

SCIENTIFIC DATA



OPEN

DATA DESCRIPTOR

Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing

Babu Valliyodan ^{1,2,18}, Anne V. Brown ^{3,18}, Juexin Wang ^{4,18}, Gunvant Patil ^{1,5,18}, Yang Liu ^{6,18}, Paul I. Otyama ⁷, Rex T. Nelson ³, Tri Vuong ¹, Qijian Song ⁸, Theresa A. Musket ¹, Ruth Wagner ⁹, Pradeep Marri ^{10,11}, Sam Reddy ¹⁰, Allen Sessions ¹², Xiaolei Wu ¹², David Grant ^{3,7}, Philipp E. Bayer ¹³, Manish Roorkiwal ¹⁴, Rajeev K. Varshney ¹⁴, Xin Liu ^{15,16}, David Edwards ¹³, Dong Xu ^{4,6}, Trupti Joshi ^{4,6,17}, Steven B. Cannon ³ & Henry T. Nguyen ¹ 

We report characteristics of soybean genetic diversity and structure from the resequencing of 481 diverse soybean accessions, comprising 52 wild (*Glycine soja*) selections and 429 cultivated (*Glycine max*) varieties (landraces and elites). This data was used to identify 7.8 million SNPs, to predict SNP effects relative to genic regions, and to identify the genetic structure, relationships, and linkage disequilibrium. We found evidence of distinct, mostly independent selection of lineages by particular geographic location. Among cultivated varieties, we identified numerous highly conserved regions, suggesting selection during domestication. Comparisons of these accessions against the whole U.S. germplasm genotyped with the SoySNP50K iSelect BeadChip revealed that over 95% of the re-sequenced accessions have a high similarity to their SoySNP50K counterparts. Probable errors in seed source or genotype tracking were also identified in approximately 5% of the accessions.

Background & Summary

Soybean [*Glycine max* (L.) Merr.] is one of the major grain legumes and oil seeds cultivated worldwide, particularly in Asia and the Americas. The cultivated soybean, *G. max*, was domesticated from its wild relative, *G. soja*, around the Eleventh Century B.C, in Eastern China¹. Cultivated soybean spread to other locations through Asia shortly following domestication, and was then introduced into the United States in 1765¹. Soybean lost genetic diversity through domestication-related genetic bottlenecks, while the wild relative *G. soja*, growing in various environmental conditions, retained significant genetic diversity^{2,3}.

¹Division of Plant Sciences and National Center for Soybean Biotechnology, University of Missouri, Columbia, MO, 65211, USA. ²Department of Agriculture and Environmental Sciences, Lincoln University, Jefferson City, MO, 65101, USA. ³USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, IA, 50011, USA. ⁴Department of Electrical Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, 65211, USA. ⁵Institute of Genomics for Crop Abiotic Stress Tolerance, Department of Plant and Soil Science, Texas Tech University, Lubbock, TX, 79409, USA. ⁶MU Institute of Data Science and Informatics, University of Missouri, Columbia, MO, 65211, USA. ⁷Department of Agronomy, Iowa State University, Ames, IA, 50011, USA. ⁸USDA-ARS, Soybean Genomics and Improvement Lab, Beltsville, MD, 20705, USA. ⁹Bayer CropScience, St. Louis, MO, 63141, USA. ¹⁰Corteva Agriscience, Indianapolis, IN, 46268, USA. ¹¹Pairwise Plants LLC, Durham, NC, 27709, USA. ¹²Bayer CropScience, Research Triangle Park, NC, 27709, USA. ¹³School of Biological Sciences, The University of Western Australia, Perth, WA, 6009, Australia. ¹⁴International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Hyderabad, Telangana, 502324, India. ¹⁵Beijing Genomics Institute-Shenzhen, Shenzhen, 518083, China. ¹⁶State Key Laboratory of Agricultural Genomics, China National GeneBank, BGI-Shenzhen, Shenzhen, 518083, China. ¹⁷Department of Health Management and Informatics, University of Missouri, Columbia, MO, 65211, USA. ¹⁸These authors contributed equally: Babu Valliyodan, Anne V. Brown, Juexin Wang, Gunvant Patil, Yang Liu.  e-mail: nguyenhenry@missouri.edu

The first reference genome for cultivated soybean was released in 2010⁴, followed by high-quality assemblies of several other soybean accessions, including elite lines and wild soybean accessions^{5–10}. This genomic sequencing has been complemented by resequencing projects (summarized in Figshare file F1¹¹), designed to assess genetic variation across broader samplings of soybean germplasm^{3,5,12,13}.

To map patterns of genome-wide variation, population structure, and to facilitate identification of the genetic basis of agronomic traits in soybean, 481 accessions from the USDA Soybean Germplasm Collection were re-sequenced. This included modern cultivars, traditional landraces, and wild species accessions, from throughout the range of both domesticated and wild soybean (Figshare file F2¹¹). All of the reads were mapped onto the Williams 82 reference genome assembly and approximately 7.8 million single nucleotide polymorphisms (SNPs) were identified from the mapped reads. The *G. max* Williams 82 assembly was used as reference because that is most widely used and accepted. We compare these results from the chip-based genotyping of the same genotypes from the SoySNP50K iSelect BeadChip¹⁴ results, helping validate both results, while also identifying several probable errors in germplasm tracking or labeling. The genotypic data was also used to evaluate genetic relationships and population structure. We provide the variant data and associated analyses as major resources for other researchers, toward better understanding of soybean diversity for continued genetic improvement of soybeans.

