D1456–D1467 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 17 September 2021*
https://doi.org/10.1093/nar/gkab794

PlantGSAD: a comprehensive gene set annotation database for plant species

Xuelian Ma, Hengyu Yan, Jiaotong Yang, Yue Liu, Zhongqiu Li, Minghao Sheng, Yaxin Cao, Xinyue Yu, Xin Yi, Wenying Xu and Zhen Su*

State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100193, China

Received July 15, 2021; Revised August 26, 2021; Editorial Decision August 27, 2021; Accepted September 01, 2021

ABSTRACT

With the accumulation of massive data sets from high-throughput experiments and the rapid emergence of new types of omics data, gene sets have become more diverse and essential for the refinement of gene annotation at multidimensional levels. Accordingly, we collected and defined 236 007 gene sets across different categories for 44 plant species in the Plant Gene Set Annotation Database (PlantGSAD). These gene sets were divided into nine main categories covering many functional subcategories, such as trait ontology, co-expression modules, chromatin states, and liquid-liquid phase separation. The annotations from the collected gene sets covered all of the genes in the *Brassicaceae* species *Arabidopsis* and *Poaceae* species *Oryza sativa*. Several GSEA tools are implemented in PlantGSAD to improve the efficiency of the analysis, including custom SEA for a flexible strategy based on customized annotations, SEACOMPARE for the cross-comparison of SEA results, and integrated visualization features for ontological analysis that intuitively reflects their parent-child relationships. In summary, PlantGSAD provides numerous gene sets for multiple plant species and highly efficient analysis tools. We believe that PlantGSAD will become a multifunctional analysis platform that can be used to predict and elucidate the functions and mechanisms of genes of interest. PlantGSAD is publicly available at http://systemsbiology.cau.edu.cn/PlantGSEAv2/.

INTRODUCTION

The continued development and widespread availability of high-throughput techniques allow for the exploration of changes and regulation at a whole genome-wide level under certain conditions. It is important to integrate the large number of data sets generated by high-throughput experiments for gene annotation, especially when working with gene sets. The gene ontology (GO) system is a common functional category widely used in plenty of analysis platforms and tools, such as DAVID (1,2) and agriGO (3,4). Due to the limitation of GO gene sets, several platforms, such as MSigDB (5), WebGestalt (6) and PlantGSEA (7), have been designed to assign genes to categories based on Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic processes, gene families, and some curated gene sets. All of these gene sets can be used to compute the overlap with a query gene list for the biological interpretation using gene set enrichment analysis (GSEA) (8).

The biological knowledge associated with gene sets is gradually extending beyond pre-existing gene set categories, with multiple diverse data sets emerging and new functional categories being introduced. For example, co-expression modules (CoMs) (9) are a new functional category that has been added into some platforms and databases, including EviNet (10), WebGestalt 2017 (6), ccNET (11) and MCENet (12). In addition, various forms of epigenetic regulation, such as DNA methylation, histone modifications, and histone variants, have been widely implicated in the control of complex biological activities (13). Specific combinations of multiple epigenetic marks at the whole genome level can be defined as chromatin states, which play a meaningful role in various biological processes. Recently, there have also been significant advances in research on liquid-liquid phase separation (LLPS), single cell RNA-seq and RNA binding proteins in biological systems, leading to new data types. It has been reported that LLPS is associated with chromatin compartmentalization in transcriptional regulation (19–21). At the same time, plant research has extended into three-dimensional (3D) genomics analysis (14) and 3D genomic data from various plant species have been published using Hi-C and ChIA-PET techniques (15–18). However, 3D-related categories for gene sets associated with Hi-C and LLPS are seldom present in databases and platforms for search and analysis.

Although some resources with gene set annotations and analyses are available, including PAGED (22) and

*To whom correspondence should be addressed. Tel: +86 10 62731380; Fax: +86 10 62731380; Email: zhensu@cau.edu.cn

GeneSetDB (23), they mainly focused on the disease-related gene sets in human and typical gene sets from model species. In addition, both PAGED (22) and GeneSetDB (23) were published in 2012 and lacked novel data types such as chromatin state-based gene sets, LLPS gene sets, and single-cell RNA-seq cluster gene sets. For plants, the functional genome database of some species covers the common types of gene sets. For example, The Arabidopsis Information Resource (TAIR) (24) database offers GO and KEGG categories for *Arabidopsis*. However, there is no existing database that broadly integrates gene sets across different categories for plant species or that provides gene annotations at multidimensional levels for plant communities. Therefore, it is necessary to expand the number of data types and gene sets for a wider range of plant species and to logically organize the gene set categories.

To meet the growing demands of researchers and users, we built the Plant Gene Set Annotation Database (PlantGSAD) and uploaded enough novel categories of gene sets from plant species, including CoMs, chromatin states, and LLPS. Currently, PlantGSAD contains 236 007 gene sets across nine main functional categories for 44 plant species. All of these gene sets can be browsed and specific detailed knowledge of gene sets can be searched for. In addition, GSEA analysis is available in the background for the collected gene sets to produce comprehensive output report and graphical visualization for each gene set category. Other useful tools in PlantGSAD include customized singular enrichment analysis (SEA), SEACOMPARE (the cross comparison of SEA results), dot plot drawer and ID conversion. We believe that PlantGSAD is beneficial for gene function analysis and as a reference platform for plant species. It is freely available at http://systemsbiology.cau.edu.cn/PlantGSEAv2.

MATERIALS AND METHODS

Data sources and integration

We collected gene sets across different types of data from various resources, and organized them using a curation workflow. The data sources used for gene sets are listed in Table 1 and Supplementary Table S1.

The ontology gene sets included gene ontology (GO), plant ontology (PO), and trait ontology (TO); these were downloaded from related annotation databases, such as agriGOv2 (3), Phytozome (25), TAIR (24), Planteome (26) and RiceData (http://www.ricedata.cn). The PO and TO gene sets were re-computed to cover every term (including the parent terms) and organized as formatted annotation files. We also collected the gene sets from KEGG (27), PlantCyc (28) and MapMan (29), and grouped them as pathway gene sets.

With regard to the gene family-based gene sets, the gene families were collected from genomics platforms (e.g. MCENet (12) for maize and ccNET (11) for cotton), and specific gene family databases (e.g. PlantTFDBv4.0 (68) for transcription factor families and iUUCDv2.0 (30) for ubiquitin proteasome system-related gene families). The collected data were manually curated to remove redundancy and conflicts.

We also integrated the gene sets related to transcription regulation. The data sets and genome-wide annotation in PCSD database (31) were organized into chromatin state (Sta) gene set, epigenetic mark related gene set (ERG), and chromatin associated factor related gene set (CRG). We grouped them together with the Hi-C gene set (HiC) to define a new category: chromatin states based gene sets. We also integrated information related to transcription factor targets and microRNA targets from multiple data sources, such as the Plant Cistrome Database (38), AGRIS (76) and PNRD (78), to define another category: target gene sets.

We defined a CoM gene set based on co-expression gene networks within the public resources. For the species such as maize and cotton, the CoM gene sets were directly downloaded from MCENet (12) and ccNET (11). For the species such as *Arabidopsis* and rice, we collected the co-expression gene network built by ATTED-II (79) and Rice-FREDN (32) and identified the functional modules by CFinder (33) with standard parameters. In addition, we manually collected reference gene sets from published literatures, which then underwent curation and organization. We also grouped some gene sets related to novel functional categories, such as single-cell RNA sequencing (scRNA-seq), liquid–liquid phase separation (LLPS), and RNA-binding proteins (RBPs). These gene sets were manually collected from specific web servers and literatures (listed in Table 1 and Supplementary Table S2).

For some species, the collected gene sets might be got from different genome versions. We converted them to the unified ID system for the same species to allow curation, browse and analysis (Supplementary Table S3).

Functional category enrichment analysis

In analysis page of PlantGSAD, users can submit a list of genes of interest for the enrichment analysis, and choose suitable categories from G1 to G9. Statistical tests and multiple test correction methods are provided. Three statistical tests are available for selection: hypergeometric test, Fisher's exact test and chi-square test. The statistical formula for Fisher's exact test, which is the default test, is presented in Equation (1):

$$\mathrm{P} = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}} \tag{1}$$

where N is the total number of genes in an organism or in the user-provided background, n is the number of genes in the query list, K is the total number of genes in one gene set and k is the number of overlapping genes.

We provide six multiple test correction methods: Yekutieli, Bonferroni, Hochberg, Benjamini-Hochberg (BH), Hommel and Holm (34). Adjusted P-values are obtained after performing correction for the false discovery rate (FDR), with the significantly enriched gene sets for each gene set category displayed using a cut-off selection.

