Genomics 114 (2022) 110364

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Identification of the accessible chromatin regions in six tissues in the soybean

Mingkun Huang [a,b], Ling Zhang [b], Limeng Zhou [a], Wai-Shing Yung [a], Zhili Wang [a], Zhixia Xiao [a], Qianwen Wang [a], Xin Wang [a], Man-Wah Li [a], Hon-Ming Lam [a,*]

[a] School of Life Sciences and Centre for Soybean Research of the State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong, SAR, PR China

[b] Lushan Botanical Garden Jiangxi Province and Chinese Academy of Sciences, NO.9 Zhiqing Road, Lushan, Jiujiang, Jiangxi, PR China

ARTICLE INFO

Keywords:
Soybean
Glycine max
ATAC-seq
Accessible chromatin region (ACR)
Genome architecture

ABSTRACT

Accessible chromatin regions (ACRs) are tightly associated with gene expressions in the genome. Conserved non-coding *cis*-regulatory elements, such as transcription factor binding motifs, are usually found in ACRs, indicating an essential regulatory role of ACRs in the plant genome architecture. However, there have been few studies on soybean ACRs, especially those focusing on specific tissues. Hence, in this study, with the convenient ATAC-seq, we identified the ACRs in six soybean tissues, including root, leaf bud, flower, flower bud, developing seed, and pod. In total, the ACRs occupied about 3.3% of the entire soybean genome. By integrating the results from RNA-seq and transcription factor (TF) ChIP-seq, ACRs were found to be tightly associated with gene expressions and TF binding capacities in soybean. Together, these data provide a comprehensive understanding of the genomic features of ACRs in soybean. As a collection of essential genomic resources, these processed data are made available at datahub.wildsoydb.org.

1. Introduction

The loosened or open chromatin in specific genomic regions, resulting from the activation of gene expression, is often sensitive to external nucleases such as transposases, and thus such regions are termed accessible chromatin regions (ACRs). Genome-wide analyses of ACRs are usually performed using the convenient Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) method, which was first reported in humans in 2013 [1] and later in plants in 2016 [2,3]. ACRs are usually enriched with non-coding *cis*-regulatory elements (CREs). Due to their tight association with gene expression, ACRs are essential for genome architecture both in mammalian and plant genomes [4,5]. In mammals, according to the Encyclopedia of DNA Elements (ENCODE) project, tens of thousands of ACRs have been experimentally identified in the human and mouse genomes [6]. Several mammalian ACRs have been found to be associated with severe diseases, suggesting their regulatory importance [7]. Similarly, in plants, sequence variations in distal ACRs or ACRs in core promoters are involved in plant domestication. For example, due to a transposon insertion in a distal ACR associated with the key transcription factor, *Teosinte Branched1* (*TB1*), the *tb1* mutant had increased apical dominance in maize [8,9], a favorable plant architecture for high-density planting. Furthermore, variations in the promoter regions of a *Teosinte branched1/cycloidea/proliferating cell factor* (*TCP*) gene are linked to adaptations to high/low soil nitrogen in rice [10]. In addition, the application of CRISPR/Cas9 to the editing of a *WUSCHEL HOMEOBOX9* (*WOX9*) gene-associated ACR resulted in new agronomic traits in tomato without changing the target protein sequences [11], demonstrating that genetic modification of ACRs is a promising strategy for trait improvement in crops.

Soybean is one of the essential crops with abundant genome assembly resources [12–16]. However, there have been limited studies on soybean ACRs, especially in a tissue-specific manner. To our knowledge, most of the available ACR datasets focused on the leaves of either cultivated or wild soybeans [5,17–19]. To facilitate a comprehensive understanding of the genomic features of ACRs in soybean, we adopted the widely used Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) to identify ACRs genome-wide in different tissues of the cultivated soybean, Williams 82 (W82, *Glycine max*). By integrating these results with those from RNA-seq and TF ChIP-seq, we

* Corresponding author.
E-mail addresses: hmk1111@qq.com (M. Huang), limanwah@cuhk.edu.hk (M.-W. Li), honming@cuhk.edu.hk (H.-M. Lam).

https://doi.org/10.1016/j.ygeno.2022.110364
Received 30 November 2021; Received in revised form 25 February 2022; Accepted 6 April 2022
Available online 12 April 2022

were able to provide a comprehensive overview of the relationships among ACRs, gene expressions, as well as TF binding capacity. Finally, information on the ACRs identified in this study is available on our website (https://datahub.wildsoydb.org), as a resource to facilitate further explorations in soybean research.

2. Materials and methods

2.1. Plant material

Soybean (*G. max*) cv Williams 82 was used in this study. Plants were grown in the greenhouse under natural conditions for the collection of flowers, floral buds, pods, leaf buds and developing seeds, where developing seeds were a mixture of immature seeds at different developmental stages. For root samples, seeds were sterilized with 75% ethanol and germinated on wet germination paper (Anchor Seed Germination Paper, 38 × 25 cm) in dark for 3–4 days, and then the primary roots (3–4 cm long) were harvested for subsequent experiments. Unless otherwise stated, the soybean plants were grown in a greenhouse in the Chinese University of Hong Kong under normal light/dark (~16/8 h) conditions and the grow temperature ranged from 20–28 °C.

2.2. RNA-seq

Soybean tissue samples mentioned above were snap-frozen in liquid nitrogen and stored at −80 °C until ready to be used. Total RNA was extracted with the RNeasy Plant Mini RNA Isolation Kit (QIAGEN). mRNA was isolated with magnetic Dynabeads™ Oligo(dT)$_{25}$ (Invitrogen) following the manufacturer's instructions. The strand-specific RNA-seq libraries were constructed with mRNA fragmentation followed by reverse transcription, second-strand synthesis with dUTP, end-repairing, dA-tailing, Y-shape adapter ligation and UNG digestion and finally library amplification. The libraries were sent for sequencing on the NovaSeq platform (Illumina) in the paired-end 150 bp mode. RNA-seq was performed with two biological replicates.

Adaptors and low-quality reads were removed by the TrimGalore software with the default parameters (http://github.com/FelixKrueger/TrimGalore). The clean reads were aligned to the soybean Williams 82 V4 reference genome (Wm82.a4.v1, https://phytozome.jgi.doe.gov) with HISAT2 under the strand-specific mode (http://daehwankimlab.github.io/hisat2/). The fragments per kilobase per million mapped reads (FPKMs) were calculated by Cuffnorm (http://cole-trapnell-lab.github.io/cufflinks/cuffnorm).

