Genome-Wide association Study Identifies Candidate Genes Related to Oleic acid content of Soybean Seed

Pi-wu Wang (✉ wangpiwu189@163.com)
Jilin Agricultural University

Qin Di
Universita degli Studi di Camerino

Xiao-Yi Liu
Guangzhou University of Chinese Medicine

Research article

Keywords: Soybean, Oleic acid, SNPs, Genome-wide association study, Candidate genes

Posted Date: March 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17853/v1

Abstract

Background: Soybeans [*Glycine max* (L.) Merr] originated in China and has been cultivated for more than 3000 years. Soybean oil is a complex mixture of five fatty acids (palmitic, stearic, oleic, linoleic, and linolenic). Soybean oil with high oleic acid content is desirable because this monounsaturated fatty acid improves the oxidative stability of the oil. To investigate the genetic architecture of oleic acid in soybean seed. 260 Soybean germplasms from Northeast China were collected as the natural population. A genome-wide association study (GWAS) was conducted in a panel of 260 germplasm resources.

Results: Phenotypic identification result showed that the content of oleic acid varied from 8.2% to 35.0%. 2,311,337 single nucleotide polymorphism (SNP) markers were obtained. GWAS analysis showed that there were many genes related to oleic acid content with contribution rate 7%. The candidate genes *Glyma.11G229600.1* on chromosome 11 and *Glyma.04G102900.1* on chromosome 4 were detected in a 2-year long GWAS. The candidate gene *Glyma.11G229600.1* showed a positive correlation with the oleic acid content, and the correlation coefficient was 0.980, while *Glyma.04G102900.1* showed a negative correlation with the oleic acid content. The correlation coefficient was -0.964.

Conclusions: *Glyma.04G102900.1* on chromosome 4 and *Glyma.11G229600.1* on chromosome 11 were detected in both anaylsis (2018 and 2019). *Glyma.04G102900.1* and *Glyma.11G229600.1* are new key candidate genes related to oleic acid in soybean seeds.

These results will be useful for High-oleic soybean breeding.

Background

Soybeans [*Glycine max* (L.) Merrill] originated in China and has been cultivated for more than 3000 years [1]. Soybean oil accounts for 20–25% of the total world fats and oil production and 30–35% of the total edible vegetable oil production [2]. In China, soybean oil is an important constituent of diet, and it is considered a major factor in preserving a healthy population. Soybean oil is a complex mixture of five fatty acids (palmitic, stearic, oleic, linoleic, and linolenic), all of which have different melting points, oxidative stabilities, and chemical functionalities [3].

Fatty acid composition of soybean oil ranged from about 5% to 11% in linolenic, 43% to 56% in linoleic, 15% to 33% in oleic, and 11% to 26% in saturated acids [4]. The palmitic acid and stearic acid belong to saturated fatty acids and they constitute 15% of the soybean oil, in recent times, there has been a running debate mainly in mainstream literature regarding the effects of palmitic acid and stearic acid consumption on the heart especially in the development of coronary artery disease [5, 6]. Linolenic acids and linoleic acid belong to polyunsaturated fatty acids and they constitute 80% of the soybean oil. Linolenic acid is needed for normal human growth and development, and linolenic acid can lower the cholesterol content in the blood, but linolenic acid is not resistant to high temperature, air oxygen and ultraviolet rays can oxidize linolenic acid, resulting in odor of soybean oil, which is not easy to preserve and lower the nutritional value of soybean oil. Oleic acid belongs to monounsaturated fatty acids, Soybean seeds with high oleic acid content can also reduce or eliminate chemical hydrogenation processes and reduce the cost of soybean oil processing [7]. The cultivation of soybean varieties with high oleic acid has become an important goal of high-quality soybean breeding [8].

The genome wide association analysis (GWAS) present a powerful tool to reconnect this trait back to its underlying genetics. With the rapid development of next-generation sequencing technology, GWAS has been successfully applied to plants such as rice and *Arabidopsis* [9, 10], a large number of genetic variations associated with complex traits were identified by the GWAS method [11]. In soybean, GWAS was performed to identify QTL controlling seed oil concentration in 298 soybean germplasm accessions exhibiting a wide range of seed protein and oil content [12]. A soybean breeding germplasm population (279 lines) was established to perform a GWAS, 8 QTLs were found that explained a range of phenotypic variance from 6.3 to 26.3% [13]. These results demonstrate that the use of GWAS with specially designed mapping populations is effective in uncovering the basis of key agronomic traits.

Scientists have successfully used the GWAS method to obtain a large number of candidate genes [14-16]. However, after the new candidate gene was discovered, how to verify the biological function of candidate genes, which has become a hot issue for researchers, RNA interference technology, the establishment and application of biochips, real-time fluorescent quantitative PCR technology (qRT-PCR) and gene editing technology provide theoretical basis for candidate gene function verification. In our study, we selected two key candidate genes to measure the expression in four different tissues (root, stem, leaf and seed) of 14 diverse soybean lines (two groups) using qRT-PCR.

In the present study, 260 Soybean germplasms from Northeast China (Heilongjiang province, Jilin province and Liaoning province) were collected as the natural population. The soybean lines were planted in the field of Jilin Agricultural University from 2018 to 2019. The contents of fatty acids in soybean seeds were determined by NIRS DS 2500 after harvest. SLAF-Seq technology was used to sequence genome of 260 soybean materials, and GWAS was used to screen candidate genes related to soybean oleic acid content.

Results

Phenotypic identification of oleic acid content in soybean seeds

From 2018 to 2019, the oleic acid content of each soybean lines was analyzed by SPSS 22.0 software. The Oleic acid content of seeds approached the normal distribution (Fig. 1a, b). The SD of oleic acid content of soybean seeds was 6.7 in 2018 and was 7.5 in 2019 (Fig. 1c). The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions. The results also indicated that the oleic acid content in different soybean lines is significantly different, the distribution of oleic acid content in soybean seed is continuous, which accorded with the genetic law of quantitative traits. The oleic acid content in soybean line q070 is 35.32% in 2019 and 34.84% in 2018, which was the largest one of all soybean lines. The oleic acid content in soybean line q024 is 5.49% in 2018 and 6.33% in 2019, which was the lowest content of all soybean lines.

Soybean fatty acid correlation analysis and Heritability Calculation

The multiple linear regression model was be used to analysis the relationship between different fatty acids. According to the regression coefficients of standardized multiple linear regression model, the content of soybean oleic acid is significantly negatively correlated with the content of linoleic acid and linolenic acid (Fig. 2). The correlation coefficient between oleic acid and linoleic acid is -0.660 and -0.532 (Table 1), there was a significant positive correlation between soybean oleic acid content and palmitic acid content, with a correlation coefficient 0.421. The heritability of the five fatty acids in 260 soybean lines is different, the heritability of oleic acid was 0.652 (Additional file 1: Table S1).

Table 1. Correlation between relative expression of oleic acid content in soybean

		Oleic acid	Linoleic acid	Linolenic acid	Palmitic acid	Stearic acid
Oleic acid	Pearson Correlation	1	-0.660**	-0.532**	0.421**	0.454**
Linoleic acid	Pearson Correlation	-0.660**	1	0.571**	0.051**	-0.893**
Linolenic acid	Pearson Correlation	-0.532**	0.571**	1	0.447**	-0.473**
Palmitic acid	Pearson Correlation	0.421**	0.051**	0.447**	1	-0.043**
Stearic acid	Pearson Correlation	0.454**	-0.893**	-0.473**	-0.043**	1

Note: * indicates significant at the 0.05, ** indicates significant at the 0.01.

SNP genotyping and SNP Annotation

In our experiment, SLAF-seq technology was used to sequence soybean genomic DNA. A total of 1,102,987 SLAF tags and 2,311,337 SNP markers were obtained (Fig. 3a). According to the number of SNP markers on different chromosomes. The results of SNP distribution on chromosomes are shown in Fig. 3b. According to the location information of SNP loci in reference genome (CDS regions, gene regions or intergenic regions), the mutate loci (non-synonymous mutations) were predicted. More than 50% of the SNP were located in the intergenic region, intergenic region is a stretch of DNA sequences located between genes, Intergenic regions are different from intragenic regions (or introns). Intragenic regions are a subset of noncoding DNA. 10% of the SNP markers were located in the upstream region of genes, and 10% of the SNPs were located in the downstream region of genes. 4.99% of the SNP loci were located in the protein coding regions. The 9.18% SNP markers were located in introns (Fig. 3c).

Phylogenetic analysis, Genetic structure analysis, Principal components analysis (PCA)

From the phylogenetic tree, it can be concluded that there are two large branches, this result suggested that 260 soybean lines are from the same ancestor. However, in the process of evolution, they evolved in two directions (Fig. 4a).

