Available online at www.sciencedirect.com
SciVerse ScienceDirect
Journal of Genetics and Genomics 39 (2012) 317−324
JGG

REVIEW

# Recent Developments of Genomic Research in Soybean

Ching Chan, Xinpeng Qi, Man-Wah Li, Fuk-Ling Wong, Hon-Ming Lam*

*State Key Laboratory of Agrobiotechnology and School of Life Sciences, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China*

Received 13 December 2011; revised 3 February 2012; accepted 4 February 2012
Available online 18 February 2012

## ABSTRACT

Soybean is an important cash crop with unique and important traits such as the high seed protein and oil contents, and the ability to perform symbiotic nitrogen fixation. A reference genome of cultivated soybeans was established in 2010, followed by whole-genome re-sequencing of wild and cultivated soybean accessions. These efforts revealed unique features of the soybean genome and helped to understand its evolution. Mapping of variations between wild and cultivated soybean genomes were performed. These genomic variations may be related to the process of domestication and human selection. Wild soybean germplasms exhibited high genomic diversity and hence may be an important source of novel genes/alleles. Accumulation of genomic data will help to refine genetic maps and expedite the identification of functional genes. In this review, we summarize the major findings from the whole-genome sequencing projects and discuss the possible impacts on soybean researches and breeding programs. Some emerging areas such as transcriptomic and epigenomic studies will be introduced. In addition, we also tabulated some useful bioinformatics tools that will help the mining of the soybean genomic data.

**KEYWORDS:** Evolution; Genome; Soybean

## 1. INTRODUCTION

Soybean is one of the most important economic crops providing ~70% dietary proteins and ~30% edible oil (Data from American Soybean Association, 2011; http://www.soystats.com/2011/). The high symbiotic nitrogen fixation ability of soybean makes it an integral component of sustainable agriculture. Improvements in yield, quality, and stress tolerance are major targets in soybean breeding program. While there is a confined narrow gene pool in domesticated cultivated soybeans (Jackson et al., 2006, 2011; Lam et al., 2010; Stupar, 2010), the undomesticated wild soybeans are promising sources of novel genes and alleles. The year 2010 has signified a major breakthrough in soybean genomic research. The first assembled reference genome of cultivated soybean was published in the beginning of 2010 (Schmutz et al., 2010), which was followed by reports on the re-sequencing of 17 wild and 14 cultivated soybean genomes (Lam et al., 2010); and a separate publication on deep re-sequencing of a wild soybean accession (Kim et al., 2010). The accumulation of these genomic data will initiate swift and intense functional and comparative genomic studies in the coming years. This review aims to summarize major findings in the soybean genomic research and discuss some possible future directions.

## 2. ESTABLISHMENT OF LINKAGE MAPS, PHYSICAL MAPS, AND A REFERENCE GENOME

It took more than a decade's effort from the early genomic survey of soybean (Marek et al., 2001) to the recent release of a reference genome (variety Williams 82) of this ancient tetraploid Phaseoloid legume (Schmutz et al., 2010). The genomic study of soybean began with tremendous efforts in building linkage maps and physical maps. Genetic maps derived from combined genetic populations (Cregan et al., 1999; Song et al., 2004; Choi et al., 2007) provide 1849 genetic markers (e.g., SSR and AFLP) (Song et al., 2004) and

* Corresponding author. Tel/fax: +852 3943 6336.
*E-mail address:* honming@cuhk.edu.hk (H.-M. Lam).


doi:10.1016/j.jgg.2012.02.002

1141 EST-based markers (Choi et al., 2007), which are quite evenly distributed in the ~2500 cM soybean genome. On the other hand, physical maps were also constructed from BAC or BIBAC libraries (Wu et al., 2004, 2008; Shoemaker et al., 2008). By the end of 2008, one *Hin*d III library containing 40,320 clones and two *Bst*Y I libraries containing 67,968 and 92,160 clones respectively, were constructed for Williams 82. A physical framework was formed by fingerprinting these libraries and associating contigs to molecular markers by overgo hybridization, RFLP hybridization and SSR amplification (Song et al., 2004; Choi et al., 2007) (e.g., by N-dimensional pools).

By integrating available genetic maps and physical maps, the Consensus Map 4.0 with a resolution of 0.6 cM mean interval was built (Hyten et al., 2010a). In addition, another genetic map was generated from a $F_5$ recombinant inbred population resulting from a cross between a wild accession (PI 468916) and Williams 82 (Hyten et al., 2010b). This map contains unevenly distributed markers enriched in regions not covered by the markers in the Consensus Map 4.0.

The whole-genome shotgun sequencing of Williams 82 was performed using the traditional Sanger's method and assembled using the Arachne algorithm (Schmutz et al., 2010). Marker information of the integrated maps (Song et al., 2004; Choi et al., 2007; Hyten et al., 2010a, 2010b) was employed to determine the order and orientation of the scaffolds. A total of 950 Mb was successfully assembled and anchored, representing 85% of the predicted 1115 Mb soybean genome.

## 3. UNIQUE FEATURES OF THE SOYBEAN GENOME

Sequencing information from the soybean reference genome revealed that there are 46,430 predicted genes (including 283 putative legume-specific gene families containing 448 high confidence soybean genes), of which 78% are located in the chromosome ends that account for most genetic recombination events (Schmutz et al., 2010).

