The use of Oxford Nanopore native barcoding for complete genome assembly.

Sion C. Bayliss[a]*, Vicky L. Hunt[a], Maho Yokoyama[a], Harry A. Thorpe[a] and Edward J. Feil[a]

[a] The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

* For correspondence: s.bayliss@bath.ac.uk. Tel: +44 (0)1225 340959

Abstract

Background

The Oxford Nanopore Technologies MinION(TM) is a mobile DNA sequencer that can produce long read sequences with a short turn-around time. Here we report the first demonstration of single contig genome assembly using Oxford Nanopore native barcoding when applied to a multiplexed library of 12 samples and combined with existing Illumina short-read data. This paves the way for the closure of multiple bacterial genomes from a single MinION(TM) sequencing run, given the availability of existing short-read data. The strain we used, MHO_001, represents the important community-acquired methicillin resistant *Staphylococcus aureus* lineage USA300.

Findings

Using a hybrid assembly of existing short read and barcoded long read sequences from multiplexed data, we completed a genome of the *S. aureus* USA300 strain MHO_001. The long-read data represented only ~5-10% of an average MinION(TM) run (~7x genomic coverage), but, using standard tools, this was sufficient to complete the circular chromosome of *S. aureus* strain MHO_001 (2.86 Mb) and two complete plasmids (27 Kb and 3 Kb). Minor differences were noted when compared to USA300 reference genome, USA300_FPR3757, including the translocation, loss and gain of mobile genetic elements.

Conclusion

Here we demonstrate that MinION(TM) reads, multiplexed using native barcoding, can be used in combination with short-read data, to fully complete a bacterial genome. The ability to complete multiple genomes, for which short-read data is already available, from a single MinION(TM) run is set to impact on our understanding of accessory genome content, plasmid diversity and genome rearrangements.

Keywords

Whole genome sequencing, *Staphylococcus aureus*, MinION, long read, hybrid assembly, bacterial genomics, multiplexing, native barcoding.

Data description

The spread of methicillin resistant *Staphylococcus aureus* (MRSA) represents a significant burden in both the health-care setting and the community. The USA300 clone is a particular cause for concern, being responsible for an increasing number of skin and soft-tissue infections within the community, particularly in North America [1]. The advent of new sequencing technologies is set to inform on novel intervention and surveillance strategies, although important technical limitations remain. Whilst short read data provides an excellent means to assay the variation within the core genome, which is useful for reconstructing hospital outbreaks, it is usually not possible to infer genome re-arrangements or to fully assemble mobile genetic elements (MGEs) such as plasmids from these data. Closure of bacterial genomes has been demonstrated on *Escherichia coli* using the Oxford Nanopore Technologies (ONT) MinION(TM) reads alone and on a range of bacteria including *Bacteriodes fragilis, Acinetobacter baylyi* and *Francisella* spp. using a hybrid approach combining error prone long reads with low error rate short reads [2–5]. Here we demonstrate that it is also possible to generate complete genomes using multiplexed reads from a single MinION(TM) run in combination with matched Illumina short reads. We used a strain of *S. aureus* of the USA300 lineage as an example.

