Publication status: Preprint has been published in a journal as an article
DOI of the published article: https://doi.org/10.1371/journal.pcbi.1010905

Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing

Ryan R. Wick, Louise M. Judd, Kathryn E. Holt

https://doi.org/10.1590/SciELOPreprints.5053

Submitted on: 2022-11-11
Posted on: 2022-11-11 (version 1)
(YYYY-MM-DD)

Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing

Ryan R. Wick[1*], Louise M. Judd[2] and Kathryn E. Holt[1,3]

1. Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Australia
2. Department of Microbiology and Immunology, University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia
3. Department of Infection Biology, London School of Hygiene & Tropical Medicine, London, United Kingdom

* rrwick@gmail.com

Abstract

A perfect bacterial genome assembly is one where the assembled sequence is an exact match for the organism's genome – each replicon sequence is complete and contains no errors of any scale. While this has been difficult to achieve in the past, improvements in long-read sequencing, assemblers and polishers have brought perfect assemblies within reach. Here we describe our recommended approach for assembling a bacterial genome to perfection using a combination of Oxford Nanopore Technologies long reads and Illumina short reads: Trycycler long-read assembly, Medaka long-read polishing, Polypolish short-read polishing, followed by other short-read polishing tools with manual curation. We also discuss potential pitfalls one might encounter when assembling challenging genomes, and we provide an online tutorial with sample data (github.com/rrwick/perfect-bacterial-genome-tutorial).

Authors' contributions

RRW (0000-0001-8349-0778): Conceptualization, Investigation, Methodology, Software, Visualization, Writing – Original Draft, Writing – Review & Editing

LMJ (0000-0003-3613-4839): Methodology, Writing – Original Draft, Writing – Review & Editing

KEH (0000-0003-3949-2471): Funding acquisition, Supervision, Writing – Review & Editing

Funding

This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation (KEH, grant number OPP1175797). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. This work was also supported by an Australian Government Research Training Program Scholarship (RRW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

Data availability

Sample data and scripts are available at: github.com/rrwick/perfect-bacterial-genome-tutorial

Keywords

bacterial genome assembly; Oxford Nanopore; long-read sequencing; hybrid assembly

Introduction

Compared to eukaryotes which have complex genomes often exceeding one billion base pairs (bp) in length, prokaryote genomes are small, typically containing a single circular chromosome a few million bp in length and often small extrachromosomal plasmids[1]. In many genomic applications, it would be most useful to know the bacterial genome sequence in its entirety, i.e. the full sequence of nucleotides for each piece of DNA in the cell. However, DNA sequencers work by fragmenting the genome and sequencing the fragments, producing reads: randomly ordered small pieces of the genome[2]. Reads are imperfect, with the frequency and type of errors depending on the platform. They are also redundant, because it is necessary to produce reads that total to many times the genome size to ensure that the genome is well covered. There is thus a disconnect between what sequencers provide (small, imperfect, redundant sequences) and what we want (a complete, error-free genome).

The solution to this problem is *de novo* assembly: the computational process of reconstructing a genome from sequencing reads. There are two broad goals to consider with genome assembly: accuracy and completeness. Accuracy refers to the number of errors present in the assembled sequences (contigs). Such errors can be small in scale (e.g. an incorrect base) or larger in scale (e.g. the addition/removal of hundreds of bases). Completeness refers to the length of the contigs relative to the corresponding genomic sequence, i.e. how fragmented the assembly is. Longer contigs are better, ideally each contig representing an entire replicon in the genome. We define a 'perfect' assembly as one which maximises both accuracy and completeness. A perfect assembly of a bacterial genome would contain one complete and error-free contig per replicon and no additional contigs.

