BMC Genomics

RESEARCH

Open Access

# Development of a versatile resource for post-genomic research through consolidating and characterizing 1500 diverse wild and cultivated soybean genomes

Hengyou Zhang[1†], He Jiang[1†], Zhenbin Hu[1†], Qijian Song[2] and Yong-qiang Charles An[1,3*]

## Abstract

**Background:** With advances in next-generation sequencing technologies, an unprecedented amount of soybean accessions has been sequenced by many individual studies and made available as raw sequencing reads for post-genomic research.

**Results:** To develop a consolidated and user-friendly genomic resource for post-genomic research, we consolidated the raw resequencing data of 1465 soybean genomes available in the public and 91 highly diverse wild soybean genomes newly sequenced. These altogether provided a collection of 1556 sequenced genomes of 1501 diverse accessions (1.5 K). The collection comprises of wild, landraces and elite cultivars of soybean that were grown in East Asia or major soybean cultivating areas around the world. Our extensive sequence analysis discovered 32 million single nucleotide polymorphisms (32mSNPs) and revealed a SNP density of 30 SNPs/kb and 12 non-synonymous SNPs/gene reflecting a high structural and functional genomic diversity of the new collection. Each SNP was annotated with 30 categories of structural and/or functional information. We further identified paired accessions between the 1.5 K and 20,087 (20 K) accessions in US collection as genomic "equivalent" accessions sharing the highest genomic identity for minimizing the barriers in soybean germplasm exchange between countries. We also exemplified the utility of 32mSNPs in enhancing post-genomics research through in-silico genotyping, high-resolution GWAS, discovering and/or characterizing genes and alleles/mutations, identifying germplasms containing beneficial alleles that are potentially experiencing artificial selection.

**Conclusion:** The comprehensive analysis of publicly available large-scale genome sequencing data of diverse cultivated accessions and the newly in-house sequenced wild accessions greatly increased the soybean genome-wide variation resolution. This could facilitate a variety of genetic and molecular-level analyses in soybean. The 32mSNPs and 1.5 K accessions with their comprehensive annotation have been made available at the SoyBase and Ag Data Commons. The dataset could further serve as a versatile and expandable core resource for exploring the exponentially increasing genome sequencing data for a variety of post-genomic research.

*Correspondence: yong-qiang.an@usda.gov
†Hengyou Zhang, He Jiang and Zhenbin Hu contributed equally to this work.
[3] US Department of Agriculture, Agricultural Research Service, Midwest Area, Plant Genetics Research Unit, 975 N Warson Rd, St. Louis, MO 63132, USA
Full list of author information is available at the end of the article

**Keywords:** Soybean, Single nucleotide polymorphism (SNP), Genetic diversity, Whole-genome resequencing, US soybean germplasm collection, Linkage disequilibrium (LD), Genomic equivalence

## Background

Soybean [*Glycine max* (L.) Merr.] is one of the most economically important field crops for its high seed protein (40%) and oil (20%), which are primarily used for animal feed, human consumption and industrial use. Soybean also plays a vital and sustainable role in agriculture through fixing atmospheric nitrogen. The worldwide soybean production has been tripled and its growing acreage has doubled since 1993 [1]. However, the demand for soybean due to its better nutritional value is predicted to continuously increase with an ever-growing world population [2]. It is critical to explore soybean genetic diversity for developing new cultivars to meet the rising demand.

With the release of the Williams 82 soybean reference genome sequence assembly in 2010 [3], a SoySNP50K iSelect Bead Chip containing 52,041 genome-wide SNPs was developed through genome resequencing of eight soybean accessions [4]. It was used to genotype 20,087 soybean accessions available in USDA soybean germplasm collection. The SoySNP50K genotyping data has become a valuable resource for a range of soybean genetics research [5–8]. In addition, transcriptome sequencing has been explored to examine the myriad transcript sequence and accumulation variations present in diverse soybean accessions. The transcriptome diversity analysis has successfully identified single nucleotide transcript variants and large DNA InDels that causes soybean seed quality and maturity variations [9–11]. With the continuous reduction in next-generation sequencing cost, a large number of soybean accessions have been re-sequenced for a wide range of soybean genome research such as developing soybean pan-genomes and understanding dynamic changes of soybean genomes [12–17]. The availability of soybean whole-genome and transcriptome sequencing data provides an unprecedented opportunity to access the genomic and genetic variation at a single nucleotide resolution in a large population and allows to explore and develop new strategies for soybean genetic improvement. Our recent analysis of 631 consolidated genome sequences from different sequencing studies facilitated uncovering of a *SWEET* gene underlying a major large-effect protein and oil QTL on chromosome 15. It supports that a two-nucleotide deletion allele in *SWEET* gene has been selected through domestication and used in breeding program to improve protein content in soybean [18]. However, most of the large-scale genome sequencing studies only focus on a specific set of accessions. Not only this, but most of the data are also released in a raw sequencing reads format in the public database like Sequence Read Archive (SRA) of the NCBI (https://www.ncbi.nlm.nih.gov/sra). However, most of laboratories lack the capacity and affordability to consolidate and systemically analyze such a huge amount of raw sequencing data available in public [19], and thus these valuable resources are relatively under-utilized. Henceforth, there is a pressing need to develop a technology platform to consolidate these massive whole-genome sequencing datasets available in the public, and systematically analyze them using the same bioinformatics pipelines and/or criteria to generate an expandable and user-friendly public-accessible resource for research community.

Soybean (*G. max*) was domesticated from *G. soja* in China about 6000 years ago [20, 21]. Artificial selection during domestication and breeding dramatically reduced the genetic diversity in modern soybean cultivars [12, 14, 22]. It has been suggested that North American soybean varieties underwent severe genetic bottlenecks [22]. Loss of genetic diversity in cultivated soybean species has been imposing a great challenge for improvement of new trait genes/variation in soybean. A recent study also revealed that genetic diversity does not increase significantly beyond approximately 800 *G. max* accessions [19]. In contrast, wild accessions possesses a high genetic diversity [12, 14, 23, 24] and is a rich genetic resource for developing improved soybean cultivars with traits like salt tolerance, cyst nematode resistance, or increased protein content [7, 18, 25, 26]. However, the number of wild soybean accessions being sequenced is very limited. To increase genome diversity for discovering novel and superior trait genes/gene variants and gain insight into soybean domestication process, it is imperative to include additional diverse wild accessions [14, 24, 27].

In this study, we analyzed the integrated publicly available genome sequences of 1465 cultivated soybean accessions and an in-house generated sequence data of 91 *G. soja* accessions representing the *G. soja* diversity in the US Soybean Germplasm Collection [12–15]. Our in-depth analysis identified and annotated 32 million SNPs (32mSNPs) across the genomes of diverse 1556 (1.5 K) soybean germplasm lines (*G. soja* accessions, landraces and improved cultivars) collected worldwide. In addition, we characterized and revealed the population structures and genetic diversity of the 1.5 K accessions. We demonstrated various utilization of the dataset for both basic and applied research in soybean. The identified variants have been made available to the research community in the public repository in a user-friendly Variant Call Format (VCF). The collection of 32mSNPs from 1.5 K accessions can serve as a valuable resource for untapping value of the huge amount of soybean genome sequencing data and exploring the genome diversity of the 1.5 K diverse soybean accessions for various soybean basic and applied research. The resource could be continuously updated with the newly generated sequencing data as additional genome sequencing data are available in the public.

## Results

### A consolidated collection of 1501 diverse soybean accessions

We retrieved approximately 15 terabytes of whole-genome sequencing reads of 1465 genomes from the NCBI SRA database. The sequencing depths ranged from 2.7- to 65.1-fold genome coverage (Table S1). Our in-house sequenced whole-genomes of 91 diverse *G. soja* accessions were at a depth ranging from 7.4 to 41.3-fold genome coverage. The 91 *G. soja* accessions represent the overall diversity of 1168 *G. soja* accessions in the US Soybean Germplasm Collection based on the genetic distances, geographic locations, and maturity groups [5]. In total, the sequencing reads of 1556 soybean genomes were generated from 1501 accessions (hereafter referred to 1.5 K).

