Published online 22 October 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D1483–D1491***
https://doi.org/10.1093/nar/gkac917

PlantExp: a platform for exploration of gene expression and alternative splicing based on public plant RNA-seq samples

Jinding Liu [1,2], Yaru Zhang[3], Yiqing Zheng[3], Yali Zhu[3], Yapin Shi[3], Zhuoran Guan[3], Kun Lang[3], Danyu Shen[4,*], Wen Huang [2,*] and Daolong Dou [1,4,*]

[1]Bioinformatics Center, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China, [2]Department of Animal Science, Michigan State University, East Lansing, MI 48824, USA, [3]College of Information Management, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China and [4]Department of Plant Pathology, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China

Received August 18, 2022; Revised September 20, 2022; Editorial Decision October 02, 2022; Accepted October 18, 2022

ABSTRACT

Over the last decade, RNA-seq has produced a massive amount of plant transcriptomic sequencing data deposited in public databases. Reanalysis of these public datasets can generate additional novel hypotheses not included in original studies. However, the large data volume and the requirement for specialized computational resources and expertise present a barrier for experimental biologists to explore public repositories. Here, we introduce PlantExp (https://biotec.njau.edu.cn/plantExp), a database platform for exploration of plant gene expression and alternative splicing profiles based on 131 423 uniformly processed publicly available RNA-seq samples from 85 species in 24 plant orders. In addition to two common retrieval accesses to gene expression and alternative splicing profiles by functional terms and sequence similarity, PlantExp is equipped with four online analysis tools, including differential expression analysis, specific expression analysis, co-expression network analysis and cross-species expression conservation analysis. With these online analysis tools, users can flexibly customize sample groups to reanalyze public RNA-seq datasets and obtain new insights. Furthermore, it offers a wide range of visualization tools to help users intuitively understand analysis results. In conclusion, PlantExp provides a valuable data resource and analysis platform for plant biologists to utilize public RNA-seq. datasets.

INTRODUCTION

High-throughput RNA sequencing (RNA-seq) has become a routine approach to exploring gene expression in a genome-wide manner. A massive amount of plant RNA-seq data across diverse tissues, developmental stages and experimental conditions are deposited in public archival repositories, such as the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) at NCBI (1), the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena) at EBI (2), the Sequence Read Archive (DRA, https://ddbj.nig.ac.jp/DRASearch) at DDBJ (3) and the Genome Sequence Archive (GSA, https://ngdc.cncb.ac.cn/gsa) at the BIG Data Center (4). Retrospective analyses of large collections of RNA-seq data can lead to new biological insights (5–7). For example, more than 10% of RNA-seq datasets from *Saccharomyces cerevisiae* are reanalyzed to generate new hypotheses regarding specific genes (8). These nucleotide archives are primarily archival repositories for storage of raw sequence reads. Without specialized computational resources and bioinformatics skills, experimental biologists cannot efficiently reuse these datasets.

The utilization of public RNA-seq datasets across diverse studies requires uniform processing, which can generate comparable gene expression data to obtain meaningful analysis results. In animals, >750 000 uniformly processed RNA-seq datasets from humans and mice are used to construct gene expression database for the research community to perform secondary analysis (9). MetazExp developed by our group includes 53 615 RNA-seq datasets from 72 metazoan species and offers differential and specific expression analysis modules (10). In plants, ePlant hosts data from 1385 samples of *Arabidopsis thaliana* for visual exploration of the spatial and temporal dynamics of gene expression (11). ARS pulled data from ∼20 000 public RNA-

*To whom correspondence should be addressed. Tel: +86 25 84396973; Email: ddou@njau.edu.cn
Correspondence may also be addressed to Wen Huang. Tel: +1 517 353 9136; Email: huangw53@msu.edu
Correspondence may also be addressed to Danyu Shen. Tel: +86 25 84396355; Email: shendanyu@njau.edu.cn

