nature reviews genetics

https://doi.org/10.1038/s41576-024-00718-w

Review article

Check for updates

Genome assembly in the telomere-to-telomere era

Heng Li[1,2] ✉ & Richard Durbin[3] ✉

Abstract

Genome sequences largely determine the biology and encode the history of an organism, and de novo assembly – the process of reconstructing the genome sequence of an organism from sequencing reads – has been a central problem in bioinformatics for four decades. Until recently, genomes were typically assembled into fragments of a few megabases at best, but now technological advances in long-read sequencing enable the near-complete assembly of each chromosome – also known as telomere-to-telomere assembly – for many organisms. Here, we review recent progress on assembly algorithms and protocols, with a focus on how to derive near-telomere-to-telomere assemblies. We also discuss the additional developments that will be required to resolve remaining assembly gaps and to assemble non-diploid genomes.

Sections

[1]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [3]Department of Genetics, Cambridge University, Cambridge, UK. ✉e-mail: hli@ds.dfci.harvard.edu; rd109@cam.ac.uk

Introduction

A complete and accurate genome assembly is essential to fully understand the genetics and evolution of an organism. For many years, automated algorithms could only produce fragmented assemblies; even the reference human genome assembly[1] had hundreds of assembly gaps comprising hundreds of megabases (Mb) of highly repetitive or recently duplicated sequences. Genome assembly has thus remained a central topic in computational biology.

Current sequencing technologies typically produce contiguous sequence reads ranging in length from a hundred base pairs (bp) to a few tens of kilobases (kb); very rarely might they exceed 1 Mb in length[2]. However, most chromosomes in multicellular organisms are more than 10 Mb-long, and they can be gigabase (Gb)-long. To obtain the genome sequence of an organism, individual sequence reads that together cover the genome multiple times need to be obtained and pieced together based on sequence overlaps between them – a process called 'de novo assembly', but often abbreviated here to 'assembly' for simplicity.

Historically, there were two major strategies for whole-genome assembly: hierarchical sequencing and whole-genome shotgun sequencing (WGS). With the hierarchical sequencing strategy, tiled genomic fragments of tens to hundreds of kilobases in length are cloned, sequenced and assembled, with the order of clone sequences confirmed against a physical or genetic map of the genome. The first assembled multicellular genome – the *Caenorhabditis elegans* genome – was assembled this way[3]. The commonly used human reference genome GRCh38[1] was also mainly derived from clone-based sequencing[4]. However, constructing and maintaining a comprehensive clone library is costly and labour-intensive, and this approach is rarely, if ever, used today. With the second strategy, WGS, a genome is sheared into random fragments, the fragments are sequenced, and the genome is then reconstructed[5,6]. Because possible overlaps between reads from across the whole genome need to be considered, rather than only between smaller sets of reads from relatively short, localized clones, WGS-based assembly is more computationally challenging than clone-based assembly. Nonetheless, with improved assembly algorithms and data quality, WGS has become the dominant sequencing strategy for genome assembly.

The three critical properties of sequencing reads for assembly are length, accuracy and evenness of representation. Many of the issues that arise in designing assembly strategies involve tradeoffs among these characteristics. The reference genomes of many model organisms were initially assembled about 20 years ago using Sanger reads of a few hundred to a thousand base pairs in length, which were low-throughput and costly. High-throughput short-read sequencing platforms reduced costs by orders of magnitude[7], but they led to more fragmented assemblies owing to unresolved repeats longer than the reduced read length (typically about 150 bp; Box 1). The advent of single-molecule-sequencing in 2010 allowed much longer reads of thousands of base pairs, and it marked a turning point in sequence assembly. Although these long reads had a higher error rate (~10%) than short reads (<1%), it was practical to routinely assemble complete bacterial genomes[8–11] and to assemble human genomes to contigs with typical lengths greater than 10 Mb[12,13]. Assemblies solely based on short reads are no longer competitive with assemblies that use long reads, in terms of completeness and continuity.

However, low-accuracy long reads were still unable to resolve many repeats longer than the read length, and so, at the end of 2019, it was still not possible to assemble the entire genome of most multicellular organisms. Furthermore, almost all eukaryotes have a diploid or polyploid genome, consisting of two or more sets of haploid genomes in each living individual. In general, unless the organism is inbred (as with laboratory models) or otherwise experimentally manipulated, these haploid copies are similar but not identical. Genome assemblers developed up until 2020 are unable to reliably separate homologous haplotypes in a diploid human genome, and they produce fragmented assemblies of much lower quality than the reference genomes[14].

The arrival of accurate long reads[15] in 2019 has revolutionized the field of sequence assembly. Recent assemblers that leverage this high accuracy outperform older long-read assemblers[16–20] and routinely

Box 1

Properties of genomes that affect assembly

The main determinant of how easy it is to assemble a genome is not its size but its repeat structure. A repetitive sequence, or repeat, is a sequence that occurs multiple times in the genome. It can be computationally resolved by a long read that bridges between non-repetitive sequences on both sides of the repeated region. However, satellite repeats and segmental duplications are often longer than reads. The pericentromeric region of human chromosome 1, for example, harbours 20 Mb of satellite repeats[68], much longer than reads produced by current sequencing technologies. Fortunately, such regions have accumulated mutations over time and do not often share an identical repeat sequence over 10 kb in length. They can be resolved by long reads that are accurate enough to distinguish between inexact repeat copies (bridging between the differences between the repeats). Ribosomal DNA may be organized as long tandem arrays consisting of highly similar copies. Long ribosomal DNA arrays are among the most difficult regions to assemble[36,66].

The two homologous haplotypes in a diploid sample can also be viewed as repeats of each other. Correctly separating these two copies (or more than two, in the case of polyploids) is known as 'phasing'. For a heterozygous diploid or polyploid sample, a telomere-to-telomere assembly also implies all chromosomes are correctly phased, because phasing haplotypes and assembling repeats are related problems. An assembler capable of resolving similar repeats naturally has a high power to separate homologous haplotypes. Conversely, an assembler incapable of haplotype phasing is unable to resolve similar repeat copies. While traditional assembly algorithms collapse homologous haplotypes, current practices often preserve haplotype phasing over megabases and can produce a chromosome-scale haplotype-resolved assembly, given multiple data types.

Table 1 | Common data types for high-quality assembly

Data type	Technologies	Description	Roles
Accurate long reads	PacBio HiFi, ONT duplex	>10 kb in length; error rate <0.5%	Initial assembly graph construction; phasing over heterozygous variants that are less than 10 kb apart
Ultra-long reads	ONT ultra-long	>100 kb in length; error rate <10%	Resolving tangles; phasing through homozygous regions over 100 kb in length
Trio data	Short-read	Standard whole-genome shotgun sequencing of parents	Whole-genome phasing
Long-range data	Hi-C, Pore-C, Strand-seq	Information over 1 kb to over 10 Mb in length	Chromosomal phasing; chromosome-scale scaffolding

deliver haplotype-resolved assemblies for diploid genomes of new non-model species[21–27]. At present, an assembly of the human genome that is produced without human intervention from high-coverage long-read data can routinely exceed the completeness and the contiguity of the human reference genome GRCh38 (ref. 25), which has been manually curated and improved over the past 20 years, yet it only has simulated centromeres[28] and still lacks most telomeres[29]. The new data type and recent algorithms enable the whole-genome de novo assembly of population samples[30,31] and have led to systematic projects to sequence entire groups of species, including ultimately all eukaryotes[32–35]. Now that a human genome with each chromosome complete from telomere to telomere (T2T)[36,37] has been obtained for the first time, we expect similar quality genome assemblies to become available for a rapidly increasing number of species in years to come.

Here, we review the current practices for the near-T2T assembly of large eukaryotic genomes. We describe common data types, discuss recent assembly algorithms and explain methods for evaluating assemblies, finishing with a discussion of open challenges and future directions. Importantly, we note that homologous haplotypes in a diploid genome are effectively repeats. An assembler unable to distinguish homologous haplotypes will have difficulties in resolving long similar repeats and is unlikely to produce a near-complete assembly, even for a homozygous sample (Box 1). We will not discuss methods developed for small genomes or methods incapable of haplotype-resolved assembly[38–51].

Sequencing technologies and data types

Current efforts towards T2T assembly of diploid samples focus on accurate long-read data in combination with multiple other data types to resolve repeats and phasing at different length scales (Table 1; Box 1).

Long-read sequencing

A long-read technology produces contiguous read sequences that are typically ≥10 kb in length. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are the two companies leading the development of long-read technologies. In 2019, PacBio introduced High-Fidelity (HiFi) reads, which are 10–20 kb in length with an error rate below 0.5%[15]. This new data type has replaced PacBio's older Continuous Long Reads that have a >10% error rate. At present, HiFi reads are the core data type for high-quality assembly[16,30–33,35,52].

