Cell Press
REVIEWS

Feature Review

Tools and Strategies for Long-Read Sequencing and *De Novo* Assembly of Plant Genomes

Hyungtaek Jung,[1,*] Christopher Winefield,[2] Aureliano Bombarely,[3,4] Peter Prentis,[5]
and Peter Waterhouse[1,6,*]

The commercial release of third-generation sequencing technologies (TGSTs), giving long and ultra-long sequencing reads, has stimulated the development of new tools for assembling highly contiguous genome sequences with unprecedented accuracy across complex repeat regions. We survey here a wide range of emerging sequencing platforms and analytical tools for *de novo* assembly, provide background information for each of their steps, and discuss the spectrum of available options. Our decision tree recommends workflows for the generation of a high-quality genome assembly when used in combination with the specific needs and resources of a project.

Highlights

Tumbling sequencing costs, improvements in bioinformatic pipelines, and increased access to high-performance computing capabilities have resulted in a perfect storm where nonspecialist genomics research groups are able to access, deploy, and generate *de novo* genome sequences in nonmodel plant systems.

However, generating a high-quality assembly for many plant species still presents significant challenges owing to genome size, complexity, and experimental and computational design.

Selecting the most appropriate sequencing and software platforms for a new genome project can be confusing and daunting because of the wide spectrum of available options and the performance quality of specific tools in different contexts.

Challenges and Progress with Plant Genomics

A genome assembly is simply the sequence produced after all of the chromosomes of a target species have been fragmented (a large number of short/long DNA sequences), sequenced, and computationally put back together again to create a representation of the original intact chromosome sequences. *De novo* genome assembly assumes no prior knowledge of the source DNA sequence length, layout, or composition. The usual aim of a genome assembly is to build a highly accurate contiguous (i.e., an uninterrupted stretch of overlapping DNA) consensus sequence representing a haploid-phase version of the genome (one for each parental haplotype) of the target species. The costs of acquiring sufficient sequence data for such an assembly have now dropped to a level that most laboratories can afford. This has led to the recent explosion of plant species being sequenced. Four questions must be considered when embarking on a new genome assembly project are: (i) how big is the genome?; (ii) is it a diploid, polyploid, and/or highly heterozygous hybrid species?; (iii) how much repetitive sequence is likely to be present in the genome; and (iv) what is the best experimental and computational design to be employed?

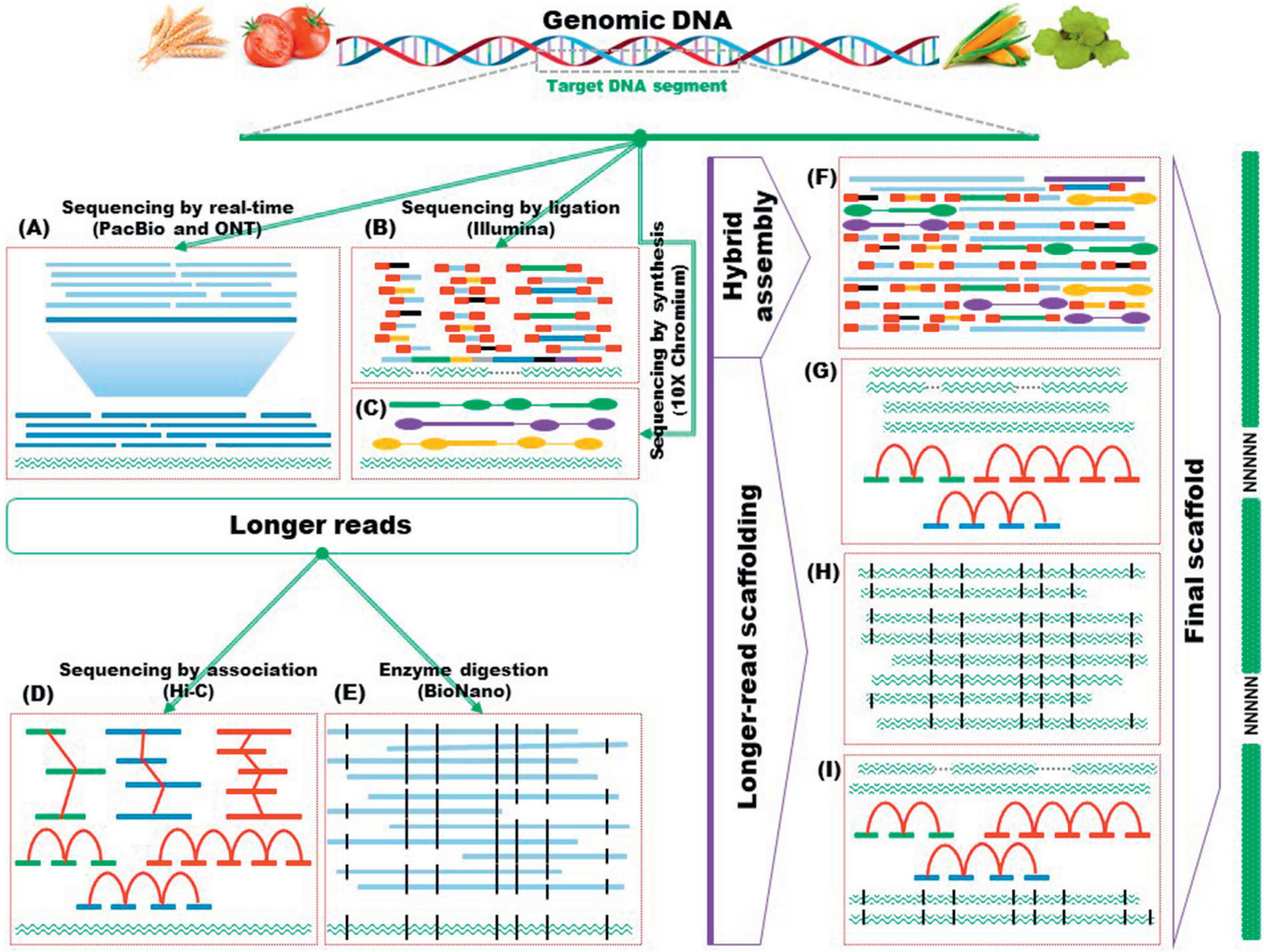
Most large plant genomes have high levels of repeated and duplicated sequences owing to whole-genome, chromosomal, subchromosomal, or tandem duplications (e.g., transposable element activity) [1,2]. With genome assemblies based on short-read (75–700 bp) data, the repeats and duplications are often not well resolved, leading to the bioinformatic formation of **chimeric sequences** (see Glossary) and fragmented **contigs**. Third-generation sequencing platforms (Pacific Biosciences, PacBio and Oxford Nanopore Technologies, ONT), that generate individual read-lengths from 8 kb to 40 kb (maximum >150 kb for PacBio and >2 Mb for ONT) [3], give much better resolution and contiguity. Nevertheless, some regions of a genome, such as the telomeric and centromeric regions of chromosomes, are often poorly resolved because they can contain megabases of repeated sequences. Current bioinformatic software does not cope well with these difficult regions, especially in the complex and polyploid genomes of many

[1]Centre for Tropical Crops and Biocommodities, Queensland University of Technology, Brisbane, QLD 4001, Australia
[2]Department of Wine, Food, and Molecular Biosciences, Lincoln University, 7647 Christchurch, New Zealand
[3]Department of Bioscience, University of Milan, Milan 20133, Italy
[4]School of Plants and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061, USA
[5]School of Earth, Environmental, and Biological Sciences, Queensland University of Technology, Brisbane, QLD, 4001, Australia
[6]School of Biological Sciences, University of Sydney, Sydney, NSW 2006, Australia

https://doi.org/10.1016/j.tplants.2019.05.003

crop species. Indeed, for this reason many reportedly 'complete' plant genome sequence assemblies have many gaps, collapsed regions, and unassigned sequences. New bioinformatic and sequencing strategies are continually being developed to overcome these problems, but none has yet been universally successful. We therefore compare the recently developed tools and give particular emphasis to their performances on a wide range of plant genomes.

*Correspondence:
h7.jung@qut.edu.au (H. Jung) and
peter.waterhouse@qut.edu.au
(P. Waterhouse).



Trends in Plant Science

Figure 1. Comparison of Different Genomic Technologies in Reconstructing Target DNA Segments. A total of eight assembly strategies are simplified and displayed from five major genomic technologies, namely short-read sequencing, long-read sequencing, **synthetic long-read (SLR)**, linked long-read (LLR), and optical mapping. (A) Long-read sequencing and assembly (PacBio and ONT). The middle blue line indicates the longest seed read used for mapping smaller reads. (B) Short-read sequencing and assembly including single-end (SE), paired-end (PE), and mate-paired (MP) reads (Illumina). The red small filled boxes indicate adaptors. The broken lines in the bottom green patterns with zigzags represent gaps in assembled contigs/scaffolds. (C) SLR and/or LLR sequencing and assembly (10X Genomics Chromium, 10xGC). (D) SLR and/or LLR sequencing and assembly (CHiCAGO and Hi-C, an extension of chromosome conformation capture, 3C). The red lines/curves indicate the LLRs that are reconstituted into chromatin via proximity ligation (Hi-C). (E) Optical mapping and assembly (BioNano). The vertical black lines indicate the enzymatic cutting and/or aligning sites. Note that BioNano is not a sequencing technology but an optical mapping technology. (F) Hybrid assembly from raw reads of A, B, and C. (G) Hybrid longer-read scaffolding and assembly from assembled contigs/scaffolds of A, B, and C (as an input), and raw reads of D. (H) Hybrid longer-read scaffolding and assembly from assembled contigs/ scaffolds of A, B, and C (as an input), and raw mapping data of E. (I) Hybrid longer-read scaffolding and assembly from assembled contigs/scaffolds of A, B, and C (as an input), raw reads of D, and mapping data of E. The bottom green patterns (in A–E) with zigzags represent assembled contigs/scaffolds for each approach. The rightmost green patterns with dots and Ns represent the final assembled scaffolds. The approaches described in A–E can be performed not only for *de novo* assembly independently but also for hybrid assembly/scaffolding approaches if merged together. Further assembly strategies are given in Figure 2.

Old and New Sequencing Technologies for Plant Genomes

The reference genome sequence of *Arabidopsis thaliana* has been invaluable to the plant science community, but it took an international effort over nearly a decade to produce the first draft and at a cost of ~100 million USD [4]. Since this initial release, generated using Sanger sequencing technology (considered to be first-generation sequencing technology [5–11]), there have been 10 major updates and the publication of a further 1135 *Arabidopsis* genomes [5]. The success of this and other model plant genome sequencing projects has been a major catalyst and inspiration for research, including the recently announced 10 000 Plant Genome Sequencing Project (10KP) [12] which will focus on nonmodel plants [13,14]. The rapid adoption of whole-genome sequencing has been facilitated by the development of **second-** and **third-generation sequencing technologies** (**SGST** and **TGST**, respectively) which have dramatically reduced sequencing costs and simplified genome assembly. Without doubt, these major new initiatives with new sequencing technology will improve our understanding of plant genomic diversity, while also acting as an important community resource for plant scientists to perform a wide range of analyses. However, to make it possible to undertake genome assembly for nonmodel plant species, the challenges will still include (i) assembling large complex genomes derived from complex whole-genome duplications, (ii) choosing the most appropriate sequencing platforms, and (iii) developing high-throughput assembly and annotation pipelines that require minimal human input.

SGSTs (including Illumina, 454, SOLiD, and Ion Torrent) are high-throughput, fast, low-cost, and highly accurate, producing reads of short length (75–700 bp). However, their limited ability to resolve complex regions with repetitive or heterozygous sequences has led to incomplete or heavily fragmented genome assemblies. This is due, in particular, to difficulties in mapping this type of data to unique positions in reference genomes and in resolving repetitive regions such as long **structural variants (SVs)**. Even after assembly, **scaffolds** will often contain many regions of unknown sequence (Figure 1 and Table 1). The TGST platforms from PacBio and ONT give long single-molecule reads (averaging >12 kb, with some ONT sequences reaching over 2 Mb [3]) with complete contiguity, facilitating assembly. However, both long-read technologies suffer from high costs per base and high error rates (Figure 1 and Table 1). Although earlier sequencing technologies and their associated assembly and mapping algorithms/software have been extensively reviewed [6–11], there are currently few comparisons or reviews of TGSTs [47–49].

