# Quick Start

For people who are new to TAIR, this page is designed to address some basic questions and illustrate some common uses of TAIR.

### Contents


- TAIR basics
  - Types of data in TAIR and their uses
  - Database and website searching
- Common Tasks
  - Finding genes by sequence or structural similarity.
  - Using gene expression data to explore gene function and regulatory networks
  - Functional classification of genes
  - Genetic analysis of gene function:from genotype to phenotype and phenotype to genotype
- User support
  - Obtaining large and custom datasets.
  - Getting more help.


---

**What is TAIR?**

The Arabidopsis Information Resource is designed to be a comprehensive resource about Arabidopsis for the research community.It incorporates information about the structure and organization of the Arabidopsis genome and the functions of its estimated 29,000 genes. It includes DNA and Seed resources from the Arabidopsis Biological Resource that can be searched and ordered via TAIR. The key elements are a relational database where biological data and community resources for Arabidopsis is curated, integrated and stored, a suite of web-based tools for querying and analyzing the stored data and simple HTML pages containing information and links of interest to plant researchers. Sources of data include large scale sequencing and functional genomics projects, individual researchers and the literature. The home page serves as a gateway to public and commercially available data and tools about Arabidopsis for plant genomics research.

**Major datatypes, examples and uses.**

- **Genes and loci** includes predicted and experimentally verified genes. Over 14,000 genes have had their predicted structures supported or updated based upon experimental evidence. Also included are predicted, hypothetical genes for which there is no known expressed transcript and unknown genes where a function could not be assigned based on sequence similarity. In addition to displaying structural annotation of genes, each locus and gene detail page displays GO annotations functional classification.Genes and annotations can be searched and browsed on the SeqViewer graphical genome viewer, or using the Gene Search

- **Sequences** include the entire genome sequence of the Columbia ecotype from the Arabidopsis Genome Initiative(AGI) genomic, mRNA, expressed sequence tag (EST) and genome survey (GSS) sequences downloaded from GenBank and protein sequences from the AGI and SwissProt. Approximately 95 Mb of shotgun sequence from the Landsberg *erecta* ecotype contributed by Monsanto(formerly Cereon) is available for for registered researchers from academic and non-profit institutions.The Arabidopsis genome sequences have been used to generate custom datasets that are available for all sequence similarity search tools.Examples of the use of these datasets and tools are given in the next section.

- **Maps** include sequence, genetic and physical maps. The SeqViewer is a graphical genome browser and is richly annotated with features that facilitate whole genome analysis, positional cloning,visualization of the genome and gene structures and functional characterization of genes. The MapVeiwer can also be used for positional cloning by aligning regions on the sequence and genetic maps.

- **DNA** includes primarily ABRC stocks. There are variety of cDNA and genomic DNA clones, ESTs and vectors for cloning, plant transformation,assaying gene expression and protein function and localization. Other stocks include filters,genomic and cDNA libraries for screening and pooled genomic DNA for PCR screening of insertion lines. You can search for all types of DNA using the DNA Search .

- **Polymorphisms** include alleles such as T-DNA and transposon insertion lines, substitutions and small insertion/deletions (INDELS) from TILLing lines. Use the Polymorphism/Allele Search to find polymorphisms by type, features and location. There are also large sets of SNPs and INDELS from the Stanford Genome Center (SGC) that are polymorphic between the frequently used Landsberg *erecta* and Columbia ecotypes and SNPs from over 10 ecotypes, provided by the MASC consortium.The Cereon Polymorphism Collection ,

available to researchers from academic and non-profit institutions, includes over 50,000 SNPS and INDELS for the Col and Ler ecotypes. The SNPS and INDELS can be used to generate PCR based markers or used directly as markers for high throughput mapping to generate dense genetic maps.

- **Keywords** Keywords consist of ontologies to describe gene product function, biological process and subcellular localization from the The Gene Ontology Consortium and Developmental Stage and Anatomy ontologies for Arabidopsis thaliana (developed as part of the Plant Ontology Consortium). Keywords are used to annotate genes, loci, microarray experiments, publications and community. Keywords can be searched or browsed graphically to view the ontology structures. The keyword search option is also available for specific searches such as Gene Search, Publication, Community and Microarray Experiments.

