Theoretical and Applied Genetics (2021) 134:1329–1341
https://doi.org/10.1007/s00122-021-03774-6

ORIGINAL ARTICLE

Comparative selective signature analysis and high-resolution GWAS reveal a new candidate gene controlling seed weight in soybean

Wei Zhang[1] · Wenjing Xu[2] · Hongmei Zhang[1] · Xiaoqing Liu[1] · Xiaoyan Cui[1] · Songsong Li[1,2] · Li Song[3] · Yuelin Zhu[2] · Xin Chen[1] · Huatao Chen[1]

Received: 10 November 2020 / Accepted: 11 January 2021 / Published online: 28 January 2021

Abstract

Key message **We detected a QTL *qHSW-16* undergone strong selection associated with seed weight and identified a novel candidate gene controlling seed weight candidate gene for this major QTL by qRT-PCT.**

Abstract Soybean [*Glycine max* (L.) Merr.] provides more than half of the world's oilseed production. To expand its germplasm resources useful for breeding increased yield and oil quality cultivars, it is necessary to resolve the diversity and evolutionary history of this crop. In this work, we resequenced 283 soybean accessions from China and obtained a large number of high-quality SNPs for investigation of the population genetics that underpin variation in seed weight and other agronomic traits. Selective signature analysis detected 78 (~25.0 Mb) and 39 (~22.60 Mb) novel putative selective signals that were selected during soybean domestication and improvement, respectively. Genome-wide association study (GWAS) identified five loci associated with seed weight. Among these QTLs, *qHSW-16*, overlapped with the improvement-selective region on chromosome 16, suggesting that this QTL may be underwent strong selection during soybean improvement. Of the 18 candidate genes in *qHSW-16*, only *SoyZH13_16G122400* showed higher expression levels in a large seed variety compared to a small seed variety during seed development. These results identify *SoyZH13_16G122400* as a novel candidate gene controlling seed weight and provide foundational insights into the molecular targets for breeding improvement of seed weight and potential seed yield in soybean.

Introduction

Soybean [*Glycine max* (L.) Merr.] is a legume crop that provides approximately 68% of the world's protein meal and 57% of vegetable oil production. Archaeological and evolutionary studies have shown that soybean was domesticated from its progenitor (*Glycine soja* Sieb. & Zucc.) approximately 5000 years ago in East Asia (Lee et al. 2011; Tengfei et al. 2019). A suite of domestication-related traits, including seed coat color (Wang et al. 2018) and seed weight (Zhou et al. 2015), has been selected for improvement in traditional landraces as well as in cultivar breeding programs.

Seed weight (SW), a major component of yield, is a complex and agronomically important trait that affects the quality of many soybean products, such as soy sprouts, soy nuts, edamame, soy sauce, and natto. This complex trait is always affected by many genetic and environmental factors including temperature and precipitation during the seed development stage (Liang et al. 2016; Wu et al. 2018). More than 200 quantitative trait loci/nucleotides (QTL/QTNs) for soybean seed weight have been reported in SoyBase (www.

Communicated by Istvan Rajcan.

Wei Zhang and Wenjing Xu have contributed equally to this work.

✉ Xin Chen
cx@jaas.ac.cn

✉ Huatao Chen
cht@jaas.ac.cn

[1] Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China

[2] College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

[3] Joint International Research Laboratory of Agriculture and Agri-Product Safety, Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding, Co-Innovation Center for Modern Production Technology of Grain Crops, Yangzhou University, Yangzhou 225009, China

soybase.org) via linkage mapping. However, integrating results from linkage mapping into breeding program is challenging due to the higher confidence interval and less genetic variation. As a result, genome-wide association study is used to take advantage of all recombination events that occur in the evolutionary history of a natural population based on linkage disequilibrium (LD) in the recent years. GWAS allows researchers to utilize natural diversity and locate valuable genes in the genome. Dissecting the genetic basis of SW is thus necessary for increasing soybean yield potential and improvement of soybean food as well.

To understand soybean domestication and improvement at a genetic level as a foundation for future soybean breeding efforts, we resequenced 283 soybean accessions to ≥ 12× depth. We then identified 10,210,329 single-nucleotide polymorphisms (SNPs), which were then used to detect 78 domestication-selective sweeps and 39 improvement-selective sweeps across the whole genome in this resequenced population. Further, a total of 689 and 1204 putative genes were identified in domestication-selective and improvement-selective sweeps, respectively. In addition, GWAS identified *qHSW-16,* a QTL significantly associated with seed weight that overlapped with the improvement-selective region on chromosome 16, suggesting that this locus underwent strong selection during soybean improvement. Among the candidate genes found in *qHSW-16*, *SoyZH13_16G122400* was previously shown to be differentially upregulated in a large seed soybean variety compared with its expression in a small seed variety during seed development. We therefore propose that this is novel candidate gene that controls seed weight. This study provides foundational insight into targets for genomics-enabled breeding for yield through the identification of useful loci, candidate genes, and beneficial alleles underlying seed weight in soybean.

Results

Genomic variation, structure, and linkage disequilibrium of soybean population

A total of 283 soybean accessions, including 19 wild accessions (*G. soja*), 52 landraces, and 212 improved cultivars, were used for resequencing in this study. The majority (17 wild accessions, 52 landraces, and 185 improved cultivars) of these 283 accessions came from the Huang-Huai region, and the southern region in China, while the others were from the northern region in China (Fig. 1a).

A total of 11.21 billion paired-end reads averaging 150 bp in length (3.3 Tb of base pairs), with an average coverage depth of more than 12.4× were generated by resequencing the 283 soybean accessions on a Novaseq-PE 150 sequencer. After mapping against the soybean Gmax_ZH13 reference genome, 10,210,329 SNPs and 2,870,410 insertions or deletions (indels) were identified from the resequenced 283 accessions. Among these SNPs, a total of 1,034,255 (10.13%) of the 10,210,329 were mapped to genes, within which 331,838 SNPs were identified in the coding sequences, 616,521 SNPs in introns, 27,919 SNPs in 5′-untranslated regions (UTRs), and 57,977 SNPs in 3′-UTRs. In addition, 7386 SNPs resulted in start codon changes, premature stop codons, or elongation of the transcripts; 1.85% and 1.34% of the total SNPs led to non-synonymous or synonymous mutations, respectively (Supplementary Table 2). We also found that the wild accessions contained more SNPs than were identified in either the landrace and cultivar accessions (Supplementary Fig. 1), indicating a richer genetic diversity among the wild accessions than in the landrace and cultivar accessions.

Removal of all SNPs with an average coverage depth < 8× and minor allele frequencies < 0.05, yielded a subset of 3,319,306 SNPs for use in subsequent analyses. We then reconstructed a SNP-based phylogeny of all 283 accessions, which revealed three groups, wild soybeans, landraces, and improved cultivars, clustered into separate clades, with a few landrace lines classified among the improved cultivar branches (Fig. 1b), suggesting a shared lineage for these accessions. Next, we performed principal component analysis (PCA) based on SNPs to better resolve the population structure of the three ecotype groups (Fig. 1c). The PCA plot showed that wild lines clustered independently from the non-wild accessions, while landraces formed a tight group that completely overlapped with improved lines. Linkage disequilibrium (LD) analysis showed that LD dropped to half of its maximum value at 106 kb for the resequenced population (Fig. 1d) but with variations among different populations. The LD in wild soybean was ~ 33 kb. In the non-wild accessions (i.e., landrace accessions and cultivars), LD increased to 120 kb, which was consistent with the findings of previous studies. Collectively, these analyses show that wild soybean populations carry distinct sets of SNPs compared to landraces and improved cultivars.

