DR YIJING ZHANG (Orcid ID : 0000-0001-9568-9389)

Article type : Technical Advance

Plant Regulomics: A Data-driven Interface for Retrieving Upstream Regulators from Plant Multi-omics Data

Xiaojuan Ran[1,2*], Fei Zhao[1,2*], Yuejun Wang[1,2*], Jian Liu[1,2], Yili Zhuang[1,2], Luhuan Ye[1,2], Meifang Qi[1,2], Jingfei Cheng[1,2], Yijing Zhang[1,2†]

[1] National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 300 Fenglin Road, Shanghai 200032, China

[2] University of the Chinese Academy of Sciences, Beijing, 100049, China

[*] **Authors contributed equally to this work.**

[†] **Author for correspondence**

E-mail: zhangyijing@sibs.ac.cn

Phone: +86-21-54924206

Fax: +86-21-54924015

Running title: Plant Regulomics for retrieving upstream regulators

Keywords: Plant Regulomics, data-driven, interface, upstream regulators, omics data

ABSTRACT

High-throughput technology has become a powerful approach for routine plant research. Interpreting the biological significance of high-throughput data has largely focused on the functional characterization of a large gene list or genomic loci, which involves the following two aspects: the functions of the genes or loci and how they are regulated as a whole, i.e. searching for the upstream regulators. Traditional platforms for functional annotation largely help resolving the first issue. Addressing the second issue is essential for a global understanding of the regulatory mechanism, but is more challenging, which requires additional high-throughput experimental evidence and a unified statistical framework for data-mining. The rapid accumulation of omics data provides a large amount of experimental data. We herein present Plant Regulomics, which integrates 19,925 transcriptomic and epigenomic data sets and diverse sources of functional evidence (58,112 terms and 695,414 protein-protein interactions) from six plant species along with the orthologous genes from 56 whole-genome sequenced plant species. All pair-wise transcriptomic comparisons with biological significance within the same study were performed, and all epigenomic data were processed to genomic loci targeted by various factors. These data were well-organized to gene modules and loci lists, which were further implemented into the same statistical framework. For any input gene list or genomic loci, Plant Regulomics retrieves the upstream factors, treatments, and experimental/environmental conditions regulating the input from the integrated omics data. Additionally, multiple tools and an interactive visualization are available through a user-friendly web interface. Plant Regulomics is available at http://bioinfo.sibs.ac.cn/plant-regulomics.

INTRODUCTION

With the rapid release of whole-genome sequences from various plant species, high-throughput technology has become a routine approach for plant research. However, interpreting the biological significance from high-throughput data is challenging. The central requirement is to deduce reliable functional information and regulatory network details from large gene lists and genomic loci derived from high-throughput technologies. Traditional way for high-throughput functional dissection of a gene list or genomic loci focused on the role of these genes or loci. The basic idea is to find the common domains (Hunter *et al.*, 2009), pathways (Kanehisa and Goto, 2000), or functional terms

(Ashburner *et al.*, 2000) shared by these genes or genes nearby the genomic loci. While the regulatory information about how these genes or loci are affected overall by other factors or perturbations is important, it is difficult to investigate. For example, it is relatively easy to obtain information about the major roles of a set of genes based on the enriched functional terms, but deducing the upstream factors regulating these genes may be problematic, because it requires additional high-throughput experimental evidence and a unified statistical framework for efficient data-mining.

The accumulation of plant omics data provides considerable experimental information for mechanistic investigations, and is a promising development for the exploration of upstream factors based on the newly generated data. For example, differentially expressed gene sets derived from transcriptomic comparisons represent genes with consistent responses to a specific treatment or perturbation. Genes commonly targeted by the same transcription factor (TF) according to chromatin immunoprecipitation sequencing (ChIP-seq) or DNA affinity purification sequencing (DAP-seq) experiments are more likely to be co-regulated by a given TF. Additionally, genes affected by a series of polymorphisms associated with a certain trait identified from genome-wide association studies tend to co-regulate the trait of interest. Moreover, if one protein interacting with a large fraction of a list of genes, it is most likely that the protein is one key regulator of the gene list. However, integrating publicly available high-throughput data to annotate newly generated data can be difficult. The major challenge is ensuring data from different studies and platforms comparable. A gene-set oriented approach, based on a group of relevant genes instead of an individual gene, has increased the chances of identifying the correct biological processes, and has become an essential part of current genomic analyses. For example, DAVID Bioinformatics Resources (Huang *et al.*, 2007), Molecular Signatures Database (MsigDB) (Liberzon *et al.*, 2011), and Metascape (Tripathi *et al.*, 2015), which are the most popular web-based platforms for the functional characterization of high-throughput data for animals, are based on the comprehensive collection and curation of gene-sets representing diverse biological processes. Major plant functional exploration platforms, including plantGSEA (Yi *et al.*, 2013) and AgriGO (Du *et al.*, 2010), are also gene-set oriented. However, the available tools for plants are mostly based on a homology with sequences or domains whose functions have been well-characterized in a few model plants. Given the common phenomenon of genome-level duplications followed by myriad fractionation processes in plants (Jiao *et al.*, 2011, Wendel *et al.*, 2016), functional dissections based only on sequence homologies are far from accurate. Additionally, the current platforms in plants is focused on determining the function of input genes instead of how they are regulated. Thus, applying a gene-set oriented approach to integrate direct experimental evidence from high-throughput data and developing reliable tools for exploring the upstream regulatory factors from multi-omics data are critical.

In addition to genes, it is now well-acknowledged that the regulatory elements present in intergenic regions are indispensable for determining phenotypes. For example, a systematic survey of human GWAS variants revealed a vast majority of common disease-associated variations lie within intergenic regulatory elements, which preferentially comprise regulatory epigenetic marks and TFs (Maurano *et al.*, 2012). On the basis of ChIP-seq experiments, genomic loci commonly targeted by specific factors contain important regulatory elements working in concert. However, analyses of these types of loci were mostly restricted to the function of nearby genes. Directly comparing these loci may provide more regulatory information, and the most challenging issue involves statistically comparing the significance of the overlap between loci derived from different studies and different platforms.

In this study, we present Plant Regulomics, which is a platform for retrieving upstream regulators from multi-omics data for 62 whole-genome sequenced plant species. A total of 19,043 transcriptomic and 1,694 epigenomic data sets from public databases were processed following a unified pipeline. All pair-wise transcriptomic comparisons with biological significance within the same study were performed, and all epigenomic data were processed to genomic loci targeted by various factors. These data sets were further organized to gene sets and genomic loci potentially involved in the same biological processes. The data analyses were integrated into the same statistical framework, which enabled cross-platform and cross-study comparisons. We further illustrate the features of Plant Regulomics using four examples, demonstrating that Plant Regulomics is a highly useful interface for the intensive data-mining of plants.

RESULTS

Plant Regulomics scheme

Figure 1 presents the data integration process and the operating principle of Plant Regulomics. Briefly, 19,043 transcriptomic and 1,694 epigenomic data sets characterizing the binding patterns of TFs and epigenetic factors in six species (*A. thaliana*, *O. sativa*, *Z. mays*, *G. max,* *S. lycopersicum* and *T. aestivum*) were collected from the GEO (Barrett *et al.*, 2013), NASCArrays (Craigon *et al.*, 2004) and DDBJ (Mashima *et al.*, 2016) databases. These data were further processed, resulting in 12,070 differentially expressed gene sets, which were derived from all pair-wise transcriptomic comparisons with biological significance within the same study, as well as 1,691 gene sets or genomic loci targeted by TFs and epigenetic factors. A total of 695,414 protein-protein interactions from three species including *A. thaliana*, *O. sativa* and *Z. mays* were downloaded from TAIR (Huala *et al.*, 2001, Lamesch *et al.*, 2012), BioGRID (Stark *et al.*, 2006, Oughtred *et al.*, 2018), DIPOS (Sapkota *et al.*, 2011), PRIN (Gu *et al.*, 2011) and PPIM (Zhu *et al.*, 2016). Orthologous gene pairs between five reference species (*A. thaliana*, *O. sativa*, *Z. mays*, *G. max*, and *S. lycopersicum*) and another 56 whole-genome sequenced species were incorporated. Supplemental Table 1 summarizes the data collection statistics. For any input gene list or genomic loci, Plant Regulomics returns the TFs, epigenetic factors, cis-elements, protein interactors, perturbation of genes, and experimental/environmental conditions regulating the input list. The omics data can be further visualized *via* a customized genome browser.

Case Studies

Case study 1. Searching for upstream regulators for single gene

The typical feature of Plant Regulomics involves the detection of treatments or perturbations, trans-acting factors, and cis-elements regulating the input gene(s) or locus/loci. We illustrate searching for upstream regulators for single gene using *FLOWERING LOCUS T* (*FT*) (Figure 2A) as the input. This gene is essential for promoting flowering in response to day-length changes (Turck *et al.*, 2008). Figure 2B lists the treatments, mutants, or ecotypes affecting *FT* expression, as well as the statistics for the transcriptional changes. The expression of *FT* is primarily regulated by day-length and light density (Figure 2B), which has been well-documented (Kardailsky *et al.*, 1999, Kim *et al.*, 2008). Factors affecting *FT* expression include *ABI1*, *ABA1*, *LHP1*, *LFY*, *AP3*, and *COP1* (Figure 2B), some of which have been confirmed by previous experiments (O'Maoileidigh *et al.*, 2014, Riboni *et al.*, 2016). Three types of cis-elements were present in the *FT* promoter region (Figure 2C), including those potentially bound by MADS box factors, C2H2 zinc finger factors, and the AP2/ERF domain. Some factors from these TF families are typically involved in flowering and floral development (Wellmer *et al.*, 2006). Figure 2D lists the TFs and epigenetic marks enriched at the *FT* locus, including LHP1, LHY, TRB1, and WRKY family TFs and the H3K27me3 modification. Previous studies revealed that LHP1 and TRB1 are the central components responsible for recruiting Polycomb group proteins catalyzing the H3K27me3 reaction (Turck *et al.*, 2007, Zhang *et al.*, 2007, Zhou *et al.*, 2018). This result is consistent with those of previous studies that indicated the activation of FT is counteracted by LHP1-mediated H3K27me3 (Adrian *et al.*, 2010, Farrona *et al.*, 2011). The genomic track in Figure 2E illustrates the co-occupancy of these factors or epigenetic marks surrounding the *FT* locus. Figure 2F listed three proteins interacting with *FT*. The interactions between GI (AT1G22770), AJH1 (AT1G22920), ATBIZP27 (AT2G17770) and *FT* have been reported previously (Abe *et al.*, 2005, Mizoguchi *et al.*, 2005, Arabidopsis Interactome Mapping Consortium, 2011). All interactors were listed in Supplemental Table 2.

