the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) **83,** 359–374 doi: 10.1111/tpj.12894

TECHNICAL ADVANCE

CARMO: a comprehensive annotation platform for functional exploration of rice multi-omics data

Jiawei Wang†, Meifang Qi†, Jian Liu† and Yijing Zhang*

National Laboratory of Plant Molecular Genetics, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 300 Fenglin Road, Shanghai 200032, China

Received 13 February 2015; revised 14 May 2015; accepted 19 May 2015; published online 4 June 2015.
*For correspondence (e-mail zhangyijing@sibs.ac.cn).
†These authors contributed equally to this work.

SUMMARY

High-throughput technology is gradually becoming a powerful tool for routine research in rice. Interpretation of biological significance from the huge amount of data is a critical but non-trivial task, especially for rice, for which gene annotations rely heavily on sequence similarity rather than direct experimental evidence. Here we describe the annotation platform for comprehensive annotation of rice multi-omics data (CARMO), which provides multiple web-based analysis tools for in-depth data mining and visualization. The central idea involves systematic integration of 1819 samples from omics studies and diverse sources of functional evidence (15 401 terms), which are further organized into gene sets and higher-level gene modules. In this way, the high-throughput data may easily be compared across studies and platforms, and integration of multiple types of evidence allows biological interpretation from the level of gene functional modules with high confidence. In addition, the functions and pathways for thousands of genes lacking description or validation may be deduced based on concerted expression of genes within the constructed co-expression networks or gene modules. Overall, CARMO provides comprehensive annotations for transcriptomic datasets, epi-genomic modification sites, single nucleotide polymorphisms identified from genome re-sequencing, and the large gene lists derived from these omics studies. Well-organized results, as well as multiple tools for interactive visualization, are available through a user-friendly web interface. Finally, we illustrate how CARMO enables biological insights using four examples, demonstrating that CARMO is a highly useful resource for intensive data mining and hypothesis generation based on rice multi-omics data. CARMO is freely available online (http://bioinfo.sibs.ac.cn/carmo).

Keywords: CARMO, *Oryza sativa*, rice omics data, functional integration, gene annotation, gene module, technical advance.

INTRODUCTION

Oryza sativa (rice) is not only a major food crop but also a valuable model plant for research. The rapid development of transcriptomic and (epi)genomic technologies has greatly promoted our understanding of the function of genes on a genome-wide scale, but this poses a major challenge with respect to data mining, which depends highly on reliable functional annotation of genes. However, 99.5% of gene ontology (GO) terms for rice in the major functional annotation resources (Ware *et al.*, 2002; Kawahara *et al.*, 2013; Sakai *et al.*, 2013) are deduced on the basis of sequence homology with other species, and thus are not sufficiently reliable. Given that current tools for functional interpretation of genes in rice are largely dependent on terms of this sort (Du *et al.*, 2010; Kawahara *et al.*, 2013; Sakai *et al.*, 2013; Yi *et al.*, 2013), incorporating direct experimental evidence from rice to develop more reliable tools for functional annotation of genes is an urgent need.

The huge amount of public rice omics data provides a good resource for functional annotation of rice genes. For example, gene sets derived from differential expression analysis represent genes with a common response to particular perturbation or treatment, while genes affected by a set of single nucleotide polymorphisms (SNPs) associated with a certein trait identified from genome-wide associa-

tion studies (GWAS) are likely to be co-regulators affecting the trait under study. However, the fact that these data are generated by different studies and different platforms, which may not be readily comparable, poses significant challenges for data integration. Furthermore, combining information extracted from omics data with other types of functional annotations is also a non-trivial task.

Various methods have been proposed for integration of different sources of information (Segal *et al.*, 2004; Subramanian *et al.*, 2005; Huang *et al.*, 2009). These focus on compilation of gene sets from diverse sources of annotation, including ontologies, pathways, domains, expression information, etc. Subramanian *et al.* (2005) described gene set enrichment analysis as a powerful knowledge-based approach for interpreting genome-wide expression profiles. However, different gene sets are tested individually in the above method, and interplay between different gene sets is not considered. Thus, users are presented with a collection of enriched functional terms without a description of the relationship between them. Huang *et al.* (2009) attempted to address this issue via clustering of over-represented terms based on the number of genes they share, such that closely related terms are organized into groups. In addition, Segal *et al.* (2004) proposed the idea of module networks to organize gene sets into higher-level modules based on their expression behavior in cancer tissues, and constructed 456 gene modules that work in concert to perform related functions in cancer; these modules have been widely used in cancer diagnostic, prognostic and therapeutic studies (Segal *et al.*, 2005; Wong *et al.*, 2008a,b). Data-driven construction of gene networks based on expression information has been described for rice (Lee *et al.*, 2011), but combination of functional information and high-throughput data for interpretation of gene function is still insufficient in rice studies, thus restricting hypothesis-driven research.

Here, we describe comprehensive annotation of rice multi-omics data (CARMO), an integrated annotation platform for functional exploration of rice multi-omics data. The current release of the rice reference genome (IRGSP 1.0; http://rapdb.dna.affrc.go.jp/download/irgsp1.html) is used, and gene models are based on Michigan State University's Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/) and the Rice Annotation Project Database (RAP-DB; http://rapdb.dna.affrc.go.jp/). We systematically collected and processed public high-throughput datasets, and curated genomics, transcriptomics, ontology, pathway and protein domain information into functional gene sets, modules and a co-expression/co-function network. CARMO provides well-organized results for data comparison and annotation, and has a user-friendly web platform and interactive interface to allow comprehensive exploration by users. We illustrate the features of CARMO using four examples, including DNA methylation-related functional modules, interplay between hormones on a genome-wide scale, the relationship between genome-wide chromatin accessibility and tissue-specific gene expression, and the functional impact of yield-related SNPs identified by GWAS. These examples demonstrate the utility of CARMO as an important and easy-to-use resource for intensive functional study in rice.

RESULTS

Data collection and integration

We started with extensive collection of two main sources of information from rice (Table S1) for further integration. First, 10 584 functional gene sets, mostly predicted based on sequence homology, were collected, representing genes belonging to the same functional category or pathway, or sharing the same domain (Table S1). Then, 1819 samples from transcriptomic and genomic studies were collected (Table S1B). Figure 1 summarizes the workflow for how this information is integrated, and the diamond-shaped boxes represent the processed data used in CARMO as introduced below.

Gene sets with differential expression in pairwise comparisons. Gene sets with differential expression represent genes with a common response to a particular perturbation or treatment. For 1776 samples in 119 transcriptomic studies (Table S1C,D), all available pairwise sample comparisons in the same study were generated. Comparisons across studies were avoided in case of confounding batch effects. After removal of samples with too few differentially expressed genes, we obtained 4589 differentially expressed gene sets (Figure 1a). Methods S1 provides details of the methods.

Genome-wide expression profile across tissues. To characterize the expression profile of genes across tissues, 27 RNA-seq samples from 13 tissues were collected (Table S1E). We calculated the fragments per kilobase of exon per million fragments mapped (FPKM) for each gene in each sample (Trapnell *et al.*, 2012) (Figure 1a).

Co-expression network and gene sets derived from the network. Genes with concerted expression generally work in coordination (Stuart *et al.*, 2003). To search for potential co-regulators of given genes, a co-expression network was constructed. A partial correlation coefficient was used to measure the tendency for co-expression (Figure 1a), which is expected to reflect the direct relationship between genes (Kolpakov *et al.*, 1998). The network is of high quality, as demonstrated by a permutation test (see Methods S1). This is a major resource for construction of a co-expression/co-function network. Furthermore, this co-expression network was partitioned into 259 gene groups, forming the basis for further gene module organization (Figure 1a).

