Genomics 116 (2024) 110842

Contents lists available at ScienceDirect

Genomics

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

Review

Advancements in long-read genome sequencing technologies and algorithms

Elena Espinosa [a], Rocio Bautista [b,*], Rafael Larrosa [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

ARTICLE INFO

2000 MSC:
0000
1111
PACS:
0000
1111
Keywords:
Genome assembly
Hybrid assembly
Long reads

ABSTRACT

The recent advent of long read sequencing technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore technology (ONT), have led to substantial improvements in accuracy and computational cost in sequencing genomes. However, *de novo* whole-genome assembly still presents significant challenges related to the quality of the results. Pursuing *de novo* whole-genome assembly remains a formidable challenge, underscored by intricate considerations surrounding computational demands and result quality. As sequencing accuracy and throughput steadily advance, a continuous stream of innovative assembly tools floods the field. Navigating this dynamic landscape necessitates a reasonable choice of sequencing platform, depth, and assembly tools to orchestrate high-quality genome reconstructions. This comprehensive review delves into the intricate interplay between cutting-edge long read sequencing technologies, assembly methodologies, and the ever-evolving field of genomics. With a focus on addressing the pivotal challenges and harnessing the opportunities presented by these advancements, we provide an in-depth exploration of the crucial factors influencing the selection of optimal strategies for achieving robust and insightful genome assemblies.

1. Introduction

In recent years, remarkable advancements in the assemblies of both genomes and transcriptomes have been driven mainly by the emergence of high-throughput sequencing technologies. *Next Generation Sequencing* (NGS) platforms, both short reads and long reads, have revolutionized the field by generating vast amounts of data in a single sequencing run. In this sense, long read sequencing, led by PacBio and ONT, offers several advantages over short read sequencing, as it allows the generation of accurate and contiguous genome assemblies. While short read sequencers such as Illumina's NovaSeq, HiSeq, NextSeq, and MiSeq instruments; BGI's MGISEQ and BGISEQ models; PacBio Onso short read platform, the Element AVITI System or Thermo Fisher's Ion Torrent sequencers produce reads of up to 600 bases, long read sequencing technologies routinely generate reads most ranging between 10 kb and 100 kb, with a current record of 2.3 Mb [1]. To illustrate the impact of NGS, consider the initial draft of the human genome achieved through Sanger Sequencing, which incurred an exorbitant cost of $3 billion and took over a decade to complete [2,3]; however, with the NGS platforms, millions or even billions of reads can be produced in a single run within a few hours or days, making it more efficient than Sanger sequencing. In this sense, in a first foray, the development of short read NGS, such as the Illumina platform, could sequence tens of thousands of genomes within a year, with a typical accuracy rate surpassing 99% [4,5]; the incorporation of unknown genomes at databases still continue to increase today. This way, Fig. 1A and B illustrate the quantity of short read (Illumina) and long read (ONT and PacBio) sequencing data archived in the NCBI, accompanied by access statistics, highlighting the importance of these technologies for the scientific community. So, these advancements in NGS technology have made large-scale genomics projects more feasible and opened up new possibilities for studying complex biological processes. The ability to obtain comprehensive genomics and transcriptomics data cost-effectively and efficiently has significantly accelerated research across various disciplines, enabling us to deepen our understanding of genetics, evolution, and disease mechanisms [6].

Despite the striking achievements of short read sequencing, as Illumina, in *de novo* assembly to constructing genomes, its limitations become evident when we need to detect long repetitive structures or long structural variants (SVs). Like that, short read methods identify rearrangements or deletions/insertions no larger than approximately

* Corresponding author.
E-mail addresses: elenamesga@uma.es (E. Espinosa), rociobm@uma.es (R. Bautista), rlarrosa@uma.es (R. Larrosa), oplata@uma.es (O. Plata).

https://doi.org/10.1016/j.ygeno.2024.110842
Received 31 October 2023; Received in revised form 1 April 2024; Accepted 6 April 2024
Available online 11 April 2024

500 bp. However, larger-size insertions constitute a significant challenge. In this sense, long reads have been shown to be more advantageous than short reads alone, because they can span longer repetitive or other problematic regions, boasting an accuracy rate of 99 9% [4,5]. Table 1 summarizes the key features and characteristics of short reads (Illumina) and long reads (PacBio and ONT).

Moving on to other issues, the cost for long read sequencing is higher to obtain the same coverage compared to short read sequencing; hence, hybrid assembly, short read plus long read, approaches have continued to be a powerful strategy for achieving highly accurate and contiguous genome assemblies without high cost [10]. The integration of these complementary data types in hybrid assembly pipelines has demonstrated performance in terms of contiguity and accuracy, facilitating the reconstruction of highly contiguous genomes with reduced gaps and misassemblies [11–14].

Given the increasing interest in long read sequencing and the rapid progress in applications and software development, the primary objective of this review is to provide a comprehensive exposition of the fundamental principles governing long read data analysis, including hybrid assembly strategies. Additionally, it aims to present a thorough survey of tools for various analytical tasks associated with long read sequencing, including hybrid assembly, while critically examining areas within long read analysis that require further refinement. In summary, long read sequencing technologies, with their ability to capture long-range information and resolve intricate genomics features, have propelled the field of genomics forward.

2. Long read sequencing technologies

The emergence of long read sequencing technologies, such as *PacBio* and ONT, has simplified genomes reconstruction and improved assembly contiguity [15–19]. A brief and comprehensive summary of the key features and characteristics of *PacBio* and *ONT* sequencing technologies is presented in Tables 1. These technologies have revolutionized the study of genomics by enabling the coverage of long repetitive regions, closing gaps in existing reference assemblies, and facilitating the characterization of structural variations (SV), many of which have been linked to various diseases [20]. So, it is especially noteworthy that the

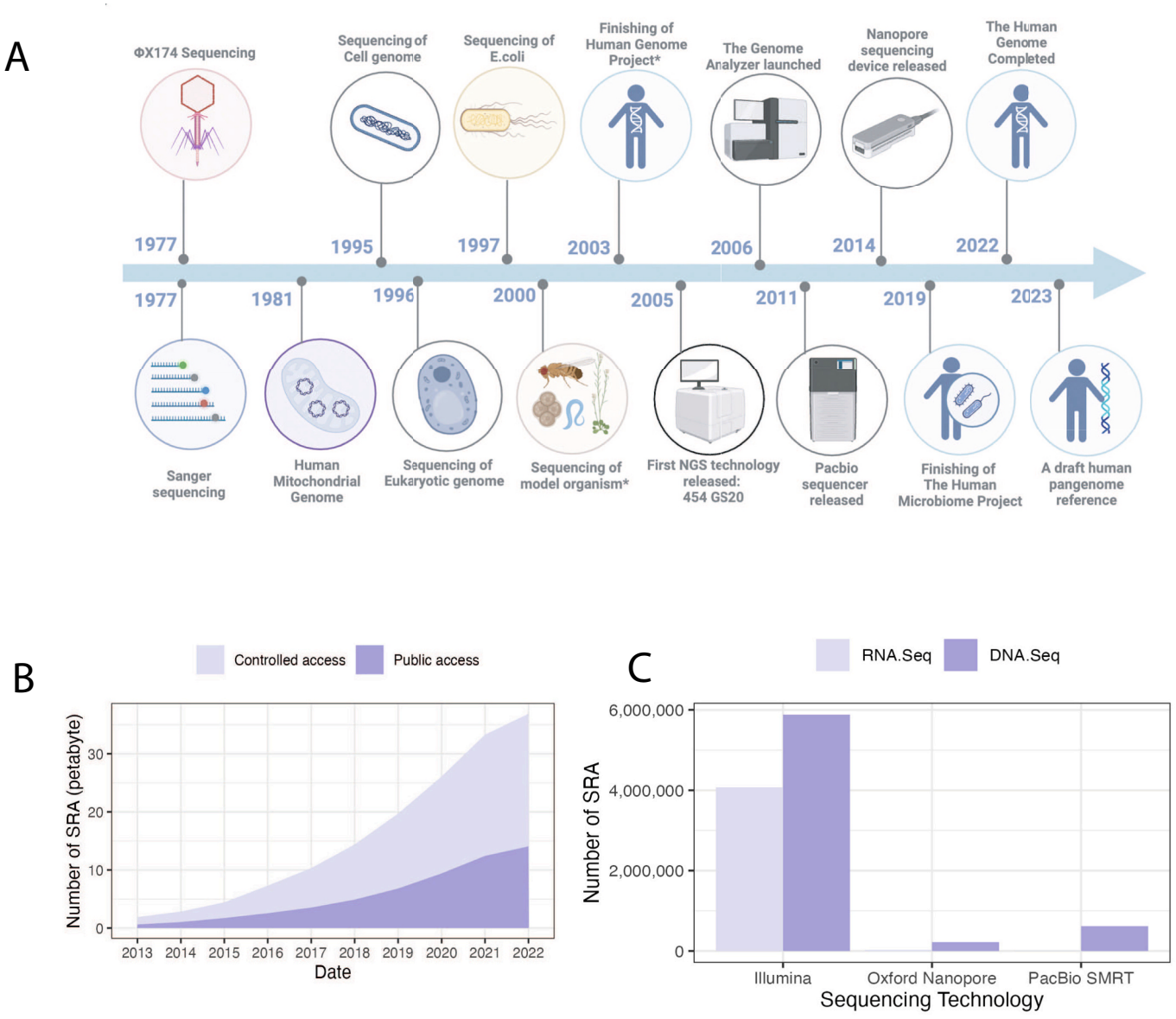

Fig. 1. Figure: Comprehensive View of Genomics Sequencing Advancements and Data Growth Trends. **A**) This figure provides an overview of the progress made in genomics sequencing technologies, specifically in the field of sequencing techniques. **B and C)** Analysis of the substantial growth in sequencing data archived within the NCBI repository.

Table 1
Overview of long read and short read sequencing technologies and platforms: An exploration of distinctive features, advantages, and applications, offering insights into their roles in genomics and research.

