Evolutionary population structure model reveals pleiotropic effects of *GmPDAT* for seed oil- and size-related traits in soybean

Jin-Yang Liu[1,2,3,6], Ya-Wen Zhang[1,6], Xu Han[1], Jian-Fang Zuo[1], Zhibin Zhang[4], Haihong Shang[4], Qijian Song[5] and Yuan-Ming Zhang[1,*]

1 Crop Information Center, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China

2 Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China

3 State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China

4 Zhengzhou Research Base, State Key Laboratory of Cotton Biology, Zhengzhou University, Zhengzhou 450000, China

5 Soybean Genomics and Improvement Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, Maryland 20705, USA

6 These authors contributed equally to this work.

Correspondence

Yuan-Ming Zhang, Crop Information Center, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China

Email: soyzhang@mail.hzau.edu.cn

Highlight

The evolutionary population structure model in genome-wide association studies identified one domestication locus for seed oil- and size-related traits and the pleiotropic effects of *GmPDAT* were confirmed by transgenic soybean.

Abstract

Soybean seed oil-related traits are related to nutritional effects linked to human health, as well as to crop domestication. These domesticated traits have significant differences across various evolutionary types. The integration of evolutionary population structure (evolutionary types) with genome-wide association studies increased the power in gene detection and identified one locus for seed oil- and size-related traits on chromosome 13. This domestication locus, along with another domestication locus in a 200 kb region, was confirmed by GEMMA and EMMAX. *GmPDAT* had higher expressional level in the high-oil and large-seed accessions than in the low-oil and small-seed accessions. Real-time qPCR analysis of *GmPDAT* showed higher expression levels in overexpression lines and lower expression levels in RNAi lines. Overexpression lines increased seed oil- and size-related traits, whereas RNAi lines decreased seed oil- and size-related traits. In addition, we deduced the molecular mechanism of *GmPDAT* based on the results from linkage analysis for triacylglycerols and histocytological comparison of transgenic soybean seeds. This result provides a new approach for identifying domestication genes with pleiotropic effects.

Keywords: evolutionary population structure, genome-wide association studies, *GmPDAT*, *GmDGAT1*, pleiotropism, soybean

Introduction

In soybean seed, linolenic acid is a major component of cell membranes; it also plays a critical role in human health as it cannot be produced within the human body and is acquired as part of the diet. Traits related to seed oil and seed size are associated with domestication and are targeted in soybean breeding programs. However, improving these traits has been a great challenge (Martin *et al.*, 2011). Although numerous studies of these traits in soybean have been reported, and these traits have been affected by domestication, knowledge about the genes responsible for both seed oil- and size-related traits in soybean is limited.

During the past few decades, efforts have been made to dissect the genetic foundation and molecular mechanism of seed oil-related traits. Approximately five hundred quantitative trait loci (QTL) (Chung *et al.*, 2003; Han *et al.*, 2012; Sun *et al.*, 2012; Eskandari *et al.*, 2013; Wang *et al.*, 2014; Cao *et al.*, 2017; Song *et al.*, 2017) or quantitative trait nucleotides (QTNs) (Niu *et al.*, 2013; Hwang *et al.*, 2014; Contreras-Soto *et al.*, 2017; Fang *et al.*, 2017; Liu *et al.*, 2020) (www.soybase.org) have been found to be associated with such traits. Genes related to acyl-lipid metabolism in soybean and *Arabidopsis* (Li-Beisson *et al.*, 2013) have been reported based on genome sequencing (Schmutz *et al.*, 2010; Li *et al.*, 2014), comparative genomics (Zhang *et al.*, 2016), and transcriptome data analyses (Yu *et al.*, 2014; Lu *et al.*, 2016; Yuan *et al.*, 2017). Some of the genes related to oil biosynthesis have been confirmed by experimental biology, such as transcription factors *GmDof11* (Wang *et al.*, 2007), *GmWRI1a* (Chen *et al.*, 2018), *GmLEC2* (Manan *et al.*, 2017), *GmbZIP123* (Song *et al.*, 2013) and *GmDREBL* (Zhang *et al.*, 2016), and functional genes *GmDGAT1 / GmDAGAT1*, *GmPLD* (lipid hydrolase), *GmLPAT* (lipid synthenase) (Zhao *et al.*, 2013; Chen *et al.*, 2016), *GmGA20OX* (Lu *et al.*, 2016), and *GmSWEET10a/b* (Wang *et al.*, 2020). Interestingly, the effect of gene *AtPDAT* on seed oil content was in conflict in previous studies. While no significant changes of total lipid content or fatty acid composition of the leaves and seeds are observed in overexpressing *AtPDAT* plants or the knockout *PDAT1* mutant in *Arabidopsis* (Ståhl *et al.*, 2004; Mhaske *et al.*, 2005), Fan *et al.* (2013) observed an increase of free fatty acids (FFAs) in *Arabidopsis* mutant with the knockout *PDAT1*, which plays a critical role in mediating triacylglycerols (TAG) synthesis and thereby protecting against FFA-induced cell death in fast-growing tissues. Pan *et al*. (2013) found an increase of *α*-linolenic acid in TAG in yeast and *Arabidopsis* overexpressing the *PDAT* gene. Zhang *et al*. (2009) observed a 70 to 80 % decrease of seed oil content in *Arabidopsis* with RNAi silencing of *PDAT1* in a *dgat1-1* background or *DGAT1* in a *pdat1-1* background. Marmon *et al*. (2017) found the compensation of *PDAT1* on TAG synthesis in the absence of *DGAT1* in cotyledons with a change in fatty acid composition but without reducing seed oil content in *Camelina sativa*. A more direct piece of evidence in forward and reverse genetics shows that *DGAT1* is a major determinant in oil accumulation (Xu *et al*., 2018a). Based on the above inconsistent functions of *PDAT1* and *DGAT1* in fatty acid biosynthesis in transgenic *Arabidopsis* (Chen *et al*., 2016), it is necessary to investigate the biological function of *GmPDAT* in soybean oil biosynthesis.

Soybean seed size is a major trait in breeding, as it is not only a component of seed yield but also an important morphological trait (Xu *et al*., 2011). There have been many reports of the genetic foundation and molecular mechanism of seed size related traits. Approximately 92 QTN regions (Contreras-Soto *et al*., 2017; Fang *et al*., 2017; Shen *et al*., 2018) and 388 QTLs for seed size-related traits in soybean have been documented at Soybase (www.soybase.org). A large number of differentially expressed genes (DEGs) associated with seed development (Zhang *et al*., 2009) were detected by differential expression analysis and comparative genomics analysis. Evolution and association studies have demonstrated the effect of *GmCYP78A10* on soybean seed size/weight (Wang *et al*., 2015). Some genes have been cloned and verified by transgenic experiments, i.e., *SBT1.1* in *Medicago truncatula* and pea (D'Erfurth *et al*., 2012); *GmMYB73* (Liu *et al*., 2014) and *GmGA20OX* (Lu *et al*., 2016) in *Arabidopsis thaliana*; *GmFAD3* (Singh *et al*., 2011), *GmCYP78A72* (Zhao *et al*., 2016), *BIG SEEDS1* (Ge *et al*., 2016), *GmWRKY15a* (Gu *et al*., 2017), *PP2C-1* (Lu *et al*., 2017), *GmLEC2* (Manan *et al*., 2017), *GmCYP78A*5 (Du *et al*., 2017), and *GmSWEET10a/b* (Wang *et al*., 2020) in soybean. Jako *et al*. (2001) showed that seed-specific over-expression of the *DGAT* cDNA in wild-type *Arabidopsis* enhanced oil deposition and increased average seed weight. Zhang *et al*. (2009) observed the overlapping functions of *PDAT1* and *DGAT1* in seed development of *Arabidopsis*. Motivated by the above findings, we decided to investigate the genetic effects of *GmPDAT* and *GmDAGAT1* on seed oil content and sizes in soybean.

Population stratification has long been recognized as a confounding factor in genome-wide association studies (GWAS). To correct for the effects of population stratification, several studies have explored the problem of how to control population structure variation. First, Pritchard *et al*. (2000) proposed a Bayesian-model-based method to infer population structure from multilocus genotype data, and named it STRUCTURE. Falush *et al*. (2003) modified the STRUCTURE approach using the linkage model and the F model. Both models can significantly improve the quality of the population structure inference. However, inferring population structure in large modern datasets imposes severe computational challenges. Thus, Raj *et al*. (2014) developed an approximated STRUCTURE method under a variational Bayesian framework. In addition, Tang *et al*. (2005) presented an extension of maximum likelihood method to estimate individual admixture, namely FRAPPE. The full maximum likelihood method increases its robustness when compared to an existing partial maximum likelihood approach. Thereafter, Alexander *et al*. (2009) developed a fast model-based estimation of ancestry in unrelated individuals, and named it ADMIXTURE. Although ADMIXTURE adopts the likelihood model embedded in STRUCTURE, it runs considerably faster. Clearly, the above methods do not provide formal significance tests for population differentiation. Thus, Patterson *et al*. (2006) combined principal component analysis with Tracy-Widom theory to investigate population structure. Meanwhile, Price *et al*. (2006) used principal components analysis to explicitly model ancestry differences between cases and controls in order to avoid spurious associations. However, it is difficult to correct the differences of oil- and size-related traits across wild, landraces and bred soybeans (evolutionary types). If the population structure is replaced by the above evolutionary types, then GWAS works well in this study. Thus, we proposed the evolutionary population structure model in GWAS.

To focus on the genetic effects of *GmPDAT* and *GmDAGAT1* on seed oil- and size-related traits in soybean, we measured six oil-related traits and four seed size traits in 286 soybean accessions over four years, and incorporated evolutionary population structure model into GWAS to investigate the association of variants on the whole genome. The gene *GmPDAT* is responsible for the changes of seed oil- and size-related traits, and its biological function was confirmed in transgenic soybeans. In addition, we also observed some signs of the interaction between *GmPDAT* and *GmDAGAT1*, although such evidence should be addressed in the future.

Materials and methods

Genetic populations, planting, genotyping and phenotyping

A total of 14 wild, 153 landrace and 119 domesticated soybeans from six geographic regions in China, were planted at the Jiangpu Field Experiment Station of Nanjing Agricultural University, Nanjing, China, with a complete randomized design with three replicates in 2011, 2012, 2014 and 2015. The plots were 1.5 m wide and 2 m long. All the 286 soybean accessions were genotyped by sequencing and a total of 106013 SNPs were identified (Zhou *et al*., 2015a). Twenty seeds from five plants in the middle row of each plot were measured for seed length (SL), seed width (SW) and seed thickness (ST) using digital vernier calipers in 2011, 2012, 2014 and 2015. The 100-seed weight (100SW) trait was measured from 100 dried samples. For each accession, the SL, SW, and ST traits were averaged based on 20 seeds and 100SW for each accession was averaged based on three replicates. Six seed oil-related traits, including seed oil content, palmitic, stearic, oleic, linoleic and linolenic acids, were measured by GC (gas chromatography) analysis. Details of the experiment were described in Lisec *et al*. (2006) and Zuo *et al*. (2019).

