# **Araport: the Arabidopsis Information Portal**

Vivek Krishnakumar<sup>1,\*</sup>, Matthew R. Hanlon<sup>2</sup>, Sergio Contrino<sup>3</sup>, Erik S. Ferlanti<sup>1</sup>, Svetlana Karamycheva<sup>1</sup>, Maria Kim<sup>1</sup>, Benjamin D. Rosen<sup>1</sup>, Chia-Yi Cheng<sup>1</sup>, Walter Moreira<sup>2</sup>, Stephen A. Mock<sup>2</sup>, Joseph Stubbs<sup>2</sup>, Julie M. Sullivan<sup>3</sup>, Konstantinos Krampis<sup>1</sup>, Jason R. Miller<sup>1</sup>, Gos Micklem<sup>3</sup>, Matthew Vaughn<sup>2</sup> and Christopher D. Town<sup>1</sup>

<sup>1</sup>Plant Genomics, J. Craig Venter Institute, Rockville, MD 20850, USA, <sup>2</sup>Texas Advanced Computing Center, The University of Texas, Austin, TX 78758, USA and <sup>3</sup>Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1QR, UK

Received September 19, 2014: Revised November 03, 2014: Accepted November 04, 2014

## ABSTRACT

The Arabidopsis Information Portal (https://www. araport.org) is a new online resource for plant biology research. It houses the Arabidopsis thaliana genome sequence and associated annotation. It was conceived as a framework that allows the research community to develop and release 'modules' that integrate, analyze and visualize Arabidopsis data that may reside at remote sites. The current implementation provides an indexed database of core genomic information. These data are made available through feature-rich web applications that provide search, data mining, and genome browser functionality, and also by bulk download and web services. Araport uses software from the InterMine and JBrowse projects to expose curated data from TAIR, GO, BAR, EBI, UniProt, PubMed and EPIC CoGe. The site also hosts 'science apps,' developed as prototypes for community modules that use dynamic web pages to present data obtained on-demand from third-party servers via RESTful web services. Designed for sustainability, the Arabidopsis Information Portal strategy exploits existing scientific computing infrastructure, adopts a practical mixture of data integration technologies and encourages collaborative enhancement of the resource by its user community.

## INTRODUCTION

The flowering plant Arabidopsis thaliana is a model organism for molecular, cellular and systems biology. Arabidopsis offers the favorable characteristics of short lifecycle, prolific seed production and genomic transformation efficiency. The A. thaliana Col-0 ecotype offers a nearly complete genome sequence assembly and high-quality gene annotation, along with many informatic and experimental resources. Other ecotypes and mutant lines, available in seed form, have been genotypically and phenotypically characterized. For over a decade, The Arabidopsis Information Resource (TAIR) served as the collector, curator and provider of pertinent information centered on the reference Col-0 genome. In 2009, the Arabidopsis community was faced with a growing number of online resources and potential loss of its central organizing resource, TAIR. After a series of workshops, members of the Arabidopsis research community published a white paper calling for a new Arabidopsis Information Portal (AIP) (1). The portal would assume responsibility for curation of the genome sequence and gene annotation. Intensive human curation of gene function, literature association etc. is not part of AIP's funded mandate. Rather, it will use an extensible architecture that will permit growth through contributions of community-curated data-centric modules embracing a wide range of data types. With NSF and BBSRC funding, we presented in 2014 a first functional release of Araport (ICAR 2014 proceedings).

Following its termination of funding and in parallel with the development of the AIP, TAIR began to require subscription fees in 2014, allowing it to continue its gene-centric data curation services (ICAR 2014 proceedings) that will complement the work of AIP. The AIP project does, and will continue to, incorporate all the pertinent data that TAIR makes public on a time-delayed schedule. Future Araport releases could act as a gateway by providing access to additional data for registered users of TAIR and other subscription-based services.

The AIP is freely available online at https://www.araport. org. It offers two core web applications: ThaleMine for data mining and JBrowse for genome browsing. It hosts an embryonic collection of 'science apps' that should serve as prototypes for community-contributed modules. It offers a reproducible and extensible framework, based on data feder-

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.

<sup>\*</sup>To whom correspondence should be addressed. Tel: +1 301 795 7363; Fax: +1 301 795 7070; Email: vkrishna@jcvi.org Present address: Konstantinos Krampis, Biological Sciences, Hunter College at City University of New York, New York, NY 10065, USA.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

ation and community-generated content, that should allow it to serve the growing and changing Arabidopsis fields of omics research.

#### **CORE FUNCTIONALITY**

The core of AIP is a database of basic genomic data within a data mining web application called ThaleMine. ThaleMine is a customized instance of the InterMine software (2), which provides parsers for many data sources, rapid retrieval of indexed data, dynamic table presentation, saved and shared list functionality, and web services (3) support. Currently, ThaleMine includes the TAIR 10 Col-0 sequence and annotation dataset (4). Assimilation of other data types (polymorphisms, phenotypes, stocks, etc.) is in progress. ThaleMine integrates data from UniProtKB (5), PubMed (6), Bio-Analytic Resource (BAR) (7), PANTHER (8), Sequence Ontology (9) and Gene Ontology (10). ThaleMine links to sources including PhytoMine implemented by Phytozome (11), FlyMine (12) and YeastMine (13) (Table 1). ThaleMine organizes 33 602 genes (including transposable element and pseudogenes), 41 671 transcripts, 31 189 transposable elements, 47 698 proteins, 16 127 publications, 125 expression studies, 6365 protein domains, 197 625 interactions and 5323 gene ontology terms. AIP loaded ThaleMine with an extensible set of template-based queries, which provide simple web interfaces to common kinds of searches (Table 2). The application supports free-text search as well as structured queries. For instance, searches can return lists of loci with specified attributes within the A. thaliana Col-0 reference genome sequence. List analysis is facilitated by widgets that generate column-wise statistics, identify overrepresented terms or publication links or compare lists by set-union and set-intersection. Capability to drill down by locus is facilitated by report pages for genes, transcripts and proteins. These pages display gene structure, functional annotation, protein domains, protein interactions, associated publications, expression patterns, co-expressed genes, orthologs in other species and links to other resources. Locus-specific pages also display an embedded BAR Electronic Fluorescent Pictograph (eFP) viewer (14) (Figure 1). The ThaleMine User Guide (https://www.araport.org/ thalemine/user-guide) has more details.

