D1060–D1063 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 29 November 2016*
doi: 10.1093/nar/gkw1029

AtPID: a genome-scale resource for genotype–phenotype associations in Arabidopsis

Qi Lv[1,2,†], Yiheng Lan[1,†], Yan Shi[1,†], Huan Wang[1,†], Xia Pan[1], Peng Li[1,*] and Tieliu Shi[1,*]

[1]Center for Bioinformatics and Computational Biology, and the Institute of Biomedical Sciences, School of Life Sciences, East China Normal University, Shanghai 200241, China and [2]School of Finance and Statistics, East China Normal University, Shanghai 200241, China

Received September 15, 2016; Revised October 16, 2016; Editorial Decision October 17, 2016; Accepted November 08, 2016

ABSTRACT

AtPID (*Arabidopsis thaliana* Protein Interactome Database, available at http://www.megabionet.org/atpid) is an integrated database resource for protein interaction network and functional annotation. In the past few years, we collected 5564 mutants with significant morphological alterations and manually curated them to 167 plant ontology (PO) morphology categories. These single/multiple-gene mutants were indexed and linked to 3919 genes. After integrated these genotype–phenotype associations with the comprehensive protein interaction network in AtPID, we developed a Naïve Bayes method and predicted 4457 novel high confidence gene-PO pairs with 1369 genes as the complement. Along with the accumulated novel data for protein interaction and functional annotation, and the updated visualization toolkits, we present a genome-scale resource for genotype–phenotype associations for Arabidopsis in AtPID 5.0. In our updated website, all the new genotype–phenotype associations from mutants, protein network, and the protein annotation information can be vividly displayed in a comprehensive network view, which will greatly enhance plant protein function and genotype–phenotype association studies in a systematical way.

INTRODUCTION

Protein functional annotation and the protein networks including protein–protein interactions and regulatory relations are essential for understanding the underlying mechanism of the biological system. AtPID is a comprehensive data resource developed using *Arabidopsis thaliana* as the model system for protein interactions and functional annotation. From the year 2005, we started to collect the protein–protein interactions (PPIs) from literature and released AtPID 1.0, which only included limited curated PPIs and protein functional annotations from TAIR (The Arabidopsis Information Resource) (1). Due to the increasing demand of the comprehensive PPI from related research communities, we extended the PPI network by different computational methods and released AtPID 2.0 in 2006. In order to further increase the coverage and overcome the false positive issue within the predicted dataset, we manually curated more PPIs from the literature and developed a Naïve Bayesian based classifier to integrate and evaluate all the predicted PPIs, which made our database updated to AtPID 3.0 in 2008 as a rich source of information for system-level understanding of gene function and biological processes (2). In order to better serve the related research communities for the mechanism studies of various physiological activities, we annotated the Arabidopsis proteins in the AtPID 4.0 database with further information (e.g. functional annotation, subcellular localization, tissue-specific expression, phosphorylation information, SNP phenotype and mutant phenotype, etc.) and interaction qualifications (e.g. transcriptional regulation, complex assembly, functional collaboration, etc.) via further literature text mining and integration of other resources (3) (Table 1).

Comparing with other organisms, plants have unique advantages on the mutagenesis and tissue culture, a large number of characterized stable *Arabidopsis* mutants have been reported in research literature, and large-scale seeds/mutant resources for plant functional studies were built for genome annotation and functional studies, e.g. uNASC Database (The European Arabidopsis Stock Centre), RAPID (RIKEN Arabidopsis Phenome Information Database), CSHL Database (the Arabidopsis Genetrap Website at Cold Spring Harbor Lab), Chloroplast Function Database, SeedGenes Database, AGRICOLA Database (Systematic RNAi knockouts in Arabidopsis), Araport (the Arabidopsis Information Portal) and TAIR (4–10). Mutant phenotypes are especially critical for functional studies of plants. Although great efforts have been made on collecting related data in plants, the mutant phenotypes are still

*To whom correspondence should be addressed. Tel: +86 21 54345020; Fax: +86 21 54344922; Email: tlshi@bio.ecnu.edu.cn
Correspondence may also be addressed to Peng Li. Tel: +86 21 54345020; Fax: +86 21 54344922; Email: PLI6@mgh.harvard.edu
†These authors contributed equally to the paper as first authors.

