UNIT 1.11

Using The *Arabidopsis* Information Resource (TAIR) to Find Information About *Arabidopsis* Genes

Leonore Reiser[1], Shabari Subramaniam[1], Peifen Zhang[1,2] and Tanya Berardini[1]

[1]Phoenix Bioinformatics, Newark, CA USA

[2] Computercraft, Washington, DC USA

ABSTRACT

The *Arabidopsis* Information Resource (TAIR; *http://arabidopsis.org*) is a comprehensive Web resource of *Arabidopsis* biology for plant scientists. TAIR curates and integrates information about genes, proteins, gene function, orthologs gene expression, mutant phenotypes, biological materials such as clones and seed stocks, genetic markers, genetic and physical maps, genome organization, images of mutant plants, protein sub-cellular localizations, publications, and the research community. The various data types are extensively interconnected and can be accessed through a variety of Web-based search and display tools. This unit primarily focuses on some basic methods for searching, browsing, visualizing, and analyzing information about *Arabidopsis* genes and genome. Additionally, we describe how members of the community can share data via JBrowse and the Generic Online Annotation Submission Tool (GOAT), in order to make their published research more accessible and visible.

Keywords: Arabidopsis ● databases ● bioinformatics ● data mining ● genomics

INTRODUCTION

The *Arabidopsis* Information Resource (TAIR; *http://arabidopsis.org*) is a comprehensive Web resource for the biology of *Arabidopsis thaliana* (Huala et al., 2001; Garcia-Hernandez et al., 2002; Rhee et al., 2003; Weems et al., 2004; Swarbreck et al., 2008, Lamesch, et al., 2012, Berardini et al., 2015). The TAIR database contains information about genes, proteins, gene expression, mutant phenotypes, germplasms, clones, genetic markers, genetic and physical maps, genome organization, publications, and the research community.

TAIR is a curated database; data are processed by Ph.D.-level plant biologists who ensure their accuracy. Curation adds value to the large-scale genomic data by incorporating information from diverse sources and making accurate associations between related data. Data from manual literature curation, such as protein localization, biochemical function, gene expression, and phenotypes, are added to the corpus of knowledge presented for each locus in the genome. TAIR aims to produce a 'gold standard' functionally annotated plant genome that plant biologists can use as a reference for understanding gene function in crop species and other plants of importance to humans (Berardini, et al., 2016).

This is the submitted version. For the final, edited version see:
Reiser, L., Subramaniam, S., Zhang, P., & Berardini, TZ., (2022) Using the Arabidopsis Information Resource (TAIR) to find information about Arabidopsis genes. Current Protocols in Bioinformatics. DOI:10.1002/cpz1.574

The database content and other information relevant to plant scientists can be accessed through dynamic Web interfaces and static hypertext (HTML) pages. Users can perform simple searches of much of the database using names or keywords. Advanced search forms for different data types are used for more complex or specialized queries. Genomic data can be accessed through text-based queries, via the graphical genome browsers (see *BASIC PROTOCOL 2* and *BASIC PROTOCOL 3*), and sequence similarity tools such as BLAST. Data from TAIR can also be obtained in bulk from selected query tools (see *BASIC PROTOCOL 5*) and downloaded from the web site. TAIR provides an extensive set of links from the database and web site to other sources of *Arabidopsis* genomic data around the world.

- The data and services of TAIR are organized into eight categories, which appear on the navigation toolbar on all TAIR pages. Text-based query tools for performing simple and complex searches of specific types of data in TAIR, such as genes (see *BASIC PROTOCOL 2*), DNA, proteins, polymorphisms (including alleles), people, laboratories, and germplasms are found in the **Search** section. The **Browse** section allows the user to browse the Gene Symbol Registry, *Arabidopsis* transposon families, *Arabidopsis* gene families, as well as Gene and Plant Ontology terms (see *BASIC PROTOCOL 3*) and recently added literature (see *BASIC PROTOCOL 9*) and other data types. Within the **Tools** section are TAIR's graphical genome browsers (SeqViewer, JBrowse and GBrowse; see *BASIC PROTOCOL 3*), MapViewer for aligning physical and genetic maps, sequence similarity software (NCBI BLAST), Motif Analysis and Patmatch (see *BASIC PROTOCOL 6*), the TAIR Synteny viewer (see *BASIC PROTOCOL 10*), the literature full-text search tool Textpresso (see Commentary), an *Arabidopsis* chromosome map tool (see Commentary), among other data analysis and visualization tools. Under the **Tools** section, one will also find tools for downloading sets of sequences, protein data, Gene Ontology assignments (see *BASIC PROTOCOL 5*), and other curated data sets for a set of genes, as well as a GO term enrichment tool for Arabidopsis and other plant species (see *BASIC PROTOCOL 4*). The **Portals** section hosts pages with links to other databases and Web sites containing useful data and tools. **Portals** also contains comprehensive lists of community resources curated by the Bioinformatics section of the Multinational Arabidopsis Steering Committee and TAIR curators (https://conf.phoenixbioinformatics.org/display/COM/Resources). The **Download** directory contains several logically organized directories containing large data sets related to genes, sequences, Gene Ontology annotations and more. The **Submit** section contains forms and documentation for submitting data to TAIR. Users can contribute functional annotation for submitted papers using the web-based Generic Online Annotation Tool (GOAT; see *BASIC PROTOCOL 8),* or by providing data in preformatted spreadsheets. TAIR also maintains the Gene Symbol Registry for Arabidopsis and registered users can submit Gene Symbols via the web. In the **News** section are links to the *Arabidopsis* community newsgroup, announcements from TAIR, meetings, and job postings.

For *Arabidopsis* to be effectively used as a reference plant species, it is essential that researchers know what data are available, how to use the information they obtain and the provenance of that data. This unit includes several basic protocols for accessing the wealth of information about *Arabidopsis* genes that has been generated by the research community and made available through TAIR. The types of data and tools at TAIR are diverse and cannot all be described in a single unit. Therefore, this unit focuses on the data and tools that are related to retrieving, mining, and visualizing information about Arabidopsis genes. These protocols are based upon data and

tools available as of June 2022. As with any actively updated Web-based informatics resource, the data and tools will change over time.

BASIC PROTOCOL 1

TAIR HOMEPAGE, SITEMAP, AND NAVIGATION

The TAIR home page (*http://arabidopsis.org*) is the main entry point to the database and Web site (Fig. 1.11.1). To facilitate navigation of the TAIR Web site, a navigation toolbar is located at the top of all TAIR pages containing headings such as **Tools**, **Search**, and **Portals**. When mousing over each item in the toolbar, a drop-down menu appears with clickable submenus that lead to a variety of datasets, tools, and external links. Several additional buttons are located above the main toolbar, including items such as **Help**, **About Us**, **Subscribe**, **Register** and **Login**. The **Help** section of the Web site (*http://arabidopsis.org/help/*) provides a quick guide to new users, frequently asked questions, a glossary of terms used on the web site, tutorials, a search help function, and user guides for database searches, specific tools, and registration. Registered users can click on **Login** to register gene symbols and update personal information. The **About Us** section has information about the project and staff. Users who access TAIR via institutional subscriptions will see their institution name displayed in the upper right side of the main toolbar. On the bottom of the home page are quick links to connect with TAIR via social media (Facebook and Twitter) as well as our YouTube channel where users can view video tutorials.

Necessary Resources

Hardware

Computer with Internet access

Software

Up-to-date Web browser. The browser must have cookies enabled to log in and submit gene symbols. See *http://www.arabidopsis.org/help/index.jsp* for information on properly configuring one's browser.

Performing a quick search

1. Go to the TAIR home page (*http://www.arabidopsis.org*). Type the search term into the text box in the upper right corner of the page and choose a category from the drop-down menu (see Fig. 1.11.1 item a). Click the Search button.

   The quick search performs a name search for most of the objects in the database (e.g., Genes, Clones, ESTs or BAC ends, People/Labs, Polymorphisms/Alleles, Germplasms, Ecotypes, Keywords, Genetic Markers, Proteins and Vectors). By default, this is a "contains" search (a search for aba1 *retrieves both ABA1 and ATRABA1A). It is also important to be aware that this search is not limited to the name field. For example, if the gene category is chosen, the gene description and keywords fields will be searched as well as the name. This is done to avoid*

missing any potentially relevant results, but may produce a large number of results. Note that searching for Metabolic Pathways sends the query term to the Plant Metabolic Network Database (http://pmn.plantcyc.org/.)

2. A list of all matching records is displayed for the data type chosen. Click on each record to access full details for that object, or download the current page of results using the download button at the top of the page. For gene search results, the additional option "download all" provides a way to download the entire result set at once, and "get all sequences" provides an option to download sequences for all the genes in the result set.
3. Alternatively, to search for any data type in TAIR by name, choose "Exact name search" from the drop-down menu to the right of the box where the search term was typed in step 1. The query will return a summary (TAIR Search Result) page listing **all** data types with matching records and the number of records for each data type. Click on any item in the list to display a summary of all the records retrieved for that data type. In this example, clicking on Proteins displays a list of the two ABA1 proteins encoded by different splice forms of the *ABA1* gene.
4. In the event that a general query returns too many results, try an Advanced Search for the specific data type (see *BASIC PROTOCOL 2* for an example of an advanced search for Genes). The advanced search parameters can be used to narrow down an overly broad query.

BASIC PROTOCOL 2

FINDING COMPREHENSIVE INFORMATION ABOUT *ARABIDOPSIS* GENES

The locus detail pages represent the most comprehensive starting point for a researcher interested in finding out what is known about a gene. The physical location of an annotated gene on the genome is called a locus in TAIR. The locus serves as a useful concept for grouping genes with other objects having the same genomic location. For convenience, genetically defined genes (i.e., those identified by linkage studies but which are not yet associated with a genomic sequence) are also included as loci that have a genetic, but no physical location. Each locus is associated with at least one gene model, which can be thought of as a version of a gene. Several gene models (labeled as splice variants in TAIR) can be associated to a gene locus based on the existence of predicted or verified alternative transcripts. Every sequenced locus is assigned a unique identifier, the Arabidopsis Genome Initiative (AGI) locus identifier. This has the format AT (for Arabidopsis thaliana) X (where X is either a number from 1-5 corresponding to one of the 5 nuclear chromosomes or C for chloroplast or M for mitochondrion) NNNNN (a 5 digit number). The locus detail page collects information such as gene symbols and full names, experimentally determined or predicted function, gene expression data, mutant phenotypes, associated germplasms, polymorphisms, clones, and publications. Because data in TAIR are highly integrated, it is possible to access the locus detail page from detail pages of almost every other type of object in the database. This protocol illustrates a commonly used way of finding genes using the Advanced Gene Search form.

Necessary Resources

See Basic Protocol 1

Searching for information about a specific gene or set of genes

1. Go to the TAIR home page (*http://www.arabidopsis.org*). In the top navigation bar click on the Search header (see Fig. 1.11.1) and select the Genes link to go to the TAIR Gene Search page (*http://www.arabidopsis.org/servlets/Search?action=new_search&type=gene*).
2. To search by name, choose "Gene name" as the option from the Search Name drop-down menu (the options include "Gene name," "description," "phenotype," "GenBank accession," "GenBank gi", "Locus TAIR object ID" or "Gene TAIR object ID"). Using the drop-down menu to the right of this, set the search to an exact match or an inexact match (the options are "contains," "starts with," "ends with," or "exactly") and type the name in the text box on the right-hand side of the same line. For example, to find a set of related genes sharing a gene symbol, such as ARF for Auxin Response Factor family members (Hagen and Guilfoyle, 2002), type in ARF as the name term and choose the "starts with" option to the left of this. Click the "submit query" button.

   Gene names include systematic names assigned based on chromosomal location (so called 'AGI locus identifiers' such as AT1G01010) or gene symbols. For more information about Arabidopsis gene nomenclature, see the Arabidopsis Gene Nomenclature Guidelines (http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp).

3. All of the loci that match the query term will be displayed in a list of results (on a page titled TAIR Gene Search Results). Click on the locus name to view the locus detail page. A sample locus detail page obtained by using the search name ABA1, and then selecting the AT5G67030 locus from the TAIR Gene Search results page, is shown in Figure 1.11.2.

   The default search only retrieves genes that are active in the database. Checking the "include obsoleted genes" check box will retrieve both active and obsoleted genes, along with the history of their status in the database. Genes may become obsolete if they are merged with other genes—or if improved genome annotation methods find inadequate evidence for their existence. TAIR retains information about obsolete genes in order to maintain a record of their histories and associations.

Using the detail pages to find information about a locus

4. On a locus detail page (Fig. 1.11.2) related data are grouped together into different sections. The following annotations (the red lettered items on the left side in Fig. 1.11.2) summarize the typical information displayed on a locus detail page. Definitions of each data type can be obtained by clicking on the adjacent question mark image to display a pop-up definition window.

   a. Representative gene model and summary information (Fig. 1.11.2A, a items). Unless there is specific experimental evidence to support one transcript over another, the default representative gene model for a protein coding gene is

the gene model with the longest coding sequence (CDS); for other gene types, the representative model is set as default to the .1 model. If there are multiple CDS of the same length the lowest numbered gene model is the representative model.

Data in this section includes Gene Model Type, Other Names and Summary. Example gene model types are protein coding, pseudogene, non-coding RNA, among others. Other names include gene symbols and full names curated from the literature or provided by researchers via the Gene Symbol Registry. The Description field is a short summary of the gene's function either manually composed by a curator or computationally generated. The latter is only shown if the locus has not yet been curated manually. Descriptions from Araport 11 were computationally generated (Cheng, et al., 2017).

b. Other Gene Models/Map Image (Fig. 1.11.2A, b items). Links to other gene models (termed splice variants in TAIR) are displayed below the representative gene model information. Clicking the gene model name will open a new window displaying the gene model detail page. View this page to see gene model specific data such as gene features in a tabular format and annotations that are specific to individual gene models. The Map detail image is a graphical display of the exon-intron boundaries of all the gene models of a locus. Clicking on the image directs the user to JBrowse (see *BASIC PROTOCOL 3*)

c. Gene function, biological role, and localization (Fig. 1.11.2A, c item).

*The **Annotations** section contains all of the controlled vocabulary terms that have been assigned to describe the molecular function, biological role, subcellular localization, and expression of the gene product. The annotations are grouped according to the type of vocabulary and summarized on the locus page. Click on the **Annotation Detail** link (located at the bottom right of the Annotations section) to display the full annotation details, which include the type of evidence supporting the annotation and the corresponding reference that is the source of the data supporting the annotation.*

d. Sequences (Fig. 1.11.2A, d item).

*Links to genomic sequence, full-length CDS, full-length cDNA, and protein sequence are located in the **Sequence** section. Clicking on the sequence name will display a new window containing the sequence, which can be uploaded directly into TAIR's BLAST tool. TAIR BLAST includes specialized Arabidopsis sequence data sets such as intergenic regions, upstream and downstream sequences, and UTRs (http://www.arabidopsis.org/help/helppages/BLAST help.jsp#datasets). BLAST can also be accessed from the TAIR homepage under the **Tools** section.*

e. Gene expression (Fig. 1.11.2A, e item).

*Information about the expression of the gene can be found in the **RNA Data** section and the lower part of the **Annotations** section. In the RNA Data section, array elements from one-channel and/or two-channel experiments that map to the locus are listed. Array element names are linked to detail pages. Note that TAIR stopped integrating and updating microarray data in 2005, see Commentary for more current datasets and tools. Lists of full-length cDNAs and expressed sequence tags (ESTs) can be found in the **Associated Transcripts** subsection within the RNA Data section. Click on the number next to the type name to see a list of all the clone records. The clone records are linked to GenBank, where information about the cDNA libraries (and therefore expression) can be found. Finally, information about gene expression, curated from the literature, is shown in the **Annotations** band along with the Plant Ontology associations (see Fig. 1.11.2A, section "c": "expressed during", "expressed in").*

f. BAR eFPbrowser image

Gene expression data are displayed using the Application Programming Interface (API) to show the electronic pictographic representation of gene expression from the BioAnalytic Resource (BAR, http://bar.utoronto.ca/). Users can toggle between different views representing different experimental datasets using the 'Data Source' drop down menu. Users can also click on the link to the BAR website to make use of the features of this resource.

g. Protein data (Fig. 1.11.2A, f item).

Structural and physical characteristics of the protein encoded by the reference gene model, including molecular weight, conserved domains, and isoelectric point, are displayed in this section. Click on the AGI name in the protein section to open a new window displaying more detailed information and the amino acid sequence itself.

h. Plant homologs (Fig. 1.11.2A, g item).

*The **Plant Homologs** data section displays information about proteins that are evolutionarily related to the locus. The homologs are phylogenetic predictions made by the PANTHER project (Mi et al. 2017). Clicking on the PhyloGenes tree view glyph will open the corresponding gene tree in PhyloGenes (Zhang et al., 2019) in a new window. PhyloGenes presents gene function data alongside phylogenetic trees to aid in gene function prediction or comparing functions in different species (see BASIC PROTOCOL 8). The Arabidopsis paralogs display includes links to the individual locus pages for each paralog. Users can also click on a button to retrieve paralog sequences, or download a list of AGI locus IDS to use in subsequent analyses. Plant homologs from the PANTHER families are listed according to their taxonomic distribution and the entire list can be downloaded and saved as a tab delimited file. The 'Search Gene Families' section uses the selected locus ID as a query to search for orthologs in external resources such as Ensembl Plants (Bolser et al., 2016), PLAZA (Proost et al.,*

2015) and Phytozome (Goodstein, et al., 2012), among others.

i. Map locations (Fig. 1.11.2A, h item).

*The **Map Locations** section displays the chromosome and coordinates of the locus for the maps on which it is found. The gene can be viewed in a whole-genome context by clicking on one of the three map options (Map Viewer, Sequence Viewer, GBrowse and JBrowse) in the Map Links section (See BASIC PROTOCOL 3).*

j. Markers, Alleles and Polymorphisms (Fig. 1.11.2A, i item).

Genetic markers that map within the locus are displayed along with the type (e.g., visible, RFLP, SSLP etc.…). All of the polymorphisms that map within the locus are shown in the Polymorphisms section, along with the type of variation. This section includes natural variations found in different ecotypes and induced mutations (e.g., T-DNA insertions) that have been mapped by sequence identity and alleles that have been curated from the literature. To find detailed information about a polymorphism, click on the name of the polymorphism.

k. Germplasm information (Fig. 1.11.2A, j item).

The Germplasm section provides information on all germplasms available for a locus, including phenotype descriptions and images of plants (if available). If a germplasm is a stock, that Stock ID and a link to ABRC will be shown. Links to the three main stock centers are provided at the bottom on the Germplasm section to facilitate searching for the stocks at those resources.

l. Clones (Fig. 1.11.2B, k item).

Clones linked to a locus may include vectors, BACS, clone ends (ESTs) that contain sequences from the locus of interest. If the clone is an available stock, that information will be displayed along with a link to order from ABRC.

m. External links (Fig. 1.11.2B, l item).

There are other web sites that provide either alternate views or different information about a locus (see Commentary). In order to provide access to as much information about a locus as possible, TAIR provides links to the corresponding locus pages in other databases and Web sites. Types of external links include other Arabidopsis genome annotation databases, gene expression databases, and functional genomics sites, as well as links to tools for further analysis. For example, all sequenced loci are linked to other Arabidopsis annotation databases including Thalemine at BAR, NCBI, and Gramene. Links are grouped by data types such as: Genomics, Expression/Localization, or Interactions. TAIR also provides links to UniProt and NCBI Reference genome from the protein detail pages.

n. Community Comments (Fig. 1.11.2B, m item).

Comments may contain additional data contributed by registered TAIR users, and are included in the display for nearly all of the TAIR detail pages. This function can be used to report new data, as well as errors or omissions related to the displayed object (see http://www.arabidopsis.org/help/helppages/addcomment.jsp).

o. Publications (Fig. 1.11.2B, n item).

Papers and conference abstracts are shown at the bottom of the detail page in the section marked Publications. Publications include published literature imported from PubMed, Agricola, and BIOSIS, along with abstracts from the International Conference on Arabidopsis Research. Only the most recent 15 papers are listed on the detail page; to retrieve the complete list, click on the View Complete List link. Clicking on the title of the publication opens a new link to the detailed record where one can read the abstract, link to the PubMed citation, associated loci and annotations, and find authors among TAIR's community. Users are encouraged to contact the TAIR curators to report missing or incorrectly associated papers.

p. Update History (Fig 1.11.2B, o item).

TAIR maintains a history of changes to the genome annotation status a locus for the purposes of tracking. Recorded changes include merges, splits or insertions.

Saving the results of a search to a file

1. Return to the list of results obtained by the query submitted in step 2 (page titled TAIR Gene Search Results). Check the box to the far left of the results summary. Each page of results must be saved separately. Only those results that are selected will be saved. Use the Check All function to save all of the results displayed on the page.

   Before downloading a large set of results, use the browser to go back to the Advanced Search page, make sure the number of records per page of results is set to the maximum (usually 200 records/page), and resubmit the query.

2. After selecting all of the desired results on a page, click on the Download Checked button (or Download All if you wish to export all results) in the upper right corner of the TAIR Gene Search Results page. The checked results will then be displayed in the browser window as tab-delimited text file. Use the Save As function under the File menu in the browser toolbar to save the results in a file on the local computer. This process must be repeated for each page of results.

3. In order to retrieve sequences for the selected results, click on the Get Checked Sequences button (or Get All Sequences if you wish to retrieve sequences for all results) on top of the TAIR Gene Search Results page. This will bring you to the

Sequence Bulk Download and Analysis page from where you can retrieve different types of sequences for your list of genes. For more information about that tool, see *BASIC PROTOCOL 5*.