2.3. ChIP-seq data processing

For TF ChIP-seq data, GmDt2 [20], GmLEC1 [21], GmNAC [22], GmABI3 [22], GmbZIP67 [22], GmAREB3 [22] and GmJAG1 [17] were downloaded from NCBI SRA database. Adaptors and low-quality reads were removed using TrimGalore. The filtered reads were aligned to the soybean Wm82.a4.v1 reference genome using Bowtie2 [23] with default parameters. Mapped reads with MAPQ value >30 were used for peak calling via MACS2 [24] with the parameters '-q 0.01 -g 1.0e+9'. Heatmaps of ChIP-seq data were processed using Deeptools [25].

2.4. Motif analysis

The MEME suite software package was used for motif analysis using the default parameters [26]. AME was used for motif enrichment analysis and FIMO was used for motif scanning. The known soybean TF motif position weight matrix (PWM), such as the G-box PWM was downloaded from the PlANTPAN3.0 database [27].

2.5. ATAC-seq

ATAC-seq libraries were prepared as described in previous studies [17,19] with two biological replicates. TrimGalore was used for adaptor trimming and low-quality read filtering. The filtered reads were mapped to soybean Wm82.a4.v1 reference genome using Bowtie2 [23] with the parameters set as: –very-sensitive -X 1000. The mapped reads containing MAPQ >30 were extracted by samtools and used for peak calling via the Genrich software in the ATAC-seq mode (−t Rep1,Rep2 -c Input -o Output -j -r -y) (https://github.com/jsh58/Genrich). The genomic DNA processed by Tn5 was used as the ATAC-seq input control. The ATAC-seq data, after removing PCR duplicates, were normalized by the input control using the Bam compare function and the profiles of ATAC-seq signals were visualized using PlotProfile in Deeptools [25]. The ACRs were further collapsed into the mACRs using BEDtools merge function with the default parameters (https://bedtools.readthedocs.io). ACR annotation was performed by the ChIPseeker [28]. GO enrichment analysis was performed by an online tool, Plantregmap (http://plantregmap.gao-lab.org/go.php), with default parameters.

2.6. Single nucleotide polymorphism (SNP)/insertion and deletion (indel) calling

The short reads of wild soybean W05 and cultivated soybean C08 resequencing data were obtained from previous works [14,29]. Adaptors and low-quality reads in raw data were filtered using TrimGalore. The filtered reads were mapped to the cultivated soybean Wm82.a4.v1 reference genome using BWA [30] with default parameters. Properly paired and uniquely mapped reads were extracted to call SNPs by SAMtools mpileup (https://github.com/samtools/) and BCFtools Call (http://samtools.github.io/bcftools) with default parameters.

3. Results

3.1. Mapping soybean ACRs in distinct tissues via ATAC-seq

To provide a comprehensive understanding of ACRs in the genome of the cultivated soybean (*G. max*, Williams 82), we employed ATAC-seq to map the ACRs in six main tissues, including flower bud (FB), flower (FL), leaf bud (LB), pod (PO), root (RO) and developing seed (SE) (Table S1). In addition, the previously released leaf (LF) ATAC-seq data were also integrated into this study [17]. After mapping, the ATAC-seq signals were found to be mainly enriched at the gene transcription start sites (TSSs). The ATAC-seq signals and the RNA-seq data (Table S1 and S2) in each tissue were well correlated (Fig. S1), which suggests that ACRs are tightly associated with gene expressions. Furthermore, the number of reproducible ACRs in each tissue in two replicates ranged from ~25,311 to ~33,456, with an average of 28,917 (Fig. S2, Table S3). Considering the relatively high correlation between two replicates (Fig. S3), these reproducible peaks in these two replicates were a confident collection for the representative ACRs in each tissue.

To obtain a general landscape of such *cis*-regulatory regions in the soybean genome, all these ACRs from different tissue types were further collapsed into 62,019 merged ACRs (mACR) as a reference set for further analyses (Table S4). The mACRs occupied about 3.3% of the soybean genome. Most of the mACRs (93.1%) were shorter than 1000 bp (Fig. 1A), with peak centers (PCs) usually located close to the TSS of a nearby gene (Fig. 1B) and a significantly higher GC content than the random genomic sequence (Fig. 1C). The proportion of mACRs situated in the promoter region was 56.2%, while about 30.3% were located in the Distal intergenic region, which was similar to the distribution in each tissue (Fig. 1D, Fig. S4).

In addition, these mACRs had a 91.1–99.7% alignment rate to 37 released soybean reference genomes, indicating that most of the ACRs were common to both cultivated and wild soybeans (*Glycine soja*) (Fig. S5). The alignment rates were in general lower in wild soybean genomes, suggesting that there has been a divergence of the ACRs of cultivated soybean from those of wild soybean.

