Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

The impact of third generation genomic technologies on plant genome assembly

Wen-Biao Jiao and Korbinian Schneeberger

Since the introduction of next generation sequencing, plant genome assembly projects do not need to rely on dedicated research facilities or community-wide consortia anymore, even individual research groups can sequence and assemble the genomes they are interested in. However, such assemblies are typically not based on the entire breadth of genomic technologies including genetic and physical maps and their contiguities tend to be low compared to the full-length gold standard reference sequences. Recently emerging third generation genomic technologies like long-read sequencing or optical mapping promise to bridge this quality gap and enable simple and cost-effective solutions for chromosomal-level assemblies.

Address
Max Planck Institute for Plant Breeding Research, Department of Plant Developmental Biology, Genome Plasticity and Computational Genomics, Cologne, Germany

Corresponding author: Schneeberger, Korbinian (schneeberger@mpipz.mpg.de)

Current Opinion in Plant Biology 2017, **36**:64–70

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

Edited by **Ian Henderson** and **Korbinian Schneeberger**

http://dx.doi.org/10.1016/j.pbi.2017.02.002

1369-5266/© 2017 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction

As a result of the drastic cost reduction for genome sequencing during the last decade, at least 183 plant reference sequences have been published so far (as of September 2016; www.plabipd.de) and first projects with *de novo* assemblies of multiple individuals of the same species appeared [1–5]. However, only the minority of the plant genome assemblies are on chromosome-level (*e.g.*, [6–8]). Most of them and this even includes some of the major crop species, are based on short-read sequencing and consist of hundreds or even thousands of fragmented contigs and scaffolds, which are usually not mapped to their chromosomal locations.

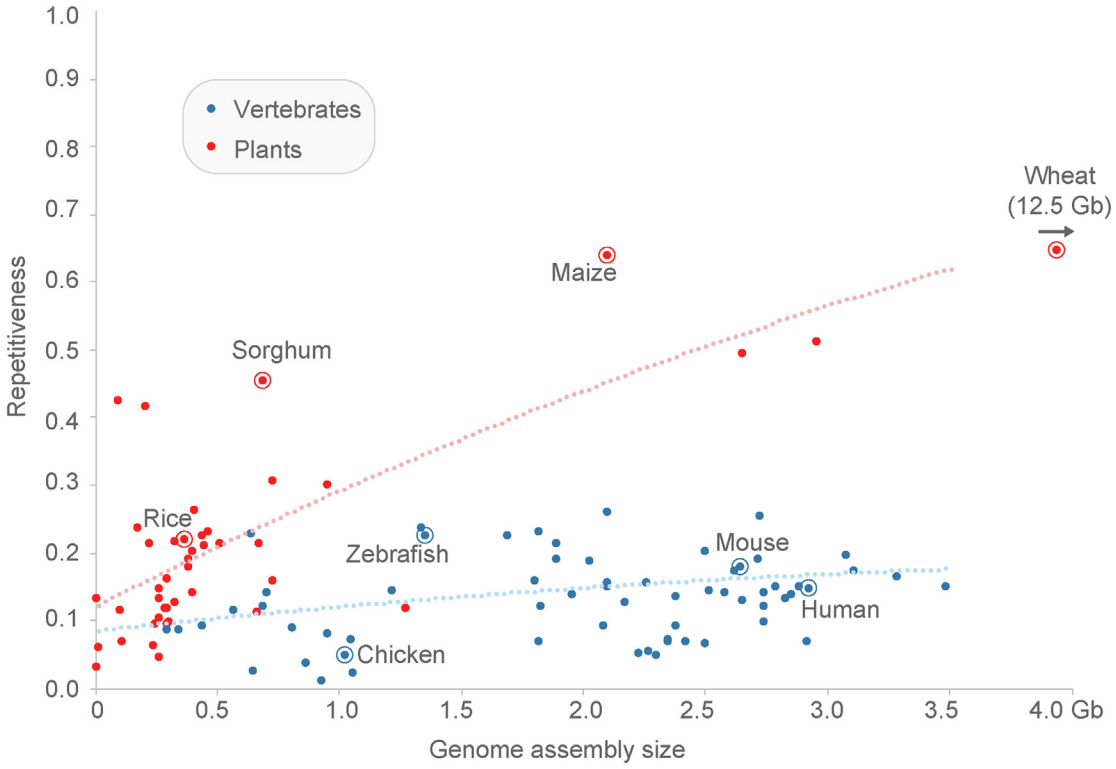
Obviously, the assembly of plant genomes is a challenging problem and presumably even more challenging than the assembly of vertebrate genomes (Figure 1). High repetitiveness due to transposable elements, extreme genome sizes, like the ones of the 22 Gb loblolly pine and 20 Gb Norway spruce genomes [9,10] and the polyploid nature of some plants and in particular of many of the crop species [11] display unique challenges [12]. Recently, long-read sequencing [13–15] and long-range scaffolding methods such as optical mapping [16], chromosome conformation capture [17], and DNA dilution-based technologies [18,19] were introduced to overcome the weaknesses of short-read assemblies and ultimately to enable the assembly of entire chromosomes [20••]. In this review, we focus on current and emerging third generation genomic technologies and their application for plant genome assembly particularly focusing on those, which are broadly available.

