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

Xiaojuan Ran¹,²*, Fei Zhao¹,²*, Yuejun Wang¹,²*, Jian Liu¹,², Yili Zhuang¹,², Luhuan Ye¹,², Meifang Qi¹,², Jingfei Cheng¹,², Yijing Zhang¹,²†

¹National 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

²University 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

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tpj.14526
This article is protected by copyright. All rights reserved.
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

This article is protected by copyright. All rights reserved.
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 those potentially bound by MADS box factors, C2H2 zinc finger factors, and the AP2/ERF domain. 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, PIFs, 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 allow 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 (fold change)| > 1, P-value < 0.01 for the microarray data and |log2 (fold change)| > 1, P-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.

This article is protected by copyright. All rights reserved.
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:

\[
\text{EASE score} = \frac{n!}{k!(N-k)!} \frac{(N-n)!}{K!(N-K)!}
\]

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 overlap ped 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.

This article is protected by copyright. All rights reserved.
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


This article is protected by copyright. All rights reserved.


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.
This article is protected by copyright. All rights reserved.