The download feature is found on all of the search results pages. Each set of results includes different information in the downloadable file. See the help documents for the specific search to view a listing and description of the downloaded fields. The files contain tab-delimited text that can be opened using a text editor or spreadsheet software such as Microsoft Excel. The download sequence option is only available on the Gene Search Results page.

BASIC PROTOCOL 3

USING THE *ARABIDOPSIS* GENOME BROWSER-JBROWSE

TAIR provides three Web applications (JBrowse, SeqViewer and GBrowse) that allow users to explore the annotated *Arabidopsis* genome sequence. SeqViewer is a graphical genome browser developed by TAIR while JBrowse (Buels et al., 2016) and its precursor, GBrowse (Stein et al., 2002) were developed by the Generic Model Organism Database project (GMOD; www.gmod.org). These tools allow the user to search for and display various sequence features such as genes, polymorphisms, T-DNA insertions, and transcripts (ESTs/cDNAs), provide a mechanism for navigating around the genome, and allow individual users to customize the type of data displayed. These tools are useful for a wide variety of tasks including positional cloning, identifying mutants in a gene of interest, finding cDNA and ESTs for a gene of interest, and finding and displaying the distribution of sequence features (e.g., polymorphisms, T-DNA insertions) in a whole-genome context. While all of these tools share some functionality, each tool has its own specific set of features. JBrowse is highly customizable, contains many data types not represented in SeqViewer or GBrowse and is more frequently refreshed with new data. SeqViewer and GBrowse are legacy software that TAIR continues to maintain.

The following protocol highlights some of the TAIR specific features of our JBrowse instance. For a more exhaustive guide to JBrowse and its features see the JBrowse User Guide (https://jbrowse.org/jbrowse1.html). As of June 2022, the current JBrowse software version is 1.16.6. Gene function data in JBrowse is updated on a quarterly basis (i.e., gene names/symbols and descriptions) and community tracks are updated as needed.

Necessary Resources

See Basic Protocol 1

Exploring JBrowse

Viewing Arabidopsis genomes and genes in JBrowse

1. Go to the TAIR home page (*http://www.arabidopsis.org*). In the Tools section of the

menu bar, click on the link to JBrowse. Alternatively, go directly to the URL https://jbrowse.arabidopsis.org/index.html?data=Araport11

2. The JBrowse display is organized into two main panels. The left side of the browser window contains the Track Selector (Fig 11.1.3 A) and the right side displays the genome browser along with controls including Genome, Track, View and Help and search functions (Fig 11.1.3 B).
3. The Genome selector (Fig 11.1.3 B item a) is used to switch between different genome versions. To change the genome, click on the named version in the drop-down menu to choose from among the preloaded options. Alternatively, one can choose to upload a custom sequence file. Once selected, the display will refresh to the new genome. This will rewrite both the browser view and available tracks displays.

   The default genome in TAIR is the Arabidopsis thaliana Col-0 genome originally sequenced in 1999. Since then, the genome has undergone 10 subsequent re-annotations up to Araport 11(Cheng et al, 2017) and the assembly (actual sequence of nucleotides along the chromosome) has been updated to the current TAIR10 version. As new Arabidopsis genomes have been sequenced, they can be added to JBrowse such as the ColCEN genome (Naish, et al., 2021).

   Altering the Genome changes the chromosome sequence, gene models, and other available tracks to those of the specified release. Note that gene models or other features present in different releases may potentially be absent, located at a different position, or otherwise altered relative to an earlier release. A description of the latest genome release can be found at http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/genome_annotation.jsp), details of earlier releases can be found on the TAIR site (http://www.arabidopsis.org/download/index-auto.jsp?dir=/download_files/Genes)

3. The names and position of genomic features such as genes or genetic markers can be entered in the search box (Fig. 1.11.3. B, item e). For genes, either the AGI code (e.g., AT1G05460) or gene symbol (e.g., SDE3) is a valid search query. Nucleotide ranges can also be entered to allow specific regions of interest to be displayed. The chromosome and start and end coordinate of the desired region must be entered in the following format Chr1:1504365..1514364.

   If a query returns multiple hits, JBrowse will display these as distinct rows with the position of each feature shown. Clicking the desired hyperlink will open the detail display for the selected region.

4. Entering a feature name or region and clicking Go (Fig 1.11.3 B item e) will update the overview and details display. The overview map shows the position of the region displayed in the detail view relative to the rest of the chromosome. The size of this region is shown in the Scroll/Zoom drop-down. The default display includes the following tracks: reference genome, protein coding loci, and T-DNA seq data.

   To highlight an area of interest in the genome, click on the highlighter icon to the left of the Go (search) button (Fig 1.11.3.X B item f). When active, the button will be yellow. With the mouse in the nucleotide numbering track just below the

controls, click on the start point of the region all the way to the end point. This highlighting provides a convenient way of keeping the feature of interest in view when expanding the display, a larger region (Fig 11.1.4 A).

11 The zoom feature (Fig 1.11. 3.B item d) can be used to adjust the viewing dimensions in order to display a larger-scale view of the genome all the way down to the nucleotide sequence level. As with the highlighter, click any start point of the nucleotide numbering track (a vertical red line will appear) and then drag the mouse to the desired end point. When the mouse is released the genome view will be redrawn to the new start and end points.

6. To move along the chromosome, click on the grey arrows (Fig 1.11.3 item b) to shift the display to the left or right. Moving the cursor over a feature and left clicking the feature glyph will bring up a pop-up window that displays additional information about that feature. For genes, this includes known symbols or common names (Fig 1.11.4 A item a). Right clicking on a feature will open a pop-up window with additional data display options. For example, clicking on a protein coding gene model will give options to view the gene info in JBrowse, view the TAIR locus page for the corresponding locus or view the sequence using SeqLighter (see below.)

Viewing pre-loaded community generated data tracks

The following section describes features and data types available when the Araport 11 Col-0 reference genome is selected

7. The Tracks menu (Fig 1.11.3 A and Fig 11.1. 4 B) allows you to customize the display of preloaded tracks within the JBrowse genome view. TAIR JBrowse includes the following major track categories: **Arabidopsis Genome Assemblies, Community Data and Whole Genome Alignments.** Within each track category there are subsections that can be expanded to display different types of data (Fig. 1.11.3 A). Clicking on the down arrow in the section header will expand to display the full list of available tracks in each category.
8. To add or remove tracks from the detail display simply check or uncheck the required tracks. The track order can be adjusted by clicking the track title in the details panel and dragging the track up or down to a new position.
9. Configure individual tracks by clicking the down arrow next to the track title. This allows the user to choose the shape and color of the glyphs, put a limit on the number of features displayed in any one region, and set preferences if a text label is displayed.

TAIR welcomes community data submissions. Community members who wish to provide data to TAIR for JBrowse should send an email to curator@arabidopsis.org and provide the data in the specified format (typically GFF or BED formats). Please specify if the data is for pre-publication review so that it can be displayed on a private server for peer review.

Adding tracks from CoGE or local sources

10. In addition to the preloaded tracks, JBrowse enables you to upload your own

annotation to visualize within the context of the *Arabidopsis* genome. To upload your own data to JBrowse go to the "Tracks" option in the Genome panel (Fig 1.11.3 B item g). Clicking on the Tracks option will open a pop-up window where you can select the custom data.

11. Users can upload any combination of data files and URLs. To upload a local file either drop in the file or use the browser to locate and select the file on your computer. To upload from a remote site, enter the complete URL address for the file. Choose the option to either open immediately or to add to tracks for future viewing. Once the file is uploaded, JBrowse will use the data to suggest the type of track to display. Note that custom tracks are only visible to the user in the current JBrowse session.

12 Users can also use the CoGE plug-in (Fig 1.11.4 B item a) to search and display publicly available tracks from CoGE (REF: https://genomevolution.org/CoGe/; Lyons and Freeling, 2008). Clicking on the Epic CoGE link will display a pop up showing all of the available community tracks. Check all the tracks you wish to include.

Other specialized tracks

13 In addition to adding data tracks from the available preloaded track menu and custom datasets, JBrowse can be configured to include some specialized track functions. These options are available from the 'Track' link.

a. 'Add combination track' allows you to create a new track that displays merged data from two tracks. Click add combination track and then drag the two tracks you wish to display into the new combination track. This creates a new track (Fig 1.11.5 A) that merges the features of both tracks.

b. Add sequence search track allows you to include a DNA or amino acid search function. Click add sequence search track, choose the parameters (e.g., DNA or AA, case sensitive or not, etc…) enter the query term and hit search. This feature is very useful for identifying sequence features within a whole genome context.

Obtaining decorated sequence files

JBrowse includes a module called SeqLighter that allows you to extract a protein coding region and highlight specific features of interest.

14 In the genome display right click on the sequence glyph of interest in the 'protein coding gene models' track and choose the 'View Sequence' option.

15 Choose the add flanking region to add 500 to 4000bp upstream and downstream sequences.

16. The SeqLighter function will display the corresponding sequence in default CODATA format (Fig 1.11.5.B). Use the option selectors to choose which features to annotate such as introns and exons and UTRs. The default format can be changed to FASTA, PRIDE, or RAW formats and the image can be saved as JPEG, SVG and PNG files.

BASIC PROTOCOL 4

USING THE GENE ONTOLOGY ANNOTATIONS FOR GENE DISCOVERY AND GENE FUNCTION ANALYSIS

Annotations (associations of controlled vocabularies or keywords to data objects) provide a richer, more complex picture of a gene that is also more computationally accessible for the purpose of querying, classification, and making correlations among seemingly unrelated data. TAIR makes extensive use of controlled vocabularies for describing data in the database. The controlled vocabularies (ontologies) that are used by TAIR are also used by other model organism databases, thereby facilitating cross-species comparisons. All of the ontologies used by TAIR are included in the Open Biological Ontologies Project (*http://www.obofoundry.org/*) where they are freely accessible.

TAIR is member of the Gene Ontology (GO) Consortium (*http://www.geneontology.org*) and participates by developing and refining the ontologies and annotating Arabidopsis gene products (The Gene Ontology Consortium, 2010). The GO controlled vocabularies describe three aspects of gene products: molecular function, biological process, and subcellular location. TAIR also imports manual and computational annotations for Arabidopsis made by other groups including UniProt, BioGrid, JCVI (formerly TIGR) and others (Wortman et al., 2003; Berardini et al., 2004). These annotations are contributed independently by each organization to the GO database, where they are accessible through the AmiGO query tool for making cross-species queries (http://amigo.geneontology.org/amigo). The other main ontology used at TAIR is the Plant Ontologies developed by the Plant Ontology Consortium (POC; *http://www.plantontology.org*). The POC has used the GO model to develop controlled vocabularies for plant structures and developmental stages (Jaiswal, et al, 2005). In TAIR, both of these ontologies are used to annotate many additional types of data such as high throughput proteomics data, low throughput gene expression data, phenotypes, and publications. TAIR also collects and displays annotations contributed by members of the community who can use a simple web tool (Generic Online Annotation Tool, GOAT) to provide GO and PO annotations for genes based on published works (see *BASIC PROTOCOL 7*).

Necessary Resources

See Basic Protocol 1

Files

WRKYFamily.txt
(available at
https://www.arabidopsis.org/download_files/Help_Documents/WRKYFamily.txt)

Using the Keyword Browser to find candidate genes

For researchers, finding candidate genes involved in a particular pathway typically involves a fishing expedition using a variety of genetic, molecular, and biochemical assays. The GO annotations can be useful in making educated guesses about what genes may act in a pathway or are members of transcriptional/signaling cascades. Because TAIR and its community contributors have focused on GO curation from the literature, Arabidopsis is the most well annotated plant genome, with a large number of experiment-based annotations (Berardini). Thus, Arabidopsis GO annotations can be particularly useful for being able to infer gene function for unknown genes in other plant species based on sequence similarity or evolutionary relatedness. Another common use of GO annotations is to identify sets of genes associated with a given function or process in Arabidopsis as a starting point to identify genes with similar functions in other species.

1. Go to the TAIR home page (*http://www.arabidopsis.org*), click Search in the toolbar (Fig. 1.11.1), and select Keywords from the drop-down menu that appears. The page shown in Figure 1.11.6A is returned (TAIR Keyword Search and Browse; can be directly accessed at *http://www.arabidopsis.org/servlets/Search?action=new_search&type=keyword*). Enter term (keyword) "root development" in the text box and choose "contains" (an inexact search) from the drop-down menu to the left of the text box. From the group of check boxes for restricting the search, choose GO Biological Process as the keyword type and click the Submit Query button.

   Many of the terms in GO exist as complex phrases. TAIR searches take the entire entered term or phrase as a complete phrase rather than a set of words. Consequently, an "exact match" search will often not retrieve any entries. Therefore, the authors recommend using the "contains" option for keyword searches.

2. On the Keyword Search Results page (Fig. 1.11.6A), each controlled vocabulary term is displayed along with a count of all data objects (e.g., loci, publications, annotations) annotated to that term. Click "loci" to display the genes annotated to "root development." The results are displayed as a Gene Search Result page (see *BASIC PROTOCOL 2*) where all of the genes associated to the term 'root development" or its children, are displayed. Click on the locus name to view the locus details or save the list as a text file (see *BASIC PROTOCOL 2*)

Finding genes annotated to related functions

3. On the Keyword Search Results page, find the listing for "root development," and click on the "treeview" link. This will open a window displaying the term in a hierarchical tree view (Fig. 1.11.6 B).

   In the Gene Ontology, terms have a parent-child relationship to one another. Parent terms are less specific than their child terms. A child term may be a part of the parent (as thylakoid is part of chloroplast) or a type of the parent (as chloroplast is a type of plastid). In contrast to simple hierarchies, a child term may have more than one parent. The ontologies are intended to be as biologically accurate as possible. Terms and their relationships are defined by what is known about the biology of the process, function, or cellular component. By examining the structure of the ontology to find related terms, related gene products can also be found via their annotations to the terms.

4. Click on the plus sign next to the parent term ("root development") to expand the node and display all of the child terms.

5. To display genes annotated to each of the parent and child terms, select the "loci" radio button from the top of the tree view page (Fig. 1.11.6 B), then click the Display button. The display will be redrawn to show a count of the number of loci annotated to each term and the number of loci annotated to the children of each term. Click on the link to list loci annotated to the term "regulation of root development" to find all loci that are annotated to this term.

Retrieving GO annotations for sets of genes

GO Annotations can also be used to rapidly classify sets of genes such as gene families or co-clustered genes revealed by analysis of high throughput expression data.

6. Go to the TAIR home page (*http://www.arabidopsis.org*), click Search in the toolbar (Fig. 1.11.1), and select Gene Ontology Annotations from the drop-down menu that appears. Alternatively, go to the URL *http://www.arabidopsis.org/tools/bulk/go/index.jsp*.

7. Upload a list of AGI locus identifiers using the sample data file WRKYFamily.txt. This file contains a list of 74 loci all belonging to the WRKY transcription factor family (Eulgem et al., 2000; http://www.arabidopsis.org/browse/genefamily/WRKY-Som.jsp). Select the Text radio button under "Select output type"; to view results locally in a table format. Click on the "Get all GO Annotations" button. The output file contains a list of all the specified loci and their annotations to all three aspects of the GO ontology.

   The annotations include the evidence code and reference for the data supporting the annotation. The file can be saved onto a local computer as a tab-delimited text file. If the HTML option is chosen, the results are hyperlinked to TAIR detail pages for loci, keywords, and publications. The Web output also has links to the corresponding keyword entry in the GO database, where one can find annotations to genes from other organisms.

Classifying sets of genes into functional categories

8. Alternatively, instead of getting a list of all annotations, the genes can be grouped into broader categories based on their annotations. After uploading the gene list (step 7 above), choose "HTML output" and click the Functional Categorization button.

   For each aspect of the GO ontologies, a subset of terms have been selected to represent major broad categories, called GO Slim categories. If a gene is annotated to a child term of one of the GO Slim terms, it is included in the broader category. The GO Slim is less specific, but presents a simpler classification. The results include gene annotations that are both experimentally supported and computationally predicted. To find sets of annotated genes based on evidence codes, use the Evidence sub option in the Search by Associated Keyword section on the Gene Search page (http://www.arabidopsis.org/servlets/Search?action=new_search&type=gene). GO Slim assignments are also included in the detailed GO annotation output

(from step 7). See http://arabidopsis.org/help/helppages/go_slim_help.jsp for a list of all GO Slim terms and their definitions.

9. The database will return a functional categorization list showing all categories represented in the genes from the input file, along with the frequency of distribution of the genes within the set (Fig. 1.11.7A). To view a list of genes in each category, click on the number in the "Gene count" column.

   Only the categories represented by the genes in the list are included; the absence of any of the GO Slim categories means that there are no genes in the list that fall into that particular group. The default option displays the list grouped by keyword type and then by categories sorted by the number of annotations in each category. The table can be re-sorted to list by gene count. Frequency refers to the number of occurrences of a gene-keyword pair in the list. Multiple annotations to the same term are essentially compressed in this view, in contrast to the Get all GO annotations option. Genes that are annotated to multiple terms that fall into different categories will be included in each of the GO Slim bins. Therefore, the total number of annotations to each aspect of the GO ontologies may be greater than the total number of genes in the query list.

Displaying the functional classification as a chart

10. The distribution of functional categories can be displayed graphically as either an annotation pie chart or gene bar chart. To display as a pie chart, above the Functional Category column (Fig. 1.11.7A), select 'Annotation Pie Chart' and click on the 'Draw' button." This will create a new page showing three separate pie charts, one for each aspect of the Gene Ontology (Fig. 1.11.7 B). Depending on how the results are sorted, the sections can be displayed from most to least frequent category, or by related categories. The percentage of the total is shown in the color key for each graph.

11. To save the graph images, hold down the Ctrl key while clicking on the image, or right click the mouse if using a PC, and save the image to the clipboard or to a file. The images are in Graphic Interchange Format (GIF), which can be opened using a variety of graphics software.

Downloading the entire set of Arabidopsis GO annotations

In some cases, it may be useful to download the set of Arabidopsis GO annotations for the entire genome. For example, a common use of TAIR's curated annotations is as a reference for annotation of other species using sequence similarity or homology based methods. In such cases it may be useful to import the Arabidopsis annotations into an analysis tool.

12. On the home page (Fig 1.11.1) go to the Downloads section of the main toolbar, choose Downloads and then GO and PO Annotations. Alternatively go directly to http://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGO_and_PO_Annotations%2FGene_Ontology_Annotations.

13. Navigate to the file named gene_association.tair.gz. This compressed file contains all of the GO annotations for Arabidopsis genes annotated by TAIR, community members, UniProt, the GO Consortium, IntAct, TIGR, and others, in

the standard GAF2.2 format (http://geneontology.org/docs/go-annotation-file-gaf-format-2.2/). The file is updated on a quarterly basis.

14. Another option is to use the ATH_GO_SLIM.txt file. This text document is a tab-delimited file that contains all the annotations to the narrow (granular) GO term as well as a column that maps the annotations to the corresponding GO Slim category. Users should consult the README file (http://www.arabidopsis.org/download_files/GO_and_PO_Annotations/Gene_Ontology_Annotations/ATH_GO.README.txt) for information on each of the data fields.

GO Term Enrichment/ Statistical over-underrepresentation test.

In addition to GO functional categorization, for any given set of genes users may also wish to determine if there are terms that are over or underrepresented in that set as a means to generate hypotheses about gene function or biological events. TAIR uses a web service, provided by PANTHER DB to facilitate GO Term statistical enrichment tests for Arabidopsis and other plants represented in the PANTHER database (http://www.arabidopsis.org/tools/go_term_enrichment.jsp; Mi, et al., 2013). Users can enter a list of locus identifiers, choose the appropriate species, and select the GO aspect (biological process, cellular component or molecular function). PANTHER's tool accesses a comprehensive list of GO annotations from the GO Consortium as well as a recent version of the ontology itself, both of which are updated monthly. Because annotations are constantly being updated as new information is obtained, the monthly updating schedule ensures that analyses done using the PANTHER tool rely on the most current annotation data.

15. Go to the TAIR home page (*http://www.arabidopsis.org*), click Tools in the upper menu bar (Fig. 1.11.1), and select GO Term Enrichment from the drop-down menu that appears. Alternatively, go to the URL *http://www.arabidopsis.org/tools/go_term_enrichment.jsp*
16. Enter in a list of gene identifiers such as AGI Locus IDs (e.g., AT5G61160), UniProt IDs (e.g., Q9FNP9) or NCBI Entrez GeneIDs (e.g., Gene: 836237), separated by newlines or commas.
17. Choose the appropriate plant species from the drop down menu.

The web service implemented at TAIR can be used to analyze Arabidopsis as well as any of the other plant species included in the PANTHER database.

18. Select the ontology aspect that you wish to analyze. The options are 'biological process', 'molecular function', and 'cellular component.'
19. Click Submit, to send the data to PANTHER.

.

Evaluating the results

The web service sends the data to PANTHER and the results are returned in a new window on the PANTHER website (Figure 1.11.8).

20. The analysis summary box (Figure 1.11.8 A) displays the analysis type (PANTHER can do several types of gene list analysis), annotation version and

annotation dataset. This information is important to record and report in your publications, as the same analysis performed with different software versions and different annotation releases may yield different results.