Long-read sequencing technologies

The most-widely used long-read sequencing technology is Pacific Biosciences' Single Molecule Real-Time (SMRT) sequencing (www.pacb.com). SMRT sequencing is performed on cells, which are patterned with tiny wells called zero-mode waveguides or ZMWs. Within each of these ZMWs a DNA polymerase/template complex gets immobilized and synthesizes a new DNA strand by incorporating phosophlinked nucleotides. Each incorporation leads to a light pulse that can be distinguished for differently labeled nucleotides [13]. Current PacBio systems generate reads with an average size of nearly 20 kb and a maximum length of over 60 kb [21,22•]. Although raw reads can have sequencing error rates of up to 15%, correction with short sequencing reads [23,24] or self-correction with sufficient sequencing data [25] enables genome assemblies with a sequence accuracy of over 99.999% simply by running freely available software, like FALCON or PBcR(MHAP) [26••,27•].

Still the costs of PacBio sequencing are quite substantial and reference sequences for large and repetitive genomes like the hexaploid wheat genome are assembled from short-reads only. However sophisticated assembly services like the one offered by NRGene (www.nrgene.com) promise convincing results even without new third generation technology, but come at certain costs. As alternative lower amounts of long reads could be used to improve short-read assemblies by gap closure or scaffolding [1,28,29••]. The computational strategies for such data integration are less straightforward as compared to *de novo* assemblies, which are typically based on single software tools [24,30]. Latest assembly tools even come

Figure 1



Comparison of size and repetitiveness of plant and vertebrate genomes. The figure shows 44 plant and 68 vertebrate genome assemblies analyzed for genome-wide repeat levels and genome size. Repetitiveness of plant genomes is generally higher and more correlated to genome size as in vertebrates. The challenges of plant genome assembly are therefore not only determined by genome size alone, but also by their increased levels of repetitiveness (Genome assemblies taken from the Ensembl Genomes release 32 [68]; low coverage genomes were excluded; repetitiveness was estimated by the percentage of non-unique 31-mers of all 31-mers in each assembly; genome sizes estimated by genome assembly size without considering ambiguous bases; dashed lines follow polynomial regression fitted to the data).

as handy push-button methods, which do not require prior adjustment of algorithmic parameters like *k*-mer sizes and therefore promise to simplify the practical assembly procedure [31••].

The first plant genomes assembled from PacBio data alone were from *Arabidopsis thaliana* [27•] and *Oropetium thomaeum* [22•]. The small genome of Arabidopsis was assembled at chromosome-arm level, while the assembly of the 245 Mb *O. thomaeum* genome featured a contig N50 of 2.4 Mb. Such contiguities were never reached by short-read assemblies, however, scaffolding (*i.e.*, ordering and orienting of contigs) using long-range read pairs could generate similar contiguities [32]. In addition to contiguity, PacBio assembly contigs feature less gaps (represented as Ns in the sequence) and their overall lengths includes more of the genomic space. For example, the PacBio-based assembly of *Arabis alpina*, a relative of Arabidopsis with an estimated genome size of 375 Mb, was 337 Mb long and with this nearly 30 Mb longer as compared to the scaffolds of an earlier short-read assembly, while the percentage of Ns was reduced from 9.2 to 3.3% [33,34••].