The group structure, refers to the sub-populations with different gene frequencies in the studied populations. The population structure analysis can quantify the number of ancestors of the studied population, and infer the source of each sample. It is a cluster analysis method that is currently more applied and is helpful to understand the evolution process of materials. Based on the SNPs. This experiment used to analyze the soybean population structure. For the study population, the number of subpopulations pre-set in this trial was 15 (Fig. 4b). We analyzed it with EIGENSTRAT in their study of 260 soybean lines. It concluded that sample we collected can be represented as an admixture of two ancestral populations (Fig. 4c).

Based on the difference in SNPs, this experiment performed principal component analysis by EIGENSOFT software, and carried out principal component analysis on 260 soybean materials to obtain clustering of 260 soybean materials. Based on the difference in SNP markers of 260 soybean lines, the PCA result showed that the 260 soybean lines can be divided into two subgroups (Fig. 4d). PC1, PC2 and PC3 accounted for 36.43%, 33.82% and 33.12%, respectively.

Genome-wide association analysis (GWAS) for seed oleic acid

Based on the oleic acid content of 260 soybean lines, the TASSEL software (Glm model, mlm model, cmlm model), fastlmmc software, and Emmax software were used for GWAS. The SNP markers significantly correlated with the oleic acid content of soybean seeds were detected, and the LD distance was set to 8.9 kb. The Manhattan and QQ (Quantile-Quantile) diagrams for oleic acid content in the 2018 and 2019 were showed in Fig. 5 and Fig. 6. The SNP markers which are significantly correlated with the oleic acid content of soybean seeds were searched (Additional file 2: Table S2). Using 9.7 kb as the linkage disequilibrium attenuation distance, candidate genes related to soybean oleic acid traits were screened in the LD distance. In 2018, 21 candidate genes related to soybean oleic acid content were screened using genome-wide association analysis (Table. 2). In 2019, 8 candidate genes were screened by genome-wide association analysis (Table. 2). Based on GO terms, the function of genes is: 1. Major CHO metabolism, 2. cell wall, 3. lipid metabolism, 4. metal handing, 5. N-metabolism, 6. amino acid metabolism, 7. secondary metabolism, 8. hormone metabolism, 9.tetrapyrrole synthesis, 10. Stress, 11. Redox, 12.misc, 13. Protein, 14. Cell, 15. Signaling, 16. Development, 17.transport. The functional distribution of candidate genes is shown in Fig. 7. The candidate gene *Glyma.11G229600.1* located on chromosome 11 was detected by GWAS during 2018 and 2019. The function of *Glyma.11G229600.1* was not reported in the soybean database, according to Swissport annotation, *Glyma.11G229600.1* belongs to the plant BAG protein family. The candidate gene *Glyma.04G102900.1* located on chromosome 4 was also detected by GWAS in two years. There is also no functional report about the candidate gene in soybean base, according to Swissport annotation. A similar gene in *Arabidopsis* belongs to the plant GRAS protein family.

Table. 2.Correlation between relative expression of oleic acid content in soybean

Year	chromosome	Gene	Predicted Function	Length	Contribution rate
2018	Chr03	*Glyma.03G054100.1*	3PREDICTED: Glycine max TMV resistance protein N-like (LOC100805036), transcript variant X3, mRNA	687	0.10
		Glyma.03G168200.3	3PREDICTED: Glycine max pleiotropic drug resistance protein 1-like (LOC100791601), mRNA	4662	0.07
	Chr04	*Glyma.04G191100.1*	3PREDICTED: Glycine max probable pectate lyase 18-like (LOC100814679), mRNA	1657	0.32
		Glyma.04G102900.1		2522	0.43
		Glyma.04G203200.1	3PREDICTED: Glycine max respiratory burst oxidase homolog protein C-like (LOC100800248), mRNA	2440	0.08
	Chr05	*Glyma.05G155300.1*	3PREDICTED: Glycine max ATP carrier protein 2, chloroplastic-like (LOC100797684), mRNA	1655	0.11
	Chr07	*Glyma.07G033100.1*	3PREDICTED: Glycine max ADP,ATP carrier protein 1, chloroplastic-like (LOC100793284), mRNA	2317	0.12
		Glyma.07G089000.1	3PREDICTED: Glycine max VIN3-like protein 1-like (LOC100780157), transcript variant X2, mRNA	2756	0.10
	Chr08	*Glyma.08G019700.1*	3PREDICTED: Glycine max calcium-dependent protein kinase 3-like (LOC100777096), transcript variant 1, mRNA	1877	0.16
		Glyma.08G185000.2	3PREDICTED: Glycine max probable plastid-lipid-associated protein 4, chloroplastic-like (LOC100803367), transcript variant 1, mRNA	979	0.15
	Chr11	*Glyma.11G229600.1*	3PREDICTED: Glycine max DNA replication complex BAG protein transcript variant 2, mRNA	1257	0.47
	Chr13	*Glyma.13G163400.1*	3PREDICTED: Glycine max protein S-acyltransferase 24-like (LOC100777470), misc_RNA	2490	0.32
	Chr14	*Glyma.14G045100.1*	3PREDICTED: Glycine max abscisic-aldehyde oxidase-like (LOC100812604), mRNA	4517	0.22
	Chr15	*Glyma.15G117700.1*	3PREDICTED: Glycine max uncharacterized LOC102666654 (LOC102666654), mRNA	693	0.17
		Glyma.15G120100.1	3PREDICTED: Glycine max tRNA methyltransferase 10 homolog A-like (LOC100779099), mRNA	1337	0.10
		Glyma.15G120200.2	3PREDICTED: Glycine max uncharacterized LOC102665381 (LOC102665381), mRNA	1227	0.08
		Glyma.15G127500.1	3PREDICTED: Glycine max polygalacturonase-like (LOC100785701), mRNA	1551	0.10
		Glyma.15G201700.1	3PREDICTED: Glycine max uncharacterized LOC100814752 (LOC100814752), mRNA	1945	0.11
		Glyma.15G210100.3	3PREDICTED: Glycine max alpha,alpha-trehalose-phosphate synthase [UDP-forming] 1-like (LOC100797320), transcript variant X6, mRNA	3542	0.12
		Glyma.15G244000.1	3PREDICTED: Glycine max uncharacterized LOC100814749 (LOC100814749), mRNA	1213	0.10
		Glyma.15G261100.1	3PREDICTED: Glycine max uncharacterized LOC100801946 (LOC100801946), transcript variant X1, mRNA	3888	0.09
2019	Chr19	*Glyma.19G110600.1*	3PREDICTED: Glycine max uncharacterized LOC102659858 (LOC102659858), mRNA	1709	0.10
	Chr02	*Glyma.02G220300.1*	2PREDICTED: Glycine max ataxin-2-like (LOC100788042), mRNA	1135	0.18
	Chr04	*Glyma.04G102900.1*		2522	0.10

Chr08	*Glyma.08G071600.1*	2PREDICTED: Glycine max metacaspase-3-like (LOC100796113), transcript variant X2, mRNA	1839	0.12
Chr11	*Glyma.11G229600.1*	3PREDICTED: Glycine max DNA replication complex BAG protein□ transcript variant 2, mRNA□	1257	0.47
Chr12	*Glyma.12G224000.1*	2PREDICTED: Glycine max uncharacterized LOC102660202 (LOC102660202), mRNA	2799	0.17
	Glyma.12G227300.1	2PREDICTED: Glycine max DNA ligase 1-like (LOC100818049), mRNA	2728	0.17
Chr20	*Glyma.20G026100.1*	2PREDICTED: Glycine max 26S proteasome non-ATPase regulatory subunit 7 homolog A-like (LOC100816479), mRNA	896	0.08

The expression of two candidate genes in diverse rapeseed accessions and tissues

In order to validate the candidate genes significantly associated oleic acids content, we

selected two genes involved and measured their gene expression in different tissues (root, stem, leaf and seed) by using qRT-PCR. The lectin gene (GenBank: A5547-127) was used as the reference gene. The results showed that the candidate gene *Glyma.11G229600.1* in soybean seedlings was expressed in different tissues, but the relative expression of genes was significantly different .The *Glyma.11G229600.1* was expressed in soybean leaves, and the relative expression level ranged from 1.23-4.31, the relative expression in stems ranged from 10.21-39.56, and the relative expression in soybean roots ranged from 16.21-43.14 (Table.3). The candidate gene *Glyma.11G229600.1* had the lowest relative expression (1.23) in the leaves of the soybean line q001, that has the lowest content of oleic acid. The relative expression of the *Glyma.11G229600.1* in the soybean line q001 seeds was also the lowest (25.26). The candidate gene *Glyma.11G229600.1* relative expression in leaves of the soybean line q353 was 4.3 times higher than in leaves of the soybean line q001 and in seeds was 10 times higher than in leaves (Additional file 3: Fig. S1). As general conclusion we can say that the relative expression of the candidate gene *Glyma.11G229600.1* in different tissues of soybean seedling is significantly different. The correlation coefficient between *Glyma.11G229600.1* and oleic acid content is 0.980-0.994 (*P□0.01*) (Additional file 4: Table. S3). The correlation coefficient strongly indicate that the candidate gene *Glyma.11G229600.1* plays a positive role in regulating the oleic acid content in the seeds.

