Genome Annotation at TAIR

The Arabidopsis genome was initially annotated by the Arabidopsis Genome Initiative (AGI) and later reannotated by TIGR in collaboration with MIPS and TAIR. TAIR assumed primary responsibility for maintaining the Arabidopsis genome annotation in North America following TIGR's final release (TIGR5), producing 5 additional genome releases, TAIR6 through TAIR10. TAIR10 was integrated into GenBank in November 2010.

The most recent release, Araport11, was produced by the Araport team at JCVI accepted by GenBank in June 2016. In 2023, TAIR is coordinating a community effort to produce v12. More information can be found at: tinyurl.com/AthalianaV12.

If you have information about misannotated or unannotated genes, please contact TAIR.

Genome Annotation Methods at TAIR

The TAIR structural annotation pipeline incorporated both manual and automated gene updates. Automated updates were carried out using the TIGR PASA annotation pipeline (Haas et al 2003). Arabidopsis cDNAs and ESTs were initially trimmed for contaminating vector sequences and poly-A tails removed (Seqclean). Initial alignments to the genome were carried out Via BLAT with a validation requirement of min 95% identity, min 90% transcript length aligned. Non-validated alignments were realigned via Sim4. Validated alignments were then assembled into distinct structures and compared to the pre-existing gene models. Strict validation criteria were applied, including min ORF size, max number of UTR exons, as well as a minimum length and similarity test. Manual annotators were provided with a web interface to review suggested automated updates. Alignments of Arabidopsis cDNAs, ESTs and plant proteins were viewed by TAIR curators in the Apollo annotation interface, where curators utilized them as evidence to support manual updates to gene structures.

Manual updates were also made following communication from the community of new or incorrectly annotated genes. Computational gene descriptions were generated using information derived from TAIR, Genbank, Uniprot and Interpro. Gene symbol, name and GO molecular function terms/s were derived from the TAIR DB. Protein similarity matches were identified from Arabidopsis and other species using BLAST. Protein domains were identified via InterproScan.

Genome annotation releases were made approximately every 9 months and were released via TAIR and NCBI.

TAIR wishes to thank Cornell University for use of the CBSU computer clusters at the Cornell Theory Center.

TAIR 10 Gene Annotation Data

TAIR10 Genome release announcement

The Arabidopsis Information Resource (TAIR) is pleased to announce the release of the latest version of the Arabidopsis genome annotation (TAIR10). The latest release builds upon the gene structures of the previous TAIR9 release using RNA-seq and proteomics datasets as well as manual updates informed by cross species alignments, peptides and community input regarding missing and incorrectly annotated genes.

TAIR10 statistics

The TAIR10 release contains 27,416 protein coding genes, 4827 pseudogenes or transposable element genes and 1359 ncRNAs (33,602 genes in all, 41,671 gene models). A total of 126 new loci and 2099 new gene models were added.

Eighteen percent (5885) of Arabidopsis genes now have annotated splice variants. Updates were made to 1184 gene models of which 707 had CDS updates. There were 41 gene splits and 37 gene merges. No changes were made to the Arabidopsis genome assembly for the TAIR10 release.

Gene annotation utilized available proteomics data (Baerenfaller et al., 2008 and Castellana et al., 2008) and RNA-seq data from the Ecker and Mockler labs (Lister et al. 2008, Filichkin et al. 2010). RNA-seq data was mapped to the Arabidopsis genome using TopHat, HashMatch or supersplat. After quality and low complexity filtering a total of ~200 million RNA-seq reads were successfully mapped to the genome. Of these, ~9 million represent spliced reads. Proteomics data and spliced RNA-seq reads were provided to Augustus and the resulting gene models categorised and manually reviewed. Validated gene updates, novel genes and novel splice variants were incorporated into the TAIR10 release. Additional spliced RNA-seq reads not already incorporated into gene models by Augustus were supplied to TAU. The resulting TAU models were again reviewed for potential novel splice variants. Transcript assemblies were generated via Cufflinks using all spliced reads and unspliced reads from the Ecker sets. Transcript assemblies were filtered and compared to existing gene models, resulting in the addition of 56 novel genes. Additional new proteome data provided to us by Katja Baerenfaller was used to directly update 24 gene models.

Gene models created using the Gnomon pipeline were provided to TAIR by NCBI. Reanalysis of these models for TAIR10 resulted in 11 additional novel genes, 67 additional alternative splice variants and 178 updates to existing genes.

Genome assembly updates (done for TAIR9)

In agreement with our reference genome policy corrections to the reference assembly were only made if supported by at least two independently derived sequence libraries from the Columbia ecotype.