To identify ACRs common to all tissues and those found only in

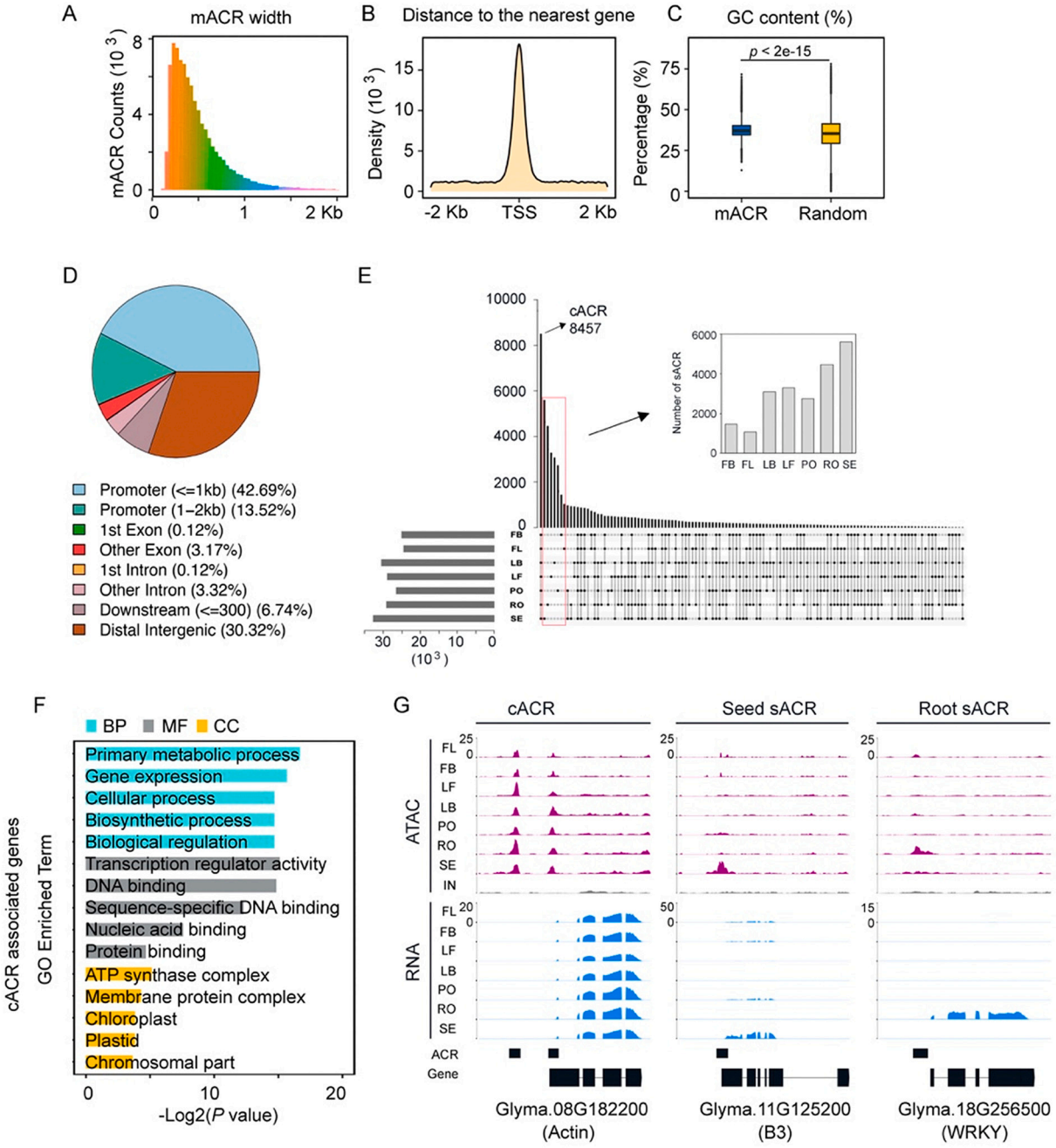

Fig. 1. Genome-wide identification of soybean accessible chromatin regions (ACRs) in different tissues.
(A) Distribution of the width of the merged ACRs (mACRs).
(B) Distance of mACR peak center (PC) to the transcription start site (TSS) of the nearest gene.
(C) Comparison of the GC content between mACRs and random genomic regions. *P*-value was calculated by the Wilcoxon test.
(D) Distribution of mACRs in the soybean genome. Distal intergenic regions are the stretches of DNA beyond the promoter, genic and downstream regions.
(E) Upset venn diagram showed the number of cACR and sACR in each tissue. The distribution of sACR (21,680 in total) in each tissue was indicated in the upper bar chart.
(F) Gene ontology (GO) enrichment of the cACR-associated genes. BP, biological progress; MF, molecular function; CC, cell component.
(G) Examples of cACRs and sACRs and an associated gene for each. FB, flower bud; FL, flower; LB, leaf bud; LF, leaf; PO, pod; RO, Root; SE, Seed; IN, ATAC-seq input.

specific tissues, we used the mACR dataset as a reference against which the ACRs identified in different tissues were compared. About 8457 mACRs (13.6% of total, with 10.2% in proximal regions encompassing the promoter, genic and downstream regions, and 3.4% in distal regions) could be found in all these tissues, and were regarded as common ACRs (cACRs) (Fig. 1E, Table S4).Then there were a total of 21,680 mACRs (34.9% of total, with 23.0% in proximal regions and 11.9% in distal regions) which were only found in one tissue type and were labeled as tissue-specific ACRs (sACRs) (Fig. 1E, Table S4).

Furthermore, Gene ontology (GO) enrichment analyses of these cACR-associated genes showed several significantly enriched GO terms including primary metabolic process, transcription regulator activity, ATP synthase complex and other related ontologies (Fig. 1F). These cACR-associated genes with broad expression were potentially involved in housekeeping functions, such as *Actin* (*Glyma.08G182200*) (Fig. 1G). In contrast, sACR-associated genes were enriched in development-

related GO terms (Fig. S6), such as "regulation of cellular process" in pod, "response to auxin" in root, "cell differentiation" in leaf bud, "developmental process" in seed, "regulation of cell communication" in flower bud, "leaf development" in leaf, and "regulation of cellular process" in flower. Particularly, genes associated with sACRs, such as those encoding a B3 domain-containing transcription factor (*Glyma.11G125200*) [31] and a WRKY (*Glyma.18G256500*) [32] (Fig. 1G) might be essential for maintaining tissue-specific functions.

3.2. Presence/absence of homoeologous mACRs in soybean affect the gene expression

As a paleopolyploid, soybean underwent at least two rounds of whole-genome duplication. Multiple copies of homoeologous coding genes (hGenes) were retained in the genome [13,14], while studies of homoeologous ACRs (hACRs) retention in soybean leaf have also suggested a tight relationship between hACRs and the soybean

