PLOS ONE

RESEARCH ARTICLE

Genome-wide association study of seed coat color in sesame (*Sesamum indicum* L.)

Chengqi Cui[1☯], Yanyang Liu[1☯], Yan Liu[2], Xianghua Cui[3], Zhiyu Sun[4], Zhenwei Du[1], Ke Wu[1], Xiaolin Jiang[1], Hongxian Mei[1]*, Yongzhan Zheng[1]*

1 Henan Sesame Research Center, Henan Academy of Agricultural Sciences, Zhengzhou, Henan, China, **2** Nanyang Academy of Agricultural Sciences, Nanyang, Henan, China, **3** Zhumadian Academy of Agricultural Sciences, Zhumadian, Henan, China, **4** College of Life Sciences, South China Normal University, Guangzhou, Guangdong, China

☯ These authors contributed equally to this work.
* meihx2003@126.com (HM); sesame168@163.com (YZ)

Citation: Cui C, Liu Y, Liu Y, Cui X, Sun Z, Du Z, et al. (2021) Genome-wide association study of seed coat color in sesame (*Sesamum indicum* L.). PLoS ONE 16(5): e0251526. https://doi.org/10.1371/journal.pone.0251526

Editor: Harsh Raman, New South Wales Department of Primary Industries, AUSTRALIA

Received: November 11, 2020

Accepted: April 27, 2021

Published: May 21, 2021

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by Agriculture Research System of China (CARS-14-1-01) to Yongzhan Zheng; the Key Project of Science and Technology of Henan Province (201300110600), and the Key R & D and Promotion Projects of Henan Province (202102110026) to Hongxian Mei; the Science-Technology Foundation for Outstanding Young Scientists (2020YQ26), and the Science & Technology Innovation and Creative

Abstract

Sesame (*Sesamum indicum* L.) is an important and ancient oilseed crop. Sesame seed coat color is related to biochemical functions involved in protein and oil metabolism, and antioxidant content. Because of its complication, the genetic basis of sesame seed coat color remains poorly understood. To elucidate the factors affecting the genetic architecture of seed coat color, 366 sesame germplasm lines were evaluated for seed coat color in 12 environments. The genome-wide association studies (GWAS) for three seed coat color space values, best linear unbiased prediction (BLUP) values from a multi-environment trial analysis and principal component scores (PCs) of three seed coat color space values were conducted. GWAS for three seed coat color space values identified a total of 224 significant single nucleotide polymorphisms (SNPs, $P < 2.34\times10^{-7}$), with phenotypic variation explained (PVE) ranging from 1.01% to 22.10%, and 35 significant SNPs were detected in more than 6 environments. Based on BLUP values, 119 significant SNPs were identified, with PVE ranging from 8.83 to 31.98%. Comparing the results of the GWAS using phenotypic data from different environments and the BLUP values, all significant SNPs detected in more than 6 environments were also detected using the BLUP values. GWAS for PCs identified 197 significant SNPs, and 30 were detected in more than 6 environments. GWAS results for PCs were consistent with those for three color space values. Out of 224 significant SNPs, 22 were located in the confidence intervals of previous reported quantitative trait loci (QTLs). Finally, 92 candidate genes were identified in the vicinity of the 4 SNPs that were most significantly associated with sesame seed coat color. The results in this paper will provide new insights into the genetic basis of sesame seed coat color, and should be useful for molecular breeding in sesame.

Introduction

Sesame (*Sesamum indicum* L., 2n = 2x = 26), which belongs to the *Sesamum* genus of the Pedaliaceae family, is one of the earliest domesticated crops [1]. It is mainly planted in tropical and

Projects (2020CX25) of Henan Academy of Agricultural Sciences to Chengqi Cui; and the Basic Scientific Research Projects (2020JC008, 2021JC013, 2021ZC69) of Henan Academy of Agricultural Sciences to Yanyang Liu. The funders had no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

subtropical regions in Asia, Africa, and South America [2]. Compared with the seeds of other main oil crops, e.g., rapeseed (*Brassica napus*), soybean (*Glycine max*), peanut (*Arachis hypogaea*) and olive (*Olea europaea*), sesame seeds not only have the highest oil content, but also are rich in proteins, vitamins, and specific antioxidants such as sesamin and sesamolin [3, 4]. Because of its high oil quality and high nutritive value, sesame seed is regarded as 'the queen of oil seeds' and one of the best choices for health foods [5].

Seed coat color is one of the most important agronomic traits of sesame. It is related to biochemical functions involved in protein and oil metabolism, antioxidant content, and disease resistance [6]. The natural color of mature sesame seeds is diverse, varying from black to white through different intermediates such as gray, dark brown, brown, pale brown, yellow and dirty white [1]. In general, pale-colored sesame seeds contain more oil than dark-colored ones [6, 7]. Therefore, white sesame seeds are usually used to produce oil, and black sesame seeds are favored as food and medication in China. Significant attention has been paid to the inheritance of seed coat color in sesame. Some early classical genetic studies have suggested that sesame seed coat color is determined by two genes [8, 9], while other reports have indicated that the genetic basis of sesame seed coat color is far more complex, which may involve multiple genes and their interactions [10, 11]. In recent years, the genotyping load and cost has been significantly reduced by the next-generation sequencing (NGS) technologies [12], several high-density genetic maps have been developed and a large number of quantitative trait loci (QTLs) for agronomically important traits have been identified in sesame [13–17], including QTLs for seed coat color [6, 15, 18]. However, QTL mapping efforts using the segregated progeny of a bi-parental cross only enable the detection of a subset of loci/alleles within the crop, and offer limited resolution owing to the small number of informative recombination events between linked genetic loci [19]. As an alternative approach to traditional QTL analysis, the genome-wide association study (GWAS), taking advantage of both the wide phenotypic variation and the high number of historical recombination events in natural populations, has been used for dissecting complex traits in crop species [20, 21], such as rice, maize, soybean, cotton, and rapeseed [22–26]. As an orphan or neglected crop, GWAS analysis in sesame is still limited. Wei et al. [27] re-sequenced 705 diverse sesame germplasm accessions and performed a comprehensive GWAS on 56 agronomic traits for the first time. Using a subset of 400 accessions from the above population, Dossa et al. [28] performed a large-scale GWAS on five traits related to drought tolerance.

In this study, seed coat color of an association-mapping panel comprising 366 sesame germplasm accessions was measured in 12 environments, and 42,781 SNPs were developed by using specific-locus amplified fragment sequencing (SLAF-seq). By performing a large-scale GWAS on seed coat color, significantly associated SNPs and candidate genes were explored. These SNPs and candidate genes will play important roles in understanding the genetic basis of seed coat color in sesame.

Materials and methods

Plant materials and experiment design

In a previous study, 366 diverse sesame lines were selected from the Henan Sesame Research Center (HSRC) sesame germplasm collection, and were assembled into an association-mapping panel [29]. In this study, the panel was used for seed coat color evaluation and marker-trait association analysis.