The simplest TGST-based whole-genome assembly approach is undertaken in three steps. First, and most importantly for these methods, extraction of high molecular weight DNA that is free of contaminants. There are many metrics to determine the quality of DNA, the most important of these are summarized below in the section on DNA extraction methods and quality measurement.

The second step requires the preparation of platform-specific libraries using kits provided by the manufacturers. Attention should be paid to the desired insert lengths in the prepared libraries because they affect the read lengths and throughput (total number of bases sequenced per run). With both platforms it is possible to obtain average read-lengths of >20 kb. However, increasing read-length often comes at the expense of throughput. We would generally recommend a blend of sequencing runs delivering smaller read-lengths with optimized throughputs followed by runs specifically aimed at long read-lengths (>50 kb) to assist scaffolding shorter reads into larger contiguous sequences.

The third step is assembly of called and quality-filtered data using overlapping sequences to generate contiguous chromosome-length sequences. When completed genomes of closely related species are available, a reference-guided/assisted genome assembly may also be an attractive option because of the lower requirement for coverage data and computational memory. Some

Glossary

Bacterial artificial chromosome (BAC): an engineered DNA molecule (vector) that is used to clone a target DNA sequence in bacterial cells.

de Bruijn graph (DBG): an efficient way to represent a sequence in terms of its K-mer components that is widely used for short-read assemblies.

Chimeric sequence: a form of sequence consisting of two or more biological sequences and/or unrelated DNA fragments that have been artificially joined together.

Contig: a continuous stretch of assembled sequence without gaps.

Contiguity: a series of contiguous sequence (contigs) that are in contact or in proximity from a set of overlapping DNA segments that together represent a consensus region of DNA.

Fourth-generation sequencing technology (FGST): a new single-cell sequencing technique that preserves the spatial coordinates of RNA and DNA sequences with potentially subcellular resolution, thus enabling mapping of sequencing reads back to the original histological context.

Error correction: the process of removing and correcting the underlying errors generated by high-throughput sequencing platforms and/or by true genetic variation and technical artefacts to increase read and sequence quality.

Gap filling: the process of reconstructing the missing and/or unknown sequences (gaps) between consecutive contigs by mapping actual sequence reads and/or introducing uncharacterized nucleotide (N) stretches of unknown or estimated lengths.

Linked long-read (LLR): a type of data that utilizes molecular barcodes to tag short reads together that come from the same long DNA fragment in 10X Genomics Chromium (10xGC).

Methylation: an epigenetic mechanism that occurs via addition of a methyl (CH_3) group to a DNA molecule, thereby often modifying the function and expression of the genes without changing the sequence.

Overlap–layout–consensus (OLC): a graph assembly algorithm for long-reads relying on three consecutive steps: (i) Overlap (build the overlap graph to find potentially overlapping reads), (ii) layout (merge reads into contigs and simplify the graph), and (iii) consensus (derive the DNA sequence and correct read errors).

caution should be exercised, however, because the resulting assemblies may contain biases toward errors and chromosomal rearrangements in the existing reference genome [50–53]. Further practical strategies and applications for reference-guided/assisted genome assembly are discussed elsewhere [50–53]. Although prokaryotic genomes have been successfully assembled with the sole use of TGST [54], this approach has been only moderately successful for plant species, mainly for those with small and less-complex diploid genomes (<300 Mb) [55,56]. For larger plant genomes, *de novo* assembly using this approach has generally delivered less than desirable results. This is due in large part to errors in the sequencing data deriving from inaccurate base calling. These errors present significant challenges to the current sequence assembly software in generating gap-closing alignments, particularly across repeat-containing regions. Some of these issues can be resolved with increased coverage. However, there appears to be an upper limit to useful read-depth because of the systemic nature of the errors in both ONT and PacBio data. This combination leaves substantial fractions of large plant genome assemblies inaccessible and, like assemblies produced by SGSTs, limits the ability to mine for important biological insights [57–59].

The regions of large plant genomes that are most challenging to accurately determine are long tracts of repeat sequence that may span >1 Mb. Even the longest read-lengths reported by either PacBio or ONT technologies will often fail to span such regions. To assemble these tracts of sequence, the development of additional assembly strategies and sequencing technologies is required. As an interim solution, the development of an advanced 'hybrid' approach, for example, incorporating 10X Genomics Chromium (10xGC) data or medium-size single-molecule DNA fragment selection and tagging before sequencing with short-read sequencing, could be a viable option to increase the **continuity** and accuracy of long reads (see Hybrid Assembly Approaches, below). Although this 'hybrid' approach increases the accuracy of long reads by mapping Illumina short reads onto them to generate a consensus sequence, and has resulted in assembled scaffolds with high accuracy, incomplete and/or unfinished assemblies still occur (e.g., gaps and fragments). Thus, additional techniques such as optical mapping (BioNano) and chromatin association (Hi-C: an extension of chromosome conformation capture, 3C) are usually required to facilitate contig joining [11,59–62] and the completion of a genome assembly. These subchromosome scaffolding assembly (SCSA) techniques often reduce the scaffold number and increase scaffold size by a factor of three–ten to give chromosome-level assemblies (Table 2).

DNA Extraction Methods and Quality Measurement

Given the potential breadth of plant species that are likely to be targeted for genomic studies, each with their own peculiarities, we are only able to provide general suggestions on extraction methods, based on our own experience. Although recent publications provide valuable guidance [80–83], users should look to develop or adapt DNA extraction methodologies along the lines we provide, paying particular attention to the quality metrics outlined below.

Aside from the obvious requirements to generate DNA preparations that are free of contaminants such as proteins, carbohydrates, and polyphenolics, users should also seek to select methods that produce high molecular weight DNA. Avoidance of column-based DNA extraction methods is recommended given the propensity of these methods to shear DNA, often to fragment sizes <8 kb. Although we have had some success with commercial magnetic bead-based DNA purification methods for plants, these methods still shear DNA. However, with care DNA prepared in this way can deliver DNA with an average size of >30 kb. In general, the most successful methods tend to be those based on cetyltrimethyl ammonium bromide (CTAB) extraction buffers combined with spooling of DNA. These approaches produce excellent quality DNA of high molecular weight, but often require larger input of tissue than the magnetic bead-based kits. Whichever approach is adopted, there will be a requirement for refinement of the method to achieve several

Polishing: improving the consensus accuracy of an assembly and/or obtaining higher sequence identity using short and/or long reads.

Scaffolds: created by joining contigs together using additional information (introducing arbitrary N letters) about the relative position and orientation of the contigs in the genome.

Second-generation sequencing technology (SGST): sequencing techniques and platforms generating short reads (<1 kb) using wash-and-scan approaches (Roche, Illumina, and Ion-Torrent).

Structural variants (SVs): large DNA alterations (generally >1 kb), often comprising inversions, balanced translocations, and copy-number variants.

Synthetic long-read (SLR): an advanced highly parallel library preparation technique to pool barcoded subsets of the genome (~20 kb) for empowering assembly and resolving highly repetitive complexes in short Illumina reads (e.g., TruSeq).

Third-generation sequencing technology (TGST): sequencing techniques and platforms that generate long reads (>10 kb) and ultra-long reads (>1 Mb) (PacBio, ONT, and BioNano).

Table 1. Summary of Selected Long-Read Sequencers for *De Novo* Assemblies of Large Eukaryotic Genomes[a,b]

Pros and Cons	10X Genomics Chromium[c] (HiSeq 4000)	Pacific Biosciences (SEQUEL/Cell)	Oxford Nanopore (MinION)	BioNano (Saphyr/Chip)	Dovetail[c] (HiSeq 4000)
Compatible platforms	Illumina	RS II	GridION[d] and PromethION[d]	Irys	Illumina
Minimum input	~3 ng	~20 μg	~1 μg	~200 ng	~5 μg
Long-read	Synthetic	True	True	True	Synthetic
Average/maximum read length	~300 bp (PE)/ ~150 kb (LLR)	~12 kb/~150 kb	~12 kb/~2 Mb	~350 kb/~1 Mb	~150 kb/~1 Mb (SLR)
Throughput	~1500 Gb	0.7 Gb–20 Gb (SEQUEL)	50 Gb–15 Tb (PromethION)	~640 Gb	~1500 Gb
Reads	~5 Billion (B)	0.07 million (M)–2 M	1.5–5 M	~2 M (image file)	~5 B
Runtime	~3 Days	6–10 h	2 h to 6 days	~1 day	~3 days
Quality scores	>30	>10	>10	NA (only nonsequence based method)	>30
Error profile	<1% (GC/AT biased and substitutions)	5–10% (indels)	5–15% (indels and substitutions)	Sizing error, false sites, and missing sites	<1% (GC/AT-biased and substitutions)
Output format	Fasta Fastq	Bam Fasta Fastq Hdf5 (RS II)	Fast5	BNX C/S/XMAP SVMerge TIFF	Fasta Fastq
General assembly software	Supernova	CANU Falcon/Falcon-Unzip Flye HGAP Minimap/Miniasm	CANU Minimap/Miniasm TULIP	RefAligner	3D-DNA HiRise LACHESIS Meraculous SALSA
Instrument cost[e]	$$$$$	$$$$$$$$$$	$$$	$$$$$$$$$$	$$$$ (different library preparations but can be used in HiSeq 2500 or above)
Cost per Gb[e]	$$	$$$$$$$$$$	$$$$$$$$	$$$$$$	$$$$
General applications	Limited testing only for human and diploid assembly	Widely tested from prokaryotic to eukaryotic organisms Can analyze DNA **methylation**	Mainly tested for prokaryotic but starting to expand to eukaryotic organisms Can analyze DNA methylation	Widely tested from prokaryotic to eukaryotic organisms Mainly applied for scaffolding improvement and chromosome-scale assemblies	Widely tested from prokaryotic to eukaryotic organisms Mainly applied for scaffolding improvement and chromosome-scale assemblies
Other pros	Moderate cost instrument and runs Low cost per Mb with high accuracy Minimal input requirement	Numerous dedicated software tools Well-established platform (SMRT Link)	Low-to-moderate cost instrument and runs Moderate cost per Mb with easy sample preparation Real-time analysis for rapid and efficient workflows (MinKNOW) Can repetitively sequence a given	No risk of PCR artefacts Real-time data monitor for quality metrics (BioNano Access/IrisView) Can create the most contiguous and accurate assemblies possible Can provide physical genome mapping	Simple assay process No separate instrument needed
Other cons	Vulnerable to	Expensive	Lower base-calling	Expensive instrument	Vulnerable to Illumina baises

Table 1. (continued)

Pros and Cons	10X Genomics Chromium[c] (HiSeq 4000)	Pacific Biosciences (SEQUEL/Cell)	Oxford Nanopore (MinilON)	BioNano (Saphyr/Chip)	Dovetail[c] (HiSeq 4000)
	Illumina biases and limitations	instrument and runs	accuracy	and runs	and limitations
	Not true long-read	Higher cost per Mb with high random error rates	Limited testing and performance for higher eukaryotic and polyploid genomes	Moderate cost per Mb with sizing errors	Not true long-read
	Limited data available	High input requirement		Limited compatible software	High input requirement
	Limited test for non-human and polyploid assembly				Limited data available
					Mainly commercial-based service (Hi-C//HiRise)

[a] This table was generated after visiting the official website of each platform and the most recent review articles [3,6,7,9]. The same acronyms (i.e., programs) are used in all Tables. For more library preparation and sequencing guides refer to the products and/or services page of the vendor.

[b] Abbreviations: LLR, linked long read; NA, not available; PE, paired-end; SLR, synthetic long read.

[c] 10X Genomics Chromium and Dovetail: focused on HiSeq 4000 platform. Although both Dovetail and Phase Genomics provide Hi-C data, we have focused on Dovetail Genomics only.

[d] GridION and PromethION are still in experimental phase (not fully accessible for commercial service).

[e] For instrument cost and cost per Gb the relative cost is indicated by the number of $ symbols.

important quality metrics that we have found to be important for both PacBio- and/or ONT-based sequencing platforms.

Generally, purified DNA should be measured/quantified using both spectrophotometric and fluorescence-based methods (such as Qubit). Optical density (OD_{260}:OD_{280}) ratios of 1.8–2.0 indicate that samples are generally free of protein contamination, and OD_{260}:OD_{230} ratios of >2.0 generally indicate the sample is free of phenolics and carbohydrates. Quantification of genomic DNA using only spectrophotometric methods is not recommended, and quantification is best performed using fluorimetric methods such as Qubit^{TM}. Achievement of a 1:1 ratio of the concentrations of DNA determined by spectrophotometry and fluorimetry, respectively, has proved to be a very good indicator of whether DNA will be sequenced efficiently.