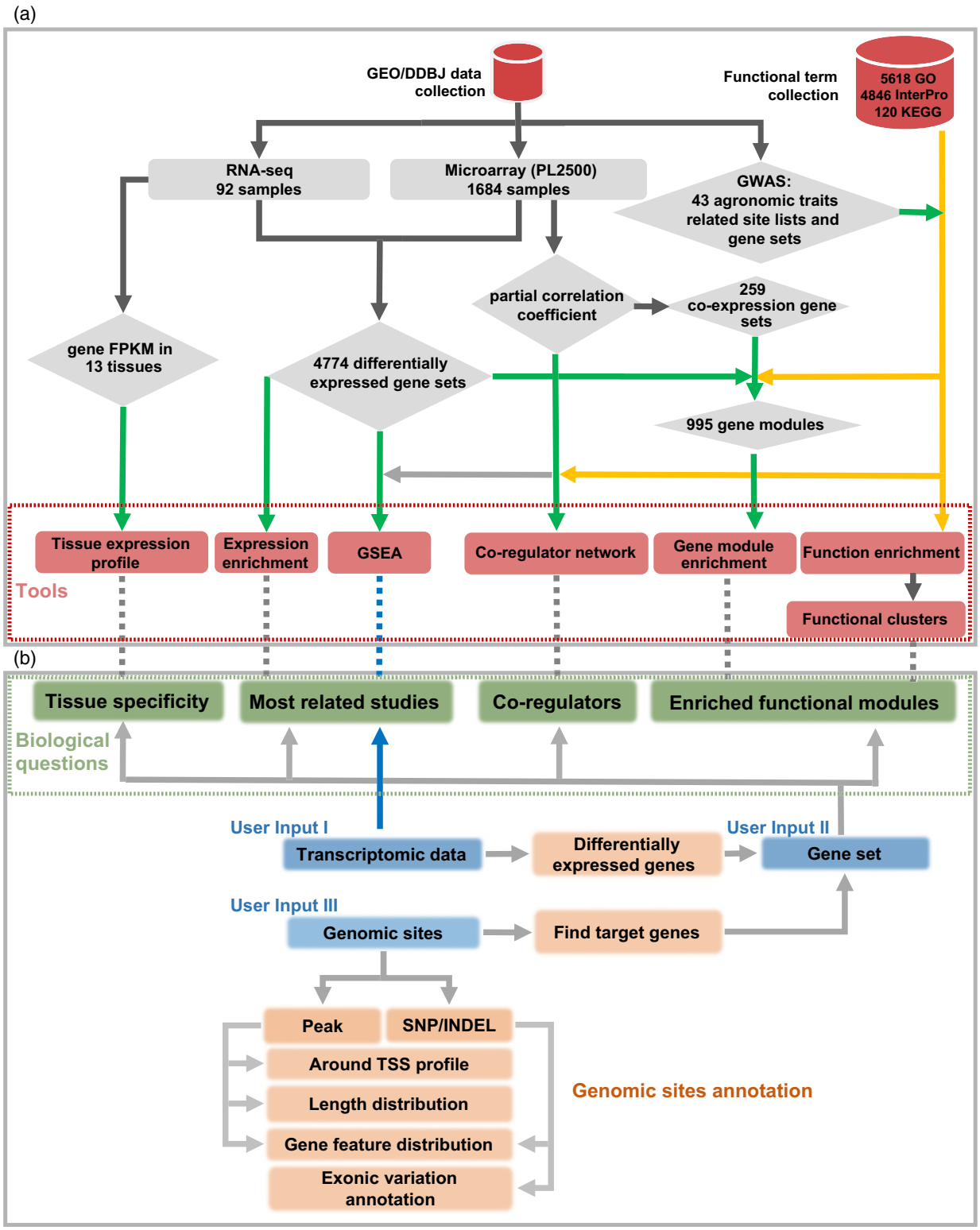

Figure 1. CARMO framework for data integration and web-based services.
(a) CARMO framework for systematic integration of gene function terms and multi-omics data from rice. The six diamond-shaped boxes represent data processed and used in CARMO; red boxes represent web-based tools that CARMO provides. The green arrows represent the flow of information from omics data; the yellow arrows indicate the flow of function term evidence.
(b) CARMO accepts gene lists, processed transcriptome data and genomic sites (regions or SNPs) as input. All corresponding tools (indicated by grey arrows for gene lists and a blue arrow for transcriptome data) may be applied independently or jointly.

Gene modules with concerted expression and related functions. As gene annotation based solely on functional homology is not sufficiently reliable, and one gene set only partially reflects the biological function, we organized gene sets of related function showing coordinated expression into higher-level modules, such that one gene module represents a group of genes expressed in concert to perform a specific function. In this way, the role of a gene with non-validated function may be deduced from its neighbors in the same module with relatively high confidence, as both sequence homology and experimental evidence are taken into account. We followed the method widely used in cancer research as proposed by Segal *et al.* (2004), and identified 995 statistically significant gene modules, representing 995 gene groups with concerted expression and related function (Figure 1a). Figure 2 shows a subset of the module map comprising 830 modules and 289 typical biological comparisons in rice. It is clear that genes preferentially expressed in callus and inflorescences are related to hormone activity and cell division, while the major role of shoot- and leaf-specific genes is photosynthesis, consistent with previous reports (Huang and Yeoman, 1984; Evans, 1989).

Genomic sites and related gene sets associated with key agronomic traits. Genomic sites and related genes associated with 43 rice traits (Table S1F) were collected from all available rice GWAS data (Huang *et al.*, 2010, 2012; Xu *et al.*, 2012) and integrated in CARMO. For any input study or gene list, it will be apparent which genes participate in regulation of key agronomic traits.

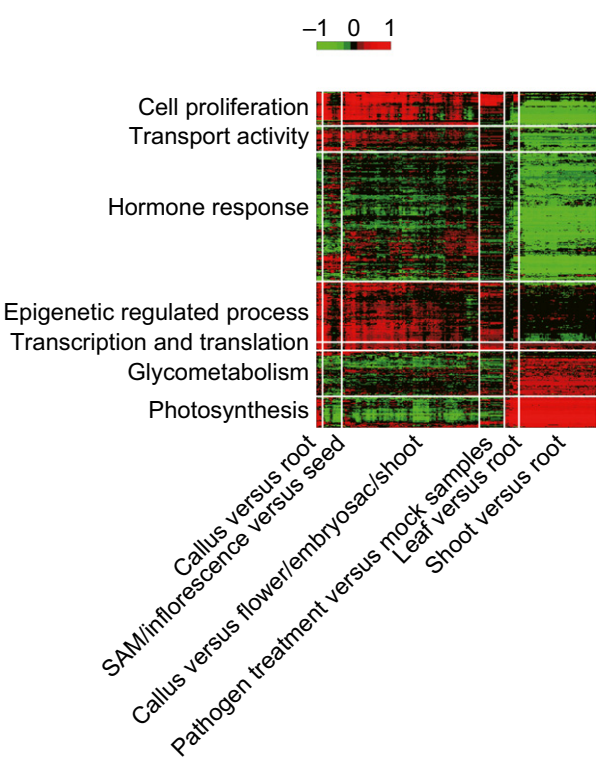

Figure 2. Matrix of modules (rows) versus array comparisons (columns). Red and green indicate genes in the module that are induced or repressed. Modules with similar expression behavior were organized into the same cluster, and each cluster is separated by white lines.

Functional exploration of gene lists: complementary approaches for integration of multiple evidence

High-throughput experiments often result in a large gene list of interest, possibly with thousands of genes. To elucidate the collective behavior of these genes and to identify key regulators may greatly facilitate downstream experiments. To address these issues, CARMO utilizes five complementary web-based tools (red boxes in Figures 1a and 3a) to integrate evidence from both omics data and that largely derived from homology comparison. Each tool has its own strength and focus, in order to provide in-depth data mining services for given gene lists. Figure 1(b) shows the general function of each tool, which may be applied individually or jointly. These tools are described according to their applications, and detailed uses are illustrated by examples in later sections.

Search for the most-related differential expression studies *to the input gene list: expression enrichment analysis*. For any given gene list, CARMO could provide the most-related pairwise transcriptomic comparisons, based on a statistical test of whether their differentially expressed genes have significant overlap with the input (Figure 3b). Details of common genes affected are also listed. From these results, users are shown which previously published treatments or gene perturbations are most closely related to their own study, and also the key genes involved. To further understand the functions or pathways affected, CARMO provides an option in the webpage of gene list annotation to submit the gene list for functional annotation as described below.

Integrative annotation of gene lists with high confidence: functional cluster and gene module enrichment. The traditional method of function enrichment analysis presents a collection of enriched functional terms for a given gene list, without description of the relationship among the terms. CARMO provides two web-based tools for integrative annotation of gene lists. The most intuitive way to integrate the enriched functions is to organize all relevant functional gene sets (sharing the same gene ontologies, pathways and domains and over-represented in input gene lists), into higher-level groups based on the number of genes they share (Huang *et al.*, 2009), as performed by the functional cluster module in CARMO (Figure 3c–e). This method helps to better elucidate the functions of the gene list based on the highly organized gene functional clusters. In addition, enrichment analyses for each individual source of functional terms are performed.

