Published online 5 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D1009–D1014***
doi:10.1093/nar/gkm965

The Arabidopsis Information Resource (TAIR): gene structure and function annotation

David Swarbreck*, Christopher Wilks, Philippe Lamesch, Tanya Z. Berardini, Margarita Garcia-Hernandez, Hartmut Foerster, Donghui Li, Tom Meyer, Robert Muller, Larry Ploetz, Amie Radenbaugh, Shanker Singh, Vanessa Swing, Christophe Tissier, Peifen Zhang and Eva Huala

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

Received September 14, 2007; Revised October 16, 2007; Accepted October 17, 2007

ABSTRACT

The Arabidopsis Information Resource (TAIR, http://arabidopsis.org) is the model organism database for the fully sequenced and intensively studied model plant *Arabidopsis thaliana*. Data in TAIR is derived in large part from manual curation of the Arabidopsis research literature and direct submissions from the research community. New developments at TAIR include the addition of the GBrowse genome viewer to the TAIR site, a redesigned home page, navigation structure and portal pages to make the site more intuitive and easier to use, the launch of several TAIR web services and a new genome annotation release (TAIR7) in April 2007. A combination of manual and computational methods were used to generate this release, which contains 27 029 protein-coding genes, 3889 pseudogenes or transposable elements and 1123 ncRNAs (32 041 genes in all, 37 019 gene models). A total of 681 new genes and 1002 new splice variants were added. Overall, 10 098 loci (one-third of all loci from the previous TAIR6 release) were updated for the TAIR7 release.

INTRODUCTION

Arabidopsis thaliana is widely used as an experimental model for plants in general and for the study of a variety of fundamental biological processes. The Arabidopsis Information Resource (TAIR) focuses mainly on Arabidopsis genomic and genetic data, including genes, clones, ecotypes, controlled vocabulary terms, markers, expression data, SNPs and other polymorphisms, mutant alleles and phenotypes, proteins, germplasms and sequences. Other data types in TAIR include publications, protocols, researchers and labs. A great strength of TAIR is the direct integration of Arabidopsis Biological Resource Center (ABRC) stocks into the database, allowing the community to discover seed and DNA stocks of interest and directly access them on TAIR pages. TAIR usage has increased steadily since the project was founded in 1999. Over a recent 1-month period (9 July–8 August, 2007) TAIR received ~110 000 visits from 31 600 unique visitors and 907 000 page views. TAIR usage has a worldwide distribution, with 28% of visits originating in the Americas, 24% in Asia and 23% in Europe.

TAIR DATA SOURCES

The major sources for data in TAIR include manual curation of the research literature, computational pipelines for annotating gene structure and function and mapping sequenced objects onto the genome, import of data from GenBank and ABRC and submissions from the research community. Manual literature curation at TAIR is at present limited to Arabidopsis research articles appearing in those journals with the highest journal impact factor due to the high time cost of manual curation. Approximately 36% of each month's average of 107 Arabidopsis research articles containing gene-related data are manually curated using this approach. The curation process includes annotation of genes with Gene Ontology (GO; function, process and cellular component) and Plant Ontology (structure and developmental stage) terms with appropriate evidence codes and references. In addition, gene symbols, alleles, phenotypes and germplasm information are captured from the literature and a free text gene description summarizing a gene's important features is composed by curators. Computational data is generated by a variety of automated pipelines. Gene structure pipelines update gene features such as exons and UTRs and add new genes based on new transcript evidence. Functional annotation pipelines assign GO terms to genes based on the presence of protein domains or signal sequences and generate a

