Benchmarking Genome Assemblers for Four Bacterial Models Based on Contiguity, Correctness, and Completeness

Hanzel Rojas-Miranda[1], Vanessa Madrigal-Ly[2] and Jose Arturo Molina Mora[1,]*

[1] Facultad de Microbiología y Centro de Investigación en Enfermedades Tropicales (CIET), Universidad de Costa Rica, Costa Rica.

[2] Universidad Nacional de Costa Rica, Costa Rica.

Emails: hansel.rojas@ucr.ac.cr, vanessa.madrigally@una.ac.cr, jose.molinamora@ucr.ac.cr

* Corresponding author: jose.molinamora@ucr.ac.cr; ORCID: 0000-0001-9764-4192

ABSTRACT

De novo genome assembly allows for the genome reconstruction of an organism without using a reference sequence. Assembly results depend on various sequencing technologies that generate data with differing fidelity, read lengths, and coverage levels, as well as on performance of a wide variety of algorithms. These attributes generate a diversity of assemblies for each single genome, which collectively defines the pan-assembly. In this study, we aimed to benchmark pan-assemblies of the prokaryotic models *Brucella henselae, Escherichia coli, Pseudomonas aeruginosa,* and *Xylella fastidiosa*, using different attributes and their impact on metrics of the 3C (contiguity, correctness, and completeness) criterion for selecting the best conditions for *de novo* genome assembly. Results showed that short-read assembly strategies presented higher accuracy with fewer errors (high correctness) and a high degree of completeness but lower contiguity due to fragmented assemblies. In contrast, long-read-based strategies showed high contiguity but lower completeness and accuracy. The hybrid strategy yielded the best overall results across all parameters by leveraging the strengths of both types of technology. Regarding assembly algorithms, Unicycler was the top assembler in 3C metrics, using any of the short-read (compared to Megahit), long-read (compared to Canu), or hybrid strategies (compared to Wengan). Overall, the hybrid approach with Unicycler proved to be the best general approach for genome assembly of the four bacterial models. Finally, regarding coverage depth, increasing depth did not significantly affect assembly quality results if a minimum data level was maintained, indicating that high-quality assemblies can be achieved using moderate coverage levels. Jointly, the results of the pan-assembly provide working conditions for *de novo* genome assembly that can be applied to bacterial models of interest, guiding the selection of optimized experimental and bioinformatics conditions while reducing sequencing costs for generating high-quality sequences.

Keywords: Pan-assembly, prokaryotes, *de novo* genome assembly, 3C criterion, benchmarking

INTRODUCCION

Whole genome assembly is the reconstruction of a nucleotide sequence that represents the actual genome from an organism, in which fragmented DNA pieces on average covered multiple times (or raw sequencing read data) are used [1], [2]. When no prior knowledge of the source DNA sequence is assumed, a strategy called *de novo* genome assembly is considered. This approach is essential for investigating species in which a reference genome is unavailable or is not considered representative of the target genome due to a spectrum of genetic variations [3].

In *de novo* assembly, sequencing reads are assembled into consensus sequences named "contigs", that together represents most of the genome at a first level of assembly [2]. Several graph-based algorithms are available for this purpose, including Overlap-Layout-Consensus, de Bruijn and greedy approaches [4], [5]. A subsequent level is the assembly of scaffolds. Scaffolding aims to bridge the gaps between the contigs using experimental data (for example mate pairs and paired end sequencing) or a reference genome [1], in which nucleotides of unsolved regions are represented with letter "N". When all gaps are solved (gap closing) by generating longer contigs [6] or Polymerase Chain Reaction [7] for example, a complete genome is assumed as a final level of assembly.

Regarding the generation of reads by sequencing technology, strategies can be separated based on the read length in short- or long-read sequencing. In short-read sequencing, such as Illumina's instruments (second-generation sequencing), reads can be produced with a length of up to 600 bases, accuracy >99.9% and cost-effective [8]. In contrast, classical long-reads or third-generation strategies, such as Pacific Biosciences' (PacBio) single-molecule real-time sequencing and Oxford Nanopore Technologies' (ONT) sequencing, reads are obtained with >10 kb and 85-90% accuracy with a higher cost [3], [8]. Advances in PacBio technology have consolidated the HiFi sequencing method, which yields highly long-read sequencing with accuracy >99.5% [9] but remains comparatively expensive. Recent advances in ONT sequencing technology have significantly enhanced both its hardware and data analysis pipelines, reducing error rates to approximately 6% [10]. Continuous methodological improvements are further increasing ONT accuracy; however, certain applications—especially those requiring single-nucleotide resolution—still benefit from the complementary use of short-read sequencing [10], [11].

Notwithstanding these technologies have been widely used to determine the DNA sequence of thousands of genomes, a diversity of candidate genome sequences can be obtained depending not only on the experimental procedures but also on the bioinformatic pipelines [5], [12]. These parameters include genome complexity (repeats, number of chromosomes, mobilome), sample

conditions, DNA extraction protocol, reads length and sequencing technology, sequencing depth (number of times that a given nucleotide has been read in an experiment), algorithm to assemble sequence (assemblers), databases, and others [1], [13]. In addition, criteria to benchmark assemblies and selecting a winner are also complicated, which depends on the aim of the study [14], [15]. Due to all this, to *de novo* genome assembly is not straightforward and is still a very classical and challenging problem in bioinformatics [2], [6].