We planted 171 recombinant inbred lines (RILs) from the orthogonal cross and 227 RILs from the reciprocal cross of LSZZH (P_1) with NN493-1 (P_2) in three-row plots with a completely randomized design at the Jiangpu Experiment Station of Nanjing Agricultural University in 2015. We genotyped for 11846 SNP markers and phenotyped for five TAGs in soybean seeds. The design and management of field experiment were the same as those in the above-mentioned GWAS. The TAGs levels were measured by LC-MS analysis (Gao *et al*., 2017).

Single- and multi-trait GWAS methods

Only the SNPs, with minor allele frequency (MAF) ≥ 0.05 and missing rate < 0.1, were used in the GWAS. Single-trait GWAS was conducted using GEMMA (Zhou & Stephens, 2012; http://www. xzlab.org/software.html), EMMAX (Kang et al. 2010; https://genome.sph.umich.edu/wiki/EMMAX) and mrMLM (Wang *et al.*, 2016; https://cran.r-project.org/web/packages/mrMLM.GUI/index.html). With the multi-locus GWAS, we treated the three evolutionary types (wild, landrace and bred soybeans) as population structure, and incorporated the structural effects into the mixed linear model of GWAS, and the critical LOD scores for suggested and significant QTNs were set as 2.50 and 3.0, respectively (Zhang *et al.*, 2019). Multi-trait GWAS was conducted using GEMMA (Zhou & Stephens, 2014). The critical P-value for significant QTNs in the GEMMA and EMMAX methods was determined by Bonferroni correction: α = 1/54,294 = 1.84e-05 (Xu *et al.*, 2018b; Zhang *et al.*, 2019). The first VanRaden kinship matrix, proposed by VanRaden (2008) and implemented via the GAPIT software (http://zzlab.net/GAPIT), was used in the mrMLM method.

All the SNP markers on Chr13 with MAF ≥ 0.05 and missing rate < 0.1 were used to detect epistasis for seed oil- and size-related traits using PEPIS (Zhang *et al.*, 2016; http://bioinfo.noble.org// PolyGenic_QTL//Home.gy). The critical LRT value for significant interaction was set at 9.21.

Mapping QTLs for five TAGs in soybean seed

171 (orthogonal) and 227 (reciprocal) RILs from the cross between LSZZH and NN493-1 were genotyped by SLAF-seq technology (Zuo *et al.*, 2019) and measured by BIOTREE (http://www.biotree.cn/) for five TAGs [TAG (20:1/18:3/18:3), TAG (18:3/18:3/18:3), TAG (14:0/18:3/18:3), TAG (20:2/18:2/18:2) and TAG (16:2/18:2/18:3)] in the soybean pods harvested at 45 days after flowering (DAF) in 2015. All the 11,846 markers were assigned by the MSTmap program into twenty soybean chromosomes (Zuo *et al.*, 2019). The GCIM and ICIM methods were used to detect QTLs for these TAGs using "QTL.gCIMapping.GUI" (Wang *et al.*, 2016; Zhang *et al.*, 2020) and "QTL IciMapping V4.1" (Li *et al.*, 2007), respectively. The orthogonal and reciprocal crosses were viewed as covariate and integrated into the genetic model of QTL mapping. The scanning step was set at 1 cM. The critical LOD scores for significant and suggested QTLs were set at 3.0 and 2.5, respectively.

Identification of candidate genes for seed oil- and size-related traits in soybean

As described by Liu et al. (2020), candidate genes for seed oil-related traits were mined in three steps. First, we found all the genes between the 100 kb upstream and downstream regions around each significant QTN. Then, we downloaded the KEGG annotation (https://soycyc.soybase.org/) and the soybean metabolic pathway database (https://soycyc.soybase.org/), and identified the genes or their *Arabidopsis* homologous genes that were annotated with fatty acid biosynthesis and seed size related pathways. Finally, four cultivated soybeans (accession nos. 101, 236, 257 and 276), with high seed oil content (20.08 ± 1.95 (%)) and large seed (16.36 ± 3.60 (g)), and two wild soybeans (accession nos. 265 and 272), with low seed oil content (13.22 ± 1.87 (%)) and small seed (2.45 ± 0.34 (g)), were used to conduct RNA-seq analysis (Zhou *et al*., 2015a). The genes, with the annotations of fatty acid biosynthesis and seed size related pathways and with differential expression levels between four cultivated and two wild soybeans, were selected as candidate genes.

Genetic variants of GmPDAT and GmDAGAT1

The *GmPDAT* and *GmDAGAT* sequences were derived from four genomic sequences of two wild soybeans (W05 and PI483463), one landrace (Williams 82, v1.1) and one cultivar (ZH13), which were downloaded from Soybase (https://www.soybase.org/). The variants of the two genes across four genomes were obtained from the MUSCLE alignment via Genious v4.8.5 software (Kearse et al. 2012). Some variants for each gene were further confirmed by the frequent changes of SNP alleles in 62 wild, 110 landrace and 130 improved soybeans in Zhou *et al*. (2015b).

Development of transgenic plants

Soybean total RNA was isolated from Williams 82 using the trizol reagent (Invitrogen, Foster city, CA, USA) according to the manufacturer's instructions and the RNAs were treated with DNase I (Promega). The first-strand cDNA was then synthesized using M-MLV reverse transcriptase (Promega). The open reading frames (ORFs) of *GmPDAT* (*Glyma.13g108100*) were amplified from the cDNA of Williams 82 by regular PCR using gene-specific primers (Table S1), and subcloned into the pMD-19 T vector (TaKaRa, Japan) for sequence verification. The verified *GmPDAT* sequence was then cloned into the dicotyledon expression vector pBWA(V)BS-ccdB vector (Biorun, China, http://www.biorun.net/), which contains a selection marker gene, *bar* (bialaphos resistance gene), using the ClonExpress Entry One

Step Cloning Kit. Then a 161-bp fragment of *GmPDAT* was amplified using a forward primer *GmPDAT*-6F containing Eco32I restriction sites, and a reverse primer *GmPDAT*-6R containing Eco32I restriction sites. A 200-bp hairpin construct fragment cloned from *GmPDAT* was connected to the two 161-bp fragments. This PCR product was also cloned into pBWA(V)BS-ccdB. The recombinant pCAMBIA3300-*GmPDAT* and pCAMBIA3300-RNAi- *GmPDAT* vectors were transformed into Williams 82 [W82 (wild-type) background] via the *Agrobacterium tumefaciens*-mediated (*GV3101*) method (Zeng *et al*., 2004). The transgenic lines were selected on MS medium containing 50 mg L^{-1} kanamycin, bar gene PCR (543bp), and LibertyLink strip detection (to detect the bar protein) (Invitrogen, Foster city, CA, USA). For LibertyLink strip detection, a total of 100 mg leaf tissue was collected and ground completely in the bottom of a conically tapered 1.5 ml tube by pestle rotation, followed by the addition of 0.5 mL of extraction buffer and a strip into the tube. After 5 minutes, those with two lines (control and test lines) were positive for bar gene expression (Gao *et al*., 2015). The T_2 homozygous transgenic lines (confirmed by the bar gene PCR, bar protein detection and target gene real-time qPCR) were used for further analysis.

Expression analysis

The total RNA was isolated from seed of the T_2 soybean transgenic lines and Williams 82 at 35 DAF. The extracted RNA was then treated with RNase-free DNase I (Promega, USA). The RevertAid first strand cDNA synthesis kit (Thermo Fisher Scientific, USA) was used to synthesize the cDNA. The first-strand cDNA mix was used as the template for RT-qPCR (Reverse Transcription - quantitative Polymerase Chain Reaction). Specific primers (Table S1) were designed based on the assembled sequences of the *GmPDAT* and *GmDAGAT1* genes. The cDNA was used as a template for RT-qPCR using SYBR premix Ex taq (Takara; http://www.takara.com). The procedure was: step 1 95°C for 3 min, step 2 95°C for 10s, step 3 55°C for 20s, step 4 72°C for 20s, step 5 75°C for 5s + Plate Read, step 6 GOTO step 2, 40 times and step 7 Melt Curve 65.0 to 95°C, increment 0.5°C, 5s + Plate Read. Reactions were run on a Bio-Rad CFX96 system. A soybean *GAPDH* (Glyceraldehyde-3-phosphate dehydrogenase) gene was amplified and used as the control in this experiment (Table S1).

Luciferase complementation image assays for the GmPDAT-by-GmDAGAT1 interaction in Nicotiana benthamiana *cells*

The LCI assays for the interaction of *GmPDAT* with *GmDAGAT1* in *Nicotiana benthamiana* cells were described in Zhang *et al*. (2018). The procedure includes material preparation, RNA and DNA isolation, primer design (Table S1) and conditions for PCR, luciferase complementation image assays, and detection of interactions *in vivo*.

Cytological analysis

To investigate the seed cell size of the *GmPDAT* transgenic lines, seeds at 45 DAF and the mature seeds were examined by serial paraffin sections. For each line, three seeds were sampled from the six *GmPDAT* transgenic representative individuals and the controls, respectively. The seeds were soaked in distilled water for 16 to 20 h at room temperature, and then all the seeds were fixed in 50% FAA solution (formalin-acetic acid-alcohol solution). After fixation for 20 to 24 h, using 60%, 70% and 80% ethanol gradient to dehydrate the FAA solution for 1 h, and using 95% and 100% gradient ethanol to dehydrate for 30 min and repeated one more time. The detailed experimental procedures were described in previous studies (Yang *et al*., 2006, 2016). Semi-thin (2.5 µm) sections were obtained using an automatic microtome (Microm HM 360, Thermo), stained with 0.1% toluidine blue O for 30–60 sec at room temperature, examined with a Nikon Eclipse 80i microscope (Nikon, Japan) and photographed under the "bright field" condition. The pictures were softened by Photoshop CS6.

Microscopy analysis

To investigate the seed oil body size of the *GmPDAT* transgenic lines, the mature seeds were examined by scanning electron microscope analyses. For each line, the seeds were first put in the FAA Fixative. Pumped with a vacuum pump until they sank to the bottom. After 2 h at room temperature, they were transferred to a 4 ℃refrigerator, before being washed three times (15 min each) in 0.1 M PBS. Then, the seeds were post-fixed with 1% OsO_4 in 0.1 M PBS (pH 7.4) for 5 h at room temperature. After removal of the OsO_4, they were rinsed three times (15 mins each) in 0.1 M PBS (pH 7.4). The blocks were then washed, dehydrated through an ethanol series of 30–100%, and embedded in EMbed 812 media. The detailed experimental procedures were described in previous studies (Zhang *et al*., 2019). Slices (60-80 nm) sections were obtained using an ultramicrotome (Leica UC7, Germany). Lastly, sections were stained in uranyl acetate in pure ethanol for 15 min, rinsed with distilled water, and then stained with lead citrate for 15 min, before rinsing with distilled water. Sections were air-dried overnight at room temperature, then photographed under a TEM (HT7700, Hitachi, Japan). The pictures were softened by Photoshop CS6.