SEACOMPARE and custom SEA

Users can compare two or more SEA results using the SEACOMPARE tool, with a heatmap showing the common

Table 1. Total numbers and sources of the categorized gene sets for comprehensively annotated species

		Category	No. of supported organisms	No. of total gene sets	No. of covered genes	Main data source
G1	GO	Gene Ontology	44	105 339	878 035	AgriGOv2 (3)/TAIR (24)/Phytozome (25)
G2	PO	Plant Ontology	3	5561	44 500	Planteome (26)/TAIR (24)
	TO	Plant Trait Ontology	4	3319	1066	RiceData /Planteome (26)
	PECO	Plant Experimental Conditions Ontology	1	564	20 392	Planteome (26)
G3	Cyc	PlantCyc	33	56 606	210 169	PlantCyc (28)
	KEGG	KEGG	39	16 494	456 730	KEGG (27)/Phytozome (25)
	Map	MapMan	14	17 426	507 408	MapMan (29)
G4	TR	Transcription Regulators/Factors	41	2851	84 041	PlantTFDBv4.0 (68)
	CAZy	Carbohydrate-Active Enzymes	9	842	10 754	CAZy database (69)
	PK	Protein Kinase	8	574	9714	MCENet (12)/ccNET(11)/iTAK (70)
	Ub	Ubiquitins	15	529	17 373	iUUCDv2.0 (30)/PlantsUPS (71)
	P450	Cytochrome P450	8	897	2555	MCENet (12) /ccNET(11) /SorghumFDB (72)
	EAR	EAR motif	35	1216	9223	PlantEAR (37)
	LIP	Lipid metabolism Enzymes	2	359	1873	ARALIP website (73)/SFGD (74)
	Caf	Chromatin associated factor	2	166	1078	ChromDB (75)
	OTH	Other gene family based gene sets	3	703	5782	TAIR/SorghumFDB (72)
G5	Sta	Chromatin states	5	159	205 647	PCSD (31)
	ERG	Epigenetic mark related genes	4	288	170 572	PCSD (31)
	CRG	Chromatin associated factor related genes	3	44	46 944	PCSD (31)
	HiC	Hi-C identified genes	2	4	4031	2 literatures
G6	TFT	Transcription Factor Targets	2	760	64 810	Plant Cistrome Database (38)/AGRIS (76)/
	MIR	MicroRNA Targets	9	4694	25 614	AraPath (77)/PNRD (78)
G7	CoM	Co-expression gene module	10	14 220	82 020	ATTED-II (79) /ccNET (11) /SorghumFDB (72) /MCENet (12)
G8	LIT	Literature/Reference gene sets	16	1950	110 503	458 literatures
G9	ScR	Single cell RNA-seq identified gene sets	4	337	32 574	7 literatures
	LLPS	Liquid-liquid phase separation related genes	31	93	194 564	DrLLPS (80)/PSPredictor (81)
	Rbp	RNA binding protein	1	12	2766	4 literatures

or unique significant gene sets. The color indicates the P-value or FDR. The custom SEA is independent of species and shares the same principle as functional category enrichment analysis. To use this custom tool, users need to prepare background datasets consisting of genes and their related gene sets. If the parent-child relationships between gene sets are submitted, the genes in the child gene sets belong to the parent gene sets by default during the enrichment computation. DAG (direct acyclic graph) graphs are also created as part of the results.

Database implementation

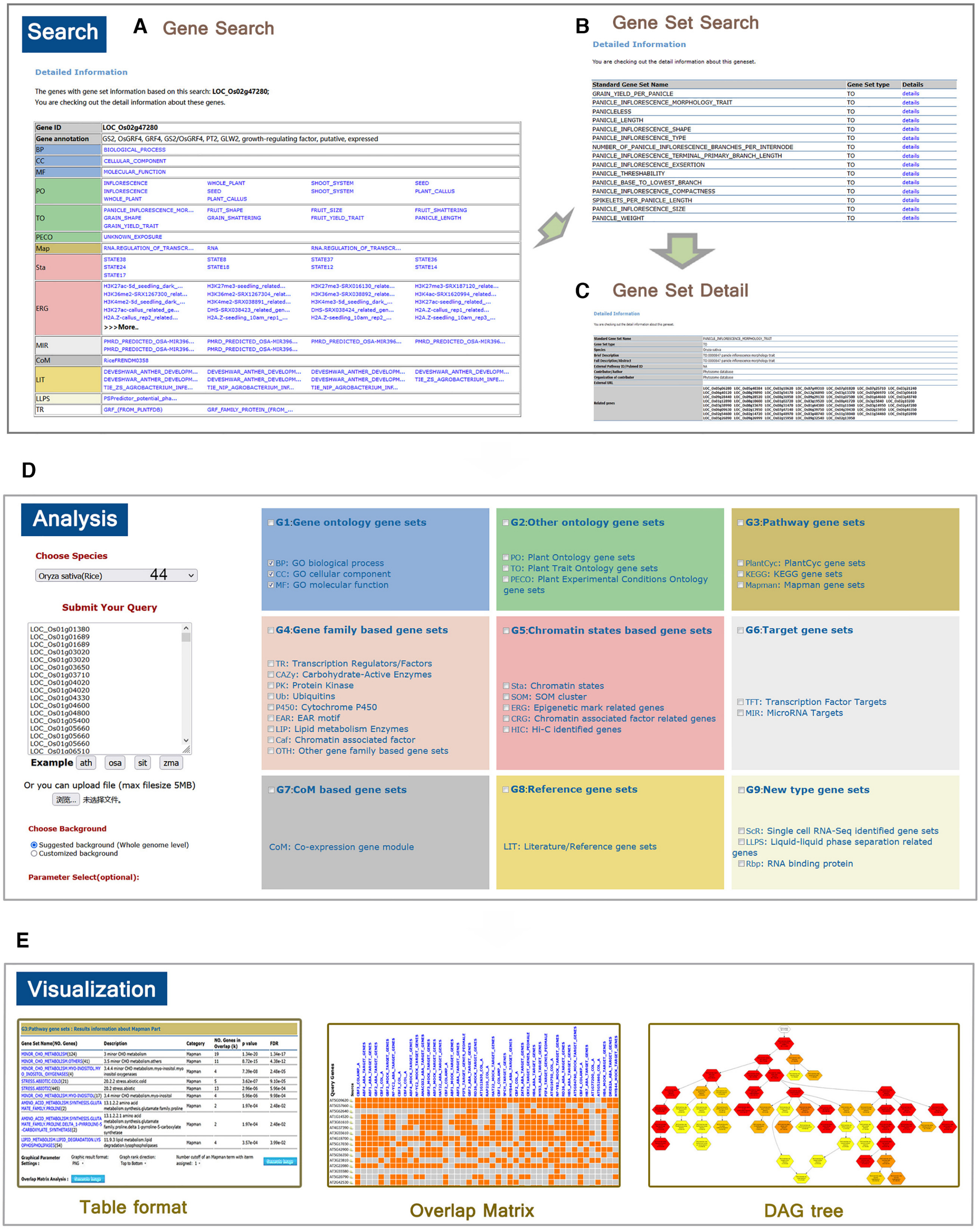
PlantGSAD was constructed on a standard LAMP (Linux + Apache + MySQL + PHP/Python) system. The data sets are stored in MySQL (www.mysql.com) and formatted text files, analysis codes were compiled using Python (www.python.org), and the web interface was built using PHP scripts (www.php.net) on Red Hat Linux powered by an Apache server (www.apache.org). No software or plug-ins are needed because the PlantGSAD tools are web-based. Users are free to access the database and there are no login requirements.

RESULTS

Gene set annotation and organization

The accumulation of multidimensional omics data and the emergence of new types of data prompted us to collect a large volume of data for various species. The resulting PlantGSAD is capable of annotating gene sets for multiple assay types, including omics data sets generated from high-throughput transcriptomics and epigenomics experiments. PlantGSAD supports 44 species across nine functional categories, including GO (G1), other ontology (G2), pathway (G3), gene family-based (G4), chromatin states-based (G5), target (G6), CoM-based (G7), reference (G8), and new type (G9) gene sets (Figure 1D and Table 1).

The number of defined gene sets is currently 236 007, covering a range of data types from public databases and manually collected literature data (Supplementary Table S2). In addition to GO, the ontology systems have been expanded to include PO, TO, and Plant Experimental Conditions Ontology (PECO), which allows the further annotation of gene sets (26). TO gene sets contain information on the agronomic traits of crops. The GO and other ontology categories are particularly useful for geneticists, biologists, and



A Gene Search

Detailed Information

The genes with gene set information based on this search: **LOC_Os02g47280;**
You are checking out the detail information about these genes.

