structure; (iv) cluster the t-sne dimensionally reduced matrix with the k-means algorithm, indicating as the number of centroids the desired number of long-read isolates; and (v) select the isolate with the closest Euclidean distance to its centroid. The proposed approach is fully available as a Snakemake pipeline at [30].

We showcase the approach on a set of 3,254 short-read sequenced isolates from a Norwegian longitudinal population genomic study of *E. coli* causing bacteraemia [32]. The number of isolates selected for long-read sequencing can be specified by fixing the number of centroids in the pipeline. For example, in this collection (Fig. 1), we estimated that selecting 96 long-read isolates would capture ~99.9% of the total variance present in the t-sne matrix computed from Jaccard distances. The selection of the 96 isolates indicated in Fig. 1 would capture all orthologous genes with a frequency >0.05 (Supplementary Fig. S1). As expected, a substantial proportion of genes with a frequency of ~0.01 are not present in these 96 isolates (Supplementary Fig. S1) because those genes are rarely shared by other isolates from the same k-means cluster.

From the same collection of 3,254 isolates, we performed an independent selection of 1,085 *E. coli* isolates for long-read sequencing using the proposed selection strategy. These isolates also belong to the Norwegian collection described above [23] for which short-read sequencing data are publicly available.

We explored the possibility of completing these 1,085 genomes by using the recently released ONT native barcode expansion kit. This kit allowed multiplexing in the same MinION flowcell up to 96 isolates and potentially recovering complete genomes in a high-throughput manner.

In the following sections, we evaluated the complete genomes obtained by multiplexing 96 of these 1,085 isolates, in the same MinION flowcell. These 96 isolates differ from those shown in Fig. 1 because they were selected considering a distinct number of centroids.

Uneven distribution of ONT reads in the 96 multiplexing approach

We used the recently released barcoding kit from ONT (SQK-NBD110-96) to multiplex 96 *E. coli* isolates (96 of 1,085 isolates) in the same MinION flow cell. The ONT sequencing run generated a total of 10.71 Gb QC-passed base-called reads (mean phred score >7) that could be confidently assigned to a barcode, with a mean N50 read length of 20.98 kb (Fig. 2). Considering a genome size of 5 Mb, the expected coverage was ~22× per isolate. From the passed reads, the mean ONT phred score corresponded to 12.48, equivalent to a read accuracy of 94.35%. As previously reported for other multiplexing kits [16], we observed an uneven distribution of reads available per barcode, which resulted in a large variation in the number of bases available per barcode (mean = 111.53 Mb, median = 103.45 Mb). This difference ranged from 3.37 Mb (coverage ~0.67×) for barcode 68 to 244.00 Mb (coverage ~48×) in barcode 73. Prior to the hybrid assemblies, ONT reads were filtered using Filtlong on the basis of quality and length (see Methods), slightly reducing the number of reads available (mean = 94.36 Mb, median = 85.39 Mb). A complete description of ONT statistics per barcode is given in Supplementary Table S1.

Evaluation of the hybrid assemblies

First, we focused on the chromosome sequence by analysing the largest component present in the hybrid assembly graph. In the 96 samples, the chromosome was present in a component with a mean size of 5.0 Mb (median = 5.03) and 11.39 contigs (median = 2.0), respectively (Fig. 3). Despite the mean number of contigs forming the chromosome, the contiguity (N50/component size) of the chromosome replicon was 0.89 (median = 1.0), which indicated that the chromosome component was assembled, for most samples, in a large contig (Fig. 3). Furthermore, in 48 samples (50%) the chromosome resulted in a single circular contig (contiguity = 1.0) (Supplementary Table S2).

We observed a positive correlation (Pearson correlation = 0.41) between the chromosome contiguity and the number of ONT reads available during the hybrid assembly (Fig. 3). We assessed that an approximate ONT depth of ~5× was minimally required to achieve a perfect chromosome assembly (contiguity = 1.0). The lowest contiguity values corresponded to 2 isolates with a poor ONT coverage (<1×) (barcodes 65 and 68). Overall, the isolates with an inferior contiguity value (threshold <0.9) (18 barcodes, ~19%) had an associated low number of ONT reads available during the hybrid assembly (mean = 42.64 Mb, coverage ~8.53×).

Next, we assessed the rest of the components present in the hybrid assembly graphs. The largest segment of each component was predicted with mlplasmids to confirm its plasmid origin. For medium and large plasmid components (size >10 kb) (n = 132), we obtained a mean contiguity of 0.98 (median 1.0) and single circular contigs were retrieved for 118 components (89.4%) (Suppl. Table S2). On average, the 132 plasmid components were formed by 1.79 contigs (median = 1.0). For the isolates with a low chromosome contiguity (<0.9), the mean plasmid contiguity was still 0.97 (median = 1.0), which indicated the suitability of the pipeline for plasmid reconstruction purposes. The correlation between plasmid contiguity and ONT reads available was weaker (Pearson correlation = 0.16) than for the chromosomal component. Even for isolates with a poor ONT coverage, the contiguity values were close to 1.0, indicating that only a few long reads are sufficient to resolve plasmid components. The small plasmids (size <10 kb, n = 78) had a mean contiguity of 0.99 (median = 1.0), and single circular contigs were retrieved for 74 plasmids (94.9%). The absence of repeat sequences in small plasmids makes it possible to already obtain circular replicons by only using short-read sequencing assemblies and thus ONT reads are not required.

Underrepresentation of small and medium plasmids in the ONT library

For each component present in the Unicycler assemblies, we determined its Illumina-read and ONT-read relative depth considering the mean chromosomal depth as the basis for the normalization, as previously performed [16]. This analysis was fundamental to confirm whether small plasmids were underrepresented in the ONT library preparation, as recently reported for ONT ligation kits [52].

