Science Bulletin 67 (2022) 1716–1719

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

News & Views

A general model for "germplasm-omics" data sharing and mining: a case study of SoyFGB v2.0

Tianqing Zheng [a,1], Yinghui Li [a,1], Yanfei Li [a], Shengrui Zhang [a], Tianli Ge [a], Chunchao Wang [a], Fan Zhang [a], Muhiuddin Faruquee [b], Lina Zhang [a], Xiangyun Wu [a], Yu Tian [a], Shan Jiang [a], Jianlong Xu [a], Lijuan Qiu [a,*]

[a] Institute of Crop Sciences/National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[b] International Rice Research Institute, Bangladesh Office, Dhaka 1213, Bangladesh

Soybean (*Glycine max* (L.) Merr.) is a crop that originated in China. Soybean is one of the most important sources of edible vegetable protein and oil, and it has become a model legume species in genomic research. Worldwide gene banks, such as the National Crop Genebank of China (NCGC) and the United States Department of Agriculture-Agricultural Research Service (USDA-ARS) Soybean Germplasm Collection, contain more than 170,000 soybean accessions that encompass genetic diversity in both the cultivated soybean (*G. max*) and its progenitor, *G. soja*. However, sharing of large germplasm-omics data sets remains a bottleneck.

At present, the field of plant genomics is transitioning from theory to application. Two barriers prevent the widespread sharing of crop Genebank data: (1) as a leading factor in both breeding and genetic studies, germplasm-omics data, especially phenotypic data, is still difficult to reuse, and (2) the balance between efficiency and cost is challenging for germplasm databases, especially those maintained by individual researchers.

The popular online resource Phytozome [1] makes a few soybean reference genomes available for plant researchers, and Soybase [2] provides soybean genetic information based on chip (SoySNP50K) data. In recent years, several studies [3–5] have reported the re-sequencing of soybean genomes from both wild and cultivated accessions. On this basis, a database called MBKbase has plans to release a set of data (https://www.mbkbase.org/soybean) based on a recent pan-genome report [4]. LegumeIP, an integrative database for comparative genomics and transcriptomics of model legumes, has recently been updated to its third version [6]. However, the previous two barriers still remain unsolved in these works. Here, we present the data sharing mode embedded in SoyFGB v2.0 (https://sfgb.rmbreeding.cn/, developed by the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, China) with the 2214-accession soybean core collection (2K-SG) as an example of our efforts.

As shown in Fig. 1a, SoyFGB v2.0 includes the 2214 soybean core collection (2K-SG) with accessions from four major soybean production and distribution areas (Asia, America, Europe, and Africa). The 2K-SG dataset comprises three major classes of soybean species: cultivated species (1993 accessions of *G. max*), annual wild species (218 *G. soja* accessions), and perennial wild species (2 *G. tomentella* accessions and 1 *G. tabacina* accession). The *G. soja* accessions and *G. max* landraces were collected from their native geographic ranges in East Asia. Improved *G. max* cultivars were sampled globally, mainly from primary soybean-producing countries such as the USA, Japan, Republic of Korea, and China. A total of 1690 out of the 1993 cultivated soybean accessions (84.8%) were selected from the Chinese primary and applied core collections based on 14 agronomic traits and the sequences of 60 single-copy genetic loci, representing the broad genetic diversity of the 23,587 cultivated soybean accessions conserved in the NCGC [7]. Whole-genome resequencing was performed using a standard procedure. DNA sequencing libraries were generated using the TruseqNano® DNA HT sample preparation kit (Illumina Inc., San Diego, USA) following the manufacturer's recommendations. PCR products were purified using the AMPure XP bead system, and the libraries were analyzed for size distribution on an Agilent2100 Bioanalyzer. Subsequently, the Illumina Hiseq X platform was used to generate 150-base paired-end (PE) sequence reads. Removal of low-quality paired reads resulted in 16.41 Tb of high-quality PE reads, of which 96.05% and 90.98% had Phred quality scores $\geq$ Q20 and $\geq$ Q30, respectively. To call sequence variants, we first mapped the reads to the soybean reference genome (Williams 82 assembly v2.0; https://www.phytozome.net/soybean) using BWA software (v0.7.17-r1188; https://bio-bwa.sourceforge.net). Duplicates were then marked with Picard tools (v2.18.15; https://broadinstitute.github.io/picard). Subsequently, we performed gVCF calling according to the best practices using the Genome Analysis Toolkit (GATK, version v4.1.2.0, https://gatk.broadinstitute.org/hc/en-us) with the HaplotypeCaller-based method. Consequently, a total of 65,374,688 single nucleotide polymorphism (SNPs, 60,153,828 are bi-allelic) and 10,952,749 InDels (8,349,613 small insertions and deletions <15 bp and fewer than 50% missing) were identified in 2K-SG. Based on a random subset of 8,785,134 highly-credible biallelic SNPs, two different levels of grouping were carried out and are presented in SoyFGB v2.0. Level one (Group 1), the SNP-only level, includes 1507 cultivated, 313 wild, and 394 admixture accessions. Level two (Group

