LETTER TO THE EDITOR

Araport Lives: An Updated Framework for Arabidopsis Bioinformatics

Conceived as a replacement for the anticipated retirement of The Arabidopsis Information Resource (TAIR), the Araport project was funded by the U.S. National Science Foundation (NSF) in 2013 to develop a new, extensible framework for Arabidopsis (Arabidopsis thaliana) bioinformatics that would facilitate data integration through the federation of distributed informatics resources. Recommended as a 5-year award, funds were initially provided for 2 years with a subsequent 6-month extension to allow for the development of a plan for the continuation of funding that was acceptable to NSF. When funds were exhausted in 2016 with the continuation award still awaiting a decision, the Araport site continued in a maintenance mode with minimal input from legacy personnel at each institution and with no data updates. The renewal request was finally declined in December 2018, leaving Araport with an uncertain future. In light of the critical importance of these services to the scientific community, a group of interested researchers (see Appendix) met in March 2019 to discuss options and propose a solution. A working group evolved from those in attendance at that meeting and has since met monthly to solidify and coordinate the execution of these nascent plans. The results have been encouraging and are described here to inform and inspire the larger plant science community.

Given the complete absence of external funding, it was agreed that, rather than try to perpetuate the entire Araport ecosystem, efforts should be directed toward maintaining the most attractive and most used features, namely ThaleMine and JBrowse, by transferring them to new ownership for perpetuation. Thus it was agreed that an updated version of ThaleMine would be established at the Bio-Analytic Resource (BAR) for Plant Biology at the University of Toronto under the leadership of Nicholas Provart and an updated version of JBrowse would be established as part of TAIR under the Phoenix Bioinformatics umbrella overseen by Tanya Berardini and Eva Huala.

JBROWSE

The JBrowse functionality provided by Araport has been successfully moved to TAIR. Araport had been running version 1,11.6 of the software with a set of tracks that included community submissions. Some of the tracks at the legacy location were no longer functioning after the underlying software (ADAMA) connecting them to outside resources lost support. TAIR installed the latest JBrowse version (1.16.6), replicated the tracks that were functional at Araport, and restored access to the nonfunctional tracks. In addition, two sets of newly integrated community-submitted tracks are now visible in this genome browser. One is a set of 41 tracks representing a multi-pronged gene expression experiment to track the response to various abiotic stresses from Lee and Bailey-Serres (2019). The other is a set of 4 tracks based on Cap Analysis of Gene Expression (CAGE) experiments to determine promoter bidirectionality performed by Thieffry et al. (2019). The CAGE data are visualized using the Stranded View plugin (Hofmeister and Schmitz, 2018), which allows separation of the display of expression values into plus and minus strands in a single track. New community tracks continue to be added, and existing track information is updated as new data become available.

THALEMINE

ThaleMine at the BAR was completely rebuilt using the latest InterMine software. The legacy ThaleMine version had not been updated since 2016 and was using the InterMine version 1.8.5, which was not forward-compatible with the latest InterMine version (4.2.0). At the time of writing, the most recent versions of publicly available data have been loaded, as listed in Table 1.

As with any instance of InterMine, the BAR’s version of ThaleMine at https://bar.utoronto.ca/thalemine/ continues to support application programming interface functionalities, in addition to the extensive web-based query options. It is also compatible with the InterMine BlueGenes interface.

GENOME CONTEXT VIEWER

As part of the unsuccessful renewal proposal, some aspects of the Araport Comparative Genomics functionalities were planned to be addressed with an instance of the Genome Context Viewer (GCV) software developed at the National Center for Genome Resources (NCGR) by Andrew Farmer and Alan Cleary (Cleary and Farmer, 2018). This viewer was originally developed as part of an NSF-funded initiative for federating disparate legume-focused information resources. It provides services to enable the dynamic comparison of multiple genomes on the basis of their shared functional elements (e.g., genes) and provides an intuitive and powerful user interface for exploring similarities and differences among a set of genomic segments with respect to element content and arrangement. A version of the GCV has been installed and is now running from NCGR (https://gcv-arabidopsis.ncgr.org) as the third component of the “second-generation” Araport (see figure). This version of the GCV provides integration of the Arabidopsis Columbia reference genome (TAIR10/Araport11) with genomes from several other data sources, including two sets of newly assembled Arabidopsis genomes of various accessions (colloquially often called ecotypes) from Jiao and Schneeberger (2020) and from the 1001 Genomes project from Detlef Weigel and colleagues (Felix Bem, Christian Kubica, and Detlef Weigel, personal communication), as well as a number of Brassicaceae genomes from Phytozome and the Brassicaceae Map Alignment Project initiative. The viewer provides convenient links to related resources for genes and genomic regions.
Table 1. Data Sources for a New Instance of ThaleMine