The soybean genome usually comprises of 20 pairs of chromosomes which are small and morphologically homogenous (Singh and Hymowitz, 1988; Findley et al., 2010). The genomic data supports the notion that the soybean genome is a palaeopolyploid, with large-scale genomic duplications occurred twice at ~59 and 13 million years ago (mya), respectively (Shoemaker et al., 2006; Schmutz et al., 2010). The duplication and diploidization events resulted in a highly duplicated genome. A total of 61.4%, 5.63%, or 21.53% of the homologous genes were identified in blocks of two, three, or four chromosomes, respectively.

Based on the reference genome, whole-genome re-sequencing was performed on 17 wild and 14 cultivated soybean accessions (with an average depth of 5× for each accession) (Lam et al., 2010). These sequencing data allow the analysis of the soybean genome using population statistics. One unique feature of the soybean genome observed is the high linkage disequilibrium (LD) in both cultivated and wild soybean genomes. The average distances for LD to decay to half of its maximum value in cultivated and wild soybeans are ~150 kb and ~75 kb, respectively. Such values are much higher than that of maize, rice, and *Arabidopsis thaliana* (Lam et al., 2010). The stringent cleistogamy of soybean may attribute to this phenomenon. The high LD in soybeans allows marker-assisted breeding of soybean using only a small subset of molecular markers, but limits the resolution for association studies using genetic populations (Lam et al., 2010). In this regard, the identification of 205,614 tag SNPs will be useful for soybean breeders (Lam et al., 2010).

Another unique feature of the soybean genome lies on its high nonsynonymous to synonymous (Nonsyn/Syn) ratio of mutations. For both cultivated and wild soybean genomes, the Nonsyn/Syn ratios are higher than rice and *A. thaliana*. The high Nonsyn/Syn ratio may partly due to the high LD that allows "hitch-hiking" alleles to accumulate. It was also observed that large-effect single nucleotide polymorphisms (SNPs) were present in an unusually high portion (10%) of the annotated genes. The high Nonsyn/Syn ratios together with high level of large-effect SNPs may cause the accumulation of deleterious mutations in the soybean genome (Lam et al., 2010).

Like other flowering plants, transposable elements (TEs) occupy a significant portion of the soybean genome. Based on a comprehensive annotation using the reference genome (Schmutz et al., 2010), a total of 32,552 class I elements (including 32,370 LTR-retrotransposons and 182 LINEs) and 6029 DNA transposons (including 9 Tc1-Mariners, 90 PIF-Harbingers, 65 hATs, 2373 Mutators, 65 CACTAs, 12 PONGs and 82 Helitrons) were identified, occupying 42% and 16% of the genome respectively (Du et al., 2010). Using the short read data from the re-sequencing project of 31 soybean germplasms (Lam et al., 2010), unique new TEs have been identified which are likely representing recent insertions during soybean domestication and diversification (J. Ma, personal communication). Based on previous researches in other flowering plants, these soybean TEs may play important roles in (i) the evolution of genome *via* recombination, rearrangement and reshuffling (Ma and Bennetzen, 2006); (ii) the regulation of the epigenome *via* methylation (Zhang et al., 2008); and (iii) the alteration of the expression of adjacent genes *via* transcriptional activation (Kashkush et al., 2003). In addition, since there is a huge amount of new TEs identified in the soybean genome (Du et al., 2010), it is anticipated that new functions of TEs will be unveiled.

## 4. GENOMIC EVOLUTION AND DIVERSITY

At least two rounds of large-scale genomic duplications have shaped the structure of the recent soybean genome. Phylogenetic study of homologous gene pairs showed that the older duplication event was shared by two major sister legume lineages, the Hologalegina (including *Medicago* and *Lotus*) and the Phaseoloides, *Glycine* (Pfeil et al., 2005; Cannon et al., 2006; Gill et al., 2009), prior to their separation ~50 mya (Pfeil et al., 2005; Jackson et al., 2006). The latter duplication was associated with the divergence of homologous

pairs in the genome of *Glycine* species, followed by the evolution of the two *Glycine* subgenera ~5 mya (Doyle and Egan, 2010). However, it was believed that the ancestral diploid genome of soybean was extinct (Gill et al., 2009). Computational and cytogenetic study on centromeric satellite indicated the occurrence of subgenome in *Glycine max*, and its wild progenitor *G. soja*, suggesting that the recent polyploidy event was allopolyploidy in nature (Gill et al., 2009).

Polyploidization and diploidization events shuffled chromosomal segments and resulted into a mosaic genome structure. During evolution and under selection, the homologous chromosome segments as a result of genomic duplication would have undergone divergent structural and functional changes. For example, by using an orthologous genomic region from *Phaseolus vulgaris* as the outgroup for comparison and subsequent experimental verification using fluorescence *in situ* hybridization (Lin et al., 2010), an inversion was found in a 1-Mb duplicated segment between the soybean chromosomes Gm08 and Gm15. Such duplication was likely an event belonging to the ancient genomic duplication 13 mya. While the inversion event represented the major structural divergence, there were also other structural bias such as gene movement, deletions, and etc. Moreover, bias in the gene expression levels and differential synonymous substitution rates were also observed between the homologous segments (Lin et al., 2010).

