 doi:10.1093/bfgp/elr045

Current state-of-art of sequencing technologies for plant genomics research

Mahendar Thudi, Yupeng Li, Scott A. Jackson, Gregory D. May and Rajeev K. Varshney

Abstract

A number of next-generation sequencing (NGS) technologies such as Roche/454, Illumina and AB SOLiD have recently become available. These technologies are capable of generating hundreds of thousands or tens of millions of short DNA sequence reads at a relatively low cost. These NGS technologies, now referred as second-generation sequencing (SGS) technologies, are being utilized for *de novo* sequencing, genome re-sequencing, and whole genome and transcriptome analysis. Now, new generation of sequencers, based on the 'next-next' or third-generation sequencing (TGS) technologies like the Single-Molecule Real-Time (SMRT™) Sequencer, Heliscope™ Single Molecule Sequencer, and the Ion Personal Genome Machine™ are becoming available that are capable of generating longer sequence reads in a shorter time and at even lower costs per instrument run. Ever declining sequencing costs and increased data output and sample throughput for NGS and TGS sequencing technologies enable the plant genomics and breeding community to undertake genotyping-by-sequencing (GBS). Data analysis, storage and management of large-scale second or TGS projects, however, are essential. This article provides an overview of different sequencing technologies with an emphasis on forthcoming TGS technologies and bioinformatics tools required for the latest evolution of DNA sequencing platforms.

Keywords: *next-generation sequencing technology; sequencing by synthesis; single molecule sequencing; plant genomics; genotyping-by-sequencing; genomic selection*

INTRODUCTION

Affordable personal genomes have been a motivation for the development of low cost, high-throughput next-generation sequencing (NGS) technologies, including Roche/454 (www.454.com/), Illumina (www.illumina.com/) and AB SOLiD, (www.appliedbiosystems.com/) which are able to generate three to four orders of magnitude more DNA sequence than Sanger-based sequencing [a first-generation sequencing (FGS) technology] on the ABI 3730xl platform. These NGS technologies have enabled the genomics community to

Corresponding author. Rajeev K. Varshney, ICRISAT, Patancheru 502 324, Greater Hyderabad, India. Tel: 0091 40 30713305; Fax: 0091 40 3071 3074/3075, E-mail: r.k.varshney@cgiar.org

Mahendar Thudi is a Post-Doctoral Visiting Scientist, actively involved in development and application of genomic resources in chickpea and leading the activities in the Centre of Excellence in Genomics at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in India.

Yupeng Li is a Ph.D. student, actively involved in next-generation sequencing data analysis including *de novo* genome sequencing and comparative genomics.

Scott Jackson is a Professor in plant genetics and genomics at the University of Georgia, Athens. His research has major focus on genomics of legumes and rice and, the exploration and utilization of genetic diversity in wild relatives of crop plants and on genome structure and evolution.

Gregory May is President of the National Centre for Genome Resource (NCGR) and Director of the New Mexico Genome Center, Santa Fe, New Mexico. He has extensive research experience in the area of genome structure and organization, gene targeting and DNA repair in plants.

Rajeev Varshney is Director, Centre of Excellence in Genomics at ICRISAT and Leader of Comparative and Applied Genomics Theme of CGIAR Generation Challenge Programe (GCP). His research has major focus on development and application of large-scale genomic and transcriptomic resources in legume crops by using next-generation sequencing and high-throughput genotyping technologies.

comprehensively characterize DNA sequence variation within a species by sequencing multiple accessions/genotypes [1, 2], *de novo* sequencing of a number of species [3, 4], detection of methlylated regions in genome [5] and gene expression profiling [6–8].

In the past, Sanger sequencing has been used to characterize the genomes of several organisms including model plants as well as major crop species like rice, soybean, sorghum, maize, grape and eucalyptus (www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_p1.shtml). The availability of NGS technologies, however, have enabled the research community to embark upon sequence the genomes of thousands of plant species through the undertaking of the 1000 Plant Genomes Project (www.onekp.com/), the 1001 Arabidopsis Genome project (www.1001genomes.org/) and the 1000 Plant and Animal Genome Project (www.ldl.genomics.cn/). Similarly, the Genome 10 K Project has been conceived to sequence and assemble 10 000 vertebrate genomes including at least one from each genus (www.genome10k.org/).