The association-mapping panel was grown at four locations in China for two to four years: Nanyang (NY, E112.52˚, N33.00˚), from 2013 to 2014; Pingyu (PY, E114.63˚, N32.97˚), from 2013 to 2016; Shangqiu (SQ, E115.65˚, N34.45˚), from 2013 to 2014; and Sanya (SY, E109.50˚,

N18.25˚), from 2012 to 2015. Field experiments were arranged by a randomized complete block design, with two replications under each environment. Each accession was grown in a plot with 23–25 plants in a single row, with a distance of 0.15 m between plants within each row and 0.4 m between rows.

Measurement of seed coat color and statistical analysis

Sesame seeds were harvested from five randomly chosen plants in each row at maturity, and were used to evaluate the seed coat color. Seed coat color was scored using a HunterLab colorimeter (ColorFlex EZ, Hunter Associates Laboratory Inc., Virginia, USA), and decomposed into L, a, and b color space values. The L-value represents brightness (black to white, 0 for black, 100 for white), the a-value represents the color from red to green (positive represents red, negative represents green), and the b-value represents the color from yellow to blue (positive represents yellow, negative represents blue) [30]. Descriptive statistics for sesame seed coat color value for each environment, were computed using the PROC UNIVARIATE procedure ($\alpha = 0.01$) of SAS 8.02 software (SAS Institute, Cary, NC, USA). Best linear unbiased predictions (BLUPs) were used to estimate seed coat color values across multiple environments using the R [31] package "lme4" [32]. The BLUP model for the phenotypic trait was $y_{ijk} = \mu + G_i + E_j + (GE)_{ij} + B_{k(ij)} + \varepsilon_{ijk}$, where μ is the total mean, G_i is the genotypic effect of the ith genotype, E_j is the effect of the jth environment, $(GE)_{ij}$ is the interaction effect between the ith genotype and the jth environment, $B_{k(ij)}$ is the effect of replication within the jth environment, and ε_{ijk} is a random error following $N(0, \sigma_\varepsilon^2)$ [33]. The analysis of variance (ANOVA) was performed using QTL IciMapping V4.0 [34]. Broad sense heritability was calculated as: $H^2 = \sigma_G^2/(\sigma_G^2 + (\sigma_{GE}^2/k) + (\sigma_\varepsilon^2/rk))$, where σ_G^2 is the genotypic variance, σ_{GE}^2 is the genotype by environment variance, σ_ε^2 is the residual variance, k is the number of environments, and r is the number of replications [33]. Principal component analysis (PCA) can transform a set of correlated variables into a substantially smaller set of uncorrelated variables as principal components (PCs), which can capture most information from the original data [35]. Borcard et al. [36] recommended that the variables used in the PCA should be scaled to zero-mean and unit-variance. Therefore, PCA for three color space values was performed using R function "prcomp" with the setting "scale = TRUE" [31]. The first 2 PCs which explained 93%~97% of the total variance in different environments, were retained for GWAS.

Marker-trait association analysis

In a previous study, the association-mapping panel was genotyped by using SLAF-seq, and 89,924 high quality SNPs (minor allele frequency (MAF) $\geq$ 0.01 and integrity $\geq$ 0.7) were identified [29]. In this study, to avoid the possible false SNP affecting the result of GWAS, a set of 42,781 SNP markers with a MAF $\geq$ 0.05 and integrity $\geq$ 0.7 was used to perform marker-trait association analysis. PCA matrix of the 42,781 SNPs was performed using the GCTA software [37]. The kinship (K) matrix was estimated using Tassel 5.0 software [38]. Marker-trait association analysis was performed for three color space values, BLUP values and two PCs of color space values using mixed linear models (PCA+K model) implemented in Tassel 5.0 software [38]. In the PCA+K model, the mixed linear model correcting for both PCA-matrix and K-matrix, were employed to reduce errors from population structure and relative kinship. The uniform Bonferroni threshold was used for the significance of associations between SNPs and traits at the significance level of 0.01. In this study, the threshold was $-\log_{10}(0.01/42{,}781) \approx 6.6$ where 42,781 is the number of SNP markers. Manhattan and QQ plots were drawn using the R package "qqman" [39].

Candidate gene prediction

To define the regions of interest for selection of potential candidate genes, the LD blocks, in which flanking SNP markers had strong LD ($r^2 > 0.6$), were defined as the candidate gene regions [40]. All genes within the same LD block ($r^2 > 0.6$) were considered as candidate genes. For significant SNPs outside of the LD blocks, the 99 kb (the LD decay distance) flanking regions on either side of the markers were used to identify candidate genes [29]. LD heatmaps surrounding peaks in the GWAS were constructed using the R package "LDheatmap" [41].

Results

Phenotypic variations of sesame seed coat color

To evaluate the phenotypic variation of seed coat color in the sesame association panel, three color space values (L-value, a-value, and b-value) in each environment and BLUP values across multiple environments were analyzed (Fig 1 and S1 Fig). Descriptive statistics for seed coat color were presented in S1 Table. The sesame association panel exhibited wide variations in seed coat color. The L-value exhibited a wide range of 10.53 to 63.40, with the coefficient of variation (CV) ranging from 14.08 to 22.94% among different environments. Similarly, the a-value ranged from 0.08 to 11.22, with CV ranging from 24.07 to 37.40%, and the b-value ranged from -0.47 to 18.75, with CV ranging from 15.51 to 24.50%. Because L-value represents brightness ranging from black to white (0 for black, 100 for white), a-value represents the color