Table 1. Summary plant genome and RNA-seq samples

Order	Species	Database	Study	Experiment	Volume (GB)
Apiales	*Daucus carota*	Ensembl	2	29	124.8
Asterales	*Helianthus annuus*	Ensembl	34	1014	6279.9
	Lactuca sativa	RefSeq	30	542	2861.8
Capparales	*Arabidopsis halleri*	Ensembl	13	1267	1344.2
	Arabidopsis lyrata	Ensembl	20	214	813.6
	Arabidopsis thaliana	Ensembl	1742	32,061	103 014.7
	Brassica napus	Ensembl	181	3,948	23 011.5
	Brassica rapa	Ensembl	140	2,328	11 614
	Raphanus sativus	RefSeq	44	338	1030.3
Caryophyllales	*Beta vulgaris*	Ensembl	23	395	2402.7
	Spinacia oleracea	Ensembl	19	272	1076.4
Cucurbitales	*Citrullus lanatus*	Ensembl	47	630	3767.5
	Cucumis melo	RefSeq	36	662	2934.6
	Cucumis sativus	Ensembl	108	1,119	6199.5
	Momordica charantia	RefSeq	1	16	56
Euphorbiales	*Hevea brasiliensis*	Ensembl	229	290	2310.8
	Manihot esculenta	Ensembl	32	693	3656.6
	Ricinus communis	Ensembl	12	52	294.2
Fabales	*Arachis hypogaea*	RefSeq	73	1,032	6189.2
	Glycine max	Ensembl	428	4,028	20 102.1
	Glycine soja	RefSeq	27	397	1622.8
	Ipomoea triloba	Ensembl	3	29	107.2
	Medicago truncatula	Ensembl	71	1,844	8176.4
	Phaseolus vulgaris	Ensembl	44	757	3371.6
	Trifolium pratense	Ensembl	7	60	255.3
	Vigna angularis	Ensembl	8	66	340.9
	Vigna radiata	Ensembl	15	99	615
	Vigna unguiculata	RefSeq	2	29	156.4
Geraniales	*Linum usitatissimum*	JGI	28	401	1855.4
Juglandales	*Juglans regia*	Ensembl	12	162	1026.2
Lamiales	*Sesamum indicum*	Ensembl	18	389	1995.9
Liliales	*Dioscorea rotundata*	Ensembl	2	19	88.4
Malvales	*Corchorus capsularis*	Ensembl	4	11	174.3
	Gossypium arboreum	Ensembl	17	224	3785
	Gossypium hirsutum	Ensembl	154	2,526	18 169.1
	Gossypium raimondii	Ensembl	9	60	245.6
	Herrania umbratica	Ensembl	1	6	124.4
	Theobroma cacao	Ensembl	12	222	968.7
Marchantiales	*Marchantia polymorpha*	Ensembl	28	256	813.4
	Physcomitrium patens	Ensembl	120	711	3040.8
Nymphaeales	*Nelumbo nucifera*	RefSeq	18	119	803.4
Poales	*Brachypodium distachyon*	Ensembl	219	1,258	5836.1
	Hordeum vulgare	Ensembl	127	5,257	18 016
	Oryza barthii	Ensembl	2	7	91.6
	Oryza glaberrima	Ensembl	4	11	147.9
	Oryza longistaminata	Ensembl	4	34	174.7
	Oryza nivara	Ensembl	4	19	169.3
	Oryza punctata	Ensembl	1	3	46.2
	Oryza rufipogon	Ensembl	19	111	786.1
	Oryza sativa Indica Group	Ensembl	106	1,140	5443
	Oryza sativa Japonica Group	Ensembl	791	9,965	51 874.4
	Panicum hallii	Ensembl	30	406	4368.6
	Saccharum spontaneum	Ensembl	13	253	1641.9
	Setaria italica	Ensembl	118	444	2522.7
	Setaria viridis	JGI	80	339	1889.7
	Sorghum bicolor	Ensembl	275	2,090	7262.5
	Triticum aestivum	Ensembl	318	4,793	40 013.5
	Triticum urartu	Ensembl	6	72	450.5
	Zea mays	Ensembl	919	21,612	83 404.2
Principes	*Elaeis guineensis*	Ensembl	22	204	1054.6
	Jatropha curcas	Ensembl	21	130	832.4
Rhamnales	*Vitis vinifera*	Ensembl	172	4,182	14,767
	Ziziphus jujuba	RefSeq	18	252	1760.1
Rosales	*Malus domestica*	JGI	87	1,442	6244.5
	Prunus avium	Ensembl	21	266	1595.5
	Prunus persica	Ensembl	56	675	3682.1
	Pyrus x bretschneideri	RefSeq	22	198	1227.1
	Rosa chinensis	Ensembl	12	177	1251.5
Rubiales	*Coffea arabica*	Ensembl	13	226	583.1
Rutales	*Citrus clementina*	Ensembl	7	50	340
	Citrus sinensis	RefSeq	49	571	3657.6

Table 1. Continued

Order	Species	Database	Study	Experiment	Volume (GB)
	Olea europaea	RefSeq	24	325	1510.4
Salicales	*Populus deltoides*	JGI	25	1,009	5967.8
	Populus euphratica	Ensembl	10	61	944.4
	Populus trichocarpa	Ensembl	78	1,926	9775.2
	Salix purpurea	JGI	10	146	1309.3
Selaginellales	*Selaginella moellendorffii*	Ensembl	8	99	166
Solanales	*Capsicum annuum*	RefSeq	50	1,004	6221.9
	Ipomoea nil	RefSeq	5	27	52.4
	Nicotiana attenuata	Ensembl	6	67	320.2
	Nicotiana tabacum	RefSeq	50	330	2175.2
	Solanum lycopersicum	Ensembl	357	7,474	22 059.9
	Solanum pennellii	RefSeq	20	547	846.3
	Solanum tuberosum	Ensembl	94	1,660	8579.3
Volvocales	*Chlamydomonas reinhardtii*	Ensembl	83	1,244	4591.8
Sum			8170	131,423	572 475.1

seq samples of *A. thaliana* to visualize tissue, developmental stage and stress condition specificity of gene expression (12). PPRD uniformly processed ∼45 000 RNA-seq datasets from five important crops such as maize, rice, soybean, wheat and cotton to provide functions of gene expression retrieval and data mining (13). These plant gene expression databases are increasingly useful to investigate gene function and generate hypotheses. Alternative splicing is a universal regulation mechanism of post transcriptional gene expression in eukaryotes and plays vital roles in diverse biological procedures. Recently, PastDB was built to provide alternative splicing and gene expression quantifications of *A. thaliana* across tissues, developmental stages and environmental conditions (14). Other than PastDB, most plant gene expression databases do not consider alternative splicing. Furthermore, the conservation of orthologous gene expression patterns is important for the investigation of gene function. To our knowledge, no plant gene expression database so far supports cross-species gene expression conservation analysis.