To determine the integrity of the DNA sample, it is strongly recommended that a sample of DNA is separated to determine the degree of degradation and the spread of molecular weight of the isolated DNA. Standard agarose gel electrophoresis is not generally recommended owing to the poor resolution of DNA above 10 kb. Contour-clamped homogeneous electric field (CHEF) or pulsed-field electrophoresis is suitable but we would recommend the use of instruments such as the TapeStation or Fragment Analyzer (Agilent Technologies) in conjunction with their high molecular weight analysis kits. Analysis of isolated DNA in this manner will assist in decisions about shearing the DNA to obtain an optimal size range for sequencing and can also be useful by assisting the identification of contaminants that may affect sequencing performance because common contaminants will often influence the mobility of DNA.

Workflow Design

The genome size, levels of ploidy and heterozygosity, and the quality of DNA that can be extracted will affect the complexity, overall quality, and cost of the genome assembly of the target species. Flow cytometry (an accurate way to determine genome size) and K-mer frequency

Table 2. Summary of Recently Published Plant *De Novo* Genome Assemblies Using Long-Read Sequences[a,b]

Scientific name	GS (Gb)	Final output			Input details and depth (×)							BND (×)	Dovetail (×)		Refs
		AGS (Gb)	TSN	N50 (Mb)	SG	454	IM	PB	10xGC	BAC	FSM		CHiCAGO	Hi-C	
Aegilops tauschii spp. *strangulata*	4.3/DP	4.22	109 861	31.73			191	35		53		90			[63]
Amaranthus hypochondriacus	0.47/DP	0.41	908	24.4			229	31		10		363		41	[64]
Arabis alpina	0.37/DP	0.34	817	4				86				722		85	[56]
Chenopodium quinoa	1.4/AT	1.39	3486	3.6			66	54				72	52		[65]
Conringia planisiliqua	0.23/DP	0.18	67	7.4				54				410			[56]
Durio zibethinus	0.74/DP	0.71	677	22.7			202	153					380	4371	[66]
Euclidium syriacum	0.26/DP	0.23	80	18.7				47				446			[56]
Hevea brasiliensis	1.4/DP	1.26	47 154	0.1		5	57	40					44		[67]
Hordeum vulgare L.	5.3/DP	4.79	4235	1.9		24	200			200		60		96	[59]
Lactuca sativa	2.5/DP	2.38	11 474	1.8			73						72		[68]
Malus domestica Borkh	0.65/DP	0.65	1081	5.6			200	35				600			[69]
Manihot esculenta	0.77/DP	0.58	2019	28.1	1	29	968				1082		125		[70]
Musa acuminata	0.53/HP	0.45	1532	3		21	91			0.2	3	60			[71]
Nicotiana attenuata	2.5/DP	2.37	37 194	0.5		5	30	10				50			[72]
Nicotiana tabacum	4.5/AT	3.69	2217	2.2		18	69			8		110			[73]
Oropetium thomaeum	0.25/DP	0.25	46	7.8				72				200			[55]
Oryza sativa Indica	0.4/DP	0.41	225	2.5			100	118			16	250			[74]
Rosa chinensis	0.56/DP	0.52	55	69.6			147	80						112	[75]
Saccharum spontaneum L.	3.36/AT	3.13	76 028	0.1			90	78		80				90	[76]
Tartary buckwheat	0.54/DP	0.45	114	7.5			175	31			25	220		195	[77]
Triticum aestivum L.	16/HxP	14.5	138 665	7.0		73	217			570		47		8	[78]
Triticum turgidum ssp. *dicoccoides*	12/AT	10.49	151 912	6.96			176							180	[57]
Triticum urartu	4.94/DP	4.86	31 559	3.67			21	19	11	297		83			[79]
Zea mays	2.3/DP	2.07	625	9.6			1000	65		1	1	60			[58]

[a] This table represents a selection of recent plant and crop genome work focusing on whole-genome assemblies using BioNano and/or Dovetail data (at least one technology used). In addition, the table does not include any individual chromosome assemblies, green alga genomes, pure TGST/SGST/hybrid genome assemblies without BioNano/Dovetail data, single-cell sequencing, or transcriptomes. If there was no estimated input depth from the original report, this was estimated from the raw data. For the most recent global statistics, we highly recommend visiting the associated GenBank BioProject.

[b] Abbreviations: 454, Roche 454; 10xGC, 10X Genomics Chromium; AGS, assembled genome size; AT, allotetraploid; BAC, bacterial artificial chromosome (including BAC-end sequence); BND, BioNano Depth; DP, diploid; FSM, fosmid; GS, genome size; HP, haploid; HxP, hexaploid; IM, Illumina [combined paired-end (PE) and mate-pair (MP) reads]; PB, PacBio; SG, Sanger; TSN, total scaffold number.

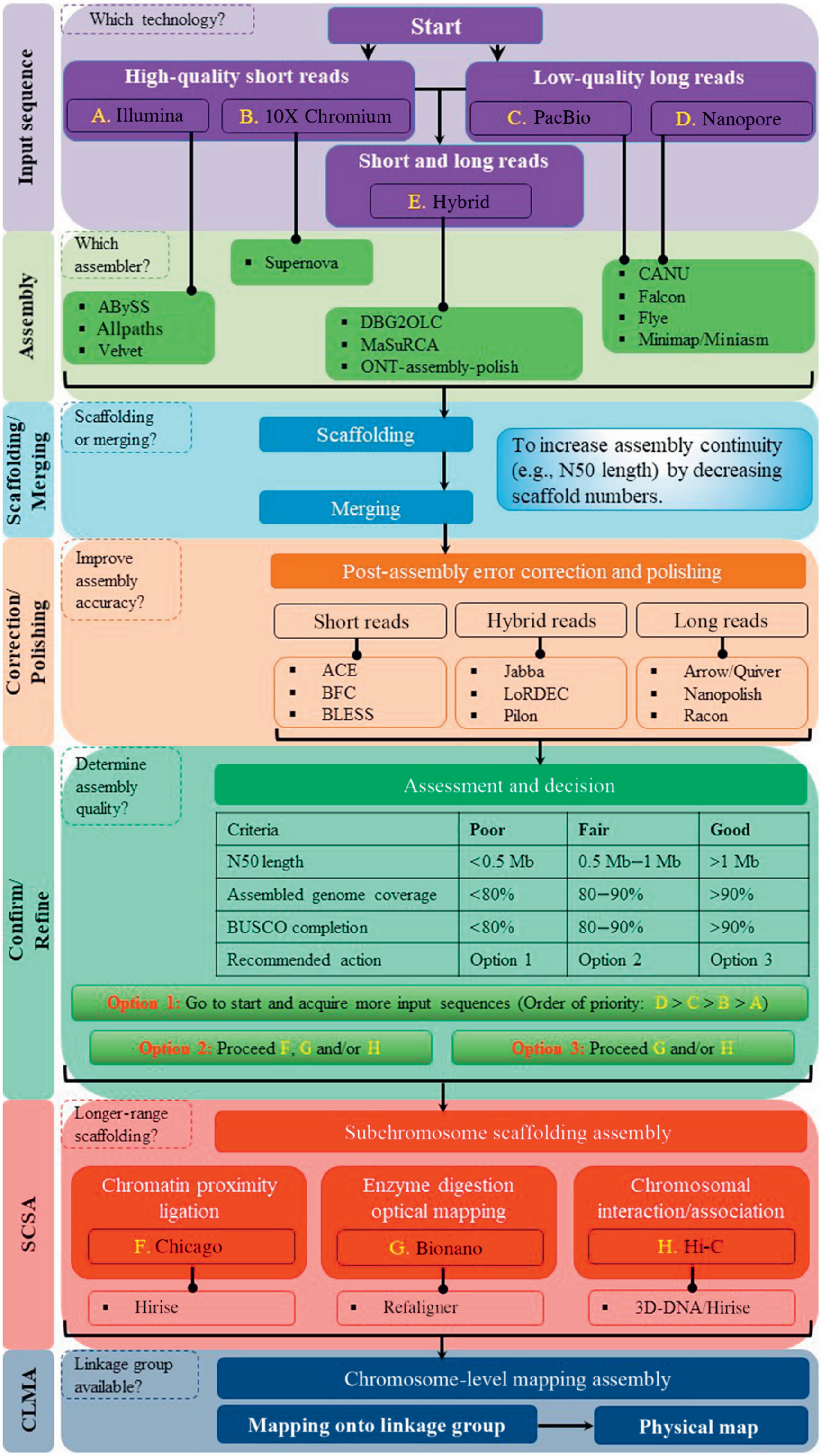
distribution (a simple approach to infer genome size, repeat content, and heterozygosity using Illumina reads) are two widely used methods to estimate the size of a genome [84], and the generation of sufficient sequence data/read coverage is a crucial starting point in a genome assembly project. If cost is not an obstacle, securing >100× coverage of long-read data can be the basis for a good genome assembly through self-correcting algorithms [e.g., in Canu, hierarchical genome assembly process (HGAP), and Falcon] that align the reads against one another without relying on any additional sequencing data. However, the cost of obtaining such high read-coverage of long-read data may not be the only problem. There are some inherent errors in the technologies. For example, both ONT and PacBio platforms struggle with homopolymeric sequences.

A hybrid approach, using a mixture of both short and long reads, can be less expensive than using long reads alone, but in general the quality of the assembly is lower [33,84–88], and several factors (e.g., genome size and the frequency of repetitive sequences) can affect its sequence contiguity. A minimum of 60× (180 Gb) long-read sequence coverage should be sufficient for a highly inbred/homozygous, small- to medium-sized (<3 Gb) diploid genome. For larger diploid genomes and/or genomes that are highly heterozygous, we recommend a minimum 60× of SGST, 30× of TGST, 50× of 10xGC, and 60× of SCSA per each haploid subgenome. Polyploid and highly repetitive genomes may require an extra 50–100% more sequence data than their diploid counterparts (Figure 2). In plants, further filtering of unwanted organelle fragments (e.g., chloroplast and mitochondrial sequences) may be necessary to increase the quality of a nuclear genome assembly. Input data usually consists of 5–20% of unwanted organelle DNA reads [89]. Thus, the apparent 60× (180 Gb) coverage of a 3 Gb plant genome may actually contain only 48× relevant coverage (144 Gb). Once high-quality sequence data (and where required SCSA) have been obtained, there remains the considerable computational task of assembling them into the best reference sequence, and this requires significant computational resources. We highlight below the typical resources available in many core facilities. However, with increasing democratization of whole-genome sequencing, more and more groups will require access to such high-performance computational resources, and the increasing availability of cloud-based solutions may offer an attractive option to many researchers (discussed below).

Each sequencing platform has different inputs (DNA, labor, and preparation), computational requirements, and costs (Table 1 and supplemental information online), with each assembly using multiple software packages and pipelines (Table 2). This article aims to provide a concise review of current and emerging TGSTs (10xGC, PacBio, ONT, BioNano, and Dovetail Genomics) and their application to *de novo* plant genome assembly. We highlight 25 analytical tools (chosen from a library of 105 readily available, open-source tools), suggest practical strategy combinations, and provide a decision tree to help researchers to select the most appropriate approach to achieve a high-quality reference genome assembly for their species of choice.