Monte Carlo simulation studies

The simulated datasets of Wang *et al*. (2016) were partly adjusted in this study in order to investigate the effect of soybean evolutionary types on QTN detection. The simulation is briefly described here; for technical detail readers are referred to the original study (Wang *et al*., 2016). The SNP genotypes were derived from the *A. thaliana* datasets (Atwell *et al*., 2010), all the SNPs between 11226256 and 12038776 bp on Chr1, between 5045828 and 6412875 bp on Chr2, between 1916588 and 3196442 bp on Chr3, between 2232796 and 3143893 bp on Chr4, and between 19999868 and 21039406 bp on Chr5 were used to conduct simulation studies. The sample size was 199, the number of lines in Atwell *et al*. (2010). Six QTNs were simulated and placed on the SNPs with an allele frequency of 0.30. The sizes of the six QTNs, measured as the proportion of phenotypic variance contributed by QTNs, were set at 0.10, 0.05, 0.05, 0.15, 0.05 and 0.05, respectively. The residual variance was set at 10.0. All simulated data sets are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.sk652. The phenotypes were simulated from the following model $\mathbf{y} = \boldsymbol{\mu} + \sum_{i=1}^{6} x_i b_i + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathrm{MVN}_n(\mathbf{0}, 10 \times \mathbf{I}_n)$.

To simulate the population structure of wild, landrace and bred soybeans in this study, the total phenotypic average of 10.0 in Wang *et al*. (2016) was adjusted into three averages of evolutionary types: 10.0 + *c* × SD (standard deviation) for bred soybean (from the first to 100th individuals), 10.0 for landraces (from the 120th to 199th individuals), and 10.0 − *c* × SD for wild soybean (from the 101th to 119th individuals), *c* = 0.5, 1.0 and 1.5, and SD = 4.6064. Each sample was analyzed by the mrMLM method. The results for determining the number of sub-populations in the matrices *Q* calculated from the markers related to the trait and all the markers are listed in Table S2, and the Q matrices are listed in Table S3. For each simulated QTN, we counted the samples in which the LOD statistic exceeded 3.0. A detected QTN within 1 kb of the simulated QTN was considered a true QTN. The ratio of the number of such samples to the total number of replicates (1000) represented the empirical power of this QTN. The Type I error was calculated as the ratio of the number of false positive effects to the total number of zero effects considered in the full model. To measure the bias of QTN effect estimate, mean squared error (MSE) was defined as $\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_i - \beta)^2$, where $\hat{\beta}_i$ is the estimate of β for each QTN in the *i*th sample, and *N* is the number of significant QTNs.

Results

Distributions and comparisons of seed oil- and seed size-related traits across wild, landrace and bred soybeans

Frequent distributions of seed oil- and size-related traits showed that these traits are typical quantitative traits with large variation (**Figs.** S1 & S2; **Table S4**). Based on evolutionary types, all the accessions were grouped into wild, landrace and bred soybeans. Thus, analysis of variance was used to compare the differences across the evolutionary types. Among 38 *F* tests, 35 were significant, indicating the existence of an evolutionary population structure (E) in soybean (Tables 1 & S5).

Detection of QTNs and their candidate genes for seed oil- and size-related traits in soybean

Single-trait GWAS and their candidate genes In Zhou *et al*. (2015a), all the 286 soybean accessions have been genotyped. In this study, we detected the marker-trait associations on the whole genome using the mrMLM, GEMMA and EMMAX methods. These results are shown in Figs. S3~S5. To reduce the false positive rate, in this study we focused on the loci, identified commonly by the three approaches, across various traits. Among these results, fifty-seven QTNs were commonly found by the three methods (Table S6).

Within the range of approximately 100 kb for each significant QTN, there were 1320 genes. Using the KEGG annotation (https://soycyc.soybase.org/) and the soybean metabolic pathway database (https://soycyc.soybase.org/), fifty-seven genes were found to be annotated with lipids-synthesis- or seed-size-related pathways (Table S6). Among the fifty-seven genes, nine were the same with those in Zhang *et al*. (2016). Among the nine genes, *Glyma.13g108100/GmPDAT*, *Glyma.06g111100* and *Glyma.06g039200* had significantly higher expression levels (P=0.004~0.041) in cultivated soybeans than in wild soybeans (Table S6). In this study, we focused on the gene *GmPDAT*. The reasons are as follows. First, *PDAT1* in transgenic *Arabidopsis* had inconsistent functions in fatty acid biosynthesis (Chen *et al*., 2016). Second, *GmPDAT* was found to be associated with soybean seed oil-related traits in our previous study (Liu *et al*., 2020).

GmPDAT is closely linked to the locus from 20676541 to 20704079 bp on Chr13 in this study, and this locus was significantly associated with oleic acid, linolenic acid, linoleic acid, seed thickness, seed length, seed width, and 100-seed weight in single-trait GWAS (*P*=9.63e-10~1.63e-05; LOD=4.27~15.94) (Fig. 1; Table 2). Thus, we suggest that *GmPDAT* is responsible for seed oil- and size-related traits.

In addition, the above locus is linked to the locus at 20532852 bp in a 200 kb region. The latter was found by GEMMA (all the years) and EMMAX (2014) to be associated significantly with linolenic acid (*P*=8.28e-09~1.58e-06; Table 2). Around this locus, there were 34 genes. In the KEGG annotation (https://soycyc.soybase.org/) and the soybean metabolic pathway database (https://soycyc.soybase.org/), we found the function of *GmDAGAT1* (*Glyma.13g106100*) to be diacylglycerol acyltransferase 1, while *Glyma.13g106300* was annotated with seed size related pathways, which are hypersensitive to ABA (Table S7). *DAGAT1* is shown to be related to seed oil content and seed development in *Arabidopsis* as reported previously. In this study, analysis of the RNA-sequences at 15, 25, 35 and 55 DAF identified *GmDAGAT1* to be differentially expressed between wild and cultivated soybeans, and the two candidate genes had significantly higher expression levels in cultivated soybeans than in wild soybeans (P=0.013; Tables S7 & S8), while *Glyma.13g106300* had significantly higher expression levels in cultivated soybeans than in wild soybeans (P-value = 0.043; Table S7).

The expression patterns of *GmDAGAT1* and *GmPDAT* at five stages in wild and cultivated soybeans were similar to those of the genes *GmWRI1a* and *GmbZIP123*, but different from those of genes *GmWRI1a*, *GmNFYA* and *GmBZR1* (**Fig. S4**). Recently, we also found the associations of *GmPDAT* and *GmDAGAT1* with soybean seed oil-related traits (Liu *et al*., 2020) and TAGs (see Discussion). Thus, *GmPDAT and GmDAGAT1* are considered as candidate genes for soybean lipid metabolism, while *Glyma.13g106300* is considered as a candidate gene for seed size related traits.

Multi-trait GWAS To further validate the results of the single-trait GWAS, a multi-trait GWAS was performed using the genome-wide efficient mixed model association (GEMMA) (Zhou & Stephens, 2014). The locus at 20532852 bp was found to be associated simultaneously with linolenic acid, ST and 100SW (P=9.45e-10~5.72e-06), and the locus between 20704034 and 20719806 bp was found to be associated simultaneously with oleic acid, oil content, seed length and width in some cases (Fig. 1b; Table S9). In summary, the above two loci on Chr13 were found to be associated with seed oil- and size-related traits. The results further confirmed the results from single-trait GWAS.

Canonical correlation analysis of seed oil- and seed size-related traits

Canonical correlation between the oil- and size-related traits was significant (P-values < 0.0001) and the coefficients were between 0.607 and 0.754, indicating the correlation of seed oil-related traits with seed size-related traits (Table S10). Meanwhile, in simple correlation analysis between each pair of seed oil- and size-related traits, the stable correlations of seed linolenic acid with seed width, seed thickness, 100-seed weight and seed length, and those of seed oleic acid with the above last three seed size-related traits support the associations of the loci at 20532852 and 20704079 (bp) simultaneously with seed oleic acid, linolenic acid, seed width, seed thickness, 100-seed weight and seed length (Table S11). The pleiotropic effects of the above two QTNs are the genetic foundation of the correlation (Tables 2 & S11).

Domesticated evidences of candidate genes for seed oil- and size-related traits To further confirm the two candidate genes of seed oil- and size-related traits, we analyzed the variants of the two genes across wild, landrace and bred soybeans. As a result, five SNPs were found to exist in *GmPDAT*. Among these variants, one was in the upstream of 26 bp in the 5’UTR, one was located in the 5’UTR, and three were located in the 3’UTR (Fig. 2a). Four SNPs located in its UTR were further confirmed in 302 soybean accessions (Zhou *et al.*, 2015b), in which there were large allelic frequency differences between wild and cultivated soybeans (Fig. 2d). In the same way, four variants were found to exist in *GmDAGAT1*. Among these variants, one was located in the upstream of 30 bp in 5'UTR, one was located in the 3'UTR, and two were located in the CDS region (Figs. 2b & 2c). The SNP in the 3'UTR was further confirmed in 302 soybean accessions (Zhou *et al.*, 2015b), in which there were large allelic frequency difference between wild and cultivated soybeans (Fig. 2d). Clearly, the two genes are involved in domestication. As we know, there are significant differences of seed oil- and size-related traits among the three kinds of soybeans. Thus, the above domestication evidence of the two genes support the associations of the two genes with seed oil- and size-related traits.

