Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in Plant Biology

Building near-complete plant genomes

Todd P Michael[1] and Robert VanBuren[2,3]

Plant genomes span several orders of magnitude in size, vary in levels of ploidy and heterozygosity, and contain old and recent bursts of transposable elements, which render them challenging but interesting to assemble. Recent advances in single molecule sequencing and physical mapping technologies have enabled high-quality, chromosome scale assemblies of plant species with increasing complexity and size. Single molecule reads can now exceed megabases in length, providing unprecedented opportunities to untangle genomic regions missed by short read technologies. However, polyploid and heterozygous plant genomes are still difficult to assemble but provide opportunities for new tools and approaches. Haplotype phasing, structural variant analysis and *de novo* pan-genomics are the emerging frontiers in plant genome assembly.

Addresses
[1] Informatics Department, J. Craig Venter Institute, La Jolla, CA, USA
[2] Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA
[3] Plant Resilience Institute, Michigan State University, East Lansing, MI 48824, USA

Corresponding author: Michael, Todd P (toddpmichael@gmail.com)

Current Opinion in Plant Biology 2020, **54**:26–33

This review comes from a themed issue on **Genome studies and molecular genetics**

Edited by **Josh T Cuperus** and **Christine Queitsch**

For a complete overview see the Issue and the Editorial

Available online 22nd January 2020

https://doi.org/10.1016/j.pbi.2019.12.009

Introduction

As we celebrate the 20th anniversary of the first plant genome for the model plant *Arabidopsis thaliana* [1], we are entering the golden age for assembling high quality, chromosome-scale genomes across a range of sizes, complexity and ploidy. While the Arabidopsis genome was sequenced using a Sanger-based bacterial artificial chromosome (BAC)-by-BAC approach by a large consortium over several years, it is now possible to create a high-quality Arabidopsis reference genome that has higher contiguity in less than a week [2,3•]. The new single-molecule sequencing technologies have reinforced the need to carefully select plant material, generate high molecular weight (HMW) DNA, and choose sequencing and assembly strategies [4]. The future of plant genome assembly is upon us, where the idea of a single reference genome is a thing of the past and pan-genome graph assemblies become the standard to enable basic plant biology, breeding and industrial applications [5•].

Many of the original high-quality plant reference genomes were assembled using a using a minimum tiling path of BACs sequenced with Sanger technology [6]. While these genomes provided high-quality chromosome scale references, they were expensive and labor intensive, usually requiring large consortium. However, it was demonstrated with *Drosophila melanogaster* that shotgun sequencing coupled to an overlap layout consensus (OLC) approach using the CELERA assembler resulted in high quality assemblies at a fraction of the cost and time [7]. Shotgun sequencing and OLC assemblers were adopted for some of the early plant genome such as papaya, soy, and poplar, among others [8–13]. The emergence of second generation sequencing technologies such as 454 and Illumina spurred the development of De Bruijn graph (DBG) assembly methods to handle shorter reads sequenced at greater depth [14–17]. The ease and cost of sequencing led to a revolution in plant genomics that resulted in many lower quality assemblies yet a precipitous increase in transformative genome-enabled discoveries about fundamental plant biology [18].