The relative expression of *Glyma.04G102900.1* was also analyzed in soybean stems, roots and seeds. The *Glyma.04G102900.1* was expressed in soybean leaves, and the relative expression level ranged from 9.62-44.41, the relative expression in stems ranged from 3.18-28.11 (Table.4). Specifically, in the soybean line q001, that has the lowest oleic acid content in seeds, the candidate gene *Glyma.04G102900.1* showed the highest expression, which was 49.01 ((Additional file 5: Fig. S2)). In general, the relative expression of *Glyma.04G102900.1* in different tissues is significantly different (*P□0.05*). The correlation coefficient between *Glyma.04G102900.1* and oleic acid content is -0.964~-0.998(Additional file 6: Table. S4). The result suggested that *Glyma.04G102900.1* is closely related to the oleic acid content, specifically showing a negative effect on the oleic acid content in seeds.

Table 3. Variance analysis of *Glyma.11G229600.1* expression in different tissues of soybean

Group	Oleic acid content		Name	relative expression in leaves		relative expression in stem		relative expression in roots		relative expression in seed	
	mean	Sig.		Mean	Sig,	mean	Sig	mean	Sig.	mean	Sig.
Soybean lines with high oleic acid	30.01	A	q318	3.41	c	35.11	b	32.24	b	56.27	b
			q20	4.08	c	36.13	b	35.79	b	56.80	b
			q073	3.67	c	35.43	b	35.98	b	56.69	b
			q176	4.13	c	37.12	a	40.64	a	62.06	a
			q070	3.85	c	35.71	b	36.42	b	57.89	b
			q68	4.32	b	38.65	a	40.23	a	62.43	a
			q353	4.61	a	38.19	a	41.01	a	62.73	a
Soybean lines with low oleic acid	14.14	B	q001	1.23	d	10.11	c	16.60	c	25.26	c
			q024	1.86	d	11.11	c	17.45	c	25.75	c
			q020	1.88	d	10.43	c	18.32	c	26.65	c
			q008	1.79	d	10.15	c	18.72	c	26.71	c
			q035	1.85	d	11.01	c	17.99	c	26.68	c
			T7287	1.34	d	10.16	c	18.11	c	26.36	c
			q033	1.90	d	10.54	c	18.23	c	26.81	c

Note: The different uppercase letters indicate significant differences at *P < 0.01*, the different lower letters indicate significant differences at *P < 0.05*, as determined by Duncan’s multiple-range test

Table 4 Variance analysis of *Glyma.04G102900.1* expression in different tissues of soybean

Group	Oleic acid content		Name	relative expression in leaves		relative expression in stems		relative expression in roots		relative expression in seeds	
	mean	Sig.		mean	Sig.	mean	Sig.	mean	Sig.	mean	Sig.
Soybean lines with low	10.31	B	q318	17.32	b	5.13	c	22.23	c	23.02	c
			q20	18.21	b	6.12	c	24.15	c	24.01	c
			q073	9.62	b	5.12	c	16.12	c	22.07	c
			q176	10.29	b	5.61	c	18.75	c	24.08	c
			q070	18.85	b	6.16	c	17.22	c	25.12	c
			q68	17.32	b	3.18	c	17.15	c	23.25	c
			q353	18.11	b	4.52	c	18.75	c	24.28	c
Soybean lines with low oleic acid	30.23	A	q001	44.41	a	28.11	b	38.34	a	49.01	a
			q024	43.21	a	23.13	b	38.46	b	46.52	b
			q020	41.38	a	22.41	b	38.43	a	48.32	a
			q008	42.15	a	23.62	b	38.89	b	45.22	b
			q035	39.25	a	24.14	b	44.56	b	45.01	b
			T7287	41.31	a	25.21	b	40.19	b	45.05	b
			q033	39.25	a	21.02	a	45.21	b	44.79	b

Note: The different uppercase letters indicate significant differences at *P < 0.01*, the different lower letters indicate significant differences at *P < 0.05*, as determined by Duncan's multiple-range test

Discussion

Soybean oil is composed of five fatty acids: palmitic acid (16:0), stearic acid (18:0), oleic acid (18:1), linoleic acid (18:2), and linolenic acid (18:3) [17]. The percentage of these five fatty acids in soybean oil averages 10%, 4%, 18%, 55%, and 13%, respectively. It was found that the content of oleic acid in grains of different soybean varieties varied greatly [18]. Japanese scientist collected 319 Japanese soybean varieties in 2016, among the 319 accessions, the oleic acid content of seeds ranged from 7.66% to 15.86%, and 101 accessions had seed oleic acid levels of 11.5% [19]. In 2008, the cultivated and wild soybean germplasms from different regions were analyzed for their fatty acid content. The result showed that the average fat content of cultivated soybean was 17.21%, 6.22% higher than that of wild soybean; oleic acid content of cultivated soybean was 23.25%, 7.75% higher than the wild; linoleic acid content was 53.53%, 2.57% lower than the wild [20]. In this study, the oleic acid content ranged from 13.5% to 38.4%. The results also showed that the average oleic acid content of soybean germplasm varied greatly, and the oleic acid content of soybean grain was significantly different between regions. In Yaduru's study, the correlation analysis clearly indicated a significant and negative correlation of oleic acid with linoleic acid (r = -0.701, P < 0.0001), stearic acid (r = -0.218, P < 0.001) . Our study is consistent with the results of

Hu Mingxiang [21]. Stearic acid was significantly positively correlated with oleic and arachidic acid, while it had an inverse association with both linoleic and linolenic acids [22]. In our study, the oleic acid content also is significantly positively correlated with linoleic acid (0.454). The results indicated that the relationship between oleic and linoleic acids may be helpful in evaluating varieties that are rich in oleic acid.

Glyma.11G229600.1 located on chromosome 11 was detected by GWAS during 2018 and 2017 simultaneously. *Glyma.11G229600.1* belongs to the plant BAG protein family. The BAG proteins are a broadly conserved gene family with homologs spanning wide evolutionary distances including yeast, animals, and plants [23]. Studies have shown that BAG protein has also been found in rice, Arabidopsis thaliana [24, 25]. The BAG protein family plays an important role in plant growth and development. Overexpression of BAG7 can increase plant sensitivity to temperature, BAG4 is anti-apoptotic genes that have been reported to confer tolerance to salinity and drought stresses in transgenic tobacco [26]. Drought treatment at different growth stages also contributed to differences in fatty acids [27]. The fatty acid composition and amino acid composition were significantly affected by drought stress [28], Severe drought increased protein content by 4.4 percentage points, while oil content decreased by 2.9 percentage points. As drought stress increased, measured by accumulating stress degree days, protein content increased linearly and oil content decreased [29]. It can be speculated that *Glyma.11G229600.1* may increase the drought tolerance of soybean, thus affecting the accumulation of oleic acid in soybean seeds.

Glyma.04G102900.1 belongs to the plant GRAS protein family. GRAS proteins constitute an important family of plant-specific proteins named after the first three members were discovered: gibberellic-acid insensitive (GAI), repressor of gai (RGA) and scarecrow (SCR). At least 33 GRAS genes have been identified in *A.thaliana* and rice [30]. Two GRAS domain proteins have recently been discovered in legumes [31]. Rhizobial bacteria enter a symbiotic interaction with legumes, activating diverse responses in roots through the lipochito oligosaccharide signaling molecule Nod factor. A study showed that GRAS protein transduces calcium signals in plants and provides a possible regulator of Nod-factor–inducible gene expression [32]. In this study, we investigated the expression of two candidate genes in different independent soybean lines by qRT-PCR, finding that the candidate gene expression varied in these lines. We discovered two genes that were correlated with oleic acid content of soybean seed in both 2018 and 2019 analyses. It is the first time that the key genes *Glyma.04G102900.1* and *Glyma.11G229600.1* had been reported to be associated with oleic acid content. Hence, in the future, further studies should support this finding. Our results provide a basis for deciphering the mechanism underlying the determination of fatty acid composition in soybean. Moreover, the SNP markers identified here demonstrate that marker-assisted selection is a powerful strategy for identifying genes of interest in soybean and can be used in breeding programs aimed at optimizing fatty acid profiles in seeds.

Conclusions

In this experiment, the genome-wide association analysis technique (GWAS) was used to find SNP markers correlated with oleic acid content. In 2018, 20 new candidate genes related to oleic acid content were detected, and in 2018, a total of new 8 candidate genes related to oleic acid content were detected. *Glyma.04G102900.1* on chromosome 4 and *Glyma.11G229600.1* on chromosome 11 were detected in both anaylsis (2018 and 2019). *Glyma.04G102900.1* and *Glyma.11G229600.1* are new key candidate genes related to oleic acid in soybean seeds.