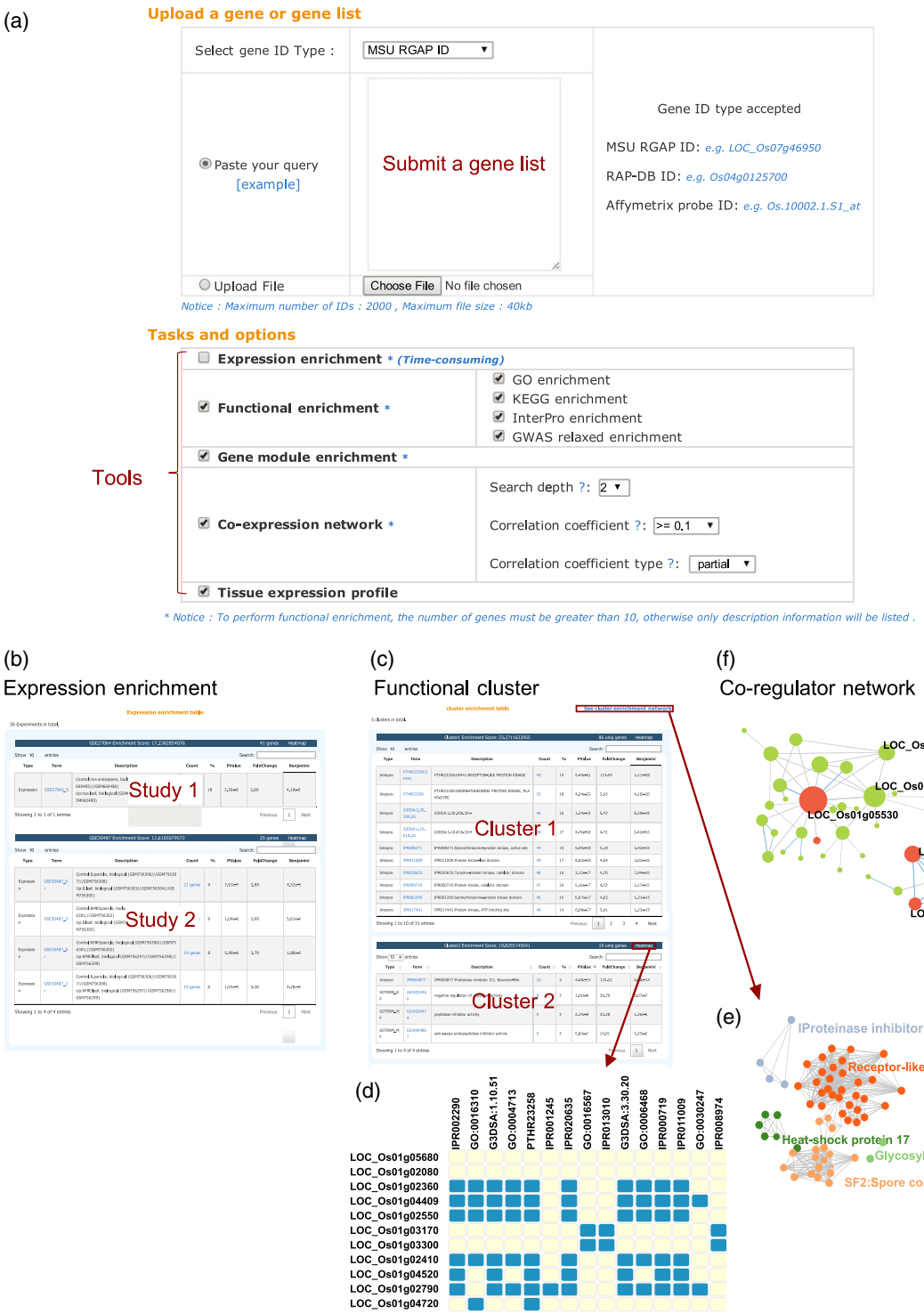

Figure 3. Input and output for integrative functional analysis of gene lists.
(a) Input and available analysis tools for gene list interpretation.
(b) Expression enrichment analysis uncovers the most-related transcriptomic studies to the input gene list. The enrichment score for each study is the geometric mean (in -log scale) of *P* values characterizing the enrichment of samples in the study.
(c) Output and related visualization tools for the functional cluster module. The enrichment score for each cluster is the geometric mean (in -log scale) of *P* values characterizing the enrichment of samples in the study.
(d) Heatmap showing the matrix of functional terms for each gene in the cluster. Each row represents a gene and each column represents a functional term.
(e) Network view of functional cluster enriched in input gene list.
(f) Visualization of a co-expression/co-function network.

The other integrative method is the gene module enrichment tool (Figure S1a), which calculates the enrichment of input gene list in the pre-compiled gene modules in CARMO. Each gene module represents a group of genes with concerted expression and related function. Given that gene modules incorporate evidence from both functional studies and omics data, annotation in this way not only provides relatively high confidence, but also helps to generate further hypotheses from the given data, as illustrated by Example I below.

It should also be noted that, for each cluster or gene module, CARMO also characterizes the functional terms for each gene via a heatmap (Figure 3d and Figure S1a), such that users may visualize the relationship between multiple genes and multiple terms directly, and are able to focus on potential key factors.

Visualization of co-regulators. Genes that show a high correlation coefficient with given gene list are extracted and presented in a network (Figure 3f). Each pair of co-expressed genes is connected by an edge, which is highlighted in blue if they also share common functions (Figure 3f), indicating relatively strong evidence for co-regulation between the pair of genes. The detailed functional annotation, pathway and domain information for all genes in the network are provided in the same web page below the network. To facilitate exploration, CARMO provides multiple interactive tools for manipulation of the network, and simultaneously highlights the genes selected from either network or the table in both sources.

Visualization of expression profile across tissues. For any given gene list, a heatmap of the expression profile across tissues, and the hierarchical clustering result, both gene-wise and tissue-wise, are presented in order to help determine the expression specificity of the genes of interest.

Interpretation of transcriptomic datasets

CARMO provides multiple tools for transcriptome comparison and annotation. Given a pair of transcriptomic datasets, two complementary methods of interpretation are applied

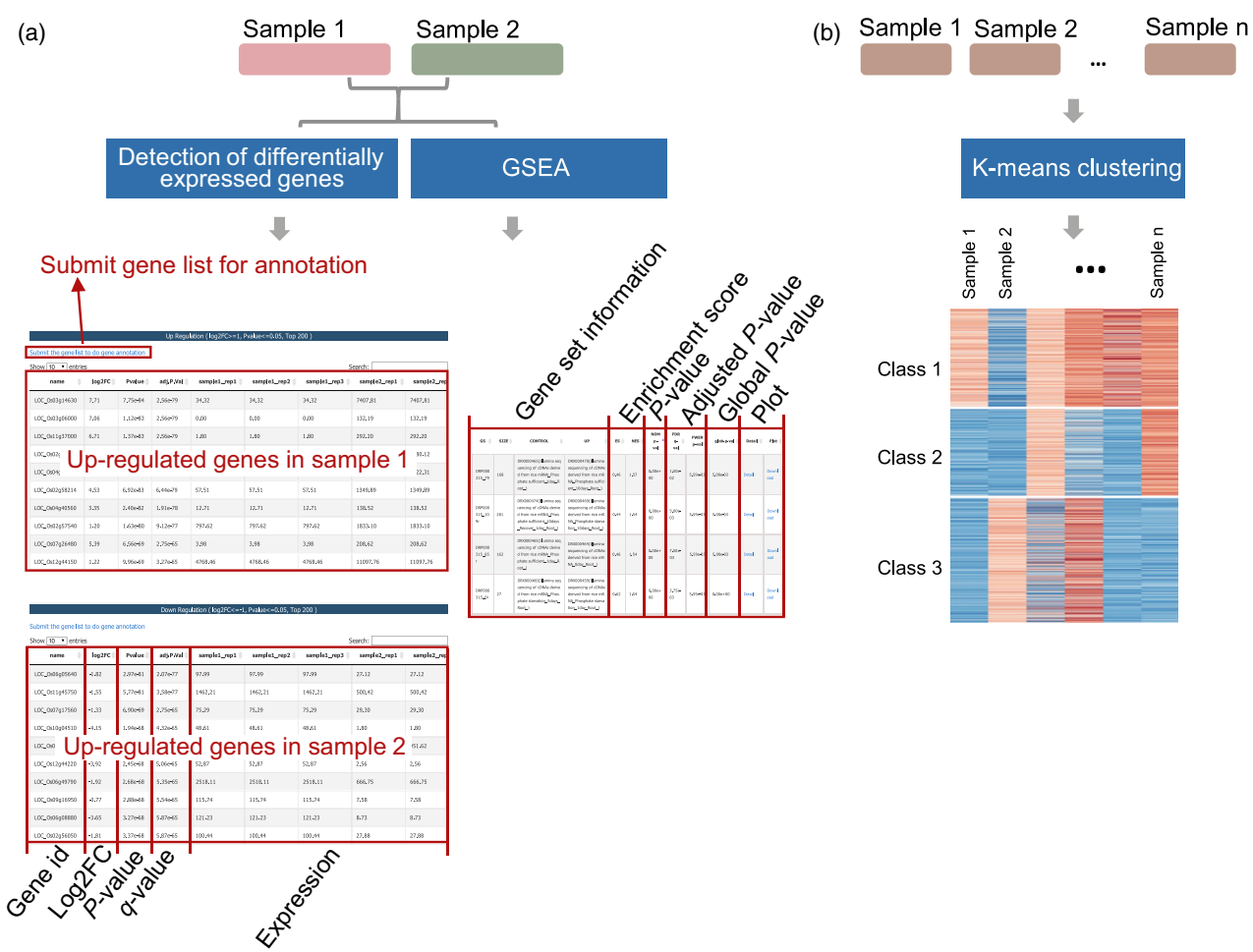

Figure 4. Input and output for transcriptomic data comparison between a pair of samples (a) and multiple samples (b).

(Figure 4a). One involves detecting differentially expressed genes followed by functional exploration of the gene list as mentioned above; the other involves using GSEA (Subramanian *et al.*, 2005) to search for pre-compiled gene sets sharing the same functions, pathways or domains that are preferentially enriched in up- or down-regulated genes. The current version of CARMO contains four types of pre-compiled gene sets for GSEA, including differentially expressed gene sets from pairwise comparison of transcriptome datasets, functional terms, domain information and pathways. Related statistics including *P* value, *q*-value (false discovery rate adjusted *P* value), fold change and enrichment score are listed in the results table (Figure 4a). The major difference between these two methods is that the former identifies enriched functional terms for up- and down-regulated genes separately, while GSEA considers the relative enrichment of gene sets between a pair of samples. Thus, the former method may recover gene sets that are enriched in both induced and repressed genes, which may be undetectable using GSEA if the enrichment is comparable between the up- and down-regulated gene lists. Users must take into account the focus of each method when interpreting their own data.

When the input contains multiple transcriptome samples (Figure 4b), CARMO applies k-means clustering (Eisen *et al.*, 1998) to group genes into groups with distinct expression profiles across samples. Genes in each cluster may further be submitted for gene list annotation. For a better understanding of the differences across samples, it is recommended that genes with differential expression be used as input.

Characterization of genomic and epigenomic datasets

For genomic datasets, CARMO accepts two types of input: genomic regions generally derived from epigenomic studies (Figure 5a), and SNPs or short insertions and deletions (INDELs) from genome-wide re-sequencing studies (Figure 5b). The results for former input are composed of two parts: (i) statistics and genomic distributions (Figure 5c), and (ii) target gene detection and annotation (Figure 5d). Notably, if a pair of region lists is given, three lists will be provided based on region overlap, resulting in one common region list and two unique region lists (Figure 5a). Next, common and unique target genes are identified, and all the gene list annotation methods mentioned above may be applied. For SNP or INDEL input, CARMO reports affected gene regions, and information for related genes. If the input is based on the current release of the rice genome (IRGSP 1.0), synonymous or non-synonymous variations may be distinguished for SNPs occurring in exons (Figure 5b). Furthermore, the function enrichment methods mentioned above are available for annotation of genes affected (see Example III).

