

FAQ

The answers to some frequently asked questions can be found on this page. Use the broad categories listed below to jump to the section relevant to your question. If you cannot find the answer you are looking for you can send email to the TAIR curators at curator@arabidopsis.org.

- [TAIR access issues](#)
 - [My institution has a subscription to TAIR but I am getting page limit messages](#)
- [How to report a problem/bug on the website](#)
- [Genome Annotations](#)
 - [Where can I find a list of coordinates for all genes \(including UTRs, introns and exons\) and other transcripts?](#)
 - [What is the difference between genome assembly and genome annotation versions?](#)
 - [Where do the gene structural annotations in TAIR come from?](#)
 - [I have identified a new gene that does not have an AGI locus identifier- how can I get an ID before I publish?](#)
 - [Why do the coordinates for a gene/BAC/5'UTR etc... differ in TAIR from what I expect?](#)
 - [Are ALL Arabidopsis genes annotated? My gene seems to be missing.](#)
 - [How can I obtain a list of functional categories for a set of genes?](#)
- [Software/Analysis Tools](#)
 - [What are the specialized datasets available at TAIR for similarity searching \(e.g. BLAST, PatMatch\)?](#)
 - [How do I interpret the error message I get when I submit a sequence?](#)
 - [Why can't I upload more sequences into TAIR's NCBI BLAST?](#)
 - [What is the most recent version of the reference genome assembly?](#)
 - [How can I convert sequence locations in base pairs to centimorgans \(cM\) and vice versa?](#)
- [Gene Expression](#)
 - [Where can I find information about experimental conditions/design of AFGC microarray experiments.](#)
 - [How do I open microarray data files I have downloaded from TAIR ?](#)
- [Markers and Polymorphisms](#)
 - [Where can I find segregation data for recombinant inbred lines?](#)
 - [Why do the BAC locations given for SNPs and INDELS from the CERION database differ from their location when I try to locate them on the BAC?](#)
 - [Bulk Downloading Data](#)
 - [Where did the bulk download tools go?](#)
 - [Where can I get the most up to date TAIR data sets?](#)
 - [How can I read files I download from TAIR?](#)
 - [How can I obtain specific datasets such as all sequenced genes, or all markers in a defined region?](#)
- [Posting Jobs](#)
 - [How do I post a job opening at TAIR?](#)
 - [How do I find out about job openings at TAIR?](#)
- [New and updated data](#)
 - [How often is data updated in TAIR?](#)

TAIR access issues

My institution has a subscription to TAIR but I am getting page limit messages

If you are accessing TAIR off site -ensure that you are accessing through your library proxy as you would any other library resource.

If you are accessing TAIR on site- please contact your librarian with the IP address of the computer you are using to ensure that your address is within the registered ranges for your institution.

How to report a problem/bug on the website

If you are having a problem with the TAIR website/database that is software related (accessing content not the content itself) please submit a bug report and include:

- The specific URL of the page you were on when you encountered the problem. If you send a screenshot please include the address bar with the URL in the image.
- A description of the problem sufficiently detailed for us to reproduce the action.
- What operating system/browser you are using.

Genome Annotations

Where can I find a list of coordinates for all genes (including UTRs, introns and exons) and other transcripts?

Please go to the Downloads directory> [Genes](#) where you will find subdirectories for all of the genome releases. These lists are available in different formats such as gff and gtf formats and some contain specific subsets of genes such as all genes whose structures changed between releases TAIR9 and TAIR10 assembly, or all new gene models.

What is the difference between genome assembly and genome annotation versions?

Genome assembly refers to the ordered nucleotide sequence of each chromosome of the reference genome. The current version is called TAIR10. Genome annotation refers to the gene calls /coordinates on the reference genome. The current version is called Araport11.

Where do the gene structural annotations in TAIR come from?

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. As of 2014, genome annotation was handled by Araport (Araport11, released 2016).

Where can I find a list of GenBank accessions that correspond to AGI locus identifiers?

A file containing the mapping can be downloaded from the TAIR ftp site [TAIR10_genome_release](#).

I have identified a new gene that does not have an AGI locus identifier- how can I get an ID before I publish?

Please do not self assign AGI locus IDs, [contact us](#) first.

Why do the coordinates for a gene/BAC/5'UTR etc... differ in TAIR from what I expect?

There are several reasons why the coordinates of various gene/genome features can differ:

- The annotation may have changed between versions of the genome release. Currently we are using the Araport11 release for the SeqViewer and many BLAST datasets.
- BAC lengths in GenBank differ from AGI BACS. Many of the BAC sequences used by the AGI were trimmed to remove overlaps whereas clones with the same name in GenBank are untrimmed and longer. Other AGI BACS were extended by TIGR, adding sequences from adjacent BACs for easier annotation of genes near BAC ends. On SeqViewer, the AGI BACs are called Assembly Units.
- Structural annotations may differ between GenBank and AGI sequences. TAIR builds gene models based on a combination of cDNA and EST sequences from GenBank. A gene's UTR may therefore be extended based on EST alignments even though cDNA is in relative terms truncated.