Case study 2. Searching for upstream regulators for gene list

We illustrate the application of the platform for searching upstream regulators for the ABA-induced genes previously published (Figure 3A) (Song *et al.*, 2016). The input gene list was compared with all 18,155 gene sets integrated in the database. Figure 3B lists the top enriched transcriptomic comparisons; the input genes were enriched for the genes whose expressions are affected by these transcriptomic comparisons, including those involving ABA treatments with the ABA-related hormones phaseic acid (PA) and dexamethasone (DEX), as well as abiotic stresses, including darkness, drought, salt, and injury. These results are consistent with the essential role of ABA in plant responses to abiotic stimuli (Figure 3B). The genes on the input list are preferentially expressed in dry seeds (Figure 3C), consistent with the essential role of ABA in seed dormancy (Finkelstein *et al.*, 2008). The top enriched TFs binding the loci surrounding these genes are ABA-related, including ABI5, ABF4, ABF1, ANAC102, HSFA6A, FBH3 and ERF48 as determined by ChIP-seq and DAP-seq (Figure 3D). The motifs over-represented in these genes were divided into two classes, namely AP2/ERF type motif and bHLH/G-box motif (Figure 3F), both of which are major cis-elements regulating ABA responses (Shen and Ho, 1995). It is worth noting that a large proportion of the input genes were enriched for H3K27me3 marks (Figure 3D, E). This is consistent with a recent report stating that H3K27me3 is responsible for attenuating ABA-induced senescence (Liu *et al.*, 2018). Figure 3G displayed the interactors associated with user input genes sorted by the number of interactions, which are the potential hub regulators of input genes. The top five interactors are shown here and all interactors could be downloaded. Among the top interactors, UBIQUITIN-CONJUGATING ENZYME 34 (UBC34, AT1G17280) is well-acknowledged to play negative role in drought response (Ahn et al., 2018), and it is interesting to examine whether the input genes interacting with UBC34 are subjected to ubiquitination-mediated degradation pathway and the association with stress responses.

Case study 3. Searching for upstream regulators of genomic loci

To illustrate the application of searching for upstream regulators of genomic loci, the previously published ChIP-seq peaks of *PRR7* were used as the input (Liu *et al.*, 2013) (Figure 4A, B); PRR7 is the central day-phased circadian factor with thousands of target loci. A comparison with ChIP-seq data integrated in Plant Regulomics revealed that 8 factors shared a significant proportion of target loci with PRR7 (Figure 4C), most of which are involved in light signaling, including *PHYA*, *HY5*, *PIF*s, *TOC1*, and *LHY*. The common and unique genes targeted by user-selected factors may be further visualized with a heatmap. The genomic track in Figure 4D illustrates the co-occupation of PRR7 and other factors. The top over-represented motifs of the PRR7-binding loci included bZIP and bHLH factors (Figure 4E). The nearby genes potentially targeted by input loci were downloaded (Figure 4A), whose function were further explored *via* gene list annotation implemented in Plant Regulomics. The PRR7 targets are mainly associated with photosynthesis and chloroplast development (Figure 4F). Consistently, the top enriched transcriptomic comparisons involved light treatments (Figure 4G). A closer examination of the transcriptomic comparisons of mutants revealed that the expression levels of the PRR7 targets are preferentially up-regulated in the *PRR5,7,9* triple mutant (Figure 4G). These observations suggested that PRRs are responsible for repressing light-induced photosynthesis, confirming a previously proposed hypothesis (Fukushima *et al.*, 2009).

Case study 4. Gene expression pattern of orthologous genes

To illustrate the application of an analysis of orthologous genes, we used pathogen-induced poplar genes as the input (Azaiez *et al.*, 2009), with *A. thaliana* and *O. sativa* serving as reference species (Figure 5A). The enriched transcriptomic comparisons were further clustered based on relevance (Method) (Figure 5B and Supplemental Table 3). The top enriched clusters in both *A. thaliana* and *O. sativa* included the transcriptomic changes induced by pathogens, as well as changes affected by the deprivation of specific factors involved in pathogen response. These results indicated that the functions of genes in these clusters are evolutionarily conserved. We next used maize pollen-specific genes as the input, and observed that the orthologous genes in *O. sativa* exhibited an apparent anther-biased expression, while the tissue specificity of the orthologous *A. thaliana* genes was unclear (Figure 5C and Supplemental Figure 1). This distinct tissue-biased expression pattern suggested pollen development differs between monocotyledonous and dicotyledonous plant species (Wilson and Zhang, 2009). Thus, a comparison of high-throughput data across species implemented in Plant Regulomics provides insightful clues from an evolutionary perspective.

Comparison of Plant Regulomics with other similar tools

We summarized the typical features of Plant Regulomics with other popular web-based databases or tools for plant gene annotation in Supplemental Table 4, including Gene-Sharing Networks (Li *et al.*, 2012), AraPath (Lai *et al.*, 2012), ShinyGO (Ge and Jung, 2018), AtCAST (Kakei and Shimada, 2014), AIV2 (Dong *et al.*, 2019), ePlant (Waese *et al.*, 2017), Expressolog TreeViewer (Patel *et al.*, 2012) and GENEVESTIGATOR (Zimmermann *et al.*, 2004, Hruz *et al.*, 2008).

DISCUSSION

Plant Regulomics focused on retrieving the upstream regulators, which is different from traditional platforms from the following aspects: Firstly, it is a data-driven platform. To understand how a gene or gene list is regulated requires comprehensive collection and organization of available experimental evidence. Instead of mere electronic annotation based on sequence homology with well-characterized genes or domains, Plant Regulomics integrates thousands sets of high throughput data from 6 plant species and corresponding orthologous genes from 56 whole-genome sequenced plant species. For any input gene list or genomic loci, Plant Regulomics retrieves the factors, treatments, and experimental/environmental conditions regulating the input from the integrated omics data. Secondly, it provides a unified statistical framework which facilitates cross-platform and cross-study comparisons. All pair-wise transcriptomic comparisons with biological significance within the same study were performed, which were further organized to gene sets regulated by gene perturbations or treatments; all epigenomic data were processed to genomic loci targeted by various factors. Further refined statistical tests allows for comparison of high-throughput data from any platforms or studies. In this way, the newly generated data may be easily compared with previously generated data, thereby enabling researchers to decipher the transcriptional regulatory network and comprehensively interpret the biological significance based on newly generated omics data with high confidence. To keep the database updated and further incorporate other plant species, a standardized and automatic pipeline was developed for all processes from data collection, data processing, and database updating. Altogether, Plant Regulomics provides the plant science community with useful resources and tools for data-mining of upstream regulators and the generation of hypotheses based on omics data.

METHODS

Data collection

A total of 19,043 transcriptomic data sets generated by RNA-seq or microarray analyses and 882 genome-binding profiling data sets generated by ChIP-seq or ChIP-chip for six plant species (*Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, *Glycine max*, *Solanum lycopersicum* and *Triticum aestivum*) were collected from the Gene Expression Omnibus (GEO), NASCArrays, and DDBJ databases (Supplemental Table 1B–E). The electronic annotation information, including pathways, Gene Ontologies, and InterPro domains, were retrieved from public databases (Supplemental Table 1A). The 489 non-redundant position weight matrices and relevant information used for motif analyses were downloaded from the latest (i.e., 7th) release of JASPAR (Khan *et al.*, 2018) (Supplemental Table 1G). Orthologous gene pairs between five reference species (*A. thaliana, O. sativa, Z. mays, G. max and S. lycopersicum*) and another 56 whole-genome sequenced species were downloaded from Phytozome12 (Goodstein *et al.*, 2012) (Supplemental Table 1F). 109,972 Arabidopsis protein-protein interactions (PPIs) were downloaded from TAIR (Garcia-Hernandez *et al.*, 2002) and BioGRID (Stark *et al.*, 2006, Oughtred *et al.*, 2018), 555,831 rice PPIs were downloaded from DIPOS (Sapkota *et al.*, 2011), PRIN (Gu *et al.*, 2011) and BioGRID (Stark *et al.*, 2006, Oughtred *et al.*, 2018) and 29,611 maize PPIs were downloaded from PPIM (Zhu *et al.*, 2016).

Processing transcriptomic data sets to differentially expressed gene sets

Transcriptomic data from microarray and RNA-seq analyses were processed, and differentially expressed gene sets were organized as described previously (Wang *et al.*, 2015). The bioconductor package limma (Ritchie *et al.*, 2015) was used for the microarray data, while DESeq (Anders and Huber, 2010) was used for the RNA-seq data. Differentially expressed genes for sample comparisons were defined based on the following criteria: $|\log2\ (\text{fold change})| > 1$, $P\text{-value} < 0.01$ for the microarray data and $|\log2\ (\text{fold change})| > 1$, $P\text{-value} < 0.05$ for the RNA-seq data. Next, transcriptomic comparisons resulted in a small number of genes (<10 genes) showing differential expression were excluded for subsequent analyses. 12,070 differentially expressed gene sets were finally obtained, which were further manually grouped into three categories, namely treatment, mutation/overexpression, and ecotype/cultivar.