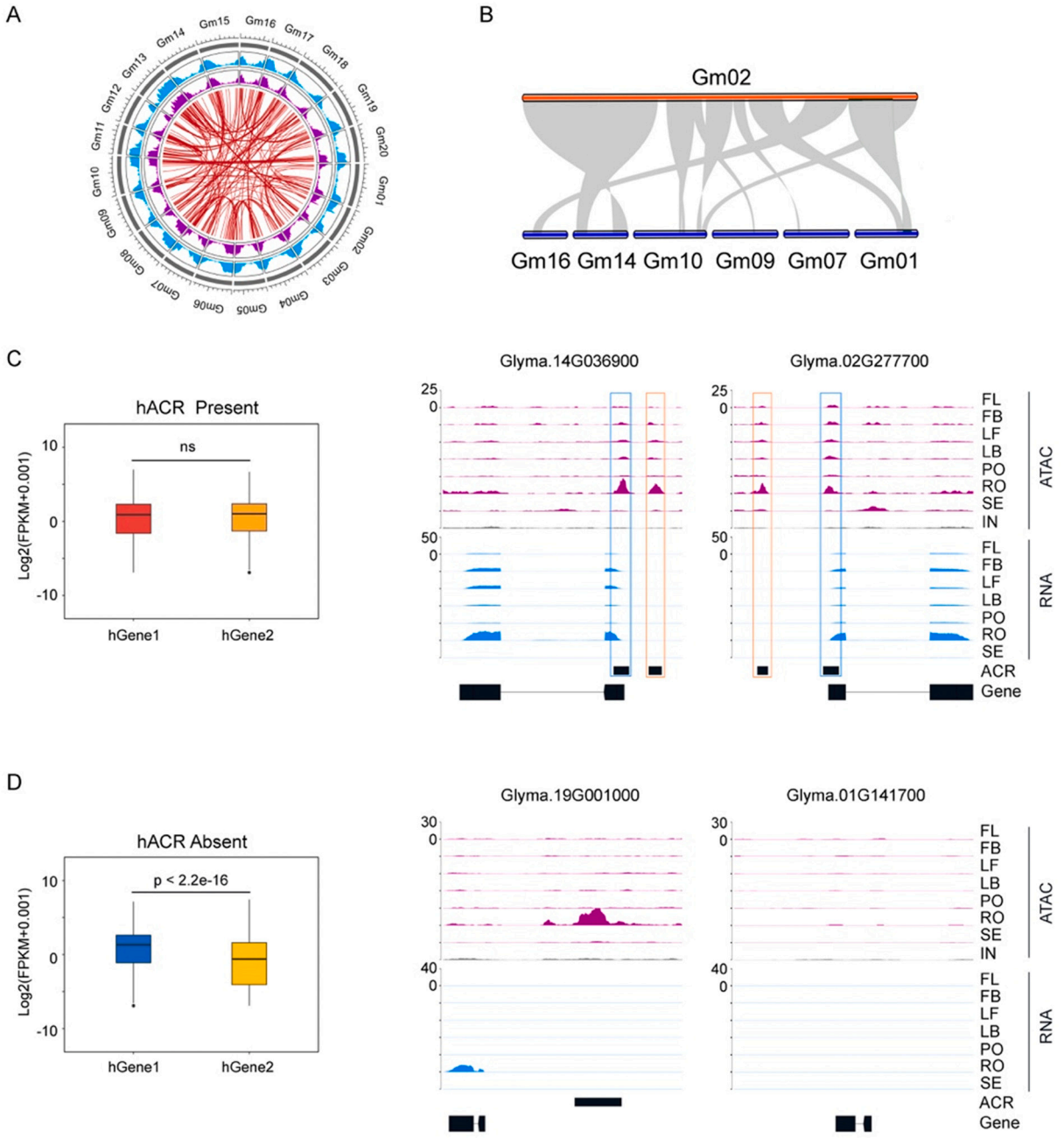

Fig. 2. The genomic feature of homoeologous ACRs (hACRs).
(A) A Circos diagram showing the distribution of hACRs in the soybean genome. The blue outer circle indicates the distribution of ACR peaks. The purple inner circle indicates the distribution of ACR-associated genes. The red links indicate 600 representative homoeologous ACR pairs.
(B) hACRs arranged in the synteny block structure between chromosomes. The linkages between hACRs (linked by grey stripes) between Chromosome 02 (Gm02) as well as other chromosomes are used as an example.
(C) The expressions of homoeologous genes (hGenes) associated with hACRs showed no significant differences (ns, left panel). Gene expression is indicated by the average fragments per kilobase per million mapped reads (FPKM) among tissues. A representative case of two hGenes associated with hACRs is shown (right panel). P-value was calculated by the Wilcoxon test. hACRs are indicated by boxes with same colors.
(D) hGenes without hACR showed significant changes in gene expressions (left panel). A representative case of two hGenes without hACRs is shown (right panel). P-value was calculated by the Wilcoxon test. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

domestication/polyploidization [18,19,33].

In this study, making use of the mACR collection in soybean genome, we found 14,633 homoeologous ACRs (hACRs) (e-value <1e-5, percentage of identity >70% and alignment length > 50% of the ACR length), which accounted for about 42.7% of mACRs in the soybean genome (Fig. 2A, Table S5). As expected, about 71.9% of the hACR pairs were associated with the hGenes retained in the soybean genome (Table S5). In addition, most of these hACRs and their associated hGenes were located in the same synteny block (Fig. 2B, Fig. S7), indicating that the retention of hACR-hGene pairs together occurs after the whole-genome duplication.

Furthermore, the overall expressions between the members of these hACR-hGene pairs showed no significant changes, such as the hGene pair of *Glyma.14G036900* and *Glyma.02G277700* (Fig. 2C). In contrast, there were 5968 hGenes in which an ACR was associated with only one member in the homoeologous pair (Table S6). Similar to recent published studies [18,19], the expression of the half of an hGene pair associated with an ACR was higher than that of its homoeologous counterpart without an associated ACR, providing evidence that the absence of an hACR could affect the expression of an hGene (Fig. 2D). Also, this suggests that subfunctionalization or non-functionalization of a functional coding gene might be resulted from the presence/absence of an associated ACR after the whole-genome duplication.

3.3. Conserved transcription factor (TF) binding motifs in the mACRs

ACRs usually contain conserved *cis*-regulatory elements, such as transcription factor (TF) binding motifs [4]. As expected, the conserved TF binding motifs from the PlantPAN3.0 database [27], such as the AP2/ERF, bHLH, bZIP, C2H2, HD-ZIP, MYB, NAC and TCP motifs, were significantly enriched in the mACRs around the peak center (PC) (Fig. 3A and B). Moreover, the motif enrichment pattern was largely different between the cACRs and the sACRs, indicating distinctly different modes of TF binding for regulating the chromatin opening in different types of ACRs (Fig. 3C).

Furthermore, the signal intensities of the previously reported soybean TF ChIP-seq, including the ones performed with GmJAG1, GmAREB3, GmABI3, GmbZIP67, GmDt2 as well as GmLEC1 [17,20–22], were also enriched at the mACR peak center (PC) (Fig. 4A). G-box element, with conserved CACGTG sequence which was reported to be bound by GmLEC1 and GmbZIP67 [21,22], also could be found in the ACRs region (Fig. 4B, Table S7). These observations partially supported the hypothesis that the binding of TFs to ACRs was important for transcriptional regulation.

Moreover, since the TF binding motifs in ACRs were essential for TF binding and the subsequent gene activation, alterations in the TF binding motif could lead to changes in plant traits. Thus, variations in ACRs might play a potential role in soybean domestication, which often involved substantial modifications in the plant physiology and morphology [18,19,33]. Indeed, using the available DNA resequencing data [14,29] from the cultivated soybean C08 and wild soybean W05, we observed on average 2.8 single nucleotide polymorphisms (SNPs) or insertion/deletion variants (indels) per mACR in quantitative trait loci (QTL) regions (Fig. S8A, Table S8). Particularly, the SNPs/indels affecting the intactness of the TF binding motifs were associated with several known domesticated genes, such as *GmGIa* [34], *GmSGR1* [35], *E3* [36] and *E4* [37] (Fig. S8B). These observations suggest a potential role of ACR variations in soybean domestication. Hence, this genomic mACR information should be taken into consideration in future large-scale investigations of phenotypic association.

