DNA Research, 2017, 24(4), 371–375
doi: 10.1093/dnares/dsx009
Advance Access Publication Date: 13 March 2017
Full Paper

Full Paper

EXPath tool—a system for comprehensively analyzing regulatory pathways and coexpression networks from high-throughput transcriptome data

Han-Qin Zheng, Nai-Yun Wu, Chi-Nga Chow, Kuan-Chieh Tseng, Chia-Hung Chien, Yu-Cheng Hung, Guan-Zhen Li, and Wen-Chi Chang*

Institute of Tropical Plant Sciences, College of Biosciences and Biotechnology, National Cheng Kung University, Tainan 701, Taiwan

*To whom correspondence should be addressed. Tel. +886 6 2757575x57322. Email: sarah321@mail.ncku.edu.tw

Edited by Prof. Hiroyuki Toh

Received 13 November 2016; Editorial decision 6 February 2017; Accepted 23 February 2017

Abstract

Next generation sequencing (NGS) has become the mainstream approach for monitoring gene expression levels in parallel with various experimental treatments. Unfortunately, there is no systematical webserver to comprehensively perform further analysis based on the huge amount of preliminary data that is obtained after finishing the process of gene annotation. Therefore, a user-friendly and effective system is required to mine important genes and regulatory pathways under specific conditions from high-throughput transcriptome data. EXPath Tool (available at: http://expathtool.itps.ncku.edu.tw/) was developed for the pathway annotation and comparative analysis of user-customized gene expression profiles derived from microarray or NGS platforms under various conditions to infer metabolic pathways for all organisms in the KEGG database. EXPath Tool contains several functions: access the gene expression patterns and the candidates of co-expression genes; dissect differentially expressed genes (DEGs) between two conditions (DEGs search), functional grouping with pathway and GO (Pathway/GO enrichment analysis), and correlation networks (co-expression analysis), and view the expression patterns of genes involved in specific pathways to infer the effects of the treatment. Additionally, the effectively of EXPath Tool has been performed by a case study on IAA-responsive genes. The results demonstrated that critical hub genes under IAA treatment could be efficiently identified.

Key words: NGS, biological pathway, co-expression network

1. Introduction

In recent years, with the development of next-generation sequencing (NGS) and single molecule real time sequencing (SMRT sequencing), the cost per megabase (Mb; a million bases) has been decreasing.[1] In particular, transcriptomic sequencing has advanced the study of gene regulatory pathways in non-model organisms. In general, assembly, annotation, gene expression calculation and systematic function analysis are the four major processes after transcriptome sequencing has been finished by biotech companies. At the step of assembly, the raw reads with low quality were removed before assembling to transcripts (contigs). Numerous software for sequencing assembly have been developed such as Trinity,[2] Oases,[3] SOAPdenovo-Trans,[4] Trans-Abyss[5] and Bridger.[6] A Linux environment is required for all of these tools. After transcripts have been assembled, the annotation is critical to define their potential function. Several tools could be applied for annotation, such as Blast2GO,[7] BLAST,[8] Trinotate,[2] MAPLE[9] and TRAPID.[10] As well as, many database can be utilized to annotate the transcripts with regard to NCBI non-redundancy proteins (NCBI NR), UniProt Swiss-Prot/TrEMBL,[11] Gene Ontology (GO)[12] and KEGG pathways.[13] Following the annotation, it is important to identify the differentially expressed transcripts when comparing different samples. Therefore, the expression value of each transcript (contig) were calculated and normalized by RPKM (reads per million per kilo base), FPKM (fragments per million per kilo base) and TPM (transcript per million) depends on sequencing methods (i.e. single end reads or paired end reads).

Finally, systematic function analysis is the most important, but trouble issue for biologists. A useful bioinformatics resource is crucial to retrieve valuable information and experimental candidates from numerous transcripts. Recently, several open source software packages (e.g. Bioconductor (based on R script)[14] and Graphical User Interface (GUI)) and commercial software (e.g. Ingenuity IPA, CLC Workbench) have been developed for functional analysis of high-throughput transcriptomic data. However, the use of commercial software might be restricted by research budgets, while open source software might be difficult for biologists to use if they do not have a programming background. Although R script is very powerful to analyze various biological data, a series courses of R language is needed before using those open resource. For some biologists, this might be a slow remedy that cannot meet an urgency. Consequently, the processing of these huge sequencing data usually becomes a bottleneck for biologists, due to the requirement of the efficient hardware and ability of programming language.

