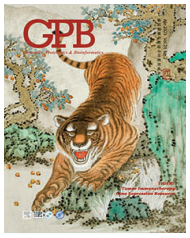
Genomics Proteomics Bioinformatics 21 (2023) 311–323

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



DATABASE

PlantCADB: A Comprehensive Plant Chromatin Accessibility Database

Ke Ding [1,2,#], Shanwen Sun [3,#], Yang Luo [2], Chaoyue Long [2], Jingwen Zhai [2], Yixiao Zhai [2], Guohua Wang [1,2,*]

[1] *State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, Harbin 150040, China*
[2] *College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China*
[3] *College of Life Science, Northeast Forestry University, Harbin 150040, China*

Received 8 January 2022; revised 25 September 2022; accepted 24 October 2022
Available online 31 October 2022

Handled by Peng Cui

KEYWORDS

Chromatin accessibility;
Plant;
Transcription factor footprint;
Regulatory network;
Stress response

Abstract **Chromatin accessibility** landscapes are essential for detecting regulatory elements, illustrating the corresponding **regulatory networks**, and, ultimately, understanding the molecular basis underlying key biological processes. With the advancement of sequencing technologies, a large volume of chromatin accessibility data has been accumulated and integrated for humans and other mammals. These data have greatly advanced the study of disease pathogenesis, cancer survival prognosis, and tissue development. To advance the understanding of molecular mechanisms regulating **plant** key traits and biological processes, we developed a comprehensive plant chromatin accessibility database (PlantCADB) from 649 samples of 37 species. These samples are abiotic stress-related (such as heat, cold, drought, and salt; 159 samples), development-related (232 samples), and/or tissue-specific (376 samples). Overall, 18,339,426 accessible chromatin regions (ACRs) were compiled. These ACRs were annotated with genomic information, associated genes, **transcription factor footprint**, motif, and single-nucleotide polymorphisms (SNPs). Additionally, PlantCADB provides various tools to visualize ACRs and corresponding annotations. It thus forms an integrated, annotated, and analyzed plant-related chromatin accessibility resource, which can aid in better understanding genetic regulatory networks underlying development, important traits, stress adaptations, and evolution. PlantCADB is freely available at https://bioinfor.nefu.edu.cn/PlantCADB/.

Introduction

Eukaryotic chromatin is an ununiformly compacted complex of DNA and proteins. Its physical compactness is referred to as chromatin accessibility, which is determined by nucleosome occupancy, topological structure, posttranslational chemical

* Corresponding author.
E-mail: ghwang@nefu.edu.cn (Wang G).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.10.005

modifications, and other chromatin binding factors. Less condensed chromatin forms accessible chromatin regions (ACRs) on the genome that can be contacted by nuclear macromolecules. Previous studies have reported that although these regions account for only 2%–3% of the total DNA sequence in humans, they accommodate important *cis*-regulatory elements that capture about 94% of all the encyclopedia of DNA elements (ENCODE) transcription factor binding sites (TFBSs) [1]. Mapping chromatin open landscape on a genome-wide scale is thus vital for detecting *cis*-regulatory elements and understanding the regulation of important biological processes [2,3]. For example, Pajoro et al. constructed the dynamic regulatory networks during *Arabidopsis* flower development by monitoring the changes in chromatin accessibility and gene expression [4]. This work helps to illustrate the mechanisms of MCM1/AGAMOUS/DEFICIENS/SRF (MADS)-domain TFs, the well-known master regulators, to regulate development and organ specification [4]. Moreover, by mapping the ACRs of control and cold-stressed samples of leaf, stem, and root tissues from three kinds of grass, Han et al. found a significant enrichment of cold-induced ACRs adjacent to cold-responsive genes and the high conservation of TF-binding motifs embedded in these regions, suggesting that common TFs may regulate the transcriptional adaptation to cold stresses across species [5]. In *Arabidopsis*, the phosphorylation of defective kernel 3 (DEK3), a chromatin architectural protein, was found to alter nucleosome occupancy and chromatin accessibility and lead to changes in gene expression, ultimately promoting salt stress tolerance [6]. Together, these results illustrate the importance of chromatin accessibility data in addressing key issues related to the molecular regulation of biological processes, development, and stress adaptations.

Experimental methods to identify such chromatin accessibility regions throughout the genome rely on combining enzymatic digestion of nuclear DNA and high-throughput sequencing, including DNase I hypersensitivity sequencing (DNase-seq) [7,8], micrococcal nuclease sequencing (MNase-seq) [9], assay for targeting accessible-chromatin with sequencing (ATAC-seq) [10,11], and formaldehyde-assisted isolation of regulatory element sequencing (FAIRE-seq) [12]. The main idea of DNase-seq and MNase-seq is to use enzymes to cut DNA double strands, and the sequencing result is accessible regions of chromatin. They are widely used in the analysis of cell-specific chromatin accessibility and to investigate the relationship between chromatin accessibility and gene expression [8,9]. ATAC-seq detects the regions bound by TFs or occupied by nucleosomes. It is faster and more sensitive than DNase-seq and MNase-seq [13,14]. FAIRE-seq overcomes the enzymatic cleavage preference that may be present in the aforementioned methods and directly detects DNA sequences occupied by nucleosomes [12]. In general, these methods can accurately and sensitively reflect the open landscape of chromatin.

With declining sequencing costs and the development of easy-to-use library construction tools, research on chromatin accessibility in humans, other mammals, and plants has become very mature. To fully exploit the large volume of chromatin accessibility data, several databases have been offered to the public. For instance, the cistrome data browser compiled chromatin accessibility data in humans and mice with homeopathic regulatory information [15]. It helped to illustrate the regulatory relationship and survival prognosis in human cancer [16,17]. The online database 'brain open chromatin atlas (BOCA)' provides an accessible atlas of human brain chromatin [18]. This database has greatly advanced research on Alzheimer's disease, neuropsychiatric disorders, and human brain development [19–22]. Chen et al. further developed OpenAnnotate to assess the chromatin accessibility of large-scale genomic regions based on features extracted from public chromatin accessibility data [23]. It builds a more comprehensive perspective to understand regulatory mechanisms in humans and mice [23]. These cases demonstrate the power of compiled chromatin accessibility database to facilitate the understanding of gene regulation in mammals and accelerate the study of pathogenesis in diseases.

Deeper analyses of chromatin accessibility data are of paramount importance for plants as well. For example, Tannenbaum et al. found that performing motif enrichment analysis on root-specific accessible regions enabled them to discover root-specific TFs that are related to root development [24]. By integrating TF chromatin immunoprecipitation sequencing (ChIP-seq) and ATAC-seq data, Tu et al. found that TF co-binding could be key for transcriptional regulation of *Zea mays*, which was conducive to the rapid diversification of the regulatory network during speciation [25]. Moreover, chromatin accessibility is crucial to illustrating how genetic variation, such as single-nucleotide polymorphisms (SNPs) in non-coding regions, leads to plant functions, adaptation, and ultimately evolution. However, public databases on plant chromatin accessibility and its subsequent analyses are missing. To provide a comprehensive chromatin accessibility data analysis platform for plants, we developed a plant chromatin accessibility database (PlantCADB). It compiled a large number of available open chromatin landscape resources and annotated their potential roles in regulation. In total, 18,339,426 ACRs from 649 samples of 37 species are available in the database (**Table 1**). Among them, 159 samples are abiotic stress-related (such as heat, cold, drought, and salt), 232 are development-related, and 376 are tissue-specific. ACR genome annotation, associated gene annotation, SNP annotation, TF footprint analysis, and motif scanning analysis are offered. Users can also perform data search, region visualization, ACR difference analysis, and overlap analysis on the web page. It additionally provides data quality control analysis, statistics, and download functions. These characteristics form integrated, annotated, and analyzed chromatin accessibility information, which can aid in better understanding the molecular mechanisms underlying key traits, biological processes, development, and stress adaptations in plants.

Data collection and database construction

Data collection

To make full use of publicly available large-scale sequencing data, we manually collected all ATAC-seq, DNase-seq, MNase-seq, and FAIRE-seq data related to plants from Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) databases from the National Center for Biotechnology Information (NCBI) [26]. These sequences are from 37 species and 19 tissues (Table 1). The reference genomes and corresponding genes of each species were downloaded from NCBI genome (https://www.ncbi.nlm.nih.gov/genome/) and Ensembl plants [27,28].