Methods

MinION(TM) library construction and sequencing

S. aureus strain MHO_001 was recovered in 2015 from asymptomatic nasal carriage via a standard nasal swab of a healthy individual with informed consent. DNA from an overnight culture was extracted using the Qiagen Genomic Tip 500/G Kit, following the manufacturer's instructions, except lysozyme was replaced with lysostaphin to a final concentration of 200 µg/ml. Sequencing library preparation was carried out with Nanopore Genomic Sequencing Kit SQK-MAP006 (Oxford Nanopore Technologies, UK) and a PCR free ‘native barcoding’ kit provided by ONT. The NEBNext Ultra II End Repair/dA Tailing kit (E7546S, NEB, USA) was used to prepare 1000 ng of sheared genomic DNA (1000 ng DNA in 50 µl nuclease free water, 7 µl of Ultra II End-Prep Buffer, 3 µl Ultra II End-Prep Enzyme Mix in a total volume of 60 ul). The reaction was incubated for 5 minutes at 20°C and heat inactivated for 5 minutes at 65°C. The DNA was purified using a 1:1 volume of Agencourt AMPure XP beads (A63880, Beckman Coulter, USA) according to manufacturer's instructions and eluted in 31 µl of nuclease free water. Blunt/TA Ligase Master Mix (M0367S, NEB, USA) was used to ligate native barcode adapters to 22.5 µl of 500ng end prepared DNA for 10 minutes at room temperature. The barcoded DNA was purified using a 1:1 volume of AMPure XP beads and eluted in 26 µl nuclease free water. Twelve barcoded samples from diverse sources including other bacterial samples were pooled, 58 ng of each sample was added to give 700 ng of pooled library DNA. Hairpin adapters were ligated using 10 µl Native Barcoding Adapter Mix, 50 µl Blunt/TA Ligase Master Mix and 2 µl Native Barcoding Hairpin Adapter (BHP) added to 38 µl the pooled library DNA to give a final reaction volume of 100 µl. The reaction mixture was incubated for 10 minutes at room temperature before the addition of 1 µl of HP tether and a further 10 min incubation. The final reaction was cleaned using pre-washed Dynabeads MyOne Streptavidin C1 beads (65001; Thermo Fisher Scientific, USA). DNA concentrations at each step were measured using a Qubit Fluorometer. 6 µl of the pooled, barcoded library was mixed with 65 µl nuclease free water, 75 µl 2x Running Buffer and 4 µl Fuel Mix (SQK-MAP006, Oxford Nanopore Technologies, UK) and immediately loaded onto a MinION(TM) Flow Cell Mk I R7.3 on a MinION(TM) MkI controlled by MinKNOW version 0.50.2.15 software (Oxford Nanopore Technologies, UK). Base calling was performed using Metrichor ONT Sequencing Workflow Software v1.19.0 with the Basecall_Barcoding workflow (Oxford Nanopore Technologies, UK). The additional DNA samples included in the pooled library were a diverse assemblage of bacterial and eukaryotic DNA samples provided by attendees during the PoreCamp Workshop 2015 at the University of Birmingham. The additional pooled library samples are being prepared for separate publication. Details on the PoreCamp Workshop and associated publications can be found at http://porecamp.github.io/. MinION reads were deposited in the European Nucleotide Archive under study accession PRJEB14152.

Illumina library construction and sequencing

An overnight culture was grown on TSB agar from a 15% glycerol stock maintained at -80 ^{o}C. An aliquot of the culture was added to tubes containing DNA beads and library preparation was carried out by MicrobesNG, University of Birmingham (http://microbesng.uk). A single 250 bp paired end library was constructed and sequenced on both MiSeq and HiSeq Illumina platforms. The reads from both sequencing runs were combined before downstream analysis. The sequenced strain is stored in the MicrobesNG indexed repository as strain 2998-174. Reads were deposited in the European Nucleotide Archive under study accession PRJEB14152.

