The 1k Soy Community SNP Panel at AgriPlex Genomics

Introduction:

The legume crop soybean (*Glycine max* [L.] Merr.) is the leading oilseed crop produced and consumed in the world today and accounts for 29% of the world's agricultural output.

The domestication of G. max from its wild progenitor (*Glycine soja Sieb*. and *Zucc.*) occurred in China approximately 5,000 years ago (Carter et al., 2004) and expanded to other parts of Asia around 2000 years ago (Kihara, 1969); The crop was likely introduced into the Americas during the 18th century.

Local adaptation throughout this global distribution resulted in a wide range of unique landraces. More than 170,000 soybean accessions are maintained by more than 160 institutions in nearly 70 countries (International Plant Genetic Resources Institute, 2001). However, only 80 accessions account for 99% of the collective parentage of North American soybean cultivars; off these, seventeen elite parents account for 86% of the collective (Li and Nelson, 2001; Carter et al., 2004). Researchers have presumed that these genetic bottlenecks have reduced the genetic diversity of modern soybean.

The size of the soy *Glycine max* genome is 1.1 Gb arranged in 20 chromosomes. The genome contains more than 46,000 protein-coding loci (Schmutz et al, 2010). Many of these gene loci have yet to be genetically and physically mapped. Exploring the soy genome, mapping all gene loci, performing functional genetic research, and translating the information into molecular breeding can be facilitated by the use of molecular markers.

Molecular markers have evolved over the past 80 years through sampling and comparing genomes. Over time, technology for interrogating genetic variation has progressed, and many DNA molecular marker systems were developed. Consequently, so did the resolution of the genomic picture the markers can depict. Single Nucleotide Polymorphisms (SNPs) have emerged as the ultimate molecular marker. SNPs are single nucleotide changes that are heritable codominant and distributed with relatively high frequency throughout eukaryotic genomes. SNPs can be the causative mutation that directly affect a phenotype or can be associated to a causative mutation.

For breeders, the use of molecular markers permits accurate and early selection of individuals of interest. The molecular markers shorten the number of selection cycles required, reduces time to market new lines, and lowers the overall cost of breeding.

The desire to create high-density marker chip arrays that can interrogate a large number of SNPs per DNA sample has dominated the evolution of SNP genotyping. Research has led to many high-resolution SNP arrays. Several high-density soy SNP arrays were developed and commercialized: SoySNP50K assay (Song et al. 2013), BARCSoySNP6K, and BARCSoySNP3K (Song et al. 2020). Researchers have made less of an effort to develop informative, high-throughput, and cost-effective mid-density genotyping solutions for applied molecular breeding programs and seed production Quality Assurance (QA). The advent of Next Generation Sequencing technology and genotyping by targeted sequencing provides an attractive method for mid-density SNP genotyping.

The Soy 1K SNP Panel

The Community soy 1K SNP panel at AgriPlex Genomics (https://agriplexgenomics.com/) is made of 1290 SNPs and consists of two parts: Genomic Screen and Trait Markers.

The Genomic Screen consists of 1213 markers. Originally, the SNPs were part of the SoySNP50K array (Song et al. 2013). Initial reduction of the array resulted in the BARCSoySNP6K (Song et al. 2020), which were reduced further to the BARCSoySNP3K. The present selection out of the 3k SNP array was made based on the following criteria:

- Reduce the number of SNPs in the same large linkage blocks in the North American elite population.
- SNP selection is based upon even spacing between SNPs in the segments of the genome outside of the major haplotype blocks.
- Location of SNPs in euchromatic vs. heterochromatic regions of the genome
- Minor allele frequency (MAF). The average MAF of the SNPs among 562 elites was 0.36, and the minimum allowed was MAF > 0.10. The average MAF in the Southern and Northern elites was 0.29 and 0.33, respectively.

The number of SNPs per chromosome ranges from 32 to 92 and averages at 61 SNPs per chromosome. The average distance between adjacent SNPs is 788 KB (Table 1).

The genomic screen portion of the panel should be revisited periodically to adjust its composition to ensure its relevance to different geographical areas, expanding germplasm, and genetic breadth.