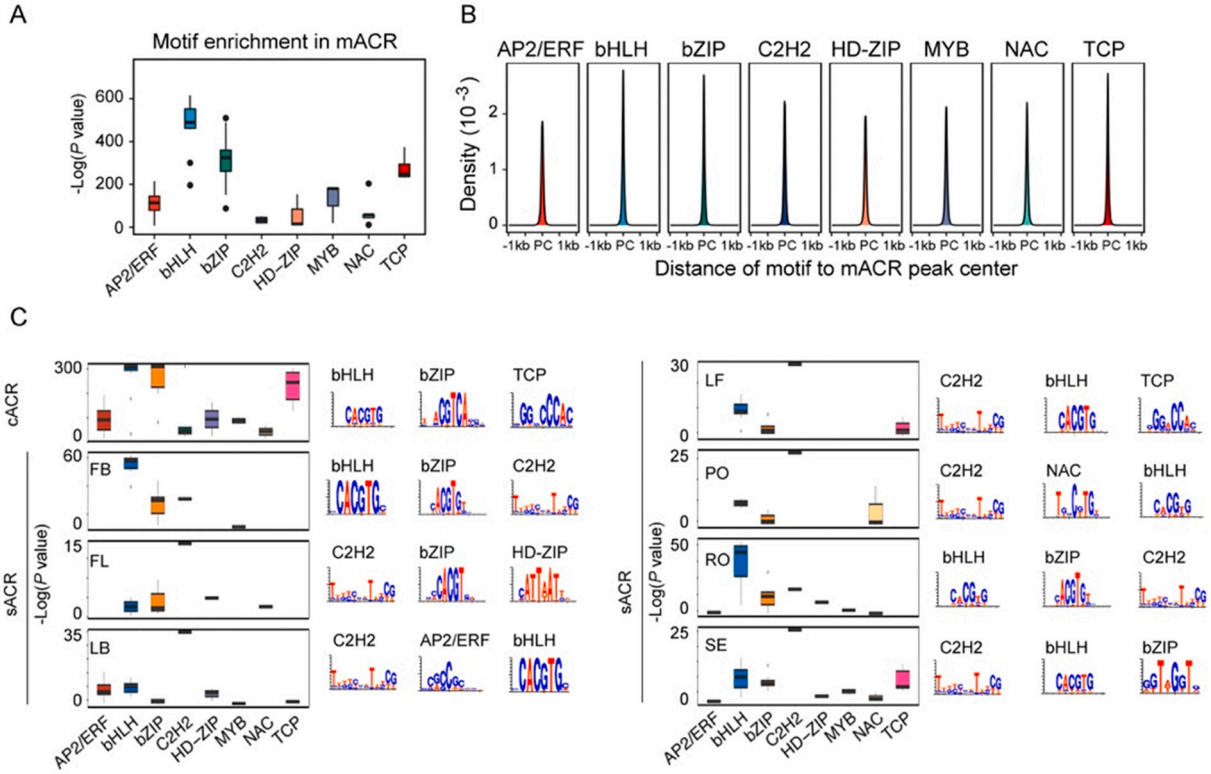

Fig. 3. Transcription factor binding motifs enriched in merged accessible chromatin regions (mACRs).
(A) The top eight families of transcription factors (TFs) with binding motifs significantly enriched in mACRs.
(B) Enrichment of the top eight TF motifs at the mACR peak center (PC).
(C) Enrichment patterns of the top eight TF motifs in the common ACR (cACR) as well as tissue-specific ACRs (sACRs) in each tissue. The top three TF motif position weight matrices (PWMs) are shown to the right of the enrichment pattern in each tissue. cACR, common ACRs that were found in all nine tissues; sACR, ACRs that were found exclusively in one specific tissue; FB, flower bud; FL, flower; LB, leaf bud; LF, leaf; PO, pod; RO, root; SE, seed.

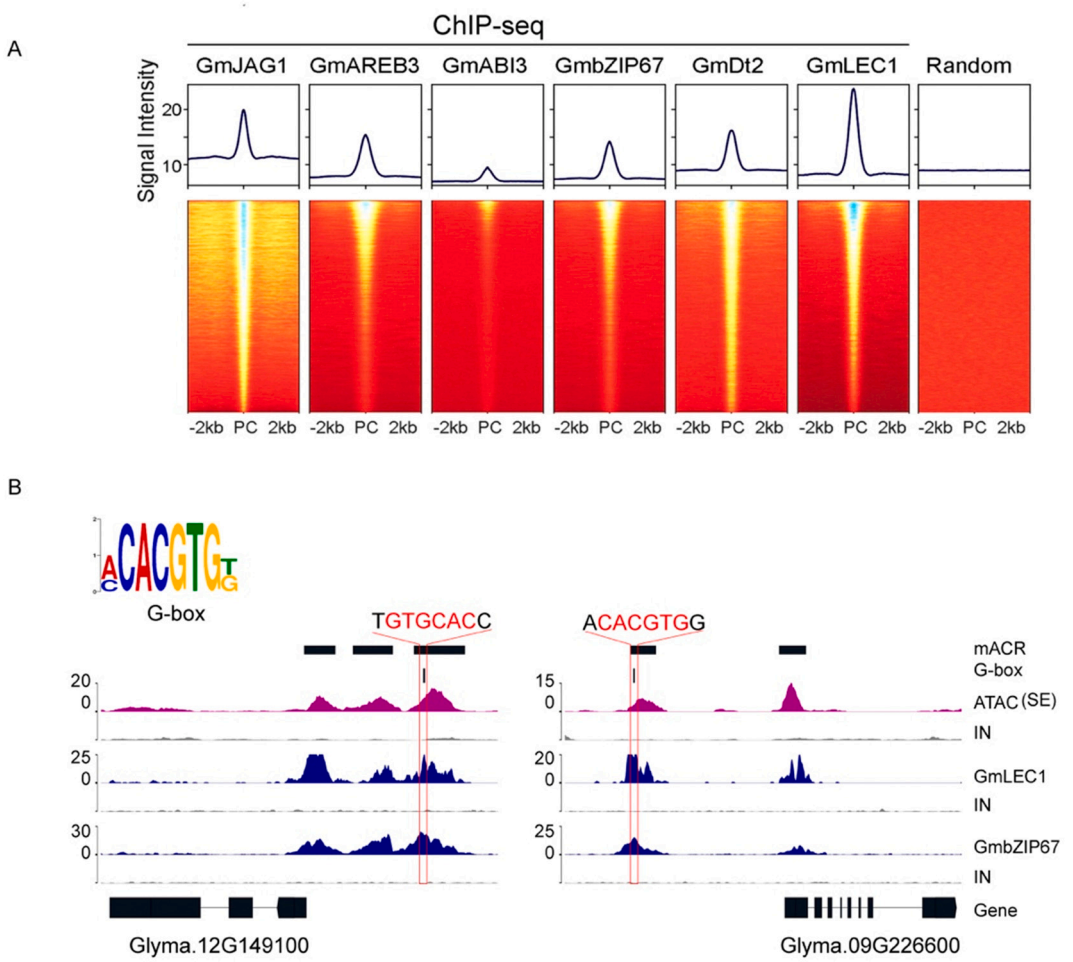

Fig. 4. TF ChIP-seq signal enriched in merged accessible chromatin regions (mACRs).
(A) Heat maps of the ChIP-seq signals of selected TFs enriched at the mACR peak center (PC) compared to the random genomic control.
(B) Examples of G-box found in the SE ACR as well as the GmLEC1 and GmbZIP67 ChIP enriched peak. IN, input control.

