Next-Generation Sequencing and Assembly of Plant Genomes

Basant K. Tiwary

Contents

Abstract

Next-generation sequencing technology produces enormous volume of accurate and inexpensive sequence data in a short span of time. Three available common next-generation sequencing (NGS) platforms for genome sequencing are discussed here. The genome assembly and scaffolding algorithms are described with special emphasis on *de novo* assembly of short-read sequences. The biological applications of next-generation sequencing in plant sciences are covered with examples from plant genomics. An account on future prospects of this technology in plant genome analysis is discussed.

Introduction

The sequence-driven research in molecular biology started with pathbreaking research by two groups led by Sanger and Gilbert (Sanger et al. 1977; Maxam and Gilbert 1977). High-throughput sequencing (HTS) techniques popularly known as next-generation sequencing (NGS) were introduced in 2005 and have revolutionized the biomedical research by substantial increase in scale and resolution of various biological applications. They provide manyfold reads at a markedly reduced cost per sequenced nucleotide than conventional Sanger sequencing. Next-generation sequencing generates a huge amount of data necessitating the development of

B.K. Tiwary, Ph.D. (✉)
Centre for Bioinformatics, Pondicherry University,
Pondicherry 605 014, India
e-mail: basant68@email.com; basant@bicpu.edu.in

D. Barh et al. (eds.), *PlantOmics: The Omics of Plant Science*,
DOI 10.1007/978-81-322-2172-2_3, © Springer India 2015

powerful computing and efficient algorithms. All commercial platforms have three common phages in their development, namely, preparation of sequencing library by adding adapters (defined sequences), immobilization of DNA fragments of sequencing library to a solid surface, and sequencing (Myllykangas et al. 2011). It can be used for whole genome sequencing, targeted resequencing, and identification of transcription factor binding sites and expression of noncoding RNA. There are several commercial platforms available such as 454 pyrosequencing (Roche Applied Science), the genome analyzer (Illumina), and SOLiD (Applied Biosystems). Next-generation sequencing can be applied to detect molecular variants such as single nucleotide variants, genomic insertions and deletions, and genomic rearrangements. RNA-seq can be used to determine the expression level of known genes and discovery of novel genes. ChIP-seq can be used for screening protein-DNA interaction at genome-wide scale. The whole genome sequencing and assembly of an organism are performed in various phases (Fig. 1). The major focus of this article is to introduce the reader with three common high-throughput sequencing platforms with more emphasis on various computational methods to analyze the next-generation sequencing data obtained from plant genomes.

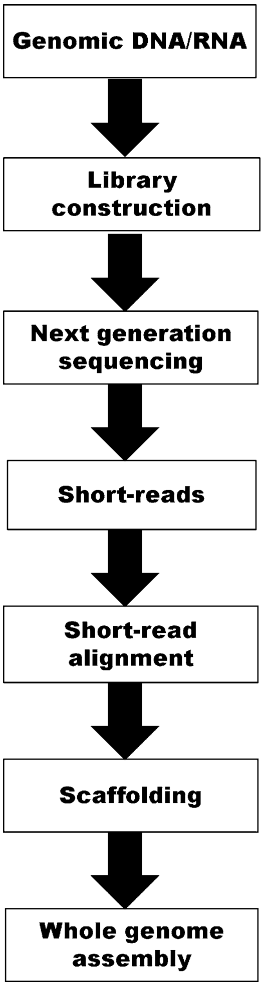

Fig. 1 Flow chart showing various steps in genome sequencing and assembly using next-generation sequencing technology

Next-Generation Sequencing Platforms

The three most popular sequencing platforms widely used to date are Roche 454 pyrosequencing, Illumina (Solexa), and SOLiD (Applied Biosystems).

454 Pyrosequencing

454 is the first next-generation technology introduced by Roche/454 Life Sciences which is based on pyrosequencing. In pyrosequencing, a double-stranded DNA is generated from a single-stranded DNA template by the addition of nucleotides. The addition of nucleotides is detected by the emission of light. It achieves a high throughput (~500 Mbp/run) with 400 bp read lengths. The major demerit associated with this platform is high error rate in homopolymer regions.

Illumina

The Illumina generates a much higher throughput (~1.5 Gbp) with a lower read length (~150 bp) when compared with 454. Although the read length is short, the platform generates high-quality sequences with an error rate less than

1 %. The principle of this method is based on sequencing by synthesis (reversible terminator chemistry). Long homopolymer runs do not affect the quality of sequence due to its chemistry. However, the major challenge associated with this technology is to process a large number of short reads, difficulty in *de novo* assembly, and in resequencing.