21. ID mapping results. Uploaded IDs are mapped to the reference proteome set in PANTHER. Click on the number to review each list to see the details.
    a. Unmapped IDs are those that could not be mapped to a corresponding UniProt reference genome protein record in the PANTHER. This list would include any non-protein coding loci.
    b. Multiple IDs. PANTHER also provides a list of IDs where multiple IDs are mapped to the same PANTHER protein entry. Typically, this occurs when more than one gene produces the same amino acid sequence.
22. Term Enrichment Results (Figure 1.11.8 B). The results are displayed in a table.
    a. Term list. The first column displays the over/underrepresented GO terms. By default, only results with a p value of less than 0.05 are displayed. The terms are presented in a hierarchical format where related terms are grouped by background color, with the most granular term at the top. Invert the sort order by clicking the arrow next to the term 'Hierarchy' in the last column header. To view it as a simple list, click 'Hierarchy'.
    b. The second column shows the number of genes (#) in the reference genome dataset that map to the terms. This is the background frequency.
    c. The third column shows the number of genes (#) in the sample gene set that map to the GO term. This is the sample frequency.
    d. The fourth column displays the number of genes mapped to the term that would be **expected** based on the whole genome representation. For example, if 113/27,352 genes in the reference set mapped to cytosolic large ribosomal subunit, then the expected frequency (0.0041) to map to that term in the sample set (0.0041 X 247=1.02). Clicking on the number will retrieve a list of the genes that map to the term.
    e. The fifth and sixth columns show the fold enrichment and a sign to show increase (+) or decrease (-). Fold change is calculated by dividing the observed by expected results.
    f. The seventh column shows the p-value. The lower the p-value, the less likely the obtained result can be explained by random distribution.

BASIC PROTOCOL 5

USING GENE LISTS TO DOWNLOAD BULK DATASETS

TAIR provides a number of tools for obtaining data in bulk for sets of genes such as gene descriptions or sequences (http://www.arabidopsis.org/tools/bulk/index.jsp). While the gene search and locus pages can provide comprehensive information on a gene by gene basis (see *BASIC PROTOCOL 2*), it is often desirable to obtain specific data for a large number of genes. TAIR's bulk download tools can be used to take a set of AGI locus identifiers as an input and obtain gene descriptions (http://www.arabidopsis.org/tools/bulk/genes/index.jsp), GO annotations (see *BASIC PROTOCOL 4*) and PO annotations (http://www.arabidopsis.org/tools/bulk/po/index.jsp), sequences (http://www.arabidopsis.org/tools/bulk/sequences/index.jsp), protein properties (http://www.arabidopsis.org/tools/bulk/protein/index.jsp) , microarray elements (http://www.arabidopsis.org/tools/bulk/microarray/index.jsp) and locus histories (http://www.arabidopsis.org/tools/bulk/locushistory/index.jsp).

Necessary Resources

See Basic Protocol 1

Downloading Gene Description/Summaries

1. On the TAIR home page (*http://www.arabidopsis.org*) select Bulk Downloads from the Tools drop-down menu. Alternatively, go directly to the URL *http://www.arabidopsis.org/tools/bulk/index.jsp*.
2. Choose Gene Descriptions (http://www.arabidopsis.org/tools/bulk/genes/index.jsp).
3. Enter in or upload a list of AGI locus identifiers or gene model identifiers.
4. Choose which data set to search against to retrieve matching records. To obtain all descriptions for a locus, choose 'get descriptions for all gene models/splice forms.
5. Choose how you want your results returned, to the browser or in a file.

Guidelines for understanding the results

6. The results will include the locus identifier, gene model name(s), description, primary gene symbol and other gene symbols.
   Each locus may be associated to one or more gene models, and each model may have distinct descriptive information that is unique for that gene product. For example, the locus AT2G42810 (http://www.arabidopsis.org/servlets/TairObject?id=33349&type=locus), encoding Protein Phosphatase5 (PP5) has a total of 5 gene models which represent different splice variants. The AT2G42810.2, or reference gene model, is an integral membrane protein whereas AT2G42810. 1 does not contain the membrane domains and is

localized to the cytoplasm.

The gene description will either be a short, computationally derived description statement attributed to Araport 11, or a curated summary written by TAIR curators.

Downloading whole genome annotations in bulk

In some cases, it may be desirable to obtain data sets for the entire Arabidopsis genome such as sequences or functional annotations. TAIR provides access to curated data (e.g., PO annotations, phenotypes, gene summaries, gene aliases, etc.) after the data have been in TAIR for one year. Year old data is released on a quarterly basis and can be found in the **Download** section (see Introduction) under Public Data Releases. Subscribers can access more recent data sets from the **Download** section (Subscriber Data Releases). Sequences from TAIR10 and Araport11 are available as BLAST data files (http://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FSequences*)*.

If you are unsure of where to find a dataset, if the dataset is the most current or for custom datasets, contact the TAIR curators (curator@arabidopsis.org).

BASIC PROTOCOL 6

USING TAIR'S ANALYSIS TOOLS TO FIND SHORT SEQUENCES AND MOTIFS

Using the Motif Analysis Tool for Identifying potential cis-regulatory motifs in upstream sequences

The Motif Finder identifies six-oligomer nucleotide sequences that are statistically over-represented in a set of input sequences when compared to the whole genome. The most common application of this tool is for identifying potential *cis*-regulatory elements in genes whose expression patterns correlate into a cluster. Consensus sequences for putative transcription factor binding sites can be used to identify additional genes having the element in the promoter using the Patmatch program

Necessary Resources

See Basic Protocol 1

Entering the search parameters

1. On the TAIR home page (*http://www.arabidopsis.org*) select Motif Analysis from the Tools drop-down menu. Alternatively, go directly to the URL *http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp*.

2. On the data entry form enter the locus identifiers of your genes of interest. This can be done manually or by uploading a list of Arabidopsis AGI identifiers from a file. In the example shown in Figure 1.11.12A, we queried for motifs in the 500 bp upstream region of 15 co-expressed genes. Note that a minimum number of 3 locus identifiers has to be entered.

3. Select length (500, 1000, or 3000 bp) of upstream sequence to be queried. Submit the query.

   The sequence data sets are either 500-, 1000-, or 3000-base-pair sequences upstream of the translation start site of each gene in the genome. The program will search for 6-mer words that are overrepresented in the upstream regions of the set of queried genes compared to upstream sequences in the entire genome. Both forward and reverse strands are queried.

Evaluating the results

4. The results are displayed in a table as shown in Figure 1.11.9 B. The columns, denoted "a" through "g" in Figure 1.11.9 B, are as follows.

   a. Oligomer.

   Each over-represented six-oligomer sequence is listed in the first column of the results table.

   b. Absolute number of oligos in the query set.

   Number of times the oligo appears in the upstream regions (of chosen length) of the query genes. This number can be higher than the number of query sequences, as some sequences contain multiple occurrences of the motif.

   c. Absolute number of oligos in the genomic set.

   Number of times the oligo appears in the upstream sequences (of chosen length) of all genes in the genome.

   d. Number of sequences in query set containing oligomer.

   Shows the ratio of the number of queried sequences containing the oligomer over the total number of queried sequences.

   e. Number of sequences (e.g., out of 34,187 in genomic set) containing oligomer.

   Shows the ratio of the number of genome sequences containing the oligomer over the total number of sequences in the genome.

   f. *p*-value.

   This score reflects the probability of the six-oligomer sequence occurring in the selected query set by chance. The lower the score (closer to zero) the greater the likelihood the match is significant.

   g. Query sequences containing this oligomer.

   All of the query genes containing the oligomer are listed here. The Patmatch tool (see next section) can be used to locate other genes that contain the oligomer in the upstream sequence.

Using the Patmatch Tool to find short sequence patterns in DNA and protein sequences

Patmatch (Yan et al., 2005) was designed for identifying patterns in a selected TAIR dataset (e.g., genes, proteins, upstream sequences, etc.) that match regular expressions. Patmatch can be useful for finding short nucleotide patterns such as *cis*-elements, Massively Parallel Signature Sequence (MPSS), Serial Analysis of Gene Expression (SAGE) tags, or small RNA binding sites. Patmatch can also be used to search for motifs in protein sequences. Other options of this tool include the selection of a target data set, strand to be queried (in case of nucleotide search), and number of results to retrieve. If one needs to process large amounts of data or increase the number of results to be included, users can download the Patmatch1.1 program (*http://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FSoftware%2FPatmatch/*) and run it locally on a Unix-based system. The BLAST data sets used by Patmatch can also be downloaded from the TAIR site (*http://www.arabidopsis.org/download/index-auto.jsp?dir=/download_files/Sequences)*.

Necessary Resources

See Basic Protocol 1

Entering the search Parameters

1. From the TAIR home page (www.arabidopsis.org) select Patmatch from the Tools drop down menu. Alternatively go directly to https://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl.
2. Enter in a query pattern and the appropriate option for a DNA or Protein search from the drop down menu. Acceptable inputs include regular expressions that include mismatches, insertions, and deletions, and apply standard IUPAC notation to indicate ambiguous sequences Supported syntax formats are displayed on the bottom of the data entry page.
3. Choose a sequence dataset to search. The program uses the same target datasets as TAIR's BLAST software.

Evaluating the results

4. Patmatch does not generate alignments or provide scores for best hits. The results are displayed in a table format that includes the following information.
    a. *Sequence name:* Name of the gene or sequence for which a hit was found.
    b. *# of hits:* Number of times the query pattern was found in that specific sequence.
    c. *Hit pattern:* Pattern used for the query.
    d. *Matching positions:* Start and end position of the hit. These coordinates are always relative to the sequence (e.g., gene, upstream region, intergenic region).
    e. *Hit sequence:* Hyperlink to the sequence for which a hit was found. The pattern match is highlighted in red letters. For nucleotide searches, coordinates shown here are always relative to the chromosome.

BASIC PROTOCOL 7

USING THE TAIR GENERIC ONLINE ANNOTATION TOOL (GOAT) TO SUBMIT FUNCTIONAL ANNOTATIONS FOR ARABIDOPSIS (OR ANY OTHER SPECIES) GENES

In order to maximize the capture of experimental information about gene function from the literature and from our expert community, TAIR has developed tools to enable researchers to curate functional annotations and make those annotations visible in TAIR. In 2021 we retired the TAIR Online Annotation Tool (TOAST, Berardini, et al, 2012) and replaced it with an easier to use Generic Online Annotation Tool (GOAT, https://goat.phoenixbioinformatics.org). Like TOAST, GOAT also enables users to submit their own GO and PO annotations, and comments. The software is more intuitive to use, allows saving in progress submissions, most importantly it allows users to annotate not only Arabidopsis genes, but genes from any species as long as they have a UniProt ID or RNA Central ID. Authentication is via the users ORCiD ID and registration at TAIR is no longer required for submission. Submitters must provide a DOI or PubMed ID. Alternatively users can download a preformatted Excel spreadsheet and email annotations to TAIR at curator@arabidopsis.org.

Necessary Resources

See Basic Protocol 1

Submitting Annotations

1. On the TAIR home page (*http://www.arabidopsis.org*) go to the section marked Submit and choose Online Submission for Authors and Others. Alternatively go to http://www.arabidopsis.org/doc/submit/functional_annotation/123.
2. Scroll to the center of the page and click the button to “Fill Online Form”. Alternatively go to https:/goat.phoenixbioinformatics.org/.

   GOAT uses ORCiD for user authentication as a way of crediting community submission, therefore before creating a submission users should register with ORCID (https://orcid.org/). Users should include a public email address in the ORCiD profile if they wish to receive automated notifications about their submission.
3. Select login from the upper right menu on the home screen follow instructions to authenticate with ORCiD.
4. On the GOAT home page click the ‘Submission” link in the upper left menu.
5. On the resulting data entry form (Fig 1.11.10) enter the PubMed ID or DOI for the article to annotate (Fig 1.11.10 item a).

   TAIR only displays annotations from peer-reviewed, published works. Users may submit annotations for articles that have been accepted for publication that have received a temporary DOI via a preformatted spreadsheet, but TAIR does not accept annotations for unpublished work.

6. Enter the identifier of the first gene to annotate. Acceptable IDs for Arabidopsis genes include AGI Locus identifiers (e.g., ATNGNNNNN.), UniProt ID (e.g.) and RNA Central ID. To add additional genes, click the "Add Another Gene" button (Fig. 1.11.10 item b).
7. Enter the annotations. The form contains separate sections for each type of annotation (GO Molecular Function, GO Process, Expression, Protein Interaction, Comment). At least one annotation must be entered in order to be able to submit (otherwise the Submit button remains grayed out).
    a. Enter the term in the left column. The auto-suggest function will offer a list of suggested terms. Choose one of the suggested terms, or if none is appropriate, enter a new term. *A TAIR curator reviews all the contributions and will approve the annotation, update to find the best term that matches, or determine if a new ontology term needs to be added.*
    b. Enter the supporting evidence in the right column. All annotations must be backed up by evidence. Choose the evidence type from the dropdown menu that most closely fits the experimental method.
8. To add additional annotations, click on the 'Add Another Annotation' button and a new data entry row will appear (Fig. 1.11.10 item d). The choose the annotation type from the drop down menu. To delete an annotation, click on the red X to the right of the annotation row (Fig. 1.11.10 item c).
9. Enter comments. At the bottom of the form there is a section to enter comments that may include information that cannot be captured in a GO or PO annotation. This section is optional.
10. Submit the annotations or add another gene. Once all of the annotations for the gene entered in step 5 are done, either submit the annotations or annotate another gene described in the same paper.
    a. To add another gene, go to the top of the form and click on the plus sign in the upper right corner (Figure 1.11.10 item b). This will append a new entry form to the bottom of the page.
    b. To submit annotations, click on the Submit Annotations button on the lower left side of the page (Figure 1.11.10 item e).
11. Once the data is submitted a curator will review the submission. If there are any questions, a curator will contact the submitter. It can take a week or two before the data is visible in TAIR.

Other ways to submit data/corrections to TAIR

One of the most fundamental aspects of science is sharing data and results with the research community. The fruits of research drive new areas of discovery, and funding agencies, such as the National Science Foundation (NSF), have invested heavily in developing community resources. Web sites and databases such as TAIR make these data accessible to anyone connected to the Internet. The long-term sustainability of databases will increasingly rely upon contributions by the research community (Reiser et al., 2016; Leonelli, et al, 2017).

TAIR encourages feedback and data submission and provides several ways for researchers to contribute their expertise and data. Instructions for submitting various types of data including gene function, interaction partners, expression patterns, markers, phenotypes, and several others, are available on the Submit Overview page (*http://arabidopsis.org/submit/index.jsp*), accessible from the Submit drop-down menu in the top navigation bar. Users can prepare data formatted according to the guidelines or download and use the preformatted Excel spreadsheets. The spreadsheets may contain macros that ensure that the proper data formats are used. To use the spreadsheets, macros must be enabled. TAIR will also accept direct submissions by email to *curator@arabidopsis.org* for small datasets and corrections to existing data, as well as very large datasets and those requiring special formats. Please contact us with any questions about data submission.

In addition, each data detail page includes a Community Comments section where community members can add additional information; click on the comment text to view the entire comment. Registered users can submit comments that are then immediately displayed in the Comments section of the detail page. On-line instructions for submitting comments are available at *http://arabidopsis.org/help/helppages/addcomment.jsp*.

BASIC PROTOCOL8

USING PHYLOGENES TO VISUALIZE GENE FAMILIES AND PREDICT FUNCTIONS

PhyloGenes (www.phylogenes.org) is a website for visualizing gene function data alongside phylogenetic trees (Zhang et al., 2019). It uses pre computed phylogenetic trees and multiple sequence alignments generated by the PANTHER project for the trees (Mi et al., 2021) and can be used to make phylogenetic based inferences about gene function for unknown members of the gene family. PHYLOGENES trees are updated annually when the PANTHER families are updated. PhyloGenes contains all plant species from PANTHER as well as 10 well annotated non plant reference species. The underlying concept of phylogenetic based inference is that gene functions that are shared by common ancestors can be attributed to their descendants. Visualizing functions alongside the tree can also indicate where evolutionary novelty arises in gene families. As described in *BASIC PROTOCOL 1*, the homology data from PhyloGenes/PANTHER is integrated into TAIR locus pages and loci in TAIR that are represented in PhyloGenes will display links to the corresponding PhyloGenes tree. The following protocol explains some of the basic features of PhyloGenes that can be performed after retrieving the PhyloGenes tree from the TAIR locus page links

Necessary Resources

See Basic Protocol 2

Viewing the PhyloGenes gene family page from a TAIR locus page

1. Clicking the link to the PhyloGenes tree viewer from a TAIR locus page (Fig 1.11.2 item h) will open a new web page on the PhyloGenes website that displays the corresponding gene family page (Figure 1.11.11 A) centered on the TAIR locus which is also highlighted. The display has three main sections.

a. Metadata section (Fig 1.11.11 A item a) contains the name of the PANTHER family along with a count of the total number of genes and the taxonomic range of the gene family (the last common ancestor).

b. Tree Panel (Fig 1.11.11). This section is where the phylogenetic tree is displayed. The default display is a compact view in which branches having genes with known functions are expanded and those without known functions are collapsed (grey triangles). The tree panel includes a search box for searching within the family (Fig 1.11.11 A item b) as well as controls for customizing the tree and downloading data (Fig 1.11.11 A item c).

c. The Data Panel (Fig 1.11.11 A). This section displays linked descriptive and gene function data. The descriptive data includes gene names, symbols, species, protein names and UniProt IDs (linked to UniProt records). Gene functions are experimentally determined and phylogenetically inferred Gene Ontology annotations to Molecular Functions or Biological Processes. The data panel can be customized to reduce or reorder columns (Fig 1.11.11 A item d).

Controlling and customizing the tree display

2. Users can collapse and expand individual nodes or expand the entire tree. Collapsed nodes are represented by a grey triangle. Clicking on a collapsed node will expand it to show all members. To expand all the nodes, click on the "Expand all" icon in the Operations menu.

   Nodes in PhyloGenes are color coded as described in the legend which can be toggled on or off using the down arrows in the operations menu (Fig 11.1.11 A item b). Speciation events are green, duplication events are orange, horizontal transfer events are light blue and subfamily nodes are represented by dark blue diamonds.

3. To reduce the complexity of the tree, users can choose to prune or remove species from the display. Clicking on the 'Tools' icon in the Operations menu will display a dropdown list of functions including 'Prune tree by organism'. Selecting that option will open a new pop up window showing all the species in the tree; by default, all species are checked (Fig 11.1.11 B). To remove a species, uncheck the box and then click the Update button. This will redraw the tree, the topology will remain the same, but the display will be less crowded. Another quick way to display a small subset is to check the box in the header to uncheck all and then check the boxes for the species you wish to display.

Controlling and customizing the data display

4. The linked gene function data is displayed in a tabular format to the right of the tree panel. Linked data are aligned with the corresponding protein in the gene tree. The data displayed include:

   a. Gene Symbol, The common name for the gene (e.g. PHOT2).

b. Gene ID. The unique identifier for that gene sequence in the reference genome (e.g. AT5G58140). For Arabidopsis, these gene IDS also link back to the corresponding TAIR locus pages.

c. Protein name. the full name for the protein (e.g. PHOTOTROPHIN-2)

d. UniProtID. The unique identifier for the corresponding protein entry in Uniprot. This is also a hyperlink to the UniProt entry (e.g.P93025)

e. PANTHER subfamily name. Subfamilies within each family are groups of genes that share a particularly high degree of similarity due to limited divergence from their common ancestor. Subfamilies are, in general, closely-related orthologs.

f. Gene Ontology Annotations are displayed for Biological Process and Molecular Function Ontologies. Phylogenetically based annotations are indicated with a green tree icon (Fig 11.1.11 A item e). Annotations that are experimentally determined are indicated by a yellow flask icon (Fig 11.1.11 A item f). Clicking on the icon will display detailed information and links to the supporting reference (Fig 11.1.11 A item g).

PhyloGenes incorporates annotations from Gene Ontology Consortium that are filtered based on evidence codes. Experimentally determined annotations include those with the following evidence codes IDA: inferred from direct assay; IMP: inferred from mutant phenotype; IGI: inferred from genetic interaction; IPI: physical interaction; IEP: expression pattern; EXP: experimental evidence. Phylogenetically inferred annotations have the evidence code IBA: inferred from biological ancestry. Annotations are retrieved from the Gene Ontology Consortium (www.geneontology.org) and updated on a quarterly basis.

5. The display of data in columns can be adjusted by clicking on the gear icon in the table header (Fig 11.1.11 A item d). The text next to the icon indicates if and how many data columns are hidden. Clicking on the icon opens a popup configuration window with the following functions (Fig 11.1.11 C).

    a. Show/hide columns by checking the box next to the name.

    b. Reorder columns using the up or down arrows next to the column name.

6. Users can also opt to toggle the display to show the multiple sequence alignment for the underlying PANTHER tree. Clicking on the text 'Show MSA>'will replace the data table with the alignment. The MSA display also includes a legend that explains the different fonts and color coding.

7. Users have several options for saving the tree data. Clicking on the Downloads icon in the Operations menu (Fig 11.1.11 B item c) displays the following download options.

    a. Download the multiple sequence alignment for the entire tree.

    b. Download a CSV file of orthologs of a given gene within the tree.

c. Download tree in PhyloXML format. The entire tree can be downloaded and saved locally in a standard data exchange format for use in other applications.

d. Save tree as SVG or PNG. These options allow you to save the tree image for use in other graphics applications or for publication.

e. Download gene table as CSV. Use this option to save all the data fields from the data panel in a single table.

Grafting a new sequence onto a gene tree

For users whose species are not represented in PhyloGenes, there is an option to graft single sequences onto the precomputed PhyloGenes trees. The TreeGrafter (Tang, 2019) tool will run HMM scoring and find the best matching gene family, if it exists, add the sequence to the MSA of that family, then run RAxML to insert the new protein sequence to the best location of the gene family tree. The inserted sequence will be labeled as 'grafted'.

8. From the PhyloGenes home page (www. phylogenes.org) click on the link 'Not seeing your species' above the list of included species (Fig 1.11.12 A) or go to http://www.PhyloGenes.org/grafting.

   The option to graft a sequence can also be accessed in the search results when no match is found (Fig 1.11.12 B).

9. Enter the raw amino acid sequence and click 'Graft'.

10. The software will return the matched tree with the new sequence included and labeled a 'Grafted'. (Fig 1.11.12 C).

    Occasionally the grafting program will produce unexpected results. If this happens, please submit a report to info@PhyloGenes.org.

