


2012

Advances in Genome Sequencing and Genotyping Technology for Soybean Diversity Analysis

David L. Hyten

Soybean Genomics and Improvement Laboratory, Beltsville, MD, david.hyten@unl.edu

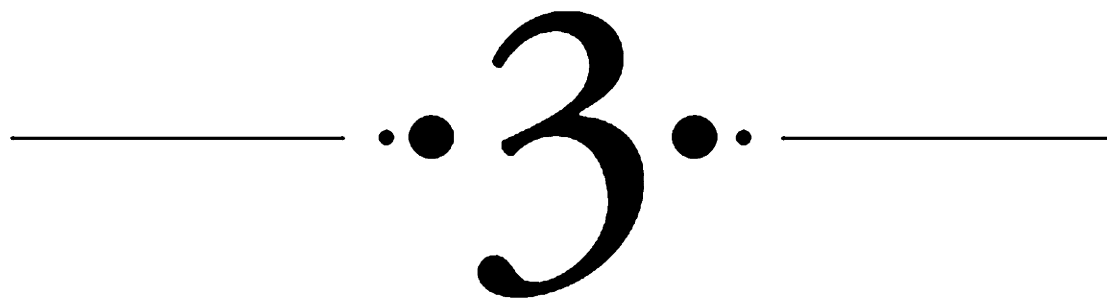
Follow this and additional works at: <https://digitalcommons.unl.edu/agronomyfacpub>

 Part of the [Agricultural Science Commons](#), [Agriculture Commons](#), [Agronomy and Crop Sciences Commons](#), [Botany Commons](#), [Horticulture Commons](#), [Other Plant Sciences Commons](#), and the [Plant Biology Commons](#)

Hyten, David L., "Advances in Genome Sequencing and Genotyping Technology for Soybean Diversity Analysis" (2012). *Agronomy & Horticulture -- Faculty Publications*. 821.

<https://digitalcommons.unl.edu/agronomyfacpub/821>

This Article is brought to you for free and open access by the Agronomy and Horticulture Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Agronomy & Horticulture -- Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.



Advances in Genome Sequencing and Genotyping Technology for Soybean Diversity Analysis

David L. Hyten

*Soybean Genomics and Improvement Laboratory,
U.S. Department of Agriculture, Agricultural Research Service, Beltsville, Maryland 20705, USA*

Overview

The completion of a soybean draft whole genome sequence along with advances in sequencing and genotyping technologies are creating a revolution in soybean genomics. The draft whole genome sequence of soybean is allowing researchers to fully take advantage of this new technology and is rapidly changing how soybean germplasm is mined. Genetic markers such as SNPs can be quickly identified by using next generation sequencing and assayed on a large number of materials using advanced technologies that can genotype tens of thousands of markers on thousands of individuals very rapidly. This ability to quickly identify and genotype genetic diversity is allowing researchers to identify beneficial genetic diversity and rapidly transfer it into elite cultivars for increased soybean production, biotic and abiotic protection, and improved seed quality.

Discussion

The Soybean Genome Sequencing Project

The soybean genome sequencing project was initiated by the Department of Energy-Joint Genome Institute (DOE-JGI) Community Sequencing Program and began large scale sequencing of soybean during the summer of 2006 (www.phytozome.net/soybean, accessed 08/19/2011). This sequencing project had the goal to create an 8x genome sequence of soybean using Sanger shotgun reads. The sequencing phase of the project was completed in January of 2008 and the first draft assembly known as Glyma1 was published in January of 2010 (Schmutz et al., 2010).

The soybean cultivar that was chosen for this first 8x draft assembly was Williams 82 (Bernard & Cremeens, 1988). Williams 82 was created through a series of backcrosses to the popular northern line Williams to introgress the phytophthora rot resistant gene *Rps1-k* from the cultivar 'Kingwa' (Bernard & Cremeens, 1988). This *Rps1-k* locus is found on chromosome 3 in the soybean genome. Due to the multiple rounds of maintaining heterozygosity during the backcrossing have led to significant genetic heterogeneity and structural variation heterogeneity being present in this chromosome in the Williams 82 cultivar. This could affect any genomic analyses which occur on chromosome 3 (Haun et al., 2010).

The draft Williams 82 sequence contains a total of 950 megabases (Mb) of assembled sequence (Schmutz et al., 2010). This sequence has been assembled into 397 sequence scaffolds and anchored to the soybean genetic consensus map and a second high resolution genetic map to create 20 pseudomolecules (Hyten et al., 2010a; Schmutz et al., 2010). The 950 Mb of assembled sequence is approximately 85% of the estimated 1.1 Mb genome of soybean (Schmutz et al., 2010).

