Genome Biology

RESEARCH

Open Access

Transcriptional regulatory network reveals key transcription factors for regulating agronomic traits in soybean

Wu Jiao[1†], Mangmang Wang[1†], Yijian Guan[1†], Wei Guo[2†], Chang Zhang[1], Yuanchun Wei[1], Zhenwei Zhao[1], Hongyu Ma[1], Longfei Wang[1], Xinyu Jiang[1], Wenxue Ye[1], Dong Cao[2*] and Qingxin Song[1*]

[†]Wu Jiao, Mangmang Wang, Yijian Guan, and Wei Guo contributed equally to this work.

*Correspondence: caodong@caas.cn; qxsong@njau.edu.cn

[1] State Key Laboratory of Crop Genetics and Germplasm Enhancement and Utilization, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, No. 1 Weigang, Nanjing, Jiangsu 210095, China
[2] Key Laboratory of Biology and Genetic Improvement of Oil Crops, Ministry of Agriculture, Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, Hubei 430062, China

Abstract

Background: Transcription factors (TFs) bind regulatory genomic regions to orchestrate spatio-temporal expression of target genes. Global dissection of the cistrome is critical for elucidating transcriptional networks underlying complex agronomic traits in crops.

Results: Here, we generate a comprehensive genome-wide binding map for 148 TFs using DNA affinity purification sequencing in soybean. We find TF binding sites (TFBSs) exhibit elevated chromatin accessibility and contain more rare alleles than other genomic regions. Intriguingly, the methylation variations at TFBSs partially contribute to expression bias among whole genome duplication paralogs. Furthermore, we construct a soybean gene regulatory network (SoyGRN) by integrating TF-target interactions with diverse datasets encompassing gene expression, TFBS motifs, chromatin accessibility, and evolutionarily conserved regulation. SoyGRN comprises 2.44 million genome-wide interactions among 3188 TFs and 51,665 target genes. We successfully identify key TFs governing seed coat color and oil content and prioritize candidate genes within quantitative trait loci associated with various agronomic traits using SoyGRN. To accelerate utilization of SoyGRN, we develop an interactive webserver (www.soytfbase.cn) for soybean community to explore functional TFs involved in trait regulation.

Conclusions: Overall, our study unravels intricate landscape of TF-target interactions in soybean and provides a valuable resource for dissecting key regulators for control of agronomic traits to accelerate soybean improvement.

Keywords: Transcription factors, Gene expression, Gene regulatory network, Soybean

Background

The intricate interplay between TFs and *cis*-acting regulatory elements (CREs) is pivotal in enabling plants to precisely modulate gene expression in a spatio-temporal manner for appropriate development and responses to the environment [1, 2]. It is now widely

believed that changes to gene expression provide important foundation for phenotypic diversity during plant evolution [3]. Novel gene expression patterns may emerge through a spectrum of mechanisms, encompassing post-translational modifications to TFs and mutations within TF binding sites (TFBSs) [4]. It is noteworthy that individual TFs are likely to exert regulatory control over multiple target genes, while individual genes may undergo cooperative regulation by multiple TFs or differential regulation by distinct TFs under different conditions. Deciphering TF-gene interactions in complex gene regulatory networks (GRNs) is an important step toward comprehending the regulatory cascades underpinning complex traits [5]. Large-scale investigations of TF-gene interactions have been conducted in both animal and plants [6–11], employing diverse methodological approaches such as DNA affinity purification sequencing (DAP-seq) and chromatin immunoprecipitation followed by sequencing (ChIP-Seq). These studies have not only unveiled the characteristics of TFBSs within genomes, but have also proffered insights into the putative functionalities of a diverse array of TFs. For instance, a substantial fraction of TFBSs has been ascertained to originate from expansions of particular transposable elements (TEs) during wheat evolution [10, 11].

Soybean (*Glycine max*) is one of the most economically important legume crops, providing a significant source of oil and protein for humans and livestock [12]. Commodity soybean contains approximately 20% oil with a favorable proportion of linoleic acid, which represents more than 50% of total oil crop production in the world (www.statista.com). The modern cultivated soybean was domesticated from wild soybean (*Glycine soja*) in China 6000–9000 years ago [13], through dramatic changes in morphological and physiological traits, including loss of pod shattering and increased oil content. TFs play a crucial role in regulation of important agronomic traits during soybean domestication and improvement, including *GmWRI1a/b* and *GmZF351* for oil content [14, 15] and *GmE1* and *GmFT2a* for photoperiodic flowering [16, 17]. Nonetheless, it is imperative to note that the comprehensive delineation of the binding landscape and target genes has thus far been conducted for limited TFs in soybean [18, 19]. The absence of an integrative regulatory network impedes systematic exploration and functional elucidation of TFs that govern key agronomic traits in soybean.

In this study, we profiled binding landscapes of 148 TFs using DAP-seq and explored the effect of TF binding divergence on expression difference in soybean. Furthermore, we constructed a comprehensive gene regulation network in soybean (SoyGRN) by integrating TF binding information derived from DAP-seq, TF motif database, chromatin accessibility, and transcriptome data of various tissues. We demonstrated the utility of SoyGRN as an instrumental tool for pinpointing key TF regulators governing agronomic traits and prioritizing causal TFs underlying QTLs. The established online platform based on SoyGRN provides a valuable resource for unraveling TF-gene interactions in soybean and will significantly expedite the discovery of functional TFs for soybean breeding.

Results

DAP-seq profiling of TFs in soybean

To investigate the transcriptional regulatory network in soybean, we cloned 230 TFs with potential important roles in development, abiotic and biotic stress response,

and nutrient utilization, and then profiled the genome-wide binding patterns of these TFs using DAP-seq (Fig. 1a and Additional file 1: Table S1). After removal of the low confidence TFs with Fraction of Reads in Peaks (FRiP) < 2%, 148 TFs from 28 families were retained for subsequent analyses (Fig. 1a, b). We further generated biological replicate data for 11 randomly selected TFs from diverse families to validate the repeatability of DAP-seq datasets. We observed more than 50% of DAP-seq peaks passing Irreproducible Discovery Rate (IDR) cutoff (0.05) for 9 (82%) TFs (Additional file 2: Fig. S1a). In addition, Pearson's correlation coefficient for 9 (82%) TFs were above 0.9 (Additional file 2: Fig. S1b), suggesting a high repeatability of the DAP-seq data. The DAP-seq success rates exhibited discernible bias among TF families, which was also reported in Arabidopsis and wheat [9, 10]. For example, high and low success rates were observed in ERF and bHLH TF families, respectively (Additional file 1: Table S2). To evaluate the reliability of DAP-seq, we compared DAP-seq and published ChIP-seq data for GmbZIP67 [18]. There was a significant overlap between ChIP-seq and DAP-seq peaks (hypergeometric test, $P = 0$) (Fig. 1c) and the similar binding motifs were also detected (Fig. 1d), which indicates high reliability of TFBS identification by DAP-seq in this study. The binding motifs of all TFs were listed in Additional file 1: Tables S3 and S4.

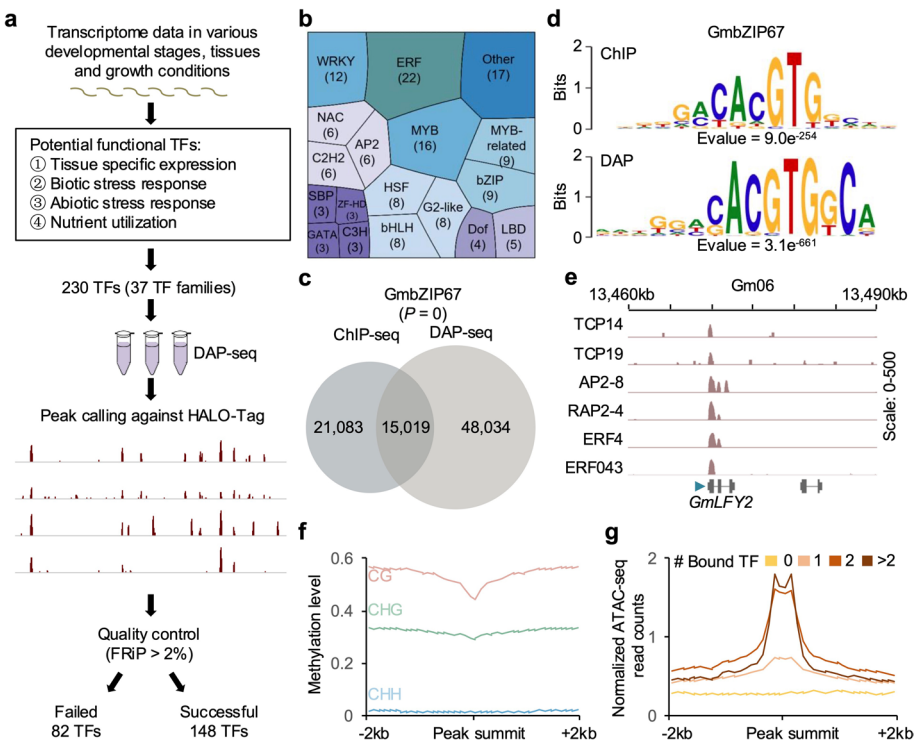

Fig. 1 Global identification of TFBSs by DAP-seq. **a** Experimental setup and quality control for DAP-seq. **b** Tree map showing the number of TFs in different families. **c** The overlap of 1-kb bins with peaks derived from ChIP-seq and DAP-seq of *GmbZIP67* across the genome. The statistic analysis was performed by hypergeometric test. **d** The similar TFBS motifs derived from ChIP-seq and DAP-seq of *GmbZIP67*. The *E* value was calculated by MEME-ChIP software. **e** Genomic tracks illustrating the promoter binding of *GmLHY2* by a subset of TFs. Triangle indicates the direction of gene transcription. **f** The average DNA methylation levels around the peak of TFs. **g** Fold enrichment of ATAC-seq reads around TFBSs bound by different numbers of TFs

A total of 3,041,762 binding peaks were identified for the 148 TFs and the number of peaks varied among TFs, with a median value around 5737 peaks per TF (Fig. 1e and Additional file 1: Table S1). Comparing these DAP-seq derived motifs to curated motif databases JASPAR, we found that the DAP-seq motifs of 65 TFs were highly consistent with published motifs in the corresponding families (Additional file 1: Table S2). In addition, noncanonical representative motifs were identified for another 58 TFs (Additional file 1: Table S2). The most enriched motifs present in the 65 TFs with canonical motifs were grouped based on their sequence similarity (Additional file 2: Fig. S2). As expected, the majority of motif sequences belonging to TFs from the same family exhibited significant similarity and clustered together, with only minor exceptions observed. Furthermore, highly similar genome-wide binding patterns were observed among TFs within the ERF, WRKY, HSF, B3, and TCP families (Additional file 2: Fig. S3). Conversely, TFs from the MYB, Dof, NAC, MYB-related, G2-like, ZF-HD, and HD-ZIP families exhibited a limited extent of shared genome-wide binding profiles (Additional file 2: Fig. S3). Compared with flanking sequences, TFBSs showed lower CG and CHG methylation levels (Fig. 1f) and much higher levels of chromatin accessibility (Fig. 1g), suggesting that epigenetic signatures may play a role in regulating TF binding, while the binding of TFs could also influence these epigenetic signatures.