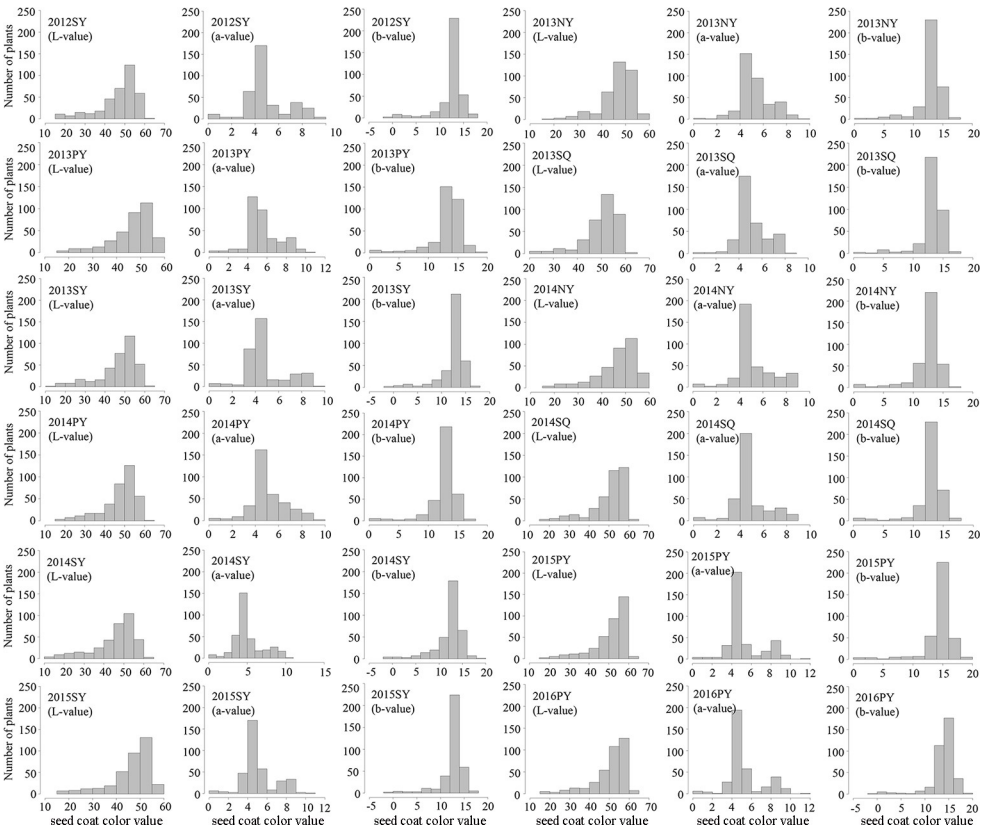

Fig 1. Histograms for the frequency distribution of three color space values (L-value, a-value and b-value).

https://doi.org/10.1371/journal.pone.0251526.g001

from red to green (positive represents red, negative represents green), and b-value represents the color from yellow to blue (positive represents yellow, negative represents blue), the measured values and distributions indicate that black, white, red, and yellow are predominant in the sesame seed coat color, which is consistent with the observation that the seed coat color distributions in the association panel (Figs 1 and 2 and S1 Fig). ANOVA was performed to reveal the effects of G (genotypes), E (environment) and G × E (interaction between G and E) for seed coat color trait in multi-environments. The results showed that there were highly significant differences among G, E, and G × E ($P < 0.01$). The broad-sense heritability of the L-value was calculated to be 98.16%, while the broad-sense heritability of the a-value and b-value was 97.55% and 96.88%, respectively.

PCA was performed for three space color values to investigate the relationships among three space color value variables. PC1 explained 56%~65% of the trait variances in different environments, and three space color values showed high negative loadings on PC1. This result suggested that seed coat color with high PC1 scores exhibited samll values for L-value, a-value and b-value. PC2 explained 34%~43% of the trait variances. Cumulative Proportion of variances for PC1 and PC2 were 93%~97%, and the PCA results were consistent with each other across different environments (S2 Table), suggesting that PC1 and PC2 can be used as quantitative indices to characterize sesame seed coat color.

Genome-wide association analysis for sesame seed coat color

To uncover the genotypic variation of seed coat color in sesame, GWAS were performed for three color space values from different environments and BLUP values across all environments. Using three color space values, a total of 224 significant SNPs ($P < 2.34\times10^{-7}$) were identified in 12 environments (Fig 3), and the R^2, the phenotypic variation explained (PVE) by SNPs, ranged from 1.01% to 22.10%. As shown in quantile-quantile plots (S2 Fig), the genomic inflation was considerably controlled. Among 224 significant SNPs, 35 were detected in more than 6 environments, 24 were detected in more than 8 environments, and 14 were detected in more than 10 environments (S3 Table). Using BLUP values, 119 significant SNPs were identified, with PVE ranging from 8.83 to 31.98% (S3 Fig). Comparing the results of the GWAS using phenotypic data from different environments and the BLUP values, all significant SNPs detected in more than 6 environments were also detected using the BLUP values (S3 Table).

Regarding L-value, 38 significant SNPs were detected on 5 linkage groups (LGs), with PVE ranging from 8.75% to 21.90%. Among these SNPs, 24 were detected using the BLUP values of

Fig 2. Seed coat color variation in sesame association panel.