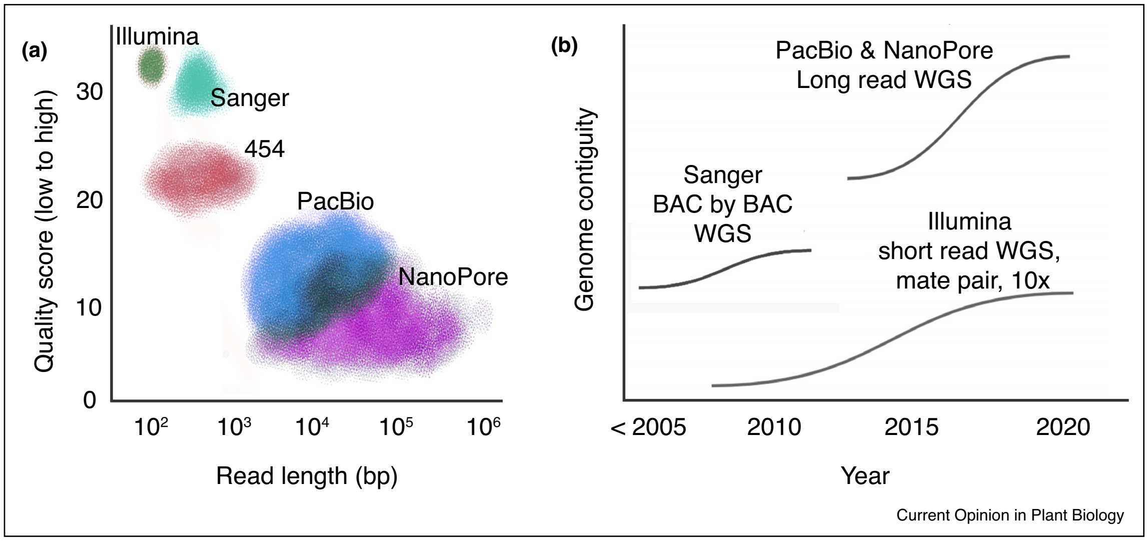
Single molecule sequencing once again changed the landscape of plant genome assembly, enabling near complete chromosomes for the first time. The first plant genome based completely on Pacific Biosciences (PacBio) single molecule real time (SMRT) sequencing was the smallest grass and desiccation tolerant *Oropetium thomaeum*, which resulted in the fourth most contiguous genome at the time including 30% complete centromeres [19••]. PacBio SMRT sequencing required the development of Falcon [20•], a new assembler specific to long error-prone reads, which was then followed by an update to the CELERA assembler (CANU) [21], both of which were critical to the flurry of genomes that followed. Notable was the update to the *Zea mays* genome, which significantly improved the contiguity of one of the most challenging genomes sequenced during the Sanger effort [22,23•]. Complementary technology emerged with the promise of bridging the near complete single molecule genomes to chromosome scale assemblies without the aid of expensive BAC-based physical maps or labor intensive genetic maps such as high throughput chromatin conformation capture (Hi-C) [24] and optical maps [25]. Finally, Oxford Nanopore Technologies (ONT) released the first nanopore sequencer that exceeded PacBio in read lengths with some exceeding megabases (Mb), and the ability to assemble more contiguous and complete versions of reference genomes of Arabidopsis, Tomato, Sorghum, Banana and Brassica [3•,26•,27,28••].

Over the past 20 years ~400 plant genomes have been published, including 333 angiosperms, 15 non-angiosperms, 2 Charophytes and 44 Chlorophyte green algae (https://www.plabipd.de/portal/web/guest/sequenced-plant-genomes). While there are some excellent recent reviews that more thoroughly cover older techniques and recipes for sequencing and assembling genomes [29–31], we are focusing on current advances in long read sequencing and assembly technologies that have enabled building near-complete plant genomes over the past couple years.

Long read single-molecule sequencing

The primary driver for improved plant genome assemblies has been the development of long read sequencing technologies (Figure 1). While assembly methods have also improved, and new physical mapping technologies have been developed (discussed below), read length is still a limiting factor for high quality plant genome assemblies. Plant genomes prove to be the most challenging to assemble due to high levels of heterozygosity, complex polyploidy, and explosive repeat content. Read length must exceed the dominant repeat length found in a genome, and nested long terminal repeats (LTR) or haplotype blocks that can span 20–200 kilobases (kb). PacBio was the first to provide long reads (>1 kb) and routinely enables the generation of read N50 lengths greater than 20 kb (Figure 1a). ONT later released the MinION and then the PromethION sequencers that enabled read length N50s in the 20–50 kb range with some reads exceeding several Mb in length. In theory, ONT read length is only limited by the input material and this has resulted in the need for new DNA extraction protocols that preserve the HMW molecules [32]. The long read tradeoff is that both PacBio and ONT have a higher error-rate than previous sequencing platforms (Figure 1a). PacBio has addressed the error-rate issue by updating their circular consensus sequencing (CCS) approach to generate long high fidelity (HiFi) 15 kb reads with accuracy upwards of 99.8% [33]. However, the high-quality reads come at a 5x cost increase per read currently, and even almost perfect 15 kb reads may not enable assembly of the nested, highly similar repeat structures often found in complex plant genomes. Many complex plant genomes have repeat structures greater than 20 kb and long reads, even those with the current error-rate, have facilitated genome assemblies with significantly increased genome contiguity, or completeness, as compared to previous technologies (Figure 1b). An update to the Arabidopsis Col-0 genome using ONT reduced the assembly to 40 contigs that spanned chromosome arms (telomere to centromere) and resolved previously identified gaps and misassemblies in the TAIR10 reference [2]. Moreover, sequencing of another Arabidopsis accession with ONT enabled the resolution of a quantitative trait loci (QTL) previously recalcitrant to BAC sequencing due to its repeat structure that required reads greater than 20 kb [3•].

Figure 1



Advances in sequencing technology have dramatically improved genome contiguity over the last two decades. **(a)** Sanger and Illumina sequencing technologies have short read lengths but high per base quality. Long read sequencing technologies (PacBio and NanoPore) have reads that can exceed 1 Mb but have a much lower per base quality. **(b)** Long read sequencing technologies have driven the largest improvements in genome contiguity, or completeness, over the last ~5 years. These are schematic depictions of sequencing technologies and not actual data.

