Shallow MinION sequencing to assist *de novo* assembly of the *Streptococcus agalactiae* genome

Tamara Hernandez-Beeftink[1], Hector Rodriguez-Perez[1], Ana Díaz-de Usera[2], Rafaela Gonzalez-Montelongo[2], José M. Lorenzo-Salazar[2], Fabián Lorenzo-Díaz[1,3], Carlos Flores[1,2,4*]

[1]Research Unit, Hospital Universitario Ntra. Sra. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain; [2]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain; [3]Departamento de Bioquímica, Microbiología, Biología Celular y Genética, Universidad de La Laguna, Santa Cruz de Tenerife, Spain; [4]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain.

*Corresponding author

E-mail: cflores@ull.edu.es (CF)

Abstract

Despite the reduced read length, the so-called Next-Generation Sequencing (NGS) of second-generation has allowed rapid and complete genome characterization of many species. MinION (Oxford Nanopore Technologies), a portable third-generation NGS device, enables sequencing of long DNA fragments at low cost. Here we used a low-coverage MinION sequencing in combination with short-read NGS to improve genome assembly. We tested this possibility by using MinION R9.0 with Rapid 1D kit and MiSeq with >300X paired-end 300 bp reads (Illumina, Inc.) for the genome assembly of a *Streptococcus agalactiae* clinical isolate (2.2 Mb). With as few as 1,171 MinION reads that covered the genome at 2.4X (the longest read being 186 Kb long), the hybrid assembly combining MinION and Illumina reads increased the N50 by 4.9-fold compared to the assembly using Illumina data alone. Almost 50% of the genome was represented into a single contig (1.02 Mb). Besides, this allowed the full reconstruction of mobile elements, including a plasmid, and improved gene annotation. Taken together, our results support that shallow MinION sequencing combined with high-throughput second-generation NGS constitutes a cost-efficient strategy for the assembly of whole genomes.

Key words: Hybrid assembly; Nanopore; artefactual reads

Background

Next-Generation Sequencing (NGS) technologies have allowed sequencing and analysis Group B *Streptococcus* (GBS) genomes from different environments and organisms[1–6]. The presence of prophages and mobile elements in GBS was associated with the bacterial adaptation and ability to cause infections, likely enhancing their diversity and spread[7–10]. In fact, the estimates suggest that roughly 50% of *Streptococcus agalactiae* genome is grouped into islands of pathogenicity, where virulence genes and mobile elements are present[11]. *S. agalactiae* genome shows a complex organization and rearrangements, where the presence of a conserved backbone and 69 variable regions have been observed[12]. These observations justify the necessity of using *de novo* assembled genomes for better comparative bacterial studies and adequate tracking of mobile elements. However, the assembly of bacterial genomes based on short-read NGS technologies (so called second-generation) usually produce discontinuous sequences[13,14], often requiring the assistance from algorithms, alternative library approaches or sequencing technologies to provide sequence continuity.

Third-generation NGS technologies allow obtaining longer DNA reads from PCR-free DNA libraries in real time, constituting a solution for resolving repeats[15] and obtaining complete assembled genomes. In this context, the portability and reduced costs of sequencing with the MinION device (Oxford Nanopore Technologies) has facilitated the adoption of this technology in many distinct applications[16]. Besides its utility for the analysis of structural variants[17,18], SNP determination[19], cytosine methylation[20], transcriptomics[21], and in-field experiments in extreme environments[22], MinION has widely used to sequence small whole genomes from

virus (10.8 kb)[23], mitochondrial DNA (16 kb)[24], and bacteria (4.6 Mb)[25], allowing their assembly into single contigs when depth of coverages were above 20-40X[26]. Nevertheless, this technology continues to be challenged by its high error rate (>10%)[27], particularly in homopolymeric regions[28], and the generation of reads that fail to align against the target[29,30]. Because of this, hybrid assemblies leveraging the per-base quality of high-coverage (>200X) second-generation data with the bridging capacity of >10X depth-of-coverage third-generation data is nowadays perceived as the optimal choice for obtaining accurate assemblies[24,27,31–35]. Given the large variability in sequence throughput offered by this technology well below product specifications, here we tested the benefit of shallow MinION sequencing data in assisting the hybrid genome assembly of a *S. agalactiae* clinical isolate (genome size of 2.2 Mb).

Methods

Bacterial isolate, growing conditions and validation

We have previously examined the IMESag-rpsI mobile element distribution in 240 whole genomes of Group B *Streptococcus* isolates. The *S. agalactiae* HRC strain was subjected to a deeper examination to assist the characterization of the mobile element in that study[36]. *S. agalactiae* HRC is a serotype V representative strain belonging to the tetracycline-resistance CC1 lineage Tn916-1[37], which was isolated in 2009 from a 60-year-old woman with abdominal sepsis at the Hospital Ramón y Cajal (Madrid, Spain)[36].

