Chapter 3

Strategies for Sequence Assembly of Plant Genomes

Stéphane Deschamps and Victor Llaca

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61927

Abstract

The field of plant genome assembly has greatly benefited from the development and widespread adoption of next-generation DNA sequencing platforms. Very high sequencing throughputs and low costs per nucleotide have considerably reduced the technical and budgetary constraints associated with early assembly projects done primarily with a traditional Sanger-based approach. Those improvements led to a sharp increase in the number of plant genomes being sequenced, including large and complex genomes of economically important crops. Although next-generation DNA sequencing has considerably improved our understanding of the overall structure and dynamics of many plant genomes, severe limitations still remain because next-generation DNA sequencing reads typically are shorter than Sanger reads. In addition, the software tools used to *de novo* assemble sequences are not necessarily designed to optimize the use of short reads. These cause challenges, common to many plant species with large genome sizes, high repeat contents, polyploidy and genome-wide duplications. This chapter provides an overview of historical and current methods used to sequence and assemble plant genomes, along with new solutions offered by the emergence of technologies such as single molecule sequencing and optical mapping to address the limitations of current sequence assemblies.

Keywords: Sequencing, Plant, Genome, Assembly

1. Introduction

Genome sequencing, assembling and annotation have been major priorities in plant genetics research during the past 20 years. The release of draft reference genomes have typically constituted major milestones and have proven to be invaluable for the analysis and characterization of genome architecture, genes and their expression, diversity and evolution [1–5]. The expansion of sequence information in a growing number of taxa has contributed to comparative studies and the implementation of molecular breeding and biotechnology approaches for crop improvement [6, 7]. The construction of the first plant genomes was made possible by applying considerable resources, coordination and effort to enabling automated Sanger-based sequencing technologies and computational algorithms. Starting in 2005, a series of technological revolutions in DNA sequencing, driven in large part by the goal of affordable personalized genome sequencing, radically changed the sequencing model. First, new technologies drastically increased throughput while reducing costs and times in data collection. Additional technologies then enabled long single-molecule reads and algorithms that were more suitable to resolve complex genomes [8, 9].

In addition to these advances, the genomics community has benefited from the development and implementation of complementary mapping technologies and methods that have facilitated the scaffolding of sequences and integration to genetic maps. This review provides a historical and technical perspective of methods and technologies applied to genome reference assembly in plants as well as current advances and future directions.

2. The development of Sanger sequencing for *de novo* assembly of plant genomes

The construction of reference genomes was initially enabled by technological advances in sequencing using the Sanger method [10]. During the 1980s and 1990s, the introduction of thermal cycle sequencing, single-tube reactions and fluorescence-tagged terminator chemistry [11] facilitated the development of high-capacity sequencing platforms. Additional improvements in parallelization, base quality assessment, read length and cost-effectiveness were achieved by the development of automatic base-calling and capillary electrophoresis [12, 13]. With no major modifications made in the past years, automated high-throughput Sanger sequencing is performed by parallel reactions that include a mixture of the DNA template, primer, DNA polymerase, and deoxynucleotides (dNTPs). A proportion of dideoxynucleotide terminators (ddNTP) are included in the reaction, each labelled with a different fluorescent dye. DNA molecules are extended from templates using a thermal cycling reaction and terminated by random incorporation of the labelled ddNTPs, which are detected by laser excitation of the fluorescent labels after capillary-based electrophoresis. The differences in dye excitation profiles are recorded and translated by a computer to generate the sequence. Primary analysis software then calls nucleotides from the raw sequences, assigning a corresponding quality score at each position [6, 14].

The complete sequencing of the first bacterial genomes [15,16] as well as the creation of initiatives aimed at sequencing the genomes of *Sacharomyces cerevisae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens* provided the technical and technological framework for the initial sequencing of genomes in plants [17–21]. These projects validated the idea of applying a scaled-up form of shotgun sequencing [22]. Shotgun sequencing relied on computer algorithms to enable *in silico* assembly of overlapping sequencing reads derived from randomly-generated subclones. The development of software suites such as Phred, Phrap and Consed [23] allowed calling bases, setting individual base quality, assembling overlapping reads, assigning assembly quality scores, viewing final assemblies and extracting consensus sequences. Two major genomic shotgun sequencing strategies were defined at that time: (1) whole-genome shotgun sequencing (WGS) and (2) clone-by-clone, also referred to as BAC-by-

BAC sequencing. In WGS, genomic DNA is randomly sheared and the ends of the cloned fragments are directly sequenced and assembled. This strategy is the simplest, and it was initially used in small bacterial and yeast genomes. Later, it was also used in *D. melanogaster* and one of two initiatives aimed at sequencing and assembling the reference human genome [19, 21]. Major improvements to *de novo* WGS assembly came from using strategies that relied on paired-end reads from multiple libraries with different average insert sizes and the optimization of software with algorithms that use end-sequence distance information from these libraries.

The second Sanger sequence assembly strategy, clone-by-clone, was successfully deployed in projects aimed at complex eukaryotic genomes. In clone-by-clone genome assembly, shotgun sequencing is performed in libraries derived from individual genomic large-insert clones, selected in a minimum tile path according to physical and genetic map information [24, 25]. The most common type of large-insert clone is the bacterial artificial chromosome (BAC), which can stably carry genomic inserts ranging from 100 to 300 kb and is relatively easy to maintain and purify. Accordingly, this method is usually referred to as BAC-by-BAC, although additional vector systems have been used in assembly projects, including yeast artificial chromosomes (YACs), P1 artificial chromosomes (PAC), transformation-competent artificial chromosomes (TACs), cosmids and fosmids. The two major genomic shotgun-sequencing approaches, WGS and BAC-by-BAC, had advantages and disadvantages when applied to Sanger-based sequencing platforms, depending on the genome of interest. The clone-by-clone approach benefited from working in small units, effectively reducing complexity and computational requirements. This approach minimized problems associated with the misassembly of highly repetitive DNA and therefore provided a better, more complete assembly in plants and other complex eukaryotic genomes. WGS projects were computationally intensive and were less effective bridging repetitive regions in complex genomes but benefited from considerably lower cost, time and logistics [14].