Out of the 1.5 K accessions, a total of 1194 accessions were annotated for their germplasm types as *G. soja* accessions (204 accessions), landraces (472), and improved cultivars (518) (Table S1). The 1.5 K accessions were from a wide range of geographic locations across the world, including major soybean growing countries (Fig. 1a and Table 1). Accessions from China and the United States accounted for majority of the accessions (71.3%) with larger set from China (967 accessions) than from United States (165 accessions). The remaining 28.7% (424 accessions) accessions were collected from the other 37 countries (Table S1). This collection also included 32 accessions from Brazil, one of those largest soybean producing countries [28]. *G. soja* accessions were mainly from East Asian regions (China (71), South Korea (61), Japan (47) and Russia (25). In addition, the 1.5 K accessions were found to be distributed into 13 maturity groups (MGs 000-X) with majority (67.93%) of the collection in MGs II, III, IV, and V (Fig. 1b and Table S1). Thus, this 1.5 K collection comprising of accessions from world over harbors abundant soybean genetic diversity. Inclusion of the in-house sequenced diverse 91 *G. soja* accessions would greatly increase genetic diversity of the soybean collection and enable us to tap diversity retained in wild soybean.

### Identification and annotation of 32mSNPs among 1.5 K diverse soybean accessions

We analyzed a total of 208.3 billion paired-end sequencing reads of the 1556 genomes with an average of 133 million reads and 14-fold genome coverage per accession (Table S1). A total of 32,456,244 SNPs (designated as 32mSNPs) were discovered at an average SNP density of 30 SNPs/kb. Of these, 16.3 million and 8.4 million SNPs had minor allele frequency (MAF) higher than 0.01 and 0.05, respectively. We revealed a significantly higher density of SNPs in euchromatic regions than heterochromatic regions (Fig. 2a). Approximately, 87% of the 32mSNPs located in the intergenic regions. The rest 13% (5,193,083) were in the genic regions with an average density of 93 SNPs per gene (Fig. 2b). Of the genic SNPs, 12.1% and 63.1% located in untranslated regions (5′ and 3’UTR) and introns, respectively (Fig. 2c). The remaining 24.8% of the genic SNPs were present in the

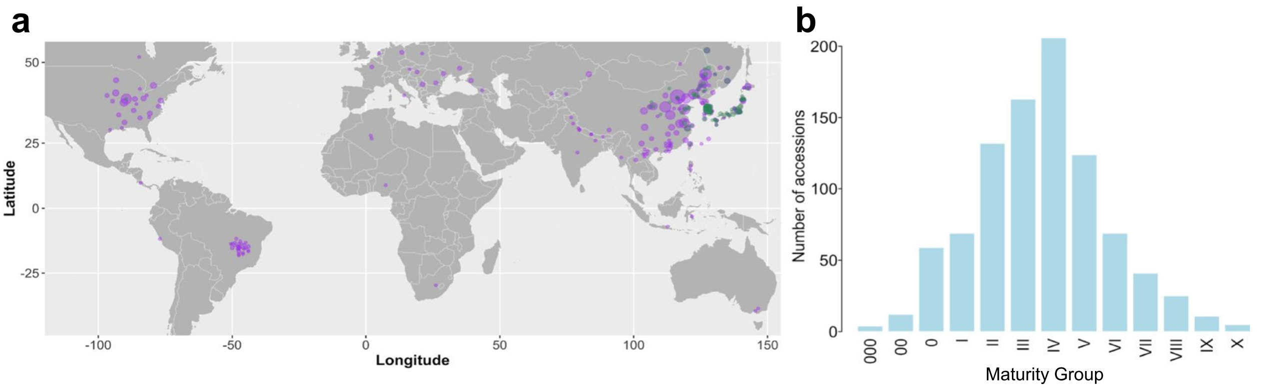


**Fig. 1** Origins and maturity groups of the 1.5 K soybean accessions. **a** Geographic distribution of the 1.5 K collection worldwide. The dots in green denote *G. soja* and those in purple denote *G. max*. Dot size denotes the sample size at the indicated geographic region. **b** Distribution of maturity group for all the accessions. Each bar represents the number of accessions belonging to the corresponding maturity group

**Table 1** The origins of 1.5 K soybean collection

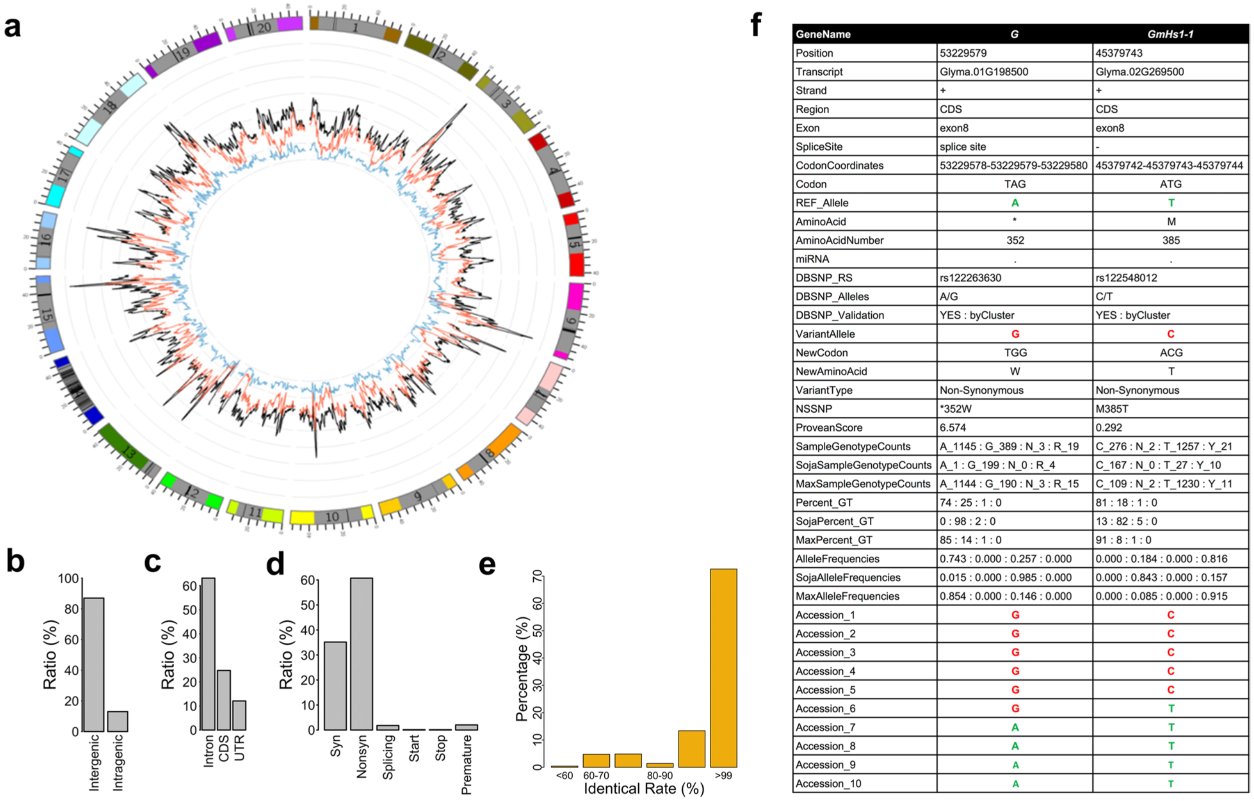
| Origins | No. of accessions | Origins | No. of accessions |
|---|---|---|---|
| China | 955 | Nepal | 2 |
| United States | 155 | Former Serbia and Montenegro | 2 |
| Japan | 105 | Serbia | 2 |
| South Korea | 93 | Indonesia | 2 |
| Russian Federation | 42 | Netherlands | 1 |
| Brazil | 32 | Peru | 1 |
| North Korea | 17 | Georgia | 1 |
| Canada | 14 | Kyrgyzstan | 1 |
| India | 8 | Belgium | 1 |
| Vietnam | 6 | Italy | 1 |
| Moldova | 4 | Nigeria | 1 |
| Ukraine | 3 | Myanmar | 1 |
| Germany | 3 | South Africa | 1 |
| Romania | 3 | Uzbekistan | 1 |
| Algeria | 2 | Sweden | 1 |
| France | 2 | Costa Rica | 1 |
| Philippines | 2 | Austria | 1 |
| Australia | 2 | Poland | 1 |
| Thailand | 2 | Unknown | 27 |
| Hungary | 2 | Total | 1501 |