The most reliable solution to a near-T2T assembly involves additionally incorporating data from ONT ultra-long reads, which are ≥100 kb in length[25,27,36]. Despite their lower accuracy (~95% for most existing data), they help to resolve remaining repetitive sequences that could not be assembled with HiFi reads alone. At present, ONT ultra-long data are more expensive to obtain than HiFi data and require large amounts of input DNA (typically tens of micrograms for a human genome), so many sequencing projects do not routinely generate ultra-long data. Most existing ONT reads are simplex – that is, only one strand is sequenced. With the latest v14 ONT chemistry released in 2023, simplex reads can reach an accuracy of ~99%. Meanwhile, ONT is actively developing duplex sequencing, which sequences both strands of a DNA fragment. ONT duplex data approach PacBio HiFi in accuracy, and reads can be much longer. It will become a compelling data type once the technology matures.

It is generally observed that an at least 15-fold accurate long-read coverage per haplotype is necessary for deriving a contiguous assembly[15,16,21,22]. Ultra-long data of >100 kb at ~5-fold coverage per haplotype noticeably improves the assembly[27], although higher coverage is still preferred. It is now possible to assemble several human chromosomes to T2T using one PacBio flowcell and a couple of ONT ultra-long flowcells, costing several thousand US dollars in reagents and a few hundred US dollars for cloud computing[27]. However, for diploid genomes over 1 Gb in length or with low heterozygosity (below approximately 1%), long reads alone are typically insufficient to reconstruct all chromosomes or to completely phase entire chromosomes. In such cases, long-range data are necessary for scaffolding and phasing.

Long-range sequencing

The most widely used long-range data type is Hi-C[53–55]. A Hi-C fragment connects two loci if they are spatially close in a nucleus. Due to chromosome packing, loci closer on the same chromosome are also more likely to interact in the 3D space. As a result, if the density of Hi-C reads is high between a pair of contigs, the pair is likely to be close on the same chromosome. This information can be used to group and order contigs into chromosomes. At the same time, because Hi-C fragments are more likely to bridge two loci on the same haploid chromosome than on different homologous chromosomes, Hi-C also provides long-range phasing information. 30-fold Hi-C data coverage is usually sufficient for assembling vertebrate genomes[24,56].

Other data types that can be used for scaffolding and phasing include Pore-C[57], which is similar to Hi-C but sequenced with ONT, and Strand-seq[58,59], which is more complex to produce than with Hi-C and is not commercially available. Trio data, in which an individual and both its parents are sequenced, can also be considered as a type of long-range data and are powerful for whole-genome phasing[14,60].

Modern linked-read technologies, including stLFR[61], TELL-seq[62] and haplotagging[63], produce clusters of short reads that come from ~100 kb fragments of the genome. They require low amounts of input DNA, and their libraries only cost a few US dollars to a few tens of US dollars to produce[63]. BioNano optical restriction digest maps also provide long-range information[64]. However, these technologies do not generate contiguous nucleotide reads and are unable to resolve repeats longer than read lengths. They are thus not as powerful as ONT ultra-long reads or Hi-C in conjunction with HiFi, and are not often used for de novo assembly.

Assembly algorithms

The current recipe for near-T2T assembly involves four key steps: error correction of accurate long reads, assembly graph construction from corrected reads, graph simplification with ultra-long reads, and phasing and scaffolding with long-range data (Fig. 1). This recipe has been broadly adopted by the Darwin Tree of Life Project (DToL)[33], the Vertebrate Genomes Project (VGP)[32], the Bovine Pangenome Consortium[35], the primate telomere-to-telomere project[65] and multiple efforts on human genome sequencing[30,31,52].

For a homozygous genome (Fig. 1a), a near-complete assembly can be produced with HiFi and ultra-long data; a small number of gaps may remain in the most challenging regions of the genome, such as ribosome DNA arrays[36,66,67], satellite repeats[68] and recent long segmental duplications[69]. The T2T-CHM13 human genome was assembled with these two data types[36]. Currently, the only assemblers that can integrate HiFi and ultra-long reads are Verkko[25] and hifiasm[27], but some CHM13 chromosomes can be assembled from T2T with HiFi reads alone using Verkko, hifiasm, HiCanu[21] or LJA[23]. These assemblers are superior to earlier tools not optimized for HiFi data, mainly owing to their use of exact sequence matches between reads to resolve repeats[16–20].

For a heterozygous diploid genome, ultra-long reads and long-range data are often necessary for a T2T assembly that correctly resolves tangles and regions of low heterozygosity[25,27,36] (Fig. 1b). It is also possible to produce a high-quality, though often not T2T, assembly with fewer data types (Fig. 2). With HiFi alone, two types of assembly pairs can be produced: a primary–alternate pair[39], in which the primary assembly is the best attempt at a representative complete haploid genome and the alternate includes alleles not present in the primary (Fig. 2a); or a dual assembly pair[24,45], which aims to represent a pair of complete haploid genomes (Fig. 2b). In both cases, there may be phase switches between the two assemblies. For deriving a single reference, the primary assembly may be preferred, because primary contigs are generally longer; the alternate assembly, which is fragmented and error-prone, is usually ignored in downstream analysis. The dual assembly pair, which represents both genomes in a diploid sample,