BASIC PROTOCOL 9

USING TAIR TO BROWSE ARABIDOPSIS LITERATURE

TAIR provides a number of ways for researchers to keep abreast of the literature. In addition to the curated links between genes and articles that are displayed on the locus detail pages (see *BASIC PROTOCOL 2*), the entire corpus of publications in TAIR (including abstracts and conference proceedings) can be searched using the Publication Search (http://www.arabidopsis.org/servlets/Search?action=new_search&type=publication) or Keyword browser (http://www.arabidopsis.org/servlets/Search?action=new_search&type=keyword). For users wanting to keep up with the latest Arabidopsis research, TAIR developed an additional tool for browsing recently added literature.

Necessary Resources

See Basic Protocol 1

Browsing Recently Added Literature

1. On the TAIR home page (*http://www.arabidopsis.org*) go to Browse and select Recently Added Literature. Alternatively go to http://www.arabidopsis.org/servlets/Search?pageNum=1&type=publication&action=search&recent=14&size=500&sort=journal.
2. The page will display a list of the research articles downloaded from PubMed and entered into TAIR during the time period specified in the header. Typically this is a one week period. Each article is displayed in a separate band of alternating background color. The contents of each band include basic citation information and links to associated resources.

   The default display is sorted alphabetically by journal name. Choose author name from the drop down selector in the upper right corner to display the results in alphabetical order by the last name of the first author.

   a. Citation. The citation includes the authors, title, journal name, and publication year.

   b. Associated genes (may be empty). These are manually curated links to genes described in the paper. Clicking on the Gene name will display the corresponding locus detail page in TAIR where you can find more information about the locus (see *BASIC PROTOCOL 2*).

   c. Associated Keywords (may be empty). Keywords are generated by automatic text matching of GO terms to the text and are not curated. To find other objects in TAIR associated with that keyword, click on the term to display the keyword detail page (see *BASIC PROTOCOL 4*).

   d. Article views. Three options are provided that offer different views of the article.

      i. Click on the Journal link to read the article on the journal's website.

      ii. Click on PubMed link to view the corresponding record and abstract at NCBI's PubMed site.

      iii. Click on the TAIR link to view the publication detail page in TAIR. The TAIR publication page displays the citation, which may include the abstract, as well as associated keywords, loci and GO and PO annotations. The annotations and linked loci are manually curated.

BASIC PROTOCOL 10

USING THE SYNTENY VIEWER TO FIND AND DISPLAY SYNTENIC

REGIONS FROM ARABIDOPSIS AND OTHER PLANT SPECIES

Synteny Viewer (https://www.arabidopsis.org/cgi-bin/syntenyviewer2/showSynteny.pl) is a tool that displays precomputed syntenic regions between *Arabidopsis thaliana* and over 35 different plant genomes. The syntenic regions between *Arabidopsis thaliana* and the other genomes have been precomputed using the SynMap tool at genomevolution.org. (https://www.arabidopsis.org/help/helppages/syntenyViewHelp/SyntenyViewHelp.pdf). The syntenic regions are displayed using the GEvo tool from CoGE (Lyons and Freeling, 2008; www.genomevolution.org). Users can search for a specific Arabidopsis gene of interest by AGI locus ID or by a chromosome region and view syntenic regions from another selected genome. It can be used to help researchers study and analyze homologous genes and other conserved elements and sequences. It can also be used to study genome duplication and evolution. By comparing newly sequenced or less studied genomes to the well-annotated *Arabidopsis* genome scientists can identify novel genes and putative regulatory elements.

Necessary Resources

See Basic Protocol 1

Searching for syntelogs

1. On the TAIR home page (http://www.arabidopsis.org) go to Tools and select Synteny Viewer. Alternatively go to https://www.arabidopsis.org/cgi-bin/syntenyviewer2/showSynteny.pl
2. To search by name enter the AGI locus ID (e.g. AT1G01010) or Gene Model ID (e.g. AT1G01010.1) into the search box in section 1. Alternatively, search by location by entering the Arabidopsis chromosomal location using the format Chromosome:Start position..Stop position (e.g. 1:3760..5630).
3. Select the species in which to identify the syntenic region from the drop-down menu in section 2. Then click submit.
4. If a syntelog is found the results will include a table displaying a list of the syntelogs, and a link to the full GEvo display at CoGE (Fig 1.11.13). Below the table results, is an iFrame containing the same High Score Segment Pairs (HSPs) and analysis functions from GEvo site. For details on using the Synteny viewer options, consult the CoGE tutorials (https://genomevolution.org/wiki/index.php/Tutorials).

GUIDELINES FOR UNDERSTANDING RESULTS

General Considerations for Using TAIR

As with any Web-based resource, some general guidelines should be observed when interpreting results. Databases are constantly changing; new information is incorporated and interfaces can also change from the time of publication of this unit.

Revisions to the Data in the Database

Over the course of genome annotation, many new genes have been added and existing genes have been made obsolete or updated (split or merged) to reflect new information (Haas et al., 2003; Swarbreck et al., 2008, Lamesch, et al., 2012; Cheng, et al., 2017). In 2004, TAIR inherited the responsibility of maintaining the genome sequence and annotation from the former Institute for Genomic Research (TIGR), now the J. Craig Venter Institute (JCVI), which provided the genome sequence and annotation from 2000 to 2004. TAIR produced five genome releases culminating in TAIR 10 (Lamesch, et al., 2012). The Arabidopsis Information Portal (Araport), which subsequently took on the responsibility of genome annotation, released Araport 11 in 2016 (Chang et al., 2017). The naming convention agreed upon by the AGI for adding new loci and updating existing loci (*http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp*) is continued in TAIR and Araport releases. Users are encouraged to submit structural annotation updates to TAIR via email and to deposit the relevant supporting sequence data to GenBank (*http://www.arabidopsis.org/submit/gene_annotation.submission.jsp*) for incorporation into future reannotation efforts. Changes in sequence annotation may affect the association of genes to related data such as protein domains, polymorphisms, and homologies. For example, domains associated to a locus that was subsequently split may then be associated to only one of the two resulting loci. The locus history, shown on the bottom of the locus page (Fig. 1.11.2, p item), summarizes all of the structural updates that have been made to the locus such as merges, splits or obsoletions. The locus history can also be searched independently by locus name using the Locus History Search (*http://www.arabidopsis.org/tools/bulk/locushistory/index.jsp*). For many data sets in Downloads section of the website, TAIR maintains older versions of the data. Users should always note the date or version information associated with any data files, such as BLAST data sets or GO annotations.

Evidence Codes in GO Annotations

When interpreting Gene Ontology annotations, it is essential to understand the process of annotation and the importance of evidence codes in interpreting the annotations. The GO Consortium has developed a set of evidence codes (The Gene Ontology Consortium, 2010) as a way of quickly assessing the basis for the assertion made in the annotation. In TAIR, annotations include an evidence description, in addition to the evidence code (Berardini et al., 2004). The evidence description is a set of controlled vocabularies that describe the type of experimental or computational evidence in greater detail. For example, an annotation having the evidence code "inferred from mutant phenotype" (IMP) may be further elaborated by including more specific information about the type of experiment done such as "RNAi experiments." Since more than one gene may be affected by RNA interference, the phenotype may be due to changes in expression of multiple loci. Thus, the GO annotation should be viewed with the understanding that the phenotype may be due to the loss of function of more than one homologous locus. When no information is found in the available published literature, annotations are made to the root terms "biological_process," "molecular_function," or "cellular_component." Such "root" annotations indicate that at the time of annotation, no information for a more specific assignment was available for the associated gene. In contrast, a gene lacking annotations altogether might have available data but has not yet been curated. At TAIR, GO annotation is an ongoing process; annotations are updated as new information about genes is published (Berardini et al., 2004). Each annotation has an associated date, which refers to the date the annotation was made. Users

should carefully evaluate the source of any tools utilizing GO annotations (e.g. Term Enrichment) to ensure that the underlying annotations are current.

COMMENTARY

Background Information

TAIR was originally a collaborative project between biologists at the Carnegie Institution, Department of Plant Biology, and computer scientists at the National Center for Genome Resources, initiated in 1999. TAIR is the third incarnation of an *Arabidopsis* community database after AAtDB (An *Arabidopsis thaliana* Database, which continued from 1991 to 1994) and AtDB (*Arabidopsis thaliana* Database, which continued from 1994 to 1999; Flanders et al., 1998; Rhee et al., 1999). TAIR arose out of the need to accommodate genomic data such as the genome sequence, gene annotations, and integration of physical and genetic maps, in the context of the experimentally verified data in the literature. From its inception until early 2014, the National Science Foundation (NSF) funded TAIR. In late 2013, anticipating the end of NSF funding, four TAIR staff members founded the nonprofit organization Phoenix Bioinformatics (www.phoenixbioinformatics.org) and transitioned TAIR to a new, sustainable user fee model (Reiser, et al., 2016). The user fee structure was carefully crafted to distribute the costs equitably among the widespread and varied user community. With the support of the research community, TAIR continues to provide up to date, continuously curated data to its global users. Curated data is updated weekly. Data that have been in TAIR for one year are released on a quarterly basis and available for download and reuse (http://www.arabidopsis.org/download/index-auto.jsp?dir=/download_files/Public_Data_Releases) under a CC-BY license. Users without a subscription can access a limited number of page views per month. Unlimited access to all TAIR pages and quarterly releases of recent data requires a subscription. ABRC stock detail and ordering pages are available free of charge. The data release policy was designed to encourage and support data reuse, while still providing an incentive to subscribe. More information on how institutions or individuals can help support this nonprofit effort is available by clicking on the Subscribe link on the TAIR home page.

Design principles and current limitations

TAIR has been designed and built as a Web tool to allow researchers to access all of the data housed in TAIR using a standard Web browser such as Google Chrome or Mozilla Firefox. It is built upon industry standards for database management systems, software architecture, and software design (Weems et al., 2004). The current system of interfaces described herein is due for a significant overhaul which will modernize the interfaces but the primary functions will remain the same. TAIR is primarily designed by biologists and, although the interfaces were created with biologists in mind, it has not always been possible to arrive at solutions that meet every user's requirements. A certain amount of familiarity with *Arabidopsis* and with basic concepts of molecular genetics and plant biology is assumed. Consequently, the breadth of information on the home page and myriad options on the search interfaces can be daunting to a novice user. More experienced users and developers may be frustrated by the difficulty in obtaining the entire database for retrieving specialized, custom data sets. Users are encouraged to contact us via email (curator@arabidopsis.org) for assistance in using any of the tools or in accessing large or specialized datasets.

Keeping up to date with TAIR and Arabidopsis research

Users can stay connected with TAIR by following the *Arabidopsis* Information Resource on Facebook (https://www.facebook.com/tairnews), or receiving tair_news twitter feeds (http://twitter.com/tair_news) or YouTube channel alerts (https://www.youtube.com/user/TAIRinfo). TAIR News and Job Postings are relayed through the TAIR Twitter feed.

Ensuring your published Arabidopsis data visible in TAIR and other resources by making it Findable, Accessible, Interoperable and Reusable

For researchers to maximize the value of published data, that data needs to be made available in standardized, machine-readable formats that are easily discoverable in accordance with the Findable, Accessible, Interoperable and Reusable (FAIR) data principles (Wilkinson et al., 2016). TAIR curators translate experimental findings into computationally accessible formats such as Gene Ontology annotations to help make these data FAIR. TAIR offers some basic guidelines for researchers to follow when preparing their data for publication to ensure that their data is FAIR (https://conf.phoenixbioinformatics.org/pages/viewpage.action?pageId=22807345). When published data is FAIR it can be more readily accessed by members of the research community and by curators from TAIR or other genomic resources who may perform further curation to integrate or add value to those data (Reiser, et al., 2018).

Additional tools at TAIR

In addition to the tools discussed in the protocols, TAIR hosts several other useful analysis tools. Some of these are briefly described below.

Textpresso

Textpresso is an information extracting and processing package for biological literature(https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2103-8). Textpresso for *Arabidopsis* (*http://www.textpresso.org/arabidopsis*) allows users to search all abstracts and over 40,527 full-text Arabidopsis publications. Keyword searches can be narrowed by searching in specific categories. The individual matches are displayed showing each of the text snippets that match the query Textpresso was initially developed by Hans-Michael Muller, Eimear Kenny, and Paul W. Sternberg, with contributions from JuanCarlos Chan and David Chen. The most recent version, Textpresso 2.0, was developed by Hans-Michael Muller with contributions from Arun Rangarajan and Tracy K. Teal. The current version of Textpresso for Arabidopsis was updated in 2022.

Chromosome map tool

This tool (*http://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp*) allows the user to map genes on top of the five *Arabidopsis* chromosomes using a list of locus names (e.g., At1g01010). The list should contain one locus name per line. To display an alternate name, append the symbol after the locus identifier in the same row (e.g. AT1g01010 ANAC001). The resulting image, which displays the location of the queried list of genes on the five chromosomes, can be saved in a variety of formats.

Critical Parameters and Troubleshooting

No data found

A frequently reported problem is that searches do not retrieve any results. In some cases the data sought are not in the database, but in other cases the data are in TAIR but are not found because of problems arising from poorly formed queries or improper use of the search forms. The temptation to fill out all of the optional fields in the advanced searches can generate too many restrictions that limit the scope of the data retrieved. This can be overcome by using fewer, rather than more options. Another reason why searches fail is that the data are not accessible through the existing search interfaces. The categories under the Advanced Search section of the Web site (*http://arabidopsis.org/servlets/Search?type=general&action=new_search*) list data types that can be searched. To obtain data that are included in the TAIR database but are not easily accessible through any of the advanced searches, please send an e-mail to the curators to request the data.

Too much data found

While "no data found" is probably the most common problem encountered, retrieving too many results can also be a problem. There are two ways to handle this problem: (1) using the advanced searches and restricting parameters to retrieve a subset of the results, or (2) manipulating the results set to select a subset of data. Restricting the search parameters can be done on all the Advanced Search pages and detailed help on using these parameters is available (*http://arabidopsis.org/help/helpcontents.jsp*). Large results sets can be downloaded and reformatted to explore the data more efficiently. All of the search results can be downloaded as tab-delimited text files (see *BASIC PROTOCOL 2*). The results can be imported into software like Excel or Google Spreadsheets that allows manipulations such as sorting, reordering, reformatting columns, and graphing the results.

Layers of connected data that are hidden

TAIR's database structure exploits the relational database design and each data type has a high degree of association to other data types. This network of associated data is not easily represented in a two-dimensional, tabular format via hyperlinks. Consequently, associated data may be separated by one or more hyperlinks. For example, all gene models are associated to a given locus, but to view information for a specific gene model, such as a list of gene features (e.g. introns, exons, UTRs) and coordinates, it is necessary to click on the link to the individual gene model.

Reporting problems and requests to curators

Perhaps the most important thing to know about troubleshooting problems with TAIR is that users are encouraged to e-mail curators (*curator@arabidopsis.org*) to report problems, ask for help or request data. Users that want a particular set of customized data should contact the curators, who can then generate the requested file. Reporting problems may also lead to improvements to TAIR's data display or addition of new data or tools that benefit the whole community.

Advanced parameters

Despite the extensive content of this unit, it still does not cover all of the functionalities of the searches and tools that are offered at TAIR. Users familiar with the basic functionalities

and who are interested in more complicated queries or specialized views are encouraged to review the help documents or contact the curators.

Suggestions for Further Analysis

While there are no complete alternatives to TAIR, there are other web sites that provide a significant amount of *Arabidopsis* data and alternative ways to view, manipulate and analyze the data. All of these sites are linked extensively from TAIR, whereby the sites in the former category are linked from each locus detail pages and sites in both categories are listed and updated in the TAIR Portal pages (*http://www.arabidopsis.org/portals/index.jsp*). Many of these resources integrate TAIR curated functional annotations (e.g. gene summaries, names, literature) however, in accordance with TAIR's data release policy, the data on these sites will be at least one year out of date.

Arabidopsis genome annotation resources

There are other resources that offer views of the Arabidopsis genome that complement the TAIR resource. Users can search, download and analyze Arabidopsis genome data with ThaleMine, an InterMine instance created by the Araport project ((Krishnakumar, et al., 2015; Cheng et al, 2017) and now maintained by the Provart group (Pasha et al.,2020)). Another view of the Arabidopsis genome can be found as one of the databases in Ensembl Plants (http://plants.ensembl.org/Arabidopsis_thaliana/Info/Index). Ensembl provides its own genome browser for visualization, as well as plant gene families generated using Compara. Users can access variant data for Arabidopsis ecotypes generated by the 1001 Genomes project (http://1001genomes.org/) within Ensembl. Experienced users will find it useful to be able to generate their own custom datasets using Ensembl's BioMart or ThaleMine. SIGnAL (*http://signal.salk.edu/*) from the Salk Institute offers a genome viewer (T-DNA Express, *http://signal.salk.edu/cgi-bin/tdnaexpress*) that is decorated with all of the T-DNA and transcript data that are generated from Salk and other laboratories around the world. Often, SIGnAL displays data that are not yet displayed at TAIR; therefore, it is a good idea to check this site to get the latest mapping of T-DNA insertions and cDNA clones. Users should pay attention to the sources of data including gene functional annotations, genome annotation versions, assembly versions and the currency of the data. Because TAIR is updated on a weekly basis (mostly for functional annotations) the information in these resources may differ from what is shown in TAIR. There are also many sites that provide detailed information about a subset of genes of *Arabidopsis* such as chromatin remodeling factors, transcription factors, and small RNAs. TAIR tries to maintain up-to-date links to these resources from the TAIR Portal pages. Please contact TAIR by e-mail (*curator@arabidopsis.org*) if there are missing or nonfunctional links.

Arabidopsis Gene Expression Resources

There are a number of databases and tools that have been developed for storing, accessing and analyzing public Arabidopsis gene expression data that includes RNA seq, microarray and single cell RNA seq. TAIR stopped accepting microarray data in 2005 as ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) and the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) emerged as centralized community repositories. TAIR still provides access to the data via the Microarray Experiment (http://www.arabidopsis.org/servlets/Search?type=expr&search_action=new_search) and Microarray Expression searches (http://www.arabidopsis.org/servlets/Search?action=new_search&type=expression) for archival

purposes. In addition to the data and tools available at ArrayExpress, The BioAnalytic Resource for Plant Biology (http://bar.utoronto.ca/) hosts a number of user friendly tools for visualizing and analyzing Arabidopsis tissue RNA seq, microarray and single cell RNA seq expression data, motif analysis and more. It provides a virtual graphical gene expression map (eFP browser;http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi, see also Fig 1.11.2 A, item f) an Expression Angler tool which can be used to find similarly expressed genes (http://bar.utoronto.ca/ExpressionAngler/) and Expressolog TreeViewer (http://bar.utoronto.ca/expressolog_treeviewer/cgi-bin/expressolog_treeviewer.cgi) for finding expression orthologs. Another popular tool is GENEVESTIGATOR (*https://genevestigator.com/*), which contains most of the publicly available high-density array data from AtGenExpress (*http://arabidopsis.org/info/expression/ATGenExpress.jsp*) and other laboratories, and allows searching and displaying of the data (Zimmermann et al., 2004). Academic users must create a basic account, after which they can search for genes that are expressed in specific conditions, growth stages, or organs, or for genes of particular interest to them, and get a comprehensive view of the expression profiles in the different environmental conditions, growth stages, and organs. GENEVESTIGATOR requires subscriptions to access additional data and tools.

Arabidopsis Metabolic Pathways

There are several excellent resources for visualizing, analyzing and accessing information about biochemical pathways in Arabidopsis and other species. AraCyc is a curated metabolic pathway database specifically for *Arabidopsis thaliana* and is included in the Plant Metabolic Network (PMN, www.plantcyc.org) database. PMN includes over 350 plant species. The AraCyc database was initially built using the Pathologic module in the Pathway Tools software developed for MetaCyc (Karp et al., 2002; Mueller et al., 2003). Pathologic predicts possible metabolic pathways based upon the set of annotated enzymes available for a particular species. Following the initial computational build of AraCyc, pathways were manually validated and some were supplemented with additional experimental evidence. AraCyc includes tools to search and browse metabolic pathways drilling down to individual reactions and products, ways to visualize gene expression data superimposed on a global pathway map, and options to save data in Smart Tables. Unification links to MetaCyc and PlantCyc facilitate comparison of pathways from different organisms. TAIR includes extensive links out to AraCyc from the locus pages (see *BASIC PROTOCOL 2*, Figure 1.11.2 item m). Another resource is the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/kegg2.html) that includes pathways, reactions, enzymes, genome and other information for Arabidopsis and many other plant species. Plant Reactome (http://plantreactome.gramene.org; Naithani, et al., 2017) contains information about biochemical, genetic and other pathways, and tools for data visualization and analysis. Reactome curates pathways for the reference genome *Oryza sativa*, which is presented along with data for many other species, including Arabidopsis.

CONFLICT OF INTEREST STATEMENT:

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT:

TAIR data and tools are available to the public at www.arabidopsis.org according to the detailed

Terms of Use (https://www.arabidopsis.org/doc/about/tair_terms_of_use/417).

ACKNOWLEDGMENTS

The authors of this unit are grateful for the continued support of members of the research community who share their expertise, ideas, data and criticisms, all of which improve TAIR immensely. We thank all of the curators and programmers past and present who helped make TAIR such a valued resource. TAIR is supported by individual, institutional, corporate and government subscriptions. TAIR is a project of Phoenix Bioinformatics (www.phoenixbioinformatics.org), which is supported, in part, by a grant from the Alfred P. Sloan Foundation. PhyloGenes was co-developed by Phoenix Bioinformatics and the PANTHER project at the University of Southern California. It was supported by the National Science Foundation (DBI-1661543).

LITERATURE CITED

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D., and Rhee, S.Y. 2004. Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol.* 135:745-755.