GmPDAT and GmDAGAT1 expression in transgenic soybeans

To identify the function of *GmPDAT*, we performed transgenic experiments by (1) overexpressing the *GmPDAT* production and (2) interfering the *GmPDAT* production under the control of the *CaMV35S* promoter in soybean cultivar Williams 82. We cloned the *GmPDAT* gene in a binary expression vector with two *CaMV35S* promoters. The *GmPDAT* gene and the bar gene (glufosinate resistant gene) were each driven by the 35S promoter. The T_2 seeds from both of three independent lines overexpressing *GmPDAT* (OX-*GmPDAT*) and the three independent lines with RNAi *GmPDAT* (RNAi-*GmPDAT*) were confirmed by the bar gene PCR, bar protein detection (**Fig.** S7) and target gene real-time qPCR (**Fig.** 3). Table S12 shows that the expression level of *GmPDAT* was 5.93~6.81, 0.80~1.29, and 2.68 in OX-*GmPDAT*, RNAi-*GmPDAT* and the control lines (wild-type, WT), respectively. The expression level is significantly higher in OX-*GmPDAT* than in WT lines (P=1.66e-04~3.26e-03) and the expression level is significantly lower in RNAi-*GmPDAT* than in WT lines (P= 5.38e-06~8.24e-05) (**Fig.** 3a). The expression level of *GmDAGAT1* was 1.84~2.39, 1.67~2.00, and 1.91 in OX-*GmPDAT*, RNAi-*GmPDAT* and WT lines, respectively, and the differences are not significant between OX-*GmPDAT* and WT lines (P=0.06~0.18) and between RNAi-*GmPDAT* and WT lines (P=0.07~0.21) (**Fig.** 3b, Table S12). In other words, the expression of *GmDAGAT1* did not change while the expression level of *GmPDAT* changed. This means that *GmPDAT* may not regulate *GmDAGAT1.*

We measured the seed oil content, the five fatty acids, and all seed size related traits (SL, SW, ST and 100SW) for the T_2 seeds of the above-mentioned *GmPDAT* transgenic, WT and negative control lines. The results are listed in Table S6. Seed oil contents in OX-*GmPDAT*, RNAi-*GmPDAT* and WT lines are 194.10~222.53, 140.03~160.89 and 177.86 mg/g, respectively; the seed linoleic acids in OX-*GmPDAT*, RNAi-*GmPDAT* and WT lines were 106.99~120.56, 72.26~91.02 and 95.93 mg/g, respectively; the seed linolenic acids in OX-*GmPDAT*, RNAi-*GmPDAT* and WT lines are 14.90~18.64, 9.40~11.98, and 13.02 mg/g, respectively. Although no significant trait difference was observed between WT and the negative control lines, seed oil content, seed linoleic acid and seed linolenic acid were significantly higher in OX-*GmPDAT* than in WT lines and significantly lower in RNAi-*GmPDAT* than in WT lines (**Fig.** 3e, Table S12).

The trait values of 100SW are 19.75~19.97, 17.54~18.15 and 18.46 g, respectively, for OX-*GmPDAT*, RNAi-*GmPDAT* and WT lines, indicating significantly larger values in OX-*GmPDAT* lines than in WT lines (P=2.27e-03~0.01) and almost significantly smaller in RNAi-*GmPDAT* lines than in WT lines (P=0.02~0.07) (**Fig.** 3d). In terms of SL, the trait values are 7.82~8.12, 7.24~7.27 and 7.32 mm, respectively, for OX-*GmPDAT*, RNAi-*GmPDAT* and WT lines, being very significantly longer in OX-*GmPDAT* lines than in WT lines (P=2.07e-11~7.2e-11) and no significant difference between RNAi-*GmPDAT* and WT lines (P=0.07~0.15). The values of trait SW in OX-*GmPDAT*, RNAi-*GmPDAT* and WT lines are 6.87~7.05, 6.48~6.54 and 6.59 mm, respectively, indicating very significantly higher values in OX-*GmPDAT* lines than in WT lines (P=2.97e-7~8.37e-7) and no significant difference between RNAi-*GmPDAT* and WT lines (P=0.06~0.22). Finally, the values of the ST trait are 5.47~5.64, 5.02~5.38 and 5.54 mm, for OX-*GmPDAT*, RNAi-*GmPDAT*, and WT lines, respectively, indicating no significant difference between OX-*GmPDAT* and WT lines (P=0.17~0.19) but significantly lower in RNAi-*GmPDAT* lines than in WT lines (P=2.3e-07~0.01) (**Fig.** 3f; Table S12).

Comparison of the E + K model with the others in GWAS using Monte Carlo simulation studies

The purpose of the simulation studies is to demonstrate the advantage of the evolutionary population structure (E) model, namely the E + K model. To simulate the soybean evolutionary population structure, we first set the trait mean at 10.0 in the first simulated dataset of Wang *et al*. (2016) and then adjusted three averages of the wild, landrace and bred soybeans, being 10.0 + *c* × standard deviation (SD), 10, and 10.0 − *c* × SD, respectively, where SD = 4.6064 and *c* takes three different values: 0.5, 1.0 and 1.5. The number of individuals is 1~100 for bred soybean, 120~199 for landraces and 101~119 for wild soybean. Under each level of *c*, all the simulations were replicated 1000 times. Each sample was analyzed under five models: (1) only polygenic background control (the K model), (2) only evolutionary population structure control (the E model), (3) population structure Q from all the markers related to the trait at the 0.05 level of significance and polygenic background controls (the Q_1 + K model), (4) population structure Q from all the markers and polygenic background controls (the Q_2 + K model), and (5) evolutionary population structure and polygenic background controls (the E + K model). The power of QTN detection, mean squared error (MSE) of QTN effect, and false positive rate (FPR) were used to measure the effect of various genetic models, and these results are listed in Table S13. To demonstrate the objective of the simulation experiments, the E + K model was compared with all the other models. In the case of *c*=0.5, as a result, the average powers of QTN detection over the six QTNs are 54.55, 38.53, 38.95, 54.08, and 67.30 (%), respectively, for the above five models; the average MSEs of QTN effects are 0.1206, 0.2003, 0.1231, 0.1410, and 0.0936, respectively; the FPRs are 0.0212, 0.1037, 0.0272, 0.0210, and 0.0176 (%), respectively (Fig. 4). Using paired *t* (or *u*) test, the E + K model has almost significantly higher statistical powers, lower MSEs, and lower false positive rate than the others (Table 3), indicating the best performances of the E + K model. The same trends are also observed in the cases of *c*= 1.0 and 1.5. The evolutionary population structure method was also used to re-analyzed a famous rice grain width dataset in Wang et al. (2018). As a result, five additionally known genes, along with two genes detected by Wang et al. (2018), were identified (Fig. S8; Table S14). Therefore, the E + K model is the best one.

Discussion

Two significant advances have been reported in this study. First, the biological function of the gene *GmPDAT* is reported for the first time in soybean. As described above, *GmPDAT* plays a critical role in the increase of seed oil- and size-related traits in the lines overexpressing *GmPDAT* and in the decrease of seed oil- and size-related traits in RNAi-*GmPDAT* lines. In *Arabidopsis*, meanwhile, the biological functions of the gene *AtPDAT* in Ståhl *et al*. (2004), Mhaske *et al*. (2005) and Marmon *et al*. (2017) are significantly different from those in Zhang *et al*. (2009), Fan *et al*. (2013) and Pan *et al*. (2013). Second, simulated and real data analyses showed that the E + K model can increase the statistical power of QTN detection in GWAS as compared with the currently-adopted Q + K model. This will provide a new approach to mine new genes in animal, plant and human genetics.

Detection of pleiotropic genes in a small region

As described by Kroymann & Mitchell-Olds (2005), our knowledge of complex trait variation is limited to QTL of relatively large effects, and it is very difficult to detect small QTNs, especially in a small region. Here we present an outstanding example on how to detect pleiotropic genes in a small region. First, we replaced the Q matrix by the E matrix in the mixed linear model of GWAS. If the *Q* matrix is incorporated into the mixed linear model of our multi-locus GWAS methods, no QTNs were identified in this small region. Then, we used several single-trait GWAS analyses, and multi-trait analyses to cross-examine these QTNs in the small region. In this study, two loci were repeatedly identified by the above approaches to be associated with seed oil- and size-related traits. Subsequent genomic function analysis, and expression level analysis in OX-*GmPDAT*, RNAi-*GmPDAT* and control lines helped us to reliably determine whether the two loci were truly associated with seed oil- and size-related traits in soybean. Recently, another pleiotropic and domestication gene

GmSWEET10a/b for seed size, oil content and protein content in soybean has been reported (Wang *et al*., 2020).

Molecular mechanism of pleiotropic effect of GmPDAT for seed oil- and size-related traits in soybean

Mapping QTLs for five TAGs in soybean RILs To understand the mechanism of pleiotropic *GmPDAT* on seed oil-related traits, we genotyped 171 (orthogonal) and 227 (reciprocal) RILs from the cross between LSZZH and NN493-1 for 11846 SNPs, and measured phenotypic values of five TAGs on seeds 45 DAF (Fig. S9; Table S15). These datasets were used to identify QTLs for the five TAGs using GCIM and ICIM. As a result, one QTL around *GmDAGAT1* and two QTLs around *GmPDAT* were identified (Fig. S10a~c; Table S14).

Histocytological comparison of GmPDAT transgenic soybean with its wild-type Because seed integument size influences final seed size, which is determined by cell proliferation and cell expansion, the outer integument cells were observed in the seeds of WT Col-0 and *GmPDAT* overexpressing lines. The cell lengths and widths increased 15.7% and 9.0% in the mature seeds, and 14.7% and 22.0% in the premature seeds (45-DAF), respectively, in OX-30 transgenic lines than in WT Col-0 lines. However, the cell lengths and widths decreased 25.2% and 23.5% in the mature seeds, and 9.0% and 12.8% in the premature seeds (45-DAF), respectively, in RNAi-16 transgenic lines than in WT Col-0 lines. The result indicates that the cell size of OX-*GmPDAT* seeds is significantly larger than that of WT Col-0 seeds (Fig. S10d~s), which is consistent with the study by Hirshfield *et al*. (1993).

Plants often store oils in the form of TAGs in seeds and fruits (Carlsson *et al*., 2011; Lu *et al*., 2011), and the TAGs are known to be produced by *DGAT* and *PDAT* (Zhang *et al*., 2009); Zhang *et al*. (2009) proved in *Arabidopsis* that *DGAT1* encodes a major enzyme contributing to the accumulation of seed TAGs, acting together mainly with *PDAT*. These results are consistent with our QTL mapping results, which show the two QTLs (20573761 and 20606048 bp; 20620950 to 20775650 bp) are very close to the candidate genes *GmDAGAT1* and *GmPDAT*. In our GWAS and *GmPDAT* transgenic soybean results, *GmDAGAT1* and *GmPDAT* are associated with seed oil- and size-related traits. In our histocytological results, the seeds of lines overexpressing *GmPDAT* are larger than their wild type equivalent. We also compared the oil bodies (OBs) of OX, WT and RNAi seeds using transmission electron microscopy. In the mature seeds, the OBs of OX seeds showed typically spherical and ovoid structures and were distributed mostly between protein bodies (Fig. S10t). Moreover, OX seed cells contained apparently bigger OBs than WT and RNAi, and the seed cells in RNAi-16 appear to contain the smallest OBs (Fig. S10u). Therefore, *GmPDAT* increased TAGs and eventually increased the seed oil. More oil requires a large space to store the oil and thus leads to increased cell length and cell width in the seeds of overexpressing *GmPDAT* lines.

The E + K model increases the statistical power of QTN detection

Population structure is widely used in GWAS to control p*opulation* stratification. Generally speaking, it is expressed by the matrix Q, which is frequently calculated from some or all the markers. In the Q matrix, each element is a posteriori probability that the *i*th accession belongs to the *k*th subpopulation. In this study, the population structure is replaced by an evolutionary population structure. In the evolutionary population structure (the E matrix), each element (e_{ik}) is composed of one or zero. If the *i*th accession belongs the second evolutionary type, so $e_{i2} = 1$ and $e_{ik} = 0$ for $k \neq 2$.