As shown in Fig. 4, small plasmids (size <10 kb) were strongly underrepresented in the ONT read output (log_2 −0.63) in contrast to the Illumina read output (log_2 4.79). For 7 small plasmids, we did not observe any ONT reads covering these replicons. These plasmids are usually present in high copy numbers in the cell as a survival and inheritance mechanism [53]. Therefore, we expected that these replicons would be overrepresented in the sequencing output, resembling the plasmid read output given by Illumina (Fig. 4). For medium plasmids (from 10 to 50 kb), we also observed the same trend even though the underrepresentation in the ONT library was less pronounced (log_2 −0.40). No-
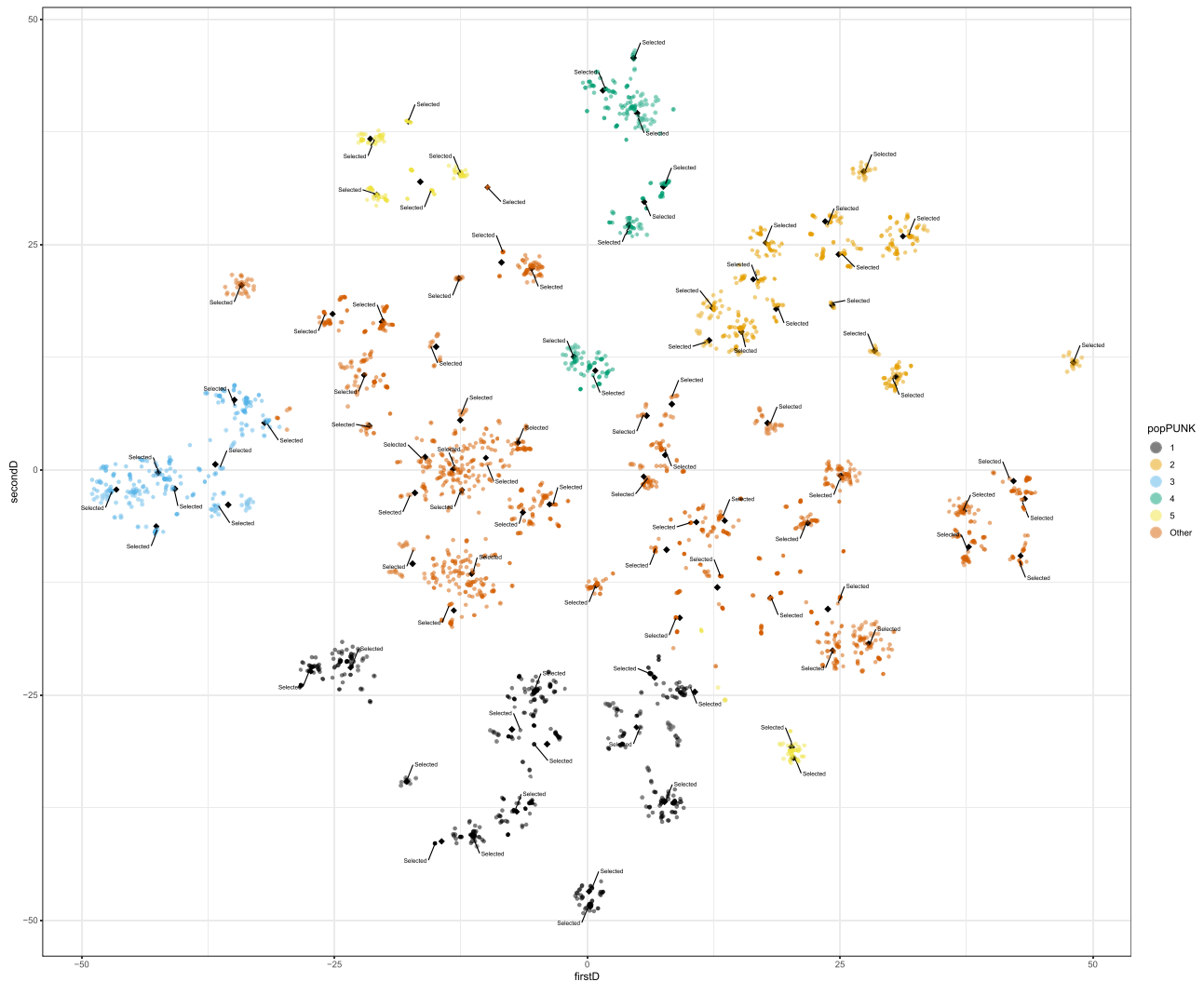
Figure 1. T-sne plot of the isolate selection, based on Jaccard distances computed from the presence/absence of genes defined by Panaroo in 3,254 *E. coli* isolates. The 5 most predominant PopPUNK lineages are indicated with distinct colours; the rest of the lineages were merged into the category "Other" (in yellow). To showcase the pipeline, we fixed the number of centroids to 96, which corresponded to the maximum number of isolates that can be multiplexed in the same MinION flow cell. The final coordinates of the centroids (n = 96) used in k-means are indicated as black diamond points. For each centroid (n = 96), the isolate closest to its respective centroid has been marked as "Selected." In total, we indicate 96 *E. coli* isolates spanning the genomic diversity inherent in the collection that could be selected for further long-read sequencing.

tably, we observed the opposite trend for large plasmids (>50 kb) for which there was an underrepresentation in the Illumina library (log$_2$ −0.60) compared to the ONT library (log$_2$ 0.18).

ONT-only assemblies

To evaluate whether Illumina reads are still required to obtain accurate genomes, we used Flye and Medaka to perform and polish assemblies based only on ONT reads. The largest component in the assembly graph had a mean size of 4.45 Mb (median = 4.98 Mb) and 1.16 contigs (median = 1.0) (Supplementary Table S3). Flye failed to produce an assembly for 2 isolates (barcodes 65 and 68) for which the number of ONT reads generated was <5 Mb (Supplementary Table S1). In 71 samples (76%), the chromosome was represented by a single circular contig. The mean chromosomal contiguity was 0.98 (median = 1.0), which indicated a higher contiguity with respect to the hybrid assemblies. However, the average size of the largest

component (mean = 4.45 Mb, median = 4.98 Mb) was shorter than for Unicycler assemblies (mean = 5.0 Mb, median = 5.03 Mb). This indicated that for some isolates the size of the largest component did not match with the expected replicon size (∼5 Mb).

