D1202–D1210 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 2 December 2011*
doi:10.1093/nar/gkr1090

The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools

Philippe Lamesch, Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, Kate Dreher, Debbie L. Alexander, Margarita Garcia-Hernandez, Athikkattuvalasu S. Karthikeyan, Cynthia H. Lee, William D. Nelson, Larry Ploetz, Shanker Singh, April Wensel and Eva Huala*

Department of Plant Biology, Carnegie Institution, 260 Panama St., Stanford, CA 94305, USA

Received September 20, 2011; Revised October 30, 2011; Accepted November 1, 2011

ABSTRACT

The Arabidopsis Information Resource (TAIR, http://arabidopsis.org) is a genome database for *Arabidopsis thaliana*, an important reference organism for many fundamental aspects of biology as well as basic and applied plant biology research. TAIR serves as a central access point for Arabidopsis data, annotates gene function and expression patterns using controlled vocabulary terms, and maintains and updates the *A. thaliana* genome assembly and annotation. TAIR also provides researchers with an extensive set of visualization and analysis tools. Recent developments include several new genome releases (TAIR8, TAIR9 and TAIR10) in which the *A. thaliana* assembly was updated, pseudogenes and transposon genes were re-annotated, and new data from proteomics and next generation transcriptome sequencing were incorporated into gene models and splice variants. Other highlights include progress on functional annotation of the genome and the release of several new tools including Textpresso for Arabidopsis which provides the capability to carry out full text searches on a large body of research literature.

INTRODUCTION

TAIR (The Arabidopsis Information Resource, http://arabidopsis.org) serves as the community database for Arabidopsis researchers and as an essential information source for the wider plant biology and model organism communities (1,2). TAIR contains genetic and genomic data for *Arabidopsis thaliana*, a well-studied plant that serves as a reference species for many aspects of plant biology (3–7). *Arabidopsis thaliana* has also served as a highly productive research organism for exploring many areas of fundamental biology including DNA repair, photobiology, protein degradation, the circadian clock, DNA methylation, RNA silencing and G-protein signaling, many of which have direct application to human health (8–11).

TAIR's usage continues to increase with 45 000 unique visitors per month in 2010 based on usage data gathered using Google Analytics and over 1.8 million visits in the past year, an increase of 6% over the previous year. Visits originated from around the world with Asia accounting for 36%, the Americas 31% and Europe 30%. Although registration is not required to view data at TAIR, users must register and log in to order seed and DNA stocks from the Arabidopsis Biological Resource Center (ABRC), enter comments on TAIR pages or submit data to TAIR via our online data submission tool. The number of registered TAIR users as of September 2011 has reached 22 000, with 9400 of these records added or

*To whom correspondence should be addressed. Tel: +1 650 739 4310; Fax: +1 650 325 6857; Email: ehuala@carnegiescience.edu
Present addresses:
David Swarbreck, The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, Norfolk, UK.
Christopher Wilks, Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, CA, USA.
Rajkumar Sasidharan, Department of Molecular, Cellular and Developmental Biology, University of California at Los Angeles, Los Angeles, CA, USA.
Debbie L. Alexander, UCSF Office of Innovation, Technology, and Alliances, San Francisco, CA, USA.
Margarita Garcia-Hernandez, Center for Evaluation and Research, University of California at Davis, Davis, CA, USA.
Athikkattuvalasu S. Karthikeyan, Department of Plant Breeding and Genetics, 240 Emerson Hall, Cornell University, Ithaca, NY 14853, USA.

modified within the past 5 years, serving as an estimate of the most active set of users.

TAIR DATA TYPES AND SOURCES

Data available from TAIR include *A. thaliana* and *A. lyrata* genomic sequences, gene structure and function annotation, *A. thaliana* metabolic pathways, gene expression patterns, DNA and seed stock data, genome maps, genetic and physical markers, ecotypes and natural variation data, publications, and information about the Arabidopsis research community. These data come from a variety of sources including manual curation of published literature and sequence data, computational pipelines for annotating gene structure and function, integration of data from other biological databases and resources (GenBank and ABRC/Arabidopsis Biological Resource Center) and submissions from the research community.

Manual literature curation by TAIR curators generates gene function and gene expression annotations based on experiments reported in the peer-reviewed research literature, using Gene Ontology (GO) terms for molecular function, biological process and cellular component, and Plant Ontology (PO) terms for plant anatomical structures and growth and developmental stages (12,13). Additional information extracted from research literature includes gene symbols and full names, alleles (including allele name, mutagen, inheritance, allele type and description) and germplasm information (parent line, associated alleles and phenotype).

In addition to extracting data from the research literature, TAIR uses several computational pipelines to integrate additional data. Functional annotation pipelines are used to assign GO terms to *A. thaliana* and *A. lyrata* genes based on the presence of protein domains or signal sequences. Gene structure pipelines are used to update gene features such as exons and UTRs (untranslated regions) for *A. thaliana* and add new genes based on new transcript evidence. Mapping pipelines are used to assign a genome position to sequenced objects including ESTs (expressed sequence tags) and cDNAs, T-DNA and transposon insertions, markers, SNPs, etc. Data import pipelines are used to download sequence data from GenBank including new ESTs, cDNAs and insertion mutant flanking sequences, and load ABRC data for seed and DNA stocks.