Here we present PlantExp (Figure 1), a web-based retrieval and analysis platform that builds upon 131 423 publicly available RNA-seq samples from 85 plant species across 24 orders. It has four important features. First, it includes by far the largest number of plant RNA-seq samples. Second, it coveres both gene expression and alternative splicing profiles. Third, in addition to the database query and retrieval functions, it offers the multiple online analysis functions including differential expression analysis, special expression analysis, co-expression network analysis and cross-species gene expression conservation analysis. Fourth, A rich diversity of visualization tools help users intuitively understand analysis results.

MATERIALS AND METHODS

RNA-seq data collection

We collected 85 species across 24 plant orders in PlantExp including the model species *Arabidopsis,* and important crops such as maize, rice, soybean, wheat and so on (Table 1). The 3 biggest plant orders in the database, Poales, Fabales and Solanales included 18, 10 and 7 species, respectively. The reference genome assemblies and annotations of the 85 species were gathered from the Ensembl (15), RefSeq (16) and JGI (17) databases.

As for RNA-seq datasets, only sequencing data generated by the Illumina platform were considered because of its ubiquity and base-calling accuracy. High-throughput RNA-seq raw datasets and metadata were queried from Sequence Read Archive databases using the combined conditions of platform = 'Illumina', Source = 'transcriptomic' and Strategy = 'RNA-seq'. A total of 131 432 RNA-seq samples from 8303 studies containing 572.4 tera-bases were collected for construction of the PlantExp database (Table 1). As expected, the most represented species were two model organisms, *A. thaliana* (thale cress) and *Zea mays* (maize), possessing 32 344 and 21 794 samples, accounting for 24% and 16%, respectively.

To conveniently perform analysis on the collected datasets by customizing sample groups, we manually curated sample attributes focusing on cultivar, genotype, tissue, developmental stage and treatment (experimental conditions) based on information embedded in the abstract, description, and published studies.

Estimation of gene expression profiles

Fastp v0.23.0 (18) was used to trim and filter raw read sequences. Hisat v2.1.0 (19) was used for sequence alignment and data quality assessment. StringTie v2.1.4 (20) was used to estimate gene expression levels. We collected gene expression levels by two metrics, TPM (Transcripts Per Million) and FPKM (Fragments Per Kilobase of gene per Million mapped fragments). In addition, raw read count numbers were obtained for online expression analysis using a python script prepDE.py accompanying StringTie2.

Gene model refinement

Gene models were refined to obtain more alternative transcript annotation by a previously described procedure (10). The RNA-seq datasets with at least 80% unique mapping rate, at least 100 bp in read length and enough sequencing bases were used to refine gene models. The requirement of sequencing volume was adjusted with plant genome size and its RNA-seq data availability to achieve at least three

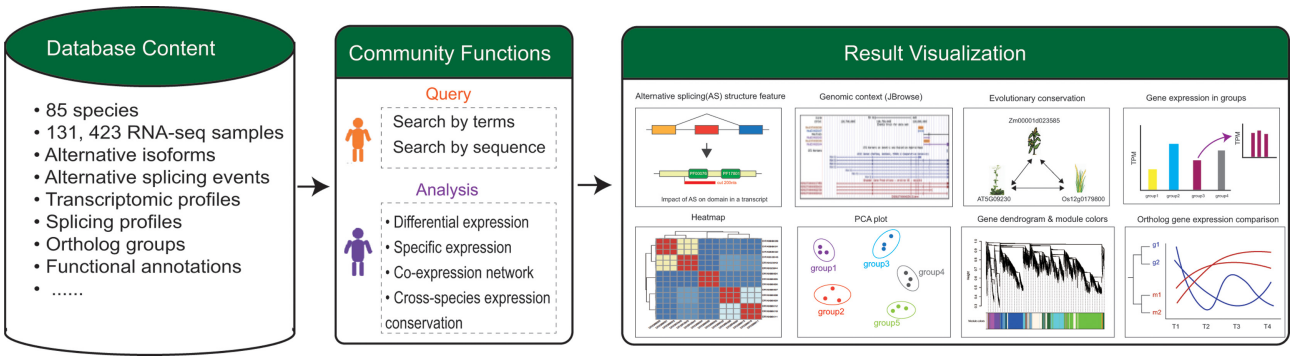


Figure 1. Summary of PlantExp contents and functions.

samples for each species. For example, at least 8G bases needed to be mapped in wheat, while at least 4G bases were needed in *Arabidopsis*. The read alignments of each RNA-seq sample were assembled into transcripts using StringTie2 with guidance of the reference gene model. For the transcriptome assembly of each RNA-seq sameple, only novel multi-exonic transcripts with at least 200 bp long, at least 2× coverage and 1× per exon for all exons were retained. Finally, the novel identified transcripts were filtered out when they didn't meet the following occurrence frequency in RNA-seq samples. The final novel transcripts must occur in at least three RNA-seq samples and account for more than a half of experiments of any tissue or at least one-third of all experiments.