Company	Systems	Data type	Read Length (Maximum)	Accuracy (%)	Maximum throughput per flow cell	Sequencing cost per Gb (USD)	Equipment cost (USD)
Pacific Biosciences (PacBio)	Sequel II/Sequel IIe	PacBio CLR	>100*kb	87–92	160Gb	13–26	approximately 525,000
		PacBio HiFi	>20 kb	>99	30Gb	43–86	
	Revio	PacBio HiFi	15–20 kb	>99.9	90Gb	11	779,000
	Onso	short read	up to 2 × 150 bp[b]	> 99.9	2400–3000 Gb	15	259,000
ONT	MinION/GridION	Long	10–100 kb		48 Gb	$~14-24[d]	from ~1000/ 69,162
		Ultra-long	>100 kb			~72[d]	
	PromethION	Long	10–100 kb	>99%	50–200 Gb	$ ~ 3–4.6[d]	from ~436,404
		Ultra-long	>100 kb				
	Flongle	Long	10–100 kb		2.8 Gb	$ ~ 9.31[d]	from ~1510
		Ultra-long	>100 kb				
Illumina	NextSeq 1000 & 2000	Single-end	1 × 50 bp[c]		60Gb[c]	30/20	210,000/335,000
		Paired-end	up to 2 × 300 bp[c]		60–180 Gb[c]		
	NovaSeq 6000 series	Single-end	1 × 35 bp[d]	>99.9%	280–350 Gb[d]	10–35	985,000
		Paired-end	up to 2 × 250 bp[d]		325–400 Gb[d e]		
	NovaSeq X Series	Paired-end	up to 2 × 150 bp[f]		up to ~8 Tb[f]	2	985,000–1.25 million

[a]All cost estimates exclude the cost of labor, instrumentation, maintenance, and computer resources. [b]Read length corresponding to 300 cycle sequencing kit. [c]Output specifications based on a single flow cell using Illumina PhiX control library at supported cluster densities [7]. [d]Specifications based on Illumina PhiX control library at supported cluster densities [8]. [e]With a maximum read length of 2 × 150, the throughput is capable of reaching 2400–3000 Gb [8]. [f]Specifications based on Illumina PhiX control library or a TruSeq DNA Library created with NA12878 at supported cluster densities [9]. [g]Current cost when performing sequencing with an SMRTbell Express Template Prep Kit 2.0 and SMRT Cell 8 M. [h]Approximate cost corresponding to the price of the flow cell R.10.4.

number of SRA files in the NCBI database has increased in recent years (Fig. 1C). Despite this, several historical limiting factors, *e.g.*, limited yield, high error rates, and high cost per base, have hindered the widespread adoption of these third-generation sequencing technologies in large-scale sequencing projects. However, significant progress in recent years has mitigated these limitations, leading to both substantial reductions in error rates and overall performance improvements.

Table 2 summarizes the significant improvement in the sequencing technology from Sanger to the third sequencing technology, PacBio and ONT. These advances have opened up new possibilities and opportunities for leveraging the unique characteristics of long reads in genome assembly. It has enabled researchers can address complex genomics regions, resolve structural variations, and gain valuable insights into diseases and biological processes. As a result, long read sequencing has significantly increased innovative approaches and tools specifically designed for analyzing and assembling long reads.

2.1. Single molecule real time sequencing (SMRT)

This technology was developed by PacBio and was the first long read sequencing technology to achieve widespread deployment [21]. It is typically differentiated by two modes of SMRT sequencing: 1) the currently deprecated Continous long reads or CLR and 2) Circular Consensus Sequencing or CCS, which generates HiFi reads.

2.1.1. PacBio continuous long read

CLR reads come from the initial construction of standard SMRTbell template libraries with DNA inserts larger than 30 kb in length. Due to the large insert size in these molecules, the polymerase only performs one or a few passes around the template, which generates subreads with a typical length of 5–60 kb but can be up to 100 kb long. It provides a lower accuracy (typically of 85–92%) [21–25] concerning Illumina short reads (which reaches 99.9%). It becomes the usage of *PacBio CLR* inappropriate to detect SNVs or indels and requires the combination with another sequencing technology type (*e.g.*, hybrid assembly with short or long reads) to detect all the different types of genetic variation, which increases the complexity and the cost of the projects.

Sequel platforms can generate CLR reads with a yield of 160Gb per flow cell (compared to the 2 Gb and 20 Gb of data per flow cell achieved on RS II and Sequel platforms) with the Sequel II platform (see Table 1). However, the new system Revio is designed explicitly for high-fidelity (HiFi) long read sequencing and does not support CLR reads.

2.1.2. PacBio high-fidelity reads (PacBio HiFi)

HiFi reads exhibit exceptional accuracy (reads over 10 kb with an accuracy of over 99%). In this case, SMRTbell template libraries are assembled with smaller inserts of 10–30 kb and later sequenced *via* CCS mode. Due to the reduced length of the insert, the polymerase can perform several passes through the SMRTbell template. This leads the polymerase to produce exceptionally long reads (an N50 read length exceeding 150 kb), which have subreads from both the forward and reverse complements of the DNA template. Ultimately, the HiFi protocol enhances DNA polymerase efficiency, increasing subread throughput (over 200 Gbases compared to 100 Gbases with CLRs). However, longer run times (30h) are required for dataset generation, as accuracy relies on more passes.

Later, these subreads are merged using the CCS algorithm to generate HiFi consensus, resulting in 15–25 Gbases of HiFi data from a single SMRT Cell 8 M, underperforming PacBio CLR (see Table 1). Usually, CCS algorithms need three or four subreads of the same molecule to remove most stochastic errors and achieve a minimum accuracy of 99% [26] (as referred to in Table 1). Nonetheless, once removed, several studies report an accuracy of up to 99.9% [25], with over 99.5% of homopolymers up to five bases in length accurately [25–27].

Table 2
Types of Errors in Genomics Sequencing Technologies: An overview of error categories in genomics sequencing.

Type of Error	Description	Predominant Technology
Random base errors	Random errors in base identification.	PacBio, ONT, Illumina, Sanger
Repetitive sequence errors	Errors related to sequencing regions of the genome containing repetitive, similar sequences.	Illumina, Sanger
End-of-read quality drop	Decrease in base quality toward the end of sequencing reads.	PacBio, ONT, Illumina, Sanger
Polymerase eruptions	Issues related to the length of reads, which can be truncated or elongated unexpectedly.	PacBio, ONT
Read correction errors	Errors that may arise if error correction strategies are not applied correctly.	PacBio, ONT, Illumina, Sanger
Base call errors	Errors in the identification of bases, more pronounced in high GC content (short reads).	PacBio, ONT, Illumina
Homonucleotide stretch errors	Errors in sequences with stretches of identical nucleotides, particularly in long reads.	PacBio, ONT
InDels errors	Insertion or deletion errors in the nucleotide sequence.	PacBio, ONT, Illumina, Sanger
Gaps	Regions in the assembly where sequences are missing, possibly due to sequencing or assembly issues.	PacBio, ONT, Illumina, Sanger
Misplaced/merged haplotype errors	Errors in identifying and placing haplotypes or genetic variants, especially in long reads.	PacBio, ONT
Gene prediction errors	Errors in gene prediction from the DNA sequence, especially in long reads.	PacBio, ONT
Missing gene errors	Errors resulting in missed gene identification, especially in long reads.	PacBio, ONT
Misplaced gene synteny	Errors in the organization and arrangement of genes in the genome, especially in long reads.	PacBio, ONT

Despite the inherent accuracy, and as an example, it needs approximately three SMRT Cells 8 M to generate a 25× coverage, enough to *de novo* assembly of a human genome [26,28]. Consequently, since each SMRT Cell 8 M is run sequentially with the currently available systems (Sequel II and Sequel IIe), this process takes several days. However, these limitations are mitigated with the new system Revio (see Table 1), which reduces the run time to 24 h and supports four high-density SMRT cells (which support 25 million ZMWs *vs.* 8 million of Sequel IIe) that can run in parallel, which involves a 15× increase in HiFi read throughput. Thus, Revio can produce up to 90 Gbase per SMRT Cell, or 360 Gbases of HiFi reads per day.

All these enhancements become HiFi reads a prominent technology. The high accuracy of PacBio HiFi sequence data has had a significant impact on 1) improving variant discovery, 2) reducing the cost of the assembly, and 3) providing access to even more complex regions of repetitive DNA, including the contiguous assembly of some human centromeres. However, the cost (see Table 1) of both sequencing (likely mitigated with the new Revio system) and equipment has limited the usage of this technology. Additionally, the computational cost employed to generate HiFi data (2000 CPU hours per SMRT Cell 8 M of data with the recent improvements) could still be considered a disadvantage.

2.2. Oxford Nanopore Technologies (ONT)

ONT presents a remarkable ability to generate extensive reads up to hundreds to thousands of kilobases in length (outperforming PacBio by at least an order of magnitude). This length is achieved by the Nanopore's chemistry, which makes the displacement of molecules through the nanopore possible, regardless of their length. Studies suggest that high molecular weight DNA extraction and preparation may determine the ONT read lengths. It differentiates two main types of ONT data: 1) the conventional long read (10–100 kb) and 2) ultra-long read (>100 kb).

1. Long read: Up to now, the accuracy of the standard long read was around 87–98% logsdon2020long, with a small portion with a precision of 69% according to some studies [22]. With the most recent Q20+ platform update and the combined application of the Ligation Sequencing Kit V14, the accuracy reaches over 99%. However, ONT raw read accuracy is highly dependent on the base-calling algorithm used [25,29,30].
2. Ultra-long read: ONT ultra-long read can be up to several megabases in length [31] and present an accuracy similar to ONT long read (see Table 1). These were crucial to complete the human genome, enabling the resolution of repetitive regions that could not be resolved with other technologies [3].

Long and ultra-long nanopore sequencing reads can span complex genomic regions, like structural variants and repeats, which present several challenges to assemble accurately using traditional short read sequencing technologies. Also, long nanopore sequencing reads enhance haplotype phasing, enabling the assignment of sequencing data to maternally or paternally inherited chromosomes.