Processing ChIP-seq, ChIP-chip and DAP-seq data to peak sets and target gene sets

The ChIP-seq data files in SRA format were downloaded from the Sequence Read Archive database maintained by the National Center for Biotechnology Information, and then converted to FASTQ format using fastq-dump (https://edwards.sdsu.edu/research/fastq-dump/). The sequence reads were cleaned by removing bases with low quality scores (< 20) and irregular GC content, cutting sequencing adapters, and filtering out short reads. The cleaned reads were mapped to the reference genome of the corresponding species using the default settings of BWA (Li and Durbin, 2009). Next, ChIP-seq peaks were detected using MACS14 (Zhang *et al.*, 2008), while the ChIP-chip peaks were detected using TileMap implemented in CisGenome (Ji *et al.*, 2011). The mapped bam files were converted to bigWig format using BEDTools (Quinlan and Hall, 2010) to configure the tracks of each study in JBrowse (Buels *et al.*, 2016). DAP-seq peaks for 540 TFs were downloaded from GEO (Guo *et al.*, 2012, O'Malley *et al.*, 2016). The target genes of each peak list were defined as those genes that have a given peak(s) in the region from 2 kb upstream to 2 kb downstream of the genic region.

Statistical test of the enrichment of the input gene list for the database gene sets

A modified Fisher's exact test defined as EASE score (Huang *et al.*, 2007) was used for testing whether the input gene list significantly overlapped the gene sets curated in the database, which is calculated as follows:

$$\text{EASE score} = \frac{\binom{n}{k-1}\binom{N-n}{K-k}}{\binom{N}{K}}$$

where N is the total number of genes as background and in different analyses the N is different, n is the number of input genes. K represents the total number of genes in one gene set and k stands for the overlapped genes between input and the pre-defined gene sets. Comparing to canonical Fisher's exact test, the EASE score is more conservative, i.e. resulted in fewer and more reliable gene sets enriched. For multiple test correction, the adjusted P values calculated by FDR, Bonferroni correction, and Benjamini and Hochberg methods were provided, which could be selected by users as criteria for definition of enriched terms.

Statistical test of the significance of the overlap between genomic loci

Analyzing the significance of the overlap between genomic loci is hampered by differences in peak lengths and the requirement that an appropriate background is selected. To address these issues, we designed the following method. First, the whole genome was divided into 1,000, 500, and 200 bp consecutive bins, selection of which for subsequent analysis is optional, depending on the pattern of input loci. For each peak list processed from the public epigenomic data or input peak list, the positions of overlapped bins were recorded. Next, given that there are some genomic loci detected as a peak region in all ChIP-seq experiments, possibly due to a mis-assembly or a repetitive feature, we generated a blacklist of genomic regions, which were defined as peaks in 124, 5, 38, and 18 negative control ChIP-seq samples in *A. thaliana*, *O. sativa*, *Z. mays*, and *G. max*, respectively. These regions were removed from further analyses. Finally, EASE score was applied to calculate the significance of the overlapped bins between the genomic loci of the input list and in the database, using bins from the whole genome sequence overlapping any peak region as the background. A multiple test correction was completed according to the FDR, Bonferroni correction, and Benjamini and Hochberg methods to lower the false-positive rate.

Motif search and enrichment analysis

We performed a motif scan using a 1,000 bp window centered at the peak center as described previously (Sun *et al.*, 2018). For each motif M, the raw motif matching score at each peak P was calculated as $\max_{S \subseteq P}\left[log\frac{P(S|M)}{P(S|B)}\right]$, in which S is a sequence fragment of the same length as the motif, and B is the background frequency of four types of nucleotides (A, C, G, and T) estimated from the genome. Regarding the motif enrichment, the enrichment of a given motif in a gene or peak list was defined as the ratio of the motif occurrence in the gene or peak list to its occurrence in random genomic regions.

Tissue expression profile

To determine the gene expression profiles across tissues, RNA-seq samples for different tissues in five species (*A. thaliana*, *O. sativa*, *Z. mays*, *S. lycopersicum* and *T. aestivum*) were selected (Supplemental Table 1E). For each gene, the number of fragments per kilobase of exon per million fragments mapped (FPKM) in each sample was calculated. Next, the mean FPKM value of samples from the same tissue was recorded.

Functional network clustering of orthologous genes

The functional annotation clustering method described by Huang et al. (Huang *et al.*, 2007) was used for constructing the network of all enriched terms. All enriched terms with a fold-change and P-value satisfying a user-defined cutoff were selected and subsequently clustered according to the number of overlapping genes between filtered terms via the MCL algorithm (Dongen, 2008). The network was visualized with Cytoscape (Shannon *et al.*, 2003). Transcriptomic data, which were used to identify genes induced by the pathogen infection of *Populus trichocarpa*, have been published (Azaiez *et al.*, 2009). For clustering, we selected the enriched terms with a P-value < 0.05 and a fold-change greater than 10 for *A. thaliana*, and a P-value < 0.05 and a fold-change greater than 5 for *O. sativa*. The lowest number of overlapped genes between selected terms was set to 5.

Database architecture

Plant Regulomics was hosted on the Apache server based on a Linux platform, with its contents maintained as a systematic database using MySQL. Python scripts were used for data processing and statistical analyses. The web interface was implemented using PHP and JavaScript.

FUNDING

This work was supported by grants from the "Strategic Priority Research Program" of the Chinese Academy of Sciences (Grant No. XDA08020102 and XDB27010302), National Natural Science Foundation of China (31770285), and sponsored by Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

DATA STATEMENT

All accession numbers of row data for transcriptomic and epigenomic data sets are listed in Supplemental Table 1.

SUPPORTING INFORMATION

Supplemental Figure 1. Expression profiles for orthologous genes of maize pollen-specific genes in different tissues of Arabidopsis **(A)** and rice **(B)**. Red stars represent 1 : many relationship of orthologs and black stars indicate the many : 1 relationship.

Supplemental Table 1. Summary of the curated data in Plant Regulomics.

(A-B) Summary of functional terms **(A)** and multi-omics data for generating gene sets **(B)** used in the Plant Regulomics platform.

(C-D) Sources of Expression profiling **(C)** and Genome binding profiling **(D)** data used for generating gene sets in Plant Regulomics.

(E) Summary of RNA-seq datasets for tissue expression profile used in Plant Regulomics.

(F-G) Summary of orthologous gene pairs and motifs used in Plant Regulomics.

Supplemental Table 2. Proteins interacting with *FT*.

Supplemental Table 3. Clusters of enriched terms in *Arabidopsis thaliana* and *Oryza sativa* for the genes induced by pathogen infection in *Populus trichocarpa*.

(A-B) Clusters of enriched terms in **(A)** *A. thaliana* and **(B)** *O. sativa*.

Supplemental Table 4. Comparison of Plant Regulomics with other web-based platforms and tools for plant gene(s) annotation.

AUTHOR CONTRIBUTIONS

YZ designed the framework. JL, XR and FZ constructed the web server. JL, YW, YZ, LY, MQ, JC and XR collected and processed public data. XR, YW and YZ wrote the manuscript with input from all authors.

REFERENCES

Abe, M., Kobayashi, Y., Yamamoto, S., Daimon, Y., Yamaguchi, A., Ikeda, Y., Ichinoki, H., Notaguchi, M., Goto, K. and Araki, T. (2005) FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. *Science*, **309**, 1052-1056.

Adrian, J., Farrona, S., Reimer, J.J., Albani, M.C., Coupland, G. and Turck, F. (2010) cis-Regulatory elements and chromatin state coordinately control temporal and spatial expression of FLOWERING LOCUS T in Arabidopsis. *Plant Cell*, **22**, 1425-1440.

Ahn, M.Y., Oh, T.R., Seo, D.H., Kim, J.H., Cho, N.H. and Kim, W.T. (2018) Arabidopsis group XIV ubiquitin-conjugating enzymes AtUBC32, AtUBC33, and AtUBC34 play negative roles in drought stress response. *Journal of plant physiology*, **230**, 73-79.

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome biology*, **11**, R106.

Arabidopsis Interactome Mapping Consortium (2011) Evidence for network evolution in an Arabidopsis interactome map. *Science*, **333**, 601-607.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25.

Azaiez, A., Boyle, B., Levee, V. and Seguin, A. (2009) Transcriptome profiling in hybrid poplar following interactions with Melampsora rust fungi. *Molecular plant-microbe interactions : MPMI*, **22**, 190-200.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S. and Soboleva, A. (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research*, **41**, D991-995.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L. and Holmes, I.H. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome biology*, **17**, 66.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J. and May, S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic acids research*, **32**, D575-D577.

Dong, S., Lau, V., Song, R., Ierullo, M., Esteban, E., Wu, Y., Sivieng, T., Nahal, H., Gaudinier, A., Pasha, A., Oughtred, R., Dolinski, K., Tyers, M., Brady, S.M., Grene, R., Usadel, B. and Provart, N.J. (2019) Proteome-wide, Structure-Based Prediction of Protein-Protein Interactions/New Molecular Interactions Viewer. *Plant physiology*, **179**, 1893-1907.

Dongen, S.V. (2008) Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis* *and Applications*, **30**, 121-141.

Du, Z., Zhou, X., Ling, Y., Zhang, Z. and Su, Z. (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic acids research*, **38**, W64-70.

Farrona, S., Thorpe, F.L., Engelhorn, J., Adrian, J., Dong, X., Sarid-Krebs, L., Goodrich, J. and Turck, F. (2011) Tissue-specific expression of FLOWERING LOCUS T in Arabidopsis is maintained independently of polycomb group protein repression. *The Plant cell*, **23**, 3204-3214.

Finkelstein, R., Reeves, W., Ariizumi, T. and Steber, C. (2008) Molecular aspects of seed dormancy. *Annual review of plant biology*, **59**, 387-415.