* Corresponding author.
E-mail address: qiulijuan@caas.cn (L. Qiu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2022.08.001

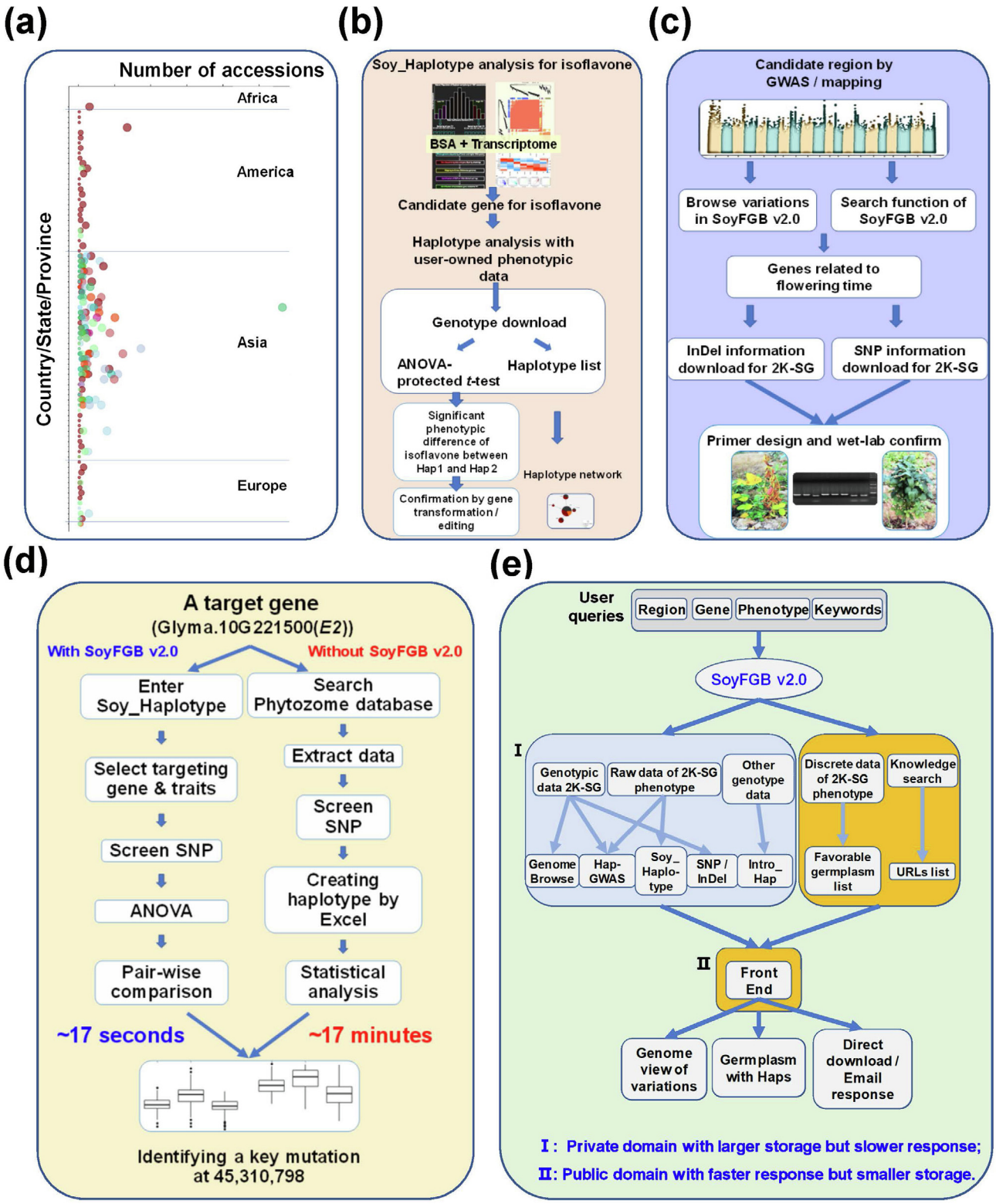

Fig. 1. Sample details, three user cases, and the structure of SoyFGB v2.0. (a) The core collection of 2214 sequenced soybean genomes (2K-SG) embedded in SoyFGB v2.0. The *x*-axis shows the origins of the accessions, and the sizes of the bubbles/circles at similar levels on the *y*-axis indicate the number of accessions. (b) A user case of mining for variations in 2K-SG in genes involved in seed isoflavone content using SoyFGB v2.0. (c) A user case of marker development based on SNP and InDel variation within a gene for maturity in soybean. (d) A comparison of the times required for the identification of key mutations using haplotype analysis with or without SoyFGB v2.0. (e) The database structure of SoyFGB v2.0.

2), based on the output of a two-step grouping, includes a species-based subgrouping and an SNP-based subgrouping within each group. In Group 2, the *G. max* group was divided into five sub-groups; the southern China region (C_SR), central China region (C_CR), northern China region plus Japan, the Korean peninsula, and the Russian far east region (C_NR), America (C_Am), and

admixtures (C_AD). The *G. soja* group was divided into four subgroups; the southern China region (W_SR), central China region (W_CR), northern China region plus Japan, the Korean peninsula, and the Russian far east region (W_NR), and the admixture (W_AD) subgroups.

It is well known that InDel and SNP marker loci tend to cluster throughout the genome [8]. Thus, in the present haplotype analysis release, SNPs are still the main factors, and the InDels shown by "-" were also taken into consideration during the analysis. Additionally, heterozygotes were regarded as an additional type. In the "Soy_Haplotype" module, a straightforward statistical analysis based on analysis of variance (ANOVA) protected *t*-test is provided. In the "Hap_GWAS" module, a linear model for GWAS [9] was adopted. Phenotypic data were obtained from data accumulated at the NCGC. Quantitative data were then transformed into discrete values based on the distributions of trait values.

