Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Comparing assembly strategies for third-generation sequencing technologies across different genomes

Elena Espinosa [a,*], Rocio Bautista [b], Ivan Fernandez [a,c], Rafael Larrosa [a,b], Emilio L. Zapata [a,b], Oscar Plata [a]

[a] *Department of Computer Architecture, University of Malaga, Louis Pasteur, 35, Campus de Teatinos, Malaga 29071, Spain*
[b] *Supercomputing and Bioinnovation Center, University of Malaga, C. Severo Ochoa, 34, Malaga 29590, Spain*
[c] *Departament d'Arquitectura de Computadors, Universitat Politècnica de Catalunya, C. Jordi Girona, 1-3, Barcelona 08034, Spain*

ARTICLE INFO

Keywords:
Genome assembly
Hybrid assembly
HiFi
ONT
PacBio

ABSTRACT

The recent advent of long-read sequencing technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore technology (ONT), has led to substantial accuracy and computational cost improvements. However, de novo whole-genome assembly still presents significant challenges related to the computational cost and the quality of the results. Accordingly, sequencing accuracy and throughput continue to improve, and many tools are constantly emerging. Therefore, selecting the correct sequencing platform, the proper sequencing depth and the assembly tools are necessary to perform high-quality assembly. This paper evaluates the primary assembly reconstruction from recent hybrid and non-hybrid pipelines on different genomes. We find that using PacBio high-fidelity long-read (HiFi) plays an essential role in haplotype construction with respect to ONT reads. However, we observe a substantial improvement in the correctness of the assembly from high-fidelity ONT datasets and combining it with HiFi or short-reads.

1. Introduction

The analysis of complete genomes and transcriptomes is a field of great interest in the bioinformatics community. One of the most famous examples is the first human genome draft, which was based on traditional Sanger Sequencing technology and took more than ten years and $3 billion to complete [1,2]. Fortunately, the incorporation of Illumina's technology (short reads) has led to significant advances in sequencing time and cost reduction. Furthermore, new technologies such as PacBio [3] or Oxford Nanopore [4] can provide millions of reads over 15,000 bp [5] long with a precision of 99.9%. This precision is due to the appearance of PacBio HiFi [6,7] reads, the latest Q20+ platform update, and the V14 kit chemistry with R10.4.1 pore [8] (see Supplementary Table 1). These technologies enable the detection of significant structural variants and challenging repetitive regions that confound short-read sequencers because their short snippets cannot be differentiated during assembly. Consequently, long reads allow for substantial advances in bioinformatics [9,10], particularly in de novo genome assembly [2,11].

However, the high costs associated with both PacBio sequencing and equipment limited the sequencing using long reads until the emergence of high-fidelity reads from Oxford Nanopore Technology. Since pricing is crucial in enabling broad adoption, most solutions utilizing long reads typically exploit a combination of Illumina short reads with Nanopore long reads. In this scenario, the questionable quality of the results has conditioned the choice of the best technology [12]. Furthermore, the choice of technology substantially impacts processing time and memory footprint, significantly as genome complexity increases. To tackle this challenge, great research efforts have been made to develop new long-read assemblers and hybrid assembly methods capable of exploiting the computational capabilities of modern systems. Therefore, selecting the appropriate sequencing platforms, depths, and genome assembly tools is fundamental for obtaining high-quality genomes.

In terms of strategies to perform an assembly, we find two main methods: *OLC (Overlap Layout Consensus)* and *DBG (De Bruijn Graph)*. The *DBG* method converts the sequence in multiple sub-sequences or *k*-mers to identify overlapping reads, and then builds an overlap graph generating connections between all *k*-mers. It was initially used

* Corresponding author.
E-mail addresses: elenamesga@uma.es (E. Espinosa), rociobm@uma.es (R. Bautista), ivanfv@uma.es (I. Fernandez), rlarrosa@uma.es (R. Larrosa), ezapata@ac.uma.es (E.L. Zapata), oplata@uma.es (O. Plata).

https://doi.org/10.1016/j.ygeno.2023.110700
Received 24 March 2023; Received in revised form 7 August 2023; Accepted 16 August 2023
Available online 18 August 2023

successfully to assemble small genomes, such as bacteria, and was later extended to process large genomes. A main computational advantage of *DBG* is its scalability with the size and complexity of the genome, as multiple overlaps between different reads do not increase the number of graph nodes, which only grow with new *k*-mers. However, sequencing errors insert additional erroneous *k*-mers in the list, which increase the number of graph edges and, consequently, the complexity of the graph. That leads to the appearance of bubbles or bifurcation in the graph within which it is easy to recognize incorrect paths based on *k*-mer frequency. It also increases the memory footprint.

On the other hand, the *OLC* approach computes the alignment between reads to identify overlaps. It consists of three steps. First, in the overlap step, the algorithm computes overlaps between all sequencing reads. It results in a *Overlap Graph* or *String Graph*. The first one [13] is a bi-directed graph whose vertices are the input reads and each edge $e = (u, v)$ represents a connection between two reads u and v if a suffix of u matches a prefix of v. Each edge in the overlap graph has two arrowheads at its endpoints and the orientations of the arrowheads are used to denote the different ways in which the two reads at the ends of an edge can overlap [14]. The *String Graph* was proposed by Myers et al. [15], and it is a simplified version of a classic overlap graph and preserves the same properties and advantages as the *DBG*. It removes the transitive edges resulting in a directed overlap graph [16]. Each edge in a string graph is bidirectional to model the double-stranded nature of DNA and labeled with the unmatched substrings of the sequence reads [17]. Second, in the layout step, the reads are laid out into the most likely contiguous sequence stretches. Third, in the final step, the consensus sequence is determined for each contig by choosing the nucleotide, which is represented by the majority of the overlapping reads for every sequence position. With respect to the previous technique (*OLC*), it allows for preserving the information contained in the reads. However, the main bottleneck in this approach is the huge time required for the alignment between every possible read combination.

When assembling long-reads, we can follow the *OLC* or *DBG* approaches, however, the former is more suitable. For a complete list of long-read assemblers and their respective characteristics, see Supplementary Table 2. *Hybrid Assembly* combines short and long reads to create an efficient algorithm for hybrid reads [18,19], utilizing *DBG* and *OLC* methods. Our study aims to empirically evaluate these new approaches, focusing on their bottlenecks and areas for improvement. Previous results in other works have mainly focused on experimental evaluations of a limited number of genomes [20–22], considering only long-read assembly from Oxford Nanopore [23] or PacBio technology [24]. In contrast, our study performs experimental evaluations on small and complex genomes assembled from long-read (Oxford Nanopore and PacBio) and short-read (Illumina) sequencing technologies for long-read and hybrid assembly. The joint selection of assembly tools and sequencing technology is crucial for the accurate reconstruction of genomes. Additionally, a comprehensive assessment of the current landscape can help us identify the significant challenges in this field. Therefore, in this study, we aim to evaluate the latest non-hybrid and hybrid assemblers from both computational and biological perspectives. Specifically, we focus on non-hybrid assemblers such as *Hifiasm* [25], *Shasta* [26], and *HiCanu* [27], which are capable of assembling PacBio and Nanopore reads. For hybrid assembly, we evaluate the new assemblers *Wengan* [28] and *Verkko* [29]. Finally, we include *Miniasm* [30], designed for noisy long reads, in our benchmark as a reference for new assemblers. A brief description of the assemblers used in our benchmark is shown in Supplementary 3.

2. Materials and methods

2.1. Sequencing data acquisition

We obtained sequencing data from Pacific Biosciences (PacBio) and Oxford Nanopore technology (ONT) of haploid and diploid genomes. For the PacBio data, we selected HiFi (high-fidelity) long-reads sequenced with Sequel long-read systems. For Oxford Nanopore (ONT) data, we selected samples sequenced with the recent ligation sequencing kit and the chemistry R9.4 and later. In addition, we downloaded short-reads from Illumina technology and long-read from PacBio CLR and other Oxford Nanopore kits of *Homo Sapiens* and *Drosophila Melanogaster*. A complete description of the sequencing samples is shown in Supplementary Table 3.

2.2. De novo assembly pipelines

Figure 1 illustrates the de novo assembly pipeline we employed for our benchmark. We evaluate four recent assemblers - *HiCanu* (v2.2), *Hifiasm* (v0.18.5), *Miniasm* (v0.3), and *Shasta* (v0.10.0) - for non-hybrid assembly. To ensure accurate comparisons, we use recommended configurations for each genome type (as described in Supplementary 5) and processed 5 HiFi subsets for the first two assemblers and 3 ONT subsets for *Shasta* and *Miniasm* (as defined in Supplementary Table 3). The ONT assembly by *Shasta* and *Miniasm* requires additional processing with *marginPolish* (v1.3). Similarly, the results of *HiCanu* and *Verkko* require additional preprocessing to separate the haplotypes and reconstruct the primary haplotype using *purge_dups* (v1.2.6). For hybrid assembly, we employ *Wengan* (v0.0.2) and *Verkko* (1.3) to combine long and short reads, as shown in Fig. 1, using the standard mode recommended by the developers.

