5 Strategies and Tools for Sequencing and Assembly of Plant Genomes

D. C. Mishra, S. B. Lal, Anu Sharma, Sanjeev Kumar, Neeraj Budhlakoti and Anil Rai

Abstract

This chapter highlights strategies and tools for sequencing and assembly of plant genomes. It discusses in brief the methods of sequencing technologies (the first, second and third generations), details the approaches of genome assembly (the de novo and reference assembly) and presents the challenges of plant genome assembly.

5.1 Introduction

In general, plant genomes have higher ploidy, higher rates of heterozygosity and repeats. Furthermore, the gene content in plants can be very complex, as shown by the presence of large gene families and abundant pseudogenes with nearly identical sequences derived from recent whole genome duplication events and transposon activity. In order to understand the complexity of the plant genomes, DNA sequencing and its assembly are very important. Thus, genome sequencing and its assembly have been major priorities in plant genetic research during the past 25 years. With rapid advancements in sequencing technologies, not only the efficiency of sequencing has greatly improved, but also significantly reduced the associated cost.

The efficiency of assembling the plant genome depends on sequencing technology and type of assembler used. Broadly, there are two types of approaches of assemblies being used by the scientific communities, i.e. de novo assembly and reference-based assembly. De novo genome assemblers are used for the reconstruction of novel genomes from a collection of reads without any reference genomes, whereas reference-based assembly is highly dependent on the availability of the reference genomes of the same or closely related species. During the last decade, efforts have been made to develop de novo assemblers to work on short read sequences generated by Next Generation Sequencing (NGS) technologies. NGS technologies are highly efficient in terms of cost and time as compared to the traditional Sanger's approach (Sanger et al. 1977). The emergence of short read sequencing imposes new challenges in assembling plant genomes due to their size and complex nature. The de novo

D. C. Mishra · S. B. Lal · A. Sharma · S. Kumar
N. Budhlakoti · A. Rai (✉)
Center for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India
e-mail: anilrai@iasri.res.in

S. K. Chakrabarti et al. (eds.), *The Potato Genome*,
Compendium of Plant Genomes, https://doi.org/10.1007/978-3-319-66135-3_5

assembly of plant genomes from these short read sequences leads to a large number of contigs and low N50 values, mainly due to the complexity of the genome and the presence of conserved regions. These limitations of short read sequencing technologies have been addressed by third generation, long read sequencing technologies. The simplicity offered by long read sequencing is often offset by low accuracy with the error rates of 10–20% of the generated sequences.

For all the above reasons, de novo assembly of a plant genome poses great challenges in spite of the availability of varied platforms of genome sequencing. Assembling a plant genome requires high coverage, long read length and high quality with a low error rate. It may be noted that it is necessary to integrate the sequences from different sequencing platforms in order to have a quality plant genome assembly.

The existing assemblers are mostly platform-dependent and are unable to handle the integration of data coming from different platforms. This further increases the computational complexity of the genome assembly process. Moreover, the available genome assemblers are either based on serial processing or based on very limited use of parallel processing technology. A number of genome assembly algorithms have been developed incorporating the benefits of short and long read sequences for de novo hybrid assembly (Jason et al. 2010).

5.2 Sequencing Technologies

DNA sequencing determines the exact order of nucleotides in a given DNA molecule, i.e. this process determines the order of the four bases: (1) adenine; (2) guanine; (3) cytosine; and (4) thymine. With the advances in DNA sequencing methods, the pace of biological research and discovery has been accelerated. Sequencing of DNA molecules started in the early 1970s with the development of the Maxam-Gilbert method, followed by the Sanger method, based on chain termination approach during the same period. Subsequently, due to the development of fluorescence-based sequencing methods along with automated analysis, this DNA sequencing process has become easier and faster. Gradually, DNA sequencing moved from traditional cloning DNA molecules to PCR based amplification analysis. Currently, the DNA sequencing is based on Single Molecule Real Time Sequencing (SMRT) methods.

5.2.1 First-Generation Sequencing Methods

The Maxam-Gilbert method was the first method of sequencing a DNA molecule. This method is based on radioactive labelling of DNA molecule at the 5′ end. In this, chemical cleavages at variable positions specific to four nucleotide reactions (G, A+G, C, C+T) are labelled (Maxam and Gilbert 1977). Therefore, a series of labelled fragments are generated by different chemical reactions. Fragments of four different reactions are electrophoresed using acrylamide gels for size separation.