Gene ID	**LOC_Os02g47280**			
Gene annotation	GS2, OsGRF4, GRF4, GS2/OsGRF4, PT2, GLW2, growth-regulating factor, putative, expressed			
BP	BIOLOGICAL_PROCESS			
CC	CELLULAR_COMPONENT			
MF	MOLECULAR_FUNCTION			
PO	INFLORESCENCE INFLORESCENCE WHOLE_PLANT	WHOLE_PLANT SEED PLANT_CALLUS	SHOOT_SYSTEM SHOOT_SYSTEM	SEED PLANT_CALLUS
TO	PANICLE_INFLORESCENCE_MOR... GRAIN_SHAPE GRAIN_YIELD_TRAIT	FRUIT_SHAPE GRAIN_SHATTERING	FRUIT_SIZE FRUIT_YIELD_TRAIT	FRUIT_SHATTERING PANICLE_LENGTH
PECO	UNKNOWN_EXPOSURE			
Map	RNA.REGULATION_OF_TRANSCR...	RNA	RNA.REGULATION_OF_TRANSCR...	
Sta	STATE38 STATE24 STATE17	STATE8 STATE18	STATE37 STATE12	STATE36 STATE14
ERG	H3K27ac-5d_seedling_dark_... H3K36me2-SRX1267300_relat... H3K4me2-5d_seedling_dark_... H3K27ac-callus_related_ge... H2A.Z-callus_rep2_related... >>>More..	H3K27me3-seedling_related... H3K36me2-SRX1267304_relat... H3K4me2-SRX038891_related... DHS-SRX038423_related_gen... H2A.Z-seedling_10am_rep1_...	H3K27me3-SRX016130_relate... H3K36me3-SRX038892_relate... H3K4me3-5d_seedling_dark_... DHS-SRX038424_related_gen... H2A.Z-seedling_10am_rep2_...	H3K27me3-SRX187120_relate... H3K4ac-SRX1620994_related... H3K27ac-seedling_related_... H2A.Z-callus_rep1_related... H2A.Z-seedling_10am_rep3_...
MIR	PMRD_PREDICTED_OSA-MIR396... PMRD_PREDICTED_OSA-MIR396...	PMRD_PREDICTED_OSA-MIR396... PMRD_PREDICTED_OSA-MIR396...	PMRD_PREDICTED_OSA-MIR396...	PMRD_PREDICTED_OSA-MIR396...
CoM	RiceFRENDM0358			
LIT	DEVESHWAR_ANTHER_DEVELOPM... DEVESHWAR_ANTHER_DEVELOPM... TIE_ZS_AGROBACTERIUM_INFE...	DEVESHWAR_ANTHER_DEVELOPM... DEVESHWAR_ANTHER_DEVELOPM... TIE_NIP_AGROBACTERIUM_INF...	DEVESHWAR_ANTHER_DEVELOPM... DEVESHWAR_ANTHER_DEVELOPM... TIE_NIP_AGROBACTERIUM_INF...	DEVESHWAR_ANTHER_DEVELOPM... TIE_ZS_AGROBACTERIUM_INFE...
LLPS	PSPredictor_potential_pha...			
TR	GRF_(FROM_PLNTFDB)	GRF_FAMILY_PROTEIN_(FROM_...		

B Gene Set Search

Detailed Information

You are checking out the detail information about this geneset.

Standard Gene Set Name	Gene Set type	Details
GRAIN_YIELD_PER_PANICLE	TO	details
PANICLE_INFLORESCENCE_MORPHOLOGY_TRAIT	TO	details
PANICLELESS	TO	details
PANICLE_LENGTH	TO	details
PANICLE_INFLORESCENCE_SHAPE	TO	details
PANICLE_INFLORESCENCE_TYPE	TO	details
NUMBER_OF_PANICLE_INFLORESCENCE_BRANCHES_PER_INTERNODE	TO	details
PANICLE_INFLORESCENCE_TERMINAL_PRIMARY_BRANCH_LENGTH	TO	details
PANICLE_INFLORESCENCE_EXSERTION	TO	details
PANICLE_THRESHABILITY	TO	details
PANICLE_BASE_TO_LOWEST_BRANCH	TO	details
PANICLE_INFLORESCENCE_COMPACTNESS	TO	details
SPIKELETS_PER_PANICLE_LENGTH	TO	details
PANICLE_INFLORESCENCE_SIZE	TO	details
PANICLE_WEIGHT	TO	details

C Gene Set Detail

Detailed Information

You are checking out the detail information about this geneset.

Standard Gene Set Name	PANICLE_INFLORESCENCE_MORPHOLOGY_TRAIT
Gene Set type	TO
Species	Oryza sativa
Brief Description	TO:0000847 panicle inflorescence morphology trait
Full Description/Abstract	TO:0000847 panicle inflorescence morphology trait
External Pathway ID/Pubmed ID	NA
Contributor/Author	Phytozome database
Organization of contributer	Phytozome database
External URL	
Related genes	LOC_Os05g06280 LOC_Os05g40384 LOC_Os03g10620 LOC_Os07g49310 LOC_Os07g01820 LOC_Os07g25710 LOC_Os03g21240 LOC_Os06g40120 LOC_Os08g39890 LOC_Os03g54170 LOC_Os12g36890 LOC_Os03g13370 LOC_Os07g06970 LOC_Os03g06410 LOC_Os09g28440 LOC_Os09g28520 LOC_Os08g36950 LOC_Os09g29130 LOC_Os01g07500 LOC_Os01g64660 LOC_Os03g48740 LOC_Os01g12890 LOC_Os08g10600 LOC_Os01g02720 LOC_Os03g19520 LOC_Os08g41720 LOC_Os3g15840 LOC_Os02g10200 LOC_Os03g38990 LOC_Os08g33670 LOC_Os08g31470 LOC_Os01g64380 LOC_Os01g11040 LOC_Os03g14950 LOC_Os02g47280 LOC_Os06g09630 LOC_Os02g13950 LOC_Os07g47140 LOC_Os06g39750 LOC_Os04g39430 LOC_Os02g15950 LOC_Os06g46350 LOC_Os02g54600 LOC_Os02g14720 LOC_Os03g48970 LOC_Os03g48740 LOC_Os11g38040 LOC_Os11g34460 LOC_Os01g02890 LOC_Os05g26890 LOC_Os09g26999 LOC_Os02g15950 LOC_Os09g32540 LOC_Os02g13950

D Analysis

Choose Species

Oryza sativa(Rice) 44

Submit Your Query

```
LOC_Os01g01380
LOC_Os01g01689
LOC_Os01g01689
LOC_Os01g03020
LOC_Os01g03020
LOC_Os01g03650
LOC_Os01g03710
LOC_Os01g04020
LOC_Os01g04020
LOC_Os01g04330
LOC_Os01g04600
LOC_Os01g04800
LOC_Os01g05400
LOC_Os01g05660
LOC_Os01g05660
LOC_Os01g05660
LOC_Os01g06510
```

Example ath osa sit zma

Or you can upload file (max filesize 5MB)

浏览... 未选择文件。

Choose Background

● Suggested background (Whole genome level)
○ Customized background

Parameter Select(optional):

□ **G1:Gene ontology gene sets**
☑ BP: GO biological process
☑ CC: GO cellular component
☑ MF: GO molecular function

□ **G2:Other ontology gene sets**
□ PO: Plant Ontology gene sets
□ TO: Plant Trait Ontology gene sets
□ PECO: Plant Experimental Conditions Ontology gene sets

□ **G3:Pathway gene sets**
□ PlantCyc: PlantCyc gene sets
□ KEGG: KEGG gene sets
□ Mapman: Mapman gene sets

□ **G4:Gene family based gene sets**
□ TR: Transcription Regulators/Factors
□ CAZy: Carbohydrate-Active Enzymes
□ PK: Protein Kinase
□ Ub: Ubiquitins
□ P450: Cytochrome P450
□ EAR: EAR motif
□ LIP: Lipid metabolism Enzymes
□ Caf: Chromatin associated factor
□ OTH: Other gene family based gene sets

□ **G5:Chromatin states based gene sets**
□ Sta: Chromatin states
□ SOM: SOM cluster
□ ERG: Epigenetic mark related genes
□ CRG: Chromatin associated factor related genes
□ HIC: Hi-C identified genes

□ **G6:Target gene sets**
□ TFT: Transcription Factor Targets
□ MIR: MicroRNA Targets

□ **G7:CoM based gene sets**
CoM: Co-expression gene module

□ **G8:Reference gene sets**
LIT: Literature/Reference gene sets

□ **G9:New type gene sets**
□ ScR: Single cell RNA-Seq identified gene sets
□ LLPS: Liquid-liquid phase separation related genes
□ Rbp: RNA binding protein

E Visualization

G3:Pathway gene sets : Results information about Mapman Part

Gene Set Name(NO. Genes)	Description	Category	NO. Genes in Overlap (k)	p value	FDR
MINOR_CHO_METABOLISM(124)	3 minor CHO metabolism	Mapman	19	1.34e-20	1.34e-17
MINOR_CHO_METABOLISM.OTHERS(41)	3.5 minor CHO metabolism.others	Mapman	11	8.72e-15	4.38e-12
MINOR_CHO_METABOLISM.MYO-INOSITOL.MYO_INOSITOL_OXYGENASES(4)	3.4.4 minor CHO metabolism.myo-inositol.myo inositol oxygenases	Mapman	4	7.39e-08	2.48e-05
STRESS.ABIOTIC.COLD(21)	20.2.2 stress.abiotic.cold	Mapman	5	3.62e-07	9.10e-05
STRESS.ABIOTIC(445)	20.2 stress.abiotic	Mapman	13	2.96e-06	5.96e-04
MINOR_CHO_METABOLISM.MYO-INOSITOL(17)	3.4 minor CHO metabolism.myo-inositol	Mapman	4	5.96e-06	9.98e-04
AMINO_ACID_METABOLISM.SYNTHESIS.GLUTAMATE_FAMILY.PROLINE(2)	13.1.2.2 amino acid metabolism.synthesis.glutamate family.proline	Mapman	2	1.97e-04	2.48e-02
AMINO_ACID_METABOLISM.SYNTHESIS.GLUTAMATE_FAMILY.PROLINE.DELTA_1-PYRROLINE-5-CARBOXYLATE_SYNTHETASE(2)	13.1.2.2.1 amino acid metabolism.synthesis.glutamate family.proline.delta 1-pyrroline-5-carboxylate synthetase	Mapman	2	1.97e-04	2.48e-02
LIPID_METABOLISM.LIPID_DEGRADATION.LYSOPHOSPHOLIPASES(54)	11.9.3 lipid metabolism.lipid degradation.lysophospholipases	Mapman	4	3.57e-04	3.99e-02

Graphical Parameter Settings : Graphic result format: PNG · Graph rank direction: Top to Bottom · Number cutoff of an Mapman term with iterm assigned: 1 · Generate Image

Overlap Matrix Analysis : Generate Image

Figure 1. Overview of PlantGSAD functions. (**A**) Example of gene search results in PlantGSAD. In the first column, the background color represents the gene set category. (**B**) Gene set search results in PlantGSAD for the TO category in rice. (**C**) Detailed information on the gene set PANICLE_INFLORESCENCE_MORPHOLOGY_TRAIT. (**D**) GSEA in PlantGSAD. The query gene list for 44 plant species can be input for enrichment analysis (left). Individual gene set categories can be selected (right). (**E**) Enrichment analysis results can be visualized using table, overlap matrix, and DAG tree.

breeders when investigating the potential molecular mechanisms associated with target genes (35).