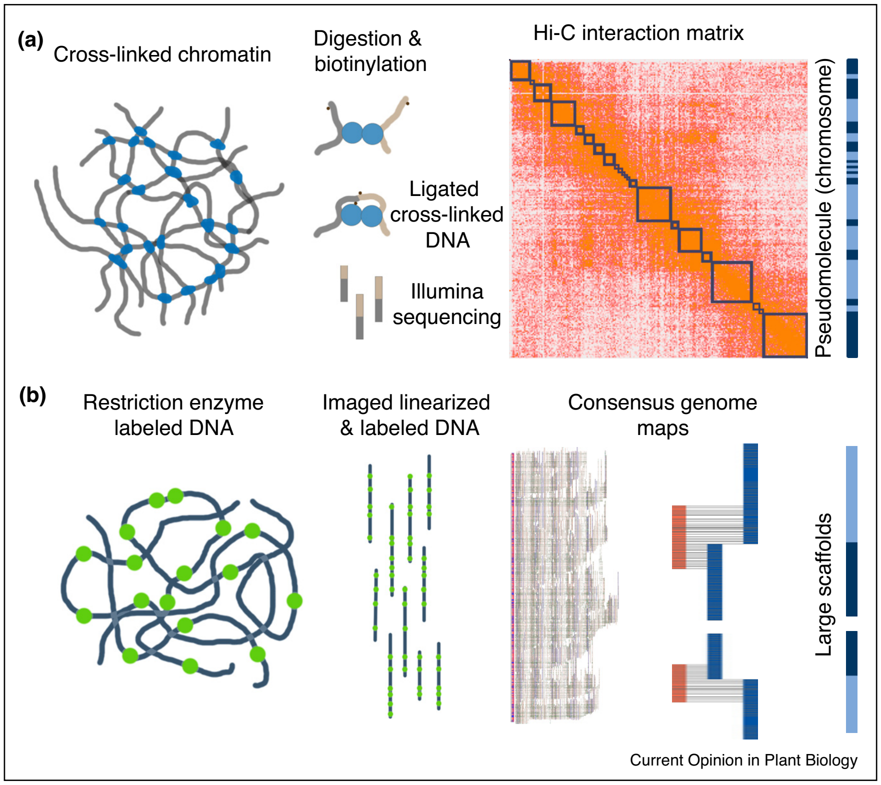
Long read error-prone genome assembly

The growth of single-molecule sequencing technologies spurred the development of new genome assembly algorithms designed to correct, overlap, and polish long reads with high error-rates. Algorithms vary in computational design, speed and memory usage, and utility for assembling complex, heterozygotic, polyploid, or large genomes. Most leading assemblers such as CANU [21], Falcon (phase and unzip) [20•,34], MARVEL [35] and MECAT [36] utilize a self-correction approach, where long reads are aligned against each other and errors are corrected with sufficiently high coverage. In contrast, several long read assemblers are 'correction-free' such as the OLC-based minimap2/miniasm [37••], or can leverage corrected reads such as the DBG-based WTDBG2 [38] and FlyE [39]. Correction-free assembly is required for some highly complex genomes like *Cannabis sativa*, where the Tetrahydrocannabinol (THC) synthase gene is nested in 70–90 kb LTR-based tandem repeats [40]. Draft long read assemblies have residual errors and must be polished using a combination of high-coverage long read and/or short read data. Quiver/Arrow (PacBio), Medaka (ONT), Nanopolish [41], and Racon [42] are designed to utilize long read data and Pilon [43] uses short read Illumina data for polishing. In general, three or more rounds of consensus and polishing are recommended for long read assemblies to reach >99.6% accuracy [3•]. Recently a new assembler based on Flacon called Peregrine was introduced that uses sparse hierarchical minimizers (SHIMMER) to leverage high-quality long reads like PacBio CCS HiFi reads and utilizes a fraction of the compute time and RAM [44].

Physical mapping technologies

Single-molecule reads produce draft assemblies with high contiguity, but chromosome scale assemblies are needed for marker assisted breeding, quantitative genetics and comparative genomics. High-density genetic maps were traditionally used to anchor contigs and scaffolds into chromosomes, but they are prone to ordering and orientation issues [45]. Genetic maps also fail to anchor

Figure 2


(a)
Cross-linked chromatin
Digestion & biotinylation
Ligated cross-linked DNA
Illumina sequencing
Hi-C interaction matrix
Pseudomolecule (chromosome)
(b)
Restriction enzyme labeled DNA
Imaged linearized & labeled DNA
Consensus genome maps
Large scaffolds
Current Opinion in Plant Biology

