American Journal of Botany

American Journal of Botany 99(2): 175–185 2012.

Applications of next-generation sequencing in plant biology[1]

Ashley N. Egan[2,5], Jessica Schlueter[3], and David M. Spooner[4,5]

[2]East Carolina University, Department of Biology, Howell Science Complex N303a, Mailstop 551, Greenville, North Carolina 27858 USA; [3]University of North Carolina–Charlotte, Department of Bioinformatics and Genomics, 9201 University City Boulevard, Bioinformatics Building 261, Charlotte, North Carolina 28223 USA; and [4]USDA, Agricultural Research Service, Department of Horticulture, University of Wisconsin, 1575 Linden Drive, Madison, Wisconsin 53706-1590 USA

The last several years have seen revolutionary advances in DNA sequencing technologies with the advent of next-generation sequencing (NGS) techniques. NGS methods now allow millions of bases to be sequenced in one round, at a fraction of the cost relative to traditional Sanger sequencing. As costs and capabilities of these technologies continue to improve, we are only beginning to see the possibilities of NGS platforms, which are developing in parallel with online availability of a wide range of biological data sets and scientific publications and allowing us to address a variety of questions not possible before. As techniques and data sets continue to improve and grow, we are rapidly moving to the point where every organism, not just select "model organisms", is open to the power of NGS. This volume presents a brief synopsis of NGS technologies and the development of exemplary applications of such methods in the fields of molecular marker development, hybridization and introgression, transcriptome investigations, phylogenetic and ecological studies, polyploid genetics, and applications for large genebank collections.

Key words: development; DNA sequencing; genomics; next-generation sequencing; plant biology.

All biological disciplines that depend on DNA sequence data have been fundamentally changed in the last few years, driven by the development and emergence of next-generation sequencing (NGS) technologies. Our knowledge of biology, particularly genomics, has grown exponentially. NGS methods have increased capabilities far beyond that of traditional Sanger sequencing (Sanger et al., 1977), allowing millions of bases to be sequenced in one round at a fraction of the cost. As the costs and capabilities of these technologies continue to improve, whole new fields of study are being opened, allowing us to analyze a variety of data sets and approach questions never possible before.

Improved sequencing is accompanied by many challenges as well as new developments. NGS technologies create a vast amount of data, presenting many problems to computational biologists, bioinformaticians, and end-users endeavoring to assemble and analyze NGS data in novel ways. A comprehensive discussion of these challenges is outside the scope of this review, but several papers in this issue address some of these issues and possible strategies for dealing with them (e.g., Cronn et al., 2012; Grover et al., 2012; Ilut et al., 2012; Kvam et al., 2012). Furthermore, these advancements have led to the requirement by most granting agencies and peer-reviewed journals that DNA sequence and other data are made publicly available online. This requirement has thus created the need for the development of databases to house and allow public access to NGS data (see McCouch et al., 2012, in this issue) as well as a sustained commitment to fund and maintain such resources.

One of us (D. Spooner) received his degree in prior rounds of "next-generation" techniques (e.g., flavonoid chemistry and isozymes for systematic studies) and has seen many techniques develop since then (such as DNA restriction site phylogenies, Sanger sequencing, molecular marker developments). In the early stages of development, NGS was initially used by researchers such as plant breeders, working with model organisms of economic importance and backed by decades of research compiled within large genetic and genomic resources. However, techniques are rapidly improving and comparative sequences are accumulating and becoming sufficient to act as reference genomes for assembly, thus allowing any organism to become a "model" organism. These and previous "next-generation" advancements only reinforce the need to continuously keep current in emerging techniques that expand our capabilities and possible avenues of research. This issue addresses recent advancements in NGS techniques and presents studies that point the way to entering this exciting field. Here we briefly review the various NGS methods currently on the market or under development and summarize the 19 papers presented in this special issue wherein a number of research studies illustrate how NGS techniques can be employed to address significant and novel questions in plant biology.

NEXT-GENERATION SEQUENCING TECHNOLOGIES

Next-generation sequencing (NGS) techniques became commercially available around 2005, the first using Solexa sequencing technology. Since then, several different sequencing methods have been developed, all of which are continually being improved at an astonishing rate. These methods can largely be grouped into three main types: sequencing by synthesis, sequencing by ligation, and single-molecule sequencing.

[1]Manuscript submitted 13 January 2012; revision accepted 20 January 2012.

The authors thank *AJB* managing editor A. McPherson and the *AJB* editorial staff for much effort and countless hours put forward to help make this special issue a success and the authors of this volume for reviewing synopses of their articles presented here. A.N.E. is grateful for a Systematics and Biodiversity Science Cluster grant (DEB-1120186) from the National Science Foundation and D.M.S. for grants from the National Science Foundation (NSF DEB 0316614) and USDA National Research Initiative (2008-35300-18669).

[5]Authors for correspondence (egana@ecu.edu; David.Spooner@ars.usda.gov)

doi:10.3732/ajb.1200020

***Sequencing by synthesis*—**Like Sanger sequencing, NGS techniques largely determine base composition through the detection of chemiluminescence created by nucleotide incorporation during synthesis of the complementary DNA strand by DNA polymerase. Sanger sequencing uses dideoxynucleotide chain termination to determine the sequence of nucleotides in a DNA strand, requiring many template fragments of varying sizes. In sequencing by synthesis, DNA is fragmented to the appropriate size, ligated to adaptor sequences, and then clonally amplified to enhance the fluorescent or chemical signal. Templates are then separated and immobilized in preparation for flow-cell cycles. Although the various methods use different chemistry, all use sequential washes of nucleotides along with varying chemistries for fluorescence or chemical detection of nucleotide incorporation. The three methods we group under sequencing by synthesis also differ by read length and how templates are amplified and immobilized.

Roche 454 pyrosequencing—In Roche 454 pyrosequencing (http://www.my454.com), a single, primed DNA template is adhered to a microbead and amplified using emulsion PCR. Each bead is then placed in a well of a PicoTiterPlate, which is put into a flow cell where it is incubated with DNA polymerase, ATP sulfurylase, luciferase, and apyrase along with the substrates luciferin and adenosine 5′-phosphosulfate (ASP). When DNA polymerase incorporates an appropriate dNTP into the new strand, pyrophosphate is released, which is converted to adenosine triphosphate (ATP) in the presence of ASP. ATP then reacts with luciferase to catalyze the conversion of luciferin to oxyluciferin, releasing light in proportion to the amount of ATP produced by dNTP incorporation (Ronaghi et al., 1998; Nyren, 2007). All unused ATP and nucleotides are then removed by apyrase, washed away, and a new chemical mixture is washed over the DNA templates. This procedure is repeated many times until the DNA template has elongated. The fluorescent light produced by nucleotide incorporation is detected by a camera and analyzed to produce the string of nucleotides that is the DNA sequence. Although Roche 454 started out with read lengths of ~100 bp, the technology has steadily improved to where read lengths are now comparable to that produced by Sanger sequencing (~800 bp), producing ~700 Mb total from ~1 million reads (Table 1; http://www.my454.com). Because of its longer reads, this platform is often used in genomic or transcriptomic sequencing when de novo assembly is involved and was employed or reviewed in several studies in this issue (e.g., Buggs et al., 2012; He et al., 2012; Lai et al., 2012; Strickler et al., 2012; Zalapa et al., 2012).