Table 1 Chromatin accessibility data summary

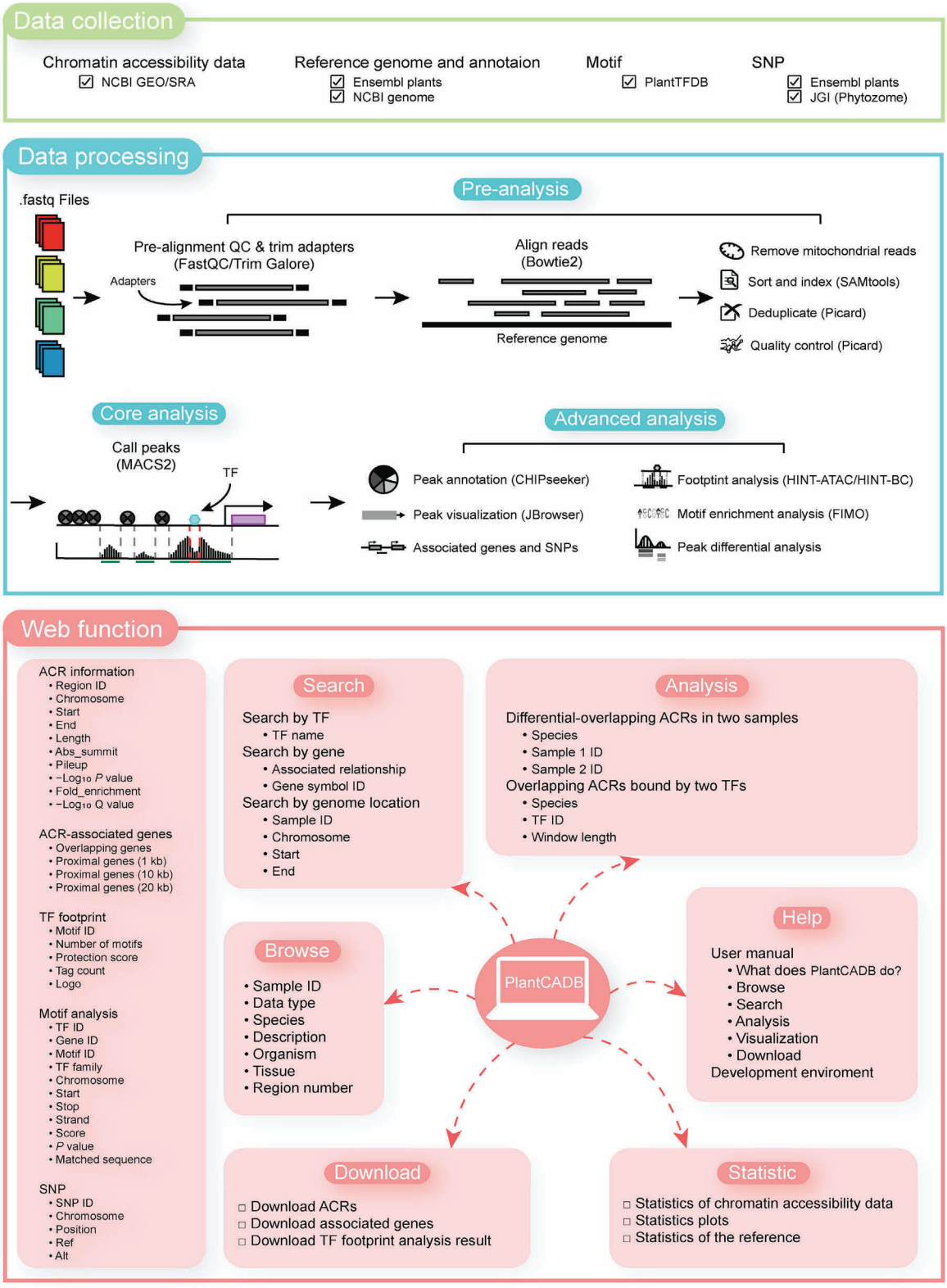
Class	Species	ATAC-seq		DNase-seq		FAIRE -seq	MNase-seq		Total
		GEO	SRA	GEO	SRA	GEO	GEO	SRA	
Dicotyledoneae	*Arabidopsis thaliana*	85	27	48	12	15	3		190
	Solanum lycopersicum	16			77				93
	Cucumis melo				30				30
	Prunus persica				21				21
	Populus trichocarpa	18							18
	Pyrus x bretschneideri				14				14
	Citrullus lanatus				10				10
	Cucumis sativus				10				10
	Medicago truncatula	9							9
	Solanum pennellii	8							8
	Vitis vinifera				8				8
	Glycine max	7							7
	Malus domestica				7				7
	Fragaria vesca				7				7
	Solanum phureja				6				6
	Carica papaya				6				6
	Arachis hypogaea		4						4
	Eutrema salsugineum	3							3
	Phaseolus vulgaris	3							3
	Gossypium arboreum				1				1
	Gossypium barbadense				1				1
	Gossypium hirsutum							1	1
	Gossypium raimondii				1				1
	Eucalyptus grandis				12				12
Monocotyledoneae	*Zea mays*	23	13	11	2				49
	Oryza sativa	58	2						60
	Sorghum bicolor	3	3	6					12
	Ananas comosus				12				12
	Setaria italica		2	6					8
	Brachypodium distachyon	3		3					6
	Musa acuminata				6				6
	Hordeum vulgare	3							3
	Setaria viridis	3							3
	Asparagus officinalis	3							3
	Spirodela polyrhiza	3							3
	Triticum aestivum	1							1
Hepaticae	*Marchantia polymorpha*		2						2
Total		249	53	74	243	15	3	1	638

Note: ATAC-seq, assay for targeting accessible-chromatin with sequencing; DNase-seq, DNase I hypersensitivity sequencing; FAIRE-seq, formaldehyde-assisted isolation of regulatory element sequencing; MNase-seq, micrococcal nuclease sequencing; GEO, Gene Expression Omnibus; SRA, Sequence Read Archive.

Data preprocessing

The downloaded raw sequencing data were first converted into FASTQ format files using fastq-dump in NCBI SRA toolkit (version 2.9.2). FastQC [29] was used to examine sequence quality, GC content, sequence length distribution, sequence duplication levels, overrepresentation sequences, and contamination of adapters in the raw sequencing data for pre-alignment quality control (QC; **Figure 1**). Adapters and low-quality reads were removed using Trim Galore (version 0.6.6; Figure 1) with the following parameters '-q 20 --phred33 --stringency 3 --length 20 -e 0.1'. After that, sequences were mapped to the reference genome (Table S1) using Bowtie2 (version 2.4.2; Figure 1) [30] with default parameters.

To improve the power of open chromatin detection and produce fewer false positives, we performed a series of post-alignment processing (Figure 1). First, mitochondrial sequences were excluded because no chromatin packaging exists [31]. We then sorted and converted files to .bam format using SAMtools (version 1.11) [32]. The low-quality reads and redundancies generated during the polymerase chain reaction (PCR) library building process were removed using Picard (version 2.25.4; https://broadinstitute.github.io/picard/). To ensure the correctness and rationality of biological conclusions, additional quality indicators were evaluated, including mean insert size, corresponding standard deviation, transcription start site (TSS) scanning score, and the fraction of reads in peaks (FRiP) score. The mean insertion size and corresponding standard deviation of the paired-end were calculated with the Picard tool. TSS scanning score and FRiP were calculated with our own codes. Based on the QC distribution of the population sample, we established the QC characteristic threshold to filter a few samples with poor quality.