SOLiD

The ABI/SOLiD platform generates maximum high throughput achieved to date by any method (~3 Gbp/run) with a read length of 75 bp. This platform requires two kinds of library preparation: fragment or mate paired. Clonal bead populations are prepared in microreactors, and modified microbeads are deposited onto a glass slide. The sequencing was done using multiple cycles of ligation, detection, and cleavage.

Genome Assembly Algorithms

Most of the biological applications of next-generation sequencing such as quantification of transcriptome, assembly of new genomes, and identification of protein binding sites and alignment of sequence reads to a reference sequence as a first step of analysis. The process of aligning short reads into longer sequences is known as assembly. It is like a jigsaw puzzle where each short read is an individual part of the puzzle, and the whole genome sequence is a finished puzzle. There are many alignment tools developed in the last 4 years which are better than classical aligners in terms of speed and accuracy. Alignment algorithms can be based on hash tables, suffix trees, and merge sorting (Miller et al. 2010). The concept of hash table started with basic local alignment search tool (BLAST) which finds significant local alignment comparing exact matches to a k-mer (seeds) in a hash table. Ma et al. (2002) improved this method by creating the spaced seed (i.e., a seed with internal mismatches) which turned out to be the most popular approach for alignment of short reads. Eland, SOAP, SeqMap, MAQ, RMAP, ZOOM, and Novoalign are various popular programs for aligning short reads to a reference genome using spaced seed. Although spaced seed has mismatches within the seed, it never permits any gap in the seed. Eland was the first program developed by Anthony Cox from Illumina that aligns short oligonucleotides against a reference genome. SOAP is an efficient program for gapped and ungapped efficient alignment of short reads onto a reference genome (Li et al. 2008a). SeqMap can map a large amount of short reads to a genome based on index-filtering algorithm (Jiang and Wong 2008). MAQ builds assemblies by mapping short reads to a reference genome using quality score (Li et al. 2008b). RMAP software package has tools to map paired-end reads using a more sophisticated quality score (Smith et al. 2009). ZOOM maps short reads onto a reference genome with improved sensitivity and speed (Lin et al. 2008). Novoalign, a commercial software developed by Novocraft Technologies, is an aligner for short reads from Illumina genome analyzer. Another seeding approach *q*-gram filter builds an index allowing a gap within the seed. Two programs SHRiMP and RazerS are based on *q*-filter algorithm. SHRiMP is highly efficient in mapping short reads to a reference genome with high polymorphism (Rumble et al. 2009). RazerS is a popular read mapper with improved performance for long reads with large numbers of indels (Weese et al. 2009).

The algorithms based on suffix/prefix tries may be represented in a form of suffix tree (McCreight 1976), enhanced suffix array (Abouelhoda et al. 2004), and FM-index (Simpson and Durbin 2010). All algorithms identify exact matches at first and then build inexact alignments based on the exact matches. The suffix trie is a data structure that stores all the suffixes of a string in order to allow fast string matching. A trie needs a huge space and is impractical for even a small genome. Thus, there are various data structures such as suffix tree, suffix array, and FM-index to reduce the space. A suffix tree requires 12–17 bytes per nucleotide and is impractical for holding human genome in memory (Li and Homer 2010). The enhanced suffix array is more space efficient than suffix tree and takes only 6.25 bytes per nucleotide. The FM-index is the most space-efficient data structure using 0.2–2 bytes per nucleotide, and an FM-index of the entire human genome needs 2–8 GB of memory. The most widely used data structure is FM-index due to its small memory footprint. Bowtie, BWA, SOAP2, BWT-SW, and BWA-SW are the most popular programs built upon FM-index. Bowtie is a very fast and memory-efficient aligner for large genomes based on Burrows-Wheeler indexing (Langmead et al. 2009). Burrows-Wheeler alignment (BWA) tool is another efficient short-read aligner for large genomes allowing mismatches and indels based on Burrows-Wheeler transform (Li and Durbin 2009). SOAP2 is a short oligonucleotide alignment program with reduced memory usage and improved alignment speed (Li et al. 2009a). BWT-SW is an efficient tool to find all local alignments (Lam et al. 2008). Burrows-Wheeler Aligner's Smith-Waterman (BWA-SW) alignment is an efficient algorithm to align long reads of up to 1 Mb against a large sequence database (Li and Durbin 2010). However, MUMmer (Kurtz et al. 2004) and OASIS (Meek et al. 2003) are based on suffix tree, whereas Segemehl (Hoffmann et al. 2009) and Vmatch (Abouelhoda et al. 2004) apply enhanced suffix array as data structure. The yet other aligner of biological sequences (YOABS) is a very efficient long-read alignment program having advantages of both hash- and trie-based algorithms (Galinsky 2012).

