Domestication and improvement genes reveal the differences of seed size- and oil-related traits in soybean domestication and improvement

Computational and Structural Biotechnology Journal 20 (2022) 2951–2964

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Domestication and improvement genes reveal the differences of seed size- and oil-related traits in soybean domestication and improvement

Jian-Fang Zuo [a], Muhammad Ikram [a], Jin-Yang Liu [b], Chun-Yu Han [a], Yuan Niu [c], Jim M. Dunwell [d], Yuan-Ming Zhang [a,*]

[a] Crop Information Center, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China
[b] Jiangsu Key Laboratory for Horticultural Crop Genetic Improvement, Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing, China
[c] School of Life Sciences and Food Engineering, Huaiyin Institute of Technology, Huaian, China
[d] School of Agriculture, Policy and Development, University of Reading, Reading, United Kingdom

ARTICLE INFO

Article history:
Received 4 March 2022
Received in revised form 7 June 2022
Accepted 7 June 2022
Available online 13 June 2022

Keywords:
Domestication
Improvement
Soybean
Seed oil content
Seed size
Genome-wide association study

ABSTRACT

To address domestication and improvement studies of soybean seed size- and oil-related traits, a series of domesticated and improved regions, loci, and candidate genes were identified in 286 soybean accessions using domestication and improvement analyses, genome-wide association studies, quantitative trait locus (QTL) mapping and bulked segregant analyses in this study. As a result, 534 candidate domestication regions (CDRs) and 458 candidate improvement regions (CIRs) were identified in this study and integrated with those in five and three previous studies, respectively, to obtain 952 CDRs and 538 CIRs; 1469 loci for soybean seed size- and oil-related traits were identified in this study and integrated with those in Soybase to obtain 433 QTL clusters. The two results were intersected to obtain 245 domestication and 221 improvement loci for the above traits. Around these trait-related domestication and improvement loci, 7 domestication and 7 improvement genes were found to be truly associated with these traits, and 372 candidate domestication and 87 candidate improvement genes were identified using gene expression, SNP variants in genome, miRNA binding, KEGG pathway, DNA methylation, and haplotype analysis. These genes were used to explain the trait changes in domestication and improvement. As a result, the trait changes can be explained by their frequencies of elite haplotypes, base mutations in coding region, and three factors affecting their expression levels. In addition, 56 domestication and 15 improvement genes may be valuable for future soybean breeding. This study can provide useful gene resources for future soybean breeding and molecular biology research.

1. Introduction

Cultivated soybeans were domesticated from wild soybeans (*Glycine soja* Sieb. & Zucc.) in China 5000 years ago [1]. Domestication is one of the most important events in agricultural development [2]. To meet human need, a limited number of elite lines are selected for breeding for the next generation, and many traits are changed during domestication (from wild to landrace soybeans) and improvement (from landrace to bred soybeans) processes [3–5]; these traits include higher yield, reduced seed dispersal and seed dormancy, larger seeds, and higher oil content [4–6]. Due to selection, genetic diversity has been greatly reduced [4], a genetic diversity bottleneck has occurred [7], and the average number of protein-coding genes per accession during domestication and improvement has been significantly reduced [8]. Thus, the identification of both genome-wide genetic diversity and genes contributing to domestication and improvement is essential for breeding elite cultivars [5].

During crop domestication and improvement, some important traits have changed significantly, and these changes may be caused

Abbreviations: QTL, quantitative trait locus; CDRs, candidate domestication regions; CIRs, candidate improvement regions; CDGs, candidate domestication genes; CIGs, candidate improvement genes; PA, palmitic acid; SA, stearic acid; OA, oleic acid; LA, linoleic acid; LNA, linolenic acid; OIL, oil content; SW, seed width; SL, seed length; ST, seed thickness; SLW, seed length to width ratio; SLT, seed length to thickness ratio; SWT, seed width to thickness ratio; 100SW, 100-seed weight; DAF, days after flowering; QTNs, quantitative trait nucleotides; LOD, logarithm of odds; PCD, potential candidate domestication; PCI, potential candidate improvement.

* Corresponding author.
E-mail address: soyzhang@mail.hzau.edu.cn (Y.-M. Zhang).

https://doi.org/10.1016/j.csbj.2022.06.014

by some major genes. For example, *tga1*, *gt1*, *tb1*, and *zfl2* encode key morphological traits associated with the domestication of teosinte (*Zea mays* ssp. *parviglumis*) to maize (*Zea mays* L.); *PROG1* alters rice from prostrate to upright growth, and *sha1* determines shattering in rice; *fw2.2* causes a large increase in tomato fruit weight (Table A.1). Thus, domestication and improvement studies in soybean should be addressed.

In previous soybean domestication and improvement studies, most focused on both diversifying selection and selective sweep and identifying domestication and improvement genes. On one hand, some candidate domestication and improvement regions (CDRs and CIRs) were identified. Li et al. [9] identified 394 CDRs and 306 CIRs using θ_π, Tajima's *D*, and F_{st}; Song et al. [10] identified 620 CDRs and 42 CIRs using Fst; Chung et al. [11] identified 206 CDRs using ROD; Zhou et al. [12] identified 121 CDRs and 109 CIRs using the XP-CLR method; Zhou et al. [13] identified 166 CDRs using Tajima's *D*; Wang et al. [14] identified one selection region on chromosome 15 using π, F_{st}, and XP-EHH. By integrating these CDRs and CIRs with previously reported loci of quantitative traits, some CDRs and CIRs were found to be associated with seed size- and oil-related traits in soybean [12–14]. However, there have been few comprehensive studies on the trait-associated CDRs and CIRs at the whole genome level, indicating the relative lack of a systematic understanding of soybean CDRs and CIRs for important traits.

On the other hand, a series of genes for domestication traits have been cloned and functionally identified (Table A.1). These genes include *SHAT1-5* and *Pdh1* for pod shattering; *GmHs1-1* for seed hardness; *GmTfl* and *Dt2* for stem growth habit; *GmPhyA3*, *GmGIa*, *GmFT2a*, *J*, *Tof11*, *Tof12*, and *GmPRR37* for flowering time; *GmNYFA* [15], *GmZF351* [16], *B1* [17], and *GmOLEO1* [18] for seed oil content; *GmGA20OX* [15], *GmCYP78A72*, *GmCYP78A5*, and *SoyWRKY15a* for seed size; *GmSWEET39* [14] and *GmPDAT* [19] for seed size and oil content (Table A.1). Sedivy et al. [20] also identified a number of candidate genes that may have played important roles in soybean domestication and improvement. With the advances of sequencing technologies and genetic analysis methodologies, genetic diversity analysis in natural populations, along with genome-wide association studies (GWAS) and linkage analysis, is frequently used to identify candidate genes for domestication and improvement traits. For example, Zhou et al. [12] identified two domestication loci for seed oil content; Zhou et al. [13] identified 18, 60, 66, and 10 candidate domestication genes for flowering time, seed development, alkaline-salt tolerance, and seed oil content, respectively; Zhang et al. [21] identified 4 candidate domestication genes and 8 candidate improvement genes for seed oil, protein, fatty acid, and amino acid content. In addition, some candidate domestication and improvement genes were identified using only genetic diversity analysis. For example, Li et al. [9] identified 928 domestication and 1,106 improvement genes; Torkamaneh et al. [21] identified 110 domestication genes; Zhou et al. [12] identified 21 fatty acid biosynthesis domestication genes, 10 of which were consistent with the loci for seed oil content in previous studies. Recently, Turquetti-Moraes et al. [23] integrated publicly available data to discover candidate genes involved in oil biosynthesis and regulation in soybean, e.g., BCCP2 and ACCase, FADs, KAS family proteins, and several transcription factors, and predicted new candidate genes, such as *Glyma.03G213300* and *Glyma.19G160700*. However, candidate domestication genes (CDGs), especially candidate improvement genes (CIGs), in soybean should be addressed owing to the limited number of soybean accessions employed and methodologies with limited efficiency, indicating the lack of comprehensive, in-depth mining of CDGs and CIGs.

To address the above issues, first, we detected the domestication and improvement regions in 286 soybean accessions and integrated them with previously reported domestication and improvement regions in soybean in order to obtain more comprehensive domestication and improvement regions. Then, we detected the genetic loci for seed size- and oil-related traits using GWAS, quantitative trait locus (QTL) mapping, and bulked segregant analysis (BSA) and integrated them with previously reported genetic loci in Soybase. Next, the above two results were integrated to obtain domestication and improvement loci for seed size- and oil-related traits, and around these loci, the CDGs and CIGs for these traits were mined using gene expression, SNP variant, miRNA binding, KEGG pathway, DNA methylation, and haplotype analyses. Finally, these genes were used to explain the differences in seed weight and oil content between wild and landrace soybeans and between landrace and bred soybeans. In addition, we summarized candidate genes available for use in future soybean breeding. This content was summarized in Fig. 1.

2. Materials and methods

2.1. Genetic populations

Genetic populations for GWAS, linkage analysis and BSA are described below.

As described in Zhou et al. [24], a total of 286 soybean accessions, including 14 wild, 153 landrace, and 119 bred soybeans, were obtained from six geographic regions of China, and planted in three-row plots in a completely randomized design at Jiangpu experimental station of Nanjing Agricultural University from 2008 to 2016 (datasets: NJ2008∼NJ2016), and at Wuhan experimental stations of Huazhong Agricultural University in 2014 and 2015 (datasets: WH2014 and WH2015), respectively. The agronomic practices were consistent with those of local production conditions.

519 recombinant inbred lines (RILs) derived from orthogonal (OC, 242) and reciprocal (RC, 277) crosses between two parents LSZZH (P_1) and NN493-1 (P_2), together with their parents, were planted in three-row plots in a completely randomized design at the Jiangpu experimental station of Nanjing Agricultural University in 2015 (NJ2015) and at the Wuhan and Ezhou experimental stations of Huazhong Agricultural University, respectively, in 2014 (WH2014) and 2015 (EZ2015).

Based on the linoleic acid and oil content of RILs across the above-mentioned three environments, 30 RILs with higher seed linoleic acid content and 30 RILs with lower seed linoleic acid content were obtained from 242 OC RILs and 277 RC RILs, namely OC_LA_H, OC_LA_L, RC_LA_H, and RC_LA_L; 30 RILs with higher seed oil content and 30 RILs with lower seed oil content were obtained from 242 OC RILs and 277 RC RILs, namely OC_OIL_H, OC_OIL_L, RC_OIL_H, and RC_OIL_L. These materials were planted at Wuhan experimental station of Huazhong Agricultural University in 2018.