Chromosomes	SNPs/Chromosomes	Average Distance Between Adjacent SNPs (kb)
1	59	912.3
2	84	611.6
3	53	889.7
4	67	732.1
5	52	772.2
6	54	891.2
7	60	739.3
8	66	672.7
9	63	708.1
10	66	722.8
11	32	1,142.90
12	39	980.9
13	64	616.7
14	55	873.5
15	87	569.3
16	45	826.1
17	60	678.3
18	92	649.1
19	70	711.1
20	45	1,057.40
Total	**1213**	
Average	61	787.9

Table 1: Genomic Screen: SNPs numbers and average spacing (KB) along the soy chromosomes

<u>The Trait Markers</u> portion of the panel is made of 88 markers that are either embedded within gene sequences of traits of interest or associated with them. Currently, the panel includes markers for diverse phenotypic characteristics (Table 2). These characteristics include:

- Growth habit and morphological features
- Biochemical qualities such as seed composition
- Environmental tolerances
- Disease resistance to life cycle attributes

The soy research and breeding community has contributed to this collection of trait markers. The effort to build this collection should be ongoing, enriching the panel with additional trait markers as new discoveries are made.

Category	No.of Markers	Traits
Agronomic	12	Flower color; Pod shatter; Pubsecence color; Stem termination; Narrow leaves/3 seeded pods
Composition	24	Seed coat color; Seed coat luster; Protein/Oil; seed antinutritional; seed carbohydrates; oil composition; Lipoxygenase
Maturity	9	Flowering time/maturity
Pathology	32	Aphid Resistance; Phytophthora sansomeana partial resistance; Phytophthora sojae resistance; Pythium irregulare partial resistance; Pythium sylvaticum partial resistance; SCN resistance;
Physiology	11	IDC tolerance; salt tolerance
Total	88	

Table 2: Selected traits, genes, and associated number of markers in the panel.

The 1K soy community SNP panel serves as a useful tool for conducting genomic selection, genomic prediction, and germplasm identification.

PlexSeq™: Mid-density multiplexed SNP genotyping

PlexSeq™ is a powerful invention of AgriPlex Genomics, serving as a highly effective platform for mid-density multiplexed SNP genotyping. The PlexSeq™ process holds a unique value as a genotyping platform due to the following attributes:

- *PlexSeq™* possesses a proprietary multiplexing algorithm called *PlexForm™.* The *PlexForm™* software is capable of designing all possible primers around any SNP a researcher may request. Through the utilization of artificial intelligence, the algorithm identifies optimal sets of primers that an individual can mix in just one PCR amplification reaction.
- Once an individual completes the amplifications, the amplicon mixture is equivalent to barcoded libraries produced from additional Next-Generation Sequencing (NGS) methods. This process is unique because the amplicon libraries produced from each sample are equivalent in concentration and do not require any additional equalization steps. A mixture of all the libraries are subjected to one bead clean-up and are loaded onto the sequencer. Thus, *PlexSeq™* saves critical time, plasticware, and expenses that researchers should invest elsewhere.

- *PlexSeq™* also proves advantageous as the process requires only minute quantities of crude DNA that can be isolated from a variety of tissues, enabling a quick and inexpensive DNA isolation process to start the genotyping workflow

The *PlexSeq™* workflow consists of:

- Crude DNA isolation
- Primary PCR: Highly multiplexed, low volume (3 ul) PCR amplifications
- Secondary, barcoding PCR amplifications
- Pooling: barcoded amplicons are combined into one tube; purified and quantitated
- Sequencing on an NGS sequencer
  - Upon completion of the sequencing process, *PlexCall™* (a proprietary allele frequency-based genotype calling analysis software) provides an automated sequencer-to-data workflow. This Java-based software is finely tuned for each assay, fully automated based on the sequencing output files and a Sample Sheet indicating sample location on the plate.
- The process is amenable to automation; all steps can be conducted on liquid handlers and high-capacity thermocyclers. This simple workflow enables high throughput genotyping.
- Molecular breeding typically requires genotyping a large number of individuals. AgriPlex Genomics' vast collection of barcode combinations permits simultaneous sequencing of up to 55,000 individuals, limited only by the sequencer's capacity.
- In some scenarios, molecular breeding may require the addition or substitution of specific SNP markers. For instance, individual breeding programs may advance or change focus regarding a parent's genetic makeup. The fact that the panel is a collection of PCR primers that are not tethered to a surface (e.g., chips) provides convenient flexibility; a researcher may dynamically customize and alter the composition of the markers in the panel, so it bests fits the germplasm or application.