Examples illustrating the power of CARMO for data mining of various types of omics data in rice

Example I: gene module enrichment analysis revealed close relationship between RNA helicase genes and DNA methylation. To understand the functional effect of DNA methylation, 98 genes grouped under GO terms (ID: 0006306) relevant for DNA methylation were used as input for gene list annotation. Some well-documented functions related to DNA methylation, including histone H3K9 methylation and cell proliferation, show up in the results for both individual GO term enrichment analysis and gene module enrichment, while only gene module enrichment analysis recovered the DEAD/DEAH RNA helicase-related genes significantly related to DNA methylation, with a *P* value of 3.92E-05 and fold enrichment of 4.2 (Table 1 and Table S2). In support of this finding, there is growing evidence demonstrating the essential role of RNA helicases in PIWI-interacting RNA-dependent DNA methylation, including Tud domain-containing (TDRD) family (Chen *et al.*, 2011), mouse VASA homolog (MVH) (He *et al.*, 2011), and DExD-box helicase MOV10-like-1 (MOV10L1) (He *et al.*, 2011) in mammals, and SILENCING DEFECTIVE 3 (SDE3) in Arabidopsis (Garcia *et al.*, 2012). A recent report in rice found that the otholog of DOMAINS rearranged methyltransferase2 (OsDRM2) interacts with the RNA helicase OseIF4A (Dangwal *et al.*, 2013), encoded by a gene that is present in the RNA helicase module collected by CARMO (Table S2). The genes in this module are thus good candidates for further functional validation of the relationship between RNA helicase and DNA methylation in rice.

Example II: analysis of multiple transcriptome datasets reveals the interplay between plant hormones. To understand the common and distinct effects of hormone treatment in rice, we collected all samples from a hormone treatment experiment (Garg *et al.*, 2012) characterizing the transcriptomic profile before and after addition of six hormones, including indole-3-acetic acid (IAA), 6-Benzylaminopurine (BAP, a synthetic cytokinin), abscisic acid (ABA), ethylene, salicylic acid and jasmonic acid (JA). Each of the six pairs of transcriptomic data (treated versus untreated) were uploaded to CARMO to detect genes regulated by various hormones. Next, k-means clustering implemented in CARMO was used to cluster the expression change pattern across the samples treated by six hormones, resulting in five classes (Figure S2). Figure 6(a) shows the four classes whose genes are obviously influenced by more than one hormone. Functional annotation implemented in CARMO was used to explore the functions of the five groups, and detailed functions and associated genes for each group are listed in Table S3.

Class 1 represents genes induced by all six hormones, with less response to BAP treatment and higher response

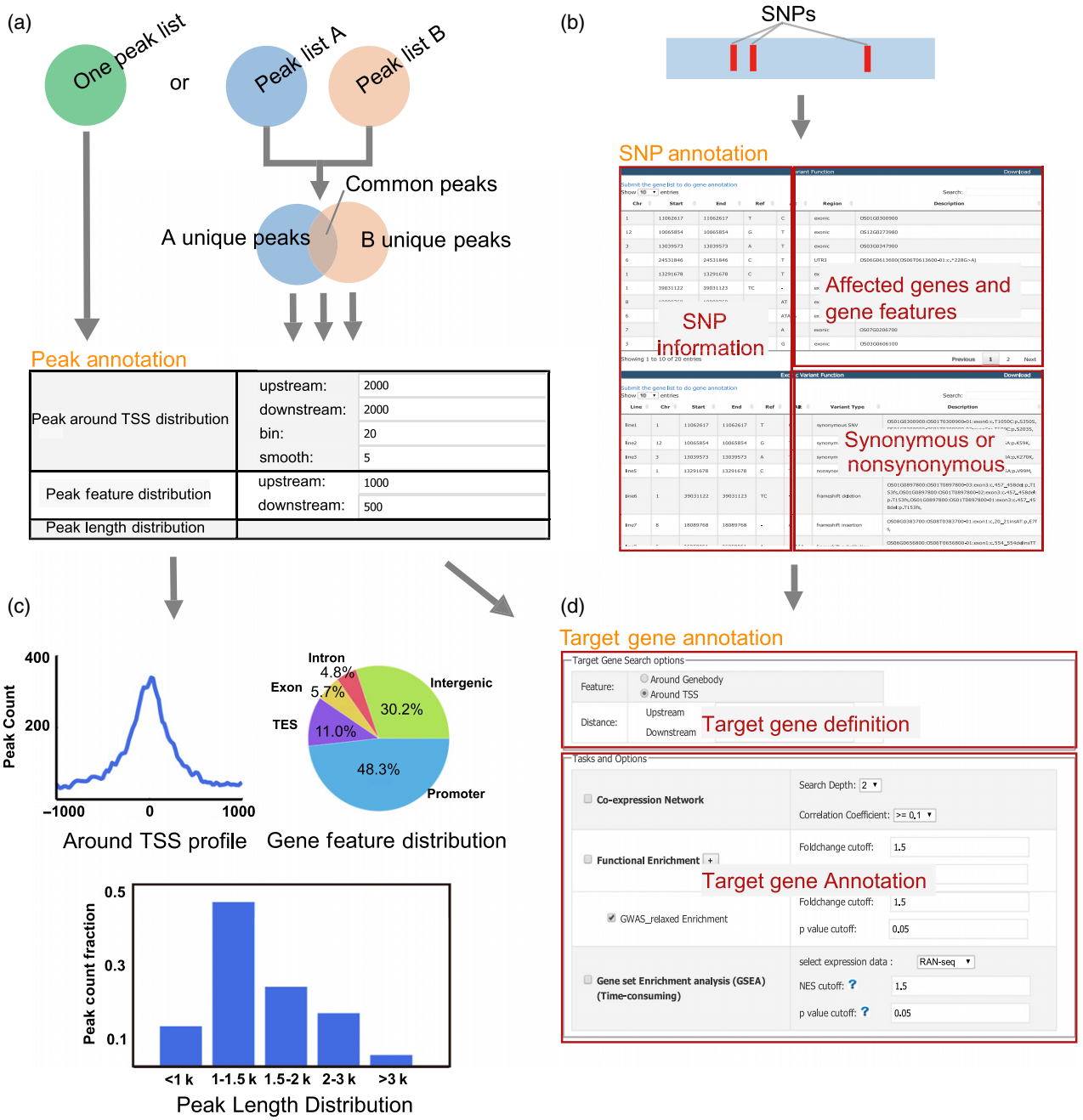

Figure 5. Input and output for genomic sites analysis pipeline.
(a) CARMO accepts an individual peak list or a pair of peaks, the latter of which are classified as common or biased peaks based on peak overlap before further annotation.
(b) Annotation results for SNPs and INDELs, including types of variant, and synonymous or non-synonymous exonic variant.
(c) Statistical results for the input peak list, including peak length distribution, gene feature distribution, and the distribution of the peak around the transcription start site.
(d) Genes surrounding the input genomic sites are defined as targets based on an optional cut-off, and functional annotation analyses are automatically performed.

to SA. The top enriched functional clusters (Figure 6b) and pathways (Figure S3) for genes in class 1 are oxidation/reduction and related enzymes, including glutathione *S*-transferase and P450 (Table S3), both of which are critical for catalyzing the oxidation of endogenous or exogenous chemicals. Consistently, studies in Arabidopsis indicate that glutathione *S*-transferase genes are induced by ethylene and salicylic acid (Xiang and Oliver, 1998; Sappl *et al.*, 2004; Yoshida *et al.*, 2009). In another report, JA was shown to induce accumulation of glutathione metabolic

Table 1 Biological processes related to DNA methylation, ranked by enrichment *P* value from low to high (details in Table S3)

Gene module	*P* value	Description
module_706	1.08E-117	DNA methylation
module_551	4.55E-73	Cell proliferation; histone H3K9 methylation
module_620	2.94E-48	Cell–cell signaling; virus-induced gene silencing; vegetative phase change; production of tasiRNAs involved in RNA interference
module_704	1.56E-40	Regulation of flower development; histone lysine methylation
module_550	2.91E-26	Epigenetic regulation of gene expression; histone binding; chromatin modification
module_737	1.26E-19	Histone phosphorylation; spindle assembly
module_568	2.81E-18	DNA repair; mitotic cell cycle; nucleotide excision repair
module_751	1.71E-13	Double-strand break repair via homologous recombination
module_774	3.00E-08	DNA endo-reduplication
module_640	7.02E-08	DNA-dependent ATPase activity; nucleolus organization
module_917	9.04E-07	Exonuclease activity; production of siRNAs involved in RNA interference; mitotic recombination
module_545	2.60E-05	Response to γ radiation; regulation of telomere maintenance; meiotic DNA double-strand break formation; telomere maintenance in response to DNA damage
module_676	3.92E-05	Nucleic acid binding; post-translational protein modification; ATP-dependent helicase activity; DEAD/DEAH box DNA/RNA helicase; N-terminal DEAD-like helicase

genes and enhance glutathione synthesizing capacity (Xiang and Oliver, 1998). Opposite expression changes were observed for genes in class 2 between the treatments of ABA and JA, and class 4 between the treatments of ABA and cytokinin (Figure 6a), representing antagonistic effects of these particular hormones, consistent with previous reports in Arabidopsis (Anderson *et al.*, 2004; Tran *et al.*, 2007; Lu *et al.*, 2014). The top enriched functional clusters of class 4 mostly include histidine kinases, chemotaxis Y protein (CheY), expansin and proteins involved in phosphorelay signal transduction (Table S3), all of which are critical components of the multi-step two-component system responsible for initiating cellular responses to environmental stimuli (Stock *et al.*, 2000). Similar results were obtained when using GSEA to identify enriched functions in genes induced by cytokinin (Figure 6c). Visual inspection of the co-regulator network of genes from class 4 identified the key factors involved in the cytokinin-responsive two-component system, with the top two sub-networks representing histidine kinases and expansins (Figure 6d). Cytokinin dehydrogenase, BTB (for BR-C, ttk and bab) protein involved in ubiquitination, and genes related to phosphor-transferase are also present in these two sub-networks, and are plausible candidates for further investigation of the downstream events following cytokinin-responsive signaling.