After gene model refinement, for the 85 species the splice junctions and exons on average were increased by 8.47% and 16.93%, respectively (Supplementary Table S1). The proportion of multi-exonic genes with alternative transcripts increased from 20.99% to 41.91% on average (Supplementary Table S1). The alternative transcript number per multi-exonic gene also on average increased from 1.52 to 2.03 (Supplementary Table S1).

Identification and estimation of alternative splicing

The five classic alternative splicing events including alternative 5′ splice sites, skipped exon, mutually exclusive exons, retained intron, and alternative 3′ splice sites were identified with rMATS v 4.0.2 (21). As expected (22), the retained intron events generally were the most abundant type in RNA-seq samples, accounting for 49.26% on average. There were 91.45% of alternative splicing events whose exon-intron structures were consistent with alternative transcripts, and 8.55% of events whose exon-intron structures were inferred from read alignments on exons. To estimate alternative splicing profiles, the PSI (percentage spliced in) values were calculated by the JCEC and JC methods (21). In the former, event counts included both reads that span junctions (Junction Counts) and reads that do not cross an exon boundary (Exon Counts). In the latter, event counts included only reads that span junctions (Junction Counts).

Identification of ortholog genes and alternative splicing groups

We identified orthologous gene groups based on the longest protein sequences of genes using orthofinder (23). Ortholog alternative splicing groups indicate alternative splicing events that occur in ortholog genes with the same exon-intron splicing structures. We use the following procedure to identify ortholog alternative splicing groups. First, the protein sequences in orthologous genes were aligned globally using mafft (24). Then, the alignments of protein sequences were converted to codon alignments using pal2nal (25). Finally, we calculated the new coordinates of exons in transcripts corresponding to the longest protein based on codon alignments. Alternative splicing events with the same coordinates based on codon alignments were classified into an orthologous group. In the 85 plants, we identified 62 897 ortholog gene groups. Based on these ortholog gene groups, we identified 225 441 putative ortholog alternative splicing groups.

Gene functional annotation

To support retrieval by gene functional terms, both gene ontology and Pfam domain annotation in Ensembl and JGI database were integrated into PlantExp. For the genomes from RefSeq genomes, Blast2GO (26) and Interproscan (27,28) were used to obtain GO and Pfam domain annotation. As for pathway annotation, protein sequences were submitted to KAAS (KEGG automatic annotation server) to obtain KO (KEGG Orthology) and KEGG pathway annotation (29). Furthermore, transcript sequences were submitted to psRNATarget (30) to predict microRNA targets.

Online analysis modules

PlantExp includes four online analysis modules of differential expression analysis, special expression analysis, co-expression network analysis and cross-species expression conservation analysis. DESeq2 (31) and edgeR (32) were used to compare overall gene expression. The statistical models implemented in rMATS(21) were used to detect differentially spliced genes. WGCNA (33) was used for weighted gene co-expression network analysis. PlantExp also contained a flexible enrichment analysis procedure based on the R package ClusterProfiler (34), including the hypergeometric test and the Gene Set Enrichment Analysis (GSEA) (35). Phylip v3.696 (36) was used to build molecular phylogenetic tree for ortholog genes. To compare ortholog gene expression profiles, the gene expression levels were log2 transformed ratios of a gene expression in a sample divided by trimmed mean expression level.

Figure 2. Database retrieval and result visualization. (**A**) The retrieved genes by terms or sequence are listed in an interactive table with links for users to open the gene page showing expression data. (**B**) The gene, transcript, splicing and interaction pages are interconnected by the inner links in pages. The gene, transcript and splicing page have similar layout, including basic annotation information (**C**), visualized genomic context (**D**) and expression/splicing profiles (**E**). (**F**) The visual effect of an alternative splicing event on a transcript is shown in the interaction page. The alternative splicing removes 72nts from the transcript.

Database usage

Web interface. PlantExp is hosted at https://biotec.njau.edu.cn/plantExp. At the top of the portal page, users can open the help and FAQ page to learn the full instructions and frequently asked questions. The body of the portal page is to introduce database contents and provide entrances to species. For each species, the data accesses can be divided into three groups. (i) The summary page provides statistics and links to download gene expression and alternative splicing data. (ii) The search and blast page offer access points to retrieve gene expression and alternative splicing data by gene terms and sequence similarity, respectively. (iii) The comparison, specificity, co-expression and cross-species pages provide users with access to analysis of the collected RNA-seq datasets.