Many downstream analyses do not require high-quality assemblies, e.g. one can identify the species of a genome or the presence/absence of a gene using a crude assembly[3]. There are, however, tasks which require extreme accuracy, e.g. estimating mutation rates and inferring transmission chains, where even a small number of errors can have consequences. Perfect assemblies offer no limits on their downstream uses, making 'is my assembly good enough?' an irrelevant question. In the absence of assembly errors, many analyses which involve interrogating reads directly (using computationally intensive approaches, e.g. variant calling) could be replaced by simpler assembly-based alternatives.

Here we describe and demonstrate a modern approach for producing a bacterial genome assembly with the goal of perfection using a combination of Oxford Nanopore Technologies (ONT) long reads and Illumina short reads (**Figure 1**). These platforms were chosen for their availability and widespread adoption in microbial genomics. While older hybrid assembly methods have used a short-read-first approach (building a short-read assembly graph and then scaffolding with long reads)[4], improvements in the yield and accuracy of long-read sequencing now mean that long-read-first hybrid assembly (making a long-read-only assembly and then polishing with short reads) can produce more accurate results[5], and that is the approach we use in this manuscript. We also provide an online tutorial (github.com/rrwick/perfect-bacterial-genome-tutorial) with sample data (hybrid sequencing of *Staphylococcus aureus* strain JKD6159[6]) so readers can try this method for themselves.

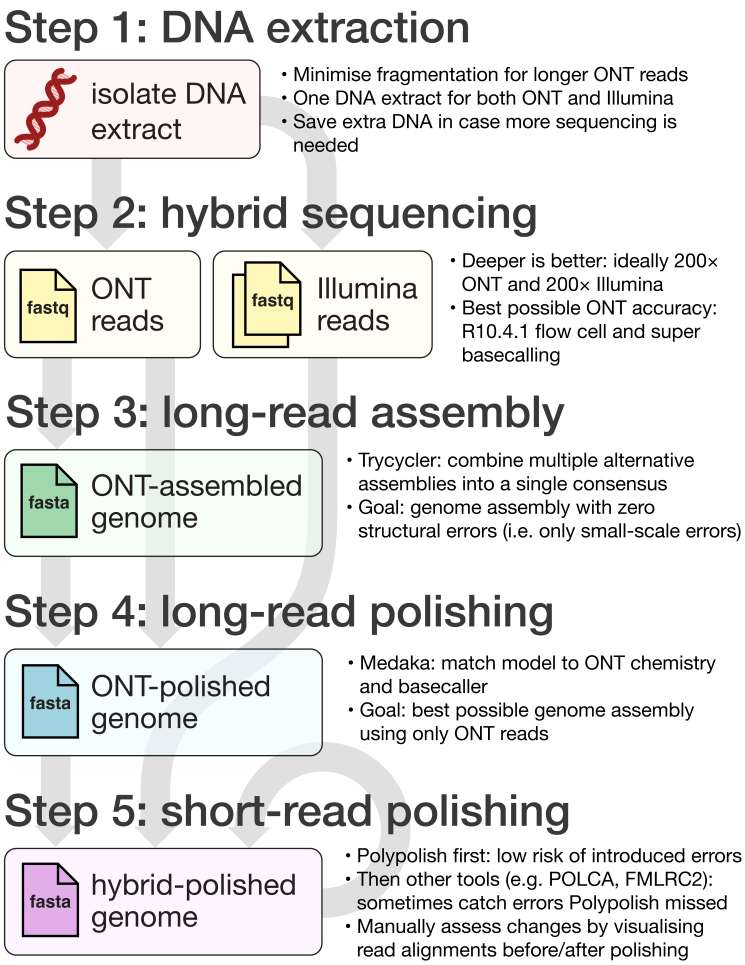

Figure 1: illustrated overview of our recommended approach to perfect bacterial whole genome assembly.

