Planta (2020) 251:39
https://doi.org/10.1007/s00425-019-03329-6

ORIGINAL ARTICLE

Genetic dissection of yield-related traits via genome-wide association analysis across multiple environments in wild soybean (*Glycine soja* Sieb. and Zucc.)

Dezhou Hu[1] · Huairen Zhang[1,2] · Qing Du[1] · Zhenbin Hu[3] · Zhongyi Yang[1] · Xiao Li[1] · Jiao Wang[1] · Fang Huang[1] · Deyue Yu[1,4] · Hui Wang[1] · Guizhen Kan[1]

Received: 19 September 2019 / Accepted: 10 December 2019

Abstract

Main conclusion **A total of 41 SNPs were identified as significantly associated with five yield-related traits in wild soybean populations across multiple environments, and the candidate gene *GsCID1* was found to be associated with seed weight. These results may facilitate improvements in cultivated soybean.**

Abstract Crop-related wild species contain new sources of genetic diversity for crop improvement. Wild soybean (*Glycine soja* Sieb. and Zucc.) is the progenitor of cultivated soybean [*Glycine max* (L.) Merr.] and can be used as an essential genetic resource for yield improvements. In this research, using genome-wide association study (GWAS) in 96 out of 113 wild soybean accessions with 114,090 single nucleotide polymorphisms (SNPs) (with minor allele frequencies ≤ 0.05), SNPs associated with five yield-related traits were identified across multiple environments. In total, 41 SNPs were significantly associated with the traits in two or more environments (significance threshold $P \leq 8.76 \times 10^{-6}$), with 29, 7, 3, and 2 SNPs detected for 100-seed weight (SW), maturity time (MT), seed yield per plant (SY) and flowering time (FT), respectively. BLAST search against the *Glycine soja* W05 reference genome was performed, 20 candidate genes were identified based on these 41 significant SNPs. One candidate gene, *GsCID1* (*Glysoja.04g010563*), harbored two significant SNPs—AX-93713187, with a non-synonymous mutation, and AX-93713188, with a synonymous mutation. *GsCID1* was highly expressed during seed development based on public information resources. The polymorphisms in this gene were associated with SW. We developed a derived cleaved amplified polymorphic sequence (dCAPS) marker for *GsCID1* that was highly associated with SW and was validated as a functional marker. In summary, the revealed SNPs/genes are useful for understanding the genetic architecture of yield-related traits in wild soybean, which could be used as a potential exotic resource to improve cultivated soybean yields.

Keywords Crop wild relatives · *GsCID1* · GWAS · Marker-assisted selection · Wild soybean

Abbreviations

DAF	Days after flowering
dCAPS	Derived cleaved amplified polymorphic sequence
FT	Flowering time
GWAS	Genome-wide association study

Dezhou Hu and Huairen Zhang contributed equally to this work

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00425-019-03329-6) contains supplementary material, which is available to authorized users.

✉ Guizhen Kan
kanguizhen@njau.edu.cn

[1] National Center for Soybean Improvement, National Key Laboratory of Crop Genetics and Germplasm Enhancement, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing 210095, China

[2] Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

[3] Department of Agronomy, Kansas State University, Manhattan, KS, USA

[4] School of Life Sciences, Guangzhou University, Guangzhou 510006, China

Published online: 06 January 2020

h^2	Broad-sense heritability
LD	Linkage disequilibrium
MT	Maturity time
PN	Number of effective pods
SNP	Single nucleotide polymorphism
SW	100-Seed weight
SY	Seed yield per plant

Introduction

Crop wild relatives are a rich source of genetic diversity and are potentially useful in plant breeding for the development of varieties with novel traits (Treuren et al. 2017). Since the beginning of the twentieth century, crop wild relatives have been increasingly used for plant breeding and have provided important genetic diversity for crop improvement in maize (Yang et al. 2013; Huang et al. 2017), rice (Tian et al. 2006; Okishio et al. 2014) and wheat (Nevo and Chen 2010; Placido et al. 2013). Cultivated soybean is an important food and economic crop worldwide. The development of high-yielding plants is one of the major breeding objectives in soybean; however, high yield is often affected by many factors, especially environmental factors (Chapman et al. 2003). Yield-related traits in soybean are typical complex quantitative traits and are controlled by multiple QTLs/genes. Wild soybean, which has higher allelic diversity than domesticated soybean and can be used as an important resource to improve domesticated soybeans, is the immediate progenitor of the domesticated soybean (Qi et al. 2014).

In recombinant inbred lines, QTL mapping was a useful method to identify the regions or genes associated with a given trait. QTL identification using linkage populations derived from wild and cultivated soybean has been widely used to determine drivers of biotic/abiotic stress responses or seed protein/oil contents. For example, genetic analysis and QTL mapping of soybean resistance to southern root-knot nematode were conducted using recombinant inbred lines derived from a cross between PI 438489B (wild soybean) and Magellan (cultivated soybean) (Xu et al. 2013). A salt tolerance gene *GmCHX1* was identified in a recombinant inbred line population of soybean, which was constructed by crossing the wild soybean accession W05 with the cultivated soybean accession C08 (Qi et al. 2014). A seed protein QTL was fine mapped using the backcross population derived from a cross between PI 468916 (wild soybean) and A81–356022 (cultivated soybean) (Nichols et al. 2006). However, QTL mapping suffers from fundamental limitations; only allelic diversity that segregates between the biparental lines can be assayed. The mapping resolution is also limited due to the amount of recombination during the creation of the recombinant inbred line population (Korte and Farlow 2013).

As next-generation sequencing has been developed, GWAS has been utilized for multiple important traits in plants, such as maize (Riedelsheimer et al. 2012; Zhang et al. 2016a), cotton (Fang et al. 2017; Ma et al. 2018) and rice (Wang et al. 2015; Shi et al. 2017). Compared to linkage mapping, GWAS can not only be used in natural populations and germplasm collections but also allows for higher resolution in determining the genomic locations of QTLs/genes (Shirasawa et al. 2013). To date, GWAS has been widely used to identify several traits in cultivated soybean. For example, 60 SNPs for soybean cyst nematode resistance were identified based on over 45,000 SNPs by GWAS (Vuong et al. 2015). The gene *GmCDF1*, which is associated with salt tolerance, was detected in the germination stage of soybean via GWAS (Zhang et al. 2019). Although GWAS has been well applied in cultivated soybean, few studies using wild soybean populations for GWAS have been reported, especially those investigating yield-related traits. GWAS using wild soybean populations has mainly been focused on seed composition or biotic stress. For example, 29 significant SNPs for seven traits located on 10 different chromosomes were discovered using a wild soybean population (Leamy et al. 2017). Using 1032 wild soybean accessions, 10 SNPs that were significantly associated with the response to soybean cyst nematode race 1 were identified (Zhang et al. 2017). Wild soybean contains favorable alleles for yield, and it is feasible to identify these alleles in wild soybean (Hu et al. 2014).

In our previous study, a preliminary association mapping of yield-related loci/regions with limited markers (85 SSR makers) was performed (Hu et al. 2014). In the present research, a large genome-wide SNP array (NJAU 355K SoySNP) was used for five yield-related traits to dissect genetic information across multiple environments (i) to detect yield-related loci in the wild soybean population and (ii) to identify candidate genes helpful for the improvement of cultivated soybean yield.

Materials and methods

Materials

Phenotypic data collection was performed using 113 wild soybean accessions of different geographical origins provided by the National Center for Soybean Improvement, Nanjing, China. These accessions were planted in two field experiments: at the Jiangpu Experimental Station, Nanjing Agricultural University (32°12 N 118°37 48 E), Nanjing, China, in 2011 (designated E1); Jiangpu Experimental Station and the Nanyang Experimental Station at Henan Agricultural University (38°7′ N 110°34′ E), Nanyang, China, in

2012 (designated E2 and E3); and at the Jiangpu Experimental Station in 2013 (designated E4).

Field trials and trait evaluation

The crop density of wild soybean is different from the cultivated soybeans, because wild soybeans have creeping habits, thin stems and plant heights as high as 3–6 m. Thus, this study utilized a completely randomized block design with three replications. For every accession, each replication consisted of a four-square meter plot containing four hills. Twenty seeds of a given accession were sown in each hill. The number of plants per hill was thinned to five about 2 weeks after germination. Irrigation, fertilization, insect and weed control were performed throughout the experiment.

Four traits were evaluated in the four environments (E1–E4), including the number of effective pods (PN), SW (g), FT (day) and SY (g). MT (day) was evaluated in three environments (E1, E2, E4). To avoid the impact of pod dehiscence and maturity differences on yield, we made daily observations to determine whether wild soybean accessions were mature in the late stage of soybean growth. The mature accessions were harvested in time. The pods of the accessions were removed by hand. The FT was calculated as the number of days from the germination to the beginning of blooming (R1, 50% of the plants in a plot had an open flower at one of the top nodes with a fully expanded leaf) (Fehr and Caviness 1977). The MT was considered the number of days from germination to the time at which 95% of the pods on a plant turned brown (Fehr and Caviness 1977). After harvesting, SY, PN and SW were evaluated. SY and SW were adjusted to 13% moisture content. For every replication, the SY value = (total seed yield of five plants)/5 and the PN value = (total pod numbers of five plants)/5. The SW value was the average value of three replications. The germplasm number and phenotypic data of the wild soybean populations are listed in Table S1.

Phenotypic data analysis

The phenotypic data from wild soybean accessions were calculated using the average values of three replications (E1-E4 for PN, SW, FT and SY; and E1, E2, E4 for MT) with SAS 9.0 software (SAS Institute 1999), including descriptive statistics, ANOVA, broad-sense heritability (h^2) and correlation analysis. Best linear unbiased prediction (BLUP) was used in linear mixed models for the estimation of random effects. In this study, to combine the data collected from different years and minimize the effects of environmental variation, BLUP values were calculated for each trait based on the combined data collected from different environments (Dhanapal et al. 2015). To derive across-environment BLUP values, accession and other factors were considered random effects (Edae et al. 2014). BLUP was estimated by fitting the following formula in the R software package "lme4" (Merk et al. 2012): $Y_{ik} = \mu + G_i + E_j + GE_{ij} + \varepsilon_{ij}$, where Y_{ik} is the trait, μ is the overall mean, G_i is the ith genotypic effect, E_j is the jth environmental effect, GE_{ij} is the effect of the genotype × environment interaction, and ε_{ij} is the residual error. The h^2 of yield-related traits was based on the formula: $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_{ge}^2/n + \sigma_e^2/nr)$. The details of h^2 formula were previously described (Hu et al. 2019). Correlation analysis among the five traits was performed using SPSS 20.0 software based on Pearson's correlations (SPSS Statistics 20).