Querying the database. Users can search the database for gene expression and alternative splicing profiles in the search page by gene ID, symbol, Pfam and pathway terms,

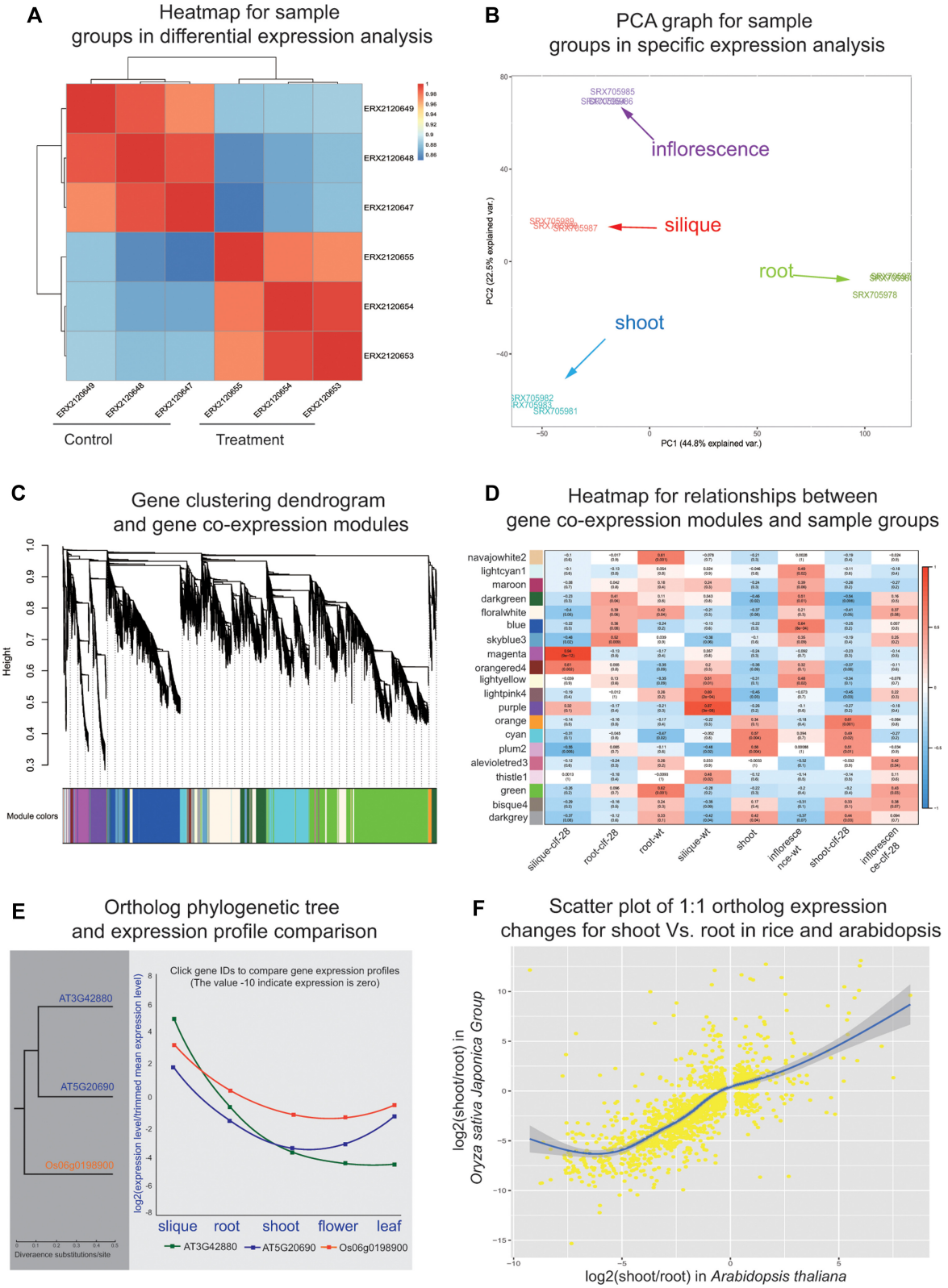
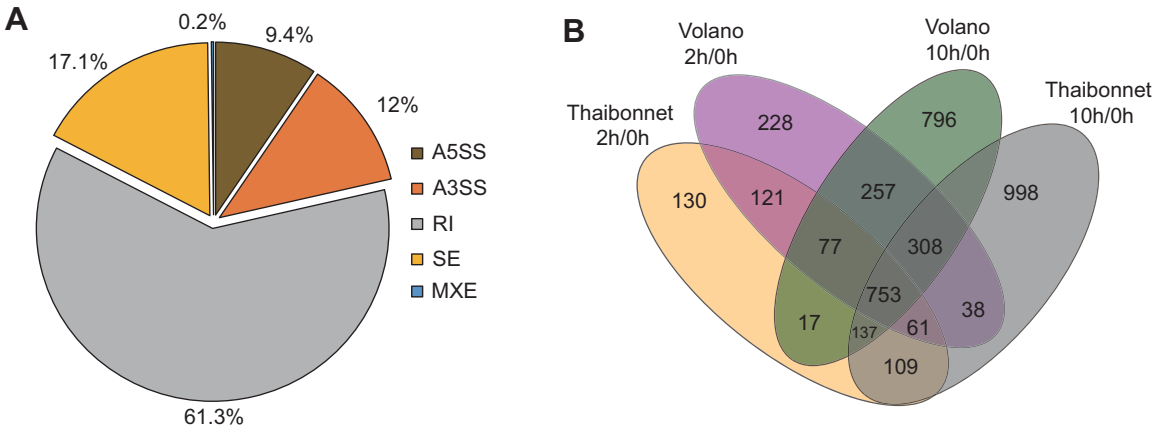


