REVIEW

SABRAO Journal
of Breeding and Genetics
45 (1) 84-99, 2013

NEXT-GENERATION SEQUENCING TECHNOLOGY FOR CROP IMPROVEMENT

K. VAN[1], K. RASTOGI[1], K.-H. KIM[2] and S.-H. LEE[1, 3,*]

[1]Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, 151-921, Korea
[2]Upland Crop Research Division, National Institute of Crop Science, Suwon, 441-857, Korea
[3]Plant Genomics and Breeding Institute, Seoul National University, Seoul, 151-921, Korea
*Corresponding author's email: sukhalee@snu.ac.kr

SUMMARY

By exploiting next-generation sequencing (NGS) technologies, many species including economically important crops, have been subjected to whole-genome sequencing by *de novo* assembly and resequencing. Now, sequencing technologies have evolved from genome sequencing projects using massive parallel sequencing technologies such as NGS to NGS of single DNA molecules (next–NGS). This NGS technology provides us with better opportunities for studying crop genomics and other post-genomics (transcriptomics, proteomics, metabolomics) more closely. Via the discovery of molecular markers generated by NGS and other analyses, we can also explore genetic diversity and crop evolution by full genome sequencing of crop species and many accessions within crop species. The increasing availability of high-throughput technology and the reduction of costs of these technologies have moved genomics from the sequencing of a few model species to sequencing any crop that is important for food security. In this paper, we introduce whole-genome sequencing technology and the status of crop genome sequencing, and we discuss the applications of NGS to crop improvement.

Keywords: crop improvement, crop molecular breeding, *de novo* assembly, marker-assisted selection, next-generation sequencing technology, resequencing

Manuscript received: July 11, 2012; Manuscript accepted: August 18, 2012.

Communicating Editor: Bertrand Collard

INTRODUCTION

Although crop production showed steady and continuous growth in recent years, further improvement in crop productivity is still necessary due to hunger and malnutrition faced by some portions of the world's population (Godfray *et al*., 2010). Also, increasing levels of wealth and the demand for high quality food affect the purchasing power of global populations. A doubling in food prices, the recent massive production of biofuels, climate change and urbanization have led to greater competition for land, water and energy, even as biodiversity and natural ecosystems are being protected (Balmford *et al*., 2005; Fargione *et al*., 2008; Godfray *et al*., 2010; Chang and Hsu, 2011; Varshney *et al*., 2011).

The world population will reach nine billion by 2050 (Godfray *et al*., 2010) and the development of agricultural biotechnology could be a key method for crop improvement to feed the world population in an environmentally and socially sustainable way. Next-generation sequencing (NGS) technology, the most advanced method of genome sequencing, has become the main tool for developing novel molecular markers and identifying genes of

agronomic importance (Edwards and Batley, 2010). Before these methods were developed, the time-consuming clone-by-clone method was used in genome sequencing with the strategy of identifying the least redundant overlapping clones (Figure 1a). However, a physical genetic map of the crop to be sequenced must be provided prior to performing these labor-intensive and time-consuming experiments (Ariyadasa and Stein, 2012). Thus, NGS platforms, such as GS-FLX and Illumina HiSeq, are the best choice for employing the whole-genome shotgun (WGS) strategy for sequencing projects of various organisms including crops because tremendous amounts of data are produced in a short period of time using these platforms. Several companies have brought different technology platforms to the market for third generation sequencing. Egan *et al*. (2012) reviewed these NGS technologies, which employ three different methods: sequencing by synthesis, sequencing by ligation and single-molecule sequencing. Roche 454 pyrosequencing, Illumina and Ion Torrent are the sequencing platforms that employ sequencing by the synthesis method. Sequencing by the ligation method is used in SOLiD and Polonator. Helicos and Pacific Biosciences use the single-molecule sequencing method, which is considered to be next-NGS (Barabaschi *et al*., 2012). Furthermore, bioinformatics tools have been developed in conjunction with the rapid development of current NGS platforms (Lee *et al*., 2012; Figure 1b). In this paper, two sequencing approaches, *de novo* assembly and resequencing (reference genome sequencing), will be introduced. The status of crop genome sequencing will also be presented, along with a discussion of the applications of these technologies to crop improvement, specifically focusing on increasing crop adaptability and productivity.

CROP GENOME SEQUENCING STATUS

After the commonly-used for sequencing changed from the Sanger method to NGS, the number of plants with complete or draft genome sequences dramatically increased. *Arabidopsis thaliana* was the first plant to be completely sequenced, and sequencing was performed by the Arabidopsis Genome Initiative (AGI, 2000). Next, rice genome sequences became available (Yu *et al*., 2002; International Rice Genome Sequencing Project, 2005). Since then, the sequences of many important crop species, such as grape, sorghum, maize and soybean, became available from studies that used the traditional Sanger method and NGS (Jallion *et al*., 2007; Paterson *et al*., 2009; Schnable *et al*., 2009; Schmutz *et al*., 2010). Genome sequencing projects involving the sequencing of many other important crop species (e.g. oil palm, banana, cotton, barley and wheat) are still in progress (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi). As of early July 2012, 42 genome sequences for 39 crop species were publicly available, and 12 crop sequencing projects were underway or not publicly available (Table 1). Of the 39 crop species with available genome sequences, most were sequenced after 2005, when NGS technology was developed. Recently, the tomato genome sequence has been published using information from studies that employed NGS as well as Sanger technology (The Tomato Genome Consortium, 2012).

Crop genome sequencing is one of the beneficiaries of the rapid development of NGS technology. With lower costs and shorter time requirements, the quality of whole-genome sequencing of crops has been improved. Also, the whole-genome sequencing of many crop plants has enabled the progression of plant evolution studies from the gene to the nucleotide level. This will be helpful for understanding the complexity of existing genomes and the strong relationships between genotypes and evolution. Because whole-genome duplications and structural variations in chromosomes played a prominent role in plant evolution, the development of NGS technology may lead to the identification of new genes with new functions by investigating the functional and evolutionary divergence among plant species. Furthermore, a direct comparison between crop genome sequences has already led to the identification of conserved elements and species-specific differences that underlie unique traits (Barabaschi *et al*., 2012).

Table 1. List of crop species that have been sequenced, along with their general information.