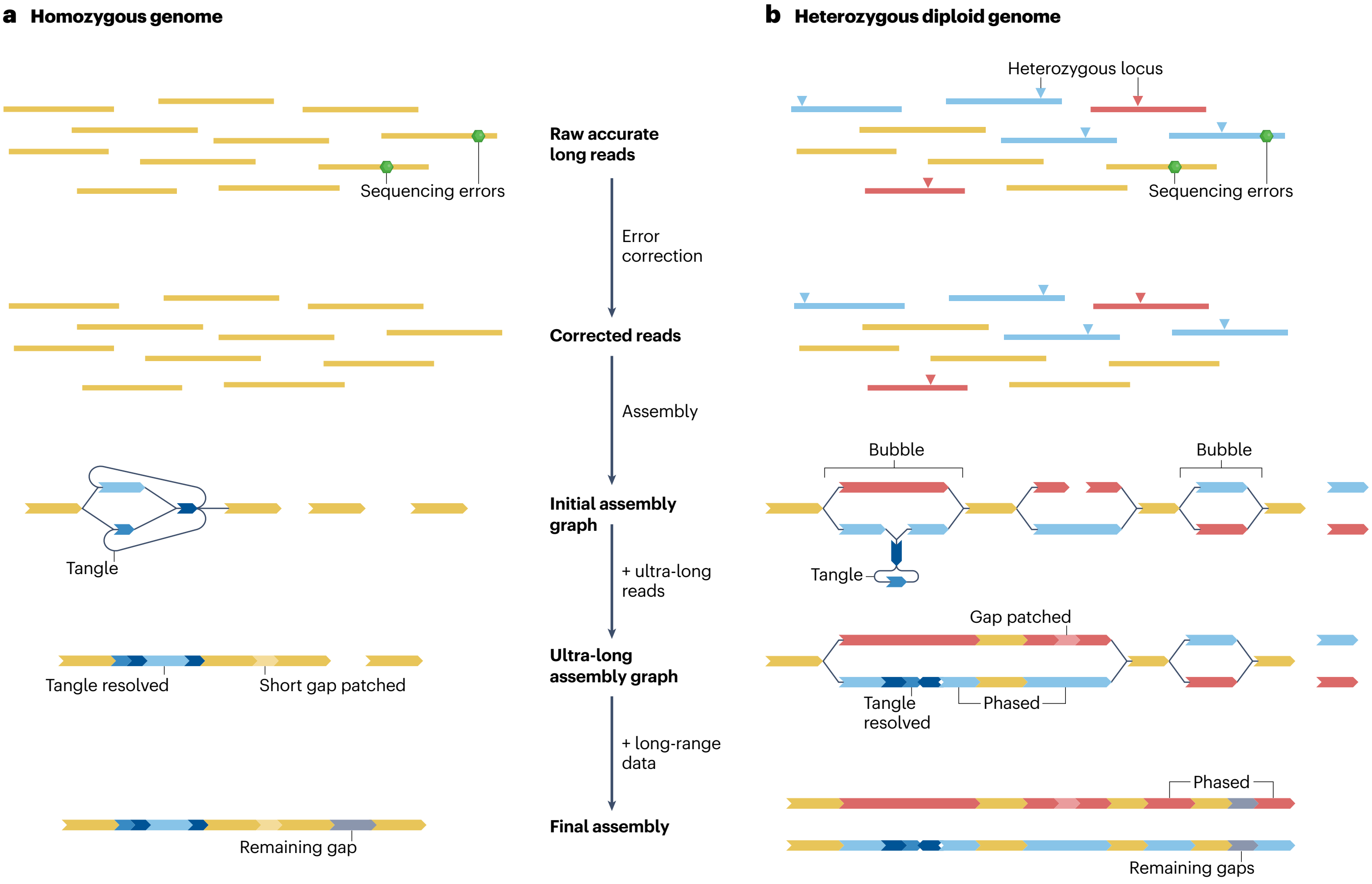

Fig. 1 | Strategy for near-telomere-to-telomere assembly. a, Assembling a haploid or homozygous genome. After sequencing errors on accurate long reads are corrected, error-free reads are assembled into an initial assembly graph, in which a thick arrow denotes a sequence and a thin line connects sequences. Ultra-long reads are then threaded through the assembly graph to resolve tangled subgraphs and patch small assembly gaps. Long-range data such as Hi-C help to scaffold across remaining gaps. **b**, Assembling a heterozygous diploid genome. Heterozygous differences between haplotypes are preserved during error correction. The assembly graphs often consist of a chain of 'bubbles', representing polymorphisms between haplotypes. Ultra-long reads and long-range data can be used to phase haplotypes as well as to resolve tangles.

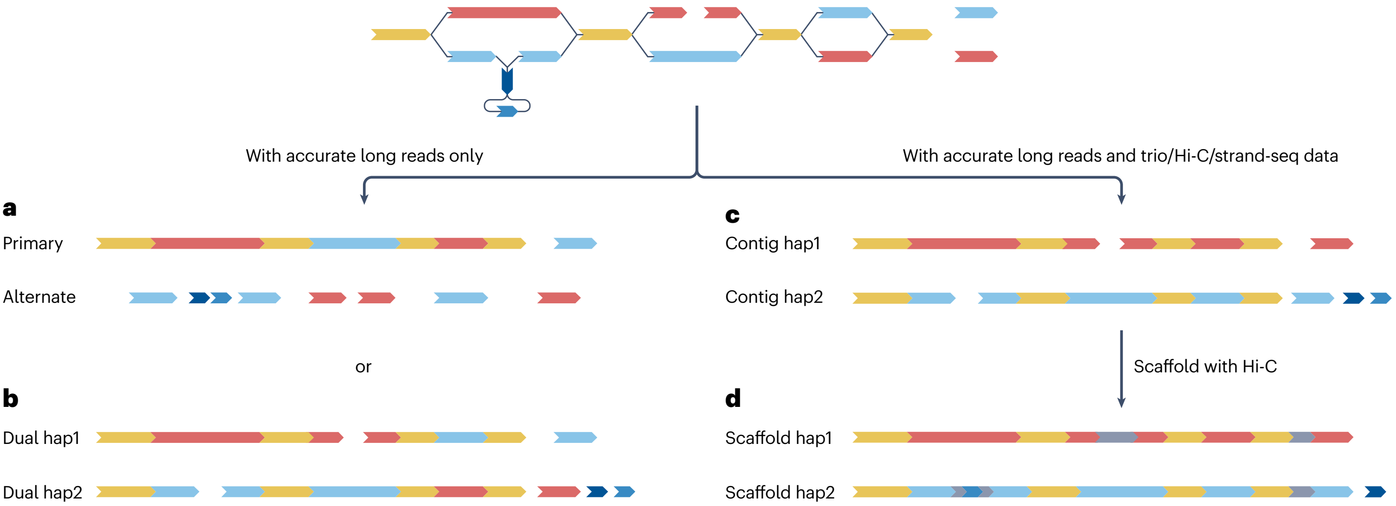

Fig. 2 | Types of phased assembly of diploid samples. The initial assembly graph can be further processed into different types of assemblies. If only accurate long-read sequence data are available, a primary–alternate assembly pair or a dual assembly pair can be generated. If long-range data (such as trio, Hi-C or strand-seq data) are available, chromosome-phased assemblies can be generated. **a**, In a primary–alternate assembly pair, the primary assembly represents a complete haploid genome with occasional phase switches whereas the alternate assembly is fragmented and contains no homozygous regions longer than the read length. **b**, In a pair of dual assemblies, each assembly is similar to a primary assembly. **c**, In a pair of chromosome-phased assemblies, long-range data are used to partition contigs from the same haploid chromosome to the same assembly. **d**, In a pair of chromosome-phased assemblies with scaffolding, Hi-C data are used to join contigs into chromosomes across assembly gaps (marked by thick grey arrows). hap, haplotype; trio data, sequencing data on an individual and both its parents.

supports assembly-based variant calling[52,70] and has been used for pangenome construction[31]. However, an assembly produced with accurate long reads alone contains shorter contigs and scaffolding can be more complex than for an assembly produced using ultra-long data. There also tends to be problems distinguishing paralogous tandem duplications within the same haplotype from homologous haplotypic duplications between haplotypes, particularly for contigs that end within long unresolved duplications or repeats. These problems can lead to false duplications[71]. For the primary assembly approach, false duplications can be found and fixed with heuristic methods that use a combination of read depth and sequence similarity to identify and discard duplicated paralogous contigs, as implemented for example in Purge Haplotigs[72] and purge_dups[73].

Correcting sequencing errors

PacBio HiFi and ONT duplex reads are accurate but not error-free. Errors are mixed with genetic variants and may impede the correct separation of homologous haplotypes or repeat copies, which would lead to fragmented assemblies. Standalone error correction tools developed for long reads ignore phasing[74–90], so all T2T-capable assemblers correct sequencing errors using their own algorithms. HiCanu, Verkko and hifiasm align all reads to each other. For each read, they correct a base if it is rarely seen among other overlapping reads aligned to the same position. LJA constructs an initial assembly graph without error correction, aligns each raw read to the graph and takes the graph path of high *k*-mer coverage as the corrected read sequence. Whereas HiCanu, Verkko and LJA compress homopolymers in reads and correct reads in the homopolymer-compressed space, hifiasm corrects errors in the original base space. All these assemblers can correct the majority of errors[91].

It is quite possible that differences in the output of these assemblers when applied to the same data depend as much on variation in error correction as variation in assembly algorithm, but this is hard to establish, because the steps are normally integrated. It would be helpful if method developers separated their error correction step from the assembly step.

Generating initial assembly graphs

Modern long-read assemblers are graph-based; in other words, they construct an assembly graph, either an overlap graph[92,93] (HiCanu and hifiasm) or a de Bruijn graph (DBG[94,95]; Verkko and LJA), from input reads. In this graph, a vertex represents a sequence, and an edge indicates a possible connection inferred from reads. An assembly graph ideally retains all information in reads without redundancy. It is, however, often nonlinear due to repeats and ploidy. Therefore, additional data types are integrated or graph traversal is implemented to resolve remaining ambiguity in the graph and obtain long linear contigs. We describe here only the basic theory of graph-based assembly. For simplicity, we assume DNA sequence only has one strand; under this assumption, assembly graphs are directed graphs, similar to classical graphs in graph theory. In practice, because DNA is double-stranded, assembly graphs are bidirected, with each edge having two directions[93].

Overlap graphs. In an overlap graph, each vertex is a read. A directed edge is added from read *A* to read *B* if a suffix of *A* can be aligned to a prefix of *B*; in this case, reads *A* and *B* are said to have an overlap. Overlapped reads can then be assembled into a longer contiguous sequence (Fig. 3a). Following other authors[5], this Review uses the word 'unitig' for these non-branching paths in the assembly graph, rather than another commonly used word, 'contig', which is a broader term that is regularly used for other purposes as well. In practice, overlap graphs will contain features that need to be further processed. In particular, there will be overlaps between different repeat copies when the repeat length is longer than the overlap length (Fig. 3b). For the human