The bacterial isolate was cultured in TSB (Tryptic Soy Broth, Becton Dickinson) liquid medium at 37ºC for 24 h. DNA extraction was performed with the

GenElute Bacterial Genomic DNA kit (Sigma-Aldrich) and was quantified on a Qubit 3.0 fluorometer using the dsDNA BR assay kit (Thermo Fisher Scientific). Its identity was first validated by 16S rRNA gene sequencing as follows. DNA was amplified by PCR with 30 cycles of 95°C for 20 s, 50°C for 30 s, and 72°C for 15 s, using the HotStarTaq master mix kit (QIAGEN), and the primers 1391R (5'-GACGGGCGGTGWGTRCA-3') and 27F (5'-AGRGTTYGATYMTGGCTCAG-3'). Amplified material was purified with ExoSAP-IT (Thermo Fisher Scientific), subject to Sanger sequencing with BigDye Terminator v3.1 Cycle Sequencing kit (Thermo Fisher Scientific), and products purified with DyeEx 2.0 Spin kit (QIAGEN), following the manufacturer's recommendations. Sequencing products were resolved and basecalled in a 3500 Genetic Analyzer (Thermo Fisher Scientific). A simple BLAST[38] search on the sequences obtained verified that the bacterial isolate had a 99% identity to *S. agalactiae*.

Illumina sequencing

DNA from *S. agalactiae* HRC (1 ng) was used to generate sequencing libraries with the Nextera XT kit (Illumina, Inc.) following the manufacturer's recommendations. A MiSeq instrument and a MiSeq Reagent kit V3 (Illumina, Inc.) was used for 300 base paired-end sequencing along with a spike-in of 20% of the PhiX Control v3 (Illumina, Inc.). Library construction and the sequencing experiment were performed at the Instituto Tecnológico y de Energías Renovables (ITER).

MinION sequencing

Rapid Sequencing kit (SQK-RAD001) (Oxford Nanopore Technologies) was used for library preparation starting from 200 ng of *S. agalactiae* HRC purified DNA.

Sequencing with MinION used SpotOn Flow Cell Mk I R9 Version (Oxford Nanopore Technologies) and followed manufacturer's recommendations, except that the experiment was left running only for 22 h.

Bioinformatics and statistical procedures

MiSeq Reporter Software v1.18.54 (Illumina, Inc.) was used to convert intensities to basecalls and generate the FASTQ file. MinION basecalling was performed with MinKNOW GUI 1.1.21 software (Oxford Nanopore Technologies) and all records were processed with poRe v0.21[39] to extract the obtained reads in FASTQ file format. Unicycler v0.4.1[40] was used with default parameters for hybrid *de novo* assembly of *S. agalactiae* HRC genome, combining Illumina and Nanopore datasets, which involved multiple rounds of short-read polishing. As a reference for comparisons, Unicycler was also used to assemble the Illumina data alone. Assembly comparisons were assessed with QUAST v4.0[41] using the *S. agalactiae* SS1 genome (NZ_CP010867.1) as the reference. The assembled contigs were visualized with Bandage v0.8.1[42]. Annotation of genes and elements in the contigs was done with IonGAP[43]. Mauve v2.4.0[44] was used to identify the insertion sites of the *S. agalactiae* HRC mobile elements in the assembly against the reference (NZ_CP010867.1).

In order to further explore raw MinION reads, per-read GC content distribution was first assessed to identify outliers (>3 SD from the mean). The existence of artefactual reads in the output[29] was evaluated through comparisons against the reconstructed *S. agalactiae* HRC genome, including the two characteristic elements identified. BLAST v2.6.0[38] alignments were used to isolate unaligned artefactual

reads. Bio.SeqUtils library of Biopython release 1.70 was then used to assess their differences in length (bp), GC content (GC%) and mean quality (Q) score against the reads aligning to the *S. agalactiae* HRC assembled genome. A non-parametric Mann-Whitney U-test was used to assess the significance in the comparisons. To test the possibility that particular pores were generating more artefacts than expected, as has been suggested elsewhere[29], unaligned reads were mapped back to the 512 nanopores of the Flow Cell and significance was tested assuming a Poisson distribution. Finally, to evaluate their impact in *de novo* assembly of *S. agalactiae* HRC genome, we then excluded all unaligned reads from the dataset, and used Unicycler under the same conditions to reassess the assemblies.