Berardini T.Z., Li D., Muller R., Chetty R., Ploetz L., Singh S., Wensel A., Huala E. 2012 Assessment of community-submitted ontology annotations from a novel database-journal partnership. *Database* (Oxford). Aug 1;2012:bas030. doi: 10.1093/database/bas030.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015 The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis.* Aug;53(8):474-85. doi: 10.1002/dvg.22877

Bolser D, Staines DM, Pritchard E, Kersey P. 2016 Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. *Methods Mol Biol.* 1374:115-40. doi: 10.1007/978-1-4939-3167-5_6.

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, Holmes IH. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol*. 2016 Apr 12;17:66. doi: 10.1186/s13059-016-0924-1. PMID: 27072794; PMCID: PMC4830012.

Cheng, C-Y, Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. 2017. Araport 11: a complete reannotation of the Arabidopsis thaliana reference genome. *The Plant J.* DOI: 10.1111/tpj.13415

Eulgem, T., Rushton, P.J., Robatzek, S., and Somssich, I.E. 2000. The WRKY superfamily of plant transcription factors. *Trends Plant Sci.* 5:199-206.

Flanders, D.J., Weng, S., Petel, F.X., and Cherry, J.M. 1998. AtDB, the *Arabidopsis thaliana* database, and graphical-web-display of progress by the *Arabidopsis* Genome Initiative. *Nucleic Acids Res.* 26:80-84.

Garcia-Hernandez, M., Berardini, T.Z., Chen, G., Crist, D., Doyle, A., Huala, E., Knee, E., Lambrecht, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Rhee, S.Y., Scholl, R., Tacklind, J.,

Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. 2002. TAIR: A resource for integrated *Arabidopsis* data. *Funct. Integr. Genomics* 2:239-253.

The Gene Ontology Consortium. 2010. The gene ontology in 2010: Extensions and refinements. *Nucleic Acids Res* Jan;38(Database issue):D331-5. doi: 10.1093/nar/gkp1018.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 2012 Jan;40 (Database issue):D1178-86. doi: 10.1093/nar/gkr944.

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., Salzberg, S.L., and White, O. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31:5654-5666.

Hagen, G. and Guilfoyle, T. 2002. Auxin-responsive gene expression: Genes, promoters and regulatory factors. *Plant Mol. Biol.* 49:373-385.

Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L.A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C.D., Somerville, C., and Rhee, S.Y. 2001. The *Arabidopsis* Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* 29:102-105.

Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, Vincent L, Ware D, Zapata F. Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comp Funct Genomics.* 2005;6(7-8):388-97. doi: 10.1002/cfg.496.

Karp, P.D., Paley, S., and Romero, P. 2002. The pathway tools software. *Bioinformatics* 18:S225-S232.

Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, Rosen BD, Cheng CY, Moreira W, Mock SA, Stubbs J, Sullivan JM, Krampis K, Miller JR, Micklem G, Vaughn M, Town CD. 2015 Araport: the Arabidopsis information portal. *Nucleic Acids Res.* Jan;43(Database issue):D1003-9. doi: 10.1093/nar/gku1200.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E. 2012 The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* Jan;40(Database issue):D1202-10. doi: 10.1093/nar/gkr1090.

Leonelli S, Davey RP, Arnaud E, Parry G, Bastow R. 2017 Data management and best practice for plant science. *Nat Plants.* Jun 6;3:17086. doi: 10.1038/nplants.2017.86

Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J*. 2008 Feb;53(4):661-73. doi: 10.1111/j.1365-313X.2007.03326.x. PMID: 18269575.

Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013 Jan;41(Database issue):D377-86. doi: 10.1093/nar/gks1118. Epub 2012 Nov 27. PMID: 23193289; PMCID: PMC3531194.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., Thomas, P.D. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data

analysis tool enhancements. *Nucleic Acids Res.* 2017:45(D1):D183-D189

Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, Thomas PD. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res*. 2021 Jan 8;49(D1):D394-D403. doi: 10.1093/nar/gkaa1106. PMID: 33290554; PMCID: PMC7778891.

Mueller, L.A., Zhang, P., and Rhee, S.Y. 2003. AraCyc: A biochemical pathway database for *Arabidopsis*. *Plant Physiol.* 132:453-460.

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, Mandáková T, Jamge B, Lambing C, Kuo P, Yelina N, Hartwick N, Colt K, Smith LM, Ton J, Kakutani T, Martienssen RA, Schneeberger K, Lysak MA, Berger F, Bousios A, Michael TP, Schatz MC, Henderson IR. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Science. 2021 Nov 12;374(6569):eabi7489. doi: 10.1126/science.abi7489. Epub 2021 Nov 12. PMID: 34762468.

Naithani S, Preece J, D'Eustachio P, Gupta P, Amarasinghe V, Dharmawardhana PD, Wu G, Fabregat A, Elser JL, Weiser J, Keays M, Fuentes AM, Petryszak R, Stein LD, Ware D, Jaiswal P. 2017 Plant Reactome: a resource for plant pathways and comparative analysis. *Nucleic Acids Res*. Jan 4;45(D1):D1029-D1039. doi: 10.1093/nar/gkw932.

Pasha A, Subramaniam S, Cleary A, Chen X, Berardini T, Farmer A, Town C, Provart N. Araport Lives: An Updated Framework for Arabidopsis Bioinformatics. *Plant Cell*. 2020 Sep;32(9):2683-2686. doi: 10.1105/tpc.20.00358. Epub 2020 Jul 22. PMID: 32699173; PMCID: PMC7474289.

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K. 2015 PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res.* Jan;43(Database issue):D974-81. doi: 10.1093/nar/gku986

Reiser, L., Berardini, T.Z., Li, D., Muller, R., Strait, E.M., Li, Q., Mezheritsky, Y., Vetushko, A., Huala, E. 2016. Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. *Database (Oxford)* 2016: baw018

Reiser L, Harper L, Freeling M, Han B, Luan S. FAIR: A Call to Make Published Data More Findable, Accessible, Interoperable, and Reusable. *Mol Plant*. 2018 Sep 10;11(9):1105-1108. doi: 10.1016/j.molp.2018.07.005. Epub 2018 Aug 1. PMID: 30076986.

Rhee, S.Y., Weng, S., Bongard-Pierce, D.K., Garcia-Hernandez, M., Malekian, A., Flanders, D.J., and Cherry, J.M. 1999. Unified display of *Arabidopsis thaliana* physical maps from AtDB, the *A.thaliana* database. *Nucleic Acids Res.* 27:79-84.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. 2003. The *Arabidopsis* Information Resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* 31:224-228.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S. 2002. The generic genome browser: A building block for a model organism system database. *Genome Res.* 12:1599-1610.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P.,

and Huala, E. 2008. The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res.* 36:D1009-D1014.

Tang H, Finn RD, Thomas PD. TreeGrafter: phylogenetic tree-based annotation of proteins with Gene Ontology terms and other annotations. Bioinformatics. 2019 Feb 1;35(3):518-520. doi: 10.1093/bioinformatics/bty625. PMID: 30032202; PMCID: PMC6361231.

Weems, D., Miller, N., Garcia-Hernandez, M., Huala, E., and Rhee, S.Y. 2004. Design, implementation, and maintenance of a model organism database for *Arabidopsis thaliana*. *Comp. Funct. Genomics* 5:362-369.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019 Mar 19;6(1):6. PMID: 26978244; PMCID: PMC4792175.

Wortman, J.R., Haas, B.J., Hannick, L.I., Smith, R.K. Jr., Maiti, R., Ronning, C.M., Chan, A.P., Yu, C., Ayele, M., Whitelaw, C.A., White, O.R., and Town, C.D. 2003. Annotation of the *Arabidopsis* genome. *Plant Physiol.* 132:461-468.

Yan T, Yoo D, Berardini TZ, Mueller LA, Weems DC, Weng S, Cherry JM, Rhee SY. 2005 PatMatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic Acids Res.* Jul 1;33(Web Server issue):W262-6.

Zhang P, Berardini TZ, Ebert D, Li Q, Mi H, Muruganujan A, Prithvi T, Reiser L, Sawant S, Thomas PD, Huala E. PhyloGenes: An online phylogenetics and functional genomics resource for plant gene function inference. *Plant Direct.* 2020 Dec 30;4(12):e00293. doi: 10.1002/pld3.293. PMID: 33392435; PMCID: PMC7773024.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. 2004. GENEVESTIGATOR: *Arabidopsis* microarray database and analysis toolbox. *Plant Physiol.* 136:2621-2632.

Figure Legends

Figure 1.11.1 TAIR's home page (*http://arabidopsis.org*) is the main entry point to the database and Web site. The general search (item a) can be used for a quick search of database entities by name

Figure 1.11.2 A sample of a locus page from TAIR showing the major data included in the detail page. A portion of the germplasm section has been deleted for simplicity. Each of the data types displayed in the alternating colored bands can be grouped into one or more the following categories: (**A**) (a) general descriptive locus information, (b) gene model information, (c)

functional annotations, (d) nucleotide and protein sequences (e) gene expression data,(f) interactive BAR eFP browser image, (g) protein data, (h) plant homolog data; **B** (i) mapping data, (j) markers, polymorphisms and alleles, (k) germplasm information, (l) clones, (m) links to resources outside of TAIR, (n) community comments about the locus, (o) papers and abstracts and (p) locus history.

Figure 1.11.3 Overview of TAIR's JBrowse instance. (**A)** JBrowse display showing the track panel (left) and genome view panel (right). (**B**) Close up view showing the genome view controls including (a) Genome selector for choosing between different genome assemblies, (b) nucleotide sequence track, (c) left and right scroll buttons, (d) zoom controls, (e) search entry box, (f) highlighter, (g) track selector for adding new tracks, and (h) view control for modifying the display of tracks.

Figure 1.11.4 JBrowse genome view and track controls. (**A**) JBrowse genome view showing highlighting function (yellow) and pop-up detail (item a). (**B**) JBrowse page with track panel expanded to show different categories and options. The Epic CoGe link (item a) is used to retrieve tracks from CoGE. In addition to genome annotation data, JBrowse includes pre-loaded community tracks (item b) and whole genome alignments (item c). The arrow links from the selected community track to the actual display in the genome view.

Figure 1.11.5 JBrowse genome view options. (**A**) Example of a combined track generated by merging two different versions of the TAIR genome annotation (TAIR 10 and Araport 11). (**B**) SeqLighter pop up display.

Figure 1.11.6 Keyword results and detail page. (**A)** Keyword search results after querying for the GO Biological Process terms containing the words "root development". (**B**) A tree view of the term "root development" with expanded children nodes and associated gene annotations.

Figure 1.11.7 TAIR Functional Categorizaton. (**A**) Results display for functional categorization of *WRKY* genes. The members of this family are grouped into pre selected high level plant specific GO Slim categories based on their annotations to more granular GO terms. A complete list of plant GO Slim terms and IDs can be found on the TAIR website (https://www.arabidopsis.org/download_files/GO_and_PO_Annotations/Gene_Ontology_Annotations/TAIR_GO_slim_categories.txt). The results list can be re-sorted by choosing Gene count from the "re-sort by:" drop-down menu and clicking on the "re-sort" button. The keywords grouped by ontology aspect; Cellular Component, Biological Process and Molecular Function. The frequency of annotations to each category is listed in the last column; the number is linked to a list of genes annotated to the terms that are children of that category. (**B**) The drop down menu is used to select a graphical output format showing the distribution and frequency of annotations to each of the GO slim terms as either a bar graph or pie chart. A different graph/chart is created for each aspect of the GO ontologies.

Figure 1.11.8 GO Term Enrichment Tool at PANTHER (www.pantherdb.org). Results display for sample data from query input form, after it has been received and processed via the PANTHER web service. Results display (**A**) query parameters and identifier mapping, and (**B**) hierarchcal ordered table of term enrichment results.

Figure 1.11.9 Motif Finder tool. (**A**) Users can type in or upload a list of genes and select the promoter length to be analyzed. (**B**) The resulting motifs are listed with the corresponding genes in which they are found. Items in columns a-g are the values columns that match the headers listed in the summary.

Figure 1.11.10 Generic Online Annotation Tool (GOAT). (**A**) GOAT data submission page after logging in using ORCiD and clicking the Submission tab (item a). Users can click to additional genes (item b) and add additional annotation data entry fields (item c). After adding all of the desired annotations click on the review submission button (item d). (**B)** GOAT submission review screen. Users are asked to review their submission before finalizing. At this point it is possible to revise the submission (item a) or continue with the submission (item b).

Figure 1.11.11 PhyloGenes tree display. (**A)** Phylogenetic tree display for PANTHER family 45637 in PhyloGenes showing the results after entering the query 'PHOT1' into the search box (item a), the matching gene names are highlighted in the tree panel. The operations menu (item b) includes controls for tree pruning, downloading data files, expanding and collapsing nodes. The data panel controls (item c) allow for showing/hiding and reordering data columns or swapping between multiple sequence alignment (MSA) and data displays. The presence of icons in the data panel indicates the type of annotation; green trees indicate phylogenetic based annotations (item d) and yellow flasks (item e) indicate experimental annotations (f) boxed area shows annotation detail pop up displayed after clicking on the annotation icon. (**B)** Expanded view of tree pruning pop up selector. (**C**) Expanded view of data display configuration pop up.

Figure 1.11.12 PhyloGenes grafting. (**A**) PhyloGenes home page showing how to access the tree grafter (red circle). (**B**) Accessing the tree grafter from a search result (red circle). (**C**) Display of grafted sequence within an existing tree.

Figure 1.11.13 Results display from SyntenyViewer when there are matches. The upper table lists the query gene (item a) linked to the corresponding detail page in TAIR, the list of matched genes (item b) and links to the full display in GEVo (item c). The iframe below (item d) previews the alignment in GEVo.

The Arabidopsis Information Resource

About TAIR

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana* . Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every week from the latest published research literature and community data submissions. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

TAIR is located at Phoenix Bioinformatics and funded by subscriptions.

Full access to TAIR requires a subscription. Please see our subscription page for further details.

Note: This site has been tested with Chrome, Firefox, Safari, and Edge browsers. Some pages may not work as expected if you are using Internet Explorer. For best results, update your browser and enable Javascript and cookies (see help). **Scheduled Maintenance:** This site may be down for maintenance on any Saturday from 8 am to 10 am PDT.

Site Map | Terms of Use

Breaking News

See you in Belfast at ICAR2022
[Jun 7, 2022]

Come find us in Belfast at **ICAR2022!** Learn more at the Arabidopsis Informatics Session Thursday June 23, 2022 4PM GMT in person or on line. Or visit our booth.

Help AgBioData chart a course towards FAIR agricultural data!
[Apr 25, 2022]

Please take 10 minutes to complete this **important survey**

Your opinion is essential and can help AgBioData define better genomic, genetic, and breeding data curation practices.

30th public release of TAIR@Phoenix data
[Apr 1, 2022]
30th public release of data curated under TAIR's subscription-based funding model. Files contain new publications, annotations, gene symbols and other data through March 31, 2021.

Locus Page Updates: Plant Homologs
[Mar 22, 2022]
TAIR's locus pages have been updated with **new ways to view and download plant homologs**.

InterPro domain update
[Feb 23, 2022]
TAIR protein domains have been updated to InterPro 87.0

ICAR2022-Funding opportunities and deadlines
[Feb 11, 2022]
Upcoming deadlines for **ICAR2022** including funding opportunities.

A

a

Locus: AT3G45780

What's new on this page | Add a Comment

Representative Gene Model AT3G45780.1

Gene Model Type protein_coding

Other names: JK224, NONPHOTOTROPIC HYPOCOTYL 1, NPH1, PHOT1, PHOTOTROPIN 1, ROOT PHOTOTROPISM 1, RPT1

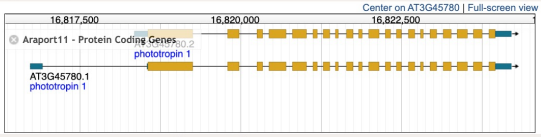
Description Blue-light photoreceptor. Contains a light activated serine-threonine kinase domain and LOV1 and LOV2 repeats. Mutants are defective in blue-light response. Mediates blue light-induced growth enhancements. PHOT1 and PHOT2 mediate blue light-dependent activation of the plasma membrane H+-ATPase in guard cell protoplasts. PHOT1 undergoes blue-light-dependent autophosphorylation. At least eight phosphorylation sites have been identified in PHOT1. Phosphorylation of serine851 in the activation loop of PHOT1 appears to be required for stomatal opening, chloroplast accumulation, leaf flattening, and phototropism, and phosphorylation of serine849 may also contribute to the regulation of these responses. Phosphorylation-dependent binding of 14-3-3 proteins to the Hinge1 region of PHOT1 appears to require serine350 and serine376.

Other Gene Models AT3G45780.2 (splice variant)

b

Map Detail Image

Center on AT3G45780 | Full-screen view



c

Annotations

category	relationship type	keyword
GO Biological Process	acts upstream of or within	blue light signaling pathway, chloroplast accumulation movement, chloroplast avoidance movement, circadian rhythm, negative regulation of anion channel activity by blue light, phototropism, protein autophosphorylation, regulation of proton transport, regulation of stomatal movement, response to blue light
GO Biological Process	involved in	protein phosphorylation
GO Cellular Component	is active in	cytoplasm, nucleus, plasma membrane
GO Cellular Component	located in	cell surface, cytoplasm, cytoplasmic side of plasma membrane, nucleus, plant-type vacuole
GO Molecular Function	enables	FMN binding, blue light photoreceptor activity, identical protein binding, kinase activity, mRNA binding, protein binding, protein kinase activity, protein serine/threonine kinase activity
Growth and Developmental Stages	expressed during	LP.02 two leaves visible stage, LP.04 four leaves visible stage, LP.06 six leaves visible stage, LP.08 eight leaves visible stage, LP.10 ten leaves visible stage, LP.12 twelve leaves visible stage, flowering stage, mature plant embryo stage, petal differentiation and expansion stage, plant embryo bilateral stage, plant embryo cotyledonary stage, plant embryo globular stage, vascular leaf senescent stage
Plant structure	expressed in	carpel, cauline leaf, collective leaf structure, cotyledon, flower, flower pedicel, guard cell, hypocotyl, inflorescence meristem, leaf apex, leaf lamina base, petal, petiole, plant embryo, pollen, root, seed, sepal, shoot apex, shoot system, stamen, stem, vascular leaf

Annotation Detail

d

Sequence full length CDS | full length genomic | full length cDNA | protein

e

RNA Data

Two-channel Arrays

array element name	avg. log ratio (std. error)	avg. intensity (std. error)
G8C11	-0.150 (0.036)	9531.351 (293.665)
H3G4	-0.208 (0.056)	6022.924 (283.107)

One-channel Arrays

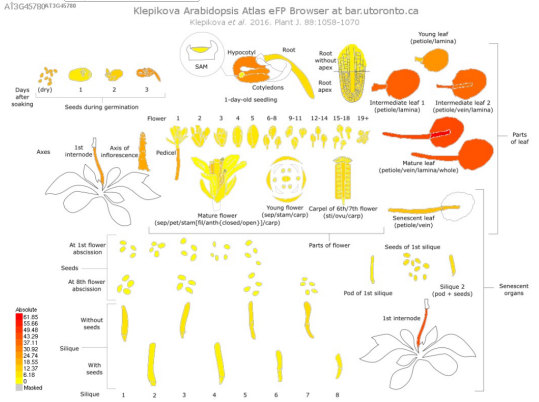
array element name	avg. signal intensity (std. error)	avg. signal percentile (std. error)
16120_AT	436.927 (102.516)	71.186 (5.158)
252543_AT	670.648 (14.924)	79.586 (0.604)

Associated Transcripts

type	number associated
EST	(187)
cDNA	(5)

f

BAR eFP Browser Data Source: Klepikova Atlas



Data from A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling: Klepikova et al., 2016, Plant J. 88:1058-1070. Total RNA was extracted with RNeasy Plant Kit and Illumina cDNA libraries were generated using the respective manufacturer's protocols. cDNA was then sequenced using Illumina HiSeq2000 with a 50bp read length. The read data are publicly available in NCBI's Sequence Read Archive under the BioProject ID 314076 (accession: PRJNA314076). Reads were aligned to the reference TAIR10 genome (Lamesch et al., 2012) using Tophat (Trapnell et al., 2009). Default Tophat settings and job resource parameters were used, with read groups unspecified. Reads per gene were counted with an in-house Python script using functions from the HTSeq package (Anders et al., 2015). Reads were filtered so that only uninterrupted reads corresponding to a region within exactly one gene were used for RPKM calculation. If a gene's expression level is not displayed, this indicates the reads for this gene did not pass the filtering criteria. RPKM values were compiled using an in-house R script.

Powered by BAR Webservices

g

Chromosome 3

Protein Data

name	length(aa)	molecular weight	isoelectric point	INTERPRO domains
AT3G45780.1	996	111687.9	7.73	PAS:IPR000014 Protein_kinase_ATP_BS:IPR017441 PAS-like_dom_sf:IPR035965 PAS-assoc_C:IPR000700 Ser/Thr_kinase_AS:IPR008271 Kinase-like_dom_sf:IPR011009 Prot_kinase_dom:IPR000719 PAC:IPR001610

h

Plant Homologs PHYLOGENES View complete family and tree

Arabidopsis paralogs Get Sequences | Download AGI IDs

AT1G16440(AGC1-6) , AT1G51170(UCN) ...