Genome-wide binding of TFs in soybean genome

To summarize the landscape of TFBSs, we merged all binding peaks from the 148 TFs into non-overlapping 2-kb windows and performed principal component analysis (PCA) based on the presence or absence of each TF at each genomic segment. PCA analysis captured global TF binding patterns with principal component 1 (PC1) explaining 18.3% of the variance, correlating strongly with the TF numbers bound to a given genomic region (Fig. 2a and Additional file 2: Fig. S4). The genomic segments with more than 74 TFs (half of all assayed TFs) within a 2-kb region were defined as TF HOT regions ($n = 823$). As expected, those TF HOT regions showed more overlap with open chromatin regions (OCRs) than other genomic regions (Fig. 2b) and were significantly enriched in genomic regions with bivalent histone marks (H3K4me3 and H3K27me3) (hypergeometric test, $P = 0$) (Fig. 2c).

TF binding at gene promoters plays critical roles in regulating their expression. An approximately normal distribution was observed in the number of genes bound by different number of TFs (Kolmogorov–Smirnov test, $P = 0.94$) (Fig. 2d and Additional file 1: Table S5). The major group (23,382, 44.2%) of genes were bound by one to five TFs, whereas only 586 (1.1%) genes were bound by more than 20 TFs (Fig. 2d). The percentages of genes bound by two TFs from the same families were significantly higher than those of random TF pairs (Wilcoxon rank sum test, $P = 2.3e^{-8}$) (Additional file 2: Fig. S5), suggesting a preferential targeting of the same genes for TFs from the same families.

We found higher density of TF binding sites in flanking regions of transcription start site (TSS), and almost no enrichment of TF binding sites in flanking regions of transcription terminal site (TTS) (Fig. 2e), which suggests an uneven distribution of TF binding sites in gene and its flanking regions. Consistent with human studies reporting that HOT regions of TF binding are associated with housekeeping genes and higher TF counts correlate with elevated gene expression [7], the genes with more TFs bound in promoters

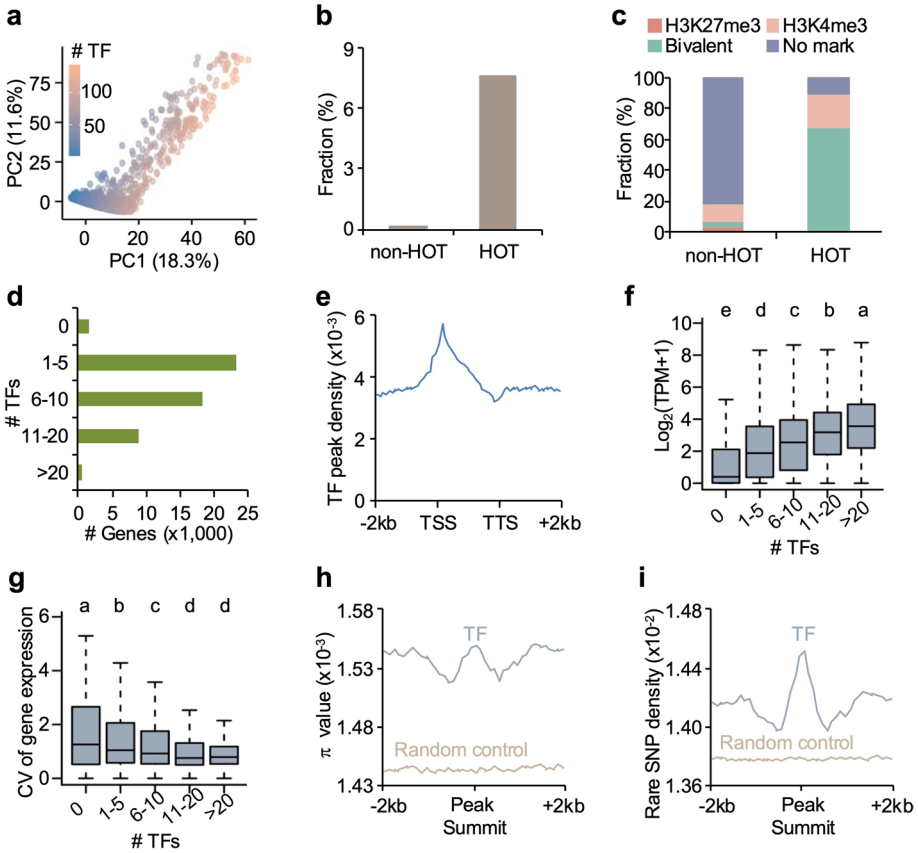

Fig. 2 Genome-wide atlas of TFBSs in soybean genome. **a** PCA of genomic segments bound by more than one TF. **b** Fractions of HOT and non-HOT regions overlapping with OCRs. **c** Fractions of HOT and non-HOT regions marked with single and bivalent (H3K4me3 and H3K27me3) histone marks or without histone marks. **d** Number of genes bound by different numbers of TFs in the promoters. **e** Density distribution of TF binding sites around genes. Expression levels (**f**) and tissue specificity (**g**) of genes bound by different numbers of TFs in the promoters. The mean expression levels (TPM) were calculated from various tissues. Tissue specificity is represented by the coefficient of expression variance (CV) across various tissues. A higher CV across tissues indicates greater tissue specificity, while a lower CV signifies reduced specificity. Different letters indicate $P < 0.01$ (Wilcoxon rank sum test). The nucleotide diversity (**h**) and density of rare SNP (**i**) around TF binding peaks compared with random regions

exhibited higher gene expression levels (Fig. 2f) and reduced tissue specificity (Fig. 2g). The genes with different number of TFs bound in promoters were enriched in distinct biological processes. For example, the genes without TFs bound in promoters were enriched in secondary metabolic process and response to stress, whereas genes with over 20 TFs bound in promoters were involved in translation, photosynthesis, generation of precursor metabolites, and energy (Additional file 2: Fig. S6). These genes, which lack TF binding in their promoters, may not be directly regulated by the TFs included in our DAP-seq assay (Additional file 1: Table S5). Alternatively, they may be regulated by TFs that were not captured in our experimental setup or by alternative regulatory mechanisms. DNA variants in TFBSs that alter gene expression contribute to variations of phenotypic traits in plants [20]. Based on published resequencing data [21], we found an obvious increase of nucleotide diversity at the center of TF binding peaks (Fig. 2h). Moreover, more rare SNPs were observed at TFBSs compared with flanking sequences

(Fig. 2i). These results suggest TFBSs are highly polymorphic in soybean populations and contain increased mutational load relative to their surrounding sequences.

Differential bindings of TFs contribute to expression bias of whole genome duplication (WGD) paralogs

Soybean is a well-documented paleopolyploid and has undergone at least two rounds of WGDs and subsequent diploidization [12]. As a result, nearly 75% of the genes in soybean genome are present in more than one copy. Following polyploidization and diploidization, there is a bias toward gene loss among subgenomes in many paleopolyploid species [22, 23]. To examine whether duplicated segments within soybean genome experienced global divergence of TF binding, we divided soybean duplicated block pairs into block1 (higher retention rate) and block2 (lower retention rate) based on the differences in retention rates of genes during diploidization (Additional file 1: Table S6) [24]. We found no significant difference in the numbers of bound TFs between block1 and block2 (Wilcoxon paired rank sum test, $P > 0.05$) (Fig. 3a, b). The levels of DNA methylation, chromatin accessibility, and histone modifications (H3K4me3, H3K27ac, and H3K27me3) at TFBSs and their flanking regions were also similar between block1 and block2 (Fig. 3c and Additional file 2: Fig. S7). These results suggest no overall differences in TF binding among the two soybean WGD blocks.

We further identified 16,634 WGD paralogs in soybean genome and observed significant divergence of promoter-bound TFs between two copies of WGD paralogs (Fig. 3d).

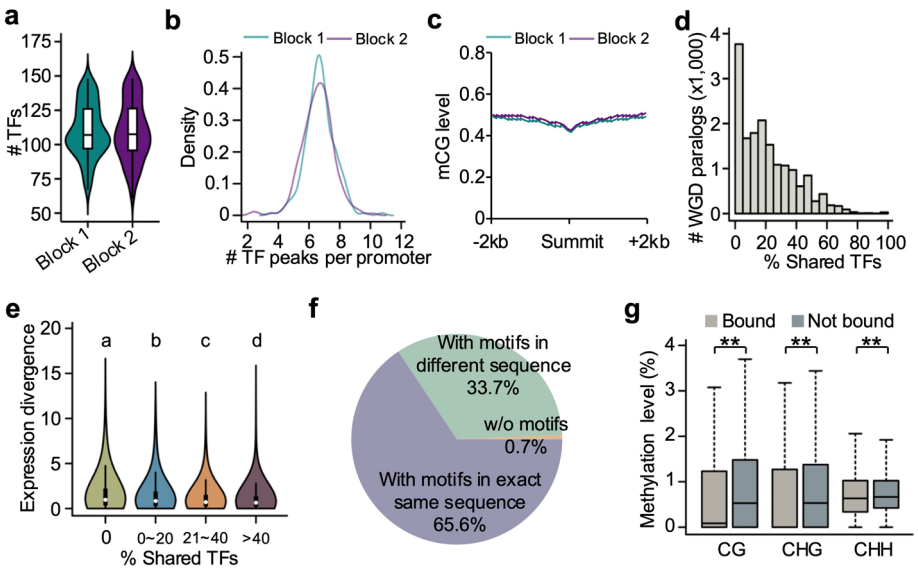

Fig. 3 Characteristics of TF bindings in WGD blocks and gene pairs in soybean. **a** Number of TFs bound to each duplicated block in soybean genome. **b** The number density of bound TFs per gene promoter in different duplicated blocks. **c** DNA methylation distribution around the binding peaks of TFs in different duplicated blocks. **d** Frequency of WGD paralogs with different percentages of shared TFs bound in promoters. **e** Gene expression divergence ($\log_2$(fold change)) between WGD paralogs with different percentages of shared TFs bounded in promoters. Different letters indicate $P < 0.01$ (Wilcoxon rank sum test). **f** Fractions of WGD paralogs containing same motifs in both copies or not. **g** Methylation levels of ± 100 bp flanking TFBSs in promoters bound by TFs in one copy of WGD paralogs relative to same TFBS motif sequences but not bound by TFs in another copy. ** indicates $P < 0.01$ (Wilcoxon paired rank sum test)