2.2. Phenotypic measurements for traits related to seed size and oil content

In all the above-mentioned three genetic populations, each accession or RIL was planted in a three-row plot in a randomized complete block design, each plot was 1.5 m wide and 2 m long, and approximately 15 plants were planted in each row. Five plants in the middle row for each line were harvested, and the seeds were prepared for the measurement of 13 traits, including six seed oil-related traits [13]: palmitic acid (PA), stearic acid (SA), oleic acid (OA), linoleic acid (LA), linolenic acid (LNA), and oil content (OIL); and seven seed size-related traits [25,26]: seed width (SW), seed length (SL), seed thickness (ST), seed length to width

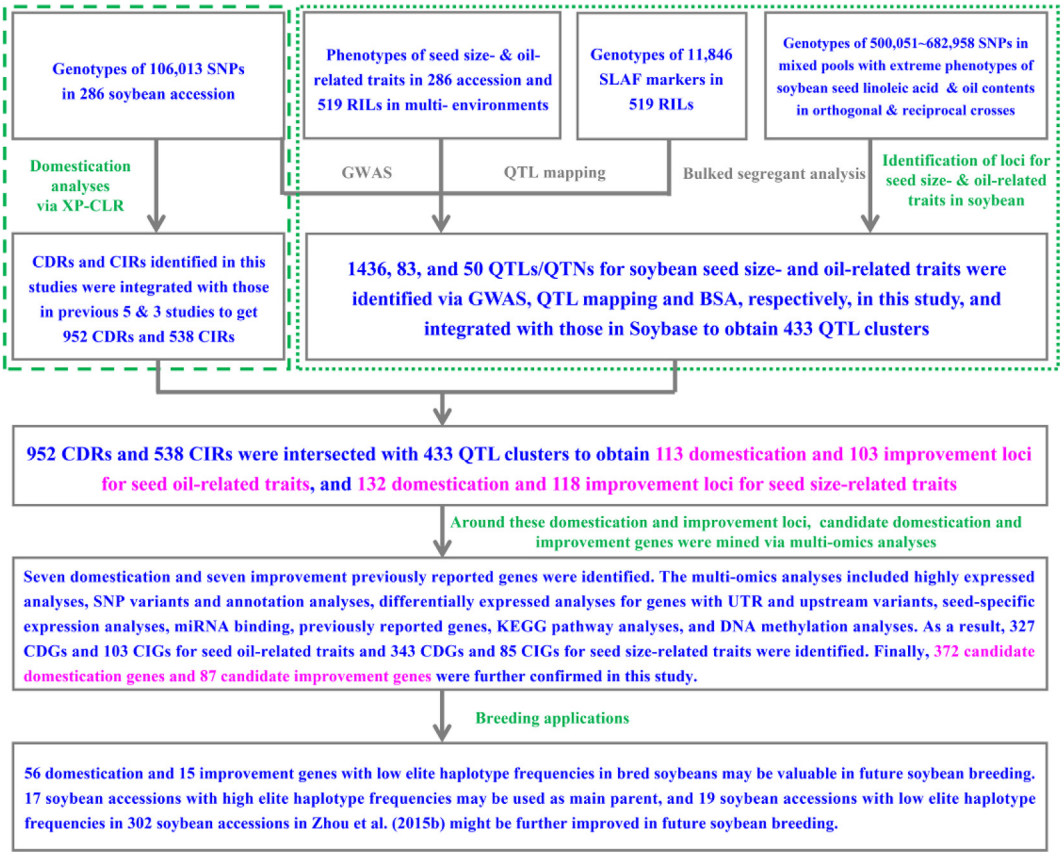

Fig. 1. The basic procedures in this study.

ratio (SLW), seed length to thickness ratio (SLT), seed width to thickness ratio (SWT), and 100-seed weight (100SW).

The phenotypes for seed size-related traits of ~250 soybean accessions in 2008~2010 and 2014~2015 were described by Niu et al. [25] and Ikram et al. [26], respectively, while the phenotypes for seed oil-related traits of 286 soybean accessions in 2011, 2012, 2014, and 2016 were described by Liu et al. [19], Zhou et al. [13], and Liu et al. [31]. The phenotypes for seed oil-related traits of 519 RILs in 2014 and 2015 were described by Zuo et al. [27]. Other phenotypic datasets in this study were new, and all the phenotypic datasets in this study were listed in Table A.2.

2.3. SNP genotyping

DNA extraction Young healthy leaves from a single plant of each accession or RIL were collected, frozen in liquid nitrogen, and used to extract DNA. Young healthy leaves from 30 RILs with extreme high (low) LA (OIL) phenotypes were equally mixed as an extreme pool to extract DNA. Total genomic DNA was extracted from each sample using the cetyltrimethylammonium bromide (CTAB) method [28].

Genotypes of SNP markers for 286 soybean accessions Through resequencing of 286 soybean accessions using RAD-seq approach, a total of 106,013 high-quality SNPs were obtained, which have been described in our previous study [24].

Genotypes of SNP markers for 519 RILs The 519 RILs were genotyped using SLAF-seq method, a total of 11,846 SLAF markers were obtained, and the detailed information was described in our previous study [27].

The resequencing and SNP calling in eight DNA pools and two parents For eight DNA pools described in "2.1 Genetic populations" and two parents, at least 6 μg of genomic DNA from each pool was used to construct a sequencing library following the manufacturer's instructions (Illumina Inc.). Paired-end sequencing libraries with an insert size of approximately 300 bp were sequenced on an Illumina HiSeq 2000 sequencer at Anoroad Gene Technology Company.

The genome of the soybean cultivar Williams 82 was used as a reference genome, and short sequences obtained from the second-generation high-throughput sequencing were compared with the reference genome using software BWA (version: 0.6.1-r104) (https://sourceforge.net/projects/bio-bwa/) [29]. The paired end-to-end resequencing sequences were aligned to the Williams 82 v1.0 reference genome of soybean. Then, SAMtools (Version: 0.1.18) [30] software was used to convert the sorted SAM files into BAM format and sort, and filter out the non-unique alignment sequences. The Picard package (http://sourceforge.net/projects/picard/) was used to remove the duplicates. The SNPs between the reference genome and samples were detected by the Genome Analysis Toolkit (GATK) software [31].

All the genotypic datasets in this study were listed in Table A.2.

2.4. Genome scanning for selective sweeps

We performed a genome scan using the composite likelihood approach of Chen et al. [32] using the software XP-CLR v1.0. Evidence for selection across the genome during domestication and improvement was evaluated in two contrasts: landraces versus wild soybeans for domestication and bred lines versus landraces for improvement. Our scan used a 0.05 cM sliding window with 100 bp steps across the whole genome. To ensure comparability of the composite likelihood score in each window, we fixed the number of SNPs assayed in each window to 250. The command line is as follows:./XPCLR -xpclr input1 input2 input3 output -w1 0.05

250 100 15 -p0 0.95. Then, we used an R script to calculate the mean likelihood score in 100 kb sliding windows with a step size of 10 kb across the genome. Finally, the regions with top 5% value were considered as selected regions, and the adjacent windows with high score were grouped into a single region to represent the effect of a single selective sweep.

To summarize the selection regions during domestication and improvement, the CDRs detected in this and previous studies were integrated, including 21 CDRs in Li et al. [9], 620 CDRs in Song et al. [10], 206 CDRs in Chung et al. [11], 121 CDRs in Zhou et al. [12], and 166 CDRs in Zhou et al. [13]; the CIRs detected in this and previous studies were integrated, including 20 CIRs in Li et al. [9], 42 CIRs in Song et al. [10], and 109 CIRs in Zhou et al. [13]. All the CDRs and CIRs from previous studies were listed in Table A.2.

2.5. Identification of genetic loci for seed oil- and size-related traits in soybean

The mrMLM v4.0.2 software [33], including the six methods of mrMLM [34], FASTmrEMMA [35], pLARmEB [36], ISIS EM-BLASSO [37], FASTmrMLM [38], and pKWmEB [39], was used to detect the association of 106,013 SNPs with seed size- and oil-related traits in multiple environments in 286 soybean accessions, where the kinship matrix K was calculated by the mrMLM software, the Q matrix was calculated by the STRUCTURE 2.3.4 software [40], the number of optimum subgroups was four [13], and the critical value of LOD $\geq$ 3 was used as the criterion for significant QTNs.

The QTL.gCIMapping.GUI v2.1 software [41] of genome-wide composite interval mapping (GCIM) [42] was used to identify the association of 11,846 SNPs with seed oil-related traits in multiple environments in 519 RILs, where the linkage maps constructed by Zuo et al. [27] were adopted in this study, and the critical value of LOD $\geq$ 2.5 was used as the criterion for significant QTLs.

Four pairs of extreme pools for seed linoleic acid and oil content in RILs were used to detect QTLs for the two traits using the BRM software [43], where the BLK was set as 30, and the $u_{\alpha/2}$ was calculated via the BRM software based on the length (cM) of linkage maps in Zuo et al. [27].

2.6. Mining candidate domestication and improvement genes for seed oil- and size-related traits in soybean

2.6.1. Expression analysis of candidate genes

There were three transcriptome datasets available in this study to conduct high expression analyses. The first transcriptome datasets of Jones and Vodkin [44] were downloaded from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42871), and used to detect high expression genes at seed oil accumulation and seed development stages, in which their expression levels at more than one stage from 5∼6 mg to 400∼500 mg were higher than the average at all the seven seed development stages [45]. These stages included whole seed 4 days after flowering (DAF), whole seed 12∼14 DAF, whole seed 22∼24 DAF, whole seed 5∼6 mg in weight, cotyledons 100∼200 mg in weight, cotyledons 400∼500 mg in weight, and dry whole seed. The second transcriptome datasets of Machado et al. [46], which are derived from the re-analyses of the transcriptome datasets at seed 4 DAF, whole seed 12∼14 DAF, whole seed 22∼24 DAF, whole seed 5∼6 mg in weight, and dry whole seed in Jones and Vodkin [44], were download from a user-friendly web interface at https://venanciogroup.uenf.br/resources/. The third transcriptome datasets at seed_10DAF, seed_14DAF, seed_21DAF, seed_25DAF, seed_28DAF, seed_35DAF, and seed_42-DAF in Severin et al. [47] were downloaded from SoyBase (http://soybase.org). If one gene was highly expressed in at least two datasets, this gene was considered to be highly expressed in this study.

The gene expression levels of two wild, two landrace and two bred soybeans at 15, 25, 35, and 55 DAF, described by Niu et al. [48] and Liu et al. [49], were used to determine differential expression genes between wild and landrace soybeans and between landrace and bred soybeans using the DEGseq package [50] with $q \leq 0.001$.

All the transcriptome datasets were listed in Table A.2.

2.6.2. SNP allele frequency and SNP annotation

Using the genotypes of 9,790,744 SNPs in 302 soybeans of Zhou et al. [12], downloaded from Figshare database (https://figshare.com/articles/Soybean_ resequencing Project/1176133), all the SNPs within each candidate gene and its 2 kb upstream were mined, and the significances for the differences of SNP allelic frequencies between wild and landrace soybeans and between landrace and bred soybeans were detected using u test with Bonferroni correction.

The genome sequences (glyma.Wm82.gnm1.FCtY.genome_main.fna.gz) and genome annotation (glyma.Wm82.gnm1.ann1.DvBy.gene_models_main.gff3.gz) were downloaded from Soybase (https://soybase.org/data/public/Glycine_max/) and used to conduct SNP annotation via the SnpEff software [51]. The SNP variants were extracted from the SnpEff-annotated VCF file using a Perl script. We retained the loss of function (LOF) mutations described in Torkamaneh et al. [22] and the variants in 5′UTR, 3′UTR, and upstream of the candidate genes.

2.6.3. Identification of candidate genes related to seed oil- and size-related traits in soybean

The candidate genes for seed size- and oil-related traits were determined using the four steps below.

First, the expression levels of genes at fourteen soybean tissues in Severin et al. [47], downloaded from SoyBase (http://soybase.org), were used to conduct specific expression analysis in seed. These tissues included seed_10DAF, seed_14DAF, seed_21DAF, seed_25DAF, seed_28DAF, seed_35DAF, seed_42DAF, young_leaf, flower, 1_cm_pod, pod_shell_10DAF, pod_shell_10DAF, root, and nodule. Then, the psRNATarget (https://plantgrn.noble.org/psRNATarget/analysis?function=3; [52]) was used to predict miRNA targets with default parameters by comparing the miRNA sequences in *Glycine max*, downloaded from the miRBase (http://www.mirbase.org/ftp.shtml), with the UTR sequence of Williams 82, extracted from the genome sequences and genome annotation. Next, the 1,123 oil-related genes in Zhang et al. [45] were compared with the potential candidate genes in this study to mine candidate genes for oil-related traits. Finally, the remaining genes were used to conduct KEGG analysis to determine the candidate genes for seed oil- and size-related traits using KOBAS (https://kobas.cbi.pku.edu.cn/kobas3).