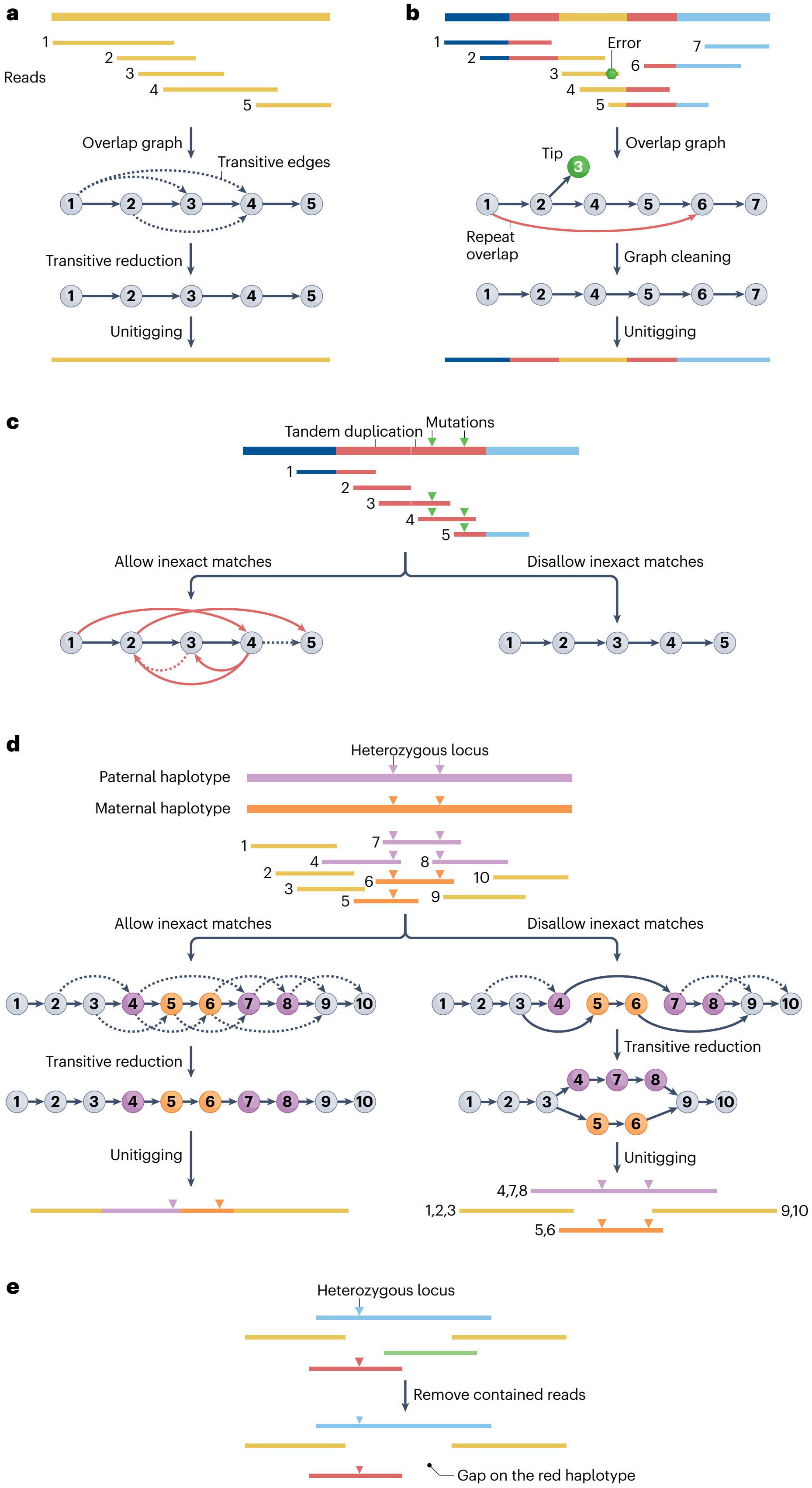

Fig. 3 | Assembly with overlap graphs. a, Simple overlap graph assembly. Find overlaps between all reads, identify transitive overlaps (dashed arrows) that can be inferred from other overlaps, remove transitive overlaps, and merge vertices with one incoming edge and one outgoing edge to get the final unitigs. **b**, Graph cleaning. An uncorrected sequencing error (green circle) may lead to a 'tip' (read 3) that should be trimmed off. Repeats (red regions) may result in overlaps between repeat copies (reads 1 and 6) that can be cut with graph cleaning. **c**, Assembling a tandem duplication longer than reads. Disallowing inexact matches removes spurious transitive edges (dashed) and unintended overlaps between repeat copies (in red). This resolves the region into a simple graph. **d**, Assembling a diploid sample. Allowing inexact overlaps leads to the loss of heterozygous differences and collapses the two haplotypes. Using only exact overlaps eliminates alignments between haplotypes and thus preserves the heterozygous alleles and their local phasing. Dashed arrows indicate transitive edges. **e**, Removing contained reads (green line) leads to assembly gaps on the red haplotype, as there is not a read path from the red read to the purple read.

genome, if overlaps shorter than a few kilobases are kept, there will be many overlaps between the ~6 kb LINE1 retrotransposon sequences that are prevalent and remain active in the human genome[96]. Nonetheless, given reads longer than LINE1, overlaps involving unique regions are expected to be longer than repetitive overlaps. If a read has two overlaps, shorter overlaps are more likely to be caused by a repeat (for example, in Fig. 3b read 1 overlaps with 6 due to a repeat) and may be discarded when graph-based algorithms attempt to simplify the initial overlap graphs later. Meanwhile, uncorrected sequencing errors may lead to extra 'tips' (for example, read 3 in Fig. 3b) or 'bubbles', which may also be removed with graph-based algorithms. Recent work demonstrated the possibility of using graph neural networks for graph layout[97], though we are not aware of extensive use of this approach in practice.

Most overlap-based assemblers follow this procedure, but HiCanu[21] and hifiasm[22], the two overlap-based assemblers optimized for HiFi reads, are distinct in that they allow only perfect overlaps. This apparently minor difference is the main source of their power to distinguish repeat copies (Fig. 3c) and phase haplotypes (Fig. 3d). In this way, they achieve a more contiguous and more accurate assembly than older assemblers, given accurate long reads[16].

It is worth noting that hifiasm implements string graphs[93], an alternative formulation of an overlap graph. As the two types of graphs can be transformed to each other without loss of information, for simplicity they are considered as the same approach.

De Bruijn graphs. There are two ways to construct a DBG: node-centric or edge-centric[98] (Fig. 4a). In a node-centric graph, DBGv(k), each vertex is a k-mer in reads, and there is an edge between two k-mers if they overlap by k−1 bases. In an edge-centric graph, DBGe(k), each edge is a k-mer in reads, and each vertex is a (k−1)-mer. Mathematically, DBGe($k+1$) is a subgraph of DBGv(k), whereas DBGv(k) is the line graph of DBGe(k), and both definitions are common in the literature. This Review takes a node-centric view of DBGs owing to its connection to overlap graphs: a node-centric DBG is an overlap graph consisting of k-mers at vertices, with edges corresponding to k−1 bp overlaps. It does not have contained reads or transitive edges and is thus simpler. Strategies used for overlap graphs are often applicable to DBGs.

A basic DBG discards information longer than the k-mer size. This reduces its power for phasing and repeat resolution. It may be tempting to choose a large k to retain long-range information, but using a large k increases the chance of contig breakpoints in low-coverage regions. There is no single best k-mer size for all situations. The multiplex DBG[99,100] provides a good solution to the dilemma of k-mer selection. Conceptually, a multiplex DBG can be thought of as the merger of multiple DBGs constructed with different k-mer sizes (Fig. 4b). It adaptively chooses a large k in repetitive regions and a small k in low-coverage regions. Nonetheless, using a multiplex DBG does not resolve all the ambiguities in DBGs (Fig. 4c). In practice, assemblers heuristically use different sets of k-mer sizes in different subgraphs[23,25]. They retrieve reads used in a subgraph and replace the subgraph with a new subgraph constructed with longer k-mers, if the new subgraph is simpler and remains contiguous. This procedure is closer to the algorithm used in the EULER assembler[95] than to the conceptual definition of multiplex DBG.

Minimizer-based sparsification[101,102] is another technique utilized in modern assemblers. Instead of storing every k-mer in reads, only a small subset of all k-mers – minimizers[103] or similar concepts[104,105] – are retained in memory. This strategy greatly reduces memory and speeds up construction. A related, but distinct, construction called the minimizer-space DBG[26,106] uses k consecutive minimizers, called k-min-mers, as k-mers to construct a DBG. MetaMDBG[106] implements multiplex minimizer-space DBG, which is multiplex DBG of k-min-mers.

Although most short-read assemblers use the DBG approach, no assemblers use DBG for assembling noisy long reads with >5% error rate, because long k-mers of 1 kb or more are needed to distinguish repeats, but even a 20-mer would have one error on average at an error rate of 5%. Nonetheless, with accurate long reads, most sequencing errors can be corrected and k-mers over 10 kb in length can be used. DBG again becomes a viable choice. Verkko[25,101] and LJA[23] are DBG-based assemblers that can construct initial assembly graphs of comparable quality to HiCanu and hifiasm, when given long-read HiFi data.