Besides PacBio, there are currently two other long-read sequencing technologies on the market. The first was introduced by Oxford Nanopore Technologies (nanoporetech.com), which provided access to their first sequencing system via an early-access program in 2014 [14,35]. In this technology, single DNA molecules are run through nanopores, where individual nucleotides create characteristic disruptions of the current of the nanopore, which reveal the sequence of the nucleotides. The system can process DNA fragments independent of their length, and thus read length is theoretically only limited by the length of the DNA molecules. While most of the reads reported so far are similar in length and accuracy to PacBio reads, the longest reads are up to 200 kb. First whole-genome assemblies using Oxford Nanopore data have reached N50 values of multiple hundred kb for fungal genomes, and bacterial genomes could be fully assembled with a nucleotide accuracy of over 99% [36,37]. However, so far there are no reports on plant genome assemblies using Oxford Nanopore data.

Another long-read technology was introduced by Illumina through a library preparation kit for Synthetic

Long-Reads (SLR) (www.illumina.com), which allows the assembly of long reads from short-read sequencing data [15,38]. The protocol starts with shearing DNA followed by size selection of fragments of ~10 kb. The fragments are diluted and distributed in multiple hundred aliquots such that each aliquot receives only a tiny fraction of a chromosome making it unlikely that two overlapping fragments from the same region are within one aliquot. After PCR amplification, sequencing libraries with unique barcodes are generated from each aliquot, pooled and sequenced using conventional Illumina short-read sequencing. The short reads of each aliquot can be identified by their barcode and can be assembled into SLRs. Reported SLR lengths range from 2 to 18 kb and have a sequence accuracy of more than 99.9%, which is much higher as compared to the accuracy of 'real' long reads of PacBio or Oxford Nanopore, but it is not clear how well SLR assembly works if the fragments contain repeats [15,39]. Generation of SLRs for *de novo* assembly requires high amounts of short read coverage (typically multiple hundred-fold genome coverage). First SLRs need to be assembled, before they can be used as reads for *de novo* assembly, which itself requires high coverage. So far, SLRs have been used to assemble a few eukaryotic genomes with genome sizes of some hundred Mb [15,38] and to improve a short-read assembly of a large maize genome [5], but despite the high quality of SLRs the assemblies hardly reached N50 statistics of more than 100 kb. The assembly of *C. elegans* even included remarkable amounts of misassemblies, which, however, could be due to shortcomings of the assembly procedure implying the need for adjusted algorithms for SLR *de novo* assembly [39], which are just being developed [40].

Long-range scaffolding technologies

Despite all of the impressive recent progress in long-read DNA sequencing, it was so far not possible to assemble a complete plant genome from sequence reads alone. To improve assembly contiguity, the contigs of an assembly need to be scaffolded, which typically starts by ordering the contigs using alignments of paired reads [41]. In particular read pairs from BAC or fosmid ends are powerful to increase contiguity and help bridging repeats, which are the main reason for breaks in the assembly [42]. Scaffolding can extend the contiguity of an assembly by orders of magnitude, however, to get a plant genome assembled on chromosome-level additional genetic or physical maps are so-far still required. Recently several novel technologies emerged, which promise to improve scaffolding and eventually eliminate the need for genetic or physical mapping.

One of these technologies is optical mapping, which was already invented at the end of last century [16], but only some years ago commercial high-throughput platforms, such as the Irys system by BioNano Genomics (www.bionanogenomics.com) or OptGen's Argus system (www.opgen.com), became available. Optical mapping generates fingerprints of DNA fragments of up to multiple hundred kb by imaging the patterns of restriction sites under light microscopes using fluorescently labeled enzymes [16,43]. Such individual fingerprints (or maps) can be assembled into genome-wide (consensus) maps [44], which can then be used to scaffold the contigs of a corresponding sequence assembly and identify large structural variations [22•,29••,45–48]. Genome assembly scaffolding and other applications of optical map data to plant genomics have been reviewed in detail in [49].