2.6.4. Validation for SNP variants of candidate genes via 30 soybean genomes

The genomes and genome annotations of four and twenty-six accessions were downloaded from Soybase (https://soybase.org/data/public/Glycine_max/) and Bigdata (https://bigd.big.ac.cn/, Project number: PRJCA002030; [53]), respectively, where the four accessions included W05, PI483463 (wild), Williams 82 (landrace), and ZH13 (cultivar). The genes of Williams 82 were used to create a local BLAST database using the NCBI-BLAST+ (v2.2.31+) software. All the genes in the 29 other genomes were compared with the genes of Williams 82 to search for the best-match genes, which are homologous to the gene of Williams 82, in each of the 29 genomes. The sequences of genes homologous to candidate genes contained within 2 kb upstream were extracted from the 30 genomic

sequences by getfasta function in BEDTools [54], and these sequences were aligned to obtain SNP variants using the MUSCLE software [55].

2.6.5. Haplotype analysis

The common 172 soybean accessions between 302 accessions of Zhou et al. [12] and the publicly available resources on the USDA GRIN database (http://www.ars-grin.gov/) were used to conduct haplotype analysis using the Haploview v4.1 software [56]. The marker genotypes were derived from Zhou et al. [12], while the phenotypes of seed weight and oil content were downloaded from the USDA GRIN database. The missing genotypes were imputed using the Beagle v5.1 software [57]. Multiple comparisons of trait differences among various haplotypes were tested using the *LSD.test* function of *agricolae* package in R.

2.6.6. Differential analysis of epigenetic regulation

4,248 DNA methylation difference regions (DMRs) during domestication and 1,164 DMRs during improvement were downloaded from Shen et al. [58] and used to mine the genes with significant differences of DNA methylation degrees between wild and landrace soybeans, and between landrace and bred soybeans (Table A.2).

2.7. GO enrichment analysis

GO enrichment analysis was conducted via KOBAS (https://kobas.cbi.pku.edu.cn/kobas3), in which the statistical method was hypergeometric test / Fisher's exact test, the false discovery rate correction method was from Benjamini and Hochberg [59], and the significant GO term was determined by the critical value of corrected P-Value at the 0.05 level.

3. Results

3.1. Phenotypic variation of thirteen traits across wild, landrace, and bred soybeans

All the 286 soybean accessions were measured in three to ten environments for seed size-related traits (SL, SW, ST, SLW, SLT, SWT, and 100SW) and oil-related traits (PA, SA, OA, LA, LNA, and OIL) in Nanjing and Wuhan, China (Table A.3). The coefficients of variation for these traits and their best linear unbiased prediction (BLUP) values ranged from 5.01 to 39.41 (%), with a mean of 13.60 %, while their heritabilities ranged from 0.55 to 0.94, with a mean of 0.77 (Table A.4). These indicate the existence of large genetic variation in the association mapping population. In correlation analysis of these traits, some known correlations were observed as well, i.e., OIL and LNA (negatively), OA and LA or LNA (negatively), and LA and LNA (positively) (Fig. A.1). Meanwhile, seed oil-related traits were found to be significantly correlated with seed size-related traits, i.e., OIL and SL (SW, ST, and100SW) (positively), OIL and SLW (SLT and SWT) (negatively), LNA and SL (SW, ST, and100SW) (negatively), and LNA and SLW (SLT and SWT) (positively) (Fig. A.1).

In two-way (environment and evolutionary type) ANOVA, significant differences were observed among wild, landrace, and bred soybeans (*P*-values = <1.00E-300∼4.37E-12), such as *P*-value = 3.55E-20 for OIL (Table A.4). Single degree of freedom analysis showed significant differences between wild and landrace soybeans (*P*-values = 9.42E-47∼1.49E-2), such as 2.67E-7 for OIL, and between landrace and bred soybeans (*P*-values = 1.65E-286∼9.50 E-3), such as 1.24E-11 for OIL (Table A.4; Fig. 2). This indicates the significant changes of the above traits in domestication and improvement (Fig. A.2).

3.2. Domesticated and improved regions in soybean genome

To identify domestication and improvement regions in soybean, the software XP-CLR v1.0 was used to calculate the XP-CLR likelihood ratios of 106,013 SNPs between wild and landrace soybeans and between landrace and bred soybeans. The first 5% of regions with the largest likelihood value were regarded as selection regions, and adjacent windows with common selection regions were merged into one large selection region. As a result, 534 CDRs (Fig. 3a) and 458 CIRs (Fig. 4a) were identified. Among these regions, 156 CDRs overlapped with 157 CIRs, indicating that these regions had undergone selection twice.

Compared with CDRs and CIRs in previous studies, 109, 63, 33, 86, and 8 CDRs in this study were found to be consistent with those in Song et al. [10], Zhou et al. [12], Zhou et al. [13], Chung et al. [11] and Li et al. [9], respectively; 5 and 34 CIRs in this study were found to be consistent with those in Song et al. [10] and Zhou et al. [12], respectively. All the CDRs and CIRs detected in all the related studies were integrated, and a total of 952 CDRs (Fig. 3a; Table A.5) and 538 CIRs (Fig. 4a; Table A.6) were obtained. Among the regions in all the related studies, 169, 46, 30 and 6 CDRs were identified 2, 3, 4, and 5 times, respectively (Table A.5); 43 and 2 CIRs were identified 2 and 3 times, respectively (Table A.6).

We found that the number of CDRs is greater than the number of CIRs in this and previous [9,10,12] studies. This is reasonable because more traits are selected and larger trait differences may exist in the domestication process as compared with those in the improvement process, e.g., seed oil content (BLUP value) in wild, landrace and bred soybeans are 15.55, 17.45, and 18.06%, respectively (Table A.2).

3.3. Mapping QTLs for seed size- and oil-related traits in soybean

To identify additional loci for seed size- and oil-related traits, 106,013 SNPs were used to associate with thirteen traits in three to ten environments and their BLUP values in 286 soybean accessions, 11,846 SLAF markers on linkage maps of Zuo et al. [27] were used to link six seed oil-related traits in three environments in 242 orthogonal and 277 reciprocal cross recombinant inbred lines (RILs), and four pairs of high and low pools of linoleic acid and oil content in the above RIL population were analyzed (Fig. 4c and 5). The results were as follows.

Genome-wide association studies for seed size- and oil-related traits 151, 164, 120, 145, 180, and 83 QTNs were found to be associated with PA, SA, OA, LA, LNA, and OIL in three to four environments and their BLUP values, respectively (Figs. 4b, 5a and A.3; Table A.7). These QTNs were distributed on all the chromosomes (Fig. A.3). The LOD scores were 3.02∼11.77 for PA, 3.00∼24.13 for SA, 3.01∼19.18 for OA, 3.02∼29.07 for LA, 3.00∼16.00 for LNA, and 3.01∼16.08 for OIL, and the corresponding mean r^2 values (%) were 4.12, 3.55, 4.03, 4.29, 3.68, and 5.01, respectively. Among these QTNs, there were 48 large QTNs (r^2 > 10%), e.g., the size of snp25032-associated QTN for LA was 21.28%.

413, 403, 418, 278, 260, 303, and 291 QTNs were found to be associated with SL, SW, ST, SLW, SLT, SWT, and 100SW in seven to ten environments and their BLUP values, respectively (Fig. 3b, 5a, and A.3; Table A.8). These QTNs were distributed on all the chromosomes (Fig. A.3). The LOD scores were 3.00∼34.82 for SL, 3.00∼38.72 for SW, 3.01∼19.64 for ST, 3.01∼12.66 for SLW, 3.00∼17.84 for SLT, 3.00∼14.36 for SWT, and 3.00∼20.93 for 100SW, and the corresponding mean r^2 values (%) were 2.22, 2.03, 2.14, 3.75, 3.81, 3.63, and 2.18, respectively. Among these QTNs, there are 87 large QTNs, e.g., the size of snp60083-associated QTN for SLT was 20.63%.

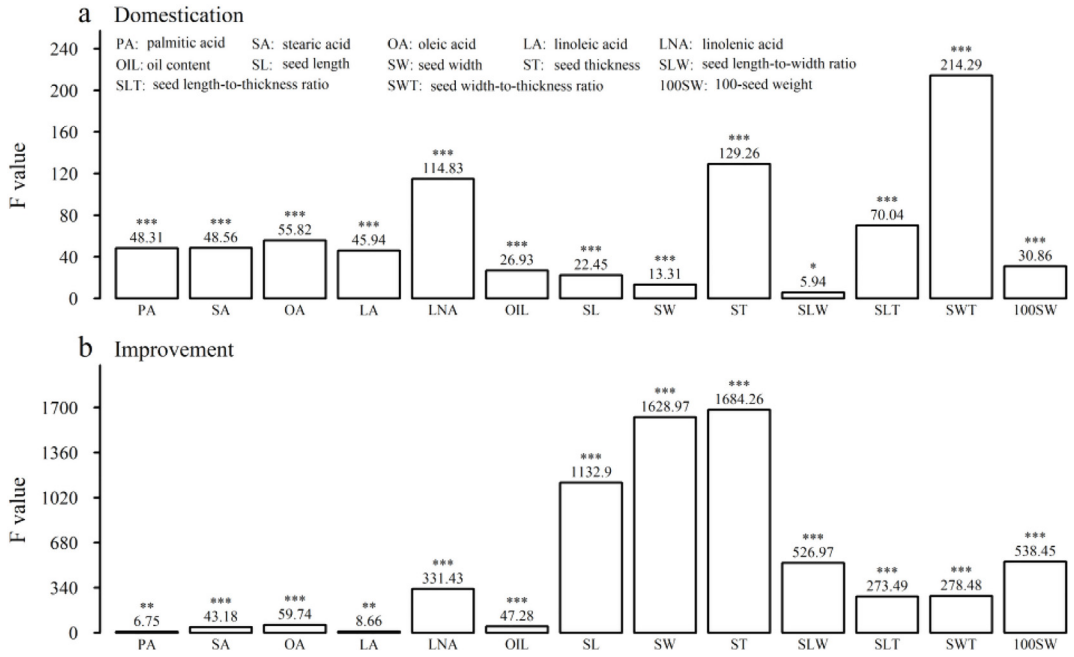

Fig. 2. The significance of the differences for thirteen traits related to seed size and oil content between wild and landrace soybeans (a), and between landrace and bred soybeans (b) using two-way ANOVA. *, ** and ***: significances of the differences at 0.05, 0.01 and 0.001 probability levels, respectively.