To provide some insights to continue overcoming these challenges, we previously proposed the 3C criterion to compare and assess *de novo* assemblies based on *Contiguity* (pieces of the assembled genome), *Correctness* (fidelity of the assembly) and *Completeness* (ability to assemble expected genes) using different assemblers with short- and long-reads for a bacterial model [14], [16]. In line with the features of each technology and previous works [1], [5], [6], [9], [17], assemblies only using short reads (Illumina) resulted with a high correctness and completeness (not considering HiFi sequencing or improved ONT), while best values for contiguity were obtained for long-reads. Best values for all metrics were obtained when both strategies were used at the same time in hybrid assemblies.

As a proof of concept, in this current work we extend the assessment of a collection of assemblies for a specific organism, a concept that we here termed “pan-assembly”, to select the best assembly based on the 3C criterion. For this purpose, we selected four prokaryotic models in which short- and long-reads were used to assemble the genome with six assemblers (two short-reads only, two long-reads only, and two hybrid assemblers), and different levels of sequencing depth of coverage. The impact of these conditions on several 3C parameters was quantified. Thus, the aim of the study was to benchmark pan-assemblies of the prokaryotic models *Brucella henselae, Escherichia coli, Pseudomonas aeruginosa,* and *Xylella fastidiosa*, using different attributes and their impact on metrics of the 3C (contiguity, correctness, and completeness) criterion for selecting the best conditions for *de novo* genome assembly.

METHODS

With the aim of providing conditions for generating a high-quality assembly for five bacterial models, a comparative assembly approach was implemented with three strategies (data source: short-reads, long-reads, or both), with two algorithms per strategy and different sequencing depth levels.

Data source and pre-processing

Sequencing data were retrieved from Sequence Read Archive database (SRA-NCBI, https://www.ncbi.nlm.nih.gov/sra) for four bacterial models: *Bartonella henselae, Xylella fastidiosa, Escherichia coli,* and *Pseudomonas aeruginosa* (Table 1). Data were selected based on the FDA-ARGOS framework (https://www.fda.gov/medical-devices/science-and-research-medical-devices/database-reference-grade-microbial-sequences-fda-argos) or similar projects.

Quality control was performed with FASTQC v.0.11.9 tool [18]. Trimmomatic v.0.39 [19] was used for eliminating low-quality bases (Q<30) and adapters.

Estimation of sequencing depth (also called depth or coverage, or simply coverage) was achieved by first mapping reads to the corresponding a reference sequence (from NCBI) with the BWA-MEM 0.7.5a-r405 software [20]. Then, alignment was examined with Qualimap 2.3 platform [21] to obtain several metrics, including sequencing depth.

Pan-assembly of bacterial genomes

A standardized bioinformatics protocol was developed to assemble and annotate genomes for each bacterium. AnalysEs were run using the High-Performance Computing Cluster of the Center for Research in Materials Science and Engineering (CICIMA-HPC), University of Costa Rica. Raw sequencing data were subsampled based on sequencing depth (Table 1) using Seqtk 1.4 software (https://github.com/lh3/seqtk). Thus, trimmed reads subsets were generated with a sequencing depth of 12.5X, 25X, 50X, 100X, 200X* and 400X* (*: when possible). Depth was classified as low when <100X, medium for 100X and high for >100X.

Using trimmed data in each level of sequencing depth, two algorithms were implemented for each of the approaches using short-reads only, long-reads only, and hybrid (both short- and long-reads). The short-reads were processed with Unicycler v0.4.7 [22] and Megahit v1.1.3 [23]. Assemblers for long-reads were Unicycler v0.4.7, and Canu v1.8 [24]. Finally, Wengan [25] and Unicycler v0.4.7 were implemented for hybrid assemblies. Jointly, all assemblies in the different conditions for a single genome were considered the pan-assembly.

Metrics related to assembly quality were calculated with QUAST 5.2.0 tool [26]. Gene prediction (structural genome annotation) was performed using Prokka v1.13.3 [27], and results (GFF files) were used to compare assemblies based on gene content (similar to pan-genome analysis) using Roary v3.12.0 [28].

Benchmarking of pan-assembly using the 3C criterion

The 3C criterion (Contiguity, Completeness and Correctness), defined previously in [14] was used to compare genome sequences in each pan-assembly. For this purpose, a variety of metrics were selected, as follows:

- Contiguity or number of assembled segments: the total number of fragments or contigs, N50 value (shortest contig length that needs to be included for covering 50% of the genome and other metrics were obtained using QUAST 5.2.0 tool [26].
- Completeness: the ability to assemble expected genes was assessed using the number of predicted genes by Prokka v1.13.3 [27], as well as the completeness score based on analysis of orthologs with BUSCO v5.4.7 [29] within the gVolante platform [30].
- Correctness: the fidelity of the assembly was estimated based on rates of mismatches and insertions/deletions of the assembly with respect to the reference sequence. Calculations were obtained for the analysis with QUAST 5.2.0 tool [26].