Computational Requirements

Genome assembly uses large amounts of sequence data and requires computation resources that are usually only available at high-performance computing (HPC) facilities. Given the vast range of potential plant genome assemblies that are likely to be undertaken in the near future, we can only give a general guide to the computational resources needed for such projects. However, the guide is scalable, based on genome size and ploidy, and our recommendations will likely apply larger and more complex genomes, but at a slower rate. As mentioned below, new innovations in the use of graphics processing units (GPUs) and other accelerator platforms will greatly improve analysis times. As a general guide, to successfully assemble a moderately sized diploid plant genome of ~1 Gb using software pipelines such as Canu or Falcon will require a minimum computing resource of 96 physical CPU cores, 1 TB of high-performance RAM, 3 TB of local storage, and 10 TB of shared storage. Polyploid, large (per 1 Gb genome size increase) and highly



Criteria	Poor	Fair	Good
N50 length	<0.5 Mb	0.5 Mb–1 Mb	>1 Mb
Assembled genome coverage	<80%	80–90%	>90%
BUSCO completion	<80%	80–90%	>90%
Recommended action	Option 1	Option 2	Option 3

Trends in Plant Science

(See figure legend at the bottom of the next page.)

repetitive genomes may require an additional 50% more computing resource than their 1 Gb diploid counterparts. Although increasing the size of the computing resources would generally be expected to reduce assembly times, this must be balanced against the costs of purchasing larger resources. To date, most plant genome assemblies and genomic analyses have been carried out by large and well-resourced teams with access to very large in-house systems. Decreasing sequencing costs have resulted in burgeoning numbers of users with projects requiring HPC resources [90]. As the number of users, the diversity of users, and the volumes of data grow, demands on these systems will also increase. Outside supercomputer-type facilities, continual growth in capacity (core numbers, RAM, and storage space) will be necessary to maintain the ability of existing in-house systems to deliver the required performance. Constant efforts are therefore required both to maintain and improve infrastructure and to drive more efficient use of these resources [91].

Other than large in-house HPC resources, two other options are available. The first and somewhat daunting option is to consider purchasing and/or constructing a moderately sized cluster. The design, construction, and maintenance of a cluster, of the size described above, is a complex and potentially expensive undertaking that requires significant IT support at all stages. Significant ongoing costs should also be expected for maintenance and future expansion in capacity, with particular attention being paid to long-term storage of pre- and post-processed data. A series of white papers covering possible options are available from insideHPC (https://insidehpc.com/2015/03/the-insidehpc-guide-to-hpc-in-life-sciences/).

The second option is to use cloud-based resources. Cloud approaches offer many advantages over fixed architecture, including customized and flexible environments that allow users to experiment and alter the computing environment without significant administrative overheads [92,93]. Although cloud-based solutions are generating significant interest among all HPC providers, several issues need to be understood before adopting this solution. Cluster architecture is crucial to achieve optimal performance, particularly where multiple nodes (typically single servers consisting of 24 cores, 256 Gb RAM, and 500 Gb of local disk space) are employed. Most large institutional systems have been built specifically to meet the demands of data-intensive computing, such as genome assembly and related analyses, and typically have an order of magnitude better performance than typical cloud-based options [92]. A considerable amount of work is often necessary to develop a workable cloud-based solution from scratch, in particular the development of software 'stacks' containing the requisite software pipelines [92]. In addition, slow internet connectivity speed can be a major impediment to efficient data transfer to the cloud and back. For

Figure 2. Simplified Workflow and Decision Tree for the Selection of Suitable Next-Generation Sequencing (NGS) Platforms, Reads, and *De Novo* Assemblies. The selection of a NGS platform requires a set of sequential decisions. First, decide on desired read-length and quality from sequencing technologies. Next, specify which assemblers, mappers, and polishers will be used for each dataset (differently colored boxes represent each stage and its related tools). Finally, determine the assembly quality and strategy to be used (if necessary, require more refinement). Note that several hybrid assemblers include gap-filling/scaffolding capabilities, and some can only be applied to reads from TFGST or SGST assemblies. In addition, a few SGST assemblers have been improved since their original inception to deal with both short and long reads simultaneously as per the hybrid assembly approach. It is highly recommended to visit the official website of each tool to verify the latest version/mode before use in case of possible recent changes and improvements. More applicable/alternative tools for each stage are given in Tables 3 and 4, and in the supplemental information online. Decision-making path to follow: black unbroken line, recommended workflow, and its tool for each approach; boxes of different colors represent each different stage. Scaffolding and merging: recommended but optional approaches for all short, long, and hybrid reads to increase assembly continuity. Confirm and refine: the recommended three options after assembly assessment. See Figure 1 and Table 1 for more sequencing technologies and reads (A–H). The very bottom dark-blue box represents the final *de novo* assembly outcome. Abbreviations: CLMA, chromosome-level mapping assembly using linkage group/map data; SCSA, subchromosome scaffolding assembly using (F–H); SGST, second-generation sequencing technology; TGST, third-generation sequencing technology.

users who find the development of a custom solution daunting, several commercial HPC service providers are emerging that offer solutions to meet the needs of genomics applications, but they may lack the flexibility of either a custom cloud solution or a large-scale institute-sponsored solution.

To increase the speed of processing massive amounts of sequence data there is a push toward parallelization of computer resource and software. The increased utilization of GPUs and field-programmable gate-arrays (FPGAs) offers much greater computer capacity and flexibility than CPU-based clusters [90–96]. Although currently utilization of such architecture requires considerable computing expertise, software solutions to utilize such massive parallelization are currently being developed [90,94–96].

In summary, access to HPC resources is crucial for genome assembly projects. Users at genome-focused institutions probably have access to in-house high-capacity systems with the appropriate software ‘stacks’. However, these resources may come under considerable pressure as the genomics research sector continues to grow and ask increasingly computationally intensive questions. Cloud-based computing is a possible solution not only to satisfying such increased demands but also as an avenue to empower genome researchers at institutions lacking HPC resources. Cloud computing provides flexibility, competitive pricing, and continually updated hardware and software. However, to set up suitable cloud-based software currently requires IT specialists, either from the user’s research institution or contracted from the many private providers with fit-for-purpose software and computation environments.

Assembly Approaches using Only SGST Data

Over the past decade the **de Bruijn Graph (DBG)** algorithm has been the method of choice for assembling plant and animal genomes from SGST short-read data [8,47,97]. Short-read sequence assembly approaches have been reviewed extensively (assembly approaches [8,47, 98–100], **error-correction** tools [101,102], and **mapping software** [103,104]). Although the short-read format is low-cost and has low error rates (<1%), it presents many technical and computational challenges for genome assembly [73,105,106]. However, the recently developed 10xGC system provides a workable solution. This emulsion-based method, utilizes the limiting dilution principle, and identifies the short reads generated from the same molecule, thus linking their sequences (**linked long-read**, **LLR**) and allowing more accurate assembly [107–109] (Figures 1 and 2). Two recent diploid genome assemblies, *Glycine latifolia* and *Capsicum annuum*, were assembled using a combination of SGST reads with 10xGC data, leading to a better than threefold improvement in scaffold N50 and a cost >20-fold lower than using SGST alone. However, this approach often still leaves many gaps and misassembled or unassembled regions in the final assemblies, particularly in repetitive regions and/or when assembling genomes from polyploid species [107,109]. Although this approach has greatly improved short-read assemblies in both large and complex genomes, TGST coupled to optical mapping and Hi-C techniques holds more promise for complete and contiguous assemblies, especially for polyploid species [53,68,110,111], as discussed in more detail below.

Assembly Approaches Using Only TGST Data

PacBio and ONT are becoming increasingly cost-effective for generating high-quality *de novo* plant genome assemblies (Tables 1, 2, and supplemental information online). The average read-length capability, which can easily exceed 30 kb [3,54–58,112,113], makes these data invaluable for large and complex plant genomes [29,58,63,88,111,113–116] (Table 2). Indeed, continual improvements in sequencing chemistry, throughput, and simplification of assembly algorithms make this approach the best choice for assembling large complex/polyploid genomes.

PacBio and ONT long-read sequencing methods use real-time observation of DNA sequencing. PacBio utilizes single molecule real-time (SMRT) sequencing using synthesis technology that harnesses single-molecule DNA replication using zero-mode waveguides (ZMWs) and phospholinked nucleotides [54,117]. ONT identifies DNA bases by observing the electrical currents generated as a single strand of DNA passes through a nanopore [113,115,118]. Both approaches have high random and systemic error rates (5–10% for PacBio and 5–15% for ONT), and thus require substantial depth of coverage to accurately determine consensus base calls [at least 30× for each haploid genome content; e.g., a 500 Mb tetraploid (4*n*) genome requires 120× coverage (30×, × 500 Mb, × 4*n* = 60 Gb)] [29,54,56,58,69,85–88,117–120] (Table 2). To work with this long-read data, **overlap–layout–consensus (OLC)** assemblers are best suited for *de novo* assembly [6,24,56,121,122].

De novo genome assembly using TGST generally consists of four stages; stage 1. raw read self-mapping; stage 2. error correction; stage 3. assembly of corrected reads; and stage 4. contig consensus **polishing**. Stage 3, that is considered to be the key OLC assembly stage, has another three internal stages: (i) find overlaps (suffix tree based or dynamic programming) and build an overlap graph, (ii) resolve the graph (layout), and (iii) call the sequence consensus. Stage 3 may also involve read-mapping again, but, because the error rate is much reduced at this step, it is easier and faster than stage 1 [24]. The most commonly employed *de novo* assemblers for long reads and their associated programs are summarized together with their functional features in Tables 3 and 4.

Drawing from recent reports of successful plant genome assemblies [55,56], we suggest the pipeline: PacBio and/or ONT read sequencing ► read-quality assessment, evaluation, and filtering (including removing organelle DNA reads [89]) ► assembly (single and/or multiple assemblers) ► a single consensus sequence ► error correction and polishing ► assessment ► subchromosome scaffolding assembly ► chromosome-level mapping assembly ► annotation (Figure 2). Relative to the previous short-read-derived reference genomes, recent PacBio-based *Triticum urartu* (wheat A subgenome) [79], *Zea mays* (maize B73) [58], and *Saccharum spontaneum* L. [76] assemblies have increased contig lengths by 101-, 52-, and sixfold, respectively. They also have notable improvements in the assembly of intergenic spaces and centromeres. Recent ONT-based assemblies of *Solanum pennellii* [29] and *Arabidopsis thaliana* (KBS-Mac-74 accession) [88] achieved contig N50 lengths of 2.5 Mb and 14.8 Mb that would be impossible using strategies based only on Illumina.

Several recent publications have reported excellent genome-assembly qualities from PacBio or ONT read-assemblies that have been polished/corrected with Illumina short-read data [29,88,115,116,122]. Given the substantial depth of coverage of TGST reads, they alone may be sufficient for consensus calling (self-polishing), but incorporation of Illumina paired-end (PE) and/or mate pair (MP) reads for extra rounds of polishing generally gives better consensus base accuracy [28,29,88,102,116,123,124].

Hybrid Assembly Approaches

Combining data from both TGST and SGST, in what has been termed a 'hybrid assembly', can compensate for the downsides of both approaches (i.e., high error rates in long-reads, and the propensity of short reads to generate fragmented assemblies). Using SGST data to correct errors in TGST reads has been very successful in producing contiguous and accurate *de novo* assemblies for both animal [125–127] and plant species [67,85–87,128–130].

Based on the results of the hybrid-based assemblies of apple [69], durian [66], Prince-of-Wales feather [64], quinoa [65], rice [74], and Tausch’s goatgrass [63] genomes, our suggested strategy

Table 3. Summary of *De Novo* Genome Assemblers for Long-Read Sequences[a,b,c,d]