Within the lineage of soybean, the cultivated soybeans (*G. max*) that were domesticated ~5000 years ago have a close genetic relationship with the undomesticated progenitor wild soybeans (*G. soja*), despite distinct differences in their morphological features. Bottleneck and human selection has reduced the genomic diversity of cultivated soybeans (Hyten et al., 2006; Lam et al., 2010). Since the two types of soybeans can inbreed, introgressions of genes from wild soybeans to cultivated soybeans have been reported, which may influence the genomic composition of cultivated soybeans (Lam et al., 2010).

Population genetic analysis using the whole-genome sequencing data showed that wild soybeans possess a much higher diversity, supporting the notion that wild soybean germplasms are important sources of novel genes/alleles for soybean improvement programs. A total of 6,318,109 SNPs were identified from 17 wild and 14 cultivated soybean genomes, including a huge array of wild-soybean-specific SNPs (Lam et al., 2010). The mapping of long LD blocks in both wild and cultivated soybean genomes, including cultivated-soybean-specific LD blocks, will facilitate the identification of genes related to domestication and human selection (Lam et al., 2010). Moreover, the cultivated-soybean-specific SNPs exhibit a higher than average Nonsyn/Syn ratio, which may attribute to the domestication-associated Hill-Robertson effect (Lu et al., 2006).

Direct comparison of deep re-sequencing data of wild soybean accessions to the reference Williams 82 genome have revealed a whole array of genomic variations between the wild and cultivated soybean genomes, including SNPs, insertions and deletions, and putative alleles/genes that may be unique to the wild soybeans (Kim et al., 2010; Lam et al., 2010). For example, over 10,000 nonsynonymous SNPs and 2398 indels were identified, potentially affecting >9000 protein structures or gene functions (Kim et al., 2010). However, since the re-sequencing approach using the cultivated soybean genome as the reference has limited the identification of novel genes in wild soybean genome, a reference genome of wild soybean built by *de novo* sequencing is needed for a more thorough analysis.

## 5. REFINING OF GENETIC MAPS

The accumulation of genomic sequencing data will expedite the refinement of genetic maps of soybean. Conventional genetic maps suffer from low resolution due to inadequate genome coverage by markers. About 100 traits including plant morphology related traits, seed quality and yield quantity were mapped on soybean genetic maps during the past 20 years (http://soybase.org/) and more than 40 soybean mapping populations with different genetic background have been published (http://soybase.org/). However, QTL studies based on these maps often resulted in a target region that contains too many candidate genes for functional analyses.

Based on the reference genome Williams 82, one recent achievement is the construction of the Universal Soy Linkage Panel (USLP 1.0) (Hyten et al., 2010a), which contains 1536 high quality SNPs selected from two GoldenGate genotyping assays (SoyOPA-2 and SoyOPA-3). This will facilitate high throughput QTL identification.

An alternative to this kind of genotyping by microarray hybridization is genotyping by direct re-sequencing. An initial attempt in soybean was performed by 4.4× re-sequencing using the Solexa system (Wu et al., 2010). This study identified 39,022 putative SNPs. However, recombinant breakpoints could not be determined in this experiment since only the two parental soybean lines were genotyped. Moreover, genotyping by sequencing also suffers from high error rate caused by relatively low coverage (Wu et al., 2010). One strategy to circumvent this error-prone approach is by adopting the "bin" concept. "Bin" is a unit between two breakpoints where unique segregation pattern is represented (Lee et al., 2004). In rice, sliding window approach (Huang et al., 2009) and Hidden Markov Model (Xie et al., 2010) were adopted for the genotyping of recombinant inbred populations to form a "bin map", which successfully identified narrow QTL regions containing genes contributing to plant height and seed width. Therefore, population scale low coverage re-sequencing followed by genotype called by "bin" instead of individual SNPs, should be a viable approach for refining soybean genetic maps.

## 6. TRANSCRIPTOME ANALYSIS

Microarray is a high throughput method in studying whole-genome transcriptome. The GeneChip Soybean Genome Array was constructed using the Affymetrix technology, allowing the analysis of 37,500 soybean transcripts. It has been successfully employed in studying the transcriptome of soybean. For

example, it was used to investigate the transcriptional responses of soybean toward the pathogen soybean cyst nematode (Puthoff et al., 2007; Mazarei et al., 2011), to identify differentially expressed gene of soybean during *Bradyrhizobium* infection (Libault et al., 2010), and to investigate the transcriptional changes of wild soybean upon $NaHCO_3$ treatment (Ge et al., 2010). Nevertheless, only 90.5% transcripts on the chip can match with the predicted transcripts in the soybean reference genome (Libault et al., 2010), suggesting a total coverage of less than 75% of the annotated genes. Whole-genome tiling arrays (Mockler and Ecker, 2005) are yet to be available for soybean.

Whole transcriptome shotgun sequencing (RNA-seq) based on the next generation sequencing platforms offers a way to detect whole-genome steady state transcriptome without being limited to pre-assigned transcripts (Wang et al., 2009; Martin and Wang, 2011). In addition to gene expression studies, RNA-seq studies also help verifying the annotation of the reference genome as well as providing more information about alternative splicing and trans-splicing RNA (Allen and Howell, 2010; Ozsolak and Milos, 2010; Martin and Wang, 2011).