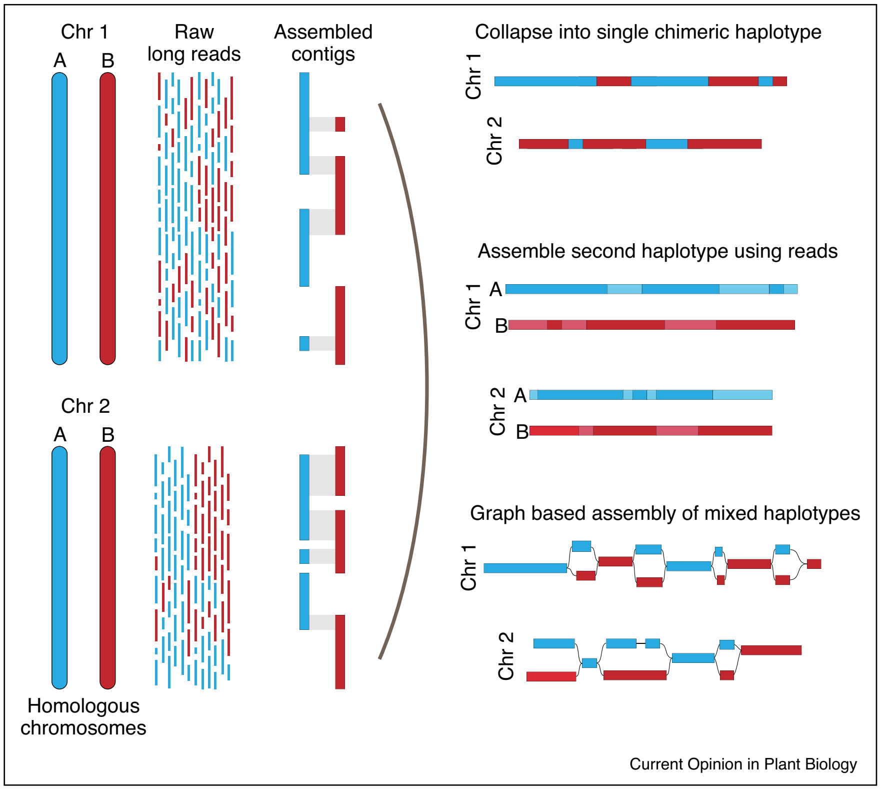
Leading strategies for scaffolding long read assemblies. **(a)** High throughput chromatin conformation capture (Hi-C) relies on the proximity of interactions from cross-linked chromatin to order contigs. Chromatin is cross-linked, digested with restriction enzymes and biotinylated, and the two chromatin ends are ligated and purified using streptaviadin beads. The resulting library is sequenced and aligned to the genome to build a Hi-C interaction matrix for scaffolding contigs into pseudomolecules. **(b)** Optical maps utilize restriction enzymes and single molecule imaging to create a physical map of the genome. Long fragments of DNA are nicked using a restriction enzyme and labeled. DNA molecules are linearized and imaged, and fingerprints for each molecule are combined to create a consensus genome map. Contigs are overlaid on the genome map based on *in silico* digestion and anchored into scaffolds or pseudomolecules.

sequences with low rates of recombination or low marker density and require access to segregating populations. The current leading scaffolding approaches rely on long-range chromatin interactions or optical maps for contig anchoring (Figure 2). Hi-C relies on the density and proximity of interactions from cross-linked chromatin to orient and order contigs [46•,47]. Chromosomes occupy distinct regions of the nucleus, and intrachromosomal interaction is much more likely to occur than interactions between chromosomes. The probability of chromatin interactions decays with linear distance, and long-range interactions (>100 Mb) are rare but more frequent than interchromosomal contacts. This principle can be used to reliably create a Hi-C interaction matrix for adjacent regions to order and scaffold even small or repetitive contigs. While Hi-C routinely enables the resolution of genomes into chromosomes, even for allotetraploids like *Eragrostis tef* [48], more complex plant genomes can prove problematic. Another physical mapping technology showing promise is optical mapping, which stretches labeled DNA molecules through nanochannels [49]. Optical mapping approaches are technically challenging and require optimization for each species. However, recent updates to the labeling techniques have enabled chromosome scale assemblies using BioNano Genomics Direct Label and Stain (DSL) coupled to ONT long reads [27,28••].