Program	Input format	Error correction	Description	Refs
ABruijn/Flye	Fasta	Yes	A very fast OLC-based *de novo* assembler using de Bruijn graphs (DBGs) for long-read data Flye (successor of ABruijn) performs an extra repeat classification and analysis step to improve the structural accuracy of the resulting sequence including a polisher module The ABruijn algorithm comprises a series of steps: K-mer counting/selection for error correction ▶ overlapping based on the ABruijn graph ▶ preassembly ▶ generating of a rough consensus from repeating graphs longer than minimum overlap ▶ draft contigs from the unbranching paths in the graph ▶ polishing to increase the final contig quality Accessible from a stand-alone command line	[15,16]
CANU	Fasta Fastq	Yes	A fork of the CA, designed for long-read data based on OLC A hierarchical assembly pipeline that has four steps: detect overlaps in high-noise sequences using MHAP ▶ generate corrected sequence consensus ▶ trim corrected sequences ▶ assemble trimmed corrected sequences Accessible from a stand-alone command line while taking full advantage of any LSF/PBS/PBSpro/Torque/Slrum/SGE grid options	[17]
FALCON	Fasta	Yes	A set of tools for fast aligning of long reads for consensus and assembly based on OLC Specifically designed for PacBio reads to efficiently assemble haploid and diploid genomes (diploid-aware assembler) 'HGAP' comprises a series of steps: raw subreads overlapping for error correction preassembly and error correction ▶ overlapping detection of the error corrected reads ▶ overlap filtering ▶ graph construction from overlaps ▶ contig construction from graph Accessible either from SMRT link or a stand-alone command line	[18]
FALCON-Unzip	Fasta	Yes	Specifically designed modules working with FALCON for fully phased diploid assembly (representing haplotype specific contigs as 'haplotigs' as assembly output) Accessible only from a stand-alone command line with limited cluster computational environments	[18]
Finisher-SC	Fasta	Yes	A repeat-aware module working with the HGAP pipeline and MUMMER alignment to produce higher-quality assemblies that can be consistently obtained after post-processing for long-read data A series of steps: error correction (HGAP) ▶ preassembly (CA) ▶ improved assembly (FinisherSC) ▶ merging of contigs (produces longer and higher-quality contigs than existing tools while maintaining high concordance) Accessible from a stand-alone command line	[19]
HGAP	Fasta	Yes	Specifically designed for PacBio data to allow the complete and accurate shotgun assembly of a wide range of genome sizes and complexity based on OLC A succession of steps: preassembly ▶ *de novo* assembly ▶ consensus polishing HGAP4 uses FALCON for *de novo* assembly Accessible either from SMRT Link or a stand-alone command line Work-friendly with SGE grip option	[20]
HINGE	Fasta	No	A *de novo* assembler tool based on the OLC with repeat-resolution capabilities of DBG assemblers using an idea called 'hinging' for long-read data A series of steps: pairwise overlaps (DAligner) ▶ read filter (remove chimeric reads and place hinges) ▶ repeat annotation ▶ overlap (use hinging to construct graph) ▶ hinge-aided greedy assembly ▶ alignment and consensus Accessible from a stand-alone command line	[21]
MARVEL	Fasta	Yes	A largely self-contained assembler consisting of a set of tools that facilitate the overlapping, patching, correction, and assembly of noisy long-read data A series of steps: overlap ▶ patch reads (in lieu of correction) ▶ overlap (align and repeat masking) ▶ scrubbing ▶ assembly graph construction and touring ▶ optional read correction ▶ Fasta file creation Accessible from a stand-alone command line	[22]
MECAT	Fasta Fastq	Yes	An ultrafast mapping, error-correction, and *de novo* assembly tool using CA for long-read data A specific of four modules: pairwise mapping (mecat2pw) ▶ reference mapping (mecat2ref) ▶ error correction based on pairwise overlaps (mecat2cns) ▶ Canu pipeline (mecat2canu) Accessible from a stand-alone command line	[23]
Minimap/	GFA	No	A very fast OLC-based *de novo* assembler for long-read data	[24]

Table 3. (continued)

Program	Input format	Error correction	Description	Refs
miniasm			A series of steps: crude read selection ► fine read selection ► generation of a string graph ► merging of unambiguous overlaps to produce unitig sequences A fast *de novo* assembler but high consensus sequence error rate similar to raw input reads Prone to collapse repeats or segmental duplications longer than input reads (difficult to fix without error correction) Accessible from a stand-alone command line	
PoreSeq	Fasta	Yes	An assembly tool for *de novo* sequencing, consensus, and variant calling on Nanopore data A series of steps: *de novo* error correction without reference using overlap alignment ► reference error correction ► scoring known sequence variants on a given dataset ► straightforward subdivision of processing for cluster/parallel tasks Accessible from a stand-alone command line	[25]
SMART-denovo	Fasta	No	A *de novo* assembler using all-versus-all raw read alignment without error correction for long-reads A useful tool to generate accurate consensus sequences using dependent consensus polish tools A series of steps: read overlapping ► rescue missing overlaps ► identification of low-quality regions and chimeras ► produce better unitig consensus Accessible from a stand-alone command line	https://github.com/ruanjue/smartdenovo
Spectrasse-mbler	Fasta	Yes	A *de novo* assembler using all-versus-all raw read mapping for long reads A useful tool to generate high quality through a coverage-based consensus generation process A series of steps: layout computation (compute alignments with minimap) ► consensus generation ► overlap-based similarity and repeats handling ► produces a better contig consensus Accessible from a stand-alone command line	[26]
SPRAI	Fastq	Yes	Specifically designed for *de novo* assembly of PacBio reads A succession of steps: prepare data ► prepare Sprai ► correct errors and assemble ► find contigs Accessible from a stand-alone command line Work-friendly with SGE grip option	http://sprai-doc.readthedocs.io/en/latest/index.html
Trio Binning	Fasta	No	Specifically designed modules working with CANU 1.7 (TrioCANU) for fully phased diploid assembly (similar to FALCON-Unzip and Supernova) Requires moderate coverage of short (30× Illumina) and long reads (80× PacBio, 40× per haplotype) to count and subtract K-mers for both parental genomes Accessible only from a stand-alone command line	[27]
TULIP	Fasta SAM	No	A prototype tool for *de novo* assembly of Nanopore reads A succession of steps using two Perl scripts (tulipseed and tulipbulb): input data ► alignment ► configuration ► TULIP seed layout ► TULIP bundling Accessible from a stand-alone command line	[28]
WTDBG	Fasta Fastq	No	A fuzzy Bruijn graph (FBG) *de novo* assembler using all-versus-all raw read alignment for long reads A novel sequence alignment (K-mer–BIN–mapping, KBM) algorithm and a new assembly graph (FBG) for efficient assembly of large genomes A series of steps: read-mapping using KBM ► FBG assembly ► produces a better contig and unitig consensus (SMARTdenovo is a progenitor of WTDBG) Accessible from a stand-alone command line	[29]

[a] This table does not include any single-cell sequencing, transcriptome, organelle genome assemblers (mitochondrial, chloroplast, and plasmid), bacterial/metagenome assemblers (microbial and smaller genomes <10 Mb), basecalling/variant calling, SV, and methylation detection. In addition, this table does not consider any of the following measurements, namely user time, system time, CPU time, real time (wall clock time), or maximum memory usage for each assembly tool and dataset because these can differ depending on sequencing coverage and the dataset. Deprecated programs have been removed from the list (Nanocorrect and pacBioToCA).

[b] Abbreviations: CA, Celera Assembler; FALCON, fast alignment and consensus for assembly; GFA, graphical fragment assembly format; HGAP, hierarchical genome assembly process; .MECAT, mapping, error correction, and *de novo* assembly tool; MHAP, MinHash alignment process; OLC, overlap–layout–consensus; PAF, pairwise read mapping format; POA, partial order alignment; SGE, Sun Grid Engine; SPRAI, single-pass read accuracy improver; and TULIP, the uncorrected long-read integration process.

[c] Long-read data: PacBio and Oxford Nanopore Technology (ONT).

[d] Polishing tools: Quiver (ideal for RS II); Arrow (ideal for SEQUEL); Pilon (ideal for Illumina data); and Nanopolish (ideal for Nanopore data) using BLASR, BWA-MEM, and pbalign.

is: ONT/PacBio and any SGSTs (recommend to use 10xGC read sequencing) ► read-quality assessment, evaluation, and filtering ► assembly (single and/or multiple assemblers) ► a single consensus sequence ► error correction and polishing ► subchromosome scaffolding ► chromosome-level mapping assembly ► annotation (Figure 2). Incorporating ONT reads from the Promethion platform appears to generate high-quality, phased *de novo* assemblies for diploid genomes of similar quality to those incorporating PacBio reads [28,29,88,116,125,131,132], but at a lower cost. These hybrid approaches have achieved much greater sequence contiguity than Illumina-only assemblies: a three- (*Pinus taeda*) [87] and sevenfold (*Malus domestica* Borkh.) [128] improvement of contig N50 size for PacBio merged with Illumina data, and 100- to 450-fold higher contig N50s (*Brassica rapa* Z1, *Brassica oleracea* HDEM, and *Musa schizocarpa*) for ONT merged with Illumina data [133].

Hybrid assemblies also greatly benefit from incorporating 10xGC (>100 kb) and/or Hi-C (>1 Mb) data. This information facilitates the ordering and linking of the scaffolds to produce whole-chromosome pseudomolecules [57,61,62,66,111] (Table 1). The 10xGC data can also provide long-range information on a genome-wide scale, including variant calling, phasing, and extensive characterization of genomic structure, giving researchers access to low-complexity and repetitive regions that were previously missed by short-read sequencing [108]. Recent studies have highlighted the efficacy and cost-effectiveness of 10xGC linked-reads in diploid plant genome *de novo* assembly by resolving long and highly similar repetitive regions [107,109]; the utility of this technology for complex and/or polyploid plant genomes is still being investigated [53]. Using single-molecule sequencing in combination with linked-reads enables a genome sequence assembly with both high sequence and scaffold contiguity, a feat not currently achievable with either technology alone.

Error Correction and Polishing of the Consensus Sequences

To increase the accuracy and assembly of a consensus sequence, sequencing errors within TGST sequences and within assembled sequences need to be corrected. This process is termed ‘polishing’. A list of programs for error correction and polishing can be found in Tables 3 and 4, and in Tables S1–S3 in the supplemental information online. Recent work has evaluated and benchmarked multiple aligners that are used to enhance the accuracy of read-mapping [102,104,121]. Although all of the tested aligners performed well with sequence read-lengths >100 bp, some tools still showed a lack of specificity, particularly in aligning tandem repeats. According to Chu and his colleagues [121], Minimap was the most computationally efficient and sensitive method (both time and memory) on ONT datasets. However, Minimap was not as sensitive or as specific as GraphMap, DALIGNER, or MHAP on the PacBio datasets. GraphMap and DALIGNER were the most specific and sensitive methods on PacBio datasets, and DALIGNER scaled better computationally. Aligner choice is largely based on factors such as genome features, and can enhance the overall accuracy in the error-correction and polishing consensus [102,104,121].

Gap Filling and Scaffolding Assembly Approach

Post-processing approaches such as **gap filling** and scaffolding can be applied to preassembled contigs to increase N50 length and decrease the total number of contigs/scaffolds (light-blue boxes in Figure 2). It is important to note the difference between contigs and scaffolds. Scaffolding is often employed for SGST assemblies to order and join short contigs in fragmented genome assemblies. Recent high-quality genome assemblies in eukaryotes have highlighted three principal deficiencies of scaffolding [18,58,69,117]. First, in forming scaffolds it is easy to join contigs ‘across’ GC-rich and repetitive sequence regions, thereby missing important structural features in these regions of the assembly. Second, the amount of sequence in any given gap that a scaffold spans often has a poor relationship to the true gap size. This lack of

Table 4. Summary of Adapter Removal, Mapping, Error Correction, and Polishing Tools for Long-Read Sequences[a,b,c]

Functionality	Program	Input format	Description	Refs
Adapter removal	BBMap/ BBTools	Fasta Fataq	A tool for finding and removing internal PacBio adapter sequences Accessible either from SMRT Link or a stand-alone command line Need to convert bax.h5 files to fasta/fastq using bash5tools.py	http://jgi.doe.gov/data-and-tools/bbtools/
	Consensus-Tools	bax.h5	Specifically designed for PacBio reads (SMRTbell) Accessible either from SMRT Link or a stand-alone command line	https://github.com/PacificBiosciences/SMRT-Analysis/wiki/ConsensusTools-v2.3.0-Documentation
	Cutadapt	Fasta Fataq	A tool for finding and removing adapter sequences from high-throughput sequencing reads including PacBio and Nanopore Accessible from a stand-alone command line	[30]
	Porechop	Fasta Fataq	A tool for finding and removing adapters from Nanopore reads Accessible from a stand-alone command line	https://github.com/rrwick/Porechop
Mapping/ alignment	BBMap/ BBTools	Fasta Fataq	A splice-aware global aligner for high-throughput sequencing reads including PacBio and Nanopore Accessible either from SMRT Link or a stand-alone command line Need to convert bax.h5 files to fasta/fastq using bash5tools.py	http://jgi.doe.gov/data-and-tools/bbtools/
	BLASR	bas.h5 Fasta Fastq	Not a splice-aware aligner but can be used to align transcript sequences to the genome Good performance for long reads Accessible either from SMRT Link or a stand-alone command line	[31]
	BWA-MEM	Fasta Fastq	An alignment tool to support long-read data and chimeric alignment for high-throughput sequencing reads including PacBio and Nanopore Need to construct the FM-index first for the reference genome using the index command Three key alignment algorithms: aln/samse/sample for BWA-backtrack, bwasw for BWA-SW, and mem for the BWA-MEM algorithm Accessible from a stand-alone command line with limited performance for queries longer than 10 Mb	https://github.com/lh3/bwa
	COSINE	Fasta	An alignment tool utilizing a new method (K-mer size) for long-read data Accessible from a stand-alone command line	[32]
	DALIGNER	Fasta	An alignment tool (embedded as the Dazzler 'Overlap' module) to find all pairwise local alignments for long-read data Accessible from a stand-alone command line	https://github.com/thegenemyers/DALIGNER
	GMAP/ GSNAP	Fasta Fataq	An alignment tool for short (spliced transcripts) and long-reads (<1 Mb) data Accessible from a stand-alone command line	http://research-pub.gene.com/gmap/
	GraphMap	Fasta Fataq	A mapper targeted at aligning long, error-prone reads including Illumina, PacBio, and Nanopore Accessible from a stand-alone command line	[33]
	HISEA	Fasta Fataq	An efficient all-versus-all read aligner for PacBio Can be integrated into the CANU assembly pipeline Accessible from a stand-alone command line	[34]
	LAST	Fasta	An alignment tool for long-read data using adaptive seeds that copes more efficiently with repeat-rich	[35]