Selection signals during domestication and improvement and GO analysis

To identify potential selective signals during soybean domestication (wild soybeans versus landraces) and improvement (landraces versus improved cultivars), the genetic parameter population-differentiation statistic F_{ST} and nucleotide diversity ($\theta\pi$) were employed as indicators of selective signals. This analysis revealed a total of 78 domestication-selective sweeps (Fig. 2a and Supplementary Table 3) and 39 improvement-selective sweeps (Fig. 2b and Supplementary Table 4), covering 25.1 Mb (~ 2.56%) and 22.60 Mb (~ 2.31%) of the assembled Gmax_ZH13 reference

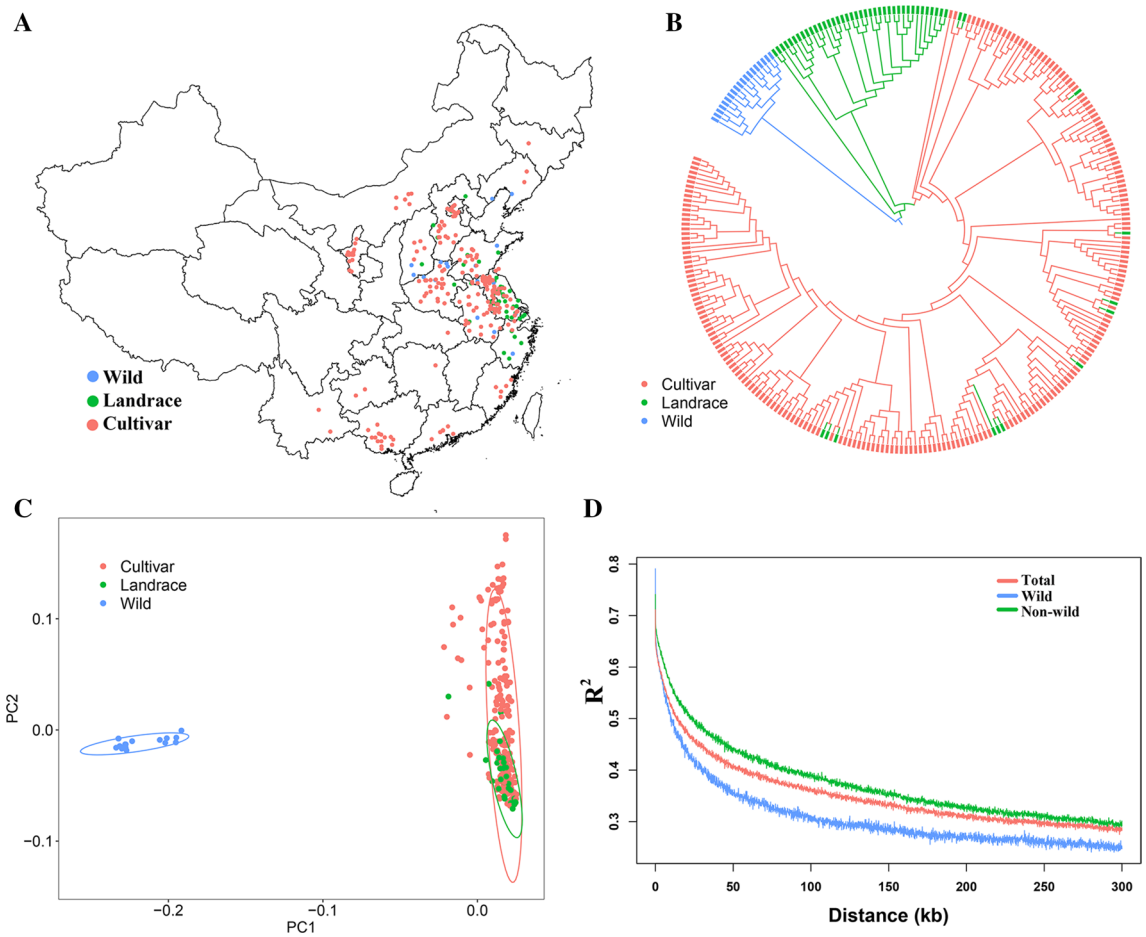

Fig. 1 Geographic distribution, population structure, and decay of linkage disequilibrium (LD) of 283 soybean accessions. **a** Geographical distribution of resequenced accessions. (ok, good to know where they were collected, but strongly consider adding the value of this information. this is part if your basic experimental design). **b** Maximum likelihood phylogenetic tree of 283 soybean accessions. **c** PCA plot showing clustering of soybean accessions into two major populations. **d** Genome-wide average Linkage Disequilibrium (LD) decay in all samples (orange), wild (blue), and non-wild accessions (green) (color figure online)

genome, respectively. The selection signals detected in the present study were almost all found in previously reported domestication-related QTL regions (Zhou et al. 2015).

A total of 689 and 1204 putative genes (Supplementary Tables 5 and 6) were identified in domestication-selective and improvement-selective sweeps, respectively. To annotate these candidate genes, BLASTp searches were performed in Arabidopsis and rice functional gene databases. The results showed that some of these genes were involved in domestication-related agronomic traits such as stress response, seed weight, plant height, and flowering time. The gene ontology enrichment analyses showed that the genes in domestication-selective sweeps were enriched in various biological processes, especially in abiotic stress responses including water deprivation, salt stress, and osmotic stress response. Besides abiotic stress response, protein modification, carbohydrate metabolic processes, and amino acid transport were also enriched in domestication-selective sweeps (Fig. 2c). In the improvement-selective regions, biological process, cellular component, and molecular functions were enriched for genes involved in protein transmembrane transport, regulation of seed germination, fatty acid metabolic/biosynthetic process, amino sugar catabolic process, amino acid biosynthetic/metabolic process, and abscisic acid biosynthetic/metabolic process (Fig. 2d), which are reported to be involved in the seed germination stage.

GWAS for color traits

We next sought to dissect the genetic mechanisms underlying specific domestication-related traits. For this purpose, we selected flower color, stem color, and seed coat color as phenotypic traits for evaluation by genome-wide associated study (GWAS), and subsequently generated a high-density map containing 2,597,425 SNPs from the landrace and cultivated accessions.