The analysis of the rest of the components present in Flye assemblies (n = 346) revealed a high number of replicons with a chromosome origin (n = 193), indicating fragmentation of the chromosome into several components. For the components with a plasmid origin (n = 153), we obtained single circular contigs for 120 plasmids (78.4%). We observed a clear difference between the hybrid and ONT-only assemblies with respect to small plasmids (size <10 kb). Only 12 small circular plasmids were recovered in the Flye assemblies, in comparison to the 78 small plasmids present in the hybrid assembly. The absence of these plasmids in the Flye assemblies could be explained by their underrepresentation or complete absence in the total ONT read output, as shown in the section above.
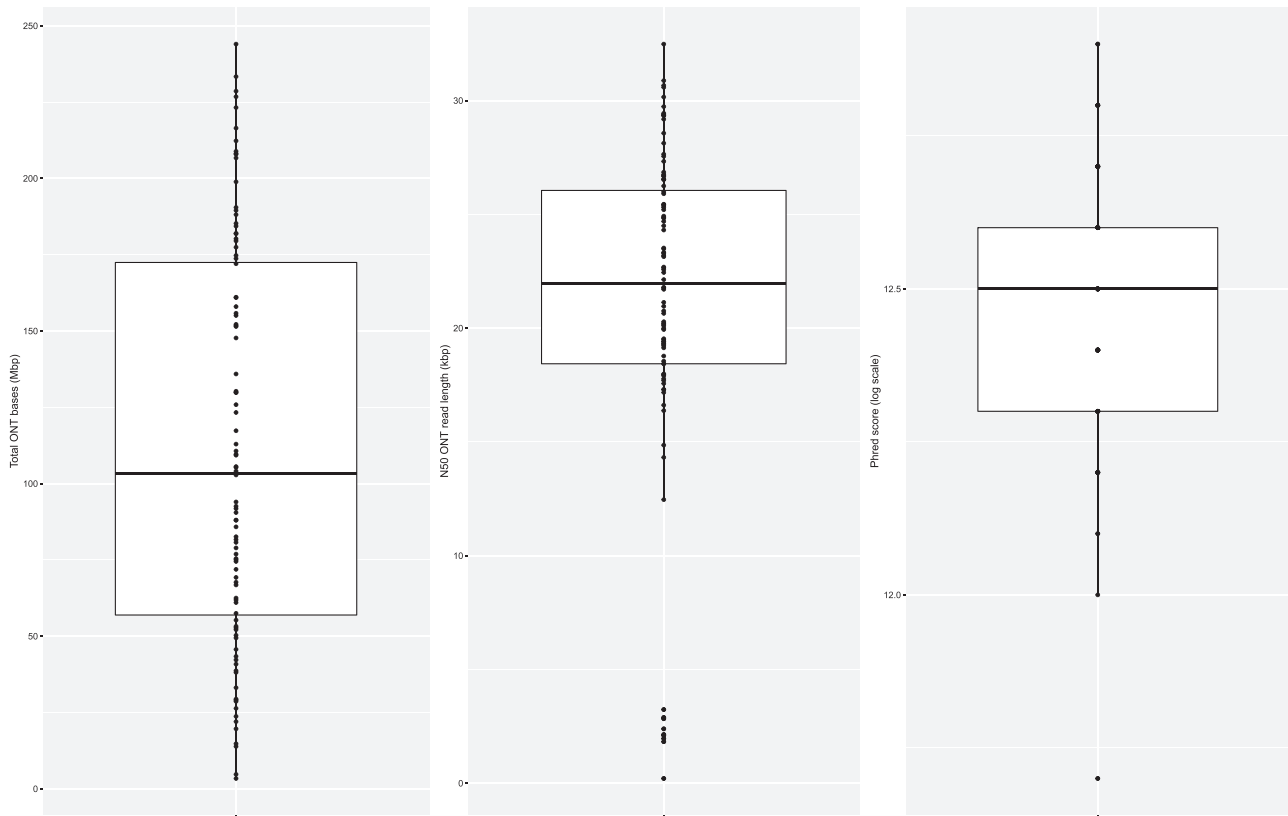
Figure 2. Oxford Nanopore Technologies (ONT) statistics per barcode (n = 96). *Left:* boxplot of the number of bases (Mb) generated in the sequencing run. *Middle:* boxplot of the N50 ONT read length (kb). *Right:* boxplot of the phred score ($\log_{10}$ scale) associated with the ONT reads.

Genome accuracy and completeness of hybrid and ONT-only assemblies

To compare the accuracy of the resulting genomes, the hybrid assemblies and Flye assemblies were compared in terms of SNPs and indels considering both reference-based and reference-free methodologies (see Methods). Despite the fact that Flye incorporates a consensus-error module to correct the resulting genome sequences, we polished the Flye assemblies using Medaka and compared the genome accuracy against hybrid (Unicycler) and stand-alone Flye assemblies. In addition, we evaluated whether a second round of polishing with Medaka could improve the accuracy of the genomes.

The hybrid assemblies showed the best accuracy stats, with a mean of 6.93 SNPs/100 kb (median = 5.24) and 0.31 indels/100 kb (median = 0.21), considering as ground truth non-repetitive alignments (>5 kb) against short-read contigs generated with Velvet. The SPAdes assemblies created by Unicycler with only short reads showed a mean of 0.85 SNPs/100 kb (median = 0.28) and 0.06 indels/100 kb (median = 0.04). This indicated that the accuracy of the hybrid assemblies was affected by the incorporation of error-prone ONT reads into the final genome sequence. The existence of dead ends in the initial SPAdes graph indicated that parts of the genome were not sequenced with short reads; thus their sequence had to be completed with long reads and could not be polished with Illumina reads. As shown in Supplementary Fig. S2, 2 barcodes (69 and 72) had an elevated number of dead ends (38 and 43), which resulted in a high number of SNPs/100 kb (barcode 69: 23.89; barcode 72: 33.36) and indels/100 kb (barcode 69: 1.25; barcode 72: 1.32).

Flye assemblies exhibited a higher number of errors, with a mean of 130.17 SNPs/100 kb (median = 14.68) and 140.82 indels/100 kb (median = 41.52). We observed a negative correlation between the number of ONT bases generated and the resulting number of SNPs (Pearson correlation = −0.27) and indels (Pearson correlation = −0.61) (Fig. 5A and B). The correction performed by Medaka on Flye assemblies showed a reduction in the number of indels (mean = 126.35, median = 22.84), but the number of mismatches remained similar (mean = 130.29, median = 12.57). A second round of polishing with Medaka did not significantly improve the number of indels (mean = 124.16, median = 22.37) or mismatches (mean = 130.66, median = 12.57). Again, we observed a negative correlation between the ONT depth and the resulting number of SNPs (Pearson correlation = −0.27) and indels (Pearson correlation = −0.61) (Fig. 5A and B). For instance, for barcode 73 with the highest ONT read depth, the ONT-only assemblies had a number of SNPs (Flye = 6.52 SNPs/100 kb, Medaka = 6.25 SNPs/100 kb) comparable to Unicycler results (5.58 SNPs/100 kb), which indicated that an increase in the ONT read depth can be highly beneficial to correct SNP errors. However, in the case of indels, we still observed 26.47 indels/100 kb and 6.46 indels/100kb for Flye and Medaka assemblies, respectively, which is far from Unicycler results (0.04 indels/100 kb).