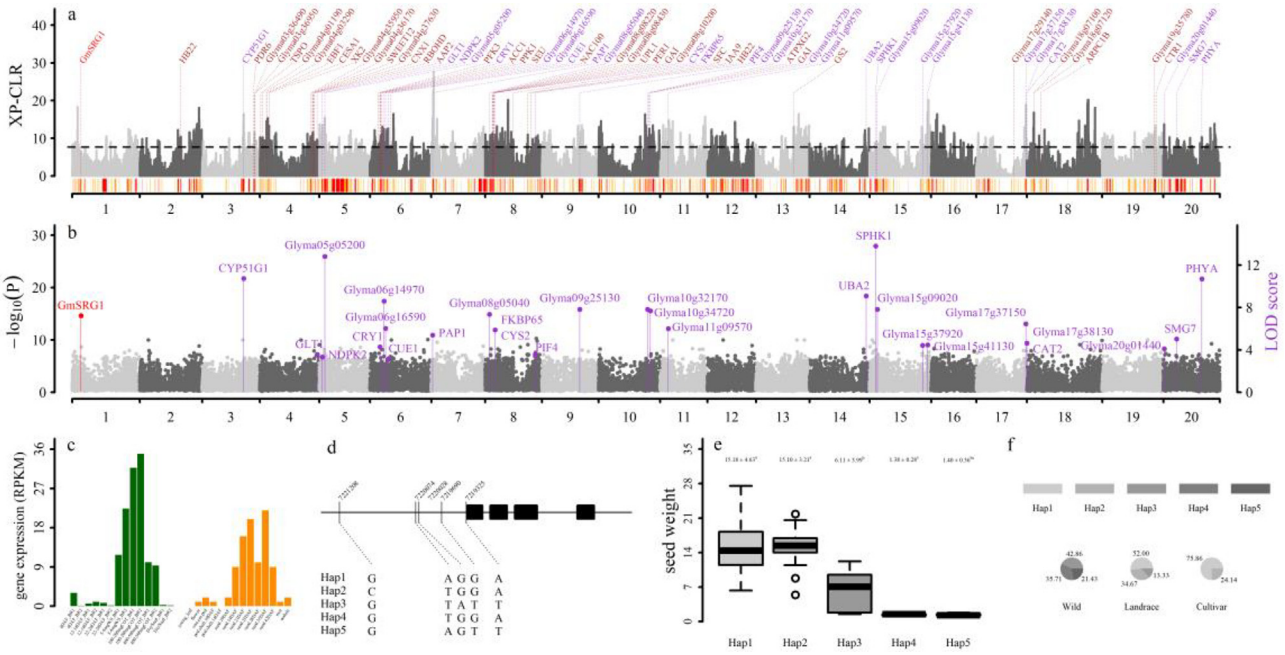

Fig. 3. The domestication genes for soybean seed length (SL). a, The XP-CLR scores between wild and landrace soybeans in 286 soybean accessions. Selection regions repeatedly identified in multiple studies are marked by red color, while the others are marked by yellow color. The candidate domestication genes, detected by genome-wide association studies, and reported QTLs are marked by darkorchid and brown colors, respectively. The gene with red color was further detailed in sub-plots c-f. b, Manhattan plot for SL. c, The expression levels of *GmSRG1* at seven seed development stages and on 14 soybean tissues. The gene expression levels from Jones and Vodkin [44] and Severin et al. [47] were marked with darkgreen and darkorange colors, respectively. d, Five SNPs and the haplotypes of *GmSRG1*. e, The seed weight of five haplotypes of *GmSRG1*, and their significant differences via multiple comparisons. f, The haplotype frequencies of *GmSRG1* in wild, landrace and bred soybeans.

Mapping QTL for seed oil-related traits in three environments Based on the high-density linkage maps of Zuo et al. [27], QTL mapping for six oil-related traits was carried out in three environments and their BLUP values in 242 orthogonal and 277 reciprocal cross RILs using genome-wide composite interval mapping (GCIM), implemented via the software QTL.gCIMapping.GUI v2.1 (Fig. 5b; [39]). As a result, 11, 8, 10, 10, 8, and 8 QTLs were found in orthogonal-cross (OC) RILs to be associated with PA, SA, OA, LA, LNA and OIL, respectively (Table A.9). The 37 QTLs were located on 18 chromosomes excluding chromosomes 1 and 4 (Fig. A.4), their LOD scores ranged from 2.53 to 6.76, and their sizes were 2.98% to 10.71%. In reciprocal-cross (RC) RILs, 10, 3, 8, 12, 19, and 6 QTLs were found to be associated with PA, SA, OA, LA, LNA, and OIL, respectively (Table A.9). The 42 QTLs were located on 18 chro-

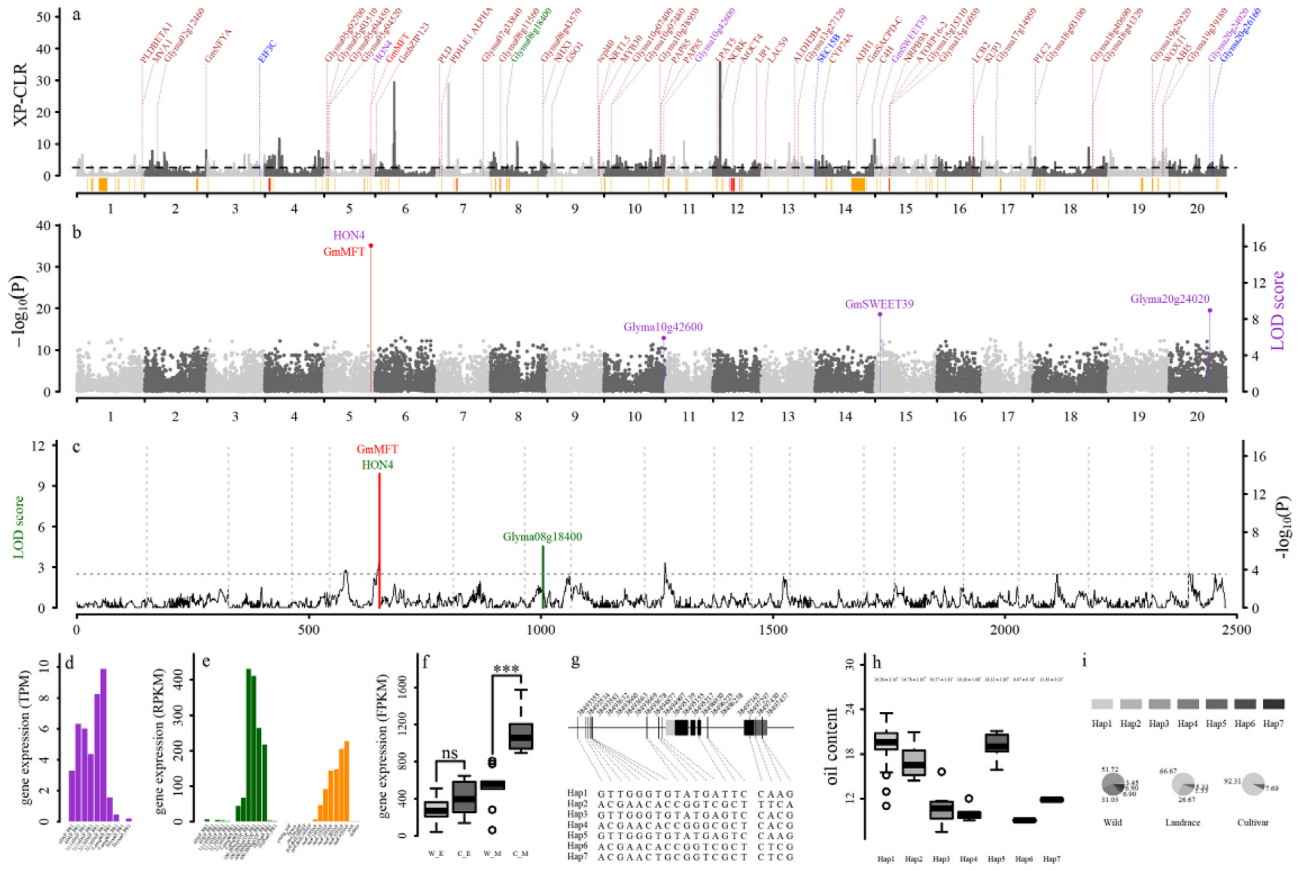

Fig. 4. The improved genes for soybean seed oil content (OIL). a, The XP-CLR scores between landrace and bred soybeans in 286 soybean accessions. Selection regions repeatedly identified in multiple studies are marked by red color, while the others are marked by yellow color. The candidate improvement genes, detected by genome-wide association studies, QTL mapping, and reported QTNs and QTLs, for OIL are marked by darkorchid, darkgreen, blue and brown colors, respectively. The gene with red color was further detailed in sub-plots d-i. b, Manhattan plot for OIL. c, QTL mapping for OIL. d, The expression levels of *GmMTF* at five seed development stages [46]. e, The expression levels of *GmMTF* at seven seed development stages and on 14 soybean tissues. The gene expression levels from Jones and Vodkin [44] and Severin et al. [47] were marked with darkgreen and darkorange colors, respectively. f, The comparison of gene expression levels on *GmMTF* between 10 wild and 10 cultivar soybeans using *t* test. ***: significance at the 0.001 probability level; ns: no significance. g, Twenty SNPs and the haplotypes of *GmMTF*. h, The seed oil content of seven haplotypes of *GmMTF*, and their significant differences via multiple comparisons. i, the haplotype frequencies of *GmMTF* in wild, landrace and bred soybeans.

mosomes excluding chromosomes 17 and 18 (Fig. A.4), their LOD scores ranged from 2.51 to 10.67, and their sizes were 1.68% to 18.30%. In all the RILs, 18, 4, 27, 22, 24, and 12 QTLs were found to be associated with PA, SA, OA, LA, LNA, and OIL, respectively (Table A.9). The 59 QTLs were located on all the chromosomes (Fig. A.4), their LOD scores ranged from 2.51 to 16.22, and their sizes were 1.11%∼13.61%.

Bulked segregant analysis for LA and OIL in OC and RC RILs Resequencing of the 8 DNA pools and 2 parents by an Illumina HiSeq 2000 sequencer generated a total of 2.65 billion paired-end reads of 150 bp in length with an average coverage depth of more than 30× in pools and 20× in parents (Table A.10). Reads were mapped on the soybean Williams 82 reference genome, and as a result, a total of 1,348,790 SNPs were identified between the two parents, and a total of 575,122, 500,051, 682,958, and 592,017 high-quality SNPs with quality value ≥ 100 and sequencing depth ≥ 30 for OC_LA, OC_OIL, RC_LA, and RC_OIL were identified in the DNA pools.

The BRM software of Huang et al. [43] was used to detect significant QTLs for LA and OIL in four pairs of high and low pools (Figs. 4c and 5b). As a result, a total of 10 and 28 QTLs were found in OC RILs to be associated with LA and OIL, respectively, while a total of 6 and 8 QTLs were found in RC RILs to be associated with LA and OIL, respectively (Table A.11; Fig. A.5).

The summarized loci for seed size- and oil-related traits in this and previous studies All the above QTNs/QTLs for seed oil-related traits in this study were integrated with 435 QTLs and 259 QTNs in Soybase, and 196 clusters were identified. Among these clusters, 79, 71, 81, 81, 79, and 87 were found to be associated with PA, SA, OA, LA, LNA, and OIL, respectively (Table A.12). All the above QTNs for seed size-related traits in this study were integrated with 359 QTLs and 127 QTNs in Soybase, and 237 clusters were identified in this study. Among these clusters, 79, 84, 83, 69, 67, 72, and 75 were associated with SL, SW, ST, SLW, SLT, SWT, and 100SW, respectively (Table A.13).

3.4. Domestication and improvement loci and their candidate genes

The above 196 and 237 clusters were compared with the above domestication and improvement regions in order to obtain domestication and improvement loci for seed size- and oil-related traits. As a result, a total of 113 domestication and 103 improvement loci for seed oil-related traits were identified (Table A.12), while a total of 132 domestication and 118 improvement loci for seed size-related traits were identified (Table A.13).