The first completed reference plant genome was from the model system *Arabidopsis thaliana*, accession Columbia [26]. At that time, it was only the third multicellular eukaryotic genome to be published, after *C. elegans* and *D. melanogaster*. The nuclear genome of Arabidopsis is distributed in five chromosomes, and it is only approximately 4% the size of the human genome. The *A. thaliana* genome initiative used multiple types of available large-insert libraries including cosmids, BACs, PACs and TACs. Shotgun clones were constructed and then mapped by restriction fragment fingerprinting as well as screening with hybridization or polymerase chain reaction (PCR) markers. End sequences for 47,788 BAC clones were further used to anchor clones, integrate contigs and help select a minimum tiling path. Each of 1,569 clones in a minimum tiling path were selected, sequenced bidirectionally and assembled at estimated error rates of less than 1 in 10,000 bases. Direct PCR products were used to close some gaps and YACs allowed the characterization of telomere sequences. As initially published, the total length of sequenced regions was 115.4 Mb, in addition to an estimated 10 Mb nonsequenced centromeric and rDNA repeat regions. Since the original publication, the Arabidopsis genome sequence reference has been subjected to several rounds of improvements, each time reducing gaps and extending the sequence towards the centromeric regions [27].

The second published plant genome reference was rice (*Oryza sativa*). While the rice genome is more than 2-fold the size of Arabidopsis, approximately 390 Mb, it is one of the smallest genomes of any major crop, less than 15% the size of the human genome. Like Arabidopsis, the rice genome was completed using a Sanger-only clone-by-clone approach [28] that required the initial construction, fingerprinting and physical mapping of a large number of random BACs and PACs. In total, 3,401 mapped clones in a minimum tiling path were selected from the physical map, randomly sheared and individually end-sequenced to approximately 10-fold coverage. Clone sequences were assembled and low-quality regions were finished using targeted sequencing. Gaps were closed and low-quality regions resolved by sequencing PCR fragments, plasmids and fosmids.

The draft reference genome of Maize, one of the most important crops in the world, is considered the last major published plant genome project based primarily on a Sanger BAC-by-BAC strategy [29]. At 2.3 Gb and spanning 10 chromosomes, the nuclear genome of maize is considerably larger than that of rice and Arabidopsis, approximately 3/4 the size of the human genome. A set of 16,848 minimally overlapping BAC clones, derived from an integrated physical and genetic map, were selected and end-sequenced. The assembly was performed after adding additional data derived from cDNA sequences and sequences from subtractive libraries with methyl-filtered DNA and high C_0t techniques, resulting in a whole-genome assembly (B73 RefGen_v1) made of 2,048 Mb in 125,325 sequence contigs and 61,161 scaffolds [29]. Unlike the completed genomes of rice and Arabidopsis, most sequenced BACs in the first version of the maize draft genome are unfinished. Gaps and low-quality regions in BACs were not systematically closed by PCR sequencing or other target approaches. Therefore, while the BACs used in the minimum tiling path were mapped, the order and orientation of individual contigs within a single BAC could be incorrect. Subsequent versions of the genome have been improved by targeting gaps and adding alternative sequencing strategies described later in this review.

Finally, it is important to mention that a significant number of plant genome sequencing initiatives have used WGS strategies, which provide a considerable reduction in time and cost associated with cloning, construction, mapping and selection. Sanger WGS genome projects included those of poplar tree, grape, and papaya [30–32]. Later refinements to the process enabled the sequencing of *Brachypodium distachyon* [33] as well as the larger genomes of *Sorghum bicolor* (~730 Mb) [34] and soybean, an ancestral tetraploid (1.1 Mb) [35]. It should be noticed that, as demonstrated by the Maize genome project, the two Sanger shotgun assembly approaches, as well as later sequence technologies, are not mutually exclusive and may be complementary to increase quality and coverage.

The high cost and logistics of plant projects based on clone-by-clone Sanger sequencing required extensive funding, the creation of large collaborative consortia and several years of fingerprinting and sequencing work. The cost of the project by the Arabidopsis Genome Initiative has been estimated at US$70 million [36]. The International Rice Genome Sequencing Project (IRGSP), which included groups from 11 different nations, took over 5 years to complete. During its early stages, IRGSP had estimated that the project would take 10 years and cost a staggering US$200 million [37]. The Maize draft genome was accomplished by multiple laboratories at an estimated cost of tens of millions in a joint NSF/DOE/USDA program. It is worth noticing that, while the cost and time required to accomplish Sanger WGS projects are in fact lower than those based on a clone-by-clone approach, they are still considerable for today's standards. The sequencing of the 1.1-Gb soybean genome, the largest

http://dx.doi.org/10.5772/61927

published plant genome based on a Sanger WGS approach, provides an example of such a cost. It was completed in less than two years although it took a group of 18 institutions several million dollars to generate and assemble more than 15 million Sanger reads from multiple libraries with average sizes ranging from 3.3 kb to 135 kb [35].

Besides cost and time considerations, these early Sanger-only projects posed considerable technical challenges. Despite the extensive resources deployed towards the sequencing of the Arabidopsis and rice genomes, which are usually considered as finished, as well as other projects mentioned in this review, they all have representation gaps. A considerable number of gaps correspond to regions that are "unclonable" under the conditions used to prepare BAC and other genomic libraries. Although many of these regions correspond to tandem repeats such as telomeric sequences and other repetitive regions, it may also include gene space [29]. Moreover, the maximum length of quality Sanger reads, usually 800–900 bp, as well as technical issues associated with the sequencing of DNA stretches with strong secondary structures or extensive homopolymers, create conditions for additional sequencing gaps, even in regions with physical coverage.

Finally, most plant genomes are characterized by elevated proportions of highly repetitive DNA and by the presence of segmental duplications or full genome duplications due to polyploidization events [38], which can be problematic during assembly. The 1C genome content in Maize, for example, is smaller than in humans but it consists of higher proportions and larger tracks of high-copy elements such as retrotransposable elements [29, 38]. At least some of the differences between the assembled and estimated genomes of the Maize line B73 could be attributed to the assembly-based collapse of highly similar long terminal repeats (LTRs) at the end of retrotransposons. It is important to note that all the Sanger-only initiatives corresponded to plant species with genomes that were considerably smaller than the average 5.8-Gb plant genome. Plant genomes have a considerably wider size range than in mammals, and in some important crops (e.g. wheat), nuclear genomes can be more than 15 Gb long, well beyond the practical realm of Sanger sequencing. Although BAC-by-BAC approaches can reduce complexity by more than 10,000 fold, Sanger-based assembly remains difficult and prohibitively expensive in plant genomes of moderate or large size. The WGS approach is even more sensitive to the complexity of plant genomes as it increases the potential for assembly artefacts due to haplotype and homeolog collapse in regions with high identity. Reductions in time and cost in WGS projects are achieved at the expense of assembly fidelity in repetitive regions and expanded need for computational resources.