Methods

Plant materials

The 260 soybean materials provided by the Biotechnology Center of Jilin Agricultural University were planted in the experimental field of Jilin Agricultural University (Changchun, China) from 2018 to 2019 (totally 2 years). A randomized complete block design was used. The field was divided into three sections (blocks) and those were subdivided into eight sections. After that we did natural drying (sun light) then we take into the seeds were threshed for oleic acid determination. 14 soybean varieties with significantly different oleic acid contents were selected to test the candidate genes expression. The names of the soybean lines and the fatty acid content are shown in Table. 3.

Table 3. Names of 14 soybean lines and average five fatty acids content

Groups	Names of lines	oil	protein	Oleic acid	Linoleic acid	Linolenic acid	Palmitic acid	Stearic acid
Low oleic acid content	q001	21.1	37.62	9.56	67.85	10.77	9.79	1.26
	q024	20.18	39.1	10.42	67.19	8.57	9.02	0.75
	q020	20.94	39.12	11.55	65.56	8.6	8.59	1.01
	q008	17.94	43.49	11.67	64.32	7.75	9.48	1.01
	q035	19.13	39.17	11.99	62.45	9.43	8.73	1.38
	T7287	20.96	37.76	12.23	60.62	7.74	8.22	2.36
	q033	21.1	38.03	12.4	62.92	9.33	9.86	0.81
high oleic acid content	q318	16.88	40.29	25.34	57.71	8.17	2.32	3.27
	q20	22.19	40.42	26.41	54.1	6.04	2.65	1.8
	q073	21.29	41.54	28.08	50.94	7.97	1.72	2.69
	q176	12.36	38.91	29.73	55.13	8.85	0.76	5.06
	q070	20.9	40.99	29.78	51.08	7.09	1.67	1.89
	q68	16.91	39.44	29.94	53.64	6.01	0.56	4.43
	q353	15.59	39.49	31.02	52.18	8.6	0.99	5.73

Determination of Fatty Acids in Soybean Seeds

The content of oleic acid and other four fatty acids (stearic acid, palmitic acid, linoleic acid and linolenic acid) in soybean seeds were determined by NIRSTM DS 2500 (FOSS, Hillerod, Denmark) after harvesting. SPSS version 22.0 software (SPSS Inc, Chicago, IL, USA) was used to calculate the correlation coefficient of fatty acid in soybean seeds.

Genotyping of soybean germplasms

Total genomic DNA was extracted from leave of each soybean line using a CTAB method according to Murray & Thompson (1980) [33]. The 260 soybean materials were genotyped by Specific-Locus Amplified Fragment Sequencing (SLAF-seq) and SNP molecular markers were developed. DNA extraction is the first step in sequencing. SNP molecular markers are used for phylogenetic analysis and genetic evolutionary correlation analysis. The restriction endonuclease combination was *RsaI-HaeIII*. The sequencing service is supported by Beijing Biomarker Biotechnology company.

Population structure evaluation

The Principal component analysis (PCA) was used to assess the population structure using the EIGENSOFT software package. Based on the neighboring method (neighbor-joining), MEGA5 software was used to construct a phylogenetic tree that included each sample.

Genome-wide association analysis (GWAS)

Based on the SNP markers obtained by SLAF-Seq technology, the correlation values between SNP markers and oleic acid content were obtained by using the five models: glm, mlm, cmlm, fastlmm and emmax of TASSEL software. TASSEL software can calculate the Q matrix of sample population structure according to the *K* matrix, and finally get a correlation value of each SNP maker. The results of each model of each trait were annotated based on the 0.000001 level of significance. In this experiment, Manhattan map and QQ map were constructed by using Haploview software. The Manhattan map was used to represent the correlation between genotype data and phenotypic data. The QQ map was used to represent the level of difference between observed and predicted values. In this study, the candidate genes were predicted by using Swiss-Prot and NR databases.

Quantitative reverse transcription-PCR

The qRT-PCR analysis was performed using a Bio-Rad CFX system (Amersham Biosciences, Little Chalfont, Buckinghamshire,UK). The total RNA was extracted using Eastep® Super total RNA extraction Kit (TaKaRa, USA). The amplification reaction conditions were pre-denaturation at 95 °C for 10 min, denaturation at 95 °C for 10 s, annealing at 53 °C for 20 s, extension at 72 °C for 15 s, and amplification reaction conditions for *Glyma.11G229600.1* gene at 95 °C for 10 min, 95 °C. Denaturation for 30s, annealing at 67°C for 30s, extension at 72°C for 30s, 35 cycles, extension at 72°C for 10min, the above reactions are all 40 cycles. After the amplification, the dissolution curve is calculated by $2^{-\Delta\Delta Ct}$ method [34]. The amplification reaction conditions of the gene *Glyma.04G102900.1* are pre-denaturation at 95 °C for 10 min, denaturation at 95 °C for 30 s, annealing at 59 °C for 30 s, extension at 72 °C for 35 s, 35 cycles, and extension at 72 °C for 10 min. Three biological replicates were used for each gene.

Data analysis

The phenotypic data was measured and recorded using Microsoft Excel 2010 software. Differential saliency analysis, analysis of variance, correlation analysis and descriptiveness were performed by using SPSS 19.0 (IBM Corp, Armonk, NY, USA) software [34]. The positive and negative maps and histograms were constructed by using Graphpad Prism software (Graphpad Company, San Diego, CA).

Abbreviations

SNP: Single nucleotide polymorphism; SLAF-seq: Specific-Locus amplified fragment sequencing; GWAS: Genome-wide association study; LD: Linkage disequilibrium; qRT-PCR: Real-time quantitative PCR; cM:

Centi morgan; QTL: Quantitative trait loci

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Availability of data and material

All the data or materials used during this study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

The project was funded by National Genetically Modified New Varieties of Major Projects of China (2018X8004-004). The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

WP designed the experiments. QD, XL planned and performed the experiments. Cristina Miceli and QD edited the manuscript. All authors discussed the results and commented on the manuscript. All authors have read and approved the manuscript.

Acknowledgements

We are grateful for the generous grant from the Biotechnology Center of Jilin Agricultural University and the Jilin Academy of Agricultural Sciences that made this work possible.

Authors' information

[1]School of Biosciences and Veterinary Medicine, University of Camerino, Camerino (MC), Italy. [2]Biotechnology Center of Jilin Agricultural University, Jilin Agricultural University, Changchun 130118, People's Republic of China. [3]College of Life Sciences, Kyung Hee University, 1732, Deogyeong daero, Giheung-gu, Yongin-si, Gyeonggi-do 17104, Republic of Korea.

References

1. Zhao T, Gai J: **The origin and evolution of cultivated soybean [{\sl Glycine max}(L.) Merr.]**. *Zhongguo nongye kexue* 2004, **37**(7):954-962.
2. Tunde-Akintunde T, Olajide J, Akintunde B: **Mass-volume-area related and mechanical properties of soybean as a function of moisture and variety**. *International journal of food properties* 2005, **8**(3):449-456.
3. Cahoon EB: **Genetic enhancement of soybean oil for industrial uses: prospects and challenges**. 2003.
4. Collins F, Sedgwick V: **Fatty acid composition of several varieties of soybeans**. *Journal of the American Oil Chemists Society* 1959, **36**(12):641-644.
5. Patil S, Balu D, Melrose J, Chan C: **Brain region-specificity of palmitic acid-induced abnormalities associated with Alzheimer's disease**. *BMC research notes* 2008, **1**(1):20.
6. Mensink RP, Zock PL, Kester AD, Katan MB: **Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: a meta-analysis of 60 controlled trials**. *The American journal of clinical nutrition* 2003, **77**(5):1146-1155.
7. Sherwin E: **Oxidation and antioxidants in fat and oil processing**. *Journal of the American Oil Chemists' Society* 1978, **55**(11):809-814.
8. Kinney AJ, Knowlton S: **Designer oils: the high oleic acid soybean**. In: *Genetic modification in the food industry.* Springer; 1998: 193-213.
9. Wang X, Pang Y, Zhang J, Wu Z, Chen K, Ali J, Ye G, Xu J, Li Z: **Genome-wide and gene-based association mapping for rice eating and cooking characteristics and protein content**. *Scientific reports* 2017, **7**(1):1-10.
10. Verslues PE, Lasky JR, Juenger TE, Liu T-W, Kumar MN: **Genome-wide association mapping combined with reverse genetics identifies new effectors of low water potential-induced proline accumulation in Arabidopsis**. *Plant physiology* 2014, **164**(1):144-159.
11. Brachi B, Morris GP, Borevitz JO: **Genome-wide association studies in plants: the missing heritability is in the field**. *Genome biology* 2011, **12**(10):232.
12. Hwang E-Y, Song Q, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB: **A genome-wide association study of seed protein and oil content in soybean**. *BMC genomics* 2014, **15**(1):1.
13. Cao Y, Li S, Wang Z, Chang F, Kong J, Gai J, Zhao T: **Identification of major quantitative trait loci for seed oil content in soybeans by combining linkage and genome-wide association mapping**. *Frontiers in plant science* 2017, **8**:1222.
14. Bastien M, Sonah H, Belzile F: **Genome wide association mapping of Sclerotinia sclerotiorum resistance in soybean with a genotyping-by-sequencing approach**. *The Plant Genome* 2014, **7**(1).
15. Kaler AS, Ray JD, Schapaugh WT, King CA, Purcell LC: **Genome-wide association mapping of canopy wilting in diverse soybean genotypes**. *Theoretical and Applied Genetics* 2017, **130**(10):2203-2217.
16. Zhang J, Wang X, Lu Y, Bhusal SJ, Song Q, Cregan PB, Yen Y, Brown M, Jiang G-L: **Genome-wide scan for seed composition provides insights into soybean quality improvement and the impacts of**

domestication and breeding. *Molecular plant* 2018, **11**(3):460-472.