Advances in nanobiology and robotics for DNA sequencing applications have also been driven by a competition to win the race of sequencing a human genome at a target price of US $1000 and therefore new sequencing technologies and platforms continue to emerge. As a result, existing NGS technologies are referred as second-generation sequencing (SGS) technologies and future or very recently available NGS technologies are referred to as third-generation sequencing (TGS) or 'next-next' generation sequencing (NNGS) technologies. This article provides an overview of different sequencing technologies with a major emphasis on the forthcoming TGS technologies. We also discuss different bioinformatics tools required for data analysis of the massive amounts of sequence data emerging from these technologies.

FGS PLATFORMS

The FGS methods include Sanger's enzymatic dideoxy DNA sequencing [9] and the Maxam and Gilbert's chemical degradation methods [10]. For commercial DNA sequencing, Applied Biosystems (www.appliedbiosystems.com/) was the first company to introduce ABI Prism 377 based on slab gel electrophoresis. Owing to the inconvenience of casting gels, the ABI Prism 3700 was developed with automated reloading of the 96 capillaries with a polymer matrix. This platform was used in the sequencing of the first human genome [11]. Sanger sequencing was also used for sequencing genomes of several plant species such as *Arabidopsis (Arabidopsis thaliana*; [12]), rice (*Oryza sativa*; [13, 14]), sorghum (*Sorghum bicolor*; [15]), grapes (*Vitis vinifera*; [16]), poplar (*Populus trichophora*; [17]) and soybean (*Glycine max*; [18]). Sequencing time and personnel costs associated with Sanger sequencing, however, prohibited the sequencing of a large number of plant species, especially those with large, complex genomes (e.g. wheat, ~16 Gb).

SGS PLATFORMS

In 2005, 454 Life Sciences (www.454.com/) launched the GS 20, the first NGS systems into the market. After acquiring 454 Life Sciences, Roche Applied Science (www.roche-applied-science.com/) extended this technology to the new version of the 454 instrument, the GS FLX titanium. Subsequently, Roche/454 launched several other platforms including GS 20/FLX, GS FLX Titanium+, GS FLX Titanium XLR70 and GS Junior (www.454.com/products/). In parallel, several other companies launched competing NGS systems that included 'Solexa 1G' (later named 'Genome Analyzer'), GA, GA II, HiSeq 2000, HiSeq 1000, Hi ScanSQ and MiSeq by Illiumina Inc. (www.illumina.com/systems.ilmn); SOLiD™ 3 and SOLiD™ 4 system by Applied Biosystems (www.appliedbiosystems.com/). Recently a new system for NGS based on multiplex polony technology [19] named as the Polonator G.007 has been introduced by Dover and Harvard Medical School (www.polonator.org/). Currently, these technologies are referred as SGS systems. Although all these systems can be used for a multitude of applications for plant genomics research [20], Illumina and Roche/454 have been the most widely adopted SGS platforms as evident by publications.

Illumina and Roche/454 employ the principle of sequencing by synthesis (SBS) i.e. they rely on PCR to amplify a given DNA template which is then attached to a solid surface and are subsequently imaged in a phased approach. On the other hand, sequencing platforms like SOLiD™ 3 and SOLiD™ 4 employ sequencing by ligation (SBL). However, the amplification process can introduce errors in the template sequence as well as introduce amplification bias. In addition, generation of NGS

data takes several days due to a large number of instrument scanning and washing cycles. Because of dephasing [21], as compared to Sanger sequencing, average read length of sequence reads produced by SGS platforms is shorter [22, 23]. Based on the throughput achieved, except MiSeq, all other sequencing platforms from Illumina (GA, GA II, HiSeq 2000, HiSeq 1000 and Hi ScanSQ), Applied Biosystems (SOLiD^{TM} 3 and SOLiD^{TM} 4) and Roche/454 (GS 20/FLX, GS FLX Titanium+, GS FLX Titanium XLR70 and GS Junior) are considered as high-throughput SGS platforms.