3.4. ACRs served as the non-coding candidate targets for genome editing in soybean

Modification of ACRs via genome editing systems has been proven as an effective way for improving the plant architecture in several crops without changing the sequencing of coding genes [11,38], thus beginning a new era in plant genetic engineering. Making use of these ACRs in each tissue, we can easily track the potential non-coding regulatory elements for each gene either in proximal or distal region. For example, we found multi-enriched ACRs peaks in the reported domestication-related genes (Fig. 5A), such as *GmGIa* [34], *GmSGR1* [35], *E3* [36], *E4* [37] and *GmJAG1* [39], which have been shown to function in flowering, seed development, etc. The proximal or distal ACRs located in each gene could be targets for genome editing systems to fine-tune their associated gene expression for potential crop improvements.

In other to facilitate further explorations in soybean research, the processed genomic information on the ACRs identified in this study is available on our website (https://datahub.wildsoydb.org)(Fig. 5B).

4. Discussion

Non-coding regulatory DNA elements are essential components shaping the entire genome architecture [6]. However, a genome-wide annotation of such important elements has been unavailable in soybean. In this study, we applied ATAC-seq to identify the main ACRs present in different tissues in soybean, which were the representative non-coding regulatory elements comprising about 3.3% of the assembled genome. Through integrative analyses with other high-throughput data, we found that ACRs were tightly associated with gene regulation in different tissue types and would dynamically change to help regulate gene expressions under different environmental conditions. Together, these data provided an overview on ACRs as the indicators of the locations of regulatory *cis* elements in the soybean genome architecture.

Recently, with advanced technologies, such as long-read sequencing and high-throughput chromosome conformation capture (Hi-C), several reference-grade cultivated and wild soybean genome assemblies has been reported [12,14–16]. For example, the pan-genome study of soybean reported 26 representative soybean assemblies [12]. It is a significant achievement that the huge amount of information on genomic variations, such as SNPs, indels and structure variations, was well interpreted and linked to the functional genes as well as the associated agronomic traits [12]. However, a large-scale association analysis between ACR variations and traits has been lacking, largely due to the absence of ACR information for soybean. According to a recent report [18] and the data provided in this study (Fig. 2), ACR presence/absence in hGenes might play an important role during soybean domestication and after polyploidization by affecting the expressions of ACR-associated genes. In addition, although a high alignment rate of mACRs from this study to the other released cultivated or wild soybean genomes was observed (Fig. S5), about 0.3–8.3% of mACRs were unable to be mapped to particular soybean genomes, indicating that the divergence of ACRs is commonplace across soybean species. Despite the

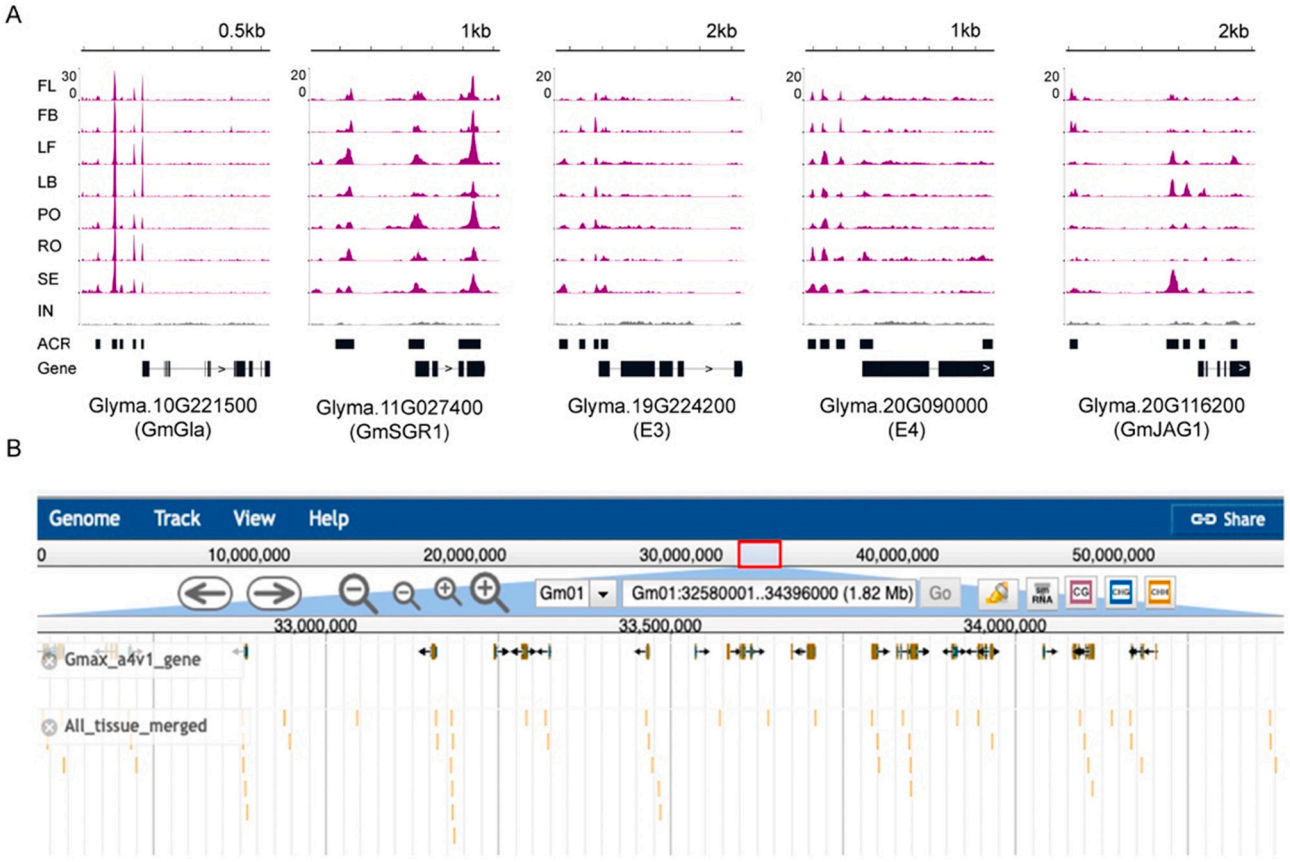

Fig. 5. Examples of the ATAC-seq data.
(A) Five examples of ACRs-associated key genes in soybean. FB, flower bud; FL, flower; LB, leaf bud; SE, seed; PO, pod; LF, leaf; IN, ATAC input control.
(B) A screenshot of the JBrowse for the genomic information of ACRs in each tissue.