Resolving complex plant genomes

Most plants have more than one copy of each chromosome and reference genomes represent a collapsed mosaic of segments from two or more homologous chromosomes or haplotypes. A haploid or monoploid reference simplifies downstream analyses but fails to capture the true genome composition of an individual. Inbred species such as maize and Arabidopsis are highly homozygotic, but many outcrossing species and clonally propagated crops are highly heterozygotic, with numerous repetitive elements, single nucleotide polymorphisms, and structural variants distinguishing haplotypes [50,51]. Accurate assembly and phasing of haplotypes are essential for allele specific analyses of complex traits such as heterosis and subgenome dominance, and cloning heterozygous loci with biological or agronomic importance. Long read assembly algorithms can accurately correct and disentangle divergent haplotypes, leading to assemblies that exceed the monoploid genome size

Figure 3


Chr 1
A
B
Raw long reads
Assembled contigs
Chr 2
A
B
Homologous chromosomes
Collapse into single chimeric haplotype
Chr 1
Chr 2
Assemble second haplotype using reads
Chr 1
A
B
Chr 2
A
B
Graph based assembly of mixed haplotypes
Chr 1
Chr 2
Current Opinion in Plant Biology

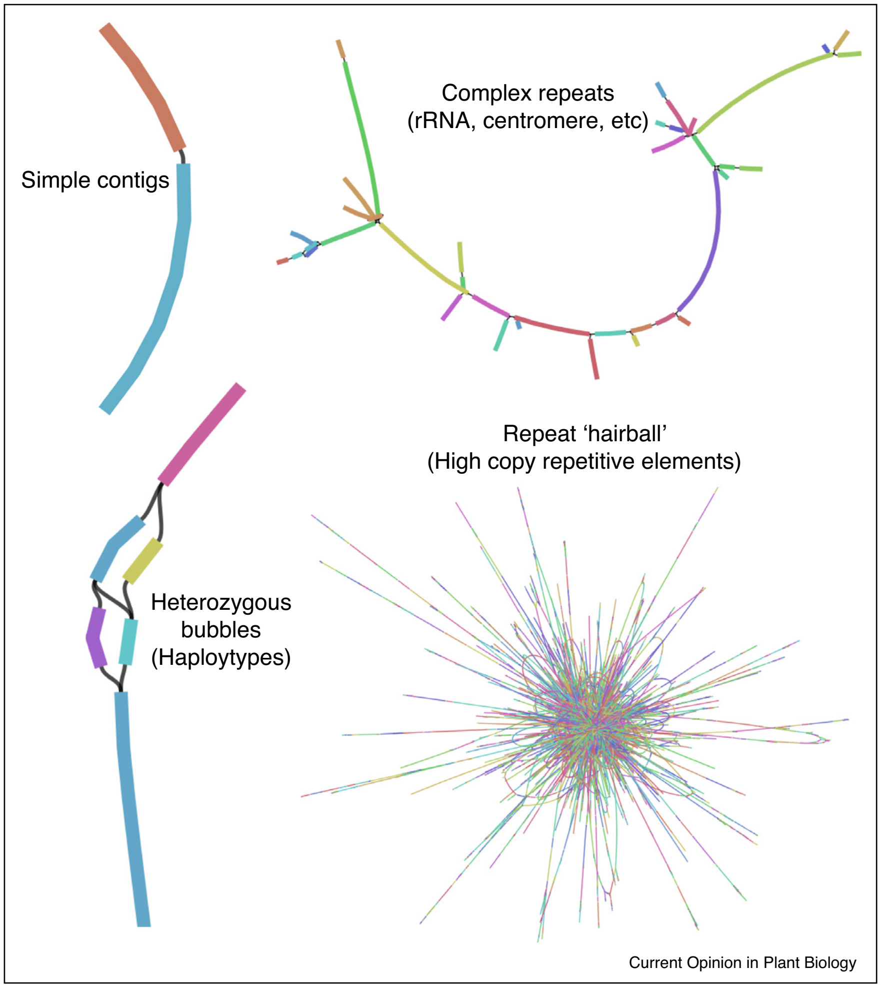
Assembly approaches for sequencing and phasing heterozygous genomes. Long read assemblies allow assembly of multiple haplotypes from homologous chromosomes in heterozygous regions. The primary and alternative haplotypes can be collapsed into a single, non-redundant but chimeric pseudomolecule for simplicity of downstream analyses (top). Raw reads can be mapped to the contigs to resolve missing haplotype regions to create a phased, diploid assembly (middle). Partial haplotypes can be retained and labeled in a graph-based assembly (bottom).

(Figure 3). These genomes have a mixture of 'duplicated' regions where divergent haplotypes are assembled separately, and single copy regions where haplotypes with few polymorphisms are collapsed into one. Duplicated regions must be marked before downstream analysis of gene dosage and duplications, and mapping of expression, resequencing, methylation, or other datasets. Partially phased genomes can be collapsed into a chimeric monoploid, phased into complete haplotypes, or be maintained as a mixture of haplotypes using a graph-based assembly structure (Figure 3). FALCON-Unzip has been used to create phased diploid assemblies of several grape species [20•,52], and a combination of 10x Genomics linked long reads and high-coverage Illumina data were used to assemble and phase all four haplotypes of autotetraploid blueberry [53]. PacBio and Hi-C enabled assembly and phasing of the octoploid sugarcane [54•], and the complex allotetraploid peanut [55], teff [48], and broomcorn millet genomes [56] among others.