Illumina—Initially developed by Solexa, the Illumina Genome Analyzer uses solid-phase bridge amplification in which 5′ and 3′ adapters are ligated to each end of a DNA template (http://www.illumina.com). One end of the fragment is then attached to the substrate. The adapters hybridize to immobilized forward or reverse primers, creating a bridge that facilitates amplification, generating amplicons that remain attached to the substrate, thus forming clusters of identical templates, which enhances chemiluminescent detection. Millions of such clusters are formed within each channel of the flow cell. Following amplification, the DNA amplicons are denatured and primed. Elongation is conducted through a series of cyclical washes, the first being the addition of a mixture of all four nucleotides, each labeled with a different fluorophore and modified as 3′-*O*-azidomethyl reversible terminators (Barnes et al., 2002; Bentley et al., 2008; Guo et al., 2008). Following image capture, elongation continues after the fluorescent dye moiety is cleaved and the 3′-OH group is restored through reaction with tris(2-carboxyethyl)phosphine. This cycle is repeated until the DNA fragment has been synthesized to its target length. On a dual flow cell, the HiSEq. 2000 system can now produce ~6 billion paired-end reads for a total of 540–600 Gb in ~11 d run time (http://www.illumina.com). This method is currently the most widely used NGS platform and is used or reviewed by most studies in this volume (Azam et al., 2012; Buggs et al., 2012; Gulledge et al., 2012; He et al., 2012; Ilut et al., 2012; Kane et al., 2012; Lai et al., 2012; McKain et al., 2012; Steele et al., 2012; Straub et al., 2012; Strickler et al., 2012; Ward et al., 2012; Zalapa et al., 2012).

Ion Torrent—The Ion Torrent system (http://www.iontorrent.com) is unique among NGS technologies in that the detection for sequencing is not based upon fluorescent dyes but rather measuring the pH change as the result of the release of a H^+ ion upon nucleotide incorporation using semiconductor technology (Rothberg et al., 2011). By sequentially adding nucleotides, the machine is able to detect which nucleotide has been incorporated into the growing strand. There are now two systems available that use this technology, the Ion PGM, for laboratory applications, and the new Ion Proton, which provides higher throughput. The new Proton system is touted to have 165 million sensors with up to a 250-bp read length upon release of the next hardware chip, projected to have 660 million sensors. For both the PGM and the new Proton systems, each hardware chip improvement increases the throughput. With the new 318 chip set, the PGM sequencer can produce over 1000 Mb of sequence with 11.1 million sensors. The other allure of the Ion systems is that sample preparation costs are relatively low compared to other systems. Publications on research that has utilized the Ion Torrent system currently focus on the shotgun sequencing of microbial genomes (e.g., Howden et al., 2011; Rothberg et al., 2011), but this system has clearly made its way into programs pursuing plant-based objectives, and we will see more publications as the Ion Torrent market continues to grow.

***Sequencing by ligation*—**Sequencing by synthesis uses DNA polymerase as the elongation engine during DNA sequence determination. Sequencing by ligation methods harness the mismatch sensitivity of DNA ligase to determine the sequence of nucleotides in a given DNA strand (Landegren et al., 1988). These methods use oligonucleotide probes of varying lengths, which are labeled with fluorescent tags, depending on the nucleotide(s) to be determined. The fragmented DNA templates are primed with a short, known anchor sequence, which allows the probes to hybridize. DNA ligase is added to the flow cell and joins the fluorescently labeled probe to the primer and template. Fluorescence imaging is performed to determine which probe was incorporated. This process is repeated using different sets of probes to query the DNA template and assess the sequence of nucleotides. The methods we describe here differ in their probe usage and read length.

SOLiD—Life Technologies/Applied Biosystems (http://www.appliedbiosystems.com) has created the support oligonucleotide ligation detection (SOLiD) platform that utilizes sequencing by ligation to determine DNA sequence composition. Fragmented or mate-paired, primed libraries are enriched using emulsion PCR on microbeads, which are then adhered to a glass

TABLE 1. Comparison of performance and advantages of various next-generation sequencing (NGS) platforms (modified and updated from Metzker, 2010).

Platform	Library	NGS chemistry	Mean read length (bases)	Reads per run (1 flow cell)	Typical output per plate	Consensus accuracy	Run time	Machine cost (US$)	Pros	Cons
Sequencing by synthesis										
Roche/454 GS FLX+ [a]	Frag, MP/emPCR	PS	700	~1 million	700 Mb	99.997%	23 h	500 000	Longer reads, fast run times; good choice for de novo assembly	Higher reagent costs, error rates in homopolymer
Illumina HiSEq. 2000 [b]	Frag, MP, solid-phase	RTs	2 × 100	>5 million	~570 Gb	>80% of bases higher than Q30	8.5 d	600 000	Currently most widely used platform, high coverage	Shorter read lengths, less feasible for de novo assembly
Ion Torrent PGM [c]	Frag, emPCR	Natural nucleotides	200	5 million	1 Gb	99.99%	2 h	50 000	Very fast run time, cost effective, open source	Not as much throughput
Sequencing by ligation										
Life/AB SOLiD 5500 Series [d]	Frag, MP/emPCR	Cleavable probe SBL	75 × 35	~1 billion	~120 Gb	99.99%	7 d	600 000	2-Base encoding error correction	Longest run times
Polonator G.007 [e]	MP only/emPCR	Noncleavable probe SBL	26	~80 million	5–12 Gb	>98%	5 d	170 000	Open source; cost effective	Users required to maintain; shortest NGS lengths
Single-molecule sequencing										
Helicos BioSciences HeliScope [f]	Frag, MP/ single-molecule	RTs	35	~1 billion	35 Gb	99.995	8 d	999 000	High multi-plexing ability, no template amplification needed	Short read lengths, high error rates compared with RT-based platforms
Pacific BioScience PacBio HRS [g]	Frag only/ single-molecule	Real-time	1300	35 000	45 Mb	99.999%	~1 h	700 000	Longest reads, no template amplification needed, real time	Highest error rates, requires multiple rounds of sequencing to reduce

Notes: Frag, fragment; MP, mate-pair; emPCR, emulsion PCR; PS, pyrosequencing; RTs, reversible terminators; SBL, sequencing by ligation.

[a] Using XL+ titanium chemistry; http://my454.com/products/gs-flx-system/index.asp [accessed 11 January 2012]

[b] For paired-end 100 bp reads on a single flow cell; http://www.illumina.com/systems/hiseq_2000/performance_specifications.ilmn [accessed 11 January 2012]

[c] For 318 Chip; http://www.iontorrent.com/technology-how-does-it-perform/ [accessed 11 January 2012]

[d] For paired end 75 × 35 bp using 1.0 μm microbeads; http://media.invitrogen.com.edgesuite.net/solid/pdf/CO18235-5500-Series-Spec-Sheet-F2.pdf [accessed 11 January 2012]

[e] Mate paired only; http://www.polonator.org/faqs.aspx [accessed 11 January 2012] and Metzker 2010

[f] http://helicosbio.com/Technology/TrueSingleMoleculeSequencing/tSMStradePerformance/tabid/151/Default.aspx [accessed 11 January 2012]

[g] 45-min movie time using SMRT technology; http://genome.hku.hk/portal/files/GRC/Events/Seminars/2011/20110513/pacbiors_overview.pdf [accessed 11 January 2012]

slide. A set of four 1,2-probes, each labeled with a different fluorophore, is added to the flow cell. The first two positions comprise a known di-base pair specific to the fluorophore; these bases query the first and second positions following the hybridized primer. Bases three to five are degenerate bases separated from bases six to eight, made up of universal inosine bases, by a phosphorothiolate linkage. A matching 1,2-probe is ligated to the primer by DNA ligase. Following fluorescence imaging to determine which 1,2-probes were ligated, silver ions cleave the phosphorothiolate link, thus regenerating the 5′ phosphate group for subsequent ligation (McKernan et al., 2005). This process of ligation, detection, and cleavage is repeated several more times, extending the complementary strand to a length determined by the number of cycles. After sufficient length is reached, the extended product is removed, the process begun anew, and the template reset with a primer complementary to the $n - 1$ position of the previous round of primers. The template is extended through the series of ligations, then reset four more times. This primer reset process results in each template base being queried twice, a check and balance system that is determined through the creation and alignment of a series of color images analyzed through space and time to determine the actual DNA sequence. This method is often used in resequencing studies (e.g., Ashelford et al., 2011), transcriptomics, or in genomic sequencing along side other technologies (e.g., Shulaev et al., 2011).