Sanger's method was developed after the Maxam-Gilbert technique. This method uses a special chemical compound, dideoxynucleoside triphosphates (ddNTPs), which lacks the hydroxyl group in the 3′ position. In this process, sequencing starts with synthetic 5′-end-labelled fragment, having oligodeoxynucleotide as a primer, polymerase and template DNA molecule. Every polymerization reaction requires the normal deoxynucleotide triphosphates (dNTPs) at a high concentration and one of the four ddNTPs at a low concentration. With the addition of DNA polymerase, the sequencing of DNA molecule is extended until a ddNTP is encountered. With an optimum dNTP: ddNTP ratio, the DNA chain will terminate at a variable length [30]. The terminated chain will always be identified with the help of a specific ddNTP. Hence, the resulting sequence can be obtained by reading the respective gel lane. Therefore, the original sequence of DNA can be obtained by reading the sequencing gel from bottom to top (Fig. 5.1).

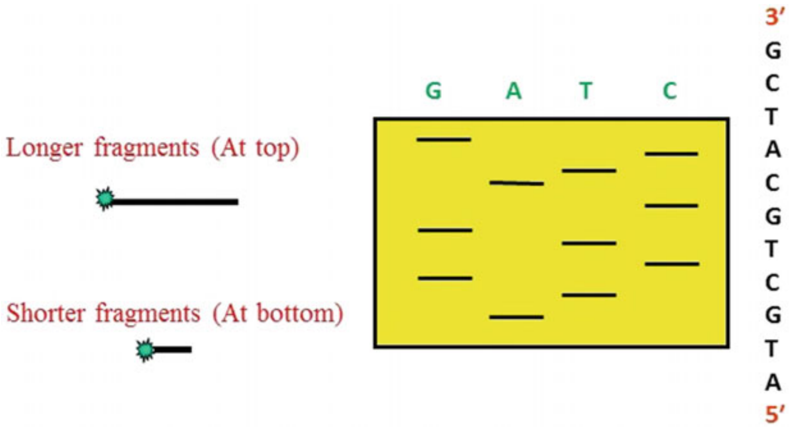

Fig. 5.1 Gel lane of Sanger's sequencing method (Sanger et al. 1977)

5.2.2 Second-Generation Sequencing

The first-generation sequencing technologies are resource-inefficient in terms of cost and time. For example, the draft assembly of the rice (*Oryza Sativa*) genome of size 390 MB took approximately around 5 years using Sanger's approach, while the wheat (*Triticum aestivum*) genome is 40 times larger and more complex than the rice genome, if Sanger's method had been used, the assembly of wheat genome would have taken more than a century. Therefore, there was an urgent need to develop a faster sequencing technology to sequence plant genomes. Keeping in mind the above need, a number of second generation sequencing technologies such as Pyrosequencing, Roche/454, Illumina, SOLiD and IonTorrent have been developed.

Pyrosequencing is a second-generation sequencing technology based on the sequencing by synthesis principle. In this sequencing method, primer is hybridized to a single-stranded DNA molecule and determination of sequence is based on the detection of pyrophosphate (PPi) released during this process. It simply requires the DNA templates to be sequenced, DNA polymerase, sequencing primer, APS (adenosine-5`-phosphosulphate), luciferin and other enzymes. Different dNTPs are sequentially added to this mixture and based on the chemical reaction and release of pyrophosphate target, a template is sequenced. The process of pyrosequencing could easily be understood through following reactions (Fig. 5.2).

This sequencing technology of Roche/454 is based on the concept of Emulsion PCR. In this sequencing process, first, the ssDNA library is prepared using suitable adapters, which was followed by its annealing with an excess of DNA beads. Also, these beads are washed to filter untethered strands, which are then subjected to a hybridization-based enrichment. Further, four DNA nucleotides were added subsequently and the DNA polymerase reaction extends the nucleotide chain by adding complementary nucleotides. Addition of these nucleotides generates a specific signal which is captured by a Charge Coupled Device (CCD) camera.

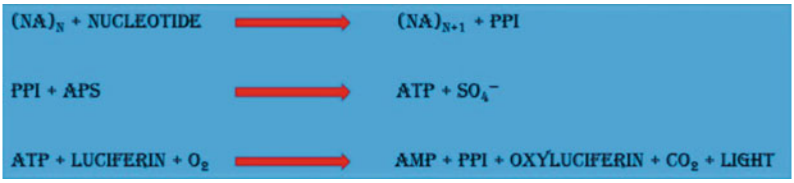

Fig. 5.2 Sequential steps of pyrosequencing

The most popular technology of second-generation sequencing is Illumina. It is based on the concept of Bridge PCR. In this, a DNA library is prepared using ligation of a suitable adapter which is further enriched by using Bridge PCR, i.e. solid phase amplification. The sequencing process starts by adding labelled modified nucleotide (reversible terminators), DNA polymerase and sequencing primer. Templates are sequenced in parallel by adding a single base at a time. These bases compete with each other to bind with templates and added nucleotide is identified using a laser. This natural competition among nucleotides improves the accuracy. This sequencing technology has overcome the drawback of Roche/454 by controlling the homopolymer error.