A total of 46,430 protein-coding genes were predicted with high-confidence while another 20,000 loci were predicted with a lower confidence level (Schmutz et al., 2010). Two separate studies have provided experimental evidence that 49,151 (Severin et al., 2010) and 55,616 (Libault et al., 2010) of the 66,430 predicted genes are transcriptionally active.

While the availability of this draft genome sequence is proving to be an invaluable tool to soybean geneticist, it is only a first step in whole genome analysis. This draft genome sequence only represents the genes and alleles that are present within a single cultivar which was developed and released over 30 years ago (Bernard & Cremeens, 1988). The sequence does not provide answers to the genetic diversity present within soybean or even the full complement of genes that are present within soybean (Haun et al., 2010). To obtain those answers, projects will be needed to determine the extent of copy number variation, genomic rearrangements, allele diversity, and epigenetic variation that are present between multiple soybean accessions.

Genetic Diversity of the Soybean Genome

Genetic diversity is one of the main factors that have helped to facilitate improvements to soybean throughout its history. The genetic diversity of the soybean genome has been affected by many factors including genetic bottlenecks and selection (Hyten et al., 2006). The history of North American soybean cultivars can be traced back to 35 ancestors which make up over 96% of the parentage of current elite public cultivars. Seventeen of those 35 ancestors account for over 86% of the parentage of current elite cultivars (Gizlice et al., 1994). These ancestors were originally derived from Asian landraces which predate modern breeding techniques. Asian landraces have been grown by Asian farmers for the past 3,000 to 5,000 years when that soybean was domesticated from its wild ancestor, *Glycine soja* (Seib. et Zucc.) (Carter et al., 2004). Over 170,000 of these ancestors to modern soybean have been collected and maintained in germplasm

collections throughout the world and form the basis of the diversity available for soybean improvement (Carter et al., 2004). Despite this large quantity of germplasm available only 1,000 have been used in applied breeding programs (Carter et al., 2004) and less than 50 have had their genome extensively sequenced.

***Glycine soja* Germplasm**

Soybean is believed to have been domesticated about 3,000 to 5,000 years ago from the wild species *Glycine soja* (Carter et al., 2004). *G. soja* grows wild throughout Asia and produces a small black seed. There are approximately 10,700 accessions which are maintained in germplasm collections throughout the world (Carter et al., 2004). The largest collection of *G. soja* is maintained at the Institute of Crop Germplasm Resources in China which contains approximately 6,200 accessions. The USDA Soybean Germplasm collection in Urbana, IL, USA is the second largest collection of wild soybean germplasm containing approximately 1,176 accessions (Carter et al., 2004). These accessions have been collected from China, Japan, South Korea, Taiwan, the Philippines, and Russia. The mating system for *G. soja* is mostly inbreeding but outcrossing rates as high as 13% have been reported (Fujita et al., 1997). Like soybean, *G. soja* has 20 chromosomes and most accessions create fully fertile hybrids when crossed with cultivated soybeans. Since *G. soja* is a wild species the genetic diversity has been found to be twice as high as domesticated soybean but still lower than other comparable wild species such as Arabidopsis, wild barley, and teosinte (Hyten et al., 2006).

***Glycine max* Germplasm**

Glycine max germplasm available for genomic analysis studies consists of Asian landraces, modern cultivars which were released after 1945, isolines, mutants, and germplasm releases which have been registered in *Crop Science* (Carter et al., 2004). Over 156,000 of these *G. max* germplasm accessions are maintained in germplasm collections throughout the world. The largest collection for *G. max* is maintained at the Institute of Crop Germplasm Resources in China which has over 23,000 accessions while the second largest collection is maintained at the USDA Soybean Germplasm collection in Urbana, IL, USA which contains over 18,000 accessions (Carter et al., 2004). While these landraces provide a large amount of genetic diversity available for crop improvement only a small fraction have been used in modern cultivars for crop improvements. Newer methods of better characterizing the genetic diversity are needed to help guide breeders in efficiently mining this germplasm for useful genetic variation.

Next-Generation Sequencing and Application to SNP Detection in Soybean

Although the Sanger dideoxy method had been the standard for sequencing DNA and was used to produce a draft sequence of a soybean cultivar, recent innovations have generated a quantum leap in sequencing technology. The so called “Next

Generation Sequencing” (NGS) methods have led to an explosion of DNA sequence information that is available for whole genome characterization. Several currently available next-generation sequencing methods include Roche 454, Illumina (previously known as Solexa), SOLiD, ion torrent, and KBioscience; many more NGS technologies are being developed. These methods have become increasingly popular because of their ability to generate orders of magnitude more sequence data than Sanger at a much lower cost per base. However, a current disadvantage of the next-generation methods is that they generally produce much shorter read lengths when compared to the Sanger dideoxy method.

