scientific reports

Check for updates

OPEN

Integration of GWAS and transcriptome and haplotype analyses to identify QTNs and candidate genes controlling oil content in soybean seeds

Kaixin Zhang[1,2,4], Bo Hu[2], Wenhao Wang[2], Wen-Xia Li[2]✉ & Hailong Ning[2,3]✉

The oil content (OC) of soybean is a critical trait with important applications in the development of both food and industrial products. Consequently, enhancing OC has consistently remained a significant objective in soybean breeding programs. In this research, a four-way recombinant inbred line (FW-RIL) population containing 144 lines developed from the cross (Kenfeng14 × Kenfeng15) × (Kenfeng19 × Heinong48) was planted in ten environments to investigate the phenotypic data for OC in seeds. On the basis of the genotype data for 109,676 nonredundant single-nucleotide polymorphism (SNP) markers obtained in previous studies, important quantitative trait nucleotides (QTNs) controlling OC in soybean were identified via five multilocus genome-wide association study (GWAS) methods. A total of 54 significant QTNs were detected by the five methods, including 21 QTNs identified by multiple methods and 2 QTNs detected in two environments. By combining pathway analysis, transcriptome sequencing and gene annotation information, two candidate genes, *Glyma.18G027100* and *Glyma.03G021800*, which are likely related to oil synthesis and metabolism in soybean seeds, were identified, and they were then verified via haplotype analysis. Additionally, 12 of the 23 important QTNs were found for the first time in this study. These results not only contribute to the understanding of the genetic control of OC in soybean seeds but are also helpful in marker-assisted selection (MAS) for breeding high-oil varieties of soybean.

Keywords Soybean, Oil content, Multilocus GWAS, QTNs, Candidate gene

Soybean (*Glycine max*(L.) Merr.) is the second largest source of vegetable oil in the world, and its consumption has increased rapidly as the population has grown, the standard of living has increased, and diets have changed[1]. Owing to the limited availability of agricultural land, increasing the oil content (OC) of soybean seeds has long been a very important goal for breeders.

OC is a typical quantitative trait that is controlled by multiple genes and is affected by the environment. To improve the efficiency of genetic breeding, studies for mapping quantitative trait loci (QTLs) and nucleotides (QTNs) and mining genes associated with OC have been conducted. To date, 315 QTLs and 93 QTNs related to the OC of soybean have been identified in SoyBase (https://soybase.org). Furthermore, 7, 7, 23, 24, 10, 5, 14, and 5 OC QTLs were detected by Cao et al., Zhang et al., Li et al., Yao et al., Huang et al., Zhu et al., Liu et al. and Li et al., respectively[2–9], and 6, 14, 15, 44, and 44 OC QTNs were detected by Li et al., Li et al., Wang et al., Li et al. and Zhao et al., respectively[10–14]. Moreover, researchers have identified several genes related to soybean oil anabolism, such as *GmDof4*, *GmDof11*, *GmbZIP123*, *FAD2-1 A*, *GmMYB73*, *GmDREBL*, *GmNFYA*, *GmZF351*, *B1*, *GmOLEO1*, *GmSWEET39* and *GmPDAT*[15–25]. These studies have enhanced our understanding of soybean lipid metabolism and provided valuable insights for improving soybean OC.

Genome-wide association study (GWAS) has gained significant traction in recent years[26–30], primarily because of their ability to refine QTL intervals at the gene level. Additionally, GWAS can be used for a diverse

[1]Yancheng Institute of Technology, Yancheng 224051, China. [2]Northeast Agricultural University, Harbin 150030, China. [3]Zhongnongfa Wudalianchi Agricultural Technology Co., Ltd, Wudalianchi 164100, China. [4]National Center for Soybean Improvement, National Key Laboratory of Crop Genetics and Germplasm Enhancement, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing, 210095, China. ✉email: liwenxianeau@126.com; ninghailongneau@126.com

Environment	Parents[a]				FW-RILs								
	P1	P2	P3	P4	Mean	SD	Max	Min	Range[b]	Kurt	Skew	Shapiro-Wilk test (*p*-value)	CV[c]
E1	21.01	21.46	18.05	18.11	20.35	0.65	21.80	18.12	3.68	0.78	−0.53	0.53	3.19
E2	21.45	21.67	17.65	17.72	20.52	0.63	22.03	19.07	2.96	−0.13	0.16	0.31	3.07
E3	21.11	22.32	18.65	18.32	20.55	0.66	22.06	18.85	3.21	0.57	−0.13	0.50	3.21
E4	21.33	21.54	18.04	16.82	20.69	0.72	22.31	18.57	3.74	1.92	−0.69	0.07	3.48
E5	21.72	21.85	19.01	18.25	20.74	0.92	22.24	19.06	3.18	0.97	−0.15	0.21	4.44
E6	21.25	22.73	17.65	18.82	20.50	0.88	22.30	18.74	3.56	1.51	0.27	0.70	4.29
E7	20.82	21.55	17.61	16.97	19.71	0.87	22.22	17.58	4.64	−0.41	0.06	0.49	4.41
E8	21.05	21.84	19.19	17.76	19.52	0.75	21.56	17.74	3.82	−0.43	0.17	0.10	3.84
E9	21.38	21.51	19.75	18.09	20.40	1.14	24.11	17.93	6.18	0.38	0.32	0.72	5.59
E10	20.57	21.26	18.94	18.03	19.74	1.03	22.40	16.71	5.69	−0.25	−0.12	0.13	5.22

Table 1. Statistical characteristics of OC in the parents and FW-RIL populations grown in ten environments. [a]Kengfeng14, Kenfeng15, Kenfeng19 and Heinong48 are indicated by P1, P2, P3 and P4, respectively. [b]Range = OCMax – OCMin. [c]CV: coefficient of variation, and the unit is %.

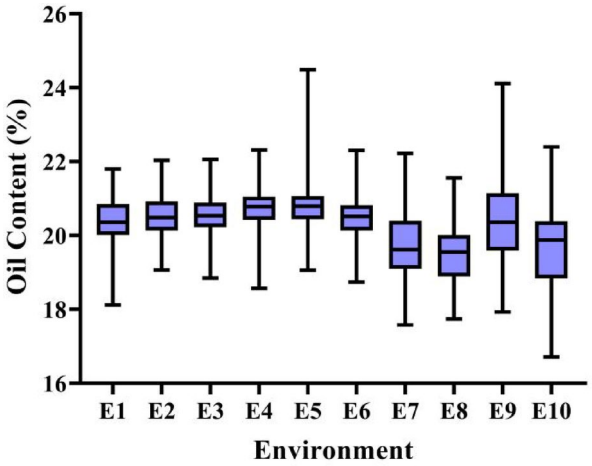

Fig. 1. Phenotypic variations in the OC of the FW-RIL population in 10 environments. The interquartile region, median, and range are represented by the box, the vertical line in the box, and the vertical lines at both ends, respectively.

range of populations, thereby overcoming the population constraints inherent in linkage analysis. Multiparent advanced generation intercross (MAGIC) populations have more diverse alleles than biparental intermated populations do, which could increase the detection power for QTLs and genes. Therefore, an increasing number of plant MAGIC populations have been reported[13,31–38]. This type of population remains amenable to analysis through GWAS, which offers greater precision in the identification of genomic regions and genes than linkage analysis does. With the expanding utilization of GWAS, a growing number of analytical methods have been developed. These methods can be broadly classified into two categories: single-locus GWAS and multilocus GWAS. Compared with single-locus GWAS, multilocus GWAS has more advantages, such as the ability to identify QTNs for small effects and a low false-positive rate, so it has been used extensively[39–43].

In this study, a soybean four-way recombinant inbred line (FW-RIL) population with 144 individuals was phenotyped for OC in 10 different environments. Combined with the SNP genotype data, the significant QTNs for seed OC in soybean were identified via 5 types of multilocus methods. In the linkage disequilibrium (LD) decay region of the common QTNs, the candidate genes involved in seed OC synthesis were selected via transcriptome sequencing and gene annotation analysis and then verified via haplotype analysis. This research could improve our understanding of the genetic control of seed OC in soybean and support marker-assisted selection for breeding soybean with high seed OC.