Methods

Germplasm selection for sequencing. The USDA Soybean Germplasm Collection contains 20,035 accessions, including 1,168 *G. soja* and 18,867 *G. max* lines. The SoySNP50K BeadChip¹⁴ was utilized for genotyping these accessions and for selecting a core set of diverse *G. max* (1,148) and *G. soja* (81) lines for further studies. The core set maintained 98.7% and 96.1% of the diversity of the entire cultivated and wild soybean collections, respectively¹⁴. To maximize diversity without losing genetic information, we selected a total of 481 diverse soybean accessions from the core mentioned above (429 of *G. max* and 52 of *G. soja*) for sequencing and analysis (Figshare file F2¹¹). The germplasm of the 481 accessions comes from 37 countries, with the largest numbers coming from China (223), Japan (56), the U.S. (54), South Korea (44), the Russian Federation (26), and North Korea (14).

Soybean sample collection, DNA sequencing, and read-mapping. Soybean seeds were obtained from the Soybean Stock Center (USDA-GRIN). For each accession 50 seeds were planted, in a greenhouse at the University of Missouri. After two weeks (V1 growth stage) a minimum of 24 plants per accession were selected based on morphological homogeneity (hypocotyl color, plant height, leaf size) for leaf sample collection. Leaf samples were flash frozen and high quality high molecular weight DNA extractions were conducted using a standard protocol¹⁵. Soybean samples were tested for the heterogeneity using the SoySNP6K Illumina Infinium BeadChips (Illumina, Inc. San Diego, CA)¹⁶ and the samples with below 10% heterogeneity were selected for whole-genome sequencing. A total of 5 µg of genomic DNA from each soybean line was used to construct the sequencing library, following the Illumina sequencing protocols. Paired-end sequencing libraries with an insert size of ~300 bp were sequenced using an Illumina HiSeq. 2000 sequencer at the Beijing Genomics Institute (BGI). The 50 most diverse lines from the core set were sequenced at a 40x genome equivalent and the remaining samples were sequenced at 15x coverage (Figshare file F3¹¹). This generated a total of 7.5 TB of raw next generation sequencing (NGS) reads data from all samples.

The reference genome for the soybean cultivar Williams 82⁴ (Wm82.a2) (a widely used reference assembly), downloaded from Phytozome (www.phytozome.net)¹⁷, was used for mapping. We built the PGen¹⁸ multi-step SNP identification workflow using the Pegasus¹⁹ workflow management system (Pegasus-WMS), for quality checks, alignment of reads, variants calling, variants filtration, and VCF merging. The workflow takes paired-end and single-end Fastq reads as input, and performs data quality checks using FastQC²⁰. Filtered, high-quality reads are then aligned against the reference genome using BWA²¹. Picard Tools²² is also used at this step to locate duplicate molecules and assign all reads into groups (using default parameters). The bwa command is 'bwa mem -t 12 -M ref.fastq paired_read1.fastq paired_read2.fastq > aln.sam'.

Variant detection. After the sequence alignment, SNPs and indels were called using the Haplotype caller algorithm from the Genome Analysis Toolkit (GATK)²³ version gatk-2.5-2-gf57256b. Filtering criteria are provided in the INFO fields in the VCF file. Important abbreviations in this section are: QD, quality by depth; FS, Fisher strand values; and MQ, mapping quality of variants. Detected variants were then filtered using the criteria "QD < 26.0 || FS > 60.0 || MQ < 40.0" for SNPs and "QD < 26.0 || FS > 200.0 || MQ < 40.0" for indels. Additional filtering can also be applied by modifying the configuration file of a PGen workflow. Outputs were generated as BAM and VCF standard formats that were stored in the CyVerse data store²⁴ and accessed via the SoyKB database²⁵ via the NGS re-sequencing data browser^{25,26}, and for download at the SoyBase Data Store²⁷. SNPs were assigned IDs using the script "assign_name.awk" available at <https://github.com/soybase/SoySNP-Names>. Missing data accounted for 1% and heterozygous SNPs accounted for 0.5% of the total dataset (Figshare file F4¹¹). SNPs were annotated using SnpEff 3.0²⁸ (Figshare file F5¹¹). In SoyKB the data is also loaded into the SNPviz 2.0 tool²⁹ for interactive exploration of accession relationships using SNPs in selected genomic regions.