There are more than 50 short-read alignment software packages available, albeit few of them are popular among users. Table 1 gives a list of popular alignment software packages available for short reads. All programs generate outputs in the form of a Sequence Alignment/Map (SAM; Li et al. 2009b) or Binary Alignment/Map (BAM; Carver et al. 2010) format which can be viewed through alignment viewers (Table 2) such as GBrowse (Stein et al. 2002), LookSeq (Manske and Kwiatkowski 2009), Tablet (Milne et al. 2010, 2013), BamView (Carver et al. 2010; Carver et al. 2013), GenomeView (Abeel et al. 2012), IGV (Thorvaldsdóttir et al. 2013), and MGAviewer (Zhu et al. 2013). The SAM format can be created and manipulated using SAMtools (Li et al. 2009b) which has extensive information regarding a read, its properties, and its alignment to a reference sequence. BAM format is the compressed binary form of SAM format which can be converted to SAM format and *vice versa* using SAMtools.

Table 1 Popular programs for short-read alignment

Program	Algorithm	References
Eland	Spaced seed	Illumina software
SOAP	Spaced seed	Li et al. (2008a)
SeqMap	Spaced seed	Jiang and Wong (2008)
MAQ	Spaced seed	Li et al. (2008b)
RMAP	Spaced seed	Smith et al. (2009)
ZOOM	Spaced seed	Lin et al. (2008)
Novoalign	Spaced seed	Novocraft Tech.
SHRiMP	Q-filter	Rumble et al. (2009)
RazerS	Q-filter	Weese et al. (2009)
BWA	FM-index	Li and Durbin (2009)
Bowtie	FM-index	Langmead et al. (2009)
SOAP2	FM-index	Li et al. (2009b)
BWT-SW	FM-index	Lam et al. (2008)
BWA-SW	FM-index	Li and Durbin (2010)
MUMmer	Suffix tree	Kurtz et al. (2004)
OASIS	Suffix tree	Meek et al. (2003)
Segemehl	Enhanced suffix array	Hoffmann et al. (2009)
Vmatch	Enhanced suffix array	Abouelhoda et al. (2004)
YOABS	Hash and trie based	Galinsky (2012)

Table 2 Alignment viewers of SAM/BAM format

Program	References
GBrowse	Stein et al. (2002)
LookSeq	Manske and Kwiatkowski (2009)
Tablet	Milne et al. (2010, 2013)
BamView	Carver et al. (2010, 2013)
GenomeView	Abeel et al. (2012)
IGV	Thorvaldsdóttir et al. (2013)
MGAviewer	Zhu et al. (2013)

De Novo Assembly of Short Reads

De novo sequence assembly is a method where individual short reads are merged into a long continuous sequence (contig) like the original template. In fact, short reads of 40 nucleotide length can be used to assemble the vast majority of protein encoding genes in most of the prokaryotic genomes albeit having many gaps. Table 3 shows common programs for assembling short reads without any reference genome. All algorithms for assembling capillary-based sequence reads of 400–1,000 nucleotides into long contiguous sequences adopts a common approach known as overlap-layout-consensus (OLC) approach (Li et al. 2012). The OLC algorithm first finds overlaps between sequence reads and then looks for most fitting pairs of reads (layout) and finally derives a consensus sequence from this layout. The overlap step is computationally expensive, and therefore various algorithmic approaches have been adopted to improve the computational efficiency. The OLC approach is adopted by many popular programs such as Arachne (Batzoglou et al. 2002), Celera Assembler (Myers et al. 2000), CAP3 (Huang and Madan 1999), PCAP (Huang et al. 2003), PHRAP (de la Bastide and McCombie 2007), Phusion (Mullikin and Ning 2003), and Newbler (a commercial assembler developed by Roche Diagnostics). Most of the assemblers designed for short-read sequences are based on De Bruijn graph (DBG; Li et al. 2012) and Eulerian path approach (Pevzner et al. 2001). Some of the popular software packages based on DBG and Eulerian paths are Euler (Pevzner et al. 2001), Euler-USR (Chaisson et al. 2009), Velvet (Zerbino and Birney 2008), ABySS (Simpson et al. 2009), ALLPATH-LG (Gnerre et al. 2011), SOAPdenovo (Li et al. 2010), and Gossamer (Conway et al. 2012). The graph-based algorithm assembly creates a model where the string is a node and the relation between strings is represented in a form of edges. In De Bruijn graph (DBG) algorithm, reads are chopped into smaller fragments (k-mers), and k-mers are converted into a DBG for final determination of genome sequence. The optimal solution is obtained through finding a Eulerian path (i.e., a path which covers a node only once) through the graph. However, the string graph assembler (SGA) is a program based on a string graph which keeps all reads intact and generates a graph based on overlaps between reads (Simpson and Durbin 2012).

