Genomic dissection of widely planted soybean cultivars leads to a new breeding strategy of crops in the post-genomic era

Xinpeng Qi a,1, Bingjun Jiang a,b,1, Tingting Wu a,1, Shi Sun a,1, Caijie Wang a, Wenwen Song a, Cunxiang Wu a, Wensheng Hou a, Qijian Song c, Hon-Ming Lam b,⇑, Tianfu Han a,⇑

a MARA Key Laboratory of Soybean Biology (Beijing), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
b Centre for Soybean Research of the State Key Laboratory of Agrobiotechnology and School of Life Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong, China
c Soybean Genomics and Improvement Lab, USDA-ARS, Beltsville, MD 20705, USA

1 These authors contributed equally to this work.

⇑ Corresponding authors.
E-mail addresses: honming@cuhk.edu.hk (H.-M. Lam), hantianfu@caas.cn (T. Han).

© 2021 Crop Science Society of China and Institute of Crop Science, CAAS. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Soybean [Glycine max (L.) Merr.] which was domesticated in Asia ~5000 years ago [1] has become a crop of global economic significance, and is listed as the top crop on the American continent in terms of planting area (USDA Foreign Agricultural Service. http://www.fas.usda.gov/). In 2019/2020, the world soybean production was 336.47 Mt, and Brazil, the United States, and Argentina are the three largest soybean producers making up 80.7% of global soybean production (USDA Foreign Agricultural Service, http://www.fas.usda.gov/). In China, which is currently the 4th largest producer, soybean is grown over a large latitudinal range and in diverse ecological environments and cropping systems that extend from the frigid regions of Northern China to the tropical regions of Southern China [2].

As the center of origin, China maintains the highest genetic diversity of soybean germplasm [3]. Modern soybean breeding began in China in 1913 [2], which was approximately 30 years earlier than in North America [4]. The first officially released soybean cultivar, ‘Huangbaozhu’, was reported in 1923 [5], and >2700 soybean cultivars have been released since then. As a result of the century-long intensive breeding efforts, approximately 5% of the released cultivars have been widely planted by farmers. These soybean cultivars are well adapted to the diverse environments and
have stable agronomic performance traits, such as plant architecture [6], seed quality [7], and photo-thermal sensitivity [8].

Since the release of the soybean genome sequence in 2010 [9], a number of studies on soybean genomics, domestication, and evolution have been reported [10–14]. These studies have established that the gene pool of wild soybean is more diverse than that of domesticated soybean. The selection footprint in the soybean genome provides new insights into domestication (from annual wild to cultivated soybean) and subsequent improvements (from landraces to modern soybean cultivars). Furthermore, genome-wide association studies (GWASs) have been undertaken to identify genomic regions associated with domestication-related and agronomically important traits such as plant height, flowering time, and maturity date [12,15–17]. Although such studies have helped us to understand the genetic basis of the increases in soybean yield [18], their practical application is limited.

In this study, we re-sequenced 134 of the most widely planted soybean cultivars released in the last 100 years in China. In order to develop a breeding platform that is highly informative and is ready to be integrated into modern breeding programs, we developed a breeding strategy called “Potalaization”, inspired by the Potala Palace situated at an altitude of ~ 3700 m above sea level in the Lhasa River Valley on the Tibetan Plateau. Briefly, “Potalaization” refers to the selection of a small collection of elite cultivars that offer the highest available agronomic performance while retaining the broadest genetic diversity. Being complementary to “pyramiding”, a strategy in which multiple desired genes are assembled into a single genotype, Potalaization identifies platform cultivars from available elite germplasm resources that are ideal for breeding applications such as generating populations for recurrent selection.

2. Materials and methods

2.1. Soybean cultivar collection

A total of 134 of the most widely planted soybean cultivars were selected for genome re-sequencing, including 80 cultivars from Northeast China, 38 cultivars from the Huang-Huai-Hai River Valley, and 16 cultivars from South China. These widely planted cultivars were officially released from 1923 to 2009, with 94 out of 134 having documented pedigrees (Table S1). They are well adapted to local environments and represent different ecotypes (North-spring sowing, Huang-Huai-Hai-spring sowing, Huang-Huai-Hai-summer sowing, South-spring sowing and South-summer sowing) in China [8,19,20]. Two wild soybean accessions collected from Northeast China (Heilongjiang, China), one wild accession from the Huang-Huai-Hai River Valley (Hebei, China), and one wild accession from South China (Guangxi, China) were also included as references (Table S1).