Duplicated-sequenced genomes were counted once here

coding sequences (CDS) (Fig. 2c). We observed that 63% (657,371) of these SNPs in the CDS were non-synonymous. PROVEN algorithm predicted 10.7% (70,143) of the non-synonymous SNPs as deleterious with PROVEN score ≤ −4.1 that have high probability of altering gene functions [29]. Additionally, we discovered 19,506 SNPs at splicing sites, 1562 at start codons, 1465 at stop codons, and 22,076 producing premature stop codons (Fig. 2d). Importantly, 99% of the 56,044 gene models in the soybean genome (*Wm82.a2 v1*) carried at least one non-synonymous SNP with an average of 12 non-synonymous SNPs/per gene. The collection of 1.5 K accessions are highly diverse, and likely contained at least one nucleotide mutation/allele for each gene that potentially alter gene functions and cause phenotypic variation. The collection of accessions also enable us to explore and discover the novel causative gene/ allelic variants for agronomically important traits for soybean improvement.

For the 56 accessions that were sequenced twice in different studies, the duplicated sequences for the same accessions had an average of 99.5% of sequence identity, indicating high reproducibility and accuracy of the SNP calling method utilized in the present study. Out of the 1.5 K accessions, 926 accessions have also been genotyped using Soy50KSNP Chip in a separate study [5]. We observed that 73% of 926 accessions had > 99% identity at SNP positions genotyped by both platforms (Fig. 2e). The remaining 27% of the accessions showed less than 99% identity between the two platforms. Having compared the 32mSNPs with 15,623,492 registered soybean SNPs available in the NCBI dbSNP database (https://www.ncbi.nlm.nih.gov/snp/), we observed that 13 million SNPs were present in the dbSNP database. Thus, 19 million out of the 32mSNPs represented novel SNPs.

To provide a comprehensive description of each SNP, especially for those non-synonymous SNPs potentially important to gene functions, each SNP was annotated with 30 categories of structural and functional information such as the SNP position, reference and alternative alleles, and their allelic frequency in *G. soja* and *G. max*. (Fig. 2f). *G* and *GmHs1-1*, are two agronomically important genes [30, 31] (Fig. 2f). *GmHs1-1* encodes a calcineurin-like protein controlling hard-seededness and *G* gene has been associated with dormancy in soybean. The causative reference and alternative alleles for *GmHs1-1* (reference T/ alternative C) and *G* (A/G) were identified as homozygous alleles in an average of 98.6% of 1.5 K collection. The annotation of each of these alleles in two genes between two subpopulations of *G. max* and *G. soja* revealed their highly biased distribution among the subpopulations. For example, the *G* gene had 85% of A allele in *G. max* and 98% of G in *G. soja*. This result is consistent with the previous reports showing that both genes and their alleles are associated with soybean domestication.



| GeneName | *G* | *GmHs1-1* |
|---|---|---|
| Position | 53229579 | 45379743 |
| Transcript | Glyma.01G198500 | Glyma.02G269500 |
| Strand | + | + |
| Region | CDS | CDS |
| Exon | exon8 | exon8 |
| SpliceSite | splice site | - |
| CodonCoordinates | 53229578-53229579-53229580 | 45379742-45379743-45379744 |
| Codon | TAG | ATG |
| REF_Allele | A | T |
| AminoAcid | * | M |
| AminoAcidNumber | 352 | 385 |
| miRNA | . | . |
| DBSNP_RS | rs122263630 | rs122548012 |
| DBSNP_Alleles | A/G | C/T |
| DBSNP_Validation | YES : byCluster | YES : byCluster |
| VariantAllele | G | C |
| NewCodon | TGG | ACG |
| NewAminoAcid | W | T |
| VariantType | Non-Synonymous | Non-Synonymous |
| NSSNP | *352W | M385T |
| ProveanScore | 6.574 | 0.292 |
| SampleGenotypeCounts | A_1145 : G_389 : N_3 : R_19 | C_276 : N_2 : T_1257 : Y_21 |
| SojaSampleGenotypeCounts | A_1 : G_199 : N_0 : R_4 | C_167 : N_0 : T_27 : Y_10 |
| MaxSampleGenotypeCounts | A_1144 : G_190 : N_3 : R_15 | C_109 : N_2 : T_1230 : Y_11 |
| Percent_GT | 74 : 25 : 1 : 0 | 81 : 18 : 1 : 0 |
| SojaPercent_GT | 0 : 98 : 2 : 0 | 13 : 82 : 5 : 0 |
| MaxPercent_GT | 85 : 14 : 1 : 0 | 91 : 8 : 1 : 0 |
| AlleleFrequencies | 0.743 : 0.000 : 0.257 : 0.000 | 0.000 : 0.184 : 0.000 : 0.816 |
| SojaAlleleFrequencies | 0.015 : 0.000 : 0.985 : 0.000 | 0.000 : 0.843 : 0.000 : 0.157 |
| MaxAlleleFrequencies | 0.854 : 0.000 : 0.146 : 0.000 | 0.000 : 0.085 : 0.000 : 0.915 |
| Accession_1 | G | C |
| Accession_2 | G | C |
| Accession_3 | G | C |
| Accession_4 | G | C |
| Accession_5 | G | C |
| Accession_6 | G | T |
| Accession_7 | A | T |
| Accession_8 | A | T |
| Accession_9 | A | T |
| Accession_10 | A | T |

**Fig. 2** Annotation of the 32mSNPs. **a** Distribution of the SNPs along 20 chromosomes compared with the dbSNPs. The outermost circle represents the 20 soybean chromosomes. 0, 20, and 40 outside the circle represent 0 Mb, 20 Mb, and 40 Mb positions on the chromosome, respectively. The solid gray boxes and black bars indicate pericentromeric regions and centromeric repeats, respectively. The black, red, and blue curves of the inner circle showed the distribution of the SNP set identified in the 1.5 K accessions, soybean dbSNPs, and the subset of SNPs with minor allele frequency ≥ 0.25, respectively. **b** Percent of the SNPs in the 32mSNPs in genome features (intergenic and intragenic). **c** Percent of genic SNPs in the 32mSNPs (Intron, CDS, UTR). **d** Percent of synonymous (Syn), non-synonymous (Nonsyn) SNPs, and the SNPs in splicing sites, start codons, stop codons, and the SNPs causing premature in the 32mSNPs. **e** Percent of identity at the coordinate-matched SNPs for the accessions that have been genotyped by genome-resequencing and SoySNP50K. **f** A SNP report exemplified with two previously identified domestication genes. *G* encodes a stay-green gene controlling soybean dormancy and *GmHs1-1* encodes calcineurin-like protein controlling hard-seededness. The SNP report contains 30 annotation categories for each identified SNP in the 32mSNPs. Reference and alternative alleles were highlighted in green and red, respectively

The comprehensive annotation is therefore a highly valuable step for the post-genomic research including population-scale characterization of genome-wide SNPs and genes for discovering domestication-associated genomic loci of interest. These results also suggest the high coverage and robustness of the 32mSNPs.