Integrating ultra-long reads

In the initial assembly graph produced above, two heterozygous positions can be phased only if there is a read that harbours both sites. The diploid genomes of humans and many other species contain many homozygous regions that are longer than HiFi reads and thus these genomes cannot be phased with HiFi reads alone[22,25]. Moreover, segmental duplications that happened within the past few thousand years are likely to remain identical over tens of kilobases and would not be resolved by HiFi reads either. ONT ultra-long reads are important to extend phased blocks and to assemble recent duplications.

To integrate ultra-long reads, Verkko aligns them to the initial assembly graph[107] and identifies paths through the graph. It then simplifies the initial graph with the same algorithm for constructing multiple DBGs. Hifiasm instead performs ultra-long-to-graph alignment with an algorithm similar to minigraph[108]. It encodes an ultra-long read as a sequence of unitigs and applies the overlap-based assembly algorithm in the unitig space. Although the two assemblers use distinct algorithms, they both produce a simplified assembly graph with fewer nodes and a more linear topology after integrating ultra-long reads. For human data, the new graph will have resolved most of the complex regions and contain phased regions averaging more than 10 Mb in length[25,27].

Integrating long-range data

When data from parents are available, assemblers can apply the trio binning algorithm[14,22,60] to achieve whole-genome phasing (Fig. 2c). The initial version of this algorithm identifies k-mers only occurring in one of the parents, marks the parental origin of each read, and assembles paternal and maternal reads separately. This strategy works for all long-read assemblers. Modern assemblers, including Verkko and hifiasm, instead mark the unitigs in the assembly graph and produce long paternal (or maternal) contigs by connecting paternal (or maternal) unitigs and unmarked unitigs in the assembly graph. Such graph-based trio binning is more accurate[16,22] and may also help to resolve subgraphs in which the within-haplotype divergence is smaller than the between-haplotype divergence. When available, trio data give extremely reliable long-range phasing[16,24]. In practice, however, parental data may be difficult to obtain due to ethical concerns in humans or because the parents are not available for wild-caught animals. In principle, data from other close relatives can provide similar information, but this type of information is rarely used for assembling the genomes of new species.

When trio data are not available, Hi-C is critical for chromosome-scale phasing (Fig. 2c). Hifiasm and GFAse (described in a recent preprint[109]) adapted reference-based Hi-C phasing algorithms[110,111] for de novo assemblies. They inspect the Hi-C-to-contig alignments and

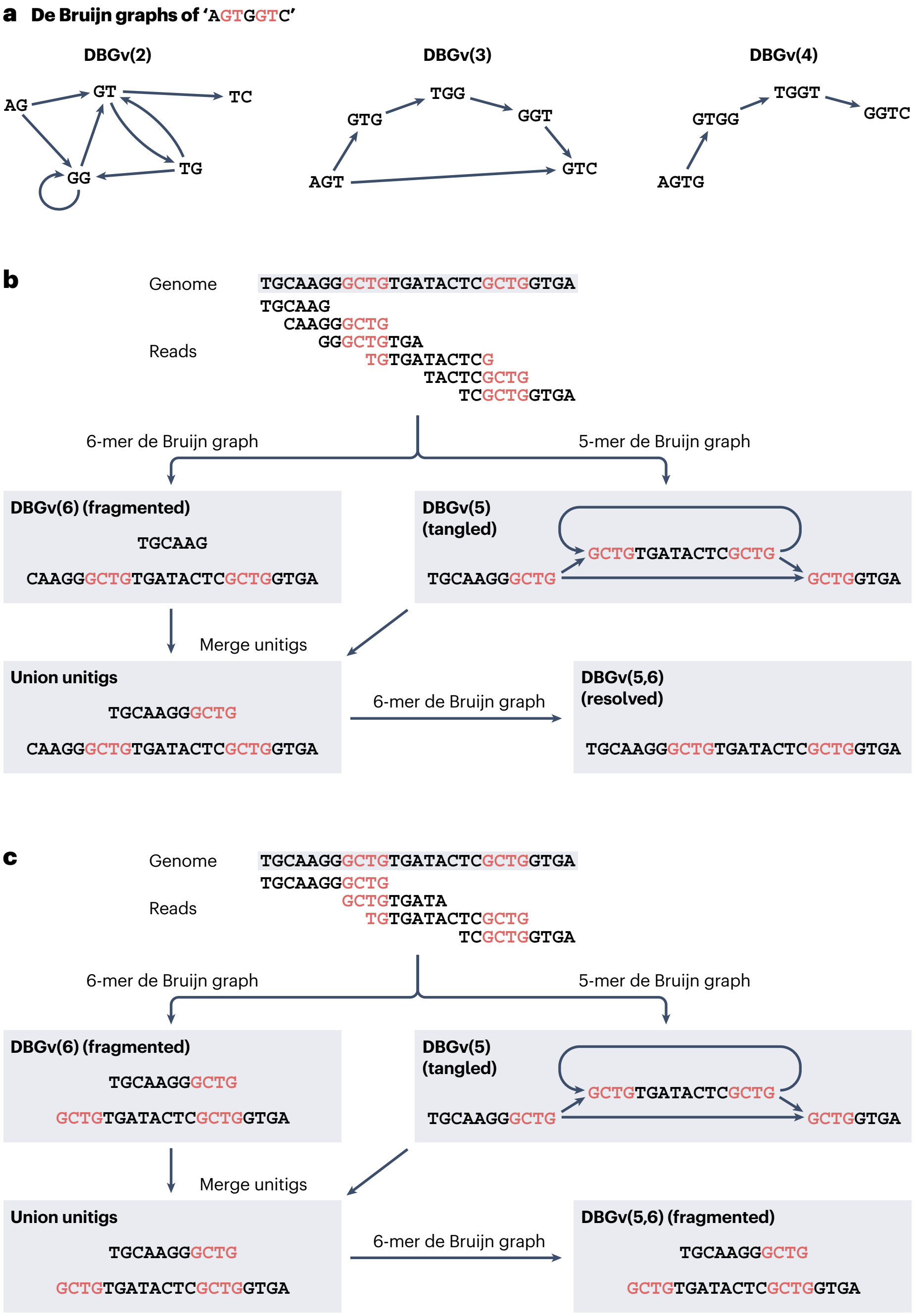

Fig. 4 | De Bruijn graphs. a, Node(vertex)-centric de Bruijn graphs (DBGs) of a string of different *k*-mer lengths. Longer *k*-mers help to resolve the 2 bp repeat in red. **b**, Multiplex DBG can improve assembly. The compacted DBG using 6-mers as nodes, DBGv(6), is fragmented into two unitigs due to the lack of 5 bp overlaps between the first two reads. DBGv(5) has one connected component, but the graph has a cycle due to the 4 bp repeat in red. A multiplex DBG, DBGv(5,6), is conceptually the 6-mer DBG on the union of unitigs from both DBGv(5) and DBGv(6), with contained unitigs removed. It has a single contig. **c**, For a different set of reads from the same locus, multiplex DBG results in two unitigs because of an assembly gap around a repeat.

apply an attractive force between two contigs if they share matches to the same Hi-C fragments, and a repulsive force if they have high sequence similarity. The attractive force tends to group together contigs with the same phase whereas the repulsive force pushes homologous contigs to opposite phases. Hifiasm and GFAse attempt to find a balance between the two types of forces to phase contigs. Verkko can optionally take GFAse phasing and phase entire chromosomes. Furthermore, it is possible to identify the parental origin of these chromosomes using imprinted methylation makers[112] if they are known and sufficiently frequent to mark each homologous pair of contigs.

In addition to phasing, Hi-C also plays a key role in scaffolding if there remain gaps in the assembly graph (Fig. 2d). There are several scaffolders optimized for high-quality HiFi assemblies[113–115]. Both the VGP[32] and the DToL[33] use YaHS[114] for scaffolding.

Polishing contig sequences

Polishing refers to the procedure implemented to improve the base accuracy of contig sequences. For a homozygous genome, wrong contig bases would not be supported by read alignment[116] or by read *k*-mers[117] and can thus be identified. For a haplotype-resolved assembly of a diploid genome, wrong contig bases can be identified by comparing variant calls from the assembly-to-reference alignment and from the read-to-reference alignment. Importantly, polishing tools developed for unphased assemblies[118–124] are not typically applied to a haplotype-resolved assembly, as they mix reads from different haplotypes and repeat copies and hence can reduce the contig base accuracy[21].

Evaluating sequence assemblies

For an assembly to be truly T2T, it must both cover the whole of each chromosome without gaps and be free from large-scale assembly errors. It is critical to rigorously assess the quality of the assembly before concluding it is T2T.