https://doi.org/10.1371/journal.pone.0251526.g002

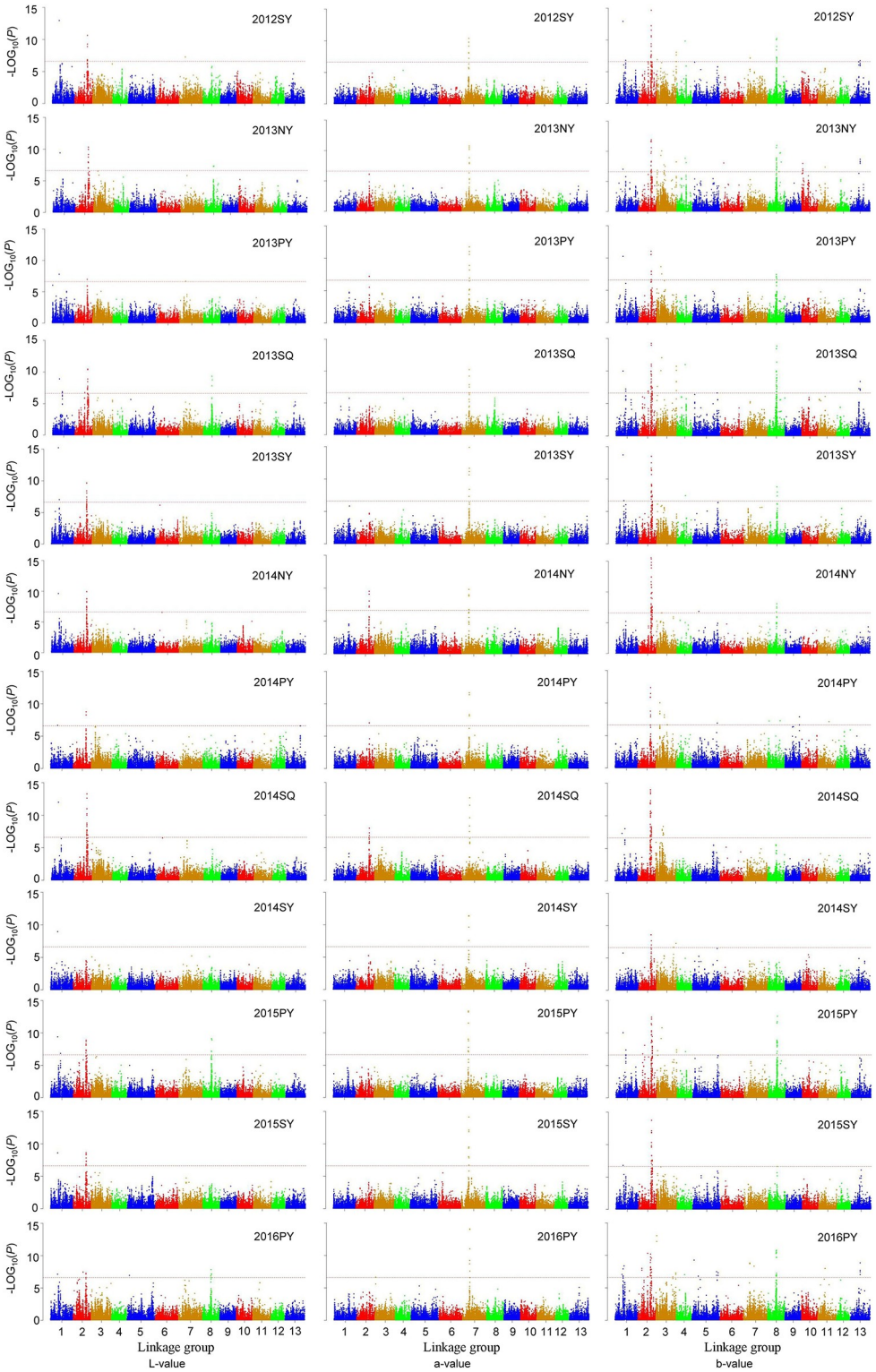

Fig 3. Genome-wide association studies (GWAS) of seed coat color in twelve environments. The red horizontal dashed lines indicate the genome-wide significance threshold ($P < 2.34 \times 10^{-7}$).

https://doi.org/10.1371/journal.pone.0251526.g003

L-value. The most significant SNP S1_6648896 on LG1 was detected in all 12 environments and was also detected using the BLUP values. On LG2, 8 multi-environment significant SNPs (S2_12167303, S2_12178804, S2_12178823, S2_12194998, S2_12232894, S2_12232938, S2_12447358, S2_12247409) were significantly associated with L-value in 7, 8, 8, 8, 7, 10, 8, and 9 environments and were also detected using the BLUP values (S3 Table). Regarding a-value, 17 significant SNPs were identified on LG2, LG3 and LG7, and 9 were detected using the BLUP values of a-value. Of all the significant SNPs, S7_6839839 was detected in all 12 environments and was also detected using the BLUP values, (S3 Table). Regarding the b-value, 169 significant SNPs distributing on LG1, LG2, LG3, LG4, LG5, LG6, LG7, LG8, LG9, LG10, LG11 and LG13 were identified, with PVE ranging from 8.68% to 31.35%. The Manhattan plots showed that 3 peaks on LG1, LG2, and LG8 were repeatedly detected in more than 6 environments and were also identified using BLUP values of b-value. Nine significant SNPs were detected on LG1. The SNP S1_6648896 with the lowest *P* value on LG1 was detected in 9 environments and was also detected using BLUP values. Seventy significant SNPs were detected on LG2. S2_12168663 and S2_12337057 were both detected in 7 environments. S2_12336812 was detected in 8 environments. S2_12167303 and S2_12247358 were detected in 9 environments. S2_12026452, S2_12178804, S2_12178823 and S2_12194998 were detected in 10 environments. S2_12015779, S2_12015820 and S2_12247409 were detected in 11 environments. S2_12232894 and S2_12232938 were detected in 12 environments. These 14 SNPs were also detected using BLUP values. On LG8, 4 multi-environment significant SNPs (S8_7910606, S8_8220220, S8_8311600, S8_8313501) were significantly associated with b-value in 7, 6, 6, and 7 environments and were also identified using BLUP values (S3 Table).

GWAS for PC1 and PC2 identified 197 significant SNPs ($P < 3.3\times10^{-7}$); however, significant SNPs were not detected for PC3 (S4 Fig; S4 Table), which indicated that PC3 might be composed of nongenetic factors. The quantile-quantile plots were shown in S5 Fig. Among 197 significant SNPs, 30 were detected in more than 6 environments, 19 were detected in more than 8 environments, and 14 were detected in more than 10 environments. For PC1, the GWAS results were consistent with those for L-value and b-value. One hundred and eighty-eight significant SNPs were identified on 12 LGs, explaining 8.68–33.93% of the phenotypic variation. Four peaks on LG1, LG2, LG4, and LG8 were repeatedly detected in more than 6 environments. The most significant SNP S1_6648896 on LG1 was repeatedly detected in 9 environments, explaining 12.93%~20.51% of the phenotypic variation. Nineteen significant SNPs on LG2 were indentified in more than 6 environments. The most significant SNP S2_12232938 on LG2 with PVE of 11.95~33.93% was detected in 12 environments. The most significant SNP S4_7766099 on LG4 was repeatedly detected in 6 environments, and explained 9.47%~15.26% of the phenotypic variation. Three significant SNPs on LG8 were detected in more than 6 environments. The most significant SNP S8_8313501 on LG8 was repeatedly detected in 8 environments, and explained 9.47%~15.26% of the phenotypic variation. The GWAS results for PC2 were consistent with those for a-value. Six significant SNPs on LG7 were detected in more than 6 environments. The most significant SNP S7_6839839 was repeatedly detected in 12 environments, and explained 14.14%~26.18% of the phenotypic variation.

Candidate genes associated with sesame seed coat color

To predict the putative genes associated with sesame seed coat color, we focused on the most reliable and stable peaks on different LGs, including S1_6648896, S2_12232938, S7_6839839 and S8_8313501 (Fig 4). The haplotype analysis showed that the SNPs S1_6648896, S2_12232938 and S7_6839839 were all in genomic regions that were in state of linkage equilibrium, while S8_8313501 was involved in a 213-kbp LD block. Within the LD block