AIP hosts a customized instance of JBrowse (15) that, like InterMine, is actively supported by the GMOD consortium (http://www.gmod.org). The browser integrates gene structure, various kinds of evidence and genomic attributes, expression data, and more. Below the reference sequence axis, the browser presents a stack of 'tracks' representing, for example, the TAIR 10 gene models. The browser currently incorporates 20 tracks, including datasets such as pseudogenes, expression, non-coding RNAs, epigenetics and T-DNA insertion lines, which are presented to the user via the hierarchical track selector (Figure 2A). Additionally, 28 more tracks (not in the hierarchical selector) are obtained on-demand from EPIC CoGe (16) via web services, which are cataloged within the faceted track selector (Figure 2B). We modified JBrowse to enable the hierarchical and faceted track selectors to co-exist. This allows users to select from our 20 standard tracks and also our 28 metadata-rich tracks organized by user-set facets of their metadata. The browser also offers bulk download of interval-specific track data in common file formats (FASTA, GFF3, BED). Similarly, it can integrate user-provided datasets, either local or hosted on third party servers (i.e. displaying tracks of read mappings without transmitting entire files, by exploiting indexed BAM files) without data transmission to AIP servers (Figure 2C). While searching or scrolling, the page URL is dynamically updated with sufficient information to fully reproduce the display. This feature enables collaboration by URL sharing between remote users.

#### EXTENSIBLE FUNCTIONALITY

The AIP design relies on data federation and software modularity for scalability and sustainability. The AIP module concept consists of a visual front end to an externally provided database or computational resource. AIP encourages the Arabidopsis research community to contribute modules as a means of disseminating their results. AIP provides the framework for developing, hosting and integrating modules as well as software and training materials. The AIP software development kit (SDK) provides boilerplate code (as an example it deploys a 'Hello World' app, the classic first task for programmers), serving as a starting point for developers. The SDK was built with node.js (http://nodejs.org), grunt (http://gruntjs.com) and yeoman (http://yeoman.io) software. Using this SDK, AIP has developed and hosts several demonstration modules. One is a BLAST (17) module offering homology search and sequence alignment visualization. Another provides graphical and tabular viewers for Arabidopsis protein-protein interactions, for which the data are obtained on demand from EBI IntAct web services (18) and the visualization relies on BioJS (19) and Cytoscape (20) open-source software. AIP science apps can access web services directly or through the AIP middleware. They can access AIP-hosted web services that expose the data, metadata and indexes associated with ThaleMine and JBrowse at AIP.

AIP hosts community information. This includes notices of news regarding the AIP project, postings syndicated from GARNet (http://www.garnetcommunity.org. uk), conferences/events and job openings. AIP has a forum for posting, and answering, questions to AIP staff and the community. AIP offers files for bulk download including the latest public TAIR release of whole genome sequence (FASTA), gene structure annotation (GFF3) and other supporting datasets.

## ARCHITECTURE

With the goal of providing the Arabidopsis research community with a sustainable resource, Araport is designed to be scalable and reproducible. For scalability, its organizing principle is federation. AIP hosts the middleware and services that enable the development and deployment of community-contributed science apps. As enablers, AIP hosts the two data-centric applications, ThaleMine and JBrowse, which provide data services, visualization tools and optional launch points for third-party modules. For reproducibility, AIP maintains public repositories hosted on GitHub (https://www.github.com), tracking the data, configuration and software code that are required to recreate