Common name	Species	Genotype	Chromosome (n)	Genome size (Mb)	Sequencing Strategy*	Sequence Coverage	Reference
Amborella	*Amborella trichopoda*	-	-	870	-	-	http://www.amborella.org/
Apple	*Malus x domestica*	Golden Delicious	17	742.3	WGS	16.9x	Velasco *et al.* (2010)
Barbados nut, purging nut, physic nut	*Jatropha curcas*	-	11	410	BAC-by-BAC & WGS	100x	Sato *et al.* (2011)
Barrel clover	*Medicago truncatula*	Mt3.5	8	~454 to 526	BAC-by-BAC & WGS (NGS)	-	Young *et al.* (2011)
Bottle gourd	*Lagenaria siceraria*	Hangzhou gourd	11	334	WGS (NGS)		Xu *et al.* (2011a)
Cacao / chocolate	*Theobroma cacao*	B97-61/B2	10	430	WGS (NGS) & BAC-by-BAC	16.5x, 44x & 0.2x	Argout *et al.* (2011)
Cassava	*Manihot esculenta*	AM560-2	18	760	WGS	-	http://www.phytozome.net/cassava.php
Castor bean	*Ricinus communis*	Hale	10	350	WGS	4.6x	Chan *et al.* (2010)
Clementine mandarin	*Citrus clementina*	-	18	296	-	6.5x	http://www.phytozome.net/clementine.php
Columbine	*Aquilegia coerulea*	Goldsmith	-	302	-	8x	http://www.phytozome.net/aquilegia.php
Common bean	*Phaseolus vulgaris*	-	11	486.9	WGS	20x	http://www.phytozome.org/commonbean.php
Cotton	*Gossypium raimonddi*	-	13	750	BAC-by-BAC & WGS (NGS)	1.52x & 14.95	http://www.phytozome.net/cotton.php
Cucumber	*Cucumis sativus*	IL 9930	7	367	BAC-by-BAC & WGS	3.9 & 68.3	Huang *et al.* (2009b)

Date palm	*Phoenix dactylifera*	Khalas, Alrijal, Khalt	18	658	WGS	53.4x,10.7 x, 11.2x	Al-Dous *et al.* (2011)
Eucalyptus/ Rose gum tree	*Eucalyptus grandis*	BRASUZ1	11	691	WGS	8x	http://www.phytozome.net/eucalyptus.php
Flax	*Linum usitatissimum*	-	15	350	WGS	50x	http://www.phytozome.net/flax http://abstracts.aspb.org/pb2010/public/P09/P09014.html
Foxtail millet	*Setaria italica*	Yugul	9	510	WGS	8.3x	Bennetzen *et al.* (2012)
Grapes	*Vitis vinifera*	PN40024	19	487	WGS	8.4x	Jaillon *et al.* (2007)
		ENTAV115		505	WGS & SBS	6.5x & 4.2x	Velasco *et al.* (2007)
Lotus	*Lotus japonicus*	Miyakojima MG-20	6	472	BAC-by-BAC	8.4x	Sato *et al.* (2008)
Maize	*Zea mays* ssp. *mays* L.	B73	10	2300	BAC-by-BAC	6x	Schnable *et al.*, 2009
	Zea mays ssp. *parviglumis*	Palomero Toluqueño EDMX-2233		2100	WGS	3.2x	Calzada *et al.* (2009)
Marijuana	*Cannabis sativa*	marijuana, hemp	10	534	WGS (NGS)	110x	Van Bakel *et al.* (2011)
Melon	*Cucumis melo* L.	DHL92 (Songwhan Charmi x Piel de Sapo)	12	450	WGS (NGS)	13.52x	Garcia-Mas *et al.* (2012)
Monkey flowers	*Mimulus guttatus*	IM62	14	430	WGS	-	http://www.phytozome.net/mimulus
Mustard, field mustard, rape mustard	*Brassica rapa*	Chiifu-401-42	10	284	WGS (NGS)	72x	Wang *et al.* (2011)
Papaya	*Carica papaya*	SunUp	9	372	BAC-by-BAC	3x	Ming *et al.* (2008)

Peach	*Prunus persica*	Lovell	8	230	WGS (NGS)	7.7x	Ahmad *et al.* (2011)
Pigeon pea	*Cajanus cajan*	ICPL 87119	11	833.07	WGS (NGS)	~163.4x	Varshney *et al.*(2012)
Poplar	*Populus trichocarpa*	Nisqually 1	19	485	WGS	7.5x	Tuskan *et al.* (2006)
Potato	*Solanum tuberosum* L.	DM1-3-516 R44	12	844	WGS	-	Xu *et al.* (2011b)
Purple false brome	*Brachypodium distachyon*	Bd21	5	272	WGS	9.4x	Vogel *et al.* (2010)
Rice	*Oryza sativa* L. ssp. *japonica*	Nipponbare	12	433	WGS	6x	Goff *et al.* (2002)
				389	BAC-by-BAC	10x	International Rice Genome Sequencing Project, 2005
	Oryza sativa L. ssp. *indica*	93-11		466	WGS	6.3x	Yu *et al.* (2002)
Shepherd's purse	*Capsella rubella*	-	8	135	WGS	22x	http://www.phytozome.net/capsella.php
Sorghum	*Sorghum bicolor*	BTx623	10	730	WGS	8.5x	Paterson *et al.* (2009)
Soybean	*Glycine max*	Williams 82	20	1,115	WGS	8.04x	Schmutz *et al.* (2010)
Strawberry	*Fragaria vesca*	Hawaii 4	7	240	WGS (NGS)	39x	Shulaev *et al.* (2011)
Sugar beet	*Beta vulgaris*	KWS2320	9	758	BAC-by-BAC & WGS	-	http://bvseq.molgen.mpg.de/Genome/start.genome.shtml
Sweet orange	*Citrus sinensis*	Ridge Pineapple	18	319	WGS (NGS) & BAC-by-BAC	- & 1.2x	http://www.phytozome.net/citrus.phphttp://www.jgi.doe.gov/sequencing/why/3128.html
Tomato	*Solanum lycopersicum*	Heinz 1706	12	900	BAC-by-BAC & WGS	22x	The Tomato Genome Consortium, 2012
Watermelon	*Citrullus lanatus*	Lanatus	11	430	WGS (NGS)	107.4x	Ren *et al.* (2012)

*BAC, bacteria artificial clone; NGS, next-generation sequencing; WGS, whole genome sequencing; SBS, sequencing by synthesis