Plant orthologs Download Orthologs

Eudicotyledons		Monocotyledons	
Brassica napus	1	Brachypodium distachyon	1
Brassica rapa	1	Hordeum vulgare	1
Citrus sinensis	1	Musa acuminata	2
Cucumis sativus	1	Oryza sativa	2
Erythranthe guttata	1	Phoenix dactylifera	1
Eucalyptus grandis	1	Setaria italica	2
Glycine max	3	Sorghum bicolor	1
Gossypium hirsutum	3	Triticum aestivum	3
Helianthus annuus	2	Zea mays	1
Juglans regia	2	Zostera marina	1
Lactuca sativa	1		
Manihot esculenta	1	**Other plants**	
Medicago truncatula	2	Amborella trichopoda	1
Nicotiana tabacum	2	Chlamydomonas reinhardtii	1
Populus trichocarpa	1	Nelumbo nucifera	1
Prunus persica	1	Ostreococcus lauri	1
Ricinus communis	1	Physcomitrella patens	7
Solanum lycopersicum	1	Selaginella moellendorffii	4
Solanum tuberosum	1		
Spinacia oleracea	1		
Theobroma cacao	1		
Vitis vinifera	1		

Search Gene Families

EnsemblPlants | PhyloGenes
InParanoid Ortholog Groups | PhylomeDB
PANTHER | Phytozome
PGDD duplications and orthologs | PLAZA

B

i

Map Locations	chrom	map	map type	coordinates	orientation	attrib
	3	AGI	nuc_sequence	16816721 - 16824210 bp	forward	details

Map Links: MapViewer | Sequence Viewer | GBrowse | JBrowse

Genetic Markers	name	type	alias	chromosome	position	mapViewer	seqViewer
	NPH1	visible		3	61.0-61.0 cM	MapViewer	

j

Polymorphism Showing 15 of 172 entries (see all)

name	type	polymorphism site	gene names	allele type
FLAG_417E06	insertion	coding_region	AT3G45780.1	unknown
FLAG_417E06	insertion	exon	AT3G45780.2	unknown
FLAG_549A02	insertion	intron	AT3G45780.1, AT3G45780.2	unknown
GABI_366F11	insertion	intron	AT3G45780.1, AT3G45780.2	unknown
GABI_503F02	insertion	coding_region	AT3G45780.1, AT3G45780.2	unknown
GABI_572D04	insertion	coding_region	AT3G45780.1, AT3G45780.2	unknown
GABI_864E02	insertion	coding_region	AT3G45780.1, AT3G45780.2	unknown
GK-378F10-017220	insertion	intron	AT3G45780.1	unknown
GK-378F10-026789	insertion	intron	AT3G45780.1	unknown
GK-378F10-026823	insertion	intron	AT3G45780.1	unknown
GK-378F10-027009	insertion	intron	AT3G45780.1	unknown
GK-378F10-027039	insertion	intron	AT3G45780.1	unknown
GK-503F02-019716	insertion	exon	AT3G45780.1, AT3G45780.2	unknown
GK-606D03-021192	insertion	intron	AT3G45780.1, AT3G45780.2	unknown
GK-864E02-025980	insertion	exon	AT3G45780.1, AT3G45780.2	unknown

k

Germplasm Showing 15 of 31 entries (see all)

name	polymorphisms	background	stock name
SAIL_774_A12	SAIL_774_A12.v1; SAILseq_774_A12.1; SAILseq_774_A12.2; SAILseq_774_A12.3; SAILseq_774_A12.4		CS834599

Images: None available
Phenotypes: None available

SALK_060667	SALK_060667.48.45.x; SALKseq_060667.0; SALKseq_060667.1	SALK_060667

Images: None available
Phenotypes: None available

SAIL_21_C10 SAIL_21_C10.v1
Images: None available
Phenotypes: None available

Search at ABRC | Search at NASC (European Users) | Search at RIKEN

l

Clones (showing 15 of 32 entries) (see all)

Clone Name	Aliases	GenBank Accession	Vector Type	Stock Number
G8C11	G8C11T7		plasmid	G8C11
H3G4	H3G4T7		plasmid	H3G4
H5E9	H5E9T7		plasmid	H5E9
U09177		AY040062	plasmid	U09177
CATMA3A38785			plasmid	CD252516
G09177			plasmid	G09177
DKLAT3G45780			plasmid	DKLAT3G45780
CSHL_049503	AMR4844-99726478		plasmid	CSHL_049503
PDEST-AD062F03			plasmid	PDEST-AD062F03
PDEST-DB062F03			plasmid	PDEST-DB062F03
HALO_SFI_62-F03			plasmid	HALO_SFI_62-F03
RZL57G08F			plasmid	
APZL27E06R			plasmid	
APZL25H06R			plasmid	
APZL19G09R			plasmid	

Search at ABRC | Search at RIKEN

m

External Link

Epigenomics
Danforth Center Small RNA/PARE/Methylation
Expression/Localization
AraQTL
ATTED-II
eFP Browser
Eukaryotic Promoter Database (EPD)
Genevisible Expression Data
The Subcellular Location of Proteins in Arabidopsis Database (SUBA)
TraVA:Transcription Variation Analysis
Gene Families Links moved to Gene Families band [show me where]
Genomics
AceView
Gramene
MIPS
NCBI-Entrez Gene
Thalemine at BAR
Interactions
BioGRID
IntAct (Protein Interaction Database at EBI)
Metabolomics
View Aracyc reaction PROTEIN-KINASE-RXN (2.7.11.1)
Other
BAR ePlant
BAR ePlant Molecule View
Polymorphisms/Alleles/Variants
Salk SNP Viewer
T-DNA Express
Proteomics
AlphaFold Protein Structure Database
AtProteome
Functional Analysis Tools for Post-Translational Modifications(FAT-PTM)
P3DB (Plant Protein Phosphorylation DataBase)
Plant Proteome Database
Plant PTM Viewer
Publications
EVEX
Reagents
Agrisera (antibody)
PhytoAB (antibody)
RIKEN BioResource Research Center (Seed/DNA)

n

Community Comments (shows only the most recent comments by default)

Add My Comment | Hide Comments | Show All Comments

o

Publication

author/title	source	associated loci	date
Zeng, Y. Schotte, S. Trinh, H K. V… Genetic Dissection of Light-Regulated Adventitious Root Induction in Arabidopsis thaliana Hypocotyls	INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES	AT2G32950 AT2G46340 AT3G15354 AT3G45780 AT4G08920 AT4G11110 AT4G14110 AT5G11260 AT5G58140	2022
Grunwald, Y. Gosa, S C. Srivasta… Out of the blue: Phototropins of the leaf vascular bundle sheath mediate the regulation of leaf hydraulic conductance by blue light	THE PLANT CELL	AT3G45780 AT4G30190 AT5G58140	2022
Beatrice, P. Chiatante, D. Scippa,… Photoreceptors' gene expression of Arabidopsis thaliana grown with biophilic LED-sourced lighting systems	PLOS ONE	AT1G04400 AT1G09570 AT3G45780 AT4G08920 AT4G16250 AT5G35840 AT5G63860	2022
Zeidler, M Physiological Analysis of Phototropic Responses to Blue and Red Light in Arabidopsis	METHODS IN MOLECULAR BIOLOGY	AT1G09570 AT2G18790 AT3G45780 AT5G58140	2022
Kimura, T. Haga, K. Sakai, T The phosphorylation status of NONPHOTOTROPIC HYPOCOTYL3 affects phot2-dependent phototropism in Arabidopsis	PLANT SIGNAL BEHAV	AT3G45780 AT5G58140 AT5G64330	2022
Eckstein, A. Grzyb, J. Hermanowi… Arabidopsis Phototropins Participate in the Regulation of Dark-Induced Leaf Senescence	INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES	AT3G45780 AT5G58140	2021
Miao, L. Zhao, J. Yang, G. Xu, P. … Arabidopsis cryptochrome 1 undergoes COP1 and LRBs-dependent degradation in response to high blue light	THE NEW PHYTOLOGIST	AT1G04400 AT1G09570 AT2G18790 AT2G32950 AT3G45780 AT3G52740 AT4G08920	2021
Liu, Y S. Song, H G. Kim, H S. K… Guard-Cell-Specific Expression of Phototropin2 C-Terminal Fragment Enhances Leaf Transpiration	PLANTS (BASEL)	AT3G45780 AT4G14480 AT5G58140	2021
Zhai, S. Cai, W. Xiang, Z X. Chen… PIN3-mediated auxin transport contributes to blue light-induced adventitious root formation in Arabidopsis	PLANT SCIENCE : AN INTERNATIONAL JOURNAL OF EXPERIMENTAL PLANT BIOLOGY	AT1G70940 AT3G45780 AT5G58140 AT5G64330	2021
Kimura, T. Haga, K. Nomura, Y. H… Phosphorylation of NONPHOTOTROPIC HYPOCOTYL3 affects photosensory adaptation during the phototropic response	PLANT PHYSIOLOGY	AT2G30520 AT3G45780 AT5G64330	2021
Wang, J. Liang, Y P. Zhu, J D. Wa… Phototropin 1 Mediates High-Intensity Blue Light-Induced Chloroplast Accumulation Response in a Root Phototropism 2-Dependent Manner in Arabidopsis phot2 Mutant Plants	FRONT PLANT SCI	AT1G75100 AT2G30520 AT3G45780 AT5G58140 AT5G67385	2021
Rusaczonek A. Czarnocka W. Wil… Phototropin 1 and 2 Influence Photosynthesis, UV-C Induced Photooxidative Stress Responses, and Cell Death.	CELLS	AT3G45780 AT5G58140	2021
?abuz, J. Sztatelman, O. Jagie??… Phototropin interactions with SUMO proteins	PLANT AND CELL PHYSIOLOGY	AT3G45780 AT5G55170 AT5G58140 AT5G60410	2021
Sullivan, S. Waksman, T. Pallogia… Regulation of plant phototropic growth by NPH3/RPT2-like substrate phosphorylation and 14-3-3 binding	NATURE COMMUNICATIONS	AT3G45780 AT5G64330	2021

View Complete List (15 of 239 displayed)