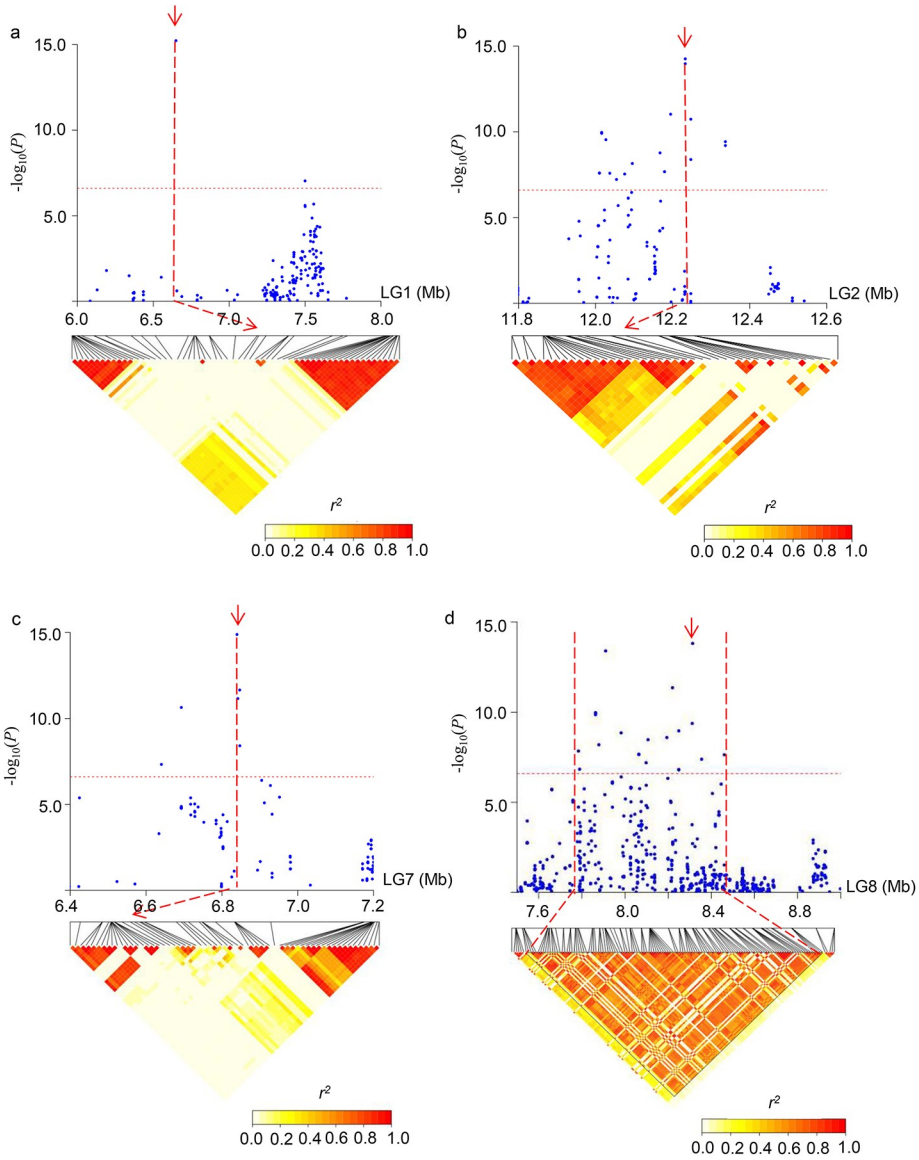

Fig 4. Local Manhattan plot (top) and LD heatmap (bottom) surrounding each peak on different linkage groups. (a) LD heatmap on LG1. The red arrow denotes the SNP S1_6648896; (b) LD heatmap on LG2. The red arrow denotes the SNP S2_12232938; (c) LD heatmap on LG7. The red arrow denotes the SNP S7_6839839; (d) LD heatmap on LG8. The red arrow denotes the SNP S8_8313501.

https://doi.org/10.1371/journal.pone.0251526.g004

(S8_8313501), or 99 kbp either side of the SNPs (S1_6648896, S2_12232938 and S7_6839839), a total of 21, 20, 31 and 20 genes were identified, respectively (S5 Table). Of the 92 genes, 26 had no definite annotation concerning their biological functions, and 12 were annotated as putative or probable proteins. The remaining 54 genes had domains of known functions. Gene ontology (GO) analysis indicated that 40, 39 and 31 genes were involved in the cellular component category, the molecular function category and the biological process category, respectively. In the cellular component category, these genes were grouped into cell (39 genes), cell part (39 genes) and organelle (36 genes) subcategories. Within the molecular function category, the majority of genes were involved in catalytic activity (14 genes), binding (15 genes), transcription regulator activity (6 genes). In the biological process category, most gene were

annotated to metabolic process (23 genes), cellular process (31 genes), response to stimulus (20 genes).

Discussion

GWAS has become an efficient and powerful tool at identifying genetic variations and loci responsible for the agronomically important traits. In 2015, a GWAS of oil quality and agronomic traits with 705 sesame lines identified several causative genes, demonstrating the feasibility of GWAS in sesame [27]. In the present study, the panel of sesame accessions with wide geographic distribution, plentiful phenotype variation, sufficient genetic variation and weak population structure is advantageous for GWAS implementation [29]. However, the reliability of GWAS is usually disturbed by phenotypic variance associated with the environment. Multi-environment analysis and unbiased predictions are practical ways to correct for this error [25]. The trait experiments were performed at four sites, which belong to three climate classifications, temperate monsoon climate (PY and SQ), subtropical monsoon climate (NY), and tropical marine monsoon climate (SY). Among four sites, there are large differences in geographic position and climate. ANOVA showed that significant variations were observed in G, E and G×E. This result suggested that sesame seed coat color was controlled by the genetic, environment effect and their interaction. Then, GWAS for coat color traits were performed in 12 environments, and many significant SNPs were only detected in a specific environment. However, the SNPs detected in more than 6 environments were detected using BLUP values in a multi-environment trial analysis. These multi-environment SNPs are reliable and will be used for further analysis. Therefore, the multi-environment trial analysis could effectively avoid influences from the environments, and is the way forward in the study of complex quantitative traits.

PCA is an effective approach for collecting information from complex, multiple traits that are highly correlated; furthermore, it is valuable for extracting underlying factors for traits by dimension reduction [35]. As PC scores represent integrated variables, they can result in robust, reliable GWAS results [35]. In this study, PCA on three space values (L-value, a-value and b-value) revealed that PC1 captured 56%~65% of variations for all values, PC2 captured 34%~43% of variations for L-value and a-value. Cumulative Proportion of variances for PC1 and PC2 were 93%~97% (S2 Table). Thus, PC1 and PC2 are good indicators for sesame seed coat color. Using the three color space values, 224 significant SNPs ($P < 2.34\times10^{-7}$) were identified. After combining the same SNPs associated with different seed coat color values (L-value, a-value and b-value), 185 SNPs were remained. Using the PC scores (PC1 and PC2) for GWAS, 201 significant SNPs associated with PCs were identified. The GWAS results for PC1 and PC2 were consistent with those for three color space values, indicating PC1 and PC2 can represent three space color values to perform GWAS.

To further confirm these significant SNPs associated with seed coat color in this paper, we compared our GWAS results with QTLs from previous linkage studies. Wang et al. [15] identified 4 QTLs (*qSCa-4.1/qSCb-4.1/qSCl-4.1*, *qSCa-8.1/qSCb-8.1/qSCl-8.1*, *qSCl-8.2*, and *qSCb-11.1/qSCl-11.1*) for seed coat color in a RIL population. Most of QTLs (3/4 QTLs) were verified by significant SNPs in the present study. Eighteen significant SNPs on LG2 were mapped to the confidence interval of the QTL *qSCa-4.1/qSCb-4.1/qSCl-4.1*. One significant SNP (S1_6648896) and three significant SNPs (S1_9324398, S1_9330855 and S1_9332327) on LG1 were mapped to the confidence intervals of QTLs *qSCa-8.1/qSCb-8.1/qSCl-8.1* and *qSCl-8.2*, respectively. These comparison results corroborated our findings. Zhang et al. [6] found 4 QTLs (*QTL1-1*, *QTL11-1*, *QTL11-2*, and *QTL13-1*) for sesame seed coat color, however, because of AFLP markers having been mainly used in the study of Zhang et al. in an

independent genetic map, it is difficult to determine the relationship of the present loci to them. The remaining SNPs, which were not mapped to the confidence intervals of reported QTLs, indicated the likely existence of new seed coat color-related sites or environment-specific SNPs.