Following on this, we considered a reference-free approach to evaluate the effect of non-corrected indels on the interruption of ORFs using the ideel test [13]. In the hybrid assemblies, the number of promptly terminated ORFs corresponded to 25.07 (median = 21.0), which may include true pseudogenes. For Flye assemblies, the mean number of interrupted ORFs was 1,380.64 (median = 772.5), and the correction with Medaka improved
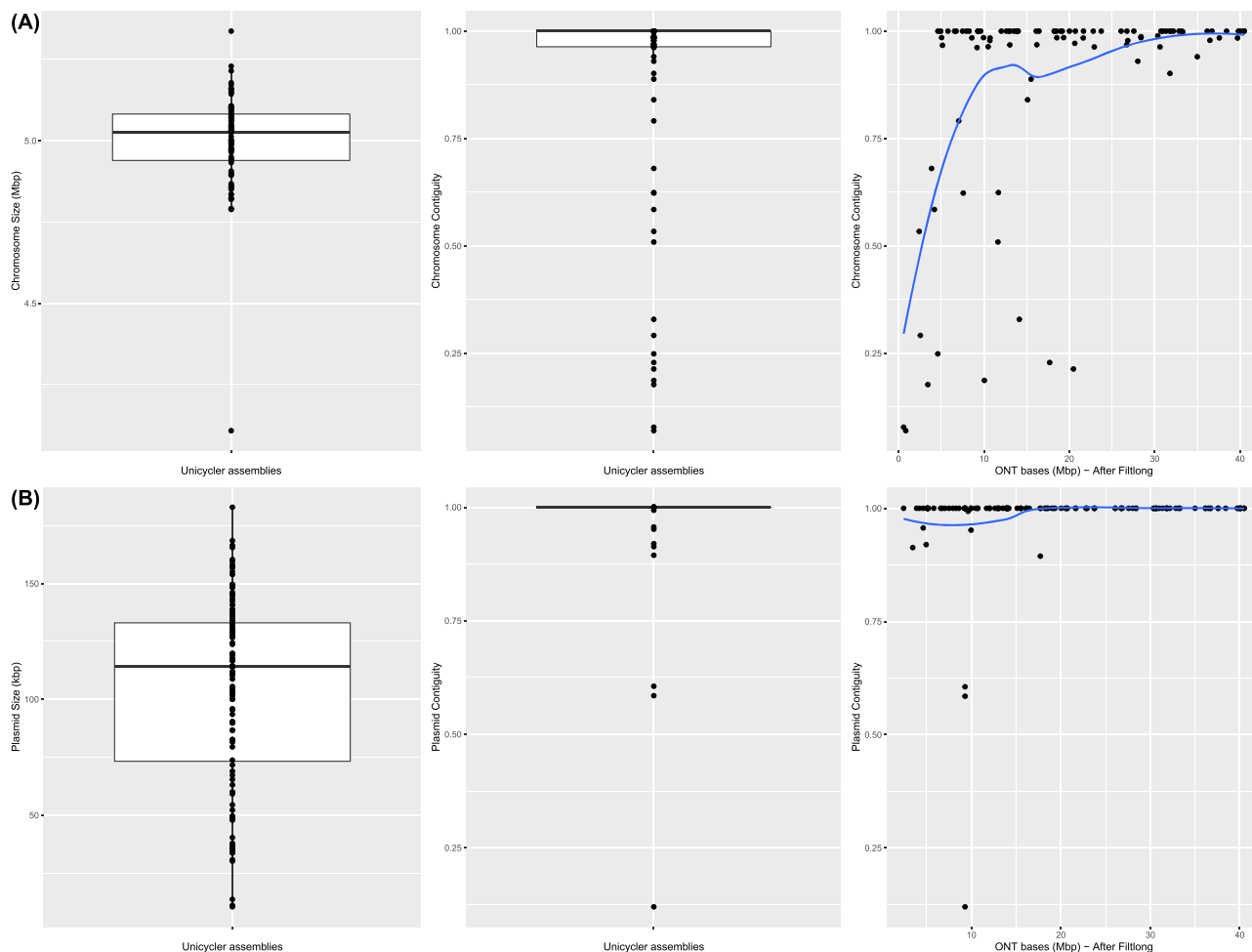
Figure 3. Unicycler statistics for the 96 isolates included in the ONT sequencing. Boxplots showing the component size (left) and contiguity values (middle) together with an evaluation of the correlation between the number of ONT bases generated and the contiguity values achieved (right). (A) Statistics for the chromosome component (largest component); the largest contig was confirmed as chromosome-derived with mlplasmids. (B) Statistics for medium and large plasmid components (size >10 kb); the largest contig was confirmed as plasmid-derived with mlplasmids.

the genome accuracy by reducing the number to 1,127.99 interrupted ORFs (median = 470.0). An additional round of polishing with Medaka resulted in a similar number of interrupted ORFs (mean = 1,103.88, median = 454.5). We observed a negative correlation between the number of ONT bases available and the number of interrupted ORFs present in Flye assemblies (Pearson correlation = −0.76) and Medaka polished assemblies (Pearson correlation = −0.78) (Fig. 4C). We observed a plateau, around ∼500 and ∼150 ORFs for Flye and Medaka polished assemblies, for which an increase in ONT depth did not translate into a lower number of early terminated ORFs. Even for the isolate with the highest ONT depth (barcode 73, ONT bases 244 Mb), the number of interrupted ORFs was still 146 for the Medaka polished assembly in comparison to 23 early terminated ORFs for the hybrid assembly.

Last, we evaluated the completeness of the genomes, searching for complete orthologous genes against a curated set of 440 Enterobacterales single-copy conserved genes (BUSCO genes). For the hybrid assemblies, we recovered all BUSCO genes (100%), in contrast to only 76.6% for Flye assemblies. Furthermore, the percentage of missing genes was 10.3%, which indicated that the *E. coli* genomes assembled by Flye were not complete. Medaka correction increased the number of recovered BUSCO genes

(mean = 80.5%) and slightly decreased the number of missing genes (mean = 9.3%). A second round of polishing with Medaka did not significantly recover more BUSCO genes (mean = 80.81) or decrease the number of absent genes (mean = 9.15%). For Flye and Medaka assemblies, a strong positive correlation of 0.78 was observed for both strategies between the ONT depth and the number of BUSCO complete genes (Fig. 4D). The completeness of the ONT assemblies also stabilized despite an increase in the ONT read depth, around ∼92% and ∼98% for Flye and Medaka assemblies.