Polonator—The Polonator G.007 system was developed in Dr. George Church's laboratory at Harvard University in collaboration with Dover Systems and is available through Azco Biotech (http://www.azcobiotech.com/instruments/polonator.php). Library preparation for sequencing on the Polonator system is accomplished using emulsion PCR for amplification of template DNA, loading of the beads onto the flow cells and fully automated polymerase colony (referred to as polony) sequencing (Shendure et al., 2005) by ligation. Currently, the output of a full eight-flow-cell Polonator-sequencing "run" is up to 240 million mappable reads of sequence with a read length of 40 bases accomplished in ~80 h. One of the key concepts of the Polonator system is to provide a benchtop, open-source platform: all aspects of the system—from the machine itself to the software available for the platform—are open source, allowing individual laboratories to develop highly specific protocols and applications that do not specifically come from a kit. While this system has been reviewed in several publications regarding next-generation sequencing technologies (Deschamps and Campbell, 2010; Moorthie et al., 2011; Myllykangas et al., 2011; Pareek et al., 2011) and is currently the major system for the Personal Genome Project (http://www.personalgenomes.org/), its use in plant genomics is limited thus far.

Single-molecule sequencing—Single-molecule sequencing (SMS), often termed "third-generation sequencing", circumvents some of the quandaries facing other NGS technologies. These methods produce a detectable signal of nucleotide incorporation via chemiluminescence during DNA sequencing from a single nucleic acid molecule, thus eliminating the need for DNA template amplification. This provides SMS with several benefits over other NGS methods, such as simplified sample preparation that can use degraded or low concentrations of starting material (e.g., Orlando et al., 2011) and the avoidance of PCR errors and biases introduced during template amplification processes. These methods have also been used for direct RNA sequencing, thus removing the biases created during cDNA amplification in RNA-seq studies (Ozsolak et al., 2009). All SMS methods to date use fluorescence imaging to detect nucleotide incorporation. The SMS technologies we discuss here differ in how emitted light is detected, how background fluorescence is minimized, the types of chemistry used, and how template or other molecules are immobilized during flow cell cycles. SMS technologies are relatively new to the market. As these methods become more readily available and further developed, applications in plant genomics are underway with publications soon to follow.

Helicos—The Helicos Genetic Analysis System (http://www.helicosbio.com) was the first commercially available single-molecule sequencing system (SMS) on the market (Harris et al., 2008). This technique uses fragmented, denatured DNA templates that are either hybridized to immobilized oligonucleotide primers adhered to a solid surface or directly immobilized by covalent bonding to a solid surface with the template then subsequently primed with a universal primer (Thompson and Steinmann, 2010). DNA polymerase and modified fluorescently labeled, "virtual terminator" nucleotides are sequentially washed over the DNA template molecules one nucleotide at a time, with incorporation events detected by fluorescence imaging. After each round, the terminating dye moiety is removed and the cycle repeated until read length is reached (Bowers et al., 2009). This method produces average read lengths of 35 bp across 600 million to 1 billion reads, totaling 21–35 Gb per run at a rate of >1 Gb/h. The Helicos platform lends itself well to multiplexing with up to 96 samples per channel or 4800 samples per run (http://www.helicosbio.com).

Pacific BioSciences—Pacific BioSciences (http://www.pacificbiosciences.com) released the PacBio *RS* SMS platform to a limited number of consumers in 2010. In this system, DNA polymerase molecules are immobilized within a zero mode waveguide (ZMW) (Levene et al., 2003) that spatially restricts the volume of nucleotides and light emission to a small area surrounding the DNA polymerase (Eid et al., 2009). The DNA template molecule is bound by the polymerase for sequencing. All four nucleotides, each labeled with a different fluorescent dye, are present in the mixture. The ZMW ensures that only the nucleotide nearest the polymerase, the one actively being incorporated, has the strongest fluorescent signal over the greatest length of time. Fluorescence imaging then allows the sequence of nucleotides to be determined. This immobilized DNA polymerase model can accommodate real-time methods and allow for much longer DNA templates to be analyzed as compared to Helicos, with lengths potentially reaching into the tens of thousands of base pairs (Metzker, 2010). The immobilized DNA polymerase eventually wears out due to degradation by laser excitation and cannot be replaced, thus potentially limiting read lengths. In addition, this method currently has the highest raw error rates of NGS methods (~15%), but this can be overcome by sequencing the same molecule multiple times to elucidate and eliminate errors (Eid et al., 2009; Metzker, 2010). This approach has been taken with their new SMRTbell protocol in which templates are circularized and sequenced repeatedly, increasing accuracy to 99.999% for 30× coverage.

Emerging technologies—NGS technologies are evolving at a very rapid pace, with established companies constantly seeking to improve performance, accessibility, and accuracy in their

quest to outpace competitors and be the first to offer the US$1000 genome. New technologies are constantly emerging, pushing the boundaries of our imaginations through interdisciplinary innovations. Several optical sequencing technologies are being explored that will enable extremely long DNA strands to be read (Thompson and Milos, 2011). Other researchers are investigating nanopores as a means of reading DNA sequence based solely on the inherent electronic or chemical properties of the native nucleotides themselves. Here we provide a brief overview of a few emerging technologies that may begin the fourth generation of NGS techniques.

Life Technologies—VisiGen/Starlight—In 2008, Life Technologies acquired VisiGen, a company that engineered another SMS platform purporting real-time DNA sequencing (Hardin, 2008). Since then, Life Tech has rechristened its platform as Starlight, an appellation originating from the use of quantum dot (qdot)-derived fluorescence resonance energy transfer (FRET) to detect nucleotide incorporation events. Only the fluorescently labeled nucleotide being incorporated is in close enough range to be excited by FRET originating from the qdot attached to the DNA polymerase (Pennisi, 2010; Thompson and Milos, 2011). DNA template strands are immobilized by ligating to oligonucleotides adhered to the glass surface and primed with DNA fragments complementary to the immobilized oligonucleotides. Like PacBio *RS*, Starlight use four-color fluorescence imaging to determine DNA sequence. Like Helicos, Starlight has the advantage of being able to replace DNA polymerase, but after each read is sequenced to completion. It is currently unknown how this technology will compare with other NGS methods, but it is likely to compete or outperform PacBio in terms of read length—because of the ability to replace spent polymerases. Quality of base calls is likely to be higher than PacBio *RS* and may be further improved by the resequencing of immobilized templates, much like PacBio does to improve the overall quality of their sequence runs.

Nanopore sequencing—At the forefront of developing nanopore sequencing is Oxford Nanopore Technologies (http://www.nanoporetech.com/) who has been developing the GridION system. Nanopore methods do not require an amplification step as part of library preparation. The allure of such a system is that a strand molecule (DNA or RNA) can be directly analyzed, as it passes thorough a nanopore, using inherent electronic or chemical properties of each nucleotide to determine order. Currently, sequencing of a DNA molecule can be accomplished by two methods: exonuclease sequencing (Howorka et al., 2001; Lieberman et al., 2010) wherein bases are cleaved one at a time and then pass through the nanopore in the order they were cleaved for detection or by strand sequencing (Clarke et al., 2009) wherein an enzyme is specifically designed to pass a single strand of DNA through a nanopore for detection. A potential advantage of nanopore sequencing is the speed with which a DNA molecule passes through a nanopore, potentially enabling incredibly fast DNA sequencing throughput. The GridION system will likely be able to produce sequence at a very fast rate (This technology is currently being used as an NHGRI-funded project in Dr. Mark Akeson's laboratory at the University of California–Santa Cruz as part of the "$1,000 genome program."). However, this potential advantage is currently an issue with this method because measures must be taken to slow down the speed with which a DNA molecule passes through the nanopore to increase accuracy of the read. Although nanopore sequencing still faces several challenges, it looks to have a promising future (Branton et al., 2008).

Comparisons and challenges—The various NGS platforms currently available or under development present several different ways to sequence DNA, each employing various strategies of template preparation, immobilization, nucleic acid chemistries, synthesis, and detection of nucleotide type and order. These methodological differences transmute into differences in key performance indicators such as read length, throughput, output, and error rates, with each NGS method having important advantages and disadvantages (Table 1). Because NGS methods differ in read length, the types and prevalence of errors, and the number of reads created per run, different approaches are needed to deal with the data in terms of quality control, assembly, and analysis. This presents a major challenge in terms of computational resources, innovation, and application.