Are ALL Arabidopsis genes annotated? My gene seems to be missing.

- Some gene models may not have been predicted by our annotation process because of limitations of the gene prediction software. If you are using BLAST or another sequence similarity search tool on TAIR, choose AGI whole genome (BAC clones), or GenBank whole genome (BAC clones) datasets. Unannotated genes will not be found in AGI transcript, protein, cds or gene data sets.
- "Missing" genes may reside in one of the few remaining known gaps in the sequence including highly repetitive regions that are difficult to sequence. See: [Genome Update](#) pages for information about gaps, incomplete clone sequences and genome monitoring status.
- If you have sequenced a gene that is not in the database, please [contact TAIR](#) to update gene structural and functional data based on experimental evidence.
- Some genes which existed in earlier annotation releases may have been obsoleted, or merged or split into 2 new gene models. To find out when a gene was added or removed, or whether a gene has been split or merged, try the [TAIR locus history search](#)

How can I obtain a list of functional categories for a set of genes?

TAIR curators annotate Arabidopsis genes using Gene Ontology terms to describe a genes molecular function, subcellular localization and biological process. GO annotations in TAIR also include contributions from the TAIR community, UniProt and the GO consortium. To obtain all of the GO annotations for a set of genes:

- Go to Advanced Search-> Gene Search
- Paste in or upload a list of locus identifiers for your genes (e.g. AT1G23030).
- From the result page choose Get GO annotations
- Download all results

Software/Analysis Tools

What are the specialized datasets available at TAIR for similarity searching (e.g. BLAST, PatMatch)?

TAIR's datasets include: AGI transcript, peptide, cds and gene sequences; introns only; insertion flank sequences; locus upstream and downstream sequences and more. For a full listing of the available data sets see: [Datasets](#). These data sets are common to BLAST and PatMatch. You can download TAIR BLAST datasets from the Downloads> [Sequences](#) directory.

How do I interpret the error message I get when I submit a sequence?

TAIR's algorithm for determining if a sequence is nucleotide or amino acid uses the following rules:

- For sequences submitted as proteins: an error is returned if the sequence contains the letters J or O, or if the sequence contains any of EFILPQXZ) and also has U.
- For sequences submitted as proteins: Sequences containing ONLY ACGTN or has the letter U but not any of (EFILPQXZ) are treated as nucleotide sequences.
- For sequences submitted as nucleotide: If the sequence has the letters J or O, or has any of (EFILPQXZ) and also has U, it will be considered not a valid nucleotide sequence.
- For sequences submitted as nucleotide: If the sequence has EFILPQXZ and does not have U it is treated as a protein sequence.

owner you will need to open the downloaded file using Safari. Alternatively, until this problem is resolved, you can choose to have the BLAST results returned via email, or use another browser on OSX such as Internet Explorer or Netscape.

Why can't I upload more sequences into TAIR's NCBI BLAST?

Currently we enforce limits on the size of queries (based on input file size and the dataset being queried) to ensure that some queries don't clog up the server. We are working on hardware updates that will enable us to modify these restrictions. If you need to process lots of sequences you can create your own local install of BLAST and use the TAIR blast datasets.

What is the most recent version of the reference genome assembly?

As of 2016 the current assembly (ordered sequence of nucleotides) for the reference genome is TAIR 10 which is the same as TAIR 9. In 2024 the assembly will be replaced with the Col-CC assembly.

How can I convert sequence locations in base pairs to centimorgans (cM) and vice versa?

Converting genetic locations to sequence locations will only give an approximate correlation. This is because the conversion depends upon both the genetic map used and the frequency of recombination which is variable within the genome. The most accurate estimate would be obtained by comparing the Lister and Dean RI map to the genome sequence (AGI map) for a global genome comparison value. You can also visually align genetic and sequence based maps using the MapViewer tool, by aligning common markers from the genetic and sequence maps.

Gene Expression

Please note that TAIR stopped accepting new microarray data submissions in June 2005. Newer and more comprehensive microarray data sets are available at [GEO](#), [ArrayExpress](#) and [NASCArrays](#).

Where can I find information about experimental conditions/design of AFGC microarray experiments.

You can now search for microarray experiments directly from the Microarray Elements Search. In addition, the Expression Viewer display now has direct links to the microarray experiment details. Click on the name of the hybridization in the Expression Viewer to display the information about the experiment.

How do I open microarray data files I have downloaded from TAIR ?