For the pathway gene sets, we integrated the KEGG, PlantCyc and MapMan gene sets. MapMan was specifically designed to cover plant-specific pathways and processes (36). We also introduced a number of gene family gene sets, including the EAR motif-related gene family and chromatin associated factors. EAR motif-containing proteins, which are highly conserved transcription repressors across a diverse range of plant species, can interact with co-repressors, affecting the structure of chromatin via histone modifications, thus repressing gene transcription and playing a role in hormone signal transduction, stress responses, and development (37). For this reason, the EAR subcategory was classified as a pivotal gene family in PlantGSAD.

Co-expression networks have emerged as an efficient way to decipher the potential function of genes. We obtained 14 220 CoMs across 10 organisms for PlantGSAD as a functional category. Every CoM was defined as a gene set. CoMs provide basic annotated information to describe the complex roles of gene sets at the transcriptomic level. The complex signaling networks associated with transcriptional changes during developmental stages and responding to environmental cues often involve many related transcription factors (TFs). Thus, transcriptional factor targets (TFTs) were classified as a subtype of target gene sets in PlantGSAD to provide a framework for understanding the regulation mechanisms during plant development processes and stress responses. A large number of gene sets for this functional subcategory were collected for PlantGSAD. For example, we integrated 608 *Arabidopsis* gene sets mainly acquired from DAP-seq (38) and 115 maize TFT data sets primarily obtained from ChIP-seq (39). These TFT data sets in PlantGSAD can be used to investigate query genes involved in complex transcription regulatory networks, providing a valuable genome annotation resource for identifying potential key TFs that transcriptionally regulate the biological processes involving the query genes.

Epigenomic data represents vital information in investigating the dynamic role of chromatin states in gene regulation, notably the linear partitioned genome segments with various epigenetic features localized in unannotated regions (31,40). Thus, we added chromatin state-based gene sets as an essential functional category in PlantGSAD. The increase in the functional categories related to chromatin states is beneficial for increasing the biological understanding of various epichromatin states during developmental processes and in response to environmental cues, and their subsequent differences in transcriptional regulation. Corresponding epigenetic marks, such as DNA methylation, post-translational histone modifications, and chromatin associated factors, are highly dynamic, with the changes in the chromatin states representing distinct epigenetic and spatial features of various transcriptional activities. The chromatin architectural capture technique Hi-C has been widely employed in the characterization of chromatin architecture in various plant and animal model species, leading to the definition of various chromatin architectural features, such as the chromosome territory, compartment A/B, topological associated domains (TADs), and chromatin loops (41). Hence, chromatin states based gene sets were added as a functional category in PlantGSAD, consisting of four subgroups: Sta (defined chromatin states) from the PCSD database, ERG (epigenetic mark related), CRG (chromatin associated factor related), and HiC (Hi-C technology based) gene sets. Every defined chromatin state in a specific species touches on some preferential epigenetic marks in preferentially located regions. The same chromatin state has genes with similar epigenetic marks in similar regions of the genes, which might jointly participate in the transcription regulation of cellular activities. In addition, the information about genome-wide chromosome conformation capture can help to identify genes related to the assembly and conformational changes of chromatin.

Reference gene sets were also curated, organized, and integrated from public research information, especially up to 1636 manually collected reference gene sets for *Arabidopsis*. We also introduced more novel functional categories to PlantGSAD to better serve the needs of biologists from different research areas. Single-cell RNA sequencing (scRNA-seq) has been extensively employed to study cell-specific gene expression in animals and plants. The genes from distinct subpopulations and rare cell types of plants, such as putative quiescent center cells, have been included for the gene set annotation. As a critical component of chromatin compartmentalization, LLPS drives the formation of miscellaneous membrane-less compartments in cells and is involved in 3D chromatin organization and transcriptional regulation. Due to the fact that LLPS is associated with chromatin compartmentalization in transcriptional regulation, we integrated the LLPS-related proteins and included related gene sets in a functional sub-category. In the process of post-transcriptional regulation, RBPs are indispensable chaperones that naturally bind to RNA via one or multiple globular RNA-binding domains, changing the function and/or fate of the bound RNAs (42–46). Because RBPs are critical components in adjusting global cellular transcript levels via the binding to and potential regulation of the transcripts, we collected these proteins as a novel gene set type.

PlantGSAD contains cereal crops, economic crops, and medicinal plants, etc. In particular, it includes vegetable and oilseed crops in the *Brassicaceae* family and major cereal crops within the *Poaceae* (grass) family. In PlantGSAD, the annotations from collected gene sets cover all of the genes in the *Brassicaceae* species *Arabidopsis* and the *Poaceae* species *Oryza Sativa* (rice), while the annotation rate for maize (*Zea Mays*) genes has reached over 99 percent. In addition, soybean (*Glycine max* L.) in *Fabaceae* produces high-quality oil from seeds, so functional annotations of the genes involved in the soybean acyl-lipid pathway have been collected and grouped into the lipid metabolism enzymes (LIP) subcategory. Medicinal plants have also attracted significant research attention (for example, *Catharanthus roseus* has been shown to produce anti-cancer agents), thus we collected CoM data sets of *C. roseus* from CroFDB database (47) for use in PlantGSAD.

The large number and the wide range of gene sets from plant species in PlantGSAD will significantly enhance the biological interpretation of genes of interest and cover a large range of unknown genes.

Gene set query and access

All of the gene sets from the 44 plant species can be easily accessed and checked individually using the free selection of categories and species on the browse page of PlantGSAD (Figure 1). The name, type, species, and annotations for these gene sets are presented and the name of the gene set links to a details page for more information (Figure 1A). Rice gene LOC_Os02g47280 (*OsGRF4*, growth-regulating factor 4) is used as an example for the search function in the PlantGSAD (Figure 1A). The search tools help users to conveniently retrieve the detailed information of collected gene sets and their constituent genes, and up to 30 genes can be searched for at the same time. A keyword-based search is provided for the gene sets (Figure 1B, C). The detailed annotations for the gene sets or individual genes serve as important selection criteria when choosing candidates for follow-up studies.

Furthermore, GSEA analysis is available in the background of all gene sets in PlantGSAD, so the query gene list can be annotated using the enriched gene sets (Figure 1D). The direct visualization of the GSEA output report is convenient for analyzing the relevant biological knowledge (Figure 1E). PlantGSAD is also capable of analyzing query lists of genes from one or more functional categories, and the GSEA results are presented clearly in a series of tables. The database provides a summary head in the result page highlighting the significant results in each category and presenting a link to the corresponding analysis table. For ontological analyses (e.g. PO and TO), we integrated visualization features (e.g. DAG) to intuitively reflect the parent-child relationships. DAGs highlight the essential and meaningful terms from the raw and partially redundant enrichment results, allowing researchers to better understand the nature of the causal relationships involved. Furthermore, the significance levels of the terms in the DAGs are directly indicated using a color range from light yellow to dark red.

Additional assistant tools for analysis

Several tools can be implemented into PlantGSAD to improve the efficiency of the analysis. Custom SEA tools have been designed to adapt to individual user-defined data ranging from specific plants to animals or other organisms, thus providing a flexible strategy based on a customized background and the optional relationship of terms. If users provided a query list and customized background, they could gain GSEA results for their area of interest. DAG graphs determining the relationship between terms are also produced if a reference file for relationships is provided. Cross-comparison is also vital for interpreting the analysis results obtained from experiments involving multiple samples, such as time-series experiments (Figure 3). Users can submit three formats for SEACOMPARE analysis, including customized datasets, result tables from GSEA analyses, and multiple session/job IDs. All of these formats contain multiple numeric values from separate experiments. In addition, an ID conversion tool is available for some species in PlantGSAD.