Data collection
Chromatin accessibility data
NCBI GEO/SRA
Reference genome and annotaion
Ensembl plants
NCBI genome
Motif
PlantTFDB
SNP
Ensembl plants
JGI (Phytozome)
Data processing
Pre-analysis
.fastq Files
Pre-alignment QC & trim adapters
(FastQC/Trim Galore)
Adapters
Align reads
(Bowtie2)
Reference genome
Remove mitochondrial reads
Sort and index (SAMtools)
Deduplicate (Picard)
Quality control (Picard)
Core analysis
Call peaks
(MACS2)
TF
Advanced analysis
Peak annotation (CHIPseeker)
Peak visualization (JBrowser)
Associated genes and SNPs
Footptint analysis (HINT-ATAC/HINT-BC)
Motif enrichment analysis (FIMO)
Peak differential analysis
Web function
ACR information
• Region ID
• Chromosome
• Start
• End
• Length
• Abs_summit
• Pileup
• −Log10 P value
• Fold_enrichment
• −Log10 Q value
ACR-associated genes
• Overlapping genes
• Proximal genes (1 kb)
• Proximal genes (10 kb)
• Proximal genes (20 kb)
TF footprint
• Motif ID
• Number of motifs
• Protection score
• Tag count
• Logo
Motif analysis
• TF ID
• Gene ID
• Motif ID
• TF family
• Chromosome
• Start
• Stop
• Strand
• Score
• P value
• Matched sequence
SNP
• SNP ID
• Chromosome
• Position
• Ref
• Alt
Search
Search by TF
• TF name
Search by gene
• Associated relationship
• Gene symbol ID
Search by genome location
• Sample ID
• Chromosome
• Start
• End
Analysis
Differential-overlapping ACRs in two samples
• Species
• Sample 1 ID
• Sample 2 ID
Overlapping ACRs bound by two TFs
• Species
• TF ID
• Window length
Browse
• Sample ID
• Data type
• Species
• Description
• Organism
• Tissue
• Region number
PlantCADB
Help
User manual
• What does PlantCADB do?
• Browse
• Search
• Analysis
• Visualization
• Download
Development enviroment
Download
□ Download ACRs
□ Download associated genes
□ Download TF footprint analysis result
Statistic
□ Statistics of chromatin accessibility data
□ Statistics plots
□ Statistics of the reference

Identification and classification of ACRs

ACRs were identified using MACS2 (version 2.2.7.1; Figure 1) [33,34] based on peak calling with the parameter '--nomodel --shift -100 --extsize 200 -g 1.2e8' [35–37]. The '-g' parameter is an effective genome size and was uniquely adjusted for each species. The example presents the '-g' setting for *Arabidopsis thaliana*. These steps identified 18,065,954 ACRs from 638 samples. To assess the reliability of our settings, we compared the results with published data. Overall, we found that these ACRs highly overlapped with the validated ACRs (overlap rate from 0.77 to 0.99; Table S2), suggesting that the peak calling for each species using the MACS2 tool is reliable. Additionally, we collected 273,472 validated ACRs from published reports [38,39].

Because the position of ACRs relative to genes is significantly divergent with different genome sizes in plants [40], we, therefore, classified them into genic (gACRs, overlapping with a gene), proximal (pACRs, within 2 kb from a gene), and distal (dACRs, distance from a gene $>$ 2 kb) according to the distance between ACRs and the nearest gene.

ACR visualization

To make it easier and more intuitive for users to compare different experimental data, we used the CHIPseeker toolset [41] to visualize ACRs for each sample, including the panorama distribution histogram of genome-wide ACRs, pie chart and combined charts of genome annotation, and heat map of ACRs near TSS. The covplot function in CHIPseeker was used to visualize the panorama of the ACR distribution, which enables users to clearly observe the location of ACRs in the whole genome. The annotatePeak function was used to analyze genome annotation, which classifies ACRs into promoter, 5'UTR, 3'UTR, 1st exon, other exons, 1st intron, other introns, downstream, and distal intergenic. The plotAnnoPie function was used to map genome annotation obtained by the AnnotatePeak function into a pie chart. We set the 1 kb upstream and downstream of TSS as the window area and used the tagHeatmap function to draw a heat map of the ACRs combined with the window area. Users can thus intuitively understand the distribution of ACRs near all gene promoters. Because ACR locations may not be unique (an ACR covers exons of one gene and, at the same time, introns of another gene), we drew the combined graph of Venn pie and upset plot using vennpie and UpSetR functions.

SNP annotation

To enable users to understand the relationship between chromatin accessibility and SNPs, we also downloaded available SNPs from Ensembl and Phytozome databases [42]. To achieve a high level of confidence in the variant calls from sequencing data, we applied a set of initial-quality filters using VCFtools (version 0.1.16) [43] to filter out low-quality SNPs. Specifically, we only kept variants with a mean read depth $\geq$ 10 to minimize spurious SNP calls due to low-coverage genomic regions and with minor allele frequency (MAF) above 0.05 [44].

Analysis of genes associated with ACRs

To further understand the function and transcriptional regulation of identified ACRs, PlantCADB identified and classified genes associated with each ACR into three categories, *i.e.*, overlapping genes, closest genes, and proximal genes. Overlapping genes are defined as the genes overlapping with each ACR. The closest genes are the ones closest to the center of each ACR. Proximal genes are defined as the genes in the upstream and downstream 1-kb area of TSS that overlap with each ACR. The associated genes were downloaded from the Ensembl plant and NCBI genome databases. The specific identification and classification method followed the steps in the ROSE script (ROSE_geneMapper.py) [45,46]. On the query interface of PlantCADB, users can search data sets and corresponding regions based on the gene identity document (ID).

Analysis of TFBSs in ACRs

Analysis of TFBSs in chromatin-accessible landscapes reflects both aggregate TF binding and the regulatory potential of a genetic locus. PlantCADB used two methods to identify TFBSs: motif scan analysis and footprint analysis. We used the MEME-suite [47] of the FIMO tool [48] to scan for a single match of the motif of each ACR in each sample [39]. Plant-related TF lists and TF-binding motifs for each species were obtained from PlantTFDB [49]. FIMO generates an ordered list of motifs as output, each with an associated log-likelihood ratio score, *P* value, and matching sequence. The log-likelihood ratio scores of each motif at each sequence position were calculated and converted into *P* values by dynamic programming. Users can set different thresholds to obtain the motif sequence.

Figure 1 Database construction and overview

PlantCADB assesses ACRs using ATAC-seq, DNase-seq, FAIRE-seq, and MNase-seq data. Genetic annotations were collected or calculated, including SNPs, TFBSs, TF footprint, and associated genes. Users can query ACRs using three strategies: TF-based query, gene-based query, and genome location-based query. PlantCADB also includes online analyzing tools and a personalized genome browser to discover the potential biological effects of ACRs. PlantCADB, plant chromatin accessibility database; ACR, accessible chromatin region; ATAC-seq, assay for targeting accessible-chromatin with sequencing; DNase-seq, DNase I hypersensitivity sequencing; FAIRE-seq, formaldehyde-assisted isolation of regulatory elements sequencing; MNase-seq, micrococcal nuclease sequencing; TF, transcription factor; TFBS, transcription factor binding site; SNP, single-nucleotide polymorphism; NCBI, National Center for Biotechnology Information; GEO, Gene Expression Omnibus; SRA, Sequence Read Archive; ID, identity document; ref, reference; alt, alternative.

Another way to decipher TF regulation rules is to use footprints. The combination of active TF and DNA will prevent the enzyme from cutting at the binding site, which is characteristic of ATAC-seq and DNase-seq experiments. This leads to the objective formation of a protected area called footprint [50]. According to the characteristics of the two sequencing technologies and current footprint analysis tools, we used different analysis software for TF footprint analysis [31]. For ATAC-seq data, there are several obstacles in footprint analysis. First, due to the 9-bp gap in the library construction process, displacement processing is required during the data handling process. In addition, Tn5 enzyme binding is biased, and the transient binding signal of TF is relatively weak. Among the existing ATAC-seq footprint analysis software, only HINT-ATAC can correct the cleavage preference of the chain-specific Tn5 enzyme [51]. For the DNase-seq data, we used the HINT-BC tool to solve DNase-seq cleavage deviation and residence time that affect the calculation of footprint [52–54]. Both tools are based on a hidden Markov model to predict TFBS with footprints, which outperform other tools.