The 1K soy SNP panel is available as a service from AgriPlex Genomics, where the turnaround time is 3-4 weeks. The panel and software are also available as a kit that can used by other NGS genotyping laboratories.

Applications:

Breeding applications were a major consideration during the 1k soy SNP panel. A study was conducted to characterize and validate the panel. AgriPlex Genomics used 2055 lines from 6 breeding programs together with the parental lines of the Nested Associated Mapping project (Song et al 2017). The average number of polymorphic markers in any pair-wise comparison ranged from 325 to 776 and averaged 662 polymorphic markers, or an average polymorphism of 55%. The marker density is sufficient to enable imputation back to the full genome level and allows for genomic selection. *PlexSeq's™* rapid turnaround time and low cost enable cost-effective genotyping of a prediction population during the last generation of line fixation, saving researchers money on the cost of field space for seed increase and allowing rapid recycling of progeny as parents.

The 1K soy SNP provides an excellent low-cost alternative for background recovery estimates in marker-assisted backcrossing programs. The combination of highly informative background markers with a wide selection of trait markers allows for an accurate estimation of background recovery, ensures valuable genes from the recipient line are recovered, and provides additional confirmation that a target gene is carried by the selected progeny. Background selection reduces the number of backcross generations required by 2 or more to achieve e.g., >95% recipient recovery.

Outside of molecular breeding applications, given the built-in level of average polymorphism, an individual can use the 1K soy SNP panel for genetic research and QTL mapping for example in bi-parental mapping efforts. The panel provides cost-effective genotyping that covers much of the genome.

In addition, the added trait markers provide informative data on the background of lines. The 1K soy SNP provides a low-cost method of analyzing 73 well characterized trait markers including major maturity genes: (E1, E2, E3), seed composition traits: (rs2 low RFO, GmSWEET 39 protein indel, Fad2 high oleic markers), physiological markers (3 salt tolerance markers and 7 IDC markers) and Soy Cyst Nematode resistance markers. These markers and other included traits (listed in full table) allow researchers to reduce costs for genotyping common markers and also allow for exploratory research for other traits in their program.

Conclusions

AgriPlex Genomics' implementation of the 1k soy community SNP panel is providing an excellent, cost-effective alternative for applications requiring mid-density SNP numbers over any number of sample throughput. The panel shapes neatly with rapid line fixation protocols due to its low cost-per-sample and fast turnaround time. This has allowed for major-locus selection and genomic selection to occur before field multiplication of seed, which saves critical time and expenses. The flexibility of the platform permits continual revision and upgrading of the marker system, ensuring the process keeps pace with current trait needs.

The panel primarily enables molecular breeding applications, however, the suite of trait markers, and genome coverage expand its usefulness in a range of other research applications

References

Carter T.E., Hymowitz T., Nelson R.L. (2004) Biogeography, Local Adaptation, Vavilov, and Genetic Diversity in Soybean. In: Werner D. (eds) Biological Resources and Migration. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-06083-4_5

Kihara, H. (1969) History of biology and other sciences in Japan in retrospect. Proc. XII Int. Cong. Genet. 3:49–70.

Zenglu Li and R.L. Nelson. 2001. Genetic Diversity among Soybean Accessions from Three Countries Measured by RAPDs. Crop Sci. 41: 1337-1347.

Schmutz, J., Cannon, S., Schlueter, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* 463**,** 178–183 (2010). https://doi.org/10.1038/nature08670

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus*, et al.*, 2013 Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PloS one* 8 (1):e54985. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0054985

Qijian Song, Long Yan, Charles Quigley, Edward Fickus, He Wei, Linfeng Chen, Faming Dong, Susan Araya, Jinlong Liu, David Hyten, Vincent Pantalone and Randall L. Nelson, Soybean BARCSoySNP6K: An assay for soybean genetics and breeding research, The Plant Journal (2020) 104, 800–811

Song, Qijian; Yan, Long; Quigley, Charles; Jordan, Brandon D.; Fickus, Edward; Schroeder, Steve; Song, BaoHua; Charles An, Yong-Qiang; Hyten, David; Nelson, Randall L.; Rainey, Katy; Beavis, William D.; Specht, Jim; Diers, Brian; and Cregan, Perry, "Genetic Characterization of the Soybean Nested Association Mapping Population" (2017). Agronomy & Horticulture -- Faculty Publications. 1212. https://digitalcommons.unl.edu/agronomyfacpub/1212