A database named EXPath[15] was developed by our group, which contains the experimental microarray data of *Arabidopsis thaliana*, *Oryza sativa* and *Zea mays*, and provides comparative expression analysis inferring metabolic pathways. It is very useful with regard to dissecting the collected microarray datasets in terms of the pathways, differential expression, co-expression and pathway/GO enrichment analysis. Unfortunately, such a valuable database cannot be used for other high-throughput expression data, and at present, the system is limited to the datasets collected in the EXPath database. In order to provide more diverse transcriptome data, EXPath Tool is developed in this work to comprehensively analyze users' transcriptome datasets. Several analysis functions were not only advanced but also added in EXPath Tool. The online interface of EXPath Tool is especially designed for biologists, and after uploading an annotation and expression profile, five main functions (Gene Search, Pathway Search, DEGs Search, Pathways/GO Enrichment and Co-expression analysis) are provided for users to dissect their data from multiple perspectives. The concept and construction of EXPath Tool is illustrated in Supplementary Fig. S1.

2. Materials and methods

2.1. Data preprocess

Due to the heavy server loading of sequence assembly, annotation and calculation of transcript expression values, several data preprocess should be finished by users before upload their data to the EXPath Tool. Usually, these processes are a part of service of sequencing company. However, the step-by-step processing flow and required tools for data preprocess were also suggested in our system. Additionally, numerous protein sequences from NCBI RefSeq and UniprotKB for the annotation of different organisms are provided in EXPath Tool. After data preprocess, the files of annotation and expression profiles could be uploaded for further analysis in the EXPath Tool. The system flow of data preprocess is shown in Supplementary Fig. S2. It briefly describes the analysis steps for sequence assembly, annotation and expression data calculation. The details about the utility of those tools and requirements are provided in the guide page of the EXPath Tool (http://expathtool.itps.ncku.edu.tw/guide_for_custom/guide/create_project_preprocess.php).

2.2. Processing of uploaded dataset in the EXPath Tool

After users create a personal account, they can login into the EXPath Tool system and upload a set of annotation and expression profiles (call a 'Project'). In the EXPath Tool system, the annotation files of the pathway and Gene Ontology (GO) are required, as well as the gene (transcript) expression data. More additional annotations from other databases, such as Pfam[16] and PANTHER[17] are encouraged. After examination of the format of the uploaded files, all data are processed individually. Figure 1 illustrates the data processing flow of uploaded files. For the expression profiles, the samples are grouped into conditions by the column name and the average expression values of each gene (transcript) under each condition are calculated. The calculation results are applied for the function of differentially expressed genes (DEGs) search (DEGs search) and co-expression analysis. For the annotation files of the pathways, if there is no assigned KEGG ortholog (KO) ID, the system automatically converts the annotated ID (i.e. NCBI GI, NCBI RefSeq, UniprotKB accession number etc.) to KO ID using the UniProt ID mapping table. The KO ID could be used to connect the KEGG pathway map, and also applied to gene annotation, pathway searches and pathway enrichment analysis. Finally, the GO term file is applied to gene annotation and GO enrichment analysis.

2.3. Algorithm and construction

2.3.1. Pathway search

A pathway describes a series links of chemical reactions. Several critical pathways are affected by environmental changes due to various expression values of numerous genes (transcripts) under different conditions. In 'Pathway Search', after setting a specific pathway and the conditions of interest, the annotated pathway map will be displayed (Fig. 2a). In addition, the annotation and expression level of the KEGG ortholog can be accessed by clicking the KEGG ortholog block (Fig. 2b). The expression of the KEGG ortholog is calculated by simply averaging the expression levels of the associated genes in each condition. Furthermore, the absolute and relative expression values of various transcripts of the KEGG ortholog are graphically illustrated (Fig. 2c and d).

2.3.2. DEGs search

To determine genes that are differentially expressed under distinct conditions, the *t*-test statistical method was applied by using function t.test() of the R package in EXPath Tool. Users can specify the treatment and control, set the cut-off values of fold change and *P* value, and the DEGs results will then be determined.