| Data                               | Source                      | Reference                                      |
|------------------------------------|-----------------------------|------------------------------------------------|
| Genome Data                        |                             |                                                |
| TAIR10 genome assembly (five chromosomes plus chloroplast and mitochondrial assemblies) | NCBI, release TAIR10 (2018/04/06) | (Arabidopsis Genome Initiative, 2000)          |
| Araport11 GFF3 data from TAIR      | TAIR, release Araport11 (2016/06/17) | (Cheng et al., 2016)                           |
| Protein Data                       |                             |                                                |
| High-quality, manually annotated, nonredundant protein sequence database | Swiss-Prot, release 2020_02 | (UniProt Consortium, 2007)                     |
| Computationally analyzed records, enriched with automatic annotation | TrEMBL, release 2020_02 | (UniProt Consortium, 2007)                     |
| Protein family and domain assignments to proteins | InterPro, release v79.0 | (Mitchell et al., 2019)                        |
| Homology Data                      |                             |                                                |
| Ortholog and paralog relationships based on the inferred speciation and gene duplication events from a phylogenetic analysis | Panther, release 14.1 | (Mi et al., 2013)                              |
| Phytozome homologs generated with InParanoid | Phytozome, real time | (Goodstein et al., 2012)                       |
| Curation                           |                             |                                                |
| Manually curated TAIR functional descriptions | TAIR, release 2019-04-02 | (Huala et al., 2001)                           |
| Manually curated TAIR gene aliases | TAIR, release 2019-04-02 | (Huala et al., 2001)                           |
| Gene Ontology                      |                             |                                                |
| GO annotations from Gene Ontology   | Gene Ontology, release 2020-06-01 | (Ashburner et al., 2000; Berardini et al., 2004) |
| Several electronic and manual GO annotation methods utilized by UniProt | UniProt, release 2020_02 | (UniProt Consortium, 2007)                     |
| Interaction Data                   |                             |                                                |
| Curated set of genetic and physical interactions for Arabidopsis | BioGRID, release 3.5.186 | (Chatr-Aryamontri et al., 2017)               |
| Curated binary and complex protein-protein interactions for Arabidopsis | IntAct, downloaded 2020-06-11 | (Kerrien et al., 2012)                        |
| Expression and Coexpression Data   |                             |                                                |
| Electronic Fluorescent Pictograph (eFP) visualization paints gene expression information from one of the AtGenExpress data sets or other compendia for a desired gene onto a diagrammatic representation of Arabidopsis plants | BAR eFP webservice, real time | (Winter et al., 2007; Brady and Provart, 2009) |
| Coexpressed gene relationships deduced from microarray and RNA-seq data via ATTED-II web services | ATTED-II coexpression, real time | (Obayashi et al., 2014)                        |
| Publications and GeneRIFs (reference-into-function) | Publications and GeneRIFs (reference-into-function) | (UniProt Consortium, 2007)                     |
| Curated associations between publications and genes from UniProt | UniProt, release 2020_02 | (UniProt Consortium, 2007)                     |
| Publications from InterPro          | InterPro, release v79.0 | (Mitchell et al., 2019)                        |
| Publications from NCBI              | NCBI, downloaded 2020-06-11 | (Maglott et al., 2007)                        |
| Concise phrase describing gene function and publication associated with NCBI gene records | NCBI, downloaded 2020-06-11 | (Maglott et al., 2007)                        |

thereby facilitating traversal into the other components of the reconfigured Araport project as well as other relevant tools. The gene family classifications utilized by the current instance are based on PANTHER 14.1 (Mi et al., 2013), and links are provided to the trees developed for these families by the PhyloGenes project (phylogenes.org).

LONG LIVE ARAPORT!

To establish continuity between the original Araport and these new functionalities, http://araport.org/ is now hosted at BAR and visitors are then presented with links to the new and maintained versions of ThaleMine, JBrowse, and the GCV. With these new sites operational, the original Araport site hosted at the Texas Advanced Computing Center has been shut down because of security issues related to the legacy versions of the packages used by the original site. We expect that the new Araport in its various component parts will continue to be widely used not just by Arabidopsis researchers but by the wider plant community.

In summary, a grassroots effort by committed community members has built upon the resources developed by the Araport project to provide continuity of Araport’s most used and useful features. It is gratifying to see that the vision of the 2012 white paper (International Arabidopsis Informatics Consortium, 2012) suggesting a future for Arabidopsis informatics as a community effort accomplished by a federation of independent community members has, in a modest way, come to pass. March 2020 saw 10,376 views of the ThaleMine landing page, showing a wide uptake by the community. That said, this rescue effort is not really a sustainable solution. Data curation and database maintenance are of vital importance and, notwithstanding TAIR’s successful subscription model, is something that is worthy of support by national funding agencies for the continued success of plant research in the United States and worldwide.

Asher Pasha1
Bio-Analytic Resource for Plant Biology, Department of Cell and Systems Biology/Centre for the
Screenshot of New Arabidopsis GCV Showing a Region with Two Clusters of Germin-Like Proteins (PTHR31238 Gene Family, Denoted Here as Purple). The central cluster shows extensive copy number variation among annotations from 14 Arabidopsis genomes and the closely related Arabidopsis lyrata genome (labeled araly.scaffold_7), as highlighted by the asterisks along the bottom. Other apparent copy number variations and presence/absence events can easily be observed.