Example III: integrative analysis of DNase-seq and transcriptomic datasets reveals a close relationship between chromatin accessibility and tissue specificity. To illustrate the use of CARMO for integrative analysis of genomic and transcriptomic datasets, we compared data of DNaseI-hyper-sensitive sites (DHSs) characterizing the chromatin open state between callus and seedlings on a genome-wide scale (Zhang *et al.*, 2012). By exploring the different functions of genes surrounding tissue specific DHSs, we identified some interesting possibilities for further investigation (Figure 7a).

In this example, CARMO accepts as input two bed files describing the genomic coordinates of sequencing read-enriched regions (DHS peaks) from callus and seedlings, and gives three peak lists based on input peak overlap, including callus-unique DHS, seedling-unique DHS and common DHS, all of which show a high percentage around the transcription start site (Figure S4). Accordingly, all peak target genes are also divided into three lists, two biased and one common. Next, expression enrichment tool was used to identify the most-related transcriptome studies. The transcriptome comparison between young leaf (refer to as seedlings hereafter in this example) and growing callus was among the top enriched comparisons (Table S4), whose stages of tissues (Fujita *et al.*, 2010) were almost the same as those in the DHS study, and was used for downstream analysis. From the expression summary page of CARMO (Figure 7b), we found that 494 genes preferentially expressed in callus show higher levels of the open chromatin state in callus as compared to seedlings, while 503 genes that showed higher levels of the open chromatin state in seedlings displayed seedling-specific expression. These two gene lists represent callus- or seedling-specific genes that are directly regulated by their surrounding chromatin state, which is affected by transcription factor binding or epigenetic modification, and thus are among the key regulators of tissue specificity.

To explore the function of these two gene lists, functional annotation tools implemented in CARMO were used. Various sets of homeobox genes are specifically enriched in seedling-specific genes or callus-specific genes (Table 2), representing key transcription factors controlling tissue specificity. Members of the seedling-specific list participate in photosynthesis, chloroplast morphogenesis and sugar transport (Figure 7c and Table S4). In contrast, the functions significantly enriched in the callus-specific gene

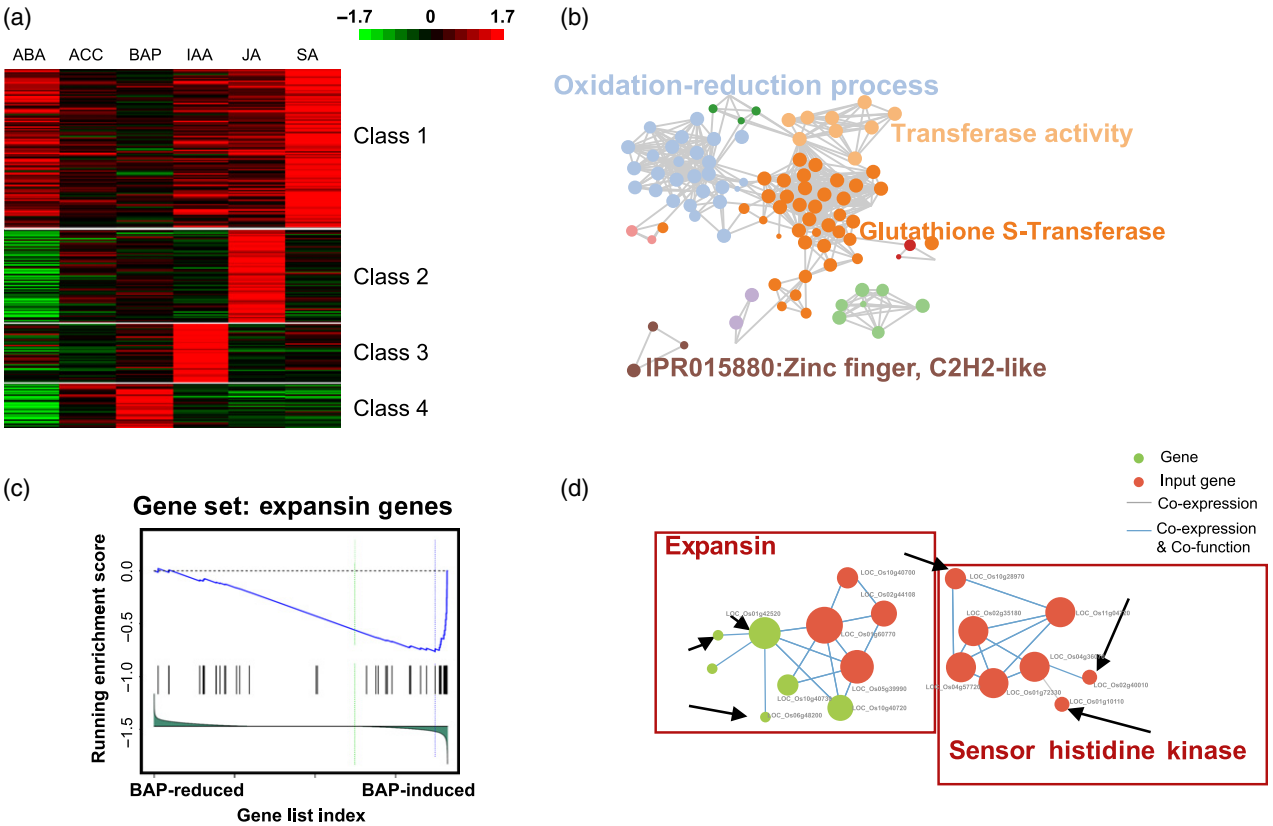

Figure 6. Interplay of plant hormones at the transcriptomic level.
(a) Four classes of genes affected by at least two hormones All five classes of hormone response genes are shown in Figure S2.
(b) Functional cluster of genes in class 1.
(c) GSEA result showing enrichment of the expansin gene set in BAP-induced genes.
(d) Co-expression co-function network of genes in class 4. Red nodes represent genes in the input list; blue edges connect pairs of genes with co-expression and shared functions; the genes in the two are mostly histidine kinases and expansins, except those indicated by arrows.

set include regulation of the cell cycle and transcription, which is expected as cell division and transcription are highly active in callus tissues (Meins and Thomas, 2003). It is interesting that genes associated with peroxidase are also significantly over-represented among callus-specific genes (Figure 7d and Table S4). It has been demonstrated that the redox state is essential for organogenesis of plant callus (Bonfill *et al.*, 2003). Similarly, in animals, redox homeostasis regulated by peroxidase is important in regulation of self-renewal and differentiation of stem cells (Hochmuth *et al.*, 2011; Wang *et al.*, 2013). No specific peroxidase has been reported to be involved in regulation of plant callus regeneration, and the peroxidases that are preferentially expressed in callus and are regulated by a surrounding open chromatin state represent good candidates for further study of redox regulation in plant callus.

Example IV: genes affected by yield-associated SNPs are preferentially expressed in panicles and involved in kinase-related processes. We collected yield-related GWAS results, including grain weight, panicle number and seed number per panicle, and obtained 60 SNPs/INDELs associated with rice yield (Huang *et al.*, 2010, 2012; Zhao *et al.*, 2011), which were further used as the input into the genomic annotation module. Genes affected by these SNPs/INDELs, gene feature distribution, and synonymous or non-synonymous variations are listed in Table S5. Function annotation modules in CARMO revealed that kinases are significantly over-represented in the gene list (Table 3), including receptor-like cytoplasmic kinases, cyclins controlling cell proliferation, lectin protein kinases and phosphatidylinositol 3-kinase-related kinase (PIKK), all of which are essential regulators of rice seed development (Fabian-Marwedel *et al.*, 2002; Gamuyao *et al.*, 2012; Liu *et al.*, 2012; Cheng *et al.*, 2013; Ramegowda *et al.*, 2014). The expression profile of the input gene list across tissues reveals that those genes are preferentially expressed in panicles and reproductive tissues (Figure 8), supporting the genetic evidence from GWAS that they are actively involved in control of rice yield.