3. Next-generation sequencing technologies applied to *de novo* assembly of plant genomes

3.1. Second-generation sequencing technologies

As indicated above, successful whole-genome sequencing projects have been achieved with the use of Sanger technology. However, such projects require dealing with several complicating factors, including high costs and relatively long turnaround times to completion. The emergence of next-generation sequencing (NGS) technologies has changed this paradigm, both by reducing costs and increasing sequencing throughputs, while at the same time introducing complexity related to the relative short reads of NGS reads. Several NGS technologies have emerged in the past 7 to 8 years [for reviews, see refs. 39–41]. All follow a relatively uniform approach to library construction and sequencing. To complete sequencing: (1) universal adapters are ligated at the end of single DNA molecule templates; (2) adapter-ligated DNA templates are amplified via PCR to create a cluster of identical isoforms and (3) clusters are loaded on sequencers and nucleotide incorporations occur in parallel on millions of clusters. These generate an amplified signal that is recognized by the platform and translated into a base call.

The most widely used NGS technology nowadays is the one commercialized by Illumina [42], whose high-throughput instrument, the HiSeq4000, can produce up to 1.5 Tb of sequencing data in approximately 3.5 days.. In the Illumina sequencing platform, sequencing templates generated during library construction are immobilized on a solid surface, and a "bridge PCR" approach allows for the localized amplification of millions of single DNA molecules, thus generating millions of clusters, each containing thousands of copies of the original DNA molecules [43]. Sequencing then is performed using a sequencing-by-synthesis approach where single-base extension allows the incorporation of a fluorescently labelled nucleotide (a blocking chemical moiety at the 3′ hydroxyl end allows the incorporation of one base only). Once incorporated, the label is detected and the resulting signal subsequently translated into a base call. Finally, the fluorescent dye and the blocking 3′ agent are cleaved, allowing the next single base incorporation event to occur. Through the use of alternating cycles of base incorporation, image capture and dye cleavage, the Illumina sequencing technology can produce reads that are up to 300 bp in length. The relatively high error rate (~0.1% or 10 times higher than Sanger sequencing) [39] can be compensated by very high sequencing coverage, thus allowing random errors at any given base position to be ignored below a certain frequency threshold. The relative short read of Illumina sequencing reads can be explained by several noise factors accumulating after each cycle, including phasing, where imperfect single-base incorporation and imperfect cleavage of the dye and 3′ hydroxyl blocking moiety lead to the accumulation of copies of various lengths within a cluster, and the subsequent increase of signal-to-noise ratio after each cycle [44].

3.2. Third-generation sequencing technologies

De novo assemblies of plant genomes have been performed with NGS reads only, either with reads generated on the Illumina platform alone or with reads generated with the Illumina platform combined with reads generated on the Roche 454 second-generation sequencing platform [45]. However, those assemblies generally are fragmented, resulting in low N50 values and a high number of contigs, mostly because of the overall short read length, the complexity of the genome and the presence of conserved regions whose length exceeds the length of NGS reads and thus cannot be extended during the *de novo* assembly process. The emergence of third-generation sequencing technologies [46, 47] has started to address some of the inherent limitations of sequencing and assembling large and complex plant genomes. Those technologies are characterized by the parallel sequencing of single molecules of DNA (rather than "clusters"), thus avoiding phasing issues, and the resulting sequences tend to be in the kb range, offering the opportunity to assemble genomes and generating longer contigs by encompassing complex and conserved genomic regions and allowing relatively high-confidence assemblies of overlapping reads. However, single sequencing reads tend to exhibit relatively high error rates (~15%–25% on average). Deep sequencing coverage or repeated sequencing of the same DNA fragments therefore are required to offset the presence of a high number of sequencing errors [48, 49]. As of today, two companies have developed and commercialized third-generation sequencing technologies, namely, Pacific Biosciences [e.g., 50] (Menlo Park, CA) and Oxford Nanopore Technologies [e.g., 51] (Oxford, UK). Each company uses vastly different approaches to sequencing. The Pacific Biosciences (PacBio) RS II system uses a sequencing-by-synthesis approach to offer up to ~40-kb reads, where base incorporation is monitored in a real-time fashion. Nanoscale holes, described as Zero Mode Waveguides ("ZMW") are located on a chip, where individual polymerases are covalently attached to the surface of each ZMW. Individual nucleotides with a fluorescent label attached to the phosphate chain are incorporated to the elongating strand and the excited dye emits a signal that is captured before diffusion of the released pyrophosphate, and translated into a specific base call. DNA fragments used as template are ligated to "bell-shaped" adapters at both ends, thus facilitating the sequencing of DNA fragments through multiple passes and the creation of a more accurate consensus sequence. The overall stability and activity of the polymerase remain limited by photo damage and the progressive dissociation of the polymerase/template complex from the surface of the ZMW. It is therefore expected that reads generated from smaller DNA fragments will exhibit higher consensus accuracy than reads from larger DNA fragments. Oxford Nanopore Technologies released the MinION sequencing device in early access mode in 2014. Like the PacBio RS II system, the MinION delivers long reads in a real-time fashion, from single molecules of DNA. In that particular case, however, sequencing is performed by measuring the change in ionic current when a single DNA strand translocates through a protein nanopore located in an insulated membrane. The resulting signal is measured and translated into a base call. Because no enzyme is involved in the DNA sequencing process, it is expected that read length will be driven mostly by the physical length of the DNA strand being sequenced. Library construction involves the ligation of two types of adapters to DNA fragment, one "Y-shaped" adapter with a bound protein that unwinds the double-stranded DNA and facilitates the translocation of a single strand through the pore, and one "bell-shaped" adapter at the other end that allows the translocation, and sequencing, of both the sense and antisense strands. Sequencing reads then are generated by aligning base calls from the two strands and producing a higher quality consensus sequence.