p

Update History AT3G45780 replaces RPT1 on 2004-02-23

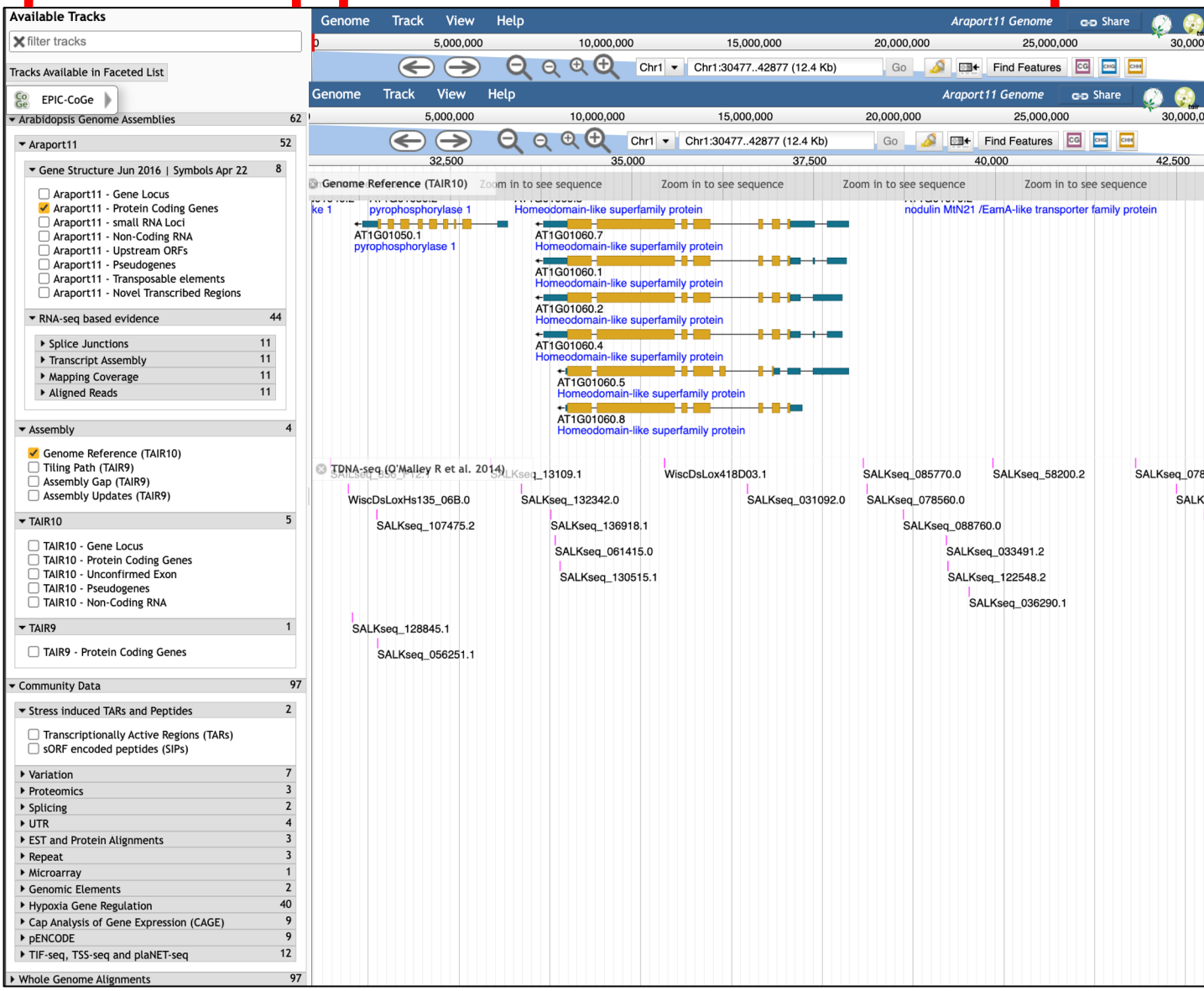
Date last modified 2015-11-30

TAIR Accession Locus:2102674

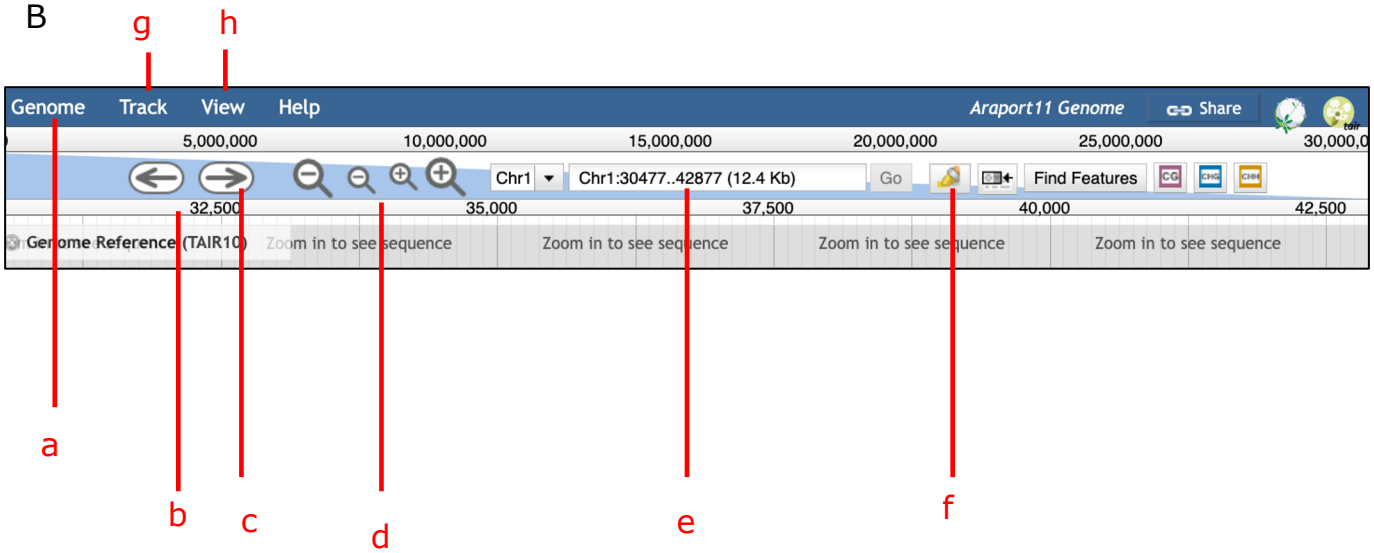
A

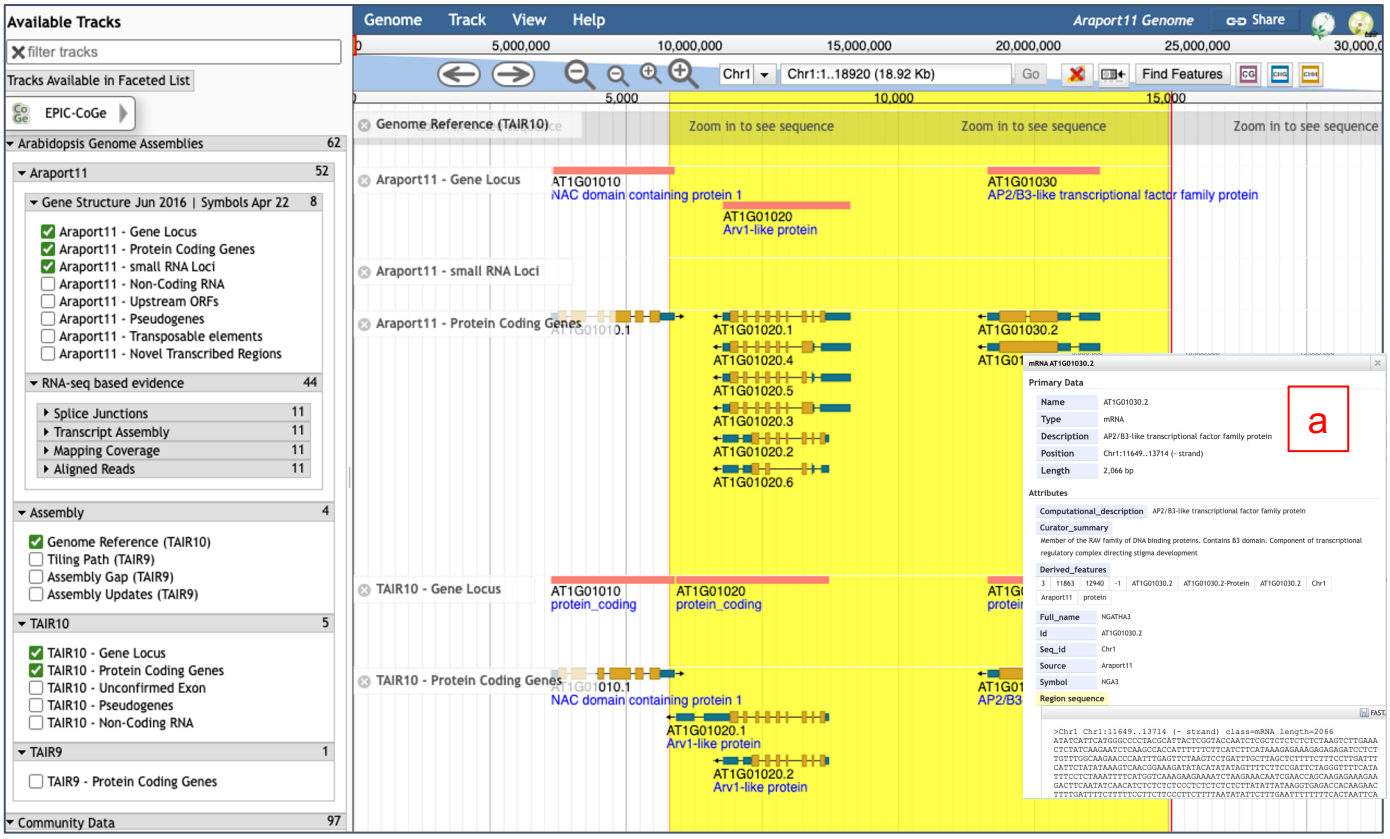
Track panel

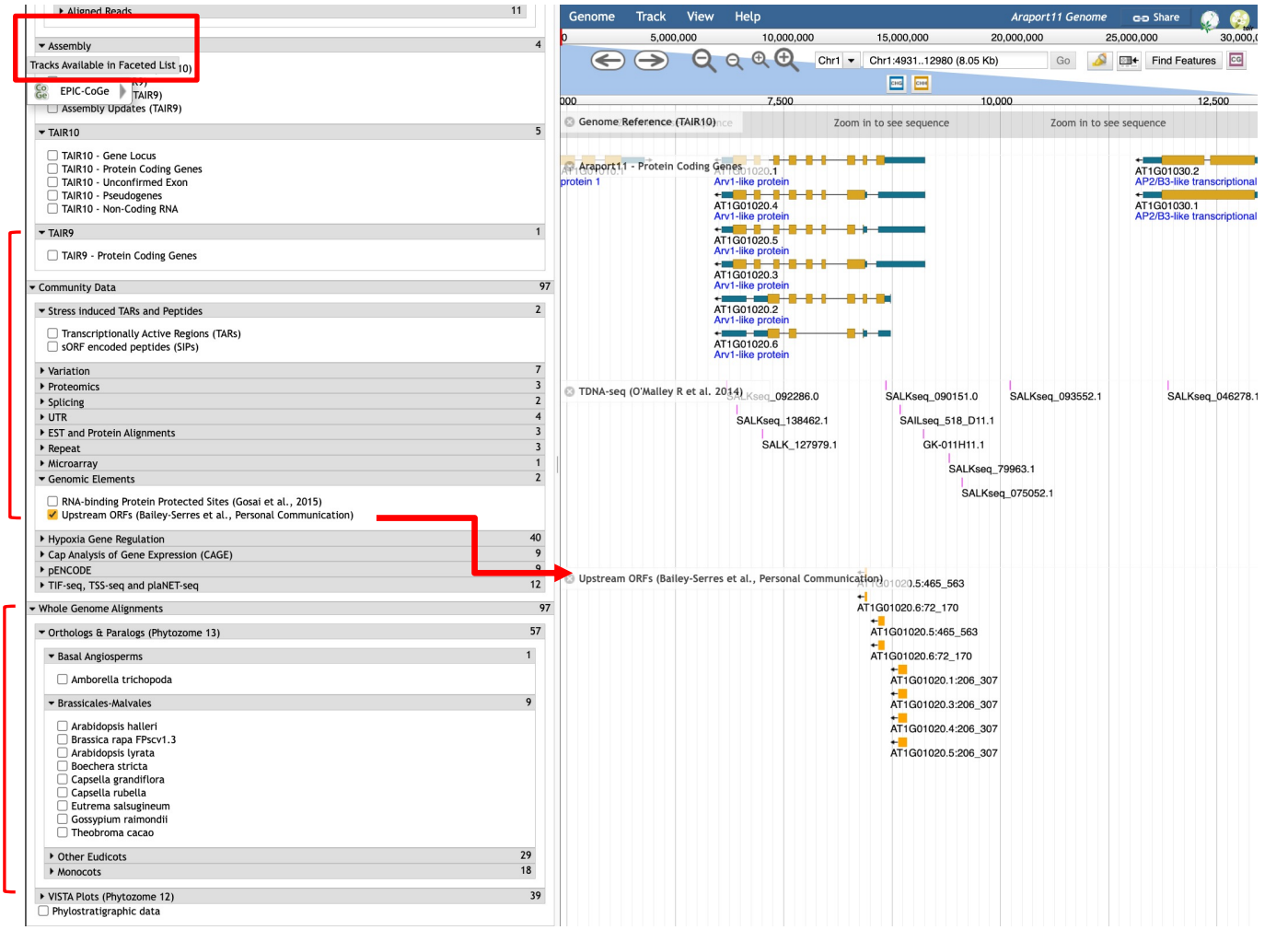
Genome view panel



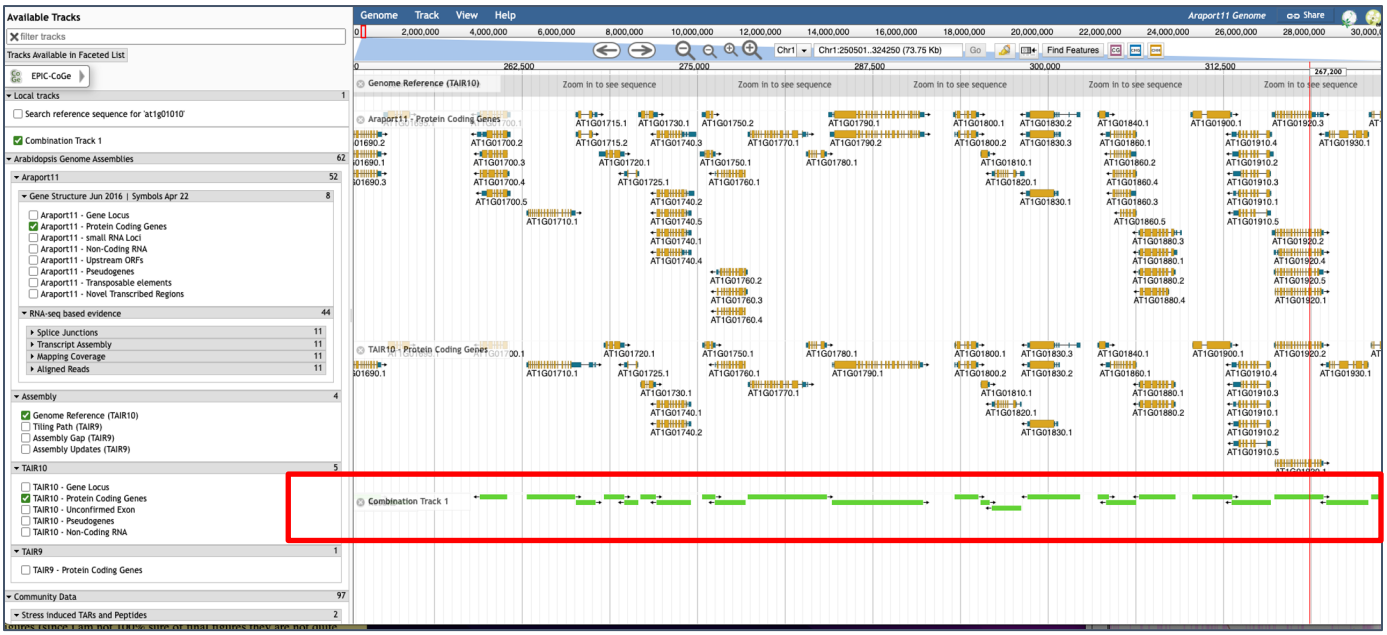
B



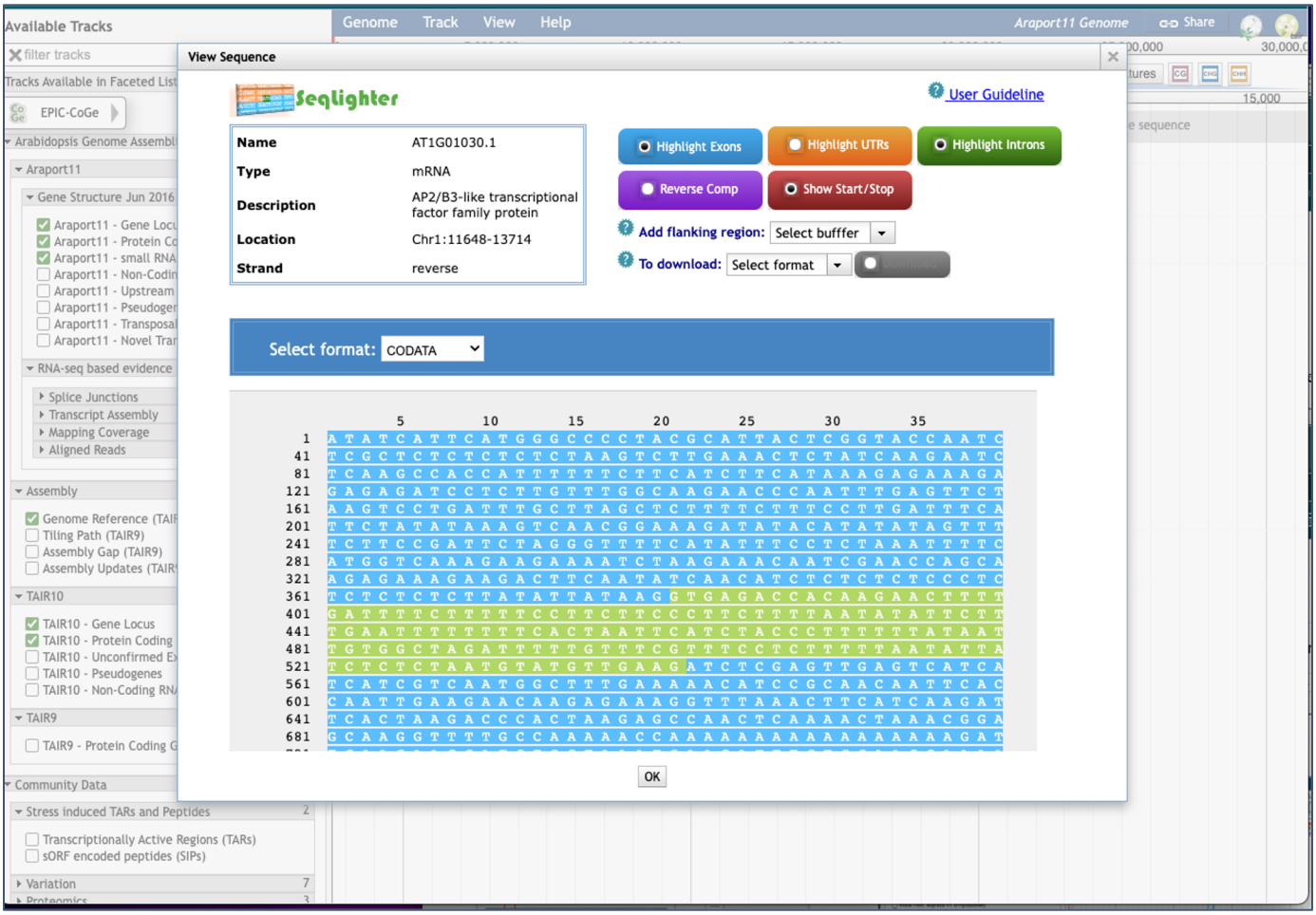

Available Tracks
filter tracks
Tracks Available in Faceted List
EPIC-CoGe
Arabidopsis Genome Assemblies 62
Araport11 52
Gene Structure Jun 2016 | Symbols Apr 22 8
Araport11 - Gene Locus
Araport11 - Protein Coding Genes
Araport11 - small RNA Loci
Araport11 - Non-Coding RNA
Araport11 - Upstream ORFs
Araport11 - Pseudogenes
Araport11 - Transposable elements
Araport11 - Novel Transcribed Regions
RNA-seq based evidence 44
Splice Junctions 11
Transcript Assembly 11
Mapping Coverage 11
Aligned Reads 11
Assembly 4
Genome Reference (TAIR10)
Tiling Path (TAIR9)
Assembly Gap (TAIR9)
Assembly Updates (TAIR9)
TAIR10 5
TAIR10 - Gene Locus
TAIR10 - Protein Coding Genes
TAIR10 - Unconfirmed Exon
TAIR10 - Pseudogenes
TAIR10 - Non-Coding RNA
TAIR9 1
TAIR9 - Protein Coding Genes
Community Data 97
Genome Track View Help
Araport11 Genome
Share
Chr1 Chr1:1..18920 (18.92 Kb) Go Find Features
Genome Reference (TAIR10)
Zoom in to see sequence
Araport11 - Gene Locus
AT1G01010
NAC domain containing protein 1
AT1G01020
Arv1-like protein
AT1G01030
AP2/B3-like transcriptional factor family protein
Araport11 - small RNA Loci
Araport11 - Protein Coding Genes
AT1G01020.1
AT1G01020.4
AT1G01020.5
AT1G01020.3
AT1G01020.2
AT1G01020.6
AT1G01030.2
TAIR10 - Gene Locus
protein_coding
TAIR10 - Protein Coding Genes
a
Primary Data
Name AT1G01030.2
Type mRNA
Description AP2/B3-like transcriptional factor family protein
Position Chr1:11649..13714 (- strand)
Length 2,066 bp
Attributes
Computational_description AP2/B3-like transcriptional factor family protein
Curator_summary
Member of the RAV family of DNA binding proteins. Contains B3 domain. Component of transcriptional regulatory complex directing stigma development
Derived_features
Full_name NGATHA3
Id AT1G01030.2
Seq_id Chr1
Source Araport11
Symbol NGA3
Region sequence


a
b
c
Aligned Reads 11
Assembly 4
Tracks Available in Faceted List
EPIC-CoGe
Assembly Updates (TAIR9)
TAIR10 5
TAIR10 - Gene Locus
TAIR10 - Protein Coding Genes
TAIR10 - Unconfirmed Exon
TAIR10 - Pseudogenes
TAIR10 - Non-Coding RNA
TAIR9 1
TAIR9 - Protein Coding Genes
Community Data 97
Stress induced TARs and Peptides 2
Transcriptionally Active Regions (TARs)
sORF encoded peptides (SIPs)
Variation 7
Proteomics 3
Splicing 2
UTR 4
EST and Protein Alignments 3
Repeat 3
Microarray 1
Genomic Elements 2
RNA-binding Protein Protected Sites (Gosai et al., 2015)
Upstream ORFs (Bailey-Serres et al., Personal Communication)
Hypoxia Gene Regulation 40
Cap Analysis of Gene Expression (CAGE) 9
pENCODE 9
TIF-seq, TSS-seq and plaNET-seq 12
Whole Genome Alignments 97
Orthologs & Paralogs (Phytozome 13) 57
Basal Angiosperms 1
Amborella trichopoda
Brassicales-Malvales 9
Arabidopsis halleri
Brassica rapa FPscv1.3
Arabidopsis lyrata
Boechera stricta
Capsella grandiflora
Capsella rubella
Eutrema salsugineum
Gossypium raimondii
Theobroma cacao
Other Eudicots 29
Monocots 18
VISTA Plots (Phytozome 12) 39
Phylostratigraphic data
Genome Track View Help
Araport11 Genome
Share
Chr1 Chr1:4931..12980 (8.05 Kb) Go Find Features
Genome Reference (TAIR10)
Zoom in to see sequence
Araport11 - Protein Coding Genes
AT1G01020.1
Arv1-like protein
AT1G01020.4
AT1G01020.5
AT1G01020.3
AT1G01020.2
AT1G01020.6
AT1G01030.2
AP2/B3-like transcriptional
AT1G01030.1
TDNA-seq (O'Malley R et al. 2014)
SALKseq_092286.0
SALKseq_090151.0
SALKseq_093552.1
SALKseq_046278.1
SALKseq_138462.1
SAILseq_518_D11.1
SALK_127979.1
GK-011H11.1
SALKseq_79963.1
SALKseq_075052.1
Upstream ORFs (Bailey-Serres et al., Personal Communication)
AT1G01020.6:72_170
AT1G01020.5:465_563
AT1G01020.6:72_170
AT1G01020.1:206_307
AT1G01020.3:206_307
AT1G01020.4:206_307
AT1G01020.5:206_307

A


Available Tracks
Genome
Track
View
Help
Araport11 Genome
Share
Genome Reference (TAIR10)
Araport11 - Protein Coding Genes
TAIR10 - Protein Coding Genes
Combination Track 1

B


View Sequence
Seqlighter
User Guideline
Name AT1G01030.1
Type mRNA
Description AP2/B3-like transcriptional factor family protein
Location Chr1:11648-13714
Strand reverse
Highlight Exons
Highlight UTRs
Highlight Introns
Reverse Comp
Show Start/Stop
Add flanking region: Select bufffer
To download: Select format
Select format: CODATA
OK

A

Keyword Search

keyword starts with

Restrict your search to keyword category by checking the box below

GO Cellular Component | GO Biological Process | Plant Growth and Developmental Stages
GO Molecular Function | Plant Anatomical Entity | Experimental Method | User defined

reset | submit query

TAIR Keyword Search Results

Your query for keywords where contains root development resulted in 15 matches.

Displaying 1 - 15 of 15 records on page 1 of 1 pages.

Keyword	Keyword Category	Tree View	Associated Data(to this term and to children terms)
root development	GO Biological Process	treeview	629 loci, 2202 publications, 1626 annotations
post-embryonic root development	GO Biological Process	treeview	162 loci, 793 publications, 199 annotations
lateral root development	GO Biological Process	treeview	145 loci, 768 publications, 182 annotations
root development stage	Growth and Developmental Stages	treeview	18 loci, 529 publications, 27 annotations
3 establishment of tissue systems stage	Growth and Developmental Stages	treeview	4 publications
4 root elongation stage	Growth and Developmental Stages	treeview	12 loci, 417 publications, 12 annotations
5 root hair formation stage	Growth and Developmental Stages	treeview	1 loci, 77 publications, 1 annotations
adventitious root development	GO Biological Process	treeview	6 loci, 20 publications, 9 annotations
primary root development	GO Biological Process	treeview	33 loci, 39 publications, 37 annotations
regulation of lateral root development	GO Biological Process	treeview	29 loci, 13 publications, 35 annotations
regulation of post-embryonic root development	GO Biological Process	treeview	31 loci, 14 publications, 37 annotations
regulation of root development	GO Biological Process	treeview	76 loci, 44 publications, 112 annotations
negative regulation of lateral root development	GO Biological Process	treeview	8 loci, 4 publications, 11 annotations
positive regulation of lateral root development	GO Biological Process	treeview	4 loci, 1 publications, 4 annotations
cell wall polysaccharide catabolic process involved in lateral root development	GO Biological Process	treeview	

B

TAIR Keyword Browser [Help]

Display | loci | publications | annotations | microarray experiments

Check the box and click the display button to see numbers of associated data

Keyword: root development
ID: GO:0048364
Definition: The process whose specific outcome is the progression of the root over time, from its formation to the mature structure. The root is the water- and mineral-absorbing part of a plant which is usually underground, does not bear leaves, tends to grow downwards and is typically derived from the radicle of the embryo.

I = 'is a' relationship | P = 'part of' relationship | D = 'develops from' relationship
R = 'regulates' relationship | R = 'positively regulates' relationship | R = 'negatively regulates' relationship

Keyword Categories - Click on the link to generate a treeview for the category.

GO Cellular Component | GO Biological Process | Plant Growth and Developmental Stages
GO Molecular Function | Plant Anatomical Entity | Experimental Method

- all
  - biological_process
    - developmental process
      - anatomical structure development
        - multicellular organism development
          - system development
            - root system development
              - root development
                - root morphogenesis (23 loci to term + 322 loci to children)
                - post-embryonic root development (14 loci to term + 148 loci to children)
                - adventitious root development (6 loci to term)
                - root cap development (15 loci to term)
                - root meristem growth (14 loci to term + 37 loci to children)
                - stele development (1 loci to term)
                - primary root development (33 loci to term)
                - root radial pattern formation (9 loci to term)
                - regulation of root development (54 loci to term + 31 loci to children)
                - root regeneration (2 loci to term)
              - root development

A

Functional Categorization Listing [Help]

new search

Annotation Pie Chart | Draw | re-sort by | Annotation Count

Displaying 99 records.

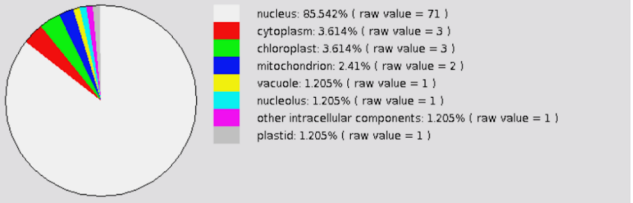
Keyword Category	Functional Category	Annotation Count	Gene Count
GO Cellular Component	nucleus	71	71
GO Cellular Component	cytoplasm	3	3
GO Cellular Component	chloroplast	3	2
GO Cellular Component	mitochondrion	2	2
GO Cellular Component	vacuole	1	1
GO Cellular Component	nucleolus	1	1
GO Cellular Component	other intracellular components	1	1
GO Cellular Component	plastid	1	1
GO Cellular Component	plasma membrane	0	0
GO Cellular Component	external encapsulating structure	0	0
GO Cellular Component	cytoskeleton	0	0
GO Cellular Component	lysosome	0	0
GO Cellular Component	nucleoplasm	0	0
GO Cellular Component	endoplasmic reticulum	0	0
GO Cellular Component	cell wall	0	0
GO Cellular Component	other membranes	0	0
GO Cellular Component	peroxisome	0	0
GO Cellular Component	thylakoid	0	0
GO Cellular Component	cytosol	0	0
GO Cellular Component	extracellular region	0	0
GO Cellular Component	unknown cellular components	0	0
GO Cellular Component	other cellular components	0	0
GO Cellular Component	ribosome	0	0
GO Cellular Component	endosome	0	0
GO Cellular Component	nuclear envelope	0	0
GO Cellular Component	Golgi apparatus	0	0
GO Molecular Function	DNA binding	93	70
GO Molecular Function	DNA-binding transcription factor activity	71	71
GO Molecular Function	protein binding	51	45
GO Molecular Function	nucleic acid binding	47	44
GO Molecular Function	other binding	3	3
GO Molecular Function	nucleotide binding	2	2
GO Molecular Function	transferase activity	1	1
GO Molecular Function	catalytic activity	1	1
GO Molecular Function	kinase activity	1	1
GO Molecular Function	translation factor activity, RNA binding	0	0
GO Molecular Function	RNA binding	0	0
GO Molecular Function	carbohydrate binding	0	0
GO Molecular Function	transporter activity	0	0
GO Molecular Function	oxygen binding	0	0
GO Molecular Function	translation regulator activity	0	0
GO Molecular Function	other molecular functions	0	0
GO Molecular Function	structural molecule activity	0	0
GO Molecular Function	chromatin binding	0	0
GO Molecular Function	enzyme regulator activity	0	0
GO Molecular Function	signaling receptor activity	0	0
GO Molecular Function	transcription regulator activity	0	0
GO Molecular Function	motor activity	0	0
GO Molecular Function	signaling receptor binding	0	0
GO Molecular Function	unknown molecular functions	0	0
GO Molecular Function	lipid binding	0	0
GO Molecular Function	nuclease activity	0	0
GO Molecular Function	hydrolase activity	0	0
GO Biological Process	other cellular processes	131	70
GO Biological Process	response to stress	116	43
GO Biological Process	other metabolic processes	90	70
GO Biological Process	response to chemical	80	36
GO Biological Process	biosynthetic process	79	70
GO Biological Process	nucleobase-containing compound metabolic process	74	70
GO Biological Process	response to external stimulus	72	32
GO Biological Process	response to biotic stimulus	69	30
GO Biological Process	anatomical structure development	35	23
GO Biological Process	response to abiotic stimulus	34	14
GO Biological Process	multicellular organism development	33	25
GO Biological Process	signal transduction	28	17
GO Biological Process	response to endogenous stimulus	26	17
GO Biological Process	cell communication	15	6
GO Biological Process	reproduction	10	9
GO Biological Process	post-embryonic development	10	9
GO Biological Process	secondary metabolic process	7	6
GO Biological Process	cell differentiation	7	7
GO Biological Process	other biological processes	6	5
GO Biological Process	catabolic process	4	4
GO Biological Process	flower development	3	3
GO Biological Process	transport	2	2
GO Biological Process	embryo development	2	2
GO Biological Process	response to light stimulus	2	2
GO Biological Process	cell death	2	2
GO Biological Process	growth	1	1
GO Biological Process	cellular component organization	1	1
GO Biological Process	pollination	1	1
GO Biological Process	protein metabolic process	0	0
GO Biological Process	cellular homeostasis	0	0
GO Biological Process	regulation of gene expression, epigenetic	0	0
GO Biological Process	lipid metabolic process	0	0
GO Biological Process	cell growth	0	0
GO Biological Process	regulation of molecular function	0	0
GO Biological Process	photosynthesis	0	0
GO Biological Process	unknown biological processes	0	0
GO Biological Process	tropism	0	0
GO Biological Process	cell-cell signaling	0	0
GO Biological Process	generation of precursor metabolites and energy	0	0
GO Biological Process	fruit ripening	0	0
GO Biological Process	carbohydrate metabolic process	0	0
GO Biological Process	cell cycle	0	0
GO Biological Process	DNA metabolic process	0	0
GO Biological Process	translation	0	0
GO Biological Process	circadian rhythm	0	0
GO Biological Process	abscission	0	0