Community data submissions to TAIR include gene families, gene structures, gene function data, mutant phenotypes, protein–protein interactions, gene expression patterns, SNPs, markers, laboratory protocols, gene symbols, metabolic pathway data and links to other resources. A recent development in community data submission to TAIR is the establishment of a novel TAIR-journal collaboration program to collect gene function information directly from authors at the time of publication and the introduction of an online author submission tool to facilitate data submission.

ARABIDOPSIS GENE FUNCTION ANNOTATION

Since joining the GO Consortium in 2001, gene function curators at TAIR have worked to capture the available experimental gene function data from the *A. thaliana* research literature in the form of GO and PO controlled vocabulary annotations. In recent years the main focus of our in-house literature curation effort has been on the annotation of newly characterized genes. An average of 260 research articles are added to TAIR each month based on PubMed searches for 'Arabidopsis' in the title, abstract or keywords and ~150 of these are linked to gene names or synonyms within TAIR each month using automated methods. These computationally generated links between articles and genes are manually reviewed and confirmed by curators if correct. During this process, abstracts that discuss a newly characterized gene are flagged as high-priority articles for curation. TAIR curators read the full text of ~40 of these high-priority articles each month and make GO and PO annotations based on the experiments reported in the article as well as extracting other types of data (gene names, allele information) as described earlier.

We have also put considerable effort into encouraging community submissions, most recently through the establishment of collaborations with 10 plant science journals and the development of a new interface for community submission of annotations (Berardini *et al.*, manuscript in preparation). Recently we have also begun to integrate external GO annotations from the UniProt Gene Ontology Annotation group at the European Bioinformatics Institute and the Reference Genome group of the GO Consortium (14). We have also integrated annotations inferred by the GO Consortium from links within the ontology. For example, gene products annotated to the molecular function term 'sodium ion transmembrane transporter activity' (GO:0015081) are also annotated to the biological process term 'sodium ion transmembrane transport' (GO:0035725) because the molecular function term is linked to the biological process term.

Each GO annotation from the sources described earlier consists of a gene identifier, a GO term, an evidence code and a reference. Although in some cases two or more annotations may contain the same gene and GO term, as long as the evidence code or reference differ these are still considered unique annotations and are retained. In other cases two separate annotations to the same gene may provide two related GO terms differing in specificity, for example 'chloroplast' versus 'chloroplast inner membrane', often because the method differed between the two experiments. In a few cases different methods or even the same method in different hands may produce a different result (e.g. location of a gene product in 'chloroplast' versus 'cytoplasm', resulting in two independent annotations, both of which are retained in order to provide a complete picture of all experimental results. We encourage all users of GO annotations to make full use of the evidence codes, evidence descriptions and links to the research article describing the experimental result that are available as part of each annotation in such cases.

To supplement the gene function information we extract manually from research literature and incorporate from the community and other resources, we use computational methods to assign GO terms based on the presence of protein domains and other conserved sequences of known function. For each genome release, we use a combination of InterProScan (15) on the proteome combined with the latest InterPro2GO mapping file (http://www.ebi.ac.uk/GOA/InterPro2GO.html) to create GO annotations for proteins based on the presence of domains with mapped GO terms. We also perform a TargetP analysis (16) with plant-specific parameters to identify proteins that are predicted to be secreted or to localize to the chloroplast or mitochondrion and created appropriate GO annotations based on these results. Annotations resulting from these computational methods are loaded into TAIR using the IEA (Inferred from Electronic Annotation) evidence code only if they provide an annotation to a GO aspect (molecular function, cellular component or biological process) not yet obtained from other annotation methods for that gene (e.g. for a gene product with an experimental annotation to 'chloroplast', an IEA annotation to 'cytoplasm' would not be loaded).

GO annotation for the *A. thaliana* genome

To date, 20% of all *A. thaliana* genes (excluding pseudogenes and genes encoded by transposable elements) have been annotated to at least one GO term for a biological process based on an experiment done directly on the gene in question or its protein or RNA product, as shown in Table 1. A slightly higher proportion (27%) have been assigned a GO cellular component term based on direct experiment, and only 13% have an experimentally based annotation to a molecular function term. When other types of evidence are included, such as sequence similarity to a gene of known function or presence of a domain with a well-defined function, for each GO aspect over half of *A. thaliana* genes have at least one annotation (Table 1, 'All evidence types' column). A total of 77% of all *A. thaliana* genes have at least one GO annotation to one of the three GO aspects.

'Unknown' genes annotated to GO root terms

The goal of the GO consortium is to assign at least one biological process, molecular function and cellular component term to every gene in an organism. In cases where there is no experimental data and no predicted function based on domains or other computational methods, curators assign the root term (e.g. 'biological process' rather than a more specific biological process term such as 'transcription' or 'leaf development') to indicate that this aspect of the gene function is unknown. The presence of an annotation to the root term serves as a way to distinguish 'unknown' genes for which a curator has examined the literature and computational outputs and found no possible GO annotation from 'uncurated' genes that lack an annotation because existing publications for that gene have not yet been examined. 'Unknown' genes account for 30–34% of the genome within different GO aspects (Table 1). However, because our computational methods for locating all publications relevant to a gene are imperfect (in particular we don't currently search for gene names in supplemental results files), it's likely that some genes currently classified as 'unknown' should be included in the 'uncurated' category.