*To whom correspondence should be addressed. Tel: +1 650 325 1521; Fax: +1 650 325 6857; Email: dswarbreck@stanford.edu

short phrase describing a gene's function. Mapping pipelines assign a genome position to sequenced objects including ESTs and cDNAs, T-DNA and transposon insertions, markers, SNPs, etc. Data import pipelines are used to download sequence data from GenBank including new ESTs and cDNAs and insertion mutant flanking sequences, and load data related to ABRC seed and DNA stocks. Community data submissions to TAIR include gene families, gene function data, new genes and updates to existing gene structures, mutant phenotypes, interaction partners, gene expression patterns, SNPs, markers, protocols, gene symbols, metabolic pathway data and links to other resources. For many of these data types, Excel-based data submission forms are available on the website. Gene symbols are handled by online submission. The data submission page (http://www.arabidopsis.org/submit/index.jsp) describes data types accepted by TAIR and provides guidelines for submission of data. Before loading into TAIR, the submissions are carefully checked by curators for completeness and correct format, and free text submissions describing gene function and experimental method are mapped to controlled vocabulary terms.

RECENT TAIR USER INTERFACE ENHANCEMENTS

To provide a simpler and more intuitive interface for accessing TAIR data, we recently redesigned the TAIR homepage using a new header containing the major functions of TAIR: Search, Browse, Tools, Stocks, Portals, Download, Submit and News. These labels suggest actions that a user of TAIR might wish to take, and provide a gateway to a variety of tools and information. In addition to the new header, lower level pages now have a left navigation bar to help orient users within the site and lead them quickly to the desired page with a minimum of clicking. Additionally, a series of portal pages were added to logically organize information relating to specific topics such as genome annotation and mapping resources. Each portal page contains a mixture of TAIR and external resources that are presented along with brief descriptions of each resource designed to guide users to the best resource for their needs. Within the genome annotation portal, TAIR's new genome snapshot page provides a summary of the state of annotation of the Arabidopsis genome from a functional and structural perspective. (http://www.arabidopsis.org/portals/genAnnotation/genome_snapshot.jsp). A site map (http://www.arabidopsis.org/sitemap.jsp) lists in hierarchical fashion the major functions and sub pages available at TAIR.

DATA ACCESS IMPROVEMENTS

As the variety and extent of data mapped to the Arabidopsis genome continues to increase there is a greater need to visualize genomic annotation and features from multiple sources in a single viewer. To facilitate this, we have recently added the popular genome viewer GBrowse (1). TAIR GBrowse (Figure 1) expands upon the capabilities of the pre-existing TAIR genome viewer (SeqViewer), offering the community a highly configurable display for viewing genomic annotations. GBrowse can be accessed from the Tools menu or TAIR detail pages (e.g. Locus and Polymorphism pages). A variety of annotation tracks are provided including gene models and their supporting transcript evidence as well as markers, clones, T-DNA and transposon insertions and polymorphisms. A track showing percent similarity between aligned Arabidopsis and *Populus trichocarpa* genomic sequences is provided through the VISTA plugin (2). Other datasets to be added to GBrowse in the near future include cross-species transcript alignments, additional genomic conservation tracks, endogenous transposable elements, expression data (e.g. SAGE) and methylation data. Additionally, users can upload their own annotation data into GBrowse from a local text file, allowing them to view private annotations in the context of existing public data.

Data files for all classes of data are accessible from the TAIR ftp site via the 'Download' link in the main header. Fully annotated chromosome sequences are provided in XML and GFF3 format, along with FASTA files of cDNA, CDS, UTR, genomic and protein sequences. Lists of newly added, deleted and updated genes for each TAIR genome release are also available on the ftp site, in addition to files mapping Arabidopsis Genome Initiative (AGI) identifiers to UniprotKB and NCBI RefSeq accessions. A variety of protein data files, microarray data files, structured vocabularies and functional gene annotation files are also provided. In contrast to the ftp site that provides genome-wide datasets, the Bulk Data Retrieval and Analysis tool provides a way to download data for a user-defined list of genes. Users can query for sequences, GO annotations, predicted protein properties, locus history and microarray elements by AGI identifier and can select either text or html output. The Gene search also provides a 'download all' option which can be used to retrieve data on a subset of genes.