RNA-seq data will often be mapped against the reference genome to bypass the need to assemble short read sequences. On the other hand, *de novo* RNA-seq without a reference genome or using a reference genome from close relatives is also possible (Lai et al., 2004; Collins et al., 2008; Wang et al., 2009; Martin and Wang, 2011). Therefore, a high quality soybean reference genome may be used as the reference for transcriptomic studies of other legumes.

Based on the William 82 Glyma1.01 genome assembly, RNA-seq was performed to study the transcriptome of fourteen diverse tissues to identify preferentially expressed transcripts in different tissues and growth stages (Severin et al., 2010). Furthermore, sequencing-based analysis also allows the study of non-protein coding RNAs. A study using the 454 platform identified 35 new families (in additional to the 20 known families) of miRNA from *Bradyrhizobium*-inoculated soybean (Subramanian et al., 2008).

## 7. EPIGENOMIC STUDIES

Epigenetics refers to the heritable transcription control without the change in genetic code stored in the genome. Mechanisms involved in epigenetics modification mainly involved DNA methylation and histone modifications.

In eukaryotes, DNA methylation occurs on C-5 position of cytosine. DNA methylation affects transcription rate through changing the affinity of DNA towards transcription factors or altering the packing density of the chromatin. Whole-genome bisulfite sequencing using the next generation sequencing platforms was employed to study the genome methylation pattern at the single-base-pair resolution (Cokus et al., 2008). A new method to detect DNA methylation at single-CpG resolution by combining the principle of bisulfite sequencing and DNA microarray was also recently introduced (Bibikova et al., 2011).

There was one early preliminary investigation of the DNA methylation patterns of the duplicated regions in the soybean genome (Zhu et al., 1994). Methylation polymorphisms were compared between cultivated and wild soybeans using 27 primer pairs to generate 984 CG/CNG methylation sites across 47 soybean accessions (Zhong et al., 2009). However, these studies are of limited scale and the whole-genome methylation pattern is still not available for soybean.

Histone modification included methylation, acetylation, ubiquitylation, phosphorylation, SUMOylation, ADP-ribosylation, and citrullination. Chromatin immunoprecipitation (ChIP) using antibodies against histone modifications is the conventional methods to study histone-DNA interaction (Spencer et al., 2003; Nelson et al., 2006; Haring et al., 2007). Protocols of ChIP have been modified to satisfy its mission in studying plant histone functions (Bowler et al., 2004; Gendrel et al., 2005; Kaufmann et al., 2010; Ricardi et al., 2010). ChIP may be accompanied with DNA microarray (ChIP-Chip) (Robyr and Grunstein, 2003; Huebert et al., 2006; Shivaswamy and Iyer, 2007) or sequencing (ChIP-seq) (Johnson et al., 2007) to census the DNA fragments interacting with the targeted histone.

Studies of histone modification in soybean are very limited. Variants of histone H3 and H4 and their post-translational modifications have been successfully identified in soybean by mass spectrometry. This study also unveiled some histone variants that have not been reported in other plants (Wu et al., 2009). There are also some evidences suggesting that transcriptional factors (such as GmPHD5) may play a role in mediating the crosstalk between different histone modifications (Wu et al., 2011).

## 8. ONLINE RESOURCES

For whole-genome sequencing projects, a huge amount of raw reads, assembled sequences, and annotation information will be generated. These data will usually be deposited in public databases or private servers of independent research groups to allow sharing of information.

In some servers, whole-genome data are presented in an interactive way using the Generic Genome Browser software (Stein et al., 2002). Gbrowse is an open source graphic based genome viewer developed by the Generic Model Organism Database project. It is widely used in displaying genome information. Information of Gbrowse is available in the following link: http://gmod.org/mediawiki/index.php?title=GBrowse&oldid=19363. This browser allows bird's view of an entire chromosome or detailed view of small chromosomal region by some simple operations in the user-friendly interface. It also supports the fast searching of genomic region by entering specific identifiers or retrieving sequence of certain region by simply clicking a few buttons. Genetic markers, topology of annotated genes, duplicated region, and other information can be viewed in the same page. Plug-in and add-on tools can also be added to the Gbrowse to facilitate instant analysis of the genome region of interest *in situ*. Gbrowse has been adopted by Phytozome, SoyBase and other soybean