(continued on next page)

Table 4. (continued)

Functionality	Program	Input format	Description	Refs
			reference sequences Accessible either from a stand-alone command line and a web service with effective performance for query sequences ranging from 100 bp to 100 Mb Need to convert fastq to fasta format	
	marginAlign	Fasta Fastq	A package tool for sequence alignment and SNVs calling of Nanopore reads Accessible from a stand-alone command line	[36]
	Mash	Fasta Fastq	A tool for fast distance estimation alignment using MinHash for high-throughput sequencing reads including PacBio and Nanopore Accessible from a stand-alone command line	[37]
	MHAP	Fasta .dat	An alignment tool (locality sensitive hashing) to detect overlaps and utilities for long-read data Accessible from a stand-alone command line	[38]
	minialign/ minimap	Fasta Fastq	An alignment tool for long-reads built on three key algorithms: minimizer-based index of the minimap overlapper, array-based seed chaining, and SIMD-parallel SWG extension Accessible from a stand-alone command line	https://github.com/ocxtal/minialign
	NanoOK	Fasta Fastq	A tool for extraction, alignment, and analysis of Nanopore reads Accessible either from a stand-alone command line or Mac OS platforms	[39]
	NGMLR	Fasta Fastq	A specifically designed tool to quickly and correctly align long reads for spanning (complex) structural variations (SVs) using an SV-aware K-mer search based on a Smith–Waterman alignment algorithm Accessible from a stand-alone command line	[40] https://github.com/philres/ngmlr
	pbalign	.h5 Fasta	A specifically designed tool for PacBio reads Accessible either from SMRT Link or a stand-alone command line Need to convert bax.h5 files to bam files using bax2bam	https://github.com/PacificBiosciences/pbalign
	STAR	Fasta Fastq	An alignment tool for short- (spliced transcripts) and long-read (<50 kb) data Accessible from a stand-alone command line	https://github.com/alexdobin/STAR
Error correction	Falcon_sense	Fasta Fastq Bam	A tool for error correction using the consensus-calling algorithm in FALCON to preserve the information from heterozygous SNPs for PacBio reads Accessible either from a stand-alone command line and SMRT Link	[18]
	Frame-Pro	Fasta m5 (BLASR) Hmm (Pfam)	A profile homology search tool using HMM and DAG for PacBio reads Accessible from a stand-alone command line	[41]
	LoRMA	Fasta	An iterative alignment-free correction method for long-read data Accessible from a stand-alone command line	[42]
	LRCstats	Fasta SAM	A novel way pipeline using SimLORD or PBSim simulator to measure the accuracy of sequencing errors for long reads Accessible from a stand-alone command line	[43]
	pbdagcon	Fasta	A tool for sequence alignment and consensus using DAGCon for PacBio reads Accessible only from a stand-alone command line	https://github.com/PacificBiosciences/pbdagcon

Table 4. (continued)

Functionality	Program	Input format	Description	Refs
	Sparc	Fasta	A sparsity-based consensus algorithm for error correction of high-throughput sequencing reads including PacBio and Nanopore Accessible from a stand-alone command line	[44]
Consensus polish	Arrow	.bam .xml	A HMM model for sequence consensus and variants for PacBio (RSII and SEQUEL) reads Accessible either from a stand-alone command line (GenomicConsensus) and SMRT Link	https://github.com/PacificBiosciences/GenomicConsensus
	Nanopolish	Fasta Fastq	A package tool for consensus sequence, methylation, and SNP calling of Nanopore reads using HMM-based consensus calling Accessible from a stand-alone command line	[45]
	Quiver	.cmp.h5 .fofn .xml	A more sophisticated tool to find the maximum quasi-likelihood template sequence for PacBio (RSII) reads Accessible either from a stand-alone command line (GenomicConsensus) and SMRT Link	https://github.com/PacificBiosciences/GenomicConsensus
	Racon	Fasta Fastq MHAP PAF	A consensus module for *de novo* assembly of long-read data based on a POA graph approach A series of steps: layout ► aligning reads and segmentation (optional error correction) ► POA graph (SIMD-accelerated) ► segment splicing ► consensus sequence Works effectively with miniasm to enable consensus genomes with similar or better quality than state-of-the-art methods while being an order of magnitude faster Accessible from a stand-alone command line	[46]

[a]This table does not include any single-cell sequencing, transcriptome, organelle genome assemblers (mitochondrial, chloroplast, and plasmid), bacterial/metagenome assemblers (microbial and smaller genomes <10 Mb), basecalling/variant calling, SV, or methylation detection. In addition, this table does not consider any measurements namely: user time, system time, CPU time, real time (wall clock time), or maximum memory usage for each assembly tool and dataset because these vary greatly depending on sequencing coverage and the dataset. In the case of cross-contamination in raw data (e.g., plastids, viruses, and bacteria), removing the contaminated reads with sequence removers (Cutadapt and Porechop) before employing genome assemblers could be helpful to improve assembly speed and accuracy.

[b]Abbreviations: DAGCon, directed acyclic graph consensus; HMM, hidden Markov model; HISEA, hierarchical seed aligner, MHAP: MinHash alignment process; NGMLR, convex gap-cost alignments for long reads; RACON, rapid consensus; SIMD, single instruction multiple data; SNP, single-nucleotide polymorphism; SNV, single-nucleotide variation; SWG, Smith–Waterman–Gotoh.

[c]Long-read data: PacBio and Oxford Nanopore Technology (ONT).

concordance can affect our ability to understand the true physical distance between functional elements in genomes. Third, the sequence flanking the newly scaffolded sequence can be of low quality, which can result in misassembly owing to the deficiencies of SGST (GC-bias or read-length limitations).

Despite the potential deficiencies of scaffolding, closing gaps in draft genomes is still an important post-processing step in genome assembly. However, if closing gaps in draft genomes intends to introduce actual nucleotide sequence (rather than 'filling' with Ns), the utility of extra 10xGC, PacBio, and ONT reads can be effective to aid gap filling before the polishing and scaffolding stages. The reason for this is that SCSA, mainly from BioNano and Hi-C data, acts to improve assembly quality by correcting misassemblies and/or ordering scaffolds based on the given input (e.g., an assembly file from a previous step).

Another post-processing approach to improve genome contiguity could be to merge assemblies from multiple assemblers (Table S3 in the supplemental material online). A recent investigation conducted by Alhakami and colleagues [134] evaluated contiguity, correctness, coverage, and the duplication ratio of the merged assemblies compared with the individual assemblies as input. For the scaffolding and meta-assembling approaches, a potential strategy to consider is:

TGST, SGST, and hybrid read sequencing ► read-quality assessment, evaluation, and filtering ► assembly from multiple assemblers (multiple parameters) ► scaffolding and/or merging ► a single consensus assembled sequence ► error correction and polishing ► assessment and decision ► subchromosome scaffolding assembly ► chromosome-level mapping assembly ► annotation (Figure 2).

An earlier review [135] gives excellent guidance for whole-genome sequencing projects using tools and technologies developed before 2015; we have focused on tools released since then (Table S2 in the supplementary material online).

Subchromosome Scaffolding Assembly

All assemblies derived only from sequence reads will contain misassemblies (inversions and translocations) that are largely caused by the inability of both sequencing and assembly pipelines to cope with long tracts of repeat sequences. These issues are further compounded by high levels of heterozygosity, as well as by polyploidization, that are common in many plant species. Two methodologies, BioNano and Hi-C, can improve the assembly quality by validating the integrity of the initial assembly, correcting misorientations, and ordering the scaffolds. These methods generally improve the scaffold N50 length by at least fivefold (Figure 2 red boxes, and Table 2). However, it is important to secure the most contiguous, complete, and minimally fragmented genome assembly to feed into the SCSA approach. If the initial assembly falls short in terms of the quality metrics discussed above, further improvement by incorporating more 10xGC, PacBio, or ONT data is highly recommended.

BioNano is a nonsequence-based scaffolding method (next-generation mapping, NGM), that uses endonucleases to nick long DNA molecules at the enzyme recognition site, upon which fluorescent nucleotides are incorporated and the long strands are repaired. This results in long (>150 kb) fragments of DNA with fluorescent labels at each endonuclease nick site in that molecule. Separation and detection of the labeled fragments allows mapping of the sequence specific labeled sites within long contiguous DNA molecules, resolving misalignments within the source DNA.

Hi-C, a chromatin-association/interaction analysis method, involves formaldehyde-mediated crosslinking of cellular contents, followed by isolation and digestion of DNA, labeling DNA ends with biotin, followed by proximity ligation of these ends, recovery of DNA, library synthesis, and Illumina based pair-end sequencing. Each pair of reads represents a single chromatin contact [136]. Subsequent computational analysis of the data allows reconstruction of chromatin interactions that reveal wider sequence structures.

Both methods have been used successfully, and a comparison of the two methods is presented in the supplemental information online. Some noticeable differences between these approaches have been reported, and in general Hi-C data have been found to resolve longer segments of chromosomes compared with BioNano, allowing near-chromosome-level assembly quickly, cheaply, and accurately [59–62,75,126]. Other studies have also pointed out that Hi-C approaches combined with PE, MP, or long-read sequencing could be effective at increasing the resolution of the spatial arrangement of chromosomes through the detection and quantification of pairwise chromatin interactions across the genome [59,64,66,75,137]. In particular, if genetic maps are available, the creation of long-distance chromatin interaction maps by Hi-C data should be considered for the final assembly step to generate pseudochromosomes from the more detailed 3D genomic structures [61,62].

Assessment of Assembly Quality

Estimating assembly quality requires several statistical and biological validations. These include overall assembly size (determining the match to the estimated genome size), measures of assembly contiguity (N50, NG50, NA50, or NGA50; number of contigs; contig length; and contig mean length), assembly likelihood scores (calculated by aligning reads against each candidate assembly), and completeness of the genome assembly (BUSCO scores and/or RNA-Seq mapping) [138]. Agreement with data on quantitative trait loci (QTL), fluorescent *in situ* hybridization (FISH) experiments using **bacterial artificial chromosome (BAC)** clones, and contiguity of the genome assembly with a chromosome-level genetic map are strong indicators of quality. If an initial assembly attempt is not satisfactory, three specific areas (contiguity, accuracy, and completeness) [138] should be considered to determine the best path forward to improve the quality of the *de novo* assembly (confirm/refine in green boxes in Figure 2). To address high contig numbers with low average size it is generally best to acquire and incorporate more TGST or 10xGC (see Hybrid Assembly Approaches) reads. Attempting to increase assembly quality through additional SCSA data is unlikely to be helpful because the data are usually ineffective in assisting hybrid assemblers to span gaps between existing contigs. Addition of more and longer TGST reads is often more productive in bridging existing contigs by increasing average contig size; subsequent addition of further SCSA data will then improve read accuracy and the overall contiguity of assemblies. However, if the resulting assembly still has >1000 contigs (i.e., is still highly fragmented), increasing the amount of SCSA data alone is unlikely to result in a dramatic improvement.