The GS FLX from 454 Life Sciences produces over a million reads of up to 1000 bases per 10 h run, for a total yield of 400–600 megabases. Thus, 454 Sequencer has longest short reads among all SGS platforms. The Illumina Genome Analyzer yields over one hundred million high-quality short reads (up to 76 bases) per 3–5 day run, totaling several gigabases of aligned sequence. To date, the majority of published NGS articles have described methods using the short sequence data produced with the Genome Analyzer. At present, the new Illumina HiSeq 2000 Genome Analyzer is capable of producing single reads of 2×100 bp (pair-end reads), and generates $\sim$200 Gb of short sequences per run. The raw base accuracy is $>$99.5%. Finally, the Applied Biosystems SOLiD system also produces hundreds of millions of short reads (up to 50 bases) per run.

The large amount of data generated by these high-throughput SGS technologies poses a challenge for data storage and transfer and informatics operations. This is especially true for the shorter reads generated by the Illumina and SOLiD systems that make sequence alignment and assembly processes challenging [23]. Nevertheless, SGS technologies are being used for a variety of applications including *de novo* sequencing of genomes, transcriptome analysis, gene expression, marker discovery and many others in plant species such as cocoa [24], chickpea [6, 25], pigeonpea [26–28], date palm [29] and pea [30].

TGS TECHNOLOGIES

In the context of challenges associated with assembling of short sequence reads, development of technologies that generate longer sequence reads will help to deliver the information required for assembling complex genomes. In addition, as SGS platforms generally require either an *in vitro* or *in vivo* amplification step, technologies those that directly sequence single molecules of DNA, eliminating the need for costly and many times problematic procedures like cloning and PCR amplification are preferred [31]. To this end, a number of academic and commercial efforts are developing ultra-low-cost 'single-molecule' sequencing (SMS) technologies. SMS technologies can be grouped into three categories: (i) fluorescence-based methods for SMS like exonucleolytic degradation, true single-molecule sequencing (tSMS^{TM}), fluorescence resonance energy transfer (FRET)-based approach, single-molecule real-time sequencing (SMRT^{TM}) and microfluidic devices; (ii) non-fluorescent sequencing systems like Nanopore's Nano-edges, sequencing using transmission electron microscopy, pyrosequencing, motion-based sequencing and scanning tunneling spectroscopy-based sequencing; and (iii) Raman-based methods such as sequencing using surface-enhanced Raman spectroscopy (SERS) and sequencing using tip-enhanced Raman spectroscopy (TERS, [32]. These technologies are referred as TGS. TGS technologies seem to be superior to SGS technologies as they generate longer sequence reads in higher throughput fashion and faster turnaround time with higher consensus accuracy. Some of these platforms that have potential for extensive use in plant genomics research are given below.

SMRT^{TM} SEQUENCER

Pacific Biosciences (www.pacificbiosciences.com/) company has recently introduced the PacBio*RS*, the TGS system that employs the SMRT^{TM} DNA sequencing technology, where in DNA sequencing is performed on SMRT cells (nanofabricated consumable substrates). In this technology, DNA fragment is sequenced by a single DNA polymerase molecule that is attached to the bottom of each zero-mode waveguides (ZMW, [33]) and as a result, each DNA polymerase resides at detection zone of ZMW [33, 34]. As per the company, PacBio*RS* requires less than a day from sample preparation to obtaining the sequence information and produces read lengths $>$ 1000 bp. The SMRT^{TM} sequencer has been available to several sequencing centers and critical assessment on its performance is ongoing.

HELISCOPE™ SINGLE MOLECULE SEQUENCER

This sequencer has been introduced by Helicos (www.helicosbio.com/) company that images billions of single molecules and produces 21–35 Gb per run, almost 100X greater than Sanger methods, and faster than many currently available NGS technologies [35, 36]. HeliScope employs true single-molecule sequencing (tSMS) chemistry [37] and direct RNA sequencing chemistries. Large numbers of strands of single DNA molecule can be sequenced simultaneously by using tSMS chemistry. tSMS has been used to sequence an individual human genome [38], re-sequence the M13 virus genome and to quantify the yeast transcriptome [39, 40]. A drawback is a relatively high raw sequence error rate that can be overcome with repetitive sequencing, but increases the cost per base for a given accuracy rate, offsetting some of the gains from lower reagent costs.