3.3. Challenges in assembling plant genomes

De novo assembly of genomes has closely mimicked the trends and improvements in sequencing technologies and accompanying sequencing assembly software over the years [45]. The emergence of next-generation sequencing technologies has allowed a much larger number of plant genomes to be sequenced and assembled than what would have been deemed possible with Sanger sequencing alone, mostly because of the costs and labor involved in such projects. However, the complexity of the majority of those genomes still makes it a challenge to resolve them with short reads alone [52, 53]. As a result, most plant genome assemblies are highly fragmented, with large number of contigs and conserved regions of the genome in an unfinished state [54]. The presence of highly conserved repeats often exceeding 10 kb in length represents a major challenge in assembling plant genomes. The most common types of repeats in plants are type II long-terminal repeat (LTR) retrotransposons and their proliferation within a genome often explains most of the structural variations between strains [55]. Their movement also results in genome expansion, where repeats represent, in some instances, more than 80–90% of the structural content of a particular genome [29]. Repeat expansion also can lead to very large genome sizes. While NGS technologies can generate enough raw data to cover an entire genome in a relatively cost-effective manner, assembling such a large amount of data often represents a major computational challenge. For example, the assembly of the loblolly pine genome (~22 Gb), which represents the largest genome assembled to date, could be solved only using condensed sets and read pooling prior to assembly [56]. Assembling large and repeat-rich genomes can also be facilitated by using supplemental layers of information, such as the physical distance between "paired" reads (end-sequences generated at both ends of a particular DNA fragment) in mate-pair libraries. Another challenge for *de novo* assembly of plant genome is the issue of polyploidy [57]. Polyploidy is an important force in plant genome evolution and it is estimated that ~80% of all living plants are polyploids [58], while close to 100% of all plant lineages have a paleo-polyploidy event in their history. As a consequence, some plants species, including economically important crop species like soybean [35], have entire gene families consisting of highly similar paralogs. Those gene families are the direct result of paleo-polyploidization events where the merger of genomes has been followed by extensive structural rearrangements, including gene loss, and the modification of gene expression for paralogs within a particular gene family. The diploid genomes of progenitor species can be used to determine the origin and structure of contigs when assembling large polyploid genomes [59]. Finally, heterozygosity may represent another important challenge when assembling plant genomes. Outcrossing species like grape, for instance, exhibit up to 13% sequence divergence between alleles, and the existence of such variation will impact contig assembly when both alleles are sequenced in a whole-genome assembly project [31].

3.4. Examples of plant genome assemblies

According to Michael and Van Buren [45], over 100 plants genomes have been sequenced since 2000, out of which 63% are genomes from various crop species. As indicated above, different Sanger sequencing strategies have been applied with varying degrees of success on several plant genomes. However, the most successful Sanger-based genome assemblies have been obtained from relatively small genomes (Arabidopsis, rice), while *de novo* assemblies for larger and complex genomes, such as maize, remains partial and unfinished (manual improvements of the maize genome were limited to nonrepetitive regions only). In addition, due to the high costs and labor associated with such approaches, and the need for (in most cases) an international consortium to complete such projects, a vast majority of the most recent genomes have been sequenced using either a hybrid approach, complementing Sanger sequencing with NGS data, or using NGS data alone, from various NGS platforms. Such platforms include Illumina, 454/Roche, and more recently, Pacific Biosciences.

The domesticated tomato genome [60] represents an example of Sanger/NGS hybrid genome assembly. A total of 30,800 BAC clones from three different BAC libraries were shotgun-sequenced and end-sequenced, generating a total of 3.3 Gb of Sanger reads. In addition, 454/Roche shotgun and mate-pair sequencing was performed, both on BAC pools and whole-genome DNA preparation, using different insert sizes and generating a total of 21 Gb of NGS data. The *de novo* assembly of Sanger and 454 data was performed using the Newbler assembly software [61] and other sequence assembly and alignment tools. Further scaffolding and polishing of the assembly were performed when integrating BAC end-sequence data and additional high-coverage Illumina and ABI/SOLiD data. Taken together, the *de novo* assembly resulted in 3,761 scaffolds totalling to 782 Mb, with 95% of the assembled scaffold sequences present in 225 scaffolds. The predicted tomato genome size is approximately 900 Mb. The correctness and integrity of the assembly were validated through different means including the alignment of clone end-sequences, publicly available tomato EST sequences, and alignment of BAC contigs from a sequence-based physical BAC map. Interestingly comparison of the tomato, potato and grape genomes supported the existence of two successive whole-genome triplication events in common ancestors that added new gene family members that mediate important fruit functions, such as enzymes involved in ethylene biosynthesis (examples of whole genome duplication or triplication events abound among plant genomes that have been sequenced to date).

Because of the relatively cheap costs involved, a large number of plant genomes have been sequenced and assembled using NGS technologies alone. This includes the assembly of the complex tetraploid genome of cultivated cotton (*Gossypium arboreum*) [62]. The tetraploid cultivated cotton genome has a genome size of approximately 1.7 Gb and is thought to have appeared 1–2 million years ago through interspecific hybridization between diploid A (*Gossypium arboretum*) and D (*Gossypium raimondii*) subgenome progenitors. A total of 371.5 Gb of shotgun Illumina data was generated with various insert sizes ranging from 180 bp to 40 kb and complemented with 33,454 BAC end sequences. The assembly was performed with SOAPdenovo [63], which resulted in 40,381 contigs, anchored and oriented in 7,914 scaffolds, ranging in length from 140 kb to 5.9 Mb with 90% of the contigs included in 3,740 scaffolds.

An example of a smaller, relatively less complex genome assembly is that of the crop species *Brassica rapa* [64]. An estimated 72× sequencing coverage of the genome was generated, corresponding to Illumina shotgun paired-end data from NGS libraries with insert sizes ranging from 200 bp to 10 kb, and assembled using SOAPdenovo [63]. The resulting assembly was made of 14,207 contigs larger than 2 kb, further assembled into 794 scaffolds, totalling approximately 283.8 Mb and estimated to cover more than 98% of the gene space, based on alignments of 214,425 *B. rapa* public EST sequences and 52,712 unigenes from the BrGP database [65]. Further assessment of the integrity of the assembly was performed by aligning BAC clone Sanger sequences reported in previous studies.

While a large number of genomes have been sequenced with NGS technologies alone, the relatively short reads of the major NGS platforms that have been used in those assembly projects, combined with the general complexity of most of those genomes, generally require the use of alternative methods to facilitate the assembly or confirm its integrity. These methods rely on the use of various types of NGS libraries, such as mate-pair large inserts, or the use of Sanger-derived sequencing data such as EST or BAC-based shotgun reads. However, scaffolding of NGS contigs, based on using pairing information between NGS reads originating from the same DNA fragment, generally leads to unresolved gaps between contigs, often due to the presence of large repeat regions whose size exceed the length and resolution of short NGS reads. As a result, significant portions of any given scaffold contain large batches of unknown sequences, and of unknown length. To address these issues and improve plant genome assemblies, researchers have developed a series of multifaceted solutions, combining alignment to known public data, such as ESTs or BAC ends, or, when available, reference genomes from related species, integration of physical and genetic map data, or new technologies. Some of these approaches have been described in the next chapter.