Table 3 Some programs for *de novo* assembly of short reads

Program	Algorithm	References
Arachne	OLC	Batzoglou et al. (2002)
Celera Assembler	OLC	Myers et al. (2000)
CAP3	OLC	Huang and Madan (1999)
PCAP	OLC	Huang et al. (2003)
PHRAP	OLC	de la Bastide and McCombie (2007)
Phusion	OLC	Mullikin and Ning (2003)
Newbler	OLC	Roche Diagnostics
Euler	DBG and Eulerian path	Pevzner et al. (2001)
Euler-USR	DBG and Eulerian path	Chaisson et al. (2009)
Velvet	DBG and Eulerian path	Zerbino and Birney (2008)
ABySS	DBG and Eulerian path	Simpson et al. (2009)
ALLPATH-LG	DBG and Eulerian path	Gnerre et al. (2011)
SOAPdenovo	DBG and Eulerian path	Li et al. (2010)
Gossamer	DBG and Eulerian path	Conway et al. (2012)
SGA	String graph	Simpson and Durbin (2012)

Scaffolding Algorithms

The large assembled regions of sequence are known as contigs which need to be joined together to get the whole genome sequence. The final process of joining multiple contigs together to form a continuous genome sequence (scaffold or supercontig) is known as scaffolding or finishing. This process is done in four consecutive steps, namely, contig orientation, contig ordering, contig distancing, and gap closing. The orientation of contigs in same direction (5′-3′ direction in prokaryotes) is done using a reverse complementary sequence. The contigs are placed in an appropriate order starting at the origin of replication and extended in 5′-3′ direction of DNA replication. The distance between contigs can be estimated after correct orientation and order. The final step of closing and filling gap can

Table 4 Popular programs for scaffolding

Program	References
SOAPdenovo	Li et al. (2010)
ABySS	Simpson et al. (2009)
Bambus	Pop et al. (2004)
SOPRA	Dayarian et al. (2010)
SSPACE	Boetzer et al. (2011)
OPERA	Gao et al. (2011)
MIP Scaffolder	Salmela et al. (2011)
GRASS	Gritsenko et al. (2012)
RACA	Kim et al. (2013)

result into a finished genome. The paired-end reads provide additional information for grouping two contigs in a genome. Scaffolding process may be based on a graph, where a contig is treated as node and matching pair contigs are connected by edges. The algorithm finds an optimal path through the graph. The scaffolding process may be more accurate using additional information such as reference sequences of related organism, restriction maps, and RNA-seq data. Some of the popular programs (Table 4) for scaffolding are SOAPdenovo (Li et al. 2010), ABySS (Simpson et al. 2009), Bambus (Pop et al. 2004), SOPRA (Dayarian et al. 2010), SSPACE (Boetzer et al. 2011), OPERA (Gao et al. 2011), MIP Scaffolder (Salmela et al. 2011), GRASS (Gritsenko et al. 2012), and RACA (Kim et al. 2013).

Biological Applications of Next-Generation Sequencing

Genome Sequencing

The worldwide effort to understand the genetic basis of common and rare genetic disorder has gained momentum with the advent of next-generation sequencing technology. It will help largely in the identification of single nucleotide polymorphisms (SNPs) and haplotype mapping (International HapMap Consortium 2003) in individual human genomes and lay a foundation for personalized medicine. The 1,000 genome project (http://www.1000genomes.org) turned into reality with the availability of NGS technology. Cancer biology is another area where next-generation sequencing can decipher the novel molecular pathways involved in tumorigenesis (Hahn and Weinberg 2002). Next-generation sequencing will also influence the highly emerging area of synthetic biology where a new enzyme or a novel genetic network may be developed (Khalil and Collins 2010).

Functional Genomics

Functional genomics is focused to apply genomics data for understanding dynamic life processes. RNA-seq is widely used to quantify gene expression levels for different genes like microarray technology (Wang et al. 2009). It has several advantages over microarray analysis such as no prior sequence information is needed; highly expressed and lowly expressed genes are equally detected and allow detailed identification of structure of transcripts including alternative promoters and alternative splicing sites. In RNA-seq technology, the relative abundance of a transcript is estimated by counting the number of times it is hit by the sequence reads. This method accurately estimates relative RNA levels under different experimental conditions or in different cell types.