Wild, landrace and bred soybeans are three evolutionary types and significant differences across these types exist for seed oil- and size-related traits (Table S4). Although the *Q* matrix calculated from all the markers can be used to correct these population structure differences, the results are unsatisfactory in this study. Thus, we replaced the *Q* matrix by the E matrix in GWAS. Its advantages have been demonstrated via Monte Carlo simulation studies and real data analysis. The possible reason is that the E model reflects the evolutionary process and the differences across various evolutionary types. Thus, more information is obtained from the E + K model. After the evolutionary population structure effects are removed, in statistics, the intra-allelic variance is smaller than that when the evolutionary population structure effects are ignored, as shown below,

$$\sigma^2 = \sum_{i=1}^{3} w_i(\mu_i^2 + \sigma_i^2) - \mu^2$$

where μ_i, σ_i^2 and w_i are the means, the variances and the proportions of the wild ($i=1$), landrace ($i=2$) and domesticated ($i=3$) soybeans, and μ is the overall mean. The above larger intra-allelic variance can significantly decrease the statistical power of QTN detection. This may be one reason for the low power in QTN detection for complex traits.

As we know, the *Q* matrix is calculated from genome-wide marker information. The purpose is to capture all genes for the trait of interest. However, we do not know where these genes are located. Therefore, one selects all markers potentially associated with the trait of interest to calculate the *Q* matrix. In genomic selection, He *et al*. (2019) compared the predicted accuracies from four marker sets, being all the unique QTNs, statistically stable QTNs, stable and large-effect QTNs and all the related markers. They found that the predicted accuracy is as high as 0.92 when all unique QTNs are used. This implies that the current method of calculating the *Q* matrix is not optimal to some extent and may explain why the two loci were not detectable when the *Q* matrix was used. More importantly, the results from the Q_1 + K model are not better than those from the E + K model.

Asian, white and black ethnic groups are three evolutionary types of humans and their complex traits and diseases have significant differences across groups (Table S16). In previous GWAS studies, these evolutionary differences were not fully controlled, although the *Q* matrix has been adopted. Evolutionary population structure exists in all species and incorporation of such structure effects into the GWAS model is necessary in animal, plant and human genetics. The evolutionary population structure is suitable for all the populations having evolutionary population structures. If all individuals in the GWAS population have a similar evolutionary type, the *Q* matrix correction may be sufficient.

Some signs of the epistasis between GmPDAT and GmDAGAT1

The PEPIS program was used to detect the epistatic effect between the above two loci (chr13-20532852 and chr13-20719806) for seed oil- and size-related traits collected. Significant epistatic effects were detected and they are responsible for the variation in linolenic acid (LRT= 24.37, 2014; LRT=11.52, 2012), and seed thickness (LRT=10.80, 2011).

To further validate the above interaction, the interaction analysis for the candidate genes *GmDAGAT1* and *GmPDAT* was performed. Molecular structure analysis of gene interaction based on the program STRING (https://string-db.org//) indicates the existence of epistasis; the statistic was 0.75, which was significantly larger than the medium confidence of 0.40. More importantly, luciferase complementation image assays (LCI) validated the protein interaction between *GmPDAT* and *GmDAGAT1*. With the co-infiltration of recombinant strain combination N-*GmPDAT* + C-*GmDAGAT1* into *N. benthamiana* fresh leaves, we observed a strong fluorescent signal in the epidermal cells, compared with those from control groups, suggesting *GmPDAT* and *GmDAGAT1* can be combined together *in vivo* (**Fig.** S11).

Conclusions

Two QTNs for seed oil- and size-related traits in soybean were identified in a small interval using an evolutionary population structure model. Around the two QTNs, two candidate genes *GmPDAT* and *GmDAGAT1* have been proved to be domestication genes. The biological function of *GmPDAT* is reported for the first time to be associated with acyl-lipid metabolism in soybean, while t**he biological function of candidate gene** *Glyma.13g106300* for seed size-related traits will be investigated in the future. Although there are some signs of the interaction between *GmPDAT* and *GmDAGAT1*, **this evidence should be addressed in the future. This study** provides a new approach for identifying pleiotropic genes.

Data availability

The following data are available at the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.sk652.

Acknowledgements

We thank Prof. Shizhong Xu at University of California, Riverside and Prof. Jim M. Dunwell at University of Reading for help with improvements to the English text. This work was supported by the National Natural Science Foundation of China (31571268 and 31871242), Huazhong Agricultural University Scientific & Technological Self-Innovation Foundation (2014RC020) and the Fundamental Research Funds for the Central Universities (2662020ZKPY017) to Y.M.Z.

The authors declare that they have no conflict of interest.

Author contributions: YMZ conceived of the project and its components. JYL and JFZ performed field experiments. YWZ, JYL, JFZ, XH and YMZ conducted the Monte Carlo simulation experiments, bioinformatics analysis and real data analysis. JYL and JFZ performed experimental LCI arrays. ZZ and HS conducted histocytological experiment. YMZ, JYL and QS wrote and revised the manuscript. All authors reviewed the manuscript.

References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research **19**, 1655-1664.

Atwell S, Huang YS, Vilhjálmsson BJ, *et al*. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature **465**, 627-631.

Cao Y, Li S, Wang Z, Chang F, Kong J, Gai J, Zhao T. 2017. Identification of major quantitative trait loci for seed oil content in soybeans by combining linkage and genome-wide association mapping. Frontiers in Plant Science **8**, 1222.

Carlsson AS, Yilmaz JL, Green AG, Stymne S, Hofvander P. 2011. Replacing fossil oil with fresh oil – with what and for what? European Journal of Lipid Science and Technology **113**, 812-831.

Chen B, Wang J, Zhang G, Liu J, Manan S, Hu H, Zhao J. 2016. Two types of soybean diacylglycerol acyltransferases are differentially involved in triacylglycerol biosynthesis and response to environmental stresses and hormones. Scientific Reports **6**, 28541.

Chen L, Zheng Y, Dong Z, Meng F, Sun X, Fan X, Zhang Y, Wang M, Wang S. 2018. Soybean (*Glycine max*) WRINKLED1 transcription factor, *GmWRI1a*, positively regulates seed oil accumulation. Molecular Genetics and Genomics **293**, 401-415.

Chung J, Babka HL, Graef GL, Staswick PE, Lee DJ, Cregan PB, Shoemaker RC, Specht JE. 2003. The seed protein, oil, and yield QTL on soybean linkage group I. Crop Science **43**, 1053-1067.

Contreras-Soto RI, Mora F, de Oliveira MA, Higashi W, Scapim CA, Schuster I. 2017. A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. PLoS One **12**, e0171105.

D'Erfurth I, Le Signor C, Aubert G, *et al*. 2012. A role for an endosperm-localized subtilase in the control of seed size in legumes. New Phytologist **196**, 738-751.

Du J, Wang S, He C, Zhou B, Ruan YL, Shou H. 2017. Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. Journal of Experimental Botany **68**, 1955-1972.

Eskandari M, Cober ER, Rajcan I. 2013. Genetic control of soybean seed oil: I. QTL and genes associated with seed oil concentration in RIL populations derived from crossing moderately high-oil parents. Theoretical and Applied Genetics **126**, 483-495.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164**, 1567-1587.

Fan J, Yan C, Xu C. 2013. Phospholipid: diacylglycerol acyltransferase-mediated triacylglycerol biosynthesis is crucial for protection against fatty acid-induced cell death in growing tissues of *Arabidopsis*. The Plant Journal **76**, 930-942.

Fang C, Ma Y, Wu S, *et al*. 2017. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biology **18**, 161.

Gao L, Ding XN, Li K, *et al*. 2015. Characterization of *soybean mosaic virus* resistance derived from inverted repeat-SMV-*HC-Pro* genes in multiple soybean cultivars. Theoretical and Applied Genetics **128**, 1489–505.

Gao XK, Zhang S, Luo J.Y, Lü LM, Zhang LJ, Cui JJ. 2017. Lipidomics and RNA-seq study of lipid regulation in *Aphis gossypii* parasitized by *Lysiphlebia japonica.* Scientific Report ***7****, 1364.*

Ge L, Yu J, Wang H, Luth D, Bai G, Wang K, Chen R. 2016. Increasing seed size and quality by manipulating *big seeds1* in legume species. *Proc*eedings of the National *Acad*emy of *Sci*ences of the United States of America **113**, 12414-12419.

Gu Y, Li W, Jiang H, Wang Y, Gao H, Liu M, Chen Q, Lai Y, He C. 2017. Differential expression of a *wrky* gene between wild and cultivated soybeans correlates to seed size. Journal of Experimental Botany **68**, 2717-2729.

Han Y, Li D, Zhu D, Li H, Li X, Teng W, Li W. 2012. QTL analysis of soybean seed weight across multi-genetic backgrounds and environments. Theoretical and Applied Genetics **125**, 671-683.

He L, Xiao J, Rashid KY, Yao Z, Li P, Jia G, Wang X, Cloutier S, You FM. 2019. Evaluation of genomic prediction for pasmo resistance in flax (*Linum usitatissimum* L.). Frontiers in Plant Science **9**, 1982.

Hirshfield KM, Flannery RL, Daie J. 1993. Cotyledon cell number and cell size in relation to seed size and seed yield of soybean. Plant Physiology and Biochemistry **31**, 395-400.

Hwang EY, Song Q, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB. 2014. A genome-wide association study of seed protein and oil content in soybean. BMC Genomics **15**, 1.

Jako C, Kumar A, Wei Y, Zou J, Barton D.L, Giblin E.M, Covello PS, Taylor DC. 2001. Seed-specific over-expression of an *Arabidopsis* cDNA encoding a diacylglycerol acyltransferase enhances seed oil content and seed weight. Plant Physiology **126**, 861-874.

Kang H M, Sul J H, Service S K, Zaitlen N A, Kong S Y, Freimer N B, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics **42**, 348-354.

Kearse M, Moir R, Wilson A, *et al.* 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**, 1647-1649.

Kroymann J, Mitchell-Olds T. 2005. Epistasis and balanced polymorphism influencing complex trait variation. Nature **435**, 95-98.

Lander ES, Kruglyak L. 1995. Genetic dissection of complex traits guidelines for interpreting and reporting linkage results. Nature Genetics **11**, 241-247.

Li H, Ye G, Wang J. 2007. A modified algorithm for the improvement of composite interval mapping. Genetics **175**, 361-374.

Li YH, Zhou G, Ma J, *et al.* 2014. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nature Biotechnology **32**, 1045-1052.

Li-Beisson Y, Shorrosh B, Beisson F, *et al.* 2013. Acyl-lipid metabolism. Arabidopsis Book **11**, e0161.

Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR. 2006. Gas chromatography mass spectrometry- based metabolite profiling in plants. Nature Protocols **1**, 387-396.

Liu JY, Li P, Zhang YW, Zuo JF, *et al.* 2020. Three-dimension genetic networks among seed oil-related traits, metabolites and genes reveal the genetic foundations of oil synthesis in soybean. The Plant Journal **103**, 1103-1124.

Liu YF, Li QT, Lu X, *et al.* 2014. Soybean *GmMYB73* promotes lipid accumulation in transgenic plants. BMC Plant Biology **14**, 73.

Lu C, Napier JA, Clemente TE, Cahoon EB. 2011. New frontiers in oilseed biotechnology: meeting the global demand for vegetable oils for food, feed, biofuel, and industrial applications. Current Opinion in Biotechnology **22**, 252-259.

Lu X, Li Q.T, Xiong Q, *et al.* 2016. The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. The Plant Journal **86**, 530-544.

Lu X, Xiong Q, Cheng T, *et al.* 2017. A *pp2c-1* allele underlying a quantitative trait locus enhances soybean 100-

seed weight. Molecular Plant **10**, 670-684.

Manan S, Ahmad MZ, Zhang G, Chen B, Haq BU, Yang J, Zhao J. 2017. Soybean *LEC2* regulates subsets of genes involved in controlling the biosynthesis and catabolism of seed storage substances and seed development. Frontiers in Plant Science **8**, 1604.

Marmon S, Sturtevant D, Herrfurth C, Chapman K, Stymne S, Feussner I. 2017. Two acyltransferases contribute differently to linolenic acid levels in seed oil. Plant Physiology **173**, 2081-2095.

Martin C, Butelli E, Petroni K, Tonelli C. 2011. How can research on plants contribute to promoting human health? The Plant Cell **23**, 1685-1699.

Mhaske V, Beldjilali K, Ohlrogge J, Pollard M. 2005. Isolation and characterization of an *Arabidopsis thaliana* knockout line for phospholipid: diacylglycerol transacylase gene (*At5g13640*). Plant Physiology and Biochemistry **43**, 413-417.

Niu Y, Xu Y, Liu XF, Yang SX, Wei SP, Xie FT, Zhang YM. 2013. Association mapping for seed size and shape traits in soybean cultivars. Molecular Breeding **31**, 785-794.

Pan X, Siloto RM, Wickramarathna AD, Mietkiewska E, Weselake RJ. 2013. Identification of a pair of phospholipid: diacylglycerol acyltransferases from developing flax (*Linum usitatissimum* L.) seed catalyzing the selective production of trilinolenin. The Journal of Biological Chemistry **288**, 24173-24188.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genetics **2**, e190.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38**, 904-909.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics **155**, 945-959.

Raj A, Stephens M, Pritchard JK. 2014. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. Genetics **197**, 573-589.

Rambaut A. 2012. FigTree_v1.4.0 2012. Available from: http://tree.bio.ed.ac.uk/software/figtree/.

Schindelin J, Rueden CT, Hiner MC, Eliceiri KW. 2015. The ImageJ ecosystem: An open platform for biomedical image analysis. Molecular Reproduction and Development **82**, 518-529.

Schmutz J, Cannon SB, Schlueter J, *et al*. 2010. Genome sequence of the palaeopolyploid soybean. Nature **463**, 178-183.

Shen Y, Liu J, Geng H, Zhang J, Liu Y, Zhang H, Xing S, Du J, Ma S, Tian Z. 2018. *De novo* assembly of a Chinese soybean genome. Science China Life Sciences **61**, 871-884.

Singh AK, Fu DQ, El-Habbak M, Navarre D, Ghabrial S, Kachroo A. 2011. Silencing genes encoding omega-3 fatty acid desaturase alters seed size and accumulation of bean pod mottle virus in soybean. Molecular Plant Microbe Interactions **24**, 506-515.

Song Q, Yan L, Quigley C, *et al*. 2017. Genetic characterization of the soybean nested association mapping population. Plant Genome **10**, doi: 10.3835/plantgenome2016.10.0109.

Song QX, Li QT, Liu YF, *et al*. 2013. Soybean *GmbZIP123* gene enhances lipid content in the seeds of transgenic *Arabidopsis* plants. Journal of Experimental Botany **64**, 4329-41.

Ståhl U, Carlsson AS, Lenman M, Dahlqvist A, Huang B, Banas W, Banas A, Stymne S. 2004. Cloning and functional characterization of a phospholipid: diacylglycerol acyltransferase from *Arabidopsis*. Plant Physiology **135**, 1324-1335.

Sun YN, Pan JB, Shi XL, *et al*. 2012. Multi-environment mapping and meta-analysis of 100-seed weight in soybean. Molecular Biology Reports **39**, 9435-9443.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. Genetic Epidemiology **28**, 289-301.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science **91**, 4414-4423.

Wang HW, Zhang B, Hao YJ, Huang J, Tian AG, Liao Y, Zhang JS, Chen SY. 2007. The soybean Dof-type transcription factor genes, *GmDof4* and *GmDof11*, enhance lipid content in the seeds of transgenic *Arabidopsis* plants. The Plant Journal **52**, 716-729.

Wang S, Liu S, Wang J, Yokosho K, Zhou B, Yu YC, Liu Z, Frommer WB, Ma JF, Chen LQ, Guan Y, Shou H, Tian Z. 2020. Simultaneous changes in seed size, oil content, and protein content driven by selection of SWEET homologues during soybean domestication. National Science Review, DOI: 10.1093/nsr/nwaa110.

Wang SB, Feng JY, Ren WL, Huang B, Zhou L, Wen YJ, Zhang J, Dunwell JM, Xu S, Zhang YM. 2016. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Scientific Report **6**, 19444.

Wang SB, Wen YJ, Ren WL, Ni YL, Zhang J, Feng JY, Zhang YM. 2016. Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. Scientific Report **6**, 29951.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557: 43–49.

Wang X, Jiang GL, Green M, Scott RA, Song Q, Hyten DL, Cregan PB. 2014. Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. Molecular Genetics and Genomics **289**, 935-949.

Wang X, Li Y, Zhang H, Sun G, Zhang W, Qiu L. 2015. Evolution and association analysis of *GmCYP78A10* gene with seed size/weight and pod number in soybean. Molecular Biology Reports **42**, 489-496.

Xie M, Chung CY, Li MW, *et al*. 2019. A reference-grade wild soybean genome. Nature Communications **10**, 1216.

Xu Y, Caldo KMP, Pal-Nath D, Ozga J, Lemieux MJ, Weselake RJ, Chen G. 2018a. Properties and biotechnological applications of Acyl-CoA: diacylglycerol Acyltransferase and Phospholipid: diacylglycerol Acyltransferase from terrestrial plants and microalgae. Lipids **53**, 663-688.

Xu Y, Li HN, Li GJ, Wang X, Cheng LG, Zhang YM. 2011. Mapping quantitative trait loci for seed size traits in soybean (*Glycine max* L. Merr.). Theoretical and Applied Genetics **122**, 581-594.

Xu Y, Yang T, Zhou Y, Yin S, Li P, Liu J, Xu S, Yang Z, Xu C. 2018b. Genome-wide association mapping of starch pasting properties in maize using single-locus and multi-locus models. Frontiers in Plant Science **9**, 1311.

Yang Y, Gehrke S, Imai Y, Huang Z, Ouyang Y, Wang JW, Yang L, Beal MF, Vogel H, Lu B. 2006. Mitochondrial pathology and muscle and dopaminergic neuron degeneration caused inactivation of *Drosophila* Pink1 is rescued by Parkin. *Pro*ceedings of the National *Acad*emy of *Sci*ences of the United States of America **103**, 10793-10798.

Yang Y, Shi J, Wang X, Liu G, Wang H. 2016. Genetic architecture and mechanism of seed number per pod in rapeseed: elucidated through linkage and near-isogenic line analysis. Scientific Report **6**, 24124.

Yu J, Zhang Z, Wei J, Ling Y, Xu W, Su Z. 2014. SFGD: a comprehensive platform for mining functional information from soybean transcriptome data and its use in identifying acyl-lipid metabolism pathways. BMC Genomics **15**, 271.

Yuan L, Li R, Mao X, Zhao K, Sun Y, Ji C, Li R. 2017. Spatio-temporal expression and stress responses of *DGAT1*, *DGAT2* and *PDAT* responsible for TAG biosynthesis in *Camelina sativa.* Emirates Journal of Food and Agriculture **29**, 274-284.

Zeng P, Vadnais DA, Zhang Z, Polacco JC. 2004. Refined glufosinate selection in *Agrobacterium*-mediated transformation of soybean [*Glycine max* (L.) Merrill]. Plant Cell Rep **22**, 478-482.

Zhang D, Zhang H, Hu Z, Chu S, Yu K, Lv L, Yu D. 2019. Artificial selection on *GmOLEO1* contributes to the increase in seed oil during soybean domestication. PLoS Genetics **15**, e1008267.

Zhang L, Liu JY, Gu H, *et al*. 2018. *Bradyrhizobium diazoefficiens* USDA 110-*Glycine max* interactome provides candidate proteins associated with symbiosis. Journal of Proteome Research **17**, 3061-3074.

Zhang L, Wang SB, Li QG, Song J, Hao YQ, Zhou L, Zheng HQ, Dunwell JM, Zhang YM. 2016. An integrated bioinformatics analysis reveals divergent evolutionary pattern of oil biosynthesis in high- and low-oil plants. PLoS One **11**, e0154882.

Zhang M, Fan J, Taylor DC, Ohlrogge JB. 2009. *DGAT1* and *PDAT1* acyltransferases have overlapping functions in *Arabidopsis* triacylglycerol biosynthesis and are essential for normal pollen and seed development. The Plant Cell **21**, 3885-3901.

Zhang W, Dai X, Wang Q, Xu S, Zhao PX. 2016. PEPIS: A pipeline for estimating epistatic effects in quantitative trait locus mapping and genome-wide association studies. PLoS Computational Biology **12**, e1004925.

Zhang YM, Jia Z, Dunwell JM. 2019. Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. Frontiers in Plant Science **10**, 100.

Zhang YQ, Lu X, Zhao FY, Li QT, Niu SL, Wei W, Zhang WK, Ma B, Chen SY, Zhang JS. 2016. Soybean *GmDREBL* increases lipid content in seeds of transgenic *Arabidopsis*. Scientific Report **6**, 34307.

Zhang YW, Wen YJ, Dunwell JM, Zhang YM. 2020. QTL.gCIMapping.GUI v2.0: An R software for detecting small-effect and linked QTLs for quantitative traits in bi-parental segregation populations. Computational and Structural Biotechnology Journal **18**, 59-65

Zhao B, Dai A, Wei H, Yang S, Wang B, Jiang N, Feng X. 2016. *Arabidopsis* KLU homologue *GmCYP78A72* regulates seed size in soybean. Plant Molecular Biology **90**, 33-47.