Discussion

In this study, we have shown that near-complete genomes can be retrieved by multiplexing 96 bacterial isolates in the same long-read sequencing run. Despite an uneven distribution of ONT reads per barcode, we observed a mean chromosomal contiguity of 0.89, which indicated that for most samples the chromosome was mostly represented by a single long contig. Furthermore, 92% of the plasmid sequences were circularized even for isolates with a low ONT read depth, which makes this high-throughput multiplexing strategy an attractive choice for plasmid studies.
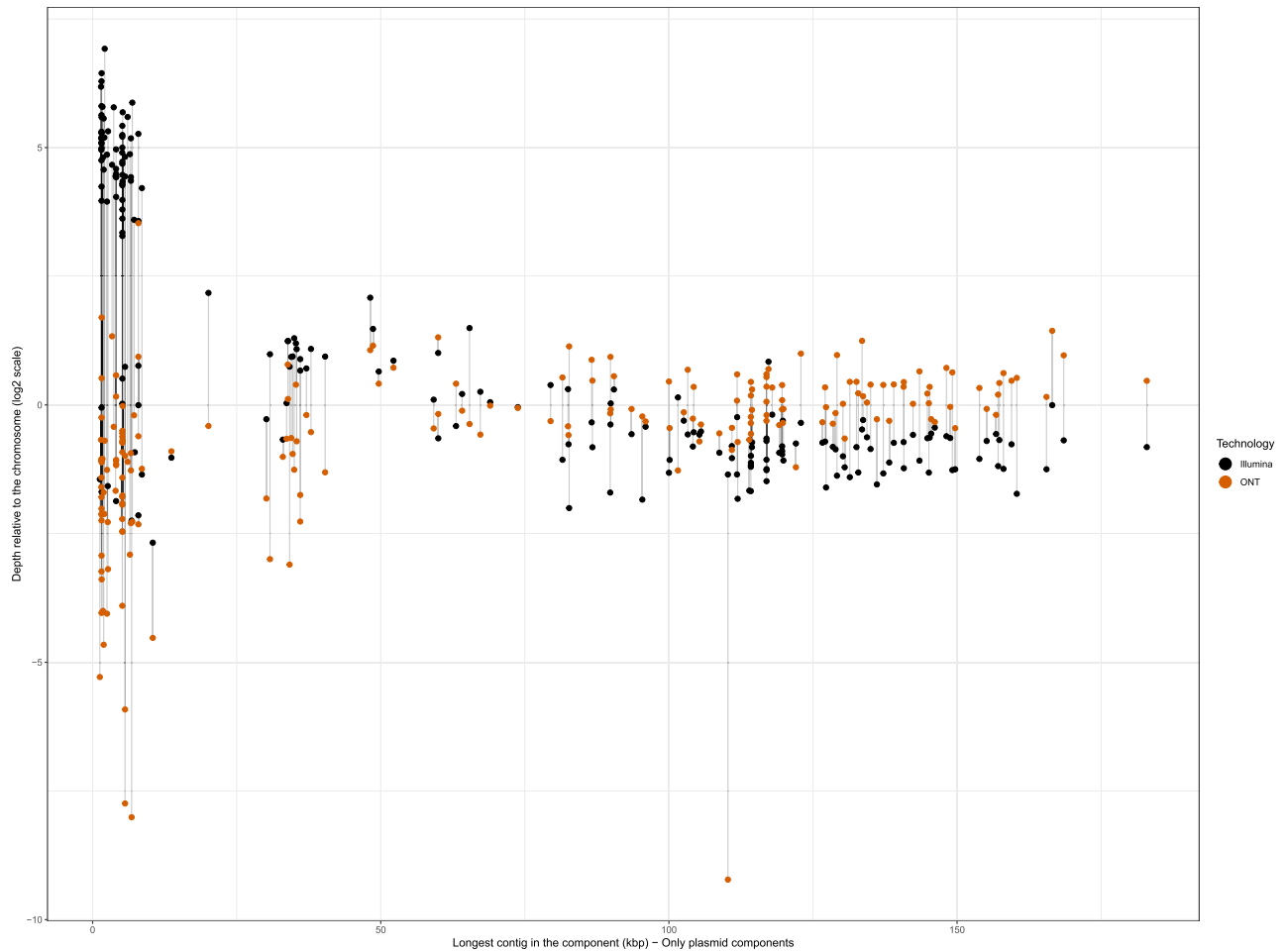
Figure 4. Illumina and ONT depth relative coverage of the plasmid components present in the Unicycler assemblies. The mean depth of the plasmid replicons was normalized against the mean depth of the chromosome to obtain a relative depth ($\log_2$ scale). Each plasmid is represented by 2 connected points depending on the sequencing technology.

The availability of short-read sequencing data from the same bacterial isolates is still strongly desirable for 3 main reasons: (i) to perform an unbiased selection of isolates for long-read sequencing, (ii) to rely on the sequence accuracy achieved by short-read technologies (phred score >30), and (iii) to include small plasmids underrepresented in the ONT library.

We proposed that given a short-read collection of isolates, the pangenome of these isolates can be computed and the presence/absence matrix of orthologous genes considered as the basis for the isolate selection. This collection should preferably include isolates with a low number of dead ends (<5) because this would ensure that the gene content can be confidently retrieved. Furthermore, a low number of dead ends is crucial to obtain an optimal genome contiguity even if the obtained ONT coverage is low.

This selection ensures that complete genomes are generated from distinct clusters that are defined by the gene content of the isolates present in the collection. Furthermore, the generated complete genomes can be used to conduct reference-based approaches relying on the short-read data of the non–long-read sequenced isolates belonging to a particular genomic cluster. A limitation of this approach is that the selected isolate may not possess other accessory genes present in clonally related isolates. This is particularly relevant for medium/small plasmids or

phage elements consisting of only a few genes and thus having a small relative weight in the distance matrix used as a basis to assign the clusters. Alternatively, choosing the isolates from the cluster with the highest distance to its centroid could select for isolates with a more distant gene content and potentially maximize the genome diversity. This issue would decrease if a higher number of centroids were selected because isolates with a more divergent gene content profile would be split into further subclusters during the k-means assignment.