Basic metrics

Assembly size (the sum of all contig lengths) and N50 (defined as the length for which contigs no shorter than this number cover half of the assembly) are commonly calculated to get a first impression of the quality of an assembly. For the autosomes of a diploid sample, the two assemblies in a pair of dual (Fig. 2b) or chromosome-phased assemblies (Fig. 2c) are expected to have similar sizes. A pair of unbalanced autosomal assemblies may indicate incomplete phasing and may benefit from manual parameter tuning or curation[32]. Of course, in the heterogametic sex (that is, XY males in mammals or ZW females in birds or other animals) the sex chromosomes will most likely have different sizes. Other ploidy variation within species can also occur, for example owing to somatic chromosome loss or diminution[125].

Evaluating gene completeness

BUSCO[126] remains a valuable tool for evaluating the completeness of an assembly. It works by aligning conserved single-copy proteins to the genome[127] and counting alignments that are missing, broken or duplicated: the less complete the assembly, the more proteins are unaligned. A caveat is that BUSCO may underestimate the completeness of large genomes owing to limitations in its ability to accurately align protein sequence to the genome. For example, the BUSCO completeness of human gene annotations is 99.2%, but the BUSCO completeness of the GRCh38 genome is only 95.7%. Compleasm[128] is a recent reimplementation of BUSCO that addresses this problem with more accurate protein-to-genome alignment[129].

The asmgene tool from the minimap2 package[22,130] is an alternative to BUSCO and also solves the low-completeness problem when a high-quality reference genome is present. This tool identifies single-copy genes based on cDNA-to-reference alignment and additionally evaluates whether a multi-copy gene in the reference genome is assembled to multiple copies in the target assembly. Noisy read assemblers may successfully assemble single-copy genes but often miss multi-copy genes[22].

Although genes are biologically important, they only contribute to a small percentage of the genome content. Gene-based evaluation, especially single-copy gene-based evaluation, omits complex genomic regions and should be complemented by alternative evaluation methods.

K-mer based evaluation

Given uniform read coverage and a perfect assembly of the reads, we expect the count of a *k*-mer in the assembly to be proportional to its count in reads. A *k*-mer occurring more often in assembly than in reads suggests a false duplication in the assembly, whereas a *k*-mer having high frequency in the reads but low frequency in the assembly suggests missing sequences. KAT[131] is a powerful tool that makes use of these simple observations to evaluate an assembly.

It is now a common practice to use *k*-mers to estimate the base accuracy of contig sequences, often measured in the Phred scale[132] as quality value. This method works by calculating the fraction of contig *k*-mers that are absent from reads. A higher fraction suggests a lower quality value. Currently, there are two implementations, Merqury[133] and yak[22]. It is worth noting that because quality value estimation depends on the quality and the depth of input reads, quality value estimates based on different input reads are not strictly comparable. It is not easy to conclude the quality value of one species to be higher than that of another species. In addition, it is common to use reads produced with different technologies from the same sample to measure quality value. Although this approach helps to reduce the effect of systematic sequencing errors, it may underestimate quality value if there are coverage gaps in reads. When there are trio data, both Merqury and yak can also use parent-specific *k*-mers to evaluate the phasing accuracy of an assembly.

Alignment-based evaluation

Ideally, when sequence reads are aligned to their assembly[130,134–137], even coverage is expected at every contig position. Excessively low or high coverage over a long region would indicate a potential assembly error. Contigs are also expected to be well supported by reads at base level. When variants are called from the read-to-assembly alignment, isolated small variant calls would indicate contig consensus errors, whereas clustered heterozygous variant calls could result from collapsed segmental duplications. Such signals played a crucial role in evaluating the homozygous CHM13 genome. For a diploid genome, the two haplotype assemblies may be merged and reads mapped to the merge. Similar signals should be observed. Flagger[30] and Inspector[19] are user-facing evaluation tools based on read-to-assembly alignment.

For a sample with a near-perfect curated assembly, such as CHM13 (ref. 36), the existing assembly can be taken as the ground truth to evaluate automated assemblies. QUAST[138] is a popular tool for this purpose. Such methods based on assembly-to-assembly alignment are invaluable for assembler developers to tune assembly algorithms but are not applicable to new species or when the 'truth' assembly and the evaluated assembly were derived from different strains or samples. In the case in which the truth assembly and the evaluated assembly had been derived from different strains or samples, QUAST would not robustly distinguish genuine differences between two samples from assembly errors. An assembly more complete in complex regions would contain more structural variants and would seem to have a higher error rate. For example, if the human reference genome GRCh38 is taken as ground truth to evaluate the T2T-CHM13 assembly, QUAST would report over 20,000 misassemblies[22].

Challenges in de novo sequence assembly

Despite the progress outlined above, de novo sequence assembly is not a solved problem. All mainstream assemblers are built upon basic assembly algorithms established by 1995 and heavily rely on hand-tuned heuristics that do not have a solid theoretical foundation. Limited by the characteristics of practical data, they cannot resolve the most complex regions in genomes. They also perform poorly with polyploid genomes or more complex cases, such as aneuploid cancer genomes with large-scale copy number changes and rearrangements.

Theoretical challenges

Each of the two assembly paradigms, overlap graph and DBG, has its own caveats. When constructing an overlap graph, a read is discarded if it is contained within longer reads. This apparently straightforward step may lead to assembly gaps when reads are variable in length (Fig. 3e). Such assembly gaps are infrequent, but, as modern assemblies are highly contiguous, additional assembly gaps caused by contained reads are noticeable. Containment removal is the Achilles' heel of overlap-based assembly algorithms[139–141] and remains an open and critical problem.

Leading DBG assemblers, including SPAdes[100] for short reads and Verkko[25] and LJA[23] for long reads, all use multiplex DBG, which is distinct from the basic DBG described in textbooks. Although constructing multiplex DBG from a fixed set of *k*-mers across all input reads has been theoretically studied[142–144], practical DBG assemblers do not use these algorithms, because one *k*-mer set may not work optimally across all subgraphs (Fig. 4c). Practical DBG assemblers resort to heuristics and walk a fine line between graph contiguity and complexity. It will be interesting to see if, moving forwards, a new assembly paradigm is developed that can combine different length scales more smoothly.

Practical challenges

Because Hi-C only provides relative phasing information between contigs, assembly using Hi-C as long-range data is more difficult than assembly using trio data. Without trio data, current assemblers may have trouble resolving complex cases such as microchromosomes[71,145,146] and residue tetraploidy[147]. Moreover, they may not cleanly separate the sex chromosomes in heterogametic samples[24,56]. Nonetheless, by inspecting Hi-C alignment, human curators from the VGP and DToL can often identify these issues and manually fix them. This suggests that there is further room for improvement in using Hi-C data, perhaps using machine learning approaches.

The current T2T strategy emphasizes the use of long reads with accuracy well above 99%. Although Canu[38], Flye[42] and Shasta[40] could produce contigs of tens of megabases for the homozygous CHM13 genome with ONT reads of >5% error rate, they generated more structural errors: 24–75% of multi-copy genes in the finished T2T-CHM13 genome were misassembled by these assemblers[22], in comparison with ~1% with HiCanu or hifiasm, which also reconstructed more bacterial artificial chromosome sequences correctly[21,22]. Assemblies of noisy long reads did not compete with assemblies of accurate long reads in general[16–20]. However, ONT is rapidly improving the accuracy of simplex reads, and we suspect it will be possible to devise error correction strategies for these new improved reads that allow use of current or adapted long read assembly algorithms that require exact matches. If this works, there is a prospect of accurate T2T assembly from a single data type, which would greatly streamline high-quality genome assembly.

In practice, experimental challenges often have a bigger impact on the quality of assembly than computational challenges. Standard protocols for long-read generation require large amounts of DNA (>1 μg), which can be hard or impossible to obtain from small organisms or clinical samples. There are unamplified low-input library protocols for PacBio that work down to 0.1 μg[148], but, below this threshold, whole-genome amplification is necessary, which introduces coverage bias and dropouts. There are also coverage biases with long reads. For example, PacBio HiFi reads struggle to sequence through long GA-rich repeats, which was shown to lead to over 25% of the assembly gaps in a recent study[149], whereas ONT reads have been reported to have trouble with telomere sequences[150].