In the case of SOLiD, the sequencing process has two choices of sample preparation, i.e. (1) single DNA fragment library generation; and (2) mate pair library generation. Then, this DNA library is amplified on beads using emulsion PCR. These enriched beads are ligated to sequencing primer and fluorescently labelled by di-base probe where each fluorescent dye represent four of sixteen di-nucleotide sequences. Sequencing reaction starts with hybridization of complementary probes and ligated. Dye is cleaved off after measurement of fluorescence. Subsequently new primer is hybridized with one length greater than earlier and this process is repeated. This sequencing platform provides the dual measurement of each base, hence accuracy is improved.

Ion-Torrent is a faster sequencing technology which is based on the release of hydrogen atoms during elongation of the DNA chain. This technology relies on a semi-conductor-based detection system. This sequencing process starts with library preparation, enrichment of reads and release of the hydrogen ion after incorporation of the nucleotides to ssDNA (Quail et al. 2012). This release of the hydrogen ion alters the pH of the chemical mixture, resulting in a change of voltage. This fluctuation in voltage indicates the incorporated nucleotide. This process occurs simultaneously in millions of wells.

5.2.3 Third-Generation Sequencing

The second-generation sequencing technologies are based on short read sequencing. However, around 200 plant genomes were sequenced by now. due to the low cost of sequencing and the faster rate of data generation. But the chromosome levels of de novo assembled sequences are available only for a few plant genomes. Most of these plant genome assemblies based on short read sequencing have a large number of contigs and scaffolds. This is due to the fact that the plant genomes are highly repetitive in nature, due to the transposable elements and are also large in genome size. For example, the genome size of the pine tree is more than 20 GB. Therefore, a number of third-generation sequencing technologies were developed to overcome the problem of short read sequencing. The four major technological developments in this area are: (1) Heliscope Sequencing; (2) Pacific Biosciences; (3) Oxford Nanopore Technologies Limited; and (4) Illumina—Synthetic Long Read (SLR).

Heliscope sequencing was developed by Helicos Biosciences. It is based on the principle of Single Molecule Real Time (SMRT) sequencing. In this, a DNA sequence is sheared into small pieces of around hundred base pair, then Poly A nucleotides with labelled fluorescence are added to 3′-end of each DNA fragment. These modified DNA sequence are then hybridized to Helicos flow cells, having millions of oligo T's immobilized in the flow cell surface. Furthermore, this hybridized molecule is loaded into the Helicos instrument. The flow cell is incorporated with the addition of fluorescently labelled nucleotide and DNA polymerase. An image is taken by CCD camera and the nucleotide is identified based on fluorescence specific to nucleotides.

The technology developed by Pacific Biosciences is most widely used for long read sequencing technology based on SMRT sequencing (Quail et al. 2012). The sequencing process starts with the addition of polymerase, nucleotide (phosphor linked with different fluorescence). Here phosphor-linked nucleotide carries the fluorescence to the phosphate rather than to the base. The activity of individual molecules is observed in Zero-Mode Waveguides (ZMW).

Nanopore sequencing technology has been developed by Oxford. In this process, the DNA molecule is passed through a nanopore membrane and voltage is applied across this membrane (Mikheyev and Tin 2014). Flow of ion through the pores creates a current and the passed nucleotide is identified through a specific pattern of current (Laver et al. 2015).

Synthetic Long Read (SLR) sequencing technology was developed by Illumina. This technology is based on the generation of long reads through short read sequences of Illumina. In this, a long fragment of DNA has been distributed into multiple tiny fragments and these tiny fragments are sequenced after bar-coding using the Illumina short read sequencing. These bar-coded tiny fragments are assembled into a synthetic long read. Different sequencing technology platforms are compared in Table 5.1 (Glenn 2011; Liu et al. 2012; Mikheyev and Tin 2014; Niedringhaus et al. 2011; Shendure and Ji 2008; Quail et al. 2012).