Results

Phenotype variation analysis

The general descriptive statistics of the OC trait for the 4 parents and the FW-RILs population are presented in Table 1; Fig. 1. For the parental lines, the OCs of Kenfeng19 and Heinong48 were lower than those of Kenfeng14 and Kenfeng15 in all 10 environments. For the FW-RILs, the "Range" (Range = OC_{Max} - OC_{Min}) of the 144 FW-RILs in 10 environments ranged from 2.96 to 6.18%, and the CV range was 3.07–5.99%. In nearly all

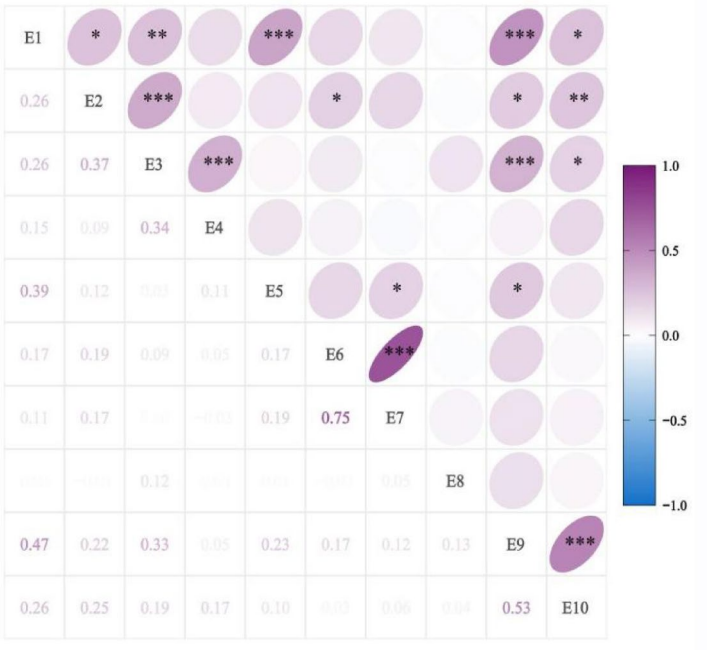

Fig. 2. Correlation analysis of OC in 10 environments. The number at the bottom left represents the corresponding correlation coefficient with the dot at the top right. * represents a significant correlation at the 0.05 level, ** represents a significant correlation at the 0.01 level, and *** represents a significant correlation at the 0.001 level.

Source	DF	SS	MS	F	*P*	Variance component
Environment	9	734.68	81.63	195.50	< 0.0001***	
Genotype	143	542.85	3.80	9.09	< 0.0001***	0.09
Genotype * Environment	1156	1660.05	1.44	3.44	< 0.0001***	0.34
Error	2618	1093.14	0.42			0.42
h^2						0.65

Table 2. Joint ANOVA of the OC of FW-RILs in multiple environments and heritability. ***Indicates significance at the 0.001 level.

environments, the absolute values of skewness and kurtosis were less than 1. Additionally, the Shapiro-Wilk test results demonstrated that *p*-values exceeded 0.05 across all environments. These findings suggest that the OC of the 144 FW-RILs adhered to a normal distribution in these environments (Supplementary Figure S1). The significant positive correlation of OC between different environments revealed that the stable expressed genetic basis among various environments (Fig. 2). The ANOVA results demonstrated significant differences among the genotypes, environments, and their interactions concerning the OC trait (Table 2). This suggests that the population is suitable for identifying QTNs associated with OC, and specific QTNs may be discerned across various environmental conditions. The higher broad-based heritability (0.65) was associated with a greater probability of identifying stable QTNs in multiple environments.

QTNs identified by multilocus GWAS methods

A total of 54 significant QTNs were identified for OC from the five multilocus GWAS methods in 10 environments (Supplementary Table S1). The numbers of significant SNPs detected by pLARmEB, ISIS EM-BLASSO, mrMLM, FASTmrMLM and FAST mrEMMA were 27, 21, 21, 21 and 10, respectively (Fig. 3A), and 22 QTNs could be detected by multiple methods (Table 3), with the PVE by each QTN ranging from 4.22 to 17.68%. In the 10 environments, the numbers of significant QTNs were 6, 6, 9, 9, 2, 9, 4, 3, 3 and 5 (Fig. 3B), and two common QTNs (*qOC-7-1*, *qOC-18-1*) were identified across multiple environments, with the PVE by each QTN ranging from 5.52 to 9.20% (Table 3). These 54 significant QTNs were distributed among 19 chromosomes of soybean, of which Chr09 had the most significant QTNs, with 9, and no significant QTN was detected on Chr12 (Fig. 3C).

A total of 23 common QTNs (Common QTN means the QTN was detected across multiple GWAS methods or in multiple environments) were found in this study (Table 3; Fig. 4). 21 common QTNs were detected by at least two methods in a single environment (listed in black font in Table 3); their LOD values ranged from

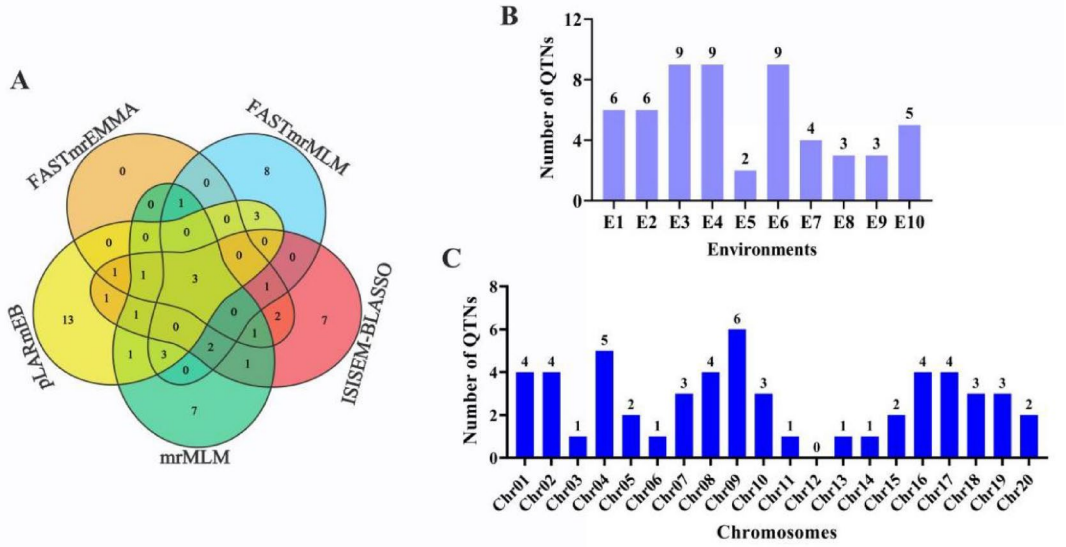

Fig. 3. (A) Venn diagram of the total numbers of significant QTNs detected by 5 multilocus GWAS methods. (B) The total number of significant QTNs detected in 10 environments across the 5 methods. (C) The total number of significant QTNs detected in 10 chromosomes.

QTN	Environment	Lead-SNP	Chromosome	Physical position (bp)	Attenuation intervals (bp)	QTN Effect (%)	LOD score	-LOG10(*P*)	r^2 (%) [a]	Method[b]	Previous research[c]
qOC-1–3	E7	AX-157,228,720	1	35,397,975	35,297,975 – 35,497,975	−0.34	5.18	5.99	14.62	1,2,3,4,5	
qOC-1–4	E9	AX-157,398,461	1	38,536,147	38,436,147 – 38,636,147	−0.44	5.21	6.01	17.68	1,2,3,4,5	
qOC-2-1	E6	AX-157,332,596	2	6,333,154	6,233,154-6,433,154	0.14	4.43	5.20	6.41	1,3,4,5	
qOC-2-2	E3	AX-157,521,230	2	7,177,292	7,077,292-7,277,292	0.29	4.24	5.00	6.44	3,4	58
qOC-2–3	E3	AX-157,296,705	2	8,264,013	8,164,013 – 8,364,013	0.19	4.56	5.34	12.19	1,2,4	58
qOC-3-1	E7	AX-157,128,483	3	2,311,853	2,211,853-2,411,853	0.21	3.24	3.95	4.22	2,4	2,62
qOC-4-2	E7	AX-117,048,711	4	8,800,699	8,700,699-8,900,699	0.41	4.80	5.59	6.79	2,4	
qOC-5-2	E2	AX-157,211,281	5	27,830,491	27,730,491 – 27,930,491	0.17	4.78	5.57	6.73	4,5	60
qOC-7-1	**E6, E7**	**AX-157,110,130**	**7**	**8,326,462**	**8,226,462-8,426,462**	**0.21**	**5.20**	**6.00**	**5.52**	**5**	63
qOC-7-3	E6	AX-157,587,035	7	38,274,432	38,174,432 – 38,374,432	0.19	5.32	6.13	11.57	1,2,4	71
qOC-8-1	E8	AX-116,922,190	8	935,430	835,430-1,035,430	−0.22	4.53	5.31	6.79	2,3,5	
qOC-9-4	E6	AX-157,054,758	9	6,546,371	6,446,371-6,646,371	0.18	4.14	4.90	7.90	1,2,3	
qOC-9-6	E6	AX-157,361,651	9	39,700,404	39,600,404 – 39,800,404	0.15	4.73	5.52	4.18	1,4	65
qOC-10-2	E3	AX-157,263,037	10	19,847,286	19,747,286 – 19,947,286	0.16	5.85	6.68	6.69	1,2,4	61
qOC-11-1	E6	AX-157,229,785	11	7,138,176	7,038,176-7,238,176	−0.13	4.36	5.13	5.35	2,4	60
qOC-14-1	E3	AX-157,557,823	14	48,498,737	48,398,737 – 48,598,737	−0.18	6.38	7.23	9.47	1,2,3,4,5	59
qOC-16-1	E1	AX-157,290,592	16	66,892	1–166,892	0.19	4.04	4.79	6.25	1,4,5	
qOC-16-2	E1	AX-157,224,595	16	6,773,991	6,673,991-6,873,991	−0.32	4.87	5.66	17.11	1,2,5	62
qOC-16-4	E1	AX-157,132,706	16	37,226,749	37,126,749 – 37,326,749	−0.39	5.78	6.61	9.50	1,3,4,5	
qOC-17-3	E8	AX-157,542,830	17	39,971,258	39,871,258 – 40,095,123	0.30	3.99	4.74	10.28	1,2	
qOC-17-4	E6	AX-157,586,271	17	41,441,825	41,341,825 – 41,541,825	0.20	5.08	5.88	10.91	1,3,5	66
qOC-18-1	**E4, E10**	**AX-157,333,937**	**18**	**2,064,407**	**1,860,377-2,164,407**	**−0.36**	**4.83**	**5.61**	**9.20**	**1,3,5**	37
qOC-18-3	E4	AX-157,083,942	18	5,362,484	5,262,484-5,462,484	0.28	7.78	8.66	12.17	3,4,5	