Epigenetics

Epigenetics deals with heritable regulatory changes in chromosomes without any change in the DNA sequence (Bird 2007). The epigenetic changes such as DNA methylation, histone modification, and ncRNA have an important role in maintaining chromosome structure. The post-translational modifications of histones such as methylation, acetylation, ubiquitination, and phosphorylation generate different “marks” for different functional properties. The DNA-binding proteins, histone modifications, or nucleosomes can be mapped across the genome using *ChIP-seq* approach where a chromatin immunoprecipitation (CHIP) is followed by sequencing (Park 2009). The DNA methylation involves methylation of the cytosine base in DNA and can be identified by a version of NGS known as *Methyl-seq* (Brunner et al. 2009). The active gene regulatory elements can be better understood by using another approach of NGS known as *DNAse-seq* (Song and Crawford 2010). Noncoding RNA (ncRNA) has been implicated in many epigenetic events such as X-chromosome inactivation and gene silencing (Mercer et al. 2009).

Current Status of Next-Generation Sequencing in Plant Genomics

NGS has been used extensively for whole genome sequencing of plants in the last 5 years (Table 5). *Arabidopsis thaliana* (125 Mb) was the first plant completely sequenced in 2000 using Sanger sequencer (*Arabidopsis* genome; Initiative 2000). It was followed by sequencing two major rice varieties, namely, *japonica* (420 Mb; Goff et al. 2002) and *indica* (466 Mb; Yu et al. 2002) in 2002 and first fruit grapevine (487 Mb; Jaillon et al. 2007) in 2007 using the same sequencing method. Subsequently, the draft genomes of papaya (372 Mb; Ming et al. 2008) and legume *Lotus japonicus* (315 Mb; Sato et al. 2008) were developed in 2008. Sorghum genome (730 Mb; Paterson et al. 2009) and maize genome (2.3 Gb; Schnable et al. 2009), and soya bean genome (1.1 Gb; Schmutz et al. 2010) were sequenced in 2009 and 2010, respectively, onto Sanger platform. However, the pace of sequencing plant genomes rapidly increased with the advent of the next-generation sequencing (NGS) technology. Cucumber genome (243.5 Mb; Huang et al. 2009) was sequenced taking advantages of both

Table 5 Overview of plant genomes sequenced applying next-generation sequencing

Plant	Genome size (Mb)	NGS platform	References
Cucumber (*Cucumis sativus*)	244	Illumina	Huang et al. (2009)
Wild grass (*Brachypodium distachyon*)	355	Illumina	International Brachypodium Initiative (2010)
Cocoa (*Theobroma cacao*)	430	Roche 454 and Illumina	Argout et al. (2011)
Apple (*Malus* X *domestica*)	604	Roche 454	Velasco et al. (2010)
Woodland strawberry (*Fragaria vesca*)	210	Roche 454, Illumina, and SOLiD	Shulaev et al. (2011)
Potato (*Solanum tuberosum*)	727	Roche 454 and Illumina	Potato Genome Sequencing Consortium (2011)
Cannabis (*Cannabis sativa*)	534	Roche 454 and Illumina	van Bakel et al. (2011)
Pigeon pea (*Cajanus cajan*)	606	Illumina	Varshney et al. (2012)
Extremophile crucifer (*Thellungiella parvula*)	140	Roche 454 and Illumina	Dassanayake et al. (2011)
Date palm (*Phoenix dactylifera*)	658	Illumina	Al-Dous et al. (2011)
Grape (*Vitis vinifera*)	505	Roche 454	Velasco et al. (2007)
Brassica rapa	288	Illumina	Wang et al. (2011)
Cotton (*Gossypium raimondii*)	775	Illumina	Wang et al. (2012)
Melon (*Cucumis melo*)	375	Roche 454 and Illumina	Garcia-Mas et al. (2012)
Tomato (*Solanum lycopersicom*)	760	Roche 454, Illumina, and SOLiD	Tomato Genome Consortium (2012)
Banana (*Musa acuminata*)	472	Roche 454 and Illumina	D'Hont et al. (2012)
Barley (*Hordeum vulgare*)	4,980	Roche 454 and Illumina	International Barley Genome Sequencing Consortium (2012)
Bread wheat (*Triticum aestivum*)	17,000	Roche 454	Brenchley et al. (2012)
Wheat A (*Triticum urartu*)	4,940	Illumina	Ling et al. (2013)
Sweet orange (*Citrus sinensis*)	367	Illumina	Qiang et al. (2013)
Chickpea (*Cicer arietinum*)	740	Roche 454	Jain et al. (2013)
Sacred lotus (*Nelumbo nucifera*)	929	Roche 454 and Illumina	Ming et al. (2013)