To improve curation efficiency and reduce the fraction of unannotated genes we have begun using a semi-automated curation process to identify papers with cellular component information and create annotations from them. Such a process has been used successfully by WormBase (17) to streamline and improve their curator's efficiency in dealing with this type of data. We have worked closely with the WormBase Textpresso team to adapt and improve the software that is used in this process for use in *A. thaliana*. A combination of user submissions, semi-automated curation, integration of annotations from collaborating groups and strategic paper selection from the most recent literature will continue to drive the updates of functional information for *A. thaliana*.

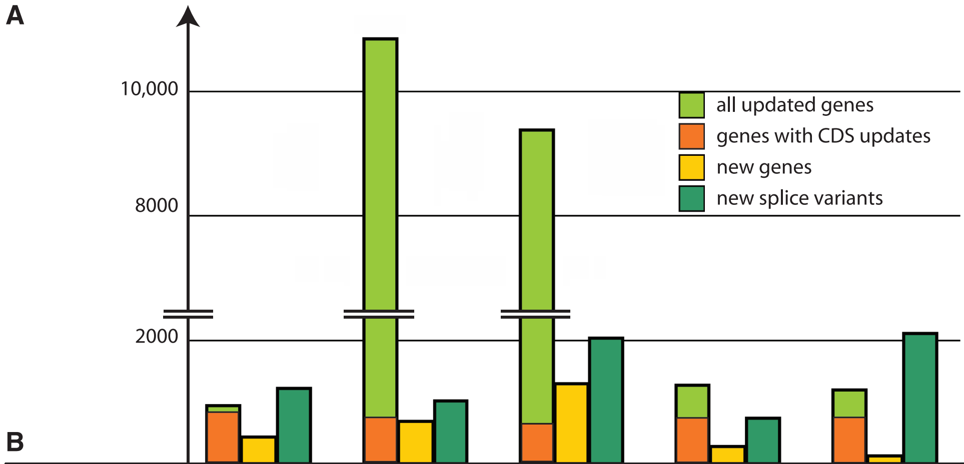
TAIR GENOME ANNOTATION

In an ongoing effort to improve the annotation of the *A. thaliana* genome, the TAIR genome annotation team has released improved versions of the *A. thaliana* gene set on a yearly basis since TAIR took over this responsibility from TIGR (The Institute for Genomic Research, now called J. Craig Venter Institute) in 2005 (18).

Table 1. *Arabidopsis thaliana* Gene Ontology Annotations

GO aspect	Experimental (%)	All evidence (%)	Unknown (%)	Not annotated (%)
Biological process (BP)	5826 (20)	15 644 (54)	9764 (34)	3367 (12)
Molecular function (MF)	3816 (13)	16 504 (57)	8732 (30)	3539 (12)
Cellular component (CC)	7762 (27)	15 383 (54)	7529 (26)	5863 (20)
BP, MF or CC	10 595 (37)	22 047 (77)	n/a	939 (3)[a]

Number of *A. thaliana* genes with annotations to the three GO aspects and their percentages relative to the total number of genes excluding pseudogenes and transposable element genes, based on the TAIR10 genome release. 'Experimental' category includes genes annotated with evidence codes IDA (inferred from direct assay), IMP (inferred from mutant phenotype), IGI (inferred from genetic interaction), IPI (inferred from physical interaction) and IEP (inferred from expression profile). 'All evidence' includes all evidence codes except ND (no biological data available). 'Unknown' includes genes annotated to the GO root term within the indicated category using the ND evidence code. 'Not annotated' includes genes with no annotation to date within the indicated GO category. Numbers as of 15 September 2011; n/a not applicable.
[a]Genes with no GO annotation of any kind.



B

TAIR release	TAIR6	TAIR7	TAIR8	TAIR9	TAIR10
main experimental datasets used for annotation	cDNAs & ESTs	cDNAs & ESTs	cDNAs & ESTs	cDNAs & ESTs MS peptides	cDNAs & ESTs MS peptides RNAseq
main gene prediction tools used	PASA	PASA	PASA Gnomon	PASA Gnomon Augustus EuGene	Augustus TAU Cufflinks
areas of focus/ comments	protein-coding genes	• UTRs • cysteine-rich peptides • other_RNA genes	• initial annotation of transposons • small ORFs • uORFs	• pseudogenes	• alternative splice-variants • B-list genes
updates made to genome sequence	none	none	• 1425 single nucl. substitutions • 14 contaminated regions substituted with Ns	• 227 single nucl. substitutions • 341 indels	none

Figure 1. Overview of TAIR genome releases. (**A**) Bar graph displaying the number of annotation updates made in each of the 5 TAIR releases. Colored bars represent four different classes of updates: updated genes (light green), genes with CDS updates (orange), new genes (yellow) and new splice variants (dark green). (**B**) Table comparing the TAIR genome releases by types of data and prediction tools used, areas of focus and genome sequence updates. The red line separating TAIR8 from TAIR9 indicates that coordinates of most genes shifted in the TAIR9 release, as a consequence of the integration of 341 Indels, and the normalization of previously identified sequence contaminations to a standard length of 100 bp. A liftover tool is available at ftp://ftp.arabidopsis.org/home/tair/Software/UpdateCoord/ for updating coordinates of objects mapped to TAIR8 or earlier releases.