Phylogenetic analysis. 25,496 SNPs in common with the SoySNP50K iSelect BeadChip-derived data were used to develop the phylogenetic tree (Fig. 1 and Data File glyma.Wm82.gnm2.div.G787.sampled_25Kpos.tree, SoyBase Data Store³⁰). To generate an alignment suitable for phylogenetic reconstruction, every 5th SNP was selected, giving an alignment length of 4,518 characters (Data File glyma.Wm82.gnm2.div.G787.sampled_25Kpos.fna, SoyBase Data Store³⁰). This moderate matrix size permits maximum likelihood phylogenetic reconstruction and also takes advantage of the SNP distribution in the SoySNP50K set, as the SoySNP50K SNPs were chosen relative



Fig. 1 Phylogenetic tree of the 481 re-sequenced accessions. The tree is rooted between *G. max* and *G. soja* accessions. Colors indicate countries of origin: blue, China; orange, Korea; green, Japan; cyan, United States; gray, all others (predominantly from Russia). Cultivars of interest are highlighted on the tree.

to LD, with greater density at chromosome ends and less density in the high-LD pericentromere, and avoiding closely-spaced SNPs. A tag was added to the genotype identifiers to indicate country of origin (Figshare file F2¹¹). A maximum likelihood phylogenetic tree was calculated using FastTree³¹, version 2.1.8, with default nucleotide parameters. Tree visualizations were generated using the Archaeopteryx tree viewer³².

Chromosome	Chromosome size (Mb)	LD decay distance (Mb)
Gm20	47.88	0.061
Gm19	50.68	0.199
Gm18	57.97	0.292
Gm17	41.62	0.093
Gm16	37.83	0.053
Gm15	51.67	0.192
Gm14	48.98	0.098
Gm13	45.81	0.080
Gm12	40.01	0.084
Gm11	34.73	0.105
Gm10	51.50	0.092
Gm09	50.15	0.118
Gm08	47.80	0.160
Gm07	44.61	0.276
Gm06	51.32	0.081
Gm05	42.19	0.139
Gm04	52.37	0.077
Gm03	45.70	0.065
Gm02	48.57	0.097
Gm01	56.83	0.105
Average	47.41	0.123
Correlation		0.33

Table 1. Average decay of LD (r^2) as a function of physical distance between pairs of loci across soybean chromosomes. Pairwise comparisons are shown for the first 1000 kb distance. Decay curves are plotted following the Hill and Weir method (Supplementary Figure 1). The curve in green represents average decay for all accessions whereas the blue and red curves represent average decay in *G. soja* and *G. max* accessions, respectively.

Linkage disequilibrium (LD) analysis. The extent and rate of LD decay was calculated as the square of the correlation coefficient (r^2) between pairwise comparisons of all SNPs across a chromosome (Table 1, Supplementary Figure 1). For LD analysis accessions were divided into two groups: *G. max* and *G. soja*, to compare the LD between wild and cultivated soybeans. Pairwise r^2 measures were calculated using PLINK v 1.90b4.4³³ using parameters `--dog --r2 --ld-window-r2 0 --ld-window 999999 --ld-window-kb 80000`. For each chromosome, r^2 values were plotted against inter-marker distances to visualize rate of decay at the baseline level where $r^2 = 0.2$. The decay curve was estimated using the Hill and Weir formula³⁴.

Structure analysis. Structure within the collection was calculated using the Bayesian clustering program FastStructure³⁵ using a logistic prior for K ranging 1 to 10. The script chooseK.py (part of the FastStructure distribution) was used to determine the best K that explained the structure in the collection based on model complexity. Structure was visualized (Fig. 2) using an R package Pophelper v2.3.0³⁶.

Chromosomal similarity heatmaps displayed in Fig. 3 were created using the tool GCViT (Genotype Comparison Visualization Tool^{37,38}), available for interactive use at <https://soybase.org/gcvit>. Soybean accession Essex (PI 548667) was selected to be the reference genotype due to its importance in soybean breeding programs. Bin size is set to 500,000 with the right side of the chromosomes set to display type “heatmap” showing SNP differences between the selected accessions and Essex. Heatmap base color set to white going from min-max.

Comparison with variants in the SoySNP50K array. The two VCF files were merged on common SNP positions using BCFtools^{39,40}. Heterozygous SNPs were treated as missing data before creating a similarity matrix. Within the merged VCF file, missing data and heterozygous SNPs combined accounted for <1.5% (Figshare file F6¹¹). The R package SNPRelate⁴¹ was used to create a similarity matrix of the merged datasets and compare the accessions to one another. The script extractTop-Match.pl, available on GitHub⁴², was used to extract the top matches for a given PI in a similarity matrix (Figshare file F7¹¹).

Data Records

The authors declare that all data reported here are fully and freely available from the date of publication. All major data products are indicated in the Data Dictionary at the National Ag Library project data collection⁴³. The sequencing data have been deposited in the NCBI Sequence Read Archive projects SRP062245⁴⁴ and SRP105183⁴⁵. We used 376 accessions raw data (331 *G. max* and 45 *G. soja*) from the SRP105183⁴⁵ and selected 105 accessions (98 *G. max* and 7 *G. soja*) from the SRP062245⁴⁴ for the analysis. Variant data and accession information are available at the SoyBase⁴⁶ project page³⁰ and the associated Data Store²⁷. The variant data are also accessible for interactive exploration at SoyKB²⁶. Also, all data including the detailed information of the accessions, phylogenetic tree, variant data are available at the National Agricultural Library Ag Data Commons (<https://doi.org/10.15482/USDA.ADC/1518301>)⁴³ and at figshare¹¹.