Results

Summary of sequencing results

Sequencing with Illumina resulted in 2,975,356 paired-end reads of 300 bp with Phred>30, which translates into 354X depth-of-coverage of *S. agalactiae* genome. In comparison, the MinION run proceeded with a total of just 276 active pores, providing 1,171 reads with a mean length of 4.5 kb and a Q score average of 10.8, and ranging from 3.8 to 25.4. While the largest proportion of reads accumulated in the 1.86-9.0 kb range (the N50 of the output was 9.2 kb) the output included reads as long as 186 kb (which represents almost 10% of the *S. agalactiae* HRC genome). The total throughput (5.2 Mb) theoretically translates into a 2.4X depth-of-coverage of the bacterial genome, which is well below product specifications, despite all initial quality controls of the platform were passed. Finally, the per-read GC content average of reads was 44.3% (ranging from 10.3% to 95.9%), noticeably larger than that of the reference sequence (35.5%; NZ_CP010867.1).

Assembly improvements by using MinION data

The Illumina data assembly generated 23 contigs (>500 bp) with a N50 of 147,211 bp, and the largest contig being 525,288 bp long. The length of the assembly from the Illumina data alone was very close to the *S. agalactiae* reference, accounting for a genome size of 2,144,732 bp, a GC content of 35.3%, and an aligning length against the reference of 2,049,610 bp (97.9% of consensus identity) (Table 1). In comparison, although the hybrid assembly including also combining the Illumina and MinION reads did not allow to obtain a fully resolved bacterial genome, mostly because of a major unresolved structure containing repetitive elements, rRNA operon copies and tRNAs (Fig 1), it reduced the number of contigs to less than a half (8 with >500 bp), having a N50 of 714,293 bp and the largest contig being 1,018,988 bp long. The hybrid assembly generated a total length of 2,159,060 bp, with 2,062,825 bp aligning against the reference (98.5% of consensus identity) (Table 1). Both assemblies also evidenced a small plasmid (2,491 bp) as an independent replicon, which perfectly aligned with the *Streptococcus oralis tigurinus* 2426 plasmid pST2426 (ASXA01000016.1), along with two other mobile elements detected in previously studies[36] (Fig 1). In terms of gene annotation, the hybrid assembly also outperformed the Illumina only assembly (2,153 vs. 2,135 genes, respectively) when compared to the reference (2,195 genes) (Table 1). Taken together, the small amount of reads provided by MinION (0.04% of total reads obtained for the bacterial isolate) allowed to increase the N50 of the genome assembly by 4.9-fold, to cover almost 50% of the bacterial genome in a single contig, and to improve the annotation of predicted genes in the assembly.

Characterization of unaligned MinION reads and impact on the assembly

A closer inspection of the MinION run records evidenced the existence of reads with large regions of limited sequence diversity (end to end) that were largely compatible with artifacts. Just by exploring the GC content of the MinION output, we detected 27 reads with large deviations (>3 SD from the mean) corresponding to GC contents below 7.2% or above 77.8%. The reads involved had a mean size of 716 bp, but were as large as 4.2 kb. However, in order to better identify and characterize the reads that were most likely artefactual, we used BLAST at 85% identity threshold to align all reads against the assembled genome. This allowed to identify 68 unaligned reads with a mean size of 2.0 kb (but as large as 21.6 kb) and a mean Q score of 6.7 (but as high as 22.9) (Fig 2). Only three of these unaligned reads (0.25% of total) aligned to sequences from other species or synthetic constructs according to a BLAST search (Table S1), most probably corresponding to experimental contaminants. Surprisingly, MinKNOW software only classified 20 of the 68 as failed based on internal manufacturer's specifications. Unaligned reads differed significantly from reads that aligned for the GC content (74.2 vs. 42.5%; $p= 2.5e^{-37}$), Q score (6.7 vs. 11.1; $p= 2.2e^{-21}$), and mean read length (2.0 vs. 4.6 kb; $p= 1.3e^{-11}$) (Fig 2). Strikingly, from the 276 active pores of the run, unaligned reads were obtained from 45, and two pores in particular experienced a significant accumulation of them (28.9%, $p<0.019$) (S1 Fig). Finally, a re-assessment of the hybrid assembly evidenced identical results with or without excluding unaligned reads, supporting that Unicycler was robust to their presence in the input. For simplicity, these results are not shown.

Discussion

Here we demonstrate that the use of very low amount of MinION reads combined with short-read sequencing data has the capacity to increase the continuity of

assembly contigs and plasmid isolation without manual intervention, having the potential to fully resolve bacterial genomes. This was illustrated with the assembly of *S. agalactiae* HRC genome based on as few as 2.4X MinION depth-of-coverage data, paving the way for obtaining cost-effective continuum whole-genome sequences as the nanopore technology throughput increases. While other species may have alternative requirements, to the best of our knowledge, this is the first study assessing the benefits of using MinION reads to assist the complete assembly of microbial genomes with coverages well below 10X (usually assisted with >14X)[14,23,33,34]. Although we were unable to fully resolve the assembly, our results are quite reassuring considering the small number of MinION reads used in the experiment (0.04% of total reads), allowing to obtain almost the 50% of the bacterial genome in a single contig. Besides, we were able to detect and fully reconstruct a plasmid, previously attained only with higher depth-of-coverage and longer MinION reads[45].