Step 1: DNA extraction

DNA should be extracted from a single bacterial colony to minimise the chance of genomic heterogeneity (see **Pitfalls**). While the best method for extracting DNA can vary by organism, one should aim to maximise purity and molecular weight. High purity will allow for better ONT yields, as chemical and biological impurities can damage or clog nanopores, shortening the life of flow cells[7]. High molecular weight will produce longer ONT reads, so one should avoid vortexing, minimise handling/pipetting and minimise freeze-thaw cycles to reduce shearing of DNA molecules[8]. Extraction methods for most bacteria should incorporate cell lysis by enzymatic digestion, using lysozyme (Sigma Aldrich, L6876) followed by proteinase K digestion (as provided in DNA extraction kits). This method is suitable for most Gram-negative and Gram-positive bacteria, but optimisation with additional enzymes may be required for difficult-to-lyse bacteria. Magnetic bead-based DNA extraction is recommended to reduce DNA shearing and maximise throughput. Recommended kits (in order of preference) are: GenFind V3 (Beckman Coulter, C34881) and MagAttract HMW DNA (Qiagen, 67563). For bacterial isolates that are difficult to lyse enzymatically, bead-beating can be used, but ONT read length may be compromised.

If culturing and DNA extraction is conducted multiple times (e.g. once for ONT sequencing and again for Illumina sequencing), there is the risk of genomic differences between the DNA samples[9]. This can lead to difficulties during polishing, so we recommend using a single DNA extract for all sequencing runs. It may also be prudent to freeze additional DNA or bacterial pellets in case further sequencing is later required.

Step 2: sequencing

Long-read ONT sequencing

One key consideration for ONT sequencing is depth: how deep must the ONT read set be? High read depth aids both assembly (allowing for more independent read sets in Trycycler, see **Step 3**) and polishing (yielding higher accuracy, see **Step 4**). When aiming for a perfect assembly, consider 100× depth to be a minimum, with 200× being ideal. Depths above 200× are better but will give diminishing returns. Using a single ONT flow cell for one bacterial isolate may provide excessive depth, so multiplexing is common in microbial genomics. This is not a problem for assembly, though barcode leakage should be considered (see **Pitfalls**).

Another consideration is length: how long must the ONT reads be? N50 length is a commonly used metric: the length-weighted median[10]. To ensure a complete assembly, the read set should have an N50 length greater than the longest repeat sequence. For many bacterial genomes, this is the rRNA operon, which is ~5 kbp and usually present in multiple copies[11], making an ONT read N50 of ~20 kbp a good target. In rare cases where the genome has an unusually long repeat (see **Pitfalls**), ultra-long DNA extraction protocols may be necessary[12].

ONT library preparation, chemistry and basecalling are also important factors. Both ligation-based and rapid preparations are appropriate for bacterial whole-genome sequencing, though ligation-based preparations can favour sequencing yield while rapid preparations can favour read length[13]. ONT currently offers MinION/GridION flow cells with two different pores: R9.4.1 (released in 2017) and R10.4.1 (released in 2022). The pores used in R10.4.1 flow cells are longer, improving homopolymer resolution and consensus accuracy, making them the better choice for assembly[14]. Basecalling, the computational process of translating the sequencer's raw signals into nucleotide sequences, is under constant development, so users should opt for the most recent version of ONT's recommended basecaller and use its highest-accuracy model. If users do not have an ONT sequencer with a GPU (e.g. a GridION), then access to a computer with a GPU will be required to perform basecalling.

After basecalling, QC filtering can improve the quality of the ONT reads. We recommend using Filtlong[15] to remove the worst reads (short length and low accuracy) with `--keep_percent 90`. If the read set has a poor N50 but is very deep, then removing short reads (e.g. <5 kbp) can help with assembly, though this may compromise small plasmid recovery (see **Pitfalls**).