Table 3. Twenty-three common QTNs for seed OC in soybean across different methods or in multiple environments. [a]r^2 (%), proportion of total phenotypic variation explained by each QTN;. [b]mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB and ISIS EM-BLASSO are indicated by 1 to 5, respectively;. [c]Reported study in the region.

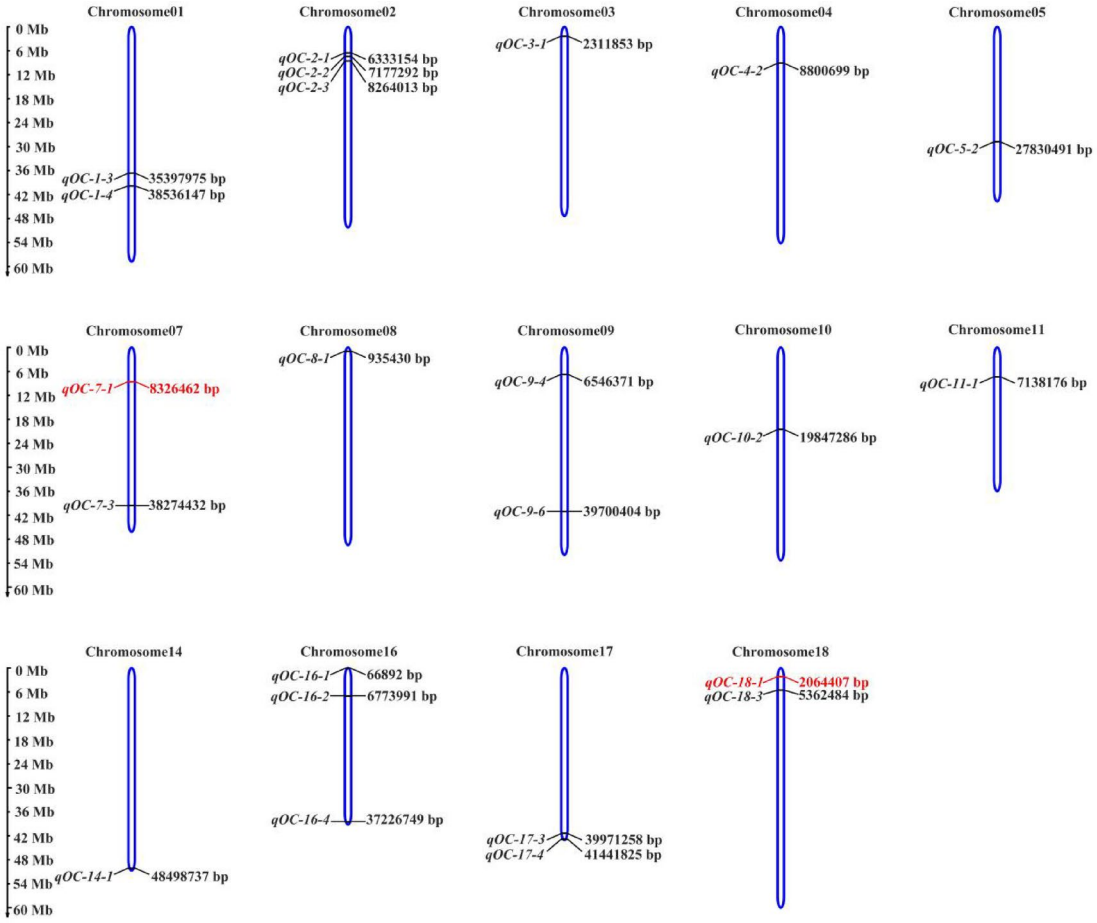

Fig. 4. Distribution of significant QTNs on 14 chromosomes. No common QTN were found in the remaining 6 chromosomes (chromosome 06, 12, 13, 15,19 and 20). The common QTNs detected in multiple environments are shown in red.

3.24 to 7.78, while the proportion of PVE by each QTN ranged from 4.22 to 17.68%. Two common QTNs were identified across multiple environments by single or multiple methods, and detailed information of these two QTNs (*qOC-7-1*, *qOC-18-1*) were shown in black bold font in Table 3; their LOD values ranged from 4.83 to 5.20, and the proportion of PVE by each QTN ranged from 5.52 to 9.20%. For all the 23 common QTNs, the direction of effect (positive or negative) of each QTN was consistent across the different methods or environments.

Transcriptome sequencing analysis and qRT-PCR validation

To identify DEGs among soybean varieties with high and low OC, transcriptome sequencing analysis was conducted on seeds from 25 days after flowering in KS01, KD31, KD48 and KD57. A total of 5.64 billion raw reads were generated from 12 samples, with individual samples ranging from 41.71 to 53.58 million reads. After filtering out low-quality and adapter reads, the number of clean reads in each sample ranged from 41.26 to 52.89 million, accounting for 98.44–99.06% of the raw reads (Supplementary Table S2 and Fig. 5A). The Q30 value and GC content ranges were 91.17–92.85% and 45.49–46.59%, respectively, which reflected the high quality of the transcriptome sequencing data (Supplementary Table S2). In addition, effective reads, after ribosomal RNA removal, were aligned with the reference genome. The results showed that the mapping ratios of the 12 samples ranged from 95.75 to 96.55%. (Supplementary Table S3). Principal component analysis (PCA) revealed that the samples had excellent repeatability and discrimination (Fig. 5B).

In total, 4559 (KD48-vs-KD31), 7013 (KD48-vs-KS01), 3491 (KD57-vs-KD31) and 11,382 (KD57-vs-KS01) DEGs were identified in the comparisons between different groups. There were more upregulated DEGs than downregulated DEGs in both "KD48-vs-KS01" and "KD57-vs-KS01". In "KD48-vs-KD31" and "KD57-vs-KD31", the number of upregulated DEGs was slightly lower than that of downregulated DEGs (Fig. 5C).

To validate the transcriptome sequencing results, we randomly selected 10 DEGs and analyzed their expression levels in KD31, KD57, KS01 and KD48 via qRT-PCR. The results validated the relevance of the transcriptome sequencing data, and the qRT-PCR results were highly consistent (Supplementary Figure S2).

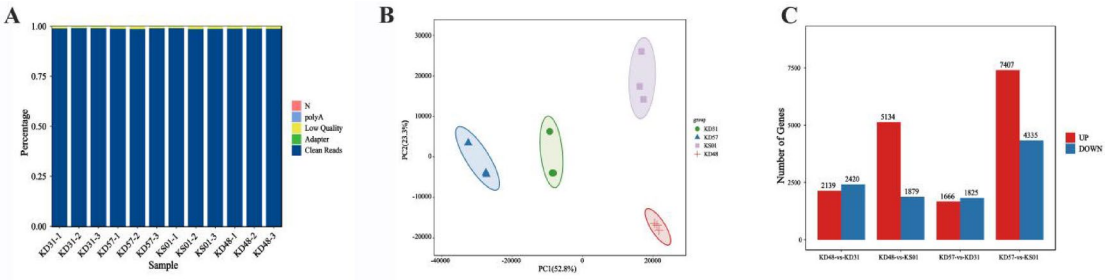

Fig. 5. (A) Statistical graph of the clean read proportions. (B) PCA of the twelve samples used for transcriptome sequencing. (C) Number of DEGs in these four groups.