Considering SNPs detected in the most environments with high genetic affect, 4 reliable and stable peaks on 4 LGs were focused on, and 92 candidate genes in the vicinity of 4 significant SNPs were identified. For the 4 SNPs (S1_6648896, S2_12232938, S7_6839839 and S8_8313501), the annotation genes included pentatricopeptide repeat-containing protein (SIN_1006005, SIN_1006010, SIN_1012034), malate dehydrogenase (SIN_1006006), basic helix-loop-helix (BHLH) DNA-binding superfamily protein (SIN_1006020 and SIN_1024895), cytochrome P450 94A2 (SIN_1006022), polyphenol oxidases (SIN_1016759 and SIN_1023237), F-box/LRR-repeat protein 3 (SIN_1023224), etc. SIN_1016759 encodes a predicted polyphenol oxidase (PPO), which participates in the oxidation step in the biosynthesis of proanthocyanidin, lignin, and melanin, and produces black pigments via the browning reaction in plants [42–44]. In sesame, Wei et al. [27] reported that SIN_1016759 was strongly associated with seed coat color, Wang et al. and Wei et al. [15, 18] showed that SIN_1016759 was located in the genomic region of a major QTL for seed coat color. qRT-PCR showed that SIN_1016759 was highly expressed in black sesame seeds from 11 to 20 days but not expressed in white sesame seeds [18], indicating that SIN_1016759 may play an important role in the formation of sesame black coat color. SIN_1023237 encodes a laccase-3 which belongs to multicopper oxidase family [45]. Laccase enzymes were shown to contribute toward cell morphology, secondary cell-wall biosynthesis, and resistance to biotic and abiotic stresses in plant [46]. They also play major roles in proanthocyanidins and lignin deposition and are involved in browning reactions on seed coat pigments [42, 43, 47]. SIN_1006022 encodes a cytochrome P450 protein, and may be related to the formation of seed coat color. Cytochromes P450 play important roles in biosynthesis of flavonoids and their coloured class of compounds, anthocyanins, which are responsible for the pigmentation pattern of vegetative parts and seed [48–51]. SIN_1023226 encodes a WRKY-type transcription factor, which is one of the WRKY family members [52]. The WRKY genes family in flowering plants encode a large group of transcription factors which play essential roles in diverse stress responses, developmental, and physiological processes [53]. SIN_1024895 encodes a bHLH transcription factor. Plant bHLHs are involved in secondary metabolism (including the flavonoid pathway), organ development and responses to abiotic stresses [54–56]. Previous reports have shown that the *WRKY* and *bHLH* genes are involved in regulation of seed coloration [57–60].

Conclusions

In this study, GWAS for sesame seed coat color were performed using 42,781 SNPs with 366 sesame germplasm lines in 12 environments. GWAS for three color space values, BLUP values from a multi-environment trial analysis and PCs of three color space values identified 224, 119, and 197 significant SNPs, respectively. The 35 significant SNPs detected in more than 6 environments were also detected using the BLUP values. Furthermore, GWAS results for PCs were consistent with those for three color space values. Multiple QTLs reported in previous studies were verified by significant SNPs in the present study, corroborating the GWAS results. Moreover, the most reliable and significant SNPs (S1_6648896, S2_12232938, S7_6839839 and S8_8313501) on 4 different LGs were focused on, and 92 candidate genes were identified. The GWAS showed great power in uncovering genetic variation in sesame seed coat color, and the results will provide new insights into the genetic basis of sesame seed coat color.

Supporting information

S1 Fig. Histograms for the frequency distribution of BLUP values for three color space values.
(TIF)

S2 Fig. Quantile-quantile plots of observed versus expected–log10(P) values of GWAS results for three seed coat color space values.
(TIF)

S3 Fig. GWAS for BLUP values of three color space values.
(TIF)

S4 Fig. GWAS for PCs in twelve environments.
(TIF)

S5 Fig. Quantile-quantile plots of observed versus expected–log10(P) values of GWAS results for PCs.
(TIF)

S1 Table. Descriptive statistics of seed coat color across 12 environments.
(XLSX)

S2 Table. Summary of the first 3 PCs (PC1, PC2, PC3) for three color space values in the dataset of 366 sesame varieties.
(XLSX)

S3 Table. SNPs significantly associated with three seed coat color space values in more than 6 environments.
(XLSX)

S4 Table. SNPs significantly associated with PCs scores across more than 6 environments.
(XLSX)

S5 Table. Candidate genes linked genomic region of SNP most highly associated with seed coat color in sesame.
(XLSX)

Author Contributions

Conceptualization: Yongzhan Zheng.

Data curation: Chengqi Cui.

Funding acquisition: Chengqi Cui, Yanyang Liu, Hongxian Mei, Yongzhan Zheng.

Investigation: Yanyang Liu, Yan Liu, Xianghua Cui, Zhiyu Sun, Zhenwei Du, Ke Wu, Xiaolin Jiang, Hongxian Mei.

Methodology: Hongxian Mei.

Project administration: Hongxian Mei.

Supervision: Yanyang Liu.

Writing – original draft: Chengqi Cui.

Writing – review & editing: Yanyang Liu, Zhenwei Du, Ke Wu, Hongxian Mei.