ION PERSONAL GENOME MACHINE™ SEQUENCER

Life Technologies company has recently launched Ion Personal Genome Machine (PGM™) Sequencer based on the ion torrent semiconductor technology (www.iontorrent.com/technology/). This technology is based on a biochemical process by which a hydrogen ion is released as a nucleotide and is incorporated into a strand of DNA by a polymerase [41]. This technology is independent of enzymatic reactions, fluorescence, chemi-luminiscence, and optics. It uses a high-density array of micro-machined wells. Each of these wells hold a different DNA template. Just beneath the wells, there is an ion-sensitive layer and a proprietary Ion sensor. During the sequencing, when a new base is added to the template, and incorporated into the strand, hydrogen ion is released. The charge from that ion changes the pH of the solution that can be detected directly by the ion sensor without imaging. The PGM™ system can perform a wide range of sequencing applications including multiplexing amplicons, transcriptome analysis, small RNA discovery, and ChIP-Seq analysis.

COMPARISION OF DIFFERENT SGS AND TGS TECHNOLOGIES

A suite of SGS and TGS technologies are currently available. Some technologies are already commercialized and are on the market, while commercialization of some sequencing technologies has not yet been realized. An effort has been made to compare different sequencing technologies in Table 1. SGS technologies rely upon SBL or SBS, including pyrosequencing and reversible chain termination. Among SGS technologies, the Genome Sequencer FLX from 454 Life Sciences/Roche, Illumina Genome Analyzer and Applied Biosystems SOLiD are widely deployed in hundreds of research laboratories across the world. These technologies vary in terms of template size and construct, read-length, and throughput thereby making comparisons difficult. In fact, some of these platforms are powerful in particular niches of the sequencing market.

The GS FLX from 454 Life Sciences produces over a million reads of up to 1000 bases per 10 h run, for a total yield of 400–600 Mb. Thus, the 454 sequencer has longest reads of the SGS platforms. The Illumina HiSeq 2000 Genome Analyzer is capable of producing single reads of 2×100 bp (pair-end reads) ~200 Gb of sequences per run. The raw base accuracy is >99.5%. Finally, the AB SOLiD system also produces hundreds of millions of short reads (up to 50 bases) per run.

Whole-genome shotgun (WGS) sequencing is challenging for larger genomes [42]. The primary reason is the abundance of repetitive sequence in larger genomes, especially true for plant genomes. However, the combination of read length and paired reads spanning quite large distances achieved through Sanger-based sequencing platforms can be used effectively by assembly algorithms to resolve many repeats and reconstruct a draft genome sequence [43]. However, current NGS platforms are unable to deliver both these features and thus cannot effectively span repeats.

Although some of the TGS technologies promise to improve read lengths, they differ significantly in their approach to sequencing and in their throughput and time taken from sample preparation to result. Of the TGS technologies, the tSMS and Ion Torrent technologies have already been commercialized and several others are in the process of commercialization. A newer approach is that taken by Complete Genomics (www.completegenomics.com/) where the sequencing platform is not commercially available but available in-house and being used extensively for human sequencing.

A majority of TGS technologies do not require cloning and amplification thereby eliminating part of