ACKNOWLEDGMENTS

We are especially grateful to the scientists at the Texas Advanced Computing Center, especially Erik Ferlanti, John Fonner, and Matt Vaughn, for continuing to host and maintain Araport long after its “use-by” date. We thank Vivek Krishnakumar for providing insights and advice on the inner workings of Araport during the transition period. Araport was supported by grants from the National Science Foundation (grant DBI-1262414) and the Biotechnology and Biological Sciences Research Council (grant BB/L027151/1). Development of the GCV was supported by USDA-ARS project funding for the Legume Information System and the National Science Foundation (grant IOS-1444806 to A.F.). The J. Craig Venter Institute workshop that launched the Araport recovery effort was supported by the U.S. National Science Foundation (MCB Award 1062348, made to the U.S. members of the International Arabidopsis Informatics Consortium Steering Committee). The BAR is supported through a grant to N.P. from the National Sciences and Engineering Research Council.

Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario M5S 3B2, Canada
ORCID ID: 0000-0002-9315-0520
Subramaniam Shabari
The Arabidopsis Information Resource/Phoenix Bioinformatics
Fremont, California 94538
ORCID ID: 0000-0002-0695-4935
Alan Cleary
National Center for Genome Resources
Santa Fe, New Mexico 87505
ORCID ID: 0000-0002-6567-5346
Xingguo Chen
The Arabidopsis Information Resource/Phoenix Bioinformatics
Fremont, California 94538
ORCID ID: 0000-0001-9589-3726
Tanya Berardini
The Arabidopsis Information Resource/Phoenix Bioinformatics
Fremont, California 94538
ORCID ID: 0000-0001-5551-7232

Resource/Phoenix Bioinformatics
Fremont, California 94538
ORCID ID: 0000-0002-3837-8864
Andrew Farmer
National Center for Genome Resources
Santa Fe, New Mexico 87505
ORCID ID: 0000-0002-4224-2433
Christopher Town
J. Craig Venter Institute
Rockville, Maryland 20850
ORCID ID: 0000-0003-4653-4262
Nicholas Provart
Bio-Analytic Resource for Plant Biology, Department of Cell and Systems Biology/Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario M5S 3B2, Canada
ORCID ID: 0000-0001-5551-7232
Ashburner, M., et al. (2000) Gene Ontology: Tool for the unification of biology. Nat. Genet. 25: 25–29.
Berardini, T.Z., et al. (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol. 135: 745–755.
Brady, S.M., and Provart, N.J. (2009). Web-queryable large-scale data sets for hypothesis generation in plant biology. Plant Cell 21: 1034–1051.
Chatr-Aryamontri, A., et al. (2017) The BioGRID interaction database: 2017 update. Nucleic Acids Res. 45: D369–D379.
Cheng, C.-Y., Krishnakumar, V., Chan, A., Schobel, S., and Town, C.D. (2016). Araport11: A complete reannotation of the Arabidopsis thaliana reference genome. bioRxiv 047308.
Cleary, A., and Farmer, A. (2018). Genome Context Viewer: Visual exploration of multiple annotated genomes using microsynteny. Bioinformatics 34: 1562–1564.
Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. 40: D1178–D1186.
Hofmeister, B.T., and Schmitz, R.J. (2018). Enhanced JBrowse plugins for epigenomics data visualization. BMC Bioinformatics 19: 159.
Huala, E., et al. (2001) The Arabidopsis Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res. 29: 102–105.
International Arabidopsis Informatics Consortium (2012). Taking the next step: Building an Arabidopsis information portal. Plant Cell 24: 2248–2256.
Jiao, W.-B., and Schneeberger, K. (2020). Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat. Commun. 11: 989.
Kerrien, S., et al. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res. 40: D841–D846.
Lee, T.A., and Bailey-Serres, J. (2019). Integrative analysis from the epigenome to translatome uncovers patterns of dominant nuclear regulation during transient stress. Plant Cell 31: 2573–2595.
Magloft, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2007). Entrez Gene: Gene-centered information at NCBI. Nucleic Acids Res. 35: D26–D31.
Mi, H., Muruganujan, A., and Thomas, P.D. (2013). PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41: D377–D386.
Mitchell, A.L., et al. (2019) InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 47: D351–D360.
Okayasahi, T., Okamura, Y., Ito, S., Takada, S., Aoki, Y., Shirota, M., and Kinoshita, K. (2014). ATTED-II in 2014: Evaluation of gene coexpression in agriculturally important plants. Plant Cell Physiol. 55: e6.
Thieffry, A., Bornholdt, J., Ivanov, M., Brodersen, P., and Sandelin, A. (2019). Characterization of Arabidopsis thaliana promoter bidirectionality and antisense RNAs by depletion of nuclear RNA decay enzymes. bioRxiv 809194.
UniProt Consortium (2007). The Universal Protein Resource (UniProt). Nucleic Acids Res. 35: D193–D197.
Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., and Provart, N.J. (2007). An “Electronic Fluorescent Pictograph” browser for exploring and analyzing large-scale biological data sets. PLoS One 2: e718.