5.3 Approaches of Genome Assembly

In order to assemble the genome from the data generated through different sequencing technologies, several genome assemblers have been developed in the past two decades. The algorithms of different assemblers differ in many ways depending on: (1) type of reads (i.e. long reads to short reads); (2) type of graph construction; (3) way of sequencing the error correction; and (4) the ability to deal with different length of fragments. Mainly two types of assembling algorithms have been developed according to the type of reads, i.e. long and short. Short read assembling algorithms can be further classified based on two approaches: (1) contig extension; and (2) a de Bruijn graph. The de Bruijn graph is a graph data structure that is particularly suitable to represent the overlap relationship of short read sequences. Several short read assemblers based on de Bruijn graphs have been developed. The most widely used assemblers in this category include ALLPATHS (Butle et al. 2008), Velvet (Zerbino and Birney 2008), ABySS (Simpson et al. 2009) and SOAPdenovo (Li et al. 2010). In the case of contig extension algorithms, a greedy algorithm is used in which overlapped areas among sequences are identified and merged, this process continues till no overlapping sequence is found. Some contig extension-based assemblers are SSAKE (Warren et al. 2007), PE-Assembler (Ariyaratne and Sung 2011) and SHARCGS (Dohm et al. 2007). However, this approach is not efficient in the case of plant genome assembly due to high repetitive regions. It may be noted that short read sequencing is useful for genome assembly of some species but is not able to resolve major repeat families of the plant genome.

Overlap Layout Consensus (OLC) is generally used for assembling long read sequences. Overlap graphs work well if there is a small number of reads with significant overlap. However, this method is computationally expensive for large plant genomes. The complexity of pairwise sequence alignment is quadratic in terms of number of reads. Recent advances in sequencing technologies such as SMRT have the capability to resolve repetitive structures in the assembly graph. A number of assembly algorithms have been developed to resolve the repetitive structures of the plant genome sequences. In this regard, MIRA is one of the OLC-based assemblers, which uses both high as well as low quality regions of the genome along with repetitive region tags. Some of the OLC-based assemblers use the MinHash Alignment algorithm. This is a probabilistic algorithm and able to detect overlaps efficiently between reads. Furthermore, it uses a dimensionality reduction technique called MinHash to create more compact representation of sequence reads and reduce space complexity. Other approaches for OLC-based assemblers use supervised learning to detect overlaps to improve the quality of contigs and classify homogeneous sequences in the data. In contrast to short read

Table 5.1 Comparison of different DNA sequencing methods

Method	Sanger	Ion Torrent	454 (Roche)	Illumina HiSeq 2500 (Rapid Run)	SOLiD	PacBio	Oxford Nanopore (MinION)
Read length (bp)	600–1000	200	700	2 × 250	2 × 60	1.0–1.5 × 10^4 on average	2–5 × 10^3 on average
Error rate (%)	0.001	1	1	0.1	5	15–20	15–20
Reads per run	96	8.2 × 10^7	1 × 10^6	1.2 × 10^9 (paired)	8 × 10^8	3.5–7.5 × 10^4	1.1–4.7 × 10^4
Time per run	0.5–3 h	2–4 h	23 h	1–6 days	6 days	0.5–4 h	50 h
Strength	High quality, long read length	Faster run time, Low cost	Long read size and Fast	High sequence yield, low error rate	Low cost per base	Longest read length	Minimal sample preparation, long read length
Weakness	High cost low throughput	Small read length	Low yield, Homopolymer error	Small read length	Small read length	Low sequence yield, high error rate	High error rate

assemblers, long read assemblers are good at resolving the repeat regions but suffer from low accuracy.

Plant genome assembly is a very complex procedure which depends on many factors such as the size of the genome, the repeat regions, heterogeneity, polyploidy and other factors. Depending upon the availability of the reference genome of closely related species and the size of the reads, genome assembly is broadly classified into three categories: (1) reference or comparative assembly; (2) de novo assembly; and (3) hybrid assembly.

5.3.1 Reference Assembly

Reference assembly is used only when the assembled genome of the same species or closely related species is available. In this approach, reads are mapped to the reference genome which forms the layout of the overlapping reads and finally a consensus sequence is produced. Reference assembly consists of three major steps:

1. Read alignment.
2. Layout refinement.
3. Consensus sequence generation.

5.3.1.1 Step 1: Read Alignment

In this step, each read is aligned with the available reference genome. In order to obtain the chains of mutually consistent matches, the Longest Increasing Subsequences (LIS) algorithm is used. The objective of this algorithm is to find the length of the longest subsequence of a given sequence such that the length of all the subsequences are arranged in ascending order. In this way the LIS algorithm produces the layout of the overlapping reads.

5.3.1.2 Step 2: Layout Refinement

The layout generated by the read alignment to the reference genome has many constraints, such as the presence of indels, the rearrangement between the target and the reference genome, etc. Due to these constraints, accurate mapping of the reads to the reference genome is difficult. Consequently, only a partial number of reads are mapped to the reference genome. Therefore, it is essential to go for the refinement of the layout formed by the read alignment step. Layout refinement is the most difficult step in reference-based assembly. In this step, reconstruction of indels information is done by following de novo assembly of these reads in the mismatched part of the target genome.