Table 1. A comprehensive comparison for the different versions of AtPID database

	Function annotation				Molecule interaction			
The version of AtPID	Protein functional description	Subcellular localization	Mutant	Phenotype annotation	Curated PPIs	Predicted PPIs	Curated transcriptional regulations	Predicted transcriptional regulations
AtPID 3 (2)	32 000	–	–	–	4666	23 396	–	–
AtPID 4 (3)	40 000	10 429	5121 mutants, 3431 genes	–	5565	98 174	8070	–
AtPID 5 (Current Version)	40 000	11 052	5609 mutants, 3916 genes	8202 mutant-PO associations	45 382	118 556	9435	31 991

largely under-annotated. AtPID has been committed to collect more mutants with significantly morphological alterations and tried to annotate all the mutants' phenotypes in a systematical way. The Plant Ontology is a controlled vocabulary (ontology) that describes plant anatomy and morphology and stages of development for all plants (11). In order to index and annotate all the mutants in AtPID into a standard semantic framework, we cooperated with Shanghai Society for Plant Biology and annotated all the mutants to more specific downstream PO categories.

In this update, the AtPID 5.0 database greatly expands the information on PPIs, mutant phenotypes obtained from published literature (12–14), public databases and computational approaches. For mutant related information, the data of mutant phenotypes were carefully curated by biologists. In addition, novel associations between genes and phenotypes were predicted through Naïve Bayes method. Furthermore, we developed a more comprehensive visualization toolkit to view all the interactions at PPI, transcriptional regulation and genotype–phenotype levels under the same framework, which could easily show/map all other annotation information in our database for selected genes. All of the improvements and updates will accelerate researchers in exploiting information in our database in a more effective and comprehensive way.

RESULTS

Summary of new data in the updated AtPID 5.0

Comparing with the other well-used PPI resources (Table 2), the updated database indexed 45 382 curated PPIs and 118 556 predicted PPIs from literature mining, public databases or computational approaches. These numbers are significantly increased due to the ravenous growth and maturing biomedical national processing language and the large-scale experiments for functional studies (15–17). We also generated a comprehensive chloroplast proteomics dataset in Arabidopsis by large-scale proteomics experiments and indexed all 3134 credible chloroplast proteins into our annotation system. Furthermore, we systematically annotated 31 991 TFBS associations to 6891 genes based on the integration of expression profiling and cis-regulatory element information. This update largely enriches protein annotations in our database by tracking the recent research progresses of related areas and will greatly assist functional experiments and systematic studies.

Comprehensive annotation of genotype–phenotype associations

Using text mining and database integration, the previous version (AtPID 4.1) collected 5121 mutants with significantly observable phenotypes related to 3431 genes. In the past few years, through in-depth cooperation with Shanghai Society for Plant Biology, we collected 488 new mutants and systematically annotated all the existed and new curated mutants' phenotypes to 167 standardized plant ontology categories (Figure 1A). Comprehensive collection on phenotype data can help phenotype mechanism studies as what have been done in systematical exploration of disease associations (18,19). Strategies or algorithms have been developed to predict gene related functions by integrating multiple level data (18–20). We integrated three different information, PPIs, co-expression from expression profiling and GO annotation with Naïve Bayes method. PPIs were quantified by the extended Czekanowski–Dice distance (21) and missing values were complemented by orthologs in other 14 species' experimental PPIs from STRING database (22). Shared Smallest Biological Processes (SSBPs) was applied to describe the possibility of gene interactions on GO annotation (23). Co-expression of gene pairs were computed over the microarrays mentioned above to predict regulatory interactions. The correlation coefficient values of the three information were low (PPIs-GO: 0.05; PPIs-co-expression: 0.08; co-expression-GO: −0.03), suggesting that features were independent from each other and satisfied with the assumption of Naïve Bayes method. Naïve Bayes was undertaken by e1071 package in R. The model showed high predictability, with average AUC 0.72. Finally, the prediction contains 4457 novel gene-PO pairs with 1369 genes, which could be a supplement to the known mutant information.