The majority (89.6%) of WGD paralogs had less than 50% of promoter-bound TFs in common (Fig. 3d). A total of 3679 (22.1%) WGD paralogs did not share any promoter-bound TFs with each other, which was significantly lower than that of random gene pairs (49.1%) (chi-square test, $P=0$) (Fig. 3d and Additional file 2: Fig. S8a, b). Strikingly, the TF binding differences of WGD paralogs were positively correlated with their expression divergences (Fig. 3e). We proposed that the motif sequence difference between WGD paralogs may mediate TF binding divergence in their promoters. To verify this, we identified TFs which specifically bound to the promoter of one copy in WGD paralogs, and then scanned motif sequences in the promoter of another gene copy. Unexpectedly, the corresponding motifs could be found in the majority (99.3%) of gene copies which were not bound by TFs (Fig. 3f). Over half (65.6%) of these gene copies shared the exact same motif sequences in their promoters (Fig. 3f). This fraction was significantly higher than that of random gene pairs (chi-square test, $P=0$) (Additional file 2: Fig. S8c). There is no significant distance divergence for TF motifs relative to TSSs between WGD copies (Additional file 2: Fig. S9a). The distance patterns between TFBSs in WGD paralogs were similar with those in random gene pairs (Additional file 2: Fig. S9). DNA methylation has been reported to repress TF binding in plants and animals [9, 25–27]. Consistent with this notion, we found the methylation levels at motifs bound by TFs in one gene copy were significantly lower than those motifs not bound by TFs in another copy (Wilcoxon paired rank sum test, $P<0.01$) (Fig. 3g), further explaining the dysregulation observed in duplicated genes with shared TF binding sites [9, 25–27]. Notably, even under relatively stringent thresholds for methylation level differences (CG > 0.1, CHG > 0.1, CHH > 0.05), there were marked increase in methylation at a substantial fraction of TFBS motif pairs with exact the same sequences (Additional file 2: Fig. S10). It can be concluded that the methylation differences at TFBSs may be involved in TF binding differences in the promoters of WGD paralogs.

Construction of GRN based on multi-omics data in soybean

To generate a comprehensive GRN in soybean, we combined our regulatory network generated from DAP-seq with multiple types of datasets, including gene co-expression networks, interaction networks based on DNA binding on the gene promoter and chromatin accessibility [9, 28–30] (see "Methods", Fig. 4a). After stringent filtering, we generated a TF regulatory network which was named as SoyGRN. The SoyGRN contained a total of 2.44 million interactions among 3188 TFs and 51,665 target genes, covering 91.0% (3188/3505) of TFs identified in soybean from PlantTFDB [29], which was used as the foundation for our entire analysis (Fig. 4a and Additional file 1: Tables S7 and S8). SoyGRN incorporated 74.7% (2773/3712) of the TFs listed in SoybeanTFDB [31], as well as 94.6% (123/130) of the bZIP family TFs and all reported TGACG-binding TFs identified in published studies [32, 33]. On average, there were 765 interactions for each TF in SoyGRN (Additional file 1: Table S7). The average number of target genes for TFs from different families showed significant variation, ranging from 115 genes in the bHLH family to 950 genes in the STAT family (Additional file 1: Table S9). The contribution of each network to SoyGRN ranges from 10.7 to 88.2% (Fig. 4b).

Multiple methods were employed to evaluate the reliability of interactions in SoyGRN. Using the published RNA-seq data of *GmMYB14* overexpression

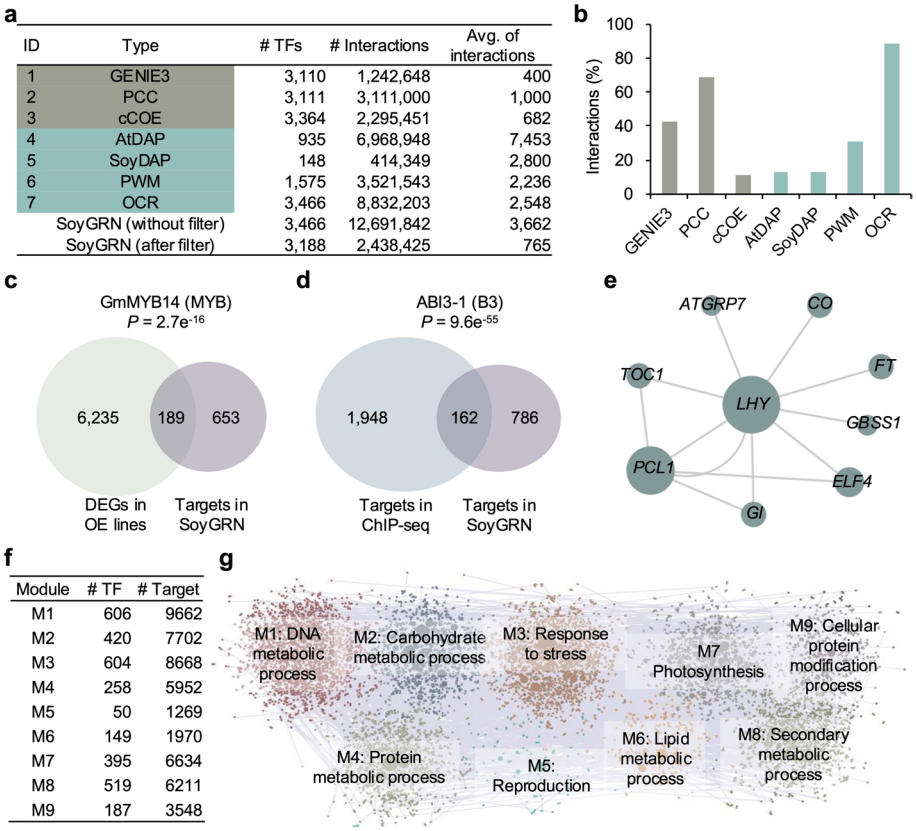

Fig. 4 Construction of SoyGRN based on multi-omics data. **a** Overview of interactions in SoyGRN. The three expression-based regulatory networks are the Pearson correlation coefficient (PCC) network, the Gene network inference with ensemble of trees (GENIE3) network, and the evolutionarily conserved regulatory network (cCOE network). The other four networks are based on TF binding occupancy of Arabidopsis (AtDAP network) and soybean (SoyDAP network), sequence-based evidence of DNA binding in promoter of targets and all available TF motifs (PWM network), and a chromatin accessibility network (OCR network). **b** Percentage of interactions supported by a certain independent network to the total number of interactions in SoyGRN. **c** Venn diagrams showing the overlap between target genes of GmMYB14 in SoyGRN and DEGs in *GmMYB14*-OE relative to wild type. The *P* value was calculated using hypergeometric test. **d** Venn diagrams showing the overlap between target genes of ABI3-1 in SoyGRN and target genes of ABI3-1 identified by ChIP-seq. The *P* value was calculated using hypergeometric test. **e** An example of known regulation relationships between *LHY* and target genes detected in SoyGRN. **f** The numbers of TFs and target genes in each module. **g** Regulatory circuit integrating information from the TF-target pairs and modules. Node colors represent different regulatory modules. The most significantly enriched biological process among genes in each module was shown

(*GmMYB14*-OE) lines [34], we identified 6424 differentially expressed genes (DEGs) between *GmMYB14*-OE and wild type (WT), which were the potential targets of *GmMYB14*. We found the target genes of *GmMYB14* in SoyGRN showed significant overlap with the DEGs (hypergeometric test, $P = 2.7e^{-16}$) (Fig. 4c). Moreover, in vivo TF binding defined by ChIP-seq data confirmed 17.1% of predicted interactions for ABI3-1 in SoyGRN (hypergeometric test, $P = 9.6e^{-55}$) (Fig. 4d). The non-overlapped genes may be attributed to the complexity of the regulatory network and the multifaceted nature of gene regulation. Some TF-target gene interactions might be restricted to specific developmental stages and environmental conditions. The systematic literature mining collects 1431 functionally confirmed TF-target interactions in the Arabidopsis transcriptional regulatory map (ATRM) [35]. Among 827

TF-target interactions from ATRM mapped to orthologs in soybean, 280 (33.9%) were predicted by SoyGRN (hypergeometric test, $P = 7.1e^{-303}$). For example, the known bindings of LHY to *ATGRP7*, *CO*, *ELF4*, *FT*, *GBSS1*, *GI*, *PCL1*, and *TOC1* were detected in SoyGRN (Fig. 4e). These results suggest high accuracy of TF-target interaction prediction in SoyGRN.

In general, genes linked to a similar biological process present a higher likelihood of physical interactions [36]. To investigate the functional modularity of SoyGRN, we applied partitioning algorithm to determine relationships between subsets of network elements and divided the SoyGRN into nine modules (Fig. 4f, g). Each module contains 50 to 606 TFs and 1269 to 9662 target genes (Fig. 4f). Functional enrichment analysis showed that genes in each module were enriched for specific biological process (Fig. 4g and Additional file 2: Fig. S11). For example, genes in module M4 were enriched in protein metabolic process, whereas genes in module M7 were involved in photosynthesis.

Identification of TFs regulating seed coat color

A TF could exert special biological function by regulating multiple downstream genes involved in the same biological process. If multiple genes in a specific metabolic pathway showed interactions with the same TF in SoyGRN, the TF may be a key regulator for the related biological process. To test it, we identified genes involved in seed coat color and predicted the potential functional TFs based on TF-target interactions in SoyGRN. The seed coat color is mainly determined by its anthocyanin and flavonoid contents [37]. A total of 3179 genes are annotated to be involved in the biosynthesis and modification pathways of flavonoids and anthocyanins in SoyBase database. The target genes of 79 TFs in SoyGRN were significantly enriched in the genes related with flavonoids and anthocyanins (Fig. 5a and Additional file 1: Table S10), indicating the potential roles of these TFs in regulating seed coat color. Consistent with the notion that MYB family TFs are involved in regulation of metabolism of flavonoid and anthocyanin [38], 25 (32%) candidate TFs for seed coat color belonged to MYB family (Fig. 5a), including *GmMYB100* that has been reported to regulate flavonoid biosynthesis in soybean [39].