AGI I Nam Brief Com Cura	e " F Description putational Descr stor Summary " nyms: ATG	AT3C54 glutathi glutathi stroma, (InterPro oxidore Arabido	i660 one reductase one reductase (G chloroplast; EXP o:IPROL3027), FA ductase, NAD-bi psis thaliana pro	R): FUNCTIONS IN RESSED IN: 24 pla D/NAD-linked res nding region Unte tein match is: glu	: glutathione-d nt structures; E) Juctase, dimeris Pro:IPR001327 tathione-disulfi	(PRESSED DURING: ation (InterPro:IPR)	activity, copper ior 14 growth stages 016156), Pyridine Ilphide reductase AT3G24170.3), 1	CONTAINS Inter nucleotide-disul InterPro: IPRO06 Fource: TAIR, Au	rPro DOMAIN Iphide oxidor 324), Pyridine	Search: e. LVED IN: glutathione //s: FAD-dependent //s: GAD-dependent eductase, class I, act e nucleotide-disulphi	9 ATIG01640 metabolic proces lyridine nucleotii ive site (interPro	de-disulphide oxid (IPR012999), Pyridi	oreductase ne nucleotide-dis	ulphide
Quid	ck Links:			Summary C	Genomics	Proteins F	unction Ex	pression	Interactio	ons Homolog	y Other			
Genomic	cs										(	Function	S	
<sup>D</sup> Gene mode	lels - GR AT30 els											<ul> <li>Gene Onto biological p</li> </ul>		
Transcripts: Transcript		rons: 10 CD5s: 1 ons				introns				CDSs			netabolic process	0 I
		R exon:1 AT3G5466			483 [FASTA]	intron_Chr3_20230	11020230305	15	96 [FASTA]			molecular fi	inction	
		R exon:2 AT3G5466 R exon:3 AT3G5466			226 FASTA	intron_Chr3_20230			09 FASTA			NADP bindin		nding " I I I
	GF	CR:exon:4 AT3G54660:exon:4			177 [FASTA] intron_Chr3_2023064							ATP binding " copper ion binding "		
CR AT3G5461 3439 FASTA	60.1	Riexon:5 AT3G5466 Riexon:6 AT3G5466			54 FASTA 149 Fasta			08 165 FASTA. AT3C54660.1_CD5				glutathione-disulfide reductase activity *		
SHOP PASTAL	GP	Rexon:7 AT3G5466	50:exon: 7		64 FASTA	intron_Chr3_20231 intron_Chr3_20231			29 FASTA 68 FASTA	1698 FASTA		mitochandrion P		
		Riexon:8 AT3G5466 Riexon:9 AT3G5466			171 FASTA	intron_Chr3_20231	886.20231970	1	85 FASTA			chloroplast stroma <sup>III</sup> III chloroplast <sup>III</sup> III		
	GR	R exon:10 AT3G546	560:exon.10		92 [FASTA]	intron_Chr3_20232 intron_Chr3_20232			85 FASTA					
	CF	Rexon:11 AT3G546	560:exon:11		366 FASTA	Diazza e Diesendaria		0				3		
Proteins	R)								Interactio	ons				
1 Proteins D8 identifie		Primary Accessio		Organism			ngth		Interaction	n Network				
GSHRP_ARA1		P42770		Arabidopsi										
Homologu Homologu	les	D malanamica	D ratio		станования 16 — 110 —	needor 9 mm-*	55 FASTA			TON CHIRBLE OF	29-4 (RX3)	TOTSA ATT	Arec 19690	
Homologu A thaliana NTRA ATMDAR2 mIL/PO1 GR1 GR1 GR1 AT4G16155 NTR8 MOAR1 LPD1 mIL/PD2	les	F8gn0031392 F8gn0032846 F8gn0032754 F8gn0036762 F8gn0025628	208-GENE-04013 208-GENE-03083 208-GENE-05053 208-GENE-04043 208-GENE-0409	26-11 ENSCOOD 22-116 ENSCOOD 26-1732 ENSCOOD 14-66 ENSCOOD 12-2282 ENSCOOD 18-96 ENSCOOD	00198431 MGI: 00197763 MGI: 00121850 MGI: 00091140 MGI: 00184470 MGI:	1919418 RGD:7350 19386711 RGD:6196 107450 RGD:6195 107450 RGD:6217 1676395 RGD:1306 1347023 RGD:6208 1349419	rus S. cerevisiae 5 5000002761 5 5000005328 7 5000005328 22 500001148 1 5 500001148			C. (	(RX2	Tencia	Henrices Hanning Hanni	>- <u>\</u> +
Homologu A thaliana ( NTRA ATMDAR2 mIL/701 GR1 ( AT4G16155 ) NTRB ( MDAR4 MDAR4 MDAR4	VBGene00014025 WBGene0004012 WBGene00008117 WBGene00010754 WBGene0001754 WBGene0001754 WBGene0007744	FBgn0037170 FBgn0031392 FBgn0032846 FBgn0032754 FBgn0036762 FBgn0025628	208-GENE-04011 208-GENE-03081 208-GENE-05051 208-GENE-04040 208-GENE-04040 208-GENE-07011 208-GENE-0911	20-4 ENSCOOD 26-11 ENSCOOD 22-116 ENSCOOD 26-1732 ENSCOOD 26-1732 ENSCOOD 12-2282 ENSCOOD 12-2282 ENSCOOD	00198431 MGI: 00197763 MGI: 00121350 MGI: 00091140 MGI: 00184470 MGI: 00156709 MGI: 00104687 MGI:	5804 919418 8CD:6196 186711 8CD:6196 8CD:6195 8CD:6295 8CD:6217 8CD:6217 8CD:6217 8CD:6205 8CD:705	rus S. cerevisiae 73 0 \$000002761 9 \$000005328 7 \$000005328 22 \$000001148 17 \$00001148			C. (	(RX2	Tencia	Henrices Hanning Hanni	>- <u>\</u> +
Homologu A thaliana ATMDAR2 miUP01 GRI AT4C1615151 NTR8 MDAR1 LPD1 miLP02 MDAR6 MDAR6 MDAR6 MDAR6 MDAR6 MDAR6	VBGene00014025 WBGene0004012 WBGene00008117 WBGene00010754 WBGene0001754 WBGene0001754 WBGene0007744	F8gn0037170 F8gn0031502 F8gn0032846 F8gn0032754 F8gn0037552 F8gn0025558 F8gn0025553	208-GENE-04011 208-GENE-03081 208-GENE-05051 208-GENE-04040 208-GENE-04040 208-GENE-07011 208-GENE-0911	20-4 ENSCOOD 26-11 ENSCOOD 2-116 ENSCOOD 26-1732 ENSCOOD 12-2282 ENSCOOD 22-2282 ENSCOOD 27-2 ENSCOOD	00198431 MGI: 00197763 MGI: 00121350 MGI: 00091140 MGI: 00184470 MGI: 00156709 MGI: 00104687 MGI:	5804 919418 8CD:6196 186711 8CD:6196 8CD:6195 8CD:6295 8CD:6217 8CD:6217 8CD:6217 8CD:6205 8CD:705	rus S. cerevisiae 73 0 \$000002761 9 \$000005328 7 \$000005328 22 \$000001148 17 \$00001148	.0*@	0 * 2	C. (	TRUE GOPTE TEOD AT BOARD AT BO		Henrices Hanning Hanni	>- <u>\</u>
Homologu A thaliana ATMDAR2 miUP01 GRI AT4C1615151 NTR8 MDAR1 LPD1 miLP02 MDAR6 MDAR6 MDAR6 MDAR6 MDAR6 MDAR6	Pes C. elegans WBConedD014825 WBConedD001553 WBConedD001553 WBConedD0019744 WBConedD0017740 WBConedD007744	F8gn0037170 F8gn0031502 F8gn0032846 F8gn0032754 F8gn0037552 F8gn0025558 F8gn0025553	208-GDNE-0401 208-GENE-0303 208-GENE-0503 208-GENE-0404 208-GENE-0404 208-GENE-0701 208-GENE-0701 208-GENE-0703	20-4 ENSCOOD 26-11 ENSCOOD 2-116 ENSCOOD 26-1732 ENSCOOD 12-2282 ENSCOOD 22-2282 ENSCOOD 27-2 ENSCOOD	00198431 MCI: 0019763 MCI: 0019763 MCI: 0019764 MCI: 00194647 MCI: 00164687 MCI: 00164687 MCI: 00164687 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00180000000000000000000000000000000000	KOD 7358     KOD 7358     KOD 7358     KOD 5196     KOD 519     KOD 5	ru         S. cerevisiae           73         \$000002711           90000012         \$00000121           900000012         \$00000121           900000012         \$00000121           20000012         \$00000127           20000012         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$000000127         \$00000127           \$000000127         \$00000127           \$0000000127         \$00000127           \$000000000000000000000000000000000000	Sample Description	Controls Name				Henrices Hanning Hanni	