4. Complementary approaches to *de novo* assembly of plant genomes

4.1. Long-read assembly

NGS assembly strategies based on the use of short reads cannot solve long and identical transposable elements abundantly present in most plant genomes. The use of long reads is expected to address some of those shortcomings and improve the overall quality of *de novo* assembly by ordering contigs, closing gaps, and improving scaffolding. As a consequence, researchers have started to adopt the single-molecule long-read sequencing technology from Pacific Biosciences in plant genome assembling projects. Spinach is an example of such genome assembly efforts. Spinach is a diploid species with a genome size estimated at 989 Mb. Van Deynze *et al*. [66] sequenced and assembled the Spinach genome using large fragment libraries of Pacific Biosciences sequence reads. They generated a 60× coverage of the genome, with 20% of the reads larger than 20 kb. Data were assembled using PacBio's Hierarchical Genome Assembly Process (HGAP) [67], which showed that long-read assemblies exhibited a 63-fold improvement in contig size over an Illumina-only assembly, derived from multiple Illumina libraries.

A distinct strategy to long-read assembly, namely, the Illumina TruSeq Synthetic Long-Read (SLR) strategy [68], is also expected to improve the quality of assemblies generated with short reads only. In SLR libraries, genomic DNA is fragmented to ~10 kb and individual indexed Illumina libraries are generated in parallel from highly diluted pools of sheared DNA fragments. After Illumina sequencing and data deconvolution, the original ~10 kb fragments can be reassembled, effectively reducing the complexity level of the assembly and generating very-high quality synthetic long reads that can subsequently be assembled together or used for haplotype resolution.

The use of long reads in *de novo* assembly is bound to become more prevalent in the near future, reducing the number of scaffolds while at the same time increasing their average length. The use of PacBio in smaller genomes, such as microbial genomes, has already demonstrated that the assemblies often result in contigs corresponding in most cases to individual chromosomes or plasmids present in the microbial cells. Likewise, it is likely that future plant studies will include such long reads, either alone or in combination with short-read NGS data to improve assembly and coverage in questionable regions, and to confirm the integrity of the assembly in a manner similar to Sanger data with current NGS assemblies.

4.2. Genetic anchoring

The emergence of NGS technologies has rapidly led researchers to develop methods and assays for variant discovery in various plant genomes. Some studies have shown that Single nucleotide polymorphisms (SNPs) can be discovered in parental inbred lines using next-generation sequencing [69]. Entire mapping populations also have been simultaneously sequenced and genotyped, in a process known as "genotyping-by-sequencing" (GBS) [70, 71], discovering in the process extensive lists of segregating markers within the mapped population [72, 73], that can be completed by using known reference maps or sequences to impute missing marker data from individual haplotypes. Various reduced-representation methods have been employed for NGS-derived SNP discovery in plant species where whole-genome shotgun sequencing still remains too expensive for sequencing more than a few individuals [71]. These methods include the use of restriction enzyme digestion–based assays with methyl-sensitive restriction endonucleases [74, 75], or methods based on sequence capture approaches [76], to sequence and map gene-rich portions of a genome, and allowing the anchoring of SNPs in a relatively unambiguous manner.

More recently, ultradense linkage maps have been created from genotyping by whole genome sequencing of a genetic mapping population. It has been used to place whole-genome sequencing contigs into a map, thus anchoring, and ordering, sequencing of contigs [77]. Such an approach requires using a genetic linkage map as a framework, into which SNPs derived from the whole genome sequencing assembly can be integrated into a genetic framework derived from low coverage whole-genome sequencing data from a segregating population. The genetic position of the sequence-derived SNPs can then be used to assign chromosomal locations to the contigs harboring them. Such an approach has been used in the context of a whole-genome assembly project in barley where genetic anchoring was applied to a whole-genome assembly [78]. SNPs discovered by sequencing individuals from two mapping populations at low coverage (~1×) were placed into genetic maps that had been previously constructed through different means, including SNP array data and GBS, or made from the whole-genome shotgun sequencing data of the population. Their genetic positions then were used to assign chromosomal locations, and integrate into the combined physical and genetic genome framework, approximately two-thirds of all whole-genome shotgun sequencing contigs. While highly effective in plants, where mapping populations are often readily available, it must be noted that such an approach is limited by the overall recombination landscape, and the subsequent relationship between physical and genetic distance within a particular region of the genome [76]. Recombination events in plants often occur in distal regions of the chromosomes, and peri-centromeric regions may require very large mapping populations to improve their resolution. In addition, recent studies have suggested that specific features of the genome, such as chromosomal inversion, translocation and duplication varying between the two parents used to generate the mapping population, may lead to errors and potentially confound genome assemblies.

4.3. BAC pool sequencing in gene-rich regions

A large number of genome assemblies have been generated with the help of physical maps and the use of a BAC-by-BAC sequencing approach. While laborious and costly, this approach still remains relevant as it offers multiple advantages over a whole-genome sequencing approach, especially in terms of assembling sequencing reads conserved in the context of a whole-genome assembly but mapping exclusively to a defined portion of a genome in the context of an individual clone assembly. Lonardi *et al*. [80] proposed a modified version of clone sequencing to take advantage of the massive sequencing capacity offered by NGS platforms. In that study, subsets of overlapping genome-tiling BAC clones were selected and pooled according to a multidimensional grid design. Each pool then was sequenced on an Illumina HiSeq2000 instrument. The resulting paired-end reads were deconvoluted by determining, for each read the intersection between the pool it originates from and the individual BAC clone(s) within that same pool covering the portion of the genome the read corresponds to, based on physical map information. Once deconvolution is achieved, reads can be assembled using an NGS assembler (Velvet) [81], to recreate the sequence of the original BAC clone. Such an approach was successfully tested in barley BAC clones selected based on BAC-unigene associations described in that same study, thus suggesting that BAC pool sequencing can be used in correlation with existing physical maps to complement or correct whole-genome sequencing assemblies, offering in the process the likelihood of higher quality contig sequence assemblies in gene-rich regions of complex plant genomes.

4.4. Optical mapping

Optical mapping is a single-molecule approach that produces fingerprints using ordered restriction maps [82] or specific nick sites [83]. After enzymatic treatment and subsequent incorporation of fluorescent labels, the DNA molecules are stretched on a glass surface or in a nanochannel array and directly imaged to locate regions corresponding to the restriction sites or nick sites within the molecule. Distances between those sites are then inferred to produce an optical map of the DNA molecule. Two commercial platforms currently are available, namely, the Opgen Argus [84] and the BioNano Genomics Irys [85] systems. Using such techniques, very large DNA molecules, in the Mb range, can be interrogated for the presence and location of short recognition sites (whose sequence will vary with the enzyme being used to treat the DNA). Consensus optical maps then can be created by determining the overlap, under highly redundant conditions, between optical maps of single DNA molecules. Such consensus maps have to take into account the possibility of errors inherent to this type of technology, including star activity and false enzyme cuts, or the possibility of chimeric maps when joining, for example, optically mapped molecules containing paralogous genomic regions.