2.3. Evaluation overview

We conduct the experiments in the *Picasso* cluster available in the SCBI (*Supercomputing and Bioinnovation Center*, Malaga Tech-Park). We run the assemblers in Bull R282-Z90 nodes, each including two 64-core AMD EPYC 7742 processors and 2 TB of RAM. We analyze assemblers regarding quality and performance, discussing computing bottlenecks that are still present. Notably, we aim to determine: 1) the quality of the final assembly and 2) the performance of the computing resources in terms of processing time and memory usage.

2.3.1. Assembly quality

Aside from sequencing errors, assembly errors can arise for various reasons: (1) genomic regions may join together in incorrect places or orientations, and (2) regions may be dismissed as repeats or inaccuracies. Unfortunately, distinguishing between errors caused by experimental artifacts or missing data can be challenging, and may result in incomplete or inaccurate assemblies. To address this issue, we assess the quality of the assembly in terms of genome contiguity, correctness, and completeness. Detailed descriptions of the parameters and datasets used for evaluation can be found in Supplementary 6, andSupplementary Table 4. Contiguity was evaluated based on the number and size of the contigs. To achieve high contiguity, genome assembly must maximize contig length while minimizing their number, to accurately reflect the number and size of chromosomes in the organism. Correctness was evaluated by assembling with a reference genome using the *Quast* (v.5.2.0) tool, which evaluates contig ordering and location. Incorrect positioning could indicate the presence of inversions, relocations, or translocations compared to the accurate genome. Finally, completeness was assessed by identifying expected genes in each genome using the *Busco* tool (v.5.4.4). This evaluates the content of the contig in terms of gene content. These errors could arise during sequencing (i.e., expected genes not sequenced) or during assembly due to discarded contigs. Additionally, we used *GenomeScope* (v.2.0) to evaluate parameters such as heterozygosity and repetitiveness based on *k*-mer frequencies.

2.4. Performance analysis

We computationally characterize the different assemblers based on the processor performance and the memory footprint. For this purpose,

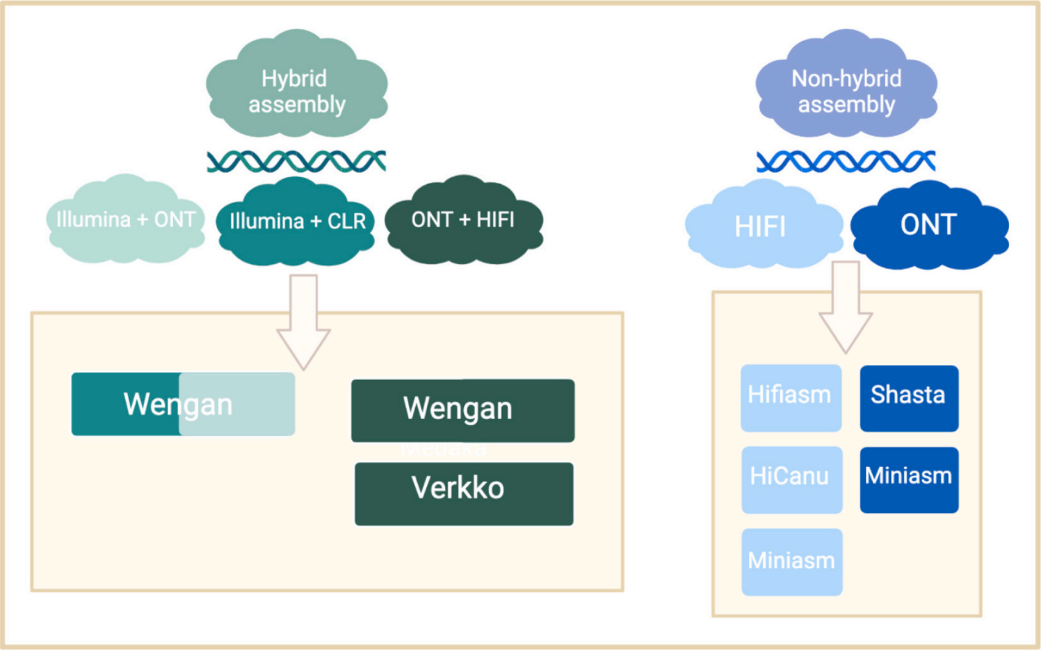

Fig. 1. Pipeline design for benchmarking hybrid and non-hybrid strategies. For the evaluation of the hybrid strategy, we combine short and long reads and the assemblers *Wengan* and *Verkko*. For the non-hybrid strategy, we selected the assemblers *Shasta* for ONT reads, *HiCanu* and *Hifiasm* for HiFi reads and *Miniasm* for both of them.

we measure the CPU workload and the RAM usage. We obtain the thread count to quantify the amount of parallelism and determine the performance in terms of CPU usage. We measure CPU and wall times to assess multithreading's inherent parallelism. We calculate the RAM usage, identifying the peak usage and the memory growth as the genome complexity and length increase.

3. Assembly quality evaluation results

The quality of the assemblies depends on several factors, such as the genome size, levels of ploidy, and heterozygosity of the genomes. Moreover, sequencing platforms, the depths of the sequencing, and the algorithms used are other essential factors. In this work, we use different algorithms and metrics to compare assembly strategies' capabilities in varying complexity genomes.

3.1. *Evaluation of non-hybrid pipelines on HiFi datasets*

Comparing two recent assembly tools designed to leverage the full potential of HiFi reads (Table 1), *HiCanu* and *Hifiasm*. We evaluate their potential to assemble different genomes, from haploid to diploid and repetitive genomes.

3.1.1. *Contiguity assembly*

Based on the length of the contigs, N50 and NG50 values obtained in *H. sapiens* in each strategy of assembly, we notice that *Hifiasm* presents higher contiguity in the assembly of diploid genomes. We observe this trend as the complexity of the genomes increases as *Solanum tuberosum* and *H. sapiens*. However, *Hifiasm* exhibits a notable number of contigs with respect to *HiCanu* in the haplotype reconstruction of the *D. melanogaster* (281 contigs vs. 59 contigs) and *S. tuberosum* (1610 contigs vs. 562 contigs) genome. On the other hand, the size of the estimated genomes resulting from the assemblies is especially larger

Table 1
Quality evaluation of different genomes on HiFi datasets measured in terms of contiguity, correctness and completeness for the assembly with *Hifiasm* and *HiCanu.*

Quality evaluation	Metric	*E. coli* k12		*S. pombe*		*D. melanogaster*		*S. tuberosum*		*H. sapiens* (HG002)	
		HiCanu	Hifiasm	HiCanu	Hifiasm	HiCanu	Hifiasm	HiCanu	Hifiasm	HiCanu	Hifiasm
Contiguity	N50 (Mbp)	4.54	4.66	4.61	9.66	20.87	22.94	3.86	24.27	36.65	87.11
	NG50 (Mbp)	4.54	4.66	4.61	9.66	20.87	23.95	5.14	28.88	38.18	78.86
	Number of contigs	217	4	72	75	59	281	562	1610	1169	934
	GC (%)	50.61	50.74	36.29	36.56	41.71	41.29	34.48	35.60	40.54	40.75
	Largest contig (Mbp)	4.54	4.66	5.62	9.66	24.41	28.21	19.10	60.22	121.36	176.22
	Genome (Mbp)	9.01	4.73	15.93	16.63	179.69	177.77	942.55	953.14	3326.85	31,126.15
Correctness	Missassemblies	10	6	139	196	4027	6255	78,237	67,167	18,844	18,715
	Missmatches	314	8	3613	4289	586,895	616,356	9,593,452	8,807,798	5,359,406	5,542,121
	Fully unaligned contig	0	0	0	0	7	55	9	62	3	12
	Partially unaligned contigs	0	0	3	1	31	92	522	828	330	187
	Genome fraction (%)	99.998	99.998	99.855	95.937	88.578	94.576	70.91	77.078	93.986	94.45
	Total aligned (Mbp)	9.01	4.73	15.81	16.41	175.02	164.29	697.13	641.38	3260.63	3002.72
Completeness	Missing genes	0	0	18.4	18.3	0.3	0.5	2.6	1.4	3.6	3.3
	Fragmented genes	0	0	1.7	1.7	0.1	0.1	0.1	0.1	1.6	1.1
	Identified genes	100	100	79.9	80	99.6	99.4	97.3	98.5	94.8	95.6
	Total	124	124	1706	1706	1367	1367	5950	5950	13,780	13,780

than their reference in repetitive genomes (observed in *S. tuberosum*). This fact is also notable in the assembly of haploid genomes by *HiCanu* which have not been previously purged (Table 1).