Homologu           A. Idaliana           NITRA           ATMOAR2           miUP01           GR1           ATAGAR2           MUR01           RR1           NTRA           MUR01           RR1           MUR02           MUR04           MUR05           MUR04           MUR04	PES         C. elegans           WBCene00014028         WBCene0000517           WBCene00001573         WBCene00005744           WBCene00007744         WBCene00007744           iion              • @ • @ ]            Experiment -	Regn0037170     Regn0031302     Regn0032546     Regn0032554     Regn0025553     Regn0025553     regn0025553     Common sector sec	208-GINE-0401 208-GINE-0208 208-GINE-0208 208-GINE-0404 208-GINE-0404 208-GINE-0701 208-GINE-0700 20	20-4 ENSCODD 20-11 ENSCODD 20-113 ENSCODD 20-1132 ENSCODD 14-66 ENSCODD 14-2322 ENSCODD 14-2328 ENSCO	00198431 MCI: 0019763 MCI: 0019763 MCI: 0019764 MCI: 00194647 MCI: 00164687 MCI: 00164687 MCI: 00164687 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00183773 MCI: 00180000000000000000000000000000000000	55804 RCD.7350 9393438 RCD.7350 870718 RCD.6355 970750 RCD.6355 970752 RCD.1306 334419 RCD.7308 3354175 RCD.1303 3354175 RCD.1303	ru         S. cerevisiae           73         \$000002711           90000012         \$00000121           900000012         \$00000121           900000012         \$00000121           20000012         \$00000127           20000012         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$00000127         \$00000127           \$000000127         \$00000127           \$000000127         \$00000127           \$0000000127         \$00000127           \$000000000000000000000000000000000000	Sample Description	Controls Name		TRUE GOPTE TEOD AT BOARD AT BO		Henrices Henrices (Hit) (Hit) (Hit)	
Homologu           A. thatiana           NTRA           ATMORE2           uturo1           qaractists           NTRA           MDARE           ATSHARE	es C. elegans WEGene0b034028 WEGene0b0353 WEGene0b0353 WEGene0b0374 WEGene0b0374 WEGene0b00374 WEGene0b003744 Common State Separiment - Tible Separiment - ABA time course in withppe seddings	F8gn0037170 F18gn003192 F18gn003192546 F18gn0032555 F18gn0020653 F18gn0020653 F18gn0020653 F18gn0020653	208-GENE-0401 208-GENE-0308 208-GENE-0308 208-GENE-0404 208-GENE-0404 208-GENE-0701 20	20-4 EHSCODD 20-11 EHSCODD 20-116 EHSCODD 20-116 EHSCODD 8-96 EHSCODD 8-96 EHSCODD 77-2 EHSCODD 77-2 EHSCODD 20-100 EHSCODD 20	00195431 MG-1 00197753 MG-1 00123550 MG-1 00123550 MG-1 00154747 MG-1 0010146477 MG-1 00101477 MG-1 000000000000000000000000000000000000	KOD 7358     KOD 7358     KOD 7358     KOD 5196     KOD 519     KOD 5	ru         S. cerevisiae           73         600002711           90000612         90006012           90000612         90006012           90000612         90006012           20000612         90006012           20000612         90006012           20000612         90006012           20000612         90006012           20000012         90000012	Sample - Description	Controls Name 1 2 Samples h III 2 Samples		TRUE GOPTE TEOD AT BOARD AT BO		Henrices Henrices (Hit) (Hit) (Hit)	
Homologu           A. thatiana           NTSA           ATMORE2           util 01           GATGAISTS           NTRA           MDARE           ATSHARE	es Es C. elegans WEGene01003112 WEGene010031 WEGene0101734 WEGene0101734 WEGene0101744 WEGene01007744 ion Caparison ABA time course in withppe seddings ABA time course in withppe seddings	Hgen031720           Figu031322           Figu032254           Hgen032554           Hgen032554           Hgen032554           Hgen032553           Demensions -           Signal           424,24           432,54	208-CDK4-0401. 208-CBK4-0308. 208-CBK4-0308. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0409. 208-CB	20-4 ProColo 2-116 ProColo 2-116 ProColo 2-116 ProColo 2-116 ProColo 2-216 ProColo 2-228	00194431 MGri 00197763 MGri 00197763 MGri 00193146 MGri 00193146 MGri 00185793 MGri 00185773 MGri 001857753 MGri 00185773 MGri 00185775 MGri 00195755 MGri 00185755 MGri 001857555 MGri 0018575555555555555555555555555555555555	S604     ACD_735     ACD	Kux         S. cerevisiae           73         000002711           900000271         900000271           9000000127         9000001271           9000000127         900001271           9000000127         900001271           900000127         900001271	Sample Description 10uM ABA for 3h 10uM ABA for 1h 10uM ABA for 30min	Controls Name 1 2 Samples 1 2 Samples 2 3 Samples	Control of the second s	TRUE GOPTE TEOD AT BOARD AT BO		Henrices Henrices (Hit) (Hit) (Hit)	