Figure 3. Examples of result visualization generated by online analysis tools. (**A**) Sample clustering heatmap based on gene expression profiles in differential expression analysis. (**B**) PCA plot of overall gene expression for samples covering four tissues in specific expression analysis. (**C**) Co-expression network analysis tool generates a gene clustering dendrogram corresponding to co-expression modules. (**D**) Heatmap relationship between gene co-expression modules and sample groups. (**E**) Cross-species expression conservation analysis tool generates a phylogenetic tree of 3 orthologs from rice and arabidopsis, and provides an interactive box to compare their expression profiles covering five tissues. The gene expression levels are log2 transformed ratios of the gene expression in a tissue divided by trimmed mean expression level. (**F**) Scatter plot of 1:1 ortholog gene expression change ratios for shoot Vs. root in rice and arabidopsis. The gene expression fold changes are log2 transformed values.



C

GO terms and pathways enriched on differentially spliced gens

		Thaibonnet 2h/0h	Thaibonnet 10h/0h	Volano 2h/0h	Volano 10h/0h
GO terms	Protein serine/threonline kinase activity	※		※	※
	Protein kinase activity	※			
	Nucleic acid binding	※	※	※	※
	Regulation of alternative mRNA splicing, via spliceosome	※		※	
	U2-type prespliceosome		※		
	mRNA binding		※		
	Nuclears speck				※
	mRNA cis splicing, via spliceosome				※
Pathways	Spliceosome	※	※	※	※
	Circadian rhythm - plant		※	※	※
	Inositol phosphate metabolism			※	
	Plant hormone signal transduction			※	
	Phosphatidylinositol signaling system			※	

Figure 4. A case of exploration alternative splicing in cold stress of rice. (**A**) Proportions of diverse DAS events in cold stress of rice. (**B**) The Venn diagram shows the number of DAS events shared among the 4 comparisons. (**C**) The enriched GO and pathway terms in the four comparisons.

or in the blast page by gene nucleic acid or amino acid sequence. The retrieved genes are listed in an interactive table with links for users to open the gene page to show gene annotation and expression profiles (Figure 2A). Through the inner links in the gene page, users can open the transcript page to show an associated transcript's expression profiles, as well as open the splicing page to show an associated alternative splicing event's profiles. Furthermore, by the inner links in the transcript or splicing page, users can open an interaction page to show the effect of an alternative splicing event on an associated transcript. The access relationships among the gene, transcript, splicing and interaction page are shown in Figure 2B.

The gene, transcript and splicing pages have similar page layout. First, the genomic position and relevant annotation are listed at the top (Figure 2C). Through the links binding with annotation terms, users can quickly jump to external databases, such as the Pfam, KEGG, AmiGO and miRbase (37) database. Most remarkably, users can explore ortholog gene expression and alternative splicing profiles in other species by the ortholog group link. The following section is a genome browser for users to explore gene, transcript and alternative splicing structure in their genomic context (Figure 2D). Then, for the gene page, there are two tables showing its associated transcripts and alternative splicing events. For the transcript page (or the splicing page), there is a table listing its associated alternative splicing events (or transcripts). Finally, a drop-down box is employed to list all collected studies and an interactive and hierarchical bar chart is used to show the expression or splicing profiles in a chosen study (Figure 2E). In addition, the effect of alternative splicing on a transcript, such as impact of protein domain and targets of microRNAs, can be visually shown in the interaction page (Figure 2F).

Online analysis. PlantExp is equipped with four online analysis tools to assist users to analyse RNA-seq datasets collected in the database (see details in Materials and Methods). The four task submission pages has similar layout. First, the current status of the analysis server is shown at the top of four task submission pages including the numbers of running and waiting tasks. The following section is a sample retrieval box. In this box, users can only load samples of interests by setting query conditions, such as RNA-seq layout, read length, data volume and data source. The retrieved samples with the information of cultivar, genotype,

tissue and so on, are presented in an interactive table. According to sample information, users can flexibly set sample groups for an analysis task. The next is an area for users to set analysis methods and parameters. Finally, at the bottom users assign job name to an analysis task and provide an email to receive notifications and results.

For a differential or specific expression analysis task, the result page first provides a heatmap (Figure 3A) and principal component analysis (PCA) graph (Figure 3B) to illustrate sample clustering based on overall gene expression and splicing profiles. Then, the differentially and specifically expressed/spliced genes are listed in interactive tables. In addition to gene information, expression/splicing change values and significance levels, the tables provide links for users to further open new pages showing gene details and expression/splicing profiles in selected samples. Finally, bar charts are presented to exhibit GO and pathway enrichment analysis on differentially and specifically expressed/spliced genes.