Potential candidate gene analysis

There were 441 genes in the search scopes for 23 common QTNs, of which a total of 196 genes were highly expressed in seeds during lipid formation. Pathway analysis of the 196 genes revealed that 132 genes were enriched in 163 pathways. Among these pathways, only 15 were related to the synthesis and metabolism of lipids, including glycolysis/gluconeogenesis, the citrate cycle (TCA cycle), fructose and mannose metabolism, galactose metabolism, starch and sucrose metabolism, amino sugar and nucleotide sugar metabolism, glyoxylate and dicarboxylate metabolism, propanoate metabolism and inositol phosphate metabolism, which belong to the carbohydrate metabolism pathway; oxidative phosphorylation, photosynthesis-antenna proteins, nitrogen metabolism and sulfur metabolism, which belong to energy metabolism; glycerolipid metabolism, which belongs to lipid metabolism; and the carbon metabolism pathway (Supplementary Table S4).

A total of 13 genes were associated with these 15 pathways, and 8 genes were differentially expressed between the high and low OC varieties of soybean compared with the DEGs of the KD48-vs-KS01, KD57-vs-KS01, KD48-vs-KD31 and KD57-vs-KD31 groups (Table 4; Fig. 6A). The eight DEGs were *Glyma.09G173200*, *Glyma.03G021800*, *Glyma.14G218800*, *Glyma.16G213200*, *Glyma.18G025500*, *Glyma.18G027100*, *Glyma.11G093200* and *Glyma.18G027200*.

For WGCNA based on transcriptome sequencing data, genes with similar expression patterns were clustered into 20 modules (Fig. 6B). An analysis of the relationship between each module and the OC phenotype revealed that MM.darkseagreen 4 ($p = 4e\text{-}07$), MM.violet ($p = 2e\text{-}04$), MM.darkred ($p = 0.01$) and MM.floralwhite ($p = 0.04$) presented extremely significant positive correlations with OC traits. The module with an extremely significant negative correlation with the OC trait was MM.brown 4 ($p = 0.003$) (Fig. 6B). A correlation analysis of the 8 DEGs with the above 5 modules was further performed, and 7 of the DEGs (all except *Glyma.16G13200*) were significantly associated with one or more of the 5 modules. *Glyma.11G093200* and *Glyma.18G027100* presented significant or extremely significant positive correlations with 4 modules (red font in Fig. 6C), which presented positive correlations with the OC trait, and *Glyma.18G027200* presented a significant positive correlation with MM.darkred and an extremely significant positive correlation with MM.floralwhite. *Glyma.03G021800*, *Glyma.09G173200*, *Glyma.18G025500* and *Glyma.14G218800* presented significant or extremely significant positive correlations with the MM.brown 4 module, which was negatively correlated with the OC trait (Fig. 6C). Therefore, the seven DEGs (all except *Glyma.16G213200*) were considered potential candidates for further haplotype analysis.

Haplotype analysis

Haplotypes were classified through a natural population that included 141 soybean accessions for which genotypic data were available (https://ngdc.cncb.ac.cn/soyomics). Through haplotype analysis of the above 7 DEGs, we found that the haplotypes of *Glyma.18G027100* and *Glyma.03G021800* presented significant differences in OC in the natural population. For *Glyma.18G027100*, two haplotypes were identified on the basis of SNPs at different positions in the exonic region: Hap1 (TAACACAT) and Hap2 (CGGTGGAC) (Fig. 7A). Hap1 presented the highest frequency (82%). In the population used for haplotype analysis, the mean OC of Hap1 was significantly greater than that of Hap2 (Fig. 7B). These results showed that Hap1 (TAACACAT) of *Glyma.18G027100* could increase the OC. Three distinct haplotypes were identified from the analysis of *Glyma.03G021800*: Hap1 (AGGA), Hap2 (AGAC) and Hap3 (CAAC) (Fig. 7C). Hap1, comprising 51% of the population, had a significantly greater mean OC than the remaining two haplotypes and was an excellent haplotype for enhancing the OC (Fig. 7D). In summary, *Glyma.18G027100* and *Glyma.03G021800* are considered candidate genes related to soybean OC.

Discussion

In this study, the distribution of seed OC followed a normal distribution in the FW-RIL population in all 10 environments, which indicates that the OC was controlled via a complex genetic basis (Supplementary Figure S1). Although the OC had a relatively high h^2, the extremely significant variance in genotype×environment interaction effects showed that it was difficult to identify stable QTNs in different environments. In the present study, only 2 QTNs (*qOC-7-1* and *qOC-18-1*) could be stably detected in multiple environments. *qOC-7-1* was detected simultaneously in the E6 and E7 environments, and *qOC-18-1* was detected simultaneously in the E4 and E10 environments (Table 3). According to the results of the correlation analysis, E4 and E10 did not exhibit

QTN Name	Lead SNP	Gene Name [a]	Chr[b]	Position (bp)	K_ID	Gene annotation	Homologous gene [c]
qOC-3-1	**AX-157,128,483**	***Glyma.03G021800***	**3**	**2,265,529.2,269,179**	**K03952**	**NADH dehydrogenase (ubiquinone) 1 alpha subcomplex subunit 8**	***AT3G06310;AT5G18800***
qOC-8-1	AX-116,922,190	*Glyma.08G011800*	8	942,038.944,988	K01784	UDP-glucose 4-epimerase	*AT1G12780(UGE1);AT1G63180(UGE3);AT1G64440(RHD1);AT4G23920(UGE2);AT4G10960(UGE5)*
qOC-9-3	AX-157,361,651	*Glyma.09G171800*	9	39,653,896.39,655,287	K01054	acylglycerol lipase	*AT1G18360;AT1G73480*
qOC-9-3	AX-157,361,651	*Glyma.09G173200*	9	39,770,769.39,775,149	K01915	glutamine synthetase	*AT5G37600(GSR_1);AT5G35630(GS2);AT3G17820(GLN1.3);AT5G16570(GLN1;4);AT1G48470(GLN1;5);AT1G66200(GSR2)*
qOC-11-1	AX-157,229,785	*Glyma.11G093200*	11	7,061,666.7,066,152	K00392	sulfite reductase (ferredoxin)	*AT5G04590(SIR)*
qOC-14-1	AX-157,557,823	*Glyma.14G218800*	14	48,395,238.48,399,660	K00844	hexokinase	*AT4G29130(HXK1);AT2G19860(HXK2);AT1G47840(HXK3);AT1G50460(HKL1)AT4G37840(HKL3);AT3G20040(ATHXK4)*
qOC-16-3	AX-157,132,706	*Glyma.16G213200*	16	37,133,392.37,137,981	K01899	succinyl-CoA synthetase alpha subunit	*AT5G08300;AT5G23250*
qOC-18-1	AX-116,877,759	*Glyma.18G025500*	18	1,879,117.1,887,759	K22920	UTP—glucose-1-phosphate uridylyltransferase	*AT3G56040(UGP3)*
qOC-18-1	**AX-157,333,937**	***Glyma.18G027100***	**18**	**2,033,839.2,038,697**	**K05857**	**phosphatidylinositol phospholipase C, delta**	***AT4G38530(PLC1);AT5G58670(PLC1);AT3G08510(PLC2);AT5G58700(PLC4);AT5G58690(PLC5);AT2G40116;AT3G55940***
qOC-18-1	AX-157,333,937	*Glyma.18G027200*	18	2,042,578.2,048,191	K05857	phosphatidylinositol phospholipase C, delta	*AT4G38530(PLC1);AT5G58670(PLC1);AT3G08510(PLC2);AT5G58700(PLC4);AT5G58690(PLC5);AT2G40116;AT3G55940*
qOC-18-1	AX-157,333,937	*Glyma.18G027300*	18	2,049,280.2,053,959	K05857	phosphatidylinositol phospholipase C, delta	*AT4G38530(PLC1);AT5G58670(PLC1);AT3G08510(PLC2);AT5G58700(PLC4);AT5G58690(PLC5);AT2G40116;AT3G55940*
qOC-18-1	AX-157,333,937	*Glyma.18G028400*	18	2,142,745.2,145,667	K08915	light-harvesting complex II chlorophyll a/b binding protein 4	*AT5G01530(LHCB4.1);AT3G08940(LHCB4.2);AT2G40100(LHCB4.3)*
qOC-18-2	AX-157,083,942	*Glyma.18G060600*	18	5,457,321.5,469,620	K01103	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	*AT1G07110(F2 KP)*

Table 4. 4 Details of 13 genes annotated in the KEGG database and information regarding homologous genes in *Arabidopsis Thaliana*. Bold font indicate the genes which correlate with the fat anabolism in soybean according to our deduction;. [a]Indicates the gene which correlates with the QTN (before the gene in the same row);. [b]Chr: Chromosome;. [c]Indicate the homologous genes in *Arabidopsis thaliana*.