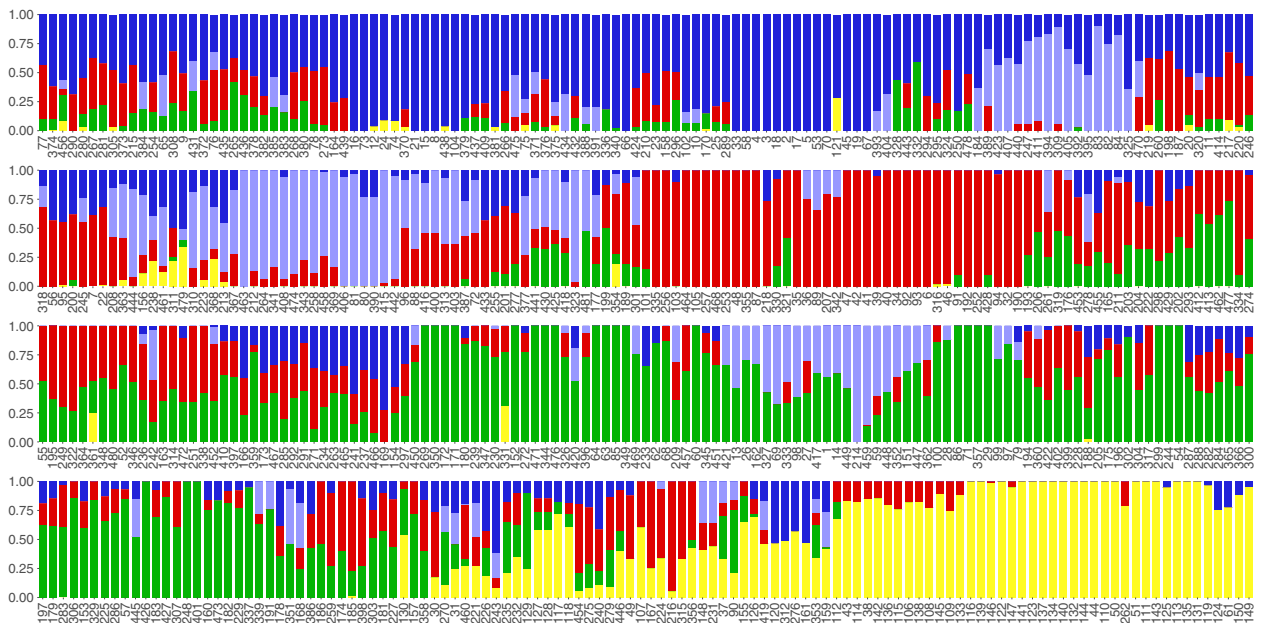


Fig. 2 Genetic Structure plot for the 481 re-sequenced accessions. Accessions are plotted in the same order as in the phylogenetic tree in Fig. 1. Numbers indicated below each structure represent the position of that accession in the VCF file. The plot was constructed using cluster size of K = 5. Accessions in yellow generally correspond with *G. soja*.

