BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 17 2012, pages 2291–2292
doi:10.1093/bioinformatics/bts421

*Databases and ontologies*

Advance Access publication July 3, 2012

# AraPath: a knowledgebase for pathway analysis in Arabidopsis

Liming Lai[1], Arthur Liberzon[2], Jason Hennessey[1], Gaixin Jiang[1], Jianli Qi[1], Jill P. Mesirov[2] and Steven X. Ge[1,*]

[1]Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007 and [2]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA

Associate Editor: Jonathan Wren

**ABSTRACT**

**Summary:** Studying plants using high-throughput genomics technologies is becoming routine, but interpretation of genome-wide expression data in terms of biological pathways remains a challenge, partly due to the lack of pathway databases. To create a knowledgebase for plant pathway analysis, we collected 1683 lists of differentially expressed genes from 397 gene-expression studies, which constitute a molecular signature database of various genetic and environmental perturbations of Arabidopsis. In addition, we extracted 1909 gene sets from various sources such as Gene Ontology, KEGG, AraCyc, Plant Ontology, predicted target genes of microRNAs and transcription factors, and computational gene clusters defined by meta-analysis. With this knowledgebase, we applied Gene Set Enrichment Analysis to an expression profile of cold acclimation and identified expected functional categories and pathways. Our results suggest that the AraPath database can be used to generate specific, testable hypotheses regarding plant molecular pathways from gene expression data.

**Availability:** http://bioinformatics.sdstate.edu/arapath/

**Contact:** gexijin@gmail.com

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online.

Received on February 28, 2012; revised on June 11, 2012; accepted on June 27, 2012

## 1 INTRODUCTION

The coverage and quality of gene set databases are critical to the effectiveness of all pathway analysis programs. The molecular Signatures Database (MSigDB) is a collection of gene sets for pathway analysis in mammals using Gene Set Enrichment Analysis (GSEA; Subramanian *et al*., 2005) or other similar methods. The most recent version (3.0) of MSigDB contains 6769 sets of genes from various sources. In addition to existing annotation databases such as Gene Ontology (Ashburner *et al*., 2000) and KEGG (Kanehisa *et al*., 2006), lists of differentially expressed genes were manually collected from hundreds of published gene expression studies of genetic and chemical perturbations. Inclusion of these gene sets in the database enables the detection of co-regulation of genes similar to those induced by chemical and genetic perturbations reported in the literature. This will be useful in identifying shared regulatory mechanisms (Ge, 2011), complementing existing tools such as AtCAST (Sasaki *et al*., 2011), HORMONOMETER (Volodarsky *et al*., 2009) and Sample Angler (bar.utoronto.ca).

To facilitate pathway analysis for plant genomics, we sought to construct a database similar to the MSigDB, so that the various pathway analysis programs could be easily used to analyze plant genomic data. Recently, (Schuler *et al*., 2011) extracted gene sets from several sources including GO, KEGG, TRASPATH and TRASFAC, so that they could use GSEA within the GeneTrail software. In this study, we compiled a large, publically available gene sets database by extracting information from existing databases and by collecting gene expression signatures from the literature. We focused on *Arabidopsis thaliana*, a widely studied plant model organism with large amounts of genetic, biochemical, physiological and genomic data available. This resource will form the basis for making similar resources for other plants and even more distant species such as *Plasmodium falciparum* (malaria parasite).

## 2 METHODS

As most authors deposit their raw genomics data into public repositories, we searched for publications reporting microarray studies of Arabidopsis in Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (www.ebi.ac.uk/arrayexpress/). The full text of the papers and their supplementary materials were retrieved and manually curated.

Gene IDs from various sources were converted to NCBI gene symbols whenever possible. AGI (Arabidopsis Genome Initiative) IDs were used for genes without official gene symbols. The conversion is based on the Arabidopsis gene information at NCBI (ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Plants/). We also used the gene aliases table at NCBI and TAIR to convert a small number of genes.

The web site was assembled using a combination of the Python programming language, Django web framework, MySQL, JavaScript and html.

## 3 RESULTS AND DISCUSSION

We succeeded in extracting pathways or functional categories information from various sources to construct a knowledgebase for pathway analysis in *Arabidopsis thaliana* (Table 1). Our database includes information from several public databases like Gene Ontology, KEGG and AraCyc. We also included predicted miRNA target genes from six studies (see data file for references). Confirmed and unconfirmed targets of transcription factors were obtained based on Yilmaz *et al*. (2011). The Plant Ontology (PO) Consortium created controlled vocabularies and association of genes with the anatomical site and developmental

* To whom correspondence should be addressed.