Discussion and Recommendations

De novo genome assembly is a rapidly evolving area of research. The speed of innovation is being driven both through technological innovations in sequence data generation and through community efforts to improve computational approaches to assemble data quickly and cost-effectively. However, the overall quality of a genome assembly is affected by all components of the pipeline, including the quality and integrity of the input DNA, genome size, genome organization, and computational design. Paajanen and colleagues [139] have benchmarked assemblies for completeness and accuracy, as well as input DNA, computational requirements, and sequencing costs (Box 1). They focused their benchmarking on a single diploid species in which the use of hybrid scaffolding (Illumina and/or PacBio + CHiCAGO and/or BioNano) was examined. Our review extends their work by addressing other approaches (Illumina and/or 10xGC + PacBio and/or ONT) and the use of SCSA approaches such as Hi-C. In doing so, we offer an updated guide for plant genome sequencing projects focusing

Box 1. An Example of a Plant Genome Sequencing Project

A good example of a recent assembly using the approaches outlined in the main text has been the assembly of *Solanum verrucosum* (a diploid wild potato from Mexico, ~722 Mb). This was derived using SGST (Illumina) and TGST (PacBio). In the assembly, various assemblers were compared that utilized both short and long reads, singly or in combination. In general, genome assembly from short reads was inferior to that from long reads. Long-read assemblies produced by Falcon were better than Canu and HGAP3 [139]. There have been only two reports of short-read assemblies after scaffolding with a mate-paired library using Soapdenovo2 (DISCOVAR-MP was better than ABySS-MP) that were comparable to the results of long-read assemblers [139]. In addition, use of linked long-read (LLR) technology from 10xGC and its associated assembler, Supernova, gave an outcome similar to that of Falcon with TGST reads. Incorporation of further SCSA data (105× depth of CHiCAGO Library, Dovetail) or an optical map (350× depth of BioNano) with multiple assembly approaches (Discovar, Falcon, and Supernova), substantially improved the selected assembly contiguity by ~fivefold.

Assembly-quality evaluation criteria for the assemblies produced using the above approaches, such as K-mer and BUSCO content, indicated that there was little difference among the final assemblies: Discovar (0.15% and 99.97%), Falcon (0.66% and 99.87%), and Supernova (1.3% and 99.40%). Gap filling by PBJelly, using a small amount of PacBio data (8× depth), slightly increased the assembled genome size and N50 values by reducing unknown sequence (Ns).

From this work, it appears that (i) the 10xGC technology gives high-quality and accurate assembly, similarly to PacBio, but at a significantly lower cost in diploid plant genomes; and (ii) incorporation of additional longer-range reads and/or optical mapping (by BioNano) assists in resolving repetitive regions and greatly increases scaffolding contiguity.

on the value and use of TGST sequencing platforms, analytical tools, and assembly strategies, and provide a decision tree that may assist researchers contemplating genome assemblies for nonmodel plants including potential polyploid species.

Regardless of the strategy chosen by researchers, the ultimate goal of genome sequencing projects is to produce the single best assembly in a cost- and time-effective manner, and if possible to a chromosome-level assembly on which scaffolds are anchored [58,65,69]. Each approach and tool that we have examined has limitations based on compromises inherent in the different algorithms and the assumptions used. Therefore, we recommend that several tools/approaches are used at each stage (assembly, correction, polishing, and scaffolding) and that their outputs are compared to determine the best combinations of tools for the data at hand. In addition, it is imperative to optimize each of the individual program parameters within the pipeline for the given dataset to produce a best-quality assembly. It is common practice to generate multiple genome assemblies from different assemblers, parameters, and algorithms (e.g., merging phase), and then try to predict the best assembly and/or to improve the contiguity and quality of each assembly into a single superior genome sequence [134,140–142]. Even with the recent advances in sequencing technology and computational analysis, this remains the best approach.

Concluding Remarks and Future Perspectives

Despite the fact that there is no perfect plant genome assembly, many high-quality plant genome assemblies have been achieved over the past 5 years thanks to the availability of high-throughput TGST and SGST data. In particular, a combination of PacBio/ONT long reads with 10xGC, Hi-C, or BioNano data has dramatically improved diploid *de novo* assembly and SV detection within organisms [29,58,66,75,79,116]. Impressive strides have been made in the production of genome assemblies for large and complex plant genomes using these combined TGST and SCSA approaches [57,65,66,143,144]. Long-read sequencing methods are facilitating the spanning of previously problematic and impenetrable repetitive regions of genome sequence, and in doing so have provided unprecedented opportunities to resolve these regions of a genome and improve both the assembly and annotation of plant genomes [76,78,79]. Long reads can also provide contiguous RNA transcript data, offering new solutions for finding new genes and precisely identifying variant splice isoforms of genes [145,146]. However, when presented with larger/polyploid genomes with a high repeat content, it is not always easy for researchers to choose the best approach for genome assembly. This often results in trade-offs between sequencing cost, assembly approaches, and accuracy due to differences among sequencing platforms and analytical tools.

Although the immediate future is arguably focused on improving TGST long-read approaches and developing **fourth-generation sequencing technologies (FGSTs)**, these are currently still expensive options (cost per base) compared with SGST approaches. Thus, for the medium term, although read accuracy and costs are improving for TGST approaches, research into hybrid methods generating a combination of data types maximizing the positive characteristics of each (e.g., cost, quality, and read-length) may be effective in achieving more complete and accurate genomes. It is likely that adoption of a hybrid approach (10xGC + ONT/PacBio + Hi-C) will often be optimal in terms of cost and accuracy when matched with an appropriate genome assembly pipeline. Unfortunately, application of FGST in plant genomes has yet not been well reported, and it will be interesting to watch its development in the coming years.

Sequence acquisition methodologies are improving rapidly, but bioinformatic approaches to deal with polyploid or aneuploid genome assemblies are virtually absent. This is a clear area for improvement, and will be greatly facilitated by the ongoing development of hybrid approaches to

Outstanding Questions

Is there a *de novo* assembler that can combine 10xGC and TGST reads from the raw data step to reach phase-separation? Will it be available for polyploid species?

Is there any alternative algorithm beyond OLC for long-read and/or hybrid-read assembly to reduce computational time and minimize storage space?

What could be the next effective hybrid algorithm for correcting read errors (i.e., increase accuracy) as well as for expanding contiguity?

What future techniques can produce accurate (>99%) and long reads (>1 Mb) that can resolve segmental duplications and heterochromatic regions in plants?

Will it be helpful to reach chromosome-level resolution without genetic markers/maps? If yes, can any practical approach be considered to produce a fast and accurate reference genome?

Will it be possible to sequence and assemble individual chromosomes after sorting chromosomes? Could this work for a wide range of plant chromosomes?

produce phase-separated chromosomal data in polyploid systems. Some progress is already being made with pipelines [76] such as Falcon-Phase [147], Trio binning [27], and highly efficient repeat assembly [148].

The obvious goal is to develop methods that produce and join sequences into accurate, contiguous, and entire-chromosome sequences, and also at low cost. Continued advances in both sequencing and bioinformatic technology hold promise that this is not very far away. Researchers will be able to spend less time in assembling genomes and focus more on exploring the biology of genomes to gain a deeper understanding of genomic diversity, evolution, epigenomics, and gene function. No doubt, this will accelerate the process of plant breeding and the production of improved varieties in a wide range of crops [10,78,149]. We hope that the decision tree we have developed, alongside our summary of analytical tools, and leading-edge technologies, will aid and encourage researchers to expand the already impressive spectrum of high-quality plant genome resources (see Outstanding Questions).

Acknowledgments

The authors are grateful to their colleagues, collaborators, and field/technical specialists of each company for their valuable comments, especially to Matthew Hodgett, Michal Lorenc, Ana Pavasovic, and two anonymous reviewers. This project is supported by an Australian Research Council (ARC) Laureate Fellowship (LF160100155) to P.W.

Supplemental Information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tplants.2019.05.003.

References

1. Pellicer, J. *et al.* (2018) Genome size diversity and its impact on the evolution of land plants. *Genes* 9, 88
2. Wang, P. *et al.* (2018) Factors influencing gene family size variation among related species in a plant family, Solanaceae. *Genome Biol. Evol.* 10, 2596–2613
3. Payne, A. *et al.* (2018) BulkVis: a graphical viewer for Oxford Nanopore bulk FAST5 files. *Bioinformatics* Published online November 20, 2018. https://doi.org/10.1093/bioinformatics/bty841
4. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
5. The 1001 Genome Consortium (2016) 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 481–491
6. Escalona, M. *et al.* (2016) A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet.* 17, 459–469
7. Goodwin, S. *et al.* (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351
8. Chen, Q. *et al.* (2017) Recent advances in sequence assembly: principles and applications. *Brief. Funct. Genomics* 16, 361–378
9. Mardis, E.R. (2017) DNA sequencing technologies: 2006–2016. *Nat. Protoc.* 12, 213–218
10. Yuan, Y. *et al.* (2017) Improvement of genomics technologies: application to crop genomics. *Trends Biotechnol.* 35, 547–558
11. Sedlazeck, F.J. *et al.* (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* 19, 329–346
12. Cheng, S. *et al.* (2018) 10KP: a phylodiverse genome sequencing plan. *GigaScience* 7, 1–9
13. Chen, F. *et al.* (2018) The sequenced angiosperm genomes and genome databases. *Front. Plant Sci.* 9, 418
14. Liu, H. *et al.* (2019) Molecular digitization of a botanical garden: high-depth whole genome sequencing of 689 vascular plant species from the Ruili Botanical Garden. *GigaScience* Published online April 1, 2019. https://doi.org/10.1093/gigascience/giz007
15. Lin, Y. *et al.* (2016) Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl. Acad. Sci.* 113, E8396–E8405
16. Kolmogorov, M. *et al.* (2019) Assembly of long error-prone reads using repeat graphs. *Nature Biotechnol.* 37, 540–546
17. Koren, S. *et al.* (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736
18. Chin, C.-S. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054
19. Lam, K.-K. *et al.* (2015) FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. *Bioinformatics* 31, 3207–3209
20. Chin, C.S. *et al.* (2013) Nonhybrids, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569
21. Kamath, G.M. *et al.* (2017) HINGE: long-read assembly achieves optimal repeat resolution. *Genome Res.* 27, 747–756
22. Grohme, M.A. *et al.* (2018) The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms. *Nature* 554, 56–61
23. Xiao, C.L. *et al.* (2017) MECAT: an ultra-fast mapping, error correction and de novo assembly tool for single-molecule sequencing reads. *Nat. Methods* 14, 1072–1074
24. Li, H. (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 32, 2103–2110
25. Szalay, T. and Golovchenko, J.A. (2015) De novo sequencing and variant calling with nanopores using PoreSeq. *Nat. Biotechnol.* 33, 1087–1091
26. Recanati, A. *et al.* (2017) A spectral algorithm for last de novo layout of uncorrected long nanopore reads. *Bioinformatics* 33, 3188–3194
27. Koren, S. *et al.* (2018) De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* 36, 1174–1182
28. Jansen, H.J. *et al.* (2017) Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Sci. Rep.* 7, 7213