2.3.3. Co-expression analysis

Co-expressed genes are a group of genes with similar expression profiles under specific conditions. Theoretically, they may be controlled

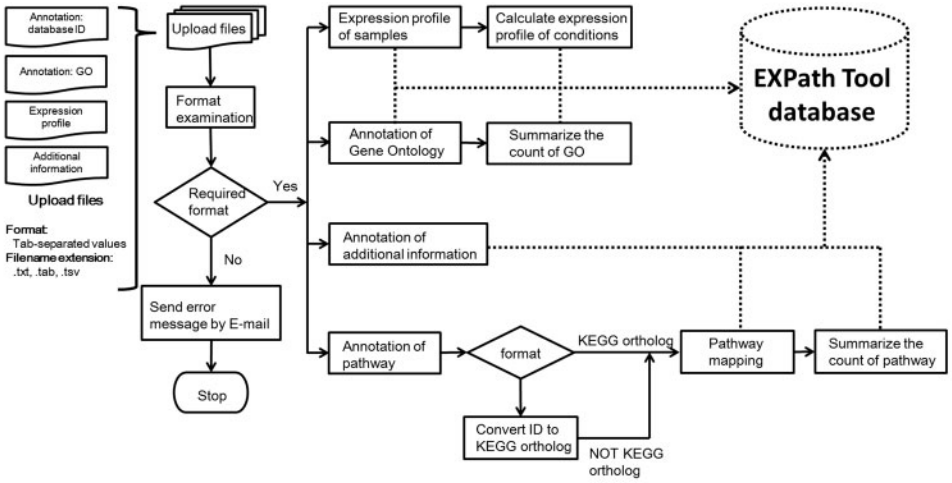

Figure 1. The analysis flowchart of data processing in EXPath Tool system.

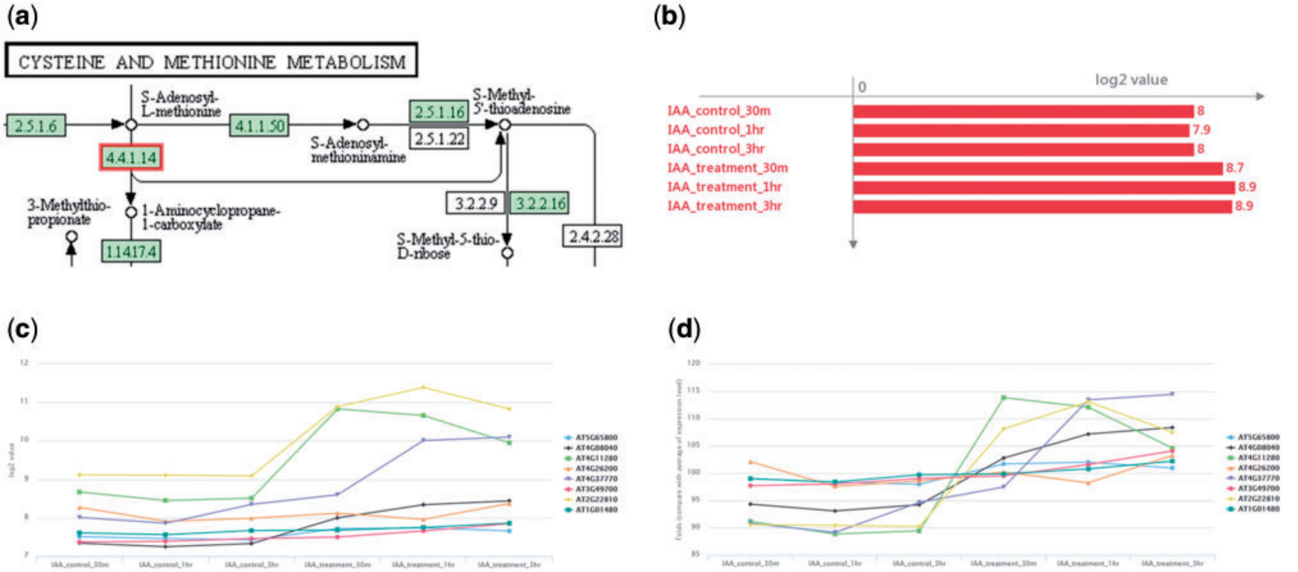

Figure 2. The web interface of pathway search result. (a) The genes annotated with the KEGG enzyme, the KEGG enzyme block is filled with green rectangle. A user selected gene is marked with red rectangle. The average (b), absolute (c) and relative (d) expression level of KEGG orthologs (including every transcripts) will be displayed when user clicks the green rectangle in figure (a).