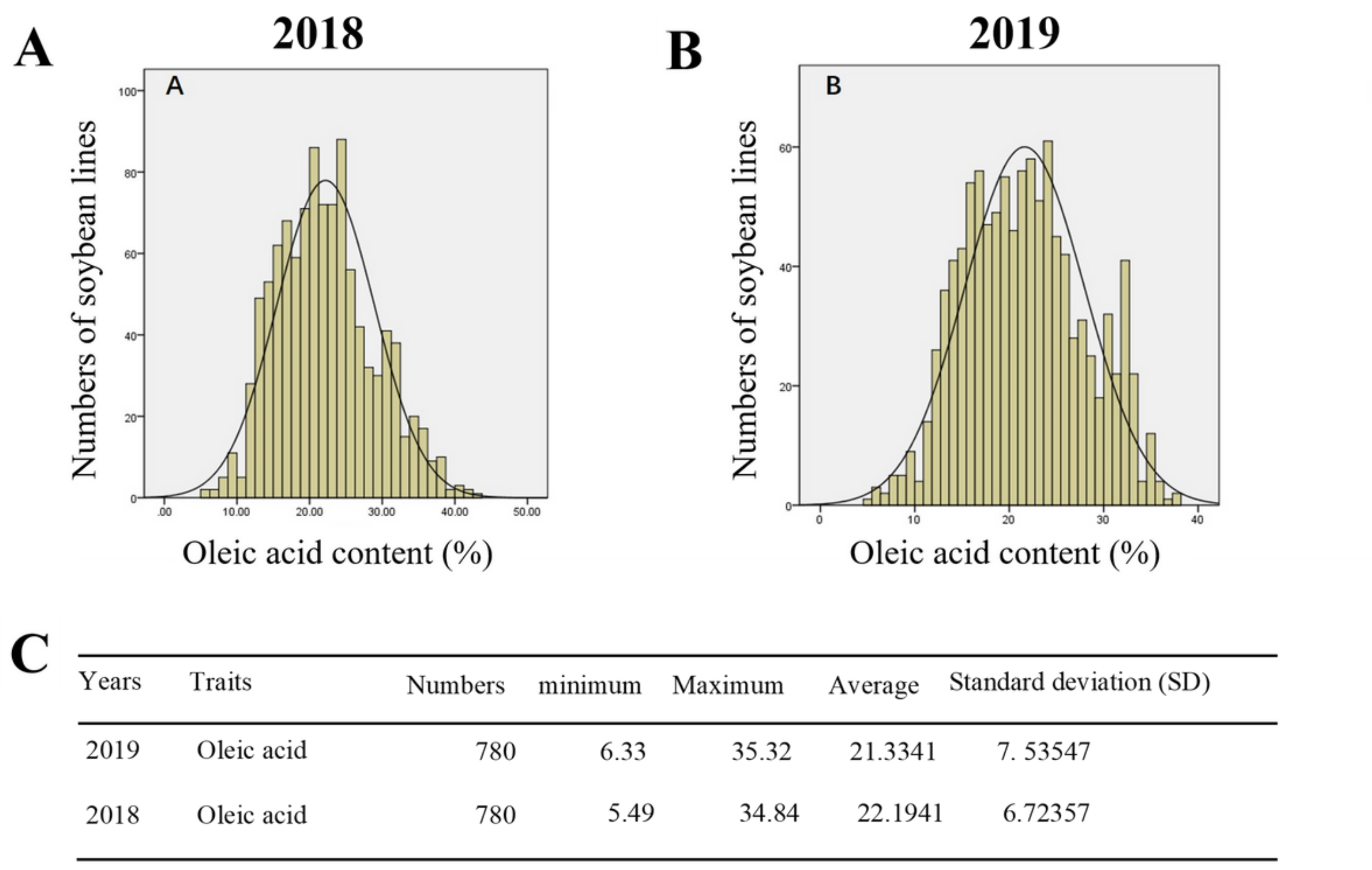
17. Akond M, Liu S, Boney M, Kantartzi SK, Meksem K, Bellaloui N, Lightfoot DA, Kassem MA: **Identification of quantitative trait loci (QTL) underlying protein, oil, and five major fatty acids' contents in soybean**. *American Journal of Plant Sciences* 2014, **2014**.
18. Diers BW, Shoemaker R: **Restriction fragment length polymorphism analysis of soybean fatty acid content**. *Journal of the american oil chemists society* 1992, **69**(12):1242-1244.
19. Wee C-D, Hashiguchi M, Anai T, Suzuki A, Akashi R: **Fatty acid composition and distribution in wild soybean (Glycine soja) seeds collected in Japan**. *Asian J Plant Sci* 2017, **16**:52-64.
20. Yongzhan Z, Junyi G, Tuanjie Z: **A study on variability of fat-related traits in cultivated and wild soybean germplasm in China**. *Scientia Agricultura Sinica* 2008.
21. Xiangxun HMYDM, Suchun Z: **GENETIC STUDIES ON SEED PROTEIN CONTENT OF HYBRID PROGENIES IN SOYBEAN [J]**. *Scientia Agricultura Sinica* 1984, **6**.
22. Kurt C: **Variation in oil content and fatty acid composition of sesame accessions from different origins**. *Grasas y aceites* 2018, **69**(1):241.
23. Williams B, Kabbage M, Britt R, Dickman MB: **AtBAG7, an Arabidopsis Bcl-2–associated athanogene, resides in the endoplasmic reticulum and is involved in the unfolded protein response**. *Proceedings of the National Academy of Sciences* 2010, **107**(13):6088-6093.
24. Rana RM, Dong S, Ali Z, Khan AI, Zhang HS: **Identification and characterization of the Bcl-2-associated athanogene (BAG) protein family in rice**. *African Journal of Biotechnology* 2012, **11**(1):88-98.
25. Yan J, He C, Zhang H: **The BAG-family proteins in Arabidopsis thaliana**. *Plant Science* 2003, **165**(1):1-7.
26. Hoang TM, Moghaddam L, Williams B, Khanna H, Dale J, Mundree SG: **Development of salinity tolerance in rice by constitutive-overexpression of genes involved in the regulation of programmed cell death**. *Frontiers in plant science* 2015, **6**:175.
27. Sánchez-Martín J, Canales FJ, Tweed JK, Lee MR, Rubiales D, Gómez-Cadenas A, Arbona V, Mur LA, Prats E: **Fatty acid profile changes during gradual soil water depletion in oats suggests a role for jasmonates in coping with drought**. *Frontiers in Plant Science* 2018, **9**:1077.
28. Dwivedi S, Nigam S, Rao RN, Singh U, Rao K: **Effect of drought on oil, fatty acids and protein contents of groundnut (Arachis hypogaea L.) seeds**. *Field crops research* 1996, **48**(2-3):125-133.
29. Dornbos D, Mullen R: **Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature**. *Journal of the American Oil Chemists Society* 1992, **69**(3):228-231.
30. Hirsch S, Oldroyd GE: **GRAS-domain transcription factors that regulate plant development**. *Plant signaling & behavior* 2009, **4**(8):698-700.
31. Heckmann AB, Lombardo F, Miwa H, Perry JA, Bunnewell S, Parniske M, Wang TL, Downie JA: **Lotus japonicus nodulation requires two GRAS domain regulators, one of which is functionally conserved in a non-legume**. *Plant physiology* 2006, **142**(4):1739-1750.

32. Smit P, Raedts J, Portyanko V, Debellé F, Gough C, Bisseling T, Geurts R: **NSP1 of the GRAS protein family is essential for rhizobial Nod factor-induced transcription**. *Science* 2005, **308**(5729):1789-1791.

33. Murray M, Thompson WF: **Rapid isolation of high molecular weight plant DNA**. *Nucleic acids research* 1980, **8**(19):4321-4326.

34. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method**. *methods* 2001, **25**(4):402-408.

Figures



Years	Traits	Numbers	minimum	Maximum	Average	Standard deviation (SD)
2019	Oleic acid	780	6.33	35.32	21.3341	7. 53547
2018	Oleic acid	780	5.49	34.84	22.1941	6.72357

Figure 1

Frequency distribution of oleic acid content in soybean seeds from 2018 to 2019. The X-axis on the graph shows the range of content of oleic acid. The Y-axis is the number of soybean lines. (a) The distribution of oleic acid content in 2018, the distribution resembles the bell-shaped curve for a normal distribution. (b) The distribution of oleic acid content of soybean seed in 2019, the distribution resembles the bell-shaped curve for a normal distribution. (c) The statistical analysis of oleic acid content in soybean seeds from 2018 to 2019.