These reads can be generated on any of the three standard ONT platforms: MinION, GridION, and PromethION, which differ in the low cell capacity, with a single flow cell for MinION and up to 5 and 48 flow cells for GridION and PromethION, respectively. Moreover, the MinION and the GridION share the same type of flow cell, with 512 nanopore channels. In contrast, PromethION uses a different type of flow cell with 2675 nanopore channels. It results in a yield for the PromethION of 50–100 Gbases of ultra-long native DNA reads, and 100–200 Gbases of native genomic DNA (gDNA) reads, in contrast to the GridION, which achieves up to 48 Gbases per MinION Flow Cell (see Table 1). Moreover, since PromethION can run up to 48 independently addressable, high-capacity Flow Cells, PromethION achieves notably higher throughput than PacBio (see Section 2.1 and Table 1).

Moreover, while the handheld MinION is already established for portable DNA sequencing, ONT has recently started the development of an even smaller device known as SmidgION. It uses the same core nanopore sensing technology as MinION and PromethION but is designed for smartphones or other mobile, low-power devices. Finally, ONT provides an adapter compatible with the MinION and GridION platforms for low-throughput applications, namely Flongle (or flow cell dongle). It presents a different type of flow cell, which has 126 channels. Each nanopore channel is controlled and measured individually by the bespoke ASIC. This allows for multiple nanopores to be run in parallel. The main advantage of Flongle is that it can perform smaller, frequent, and rapid tests at a significantly lower cost than MinION or GridION flow cells.

3. *De novo* assembly

Regarding genome assembly using the advancements in sequencing technology, we encounter two strategies: 1) *reassembly*, accomplished by aligning reads to an existing reference genome, and 2) *de novo* whole-genome assembly. *De novo* whole-genome assembly can be likened to the meticulous assembly of an intricate jigsaw puzzle, where each piece corresponds to a nucleotide sequence (read) of the genome. This complex process involves the assembly of substrings or contigs, which are ingeniously pieced together to construct complete chromosomes by identifying overlapping regions between them. This intricate puzzle-solving underlines why obtaining the entirety of a genome's sequence in one go remains beyond the reach of current sequencing technologies. A key advantage of *de novo* assembly concerning the mapping against a reference is that it avoids biases arising from evolutionary differences or

genetic diversity between the reference and the sequenced genome. Remarkable projects such as the 1000 Genome Project [32], 10 k UK Genome Project [33], International Cancer Genome Consortium [34], Vertebrate Genome Project [35], Darwin Tree of Life [36], European reference Genome Atlas or ATLASeaand [37], Atlas des Génomes Marins or ATLASea [38] and 1001 Arabidopsis Genome Project [39] have effectively showcased the vast genetic variability present among individuals and cells types. These endeavors have successfully unveiled single-nucleotide variations and structural differences.

Nevertheless, *de novo* assembly has its substantial challenges. The process of genome assembly demands significant time and computational resources. While assembling small genomes, such as those of prokaryotes, might be manageable with relatively modest computing resources and time investments, the story is quite different for eukaryote genome projects. Moreover, achieving high-quality genome assembly still grapples with biological complexities. Challenges arise when the genome reconstruction involves scenarios such as 1) significant heterozygosity within the genome, 2) non-random repeat elements like long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), long terminal repeats (LTRs), and simple tandem repeats (STRs), 3) organisms with polyploid genomes. Repeated regions within the genome can introduce issues, leading to misassemblies between distant genomic regions or incorrect estimations of repeat counts. Frequently, regions of high repetition result in fragmented assemblies, as existing tools need help to navigate these regions and cease extending contigs at the boundaries of the repeats [40].

Interestingly, long read sequencing platforms, like those offered by PacBio or ONT, demonstrate improved capabilities in assembling genomes containing extensive repetitive elements. Genomes with high levels of heterozygosity can also lead to fragmented assemblies or introduce uncertainty regarding the homology of contigs, *i.e.*, whether certain contigs share an evolutionary relationship or belong to different genomic regions. Furthermore, the challenges are compounded when dealing with highly polyploid genomes.

Within the existing literature, two distinct types of long read assembly stand out: *long read-only assembly* and *hybrid assembly* (short reads and long reads). In terms of assembly strategies, three prominent methods emerge: *OLC (Overlap Layout Consensus)*, *DBG (De Bruijn Graph)*, and *SG (String Graph)*. These approaches each offer unique advantages and considerations, contributing to the complex landscape of genome assembly using the cutting-edge technologies at our disposal.

3.1. Assembly algorithms

There are two widely recognized methods for genome assembly: the *De Bruijn Graph* (DBG) approach and the *Overlap-Layout-Consensus* (OLC) method. Fig. 2 offers a visual representation of these two prominent techniques for genome reconstruction, providing a comprehensive overview of their respective processes in assembling genomes from sequencing data.

3.1.1. De Bruijn Graph (DBG)

The *De Bruijn Graph (DBG)* algorithm was initially proposed by Idbury and Waterman in 1946 to assemble sequence fragments. It involves breaking down the input sequence into multiple sub-sequences, known as *k*-mers, to identify overlaps between reads. These overlaps

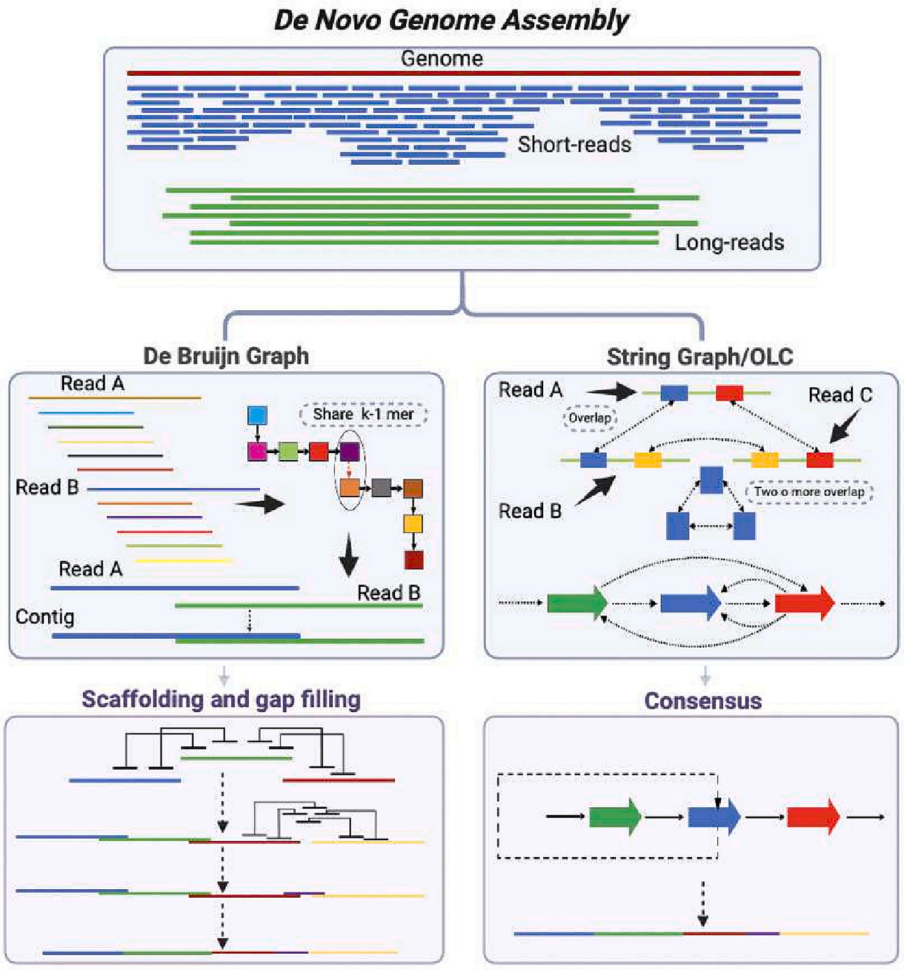

Fig. 2. Overview of *de novo* Genome Assembly Methods: An exploration of the methodologies and strategies.

construct a graph that establishes connections among all the *k*-mers. In the context of *DBG* assemblers, the graph nodes correspond to overlapping regions, while the *k*-mers form the edges. It's important to note that repeated *k*-mers create multiple edges, linking the same pair of nodes. Regions in the sequence with repetitive patterns and varying copy numbers are represented as individual nodes with numerous incoming and outgoing edges. The desired outcome is the formation of a single, unambiguous contig through which a singular path traverses the graph, visiting each edge exactly once. However, in practice, the graph often becomes divided into multiple disjointed sub-graphs due to inherent sequencing errors and incomplete coverage of *k*-mers.

The *DBG* method gained more traction with the advent of Illumina/Solexa sequencing technology. It initially successfully assembled smaller genomes, such as bacteria, and subsequently evolved to accommodate larger genomes. A critical computational advantage of *DBG* lies in its scalability, adapting to the size and intricacy of the genome. The quantity of graph nodes does not increase with multiple overlaps between different reads; instead, it grows solely by adding new *k*-mers. Nonetheless, sequencing errors introduce erroneous *k*-mers into the dataset, leading to a higher count of graph edges and increased graph complexity. This phenomenon results in bubbles or bifurcations within the graph, where incorrect paths can be identified based on *k*-mer frequency. Furthermore, this elevated complexity necessitates a larger memory footprint. Finally, an inherent drawback of the *DBG* approach is the loss of information inherent in the original reads.

3.1.2. Overlap Layout Consensus (OLC)

The Overlap Layout Consensus (*OLC*) algorithm was introduced by Staden in 1980 and became a key issue with the widespread adoption of *Sanger* sequencing technology. It consists of three steps: (1) overlap, (2) layout and (3) consensus. First, overlaps (O) between all sequencing reads are identified. Second, the OLC algorithm creates the layout (L) of all reads and overlaps information on a graph. It is a Hamiltonian path problem NP-hard in contrast to DBG assembly where inferring the contigs sequence is an Euler path problem that is easier to resolve. Finally, in the consensus (C) step, sequence is inferred.