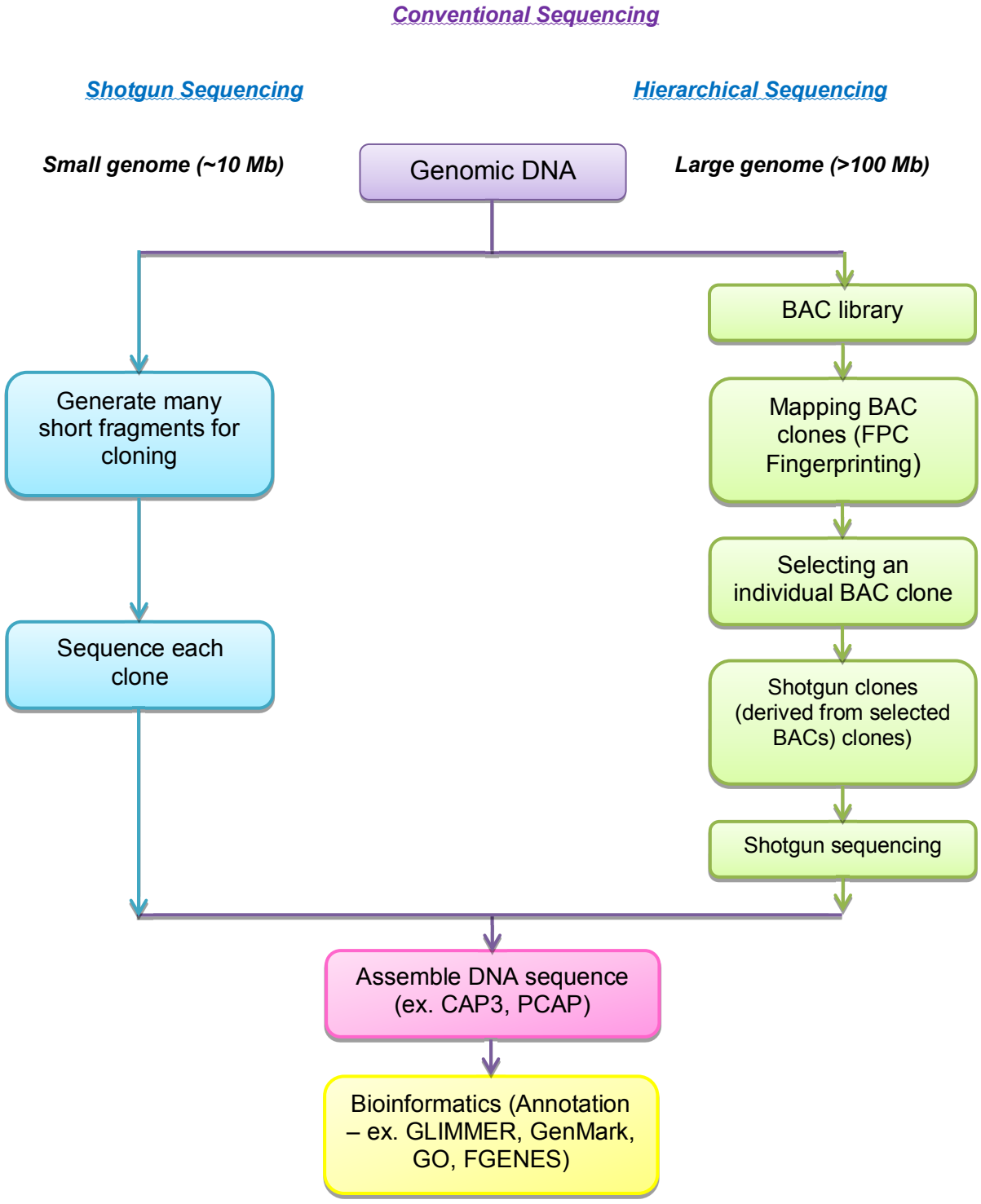
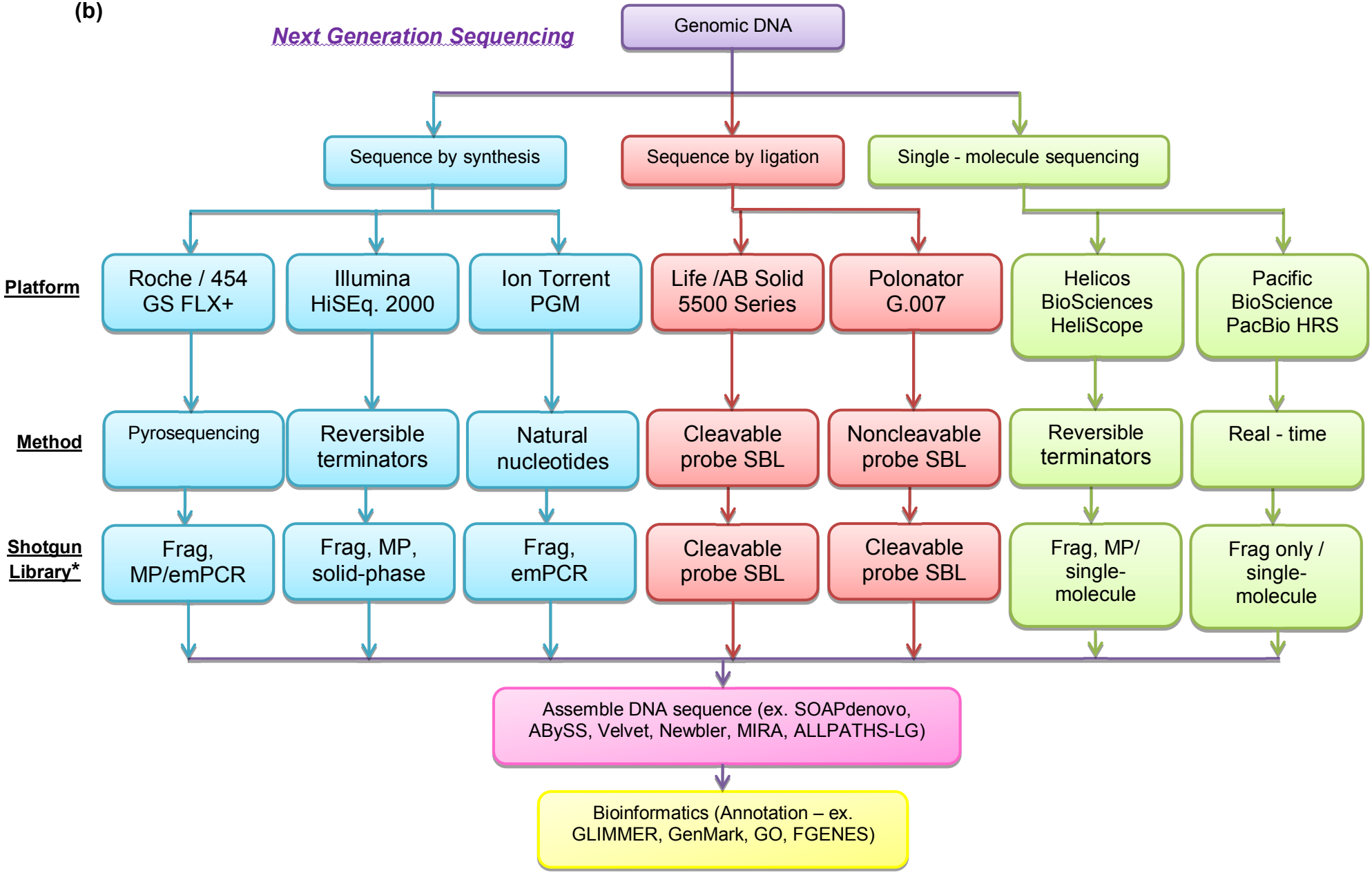


Figure 1. Overview of whole-genome sequencing by (a) conventional sequencing method and (b) next-generation sequencing method. *Frag - fragment; MP - mate-pair; emPCR - emulsion PCR; SBL - sequencing by ligation.