Notably, we found a bHLH TF (*Glyma.10G026000*) regulated the expression of seven genes involved in the four major steps of the anthocyanin biosynthesis from naringenin chalcone, including *GmCHI04*, *GmF3H*, *GmF3'H*, *GmDFR1*, *GmDFR2*, *GmANS1*, and *GmANS2* (Fig. 5b). Phylogenetic analysis showed that *Glyma.10G026000* grouped with *AtTT8* in Arabidopsis (Additional file 2: Fig. S12a), which was named as *GmTT8b*. *AtTT8* regulates the seed coat color by effecting expression of flavonoid biosynthetic gene *DFR* in Arabidopsis [40]. To confirm the function of *GmTT8b* in soybean, we generated overexpression lines of *GmTT8b* (Additional file 2: Fig. S12b). The seeds of *GmTT8b* overexpression lines (OE1-3) showed intense pigmentations (Fig. 5c), consistent with its predicted function in the regulation of seed coat color. *GmTT8b* and its seven target genes in the anthocyanin biosynthesis pathway were preferentially expressed in seeds relative to other tissues (Fig. 5d). Interestingly, the proximal regions of these genes showed higher chromatin accessibility in seeds compared with other tissues based on published ATAC-seq datasets (Fig. 5e) [30].

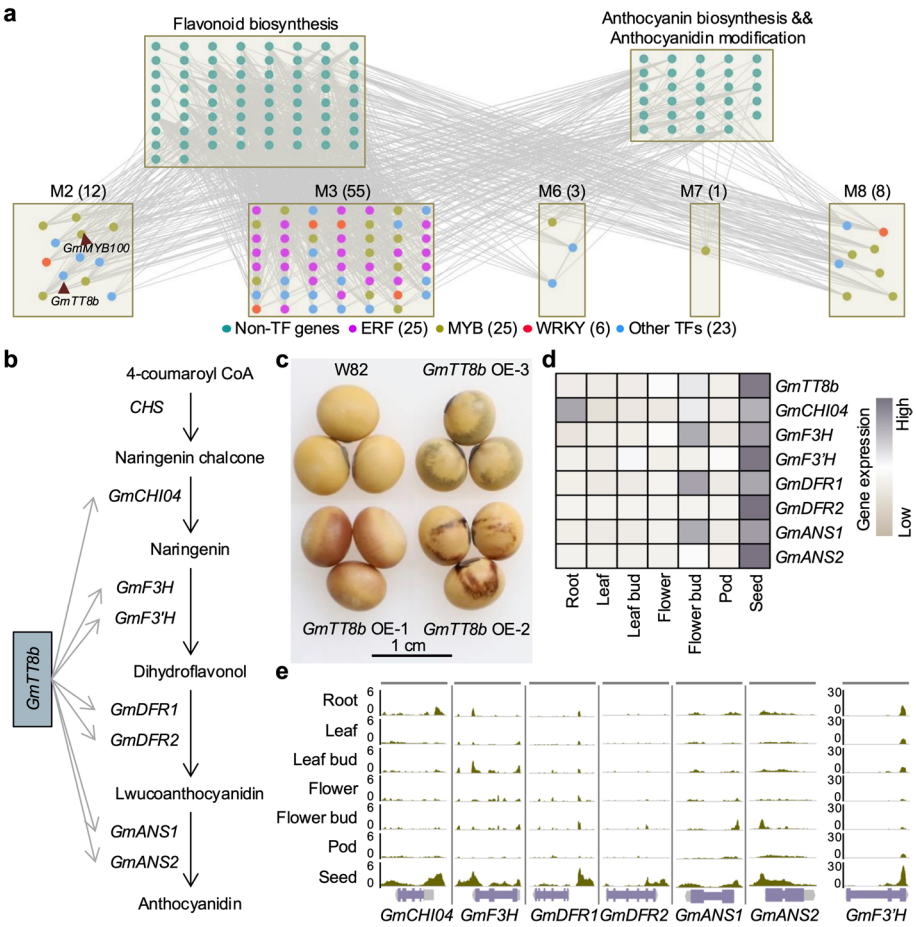

Fig. 5 Identification of TFs regulating seed coat color. **a** The TFs regulating genes involved in seed coat color including flavonoid biosynthesis, anthocyanin biosynthesis, and modifications. **b** *GmTT8b* regulating genes involved in anthocyanin biosynthesis pathway. **c** Altered seed coat color in *GmTT8b* overexpression lines compared with wild type. **d** Heatmap showing higher expression levels of *GmTT8b* and its target genes involved in anthocyanin biosynthesis pathway in seed relative to other tissues. **e** The chromatin accessibility in proximal regions of *GmTT8b* target genes in different soybean tissues

Identification of TFs regulating seed oil content

Increasing seed oil content is a major objective for soybean breeding due to a high global demand for edible vegetable oil [41]. To further dissect the key regulators controlling seed oil content, we collected 723 genes involved in oil accumulation including fatty acid synthesis, fatty acid elongation, and triacylglycerol biosynthesis. Based on interactions with these oil-related genes in SoyGRN, 279 TFs were predicted to regulate seed oil content (Additional file 1: Table S11), which included several well-known seed oil associated TFs such as *GmWRI1b* [42], *GmZF351*, and *NFYA* [15, 43]. Using published RNA-seq data from seeds at five different developmental stages corresponding to 3, 5, 6, 8, and 10 weeks after flowering [44], we found these candidate TFs exhibited divergence of expression pattern during seed development (Fig. 6a), indicating they may play roles in different developmental stages. To confirm the roles of these TFs in regulation of oil synthesis, we identified stop-gain mutations in 31 TFs from our previously generated mutant library [45] (Additional file 1: Table S12).

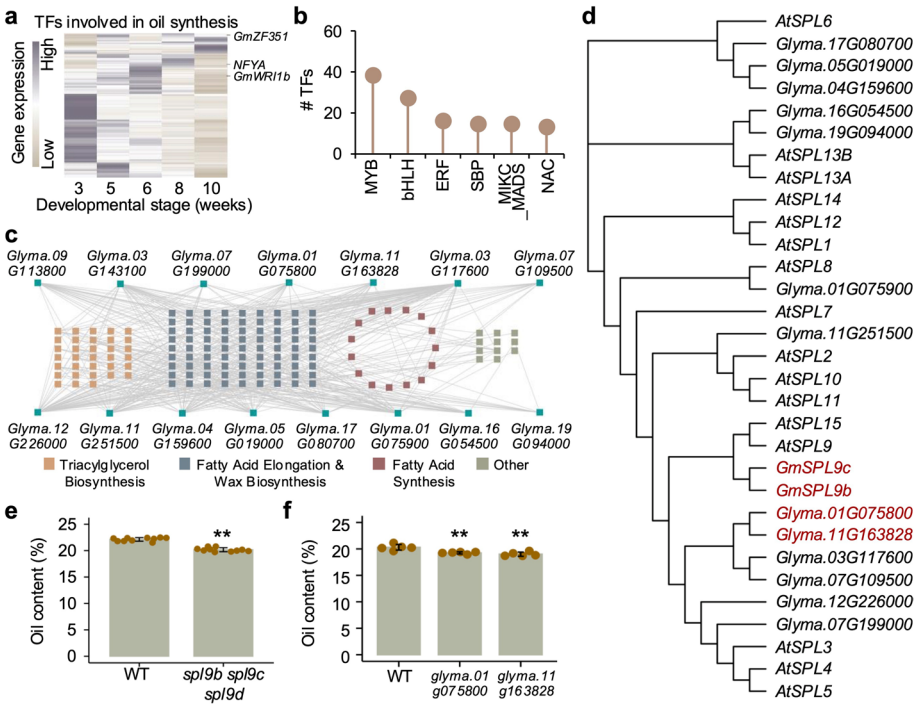

Fig. 6 Identification of TFs regulating seed oil content. **a** Heatmap showing the expression of 184 TFs predicted to regulate seed oil content in seeds at different developmental stages. **b** The top six TF families predicted to be involved in regulation of seed oil content. **c** Regulation network of SBP TFs and genes related to seed oil content. **d** Phylogenetic analysis of SBP family TFs regulating seed oil content. The TFs used for functional validation were highlighted in red. **e** Seed oil content of the wild type and *spl9b spl9c spl9d* mutant (mean ± SD, $n = 10$). ** indicates $P < 0.01$ (Student's *t*-test). **f** Seed oil content of the wild type and knockout mutants of *Glyma.01G075800* and *Glyma.11G163828* (mean ± SD, $n = 5$). ** indicates $P < 0.01$ (Student's *t*-test)

Disruption of 14 (45.2%) TFs displayed decreased seed oil content relative to wild type (Student's *t*-test, $P < 0.01$) (Additional file 2: Fig. S13), although we cannot exclude the effect of other unrelated mutations in mutant lines.

Among TFs predicted to regulate seed oil content, MYB, bHLH, ERF, SBP, MIKC_MADS, and NAC families were top six TF families (Fig. 6b). SBP genes are plant-specific TFs that control many important biological functions including stress responses and plant growth [4]. To our knowledge, SBP TFs are rarely reported to participate in lipid metabolism. The 15 SBP TFs regulated the expression of 81 (81/378, 21.4%) genes involved in the "fatty acid elongation & wax biosynthesis" pathway and 33 (33/137, 24.1%) genes involved in the "triacylglycerol biosynthesis" pathway (Fig. 6c). The *Glyma.09G113800* (*GmSPL9b*), *Glyma.03G143100* (*GmSPL9c*), and *Glyma.19G146000* (*GmSPL9d*) have been found to regulate plant architecture in our previous study [46], of which *GmSPL9b* and *GmSPL9c* were predicted to control oil synthesis in this study (Fig. 6d). To validate the roles of SBP TFs in regulating seed oil content, we measured the seed oil content of the previous generated *spl9b spl9c spl9d* mutant [46]. The *spl9b spl9c spl9d* mutant showed significant decreased seed oil content compared with wild type (Student's *t*-test, $P = 5.4e^{-10}$) (Fig. 6e). It is worth noting that four SBP genes (*Glyma.03G117600*, *Glyma.07G109500*, *Glyma.01G075800*, and *Glyma.11G163828*) clustered together in the phylogenetic tree of SBP genes

potentially involved in lipid metabolism (Fig. 6d). We further generated knockout mutants for *Glyma.01G075800* and *Glyma.11G163828* using the CRISPR/Cas9 technology (Additional file 2: Fig. S14). The seed oil contents of the two mutants were also significantly decreased relative to the wild type (Student's *t*-test, $P < 0.01$) (Fig. 6f). We observed more branches and more seeds per plant in knockout mutants compared with wild type (Student's *t*-test, $P < 0.05$) (Additional file 2: Fig. S15a–c). However, there were no discernible differences in 100-seed weight between wild type and the two mutants (Student's *t*-test, $P > 0.05$) (Additional file 2: Fig. S15d, e). These results suggest the important role of SBP TFs in regulation of seed oil content and demonstrate the predication reliability of functional TFs for agronomic traits using SoyGRN.

SoyGRN contributes to pinpoint causal TFs in QTLs for agronomic traits

The linkage disequilibrium has strongly hindered in the exploration of candidate genes within QTL intervals. Using TF-target interaction information, GRN can contribute to prioritize candidate genes in QTLs associated with complex traits [47]. For each TF within QTLs for a specific trait, we summed up the total interaction scores to its target genes within the rest of QTLs and compared with the 1000 random TFs (Fig. 7a). The TFs with total interaction scores higher than the top 5% random TFs were regarded as potential candidate trait-related TFs within QTLs.