## Population structure of 1.5 K soybean accessions

We assessed the population structure using principal component analysis (PCA) and neighbor-joining (NJ) phylogenetic analysis. The PCA revealed that the first 20 principal components (PCs) of the genetic data captured 32.53% of the total variance among the 1.5 K accessions. The first PC captured 9.25% of the variation and mainly explained the divergence between *G. max* and *G. soja* (Fig. 3a). The second PC captured 5.23% of the variation, mainly explaining variation within the germplasms, and the third PC captured 3.09% variation and mainly explained the variation within the remaining variation within the germplasms (Fig. 3b). The neighbor-joining phylogenetic tree constructed with the 1.5 K accessions showed general consistency of population structure corresponding with PCA, where accessions were clustered based on the germplasm types with some percentage of admixture between germplasms (Fig. 3a, b, c). *G. soja* accessions were mainly clustered independent of *G. max* clusters. As expected, we observed more diverse between *G. soja* and *G. max* groups than between landrace and cultivar within *G. max* group. Therefore, this collection has a clear, germplasm type-inferred population structure

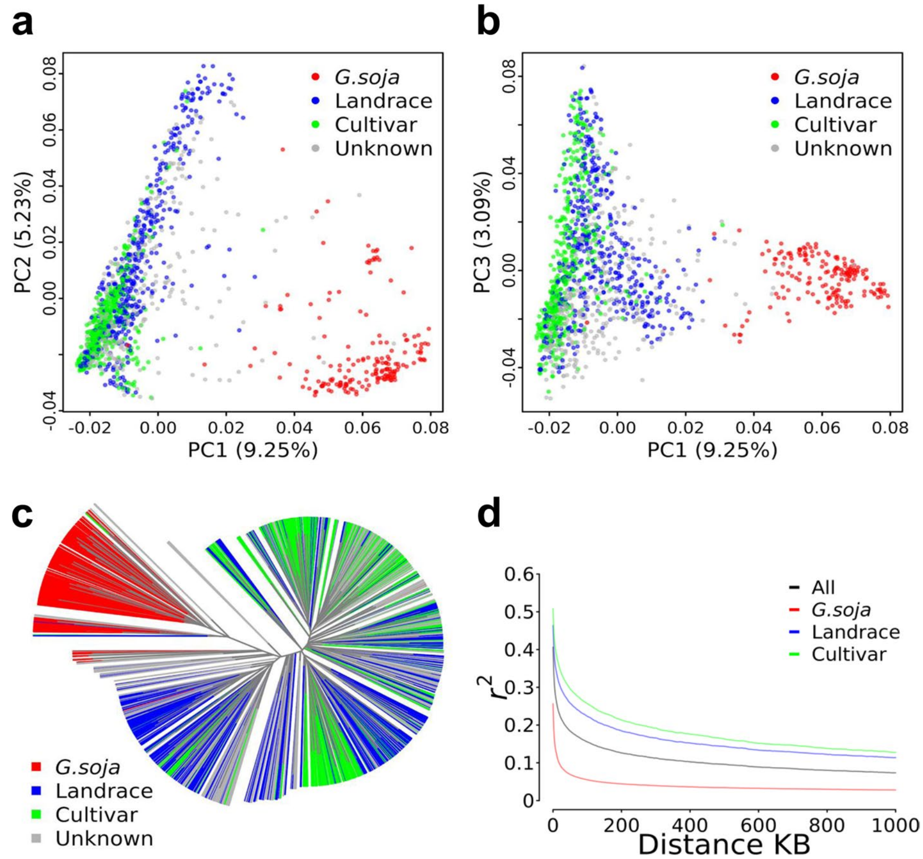


**Fig. 3** Population genomic analysis of the 1.5 K collection. **a** Scatter plots for the first two components (PC1, PC2) of the 1556 (1.5 K) accessions, and **b** for PC1 and PC3. **c** A neighbor-joining phylogenetic tree of the 1.5 K accessions. The germplasm types (*G. soja*, landrace, cultivar) are correspondingly labeled. **d** Comparison of LD decay among three germplasm types (*G. soja*, landrace, cultivar)

that allows to leverage the high-resolution SNPs to investigate complex processes of soybean domestication and improvement.

The estimation of linkage disequilibrium (LD) across the three germplasm types (Fig. 3d) revealed an overall rapid LD decay in *G. soja* compared to domesticated *G. max* landraces and cultivars analyzed independently. Overall LD decay of all accessions (all three germplasm types) was ~ 35 kb at $r^2$ of 0.2. The LD decayed in *G. soja* at ~ 2 kb at $r^2$ of 0.2, which is dramatically shorter than ~ 151 kb for landraces and much shorter than ~ 255 kb for cultivars (Fig. 3d). This result indicates that the highly diverse *G. soja* accessions in the collection greatly reduced the sizes of LD blocks, thereby, it should significantly increase the resolution of association mapping by breaking long LD in *G. max* population alone.

To determine if the 1.5 K diverse collection can reflect genetic diversity of the 20 K wild and cultivated accessions in the US Soybean Germplasm Collection, which have been genotyped by SoySNP50K Chip. PCA analysis showed that the 1.5 K accessions spread in major clusters among the 20 K accessions (Fig. 4a). This observation was supported by the presence of the 1.5 K accessions (orange) in almost all major clusters in the neighbor-joining phylogenetic tree of the 20 K accessions (Fig. 4b). Having conducted a pairwise comparison between the two collections of accessions, we identified the accessions in the 1.5 K collection that shared the highest sequence identity with each of the 20 K accessions (Fig. 4c, Table S1). These accessions were treated as potentially genomic "equivalent" accessions in US collection. These can therefore serve as a reference for an inaccessible sequenced accession or may be the genome sequence of the un-sequenced accessions to maximize the use of the genome sequences generated worldwide for soybean research (Table S2).

## Diverse application of the 32mSNPs among 1.5 K soybean genomes

The 32mSNPs among 1.5 K diverse wild and cultivated soybean accessions offers an excellent resource for a variety of post-genomic research. Here we demonstrated that

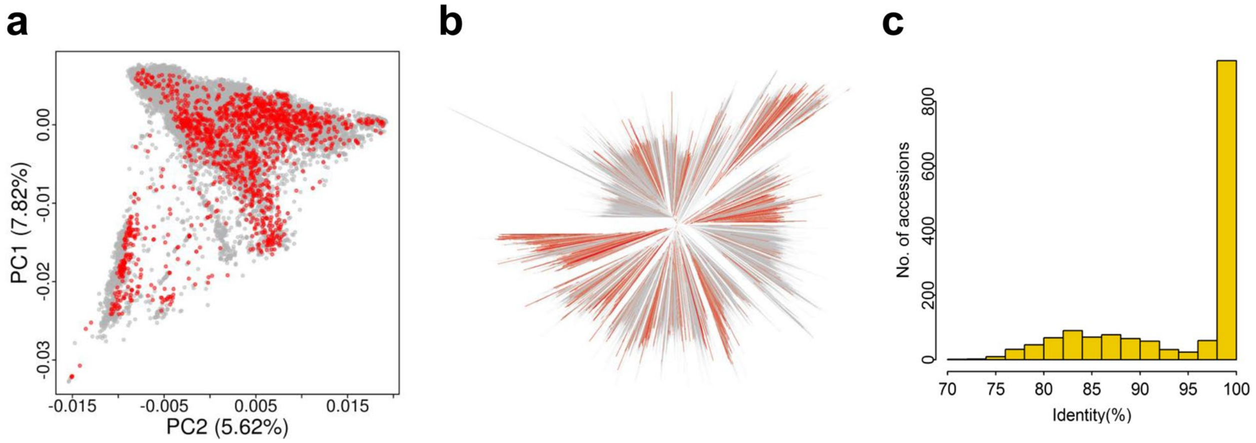


**Fig. 4** Representation of the 1556 accessions in the US 20 K Soybean Collection. **a** Distribution of the whole-genome sequenced 1556 (1.5 K) accessions and the 20,087 (20 K) accessions using principal component analysis. Red dots denote the 1.5 K accessions and gray dots denote the 20 K accessions. **b** A neighbor-joining phylogenetic tree constructed by the 1.5 K accessions and 20 K accessions. Branches in orange denote the 1.5 K whole genomes sequenced accessions and those in gray denote the 20 K accessions. **c** The genomic similarity between the 1.5 K accessions corresponding to the 20 K accessions as determined by comparing the identity of coordinated-matched SNP sites

it can be effectively used for high-resolution association analysis and reverse genetic study.