For a co-expression network (WGCNA) analysis task, the result page provides diverse graphs or tables to display analysis results. First, a dendrogram exhibits the clustering relationships of samples based on overall gene expression profiles. According the dendrogram, users can detect whether outlier samples exist. The identified gene co-expression modules (networks) are listed in an interactive table with links for users to open new pages showing function enrichment bar charts and visual networks. To help users understand the co-expressed genes, PlantExp generates a dendrogram of gene clustered using dissimilarity measure based on topological overlap matrix (Figure 3C). Finally, a heatmap is used to characterize relationships between sample groups and gene co-expression modules (Figure 3D).

For cross-species expression conservation analysis, users can customize sample groups with similar attributes in any specified two species to explore gene expression conservation. In the returned result page, an interactive table is used to list the consistency of ortholog gene expression intra- and inter-species. By clicking on the link icons in the table, users can open a new page to view diagram of curves representing gene expression profiles covering multiple sample groups, and an evolutionary tree exhibiting molecular phylogenetic relationships based on ortholog protein sequences (Figure 3E). The 1:1 ortholog groups are more conserved because of the importance for exploration of species phylogeny. PlantExp presents a table showing 1:1 ortholog genes differentially expressed in all two-group comparisons. Furthermore, scatter plots are used to show gene expression ratios of 1:1 ortholog gene pairs (Figure 3F).

A case study to explore alternative splicing induced by cold stress. To illustrate the power of PlantExp, we present here a case study to explore alternative splicing induced by cold stress in rice. Two different rice cultivars, Thaibonnet and Volano, are respectively sensitive and tolerant to cold stress, and the gene expression level changes at 0, 2 and 10 h at 10°C cold stress have been explored (PRJEB22031) (38). After running comparative analysis of selected SRA samples representing these conditions, we retrieved the result page of the analysis. From the heatmaps, the samples in four comparisons (Thaibonnet 2 h/0 h and 10 h/0 h; Volano 2 h/0 h and 10 h/0 h) were well clustered into control and treatment groups based on both gene expression levels (Supplemental Figure S1A) and alternative splicing profiles (Supplemental Figure S1B). This implied that both overall gene expression and alternative splicing profiles were altered by cold stress.

MATS_LRT with default parameters was used to detect gene differentially splicing. A total of 4713 differentially alternative splicing (DAS) events were identified after cold stress in the two rice cultivars (Figure 4B). In the identified DAS events, the retained intron was the most abundant alternative splicing type, accounting for 61.3% (Figure 4A). After 2 h of cold stress, there were a total of 2236 DAS events detected in two cultivars, in which 393 DAS specifically occurred in the sensitive cultivar Thaibonnet and 831 DAS specifically occurred in the tolerant cultivar Volano. After 10 h cold stress, the DAS events were almost increased to twice, reaching 4234. There were respectively 1147 and 891 DAS events specifically occurring in the Thaibonnet and Volano cultivar. In addition, The GO and pathway enrichment analysis revealed functional commonalities as well as differences in the differentially spliced genes of the two cultivars after 2 and 10 h cold stress (Figure 4C). These findings about alternative splicing, not reported by the original study (38) that generated the RNA-Seq data, may generate new testable hypotheses regarding genes involved in cold stress of rice.

CONCLUSIONS

In summary, we have constructed the most comprehensive plant gene expression and alternative splicing database. The diverse retrieval, analysis and visualization functions make it a one-stop resource for botanists to explore large public RNA-seq datasets. As for future directions, the database will be continuously updated (12–18 months) when more RNA-seq and other types (e.g. RNA editing, small RNAs) data become available.

DATA AVAILABILITY

All data are available at: https://biotec.njau.edu.cn/plantExp.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge the works of all the genome and RNA-seq data producers.

FUNDING

National Natural Science Foundation of China [32230089 to D.D., 32270208 to J.L., 32070139 to D.S.]; Technical System of Chinese Herbal Medicine Industry [CARS-21 to D.D.]; Jiangsu Agricultural Science and Technology Innovation Fund [CX(21)3085 to D.S.]; Innovative Experimental Program for College Students [202110307064 to Y.Z.];

Michigan State University (to W.H.); MSU AgBioResearch USDA Hatch project [MICL02560 to W.H.]. Funding for open access charge: National Natural Science Foundation of China [32230089, 32230089, 32070139].
Conflict of interest statement. None declared.