In addition, assessed conditions for each pan-assembly (depth level, assembly approach and algorithm/assembler) were compared and used to select the parameters for the reconstruction of each genome with the best quality possible. For this purpose, tests of statistical significance were performed with R v4.3.1 software (www.r-project.org/) using RStudio interface (www.rstudio.com). Statistical analyzes were run with parametric and non-parametric tests, as appropriate, including ANOVA tests or Kruskal–Wallis tests, T or U tests, multiple linear regression or generalized linear models (see Results). Additionally, a Hierarchical clustering and Principal Component Analysis were performed in R software to study each pan-assembly based on all metrics of the 3C criterion.

RESULTS

Sequencing data obtained by short- and long-reads for four bacterial models were used to generate *de novo* genome assemblies. The pan-assembly was established for each model using three strategies (based on a single technology separately or in a hybrid mode), two algorithms per strategy, and several levels of sequencing depth. The number of assemblies obtained for the organisms was 67 for *B. henselae*, 76 for *E. coli*, 69 for *P. aeruginosa*, 89 for *X. fastidiosa*.

Metrics of the 3C criterion were calculated for each assembly, including those obtained from the comparison against the correspondent reference (consensus) sequence. Distribution of values for all assemblies evidenced a non-gaussian pattern, in which the median was used as measure of

central tendency, as well as non-parametric tests (including multivariable models) for comparisons. Six metrics (two for each 3C category) were used as key parameters for in-depth comparisons: number of contigs, N50, mismatches/100kbp, indels/100kbp, BUSCO score, and number of CDS. Using the whole set or selected metrics, pan-assembly was then described depending on sequencing strategy, assembler, and sequencing depth.

Regarding sequencing strategy, clear patterns of segregation depending on technology were found for all the models. This assessment of each pan-assembly was first done using clustering, based on the similarity among the whole set of 15 metrics of the 3C criterion (details in Supplementary file). For *X. fastidiosa*, shown in Figure 1-A-B, hierarchical clustering and PCA were implemented. These analyses not only assess the variation of metrics among assemblies, but also which sequences were more or less similar to each other. It is observed that clusters are defined by the sequencing strategy and then by algorithm for the assembly (assembler), rather than depth sequencing. In this line, distribution of values for the six key 3C metrics defines a particular pattern depending on the strategy, as shown in Table 2 and Figure 2. Fragmented genome and lower N50 (lower contiguity) are reported for short-reads only, with median values of 90 and 101 890 respectively, in contrast to long-reads only (2 contigs and N50=2 513 184) or hybrid approaches (4 contigs and N50=1 441 411). In correctness, mismatches and indels are more variable with higher values for long-reads and hybrid methods. A higher completeness, based on BUSCO score and appropriate CDS number, was measured for short-reads only assemblies.

For *B. henselae*, *E. coli*, and *P. aeruginosa*, similar results were obtained during the clustering analysis (Supplementary Figures 1-2-3) and median comparison among strategies, shown in Table 2 for each pan-assembly. Interestingly, for all the four models, the number of CDS were higher and BUSCO score were lower under long-reads only approaches, unlike other methods which remained homogeneous and closer to the values of the reference sequence.

In the comparison of assemblers, differences were revealed even using the same sequencing strategy. For *X. fastidiosa*, results for Megahit and Unicycler (short-reads) were similar among clustering analysis (Figure 1-A-B) and key 3C metrics (Figure 3). However, differences in values and dispersion were evidenced using long-reads (Canu vrs Unicycler) or Hybrid (Wengan vrs Unicycler), indicating that the algorithm influences the final assembly. This appreciation was also verified for the other prokaryotic models, as shown in Supplementary Figures 1-2-3.

Besides, conditions of those assemblers belonging to the vicinity of the reference sequence were considered as the relevant parameters to obtain the expected sequence. In *X. fastidiosa*, these

conditions were the use of Unicycler as assembler under a hybrid strategy (Figure 1-A-B). This is supported by the high value for the explained variance of 80.7% (PC1+PC2) in the PCA analysis. The profile of long-reads only approaches (with Canu or Unicycler), as well as a few hybrid cases with Wengan, resulted with the more discordant profiles. Again, proximity to the reference sequence is not associated with the sequencing depth for short- or long-reads used for the assembly (Figure 1-A).

For the other bacterial models (Supplementary Figures 1-2-3), hybrid approaches with Unicycler also resulted closer to the reference than other assemblies. Moreover, unlike *X. fastidiosa,* short-reads approaches tended to present a more different profile of 3C metrics from reference when compared to the long-reads approaches. In all cases, these patterns were provided with high support for the explained variance of the PCA, with 72.9%, 83.5% and 78% for *B. heneselae*, *E. coli*, and *P. aeruginosa*, respectively.

On the other hand, a comparative genomics approach was implemented to describe the pan-assembly based on gene content. As found in Figures 4 and 5 for the four models (blue lines: presence of the gene in each assembly), the well-defined clusters suggest that the gene content profile is established -again- by sequencing strategy and assembler, but independent on sequencing depth (after a minimal value). For *X. fastidiosa* (Figure 4), gene fragmentations are identified for long-reads only and some hybrids with Wengan (when sequencing depth for short-reads is 12.5X). Gene fragmentations for long-reads, as well as some specific hybrid cases with Wengan, are also present in the other models (Figure 5). In *P. aeruginosa*, loss of genes is evidenced when using Wengan even for >100X depth for short-reads, besides no assemblies were established for lower levels with the same assembler. Unlike this case, using Unicycler in a hybrid mode, sequences were built using at least 25X depth for short-reads.