Soybean seeds were obtained from breeders or from the National Crop Genebank of China, Institute of Crop Sciences, the Chinese Academy of Agricultural Sciences (CAAS), Beijing, China. The seeds were planted in pots at the CAAS campus. A typical individual plant for each cultivar was selected and propagated; DNA was extracted from each cultivar.

2.2. Sequencing, reads mapping and variant calling

Genome sequencing libraries were constructed according to the Illumina manufacturer’s instructions. We generated > 1.26 Tb of data (13.98 billion paired-end reads of 90 bp each) on an Illumina HiSeq2000 sequencer. Raw reads were quality-trimmed using TRIMMOMATIC [21] (ILLUMINACLIP: adapter.seq:2:30:10 TRIMMING:3 SLIDINGWINDOW:4:10 MINLEN:20). Reads from the 302-soybean accessions re-sequencing project [12] were processed using the same parameters. Clean reads were then mapped to the soybean reference genome [9] (version Wm82.a2.v1) using BWA-mem [22] with the default parameters.

SNPs and InDels were called using the Genome Analysis Toolkit (GATK) HaplotypeCaller [23] (–stand_call_conf set to 30.0, –stand_emit_conf set to 10.0, and -glm set to BOTH). The variants were then recalibrated by the Gaussian mixture model, and outliers were discarded. Variants were further filtered using BCTools (version 1.2, QUAL ≥ 50.0, DP ≥ 5.0, QD ≥ 5.0, MQ ≥ 30, MAF ≥ 0.03, Coverage ≥ 90%). InDels longer than 6 bp were discarded. After recalibration using the GATK VariantRecalibrator, the variants were further filtered by MAC ≥ 20. SNP and InDel loci distributed over the 20 soybean chromosomes were displayed using visualization tool Circos [24].

2.3. Phylogenetic analysis

Bi-allelic SNPs identified from the 134 widely planted soybean cultivars and from the 302-soybean accessions re-sequencing project were further filtered by minor allele frequency ≥ 0.05, and missing data frequency ≤ 0.2. Four-fold degenerate SNP sites [25] were extracted (MAC ≥ 100 and AN ≥ 800) and pruned by PLINK [26] (version 1.90b3b) using the default parameters. A maximum-likelihood phylogenetic tree was constructed using PHYLIp [27] (version 3.695, 1000 permutations) based on these SNPs. Tightly-linked SNPs were filtered using the R function snpGdsLDpruning (LD threshold equal to 0.2) in the SNPRelate package [28] (version 1.2.0). Principle component analysis (PCA) was also performed using the SNPRelate package (version 1.2.0).

2.4. Population structures and genomic features of the widely planted soybean cultivars

Population structure analysis was performed using STRUCTURE [29] (version 2.3.4, 500,000 burn-in period, 1,000,000 repeats, K values of 2, 3, 4, and 5). After pruning, the SNPs were used to identify linkage disequilibrium patterns using Haploview [30] and to calculate genetic diversity by VCFtools [31].

Copy number variations (CNVs) were identified by CNV-seq [32] based on pooled reads from cultivars planted in the same geographic region. Genes affected by CNVs were annotated based on the soybean reference genome [9].

2.5. Selective region identification

Whole genome scanning of selective regions was performed by XP-CLR [33] (parameters: gWin = 0.005; snpWin = 100; grid-Size = 100). The genetic map required by XP-CLR was a pseudo-map constructed with reference to the 302 soybean accessions re-sequencing project [12]. In brief, the genetic position of each SNP was assigned linearly along the genetic map from SoyBase [34] under the assumption that recombination between mapped markers is uniform [12].