3.1.2. Correctness assembly

Concerning the correctness, and according to the results of the alignment to the references, we notice that the assemblies led by *HiCanu* and *Hifiasm* present a similar number of mismatches and misassemblies, except *Escherichia coli* which shows an exacerbated number of mismatches in the assembly with *HiCanu* (314 mismatches vs. 8 mismatches on *Hifiasm*). In the same way, both present a high number of aligned bases at the references, 3.2 Gbp and 3 Gbp for *HiCanu* and *Hifiasm*, respectively, for the assembly of *H.Sapiens*. *S.tuberosum* is an exception considering that it exhibits a low value of percentage genome fraction, 70.91% for *HiCanu* and 77.078% for *Hifiasm* (Table 1).

3.1.3. Functional completeness assembly

Upon identifying the expected gene content in the assemblies, we observe that both assemblers, *HiCanu* and *Hifiasm*, yield similar numbers of missing and fragmented genes across most datasets, regardless of genome complexity. Notably, both assemblers successfully identify a high percentage of expected genes, with less than 4% missing, except for *S. pombe*, where *HiCanu* and *Hifiasm* identify 18.4% and 18.3% missing genes, respectively (Table 1).

3.2. Evaluation of non-hybrid pipelines on Nanopore (ONT-bases) datasets

To investigate the impact of sequencing technology on genome reconstruction, we assemble the haploid genome of *S. pombe* and the diploid genomes of *D. melanogaster* and *H. sapiens* using *Shasta* and *Miniasm* pipelines. We obtain these genomes through different protocols of the Nanopore sequencing technology, and all results are shown in Table 2. Overall, we observe a significant effect of the new Q20 sequencing, combined with the new R10.4.1 flow cell, on the assemblies of *D. melanogaster* and *H. sapiens*, compared to sequencing with the standard ligation kit protocol used for *S. pombe* sequencing. We note this trend in the assemblies performed by both *Shasta* and *Miniasm* pipelines on all three datasets. However, the completeness of the *S. pombe* genome was lower, with only 67.8% of genes identified using *Shasta* and 55.8% using *Miniasm*. In contrast, there is a clear improvement in the percentage of genes identified in the genomes of *D. melanogaster* and *H. sapiens*, especially in the assemblies performed by Shasta, with 97.3% for *D. melanogaster* and 87.2% for *H. sapiens*. Nevertheless, there is a significant difference in contiguity among the three assemblies when comparing with *HiCanu* and *Hifiasm*'s HiFi pipelines. *Hifiasm* exhibits the highest contiguity with an N50 and NG50 value of 87.11 Mbp and 78.86 Mbp, respectively, for *H. sapiens*, whereas the other two assemblies, with *Shasta* and *Miniasm*, have N50 and NG50 values of less than 2 Mbp for the same genome. When comparing *Shasta* and *Miniasm*'s results (Table 2), diploid genomes assembled with *Miniasm* exhibit higher contiguity with N50 values of 3.33 Mbp for *D. melanogaster* and 1.14 Mbp for *H. sapiens*. In contrast, *Shasta*'s N50 values are only 0.026 MB for *D. melanogaster* and 0.78 Mbp for *H. sapiens*, indicating lower contiguity. Furthermore, *Shasta* has a higher number of fully unaligned contigs (3069) compared to *Minisam* (59) in *H. sapiens*. However, the number of partially unaligned contigs is similar in both assemblies for the *H. sapiens* dataset (3326 for *Shasta* and 3354 for *Minisam*). Despite these differences, *Shasta* outperforms *Miniasm* in terms of completeness, as it has identified fewer missing and fragmented genes in the assembly of *H. sapiens* (7.5% and 3.9% for *Shasta*, compared to 29.9% and 6.9% for *Miniasm*, respectively).

3.3. Impact of HiFi sets in the improvement of the non-hybrid assembly

To demonstrate the impact of sequencing technologies on assembly accuracy, we evaluate the assembly of *S. pombe* using *Miniasm* with both *HiFi* and *ONT* sequencing data. We assess two levels of ONT assembly: raw (uncorrected assembly) and corrected with *marginPolish*. Our results show a significant improvement in assembly quality, including completeness, correctness, and contiguity (seeSupplementary Table 5). Cleaning the *ONT* assembly substantially improves the results (88.2% versus 28.4% of missing genes). However, the corrected ONT assembly still has many absent genes and is inferior to the HiFi assembly in terms of accuracy (61.8% versus 79.9% and 80% of identified genes in *HiCanu* and *Hifiasm* assemblies). This underscores the importance of HiFi reads in improving assembly quality. Additionally, comparing the completeness of the HiFi and uncleaned ONT assemblies highlights the significance of high-quality PacBio reads (36.6% versus 88.2% of missing genes). Despite this, the ONT assembly's completeness remains lower than that of the HiFi assembly.

3.4. Evaluation of hybrid strategy

In this work, we also explore the benefits of employing a hybrid approach for genome assembly and investigate its performance on diploid genomes. Specifically, we assemble the genomes of *H. sapiens* from the diploid cell line HG002 and *D. melanogaster* using *Wengan* and *Verkko* algorithms, with the aim of assessing whether incorporating two types of reads (long and short reads) could enhance the quality of the final results. The outcomes of our evaluation are summarized in Table 3.

3.4.1. Hybrid assembly from short and long reads

We estimate the size of two genomes and found that they were

Table 2
Quality evaluation of different genomes on ONT datasets measured in terms of contiguity, correctness, and completeness for the assembly with *Shasta* and *Miniasm.*

Quality evaluation	Metric	*S. pombe*		*D. melanogaster*		*H. sapiens* (HG002)	
		Shasta	Miniasm	Shasta	Miniasm	Shasta	Miniasm
Contiguity	N50 (Mbp)	4.57	4.45	0.026	3.3	0.78	1.14
	NG50 (Mbp)	4.57	4.45	0.35	3.64	0.66	0.96
	Number of contigs	6	50	19,889	192	28,310	6599
	Largest contig (Mbp)	5.44	5.50	5.63	9.01	4.09	5.90
	GC (%)	36.11	35.18	42.35	41.2	40.96	40.91
	Genome (Mbp)	12.59	16.09	313.05	164.22	2785.73	2823.56
Correctness	Missassemblies	76	125	4922	4596	5396	4561
	Missmatches	69,060	85,720	1,927,137	1,448,092	3,747,738	20,904,220
	Fully unaligned contig	0	12	591	6	3069	59
	Partially unaligned contigs	3	38	379	157	3326	3354
	Genome fraction (%)	97.006	97.094	96.333	94.394	89.36	90.227
	Total aligned (Mbp)	12.36	13.27	304.18	152.74	2755.54	2784.83
Completeness	Missing genes (%)	24.3	30.3	1.5	6.9	7.5	29.9
	Fragmented genes (%)	7.9	13.9	1.2	5.8	3.9	6.9
	Identified genes (%)	67.8	55.8	97.3	87.3	87.2	63.2

Table 3
Quality evaluation of hybrid approaches on diploid datasets measured in terms of contiguity, correctness, and completeness for the assembly with *Verkko* and *Wengan*.