The Sanger dideoxy method can generally produce reads that are in excess of 1 kb. The next-generation methods are able to produce read lengths which range from 36 bp up to 500 bp depending upon the technology. For example Roche 454 sequencing produces a total amount of sequence of approximately 400 to 700 million bp for one run which is much less than some of the other NGS methods. The lack of sequence depth from Roche 454 sequencing is compensated for by the ability to obtain over 85% of the reads at greater than 500 bp (www.454.com, verified August 1, 2011).

While the 454 method is useful for its longer reads several of the other next generation sequencing technologies can obtain much more total sequence per run but at a cost of having shorter read lengths. The NGS technology that has been very commonly used in soybean is the Illumina technology. This technology started with the Illumina Genome Analyzer (GA) 1G that was capable of producing a total of 1 billion bp of total sequence with an average read length of 36 bp. While at the time this seemed like an impressive amount of sequence data the throughput of the Illumina technology has very quickly advanced and currently dwarfs that of the GA 1G. The latest machine that uses the Illumina technology is the HiSeq 2000. The HiSeq 2000 has a capability of producing 100 bp paired end reads with a total throughput of 540–600 billion bp (www.illumina.com, verified August 1, 2011).

Reduced Representation Libraries

Next-generation sequencing has greatly enhanced efforts to find large numbers of single nucleotide polymorphisms (SNPs) in soybean. SNPs are the most common form of genetic variation between soybean lines. Between any two distantly related soybean lines one can estimate that there is an average of 1.6 to 3 million SNPs that are polymorphic. Before next-generation sequencing, discovering SNPs was very labor intensive and only a small fraction contained within soybean were found (Choi et al., 2007).

Once the soybean genome was obtained several groups were able to use a method developed by Van Tassel et al. (2008) which utilized NGS to efficiently discover SNPs. The method developed by Van Tassel et al. (2008) utilized reducing the fraction of the genome that was sequenced by digesting the genome with a restriction enzyme and size selection on an agarose gel so that a small proportion of the genome would be sequenced with NGS and aligned to a sequenced genome.

This method in soybean, first reported by Hyten et al. (2010a), utilized five restriction enzymes to obtain a larger proportion of the genome in the DNA size range that they were isolating. With a partial run of the Illumina GA 1G, Hyten et al. (2010a) obtained approximately 20,000 SNPs which had a validation rate of 92%. Subsequently Deschamps et al. (2010) used a methylation-sensitive restriction endonuclease which was then followed by a second digestion with another restriction endonuclease to create a reduced representation library which was enriched for hypomethylated genomic DNA. Using this method Deschamps et al. (2010) discovered 1,682 SNPs which had a validation rate of 97%. These methods using NGS transformed and greatly accelerated the pace at which SNPs could be discovered in soybean germplasm.

Whole Genome Sequencing

Although using reduced representation libraries for SNP discovery was useful when the next-generation sequencers were only capable of obtaining 1 Gb of sequence, the technology has quickly advanced. Once the capabilities of the next-generation sequencers increased to 20–40 Gb per run it became possible to perform whole genome resequencing. This was demonstrated by the resequencing of 31 wild and cultivated soybean genomes using the Illumina Genome Analyzer II platform (Lam et al., 2010). Lam et al. (2010) sequenced these 31 lines to an average depth of 5x which led to a total of 180Gb of sequence, each line having an average coverage >90%. From this resequencing project, Lam et al. (2010) identified 6.3 million SNPs and 186,177 variations.

Another soybean genome sequence using next generation sequencing technology was recently reported which resequenced a single *Glycine soja* line (Kim et al., 2010). This *G. soja* line was sequenced to a 43x depth and covered approximately 97.65% of the soybean reference genome sequence. This genome sequence discovered approximately 2.5 million SNPs and 406 kb worth of insertion/deletion sequence.

The main disadvantage to both of these projects is that they had to rely on the reference Williams 82 genome sequence. Neither sequencing project was able to create a de novo assembly of the material sequence to adequately explore structural variation between Williams 82 and other *Glycine max* lines or the wild *G. soja* lines. As NGS technology improves for longer read lengths then quick sequencing and de novo assembly of multiple soybean lines may become possible. This will greatly enhance our knowledge of copy number variation and genomic rearrangements that exist between soybean lines.