Selective regions were then identified reference to a prior study [35]. In brief, the soybean genome was divided into 10 kb segments, and the average cross-population composite likelihood ratio (XP-CLR) score was calculated for each segment. Segments in the top 20 percentile of XP-CLR scores were then selected. Selected segments that were separated by no more than one low XP-CLR score segment were merged to form candidate selective regions. The candidate selective regions with the top 10 percentile highest segmental XP-CLR scores were then identified. Selective regions of <30 kb in length were discarded because of the short history of modern soybean breeding.

The windowed Weir and Cockerham fixation index (Fst) was calculated by VCFTools [31] with a window size of 30 kb and step
size of 30 kb. Nucleotide fixation locus was defined as where one sub-population maintains a monotypic genotype while the other exhibits polymorphic genotypes [36]. Genes in selective regions were categorized by WEGO [37] according to their gene ontology (GO) annotation.

2.6. Pedigree analysis

Pedigree analysis was conducted using IBDseq [38] (LOD cutoff of 3.0) on the 94 widely planted cultivars that have documented pedigrees.

2.7. GO and kyoto encyclopedia of genes and genomes (KEGG) enrichment analysis

GO term enrichment analysis was performed on the AgriGO web server [39]. KEGG pathway module enrichment analysis was performed with the R package clusterProfiler [40].

2.8. Phenotypic data collection

The detailed phenotyping method was described in our previous study where we evaluated the changing trend of photothermal sensitivity over the modern breeding history of soybean based on 63 widely planted cultivars [8]. In brief, the experiment was conducted using a completely randomized design. Plants were sown in the spring and summer of 2010 and 2011 in Beijing, China. The number of days from the soybean plant development stage [41] emergence (VE) to the first flower (R1) was recorded as the trait of vegetative period [34].

3. Results

3.1. Population re-sequencing and genomic variation in the widely planted soybean cultivars

A total of 13.98 billion raw reads (90 bp in length) were generated using the Illumina Hiseq2000 platform, resulting in an average of ~8X depth of clean reads for each of the 134 widely planted cultivars. We also sequenced four wild soybean accessions (Glycine soja Siebold and Zucc.) that were used as the out-group for subsequent phylogenetic analyses. For each of the re-sequenced accessions, the clean reads were mapped to the soybean reference genome [9], resulting in an average mapping coverage of 93.81% (Table S1). After quality filtering, variants including 4,163,977 small insertions/deletions (InDels) were identified (Tables S1). After quality filtering, variants including 4,163,977 small insertions/deletions (InDels) were identified (Tables S1). After quality filtering, variants including 4,163,977 small insertions/deletions (InDels) were identified (Tables S1). After quality filtering, variants including 4,163,977 small insertions/deletions (InDels) were identified (Tables S1).

3.2. Uniqueness of the widely planted soybean cultivar collection

We examined the phylogenetic relationships within the 134 widely planted cultivars combined with the 302 previously re-sequenced soybean accessions [12] based on the SNPs identified in the two datasets. The three major sub-populations of wild soybeans, landraces, and improved cultivars identified in this combined analysis were consistent with sub-groups reported in the 302 re-sequenced soybean accessions (Fig. S2). The 134 cultivars sequenced in our study were distributed in the sub-groups of landrace and improved cultivar (Fig. S2), indicating the broad genomic diversity of widely planted cultivars.

PCA based on 6666 four-fold degenerate SNPs showed that no obvious clusters emerged from combinations of any two of the top three PCs (Fig. 1). Individually, the plots based on the PC1 value (upper left in the diagonal) resulted in major peaks that were consistent with the tree branches in the phylogenetic analysis (Fig. S2). The plots based on PC2 (middle in the diagonal) overlapped with the peaks between the widely planted cultivars and the 302 previously re-sequenced soybean accessions, while the plots based on PC3 (lower right in the diagonal) differentiated the widely planted cultivars from the 302 re-sequenced accessions, providing evidence of genomic uniqueness in the collection of widely planted cultivars.