References

1. Bedigian D and Harlan JR. Evidence for cultivation of sesame in the ancient world. Econ Bot. 1986; 40 (2):137–154. https://doi.org/10.1007/BF02859136
2. Ashri A. Sesame breeding. In: Janick J. Editor(s). Plant Breeding Reviews. New York: John Wiley & Sons Inc; 1998. pp.179–228.
3. Moazzami AA and Kamal-Eldin A. Sesame seed is a rich source of dietary lignans. J Am Oil Chem Soc. 2006; 83(8):719–723. https://doi.org/10.1007/s11746-006-5029-7
4. Li C, Miao H, Wei L, Zhang T, Han X, Zhang H. Association mapping of seed oil and protein content in *Sesamum indicum* L. using SSR markers. PLoS One. 2014; 9(8):e105757. https://doi.org/10.1371/journal.pone.0105757 PMID: 25153139
5. Wang L, Yu S, Tong C, Zhao Y, Liu Y, Song C, et al. Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. Genome Biol. 2014; 15(2):R39. https://doi.org/10.1186/gb-2014-15-2-r39 PMID: 24576357
6. Zhang H, Miao H, Wei L, Li C, Zhao R, Wang C. Genetic analysis and QTL mapping of seed coat color in sesame (*Sesamum indicum* L.). PLoS One. 2013; 8(5):e63898. https://doi.org/10.1371/journal.pone.0063898 PMID: 23704951
7. Mei H, Wei A, Liu Y, Wang C, Du Z, Zheng Y. Variation and correlation analysis of sesamin, oil and protein contents in sesame germplasm resources. China Oils & Fats. 2013; 38(4):87–90. https://doi.org/10.3969/j.issn.1003-7969.2013.04.023
8. Gutierrez E, Monteverde E, Quijada P. Inheritance of seed coat color and number of locules per capsule in three cultivars of sesame Sesamun indicum L. Agronomia Trop. 1999; 44:513–527.
9. Baydar H and Turgut I. Studies on genetics and breeding of sesame (*Sesamum indicum* L.) I. Inheritance of the characters determining the plant type. Turk J Biol. 2000; 24:503–512. https://doi.org/10.1117/12.775875
10. Falusi OA. Segregation of genes controlling seed colour in sesame (*Sesamum indicum* linn.) from Nigeria. Afr J Biotech. 2007; 6(24):2780–2783. https://doi.org/10.5897/AJB2007.000–2444
11. Pandey SK, Das A, Dasgupta T. Genetics of seed coat color in sesame (*Sesamum indicum* L.). Afr J Biotech. 2013; 12(42):6061–6067. https://doi.org/10.5897/AJB2013.13055
12. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011; 12(7):499–510. https://doi.org/10.1038/nrg3012 PMID: 21681211
13. Zhang Y, Wang L, Xin H, Li D, Ma C, Ding X, et al. Construction of a high-density genetic map for sesame based on large scale marker development by specific length amplified fragment (SLAF) sequencing. BMC Plant Biol. 2013; 13:141. https://doi.org/10.1186/1471-2229-13-141 PMID: 24060091
14. Wu K, Liu H, Yang M, Tao Y, Ma H, Wu W, et al. High-density genetic map construction and QTLs analysis of grain yield-related traits in sesame (*Sesamum indicum* L.) based on RAD-Seq techonology. BMC Plant Biol. 2014; 14:274. https://doi.org/10.1186/s12870-014-0274-7 PMID: 25300176
15. Wang L, Xia Q, Zhang Y, Zhu X, Zhu X, Li D, et al. Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map. BMC Genomics. 2016; 17:31. https://doi.org/10.1186/s12864-015-2316-4 PMID: 26732604
16. Zhang H, Miao H, Li C, Wei L, Duan Y, Ma Q, et al. Ultra-dense SNP genetic map construction and identification of *SiDt* gene controlling the determinate growth habit in *Sesamum indicum* L. Sci Rep. 2016; 6:31556. https://doi.org/10.1038/srep31556 PMID: 27527492
17. Mei H, Liu Y, Du Z, Wu K, Cui C, Jiang X, et al. High-density genetic map construction and gene mapping of basal branching habit and flowers per leaf axil in sesame. Front Plant Sci. 2017; 8:636. https://doi.org/10.3389/fpls.2017.00636 PMID: 28496450
18. Wei X, Zhu X, Yu J, Wang L, Zhang Y, Li D, et al. Identification of sesame genomic variations from genome comparison of landrace and variety. Front Plant Sci. 2016; 7:1169. https://doi.org/10.3389/fpls.2016.01169 PMID: 27536315
19. Nordborg M and Weigel D. Next-generation genetics in plants. Nature. 2008; 456(7223):720–723. https://doi.org/10.1038/nature07629 PMID: 19079047
20. Guo B, Wang D, Guo Z, Beavis WD. Family-based association mapping in crop species. Theor Appl Genet. 2013; 126(6):1419–1430. https://doi.org/10.1007/s00122-013-2100-2 PMID: 23620001
21. Huang X and Han B. Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol. 2014; 65:531–551. https://doi.org/10.1146/annurev-arplant-050213-035715 PMID: 24274033

22. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 2010; 42(11):961–967. https://doi.org/10.1038/ng.695 PMID: 20972439

23. Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet. 2013; 45(1):43–50. https://doi.org/10.1038/ng.2484 PMID: 23242369

24. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015; 33 (4):408–414. https://doi.org/10.1038/nbt.3096 PMID: 25643055

25. Huang C, Nie X, Shen C, You C, Li W, Zhao W, et al. Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. Plant Biotechnol J. 2017; 15(11):1374–1386. https://doi.org/10.1111/pbi.12722 PMID: 28301713

26. Zhou Q, Han D, Mason AS, Zhou C, Zheng W, Li Y, et al. Earliness traits in rapeseed (*Brassica napus*): SNP loci and candidate genes identified by genome-wide association analysis. DNA Res. 2018; 25 (3):229–244. https://doi.org/10.1093/dnares/dsx052 PMID: 29236947

27. Wei X, Liu K, Zhang Y, Feng Q, Wang L, Zhao Y, et al. Genetic discovery for oil production and quality in sesame. Nat Commun. 2015; 6:8609. https://doi.org/10.1038/ncomms9609 PMID: 26477832

28. Dossa K, Li D, Zhou R, Yu J, Wang L, Zhang Y, et al. The genetic basis of drought tolerance in the high oil crop *Sesamum indicum*. Plant Biotechnol J. 2019; 17(9):1788–803. https://doi.org/10.1111/pbi.13100 PMID: 30801874

29. Cui C, Mei H, Liu Y, Zhang H, Zheng Y. Genetic diversity, population structure, and linkage disequilibrium of an association-mapping panel revealed by genome-wide SNP markers in sesame. Front Plant Sci. 2017; 8:1189. https://doi.org/10.3389/fpls.2017.01189 PMID: 28729877

30. Champa WAH, Gill MIS, Mahajan BVC, Aroraa NK. Postharvest treatment of polyamines maintains quality and extends shelf-life of table grapes (*Vitis vinifera* L.) cv. Flame Seedless. Postharvest Biol Tec. 2014; 91:57–63. https://doi.org/10.1016/j.postharvbio.2013.12.014

31. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020.

32. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw, 2015; 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

33. Kaler AS, Ray JD, Schapaugh WT, King CA, Purcell LC. Genome-wide association mapping of canopy wilting in diverse soybean genotypes. Theor Appl Genet. 2017; 130(10):2203–2217. https://doi.org/10.1007/s00122-017-2951-z PMID: 28730464

34. Meng L, Li H, Zhang L, Wang J. QTL Icimapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. Crop J. 2015; 3:269–283. https://doi.org/10.1016/j.cj.2015.01.001

35. Yano K, Morinaka Y, Wang F, Huang P, Takehara S, Hirai T, et al. GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. Proc Natl Acad Sci U S A. 2019; 116(42):21262–21267. https://doi.org/10.1073/pnas.1904964116 PMID: 31570620

36. Borcard D, Gillet F, Legendre P. Numerical ecology with R. 1st ed. New York: Springer New York; 2011.

37. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88(1):76–82. https://doi.org/10.1016/j.ajhg.2010.11.011 PMID: 21167468

38. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007; 23(19):2633–2635. https://doi.org/10.1093/bioinformatics/btm308 PMID: 17586829

39. Turner SD. qqman: An R package for visualizing GWAS results using Q−Q and Manhattan plots. Journal of Open Source Software. 2018; 3(25):1731. https://doi.org/10.21105/joss.00731

40. Yano K, Yamamoto E, Aya K, Takeuchi H, Lo PC, Hu L, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 2016; 48(8):927–934. https://doi.org/10.1038/ng.3596 PMID: 27322545

41. Shin JH, Blay S, McNeney B, Graham J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Softw. 2006; 16:Code Snippet 3. https://doi.org/10.18637/jss.v016.i11 PMID: 21451741

42. Pourcel L, Routaboul JM, Kerhoas L, Caboche M, Lepiniec L, Debeaujon I. *TRANSPARENT TESTA10* encodes a laccase-like enzyme involved in oxidative polymerization of flavonoids in *Arabidopsis* seed coat. Plant Cell. 2005; 17(11):2966–2980. https://doi.org/10.1105/tpc.105.035154 PMID: 16243908

43. Pourcel L, Routaboul JM, Cheynier V, Lepiniec L, Debeaujon I. Flavonoid oxidation in plants: from biochemical properties to physiological functions. Trends Plant Sci. 2007; 12(1):29−36. https://doi.org/10.1016/j.tplants.2006.11.006 PMID: 17161643

44. Yu CY. Molecular mechanism of manipulating seed coat coloration in oilseed *Brassica* species. J Appl Genet. 2013; 54(2):135–145. https://doi.org/10.1007/s13353-012-0132-y PMID: 23329015

45. Turlapati PV, Kim KW, Davin LB, Lewis NG. The laccase multigene family in *Arabidopsis thaliana*: towards addressing the mystery of their gene function(s). Planta. 2011; 233(3):439–470. https://doi.org/10.1007/s00425-010-1298-3 PMID: 21063888

46. Wang Q, Li G, Zheng K, Zhu X, Ma J, Wang D, et al. The soybean laccase gene family: evolution and possible roles in plant defense and stem strength selection. Genes. 2019; 10(9):701. https://doi.org/10.3390/genes10090701 PMID: 31514462

47. Liu L, Stein A, Wittkop B, Sarvari P, Li J, Yan X, et al. A knockout mutation in the lignin biosynthesis gene *CCR1* explains a major QTL for acid detergent lignin content in *Brassica napus* seeds. Theor Appl Genet. 2012; 124(8):1573–1586. https://doi.org/10.1007/s00122-012-1811-0 PMID: 22350089

48. de Vetten N, ter Horst J, van Schaik HP, de Boer A, Mol J, Koes R. A cytochrome b5 is required for full activity of flavonoid 3',5'-hydroxylase, a cytochrome P450 involved in the formation of blue flower colors. Proc Natl Acad Sci U S A. 1999; 96(2):778–783. https://doi.org/10.1073/pnas.96.2.778 PMID: 9892710

49. Matsubara K, Kodama H, Kokubun H, Watanabe H, Ando T. Two novel transposable elements in a cytochrome P450 gene govern anthocyanin biosynthesis of commercial petunias. Gene. 2005; 358:121–126. https://doi.org/10.1016/j.gene.2005.05.031 PMID: 16051450

50. Guo Y and Qiu LJ. Allele-specific marker development and selection efficiencies for both flavonoid 3'-hydroxylase and flavonoid 3',5'-hydroxylase genes in soybean subgenus soja. Theor Appl Genet. 2013; 126(6):1445–1455. https://doi.org/10.1007/s00122-013-2063-3 PMID: 23463490

51. Tanaka Y and Brugliera F. Flower colour and cytochromes P450. Philos Trans R Soc Lond B Biol Sci. 2013; 368(1612):20120432. https://doi.org/10.1098/rstb.2012.0432 PMID: 23297355

52. Li D, Liu P, Yu J, Wang L, Dossa K, Zhang Y, et al. Genome-wide analysis of WRKY gene family in the sesame genome and identification of the WRKY genes involved in responses to abiotic stresses. BMC Plant Biol. 2017; 17(1):152. https://doi.org/10.1186/s12870-017-1099-y PMID: 28893196

53. Chen F, Hu Y, Vannozzi A, Wu K, Cai H, Qin Y, et al. The WRKY transcription factor family in model plants and crops. Crit Rev Plant Sci. 2018; 36:311–335. https://doi.org/10.1080/07352689.2018.1441103

54. Dong Y, Wang C, Han X, Tang S, Liu S, Xia X, et al. A novel bHLH transcription factor PebHLH35 from Populus euphratica confers drought tolerance through regulating stomatal development, photosynthesis and growth in *Arabidopsis*. Biochem Biophys Res Commun. 2014; 450(1):453–458. https://doi.org/10.1016/j.bbrc.2014.05.139 PMID: 24909687

55. Li P, Chen B, Zhang G, Chen L, Dong Q, Wen J, et al. Regulation of anthocyanin and proanthocyanidin biosynthesis by *Medicago truncatula* bHLH transcription factor MtTT8. New Phytol. 2016; 210(3):905–921. https://doi.org/10.1111/nph.13816 PMID: 26725247

56. Makkena S and Lamb RS. The bHLH transcription factor SPATULA is a key regulator of organ size in *Arabidopsis thaliana*. Plant Signal Behav. 2013; 8(5):e24140. https://doi.org/10.4161/psb.24140 PMID: 23470719

57. Nesi N, Debeaujon I, Jond C, Pelletier G, Caboche M, Lepiniec L. The *TT8* gene encodes a basic helix-loop-helix domain protein required for expression of *DFR* and *BAN* genes in *Arabidopsis* siliques. Plant Cell. 2000; 12(10):1863–1878. https://doi.org/10.1105/tpc.12.10.1863 PMID: 11041882

58. Johnson CS, Kolevski B, Smyth DR. *TRANSPARENT TESTA GLABRA2*, a trichome and seed coat development gene of *Arabidopsis*, encodes a WRKY transcription factor. Plant Cell. 2002; 14(6):1359–1375. https://doi.org/10.1105/tpc.001404 PMID: 12084832

59. Xu W, Dubos C, Lepiniec L. Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. Trends Plant Sci. 2015; 20(3):176–185. https://doi.org/10.1016/j.tplants.2014.12.001 PMID: 25577424

60. Gonzalez A, Brown M, Hatlestad G, Akhavan N, Smith T, Hembd A, et al. TTG2 controls the developmental regulation of seed coat tannins in Arabidopsis by regulating vacuolar transport steps in the proanthocyanidin pathway. Dev Biol. 2016; 419(1):54–63. https://doi.org/10.1016/j.ydbio.2016.03.031 PMID: 27046632