Differences between Essex and a selection of Elite, Landrace and Wild Cultivars

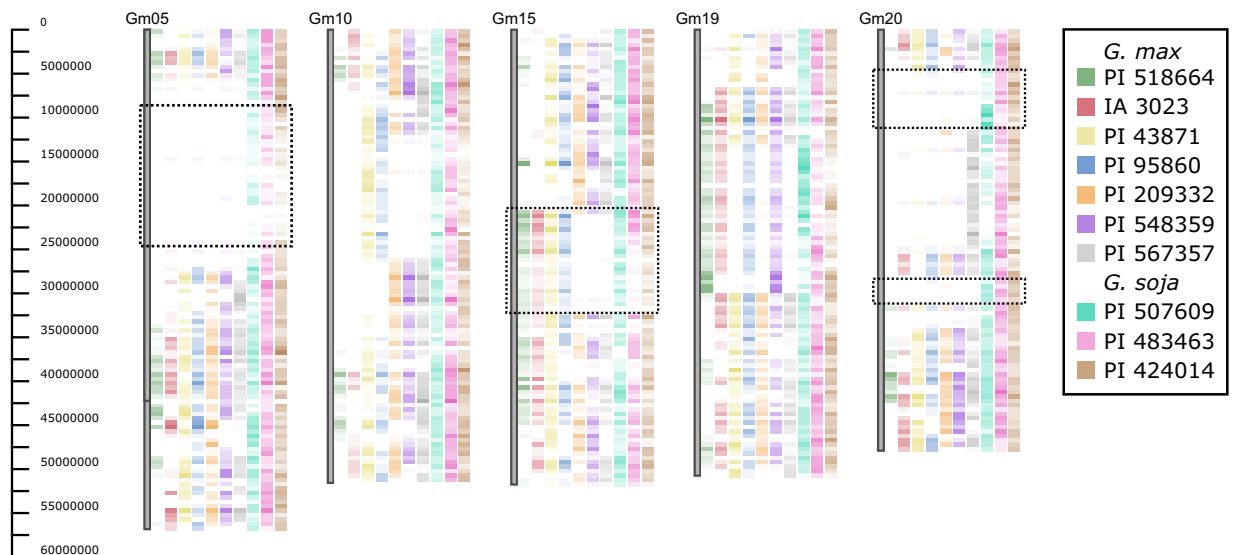


Fig. 3 Heatmap comparison of cultivated and wild accessions to elite breeding line Essex. The darker the color, the more differences between the given accession and Essex. White regions indicate no differences between the given accession and Essex (PI 548667). Accessions are ordered and color-coded from left to right as follows: green, PI518664 (Hutcheson); red, IA3023; yellow, PI438471 (Fiskeby III); blue, PI 95860 (*G. max* line from Korea); orange, PI 209332 (*G. max* line from Japan); purple, PI 548359 (Kingwa); gray, PI 567357 (top of phylogenetic tree); turquoise, PI 507609 (*G. soja* from Japan); pink, PI 483463 (*G. soja* from China); brown, PI 424014 (*G. soja* from Korea).

Technical Validation

Assessment of variant calls and accessions relative to the U.S. soybean germplasm collection. All but 5 of the 481 accessions sequenced in this study have a counterpart in the U.S. germplasm collection (sharing the same PI accession number). The VCF file was merged with the SoySNP50K VCF file from the SoyBase Data Store⁴⁷, which contains genotype information for the entire USDA soybean germplasm collection⁴⁸, based on common SNP positions. The R program SNPRelate⁴¹ was used to create a similarity matrix of all of the

lines in the VCF file. In soybean, there is known variation between accessions with the same name⁴⁹, however, it is expected that the top match for each of the 481 accessions would be the accession with the same PI name from the SoySNP50K dataset. The similarity scores between the re-sequenced accessions and its SoySNP50K counterpart ranged from 99–55%, with an average similarity score of 0.987. Of the 481 accessions, 441 had a 99% or greater similarity to their expected match, while 20 lines had a similarity between 90–99% to their expected match (Figshare file F5¹¹). Of the 461 accessions that had a similarity >90% to their counterpart, 407 accessions had the same PI identifier as the top match from the SoySNP50K. In 54 cases, the top hit from the SoySNP50K was not the expected PI accession number, but the difference in percent identity between the expected accession and top match was within 3%. These results indicate genetic redundancy for some groups of genotypes in the USDA soybean germplasm collection.

The 5 accessions that did not have a counterpart in the SoySNP50K dataset had similarity <90% to all other lines, indicating that these lines are unique to this study. Fifteen accessions were labeled as mis-assigned, as these accessions had a >3% similarity difference between the expected top match and what was observed, indicating an error in the accession's identity. These differences could be due to the fact that some accessions were collected at markets that included a mixture of seed. This explanation may apply for PIs PI 407262, PI 424079, PI 437160, PI 639586, PI 360957, PI 628913, PI 639559B and their respective top matches: PI 407264, PI 424048, PI 437153 A, PI 639610, PI 379563, PI 628917, PI 639558. For all of these cases, the PIs and their top matches all have the same country of origin. For example, PI 407262 has common name K42-A and is from South Korea, its top match is PI 407264, also from South Korea, and with a similar common name of K42-C. The top match of accession PI 548402 is to PI 438497. These accessions have the same common name, 'Peking', which could have led to a mis-labeling, as there are 7 lines in the USDA Germplasm Resources Information Network (GRIN) with the common name 'Peking'. This also illustrates that two accessions with the same common name may not be genetically identical – or even particularly similar. Considering accessions that had a 90–99% similarity to their U.S. counterpart, we found that in the SoySNP50K data, there are two accessions that reference soybean line Hutcheson ('Hutcheson' and 'PI 518664'). There are differences between these lines due to 'Hutcheson' being grown out for multiple generations in a lab after seed was obtained from GRIN, while PI 518664 was sourced directly from GRIN (Qijian Song, personal communication.). We suspect that a 90–99% similarity score to the SoySNP50K counterpart could be due to this event when 2 accessions have been grown out separately for multiple generations and inadvertent selection occurs.

Genomic diversity. Resequencing of 376 accessions yielded 680 billion 125-bp paired-end reads, resulting in 6.5Tb of high-quality raw data (Figshare file F3¹¹). This 6.5 Tb data was combined with 1 Tb (105 selected accessions) of publicly available raw data³. Sequence reads were aligned to soybean reference genome Wm82.a2 (downloaded from Phytozome), using the BWA aligner²¹. The mapping rate varied from 97.52%–99.50%, averaging 99.14% in the *G. max* lines and 97% in *G. soja* (Figshare file F3¹¹). From the mapped sequence a total of 7,869,806 SNPs were identified.