**Table 1.** Sources for gene sets in AraPath

| Source | No. of gene sets ($N \geq 5$) | References |
|---|---|---|
| Literature | 1683 | Present study (397 references) |
| Gene Ontology | 941 | (Ashburner *et al.*, 2000) |
| AraCyc pathways | 224 | (Mueller *et al.*, 2003) |
| KEGG pathways | 109 | (Kanehisa *et al.*, 2006) |
| microRNA target genes | 309 | 6 References (see data) |
| TF target genes | 33 | (Yilmaz *et al.*, 2011) |
| Computational gene sets | 48 | (Atias *et al.*, 2009) |
| | 15 | (Wilson *et al.*, 2012) |
| Plant Ontology | 230 | (Jaiswal *et al.*, 2005) |
| Total | 3692 | |

stage at which a gene is expressed (Jaiswal *et al.*, 2005). The information in PO was used to derive gene lists. We also collected sets of genes that are found to be co-expressed in meta-analyses of expression datasets (Atias *et al.*, 2009; Wilson *et al.*, 2012). From these seven sources we collected 1909 gene lists (Table 1).

The most challenging part was the extraction of gene lists from hundreds of published gene expression studies. We examined 1039 datasets in the GEO and 590 in ArrayExpress and were able to retrieve 783 corresponding papers (52% of the datasets lacked publication information). After reading the 783 papers, we found that 397 of them contained 1683 useful sets of differentially expressed genes. Each gene set was annotated with a unique name and a brief description, accompanied by detailed information on the source publication. Such information will be useful in finding similarities in expression signatures of a user's data with those published.

An additional 740 gene sets with less than five genes were also collected from the sources mentioned above. All together we collected 4332 gene sets. These include gene sets similar to all the categories of MSigDB excluding the C1: position gene sets. Besides pathway analysis, the databases of differentially expressed gene lists also serve as an information source that can be queried or analyzed. Similar attempts have been made for the human genome to create databases of reported gene lists (Cahan *et al.*, 2005; Newman and Weiner, 2005). As lists of affected genes are often the final results of genome-wide studies, the archival of these results will facilitate the easy access to this accumulated knowledge. Our data can be downloaded or queried at http://bioinformatics.sdstate.edu/arapath/.

To demonstrate the utility of the database, we used GSEA to perform pathway analysis of a dataset (GEO accession number GSE5534) that measures the response of Arabidopsis seedlings to cold acclimation (4°C for a day). See Supplementary Materials 1–3 for details. We identified 12 significant GO terms, including 'Response to cold', 'structural constituent of ribosome', 'ribosome' and 'translation'. Proper function of the translation machinery is necessary at low temperatures, and a previous genome-wide expression study found the up-regulation of ribosomal proteins during cold acclimation (Hannah *et al.*, 2005). When we ran GSEA with the literature gene sets, we identified 128 significant gene sets, many more than the 12 identified using GO gene sets. Out of the top 10 up-regulated gene sets, 5 of them are cold related. This suggests the genomic response to cold in the new study is similar to these reported in previous studies.

In summary, we have compiled a gene set database for pathway analysis in Arabidopsis by extracting information from various databases of gene functional categories and pathways. It also includes over a thousand lists of differentially expressed genes collected from the published genome-wide studies. We will continue to add to this database and hope this resource will facilitate various genomics studies and generate novel hypotheses.

## ACKNOWLEDGEMENTS

The authors thank Fedora Sutton for providing valuable comments.

*Funding*: NIH grant to SXG (GM083226, in part).

*Conflict of Interest*: None declared.

## REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Atias,O. *et al.* (2009) Large-scale analysis of Arabidopsis transcription reveals a basal co-regulation network. *BMC Syst. Biol.*, **3**, 86.

Cahan,P. *et al.* (2005) List of lists-annotated (LOLA): a database for annotation and comparison of published microarray gene lists. *Gene*, **360**, 78–82.

Ge,S.X. (2011) Large-scale analysis of expression signatures reveals hidden links among diverse cellular processes. *BMC Syst. Biol.*, **5**, 87.

Hannah,M.A. *et al.* (2005) A global survey of gene regulation during cold acclimation in Arabidopsis thaliana. *PLoS Genet.*, **1**, e26.

Jaiswal,P. *et al.* (2005) Plant ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp. Funct. Genom.*, **6**, 388–397.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Mueller,L.A. *et al.* (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol.*, **132**, 453–460.

Newman,J.C. and Weiner,A.M. (2005) L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol.*, **6**, R81.

Sasaki,E. *et al.* (2011) AtCAST, a tool for exploring gene expression similarities among DNA microarray experiments using networks. *Plant Cell Physiol.*, **52**, 169–180.

Schuler,M. *et al.* (2011) Transcriptome analysis by GeneTrail revealed regulation of functional categories in response to alterations of iron homeostasis in Arabidopsis thaliana. *BMC Plant Biol.*, **11**, 87.

Subramanian,A. *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.

Volodarsky,D. *et al.* (2009) HORMONOMETER: a tool for discerning transcript signatures of hormone action in the Arabidopsis transcriptome. *Plant Physiol.*, **150**, 1796–1805.

Wilson,T.J. *et al.* (2012) Identification of metagenes and their interactions through large-scale analysis of Arabidopsis gene expression data. *BMC Genom.*, **13**, 237.

Yilmaz,A. *et al.* (2011) AGRIS: the Arabidopsis gene regulatory information server, an update. *Nucleic Acids Res.*, **39**, D1118–D1122.