Table 1
Summary of the available online resources for soybean genome analysis

| Host | Datasets | Major bioinformatics tools |
|---|---|---|
| NCBI<br>http://www.ncbi.nlm.nih.gov/<br>The National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH) | • Genome sequences<br>• Physical and genetic map<br>• Gene annotation<br>• Transcriptome<br>• Gene expression<br>• Genomic libraries and clones<br>• Genome re-sequencing datasets<br>• Sequence of individual chromosome<br>• SNPs<br>• ESTs | • BLAST<br>• Conserved Domain Database (CDD)<br>• Map viewer |
| Phytozome<br>http://www.phytozome.net/<br>Joint Genome Institute and Center for Integrative Genomics, University of California (Goodstein et al., 2012) | • Assembled sequences<br>• High confidence annotation | • Genome browser<br>• BLAST<br>• Cross species homology search |
| SoyBase<br>http://soybase.org/<br>USDA and Iowa State University (Grant et al., 2010; Du et al., 2010) | • Genetic map<br>• Genome sequence<br>• Gene annotation<br>• Physical map (WebFPC)<br>• QTL and loci collections<br>• Williams 82 transposable element (TE) database | • BLAST<br>• EST library<br>• SoyChip annotation<br>• SoyChip probe analysis<br>• SoyChip probe viewer<br>• Metabolic pathway database<br>• Haplotype identifier<br>• Tissue based expression profile<br>• Hierarchical clustering<br>• TE analysis |
| SoyKB (Soybean Knowledge Base)<br>http://soykb.org/<br>University of Missouri, Columbia, MO 65211-2060, USA (Joshi et al., 2012) | • GBrowse for annotated genes, miRNA, metabolites, SNPs | • BLAST<br>• Sequence evidence (EST, 5′RATE, full length cDNA)<br>• Experimental data (transcriptomics, microarray, proteomic)<br>• 3D protein structure<br>• Gene pathway viewer<br>• Metabolic pathway viewer<br>• Affymetrix probe ID Mapper<br>• Motifsampler by WebLOGO |
| SoyDB (for soybean transcription factors)<br>http://casp.rnet.missouri.edu/soydb/<br>National Science Foundation and University of Missouri (Wang et al., 2010) | • Amino acid sequences<br>• Predicted tertiary structures<br>• DNA binding sites<br>• Domain predictions<br>• Homologous proteins from the Protein Data Bank<br>• Protein family classifications<br>• Multiple sequence alignments<br>• Consensus DNA binding motifs and web logo of each family | • Text search<br>• PSI-BLAST<br>• Browse database<br>• Family prediction by HMM |
| Soy-TFKB (Soybean Transcription Factor Knowledge Base)<br>http://www.igece.org/Soybean_TF/<br>Institute for Green Energy and Clean Environment | • List of transcription factors (76 classified transcription factors families, 4452 predicted putative transcription factors)<br>• Protein sequences<br>• Transcript sequence | • TF families browser |
| PMRD: plant microRNA database<br>http://bioinformatics.cau.edu.cn/PMRD<br>China Agricultural University (Zhang et al., 2010) | • microRNA sequences<br>• Secondary structure<br>• Expression profiling<br>• Promoter sequence<br>• Target prediction | • microRNA search by name, location, target gene ID or stem-loop sequence<br>• Genome browser<br>• microRNA prediction |
| SGMD (The soybean genomics and microarray database)<br>http://bioinformatics.towson.edu/SGMD/ (Alkharouf and Matthews, 2004) | • EST library<br>• Microarray database | • LOBSA |
| BGI-Shenzhen<br>ftp://public.genomics.org.cn/BGI/soybean_re-sequencing/<br>(Lam et al., 2010) | • Genome re-sequencing dataset (17 wild and 14 cultivated soybean genome)<br>• SNP datasets<br>• PAV datasets | • N/A |

databases to display the information of soybean. These websites often accompany with sequence analysis tools and other soybean genome information.

The sharing of information through the internet allows different research groups having different expertise to utilize and analyze the soybean genome sequences. Information or knowledge generated from these researches is also made accessible on the internet. Here, we summarized some useful online resources for public access in Table 1.

## 9. CONCLUSIONS AND PERSPECTIVES

The release of a reference genome of cultivated soybean has generated important information and tools for soybean researches and breeding programs. However, due to the complex and highly repetitive nature of the soybean genome, there are still genomic regions that are not covered in this reference genome and gaps are still present between aligned scaffolds. Future breakthroughs in genomic sequencing technology and/or sequencing of more soybean accessions will help to refine this reference genome.

Wild soybeans exhibit a much higher genomic diversity. It is therefore an important source for novel genes/alleles. Population genetics analysis of wild and cultivated soybean genomes has revealed genomic regions that may be related to domestication and human selections. However, since the genomic data of wild soybeans are still based on re-sequencing, a reference genome of wild soybeans is needed to allow a more detailed analysis of these events.

The accumulation of genomic sequencing data helps to increase the resolution of genetic maps. The availability of more sequencing information of different genetic population and germplasms in the future will provide great tools to identify useful genes. However, due to the high LD of the soybean genome, genome-wide association studies may require deep sequencing of a large collection of germplasms. On the other hand, such exceptionally high LD in soybean and the identification of tag SNPs may expedite the marker-assisted breeding programs.

To assist researchers and breeders to make use of the genomic data, various online sources are now made available for data and information sharing. A concerted effort worldwide in the soybean genomic study has brought the research of this important crop to a more advanced level.

## ACKNOWLEDGEMENTS

The work is supported by the Hong Kong RGC General Research Fund (Nos. 468409 and 468610).

## REFERENCES

Alkharouf, N., Matthews, B.F., 2004. SGMD: the Soybean Genomics and Microarray Database. Nucleic Acids Res. 32, 1−3.

Allen, E., Howell, M.D., 2010. miRNAs in the biogenesis of trans-acting siRNAs in higher plants. Semin. Cell Dev. Biol. 21, 798−804.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., Fan, J.B., Shen, R., 2011. High density DNA methylation array with single CpG site resolution. Genomics 98, 288−295.