After selecting appropriate methods, we downloaded the position weight matrix (PWM) of the motif from the PlantTFDB database and created the regulatory genomics toolbox (RGT) data folder for each species. The folder had five files: gene, chromosome sizes, gene regions, annotation information, and gene alias. The gene alias file allows for translation between multiple different gene IDs. The PWM was matched with the reference genome of the corresponding species, which outputted protection score, tag count (TC), number of binding sites, and footprint logo combined with TF. The protection score measures the difference between enzyme digestion region counts in the flanking region and in the motif-predicted binding site. It can detect TF with a potentially short residence time [53]. TC is used to represent the number of reads near the TFBSs that are ranked by footprint prediction. We also offer threshold conditions so that users can filter data according to their own criteria.

Database and website implementation

The current version of PlantCADB is developed using Java 8 and HTML 5 and deployed to run on a Linux-based Apache web server. The website page framework was designed and constructed using Bootstrap (version 3.3.7), and the front and back data interaction was realized by JQuery (version 3.6.0). Echarts (version 3.7.0) was used to achieve data visualization. JBrowse2 browser framework was used for genomic visualization. We recommend modern web browsers that support the HTML5 standard for the best display.

Database content and usage

Statistics of PlantCADB

In the current version of PlantCADB, we collected a total of 649 samples from 37 species (**Figure 2**A; Table 1), covering four types of sequencing data from 19 tissues (Figure 2B). The plant species include angiosperms (monocotyledons and dicotyledons) and bryophytes with different genome sizes (122–14,790 Mb), genome structures, and gene densities (7.5–124 genes/Mb). Overall, 18,339,426 ACRs were identified (7,972,702 from ATAC-seq; 9,472,599 from DNase-seq; 568,255 from FAIRE-seq; and 325,870 from MNase-seq). The total sequence length of ACRs in each species ranges from 0.4 to 104.3 Mb and accounts for 0.003%–11.3% of the genome size across species. The numbers of ACRs, its total sequence length, and the percentage of the genome size occupied by all ACRs in each species do not significantly increase with the increased genome size (Spearman rank correlation $\rho_{number} = 0.19$, $\rho_{length} = 0.14$, $\rho_{percent} = -0.48$, all P values > 0.05; Figure S1). After motif scanning analysis on 624 samples of 33 species (four species without available motif data were excluded from the analysis), we found that approximately 99.1% of ACRs have the potential to be bound by TFs. After analyzing the distribution of SNPs in different sequencing data types of each species from a total of eight species for which SNP data are available, we found that the density of SNPs are significantly higher in ACRs than in the whole genome (the ratios between densities ranging from 1.43 to 25.82; Figure 2C).

We found that the number of dACRs (8,013,939) is positively correlated with increasing genome sizes (Spearman rank correlation $\rho = 0.66$, P value < 0.001; Figure 2D), whereas the numbers of gACRs (5,488,948) and pACRs (4,157,698) are both negatively correlated with increasing genome sizes (Spearman rank correlation $\rho_{pACRs} = -0.62$, $\rho_{gACRs} = -0.57$, both P values < 0.001; Figure 2D). In addition, the numbers of the three ACR types are also significantly correlated with gene density across species (Spearman rank correlation $\rho_{dACRs} = -0.73$, $\rho_{pACRs} = 0.63$, $\rho_{gACRs} = 0.62$, all P values < 0.001; Figure S2).

The web interface of PlantCADB

The search interface for retrieving ACRs

PlantCADB offers three user-friendly search options to retrieve chromatin accessibility data (**Figure 3**A). With a TF-based query ('search ACR by TF'), users can obtain all ACRs that are potentially bound by the query TF by selecting species and TF ID. With a gene-based query ('search ACR by gene'), users can obtain all ACRs associated with the gene of interest after selecting species and identification strategy (overlapping, proximal, and closest). With a genomic region-based query ('search ACR by genome location'), users can get all ACRs that overlap (at least 1 kb) with the submitted region by selecting species and sample ID and inputting genomic location.

The browsing interface for retrieving ACRs

Users can browse all ACRs belonging to a specific data type, experimental classification, layout (the construction method of sequencing library), species, or tissue (Figure 3B). The result shows samples that match the filter conditions. All samples from four sequencing technologies were named with different prefixes ('sample_00' for ATAC-seq, 'sample_01' for DNase-seq, 'sample_02' for FAIRE-seq, and 'sample_03' for MNase-seq). Users can further click the 'sample ID' to get detailed information about the sample, including sample overview, ACR result table, TF footprint annotation, and peak annotation visualization (Figure 3C–F). The sample overview provides general information, values of four QC metrics, measurement indicators, and pie charts of statistical information about the ACRs of the sample (Figure 3C). Based on two

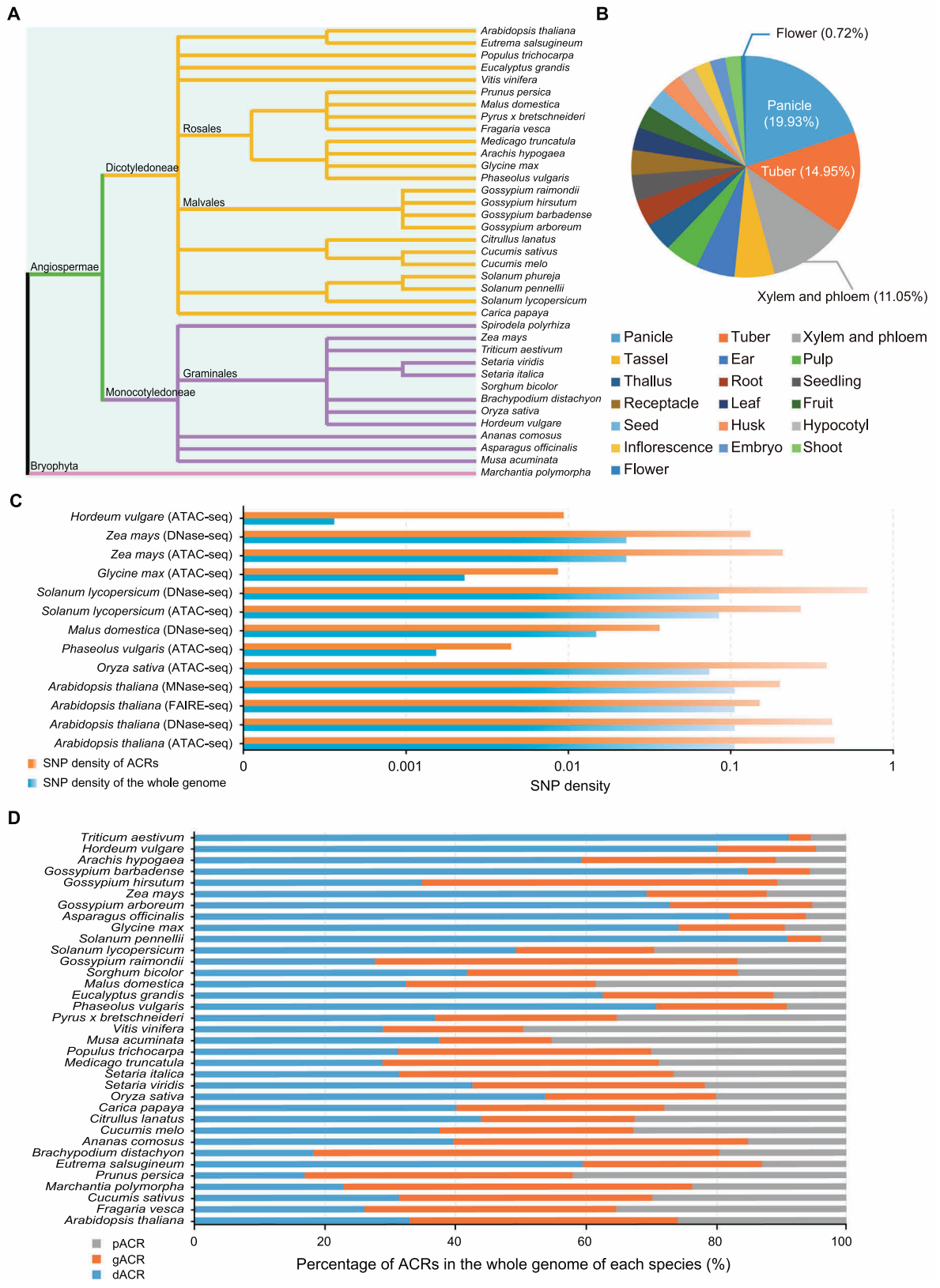
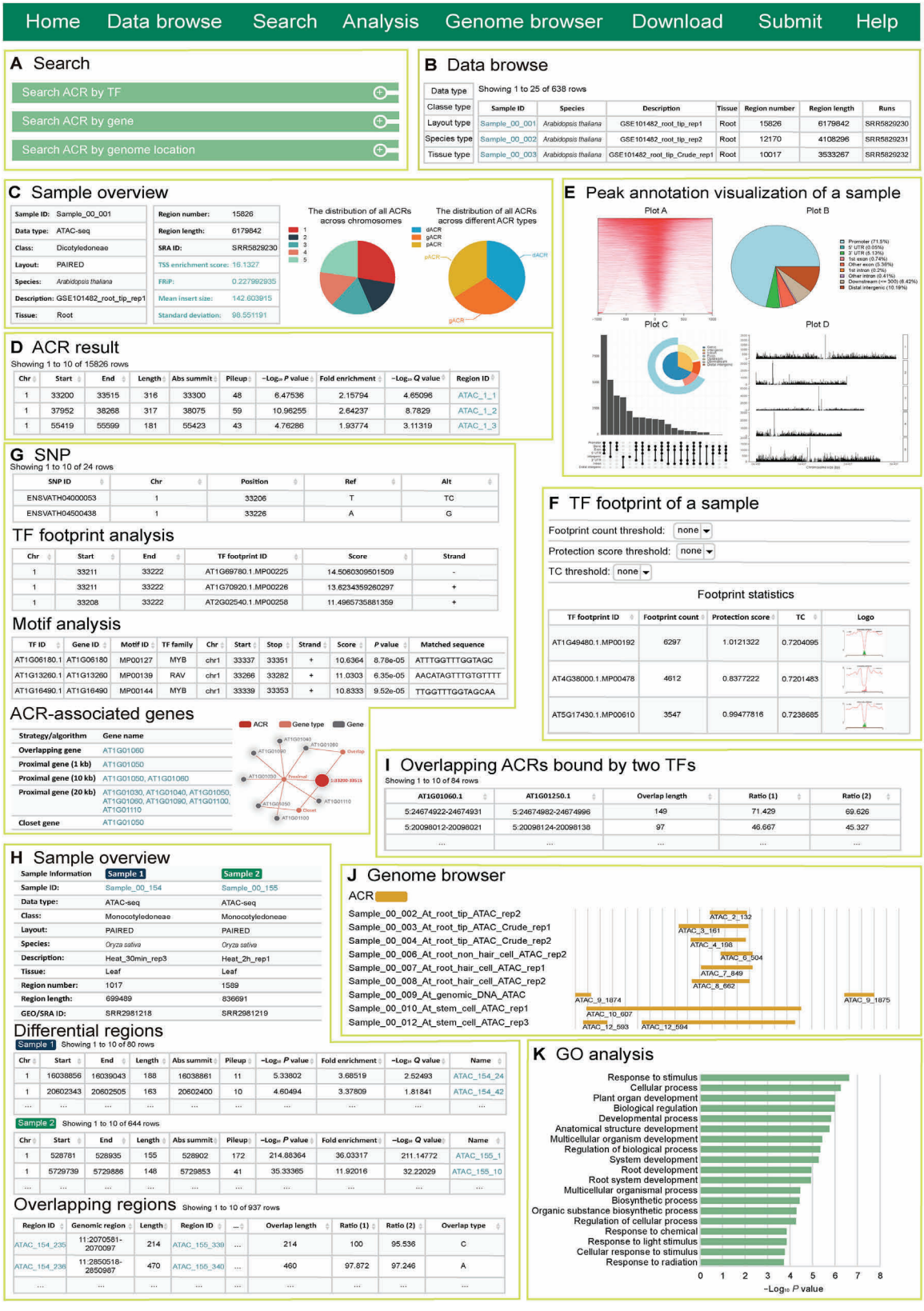