TAIR has also recently added several web services using the BioMoby (3) framework to increase the options for automation and customization of bulk data retrieval. A significant benefit of using web services technology is the ability to simultaneously query several different data repositories and combine the resulting data into a single dataset. We have chosen to make our web services quite granular, to maximize the ability to construct any desired workflow by selecting the appropriate web service components. Currently TAIR has four web services in the production BioMoby registry: Locus2SpliceVariants, Locus2GOIDs, Locus2Publications and Locus2GeneAliases. We plan to add additional web services covering all major TAIR data types. These web services can be accessed through custom scripts or by using tools such as Taverna (http://taverna.sourceforge.net/) (4) or aggregator pages such as tAIGa (http://mips.gsf.de/proj/planet/araws/tAIGaSearch.html).

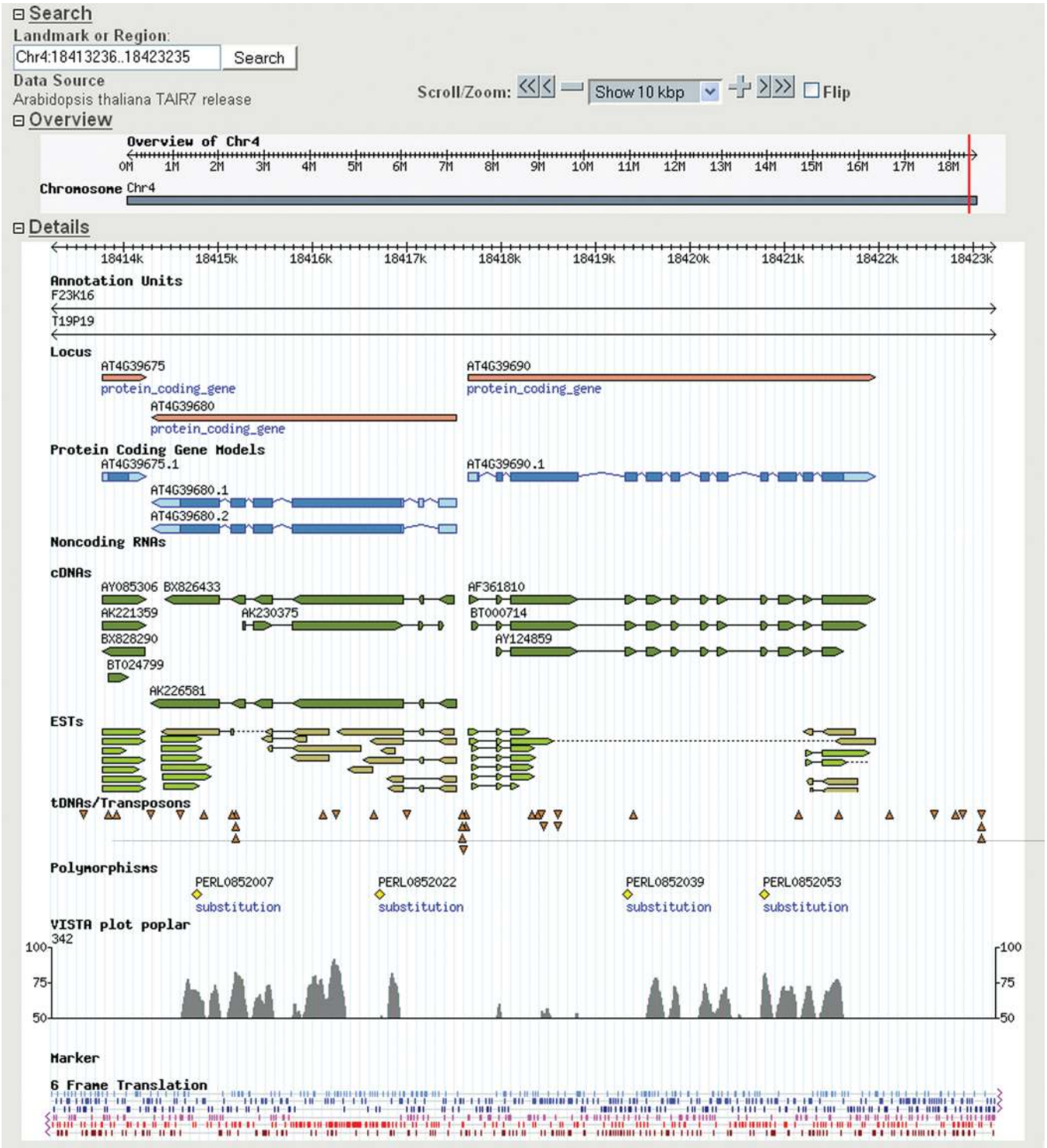

Figure 1. TAIR GBrowse. The TAIR GBrowse tool allows navigation of the five *A. thaliana* nuclear chromosomes plus the mitochondrial and chloroplast genomes. A 10 kb region including AT4G39680 and AT4G39690 is shown (coding regions in dark blue and UTRs in light blue). Selected tracks shown here include cDNAs (dark green), ESTs (light green for forward orientation and light brown for reverse), T-DNA and transposon insertions (orange triangles), polymorphisms (yellow diamonds) and the VISTA plot showing sequence similarity with poplar. Additional tracks (data not shown) include CDS segments, markers and GC content. All elements shown in GBrowse can be clicked on to access the TAIR detail page for that object.

Table 1. Evolution of the *A. thaliana* genome annotation from the initial *A. thaliana* sequencing project to the latest TAIR release

	Nature	TIGR1	TIGR2	TIGR3	TIGR4	TIGR5	TAIR6	TAIR7
Release date	12/14/00	1/17/01	9/11/01	8/2/02	4/18/03	1/29/04	11/11/05	4/24/07
Genome size (Mb)	115.410	116.238	117.227	117.077	119.055	118.998	119.186	119.186
Protein-coding genes	25 498	25 554	26 156	27 117	27 170	26 207	26 541	26 819
Transposons and pseudogenes	n/a	1274	1305	1967	2218	3786	3818	3889
Genes annotated with alternative splice-variants	n/a	0	28	162	1267	2330	3159	3866
Gene density (kb per gene)	4.5	4.55	4.48	4.32	4.38	4.54	4.48	4.44
Exons/gene model	5.2	5.23	5.25	5.24	5.31	5.42	5.64	5.79
Average exon length	250	256	265	266	279	276	269	268