Using the QTLs related to drought susceptibility index as input, SoyGRN prioritized *GmMYB306* (*Glyma.17G099800*) as a high confidence candidate gene associated with drought tolerance based on its interactions with target genes in QTLs (Fig. 7b). Indeed, Gene Ontology (GO) analysis showed the "response to abiotic stimulus" process was significantly enriched in the target genes of *GmMYB306* (Fig. 7c). The targets of *GmMYB306* included several drought stress-related genes such as *GmRVE8a* [48], *GmLCLa2*, and *GmLCLb1* [49] (Fig. 7d). Consistent with these results, overexpression of *MYB94*, an Arabidopsis homolog of *GmMYB306*, was shown to enhance drought tolerance, concomitantly promoting cuticular wax accumulation and mitigating cuticular transpiration in leaves [50]. These results demonstrate the potential of SoyGRN to prioritize candidate genes in QTL analysis. The same method was also applied to QTLs of other various traits, and many TFs were identified as candidate genes in regulating corresponding traits (Additional file 1: Table S13).

To facilitate the utilization of gene regulatory networks in this study, we integrated SoyGRN information to develop an interactive web platform, SoyTFBase (www.soytfbase.cn), for soybean community to explore the TF-gene relationships and dissect functional TFs associated with agronomic traits (Fig. 7e). Users can search the target genes of a defined TF, or search the TFs regulating a specific gene in SoyTFBase (Fig. 7f). We also implemented a "compare" tool to help users to discover common TF regulators for different genes, or common target genes of different TFs (Fig. 7g).

Discussion

TFs play important roles in regulating spatio-temporal specificity of gene expression in plants and animals through interaction with CREs [1]. Large-scale investigation of TF binding across genome can provide a comprehensive view of transcriptional regulatory network. Here, we generated a genome-wide view of TF binding patterns in the

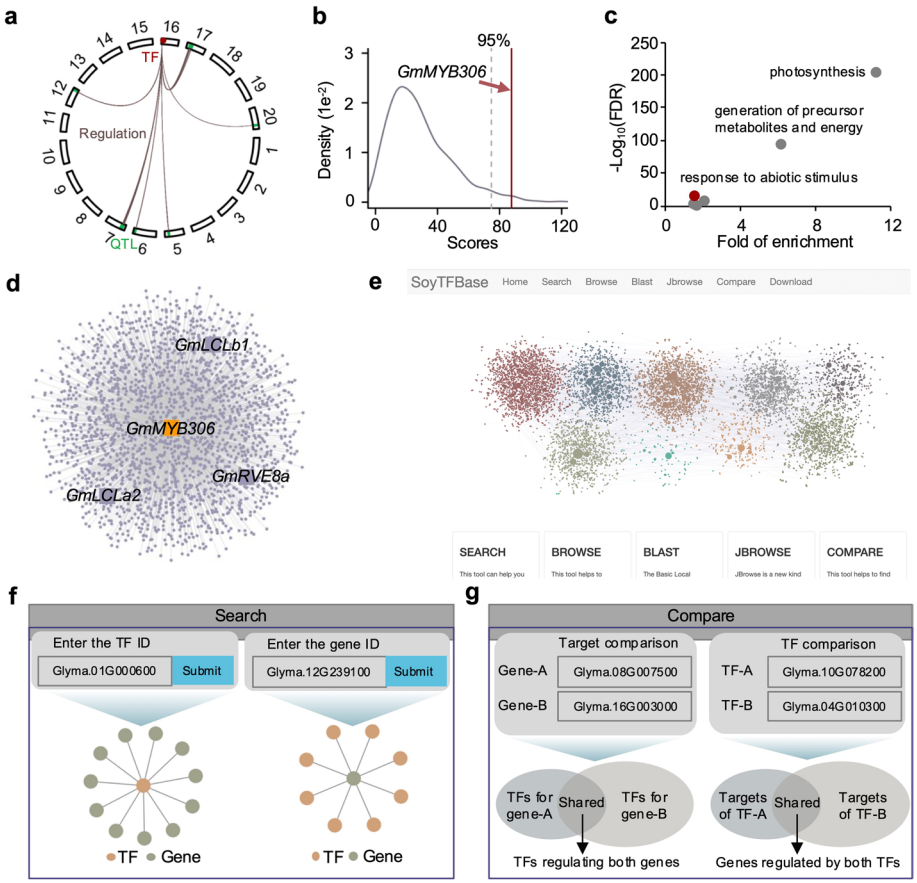

Fig. 7 Identification of candidate TFs within QTLs based on SoyGRN. **a** Examples illustrating the methods to calculate the total weight of TFs to its targets within QTLs. **b** The score distribution of random TFs and *GmMYB306* to their targets within QTLs associated with drought susceptibility index. Dashed line indicates the cutoff score at $P < 0.05$. Red line indicates *GmMYB306*. **c** Significantly enriched biological process in the target genes of *GmMYB306*. The top three most significantly enriched biological processes are labeled. **d** A network showing the regulation of *GmMYB306* on its target genes. The known genes related to drought stress response were highlighted and labeled. **e** The snapshot of the SoyTFBase website. The function description of search (**f**) and compare (**g**) tools in SoyTFBase website

soybean genome for 148 TFs from 28 different families using DAP-seq. We observed a discernible bias in the DAP-seq success rates among TF families, which is also reported in Arabidopsis and wheat [9, 10]. On the one hand, some TFs exhibit instability when recombinantly expressed and are thus not compatible with DAP-seq. On the other hand, the DNA-binding activity of numerous TFs is contingent upon the presence of specific protein partners or co-factors. Optimizing assay conditions and co-expressing TFs along with their interacting partners have the potential to enhance the efficacy of DAP-seq.

Our DAP-seq data for *GmbZIP67* in soybean showed a 41.6% overlap with ChIP-seq peaks. This is comparable to the overlap reported for wheat (37%) and Arabidopsis (36–81%) [9, 11]. The similarity in enriched motifs between DAP-seq and ChIP-seq peaks for *GmbZIP67* further indicates the high quality of our DAP-seq results. Despite the scarcity of ChIP-seq data for soybean, our findings demonstrate the value of DAP-seq in elucidating TF binding patterns. Specifically, we found that 65 out of 148 TF motifs (43.9%) exhibited high consistency with previously reported

motifs in their corresponding families. This substantial agreement underscores the robustness and reliability of our DAP-seq approach in identifying known TF binding motifs, thereby validating our experimental findings. It is worth noting that the remaining 58 (39.2%) TF motifs represent noncanonical representative motifs that were not previously reported. These novel motifs may represent TF-DNA interactions that are specific to experimental conditions, the soybean organism, or alternative binding modes of known TFs. They may also suggest a requirement for co-factors in TF binding. Additionally, it is possible that these motifs exhibit similarity to other TF motifs, as previously observed in ChIP-seq data, where sequences enriched do not always directly correspond to assayed TF motifs [51]. This highlights the complexity and diversity of TF-DNA interactions in soybean and underscores the value of our DAP-seq approach in uncovering novel regulatory elements.

The TFBSs were not randomly distributed in the soybean genome, but instead clustered into a set of HOT regions, which were enriched in OCRs (Fig. 2a, b). This is consistent with the report in wheat that TF HOT regions show high levels of chromatin accessibility and active histone modification H3K9ac [10]. Clustered TFBSs are also observed in human and the majority (92%) of the TF HOT regions are located in promoters or strong enhancer-like regions [7, 52]. The number of bound TFs correlates with the expression levels of the nearest genes in human [7]. Similarly, the number of bound TFs at gene promoters also positively correlated with the gene expression levels in soybean, which was in agreements with the higher chromatin accessibility of regions bound by more TFs (Figs. 1g and 2f). These suggest that the clustered binding distributions of TFs and their regulation on gene expression are conserved in plants and animals. On the other hand, these HOT regions may be bound by distinct TFs in vivo under different conditions or in different tissues, but not at the same time or in same cells. The chromatin accessibility, histone modifications, and expression of TF are involved in determining interaction specificity between CREs and TFs [1]. The in vitro DAP-seq using naked genomic DNA could capture TF-gene interactions at once which may occur under different conditions or in different tissues.

Uncovering GRNs can greatly promote our understanding of gene regulations and the key regulators of many important biological processes. However, the integrative GRN is lacking in soybean due to technical challenges. The constructed SoyGRN in this study captured TF-target interactions for 91.0% (3188/3505) of TFs identified in soybean from PlantTFDB [29], which was used as the foundation for our entire analysis. Notably, SoyGRN integrates 74.7% (2773/3712) of the TFs cataloged in another TF database SoybeanTFDB [31], demonstrating its substantial coverage of known TFs in soybean. To fully unlock the regulatory landscape of soybean, future endeavors should strive to integrate additional multi-omics datasets, which would enable the inclusion of regulatory information for the remaining TFs. The researchers could use SoyGRN to explore potential target genes for the interested TF or candidate TF regulators for the specific genes to accelerate functional analysis in soybean. In addition to seed coat color and oil content, SoyGRN would contribute to explore functional TFs in other agronomic traits with the input of related gene information.

Conclusions

In summary, we provided an integrative TF regulatory network SoyGRN for the exploration of gene regulatory relationships in soybean, which is a valuable resource for functional genomics and molecular breeding.

Methods

Plant materials and grouth conditions

Soybeans were planted in a greenhouse at 25 °C with a photoperiod of 12 h light/12 h dark. Two-week-old soybean leaves were harvested and stored at − 80 °C for DNA/RNA extraction. Wild type and transgenic mutants were planted in the experimental field at Hanchuan, Hubei, in 2023 for measurement of oil content. The wild type ("Williams 82") and ethyl methanesulfonate mutant lines were planted in Nanjing in 2022 for measurement of oil content. The mutant lines used were listed in Additional file 1: Table S12.