To maximize both efficiency and accuracy of the genome annotation process TAIR has made use of a combination of computational methods to identify genes requiring updates and carry out simpler updates, and manual curation using the Apollo gene editing tool (19) to review and carry out more complex updates.

Figure 1 illustrates the consecutive steps that we have taken to gradually improve the annotation of the *A. thaliana* genome. After an initial effort (TAIR6) focused on the annotation of protein-coding genes, more extensive reannotation projects were undertaken in subsequent releases to improve the annotation of UTRs, short protein-coding genes, non-coding RNAs (ncRNAs), transposon genes, pseudogenes and splice variants. As shown in Figure 1B, while the genome annotation tool PASA (Program to Assemble Spliced Alignments) (20) was the only gene prediction tool used in the first releases, curators increasingly relied on additional gene prediction tools to make use of newly available transcript profiling (RNA-seq) and peptide datasets generated using high-throughput methods.

TAIR8 genome release (April 2008)

As with the previously described TAIR6 and TAIR7 releases (21), PASA was used to incorporate all available *A. thaliana* ESTs and cDNAs into transcript assemblies and generate lists of suggested updates to existing gene models for the TAIR8 release. These updates were categorized by PASA into different groups depending on the type of change required (i.e. extension of the 3′-UTR, altered protein coding sequence, etc.) All but the most straightforward update categories, such as extension of the 5′- or 3′-UTRs, were individually reviewed by curators using a modified PASA interface and marked for manual curation, computational update, or rejection.

In addition to making genome-wide computational updates based on new transcript evidence, for each TAIR release we have targeted a specific subclass of genes for review and update. For the TAIR8 release, we integrated a large new set of transposable elements provided by Quesneville and co-authors (22) and used this new information to update the gene type for many genes contained within these newly mapped transposable elements from protein-coding or pseudogene to transposable element gene (see Supplementary Data for more information). Other novel genes introduced in TAIR8 include conserved uORFs (upstream open reading frames located within the UTR of other, larger genes) (23), and very short protein-coding genes with substantial supporting evidence (24). Datasets from several other groups were also used to annotate new genes and splice variants and make gene structure updates (25–27, T. Tatusova, personal communication).

The TAIR8 release contained 27 235 protein coding genes, 859 pseudogenes, 3900 transposable element genes and 1288 ncRNAs (33 282 genes in all, 38 963 gene models). A total of 1291 new genes and 2009 new gene models were added. Thirteen percent (4330) of *A. thaliana* genes had annotated splice variants in this release. Updates were made to 3811 gene structures of which 625 gene models had coding sequence (CDS) updates; a total of 4007 exons were modified and 683 new exons incorporated. There were 33 gene splits and 41 gene merges. Overall 23% of all existing TAIR7 genes (7380 genes) were updated, including updates to gene structure and/or gene type.

The TAIR8 release also included changes to the genome sequence. In 14 regions identified as contaminating sequences from vectors, *E. coli* or rice, the contaminating sequence was replaced with a run of 'N' of the same length to avoid changes to chromosome length. In addition, 1425 single nucleotide substitutions were made to the assembly sequence based on high-confidence resequencing data provided to TAIR (28). The sequences of 518 genes overlapping these substitutions were also updated. Because all assembly changes for the TAIR8 release were substitutions rather than insertions or deletions, the chromosome lengths and gene coordinates were unchanged from the previous releases.

TAIR9 genome release (June 2009)

With the TAIR9 release, the set of data used as evidence for updates to gene structures was expanded to include cross-species alignments and peptide data from two large-scale proteomics experiments (29,30). These proteome datasets were used to reclassify 99 pseudogenes as protein coding and merge nine pseudogenes with existing protein coding genes. In addition, 158 peptides were used to update TAIR gene structures. A set of predicted Augustus gene models based on proteome data (30) were evaluated to identify potential exons missing from TAIR8. Of 591 Augustus models examined, 339 were incorporated into TAIR9 gene models, with 175 new splice variants added, 118 modifications to existing TAIR models and 46 new gene models.

The TAIR9 release also included a genome-wide reannotation of pseudogenes based on output from the PseudoPipe software package, provided by the Gerstein lab (31). Further analysis was undertaken to identify a subset of pseudogene models exhibiting CDS disablements or truncations relative to the parent gene. A total of 168 novel pseudogenes were added for the TAIR9 release.