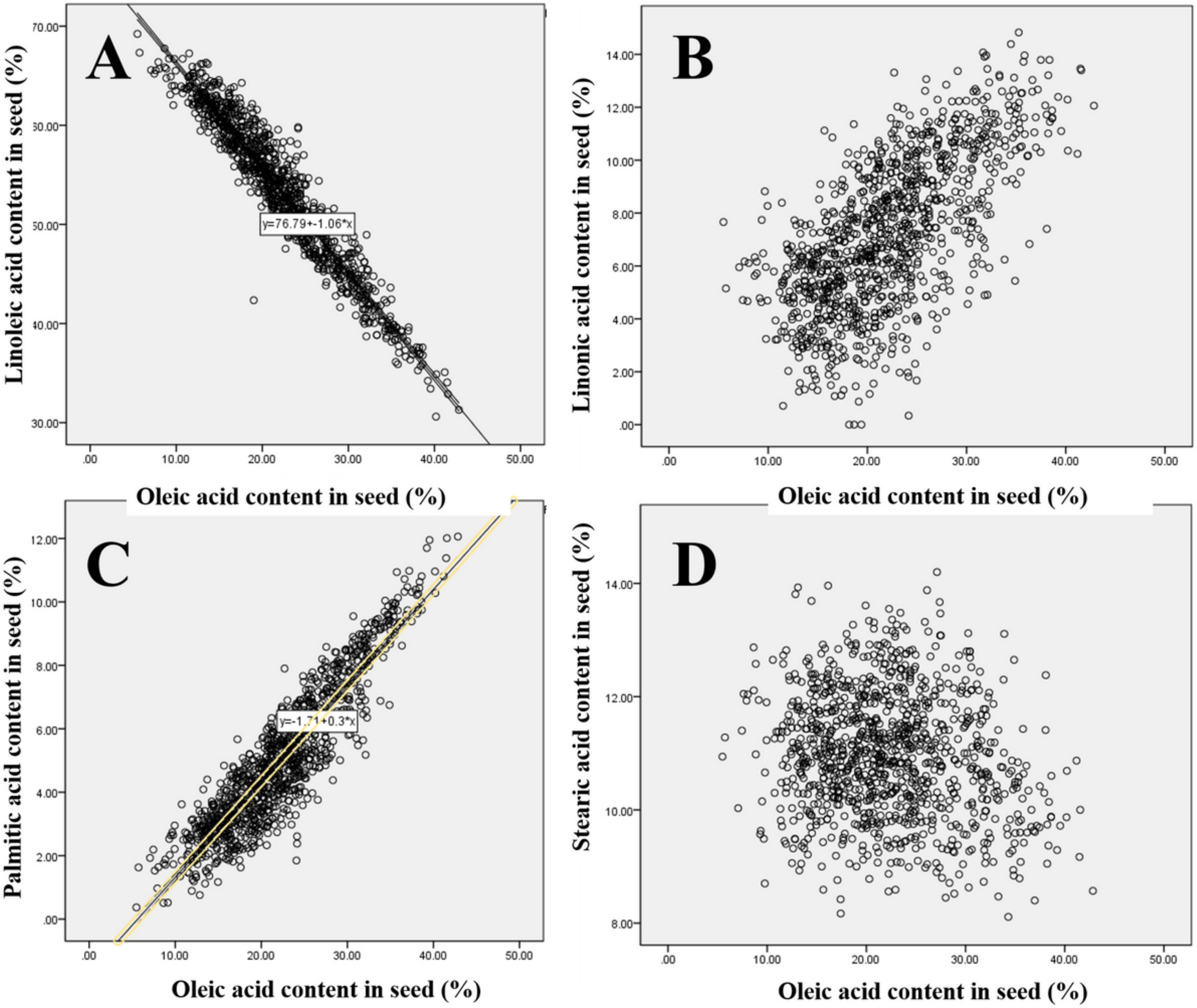


Figure 2

The relationship between oleic acid and other fatty acids. (a) Association between percentages of oleic acid and linoleic acid, y = 1.06x + 76.79, is a linear equation, where y represents linoleic acid content and x represents oleic acid content. (b) Association between percentages of oleic acid and linolenic acid. (c) Association between percentages of oleic acid and palmitic acid, y = 0.3x + 1.71, is a linear equation, where y represents palmitic acid content and x represents oleic acid content. (d) Association between percentages of oleic acid and stearic acid.

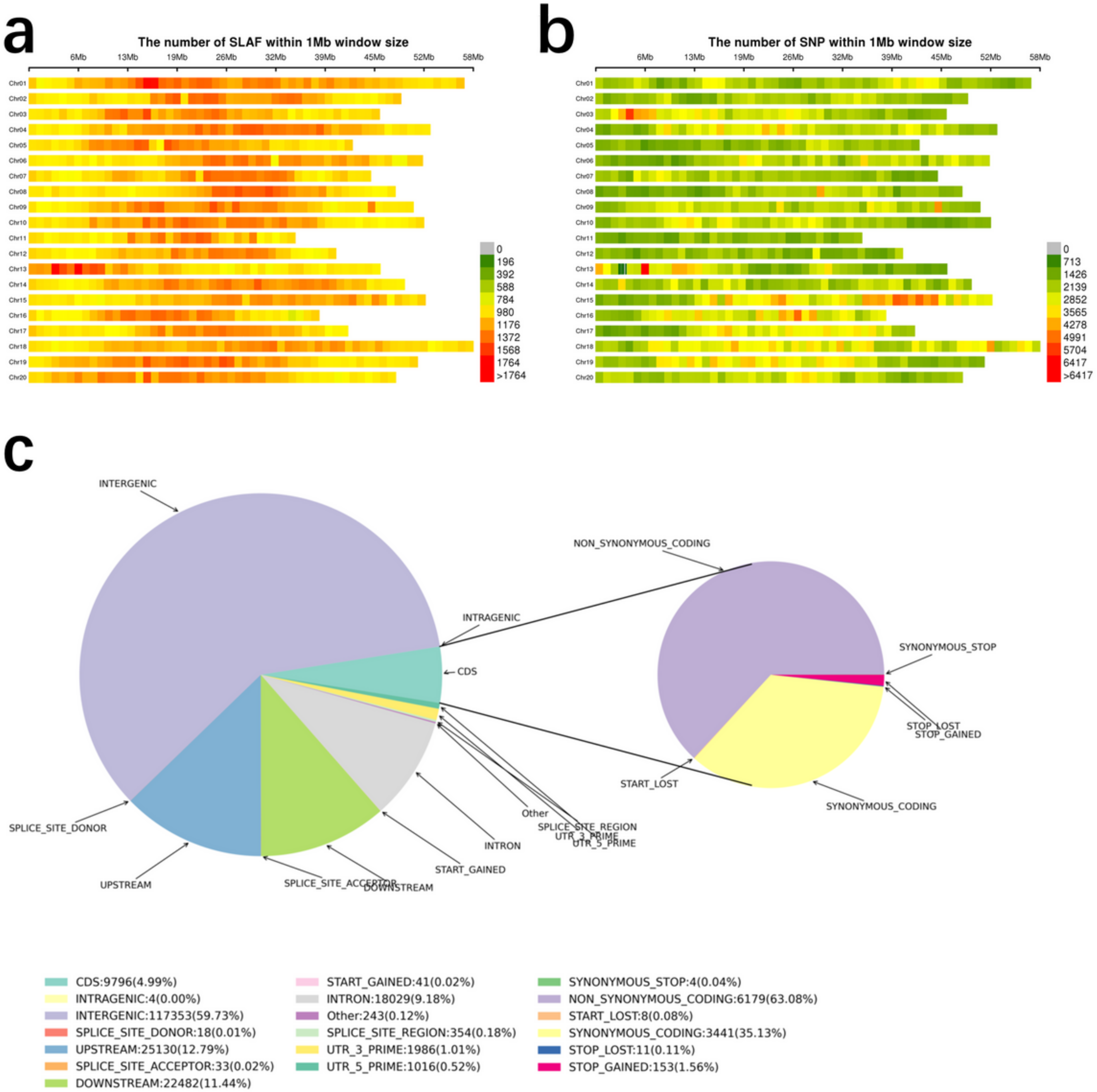


Figure 3

Genotyping and SNP Annotation. (a) Distribution Map of SLAF on chromosomes. The abscissa is the length of the chromosome. Each band represents one chromosome. The genome is divided according to the size of 1Mb. The SLAF label is concentrated in the red area of map. (b) Distribution Map of SNPs on different chromosome. The abscissa is the length of the chromosome. Each band represents one chromosome. The red color represented the number of SNP markers in the region is more than 5706, the yellow color represented the number of SNP markers in the region of the chromosome is between 2852 and 3565. (c) Pie chart of SNPs Annotation. Left panel: SNPs percentages generally associated to genes. Right panel: SNPs percentages in coding regions.