Quality evaluation	Metric	D.Melanogaster (PacBio CLR + ILL)	D.Melanogaster (ONT + ILL)	H.Sapiens (ONT + ILL)	H.Sapiens (HiFi + ONT)	
		Wengan	Wengan	Wengan	Wengan	Verkko
Contiguity	N50 (Mbp)	4.68	6.37	16.69	0.78	0.86
	NG50 (Mbp)	3.33	1.84	13.9	0.55	1.47
	Number of contigs	263	418	2605	37,277	23,945
	Largest contig (Mbp)	16.62	19.38	102.26	45.36	50.03
	GC (%)	42.36	42.29	40.81	40.85	40.84
	Genome (Mbp)	121.18	118.39	2734.80	2562.06	5838.96
Correctness	Missassemblies	1305	1766	2576	2145	27,130
	Missmatches	710,370	682,356	3,707,472	3,491,463	9,301,410
	Fully unaligned contig	73	79	21	199	187
	Partially unaligned contigs	78	78	488	1417	2857
	Genome fraction (%)	83.63	81.18	89.24	79.83	95.389
	Total aligned (Mbp)	119.21	116.29	2722.49	2545.64	5681.68
Completeness	Missing genes	0.7	2.9	6.9	16.5	5.3
	Fragmented genes	0.3	0.	1.6	2.9	3.7
	Identified genes	98.3	95.6	91.5	80.6	91.0

shorter than the reference genome. *D. melanogaster* has a length of 121.18 Mbp and 118.39 Mbp, while *H. sapiens* has a length of 2734.80 Mbp, both genomes shorter than the respective reference genomes. The N50 and NG50 values predict low contiguity, from *Wegan*, in both genomes when compared with the results obtained by *HiCanu* and *Hifiasm* (Fig. 2). On the other hand, we also observe that the quality values of the hybrid assembly from PacBio CLR reads decrease as the genome's complexity increase, the assembly of *H. sapiens* exhibiting an incomplete construction of 0.55 Gbp (data not shown). When we measure the mapping against the reference genome, we find that the number of misassemblies is lower for PacBio and ONT (1305 and 1766, respectively) compared to *HiCanu* and *Hifiasm* (4027 for *HiCanu*, and 6255 for *Hifiasm*) in the assembly of *D. melanogaster*. Also, these values and the number of missmatches are significantly better in the assembly of *H. sapiens* from ultra-long reads and Illumina respect to the assemblies from HiFi pipelines (e.g 2576 and 3,707,472 vs. 18,715 and 5,542,121 identified in the assembly of *Hifiasm*). However, the number of aligned bases with the reference is insufficient for the two genomes concerning the two *HiFi* assemblers (less than 120 Mbp for *D. melanogaster* and less than 2.8 Gbp for *H. sapiens*, with respect to 164.29 Mbp and 3 Gbp denoted by *Hifiasm* for the two assemblies). The percentage of the reconstructed genome fraction is also lower in both assemblies

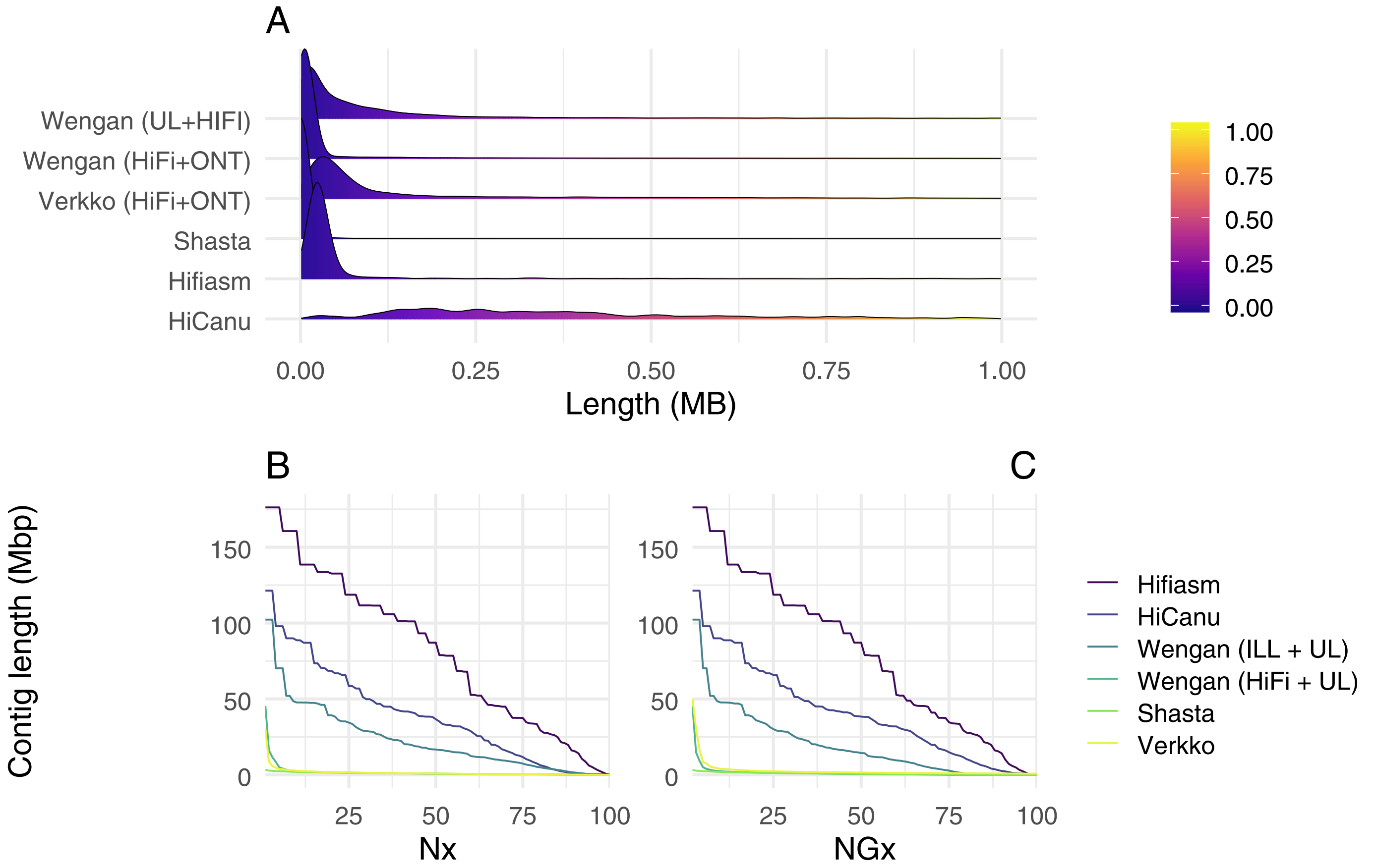

Fig. 2. Contiguity analysis for the evaluation of hybrid and non-hybrid strategies based on the assembly of the diploid cell line of *H. sapiens* (HG002). (A) Length distribution of contigs. (B)-(C) Nx (left) and NGx (right) values as x varies from 0 to 100%.

compared to those obtained with *HiCanu* or *Hifiasm* on both models (less than 90% in both assemblies). It decreases dramatically for the assembly of *H. sapiens* from PacBio CLR reads, which does not exceed 10% (data not shown). Finally, according to the completeness of the assembly, *Wengan* denotes a similar number of missing and fragmented genes when compared to *Hifiasm* and *HiCanu*, except for the hybrid assembly of *H. sapiens* from PacBio CLR reads (data not shown).

3.4.2. Hybrid assembly from only long reads

On the other hand, the combination of high-quality reads (i.e., ONT and HiFi reads) does not result in an improvement in the quality parameters if we compare it to the hybrid assembly of long and short reads, and the contiguity remains very low, as shown in Fig. 2. Only a slightly higher percentage of the genome fraction is observed in the *Verkko* assembly due to the increased number of aligned bases (5.8 Gbp vs. 2.7 Gbp obtained by *Wengan*). However, there is a significant improvement in the completeness of the hybrid assembly with *Verkko* compared to *Wengan* (5.3% of missing genes vs. 16.5%). Supplementary Table 6, Supplementary Table 7 and Supplementary Table 8 provide additional *QUAST* parameters, offering further insights into the results of the analysis for the evaluation from HiFi, ONT and hybrid assembly. Supplementary Figure 3 addresses additional contiguity and completeness analysis for the evaluation of the primary assembly following hybrid and non-hybrid approaches. Furthermore, Supplementary Table 9 and Supplementary Table 10 show quality evaluation with *QUAST* of the diploid cell line HG002 for the primary and phased assembly with hybrid and non-hybrid approaches using the T2T CHM13 v2.0 and the recently finished HG002 ChrX and ChrY as reference genome respectively. Finally, to assess the correctness based on the reference, we show in Fig. 3 the alignment of the assembly results, both for the long-read assembly and the hybrid assembly, on the HG002 model. In this graph, we can observe that, despite *HiCanu* and *Hifiasm* exhibiting better quality results in the assembly of *H. sapiens*, the assembly from ONT reads, and the assembly of the combined reads, show a lower number of translocations. In this sense, *Shasta*, *Wengan*, and *Verkko* show a less contiguous and complete genome but with more inversions. In addition, results on the alignment (see Supplementary Figure 1 and Supplementary Figure 2) against the T2T CHM13 v2.0 and the recently finished HG002 ChrX and ChrY yield the same results. Finally, Supplementary Figures 4-8 show additional analysis with *Merqury*.