Bowler, C., Benvenuto, G., Laflamme, P., Molino, D., Probst, A.V., Tariq, M., Paszkowski, J., 2004. Chromatin techniques for plant cells. Plant J. 39, 776−789.

Cannon, S.B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., Wang, X.H., Mudge, J., Vasdewani, J., Scheix, T., Spannagl, M., Monaghan, E., Nicholson, C., Humphray, S.J., Schoof, H., Mayer, K.F.X., Rogers, J., Quetier, F., Oldroyd, G.E., Debelle, F., Cook, D.R., Retzel, E.F., Roe, B.A., Town, C.D., Tabata, S., van de Peer, Y., Young, N.D., 2006. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. Proc. Natl. Acad. Sci. USA 103, 14959−14964.

Choi, I.Y., Hyten, D.L., Matukumalli, L.K., Song, Q.J., Chaky, J.M., Quigley, C.V., Chase, K., Lark, K.G., Reiter, R.S., Yoon, M.S., Hwang, E.Y., Yi, S.I., Young, N.D., Shoemaker, R.C., van Tassell, C.P., Specht, J.E., Cregan, P.B., 2007. A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176, 685−696.

Cokus, S.J., Feng, S., Zhang, X.C., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., Jacobsen, S.E., 2008. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature 452, 215−219.

Collins, L.J., Biggs, P.J., Voelckel, C., Joly, S., 2008. An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome Inform. 21, 3−14.

Cregan, P.B., Jarvik, T., Bush, A.L., Shoemaker, R.C., Lark, K.G., Kahler, A.L., Kaya, N., VanToai, T.T., Lohnes, D.G., Chung, L., Specht, J.E., 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39, 1464−1490.

Doyle, J.J., Egan, A.N., 2010. Dating the origins of polyploidy events. New Phytol. 186, 73−85.

Du, J.C., Grant, D., Tian, Z.X., Nelson, R.T., Zhu, L.C., Shoemaker, R.C., Ma, J.X., 2010. SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 11, 113.

Findley, S.D., Cannon, S., Varala, K., Du, J.C., Ma, J.X., Hudson, M.E., Birchler, J.A., Stacey, G., 2010. A fluorescence *in situ* hybridization system for karyotyping soybean. Genetics 185, 727−744.

Ge, Y., Li, Y., Zhu, Y., Bai, X., Lv, D., Guo, D., Ji, W., Cai, H., 2010. Global transcriptome profiling of wild soybean (*Glycine soja*) roots under $NaHCO_3$ treatment. BMC Plant Biol. 10, 153.

Gendrel, A.V., Lippman, Z., Martienssen, R., Colot, V., 2005. Profiling histone modification patterns in plants using genomic tiling microarrays. Nat. Methods 2, 213−218.

Gill, N., Findley, S., Walling, J.G., Hans, C., Ma, J.X., Doyle, J., Stacey, G., Jackson, S.A., 2009. Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol. 151, 1167−1174.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D.S., 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40, D1178−D1186.

Grant, D., Nelson, R.T., Cannon, S.B., Shoemaker, R.C., 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38, D843−D846.

Haring, M., Offermann, S., Danker, T., Horst, I., Peterhansel, C., Stam, M., 2007. Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. Plant Methods 3, 11.

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

Huebert, D.J., Kamal, M., O'Donovan, A., Bernstein, B.E., 2006. Genome-wide analysis of histone modifications by ChIP-on-chip. Methods 40, 365−369.

Hyten, D.L., Song, Q.J., Zhu, Y.L., Choi, I.Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C., Cregan, P.B., 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. USA 103, 16666−16671.

Hyten, D.L., Choi, I.Y., Song, Q.J., Specht, J.E., Carter, T.E., Shoemaker, R.C., Hwang, E.Y., Matukumalli, L.K., Cregan, P.B., 2010a. A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci. 50, 960−968.

Hyten, D.L., Cannon, S.B., Song, Q.J., Weeks, N., Fickus, E.W., Shoemaker, R.C., Specht, J.E., Farmer, A.D., May, G.D., Cregan, P.B., 2010b. 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.

Jackson, S.A., Iwata, A., Lee, S.H., Schmutz, J., Shoemaker, R., 2011. Sequencing crop genomes: approaches and applications. New Phytol. 191, 915−925.

Jackson, S.A., Rokhsar, D., Stacey, G., Shoemaker, R.C., Schmutz, J., Grimwood, J., 2006. Toward a reference sequencing of the soybean genome: a multiagency effort. Crop Sci. 46, S55−S61.

Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-wide mapping of *in vivo* protein−DNA interactions. Science 316, 1497−1502.

Joshi, T., Patil, K., Fitzpatrick, M.R., Franklin, L.D., Yao, Q., Cook, J.R., Wang, Z., Libault, M., Brechenmacher, L., Valliyodan, B., Wu, X., Cheng, J., Stacey, G., Nguyen, H.T., Xu, D., 2012. Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. BMC Genomics 13 (Suppl. 1), S15.

Kashkush, K., Feldman, M., Levy, A.A., 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat. Genet. 33, 102−106.