current data supporting the hypothesis that these variations in ACRs could play roles during soybean domestication and after polyploidization, to our knowledge, no association between soybean ACRs and any important agronomic traits has been reported, probably due to the lack of genomic information on ACRs. Thus, the ACRs in different tissues identified in this study could facilitate association studies between large-scale phenotypic and genomic data to investigate whether ACRs have played any important role during human selection.

In addition, root nodules are special tissues formed during symbiosis with nitrogen-fixing bacteria, such as rhizobia, in most legumes including soybean [40]. The nitrogen-fixing ability of rhizobia varies among soybean germplasms and is linked to several QTL regions in soybean [41,42]. However, in this study, the current ATAC-seq method was not able to predict the nodule ACRs in soybean, because the conventional procedures for isolating nuclei cannot exclude rhizobial genomic DNA, which strongly competes with the plant genomic DNA during Tn5 transposase tagmentation. Identification of the ACRs in nodule might be achieved through another optimized method in the future and such information can then be used for phenotypic association analyses.

Since genome editing of the non-coding region provides a new approach for the crop improvement [11,38], understanding the regulatory role of ACRs as well as their associated genes in soybean is valuable for guiding soybean trait improvements using this technology. In addition, single-cell ATAC-seq (scATAC-seq) was recently used to generate a *cis*-regulatory element (CRE) atlas at the single-cell resolution in six organs in maize [43], supporting the claim that non-coding CREs were essential for both tissue and cell differentiation. A detailed map of *cis*-regulatory regions allowed genome editing at the single-cell resolution for agronomic trait improvements [43]. Although ACRs in the major tissues in soybean were identified in this study, the application of scATAC-seq in soybean to trace the dynamic changes in ACRs at the single-cell level has yet to be carried out. The information thus generated would help us decipher the role of those CREs in the different cell types of the meristem, which determines the final plant architecture.

Besides, with better understanding of the regulatory role of ACRs in different tissues in soybean, new strategies for biotechnological applications can be developed. For example, ACRs, especially the tissue-specific ones located in the core promoter or distal region, could be used for developing a new promoter or enhancer specifically for driving a gene candidate in the target tissue while avoiding undesirable side effects in other tissues.

Taken together, the genomic information on ACRs in soybean provided by this study would pave the way for deeper understanding on how non-coding *cis*-elements shape the soybean genome architecture.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110364.

Accession numbers

The raw data in this study have been submitted to NCBI under the BioProject accession number: PRJNA751745.

Author contributions

HML coordinated this research and designed the details of the experiments. MKH, LZ, LMZ and ZLW performed the wet-lab experiments and ZXX, QWW, XW conducted the bioinformatics analyses. MKH, MWL, WSY and HML co-wrote the manuscript.

Declaration of Competing Interest

There are no interests to declare.

Acknowledgments

This research was supported by the Hong Kong Research Grants Council Area of Excellence Scheme (grant number AoE/M-403/16), Lo Kwee-Seong Biomedical Research Fund awarded to H.-M.L and the Lushan Botanical Garden Research Funds (2021ZWZX31 and 2022ZWZX01). M.H. was supported by the Impact Postdoctoral Fellowship Scheme of The Chinese University of Hong Kong. Ms. Jee Yan Chu copy-edited the manuscript. Any opinions, findings, conclusions or recommendations expressed in this publication do not reflect the views of the Government of the Hong Kong Special Administrative Region or the Innovation and Technology Commission.