Нотолоди А. thatiana МТА АТМОА22 тиП201 ССС ССС ССС ССС ССС ССС ССС С	es         C. elegans           WEGene0b014028         WEGene0b00174           WEGene0b01754         WEGene0b01754           WEGene0b01754         WEGene0b01754           WEGene0b01754         WEGene0b01754           WEGene0b01754         WEGene0b01754           ABA Rescuedon1764         TELE           BAR Rescuedon1764	Herroll 7170           Fleynoll 1322           Fleynoll 1322           Herroll 7170           Fleynoll 2324           Herroll 702           Fleynoll 2524           Herroll 702           Fleynoll 702           Fleynoll 702           Fleynoll 702           Fleynoll 702           Signal           424.24	208-CDK4-0401. 208-CBK4-0308. 208-CBK4-0308. 208-CBK4-0309. 208-CBK4-0701. 208-CDK4-0701. 208-CDK4-0701. 208-CDK4-0303. 208-CDK4-0303. 208-CDK4-0303. 208-CDK4-0303. 208-CDK4-0303. 208-CDK4-0303. 208-CDK4-0303. 208-CDK4-0401. 208-CDK4-0400. 208-CDK4-0400. 208-CDK4-0400. 208-CDK4-0400. 208-CD	20-4 EPGC000 20-11 EPGC000 20-116 EPGC00 20-216 EPGC00 20-22 EPGC00	00194431 MGri 00197763 MGri 00197763 MGri 00193146 MGri 00193146 MGri 00185793 MGri 00185773 MGri 001857753 MGri 00185773 MGri 00185775 MGri 00195755 MGri 00185755 MGri 001857555 MGri 0018575555555555555555555555555555555555	S604     ACD_735     ACD	Kus S. cerevisiae     73     5     5000002761     9     500000013     9     500000013     9     500001876     17     5100001876     170_10uM-ABA.3     176_10uM-ABA.3     176_10uM-ABA.3	Sample Description 10uM ABA for 3h 10uM ABA for 1h 10uM ABA for 30min	Controls Name 1 2 Samples 1 2 Samples 2 3 Samples	Canal of Can	TRUE GOPTE TEOD AT BOARD AT BO		Henrices Henrices (Hit) (Hit) (Hit)	
Homologu           A. thatiana           NTRA           ATMORA2           mitro1           ATAGAR2           NTRA           ATMORA2           MURA1           MURA1           NTRA           MURA1           MURA1           NTRC           NTRA           MURA1           MURA2           NTRC           MURA1           MURA2           NTRC           MURA2           NTRC           MURA2           NTRC           ATMORAN	es C. elegans WEGene010014228 WEGene01001734 WEGene0101734 W	Hgen001720           Figu0011822           Figu0012754           Hgen001756           Figu0012754           Hgen0016762           Figu0012764           Hgen0016762           Figu0012764           Hgen0016762           Figu0016762           Figu0016763           Signal           424,24           422,54	208-CDK4-0401. 208-CBK4-0308. 208-CBK4-0308. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0409. 208-CB	20-4 ProColo 2-116 ProColo 2-116 ProColo 2-116 ProColo 2-116 ProColo 2-216 ProColo 2-228	00194431 MGri 00197763 MGri 00197763 MGri 00193146 MGri 00193146 MGri 00185793 MGri 00185773 MGri 001857753 MGri 00185773 MGri 00185775 MGri 00195755 MGri 00185755 MGri 001857555 MGri 0018575555555555555555555555555555555555	S604     ACD_735     ACD	Kux         S. cerevisiae           73         000002711           900000271         900000271           9000000127         9000001271           9000000127         900001271           9000000127         900001271           900000127         900001271	Sample Description 10uM ABA for 3h 10uM ABA for 1h 10uM ABA for 30min	Controls Name 2 Samples 1 2 Samples 1 2 Samples 1 1 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3	Carrier of the second s				
Нотолоди А. thatiana МТА АТМОА22 МТОА22 ПССССССССССССССССССССССССССССССССССС	es C. elegans WEGene010014228 WEGene01001734 WEGene0101734 W	Hgen001720           Figu0011822           Figu0012754           Hgen001756           Figu0012754           Hgen0016762           Figu0012764           Hgen0016762           Figu0012764           Hgen0016762           Figu0016762           Figu0016763           Signal           424,24           422,54	208-CDK4-0401. 208-CBK4-0308. 208-CBK4-0308. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0309. 208-CBK4-0409. 208-CB	20-4 ProColo 20-14 ProColo 2-116 ProColo 2-116 ProColo 2-216 ProColo 2-216 ProColo 2-228	00194431 MGri 00197763 MGri 00197763 MGri 00193146 MGri 00193146 MGri 00185793 MGri 00185773 MGri 001857753 MGri 00185773 MGri 00185775 MGri 00195755 MGri 00185755 MGri 001857555 MGri 0018575555555555555555555555555555555555	S604     ACD_735     ACD	Kux         S. cerevisiae           73         000002711           900000271         900000271           9000000127         9000001271           9000000127         900001271           9000000127         900001271           900000127         900001271	Sample Description 10uM ABA for 3h 10uM ABA for 1h 10uM ABA for 30min	Controls Name Name Samples Di 2 Samples Di 2 Samples Di 2 Samples Di 2 Samples	Cange of Canada		or trees a		
Homologu           A. thatiana           NTRA           ATMORA2           mitrol           ATAGAR2           NTRA           ATMORA2           MURA1           MURA1           NTRA           MURA1           MURA1           MURA1           MURA1           MURA2           NTRC           MURA1           MURA2           NTRC           MURA1           MURA2           NTRC           MURA2           NTRC           MURA2           NTRC           MURA2           NTRC           MURA2           NTRC           ATJOS4660           ATJOS4660	es Es C. elegans WEGene000514228 WEGene0005317 WEGene0001794 WEGene0000794 WEGene00007746 ion Dependent - 0 * 0 Dependent - 10 * 0 ABA fem course in withpe sedSings ABA fem course in withpe sedSings ABA fem course in withpe sedSings aABA fem course in withpe sedSings aBA fem course in withpe sedSings aBA fem course in withpe and a fem course in withpe and fem course in the fe	Itegen017170           Flig=n0217170           Flig=n0217170           Flig=n0217170           Flig=n0217170           Flig=n0217170           Flig=n0217553           Flig=n0220553           Signal           366.31           424.24           435.44	208-GN4-0401. 208-GN4-0505. 208-GN4-0505. 208-GN4-0505. 208-GN4-0505. 208-GN4-0501. 208-GN4-0501. 208-GN4-0501. 208-GN4-0501. 208-GN4-0505. 20	20-4 PR5C00 0-113 PR5C00 0-172 PR5C00 0-172 PR5C00 0-172 PR5C00 0-172 PR5C00 0-172 PR5C00 0-172 PR5C00 2-232 PR5C00 2-	00199431 MG1 00197763 MG1 00197763 MG1 001913470 MG1 0018470 MG1 0018470 MG1 0018470 MG1 0018470 MG1 00183773 MG1 0019000000000000000000000000000000000	S604     ACD_735     ACD	Kux         S. cerevisiae           73         500002761           73         500002761           74         5000002761           75         5000012761           76         5000012876           253         5000012876           253         5000012876           176         10u/A ABA-3           176         10u/A ABA-1           176         10u/A ABA-1           176         10u/A ABA-1	Sample Description 10ukl ABA for 3h 10ukl ABA for 1h 10ukl ABA for 1h 10ukl ABA for 1h 10ukl ABA for 1h	Controls Name 2 Samples 1 2 Samples 1 2 Samples 1 1 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3	Cange of Canada				