REFERENCES

1. Katz,K., Shutov,O., Lapoint,R., Kimelman,M., Brister,J.R. and O'Sullivan,C. (2022) The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res.*, **50**, D387–D390.
2. Cummins,C., Ahamed,A., Aslam,R., Burgin,J., Devraj,R., Edbali,O., Gupta,D., Harrison,P.W., Haseeb,M., Holt,S. *et al.* (2022) The european nucleotide archive in 2021. *Nucleic Acids Res.*, **50**, D106–D110.
3. Okido,T., Kodama,Y., Mashima,J., Kosuge,T., Fujisawa,T. and Ogasawara,O. (2022) DNA data bank of japan (DDBJ) update report 2021. *Nucleic Acids Res.*, **50**, D102–D105.
4. Members,C.-N. and PartnersPartners. (2021) Database resources of the national genomics data center, china national center for bioinformation in 2021. *Nucleic Acids Res.*, **49**, D18–D28.
5. Collado-Torres,L., Nellore,A., Kammers,K., Ellis,S.E., Taub,M.A., Hansen,K.D., Jaffe,A.E., Langmead,B. and Leek,J.T. (2017) Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.*, **35**, 319–321.
6. Lachmann,A., Torre,D., Keenan,A.B., Jagodnik,K.M., Lee,H.J., Wang,L., Silverstein,M.C. and Ma'ayan,A. (2018) Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.*, **9**, 1366.
7. Moreno,P., Fexova,S., George,N., Manning,J.R., Miao,Z.C., Mohammed,S., Munoz-Pomer,A., Fullgrabe,A., Bi,Y.L., Bush,N. *et al.* (2022) Expression atlas update: gene and protein expression in multiple species. *Nucleic Acids Res.*, **50**, D129–D140.
8. Doughty,T. and Kerkhoven,E. (2020) Extracting novel hypotheses and findings from RNA-seq data. *FEMS Yeast Res.*, **20**, foaa007.
9. Wilks,C., Zheng,S.C., Chen,F.Y., Charles,R., Solomon,B., Ling,J.P., Imada,E.L., Zhang,D., Joseph,L., Leek,J.T. *et al.* (2021) recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.*, **22**, 323.
10. Liu,J., Yin,F., Lang,K., Jie,W., Tan,S., Duan,R., Huang,S. and Huang,W. (2022) MetazExp: a database for gene expression and alternative splicing profiles and their analyses based on 53 615 public RNA-seq samples in 72 metazoan species. *Nucleic Acids Res.*, **50**, D1046–D1054.
11. Waese,J., Fan,J., Pasha,A., Yu,H., Fucile,G., Shi,R., Cumming,M., Kelley,L.A., Sternberg,M.J., Krishnakumar,V. *et al.* (2017) ePlant: visualizing and exploring multiple levels of data for hypothesis generation in plant biology. *Plant Cell*, **29**, 1806–1821.
12. Zhang,H., Zhang,F., Yu,Y., Feng,L., Jia,J., Liu,B., Li,B., Guo,H. and Zhai,J. (2020) A comprehensive online database for exploring approximately 20,000 public arabidopsis RNA-Seq libraries. *Mol. Plant.*, **13**, 1231–1233.
13. Yu,Y., Zhang,H., Long,Y., Shu,Y. and Zhai,J. (2022) Plant public RNA-seq database: a comprehensive online database for expression analysis of ∼45 000 plant public RNA-Seq libraries. *Plant Biotechnol. J.*, **20**, 806–808.
14. Martin,G., Marquez,Y., Mantica,F., Duque,P. and Irimia,M. (2021) Alternative splicing landscapes in *Arabidopsis thaliana* across tissues and stress conditions highlight major functional differences with animals. *Genome Biol.*, **22**, 35.
15. Bolser,D., Staines,D.M., Pritchard,E. and Kersey,P. (2016) Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. *Methods Mol. Biol.*, **1374**, 115–140.
16. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
17. Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
18. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
19. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
20. Kovaka,S., Zimin,A.V., Pertea,G.M., Razaghi,R., Salzberg,S.L. and Pertea,M. (2019) Transcriptome assembly from long-read RNA-seq alignments with stringtie2. *Genome Biol.*, **20**, 278.
21. Shen,S., Park,J.W., Lu,Z.X., Lin,L., Henry,M.D., Wu,Y.N., Zhou,Q. and Xing,Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5593–E5601.
22. Syed,N.H., Kalyna,M., Marquez,Y., Barta,A. and Brown,J.W. (2012) Alternative splicing in plants–coming of age. *Trends Plant Sci.*, **17**, 616–623.
23. Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.
24. Nakamura,T., Yamada,K.D., Tomii,K. and Katoh,K. (2018) Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics*, **34**, 2490–2492.
25. Suyama,M., Torrents,D. and Bork,P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609–W612.
26. Conesa,A. and Gotz,S. (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics.*, **2008**, 619832.
27. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
28. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
29. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
30. Dai,X., Zhuang,Z. and Zhao,P.X. (2018) psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.*, **46**, W49–W54.
31. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
32. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
33. Langfelder,P. and Horvath,S. (2008) WGCNA: an r package for weighted correlation network analysis. *BMC Bioinf.*, **9**, 559.
34. Yu,G., Wang,L.G., Yan,G.R. and He,Q.Y. (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, **31**, 608–609.
35. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
36. Shimada,M.K. and Nishida,T. (2017) A modification of the PHYLIP program: a solution for the redundant cluster problem, and an implementation of an automatic bootstrapping on trees inferred from original data. *Mol. Phylogenet. Evol.*, **109**, 409–414.
37. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
38. Buti,M., Pasquariello,M., Ronga,D., Milc,J.A., Pecchioni,N., Ho,V.T., Pucciariello,C., Perata,P. and Francia,E. (2018) Transcriptome profiling of short-term response to chilling stress in tolerant and sensitive oryza sativa ssp japonica seedlings. *Funct. Integr. Genomic.*, **18**, 627–644.