Average intron length	168	168	167	166	166	164	164	165

TIGR values from Haas *et al.* (5). Numbers of protein-coding genes do not include those present on mitochondrial and chloroplast genomes.

GENOME ANNOTATION

Arabidopsis annotation past and present

Although the Arabidopsis genome sequence was completed in 2000, much work remains to be done to incorporate all available experimental data on gene structure and function into the genome annotation. The Arabidopsis genome is the only dicot genome finished and annotated to a high standard, and (with rice) one of only two finished, rather than draft quality, plant genome sequences. Because annotation of new genomes is in large measure based on the annotation of existing complete genomes, improvements to the Arabidopsis genome annotation will directly aid the annotation of future plant genomes.

TAIR assumed primary responsibility for updating the Arabidopsis annotation following the Institute for Genomic Research's (TIGR's) final genome release in 2004 (5). Since the original Arabidopsis genome annotation in which 25 498 genes were reported (6), the number of annotated genes has steadily increased as new sequencing and array-based technologies have provided evidence for many previously unannotated genes (Table 1). In particular, new sequence data deposited in the nucleotide sequence databases (EMBL/GenBank/DDBJ), along with user submissions from the Arabidopsis community, has resulted in the addition of over a thousand new genes since the final TIGR release in January 2004.

TAIR annotation pipeline

The TAIR gene structure annotation pipeline incorporates both manual and automated updates. Automated updates are carried out using the Program to Assemble Spliced Alignments (PASA) optimized for Arabidopsis (7). Updates are manually reviewed before incorporation by examining available transcript evidence. For the latest genome release (TAIR7), 581 (25%) of 2298 suggested updates (excluding simple UTR extensions) were rejected. Rejected updates included cases where genomic sequencing errors or misalignments of poor-quality ESTs or other aberrant transcripts resulted in incorrect gene structures. Additional rejected updates included gene splits based on partial cDNAs. The occurrence of such a large fraction of incorrect automated updates highlights the continued benefits of manual annotation to achieve a gold standard of annotation.