Web-based integrative services

To make the platform easy to use, we packaged various analysis procedures for various data types into an auto-

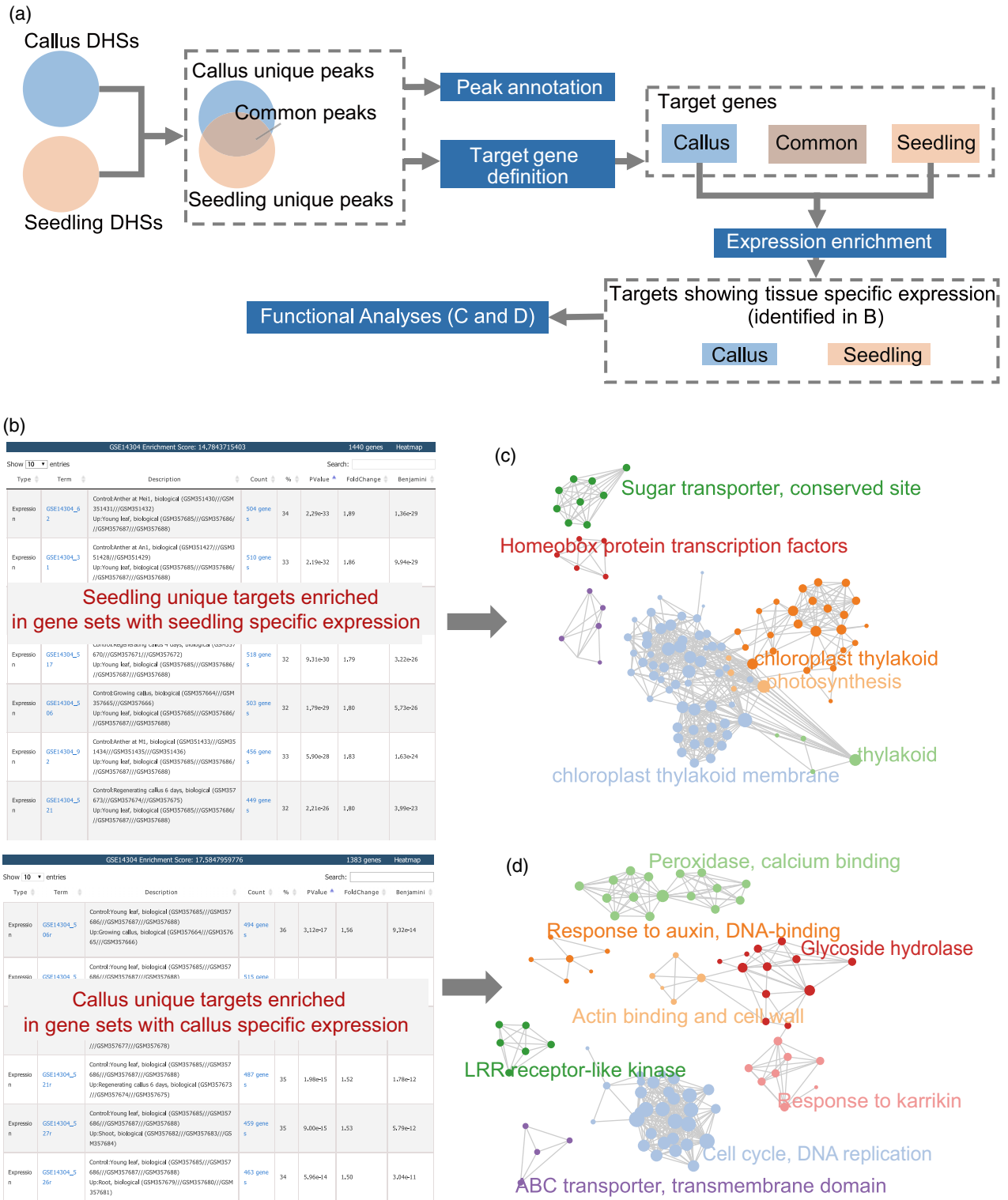

Figure 7. Integrative comparison of tissue-specific DHS and transcriptomic data.
(a) Two bed files listing coordinates of callus DHS and seedlings DHS are submitted to CARMO, and the following pipeline is automatically deployed: divide the two inputs to common and unique peak files, summarize peak statistics, detect common and unique target genes, and perform integrative functional analyses for each gene list.
(b) Layout of the expression enrichment result, which identified gene sets with tissue-specific expression enriched in tissue-specific DHS targets.
(c) Cluster view of enriched functions of callus-specific targets also showing callus-specific expression.
(d) Cluster view of enriched functions of seedling-specific targets also showing seedling-specific expression.

Table 2 Homeobox genes specifically expressed in seedling (12 genes) or callus (five genes) that are regulated by the open state of the surrounding chromatin

	Gene		Annotation
Seedling	LOC_Os10 g39030	Os10 g0534900	Homeobox domain-containing protein
	LOC_Os01 g74020	Os01 g0971800	MYB family transcription factor
	LOC_Os03 g47740	Os03 g0680800	Homeodomain protein
	LOC_Os03 g20900	Os03 g0325500	MYB transcription factor
	LOC_Os07 g30130	Os07 g0484700	MYB family transcription factor
	LOC_Os10 g01470	Os10 g0103700	Homeobox-associated leucine zipper
	LOC_Os08 g19650	Os08 g0292900	Homeobox protein knotted-1
	LOC_Os05 g35500	Os05 g0429900	MYB family transcription factor
	LOC_Os03 g52239	Os03 g0732100	Homeobox domain-containing protein
	LOC_Os12 g06340	Os12 g0160500	BEL1-like homeodomain transcription factor
	LOC_Os01 g44390	Os01 g0635200	MYB family transcription factor
	LOC_Os06 g24070	Os06 g0348800	MYB-like DNA-binding domain-containing protein
Callus	LOC_Os03 g10210	Os03 g0198600	Homeobox domain-containing protein
	LOC_Os07 g39320	Os07 g0581700	Homeobox domain-containing protein
	LOC_Os01 g62310	Os01 g0840300	Homeobox domain-containing protein
	LOC_Os01 g63510	Os01 g0854500	Homeobox domain-containing protein
	LOC_Os07 g48560	Os07 g0684900	Homeobox domain-containing protein

Table 3 Nine yield-associated genes from GWAS involved in kinase-related processes

Gene		Annotation
LOC_Os02 g43550	Os02 g0652000	Cyclin
LOC_Os02 g48360	Os02 g0714200	Pyrophosphate-fructose 6-phosphate 1-phosphotransferase subunit α
LOC_Os03 g27990	Os03 g0397700	Strubbelig receptor family 7 precursor
LOC_Os03 g30130	Os03 g0415200	Phospholipase C
LOC_Os05 g09500	Os05 g0187100	Hexokinase
LOC_Os06 g29810	Os06 g0494100	Lectin protein kinase family protein
LOC_Os08 g38200	Os08 g0489800	Phosphatidylinositol 3- and 4-kinase family protein
LOC_Os09 g37890	Os09 g0551500	Serine/threonine protein kinase receptor precursor
LOC_Os12 g41180	Os12 g0604700	LSTK-1-like kinase

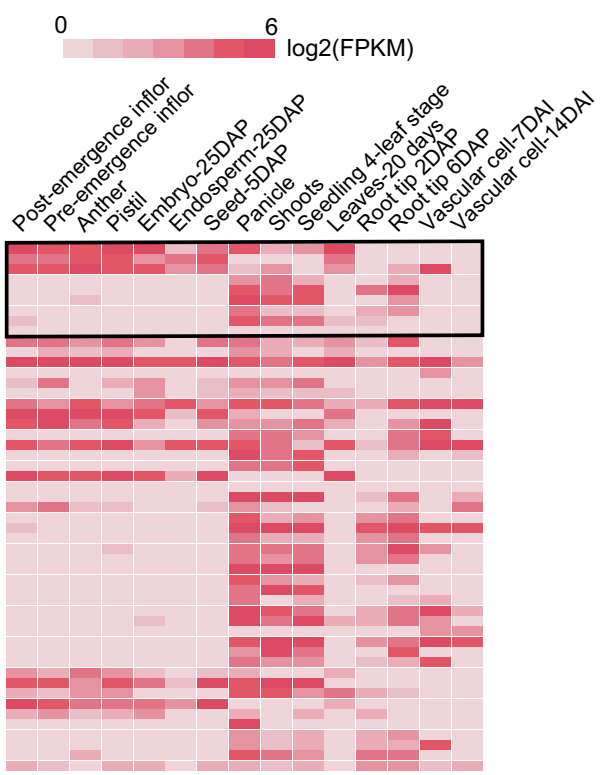

Figure 8. Heatmap showing the expression profile of yield-related genes across tissues.

matic pipeline. Only one input file is required, and all relevant results, including the statistic information, genes affected, related functions and comparison with previously published studies, are well-organized (Figure 1b).

It should also be noted that, to make images user-friendly and interactive, we present figures using d3.js, which is a widely used JavaScript library for web data visualization. Importantly, with d3.js, we are able to draw interactive graphs with high resolution in the browser. Co-expression networks, functional networks, heatmaps and bar graphs illustrating *P* values were dynamically generated by d3.js, and these images may be manipulated interactively (Figure 3 and Figure S1), including zooming in and out, moving within the image, highlighting and so forth. Additionally, the images may be conveniently downloaded in publishable high-quality format.

DISCUSSION

CARMO is a web-based platform for intensive functional exploration of rice omics data, whose power lies in the comprehensive collection and integration of information from both multi-omics data and diverse functional evidence from rice, which is further organized into gene sets and higher-level gene modules. It has four major features. The first is the ability to search for the gene lists derived from 1819 published rice omics samples that have significant overlap with input gene list. Second, 15 401 functional gene sets covering 49 469 genes are integrated from various sources, and, for any given gene list, all enriched functional gene sets are organized into clusters or modules, which help to elucidate the role of a given gene list on the level of gene functional unit. Third, it provides the first annotation platform for multi-omics rice data, and may be applied for interpretation of datasets characterizing transcriptome or genomic sites under comparison, open chromatin, epigenetic modifications, and genomic re-sequencing results. Fourth, use of various interactive visualization tools, including networks of co-regulators, cluster views of functional groups and heatmaps of genes in functional clusters or modules, make CARMO a user-friendly exploration platform.

We demonstrated the performance of CARMO on multiple transcriptomic and genomic datasets, and found that CARMO not only reproduced evidence previously reported, but also proposed useful functional insights for further experimental exploration, suggesting that CARMO is an invaluable resource for extracting biological insight from rice omics data.