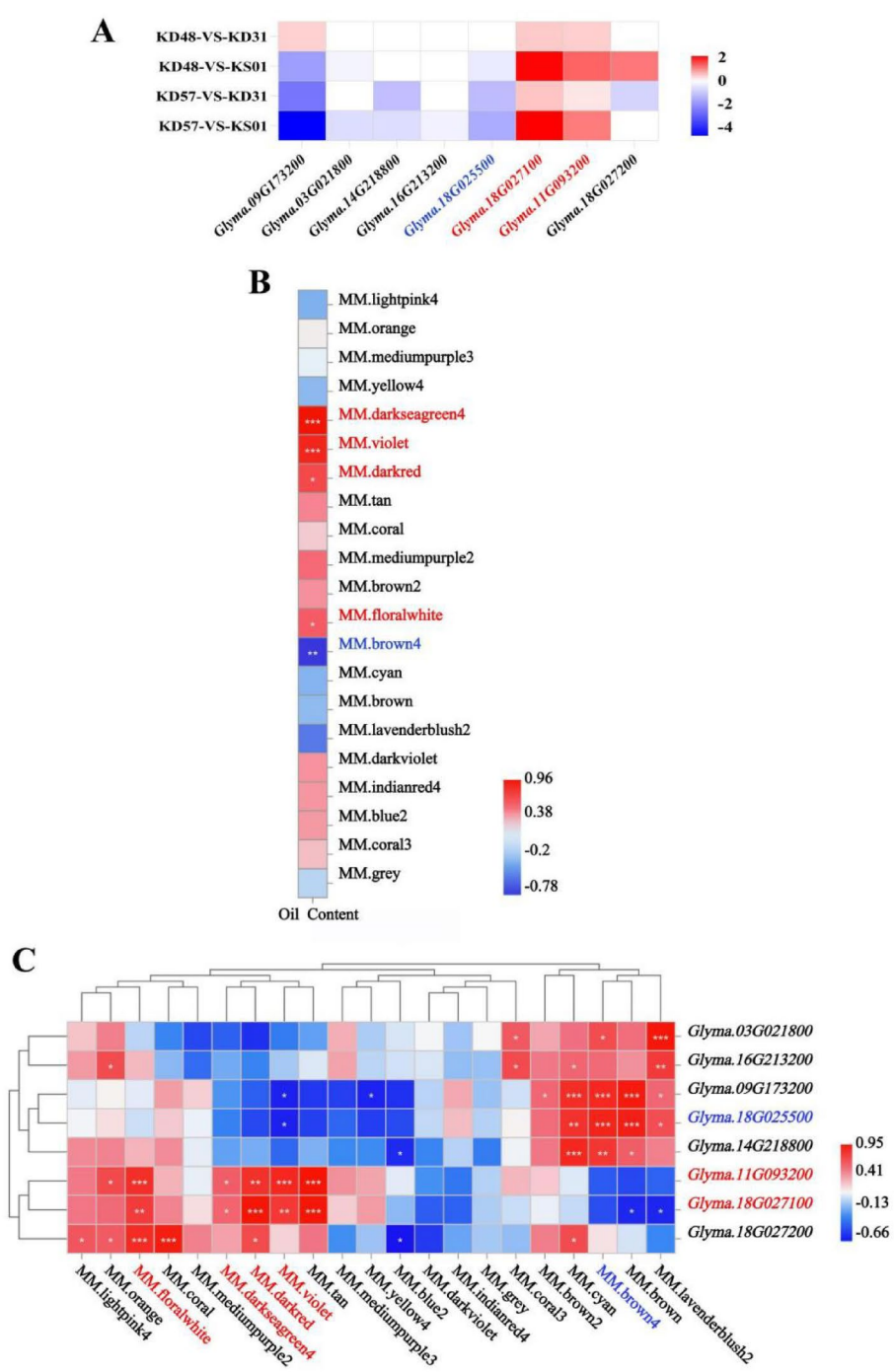

Fig. 6. (A) Expression heatmap of the 8 DEGs in these four groups. The blue represents downregulated genes, and the red represents upregulated genes. The white blank represents a nonsignificant difference. The genes in red font were upregulated in all the groups, and the genes in blue font were downregulated in all the groups, indicating a significant difference in expression. (B) Correlations of the OC trait with WGCNA modules. The red color represents a positive correlation, and the blue color represents a negative correlation. * represents a significant correlation at the 0.05 level, ** represents a significant correlation at the 0.01 level, and *** represents a significant correlation at the 0.001 level. (C) Correlations between modules and 8 DEGs. The red color represents a positive correlation, and the blue color represents a negative correlation. * represents a significant correlation at the 0.05 level, ** represents a significant correlation at the 0.01 level, and *** represents a significant correlation at the 0.001 level.

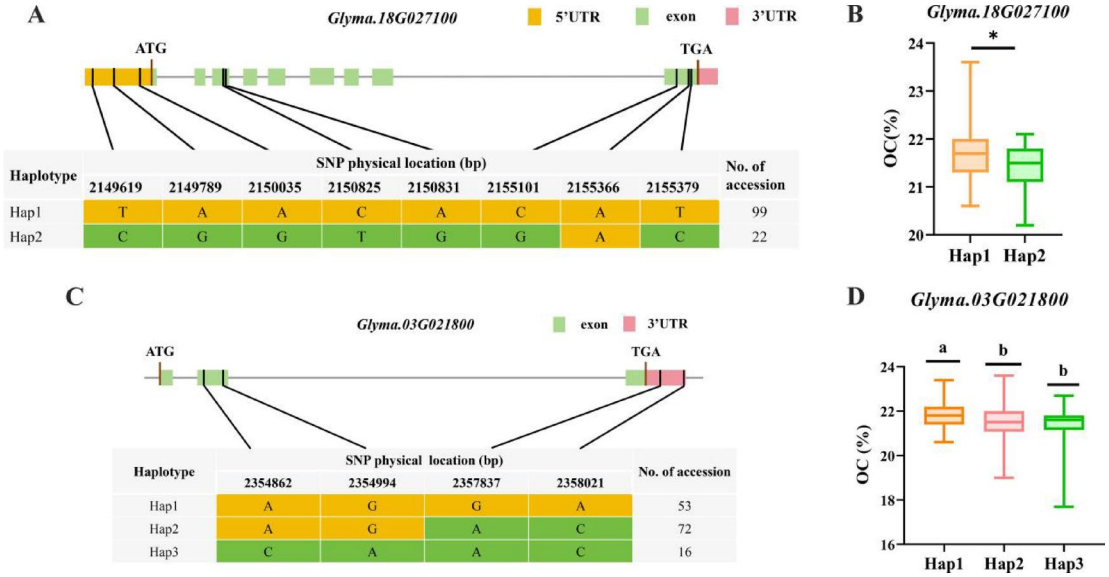

Fig. 7. (A) Haplotype information for *Glyma.18G027100* and the corresponding numbers of accessions. The upper part shows the gene structure diagram of *Glyma.18G027100*. (B) Boxplot showing the OC (percentage of dry weight) of the Hap1 and Hap2 haplotypes of *Glyma.18G027100* among 121 accessions. (C) Haplotype information for *Glyma.03G021800* and the corresponding numbers of accessions. The upper part shows the gene structure diagram for *Glyma.03G021800*. (D) Boxplot showing the OC (percentage of dry weight) of the Hap1, Hap2 and Hap3 haplotypes of *Glyma.03G021800* among 141 accessions.

a significant correlation (Fig. 2), which indicated that the effects of the E4 and E10 environments on OC differed greatly. Therefore, we believe that *qOC-18-1* was more stable and more reliable for searching for candidate genes.

However, to avoid the omission of important QTNs due to this environmental impact, we also treated the QTNs identified by multiple methods as important target QTNs. In this way, stable environment-specific QTNs could be obtained by further reducing false positives, thereby preventing valuable genes from being overlooked. Many studies have shown that QTNs identified by multiple methods are reliable and suitable for identifying candidate genes[42,53–56]. The candidate gene *Glyma.03G021800* found in this study was identified in *qOC-3-1*, which was detected via two methods.