High quality noise reads generated by MinION have been described recently in the literature[29], consisting on reads that did not map to the target, lab contaminants, or spike-in sequences. Many of those appeared to be artefacts of the technology that were associated with specific active pores[29], to which some authors have previously attributed an insufficient read quality[30]. In this study, MinION generated around 5% of reads of variable quality that did not map to the target. We found that despite a few of them had partial identity with bacterial sequences that may indicate minor experimental contaminations as has been shown by others[15,46], most of them consisted in low-complexity sequences from end to end. We demonstrate that a number of them were generated from specific active pores during our experiment. Based on these evidences, we speculate that most of these reads are artefacts produced during the sequencing

process and that they are most likely independent from the basecaller being used. In this study, we demonstrate that they did not negatively affect the hybrid assembly process in the conditions assessed. However, their impact in other projected applications (such as shotgun metagenomics or whole-genome sequencing of eukaryotes) or algorithms other than Unicycler remain unknown. Solutions to reduce or identify and filter out these reads will need to be implemented.

With sufficient MinION yield providing enough depth-of-coverage, the standard analysis is based on long-read and error-tolerant assemblers such as Canu[47]. Because of the high error rate of this technology at the moment (10-20%)[27,46,48], and the problems resolving the homopolymers[28], the assembly process from MinION data alone typically necessitates of a number of error correction steps to produce high quality genome sequences[49]. Most importantly, this assembly process is computationally intensive[49,50]. Theoretically, MinION can produce many thousands of reads and a few dozen Gb per run[49]. However, in practice MinION outputs are highly variable translating sometimes in a few thousand reads (3,400 in Benitez-Paez et al[51]) and Mb scale outputs. In our hands, regular experiments are roughly in the scale of 100 Mb per run. This is because the number of active pores changes during the experiment (between 105 and 426 in Greninger et al[46]), below theoretical maximum in distinct runs, and the effective number of useful reads represents a fraction of the output[15,46]. Some studies have reported that a large proportion of the output (>50%) have failed to align against the target[30,52]. Besides, sudden crashes of the MinION software controller[14] or the computer connection failures[53] have been reported during a sequencing experiment. In addition, unofficial information provided by users in The Nanopore Community report extremely low occupancy of pores (<10%) during the

sequencing experiments with Rapid 1D kits, which have been suggested to be related to the kit production itself. These numbers are comparable to our observations in this study and all these issues have obvious consequences in the throughput of a MinION experiment. Therefore, although not the standard, the limited yield obtained by the MinION sequencing in this study is not uncommon. In this context, a hybrid assembly with a very shallow coverage of MinION data constitutes an advantage over the approaches based on MinION data alone for obtaining cost-effective high-quality continuous bacterial genomes. Besides being computationally faster, this hybrid option reduces the burden of generating the thousands of long reads that are necessary to obtain high quality bacterial genome assemblies with MinION data alone (>20X)[26].

Conclusions

We demonstrate that the combination of a small proportion of long reads with high-coverage short-read data constitutes a promising strategy to generate high quality, highly contiguous assemblies, here allowing to nearly complete the *S. agalactiae* HRC genome. Besides, we detected MinION reads consisting of low-complexity sequences compatible with artefacts of the technology but that did not affect the assembly.

Acknowledgments

Funded by Instituto de Salud Carlos III (PI14/00844; PI17/00610; FI17/00177) and Ministerio de Ciencia, Innovación y Universidades (RTC-2017-6471-1; MINECO/AEI/FEDER, UE) co-financed by the European Regional Development Funds, “A way of making Europe” from the European Union; and by the agreement OA17/008 with Instituto Tecnológico y de Energías Renovables (ITER) to strengthen

scientific and technological education, training, research, development and innovation in Genomics, Personalized Medicine and Biotechnology.

Authors' contributions

T.H.B.: performed experiments, data analysis, interpretation, and manuscript drafting; H.R.P., A.D.U., R.G.M., J.L.S.: performed experiments and data analysis; F.L.D.: performed experiments, study conception, interpretation, revising the manuscript critically; C.F.: Study conception and design, data analysis, interpretation, revising the manuscript critically and conceived the project.

The authors declare that there are no competing interests.