Fukushima, A., Kusano, M., Nakamichi, N., Kobayashi, M., Hayashi, N., Sakakibara, H., Mizuno, T. and Saito, K. (2009) Impact of clock-associated Arabidopsis pseudo-response regulators in metabolic coordination. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 7251-7256.

Garcia-Hernandez, M., Berardini, T.Z., Chen, G., Crist, D., Doyle, A., Huala, E., Knee, E., Lambrecht, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Rhee, S.Y., Scholl, R., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P. (2002) TAIR: a resource for integrated Arabidopsis data. *Functional & integrative genomics*, **2**, 239-253.

Ge, S.X. and Jung, D. (2018) ShinyGO: a graphical enrichment tool for ani-mals and plants. *bioRxiv*, 315150.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. and Rokhsar, D.S. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, **40**, D1178-1186.

Gu, H., Zhu, P., Jiao, Y., Meng, Y. and Chen, M. (2011) PRIN: a predicted rice interactome network. *BMC Bioinformatics*, **12**, 161.

Guo, Y., Mahony, S. and Gifford, D.K. (2012) High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. *PLOS Computational Biology*, **8**, e1002638.

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W. and Zimmermann, P. (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv Bioinformatics*, **2008**, 420747-420747.

Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L.A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C.D., Somerville, C. and Rhee, S.Y. (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic acids research*, **29**, 102-105.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C. and Lempicki, R.A. (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, **35**, W169-W175.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. and Yeats, C. (2009) InterPro: the integrative protein signature database. *Nucleic acids research*, **37**, D211-215.

Ji, H., Jiang, H., Ma, W. and Wong, W.H. (2011) Using CisGenome to Analyze ChIP-chip and ChIP-seq Data.

Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], **CHAPTER**, Unit2.13-Unit12.13.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., Soltis, D.E., Clifton, S.W., Schlarbaum, S.E., Schuster, S.C., Ma, H., Leebens-Mack, J. and dePamphilis, C.W. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97.

Kakei, Y. and Shimada, Y. (2014) AtCAST3.0 Update: A Web-Based Tool for Analysis of Transcriptome Data by Searching Similarities in Gene Expression Profiles. *Plant and Cell Physiology*, **56**, e7-e7.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**, 27-30.

Kardailsky, I., Shukla, V.K., Ahn, J.H., Dagenais, N., Christensen, S.K., Nguyen, J.T., Chory, J., Harrison, M.J. and Weigel, D. (1999) Activation tagging of the floral inducer FT. *Science*, **286**, 1962-1965.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., Baranasic, D., Arenillas, D.J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W.W., Parcy, F. and Mathelier, A. (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, **46**, D260-D266.

Kim, S.Y., Yu, X. and Michaels, S.D. (2008) Regulation of CONSTANS and FLOWERING LOCUS T expression in response to changing light quality. *Plant physiology*, **148**, 269-279.

Lai, L., Liberzon, A., Hennessey, J., Jiang, G., Qi, J., Mesirov, J.P. and Ge, S.X. (2012) AraPath: a knowledgebase for pathway analysis in Arabidopsis. *Bioinformatics*, **28**, 2291-2292.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S., Wensel, A. and Huala, E. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research*, **40**, D1202-1210.

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

Li, S., Pandey, S., Gookin, T.E., Zhao, Z., Wilson, L. and Assmann, S.M. (2012) Gene-Sharing Networks Reveal Organizing Principles of Transcriptomes in <em>Arabidopsis</em> and Other Multicellular Organisms. *The Plant cell*, **24**, 1362-1378.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P. and Mesirov, J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739-1740.

Liu, C., Cheng, J., Zhuang, Y., Ye, L., Li, Z., Wang, Y., Qi, M. and Zhang, Y. (2018) Polycomb repressive complex 2 attenuates ABA-induced senescence in Arabidopsis. *The Plant Journal*, **0**.

Liu, T., Carlsson, J., Takeuchi, T., Newton, L. and Farre, E.M. (2013) Direct regulation of abiotic responses by the Arabidopsis circadian clock component PRR7. *The Plant journal : for cell and molecular biology*, **76**, 101-114.

Mashima, J., Kodama, Y., Kosuge, T., Fujisawa, T., Katayama, T., Nagasaki, H., Okuda, Y., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y. and Takagi, T. (2016) DNA data bank of Japan (DDBJ) progress report. *Nucleic acids research*, **44**, D51-57.

**Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R.,

Kaul, R. and Stamatoyannopoulos, J.A. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.)*, **337**, 1190-1195.

Mizoguchi, T., Wright, L., Fujiwara, S., Cremer, F., Lee, K., Onouchi, H., Mouradov, A., Fowler, S., Kamada, H., Putterill, J. and Coupland, G. (2005) Distinct roles of GIGANTEA in promoting flowering and regulating circadian rhythms in Arabidopsis. *The Plant cell*, **17**, 2255-2270.

O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A. and Ecker, J.R. (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, **165**, 1280-1292.

O'Maoileidigh, D.S., Graciet, E. and Wellmer, F. (2014) Gene networks controlling Arabidopsis thaliana flower development. *The New phytologist*, **201**, 16-30.

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K. and Tyers, M. (2018) The BioGRID interaction database: 2019 update. *Nucleic acids research*, **47**, D529-D541.

Patel, R.V., Nahal, H.K., Breit, R. and Provart, N.J. (2012) BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. *The Plant Journal*, **71**, 1038-1050.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

Riboni, M., Robustelli Test, A., Galbiati, M., Tonelli, C. and Conti, L. (2016) ABA-dependent control of GIGANTEA signalling enables drought escape via up-regulation of FLOWERING LOCUS T in Arabidopsis thaliana. *Journal of experimental botany*, **67**, 6309-6322.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, **43**, e47.

Sapkota, A., Liu, X., Zhao, X.M., Cao, Y., Liu, J., Liu, Z.P. and Chen, L. (2011) DIPOS: database of interacting proteins in Oryza sativa. *Molecular bioSystems*, **7**, 2615-2621.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, **13**, 2498-2504.

Shen, Q. and Ho, T.H. (1995) Functional dissection of an abscisic acid (ABA)-inducible gene reveals two independent ABA-responsive complexes each containing a G-box and a novel cis-acting element. *Plant Cell*, **7**, 295-307.

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**.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic acids research*, **34**, D535-539.

Sun, H., Wang, J., Gong, Z., Yao, J., Wang, Y., Xu, J., Yuan, G.C., Zhang, Y. and Shao, Z. (2018) Quantitative integration of epigenomic variation and transcription factor binding using MAmotif toolkit identifies an important role of IRF2 as transcription activator at gene promoters. *Cell discovery*, **4**, 38.

Tripathi, S., Pohl, M.O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D.A., Moulton, H.M., DeJesus, P., Che, J., Mulder, L.C., Yanguez, E., Andenmatten, D., Pache, L., Manicassamy, B., Albrecht, R.A., Gonzalez, M.G., Nguyen, Q., Brass, A., Elledge, S., White, M., Shapira, S., Hacohen, N., Karlas, A., Meyer, T.F., Shales, M., Gatorano, A., Johnson, J.R., Jang, G., Johnson,

T., Verschueren, E., Sanders, D., Krogan, N., Shaw, M., Konig, R., Stertz, S., Garcia-Sastre, A. and Chanda, S.K. (2015) Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. *Cell host & microbe*, **18**, 723-735.

Turck, F., Fornara, F. and Coupland, G. (2008) Regulation and identity of florigen: FLOWERING LOCUS T moves center stage. *Annual review of plant biology*, **59**, 573-594.

Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M.L., Guillaume, E., Buisine, N., Gagnot, S., Martienssen, R.A., Coupland, G. and Colot, V. (2007) Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. *PLoS genetics*, **3**, e86.

Waese, J., Fan, J., Pasha, A., Yu, H., Fucile, G., Shi, R., Cumming, M., Kelley, L.A., Sternberg, M.J., Krishnakumar, V., Ferlanti, E., Miller, J., Town, C., Stuerzlinger, W. and Provart, N.J. (2017) ePlant: Visualizing and Exploring Multiple Levels of Data for Hypothesis Generation in Plant Biology. *The Plant cell*, **29**, 1806-1821.

Wang, J., Qi, M., Liu, J. and Zhang, Y. (2015) CARMO: a comprehensive annotation platform for functional exploration of rice multi-omics data. *The Plant journal : for cell and molecular biology*, **83**, 359-374.

Wellmer, F., Alves-Ferreira, M., Dubois, A., Riechmann, J.L. and Meyerowitz, E.M. (2006) Genome-wide analysis of gene expression during early Arabidopsis flower development. *Plos Genet*, **2**, 1012-1024.

Wendel, J.F., Jackson, S.A., Meyers, B.C. and Wing, R.A. (2016) Evolution of plant genome architecture. *Genome biology*, **17**, 37.

Wilson, Z.A. and Zhang, D.B. (2009) From Arabidopsis to rice: pathways in pollen development. *J Exp Bot*, **60**, 1479-1492.

Yi, X., Du, Z. and Su, Z. (2013) PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic acids research*, **41**, W98-103.

Zhang, X., Germann, S., Blus, B.J., Khorasanizadeh, S., Gaudin, V. and Jacobsen, S.E. (2007) The Arabidopsis LHP1 protein colocalizes with histone H3 Lys27 trimethylation. *Nature structural & molecular biology*, **14**, 869-871.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. and Liu, X.S. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**, R137.

Zhou, Y., Wang, Y., Krause, K., Yang, T., Dongus, J.A., Zhang, Y. and Turck, F. (2018) Telobox motifs recruit CLF/SWN-PRC2 for H3K27me3 deposition via TRB factors in Arabidopsis. *Nat Genet*, **50**, 638-644.