User friendly visualization toolkit for comprehensive genotype–phenotype network

For the phenotype annotation information, we redeveloped the network visualization application (Figure 1B and C) with JavaScript, which inherited all the functions of the old java applet, and added phenotype as a new node type. The new visualization application has better compatibility and performance due to the optimization of the database structure and the network generation methods. Meanwhile, it presents the network in a more interactive and comprehensive way. All the protein annotation information and protein relations in AtPID 5.0 can be presented simultaneously on the same view, and users can easily extend the network by double clicking any node on the border of current network. The combination

Table 2. Numbers of interactions in AtPID 5.0 compared with the other well-used data resources

PPI-related database	Description for the PPI database	Curated PPIs	Predicted PPIs
AtPID 5.0	An integrated database resource for protein interaction network and functional annotation proteome.(http://www.megabionet.org/atpid)	45 382	118 556
PAIR	The predicted *Arabidopsis* interactome resource(http://www.cls.zju.edu.cn/pair/) (24)	5990	137 986
TAIR	A database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana*.(http://www.arabidopsis.org) (25)	6503	
BioGRID	An interaction repository with data compiled through comprehensive curation efforts.(http://thebiogrid.org/) (26)	42 216	
STRING	A database of predicted functional associations between proteins.(http://string-db.org/) (27)		>1 000 000