Assembly, Annotation and Analysis

The full informatics analysis and associated data is available as a step-by-step walk-through at https://github.com/SionBayliss/MHO_analysis. Illumina reads were trimmed using Trimmomatic-0.33 [6]. Reads were trimmed to a minimum read quality of Q15. Reads below 30 bp in length were excluded and sequencing adapters were removed. MinION(TM) 2D reads were filtered into pass and fail reads by the Metrichore basecaller, hereafter these two categories of reads will be referred to as “2D pass” and “2D fail” reads, following the terminology adopted by the manufacturer and used in Karlsson et al. and Ip et al. [4,7]. These are equivalent to the “high quality” and “low quality” read groups from Oikonomopoulos et al. [8]. MinION(TM) 1D reads were not used for this analysis. The 2D fail reads, those which did not pass the basecaller quality threshold, were demultiplexed using an in-house script (FilterBarcodes.pl). The twelve 40 bp barcodes used for library construction were compared in a moving 40 bp window to the sequence in the first and last 150 bp of each read. The barcode requiring the least insertions, deletions or substitutions to be permuted into a sequence in the beginning or end of a read, with a maximum cut-off of 14 permutations, was considered a match. Each read could only be assigned to one individual sample, in the case of a tie the reads were discarded. Sequence preceding or following the presence of a barcode at the beginning or end or a read, respectively, were trimmed as adapter sequence. After quality trimming, 439,480 paired short reads, 1324 2D pass reads and 1499 demultiplexed 2D fail reads (2823 total) nanopore long reads were passed as input files to SPAdes v3.6.1 using the --nanopore, --cov-cutoff 5 and --careful options [9]. The nanopore reads had a median read length of 7,577 bp, a maximum length of 23,380 bp and a minimum length of 250 bp (Figure 1A). After assembly, all contigs less than 300 bp were removed. This resulted in three contigs, the complete chromosome of MHO_001, and two complete plasmids. The contigs were circularised by MUSCLE v3.8.31 alignment (default parameters) of identical overlapping regions at the end of contigs and removal of one alternative overlapping sequence using an in-house script (CirculariseOnOverlaps.pl) [10]. Start sites were fixed relative to the beginning of the relevant reference sequence. A BLAST search against the nt/nr database using default megablast settings revealed the closest, well studied, reference genome was USA300_FPR3757 (Genbank:CP000255) [11]. The two smaller contigs were 100% identical in both aligned sequence and alignment length to previously sequenced *S. aureus* lineage USA300 plasmids, SAP046A (Genbank:GQ900404.1) and SAP046B (Genbank:GQ900403.1). The smallest plasmid was also identical to USA300_FPR3757 plasmid pUSA01 (CP000256). The complete genome of MHO_001 was annotated using Prokka 1.11 [12].

In order to calculate per base read coverage short and long reads were mapped to MHO_001 using BWA 0.7.12-r1039 and coverage was calculated using samtools 1.2 [13,14]. Nanopore reads were mapped using the 'bwa mem -x ont2d' option. In order to assess the sequence similarity and number of reads mapped between the long reads and the MHO_001 assembly the nanopore 2D pass, demultiplexed 2D fail reads and 2D fail reads in which no barcodes were identified were aligned to the MHO_001, including plasmids, using BLASR (Figure 1, Table 1) [15]. SNPs were called between the chromosome and reference genome using MAUVE [16]. SNPs were further confirmed by mapping short reads independently to USA300_FPR3757 and calling variants. Mapping was performed using BWA, reads at indel sites were realigned using the GATK toolbox and SNPs were called using samtools [14,17]. The variant call file (VCF) was filtered for variants supported by a minimum read depth of 4 (minimum 2 per strand), >30 map quality, >50 average base quality, no significant strand bias and >75% of reads supporting the variant. Indels were additionally confirmed using pindel [18]. The VCF file was filtered to remove regions unique to MHO_001 or USA300_FPR3757. Repeat regions of >50bp, which are notoriously problematic for short read mapping, were identified using nucmer and removed from the comparison [19] [Supplementary Table 1]. The absence of SAPI5 in MHO_001 and expansion of the tRNA island at 554,826 were confirmed using PCR and Sanger sequencing (Supplementary Analysis).

Results and Discussion

A hybrid assembly using a low coverage of MinION(TM) reads (6-8x) combined with moderate coverage Illumina reads (~50x) was used to generate a complete genome. The assembly resolved regions of the genome that were problematic for short read assembly alone, such as chromosomal rRNA operons. The generation of a complete genome from only ~5% of the possible current yield of a MinION(TM) run using a multiplexed library should represent a cost effective means to complete multiple genomes during a single MinION(TM) sequencing run, although the approach also requires matching short-read Illumina data. Larger or more complex bacterial genomes may require higher coverage read data alongside additional bioinformatics analyses to generate comparably polished, complete genomes [3].