3.5. Runtime, memory and CPU utilization

3.5.1. CPU workloads characterization

The length and complexity of a genome have a direct impact on the performance evaluation of assemblers in terms of computational resource utilization. With this in mind, we measure the efficiency of evaluated assemblers using small and complex genomes in a multicore system. Generally, *Shasta*'s assembly of ONT datasets results in lower CPU usage while maintaining a significant commitment to quality. It also shows improvement over *Miniasm*, whose computation time approaches that of HiFi pipelines (Fig. 4). For the assembly of HiFi datasets, we observe that *Hifiasm* has a lower CPU consumption for the assembly of long and complex genomes (Table 4). *HiCanu* exhibits lower CPU usage for less complex genomes like *E. coli* and *S. pombe* but increases dramatically with more complex genomes like *H. sapiens*, as shown in Fig. 4. Conversely, the hybrid assembly using long and short reads reported by *Wengan* for the assembly of the diploid genome of *D. melanogaster* does not entail an additional computational cost compared to *HiCanu* and *Hifiasm* (less than 40 CPU hours compared to 246 and 460 reported by *Hifiasm* and *HiCanu*, respectively). However, this trend changes when the genome's complexity increases, as seen

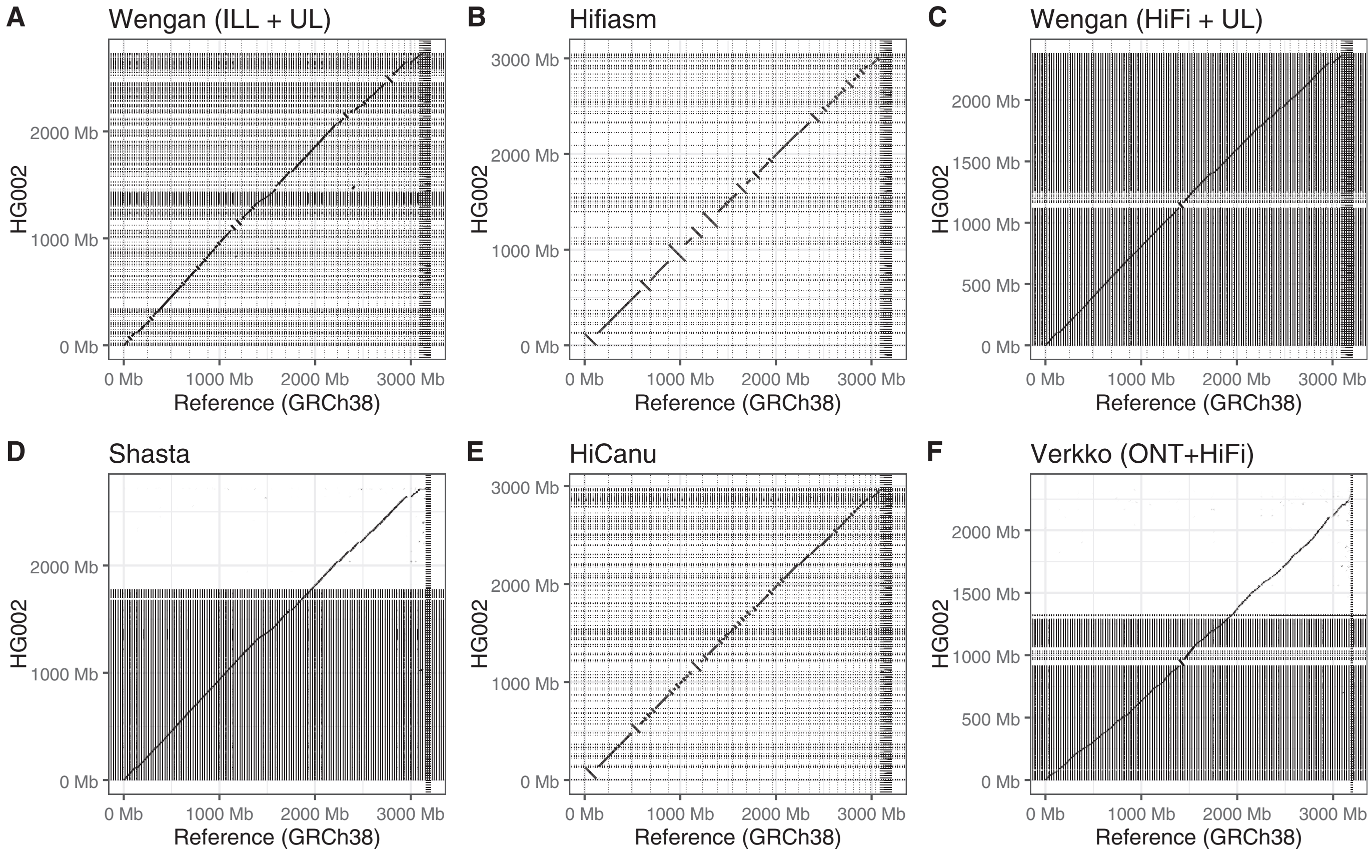

Fig. 3. Correctness analysis for the evaluation of hybrid and non-hybrid strategies based on alignments between the human genome reference (GRCh38) and the assembly of the diploid cell line of *H. sapiens* (HG002).

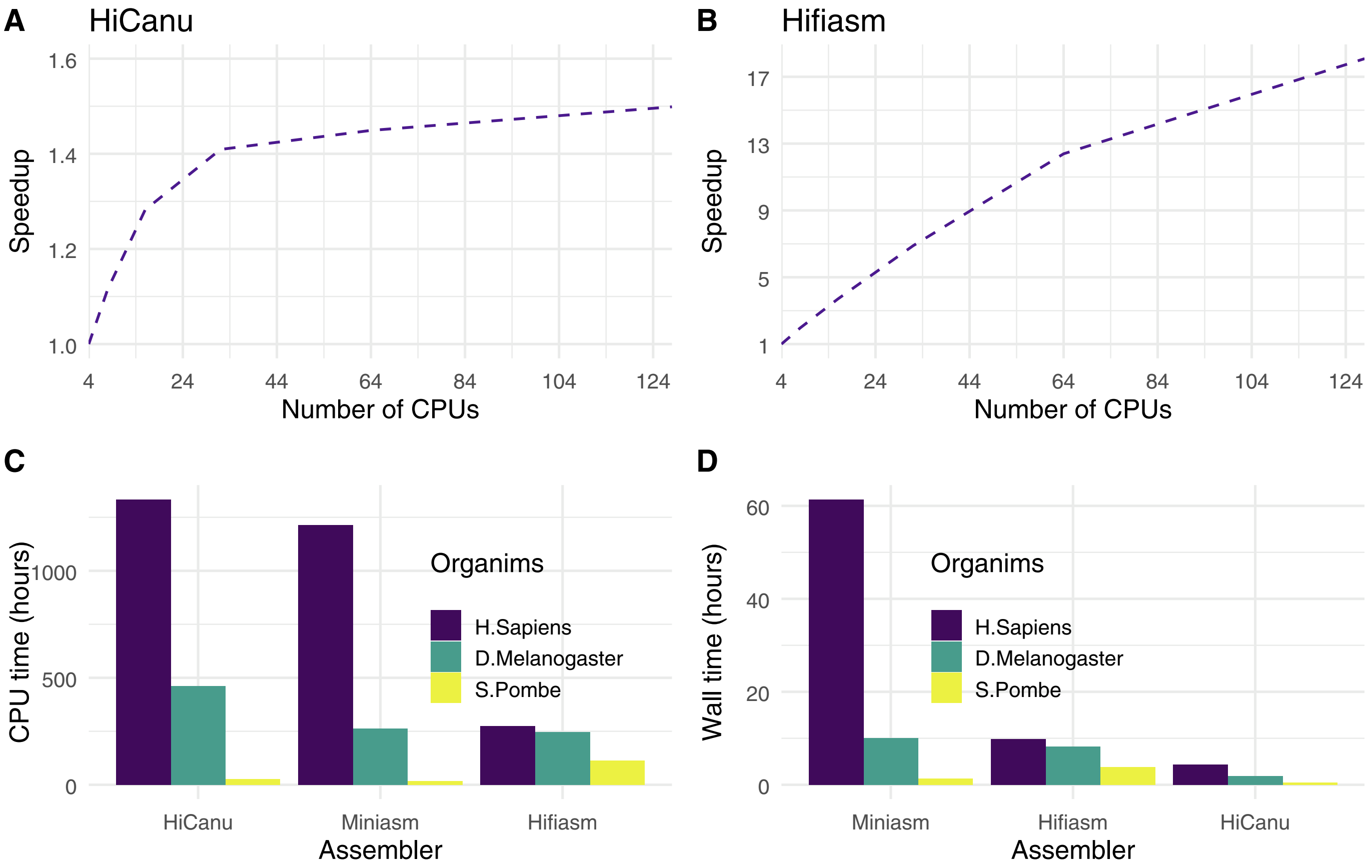

Fig. 4. Computational analysis of resources consumed by *Hifiasm*, *HiCanu* and *Miniasm*. (A)-(B) Scalability analysis for the assembly of *H. sapiens* using CPUs count from 8 to 128 (speedup measured with respect to 4 CPUs). (C)-(D) CPU consumption and elapsed time in the assembly of the organisms *H. sapiens*, *D. melanogaster* and *S. pombe* from ONT and HiFi datasets.