3.3. Population structure and linkage disequilibrium in the widely planted soybean cultivars

The 134 widely planted soybean cultivars were grown in a latitudinal range from 50°N in Heilongjiang to about 21°N in Guangxi. A phylogenetic analysis based solely on this collection showed two major geographically structured clusters that correspond to the Northeast China (NE) sub-population and the Huang-Huai-Hai River Valley and South China (HS) sub-population (Fig. 2A). This result is consistent with the phylogenetic tree and the PCA clusters based on the combined dataset of the 134 widely planted cultivars and the 302 re-sequenced soybean accessions. We checked the origins of 539 ancestors of the 134 widely planted cultivars, and the majority of these ancestors originated from the same geographical regions as their respective descendant cultivars (179 out of 191 ancestors were recorded as “1/2” contributors, Table S3). Therefore, geographically structured sub-population clustering was the major characteristic of the widely planted soybean cultivar collection.

The NE sub-population comprises 75 cultivars planted in Northeast China except for CA086 (“Qihuang 10”), a cultivar from Shandong province in the Huang-Huai-Hai River Valley. The HS sub-population comprises 53 cultivars from the Huang-Huai-Hai River Valley and South China and six cultivars from HS-adjacent NE areas (CA054, CA064, CA071, CA074, and CA077 from Liaoning, and CA079 from Beijing). Analysis of the admixture of the cultivars in the two sub-populations using STRUCTURE [29] also showed the geographically based clustering of the cultivars (Fig. 2B). Increasing the K value resulted in the same major sub-groups (Fig. S3). Not surprisingly, admixture was observed for cultivars that originated from the HS-adjacent NE areas. PCA based on 5889 SNPs identified from the widely planted cultivars showed that the NE sub-population is tightly clustered together, while the HS sub-population is scattered along both the horizontal axis and the vertical axis (Fig. 2D). Nuclearotide diversity calculated in a 10 kb window along each chromosome is substantially higher in the HS sub-population than in the NE sub-population (Table 1). Linkage disequilibrium decayed more rapidly in the HS sub-population than in the NE sub-population (Fig. S4). Collectively, the HS sub-population has a broader genetic background than does the NE sub-population.

To further explore structural variations, we identified 3705 CNVs with an average length of 6.49 kb in cultivars that share the same geographical origins (Table S4). The CNVs were distributed unevenly among the chromosomes and in different chromosomal regions (Figs. S5, S6). Chromosome 18 had the most CNVs (621), while Chromosome 11 had the fewest (36). CNV gains were mostly observed in Chromosomes 14, 15, and 18 in cultivars from the HS sub-population, and in Chromosome 6 in cultivars from the NE sub-population. Gene ontology (GO) analyses of genes affected by CNVs showed enrichment of phosphorylation-related and binding-related functions (Table S5).

3.4. Selective regions between widely planted soybean cultivars from the two sub-populations

To capture the chromosomal regions that have undergone selection between the NE and HS sub-populations, we calculated
the cross-population composite likelihood ratios (XP-CLR) along the soybean genome. A total of 167 genomic regions (those with the top 10% highest XP-CLR scores) were identified as selective regions that differed between the HS and NE sub-populations (HS vs. NE), and 186 selective regions were identified in the reciprocal comparison (NE vs. HS; Figs. 3, S7). These selective regions comprise 2.51% and 2.64% of the total genome length and harbor 1600 and 1251 genes, respectively (Fig. S8). The mean genomic sequence length of the selective regions was 142.99 kb and 134.73 kb, with the largest regions of 900 kb and 580 kb in the HS vs. NE and the NE vs. HS comparisons, respectively. A total of 21 selective regions identified from the HS vs. NE comparison overlapped with 23 selective regions identified from the NE vs. HS comparison (Tables S6, S7). These overlapping selective regions are the regions of the genome that were commonly impacted during breeding, while the unique selective regions from each of the comparisons represent selections that occurred specifically in a single sub-population. A total of 309 non-redundant genomic regions were identified as selective regions that were impacted by geographical origins. These non-redundant genomic regions are 42,790 kb in length (4.5% of the genome sequence), and harbor 2434 annotated genes.