Alternative genome annotation datasets derived from several different software packages, including Gnomon (http://www.ncbi.nlm.nih.gov/projects/genome/guide/gnomon.shtml), (predictions provided by Tatiana Tatusova and Alexandre Souvorov, NCBI), EuGene (27,32) (gene predictions provided by Sébastien Aubourg, Unité de Recherche en Génomique Végétale) and AceView (33), predictions provided by Jean Thierry-Mieg, NCBI) were also used as a source of gene structure updates for the TAIR9 release. An analysis of these gene prediction sets was undertaken to identify a set of exons absent from TAIR8 annotation but supported by transcript, peptide or cross-species evidence, resulting in the addition or modification of over a thousand exons for TAIR9. The full set of alternative gene models submitted to TAIR for all three software packages can be viewed in TAIR's genome browser as tracks within the Community Alternative Annotation section (http://gbrowse.arabidopsis.org/cgi-bin/gbrowse/arabidopsis).

The TAIR9 release contained 27 379 protein coding genes, 926 pseudogenes, 3901 transposable element genes and 1312 ncRNAs (33 518 genes in all, 39 640 gene models). A total of 282 new genes and 739 new splice variants were added. Fourteen percent (4626) of *A. thaliana* genes had annotated splice variants in this release. Updates were made to 1254 gene models of which 774 had CDS updates; a total of 1144 exons were modified and 1056 new exons incorporated. There were 13 gene splits and 46 gene merges.

Genome assembly updates made for the TAIR9 release included 227 single nucleotide substitutions based on re-sequencing data provided to TAIR (28,34). A set of 341 insertions or deletions were made based on re-sequencing data (34) and EST or cDNA sequences deposited in GenBank that supported the change. In accordance with our reference genome policy (http://arabidopsis.org/doc/portals/genAnnotation/gene_structural_annotation/ ref_genome_sequence/11413) corrections to the reference assembly were only made if supported by at least two independently derived sequence libraries from the Columbia ecotype. In addition to these changes, the 14 regions previously identified in TAIR8 as either vector, *E. coli* or rice contamination and substituted with the equivalent number of IUPAC ambiguity code 'N's were standardized (via deletion) to a set size of 100 bp for TAIR9. As a consequence of these assembly updates, the coordinates of most genes, as well as other mapped features such as transcripts, polymorphisms, T-DNAs, etc. were modified between the TAIR8 and TAIR9 releases.

TAIR10 genome release (December 2010)

For the TAIR10 release, RNA-seq data were incorporated as evidence for gene model updates. Data used for this release included a total of 538 million reads obtained from two groups (28,35). RNA-seq reads were mapped to the genome using TopHat (36), HashMatch (http://mocklerlab-tools.cgrb.oregonstate.edu/HashMatch.html) and Supersplat (37). After quality and low complexity filtering, we mapped >200 million reads to the genome, including about nine million spliced reads. Spliced aligned reads can be viewed within TAIR's genome browser, in the 'Spliced RNA-Seq Reads' track within the Sequence Similarity section (http://gbrowse.arabidopsis.org/cgi-bin/gbrowse/arabidopsis). These spliced read alignments plus peptide data obtained for the TAIR9 release were used as an input for the Augustus gene prediction package (38) and the resulting gene models were categorized and manually reviewed (see Supplementary Figure S1). Validated gene updates, novel genes and novel splice variants from the Augustus output were incorporated into the TAIR10 release. Spliced reads not incorporated into gene models by Augustus were supplied to TAU (http://mocklerlab-tools.cgrb.oregonstate.edu/TAU.html), and resulting models were reviewed by TAIR curators for the addition of novel splice variants. Transcript assemblies were also generated independently via Cufflinks (39), using both spliced RNA-seq reads and a subset of unspliced reads generated by the Ecker lab. Transcript assemblies were filtered and compared to existing gene models, resulting in the addition of 56 novel genes.

In addition to the updates resulting from incorporation of RNA-seq data, new proteome data provided to TAIR (40) was used to directly update 24 gene models. Also, gene models created using the Gnomon pipeline were provided to TAIR by NCBI and reanalysis of these models resulted in 11 additional novel genes, 67 additional alternative splice variants and 164 updates to existing genes. Finally, a set of 125 updates provided by curators at Swiss-Prot (http://www.uniprot.org/) were reviewed and 104 of these updates were incorporated into this release.

The TAIR10 release is summarized in Table 2. For this release, a total of 126 new genes and 2099 new splice variants were added. Updates were made to 1184 gene models of which 707 had CDS updates. There were 41 gene splits and 37 gene merges. No updates were made to the genome assembly for the TAIR10 release. Eighteen percent of *A. thaliana* genes (total of 5885) now have annotated splice variants and 65.1% (22 982) of protein coding gene model structures are fully confirmed by EST or cDNA data (every exon is supported by an *A. thaliana* EST or cDNA). A further 10 829 gene models are partially supported. Thus, a total of 33 811 (95.7%) protein-coding gene models have at least partial transcript support.