A final assessment of the parameters studied here (predictors: strategy, assembler, and sequencing depth) and their effect on the key 3C metrics (response variable) was done using generalized linear models (GLM) for the pan-assembly of each organism. As summarized in Table 3 for all the four bacteria, significant values (<0.01) were obtained for the very most cases of the 3C metrics using strategy and assembler as predictors. Only scarce significant values were evidenced for sequencing depth for short- or long-reads among response conditions.

Finally, the subsequent analysis was conducted to determine the minimal sequencing depth to achieve an assembly with a similar quality to assemblies obtained with depth ≥100X data under the same strategy and assembler (Table 4). Using short-reads only approaches, requirements of

depth for Unicycler and Megahit were always the same for all the bacterial models (12.5X for *E. coli* and *X. fastidiosa*, and 25X for others). A similar situation was observed for long-reads only approaches, in which 25X is enough to assemble the sequence for all bacteria but *X. fastidiosa* (with 12.5X). In the case of hybrid approaches, Wengan was demonstrated to always need more sequencing depth in comparison to Unicycler, including requirements of >100X depth for short-reads (*B. henselae* and *P. aeruginosa*). In *B. henselae*, Wengan also demanded 100X depth for long-reads. When sequencing depth ≥25X (in some cases even with lower values), hybrid Unicycler was always able to assemble the sequence.

Jointly, whole results of pan-assembly analyses indicate that the assemble sequence and the 3C metrics are significantly impacted by sequencing strategy and assembler, but not by sequencing depth after a minimal level is considered for the four bacterial models.

DISCUSSION

De novo genome assembly allows for the genome reconstruction of an organism without using a reference sequence [31]. However, the assembly results depend on various sequencing technologies that generate data with differing fidelity, read lengths, and coverage levels, as well as on performance of a wide variety of assemblers (algorithms) [13], [32]. In this study, we compared pan-assemblies under these conditions and their impact on *de novo* genome assembly for four bacterial models, using the 3C criteria -contiguity, correctness, and completeness- for evaluation.

Regarding contiguity, our results indicate that short reads present significant variability in the number of contigs, suggesting high fragmentation in the resulting assemblies and reduced N50 values (low contiguity), in contrast to when long reads are used either alone or in hybrid formats. Thus, the use of long-read technology improves contiguity and potentially leads to better genome reconstruction under this criterion, even allowing for the potential circularization of bacterial genomes, as previously demonstrated [22], [33]. In another benchmarking strategy, the performance of long reads was also outstanding, with a low number of contigs and high N50 values [34]. However, a detailed inspection of the fragments revealed assembly errors that were not apparent when only contiguity was assessed. These findings have also been highlighted in other comparative studies [35], [36]. This underscores the need for diverse metrics to evaluate assemblies, as we propose with the 3C criteria [14], [16].

In the context of correctness or fidelity evaluation, significant differences were reported in the error rates for the four bacterial models. Short reads demonstrated a lower incidence of errors compared to long reads. In other studies, using Illumina and Oxford Nanopore data for various bacterial models, the results support our findings, with a low error rate detected when working with short-read data, including regions containing repeats [14], [35], [36].Thus, in applications where fidelity is a priority, short reads may be preferable to minimize specific errors such as those caused by indels and technical sequencing errors [33], [37].

Regarding completeness, the best performance was identified in assemblies with short reads and hybrid strategies. This effect was also evident in the pan-assembly analyses based on CDS gene content, with significant reductions mainly in hybrids assemblies using Wengan and low-coverage conditions. Although long reads can span more complex and broader genomic regions, gene prediction is incomplete due to assembly errors, a characteristic of these strategies with lower fidelity [38]. These findings reflect the capability of short reads to generate assemblies with high completeness, despite inherent limitations such as raw data length and low contiguity [33].

For hybrid strategies, the high completeness is associated with the resolution of ambiguities in genomic regions characterized by repetitive or highly complex sequences provided by long reads [39], [40]. In other works, the best results are justified by the resolution achieved with long assembled fragments [37], [41], although this contrasts with other cases where long reads are associated with lower completeness and gene identification due to frameshifts from sequencing errors and their impact on gene prediction [14]. This latter effect is also evidenced by the high number of CDS (compared to the reference sequence) found for long reads but not for other strategies.

Due to the lower fidelity of long-read strategies, the need to continue optimizing long-read sequencing techniques and minimizing errors without compromising assembly integrity is emphasized [42]. Notable advances in this area include the implementation of polishing strategies for assemblies generated from long-read data, which have significantly improved assembly completeness [33], [43]. In parallel, the development of high-fidelity long-read sequencing technologies—such as PacBio HiFi [9], [38], [44] —and optimized ONT platforms incorporating new flow cells and deep learning–based data processing have further enhanced sequencing accuracy and reliability [10], [11]. These two scenarios were not included in this study.