Homologu           A thatiana           NTRA           ATMORE2           mUV01           GAT           MORR1           LP01           NTRA           MORR1           MORR1           MORR1           NTRC           MORR1           MORR1           MORR1           MORR1           MORR2           NTRC           MORR1           MORR1           MORR2           NTRC           MORR1           MORR2           NTRC           MORR1           NTRC           MORR2           NTRC           MORR2           ATTOGE           AGI ID           ATTOGE           ATTOGE	es C. elegans WEGene00014028 WEGene0001734 WEGene0001734 WEGene0001734 WEGene0001734 WEGene00007746 icon Con Con Con Con Con Con Con C	Itegen017170           Flig=n0217170           Flig=n0217170           Flig=n0217170           Flig=n0217170           Flig=n0217170           Flig=n0217553           Flig=n0220553           Signal           366.31           424.24           435.44	208-GN4-0401. 208-GN4-0505. 208-GN4-0505. 208-GN4-0505. 208-GN4-0505. 208-GN4-0501. 208-GN4-0501. 208-GN4-0501. 208-GN4-0501. 208-GN4-0505. 30gnat 441.05 457.3 431.84	20-4 PR5C00 2-111 PR5C00 2-111 PR5C00 2-111 PR5C00 2-111 PR5C00 2-112 PR5C00 2-232 PR5C00 2-232 PR5C00 2-323 PR5C00 2-324 PR5C00 2-	00199431 MG1 0019743 MG1 0019743 MG1 0013135 MG1 0013135 MG1 00131470 MG1 0013173 MG1 0013	CO 350     CO 450     CO 450     SO 45	Kux         S. cerevisiae           73         S00002761           S00002761         S00002761           7         S000005261           7         S000005261           7         S000005261           7         S000005261           7         S000005261           7         S000001876           2         O ● (1)           Simple -         Replication           176_10u/M_ABA-1         176_10u/M_ABA-1           176_10u/M_ABA-1         176_10u/M_ABA-1           176_10u/M_ABA-1         2004	Sample Description 10ukl ABA for 3h 10ukl ABA for 3h 10ukl ABA for 1h 10ukl ABA for 1h	Controls Name Name Samples M El 2 Samples M El 2 Samples M 2 Samples M El 2 Samples M El 2 Samples	Constant of the second se				* Drawn a law de
Homologu           A thatiana of NTRA           NTRA           ATMORA2           mtUP01           Galiana of NTRA           MOARI           ATJOGAGO           ATJOGAGOO           ATJOGAGOO           S Publica           First Autho	es Es C. elegans WEGene000514028 WEGene0005517 WEGene0001794 WEGene0001794 WEGene00007746 ion Content	Flepro21777     Flepro22777     Flepro2246     Flepro2246     Flepro2246     Flepro2246     Flepro24762     Flepro24762     Flepro24762     Signal     424.24      424.24      424.24	208-CIN4 -0401. 208-CIN4 -0405. 208-CIN4 -0305. 208-CIN4 -0305	20-4 (PSC000 2-11) (PSC000 2-11) (PSC000 2-11) (PSC000 2-11) (PSC000 2-11) (PSC000 2-11) (PSC000 2-232) (PSC000 2-232	00199431 MG1 00197763 MG1 00197763 MG1 00197763 MG1 00131370 MG1 00138470 MG1 00138470 MG1 00138470 MG1 00138477 MG1 00138773 MG1 00138773 MG1 00138773 MG1 00138773 MG1 0013973 0.97 0.97 1 0.97	CO 350     CO 450     CO 450     SO 45	Kux         S. cerevisiae           73         500002761           73         5000002761           74         50000002761           75         50000002761           75         500000002761           76         5000000000000000000000000000000000000	Sampa . Description 10ukt ABA tor 3h 10ukt ABA tor 3h 10ukt ABA tor 1h 10ukt ABA tor 1h 10ukt ABA tor 1h 10ukt ABA for 1h 10ukt ABA for 1h 10ukt ABA for 1h	Controls Name Name Samples M El 2 Samples M El 2 Samples M El 2 Samples M El 2 Samples Volume 14	Carpie Carpie Carpie Carpie Properties E 6 SamplePropertys E 7 SamplePropertys E 6 SamplePropertys E 7 SamplePropertys E 7 Sa		Link to other N Extense D. melanogaser C. C. C. C. S.		* Diverse al bia de * Diverse al bia de
Homologu           A. thatiana           NTRA           ATMORA2           mtUP01           GR. TISISS           MDARI           ATJOGIO           ATJOGIO           ATJOGIO           S Publica           First Autho           Kieffmann 1           Pelter /B	es es C. elegans WEGene01014028 WEGene0101754	Piegeo27770           Piegeo27770           Piegeo23246           Piegeo23246           Piegeo232346           Piegeo232346           Piegeo26723           Piegoo26723           Piego	208-CIN4 -0401. 208-CIN4 -0405. 208-CIN4 -0405	20-4 (PSC000 2-11) (PSC000 2-11) (PSC000 2-11) (PSC000 2-11) (PSC000 2-11) (PSC000 2-11) (PSC000 2-232 (PSC000 2-232 (PSC000 2-232 (PSC000 2-232 (PSC000 2-232 (PSC000 2-232 (PSC000 2-232 (PSC000 2-332 (PSC0000 2-332 (PSC	00199431 MG1 00197763 MG1 00197763 MG1 00197763 MG1 00191470 MG1 0018470 MG1 0018477 MG1 0018477 MG1 0018477 MG1 0018477 MG1 0018477 MG1 0018477 MG1 0018477 MG1 0018477 MG1 0018477 MG1 001847 0008000000	CO	Kux         S. cerevisiae           73         500002761           74         5000002761           75         50000002761           76         50000002761           77         50000012761           78         50000012761           79         50000012761           70         5000012761           710         500001876           7176_10uM-ABA-3         176_10uM-ABA-1           7176_10uM-ABA-3         176_10uM-ABA-1           7176_10uM-ABA-1         2004           2006         ccloside         2006	Sampa - Description 10ukl ABA tor 3h 10ukl ABA tor 3h 10ukl ABA tor 1h 10ukl ABA tor 1h	Controls Name 1 ED 2 Samples 1 ED 2 Samples ED 2 Samples ED 2 Samples ED 2 Samples Volume 14 S	Cast		Coloros A AF		The second seco