GWAS analysis

A set of 96 accessions from 113 wild soybeans was genotyped with 282,469 SNPs (NJAU 355K SoySNP array) (Wang et al. 2016). A total of 114,090 SNPs remained after filtering out SNPs with minor allele frequencies ≤ 0.05. GWAS was performed using the GLM (PCA), GLM (Q), MLM (PCA + K) and MLM (Q + K) in TASSEL software V5.0 (Bradbury et al. 2007). The kinship matrix (K) and population structure (PCA = 5) were assessed by TASSEL software V5.0. The population structure (Q) was calculated using STRUCTURE software version 2.3.4 (Falush et al. 2003). In this study, SNPs were identified as being significantly associated with yield-related traits ($P \le 1/114{,}090 = 8.76 \times 10^{-6}$ or $-\log_{10}(P) \ge 5.05$) using the adjusted Bonferroni method ($P \le 1/n$, n is the total number of genome-wide SNPs) (Zhang et al. 2016a, 2019). To obtain the position of significant SNPs, a BLAST search was performed on the current *Glycine soja* W05 reference genome (https://www.wildsoydb.org/Gsoja_W05/) using 16-bp sequences flanking the significant SNPs (Xie et al. 2019). Manhattan plots and quantile–quantile plots were generated using the R software package "qqman" (Turner 2014).

Putative gene prediction and expression pattern analysis

A gene model was considered to be a putative gene for yield-related traits if the gene model satisfied the following conditions: (i) the SNP was significantly associated with yield-related traits in two or more environments, (ii) the significant SNP was detected in the genomic region of the gene model, and (iii) the gene model was predicted to be related to yield or yield-related traits based on the wild soybean genome annotation. Using the amino acid sequences of candidate genes, we employed BLAST-P analysis against cultivated soybean and *Arabidopsis* proteins.

The expression patterns of putative genes were identified based on two databases obtained from cultivated soybean.

One database included RNA-seq data downloaded from SoyBase (https://www.soybase.org/soyseq/), containing the data from three vegetative tissues and 11 stages of reproductive tissue development; the other database included expression data from different seed growth stages and was downloaded from the Plant Expression Database (https://www.plexdb.org, GEO Accession number: GSE42871). The robust multiarray analysis method is the most widely used preprocessing algorithm for Affymetrix and gene expression microarrays. The robust multiarray analysis-normalized values from the microarray data were $\log_2$-transformed. Two heat maps based on gene expression data were drawn with TBtools software (https://cj-chen.github.io/tbtools/).

GsCID1-based association analysis and dCAPS marker development

Six SNPs located in the genomic region of *GsCID1* were used to identify its haplotypes. The linkage disequilibrium (LD) level was evaluated by the R package 'LDheatmap' (Shin et al. 2006). Six SNPs in *GsCID1* were used to perform the association mapping with the mean SW values. The SNP AX-94009457 (T/C) of *GsCID* was used to develop the dCAPS marker. The materials with the top 20%, middle 60% and lowest 20% of the SW values were defined as high-SW materials, moderate-SW materials and low-SW materials, respectively. The primers were designed by dCAPS Finder 2.0 (https://helix.wustl.edu/dcaps/dcaps.html) (forward: 5′-AGTGTTCACATGTATGTATGGTAC-3′, reverse: 5′-ATTAATTTTGATATTGATCAATGG -3′). Ten high-SW varieties, three moderate-SW varieties and five low-SW varieties were selected to identify the validity of the dCAPS marker (Table S2). The PCR products were digested with the restriction endonuclease *KpnI* (37 °C, 2 h) and assayed by 8% polyacrylamide gel electrophoresis to distinguish the sizes of the polymorphic fragments.

Results

Phenotypic variation of yield-related traits in wild soybean

ANOVA, h^2 and descriptive statistics of five traits were identified for 113 wild soybean accessions in three or four environments (Table 1). The mean values of SY, SW, PN, FT and MT were 14.12, 2.35, 341.65, 59.23 and 106.24, respectively. These five traits exhibited wide variation among different environments, and the mean coefficients of variation for the five traits were 54.24%, 60.02%, 46.94%, 23.09% and 11.71%, respectively. The frequency distribution of the wild soybean population showed a normal distribution (Fig. 1), indicating that these five traits were complex quantitative traits controlled by multiple QTLs/genes. ANOVA results revealed that these five traits were significantly affected by the genotype and environment effects ($P < 0.01$).

The h^2 value of the PN was 50.27%, which was lower than the previously reported results in cultivated soybean (Hao et al. 2012), suggesting that the PN in wild soybean was more sensitive to the environment than that in cultivated soybean. Similar to the cultivated soybean, the h^2 values of SY, SW, FT and MT in wild soybean were also high (72.69%, 90.62%, 88.58% and 88.40%, respectively), indicating that these four traits were relatively stable and not particularly affected by environmental factors.

Correlation analysis among five yield-related traits

Pearson correlation coefficients between five traits were analyzed for the 113 wild accessions (Fig. 2). The results showed that MT was positively correlated with SW ($r = 0.15$, not significant) and strongly positively correlated with SY ($r = 0.27$, $P < 0.05$), PN ($r = 0.29$, $P < 0.05$) and FT ($r = 0.91$, $P < 0.001$). SY was positively correlated with PN ($r = 0.2$, not significant) and FT ($r = 0.13$, not significant) and significantly positively correlated with SW ($r = 0.55$, $P < 0.001$). However, SW was negatively correlated with FT ($r = -0.15$, not significant) and PN ($r = -0.42$, $P < 0.001$, significant). Additionally, PN was significantly positively correlated with FT ($r = 0.44$, $P < 0.001$).

GWAS for SY, SW, PN, FT and MT

Significant SNPs associated with yield-related traits were identified using the NJAU 355K SoySNP array with Tassel 5.0 software. To further determine which model was suitable for these traits, four different models, including GLM (PCA), MLM (PCA + K), GLM (Q) and MLM (Q + K), were compared. As shown in Fig. 3, the quantile–quantile plot of GLM (PCA) was much closer to the diagonal line than that of GLM (Q), and the quantile–quantile plot of MLM (PCA + K) was much closer to the diagonal line than that of MLM (Q + K), suggesting that PCA as a covariate in GWAS was better than Q for this wild soybean population. The kinship matrix was used to represent the covariance structure of random polygenic effects. Based on the quantile–quantile plot distributions of different models, the MLM (PCA + K) was identified as more suitable for SY, SW, FT and MT, whereas the GLM (PCA) was more appropriate for PN (Fig. 3). Using these two models, a total of 218 SNPs significantly (with a significance threshold of $-\log_{10}(P) \geq 5.05$) associated with yield-related traits are listed (Fig. 4; Table S3). Additionally, 41 SNPs significantly associated with yield-related traits in two or more environments are summarized and detailed in Table 2.

Table 1 Descriptive statistics, ANOVA and h^2 of the yield-related traits in wild soybean population

Trait	Environment	Mean	SD[a]	Median	Minimum	Maximum	CV[b] (%)	Skewness	Kurtosis	G[c]	G×E[d]	h^2 (%)
SY	E1	15.90	7.56	15.63	3.62	44.65	47.52	1.02	2.44	**	**	72.69
	E2	16.40	9.50	14.68	3.19	54.35	57.93	1.30	2.13			
	E3	17.20	8.24	16.98	1.66	43.38	47.92	1.05	1.45			
	E4	6.97	4.43	6.09	0.37	21.63	63.58	0.85	0.51			
	Mean	14.12	7.43	13.35	2.21	41.00	54.24	1.06	1.63			
SW	E1	2.51	1.43	2.23	0.90	10.41	57.00	2.51	9.04	**	**	90.62
	E2	2.39	1.44	2.01	0.90	9.74	60.24	2.28	6.71			
	E3	2.15	1.45	1.76	0.66	8.48	67.37	2.47	6.97			
	E4	2.35	1.30	2.05	0.95	8.89	55.48	2.45	7.30			
	Mean	2.35	1.41	2.01	0.85	9.38	60.02	2.43	7.51			
PN	E1	268.43	111.42	261.55	93.40	611.23	41.51	0.60	0.00	**	**	50.27
	E2	220.11	81.14	211.33	50.33	419.58	36.86	0.42	− 0.37			
	E3	726.33	355.61	706.38	124.13	1621.50	48.96	0.29	− 0.56			
	E4	151.72	91.65	134.97	3.92	435.78	60.41	0.80	0.05			
	Mean	341.65	159.96	328.56	67.95	772.02	46.94	0.53	− 0.22			
FT	E1	55.61	13.02	54.67	27.00	84.00	23.42	− 0.08	− 1.00	**	**	88.58
	E2	57.28	12.70	59.10	26.00	84.50	22.16	− 0.18	− 0.75			
	E3	60.52	13.04	59.33	35.50	92.00	21.54	0.05	− 0.92			
	E4	63.49	16.02	65.00	31.00	92.00	25.23	− 0.19	− 1.17			
	Mean	59.23	13.70	59.53	29.88	88.13	23.09	− 0.10	− 0.96			
MT	E1	105.70	12.76	106.95	81.28	138.33	12.08	− 0.04	− 0.42	**	**	88.40
	E2	102.61	11.39	100.40	78.50	131.00	11.10	0.42	− 0.13			
	E4	110.42	13.18	109.33	79.67	145.00	11.94	0.13	0.19			
	Mean	106.24	12.44	105.56	79.82	138.11	11.71	0.17	− 0.12			

**Significant at $P < 0.01$

[a]Standard deviation

[b]Coefficient of variation

[c]Genotype

[d]Genotype × environment

Of these 41 SNPs, nine significant SNPs for MT and FT were distributed on chromosome 1, 11, 14 and 19, with R^2 values ranging from 9.62% to 13.04%. Three SNPs associated with SY were located on chromosome 20, and these three SNPs were close to each other and located within the previously identified QTLs *Seed yield 21-7*, *Seed weight 36-5*, *Seed weight 24-3* and *Seed weight 50-13* (Reinprecht et al. 2006; Han et al. 2012; Kato et al. 2014). Remarkably, the R^2 values of these three adjacent SNPs ranged from 20.96 to 23.96%, suggesting that a major QTL exists in this region. A total of 29 SNPs significantly associated with SW were uncovered. These significant SNPs were located on chromosome 1, 2, 3, 4, 5, 6, 9, 11, 12, 13, 14, 15, 16 and 19 and comprised two, three, one, nine, one, one, one, two, two, one, one, one, three and one SNP, respectively. The 29 SNPs explained 6.78–13.83% of the total phenotypic variation. Most of these 29 significant SNPs were mapped within the previously reported QTLs. For example, two significant SNPs, AX-93961487 and AX-93616455 on chromosome 1, were located within two previously mapped QTLs or SNPs, *Seed weight 18-1.2* or *cqSeed weight-010* (Panthee et al. 2005; Pathan et al. 2013). AX-93687543 and AX-93985010 on chromosome 2 were located within a known QTL, *Seed weight 50-12* (Kato et al. 2014). Six SNPs on chromosome 4 were located within the region of *Seed weight 5-g4* (Zhang et al. 2016b).