With the accelerated accumulation of multi-omics data, it is of particular importance to keep data updated frequently. Here we developed semi-automatic pipeline from data downloading to result processing, which make it easy to keep the database updated, as well as to extend it to other species as required. The omics data used in the current version of CARMO are mainly from transcriptome studies, which are the predominant type of public rice omics data. However, with the rapid generation of other types of omics data in the future, CARMO may expand accordingly, especially to include epigenomic data characterizing the binding profile of transcription factors and epigenetic modifications, which are important for elucidation of regulatory networks. Meanwhile, more bioinformatics tools for intensive analysis of (epi)genomics data will be incorporated into CARMO, including tools for quantitative comparison of epigenomic data from various samples, and for detection of transcription factor binding motifs as developed previously (Shao *et al.*, 2012). Taken together, CARMO aims at systematic organization of evidence from diverse sources, to provide comprehensive and reliable interpretation for multi-omics data in rice, and thereby help to facilitate subsequent hypothesis-driven research.

EXPERIMENTAL PROCEDURES

Data collection

Rice microarray data are retrieved from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/), including 111 Affymetrix microarray experiments (GEO Series, GSE) under GEO Platform (GPL) 2025, with 1684 microarray samples (GEO SOFT format Sample, GSM) in total (Edgar *et al.*, 2002) (Table S1). For each experiment, all potential pairwise comparisons were performed, resulting in 2712 comparisons. All eight RNA-seq experiments performed in Nipponbare (*Oryza sativa* L. ssp. *japonica*), including 92 samples and 267 comparisons, were collected from the DDBJ database (Ogasawara *et al.*, 2013). A total of 4846 GO terms were integrated from the Affymetrix rice annotation (http://www.affymetrix.com), Michigan State University's Rice Genome Annotation Project (Kawahara *et al.*, 2013) and Gramene (Ware *et al.*, 2002), the last of which is mainly based on information from the RAP-DB (http://rapdb.dna.affrc.go.jp/). We also incorporated 120 rice pathways from the KEGG database (Kanehisa and Goto, 2000) and 5618 domains from the EMBL InterPro database (Hunter *et al.*, 2012).

Pre-processing of RNA-seq data

We started by cleaning the sequencing reads, including removing bases with a low quality score (<20) and irregular GC content, cutting out sequencing adaptors and filtering short reads. Then TopHat (Trapnell *et al.*, 2009) was used to map the read to genomic regions, followed by counting the number of reads in each gene.

Preparation of gene lists with differential expression

For detection of differentially expressed genes, the Bioconductor package *limma* (Smyth, 2004, 2005; Ritchie *et al.*, 2015) was used for microarray data, and DESeq (Anders and Huber, 2010) was used for RNA-seq data. Differentially expressed genes for microarray sample comparison were defined based on the following criteria: $|\log_2$ (fold change)$|$ >3, q-value <0.001. For RNA-seq data, which are expected to be more sensitive with respect to detecting differential expression, the combined criteria of $|\log_2$ (fold change)$|$ >1 and q-value <0.05 was used. Finally, gene lists with gene number <14 were excluded for further analysis, resulting in 4589 gene sets in total.

Detection of genes with tissue-specific expression

We performed statistical tests to detect tissue-specific expression based on the method proposed by Ge *et al.* (2005). Twenty-seven RNA-seq samples from 13 tissues were collected (Table S1E). Cufflinks (Trapnell *et al.*, 2012) was used to quantify gene expression levels (fragments per kilobase of exon per million fragments mapped, FPKM). For each gene, the mean FPKM of samples from the same tissue is recorded. Tissue-specific genes are defined based on the combined criteria: (i) the FPKM is more than three standard deviations above the FPKM in the remaining tissues, (ii) the ratio between the FPKM of the tissue-specific gene and that of the second most highly expressed gene is >2, and (iii) the FPKM is >5.

In addition to genes specifically expressed in one tissue, we also defined genes that are highly expressed in two or three tissues as tissue-specific genes as proposed by Ge *et al.* (2005). In

the heatmap online showing the expression intensity of input genes across tissues, tissue-specific expression is highlighted by a blue box.

Construction of co-expression networks based on partial correlation coefficient

For all 57 381 probes on the Affymetrix platform (GPL2025 in GEO), we first removed probes representing multiple genes, and filtered redundant probes corresponding to the same gene, such that only the probe with highest mean intensity across all samples was retained, resulting in 18 212 probes. Next, sample pairs without large differences were removed, such that only comparisons with more than 100 differentially expressed probes ($|\log_2$ (fold change)$|$ >1 and q-value <0.001) were retained, leading to 2185 sample pairs.

GeneNet (Schafer and Strimmer, 2005) was used to construct co-expression networks of genes based on partial correlation of the expression ratio (Figure 1a). We used the method proposed by Ma *et al.* (2007), using an iterative process with 1000 iterations for calculation of partial correlation. In each iteration, a partial correlation coefficient was calculated among 2000 genes that were randomly chosen from the total genes. The ultimate partial correlation coefficient of each gene pair was the minimum value over 1000 cycles of calculation. We tested the accuracy of the network by determining whether genes with a high correlation coefficient participate in the same KEGG pathways. Using a partial correlation coefficient >0.01 as the cut-off, 5708 gene pairs were obtained, among which 254 pairs of genes shared the same KEGG pathway, which is significantly higher than in random permutations (Figure S5). Specifically, we randomly permutated the genes in the KEGG gene lists (repeated 1000 times). For each iteration, 5708 gene pairs are randomly selected from all annotated rice genes, and the number of pairs in the same KEGG pathway is recorded. None of the 1000 random result is higher than 254, thus the permutation P value is smaller than 1E-3 (Figure S5).

Construction of gene modules

Our gene module construction method was a modified version of Segal's procedure (Segal *et al.*, 2004). We first collected all 10 584 gene sets from GO, KEGG and InterPro, as well as 259 gene sets showing co-expression based on partition of the co-expression network via the MCL algorithm (Enright *et al.*, 2002). Next, differentially expressed genes sets were defined based on the criteria $|\log_2$ (fold change)$|$ >2 and q-value <0.01. All functional gene sets with similar expression behavior were clustered based on whether a significant proportion of the genes in the set showed a consistent expression change (either up- or down-regulated) across arrays (Figure 2). After removing genes with no coordinated expression with the gene set they belong to, and further removal of modules with a false discovery rate <0.05, in 'leave one out' cross-validation (Segal *et al.*, 2004), we finally obtained 995 gene modules containing 20 476 genes. For each module, we manually removed redundant descriptions.

Functional classification based on the network clustering method

To obtain an input gene list, the EASE score (a modified Fisher exact test) (Huang *et al.*, 2009) was used to detect enriched gene sets sharing the same features. The EASE score was used to make the detection more conservative with fewer enriched gene sets, as the canonical Fisher exact test is too sensitive. For example, when the size of a pre-compiled gene set is small, the canonical Fisher exact test usually shows significant enrichment, even if only a few genes in the input gene list are present in the gene set, possibly due to random effect.

Next, all enriched functions were clustered such that related functional terms are organized into the same cluster. Such a simplified approach allows users to quickly obtain an overview of the functions of the input genes (Huang *et al.*, 2009). Briefly, a functional network is organized such that each node represents a functional term, and the edge between nodes represents a connection between two functional terms sharing at least three genes. Then, the MCL algorithm was applied to divide the network graph based on the bootstrapping method. The results are presented in a user-friendly interactive interface via an in-house script using d3 JavaScript library.

GSEA data preparation

GSEA-R (Subramanian *et al.*, 2005), the R implementation of GSEA, with minor modifications was used for data mining from pairwise comparison of transcriptomic datasets. The purpose of GSEA is to test whether given gene sets are enriched in up- or down-regulated genes from pairwise comparisons. In addition, the rank may be taken into account by proper choice of the scoring metric (Subramanian *et al.*, 2005).

Characterization of genomic sites: statistics, distribution, target gene definition and function enrichment analyses

For each set of genomic regions, statistics including length, density around the transcription start site, and distribution in relation to gene annotation are provided. If two regions are uploaded, a common region list and two unique region lists are given based on whether the two input lists have overlap with each other. Target genes are defined as the nearest genes to the given genomic sites. For input list of SNPs/INDELs based on the current release of rice genome (IRGSP 1.0), ANNOVAR (Wang *et al.*, 2010) was applied to determine whether the mutation sites are synonymous substitutions or non-synonymous substitutions. The codon frame information in the RAP-DB (Sakai *et al.*, 2013) was used. For previous releases, no codon frame information is available, and only genes affected are listed without describing the substitution type of SNPs/INDELs in exons. Functional analyses modules are automatically executed for functional interpretation of target genes.

Web server implementation

CARMO was designed as a relational database using a typical LAMP (Linux, Apache, MySQL and PHP) platform aided by JavaScript. An overview of the scheme underlying CARMO is shown in Figure 1(b). The in-house scripts for data processing were written in Python, and are available upon request.

ACKNOWLEDGEMENTS

We thank Dr. A. Sugathan from Five Prime Therapeutics for helpful revisions, and Dr. L. Xu from our institute and Z. Shao from CAS-MPG Partner Institute for Computational Biology for useful comments on this study. This work was supported by the 'Strategic Priority Research Program' of the Chinese Academy of Sciences (grant number XDA01020304) and sponsored by Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Output for functional annotation tools in addition to those shown in Figure 3.

Figure S2. Five groups of hormone response genes were identified via k-means clustering. The fifth group representing genes specifically regulated by ABA is not shown in Figure 6.

Figure S3. CARMO allows visualization of genes in enriched pathways.

Figure S4. Statistics for DHS peaks.