SoyFGB v2.0 includes three major tabs; "Search", "Browse", and "Analysis". The "Search" tab contains four modules including "Germplasm", "Phenotype", "Gene (SNP & InDel)", and "Knowledge". Users can select favourable germplasm by discrete phenotypic data in "Phenotype" or by target gene variations embedded in the "Gene (SNP & InDel)" tab. More information about 2K-SG and soybean research is provided by the "Germplasm" and "Knowledge" modules. With the "Browse" tab, the SNP or InDel variations are accessible in a genome browser view embedded in the "SNP" and "InDel" modules, respectively. In the "Analysis" tab, three modules named "Hap-GWAS", "Soy_Haplotype", and "Intro_Hap" are provided. With these tools, users may perform a deep mining for haplotypes in genotyped soybean accessions. Typical uses of SoyFGB v2.0 are demonstrated in the following user cases:

(1) *Exploring soybean germplasm based on discrete-phenotype or accession information.* A typical pre-breeding/forward genetics scheme starts with phenotyping. In SoyFGB v2.0, a set of discrete-phenotype data covering 42 traits, including nine quality and 33 quantitative traits are embedded in the Search tab based on the "Phenotype" module. The user can screen the 2K-SG germplasm collection with discrete-phenotype data. A three-step route can be followed to explore elite donors for a breeding scheme: (a) target trait scaling, demonstrated herein by screening the top 30% in protein content as an example, which includes 13 samples; (b) from these samples, favourable early-maturing (top 50 %) samples were further screened, and favourable samples may be added to create a list of candidate germplasm (3 samples); (c) the user can then export a list of candidate donors for different breeding schemes based on the two grouping levels. An easy way to access the "Seed Request" module is available via a single click on a key called "Request Germplasm".

(2) *Haplotype mining with embedded/user-owned 2K-SG phenotypic data.* In the "Soy_Haplotype" module, the user can mine the haplotype variations from the 2K-SG collection in a defined target region using gene name, physical range, or even a set of SNPs. With the SoyFGB-embedded or user-owned phenotypic data, the phenotypic effects of different haplotypes for target traits are available to the user. The donor lists of different haplotypes are also provided for users with supporting evidence from statistical analyses based on the ANOVA protected *t*-test.