This strategy exits two data structures: (1) the *overlap graph* and (2) the *string graph*. Kececioglu and Myers proposed the Overlap Graph (OG) and adopted a bidirectional architecture [41]. In this graph, vertices correspond to the input reads, and edges are defined by connections between reads where a suffix of one matches a prefix of another. Moreover, arrowheads on each edge denote the different ways in which two reads can overlap [42]. Conversely, the *String Graph* (SG), introduced by Myers and his collaborators [43], simplifies the classic overlap graph and removes the transitive edges. It results in a direct overlap graph [44]. Thus, a *String Graph* can be obtained from the overlap graph by eliminating duplicate and contained reads, followed by removing transitive edges from the graph.

Since identifying overlap regions between each pair of reads involves all-*versus*-all pairwise alignment between all reads, this step represents the major performance bottleneck in *OLC* assemblers. A joint proposal to mitigate this limitation has been to create a mapping from the constituent *k*-mers (subsequences of fixed length *k*) to the sequences [45,46], and only compare sequences that share common *k*-mers. However, this proposal has a long memory footprint due to the many *k*-mers generated (potentially up to 4 k *k*-mers). To address this limitation, the usage of minimizers [47] has been proposed. The main idea is to select the minimal *k*-mer (minimizer) in each sliding window of each sequence and generate a fingerprint of each one, much smaller than the original sequence. If the fingerprint of a pair of sequences presents a large overlap, an overlap is likely to exist at the sequence level. Moreover, *de novo* assembly leverages efficient data structures to compute all exact maximal pairwise suffix-prefix overlaps, *e.g.*, Burrows–Wheeler transform (BWT) [48] and FMindex [44,49]. Other recent research [50] proposes instead of computing all (irreducible) pairwise overlaps upfront, introduces multiple queries: (1) assessing one-to-one overlaps, (2) evaluating one-to-all overlaps, (3) reporting all overlaps longer than a given constant, (4) counting overlaps longer than a specified length, and (5) identifying and returning the top longest overlaps.

One of the significant advantages of *OLC* approaches is that any repetitive genome region shorter than the read length is automatically resolved. However, repeats larger than the read length can generate unresolvable ambiguous overlaps between reads from different genome parts. *OLC* algorithm addresses this issue by masking or hiding reads associated with repetitive regions. Consequently, ambiguous connections between reads from different genomics regions are ignored, leading to the breaking of the assembly graph at the start and end of extensive repetitive sequences.

3.2. Assembly strategies

We can find two approaches to performing an assembly with long reads: 1) based on long reads only, or 2) by combining shorts and/or long reads to enhance the assembly process. The assembly of long read requires the technology of PacBio (continuous long reads (*CLRs*) or high-fidelity (*HiFi*) reads) or ONT (long and ultra-long reads). In the hybrid assembly, short reads from Illumina are usually combined with long reads from PacBio and Nanopore; alternatively, long reads from both companies can be employed. Fig. 3 details the typical pipeline used in the genome assembly process, outlining the key stages and methodologies involved.

3.2.1. Long read only assembly

Two primary approaches emerge when tackling the assembly of solely long reads: Overlap Layout Consensus (*OLC*) and De Bruijn Graph (*DBG*) methods. Nevertheless, current advancements in the field tend to favor the former approach due to its compatibility with the distinct characteristics of long reads. Table 3 presents an expanded list of long read assemblers and their distinguishing features.

One notable OLC-based assembler is *Canu* [51], a fork of Celera assembler [52]. *Canu* introduces the *MinHash* algorithm and employs a sparse assembly graph construction technique. This strategy effectively avoids the difficulties of collapsing divergent repeats and haplotypes, enhancing the accuracy of the final assembly. The latest iterations of *Canu* include *HiCanu*, which was developed to assemble *HiFi* reads. *HiCanu* improves the assembly process through homopolymer compression, overlap-based error correction, and meticulous false overlap filtering steps. Similarly, another assembler, the *Hifiasm*, is explicitly designed for *HiFi* reads [18]. It ensures information preservation and the contiguity of haplotypes. *Hifiasm* accelerates the pairwise alignment by implementing a windowed version of Myer's algorithm [53], which leverages data-level parallelism. Moreover, it implements a haplotype-aware long read error correction, preserving allelic heterozygosity. Thus, the phased assembly graph constructed by *Hifiasm* presents a detailed representation of heterozygous alleles.

Likewise, *Shasta* outperforms *Canu* for the assembly of ONT data, providing a faster and cheaper way to assemble human-scale genomes. It is an example of a *DBG*-based assembler that employs a compact representation of the marker graph, with edges connecting markers in the same reads. The weight of each edge corresponds to the number of reads containing that specific marker sequence. On the other hand, *Flye* (FALTA REFERENCIA), like *Canu*, takes a versatile approach, accommodating both ONT and PacBio reads. It introduces a repeat graph framework, which approximates the behavior of *DBG* when a large *k*-mer size is utilized. This method proves particularly effective in resolving unbridged repeats, those not spanned by any individual reads. *NextDenovo* [54] also supports both PacBio and ONT reads. It follows a string graph approach and adopts a "correct-then-assemble" strategy similar to *Canu* but with a reduction of computing resources and storage. However, after assembly, *NextPolish* [55] is recommended to improve single base accuracy further.

Finally, a widely used *OLC*-based *de novo* assembler for both PacBio

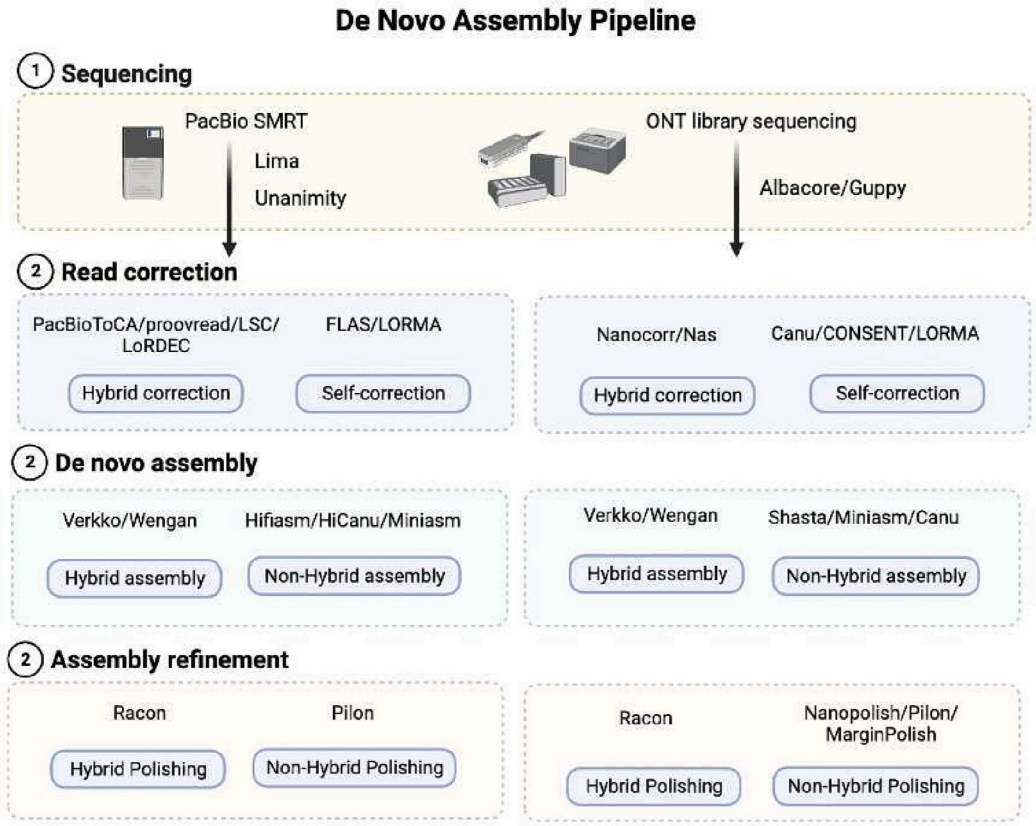

Fig. 3. Overview of *de novo* Genome Assembly Pipeline: An insight into the workflow and key components of *de novo* Genome Assembly.

and ONT reads is *Miniasm* [47]. It implements an overlap graph and leverages the Minimap [56] aligner for mapping all pairs of reads. Subsequently, it employs MinHash [57] sketches for *k*-mer comparison. Furthermore, combining different types of long reads in genome assembly provides a novel, powerful strategy to enhance genome assembly. *Verkko* [58] is a prominent assembler developed to address complex repetitive regions in the human genome by combining ONT and *HiFi* reads. *Verkko* corrects remaining errors in the *HiFi/duplex* reads using *Canu*, builds a multiplex De Bruijn Graph using *MBG* [59], and aligns the ONT reads to the graph using *GraphAligner* [60]. Progressively resolves loops and tangles using first *HiFi* reads, then aligned ONT reads, and finally, generates contig consensus sequences using the consensus module of *Canu*. It results in a phased, diploid assembly of both haplotypes, with many chromosomes automatically assembled from telomere to telomere. *Verkko* successfully assembled 20 out of 46 chromosomes using the HG002 human genome without any gaps, with an accuracy of 99.9997% [58]Also, *Wengan*, the hybrid assembler described in 3.2.2, allows the long read-only assembly of PaBio and ONT data using the mode *WenganM* and the option *-ccsont*. However, it is specifically designed for the hybrid assembly with short and long reads.