*Bioinformatics—*As next-generation sequencing has continued to improve with higher sequencing depth, reduction in cost, and a broadening of application to a wide range of projects from ecology to marker-assisted breeding, the computational challenges have correspondingly grown. Creating 180 million reads has become somewhat straightforward (assuming you use a commercially available kit), but what to do with such depth of data is a challenge. Biologists are now faced with "I have a hard drive full of data but what do I do now?" The challenge of dealing with NGS data is compounded by the fact that each sequencing platform presents its own unique set of challenges for assembly and analysis. A full review of these challenges and the software endeavoring to overcome them are beyond the scope of this paper, but have been the subject of several excellent reviews that cover the incredible depth of software available for quality control, assembly, and quantitative analysis of next-generation sequence (e.g., Metzker, 2010; Miller et al., 2010; Horner et al., 2009). In 2009, the journal *Bioinformatics* devoted an entire issue to the bioinformatic tools and algorithms that have been developed for next-generation sequence analytical challenges (Bateman and Quackenbush, 2009). These bioinformatic tools and programs are continually evolving and improving to keep pace with NGS technical advances, with new software being created all the time.

While many of the early software packages available ran via command line in a UNIX environment, several packages have arrived on the market that allow the development of pipelines for analysis or allow a scientist to use existing computational pipelines within the framework of a user-friendly graphical interface. Many of these platforms incorporate the algorithms that have been developed that address the challenges of mapping reads to a genome or performing de novo assemblies in the absence of a reference genome. One such platform is Galaxy (http://main.g2.bx.psu.edu/) (Giardine et al., 2005; Blankenberg et al., 2010; Goecks et al., 2010). Galaxy is a fully open-source platform that allows a scientist to create custom analysis pipelines or to use other developer's pipelines for analysis. This inclusive platform allows the user to traverse from quality control of data through statistical analysis and visualization of the output. Galaxy also has developed a full NGS analytical toolbox as part of the system. Of particular interest to plant biologists is the iPlant Collaborative (http://www.iplantcollaborative.org) (Goff et al., 2011), a community-driven approach to solving such grand challenges as dealing with NGS data. This initiative has

resulted in the creation of a data analysis portal including tools for NGS analyses and offers many of the same advantages as Galaxy.

For those familiar with the R statistical analysis environment, Bioconductor (http://www.bioconductor.org/) (Gentleman et al., 2004) provides a fully open-source analysis pipeline for next-generation sequence data, and many of the open-source command-line software packages are built in. The only potential drawback to Bioconductor is that it requires a fair amount of knowledge of the UNIX environment to navigate through the system. CLC Bio has also entered the arena of providing user-friendly software for next-generation sequencing with the CLC Genomics Workbench (http://www.clcbio.com/genomics). Analytical toolkits are available for shotgun sequence assembly, transcriptomics, and epigenomics. The JMP interface for SAS is under development for the JMP Genomics (http://www.jmp.com/software/genomics/) platform that will allow scientists to take advantage of the statistical power of SAS in a point-and-click environment with unique visualization output. Softgenetics, one of the leading software developers for microarray analyses, has also developed a software package, NextGENe (http://softgenetics.com/NextGENe.html), for several applications of next-generation sequence data. Similarly, DNASTAR has developed SeqMan NGen (http://www.dnastar.com/t-products-seqman-ngen.aspx) as a fully integrated platform for assembling 454 and Illumina data. While many of these analytical toolkits are available only for purchase, one can easily obtain trial versions to see which platform will provide the needed tools.

SYNOPSES OF THE STUDIES IN THIS SPECIAL ISSUE

Next-generation sequencing technologies are paving the way to a new era of scientific discovery. As genome sequencing becomes easier, more accessible, and more cost effective, genomics will become an integral part of every branch of the life sciences; plant biology is no exception. As exemplified in this special issue, scientists are seeking and finding novel and interdisciplinary applications for NGS technologies, advancing plant sciences in ways previously thought impossible. Here, we summarize the body of research presented in this special issue.

***Molecular marker development*—**Single nucleotide polymorphisms (SNPs; or point mutations) refer to changes in a single base in DNA relative to that expected at that position. They can be used to determine genetic variation, construct genetic linkage maps, or associate SNPs to phenotypic variants, for example, disease associations in humans and traits for breeding in crop plants. Sanger sequencing of amplicons or the mining of Sanger expressed sequence tags (ESTs) was the most common approach of SNP identification earlier, but NGS analysis of DNA/RNA of individuals is an emerging approach for SNP discovery in plant and animal species. NGS-based SNP discovery is very challenging in the species that do not have a reference genome because of poor alignment of short sequence reads of different individuals and genotypes generated by current NGS technologies. In this context, Azam et al. (2012) compared four commonly used short read alignment tools (Maq, BowTie, Novoalign, and SOAP2) with their new approach called coverage-based consensus calling (CbCC) for SNP discovery as a case study in chickpea, *Cicer arietienum* L., a crop lacking a reference genome. They found Maq to be most accurate and sensitive, even at low read depth. All four tools demonstrated greater accuracy at higher read depth, and SNPs predicted by three or four tools were more likely to be correct. SNP prediction accuracy generally increased with increasing read depth. The results obtained in this study are applicable for NGS-based SNP discovery in any other plant species that does not have a reference genome. In addition, the study identified 4543 putative SNPs in chickpea that will be useful for advancing chickpea genetics research and breeding applications.

Simple sequence repeats (SSRs or microsatellites) are repeating DNA sequences (tandem arrays) of 1–6 nucleotides that occur in all prokaryotic and eukaryotic genomes. Their high mutation rate and consequent high polymorphism, multiallelic and codominant nature, and need for little DNA for gathering data makes them ideal for a variety of applications involving closely related accessions, including linkage map development, quantitative trait loci (QTL) mapping, marker-assisted selection, parentage analysis, cultivar fingerprinting, genetic diversity studies, gene flow, and evolutionary studies. Once developed, microsatellite primers are quick and easy to use. Before NGS, SSRs were developed through the laborious process of constructing genomic libraries using recombinant DNA enriched for a few targeted SSR motifs, followed by the isolation and sequencing of clones containing the SSRs. Zalapa et al. (2012) show the power of NGS for developing SSRs in plants through a review of their work in cranberry and 95 other studies that developed SSRs using Sanger, Illumina, and 454 technologies, with many of these studies published in *American Journal of Botany Primer Notes and Protocols in Plant Sciences*.

***Hybridization and introgression*—**Weeds are economically important yet have received little attention in the development of their genome resources. Lai et al. (2012) generated 22 EST libraries from a variety of tissues for 11 weeds in the sunflower family using Sanger, 454, and Illumina sequencing, compared the coverage and quality of sequence assemblies, and developed NimbleGen microarrays for expression analyses in five of them. For some taxa, they also compared the distributions of Ks values (number of synonymous substitutions per synonymous site) between orthologs of congeneric taxa to detect and quantify hybridization and introgression. The distribution of Ks values for orthologs should be centered around a Ks value corresponding to the time since the last common ancestor of the taxa involved, but a secondary peak at a lower Ks value can be attributed to more recent gene flow. Using this method, substantial introgression was detected among congeneric taxa of *Centaurea* and *Helianthus*, but not in *Ambrosia* and *Lactuca*; such data will be useful for long-standing questions of whether hybridization is a cause or consequence of range expansions. They found that gene discovery was enhanced by sequencing from multiple tissues, normalization of cDNA libraries, and especially greater sequencing depth. They also discovered that assemblies from short sequence reads sometimes failed to resolve close paralogs.