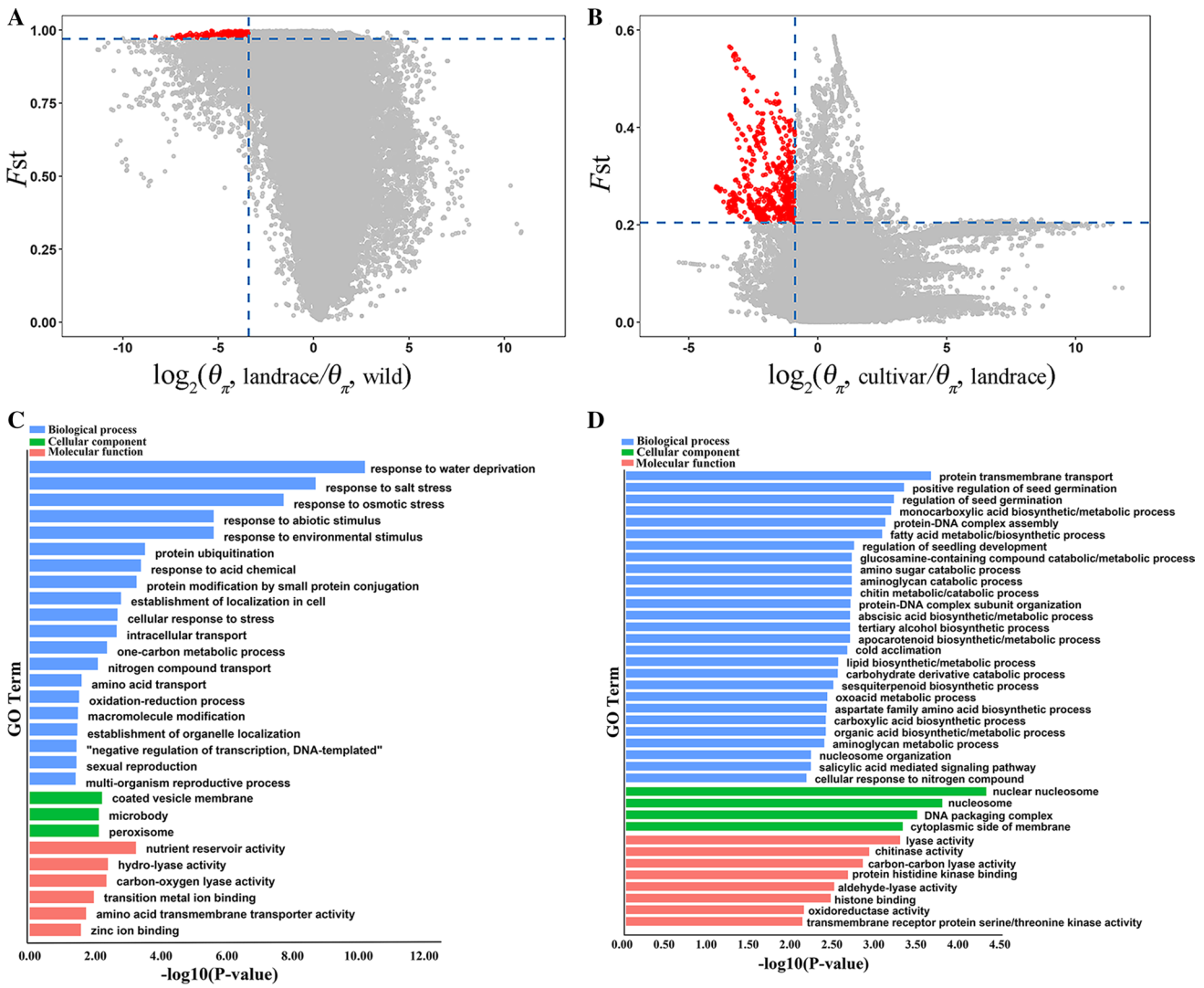

Fig. 2 Genome-wide selective sweep analysis for the resequenced population and GO enrichment analysis of candidate genes in selected regions. The horizontal and vertical lines represent threshold lines of the top 5% of the F_{st} and $\theta\pi$ ratio values. Points (red) located in the top left sector represent selective signatures for landrace versus wild (**a**) and cultivar versus landrace (**b**), respectively. **c** Enrichment GO terms in domestication-selective sweeps and improvement-selective sweeps (**d**). Blue, green, and red bars indicate biological process, cellular component, and molecular function, respectively (color figure online)

For flower color and stem color (purple or white), significant GWAS signals were detected for both traits at the *W1* locus on Chr. 13 (Fig. 3a, b), in which *GmF3′5′H* (*SoyZH13_13G05760*/*Glyma.13g072100*), encoding a flavonoid 3′, 5′-hydroxylase, was reported to control both the flower color and stem color (Zabala and Vodkin 2007). In this GWAS signal, the lead SNP S13_18801271 explained 43.53% of the phenotypic variation for flower color and the SNP S13_17908259 explained 41.35% of the phenotypic variation for stem color. In addition to the *W1* locus, we identified a new GWAS signal associated with flower color located on Chr. 03 with $-\log_{10}(p) > 8.1$ (Fig. 3b).

Soybean seed coat color varies significantly among the accessions used in the present population, including black, yellow, green, and brown. For these four colors, a GWAS signal at the *I* locus on Chr.08 was detected to corresponding to seed coat color variation. In addition to these previously characterized loci, we identified other new GWAS signals responsible for seed coat color on chromosomes 3, 10, 11, and 19 (Fig. 3c). Another GWAS was taken for seed coat color with green or yellow seed coats from our resequenced germplasms and a significant GWAS signal consisting of 116 SNPs for seed coat color was detected on chromosome 1 (Fig. 3d), among which the lead SNP S01_56082377 with $-\log_{10}(p) = 27.2$, explained 38.76% of the phenotypic variation for seed coat color. *SoyZH13_01G182000* (*GmG*/*Glyma.01g198500*) was proven to control the green seed coat in soybean (Wang et al. 2018). The above results

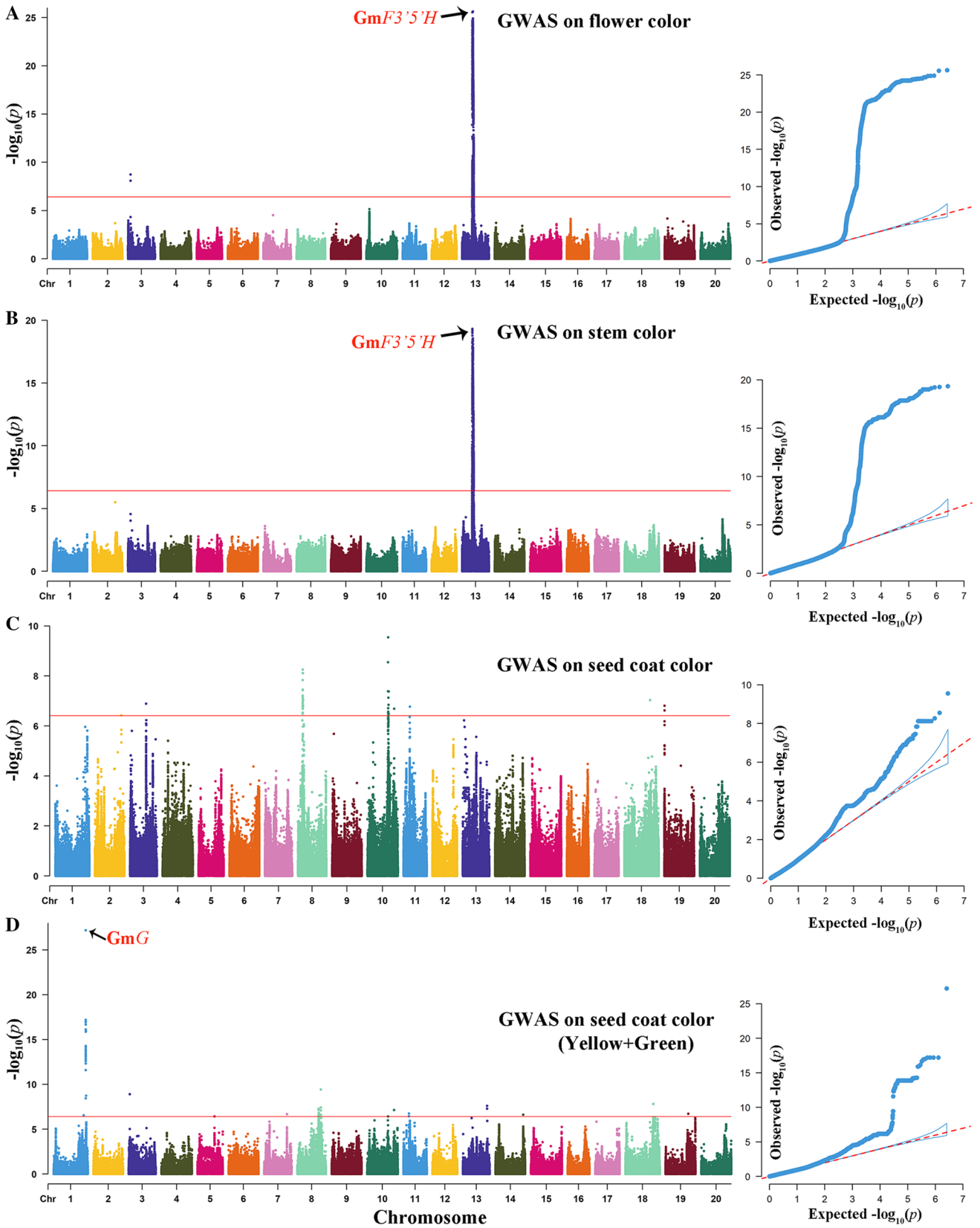

Fig. 3 Manhattan plots and quantile–quantile for the GWAS for color traits. **a** GWAS for flower color, **b** GWAS for stem color, **c** GWAS for seed coat color, **d** GWAS for seed coat color (yellow + green). The red line indicates the significance threshold ($-\log_{10}(P) = 6.4$) (color figure online)

prove that this set of high-density SNP markers is accurate and effective for gene mapping via GWAS.