3.2.2. Hybrid assembly: integrating short and long reads

The hybrid assembly approach leverages the advantages of short and long reads for comprehensive genomics reconstructions. Combining elements from the De Bruijn Graph (*DBG*) and Overlap Layout Consensus (*OLC*) methodologies has emerged as a powerful strategy for efficiently assembling hybrid datasets. We distinct four different approaches [61,62]:

1. *Direct Mapping and Ambiguity Resolution*: Long reads can be directly mapped onto a *DBG* constructed from short reads. This initial alignment enables tackling *DBG* ambiguities, enhancing the resultant sequences' overall coherence.
2. *De Novo Assembly and Error Correction*: Using specialized assemblers, long reads are subjected to *de novo* assembly. After this assembly step, short reads are utilized to map and correct potential errors within the formed contigs.
3. *Short Read Correction and Joint Assembly*: Short reads are employed to correct long read sequences, enabling more accurate representations. The corrected long reads and original short reads are assembled using algorithms designed for third-generation sequencing data.
4. *De Novo Short Read Assembly with Long Read Linkage*: Short reads are independently assembled using specialized second-generation sequencing assemblers. Long reads are introduced to bridge the gaps and link the generated contigs.

A prime example of the first strategy is the assembler *Meraculous* [63]. The Super-Read Celera Assembler (*MaSuRCA* [64]) represents the second strategy, emphasizing the independent *de novo* assembly of long reads before enhancing contig accuracy through short read mapping. *DBG2OLC* [65] and *HASLR* use the third approach. Finally, *HybridSPAdes* [66] and *Wengan* [67], employ the last approach.

Within the realm of hybrid assembly, it is noteworthy to emphasize the advances enabled by single cell Strand-seq (Strand sequencing). Strand-seq preserves the directionality of DNA in short read sequencing libraries. This has proven to be advantageous in overcoming challenges related to genomic variability and polymorphic inversions. Chromosome-length phasing enabled by Strand-seq was shown to enable more accurate and reliable genomic reconstructions [12–14].

3.2.3. Hi-C enhanced assembly

The advances in sequencing technology have enabled assemblies to achieve long and accurate contigs. However, since they cannot deliver chromosome-scale contiguity, they can not generate quality genomes on their own. Using Hi-C data has improved the assembly in this context, leading to a genome chromosome-level assembly [68,69]. The Hi-C data

Table 3
An Overview of long read Assemblers: Techniques and Sequencing Read Types Employed for Genome Assembly.

Assembler	Algorithm	Graph Structure	Sequencing data	Documentation
Miniasm	OLC-based	Overlap graph	ONT, PacBio	Miniasm
Canu	OLC-based	Overlap graph	ONT, PacBio	Canu
HiCanu	OLC-based	Overlap graph	PacBio	Canu
Flye	DBG-based	Repeat graph	ONT, PacBio	Flye
Shasta	DBG-based	Marker graph	ONT	Shasta
Hifiasm	OLC-based	String graph	PacBio	Hifiasm
Wtdbg2	DBG-based	Fuzzy Bruijn graph	ONT, PacBio	Wtdbg2
ABruijn	DBG-based	A-Bruijn graph	ONT, PacBio	ABruijn
Marvel	OLC-based	Overlap graph	PacBio	Marvel
Raven	OLC-based	Overlap graph	ONT, PacBio	Raven
NECAT	OLC-based	String graph	ONT	NECAT
Peregrine	OLC-based	Overlap graph	PacBio	Peregrine
HINGE	OLC-based	Overlap graph	PacBio	HINGE
FALCON	OLC-based	String graph	PacBio	FALCON
FALCON-Unzip	OLC-based	String graph	PacBio	FALCON-Unzip
SMARTdenovo	OLC-based	Overlap graph	ONT + PacBio	SMARTdenovo
Verkko	DBG-based	Sparse graph	PacBio+ONT, PacBio	Verkko

allow for correcting assembly errors, complementing linked reads and optical maps for improved scaffolding of contigs, and providing chromosome-spanning contiguity to the assembly. Specifically, Hi-C data can improve genome assemblies in four paramount manners: (i) Ordering and orienting contigs, (ii) Correcting misassemblies and/or identifying structural variation, (iii) linking contigs to chromosomes, and (iv) Generating phased assemblies by leveraging predictable patterns of intra-chromosomal or inter-chromosomal interactions to group and scaffold individual haplotypes. For instance, the Telomere-to-Telomere (T2T) consortium, a few years ago, completed the sequencing of the human X chromosome from telomere to telomere (T2T), nearly twenty years after the first sequencing of the human genome in 2001 [70], thanks to the employment of ultra-long read nanopore technology, single-molecule high-fidelity (HiFi) sequencing technology, and chromosome-spanning connectivity information from the Arima High Coverage Hi-C kit. Moreover, Hi-C was used to assemble and validate the complete sequence of the human genome [3]. However, the combination of PacBio HiFi and Hi-C has been employed for the chromosome-level of multiple organisms in recent research [71–82], and the integration with long and short reads have enhanced the understanding of the structure of germline rearranged human genomes [83]. Thus, using Hi-C in the genome assembly pipeline may reduce the steps required to generate a chromosome-scale, phased genome assembly. The recent assembler Hifiasm [18,84] recently added the Hi-C data, improving the phasing and post-assembly scaffolding.

4. Error correction strategies in sequence assembly

The historical inaccuracies associated with long sequencing reads have led to the development of methods for correcting these noisy reads. Now, with the inherent improvement of long read sequencing, the usage of correction methods is controversial and relies on multiple factors such as characteristics of the sequencing sample, study type, and the selected software. This assessment is specially significant in applications like *de novo* genome assembly, where a high degree of accuracy is required, and may arise several challenges, *e.g.*, repeats, homopolymers, or regions close to the centromere.

4.1. Sequencing read error correction

Long read sequencing technologies can exhibit lower per-read accuracy. To overcome this challenge, error correction emerges as a pivotal step, ensuring the precision and reliability of the sequenced genomic data. We outline the two primary methods employed for error correction: hybrid correction and self-correction.

4.1.1. Hybrid error correction: leveraging short reads

This hybrid correction approach utilizes short read data for error correction and relies on four distinct techniques: (1) Alignment of Short Reads to Long Reads: Examples include *CoLoRMAP* [85] and HECiL [86]. This technique corrects long reads by aligning short reads to them, determining a consensus sequence from the subset of short reads linked to each long read.(2) De Bruijn Graph Exploration: Implemented by tools like *LoRDEC* [87] and *Jabba* [88], this method involves constructing a *De Bruijn Graph* from short reads. Once built, long reads can be aligned to the graph. Traversal of the graph helps identify paths linking anchored regions of long reads, facilitating correction in unanchored regions. (3) Contigs Generation and Alignment: Tools like *MiRCA* [89] and *HALC* [90] use this technique to generate contigs from short reads and align long reads to these contigs. Long reads are corrected by leveraging the consensus sequences obtained from the aligned contigs. (4) Hidden Markov Models (*e.g.*, Hercules): This model is initialized with long reads and trained with short reads, aiming to extract consensus sequences representing corrected long reads.

Additionally, some algorithms combine different strategies. For instance, *NaS* [91] combines strategies 1 and 3, while *HG-CoLoR* [92] relies on strategies 1 and 2. So, in the era of first-generation long reads, characterized by low accuracy (approximately 15–30% error rates on average), the predominant approach was using short reads due to their wider availability. However, with advancements in long read technology, self-correction has become a viable and effective alternative.

4.1.2. Self-correction: exploiting redundancy

Self-correction of long reads encompasses two primary strategies: (1) Multiple Sequence Alignment of Long Reads: This strategy parallels the hybrid approach, as discussed in Section 4.1.1, where short reads or contigs are aligned. Following alignment, long reads undergo correction by estimating a consensus sequence for each, employing a similar strategy to the hybrid approach. Several tools leverage this method, such as *PBDAGCon* [93] (the correction module of *HGAP* [93] assembler), *PBcR-BLASR* [94], *Sprai* [95], *PBcR-MHAP* [46], *FalconSense* [96] (the correction module of *Falcon* assembler), *Sparc* [97], *MECAT* [98], *FLAS* [99], the correction module used in *Canu* [51]. (2) *De Bruijn Graphs*: Similar to the hybrid strategy outlined in Section 4.1.1, this method is employed to anchor long reads once the graph is constructed. Subsequently, the graph is traversed to identify paths that join anchored regions of long reads, correcting unanchored regions.

4.2. Assembly polishing

Assembly polishing refers to the refinement of a draft sequence assembly, often a preliminary genome. This process entails analyzing the draft assembly (or region assembly) to eliminate artifacts introduced during the assembly process, thereby enhancing both local accuracy and the overall consensus accuracy of the assembled sequence. *Pilon* [100], a well-established polishing algorithm, is particularly effective when

provided with paired-end data from two Illumina libraries with small (*e.g.*, 180 bp) and large (*e.g.*, 3–5 Kb) inserts. *Nanopolish* [101–103] serves as a traditional Nanopore-based polishing tool. However, it's worth noting that *Nanopolish* does not support R10.4 flowcells (ONT), as its variant and methylation calling accuracy do not necessitate signal-level analysis. On the other hand, algorithms such as *Racon* [104] and *Medaka* [105] represent widely adopted standards for nanopore-based polishing [106]. *Racon* selects high-quality segments of reads and subsequently refines the genome through a *Partial Order Alignment* (POA) with vectorization. However, despite its correction of numerous errors, a considerable number of systematic errors persist in the genome, where the correct allele is a minority at specific loci [106]. To tackle this issue, ONT introduced Medaka, which employs a bidirectional *Long-Short-Term Memory* (LSTM) trained to rectify the systematic errors overlooked by *Racon*. As a result, the official protocol for genomes assembled from ONT reads involves initial polishing by *Racon*, followed by *Medaka*. *Racon* can also be applied to the reads from PacBio [106], although assemblers such as *Hifiasm* do not necessitate polishing, thereby streamlining the assembly pipeline and reducing processing time. Furthermore, alternative polishing pipelines for ONT data include *MarginPolish* and *HELEN* [15]. *MarginPolish* utilizes a hidden Markov model to estimate alignment statistics and constructs a weighted POA graph for processing by *HELEN*. Consequently, *HELEN* incorporates a multi-task recurrent neural network (RNN) to predict the nucleotide base and run length for each genomic position, leveraging contextual genomic features and POA weights.