- **Microarray Data** Public microarray data is available at TAIR as both raw and analyzed data.The primary source of the data is the Arabidopsis Functional Genomics Consortium (AFGC) cDNA arrays. Search for genes by expression value using the Microarray Expression Search , or find specific experiments with the Microarray Experiment Search . The Microarray Elements Search uses locus ids, Genbank accessions or element names to access clustered data and spot histories. Download entire datasets and microarray element-gene mapping files in tab delimited format from the FTP site. Analyzed data from over 370 microarray experiments can be viewed using hierarchical Java Tree Viewer or the VxInsight 'mountain' viewer .

- **Genetic markers** include PCR based markers such as CAPS, SSLPs and other types of markers used for mapping genes and QTLs and positional cloning.Details for the markers include primer sequences for generating markers, and information about polymorphisms detected in different ecotypes. The Marker search can be used to find markers; or search/browse markers on the SeqViewer.

- **Germplasms** are primarily ABRC stocks.They include mapping populations, uncharacterized and characterized mutant strains including TILLed lines, sequence tagged insertion lines from the SALK and other sources for analyzing the functions of genes using knockout mutations,natural accessions (ecotypes) which can be used for QTL identification and mapping and a number of transgenic lines containing reporter genes for assaying gene expression and protein localization. Many entries for mutant lines are now linked to images showing the mutant phenotype.

- Arabidopsis **proteins** have been associated to a variety of physiochemical features such as molecular weight and PI which can be used to find candidates from 2-D gels. Proteins have also been associated to known domains and motifs . The SCOP structural classification can be used to search and group Arabidopsis proteins.Types and uses of the structural annotations are described in the next section. There are two tools for accessing protein data, the Protein Search and Bulk Protein Download

- **Community** primarily consists of researchers and organizations working on Arabidopsis or other plant species. You can search for researchers who are working on related projects,or locate a potential collaborator using the Community search . Another way to find researchers working on a particular gene or process is to browse or search the projects on the Functional Genomics pages. Members of the community have been very pro-active about making data available through TAIR and we happily accept contributions.Data is linked the community via publications and to specific data they have contributed.These links can be used to identify sources of materials not found in the stock centers.

- **References** are primarily publications, personal communications and meeting abstracts. References are associated to many types of data to facilitate access to relevant information such as publications about a gene.References can be searched using the Publication Search

Data in TAIR is extensively interconnected through hyperlinks on the search results and the detail pages. For example, the locus detail page integrates and provides links to sequences, associated gene models, maps, protein properties, alleles, clones, germplasms, genetic markers and related publications. In addition,these pages facilitate ordering of stocks from the ABRC such as clones and germplasms.

**Database and website search basics.**

The simplest way search to search the database is to use the simple search located in the upper right corner. Select the type of data you are searching for from the drop down menu. For example, you can search for genes, germplasms, clones and other data types in TAIR. If you choose the exact name search, this will find ANYTHING in TAIR having the name you submit. For a complete list of fields searched with the simple search see the Search and Navigation Help documents.

You can also use the QuickSearch to perform a Google search of the static web pages. Enter a word (such as a gene name) or a short phrase (such as "gene knockouts" or "protein localization") and select Google TAIR Website from the drop down menu.

The Advanced Searches allow you to restrict your search to specific data types and select from available parameters to formulate complex queries.

SeqViewer (TAIR's genome browser) displays annotation units, genes, transcripts, ESTs,full length cDNAs, genetic markers, polymorphisms (including substitutions, small insertions and deletions and T-DNA and transposon insertions) from the whole chromosome down to the nucleotide level. This is a multi functional tool with many applications some of which are described in the following section.

You can use TAIR's AraCyc database to search or browse plant metabolic pathways and obtain detailed information about enzymes and pathways including reactions, compounds, intermediates,and co-factors.This detailed picture of the pathways can be used to find candidate genes and pathways for metabolic engineering and aspects of plant metabolism that need to be elucidated.

**For more information about the database and data sources and types see:**
TAIR Data Sources: Information about the major sources of data.
About TAIR: for information and publications where you can find more detailed descriptions of TAIR database and tools.

**Common Tasks**

Many people use TAIR because they are interested in analyzing a particular gene from Arabidopsis or another organism or classes of proteins or specific processes.