qHSW-16 associated with seed weight was selected during soybean improvement

As a yield component, seed weight is a complex and agronomically important trait in soybean. We surveyed the 100-seed weight (HSW) among the resequenced landrace and cultivar population grown in Nanjing in 2018 (E1), 2019 (E2), and grown in Yancheng in 2019 (E3). Through GWAS with MLM model, we identified a total of four associations with 152 significant SNPs ($-\log_{10}(p) > 6.4$) for HSW located on chromosomes 8, 10, 11, and 20 (Fig. 4, Supplementary Table 6). Individual SNPs explained between 4.3 and 11.55% of the phenotypic variation. For convenience in further analysis, we grouped the significant trait-associated SNPs that were located in close proximity at LD $r^2 > 0.70$, and the lead SNP was used to represent the locus (Table 1). From group E3, *qHSW-8* was identified on chromosome 8, consisting of 131 significant SNPs, which showed a significant marker-trait association with $-\log_{10}(P)$ as high as 8.5

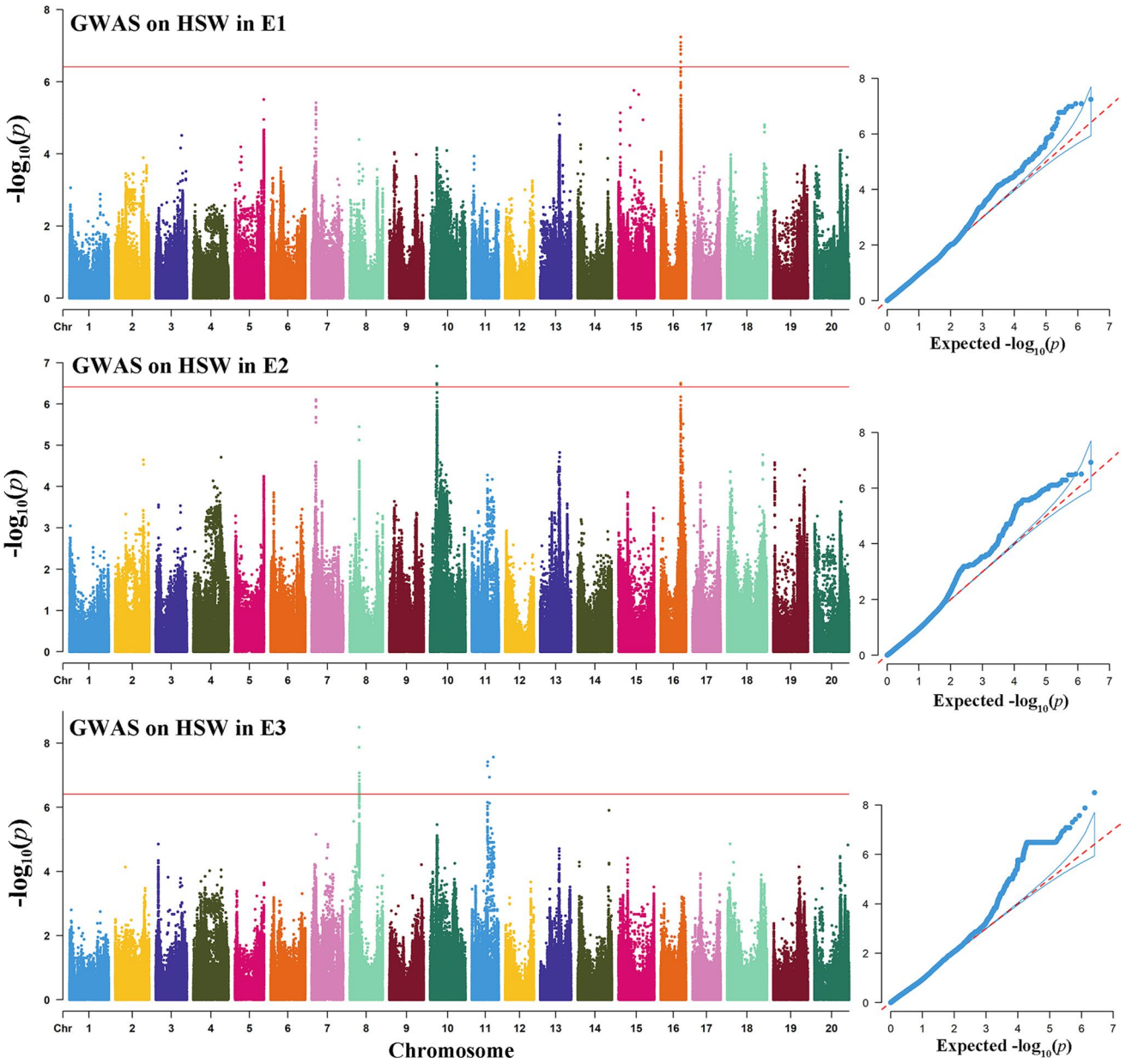

Fig.4 Manhattan plots and quantile–quantile for the GWAS for seed weight in E1 (**a**), E2 (**b**), and E3 (**c**). The red line indicates the significance threshold ($-\log_{10}(P) = 6.4$) (color figure online)

Table 1 Loci and SNPs significantly associated with seed weight, predicted candidate genes and previously reported QTLs for seed weight at similar genome regions

Loci	Chr	Position	Lead SNP	Alleles	$-\log_{10}(p)$	R^2	Known QTLs
qHSW-8	Gm8	14087453	S08_14087453	T/A	8.5	0.116	Seed weight 35-1 (Han et al. 2012) and 34-13 (Han et al. 2012)
qHSW-10	Gm10	10272204	S10_10272204	C/T	6.92	0.046	Seed weight 50-10 (Kato et al. 2014)
qHSW-11-1	Gm11	23895170	S11_23895170	C/T	7.42	0.099	Seed weight 10-3 (Specht et al. 2001) and 36-11 (Han et al. 2012)
qHSW-11-2	Gm11	32685975	S11_32685975	C/T	7.57	0.101	Seed weight 35-9 (Han et al. 2012) and 32-1 (Li and Zheng 2008)
qHSW-16	Gm16	30298930	S16_30298930	A/G	7.24	0.066	–

(Fig. 4c; Table 1). In E2, *qHSW-10* was detected as a significant QTL that explained 4.6% of the variation in HSW, while in E3, *qHSW-11-1* and *qHSW-11-2* loci on chromosome 11 explained 9.9% and 10.1% of variation for HSW, respectively. A highly significant SNP cluster on chromosome 16, *qHSW-16* (including 11 SNPs, $P = 5.72 \times 10^{-8}$) was detected in both E1 and E2, indicating that this QTL likely harbored a candidate gene responsible for HSW.