Table 4
CPU and memory usage for hybrid and non-hybrid assemblies measured for the assembly of haploid and diploid genomes.

Organims	Assembler	CPU time (h)	Elapsed time (h)	CPU Efficiency (%)	Peak RSS (GB)
E.Coli k12	HiCanu	5:19	0:20	67	4.06
E. coli k12	Hifiasm	16:52	0:35	90.35	4.654
S. pombe	Shasta	1:16	0:05	48.73	22.81
S. pombe	Miniasm	17:50	1:18	*81.19	64.04
S. pombe	HiCanu	27:43	0:31	78.39	4.23
S. pombe	Hifiasm	113:42	3:51	92.45	17.382
D. melanogaster	Shasta	165:11	5:59	86.21	80.74
D. melanogaster	Miniasm	263:34	10:03	*82.36	71.54
D. melanogaster	HiCanu	460:53	1:55	77	18.82
D. melanogaster	Hifiasm	246:08	8:15	93.17	47.877
D. melanogaster	Wegan (PacBio CLR + ILL)	30:27	1:07	84.90	24.20
D. melanogaster	Wegan (Nanopore+ILL)	20:13	1:06	57.82	7.17
D. melanogaster	Wengan (Nanopore+HiFi)	176:58	24:28	22.60	24.82
S. tuberosum	HiCanu	482:06	2:19	74.81	12,74
S. tuberosum	Hifiasm	75:03	2:36	90.37	34.88
H. sapiens	Shasta	33:39	1:44	60.88	354.77
H. sapiens	Miniasm	1212:50	61:24	*86.53	772.25
H. sapiens	HiCanu	1333:52	4:21	67,25	125.1
H. sapiens	Hifiasm	275:15	9:49	87.48	118.04
H. sapiens	Wegan (Nanopore+ILL)	372:00	15:44	73.98	1320
H. sapiens	Wegan (Nanopore+HiFi)	869:00	108:58	24.92	94.25
H. sapiens	Verkko (Nanopore+HiFi)	495:14	51:15	77.11	40.37

with *H. sapiens*, where the computational cost increases dramatically (372 CPU hours for the assembly from ONT and Illumina vs. 275 h spent by *Hifiasm*). Similarly, assembly from only long reads with improved sequencing quality and sequencing depth results in a remarkable increase in computational cost (869 CPU hours compared to 372 h for the assembly from ONT and Illumina). Despite this increase, the computational cost involved in the hybrid assembly is lower than that used by *HiCanu*. This trend is also evident in the assembly with *Verkko*, where the CPU time is kept at 495 h when we omit the correction step with *Canu*. Moreover, *HiCanu* has higher thread parallelism with respect to *Hifiasm* and the two hybrid assemblers by taking advantage of process-level parallelism through the use of *array jobs*. It is more prominent in the *overlap* step where the number of used cores is configurable and the size of the array jobs is significantly high and proportional to the genome size.

In contrast, *Hifiasm* takes advantage of data-level parallelism through the use of Intel SSE vector instructions which allows an efficient implementation of the *Myer* bit-vector algorithm used to compute the distance between reads. It decreases the latency and consequently the memory access time by the CPU which is reflected in CPU consumption used in the different assemblies. In the same way, despite the high thread-level parallelism presented by *HiCanu*, due to the large number of processes generated in the array jobs, it is noted that the speed up grows slowly as the number of CPUs increases respect to the results of *Hifiasm* (Fig. 4). Likewise, we observe that the CPU efficiency with *Shasta* is in general low (reached 86.21% in *D.melanogaster* and decrease to 60.88% and 48.73% in *H.Sapiens* and *S.pombe* respectively). This behavior is also observed in *Wengan* which presents different values in based on the mode and the genome. In this sense, we conclude that the complexity of the genome as well as the type of reads have a direct influence on it. In the case of *Miniasm*, it remains constant and *Verkko* reaches values close to *Wengan* (77.11% vs. 74%). On the other hand, in HiFi pipelines, it tends to decrease as the complexity increases. Also, we appreciate that the CPU performance is lower in *HiCanu* than *Hifiasm* (which maintains good CPU performance). We observe that it is explained by the existence of tasks with low parallelism levels, which do not have an impact on the final time due to the low CPU consumption. It also happens with *Verkko*, which reaches 99% CPU performance in some stages (e.g., graph processing) and decreases to 4% in the index of the graph.

Finally, in the hotspot analysis of the most time-consuming steps in *Hifiasm* and *HiCanu*, we find that the overlap computation consumes most of the CPU time, growing larger with the genome size. In this sense, we note that even though *Hifiasm* achieves a remarkable speed up in the chaining step with respect to *HiCanu*, it is still very time-consuming. We observe that the CPU time increases dramatically as the complexity of the genome grows, due to the large rise in the number of read comparisons. This fact can lead to a computing bottleneck when processing large and complex genomes in systems with limited hardware support for thread-level parallelism.

3.5.2. Memory footprint

The length of the reads as well as the number of reads have a direct effect on the memory footprint depending on the assembly algorithm used. This fact not only has an impact on the choice of the algorithm but also on the choice of the most suitable architecture. Regarding DRAM memory usage, the assembly, in general, leads to an increase in the memory footprint as the length of the genome or ploidy level grows (shown in Table 4). The memory footprint of *HiCanu* is relatively small regarding disk usage, as it works mainly with data on disk. Otherwise, the memory usage would be extremely huge. On the contrary, *Hifiasm* works mainly with data on memory, limiting the disk usage to the initial reading of sequencing files and the final writing of the resulting assembly. This fact boosts performance. It is also observed that, in some steps, *HiCanu* have an inefficient behavior when using files that only need to be accessed locally and are stored in a distributed file system.

In the assembly of ONT datasets, we observe that *Shasta* and *Miniasm* impact on the memory footprint with a consumption of 354.77 GB and 772.25 GB respectively in the assembly of *H.sapiens*, versus 118.04 GB consumed by *Hifiasm*. It is exceeded by *Wengan* which reaches 1.32 TB in the assembly oh *H. sapiens* from ONT and Illumina datasets, five times larger than *Shasta*. Also, it can be seen that the memory footprint increases dramatically with the length of the genome and the sequencing deep. We observe it in the assembly of *D.melanogaster*, where the DRAM memory usage remains below the DRAM spent by *Hifiasm*. We can conclude that *Hifiasm* presents a more affordable memory footprint. However, the processing of generated sub-sequences and the post-processing of overlapping sequences result in a huge memory footprint. We note that the read comparison and the building of the assembly graph result in a peak memory explosion, generating a large data movement between the memory and processor.

The jobs were configured with 32 CPUs. *Because Miniasm has no thread-level parallelism, the actual time of Miniasm is determined by the performance of Minimap2 which establishes the overlaps between the reads*.

4. Discussion

We find that the emergence of *high-quality* long reads from PacBio has a significant impact on genome assembly in terms of both assembly quality and computational cost [31–33]. However, assembly from short reads or noisy long reads with high error levels presents significant computational challenges [33–35]. Hybrid assembly and the new ONT technology [36] have emerged as potential alternatives that could improve the accuracy of the assembly. When evaluating assembly quality, we observe that the assemblies produced by *Hifiasm* and *HiCanu* have higher contiguity than those produced by ONT datasets and hybrid assemblies, particularly in the assembly of *H. sapiens*. This may be due to the inability to resolve the presence of highly repetitive regions in complex genomes where the presence of very similar copies could lead to a high number of ambiguous and unresolvable paths in the assembly graph [37]. Also, highly heterozygous regions may be difficult to resolve and may result in fragmented regions or gaps in the assembly. *Hifiasm* generally reportes longer and more uniformly distributed contigs than *HiCanu*, but in homozygous genomes (e.g., *D. melanogaster* and *S. tuberosum*), the number of contigs reported by *Hifiasm* is notably higher. However, *Hifiasm* also produces contigs larger than the reference, potentially indicating misjoins that could be misleading [38].