Figure 2 Statistics of PlantCADB and ACRs

A. A phylogenetic map of the plant species that were investigated. **B.** A pie chart of the average number of ACRs in different tissues. **C.** The density of SNPs in eight species in the whole genome and in ACRs. **D.** The percentage of dACRs, gACRs, and pACRs to total ACRs in each species (species are ordered based on the reference-genome size). dACR, distal accessible chromatin region; gACR, genic accessible chromatin region; pACR, proximal accessible chromatin region.


Home
Data browse
Search
Analysis
Genome browser
Download
Submit
Help
A Search
Search ACR by TF
Search ACR by gene
Search ACR by genome location
B Data browse
Data type
Classe type
Layout type
Species type
Tissue type
Showing 1 to 25 of 638 rows
Sample ID
Species
Description
Tissue
Region number
Region length
Runs
Sample_00_001
Arabidopsis thaliana
GSE101482_root_tip_rep1
Root
15826
6179842
SRR5829230
Sample_00_002
Arabidopsis thaliana
GSE101482_root_tip_rep2
Root
12170
4108296
SRR5829231
Sample_00_003
Arabidopsis thaliana
GSE101482_root_tip_Crude_rep1
Root
10017
3533267
SRR5829232
C Sample overview
Sample ID: Sample_00_001
Data type: ATAC-seq
Class: Dicotyledoneae
Layout: PAIRED
Species: Arabidopsis thaliana
Description: GSE101482_root_tip_rep1
Tissue: Root
Region number: 15826
Region length: 6179842
SRA ID: SRR5829230
TSS enrichment score: 16.1327
FRiP: 0.227992935
Mean insert size: 142.603915
Standard deviation: 98.551191
The distribution of all ACRs across chromosomes
The distribution of all ACRs across different ACR types
dACR
pACR
gACR
D ACR result
Showing 1 to 10 of 15826 rows
Chr
Start
End
Length
Abs summit
Pileup
−Log10 P value
Fold enrichment
−Log10 Q value
Region ID
1 33200 33515 316 33300 48 6.47536 2.15794 4.65096 ATAC_1_1
1 37952 38268 317 38075 59 10.96255 2.64237 8.7829 ATAC_1_2
1 55419 55599 181 55423 43 4.76286 1.93774 3.11319 ATAC_1_3
E Peak annotation visualization of a sample
Plot A
Plot B
Plot C
Plot D
G SNP
Showing 1 to 10 of 24 rows
SNP ID
Chr
Position
Ref
Alt
ENSVATH04000053 1 33206 T TC
ENSVATH04500438 1 33226 A G
TF footprint analysis
Chr
Start
End
TF footprint ID
Score
Strand
1 33211 33222 AT1G69780.1.MP00225 14.5060309501509 -
1 33211 33222 AT1G70920.1.MP00226 13.6234359260297 +
1 33208 33222 AT2G02540.1.MP00258 11.4965735881359 +
Motif analysis
TF ID
Gene ID
Motif ID
TF family
Chr
Start
Stop
Strand
Score
P value
Matched sequence
AT1G06180.1 AT1G06180 MP00127 MYB chr1 33337 33351 + 10.6364 8.78e-05 ATTTGGTTTGGTAGC
AT1G13260.1 AT1G13260 MP00139 RAV chr1 33266 33282 + 11.0303 6.35e-05 AACATAGTTTGTGTTTT
AT1G16490.1 AT1G16490 MP00144 MYB chr1 33339 33353 + 10.8333 9.52e-05 TTGGTTTGGTAGCAA
ACR-associated genes
Strategy/algorithm
Gene name
Overlapping gene AT1G01060
Proximal gene (1 kb) AT1G01050
Proximal gene (10 kb) AT1G01050, AT1G01060
Proximal gene (20 kb) AT1G01030, AT1G01040, AT1G01050, AT1G01060, AT1G01090, AT1G01100, AT1G01110
Closet gene AT1G01050
ACR
Gene type
Gene
F TF footprint of a sample
Footprint count threshold: none
Protection score threshold: none
TC threshold: none
Footprint statistics
TF footprint ID
Footprint count
Protection score
TC
Logo
AT1G49480.1.MP00192 6297 1.0121322 0.7204095
AT4G38000.1.MP00478 4612 0.8377222 0.7201483
AT5G17430.1.MP00610 3547 0.99477816 0.7238685
I Overlapping ACRs bound by two TFs
Showing 1 to 10 of 84 rows
AT1G01060.1
AT1G01250.1
Overlap length
Ratio (1)
Ratio (2)
5:24674922-24674931 5:24674982-24674996 149 71.429 69.626
5:20098012-20098021 5:20098124-20098138 97 46.667 45.327
H Sample overview
Sample information
Sample 1
Sample 2
Sample ID: Sample_00_154 Sample_00_155
Data type: ATAC-seq ATAC-seq
Class: Monocotyledoneae Monocotyledoneae
Layout: PAIRED PAIRED
Species: Oryza sativa Oryza sativa
Description: Heat_30min_rep3 Heat_2h_rep1
Tissue: Leaf Leaf
Region number: 1017 1589
Region length: 699489 836691
GEO/SRA ID: SRR2981218 SRR2981219
Differential regions
Sample 1 Showing 1 to 10 of 80 rows
Chr
Start
End
Length
Abs summit
Pileup
−Log10 P value
Fold enrichment
−Log10 Q value
Name
1 16038856 16039043 188 16038861 11 5.33802 3.68519 2.52493 ATAC_154_24
1 20602343 20602505 163 20602400 10 4.60494 3.37809 1.81841 ATAC_154_42
Sample 2 Showing 1 to 10 of 644 rows
1 528781 528935 155 528902 172 214.88364 36.03317 211.14772 ATAC_155_1
1 5729739 5729886 148 5729853 41 35.33365 11.92016 32.22029 ATAC_155_10
Overlapping regions Showing 1 to 10 of 937 rows
Region ID
Genomic region
Length
Region ID
Overlap length
Ratio (1)
Ratio (2)
Overlap type
ATAC_154_235 11:2070581-2070097 214 ATAC_155_339 214 100 95.536 C
ATAC_154_236 11:2850518-2850987 470 ATAC_155_340 460 97.872 97.246 A
J Genome browser
ACR
Sample_00_002_At_root_tip_ATAC_rep2
Sample_00_003_At_root_tip_ATAC_Crude_rep1
Sample_00_004_At_root_tip_ATAC_Crude_rep2
Sample_00_006_At_root_non_hair_cell_ATAC_rep2
Sample_00_007_At_root_hair_cell_ATAC_rep1
Sample_00_008_At_root_hair_cell_ATAC_rep2
Sample_00_009_At_genomic_DNA_ATAC
Sample_00_010_At_stem_cell_ATAC_rep1
Sample_00_012_At_stem_cell_ATAC_rep3
ATAC_2_132
ATAC_3_161
ATAC_4_198
ATAC_6_504
ATAC_7_849
ATAC_8_662
ATAC_9_1874
ATAC_9_1875
ATAC_10_607
ATAC_12_593
ATAC_12_594
K GO analysis
Response to stimulus
Cellular process
Plant organ development
Biological regulation
Developmental process
Anatomical structure development
Multicellular organism development
Regulation of biological process
System development
Root development
Root system development
Multicellular organismal process
Biosynthetic process
Organic substance biosynthetic process
Regulation of cellular process
Response to chemical
Response to light stimulus
Cellular response to stimulus
Response to radiation
0 1 2 3 4 5 6 7 8
−Log10 P value