A total of 23 important QTNs were identified in this study. Among them, 12 QTNs existed in the QTL intervals associated with OC in soybean that have been mapped in previous studies: *qOC-2-2*, *qOC-2–3*, *qOC-3-1*, *qOC-5-2*, *qOC-7-1*, *qOC-7-3*, *qOC-9-6*, *qOC-10-2*, *qOC-11-1*, *qOC-14-1*, *qOC-16-2* and *qOC-17-4* (Table 3)[2,44–52]. In addition, we predicted that one significant SNP, AX-157,333,937, is associated with protein content (PC) in a previous analysis[37], and it was found to be related to OC in soybean seeds in this study (Table 3); thus, AX-157,333,937, which belongs to *qOC-18-1*, is related to both PC and OC in soybean seeds. PC and OC are a pair of antagonistic traits, and AX-157,333,937 showed that the QTN effect was positive for PC (QTN effect: 0.48%~0.96%) and negative for OC (−0.23%~−0.16%). It is predicted that AX-157,333,937 is an important QTN related to soybean quality traits and may play an important role in the breeding of high-quality soybean.

Through GWAS, transcriptomic analysis and haplotype analysis, the genes *Glyma.18G027100* and *Glyma.03G021800* have been identified as candidate genes related to soybean OC. *Glyma.18G027100*, which was found across *qOC-18-1*, is a *GmPLCδ* gene belonging to the PLC gene family and encodes phosphatidylinositol phospholipase C delta (PLCδ). Gene Ontology annotation revealed that these genes are involved in the lipid metabolism process (GO:0006629). Twelve *PLC* genes were detected in the soybean genome, and they were distributed across chromosomes 2, 11, 14 and 18 equally. Interestingly, all three *PLC* genes on each chromosome were located adjacent to one another, except *Glyma.14G193800* (*GmPLC1*) on chromosome 14. A similar phenomenon has been observed in *Arabidopsis thaliana*, in which *AtPLC1*, *AtPLC4* and *AtPLC5*are located in a 12 kb fragment on chromosome 5[57]. The PLC gene family has been confirmed in *A. thaliana* mutants to be associated with acyl-lipid metabolism (http://aralip.plantbiology.msu.edu)[58]. Some studies on *PLC*genes in soybean have also been reported, for example, studies on the induction of pathogenesis-related genes in soybean and the regulation of abiotic stress responses[59,60]. However, a *PLC* gene has not been reported to be involved in acyl-lipid metabolism in soybean. We can perform further studies on *PLC* genes in soybean to determine the relationship between *PLC* genes and OC in soybean.

Glyma.03G021800 encodes a NADH-ubiquinone oxidoreductase that contains a CHCH domain, which acts on the oxidative phosphorylation pathway and is involved in mitochondrial ATP synthesis coupled with electron transport (GO:0042775). The synthesis of oils requires the combination of glycerol triphosphate with long-chain fatty acids, which are synthesized from short chains on the endoplasmic reticulum. Short-chain fatty acids are synthesized in the mitochondria by ATP as the main energy source and then transported out of the mitochondria. Since oxidative phosphorylation is the main pathway for ATP formation in mitochondria, we predicted that *Glyma.03G021800* is closely related to the synthesis of soybean oil.

In summary, A total of 23 common QTNs associated with OC were detected via five multilocus GWAS methods in 10 environments. Among them, 12 QTNs were found for the first time in this study. Among these QTNs, 2 candidate genes, *Glyma.18G027100* and *Glyma.03G021800*, were identified. These results contribute to the analysis of the regulatory network of OC in soybean seeds and are valuable for MAS in breeding of high-oil soybean varieties.

Materials and methods

Plant materials and collection of phenotypic data

A population with 144 FW-RILs, derived from a four-way cross of the parental lines Kenfeng14 (OC 20.88%), Kenfeng15 (OC 21.66%), Kenfeng19 (OC 18.33%), and Heinong48 (OC 17.76%), was used in the present study. Details regarding FW-RIL population construction were described in a previous paper published by our research team[37].

The 144 FW-RILs and 4 parental lines were planted in 10 environments (hereafter expressed as E1-E10) in different locations and years (Supplementary Table S5). All the materials in each environment were planted in a completely randomized block design with three replications. Three lines were set in a block, and the length and width of the lines were 5 m and 0.7 m, respectively. The experimental plots were managed in a manner consistent with local soybean production. At maturity, ten plants from the middle of each line were harvested and threshed separately. The OC of dry seeds (with approximately 10% moisture) of each line was determined via a near-infrared analyzer (Infratec 1241, Foss, Denmark) at the Key Laboratory of Soybean Biology of the Chinese Education Ministry at Northeast Agricultural University in China. The calibration regression technique used for this near-infrared analyzer was the partial least square (PLS) method, which involves combining spectral data with laboratory data (Kjeldahl method) to calculate the seed OC, described as the percentage of seed weight. The phenotypic values given for each parental and FW-RIL individual used in this study were all the averages of three repetitions.

Analysis of phenotypic data

The analysis of phenotypic data included determination of the mean, standard deviation, range, minimum, maximum, skewness, kurtosis, coefficient of variation (CV) values, variance (ANOVA) and broad-sense heritability (h^2). All the statistical analyses were implemented in SAS 9.2 (SAS Institute, Cary, USA). The ANOVA for multiple environments was conducted jointly via the generalized linear model (GLM) method, and the variance components were estimated via the mixed linear model (MLM). Then, h^2 was calculated via the following equation:

$$h^2 = \sigma_g^2 / \left(\sigma_g^2 + \sigma_{ge}^2/n + \sigma_e^2/nr\right)$$

where σ_g^2, σ_{ge}^2 and σ_e^2 represent the variance in genotype, genotype-by-environment interaction and error variance, respectively; r represents the number of replications within an environment; and n represents the number of environments.

Genotyping, structure and linkage disequilibrium (LD)

Genomic DNA was extracted from the juvenile leaves of parent and FW-RIL plants via the CTAB method[61]. The DNA concentration was determined via a UV752 N spectrophotometer (Shanghai Jingke Science Instrument Co. Ltd.), and the DNA was diluted to 100 ± 1 ng in deionized water. SNP genotyping was based on the SoySNP660 K BeadChip method and performed at Beijing Boao Biotechnology Co., Ltd. By performing quality filtration with the criteria maximum missing site ratio < 10% and minor allele frequency (MAF) > 5%[62], 109,676 SNPs were selected for GWAS analysis. The total number of SNPs on each chromosome is shown in Supplementary Table S6.

By analyzing population structure, the FW-RIL population was divided into two subpopulations, and a Q matrix was obtained and used for multilocus GWAS. The LD region decayed fastest before 200 kb and then tended to flatten, so the 200 kb regions flanking the targeted SNP were determined to search for potential candidate genes. The results have been described in a previously published paper[37].

Genome-wide association study (GWAS)

Five kinds of multilocus GWAS methods, including mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, and ISIS EM-BLASSO[63–66], were used to perform GWAS of the OC in soybean seeds. Calculations were performed via mrMLM.GUI (version 3.0) software. The critical P value parameters for these methods at the first stage were set to 0.01, except for FASTmrEMMA, where the critical P value was set to 0.005; the critical LOD score was set to 3, and the P value was set to 0.0002 for a significant QTN at the last stage. The parameter settings were Likelihood="REML", SearchRadius = 100, CriLOD = 3, SelectVariable = 143, and Bootstrap = FALSE. The kinship matrix was also obtained via mrMLM.GUI 3.0 software.

The significant SNPs were considered the same QTN if they met one of the following conditions: (1) the same SNP markers were detected; (2) the attenuation intervals of different SNP markers intersected, and the same QTNs detected repeatedly by multiple methods were calculated only once.

The QTNs were named as follows: *q* + *trait name-chromosome name-QTN number*, where *q* represents the QTN, the trait name is the OC, and the QTN number represents the ranking of the QTN on the corresponding chromosome.

The QTNs detected via multiple GWAS methods or in multiple environments were considered common QTNs.

Transcriptome sequencing analysis

The soybean materials used for transcriptome sequencing were Keshan 01 (KS01), Kedou 31 (KD31), Kedou 48 (KD48) and Kedou 57 (KD57). KS01 and KD31 were high-oil cultivars with 21.82% and 21.26% OC, respectively, and KD48 and KD57 were low-oil cultivars with 15.78% and 16.44% OC, respectively. These cultivars were provided by the Keshan Branch of Heilongjiang Academy of Sciences, Keshan, Heilongjiang, China.

Uniform seeds of the four cultivars were chosen and sown in pots (five seeds/pot, 25*18 cm pots), and after seed germination, the number of seedlings was reduced to 1 per pot. At 25 days after flowering (the rapid oil accumulation stage l), tender seeds collected from three replicate plants were frozen in liquid nitrogen and stored in an ultralow-temperature freezer at −80 °C until transcriptome sequencing analysis. All four samples were subjected to transcriptome sequencing with three biological repeats.