by specific transcription factors and involved in identical pathways. To determine a group of genes which are co-expressed, Pearson's correlation coefficient and Spearman's rank methods are applied by using function cor() of the R package in EXPath Tool. There are two analysis modes in the co-expression analysis:

(A) Correlation of a single gene

This function is identifying positive/negative correlation gene groups of a query gene. The expression pattern of co-expression gene groups is illustrated based on the z-score transformation (Supplementary Fig. S3):

$$\mathrm{z} = \frac{x - \mu}{\sigma}$$

z denotes the *z*-score value, while x, μ and σ represent the raw intensity, mean and standard deviation of the gene expression levels, respectively.

(B) Correlation network

This function is identifying the correlation network of a group of query genes. The algorithm of the correlation coefficient and cut-off value should be set first. The correlation networks are constructed by Cytoscape Web[18] based on the setting parameters (Supplementary Fig. S4).

2.3.4. KEGG pathway and GO enrichment analysis

Enrichment analysis helps researchers to find the important pathways or functional GO terms from a group of query genes. To determine the significant pathways or GO terms, the cumulative probability (p-value) of the hypergeometric distribution, which is applied by dhyper() and phyper() of the R package, are calculated to evaluate the KEGG/GO enrichment of a group of input genes. The formula is as follows:

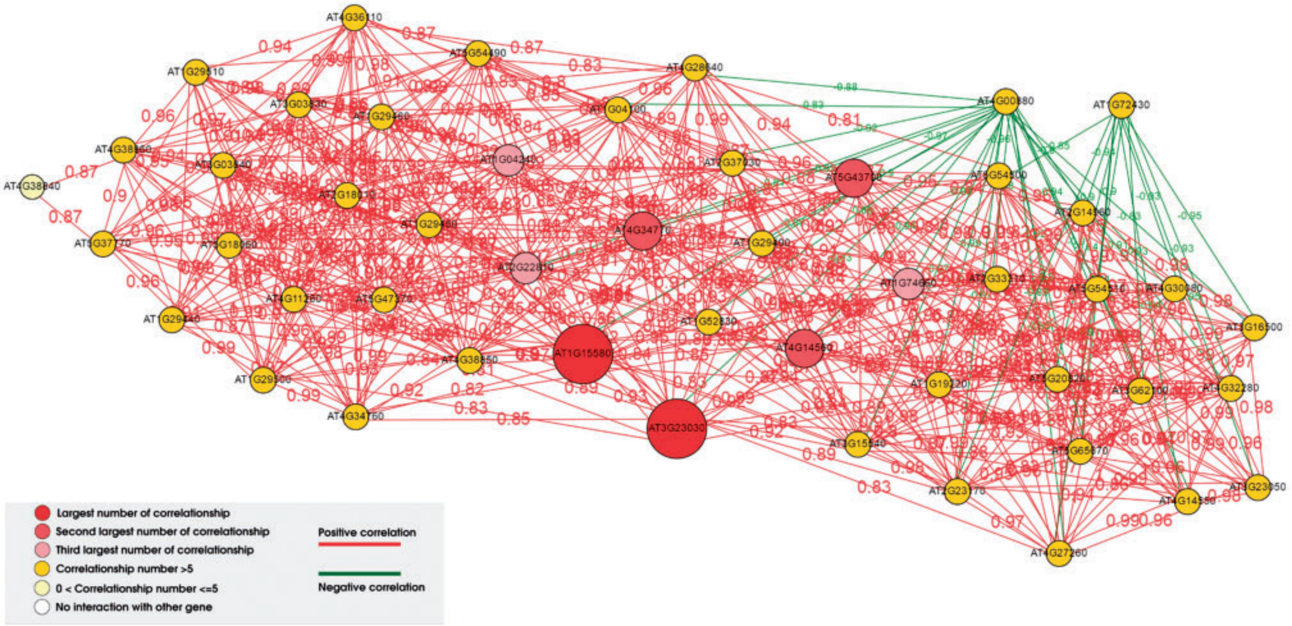
Largest number of correlationship
Second largest number of correlationship
Third largest number of correlationship
Correlationship number >5
0 < Correlationship number <=5
No interaction with other gene
Positive correlation
Negative correlation
AT1G15580
AT3G23030
AT1G72430
AT4G00880

Figure 3. The correlation networks of IAA-responsive genes. Two down-regulated genes (AT1G72430 and AT4G00880) show the negative co-expressions with other genes.

$$p(X \leq k) = \sum_{i=x}^{n} \frac{\binom{M}{i}\binom{N - M}{n - i}}{\binom{N}{n}}$$

N and *M* denote the number of background genes and total genes involved in specific KEGG pathways or GO terms, respectively. *i* is the number of input genes (defined as *n*) which belong to those KEGG pathways or GO terms. An example of the results is presented in Supplementary Fig. S5 in the Supplementary data.