### *High-resolution association analysis*

SNPs have been one of the most widely used molecular markers in genetic research. The 32mSNPs among 1.5 K diverse soybean accessions will not only be advantageous in high-resolution SNPs-based large and diverse population diversity study but also in the QTL discovery studies for traits of interest using genome-wide association study (GWAS). The determine growth was phenotyped in 642 accessions out of the 1.5 K accessions. GWAS with the 642 accessions revealed two major QTLs with strong associations ($p < 7.10 \times e^8$) on chromosomes 3 and 19 (Fig. 5a). A cluster of SNPs spanning 200 kb (chr19: 45105190-45305190) were significantly associated with determinate growth at the QTL on chr19. These significantly associated SNPs were found to be co-located with a previously reported *Dt1* gene (*Glyma.19G194300*) underlying a stem determination QTL [32, 33] (Fig. 5c). Interestingly, a cluster of strongly associated SNPs at the QTL region on chr 3 nearly coincided with a *Dt1* paralog (*Glyma.03G194700*) (Fig. 5b). The GWAS result strongly supports that the *Dt1* paralog is likely to preserve its function in regulating the determination growth after gene duplication [32]. The successful pinpointing of the determinate growth QTLs to a previously known *Dt1* gene and its paralog using 32mSNPs genotyping data among 1.5 K accessions indicates that these dataset are quite effective for performing high-resolution GWAS of a trait of interest to discover their causative QTL genes and alleles.

### *Reverse genetics to explore variation of trait genes*

Nucleotide substitution is one of the major DNA variants leading to gene function alteration and plant phenotypic variation. The 32mSNPs among 1.5 K diverse soybean accessions with comprehensive structural and functional annotation offers an excellent resource to study gene functions and identify causative alleles for a trait gene using a reverse genetic approach. It also allows us to develop and apply a facile in-silico approach to genotype given genes/features or defined genomic regions of interest for 1.5 K soybean accessions. For example, *FATTY ACID DESATURASE 2* (FAD2) is a key enzyme in fatty acid biosynthetic pathways and plays an important role in regulating fatty acid profile in soybean seeds. This FAD2 enzymes are encoded by seven *FAD2* genes in soybean [34]. We identified a total of 340 SNPs in *FAD2* genes. Out of the 340 SNPs, 53 non-synonymous SNPs, 2 nonsense SNPs, and 2 SNPs at splicing sites were identified (Table 2). Four SNPs (S86F, M126V, P137R, and I143T) were previously demonstrated to be highly correlated with high oleic acid content [35, 36]. We identified three (S86F, M126V, P137R) of the four SNPs, indicating a high diversity and coverage of the SNPs among 1.5 K collection. Of all these variants, the missense mutation, P137R in *FAD2-1B* in PI283327 (Pingtung Pearl, maturity group V) is a critical allele used for breeding

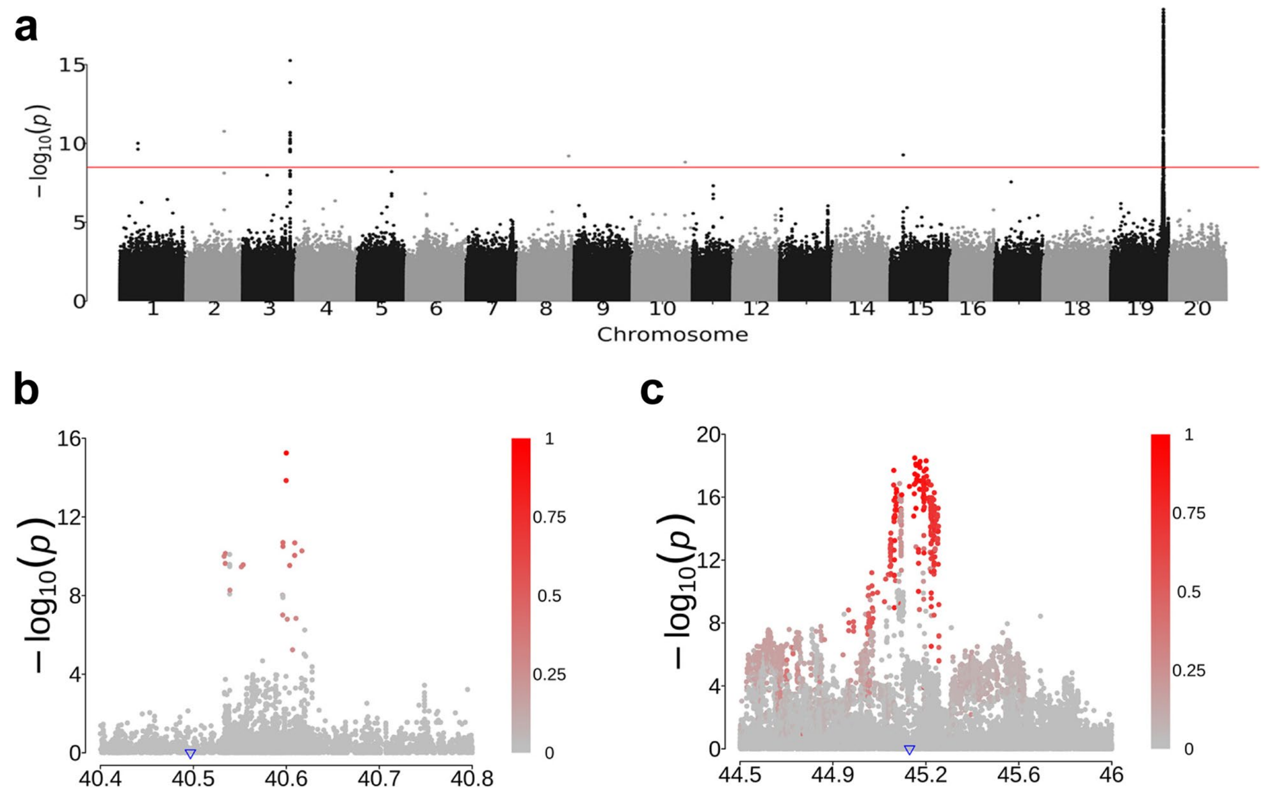


**Fig. 5** Two associations for stem determinate using the 32mSNPs. **a** A Manhattan plot illustrating two major QTLs associated with stem determinate on chromosomes 3 and 19. **b** and **c** Zoomed-in Manhattan plot for the association regions on chr3 and chr19, respectively. Blue triangles denote the physical positions of the *Dt1* and its homologous gene. The color intensity of each SNP indicates its $r^2$ value with the peak association SNP. The color scale was shown beside the panel

**Table 2** Total SNPs identified in *FAD2* family genes using the 32mSNPs

| | Total | FAD2-1A | FAD2-1B | FAD2-2A | FAD2-2B | FAD2-2C | FAD2-2D | FAD2-2E |
|---|---|---|---|---|---|---|---|---|
| Total Genic SNPs | 340 | 16 | 60 | 22 | 23 | 144 | 67 | 8 |
| CDS[a] | 91 | 7 | 19 | 19 | 9 | 20 | 8 | 9 |
| Non-Synonymous | 53 | 6 | 8 | 13 | 3 | 11 | 5 | 7 |
| Synonymous | 38 | 1 | 11 | 6 | 6 | 9 | 3 | 2 |
| Splice Site | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| Premature Termination Codon | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| Five Prime UTR[b] | 55 | 5 | 25 | 0 | 10 | 10 | 5 | 0 |
| Three Prime UTR | 11 | 1 | 3 | 0 | 3 | 4 | 0 | 0 |
| Intron | 191 | 3 | 13 | 4 | 1 | 116 | 14 | 0 |

[a] Coding sequence

[b] Untranslated region

high oleic acid [35]. Remarkably, despite rare alleles, we also successfully identified another accession (PI506933, Kouiku at maturity group IV) carrying the same P137R allele (Table 3). This result suggests the importance of keeping all identified SNPs, including rare alleles among 1.5 K collection for further studies. These altogether demonstrated that the collection of highly diverse 1.5 K soybean accessions along with in-silico genotyping offers a highly effective approach for discovering novel or known germplasm lines containing previously identified or new trait gene mutations/alleles.