Illumina GA technology (high sequencing depth and low unit cost) and Sanger technology (long read and clone length). In 2010, wild grass, *Brachypodium distachyon* was sequenced using both methods (International Brachypodium Initiative 2010). The cocoa genome (430 Mb; Argout et al. 2011) was sequenced applying two NGS platforms, namely, Roche 454 and Illumina along with Sanger sequencing. The apple genome (604 Mb; Velasco et al. 2010) was sequenced using both Roche 454 technology and Sanger technology. The woodland strawberry (209.8 Mb) was sequenced using three NGS platforms: Roche 454, Illumina Solexa, and Life Technologies SOLiD (Shulaev et al. 2011). In 2011, the potato genome (727 Mbp) was sequenced using two major NGS platforms: Roche 454 and Illumina Genome Analyzer along with conventional Sanger sequencing technology (Potato Genome Sequencing Consortium 2011). The cannabis genome (534 Mb) was sequenced using Roche 454 and Illumina Genome Analyzer IIx or HiSeq platforms (van Bakel et al. 2011). The draft genome of pigeon pea (606 Mb) was sequenced with Illumina technology along with Sanger technology (Varshney et al. 2012). A close relative of *Arabidopsis*, *Thellungiella parvula*, is endemic to saline habitat, and its genome (140 Mb) was investigated using Roche 454 and Illumina GA2 (Dassanayake et al. 2011). The date palm genome (658 Mb) was sequenced *de novo* using Illumina GA2 and Sanger sequencer (Al-Dous et al. 2011). A draft consensus sequence of grape genome (504 Mbp) was developed with 1.7 million SNPs (Velasco et al. 2007). *Brassica rapa* genome was sequenced by *Brassica rapa* genome sequencing project consortium (Wang et al. 2011). The cotton plant draft genome (775 Mb) was sequenced using Illumina HiSeq 2000 platform (Wang et al. 2012). Melon, a close relative of cucumber, was covered for genome (375 Mb) using Roche 454 pyrosequencing, Illumina, and Sanger technologies (Garcia-Mas et al. 2012). The tomato genome (760 Mb) was sequenced using Roche 454 GS FLX, Illumina GA2, and SOLiD along with Sanger sequencing (Tomato Genome Consortium 2012). The banana genome (472 Mb) was sequenced with combined application of Roche 454, Illumina GA2, and Sanger technologies (D'Hont et al. 2012). Recently, both Roche 454 (GS FLX or FLX Titanium) and Illumina (GA2 or HiSeq 2000) have been applied to decipher the genome sequence of barley (4.98 Gbp) (The International Barley Genome Sequencing Consortium 2012). Since the bread wheat has a large genome size (17 Gb) than other cereals and is hexaploid in nature, the successful completion of bread wheat genome sequencing using 454 pyrosequencing and wheat A-genome (4.94 Gb) sequencing on Illumina HiSequation platform is a significant event in the next-generation sequencing of crops (Brenchley et al. 2012; Ling et al. 2013). The completion of wheat genome will not only pave the way for better productivity of wheat crop but decipher the role of polyploidy in plant genome evolution as well. Recently, the whole genome sequencing of sweet orange (*Citrus sinensis*; 367 Mb) was done on Illumina GAII sequencer (Qiang et al. 2013). The draft genome sequence of chickpea (*Cicer arietinum*; 740 Mb) was completed on 454/Roche GS FLX Titanium platform (Jain et al. 2013). The complete genome of sacred lotus (929 Mb) was sequenced with combined application of Illumina and 454 technologies (Ming et al. 2013). Other whole genome sequencing projects are underway in many plant species such as amborella (*Amborella trichopoda*), columbine (*Aquilegia* sp.), sugar beet (*Beta vulgaris*), monkey flower (*Mimulus guttatus*), rose gum tree (*Eucalyptus grandis*), flax (*Linum usitatissimum*), cassava (*Manihot esculenta*), and pear (*Pyrus bretschneideri*). Some species of the lower plant species were sequenced in order to understand the evolution of vascular plants on land. The genome of green alga (Chlamydomonas reinhardtii; 120 Mb) (Merchant et al. 2007), genome of moss (480 Mb; *Physcomitrella patens*) (Rensing et al. 2008), and genome of lycophyte (*Selaginella moellendorffii*; 213 Mb) (Banks et al. 2011) were sequenced using conventional Sanger sequencing and have revealed insights into genomic evolution of land plants.