Data availability

All raw reads from MinION and Illumina are available from the SRA database (accession number(s) SRP141332, SRP141319).

References

1. Zubair S, de Villiers EP, Younan M, Andersson G, Tettelin H, Riley DR, et al. Genome Sequences of Two Pathogenic Streptococcus agalactiae Isolates from the One-Humped Camel Camelus dromedarius. Genome Announc. 2013;1. doi:10.1128/genomeA.00515-13

2. Zubair S, de Villiers EP, Fuxelius HH, Andersson G, Johansson K-E, Bishop RP, et al. Genome Sequence of Streptococcus agalactiae Strain 09mas018883, Isolated from a Swedish Cow. Genome Announc. 2013;1. doi:10.1128/genomeA.00456-13

3. Kropp KA, Lucid A, Carroll J, Belgrudov V, Walsh P, Kelly B, et al. Draft Genome Sequence of a Streptococcus agalactiae Strain Isolated from a Preterm Neonate Blood Sepsis Patient at the Royal Infirmary, Edinburgh, Scotland. Genome Announc. 2014;2. doi:10.1128/genomeA.00875-14

4. Areechon N, Kannika K, Hirono I, Kondo H, Unajak S. Draft Genome Sequences of Streptococcus agalactiae Serotype Ia and III Isolates from Tilapia Farms in Thailand. Genome Announc. 2016;4. doi:10.1128/genomeA.00122-16

5. Barony GM, Tavares GC, Pereira FL, Carvalho AF, Dorella FA, Leal CAG, et al. Large-scale genomic analyses reveal the population structure and evolutionary trends of Streptococcus agalactiae strains in Brazilian fish farms. Sci Rep. 2017;7: 13538.

6. Bodi Winn C, Dzink-Fox J, Feng Y, Shen Z, Bakthavatchalu V, Fox JG. Whole-Genome Sequences and Classification of Streptococcus agalactiae Strains Isolated from Laboratory-Reared Long-Evans Rats (Rattus norvegicus). Genome Announc. 2017;5. doi:10.1128/genomeA.01435-16

7. Otaguiri ES, Morguette AEB, Tavares ER, dos Santos PMC, Morey AT, Cardoso JD, et al. Commensal Streptococcus agalactiae isolated from patients seen at University Hospital of Londrina, Paraná, Brazil: capsular types, genotyping, antimicrobial susceptibility and

virulence determinants. BMC Microbiol. 2013;13: 297.

8. Fléchard M, Gilot P. Physiological impact of transposable elements encoding DDE transposases in the environmental adaptation of Streptococcus agalactiae. Microbiology. 2014;160: 1298–1315.

9. Morici E, Simoni S, Brenciani A, Giovanetti E, Varaldo PE, Mingoia M. A new mosaic integrative and conjugative element from Streptococcus agalactiae carrying resistance genes for chloramphenicol (catQ) and macrolides [mef(I) and erm(TR)]. J Antimicrob Chemother. Oxford University Press; 2017;72: 64–67.

10. van der Mee-Marquet N, Diene SM, Barbera L, Courtier-Martinez L, Lafont L, Ouachée A, et al. Analysis of the prophages carried by human infecting isolates provides new insight into the evolution of Group B Streptococcus species. Clin Microbiol Infect. 2017; doi:10.1016/j.cmi.2017.08.024

11. Glaser P, Rusniok C, Buchrieser C, Chevalier F, Frangeul L, Msadek T, et al. Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease. Mol Microbiol. 2002;45: 1499–1513.

12. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A. 2005;102: 13950–13955.

13. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. J Exp Bot. 2017; doi:10.1093/jxb/erx289

14. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing [Internet]. bioRxiv. 2017. p. 160614. doi:10.1101/160614

15. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nat Methods. 2015;12: 351–356.

16. de Lannoy C, de Ridder D, Risse J. The long reads ahead: de novo genome assembly using the MinION. F1000Res. 2017;6: 1083.

17. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W. Nanopore sequencing detects structural variants in cancer. Cancer Biol Ther. 2016;17: 246–253.

18. Stancu MC, van Roosmalen MJ, Renkens I, Nieboer M, Middelkamp S, de Ligt J, et al. Mapping And Phasing Of Structural Variation In Patient Genomes Using Nanopore Sequencing [Internet]. bioRxiv. 2017. p. 129379. doi:10.1101/129379

19. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17: 239.

20. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods. 2017;14: 411–413.

21. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, et al. Nanopore Long-Read RNAseq Reveals Widespread Transcriptional Variation Among the Surface Receptors of Individual B cells [Internet]. bioRxiv. 2017. p. 126847. doi:10.1101/126847

22. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre ABR, et al. Nanopore DNA Sequencing and Genome Assembly on the International Space Station [Internet]. bioRxiv. 2016. p. 077651. doi:10.1101/077651

23. Márquez S, Carrera J, Pullan ST, Lewandowski K, Paz V, Loman N, et al. First Complete Genome Sequences of Zika Virus Isolated from Febrile Patient Sera in Ecuador. Genome Announc. 2017;5. doi:10.1128/genomeA.01673-16

24. Chandler J, Camberis M, Bouchery T, Blaxter M, Le Gros G, Eccles DA. Annotated mitochondrial genome with Nanopore R9 signal for Nippostrongylus brasiliensis. F1000Res. 2017;6: 56.

381 25. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using
382 only nanopore sequencing data. Nat Methods. 2015;12: 733–735.

383 26. Sovic I, Krizanovic K, Skala K, Sikic M. Evaluation of hybrid and non-hybrid methods for
384 de novo assembly of nanopore reads [Internet]. bioRxiv. 2015. p. 030437.
385 doi:10.1101/030437

386 27. Salazar AN, Gorter de Vries AR, van den Broek M, Wijsman M, de la Torre Cortés P,
387 Brickwedde A, et al. Nanopore sequencing enables near-complete de novo assembly of
388 Saccharomyces cerevisiae reference strain CEN.PK113-7D. FEMS Yeast Res. 2017;17.
389 doi:10.1093/femsyr/fox074

390 28. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time,
391 portable genome sequencing for Ebola surveillance. Nature. 2016;530: 228–232.

392 29. Mojarro A, Hachey J, Ruvkun G, Zuber MT, Carr CE. CarrierSeq: a sequence analysis
393 workflow for low-input nanopore sequencing [Internet]. bioRxiv. 2017. p. 175281.
394 doi:10.1101/175281

395 30. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR.
396 Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a
397 eukaryotic genome. Genome Res. 2015;25: 1750–1756.

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

400 32. Batovska J, Lynch SE, Rodoni BC, Sawbridge TI, Cogan NO. Metagenomic arbovirus
401 detection using MinION nanopore sequencing. J Virol Methods. 2017;249: 79–84.

402 33. Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G, Blaxter M, et al. A single
403 chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION
404 nanopore sequencing data. Gigascience. 2015;4: 60.

405 34. Fournier T, Gounot J-S, Freel K, Cruaud C, Lemainque A, Aury J-M, et al. High-Quality
406 de Novo Genome Assembly of the Dekkera bruxellensis Yeast Using Nanopore MinION
407 Sequencing. G3 . 2017;7: 3243–3250.

408 35. Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien F-A, Swinkels W, et al. Rapid
409 de novo assembly of the European eel genome from nanopore sequencing reads. Sci Rep.
410 2017;7: 7213.

411 36. Lorenzo-Diaz F, Fernández-Lopez C, Douarre P-E, Baez-Ortega A, Flores C, Glaser P, et
412 al. Streptococcal group B integrative and mobilizable element IMESag-rpsI encodes a
413 functional relaxase involved in its transfer. Open Biol. 2016;6. doi:10.1098/rsob.160084

414 37. Da Cunha V, Davies MR, Douarre P-E, Rosinski-Chupin I, Margarit I, Spinali S, et al.
415 Streptococcus agalactiae clones infecting humans were selected and fixed through the
416 extensive use of tetracycline. Nat Commun. 2014;5: 4544.

417 38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool.
418 J Mol Biol. 1990;215: 403–410.

419 39. Stewart RD, Watson M. poRe GUIs for parallel and real-time processing of MinION
420 sequence data. Bioinformatics. 2017;33: 2207–2208.

421 40. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome
422 assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13: e1005595.

423 41. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome
424 assemblies. Bioinformatics. 2013;29: 1072–1075.

425 42. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo
426 genome assemblies. Bioinformatics. 2015;31: 3350–3352.

427 43. Baez-Ortega A, Lorenzo-Diaz F, Hernandez M, Gonzalez-Vila CI, Roda-Garcia JL,
428 Colebrook M, et al. IonGAP: integrative bacterial genome analysis for Ion Torrent

sequence data. Bioinformatics. 2015;31: 2870–2873.

44. Darling AE, Tritt A, Eisen JA, Facciotti MT. Mauve assembly metrics. Bioinformatics. 2011;27: 2756–2757.

45. George S, Pankhurst L, Hubbard A, Votintseva A, Stoesser N, Sheppard AE, et al. Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. Microbial Genomics. 2017;3. doi:10.1099/mgen.0.000118

46. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 2015;7: 99.

47. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation. Genome Res. 2017;27: 722–736.

48. Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. Gigascience. 2014;3: 22.