statistical pie charts, users can view the distribution of ACRs on each chromosome and the number of various types of ACRs. The ACR result table shows all ACRs of the sample. It describes region ID, genome location, region length, summit site (abs_summit), summit height (pileup), fold enrichment, and $-\log_{10} P$ value (Figure 3D). The peak annotation visualization displays the annotation information of ACRs in different ways, including peak distribution heat maps near TSS regions, pie charts annotating genomic features, panoramic maps of ACRs in the genome distribution, and a more detailed combination chart (Figure 3E). The TF footprint annotation shows the results of the HINT software analysis. TC, protection score, number of binding sites, and footprint logo were identified for each sample. We also offer a 'threshold' option, which allows users to screen for TFs with high activity by setting thresholds (Figure 3F).

To view more detailed information on the designated ACR, we provide SNPs, TF footprint corresponding scores, the results of motif scanning, and associated genes (Figure 3G). Motif scanning includes location information, sequence scores, and matched sequences. The associated genes are further classified into overlapping genes, proximal genes within ± 1 kb, proximal genes within ± 10 kb, proximal genes within ± 20 kb, and the closest genes.

Online analysis tools

Analyzing the dynamics of chromatin accessibility can allow us to understand the changes in molecular regulation in response to developmental cues and external stimuli [55]. PlantCADB provides two online analysis tools. The first one is a differential-overlapping analysis of ACRs. In the 'analysis of differential-overlapping ACRs' panel, the user can submit 'experimental classification' and two 'sample IDs' of interest and get the analyzed differential and overlapping ACRs between the two samples (Figure 3H). For overlapping regions (at least 1-bp overlap between them), we divide them into four types. Type A indicates that the right wing of ACR in 'sample 1' overlaps at least 1 bp with the left wing of ACR in 'sample 2'. Type B is opposite to type A, indicating that the left wing of ACR in 'sample 1' overlaps the right wing of ACR in 'sample 2'. Type C means that a certain ACR of 'sample 1' is completely covered by a certain ACR of 'sample 2'. Type D is completely opposite to type C, which means that a certain ACR of 'sample 1' completely covers the ACR of 'sample 2'. We additionally provide the genomic positions of the two overlapping regions, overlap position, overlap length, and ratios of overlap (ratio of overlapping areas in each ACR). For differential regions, we define them as different regions and output these regions of the two samples, respectively.

Users can also analyze the overlapping ACRs bound by two TFs. In the 'analysis of overlapping ACRs bound by two TFs' panel, users can submit two interested 'TF name' and 'window length' (Figure 3I). The tool can fetch all regions that are bound by both TFs and calculate the two overlapping areas according to the submitted window length. The results of the analysis are briefly displayed in a table, including the genomic location of TF, length of the overlap region, and overlap rate (ratio of the overlap length to total length, where the total length = the length of the TF bound to the genome location + 2 × the length of the window).

Data visualization and personalized genome browser

To help users better view the genomic information of chromatin accessibility, we also provide a personalized genome browser, which is developed using the latest version of JBrowser2 [56]. Users can intuitively see the positional relationship between chromatin accessibility and nearby genes, mRNA, tRNA, lncRNA, rRNA, and other genomic fragments (Figure 3J). In addition, we also provide annotated pie charts and distribution maps of ACRs, and the network between ACRs and genes is drawn online using Echarts software.

Case studies

To provide an example of how regulators can be used in PlantCADB to retrieve the putative corresponding regulatory network, the basic helix–loop–helix (bHLH) TF family transcription factor myelocytomatosis 2 (MYC2, AT1G32640.1) in *Arabidopsis thaliana* is used as input to our database for 'search ACR by TF'. After clicking the 'start search' button, a total of 922 ACRs with 110 nearest neighbor genes that are potentially bound by MYC2 were retrieved. To characterize the potential functions of these associated genes, we performed a Gene Ontology (GO) enrichment analysis using PANTHER [57]. Here, we obtained 19 significant biological processes (Q value < 0.05; Figure 3K), including root development, response to light stimulus, organic substance biosynthetic process, response to chemical, cellular response to stimulus, and regulation of the cellular process. These results are consistent with previous findings, which indicate that MYC2 is an important regulator of lateral root formation

Figure 3 The main functions and usage of PlantCADB

A. Users can query ACRs in three ways: 'search ACR by TF', 'search ACR by gene', and 'search ACR by genomic location'. **B.** Browsing the sample details. **C.** Sample information including sample ID, data type, class, layout, species, description, tissue, region number, region length, SRA ID, quality control report, and a pie chart of the ACR distribution. **D.** ACR result table for a sample, including Chr, start, end, length, abs summit, pileup, $-\text{Log}_{10} P$ value, fold enrichment, $-\text{Log}_{10} Q$ value, and region ID. **E.** Peak annotation visualization of a sample. **F.** TF footprint statistics of a sample. **G.** SNP panel: the detailed information of SNP. TF footprint analysis panel: the detailed information on each region of TF analysis, including TF footprint and motif scanning. ACR-associated gene panel: potential ACR-associated genes are identified by three strategies. Their relationships are displayed using a network diagram. **H.** Analysis of differential and overlapping ACRs between two samples in the same species. **I.** Analysis of overlapping ACRs bound by two TFs in the same species. **J.** Visualization of a genome browser. **K.** GO analysis of biological processes associated with ACRs of MYC2 TF in *Arabidopsis thaliana*. Chr, chromosome; TC, tag count; GO, Gene Ontology; MYC2, myelocytomatosis 2.

and light responses [58,59] and suggest that MYC2 may perform other regulatory roles in biosynthesis and/or responses to stimulus.