Candidate genes associated with the yield component traits

GWAS based on high-density SNP markers can be used to finely map quantitative trait regions, even to the genes themselves (Chu et al. 2017). For example, a soybean isoflavone content-related gene, *GmMYB29,* was identified based on a significant SNP located in the 5′-UTR (Chu et al. 2017). Therefore, to further understand the molecular mechanism of the yield-related traits in wild soybean, we evaluated the genes within or close to the significant SNPs.

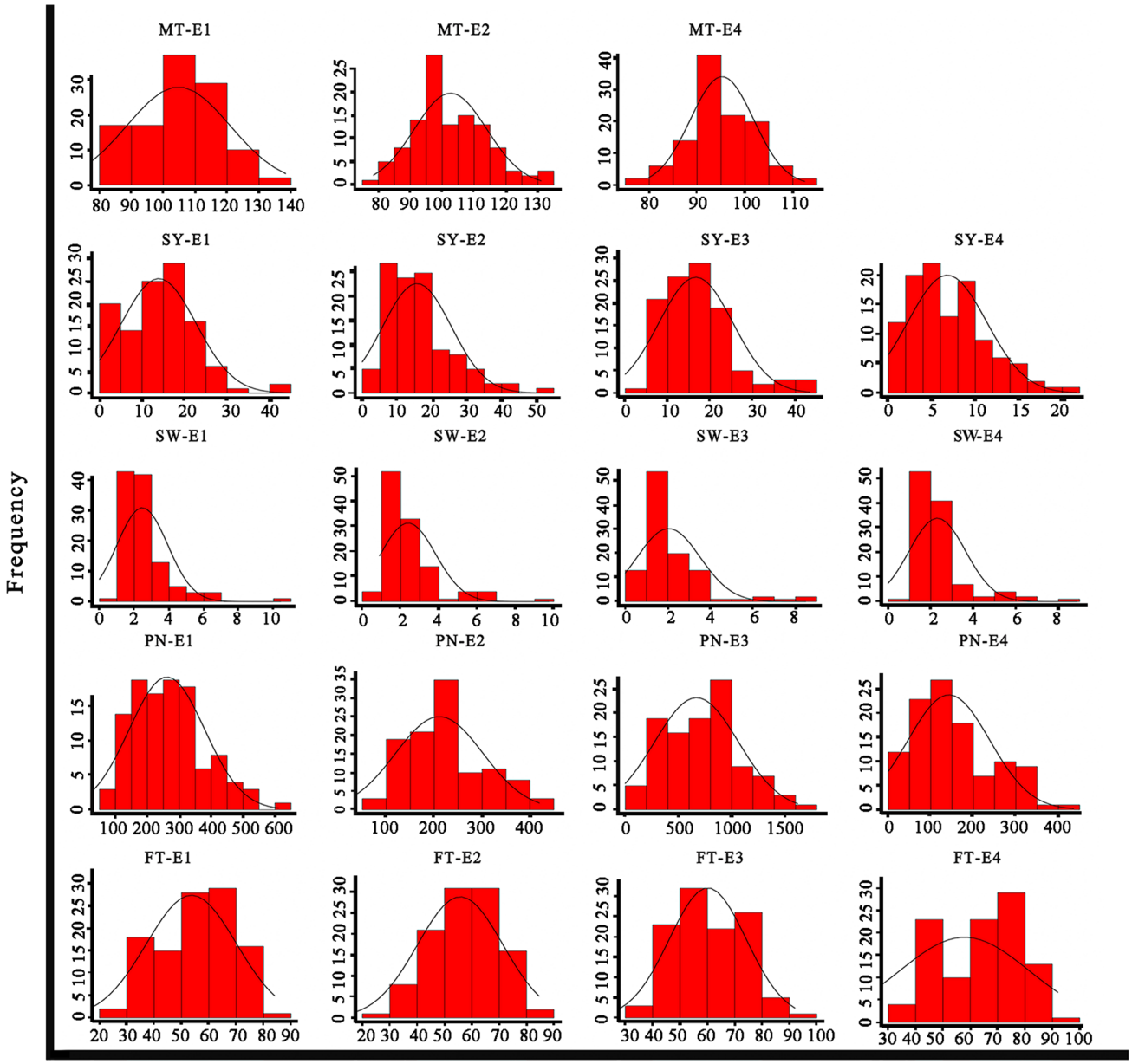


Fig. 1 Frequency distribution of five yield traits in the wild soybean population across multiple environments

As mentioned above, 41 significant SNPs were associated with the traits in two or more environments. Of these, 11 SNPs were located in the intergenic regions (AX-93959615 was coidentified with FT and MT); eight SNPs were located in the 3′-UTR of the genes; three SNPs were located in the promoter of the genes; one SNP was located in the 5′-UTR of the genes; ten SNPs were located in the intron of the genes; and eight SNPs were located in conservative coding regions, including seven non-synonymous mutations altering the amino acid sequence and one synonymous mutation (Table 3). As a result, there were 30 significant SNPs located in the genomic region of 24 gene models. The functional annotation of the 24 genes is listed in Table 3. Four genes were predicted without biological functional annotation (Table 3). For the other 20 genes, *Glysoja.04g010563* and *Glysoja.15g042025* were predicted to be associated with cell division and gibberellin biosynthesis, respectively. The remaining genes were predicted to be involved in protein transport, metabolic processes, biotic stress response and other functions (Table S4). Among these two genes, only *Glysoja.04g010563* possessed one SNP with a non-synonymous mutation and was located in the yield-related QTL *Seed weight 5-g4* (Zhang et al. 2016b). *Glysoja.04g010563* encodes a polyadenylate-binding protein-interacting protein

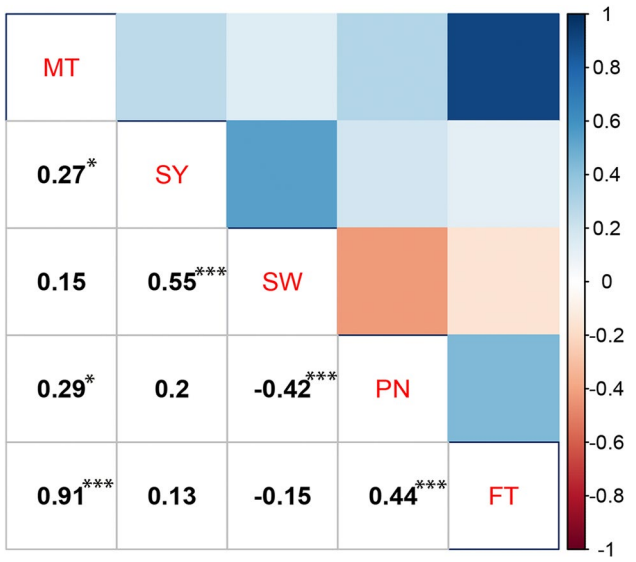


Fig. 2 Phenotypic correlations among five yield traits based on mean values of traits in all wild soybean lines. *Significant at $0.01 < P < 0.05$; ***significant at $P < 0.001$

with a conserved RNA recognition motif (RRM) domain. In rice, overexpression of the RRM domain of *OsFCA* increased cell size and yield (Hong et al. 2007).

Expression patterns of corresponding genes in cultivated soybean revealed the functional role of *Glysoja.04g010563* in seed development

We then compared the similarity of genomic and amino acid sequences between these 20 genes (with functional annotation) and the corresponding genes in cultivated soybean (Table S5). Five and nine of the 20 genes had a similarity of 100% in their genomic and amino acid sequences, respectively. In the remaining genes, the similarity of amino acid sequences ranged from 95.10 to 99.96%, and the similarity of genomic sequences ranged from 96.80 to 99.90%. We also listed the functional annotation and name of the corresponding genes (Table S5). These results indicated that the potential functions of these 20 genes can be partially predicted based on the expression patterns of the corresponding genes in cultivated soybean.

To further elucidate the potential functions of the 20 genes with functional annotation, the expression patterns of the corresponding genes in cultivated soybean were detected in public RNA-seq data. As shown in Fig. 5a and Table S6, the five genes, including *Glyma.15g252100* (*Glysoja.15g042025*), were weakly expressed in all tissues, and the other 15 genes were highly expressed in several tissues. For example, *Glyma.01g015000* (*Glysoja.01g000148*) was expressed in leaves, flowers, and seeds at 10 days after flowering (DAF) and roots. The expression level of *Glyma.03g125400* (*Glysoja.03g006986*) was high in seeds at 10 and 14 DAF. *Glyma.16g057500* (*Glysoja.16g042837*) was barely expressed in roots, young leaves, flowers, pods and seeds, while it was mainly expressed in nodules. Notably, *Glyma.04g229500* (*Glysoja.04g010563*) was expressed in nearly all tissues and was highly expressed in seeds at 35 DAF.

Considering that seed development is essential for seed yield in soybean, the expression patterns of the 20 genes at different seed development periods were also examined in previously reported RNA-seq data (Fig. 5b; Table S7). The expression profiles revealed that ten genes, including *Glyma.15g252100* (*Glysoja.15g042025*), showed low expression levels during seed development. Four of the remaining 10 genes, *Glyma.01g015000* (*Glysoja.01g000148*), *Glyma.03g125400* (*Glysoja.03g006986*), *Glyma.09g210000* (*Glysoja.09g024829*) and *Glyma.12g222000* (*Glysoja.12g035175*), were strongly expressed in the early seed development period and weakly expressed in dry seeds, whereas *Glyma.04g229500* (*Glysoja.04g010563*) was continuously expressed at high levels. Since *Glyma.04g229500* (*Glysoja.04g010563*) exhibited high expression in all tissues, especially during seed development, and it was detected based on two SNPs that were significantly associated with SW in two environments, *Glysoja.04g010563* was selected as a candidate gene of SW in wild soybean. The homolog of Glysoja.04g010563 in *Arabidopsis* is AtCID11 (At1g32790). Thus, *Glysoja.04g010563* was named *GsCID1*.