By demultiplexing the 2D fail reads we were able to double the number of nanopore reads for assembly from 1324 to 2823 reads. The nanopore reads were aligned to the complete MHO_001 genome using BLASR (Figure 1, Table 1). 1320/1324 (99.70%) 2D pass reads demultiplexed by Metrichor aligned to the assembly with an average percentage similarity of 85.87% and a mean alignment length of 96.79% of the input read. 1292/1499 (99.70%) 2D fail reads demultiplexed by in-house scripts aligned to the assembly with an average percentage similarity of 77.76% and a mean alignment length of 92.90%. The fail reads in which we failed to find a barcode contained 722/9501 (7.60%) reads that aligned to the MHO_001 genome. In summary, a considerable amount of useful information was contained within the demultiplexed 2D fail reads without which we would have been unable to produce a complete genome. We can conclude that we were able to correctly identify the ONT barcodes in ~85% of the 2D fail reads used for assembly.

The chromosome showed minor differences to the USA300 reference genome USA300_FPR3757 including 155 SNP differences and the loss and gain of mobile genetic elements (Figure 2). In order to provide an independent confirmation of the 155 SNP differences identified by MAUVE between aligned regions of MHO_001 and USA300_FPR3757 the short reads were mapped to USA300_FPR3757 and variants were called using strict parameters. Of the 155 MAUVE SNPs 41 (26.5%) were present in repeat regions and excluded from the comparison. Of the remaining 114

SNPs, 111 (97.36%) were supported by short read mapping to USA300_FPR3757. The remaining 3 SNPs (2.6%) were unsupported. No indels were identified by short read mapping to MHO_001 by either GATK/samtools or pindel. In summary, of the 114 SNPs identified by MAUVE that could be robustly investigated by short read mapping 111 (97.4 %) were confirmed using low error rate short reads. Furthermore, the long and short read coverage support at the edge of each of the large structural variants in MHO_001 was 8-10x for nanopore reads, with the exception of the 3' edge of the transposed 13,356 bp insertion sequence (IS) which had a read coverage of 3x, compared to the genomic average of 6.8x coverage. The edge of each structural variant was supported by >25 short reads.

There was minor sequence dissimilarity, including a small deletion, in ribosomal RNA operons. This could either reflect evolutionary changes in these highly conserved sequences or minor misassembly; these regions are typically difficult to assemble. MHO_001 lacked Staphylococcal pathogenicity island 5 (SAPI5), a 13,960 bp exotoxin encoding transposon observed at position 881,852 in the reference. MHO_001 also lacked the prophage phiSA3USA which harbours the important virulence factor staphylokinase. As the integration site of this phage (the *hlb* gene) is intact it is possible that MHO_001 has never acquired this phage. MHO_001 contained a 42,297 bp tyrosine recombinase bacteriophage integrated at position 867,385. This bacteriophage contained a beta-lactamase and a putative Panton-Valentine-like leuckocidin and several hypothetical genes. The position of an insertion sequence containing ftsK translocase differs between MHO_001 and the reference genome, consistent with a translocation event (USA300_FPR3757:1630720-1644076 to MHO_001:679522-692877). The location of this element in MHO_001 truncates a gene of unknown function. There is a short 1282 bp deletion of a gene encoding an exotoxin at position 448,767 in MHO_001. MHO_001 also has an extended tRNA cluster at 554,826 containing 7 additional tRNAs (val, thr, lys, gly, leu, arg, pro) relative to USA300_FPR3757, representing either gene expansion, or reduction of this gene cluster in USA300_FPR3757.

A BLAST search revealed that the two smaller contigs were identical to previously sequenced plasmids associated with USA300 [20]. The larger of the plasmids contained an N-type replication system (repA) with a pSK1 type plasmid partitioning system. It encoded a host of resistance genotypes including macrolide (mac), erythromycin (ery), cadmium (cadX and cadD), streptothricin (sta), aminoglycoside (aad), neomycin and kanamycin (aph) resistance genes. In addition to this the plasmid contained a Tn552-like transposon containing a beta-lactam resistance (bin, blaI, blaR1, blaZ) operon and a sin recombinase. The smaller of the two plasmids encoded three hypothetical proteins and a replicase. Both plasmids have been previously observed to occur concurrently in the same host.