Optical maps can be used for multiple applications, including comparative genomics and structural variation detection, as well as the development of optical map-guided genome assemblies, where the optical map is aligned and compared to *in silico* digested contig sequences. Optical map-guided genome assemblies can assist in building high-quality genome assemblies by providing evidence of the ordering of adjacent contigs and scaffolds, or by assessing the overall sequence accuracy of contigs and suggesting potential errors in an assembly, such as inversions, translocations or chimeric contig or scaffold sequences. The addition of optical maps to a genome assembly often results in a significant increase in the scaffold N50 value. For example, Hastie *et al*. [86] used the mapping of tiling BAC clones in a 2.1 Mb highly repetitive region of *Aegilops tauschii* (the D-genome donor of hexaploid wheat) to correct several misassemblies and improve the assembly from 75% to 95% complete. In another study [87], a high-resolution optical map, spanning 91% of the maize genome, was built, and used to characterize gaps within contigs, the maize genetic-physical (FPC) map and the reference pseudomolecules. Results also suggested that the placement of 12 FPC contigs on the maize genetic-physical map required re-evaluation.

4.5. Long-range Hi-C interactions

High-throughput Chromosome capture (Hi-C) is a method that uses cross-linking of DNA-binding protein to DNA followed by restriction digestion and self-ligation of protein-bound DNA fragments, to probe genome-wide three-dimensional chromatin interactions between chromosomal regions bound to the same proteins (such as enhancer and promoter regions) [88]. There is a statistically higher probability that those regions are located on the same chromosome rather than on different chromosomes, as expected within the context of chromosomes located in distinct three-dimensional spaces within the nucleus. As a result, a vast majority of Hi-C read pairs (where each paired reads correspond to reads that may be millions of bases apart from each other on the same chromosome) can be used to determine what two contigs can be linked together on the same chromosome, based on the Hi-C paired reads they each contain.

Burton *et al*. [89] evaluated the use of Hi-C datasets for long-range scaffolding of *de novo* whole-genome assemblies. This approach works, first, by aligning Hi-C reads to *de novo* assembly contig sequences and indexing each contig to their respective chromosomes, ordering contigs within each respective chromosome group by using higher Hi-C interaction densities expected between closely located contigs, and orienting ordered contigs using the location and orientation of Hi-C reads within each contig. The approach tested on existing human and mouse contig datasets generated from next-generation shotgun and mate-pair sequencing reads showed that a vast majority of the contigs could be grouped (98.2% and 98% of all sequences, in human and mouse, respectively) and ordered (94.4% and 86.7% of all grouped sequences, in human and mouse, respectively) within individual chromosomes when combined with Hi-C sequencing reads. Similar studies, where Hi-C datasets were used to complement *de novo* assembly generated with next-generation sequencing reads have been performed in human and mouse by Kaplan and Dekker [90] and Selvaraj *et al*. [91].

4.6. Long-range scaffolding

Two companies, namely, 10X Genomics [92] (Pleasanton, CA) and Dovetail Genomics [93] (Santa Cruz, CA), recently presented new ways to assemble short reads delivered by the Illumina technology. The GemCode instrument from 10X Genomics is a microfluidic device used to partition very long DNA molecules (typically 50 kb or more) into oil-based droplets and to prepare Illumina-compatible libraries in combination with "gel beads", each containing a unique 14-bp indexing barcode. Once sequencing is performed, in-house software deconvolutes the barcodes and reconstructs the sequence of the original DNA subfragments as to where they originate from on the original long DNA molecule. In contrast to 10X Genomics, Dovetail Genomics approach does not necessarily require an instrument but requires larger amount of starting material for preparing samples. Dovetail's approach works essentially by *in vitro* making a Hi-C library from chromatin-free purified DNA, thus recreating intramolecular interactions while reducing intermolecular ones. The resulting fragments can then be selected for mate-pair sets capturing long-range intramolecular interactions for genome scaffolding. While not yet applied on plant genome assemblies, it is presumed that the strategies and technologies highlighted above could potentially assist in grouping and ordering contigs and scaffolds from gene-rich regions of diploid plant genomes.

5. Conclusion

Reference genomes are now available for a significant number of plant species. The emergence of NGS technologies has made it possible to sequence genomes not only from economically important crop species but also from nonstandard model and special plants whose genomes otherwise might not have been sequenced due to the requirements for large funds, instrumentation and personnel that was witnessed in earlier pre-NGS days. While great progress has been made, assembling such genomes still remains challenging due to their inherent complexity and the relative absence of long-range connectivity, lost during DNA fragmentation and short-read sequencing. As a result, plant genome assemblies tend to be highly fragmented, and focused essentially on unique "gene-rich" regions, while large fractions of the genomes, namely, complex repeat and conserved regions, remain unassembled. Researchers have come up with creative ways to address those shortcomings, including the use of mate-pair NGS libraries, the complementation of physical assemblies with genetic maps, or the use of new technologies for sequencing, physical mapping or scaffolding. It is hoped that the routine use of such novel approaches will help in elucidating the biological aspects of genomes by allowing true comparative and structural analysis between species, strains, tissue or environment.

Acknowledgements

The authors would like to acknowledge Gregory May for his support and contribution to this book chapter.

Author details

Stéphane Deschamps* and Victor Llaca

*Address all correspondence to: stephane.deschamps@cgr.dupont.com

DuPont Pioneer, Wilmington, Delaware, USA

References

[1] Varshney RK, Navak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding.Trends Biotechnol. 2009;27:522–530. DOI: 10.1016/j.tibtech.2009.05.006

[2] Messing J, Llaca V.Importance of anchor genomes for any plant genome project. Proc. Natl. Acad. U.S.A.1998;95:2017–2020. DOI: 10.1073/pnas.95.5.2017

[3] Edwards D, Batley J.Plant genome sequencing: applications for crop improvement. Plant Biotechnol. J.2010;8:2–9. DOI: 10.1111/j.1467-7652.2009.00459.x

[4] Jackson SA, Iwata A, Lee SH, Schmutz J, Shoemaker R. Sequencing crop genomes: approaches and applications. New Phytol.2011;191:915–925. DOI: 10.1111/j.1469-8137.2011.03804.x

[5] Hou H, Atlihan N, Lu ZX. New biotechnology enhances the application of cisgenesis in plant breeding. Front. Plant Sci. 2014;11:389. DOI: 10.3389/fpls.2014.00389

[6] Green ED. Strategies for the systematic sequencing of complex genomes. Nat. Rev. Genet. 2001;2:573–583. DOI: 10.1038/35084503

[7] Liu, W, Yuan, JS, Stewart, CN. Advanced genetic tools for plant biotechnology. Nat Rev Genet. 2013;14:781–793. DOI: 10.1038/nrg3583