DAP-seq library preparation and sequencing

DAP-seq was performed as previously described [53]. In brief, about 2 μg genomic DNA was fragmented (300–500 bp), end-repaired, and 3′-end adenylated to ligate adapters using NEBNext Ultra™ II DNA Library Prep Kit for Illumina (NEB, Ipswich, USA). After purification using VAHTS DNA Clean Beads (Vazyme, Nanjing, China), the adapter-ligated DNA fragments was used for DAP-seq. Total RNA was isolated from various soybean tissues, including seeds, leaves, stems, roots, and flowers using TRIzol reagent (TaKaRa, Shiga, Japan). Following RNA quantification and quality assessment, the first-strand cDNA was synthesized using PrimeScript RT Reagent Kit with gDNA Eraser (TaKaRa, Shiga, Japan). The full list of TFs in the soybean genome was downloaded from PlantTFDB and used throughout the analysis [29]. The coding sequences of TFs were individually amplified by RT-PCR using the above cDNA as templates, based on the preferential expression patterns of each TF across different tissues (Additional file 1: Table S14). Subsequently, the amplified coding sequences were cloned into the pUC18-Halo-ORF vector to generate pUC18-Halo-TF vectors with ClonExpress®II One Step Cloning Kit (Vazyme, Nanjing, China). Each recombinant protein was translated from each vector using TNT® SP6 Coupled Wheat Germ Extract System (Promega, Madison, USA). A 50 μl protein expression reaction was performed with 1 μg of pUC18-Halo-TF plasmids, which was incubated at 30 °C for 2 h. Each 50 μl reaction yielded approximately 150–300 ng of Halo-tagged TF protein. The resulting proteins were immobilized onto 10 μl Magne HaloTag Beads (Promega, Madison, USA) with 40 μl wash buffer (pH = 7.4, 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO4, 1.8 mM KH_2PO4, and 0.05% NP40) on a rotator at room temperature for 1 h. The proteins immobilized into beads were subjected to three rigorous washes with 85 μl of wash buffer. Next, the protein-bound beads were incubated with 100 ng of an adapter-ligated gDNA library in 80 μl of wash buffer on a rotator at room temperature for 1 h. The beads were thoroughly washed three times with wash buffer, resuspended in 30 μl of elution buffer, and heated to 98 °C for 10 min. Immediately afterward, the mixture was cooled on ice for 5 min to stabilize the released DNA. The DNA fragments were directly amplified by 20 cycles of PCR using Q5 HiFi HotStart DNA Polymerase (NEB, Ipswich, USA). After purification, the

DAP-seq libraries were sequenced on a NovaSeq platform (Illumina, San Diego, USA) for 150-bp paired-end reads. The DAP-seq libraries were constructed in one replicate for 219 TFs and in two replicates for 11 TFs.

DAP-seq data processing

Raw reads were trimmed to remove adaptors and low-quality reads using Trimmomatic (version 0.39) [54]. Clean reads were mapped to the reference genome of *Glycine max* (Wm82.a4) by Bowtie2 (version 2.2.9) using the parameter "-X 1000" [55], retaining only concordantly mapped reads with MAPQ > 20. Pearson's correlation coefficient for read distribution between two biological replicates was computed genome wide using deepTools [56]. Peak calling was performed by GEM peak caller (version 3.4) using the Halo-GST sample as negative control with the following parameters "–f SAM –k_min 6 –kmax 20 -t 1 –outNP –outBED –outJASPAR –outMEME –outHOMER –s1 –k_neg_dinu_shuffle" [57], which generated peaks with a fixed width of ± 100 bp from the peak summit. A total of 148 DAP-seq datasets with FRiP > 2% were used in subsequent analysis [58]. The peaks were separated into different groups of 600 peaks based on the *P* value from the most significant peaks to the least significant peaks. The peaks of each group were extracted to detect motifs using MEME-ChIP (v5.4.1) with the default parameters [59]. The most significantly enriched motif in each group was compared to the top group. Once the most significantly enriched motif in a specific group belongs to a TF family distinct from the one identified in the top group, the peaks in that group and subsequent groups were discarded to ensure the high reliability of the peaks. The remaining peaks were used for further analysis. Read coverage across the genome was normalized by counts per million mapped reads (CPM) using deepTools [56]. The genes with TF peak summits located within the promoter regions (1 kb upstream to 500 bp downstream of transcription start site) were defined as target genes of each TF. The expected binding sites of TFs at the promoter regions were scanned using the k-mer set memory (KSM) analysis with a *P* value threshold of $1e^{-5}$ [60]. To assess the frequency of co-occurrence between peaks of two TFs, the presence and absence of each TF within 2-kb windows was evaluated across the genome. For this purpose, the presence of a TF in a given window was designated a value of 1, whereas its absence was assigned a value of 0. Pearson's correlation coefficients were computed for each pair of TFs.

Analysis of published ChIP-seq and ATAC-seq data

Published ChIP-seq and ATAC-seq data derived from leaves of soybean cultivar Williams 82 were downloaded from NCBI under accession numbers (PRJNA395102 for bZIP67, PRJNA395064 for ABI3-1, PRJNA657728 for H3K27ac, H3K4me3, and H3K27me3). After filtering low-quality reads and adapter sequences by TrimGalore using the parameters "–paired –stringency 3 –trim-n –max_n 7," clean reads were mapped to the reference genome of *Glycine max* (Wm82.a4) by Bowtie2 (version 2.2.9) using the parameter "-X 1000" [55], retaining only concordantly mapped reads with MAPQ > 20. Peak calling was performed using MACS2 with parameters "-f BAM -g 978,386,919 -n –keep-dup auto –call-summits" [61]. Consensus peaks from two biological replicates were extracted using the IDR pipeline with threshold of 0.01, retaining the overlapping regions between the two biological replicates within the consensus peaks

[62]. Duplicated reads were discarded using the Picard tools and the two replicates were merged. Normalized fold enrichment tracks were generated by MACS2 using the callpeak function with the -SPMR flag, followed by passing the bedgraph outputs into the bdgcmp function with the setting "-m FE" to calculate the fold enrichment (FE) values relative to the input control library.

To assess the overlap of peaks derived from ChIP-seq and DAP-seq experiments targeting *GmbZIP*I67, we processed the DAP-seq data using methods identical to those described for ChIP-seq. We segmented the soybean genome into 1-kb windows and quantified the number of windows overlapping with peaks from each technique. To assess the statistical significance of this overlap, we employed the hypergeometric test with the following parameters:

a = the number of 1-kb windows overlapping with peaks from ChIP-seq
b = the number of 1-kb windows overlapping with peaks from DAP-seq
total = the total number of 1-kb windows in the soybean genome
inter = the number of 1-kb windows overlapping with peaks from both ChIP-seq and DAP-seq

The hypergeometric test was conducted in R using the formula "phyper(inter-1, a, total-a, b, lower.tail = F)" to test against the null hypothesis that the observed overlap is no greater than expected by chance.

The motif enrichment analysis was performed using MEME-ChIP (v5.4.1) with the default parameters [59]. The *E* value calculated by MEME-ChIP (v5.4.1) quantifies the statistical significance of the observed motifs by assessing the probability that their distinctive features could have arisen by chance alone, rather than representing biologically meaningful signals.

Analysis of published MethylC-seq data

Published MethylC-seq data from leaves of soybean cultivar Williams 82 were downloaded from NCBI under accession numbers (PRJNA657728). After filtering by Trimmomatic (version 0.39) with default parameters [54], clean MethylC-seq reads were mapped to the reference genome of *Glycine max* (Wm82.a4) using Bismark (v0.15.0) with options (–score_min L,0,-0.2 -X 1000) [63]. The reads mapping to the same sites were collapsed into a single consensus read to reduce clonal bias. Then, each cytosine site covered by at least 2 reads in CG context and 3 reads in CHG and CHH contexts was retained for analysis. The methylation level of each cytosine site was calculated as C/(C + T). C indicates the number of reads with cytosine for this site and T indicates the number of reads with thymine for this site.

Analysis of published RNA-seq data

RNA-seq data generated from 28 different tissues and developmental stages of soybean cultivar Williams 82 were obtained from previous study with the NCBI accession number of PRJNA238493 [44]. The samples were collected from roots, shoots, leaves, flowers, and seeds at different developmental stages. Raw data were first cleaned using fastp [64] and mapped by HISAT2 (version 2.1.0) with the parameter of "-dta" [65]. Reads with

MAPQ > 60 were retained. The expression levels (transcripts per million, TPM) of genes were calculated by StringTie [66].

RNA-seq data of *GmMYB14* overexpression lines were downloaded from NCBI (accession number: PRJNA626031). Gene expression quantification and differential expression analysis were performed using EdgeR [67]. Fold changes (> 2) and FDR values (< 0.05) were used for identification of DEGs.

Analysis of genetic variation at TFBSs

The resequencing data of 302 soybeans were downloaded from previous study with the NCBI accession number of SRP045129 [21], including 62 wild soybeans (*G. soja*), 130 landraces, and 110 improved cultivars. The raw reads were cleaned using TrimGalore, and then aligned to the reference genome of *Glycine max* (Wm82.a4) by BWA using default parameters [68]. The duplicated reads were removed using Picard tools. Read pairs with a mapping quality lower than 10 were removed, and only coordinated mapping reads were retained. SNPs were called using GATK [69]. The SNPs were further filtered according to the following threshold "QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 4.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0." The SNPs with missing rates over 10% were discarded. Those SNPs with minor allele frequency less than 0.05 were regarded as rare SNPs, and the rest were kept to calculate the nucleotide diversity using VCFtools with parameter "–site-pi" [70]. The distribution of nucleotide diversity in the 2-kb flanking regions of TF peak summit was calculated in a sliding window of 100 bp. The randomly selected regions across the genome were used as the control.

Construction of SoyGRN based on combined networks

We collected various datasets to construct seven independent gene regulation networks (GENIE3, PCC, cCOE, AtDAP, SoyDAP, PWM, and OCR networks) as described in wheat [47].

For GENIE3 network, we calculated gene expression levels from 1461 soybean RNA-seq datasets using publicly available data in NCBI (Additional file 1: Table S15), covering gene expression in a wide range of genotypes, growth conditions, tissues, and developmental stages. The gene expression levels were calculated as described above, keeping only genes with > 1 TPM in at least 5% samples. These genes were used to construct the GRN using GENIE3 with a weight cutoff of 0.005 [71]. The score of each GENIE3 interaction was defined as follows: $GENIE3score(i) = \frac{weight(i)}{2\times \max weight(j)} + 0.5$, where j ranges from 1 to the total number of interactions.

For PCC network, we calculated the Pearson correlation coefficient (PCC) of each TF to expressed genes identified above. The interaction scores were defined as the absolute PCC, and the top 1000 genes with the highest score were considered as the potential targets.

For cCOE network, we analyzed the expression data of Arabidopsis to construct an evolutionarily conserved coexpression-based regulatory network. This included 49 RNA-seq datasets generated from various tissues, developmental stages, and growth conditions, which were used to construct the network in Arabidopsis [28]. Raw data were downloaded from NCBI and processed as described above for analysis of RNA-seq, followed by the construction of GENIE3 network and PCC network. For each TF, the

top 350 genes with the highest PCC score were retained. The total interaction score of each TF and gene pair were calculated based on scores from GENIE3 and PCC: $cCOEscore(i) = (\frac{GENIE3weight(i)}{2\times \max GENIE3weight(j)} + \mathrm{PCC(i)})/2 + 0.25$, where j ranges from 1 to the total number of interactions. The interactions were converted to soybean genes based on the sequence homology, and those without orthologs in soybean were discarded.