Zhao JZ. 2013. Phospholipase gene *GmPLD* and lipid synthase genes *GmDGAT* and *GmLPAT* play important role in regulating *Arabidopsis* seed oil content and growth [D]. Nanjing Agricultural University.

Zhou L, Wang SB, Jian J, *et al*. 2015a. Identification of domestication-related loci associated with flowering time and seed size in soybean with the RAD-seq genotyping method. Scientific Reports **5**, 9350.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed model analysis for association studies. Nature Genetics **44**, 821-824.

Zhou X, Stephens M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods **11**, 407-409.

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, *et al*. 2015b. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nature Biotechnology **33**, 408-414.

Zuo JF, Niu Y, Cheng P, Feng JY, Han SF, Zhang YH, Shu G, Wang Y, Zhang YM. 2019. Effect of marker segregation distortion on high density linkage map construction and QTL mapping in soybean (*Glycine max* L.). Heredity **123**, 579-592.

Table 1. ANOVA for seed oil- and size-related traits across wild, landrace and bred soybeans

Year	Palmitic acid (%)		Stearic acid (%)		Oleic acid (%)		Linoleic acid (%)		Linolenic acid (%)	
	F-value	P-value	F-value	P-value	F-value	P-value	F-value	P-value	F-value	P-value
2011	8.11**	3.4e-4	20.41**	6.5e-09	21.68**	2.2e-09	7.11**	9.9e-4	74.13**	9.8e-26
2012	31.36**	8.2e-13	5.19**	6.2e-3	22.01**	1.70e-09	2.89^{NS}	0.057	92.28**	2.29e-30
2014	1.46^{NS}	0.23	14.99**	7.40e-07	27.16**	2.39e-11	9.72**	8.69e-05	73.67**	1.31e-25
2015	1.28^{NS}	0.27	16.85**	1.42e-07	50.03**	7.44e-19	10.02**	6.63e-05	126.57**	4.02e-38
	Oil content (%)		Seed length (mm)		Seed width (mm)		Seed thickness (mm)		100-seed weight (g)	
	F-value	P-value	F-value	P-value	F-value	P-value	F-value	P-value	F-value	P-value
2011			71.7**	4.2e-25	100.66**	2.2e-32	132.61**	2.25e-39	42.48**	1.70e-16
2012			70.12**	1.19e-24	106.97**	7.84e-34	123.20**	2.07e-37	45.68**	1.65e-17
2014	40.70**	6.37e-16	86.33**	6.82e-29	117.63**	3.29e-36	127.62**	2.43e-38	42.30**	1.94e-16
2015	40.42**	7.82e-16	77.51**	1.25e-26	99.41**	4.45e-32	116.19**	6.79e-36	49.59**	1.02e-18

NS: no significance at the 0.05 level; * and **: significance at the 0.05 and 0.01 levels, respectively.

Table 2. QTNs from single-trait genome-wide association studies of soybean seed oil- and size-related traits on a small region of chromosome 13 and their candidate genes

Locus	Trait	Single-trait genome-wide association studies				r^2 (%)	Comparative genomics analysis			P-value§
		Chr	Position (bp)	LOD score or P-value	Method (year)		Candidate genes	Arabidopsis homologs	Functional annotation	
1	Linolenic acid	13	20532852	8.28E-09~1.58E-06	GEMMA (all years)	NA	*Glyma.13g106100, GmDAGAT*	*AT2G19450.1*	diacylglycerol acyltransferase 1	0.013*
				4.03E-06	EMMAX (2014)					
2	Linolenic acid	13	20704034	1.39E-06	GEMMA (2014)	NA	*Glyma.13g108100, GmPDAT*	*AT5G13640.1*	triacylglycerol biosynthesis	0.016*
			20704079	1.97E-08						
	Oleic acid	13	20704062	15.94	mrMLM (2011)	0.15	*Glyma.13g108100, GmPDAT*	*AT5G13640.1*	triacylglycerol biosynthesis	0.016*
	Linoleic acid	13	20704062	10.74	mrMLM (2011)	0.02	*Glyma.13g108100, GmPDAT*	*AT5G13640.1*	triacylglycerol biosynthesis	0.016*
	Seed length	13	20704034	5.52E-06	GEMMA (2014)	NA	*Glyma.13g108100, GmPDAT*	*AT5G13640.1*	triacylglycerol biosynthesis	0.016*
	Seed width	13	20704034	5.43E-08~1.78E-06	GEMMA (2014, 2015)	NA	*Glyma.13g108100, GmPDAT*	*AT5G13640.1*	triacylglycerol biosynthesis	0.016*
			20704034	4.04E-07~2.86E-07	EMMAX (2011, 2014)	NA				
	Seed thickness	13	20676541	5.69	mrMLM (2015)	1.87	*Glyma.13g108100, GmPDAT*	*AT5G13640.1*	triacylglycerol biosynthesis	0.016*
			20704034	3.61E-08~2.10E-07	GEMMA (2014, 2015)	NA				
			20704034	9.63E-10~1.63E-05	EMMAX (2014, 2015)	NA				
			20704034	4.54	mrMLM (2014)	1.22				
			20704079	8.47E-06	GEMMA (2014)	NA				
			20704079	7.33E-06	EMMAX (2011)	NA				
	100-seed weight	13	20704034	3.07E-08	GEMMA (2014)	NA	*Glyma.13g108100, GmPDAT*	*AT5G13640.1*	triacylglycerol biosynthesis	0.016*
			20704034	1.40E-08~6.79E-07	EMMAX (2011, 2014, 2015)	NA				
			20704034	4.27~11.15	mrMLM (2014, 2015)	1.25~1.78				
			20704079	3.12E-06	GEMMA (2014)	NA				
			20704079	5.46E-06	EMMAX (2011)	NA				

§: The P-values were calculated using the paired Student's *t*-test from the average RPKM values at four stages between cultivated (high seed oil, $n_1 = 4$) and wild (low seed oil, $n_2 = 2$) soybeans, with the 0.05 level of significance indicated by an asterisk.

Table 3. Paired *t* tests and their P-values of the differences (A–B) between the E + K model (A) and the others (B) for power and mean squared error (MSE) and false positive rate in Monte Carlo simulation studies

Differences among evolutionary populations types	Comparison	Statistical power of QTN detection (%)		MSE of QTN effect		False positive rate (%)	
		t	P-value	*t*	P-value	*u*	P-value
0.5 SD	K	3.75^{*}	1.33e-2	-2.42^{NS}	6.00e-2	-5.78^{***}	7.56e-09
	E	4.76^{**}	5.04e-3	-2.88^{*}	3.44e-2	-78.18^{***}	<1.00e-300
	Q_1+ K	6.45^{**}	1.34e-3	-1.92^{NS}	1.13e-1	-14.34^{***}	1.22e-46
	Q_2+ K	3.62^{*}	1.52e-2	-2.69^{*}	4.32e-2	-5.47^{***}	4.47e-08
1.0 SD	K	4.75^{**}	5.10e-3	-3.17^{*}	2.48e-2	-21.65^{***}	6.69e-104
	E	4.82^{**}	4.80e-3	-2.77^{*}	3.95e-2	-77.94^{***}	<1.00e-300
	Q_1+ K	6.16^{**}	1.64e-3	-2.58^{*}	4.93e-2	-18.22^{***}	3.48e-74
	Q_2+ K	5.12^{**}	3.69e-3	-2.53^{NS}	5.26e-2	-19.22^{***}	2.61e-82
1.5 SD	K	7.55^{***}	6.45e-4	-3.15^{*}	2.54e-2	-30.67^{***}	1.60e-206
	E	4.89^{**}	4.53e-3	-2.79^{*}	3.85e-2	-77.94^{***}	<1.00e-300
	Q_1+ K	7.31^{***}	7.51e-4	-1.97^{NS}	1.05e-1	-27.24^{***}	1.97e-163
	Q_2+ K	7.75^{***}	5.73e-4	-1.97^{NS}	1.05e-1	-26.25^{***}	7.22e-152

E: Evolutionary population structure; Q_1: Q matrix calculated from the markers related to the trait; Q_2: Q matrix calculated from all the markers; K: kinship matrix; NS: no significant at the 0.05 probability level; * and **: significance at the 0.05 and 0.01 probability levels, respectively.

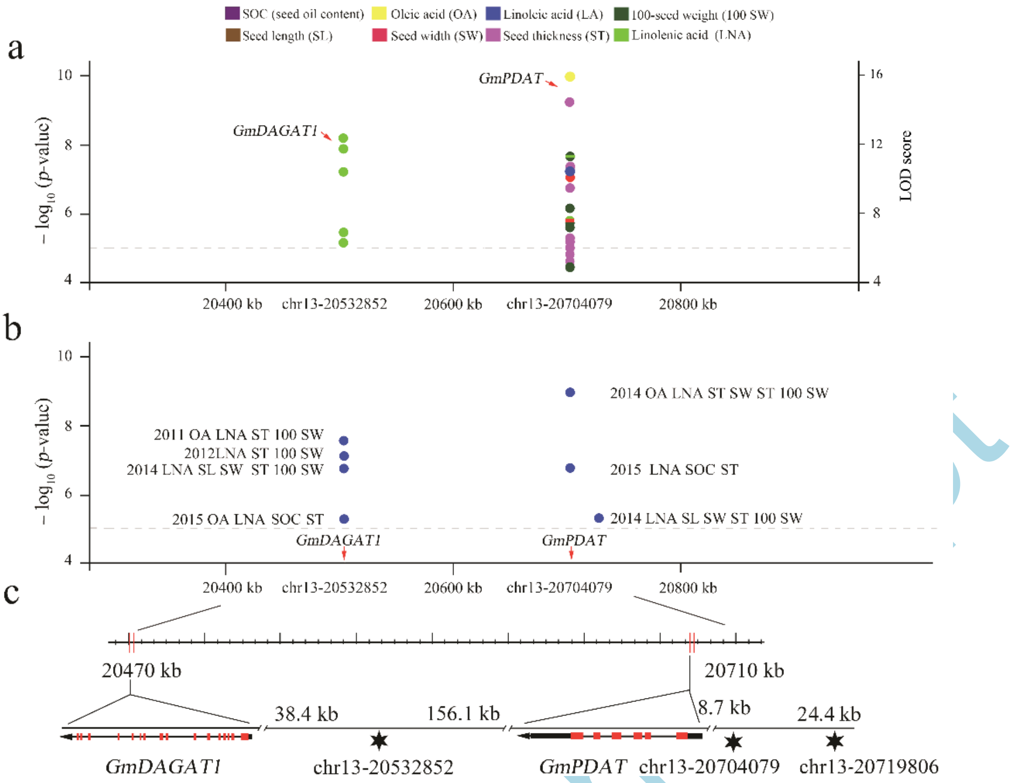
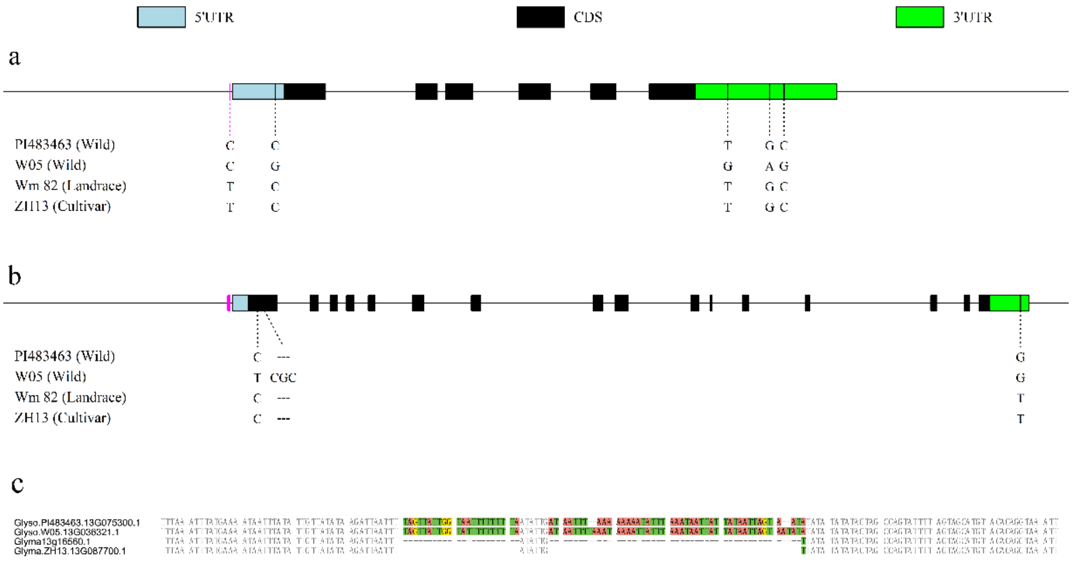