Beyond diploid samples

Although excellent progress has been made recently on T2T assembly of diploid genomes, there is not a satisfactory solution for polyploid genomes. There have been two recent haplotype-phased assemblies of the tetraploid potato, which used single-cell sequencing data[151] or genetic maps[152] for phasing, but these are not methods that can be deployed at scale. It would be preferable to derive polyploid assemblies using common data types; in principle, the necessary information should be present in HiFi long reads, ultra-long reads and Hi-C data. Cancer genomes are polyploid to some extent, with ploidy varying between or within chromosomes, making them even harder to assemble. Beyond cancer, it would be beneficial to be able to assemble metagenomic samples, which contain a large variety of species, typically microbial, at very different relative abundances. A metagenome can also be considered as a polyploid genome with even higher ploidy variation. Nonetheless, when assembling a metagenome sample, the bar is lower in comparison with polyploid genome assembly: for example, it is not normally expected to phase highly similar genomes. There are dedicated metagenome assemblers, such as MetaFlye[153], hifiasm-meta[154] and metaMDBG[106], that can reconstruct up to a few hundred closed bacterial genomes from a deeply sequenced metagenome sample, although 16S or *k*-mer profiling indicate that there are missing species[155]. There is still a long way to go to achieve complete metagenome assembly.

Conclusions

Thanks to the availability of PacBio HiFi reads and ONT ultra-long reads, the quality of de novo assembly has improved dramatically over the past two years. Now, a fully automated assembler can phase and assemble some chromosomes from T2T for diploid mammals and other species with large genomes. This was unthinkable in mid-2020.

Can we automatically assemble all chromosomes in most genomes from T2T with current data? We think the answer to be generally no. We believe that most of the advances over the past few years have been made because of improvements in data quality. Current assemblers can correctly reconstruct long segments of genome sequence, but the remaining assembly gaps, especially in diverse non-human genomes, remain hard to patch automatically. Algorithm improvement alone may not reliably resolve all genomes with the currently available data. We look forward to continuing new advances in sequencing technologies to truly complete a genome without human intervention.

It is important to note that a complete assembly is only the first step towards downstream biological discoveries. Although genome assembly has progressed rapidly, whole-genome alignment and gene annotation continue to have major challenges[156–158]. We hope to see continued development of downstream genome analysis tools in the future to realize the full power of (near-) complete assembly.

Published online: 22 April 2024

References

1. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
2. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
3. C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
4. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Myers, E. W. et al. A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
6. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
7. Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
8. Chin, C.-S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
9. Koren, S. et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**, R101 (2013).

10. Koren, S. et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
11. Koren, S. & Phillippy, A. M. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* **23**, 110–120 (2015).
12. Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
13. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
14. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
15. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
16. Jarvis, E. D. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
    This work evaluates 23 developer-submitted assemblies of a diploid human sample and demonstrates the advantage of accurate long-read assembly.
17. Espinosa, E. et al. Comparing assembly strategies for third-generation sequencing technologies across different genomes. *Genomics* **115**, 110700 (2023).
18. Gavrielatos, M., Kyriakidis, K., Spandidos, D. A. & Michalopoulos, I. Benchmarking of next and third generation sequencing technologies and their associated algorithms for de novo genome assembly. *Mol. Med. Rep.* **23**, 251 (2021).
19. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with inspector. *Genome Biol.* **22**, 312 (2021).
20. Eché, C. et al. A *Bos taurus* sequencing methods benchmark for assembly, haplotyping, and variant calling. *Sci. Data* **10**, 369 (2023).
21. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
    This seminal paper reports the first T2T human genome.
22. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
    This paper describes hifiasm, a widely used assembler that produces high-quality assembly by integrating multiple data types.
23. Bankevich, A., Bzikadze, A. V., Kolmogorov, M., Antipov, D. & Pevzner, P. A. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat. Biotechnol.* **40**, 1075–1081 (2022).
    This paper describes the application of multiplex DBG to accurate long-read assembly.
24. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
25. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
    This paper describes Verkko, a tool that integrates PacBio HiFi and ONT ultra-long data for automated high-quality assembly.
26. Ekim, B., Berger, B. & Chikhi, R. Minimizer-space de Bruijn graphs: whole-genome assembly of long reads in minutes on a personal computer. *Cell Syst.* **12**, 958–968.e6 (2021).
27. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2306.03399 (2023).
28. Miga, K. H. et al. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014).
29. Stong, N. et al. Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline. *Genome Res.* **24**, 1039–1050 (2014).
30. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
31. Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
32. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
    This paper presents 16 chromosomal assemblies of diverse vertebrate species, highlighting the improvements in assembly quality derived from long-read assembly.
33. Darwin Tree of Life Project Consortium. Sequence locally, think globally: the Darwin Tree of Life Project. *Proc. Natl Acad. Sci. USA* **119**, e2115642118 (2022).
34. Lewin, H. A. et al. The Earth Biogenome Project 2020: starting the clock. *Proc. Natl Acad. Sci. USA* **119**, e2115635118 (2022).
35. Smith, T. P. L. et al. The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. *Genome Biol.* **24**, 139 (2023).
36. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
37. Rhie, A. et al. The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
38. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
39. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
40. Shafin, K. et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).
41. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).
42. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
43. Vaser, R. & Šikić, M. Time- and memory-efficient genome assembly with Raven. *Nat. Comput. Sci.* **1**, 332–336 (2021).
44. Di Genova, A., Buena-Atienza, E., Ossowski, S. & Sagot, M.-F. Efficient hybrid de novo assembly of human genomes with WENGAN. *Nat. Biotechnol.* **39**, 422–430 (2021).
45. Chin, C.-S. & Khalak, A. Human genome assembly in 100 minutes. Preprint at *bioRxiv* https://doi.org/10.1101/705616 (2019).
46. Xiao, C.-L. et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).
47. Chen, Y. et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat. Commun.* **12**, 60 (2021).
48. Hu, J. et al. An efficient error correction and accurate assembly tool for noisy long reads. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.09.531669 (2023).
49. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
50. Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A. & Tse, D. N. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Res.* **27**, 747–756 (2017).
51. Lin, Y. et al. Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl Acad. Sci. USA* **113**, E8396–E8405 (2016).
52. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
53. Selvaraj, S., R. Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).
54. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
55. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).
56. Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**, 309–312 (2021).
57. Deshpande, A. S. et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nat. Biotechnol.* **40**, 1488–1499 (2022).
58. Falconer, E. et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
59. Porubsky, D. et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308 (2021).
60. Malinsky, M., Simpson, J. T. & Durbin, R. trio-sga: facilitating de novo assembly of highly heterozygous genomes with parent-child trios. Preprint at *bioRxiv* https://doi.org/10.1101/051516 (2016).
61. Wang, O. et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.* **29**, 798–808 (2019).
62. Chen, Z. et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. *Genome Res.* **30**, 898–909 (2020).
63. Meier, J. I. et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proc. Natl Acad. Sci. USA* **118**, e2015005118 (2021).
64. Lam, E. T. et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
65. Makova, K. D. et al. The complete sequence and comparative analysis of ape sex chromosomes. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.30.569198 (2023).
66. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, eabi7489 (2021).
67. Wang, B. et al. High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. *Genom. Proteom. Bioinform.* **20**, 4–13 (2022).
68. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
69. Vollger, M. R. et al. Increased mutation and gene conversion within human segmental duplications. *Nature* **617**, 325–334 (2023).
70. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
71. Ko, B. J. et al. Widespread false gene gains caused by duplication errors in genome assemblies. *Genome Biol.* **23**, 205 (2022).
72. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform.* **19**, 460 (2018).
73. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
74. Das, A. K., Goswami, S., Lee, K. & Park, S.-J. A hybrid and scalable error correction algorithm for indel and substitution errors of long reads. *BMC Genom.* **20**, 948 (2019).
75. Holley, G. et al. Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biol.* **22**, 28 (2021).
76. Au, K. F., Underwood, J. G., Lee, L. & Wong, W. H. Improving PacBio long read accuracy by short read alignment. *PLoS ONE* **7**, e46679 (2012).
77. Salmela, L. & Rivals, E. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* **30**, 3506–3514 (2014).
78. Hackl, T., Hedrich, R., Schultz, J. & Förster, F. *proovread*: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004–3011 (2014).