Furthermore, within the realm of polishing tools for ONT data, *Homopolish* [106] emerges as an innovative approach that rectifies systematic errors, particularly indel errors in homopolymers, which existing polishing tools fail to correct. *Homopolish* achieves this by leveraging homologous sequences from related genomes. This novel polishing tool surpasses existing pipelines like *Medaka* and *HELEN* across diverse microbial genomes, providing superior accuracy in error correction.

Results demonstrate that combining *Homopolish* with *Medaka/HELEN* enhances genome quality, surpassing Q50 on R9.4 flow cells of ONT [106]. Other tools developed for ONT data in the past 1–2 years include *Nextpolish* [55], *PEPPER* [107], *Apollo* [108], and *NeuralPolish*. Despite these advancements, correcting diploid and polyploid assemblies poses a significant challenge. In diploid genomes, the consensus of a given gene could involve a mixture of the two haplotypes, potentially resulting in a premature stop codon. Addressing these challenges,*HapoG*(Haplotype Aware Polishing Of Genomes) [109] proposes a new method allowing the incorporation of phasing information from high-quality reads (short or long reads) to polish assemblies, particularly those of diploid and heterozygous genomes.

Other tools such as *Purge_dups* aims to mitigate the issues arising from haplotype divergence in regions of high heterozygosity during assembly. It leads to generate two copies rather than one copy of a region, breaking the contiguity and compromising downstream steps such as gene annotation. This tools is essential as an additional processing step in some assemblers such as *Canu*. Since, it generates a set of contigs representing all resolved alleles regardless of ploidy, you need to partition the contigs to obtain the primary and alternate allele sets.

In any case, both raw read correction and assembly polishing can be controversial and require in-depth study. *Hifieval* aims to assist HiFi assemblers in enhancing assembly quality over the long term by evaluating the over- and under-corrections produced by error-correction tools, especially in challenging regions such as homopolymer regions, centromeric regions, and segmental duplications.

5. Assessing the quality of *de novo* assembly

Beyond the obstacle posed by sequencing errors, *de novo* assemblies can harbor inaccuracies arising from factors such as the incorrect fusion of genomic regions in improper orientations or locations or the inadvertent dismissal of authentic regions as repeats, inversion, or translocation. Distinguishing between genuine errors and experimental artifacts, or even instances of missing information, can be an arduous task. Therefore, a thorough evaluation of the assembly becomes paramount. This evaluation should delve into three critical dimensions: contiguity, correctness, and completeness, providing a comprehensive understanding of the assembly's overall quality.

5.1. Contiguity assembly

This characteristic evaluates both the size and quantity of contigs. Higher contiguity suggests that genomic sequences are represented by fewer long contigs than numerous smaller contigs. The N50 parameter is the most commonly employed metric for assessing assembly quality regarding contiguity. It is defined as the sequence length of the shortest contig that covers 50% of the total genome length. In some cases, the N90 and N10 parameters are also used. Similarly, NG10, NG50, and NG90 use 10%, 50%, and 90%, respectively, of the reference genome instead of the total size of the assembled genome (such as N10, N50, and N90). Other parameters are L90, L50, and L10, which determine the count of the smallest number of contigs whose length sum makes up half of the genome size.

5.2. Completeness assembly

Completeness determines the content of the contigs, especially regarding the gene content. Completeness errors can come from the sequencing process (important genes may have yet to be sequenced) or arise during the assembly process (genes may end up in discarded contigs). Completeness is usually measured using *BUSCO* [110] (Benchmarking Universal Single-Copy Orthologs), which aims to determine the presence or absence of highly conserved genes. A *BUSCO* score above 95% is considered good. Another way to assess the completeness is by comparing the k-mers present in the short reads from the same individual that are missing in the assembly. *Merqury* [111], developed by Arang Rhie in Adam Phillippy's group, is one tool to do this. It is based on KAT [112] tool ideas, which introduced the genome assembly validation using k-mer copy number analysis. Other tools such as *GenomeScope* [113] use the k-mer information of the sequencing data to determine pre-assembly genome characteristics such as size, heterozygosity, and repetitiveness.

5.3. Correctness assembly

Correctness refers to the accuracy of each base pair in the assembled sequences. A correct assembly ensures that the order and the location of the contigs are correct. This evaluation is most often conducted by employing a gold standard reference, as an accurate genome [114,115] and covers parameters such as single or a few nucleotide polymorphisms (*e.g.*, insertions, deletions, and substitutions) or structural variations (*e.g.*, inversions, relocations, or translocations). A commonly used tool to do this analysis is *QUAST* [116]. It works with or without a genome reference and yields metrics to assess the correctness, contiguity, and completeness of the assembled genome. Other tools include *misFinder* [117], *Mauve* [118], and *REAPR* [119], which evaluate the accuracy of a genome assembly using mapped paired end reads without the use of a reference genome for comparison. Another measure of correctness is the number of frameshifting indels in coding genes. Frameshift mutations frequently interfere with the production of the protein encoded by the gene and are not usual. Therefore, most of the observed frameshifts correspond to assembly errors. This proposal is similar to *BUSCO*, but a more extensive set is used instead of analyzing a conserved set of genes. It requires a set of transcripts from the same (or very closely related) sample, commonly generated in the scope of a genome annotation. So, PacBio RNA sequencing, using the Iso-Seq method, represents a promising approach for genome annotation. Finally, *Merqury* [111] also can track error bases in the assembly. It can generate files that are uploaded

as IGV tracks, allowing users to visualize misassembles or other errors.

Since misjoins in the assembled genome result in misleading high contiguity parameters, the assessment of assembly correctness becomes crucial. Likewise, some measures of completeness are related to contiguity. For example, fragmentation of the genome (a measure of contiguity) is related to fragmentation of the genes.

6. Future perspectives

Long read sequencing technologies have transformed genome assembly, requiring careful tool and platform selection. Assessing the current state reveals progress and challenges in hybrid and non-hybrid assemblers. Recent advancements in tools like *Hifiasm*, *Shasta*, and *HiCanu* focus on PacBio and ONT reads, offering insights into their potential. Moreover, *Wengan* and *Verkko* aims to enhance the accuracy of the assembly by the integration of long reads (ONT + PacBio). Benchmarking with Miniasm anchors innovation, ensuring accuracy remains paramount. However, the chromosome-level assembly remains a challenge without introduce long-range data. Nonetheless, A promising future for long read assembly emerges. Algorithm refinements, guided by genomic insights, promise more accurate and comprehensive results. Evolving sequencing technologies could unlock solutions for complex genomic regions and structural variations. In the foreseeable future, the synergy of advanced assemblers, robust benchmarking, and cutting-edge sequencing technologies will likely drive the field toward more accurate, comprehensive, and insightful genome assemblies. The journey promises a continued unraveling of the genome's secrets, fueling discoveries in diverse fields from biology to medicine.

CRediT authorship contribution statement

Elena Espinosa: Writing – original draft, Methodology, Investigation, Conceptualization. **Rocio Bautista:** Writing – review & editing, Validation, Supervision, Methodology. **Rafael Larrosa:** Supervision. **Oscar Plata:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

No conflict of interest exists.

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors thank 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, and PID2022-136575OB-I00 projects. Funding for open access charge: Universidad de Málaga/CBUA.