As an example, the candidate genes for isoflavone content in soybean were identified by a combination of bulked segregant analysis (BSA) with a natural population and weighted gene co-expression network analysis (WGCNA) using the transcriptomes of different seed developmental stages. SoyFGB v2.0 provided the haplotype analysis function for the candidate genes. Firstly, the locus number of one candidate gene ID and the phenotypic data for isoflavone content from 2K-SG from user data were submitted to the "Soy_Haplotype" module embedded in the Analysis tab of SoyFGB v2.0. All the haplotypes of this gene were then presented. Subsequently, with the aid of a straightforward statistical analysis between different haplotypes, germplasm accessions harboring the different haplotypes were found to be significantly distinct from one another in isoflavone content. This implies the possible contribution of the candidate gene in regulating the isoflavone content of soybean grain. Finally, the haplotype variations and the germplasm list for the candidate gene were also downloaded for further laboratory work (Fig. 1b). Alternatively, an enhanced correlation between the phenotypes and haplotypes could be explored with the "Hap-GWAS" module, which uses the methodology described recently [9]. In order to save the possible waiting time for this analysis, an email reminder system was adopted. Once the results of the analysis are ready, a reminder email containing a direct access link to the output is sent to a mailbox defined by the user. Together with the instant screening using the "Soy_Haplotype" module, correlations between the phenotypes and haplotypes may be mined at different levels. A number of data sets have been generated using relatively low-density genotyping methods, such as the SNP chip. A tool for haplotype analysis in target regions using this type of data in populations with or without known parents is also provided in the "Intro_Hap" module. Since a request for this tab was recently made by users during an indoor testing period, we are still looking for more user responses since this release of SoyFGB v2.0.

(3) *Searching for variation within candidate genes for favorable germplasm.* A route for shortlisting candidate genes using SoyFGB v2.0 is shown in Fig. 1c and involves the following: (a) identifying a target region using mapping methods such as GWAS or sorting accessions with favorable target traits; (b) using the "SNP" or "InDel" modules in the Browse tab to explore the variations within a target gene/region; (c) inputting the target region or gene locus ID into the "Gene(SNP/InDel)" module of the Search tab; and (d) with the downloaded genotype information (SNP or InDel), users can perform further work with primer design and laboratory confirmation. In the example shown in Fig. 1c, a marker for maturity time in soybean was developed based on the above description. We have compared the efficiencies of identifying key mutations with or without SoyFGB v2.0. As shown in Fig. 1d, for a known target gene (*E2*) with the ID Glyma.10G221500, the key mutation at position 45,310,798 can be found in less than 17 s with SoyFGB v2.0. However, it could take more than 60 times as long (17 min) using other known tools.

In summary, a workflow is provided in the Flowchart page of the Introduction module in the "About" tab through the following URL: https://sfgb.rmbreeding.cn/about/introduction for users to follow. More details are also accessible through this link.

In the "omics" era, a suitable mode for phenotypic data sharing of Genebank germplasm is urgently needed [10]. In contrast to genomic and other omics data, phenotypic data are rarely reused. Much of the phenotypic data in gene banks is not openly accessible, even though it complies with the FAIR criteria [11].

Our objective in designing SoyFGB v2.0 was to set up a general phenotypic data sharing mode for germplasm accessions of a crop (soybean) that is under strict control for data sharing, and which is designed to be adaptive and responsive to the overwhelming quantity of genomic and phenotypic data. The FGB general data sharing mode in SoyFGB v2.0 has the following characteristics:

(1) Mining elite donor lines with favourable haplotypes is of high value to breeders. The correlations between phenotypes and SNPs, such as from GWAS results, can only be accessible through search functions on other websites [12,13]. Instead of providing a direct download link to raw phenotypic data, SoyFGB v2.0 not only employs a discrete-phenotype data-led germplasm sharing mode, but also performs online analyses through the "Hap-GWAS", "Soy_Haplotype", and "Intro_Hap" modules. This has provided a platform that is more conducive to data contributors sharing their own unpublished data with public users.

(2) To keep up with the development of multiple-client ends, a development framework different from the previous FGB website [14] was employed in SoyFGB v2.0. The website is driven by Nginx, including the front end developed using the Vue-Element-Axios tool, and the back end was developed with the java-based tool. The RESTful API facilitates easy data access through different client platforms. Additionally, in order to balance efficiency and cost, a distributed database structure was designed for SoyFGB v2.0 (Fig. 1e). Phenotypic and genotypic data are stored on servers with different capacities. Phenotypic data are managed by an instant-response server with a relatively small storage capacity using the MySQL database, and the genotypic data are stored on a server with a slower response but a larger storage capacity. All these features are aimed at meeting developing trends including decentralization and multiple ways of accessing the ever-increasing amount of biological data in the near future.