Fig. 1. Significant QTNs (a and b) and their candidate genes (c) for soybean seed oil- and size-related traits on a small region of chromosome 13 using single- (a) and multi-trait (b) genome-wide association studies. The critical P-value of significant QTNs (a and b) was set at 1/*m* where *m* = 54294 and marked by horizontal lines. These QTNs are marked by dots with various colors (a and b). In the structures of *GmPDAT* and *GmDAGAT1*, exons, introns and untranslated regions (UTRs) are indicated by red boxes, thin black lines and black boxes, respectively (c).



d

Gene	SNP information				Allelic frequency in 302 accessions			SNP annotation
	SNP name	Chr	Posi (bp)	Ref	Wild	Landrace	Cultivar	
Glyma13g16790 (*GmPDAT*)	13_20689526	13	20689526	G	0.61	0.91	0.91	3_prime_UTR_variant
	13_20689679	13	20689679	C	0.60	0.91	0.91	3_prime_UTR_variant
	13_20690120	13	20690120	A	0.63	0.91	0.89	3_prime_UTR_variant
	13_20694893	13	20694893	T	0.48	0.91	0.91	5_prime_UTR_variant
Glyma13g16560 (*GmDAGAT*)	13_20486122	13	20486122	A	0.05	0.57	0.60	3_prime_UTR_variant

Fig. 2. The variants of *GmPDAT* (a) and *GmDAGAT* (b and c) across four genomes of two wild soybeans (W05 and PI483463), one landrace (Williams 82, v1.1) and one cultivar (ZH13), and the SNP allele frequencies (d) in 62 wild, 110 landrace and 130 improved soybeans of Zhou *et al*. (2015b). The *GmPDAT* and *GmDAGAT* sequences were derived from four genomic sequences of the above four accessions, which were downloaded from Soybase (https://www.soybase.org/). The variants of the two genes across four genomes were obtained from the MUSCLE alignment of Genious v4.8.5 software (Kearse *et al*. 2012). *GmDAGAT* in genome v1.1 is *GmDAGAT1* in genome v2.0.

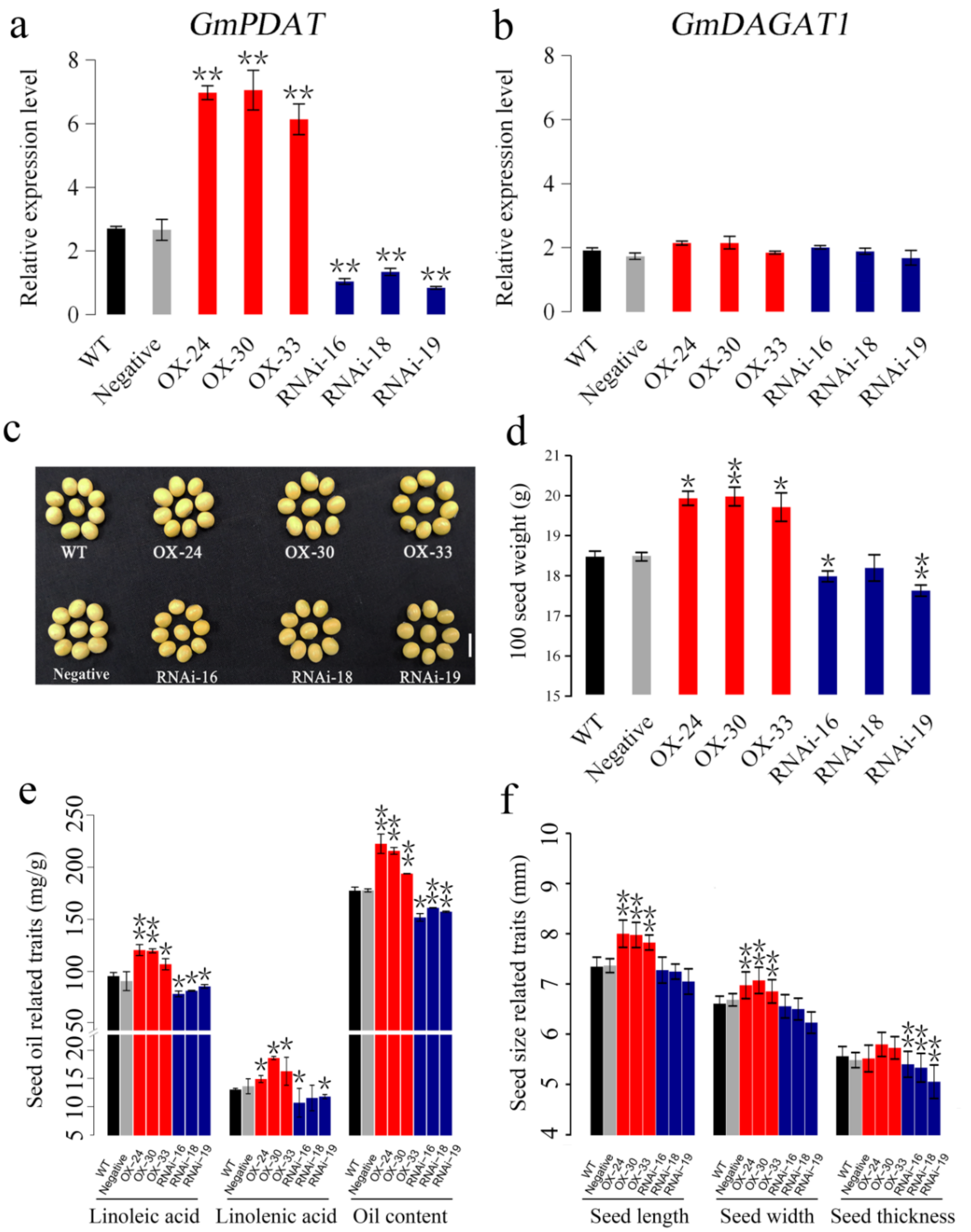


Fig. 3. Real-time PCR analysis (a: *GmPDAT*; b: *GmDAGT1*), seeds (c; scale bar: 10 mm), and their oil- and size-related traits (d~f) of *GmPDAT* transgenic soybean. WT: non-transgenic cultivar Williams 82; negative: control plants transformed with empty vector; OX: *GmPDAT* overexpressed lines; RNAi: RNAi transgenic lines. Seven to eleven seeds from each T_2 transgenic line were used to measure seed oil-related traits, the results from three lines are shown as mean ± standard deviation, and *n* = 3. * and **: the 0.05 and 0.01 levels of significance, respectively. The *t*-test was used to test the significant differences of 100-seed weight (d), linolenic acid, linoleic acid, oil content (e), seed length, width and thickness (f) between OX (or RNAi) and WT. The raw datasets are listed in Supplemental Datasets 1~3.

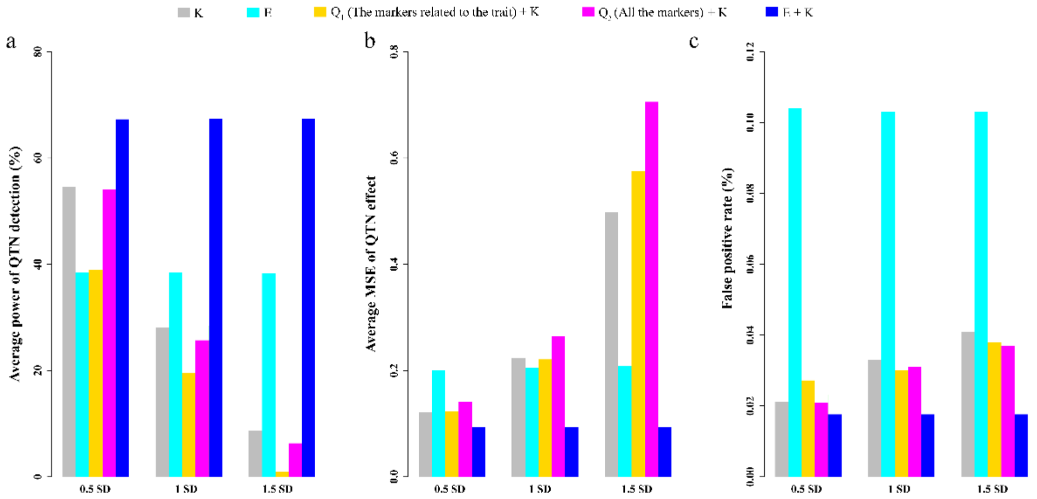


Fig. 4. Comparison of evolutionary population structure (E) with frequently-used population structures in genome-wide association studies under various subpopulation differences (SD). a: average power of QTN detection across six simulated QTNs; b: average mean squared error (MSE) of QTN effects across six simulated QTNs; c: false positive rate. K: kinship matrix; Q: Q matrix using the software STRUCTURE; SD = 4.6064. Sample size was 199, and the number of replicates was set at 1000. The simulated datasets in this study were derived from the first Monte Carlo simulation experiment in Wang *et al*. (2016) Sci Rep 6: 19444.