Short-read Illumina sequencing

Since Illumina reads will only be used for final polishing (see **Step 5**), they carry less importance than ONT reads. Most current Illumina platforms produce similar data (e.g. 150-bp paired-end reads) and will function equally well for bacterial whole-genome sequencing, with instrument choice driven by cost and multiplexing needs. Nextera XT library preparations produce variable read depth (i.e. some regions of the genome may have low depth), so Illumina DNA Prep (a.k.a. Nextera DNA Flex) and TruSeq are preferable[16]. If Nextera XT is used, aim for a high mean depth (e.g. 300×) to compensate for depth variation, otherwise 100× should be sufficient. For highly repetitive genomes, mate-pair preparations may improve short-read polishing performance (see **Pitfalls**). After Illumina reads are produced, we recommend using a QC tool such as fastp[17] to remove low-quality bases and adapter sequences.

Step 3: long-read assembly

The goal of long-read assembly is to produce complete sequences with no structural errors, i.e. the only errors in the assembly should be small-scale, e.g. single-bp substitutions, insertions or deletions. This is because later polishing steps can repair small-scale errors but may not be able to fix larger structural errors.

Several long-read assemblers have been developed that are suitable for bacterial genomes, including Canu[18], Flye[19], NECAT[20], NextDenovo[21] and Raven[22], each of which uses different methods and thus has advantages/disadvantages. Regardless of the assembler used, most long-read bacterial genome

assemblies contain avoidable errors, and given the same read set, different assemblers are likely to produce assemblies with different errors[23]. Trycycler exploits this fact by building a consensus from multiple alternative assemblies of the same genome, allowing it to avoid structural errors, remove spurious contigs and ensure that circular sequences have no missing/duplicated bases at their ends[5]. We therefore recommend using Trycycler to produce long-read bacterial genome assemblies. However, note that Trycycler is not an automated tool – it requires human judgement and interaction.

Step 4: long-read polishing

This step aims to fix as many remaining errors as possible using only long reads. We recommend using Medaka[24], which we have found to produce more accurate results than Nanopolish[25,26]. Medaka uses a neural network and comes with trained models that correspond to specific combinations of ONT chemistry and basecaller, so one should choose the Medaka model which most closely matches their ONT reads. Alternatively, long-read variant callers such as Clair3[27] can be used as polishers by applying the called variants to the assembly.

Long-read polishing is done before short-read polishing because it is less influenced by genomic repeats. A 'repeat' in this context is a sequence which causes reads to align to multiple and/or incorrect positions of the genome. For example, some 150-bp short reads will be contained within the rRNA operon and will therefore align to multiple places, making the operon a repeat and impairing the ability of polishers to repair errors. With 20 kbp long reads, however, all can span the rRNA operon and therefore align uniquely, so the operon is not a repeat, ensuring that polishing changes occur in the correct instances of the operon.

Long-read polishing usually improves assembly accuracy, but a drop in accuracy is sometimes possible. It can therefore be unclear at this step whether the unpolished assembly, Medaka-polished assembly or some alternative (e.g. Clair3-polished) is best. ALE is a tool which quantifies the concordance between an assembly and a short-read set[28], allowing one to assess the relative accuracy of different assemblies. We therefore recommend using ALE to guide the decision regarding which version of the assembly should progress to the next step (short-read polishing).

Step 5: short-read polishing

The previous steps have generated a long-read-only assembly of maximal accuracy, likely ~Q50 (one error per 100 kbp) if R10.4.1 ONT reads were used. The final step is to repair any remaining errors with short reads. For example, long homopolymers can be difficult for ONT sequencing to resolve[14], but Illumina sequencing does not suffer from this problem[12,29], so homopolymer-length errors which persist after long-read polishing can be fixed by short-read polishing.

Our tool Polypolish[30] was designed with two goals in mind. The first was to use all-per-read alignments to overcome some of the constraints imposed by repeats. The second was to be very conservative, i.e. to minimise the chance of introducing errors during polishing. Polypolish only makes changes that are unambiguously supported by the read alignments, so when there are multiple possibilities at a locus (e.g. a base could be `A` or `C` with some alignments supporting each), Polypolish will not change the sequence. For this reason, we recommend running Polypolish before any other short-read polisher.