2.3.5. Project management

After logging into EXPath Tool, users can import their data via ‘Create New Project’, and all information can be modified via project management. In order to carry out different sets of experiments, numerous projects are allowed in EXPath Tool. After finishing all the requirements for the system, the functions of Gene Search, Pathway Search, DEGs Search, Pathways/GO Enrichment and Co-expression analysis can be carried out by clicking ‘ENTER PROJECT’. It is easy to access differentially expressed genes (transcripts), critical pathways, and coexpression networks under a specific treatment by using EXPath Tool. The detailed introduction and guidelines for each function can be retrieved from the main page.

3. Results and discussion

3.1. A case study: co-expression analysis of IAA-responsive genes

EXPath tool provides a platform to assess co-expressions among a set of genes or transcripts by using users’ expression data. EXPath tool offers not only the visualization of correlation networks but also the enriched GO/Pathway of co-expressed genes. The following case study demonstrates the application of EXPath tool.

As is well known, auxins are important plant hormones involved in a wide range of cellular processes such as hypocotyl elongation, organ development and gravitropism. Nemhauser et al.[19] revealed that 51 Arabidopsis genes were vital IAA-responsive genes by using Affymetrix GeneChips as gene expression measurement (Supplementary Table S1). Among these genes, 49 genes showed significant up-regulation under IAA treatments. The other two genes (AT1G72430 and AT4G00880) were down-regulated. All genes were related to signaling and transcription. There is a series of microarray expression data (GSE39384) retrieved from AtGeneExpress were uploaded to demonstrate the usage of EXPath Tool.[20] The samples included three time points of IAA and mock treatment. First, 51 genes were utilized to construct correlation networks. Figure 3 demonstrates that AT1G72430 and AT4G00880 were identified the negative co-expressions (marked in green), which consisted with the expression profile in the previous study. Moreover, two IAA genes, AT1G15580 (IAA5) and AT3G23030 (IAA2) showed the highest connection with other IAA-responsive genes. Afterwards, all these genes were sent to the pathways/GO enrichment analysis. As expected, the most significant pathways and GO terms were plant hormone signal transduction, response to auxin, and auxin-activated signaling pathway (Supplementary Tables S2 and S3). This case described above reveals that EXPath Tool is an effective webserver for users to explore an important regulatory mechanism from their own high-throughput transcriptome data.

3.2.The utility and significance of EXPath Tool

There are several useful webservers or software packages that have been developed to help biologists to systematically analyze their high-throughput transcriptome data. Supplementary Table S4 compares the EXPath Tool and other similar resources. BioWardrob[21] and T-REx[22] conduct differentially expressed analysis and co-expression analysis according to the expression levels of genes. BioWardrob can conduct additional analyses (i.e., enrichment analysis) by executing R-scripts. However, code-editing is required for this, and R-script is usually difficult for a biologist with only wet lab experience. S-MART[23] focuses on the detection of alternative splicing and transcript isoforms, and differentially expressed genes also can be identified. Although FunRich,[24] GeneProf and Microscope can conduct GO enrichment analysis, differential expression analysis and network construction, visual pathway and co-expression analysis are unavailable in these systems. Degust not only displays the normalized expression level of each condition, but also combines the expression level and KEGG pathway map.[25] Unfortunately, Degust is only supplied for RNA-seq data. Additionally, wapRNA is a good tool to analyze GO terms and KEGG pathways of the DEGs.[26] However, the significance of GO and pathways could not be easily identified in the output results, neither can the number of related transcripts. Nevertheless, all of the tools mentioned above are useful for bioinformatics scientists, while some are inefficient for biologists. EXPath Tool is constructed especially for biologists, and the web interface is very user-friendly and effective. Comprehensive analysis functions for transcriptome data (including microarray and RNA-seq) such as Gene Search, Pathway Search, DEGs Search, Pathways/GO Enrichment and Co-expression analysis are freely provided. It is thus easy to identify critical candidate genes/pathways for further experiments. This efficient tool is now available at http://expathtool.itps.ncku.edu.tw/.