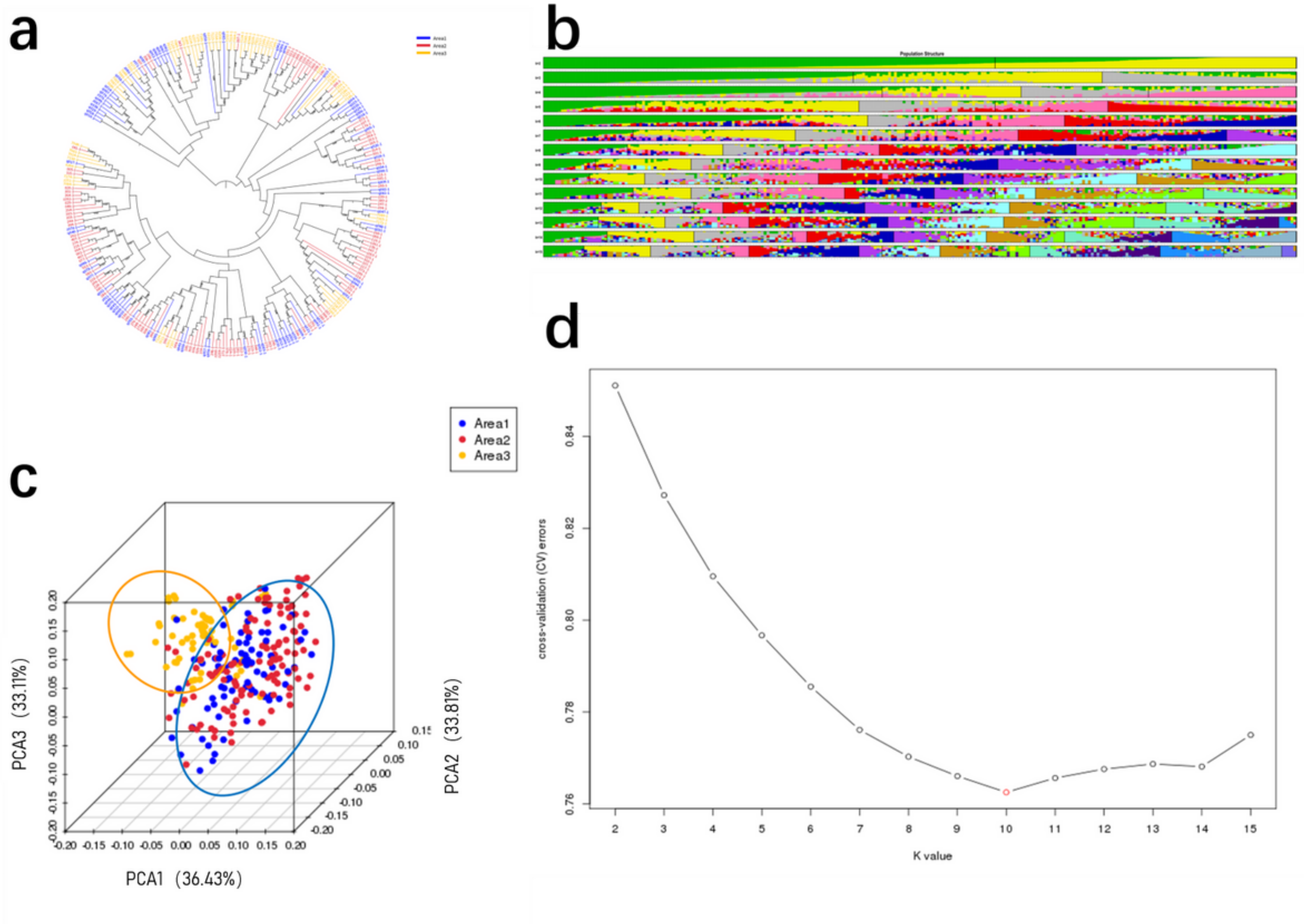


Figure 4

Group Structure of 260 soybean lines (a) Phylogenetic tree of 260 soybean lines. (b) The clustering analysis when the number of subgroups is in the range 2-13, each color is one subgroup. (c) Diagram showing the value of 260 samples based on clustering from 1 to 15; Cross validation error rate for each K value of 1-15, X-axis is K value 1-15, Y-axis is cross-validation errors. (d) Three-dimensional score plot (PC1, PC2, PC3) to discriminate between soybeans lines from three provinces of China. Aea1 represents the soybean lines from Jilin province, Area2 represents the soybean lines from Heilongjiang province, Area3 represents the soybean lines from Liaoning province.

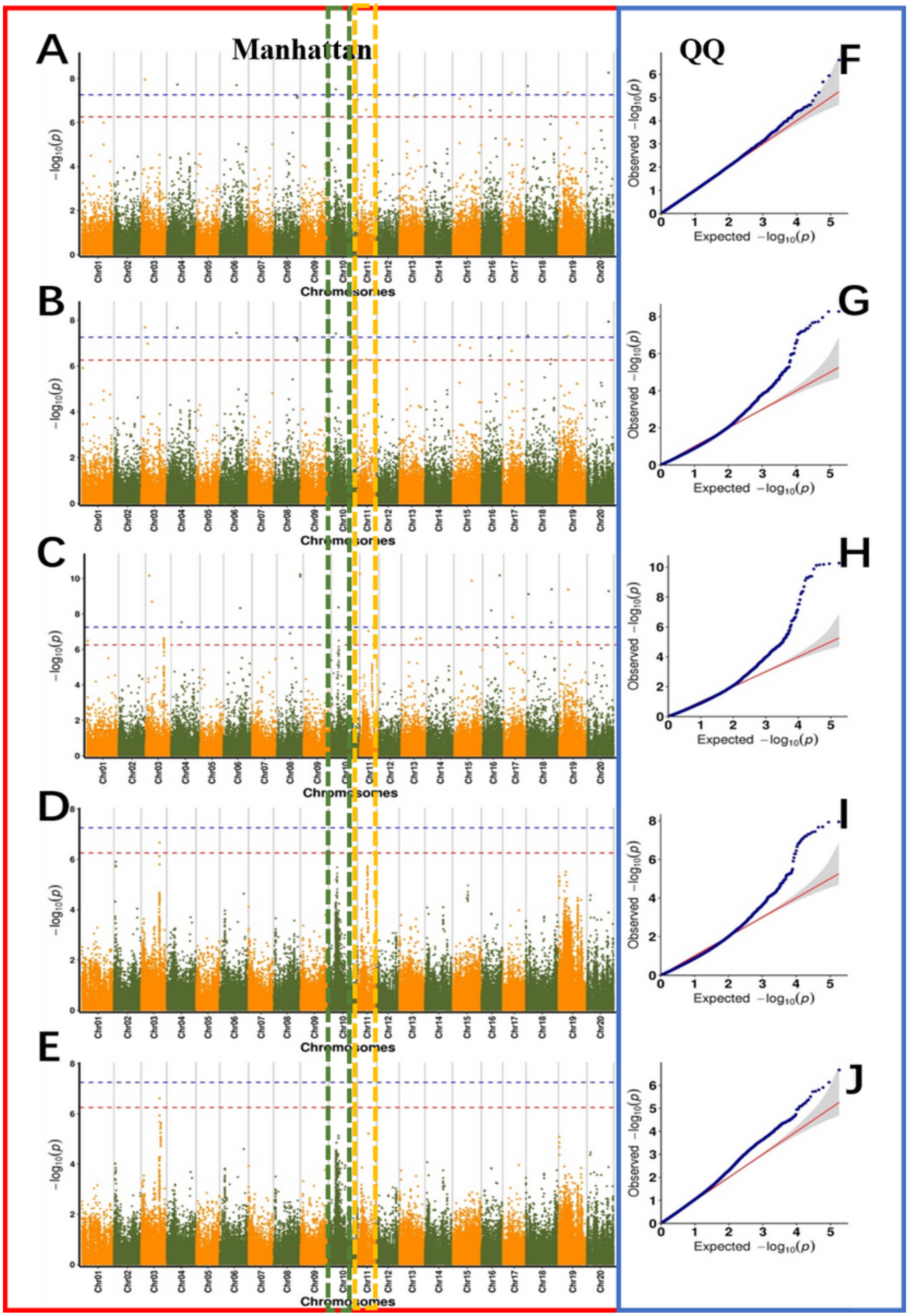


Figure 5

Genome-wide Manhattan plots of associations for oleic acid content for 2019 analysis. In the left panel, the X-axis indicates the SNPs along each chromosome; the Y-axis is the −log 10 (P-value) for the association, the threshold value was set at −log(p) > 6.20 (red) and −log(p) > 7.20 (blue). (a) GWAS result based on the cmlm model, (b) GWAS result based on the EmMax model, (c) GWAS result based on the fastlmm model, (d) GWAS result based on the GLM model, (e) GWAS result based on the MLM model. Q-

Q plots for oleic acid using cmlm model (f), EmMax model (g), fastlmm model (h), GLM (I), MLM (j). The grey area represents the 95% concentration band. Each dot represents a SNP.