***Transcriptome investigations*—**RNA-seq is a NGS method that sequences the transcriptome (all RNA transcripts). It can show the expressed sequences in specific tissues at a specific time and is rapidly replacing other methods of studying gene expression such as microarrays. It is practical in nonmodel plant species mainly because a reference genome is not required. Strickler et al. (2012) review methods to design RNA-seq

projects and to analyze and interpret the data. They discuss (1) the need for consideration of organism-specific features, such as level of heterozygosity and availability of a reference genome, and the consequences of organism choice on analysis; (2) tissue treatment and selection of tissue types to obtain the transcriptome desired and the desirability of replicates; (3) techniques for efficiently selecting transcripts for sequencing; (4) normalizing transcripts to avoid over-representation of highly abundant transcripts; (5) choice of sequencing platforms; and (6) methods of data assembly, with a useful table of assembly programs. They then provide a set of ways in which transcriptome data can be used, such as characterizing differential expression or tissue-specific transcripts, and SNP identification that can be useful in designing markers for mapping and studying evolution. They also discuss ways to share the data with publicly available databases.

Alternative splicing is a mechanism by which different forms of mature mRNAs are generated from the same gene. Conservative estimates of alternative splicing project that around one fifth or more of plant genes undergo alternative splicing, in which a single pre-mRNA can be processed into diverse transcripts, often encoding protein isoforms with distinct or even antagonistic functions. RNA splicing operates by different mechanisms for exons to be incorporated into mRNA. Gulledge et al. (2012) describe a method to analyze NGS data to assess alternative splicing in plant genes. Until recently, it was difficult to study alternative splicing prevalence or regulation, but NGS RNA-seq now makes it possible to study splicing on a whole genome scale and assess how splicing changes in response to diverse treatments. Gulledge et al. (2012) added a new RNA-seq visualization capability to Integrated Genome Browser, an interactive tool to view and explore genomic data. This paper demonstrates how scientists can use Integrated Genome Browser to investigate splicing using a single locus, using SR45a (encoding a putative splicing regulator from *Arabidopsis*) and *Arabidopsis* RNA-seq data sets from their group as examples. SR45a is annotated as producing two splice variants; the V1 form encodes the full-length protein, while the V2 form includes an alternatively spliced exon that introduces a premature stop codon and encodes a truncated product. Their paper describes how they used IGB to explore RNA-seq data sets from *Arabidopsis* plants subjected to heat and dehydration treatments. Using Integrated Genome Browser, they found that stresses increased exon skipping in SR45a, causing increased production of the full-length form. This result supports the hypothesis that SR45a plays a role in adaptation to heat and water deprivation stress in *Arabidopsis* and suggests the possibility that it may participate in regulating its own splicing.

Next-generation sequencing has greatly advanced our opportunities to obtain genome sequences, but its application to large-scale population and phylogenetic studies is hindered by large genome sizes, variation in the proportion of organellar DNA in total DNA, polyploidy, and gene number/redundancy. Targeted enrichment strategies can alleviate many of these problems by reducing the complexity of plant genomes. Cronn et al. (2012) summarize the many available targeted enrichment strategies for organellar and nuclear genomes. They classify these strategies into four categories: (1) PCR-based enrichment, (2) hybridization-based enrichment, (3), restriction-enzyme-based enrichment, and (4) enrichment of expressed gene sequences. They conclude that PCR-based enrichment provides a reasonable strategy for accessing small genomic targets (e.g., <50 kbp), but that hybridization and transcriptome sequencing are most efficient if larger genomic targets are desired. They provide detailed protocols (as online appendices) to generate hybridization-enriched libraries of organellar genomes and of estimated costs to produce enrichment libraries.

A major challenge to those beginning to work with NGS data is retooling for methods to store and analyze huge amounts of data. Statistical analysis is an essential component for RNA-seq data, but due to the short history of the technology and its continuous development, there are as yet no standard methods available to detect and analyze differentially expressed genes based on NGS data. Analytical programs for these data are just emerging and need to be evaluated. Kvam et al. (2012) review and compare the currently available methods to detect differentially expressed genes and provide information on how to download the corresponding packages in freely available R software (R Development Core Team, 2010) They compare the performance of four recently proposed statistical methods (edgeR, DESeq, baySeq, and a method with two-stage Poisson model [TSPM]) on significance ranking of genes and false discovery rate control through simulation studies under various settings mimicking real data. The results show that the performances of different methods vary and that baySeq performs best in terms of significance ranking of genes. The false discovery rate may not be controlled well in practice, and they suggest applying a relatively stringent level to avoid too many false positives. In addition, the flexibility of handling different experimental designs varies among the current versions of the different packages. Plant biologists may want to choose the one that best fits their experimental design and goal.

Ward et al. (2012) review approaches for the analysis of short-read transcriptome data for nonmodel species for which the genome of a close relative is used in place of a true reference genome. They compare two approaches (align then assemble, which is dependent on a reference genome; and assemble then align, a de novo assembly method). While both are powerful methods for transcriptome analysis in nonmodel organisms, the former requires a close reference genome, while the latter does not require such sequences and, in their system, identified over twice the number of transcripts. They conclude that the use of both methods provides a powerful combination. While align then assemble methods provide a statistical framework for differential expression testing and hypothesis generation, the de novo approach holds more power for discovering unique sequences and also provides the possibility of simultaneously querying the transcripts and expression levels in multiple organisms in a system.

Transcription factors are proteins that bind to regulatory regions in the genome and help govern gene expression. They control when genes are switched on or off and whether genes are transcribed or not. In addition, different transcription factors can act as repressors and some as activators. While transcription factors bind to specific DNA motifs adjacent to the genes they regulate, there can be more than one functional transcription-factor-binding site per gene. Further, single transcription factors sometimes directly govern multiple downstream processes, resulting in the highly complex gene regulatory networks that control multiple biological processes. NGS provides ideal tools to disentangle the many interactions that constitute these gene regulatory networks. Yant (2012) reviews recent work using the in vivo transcription-factor-binding mapping technique chromatin immunoprecipitation (ChIP), combined with ultra-high-throughput sequencing (ChIP-seq). This approach provides a snapshot of a single protein's direct physical

interactions with DNA at a particular time in a particular tissue, on a genome-wide scale. This allows the in vivo mapping of the interactions of particular proteins and their transcriptional targets to show interconnections of gene regulatory networks. All work to date is in *Arabidopsis thaliana*, but NGS will soon open this up to other organisms. Yant (2012) reviews the mapping of regulatory networks governing the switch from the vegetative to reproductive phase, pointing out widespread integration in complex sets of signaling networks.

Common reed (*Phragmites australis* L.) is one of the most widely distributed angiosperms and a significant invasive species worldwide that outcompetes most species it encounters. Common reed, as many vascular plants, including important hay, forage, pasture crops, and wild grain species, uses rhizomes as the primary energy storage and propagation organ. He et al. (2012) compare the transcriptomes and proteomes of developing rhizomes in common reed to identify candidate genes and proteins involved in rhizome growth, development, metabolism, and invasiveness. They found that a considerable number of phytohormone-related transcripts/genes that are important to rhizome growth and development are expressed in the developing tips of reed rhizomes, genes that encode proteins with known roles in the metabolism of or response to phytohormones such as auxin, abscisic acid (ABA), cytokinin, gibberellic acid (GA), and ethylene. Additionally, a number of putative transcription factors and regulators were identified that may also play important roles in rhizome development and function. Common reed competitiveness possibly is due to allelopathic agents as yet unidentified, and the transcriptome database will be useful for further investigations.