49. Jain M, Koren S, Quick J, Rand AC, Sasani TA, Tyson JR, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads [Internet]. bioRxiv. 2017. p. 128835. doi:10.1101/128835

50. Judge K, Hunt M, Reuter S, Tracey A, Quail MA, Parkhill J, et al. Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology. Microb Genom. 2016;2: e000085.

51. Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. Gigascience. 2016;5: 4.

52. Shin J, Lee S, Go M-J, Lee SY, Kim SC, Lee C-H, et al. Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. Sci Rep. 2016;6: 29681.

53. Zaaijer S, Columbia University Ubiquitous Genomics 2015 class, Erlich Y. Using mobile sequencers in an academic classroom. Elife. 2016;5. doi:10.7554/eLife.14258.

462 **Fig 1.** Schematic representation of the *S. agalactiae* HRC hybrid assembly. The
463 flanking sequences of the two characteristics elements of this isolate are indicated at the
464 bottom: 1) A conjugative element attached to the Tn916 transposon located in the
465 largest contig (1,018,988 bp); 2) A prophage located in the third largest contig (224,154
466 bp). Most copies of the rRNA (16S, 23S, 5S) and tRNA genes were positioned within
467 the largest unresolved assembly structure.

468

469

470 **Fig 2.** Distribution of MinION reads by quality (Q) scores and length (bp). Aligned
471 reads are represented in grey, whereas the 68 unaligned reads are overlaid in black. For
472 simplicity, the plot was capped to represent reads <60 kb length and Q scores <15.

473

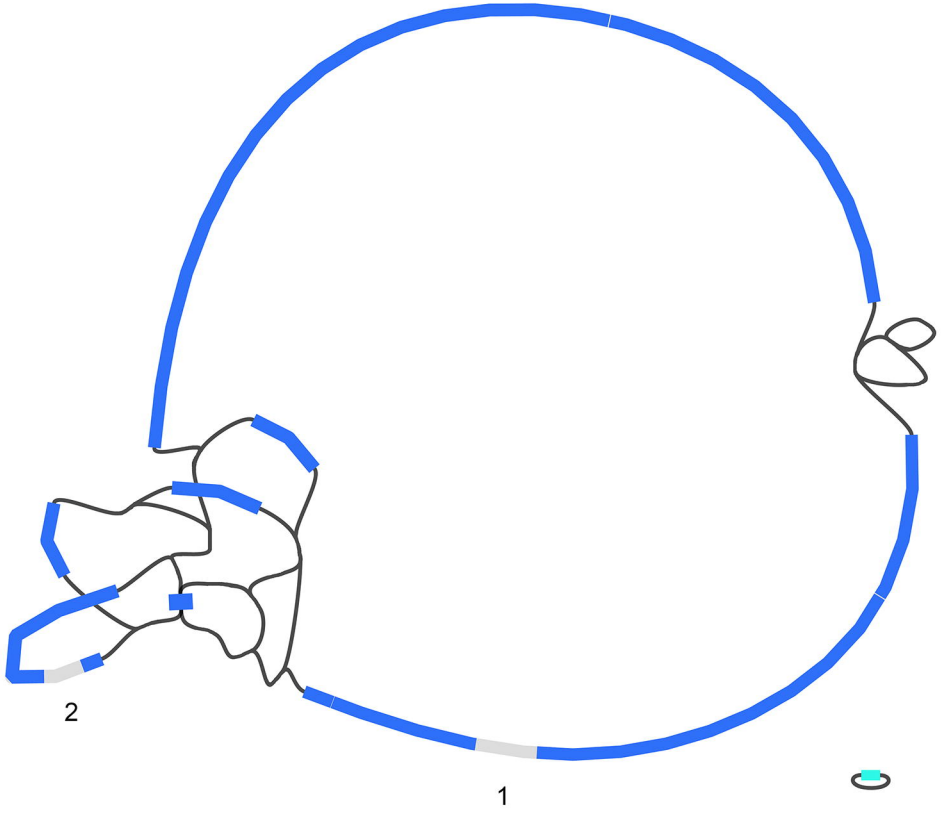
Table 1. Summary of the hybrid and Illumina-only assemblies of *S. agalactiae* HRC genome.

Parameters	**Illumina**	**Hybrid**
Total length assembled (bp)	2,144,732	2,159,060
Contigs (>=500 bp)	23	8
Largest contig (bp)	525,288	1,018,988
N50	147,211	714,293
GC content (%)	35.3	35.4
Genome fraction (%)*	97.9	98.5
Total aligned length (bp)*	2,049,610	2,062,825
Largest alignment (bp)*	524,861	1,018,010
Mismatches (#)*	151	193
Indels (#)*	33	77
Genes**	2135	2153
Coding DNA sequences**	2104	2117
tRNAs**	31	36

*Against the reference sequence: NZ_CP010867.1.