Table 1: Comparison of different sequencing platforms*

Platform	Sequencer	Webpage	Throughput	Read length	Accuracy (%)	Run time
SGS platforms						
Roche/454 sequencing	GS Junior System	http://454.com/products/gs-junior-system/index.asp	~35 Mb	~400 bp	99	10 h
	GS FLX Titanium XL+	http://454.com/products/gs-flx-system/index.asp	700 Mb	Up to 1000 bp	99.997	23 h
	GS FLX Titanium XLR70	http://454.com/products/gs-flx-system/index.asp	450 Mb	Up to 600 bp	99.995	10 h
Soleaxa/Illumina sequencing	HiSeq 2000	http://www.illumina.com/systems/hiseq.2000.ilmn	Up to 600 Gb	2 × 100 bp	>85 (2 × 50 bp); >80 (2 × 100 bp)	1.5–11 days
	HiSeq 1000	http://www.illumina.com/systems/hiseq.1000.ilmn	Up to 300 Gb	2 × 100 bp	>85 (2 × 50 bp); >80(2 × 100 bp)	1.5–8.5 days
	Genome Analyzer IIx	http://www.illumina.com/systems/genome.analyzer.iix.ilmn	95 Gb	2 × 150 bp	>85 (2 × 50 bp); >80 (2 × 100 bp)	2–14 days
	MiSeq	http://www.illumina.com/systems/miseq.ilmn	>1 Gb	2 × 150 bp	>85 (2 × 50 bp); >80 (2 × 100 bp)	8 h
Applied Biosystems sequencing	5500 System	http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html	7–9 Gb/day	Mate-paired: 2 × 60 bp; Paired-end: 75 bp × 35 bp; Fragment: 75 bp	Up to 99.99	2–8 days
	5500xl System (1.0 μm microbeads)	http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html	10–15 Gb/day	Mate-paired: 2 × 60 bp; Paired-end: 75 bp × 35 bp; Fragment: 75 bp	Up to 99.99	2–8 days
	5500xl System (0.75 μm nanobeads available 2nd half of 2011)	http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html	>20 Gb/day	Fragment: 50 bp	Up to 99.99	2-8 days
Multiplex Polony technology	Polonator G.007	http://www.polonator.org/	10–35 Gb	28 bases (14 + 14 paired-end reads)	98	4 days
TGS platforms						
SMRT^{TM} sequencer	PACBIO *RS*	http://www.pacificbiosciences.com/	100 Mb	2000 bp	85	1 h
Heliscope^{TM} single molecule sequencer	Heliscope^{TM}Sequencer	http://www.helicosbio.com/	21–35 Gb	55 bp	–	8 days
Ion Torrent sequencing technology	PGM^{TM} Sequencer	http://www.iontorrent.com/technology/	>10 Mb (314 chip) >100 Mb (316 chip) >1 Gb (318 chip)	>400 bp (in 2012) >200 bp (in 2011)	99.99 consensus accuracy; 99.5 raw accuracy	<2 h

*Features of different sequencing platforms have been compiled from the websites of respective companies.

Table 2: Features of some important tools for analysis of NGS data*

Tool/program	Features	References
De novo alignment		
ABySS	*De novo* sequence assembler designed for aligning very short reads. The single-processor version is useful for assembling genomes up to 40–50 Mb in size.	http://www.bcgsc.ca/platform/bioinfo/software/abyss
EULER-SR	Short read *de novo* assembly, uses a *de Bruijn* graph approach.	http://euler-assembler.ucsd.edu/portal/
MIRA2	MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid *de novo* assemblies using reads gathered through 454 sequencing technology (GS20 or GS FLX). Compatible with 454, Illumina and Sanger data. Linux OS required.	http://chevreux.org/projects_mira.html
SSAKE	Short Sequence Assembly by K-mer search and 3′-read Extension (SSAKE) for aggressively assembling millions of short nucleotide sequences by progressively searching for perfect 3′-most k-mers using a DNA prefix tree.	http://www.bcgsc.ca/platform/bioinfo/software/ssake
SOAP*denovo*	Part of the SOAP suite (see below).	http://soap.genomics.org.cn/
VCAKE	*De novo* assembly of short reads with robust error correction. An improvement on early versions of SSAKE.	http://sourceforge.net/projects/vcake/
Velvet	*De novo* genomic assembler specially designed for short read sequencing technologies, such as Illumina or 454. Need ∼20–25 × coverage and paired reads.	http://www.ebi.ac.uk/%7Ezerbino/velvet/
Alignment to a reference genome		
Bowtie	Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per h on a typical workstation with 2 GB of memory. Uses a Burrows–Wheeler-Transformed (BWT) index.	http://bowtie-bio.sourceforge.net/index.shtml
Exonerate	Offers various forms of pairwise alignment of DNA/protein against a reference.	http://www.ebi.ac.uk/∼guy/exonerate/
GenomeMapper	A short read mapping tool designed for accurate read alignments. It quickly aligns millions of reads either with ungapped or gapped alignments.	http://100genomes.org/downloads/genomemapper.html
GMAP	For aligning mRNA and EST Sequences.	http://research-pub.gene.com/gmap/
MAQ	Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina with preliminary functions to handle ABI SOLiD data.	http://sourceforge.net/projects/maq/
PASS	Allows the users to modulate very finely the sensitivity of the alignments.	http://pass.cribi.unipd.it/cgi-bin/pass.pl
RMAP	Assembles 20–64 bp reads to a FASTA reference genome.	http://rulai.cshl.edu/rmap/
SeqMap	Supports up to 5 or more bp mismatches/INDELs. Highly tuneable.	http://seqmap.compbio.iupui.edu/
SHRiMP	Assembles to a reference sequence.	http://compbio.cs.toronto.edu/shrimp/
Slider	An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a reference sequence or a set of reference sequences.	http://www.bcgsc.ca/platform/bioinfo/software/slider
SOAP	SOAP (Short Oligonucleotide Alignment Program) is a program for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The updated version uses a BWT. Can call SNPs and INDELs.	http://soap.genomics.org.cn/
SSAHA	SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases using a hash table. Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.	http://www.sanger.ac.uk/resources/software/
Vmatch	A versatile software tool for efficiently solving large-scale sequence matching tasks.	http://www.vmatch.de/
Zoom	ZOOM is highly accurate, flexible, and user-friendly with speed being a critical priority. Enables to map millions of short reads, emerged by NGS technology, back to the reference genomes, and carry out post-analysis.	http://www.bioinformaticssolutions.com/all-products/zoom/index.php
SNP/Indel Discovery		
ssahaSNP	A polymorphism detection tool that detects homozygous SNPs and indels by aligning shotgun reads to the finished genome sequence. Highly repetitive elements are filtered out by ignoring those k-mer words with high occurrence numbers. More tuned for ABI Sanger reads.	http://www.sanger.ac.uk/resources/software/
PolyBayesShort	This version is specifically optimized for the analysis of large numbers (millions) of high-throughput next-generation sequencer reads, aligned to whole chromosomes of model organism or mammalian genomes.	http://bioinformatics.bc.edu/marthlab/PbShort