[8] Shendure, J, Aiden, EL.The expanding scope of DNA sequencing. Nature Biotechnol. 2012;30:1084–1094.DOI:10.1038/nbt.2421

[9] Mardis ER.A decade’s perspective on DNA sequencing technology. Nature. 2011;470:198–203. DOI: 10.1038/nature09796

[10] Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.J. Mol. Biol. 1975;94:441–448. DOI: 10.1016/0022-2836(75)90213-2

[11] Trainor GL. DNA sequencing, automation and the human genome. Anal. Chem. 1990;62:418–426. DOI: 10.1021/ac00204a001

[12] Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res.1998;8:175–185. DOI: 10.1101/gr.8.3.175

[13] Karger BL, Guttman A. DNA sequencing by CE. Electrophoresis. 2009;30:S196–202. DOI: 10.1002/elps.200900218

[14] Llaca V.Sequencing technologies and their use in plant biotechnology and breeding. In: Dr. Anjana Munshi, editor. DNA Sequencing—Methods and Applications. 2012. p. 35–60. DOI: 10.5772/37918

[15] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, *et al*.Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;269:496–512. DOI: 10.1126/science.7542800

[16] Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R.Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 1996;24:4420–4449. DOI: 10.1093/nar/24.22.4420

[17] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, *et al*.Life with 6000 genes. Science. 1996;274:563–567. DOI: 10.1126/science.274.5287.546

[18] *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science. 1998;282:2012–2018. DOI: 10.1126/science.282.5396.2012

[19] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, *et al*.The genome sequence of *Drosophila melanogaster*. Science. 2000;287:2185–2195. DOI: 10.1126/science.287.5461.2185

[20] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. DOI: 10.1038/35057062

[21] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al*.The sequence of the human genome. Science. 2001;291:1304–1351. DOI: 10.1126/science.1058040

[22] Staden R.A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. 1979;6:2601–2610. DOI: 10.1093/nar/6.7.2601

[23] Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res.1998;8:195–202. DOI: 10.1101/gr.8.3.195

[24] Soderlund C, Longden I, Mott R. FPC: a system for building contigs from restriction fingerprinted clones. Comput. Appl. Biosci. 1997;13:523–535. DOI: 10.1093/bioinformatics/13.5.523

[25] Ding Y, Johnson MD, Chen WQ, Wong D, Chen YJ, Benson SC, *et al*.Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. Genomics. 2001;74:142–154. DOI: 10.1006/geno.2001.6547

[26] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 2000;408:796–815. DOI: 10.1038/35048692

[27] Hosouchi, T, Kumekawa N, Tsuruoka H, Kotani H. Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1, 2, and 3. DNA Res. 2002;9:117–121. DOI: 10.1093/dnares/9.4.117

[28] International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005;436:793–800. DOI: 10.1038/nature03895

[29] Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, *et al*.The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326:1112–1115. DOI: 10.1126/science.1178534

[30] Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, *et al*.The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313:1596–1604. DOI: 10.1126/science.1128691

[31] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al.The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449:463–467. DOI: 10.1038/nature06148

[32] Ming R, Hou S, Feng Y, Yu Q, Fionne-Laporte A, Saw JH, *et al*.The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature. 2008;452:991–996. DOI: 10.1038/nature06856

[33] International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010;463:763–768. DOI: 10.1038/nature08747

[34] Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, *et al*.The *Sorghum bicolor* genome and the diversification of grasses. Nature. 2009;457:551–556. DOI: 10.1038/nature07723

[35] Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, *et al*.Genome sequence of the palaeopolyploid soybean. Nature. 2010;463:178–183. DOI: 10.1038/nature08670

[36] Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. Crop genome sequencing: lessons and rationales. Trends Plant Sci. 2011;16:77–88. DOI: 10.1016/j.tplants. 2010.10.005

[37] Saegusa A. US firm's bid to sequence rice genome causes stir in Japan. Nature. 1999;398:545. DOI: 10.1038/19123

[38] Llaca V, Deschamps S, Campbell M. Genome diversity in maize. J. Botany. 2011;104172. DOI: 10.1155/2011/104172

[39] Shendure J, Ji H.Next-generation DNA sequencing. Nat. Biotechnol.2008;28:1135–1145. DOI: 10.1038/nbt1486

[40] Ansorge WJ.Next-generation DNA sequencing techniques. N. Biotechnol. 2009;25:195–203. DOI: 10.1016/j.nbt.2008.12.009

[41] Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE.Landscape of next-generation sequencing technologies. Anal. Chem.2011;83:4327–4341. DOI: 10.1021/ac2010857

[42] Bentley DR, Balasubramanjan S, Swerdlow HP, Smith GP, Milton J, Brown CG, *et al*.Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53–59. DOI: 10.1038/nature07517

[43] Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G.BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 2006;34:e22. DOI: 10.1093/nar/gnj023

[44] Erlich Y, Mitra PP, de la Bastide M, McCombie WR, Hannon GJ. Alta-cyclic: a self-optimizing base caller for next-generation sequencing. Nat. Methods. 2008;5:679–682. DOI: 10.1038/nmeth.1230

[45] Michael TP, Van Buren R. Progress, challenges and the future of crop genomes. Curr. Opin. Plant Biol.2015;24:71–81. DOI: 10.1016/j.pbi.2015.02.002

[46] Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum. Mol. Genet.2010;19:R227–240. DOI: 10.1093/hmg/ddq416

[47] Feng Y, Zhang Y, Ying C, Wang D, Du C. Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics. 2015;13:4–16. DOI: 10.1016/j.gpb.2015.01.009

[48] Anton BP, Mongodin EF, Agrawal S, Fomenkov A, Byrd DR, Roberts RJ, *et al*.Complete genome sequence of ER2796, a DNA methyltransferase-deficient strain of *Escherichia coli* K-12. PLoS One. 2015;10:e0127446. DOI: 10.1371/journal.pone.0127446

[49] Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods. 2015;12:733–735. DOI: 10.1038/nmeth.3444

[50] Plant and animal whole genome sequencing [Internet]. 2015. Available from: http://www.pacb.com/applications/whole-genome-sequencing/plant-animal/ [Accessed 2015-10-14]

[51] DNA: nanopore sequencing [Internet]. 2015. Available from: https://nanoporetech.com/applications/dna-nanopore-sequencing [Accessed 2015-10-14]

[52] Bennetzen JL, Ma J, Devos KM. Mechanisms of recent genome size variation in flowering plants. Ann. Bot.2005;95:127–132. DOI: 10.1093/aob/mci008