Transcriptome sequencing analysis was performed on an Illumina HiSeq2500/4000 system by Gene Denovo Biotechnology Co. (Guangzhou, China). To obtain high-quality clean reads, the sequencing data were processed via fastp[67](version 0.18.0) for read filtering. The parameters used in this analysis were as follows: (1) reads containing adapters were removed; (2) reads containing more than 10% unknown nucleotides (N) were removed; and (3) low-quality reads containing more than 50% low-quality (Q value ≤ 20) bases were removed. The mapped reads of each sample were assembled via StringTie v1.3.1[68,69]. For each transcription region, an FPKM (fragment per kilobase of transcript per million mapped reads) value was calculated to quantify its expression abundance and variations via RSEM[70] software. The FPKM formula is as follows:

$$\mathrm{FPKM} = \frac{10^6 \mathrm{C}}{\mathrm{NL}/10^3}$$

Given FPKM(A) to be the expression of gene A, C to be number of fragments mapped to gene A, N to be total number of fragments that mapped to reference genes, and L to be number of bases on gene A. The FPKM method is able to eliminate the influence of different gene lengths and sequencing data amount on the calculation of gene expression. Therefore, the calculated gene expression can be directly used for comparing the difference of gene expression among samples.

Differential expression analysis was performed using DESeq2[71]software between two different groups (and by edgeR[72] between two samples). The genes/transcripts with a false discovery rate (FDR) of less than 0.05 and an absolute fold change of ≥ 2 were considered differentially expressed genes/transcripts.

Weighted gene correlation network analysis (WGCNA) was performed using Omicsmart, a real-time interactive online platform for data analysis (http://www.omicsmart.com). The expression similarity between genes was calculated, and the Pearson correlation coefficient was used to measure the correlation between gene pairs. The similarity matrix was converted into a weighted adjacency matrix, and a soft thresholding method was used to enhance strong and weaken weak correlations. The appropriate soft threshold parameter (power value) was selected to assign the network to the characteristics of a scale-free network. Genes were divided into different modules via hierarchical clustering and dynamic tree cut methods. Each module consists of a set of highly coexpressed genes, and the genes within the module have similar expression patterns. Module-trait associations were estimated via the correlation (Pearson correlation) between the module and the OC trait.

Real-time PCR verification

Total RNA extraction, cDNA synthesis and qRT-PCR analysis for each sample were performed according to Zhang et al.[73]. The internal reference gene was *Acint7*, and data analysis was performed via the $2^{-\Delta\Delta CT}$ method. The primers used for qRT-PCR are shown in Supplementary Table S7.

Search for potential candidate genes

On the basis of the rate of LD decay, the 100 kb intervals on either side of each common QTN were set as the range for seeking potential candidate genes. On the Phytozome website (https://phytozome.jgi.doe.gov), all the genes in these intervals were found out. According to the transcriptomic datasets for the seed developmental stages in soybean (GES42871) downloaded from the GEO (Gene Expression Omnibus) database, genes that are highly expressed in seeds during oil formation were subsequently chosen. All the highly expressed genes were subsequently combined for pathway analysis via the Kyoto Encyclopedia of Genes and Genomes (KEGG) website (http://www.kegg.jp)[74–76], after which genes enriched in lipid anabolism-related pathways were identified. Finally, according to the transcriptome sequencing results, differentially expressed genes (DEGs) between the high-oil and low-oil cultivars were identified in lipid anabolism-related pathways. By comparing these DEGs with the results of WGCNA and annotation information for the genes, potential candidate genes were identified.

Haplotype analysis

The haplotype analysis of genes among 141 soybean accessions was performed using the available genotypic data (https://ngdc.cncb.ac.cn/soyomics). First, the markers with missing genotypes more than 20% were deleted. Next, imputation using markov function of NAM package in R was used to fill with missing genotypes. Finally, markers with minor allele frequency less than 0.05 were deleted. The haplotypes were constructed according to SNPs from the untranslated region (UTR) and exon regions via Haploview v4.2 software[77]. Differences between haplotypes for each gene were tested by analysis of variance (contains more than two haplotypes of genes) or unpaired t-test (contains two haplotypes of genes). Multiple comparisons of phenotype differences among various haplotypes were conducted using the LSD.test function of agricolae package in R.

Data availability

Transcriptomic data that support the findings of this study have been deposited in the NCBI (BioProject ID PRJNA1185960, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1185960).

Received: 17 October 2024; Accepted: 2 May 2025
Published online: 14 May 2025

References

1. Marchive, C., Nikovics, K., To, A., Lepiniec, L. & Baud, S. Transcriptional regulation of fatty acid production in higher plants: molecular bases and biotechnological outcomes. *Eur. J. Lipid Sci. Tech.* **116**, 1332–1343 (2014).
2. Cao, Y. C. et al. Identification of major quantitative trait loci for seed oil content in soybeans by combining linkage and Genome-Wide association mapping. *Front. Plant. Sci.* **8** (2017).
3. Zhang, T., Wu, T., Wang, L., Jiang, B. & Sun, S. A. Combined linkage and GWAS analysis identifies QTLs linked to soybean seed protein and oil content. *Int. J. Mol. Sci.* **20**, 5915 (2019).
4. Li, S. G., Xu, H. F., Yang, J. Y. & Zhao, T. J. Dissecting the genetic architecture of seed protein and oil content in soybean from the Yangtze and Huaihe river valleys using Multi-Locus Genome-Wide association studies. *Int. J. Mol. Sci.* **20** (2019).
5. Yao, Y. J. et al. Quantitative trait loci analysis of seed oil content and composition of wild and cultivated soybean. *BMC Plant. Biol.* **20** (2020).
6. Huang, J. H. et al. Identification and mapping of stable QTLs for seed oil and protein content in soybean [*Glycine max* (L.) Merr]. *J. Agric. Food Chem.* **68**, 6448–6460 (2020).
7. Zhu, X. T., Leiser, W. L., Hahn, V. & Würschum, T. Identification of seed protein and oil related QTL in 944 RILs from a Diallel of early-maturing European soybean. *Crop J.* **9**, 238–247 (2021).
8. Liu, J. Q. et al. QTL mapping for seed quality traits under multiple environments in soybean (*Glycine max* L). *Agronomy-Basel* **13** (2023).
9. Li, B. et al. Identification of an important QTL for seed oil content in soybean. *Mol. Breeding* **43** (2023).
10. Li, Y. H. et al. Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. *Plant. Sci.* **266**, 95–101 (2018).
11. Li, D. M., Zhao, X., Han, Y. P., Li, W. B. & Xie, F. T. Genome-wide association mapping for seed protein and oil contents using a large panel of soybean accessions. *Genomics* **111**, 90–95 (2019).
12. Wang, Y. Y. et al. Genotyping of soybean cultivars with Medium-Density array reveals the population structure and QTNs underlying maturity and seed traits. *Front. Plant. Sci.* **9** (2018).
13. Li, X. Y. et al. Detection of QTL and QTN and candidate genes for oil content in soybean using a combination of four-way-RIL and germplasm populations. *Crop J.* **8**, 802–811 (2020).
14. Zhao, X. C. et al. GWAS and WGCNA analysis uncover candidate genes associated with oil content in soybean. *Plants-Basel* **13** (2024).
15. Wang, H. W. et al. The soybean Dof-type transcription factor genes, GmDof4 and GmDof11, enhance lipid content in the seeds of Transgenic Arabidopsis plants. *Plant J.* **52**, 716–729 (2007).
16. Song, Q. X. et al. Soybean *GmbZIP123* gene enhances lipid content in the seeds of Transgenic *Arabidopsis* plants. *J. Exp. Bot.* **64**, 4329–4341 (2013).
17. Haun, W. et al. Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. *Plant Biotechnol. J.* **12**, 934–940 (2014).
18. Liu, Y. F. et al. Soybean *GmMYB73* promotes lipid accumulation in Transgenic plants. *BMC Plant. Biol.* 14 (2014).
19. Zhang, L. et al. An integrated bioinformatics analysis reveals divergent evolutionary pattern of oil biosynthesis in High- and Low-Oil plants. *Plos One* **11** (2016).
20. Lu, X. et al. The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. *Plant. J.* **86**, 530–544 (2016).
21. Li, Q. T. et al. Selection for a Zinc-Finger protein contributes to seed oil increase during soybean domestication. *Plant. Physiol.* **173**, 2208–2224 (2017).
22. Zhang, D. J. et al. Elevation of soybean seed oil content through selection for seed coat shininess. *Nat. Plants*. **4**, 30–35 (2018).
23. Zhang, D. et al. Artificial selection on *GmOLEO1* contributes to the increase in seed oil during soybean domestication. *PLoS Genet.* **15** (2019).
24. Zhang, H. Y. et al. Dual use and selection of *GmSWEET39* for oil and protein improvement in soybean. *PLoS Genet.* 16 (2020).
25. Liu, J. Y. 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.* **71**, 6988–7002 (2020).
26. Zhang, J. et al. Identification of QTNs, QTN-by-environment interactions and genes for yield-related traits in rice using 3VmrMLM. *Front. Plant. Sci.* **13** (2022).
27. Zhao, Q. et al. Identification of QTNs, QTN-by-environment interactions, and their candidate genes for grain size traits in main crop and Ratoon rice. *Front. Plant. Sci.* **14** (2023).
28. Sui, Y., Che, Y. R., Zhong, Y. & He, L. Q. Genome-wide association studies using 3VmrMLM model provide new insights into branched-chain amino acid contents in rice grains. *Plants-Basel* **12** (2023).
29. Zhang, P. et al. Genome-Wide Association Study and Candidate Gene Mining of Seed Size Traits in Soybean. *Agronomy-Basel* **14** (2024).
30. Yang, B. et al. Genome-wide association studies for rice seed germination under drought stress using 3VmrMLM. *Food Energy Secur.* **13** (2024).
31. Huang, B. E. et al. A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnol. J.* **10**, 826–839 (2012).
32. Bandillo, N. et al. Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice* **6** (2013).
33. Mackay, I. J. et al. An Eight-Parent multiparent advanced generation Inter-Cross population for Winter-Sown wheat: creation, properties, and validation. *G3-Genes Genomes Genet.* **4**, 1603–1610 (2014).
34. Dell'Acqua, M. et al. Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biol.* **16** (2015).
35. Pascual, L. et al. Potential of a tomato MAGIC population to Decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnol. J.* **13**, 565–577 (2015).
36. Meng, L. J., Guo, L. B., Ponce, K., Zhao, X. Q. & Ye, G. Y. Characterization of Three Rice Multiparent Advanced Generation Intercross (MAGIC) Populations for Quantitative Trait Loci Identification. *Plant Genome* **9** (2016).
37. Zhang, K. X. et al. Identification of QTNs controlling seed protein content in soybean using Multi-Locus Genome-Wide association studies. *Front. Plant. Sci.* **9** (2018).
38. Li, X. Y. et al. Fine mapping QTL and mining genes for protein content in soybean by the combination of linkage and association analysis. *Theor. Appl. Genet.* **134**, 1095–1122 (2021).