Zhu, G., Wu, A., Xu, X.J., Xiao, P.P., Lu, L., Liu, J., Cao, Y., Chen, L., Wu, J. and Zhao, X.M. (2016) PPIM: A Protein-Protein Interaction Database for Maize. *Plant physiology*, **170**, 618-626.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Gruissem, W. (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant physiology*, **136**, 2621-2632.

FIGURE LEGENDS

Figure 1. Overview and workflow of Plant Regulomics.

Plant Regulomics integrated 19,043 transcriptomic and 1,694 epigenomic data sets, which were processed and organized into 12,070 differentially expressed gene sets and 1,691 gene sets or genomic loci targeted by transcription factors and epigenetic factors. For any input gene list or genomic loci derived from high-throughput technology, Plant Regulomics returns the transcription factors, epigenetic factors, cis-elements, interactors, perturbation of genes, and experimental/environmental conditions regulating the input genes or genomic loci.

Figure 2. Searching for upstream factors for *FT* gene.

(A) Input page for a single gene. **(B)** Upstream factors of *FT* determined by comparisons of differentially expressed genes based on transcriptomic data. **(C)** Motifs present in the *FT* promoter and gene body regions. **(D)** Transcription and epigenetic factors surrounding *FT*, as determined by ChIP-seq studies. **(E)** Genomic tracks illustrating the binding of factors selected in **(D)** surrounding the *FT* locus. **(G)** The protein interactor(s) of *FT*.

Figure 3. Searching for upstream factors for a set of ABA-induced genes.

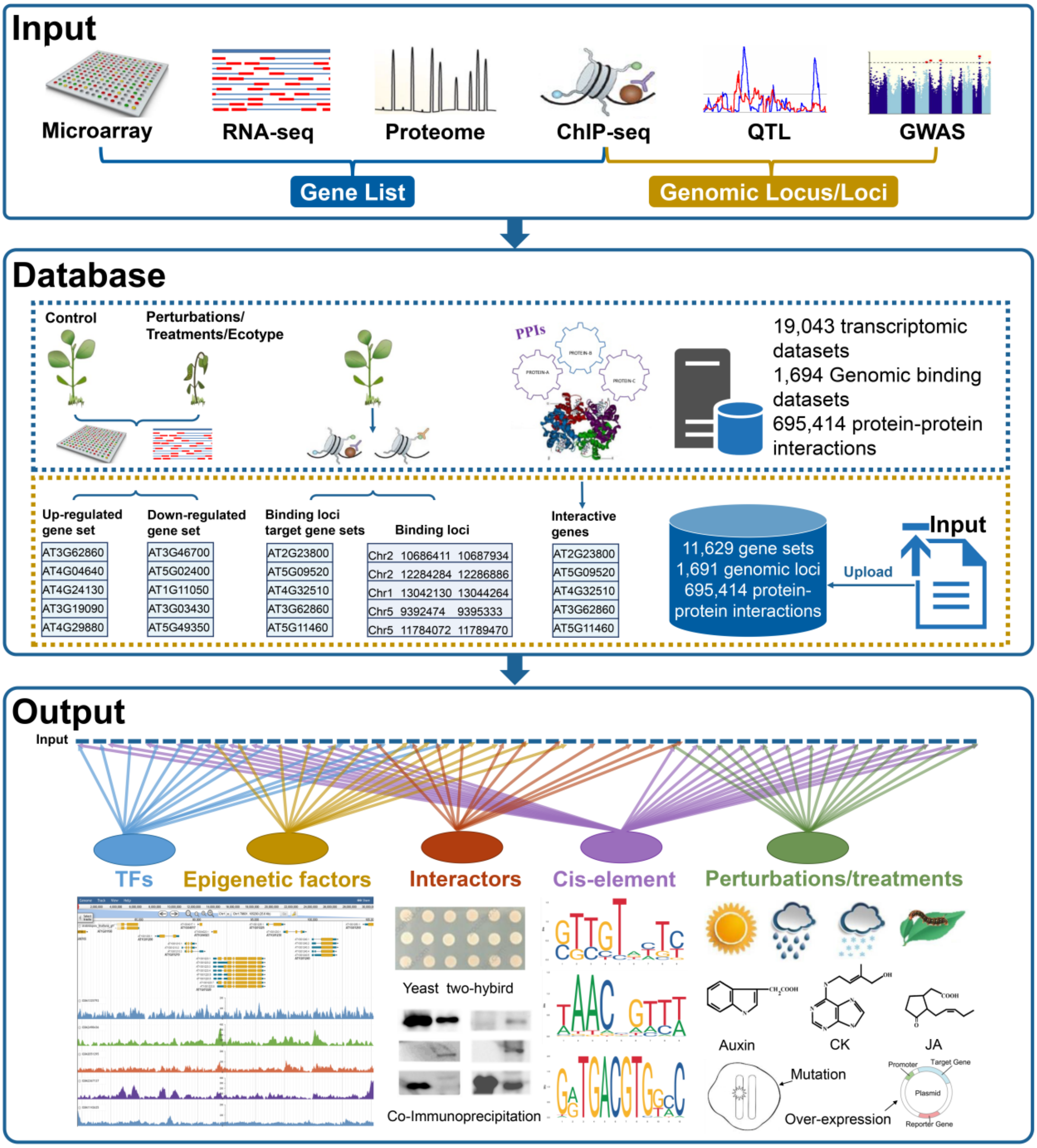
(A) Result summary page for searching the upstream factors of ABA-induced genes. **(B)** The table in the top panel lists the top enriched transcriptomic comparisons for the input gene list. The heatmap under the table indicates whether the expression of a given gene displayed significant changes in the transcriptomic comparison selected from the upper table. **(C)** Tissue expression profiles for the input gene list. **(D)** Transcription factors and epigenetic factors whose bindings are over-represented surrounding input genes. **(E)** Genomic tracks illustrating the binding of factors selected in **(D)**. **(F)** Cis-elements over-represented surrounding input genes. **(G)** The protein interactors associated with input genes. The top five interactors with higher numbers of interactions with input gene list are listed, and all interactors could be downloaded.

Figure 4. Searching for upstream factors for PRR7-binding genomic loci determined by ChIP-seq.

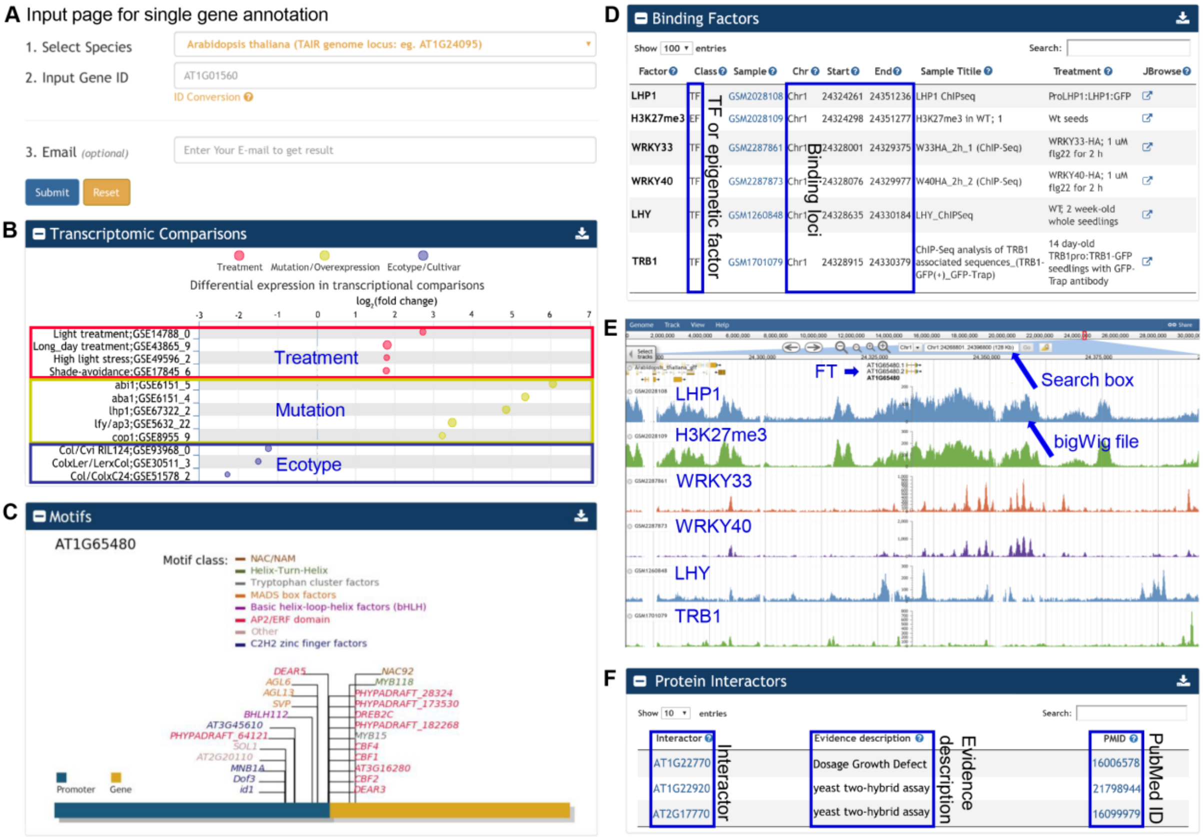
(A) Result summary page for genomic loci. **(B)** Details regarding the distribution and length of input genomic loci. **(C)** The table in the top panel lists the transcription factors and epigenetic factors whose bindings are over-represented surrounding input loci. The heatmap under the table indicates whether the factors selected from the table are present surrounding each locus. Genomic tracks for selected factors were visualized *via* a genome browser **(D). (E)** Cis-elements over-represented surrounding input loci. **(F–G)** Gene list functional annotation result for PRR7 peak targets, including the top enriched GO terms **(F)**, transcriptomic comparisons between the treatment and control and transcriptomic changes in mutants **(G)**.