Genes around the above domestication and improvement loci We searched all the genes within each domestication or improvement locus. As a result, 11,731 and 8,214 genes were

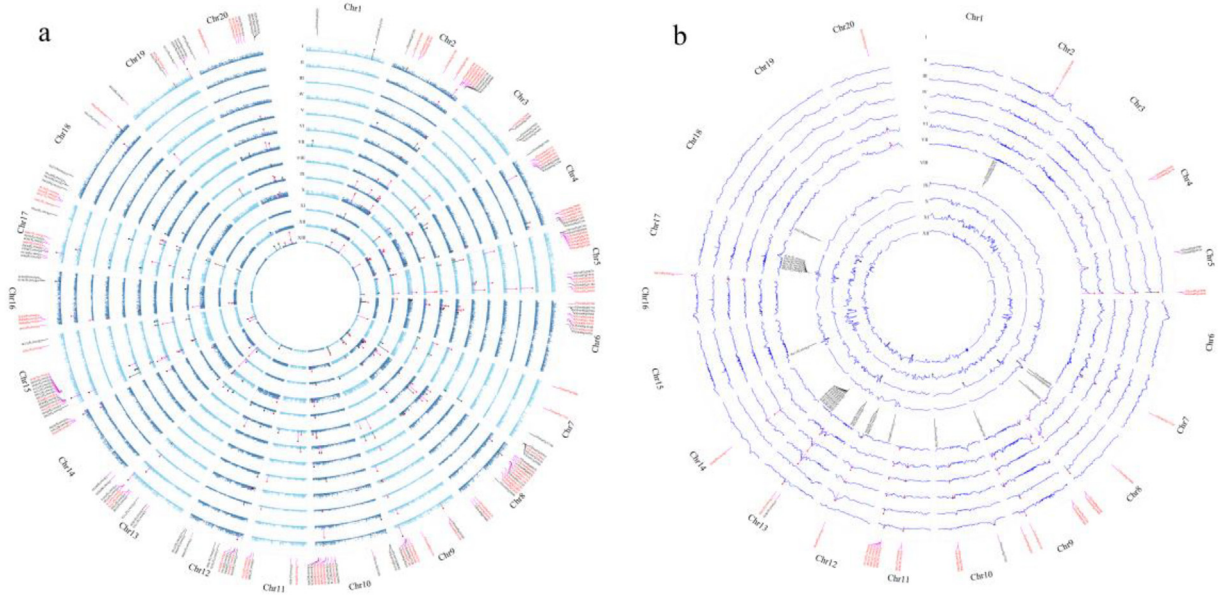

Fig. 5. The loci for seed oil- and size-related traits in soybean. a, QTNs and candidate genes for palmitic acid (I), stearic acid (II), oleic acid (III), linoleic acid (IV), linolenic acid (V), oil content (OIL, VI), seed length (VII), seed width (VIII), seed thickness (IX), length-to-width ratio (X), length-to-thickness ratio (XI), width-to-thickness ratio (XII) and 100-seed weight (XIII) using mrMLM v4.0.2 software. b, QTLs and candidate genes for seed oil-related by QTL mapping and bulked segregant analysis (BSA). I: candidate genes for seed oil-related traits using QTL mapping; II∼VII: QTLs for palmitic acid (II), stearic acid (III), oleic acid (IV), linoleic acid (V), linolenic acid (VI) and oil content (VII) using QTL mapping; VIII: candidate genes for seed linoleic acid (LA) and oil content (OIL) using BSA; IX∼XII: QTLs detected in LA_OC (IX), LA_RC (X), OIL_OC (XI) and OIL_RC (XII) extreme pools. OC: orthogonal cross; RC: reciprocal cross. The loci or candidate genes repeatedly identified are marked by magenta line, while the others are marked by black line. The loci or candidate genes associated with multiple traits are marked by red color, while the others are marked by black color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

found to be within 113 domestication and 103 improvement loci for seed oil-related traits, respectively, and 12,999 and 8,941 genes were found to be within 132 and 118 loci for seed size-related traits, respectively. Among these genes, eleven were previously verified through functional experiments to regulate seed oil metabolism, such as *GmNFYA* [15], *GmFATB2B* [60], *GmDGAT1A* [61], *B1* [17], *GmFATB1B* [60], *GmFATA1A* [60], and *Glyma20g21376* [62] around domestication loci, and *GmNFYA* [15], *GmbZIP123* [63], *FAD2-1A* [64], *GmSACPD-C* [65], and *GmSWEET39* / *GmSWEET10a* [14] around improvement loci, whereas three were previously verified through functional experiments to regulate seed size / weight traits, such as *GmBS1* [66], *GmSWEET39* / *GmSWEET10a* [14], and *GmCIF1* [67] around improvement loci (Table 1).

The genotypic datasets of 302 soybean accessions in Zhou et al. [12] were used to detect the significances of allelic frequency differences for these known genes during soybean domestication and improvement. As a result, at least one SNP within each known gene was found to have significant differences ($P \leq 0.05$; Table A.14). These SNP variants were used to conduct haplotype analysis for seed oil content (172 accessions) and weight (134

Table 1
The previously reported genes around domestication and improvement loci for seed oil- and size-related traits.

Gene	Symbol	SNP				Wild soybean		Landrace		Bred soybean		*P*-value	Reference
		No.	Genome region	NoH	EH	NoH	% EH	NoH	% EH	NoH	% EH		
Known genes around domestication loci for seed oil-related traits													
Glyma02g47380	*GmNFYA*	9	3′UTR; 5′UTR; UP	7	Hap1	6	3.23	5	47.83	2	80.56	1.16E-23	[15]
Glyma06g23560	*GmFATB2B*	6	3′UTR; 5′UTR; UP	6	Hap2	5	3.12	4	11.54	4	58.33	2.60E-22	[60]
Glyma13g16560	*GmDGAT1A*	2	3′UTR; CDS	3	Hap1	3	4.65	3	47.37	2	73.68	1.01E-29	[61]
Glyma13g31540	*B1*	8	3′UTR; 5′UTR; UP; CDS	5	Hap2	3	0.00	3	6.10	2	46.15	6.79E-46	[17]
Glyma17g12940	*GmFATB1B*	3	3′UTR; 5′UTR; UP	5	Hap3	5	3.03	4	5.06	2	32.43	2.09E-36	[60]
Glyma18g36130	*GmFATA1A*	2	UP	3	Hap1	3	8.11	3	79.75	2	97.50	3.45E-26	[60]
Glyma20g21376		3	UP	3	Hap1	3	43.75	3	95.18	1	100.00	2.41E-13	[62]
Known genes around improvement loci for seed oil-related traits													
Glyma02g47380	*GmNFYA*	11	3′UTR; 5′UTR; UP	7	Hap1	6	3.45	5	48.57	2	80.56	1.16E-22	[15]
Glyma06g01240	*GmbZIP123*	1	CDS	2	Hap2	2	2.22	2	49.37	2	4.88	4.37E-02	[63]
Glyma10g42470	*FAD2-1A*	2	3′UTR; CDS	3	Hap1	3	33.33	3	65.82	2	94.74	1.20E-09	[64]
Glyma14g27990	*GmSACPD-C*	3	3′UTR; 5′UTR; UP	3	Hap1, Hap2	3	89.74	2	100.00	2	100.00	9.54E-04	[65]
Glyma15g05470	*GmSWEET39*	7	3′UTR; CDS	3	Hap1	2	6.90	3	82.50	1	100.00	8.65E-33	[14]
Known genes around improvement loci for seed size-related traits													
Glyma10g38970	*GmBS1*	2	UP	2	Hap1	2	35.29	2	89.74	2	90.32	1.59E-06	[66]
Glyma15g05470	*GmSWEET39*	7	3′UTR; CDS	3	Hap1	2	0	3	82.28	1	100	1.16E-19	[14]
Glyma17g04040	*GmCIF1*	2	CDS	4	Hap1	3	5.88	3	87.65	1	100	8.19E-12	[67]

3′UTR: 3′ untranslated region; 5′UTR: 5′ untranslated region; UP: upstream; CDS: Coding sequence. NoH: No. of haplotypes; EH: elite haplotype. *P*-value is obtained from ANOVA for the traits of interest across various haplotypes.

accessions), for which trait phenotypes were downloaded from the USDA GRIN database. The result showed that all the differences of seed weight and oil content across various haplotypes for all the known genes were significant (Table 1).

High expression genes The expression datasets at seed developmental stages from Jones and Vodkin [44], Machado et al. [46] and Severin et al. [47] were used to conduct high expression analysis (Fig. 3c and 4d-e). Within domestication and improvement loci, as a result, there were a total of 5,709 and 4,049 high expression genes at seed oil accumulation stages for seed oil-related traits and a total of 6,415 and 4,506 high expression genes at seed development stages for seed size-related traits.

The SNP variance and annotation The genotypic datasets of 302 soybean accessions of Zhou et al. [12] were used to detect the significant differences of SNPs within high expression genes and their upstream regulation regions between wild and landrace soybeans and between landrace and bred soybeans. As a result, a total of 5,709 (100%) potential candidate domestication (PCD) and 2,647 (65.37%) potential candidate improvement (PCI) high expression genes for seed oil-related traits were found to have significant differences of allelic frequencies ($P \leq$ 8.61e-8 for domestication; $P \leq$ 1.17e-7 for improvement), and a total of 6,415 (100%) PCD and 2,905 (64.47%) PCI high expression genes for seed size-related traits were found to have significant differences of allelic frequencies ($P \leq$ 7.81e-8 for domestication; $P \leq$ 1.08e-7 for improvement).

We used the software SnpEff [51] to annotate the SNP variations in the above significantly different genes. As a result, a total of 3,053 PCD and 683 PCI genes with LOF variants, 4,084 PCD and 982 PCI genes with UTR variants, and 5,534 PCD and 2,177 PCI genes with upstream variants were identified for seed oil-related traits, while a total of 3,402 PCD and 753 PCI genes with LOF variants, 4,603 PCD and 1,092 PCI genes with UTR variants, and 6,232 PCD and 2,400 PCI genes with upstream variants were identified for seed size-related traits.

Differential expression analysis The expression datasets of two wild, two landrace, and two bred soybeans from Niu et al. [25] and Liu et al. [49] were used to mine all the PCD and PCI genes with UTR and upstream variants. As a result, 2,060 (2,567) PCD and 323 (716) PCI genes with UTR (upstream) variants were identified for seed oil-related traits to have differential expression levels between wild and landrace soybeans and between landrace and bred soybeans, respectively, and 2,309 (2,895) PCD and 358 (797) PCI genes with UTR (upstream) variants were identified for seed size-related traits to have differential expression levels between wild and landrace soybeans and between landrace and bred soybeans, respectively (Tables A.15 and A.16).

These differentially expressed genes, along with the above LOF variant genes, were used to conduct the following analysis, including 4,338 PCD and 1,260 PCI genes for seed oil-related traits, and 4,886 PCD and 1,376 PCI genes for seed size-related traits.

Further selection of PCD and PCI genes The above PCD and PCI genes were further selected via special expression analysis in seed, micoRNA (miRNA) regulation analysis, the common candidate lipid-metabolism-related genes with those in Zhang et al. [45], and KEGG analysis. First, the RNA-seq datasets of 14 soybean tissues from Severin et al. [47] were used to identify genes with seed-specific expression. As a result, 85 PCD and 35 PCI genes for seed oil-related traits (Fig. A.6a) and 96 PCD and 39 PCI genes for seed size-related traits (Fig. A.6b) were specifically expressed in seed rather than in other tissues (Fig. 4e).

Then, the website psRNATarget was used to predict the binding of miRNA with the above UTR variant genes. As a result, some SNPs among 23 PCD and 7 PCI genes for seed oil-related traits and among 25 PCD and 7 PCI genes for seed size-related traits were identified to be located on their miRNA binding regions (Table A.17). Next, 104 PCD and 41 PCI genes for seed oil-related traits were found to be consistent with the candidate lipid-metabolism-related genes in Zhang et al. [45]. Finally, all the remaining PCD and PCI genes were used to conduct KEGG analysis. As a result, 34 PCD and 13 PCI genes for seed oil-related traits and 122 PCD and 31 PCI genes for seed size-related traits were identified to be associated with lipid metabolism and seed development, respectively (Table A.18).

DNA methylation The above differential expression and LOF variant genes were compared to 4248 differentially methylated domestication and 1164 differentially methylated improvement regions in Shen et al. [58]. As a result, a total of 92 PCD and 10 PCI genes for seed oil-related traits and 110 PCD and 10 PCI genes for seed size-related traits were located in differentially methylated regions (Table A.19).

In summary, 327 CDGs and 103 CIGs for seed oil-related traits and 343 CDGs and 85 CIGs for seed size-related traits were identified.