(continued)

Table 2: Continued

Tool/program	Features	References
PyroBayes	A novel base caller for pyrosequences from the 454 Life Sciences sequencing machines. It was designed to assign more accurate base quality estimates to the 454 pyrosequences.	http://bioinformatics.bc.edu/marthlab/PyroBayes
Alpheus	Pair-wise alignments use BioJava MegaBLAST and Java GMAP parsers; alignments to reference databases; variant detection (SNPs and indels)	http://alpheus.ncgr.org/technical-overview.jsp
Transcriptomics		
G-Mo.R-Se	G-Mo.R-Se is a method aimed at using RNA-Seq short reads to build *de novo* gene models.	http://www.genoscope.cns.fr/externe/gmorse/
MapNext	Useful for (i) unspliced alignment and clustering of reads, (ii) spliced alignment of transcriptomic reads, (iii) SNP detection and calculation of SNP frequency from population sequences and (iv) storage of result data into database to make it available for more flexible query and further analyses.	http://evolution.sysu.edu.cn/english/software/mapnext.htm
QPalma	Optimal Spliced Alignments of Short Sequence Reads. Is an easy-to-use and flexible tool to accurately and efficiently align both transcriptome reads (spliced and unspliced) from RNA-Seq experiments against a reference genome.	http://www.fml.tuebingen.mpg.de/raetsch/suppl/qpalma
TopHat	TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions.	http://tophat.cbcb.umd.edu/
Genome annotation/genome browser/alignment viewer/assembly database		
EagleView	An information-rich genome assembler viewer. It can display a dozen different types of information including base quality and flowgram signal.	http://bioinformatics.bc.edu/marthlab/EagleView
LookSeq	LookSeq is a web-based application for alignment visualization, browsing and analysis of genome sequence data. Supports multiple sequencing technologies, alignment sources, and viewing modes; low or high-depth read pileups; and easy visualization of putative single nucleotide and structural variation.	http://www.sanger.ac.uk/resources/software/
MapView	Enables visualization of short reads alignment on desktop computer.	http://evolution.sysu.edu.cn/mapview/
SAM	Sequence Assembly Manager (SAM) is a whole-genome assembly (WGA) management and visualization Tool. It provides a generic platform for manipulating, analyzing and viewing WGA data, regardless of input type.	http://www.bcgsc.ca/platform/bioinfo/software/sam
XMatchView	A visual tool for analyzing cross match alignments.	http://www.bcgsc.ca/platform/bioinfo/software/xmatchview
Miscelleneous		
CNV-Seq	For detection of copy number variation using high-throughput sequencing.	http://tiger.dbs.nus.edu.sg/cnv-seq/
FindPeaks	Perform analysis of ChIP-Seq experiments.	http://www.bcgsc.ca/platform/bioinfo/software/findpeaks
MACS	Model-based Analysis for ChIP-Seq.	http://liulab.dfci.harvard.edu/MACS/
PeakSeq	PeakSeq is a program for identifying and ranking peak regions in ChIP-Seq experiments.	http://info.gersteinlab.org/PeakSeq
SISSRs	For precise identification of genome-wide transcription factor binding sites from ChIP-Seq data.	http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/