The following updates were made to the chromosome sequences for the TAIR9 release:

- 227 single nucleotide substitutions were made to the assembly sequence based on re-sequencing data provided by Richard Clark (Ossowski et al. 2008) and Joe Ecker.
- 341 indels were made to the assembly sequence based on re-sequencing data provided by Richard Clark and EST and cDNA sequences deposited in Genbank that supported the insertion/deletion.
- 14 regions previously identified in TAIR8 as either vector, E.coli or rice contamination, and where the existing sequence had been substituted with the equivalent number of IUPAC ambiguity code 'N's were standardized (via deletion) to a set size of 100bp.
- All five nuclear chromosomes were updated for TAIR9 details of the golden path length of each chromosome can be found at here.

Further details of these TAIR 9 assembly changes and earlier TAIR8 updates are linked.

We would like to thank all those who contributed to the latest release by providing submissions for new and incorrectly annotated genes. TAIR wishes to thank Cornell University for use of the computer clusters at the Cornell Center for Advanced Computing (CAC).

Data availability

The fully annotated chromosome sequences in TIGR xml format or GFF format, along with FASTA files of cDNA, CDS, genomic and protein sequences, and lists of added, deleted and updated genes are available from the TAIR Download site.

Previous TIGR annotation is available from TAIR Download site.

For a summary of the different genome version statistics see table All Genome Versions Statistics below.

FASTA formatted files for all TAIR sequence analysis datasets including sets of intron, intergenic, UTR, upstream and downstream sequences are also available in the BLAST datasets directory of the TAIR Download Section.

Datasets are also available from TAIR's Advanced Gene Search; paste in or upload a list of AGI identifiers (such as At1g01010) and download the corresponding sequences. A graphic display of the Arabidopsis sequence and annotation can be viewed using TAIR's genome browsers.

Transposon genes and Transposable elements

Prior to TAIR8 all non transposon related pseudogenes and transposon genes were categorised as locus type pseudogene. For TAIR8, transposon related genes were reclassified into a distinct transposable element gene class.

Transposable element annotations provided by Hadi Quesneville were combined with pre-existing annotations to create a composite set of Arabidopsis transposons. These have been assigned a unique identifier (e.g. AT3TE53245) that indicates relative position on the chromosome. Under defined criteria (see additional readme-transposons) we have associated transposons to overlapping transposable element genes e.g. genes AT3G32022, AT3G32024, AT3G32026, AT3G32027 and AT3G32028 are associated to transposon AT3TE53245. Transposons can be viewed in TAIR's genome browsers and additional information can be found on the Transposon and Transposon family detail pages.

TAIR10 Genome Statistics

Chr	Protein coding	pre-tRNA	rRNA	snRNA	snoRNA	miRNA	Other RNA	Pseudo gene	TE gene	Total
1	7,078	240	0	2	18	53	118	241	683	8,433
2	4,245	96	2	0	15	29	83	217	826	5,513
3	5,437	93	2	7	15	30	66	202	878	6,730
4	4,124	79	0	0	11	28	62	121	711	5,410
5	6,318	123	0	4	12	37	65	143	805	7,507
Total	**27,206**	**631**	**4**	**13**	**71**	**177**	**394**	**924**	3,903	**33,323**
C	88	37	8	0	0	0	0	0	0	133
M	122	21	3	0	0	0	0	0	0	146
Total	**27,416**	**689**	**15**	**13**	**71**	**177**	**394**	924	**3,903**	**33,602**

All Genome Versions Statistics

	Protein Coding Genes	Transposons and pseudogenes	Alternatively spliced genes	Gene density (Kb /gene)	Avg. exons per gene	Avg. exon length	Avg. intron length

Araport11 (06 /16)	27,655	4,853	10,695		6.7	335.5	
TAIR10 (11 /10)	27,411	4,827	5,885	4.35	5.89	296	165
TAIR9 (6/09)	27,379	4,827	4,626	4.35	5.67	304	165
TAIR8 (4/08)	27,235	4,759	4,330	4.37	5.62	306	165
TAIR7 (4/07)	26,819	3,889	3,866	4.44	5.79	268	165
TAIR6 (11/05)	26,541	3,818	3,159	4.48	5.64	269	164
TIGR5 (1/04)	26,207	3,786	2,330	4.54	5.42	276	164
TIGR4 (4/03)	27,170	2,218	1,267	4.38	5.31	279	166
TIGR3 (8/02)	27,117	1,967	162	4.32	5.24	266	166
TIGR2 (1/02)	26,156	1,305	28	4.48	5.25	265	167
TIGR1 (8/01)	25,554	1,274	0	4.55	5.23	256	168
Nature (12/00)	25,498	NA	NA	4.50	5.20	250	168