In the correctness evaluation, we find that the number of misassemblies in the assembly's alignment to a reference is similar in both assemblies. However, these values increase dramatically in the assembly of *S. tuberosum* for both *Hifiasm* and *HiCanu*, possibly due to the presence of complex repeat structures. The genome of *S. tuberosum* [39,40] contains a high proportion of repetitive sequences and the presence of these repeat structures can lead to misassemblies, as the assembler may have difficulty in resolving the correct order and orientation of the repeats. It also could be influenced by a large number of single nucleotide polymorphisms (SNPs) and structural variations (SVs) between the two homologous chromosomes present in the genome. Interestingly, although the number of contigs that do not align partially and completely is high in ONT and hybrid assemblies, the number of misassemblies is greatly reduced in the assembly of *H. sapiens*, with fewer mismatches reported. These results are confirmed by the alignment against the reference, where the number of low translocations found in the alignment of the assembly performed by hybrid pipelines and *Shasta* with the reference indicates quality in the correctness analysis. However, repeats may also cause the assembler to over-split the genome, resulting in a higher number of contigs but with fewer inversions due to the fragmentation of the genome. In this sense, the low contiguity presented by these assemblers could mask the presence of these translocations.

When searching for universal single-copy orthologs in the assembled genomes, we find that *Hifiasm* and *HiCanu* in general show low fragmentation with respect to the other assemblers. However, we observe low completeness in the assembly of the genome of *S. pombe* from Hifi and ONT pipelines. This fact combined with the number of misassemblies could be explained by the presence of repetitive DNA sequences and transposable elements in the genome or the heterozygosity which can occur in haploid genomes due to the accumulation of mutations and may lead to difficulties in assembling the genome.

From a computational standpoint, high-fidelity long reads have made it possible to reduce the computational cost of genome assembly [41,42], which was previously unapproachable for long and complex genomes. One example of this improvement is the addition of *HiCanu* to the *Canu* package [27]. The assemblers *Hifiasm* and *HiCanu* have both successfully overcome significant computational challenges. Our analysis of both assemblers shows that while *Hifiasm* uses less CPU as the length and complexity of the genome increases, it also has a reduced memory footprint, particularly in the overlap step. Despite the compulsive parallelism implemented by *HiCanu* using array jobs, *Hifiasm* still exhibits moderate CPU usage. We also note that the modified Myer algorithm implemented by *Hifiasm*, which performs overlap alignment of all-vs-all, shows improvements compared to the MHAP [43] strategy implemented by *Canu* and *HiCanu*, involving the exploitation of bit-level parallelism [44]. However, the overlap step is still the major performance bottleneck. Furthermore, the combined use of long and short reads at low coverage only slightly increases computation time compared to *Hifiasm*, albeit with a larger memory footprint. However, the combined use of long reads significantly increases CPU usage, especially in the hybrid assembly of *H. sapiens* from HiFi and ONT reads, resulting in substantially longer wall time (e.g., 108 h for the assembly of *H. sapiens* from ONT and HiFi reads, versus 9 h for *Hifiasm*). As a result, genome assembly using high-fidelity long reads still presents computational challenges, which are further exacerbated when implementing a hybrid assembly to increase fidelity and sequencing depth.

In general, the performance analysis of different assemblers highlights the limitations of available assemblers for processing large and complex genomes. Additionally, the study of scalability demonstrates the application of *Amdahl's* law. However, as sequencing data grows, solutions that exploit multicore systems will inevitably be limited by *Moore's* law [45]. Therefore, it is necessary to approach the inherent computational cost of genome assembly from a different perspective. This includes implementing and improving new techniques, as well as designing new architectures that efficiently support applications. One potential solution is to adopt computing paradigms such as data-centric computing, near-data processing, and processing in memory. These approaches have been proposed for processing large, data-intensive applications like machine learning [46–49] and graph processing algorithms [50–52]. They could mitigate data movement and reduce memory latency. Additionally, data-level parallelism could be increased by using larger registers available in AVX-512-capable processors [53–55], which would allow for more comparisons per sequence processed simultaneously. This would be particularly useful for long sequences, which are currently limited in the number of comparisons that can be made simultaneously. By shifting to these paradigms, it is possible to accelerate the genome assembly pipeline and increase scalability with long and complex genomes. Utilizing memories with ultra-high bandwidth could also help to achieve these goals.

5. Conclusion

Genome assembly is a resource-intensive process that requires careful consideration of the sequencing technology, depth, and assembler used. Long-read sequencing technologies have shown promise in improving assembly quality while reducing computational costs, but their higher sequencing costs and lower quality raise questions about the need for hybrid strategies. Our study find that *Hifiasm* and *HiCanu* have an higher computational cost compared to the novel assembler *Shasta*. However, *Hifiasm* is more suitable for complex genome assembly from both biological and computational perspectives. We also observe that hybrid assembly strategies, such as *Wengan* and *Verkko*, show lower contiguity and higher computational costs as the genome complexity increased. While HiFi reads can provide high-quality de novo assembly at an affordable cost, our results suggest the need for further improvements in hybrid strategies to achieve more precise assemblies.

Author declaration

[Instructions: Please check all applicable boxes and provide additional information as requested.]

Funding

Funding was received for this work.
Funding for open access charge: Universidad de Málaga/CBUA.

Intellectual property

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Research ethics

Not applicable. We confirm that this study has not involved human patients.

Authorship

We confirm that the manuscript has been read and approved by all named authors.

We confirm that the order of authors listed in the manuscript has been approved by all named authors.

Declaration of Competing Interest

No conflict of interest exists.

Data availability

The input data as well as the results generated can be found in this article. Complementary data used to carry out the evaluation can also be found in supplementary

Acknowledgments

The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the Supercomputing and Bioinnovation Center (SCBI) of the University of Malaga. Also, this work has been partially supported by the Spanish MINECO PID2019-105396RB-I00, Junta de Andalucia JA2018 P18-FR-3433, and UMA18-FEDERJA-197 projects. Funding for open access charge: Universidad de Málaga/CBUA.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2023.110700.

References

[1] I. H. G. S. Consortium, Initial sequencing and analysis of the human genome, Nature 409 (6822) (2001) 860–921.
[2] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A.V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, et al., The complete sequence of a human genome, Science 376 (6588) (2022) 44–53.
[3] P. Biosciences, Pacific Biosciences. https://www.pacb.com/.
[4] O. Nanopore, Oxford Nanopore. https://nanoporetech.com/.
[5] T. Hu, N. Chitnis, D. Monos, A. Dinh, Next-generation sequencing technologies: an overview, Hum. Immunol. 82 (11) (2021) 801–811.
[6] PacBio, HIFI SEQUENCING. https://www.pacb.com/technology/hifi-sequencing/.
[7] A.M. Wenger, P. Peluso, W.J. Rowell, P.-C. Chang, R.J. Hall, G.T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N.D. Olson, et al., Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome, Nat. Biotechnol. 37 (10) (2019) 1155–1162.
[8] O. Nanopore, The power of Q20+ chemistry. https://nanoporetech.com/q20plus-chemistry.

[9] M.R. Vollger, G.A. Logsdon, P.A. Audano, A. Sulovari, D. Porubsky, P. Peluso, A. M. Wenger, G.T. Concepcion, Z.N. Kronenberg, K.M. Munson, et al., Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads, Ann. Hum. Genet. 84 (2) (2020) 125–140.

[10] G.A. Logsdon, M.R. Vollger, E.E. Eichler, Long-read human genome sequencing and its applications, Nat. Rev. Genet. 21 (10) (2020) 597–614.

[11] T. Hon, K. Mars, G. Young, Y.-C. Tsai, J.W. Karalius, J.M. Landolin, N. Maurer, D. Kudrna, M.A. Hardigan, C.C. Steiner, et al., Highly accurate long-read hifi sequencing data for five complex genomes, Scientific Data 7 (1) (2020) 1–11.

[12] J. Foox, S.W. Tighe, C.M. Nicolet, J.M. Zook, M. Byrska-Bishop, W.E. Clarke, M. M. Khayat, M. Mahmoud, P.K. Laaguiby, Z.T. Herbert, et al., Performance assessment of dna sequencing platforms in the abrf next-generation sequencing study, Nat. Biotechnol. 39 (9) (2021) 1129–1140.

[13] J. Kececioglu, E. Myers, Exact and approximate algorithms for the sequence reconstruction problem, Algorithmica 13 (7) (1995).

[14] S. Draghici, P. Khatri, A.L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, R. Romero, A systems biology approach for pathway level analysis, Genome Res. 17 (10) (2007) 1537–1545.

[15] E.W. Myers, The fragment assembly string graph, Bioinformatics 21 (suppl_2) (2005) ii79–ii85.

[16] J.T. Simpson, R. Durbin, Efficient construction of an assembly string graph using the fm-index, Bioinformatics 26 (12) (2010) i367–i373.

[17] J.T. Simpson, R. Durbin, Efficient de novo assembly of large genomes using compressed data structures, Genome Res. 22 (3) (2012) 549–556.