For AtDAP network, we downloaded the positions of TF binding peaks generated from various TF DAP-seq datasets in Arabidopsis [9]. The genes with peak summits located within the promoter regions were defined as the target genes of TFs. The regulatory networks were converted to soybean genes based on the sequence homology, and the interaction scores were defined as one.

For SoyDAP network, TF peak summits from DAP-seq data in this study within gene promoter regions were used to generate regulation relationships between TF and target genes. The interaction scores were defined as one.

For PWM network, we downloaded the non-redundant and high-quality position weight matrix (PWM) of Arabidopsis from the PlantTFDB [29] and scanned the PWM in gene promoter regions of soybean using FIMO with the default parameters from MEME software [72]. The interaction score was calculated based on score from FIMO. $PWMscore(i) = \frac{score(i)}{2\times \max score(j)} + 0.5$, where j ranges from 1 to the total number of interactions.

For OCR network, we downloaded ATAC-seq data of different tissues including root, leaf, leaf bud, flower, flower bud, pod, and seed from previous study [30] and identified peaks as described above with two biological replicates combined. Peak with the highest score within promoter was kept for each gene. The chromatin accessibility for a given gene was defined as follows: $chromatinscore(i) = \frac{\log_2 peakscore(i)+1}{2\times \max \log_2 peakscore(j)+1} + 0.5$, where j ranges from 1 to the total number of genes. The OCR network served as a supplement to the above networks. If OCRs exist in the promoter of target genes, we selected TF-target gene interactions from the aforementioned networks and assigned a new score based on chromatin accessibility.

Finally, we combined the all independent networks and summed up the interaction scores for each TF-gene pair. We removed the interactions supported only by homolog-map approaches in cCOE, AtDAP, and PWM. Interactions with scores lower than 1.3 were also discarded to obtain relative high confidence regulation relationships.

Network modularity and functional analysis

The TF-gene regulatory networks in SoyGRN were partitioned into different modules based on the connectivity via louvain algorithm in python-igraph (version 0.8.2) (https://igraph.org/). The networks were visualized using Cytoscape [73]. Functional enrichment of genes within each module and the target genes of each TF was performed using GOSlim. GO terms with FDR < 0.05 were considered as significantly enriched GO terms.

Prediction of TF regulators for seed coat color and oil content

We collected different sets of genes with potential roles in the regulation of seed coat color and oil content. For seed coat color, the 3122 genes predicted to be involved in the flavonoid biosynthesis, anthocyanin biosynthesis, and modification pathways in SoyBase

were extracted as a reference gene set [74]. The genes reported to be involved in the pathways including "Fatty Acid Elongation, Desaturation & Export From Plastid," "Fatty Acid Elongation & Wax Biosynthesis," "Fatty Acid Synthesis," "Triacylglycerol Biosynthesis," and "Triacylglycerol & Fatty Acid Degradation" in Arabidopsis were extracted from ARALIPmutantDB [75]. These genes were converted to 723 genes in the soybean genome based on homology and used as genes predicted to be related to seed oil content. The hypergeometric test was used to calculate the *P* value of overlap between candidate genes and the target genes of each TF. Candidate TF regulators were determined by a FDR < $1e^{-3}$ and fold enrichment > 2 for seed coat color and oil content.

Identification of candidate TFs within QTLs

QTLs for various traits of soybean were acquired from the SoyBase [74]. QTLs for the same trait were merged. For each TF within QTLs, we added up the total interaction scores to its targets within the rest of QTLs and compared the score with 1000 random selected TFs. The TFs with total interaction scores higher than the top 5% random TFs were retained, and the TFs with the highest score within each QTL were regarded as the potential candidate TF regulator for each trait.

Plant transformation

The full-length open reading frame (ORF) of *GmTT8b* was amplified by PCR from cDNA of soybean cultivar Williams 82 seeds using Phanta Max Super-Fidelity DNA Polymerase (Vazyme, Nanjing, China). The PCR fragments were recombined into pFGC5941 plasmid to generate the *35S::GmTT8b* construct. The vector was introduced into *Agrobacterium tumefaciens* strain EHA101 and used to transform soybean cultivar Williams 82 by *Agrobacterium tumefaciens*-mediated transformation. The gene expression levels in transgenic plants were tested by RT-qPCR analysis. The T2 seeds of *GmTT8b* were used for the investigation of phenotypes.

To knock out *Glyma.01G075800* and *Glyma.11G163828*, two gRNAs (AGTTGT AGGAGGAGACTAGC and GCTCCCATGATTCTGTTGCA) were selected using the web tool CRISPR-P (http://skl.scau.edu.cn/targetdesign/), and then built into the pYL-CRISPR/Cas9P35S-BS vector. The resulting constructs were transferred into the *Agrobacterium tumefaciens* EHA105. The EHA105 recombinant strain with corresponding plasmid was used to transform soybean cultivar Tianlong No. 1 according to the method as previously described [46]. The transgenic plants were genotyped to identify gene editing events near the targeted sites using Sanger sequencing. The homozygous T3 transgenic lines for *Glyma.01G075800* and *Glyma.11G163828* were used for phenotypic analysis. The primers used for soybean transformation were listed in Additional file 1: Table S16.

RNA extraction and RT-qPCR analysis

Total RNA was isolated using TRIzol reagent (TaKaRa, Shiga, Japan). The first-strand cDNA was synthesized using PrimeScript RT Reagent Kit with gDNA Eraser (Perfect Real Time) (TaKaRa, Shiga, Japan). The RT-qPCR was performed using SYBR Green Master Mix (Vazyme, Nanjing, China). *GmTUBULIN* was used as internal reference

gene for RT-qPCR. The resulting data was analyzed using the $2^{-\Delta\Delta CT}$ method. Primers used for RT-qPCR are listed in Additional file 1: Table S16.

Measurement of seed oil content in wild type and mutants

The oil content in soybean was measured as previously described [76]. In brief, about 20 dry seeds from wild type and mutants were ground to a fine powder and 100 mg of seed powder was added to 500 μl of the prepared 95% isopropanol followed by thorough mixing. After overnight rotation at 52 °C in incubator and centrifuge, the isopropanol supernatant was transferred to a new centrifuge tube and the sediment was washed by isopropanol again. After the complete evaporation of the isopropanol, the oil content was calculated according to the weight change of the centrifuge tube divided by original weight of seed powder.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03454-w.

Additional file 1: Supplementary Tables S1–S16.

Additional file 2: Supplementary Figures S1–S15.

Acknowledgements
We thank Dr. Yongming Chen at China Agricultural University for helpful suggestions in SoyGRN construction. We also thank Bioinformatics Center at Nanjing Agricultural University for data analysis.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
Q.S. and D.C. conceived the research. M.W., W.G., C.Z., Y.W., and Z.Z. performed experiments. W.J., Y.G., H.M., L.W., X.J. and W.Y. analyzed the data. W.J. and Q.S. wrote the manuscript.

Funding
This work was financially supported by the STI 2030-Major Projects (2023ZD04069) and National Key Research and Development Program of China (2021YFF1001204 and 2022YFD1201400).

Data availability
Sequencing data of DAP-seq are available at Genome Sequence Archive in National Genomics Data Center (https://bigd.big.ac.cn/gsa) under accession number PRJCA017519 [77]. The published datasets about ChIP-seq for H3K27ac, H3K4me3 and H3K27me3, ATAC-seq and MethylC-seq derived from leaves of soybean cultivar Williams 82 are available in NCBI under accession number PRJNA657728 [24]. The published ChIP-seq datasets for GmbZIP67 and GmABI3-1 are available in NCBI under accession number PRJNA395102 and PRJNA395064, respectively [18]. The published ATAC-seq datasets derived from different tissues of soybean cultivar Williams 82 are available in NCBI under accession number PRJNA751745 [30]. The RNA-seq data generated from 28 different tissues and developmental stages of soybean cultivar Williams 82 are available in NCBI under accession number PRJNA238493 [44]. The published RNA-seq data of ***GmMYB14*** overexpression lines and the wild type are available in NCBI under accession number PRJNA626031 [34]. The previously generated 1,461 soybean RNA-seq datasets covering gene expression in a wide range of genotypes, growth conditions, tissues and developmental stages were downloaded from NCBI under the accession numbers listed in Additional file 1: Table S15. The published resequencing data of 302 soybean accessions are available in NCBI under accession number SRP045129 [21]. No other scripts and software were used other than those mentioned in the "Methods" section.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Qingxin Song is an Editorial Board Member for Genome Biology but was not involved in the editorial process of this manuscript.

Received: 13 June 2024 Accepted: 4 December 2024
Published online: 18 December 2024