Leveraging the assembly graph

The quality of a draft genome assembly is typically assessed by the contig N50 or the shortest sequence length containing >50% of the assembly. The 'best' assembly is often chosen by tweaking parameters or testing different algorithms to produce the highest N50. This approach is problematic as this simplistic metric ignores ambiguities that were encountered during the overlap steps of assembly [57]. A genome assembly graph can be used to visualize complexities and overlaps between adjacent contigs [58••,59] (Figure 4). An ideal graph would have a single edge for each contig (node) with connections representing adjacent sequences along a chromosome, and this often is the case for simple

Figure 4



Assessing genome quality using genome assembly graphs. Assembled contigs are shown as nodes and the connections between those contigs are represented by edges. The simplest instance would be one or more linear contigs connected by a single edge. Heterozygous or homoeologous regions have bubbles where nodes (haplotypes) are connected by multiple edges. Complex repeats such as ribosomal RNA (rRNA) or centromeric satellite DNA can create higher-order ambiguities in the graph structure. 'Hairballs' with thousands of nodes and edges are common and are likely driven by complex genome features and high copy number repeats.

homozygous genomes like Arabidopsis. Assembled haplotypes create 'bubbles' on the graph where one collapsed region is connected to two adjacent phased haplotypes, which is often the case for tree genomes that have a high level of heterozygosity and older repeats [60]. Complex sequences and high copy number repeats (LTRs, centromeres, etc.) can create indiscernible 'hairballs' of thousands of interconnected nodes with no clear paths. Genome graphs can help identify parameters that should be modified in future assembly iterations or test if more coverage or other technologies are needed to resolve assembly issues. Ultimately, genome graphs may be a better way to represent the complexity of genomes, especially as the concept of the reference genome is replaced by the pan-genome [61].

Ongoing challenges and future prospects

Polyploidy and heterozygosity are ongoing challenges in genome assembly, but complete, gapless, and fully phased plant genomes are on the horizon. The throughput of single-molecule sequencing is rising in parallel with plummeting costs, mirroring the trends observed in next generation sequencing during the early 2010s. This will facilitate not only sequencing virtually any plant species, but also generating numerous *de novo* references within a single species to capture the true genetic diversity. In the next few years, *de novo* assembly will replace whole genome resequencing for population genetics and pan-genome analyses [62]. Advances in gene annotation have lagged behind improvements in genome assembly and generating accurate gene predictions is still a major limitation. Improving annotation quality will require not only new technologies such as nanopore full length cDNA sequencing or PacBio Iso-seq, but also new algorithms to better predict functional genomic elements.

Funding

This work was supported in part by funding from the National Science Foundation (NSF-MCB 1817347 to R.V.), and the J. Craig Venter Institute (T.P.M).

Conflict of interest statement

Nothing declared.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**:796-815.

2. Jupe F, Rivkin AC, Michael TP, Zander M, Motley ST, Sandoval JP, Slotkin RK, Chen H, Castanon R, Nery JR *et al.*: **The complex architecture and epigenomic impact of plant T-DNA insertions**. *PLoS Genet* 2019, **15**:e1007819.

3. • Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR: **High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell**. *Nat Commun* 2018, **9**:541

The authors demonstrate that long read ONT data can be used to assemble a near complete genome (Arabidopsis) in the span of a week for relatively little cost.

4. Li F-W, Harkess A: **A guide to sequence your favorite plant genomes**. *Appl Plant Sci* 2018, **6**:e1030.

5. • Hurgobin B, Edwards D: **SNP discovery using a pangenome: has the single reference approach become obsolete?** *Biology* 2017, **6**

The authors discuss using pangenomes for variant discovery to capture the true variability within genomes.

6. Michael TP, Jackson S: **The first 50 plant genomes**. *Plant Genome* 2013, **6**.

7. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.

8. Sutton GG, White O, Adams MD, Kerlavage AR: **TIGR assembler: a new tool for assembling large shotgun sequencing projects**. *Genome Sci Technol* 1995, **1**:9-19.

9. Huang X, Wang J, Aluru S, Yang S-P, Hillier L: **PCAP: a whole-genome assembly program**. *Genome Res* 2003, **13**:2164-2170.

10. Batzoglou S: **ARACHNE: a whole-genome shotgun assembler**. *Genome Res* 2002, **12**:177-189.

11. Mullikin JC, Ning Z: **The phusion assembler**. *Genome Res* 2003, **13**:81-90.

12. Shapiro H: *Outline of the Assembly process: JAZZ, the JGI In-House Assembler*. 2005 http://dx.doi.org/10.2172/843143.

13. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly**. *Proc Natl Acad Sci U S A* 2001, **98**:9748-9753.

14. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376-380.

15. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res* 2008, **18**:821-829.

16. Chaisson MJ, Pevzner PA: **Short read fragment assembly of bacterial genomes**. *Genome Res* 2008, **18**:324-330.

17. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K *et al.*: **De novo assembly of human genomes with massively parallel short read sequencing**. *Genome Res* 2010, **20**:265-272.

18. Michael TP, VanBuren R: **Progress, challenges and the future of crop genomes**. *Curr Opin Plant Biol* 2015, **24**:71-81.

19. •• VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E *et al.*: **Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum***. *Nature* 2015, **527**:508-511

The authors describe the first plant genome sequenced with only single molecule long reads, which resulted in the fourth most continuous genome at the time with several full centromeres.

20. • Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A *et al.*: **Phased diploid genome assembly with single-molecule real-time sequencing**. *Nat Methods* 2016, **13**:1050-1054

The authors describe FALCON-Unzip, a new iteration of FALCON that utilizes long read data to phase heterozygotic regions to produce true diploid assemblies.

21. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation**. *Genome Res* 2017, **27**:722-736.

22. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al.*: **The B73 maize**

genome: complexity, diversity, and dynamics. *Science* 2009, **326**:1112-1115.

23. • Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S *et al.*: **Improved maize reference genome with single-molecule technologies**. *Nature* 2017, **546**:524-527

The authors use long read PacBio data to reassemble the large, repeat dense maize genome. The updated maize reference had significant improvements in contiguity and gene content.

24. Xie T, Zheng J-F, Liu S, Peng C, Zhou Y-M, Yang Q-Y, Zhang H-Y: **De novo plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana***. *Mol Plant* 2015, **8**:489-492.

25. Michael TP, Bryant D, Gutierrez R, Borisjuk N, Chu P, Zhang H, Xia J, Zhou J, Peng H, El Baidouri M *et al.*: **Comprehensive definition of genome features in *Spirodela polyrhiza* by high-depth physical mapping and short-read DNA sequencing strategies**. *Plant J* 2017, **89**:617-635.

26. • Schmidt MH-W, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Maß J, Pfaff C *et al.*: **De novo assembly of a new *Solanum pennellii* accession using nanopore sequencing**. *Plant Cell* 2017, **29**:2336-2348

The authors present the first large plant genome sequenced with ONT.

27. Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H: **A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping**. *Nat Commun* 2018, **9**:4844.

28. •• Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, Genete M, Berrabah W, Chèvre A-M, Delourme R *et al.*: **Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps**. *Nat Plants* 2018, **4**:879-887

The authors demonstrate high quality assembly for three larger plant genomes using a combination of ONT and BioNano DLS.

29. Li C, Lin F, An D, Wang W, Huang R: **Genome sequencing and assembly by long reads in plants**. *Genes* 2017, **9**.

30. Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV: **Current strategies of polyploid plant genome sequence assembly**. *Front Plant Sci* 2018, **9**:1660.

31. Jiao W-B, Schneeberger K: **The impact of third generation genomic technologies on plant genome assembly**. *Curr Opin Plant Biol* 2017, **36**:64-70.

32. Lutz KA, Wang W, Zdepski A, Michael TP: **Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing**. *BMC Biotechnol* 2011, **11**:54.

33. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND *et al.*: **Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome**. *Nat Biotechnol* 2019, **37**:1155-1162.

34. Kronenberg ZN, Hall RJ, Hiendleder S, Smith TPL, Sullivan ST, Williams JL, Kingan SB: **FALCON-Phase: integrating PacBio and Hi-C data for phased diploid genomes**. *bioRxiv* 2018 http://dx.doi.org/10.1101/327064.

35. Nowoshilow S, Schloissnig S, Fei J-F, Dahl A, Pang AWC, Pippel M, Winkler S, Hastie AR, Young G, Roscito JG *et al.*: **The axolotl genome and the evolution of key tissue formation regulators**. *Nature* 2018, **554**:50-55.

36. Xiao C-L, Chen Y, Xie S-Q, Chen K-N, Wang Y, Han Y, Luo F, Xie Z: **MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads**. *Nat Methods* 2017, **14**:1072-1074.

37. •• Li H: **Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences**. *Bioinformatics* 2016, **32**:2103-2110

The author demonstrates a fast and accurate set of tools for genome assembly with long reads.

38. Ruan J, Li H: **Fast and accurate long-read assembly with wtdbg2**. *bioRxiv* 2019 http://dx.doi.org/10.1101/530972.

39. Kolmogorov M, Yuan J, Lin Y, Pevzner P: Assembly of Long Error-Prone Reads Using Repeat Graphs. [date unknown], DOI:10.1101/247148.