Due to its conservativeness, Polypolish may miss errors that other short-read polishers can fix, e.g. in regions of low Illumina depth. However, other polishers can introduce new errors[30], which is unacceptable when aiming for perfection. We therefore recommend trying other short-read polishers but manually assessing any and all changes using a tool such as IGV[31]. Viewing the read alignments at a particular locus before/after polishing can clarify whether the change fixed an error (in which case it should be retained) or introduced an error (in which case it should be rejected)[32]. Polishers to try

include POLCA[33] (due to its low rate of introduced errors) and FMLRC2[34] (due to its ability to fix errors other polishers cannot).

Automation

The above-described method requires human judgement and interaction, particularly at the Trycycler and short-read polishing steps. This allows users to catch unexpected results, ensuring that poor data does not proceed to the next step. This method is appropriate where accuracy is paramount (e.g. reference genome assembly), but it cannot be run in an automated manner (e.g. with Nextflow[35]) and is thus not suitable for high-throughput assembly.

If automation is required, changes in the workflow are needed. Flye[19] is less likely than other long-read assemblers to produce large-scale errors which downstream polishers may not be able to fix[23], making it a good replacement for Trycycler. Before polishing with Medaka, circular Flye contigs should be 'rotated' to a consistent starting sequence (e.g. *dnaA*[36]) or random starting sequence. This will serve to move any duplicated/missing bases at the start/end of circular contigs to the middle of the sequence where polishing tools can repair the error. For short-read polishing, we recommend Polypolish followed by POLCA, as these tools are the least likely to introduce errors[30].

Users should not assume that automated assemblies are error-free. In particular, structural errors (e.g. fragmented replicons, doubled plasmids, etc.) are possible, as these are what Trycycler aims to avoid.

Pitfalls

Small plasmids (<20 kbp) can be underrepresented in ONT read sets, due to either ligation preparations (where circular sequences fail to acquire adapters[13]) or overly aggressive QC (e.g. discarding all reads <10 kbp). This can be avoided by using rapid preparations and less stringent QC (e.g. only discarding reads <1 kbp). Alternatively, small plasmids can be recovered from an Illumina-only or short-read-first-hybrid assembly graph (e.g. from Unicycler[4]) where they usually appear as circular contigs separate from the rest of the genome (**Figure 2A**).

Some bacterial taxa have undergone proliferation of insertion sequence elements in their evolution, resulting in genomes with hundreds of 1–2 kbp repeats[37,38]. Perfect assembly of such genomes can be challenging because short-read polishers struggle to repair errors in high-copy-number repeats (**Figure 2B**). For this reason, it is crucial to maximise ONT-only accuracy (using high ONT depth, R10.4.1 pores, 'super' basecalling and Medaka polishing) to minimise the number of errors left for short-read polishing to fix. Additionally, mate-pair Illumina sequencing may enable Polypolish to fix errors within repeat sequences by reducing the number of ambiguous short-read alignments[39].

While the ~5 kbp rRNA operon is the longest repeat in many bacterial genomes, longer repeats are possible. For example, *Mycobacterium smegmatis* mc²155 contains a 56 kbp duplication in its chromosome[40]. In such cases, typical ONT read lengths (~20 kbp) can be insufficient for assembly and ultra-long reads (~100 kbp) are needed (**Figure 2C**).

Demultiplexing errors can occur in multiplexed sequencing runs: reads from one barcode can 'leak' into another, resulting in low-level contamination[41]. When a sequence in one barcode is very high depth, it may appear in other barcodes at sufficient depths to be assembled. This most often occurs with high-copy-number plasmids (**Figure 2D**), so when multiple genome assemblies from the same sequencing run contain identical plasmids, cross-barcode contamination should be considered as a possible cause.