Commonly, wild soybeans have small seeds, whereas landraces and improved cultivars exhibit large seed sizes, suggesting that seed size was selected during domestication and improvement. In the present study, we found that an improvement-selective sweep region on Chr.16 overlapped with a significant GWAS signal, *qHSW-16* (Fig. 5a, b), thus indicating that seed weight was selected during soybean improvement. As shown, 11 SNPs were significantly associated with seed weight (Fig. 5c, Supplementary Table 7), and exhibited strong linkage disequilibrium ($r^2 > 0.8$) (Fig. 5d). The lead SNP S16_30298930 hereafter represents this significant QTL ($P = 5.72 \times 10^{-8}$). Among the landraces and cultivars used in this study, soybean accessions carrying the SNP S16_30298930-G allele exhibited significantly higher average seed weight than those carrying the S16_30298930-A allele (Fig. 5e). In addition, we observed that the frequency of S16_30298930-G allele in landraces is 38.5%, higher than 9.7% of cultivars (Fig. 5f), resulting in higher average seed weight for landraces than that of cultivars (Fig. 5g). Taken together, these results suggest that *qHSW-16* is significantly associated with seed weight and has undergone strong selection during improvement.

Identification and expression patterns of candidate genes in *qHSW-16*

Identification and utilization of candidate genes are one of the key objectives of GWAS. The LD decay distance is 120 kb on average across all chromosomes in the landrace and cultivar genomes (Fig. 1d). Using the annotation of the Gmax_ZH13 reference genome, we found 18 gene models within the LD decay distance of SNP S16_30298930 in *qHSW-16* (Table 2). To confirm which of genes related to seed weight, real-time quantitative PCR (qRT-PCR) was performed to analyze the expression patterns of the 18 candidate genes in NPS255 (large seed variety) and NPS270 (small seed variety). The results demonstrated that the expression levels of six of these 18 genes (*SoyZH13_16G122600*, *122700*, *122900*, *123100*, *123200,* and *124000*) were too low to be detected, while 11 genes showed no significant differences in transcriptional expression between the large seed and small seed accessions during seed development after flowering 7 to 21 days (Fig. 6). In fact, only *SoyZH13_16G122400* in NPS255 showed approximately 2.5-fold and 2.9-fold higher mRNA levels than that of NPS270 at 7 and 14 days after flowering (DAF), respectively (Fig. 6). Thus, these observations strongly suggest that *SoyZH13_16G122400* is most likely the causal gene involved in the regulation of seed weight in QTL *qHSW-16* detected in our study.

Discussion

Cultivated soybean was domesticated from wild soybean in China ~ 5000 years ago (Dashiell 2005). During domestication and artificial selection processes, various allelic variations of wild soybeans were lost, which led to much lower genetic diversity in landrace and cultivar accessions than that of their wild counterparts (Lam et al. 2010; Zhou et al. 2015). This reduced variation has potentially resulted in the loss of genes from soybean important for adaptation to different environments (Qi et al. 2014; Lu et al. 2020). In the present study, GO analysis showed that the most significant GO terms enriched with genes in domestication-selective sweep regions included "response to water deprivation," "response to salt stress," and "response to osmotic stress" (Fig. 2c), which suggested that soybean likely lost its adaptability to environmental stress during the domestication process from *G. soja* to landraces. Wild soybeans that exhibit high stress tolerance may therefore serve as a resource for germplasm improvement in the development of cultivars adapted to certain environmental conditions. Rapid and uniform seed germination is a crucial prerequisite for successful seedling establishment and high yields in soybean production (Caverzan et al. 2018). In the present study, genes

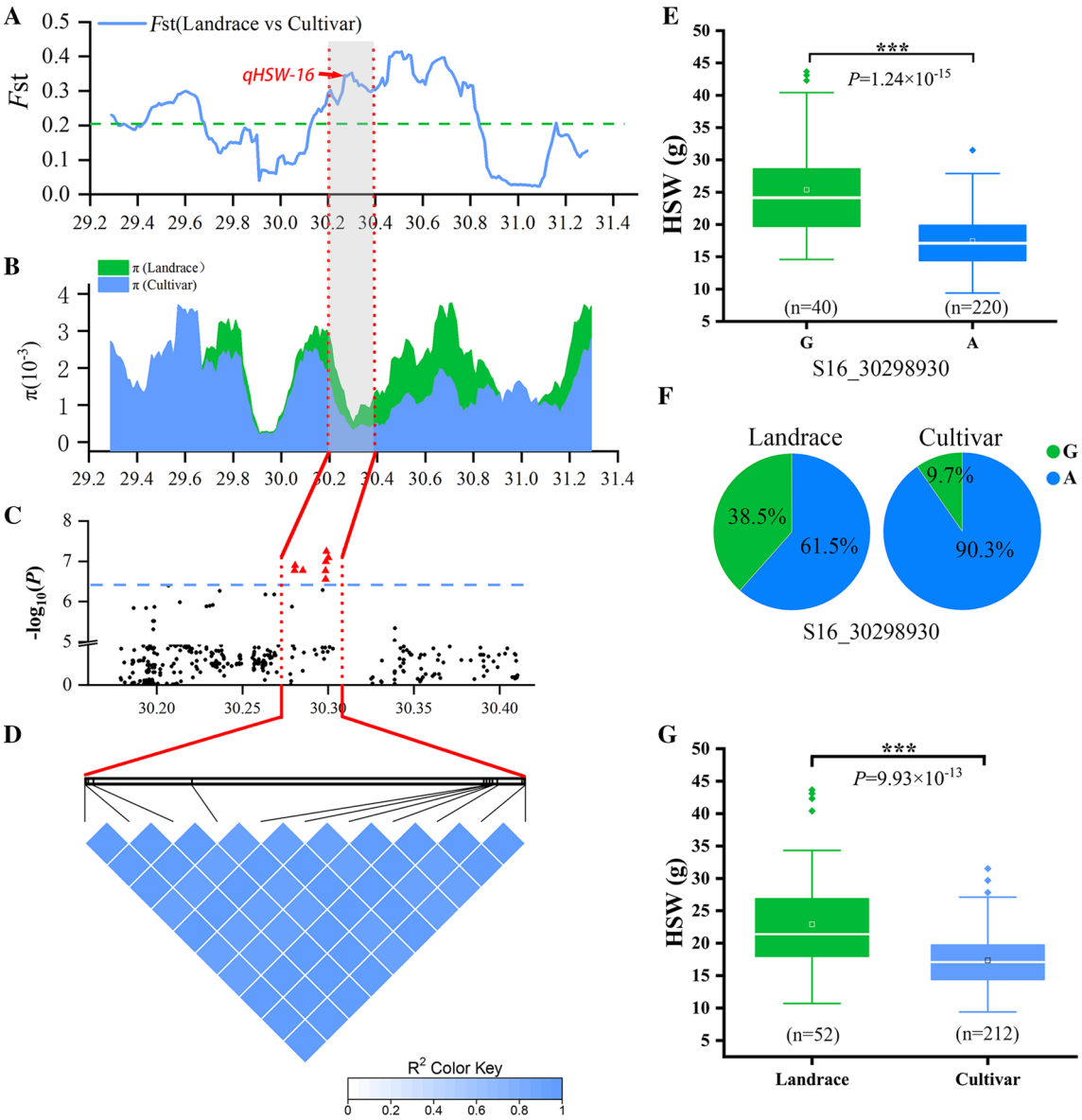

Fig. 5 **a** F_{ST} plot of landraces versus cultivars and **b** π values in landraces and improved cultivars across the 2 Mb genomic regions surrounding the lead SNP S16_30298930 of *qHSW-16*. **c** Significant GWAS signal for seed weight on chromosome 16. **d** Pairwise LD analysis between the 11 SNPs. The LD plot is represented by the inverted triangle. The LD level between 11 SNPs was indexed by the R^2 value. **e** Comparison of seed weight in accessions with different alleles of S16_30298930 by boxplot. **f** Allele frequencies of S16_30298930 in soybean landraces and cultivars. **g** Boxplot of soybean seed weight in landraces and cultivars

annotated for GO terms related to seed germination, such as "regulation of seed germination," "fatty acid metabolic/biosynthetic process," "amino sugar catabolic process," and "protein-DNA complex assembly" (Rajjou et al. 2012; Penfield 2017), were significantly enriched through the process of soybean improvement (Fig. 2d), indicating that seed germination traits have undergone strong selection during improvement from landrace to cultivar.