There was a discrepancy observed between the coverage of short and long reads of plasmidic and chromosomal contigs (Figure 2, top and middle panels). The average chromosomal coverage was 49.6x (7.0 SD) with short read data and 6.8x (2.6 SD) with nanopore reads. The average short read coverage of plasmids A and B was 78.35 (8.9 SD) and 7302.04 (85.4 SD) respectively. This represents a coverage increase of 1.5- and 150-fold relative to the chromosome. The opposite trend was observed with long reads; plasmids A and B had and average coverage of 4.05 (2.0 SD) and 2.9 (1.7 SD) respectively, which represents a 40% and 60% decrease in coverage relative of the chromosome. In addition to this the smaller of the two plasmids was only intermittently covered by nanopore reads. The reduced number of mappable nanopore reads was likely due to the fragment size selection steps during library preparation. The inherent problems of aligning long error-prone reads to reference sequences may also have contributed. It is thus important that future studies attempting to reconstruct plasmids or studying plasmid diversity consider the impact of size selection on downstream analysis or to prepare multiple DNA libraries with differential size selection as previously discussed by Koren and Phillippy [21]. However, the clear benefit of hybrid sequencing is that it allows for the generation of larger assemblies with less uncertainties than by using a single sequencing technology preferentially over another.

Competing interests

No competing interests.

Funding

The authors would like to acknowledge BBSRC/NERC grant number BB/M026388/1 for providing funding for SB. SB and VH were also funded by a grant from the United Kingdom Clinical Research Collaboration (UKCRC) Translational Infection Research (TIR) initiative, and the Medical Research Council (Grant Number G1000803, held by Prof. Sharon Peacock) with contributions from the Biotechnology and Biological Sciences Research Council, the National Institute for Health Research on behalf of the Department of Health, and the Chief Scientist Office of the Scottish Government Health Directorate. The authors are grateful for travel funds provided by NERC (NE/N000501/1) for SB and Medical Research Council Cloud Infrastructure for Microbial Bioinformatics (CLIMB) for VH to attend.

Authors Contributions

SB and VH were responsible for the conception and design of study and data acquisition. SB performed the analysis and interpretation of data and manuscript drafting. MY carried out the supplementary analysis. HAT and EF revised the manuscript critically for important intellectual content. SB and EF approved the version of the manuscript to be published.

Acknowledgements

The authors would like to acknowledge the contribution of Nick Loman and the rest of the organizing committee of the PoreCamp 2015 nanopore training workshop which was hosted in the Centre for Computational Biology at the University of Birmingham during December 2015. We would like to thank Oxford Nanopore Technologies for allowing the University of Bath access to the MinION(TM) Access Programme (MAP).

Data Availability

The dataset supporting the conclusions of this article is available in the European Nucleotide Archive repository under project number PRJEB14152. Further supporting data is also available from the *GigaScience* GigaDB repository [22].

Availability and requirements

- Project name: MHO_001 hybrid read assembly and analysis
- Project home page: https://github.com/SionBayliss/MHO_analysis
- Operating system: Unix
- Programming language: R, perl
- Other requirements: Dependencies include Samtools (>=1.18), Trimmomatic, SPAdes v3.6.1, BWA (0.7.5a-r405), BioPerl, MAUVE, BLASR, prokka, Tablet/Artemis
- License: GNU GPL v3

References

1. Glaser P, Martins-Simões P, Villain A, Barbier M, Tristan A, Bouchier C, et al. Demography and Intercontinental Spread of the USA300 Community-Acquired Methicillin-Resistant *Staphylococcus aureus* Lineage. MBio. 2016;7:e02183-15.

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

3. Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G, Blaxter M, et al. A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore

sequencing data. Gigascience. 2015;4:60.

4. Karlsson E, Lärkeryd A, Sjödin A, Forsman M, Stenberg P. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci. Rep. 2015;5:11996.