Kaufmann, K., Muino, J.M., Osteras, M., Farinelli, L., Krajewski, P., Angenent, G.C., 2010. Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nat. Protoc. 5, 457−472.

Kim, M.Y., Lee, S., Van, K., Kim, T.H., Jeong, S.C., Choi, I.Y., Kim, D.S., Lee, Y.S., Park, D., Ma, J., Kim, W.Y., Kim, B.C., Park, S., Lee, K.A., Kim, D.H., Kim, K.H., Shin, J.H., Jang, Y.E., Kim, K.D., Liu, W.X., Chaisan, T., Kang, Y.J., Lee, Y.H., Moon, J.K., Schmutz, J., Jackson, S.A., Bhak, J., Lee, S.H., 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.

Lai, J.S., Dey, N., Kim, C.S., Bharti, A.K., Rudd, S., Mayer, K.F.X., Larkins, B.A., Becraft, P., Messing, J., 2004. Characterization of the maize endosperm transcriptome and its comparison to the rice genome. Genome Res. 14, 1932−1937.

Lam, H.-M., Xu, X., Liu, X., Chen, W.B., Yang, G.H., Wong, F.L., Li, M.W., He, W.M., Qin, N., Wang, B., Li, J., Jian, M., Wang, J.A., Shao, G., Wang, J., Sun, S.S.M., Zhang, G.Y., 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42, 1053−1059.

Lee, G.J., Boerma, H.R., Villagarcia, M.R., Zhou, X., Carter, T.E., Li, Z., Gibbs, M.O., 2004. A major QTL conditioning salt tolerance in S-100 soybean and descendent cultivars. Theor. Appl. Genet. 109, 1610−1619.

Libault, M., Farmer, A., Brechenmacher, L., Drnevich, J., Langley, R.J., Bilgin, D.D., Radwan, O., Neece, D.J., Clough, S.J., May, G.D., Stacey, G., 2010. Complete transcriptome of the soybean root hair cell, a single-cell model, and its alteration in response to *Bradyrhizobium japonicum* infection. Plant Physiol. 152, 541−552.

Lin, J.Y., Stupar, R.M., Hans, C., Hyten, D.L., Jackson, S.A., 2010. Structural and functional divergence of a 1-Mb duplicated region in the soybean (*Glycine max*) genome and comparison to an orthologous region from *Phaseolus vulgaris*. Plant Cell 22, 2545−2561.

Lu, J., Tang, T., Tang, H., Huang, J.Z., Shi, S.H., Wu, C.I., 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet. 22, 126−131.

Ma, J.X., Bennetzen, J.L., 2006. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc. Natl. Acad. Sci. USA 103, 383−388.

Marek, L.F., Mudge, J., Darnielle, L., Grant, D., Hanson, N., Paz, M., Huihuang, Y., Denny, R., Larson, K., Foster-Hartnett, D., Cooper, A., Danesh, D., Larsen, D., Schmidt, T., Staggs, R., Crow, J.A., Retzel, E., Young, N.D., Shoemaker, R.C., 2001. Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome 44, 572−581.

Martin, J.A., Wang, Z., 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12, 671−682.

Mazarei, M., Liu, W.X., Al-Ahmad, H., Arelli, P., Pantalone, V., Stewart, C., 2011. Gene expression profiling of resistant and susceptible soybean lines infected with soybean cyst nematode. Theor. Appl. Genet. 123, 1193−1206.

Mockler, T.C., Ecker, J.R., 2005. Applications of DNA tiling arrays for whole-genome analysis. Genomics 85, 1−15.

Nelson, J.D., Denisenko, O., Bomsztyk, K., 2006. Protocol for the fast chromatin immunoprecipitation (ChIP) method. Nat. Protoc. 1, 179−185.

Ozsolak, F., Milos, P.M., 2010. RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. 12, 87−98.

Pfeil, B.E., Schlueter, J.A., Shoemaker, R.C., Doyle, J.J., 2005. Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Syst. Biol. 54, 441−454.

Puthoff, D.P., Ehrenfried, M.L., Vinyard, B.T., Tucker, M.L., 2007. GeneChip profiling of transcriptional responses to soybean cyst nematode, *Heterodera glycines*, colonization of soybean roots. J. Exp. Bot. 58, 3407−3418.

Ricardi, M.M., Gonzalez, R.M., Iusem, N.D., 2010. Protocol: fine-tuning of a chromatin immunoprecipitation (ChIP) protocol in tomato. Plant Methods 6, 11.

Robyr, D., Grunstein, M., 2003. Genomewide histone acetylation microarrays. Methods 31, 83−89.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J.X., Mitros, T., Nelson, W., Hyten, D.L., Song, Q.J., Thelen, J.J., Cheng, J.L., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S.Q., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J.C., Tian, Z.X., Zhu, L.C., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., Jackson, S.A., 2010. Genome sequence of the palaeopolyploid soybean. Nature 463, 178−183.

Severin, A.J., Woody, J.L., Bolon, Y.T., Joseph, B., Diers, B.W., Farmer, A.D., Muehlbauer, G.J., Nelson, R.T., Grant, D., Specht, J.E., Graham, M.A., Cannon, S.B., May, G.D., Vance, C.P., Shoemaker, R.C., 2010. RNA-seq atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biol. 10, 160.