In sharp contrast with stress-associated traits, during soybean domestication and improvement, genetic variation increased for traits associated with seed protein, oil content, flowering, seed weight, and yield through both natural and artificial selection (Young et al. 2014; Lu et al. 2020; Miao et al. 2020). These selection processes left differential genomic footprints across germplasm groups. Selective signature analysis identifies the signatures left by positive selection that can be used to screen for candidate genes associated with specific traits (Nielsen et al. 2005). However, the selective signatures always span several or even dozens of kilobases, and thus it may not be possible to identify all candidate genes within a selective signature. GWAS using the GLM module takes into account both familial relatedness

Table 2 Candidate genes for seed weight in *qHSW-16*

Gene ID	Annotation
SoyZH13_16G122400	Anaphase-promoting complex 10
SoyZH13_16G122500	Unknown protein
SoyZH13_16G122600	U-box domain-containing protein 5-like
SoyZH13_16G122700	Uncharacterized protein LOC100779111
SoyZH13_16G122800	ZF-HD homeobox protein At4g24660-like
SoyZH13_16G122900	Protein kinase family protein
SoyZH13_16G123000	Sugar porter (SP) family MFS transporter
SoyZH13_16G123100	Uncharacterized mitochondrial protein
SoyZH13_16G123200	Sugar porter (SP) family MFS transporter
SoyZH13_16G123300	Alpha/beta-Hydrolases superfamily protein
SoyZH13_16G123400	Alpha/beta-Hydrolases superfamily protein;
SoyZH13_16G123500	GATA transcription factor 19
SoyZH13_16G123600	Maf-like protein
SoyZH13_16G123700	FAD/NAD(P)-binding oxidoreductase family protein
SoyZH13_16G123800	Golgi to vacuole transport-related protein
SoyZH13_16G123900	Pentatricopeptide repeat (PPR) superfamily protein
SoyZH13_16G124000	Tryptophan aminotransferase related 2
SoyZH13_16G124100	Aminopeptidase M1

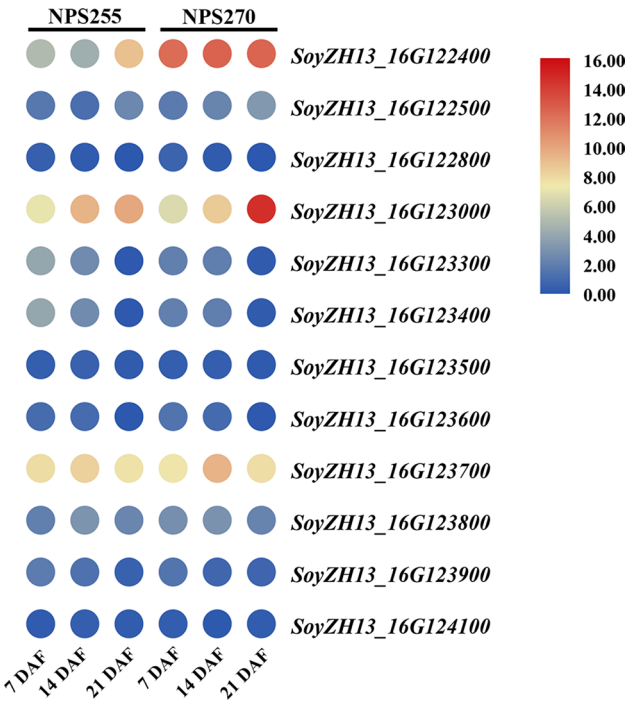

Fig. 6 Heat map of candidate genes expression levels in *qHSW-16*

and population structure and can thus serve as an informative tool to connect complex phenotypic traits with their underlying genetic factors (Zhang et al. 2019; Li et al. 2020). Although GWAS is a sensitive means of identifying candidate genes that have experienced selection, selective signature analysis is preferred due to its utility is circumventing the limitations of GWAS. In the present study, both GWAS and selective signature analysis were applied to identify 18 candidate genes associated with seed weight.

Seed weight, controlled by multiple genes, is a major component of seed yield in soybean. To improve yield, it is thus necessary to understand the genetic basis of seed weight through identification of candidate QTLs that significantly affect this trait. In soybean, over 200 QTLs for seed weight have been reported across 20 chromosomes (SoyBase, http://www.soybase.org/). In this study, 283 wild, landrace, and cultivar soybean accessions were collected from across a wide geographic range in China to dissect the genetic architecture controlling seed weight. Through GWAS with high-density SNP markers, 152 SNPs (five loci) were significantly associated with seed weight. Individual SNPs accounted for 4.3 to 11.55% of the phenotypic variation (Supplementary Table 7). Among these loci, four QTLs have been previously reported at least once (Table 1). *qHSW-10* on chromosome 10 were located within QTL, *seed weight 50-10* (Kato et al. 2014). The lead SNP of *qHSW-8*, S08_14087453 on chromosome 8, was located within two previously mapped QTLs, *seed weight 35-1* and *34-13* (Han et al. 2012). *qHSW-11-1* was located within QTL, *seed weight 10-3* and *36-11* (Specht et al. 2001; Han et al. 2012). *qHSW-11-2* was mapped within QTL, *seed weight 35-9*, and *32-1* (Li and Zheng 2008; Han et al. 2012). Interestingly, selective sweep analyses showed that these three identified loci were overlapped with selective-sweep regions during soybean improvement (Supplementary Table 4), suggesting that these three loci associated with seed weight

have undergone further selection from landrace to cultivar. The high repeatability of these loci across various environments and genetic backgrounds implies a great potential for marker-based breeding for seed weight in soybean. As seed weight plays an important role in soybean yield, these loci are also useful for soybean yield improvement. Among these loci, the *qHSW-16* locus at 30.3 Mb position on chromosome 16 was identified in two environments exhibiting a strong GWAS signal.

To date, some of these previously described QTLs were mapped to chromosome 16 in soybean (Maughan et al. 1996; Mian et al. 1996; Han et al. 2012; Hu et al. 2014, 2020), but no QTL has yet been reported in the *qHSW-16* genomic region of chromosome 16. Among these reported QTLs or SNPs, *seed weight 34-18* was identified on chromosome 16 from 25.1 to 26.4 Mb (Han et al., 2012), approximately 3.6 Mb away from *qHSW-16*. *Seed weight 2-6* and *Seed weight 4-4* were about 25 Mb away from *qHSW-16* (Maughan et al. 1996; Mian et al. 1996). These results suggested that *qHSW-16* is a novel QTL regulating seed weight in soybean. Further study of the interaction between *qHSW-16* and other seed weight-related QTLs will greatly improve our understanding of the genetic mechanisms governing seed weight in soybean. For this novel QTL, *qHSW-16, SoyZH13_16G122400* showed higher transcription levels in a large seed variety than that in a small seed variety during seed development, suggesting that *SoyZH13_16G122400* is the most likely causal gene for regulation of seed weight. Bioinformatics and sequence homology analysis revealed that *SoyZH13_16G122400* encodes anaphase-promoting complex 10 (*GmAPC10*), a multi-subunit E3 ubiquitin ligase, which is an anaphase-promoting complex/cyclosome (APC/C) that has been well studied in model species such as Arabidopsis and rice for its major contributions to cell-cycle regulation. In Arabidopsis, *AtAPC10* is essential for cell proliferation during leaf development (Eloy et al. 2011). *AtAPC8* is required for male meiosis (Xu et al. 2019), deficiency of *AtAPC8* showed abnormal development of the meristem, leaves and shoots, as well as reduced fertility with short siliques and mature pollen with no or single sperm-like cells (Zheng et al. 2011). In rice (*Oryza sativa* L.), OsTAD1 and OsMOC1 interact to form a complex together with OsAPC10 that regulates rice tillering(Lin et al. 2012; Xu et al. 2012). In tobacco (*Nicotiana tabacum*), overexpression of *AtAPC10* promotes high biomass accumulation, including fresh and dry weight, root length, and number of seeds per plant (Lima et al. 2013). The influence of anaphase-promoting complex 10 on seed weight for *GmAPC10* awaits further study. Our results may support the cloning of a gene influencing seed weight and the development of functional markers for marker-assisted selection in soybean breeding.