*Features of different tools/programs have been compiled from their respective websites.

the cost relative to SGS technologies. In addition, read lengths from TGS technologies are expected to be around 1 kb and longer read lengths will ease many of the informatics challenges relating to *de novo* assembly that are currently encountered.

BIOINFORMATICS TOOLS

The increase in sequence throughput from different sequencing platforms is exponential (Table 2). In this context, storage and management of humongous datasets is very challenging. In addition to data storage and management, primary, secondary and tertiary analysis solutions like quality control, base calling, *de novo* assembly, alignment to a reference genome, variant calling, Chip-Seq, transcriptome analysis are necessary to make sense of the larger volumes of sequence data. As existing sequence analysis tools were not appropriate for analysis of sequence data coming out from new sequencing technologies, a number of tools/software packages have been developed in last few years. Some of these tools are listed in Table 2. The bioinformatics community needs to be ready continually develop new tools as well as data storage and management systems in anticipation of even larger amount of sequence data coming out from TGS technologies. Moreover, the types of data and quality associated with each will complicate analyses and the use of existing tools. Cloud computing is a potential solution to the question of massive data storage as well as analysis [44]. Cloud computing provides computation, software, data access and storage services that do not require end-user knowledge of the physical location and configuration of the system that delivers the services.

SUMMARY AND OUTLOOK

The last 5 years have witnessed the rise of massively parallel sequencing technologies and a revolution in both plant and animal genomics research. These technologies are continuously evolving resulting in a continuous decline in sequencing cost and an increase in sequence read lengths. The result of this evolution is that genotyping-by-sequencing (GBS) will be routine in a few years [2]. As a result, plant genetics and breeding will benefit from modern genetics and breeding approaches like association mapping [45], allele mining, domestication and genomic selection [46]. The potential as well as proof-of-concept of these new sequencing technologies for a variety of applications has been discussed in several other articles in this Special Issue of this journal.

Key Points

- SGS technologies dramatically reduced the cost of sequencing.
- TGS technologies are poised to generate longer sequence reads at a very low cost in less time.
- A new generation of DNA sequencing platforms is ready for commercialization that will change the landscape of sequencing in plant genomics research.
- Although a large number of tools/software packages are available for analysis, visualization and storage of sequence data, there is a need to develop more powerful and efficient tools/platforms.

FUNDING

CGIAR Generation Challenge Programme, Mexico; US National Science Foundation (BIO 0822258) and Centre of Excellence (CoE) grant from Department of Biotechnology, Government of India.

References

1. Siu H, Zhu Y, Jin L, *et al*. Implication of next-generation sequencing on association studies. *BMC Genomics* 2011;**12**:322.
2. Elshire RJ, Glaubitz JC, Sun Q, *et al*. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 2011;**6**:e19379.
3. Li R, Fan W, Tian G, *et al*. The sequence and *de novo* assembly of the giant panda genome. *Nature* 2010;**463**:311–17.
4. Kuczynski J, Costello EK, Nemergut DR, *et al*. Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol* 2010;**11**:210.
5. Wu G, Yi N, Absher D, *et al*. Statistical quantification of methylation levels by next-generation sequencing. *PLoS ONE* 2011;**6**:e21034.
6. Hiremath PJ, Farmer A, Cannon SB, *et al*. Large-scale transcriptome analysis in chickpea (*Cicer arietinum* L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. *Plant Biotech J* 2011;**8**:922–31.
7. Varshney RK, Hiremath PJ, Lekha PT, *et al*. A comprehensive resource of drought- and salinity- responsive ESTs for gene discovery and marker development in chickpea (*Cicer arietinum* L.). *BMC Genomics* 2009a;**10**:523.
8. Wall PK, Leebens-Mack J, Chanderbali AS, *et al*. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* 2009;**10**:347.
9. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain terminating inhibitors. *Proc Natl Acad Sci USA* 1977;**74**:5463–7.