References

[1] A. Payne, N. Holmes, V. Rakyan, M. Loose, Whale Watching with Bulkvis: A Graphical Viewer for Oxford Nanopore Bulk fast5 Files, bioRxiv, 2018.
[2] I. H. G. S. Consortium, Initial sequencing and analysis of the human genome, Nature 409 (6822) (2001) 860–921.
[3] 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.
[4] Illumina, Illumina. https://www.illumina.com/.
[5] N. Stoler, A. Nekrutenko, Sequencing error profiles of Illumina sequencing instruments, NAR Genom. Bioinform. 3 (1) (2021) lqab019.
[6] E. Espinosa Garcia, M. Arroyo Varela, R. Larrosa Jimenez, J. Gomez-Maldonado, M.A. Cobo Dols, M.G. Claros, R. Bautista Moreno, Construction of mirna–mrna networks for the identification of lung cancer biomarkers in liquid biopsies, Clin. Transl. Oncol. 25 (3) (2023) 643–652.
[7] Illumina, NextSeq 1000 and NextSeq 2000 Sequencing Systems. https://www.illumina.com/systems/sequencing-platforms/nextseq-1000-2000.html.
[8] Illumina, NovaSeq 6000 System. https://www.illumina.com/systems/sequencing-platforms/novaseq.html.
[9] Illumina, NovaSeq X Series. https://www.illumina.com/systems/sequencing-platforms/novaseq-x-plus.html.
[10] E. Espinosa, R. Bautista, I. Fernandez, R. Larrosa, E.L. Zapata, O. Plata, Comparing assembly strategies for third-generation sequencing technologies across different genomes, Genomics 110700 (2023).
[11] M. Rautiainen, S. Nurk, B.P. Walenz, G.A. Logsdon, D. Porubsky, A. Rhie, E. E. Eichler, A.M. Phillippy, S. Koren, Telomere-to-telomere assembly of diploid chromosomes with verkko, Nat. Biotechnol. (2023) 1–9.
[12] D. Porubsky, P. Ebert, P.A. Audano, M.R. Vollger, W.T. Harvey, P. Marijon, J. Ebler, K.M. Munson, M. Sorensen, A. Sulovari, M. Haukness, M. Ghareghani, Human Genome Structural Variation Consortium, P.M. Lansdorp, B. Paten, S. E. Devine, A.D. Sanders, C. Lee, M.J.P. Chaisson, J.O. Korbel, E.E. Eichler, T. Marschall, Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads, Nat. Biotechnol. 39 (3) (2021) 302–308, https://doi.org/10.1038/s41587-020-0719-5.
[13] P. Ebert, P.A. Audano, Q. Zhu, B. Rodriguez-Martin, D. Porubsky, M.J. Bonder, A. Sulovari, J. Ebler, W. Zhou, R. Serra Mari, F. Yilmaz, X. Zhao, P. Hsieh, J. Lee, S. Kumar, J. Lin, T. Rausch, Y. Chen, J. Ren, M. Santamarina, W. Höps, H. Ashraf, N.T. Chuang, X. Yang, K.M. Munson, A.P. Lewis, S. Fairley, L.J. Tallon, W. E. Clarke, A.O. Basile, M. Byrska-Bishop, A. Corvelo, U.S. Evani, T.-Y. Lu, M.J. P. Chaisson, J. Chen, C. Li, H. Brand, A.M. Wenger, M. Ghareghani, W.T. Harvey, B. Raeder, P. Hasenfeld, A.A. Regier, H.J. Abel, I.M. Hall, P. Flicek, O. Stegle, M. B. Gerstein, J.M.C. Tubio, Z. Mu, Y.I. Li, X. Shi, A.R. Hastie, K. Ye, Z. Chong, A. D. Sanders, M.C. Zody, M.E. Talkowski, R.E. Mills, S.E. Devine, C. Lee, J. O. Korbel, T. Marschall, E.E. Eichler, Haplotype-resolved diverse human genomes and integrated analysis of structural variation, Science 372 (6537) (Apr 2021), https://doi.org/10.1126/science.abf7117.
[14] M. Hills, E. Falconer, K. O'Neill, A.D. Sanders, K. Howe, V. Guryev, P. M. Lansdorp, Construction of whole genomes from scaffolds using single cell strand-seq data, Int. J. Mol. Sci. 22 (7) (Mar 2021), https://doi.org/10.3390/ijms22073617.
[15] 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.
[16] M. Kolmogorov, J. Yuan, Y. Lin, P.A. Pevzner, Assembly of long, error-prone reads using repeat graphs, Nat. Biotechnol. 37 (5) (2019) 540–546.
[17] 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.
[18] 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.
[19] A. Rhie, S.A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim, et al., Towards complete and error-free genome assemblies of all vertebrate species, Nature 592 (7856) (2021) 737–746.
[20] T.J. Treangen, S.L. Salzberg, Repetitive dna and next-generation sequencing: computational challenges and solutions, Nat. Rev. Genet. 13 (1) (2012) 36–46.
[21] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al., Real-time dna sequencing from single polymerase molecules, Science 323 (5910) (2009) 133–138.
[22] J. Korlach, Understanding accuracy in SMRT® sequencing. https://www.pacb.com/wp-content/uploads/2015/09/Perspective_UnderstandingAccuracySMRTSequencing1.pdf.
[23] A. Rhoads, K.F. Au, Pacbio sequencing and its applications, Genom. Proteom. Bioinform. 13 (5) (2015) 278–289.
[24] J.L. Weirather, M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X.-J. Wang, D. Buck, K.F. Au, Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis, F1000Research 6 (2017).
[25] 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.
[26] 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.
[27] M.R. Vollger, P.C. Dishuck, M. Sorensen, A.E. Welch, V. Dang, M.L. Dougherty, T. A. Graves-Lindsay, R.K. Wilson, M.J. Chaisson, E.E. Eichler, Long-read sequence and assembly of segmental duplications, Nat. Methods 16 (1) (2019) 88–94.
[28] 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.
[29] F.J. Rang, W.P. Kloosterman, J. de Ridder, From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy, Genome Biol. 19 (1) (2018) 90.
[30] R.R. Wick, L.M. Judd, K.E. Holt, Performance of neural network basecalling tools for oxford nanopore sequencing, Genome Biol. 20 (2019) 1–10.

[31] A. Payne, N. Holmes, V. Rakyan, M. Loose, Bulkvis: a graphical viewer for oxford nanopore bulk fast5 files, Bioinformatics 35 (13) (2019) 2193–2198.
[32] I. T. I. G. S. Resource, GThe 1000 Genome Project. http://www.1000genomes.org/.
[33] UK10K, 10k UK Genome Project. http://www.uk10k.org/.
[34] I. D. Portal, International Cancer Genome Consortium. http://icgc.org/.
[35] G. Consortium, The Vertebrate Genomes Project introduces a new era of genome sequencing. http://vertebrategenomesproject.org/.
[36] M. B. A. N. H. M. R. B. G. E. R. B. G. K. W. S. I. U. o. C. U. o. E. U. o. O. Earlham Institute, EMBL-EBI, Darwin Tree of Life. https://www.darwintreeoflife.org/.
[37] E. R. G. Atlas, The European Reference Genome Atlas (ERGA). https://www.erga-biodiversity.eu/.
[38] I. de Biologie de l''ENS (IBENS), ATLASea: atlas des génomes marins. https://www.atlasea.fr/en/home-page-en/.
[39] genomes, 1001 Arabidopsis Genome Project. http://1001genomes.org/.
[40] M.J. Chaisson, R.K. Wilson, E.E. Eichler, Genetic variation and the de novo assembly of human genomes, Nat. Rev. Genet. 16 (11) (2015) 627–640.
[41] J. Kececioglu, E. Myers, Exact and approximate algorithms for the sequence reconstruction problem, Algorithmica 13 (7) (1995).
[42] 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.
[43] E.W. Myers, The fragment assembly string graph, Bioinformatics 21 (suppl_2) (2005) ii79–ii85.
[44] J.T. Simpson, R. Durbin, Efficient construction of an assembly string graph using the fm-index, Bioinformatics 26 (12) (2010) i367–i373.
[45] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[46] 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.
[47] H. Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, Bioinformatics 32 (14) (2016) 2103–2110.
[48] M. Burrows, A block-sorting lossless data compression algorithm, SRS Res. Rep. 124 (1994).
[49] P. Ferragina, G. Manzini, Indexing compressed text, J. ACM (JACM) 52 (4) (2005) 552–581.
[50] G. Loukides, S.P. Pissis, S.V. Thankachan, W. Zuba, Suffix-prefix queries on a dictionary, in: 34th Annual Symposium on Combinatorial Pattern Matching (CPM 2023), 2023. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
[51] S. Koren, B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman, A.M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, Genome Res. 27 (5) (2017) 722–736.
[52] G. Denisov, B. Walenz, A.L. Halpern, J. Miller, N. Axelrod, S. Levy, G. Sutton, Consensus generation and variant detection by celera assembler, Bioinformatics 24 (8) (2008) 1035–1040.
[53] 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.
[54] J. Hu, Z. Wang, Z. Sun, B. Hu, A.O. Ayoola, F. Liang, J. Li, J.R. Sandoval, D. N. Cooper, K. Ye, et al., An efficient error correction and accurate assembly tool for noisy long reads, bioRxiv, 2023, 2023–03.
[55] J. Hu, J. Fan, Z. Sun, S. Liu, Nextpolish: a fast and efficient genome polishing tool for long-read assembly, Bioinformatics 36 (7) (2020) 2253–2255.
[56] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics 34 (18) (2018) 3094–3100.
[57] A.Z. Broder, On the resemblance and containment of documents, in: Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171), IEEE, 1997, pp. 21–29.
[58] 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.
[59] M. Rautiainen, T. Marschall, Mbg: minimizer-based sparse de bruijn graph construction, Bioinformatics 37 (16) (2021) 2476–2478.
[60] M. Rautiainen, T. Marschall, Graphaligner: rapid and versatile sequence-to-graph alignment, Genome Biol. 21 (1) (2020) 253.
[61] 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).
[62] J.-I. Sohn, J.-W. Nam, The present and future of de novo whole-genome assembly, Brief. Bioinform. 19 (1) (2018) 23–40.
[63] J.A. Chapman, I. Ho, S. Sunkara, S. Luo, G.P. Schroth, D.S. Rokhsar, Meraculous: de novo genome assembly with short paired-end reads, PLoS One 6 (8) (2011) e23501.
[64] A.V. Zimin, G. Marçais, D. Puiu, M. Roberts, S.L. Salzberg, J.A. Yorke, The masurca genome assembler, Bioinformatics 29 (21) (2013) 2669–2677.
[65] C. Ye, C.M. Hill, S. Wu, J. Ruan, Z.S. Ma, Dbg2olc: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies, Sci. Rep. 6 (1) (2016) 1–9.
[66] D. Antipov, N. Hartwick, M. Shen, M. Raiko, A. Lapidus, P.A. Pevzner, Plasmidspades: assembling plasmids from whole genome sequencing data, Bioinformatics 32 (22) (2016) 3380–3387.
[67] 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.
[68] J.O. Korbel, C. Lee, Genome assembly and haplotyping with hi-c, Nat. Biotechnol. 31 (12) (2013) 1099–1101.
[69] X. Zhang, S. Zhang, Q. Zhao, R. Ming, H. Tang, Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on hi-c data, Nat. Plants 5 (8) (2019) 833–845.
[70] K.H. Miga, S. Koren, A. Rhie, M.R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G.A. Logsdon, et al., Telomere-to-telomere assembly of a complete human x chromosome, Nature 585 (7823) (2020) 79–84.
[71] W. Zhang, Y. Yang, S. Hua, Q. Ruan, D. Li, L. Wang, X. Wang, X. Wen, X. Liu, Z. Meng, Chromosome-level genome assembly and annotation of the yellow grouper, epinephelus awoara, Scientific Data 11 (1) (2024) 151.
[72] Y.-X. Huang, X.-S. Zhu, X.-N. Chen, X.-Y. Zheng, B.-S. Su, X.-Y. Shi, X. Wang, S.-A. Wu, H.-Y. Hu, J.-P. Yu, et al., A chromosome-level genome assembly of the forestry pest coronaproctus castanopsis, Scientific Data 11 (1) (2024) 218.
[73] Z. Wei, L. Zhang, L. Gao, J. Chen, L. Peng, L. Yang, Chromosome-level genome assembly and annotation of the yunling cattle with pacbio and hi-c sequencing data, Scientific Data 11 (1) (2024) 233.
[74] J. Zheng, J. Jiang, Q. Rui, F. Li, S. Liu, S. Cheng, M. Chi, W. Jiang, Chromosome-level genome assembly of acrossocheilus fasciatus using pacbio sequencing and hi-c technology, Scientific Data 11 (1) (2024) 166.
[75] J. Liu, H. Sun, L. Tang, Y. Wang, Z. Wang, Y. Mao, H. Huang, Q. Zhang, Chromosome-level genome assembly of humpback grouper using pacbio hifi reads and hi-c technologies, Scientific Data 11 (1) (2024) 51.
[76] Q. Zeng, Z. Zhou, Q. He, L. Li, F. Pu, M. Yan, P. Xu, Chromosome-level haplotype-resolved genome assembly for takifugu ocellatus using pacbio and hi-c technologies, Scientific Data 10 (1) (2023) 22.
[77] C. Bian, C. Liu, G. Zhang, M. Tao, D. Huang, C. Wang, S. Lou, H. Li, Q. Shi, Z. Hu, A chromosome-level genome assembly for the astaxanthin-producing microalga haematococcus pluvialis, Scientific Data 10 (1) (2023) 511.
[78] Y. Chang, R. Zhang, Y. Ma, W. Sun, A haplotype-resolved genome assembly of rhododendron vialii based on pacbio hifi reads and hi-c data, Scientific Data 10 (1) (2023) 451.
[79] J. Yan, C. Zhang, M. Zhang, H. Zhou, Z. Zuo, X. Ding, R. Zhang, F. Li, Y. Gao, Chromosome-level genome assembly of the Colorado potato beetle, leptinotarsa decemlineata, Scientific Data 10 (1) (2023) 36.
[80] Z. Zheng, Z. Lai, B. Wu, X. Song, W. Zhao, R. Zhong, J. Zhang, Y. Liao, C. Yang, Y. Deng, et al., The first high-quality chromosome-level genome of the sipuncula sipunculus nudus using hifi and hi-c data, Scientific Data 10 (1) (2023) 317.
[81] J. Jin, Y. Zhao, G. Zhang, Z. Pan, F. Zhang, The first chromosome-level genome assembly of entomobrya proxima folsom, 1924 (collembola: Entomobryidae), Scientific Data 10 (1) (2023) 541.
[82] V. Jayakumar, O. Nishimura, M. Kadota, N. Hirose, H. Sano, Y. Murakawa, Y. Yamamoto, M. Nakaya, T. Tsukiyama, Y. Seita, et al., Chromosomal-scale de novo genome assemblies of cynomolgus macaque and common marmoset, Scientific Data 8 (1) (2021) 159.
[83] R. Schöpflin, U.S. Melo, H. Moeinzadeh, D. Heller, V. Laupert, J. Hertzberg, M. Holtgrewe, N. Alavi, M.-K. Klever, J. Jungnitsch, et al., Integration of hi-c with short and long-read genome sequencing reveals the structure of germline rearranged genomes, Nat. Commun. 13 (1) (2022) 6470.
[84] H. Cheng, E.D. Jarvis, O. Fedrigo, K.-P. Koepfli, L. Urban, N.J. Gemmell, H. Li, Haplotype-resolved assembly of diploid genomes without parental data, Nat. Biotechnol. 40 (9) (2022) 1332–1335.
[85] E. Haghshenas, F. Hach, S.C. Sahinalp, C. Chauve, Colormap: correcting long reads by mapping short reads, Bioinformatics 32 (17) (2016) i545–i551.
[86] O. Choudhury, A. Chakrabarty, S.J. Emrich, Hecil: a hybrid error correction algorithm for long reads with iterative learning, Sci. Rep. 8 (1) (2018) 9936.
[87] L. Salmela, E. Rivals, Lordec: accurate and efficient long read error correction, Bioinformatics 30 (24) (2014) 3506–3514.
[88] G. Miclotte, M. Heydari, P. Demeester, S. Rombauts, Y. Van de Peer, P. Audenaert, J. Fostier, Jabba: hybrid error correction for long sequencing reads, Algorithms Mol. Biol. 11 (1) (2016) 1–12.
[89] M. Kchouk, M. Elloumi, Efficient hybrid de novo error correction and assembly for long reads, in: 2016 27th International Workshop on Database and Expert Systems Applications (DEXA), IEEE, 2016, pp. 88–92.
[90] E. Bao, L. Lan, Halc: high throughput algorithm for long read error correction, BMC Bioinform. 18 (2017) 1–12.
[91] M.-A. Madoui, S. Engelen, C. Cruaud, C. Belser, L. Bertrand, A. Alberti, A. Lemainque, P. Wincker, J.-M. Aury, Genome assembly using nanopore-guided long and error-free dna reads, BMC Genomics 16 (2015) 1–11.
[92] P. Morisse, T. Lecroq, A. Lefebvre, Hg-color: hybrid graph for the error correction of long reads, Comité de programme 67 (2017).
[93] C.-S. Chin, D.H. Alexander, P. Marks, A.A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E.E. Eichler, et al., Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data, Nat. Methods 10 (6) (2013) 563–569.
[94] S. Koren, G.P. Harhay, T.P. Smith, J.L. Bono, D.M. Harhay, S.D. Mcvey, D. Radune, N.H. Bergman, A.M. Phillippy, Reducing assembly complexity of microbial genomes with single-molecule sequencing, Genome Biol. 14 (9) (2013) 1–16.
[95] M. Miyamoto, D. Motooka, K. Gotoh, T. Imai, K. Yoshitake, N. Goto, T. Iida, T. Yasunaga, T. Horii, K. Arakawa, et al., Performance comparison of second-and third-generation sequencers using a bacterial genome with two chromosomes, BMC Genomics 15 (1) (2014) 1–9.
[96] C.-S. Chin, P. Peluso, F.J. Sedlazeck, M. Nattestad, G.T. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz, et al., Phased diploid genome assembly with single-molecule real-time sequencing, Nat. Methods 13 (12) (2016) 1050–1054.