3.5. Validation of CDGs and CIGs for seed size- and oil-related traits

Further identification of SNP variants in 30 genomic sequences We downloaded 26 soybean genomic sequences of Liu et al. [53] and 4 soybean genomic sequences from Soybase, including 5 wild, 10 landrace, and 15 bred soybeans, and aligned the sequences of the above CDGs and CIGs in 30 genomes. As a result, most SNP variants in 302 accessions were also found in 30 genomes, and 220 CDGs and 90 CIGs for seed oil-related traits and 164 CDGs and 57 CIGs for seed size-related traits were found to have common SNP variants between 30 genomes and 302 accessions.

Haplotype analysis for CDGs and CIGs The further identified SNP variants in 30 genomes were used to conduct haplotype analysis, and the phenotypic values of haplotypes for seed oil content and weight were calculated, respectively, from 172 soybean accessions and 134 soybean accessions, which were downloaded from the USDA GRIN database. As a result, significant differences in seed weight or oil content were observed among haplotypes in 197 CDGs and 65 CIGs for seed oil-related traits and in 139 CDGs and 36 CIGs for seed size-related traits (Table A.20; Figs. 3d-f, 4g-i and A.7 to A.10).

For the above-mentioned 196 CDGs for seed oil-related traits and 138 CDGs for seed size-related traits (Table A.20; Figs. 3, A.7, and A.9), the frequencies of elite haplotypes increased from wild to landrace soybeans, such as *GmGA2OX2* for seed size-related traits; the frequency of elite haplotype Hap1 increased from 0.00% (wild) to 66.67% (landrace) (Fig. 3). The same situation could be found for *DGK1*, *mtACP1*, *LOX*, *LACS2*, *LACS9*, *DES1.2*, *GmOLE9*, and *BCCP2*. For the above-mentioned 61 CIGs for seed oil-related traits and 27 CIGs for seed size-related traits (Table A.20; Figs. 4, A.8 and A.10), the frequencies of elite haplotypes increased from landrace to bred soybeans, such as *GmMFT* for seed oil-related traits; the frequency of elite haplotype Hap1 increased from 66.67% (landrace) to 92.31% (bred) (Fig. 4). Interestingly, 13 seed oil-related and 2 seed size-related candidate genes simultaneously underwent both domestication and improvement processes. The frequencies of elite haplotypes increased from wild to landrace soybeans and from landrace to modern cultivars (Table A.20; Figs. A.7-A.10), such as *Glyma04g36170* for seed oil-related traits; the frequency of elite haplotype Hap1 increased from 0.00% (wild) to 46.97% (landrace), and from 46.97% (landrace) to 78.95% (bred).

GO enrichment analysis for CDGs and CIGs The above CDGs and CIGs were used to conduct GO enrichment analysis via the KOBAS software. The results were listed in Table A.21. As a result, 11 and 11 out of 23 and 49 terms significantly enriched from the CDGs and CIGs, respectively, for seed oil-related traits that were

found to be associated with lipid metabolism; the former includes "fatty acid biosynthetic process", "triglyceride biosynthetic process", and "lipid transporter activity", while the latter includes "fatty acid elongation", "phospholipase D activity", "phospholipid catabolic process", "acetyl-CoA biosynthetic process from pyruvate", and "long-chain fatty acid-CoA ligase activity"; 2 and 3 out of 8 and 37 terms significantly enriched from the CDGs and CIGs, respectively, for seed size-related traits that were found to be associated with seed development; the former includes "ubiquitin-dependent protein catabolic process" and "ubiquitin conjugating enzyme activity", while the latter includes "protein polyubiquitination", "cell tip growth", "brassinosteroid mediated signaling pathway", and "response to abscisic acid". Among the above 28 terms, there are 32 seed oil and 10 seed size previously reported genes (Table A.21). These results confirmed the reliability of these CDGs and CIGs in this study.

4. Discussion

As shown in Fig. 2, there are significant differences of seed size- and oil-related traits in the domestication and improvement processes. Although some candidate genes for these traits had been mined in previous studies [12,15,22], the knowledge is limited owing to materials, genetic diversity indicators, and genetic analysis approaches. To overcome these issues, 534 CDRs and 458 CIRs from 286 soybean accessions via the XP-CLR method in this study were integrated with those in five [9–13] and three [9,10,12] previous studies, respectively, to obtain the 952 CDRs and 538 CIRs (Tables A.5 and A.6; Figs. 3-4), while 1469 loci for soybean seed size- and oil-related traits via GWAS, QTL mapping, and BSA in this study were integrated with those in Soybase to obtain the 433 QTL clusters for these traits. All the above CDRs and CIRs were integrated with all the above loci to obtain the trait-related 245 domestication and 221 improvement loci (Tables A.11 and A.12). Around these trait-related domestication and improvement loci, all the genes were scanned by using gene expression, SNP variant in genotype and genome, miRNA binding, KEGG pathway, DNA methylation, and haplotype analysis in order to obtain 372 CDGs and 87 CIGs for these traits (Table A.22; Figs. 3-4 and A.11-34). Thus, systematic and summary results were reported in this study. This technique route is different from those in previous studies. Most previous studies focus on genetic diversity at the genome-wide level to determine CDRs and CIRs, such as Li et al. [9], Song et al. [10], Chung et al. [11], Zhou et al. [12], and Zhou et al. [13], and some candidate genes in association and linkage studies were found to be domesticated or improved, such as in Miao et al. [68] and Zhang et al. [18]. In addition, newly developed methodologies in GWAS, QTL mapping, and BSA were adopted in this study to identify more genetic loci [34,43,69]. Although domestication/improvement regions and trait-related loci, along with those in previous studies, had been identified as far as possible, some domestication and improvement genes may not be found due to population limitations.

4.1. Candidate domestication genes can explain the differences of seed oil- and size-related traits during soybean domestication

The known genes and the above-mentioned CDGs were used to dissect why seed weight and oil content increased during soybean domestication in two aspects.

First, the frequencies of elite haplotypes in seven known genes (*GmNFYA* [15], *GmFATB2B* [60], *GmDGAT1A* [61], *B1* [17], *GmFATB1B* [60], *GmFATA1A* [60], and *Glyma20g21376* [62]) around domestication loci for seed-oil-related traits and 196 seed-oil-related and 138 seed-size-related CDGs are higher in landrace soybean than in wild soybean (Table A.20; Figs. A.7 and A.9), such as known gene *GmNFYA* [15] for seed oil-related traits, for which the frequency of its elite haplotype Hap1 increased from 3.23% (wild) to 47.83% (landrace); *GmOLE9* for seed oil-related traits, for which the frequency of its elite haplotype Hap1 increased from 26.92% (wild) to 98.81% (landrace) (Fig. A.11h, i). Among these CDGs, *GmOLE9* (homologous to *GmOLEO1*; [18]) has been found to increase seed size or oil content during soybean domestication. This indicates the potential of these CDGs for increasing seed size and oil content during soybean domestication.

Second, the variants of amino acid sequences and expression levels of the above-mentioned CDGs may result in phenotypic changes of seed size- and oil-related traits during soybean domestication. In this study, on one hand, we observed the variants of amino acid sequences in two known genes (*GmDGAT1A* [61] and *B1* [17]) around domestication loci for seed-oil-related traits and 82 seed-oil-related and 67 seed-size-related CDGs (Table A.20); e.g., three non-synonymous mutations in known gene *B1* [17] and one non-synonymous mutation in *GmLPTA2* changed amino acid sequences. LOF mutation in a unique gene necessarily results in different phenotypes [70]. In previous studies, the changes of amino acid sequences have been found to affect trait phenotypes [4,17,21]. Thus, we deduce that the variations of amino acid sequences in our 117 CDGs may change the phenotypes of seed size or oil content during soybean domestication (Table A.20). On the other hand, gene expression levels can be affected by variations of gene regulatory and miRNA binding regions and epigenetic modification. In this study, we detected the variants of regulatory regions in seven known genes (*GmNFYA* [15], *GmFATB2B* [60], *GmDGAT1A* [61], *B1* [17], *GmFATB1B* [60], *GmFATA1A* [60], and *Glyma20g21376* [62]) around domestication loci for seed-oil-related traits and 196 seed oil-related and 139 seed size-related CDGs (Table A.20), such as, five SNP domestication loci in 3′UTR, 5′UTR, and upstream regulatory regions of known gene *B1* [17]; and seven SNP domestication loci in 3′UTR, 5′UTR, and upstream regulatory regions of *GmOLE9* (Fig. A.11g). This gene was highly and specifically expressed, in seed rather than other tissues, at oil accumulation stage (Fig. A.11d-e), and differentially expressed at the middle seed development stage between ten wild soybeans and ten cultivated soybeans (P-value < 0.05; Fig. A.11f; [14]). Thus, these variations may lead to the changes of gene expression levels. In this study, we also found the variants of miRNA binding regions in 24 seed oil-related and 25 seed size-related CDGs in 302 soybean accessions [12] and 30 soybean genomes ([53]; Soybase; Table A.17), such as one SNP domestication locus in 3′UTR and miRNA binding region of *UPL3*. This gene was highly expressed at oil accumulation stage, and differentially expressed at early seed development stage between ten wild and ten cultivated soybeans (P-value < 0.01; [14]). Thus, these variations may lead to the changes of gene expression levels. In this study, we still found significant differences of DNA methylation degrees between wild and landrace soybeans in 92 seed oil-related and 110 seed size-related CDGs (Table A.19); e.g., the DNA methylation degrees of domestication genes *ACCase1* and *CYP51G1* in landrace soybean are significantly lower than those in wild soybean [58]. All the above-mentioned CDGs were differentially expressed between two wild and two landrace soybeans (Table A.15). In previous studies, it has been shown that the changes of gene expression levels resulted in significant differences of trait phenotypes [71]. Thus, we deduce that the variations of gene regulatory and miRNA binding regions and epigenetic modification may result in the phenotypic changes of seed size- and oil-related traits during soybean domestication.

In addition, 26 seed-oil-related and 25 seed-size-related CDGs in the above-mentioned CDGs were found to encode proteins with unknown functions (Table A.21).

4.2. Candidate improvement genes can explain the differences of seed oil- and size-related traits during soybean improvement

The known genes and the above-mentioned CIGs were used to dissect why seed weight and oil content increased during soybean improvement in two aspects.

First, the frequencies of elite haplotypes in three known genes (*GmNFYA* [15], *FAD2-1A* [64], and *GmSWEET39* / *GmSWEET10a* [14]) around improvement loci for seed oil-related traits, three known genes (*GmBS1* [66], *GmSWEET39* / *GmSWEET10a* [14], and *GmCIF1* [67]) around improvement loci for seed size-related traits, and 61 seed-oil-related and 27 seed-size-related CIGs are higher in bred soybean than in landrace soybean (Table A.20; Figs. A.8 and A.10), such as known gene *GmNFYA* [15] for seed oil-related trait, which has an increased frequency of its elite haplotype Hap1 from 48.57% (landrace) to 80.56% (bred) and *KAS III* for seed oil-related traits, which has an increased frequency of its elite haplotype Hap1 from 38.24% (landrace) to 88.24% (bred) (Fig. A.8). Among these CIGs and their homologous genes, *GmLPAT5* [72] and *GmKAS III* (homologous to *KAS III* in many plants; [73,74]) have been found to increase seed size and/or oil content during soybean improvement. This indicates the potential of these CIGs for increasing seed size and oil content during soybean improvement.