As part of the TAIR10 release two new genome browser tracks, 'B-List Genes' and 'TAIR10 Unconfirmed Exons', have been added to the TAIR genome browser (http://gbrowse.arabidopsis.org/cgi-bin/gbrowse/arabidopsis/). 'B-List Genes' displays a set of 1737 gene models not included in the TAIR10 release because curators were unable to find an appropriate open reading frame that extended through the gene model. Many of these gene models are associated with previously annotated protein-coding genes and may be non-coding splice variants of these genes. Gene models were only included on the B-List if they had sufficient experimental data (typically >50% of exons with overlapping evidence and at least two different types of evidence such as RNA-seq, ESTs, cDNAs or peptides) to suggest that they are expressed. A second new track, 'Unconfirmed Exons', displays TAIR10 gene exons that lack confirming experimental evidence for one or both splice sites that flank the exon. The text displayed below each exon within the track indicates whether the donor or acceptor site of the exon is unsupported. Documentation on how we generated the confidence score for each exon can be found on the TAIR ftp site at: ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR_gene_confidence_ranking/DOCUMENTATION_TAIR_Gene_Confidence.pdf.

NEW TOOLS AT TAIR

TAIR provides access to a variety of in-house and external tools that help the user query and analyze Arabidopsis data. All tools are available from every TAIR page under the Tools dropdown menu. Recently added tools include the Textpresso literature search tool, the N-Browse interaction viewer, the synteny viewer GBrowse_syn, the Integrated Genome Browser (IGB) and GBrowse viewers for nine new plant genomes.

Textpresso is an information extracting and processing package for biological literature (41). Textpresso for Arabidopsis allows users to search over 40 000 abstracts and 27 000 full-text publications in TAIR. Keyword searches can be narrowed by searching in specific keyword categories including *A. thaliana* gene names, Gene Ontology and Plant Ontology terms or combinations of keywords. This tool is extremely useful in tracking down specific information like the mutation sites in certain alleles. Sentences that contain matching keywords are displayed together with bibliographic

Table 2. TAIR10 genome statistics

	Protein coding	pre-tRNA	rRNA	snRNA	snoRNA	miRNA	Other RNA	Pseudo gene	TE gene	Total
Nuclear	27 206	631	4	13	71	177	394	924	3903	33 323
Chloroplast	88	37	8	0	0	0	0	0	0	133
Mitochondrial	122	21	3	0	0	0	0	0	0	146
Total	27 416	689	15	13	71	177	394	924	3903	33 602

Number of genes of each category in the TAIR10 genome release.

information so that users can quickly confirm the usefulness of a particular paper and link directly to the full text, if they have the appropriate subscriptions to the journals in question.

N-Browse is an interactive graphical browser for biological networks (42). Users can launch N-Browse using Java Web Start with or without an initial query gene. N-Browse contains 8626 protein–protein interactions based on experimental data curated by TAIR or the protein interaction databases BioGRID (http://thebiogrid.org) (43) or IntAct (http://www.ebi.ac.uk/intact) (44). N-Browse does not currently contain any predicted protein interaction data. Interaction data is available for download at ftp://ftp.arabidopsis.org/home/tair/Proteins/Protein_interaction_data/.

GBrowse_syn is a GBrowse-based synteny browser designed to display multiple genomes, with a central reference species compared to several additional species (45). This tool uses a central 'joining' database that contains information about the multiple sequence alignments as well as additional databases for each species represented in the alignments. GBrowse_syn was built to help researchers study and analyze syntenic regions, homologous genes and other conserved elements between sequences. It can also be used to study genome duplication and evolution. By comparing newly sequenced or less studied genomes to the well annotated *A. thaliana* genome in Gbrowse_syn, scientists can identify novel genes and putative regulatory elements. The first version of the Gbrowse_syn tool at TAIR includes the genomes of *A. thaliana*, *A. lyrata* and *Populus trichocarpa*.

Integrated Genome Browser (IGB) is an interactive genome browser tool (46). IGB is different from other genome browsers in that it lets the user open, visualize and analyze their own large-scale data sets (i.e. RNA-Seq, ChIP-Seq, epigenetics, tiling array, etc), displaying these data alongside publicly available data sets, including gene models and the reference sequence itself. Using IGB's QuickLoad system, users can also use IGB to share data with collaborators and members of the community. IGB runs on the user's local computer rather than on the TAIR servers.

New plant genomes in GBrowse In addition to GBrowse for *A. thaliana*, TAIR has made GBrowse viewers available for the following plant genomes: *A. lyrata, Brachypodium distachyon, Oryza sativa* ssp. *japonica, O. sativa* ssp. *indica, P. trichocarpa, Physcomitrella patens, Sorghum bicolor, Vitis vinifera, Zea mays*. Gene models for each species were obtained from Ensembl, while transcript data were retrieved from GenBank and aligned to each genome using the GMAP alignment tool (47). *Arabidopsis thaliana* gene models were aligned to each plant genome using CAT (48), and the alignments are displayed in a GBrowse ortholog track.

ARACYC

AraCyc contains information about *A. thaliana* metabolic pathways, reactions and enzymes. This resource was originally developed at TAIR but is now maintained and updated by the Plant Metabolic Network (49) (http://www.plantcyc.org). TAIR users can still directly query the AraCyc database by using the quick search tool at TAIR and selecting the 'Metabolic Pathways' option from the drop-down menu. Searching using a simple term such as 'tryptophan' returns an array of pathways, enzymes, reactions and compounds, but more specific searches, for example using an AGI locus code, can bring users directly to the information they want related to a specific enzyme. In addition, links to relevant AraCyc pathways and reactions can be accessed in the 'External Link' section of the locus page for enzymes in TAIR.