The experimental set-up used in this study (96 barcodes) and the uneven distribution of ONT reads per isolate make it challenging to obtain full and high-quality genome sequences with only long reads. Thus, the reported performance of Flye and Medaka is only informative considering the study constraints. We observed that an increase in the ONT read depth is critical to improving the accuracy of ONT-only assemblies, in particular for SNP calling. However, in the case of indels, the systematic and non-random ONT read errors reported in homopolymer sequences resulted in a high number of early terminated ORFs, even for the isolates with the highest ONT read depth (~150 ORFs). Given the uneven distribution of ONT reads observed in the multiplexing approach, the accuracy and completeness of ONT-only assemblies for the isolates with a low read depth can result in a high number of SNPs and indels in their complete
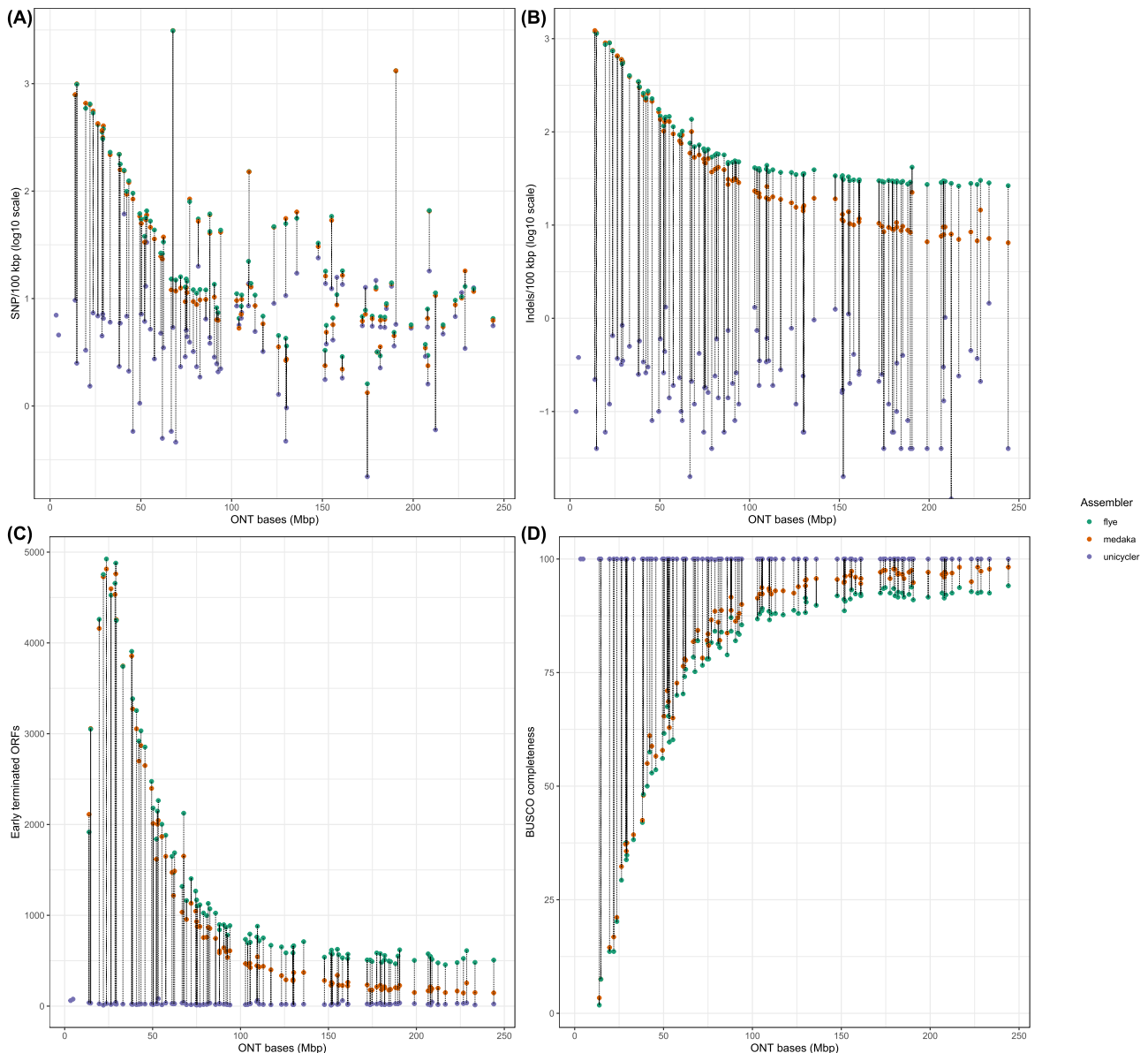
Figure 5. Correlation between the number of ONT bases generated per isolate and genome accuracy statistics for Flye assemblies (green), Medaka-polished assemblies (red), and Unicycler assemblies (blue). (A) Number of SNPs per 100 kb ($\log_{10}$ scale) computed considering as reference short-read contigs generated by an independent assembler. (B) Number of indels per 100 kb ($\log_{10}$ scale) computed considering as reference short-read contigs generated by an independent assembler. (C) Number of early terminated ORFs based on the ideel test. (D) Completeness of the assemblies as a percentage of complete single-copy conserved BUSCO genes (n = 440).

genome sequences. This drawback still allows the genomic context to be identified from a gene-of-interest (e.g., AMR genes) but can limit studies based on SNP signatures such as outbreak investigations. Of note, the use of the newer Nanopore R10.3 chemistry with FLO-MIN111 flow cells would likely increase the base-pair accuracy of the ONT-only assemblies.

In the hybrid assemblies, the accuracy of the complete genomes is unaffected by the ONT read depth because, in general, the long reads are only used as bridges to unequivocally connect short-read contigs. With fewer long reads, we could still obtain a complete genome and its accuracy would be determined by the error read associated with the short-read technology. However, the accuracy of the hybrid assemblies can be affected by the quality of the initial short-read graph. If there is an elevated number of dead ends in the short-read assembly graph,

the sequence of ONT reads would be considered to complete the resulting genome and polishing that particular genomic region with Illumina reads would not be possible.