Application of GSEA with rice OsGRF4 binding genes

In this section, we present an example of the use of PlantGSAD for gene set enrichment analysis (GSEA) of a gene list of interest to illustrate the features of the multiple functional categories in the database. Rice GROWTH-REGULATING FACTOR 4 (OsGRF4) is a positive transcriptional regulator of multiple nitrogen-metabolism genes and coordinates carbon metabolism and growth (48). We utilized a group of target genes associated with OsGRF4 binding peaks identified in a ChIP-seq experiment (48) as a query gene list for GSEA. The OsGRF4 binding peaks cover 387 genes spreading all over the 12 chromosomes of rice (Figure 2A). GSEA was initially conducted with these 387 genes. Gene sets from the GO, TO, PECO, metabolic pathways, and chromatin states categories were significantly enriched.

In the GO (G1) category, there were some enriched biological processes related to key nitrogen-assimilation enzymes (Figure 2B), such as nitrate assimilation (related to assimilation), glutamine family amino acid biosynthetic process, and glutamine biosynthetic process (related to assimilation), which were confirmed by the enrichment analysis results derived from agriGOv2 (Supplementary Table S4). In the TO category, a number of agronomic traits were enriched, especially plant growth and development trait, grain number and grain yield trait (Figure 2C). The pathways such as nitrate reduction II (assimilatory), Nitrogen metabolism, PS.lightreaction were also enriched (Figure 2D). In particular, the enrichment of PS.lightreaction indicated that these query genes had effects on photosynthesis and carbon-assimilation. The GO and TO gene set enrichment analysis results indicated that OsGRF4 regulates downstream genes and coordinates both nitrogen and carbon assimilation, which is consistent with previous research that has reported that OsGRF4 is a positive regulator of plant carbon and nitrogen metabolism, photosynthesis, maintaining the C:N ratio, and promoting plant growth and development (48).

Interestingly, in the analysis of chromatin states based gene sets, we found that CENH3-related and H3K27me3-related genes were significantly enriched (Supplementary Table S4). This may be related to the recently reported enhancement of the sustainable green revolution yield via nitrogen-responsive chromatin modulation in rice (49). NITROGEN-MEDIATED TILLER GROWTH RESPONSE 5 (NGR5) interacts with a component of the polycomb repressive complex 2 (PRC2) and alters the H3K27me3 pattern response to changes in nitrogen availability. There exists a genetic interaction between GRF4 and NGR5, and the accumulation of these proteins can improve the yield and nitrogen-use efficiency of main rice varieties under reduced nitrogen fertilizer loadings (49). Thus, our enrichment analysis results for chromatin states-based gene sets suggest that OsGRF4 might regulate downstream genes via the NGR5-dependent recruitment of PRC2 and the reprogramming of H3K27me3 methylation.

We then conducted overlap matrix analysis for the enriched gene sets to isolate the related essential genes (Figure 2E). These included *SLR1* (50,51)*, sd1* (52,53)*, OsNRT1.1B*

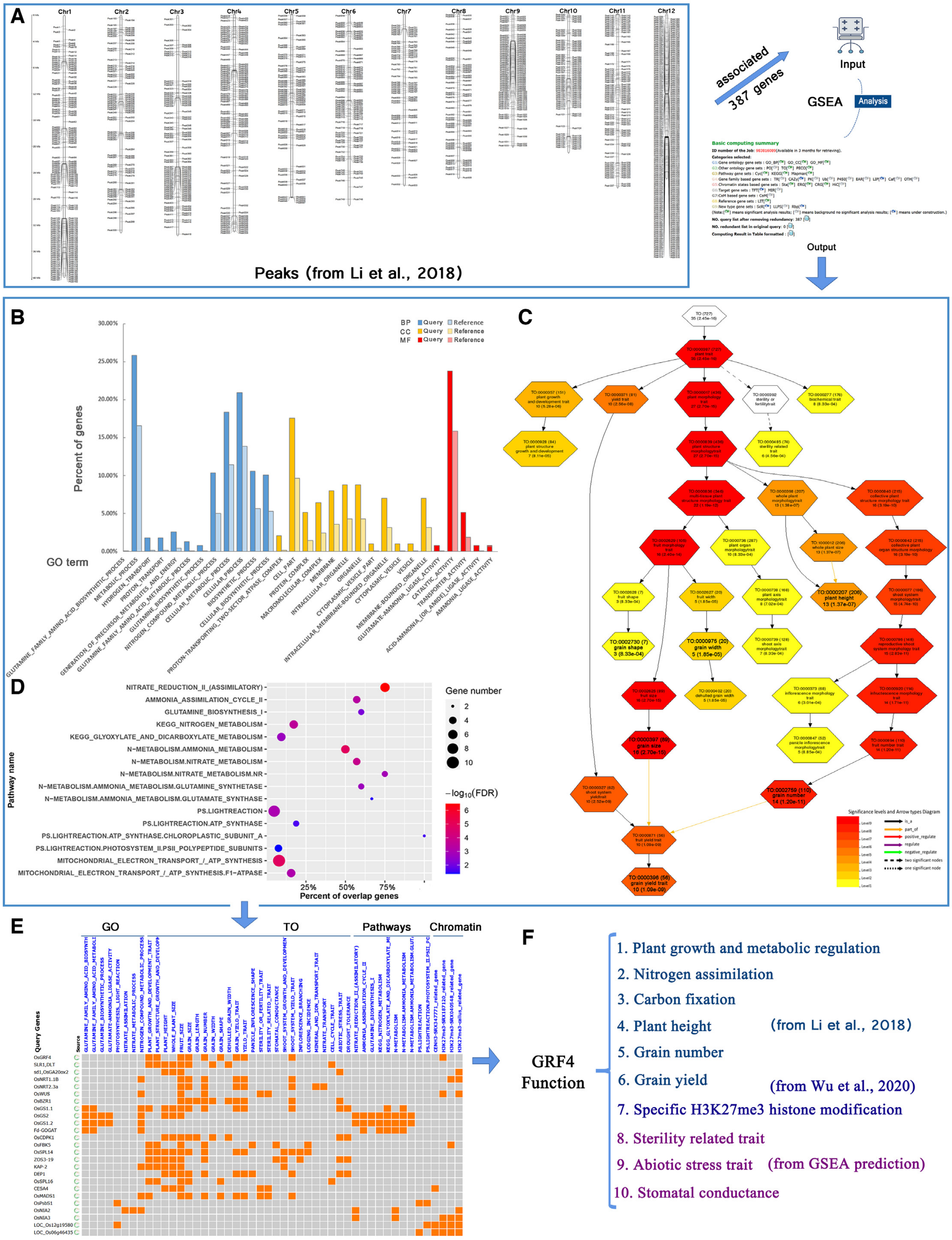


Figure 2. Gene sets enrichment analysis of OsGRF4 binding genes. (**A**) Distribution of OsGRF4 binding peaks in the 12 rice chromosomes. (**B**) Enriched GO terms from the GSEA results. (**C**) Enriched pathways from the GSEA results. (**D**) DAG for the enriched TO terms with a cutoff FDR of 0.001. (**E**) Overlap matrix for GO-, TO-, pathway- and chromatin state-related gene sets. (**F**) Overview of GRF4 function.

(54), *OsNRT2.3a* (55), *OsWUS, OsBZR1* (56), *OsGS1.1, OsGS1.2, OsGS2* (57) and *OsCDPK1* (58,59). These genes physically interact with the rice GRF4 and DELLA protein SLR1 to modulate plant growth and metabolic coregulation. Moreover, *OsNRT1.1B* and *OsNRT2.3a* encode uptake transporters, and *sd1* and *OsBZR1* conferred semi-dwarf leads to GRV resistance to yield-reducing lodging (i.e. the flattening of plants by the wind and rain).

Finally, we summarized known and unknown GRF4 functions from published papers (Figure 2F). Using GSEA in PlantGSAD and follow-up prediction, it was shown that GRF4 not only promotes nitrogen assimilation, carbon fixation, plant height, and grain yield and growth, but also takes part in the response to abiotic stress and other predicted biological processes. In addition, the target genes for OsGRF4 are possibly involved in H3K27me3 histone modification.

The above functional analysis results demonstrated that PlantGSAD is a powerful online database with a large gene set coverage and multiple visual graph tools for the enrichment analysis of functional categories.

Application of SEACOMPARE tool

In PlantGSAD, several singular enrichment analysis (SEA) results from different samples can be compared using the SEACOMPARE tools. We thus investigated dynamic changes in chromatin states and the epigenetic and transcriptional regulators of up-regulated genes under cold treatment with time series in the model plant *Arabidopsis thaliana* (60). The numbers of gene associated with cold treatment of 0.5, 1, 3, 6, 12 and 24 h were 302, 594, 768, 1055, 1214 and 1205, respectively (Figure 3A and Supplementary Table S5). Using enrichment analysis of chromatin state and TFT gene sets, the temporally dynamic response to cold stress at the epigenetic (Figure 3B, C) and transcriptional level (Figure 3D) was determined.

For convenience, we describe here only a selection of the gene sets from the chromatin, ERG and TFT categories (complete SEA comparison results are presented in Supplementary Table S6). Intriguingly, the STATES 11, 12, 14 and 15, mainly corresponding to the repressed epigenetic mark H3K27me3, were enriched for the up-regulated genes at the early stage of treatment (0.5 h). Over time, however, STATES 17, 18, and 19 became enriched (1-24 h), followed by STATES 16, 20 and 21 (i.e. 3, 6, 12 and 24 h). The SEACOMPARE results indicate that the genes modified with the repressed epigenetic mark H3K27me3 tend to function at the beginning, followed by genes with the accessible regions. In the ERG category, the STATES 27, 26 and 22 conformed to genes with H3K4 tri-methylation or dimethylation modifications. STATES 22-24 and 2 are mainly associated with histone acetylation. After 24 h of cold treatment, the STATES 7 and 3-4 were significantly enriched; these were associated with H3K4me1 and H3K36me3 modification (Figure 3B, C).