**Finding genes by sequence or structural similarity**.

Sequence Similarity SearchingTAIR provides BLAST for finding sequences based on nucleotide or amino acid similarity. Many BLAST data sets are available for searching to find genes, proteins, transcripts, 3' and 5' UTRs, introns and intergenic regions and analyze genome composition. These tools provide an entry point for analyzing characteristics and functions of Arabidopsis genes and gene families. Match results to Arabidopsis genes are hyperlinked to the locus detail pages which are launch point for obtaining structural and functional data.Gene familiesHomologous genes (genes that share a common ancestor), often have similar functions. The relationships among members of a gene family can be analyzed using algorithms that evaluate the evolutionary relationships and graphically display in a variety of ways such as phylogenetic trees or dendrograms.Sets of locus identifiers for genes from Arabidopsis can be imported into the Bulk Sequence Retrieval tool to obtain a set of FASTA formatted sequences which can be uploaded into multiple sequence alignment tools such as ClustalW. Gene families contributed by members of the research community can be browsed and downloaded from the Genes/Gene_families directory on the FTP site.You can visualize the distribution of members of a gene family by uploading the set of locus identifiers to into the Chromosome Map Tool. This whole genome view can reveal sites of potential tandem duplication (leading to gene duplication).Motif/Domain findingAnalysis of protein structural features can be used to find common functions for proteins having shared domains as well as novel domains and functions. Arabidopsis proteins can be searched and grouped by Interpro, SMART, Pfam, ProDom, PRINTS and PROSITE domains , SCOP structural classification and sub-cellular localization based on predicted target sequences using the Protein Search.The Bulk Protein Search has many of the same parameters as the Protein search. You can search all Arabidopsis proteins or within your own defined set of proteins. Sets of protein ids can be uploaded to generate a custom list of proteins and along with selected physical-chemical features. Comparison of protein sequences may also reveal the presence of novel domains and motifs. The PatMatch tool can be used to find proteins having exact or degenerate matches to sequences corresponding to known and novel domains and motifs.

**Using gene expression data.**

Finding patterns of expression for gene(s) of interest.The Microarray Expression Search can be used to search and display global patterns of gene expression for individual or sets of genes. or find specific experiments with the Microarray Experiment Search .Using public microarray data.TAIR makes public microarray data submitted by users available in both raw and analyzed formats. The raw data can be searched and downloaded using the the Microarray Experiment Search , or from the FTP site. The data files can then be imported into microarray analysis software of your choice.Finding and analyzing co-clustered genes.TAIR has taken the data from public microarray experiments and generated cluster data using a subset of these experiments. In order to ensure the quality of the analyzed data, some hybridizations were not included (for example, to eliminate spatial bias). The resulting clusters can be accessed in a number of ways.The Microarray Elements Search uses locus ids, Genbank accessions or element names to access clustered data and spot histories. Two other tools are available for viewing TAIR's analyzed microarray data.The Java Tree Viewer displays the data in hierarchical cluster format. The VxInsight viewer presents the same data in a more intuitive way. Each cluster appears as a mountain; the view can be zoomed from a birds eye over view down to the level of a single element. Each mountain can be interrogated visually. Searching include GO terms, gene names, enzyme names (EC number) and pathway names which allows you to find the distribution of processes, gene families etc... within the clusters. Sets of clustered genes can be downloaded for further analysis. For example, genes of unknown function may cluster with a group of genes involved in carotenoid biosynthesis. All members of the cluster 'mountain' can be downloaded as a list. The list in turn can be used to search for insertion lines to analyze mutations in the unknown genes and determine if carotenoid biosynthesis is affected by the mutated genes. To find potential regulatory regions in co-expressed genes use the Motif Finder which finds short (6 base pair) motifs that are over-represented in the 500 base pairs of upstream sequence.

You can also overly gene expression data on the AraCyc metabolic map to find pathways that are similarly or differentially affected by a treatment. If your data has multiple time points, you can display an animated overview. The animation shows how each pathway changes over time.

**Functional classification of genes.**

Both TAIR and TIGR are annotating Arabidopsis gene products with molecular function,biological process and sub-cellular localization using the Gene Ontology controlled vocabularies. The associations are made using experimental data from the literature and computational methods. The annotations can be used to find groups of genes with similar features, as well as distinguishing distinct functions among homologous genes. In addition, TAIR is using controlled vocabularies for Arabidopsis anatomy and growth/developmental stages to annotate patterns of gene expression and mutant phenotypes.