Figure 5. Functional annotation for orthologous genes.

(A) For the input pathogen-induced poplar genes, the orthologous genes in any of the five reference species were returned. **(B)** The top enriched clusters of transcriptomic comparisons in *Arabidopsis thaliana* and *Oryza sativa* for the input poplar genes. **(C)** For the input maize pollen-specific genes, the tissue expression profiles of orthologous genes in *A. thaliana* and *O. sativa* are presented in the heatmap.



Up-regulated gene set	Down-regulated gene set	Binding loci target gene sets	Binding loci	Interactive genes
AT3G62860	AT3G46700	AT2G23800	Chr2 10686411 10687934	AT2G23800
AT4G04640	AT5G02400	AT5G09520	Chr2 12284284 12286886	AT5G09520
AT4G24130	AT1G11050	AT4G32510	Chr1 13042130 13044264	AT4G32510
AT3G19090	AT3G03430	AT3G62860	Chr5 9392474 9395333	AT3G62860
AT4G29880	AT5G49350	AT5G11460	Chr5 11784072 11789470	AT5G11460



A Input page for single gene annotation

1. Select Species: Arabidopsis thaliana (TAIR genome locus: eg. AT1G24095)

2. Input Gene ID: AT1G01560

ID Conversion

3. Email (optional): Enter Your E-mail to get result

Submit Reset

Binding Factors

Show 100 entries Search:

Factor	Class	Sample	Chr	Start	End	Sample Titile	Treatment	JBrowse
LHP1	TF	GSM2028108	Chr1	24324261	24351236	LHP1 ChIPseq	ProLHP1:LHP1:GFP	↗
H3K27me3	EF	GSM2028109	Chr1	24324298	24351277	H3K27me3 in WT; 1	Wt seeds	↗
WRKY33	TF	GSM2287861	Chr1	24328001	24329375	W33HA_2h_1 (ChIP-Seq)	WRKY33-HA; 1 uM flg22 for 2 h	↗
WRKY40	TF	GSM2287873	Chr1	24328076	24329977	W40HA_2h_2 (ChIP-Seq)	WRKY40-HA; 1 uM flg22 for 2 h	↗
LHY	TF	GSM1260848	Chr1	24328635	24330184	LHY_ChIPSeq	WT; 2 week-old whole seedlings	↗
TRB1	TF	GSM1701079	Chr1	24328915	24330379	ChIP-Seq analysis of TRB1 associated sequences_(TRB1-GFP(+)_GFP-Trap)	14 day-old TRB1pro:TRB1-GFP seedlings with GFP-Trap antibody	↗

TF or epigenetic factor

Binding loci

Protein Interactors

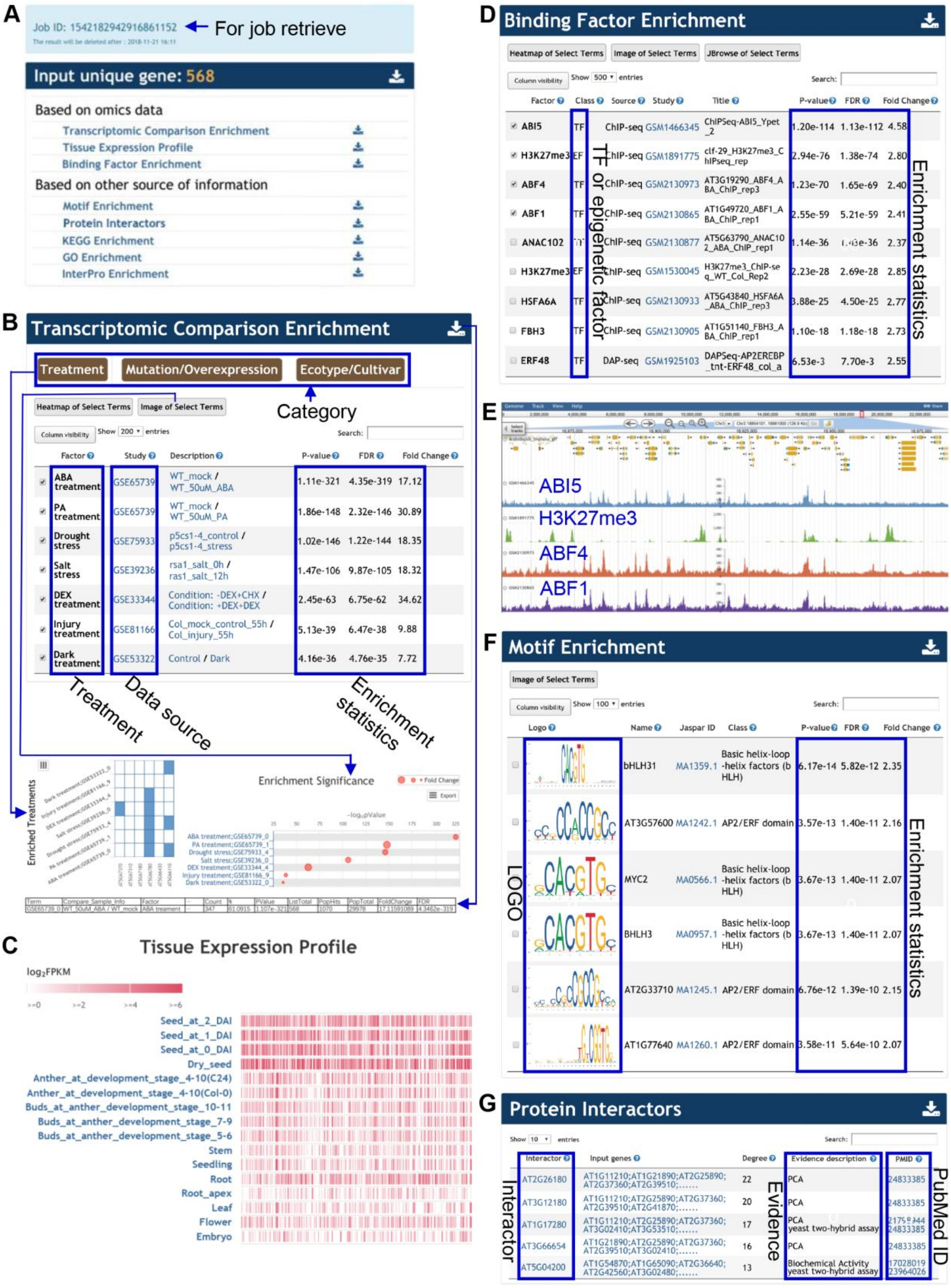
Show 10 entries Search:

Interactor	Evidence description	PMID
AT1G22770	Dosage Growth Defect	16006578
AT1G22920	yeast two-hybrid assay	21798944
AT2G17770	yeast two-hybrid assay	16099979

Interactor

Evidence description

PubMed ID



Job ID: 1542182942916861152

The result will be deleted after : 2018-11-21 16:11

Input unique gene: 568

Based on omics data

- Transcriptomic Comparison Enrichment
- Tissue Expression Profile
- Binding Factor Enrichment

Based on other source of information

- Motif Enrichment
- Protein Interactors
- KEGG Enrichment
- GO Enrichment
- InterPro Enrichment

Transcriptomic Comparison Enrichment

Treatment | Mutation/Overexpression | Ecotype/Cultivar

Heatmap of Select Terms | Image of Select Terms

Column visibility Show 200 entries Search:

	Factor	Study	Description	P-value	FDR	Fold Change
☑	ABA treatment	GSE65739	WT_mock / WT_50uM_ABA	1.11e-321	4.35e-319	17.12
☑	PA treatment	GSE65739	WT_mock / WT_50uM_PA	1.86e-148	2.32e-146	30.89
☑	Drought stress	GSE75933	p5cs1-4_control / p5cs1-4_stress	1.02e-146	1.22e-144	18.35
☑	Salt stress	GSE39236	rsa1_salt_0h / ras1_salt_12h	1.47e-106	9.87e-105	18.32
☑	DEX treatment	GSE33344	Condition: -DEX+CHX / Condition: +DEX+DEX	2.45e-63	6.75e-62	34.62
☑	Injury treatment	GSE81166	Col_mock_control_55h / Col_injury_55h	5.13e-39	6.47e-38	9.88
☑	Dark treatment	GSE53322	Control / Dark	4.16e-36	4.76e-35	7.72

Term	Compare_Sample_Info	Factor	...	Count	%	PValue	ListTotal	PopHits	PopTotal	FoldChange	FDR
GSE65739_0	WT_50uM_ABA / WT_mock	ABA treament	...	347	61.0915	1.107e-321	568	1070	29978	17.11591089	4.3462e-319

Binding Factor Enrichment

Heatmap of Select Terms | Image of Select Terms | JBrowse of Select Terms

Column visibility Show 500 entries Search:

	Factor	Class	Source	Study	Title	P-value	FDR	Fold Change
☑	ABI5	TF	ChIP-seq	GSM1466345	ChIPSeq-ABI5_Ypet_2	1.20e-114	1.13e-112	4.58
☑	H3K27me3	EF	ChIP-seq	GSM1891775	clf-29_H3K27me3_ChIPseq_rep	2.94e-76	1.38e-74	2.80
☑	ABF4	TF	ChIP-seq	GSM2130973	AT3G19290_ABF4_ABA_ChIP_rep3	1.23e-70	1.65e-69	2.40
☑	ABF1	TF	ChIP-seq	GSM2130865	AT1G49720_ABF1_ABA_ChIP_rep1	2.55e-59	5.21e-59	2.41
□	ANAC102	TF	ChIP-seq	GSM2130877	AT5G63790_ANAC102_ABA_ChIP_rep1	1.14e-36	1.43e-36	2.37
□	H3K27me3	EF	ChIP-seq	GSM1530045	H3K27me3_ChIP-seq_WT_Col_Rep2	2.23e-28	2.69e-28	2.85
□	HSFA6A	TF	ChIP-seq	GSM2130933	AT5G43840_HSFA6A_ABA_ChIP_rep3	3.88e-25	4.50e-25	2.77
□	FBH3	TF	ChIP-seq	GSM2130905	AT1G51140_FBH3_ABA_ChIP_rep1	1.10e-18	1.18e-18	2.73
□	ERF48	TF	DAP-seq	GSM1925103	DAPSeq-AP2EREBP_tnt-ERF48_col_a	6.53e-3	7.70e-3	2.55