The selective regions identified by the XP-CLR analyses had markedly higher \( F_{st} \) values compared to the whole genome (Fig. S9). When we examined the genetic variations between the two sub-populations by scanning for nucleotide fixation, we found that the selective regions had a higher proportion of fixed SNPs than did the whole genome (Fig. S10). In other words, although the selective regions contain a relatively larger number of polymorphisms, the diversity of these polymorphisms in such regions is actually reduced.

3.5. Genetic variations relating to soybean trait of vegetative period

Soybean plant development stage from VE to R1, described as the trait of “vegetative period” by Soybase [34], was recorded...
Fig. 2. Analyses of population structures and geographical distributions of the widely planted soybean cultivars. (A) Maximum-likelihood phylogenetic tree. (B) Bayesian clustering based on the genomic composition of population structure analyzed by STRUCTURE. (C) Geographical distribution of the widely planted cultivars. Total production during 1995–2011 is indicated by scaled colors according to different Chinese provinces. The blue circles indicate the number of cultivars from the corresponding province. (D) Principal component analysis of widely planted cultivars. PC1/2, the first/second principle component.

Table 1
Statistics of the genomic variations identified in the 134 widely planted soybean cultivars and the four wild soybean accessions.

<table>
<thead>
<tr>
<th>Group</th>
<th>No. of accessions</th>
<th>SNP Total</th>
<th>Exonic</th>
<th>Syn/Nonsyn Diversity</th>
<th>InDel Total</th>
<th>Exonic</th>
<th>FS/NonFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS</td>
<td>54</td>
<td>4,039,080</td>
<td>161,393</td>
<td>71,306/92,878</td>
<td>1.31E-03</td>
<td>465,182</td>
<td>5978</td>
</tr>
<tr>
<td>NE</td>
<td>80</td>
<td>3,580,848</td>
<td>144,682</td>
<td>63,942/83,217</td>
<td>1.04E-03</td>
<td>417,030</td>
<td>5343</td>
</tr>
<tr>
<td>Wild</td>
<td>4</td>
<td>3,239,102</td>
<td>129,110</td>
<td>58,362/72,969</td>
<td>1.30E-03</td>
<td>376,670</td>
<td>4471</td>
</tr>
<tr>
<td>Total</td>
<td>138</td>
<td>4,163,977</td>
<td>165,613</td>
<td>73,098/95,393</td>
<td>1.28E-03</td>
<td>478,985</td>
<td>6191</td>
</tr>
</tbody>
</table>

Diversity was calculated based on the p value. Syn, synonymous substitution; Nonsyn, non-synonymous substitution; FS, frame shift; NonFS, non-frame shift; SNP, single nucleotide polymorphism; InDel, insertion/deletion; HS, Huang-Huai-Hai River Valley and South China; NE, Northeast China.

Fig. 3. Selective sweeps identified between the two sub-populations by whole genome XP-CLR scanning. Upper panel, the NE sub-population was used as the reference population. Lower panel, the HS sub-population was used as the reference population. The XP-CLR scores >200 were marked as 200 to facilitate display.
under short-day (SD) and long-day (LD) conditions in two different seasons for our widely planted soybean collection. All four records showed differences between the groups of cultivars based on the two sub-populations (Tables S8, S9). K-means clustering analysis with \( k \) equal to 2 showed that the vegetative period recorded under spring-planting SD conditions (SP-SD) and summer-planting LD conditions (SU-LD) successfully separated the widely planted cultivars into two groups. This grouping was consistent with the origins of the cultivars in the NE and HS sub-populations, with Spearman’s rho correlation coefficients of 0.83 and 0.89, respectively (Fig. 4A, B). While, the K-means clustering of vegetative period recorded in spring-planting LD conditions (SP-LD) and summer-planting SD conditions (SU-SD) was not consistent with geographical origins, with Spearman’s rho correlation coefficients of 0.68 and 0.51, respectively (Fig. 4C, D). Even though larger genotype-environment interactions and environmental effects were reported for soybean vegetative period [8], the k-means clustering results provide evidence for the genotypic contribution to this trait.