Regarding algorithms, this study observed statistically significant differences in the performance of assemblers for genome reconstruction. Their effect was evident in the size of the contigs produced (N50 and number of contigs), the number of expected genes reconstructed or completeness (BUSCO Score and CDS), fidelity or correctness (indels and mismatches), and the resolution of pan-assemblies based on gene content. Overall, the choice of algorithm has direct implications for the quality and utility of the assembled genome [5], [45], [46]. Additionally, as evidenced here, using exactly the same input data within the same assembly strategy, fidelity results increase for certain algorithms [47].

Considering all parameters, assemblies using Unicycler performed best in each strategy for the bacterial models studied. As supported by the literature in various benchmarking studies, Unicycler is currently one of the algorithms with the best reported performance for bacterial genome assembly, whether using long reads, short reads, or a hybrid mode [14], [35], [43], [48]. In studies with only long reads, Unicycler also stood out as the top-performing algorithm [35], though in other studies, Canu has been reported as performing better [46], [47], [48]. In our study, Canu showed good values in contiguity and moderate accuracy and completeness, but its performance did not surpass that of Unicycler. However, it is worth noting that Canu is optimized for maximizing performance in cluster computing environments [40]. Regarding the use of Megahit for assemblies with only short reads, it has been reported as a high-performing assembler compared to other strategies [48], [49], [50], even comparable to Unicycler, mainly in terms of completeness and accuracy. A strength of Megahit is its execution time, which is much faster compared to Unicycler. For hybrid assemblies, the other algorithm used was Wengan. In this study, it performed well but depended on high coverage to produce high-quality assemblies, unlike Unicycler in hybrid mode. This observation contrasts with a previous report where Wengan was highlighted for generating quality assemblies even in low-coverage situations [25]. In other metrics, Wengan has been reported with superior performance compared to other algorithms [25], [46]. However, Wengan showed suboptimal performance in contiguity, with low N50 values in another study [38]. In summary, Unicycler showed the highest concordance with reference genome data compared to other algorithms, regardless of whether long reads, short reads, or hybrid modes were used. The hybrid mode demonstrated the best overall performance.

Lastly, the final evaluation of the pan-assembly focused on the possible correlation between quality of the assembly and coverage depth. The results suggest that there is no statistical correlation between coverage depth (after a minimum level) and 3C criteria metrics, with a few exceptions. These results are consistent with several reports in the literature regarding this parameter in genome assembly analyses, showing no association between increased depth (in both long and short reads) and improved assembly quality [43], [51], [52]. One study highlighted that values of 40X do not change the results in cases of lower coverage, and ultra-deep sequencing can lead to algorithm saturation and a significant increase in computational resource usage [43]. It has also been reported that very low coverage levels of 16x are sufficient to assemble a complete genome, but accuracy improves with values around 30X [46]. In other studies, the minimum coverage level has even been reported as low as 10X [51], [53]. These minimum values are very similar to our results for the four organisms studied. It should be noted that in certain applications, such as genetic variant imputations, there are recommendations for a minimum coverage level for decision-making [53], [54], but these are not focused on *de novo* genome assembly. From a practical perspective, this aspect of coverage depth is highly relevant because it means that high-quality assemblies can be generated using moderate coverage levels, without the need for ultra-deep sequencing (>100X), translating into significantly lower sequencing costs.

In summary, once a minimum data level is reached, coverage depth does not significantly affect assembly quality results across the various characteristics evaluated by the 3C criteria, for which no statistical correlation was observed. This contrasts with the influence of strategy (technology) type and assembly algorithms on the quality of the resulting assembly. These results are based on genome sequencing, and our previous works on other molecular strategies [55], [56], will be considered to continue working on these bacterial models at the local biological context.

It should be mentioned that this study has some limitations. This study focused on four bacterial models, each with their own genomic complexity, but it could be extended to other models, including eukaryotic organisms in further analyses. Other methodological strategies, such as genome polishing or high-fidelity long-reads data obtained using HiFi sequencing ot enhanced ONT platforms, were not assessed in this comparison. Additionally, other 3C criteria metrics, execution time and computational resource usage could be valuable for comparing assembly conditions in other studies.

CONCLUSIONS

In this study, the pan-assembly of four bacterial models (*B. henselae, E. coli, P. aeruginosa,* and *X. fastidiosa*) was assessed using contiguity, accuracy, and completeness metrics (the 3C criteria). The benchmarking strategy showed that short-read assemblies presented higher accuracy with fewer errors (high correctness) and a high degree of completeness but lower contiguity due to fragmented assemblies. In contrast, long-read-based strategies showed high contiguity but lower completeness and accuracy. The hybrid strategy yielded the best overall results across all parameters by leveraging the strengths of both types of technology. Regarding assembly algorithms, Unicycler was the top assembler in 3C metrics, using any of the short-read (compared to Megahit), long-read (compared to Canu), or hybrid strategies (compared to Wengan). Overall, the hybrid approach with Unicycler proved to be the best general approach for genome assembly of the four bacterial models. Finally, regarding coverage depth, increasing depth did not significantly affect assembly quality results if a minimum data level was maintained, indicating that high-quality assemblies can be achieved using moderate coverage levels. In summary, these results of the pan-assembly provide working conditions for *de novo* genome assembly that can be applied to bacterial models of interest, guiding the selection of optimized experimental and bioinformatics conditions while reducing sequencing costs for generating high-quality sequences.