5. Madoui M-A, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, et al. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics. 2015;16:327.

6. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

7. Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research. 2015;4:1–35.

8. Oikonomopoulos S, Wang YC, Djambazian H, Badescu D, Ragoussis J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Sci. Rep. 2016;6:31602.

9. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 2012;19:455–77.

10. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

11. Diep BA, Gill SR, Chang RF, Phan TH, Chen JH, Davidson MG, et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. Lancet . 2006;367:731–9.

12. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.

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

14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

15. Chaisson MJ, Tesler G, Smith T, Waterman M, Zhang Z, Schwartz S, et al. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13:238.

16. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5:e11147.

17. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.

Genome Res. 2010;20:1297–303.

18. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865–71.

19. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12

20. Shearer JES, Wireman J, Hostetler J, Forberger H, Borman J, Gill J, et al. Major families of multiresistant plasmids from geographically and epidemiologically diverse staphylococci. G3. 2011;1:581–91.

21. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr. Opin. Microbiol. 2015;23:110–20.

22. Bayliss SC, Hunt VL, Yokoyama M, Thorpe HA, Feil EJ. Supporting data for "The use of Oxford Nanopore native barcoding for complete genome assembly". GigaScience Database. 2016. http://dx.doi.org/10.5524/100269.

Figure Legends

Figure 1. Figure summarising read statistics for the 2D nanopore pass (red) and fail (green) reads. A) Read length distributions of pass and fail reads. Data was binned every 500 bp. B) Box and whisker plot of the sequence similarity of nanopore reads to the genome of MHO_001 as determined by BLASR. Only the alignment with the highest percentage similarity was considered for each read. The lower and upper "hinges" correspond to the first and third quartiles. The upper and lower whiskers extend from the hinge to the most extreme value that is within 1.5 x interquartile range. Data beyond the end of the whiskers are outliers and plotted as points. C) The distribution of BLASR alignment lengths of nanopore reads as a percentage of the original read length. Only the alignment with the highest percentage similarity was considered for each read. Nanopore 2D reads with a phred score greater than 8 were classified by Metrichor as pass reads (blue), all other 2D reads were classified as fail reads (blue).

Table 1. Table summarising the BLASR analysis of demultiplexed 2D pass and fail nanopore long reads assigned to sample MHO_001. Reads were aligned to the assembled MHO_001 reference genome using BLASR with default parameters. Only the alignment with the highest percentage similarity was considered for each read. The average alignment length was calculated from the length of the top BLASR alignment relative to the length of the input read.

Figure 2. Alignment of MHO_001 chromosome (A), plasmid A (B) and plasmid B (C) to the USA300_FPR3757 genome and reference plasmids alongside long and short read coverage. The bottom panels show alignments between MHO_001 and the reference sequences. Contiguous sequences are shown by connecting red lines and inversions are depicted in blue. Coding sequences (CDS) are annotated as blue rectangles with the exception of ribosomal RNA operons which are represented by red rectangles. Those above the line represent open reading frames on the forward strand and those under the line on the reverse strand. Notable mobile genetic elements or genomic features are annotated. A scale bar in basepairs (bp) is presented underneath each contig. The middle panels represent per base read coverage of short reads across the MHO_001 genome. The data was binned every 1000 bp. The y-axis, representing per bin read coverage, has been constrained to 200, 350 and 8000 reads per bin for the MHO_001 chromosome, plasmid A and plasmid B respectively. The top panel represents the per base read coverage of nanopore long reads across the MHO_001 genome. The data was binned every 1000 bp. The y-axis, representing per bin read coverage, has been constrained to 20 reads per bin for each contig.

Supplementary Table 1. Table summarising the BLASR analysis of demultiplexed non-target sample 2D nanopore long reads and 2D fail reads in which no barcode was detected. Reads were aligned to the assembled MHO_001 reference genome using BLASR with default parameters. Only the alignment with the highest percentage similarity was considered for each read. The average alignment length was calculated from the length of the top BLASR alignment relative to the length of the input read.