**Table 3** Accessions carrying the DNA variant (reference allele C to alternative allele G) in *FAD2-1B* that results in amino acid change at P137R

| PI | Common Name | Species | MG | P137R |
|---|---|---|---|---|
| PI518671 | Williams82 | *G. max* | III | C |
| PI283327 | Pingtung Pearl | *G. max* | V | G |
| PI506933 | Kouiku 1 | *G. max* | IV | G |

## Discussion

The economic importance of soybean in agriculture and food industry has promoted the whole-genome sequencing of thousands of cultivated and wild soybean accessions globally. To facilitate applying the huge amount of sequencing data into soybean research and improvement, this study consolidated the large quantity of genome sequencing data generated in independent studies into one dataset, and made them publicly available in a user-friendly format and demonstrated their versatile applications [37]. It has been long acknowledged that soybean has a vast amount of genetic diversity [31]. The increased size and diversity of the consolidated population should greatly enhance statistic power for post-genomic population studies such as GWAS to identify trait genes and causative alleles. However, analysis of such massive genome sequences requires an investment of tremendous time and effort, the proficient bioinformatics skills, and high-performance computing. These have been limiting the use of the valuable genome sequencing data in post-genomic research. Identification and comprehensive annotation of the32 million SNPs among the 1.5 K soybean accessions using the same bioinformatic pipelines and parameters will greatly enhance the translational genomic research.

The collection of 1.5 K accessions include many soybean germplasm lines widely used in soybean genetic studies and breeding programs such as parental lines of Nested Association Mapping and many RIL mapping populations, important landrace used in the soybean breeding, and elite cultivars [18, 38–41]. For example, Lee (PI548656), Essex (PI548667) and Harosoy (PI548573) that have served as important breeding materials in Southern and Northern US breeding programs are part of the collection. Thus, the collection of accessions should contain genetic diversity important to current soybean breeding and genetic studies. Incorporation of 91 representative wild accessions in the collection can enriched the gene pools for new allele discovery and evolutionary research [12, 17, 18]. The genotyping data at 32mSNP positions among 1500 diverse soybean accessions with an average density of 30 SNPs/Kb genome sequence enable detailed genomic analyses at a single nucleotide resolution. We observed an average of 12 non-synonymous SNPs per gene indicating that the 1.5 K accessions is a rich resource of gene mutants and can be useful for gene function discovery.

The usage of the 32mSNPs genotyping data in the 1.5 K soybean germplasm population may beyond what is presented here (Fig. 6). For example, the 1.5 K covers a significant amount of genetic diversity in the US collection of 20 K accessions. Therefore, imputing the Soy50KSNP chip data based on 20 K accessions with the 32mSNPs among 1.5 K accessions may enable in-depth exploration of the entire USDA Soybean Collection [42]. The 32mSNPs may also contribute to further analysis of genome patterns such as structural variation [43], selective sweeps, and deleterious mutations in soybean [14, 16, 17] or comparative studies with a wider breadth of species. Combining our detailed analysis with other

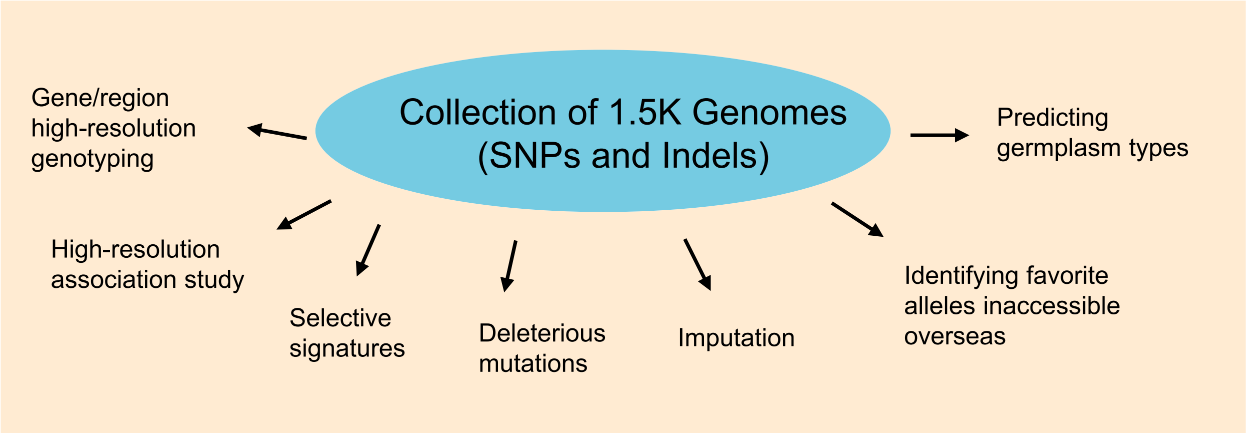

Collection of 1.5K Genomes (SNPs and Indels)
Gene/region high-resolution genotyping
Predicting germplasm types
High-resolution association study
Selective signatures
Deleterious mutations
Imputation
Identifying favorite alleles inaccessible overseas

**Fig. 6** Versatile Usages of the 32mSNPs in the collection of 1.5 K genomes

studies such as haplotype map-based GmHapMap [19] and other types of variation such as structural variation in soybean accessions [16] could be more powerful to address a broad range of basic and applied research questions. To leverage full potential of the SNP data among 1.5 K accessions, the dataset with its detailed annotation information was made public at SoyBase (https://soybase.org) and Ag Data Commons (https://doi.org/10.15482/USDA.ADC/1519167) for extensive uses. Identification of genomic "equivalent"/closely related accessions between 1.5 K and 20 K accessions of US soybean collection offers an opportunity to maximize the use of genomic data, and allowsto access useful variation existing in the accessions present overseas and are inaccessible due to international germplasm exchange policy barriers. In addition, the phylogenetic distances as indicated in the phylogenetic tree may provide guidance in appropriate selection of representative parents to develop population for genetic studies or breeding [44]. Despite an estimated average 2.63% false positive/negative rate [12] and the dramatic drop in the number of high-quality SNPs after MAF-based filtration (0.01 and 0.05), it is advisable to retaining rare alleles in the collection. This is useful to identify not only rare but valuable alleles or accessions as exampled by the identification of two P137R-carrying accessions (MAF = 0.12%). Therefore, the 32 million SNPs among the 1.5 K diverse soybean accessions may be an indispensable rich source for mining valuable trait genes and their variants. Recently, additional genome sequences were released for examining the pan-genome and domestication-associated variation [16, 17, 45]. With the genome sequencing analysis pipeline developed in the study, it could serve a core resource for continuous integration of those genome sequencing data newly available in the public to maximize application of the huge amount of genome sequencing data into soybean research and product development.

In summary, this study provides a user-friendly resource of soybean genomic variants at a single nucleotide-resolution and examples for its versatile usages such as trait and gene variation discovery through genetic association studies, thereby aiding soybean fundamental research and crop improvement (Fig. 6).