Variant effects. The snpEff program²⁸, used to predict SNP effects, labeled 7,590,330 SNPs as “modifier” (falling in intergenic regions), 125,602 as having a “low” effect (causing a synonymous mutation), 146,236 as having a “moderate” effect (causing a non-disruptive change in the protein), and 7,638 as having a “high” effect (causing a disruptive change in the protein). The 7,638 SNPs predicted as having a “high” effect are in 5,987 genes (Figshare file F5¹¹ and file glyma.Wm82.gnm2.div.G787.snpEff.gff3 at the SoyBase Data Store²⁷). The SNPs, along with their snpEff annotation, can be viewed on the SoyBase genome browser (<https://soybase.org/gb2/gbrowse/gmax2.0/>) under the “naturally occurring sequence variants” track, as “USB481”.

SnpEff was run on the *G. max* and *G. soja* data files glyma.Wm82.gnm2.div.G787.USB481_nosoja.vcf.gz and glyma.Wm82.gnm2.div.G787.Soja.vcf.gz at the SoyBase Data Store²⁷, giving the respective snpEFF files at that location. The *G. max* results indicated 4,661,844 SNPs as a “modifier”, 76,352 as having a “low” effect, 82,567 as having a “moderate” effect, and 4,041 as having a “high” effect. The 4,041 SNPs as having a “high” effect fall into 3,420 genes. The *G. soja* results indicated 9,356,860 SNPs as a “modifier”, 187,992 as having a “low” effect, 212,196 as having a “moderate” effect, and 10,326 as having a “high” effect.

Usage Notes

The variant data for the 481 diverse soybean accessions have many potential uses. We illustrate with four analyses: calculation of linkage disequilibrium decay; visualization of regional conservation and difference between accessions; and phylogenetic and structural analyses. Online tools for visualizing similarities and differences for the 481-accession data set are described below. The phylogenetic analysis and associated files are also available for interactive exploration.

Linkage disequilibrium analysis. A set of 25,495 SNPs in-common between the resequencing results and the SoySNP50K haplotypes⁴⁸ were used to survey the rate of LD decay and population structure in the collection. The LD for each chromosome and the average LD across all 20 chromosomes is indicated in Table 1. The average LD decay distance value for the 429 *G. max* accessions is 173 kb, at $r^2 = 0.2$. (The 45 accessions of *G. soja* were too few to calculate LD decay that is meaningfully comparable to LD for the *G. soja* accessions). There is no clear correlation ($r^2 = 0.33$) between chromosome size and LD decay distance (Table 1, Supplementary Figure 1). The LD decays at approximately the same distance (0.1 Mb) in both the shortest and second longest chromosomes (chromosomes 11 and 1).

Conservation and divergence of genomic regions. The tool GCViT (Genotype Comparison Visualization Tool)^{37,38} available at SoyBase (<https://soybase.org/gcvit/>) was used to identify genomic regions that

differ between accessions. A small subset of cultivated and wild accessions were selected randomly to compare against the elite cultivar Essex (PI 548677) (Fig. 3). The three *G. soja* accessions were selected to include one from each county of China, Japan, and Korea. Essex was chosen to be used as a reference because of its importance in soybean breeding programs. Five of the 20 chromosomes are displayed in the figure. The heatmaps show blocks of conserved regions between cultivated accessions on chromosomes 5 and 20. Similar results are produced when comparing any cultivated soybean to wild within this dataset. Darker regions correspond to genomic regions that differ greatly between Essex and the accession of comparison whereas the lighter regions show shared genomic regions. Conserved regions could indicate genomic regions that were selected during domestication, as similar results (e.g. conservation on chromosomes 5 and 20) were reported by Han *et al.*⁵⁰. Also noteworthy is a region in the middle of chromosome 15, where Essex shares a genomic region with accessions PI 209332, PI 548359 (Kingwa), and PI 567357, but is different between all other accessions, suggesting selective introgressions. Highly similar results are produced using different elite cultivars as a reference and comparing other landrace and wild cultivars. For example, if Lee (PI 548656) is used as the reference instead of Essex, we see the exact same results on Gm05 and Gm20. Results can be tested/confirmed/explored using the GCViT tool^{37,38} at SoyBase and selecting the USB481 dataset.

Phylogenetic and structure analysis. 25,496 SNPs from this study that were in common with the SoySNP50K iSelect BeadChip-derived data¹⁴ were used to develop a phylogenetic tree showing similarity relationships among accessions (Fig. 1). The tree is rooted between *G. max* and *G. soja* accessions. In both the *G. max* and *G. soja* clades, the accessions generally cluster by country of origin, as indicated by groupings of colors (countries): Japanese (green), Korean (orange), Chinese (blue), U.S. (light blue), and all other countries in gray. This suggests that phylogenetically distinct lineages arose within particular geographic locations, with relatively limited genetic exchange between, for example, central China and Japan. The accessions near the top (Fig. 1, predominantly dark blue clade, extending from PI 567357 and ending prior to Fiskeby III) are primarily from China while the U.S. elite lines are mostly in the lower middle (green clade, extending from Fiskeby III through PI 95860). A few accessions are highlighted and named in the figure, in view of their importance in U.S. breeding programs. Several cultivars near the top of the tree have been important sources of disease resistance: PI 88788 and Peking for SCN resistance, and Kingwa for Phytophthora resistance. Fiskeby III and Lee are tolerant to salt⁵¹, while Essex, Hutcheson, and IA3023 are common parents in many breeding programs. *G. soja*, PI 483463, is highlighted in the figure due to its genome having been recently sequenced and made available⁹.