(b)


Next Generation Sequencing
Genomic DNA
Sequence by synthesis
Sequence by ligation
Single - molecule sequencing
Platform
Roche / 454 GS FLX+
Illumina HiSEq. 2000
Ion Torrent PGM
Life /AB Solid 5500 Series
Polonator G.007
Helicos BioSciences HeliScope
Pacific BioScience PacBio HRS
Method
Pyrosequencing
Reversible terminators
Natural nucleotides
Cleavable probe SBL
Noncleavable probe SBL
Reversible terminators
Real - time
Shotgun Library*
Frag, MP/emPCR
Frag, MP, solid-phase
Frag, emPCR
Cleavable probe SBL
Cleavable probe SBL
Frag, MP/ single-molecule
Frag only / single-molecule
Assemble DNA sequence (ex. SOAPdenovo, ABySS, Velvet, Newbler, MIRA, ALLPATHS-LG)
Bioinformatics (Annotation – ex. GLIMMER, GenMark, GO, FGENES)

WHOLE-GENOME SEQUENCING TECHNOLOGY

De novo assembly

Because reference genome sequences were previously not available, gene discovery in crop species such as wheat and barley was totally dependent on unassembled genome sequences and expressed sequence tags (ESTs). Also, single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) molecular markers could be developed from these sequences in the absence of a reference genome sequence. This strategy is therefore suitable for obtaining sufficient data from orphan or less-studied crops (Edwards and Batley, 2010).

The early clone-by clone strategy, which employed a physical map of cosmid or bacterial artificial chromosome (BAC) clones, was the typical *de novo* assembly method for whole-genome sequencing projects before the development of NGS (Figure 1a). But now, with the advent of the WGS strategy, different sizes of inserts from genomic DNA are constructed as mate pairs and sequenced from both ends. Due to the rapid expansion of NGS capacity compared to Sanger sequencing, hybrid methods based on pyrosequencing and Sanger long pairs have become the current strategy for *de novo* assembly (Jackson *et al*., 2011).

This current strategy is much faster than conventional BAC library-based sequencing and has helped lead to the development of many *de novo* assembly algorithms, such as ABySS (Simpson *et al*., 2009), Velvet (Zerbino and Birney, 2008), ALLPATHS (Butler *et al*., 2008), SOAPdenovo (http://soap.genomics.org.cn), CLC bio's *de novo* assembler (http://clcdenovo.com/index.php), and others. Because WGS has been performed on a massive scale using different sequencing platforms, Lim *et al*. (2012) suggested that sequences could be effectively integrated by *de novo* assembly. After contigs were generated from different *de novo* algorithms, scaffolds could be constructed using SSPACE software (Boetzer *et al*., 2011) for contig ordering via hybrid contigs using MIRA assembler (http://sourceforge.net/apps/mediawiki/mira-assembler/index.php?title=Main_Page). By placing short reads into larger scaffolds using these bioinformatics tools, many crop genomics researchers have been attempting to sequence various crops. For example, the woodland strawberry genome (240 Mb) was *de novo* assembled with 39x sequence depth, and assembled contigs were successfully positioned on seven pseudo-chromosomes (Shulaev *et al*., 2011). Thus, draft genome sequences by *de novo* assembly are adequate for building gene catalogs and studying interspecific comparative genomics (Feuillet *et al*., 2011).

Resequencing: reference mapping

Because reference genome sequences of several crop species via *de novo* assembly of whole-genome sequences are now available, candidate genes of agronomic importance and SNPs between the reference genome and sequences from different cultivars can more rapidly and easily be identified using bioinformatics tools. This rapid and efficient resequencing method accelerates whole-genome sequencing of not only individuals but also populations, thus leading to the development of molecular markers and the construction of saturated molecular genetic maps (Gao *et al*., 2012).

In terms of population genetics, the WGS strategy using NGS is very helpful for studying genetic variations among populations, population structure and linkage disequilibrium, all of which are important for crop breeding programs. By performing NGS on an individual of a crop species that has a reference genome, WGS at the individual and population levels can easily be used to identify genetic variations such as SNPs, insertions/deletions (indels), structural variation (SV) including translocations and chromosome fusions, and copy number variations (CNVs) (Feuillet *et al*., 2011; Gao *et al*., 2012). All of this information is also beneficial for studying the evolutionary history of a crop species, adaptation to the various environmental conditions, and natural selection at the population level. For example, Kim *et al*. (2010) resequenced wild soybean (*Glycine soja* var. IT182932) and were able to predict about 2.5 million SNPs between cultivated soybean (*G. max* var. Williams 82, Schmutz *et al*., 2010) and wild soybean. Among 2.5 million SNPs,

86,236 SNPs were classified as coding sequence variants, and more than 196,000 indels (-35 to +14-bp) were identified and located throughout the *G. soja* genome. In terms of SV, 5,794 deletions and 194 inversions in the range of 0.1–100 kb were detected, and 8,554 insertions were predicted in the *G. soja* genome. A difference of 0.31% was found between *G. max* and *G. soja* in the 937.5 Mb genome sequences that were examined, and the estimated theoretical divergence time suggested that *G. soja* and *G. max* diverged at 0.267 ± 0.03 MYA and that the divergence between IT182932 (*G. soja*) and Williams 82 (*G. max*) pre-dated soybean domestication. Thus, in this study, the resequencing strategy was used to elucidate the genetic history of soybean by identifying SNPs, indels and SVs.