Availability

The EXPath Tool webserver is freely available at http://expathtool.itps.ncku.edu.tw/.

Acknowledgements

The authors would like to thank the Ministry of Science and Technology (MOST 105-2311-B-006 -004-MY3 and MOST 104-2313-B-006-002) and Academia Sinica (Innovative Translational Agricultural Research Grant) of the Republic of China for financially supporting this research.

Supplementary data

Supplementary data are available at *DNARES* Online.

Conflict of interest

The authors have no conflict of interest to declare.

References

1. Wetterstrand, K. 2013, DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). *National Human Genome Research Institute*, Available at: http://www.genome.gov/sequencingcosts/.
2. Haas, B.J., Papanicolaou, A., Yassour, M., et al. 2013, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols*, **8**, 1494–512.
3. Schulz, M.H., Zerbino, D.R., Vingron, M. and Birney, E. 2012, Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–92.
4. Xie, Y., Wu, G., Tang, J., et al. 2014, SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30**, 1660–6.
5. Robertson, G., Schein, J., Chiu, R., et al. 2010, De novo assembly and analysis of RNA-seq data. *Nat. Methods*., **7**, 909–12.
6. Chang, Z., Li, G., Liu, J., et al. 2015, Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol*., **16**, 1.
7. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–6.
8. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990, Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–10.
9. Takami, H., Taniguchi, T., Arai, W., Takemoto, K., Moriya, Y. and Goto, S. 2016, An automated system for evaluation of the potential functionome: MAPLE version 2.1. 0. *DNA Res*., **23**, 467–75.
10. Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y. and Vandepoele, K. 2013, TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. *Genome Biol*., **14**, 1–10.
11. Consortium, U. 2014, UniProt: a hub for protein information. *Nucleic Acids Res*., gku989.
12. Ashburner, M., Ball, C.A., Blake, J.A., et al. 2000, Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–9.
13. Kanehisa, M. and Goto, S. 2000, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.
14. Gentleman, R.C., Carey, V.J., Bates, D.M., et al. 2004, Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.
15. Chien, C.-H., Chow, C.-N., Wu, N.-Y., Chiang-Hsieh, Y.-F., Hou, P.-F. and Chang, W.-C. 2015, EXPath: a database of comparative expression analysis inferring metabolic pathways for plants. *BMC Genomics*, **16**, S6.
16. Finn, R.D., Coggill, P., Eberhardt, R.Y., et al. 2015, The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*., gkv1344.
17. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T. and Thomas, P.D. 2016, PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res*., **44**, D336–42.
18. Lopes, C.T., Franz, M., Kazi, F., Donaldson, S.L., Morris, Q. and Bader, G.D. 2010, Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–8.
19. Nemhauser, J.L., Hong, F. and Chory, J. 2006, Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses. *Cell*, **126**, 467–75.
20. Goda, H., Sasaki, E., Akiyama, K., et al. 2008, The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. *Plant J*., **55**, 526–42.
21. Kartashov, A.V. and Barski, A. 2015, BioWardrobe: an integrated platform for analysis of epigenomics and transcriptomics data. *Genome Biol*., **16**, 1–8.
22. de Jong, A., van der Meulen, S., Kuipers, O.P. and Kok, J. 2015, T-REx: transcriptome analysis webserver for RNA-seq expression data. *BMC Genomics*, **16**, 663.
23. Zytnicki, M. and Quesneville, H. 2011, S-MART, a software toolbox to aid RNA-Seq data analysis. *PLoS One*, **6**, e25988.
24. Pathan, M., Keerthikumar, S., Ang, C.S., et al. 2015, FunRich: an open access standalone functional enrichment and interaction network analysis tool. *Proteomics*., **15**, 2597–601.
25. Powell, D.R. 2015, Degust: Visualize, explore and appreciate RNA-seq differential gene-expression data. (http://victorian-bioinformatics-consortium.github.io/degust/).
26. Zhao, W., Liu, W., Tian, D., et al. 2011, wapRNA: a web-based application for the processing of RNA sequences. *Bioinformatics*, **27**, 3076–7.