To complement the phylogenetic analysis, a genomic Structure plot for all 481 accessions was generated based on a cluster size of $K = 5$ (Fig. 2). The order of the accessions in the Structure plot is the same as in the phylogenetic tree (Fig. 1). The Structure results generally correspond with the phylogenetic results – for example, with *G. soja* lines (yellow) occur together, as expected considering their genetic distance from *G. max*.

In this study, we have presented re-sequencing data of 481 diverse accessions, including 52 wildtypes (*G. soja*). 7.8 million SNPs have been identified and more than 5,900 genes with high effect changes have been discovered among the germplasm collection. These changes will be of use in soybean breeding programs.

Among the USDA soybean germplasm collection, we have identified 15 accessions that are potentially misnamed which will help other researchers avoid errors in their analyses. Illustration of the data with phylogenetic and structure analyses highlights the history of soybean domestication through mostly independent selection in numerous locations across Asia.

The results presented here help build a more complete history of the US soybean breeding programs, which in turn will guide future efforts in soybean breeding.

Code availability

Scripts used to extract information from similarity matrix can be found on Github⁴².

Received: 9 April 2020; Accepted: 6 January 2021;

Published online: 08 February 2021

References

- Hymowitz, T. & Harlan, J. R. Introduction of soybean to North America by Samuel Bowen in 1765. *Economic Botany* **37**, 371–379 (1983).
- Hyten, D. L. *et al.* Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci USA* **103**, 16666–16671 (2006).
- Valliyodan, B. *et al.* Landscape of genomic diversity and trait discovery in soybean. *Scientific reports* **6**, 1–10 (2016).
- Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
- Li, Y. H. *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature biotechnology* **32**, 1045–1052 (2014).
- Liu, Q., Chang, S., Hartman, G. L. & Domier, L. L. Assembly and annotation of a draft genome sequence for *Glycine latifolia*, a perennial wild relative of soybean. *The Plant journal: for cell and molecular biology* **95**, 71–85 (2018).
- Shen, Y. *et al.* De novo assembly of a Chinese soybean genome. *Science China Life Sciences* **61**, 871–884 (2018).
- Shimomura, M. *et al.* The Glycine max cv. Enrei Genome for Improvement of Japanese Soybean Cultivars. *International journal of genomics* **2015**, 358127 (2015).
- Valliyodan, B. *et al.* Construction and comparison of three reference-quality genome assemblies for soybean. *The Plant journal: for cell and molecular biology* **100**, 1066–1082 (2019).
- Xie, M. *et al.* A reference-grade wild soybean genome. *Nat Commun* **10**, 1216 (2019).
- Valliyodan, B. *et al.* Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. *figshare* <https://doi.org/10.6084/m9.figshare.c.4863006> (2021).
- Lam, H.-M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature genetics* **42**, 1053–1059 (2010).
- Zhou, Z. *et al.* Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature biotechnology* **33**, 408–414 (2015).