## Materials and methods

### Data collection and genome sequencing

Genome sequencing reads of all *G. soja* and *G. max* accessions that were generated and made available in previous studies until 2020 [12–15] were retrieved from the Short Read Archive (SRA) database at NCBI (www.ncbi.nlm.nih.gov) using the fastq-dump sra toolkit (https://github.com/ncbi/sra-tools). The sequencing reads of 91 representative *G. soja* accessions (Table S1) [4] were generated using the Illumina HiSeq2000 sequencer. All sequencing data were combined for further analysis. Based on PI (Plant Introduction) information, we identified that sequencing of 56 accessions were performed twice and one was sequenced thrice from different laboratories (Table S1). Considering the possible genetic variation in the duplicated accessions from different laboratories, we retained all the multiple sequences of these accessions in our analysis but accounted different sequences corresponding to one of these accession as individual accessions. The information on accession ID or name, species, maturity group, origin, and geographical origin was retrieved from the Germplasm Resources Information Network (GRIN, https://www.ars-grin.gov/) and listed in Table S1. The Williams 82 soybean reference genome sequence was downloaded from the Phytozome v12 (https://phytozome.jgi.doe.gov/pz/portal.html). The NCBI's dbSNP Data were downloaded at https://www.ncbi.nlm.nih.gov/SNP. The SoySNP50K iSelect Bead Chip for 20,087 (20 K) soybean accessions in the US Soybean Collection were downloaded from the SoyBase (https://soybase.org).

### SNP calling and annotation

The raw sequencing reads were aligned with soybean reference genome (*G. max* cv. Williams 82.a2 v1) available at Phytozome v12 [3, 46] using burrows-wheeler aligner (BWA) (version: 0.7.17-r1188) [47]. Picard tool was used to add reads group, reorder and sort the reads, and mark the duplicated ones (version: 2.9.2, https://broadinstitute.github.io/picard/). SNPs were called using UnifiedGenotyper function of Genome Analysis Toolkit (GATK) (Version: 3.4) with default parameters [48]. SNPs in each sample were filtered as follows: a minimal coverage of 20%, read depth ≥ 5 reads, SNPs quality score of at least 50 and, maximum of 2 SNPs across 10-bp window. Each sample was genotyped individually at each of the SNP sites passing the filtering criteria. SNP sites for the 1.5 K collection were merged and further filtered based on minor allele frequency > 0.001 and missing rate < 0.5. We used the relaxed filtering criteria to retain rare alleles for downstream research. The genome-wide distribution of these high quality SNPs was illustrated using Circos [49]. A custom Perl script was developed to annotate SNPs based on the coordinates of features (intron, exon, UTR, splicing site, *etc*) of all 56,044 gene models from *G. max* cv. Wm82.a2.v1 reference genome [3, 46]. The effects of non-synonymous SNPs on gene functions were predicted by Protein Variation Effect Analyzer (PROVEAN version: 1.1.5) [29].

### SNP data comparison

SoySNP50K iSelect Bead Chip based genotyping results for 20,087 (20 K) accessions of the US Soybean Collection were downloaded from the SoyBase (https://soybase.org) [4]. Using custom Perl scripts, the SNP coordinates obtained from both SoySNP50K iSelect Bead Chip and our 32mSNPs were identified and used to compare the identity rate of accessions genotyped by these two technology platforms. Coordinates with either no call or N were not included in the comparison. The percent identity of SNP coordinates among accessions was calculated based on the ratio of Sum of Identical SNPs and All Coordinates Compared Count.

The 15,623,492 registered soybean SNP coordinates were downloaded from the NCBI's dbSNP Database (https://www.ncbi.nlm.nih.gov/SNP) and compared with the 32mSNPs from 1.5 K genomes. The coordinates of 32mSNPs that did not correspond with those in dbSNP database were regarded as newly identified SNPs.

### PCA analysis

Principal component analysis (PCA) was used to analyze population structure and relation between re-sequenced 1556 accessions and the 20,087 accessions in the GRIN database using filtered 32mSNPs and SoySNP50K SNPs, respectively. PCA was conducted using the SNPRelate package [50]. An identity-by-state (IBS)-based neighbor-joining (NJ) phylogenetic tree was constructed for all accessions using an R package SNPRelate [50]. The phylogenetic tree was visualized using *ape* packages [51]. All SNPs used in this analysis were filtered using *snpdsLDpruning* function of SNPRelate based on ld. Threshold $\geq$0.5, missing rate $\leq$ 0.2. LD decay was analyzed using PopLDdecay with -MaxDist of 1000 and MAF $\geq$0.05 [52].

### Genome-wide association study

The phenotypic data for stem determinate was downloaded from the GRIN database. SNP data were filtered to keep only biallelic SNP and minor allele frequency $\geq$ 0.01. GWAS was performed using a linear mixed model in GAPIT [53, 54] with Model.selection = T. Kinship was calculated using the default setting in GAPIT and included in the analysis. The threshold for significant association was determined with $p < 0.05/$ total amount of SNPs used in the analysis. The regional LD heatmap was generated using the LDheatmap package [55].

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08326-w.

**Additional file 1: Table S1.** Summary for the collection of 1.5 K accessions: For each accession, the columns showed its common name, PI, maturity group (MG), origin and geographic location (latitude and longitude), coverage of its genome sequencing, percent aligned to the reference genome sequence, total number of reads (Total), percentage of identity in comparison with its own 50 K SNP genotype, the PI with the highest Percentage of Identity (PID) in comparison with 50 K SNP genotypes of all accessions including itself and not including itself. **Table S2.** Best match of each of the 20 K accessions to the 1.5 K accessions: The best match accession and the percentage of identity is listed for each accession in US collection of 20 K accession.

**Acknowledgments**

The authors would like to acknowledge Rick Meyer for his critical technical support in computationally data processing and analysis, Drs. Gunvant Patil, Alice Kujur, Ying Li and ZhenZhen Qiao for helps in drafting and revising the manuscript, Drs. Anne Brown, Steven Cannon, Nelson Rex to release the data on SoyBase, and many researchers for deposing the genome sequencing data in the public to make the research possible, and Dr. Dilip Shah for his kind help with this project.

#### Disclaimer note

Names are necessary to report factually on available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by USDA implies no approval of the product to the exclusion of others that may also be suitable. USDA is an equal opportunity provider and employer.

#### Authors' contributions

YQA concepted the study, participated in data analysis and drafting the manuscript. HZ, HJ and ZH drafted manuscript and participated in data analysis; and QS generated genome sequencing and participated in revising the manuscript. The authors read and approved the final manuscript.

**Funding**

The research is funded by the United Soybean Board (USB #2020-162-0202), USDA-ARS (Project #: 5070-21000-042-00-D) and Bayer AG (Appl. No: 2019-01-058).

#### Availability of data and materials

The dataset supporting the collusions of this article is available in the SoyBase (https://soybase.org) and Ag Data Commons (https://doi.org/10.15482/USDA.ADC/1519167).

## Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

**Author details**

[1]Donald Danforth Plant Science Center, St Louis, MO 63132, USA. [2]US Department of Agriculture, Agricultural Research Service, Soybean Genomics and Improvement Laboratory, Beltsville, MD 20705, USA. [3]US Department of Agriculture, Agricultural Research Service, Midwest Area, Plant Genetics Research Unit, 975 N Warson Rd, St. Louis, MO 63132, USA.