For the purpose of uncovering related genes, we examined all four vegetative period QTL [42] documented at Soybase [34]. These four vegetative period QTL are located in 28 identified selective regions. A total of 121 annotated genes were found in these genomic regions (Tables S6, S7), including genes annotated as having binding and cellular process functions (Fig. S11). We also examined other soybean plant development traits, and eight genes related to flowering time [43–46] are located in selective regions (Tables S6, S7).

3.6. Pedigree analyses of the widely planted soybean cultivars

Of the 134 widely planted cultivars, 94 cultivars have at least one ancestor that is also a widely planted cultivar (Table S1). CA056 (‘Huangbaozhu’), CA025 (‘Mancangjin’) and CA135 (‘Zihua 4’) are the top three most frequently used ancestral cultivars in Northeast China, while CA085 (‘Juxuan 23’), CA084 (‘Xinhuangdou’) and CA083 (‘Yidupingdinghuang’) are the top three most frequently used in the Huang-Huai-Hai River Valley and South China (Fig. S12). The pedigrees of 25 cultivars could be traced back at least 10 generations (Table S1), and four cultivars (CA020, CA048, CA045, and CA076) have 10 or more ancestors that were also re-sequenced in this study. IBDseq [38] was performed to identify the identity-by-descent (IBD) segments in these four cultivars based on their pedigrees. In brief, 110, 134, 115, and 112 IBD segments were identified that cover 26.72%, 45.78%, 26.10%, and 36.05% of the whole genome, respectively (Table S10). A majority of the detected IBD segments are shared by cultivars released in adjacent generations, but only 18 IBD segments were identified from the earliest recorded ancestors to the current widely planted cultivars (Figs. S13, S14).

Fig. 4. K-means clustering of the soybean trait of vegetative period. (A) Spring-planted short-day (SP-SD) conditions. (B) Summer-planted long-day (SU-LD) conditions. (C) Spring-planted long-day (SP-LD) conditions. (D) Summer-planted short-day (SU-SD) conditions. Clusters differentiated by K-means are indicated by red and black. The NE and HS sub-populations are indicated by triangles and circles, respectively.
3.7. Defining a platform panel of widely planted cultivars and the new breeding strategy of “Potalaization”

We sought to identify the most diverse representative cultivars from the 134 widely planted cultivars using an iterative calculation of accumulated genetic diversity based on SNP data. Specifically, the two cultivars with the highest percentage of segregated SNPs among the widely planted cultivar collection were selected as the initial pair in the platform panel, then each of the remaining 132 cultivars was assessed and the cultivars contributing the largest proportion of segregated SNPs would be selectively added to the platform panel. This iterative selection continued until the platform panel captured 95% of total segregated SNPs in the whole collection. Eventually, a total of 14 of the widely planted cultivars were defined as the platform panel (Fig. 5A). Eight of the platform cultivars are from the HS sub-population and six are from the NE sub-population. In addition to their geographical diversity, these platform cultivars were released in different years (Table S11). Our platform panel complements the current Chinese soybean mini-core collection consisting of 299 soybean accessions that were selected based on morphological traits and traditional molecular markers [47].

To document this protocol, we conceived a breeding strategy that we call “Potalaization” for selecting, managing, and exploiting elite crop germplasm in the post-genomic era (Fig. 5B). The whole germplasm collection and the released cultivars are analogous to the high altitude of the Tibetan Plateau. Metaphorically, we recapitulated the impressive breadth of the “Red Hill” in a demonstration case study by iteratively identifying the most genetically diverse cultivars from amongst the elite cultivars. The cultivars identified thus will be ideal as starting materials for gene pyramiding while retaining the highest genetic diversity for future breeding efforts.

4. Discussion

Although there have been large-scale genomic re-sequencing studies reported for soybean, none has focused primarily on identifying the genetic diversity present within the genomes of the most popular cultivars. Here we collected 134 of the most widely planted soybeans from the last hundred years of Chinese soybean breeding and conducted a re-sequencing study. Cultivars in this collection have been widely adopted in soybean production regions and are extensively used in current soybean breeding programs [20], indicating that they have excellent agronomic performance that is appreciated by producers while retaining unique genomic information as demonstrated by our phylogenetic analyses. These widely planted cultivars were divided into two geographically distinct sub-populations, NE and HS. We identified selective regions in each sub-population, and found that the cultivars clustered into two consistent groups based on the vegetative period when grown under SP-SD and SU-LD conditions.