In addition to using the PASA pipeline to fold in new transcript evidence, TAIR has employed a targeted approach to identify and resolve structural errors in existing gene models. For example, some UTRs were falsely extended in previous releases based on ESTs of ambiguous orientation (i.e. non-splicing transcripts without poly(A) features, which originated from genes on the opposite strand). We manually reviewed candidate genes potentially containing falsely extended UTRs and corrected a total of 1098 gene models (909 genes) for the TAIR7 release. This has resulted in several hundred array elements no longer mapping to previously associated genes. In addition, the average 3' UTR length decreased from 250 to 233 bp in the TAIR7 release, contrary to the usual expectation that annotated UTRs are extended between releases due to the availability of more transcript data. In contrast, the average 5′ UTR size increased from 139 to 146 bp for this release.

TAIR7 genome release

The latest genome release, TAIR7, contains annotations for 27 029 protein-coding genes (including mitochondrial and chloroplast genomes), 3889 pseudogenes or transposable elements and 1123 ncRNAs (32 041 genes in all, with 37 019 gene models 31 921 of which are protein-coding models). Of the 27 029 protein-coding genes, 3799 (14%) are annotated with alternatively spliced gene models and 70% of all protein-coding models have annotated 5′ and 3′ UTRs. Seventy-one percent (22 596) of model structures are confirmed by transcript data (where every coding exon is supported by an *A. thaliana* EST or cDNA). A further 5819 gene models are partially supported. Thus, a total of 28 415 (89%) protein-coding gene models have at least partial transcript support, an increase of 5001 compared to November 2005 (the time of the TAIR6 release). In the main, this is due to the increased number of EST sequences deposited in GenBank, particularly those originating from high-throughput sequencing technologies (8). While the total number of genes continues to rise with each release (Table 1) (681 new genes were added for the latest release), the most common updates are refinements to existing gene structures. For the TAIR7 release 10 792 gene structures were updated, of which 797 gene models had updated coding exons. A total of 14 050 exons were modified and 828 new exons were incorporated. There were 41 gene splits and 34 gene merges. Overall, one-third

of all existing TAIR6 genes (10 098 genes) were updated for the TAIR7 release.

Expanding the Arabidopsis annotation by inclusion of small proteins and non-coding transcripts

Whereas the original Arabidopsis genome annotation focused solely on identifying protein-coding genes, subsequent re-annotations have added additional gene classes (i.e. pseudogenes, ncRNAs and transposable elements). However, certain gene classes may still be underrepresented. Transcriptome sequencing and whole-genome tiling array studies have revealed significant levels of expression in unannotated 'intergenic' regions (9,10). Not all such transcripts will be of functional significance but at least some are likely to represent unannotated protein-coding genes, ncRNAs or transposable elements. The problem of false-negative prediction is particularly serious for smaller CDSs where the decreasing signal-to-noise ratio makes distinguishing real genes from biologically meaningless open reading frames (ORFs) a more problematic issue (11). In order to reduce the number of short false-positive gene predictions TIGR applied a minimum cut-off of 110 amino acids (5). Unannotated protein-coding genes are therefore likely to fall below this threshold. Some previously unannotated small genes including 467 cysteine-rich peptides (12,13) have now been incorporated into the most recent genome releases (TAIR6 and TAIR7) while others are currently being evaluated for potential inclusion in the forthcoming release (14). In addition many unannotated transcripts excluded from earlier releases due to a lack of clear coding potential or assignment within one of the small RNA classes have now been incorporated. A total 213 genes were added as gene type 'other_RNA' where the gene cannot be assigned an unambiguous ORF and may therefore function as an ncRNA. One hundred and eighteen of these gene models overlap antisense to existing protein-coding genes and may represent natural antisense genes.