The rapid characterization of genetic variations by NGS will contribute to the identification of agronomically important traits and to the shortening of crop breeding times with the use of marker-assisted selection (MAS) (Gao *et al*., 2012). Until NGS was developed, quantitative trait loci (QTL) mapping and analysis using genetic linkage maps were common strategies used to study agronomic traits. Because of the low densities of molecular markers, restriction fragment length polymorphisms (RFLPs) and SSRs used to produce genetic linkage maps, the locations of QTLs in these maps could not provide enough information to accurately map the positions of genes that regulate the QTLs. Now, genome resequencing is the strategy of choice for marker development and QTL mapping (Gao *et al*., 2012). Using this high-throughput method, rice recombination inbred lines (RILs) were genotyped by resequencing the entire genome (Huang *et al*., 2009a) or by using lower coverage sequencing (0.06X) to construct a high-density genetic map (Yu *et al*., 2011). A chromosome segment substitution line population in rice was also resequenced at the lower coverage (0.13X) to construct a genetic map (Xu *et al*., 2010). Since numerous molecular markers can be produced in a short period of time, the genome-wide genotyping method, genotyping-by-sequencing (GBS), using reference mapping may become a popular approach for MAS in crop breeding programs.

The resequencing method is also a powerful tool for identifying mutation sites in mutant populations (Gao *et al*., 2012). Because of the rapid development of the NGS and bioinformatics software, it has become possible to accurately and efficiently identify mutated sites by comparing mutants with parental strains. SHOREmap was developed for identifying mutations in EMS-mutagenized F_2 Arabidopsis populations. In addition, by performing QTL mapping, large deletions and recessive mutations could be identified using this software (Schneeberger *et al*., 2009). Using another EMS-mutagenized F_2 Arabidopsis population, WGS was performed, and the mutation sites were localized (Cuperus *et al*., 2010). An Arabidopsis mutant backcross line was genome resequenced, and SNPs were detected by comparing the expression data for the mutant with that of the parental line (Ashelford *et al*., 2011). In maize, SNPs were identified among RILs using NGS. These SNPs could be used for developing genetic markers, constructing a genetic map via genotyping, and mapping mutants and QTLs (Liu *et al*., 2010).

APPLICATIONS TO CROP IMPROVEMENT

Whole-genome analysis

Due to the rapid development of bioinformatics tools, sequences generated by NGS can be analyzed at the whole-genome level. Krzywinski *et al*. (2009) developed a visualization tool called Circos written in Perl. Using data generated from NGS, sequence alignment, hybridization arrays, genome mapping and genotyping studies, this program is able to identify and analyze similarities and differences between genomes. By identifying BACs spanning rearrangement breakpoints and sequence contigs containing breakpoints, breakpoint structures can be explored on a local scale. Figure 2 shows the example of the use of Circos. The positions of QTLs related to seed yield can be shown within one circular map of all 20 soybean chromosomes. And, the recent duplication of the regions nearby to these QTLs was observed within one map of soybean whole-

genome. Some recent papers also used Circos for whole-genome analysis. Van *et al*. (2012) used Circos for visualizing the contigs of *G. max* var. Sinpaldalkong 2 mapped onto the reference genome (*G. max* var. Williams 82), suggesting that there was a recent duplication of the Sinpaldalkong 2 genome. The whole genome can be magnified and analyzed on a global scale by exploring whole-genome syntenic profiles among chromosomes within a species or between two genomes. After homologous blocks were identified within the soybean genome, the homologous relationships between 20 soybean chromosomes were visualized by Circos (Schmutz *et al*., 2010). The circular maps presented an extremely high level of homologous relationships, and homologous blocks were found to be located on two or more soybean chromosomes. Varshney *et al*. (2012) compared the pigeonpea and soybean genomes using Circos. Each pigeonpea chromosome was similar to two or more than two chromosomes in soybean, showing the extensive synteny between these two genomes.

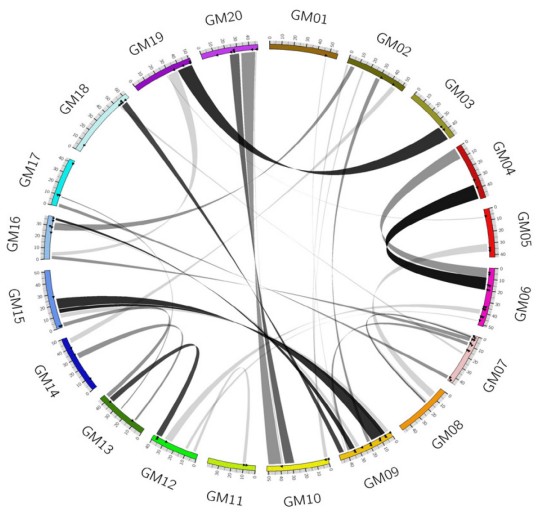


Figure 2. Whole genome analysis of 20 soybean (G. max, var. Williams 82) chromosomes with QTLs related to seed yield (black triangles). The information of QTLs were obtained from SoyBase (http://soybase.org). Using duplication region data of Williams 82 at Phytozome (http://www.phytozome.net/soybean.php), the inside of the outer layer visualizes duplicated positions of the *G. max* genome, represented as ribbons after similar duplicated regions are grouped as bundles.

Transcriptome analysis

Microarray and serial analysis of gene expression (SAGE) were used for studying gene expression before the NGS era. Now, however, RNA-Seq is the popular choice for gene expression profiling via the sequencing of a whole-transcriptomes using NGS (Varshney *et al*., 2009; Jain, 2011; Strickler *et al*., 2012). Because the depth of sequence coverage is considered to be proportional to the expression level of the corresponding gene, even rare and novel transcripts can now be identified. Many researchers have therefore tried to elucidate a nearly complete picture of gene expression profiles under different environmental conditions using transcriptome analysis. In addition to characterizing genes under various conditions, SNPs, alternative splicing and SV could be also studied (Alagna *et al*., 2009; Lister *et al*., 2009; Filichkin *et al*., 2010).

RNA-Seq experiments are carefully designed by addressing the questions of interest. Making good sample choices based on factors such as species background information is a very important step in pre-sequencing because highly homozygous lines can discriminate sequencing errors, heterozygosity and duplicate genes (Strickler *et al*., 2012). Kim *et al*. (2011) used near isogenic lines (NILs) as sample species to ensure that the same genetic background was present between the control and treatment plants. Tissue treatment and selection and the use of appropriate sequencing platforms are also important considerations (Strickler *et al*., 2012). After NGS and quality control were performed, high-quality trimmed reads were mapped onto the reference genome or *de novo* assembled (Jain, 2011; Strickler *et al*., 2012). Computer programs, such as Eland, SOAP, MAQ, RMAP, SSAHA2, SHRiMP, Stampy, TopHat, RNA-MATE, BWA and Bowtie, were available for mapping the assembly of transcriptomes. CAP3, CLC Bio, MIRA, TGICL, BLAT and other programs were used for *de novo* assembly and clustering analysis of transcriptomes. Reads per kilobase or per million of mapped reads was used for normalizing read counts as a quantitative normalized measure between samples/conditions. Differential gene expression was then profiled using Cufflinks, ALEXA-seq,

DESeq, DEGseq, Myrna, MMSEQ, rQuant, edgeR and ERANGE software. Using a set of differentially expressed genes, transcriptome characterization and gene annotation were performed as transcriptomics downstream analysis. BLAST and Blast2GO were typically used for gene annotation, and this gene annotation can be integrated with metabolic pathways. Also, SNP detection is a common application of RNA-Seq (Strickler *et al*., 2012).