Future Prospects in Next-Generation Sequencing and Assembly

NGS-based technology has a wide scope for solving many existing problems in genomics. However, the low read length with intrinsic error rate of NGS is a major problem and is a prohibitive factor for *de novo* assembly of large genomes. Therefore, this technology is largely based on the availability of a reference genome. However, this problem will be solved in the future with an increase in size of the read length. Although NGS provides a deep coverage, it has a low throughput in comparison to microarrays. However, this problem may be alleviated by developing large-scale parallel NGS. With an increase in the number of reference genomes, it is expected that whole genome resequencing will become more popular in order to interrogate the diversity of crop genomes. New dedicated algorithms are needed to deal with complex repeats in the plant genome for better quality of assembly. Along with assembly algorithms, the next-generation data quality and quantity should be improved in the near future.

Conclusion

In this work, three common NGS platforms and various computational methods for analysis of NGS-derived sequence data are discussed. The impact of NGS technology on plant genome sequencing especially on crop genomes, till date, is elaborated. It is expected that NGS technology will grow further in sensitivity and speed and will decipher the genomes of other plants to understand the genome evolution and help in revealing key genomic features to agricultural productivity.

References

Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y (2012) GenomeView: a next-generation genome browser. Nucleic Acids Res 40:e12

Abouelhoda MI, Kurtz S, Ohlebusch E (2004) Replacing suffix trees with enhanced suffix arrays. J Discrete Algorithms 2:53–86

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J et al (2011) De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nat Biotechnol 29:521–527

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN et al (2011) The genome of *Theobroma cacao*. Nat Genet 43:101–108

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, Ashton NW et al (2011) The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science 332:960–963

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES (2002) ARACHNE: a whole-genome shotgun assembler. Genome Res 12:177–189

Bird A (2007) Perceptions of epigenetics. Nature 447:396–398

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579

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

Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, Neff NF, Anton E, Medina C, Nguyen L, Chiao E et al (2009) Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res 19:1044–1056

Carver T, Bohme U, Otto T, Parkhill J, Berriman M (2010) BamView: viewing mapped read alignment data in the context of the reference sequence. Bioinformatics 26:676–677

Carver T, Harris SR, Otto TD, Berriman M, Parkhill J, McQuillan JA (2013) BamView: visualizing and interpretation of next-generation sequencing read alignments. Brief Bioinform 14:203–212

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

Conway T, Wazny J, Bromage A, Zobel J, Beresford-Smith B (2012) Gossamer—a resource-efficient *de novo* assembler. Bioinformatics 28:1937–1938

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M et al (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488:213–217

Dassanayake M, Oh D-H, Haas JS, Hernandez A, Hong H, Ali S, Yun D-J, Bressan RA, Zhu J-K, Bohnert HJ et al (2011) The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet 43:913–918

Dayarian A, Michael TP, Sengupta AM (2010) SOPRA: scaffolding algorithm for paired reads via statistical optimization. BMC Bioinf 11:345

de la Bastide M, McCombie WR (2007) Assembling genomic DNA sequences with PHRAP. Curr Protoc Bioinform, Chapter 11:Unit 11.4

Galinsky VL (2012) YOABS: yet other aligner of biological sequences—an efficient linearly scaling nucleotide aligner. Bioinformatics 28:1070–1077

Gao S, Sung WK, Nagarajan N (2011) Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. J Comput Biol 18:1681–1691

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, Gonzalez VM, Henaff E, Camara F, Cozzuto L, Lowy E et al (2012) The genome of melon (*Cucumis melo* L.). Proc Natl Acad Sci U S A 109:11872–11877

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108:1513–1518

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. *ssp japonica*). Science 296:92–100

Gritsenko AA, Nijkamp JF, Reinders MJT, de Ridder D (2012) GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. Bioinformatics 28:1429–1437

Hahn WC, Weinberg RA (2002) Mechanisms of disease: rules for making human tumor cells. N Engl J Med 34:1593–1603

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5:e1000502

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868–877

Huang X, Wang J, Aluru S, Yang SP, Hillier L (2003) PCAP: a whole-genome assembly program. Genome Res 13:2164–2170

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y et al (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41:1275–1281

International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Jain M, Misra G, Patel RK, Priya P, Jhanwar S, Khan AW, Shah N, Singh VK, Garg R, Jeena G et al (2013) A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). Plant J. doi:10.1111/tpj.12173

Jiang H, Wong WH (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24:2395–2396

Khalil AS, Collins JJ (2010) Synthetic biology: applications come of age. Nat Rev Genet 11:367–379

Kim J, Larkin DM, Cai Q, Asan ZY, Ge R-L, Auvil L, Capitanu B, Zhang G, Lewin HA, Ma J (2013) Reference-assisted chromosome assembly. Proc Natl Acad Sci U S A 110:1785–1790