References

[1] J.D. Buenrostro, P.G. Giresi, L.C. Zaba, H.Y. Chang, W.J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position, Nat. Methods 10 (2013) 1213–1218.
[2] O. Wilkins, C. Hafemeister, A. Plessis, M.M. Holloway-Phillips, G.M. Pham, A. B. Nicotra, et al., EGRINs (environmental gene regulatory influence networks) in Rice that function in the response to water deficit, high temperature, and agricultural environments, Plant Cell 28 (2016) 2365–2384.
[3] Z. Lu, B.T. Hofmeister, C. Vollmers, R.M. DuBois, R.J. Schmitz, Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes, Nucleic Acids Res. 45 (2016) e41-e.
[4] S.L. Klemm, Z. Shipony, W.J. Greenleaf, Chromatin accessibility and the regulatory epigenome, Nat. Rev. Genet. 20 (2019) 207–220.
[5] Z. Lu, A.P. Marand, W.A. Ricci, C.L. Ethridge, X. Zhang, R.J. Schmitz, The prevalence, evolution and chromatin signatures of plant regulatory elements, Nat Plants 5 (2019) 1250–1259.
[6] E.P. Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74.
[7] M.A. Schaub, A.P. Boyle, A. Kundaje, S. Batzoglou, M. Snyder, Linking disease associations with regulatory information in the human genome, Genome Res. 22 (2012) 1748–1759.
[8] A. Studer, Q. Zhao, J. Ross-Ibarra, J. Doebley, Identification of a functional transposon insertion in the maize domestication gene tb1, Nat. Genet. 43 (2011) 1160–1163.
[9] W.A. Ricci, Z. Lu, L. Ji, A.P. Marand, C.L. Ethridge, N.G. Murphy, et al., Widespread long-range cis-regulatory elements in the maize genome, Nat Plants 5 (2019) 1237–1249.
[10] Y. Liu, H. Wang, Z. Jiang, W. Wang, R. Xu, Q. Wang, et al., Genomic basis of geographical adaptation to soil nitrogen in rice, Nature 590 (2021) 600–605.
[11] A. Hendelman, S. Zebell, D. Rodriguez-Leal, N. Dukler, G. Robitaille, X. Wu, et al., Conserved pleiotropy of an ancient plant homeobox gene uncovered by cis-regulatory dissection, Cell 184 (2021) 1724–1739.
[12] Y. Liu, H. Du, P. Li, Y. Shen, H. Peng, S. Liu, et al., Pan-genome of wild and cultivated soybeans, Cell 182 (2020) 162–176 (e13).
[13] J. Schmutz, S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, et al., Genome sequence of the palaeopolyploid soybean, Nature 463 (2010) 178–183.
[14] M. Xie, C.-Y. Chung, M.-W. Li, F.-L. Wong, X. Wang, A. Liu, et al., A reference-grade wild soybean genome, Nat. Commun. 10 (2019) 1216.
[15] Y. Shen, H. Du, Y. Liu, L. Ni, Z. Wang, C. Liang, et al., Update soybean Zhonghuang 13 genome to a golden reference, Sci. China Life Sci. 62 (2019) 1257–1260.
[16] B. Valliyodan, S.B. Cannon, P.E. Bayer, S. Shu, A.V. Brown, L. Ren, et al., Construction and comparison of three reference-quality genome assemblies for soybean, Plant J. 100 (2019) 1066–1082.
[17] M. Huang, L. Zhang, L. Zhou, M. Wang, W.-S. Yung, Z. Wang, et al., An expedient survey and characterization of the soybean JAGGED 1 (GmJAG1) transcription factor binding preference in the soybean genome by modified ChIPmentation on soybean protoplasts, Genomics 113 (2021) 344–355.
[18] L. Wang, G. Jia, X. Jiang, S. Cao, Z.J. Chen, Q. Song, Altered chromatin architecture and gene expression during polyploidization and domestication of soybean, Plant Cell 33 (2021) 1430–1446.
[19] M.-K. Huang, L. Zhang, L.-M. Zhou, W.-S. Yung, M.-W. Li, H.-M. Lam, Genomic features of open chromatin regions (OCRs) in wild soybean and their effects on gene expressions, Genes 12 (2021).
[20] D. Zhang, X. Wang, S. Li, C. Wang, M.J. Gosney, M.V. Mickelbart, et al., A post-domestication mutation, Dt2, triggers systemic modification of divergent and convergent pathways modulating multiple agronomic traits in soybean, Mol. Plant 12 (2019) 1366–1382.
[21] J.M. Pelletier, R.W. Kwong, S. Park, B.H. Le, R. Baden, A. Cagliari, et al., LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development, Proc. Natl. Acad. Sci. U. S. A. 114 (2017). E6710-E9.
[22] L. Jo, J.M. Pelletier, S.W. Hsu, R. Baden, R.B. Goldberg, J.J. Harada, Combinatorial interactions of the LEC1 transcription factor specify diverse developmental programs during soybean seed development, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 1223–1232.
[23] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with bowtie 2, Nat. Methods 9 (2012) 357–359.
[24] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, et al., Model-based analysis of ChIP-Seq (MACS), Genome Biol. 9 (2008) R137.
[25] F. Ramirez, F. Dundar, S. Diehl, B.A. Gruning, T. Manke, deepTools: a flexible platform for exploring deep-sequencing data, Nucleic Acids Res. 42 (2014) W187–W191.
[26] T.L. Bailey, M. Boden, F.A. Buske, M. Frith, C.E. Grant, L. Clementi, et al., MEME SUITE: tools for motif discovery and searching, Nucleic Acids Res. 37 (2009) W202–W208.
[27] C.N. Chow, T.Y. Lee, Y.C. Hung, G.Z. Li, K.C. Tseng, Y.H. Liu, et al., PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants, Nucleic Acids Res. 47 (2019). D1155-D63.
[28] G. Yu, L.G. Wang, Q.Y. He, ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization, Bioinformatics 31 (2015) 2382–2383.
[29] H.-M. Lam, X. Xu, X. Liu, W. Chen, G. Yang, F.-L. Wong, et al., Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection, Nat. Genet. 42 (2010) 1053–1059.
[30] H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[31] K. Swaminathan, K. Peterson, T. Jack, The plant B3 superfamily, Trends Plant Sci. 13 (2008) 647–655.
[32] P.J. Rushton, I.E. Somssich, P. Ringler, Q.J. Shen, WRKY transcription factors, Trends Plant Sci. 15 (2010) 247–258.
[33] M. Huang, M.-W. Li, H.-M. Lam, How noncoding open chromatin regions shape soybean domestication, Trends Plant Sci. 26 (2021) 876–878.
[34] S. Watanabe, Z. Xia, R. Hideshima, Y. Tsubokura, S. Sato, N. Yamanaka, et al., A map-based cloning strategy employing a residual heterozygous line reveals that the GIGANTEA gene is involved in soybean maturity and flowering, Genetics 188 (2011) 395–407.
[35] S.Y. Shi, H.Y. Miao, X.M. Du, J.T. Gu, K. Xiao, GmSGR1, a stay-green gene in soybean (Glycine max L.), plays an important role in regulating early leaf-yellowing phenotype and plant productivity under nitrogen deprivation, Acta Physiol. Plant. 38 (2016).
[36] S. Watanabe, R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, Y. Nakamoto, et al., Map-based cloning of the gene associated with the soybean maturity locus E3, Genetics 182 (2009) 1251–1262.
[37] B. Liu, A. Kanazawa, H. Matsumura, R. Takahashi, K. Harada, J. Abe, Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome a gene, Genetics 180 (2008) 995–1007.
[38] L. Liu, J. Gallagher, E.D. Arevalo, R. Chen, T. Skopelitis, Q. Wu, et al., Enhancing grain-yield-related traits by CRISPR-Cas9 promoter editing of maize CLE genes, Nat Plants 7 (2021) 287–294.
[39] N. Jeong, S.J. Suh, M.H. Kim, S. Lee, J.K. Moon, H.S. Kim, et al., Ln is a key regulator of leaflet shape and number of seeds per pod in soybean, Plant Cell 24 (2012) 4807–4818.
[40] M. Crespi, S. Galvez, Molecular mechanisms in root nodule development, J. Plant Growth Regul. 19 (2000) 155–166.
[41] J. Zhu, J. Wang, Q. Li, J. Wang, Y. Liu, J. Li, et al., QTL analysis of nodule traits and the identification of loci interacting with the type III secretion system in soybean, Mol. Gen. Genomics. 294 (2019) 1049–1058.
[42] B. Zhang, M. Wang, Y. Sun, P. Zhao, C. Liu, K. Qing, et al., Glycine max NNL1 restricts symbiotic compatibility with widely distributed bradyrhizobia via root hair infection, Nat Plants 7 (2021) 73–86.
[43] A.P. Marand, Z. Chen, A. Gallavotti, R.J. Schmitz, A cis-regulatory atlas in maize at single-cell resolution, Cell 184 (2021) 3041–3055 (e21).