Marker development and association studies

Compared to the traditional Sanger method, NGS is helpful for discovering and developing SSR or microsatellite loci efficiently. These markers are still commonly used for the construction of linkage maps, QTL mapping, MAS, cultivar fingerprinting, and studying gene flow. Zalapa *et al*. (2012) listed plant SSR markers that were recently developed using the Sanger and NGS methods, but still, the majority of SSR markers were identified using Sanger technology. Also, using cranberry (*Vaccinium macrocarpon*) NGS sequences, SSRs were identified from the raw data before contig assembly, using bioinformatics methods to efficiently design primers. Due to difficulties of DNA sequence assembly with repeats, the use of GS-FLX 454 technology, rather than Illumina technology, is preferred for the isolation of SSR loci in plants due to its longer read length. The development of bioinformatics tools for the identification of SSRs is also a challenge (Zalapa *et al*. 2012).

With the rapid development of NGS technologies, tremendous numbers of molecular markers like SNPs have been identified, and SNP-based resources are publically available for crop breeding programs (Kilian and Graner, 2012). Genome-wide marker discovery by NGS has become more feasible using new methods, such as reduced-representation libraries (Hyten *et al*., 2010), complexity reduction of polymorphic sequences (van Orsouw *et al*., 2007), restriction site-associated DNA sequencing (Baxter *et al*., 2011), and low-coverage sequencing for genotyping (Huang *et al*., 2009a; Elshire *et al*., 2011). Since genome-wide markers were quickly developed in large quantities using NGS technologies, association mapping, patterns of natural population structure and the decay of linkage disequilibrium (LD) can be studied more easily by whole-genome scanning using NGS (Varshney *et al*., 2009; Kilian and Graner, 2012). Also, whole-genome scanning has been performed using specially designed mapping populations. For example, Tian *et al*. (2011) studied the genetic basis of traits and identified some genes related to leaf architecture, which is important for efficient light capture. Leaf architecture changes depending on leaf size, leaf angle and time of day. In another study in maize, which employed nested association mapping, QTLs for resistance to southern leaf blight were identified and limited LD was shown to occur around the regions of some SNPs linked to this disease (Kump *et al*., 2011).

Marker-assisted breeding

As previously described, the selection of material to be sequenced is very important for crop improvement. Different MAS strategies are used depending on the specific types of traits and breeding programs (Xu *et al*., 2012). Two major marker-assisted backcrossing (MABC) methods, marker-assisted foreground selection and background selection, are commonly used for breeding major gene-controlled traits. Marker-assisted foreground selection uses two-to-ten markers for each target trait; both single and multiple traits are used for introgression with a population size of several hundred. However, marker-assisted background selection requires at least 200 additional markers for the selection of target genes from the same population size (Xu *et al*., 2012).

As one of the selection methods for complex traits, marker-assisted recurrent selection (MARS) uses markers from each generation of the population and is considered to be the most beneficial method for breeding traits controlled by a moderately large number of QTLs (Bernardo and Charcosset, 2006). The prior information of QTL is very useful and selection must be made by significant marker-trait association in the process of MARS. Genomic selection (GS) is another selection scheme for breeding complex traits with three steps, including prediction model training and

validation, breeding value prediction of single-crosses and selection based on these predictions (Xu *et al*., 2012). However reports of MARS and GS have largely been through simulation studies rather than empirical results.

Throughout selection using these methods, a large number of molecular markers and individual lines in a population should be studied. By using advanced NGS technologies rather than earlier methods, the cost of genotyping with these markers and populations are dramatically decreased. But, many breeding programs still demand much lower cost per samples. GBS is performed using libraries of reduced genome complexity that are created with the use of restriction enzymes. This simple, specific and reproducible method has become a popular tool for population genotyping using NGS. High-throughput, large-scale genotyping methods using GBS have been introduced, and these strategies have already been applied to recombinant inbred lines (Huang *et al*., 2010; Elshire *et al*., 2011).

After genotyping by NGS, high-throughput and precise phenotyping is required for the genetic analysis of traits examined by MAS in crop breeding programs. Automated platforms in growth chambers or greenhouses are designed for phenotyping throughout the life cycle of the plant, and these plant materials are good resources for metabolomics and quantitative phenotyping (Bergelson and Roux, 2010; Massonnet *et al*., 2010).

Genetic diversity

To help counteract the loss of genetic diversity caused by agricultural practices, plant genetic resources (PGRs) including cultivars, landraces, wild species closely related to cultivated varieties, breeder's elite lines and mutants have been collected to increase the genetic variability of plants used in crop breeding programs. The collection of these PGRS was also performed to enhance future food security (Van *et al*., 2011; Barabaschi *et al*., 2012; Kilian and Graner, 2012). Since a barcoding system was developed for use with NGS technology, many individual plants could now be sequenced simultaneously at a lower cost. Sequencing at lower levels of coverage or sequencing only targeted regions of DNA are practical strategies for studying population genetics, conservation genetics and molecular ecology. The genomics era provides a golden opportunity for categorizing PGRs by SNP marker instead of by phenotype. Therefore, the resequencing method using a wide range of PCR products is now affordable and enables genome-wide marker development, genotyping within populations and the evaluation of genetic diversity (Barabaschi *et al*., 2012; Kilian and Graner, 2012).

FUTURE DIRECTIONS

NGS technology provides a golden opportunity for understanding biological systems in crops. Compared to the traditional Sanger sequencing method, the cost of NGS is dramatically decreased, and employing advanced NGS technology is more feasible for many researchers who wish to sequence crop genomes. Some cash crops were considered to be less-studied/orphan crops due to a lack of sequence and marker information. But now, by employing *de novo* assembly strategies, whole-genome sequences of less-studied/orphan crops are becoming feasible for crop improvement. Also, more molecular markers like SNPs and indels have been rapidly developed at lower cost, and these markers are easily applicable to MAS in crop breeding programs.