Analyzing the changes in chromatin accessibility can help to reveal the dynamics of the transcriptional regulatory landscape during the development or in response to external stimuli [60]. Here is an example to show how to use overlapping-differential analysis tools to identify dynamic changes in chromatin accessibility to respond to heat. In 'analysis of differential-overlapping ACRs' interface, after selecting '*Oryza sativa*' species, 'abiotic stress' experimental classification, 'sample_00_154: heat_30min_3', and 'sample_00_155: heat_2h_1', click the 'start analysis' button to start an analysis. These two samples are 14 days old rice leaves (second leaf) under heat stress (transferred from 30 ℃ to 40 ℃) for 0.5 h and 2 h, respectively. The upper interface shows the detailed information of the two samples as well as the pie chart of overlap rate (percentage of overlapping ACRs of all ACRs in this sample) and differential rate (percentage of different ACRs of all ACRs in this sample). The lower interface displays differential regions and overlapping regions between the two samples, respectively. In differential regions part, 'sample_00_155: heat_2h_1' has more ACRs than 'sample_00_154: heat_30min_3'. In the overlapping regions section, we divided overlapping types into four types. In this example, type C accounts for the most ($\sim$ 51.5%) and type D the least ($\sim$ 15.7%). Type A and type B account for $\sim$ 16.2% and $\sim$ 16.8%, respectively. These results suggest that the chromatin accessibility landscape expands with increasing exposure to high temperatures in *Oryza sativa*, which cooccurs with the expression of about 500 more genes [61], indicating the transcriptional reprogramming in response to heat stress.

To compare tissue-specific ACRs, we used 'sample_00_237: maize_B73Leaf_rep1' (leaf tissue) and 'sample_00_238: maize_B73Ear_rep1' (ear tissue) from *Zea mays*. We identified a total of 67,532 and 34,270 ACRs in the leaf and ear, respectively. Among them, 23,443 are shared ACRs, and 43,949 and 9707 are tissue-specific ACRs in the leaf and ear, respectively. To characterize the potential functions of ACRs, we extracted the nearest neighbor gene of each ACR for GO enrichment analysis. For shared ACRs, the results of the GO enrichment analysis included 87 terms, such as regulation of RNA biosynthetic process, regulation of nucleic acid-templated transcription, and regulation of DNA-templated transcription. For tissue-specific ACRs, we obtained 19,450 and 7384 nearest neighbor genes from leaf and ear tissue, respectively. To further ensure that these genes function in a tissue-specific manner, we integrated RNA-seq data from leaf and ear tissues. The corresponding RNA-seq [62] data were analyzed with the limma test to obtain differentially expressed genes. There are 2382 and 795 nearest neighbor genes differentially expressed in leaf and ear, respectively. Interestingly, regulation of RNA biosynthetic process, regulation of nucleic acid-templated transcription, and regulation of DNA-templated transcription are the top three significant GO groups in both tissues, which suggest that different sets of ACRs and genes may play an important role in regulating transcription in different tissues. Moreover, we found that the genes related to leaf-specific ACRs are heavily enriched in chloroplast rRNA processing and protein localization to the chloroplast.

Discussion

Profiling chromatin accessibility on a genome-wide scale is widely used to understand the transcriptional regulation, tissue specificity, stress responses, and developmental dynamics of plants [38]. For example, Wu et al. systematically studied the combined effects of multiple epigenome features on gene expression in *Arabidopsis thaliana* and *Oryza sativa* based on histone modifications and chromatin accessibility data [63]. Wang et al. studied ATAC-seq data at different stages of somatic embryogenesis induced by auxin and found that auxin can rapidly reconnect the totipotent network of cells by altering chromatin accessibility in *Arabidopsis thaliana* [38]. Moreover, based on the dynamic analysis of chromatin accessibility, they also revealed the hierarchical gene regulatory network in the process of somatic embryogenesis [38]. These applications benefit from the high genomic resolution of chromatin accessibility analysis, reasonable cost, and the ability to process many samples in a fast manner. Database, such as ENCODE [64,65], TCGA [66], and Cistrome [15], all focus on providing original chromatin accessibility data in humans and are being extensively used by tools to annotate *cis*-regulatory elements, such as enhancers [67] and silencers [68]. There is currently no database that provides a collection of complete chromatin accessibility regions and detailed annotation information and analyses of ACRs in plants. Here, we provide the PlantCADB, which can make it easier for users to use ACRs and investigate the mechanisms underlying biological functions.

PlantCADB is a comprehensive database of ACRs and provides a convenient interface to browse, query, analysis, visualize, and download ACRs. The advantages of PlantCADB include: (1) comprehensive ACRs of plant species; (2) inferred TF binding in ACRs using TF footprint analysis; (3) options to query the associated ACRs with user-submitted genome location, gene, or transcription factor; (4) useful online analysis tools for ACRs, such as 'analysis of differential-overlapping ACRs' and 'analysis of overlapping ACRs bound by two TFs'; (5) personalized genome browser for intuitively viewing information of ACRs and adding other useful tracks; (6) conveniently displaying and downloading of ACRs and related annotation information via interactive tables. As illustrated in three case studies, PlantCADB provides convenient tools to explore the relationship between genes, transcription factors, and chromatin accessibility regions to decipher the key questions in plant science. Although till now plant chromatin accessibility is only assessed with bulk sequencing data, with the development of single-cell sequencing technology, users will soon be able to construct the epigenetic landscape of single cell and cell differentiation trajectories with tools, such as scDEC [69], RA3 [70], and epiAnno [71]. PlantCADB will be updated in time to add new datasets and be applied to more plant species.

Data availability

PlantCADB is freely available to the research community without login at https://bioinfor.nefu.edu.cn/PlantCADB/.

Competing interests

The authors have declared no competing interests.

CRediT authorship contribution statement

Ke Ding: Methodology, Formal analysis, Data curation, Conceptualization, Writing – original draft, Writing – review & editing. **Shanwen Sun:** Project administration, Writing – review & editing, Funding acquisition. **Yang Luo:** Software, Visualization. **Chaoyue Long:** Software, Visualization. **Jingwen Zhai:** Software. **Yixiao Zhai:** Visualization. **Guohua Wang:** Conceptualization, Supervision, Funding acquisition, Resources. All authors have read and approved the final manuscript.

Acknowledgments

This work was supported by the Innovation Project of State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, China (Grant No. 2019A04), the National Natural Science Foundation of China (Grant Nos. 62225109, 62001088, and 62072095), and the Fundamental Research Funds for the Central Universities, China (Grant No. 2572022BD04).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.10.005.