Polymorphisms of the *GsCID1* gene are associated with SW in wild soybean

A *GsCID1*-based association analysis was performed to analyze the correlation between the allelic variation of *GsCID1* and SW. Six SNPs, AX-93623918, AX-94009453, and AX-93713186 located in the promoter, AX-93713187 located in the exon, AX-93713188 located in the intron and AX-94009457 located in the 3′-UTR of *GsCID1*, were used for haplotype analysis (Table S8). The association study showed that these six SNPs exhibited strong LD (Fig. 6b). Four of these six SNPs were significantly associated with variation in SW (Fig. 6a). Based on these six variants, the genotypes of this wild soybean population were classified into six haplotypes (Hap1–Hap6) (Fig. 6d): Hap1 (ATCAGT), Hap2 (ATTAGT), Hap3 (GTCAGT), Hap4 (ACTCAC), Hap5 (GTTAGT) and Hap6 (ATCAGC) which contained 17, 24, 12, 5, 3 and 3 accessions, respectively. The phenotypic data of six haplotypes were further analyzed. As shown in Fig. 6c, the SW values of Hap4 and Hap1 were the highest and lowest, respectively. Among the six haplotypes, Hap4 exhibited significantly higher SW values than the other five haplotypes. While the most

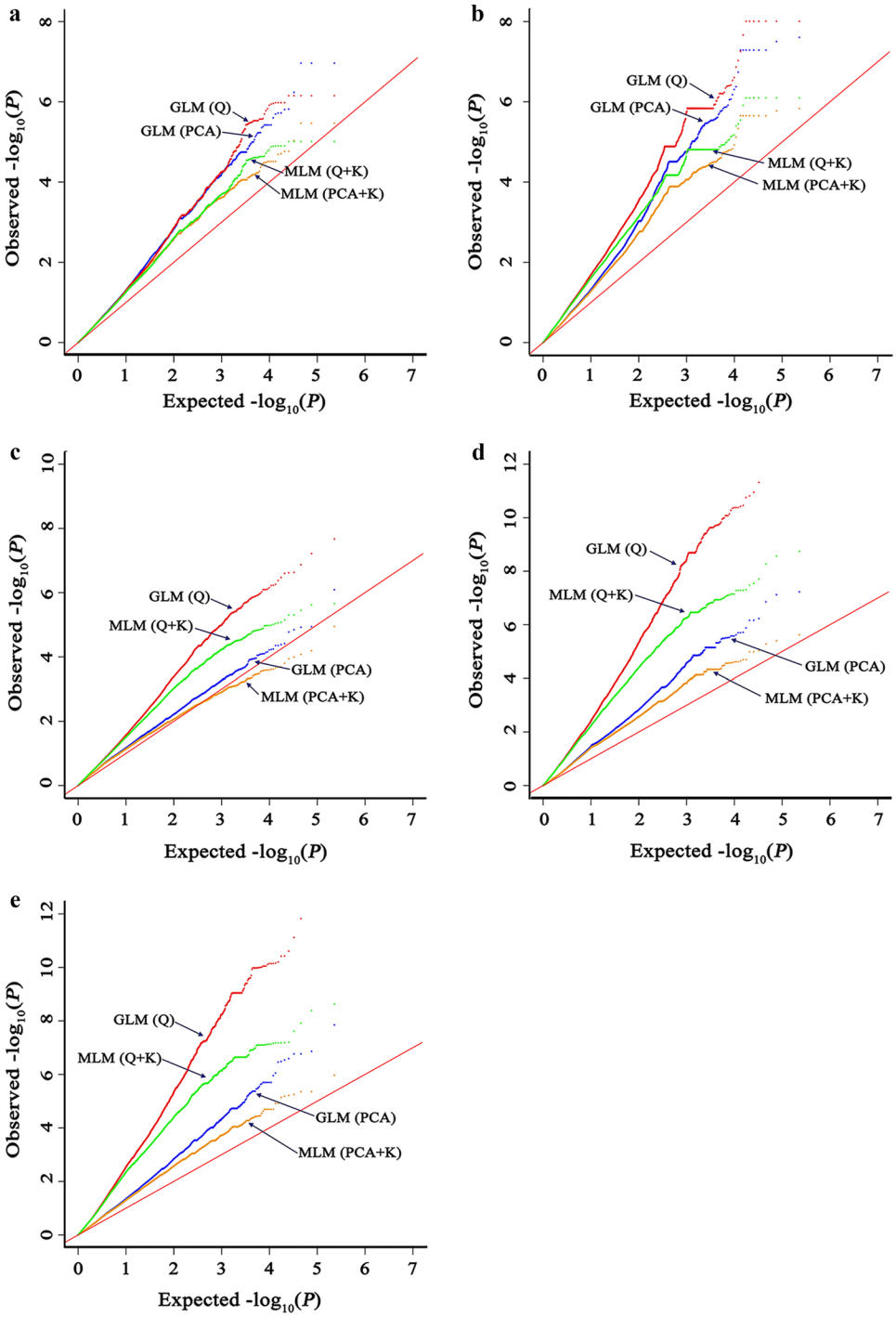


Fig. 3 Quantile–quantile plots of four GWAS models for five yield traits using the mean values of the traits. **a** SY. **b** SW. **c** PN. **d** FT. **e** MT

significant difference was seen between Hap4 and Hap1 ($P = 4.94 \times 10^{-3}$). Consequently, Hap4 should be the most favorable haplotype for high SW, while Hap1 was the most unfavorable haplotype for high SW. These data suggested that *GsCID1* polymorphisms are associated with SW and that the elite haplotype in wild soybean might be helpful in improving seed yield in cultivated soybean.

Development of a functional marker for wild soybean SW

In this study, six SNPs were located in the genomic region of *GsCID1*. AX-94009453-C, AX-93713187-C, AX-93713188-A and AX-94009457-C exhibited significantly higher SW than AX-94009453-T, AX-93713187-A,

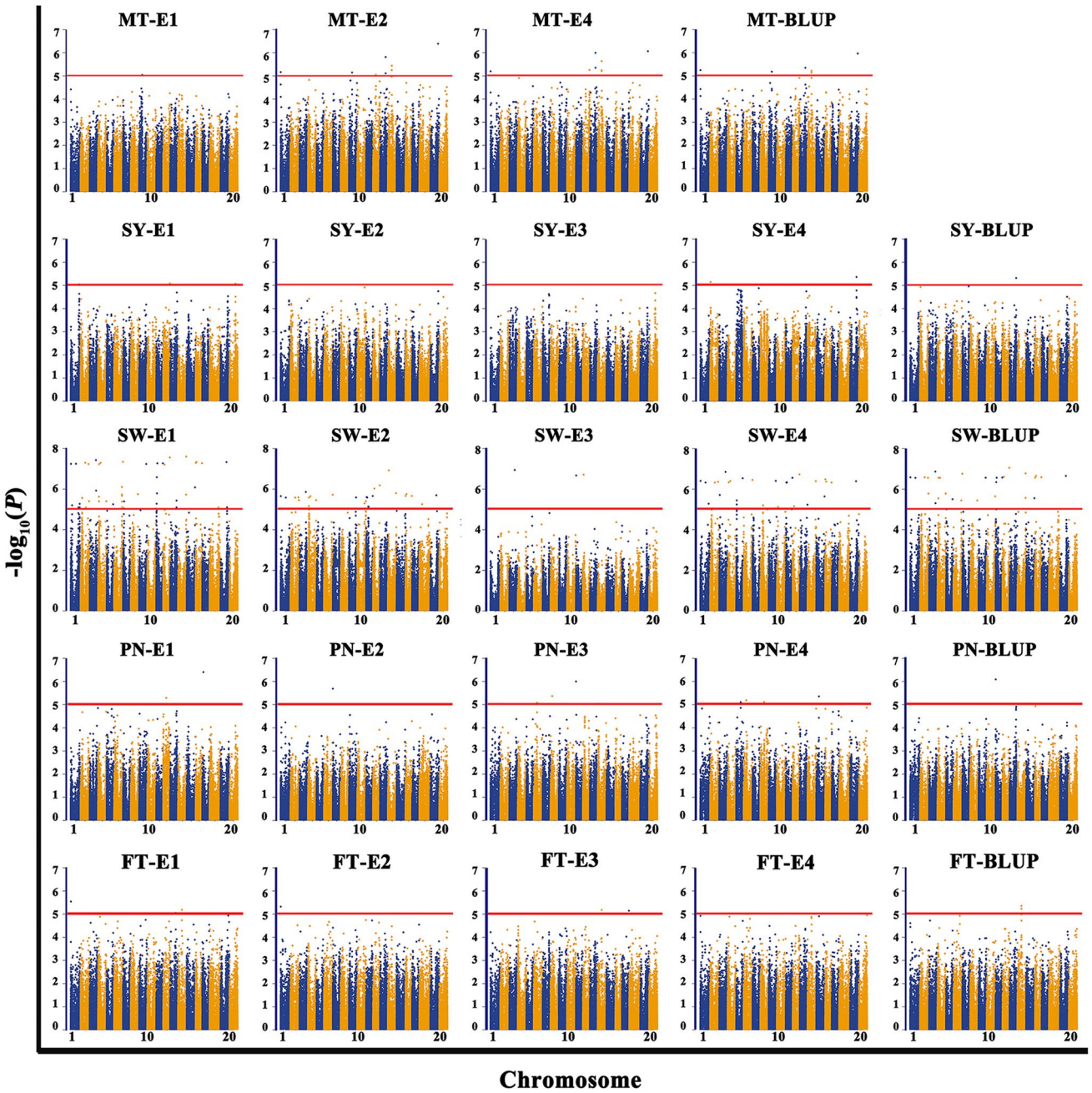


Fig. 4 Genome-wide association study of five yield traits in wild soybean populations across multiple environments. The horizontal line depicts the significant threshold (8.76×10^{-6})

AX-93713188-G and AX-94009457-T, respectively (Fig. 6c, d). Therefore, the dCAPS marker that was developed based on one (AX-94009457) of these three SNPs was assayed in 18 wild soybean accessions, representing materials with high SW, moderate SW and low SW. As shown in Fig. 7, all ten materials with low SW produced 201-bp products by enzyme digestion. Five materials with high SW produced 177-bp products by enzyme digestion. However, the remaining three moderate-SW materials also produced 177-bp products. It is possible that the SW in wild soybean is a complex quantitative trait, thus it is hard to distinguish all materials based on a single gene marker. These results confirmed that this dCAPS marker is strongly associated with SW and validated dCAPS as a functional marker.