5.3.1.3 Step 3: Consensus Sequence Generation

In this step, a Multiple Sequence Alignment (MSA) algorithm is used for the generation of a consensus sequence. For each of the refined layouts, MSA is applied to find the overlapped reads for the generation of the consensus sequence. Here, the MSA algorithm follows an iterative approach to find the final consensus sequence. In each iteration, a pairwise alignment between each read and current consensus sequence is carried out to find the next consensus sequence. This process is repeated until the new consensus sequence is same as the previous one. This consensus sequence is called a contig.

5.3.1.4 Step 4: Scaffold Generation

The contigs obtained lack the information regarding their order and orientation. Scaffolds are generated by combining contigs together in the proper order and orientation. Scaffold generation requires the information of mate pair, physical/genetic map or some additional information, such as BAC library, optical mapping, long-range HI-C interaction, etc.

5.3.2 De Novo Assembly

The de novo assembly needs to be done in the absence of the availability of a reference. Therefore, the assembly of the plant genome is done right from scratch (Chaisson et al. 2004, 2009). There are two broad approaches: (1) OLC; and (2) a de Bruijn graph (DBG) approach based on the read length for the de novo assembly.

5.3.2.1 Approach 1: Overlap Layout Consensus (OLC)

It is desirable to use OLC for assembling plant genomes using long read sequences due to the fact of obtaining desirable overlap regions among the sequences. The performance of the OLC approach is poor in the case of short read sequencing as in the case of assembling short read sequences both time and space complexities are very high. Also, this approach is highly computational-intensive and not suitable for large numbers of reads (i.e. short reads of large genomes). Assembly based on OLC is performed using following steps:

Step 1: Identification of Candidate Overlap
Sequence overlaps are identified by constructing an overlap graph by pair-wise comparison of each read to other reads. Nodes in the overlap graph represent reads while sequence overlaps are shown by edges.

Step 2: Fragment Layout Formation
Fragment layout formation is done through bundling stretches of overlaps, which satisfies the prefixed criterion of (1) minimum length of overlaps; (2) maximum length of overhangs; (3) minimum similarity in the overlapping region; and (4) maximum number of local errors.

Step 3: Consensus Sequence Generation
An overlap graph is traversed to find the simple path for a consensus sequence generation. This path is obtained by traversing through all the nodes and edges and keeping the node in the path at most once.

5.3.2.2 Approach 2: De-Bruijn Graph (DBG)

The DBG-based approach is also known as the k-mer graph approach, and requires the generation of k-mers of reads for graph construction (Chaisson et al. 2004). A de Bruijn graph is a form of directed graph of the same in and out degrees where each node represents k-mer and edges represent the overlaps between the reads. In this way, contigs are generated by traversing the Eulerian path in the graph. This approach is comparatively faster than OLC. However, in the case of long reads, its performance is not satisfactory due to the increase in computational complexity. The following problems may occur during the formation of contigs from a k-mer graph:

1. In the case of a low coverage sequencing, a tip, i.e. a short, dead-end divergence from the main path is formed (Fig. 5.3a).
2. Plant genomes are often very heterozygous or polymorphic. In such cases, bubbles may be formed in k-mer graph (Fig. 5.3b).
3. Plant genomes are highly repetitive in nature, due to which a frayed rope-like structure, i.e. convergent and divergent paths may be formed (Fig. 5.3c).

The contig formation is followed by scaffold generation as discussed in step 4 of the OLC approach.

5.3.2.3 Approach 3: Hybrid Assembly

Various sequencing platforms have some inherent advantages and disadvantages. As already discussed, some of these platforms produce large reads with a high error rate (Pacific Biosciences) while others generate short reads with high accuracy (Illumina). Therefore, it is desirable to use the sequencing data from both types of