The raw data files for microarray experiments are large and therefore have been compressed. To uncompress the microarray datafiles (with .gz suffix) do the following: For MAC/UNIX/LINUX from the command line in a terminal window type in `gzip -d /home/yourname/yourpath/filename.gz`. For example: `gzip -d /home/frank/frankfiles/ciw_2000.gz`. For PC's you should download the WinZip utility to decompress the files.

Markers and Polymorphisms

Where can I find segregation data for recombinant inbred lines?

In the Downloads section under the [Maps](#) directory.

Why do the BAC locations given for SNPs and INDELS from the CEREON database differ from their location when I try to locate them on the BAC?

The difference position of the SNP provided by Cereon may be due to a difference in the length of the BAC reported in different versions of the BAC sequence. For example, many of the BAC sequences used by the AGI for the genome sequence had sequences at the ends trimmed resulting in a shorter length BAC. To obtain an accurate location on the genome try one of the following:

- Use TAIR's BLAST to match the SNP sequence against the AGI whole genome or Genbank whole genome datasets.

- Use the SeqViewer and
 - Paste the SNP/INDEL sequence into the input box.
 - Choose search by sequence.
 - Hits will be displayed as red lines on the chromosome bars.
 - Click on the chromosome bar to zoom in on the region.

Bulk Downloading Data

Where did the bulk download tools go?

Bulk download functions have been integrated into the [Gene Search](#) and [Protein Search](#). See the [Bulk Downloading Gene Data help document](#) for specifics.

Where can I get the most up to date TAIR data sets?

If you are a subscriber the most up to date data is available through the website on locus and other detail pages as well as bulk download tools. We also prepare quarterly data releases which include up to date gene names, locus summaries, Gene Ontology annotations, Plant Ontology annotations, and publications. Subscribers can access the most recent releases [here](#). Year old quarterly releases are available [here](#) for non subscribed institutions/users.

How can I read files I download from TAIR?

Tab-delimited files from the search results or Downloads site can easily be opened using a spreadsheet program such as Microsoft Excel or a text editor such as BBEdit.

- Open the file in the application and follow the instructions for choosing column delimiters (tabs) and column format (use text as a default for all).

How can I obtain specific datasets such as all sequenced genes, or all markers in a defined region?

Each of the search results pages includes the option to obtain a listing of specific records or a set of records that you can download and open in a spreadsheet. Information about the downloaded fields can be found in the help pages. If there is a specific set of data that you would like, contact the curators and we will do our best to accommodate your request.

Posting Jobs

How do I post a job opening at TAIR?

Please email a PDF/Word description of your job listing and/or URL to the posting in your website to curator@arabidopsis.org. We will put up the ad for no charge. Job postings are also routinely obtained as they are listed on the TAIR newsgroup. Your posting will appear on TAIR shortly after it is sent out to the newsgroup.

How do I find out about job openings at TAIR?

Jobs at TAIR are posted in the TAIR news section.

New and updated data

How often is data updated in TAIR?

Data in TAIR is constantly being updated and new data is constantly being added. Information about significant updates and additions is posted in the Breaking News section of the website as well as in the TAIR news section. If you are registered at TAIR, you can choose to receive quarterly news updates sent to your email address (see: [Registration Help](#) for information on registration and opting into the email updates).

Datasets that are frequently updated include:

- BLAST, PATMATCH datasets-AGI datasets and other datasets derived from the AGI sequences are updated after each major genome release. The current datasets reflects changes from the last TAIR release Araport 11 (April 2016), the next release will be TAIR12.
- Publications -downloaded weekly from PubMed.
- Gene/Locus summary updates-weekly
- Gene Ontology Annotations-weekly updates from TAIR curation, monthly from GO consortium
- Polymorphisms and Phenotypes - weekly updates from TAIR literature curation

Data Submissions and Corrections

How do I correct information about an incorrectly annotated gene?

- Send an email to curator@arabidopsis.org.
- Add a comment to the gene detail page for the gene in question. A curator will review the comments and contact you for more information if needed.

I have published a paper. How do I make associated software/ supplementary data available at TAIR?

Yes, please contact us and to arrange data/software submission.

How do I submit data to TAIR?

See [Data Submission section](#) for information on what data types TAIR accepts directly.

Gene nomenclature

Before I name my gene, how can I find out if a gene name is in use?

Consult the Gene Symbol List for a list of gene names that have been reserved and are not available. You should also search for the name in the TAIR database because the registry only contains a fraction of the names in use. UniProt is another source of curated gene names.

Related pages:

TAIR nomenclature guide

What is the difference between a hypothetical, unknown and putative protein?

Putative proteins are similar to a known gene. Unknown proteins are not similar to a known gene but do have EST or cDNA matches showing that they are expressed. Hypothetical proteins have no EST or cDNA matches and are not similar to a known gene, so there is no evidence that they are expressed genes.

FAQs for Developers

Frequently asked questions for software developers and casual programmers

What are the proper procedures for creating hyperlinks to TAIR pages?