Table 2 SNP loci that significantly associated with SY, SW, FT and MT in two or more environments

SNP ID	Chr	Position (W05)	Trait	*P* value					R^2 (%)	Related QTLs/SNPs
				E1	E2	E3	E4	BLUP		
AX-93914172	1	954,964	MT	ns	6.87×10^{-6}	/	6.33×10^{-6}	5.64×10^{-6}	9.66–9.96	/
AX-93665227	1	1,205,496	FT	2.86×10^{-6}	4.76×10^{-6}	ns	ns	ns	11.11–12.87	/
AX-93961487	1	1,520,619	SW	5.69×10^{-8}	2.64×10^{-7}	ns	3.96×10^{-7}	2.64×10^{-7}	9.32–13.04	*Seed weight 18-1.2* (Panthee et al. 2005) *cqSeed weight-010* (Pathan et al. 2013)
AX-93616455	1	27,635,741	SW	5.63×10^{-8}	2.63×10^{-6}	ns	4.74×10^{-7}	2.79×10^{-7}	9.18–13.05	/
AX-93979681	2	26,610,200	SW	5.02×10^{-8}	2.62×10^{-6}	ns	4.76×10^{-7}	2.62×10^{-7}	9.18–13.16	*Seed yield 28-10* (Rossi et al. 2013)
AX-93687543	2	44,233,867	SW	5.80×10^{-8}	2.57×10^{-6}	ns	4.36×10^{-7}	2.78×10^{-7}	9.25–13.03	*Seed weight 50-12* (Kato et al. 2014)
AX-93985010	2	47,365,724	SW	3.99×10^{-6}	3.52×10^{-6}	ns	ns	1.66×10^{-6}	7.51–8.27	
AX-94274381	3	35,831,541	SW	3.71×10^{-8}	1.39×10^{-6}	ns	1.42×10^{-7}	1.37×10^{-7}	10.09–13.45	/
AX-94000466	4	5,303,432	SW	2.38×10^{-6}	6.89×10^{-6}	ns	ns	2.84×10^{-6}	7.13–8.17	*Seed weight 38-2* (Yang et al. 2011)
AX-94000835	4	6,362,101	SW	5.81×10^{-8}	2.62×10^{-6}	ns	4.69×10^{-7}	2.75×10^{-7}	9.19–13.03	
AX-93706936	4	14,949,007	SW	5.07×10^{-8}	2.25×10^{-6}	ns	3.99×10^{-7}	2.78×10^{-7}	9.31–13.15	*Seed weight 50-9* (Kato et al. 2014)
AX-94009450	4	49,444,067	SW	3.99×10^{-6}	3.52×10^{-6}	ns	ns	1.66×10^{-6}	7.51–8.27	*Seed weight 5-g4* (Zhang et al. 2016b)
AX-94009451	4	49,445,786	SW	3.99×10^{-6}	3.52×10^{-6}	ns	ns	1.66×10^{-6}	7.51–8.27	
AX-93713187	4	49,453,295	SW	3.99×10^{-6}	3.52×10^{-6}	ns	ns	1.66×10^{-6}	7.51–8.27	
AX-93713188	4	49,453,902	SW	3.99×10^{-6}	3.52×10^{-6}	ns	ns	1.66×10^{-6}	7.51–8.27	
AX-94009464	4	49,472,619	SW	3.99×10^{-6}	3.52×10^{-6}	ns	ns	1.66×10^{-6}	7.51–8.27	
AX-94009465	4	49,473,889	SW	3.99×10^{-6}	3.52×10^{-6}	ns	ns	1.66×10^{-6}	7.51–8.27	
AX-93922119	5	38,959,835	SW	5.60×10^{-6}	ns	ns	2.92×10^{-7}	3.15×10^{-6}	7.05–8.19	*Seed weight 34-9* (Han et al. 2012)
AX-93737581	6	48,230,610	SW	4.55×10^{-8}	1.90×10^{-6}	ns	3.51×10^{-7}	1.74×10^{-7}	9.41–13.26	/
AX-94068012	9	41,819,465	SW	5.77×10^{-8}	2.60×10^{-6}	ns	3.94×10^{-7}	2.75×10^{-7}	9.32–13.03	*Seed weight per plant 2-1* (Vieira et al. 2006)
AX-94083659	11	3,380,445	SW	2.60×10^{-7}	2.50×10^{-6}	ns	4.81×10^{-7}	4.56×10^{-7}	7.84–9.99	/
AX-93787863	11	6,125,111	SW	5.35×10^{-8}	2.61×10^{-6}	ns	4.63×10^{-7}	2.75×10^{-7}	9.20–13.10	*Seed weight 21-1* (Gai et al. 2007)
AX-94112641	11	40,797,719	MT	ns	1.54×10^{-6}	/	1.01×10^{-6}	4.46×10^{-6}	9.93–11.57	/
AX-93815870	11	40,799,505	MT	ns	1.54×10^{-6}	/	1.01×10^{-6}	4.46×10^{-6}	9.93–11.57	*Seed weight 23-2* (Li et al. 2008)
AX-93815871	11	40,801,080	MT	ns	7.77×10^{-6}	/	4.45×10^{-6}	ns	11.72–11.93	/
AX-94276811	12	8,022,185	SW	6.15×10^{-8}	7.05×10^{-6}	ns	9.35×10^{-7}	2.49×10^{-7}	8.89–11.23	/
AX-93938118	12	42,554,533	SW	2.75×10^{-8}	6.54×10^{-7}	ns	1.88×10^{-7}	8.91×10^{-8}	9.88–13.74	/
AX-94091961	13	29,984,153	SW	5.36×10^{-8}	2.21×10^{-6}	ns	2.78×10^{-7}	2.69×10^{-7}	9.58–13.10	/
AX-93825970	14	21,490,937	MT	ns	5.62×10^{-6}	/	6.19×10^{-6}	7.26×10^{-6}	11.29–12.13	/
AX-93959615	14	21,888,250	FT	6.51×10^{-6}	ns	6.62×10^{-6}	ns	5.87×10^{-6}	10.21–11.88	/
			MT	ns	3.62×10^{-6}	/	2.30×10^{-6}	6.08×10^{-6}	9.62–10.71	/
AX-93832016	14	48,834,625	SW	2.50×10^{-8}	1.59×10^{-6}	ns	4.06×10^{-7}	1.68×10^{-7}	9.30–13.83	*Seed yield 31-1* (Wang et al. 2014)
AX-93945346	15	48,155,287	SW	8.21×10^{-7}	ns	ns	2.31×10^{-6}	2.81×10^{-6}	6.78–9.03	/
AX-93845958	16	875,833	SW	5.02×10^{-8}	2.17×10^{-6}	ns	3.28×10^{-7}	2.71×10^{-7}	9.46–13.16	/

Table 2 (continued)

SNP ID	Chr	Position (W05)	Trait	*P* value					R^2 (%)	Related QTLs/SNPs
				E1	E2	E3	E4	BLUP		
AX-93847768	16	5,595,022	SW	4.63×10^{-8}	1.77×10^{-6}	ns	4.29×10^{-7}	2.21×10^{-7}	9.26–13.24	*Seed yield 23-6* (Guzman et al. 2007)
AX-93856062	16	37,039,841	SW	5.25×10^{-8}	2.19×10^{-6}	ns	4.76×10^{-7}	2.60×10^{-7}	9.18–13.12	
AX-93893297	19	34,156,991	SW	4.68×10^{-8}	2.01×10^{-6}	ns	4.10×10^{-7}	2.24×10^{-7}	9.29–13.23	*Seed yield 11-5* (Specht et al. 2001)
AX-93896805	19	44,004,764	MT	ns	4.10×10^{-7}	ns	8.60×10^{-7}	1.08×10^{-6}	11.40–13.04	*First flower 24-3* (Kuroda et al. 2013)
AX-93907195	20	38,828,704	SY	8.42×10^{-6}	ns	3.41×10^{-6}	ns	ns	20.96–23.96	*Seed yield 21-7*, *Seed weight 24-3* (Reinprecht et al. 2006); *Seed weight 36-5* (Han et al. 2012); *Seed weight 50-13* (Kato et al. 2014)
AX-94205175	20	38,829,894	SY	8.42×10^{-6}	ns	3.41×10^{-6}	ns	ns	20.96–23.96	
AX-93956919	20	38,830,442	SY	8.42×10^{-6}	ns	3.41×10^{-6}	ns	ns	20.96–23.96	

BLUP best linear unbiased prediction, *Chr* chromosome, *ns* not significant

Discussion

New markers related to yield-related traits in wild soybean by high-density SNP array

Unique germplasm resources, such as wild accessions, are essential to identify important loci and genes for the improvement of crop plants. In the past 2 decades, crop wild relatives have gradually gained attention due to their higher genetic diversity than that of their domesticated descendants. Wild soybean is the closest relative of cultivated soybean. To date, studies on wild soybean have mainly focused on biotic or abiotic stresses, such as resistance to root-knot nematode (Xu et al. 2013) and soybean cyst nematode (Zhang et al. 2017). A small number of studies identified QTLs or genes related to yield-related traits (Hu et al. 2014), whereas few studies used GWAS with high-density SNP markers to understand the genetic architecture of yield-related traits.

In this study, GWAS was performed with four traits (SY, SW, PN, FT) in four environments. One trait (MT) was performed within three environments. Overall, 218 SNPs distributed on 20 chromosomes were associated with yield-related traits (Table S3). Of these, 41 SNPs were identified in two or more environments (Table 2). Twenty-nine of 41 SNPs were associated with SW. These SNPs were close to or within the previously identified QTLs (Table 2). Three consecutive SNPs located on chromosome 20 were associated with SY with R^2 values as high as 23.96%. These three SNPs were located in the region of *Seed yield 21-7 Seed weight 24-3* (Reinprecht et al. 2006), *Seed weight 36-5* (Han et al. 2012) and *Seed weight 50-13* (Kato et al. 2014), indicating that these SNPs are critical for seed yield and that candidate genes for seed yield can likely be identified. AX-93959615, located on chromosome 14, was the unique SNP co-associated with FT and MT in two different environments (Table 2; Table S3). The lower h^2 value of PN indicates that PN was easily affected by environmental factors, which might be the reason why no SNP was identified in two different environments for PN.

Overall, we identified several SNPs that existed both in the cultivated and wild soybean populations. The flanking regions of these SNPs could be narrowed down to previously reported QTL positions and were helpful in identifying the candidate genes underlying the QTLs. We also identified many novel SNPs that only exist in wild soybean, and these SNPs will enrich the knowledge of the genetic mechanism underlying yield in wild soybean.