Materials and methods

Plant materials and phenotypic evaluation

A total of 283 soybean accessions, including 19 wild accessions, 52 landraces, and 212 improved cultivars from China were selected to construct an association mapping panel. All materials were planted in Nanjing City, Jiangsu Province in 2018 and 2019, and Yancheng City, Jiangsu Province in 2019. The experiment followed a randomized complete block design with a single row plot and three replications. For description purposes, the three environments, 2018 Nanjing, 2019 Nanjing, and 2019 Yancheng were designated as E1, E2, and E3, respectively.

The flower color, stem color, and seed coat color of each accession were surveyed in Nanjing in 2018. For 100-seed weight, the seeds were dried in an air dryer and a sample of 100 cleaned seeds in E1, E2, and E3 was randomly taken and weighed with three replications.

DNA extraction and whole-genome resequencing

Young leaves were collected 4 weeks after planting and quickly frozen in liquid nitrogen for sequencing. Total DNA was extracted using the CTAB method (Paterson et al. 1993). Paired-end sequencing libraries (150 bp × 2) were constructed for all the 283 lines following the manufacturer's instructions (Illumina). The soybean reference genome Gmax_ZH13 (Shen et al. 2018) and its annotation were downloaded online (https://bigd.big.ac.cn/gwh/Assembly/125/show).

Identification of variation and filtering

All paired-end sequence reads were mapped to the Gmax_ZH13 reference genome using BWA software(Li and Durbin 2009) with default parameters. Only reads with unique mapping position in the Gmax_ZH13 reference genome and mapping quality value greater 30 were retained in BAM format by SAMtools (Li et al. 2009). Additionally, we improved the alignment performance by realignment of reads around Indels from the BWA mapping results with the IdnelRealigner package in the Genome Analysis Toolkit (GATK) (McKenna et al. 2010). High-quality SNPs and Indel variations were obtained according to the following criteria. (a) Only concordant sites identified by GATK and VCFtools (Danecek et al. 2011) with the Select Variants packages were retained. (b) The SNP quality value should be greater than 30. (c) SNPs and Indels with a MAF < 5% and missing rate < 10% were discarded. (d) The average sequencing depth was greater than 8×. (f) Insertions and deletions with

a maximum length 10 bp were taken into account. The annotation information of variants was obtained by ANNOVAR (Wang et al. 2010).

Population structure

The phylogenetic tree was constructed with maximum likelihood using SNPhylo based on the high-density SNPs (Lee et al. 2014) and visualized with iTOL (Letunic and Bork 2019). LD was calculated for wild accessions and non-wild accessions by software PopLDdecay (Chi et al. 2019). PCA was performed using GCAT (Yang et al. 2011), and two-dimensional coordinates were plotted for the 283 soybean accessions in R (www.r-project.org).

Selective sweep analyses

The 3,319,306 high-quality SNPs were used for selective sweep analysis. The fixation index (F_{ST}) between landraces and wild accessions, landraces and cultivars, and the nucleotide diversity (π) were analyzed using the VCFtools package (Danecek et al. 2011) with a step size of 10 kb and a 100-kb sliding window. The top 5% values of $\theta\pi$ ratio were adopted to identify putative selective regions. Likewise, the top 5% values of F_{ST} were applied to confirm highly differentiated regions. The intersections of windows based on F_{ST} and $\theta\pi$ ratio were assigned as potential selective regions. The thresholds of $F_{ST} \geq 0.970$ and $\theta\pi$ ratio < 0.0947 for wild_vs_landrace, and $F_{ST} \geq 0.204$ and $\theta\pi$ ratio < 0.545 for landrace_vs_cultivar were used to identify selective sweeps.

Go enrichment analysis for candidate genes

To identify biological functions of candidate genes associated with the selection, annotation and enrichment analysis were performed by submitting gene information to Gene Ontology (GO) using the software TBtools (Chen et al. 2020). GO terms with P value < 0.01 were taken as those in which candidate genes were significantly enriched.

Genome-wide association study (GWAS) and identification of candidate gene

Genome-wide association study of the landrace panel and the cultivated panel was conducted with 2,597,425 SNPs. To minimize false positives and increase statistical power, a mixed linear model (MLM) accounting for population structure and kinship were implemented in the Genomic Association and Prediction Integrated Tool (GAPIT) R package. (Lipka et al. 2012). The threshold for a significant association was set to 1/n (n is the number of markers, $P < 3.85 \times 10^{-7}$ or $-\log_{10}(P) > 6.4$). The LD heatmaps with surrounding peaks in the GWAS results were visualized using the R package of "LDheatmap" (Shin et al. 2006).

To identify potential candidate genes for seed weight, genes that were located within the LD decay distance upstream and downstream of peak SNPs (the most significant SNPs with a maximum of $-\log_{10} P$ values) were identified.

Quantitative RT-PCR for candidate genes

Two soybean varieties with contrasting seed size were used in this study. The large seed variety NPS255 is an elite variety with a HSW of 29.2 g carrying the S16_30298930-G allele, whereas the small seed variety NPS270 with a HSW of 13.2 g, carries the S16_30298930-A allele. The sampling points of early seed maturation were at 7, 14, and 21 days after flowering (DAF). After sampling, the tissues were quickly frozen in liquid nitrogen and stored at − 70 ℃ until RNA isolation. Three biological replicates were used for each of the sampling points.

Total RNA was isolated using the RNA simple Total RNA Kit (TIANGEN Beijing, China), and first-strand cDNA was reverse-transcribed using a TaKaRa Primer Script RT reagent kit with gDNA Eraser. Gene expression was determined by RT-PCR using an ABI 7500 system (Applied Biosystems, Foster City, CA, USA) with the SYBR Green Real-time Master Mix (Toyobo), and the data were analyzed using ABI 7500 Real-Time PCR System. The relative expression level against the Actin11 gene was quantified using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen 2001). Three replicates were run for each sample.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-021-03774-6.Acknowledgements This work was supported by the National Key Research and Development Program of China (2018YFE0112200), the Key R&D project of Jiangsu Province (BE2019376).

Author Contribution statement ZW, WX, HZ, and SL contributed to field design and phenotypic data collection; ZW and XC performed phenotypic analysis; WX, LS, XL, YZ, and CX assisted in revising the manuscript; ZW analyzed the experimental results; ZW and HT wrote the manuscript. All authors reviewed the manuscript and provided suggestions.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Availability of data and materials Data provided in supplementary files.