The combination of sequencing data and optical maps works particularity well, as the typical breaks in sequence assemblies are at repeats, while optical map assemblies have a bias to break at regions with closely-spaced restriction sites [50]. The contiguity improvement gained by integration of optical maps depends on different factors including the contiguity of the assembly before the integration as well as contiguity of the optical consensus maps themselves, which in the best cases led to the reconstruction of entire chromosomes [47,51]. In addition, consensus maps can also be used to identify misassembled contigs [29••,52]. Though breaking of contigs at misassembled sites shortens the contigs, it can even lead to longer scaffolds compared to scaffolding without prior contig correction [34••]. Interestingly, like optical maps can be used to control for errors in the sequence assembly, the sequence contigs can be used to control for errors in the optical maps. In theory this should allow to aggressively assemble the sequence data and the optical maps without stringent error control. However, this might even be unnecessary if optical map information is integrated during sequence assembly already using data structures that can include both types of data [53].

Another elegant solution to the challenges of chromosome-scale assembly is based on chromosome conformation capture sequencing (Hi-C), a method originally developed to study the three-dimensional folding of the genome [54]. Hi-C is based on proximity ligation of DNA fragments that are physically close in their natural conformation. They are ligated *in situ* before they are cleaved by restriction enzymes and isolated from cells. The ligated DNA fragments are then amplified and sequenced. Though not all DNA, which is in close proximity, is also closely linked, the majority of the read pairs generated from the two ends of such fragments comes from closely linked DNA and therefore these read pairs can be used for scaffolding [17,55]. A modified Hi-C protocol called Chicago is provided as a service by Dovetail Genomics since 2014 (www.dovetailgenomics.com). The Chicago method captures chromatin contacts after confounding biological signals are removed by re-constituting *in vitro* chromatin [56]. Like for optical mapping, data integration follows a two-step approach, first assembly errors are identified and resolved, and then

the broken contigs are scaffolded. Integration of Dovetail Genomics read pairs into the PacBio assembly of *A. alpina* improved assembly contiguity similar well as compared to optical mapping data [34••]. Combined integration of optical mapping and Dovetail Genomics data could further advance the contiguity supporting the idea that optical mapping and Hi-C data are complementary and can help bridging different regions in the genome [34••].

The last technology highlighted here was introduced by 10X Genomics in 2015 who integrated their proprietary GemCode technology in their latest system called Chromium (www.10xgenomics.com). In this technology, diluted DNA fragments of up to 100 kb are dispersed into more than 10 000 gel bead partitions, then barcoded using unique tags and finally pooled together to perform usual Illumina short-read sequencing [19]. This is similar to Illumina's SLRs and also similar to the Long Fragment Read technology service offered by Complete Genomics (www.completegenomics.com) (which however is only available for human genomes and thus not further considered here). In contrast to Illumina's SLR protocol, however, the Chromium system can process many more and considerably longer fragments and the workflow does not require the sequencing depth necessary to assemble the individual fragments [19]. Instead reads with the same barcode (called linked reads), which are amplified from the same DNA molecules, can be used for scaffolding [57•]. If used like this, the 10X Genomics data relies on the presence of a prior sequence assembly (whereas the Illumina SLRs are the actual basis for a sequence assembly). However, in an impressive recent effort to *de novo* assemble the genomes of seven humans, the 10X Genomics linked read data were first used for whole-genome assemblies (without considering that they were sequenced from individually tagged molecules), while the barcode information was afterwards used for scaffolding of the contigs [31••]. This two-step method was implemented in a single, simple-to-use software and yielded contig N50 values of over 100 kb, which were an order of magnitude lower as compared to long-read assemblies of human genomes. However, scaffold N50 values of close to 20 Mb outperformed the contiguity of long-read assembly contigs at a fraction of the costs, while only 2–3% of the nucleotide sequences remained unresolved and despite the fact that only short-read data of a single sequencing library at modest coverage was used. So far, there are no published records of plant genomes assembled with the 10X Genomics technology, however, there is no reason that these promising results could not be achieved for a plant genome as well.