Motif Enrichment

Image of Select Terms

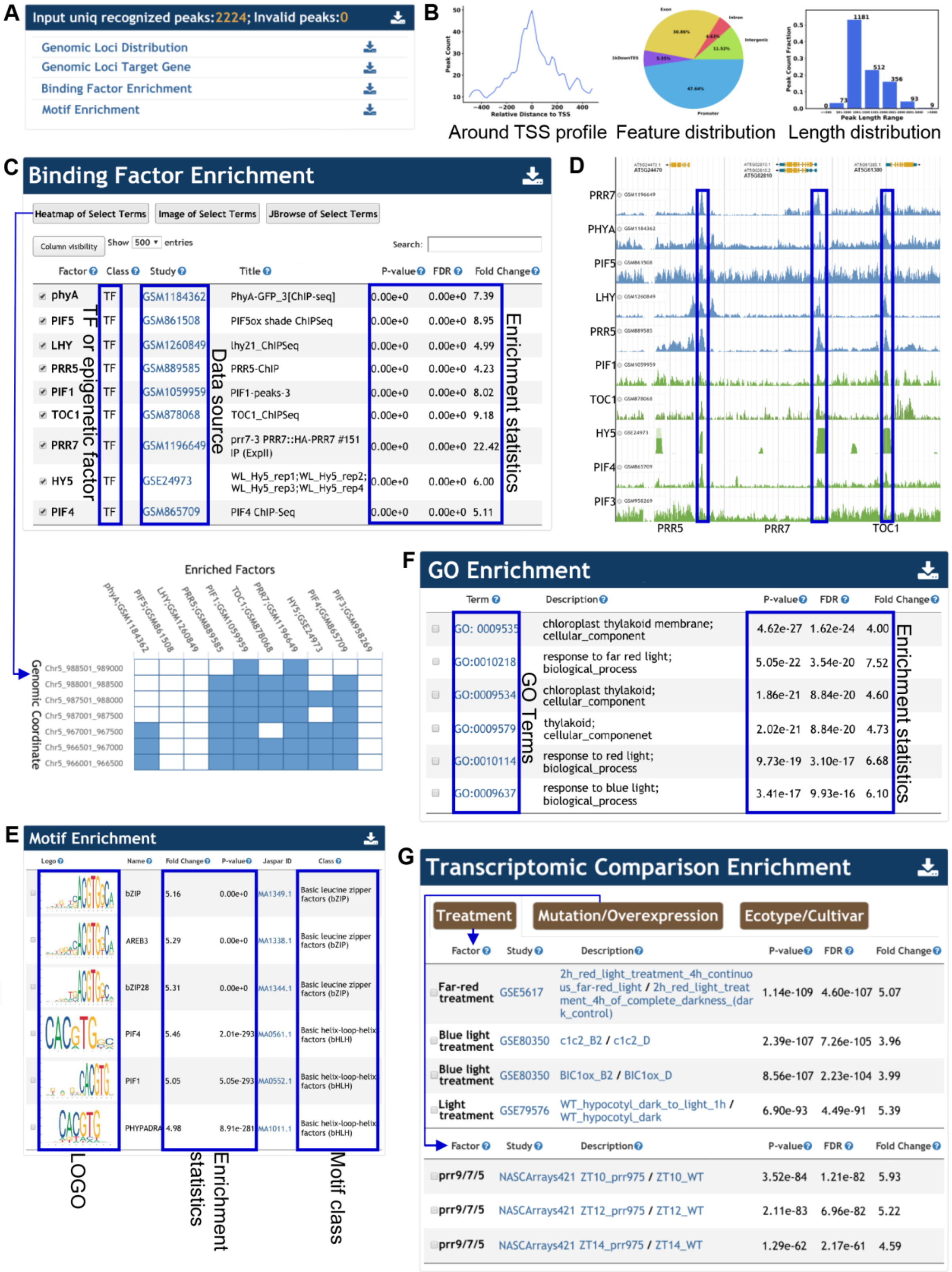
Column visibility Show 100 entries Search:

	Logo	Name	Jaspar ID	Class	P-value	FDR	Fold Change
□		bHLH31	MA1359.1	Basic helix-loop-helix factors (bHLH)	6.17e-14	5.82e-12	2.35
□		AT3G57600	MA1242.1	AP2/ERF domain	3.57e-13	1.40e-11	2.16
□		MYC2	MA0566.1	Basic helix-loop-helix factors (bHLH)	3.67e-13	1.40e-11	2.07
□		BHLH3	MA0957.1	Basic helix-loop-helix factors (bHLH)	3.67e-13	1.40e-11	2.07
□		AT2G33710	MA1245.1	AP2/ERF domain	6.76e-12	1.39e-10	2.15
□		AT1G77640	MA1260.1	AP2/ERF domain	3.58e-11	5.64e-10	2.07

Protein Interactors

Show 10 entries Search:

Interactor	Input genes	Degree	Evidence description	PMID
AT2G26180	AT1G11210;AT1G21890;AT2G25890;AT2G37360;AT2G39510;......	22	PCA	24833385
AT3G12180	AT1G11210;AT2G25890;AT2G37360;AT2G39510;AT2G41870;......	20	PCA	24833385
AT1G17280	AT1G11210;AT2G25890;AT2G37360;AT3G02410;AT3G53510;......	17	PCA yeast two-hybrid assay	21798944 24833385
AT3G66654	AT1G21890;AT2G25890;AT2G37360;AT2G39510;AT3G02410;......	16	PCA	24833385
AT5G04200	AT1G54870;AT1G65090;AT2G36640;AT2G42560;AT3G02480;......	13	Biochemical Activity yeast two-hybrid assay	17028019 23964026



Input uniq recognized peaks:2224; Invalid peaks:0

- Genomic Loci Distribution
- Genomic Loci Target Gene
- Binding Factor Enrichment
- Motif Enrichment

Binding Factor Enrichment

Heatmap of Select Terms | Image of Select Terms | JBrowse of Select Terms

Column visibility Show 500 entries Search:

	Factor	Class	Study	Title	P-value	FDR	Fold Change
☑	phyA	TF	GSM1184362	PhyA-GFP_3[ChIP-seq]	0.00e+0	0.00e+0	7.39
☑	PIF5	TF	GSM861508	PIF5ox shade ChIPSeq	0.00e+0	0.00e+0	8.95
☑	LHY	TF	GSM1260849	lhy21_ChIPSeq	0.00e+0	0.00e+0	4.99
☑	PRR5	TF	GSM889585	PRR5-ChIP	0.00e+0	0.00e+0	4.23
☑	PIF1	TF	GSM1059959	PIF1-peaks-3	0.00e+0	0.00e+0	8.02
☑	TOC1	TF	GSM878068	TOC1_ChIPSeq	0.00e+0	0.00e+0	9.18
☑	PRR7	TF	GSM1196649	prr7-3 PRR7::HA-PRR7 #151 IP (ExpII)	0.00e+0	0.00e+0	22.42
☑	HY5	TF	GSE24973	WL_Hy5_rep1;WL_Hy5_rep2; WL_Hy5_rep3;WL_Hy5_rep4	0.00e+0	0.00e+0	6.00
☑	PIF4	TF	GSM865709	PIF4 ChIP-Seq	0.00e+0	0.00e+0	5.11

GO Enrichment

	Term	Description	P-value	FDR	Fold Change
□	GO: 0009535	chloroplast thylakoid membrane; cellular_component	4.62e-27	1.62e-24	4.00
□	GO:0010218	response to far red light; biological_process	5.05e-22	3.54e-20	7.52
□	GO:0009534	chloroplast thylakoid; cellular_component	1.86e-21	8.84e-20	4.60
□	GO:0009579	thylakoid; cellular_componenet	2.02e-21	8.84e-20	4.73
□	GO:0010114	response to red light; biological_process	9.73e-19	3.10e-17	6.68
□	GO:0009637	response to blue light; biological_process	3.41e-17	9.93e-16	6.10