29. Schmidt, M.H.W. *et al.* (2017) De novo assembly of new *Solanum pennellii* accession using nanopore sequencing. *Plant Cell* 29, 2336–2348
30. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12
31. Chaisson, M.J. and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application an theory. *BMC Bioinformatics* 13, 238
32. Afshar, P.T. and Wong, W.H. (2017) COSINE: non-seeding method for mapping long noisy sequences. *Nucleic Acids Res.* 45, e132
33. Sović, I. *et al.* (2016) Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. *Bioinformatics* 32, 2582–2589
34. Khiste, N. and Ilie, L. (2017) HISEA: HIerachical SEed Aligner for PacBio data. *BMC Bioinformatics* 18, 564
35. Kielbasa, S.M. *et al.* (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493
36. Jain, M. *et al.* (2015) Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* 12, 351–356
37. Ondov, B.D. *et al.* (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17, 132
38. Berlin, K. *et al.* (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* 33, 623–630
39. Leggett, R.M. *et al.* (2016) NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics.* 32, 142–144
40. Sedlazeck, F.J. *et al.* (2018) Accurate detection of complex structural variations using single molecule sequencing. *Nat. Methods* 15, 461–468
41. Du, N. and Sun, Y. (2016) Improved homology search sensitivity of PacBio data by correcting frameshifts. *Bioinformatics* 32, i529–i537
42. Salmela, L. *et al.* (2017) Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* 33, 799–806
43. La, S. *et al.* (2017) LRCstats, a tool for evaluating long reads correction methods. *Bioinformatics* 33, 3652–3654
44. Ye, C. and Ma, Z. (2016) Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *Peer J.* 4, e2016
45. Loman, N.J. *et al.* (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735
46. Vaser, R. *et al.* (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746
47. Sohn, J.I. and Nam, J.W. (2018) The present and future of de novo whole-genome assembly. *Brief. Bioinform.* 19, 23–40
48. van Dijk, E.L. *et al.* (2018) The third revolution in sequencing technology. *Trends Genet.* 34, 666–681
49. Wee, Y.K. *et al.* (2019) The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. *Brief. Funct. Genomics* 18, 1–12
50. Lischer, H.E.L. and Shimizu, K.K. (2017) Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics* 18, 474
51. Garg, S. *et al.* (2018) A graph-based approach to diploid genome assembly. *Bioinformatics* 34, i105–i114
52. Kolmogorov, M. *et al.* (2018) Chromosome assembly of large and complex genomes using multiple references. *Genome Res.* 28, 1720–1732
53. Kyriakidou, M. *et al.* (2018) Current strategies of polyploidy plant genome sequence assembly. *Front. Plant Sci.* 9, 1660
54. Rhoads, A. and Au, K.F. (2015) PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* 13, 278–289
55. VanBuren, R. *et al.* (2015) Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* 527, 508–511
56. Jiao, W.B. *et al.* (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.* 27, 778–786
57. Avni, R. *et al.* (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357, 93–97
58. Jiao, Y. *et al.* (2017) Improved maize reference genome with single-molecule technologies. *Nature* 546, 524–527
59. Mascher, M. *et al.* (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433
60. Moll, K.M. *et al.* (2017) Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, *Medicago truncatula*. *BMC Genomics* 18, 578
61. Lin, D. *et al.* (2018) Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture. *Nat. Genet.* 50, 754–763
62. Wang, M. *et al.* (2018) Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nat. Plants* 4, 90–97
63. Luo, M.C. *et al.* (2017) Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* 551, 498–502
64. Lightfoot, D.J. *et al.* (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biol.* 15, 74
65. Jarvis, D.E. *et al.* (2017) The genome of *Chenopodium quinoa*. *Nature* 542, 307–312
66. Teh, B.T. *et al.* (2017) The draft genome of tropical fruit durian (*Durio zibethinus*). *Nat. Genet.* 49, 1633–1641
67. Pootakham, W. *et al.* (2017) De novo hybrid assembly of the rubber tree genome reveals evidence of paleotetraploidy in *Hevea* species. *Sci. Rep.* 7, 41457
68. Reyes-Chin-Wo, S. *et al.* (2017) Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nat. Commun.* 8, 14953
69. Daccord, N. *et al.* (2017) High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* 49, 1099–1106
70. Bredeson, J.V. *et al.* (2016) Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* 34, 562–570
71. Martin, G. *et al.* (2016) Improvement of the banana '*Musa acuminate*' reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics* 17, 243
72. Xu, S. *et al.* (2017) Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proc. Natl. Acad. Sci.* 114, 6133–6138
73. Edwards, K.D. *et al.* (2017) A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. *BMC Genomics* 18, 448
74. Du, H. *et al.* (2017) Sequencing and de novo assembly of a near complete *indica* rice genome. *Nat. Commun.* 8, 15324
75. Raymond, O. *et al.* (2018) The *Rosa* genome provides new insights into the domestication of modern roses. *Nat. Genet.* 50, 772–777
76. Zhang, J. *et al.* (2018) Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* 50, 1565–1573
77. Zhang, L. *et al.* (2017) The Tartary Buckwheat genome provides insights into rutin biosynthesis and abiotic stress tolerance. *Mol. Plant* 10, 1224–1237
78. International Wheat Genome Sequencing Consortium (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, eaar7191
79. Ling, H.Q. *et al.* (2018) Genome sequence of the progenitor of wheat A subgenome *Triticum Urartu*. *Nature* 557, 424–428
80. Mayjonade, B. *et al.* (2016) Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *BioTechniques* 61, 203–205
81. Denis, E. *et al.* (2018) Extracting high molecular genomic DNA from *Saccharomyces cerevisiae*. *Protocol Exchange* https://doi.org/10.1038/protex.2018.076
82. Workman, R. (2018) High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing.

Protoc. Exch. Published online April 27, 2018. https://www.nature.com/protocolexchange/protocols/6785

83. Schalamun, M. *et al.* (2019) Harnessing the MinION: an example of how to establish long-read sequencing in a laboratory using challenging plant tissue from *Eucalyptus pauciflora*. *Mol. Ecol. Resour.* 19, 77–89
84. Li, F.W. and Harkess, A. (2018) A guide to sequence your favourite plant genomes. *Appl. Plant Sci.* 6, e1030
85. Zimin, A.V. *et al.* (2017) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread, with the MaSuRCA mega-reads algorithm. *Genome Res.* 27, 787–792
86. Zimin, A.V. *et al.* (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* 6, 1–7
87. Zimin, A.V. *et al.* (2017) An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *Gigascience* 6, 1–4
88. Michael, T.P. *et al.* (2018) High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat. Commun.* 9, 541
89. Soorni, A. *et al.* (2017) Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC Genomics* 18, 49
90. Liu, W. *et al.* (2017) Computing platforms for big biological data analytics: perspectives and challenges. *Comput. Struct. Biotechnol. J.* 15, 403–411
91. Dahlö, M. *et al.* (2018) Tracking the NGS revolution: managing life science research on shared high-performance computing clusters. *Gigascience* 7, 1–11
92. Yelick, K. *et al.* (2011) *The Magellan Report on Cloud Computing for Science*. Office of Advanced Scientific Computing Research
93. Langmead, B. and Nellore, A. (2018) Cloud computing for genomic data analysis and collaboration. *Nat. Rev. Genet.* 19, 208–219
94. Ocaña, K. and De Oliveira, D. (2015) Parallel computing in genomic research: advances and applications. *Adv. Appl. Bioinforma. Chem.* 8, 23–35
95. Kawalia, A. *et al.* (2015) Leveraging the power of high performance computing for next generation sequencing data analysis: tricks and twists from a high throughput exome workflow. *PLoS One* 10, 1–16
96. Kulkarni, P. and Frommolt, P. (2017) Challenges in the setup of large-scale next-generation sequencing analysis workflows. *Comput. Struct. Biotechnol. J.* 15, 471–477
97. Compeau, P.E. *et al.* (2011) How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29, 987–991
98. Kajitani, R. *et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24, 1384–1395
99. Liu, B. *et al.* (2016) BASE: a practical de novo assembler for large genomes using long NGS reads. *BMC Genomics* 17, 499
100. Pryszcz, L.P. and Gabaldón, T. (2016) Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* 44, e113
101. Utturkar, S.M. *et al.* (2014) Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics* 30, 2709–2716
102. Heydari, M. *et al.* (2017) Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC Bioinformatics* 18, 374
103. Smith, H.E. and Yun, S. (2017) Evaluating alignment and variant-calling software for mutation identification in *C. elegans* by whole-genome sequencing. *PLoS One* 12, e0174446
104. Thankaswamy-Kosalai, S. *et al.* (2017) Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics* 109, 186–191
105. Schatz, M.C. *et al.* (2012) Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.* 13, 243
106. Michael, T.D. and Jackson, S. (2013) The first 50 plant genomes. *Plant Genome* 6, 1–7
107. Hulse-Kemp, A.M. *et al.* (2018) Reference quality assembly of the 3.5 Gb genome of *Capsicum annuum* from a single linked-read library. *Hortic. Res.* 5, 4
108. Jackman, S.D. *et al.* (2018) Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* 19, 393
109. Liu, Q. *et al.* (2018) Assembly and annotation of a draft genome sequence for *Glycine latifolia*, a perennial wild relative of soybean. *Plant J.* 95, 71–85
110. Ott, A. *et al.* (2018) Linked read technology for assembling large complex and polyploidy genomes. *BMC Genomics* 19, 651
111. Marks, P. *et al.* (2019) Resolving the full spectrum of human genome variation using linked-reads. *Genome Res.* 29, 635–645
112. Ashton, P.M. *et al.* (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* 33, 296–300
113. Jain, M. *et al.* (2017) MinION analysis and reference consortium: phase 2 data release and analysis of R9.0 chemistry. *F1000Res* 6, 760
114. Debladis, E. *et al.* (2017) Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore sequencing technology. *BMC Genomics* 18, 537
115. Leggett, R.M. and Clark, M.D. (2017) A world of opportunities with nanopore sequencing. *J. Exp. Bot.* 68, 5419–5429
116. Jain, M. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345
117. Gordon, D. *et al.* (2016) Long-read sequence assembly of the gorilla genome. *Science* 352, aae0344
118. Magi, A. *et al.* (2018) Nanopore sequencing data analysis: state of art, applications and challenges. *Brief. Bioinform.* 19, 1256–1272
119. Rang, F.J. *et al.* (2018) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* 19, 90
120. Volden, R. *et al.* (2018) Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci.* 39, 9726–9731
121. Chu, J. *et al.* (2017) Innovations and challenges in detection long read overlaps: an evaluation of the state-of-the-art. *Bioinformatics* 33, 1261–1270
122. Kchouk, M. and Elloumi, M. (2016) Hybrid error correction approach and de novo assembly for minion sequencing long reads. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 122–125, IEEE
123. Carvalho, A.B. *et al.* (2016) Improved assembly of noisy long reads by k-mer validation. *Genome Res.* 26, 1710–1720
124. Cao, M.D. *et al.* (2017) Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat. Commun.* 8, 14515
125. Mostovoy, Y. *et al.* (2016) A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods* 13, 587–590
126. Bickhart, D.M. *et al.* (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* 49, 43–650
127. Weissensteiner *et al.* (2017) Combination of short-read, long-read and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res.* 27, 697–708
128. Li, X. *et al.* (2016) Improved hybrid de novo genome assembly of domesticated apple (*Malus* x *domestica*). *GigaScience* 5, 35
129. Clavijo, B.J. *et al.* (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* 27, 885–896
130. Miller, J.R. *et al.* (2017) Hybrid assembly with long and short reads improves discovery of gene family expansions. *BMC Genomics* 18, 541
131. Goodwin, S. *et al.* (2015) Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* 25, 1750–1756

132. Madoui, M.-A. *et al.* (2015) Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* 16, 327
133. Belser, C. *et al.* (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* 4, 879–887
134. Alhakami, H. *et al.* (2017) A comparative evaluation of genome assembly reconciliation tools. *Genome Biol.* 18, 93
135. Hunt, M. *et al.* (2014) A comprehensive evaluation of assembly scaffolding tools. *Genome Biol.* 15, R42
136. Belaghzal, H. *et al.* (2017) Hi-C 2.0: an optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* 123, 56–65
137. Ghurye, J. *et al.* (2017) Scaffolding of long read assemblies using long range contract information. *BMC Genomics* 18, 527
138. Conte, M.A. *et al.* (2017) A high quality assembly of the Nile tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics* 18, 341
139. Paajanen, P. *et al.* (2019) A critical comparison of technologies for a plant genome sequencing project. *GigaScience* 8, 1–12
140. Wences, A.H. and Schatz, M.C. (2015) Metassembler: merging and optimizing de novo genome assemblies. *Genome Biol.* 16, 207
141. Chakraborty, M. *et al.* (2016) Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* 44, e147
142. Lam, K.-K. *et al.* (2016) BIGMAC: breaking inaccurate genomes and merging assembled contigs for long read metagenomic assembly. *BMC Bioinformatics* 17, 435
143. Thind, A.K. *et al.* (2017) Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. *Nat. Biotechnol.* 35, 793–796
144. Thind, A.K. *et al.* (2018) Chromosome-scale comparative sequence analysis unravels molecular mechanisms of genome dynamics between two wheat cultivars. *Genome Biol.* 19, 104
145. Chen, X. *et al.* (2018) Transcriptome-referenced association study of clove shape traits in garlic. *DNA Res.* 25, 587–596
146. Wang, B. *et al.* (2018) A comparative transcriptional landscape of maize and sorghum obtained by single-molecular sequencing. *Genome Res.* 28, 921–932
147. Kronenberg, Z.N. *et al.* (2018) Extended haplotype phasing of de novo genome assemblies with FALCON-Phase. *bioRxiv.* Published online April 19, 2019. https://doi.org/10.1101/327064
148. Du, H. and Liang, C. (2018) Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *bioRxiv.* Published online June 13, 2018. http://doi.org/10.1101/345983
149. Schreiber, M. *et al.* (2018) Genomic approaches for studying crop evolution. *Genome Biol.* 19, 140