Alternatively spliced genes

While our latest annotation includes 1002 new splice variants, numerous sequences in GenBank that represent alternatively spliced transcripts with disrupted ORFs are not presently incorporated into TAIR gene structures. Many of these transcripts contain retained introns or alternative splice sites that generate premature termination codons (PTCs) and heavily truncated peptides. It remains an open question if such transcripts encode functional peptides. Some will undoubtedly represent mistakes by the splicing machinery or cloning and sequencing errors, while others may have regulatory roles via the nonsense-mediated decay (NMD) pathway, a surveillance mechanism that selectively degrades nonsense mRNAs. Coupling of regulated alternative splicing and NMD may therefore provide an alternative mechanism for regulating protein expression (15). We plan to incorporate and categorize these and other unusual transcripts encoding fused proteins or dicistronic transcripts containing two complete ORFs in future releases.

Functional annotation

At the level of gene function, we continue to refine our annotations using information from the literature and community-submitted annotations, as well as sequence similarity and other computational methods. Of the 28 152 genes in TAIR (excluding 3889 transposon genes and pseudogenes and 793 genetically defined, unsequenced genes), ~60% have been annotated to a GO molecular function, 50% to a GO biological process and 49% to a GO cellular component (excluding annotations to the 'unknown' terms). In addition to updating GO annotations, TAIR updates the gene description field either manually, using literature data, or computationally, using BLAST and InterProScan to identify similar proteins and protein domains. Approximately 4000 genes currently have similarity only to uncharacterized proteins (i.e. hypothetical, predicted, unknown, etc), while 758 have no significant protein similarity to any proteins deposited in GenBank. Of these, 286 in addition have no supporting EST/cDNA evidence and may represent erroneous gene predictions incorporated in earlier releases.

Future releases

Future TAIR genome releases will contain corrections to the chromosome sequences based on newly available sequence data, improved annotation of endogenous transposable elements and pseudogenes, more complete annotation of splice variants including those lacking coding potential and utilization of genome comparisons to further refine the existing gene structures, particularly those with little or no transcript support. We plan to produce new genome releases at least once per year, with the next release (TAIR8) expected in early 2008.

ACKNOWLEDGEMENTS

TAIR is supported by grant DBI-0417062 from the National Science Foundation. Funding to pay the Open Access publication charges for this article was provided by grant DBI-0417062 from the National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Stein,L., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J., Harris,T. *et al.* (2002) The generic genome browser: a building block for a model organism database. *Genome Res.*, **12**, 1599–1610.
2. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
3. Wilkinson,M.D. and Links,M. (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform.*, **3**, 331–341.
4. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, 729–732.
5. 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. *et al.* (2005)

Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release. *BMC Biol.*, **3**, 7.
6. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
7. 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. *et al*. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.
8. Weber,A.P., Weber,K.L., Carr,K., Wilkerson,C. and Ohlrogge,J.B. (2007) Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol.*, **144**, 32–42.
9. Yamada,K., Lim,J., Dale,J.M., Chen,H., Shinn,P., Palm,C.J., Southwick,A.M., Wu,H.C., Kim,C. *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*, **302**, 842–846.
10. Stolc,V., Samanta,M.P., Tongprasit,W., Sethi,H., Liang,S., Nelson,D.C., Hegeman,A., Nelson,C., Rancour,D. *et al*. (2005) Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc. Natl. Acad. Sci. USA*, **102**, 4453–4458.
11. Wang,J., Li,S., Zhang,Y., Zheng,H., Xu,Z., Ye,J., Yu,J. and Wong,G.K. (2003) Vertebrate gene predictions and the problem of large genes. *Nat. Rev. Genet.*, **4**, 741–749.
12. Silverstein,K.A., Graham,M.A., Paape,T.D. and VandenBosch,K.A. (2005) Genome organization of more than 300 defensin-like genes in *Arabidopsis*. *Plant Physiol.*, **138**, 600–610.
13. Silverstein,K.A., Moskal,W.A.Jr, Wu,H.C., Underwood,B.A., Graham,M.A., Town,C.D. and VandenBosch,K.A. (2007) Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. *Plant J.*, **51**, 262–280.
14. 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.*, **17–640.**
15. Lewis,B.P., Green,R.E. and Brenner,S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.