Recently, Dilthey et al. postulated a new method that allows the multiplexing step to be skipped by pooling together multiple non-barcoded samples in the same ONT flow cell [17]. This method relies on the availability of Illumina data and intersample genetic differences to *in silico* assign ONT reads to particular isolates. This strategy could reduce even further the sequencing costs per isolate and labour time associated with the molecular barcoding preparation. However, a current limitation of the tool is imposed by the high computational requirements required to perform the *in silico* assignment, which can limit its applicability when a large number of samples (e.g., 96) are pooled in the same flow cell.

As previously reported [52], we also observed an underrepresentation of small plasmids in the ONT library, which resulted in their absence in ONT-only assemblies. These plasmids can carry AMR or virulence genes, and overlooking their presence in ONT-only assemblies when using ligation kits can affect subsequent analyses [52]. Small plasmids, however, are present in the initial short-read graph, usually as single circular contigs, and thus the absence of ONT reads covering these replicons does not affect the true representation of the genome. Ideally, the same genomic DNA would be used for both short- and long-read sequencing to avoid any potential bias and losing any plasmids due to the growth and extraction procedures used. In this study, a short-read dataset previously created by Gladstone et al. [32] was used and a renewed growth and DNA extraction step was thereby needed for the ONT sequencing. This limitation could affect the stability of plasmids and may partly contribute to the underrepresentation of some extrachromosomal elements in the ONT library. However, owing to the similarity of the growth conditions used, we expect the potential bias to be minimal in our study.

In conclusion, we have shown the potential of using the recently released ONT nanopore barcoding kit for 96 bacterial isolates to recover near-complete genomes in combination with prior short-read sequencing data. We propose a long-read isolate selection based on the gene content to ensure that the resulting complete genomes span the diversity present in the collection. Finally, the possibility of generating complete genomes on a high-throughput basis will likely continue to significantly advance the field of microbial genomics.

Data Availability

ONT and Illumina sequencing data are available through the ENA Bioprojects PRJEB45354 and PRJEB32059, respectively. For each ONT barcode, individual ONT and Illumina accessions are indicated in Supplementary Table S1. ONT reads and assemblies are available through the following permanent Figshare datasets: ONT reads [54], Unicycler assemblies [55], Flye assemblies [56], and Medaka polished assemblies [57]. The isolate selection pipeline is available at [30]. An Rmarkdown document with the code and files required to reproduce the results presented in this article is available at [58]. Snapshots of the code are available in the *GigaScience* GigaDB repository [26].

Additional Files

Supplementary Figure S1. Histogram with the proportion of orthologous genes captured by the selected 96 *E. coli* isolates. In white, the total number of orthologous genes (n = 33,508) present at the 3,254 *E. coli* isolates were split into 100 bins based on their frequency. In green, the orthologous genes (n = 16,308) present at the 96 selected isolates were split into the same 100 bins. For each white bin, we can observe the proportion of genes (in green) that would be recovered by sequencing the 96 *E. coli* isolates selected in Fig. 1.
Supplementary Figure S2. Genome accuracy comparison between SPAdes assemblies and Unicycler (hybrid) assemblies considering as reference short-read contigs assembled with Velvet. (A) Number of SNP differences per 100 kb observed in the SPAdes assemblies (black circles) and Unicycler assemblies (yellow circles). For each barcode, the difference in SNPs/100 kb is represented by a line connecting the results of the 2 assemblies. (B) Number of indel differences per 100 kb observed in the

SPAdes assemblies (black circles) and Unicycler assemblies (yellow circles). For each barcode, the difference in indels/100 kb is represented by a line connecting the results of the 2 assemblies.
Supplementary Figure S3. Overview of the hybrid assembly accuracy based on the quality of the initial SPAdes assembly graph in terms of dead ends (x-axis). (A) Number of SNP differences per 100 kb in comparison to the number of dead ends present in the SPAdes assembly graph. Each dot represents a barcode, and its size and colour differ depending on the number of contigs (SR contigs) present at the SPAdes assembly graph. (B) Number of indel differences per 100 kb in comparison to the number of dead ends present in the SPAdes assembly graph. Each dot represents a barcode, and its size and colour differ depending on the number of contigs (SR contigs) present at the SPAdes assembly graph.
Supplementary Table S1. Summary of the ONT read statistics and ENA read accessions.
Supplementary Table S2. Assembly statistics and mlplasmids prediction of the components present in the Unicycler assembly graph.
Supplementary Table S3. Assembly statistics and mlplasmids prediction of the components present in the Flye assembly graph.

Abbreviations

AMR: antimicrobial resistance; BLAST: Basic Local Alignment Search Tool; bp: base pair; BUSCO: Benchmarking Universal Single-Copy Orthologs; BWA: Burrows-Wheeler Alignment Tool; HMW: high-molecular-weight; indels: insertions and deletions; IS: insertion sequence; kb: kilobase pairs; Mb: megabase pairs; ONT: Oxford Nanopore Technologies; ORF: open reading frame; QC: quality control; SNP: single-nucleotide polymorphism; t-sne: t-distributed stochastic neighbour embedding; WGS: whole-genome sequencing.

Competing Interests

The authors declare that they have no competing interests.

Funding

Authors' Contributions

P.J.J., Ø.S., and J.C. designed and sought funding for the study. S.A.-A., A.C.S., and J.C. designed and implemented the isolate selection pipeline. A.K.P and F.C. performed the HMW DNA extractions. S.A.-A. performed the computational analyses: read processing, assembly, and genome statistics. R.A.G. facilitated the short-read sequencing data, assemblies, and provided the popPUNK lineages. S.A.-A., A.K.P., and J.C. wrote the first draft

of the manuscript. All authors contributed to and reviewed the manuscript.

References

1. Köser CU, Ellington MJ, Cartwright EJP, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 2012;**8**:e1002824.

2. Köser CU, Holden MTG, Ellington MJ, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med 2012;**366**(24):2267–75.

3. Hendriksen RS, Bortolaia V, Tate H, et al. Using genomics to track global antimicrobial resistance. Front Public Health 2019;**7**:242.

4. Schürch AC, van Schaik W. Challenges and opportunities for whole-genome sequencing-based surveillance of antibiotic resistance. Ann N Y Acad Sci 2017;**1388**(1):108–20.

5. Didelot X, Bowden R, Wilson DJ, et al. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 2012;**13**(9):601–12.

6. Robinson ER, Walker TM, Pallen MJ. Genomics and outbreak investigation: from sequence to consequence. Genome Med 2013;**5**(4):36.

7. Arredondo-Alonso S, Willems RJ, van Schaik W, et al. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microb Genom 2017;**3**(10):doi:10.1099/mgen.0.000128.