High-Throughput Genotyping

The ability to discover large numbers of SNP markers has allowed soybean researchers to take advantage of the most recent technological advances in SNP detection. These new high-throughput genotyping technologies have greatly accelerated the pace at which germplasm and genetic populations can be genotyped. The first real breakthrough was with the Illumina GoldenGate assay which is currently being extensively

used to genotype genetic populations. The second breakthrough was with the design of an Illumina Infinium assay for soybean which is capable of genotyping 52,000 SNPs on a large number of soybean lines at once.

GoldenGate Assay

High-throughput SNP genotyping has been made possible recently due to advancement in technology. One new technology that has been extensively used in soybean is the GoldenGate assay (Illumina, Inc.) which is capable of allowing a single technician to genotype 192 DNA samples with 3,072 SNPs. The GoldenGate assay uses allele specific hybridization, then ligation/extension, followed by a universal primer amplification, which labels one allele of the SNP with a Cy3 fluorescent dye and the other allele with a Cy5 fluorescent dye. These amplification products are then hybridized on either a syntrix array matrix or a universal bead chip and scanned on an Illumina BeadStation or an iScan (Fan et al., 2003).

The first GoldenGate assay designed to soybean was a 384 SNP-plex assay (Hyten et al., 2008). This initial test in soybean demonstrated that the GoldenGate assay could be successfully used to genotype the complex genome of soybean. Out of the 384 SNPs it was shown that 89% of the SNPs successfully genotyped three recombinant inbred populations along with 96 diverse Asian landraces (Hyten et al., 2008). Subsequently two 1,536 SNP GoldenGate assays were designed and tested (Hyten et al., 2010b). All three GoldenGate assays were tested with 96 diverse Asian landraces and 96 diverse elite cultivars. By using map positioning of the SNPs along with their allele frequencies in the diverse germplasm a total of 1,536 SNPs were selected to use as a universal array which could be used to genotype any mapping population or to genotype diverse germplasm (Hyten et al., 2010b).

One surprise that was found with the original 384 GoldenGate assay was that the assay was copy number sensitive (Hyten et al., 2008). The observed sensitivity of the GoldenGate assay to copy number lead to testing of the first 1,536 GoldenGate assay with three bulked DNA samples which consisted of F₂ plants which were found to be susceptible to soybean rust. These F₂ plants came from a population which segregated for the *Rpp3* resistance locus. The use of bulked segregant analysis coupled with the GoldenGate assay allowed for the rapid identification of a 14 cM *Rpp3* candidate region. This region was confirmed to be the *Rpp3* locus through mapping with SSR loci in the region on the entire F₂ population. This application of the GoldenGate assay to genotype bulks demonstrated it to be a powerful tool for mapping genes responsible for single gene traits quickly (Hyten et al., 2010b). Kendrick et al. (2011) have used the 1536 SNP Universal Soybean Linkage Panel (USLP 1.0) to putatively map multiple populations which segregated for soybean rust resistance using the bulked segregant analysis.

Infinium Assay

While the GoldenGate assay can be considered high-throughput, the Infinium assay is capable of much greater multiplexing levels. Recently a 52,000 SNP soybean

Infinium plex has been designed and is currently being used for a large soybean HapMap project (Hyten et al., 2010c). The HapMap project has a goal of genotyping the 18,000 soybean accessions that are maintained at the USDA Soybean Germplasm collection in Urbana, IL, USA. This project should provide an in depth look into the population structure and haplotype diversity that is contained within this larger germplasm collection.

Conclusion

There has been great progress made in using soybean genomics to understand the genetic diversity which is contained within soybean. A combination of funding and technology advancement has led to several important advancements. Unquestionably, the largest of these advancements was the production of an 8x draft soybean sequence. This draft sequence has allowed the extensive use of next generation sequencing to help explore the genetic diversity contained within soybean. High-throughput SNP discovery has now become commonplace and the ability to completely resequence many soybean genomes has become a reality. The use of high-throughput genotyping methods has allowed researchers to explore the diversity structure in larger germplasm collections and to start linking genetic diversity with phenotypic diversity on a much faster pace than in the past.

While these advancements are enabling many avenues of research there is still much to be accomplished. As NGS evolves into third generation sequencing the capability to sequence and de novo assemble many soybean genomes will become a reality. In addition, as all the genetic variation is mapped out within soybean germplasm, new and novel methods will be developed that will better link this genetic variation to phenotype variation. This will help in our abilities to improve soybean.