The SEACOMPARE results for the TFT subcategory showed that, among the genes induced by 1 h of cold treatment, there was an enrichment in the binding sites for CALMODULIN-BINDING TRANSCRIPTIONAL ACTIVATOR 1 (CAMTA1). Previous studies (61,62) have suggested that CAMTA1 and CAMTA2 work in concert with CAMTA3 at low temperature (4°C) to increase the transcript levels of CRT/DRE BINDING FACTOR (CBF) genes after 2 h. The key regulatory CBF pathway confers freezing tolerance in Arabidopsis and other plants (63,64). Our SEACOMPARE results showed that the targets of CBF1, CBF2 and CBF3 were enriched after from 3 h of cold treatment (Figure 3D). Interestingly, the enrichment of target genes of CBFs reached a maximum level at ∼24 h. In addition, several other transcription factors, such as BZIPs, ANACs, and WRKYs may contribute to genes induced at 0.5-24 h cold treatment to enhance tolerance to freezing. For example, WRKY6 has been reported to positively regulate freezing tolerance (65).

Overall, using SEACOMPARE tool on of chromatin states, ERGs and TFTs, the temporally dynamic regulation changes of genes and their progressive waves of transcriptional responses during cold stress were able to be monitored. Thus, this tool allows users to quickly and efficiently decipher the dynamic changes in regulation and biological processes during different developmental stages and in relation to stress responses over time.

DISCUSSION

The emergence of numerous novel data types and high-throughput data sets for plant species over recent years has raised the urgent need to comprehensively integrate and effectively process annotated gene sets. The gene set annotation databases for human and model species PAGED (22) and GeneSetDB (23) were published in 2012, and the platforms DAVID (2,66), MSigDB (5), and WebGestalt (6) offer some functional categories for gene sets. However, they primarily focus on the gene-set annotation for human and model species, and there are a very limited number of plant species in these platforms. Therefore, we created PlantGSAD, a comprehensive gene set annotation database for multiple plant species. PlantGSAD integrates 236 007 gene sets across nine different categories for 44 plant species, consisting primarily of cereal crops, economic crops, and medicinal plants. Compared with other platforms, PlantGSAD covers a more comprehensive range of categories, including chromatin states based gene sets (G5) and new type gene sets (G9) such as LLPS and scRNA-seq gene sets. In addition, PO, TO and other plant-related categories are available in PlantGSAD, increasing the annotating possibilities.

In PlantGSAD, we provide the basic GSEA, which is a powerful functional analysis method for interpreting the biological meaning of a group of genes. We originally developed the web-based PlantGSEA (7) in 2013, which has a user-friendly interface and a performance-efficient framework. We adapted the original GSEA tools from PlantGSEA for use in PlantGSAD, while several new tools were also developed to improve the analytical efficiency, including DAG for ontology categories, SEACOMPARE, and custom SEA. The DAGs intuitively reflect the parent-child relationships in ontology analysis, while SEACOMPARE was developed to allow the cross-comparison of SEA results. The present paper presented an example of this tool being used to investigate the dynamic changes in chromatin

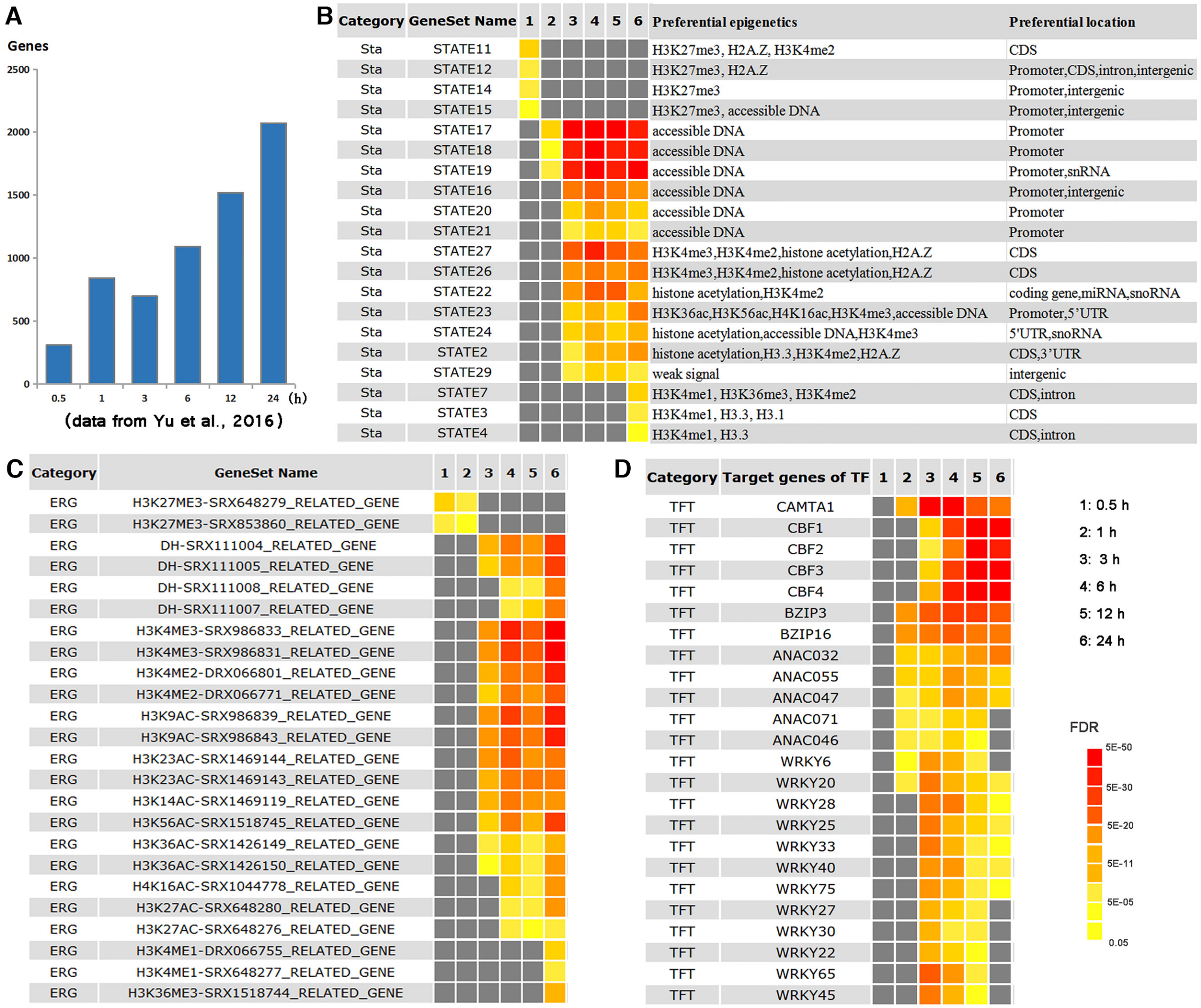


Figure 3. SEACOMPARE analysis of GSEA results for the up-regulated gene list after a 0.5-, 1-, 3-, 6-, 12- or 24-h of cold treatment in *Arabidopsis thaliana*. (**A**) Number of up-regulated genes after a 0.5-, 1-, 3-, 6-, 12- or 24-h of cold treatment in *A. thaliana.* (**B-D**) SEACOMPARE analysis of GSEA results for chromatin states (Sta), epigenetic mark related (ERG), and transcription factor targets (TFT) gene sets. The columns present, from left to right, the subcategory, gene set name, and a heatmap (columns 3–9). The Sta subcategory has two additional annotated columns for preferential epigenetics (right). The colored blocks in the heatmap represent the level of significance for each term. Moving from yellow to red indicates a movement from low to high significance; grey indicates not significant.

states and transcription regulation from time-series transcriptomic data during cold stress in *Arabidopsis*. The custom SEA tool was established to process individual user-defined data on a customized background and the optional relationship of gene sets.

It is important to note that PlantGSAD has some limitations and room for improvement. In particular, there are plans to add more gene sets, data types, species, visualizations, and tools. We will increase the types of gene sets for the supported species from the published literatures and new databases. For example, Li *et al.* (67) recently reported a high-resolution cell atlas for the xylem in *Populus* using scRNA-seq, which has already been added to our database as a key ScR gene set for *Populus*. We plan to routinely update the genome version of the available plant species around every six months. We also plan to add gene sets for new plant species in PlantGSAD, and this process may take longer time due to the accumulation of various types of gene sets from different resources. Furthermore, we have set up the custom SEA tool to satisfy the urgent requirements for the data types or the species that were not included in the existing data sets.

In PlantGSAD, the large number and the wide variety of gene sets from multiple species, together with the high-efficiency analysis tools and visualization features, greatly assist the biological interpretation of genes of interest in plants. We believe that the multidimensional plant gene annotations in PlantGSAD can be used to conveniently interpret the biological meaning of a gene via its connectivity and relationships, thus it represents a promising multifunctional platform for the functional analysis of plant genes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [31970629, 32170673, 31771467]. Funding for open access charge: National Natural Science Foundation of China [31970629, 32170673, 31771467].