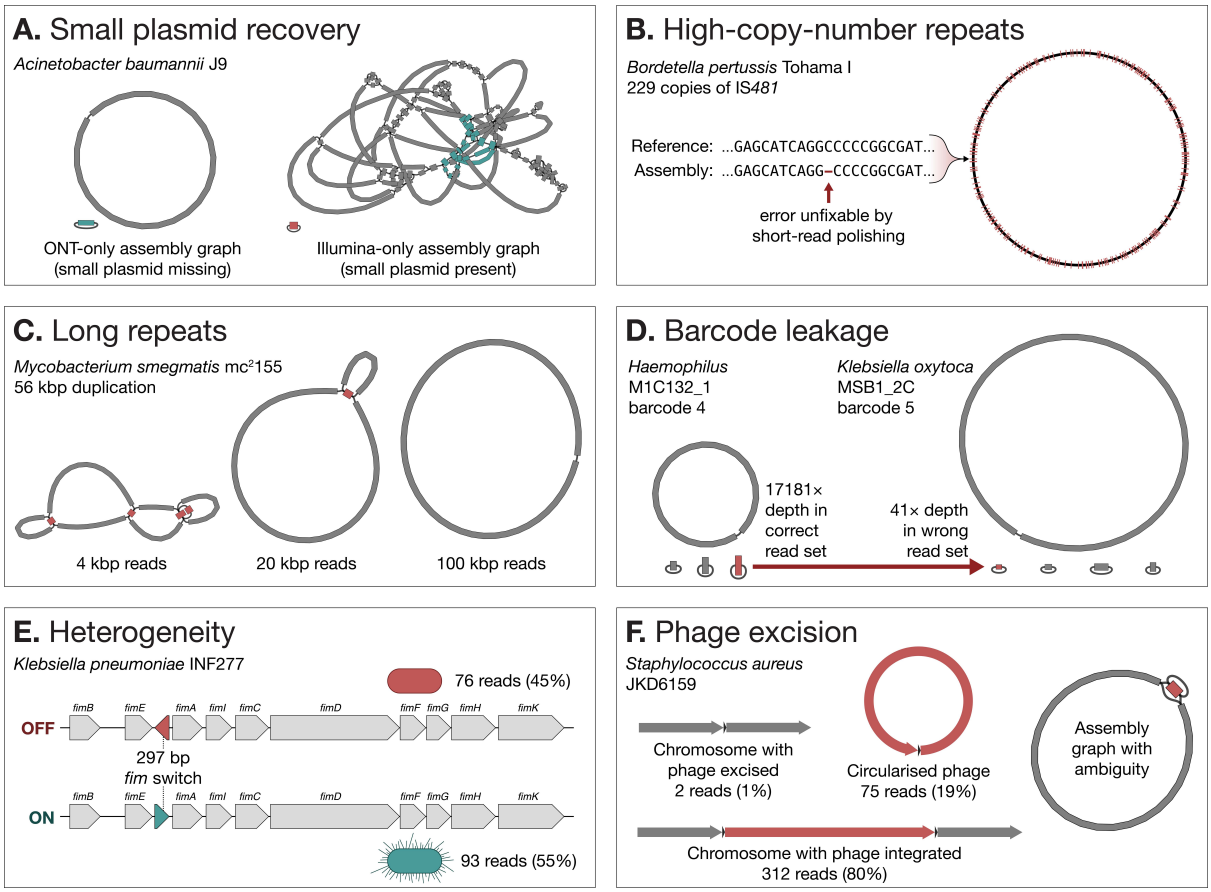

Figure 2: examples of pitfalls in bacterial genome assembly and polishing.

A. *A. baumannii* J9[13] contains one large 145 kbp plasmid (blue) and one small 6 kbp plasmid (red). The small plasmid is missing from an ONT-only assembly of this genome (left). However, it assembled completely in an Illumina-only assembly (right), enabling its recovery.

B. IS*481* is a repeat in the *B. pertussis* Tohama I genome[42]. Due to its high copy-number, some errors in this repeat are not fixable using paired-end Illumina reads and short-read polishers.