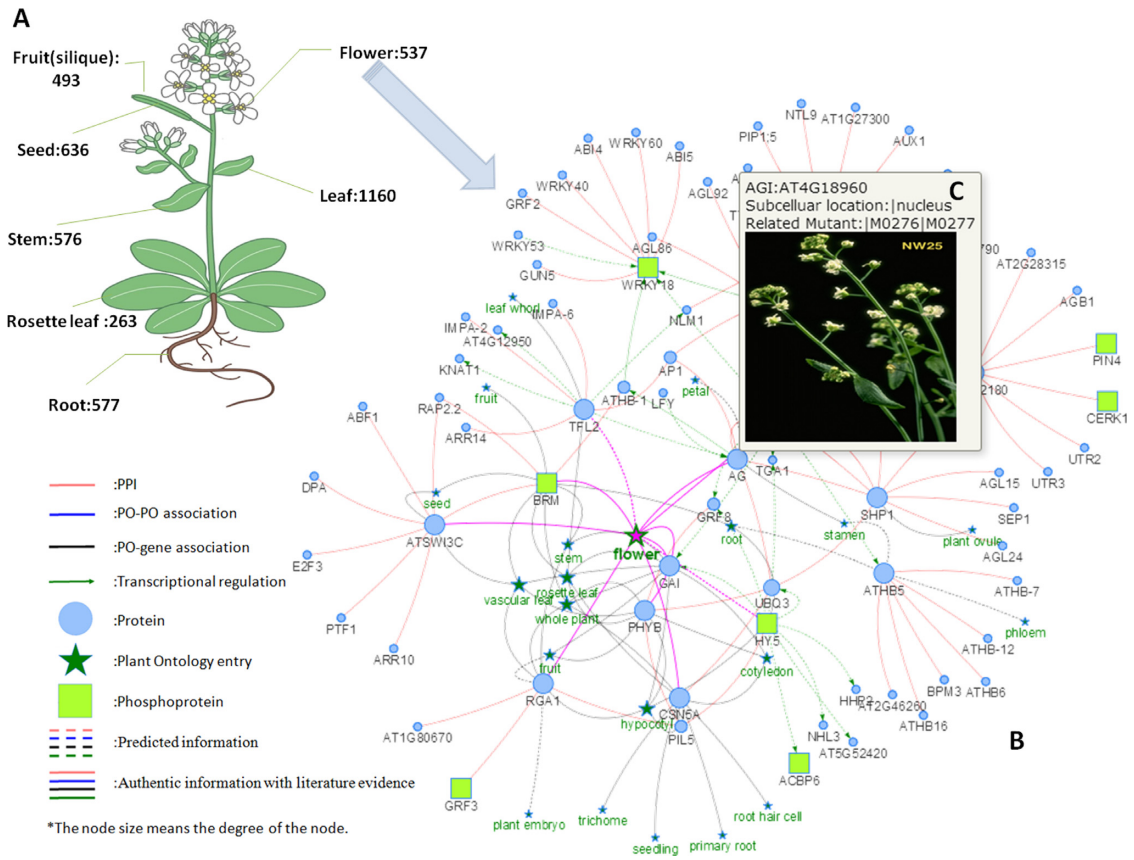

Figure 1. The overview and network display of the curated genotype–phenotype associations in AtPID 5.0. Top left corner (**A**) exhibits the top level Plant Ontology (PO) entries in *Arabidopsis* and the annotated gene numbers related to this PO. Bottom right corner (**B**) shows the flower-associated network. (**C**) Node with mouse hovering annotation.

of genotype–phenotype associations and the protein interaction information can provide existing knowledge of selected proteins to biologists in a very intuitive way and help them easily understand the functional relations to confirm their hypotheses or inspire them on new study designs.

CONCLUSIONS

Here, we have made great efforts to provide a significantly improved resource for genotype–phenotype associations, which could serve as a resource for experimental design and facilitate genome-wide systematical studies in Arabidopsis. The AtPID 5.0 also provides illustrations of the functional annotation and protein network with a friendly web-based interface. We have largely extended the current annotation information by literature curation, bioinformatics predictions and also the high-throughput experimental data in the AtPID 5.0, e.g. we generated a comprehensive chloroplast proteomics dataset in Arabidopsis by large-scale proteomics experiments and indexed all the data as the evidence for subcellular localization in current AtPID. We will continue to accumulate more genome-wide data to better serve the research community.

ACKNOWLEDGEMENTS

We are grateful to Dr. Wenzhong Xiao from Harvard Medical School for assistance in revising manuscript, Wubin Ding and Huanlong Liu from East China Normal University for giving technical supports on this research. We are grateful to Drs. Jin Xu and Xiaolong Pu from East China Normal University for their support on this project.

FUNDING

National Key Basic Research and Development Plan 973 [2013CB127005]; National High Technology Research and Development Program of China [2015AA020108]; National Science Foundation of China [31171264, 31401133, 31671377]; 111 Project [B14019]; Science and Technology Commission of Shanghai Municipality [14YF1404400]; Ernst Mach-Stipendien Eurasia-Pacific Uninet programme; China Postdoctoral Science Foundation funded project and the Supercomputer Center of East China Normal University. Funding for open access charge: National Science Foundation of China [31171264, 31401133, 31671377].

Conflict of interest statement. None declared.