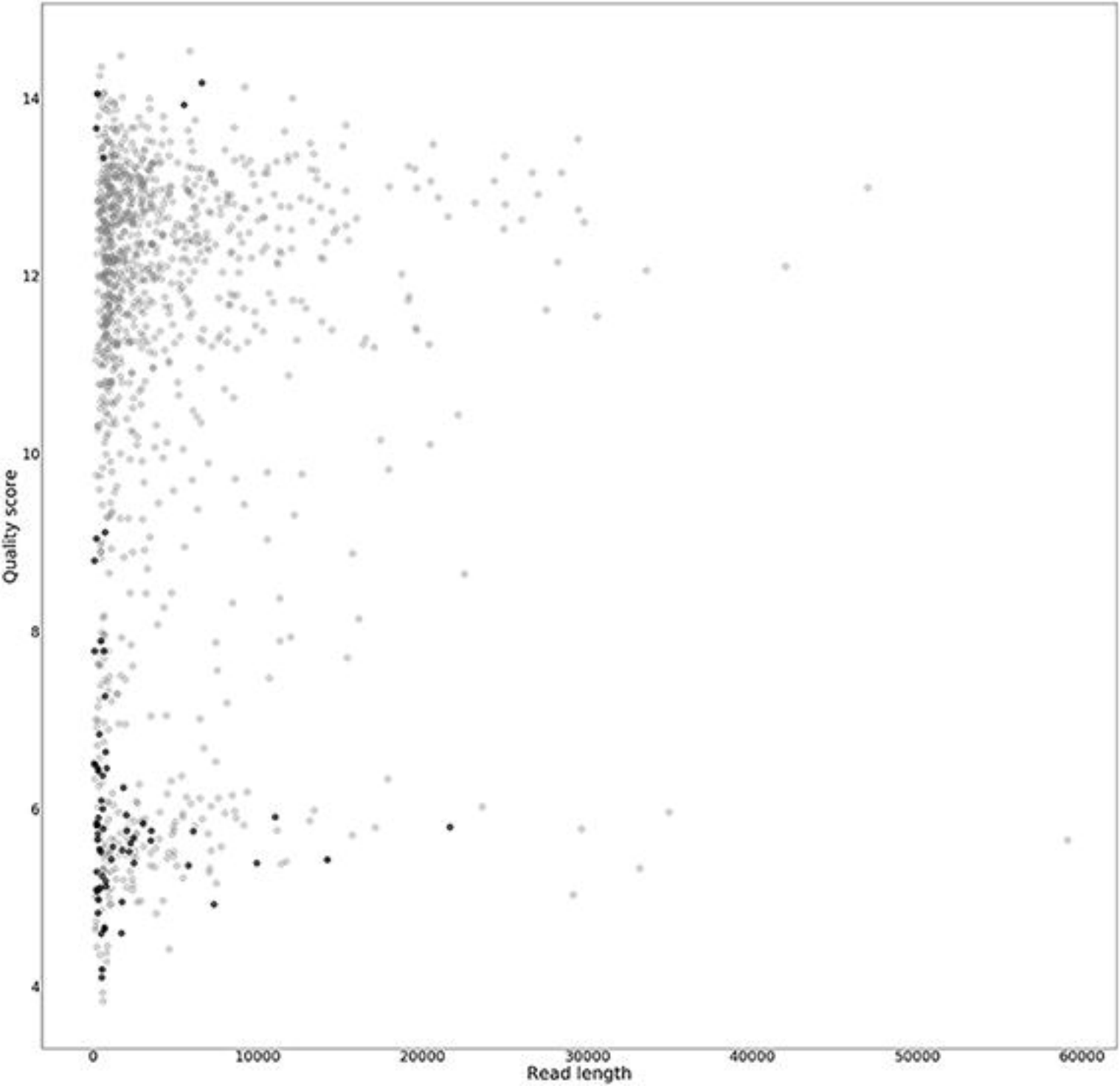
**For comparison, IonGAP predicted 2,195 genes, 2,115 coding DNA sequences, and 80 tRNAs in the reference.

474



1 …CAATGACACATCATGTCGAG***ACGGTAGCACTTTTGTCCAA…TGGACAATGATGAACAAGTT***GAGTGTGTTGCTCTGCTTGT…

2 …GCATTTACACTTTTAGAAAT***CAAGGATAGTAAATTTCTTT…TAATTTTAGTAACCTTGATT***GTGGTTAGTGTAGAGCCTTT…


Quality score
14
12
10
8
6
4
0
10000
20000
30000
40000
50000
60000
Read length