C. If a genome contains a very long repeat, as is the case with *M. smegmatis* mc²155[40], typical ONT read lengths of ~20 kbp may not be sufficient for complete assembly.

D. As occurred with *Haemophilus* M1C132_1 and *K. oxytoca* MSB1_2C[5], read demultiplexing errors can cause a deeply sequenced replicon in one genome (left) to erroneously appear in the assembly of another genome from the same sequencing run (right).

E. ONT sequencing of *K. pneumoniae* INF277[43] contained a near-50:50 mixture of *fim* switch orientations, causing problems during long-read and short-read polishing.

F. *S. aureus* JKD6159[6] read sets contained structural heterogeneity around the ΦSa3 bacteriophage sequence (left), causing an incomplete Flye assembly graph (right).

Heterogeneity occurs when there is not a single underlying genome but rather a mixture of two or more alternatives. This can occur at small scales (e.g. a mixture of different bases at a locus) or large scales (e.g. a mixture of structural configurations). When heterogeneity occurs at a low level (e.g. 95% of the reads support one sequence and 5% another), it does not typically cause problems as assemblers/polishers will use the more common alternative. However, balanced heterogeneity (e.g. a 50:50 mixture) can cause misassemblies and polishing mistakes. The phase variation of the *fim* switch is one cause of heterogeneity in *Enterobacteriaceae*[44] (**Figure 2E**). Another common example occurs with bacteriophages, which can integrate into and excise from bacterial chromosomes[45] (**Figure 2F**). Heterogeneity can be identified by incomplete assembly graphs and dense clusters of changes made by a polisher. It may then be necessary to manually exclude reads which support one alternative, allowing the other alternative to assemble/polish cleanly.

Conclusions

In contrast to short-read-first hybrid assembly approaches of the past (e.g. Unicycler), our recommended method follows a long-read-first paradigm. Due to their improved handling of repeats, long reads form a solid assembly foundation, with short reads only used for final polishing. Using this approach, we believe perfect genome assemblies with zero errors are achievable. However, it is not easy to establish a ground truth genome sequence, so when assembly accuracy is critical, we recommend performing multiple alternative assemblies that vary in data/methods: sequencing platforms, assemblers in the Trycycler pipeline, read QC thresholds, short-read polishing tools, etc. When alternative data/methods produce identical assemblies, this builds confidence in their correctness. When alternative assemblies are not identical, further investigation (e.g. visualising read alignments in IGV) is warranted.

While perfect bacterial genome assemblies are now possible, they are not yet simple to produce. The future will undoubtedly bring improvements to ONT chemistry, basecallers and polishers, but whether these will be sufficient for perfect ONT-only assemblies (negating the need for Illumina reads) remains to be seen. Further software developments are needed to remove the human-interaction elements, enabling perfect assemblies from a fully automated pipeline, even in complicated cases (e.g. genomes with heterogeneity). The ultimate goal is a future where genomes can be assembled to perfection with enough ease and reliability that it is taken for granted.

References

1. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15(2):141–61.
2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
3. Foster-Nyarko E, Cottingham H, Wick RR, Judd LM, Lam MMC, Wyres KL, et al. Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, *Klebsiella pneumoniae*. bioRxiv. 2022;
4. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, editor. PLOS Comput Biol. 2017;13(6):e1005595.
5. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Méric G, Vezina B, et al. Trycycler: consensus long-read assemblies for bacterial genomes. Genome Biol. 2021;22(1):266.
6. Chua K, Seemann T, Harrison PF, Davies JK, Coutts SJ, Chen H, et al. Complete genome sequence of *Staphylococcus aureus* strain JKD6159, a unique Australian clone of ST93-IV community methicillin-resistant *Staphylococcus aureus*. J Bacteriol. 2010;192(20):5556–7.