References

- Bernard, R.L.; Cremeens, C.R. Registration of 'Williams 82' soybean. *Crop Sci.* 1988, 28, 1027–1028.
- Carter, T.E.; Nelson, R.; Sneller, C.H.; Cui, Z. Genetic Diversity in Soybean. In *Soybeans: Improvement, production, and uses*; Boerma, H. R.; Specht, J. E. Eds.; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI 2004; 303–416.
- Deschamps, S.; Rota, M.L.; Ratashak, J.P.; Biddle, P.; Thureen, D.; Farmer, A.; Luck, S.; Beatty, M.; Nagasawa, N.; Michael, L.; et al. Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina genome analyzer. *The Plant Genome* 2010, 3(1), 53–68.
- Fan, J.B.; Oliphant, A.; Shen, R.; Kermani, B.G.; Garcia, F.; Gunderson, K.L.; Hansen, M.; Steemers, F.; Butler, S.L.; Deloukas, P.; et al. Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol.* 2003, 68, 69–78.
- Fujita, R.; Ohara, M.; Okazaki, K.; Shimamoto, Y. The extent of natural cross-pollination in wild soybean (*Glycine soja*). *J. Hered.* 1997, 88, 124–128.

- Gizlice, Z.; Carter, T.E.; Burton, J.W. Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Sci.* 1994, 34, 1143–1151.
- Haun, W.J.; Hyten, D.L.; Xu, W.W.; Gerhardt, D.J.; Albert, T.J.; Richmond, T.; Jeddeloh, J.A.; Jia, G.; Springer, N.M.; Vance, C.P.; Stupar, R.M. The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol.* 2011, 155, 645–655.
- Hyten, D.L.; Cannon, S.B.; Song, Q.; Weeks, N.; Fickus, E.W.; Shoemaker, R.C.; Specht, J.E.; Farmer, A.D.; May, G.D.; Cregan, P.B. 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* 2010a, 11, 38.
- Hyten, D.L.; Choi, I.Y.; Song, Q.; Specht, J.E.; Carter, T.E.; Shoemaker, R.C.; Hwang, E.Y.; Matukumalli, L.K.; Cregan, P.B. A high density integrated genetic linkage map of soybean and the development of a 1,536 Universal Soy Linkage Panel for QTL mapping. *Crop Sci.* 2010b, 960–968.
- Hyten, D.L.; Song, Q.; Jia, G.; Nelson, R.; Pantalone, V.; Specht, J.; Cregan, P. The Soybean HapMap Phase I. 13th Biennial Molecular & Cellular Biology of the Soybean Conference 2010, Durham, North Carolina, 2010c.
- Hyten, D.L.; Song, Q.; Choi, I.Y.; Yoon, M.S.; Specht, J.E.; Matukumalli, L.K.; Nelson, R.L.; Shoemaker, R.C.; Young, N.D.; Cregan, P.B. High-throughput genotyping with the Golden-Gate assay in the complex genome of soybean. *Theor. Appl. Genet.* 2008, 116, 945–952.
- Hyten, D.L.; Song, Q.; Zhu, Y.; 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* 2006; 103, 16666–71.
- Kendrick, M.D.; Harris, D.K.; Ha, B.K.; Hyten, D.L.; Cregan, P.B.; Frederick, R.D.; Boerma, H.R.; Pedley, K.F. Identification of a second asian soybean rust resistance gene in Hyuuga soybean. *Phytopathology* 2011, 101, 535–543. DOI: 10.1094/PHYTO-09-10-0257.
- 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.; et al. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proceedings of the National Academy of Sciences of the United States of America* 2010; 107, 22032–37. DOI: 10.1073/pnas.1009526107.
- Lam, H.M.; Xu, X.; Liu, X.; Chen, W.; Yang, G.; Wong, F.L.; Li, M.W.; He, W.; Qin, N.; Wang, B.; et al. (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature genetics* 2010, 42, 1053–9. DOI: 10.1038/ng.715.
- Libault, M.; Farmer, A.; Joshi, T.; Takahashi, K.; Langley, R.J.; Franklin, L.D.; He, J.; Xu, D.; May, G.; Stacey, G. An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *Plant J.* 2010; 63, 86–99.
- Schmutz, J.; Cannon, S.B.; Schlueter, J.; Ma, J.; Mitros, T.; Nelson, W.; Hyten, D.L.; Song, Q.; Thelen, J.J.; Cheng, J.; et al. Genome sequence of the palaeopolyploid soybean. *Nature* 2010, 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.; et al. RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biol.* 2010, 10, 160.
- Van Tassell, C.P.; Smith, T.P.; Matukumalli, L.K.; Taylor, J.F.; Schnabel, R.D.; Lawley, C.T.; Haudenschild, C.; Moore, S.S.; Warren, W.C.; Sonstegard, T.S. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* 2008, 5, 247–252.