Assembly of heterozygous and polyploid genomes

Assembly of plant genomes is typically performed on inbred, homozygous individuals. While homozygous genomes can be assembled like haploids, the assembly of heterozygous individuals requires correct handling and ideally also reconstruction of the different chromosome sets. An elegant solution for this is provided by a recent algorithm, FALCON-Unzip, which uses PacBio sequencing data to phase the variation between individual chromosomes already during the assembly of the reads [26••]. The assemblies consist of haplotigs, which are contigs that represent individual chromosomes, and reached N50 values of 6.9 and 0.8 Mb when used for the *de novo* assembly of an Arabidopsis F1 hybrid and a heterozygous grapevine accession [26••]. Similar to inferring the haplotypes from long reads they can also be assembled from long-range scaffolding methods like proximity ligation-based methods [17,55,56] or long DNA fragments derived from dilution methods [18,58–60]. For example, the seven human genome sequences assembled with short reads from libraries generated with the 10X Genomics system were also assembled to haplotigs and achieved N50 values of up to 9 Mb for the assembly of individual sets of chromosomes [31••].

It seems obvious how third generation technologies combined with similar algorithms could help during the assembly of polyploid genomes as well, as local similarity between homeologous chromosomes can introduce similar challenges as heterozygosity. However, despite the increased repetitiveness of polyploid genomes, the assembly of allopolyploids (polyploids which evolved by the merger of two or more distinct species) works surprisingly well even without dedicated assembly methods [29••,61–63]. For example, the recent assembly of *Brassica juncea*, which evolved by the merge of *B. rapa* and *B. nigra* to a 922 Mb tetraploid genome, featured a scaffold N50 of 1.5 Mb. This assembly was generated with short and long-read sequencing data combined with optical maps, however, none of the algorithms used was particularly developed for polyploid genome assembly suggesting that the level of divergence between the homeologous chromosomes was sufficiently high to be assembled separately. Even the assembly of the gene space of the gigantic 17 Gb hexaploid genome of wheat including three homeologs of most of the genes was first attempted with simple whole-genome shot-gun approaches [64].

To our knowledge, there is so far no *de novo* assembly of an autoploid plant, which reconstructs each of the homeologous chromosomes separately. Instead diploid or even haploid individuals have been used for assembly of autoploid species [65]. Dependent on the level of divergence between the homeologous chromosomes, autoploid genomes could be assembled into a 'pseudo-haploid' sequence, where polymorphic sites could be annotated in subsequent steps. Alternatively, third generation technologies might allow to bridge between neighbored polymorphisms and thereby distinguish the homeologous chromosomes. Obviously, longer reads or molecules

will help improving the reconstruction of individual homeologs.

Conclusions

New third generation genomic technologies reach out to bridge the quality gap between high-quality reference sequences and low-cost short-read assemblies. Long-read assemblies on their own already outperform short-read assemblies and combination with optical mapping or other long-range data emerges as a strategy to assemble genomes at unprecedented completeness and quality. The availability of many good assemblies will question the role of species-representing reference sequences, and in particular research on non-standard lines might not need to rely on common reference sequences anymore. While first study-specific references are currently generated, there are also first algorithms that can handle multiple reference sequences (*e.g.*, [66,67]) to support simultaneous alignments against multiple references or to facilitate identification of polymorphisms within assembly graphs build from sequencing data of multiple individuals.

In practice, however, projects that involve population-scale sequencing are still commonly using Illumina's short-read technology and not long-read sequencing and thus they have to deal with all the short-comings of reference-biased resequencing even if study-specific reference sequences are generated. Interestingly though, the advent of the 10X Genomics platform has the potential to revolutionize such population-scale sequencing projects. Instead of 'assembling' short-read data using alignments against a reference sequence, the linked short read data enable *de novo* assemblies for each individual of a population with only a minor increase in costs. So far, there are hardly any methods to analyze such sets of assemblies, but this might change soon and open the door for analyses of entire genomes of large panels of plants including hybrids and polyploids in the near future. Eventually this will end the era of re-sequencing and enable new whole-genome comparisons that reveal and utilize essentially all of the differences between genomes independent of their size or complexity.