**Figure 1.** A Gene report page showing: (**A**) attributes like standard gene locus identifier, symbols/synonyms, TAIR curator summary, confidence rating, etc. Other useful information is segregated into the aspects, as follows: (**B**) 'Genomics' showing gene structure information; (**C**) 'Function' displaying the Gene Ontology annotation; (**D**) 'Proteins' with links to the relevant protein records populated with data from UniProt; (**E**) 'Homology' listing the computed orthologs and paralogs across a diverse set of species; (**F**) 'Interactions' showing a visual representation of the genes' physical/genetic interactions (tabular format is also available); (**G**) 'Expression' reporting gene expression levels based on AtGenExpress project data, also visualized using an embedded eFP view; (**H**) 'Publications' reports papers associated with the current gene; (**I**) 'Links' to other InterMine databases where homologs of the current gene exist, and external links to several important Arabidopsis and plant genomics data providers.

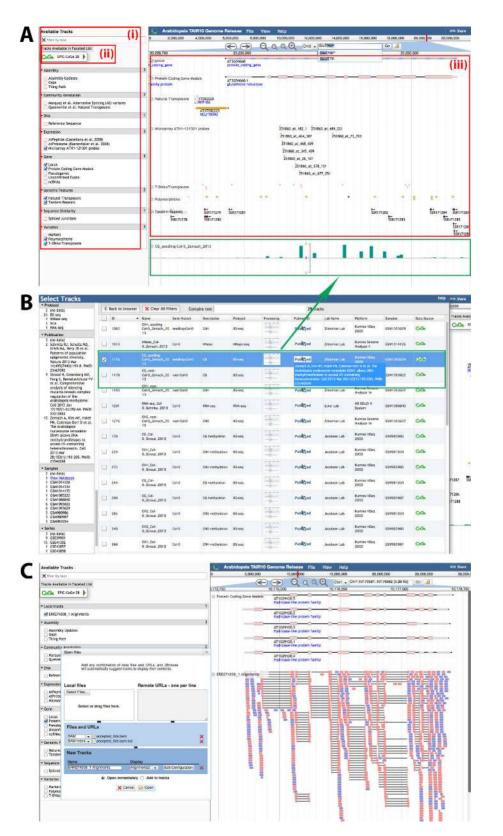


Figure 2. (A) (i) List of tracks available via the hierarchical track selector; (ii) a link to access the faceted track selector; (iii) track display panel showing the currently selected tracks for viewing (green inset: track chosen from the faceted selector). (B) Faceted track selector offers a search and filtering interface to enable choosing tracks based on the experiment metadata. (C) File upload capability ensures privacy to view ones own datasets alongside the Arabidopsis genome, and flexibility to work with large datasets such as short read alignments.

Table 1. Summary of the datasets used to populate ThaleMine and their corresponding sources

Dataset	Data source		
Genome sequence and annotation Protein sequence and properties Protein interactions Affymetrix expression data Electronic Fluorescent Pictographs Publications Orthologs	TAIR (version 10; 8/31/13) UniprotKB Bio-Analytic Resource (BAR) AtGenExpress via BAR Bio-Analytic Resource UniprotKB and NCBI Entrez PANTHER and PhytoMine		

Table 2. List of commonly executed searches provided as 'Template searches'

Input (accepts wildcard characters)	Output (reported as a table, exportable in different file formats)				
Gene (identifier/alias)	Related protein identifiers				
	List of interactors				
	Set of homologous genes				
	Expression values				
	Publications referencing this gene				
	FASTA sequences of the CDS				
	FASTA sequence of the proteins				
	FASTA sequences of the $5'/3'$ UTRs				
Protein (identifier/domain)	Related gene identifiers				
	Publications referencing this protein				
Ontology term	List of genes				

the Araport website. We already rebuild parts of Araport (e.g. ThaleMine) automatically from scripts in the repository. We aim to automatically deploy the complete site at two locations (JCVI, TACC) and possibly others.

The AIP core applications, ThaleMine and JBrowse, were both implemented by customizing open-source GMOD software. The core applications represent a mixture of the data federation and data hosting approaches to website development. Both applications use federation to obtain data on demand from remote sites. ThaleMine displays links to orthologs and eFP expression pattern images, both dependent on run-time data requests to third-party web services. ThaleMine and JBrowse not only consume but also expose web services. These web services have RESTful interfaces, as first described by Fielding (21). They are exposed through an AIP mediation layer in an effort to increase their uniformity, stability and ease of use (e.g. by precluding crossorigin resource sharing (CORS) security restrictions). Both ThaleMine and JBrowse also rely on AIP-hosted data, currently ~150 GB, sufficient to provide rapid displays and search responses. Hosted data include genome sequence; IDs and functional descriptions of genes, transcripts, proteins, domains and ontological terms; and metadata used to identify genome browser tracks. A portion of ThaleMine data is obtained from TAIR on a delayed-release schedule. Other data are downloaded from FTP sites and web services shortly before deploying the applications. A Chado database (22), with Tripal modules (23), is under construction. It will act as a data warehouse in support of future application deployments.

The Araport software employs a strongly decoupled three-tier design in which front end presentation software communicates through a middle layer to back end databases. This architecture was adopted to promote site sustainability by enabling future re-engineering of each tier independently. The AIP presentation layer uses HTML5 (http://www.w3.org/TR/html5/), CSS (http://www.w3.org/ Style/CSS/), JavaScript, AJAX and jQuery (http://jquery. com) delivered via HTTPS to provide a responsive and secure experience to clients using modern web browsers on any platform. AIP's middle layer employs the Apache Tomcat web server, InterMine software, the Drupal content management system, the iPlant Collaborative-developed Agave API platform (24) and OAuth2 (http://tools.ietf.org/ html/rfc6749) authentication. On the back end, AIP employs PostgreSQL (http://www.postgresql.org) databases and the iPlant Data Store (25) for data persistence. Modules of AIP communicate with databases via AIP's mediating layer that provides authentication and other services (logging, throttling, load balancing, data transformation, caching). Communications between front-to-middle layers and middle-to-back layers are implemented as web services. These exchange JSON-formatted data over the HTTPS protocol using RESTful URL standards being developed and documented at AIP.