8. Orlek A, Stoesser N, Anjum MF, et al. Plasmid classification in an era of whole-genome sequencing: Application in studies of antibiotic resistance epidemiology. Front Microbiol 2017;**8**:182.

9. Amarasinghe SL, Su S, Dong X, et al. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 2020;**21**(1):30.

10. Nanopore Sequencing Accuracy. https://nanoporetech.com/accuracy. Accessed 28 April 2021.

11. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 2015;**12**(8):733–5.

12. Risse J, Thomson M, Patrick S, et al. A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. Gigascience 2015;**4**:60.

13. Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. Nat Biotechnol 2019;**37**(2):124–6.

14. Medaka: Sequence correction provided by ONT Research. https://github.com/nanoporetech/medaka. Accessed 21 September 2021.

15. De Maio N, Shaw LP, Hubbard A, et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom 2019;**5**(9):doi:10.1099/mgen.0.000294.

16. Wick RR, Judd LM, Gorrie CL, et al. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom 2017;**3**(10):e000132.

17. Dilthey AT, Meyer SA, Kaasch AJ. Ultraplexing: increasing the efficiency of long-read sequencing for hybrid assembly with k-mer-based multiplexing. Genome Biol 2020;**21**(1):68.

18. Lipworth S, Pickford H, Sanderson N, et al. Optimized use of Oxford Nanopore flowcells for hybrid assemblies. Microb Genom 2020;**6**(11):doi:10.1099/mgen.0.000453.

19. Wick RR, Judd LM, Gorrie CL, et al. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017;**13**:e1005595.

20. Nguyen SH, Cao MD, Coin LJM. Real-time resolution of short-read assembly graph using ONT long reads. PLoS Comput Biol 2021;**17**:e1008586.

21. Arredondo-Alonso S, Top J, McNally A, et al. Plasmids shaped the recent emergence of the major nosocomial pathogen *Enterococcus faecium*. MBio 2020;**11**(1):doi:10.1128/mBio.03284-19.

22. Pöntinen AK, Top J, Arredondo-Alonso S, et al. Apparent nosocomial adaptation of *Enterococcus faecalis* predates the modern hospital era. Nat Commun 2021;**12**(1):1523.

23. Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;**31**(22):3691–3.

24. Tonkin-Hill G, MacAlasdair N, Ruis C, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol 2020;**21**(1):180.

25. R Package: Parallel Distance Matrix Computation using Multiple Threads. https://github.com/alexeckert/parallelDist. Accessed 2 December 2020.

26. Arredondo-Alonso S, Pöntinen AK, Cléon F, et al. Supporting data for "A high-throughput multiplexing and selection strategy to complete bacterial genomes." GigaScience Database 2021. http://dx.doi.org/10.5524/100942.

27. Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;**9**:2579–605.

28. R wrapper for Van der Maaten's Barnes-Hut implementation of t-Distributed Stochastic Neighbor Embedding. https://github.com/jkrijthe/Rtsne. Accessed 2 December 2020.

29. Kanungo T, Mount DM, Netanyahu NS, et al. An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 2002;**24**(7):881–92.

30. Snakemake pipeline to select isolates for long-read sequencing based on a gene presence/absence matrix. https://gitlab.com/sirarredondo/long_read_selection. Accessed 9 June 2021.

31. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 2012;**28**(19):2520–2.

32. Gladstone RA, McNally A, Pöntinen AK, et al. Emergence and dissemination of antimicrobial resistance in E*scherichia coli* causing causing bloodstream infections in Norway in 2002–17: a nationwide, longitudinal, microbial population genomic study. Lancet Microb 2021;**2**(7):e331–41.

33. Velvet Optimiser VBC website https://bioinformatics.net.au/software.velvetoptimiser.shtml.

34. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;**18**(5):821–9.

35. Boetzer M, Henkel CV, Jansen HJ, et al. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 2011;**27**(4):578–9.

36. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol 2012;**13**:R56.

37. Lees JA, Harris SR, Tonkin-Hill G, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res 2019;**29**(2):304–16.

38. Porechop: adapter trimmer for Oxford Nanopore reads. https://github.com/rrwick/Porechop. Accessed 7 February 2021.

39. Filtlong: quality filtering tool for long reads. https://github.com/rrwick/Filtlong. Accessed 7 February 2021.

40. Arredondo-Alonso S, Rogers MRC, Braat JC, et al. mlplasmids: A user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microb Genom 2018;**4**(11):doi:10.1099/mgen.0.000224.

41. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**(4):357–9.

42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;**25**(14):1754–60.

43. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;**10**(2):doi:10.1093/gigascience/giab008.

44. Kolmogorov M, Yuan J, Lin Y, et al. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 2019;**37**(5):540–6.

45. Tricycler documentation, guide to bacterial genome assembly. https://github.com/rrwick/Trycycler/wiki/Guide-to-bacterial-genome-assembly#13-medaka. Accessed 8 March 2021.

46. Medaka: sequence correction provided by ONT Research. https://github.com/nanoporetech/medaka. Accessed 8 March 2021.

47. Gurevich A, Saveliev V, Vyahhi N, et al. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 2013;**29**(8):1072–5.

48. Ideel: Indels are not ideal - quick test for interrupted ORFs in bacterial/microbial genomes. https://github.com/mw55309/ideel. Accessed 4 April 2021.

49. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;**12**(1):59–60.

50. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;**215**(3):403–10.

51. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**(19):3210–2.

52. Wick RR, Judd LM, Wyres KL, et al. Recovery of small plasmid sequences via Oxford Nanopore sequencing. Microb Genom 2021;**7**(8):doi:10.1099/mgen.0.000631.

53. Million-Weaver S, Camps M. Mechanisms of plasmid segregation: Have multicopy plasmids been overlooked? Plasmid 2014;**75**:27–36.

54. ONT passed reads. https://doi.org/10.6084/m9.figshare.14778333. Accessed 14 June 2021.

55. Unicycler assemblies. https://doi.org/10.6084/m9.figshare.14705979. Accessed 31 May 2021.

56. Flye assemblies. https://doi.org/10.6084/m9.figshare.14706015. Accessed 31 May 2021.

57. Medaka polished assemblies. https://doi.org/10.6084/m9.figshare.14706108. Accessed 31 May 2021.

58. Code and documentation to reproduce the results presented in the manuscript. https://gitlab.com/sirarredondo/highthroughput_strategy. Accessed 14 June 2021.