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12

Lam TW, Sung WK, Tam SL, Wong CK, Yiu SM (2008) Compressed indexing and local alignment of DNA. Bioinformatics 24:791–797

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinformatics 11:473–483

Li H, Ruan J, Durbin R (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858

Li R, Li Y, Kristiansen K, Wang J (2008b) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J (2009a) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009b) The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078–2079

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 J, Wang J (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B, Yang B, Fan W (2012) Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. Brief Funct Genomics 11:25–37

Lin H, Zhang Z, Zhang MQ, Ma B, Li M (2008) ZOOM! Zillions of oligos mapped. Bioinformatics 24:2431–2437

Ling H-Q, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y et al (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496:87–90

Ma B, Tromp J, Li M (2002) PatternHunter: faster and more sensitive homology search. Bioinformatics 18:440–445

Manske HM, Kwiatkowski DP (2009) LookSeq: a browser-based viewer for deep sequencing data. Genome Res 19:2125–2132

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74:560–564

McCreight EM (1976) A space-economical suffix tree construction algorithm. J ACM 23:262–272

Meek C, Patel JM, Kasetty S (2003) OASIS: an online and accurate technique for local-alignment searches on biological sequences. In: Proceedings of 29th international conference on Very Large Data Bases (VLDB 2003), Berlin, pp 910–921

Mercer TR, Dinger ME, Mattick JS (2009) Long noncoding RNAs: insights into functions. Nat Rev Genet 10:155–159

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L et al (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318:245–250

Miller J, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. Genomics 14:315–327. doi:10.1016/j.ygeno.2010.03.001

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D (2010) Tablet–next generation sequence assembly visualization. Bioinformatics 26:401–402

Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D (2013) Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform 14:193–202

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL et al (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452:991–996

Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li L-T, Zhang Q, Kim M-J, Schatz MC, Campbell M et al (2013) Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn). Genome Biol 14:R41

Mullikin JC, Ning Z (2003) The phusion assembler. Genome Res 13:81–90

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH et al (2000) A whole-genome assembly of *Drosophila*. Science 287:2196–2204

Myllykangas S, Buenrostro J, Ji HP (2011) Overview of sequencing technology platforms. In: Rodriguez-Ezpeleta N, Hackenberg M, Aransayet AM (eds) Bioinformatics for high throughput sequencing. Springer, New York

Park PJ (2009) Chip-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

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

Pop M, Kosack DS, Salzberg SL (2004) Hierarchical scaffolding with Bambus. Genome Res 14:149–159

Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195

Qiang X, Chen L-L, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao W-B, Hao B-H, Lyon MP et al (2013) The draft genome of sweet orange (*Citrus sinensis*). Nat Genet 45:59–66

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist EA, Kamisugi YA et al (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319:64–69

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol 5(5):e1000386. doi:10.1371/journal.pcbi.1000386

Salmela L, Mäkinen V, Välimäki N, Ylinen J, Ukkonen E (2011) Fast scaffolding with small independent mixed integer programs. Bioinformatics 27:3259–3265

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

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K et al (2008) Genome structure of the legume, *Lotus japonicus*. DNA Res 15:227–239

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA et al (2009) The B73 maize genome: complexity, diversity and dynamics. Science 326:1112–1115

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP et al (2011) The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 43:109–116

Simpson JT, Durbin R (2010) Efficient construction of an assembly string graph using the FM-index. Bioinformatics 26:i367–i373

Simpson JT, Durbin R (2012) Efficient de novo assembly of large genomes using compressed data structures. Genome Res 22:549–556

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

Smith AD, Chung WY, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ (2009) Updates to the RMAP short-read mapping software. Bioinformatics 25:2841–2842

Song L, Crawford GE (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. doi:10.1101/pdb.prot5384

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12:1599–1610

The International Barley Genome Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641

van Bakel H, Stout J, Cote A, Tallon C, Sharpe A, Hughes T, Page J (2011) The draft genome and transcriptome of *Cannabis sativa*. Genome Biol 12:R102

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan G, Whaley AM et al (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotech 30:83–89

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2(12):e1326. doi:10.1371/journal.pone.0001326

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D et al (2010) The genome of the domesticated apple (*Malus* X *domestica* Borkh.). Nat Genet 42:833–839

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S et al (2012) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44:1098–1103

Weese D, Emde AK, Rausch T, Döring A, Reinert K (2009) RazerS–fast read mapping with sensitivity control. Genome Res 19:1646–1654

Yu J, Hu SN, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. *ssp indica*). Science 296:79–92

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

Zhu Z, Niu B, Chen J, Wu S, Sun S, Li W (2013) MGAviewer: a desktop visualization tool for analysis of metagenomics alignment data. Bioinformatics 29:122–123