Conflict of interest

The authors declare that no conflicts of interest exist.

Acknowledgements

The work on plant genome assembly methods in the Schneeberger lab is financially supported by the Max Planck Society and the German Federal Ministry of Education and Research in the frame of RECONSTRUCT (FKZ 031B0200A-E).

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. Zapata L, Ding J, Willing E, Hartwig B, Bezdan D, Jiao W, Patel V: **Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms**. *Proc Natl Acad Sci* 2016, **113**:E4052-E4060.

2. Schatz MC, Maron LG, Stein JC, Hernandez Wences A, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E *et al.*: **Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica**. *Genome Biol* 2014, **15**:506.

3. Zhang J, Chen L-L, Xing F, Kudrna DA, Yao W, Copetti D, Mu T, Li W, Song J-M, Xie W *et al.*: **Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63**. *Proc Natl Acad Sci* 2016, **113**:E5163-E5171.

4. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L *et al.*: **De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits**. *Nat Biotechnol* 2014, **32**:1045-1052.

5. Hirsch C, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, Shem-Tov D, Baruch K, Lu F, Hernandez AG *et al.*: **Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize**. *Plant Cell* 2016, **28**:2700-2714.

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

7. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome**. *Nature* 2005, **436**:793-800.

8. 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.

9. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD *et al.*: **Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies**. *Genome Biol* 2014, **15**:R59.

10. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A *et al.*: **The Norway spruce genome sequence and conifer genome evolution**. *Nature* 2013, **497**:579-584.

11. Salman-Minkov A, Sabath N, Mayrose I: **Whole-genome duplication as a key factor in crop domestication**. *Nat Plants* 2016, **2**:16115.

12. Schatz MC, Witkowski J, McCombie WR: **Current challenges in de novo plant genome sequencing and assembly**. *Genome Biol* 2012, **13**:243.

13. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al.*: **Real-time DNA sequencing from single polymerase molecules**. *Science* 2009, **323**:133-138.

14. Deamer D, Akeson M, Branton D: **Three decades of nanopore sequencing**. *Nat Biotechnol* 2016, **34**:518-524.

15. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS: **Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements**. *PLoS One* 2014, **9**:e106689.

16. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK: **Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping**. *Science* 1993, **262**:110-114.

17. 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.

18. Amini S, Pushkarev D, Christiansen L, Kostem E, Royce T, Turk C, Pignatelli N, Adey A, Kitzman JO, Vijayan K *et al.*: **Haplotype-resolved whole-genome sequencing by contiguity-preserving**

transposition and combinatorial indexing. *Nat Genet* 2014, **46**:1343-1349.

19. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM *et al.*: **Haplotyping germline and cancer genomes with high-throughput linked-read sequencing**. *Nat Biotechnol* 2016, **34**:303-311.

20. •• Koren S, Phillippy AM: **One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly**. *Curr Opin Microbiol* 2015, **23**:110-120.
A overview about long-read sequencing technologies and genome assembly algorithms for long read data.

21. Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin C-S, Rapicavoli NA, Rank DR, Li J *et al.*: **Long-read, whole-genome shotgun sequence data for five model organisms**. *Sci Data* 2014, **1**:140045.

22. • 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 first plant genome sequence assembled using PacBio long reads and BioNano optical maps.

23. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED *et al.*: **Hybrid error correction and de novo assembly of single-molecule sequencing reads**. *Nat Biotechnol* 2012, **30**:693-700.

24. Bashir A, Klammer AA, Robins WP, Chin C-S, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P *et al.*: **A hybrid approach for the automated finishing of bacterial genomes**. *Nat Biotechnol* 2012, **30**:701-707.

25. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE *et al.*: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data**. *Nat Methods* 2013, **10**:563-569.

26. •• Chin C, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Dunn C, Malley RO, Figueroa-balderas R, Morales-cruz A, Grant R *et al.*: **Phased diploid genome assembly with single molecule real-time sequencing**. *Nat Methods* 2016, **13**:1050-1054.
This paper introduced a phased diploid genome assembler called FALCON-unzip based on SMRT sequencing reads. The authors applied it for the assembly of diploid genomes including an Arabidopsis F1 hybrid and a heterozygous grape genome.

27. • Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM: **Assembling large genomes with single-molecule sequencing and locality-sensitive hashing**. *Nat Biotechnol* 2015, **33**:623-630.
An efficent method MHAP was introduced to process overlapping noisy, long reads using probabilistic, locality-sensitive hashing enabling assemble of large eukaryotic genomes.

28. Bombarely A, Moser M, Amrad A, Bao M, Bapaume L, Barry C, Bliek M, Boersma M, Borghi L, Bruggmann R *et al.*: **Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida***. *Nat Plants* 2016, **2**:16074.

29. •• Yang J, Liu D, Wang X, Ji C, Cheng F: **The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection**. *Nat Genet* 2016, **48**:1225-1232.
By combining sequencing reads from multiple Illumina paired-end libraries and PacBio SMRT, and BioNano optical maps, a high-contiguity genome assembly of the complicated allopolyploid genome of *Brassica juncea* was obtained.

30. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC *et al.*: **Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology**. *PLoS One* 2012, **7**:e47768.

31. •• Weisenfeld NI, Kumar V, Shah P, Church D, Jaffe DB: **Direct determination of diploid genome sequences**. *bioRxiv* 2016 http://dx.doi.org/10.1101/070425.
The authors developed a new algorithm Supernovo to utilize 10X Genomics data for *de novo* assembly of seven human genomes without providing primary assembly contigs, which could reduce sequence costs and improve assembly contiguity simultaneously.

32. Hoshino A, Jayakumar V, Nitasaka E, Toyoda A, Noguchi H, Itoh T, Shin-I T, Minakuchi Y, Koda Y, Nagano AJ *et al.*: **Genome sequence and analysis of the Japanese morning glory Ipomoea nil**. *Nat Commun* 2016, **7**:13295.

33. Willing E-M, Rawat V, Mandáková T, Maumus F, James GV, Nordström KJV, Becker C, Warthmann N, Chica C, Szarzynska B *et al.*: **Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation**. *Nat Plants* 2015, **1**:14023.

34. •• Jiao W-B, Garcia Accinelli G, Hartwig B, Kiefer C, Baker D, Severing E, Willing E-M, Piednoel M, Woetzel S, Madrid-Herrero E *et al.*: **Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data**. *Genome Res* 2017 http://dx.doi.org/10.1101/gr.213652.116.
The authors developed different scaffolding workflows using optical mapping and chromosome conformation capture data, which not only improve assembly contiguity but also help correcting misassemblies within PacBio assemblies.

35. Quick J, Quinlan AR, Loman NJ: **A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer**. *Gigascience* 2014, **3**:22.

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

37. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR: **Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome**. *Genome Res* 2015, **25**:1750-1756.

38. Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ *et al.*: **The genome sequence of the colonial chordate, *Botryllus schlosseri***. *Elife* 2013, **2013**:1-24.

39. Li R, Hsieh C-L, Young A, Zhang Z, Ren X, Zhao Z: **Illumina synthetic long read sequencing allows recovery of missing sequences even in the finished *C. elegans* genome**. *Sci Rep* 2015, **5**:10814.

40. Bankevich A, Pevzner PA: **TruSPAdes: barcode assembly of TruSeq synthetic long reads**. *Nat Methods* 2016, **13**:248-250.

41. Roach JC, Boysen C, Wang K, Hood L: **Pairwise end sequencing: a unified approach to genomic mapping and sequencing**. *Genomics* 1995, **26**:345-353.

42. Nagarajan N, Pop M: **Sequence assembly demystified**. *Nat Rev Genet* 2013, **14**:157-167.

43. 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.

44. Valouev A, Schwartz DC, Zhou S, Waterman MS: **An algorithm for assembly of ordered restriction maps from single DNA molecules**. *Proc Natl Acad Sci* 2006, **103**:15770-15775.

45. Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y *et al.*: **Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions**. *Cell* 2016, **166**:492-505.