Table 3 Candidate genes identified based on the position between significant SNPs and genomic region

SNP ID	Chr	Position (W05)	Location site	Gene ID	Functional annotation
AX-93914172	1	954,964	Intron	*Glysoja.01g000095*	Protein unc-13-like
AX-93665227	1	1,205,496	3′-UTR	*Glysoja.01g000124*	Sn1-specific diacylglycerol lipase alpha
AX-93961487	1	1,520,619	3′-UTR	*Glysoja.01g000148*	Inactive poly polymerase RCD1
AX-93616455	1	27,635,741	Intrergenic	/	/
AX-93979681	2	26,610,200	Intrergenic	/	/
AX-93687543	2	44,233,867	Non-synonymous coding	*Glysoja.02g004846*	No description
AX-93985010	2	47,365,724	5′-UTR	*Glysoja.02g005196*	E3 ubiquitin-protein ligase COP1
AX-94274381	3	35,831,541	3′-UTR	*Glysoja.03g006986*	Serine carboxypeptidase-like 45
AX-94000466	4	5,303,432	Intrergenic	/	/
AX-94000835	4	6,362,101	Intrergenic	/	/
AX-93706936	4	14,949,007	Intrergenic	/	/
AX-94009450	4	49,444,067	Intrergenic	/	/
AX-94009451	4	49,445,786	Intrergenic	/	/
AX-93713187	4	49,453,295	Non-synonymous coding	*Glysoja.04g010563*	Polyadenylate-binding protein-interacting protein
AX-93713188	4	49,453,902	Intron	*Glysoja.04g010563*	Polyadenylate-binding protein-interacting protein
AX-94009464	4	49,472,619	Non-synonymous coding	*Glysoja.04g010566*	Small heat shock protein, chloroplastic
AX-94009465	4	49,473,889	Non-synonymous coding	*Glysoja.04g010566*	Small heat shock protein, chloroplastic
AX-93922119	5	38,959,835	3′-UTR	*Glysoja.05g012674*	Protein activity of BC1 and complex kinase 8, chloroplastic
AX-93737581	6	48,230,610	Non-synonymous coding	*Glysoja.06g016409*	WRKY transcription factor 2
AX-94068012	9	41,819,465	Intron	*Glysoja.09g024829*	26S proteasome non-ATPase regulatory subunit 13-like A
AX-94083659	11	3,380,445	Intron	*Glysoja.11g028982*	Branched-chain amino acid aminotransferase 2, chloroplastic
AX-93787863	11	6,125,111	3′-UTR	*Glysoja.11g029328*	No description
AX-94112641	11	40,797,719	Intron	*Glysoja.11g031433*	Cullin-4
AX-93815870	11	40,799,505	Intron	*Glysoja.11g031433*	Cullin-4
AX-93815871	11	40,801,080	Intron	*Glysoja.11g031433*	Cullin-4
AX-94276811	12	8,022,185	3′-UTR	*Glysoja.12g033845*	Protein FAF-like, chloroplastic
AX-93938118	12	42,554,533	Intron	*Glysoja.12g035175*	Hypothetical protein
AX-94091961	13	29,984,153	Non-synonymous coding	*Glysoja.13g037056*	DNA polymerase epsilon catalytic subunit A
AX-93825970	14	21,490,937	Intrergenic	/	/
AX-93959615	14	21,888,250	Intrergenic	/	/
AX-93832016	14	48,834,625	Promoter	*Glysoja.14g039290*	UDP-glycosyltransferase 87A1
AX-93945346	15	48,155,287	3′-UTR	*Glysoja.15g042025*	Gibberellin 2-beta-dioxygenase 1
AX-93845958	16	875,833	Synonymous coding	*Glysoja.16g042366*	Exocyst complex component EXO70A1
AX-93847768	16	5,595,022	Promoter	*Glysoja.16g042837*	Hypothetical protein
AX-93856062	16	37,039,841	Promoter	*Glysoja.16g044294*	No description
AX-93893297	19	34,156,991	3′-UTR	*Glysoja.19g051042*	No description
AX-93896805	19	44,004,764	Intergenic	/	/
AX-93907195	20	38,828,704	Intron	*Glysoja.20g053860*	Caffeoylshikimate esterase
AX-94205175	20	38,829,894	Intron	*Glysoja.20g053860*	Caffeoylshikimate esterase
AX-93956919	20	38,830,442	Non-synonymous coding	*Glysoja.20g053860*	Caffeoylshikimate esterase

Chr chromosome

Candidate genes for SW in wild soybean

The yield and yield-related traits of soybean are complex and quantitative, and environmental variations can trigger and modify the actions of related genes (Li et al. 2005). h^2 is usually used to evaluate the impact of environmental factors on a given trait. Only the traits with high h^2 values could be mapped stably (Hao et al. 2012). The h^2 value of SW was the highest in yield-related traits in both wild soybean and cultivated soybean, and several

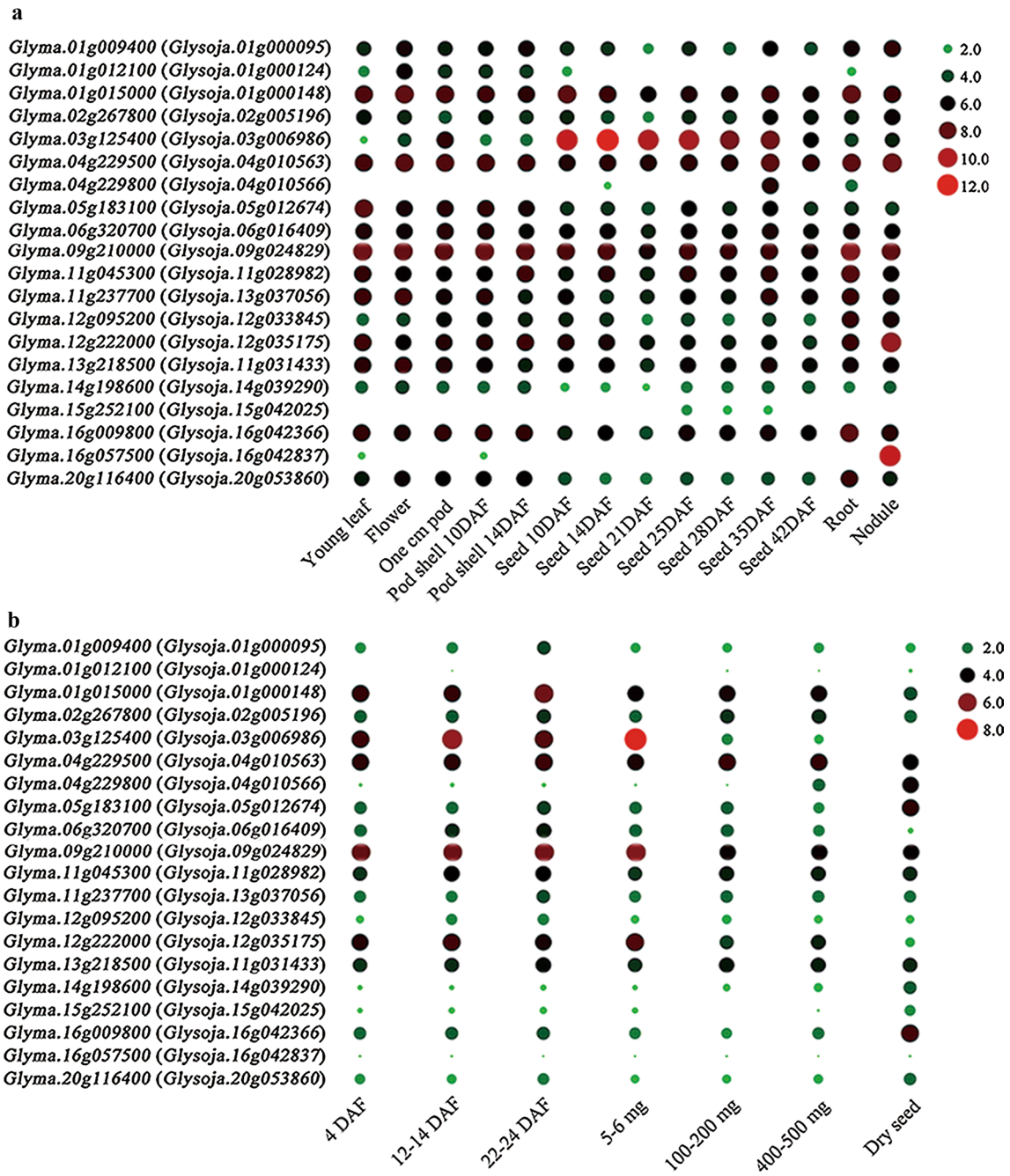
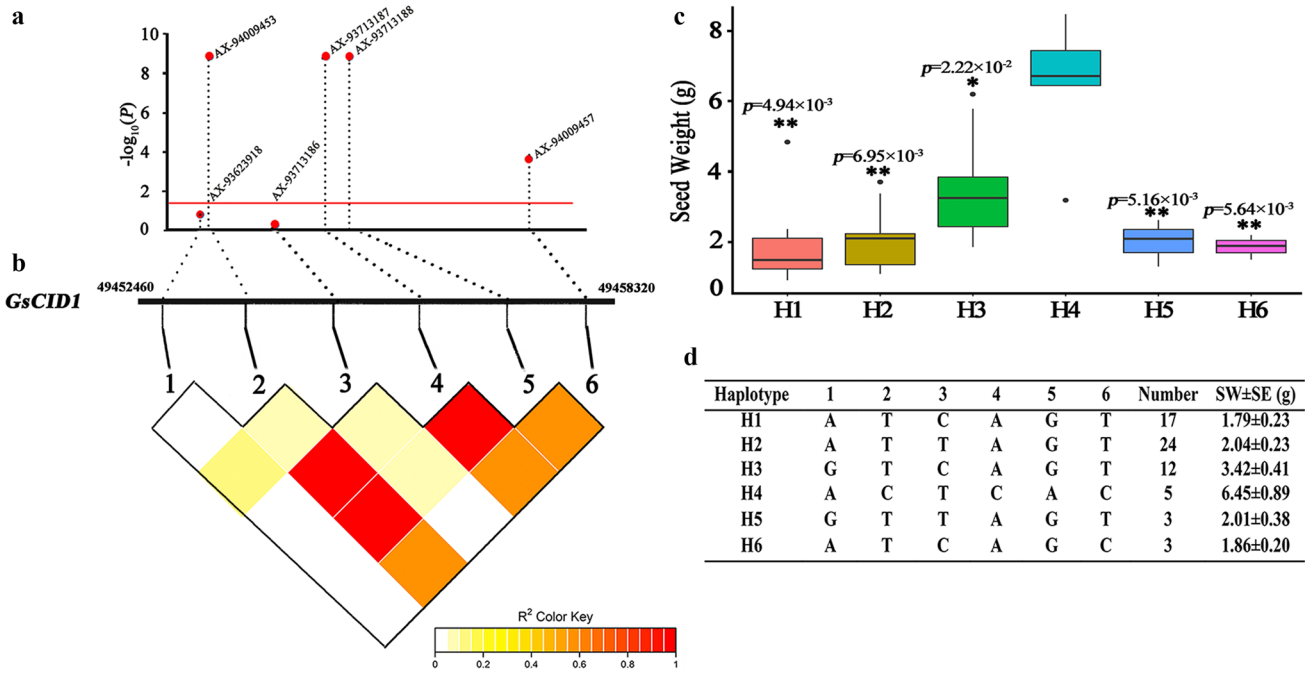


Fig. 5 Digital expression profiles for 20 genes in various tissues. **a** Expression levels of 20 genes in three vegetative tissues and at different seed development stages based on RNA-seq data (downloaded from the SoyBase website https://www.soybase.org/soyseq). The reads per kilobase million-normalized values were $\log_2$-transformed. **b** Expression of 20 genes in four different pod development stages from microarray data (GEO Accession number: GSE42871). The robust multiarray analysis-normalized values from the microarray data were $\log_2$-transformed

candidate genes underlying SW were reported. For example, *GmGA20ox* was identified by analyzing the transcriptional characteristics of soybean seed development (Lu et al. 2016). Overexpression of *GmGA20ox* increases the SW in *Arabidopsis thaliana*. Zhang et al. (2015) identified 39 candidate genes involved in SW by establishing a 100-seed weight QTL-allele matrix in cultivated soybean (Zhang et al. 2015).