Supplementary Table 2. Spreadsheet summarising the comparison between SNPs called by MAUVE alignment of assemblies created using long and short reads and SNPs called via mapping short reads to USA300_FPR3757.

Supplementary Figure 1. MAUVE alignment of the overlapping region included in the circularised single chromosomal contig aligned to USA300_FPR3757.

Supplementary Figure 2. MAUVE alignment of the overlapping region not included in the circularised single chromosomal contig aligned to USA300_FPR3757.

Supplementary Figure 3. CLUSTAL visualisation of the MUSCLE alignment between the two overlapping regions at the edge of the single chromosomal contig.

Supplementary Figure 4. Tablet visualisation of the nanopore long reads that span the overlapping regions at the edge of the circularised single chromosomal contig.

Supplementary Analysis. PCR and Sanger sequencing analysis of large structural variants SAPI5 and tRNA expansion.

Table

Click here to download Table Table1.pdf

	Pass	**Fail**
# Reads	1324	1499
# BLASR Hits (% # Reads)	1320 (99.70%)	1292 (86.19%)
Mean Alignment Length (%)	96.79	92.90
Mean Percentage Similarity (%)	85.87	77.76
# Hits < 75% Read Length (%)	11 (0.83%)	93 (7.20%)
# Hits >= 75% Read Length (%)	1309 (99.17%)	1199 (92.80%)