References

1. Schmitz RJ, Grotewold E, Stam M. Cis-regulatory sequences in plants: their importance, discovery, and future challenges. Plant Cell. 2022;34:718–41.
2. Marand AP, Eveland AL, Kaufmann K, Springer NM: cis-Regulatory Elements in Plant Development, Adaptation, and Evolution. Annu Rev Plant Biol. 2023;74:111-137.
3. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012;13:505–16.
4. Yamasaki K, Kigawa T, Seki M, Shinozaki K, Yokoyama S. DNA-binding domains of plant-specific transcription factors: structure, function, and evolution. Trends Plant Sci. 2013;18:267–76.
5. Ko DK, Brandizzi F. Network-based approaches for understanding gene regulation and function in plants. Plant J. 2020;104:302–17.
6. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012;22:1798–812.
7. Partridge EC, Chhetri SB, Prokop JW, Ramaker RC, Jansen CS, Goh ST, Mackiewicz M, Newberry KM, Brandsmeier LA, Meadows SK, et al. Occupancy maps of 208 chromatin-associated proteins in one human cell type. Nature. 2020;583:720–8.
8. Tu X, Mejia-Guerra MK, Valdes Franco JA, Tzeng D, Chu PY, Shen W, Wei Y, Dai X, Li P, Buckler ES, Zhong S. Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. Nat Commun. 2020;11:5089.
9. O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. Cistrome and epicistrome features shape the regulatory DNA landscape. Cell. 2016;165:1280–92.
10. Zhang Y, Li Z, Liu J, Zhang Y, Ye L, Peng Y, Wang H, Diao H, Ma Y, Wang M, et al. Transposable elements orchestrate subgenome-convergent and -divergent transcription in common wheat. Nat Commun. 2022;13:6940.
11. Zhang Y, Li Z, Zhang Y, Lin K, Peng Y, Ye L, Zhuang Y, Wang M, Xie Y, Guo J, et al. Evolutionary rewiring of the wheat transcriptional regulatory network by lineage-specific transposable elements. Genome Res. 2021;31:2276–89.
12. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463:178–83.
13. Sedivy EJ, Wu F, Hanzawa Y. Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol. 2017;214:539–53.
14. Zhang D, Zhao M, Li S, Sun L, Wang W, Cai C, Dierking EC, Ma J. Plasticity and innovation of regulatory mechanisms underlying seed oil content mediated by duplicated genes in the palaeopolyploid soybean. Plant J. 2017;90:1120–33.
15. Li QT, Lu X, Song QX, Chen HW, Wei W, Tao JJ, Bian XH, Shen M, Ma B, Zhang WK, et al. Selection for a zinc-finger protein contributes to seed oil increase during soybean domestication. Plant Physiol. 2017;173:2208–24.
16. Cai Y, Wang L, Chen L, Wu T, Liu L, Sun S, Wu C, Yao W, Jiang B, Yuan S, et al. Mutagenesis of GmFT2a and GmFT5a mediated by CRISPR/Cas9 contributes for expanding the regional adaptability of soybean. Plant Biotechnol J. 2020;18:298–309.
17. Xia Z, Watanabe S, Yamada T, Tsubokura Y, Nakashima H, Zhai H, Anai T, Sato S, Yamazaki T, Lu S, et al. Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proc Natl Acad Sci U S A. 2012;109:E2155-2164.
18. Jo L, Pelletier JM, Hsu SW, Baden R, Goldberg RB, Harada JJ. Combinatorial interactions of the LEC1 transcription factor specify diverse developmental programs during soybean seed development. Proc Natl Acad Sci U S A. 2020;117:1223–32.
19. Wang X, Qiu Z, Zhu W, Wang N, Bai M, Kuang H, Cai C, Zhong X, Kong F, Lu P, Guan Y. The NAC transcription factors SNAP1/2/3/4 are central regulators mediating high nitrogen responses in mature nodules of soybean. Nat Commun. 2023;14:4711.
20. Tian J, Wang C, Xia J, Wu L, Xu G, Wu W, Li D, Qin W, Han X, Chen Q, et al. Teosinte ligule allele narrows plant architecture and enhances high-density maize yields. Science. 2019;365:658–64.
21. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33:408–14.
22. Zhao M, Zhang B, Lisch D, Ma J. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. Plant Cell. 2017;29:2974–94.
23. Song Q, Chen ZJ. Epigenetic and developmental regulation in plant polyploids. Curr Opin Plant Biol. 2015;24:101–9.
24. Wang L, Jia G, Jiang X, Cao S, Chen ZJ, Song Q. Altered chromatin architecture and gene expression during polyploidization and domestication of soybean. Plant Cell. 2021;33:1430–46.
25. Kaluscha S, Domcke S, Wirbelauer C, Stadler MB, Durdu S, Burger L, Schubeler D. Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. Nat Genet. 2022;54:1895–906.
26. Cao S, Chen K, Lu K, Chen S, Zhang X, Shen C, Zhu S, Niu Y, Fan L, Chen ZJ, et al. Asymmetric variation in DNA methylation during domestication and de-domestication of rice. Plant Cell. 2023;35:3429–43.
27. Han T, Wang F, Song Q, Ye W, Liu T, Wang L, Chen ZJ: An epigenetic basis of inbreeding depression in maize. Sci Adv. 2021;7:eabg5442.

28. De Clercq I, Van de Velde J, Luo X, Liu L, Storme V, Van Bel M, Pottie R, Vaneechoutte D, Van Breusegem F, Vandepoele K. Integrative inference of transcriptional networks in Arabidopsis yields novel ROS signalling regulators. Nat Plants. 2021;7:500–13.
29. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017;45:D1040–5.
30. Huang M, Zhang L, Zhou L, Yung WS, Wang Z, Xiao Z, Wang Q, Wang X, Li MW, Lam HM. Identification of the accessible chromatin regions in six tissues in the soybean. Genomics. 2022;114:110364.
31. Mochida K, Yoshida T, Sakurai T, Yamaguchi-Shinozaki K, Shinozaki K, Tran LP. In silico analysis of transcription factor repertoire and prediction of stress responsive transcription factors in soybean. DNA Res. 2009;16:353–69.
32. Yang Y, Yu TF, Ma J, Chen J, Zhou YB, Chen M, Ma YZ, Wei WL, Xu ZS: The Soybean bZIP Transcription Factor Gene GmbZIP2 Confers Drought and Salt Resistances in Transgenic Plants. Int J Mol Sci. 2020;21:670.
33. Ullah I, Magdy M, Wang L, Liu M, Li X. Genome-wide identification and evolutionary analysis of TGA transcription factors in soybean. Sci Rep. 2019;9:11186.
34. Chen L, Yang H, Fang Y, Guo W, Chen H, Zhang X, Dai W, Chen S, Hao Q, Yuan S, et al. Overexpression of GmMYB14 improves high-density yield and drought tolerance of soybean through regulating plant architecture mediated by the brassinosteroid pathway. Plant Biotechnol J. 2021;19:702–16.
35. Jin J, He K, Tang X, Li Z, Lv L, Zhao Y, Luo J, Gao G. An Arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors. Mol Biol Evol. 2015;32:1767–73.
36. Alcala-Corona SA, Sandoval-Motta S, Espinal-Enriquez J, Hernandez-Lemus E. Modularity in biological networks. Front Genet. 2021;12:701331.
37. Gao R, Han T, Xun H, Zeng X, Li P, Li Y, Wang Y, Shao Y, Cheng X, Feng X, et al. MYB transcription factors GmMYBA2 and GmMYBR function in a feedback loop to control pigmentation of seed coat in soybean. J Exp Bot. 2021;72:4401–18.
38. Liu J, Osbourn A, Ma P. MYB transcription factors as regulators of phenylpropanoid metabolism in plants. Mol Plant. 2015;8:689–708.
39. Yan J, Wang B, Zhong Y, Yao L, Cheng L, Wu T. The soybean R2R3 MYB transcription factor GmMYB100 negatively regulates plant flavonoid biosynthesis. Plant Mol Biol. 2015;89:35–48.
40. Nesi N, Debeaujon I, Jond C, Pelletier G, Caboche M, Lepiniec L. The TT8 gene encodes a basic helix-loop-helix domain protein required for expression of DFR and BAN genes in Arabidopsis siliques. Plant Cell. 2000;12:1863–78.
41. Clemente TE, Cahoon EB. Soybean oil: genetic approaches for modification of functionality and total content. Plant Physiol. 2009;151:1030–40.
42. Guo W, Chen L, Chen H, Yang H, You Q, Bao A, Chen S, Hao Q, Huang Y, Qiu D, et al. Overexpression of GmWRI1b in soybean stably improves plant architecture and associated yield parameters, and increases total seed oil production under field conditions. Plant Biotechnol J. 2020;18:1639–41.
43. Lu X, Li QT, Xiong Q, Li W, Bi YD, Lai YC, Liu XL, Man WQ, Zhang WK, Ma B, et al. The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. Plant J. 2016;86:530–44.
44. Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, Ma Y, Liu T, Kong LA, Peng DL, Tian Z. Global dissection of alternative splicing in paleopolyploid soybean. Plant Cell. 2014;26:996–1008.
45. Zhang M, Zhang X, Jiang X, Qiu L, Jia G, Wang L, Ye W, Song Q. iSoybean: a database for the mutational fingerprints of soybean. Plant Biotechnol J. 2022;20:1435–7.
46. Bao A, Chen H, Chen L, Chen S, Hao Q, Guo W, Qiu D, Shan Z, Yang Z, Yuan S, et al. CRISPR/Cas9-mediated targeted mutagenesis of GmSPL9 genes alters plant architecture in soybean. BMC Plant Biol. 2019;19:131.
47. Chen Y, Guo Y, Guan P, Wang Y, Wang X, Wang Z, Qin Z, Ma S, Xin M, Hu Z, et al. A wheat integrative regulatory network from large-scale complementary functional datasets enables trait-associated gene discovery for crop improvement. Mol Plant. 2023;16:393–414.
48. Bao G, Sun G, Wang J, Shi T, Xu X, Zhai L, Bian S, Li X. Soybean RVE8a confers salt and drought tolerance in Arabidopsis. Biochem Biophys Res Commun. 2024;704:149660.
49. Yuan L, Xie GZ, Zhang S, Li B, Wang X, Li Y, Liu T, Xu X. GmLCLs negatively regulate ABA perception and signalling genes in soybean leaf dehydration response. Plant Cell Environ. 2021;44:412–24.
50. Lee SB, Suh MC. Cuticular wax biosynthesis is up-regulated by the MYB94 transcription factor in Arabidopsis. Plant Cell Physiol. 2015;56:48–60.
51. Worsley Hunt R, Wasserman WW. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. Genome Biol. 2014;15:412.
52. Yan J, Enge M, Whitington T, Dave K, Liu J, Sur I, Schmierer B, Jolma A, Kivioja T, Taipale M, Taipale J. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. Cell. 2013;154:801–13.
53. Bartlett A, O'Malley RC, Huang SC, Galli M, Nery JR, Gallavotti A, Ecker JR. Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat Protoc. 2017;12:1659–72.
54. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
55. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
56. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44:W160-165.
57. Guo Y, Mahony S, Gifford DK. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput Biol. 2012;8:e1002638.
58. Galli M, Khakhar A, Lu Z, Chen Z, Sen S, Joshi T, Nemhauser JL, Schmitz RJ, Gallavotti A. The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. Nat Commun. 2018;9:4526.
59. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011;27:1696–7.
60. Guo Y, Tian K, Zeng H, Guo X, Gifford DK. A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. Genome Res. 2018;28:891–900.

61. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
62. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. Ann of Appl Stat. 2011;5(1752–1779):1728.
63. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.
64. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
65. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
66. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
67. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
68. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
69. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
70. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
71. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P: Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010;5:e12776.
72. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202-208.
73. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
74. Grant D, Nelson RT, Cannon SB, Shoemaker RC. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 2010;38:D843-846.
75. McGlew K, Shaw V, Zhang M, Kim RJ, Yang W, Shorrosh B, Suh MC, Ohlrogge J. An annotated database of Arabidopsis mutants of acyl lipid metabolism. Plant Cell Rep. 2015;34:519–32.
76. Liu YF, Li QT, Lu X, Song QX, Lam SM, Zhang WK, Ma B, Lin Q, Man WQ, Du WG, et al. Soybean GmMYB73 promotes lipid accumulation in transgenic plants. BMC Plant Biol. 2014;14:73.
77. Jiao W, Wang MM, Guan YJ, Guo W, Zhang C, Wei YC, Zhao ZW, Ma HY, Wang LF, Jiang XY, Ye WX, Cao D and Song QX. Transcriptional regulatory network reveals key transcription factors for regulating agronomic traits in soybean. Datasets. National Genomics Data Center (NGDC). https://ngdc.cncb.ac.cn/gsa/search?searchTerm=PRJCA017519. 2023.

Publisher's Note