In this study, the h^2 value of SW was also the highest among the five traits ($h^2 = 90.62\%$) (Table 1). By genotyping the data from four or three environments and the NJAU 355K SoySNP array, we identified 29 SNPs for SW in two



Haplotype	1	2	3	4	5	6	Number	SW±SE (g)
H1	A	T	C	A	G	T	17	1.79±0.23
H2	A	T	T	A	G	T	24	2.04±0.23
H3	G	T	C	A	G	T	12	3.42±0.41
H4	A	C	T	C	A	C	5	6.45±0.89
H5	G	T	T	A	G	T	3	2.01±0.38
H6	A	T	C	A	G	C	3	1.86±0.20

Fig. 6 Polymorphisms of *GsCID1* are significantly associated with SW in wild soybean. **a** *GsCID1*-based association mapping with SW (GLM model, Tassel software 5.0, $P<0.05$). The horizontal line depicts the $-\log_{10}(P)=1.30$. **b** Pairwise LD analysis between the six *GsCID1* SNPs. The LD plot is represented by the inverted triangle. The gene model of *GsCID1* is indicated by a black rectangle, and the physical position of the SNP is drawn above the plot. The LD level between six SNPs was indexed by the R^2 value. 1–6 represents AX-93623918, AX-94009453, AX-93713186, AX-93713187, AX-93713188 and AX-94009457, respectively. **c** The boxplot of SW values between Hap1–Hap6. Significance analysis between H4 and H1, H2, H3, H4, H5, and H6 was performed. Statistical significance was detected by a two-tailed *t* test. *Significant at $0.01<P<0.05$; **significant at $0.001<P<0.01$. **d** Haplotypes of *GsCID1* among 64 wild soybean accessions

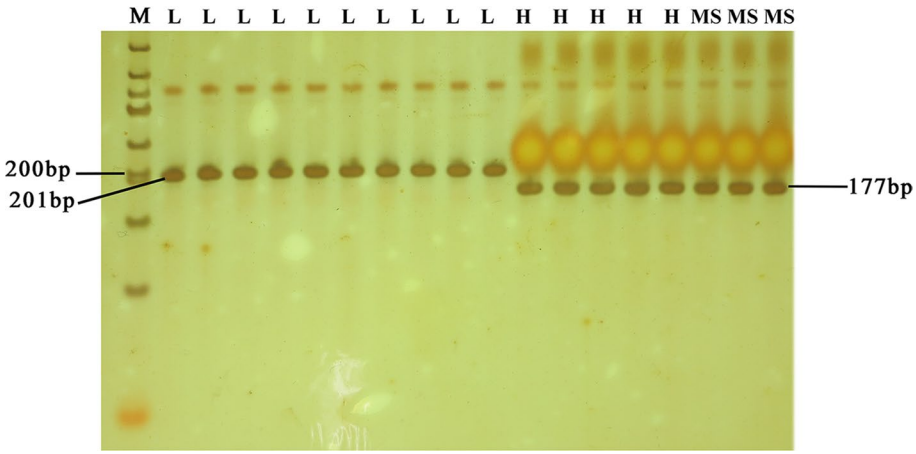


Fig. 7 PCR products of 18 wild soybean varieties using the dCAPS marker. The samples are shown in Table S2. Lanes: *M* marker, *L* the low-SW materials, *H* the high-SW materials, *MS* the moderate-SW materials

or more environments (Table 2). Six consecutive SNPs spanning 49,444,067–49,473,889 and located on chromosome 4 were detected in two environments. Two of these six SNPs (AX-93713187 and AX-93713188) were located in the region of *GsCID1* (*Glysoja.04g010563*) and caused a non-synonymous mutation (Table 3). Tissue expression patterns based on public information resources showed that *GsCID1* was expressed in nearly all tissues and was highly expressed in seeds at 35 DAF (Fig. 5a; Table S6). In the RNA-seq data from different seed development periods, *GsCID1* was continuously expressed at high levels through all stages of seed development (Fig. 5b; Table S7). Polymorphism analysis suggested that *GsCID1* was associated with SW in wild soybean using SNP information from the NJAU 355K SoySNP array (Fig. 6). *GsCID1* contains a conserved RNA recognition motif (RRM) domain, which was reported to participate in constitutive pre-mRNA splicing and regulate posttranscriptional gene expression

(Birney et al. 1993; Maris et al. 2005). In rice, overexpression of the RRM domain of *OsFCA* increases cell size and rice yield (Hong et al. 2007). In *Brassica napus*, *BncsRRM2*, encoding an RRM domain protein, positively regulates yield components in both transgenic cotton and *Brassica napus* (Qi et al. 2012; Sun et al. 2012). In barley, a glycine-rich RNA-binding protein participates in the regulation of barley development and the stress response (Tripet et al. 2014). RRMs are highly conserved in plants and animals as important regulators of gene expression, and we speculate that the soybean RRM domain gene may have the same function in improving SW. In summary, *GsCID1* might be a candidate gene related to SW in wild soybean and should be investigated further.

In our GWAS results, only one SNP, AX-93959615, located on chromosome 14, was significantly co-associated with MT and FT. Four genes were identified in the 80 kb (LD decay distance of this wild soybean population) flanking regions of AX-93959615. Among them, Glysoja.14g038573 is homologous to AtARR10 (At4g31920), which takes part in the cytokinin signaling pathway in *Arabidopsis* (Nguyen et al. 2016; Zubo et al. 2017; Xie et al. 2018). We also found that three consecutive SNPs on chromosome 20 were significantly associated with SY, and their R^2 values were as high as 23.96%. Of these three significant SNPs, AX-93907195 and AX-94205175 were located on the intron of *Glysoja.20g053860*, and AX-93956919 was located in the exon of *Glysoja.20g053860*, with a non-synonymous mutation. *Glysoja.20g053860*, encoding a monoglyceride lipase-like protein, was involved in the acylglycerol degradation process (https://www.kegg.jp/module/M00098) and glycerolipid metabolism pathway (https://www.kegg.jp/pathway/ko00561). The biological functions of these genes will be further investigated in future works.

Author contribution statement GK and DY designed this research; DH, QD, JW and FH conducted GWAS; HZ and ZH performed the field experiments; DH, QD, XL, ZY and HW developed the dCAPS marker; DH and GK wrote this manuscript. All authors approved the manuscript.

Acknowledgements This work was supported in part by the Ministry of Science and Technology (2016YFD0100304, 2017YFE0111000), the Key Transgenic Breeding Program of China (2016ZX08004-003, 2016ZX08009003-004), and the National Natural Science Foundation of China (31871649, 31671715).

Compliance with ethical standards

Ethical standards This research complied with ethical standards.

Conflict of interest The authors declare that they have no conflict of interest.

References

Birney E, Kumar S, Krainer AR (1993) Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. Nucleic Acids Res 21(25):5803–5816

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633–2635

Chapman A, Pantalone VR, Ustun A, Allen FL, Landauellis D, Trigiano RN, Gresshoff PM (2003) Quantitative trait loci for agronomic and seed quality traits in an F2 and F4:6 soybean population. Euphytica 129(3):387–393

Chu SS, Wang J, Zhu Y, Liu SL, Zhou X, Zhang HR, Wang CE, Yang WM, Tian ZX, Cheng H, Yu DY (2017) An R2R3-type MYB transcription factor, GmMYB29, regulates isoflavone biosynthesis in soybean. PLoS Genet 13(5):e1006770

Dhanapal AP, Ray JD, Singh SK, Hoyos-Villegas V, Smith JR, Purcell LC, King CA, Cregan PB, Song Q, Fritschi FB (2015) Genome-wide association study (GWAS) of carbon isotope ratio (δ 13 C) in diverse soybean [Glycine max (L.) Merr.] genotypes. Theor Appl Genet 128(1):73–91

Edae EA, Byrne PF, Haley SD, Lopes MS, Reynolds MP (2014) Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theor Appl Genet 127(4):791–807

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164(4):1567–1587

Fang L, Wang Q, Hu Y, Jia Y et al (2017) Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. Nat Genet 49(7):1089–1098

Fehr WR, Caviness CE (1977) Stages of soybean development. Ames, IA, USA: Cooperative Extension Service, Agriculture and Home Economics Experiment Station, Iowa State University of Science and Technology

Gai JY, Wang YJ, Wu XL, Chen SY (2007) A comparative study on segregation analysis and QTL mapping of quantitative traits in plants—with a case in soybean. Front Agric Chin 1(1):1–7

Guzman PS, Diers BW, Neece DJ, Martin SKS, Leroy AR, Grau CR, Hughes TJ, Nelson RL (2007) QTL associated with yield in three backcross-derived populations of soybean. Crop Sci 47(1):111–122

Han YP, Li DM, Zhu D, Li HY, Li XP, Teng WL, Li WB (2012) QTL analysis of soybean seed weight across multi-genetic backgrounds and environments. Theor Appl Genet 125(4):671–683

Hao DR, Cheng H, Yin ZT, Cui SY, Zhang D, Wang H, Yu DY (2012) Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. Theor Appl Genet 124(3):447–458

Hong F, Attia K, Wei C, Li KG, He GM, Su W, Zhang QH, Qian XY, Yang JS (2007) Overexpression of the rFCA RNA recognition motif affects morphologies modifications in rice (Oryza sativa L.). Biosci Rep 27(4–5):225–234

Hu ZB, Zhang D, Zhang GZ, Kan GZ, Hong DR, Yu DY (2014) Association mapping of yield-related traits and SSR markers in wild soybean (Glycine soja Sieb. and Zucc.). Breeding Sci 63(5):441–449

Hu DZ, Kan GZ, Hu W, Li YL, Hao DR, Li X, Yang H, Yang ZY, He XH, Huang F, Yu DY (2019) Identification of loci and candidate genes responsible for pod dehiscence in soybean via genome-wide association analysis across multiple environments. Front Plant Sci 10:811–811

Huang C, Sun H, Xu DY, Chen QY, Liang YM, Wang XF, Xu GH, Tian JG, Wang CL, Li D, Wu LS, Yang XH, Jin WW, Doebley JF, Tian