Phylogenetic and ecological studies—Before NGS, opportunities for addressing a wide range of large-scale genome-level questions were restricted to those working on well-studied model organisms (or their close relatives) possessing a wide scope of genomic resources ranging from sequenced EST libraries to whole genome sequences. Ecologists and evolutionary biologists need data from large numbers of individuals, and until recently, those working on nonmodel organisms were limited to slow and costly gene-by-gene approaches. Grover et al. (2012) describe how targeted sequence capture, coupled with NGS, opens up genomic resources to nonmodel organisms, allowing us to address questions such as parentage, gene flow, population divergence, phylogeography, diversity, domestication and improvement, phylogeny, hybrid identification, introgression, and polyploid parentage. Targeted sequencing refers to a range of technologies designed to isolate specific genomic regions for subsequent NGS. The reduced genomic portion of the specifically targeted sample of sequences generated by these techniques allows multiplexing of reactions and greatly simplifies analysis and costs. Grover et al. (2012) mention three methods for targeted enrichment: (1) hybridization-based sequence capture, (2) PCR-based amplification, and (3) molecular inversion probe-based amplification (reviewed in detail in Mamanova et al., 2010), and elaborate further on sequence capture, which is quick, simple, and relatively inexpensive. They then provide examples of such techniques applicable to a wide range of questions addressed by ecologists and evolutionary biologists.

Steele et al. (2012) investigate the use of NGS in phylogenetic analysis of two lineages of monocots, the Asparagales and the grasses, using Illumina data (80–120-bp reads). They make the point that even low-coverage data, which does not aim to assemble complete nuclear sequences, provide genomic sequences of high-copy regions (plastids, mitochondria, nuclear ribosomal DNA) sufficiently well to provide high quality assemblies. These can provide a sufficient quantity of phylogenetically informative characters to produce robust phylogenies, even of closely related and recently diverged taxa that were recalcitrant to phylogenetic analysis before. Their results were not dependent on genome size (ranging from 1.3 pg/2C to 50.9 pg/2C), amount of plastid present in the total genomic DNA template (as determined by real-time PCR Ct values), or relatedness of available reference sequences for assembly. Costs are dramatically lower for generating the data, and much laboratory time is saved. In addition, perhaps 90% of the data from the nuclear genome remains unanalyzed and presents a potentially valuable resource for analysis of repetitive assemblies. The biggest challenge is new training to handle such large quantities of data.

Kane et al. (2012) similarly sequence plastid and ribosomal DNA but with the goal of producing "barcodes" (taxon-specific molecular profiles) below the species level. Using Illumina sequencing, they examined whole plastid genomes and nearly 6000 bases of nuclear ribosomal DNA sequences. Their large numbers of characters vastly exceed that of traditional barcoding, which uses short sequences from defined regions of the genome. They term their approach "ultra-barcoding" and use it to examine nine genotypes of three varieties of *Theobroma cacao* L. and an individual of the related species *T. grandiflorum* (Sprengel) Schumann, and *T. cacao*. They analyzed the plastid data by maximum likelihood and the ribosomal data by a network-based approach because ribosomal DNA undergoes recombination and can violate bifurcating models. They obtained 4.2–11 times coverage of the nuclear genome and had more than enough coverage for plastid and nuclear ribosomal DNA assembly. The data clearly separated *T. cacao* from *T. grandiflorum*. The plastid data showed two strongly supported clades within *T. cacao* corresponding to two of the three varieties of *T. cacao* (Forastero and Criollo), and showed that maternal lineages of the third variety (Trinitario) came from both Forastero and Criollo; the ribosomal data showed a similar pattern. The authors suggest that ultra-barcoding will be very useful as a supplement to traditional barcoding methods and show that taxon-specific profiling can be successful below the species level.

Straub et al. (2012) describe the use of Illumina technology to "skim" the high copy fraction of the genome to obtain nucleotide sequences of nearly complete plastid genomes and nuclear ribosomal DNA (rDNA), as well as kilobase portions of the mitochondrial genome. Ironically, such sequences are regarded as "contaminants" to be minimized in protocols targeting the nuclear genome, but that are very useful for systematists and ecologists. Genome skimming can also provide partial sequences of low-copy nuclear loci, sufficient for designing PCR primers or probes for hybridization-based genome reduction approaches as described by Cronn et al. (2012) in this issue. This paper examines, through simulations, the optimal sequencing depth needed for single-end and paired-end sequence data sets for ribosomal and plastid sequences and for conserved low-copy nuclear loci. They describe efficiency relative to increasing divergence from a plastome assembly. They demonstrate the utility of these approaches within a clade of milkweeds, *Asclepias*.

Polyploid genetics—Buggs et al. (2012) review NGS approaches for investigating genomic and transcriptomic changes in polyploids, with application to *Tragopogon* (Compositae)

and *Nicotiana* (Solanaceae), evolutionary models for the study of allopolyploidy. Genomic changes in *Tragopogon* polyploids were examined at the level of genes and repetitive DNA (by quantifying the rapidity of gene and repeat loss and documenting parental and functional biases) and the level of the chromosome (by developing FISH and GISH markers for the study of genomic structural changes). Transcriptomic changes—gene silencing and tissue-specific expression—were also detected in *Tragopogon* polyploids. The general experimental approach begins with extensive genomic and cDNA sequencing of the diploid parents of allopolyploids using a combination of 454 and Illumina platforms. These data were used to discover parent-specific differences within protein-coding and repetitive regions. Parent-specific SNPs in *Tragopogon* allopolyploids were assayed in both the genome and transcriptome using Sequenom MassARRAY iPLEX technology, a method used in maize genomics, which is especially suited for assaying genetic variants among highly similar sequences. Sequenom experiments on *Tragopogon miscellus* led to two significant discoveries. First, genes were repeatedly retained or lost in clusters—and the gene ontology categories of the genes missing from the genome correspond to those lost after ancient whole-genome duplication in the same family (Compositae) and with gene-dosage sensitivity; these results provide evidence for the gene balance hypothesis and also suggest that the outcomes of polyploidy are predictable, even in young polyploids after just 40 generations. Second, tissue-specific expression of genes evolves in as few as 40 generations after allopolyploid formation, following an apparent deregulation of expression upon hybridization. In both *Tragopogon* and *Nicotiana*, differences found between parents in repetitive elements enabled the design of FISH probes, contributing to the discovery of extensive chromosomal changes in young *T. miscellus* allopolyploids and changes in repetitive element composition in *Nicotiana* allopolyploids. Buggs et al. (2012) stress that NGS technologies can be easily and inexpensively applied to many plant species, making any evolutionarily provocative system a potential new "model" system. They also point out the need for training the next generation of biologists in bioinformatics to handle, analyze, and interpret the huge quantities of data generated by these new approaches.

Bread wheat (*Triticum aestivum* L.) is a crop plant of great economic importance, providing 20% of the calories consumed by the world's population. It has a huge and complex allohexaploid genome making complete genome sequencing a major challenge. The bread wheat genome is 17 Gbp in size and consists predominantly of repetitive elements. In comparison, the rice genome is 400 Mbp and maize, 2.3 Gbp. The availability of genome sequences for rice, sorghum, and maize has greatly enhanced our ability to understand the physiology, QTL for genes associated with domestication and drought tolerance, and pest and disease resistance of these crops. The lack of a genome sequence for wheat has hampered efforts to determine the genetic basis of phenotypic traits. Berkman et al. (2012) describe the progress of resolving the bread wheat genome. This problem has been approached by sequencing isolated chromosome arms, thereby reducing the confounding effects of allopolyploid complexity, through the application of NGS technologies. They also describe transcriptome sequence alignment and the development of SNPs that have been respectively useful in investigating gene expression patterns and association genetics in wheat.

Ilut et al. (2012) explore the problems associated with conducting expression profiling on ancient and recent polyploids, using the legume genus *Glycine* (soybean and allies) as a model. Soybean and other "diploid" *Glycine* are actually polyploids from a whole-genome duplication event that occurred within the last 10 Myr. Much more recent hybridization and duplication has produced several allopolyploid species, some of which show greater adaptability and colonizing ability (polyploid advantage) relative to diploids that could be due to enhanced photosynthesis. Transcript assemblies from short (36 bp) single-end Illumina reads are complicated by the fact that transcripts belonging to duplicated genes in a polyploid genome (homoeologues) will assemble together, making the measurement of a gene's transcript counts prone to large errors. As well, many short reads are likely to align to identical sequence fragments at several loci in the duplicated soybean reference genome and thus may be discarded from transcript counts, also affecting the estimated expression level for a given gene. These authors conducted an RNA-seq experiment comparing leaf transcriptome profiles of a recently formed allopolyploid relative of soybean with diploid species contributing to its genome. They developed a transcript-specific metric that allows an investigator to account for the ambiguously aligned sequences when measuring the transcript level of a gene against a highly duplicated scaffold. Another problem arises when the system under investigation is polyploid with respect to the reference genome, but if the diploid progenitor species are extant and their allele sequences are sampled, it is possible to determine the relative contribution of the duplicated loci in the polyploid for a large number of genes. Finally, the overall expression patterns of different genotypes can be similar, and it can be difficult to quantify the differences between these distributions. These authors present a metric borrowed from information theory, the Kullback–Leibler divergence, which allows a useful quantification of the statistical distance between gene expression level distributions.