46. Chamala S, Chanderbali AS, Der JP, Lan T, Walts B, Albert VA, dePamphilis CW, Leebens-Mack J, Rounsley S, Schuster SC *et al.*: **Assembly and validation of the genome of the nonmodel basal angiosperm *Amborella***. *Science* 2013, **342**:1516-1517.

47. Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, Gentzbittel L, Childs KL, Yandell M, Gundlach H *et al.*: **An improved genome release (version Mt4.0) for the model legume *Medicago truncatula***. *BMC Genom* 2014, **15**:312.

48. Nagarajan N, Read TD, Pop M: **Scaffolding and validation of bacterial genome assemblies using optical restriction maps**. *Bioinformatics* 2008, **24**:1229-1235.

49. Tang H, Lyons E, Town CD: **Optical mapping in plant comparative genomics**. *Gigascience* 2015, **4**:3.

50. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A *et al.*: **Assembly and diploid architecture of an individual human genome via single-molecule technologies**. *Nat Methods* 2015, **12**:780-786.

51. Zhou S, Bechner MC, Place M, Churas CP, Pape L, Leong SA, Runnheim R, Forrest DK, Goldstein S, Livny M *et al.*: **Validation of rice genome sequence by optical mapping**. *BMC Genom* 2007, **8**:278.

52. Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, Cao H, Kwok PY, Deal KR, Dvorak J *et al.*: **Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex *Aegilops tauschii* genome**. *PLoS One* 2013, **8**:e55864.

53. Lin HC, Goldstein S, Mendelowitz L, Zhou S, Wetzel J, Schwartz DC, Pop M: **AGORA: assembly guided by optical restriction alignment**. *BMC Bioinform* 2012, **13**:189.

54. Lieberman-aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al.*: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science* 2009, **326**:289-293.

55. Selvaraj S, Dixon JR, Bansal V, Ren B: **Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing**. *Nat Biotechnol* 2013, **31**:1111-1118.

56. Putnam NH, Connell BO, Stites JC, Rice BJ, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS: **Chromosome-scale shotgun assembly using an in vitro method for long-range linkage**. *Genome Res* 2016, **26**:342-350.

57. • Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Dakula *et al.*: **A hybrid approach for de novo human genome sequence assembly and phasing**. *Nat Methods* 2016, **13**:587-590.

This was the first study combining 10X Genomics data with Illumina short reads and BioNano optical maps to assemble a large human genome.

58. Kaper F, Swamy S, Klotzle B, Munchel S, Cottrell J, Bibikova M, Chuang H-Y, Kruglyak S, Ronaghi M, Eberle MA *et al.*: **Whole-genome haplotyping by dilution, amplification, and sequencing**. *Proc Natl Acad Sci* 2013, **110**:5552-5557.

59. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE *et al.*: **Haplotype-resolved genome sequencing of a Gujarati Indian individual**. *Nat Biotechnol* 2011, **29**:59-63.

60. Snyder MW, Adey A, Kitzman JO, Shendure J: **Haplotype-resolved genome sequencing: experimental methods and applications**. *Nat Rev Genet* 2015, **16**:344-358.

61. Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B *et al.*: **Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome**. *Science* 2014, **345**:950-953.

62. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM *et al.*: **Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement**. *Nat Biotechnol* 2015, **33**:531-537.

63. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J *et al.*: **Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution**. *Nat Biotechnol* 2015, **33**:524-530.

64. Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D *et al.*: **Analysis of the bread wheat genome using whole-genome shotgun sequencing**. *Nature* 2012, **491**:705-710.

65. Potato Genome Consortium, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R *et al.*: **Genome sequence and analysis of the tuber crop potato**. *Nature* 2011, **475**:189-195.

66. Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, Weigel D: **Simultaneous alignment of short reads against multiple genomes**. *Genome Biol* 2009, **10**: R98.

67. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G: **De novo assembly and genotyping of variants using colored de Bruijn graphs**. *Nat Genet* 2012, **44**:226-232.

68. Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Falin LJ, Grabmueller C *et al.*: **Ensembl Genomes 2016: more genomes, more complexity**. *Nucleic Acids Res* 2016, **44**:D574-D580.