39. Shang, G. X. et al. Genome-Wide Association Study of Phenylalanine Derived Glucosinolates in (*Brassica rapa*). *Plants-Basel* **11** (2022).
40. Asfaw, A., Mondo, J. M., Agre, P. A., Asiedu, R. & Akoroda, M. O. Association mapping of plant sex and cross-compatibility related traits in white Guinea Yam (*Dioscorea rotundata* Poir.) clones. *BMC Plant. Biol.* **22** (2022).
41. Guo, S. J. et al. High-Throughput Phenotyping Accelerates the Dissection of the Phenotypic Variation and Genetic Architecture of Shank Vascular Bundles in Maize (*Zea mays* L.). *Plants-Basel* **11** (2022).
42. Park, M. et al. Chromosome-level genome sequence assembly and genome-wide association study of (*Muscadinia rotundifolia*) reveal the genetics of 12 berry-related traits. *Hortic. Res.* **9** (2022).
43. Delfan, S., Bihamta, M. R., Dadrezaei, S. T., Abbasi, A. & Alipoor, H. Exploring genomic regions involved in bread wheat resistance to leaf rust at seedling/adult stages by using GWAS analysis. *BMC Genom.* **24** (2023).
44. Kabelka, E. A. et al. Putative alleles for increased yield from soybean plant introductions. *Crop Sci.* (2004).
45. Akond, M., Liu, S., Boney, M., Kantartzi, S. K. & Kassem, M. A. Identification of quantitative trait loci (QTL) underlying protein, oil, and five major fatty acids' contents in soybean. *Am. J. Plant. Sci.* **5**, 158–167 (2014).
46. Qi, Z. M. et al. Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. *Euphytica* **179**, 499–514 (2011).
47. Hwang, E. Y. et al. A genome-wide association study of seed protein and oil content in soybean. *BMC Genom.* **15** (2014).
48. Tajuddin, T., Watanabe, S., Yamanaka, N. & Harada, K. Analysis of quantitative trait loci for protein and lipid contents in soybean seeds using Recombinant inbred lines. *Breed. Sci.* **53**, 133–140 (2003).
49. Leite, D. C., Pinheiro, J. B., Campos, J. B., Di Mauro, A. O. & Unêda-Trevisoli, S. H. QTL mapping of soybean oil content for marker-assisted selection in plant breeding program. *Genet. Mol. Res.* **15** (2016).
50. Mao, T. T. et al. Identification of quantitative trait loci underlying seed protein and oil contents of soybean across multi-genetic backgrounds and environments. *Plant. Breed.* **132**, 630–641 (2013).
51. Gai, J. Y. et al. A comparative study on segregation analysis and QTL mapping of quantitative traits in plants-with a case in soybean. *Front. Agric. China*. **1**, 1–7 (2007).
52. Seo, J. H. et al. Quantitative trait locus analysis for soybean (*Glycine max*) seed protein and oil concentrations using selected breeding populations. *Plant. Breed.* **138**, 95–104 (2019).
53. Cui, Y. R., Zhang, F. & Zhou, Y. L. The application of Multi-Locus GWAS for the detection of Salt-Tolerance loci in rice. *Front. Plant. Sci.* **9** (2018).
54. Ma, L. L. et al. Genetic dissection of maize embryonic callus regenerative capacity using Multi-Locus Genome-Wide association studies. *Front. Plant. Sci.* **9** (2018).
55. Zhang, Y. M., Jia, Z. Y. & Dunwell, J. M. Editorial: The applications of new Multi-Locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant. Sci.* **10** (2019).
56. Ikram, M. 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* **11** (2020).
57. Tasma, I. M., Brendel, V., Whitham, S. A. & Bhattacharyya, M. K. Expression and evolution of the phosphoinositide-specific phospholipase C gene family in Arabidopsis thaliana. *Plant Physiol. Biochem.* **46**, 627–637 (2008).
58. Mcglew, K. et al. An annotated database of *Arabidopsis* mutants of acyl lipid metabolism. *Plant Cell Rep.* **34**, 519–532 (2015).
59. Chou, W. M., Shigaki, T., Dammann, C., Liu, Y. Q. & Bhattacharyya, M. K. Inhibition of phosphoinositide-specific phospholipase C results in the induction of pathogenesis-related genes in soybean. *Plant Biol.* **6**, 664–672 (2004).
60. Wang, F. W. et al. Genome-Wide analysis and expression profiling of the phospholipase C gene family in soybean (*Glycine max*). *Plos One* **10** (2015).
61. Doyle, J. J., Doyle, J. L. & Brown, A. Analysis of a polyploid complex in Glycine with Chloroplast and nuclear DNA. *Aust. Syst. Bot.* **3** (1990).
62. Belamkar, V. et al. Genomics-assisted characterization of a breeding collection of *Apios americana*, an edible tuberous legume. *Sci. Rep.* **6** (2016).
63. Wang, S. B. et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* **6** (2016).
64. Tamba, C. L., Ni, Y. L. & Zhang, Y. M. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* **13** (2017).
65. Wen, Y. J. et al. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* **19**, 700–712 (2018).
66. Zhang, J. et al. pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* **118**, 517–524 (2017).
67. Chen, S. et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *bioRxiv*, 274100 (2018).
68. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290 (2015).
69. Pertea, M. et al. Transcript-level expression analysis ofRNA-seq experiments with HISAT, stringtie and ballgown. *Nat. Protoc.* **11**, 650 (2016).
70. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
71. Love, M. I., Huber, W. & Anders, S. Moderated Estimation of fold change and dispersion for RNAseq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
72. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
73. Zhang, W. et al. A cation diffusion facilitator, *GmCDF1*, negatively regulates salt tolerance in soybean. *PLoS Genet.* 15 (2019).
74. Kanehisa, M. & Goto, S. K. E. G. G. Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
75. Kanehisa, M. Toward Understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).
76. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
77. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).

Acknowledgements

This work was supported in part by theNational Natural Science Foundation of China (32201874), the grant from Opening Competition Mechanism to Select the BestCandidates Project of Heilongjiang Province Department of Science and Technology (2023ZXJ02B02) and ScienceFoundation of Heilongjiang Province, China (LH2021C028).

Author contributions

H.N. and W-X. L. designed the research and applied funds. K.Z. and B.H. analyzed the data and wrote the

manuscript, and measured the phenotype data. W. W. conducted phenotypic collection. All authors revised the manuscript and approved the final version to be published.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00869-8.

Correspondence and requests for materials should be addressed to W.-X.L. or H.N.

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

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