(3) Searching plant omics databases for functional information has grown in popularity [15]. Accordingly, the "Search" function in SoyFGB v2.0 is important for helping users search for useful information inside and/or outside of SoyFGB. Through the three major embedded tabs, users can access 2K-SG data in an effective and efficient manner.

(4) Many correlations between phenotypes and haplotypes can be directly mined in soybean using SoyFGB v2.0. Moreover, the efficiency of haplotype mining in soybean should be largely improved (by ~60 folds).

In summary, SoyFGB v2.0 is an example of a portal that was established for sharing and mining big germplasm-omics datasets at the phenotypic and genotypic levels from more than 2200 soybean accessions. The NCGC has now provided a phenotype-led germplasm-omics data sharing platform in SoyFGB, which may inspire new ideas for data sharing and mining in other crops.

Data availability

All data provided by SoyFGB v2.0 are accessible through the following URL: https://sfgb.rmbreeding.cn/.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2021YFD1201601, 2020YFE0202300, and 2016YFD0100201), the National Natural Science Fundation of China (31871715), the International Science & Technology Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-ZDRW202109), Central Public-interest Scientific Institution Basal Research Fund (Y2020PT24 and Y2020YJ09), the Science & Technology Innovation Program of Chinese Academy of Agricultural Sciences (ICS2020YJ07BX), Hainan Yazhou Bay Seed Lab (B21HJ0216), and the Bill & Melinda Gates Foundation (OPP1130530).

References

[1] Goodstein DM, Shu S, Howson R, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 2012;40:D1178–86.
[2] Grant D, Nelson RT, Cannon SB, et al. Soybase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res 2010;38:D843–6.
[3] Li YH, Zhou G, Ma J, et al. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 2014;32:1045–52.
[4] Liu YC, Du HL, Li PC, et al. Pan-genome of wild and cultivated soybeans. Cell 2020;182:162–76.
[5] Torkamaneh D, Laroche J, Valliyodan B, et al. Soybean (*Glycine max*) haplotype map (GmHapMap): a universal resource for soybean translational and functional genomics. Plant Biotechol J 2021;19:324–34.
[6] Dai X, Zhuang Z, Boschiero C, et al. LegumeIP V3: from models to crops-an integrative gene discovery platform for translational genomics in legumes. Nucleic Acids Res 2021;49:D1472–9.
[7] Wang LX, Guan Y, Guan RX, et al. Establishment of Chinese soybean *Glycine max* core collections with agronomic traits and SSR markers. Euphytica 2006;151:215–23.
[8] Montgomery SB, Goode DL, Kvikstad E, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res 2013;23:749–61.
[9] Zhang F, Wang C, Li M, et al. The landscape of gene–CDS–haplotype diversity in rice: Properties, population organization, footprints of domestication and breeding, and implications for genetic improvement. Mol Plant 2021;14:787–804.
[10] Yang WN, Feng H, Zhang XH, et al. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. Mol Plant 2020;13:187–214.
[11] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The fair guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018.
[12] Zhao H, Li J, Yang L, et al. An inferred functional impact map of genetic variants in rice. Mol Plant 2021;14:1584–99.
[13] Li C, Tian D, Tang B, et al. Genome variation map: a worldwide collection of genome variations across multiple species. Nucleic Acids Res 2020;49:D1186–91.
[14] Wang CC, Yu H, Huang J, et al. Towards a deeper haplotype mining of complex traits in rice with RFGB v2.0. Plant Biotechol J 2020;18:14–6.
[15] Gui S, Yang L, Li J, et al. ZEAMAP, a comprehensive database adapted to the maize multi-omics era. iScience 2020;23:101241.

Tianqing Zheng received his Ph.D. degree in 2006 after working jointly at International Rice Research Institute and Nanjing Agricultural University. He currently works at the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences with a major research focus on big data-driven molecular breeding.

Yinghui Li received her Ph.D. degree in Agronomy from the China Agricultural University in 2003. Currently, she is a professor at the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences. Her research interest includes (1) understanding the genetic basis of adaptation and domestication using soybean and its wild relatives; (2) identifying genes associated with important traits in soybean germplasm using combined phenomics, genomics, and transcriptomics approaches for improving modern cultivars.

Lijuan Qiu received her Ph.D. degree in Agronomy from the Northeast Agricultural University in 1989. Currently, she is a professor at the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences. Her current research focuses on maintaining the national soybean germplasm resource collection, identifying elite soybean germplasm, discovering and elucidating the function of genes and the molecular mechanisms governing important traits, and developing elite cultivars using biotechnology.