In this paper, several different applications for crop improvement were discussed. The identification of genes related to agronomic traits by crop breeding is important, but experiments for understanding the functions of these identified genes should be performed and could be applied to crop improvement in breeding programs. Although the development of bioinformatics tools and storage space for huge sequence data are still a challenge for NGS, the speed of crop improvement will be much faster than before because the third generation of sequencing platforms, such as HeliScope, Ion Torrent, single molecular real-time sequencing and Oxford Nanopore, have already been developed.

ACKNOWLEDGMENTS

This research was supported by a grant from the Next-Generation BioGreen 21 Program (No. PJ008117) of the Rural Development Administration, Republic of Korea.

REFERENCES

AGI TAGI (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, Arondel V, Ohlrogge J, Saie IJ, Suliman-Elmeer KM, Bennetzen JK, Kruegger RR, Malek JA (2011). *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat. Biotechnol.* 29: 521-527.

Alagna F, D'Agostino N, Torchia L, Servilli M, Rao R, Pietrella M, Giuliano G, Chiusano ML, Baldoni L, Perrotta G (2009). Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* 10: 399.

Argout X, Salse J, Aury JM, Guiltinan MJ, *et al*. [61 authors] (2011). The genome of Theobroma cacao. *Nat. Genet.* 43: 101-108.

Ariyadasa R, Stein N (2012). Advances in BAC-based physical mapping and map integration strategies in plants. *J. Biomed. Biotechnol.* doi:10.1155/2012/184854.

Ashelford K, Eriksson ME, Allen CM, D'Amore R, Johansson M, Gould P, Kay S, Miller AJ, Hall N, Hall A (2011). Full genome resequencing reveals a novel circadian clock mutation in Arabidopsis. *Genome Biol.* 12: R28.

van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, Page JE (2011). The draft genome and transcriptome of *Cannabis sativa*. *Genome Biol* 12: R102.

Balmford A, Green RE, Scharlemann JPW (2005). Sparing land for nature: exploring the potential impact of changes in agricultural yield on the area needed for crop production. *Global Change Biol.* 11: 1594-1605.

Barabaschi D, Guerra D, Lacrima K, Laino P, Michelotti V, Urso S, Vale G, Cattivelli L (2012). Emerging knowledge from genome sequencing of crop species. *Mol. Biotechnol.* 50: 250-266.

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD, Blaxter ML (2011). Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* 6: e19315.

Bennetzen JL, Schmutz J, Wang H, Percifield R, et al. [34 authors] (2012). *Nat. Biotechnol.* 30: 555-561.

Bergelson J, Roux F (2010). Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nat. Rev. Genet.* 11: 867-879.

Bernardo R, Charcosset A (2006). Usefulness of gene information in marker-assisted recurrent selection: a simulation appraisal. *Crop Sci.* 46: 614-621.

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

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

Calzada JPV, de la Vega OM, Hernandez-Guzman G, Ibarra-Laclette E, Alvarez-Mejia C, Vega-Arreguin JC, Jimenez-Moraila B, Fernandez-Cortes A, Corona-Armenta G, Herrera-Estrella L, Herrera-Estrella A (2009). The Palomero genome suggests metal effects on domestication. *Science* 326: 1078.

Chan AP, Crabtree J, Zhao Q, Lorenzi H, *et al*. [18 authors] (2010). Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.* 28: 951-956.

Chang CC, Hsu SH (2011). Food security – Global trends and region perspective with reference to East Asia. Agricultural & Applied Economics Association 2011 AAEA & NAREA Joint Annual Meeting, Pittsburgh, PA, USA.

Cuperus JT, Montgomery TA, Fahlgren N, Burke RT, Townsend T, Sullivan CM, Carrington JC (2010). Identification of MIR390a precursor processing-defective mutants in Arabidopsis by direct genome sequencing. *Proc. Natl. Acad. Sci. USA* 107: 466-471.

Edwards D, Batley J (2010). Plant genome sequencing: applications for crop improvement. *Plant Biotechnol. J.* 8: 2-9.

Egan AN, Schlueter J, Spooner DM (2012). Applications of next-generation sequencing in plant biology. *Amer. J. Bot.* 99: 175-185.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011). A robust, simple genotyping-by sequencing (GBS) approach for high diversity species. *PLoS ONE* 6: e19379.

Fargione J, Hill J, Tilman D, Polasky S, Hawthorne P (2008). Land clearing and the biofuel carbon debt. *Science* 319: 1235-1238.

Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K (2011). Crop genome sequencing: lessons and rationales. *Trend Plant Sci.* 16: 77-88.

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* 20: 45-58.

Gao Q, Yue G, Li W, Wang J, Xu J, Yin Y (2012). Recent progress using high-through put sequencing technologies in plant molecular breeding. *J. Integr. Plant Biol.* 54: 215-227.

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, *et al.* [34 authors] (2012). The genome of melon (*Cucumis melo* L.). *Proc. Natl. Acad. Sci. USA* 109: 11872-11877.

Godfray HCJ, Beddubgton JR, Crute IR, Haddad L, Lawrence D, Muir JF, Pretty J, Robinson S, Thomas SM, Toulmin C (2010). Food security: the challenge of feeding 9 billion people. *Science* 327: 812-818.

Goff SA, Ricke D, Lan TH, Presting G, *et al.* [55 authors] (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92-100.

Harmon-Smith K, Lail H, Tice J, Schmutz, *et al.* [162 authors] (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463: 763-768.

Huang S, Li R, Zhang Z, Li L, *et al.* [85 authors] (2009a). The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* 41: 1275-1281.

Huang X, Feng Q, Qian Q, Zhao Q, Wang L,Wang A, Guan J, Fan D, Weng Q, Huang T, Dong G, Sang T, Han B (2009b). High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19: 1068-1076.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42: 961-967.

Hyten D, Cannon S, Song Q, Weeks N, Fickus EW, Shoemaker RC, Specht JE, Farmer AD, May GD, Cregan PB (2010). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* 11: 38.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* 436: 793-800.

Jackson SA, Iwata A, Lee SH, Schmutz J, Shoemaker R (2011). Sequencing crop genomes: approaches and applications. *New Phytologist* 191: 915-925.

Jain M, (2011). Next-generation sequencing technologies for gene expression profiling in plants. *Brief. Func. Genomics* 2: 63-70.

Jallion O, Aury JM, Noel B, Policriti A, *et al.* [55 authors] (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463-467.

Kilian B, Graner A (2012). NGS technologies for analyzing germplasm diversity in genebanks. *Brief. Func. Genomics* 2: 38-50.