10. Maxam AM, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci USA* 1977;**74**:560–4.
11. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–73.
12. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000;**408**:796–815.
13. Goff S, Ricke D, Lan TH, *et al*. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 2002;**296**: 92–100.
14. Yu J, Hu S, Wang S, *et al*. A draft sequence of the rice genome (*Oryza sativa L. ssp. indica)*. *Science* 2002;**296**:79–92.
15. Paterson A, Bowers J, Bruggmann R, *et al*. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 2009;**457**:551–6.
16. Jaillon O, Aury JM, Noel B, *et al*. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007;**449**:463–7.
17. Tuskan G, DiFazio , Jannson S, *et al*. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006;**313**:1596–1604.
18. Schmutz J, Cannon S, Schlueter J, *et al*. Genome sequence of the palaeopolyploid soybean. *Nature* 2010; **463**:178–83.
19. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45.
20. Varshney RK, Nayak SN, May GD, *et al*. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol* 2009b;**27**:522–30.
21. Whiteford N, Skelly T, Curtis C, *et al*. Swift: primary data analysis for the Illumina/Solexa sequencing platform. *Bioinformatics* 2009;**25**:2194–9.
22. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;**1**:31–46.
23. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. *Genome Res* 2010;**20**:1165–73.
24. Argout X, Salse J, Aury J-M, *et al*. The genome of *Theobroma cacao*. *Nat Genet* 2011;**43**:101–8.
25. Garg R, Patel RK, Tyagi AK, *et al*. *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res* 2011b;**18**:53–63.
26. Varshney RK, Chen W, Li Y, *et al*. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat Biotechnol* 2012;**30**:83–89.
27. Dubey A, Farmer A, Schlueter J, *et al*. Defining the transcriptome assembly and its use for genome dynamics and transcriptome profiling studies in pigeonpea (*Cajanus cajan* L.). *DNA Res* 2011;**18**:153–64.
28. Kudapa H, Bharti AK, Cannon SB, *et al*. A comprehensive transcriptome assembly of pigeonpea (*Cajanus cajan* L.) using Sanger and second-generation sequencing platforms. *Mol Plant* 2012; doi:10.1093/mp/ssr111.
29. Al-Dous EK, George B, Al-Mahmoud ME, *et al*. *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat Biotechnol* 2011;**29**:521–7.
30. Franssen SU, Shrestha RP, Bräutigam A, *et al*. Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. *BMC Genomics* 2011;**12**:227.
31. Gupta PK. Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* 2009;**26**: 602–11.
32. Treffer R, Deckert V. Recent advances in single-molecule sequencing. *Curr Opin Biotechnol* 2010;**21**:1–8.
33. Levene MJ, Korlach J, Turner SW, *et al*. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 2003;**299**:682–6.
34. Eid J, Fehr A, Gray J, *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* 2009;**323**:133–8.
35. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011a;**12**: 87–98.
36. Ozsolak F, Milos PM. Transcriptome profiling using single molecule direct RNA sequencing. *Methods Mol Biol* 2011b; **733**:51–61.
37. Braslavsky I, Hebert B, Kartalov E, *et al*. Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci USA* 2003;**100**:3960–4.
38. Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 2009;**27**:847–50.
39. Harris TD, Buzby PR, Babcock H, *et al*. Single molecule DNA sequencing of a viral genome. *Science* 2008;**320**: 106–9.
40. Lipson D, Raz T, Kieu A, *et al*. Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* 2009;**27**:652–8.
41. Rothberg JM, Hinz W, Rearick TM, *et al*. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011;**475**:348–52.
42. Feuillet C, Leach JE, Rogers J, *et al*. Crop genome sequencing: lessons and rationales. *Trends Plant Sci* 2011;**16**:77–88.
43. Jaffe DB, Butler J, Gnerre S, *et al*. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 2003;**13**:91–96.
44. Stein LD. The case for cloud computing in genome informatics. *Genome Biol* 2010;**11**:207.
45. Poland JA, Bradbury PJ, Buckler ES, *et al*. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc Natl Acad Sci USA* 2011; **108**:6893–8.
46. Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* 2010;**9**:166–77.