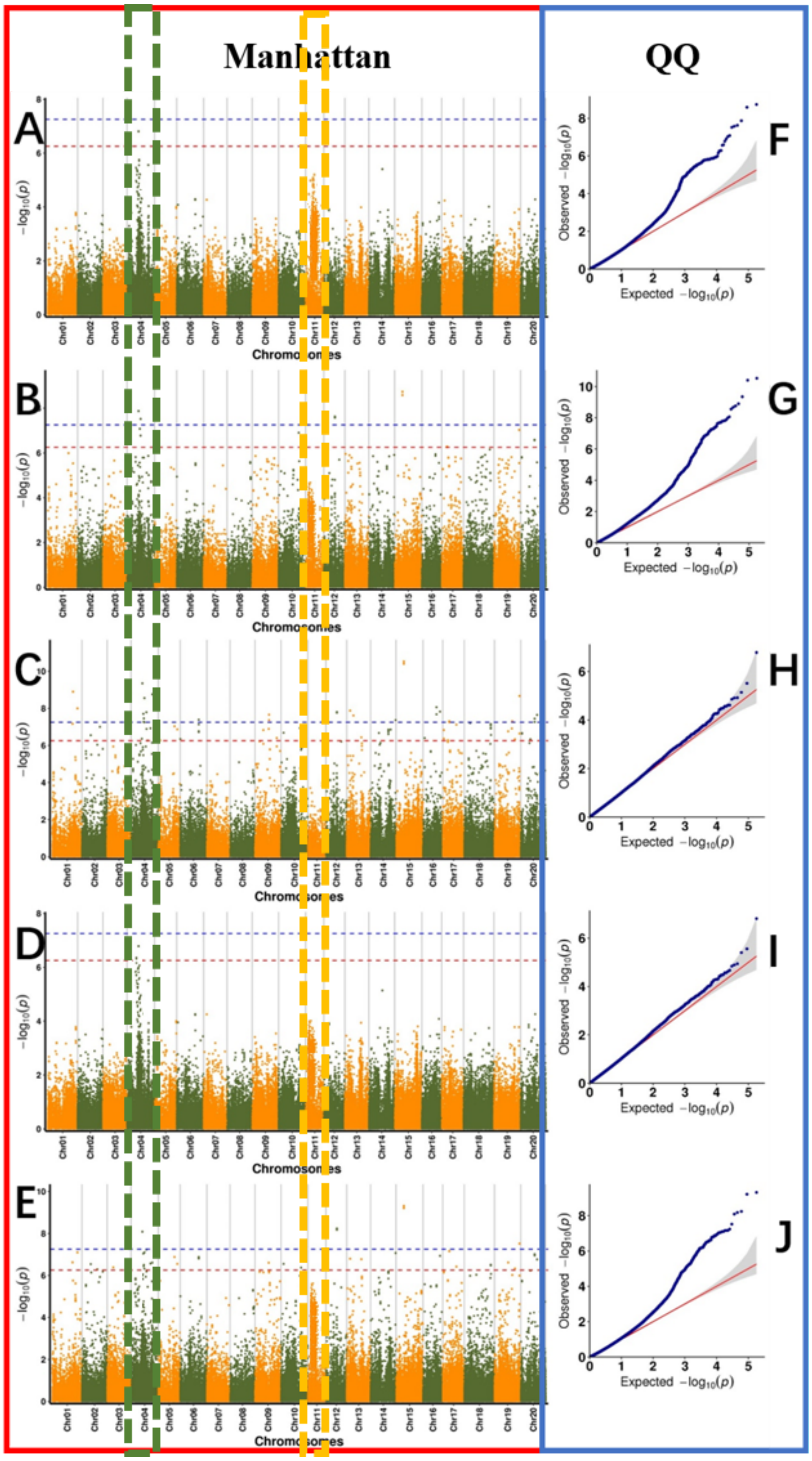


Figure 6

Genome-wide Manhattan plots of associations for oleic acid content for 2019 analysis. In the left panel, the X-axis indicates the SNPs along each chromosome; the Y-axis is the −log 10 (P-value) for the association, the threshold value was set at −log(p) > 6.20 (red) and −log(p) > 7.20 (blue). (a) GWAS result based on the cmlm model, (b) GWAS result based on the EmMax model, (c) GWAS result based on the fastlmm model, (d) GWAS result based on the GLM model, (e) GWAS result based on the MLM model. Q-Q plots for oleic acid using cmlm model (f), EmMax model (g), fastlmm model (h), GLM (I), MLM (j). The grey area represents the 95% concentration band. Each dot represents a SNP.

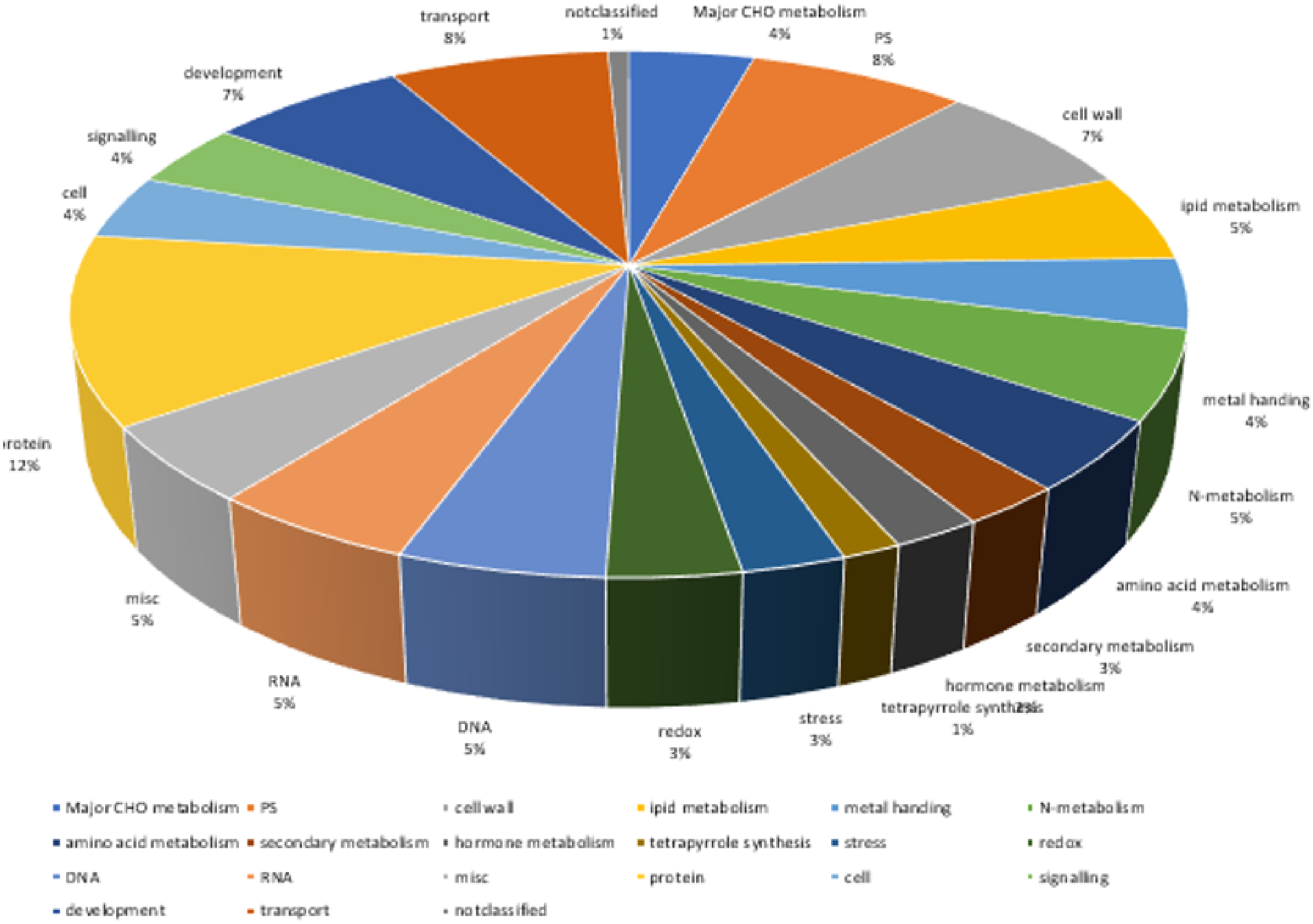


Figure 7

Distribution Map of Candidate Genes in the different functional process.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile5Fig.S2.pptx
- Additionalfile1TableS1.pdf
- Additionalfile3Fig.S1.pptx
- Additionalfile2TableS2.pdf
- Additionalfile6TableS4.pdf

- Additionalfile4.TableS3.pdf