Received: 17 July 2021 Accepted: 20 January 2022
Published online: 31 March 2022

## References

1. USDA-OCE. US Department of Agriculture - USDA Agricultural Projections to 2026 February. 2017. https://www.usda.gov/oce/commodity/projections/.
2. Tilman D, Balzer C, Hill J, Befort BL. Global food demand and the sustainable intensification of agriculture. P Natl Acad Sci USA. 2011;108(50):20260–4.
3. Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463(7278):178–83.
4. Song QJ, Hyten DL, Jia GF, Quigley CV, Fickus EW, Nelson RL, et al. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One. 2013;8(1):e54985.
5. Song QJ, Hyten DL, Jia GF, Quigley CV, Fickus EW, Nelson RL, et al. Fingerprinting soybean germplasm and its utility in genomic research. G3-Genes Genom Genet. 2015;5(10):1999–2006.
6. Leamy LJ, Zhang HY, Li CB, Chen CY, Song BH. A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). BMC Genomics. 2017;18:18.
7. Zhang HY, Song QJ, Griffin JD, Song BH. Genetic architecture of wild soybean (*Glycine soja*) response to soybean cyst nematode (*Heterodera glycines*). Mol Gen Genomics. 2017;292(6):1257–65.
8. Bandillo NB, Lorenz AJ, Graef GL, Arquin D, Hyten DL, Nelson RL, et al. Genome-wide association mapping of qualitatively inherited traits in a germplasm collection. Plant Genome. 2017;10(2). https://doi.org/10.3835/plantgenome2016.3806.0054.
9. Goettel W, Ramirez M, Upchurch RG, An YQC. Identification and characterization of large DNA deletions affecting oil quality traits in soybean seeds through transcriptome sequencing analysis. Theor Appl Genet. 2016;129(8):1577–93.
10. Goettel W, Xia E, Upchurch R, Wang ML, Chen PY, An YQC. Identification and characterization of transcript polymorphisms in soybean lines varying in oil composition and content. BMC Genomics. 2014;15.
11. Wolfgang G, Charles YQ. Genetic separation of southern and northern soybean breeding programs in North America and their associated allelic variation at four maturity loci. Mol Breed. 2017;37(1).
12. Lam HM, Xu X, Liu X, Chen WB, Yang GH, Wong FL, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 2010;42(12):1053–9.
13. Valliyodan B, Qiu D, Patil G, Zeng P, Huang JY, Dai L, et al. Landscape of genomic diversity and trait discovery in soybean. Sci Rep-Uk. 2016;6:23598.
14. Zhou ZK, Jiang Y, Wang Z, Gou ZH, Lyu J, Li WY, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33(4):408–14.
15. Fang C, Ma YM, Wu SW, Liu Z, Wang Z, Yang R, et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol. 2017;18:161.
16. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-genome of wild and cultivated soybeans. Cell. 2020;182(1):162-+.
17. Kim M-S, Lozano R, Kim JH, Bae DN, Kim S-T, Park J-H, et al. The patterns of deleterious mutations during the domestication of soybean. Nat Commun. 2021;12:97.
18. Zhang HY, Goettel W, Song QJ, Jiang H, Hu ZB, Wang ML, et al. Selection of GmSWEET39 for oil and protein improvement in soybean. PLoS Genet. 2020;16(11).
19. Torkamaneh D, Laroche J, Valliyodan B, O'Donoughue L, Cober E, Rajcan I, et al. Soybean (*Glycine max*) haplotype map (GmHapMap): a universal resource for soybean translational and functional genomics. Plant Biotechnol J. 2020. https://doi.org/10.1111/pbi.13466.
20. Jeong SC, Moon JK, Park SK, Kim MS, Lee K, Lee SR, et al. Genetic diversity patterns and domestication origin of soybean. Theor Appl Genet. 2019;132(4):1179–93.
21. Lee GA, Crawford GW, Liu L, Sasaki Y, Chen X. Archaeological soybean (Glycine max) in East Asia: does size matter? PLoS One. 2011;6(11):e26720.
22. Hyten DL, Song QJ, Zhu YL, Choi IY, Nelson RL, Costa JM, et al. Impacts of genetic bottlenecks on soybean genome diversity. P Natl Acad Sci USA. 2006;103(45):16666–71.
23. Kofsky J, Zhang HY, Song BH. The untapped genetic reservoir: the past, current, and future applications of the wild soybean (*Glycine soja*). Front Plant Sci. 2018;9.
24. Li YH, Zhou GY, Ma JX, Jiang WK, Jin LG, Zhang ZH, et al. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol. 2014;32(10):1045–52.
25. Qi XP, Li MW, Xie M, Liu X, Ni M, Shao GH, et al. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat Commun. 2014;5:4340.
26. Zhang HY, Li CY, Davis EL, Wang JS, Griffin JD, Kofsky J, et al. Genome-wide association study of resistance to soybean cyst nematode (*Heterodera glycines*) HG type 2.5.7 in wild soybean (*Glycine soja*). Front Plant Sci. 2016;7(1214).
27. Kim MY, Lee S, Van K, Kim TH, Jeong SC, Choi IY, et al. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. P Natl Acad Sci USA. 2010;107(51):22032–7.
28. USDA-FAS. USDA-Foreign Agricultural Service (FAS): world agricultural production. Table 11. 2020. p. 26.
29. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7(10):e46688.
30. Wang M, Li WZ, Fang C, Xu F, Liu YC, Wang Z, et al. Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat Genet. 2018;50(10):1435–41.
31. Sun LJ, Miao ZY, Cai CM, Zhang DJ, Zhao MX, Wu YY, et al. *GmHs1-1*, encoding a calcineurin-like protein, controls hard-seededness in soybean. Nat Genet. 2015;47(8):939–43.
32. Tian ZX, Wang XB, Lee R, Li YH, Specht JE, Nelson RL, et al. Artificial selection for determinate growth habit in soybean. P Natl Acad Sci USA. 2010;107(19):8563–8.
33. Liu BH, Watanabe S, Uchiyama T, Kong FJ, Kanazawa A, Xia ZJ, et al. The soybean stem growth habit gene *Dt1* is an ortholog of *Arabidopsis TERMINAL FLOWER1*. Plant Physiol. 2010;153(1):198–210.
34. Lakhssassi N, Zhou Z, Liu SM, Colantonio V, AbuGhazaleh A, Meksem K. Characterization of the *FAD2* gene family in soybean reveals the limitations of gel- based TILLING in genes with high copy number. Front Plant Sci. 2017;8:324.
35. Pham AT, Lee JD, Shannon JG, Bilyeu KD. Mutant alleles of *FAD2-1A* and *FAD2-1B* combine to produce soybeans with the high oleic acid seed oil trait. BMC Plant Biol. 2010;10:195.
36. Schlueter JA, Vaslenko-Sanders IF, Deshpande S, Yi J, Siegfried M, Roe BA, et al. The *FAD2* gene family of soybean: insights into the structural and functional divergence of a paleoplyploid genome. Crop Sci. 2007;47(1):S14–26.
37. Kersey PJ. Plant genome sequences: past, present, future. Curr Opin Plant Biol. 2019;48:1–8.
38. Diers BW, Keim P, Fehr WR, Shoemaker RC. Rflp analysis of soybean seed protein and oil content. Theor Appl Genet. 1992;83(5):608–12.
39. Hyten DL, Pantalone VR, Sams CE, Saxton AM, Landau-Ellis D, Stefaniak TR, et al. Seed quality QTL in a prominent soybean population. Theor Appl Genet. 2004;109(3):552–61.
40. Warrington CV, Abdel-Haleem H, Hyten DL, Cregan PB, Orf JH, Killam AS, et al. QTL for seed protein and amino acids in the Benning x Danbaekkong soybean population. Theor Appl Genet. 2015;128(5):839–50.
41. Song QJ, Yan L, Quigley C, Jordan BD, Fickus E, Schroeder S, et al. Genetic characterization of the soybean nested association mapping population. Plant Genome-Us. 2017;10:2.
42. Arouisse B, Korte A, van Eeuwijk F, Kruijer W. Imputation of 3 million SNPs in the Arabidopsis regional mapping population. Plant J. 2020;102(4):872–82.
43. Bayer PE, Valliyodan B, Hu HF, Marsh JI, Yuan YX, Vuong TD, et al. Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. Plant Genome-Us. 2021.
44. Yu JM, Holland JB, McMullen MD, Buckler ES. Genetic design and statistical power of nested association mapping in maize. Genetics. 2008;178(1):539–51.

45. Valliyodan B, Brown AV, Wang J, Patil G, Liu Y, Otyama PI, et al. Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. Sci Data. 2021;8(1):50.
46. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(Database issue):D1178–86.
47. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25(14):1754–60.
48. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
49. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.
50. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28(24):3326–8.
51. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289–90.
52. Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35(10):1786–8.
53. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: genome association and prediction integrated tool. Bioinformatics. 2012;28(18):2397–9.
54. Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38(2):203–8.
55. Shin JH, Blay S, McNeney B, Graham J. LDheatmap: An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Softw. 2006;16:1–10.

## Publisher's Note

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