F (2017) *ZmCCT9* enhances maize adaptation to higher latitudes. Proc Natl Acad Sci USA 115(2):E334–E341
Hyten DL, Pantalone VR, Sams CE, Saxton AM, Landauellis D, Stefaniak TR, Schmidt ME (2004) Seed quality QTL in a prominent soybean population. Theor Appl Genet 109(3):552–561
Kato S, Sayama T, Fujii K, Yumoto S, Kono Y, Hwang TY, Kikuchi A, Takada Y, Tanaka Y, Shiraiwa T, Ishimoto M (2014) A major and stable QTL associated with seed weight in soybean across multiple environments and genetic backgrounds. Theor Appl Genet 127(6):1365–1374
Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9(1):29
Kuroda Y, Kaga A, Tomooka N, Yano H, Takada Y, Kato S, Vaughan D (2013) QTL affecting fitness of hybrids between wild and cultivated soybeans in experimental fields. Ecol Evolution 3(7):2150–2168
Leamy LJ, Zhang HY, Li CB, Chen CY, Song BH (2017) A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). BMC Genomics 18(1):18
Li JZ, Huang XQ, Heinrichs F, Ganal MW, Roder MS (2005) Analysis of QTLs for yield, yield components, and malting quality in a BC3-DH population of spring barley. Theor Appl Genet 110(2):356–363
Li DD, Pfeiffer TW, Cornelius PL (2008) Soybean QTL for yield and yield components associated with *Glycine soja* alleles. Crop Sci 48(2):571–581
Lu X, Li QT, Xiong Q, Li W, Bi YD, Lai YC, Liu XL, Man WQ, Zhang W, Ma B, Chen SY, Zhang JS (2016) The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. Plant J 86(6):530–544
Ma ZY, He SP, Wang XF et al (2018) Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. Nat Genet 50(6):803–813
Maris C, Dominguez C, Allain FHT (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J 272(9):2118–2131
Merk HL, Yarnes SC, Deynze VA, Tong N, Menda N, Mueller LA, Mutschler MA, Loewen SA, Myers JR, Francis DM (2012) Trait diversity and potential for selection indices based on variation among regionally adapted processing tomato germplasm. J Amer Soc Horticult Sci 137(6):427–437
Nevo E, Chen GX (2010) Drought and salt tolerances in wild relatives for wheat and barley improvement. Plant Cell Environ 33(4):670–685
Nguyen KH, Ha VC, Nishiyama R, Watanabe Y, Leyvagonzalez MA, Fujita Y, Tran UT, Li WQ, Tanaka M, Seki M, Schaller GE, Herreraestrella L, Tran LP (2016) *Arabidopsis* type B cytokinin response regulators ARR1, ARR10, and ARR12 negatively regulate plant responses to drought. Proc Natl Acad Sci USA 113(11):3090–3095
Nichols DM, Glover KD, Carlson SR, Specht JE, Diers BW (2006) Fine mapping of a seed protein QTL on soybean linkage group i and its correlated effects on agronomic traits. Crop Sci 46(2):834–839
Okishio T, Sasayama D, Hirano T, Akimoto M, Itoh K, Azuma T (2014) Growth promotion and inhibition of the Amazonian wild rice species *Oryza grandiglumis* to survive flooding. Planta 240(3):459–469
Panthee DR, Pantalone VR, West DR, Saxton AM, Sams CE (2005) Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. Crop Sci 45(5):2015–2022
Pathan SM, Vuong TD, Clark KM, Lee JD, Shannon JG, Roberts CA, Ellersieck MR, Burton JW, Cregan PB, Hyten DL, Nguyen NT (2013) Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. Crop Sci 53(3):765–774
Placido D, Campbell MT, Folsom JJ, Cui XP, Kruger GR, Baenziger PS, Walia H (2013) Introgression of novel traits from a wild wheat relative improves drought adaptation in wheat. Plant Physiol 161(4):1806–1819
Qi WW, Zhang FQ, Sun F, Huang YJ, Guan RZ, Yang JS, Luo XJ (2012) Over-expression of a conserved RNA-binding motif (RRM) domain (*csRRM2*) improves components of *Brassica napus* yield by regulating cell size. Plant Breed 131(5):614–619
Qi XP, Li MW, Xie M et al (2014) Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat Commun 5(1):4340–4340
Reinprecht Y, Poysa V, Yu KY, Rajcan I, Ablett GR, Pauls KP (2006) Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (Glycine max (L.) Merrill) germplasm. Genome 49(12):1510–1527
Riedelsheimer C, Lisec J, Czedikeysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci USA 109(23):8872–8877
Rossi ME, Orf JH, Liu LJ, Dong ZM, Rajcan I (2013) Genetic basis of soybean adaptation to North American vs Asian mega-environments in two independent populations from Canadian Chinese crosses. Theor Appl Genetics 126(7):1809–1823
Shi YY, Gao LL, Wu ZC, Zhang XJ, Wang MM, Zhang CS, Zhang F, Zhou YL, Li ZK (2017) Genome-wide association study of salt tolerance at the seed germination stage in rice. BMC Plant Biol 17(1):92
Shin J-H, Blay S, McNeney B, Graham J (2006) LDheatmap: An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Softw 16(1):1–9
Shirasawa K, Fukuoka H, Matsunaga H, Kobayashi Y, Kobayashi I, Hirakawa H, Isobe S, Tabata S (2013) Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. DNA Res 20(6):593–603
Specht JE, Chase K, Macrander M, Graef GL, Chung J, Markwell JP, Germann M, Orf JH, Lark KG (2001) Soybean response to water: A QTL analysis of drought tolerance. Crop Sci 41(2):493–509
Sun F, Liu CL, Zhang CJ, Qi WW, Zhang XY, Wu ZX, Kong DP, Wang QH, Shang HH, Qian XY, Li FG, Yang JS (2012) A conserved RNA recognition motif (RRM) domain of *Brassica napus* FCA improves cotton fiber quality and yield by regulating cell size. Mol Breed 30(1):93–101
Tian F, Zhu ZF, Zhang BS, Tan LB, Fu YC, Wang XK, Sun CQ (2006) Fine mapping of a quantitative trait locus for grain number per panicle from wild rice (Oryza rufipogon Griff.). Theor Appl Genetics 113(4):619–629
Treuren RV, Hoekstra R, Hintum TV (2017) Inventory and prioritization for the conservation of crop wild relatives in The Netherlands under climate change. Biol Cons 216:123–139
Tripet BP, Mason KE, Eilers BJ, Bruns J, Powell P, Fischer AM, Valérie C (2014) Structural and biochemical analysis of the Hordeum vulgare L HvGR-RBP1 protein, a glycine-rich RNA-binding protein involved in the regulation of barley plant development and stress response. Biochemistry 53(50):7945–7960
Turner SD (2014) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Biorxiv
Vieira AJD, Oliveira DAA, Soares TCB, Schuster I, Piovesan ND, Martinez CA, Barros EDG, Moreira MA (2006) Use of the QTL approach to the study of soybean trait relationships in two populations of recombinant inbred lines at the F7 and F8 generations. Brazilian J Plant Physiol 18(2):281–290
Vuong TD, Sonah H, Meinhardt CG, Deshmukh R, Kadam S, Nelson RL, Shannon JG, Nguyen HT (2015) Genetic architecture of cyst

nematode resistance revealed by genome-wide association study in soybean. BMC Genomics 16(1):593–593

Wang XZ, Jiang GL, Green M, Scott RA, Song QJ, Hyten DL, Cregan PB (2014) Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. Mol Genetics Genomics 289(5):935–949

Wang QX, Xie WB, Xing HK, Yan J, Meng XZ, Li XH, Fu XK, Xu JY, Lian XM, Yu SB, Xing YZ, Wang GW (2015) Genetic architecture of natural variation in rice chlorophyll content revealed by a genome-wide association study. Mol Plant 8(6):946–957

Wang J, Chu SS, Zhang HR, Zhu Y, Cheng H, Yu DY (2016) Development and application of a novel genome-wide SNP array reveals domestication history in soybean. Sci Rep 6(1):20728–20728

Xie MT, Chen HY, Huang L, O'Neil RC, Shokhirev MN, Ecker JR (2018) A B-ARR-mediated cytokinin transcriptional network directs hormone cross-regulation and shoot development. Nat Commun 9(1):1604

Xie M, Chung CYL, Li MM et al (2019) A reference-grade wild soybean genome. Nat Commun 10(1):1–12

Xu XY, Zeng L, Tao Y, Vuong T, Wan JR, Boerma R, Noe J, Li Z, Finnerty S, Pathan SM, Shannon JG, Nguyen HT (2013) Pinpointing genes underlying the quantitative trait loci for root-knot nematode resistance in palaeopolyploid soybean by whole genome resequencing. Proc Natl Acad Sci USA 110(33):13469–13474

Yang KW, Moon JK, Jeong NH, Chun HK, Kang ST, Back KW, Jeong SC (2011) Novel major quantitative trait loci regulating the content of isoflavone in soybean seeds. Genes Genom 33(6):685–692

Yang Q, Li Z, Li WQ, Ku LX, Wang C, Ye JR, Li K, Yang N, Li YP, Zhong T, Li JS, Chen YH, Yan JB, Yang XH, Xu ML (2013) CACTA-like transposable element in *ZmCCT* attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proc Natl Acad Sci USA 110(42):16969–16974

Zhang YH, He JB, Wang YF, Xing GN, Zhao JM, Li Y, Yang SP, Palmer RG, Zhao TJ, Gai JY (2015) Establishment of a 100-seed weight quantitative trait locus–allele matrix of the germplasm population for optimal recombination design in soybean breeding programmes. J Exp Bot 66(20):6311–6325

Zhang JP, Song QJ, Cregan PB, Jiang GL (2016a) Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). Theor Appl Genet 129(1):117–130

Zhang X, Warburton ML, Setter T, Liu H, Xue YD, Yang N, Yan JB, Xiao YJ (2016b) Genome-wide association studies of drought-related metabolic changes in maize using an enlarged SNP panel. Theor Appl Genet 129(8):1449–1463

Zhang HY, Song QJ, Griffin JD, Song BH (2017) Genetic architecture of wild soybean (*Glycine soja*) response to soybean cyst nematode (*Heterodera glycines*). Mol Genet Genomics 292(6):1257–1265

Zhang W, Liao XL, Cui YM, Ma WY, Zhang XN, Du HY, Ma YJ, Ning LH, Wang H, Huang F, Yang H, Kan GZ, Yu DY (2019) A cation diffusion facilitator, *GmCDF1*, negatively regulates salt tolerance in soybean. PLoS Genet 15(1):e1007798

Zubo YO, Blakley IC, Yamburenko MV, Worthen JM, Street IH, Francozorrilla JM, Zhang WJ, Hill K, Raines T, Solano R, Kieber JJ, Loraine AE, Schaller GE (2017) Cytokinin induces genome-wide binding of the type-B response regulator ARR10 to regulate growth and development in *Arabidopsis*. Proc Natl Acad Sci USA 114(29):E5995–E6004