The re-sequencing of other soybean accessions reported to-date, including wild soybean relatives, landraces, and modern cultivars, has helped deepen our understanding of the process of soybean domestication and its subsequent improvements due to continued soybean breeding efforts [10–12,17]. Our analyses of the widely planted cultivars confirmed that geographical origin is a major factor in differentiating elite Chinese soybean cultivars. Specifically, we found that the genetic diversity (calculated PI value) of the HS sub-population was comparable to that previously reported for the landraces (1.31 × 10^{-3} versus 1.40 × 10^{-3}), whereas the PI value for the NE sub-population was comparable to that of the improved cultivars (1.04 × 10^{-3} versus 1.05 × 10^{-3}). The overall diversity was lower than that in the 302 re-sequenced soybean accessions [12]. Linkage disequilibrium decayed more quickly among the 302 soybean accessions than among the widely planted soybean cultivars (Fig. 5A).

We have devised the new breeding strategy of “Potalaization”, which takes advantage of traditional breeding, because the widely planted cultivars we re-sequenced represent the highest levels of achievement in terms of overall agronomic performance, especially in terms of yield and plant morphology [6,48], of Chinese soybean breeding over the last 100 years. Using this approach, we defined a platform panel of cultivars that could be used as parental materials or receptors for advanced cultivar breeding so that their overall superiority will remain and only the desired traits will be targeted. A recent study in wheat [49] describes a strategy evaluating breeding value based on the accumulation of beneficial haplotypes in centromere-spanning linkage disequilibrium blocks. In the “Potalaization” strategy, we used genomic diversity based on the proportion of segregated SNPs in the entire genome. Besides the statistical index differences that are due to different genome features of these two crops, the wheat strategy can be applied to individual cultivars, while the “Potalaization” strategy is optimized for a collection of elite cultivars. In the case of soybean, the National Crop Genebank of China contains the largest number of soybean accessions in the world that represent extensive genetic diversity [4]. The “Potalaization” strategy we propose here would be helpful...
in germplasm manipulation, which will increase the efficiency of advanced soybean breeding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Dr. Lijuan Qiu and Zhaoxiong Liu (Institute of Crop Sciences, Chinese Agricultural Academy of Sciences, China) for providing some of the soybean germplasm used in this study. The authors would also like to thank Jee-Yan Chu for copy-editing the manuscript.

This work was supported by the National Key Research and Development Program of China (2017YFD0101400), China Agriculture Research System (CARS-04), and the Agricultural Science and Technology Innovation Program of CAAS. This work was also supported by a grant from the Hong Kong Research Grants Council Area of Excellence Scheme (AoE/M-403/16) awarded to Hon-Ming Lam.

Accession numbers

The datasets supporting the conclusions given in this article are available in the National Center for Biotechnology Information (NCBI) repository. The short re-sequencing reads were deposited in the Short Read Archive (SRA) under accession number SRP062560.

Appendix A. Supplementary data

Supplementary data for this article can be found online at https://doi.org/10.1016/j.cj.2021.01.001.

References

[12] Tao Lin, Guangtao Zhu, Junhong Zhang, Xiangyang Xu, Qinhui Yu, Zheng Zheng, Zhonghua Zhang, Yaoyuan Luo, Shuai Li, Xiaowuan Xiang, Zejun Huang, Junming Li, Chunzhi Zhang, Taoao Wang, Yuyang Zhang, Aoxue Wang, Yangyang Zhang, Kui Li, Chuan Liu, Guosheng Xiong, Yousao Hao, Xuehua Xue, Andrea Mazzuccato, Mathilde Causse, Zhaiping Fei, James G Giovannoni, Roger T Chetelat, Dani Zamir, Thomas Städler, Jingfu Li, Zhibiao Ye, Yongchen Du, Sanwen Huang, Genetic analyses provide insights into the history of tomato breeding, Nat. Genet. 46 (2014) 1220–1226.
X. Qi, B. Jiang, T. Wu et al.