Whereas most karyotypes exhibit a continuous range of chromosome sizes, karyotypes of some taxa are bimodal with chromosomes falling into two distinct size classes often described as L for large and S for small. Bimodal karyotypes are most often limited to single genera, small groups of closely related species, or single species, but bimodal chromosome size distributions are shared among multiple genera in Asphodeloideae (Xanthorrhoeaceae) and Agavoideae (Asparagaceae). Within these subfamilies, bimodal karyotypes are synapomorphies for species-rich clades that may be millions of years old. Chromosome sizes are uniformly distributed in karyotypes for Agavoideae species outside of a clade of Agavaceae s.l. comprised of 15 genera and 377 species, here referred to as the ABK clade. Two processes have been hypothesized to give rise to bimodal karyotypes, fusion–fission events possibly caused by genomic shock following an allopolyploid event or the retention of distinct chromosome size distributions following allopolyploidy of parental diploids having different size distributions. McKain et al. (2012) use NGS of transcriptomes to test the hypothesis that chromosome bimodality in this group coincides with a polyploid event, by examining divergence between duplicate genes as measured by the number of nonsynonymous substitutions per synonymous site (K_s) to identify whole-genome duplications. These authors combined analyses of K_s plots and gene family phylogenies to test whether the origin of bimodal karyotypes in this clade is associated with whole genome duplication. K_s frequency plots suggested paleopolyploid events in the history of the genera *Agave*, *Hosta*, and *Chlorophytum*. Phylogenetic analyses of gene families estimated from transcriptome data revealed two polyploid events: one pre-dating the last common ancestor of *Agave* and *Hosta* and one within the lineage

leading to *Chlorophytum*. They conclude that allopolyploidy and the origin of the *Yucca–Agave* bimodal karyotype co-occurs on the same lineage, consistent with the hypothesis that the bimodal karyotype is a consequence of allopolyploidy.

Applications for large gene bank collections—A total of 1750 national and international gene banks worldwide preserve ~7 million accessions of advanced cultivars, landraces, and wild species relatives of plants that the world depends on for food, fiber, and fuel (FAO, 2010). McCouch et al. (2012) present a vision for the potential of large-scale genotyping to help characterize, use, and manage gene bank collections, from their perspectives as scientists working with large-scale rice collections. Gene banks have many pressing challenges due to the large size of their collections and the need to characterize them properly for a wide variety of users. They also face legal constraints (and opportunities) imposed in today's climate of ownership of genetic resources. The challenges include the need to correctly identify accessions, track seed lots, varieties, and alleles, identify and eliminate duplicate accessions, justify adding new accessions to the collection, identify a small subset of the collection that represents a majority of the variation in the entire collection (a "core collection"), identify geographic areas holding useful sets of diverse alleles, associate genotypes with phenotypes, and motivate innovative collaborations to place useful materials into the hands of plant breeders. McCouch et al. (2012) outline these challenges and show how NGS can vastly improve genetic characterization efforts in genebanks. Initial NGS projects with the rice collections include identification of SNPs and other polymorphisms (http://www.oryzasnp.org/; http://www.ricediversity.org/; http://www.ricesnp.org/) based on large-scale resequencing and genotyping projects.

CLOSING REMARKS

As NGS technologies continue to improve, their scope and application will correspondingly expand within and across scientific disciplines. Plant biology has much to gain from increasing our technological capacity in genomics, with applications reaching from plant breeding to evolutionary studies. In terms of comparative genomics, the increasing number of fully sequenced plant genomes will enable greater understanding of genetic, genomic, developmental, and evolutionary processes that create the diversity of plant life on earth. The innovation and application of NGS technologies paints a bright future for plant biology and all areas of life science research, as illustrated in the body of work within this special issue.

LITERATURE CITED

Ashelford, K., M. E. Eriksson, C. M. Allen, R. D'Amore, M. Johansson, P. Gould, S. Kay, et al. 2011. Full genome re-sequencing reveals a novel circadian clock mutation in Arabidopsis. *Genome Biology* 12: R28.

Azam, S., V. Thakur, P. Ruperao, T. Shah, J. Balaji, B. Amindala, A. D. Farmer, et al. 2012. Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results for the identification of SNPs in chickpea (*Cicer arietinum;* Fabaceae), a crop species without a reference genome. *American Journal of Botany* 99: 186–192.

Barnes, C., S. Balasubramanian, X. Liu, H. Swerdlow, and J. Milton. 2002. Labelled nucleotides. US Patent 7,057,026. U. S. Patent and Trademarks Office, Department of Commerce, Washington, D.C., USA.

Bateman, A., and J. Quackenbush. 2009. Bioinformatics for next generation sequencing. *Bioinformatics (Oxford, England)* 25: 429.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53–59.

Berkman, P. J., K. Lai, M. T. Lorenc, and D. Edwards. 2012. Next-generation sequencing applications for wheat crop improvement. *American Journal of Botany* 99: 365–371.

Blankenberg, D., G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor. 2010. Galaxy: A web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology* 19: Unit 19.10.1-21.

Bowers, J., J. Mitchell, E. Beer, P. R. Buzby, M. Causey, J. W. Efcavitch, M. Jarosz, et al. 2009. Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods* 6: 593–595.

Branton, D., D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, et al. 2008. The potential and challenges of nanopore sequencing. *Nature Biotechnology* 26: 1146–1153.

Buggs, R. J. A., S. Renny-Byfield, M. Chester, I. E. Jordon-Thaden, L. F. Viccini, S. Chamala, A. R. Leitch, et al. 2012. Next-generation sequencing and genome evolution in allopolyploids. *American Journal of Botany* 99: 372–382.

Clarke, J., H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* 4: 265–270.

Cronn, R., B. Knaus, A. Liston, P. J. Maughan, M. Parks, J. Syring, and J. Udall. 2012. Targeted enrichment strategies for next generation plant biology. *American Journal of Botany* 99: 291–311.

Deschamps, S., and M. A. Campbell. 2010. Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding* 25: 553–570.

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

FAO. 2010. The second report on the state of the world's plant genetic resources for food and agriculture. Food and Agriculture Organization, Rome, Italy.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, et al. 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* 5: R80.

Giardine, B., C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, et al. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* 15: 1451–1455.

Goecks, J., A. Nekrutenko, and J. Taylor, and The Galaxy Team. 2010. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11: R86.

Goff, S. A., M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, et al. 2011. The iPlant collaborative: Cyberinfrastructure for plant biology. *Frontiers in Plant Science* 2: 34.

Grover, C. E., A. Salmon, and J. F. Wendel. 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany* 99: 312–319.

Gulledge, A. A., A. D. Roberts, H. Vora, K. Patel, and A. E. Loraine. 2012. Mining *Arabidopsis thaliana* RNA-seq data with Integrated Genome Browser reveals stress-induced alternative splicing of the putative splicing regulator SR45a. *American Journal of Botany* 99: 219–231.

Guo, J., N. Xu, Z. Li, S. Zhang, J. Wu, D. H. Kim, M. S. Marma, et al. 2008. Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences, USA* 105: 9145–9150.

Hardin, S. H. 2008. Real-time DNA sequencing. In M. Janitz [ed.], Next generation genome sequencing: Towards personalized medicine, 97–102. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

HARRIS, T. D., P. R. BUZBY, H. BABCOCK, E. BEER, J. BOWERS, I. BRASLAVSKY, M. CAUSEY, ET AL. 2008. Single-molecule DNA sequencing of a viral genome. *Science* 320: 106–109.