Second, the variants of amino acid sequences and expression levels of the above-mentioned CIGs may result in phenotypic changes of seed size- and oil-related traits during soybean improvement. In this study, on one hand, we observed the variants of amino acid sequences in three known genes (*GmbZIP123* [63], *FAD2-1A* [64], and *GmSWEET39* / *GmSWEET10a* [14]) around improvement loci for seed oil-related traits, two known genes (*GmSWEET39* / *GmSWEET10a* [14], and *GmCIF1* [67]) around improvement loci for seed size-related traits, and 31 seed oil- and 20 size-related CIGs (Table A.20), for example, two non-synonymous mutations in known gene *GmSWEET39* / *GmSWEET10a* [14] changed amino acid sequences. Here we found the change of amino acid sequences owing to two non-synonymous mutations of *GmLPAT5*, while Angkawijaya et al. [72] confirmed the regulation of *LPAT5* in *Arabidopsis* involved in glycerolipid metabolism. As described above, the changes of amino acid sequences have been found to affect trait phenotypes. Thus, we deduced that the variations of amino acid sequences in our 40 CIGs may change the phenotypes of seed size or oil content during soybean improvement (Table A.20). On the other hand, gene expression levels can be affected by epigenetic modification and the variations of both gene regulatory and miRNA binding regions. In this study, we detected variants of regulatory regions in four known genes (*GmNFYA* [15], *FAD2-1A* [64], *GmSACPD-C* [65], and *GmSWEET39* / *GmSWEET10a* [14]) around improvement loci for seed oil-related traits, two known genes (*GmBS1* [66], and *GmSWEET39* / *GmSWEET10a* [14]) around improvement loci for seed size-related traits, and 67 seed oil-related and 39 seed size-related CIGs (Table A.20), such as five SNP improvement loci in 3′UTR of known gene *GmSWEET39* / *GmSWEET10a* [14]; 18 SNP improvement loci in upstream regulatory regions; 3′UTR of *GmMFT* (Fig. 4g), which was highly and specifically expressed, in seed rather than other tissues, at oil accumulation stage (Fig. 4d-e), and differentially expressed at the middle seed development stage between ten wild soybeans and ten cultivated soybeans (*P*-value < 0.001; Fig. 4f; [15]). Thus, these variations may lead to the changes of gene expression levels. We also found variants of miRNA binding regions in 7 seed oil-related and 7 seed size-related CIGs in 302 soybean accessions [12] and 30 soybean genomes ([53]; Soybase; Table A.17), such as one SNP improvement locus in 5′UTR and miRNA binding region of *SCPL40*, which was highly and specifically expressed at oil accumulation stage (Fig. A.6a), and differentially expressed between landrace and bred soybeans (*P*-value < 0.001; Table A.16). Thus, these variations may lead to the changes of gene expression levels. We also found significant differences of DNA methylation degrees between landrace and bred soybeans in 10 seed oil-related and 10 seed size-related CIGs (Table A.19); e.g., the DNA methylation degree of improvement gene *Glyma02g12460* in bred soybean is significantly lower than that of landrace soybean [58]. All the above-mentioned CIGs were differentially expressed between two landrace and two bred soybeans (Table A.16). In previous studies, the changes of gene expression levels resulted in significant differences of trait phenotypes [71]. Thus, we deduced that epigenetic modification and the variations of gene regulatory and miRNA binding regions may result in the phenotypic changes of seed size- and oil-related traits during the soybean improvement process.

In addition, 5 seed-oil-related and 5 seed-size-related CIGs in the above-mentioned CIGs were found to encode proteins with unknown functions (Table A.22).

4.3. Pleiotropic genes for seed size- and oil-related traits in soybean

Gene pleiotropy is a common phenomenon in the genetic dissection of complex, domesticated and improved traits [75,76]; e.g., *Q* regulates free threshing ability, panicle length, plant height, and heading date in *Triticum aestivum* [75], and *GAD1* regulates grain number, grain length, and awn development in rice domestication [76].

Among all the 372 CDGs and 87 CIGs in this study, 356 CDGs and 81 CIGs were found to be associated with at least two traits. 258 CDGs and 70 CIGs were found to be associated with at least two seed oil-related traits, explaining the correlation of seed oil-related traits. The same phenomenon was also found by Zhang et al. [21] and Zhou et al. [77]. 237 CDGs and 40 CIGs were found to be associated with at least two seed size-related traits, explaining the correlation of seed size-related traits. The same phenomenon was also found by Xu et al. [78] and Niu et al. [25]. 152 CDGs and 32 CIGs were found to be associated with both at least one seed size-related trait and at least one seed oil-related trait, explaining the correlation between seed oil-related traits and seed size-related traits. The same phenomenon was also found by the canonical correlation analyses of Liu et al. [19].

4.4. Available genes for future soybean improvement

In our opinion, the above-mentioned CDGs and CIGs with low elite haplotype frequency and unknown functions (or coding proteins) are relevant to future soybean improvement of seed size- and oil-related traits (Table A.23). The discovery of important crop domestication/improvement genes can accelerate breeding selections and facilitate ideal crop designs [79].

The elite haplotype frequencies of two known genes (*GmFATB1B* [60] and *Glyma20g21376* [62]) around domestication loci for seed oil-related traits, three known genes (*FAD2-1A* [64], *GmSACPD-C* [65], and *GmSWEET39* / *GmSWEET10a* [14]) around improvement loci for seed oil-related traits, three known genes (*GmBS1* [66], *GmSWEET39*/*GmSWEET10a* [14], and *GmCIF1* [67]) around improvement loci for seed size-related traits, 147 CDGs and 32 CIGs for seed oil-related traits, and 105 CDGs and 25 CIGs for seed size-related traits were more than 90% in bred soybeans, while one known gene (*B1* [17]) around domestication loci for seed oil-related trait, one known gene (*GmbZIP123* [63]) around improvement loci for seed oil-related trait, 17 seed-oil-related CDGs, 5 seed-oil-related CIGs, 12 seed-size-related CDGs, and 4 seed-size-related CIGs with elite haplotype frequencies<50% in bred soybeans (Table 1 and A.20; Figs. A.7-A.10 and A.35) can be exploited for the improvement of soybean cultivars, e.g., known genes *B1* (46.15% in bred soybeans; [17]) and *Glyma02g07250* (43.59% in

bred soybeans). The same situation can be found for *GmbZIP123* [63], *LACS9*, *PLC2*, *HCD1*, and *MYB30* (Table 1 and A.23; Figs. A.7-A.10). These elite haplotypes of CDGs and CIGs may be transferred into main cultivars in soybean production without the elite haplotypes in order to improve their seed size and oil content (Fig. 1).

26 CDGs and 5 CIGs for seed oil-related traits and 25 CDGs and 5 CIGs for seed size-related traits with unknown gene function were identified in this study (Tables A.22-A.23), including *Glyma02g16660* and *Glyma04g36170*. Although the function of these genes is unknown, these candidate genes may regulate seed oil-related traits. Thus, these genes may be of breeding value.

By counting the number of elite haplotypes in each of the 302 soybean accessions, 3 landrace and 14 improved soybeans were identified to have more than 90% elite haplotypes. Among these cultivars, there are one landrace and seven improved cultivars for seed oil content and 2 landrace and 11 improved cultivars for seed weight, such as PI547716 (Table A.24). These accessions may be used as main parents in soybean breeding. In addition, we found that almost all the 62 wild soybeans had less than 30% elite haplotypes, while 17 landrace and 2 improved soybeans had less than 30% elite haplotypes, such as Hu PI Dou (Table A.25). Thus, we speculated that these accessions may be improved by the introduction of elite genes for seed size- and oil-related traits, although their functions in molecular biology and roles in soybean breeding need to be further verified.

Recently, Zhuang et al. [80] assembled chromosome-level genomes of representative perennial species and constructed a *Glycine* super-pangenome framework. Zhuang et al. [80] compared the differences between annual and perennial diploid soybeans, and one of their purposes was to mine candidate genes responsible for the transition from perennial to annual soybeans. In contrast, our study focused on mining candidate domestication and improvement genes for seed oil- and size-related traits; these genes are used to explain the trait differences between wild and landrace soybeans and between landrace and bred soybeans, in which the wild, landrace, and bred soybeans are annual diploid soybean.

Recently, our team established a new compressed variance component mixed model framework in GWAS [81,82]. Although this method can detect all types of loci and estimate their effects conditional on fully controlling all the possible polygene backgrounds, we adopted our multi-locus GWAS approaches in this study for the following reasons. First, the estimated effects of QTNs in our multi-locus GWAS methods are additive effects when the marker genotypes are homozygous in our association mapping population. Second, QTN-by-environment and QTN-by-QTN interactions are not involved in this study. Thus, it is unnecessary to adopt our new 3VmrMLM method.

5. Conclusion

Among 952 domestication and 538 improvement regions in this and previous studies, 300 domestication and 408 improvement regions were newly identified, while among 196 seed oil-related trait and 237 seed size-related trait loci in this and previous studies, 103 and 173 were newly detected, respectively. Around 66 domestication and 56 improvement loci, seven known domestication and seven known improvement genes were identified, and 372 candidate domestication and 87 candidate improvement genes were mined from multi-omics analysis. Among these candidate genes, their frequencies of elite haplotypes, base mutations in coding region, and three factors affecting their expression levels were used to elucidate the trait changes in domestication and improvement processes. In addition, we found that 56 domestication and 15 improvement genes may be valuable, 17 soybean accessions may be used as main parents, and 19 soybean accessions could be further improved in future soybean breeding.

Conflict of interest

The authors declared that they have no conflict of interest to this work.

Author statement

Y.M.Z. conceived and managed the research. J.F.Z., and M.I. analyzed datasets. J.F.Z., J.Y.L., C.Y.H., and Y.N. measured the phenotypes of the traits. J.F.Z. carried out bulked segregant analysis and wrote the draft. Y.M.Z., J.F.Z., and J.M.D. revised the manuscript.

Acknowledgments

We thank Mr. Hanwen Zhang (Hence Education Ltd., Vancouver, Canada; hywenzhang@henceedu.com) for improving the language within the manuscript, and senior engineer Ms Ju Huang for the help in data analysis. This work was supported by the National Natural Science Foundation of China (32070557 and 31871242), the Fundamental Research Funds for the Central Universities (2662020ZKPY017), and Huazhong Agricultural University Scientific & Technological Self-Innovation Foundation (2014RC020).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.014.

References

[1] Carter TE, Nelson R, Sneller CH, Cui Z. Soybeans: improvement, production and uses. 3rd edn. Wisconsin, USA: Madison; 2004.
[2] Diamond J. Evolution, consequences and future of plant and animal domestication. Nature 2002;418:700–7.
[3] Purugganan MD. The molecular population genetics of regulatory genes. Mol Ecol 2000;9:1451–61.
[4] Doebley JF, Gaut BS, Smith BD. The molecular genetics of crop domestication. Cell 2006;127:1309–21.
[5] Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia JM, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nat Genet 2012;44:808–11.
[6] Zhang Z, Li A, Song G, Geng S, Gill BS, Faris JD, et al. Comprehensive analysis of Q gene near-isogenic lines reveals key molecular pathways for wheat domestication and improvement. Plant J 2020;102:299–310.
[7] Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, et al. Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci U S A 2006;103:16666–71.
[8] Bayer PE, Valliyodan B, Hu H, Marsh JI, Yuan Y, Vuong TD, et al. Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. Plant Genome 2022;15(1):e20109.
[9] Li YH, Zhao SC, Ma JX, Li D, Yan L, Li J, et al. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 2013;14:579.
[10] Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, et al. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS ONE 2013;8:e54985.
[11] Chung WH, Jeong N, Kim J, Lee WK, Lee YG, Lee SH, et al. Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. DNA Res 2014;21:153–67.
[12] Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 2015;33:408–14.
[13] Zhou L, Luo L, Zuo JF, Yang L, Zhang L, Guang X, et al. Identification and validation of candidate genes associated with domesticated and improved traits in soybean. Plant Genome 2016;9. https://doi.org/10.3835/plantgenome2015.09.0090. PMID: 27898807.
[14] Wang S, Liu S, Wang J, Yokosho K, Zhou B, Yu YC, et al. Simultaneous changes in seed size, oil content and protein content driven by selection of *SWEET* homologues during soybean domestication. Natl Sci Rev 2020;7:1776–86.
[15] Lu X, Li QT, Xiong Q, Li W, Bi YD, Lai YC, et al. The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. Plant J 2016;86:530–44.