[97] C. Ye, Z.S. Ma, Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads, PeerJ 4 (2016) e2016.
[98] C.-L. Xiao, Y. Chen, S.-Q. Xie, K.-N. Chen, Y. Wang, Y. Han, F. Luo, Z. Xie, Mecat: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads, Nat. Methods 14 (11) (2017) 1072–1074.
[99] E. Bao, F. Xie, C. Song, D. Song, Flas: fast and high-throughput algorithm for pacbio long-read self-correction, Bioinformatics 35 (20) (2019) 3953–3960.
[100] B.J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S.K. Young, et al., Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, PLoS One 9 (11) (2014) e112963.
[101] N.J. Loman, J. Quick, J.T. Simpson, A complete bacterial genome assembled de novo using only nanopore sequencing data, Nat. Methods 12 (8) (2015) 733–735.
[102] J. Quick, N.J. Loman, S. Duraffour, J.T. Simpson, E. Severi, L. Cowley, J.A. Bore, R. Koundouno, G. Dudas, A. Mikhail, et al., Real-time, portable genome sequencing for ebola surveillance, Nature 530 (7589) (2016) 228–232.
[103] J.T. Simpson, R.E. Workman, P. Zuzarte, M. David, L. Dursi, W. Timp, Detecting dna cytosine methylation using nanopore sequencing, Nat. Methods 14 (4) (2017) 407–410.
[104] R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads, Genome Res. 27 (5) (2017) 737–746.
[105] O. N. T. Ltd, Medaka. https://github.com/nanoporetech/medaka.
[106] Y.-T. Huang, P.-Y. Liu, P.-W. Shih, Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing, Genome Biol. 22 (1) (2021) 1–17.
[107] K. Shafin, T. Pesout, P.-C. Chang, M. Nattestad, A. Kolesnikov, S. Goel, G. Baid, J. M. Eizenga, K.H. Miga, P. Carnevali, et al., Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks, BioRxiv (2021), 2021–03.
[108] C. Firtina, J.S. Kim, M. Alser, D. Senol Cali, A.E. Cicek, C. Alkan, O. Mutlu, Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm, Bioinformatics 36 (12) (2020) 3669–3679.
[109] J.-M. Aury, B. Istace, Hapo-g, haplotype-aware polishing of genome assemblies with accurate reads, NAR Genom. Bioinform. 3 (2) (2021) lqab034.
[110] M. Seppey, M. Manni, E.M. Zdobnov, Busco: assessing genome assembly and annotation completeness, Gene prediction: methods and protocols, 2019, pp. 227–245.
[111] A. Rhie, B.P. Walenz, S. Koren, A.M. Phillippy, Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies, Genome Biol. 21 (1) (2020) 1–27.
[112] D. Mapleson, G. Garcia Accinelli, G. Kettleborough, J. Wright, B.J. Clavijo, Kat: a k-mer analysis toolkit to quality control ngs datasets and genome assemblies, Bioinformatics 33 (4) (2017) 574–576.
[113] T.R. Ranallo-Benavidez, K.S. Jaron, M.C. Schatz, Genomescope 2.0 and smudgeplot for reference-free profiling of polyploid genomes, Nat. Commun. 11 (1) (2020) 1432.
[114] S.L. Salzberg, A.M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M.C. Schatz, A.L. Delcher, M. Roberts, et al., Gage: a critical evaluation of genome assemblies and assembly algorithms, Genome Res. 22 (3) (2012) 557–567.
[115] A. Thrash, F. Hoffmann, A. Perkins, Toward a more holistic method of genome assembly assessment, BMC Bioinform. 21 (4) (2020) 1–8.
[116] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, Quast: quality assessment tool for genome assemblies, Bioinformatics 29 (8) (2013) 1072–1075.
[117] X. Zhu, H.C. Leung, R. Wang, F.Y. Chin, S.M. Yiu, G. Quan, Y. Li, R. Zhang, Q. Jiang, B. Liu, et al., misfinder: identify mis-assemblies in an unbiased manner using reference and paired-end reads, BMC Bioinform. 16 (1) (2015) 1–16.
[118] A.E. Darling, A. Tritt, J.A. Eisen, M.T. Facciotti, Mauve assembly metrics, Bioinformatics 27 (19) (2011) 2756–2757.
[119] M. Hunt, T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, T.D. Otto, Reapr: a universal tool for genome assembly evaluation, Genome Biol. 14 (5) (2013) 1–10.