14. Song, Q. *et al.* Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One* **8**, e54985 (2013).
15. Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res* **8**, 4321–4325 (1980).
16. Hyten, D. L. *et al.* 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 (2010).
17. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40**, D1178–D1186 (2012).
18. Liu, Y. *et al.* PGen: large-scale genomic variations analysis workflow and browser in SoyKB. *BMC Bioinformatics* **17**, 337 (2016).
19. Deelman, E. *et al.* Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming* **13**, 219–237 (2005).
20. Andrews, S. *FastQC: A quality control tool for high throughput sequence data* <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).
21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
22. Broad Institute. *Picard Tools* <http://broadinstitute.github.io/picard/> (2020).
23. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297–1303 (2010).
24. Goff, S. A. *et al.* The iPlant collaborative: cyberinfrastructure for plant biology. *Frontiers in plant science* **2**, 34 (2011).
25. Joshi, T. *et al.* Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. *BMC genomics* **13**, 1 (2012).
26. Joshi, T., Xu, D., Valliyodan, B. & Stacey, G. *NGS Resequencing Browsr* http://soykb.org/NGS_Resequencing/NGS_index.php (2020).
27. Valliyodan, B., Brown, A. V., Nguyen, H. T. & Cannon, S. B. Variant data for 481 accessions. *SoyBase Data Store*. https://soybase.org/data/public/Glycine_max/Wm82.gnm2.div.G787 (2020).
28. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
29. Langewisch, T. *et al.* Major soybean maturity gene haplotypes revealed by SNPviz analysis of 72 sequenced soybean genomes. *PLoS one* **9**, e94150 (2014).
30. Nelson, R. T., Cannon, S. B. & Grant, D. Large Scale Sequencing of Germplasm to Develop Genomic Resources for Soybean Improvement *SoyBase Data Store*. <https://soybase.org/projects/SoyBase.B2014.02.php> (2020).
31. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
32. Han, M. V. & Zmasek, C. M. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* **10**, 356 (2009).
33. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559–575 (2007).
34. Hill, W. G. & Weir, B. S. Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical population biology* **33**, 54–78 (1988).
35. Raj, A., Stephens, M. & Pritchard, J. K. FastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
36. Francis, R. M. pophelper: an R package and web app to analyse and visualize population structure. *Molecular ecology resources* **17**, 27–32 (2017).
37. Wilkey, A. P., Brown, A. V., Cannon, S. B. & Cannon, E. K. S. GCViT: a method for interactive, genome-wide visualization of resequencing and SNP array data. *BMC Genomics* **21**, 822 (2020).
38. Wilkey, A., Weeks, N. & Brown, A. V. LegumeFederation/gcvit: GCViT 1.0. *Zenodo* <https://doi.org/10.5281/zenodo.4008713> (2020).
39. Narasimhan, V. *et al.* BCftools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**, 1749–1751 (2016).
40. Danecek, P. & McCarthy, S. A. BCftools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039 (2017).
41. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
42. Brown, A. V. *SimMatrix-Analysis* <https://github.com/avbrown1/SimMatrix-Analysis> (2020).
43. Valliyodan, B., Brown, A. V., Cannon, S. B. & Nguyen, H. T. Data from: Genetic variation among 481 diverse soybean accessions. *Ag Data Commons* <https://doi.org/10.15482/USDA.ADC/1518301> (2020).
44. *NCBI Sequence Read Archive* <https://identifiers.org/ncbi/insdc.sra:SRP062245> (2020).
45. *NCBI Sequence Read Archive* <https://identifiers.org/ncbi/insdc.sra:SRP105183> (2020).
46. Grant, D., Nelson, R. T., Cannon, S. B. & Shoemaker, R. C. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res* **38**, D843–846 (2010).
47. Song, Q. *et al.* Fingerprinting soybean germplasm and its utility in genomic research. *SoyBase Data Store*. https://soybase.org/data/public/Glycine_max/Wm82.gnm2.div.892R/ (2015).
48. Song, Q. *et al.* Fingerprinting Soybean Germplasm and Its Utility in Genomic. *Research G3 (Bethesda)* **5**, 1999–2006 (2015).
49. Haun, W. J. *et al.* The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol* **155**, 645–655 (2011).
50. Han, Y. *et al.* Domestication footprints anchor genomic regions of agronomic importance in soybeans. *The New phytologist* **209**, 871–884 (2016).
51. Do, T. D. *et al.* Mapping and confirmation of loci for salt tolerance in a novel soybean germplasm, Fiskeby III TAG. *Theoretical and applied genetics. Theoretische und angewandte Genetik* **131**, 513–524 (2018).

Acknowledgements

We acknowledge funding support from the United Soybean Board for the large-scale sequencing of soybean genomes (project #1320-532-5615) and two private companies: Bayer (previously Monsanto and Bayer) and Corteva (previously Dow AgroSciences). The authors would like to thank Andrew Wilkey for his help with GCViT and figure production and Prasanna Vaduvathiryan, Biomedical Librarian, University of Kansas Medical Center for her help with formatting the references. We acknowledge Dr. Paul Chomet for his comments and suggestions on the soybean re-sequencing project plan, while he was working at the Monsanto Company. In-kind support for analysis was provided by the USDA Agricultural Research Service project 5030-21000-062-00D. The USDA is an equal opportunity provider and employer.

Author contributions

H.T.N. is the principal investigator for the project. H.T.N. and B.V. conceived the idea, designed the experiments, and managed the project. A.V.B., J.W., Y.L., P.I.O., R.T.N., P.E.B., M.R., B.V., Q.S., D.X., T.J., S.B.C. conducted the genome sequence and variant analysis. R.W., P.M., S.R., A.S., X.W., X.L., R.K.V. and D.E. contributed to the project planning and key discussions. B.V., G.P., T.V. and T.M. contributed to the plant growth, sample preparation, DNA

quality tests and data generation. S.B.C., A.V.B., R.T.N. and D.G. contributed to data storage and prepared the SoyBase page for the project data. B.V., A.V.B., S.B.C., G.P., J.W., Y.L., D.X., T.J. and H.T.N. participated in the first draft of the article, and S.B.C., A.V.B. and B.V. wrote the final draft.

Competing interests

Babu Valliyodan, Anne V. Brown, Juexin Wang, Gunvant Patil, Yang Liu, Paul I. Otyama, Rex T. Nelson, Tri Vuong, Qijian Song, Theresa A. Musket, Pradeep Marri, Sam Reddy, Allen Sessions, Xiaolei Wu, David Grant, Phillip E. Bayer, Manish Roorkiwal, Rajeev K. Varshney, Xin Liu, David Edwards, Dong Xu, Trupti Joshi, Steven B. Cannon, and Henry T. Nguyen declare no competing financial interests. Ruth Wagner is an employee of Bayer Crop Sciences, a developer of agricultural seeds.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-021-00834-w>.

Correspondence and requests for materials should be addressed to H.T.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2021