79. Madoui, M.-A. et al. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genom.* **16**, 327 (2015).
80. Goodwin, S. et al. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* **25**, 1750–1756 (2015).
81. Miclotte, G. et al. Jabba: hybrid error correction for long sequencing reads. *Algorithms Mol. Biol.* **11**, 10 (2016).
82. Haghshenas, E., Hach, F., Sahinalp, S. C. & Chauve, C. CoLoRMap: correcting long reads by mapping short reads. *Bioinformatics* **32**, i545–i551 (2016).
83. Salmela, L., Walve, R., Rivals, E. & Ukkonen, E. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* **33**, 799–806 (2017).
84. Bao, E. & Lan, L. HALC: high throughput algorithm for long read error correction. *BMC Bioinform.* **18**, 204 (2017).
85. Bao, E., Xie, F., Song, C. & Song, D. FLAS: fast and high-throughput algorithm for PacBio long-read self-correction. *Bioinformatics* **35**, 3953–3960 (2019).
86. Wang, J. R., Holt, J., McMillan, L. & Jones, C. D. FMLRC: hybrid long read error correction using an FM-index. *BMC Bioinform.* **19**, 50 (2018).
87. Mak, Q. X. C., Wick, R. R., Holt, J. M. & Wang, J. R. Polishing de novo nanopore assemblies of bacteria and eukaryotes with FMLRC2. *Mol. Biol. Evol.* **40**, msad048 (2023).
88. Morisse, P., Lecroq, T. & Lefebvre, A. Hybrid correction of highly noisy long reads using a variable-order de Bruijn graph. *Bioinformatics* **34**, 4213–4222 (2018).
89. Firtina, C., Bar-Joseph, Z., Alkan, C. & Cicek, A. E. Hercules: a profile HMM-based hybrid error correction algorithm for long reads. *Nucleic Acids Res.* **46**, e125 (2018).
90. Zhang, H., Jain, C. & Aluru, S. A comprehensive evaluation of long read error correction methods. *BMC Genom.* **21**, 889 (2020).
91. Guo, Y., Feng, X. & Li, H. Evaluation of haplotype-aware long-read error correction with hifieval. *Bioinformatics* **39**, btad631 (2023).
92. Myers, E. W. Toward simplifying and accurately formulating fragment assembly. *J. Comput. Biol.* **2**, 275–290 (1995).
93. Myers, E. W. The fragment assembly string graph. *Bioinformatics* **21**, ii79–ii85 (2005).
94. Idury, R. M. & Waterman, M. S. A new algorithm for DNA sequence assembly. *J. Comput. Biol.* **2**, 291–306 (1995).
95. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA* **98**, 9748–9753 (2001).
96. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
97. Vrček, L., Bresson, X., Laurent, T., Schmitz, M. & Šikić, M. Learning to untangle genome assembly with graph convolutional networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2206.00668 (2022).
98. Chikhi, R., Limasset, A. & Medvedev, P. Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics* **32**, i201–i208 (2016).
99. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
100. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
101. Rautiainen, M. & Marschall, T. MBG: minimizer-based sparse de Bruijn Graph construction. *Bioinformatics* **37**, 2476–2478 (2021).
102. Ye, C., Ma, Z. S., Cannon, C. H., Pop, M. & Yu, D. W. Exploiting sparseness in de novo genome assembly. *BMC Bioinform.* **13**, S1 (2012).
103. Roberts, M., Hayes, W., Hunt, B. R., Mount, S. M. & Yorke, J. A. Reducing storage requirements for biological sequence comparison. *Bioinformatics* **20**, 3363–3369 (2004).
104. Edgar, R. Syncmers are more sensitive than minimizers for selecting conserved *k*-mers in biological sequences. *PeerJ* **9**, e10805 (2021).
105. Kille, B., Garrison, E., Treangen, T. J. & Phillippy, A. M. Minmers are a generalization of minimizers that enable unbiased local Jaccard estimation. *Bioinformatics* **39**, btad512 (2023).
106. Benoit, G. et al. High-quality metagenome assembly from long accurate reads with metaMDBG. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01983-6 (2024).
107. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
108. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
109. Lorig-Roach, R. et al. Phased nanopore assembly with Shasta and modular graph phasing with GFAse. Preprint at *bioRxiv* https://doi.org/10.1101/2023.02.21.529152 (2023).
110. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
111. Tourdot, R. W., Brunette, G. J., Pinto, R. A. & Zhang, C.-Z. Determination of complete chromosomal haplotypes by bulk DNA sequencing. *Genome Biol.* **22**, 139 (2021).
112. Akbari, V. et al. Parent-of-origin detection and chromosome-scale haplotyping using long-read DNA methylation sequencing and Strand-seq. *Cell Genom.* **3**, 100233 (2023).
113. Zeng, X. et al. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.18.567668 (2023).
114. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
This paper describes the current state of the art Hi-C scaffolding method.
115. Garg, S. Towards routine chromosome-scale haplotype-resolved reconstruction in cancer genomics. *Nat. Commun.* **14**, 1358 (2023).
116. Mc Cartney, A. M. et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).
117. Formenti, G. et al. Merfin: improved variant filtering, assembly evaluation and polishing via *k*-mer validation. *Nat. Methods* **19**, 696–704 (2022).
118. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
119. Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput. Biol.* **16**, e1007981 (2020).
120. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
121. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
122. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
123. Morisse, P., Marchet, C., Limasset, A., Lecroq, T. & Lefebvre, A. Scalable long read self-correction and assembly polishing with multiple sequence alignment. *Sci. Rep.* **11**, 761 (2021).
124. Hu, J. et al. NextPolish2: a repeat-aware polishing tool for genomes assembled using HiFi long reads. *Genom. Proteom. Bioinform.* https://doi.org/10.1093/gpbjnl/qzad009 (2024).
125. Du, K. et al. The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. *Nat. Ecol. Evol.* **4**, 841–852 (2020).
126. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
127. Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48 (2020).
128. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, btad595 (2023).
129. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
130. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
131. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
132. Ewing, B. & Green, P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
133. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
134. Jain, C. et al. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
135. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods* **19**, 705–710 (2022).
136. Mikheenko, A., Bzikadze, A. V., Gurevich, A., Miga, K. H. & Pevzner, P. A. TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. *Bioinformatics* **36**, i75–i83 (2020).
137. Bzikadze, A. V., Mikheenko, A. & Pevzner, P. A. Fast and accurate mapping of long reads to complete genome assemblies with VerityMap. *Genome Res.* **32**, 2107–2118 (2022).
138. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
139. Hui, J., Shomorony, I., Ramchandran, K. & Courtade, T. A. Overlap-based genome assembly from variable-length reads. In *2016 IEEE International Symposium on Information Theory (ISIT)* 1018–1022 (IEEE, 2016).
140. Jain, C. Coverage-preserving sparsification of overlap graphs for long-read assembly. *Bioinformatics* **39**, btad124 (2023).
141. Kamath, S. S., Bindra, M., Pal, D. & Jain, C. Telomere-to-telomere assembly by preserving contained reads. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.07.565066 (2023).
142. Boucher, C., Bowe, A., Gagie, T., Puglisi, S. J. & Sadakane, K. Variable-order de Bruijn graphs. In *2015 Data Compression Conference* 383–392 (IEEE, 2015).
143. Belazzougui, D., Gagie, T., Mäkinen, V., Previtali, M. & Puglisi, S. J. Bidirectional variable-order de Bruijn graphs. In *LATIN 2016: Theoretical Informatics* (eds Kranakis, E. et al.) 164–178 (Springer, 2016).
144. Díaz-Domínguez, D., Onodera, T., Puglisi, S. J. & Salmela, L. Genome assembly with variable order de Bruijn graphs. Preprint at *bioRxiv* https://doi.org/10.1101/2022.09.06.506758 (2022).
145. Ohno, S., Christian, L. C. & Stenius, C. Nucleolus-organizing microchromosomes of *Gallus domesticus*. *Exp. Cell Res.* **27**, 612–614 (1962).
146. Smith, J. et al. Differences in gene density on chicken macrochromosomes and microchromosomes. *Anim. Genet.* **31**, 96–103 (2000).
147. Allendorf, F. W. et al. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J. Hered.* **106**, 217–227 (2015).
148. Lawniczak, M. K. N. et al. Standards recommendations for the Earth BioGenome Project. *Proc. Natl Acad. Sci. USA* **119**, e2115639118 (2022).
149. Porubsky, D. et al. Gaps and complex structurally variant loci in phased genome assemblies. *Genome Res.* **33**, 496–510 (2023).
150. Tan, K.-T., Slevin, M. K., Meyerson, M. & Li, H. Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres. *Genome Biol.* **23**, 180 (2022).
151. Sun, H. et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **54**, 342–348 (2022).

152. Bao, Z. et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. *Mol. Plant* **15**, 1211–1226 (2022).
153. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
154. Feng, X., Cheng, H., Portik, D. & Li, H. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nat. Methods* **19**, 671–674 (2022).
155. Feng, X. & Li, H. Towards complete representation of bacterial contents in metagenomic samples. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.00098 (2022).
156. Song, B., Buckler, E. S. & Stitzer, M. C. New whole-genome alignment tools are needed for tapping into plant diversity. *Trends Plant Sci.* **29**, 355–369 (2024).
157. Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O. & Thompson, J. D. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genom.* **21**, 293 (2020).
158. Gabriel, L. et al. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.10.544449 (2023).

Acknowledgements

We are grateful to H. Cheng and C. Zhou for their helpful comments on the manuscript.

Author contributions

Both authors contributed to all aspects of the Review.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Genetics* thanks Zechen Chong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.