[53] Rostoks N, Park YJ, Ramakrishna W, Ma J, Druka A, Shiloff BA, *et al*.Genomic sequencing reveals gene content, genomic organization, and recombination relation-

ships in barley. Funct. Integr. Genomics. 2002;2:51–59. DOI: 10.1007/s10142-002-0055-5

[54] Claros MG, Bautista R, Guererro-Fernandez D, Benzerki H, Seoane P, Fernandez-Pozo N. Why assembling plant genome sequences is so challenging. Biology (Basel). 2012;1:439–459. DOI: 10.3390/biology1020439

[55] Proost S, Pattyn P, Gerats T, Van de Peer Y. Journey through the past: 150 million years of plant genome evolution. Plant J.2011;66:58–65. DOI: 10.1111/j.1365-313X.2011.04521.x

[56] Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marcais G, *et al*. Sequencing and assembly of the 22-Gb loblolly pine genome. Genetics. 2014;196:875–890. DOI: 10.1534/genetics.113.159715

[57] Comai L.The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 2005;6:836–846. DOI: 10.1038/nrg1711

[58] Meyers LA, Levin DA. On the abundance of polyploids in flowering plants. Evolution. 2006;60:1198–1206. DOI: 10.1554/05-629.1

[59] Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, *et al*.Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature. 2013;496:87–90. DOI: 10.1038/nature11997

[60] Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485:635–641. DOI: 10.1038/nature11119

[61] Nederbragt AJ.On the middle ground between open source and commercial software—the case of the Newbler program. Genome Biol. 2014;15:113. DOI: 10.1186/gb4173

[62] Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, *et al*.Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat. Genet.2014;46:567–572. DOI: 10.1038/ng.2987

[63] Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, *et al*.De novo assembly of human genomes with massively parallel short read sequencing. Genome Res.2010;20:265–272. DOI: 10.1101/gr.097261.109

[64] Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, *et al*.The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet.2011;43:1035–1039. DOI: 10.1038/ng.919

[65] Brassica database [Internet]. 2010. Available from: http://brassicadb.org/brad/index.php [Accessed 2015-10-14]

[66] Van Deynze A, Ashrafi H, Hickey L, Peluso P, Rank D, Chin J, *et al*.Using spinach to compare technologies for whole genome assemblies. In: Plant & Animal Genome XXIII; 10–14 January 2015; San Diego, CA.

[67] Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, *et al.*Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods. 2013;10:563–569. DOI: 10.1038/nmeth.2474

[68] McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, *et al*. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS One. 2014;9:e106689. DOI: 10.1371/journal.pone.0106689.

[69] Deschamps S, Campbell M. Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Mol. Breed.2010;25:553–570. DOI: 10.1007/s11032-009-9357-9

[70] Elshire RJ, Glaubitz JC, Suri Q, Poland JA, Kawamoto K, Buckler ES, *et al.*A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6:e19379. DOI: 10.1371/journal.pone.0019379

[71] Deschamps S, Llaca V, May GD. Genotyping-by-sequencing in plants. Biology (Basel). 2012;1:460–483. DOI: 10.3390/biology1030460

[72] Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, *et al.*High-throughput genotyping by whole-genome resequencing. Genome Res.2009;19:1068–1076. DOI: 10.1101/gr.089516.108

[73] Trick M, Adamski NM, Mugford SG, Jiang CC, Febrer M, Uauy C. Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploidy wheat. BMC Plant Biol.2012;12. DOI: 10.1186/1471-2229-12-14

[74] Fellers JP. Genome filtering using methylation-sensitive restriction enzymes with six base pair recognition sites. The Plant Genome. 2008;1:146–152. DOI: 10.3835/plantgenome2008.05.0245

[75] Gore MA, Wright MH, Ersoz ES, Bouffard P, Szekeres ES, Jarvie TP, *et al.*Large-scale discovery of gene-enriched SNPs. The Plant Genome. 2009;2:121–133. DOI: doi: 10.3835/plantgenome2009.01.0002

[76] Muraya MM, Schmutzer T, Ulpinnis C, Scholz U, Altmann T. Targeted sequencing reveals large-scale sequence polymorphism in maize candidate genes for biomass production and composition. PLoS One. 2015;10:e0132120. DOI: 10.1371/journal.pone.0132120

[77] Mascher M, Muehlbauer GJ, Rokhsar DS, Chapman J, Schmutz J, Barry K, *et al.*Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J. 2013;76:718–727. DOI: 10.1111/tpj.12319

[78] Ariyadasa R, Mascher M, Nussbaumer T, Schulte D, Frenkel Z, Poursarebani N, *et al.*A sequence-ready physical map of barley anchored genetically by two million sin-

gle-nucleotide polymorphisms. Plant Physiol.2014;164:412–423. DOI: 10.1104/pp.113.228213

[79] Mascher M, Stein N. Genetic anchoring of whole-genome shotgun assemblies. Front. Genet.2014;5:208. DOI: 10.3389/fgene.2014.00208

[80] Lonardi S, Duma D, Alpert M, Cordero F, Beccuti M, Bhat PR, *et al*.Combinatorial pooling enables selective sequencing of the barley gene space. PLoS Comput. Biol. 2013;9:e1003010. DOI: 10.1371/journal.pcbi.1003010

[81] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res.2008;18:821–829. DOI: 10.1101/gr.074492.107

[82] 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. DOI: 10.1126/science.8211116

[83] Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, *et al*.Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat. Biotechnol. 2012;30:771–776. DOI: 10.1038/nbt.2303

[84] Argus system [Internet]. 2015. Available from: http://opgen.com/genomic-services/argus-system [Accessed 2015-10-14]

[85] Irys system [Internet]. 2015. Available from: http://www.bionanogenomics.com/products/ [Accessed 2015-10-14]

[86] Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, *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. DOI: 10.1371/journal.pone.0055864

[87] Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, *et al*.A single molecule scaffold for the maize genome. PLoS Genet.2009;5:e1000711. DOI: 10.1371/journal.pgen.1000711

[88] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, *et al*.Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–293. DOI: 10.1126/science.1181369

[89] 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. DOI: 10.1038/nbt.2727

[90] Kaplan N, Dekker J. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat. Biotechnol.2013;31:1143–1147. DOI: 10.1038/nbt.2768

[91] Selvaraj, S, R Dixon J, Bansai V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. Biotechnol.2013;31:1111–1118. DOI: 10.1038/nbt.2728

[92] Megabase scale phasing from 10× Genomics [Internet]. 2015. Available from: http://10xgenomics.com/sites/default/files/downloads/10x-0009_app_note_phasing_press_ready.pdf [Accessed 2015-10-14]

[93] Chromosome-scale shotgun assembly using an in vitro method for long-range linkage [Internet]. 2015. Available from: http://arxiv.org/abs/1502.05331 [Accessed 2015-10-14]