The AraCyc 8.0 release from April 2011 contains 446 pathways, 5520 enzymes, 2689 reactions, 2825 compounds and includes information from 3346 references (http://www.plantcyc.org/ release_notes/content_statistics.faces). New releases are typically produced twice a year. Annotations of enzymes to specific reactions and pathways are made using evidence codes to enable users to easily distinguish which enzyme-reaction pairings are supported by experimental or computational evidence. Specific pages give more detailed information about each enzyme, reaction and compound, plus these different data types are brought together on information-rich pathway pages. AraCyc also provides access to tools that allow users to generate complex queries, to compare metabolism across different species, and to overlay experimental data from large-scale transcriptomic, proteomic and/or metabolomic studies onto a zoomable metabolic map (http://pmn.plantcyc.org/overviewsWeb/celOv.shtml?orgid = ARA).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Methods and Supplementary Figure S1.

ACKNOWLEDGEMENTS

We wish to thank Richard Clark and Ronan O'Malley for providing genomic resequencing data, Katja Bärenfaller and Natalie Castellana for providing peptide data, Michael Tognolli, Tatiana Tatusova, Anjana Raina and Volker Brendel for providing community-submitted gene structure updates through yrGATE, Noah Fahlgren for providing miRNAs, Alexander Poliakov for providing conserved non-coding regions across several plant genomes, and many members of the Arabidopsis research community for gene structure and function updates, Mark Gerstein for providing PseudoPipe results, Hans-Michael Müller for providing Textpresso for Arabidopsis, Kris Gunsalus for providing the N-Browse interaction viewer, Sheldon McKay for providing GBrowse_syn, Pedro Pattyn for providing *A. lyrata* and *P. trichocarpa* genomic alignments to *A. thaliana* displayed within GBrowse_syn, Ann Loraine for providing the IGB genome browser, and Sue Rhee and Peifen Zhang for providing AraCyc.

FUNDING

National Science Foundation (grant DBI-0850219); National Institutes of Health National Human Genome Research Institute (NHGRI) (grant 5P41HG002273-09 for gene function curation, partial). Additional support for gene function curation comes from the TAIR sponsorship program (see http://arabidopsis.org/doc/about/tair_sponsors/413 for a complete list of sponsors). Part of this work was carried out by using the resources of the Computational Biology Service Unit from Cornell University which is partially funded by Microsoft Corporation. Funding for open access charge: National Science Foundation (grant DBI-0850219).

Conflict of interest statement. None declared.