Author Contributions

J.A.M.M. participated in the conception and design of the study. H.R.M., V.M.L. and J.A.M.M. performed experimental assays and data analysis. J.A.M.M. drafted the manuscript. All authors were involved in its revision and the final approbation of the manuscript.

Funding

This work was funded by projects “C1163 pro-NGS 2.0: Protocolos operativos estandarizados de análisis de datos moleculares obtenidos por NGS o afines y de algoritmos de inteligencia artificial en modelos biológicos”, Vicerrectoría de Investigación, Universidad de Costa Rica (period 2021–2023) and “C4604 iPAT: Plataforma genómica, bioinformática y de inteligencia artificial para la vigilancia de patógenos”, Vicerrectoría de Investigación, Universidad de Costa Rica (period 2024–2026).

Informed Consent Statement

Not applicable.

Data Availability

The raw sequencing data (short and long reads), as well as the assembled genomes used in this study, are available in GenBank and SRA databases (NCBI), with the accession number detailed in Table 1.

Conflicts of Interest

The authors declare no conflicts of interest.

REFERENCES

[1] J. T. Simpson and M. Pop, "The Theory and Practice of Genome Sequence Assembly," *Annu Rev Genomics Hum Genet*, vol. 16, pp. 153–172, 2015, doi: 10.1146/annurev-genom-090314-050032.

[2] B. Segerman, "The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases," *Front Cell Infect Microbiol*, vol. 10, p. 527102, Oct. 2020, doi: 10.3389/FCIMB.2020.527102/BIBTEX.

[3] Y. Chen, Y. Zhang, A. Y. Wang, M. Gao, and Z. Chong, "Accurate long-read de novo assembly evaluation with Inspector," *Genome Biol*, vol. 22, no. 1, pp. 1–21, 2021, doi: 10.1186/s13059-021-02527-4.

[4] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithm for Next-Generation Sequencing data," *Genomics*, vol. 95, no. 6, pp. 315–327, 2010, doi: 10.1016/j.ygeno.2010.03.001.Assembly.

[5] F. Dida and G. Yi, "Empirical evaluation of methods for de novo genome assembly," *PeerJ Comput Sci*, vol. 7, p. e636, Jul. 2021, doi: 10.7717/PEERJ-CS.636.

[6] S. Schmeing and M. D. Robinson, "Gapless provides combined scaffolding, gap filling, and assembly correction with long reads," *Life Sci Alliance*, vol. 6, no. 7, Jul. 2023, doi: 10.26508/LSA.202201471.

[7] R. Beigel, N. Alon, M. S. Apaydin, L. Fortnow, and S. Kasif, "An optimal procedure for gap closing in whole genome shotgun sequencing," *Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB*, pp. 22–30, 2001, doi: 10.1145/369133.369152.

[8] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil, "Opportunities and challenges in long-read sequencing data analysis," *Genome Biol*, vol. 21, no. 1, Feb. 2020, doi: 10.1186/S13059-020-1935-5.

[9] T. Hon *et al.*, "Highly accurate long-read HiFi sequencing data for five complex genomes," *Sci Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1038/S41597-020-00743-4.

[10] C. Delahaye and J. Nicolas, "Sequencing DNA with nanopores: Troubles and biases," *PLoS One*, vol. 16, no. 10, p. e0257521, Oct. 2021, doi: 10.1371/JOURNAL.PONE.0257521.

[11] N. Lerminiaux, K. Fakharuddin, M. R. Mulvey, and L. Mataseje, "Do we still need Illumina sequencing data? Evaluating Oxford Nanopore Technologies R10.4.1 flow cells and the Rapid v14 library prep kit for Gram negative bacteria whole genome assemblies," *Can J Microbiol*, vol. 70, no. 5, pp. 178–189, 2024, doi: 10.1139/CJM-2023-0175/SUPPL_FILE/CJM-2023-0175SUPPLA.DOCX.

[12] H. Lantz *et al.*, "Ten steps to get started in Genome Assembly and Annotation," *F1000Res*, vol. 7, 2018, doi: 10.12688/f1000research.13598.1.

[13] A. Bellec, A. Courtial, S. Cauet, and N. Rodde, "Long Read Sequencing Technology to Solve Complex Genomic Regions Assembly in Plants," *Journal of Next Generation Sequencing & Applications*, vol. 3, no. 2, 2016, doi: 10.4172/2469-9853.1000128.

[14] J.-A. Molina-Mora, R. Campos-Sánchez, C. Rodríguez, L. Shi, and F. García, "High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers," *Sci Rep*, vol. 10, no. 1, p. 1392, Dec. 2020, doi: 10.1038/s41598-020-58319-6.

[15] J. Wang *et al.*, "Systematic Comparison of the Performances of De Novo Genome Assemblers for Oxford Nanopore Technology Reads From Piroplasm," *Front Cell Infect Microbiol*, vol. 11, p. 1, Aug. 2021, doi: 10.3389/FCIMB.2021.696669/FULL.