ORCID

ORCID 0000-0001-6830-4578 (Ke Ding)
ORCID 0000-0003-4358-8636 (Shanwen Sun)
ORCID 0000-0001-9355-603X (Yang Luo)
ORCID 0000-0001-5293-9536 (Chaoyue Long)
ORCID 0000-0001-8847-0434 (Jingwen Zhai)
ORCID 0000-0002-5068-5672 (Yixiao Zhai)
ORCID 0000-0001-7381-2374 (Guohua Wang)

References

[1] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature 2012;489:75–82.
[2] Bajic M, Maher KA, Deal RB. Identification of open chromatin regions in plant genomes using ATAC-seq. Methods Mol Biol 2018;1675:183–201.
[3] Minnoye L, Marinov GK, Krausgruber T, Pan L, Marand AP, Secchia S, et al. Chromatin accessibility profiling methods. Nat Rev Methods Primers 2021;1:11.
[4] Pajoro A, Madrigal P, Muiño JM, Matus JT, Jin J, Mecchia MA, et al. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biol 2014;15:R41.
[5] Han J, Wang P, Wang Q, Lin Q, Chen Z, Yu G, et al. Genome-wide characterization of DNase I-hypersensitive sites and cold response regulatory landscapes in grasses. Plant Cell 2020;32:2457–73.
[6] Waidmann S, Petutschnig E, Rozhon W, Molnár G, Popova O, Mechtler K, et al. GSK3-mediated phosphorylation of *DEK3* regulates chromatin accessibility and stress tolerance in *Arabidopsis*. FEBS J 2021;289:473–93.
[7] Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 2008;132:311–22.
[8] Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 2010;2010:pdb.prot5384.
[9] Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell 2008;132:887–98.
[10] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 2013;10:1213–8.
[11] Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol 2015;109:21.29.1–9.
[12] Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. Genome Res 2007;17:877–85.
[13] Shashikant T, Ettensohn CA. Genome-wide analysis of chromatin accessibility using ATAC-seq. Methods Cell Biol 2019;151:219–35.
[14] Cui Y, Sheng Li J, Li W. From reads to insights: integrative pipelines for biological interpretation of ATAC-seq data. Genomics Proteomics Bioinformatics 2021;19:519–21.
[15] Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome data browser: a data portal for ChIP-seq and chromatin accessibility data in human and mouse. Nucleic Acids Res 2017;45:D658–62.
[16] Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res 2019;47:D729–35.
[17] Liu Y, Liu X, Gu Y, Lu H. A novel RNA binding protein-associated prognostic model to predict overall survival in hepatocellular carcinoma patients. Medicine 2021;100:e26491.
[18] Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru MD, Reach SM, et al. An atlas of chromatin accessibility in the adult human brain. Genome Res 2018;28:1243–52.
[19] Iatrou A, Clark EM, Wang Y. Nuclear dynamics and stress responses in Alzheimer's disease. Mol Neurodegener 2021;16:65.
[20] Egervari G. Chromatin accessibility in neuropsychiatric disorders. Neurobiol Learn Mem 2021;181:107438.
[21] Playfoot CJ, Duc J, Sheppard S, Dind S, Coudray A, Planet E, et al. Transposable elements and their KZFP controllers are drivers of transcriptional innovation in the developing human brain. Genome Res 2021;31:1531–45.
[22] Rizzardi LF, Hickey PF, Idrizi A, Tryggvadottir R, Callahan CM, Stephens KE, et al. Human brain region-specific variably methylated regions are enriched for heritability of distinct neuropsychiatric traits. Genome Biol 2021;22:116.
[23] Chen S, Liu Q, Cui X, Feng Z, Li C, Wang X, et al. OpenAnnotate: a web server to annotate the chromatin accessibility of genomic regions. Nucleic Acids Res 2021;49:W483–90.
[24] Tannenbaum M, Sarusi-Portuguez A, Krispil R, Schwartz M, Loza O, Benichou JIC, et al. Regulatory chromatin landscape in *Arabidopsis thaliana* roots uncovered by coupling INTACT and ATAC-seq. Plant Methods 2018;14:113.
[25] Tu X, Mejia-Guerra MK, Valdes Franco JA, Tzeng D, Chu PY, Shen W, et al. Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. Nat Commun 2020;11:5089.

[26] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Res 2011;39:D1005–10.
[27] Kersey PJ, Allen JE, Christensen M, Davis P, Falin LJ, Grabmueller C, et al. Ensembl genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res 2014;42:D546–52.
[28] Bolser DM, Staines DM, Perry E, Kersey PJ. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. Methods Mol Biol 2017;1533:1–31.
[29] de Sena BG, Smith AD. Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Res 2019;1874:8.
[30] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.
[31] Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol 2020;21:22.
[32] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
[33] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137.
[34] Zhuo B, Yu J, Chang L, Lei J, Wen Z, Liu C, et al. Quantitative analysis of chromatin accessibility in mouse embryonic fibroblasts. Biochem Biophys Res Commun 2017;493:814–20.
[35] Yan W, Chen D, Schumacher J, Durantini D, Engelhorn J, Chen M, et al. Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. Nat Commun 2019;10:1705.
[36] Wang F, Bai X, Wang Y, Jiang Y, Ai B, Zhang Y, et al. ATACdb: a comprehensive human chromatin accessibility database. Nucleic Acids Res 2021;49:D55–64.
[37] Shah RN, Ruthenburg AJ. Sequence deeper without sequencing more: Bayesian resolution of ambiguously mapped reads. PLoS Comput Biol 2021;17:e1008926.
[38] Wang FX, Shang GD, Wu LY, Xu ZG, Zhao XY, Wang JW. Chromatin accessibility dynamics and a hierarchical transcriptional regulatory network structure for plant somatic embryogenesis. Dev Cell 2020;54:742–57.
[39] Maher KA, Bajic M, Kajala K, Reynoso M, Pauluzzi G, West DA, et al. Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. Plant Cell 2018;30:15–36.
[40] Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015;31:2382–3.
[41] Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 2012;40:D1178–86.
[42] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156–8.
[43] Bryois J, Garrett ME, Song L, Safi A, Giusti-Rodriguez P, Johnson GD, et al. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nat Commun 2018;9:3121.
[44] Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell 2013;153:320–34.
[45] Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 2013;153:307–19.
[46] Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME suite: tools for motif discovery and searching. Nucleic Acids Res 2009;37:W202–8.
[47] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011;27:1017–8.
[48] Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 2017;45:D1040–5.
[49] Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res 2012;22:1711–22.
[50] Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biol 2019;20:45.
[51] Gusmao EG, Dieterich C, Zenke M, Costa IG. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics 2014;30:3143–51.
[52] Gusmao EG, Allhoff M, Zenke M, Costa IG. Analysis of computational footprinting methods for DNase sequencing experiments. Nat Methods 2016;13:303–9.
[53] Karabacak Calviello A, Hirsekorn A, Wurmus R, Yusuf D, Ohler U. Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. Genome Biol 2019;20:42.
[54] Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, Schmitz RJ. The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants 2019;5:1250–9.
[55] Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 2019;20:207–20.
[56] Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 2016;17:66.
[57] Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 2019;47:D419–26.
[58] Yadav V, Mallappa C, Gangappa SN, Bhatia S, Chattopadhyay S. A basic helix-loop-helix transcription factor in *Arabidopsis*, MYC2, acts as a repressor of blue light-mediated photomorphogenic growth. Plant Cell 2005;17:1953–66.
[59] Chen Q, Sun J, Zhai Q, Zhou W, Qi L, Xu L, et al. The basic helix-loop-helix transcription factor MYC2 directly represses *PLETHORA* expression during jasmonate-mediated modulation of the root stem cell niche in *Arabidopsis*. Plant Cell 2011;23:3335–52.
[60] Ho L, Crabtree GR. Chromatin remodelling during development. Nature 2010;463:474–84.
[61] Wilkins O, Hafemeister C, Plessis A, Holloway-Phillips MM, Pham GM, Nicotra AB, 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 2016;28:2365–84.
[62] Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, et al. Widespread long-range *cis*-regulatory elements in the maize genome. Nat Plants 2019;5:1237–49.
[63] Wu Z, Tang J, Zhuo J, Tian Y, Zhao F, Li Z, et al. Chromatin signature and transcription factor binding provide a predictive basis for understanding plant gene expression. Plant Cell Physiol 2019;60:1471–86.
[64] Consortium EP, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 2020;583:699–710.
[65] Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[66] Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science 2018;362:eaav1898.
[67] Chen S, Gan M, Lv H, Jiang R. DeepCAPE: a deep convolutional neural network for the accurate prediction of enhancers. Genomics Proteomics Bioinformatics 2021;19:565–77.

[68] Zeng W, Chen S, Cui X, Chen X, Gao Z, Jiang R. SilencerDB: a comprehensive database of silencers. Nucleic Acids Res 2021;49: D221–8.
[69] Liu Q, Chen S, Jiang R, Wong WH. Simultaneous deep generative modeling and clustering of single cell genomic data. Nat Mach Intell 2021;3:536–44.
[70] Chen S, Yan G, Zhang W, Li J, Jiang R, Lin Z. RA3 is a reference-guided approach for epigenetic characterization of single cells. Nat Commun 2021;12:2177.
[71] Chen X, Chen S, Song S, Gao Z, Hou L, Zhang X, et al. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. Nat Mach Intell 2022;4:116–26.