PlexSeq™: The mid-density multiplexed SNP genotyping solution

The revised SNP collection was again multiplexed and validated as a PlexSeq panel. The most important attribute of PlexSeq is its simplicity. PlexSeq™ workflow consists of:

- Primer design and multiplex prediction: A proprietary algorithm that uses artificial intelligence to identify the optimal sets of compatible primers that can be mixed in one PCR amplification.
- Crude DNA isolation.
- Thermocycling: highly multiplexed Primary PCR followed by secondary barcoding PCR.
- Pooling and Sequencing: barcoded amplicons are combined, purified, quantitated, and uploaded on an NGS sequencers.

This workflow is amenable to automation; all steps are carried out on liquid handlers and high-capacity thermocyclers. The efficiency and usefulness of the panel is further supported by AgriPlex’ s large collection of barcode combinations. These allow simultaneous sequencing of up to 55,000 individuals; thus, effectively, the number of individuals tested simultaneously is limited only by the sequencer’s capacity.

In addition, the panel is a collection of PCR primers not tethered to a surface (e.g.: chips). This provides the flexibility to dynamically customize and alter the composition of the markers in the panel so it best fits the germplasm or the application. Once the sequencing is complete, a proprietary genotype calling analysis software, provides an automated sequencer to data report workflow.

The 1k RiCA V4 is available as a service from AgriPlex Genomics. The panel and software are also available as a kit to be used by in-house genotyping laboratories.

References:

1. Chen H, He H, Zou Y, Chen W, Yu R, Liu X, Yang Y, Gao YM, Xu JL, Fan LM, Li Y, Li ZK, Deng XW (2011) Development and application of a set of breeder-friendly SNP markers for genetic analyses and molecular breeding of rice (Oryza sativa L.) Theor Appl Genet 123:869–879. https://doi.org/10.1007/s00122-011-1633-5
2. Thomson MJ, Singh N, Dwiyanti MS, Wang DR, Wright MH, Perez FA, DeClerck G, Chin JH, Malitic-Layaoen GA, Juanillas VM, Dilla-Ermita CJ, Mauleon R, Kretzschmar T, McCouch SR (2017) Large-scale deployment of a rice 6 K SNP array for genetics and breeding applications. Rice 10-40. https://doi.org/10.11 86/s12284-017-0181-2
3. McCouch SR, Wright MH, Tung C-W, Maron LG, McNally KL, Fitzgerald M, Singh N, DeClerck G, Agosto-Perez F, Korniliev P, Greenberg AJ, Naredo MEB, Mercado SMQ, Harrington SE, Shi Y, Branchini DA, Kuser-Falcão PR, Leung H, Ebana K, Yano M, Eizenga G, McClung A, Mezey J (2016)

Open access resources for genome-wide association mapping in rice. Nat Commun 7: 10532. https://doi.org/10.1038/ncomms10532

4. Juan David Arbelaez, Maria Stefanie Dwiyanti, Erwin Tandayu, Krizzel Llantada, Annalhea Jarana, John Carlos Ignacio, John Damien Platten, Joshua Cobb, Jessica Elaine Rutkoski, Michael J. Thomson and Tobias Kretzschmar (2019) 1k-RiCA (1K-Rice Custom Amplicon) a novel genotyping amplicon-based SNP assay for genetics and breeding applications in rice. Rice 12-55. https://doi.org/10.1186/s12284-019-0311-0
5. Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, Ulat VJ, Chebotarov D, Zhang G, Li Z, Mauleon R, Hamilton RS, McNally KL (2015) SNP-seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res 43:D1023–D1027. https://doi.org/10.1093/nar/gku1039
6. Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, Sanciangco M, Palis K, Copetti D, Poliakov A, Dubchak I, Solovyev V, Wing RA, Hamilton RS, Mauleon R, McNally KL, Alexandrov N (2017) Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res 45: D1075– D1081. https://doi.org/10.1093/nar/gkw1135
7. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann JC, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49. https://doi.org/10.1038/s41586-018- 0063-