40. Grassa CJ, Wenger JP, Dabney C, Poplawski SG, Timothy Motley S, Michael TP, Schwartz CJ, Weiblen GD: A complete Cannabis chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. [date unknown], DOI:10.1101/458083.

41. Loman NJ, Quick J, Simpson JT: **A complete bacterial genome assembled de novo using only nanopore sequencing data**. *Nat Methods* 2015, **12**:733-735.

42. Vaser R, Sović I, Nagarajan N, Šikić M: **Fast and accurate de novo genome assembly from long uncorrected reads**. *Genome Res* 2017, **27**:737-746.

43. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK *et al.*: **Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement**. *PLoS One* 2014, **9**: e112963.

44. Chin C-S, Khalak A: Human Genome Assembly in 100 Minutes. [date unknown], doi:10.1101/705616.

45. Mascher M, Stein N: **Genetic anchoring of whole-genome shotgun assemblies**. *Front Genet* 2014, **5**:208.

46. • Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J: **Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions**. *Nat Biotechnol* 2013, **31**:1119-1125

The authors adapt Hi-C for generating long-range interaction matrixes to build chromosome-scale genomes. Hi-C is arguably the best scaffolding method and it allows anchoring of small or repetitive contigs quickly with relatively low cost and input.

47. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW *et al.*: **Chromosome-scale shotgun assembly using an in vitro method for long-range linkage**. *Genome Res* 2016, **26**:342-350.

48. VanBuren R, Wai CM, Pardo J, Yocca AE, Wang X, Wang H, Chaluvadi SR, Bryant D, Edger PP, Bennetzen JL *et al.*: **Exceptional subgenome stability and functional divergence in allotetraploid teff, the primary cereal crop in Ethiopia**. *bioRxiv* 2019 http://dx.doi.org/10.1101/580720.

49. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M *et al.*: **Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly**. *Nat Biotechnol* 2012, **30**:771-776.

50. VanBuren R, Wai CM, Ou S, Pardo J, Bryant D, Jiang N, Mockler TC, Edger P, Michael TP: **Extreme haplotype variation in the desiccation-tolerant clubmoss Selaginella lepidophylla**. *Nat Commun* 2018, **9**:13.

51. Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS: **Structural variants, hemizygosity and clonal propagation in grapevines**. *bioRxiv* 2019 http://dx.doi.org/10.1101/508119.

52. Girollet N, Rubio B, Bert P-F: **De novo phased assembly of the *Vitis riparia* grape genome**. *bioRxiv* 2019 http://dx.doi.org/10.1101/640565.

53. Colle M, Leisner CP, Wai CM, Ou S, Bird KA, Wang J, Wisecaver JH, Yocca AE, Alger EI, Tang H *et al.*: **Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry**. *Gigascience* 2019, **8**.

54. • Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, Zhu F, Jones T, Zhu X, Bowers J *et al.*: **Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L.** *Nat Genet* 2018, **50**:1565-1573

The authors assembled the sugarcane genome. Sugarcane is octoploid, and perhaps the most complex plant genome sequenced to data. A novel Hi-C scaffolding algorithm was developed to phase sugarcane haplotypes.

55. Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, Leal-Bertioli SCM, Ren L, Farmer AD, Pandey MK *et al.*: **The**

genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat Genet* 2019, **51**:877-884.

56. Zou C, Li L, Miki D, Li D, Tang Q, Xiao L, Rajput S, Deng P, Peng L, Jia W *et al.*: **The genome of broomcorn millet**. *Nat Commun* 2019, **10**:436.

57. Sedlazeck FJ, Lee H, Darby CA, Schatz MC: **Piercing the dark matter: bioinformatics of long-range sequencing and mapping**. *Nat Rev Genet* 2018, **19**:329-346.

58. Wick RR, Schultz MB, Zobel J, Holt KE: **Bandage: interactive visualization of de novo genome assemblies**. *Bioinformatics* 2015, **31**:3350-3352
••
The authors report the visualization tool Bandage, which has a simple graphic user interface that allows users to view the genome assembly graph and inform future improvement strategies.

59. Mikheenko A, Kolmogorov M: **Assembly Graph Browser: interactive visualization of assembly graphs**. *Bioinformatics* 2019 http://dx.doi.org/10.1093/bioinformatics/btz072.

60. Tischler G: **Haplotype and repeat separation in long reads**. *bioRxiv* 2017 http://dx.doi.org/10.1101/145474.

61. Tao Y, Zhao X, Mace E, Henry R, Jordan D: **Exploring and exploiting pan-genomics for crop improvement**. *Mol Plant* 2019, **12**:156-169.

62. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL *et al.*: **The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor**. *Nat Genet* 2019, **51**:1044-1051.