#### **FUTURE DIRECTIONS**

As TAIR'S NSF funding ended, responsibility for the *A. thaliana* Col-0 reference genome and annotation passed to AIP. We plan to update the standard 'TAIR10' sequence and annotation with releases labeled 'AIP11', 'AIP12', etc. We are implementing an automated re-annotation pipeline that combines publicly available RNA-seq reads, their Trinity *de novo* and reference-guided assemblies (26) and PASA annotation comparison (27). One goal is to delineate tissue-specific transcript isoforms. We will solicit expert community review to be implemented with Tripal and WebApollo (28) software.

Araport is designed to respond to a recent whitepaper (1) that called for a sustainable community portal whose growth would depend on modular contributions from the Arabidopsis research community. Now entering its second year of funding, Araport provides two fully functional web applications for browsing and data mining A. thaliana genomics data. More importantly for growth, the AIP staff is working with the community to extend its documentation, middleware and SDKs that enable third-party module development and dissemination. By early November 2014, AIP will have hosted its first workshop for community developers. The workshop is expected to generate about a dozen modules addressing research areas including expression, regulation and epigenetics. Each module will offer dynamic web pages providing visualization of remote data that is accessed on demand through RESTful web services. This and other workshops should enhance the repertoire of available modules, increase community involvement, lower the technical hurdles for contribution, and eventually enable increasing levels of integration and sophistication of module design. AIP will continue to enhance its presentation layer to aid users in the discovery, use and combination of modules. AIP will integrate third-party modules into its core applications and simultaneously promote the integration of core application web services and visualizations within third-party modules. This federated model of data sharing and software development should help AIP continuously adapt to the omics data explosion and the resulting needs of the broader Arabidopsis user community.

## ACKNOWLEDGEMENTS

We would like to thank Eva Huala and Bob Muller (Phoenix Bioinformatics, formerly TAIR) for their assistance with the migration of *Arabidopsis* data to AIP; summer interns Eleanor Pence and Jane Smitham for their contributions; Nicholas Provart and the Bio-Analytic Resource group for providing eFP web services and databases of expression and protein interaction data; Eric Lyons and the CoGe team for their assistance with integrating EPIC-CoGe data; the iPlant Collaborative for assistance in developing BAM file streaming; Joe Carlson and David Goodstein for assistance with integrating Phytozome's PhytoMine.

## FUNDING

National Science Foundation [DBI-1262414 to C.D.T., G.M., M.V., J.R.M., K.K.]; Biotechnology and Biological Sciences Research Council [BB/L027151/1 to G.M.]. Funding for open access charge: National Science Foundation [DBI-1262414].

Conflict of interest statement. None declared.

## REFERENCES

- 1. International Arabidopsis Informatics Consortium. (2012) Taking the next step: building an Arabidopsis information portal. *Plant Cell*, **24**, 2248–2256.
- Smith,R.N., Aleksic,J., Butano,D., Carr,A., Contrino,S., Hu,F., Lyne,M., Lyne,R., Kalderimis,A., Rutherford,K. *et al.* (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, 28, 3163–3165.
- Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., Hu, F., Smith, R., Stepan, R., Sullivan, J. et al. (2014) InterMine: extensive web services for modern biology. *Nucleic Acids Res.*, 42, W468–W472.

- 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.
- 5. UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198
- Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, 39, D52–D57.
- 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.
- 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.
- Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, 6, R44.
- Gene Ontology Consortium., Blake, J.A., Dolan, M., Drabkin, H., Hill, D.P., Li, N., Sitnikov, D., Bridges, S., Burgess, S., Buza, T. *et al.* (2013) Gene Ontology Annotations and Resources. *Nucleic Acids Res.*, **41**, D530–D535.
- 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.
- Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., McLaren, P., North, P. *et al.* (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.*, 8, R129.
- Balakrishnan, R., Park, J., Karra, K., Hitz, B.C., Binkley, G., Hong, E.L., Sullivan, J., Micklem, G. and Michael Cherry, J. (2012) Yeast Mine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database*, 2012, bar062.
- 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.
- Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, 19, 1630–1638.
- Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D. *et al.* (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.*, 148, 1772–1781.
- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410.
- Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
- Gomez, J., Garcia, L.J., Salazar, G.A., Villaveces, J., Gore, S., Garcia, A., Martin, M.J., Launay, G., Alcantara, R., Del-Toro, N. *et al.* (2013) BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics*, 29, 1103–1104.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13, 2498–2504.
- Fielding, R.T. (2000). *PhD Thesis*. University of California, Irvine. AAI9980887, 1–162.
- Zhou, P., Emmert, D. and Zhang, P. (2006) Using Chado to store genome annotation data. *Curr. Protoc. Bioinformatics*, Chapter 9, Unit 9.6.
- Sanderson, L.A., Ficklin, S.P., Cheng, C.H., Jung, S., Feltus, F.A., Bett, K.E. and Main, D. (2013) Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database*, 2013, bat075.

- Dooley, R. and Hanlon, M.R. (2014) Recipes 2.0: building for today and tomorrow. *Concurrency Computat.: Pract. Exper.*, doi:10.1002/cpe.3285.
- Goff,S.A., Vaughn,M., McKay,S., Lyons,E., Stapleton,A.E., Gessler,D., Matasci,N., Wang,L., Hanlon,M., Lenards,A. *et al.* (2011) The iPlant Collaborative: cyberinfrastructure for plant biology. *Front. Plant Sci.*, 2, 1–16.
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M. et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the

Trinity platform for reference generation and analysis. *Nat. Protoc.*, **8**, 1494–1512.

- Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K. Jr, Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D. *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.
- Lee, E., Helt, G.A., Reese, J.T., Munoz-Torres, M.C., Childers, C.P., Buels, R.M., Stein, L., Holmes, I.H., Elsik, C.G. and Lewis, S.E. (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, 14, R93.