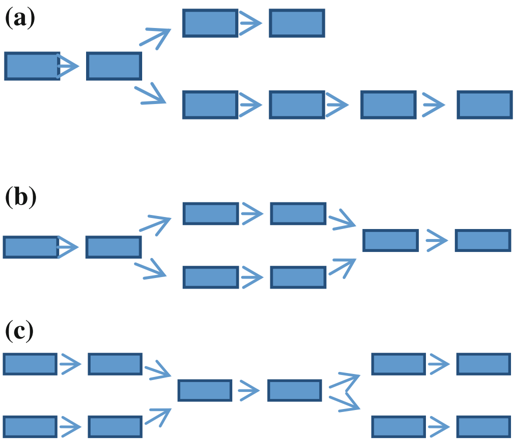

Fig. 5.3 **a** Tip formation in k-mer graph, **b** Bubble formation in k-mer graph, **c** Frayed rope formation in k-mer graph

platforms to improve the quality of plant genome assembly. This approach is known as hybrid assembly. There are various methods of hybrid assembly where the data from different platforms are combined either at the level of reads or at the level of contigs. A method has been developed for hybrid error correction and de novo assembly of single-molecule sequencing reads (Koren et al. 2012). A hybrid assembly pipeline has been developed using primary and secondary assembly steps (Wang et al. 2012). Similarly, a method of hybrid assembly of Illumina and Roche/454 data has been developed (Utturkar et al. 2014). Also, there exists a hybrid error correction method known as LoRDEC, that builds a succinct DBG representing the short reads, and seeks a corrective sequence for each erroneous region in the long reads by traversing chosen paths in the graph by Salmela and Rivals (2014). Further, a novel hybrid error correction algorithm for long PacBio sequencing reads that uses pre-assembled Illumina sequences for the error correction has been developed (Lee et al. 2014). A popular hybrid assembler named Jabba has been developed, in which the hybrid method is used to correct long third-generation reads by mapping them onto a corrected DBG that was constructed from second-generation data (Miclotte et al. 2016). Recently one efficient approach called DBG2OLC (Ye et al. 2016) has been developed and used extensively. This pipeline is executed through following steps:

- Generate contigs using de Bruijn graph (DBG) from highly accurate NGS short reads. The generated contigs are mapped to the long reads and long reads are further compressed into a list of contig identifiers.
- Multiple sequence alignment is used to clean the errors present in the long reads.
- Following the OLC approach, a best overlap graph of the cleaned compressed reads is generated.
- A final consensus sequence is obtained by decompressing the compressed long reads and obtaining the simple path from the generated best overlap graph.

A description of some widely used plant genome assembly tools is given in Table 5.2.

5.4 Issues and Challenges of Plant Genome Assembly

Three major issues associated with the genome assembly are: (1) computational complexity; (2) biological complexity; and (3) the quality genome finishing. The issues and challenges in these areas are discussed below:

5.4.1 Computational Complexity

The plant genome assembly needs high-end computational resources to assemble the sequence fragment of DNA. Also, the assembly programs should be able to handle large data sets efficiently. Two major algorithms employed by existing assemblers are based on OLC and the DBG approach (Li et al. 2011). Each of these has associated memory and space requirements. The OLC-based approach was implemented in many assemblers, namely, Celera, CABOG (Miller 2008) and MaSuRCA (Zimin et al. 2013). This approach is computationally constrained by the complicacy in the identification of overlaps between reads. This step requires O(n^2) pairwise alignments, where, n is the number of reads, and each pairwise alignment is O(nm) where n and m are the lengths of the reads. Many variants of OLC algorithm have been proposed by scholars and researchers to reduce the time complexity in the original algorithm, such as the use of dynamic programming and indexing (Li et al. 2011).

Eulerian path-based DBG graphs using k-mers are a much faster approach compared to OLC, given the same computational memory (Li et al. 2011). A major issue is the space complexity which requires optimization. Some of the widely used assemblers based on DBG are Velvet for very short reads (Zerbino and Birney 2008), SOAPdenovo2 (Short Oligonucleotide Analysis Package) (Luo et al. 2012), Minia (Chikhi and Rizk 2012), Ray for parallel genome

Table 5.2 List of plant genome assembly tools

Assembler	Input	Acceptable technologies	Year	Assembler type
ABySS	Genomic reads	Solexa, SOLiD	2008	De novo
ALLPATHS-LG	Genomic reads	Solexa, SOLiD	2011	De novo
Celera WGA Assembler	Genomic reads	Sanger, 454, Solexa	2004	De novo and reference assembly
CLC Genomics Workbench	Genomic reads	Sanger, 454, Solexa, SOLiD	2008	De novo and reference assembly
DNASTAR	Genomic reads, exomes, transcriptomes, metagenomes, ESTs	Illumina, ABI SOLiD, Roche 454, Ion Torrent, Solexa, Sanger	2007	De novo
Newbler	Genomic reads, ESTs	454, Sanger	2004	De novo
PASHA	Genomic reads	Illumina	2011	De novo
Phrap	Genomic reads	Sanger, 454, Solexa	1994	Reference assembly
TIGR Assembler	Genomic reads	Sanger	1995	Reference assembly
Trinity	Transcriptomes	short reads (paired, oriented, mixed) Illumina, 454, Solid, …	2011	De novo
SOAPdenovo	Genomic reads	Solexa	2009	De novo
SPAdes	Genomic reads	Illumina, Solexa, Sanger, 454, Ion Torrent, PacBio, Oxford Nanopore	2012	De novo
Velvet	Genomic reads	Sanger, 454, Solexa, SOLiD	2007	De novo
LoRDEC	Genomic reads	Illumina, PacBio	2014	Hybrid assembler
DBG2OLC	Genomic reads	Illumina, PacBio, Oxford Nanopore	2016	Hybrid assembler
Jabba	Genomic reads	Illumina, PacBio	2016	Hybrid assembler