HE, R., M.-J. KIM, W. NELSON, T. S. BALBUENA, R. KIM, R. KRAMER, J. A. CROW, ET AL. 2012. Next-generation sequencing-based transcriptomic and proteomic analysis of the common reed, *Phragmites australis* (Poaceae), reveals genes involved in invasiveness and rhizome specificity. *American Journal of Botany* 99: 232–247.

HORNER, D. S., G. PAVESI, T. CASTRIGNANO, P. D'ONORIO DE MEO, S. LIUNI, M. SAMMETH, E. PICARDI, AND G. PESOLE. 2009. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics* 11: 181–197.

HOWDEN, D. P., C. R. E. MCEVOY, D. L. ALLEN, K. CHUA, W. GAO, P. F. HARRISON, J. BELL, ET AL. 2011. Evolution of multidrug resistance during Staphylococcus aureus infection involves mutation of the essential two component regulator WalKR. *PLoS Pathogens* 7: e1002359.

HOWORKA, S., S. CHELEY, AND H. BAYLEY. 2001. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nature Biotechnology* 19: 636–639.

ILUT, D. C., J. E. COATE, A. K. LUCIANO, T. G. OWENS, G. D. MAY, A. FARMER, AND J. J. DOYLE. 2012. A comparative transcriptomic study of an allotetraploid and its diploid progenitors illustrates the unique advantages and challenges of RNA-seq in plant species. *American Journal of Botany* 99: 383–396.

KANE, N., S. SVEINSSON, H. DEMPEWOLF, J. Y. YANG, D. ZHANG, J. M. M. ENGELS, AND Q. CRONK. 2012. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* 99: 320–329.

KVAM, V. M., P. LIU, AND Y. SI. 2012. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany* 99: 248–256.

LAI, Z., N. C. KANE, A. KOZIK, K. A. HODGINS, K. M. DLUGOSCH, M. S. BARKER, M. MATVIENKO, ET AL. 2012. Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression. *American Journal of Botany* 99: 209–218.

LANDEGREN, U., R. KAISER, J. SANDERS, AND L. HOOD. 1988. A ligase-mediated gene detection technique. *Science* 241: 1077–1080.

LEVENE, M. J., J. KORLACH, S. W. TURNER, M. FOQUET, H. G. CRAIGHEAD, AND W. W. WEBB. 2003. Zero-mode waveguides for single-molecule analysis at high concentration. *Science* 299: 682–686.

LIEBERMAN, K. R., G. M. CHERF, M. J. DOODY, F. OLASAGASTI, Y. KOLODJI, AND M. AKESON. 2010. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *Journal of the American Chemical Society* 132: 17961–17972.

MAMANOVA, L., A. J. COFFEY, C. E. SCOTT, I. KOZAREWA, E. H. TURNER, A. KUMAR, E. HOWARD, ET AL. 2010. Target-enrichment strategies for next generation sequencing. *Nature Methods* 7: 111–118.

MCCOUCH, S. R., K. L. MCNALLY, W. WANG, AND R. S. HAMILTON. 2012. Genomics of gene banks: A case study in rice. *American Journal of Botany* 99: 407–423.

MCKAIN, M. R., N. WICKETT, Y. ZHANG, S. AYYAMPALAYAM, W. R. MCCOMBIE, M. W. CHASE, J. C. PIRES, ET AL. 2012. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *American Journal of Botany* 99: 397–406.

MCKERNAN, K., A. BLANCHARD, L. KOTLER, AND G. COSTA. 2005. Reagents, methods, and libraries for bead-based sequencing. US Patent application 11/345,979. U. S. Patent and Trademarks Office, Department of Commerce, Washington, D.C., USA.

METZKER, M. L. 2010. Sequencing technologies—The next generation. *Nature Reviews Genetics* 11: 31–46.

MILLER, J. R., S. KOREN, AND G. SUTTON. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* 95: 315–327.

MOORTHIE, S., C. J. MATTOCKS, AND C. F. WRIGHT. 2011. Review of massively parallel DNA sequencing technologies. *HUGO Journal* 5: 1–12.

MYLLYKANGAS, S., J. BUENROSTRO, AND H. P. JI. 2011. Overview of sequencing technology platforms. In N. Rodriguez-Ezpeleta, M. Hackenberg, and A. M. Aransayet [eds.], Bioinformatics for high throughput sequencing, 11–26. Springer-Verlag, New York, New York, USA.

NYRÉN, P. 2007. The history of pyrosequencing. *Methods in Molecular Biology (Clifton, N.J.)* 373: 1–14.

ORLANDO, L., A. GINOLHAC, M. RAGHAVAN, J. VILSTRUP, M. RASMUSSEN, K. MAGNUSSEN, K. STEINMANN, ET AL. 2011. True single-molecule DNA sequencing of a Pleistocene horse bone. *Genome Research* 21: 1705–1719.

OZSOLAK, F., A. R. PLATT, D. R. JONES, J. G. REIFENBERGER, L. E. SASS, P. MCINERNEY, J. F. THOMPSON, ET AL. 2009. Direct RNA sequencing. *Nature* 461: 814–818.

PAREEK, C. S., R. SMOCZYNSKI, AND A. TRETYN. 2011. Sequencing technologies and genome sequencing. *Journal of Applied Genetics* 52: 413–435.

PENNISI, E. 2010. Semiconductors inspire new sequencing technologies. *Science* 327: 1190.

R DEVELOPMENT CORE TEAM. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RONAGHI, M., M. UHLÉN, AND P. NYRÉN. 1998. A sequencing method based on real-time pyrophosphate. *Science* 281: 363–365.

ROTHBERG, J. M., W. HINZ, T. M. REARICK, J. SCHULTZ, W. MILESKI, M. DAVEY, J. H. LEAMON, ET AL. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475: 348–352.

SANGER, F., S. NICKLEN, AND A. R. COULSON. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, USA* 74: 5463–5467.

SHENDURE, J., G. J. PORRECA, N. B. REPPAS, X. LIN, J. P. MCCUTCHEON, A. M. ROSENBAUM, M. D. WANG, ET AL. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728–1732.

SHULAEV, V., D. J. SARGENT, R. N. CROWHURST, T. C. MOCKLER, O. FOLKERTS, A. L. DELCHER, P. JAISWAL, ET AL. 2011. The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics* 43: 109–116.

STEELE, P. R., K. L. HERTWECK, D. MAYFIELD, M. R. MCKAIN, J. H. LEEBENS-MACK, AND J. C. PIRES. 2012. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *American Journal of Botany* 99: 330–348.

STRICKLER, S. R., A. BOMBARELY, AND L. A. MUELLER. 2012. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *American Journal of Botany* 99: 257–266.

STRAUB, S. C. K., M. PARKS, K. WEITEMIER, M. FISHBEIN, R. C. CRONN, AND A. LISTON. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349–364.

THOMPSON, J. F., AND P. M. MILOS. 2011. The properties and applications of single-molecule DNA sequencing. *Genome Biology* 12: 217.

THOMPSON, J. F., AND K. E. STEINMANN. 2010. Single-molecule sequencing with a HeliScope Genetic Analysis System. *Current Protocols in Molecular Biology* 92: 7.10.1–7.10.14.

WARD, J. A., L. PONNALA, AND C. A. WEBER. 2012. Strategies for transcriptome analysis in nonmodel plants. *American Journal of Botany* 99: 267–276.

YANT, Y. 2012. Genome-wide mapping of transcription factor binding reveals developmental process integration and a fresh look at evolutionary dynamics. *American Journal of Botany* 99: 277–290.

ZALAPA, J. E., H. CUEVAS, H. ZHU, S. STEFFAN, D. SENALIK, E. ZELDIN, B. MCCOWN, ET AL. 2012. Using next-generation sequencing approaches for the isolation of simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany* 99: 193–208.