Motif Enrichment

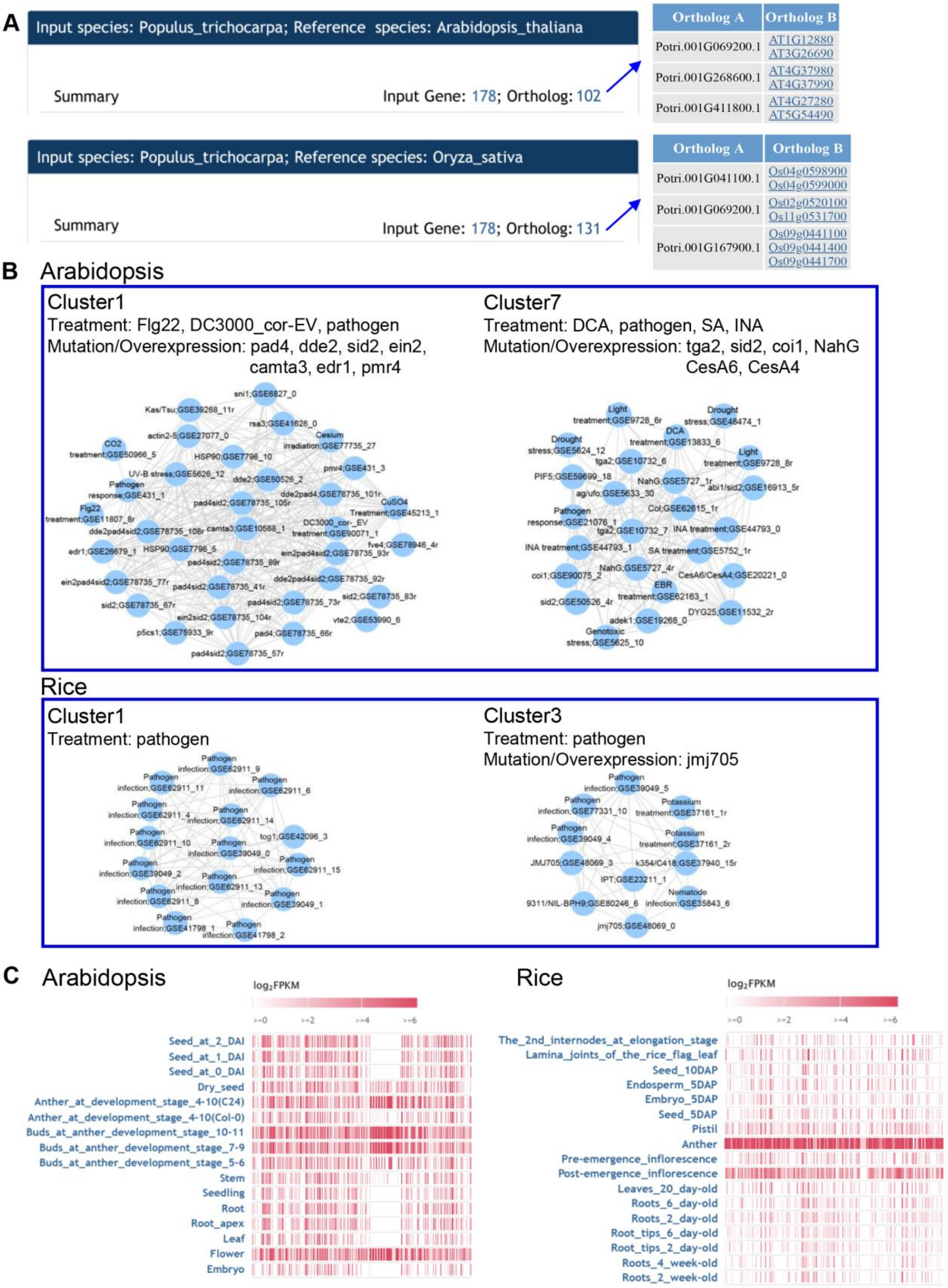
	Logo	Name	Fold Change	P-value	Jaspar ID	Class
□		bZIP	5.16	0.00e+0	MA1349.1	Basic leucine zipper factors (bZIP)
□		AREB3	5.29	0.00e+0	MA1338.1	Basic leucine zipper factors (bZIP)
□		bZIP28	5.31	0.00e+0	MA1344.1	Basic leucine zipper factors (bZIP)
□		PIF4	5.46	2.01e-293	MA0561.1	Basic helix-loop-helix factors (bHLH)
□		PIF1	5.05	5.05e-293	MA0552.1	Basic helix-loop-helix factors (bHLH)
□		PHYPADRA	4.98	8.91e-281	MA1011.1	Basic helix-loop-helix factors (bHLH)

Transcriptomic Comparison Enrichment

Treatment | Mutation/Overexpression | Ecotype/Cultivar

	Factor	Study	Description	P-value	FDR	Fold Change
□	Far-red treatment	GSE5617	2h_red_light_treatment_4h_continuous_far-red_light / 2h_red_light_treatment_4h_of_complete_darkness_(dark_control)	1.14e-109	4.60e-107	5.07
□	Blue light treatment	GSE80350	c1c2_B2 / c1c2_D	2.39e-107	7.26e-105	3.96
□	Blue light treatment	GSE80350	BIC1ox_B2 / BIC1ox_D	8.56e-107	2.23e-104	3.99
□	Light treatment	GSE79576	WT_hypocotyl_dark_to_light_1h / WT_hypocotyl_dark	6.90e-93	4.49e-91	5.39

	Factor	Study	Description	P-value	FDR	Fold Change
□	prr9/7/5	NASCArrays421	ZT10_prr975 / ZT10_WT	3.52e-84	1.21e-82	5.93
□	prr9/7/5	NASCArrays421	ZT12_prr975 / ZT12_WT	2.11e-83	6.96e-82	5.22
□	prr9/7/5	NASCArrays421	ZT14_prr975 / ZT14_WT	1.29e-62	2.17e-61	4.59



A

Input species: Populus_trichocarpa; Reference species: Arabidopsis_thaliana

Summary Input Gene: 178; Ortholog: 102

Ortholog A	Ortholog B
Potri.001G069200.1	AT1G12880 AT3G26690
Potri.001G268600.1	AT4G37980 AT4G37990
Potri.001G411800.1	AT4G27280 AT5G54490

Input species: Populus_trichocarpa; Reference species: Oryza_sativa

Summary Input Gene: 178; Ortholog: 131

Ortholog A	Ortholog B
Potri.001G041100.1	Os04g0598900 Os04g0599000
Potri.001G069200.1	Os02g0520100 Os11g0531700
Potri.001G167900.1	Os09g0441100 Os09g0441400 Os09g0441700

B
Arabidopsis
Cluster1
Treatment: Flg22, DC3000_cor-EV, pathogen
Mutation/Overexpression: pad4, dde2, sid2, ein2, camta3, edr1, pmr4
sni1;GSE6827_0
Kas/Tsu;GSE39268_11r
rsa3;GSE41626_0
actin2-5;GSE27077_0
Cesium irradiation;GSE77735_27
CO2 treatment;GSE50966_5
HSP90;GSE7796_10
pmr4;GSE431_3
UV-B stress;GSE5626_12
dde2;GSE50526_2
Pathogen response;GSE431_1
dde2pad4;GSE78735_101r
pad4sid2;GSE78735_105r
CuSO4 Treatment;GSE45213_1
Flg22 treatment;GSE11807_8r
DC3000_cor-_EV treatment;GSE90071_1
dde2pad4sid2;GSE78735_108r
camta3;GSE10568_1
fve4;GSE78946_4r
edr1;GSE26679_1
HSP90;GSE7796_5
ein2pad4sid2;GSE78735_93r
pad4sid2;GSE78735_89r
ein2pad4sid2;GSE78735_77r
pad4sid2;GSE78735_41r
dde2pad4sid2;GSE78735_92r
sid2;GSE78735_67r
pad4sid2;GSE78735_73r
sid2;GSE78735_83r
ein2sid2;GSE78735_104r
vte2;GSE53990_6
p5cs1;GSE75933_9r
pad4;GSE78735_66r
pad4sid2;GSE78735_57r
Cluster7
Treatment: DCA, pathogen, SA, INA
Mutation/Overexpression: tga2, sid2, coi1, NahG CesA6, CesA4
Light treatment;GSE9728_6r
Drought stress;GSE48474_1
DCA treatment;GSE13833_6
Drought stress;GSE5624_12
tga2;GSE10732_6
Light treatment;GSE9728_8r
PIF5;GSE59699_18
NahG;GSE5727_1r
abi1/sid2;GSE16913_5r
ag/ufo;GSE5633_30
Col;GSE62615_1r
Pathogen response;GSE21076_1
tga2;GSE10732_7
INA treatment;GSE44793_0
INA treatment;GSE44793_1
SA treatment;GSE5752_1r
NahG;GSE5727_4r
coi1;GSE90075_2
CesA6/CesA4;GSE20221_0
EBR treatment;GSE62163_1
sid2;GSE50526_4r
DYG25;GSE11532_2r
adek1;GSE19268_0
Genotoxic stress;GSE5625_10
Rice
Cluster1
Treatment: pathogen
Pathogen infection;GSE62911_9
Pathogen infection;GSE62911_11
Pathogen infection;GSE62911_6
Pathogen infection;GSE62911_4
Pathogen infection;GSE62911_14
Pathogen infection;GSE62911_10
tog1;GSE42096_3
Pathogen infection;GSE39049_0
Pathogen infection;GSE39049_2
Pathogen infection;GSE62911_15
Pathogen infection;GSE62911_13
Pathogen infection;GSE62911_8
Pathogen infection;GSE39049_1
Pathogen infection;GSE41798_1
Pathogen infection;GSE41798_2
Cluster3
Treatment: pathogen
Mutation/Overexpression: jmj705
Pathogen infection;GSE39049_5
Pathogen infection;GSE77331_10
Potassium treatment;GSE37161_1r
Pathogen infection;GSE39049_4
Potassium treatment;GSE37161_2r
JMJ705;GSE48069_3
k354/C418;GSE37940_15r
IPT;GSE23211_1
9311/NIL-BPH9;GSE80246_6
Nematode infection;GSE35843_6
jmj705;GSE48069_0
C
Arabidopsis
log2FPKM
>=0 >=2 >=4 >=6
Seed_at_2_DAI
Seed_at_1_DAI
Seed_at_0_DAI
Dry_seed
Anther_at_development_stage_4-10(C24)
Anther_at_development_stage_4-10(Col-0)
Buds_at_anther_development_stage_10-11
Buds_at_anther_development_stage_7-9
Buds_at_anther_development_stage_5-6
Stem
Seedling
Root
Root_apex
Leaf
Flower
Embryo
Rice
log2FPKM
>=0 >=2 >=4 >=6
The_2nd_internodes_at_elongation_stage
Lamina_joints_of_the_rice_flag_leaf
Seed_10DAP
Endosperm_5DAP
Embryo_5DAP
Seed_5DAP
Pistil
Anther
Pre-emergence_inflorescence
Post-emergence_inflorescence
Leaves_20_day-old
Roots_6_day-old
Roots_2_day-old
Root_tips_6_day-old
Root_tips_2_day-old
Roots_4_week-old
Roots_2_week-old