[16] J. A. Molina-Mora and F. Garcia, "The 3C criterion: Contiguity, Completeness and Correctness to assess de novo genome assemblies.," *BMC Bioinformatics, Bioinformatics: from Algorithms to Applications*, vol. 21, no. S20: O7, p. 5, Dec. 2020, doi: 10.1186/s12859-020-03838-2.

[17] A. V. Zimin and S. L. Salzberg, "The SAMBA tool uses long reads to improve the contiguity of genome assemblies," *PLoS Comput Biol*, vol. 18, no. 2, Feb. 2022, doi: 10.1371/JOURNAL.PCBI.1009860.

[18] S. Andrews, "FastQC A Quality Control tool for High Throughput Sequence Data." Accessed: Apr. 10, 2018. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[19] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[20] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform.,” *Bioinformatics*, vol. 25, no. 14, pp. 1754–60, Jul. 2009, doi: 10.1093/bioinformatics/btp324.

[21] K. Okonechnikov, A. Conesa, and F. García-Alcalde, “Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data.,” *Bioinformatics*, vol. 32, no. 2, pp. 292–4, Jan. 2016, doi: 10.1093/bioinformatics/btv566.

[22] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, “Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads,” *PLoS Comput Biol*, vol. 13, no. 6, pp. 1–22, 2017, doi: 10.1371/journal.pcbi.1005595.

[23] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph,” *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, May 2015, doi: 10.1093/bioinformatics/btv033.

[24] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, “Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.,” *Genome Res*, vol. 27, no. 5, pp. 722–736, Mar. 2017, doi: 10.1101/gr.215087.116.

[25] A. Di Genova, E. Buena-Atienza, S. Ossowski, and M. F. Sagot, “Efficient hybrid de novo assembly of human genomes with WENGAN,” *Nat Biotechnol*, vol. 39, no. 4, pp. 422–430, 2021, doi: 10.1038/s41587-020-00747-w.

[26] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, “QUAST: quality assessment tool for genome assemblies,” *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, Apr. 2013, doi: 10.1093/bioinformatics/btt086.

[27] T. Seemann, “Prokka: rapid prokaryotic genome annotation,” *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014, doi: 10.1093/bioinformatics/btu153.

[28] A. J. Page *et al.*, “Roary: rapid large-scale prokaryote pan genome analysis,” *Bioinformatics*, vol. 31, no. 22, pp. 3691–3693, Nov. 2015, doi: 10.1093/bioinformatics/btv421.

[29] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs,” *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, Oct. 2015, doi: 10.1093/bioinformatics/btv351.

[30] O. Nishimura, Y. Hara, and S. Kuraku, “gVolante for standardizing completeness assessment of genome and transcriptome assemblies,” *Bioinformatics*, vol. 33, no. 22, pp. 3635–3637, Nov. 2017, doi: 10.1093/bioinformatics/btx445.

[31] J. Il Sohn and J. W. Nam, “The present and future of de novo whole-genome assembly,” *Brief Bioinform*, vol. 19, no. 1, pp. 23–40, Jan. 2018, doi: 10.1093/bib/bbw096.

[32] R. Ekblom and J. B. W. Wolf, “A field guide to whole-genome sequencing, assembly and annotation,” *Evol Appl*, vol. 7, no. 9, pp. 1026–1042, 2014, doi: 10.1111/eva.12178.

[33] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, “Completing bacterial genome assemblies with multiplex MinION sequencing,” *Microb Genom*, vol. 3, no. 10, 2017, doi: 10.1099/mgen.0.000132.

[34] G. Narzisi and B. Mishra, “Comparing De Novo Genome Assembly: The Long and Short of It,” *PLoS One*, vol. 6, no. 4, p. e19175, 2011, doi: 10.1371/JOURNAL.PONE.0019175.

[35] G. L. Breckell and O. K. Silander, “Do You Want to Build a Genome? Benchmarking Hybrid Bacterial Genome Assembly Methods,” *bioRxiv*, p. 2021.11.07.467652, Nov. 2021, doi: 10.1101/2021.11.07.467652.

[36] Z. Chen, D. L. Erickson, and J. Meng, “Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing,” *BMC Genomics*, vol. 21, no. 1, pp. 1–21, Sep. 2020, doi: 10.1186/S12864-020-07041-8/FIGURES/6.

[37] V. Jayakumar and Y. Sakakibara, “Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data,” *Brief Bioinform*, vol. 20, no. 3, pp. 866–876, 2019, doi: 10.1093/bib/bbx147.

[38] E. Espinosa, R. Bautista, I. Fernandez, R. Larrosa, E. L. Zapata, and O. Plata, “Comparing assembly strategies for third-generation sequencing technologies across different genomes,” *Genomics*, vol. 115, no. 5, p. 110700, Sep. 2023, doi: 10.1016/J.YGENO.2023.110700.

[39] E. M. Batty *et al.*, “Long-read whole genome sequencing and comparative analysis of six strains of the human pathogen Orientia tsutsugamushi,” *PLoS Negl Trop Dis*, vol. 12, no. 6, 2018, doi: 10.1371/journal.pntd.0006566.