[18] W. Kuśmirek, W. Franus, R. Nowak, Linking de novo assembly results with long dna reads using the dnaasm-link application, Biomed. Res. Int. 2019 (2019).

[19] J.-I. Sohn, J.-W. Nam, The present and future of de novo whole-genome assembly, Brief. Bioinform. 19 (1) (2018) 23–40.

[20] M. Gavrielatos, K. Kyriakidis, D.A. Spandidos, I. Michalopoulos, Benchmarking of next and third generation sequencing technologies and their associated algorithms for de novo genome assembly, Mol. Med. Rep. 23 (4) (2021) 1.

[21] R.R. Wick, K.E. Holt, Benchmarking of long-read assemblers for prokaryote whole genome sequencing, F1000Research (2019) 8.

[22] M. Mascher, T. Wicker, J. Jenkins, C. Plott, T. Lux, C.S. Koh, J. Ens, H. Gundlach, L. B. Boston, Z. Tulpová, et al., Long-read sequence assembly: a technical evaluation in barley, Plant Cell 33 (6) (2021) 1888–1906.

[23] S. Goldstein, L. Beka, J. Graf, J.L. Klassen, Evaluation of strategies for the assembly of diverse bacterial genomes using minion long-read sequencing, BMC Genomics 20 (1) (2019) 1–17.

[24] V. Jayakumar, Y. Sakakibara, Comprehensive evaluation of non-hybrid genome assembly tools for third-generation pacbio long-read sequence data, Brief. Bioinform. 20 (3) (2019) 866–876.

[25] H. Cheng, G.T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm, Nat. Methods 18 (2) (2021) 170–175.

[26] K. Shafin, T. Pesout, R. Lorig-Roach, M. Haukness, H.E. Olsen, C. Bosworth, J. Armstrong, K. Tigyi, N. Maurer, S. Koren, et al., Nanopore sequencing and the shasta toolkit enable efficient de novo assembly of eleven human genomes, Nat. Biotechnol. 38 (9) (2020) 1044–1053.

[27] S. Nurk, B.P. Walenz, A. Rhie, M.R. Vollger, G.A. Logsdon, R. Grothe, K.H. Miga, E. E. Eichler, A.M. Phillippy, S. Koren, Hicanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads, Genome Res. 30 (9) (2020) 1291–1305.

[28] A. Di Genova, E. Buena-Atienza, S. Ossowski, M.-F. Sagot, Efficient hybrid de novo assembly of human genomes with wengan, Nat. Biotechnol. 39 (4) (2021) 422–430.

[29] M. Rautiainen, S. Nurk, B.P. Walenz, G.A. Logsdon, D. Porubsky, A. Rhie, E. E. Eichler, A.M. Phillippy, S. Koren, Verkko: telomere-to-telomere assembly of diploid chromosomes, BioRxiv (2022), 2022–06.

[30] H. Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, Bioinformatics 32 (14) (2016) 2103–2110.

[31] A. Rhoads, K.F. Au, Pacbio sequencing and its applications, Genomics, Proteomics Bioinform. 13 (5) (2015) 278–289.

[32] S.C. Shin, D.H. Ahn, S.J. Kim, H. Lee, T.-J. Oh, J.E. Lee, H. Park, Advantages of single-molecule real-time sequencing in high-gc content genomes, PLoS One 8 (7) (2013), e68824.

[33] C. Alkan, S. Sajjadian, E.E. Eichler, Limitations of next-generation genome sequence assembly, Nat. Methods 8 (1) (2011) 61–65.

[34] K. Paszkiewicz, D.J. Studholme, De novo assembly of short sequence reads, Brief. Bioinform. 11 (5) (2010) 457–472.

[35] T. Laver, J. Harrison, P. O'neill, K. Moore, A. Farbos, K. Paszkiewicz, D. J. Studholme, Assessing the performance of the oxford nanopore technologies minion, Biomol. Detect. Quantific. 3 (2015) 1–8.

[36] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K.F. Au, Nanopore sequencing technology, bioinformatics and applications, Nat. Biotechnol. 39 (11) (2021) 1348–1365.

[37] J.H. Grau, T. Hackl, K.-P. Koepfli, M. Hofreiter, Improving draft genome contiguity with reference-derived in silico mate-pair libraries, GigaScience 7 (5) (2018) giy029.

[38] A. Thrash, F. Hoffmann, A. Perkins, Toward a more holistic method of genome assembly assessment, BMC Bioinform. 21 (4) (2020) 1–8.

[39] L.A. Diambra, Genome sequence and analysis of the tuber crop potato, Nature 475 (2011).

[40] D. Tang, Y. Jia, J. Zhang, H. Li, L. Cheng, P. Wang, Z. Bao, Z. Liu, S. Feng, X. Zhu, et al., Genome evolution and diversity of wild and cultivated potatoes, Nature 606 (7914) (2022) 535–541.

[41] S.L. Amarasinghe, S. Su, X. Dong, L. Zappia, M.E. Ritchie, Q. Gouil, Opportunities and challenges in long-read sequencing data analysis, Genome Biol. 21 (1) (2020) 1–16.

[42] W.-B. Jiao, K. Schneeberger, The impact of third generation genomic technologies on plant genome assembly, Curr. Opin. Plant Biol. 36 (2017) 64–70.

[43] K. Berlin, S. Koren, C.-S. Chin, J.P. Drake, J.M. Landolin, A.M. Phillippy, Assembling large genomes with single-molecule sequencing and locality-sensitive hashing, Nat. Biotechnol. 33 (6) (2015) 623–630.

[44] H. Cheng, H. Jiang, J. Yang, Y. Xu, Y. Shang, Bitmapper: an efficient all-mapper based on bit-vector computing, BMC Bioinform. 16 (1) (2015) 1–16.

[45] P. Muir, S. Li, S. Lou, D. Wang, D.J. Spakowicz, L. Salichos, J. Zhang, G. M. Weinstock, F. Isaacs, J. Rozowsky, et al., The real cost of sequencing: scaling computation to keep pace with data generation, Genome Biol. 17 (1) (2016) 1–9.

[46] D. Kim, C. Yu, S. Xie, Y. Chen, J.-Y. Kim, B. Kim, J.P. Kulkarni, T.T.-H. Kim, An overview of processing-in-memory circuits for artificial intelligence and machine learning, IEEE J. Emerg. Select. Topics Circ. Syst. 12 (2) (2022) 338–353.

[47] M.S. Akhoon, S.A. Suandi, A. Alshahrani, A.-M.H. Saad, F.R. Albogamy, M.Z. B. Abdullah, S.A. Loan, High performance accelerators for deep neural networks: a review, Expert. Syst. 39 (1) (2022), e12831.

[48] S. Kim, H. Genc, V.V. Nikiforov, K. Asanović, B. Nikolić, Y.S. Shao, Moca: Memory-centric, adaptive execution for multi-tenant deep neural networks, in: IEEE International Symposium on High-Performance Computer Architecture (HPCA) 2023, IEEE, 2023, pp. 828–841.

[49] D.E. Kim, A. Ankit, C. Wang, K. Roy, Samba: sparsity aware in-memory computing based machine learning accelerator, IEEE Trans. Comput. 72 (2023) 2615–2627.

[50] V. Elisseev, L.-J. Gardiner, R. Krishna, Scalable in-memory processing of omics workflows, computational and structural, Biotechnol. J. 20 (2022) 1914–1924.

[51] M. Zhou, M. Li, M. Imani, T. Rosing, Hygraph: Accelerating graph processing with hybrid memory-centric computing, in: Design, Automation & Test in Europe Conference & Exhibition (DATE) 2021, IEEE, 2021, pp. 330–335.

[52] G. Dai, T. Huang, Y. Chi, J. Zhao, G. Sun, Y. Liu, Y. Wang, Y. Xie, H. Yang, Graphh: a processing-in-memory architecture for large-scale graph processing, IEEE Transactions on Comp.-Aided Design Integr. Circ. Syst. 38 (4) (2018) 640–653.

[53] R. Rahn, S. Budach, P. Costanza, M. Ehrhardt, J. Hancox, K. Reinert, Generic accelerated sequence alignment in seqan using vectorization and multi-threading, Bioinformatics 34 (20) (2018) 3437–3445.

[54] S. Gálvez, F. Agostini, J. Caselli, P. Hernandez, G. Dorado, Blvector: fast blast-like algorithm for manycore cpu with vectorization, Front. Genet. 12 (2021), 618659.

[55] T.T. Tran, Y. Liu, B. Schmidt, Bit-parallel approximate pattern matching: Kepler gpu versus xeon phi, Parallel Comput. 54 (2016) 128–138.