REFERENCES

1. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
2. Cui,J., Li,P., Li,G., Xu,F., Zhao,C., Li,Y., Yang,Z., Wang,G., Yu,Q., Li,Y. *et al.* (2008) AtPID: Arabidopsis thaliana protein interactome database–an integrative platform for plant systems biology. *Nucleic Acids Res.*, **36**, D999–D1008.
3. Li,P., Zang,W., Li,Y., Xu,F., Wang,J. and Shi,T. (2011) AtPID: the overall hierarchical functional protein interaction network interface and analytic platform for Arabidopsis. *Nucleic Acids Res.*, **39**, D1130–D1133.
4. Kuromori,T., Wada,T., Kamiya,A., Yuguchi,M., Yokouchi,T., Imura,Y., Takabe,H., Sakurai,T., Akiyama,K., Hirayama,T. *et al.* (2006) A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of Arabidopsis. *Plant J.*, **47**, 640–651.
5. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
6. Nakayama,N., Arroyo,J.M., Simorowski,J., May,B., Martienssen,R. and Irish,V.F. (2005) Gene trap lines define domains of gene regulation in Arabidopsis petals and stamens. *Plant Cell*, **17**, 2486–2506.
7. Myouga,F., Akiyama,K., Tomonaga,Y., Kato,A., Sato,Y., Kobayashi,M., Nagata,N., Sakurai,T. and Shinozaki,K. (2013) The Chloroplast Function Database II: a comprehensive collection of homozygous mutants and their phenotypic/genotypic traits for nuclear-encoded chloroplast proteins. *Plant Cell Physiol.*, **54**, e2.
8. Meinke,D., Muralla,R., Sweeney,C. and Dickerman,A. (2008) Identifying essential genes in Arabidopsis thaliana. *Trends Plant Sci.*, **13**, 483–491.
9. Hilson,P., Allemeersch,J., Altmann,T., Aubourg,S., Avon,A., Beynon,J., Bhalerao,R.P., Bitton,F., Caboche,M., Cannoot,B. *et al.* (2004) Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications. *Genome Res.*, **14**, 2176–2189.
10. Krishnakumar,V., Hanlon,M.R., Contrino,S., Ferlanti,E.S., Karamycheva,S., Kim,M., Rosen,B.D., Cheng,C.Y., Moreira,W., Mock,S.A. *et al.* (2015) Araport: the Arabidopsis information portal. *Nucleic Acids Res.*, **43**, D1003–D1009.
11. Jaiswal,P., Avraham,S., Ilic,K., Kellogg,E.A., McCouch,S., Pujar,A., Reiser,L., Rhee,S.Y., Sachs,M.M., Schaeffer,M. *et al.* (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp. Funct. Genomics*, **6**, 388–397.
12. Fan,T., Yang,L., Wu,X., Ni,J., Jiang,H., Zhang,Q., Fang,L., Sheng,Y., Ren,Y. and Cao,S. (2016) The PSE1 gene modulates lead tolerance in Arabidopsis. *J. Exp. Bot.*, **67**, 4685–4695.
13. Fu,X., Li,C., Liang,Q., Zhou,Y., He,H. and Fan,L.M. (2016) CHD3 chromatin-remodeling factor PICKLE regulates floral transition partially via modulating LEAFY expression at the chromatin level in Arabidopsis. *Sci. China. Life Sci.*, **59**, 516–528.
14. Zhao,S., Zhao,Y. and Guo,Y. (2015) 14-3-3 λ protein interacts with ADF1 to regulate actin cytoskeleton dynamics in Arabidopsis. *Sci. China. Life Sci.*, **58**, 1142–1150.
15. Hunter,L., Lu,Z., Firby,J., Baumgartner,W.A. Jr., Johnson,H.L., Ogren,P.V. and Cohen,K.B. (2008) OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, **9**, 78.
16. Yang,Z., Zhao,Z., Li,Y., Hu,Y. and Lin,H. (2013) PPIExtractor: a protein interaction extraction and visualization system for biomedical literature. *IEEE Trans. Nanobiosci.*, **12**, 173–181.
17. Meyer,P., Alexopoulos,L.G., Bonk,T., Califano,A., Cho,C.R., de la Fuente,A., de Graaf,D., Hartemink,A.J., Hoeng,J., Ivanov,N.V. *et al.* (2011) Verification of systems biology research in the age of collaborative competition. *Nat. Biotechnol.*, **29**, 811–815.
18. Calvo,S., Jain,M., Xie,X., Sheth,S.A., Chang,B., Goldberger,O.A., Spinazzola,A., Zeviani,M., Carr,S.A. and Mootha,V.K. (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet.*, **38**, 576–582.
19. Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res.*, **18**, 644–652.
20. Davis,D.A. and Chawla,N.V. (2011) Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PLoS One*, **6**, e22670.
21. Brun,C., Herrmann,C. and Guenoche,A. (2004) Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics*, **5**, 95.
22. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
23. Rhodes,D.R., Tomlins,S.A., Varambally,S., Mahavisno,V., Barrette,T., Kalyana-Sundaram,S., Ghosh,D., Pandey,A. and Chinnaiyan,A.M. (2005) Probabilistic model of the human protein–protein interaction network. *Nat. Biotechnol.*, **23**, 951–959.
24. Lin,M., Shen,X. and Chen,X. (2011) PAIR: the predicted Arabidopsis interactome resource. *Nucleic Acids Res.*, **39**, D1134–D1140.
25. Reiser,L., Berardini,T.Z., Li,D., Muller,R., Strait,E.M., Li,Q., Mezheritsky,Y., Vetushko,A. and Huala,E. (2016) Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. *Database (Oxford)*, **2016**.
26. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
27. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.