Figure 1

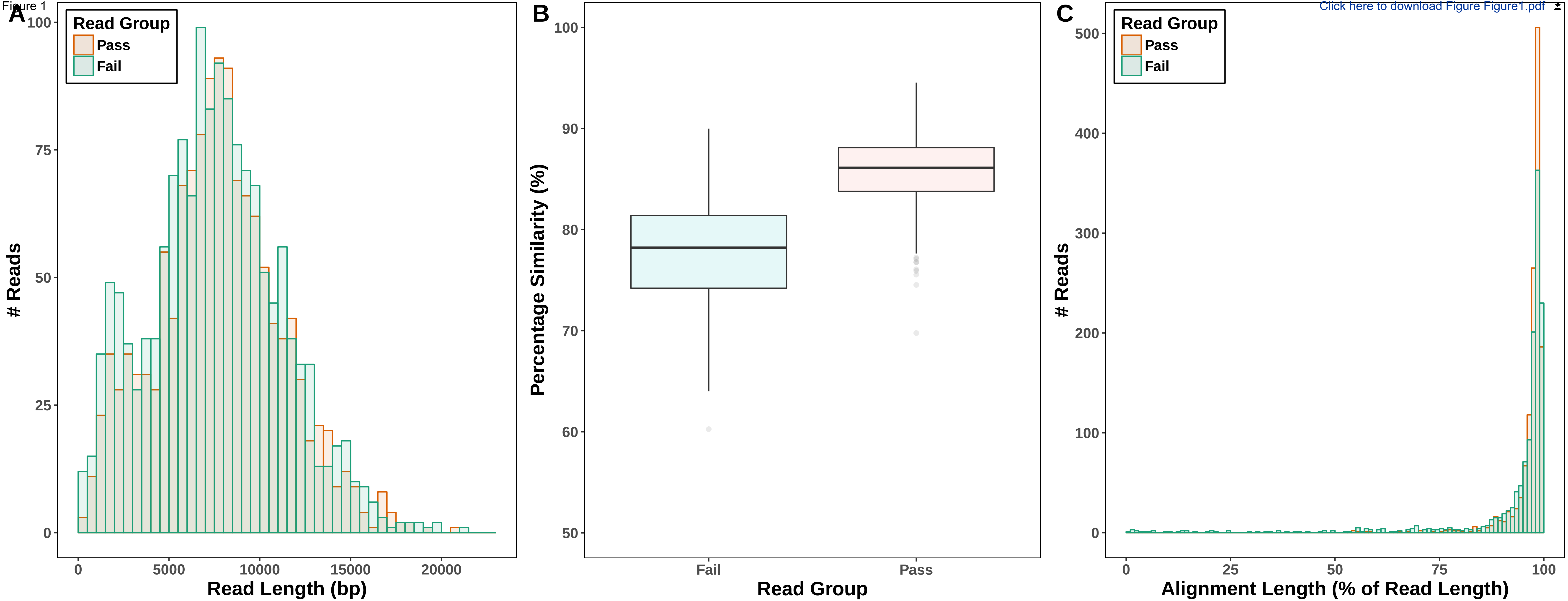
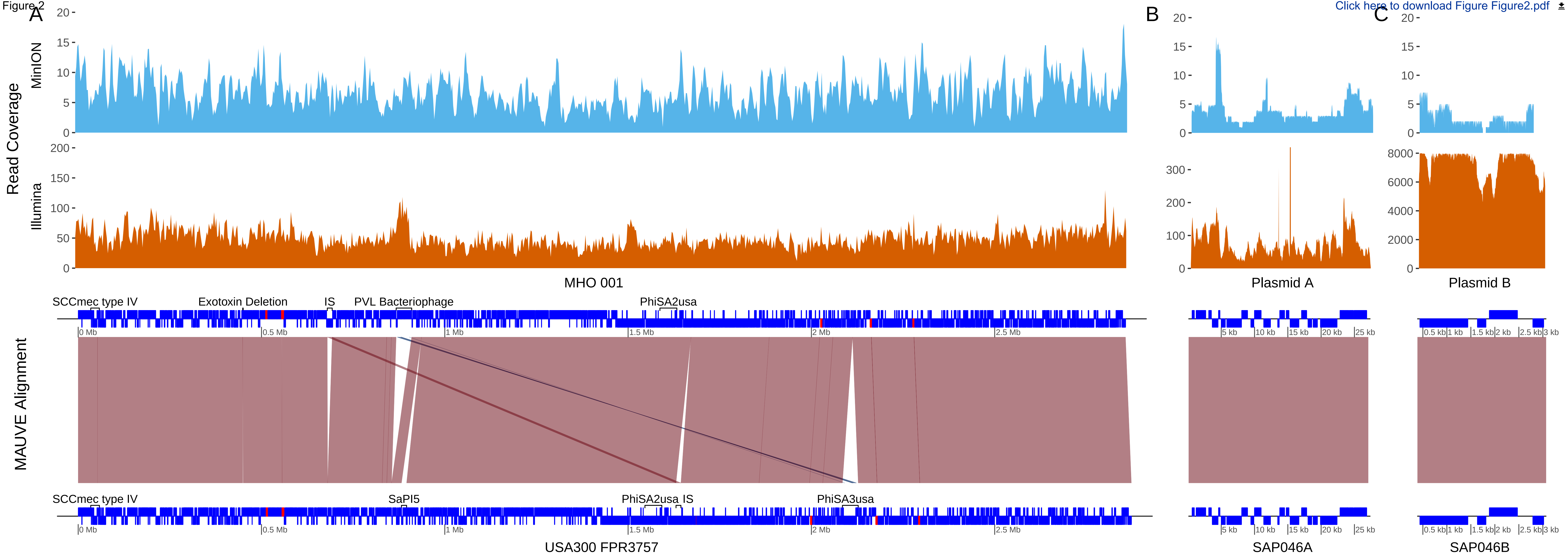

A
Read Group
Pass
Fail
# Reads
100
75
50
25
0
0
5000
10000
15000
20000
Read Length (bp)
B
Percentage Similarity (%)
100
90
80
70
60
50
Fail
Pass
Read Group
C
Read Group
Pass
Fail
# Reads
500
400
300
200
100
0
0
25
50
75
100
Alignment Length (% of Read Length)

Figure 2

Click here to download Figure Figure2.pdf



Supplementary Table 1

Supplementary Table 2

Supplementary Figure 1

Click here to access/download

Supplementary Material

Supp_Figure1.jpg

Supplementary Figure 2

Supplementary Figure 3

Supplementary Figure 4

Supplementary Analysis

Data Description

Response to Reviewers incl. Figures

Readme