[16] Li QT, Lu X, Song QX, Chen HW, Wei W, Tao JJ, et al. Selection for a zinc-finger protein contributes to seed oil increase during soybean domestication. Plant Physiol 2017;173:2208–24.
[17] Zhang D, Sun L, Li S, Wang W, Ding Y, Swarm SA, et al. Elevation of soybean seed oil content through selection for seed coat shininess. Nat Plants 2018;4:30–5.
[18] Zhang D, Zhang H, Hu Z, Chu S, Yu K, Lv L, et al. Artificial selection on *GmOLEO1* contributes to the increase in seed oil during soybean domestication. PLoS Genet 2019;15:e1008267.
[19] Liu JY, Zhang YW, Han X, Zuo JF, Zhang Z, Shang H, et al. An evolutionary population structure model reveals pleiotropic effects of *GmPDAT* for traits related to seed size and oil content in soybean. J Exp Bot 2020;71:6988–7002.
[20] Sedivy EJ, Wu F, Hanzawa Y. Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol 2017;214:539–53.
[21] Zhang J, Wang X, Lu Y, Bhusal SJ, Song Q, Cregan PB, et al. Genome-wide scan for seed composition provides insights into soybean quality improvement and the impacts of domestication and breeding. Mol Plant 2018;11:460–72.
[22] Torkamaneh D, Laroche J, Rajcan I, Belzile F. Identification of candidate domestication- related genes with a systematic survey of loss-of-function mutations. Plant J 2018;96:1218–27.
[23] Turquetti-Moraes DK, Moharana KC, Almeida-Silva F, Pedrosa-Silva F, Venancio TM. Integrating omics approaches to discover and prioritize candidate genes involved in oil biosynthesis in soybean. Gene 2022;808:145976.
[24] Zhou L, Wang SB, Jian J, Geng QC, Wen J, Song Q, et al. Identification of domestication-related loci associated with flowering time and seed size in soybean with the RAD-seq genotyping method. Sci Rep 2015;5:9350.
[25] Niu Y, Xu Y, Liu XF, Yang SX, Wei SP, Xie FT, et al. Association mapping for seed size and shape traits in soybean cultivars. Mol Breeding 2013;31:785–94.
[26] Ikram M, Han X, Zuo JF, Song J, Han CY, Zhang YW, et al. Identification of QTNs and their candidate genes for 100-seed weight in soybean (*Glycine max* L.) using multi-locus genome-wide association studies. Genes 2020;11:714.
[27] Zuo JF, Niu Y, Cheng P, Feng JY, Han SF, Zhang YH, et al. Effect of marker segregation distortion on high density linkage map construction and QTL mapping in soybean (*Glycine max* L.). Heredity 2019;123:579–92.
[28] Doyle JJ, Doyle JL. Isolation of plant DNA from fresh tissue. Focus 1990;12:39–40.
[29] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
[30] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
[31] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
[32] Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res 2010;20:393–402.
[33] Zhang YW, Tamba CL, Wen YJ, Li P, Ren WL, Ni YL, et al. mrMLM v4.0.2: An R platform for multi-locus genome-wide association studies. Genom Proteom Bioinf 2020;18:481–7.
[34] Wang SB, Feng JY, Ren WL, Huang B, Zhou L, Wen YJ, et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep 2016;6:19444.
[35] Wen YJ, Zhang H, Ni YL, Huang B, Zhang J, Feng JY, et al. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Brief Bioinform 2018;19:700–12.
[36] Zhang J, Feng JY, Ni YL, Wen YJ, Niu Y, Tamba CL, et al. PLARmEB: Integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. Heredity 2017;118:517–24.
[37] Tamba CL, Ni YL, Zhang YM. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. PLoS Comput Biol 2017;13:e1005357.
[38] Tamba CL, Zhang YM (2018) A fast mrMLM algorithm for multi-locus genome-wide association studies. bioRxiv. doi: 10.1101/341784.
[39] Ren WL, Wen YJ, Dunwell JM, Zhang YM. pKWmEB: integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. Heredity 2018;120:208–18.
[40] Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 2009;9:1322–32.
[41] Zhang YW, Wen YJ, Dunwell JM, Zhang YM. QTL.gCIMapping.GUI v2.0: An R software for detecting small-effect and linked QTLs for quantitative traits in bi-parental segregation populations. Comput Struct Biotechnol J 2019;18:59–65.
[42] Wang SB, Wen YJ, Ren WL, Ni YL, Zhang J, Feng JY, et al. Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. Sci Rep 2016;6:29951.
[43] Huang L, Tang W, Bu S, Wu W. BRM: a statistical method for QTL mapping based on bulked segregant analysis by deep sequencing. Bioinformatics 2020;36:2150–6.
[44] Jones SI, Vodkin LO. Using RNA-Seq to profile soybean seed development from fertilization to maturity. PLoS ONE 2013;8:e59270.
[45] Zhang L, Wang SB, Li QG, Song J, Hao YQ, Zhou L, et al. An integrated bioinformatics analysis reveals divergent evolutionary pattern of oil biosynthesis in high- and low-oil plants. PLoS ONE 2016;11:e0154882.
[46] Machado FB, Moharana KC, Almeida-Silva F, Gazara RK, Pedrosa-Silva F, Coelho FS, et al. Systematic analysis of 1298 RNA-Seq samples and construction of a comprehensive soybean (*Glycine max*) expression atlas. Plant J 2020;103:1894–909.
[47] Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, et al. RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biol 2010;10:160.
[48] Niu Y, Zhang G, Wan F, Zhang YM. Integration of RNA-Seq profiling with genome-wide association study predicts candidate genes for oil accumulation in soybean. Crop Pasture Sci 2020;71:996–1009.
[49] Liu JY, Li P, Zhang YW, Zuo JF, Li G, Han X, et al. Three-dimensional genetic networks among seed oil-related traits, metabolites and genes reveal the genetic foundations of oil synthesis in soybean. Plant J 2020;103:1103–24.
[50] Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 2010;26:136–8.
[51] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, 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 2012;6:80–92.
[52] Dai X, Zhuang Z, Zhao PX. psRNATarget: a plant small RNA target analysis server (2017 release). Nucleic Acids Res 2018;46:W49–54.
[53] Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-genome of wild and cultivated soybeans. Cell 2020;182:162–176.e13.
[54] Quinlan AR (2014) BEDTools: The Swiss-Army tool for genome feature analysis. Curr Protoc Bioinformatics 47: 11.12.1–11.12.34.
[55] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7.
[56] Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–5.
[57] Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet 2018;103:338–48.
[58] Shen Y, Zhang J, Liu Y, Liu S, Liu Z, Duan Z, et al. DNA methylation footprints during soybean domestication and improvement. Genome Biol 2018;19:128.
[59] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57:289–300.
[60] Zhou Z, Lakhssassi N, Knizia D, Cullen MA, El Baz A, Embaby MG, et al. Genome-wide identification and analysis of soybean acyl-ACP thioesterase gene family reveals the role of *GmFAT* to improve fatty acid composition in soybean seed. Theor Appl Genet 2021;134:3611–23.
[61] Chen B, Wang J, Zhang G, Liu J, Manan S, Hu H, et al. Two types of soybean diacylglycerol acyltransferases are differentially involved in triacylglycerol biosynthesis and response to environmental stresses and hormones. Sci Rep 2016;6:28541.
[62] Fliege CE, Ward RA, Vogel P, Nguyen H, Quach T, Guo M, et al. Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20. Plant J 2022;110:114–28.
[63] Song QX, Li QT, Liu YF, Zhang FX, Ma B, Zhang WK, et al. Soybean *GmbZIP123* gene enhances lipid content in the seeds of transgenic *Arabidopsis* plants. J Exp Bot 2013;64:4329–41.
[64] Haun W, Coffman A, Clasen BM, Demorest ZL, Lowy A, Ray E, et al. Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. Plant Biotechnol J 2014;12:934–40.
[65] Kachroo A, Fu DQ, Havens W, Navarre D, Kachroo P, Ghabrial SA. An oleic acid-mediated pathway induces constitutive defense signaling and enhanced resistance to multiple pathogens in soybean. Mol Plant Microbe Interact 2008;21(5):564–75.
[66] Ge L, Yu J, Wang H, Luth D, Bai G, Wang K, et al. Increasing seed size and quality by manipulating *BIG SEEDS1* in legume species. Proc Natl Acad Sci U S A 2016;113:12414–9.
[67] Tang X, Su T, Han M, Wei L, Wang W, Yu Z, et al. Suppression of extracellular invertase inhibitor gene expression improves seed weight in soybean (*Glycine max*). J Exp Bot 2017;68:469–82.
[68] Miao L, Yang S, Zhang K, He J, Wu C, Ren Y, et al. Natural variation and selection in *GmSWEET39* affect soybean seed oil content. New Phytol 2020;225:1651–66.
[69] Zhang YM, Jia Z, Dunwell JM. Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. Front Plant Sci 2019;10:100.
[70] 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 2021;19:324–34.
[71] Doebley J, Stec A, Hubbard L. The evolution of apical dominance in maize. Nature 1997;386:485–8.
[72] Angkawijaya AE, Nguyen VC, Nakamura Y. LYSOPHOSPHATIDIC ACID ACYLTRANSFERASES 4 and 5 are involved in glycerolipid metabolism and nitrogen starvation response in *Arabidopsis*. New Phytol 2019;224:336–51.
[73] Dehesh K, Tai H, Edwards P, Byrne J, Jaworski JG. Overexpression of 3-ketoacyl-acyl-carrier protein synthase IIIs in plants reduces the rate of lipid synthesis. Plant Physiol 2001;125:1103–14.
[74] Cardinal AJ, Whetten R, Wang S, Auclair J, Hyten D, Cregan P, et al. Mapping the low palmitate fap1 mutation and validation of its effects in soybean oil and agronomic traits in three soybean populations. Theor Appl Genet 2014;127:97–111.
[75] Simons KJ, Fellers JP, Trick HN, Zhang Z, Tai YS, Gill BS, et al. Molecular characterization of the major wheat domestication gene *Q*. Genetics 2006;172:547–55.

[76] Jin J, Hua L, Zhu Z, Tan L, Zhao X, Zhang W, et al. *GAD1* encodes a secreted peptide that regulates grain number, grain length, and awn development in rice domestication. Plant Cell 2016;28:2453–63.
[77] Zhou Z, Lakhssassi N, Cullen MA, El Baz A, Vuong TD, Nguyen HT, et al. Assessment of phenotypic variations and correlation among seed composition traits in mutagenized soybean populations. Genes 2019;10:975.
[78] Xu Y, Li HN, Li GJ, Wang X, Cheng LG, Zhang YM. Mapping quantitative trait loci for seed size traits in soybean (*Glycine max* L. Merr.). Theor Appl Genet 2011;122:581–94.
[79] Huang X, Huang S, Han B, Li J. The integrated genomics of crop domestication and breeding. Cell 2022. https://doi.org/10.1016/j.cell.2022.04.036. in press.
[80] Zhuang Y, Wang X, Li X, Hu J, Fan L, Landis JB, et al. Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. Nat Plants 2022;8:233–44.
[81] Li M, Zhang YW, Zhang ZC, Xiang Y, Liu MH, Zhou YH, et al. A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. Mol Plant 2022;15(4):630–50.
[82] Li M, Zhang YW, Xiang Y, Liu MH, Zhang YM. IIIVmrMLM: the R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. Mol Plant 2022. https://doi.org/10.1016/j.molp.2022.06.002. In press.