assemblies for parallel DNA sequencing (Boisvert et al. 2010), etc. The most computational-intensive and space-intensive task is the construction of the DBG graph. Algorithms like Minia and SparseAssembler (Ye et al. 2012) tackle the space complexity problem of DBG algorithms, however, a sacrifice on accuracy and runtime is made.

Assemblers based on a greedy algorithm make use of a graph structure and the construction of a graph, which is a computationally complex task. Even the greedy assembly algorithms like SSAKE (Warren et al. 2007), VCAKE (Jeck et al. 2007) and others are not computationally efficient at graph construction.

5.4.2 Biological Complexity

Sequencing of large genomes scales up both the biological and computational complexity during the assembly process. Increasing the genome size results in the increase in the number and type of sub-clones, the number of sequence reads, the computational resources requirement and the demand for better assembly algorithms. Further, this complexity increases with an increase in the depth of coverage.

The presence of large repetitive/duplicated regions enhances the generation of redundant sequences in plant genomes, which may lead to poor assembly of the genome. One of the primary difficulties in computational genome assembly is to develop an algorithmic approach capable of detecting stretches of repetitive DNA without compromising the quality of the assembly. Repetitive sequences complicate the assembly as different pieces of sequence can share the same repeat sequence but originating from different genomic locations. Since the pieces are put together by searching for matching overlapping nucleotides, these repeats can be put together erroneously. Typically, for shotgun data, repetitive sequences are revealed by clusters containing more overlapping reads than expected by chance.

Another biological complexity of the plant genomes arises due to polymorphism in plants. A high degree of heterozygosity in plants can complicate the assembly, depending on the sequencing strategy and the assembly algorithm. Some assemblers, such as Platanus (Kajitani et al. 2014) or Spades (Bankevich et al. 2012), perform comparatively better than others. Apart from this, the assembly process of plant genomes is further complicated by the chromosomal structure. During the sequencing process, the stem-loop structure of the centromere region of the chromosome is generally ignored. Also, parts of the telomere region are not properly sequenced. These biological complexities create problems during the assembly process of the genome.

5.4.3 Genome Finishing

Genome assemblies produced by different assemblers must be re-examined and reconsidered with respect to low coverage of reads, poor quality of data and inadequate handling of repeat regions. This re-examination is performed manually as well as with the help of automated tools to elucidate the specific ambiguities. This procedure is known as genome finishing and it consists of three main sub-processes, namely, gap closure, assembly validation and genome refinement.

The main approaches used for gap closure are: (1) directed-PCR; (2) mate-pair libraries; and (3) primer-walking. Mate-pair libraries are used to infer the adjacency of contigs and filling the gaps in the assembly. PCR experiments are used in those situations where mate-pair libraries cannot be used, such as regions with stem-loop structure on the chromosome. Other relevant information, such as BAC libraries, EST, mRNA, physical map, etc. is used for the gap closure. The recent techniques, such as optical mapping, long-range-HI-C technique, are now becoming popular for gap closure.

Analysis of the assembled contigs can be performed using a number of tools. One of these is Consed (Gordon et al. 1998), which allows the navigation of the assembled contigs and reads. Using this tool, problematic regions of the genome can be searched and tagged, based on different criteria for further inspection. Other tools for a similar task are Autofinish (Gordon et al. 2001), BAC cardi (Bartels et al. 2005) and GAP4 (Bonfield et al. 1995).

Genome assembly validation and refinement can be done using physical/genetic maps which provide a context or scaffold for the sequence assembly contigs. Genetic maps typically provide context in terms of simple sequence repeats that generally occur near genic regions. The availability of the genome sequence for a closely related organism can also provide some support for assembly validation. The assembly can also be validated using molecular markers like SNPs, SSRs, AFLP, RAPD, RFLP, etc. ESTs are also useful for checking quality genome assembly as well as genome refinement.