References

Caverzan A, Giacomin R, Müller M, Biazus C, Lângaro NC, Chavarria G (2018) How does seed vigor affect soybean yield components? Agron J 110:1318–1327

Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R (2020) TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant 13:1194–1202

Chi Z, Shan-Shan D, Jun-Yang X, Wei-Ming H, Tie-Lin Y (2019) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 10:1786–1788

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Dashiell K (2005) SOYBEANS: improvement, production, and uses. Third Edition: Boerma, H.R., Specht, J.E. (Eds), American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, Wisconsin, USA, 2004, 1144 pp. Price: US$155.00 (hardback). ISBN 0-89118-154-7. Agric Syst 83:110–111

Eloy NB, de Freitas Lima M, Van Damme D, Vanhaeren H, Gonzalez N, De Milde L, Hemerly AS, Beemster GT, Inzé D, Ferreira PC (2011) The APC/C subunit 10 plays an essential role in cell proliferation during leaf development. Plant J 68:351–363

Han Y, Li D, Zhu D, Li H, Li X, Teng W, Li W (2012) QTL analysis of soybean seed weight across multi-genetic backgrounds and environments. Theor Appl Genet 125:671–683

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

Hu D, Zhang H, Du Q, Hu Z, Yang Z, Li X, Wang J, Huang F, Yu D, Wang H, Kan G (2020) Genetic dissection of yield-related traits via genome-wide association analysis across multiple environments in wild soybean (*Glycine soja* Sieb. and Zucc.). Planta 251:39

Hwang EY, Song Q, Jia G et al (2014) A genome-wide association study of seed protein and oil content in soybean[J]. BMC Genom 15(1):1–12

Kato S, Sayama T, Fujii K, Yumoto S, Kono Y, Hwang T, 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 127:1365–1374

Lam H, Xu X, Liu X, Chen W, Yang G, Wong F, Li M, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SS, Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059

Lee G, Crawford GW, Liu L, Sasaki Y, Chen X (2011) Archaeological soybean (*Glycine max*) in East Asia: does size matter? PLoS ONE 6:e26720

Lee T, Guo H, Wang X, Kim C, Paterson AH (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15:162

Letunic I, Bork P (2019) Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760

Li W, Zheng D (2008) QTL mapping for major agronomic traits across 2 years in soybean. J Crop Sci Biotechnol 11(3):171–190

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome PDPS (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li YH, Li D, Jiao YQ, Schnable JC, Li YF, Li HH, Chen HZ, Hong HL, Zhang T, Liu B (2020) Identification of loci controlling adaptation in Chinese soya bean landraces via a combination of conventional and bioclimatic GWAS. Plant Biotechnol J 18:389–401

Liang H, Xu L, Yu Y, Yang H, Dong W, Zhang H (2016) Identification of QTLs with main, epistatic and QTL by environment interaction effects for seed shape and hundred-seed weight in soybean across multiple years. J Genet 95:475–477

Lima MDF, Eloy NB, Bottino MC, Hemerly AS, Ferreira PCG (2013) Overexpression of the anaphase-promoting complex (APC) genes in *Nicotiana tabacum* promotes increasing biomass accumulation. Mol Biol Rep 40:7093–7102

Lin Q, Wang D, Dong H, Gu S, Cheng Z, Gong J, Qin R, Jiang L, Li G, Wang JL, Wu F, Guo X, Zhang X, Lei C, Wang H, Wan J (2012) Rice APC/C-TE controls tillering by mediating the degradation of MONOCULM 1. Nat Commun 3:752

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta}$CT method. Methods 25:402–408

Lu S, Dong L, Fang C, Liu S, Kong L, Cheng Q, Chen L, Su T, Nan H, Zhang D, Zhang L, Wang Z, Yang Y, Yu D, Liu X, Yang Q, Lin X, Tang Y, Zhao X, Yang X, Tian C, Xie Q, Li X, Yuan X, Tian Z, Liu B, Weller JL, Kong F (2020) Stepwise selection on homeologous PRR genes controlling flowering and maturity during soybean domestication. Nat Genet 52:428–436

Maughan PJ, Maroof MAS, Buss GR (1996) Molecular-marker analysis of seed-weight: genomic locations, gene action, and evidence for orthologous evolution among three legume species. Theor Appl Genet 93:574–579

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

Mian MAR, Bailey MA, Tamulonis JP, Shipe ER, Carter TE, Parrott WA, Ashley DA, Hussey RS, Boerma HR (1996) Molecular markers associated with seed weight in two soybean populations. Theor Appl Genet 93:1011–1016

Miao L, Yang S, Zhang K, He J, Wu C, Ren Y, Gai J, Li Y (2020) Natural variation and selection in *GmSWEET39* affect soybean seed oil content. New Phytol 225:1651–1666

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15:1566–1575

Paterson AH, Brubaker CL, Wendel JF (1993) A rapid method for extraction of cotton (Gossypium spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol Biol Rep 11:122–127

Penfield S (2017) Seed dormancy and germination. Curr Biol 27:R874–R878

Qi X, Li M, Xie M, Liu X, Ni M, Shao G, Song C, Kay-Yuen Yim A, Tao Y, Wong F, Isobe S, Wong C, Wong K, Xu C, Li C, Wang Y, Guan R, Sun F, Fan G, Xiao Z, Zhou F, Phang T, Liu X, Tong S, Chan T, Yiu S, Tabata S, Wang J, Xu X, Lam H (2014) Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat Commun 5:4340

Rajjou L, Duval M, Gallardo K, Catusse J, Bally J, Job C, Job D (2012) Seed germination and vigor. Annu Rev Plant Biol 63:507–533

Shen Y, Liu J, Geng H, Zhang J, Liu Y, Zhang H, Xing S, Du J, Ma S, Tian Z (2018) De novo assembly of a Chinese soybean genome. Sci China Life Sci 61:871–884

Shin J, Blay S, McNeney B, Graham J (2006) LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single. J Stat Softw 16:1–10

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 41:493–509

Tengfei Z, Tingting W, Liwei W, Bingjun J, Caixin Z, Shan Y, Wensheng H, Cunxiang W, Tianfu H, Shi S (2019) A Combined linkage and GWAS analysis identifies QTLs linked to soybean seed protein and oil content. Int J Mol Sci 20:5915

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164

Wang M, Li W, Fang C, Xu F, Liu Y, Wang Z, Yang R, Zhang M, Liu S, Lu S, Lin T, Tang J, Wang Y, Wang H, Lin H, Zhu B, Chen M, Kong F, Liu B, Zeng D, Jackson SA, Chu C, Tian Z (2018) Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat Genet 50:1435

Wu D, Zhan Y, Sun Q, Xu L, Lian M, Zhao X, Han Y, Li W (2018) Identification of quantitative trait loci underlying soybean (*Glycine max* [L.] Merr.) seed weight including main, epistatic and QTL × environment effects in different regions of Northeast China. Plant Breeding 137:194–202

Xu C, Wang Y, Yu Y, Duan J, Liao Z, Xiong G, Meng X, Liu G, Qian Q, Li J (2012) Degradation of MONOCULM 1 by APC/CTAD1 regulates rice tillering. Nat Commun 3:750

Xu R, Xu J, Wang L, Niu B, Copenhaver GP, Ma H, Zheng B, Wang Y (2019) The Arabidopsis anaphase-promoting complex/cyclosome subunit 8 is required for male meiosis. New Phytol 224:229–241

Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82

Zabala G, Vodkin LO (2007) A rearrangement resulting in small tandem repeats in the F3′5′H gene of white flower genotypes is associated with the Soybean W1 Locus. CROP SCI 47:113–124

Zhang W, Liao X, Cui Y, Ma W, Zhang X, Du H, Ma Y, Ning L, Wang H, Huang F, Yang H, Kan G, Yu D (2019) A cation diffusion facilitator, GmCDF1, negatively regulates salt tolerance in soybean. PLoS Genet 15:e1007798

Zheng B, Chen X, McCormick S (2011) Zheng B, Chen X, McCormick S. The anaphase-promoting complex is a dual integrator that regulates both MicroRNA-mediated transcriptional regulation of cyclin B1 and degradation of Cyclin B1 during Arabidopsis male gametophyte development. Plant Cell 23:1033–1046

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, Fang C, Shen Y, Liu T, Li C, Li Q, Wu M, Wang M, Wu Y, Dong Y, Wan W, Wang X, Ding Z, Gao Y, Xiang H, Zhu B, Lee S, Wang W, Tian Z (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33:125–408