Conflict of interest statement. None declared.

REFERENCES

1. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
2. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, R60.
3. Tian,T., Liu,Y., Yan,H., You,Q., Yi,X., Du,Z., Xu,W. and Su,Z. (2017) agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.*, **45**, W122–W129.
4. Du,Z., Zhou,X., Ling,Y., Zhang,Z. and Su,Z. (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.*, **38**, W64–W70.
5. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
6. Wang,J., Vasaikar,S., Shi,Z., Greer,M. and Zhang,B. (2017) WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.*, **45**, W130–W137.
7. Yi,X., Du,Z. and Su,Z. (2013) PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res.*, **41**, W98.
8. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
9. Alexeyenko,A., Lee,W., Pernemalm,M., Guegan,J., Dessen,P., Lazar,V., Lehtio,J. and Pawitan,Y. (2012) Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, **13**, 226.
10. Jeggari,A., Alekseenko,Z., Petrov,I., Dias,J.M., Ericson,J. and Alexeyenko,A. (2018) EviNet: a web platform for network enrichment analysis with flexible definition of gene sets. *Nucleic Acids Res.*, **46**, W163–W170.
11. You,Q., Xu,W., Zhang,K., Zhang,L., Yi,X., Yao,D., Wang,C., Zhang,X., Zhao,X., Provart,N.J. *et al.* (2017) ccNET: Database of co-expression networks with functional modules for diploid and polyploid Gossypium. *Nucleic Acids Res.*, **45**, D1090–D1099.
12. Tian,T., You,Q., Yan,H., Xu,W. and Su,Z. (2018) MCENet: A database for maize conditional co-expression network and network characterization collaborated with multi-dimensional omics levels. *J. Genet. Genomics*, **45**, 351–360.
13. Deng,X., Song,X.W., Wei,L.Y., Liu,C.Y. and Cao,X.F. (2016) Epigenetic regulation and epigenomic landscape in rice. *Natl. Sci. Rev.*, **3**, 309–327.
14. Zheng,Y. and Liu,X. (2019) Review: Chromatin organization in plant and animal stem cell maintenance. *Plant Sci.*, **281**, 173–179.
15. Dong,Q., Li,N., Li,X., Yuan,Z., Xie,D., Wang,X., Li,J., Yu,Y., Wang,J., Ding,B. *et al.* (2018) Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. *Plant J.*, **94**, 1141–1156.
16. Dong,P., Tu,X., Chu,P.Y., Lu,P., Zhu,N., Grierson,D., Du,B., Li,P. and Zhong,S. (2017) 3D chromatin architecture of large plant genomes determined by local A/B compartments. *Mol. Plant*, **10**, 1497–1509.
17. Dogan,E.S. and Liu,C. (2018) Three-dimensional chromatin packing and positioning of plant genomes. *Nat. Plants*, **4**, 521–529.
18. Dong,P., Tu,X., Li,H., Zhang,J., Grierson,D., Li,P. and Zhong,S. (2020) Tissue-specific Hi-C analyses of rice, foxtail millet and maize suggest non-canonical function of plant chromatin domains. *J. Integr. Plant Biol.*, **62**, 201–217.
19. Boija,A., Klein,I.A., Sabari,B.R., Dall'Agnese,A., Coffey,E.L., Zamudio,A.V., Li,C.H., Shrinivas,K., Manteiga,J.C., Hannett,N.M. *et al.* (2018) Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell*, **175**, 1842–1855.
20. Hnisz,D., Shrinivas,K., Young,R.A., Chakraborty,A.K. and Sharp,P.A. (2017) A phase separation model for transcriptional control. *Cell*, **169**, 13–23.
21. Sabari,B.R., Dall'Agnese,A., Boija,A., Klein,I.A., Coffey,E.L., Shrinivas,K., Abraham,B.J., Hannett,N.M., Zamudio,A.V., Manteiga,J.C. *et al.* (2018) Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, **361**, eaar3958.
22. Huang,H., Wu,X., Sonachalam,M., Mandape,S.N., Pandey,R., MacDorman,K.F., Wan,P. and Chen,J.Y. (2012) PAGED: a pathway and gene-set enrichment database to enable molecular phenotype discoveries. *BMC Bioinformatics*, **13**, S2.
23. Araki,H., Knapp,C., Tsai,P. and Print,C. (2012) GeneSetDB: a comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, **2**, 76–82.
24. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
25. Goodstein,D.M., Shu,S.Q., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
26. Cooper,L., Meier,A., Laporte,M.A., Elser,J.L., Mungall,C., Sinn,B.T., Cavaliere,D., Carbon,S., Dunn,N.A., Smith,B. *et al.* (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Res.*, **46**, D1168–D1180.
27. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
28. Schlapfer,P., Zhang,P.F., Wang,C.A., Kim,T., Banf,M., Chae,L., Dreher,K., Chavali,A.K., Nilo-Poyanco,R., Bernard,T. *et al.* (2017) Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.*, **173**, 2041–2059.
29. Usadel,B., Poree,F., Nagel,A., Lohse,M., Czedik-Eysenberg,A. and Stitt,M. (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant Cell Environ.*, **32**, 1211–1229.
30. Zhou,J.Q., Xu,Y., Lin,S.F., Guo,Y.P., Deng,W.K., Zhang,Y., Guo,A. and Xue,Y. (2018) iUUCD 2.0: an update with rich annotations for ubiquitin and ubiquitin-like conjugations. *Nucleic Acids Res.*, **46**, D447–D453.
31. Liu,Y., Tian,T., Zhang,K., You,Q., Yan,H., Zhao,N., Yi,X., Xu,W. and Su,Z. (2018) PCSD: a plant chromatin state database. *Nucleic Acids Res.*, **46**, D1157–D1167.
32. Sato,Y., Namiki,N., Takehisa,H., Kamatsuki,K., Minami,H., Ikawa,H., Ohyanagi,H., Sugimoto,K., Itoh,J., Antonio,B.A. *et al.* (2013) RiceFREND: a platform for retrieving coexpressed gene networks in rice. *Nucleic Acids Res.*, **41**, D1214–D1221.
33. Adamcsek,B., Palla,G., Farkas,I.J., Derényi,I. and Vicsek,T. (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**, 1021–1023.
34. Chen,S.Y., Feng,Z. and Yi,X. (2017) A general introduction to adjustment for multiple comparisons. *J Thorac Dis*, **9**, 1725–1729.
35. Pan,Q., Wei,J., Guo,F., Huang,S., Gong,Y., Liu,H., Liu,J. and Li,L. (2019) Trait ontology analysis based on association mapping studies bridges the gap between crop genomics and Phenomics. *BMC Genomics*, **20**, 443.