7. Maghini DG, Moss EL, Vance SE, Bhatt AS. Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome. Nat Protoc. 2021;16(1):458–71.

8. Branton D, Deamer DW. Nanopore Sequencing: An Introduction. World Scientific Publishing Company; 2019.

9. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genomics. 2017;3(10).

10. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020;38(9):1044–53.

11. Espejo RT, Plaza N. Multiple ribosomal RNA operons in bacteria; their concerted evolution and potential consequences on the rate of evolution of their 16S rRNA. Front Microbiol. 2018;9:1232.

12. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45.

13. Wick RR, Judd LM, Wyres KL, Holt KE. Recovery of small plasmid sequences via Oxford Nanopore sequencing. Microb Genomics. 2021;7(8).

14. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022;19(7):823–6.

15. Wick RR. Filtlong. 2021.

16. Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, et al. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Res. 2019;26(5):391–8.

17. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.

18. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.

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

20. Chen Y, Nie F, Xie SQ, Zheng YF, Dai Q, Bray T, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun. 2021;12(1):60.

21. Hu J. NextDenovo. 2021.

22. Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. Nat Comput Sci. 2021;1(5):332–6.

23. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Research. 2019;8(2138).

24. Wright C, Wykes M. Medaka.

25. 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.

26. Wick RR. Perfecting bacterial genome assembly. Monash University; 2022.

27. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. :14.

28. Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics. 2013;29(4):435–43.

29. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. 2011;12(11):R112.

30. Wick RR, Holt KE. Polypolish: short-read polishing of long-read bacterial genome assemblies. Schneidman-Duhovny D, editor. PLOS Comput Biol. 2022;18(1):e1009802.

31. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.

32. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant review with the Integrative Genomics Viewer. Cancer Res. 2017;77(21):e31–4.

33. Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. Ouzounis CA, editor. PLOS Comput Biol. 2020;16(6):e1007981.

34. Mak QC, Wick RR, Holt JM, Wang JR. Polishing *de novo* nanopore assemblies of bacteria and eukaryotes with FMLRC2. bioRxiv. 2022;

35. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35(4):316–9.

36. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol. 2015;16(1):294.

37. Register KB, Sanden GN. Prevalence and Sequence Variants of IS*481* in *Bordetella bronchiseptica*: Implications for IS*481*-Based Detection of *Bordetella pertussis*. J Clin Microbiol. 2006;44(12):4577–83.

38. Hawkey J, Monk JM, Billman-Jacobe H, Palsson B, Holt KE. Impact of insertion sequences on convergent evolution of *Shigella* species. Hughes D, editor. PLOS Genet. 2020;16(7):e1008931.

39. Wetzel J, Kingsford C, Pop M. Assessing the benefits of using mate-pairs to resolve repeats in *de novo* short-read prokaryotic assemblies. BMC Bioinformatics. 2011;12(1):95.

40. Wang XM, Galamba A, Warner DF, Soetaert K, Merkel JS, Kalai M, et al. IS*1096*-mediated DNA rearrangements play a key role in genome evolution of *Mycobacterium smegmatis*. Tuberculosis. 2008;88(5):399–409.

41. Wick RR, Judd LM, Holt KE. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. Pertea M, editor. PLOS Comput Biol. 2018;14(11):e1006583.

42. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, et al. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet. 2003;35(1):32–40.

43. Wyres KL, Hawkey J, Mirčeta M, Judd LM, Wick RR, Gorrie CL, et al. Genomic surveillance of antimicrobial resistant bacterial colonisation and infection in intensive care patients. BMC Infect Dis. 2021;21(1):683.

44. Schwan WR. Regulation of *fim* genes in uropathogenic *Escherichia coli*. World J Clin Infect Dis. 2011;1(1):17.

45. Fogg PCM, Colloms S, Rosser S, Stark M, Smith MCM. New applications for phage integrases. J Mol Biol. 2014;426(15):2703–16.

This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.