5.5 Conclusion

Plant genomes are relatively very complex in nature and sequencing plant genome as well as its assembly are still challenging tasks. However, NGS technology has accelerated the process of plant genome sequencing, but most of the assembled plant genomes are highly fragmented due to lack of a proper algorithm dealing with biological and computational complexity. The majority of existing algorithms are not able to perfectly preserve the repetitive regions, the regions of heterozygosity, structural variants, etc. Thus, there is a need to develop better computational approaches to preserve the additional information along with an efficient algorithm for searching the shortest common superstring/ sequence and finding Eulerian walks in a DBG. With the advances in the computing industry, the cost of memory and cores has dropped remarkably and biologists are using cluster and GPU-based systems for genome assembly. This has considerably reduced the problem of computational complexity. The process of plant genome assembly can be further accelerated by developing efficient algorithms using a parallelized computing framework on GPU clusters. Further, the big data analytic approach is another promising area that may be applied for faster genome assembly. The major challenge faced by researchers is the modification of the existing algorithm to recreate the biological truth.

References

Ariyaratne PN, Sung WK (2011) PE-Assembler: de novo assembler using short paired-end reads. Bioinformatics 27(2):167–174

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477

Bartels D, Kespohl S, Albaum S, Druke T, Goesmann A, Herold J, Kaiser O, Puhler A, Pfeiffer F, Raddatz G, Stoye J, Meyer F, Schuster SC (2005) BACCardI—a tool for the validation of genomic assemblies, assisting genome finishing and inter genome comparison. Bioinformatics 21(7):853–859

Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol 17:1519–1533

Bonfield J, Smith K, Staden R (1995) A new DNA sequence assembly program. Nucleic Acids Res 23 (24):4992–4999

Butle J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18(5):810–820

Chaisson M, Pevzner P, Tang H (2004) Fragment assembly with short reads. Bioinformatics 20:2067–2074

Chaisson MJ, Brinza D, Pevzner PA (2009) De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res 19:336–346

Chikhi R, Rizk G (2012) Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms Mol Biol 8:22

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res 17(11):1697–1706

Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11:759–769

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8 (3):195–202

Gordon D, Desmarais C, Green P (2001) Automated finishing with auto finish. Genome Res 11(4):614–625

Jason R, Miller SK, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327

Jeck WR, Reinhardt JA, Baltrus DA (2007) Extending assembly of short DNA sequences to handle error. Bioinformatics 23:2942–2944

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res 24:1384–1395

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotech 30:693–700

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif 3:1–8

Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M (2014) Error correction and assembly complexity of single molecule sequencing reads. bioRxiv doi:10.1101/006395

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang JI, Wang JU (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20(2):265–272

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B, Yang B, Fan W (2011) Comparison of the two major classes of assembly algorithms: overlap layout consensus and de-bruijn graph. Brief Funct Genomics 11(1):25–37

Liu L, Li Y, Li S, Hu N, He Y, Pong R (2012) Comparison of next-generation sequencing systems. J Bio Med Res Int 2012:e251364

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G et al (2012) SOAPdenovo2: an empirically improved memory-efficient short read de novo assembly. GigaScience 1:18

Maxum A, Gilbert W (1977) A new method for sequencing DNA. Procd Natl Acad Sci 74:560–564

Miclotte G, Heydari M, Demeester P, Rombauts S, Van de Peer Y, Audenaert P, Fostier J (2016) Jabba: hybrid error correction for long sequencing reads. Algorithms Mol Biol 11:10

Mikheyev AS, Tin MMY (2014) A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour 14:1097–1102

Miller JR (2008) Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24(24):2818–2824

Niedringhaus TP, Milanova D, Kerby MB, Snyder MP (2011) Barron Landscape of next-generation sequencing technologies. Anal Chem 83:4327–4341

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genom 13:341

Salmela L, Rivals E (2014) LoRDEC: accurate and efficient long read error correction. Bioinformatics 30:3506–3514

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Procd Natl Acad Sci 74:5463–5467

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: A parallel assembler for short read sequence data. Genome Res 19:1117–1123

Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD (2014) Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics 30:2709–2716

Wang Y, Yu Y, Pan B, Hao P, Li Y, Shao Z, Xu X, Li X (2012) Optimizing hybrid assembly of next-generation sequence data from Enterococcus faecium: a microbe with highly divergent genome. BMC Systems Biol 6 (Suppl 3):S21

Warren RL, Sutton GG, Jones SJ, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501

Ye C, Cannon CH, Ma ZS, Yu DW, Pop M (2012) Sparseassembler2: Sparse k-mer graph for memory efficient genome assembly. ArXiv:1108.3556

Ye C, Hill CM, Wu S, Ruan J, Ma Z (2016) DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Scientific Rep 6:31900–31906

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29(21):2669–2677