Kim KH, Kang YJ, Kim DH, Yoon MY, Moon JK, Kim MY, Van K, Lee SH (2011). RNA-Seq analysis of a soybean near-isogenic line carrying bacterial leaf pustule-resistant and – susceptible alleles. *DNA Res.* 18: 483-497.

Kim MY, Lee S, Van K, Kim TH, *et al.* [29 authors] (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc. Natl. Acad. Sci. USA* 107: 22032-22037.

Kim MY, Van K, Kang YJ, Kim KH, Lee SH (2012). Tracing soybean domestication history: From nucleotide to genome. *Breed. Sci.* 61: 445-452.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009). Circos: An information aesthetic for comparative genomics. *Genome Res.* 19: 1639-1645.

Kump KL, Bradbury PJ, Wisser RJ, Bickler ES, Belcher AR, Oropeza-Rosas MA, Zwonitazer JC, Kresovish S, McMullen MD, Ware D, Balint-Kurti PJ and Holland JB (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43: 163-168.

Lee HC, Lai K, Lorenc MT, Imelfort M, Duran C, Edwards D (2012). Bioinformatics tools and databases for analysis of next-generation

sequence data. *Brief. Func. Genomics* 11: 12-24.

Lim JS, Choi BS, Lee JS, Shin C, Yang TJ, Rhee JS, Lee JS, Choi IY (2012). Survey of the applications of NGS to whole-genome sequencing and expression profiling. *Genomics & Informatics* 10: 1-8.

Lister R, Gregory BD, Ecker JR (2009). Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol*. 12: 107-118.

Liu S, Chen HD, Marketvitch I, Shirmer R, Emrich SJ, Dietrich CR, Barbazuk WB, Springer NM, Schnable PS (2010). High-throughput genetic mapping of mutants via quantitative single nucleotide polymorphism typing. *Genetics* 184: 19-26.

Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrett Palanisamy TN, Chinnaiyan AM (2009). Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458: 97-101.

Massonnet C, Vile D, Fabre J, Hannah MA, *et al.* [27 authors] (2010). Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three *Arabidopsis* accessions cultivated in ten laboratories. *Plant Physiol.* 152: 2142-2157.

Matsumoto T, Wu J, Kanamori H, Katayose Y, *et al.* [92 authors] (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452: 991-996.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, *et al.* [45 authors] **(**2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457: 551-556.

Ren Y, Hong Z, Kou Q, Jiang J, Guo S, Zhang H, Hou W, Zou X, Sun H, Gong G, Levi A, Xu Y (2012). A high resolution genetic map anchoring scaffolds of the sequenced watermelon genome. *PLoS One* 7: e29453.

Sato S, Hirakawa H, Isobe S, Fukai E, *et al.* [36 authors] (2011). Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Res.* 18: 65-76.

Sato S, Nakamura Y, Kaneko T, Asamizu E, *et al.* [29 authors] (2008). Genome structure of the legume, *Lotus japonicus*. *DNA Res.* 15: 227-239.

Schmutz J, Cannon SB, Schlueter J, Ma J, *et al.* [44 authors] (2010). Genome sequence of the paleopolyploid soybean. *Nature* 463: 178-183.

Schnable PS, Ware D, Fulton RS, Stein JC, *et al.* [152 authors] (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326: 1112-1115.

Schneeberger K, Ossowski S, Lanz C, Juul T, Paterson AH, Nielsen KL, Jorgensen JE, Weigel D, Andersen SU (2009). SHOREmap: Simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* 6: 550-551.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, *et al.* [72 authors] (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43: 109-116.

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

Strickler SR, Bombarely A, Mueller LA (2012). Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *Amer. J. Bot*. 99: 257-266.

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

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43: 159-162.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, *et al.* [111 authors] (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596-1604.

Van K, Kang YJ, Shim SR, Lee SH (2012). Genome-wide scan of the soybean genome using degenerate oligonucleotide primed PCR: an example for studying large complex genome structure. *Genes & Genomics* 34: 467-474.

Van K, Kim DH, Shin JH, Lee SH (2011). Genomics of plant genetic resources: past, present, and future. *Plant Genet. Resour.* 9: 155-158.

Van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, van der Poel H, van Oeveren J, Verstegen H, van Eijk MJT (2007). Complexity reduction of polymorphic sequences (CRoPS^{TM}): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* 2: e1172.

Varshney RK, Bansal KC, Aggarwal PK, Datta SK, Craufurd PQ (2011). Agricultural biotechnology for crop improvement in a variable climate: hope or hype? *Trend. Plant Sci.* 16: 363-371.

Varshney RK, Chen W, Li Y, Bharti AK, *et al.* [30 authors] (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* 30: 83-89.

Varshney RK, Nayak SN, May GD, Jackson SA (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trend. Biotechnol.* 27: 522-530.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, *et al.* [86 authors] (2010). The genome of the domesticated apple (*Malus* x *domestica* Borkh.). *Nat. Genet.* 42: 833-839.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, *et al.* [57 authors] (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2: e1326.

Vogel JP, Garvin DF, Mockler TC, Schmutz J, *et al.* [106 authors] (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43: 1035-1039.

Xu J, Zao Q, Du P, Xu C, Wang B, Feng Q, Liu Q, Tang S, Gu M, Han B, Liang G (2010). Developing high throughput genotyped chromosome segment substitution lines based on population whole-genome re-sequencing in rice (*Oryza sativa* L.). *BMC Genomics* 11: 656.

Xu P, Wu X, Luo J, Wang B, Liu Y, Ehlers JD, Wang S, Lu Z, Li G (2011a). Partial sequencing of the bottle gourd genome reveals markers useful for phylogenetic analysis and breeding. *BMC Genomics* 12: 467.

Xu X, Pan S, Cheng S, Zhang B, *et al.* [94 authors] (2011b). Genome sequence and analysis of the tuber crop potato. *Nature* 475: 189-195.

Xu Y, Lu Y, Xie C, Gao S, Wan J and Prasanna BM (2012). Whole-genome strategies for marker-assisted plant breeding. *Mol. Breed.* 29: 833-854.

Young ND, Debelle F, Oldroyd GE, Geurts R, et al. [124 authors] (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* 480: 520-524.

Yu H, Xie W, Wang J, Xing Y, Xu C, Li X, Xiao J, Zhang Q (2011). Gains in QTL detection using an ultra-high density SNP map based on population sequencing relate to traditional RFLP/SSR markers. *PLoS One* 6: e17595.

Yu J, Hu S, Wang J, Wong GK, *et al.* [85 authors] (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79-92.

Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik Zeldin DE, McCown B, Harbut R, Simon P (2012). Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Amer. J. Bot.* 99: 193-208.

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