[40] D. Southwood, R. V Rane, S. F. Lee, J. G. Oakeshott, and S. Ranganathan, “Exhaustive benchmarking of de novo assembly methods for eukaryotic genomes,” *bioRxiv*, p. 2023.04.18.537422, Apr. 2023, doi: 10.1101/2023.04.18.537422.

[41] A. Di Genova, E. Buena-Atienza, S. Ossowski, and M. F. Sagot, “Efficient hybrid de novo assembly of human genomes with WENGAN,” *Nature Biotechnology 2020 39:4*, vol. 39, no. 4, pp. 422–430, Dec. 2020, doi: 10.1038/s41587-020-00747-w.

[42] H. Alhakami, H. Mirebrahim, and S. Lonardi, “A comparative evaluation of genome assembly reconciliation tools,” *Genome Biol*, vol. 18, no. 1, pp. 1–14, 2017, doi: 10.1186/s13059-017-1213-3.

[43] X. Zhang, C. G. Liu, S. H. Yang, X. Wang, F. W. Bai, and Z. Wang, “Benchmarking of long-read sequencing, assemblers and polishers for yeast genome,” *Brief Bioinform*, vol. 23, no. 3, pp. 1–13, May 2022, doi: 10.1093/BIB/BBAC146.

[44] A. M. Wenger *et al.*, “Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome,” *Nature Biotechnology 2019 37:10*, vol. 37, no. 10, pp. 1155–1162, Aug. 2019, doi: 10.1038/s41587-019-0217-9.

[45] W. T. Harvey *et al.*, “Whole-genome long-read sequencing downsampling and its effect on variant calling precision and recall,” *bioRxiv*, May 2023, doi: 10.1101/2023.05.04.539448.

[46] M. Gavrielatos, K. Kyriakidis, D. A. Spandidos, and I. Michalopoulos, “Benchmarking of next and third generation sequencing technologies and their associated algorithms for de novo genome assembly,” *Mol Med Rep*, vol. 23, no. 4, Apr. 2021, doi: 10.3892/MMR.2021.11890.

[47] A. Latorre-Pérez, P. Villalba-Bermell, J. Pascual, and C. Vilanova, “Assembly methods for nanopore-based metagenomic sequencing: a comparative study,” *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 1–14, Aug. 2020, doi: 10.1038/s41598-020-70491-3.

[48] Z. Zhang, C. Yang, W. P. Veldsman, X. Fang, and L. Zhang, “Benchmarking genome assembly methods on metagenomic sequencing data,” *Brief Bioinform*, vol. 24, no. 2, pp. 1–17, Mar. 2023, doi: 10.1093/BIB/BBAD087.

[49] A. Fuentes-Trillo *et al.*, “Benchmarking different approaches for Norovirus genome assembly in metagenome samples,” *BMC Genomics*, vol. 22, no. 1, pp. 1–12, Dec. 2021, doi: 10.1186/S12864-021-08067-2/FIGURES/2.

[50] F. Meyer *et al.*, “Critical Assessment of Metagenome Interpretation: the second round of challenges,” *Nature Methods 2022 19:4*, vol. 19, no. 4, pp. 429–440, Apr. 2022, doi: 10.1038/s41592-022-01431-4.

[51] G. Song *et al.*, “Integrative Meta-Assembly Pipeline (IMAP): Chromosome-level genome assembler combining multiple de novo assemblies,” *PLoS One*, vol. 14, no. 8, p. e0221858, Aug. 2019, doi: 10.1371/JOURNAL.PONE.0221858.

[52] I. A. Babarinde and A. P. Hutchins, “The effects of sequencing depth on the assembly of coding and noncoding transcripts in the human genome,” *BMC Genomics*, vol. 23, no. 1, pp. 1–14, Dec. 2022, doi: 10.1186/S12864-022-08717-Z/FIGURES/5.

[53] Y. Jiang, Y. Jiang, S. Wang, Q. Zhang, and X. Ding, “Optimal sequencing depth design for whole genome re-sequencing in pigs,” *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–12, Nov. 2019, doi: 10.1186/S12859-019-3164-Z/FIGURES/7.

[54] J. R. Homburger, C. L. Neben, G. Mishne, A. Y. Zhou, S. Kathiresan, and A. V. Khera, “Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores,” *Genome Med*, vol. 11, no. 1, pp. 1–12, Nov. 2019, doi: 10.1186/S13073-019-0682-2/FIGURES/5.

[55] J. A. Molina-Mora, M. Sibaja-Amador, L. Rivera-Montero, D. Chacón-Arguedas, C. Guzmán, and F. García, “Assessment of Mathematical Approaches for the Estimation and Comparison of Efficiency in qPCR Assays for a Prokaryotic Model,” *DNA*, vol. 4, no. 3, pp. 189–200, Jun. 2024, doi: 10.3390/DNA4030012.

[56] J. A. Molina-Mora and F. García, “Molecular Determinants of Antibiotic Resistance in the Costa Rican Pseudomonas aeruginosa AG1 by a Multi-omics Approach: A Review of 10 Years of Study,” *Phenomics*, vol. 1, p. 3, Jun. 2021, doi: 10.1007/s43657-021-00016-z.