REFERENCES

1. National Research Council. (2008) *Funding a Revolution: Achievements of the National Plant Genome Initiative and New Horizons in Plant Biology*. National Academy Press, Washington, DC.
2. Xu,X.M. and Møller,S.G. (2011) The value of *Arabidopsis* research in understanding human disease states. *Curr. Opin. Biotechnol.*, **22**, 300–307.
3. Koornneef,M. and Meinke,D. (2010) The development of *Arabidopsis* as a model plant. *Plant J.*, **61**, 909–921.
4. Buell,C.R. and Last,R.L. (2010) Twenty-first century plant biology: impacts of the *Arabidopsis* genome on plant biology and agriculture. *Plant Physiol.*, **154**, 497–500.
5. Avni,A. and Blázquez,M.A. (2011) Can plant biotechnology help in solving our food and energy shortage in the future? *Curr. Opin. Biotechnol.*, **22**, 220–223.
6. Chew,Y.H. and Halliday,K.J. (2011) A stress-free walk from *Arabidopsis* to crops. *Curr. Opin. Biotechnol.*, **22**, 281–286.
7. Zhang,J., Elo,A. and Helariutta,Y. (2011) *Arabidopsis* as a model for wood formation. *Curr. Opin. Biotechnol.*, **22**, 293–299.
8. Hays,J.B. (2002) *Arabidopsis thaliana*, a versatile model system for study of eukaryotic genome-maintenance functions. *DNA Repair*, **1**, 579–600.
9. van Baarlen,P., van Belkum,A. and Thomma,B.P. (2007) Disease induction by human microbial pathogens in plant-model systems: potential, problems and prospects. *Drug Discov. Today*, **12**, 167–173.
10. Jones,A.M., Chory,J., Dangl,J.L., Estelle,M., Jacobsen,S.E., Meyerowitz,E.M., Nordborg,M. and Weigel,D. (2008) The impact of *Arabidopsis* on human health: diversifying our portfolio. *Cell*, **133**, 939–943.
11. Schlaich,N.L. (2011) *Arabidopsis thaliana* – the model plant to study host-pathogen interactions. *Curr. Drug Targets*, **12**, 955–966.
12. Gene Ontology Consortium. (2010) (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
13. 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.
14. Reference Genome Group of the Gene Ontology Consortium. (2009) (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput. Biol.*, **5**, e1000431.
15. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
16. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.
17. Van Auken,K., Jaffery,J., Chan,J., Müller,H.M. and Sternberg,P.W. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*, **10**, 228.
18. Haas,B.J., Wortman,J.R., Ronning,C.M., Hannick,L.I., Smith,R.K. Jr, Maiti,R., Chan,A.P., Yu,C., Farzad,M., Wu,D. *et al.* (2005) Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol.*, **3**, 7.
19. Lewis,S.E., Searle,S.M.J., Harris,N., Gibson,M., Iyer,V., Ricter,J., Wiel,C., Bayraktaroglu,L., Birney,E., Crosby,M.A. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, research0082.
20. 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.
21. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
22. Buisine,N., Quesneville,H. and Colot,V. (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*, **91**, 467–475.
23. Hayden,C.A. and Jorgensen,R.A. (2007) Identification of novel conserved peptide uORF homology groups in Arabidopsis and rice reveals ancient eukaryotic origin of select groups and preferential association with transcription factor-encoding genes. *BMC Biol.*, **5**, 32.
24. Hanada,K., Zhang,X., Borevitz,J.O., Li,W.H. and Shiu,S.H. (2007) A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Res.*, **17**, 632–640.
25. Alexandrov,N.N., Troukhan,M.E., Brover,V.V., Tatarinova,T., Flavell,R.B. and Feldmann,K.A. (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. *Plant Mol. Biol.*, **60**, 69–85.
26. Backman,T.W., Sullivan,C.M., Cumbie,J.S., Miller,Z.A., Chapman,E.J., Fahlgren,N., Givan,S.A., Carrington,J.C. and Kasschau,K.D. (2008) Update of ASRP: the Arabidopsis Small RNA Project database. *Nucleic Acids Res.*, **36**, D982–D985.
27. Aubourg,S., Martin-Magniette,M.L., Brunaud,V., Taconnat,L., Bitton,F., Balzergue,S., Jullien,P.E., Ingouff,M., Thareau,V., Schiex,T. *et al.* (2007) Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. *BMC Genomics*, **8**, 401.
28. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.
29. Baerenfaller,K., Grossmann,J., Grobei,M.A., Hull,R., Hirsch-Hoffmann,M., Yalovsky,S., Zimmermann,P., Grossniklaus,U., Gruissem,W. and Baginsky,S. (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science*, **320**, 938–941.
30. Castellana,N.E., Payne,S.H., Shen,Z., Stanke,M., Bafna,V. and Briggs,S.P. (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc. Natl Acad. Sci. USA*, **105**, 21034–21038.
31. Zhang,Z., Carriero,N., Zheng,D., Karro,J., Harrison,P.M. and Gerstein,M. (2006) PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*, **22**, 1437–1439.
32. Schiex,T., Moisan,A. and Rouzé,P. (2001) Eugène, an eukaryotic gene finder that combines several sources of evidence. *Lect. Notes Comp. Sci.*, **2066/2001**, 111–125.
33. Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7(Suppl. 1)**, S12.1–S12.14.
34. Ossowski,S., Schneeberger,K., Clark,R.M., Lanz,C., Warthmann,N. and Weigel,D. (2008) Sequencing of natural

strains of *Arabidopsis thaliana* with short reads. *Genome Res.*, **18**, 2024–2033.

35. Filichkin,S.A., Priest,H.D., Givan,S.A., Shen,R., Bryant,D.W., Fox,S.E., Wong,W.K. and Mockler,T.C. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.*, **20**, 45–58.
36. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
37. Bryant,D.W. Jr, Shen,R., Priest,H.D., Wong,W.K. and Mockler,T.C. (2010) Supersplat – spliced RNA-seq alignment. *Bioinformatics*, **26**, 1500–1505.
38. Stanke,M., Keller,O., Gunduz,I., Hayes,A., Waack,S. and Morgenstern,B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, **34**, W435–W439.
39. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
40. Baerenfaller,K., Hirsch-Hoffmann,M., Svozil,J., Hull,R., Russenberger,D., Bischof,S., Lu,Q., Gruissem,W. and Baginsky,S. (2011) pep2pro: a new tool for comprehensive proteome data analysis to reveal information about organ-specific proteomes in *Arabidopsis thaliana*. *Integr. Biol.*, **3**, 225–237.
41. Müller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.
42. Kao,H.L. and Gunsalus,K.C. (2008) Browsing multidimensional molecular networks with the generic network browser (N-Browse). *Curr. Protoc. Bioinformatics*, Chapter 9, Unit 9–11, 1–21.
43. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
44. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
45. McKay,S.J., Vergara,I.A. and Stajich,J.E. (2010) Using the Generic Synteny Browser (GBrowse_syn). *Curr. Protoc. Bioinformatics*, Chapter 9, Unit 9.12, 1–25.
46. Nicol,J.W., Helt,G.A., Blanchard,S.G. Jr, Raja,A. and Loraine,A.E. (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, **25**, 2730–2731.
47. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
48. Li,H., Guan,L., Liu,T., Guo,Y., Zheng,W.M., Wong,G.K. and Wang,J. (2007) A cross-species alignment tool (CAT). *BMC Bioinformatics*, **8**, 349.
49. Zhang,P., Dreher,K., Karthikeyan,A., Chi,A., Pujar,A., Caspi,R., Karp,P., Kirkup,V., Latendresse,M., Lee,C. *et al.* (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.*, **153**, 1479–1491.