Finding sets of genes with shared functions, processes and localization.Genes can be searched based on GO annotation using the Gene Search which allows you to limit your search based on evidence for displaying annotations. For example, you can limit a search for genes localized to the chloroplast to only those that are experimentally verified. Correlation of co-clustered genes with Gene Ontology Annotations can be used to predict functions of unknown genes or new roles for characterized genes. The organization of keywords into ontologies makes it possible to use general terms for a search and find not only genes associated to the term, but to the children terms as well.The Keyword Search/Browser can be used to find any data associated with a given keyword and to view the organization of the ontologies. The keyword search can also be used to find genes expressed in the same body part (anatomy) or at the same time (same growth or developmental stage).

Functional categorization of a set of genes:Groups of genes such as co-expressed genes identified in microarray experiements, can be ordered into functional categories using their Gene Ontology associations using the Bulk GO Annotation Download Tool. Selecting the Functional Categorization option allows you to generate table showing the distribution of functions associated with genes in your list. The classification is broken down according to each aspect of the GO: molecular function, subcellular localization and biological process. Annotations are grouped as broad categories to provide a general overview. The data can be displayed graphically in the form of a pie chart that can be saved as an image file and used for publications. The categorization can also be saved as a tab-delimited list which can be imported into a spread sheet program and used to generate custom graphs and charts.

**Using forward and reverse genetic resources to dissect a biological process or the functions of specific genes.**

Genetic analysis in Arabidopsis has been greatly enhanced by systematic efforts to generate mutations in all Arabidopsis genes and an extensive collection of natural variants of Arabidopsis.

To find mutations in a specific gene use the gene name or locus identifier in the Polymorphism/Allele search. The Germplasm search can be used to find strains T-DNA or transposon insertions in specific genes, or mutant strains having a common phenotype, or mutagenized lines that you can screen for new phenotypes. Analysis of quantitative trait loci can reveal genetic mechanisms underlying adaptation to local environments. The Ecotype/Species search can be used to select appropriate accessions for QTL analysis.

The SeqViewer is a useful tool for both forward and reverse genetics approaches.You can find and obtain lists of genes between flanking markers, find closely linked genetic markers or find sequence polymorphisms for generating new markers for positional cloning of genes defined by mutation or quantitative trait loci. For reverse genetics, it can be used to find polymorphisms including T-DNA insertions from the SALK Institute and substitution/small indel alleles generated by the TILLing project and see their locations in specific genes or regions of the genome.

Another way to find knockouts that is especially useful for finding mutations in multiple genes is to use the sequence similarity search tools where you can upload a sequence or set of locus identifiers and search the custom dataset of insertion flanking sequences. Note that it is best to use sequence such as AGI genes (plus UTR and introns) to maximize the chance of identifying an insertion anywhere in a gene (not just in exons).

**User support**

**Obtaining large or custom datasets**.

You can obtain datasets in a tab-delimited format which can be used with spreadsheet programs such as Microsoft Excel or imported into other tools for further analysis. For example you can can use the Advanced database searches to d sets of results that meet your selection criteria (e.g. a list of polymorphisms between two ecotypes in a specific location on a chromosome) . If you have a set of locus identifiers for a group of genes, you can use any of the bulk download tools such as the bulk GO annotation download , Sequence Download , bulk Protein download , and the Microarray Elements Download . Many frequently requested datasets such as mappings between microarray elements to genes, all Gene Ontology annotations, and datasets used to generate the SeqViewer maps, can be downloaded from the FTP site. If possible, we will accommodate requests for specific datasets which are then made available in the User Requests directory on the FTP site.

**How can I get more help?**

To access the help documents for TAIR, you can click on "Help" in the toolbar. This will take you to the main help page. Most pages also link to a help document specific to the tool or data type (e.g. Germplasm searching, BLAST). You can also email us with your specific questions/concerns:

- curator@arabidopsis.org: for general questions/comments.
- abrc@osu.edu:for questions about Seed and DNA stocks from the ABRC.
- curator@arabidopsis.org:for problems with/updates to gene annotations.
- curator@arabidopsis.org: for questions and problems with software/tools.