36. Klie,S. and Nikoloski,Z. (2012) The choice between mapman and gene ontology for automated gene function prediction in plant science. *Front. Genet.*, **3**, 115.
37. Yang,J.T., Liu,Y., Yan,H.Y., Tian,T., You,Q., Zhang,L.W., Xu,W.Y. and Su,Z. (2018) PlantEAR: functional analysis platform for plant EAR Motif-Containing proteins. *Front. Genet.*, **9**, 590.
38. Song,L., Huang,S.C., Wise,A., Castanon,R., Nery,J.R., Chen,H., Watanabe,M., Thomas,J., Bar-Joseph,Z. and Ecker,J.R. (2016) A transcription factor hierarchy defines an environmental stress response network. *Science*, **354**, aag1550.
39. Tu,X.Y., Mejia-Guerra,M.K., Franco,J.A.V., Tzeng,D., Chu,P.Y., Shen,W., Wei,Y.Y., Dai,X.R., Li,P.H., Buckler,E.S. *et al.* (2020) Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nat. Commun.*, **11**, 5089.
40. Yen,A. and Kellis,M. (2015) Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nat. Commun.*, **6**, 7973.
41. Dong,Q.L., Wang,J.B., Li,X.C. and Gong,L. (2020) Progresses in the plant 3D chromatin architecture. *Yi Chuan*, **42**, 73–86.
42. Marondedze,C. (2020) The increasing diversity and complexity of the RNA-binding protein repertoire in plants. *Proc. Biol. Sci.*, **287**, 20201397.
43. Foley,S.W., Gosai,S.J., Wang,D., Selamoglu,N., Sollitti,A.C., Köster,T., Steffen,A., Lyons,E., Daldal,F., Garcia,B.A. *et al.* (2017) A global view of RNA-Protein interactions identifies post-transcriptional regulators of root hair cell fate. *Dev. Cell*, **41**, 204–220.
44. Bach-Pages,M., Homma,F., Kourelis,J., Kaschani,F., Mohammed,S., Kaiser,M., van der Hoorn,R.A.L., Castello,A. and Preston,G.M. (2020) Discovering the RNA-binding proteome of plant leaves with an improved RNA interactome capture method. *Biomolecules*, **10**, 661.
45. Marondedze,C., Thomas,L., Serrano,N.L., Lilley,K.S. and Gehring,C. (2016) The RNA-binding protein repertoire of *Arabidopsis thaliana*. *Sci. Rep.*, **6**, 29766.
46. Koster,T., Marondedze,C., Meyer,K. and Staiger,D. (2017) RNA-Binding proteins revisited - the emerging arabidopsis mRNA interactome. *Trends Plant Sci.*, **22**, 512–526.
47. She,J., Yan,H., Yang,J., Xu,W. and Su,Z. (2019) croFGD: *Catharanthus roseus* functional genomics database. *Front. Genet.*, **10**, 238.
48. Li,S., Tian,Y., Wu,K., Ye,Y., Yu,J., Zhang,J., Liu,Q., Hu,M., Li,H., Tong,Y. *et al.* (2018) Modulating plant growth-metabolism coordination for sustainable agriculture. *Nature*, **560**, 595–600.
49. Wu,K., Wang,S., Song,W., Zhang,J., Wang,Y., Liu,Q., Yu,J., Ye,Y., Li,S., Chen,J. *et al.* (2020) Enhanced sustainable green revolution yield via nitrogen-responsive chromatin modulation in rice. *Science*, **367**, eaaz2046.
50. Harberd,N.P., Belfield,E. and Yasumura,Y. (2009) The angiosperm gibberellin-GID1-DELLA growth regulatory mechanism: how an "inhibitor of an inhibitor" enables flexible response to fluctuating environments. *Plant Cell*, **21**, 1328–1339.
51. Peng,J., Richards,D.E., Hartley,N.M., Murphy,G.P., Devos,K.M., Flintham,J.E., Beales,J., Fish,L.J., Worland,A.J., Pelica,F. *et al.* (1999) 'Green revolution' genes encode mutant gibberellin response modulators. *Nature*, **400**, 256–261.
52. Sasaki,A., Ashikari,M., Ueguchi-Tanaka,M., Itoh,H., Nishimura,A., Swapan,D., Ishiyama,K., Saito,T., Kobayashi,M., Khush,G.S. *et al.* (2002) Green revolution: a mutant gibberellin-synthesis gene in rice. *Nature*, **416**, 701–702.
53. Spielmeyer,W., Ellis,M.H. and Chandler,P.M. (2002) Semidwarf (sd-1), "green revolution" rice, contains a defective gibberellin 20-oxidase gene. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 9043–9048.
54. Zhang,J.Y., Liu,Y.X., Zhang,N., Hu,B., Jin,T., Xu,H.R., Qin,Y., Yan,P.X., Zhang,X.N., Guo,X.X. *et al.* (2019) NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. *Nat. Biotechnol.*, **37**, 676–684.
55. Liu,X.Q., Huang,D.M., Tao,J.Y., Miller,A.J., Fan,X.R. and Xu,G.H. (2014) Identification and functional assay of the interaction motifs in the partner protein OsNAR2.1 of the two-component system for high-affinity nitrate transport. *New Phytol.*, **204**, 74–80.
56. Fang,Z., Ji,Y., Hu,J., Guo,R., Sun,S. and Wang,X. (2020) Strigolactones and brassinosteroids antagonistically regulate the stability of the D53-OsBZR1 complex to determine FC1 expression in rice tillering. *Mol. Plant*, **13**, 586–597.
57. Hu,J., Wang,Y., Fang,Y., Zeng,L., Xu,J., Yu,H., Shi,Z., Pan,J., Zhang,D., Kang,S. *et al.* (2015) A rare allele of GS2 enhances grain size and grain yield in rice. *Mol. Plant*, **8**, 1455–1465.
58. Ho,S.L., Huang,L.F., Lu,C.A., He,S.L., Wang,C.C., Yu,S.P., Chen,J.C. and Yu,S.M. (2013) Sugar starvation- and GA-inducible calcium-dependent protein kinase 1 feedback regulates GA biosynthesis and activates a 14-3-3 protein to confer drought tolerance in rice seedlings. *Plant Mol. Biol.*, **81**, 347–361.
59. Asano,T., Tanaka,N., Yang,G.X., Hayashi,N. and Komatsu,S. (2005) Genome-wide identification of the rice calcium-dependent protein kinase and its closely related kinase gene families: comprehensive analysis of the CDPKs gene family in rice. *Plant Cell Physiol.*, **46**, 356–366.
60. Yu,J., Zhang,Y., Di,C., Zhang,Q., Zhang,K., Wang,C., You,Q., Yan,H., Dai,S.Y., Yuan,J.S. *et al.* (2016) JAZ7 negatively regulates dark-induced leaf senescence in Arabidopsis. *J. Exp. Bot.*, **67**, 751–762.
61. Kim,Y., Park,S., Gilmour,S.J. and Thomashow,M.F. (2013) Roles of CAMTA transcription factors and salicylic acid in configuring the low-temperature transcriptome and freezing tolerance of Arabidopsis. *Plant J.*, **75**, 364–376.
62. Doherty,C.J., Van Buskirk,H.A., Myers,S.J. and Thomashow,M.F. (2009) Roles for *Arabidopsis* CAMTA transcription factors in cold-regulated gene expression and freezing tolerance. *Plant Cell*, **21**, 972–984.
63. Thomashow,M.F. (2010) Molecular basis of plant cold acclimation: insights gained from studying the CBF cold response pathway. *Plant Physiol.*, **154**, 571–577.
64. Knight,M.R. and Knight,H. (2012) Low-temperature perception leading to gene expression and cold tolerance in higher plants. *New Phytol.*, **195**, 737–751.
65. Li,H., Ye,K., Shi,Y., Cheng,J., Zhang,X. and Yang,S. (2017) BZR1 positively regulates freezing tolerance via CBF-dependent and CBF-independent pathways in arabidopsis. *Mol. Plant*, **10**, 545–559.
66. Huang,D., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
67. Li,H., Dai,X., Huang,X., Xu,M., Wang,Q., Yan,X., Sederoff,R.R. and Li,Q. (2021) Single-cell RNA sequencing reveals a high-resolution cell atlas of xylem in Populus. *J. Integr. Plant Biol.*, https://doi.org/10.1111/jipb.13159.
68. Jin,J., Tian,F., Yang,D.C., Meng,Y.Q., Kong,L., Luo,J. and Gao,G. (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.*, **45**, D1040–D1045.
69. Lombard,V., Ramulu,H.G., Drula,E., Coutinho,P.M. and Henrissat,B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42**, D490–D495.
70. Zheng,Y., Jiao,C., Sun,H., Rosli,H.G., Pombo,M.A., Zhang,P., Banf,M., Dai,X., Martin,G.B., Giovannoni,J.J. *et al.* (2016) iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant*, **9**, 1667–1670.
71. Du,Z., Zhou,X., Li,L. and Su,Z. (2009) plantsUPS: a database of plants' ubiquitin proteasome system. *BMC Genomics*, **10**, 227.
72. Tian,T., You,Q., Zhang,L., Yi,X., Yan,H., Xu,W. and Su,Z. (2016) SorghumFDB: sorghum functional genomics database with multidimensional network analysis. *Database (Oxford)*, **2016**, baw099.
73. Li-Beisson,Y., Shorrosh,B., Beisson,F., Andersson,M.X., Arondel,V., Bates,P.D., Baud,S., Bird,D., Debono,A., Durrett,T.P. *et al.* (2013) Acyl-lipid metabolism. *Arabidopsis Book*, **11**, e0161.
74. Yu,J., Zhang,Z.H., Wei,J.G., Ling,Y., Xu,W.Y. and Su,Z. (2014) SFGD: a comprehensive platform for mining functional information from soybean transcriptome data and its use in identifying acyl-lipid metabolism pathways. *BMC Genomics*, **15**, 271.
75. Gendler,K., Paulsen,T. and Napoli,C. (2008) ChromDB: the chromatin database. *Nucleic Acids Res.*, **36**, D298–D302.
76. Yilmaz,A., Mejia-Guerra,M.K., Kurz,K., Liang,X., Welch,L. and Grotewold,E. (2011) AGRIS: the arabidopsis gene regulatory information server, an update. *Nucleic Acids Res.*, **39**, D1118–D1122.

77. Lai,L.M., Liberzon,A., Hennessey,J., Jiang,G.X., Qi,J.L., Mesirov,J.P. and Ge,S.X. (2012) AraPath: a knowledgebase for pathway analysis in Arabidopsis. *Bioinformatics*, **28**, 2291–2292.
78. Yi,X., Zhang,Z., Ling,Y., Xu,W. and Su,Z. (2015) PNRD: a plant non-coding RNA database. *Nucleic Acids Res.*, **43**, D982–D989.
79. Obayashi,T., Aoki,Y., Tadaka,S., Kagaya,Y. and Kinoshita,K. (2018) ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant Cell Physiol.*, **59**, e3.
80. Ning,W., Guo,Y., Lin,S., Mei,B., Wu,Y., Jiang,P., Tan,X., Zhang,W., Chen,G., Peng,D. *et al.* (2019) DrLLPS: a data resource of liquid–liquid phase separation in eukaryotes. *Nucleic Acids Res.*, **48**, D288–D295.
81. Sun,T., Li,Q., Xu,Y., Zhang,Z., Lai,L. and Pei,J. (2019) In: *Prediction of Liquid-Liquid Phase Separation Proteins Using Machine Learning*. Social Science Electronic Publishing.