Shivaswamy, S., Iyer, V.R., 2007. Genome-wide analysis of chromatin status using tiling microarrays. Methods 41, 304−311.

Shoemaker, R.C., Schlueter, J., Doyle, J.J., 2006. Paleopolyploidy and gene duplication in soybean and other legumes. Curr. Opin. Plant Biol. 9, 104−109.

Shoemaker, R.C., Grant, D., Olson, T., Warren, W.C., Wing, R., Yu, Y., Kim, H., Cregan, P., Joseph, B., Futrell-Griggs, M., Nelson, W., Davito, J., Walker, J., Wallis, J., Kremitski, C., Scheer, D., Clifton, S.W., Graves, T., Nguyen, H., Wu, X.L., Luo, M.C., Dvorak, J., Nelson, R., Cannon, S., Tomkins, J., Schmutz, J., Stacey, G., Jackson, S., 2008. Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. Genome 51, 294−302.

Singh, R.J., Hymowitz, T., 1988. The genome relationship between *Glycine max* (L) Merr. and *Glycine soja* Sieb. and Zucc. as revealed by pachytene chromosome analysis. Theor. Appl. Genet. 76, 705−711.

Song, Q.J., Marek, L.F., Shoemaker, R.C., Lark, K.G., Concibido, V.C., Delannay, X., Specht, J.E., Cregan, P.B., 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109, 122−128.

Spencer, V.A., Sun, J.M., Li, L., Davie, J.R., 2003. Chromatin immunoprecipitation: a tool for studying histone acetylation and transcription factor binding. Methods 31, 67−75.

Stein, L.D., Mungall, C., Shu, S.Q., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., Lewis, S., 2002. The generic genome browser: a building block for a model organism system database. Genome Res. 12, 1599−1610.

Stupar, R.M., 2010. Into the wild: the soybean genome meets its undomesticated relative. Proc. Natl. Acad. Sci. USA 107, 21947−21948.

Subramanian, S., Fu, Y., Sunkar, R., Barbazuk, W.B., Zhu, J.-K., Yu, O., 2008. Novel and nodulation-regulated microRNAs in soybean roots. BMC Genomics 9, 160.

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

Wang, Z., Libault, M., Joshi, T., Valliyodan, B., Nguyen, H.T., Xu, D., Stacey, G., Cheng, J.L., 2010. SoyDB: a knowledge database of soybean transcription factors. BMC Plant Biol. 10, 14.

Wu, C.C., Sun, S.K., Nimmakayala, P., Santos, F.A., Meksem, K., Springman, R., Ding, K., Lightfoot, D.A., Zhang, H.B., 2004. A BAC and BIBAC-based physical map of the soybean genome. Genome Res. 14, 319−326.

Wu, T., Yuan, T.Z., Tsai, S.N., Wang, C.M., Sun, S.M., Lam, H.M., Ngai, S.M., 2009. Mass spectrometry analysis of the variants of histone H3 and H4 of soybean and their post-translational modifications. BMC Plant Biol. 9, 98.

Wu, T., Pi, E.-X., Tsai, S.-N., Lam, H.-M., Sun, S.S.-M., Kwan, Y.W., Ngai, S.-M., 2011. GmPHD5 acts as an important regulator for crosstalk between histone H3K4 di-methylation and H3K14 acetylation in response to salinity stress in soybean. BMC Plant Biol. 11, 178.

Wu, X.L., Zhong, G.H., Findley, S.D., Cregan, P., Stacey, G., Nguyen, H.T., 2008. Genetic marker anchoring by six-dimensional pools for development of a soybean physical map. BMC Genomics 9, 28.

Wu, X.L., Ren, C.W., Joshi, T., Vuong, T., Xu, D., Nguyen, H.T., 2010. SNP discovery by high-throughput sequencing in soybean. BMC Genomics 11, 469.

Xie, W.B., Feng, Q., Yu, H.H., Huang, X.H., Zhao, Q.A., Xing, Y.Z., Yu, S.B., Han, B., Zhang, Q.F., 2010. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc. Natl. Acad. Sci. USA 107, 10578−10583.

Zhang, W.L., Lee, H.R., Koo, D.H., Jiang, J.M., 2008. Epigenetic modification of centromeric chromatin: hypomethylation of DNA sequences in the CENH3-associated chromatin in *Arabidopsis thaliana* and maize. Plant Cell 20, 25−34.

Zhang, Z.H., Yu, J.Y., Li, D.F., Zhang, Z.Y., Liu, F.X., Zhou, X., Wang, T., Ling, Y., Su, Z., 2010. PMRD: plant microRNA database. Nucleic Acids Res. 38, D806−D813.

Zhong, X.F., Wang, Y.M., Liu, X.D., Gong, L., Ma, Y., Qi, B., Dong, Y.S., Liu, B., 2009. DNA methylation polymorphism in annual wild soybean (*Glycine soja* Sieb. et Zucc.) and cultivated soybean (*G. max* L. Merr.). Can. J. Plant Sci. 89, 851−863.

Zhu, T., Schupp, J.M., Oliphant, A., Keim, P., 1994. Hypomethylated sequences: characterization of the duplicate soybean genome. Mol. Gen. Genet. 244, 638−645.