Figure S5. The distribution of random permutation results.

Table S1. Detailed information on datasets used in this study.

Table S2. Gene modules enriched in genes relevant for DNA methylation.

Table S3. Functional clusters and separate terms enriched in each of the five classes of hormone-responsive genes shown in Figure S2.

Table S4. Gene list annotation result for callus-/seedling-specific DHS targets.

Table S5. Annotation for yield-related SNPs.

Methods S1. Calculation of the percentage of rice GO annotations with deduced functions instead of from direct experimental evidence.

REFERENCES

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

Anderson, J.P., Badruzsaufari, E., Schenk, P.M., Manners, J.M., Desmond, O.J., Ehlert, C., Maclean, D.J., Ebert, P.R. and Kazan, K. (2004) Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in Arabidopsis. *Plant Cell*, **16**, 3460–3479.

Bonfill, M., Cusidó, R., Palazón, J., Canut, E., Piñol, M.T. and Morales, C. (2003) Relationship between peroxidase activity and organogenesis in *Panax ginseng* calluses. *Plant Cell Tissue Organ Cult.* **73**, 37–41.

Chen, C., Nott, T.J., Jin, J. and Pawson, T. (2011) Deciphering arginine methylation: Tudor tells the tale. *Nat. Rev. Mol. Cell Biol.* **12**, 629–642.

Cheng, X., Wu, Y., Guo, J., Du, B., Chen, R., Zhu, L. and He, G. (2013) A rice lectin receptor-like kinase that is involved in innate immune responses also contributes to seed germination. *Plant J.* **76**, 687–698.

Dangwal, M., Malik, G., Kapoor, S. and Kapoor, M. (2013) *De novo* methyltransferase, OsDRM2, interacts with the ATP-dependent RNA helicase, OseIF4A, in rice. *J. Mol. Biol.* **425**, 2853–2866.

Du, Z., Zhou, X., Ling, Y., Zhang, Z. and Su, Z. (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* **38**, W64–W70.

Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584.

Evans, J. (1989) Photosynthesis and nitrogen relationships in leaves of C3 plants. *Oecologia*, **78**, 9–19.

Fabian-Marwedel, T., Umeda, M. and Sauter, M. (2002) The rice cyclin-dependent kinase-activating kinase R2 regulates S-phase progression. *Plant Cell*, **14**, 197–210.

Fujita, M., Horiuchi, Y., Ueda, Y., et al. (2010) Rice expression atlas in reproductive development. *Plant Cell Physiol.*, **51**, 2060–2081.

Gamuyao, R., Chin, J.H., Pariasca-Tanaka, J., Pesaresi, P., Catausan, S., Dalid, C., Slamet-Loedin, I., Tecson-Mendoza, E.M., Wissuwa, M. and Heuer, S. (2012) The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. *Nature*, **488**, 535–539.

Garcia, D., Garcia, S., Pontier, D., Marchais, A., Renou, J.P., Lagrange, T. and Voinnet, O. (2012) Ago hook and RNA helicase motifs underpin dual roles for SDE3 in antiviral defense and silencing of nonconserved intergenic regions. *Mol. Cell*, **48**, 109–120.

Garg, R., Tyagi, A.K. and Jain, M. (2012) Microarray analysis reveals overlapping and specific transcriptional responses to different plant hormones in rice. *Plant Signal Behav.* **7**, 951–956.

Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S.M. and Aburatani, H. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, **86**, 127–141.

He, X.J., Chen, T. and Zhu, J.K. (2011) Regulation and function of DNA methylation in plants and animals. *Cell Res.* **21**, 442–465.

Hochmuth, C.E., Biteau, B., Bohmann, D. and Jasper, H. (2011) Redox regulation by Keap1 and Nrf2 controls intestinal stem cell proliferation in Drosophila. *Cell Stem Cell*, **8**, 188–199.

Huang, B.C. and Yeoman, M.M. (1984) Callus proliferation and morphogenesis in tissue cultures of *Arabidopsis thaliana* L. *Plant. Sci. Lett.* **33**, 353–363.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57.

Huang, X., Wei, X., Sang, T. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967.

Huang, X., Zhao, Y., Wei, X. *et al.* (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* **44**, 32–39.

Hunter, S., Jones, P., Mitchell, A. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306–D312.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30.

Kawahara, Y., de la Bastide, M., Hamilton, J.P. *et al.* (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.

Kolpakov, F.A., Ananko, E.A., Kolesov, G.B. and Kolchanov, N.A. (1998) GeneNet: a gene network database and its automated visualization. *Bioinformatics*, **14**, 529–537.

Lee, I., Seo, Y.S., Coltrane, D., Hwang, S., Oh, T., Marcotte, E.M. and Ronald, P.C. (2011) Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proc. Natl Acad. Sci. USA*, **108**, 18548–18553.

Liu, J., Zhou, J. and Xing, D. (2012) Phosphatidylinositol 3-kinase plays a vital role in regulation of rice seed vigor via altering NADPH oxidase activity. *PLoS One*, **7**, e33817.

Lu, Y., Tarkowska, D., Tureckova, V., Luo, T., Xin, Y., Li, J., Wang, Q., Jiao, N., Strnad, M. and Xu, J. (2014) Antagonistic roles of abscisic acid and cytokinin during response to nitrogen depletion in oleaginous microalga *Nannochloropsis oceanica* expand the evolutionary breadth of phytohormone function. *Plant J.* **80**, 52–68.

Ma, S., Gong, Q. and Bohnert, H.J. (2007) An Arabidopsis gene network based on the graphical Gaussian model. *Genome Res.* **17**, 1614–1625.

Meins, F. Jr and Thomas, M. (2003) Meiotic transmission of epigenetic changes in the cell-division factor requirement of plant cells. *Development*, **130**, 6201–6208.

Ogasawara, O., Mashima, J., Kodama, Y., Kaminuma, E., Nakamura, Y., Okubo, K. and Takagi, T. (2013) DDBJ new system and service refactoring. *Nucleic Acids Res.* **41**, D25–D29.

Ramegowda, V., Basu, S., Krishnan, A. and Pereira, A. (2014) Rice GROWTH UNDER DROUGHT KINASE is required for drought tolerance and grain yield under normal and drought stress conditions. *Plant Physiol.* **166**, 1634–1645.

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 Res.* **43**, e47.

Sakai, H., Lee, S.S., Tanaka, T. *et al.* (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* **54**, e6.

Sappl, P.G., Onate-Sanchez, L., Singh, K.B. and Millar, A.H. (2004) Proteomic analysis of glutathione S-transferases of *Arabidopsis thaliana* reveals differential salicylic acid-induced expression of the plant-specific phi and tau classes. *Plant Mol. Biol.* **54**, 205–219.

Schafer, J. and Strimmer, K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.

Segal, E., Friedman, N., Koller, D. and Regev, A. (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* **36**, 1090–1098.

Segal, E., Friedman, N., Kaminski, N., Regev, A. and Koller, D. (2005) From signatures to models: understanding cancer using microarrays. *Nat. Genet.* **37**, S38–S45.

Shao, Z., Zhang, Y., Yuan, G.C., Orkin, S.H. and Waxman, D.J. (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.* **13**, R16.

Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article 3.

Smyth, G.K. (2005) *limma:* linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A. and Dudoit, S., eds). Berlin: Springer, pp. 397–420.

Stock, A.M., Robinson, V.L. and Goudreau, P.N. (2000) Two-component signal transduction. *Annu. Rev. Biochem.* **69**, 183–215.

Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Subramanian, A., Tamayo, P., Mootha, V.K. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tran, L.S., Urao, T., Qin, F., Maruyama, K., Kakimoto, T., Shinozaki, K. and Yamaguchi-Shinozaki, K. (2007) Functional analysis of AHK1/ATHK1 and cytokinin receptor histidine kinases in response to abscisic acid, drought, and salt stress in Arabidopsis. *Proc. Natl Acad. Sci. USA*, **104**, 20623–20628.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164.

Wang, K., Zhang, T., Dong, Q., Nice, E.C., Huang, C. and Wei, Y. (2013) Redox homeostasis: the linchpin in stem cell self-renewal and differentiation. *Cell Death Dis.* **4**, e537.

Ware, D.H., Jaiswal, P., Ni, J. *et al.* (2002) Gramene, a tool for grass genomics. *Plant Physiol.* **130**, 1606–1613.

Wong, D.J., Liu, H., Ridky, T.W., Cassarino, D., Segal, E. and Chang, H.Y. (2008a) Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell*, **2**, 333–344.

Wong, D.J., Nuyten, D.S., Regev, A., Lin, M., Adler, A.S., Segal, E., van de Vijver, M.J. and Chang, H.Y. (2008b) Revealing targeted therapy for human cancer by gene module maps. *Cancer Res.* **68**, 369–378.

Xiang, C. and Oliver, D.J. (1998) Glutathione metabolic genes coordinately respond to heavy metals and jasmonic acid in Arabidopsis. *Plant Cell*, **10**, 1539–1550.

Xu, X., Liu, X., Ge, S. *et al.* (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111.

Yi, X., Du, Z. and Su, Z. (2013) PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res.* **41**, W98–W103.

Yoshida, S., Tamaoki, M., Ioki, M. *et al.* (2009) Ethylene and salicylic acid control glutathione biosynthesis in ozone-exposed *Arabidopsis thaliana*. *Physiol. Plant.* **136**, 284–298.

Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E. and Jiang, J. (2012) High-resolution mapping of open chromatin in the rice genome. *Genome Res.* **22**, 151–162.

Zhao, K., Tung, C.W., Eizenga, G.C. *et al.* (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* **2**, 467.