B

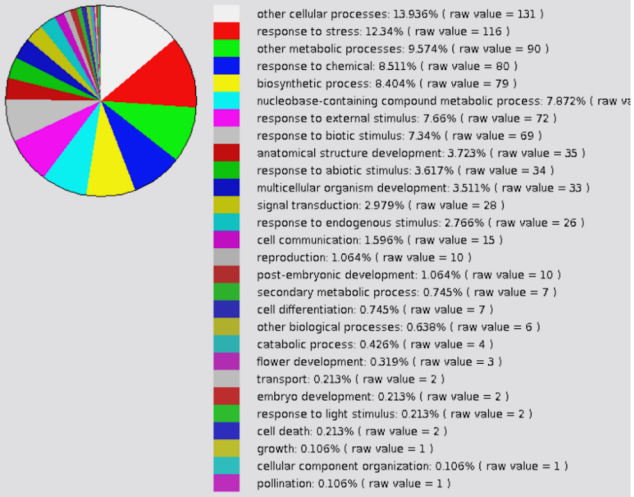
Charts for Functional Categorization [Help]

$$\frac{\text{\# of annotations to terms in this GOslim category} \times 100}{\text{\# of total annotations to terms in this ontology}} = \%$$

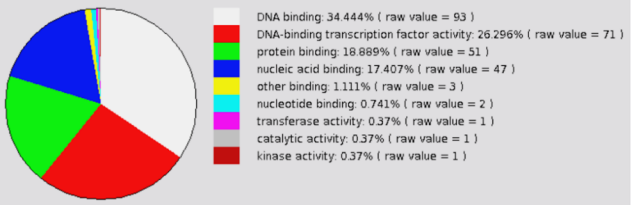
Functional Categorization by annotation for : GO Cellular Component



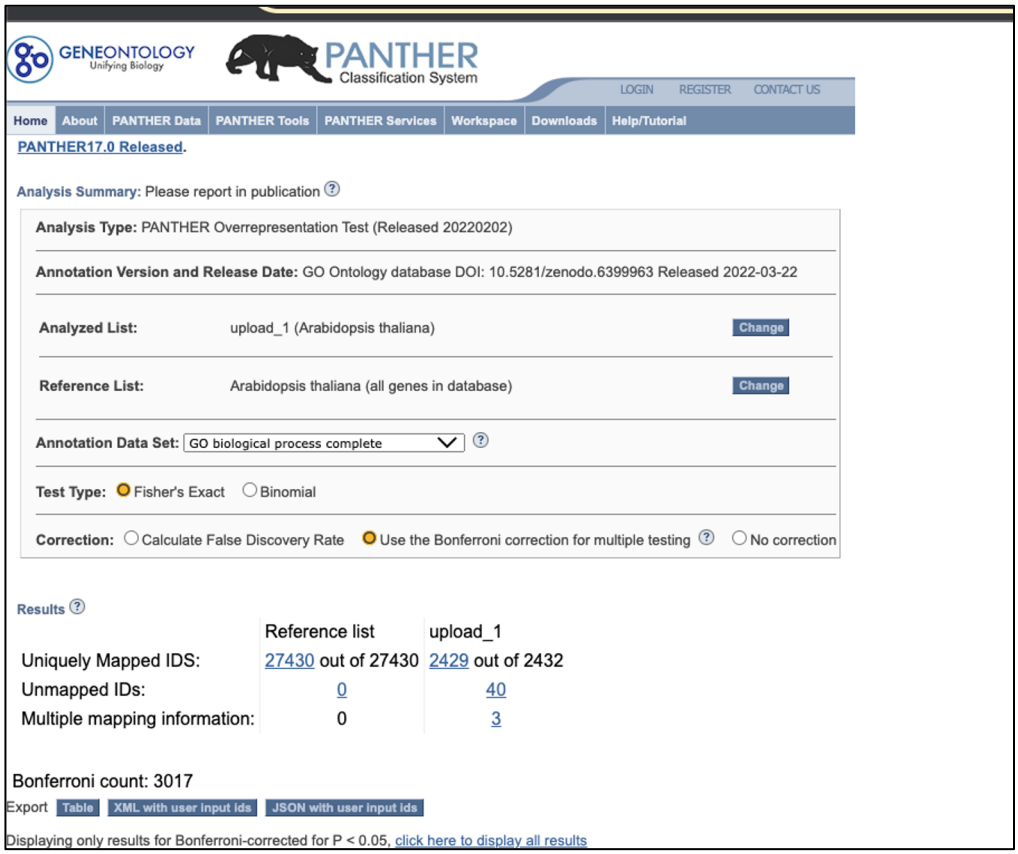
Functional Categorization by annotation for : GO Biological Process



Functional Categorization by annotation for : GO Molecular Function



A



B

	Arabidopsis thaliana (REF)	upload_1 (▽ Hierarchy NEW!)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
cellular response to hypoxia	238	119	21.10	5.64	+	5.16E-39
↳cellular response to stress	1200	229	106.39	2.15	+	3.14E-20
↳cellular response to stimulus	3901	680	345.87	1.97	+	6.08E-59
↳cellular process	15276	1647	1354.40	1.22	+	1.03E-27
↳response to stimulus	9347	1398	828.72	1.69	+	8.55E-109
↳response to stress	5305	971	470.35	2.06	+	7.50E-106
↳response to hypoxia	325	142	28.82	4.93	+	1.09E-41
↳response to decreased oxygen levels	333	143	29.52	4.84	+	2.62E-41
↳response to oxygen levels	335	143	29.70	4.81	+	4.54E-41
↳response to abiotic stimulus	4145	698	367.50	1.90	+	1.72E-55
↳cellular response to decreased oxygen levels	240	120	21.28	5.64	+	2.31E-39
↳cellular response to oxygen levels	241	120	21.37	5.62	+	3.19E-39
↳cellular response to chemical stimulus	2081	479	184.51	2.60	+	2.69E-69
↳response to chemical	5130	894	454.84	1.97	+	3.42E-84
response to chitin	320	159	28.37	5.60	+	8.81E-53
↳response to oxygen-containing compound	3119	616	276.54	2.23	+	1.50E-69
↳response to organonitrogen compound	681	215	60.38	3.56	+	1.56E-45
↳response to organic substance	3557	665	315.37	2.11	+	8.66E-68
↳response to nitrogen compound	825	242	73.15	3.31	+	5.17E-47

A

Home > Tools > Motif Analysis

Statistical Motif Analysis in Promoter or Upstream Gene Sequences

The program compares the frequencies of 6-mer "words" in your query set of sequences (on both strands) with the frequencies of the words in the current genomic sequence set of 33518 sequences, each containing 500 (or 1000) bp upstream of the start codon of each gene. You can type in sets of AGI locus identifiers (e.g. At1g01030) or sets of fasta sequences. Make sure each fasta header is formated as such, fasta symbol (>), immediatly followed by a unique ID, a space, then all other descriptions (e.g. >ABCD1.1 my gene). Ensure that there are no sequences appearing more than once in your query set.

```
At3g46230
At5g12020
At4g10250
At5g12030
At1g69920
At5g52760
At2g26150
At1g59860
At2g28210
At1g66090
At3g02840
At3g54150
At1g53540
At5g42380
```

Upload file: Choose File No file chosen

Dataset:
(•) 500 bp upstream ( ) 1000 bp upstream ( ) 3000 bp upstream

Output type:
(•) HTML ( ) Text Reset submit

B

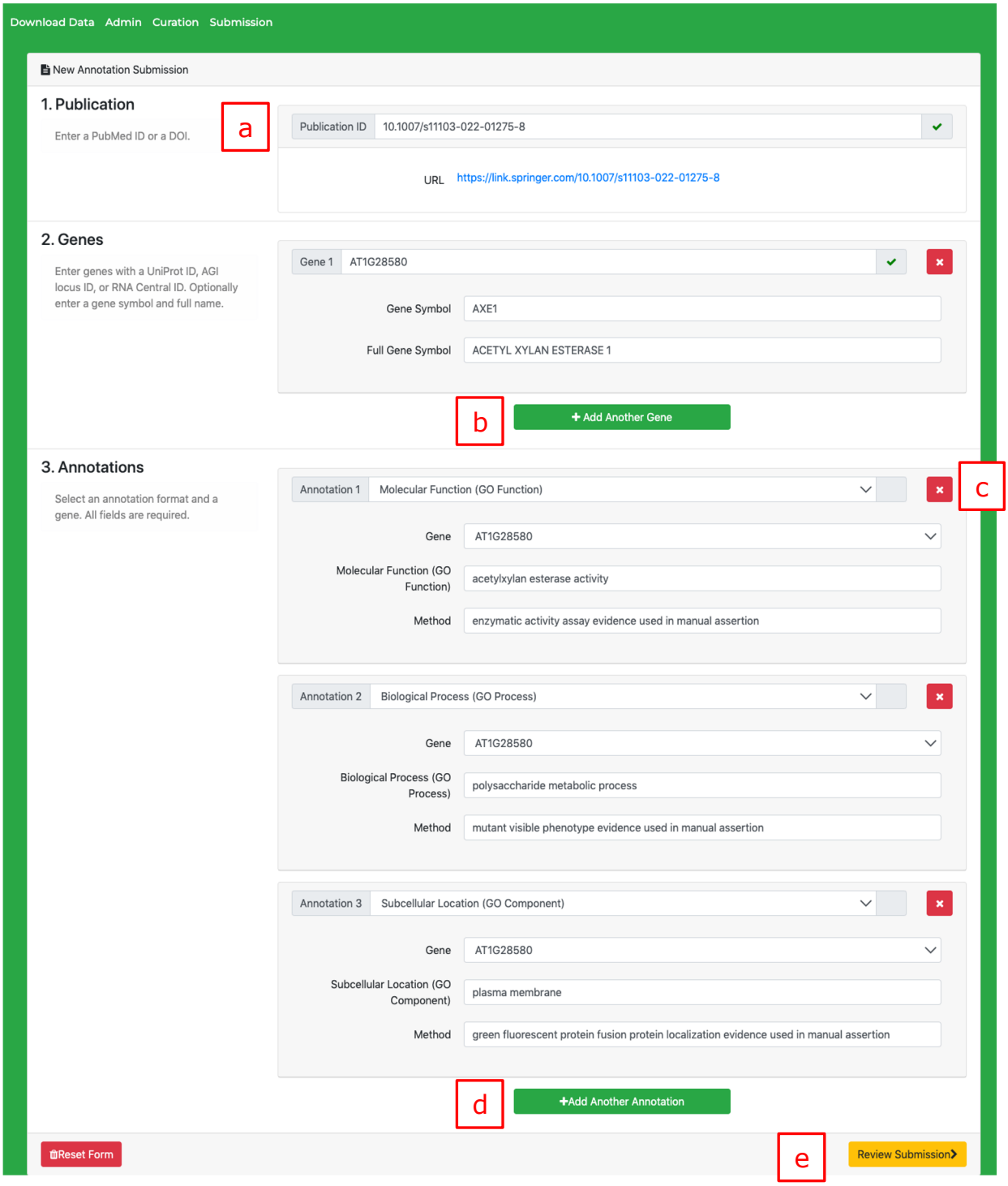
Motif Analysis in Promoter/Upstream Sequences

Only oligos occurring in 3 or more of sequences in the query set are reported, and are sorted by p-value. Columns are as follows (left to right):

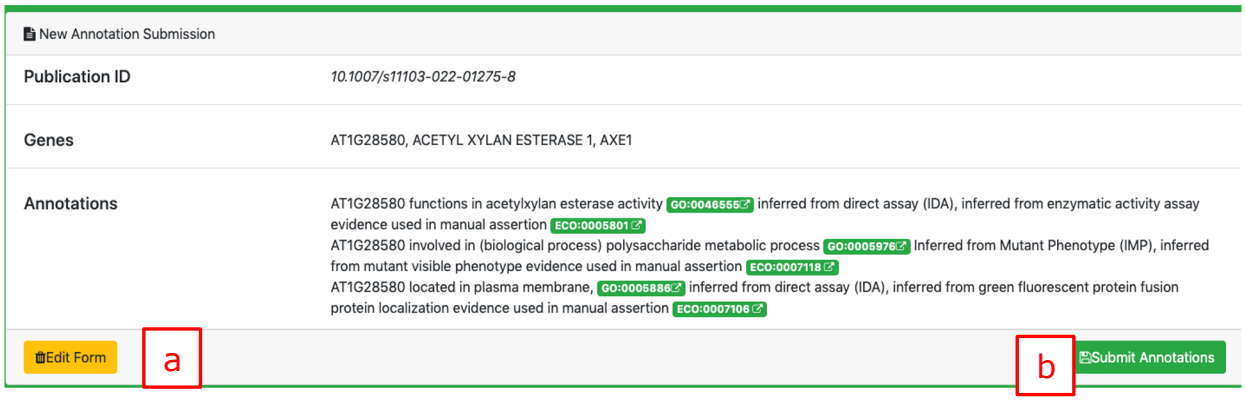
```
oligoMer
Absolute number of this oligoMer in query set
Absolute number in genomic set
Number of sequences in query set containing oligoMer
Number of sequences (out of 34187 in genomic set) containing oligoMer
p-value from binomial distribution
Query sequences containing this oligoMer
```

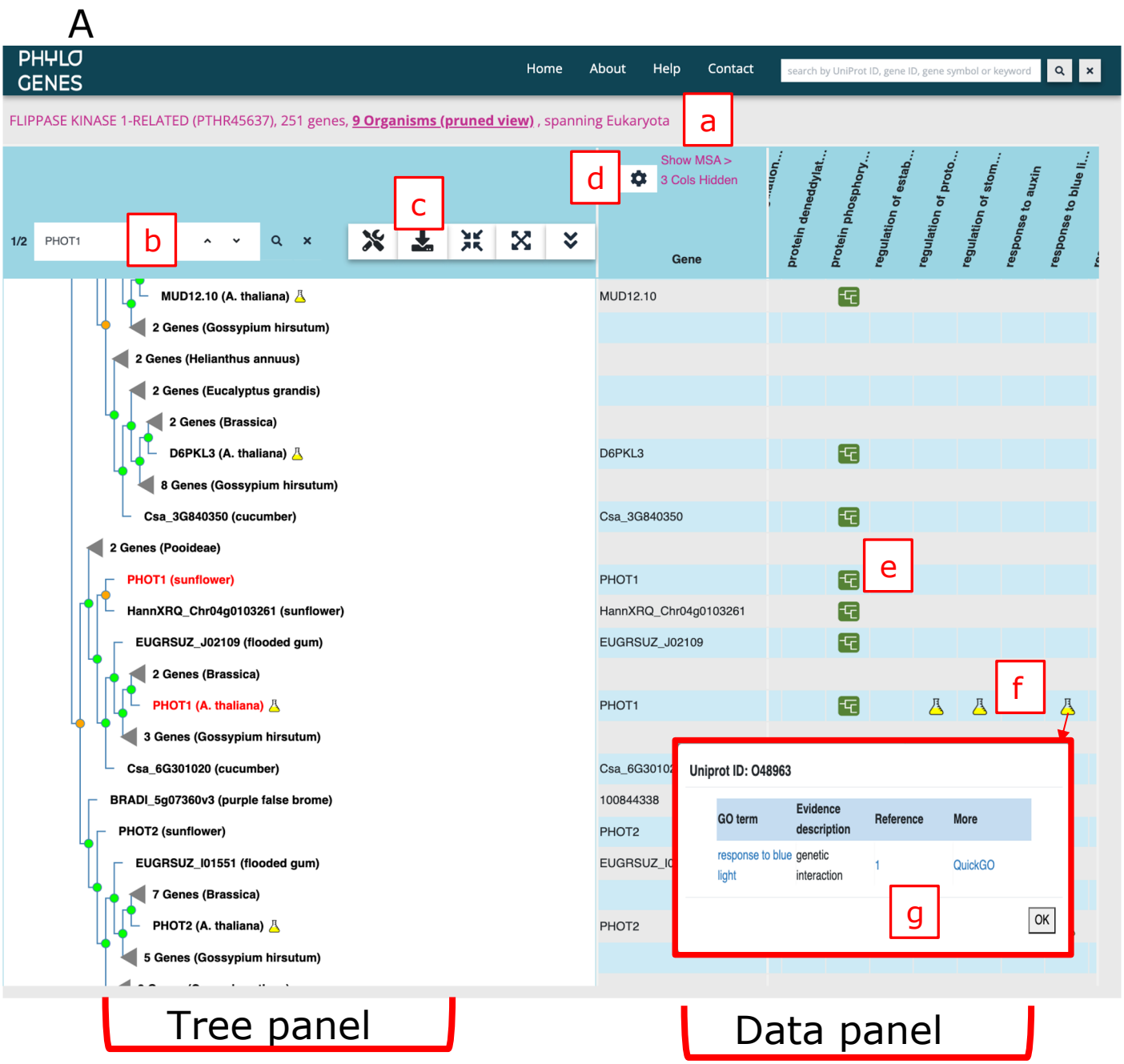
a	b	c	d	e	f	g
AGGCCC	9	4917	8/15	3948/34187	8.62e-05	AT3G46230 AT5G12020 AT4G10250 AT5G12030AT2G26150 AT1G59860 AT3G54150 AT1G53540
GGGCCT	9	4917	8/15	3948/34187	8.62e-05	AT3G46230 AT5G12020 AT4G10250 AT5G12030AT2G26150 AT1G59860 AT3G54150 AT1G53540
CCAAGA	11	8014	10/15	6906/34187	1.10e-04	AT1G69930 AT3G46230 AT5G12030 AT1G69920AT5G52760 AT1G59860 AT2G28210 AT3G02840AT3G54150 AT1G53540
TCTTGG	11	8014	10/15	6906/34187	1.10e-04	AT1G69930 AT3G46230 AT5G12030 AT1G69920AT5G52760 AT1G59860 AT2G28210 AT3G02840AT3G54150 AT1G53540
GGCCCA	13	8386	9/15	5644/34187	1.54e-04	AT3G46230 AT5G12020 AT4G10250 AT5G12030AT2G26150 AT1G59860 AT3G54150 AT1G53540AT5G42380
TGGGCC	13	8386	9/15	5644/34187	1.54e-	AT3G46230 AT5G12020 AT4G10250 AT5G12030AT2G26150

A

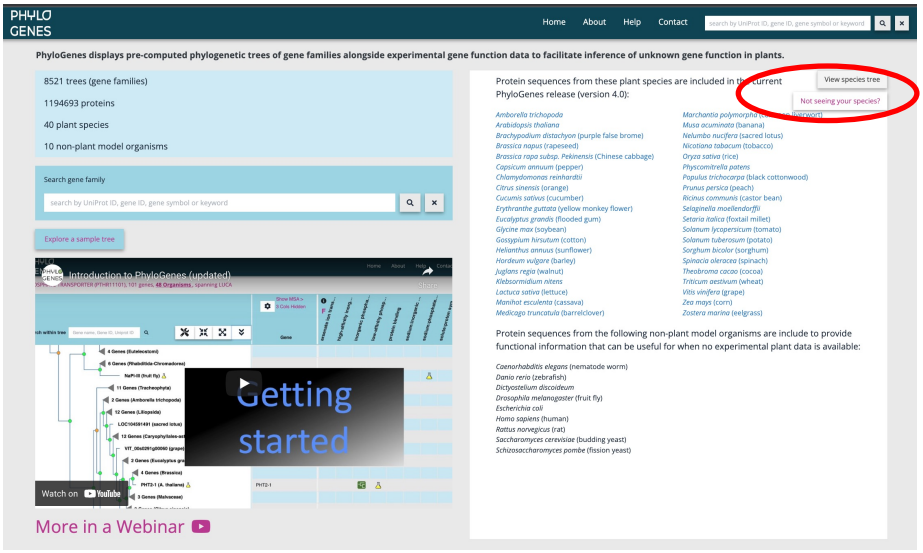

Download Data
Admin
Curation
Submission
New Annotation Submission
1. Publication
Enter a PubMed ID or a DOI.
a
Publication ID
10.1007/s11103-022-01275-8
URL
https://link.springer.com/10.1007/s11103-022-01275-8
2. Genes
Enter genes with a UniProt ID, AGI locus ID, or RNA Central ID. Optionally enter a gene symbol and full name.
Gene 1
AT1G28580
Gene Symbol
AXE1
Full Gene Symbol
ACETYL XYLAN ESTERASE 1
b
+ Add Another Gene
3. Annotations
Select an annotation format and a gene. All fields are required.
Annotation 1
Molecular Function (GO Function)
c
Gene
AT1G28580
Molecular Function (GO Function)
acetylxylan esterase activity
Method
enzymatic activity assay evidence used in manual assertion
Annotation 2
Biological Process (GO Process)
Gene
AT1G28580
Biological Process (GO Process)
polysaccharide metabolic process
Method
mutant visible phenotype evidence used in manual assertion
Annotation 3
Subcellular Location (GO Component)
Gene
AT1G28580
Subcellular Location (GO Component)
plasma membrane
Method
green fluorescent protein fusion protein localization evidence used in manual assertion
d
+Add Another Annotation
Reset Form
e
Review Submission

B

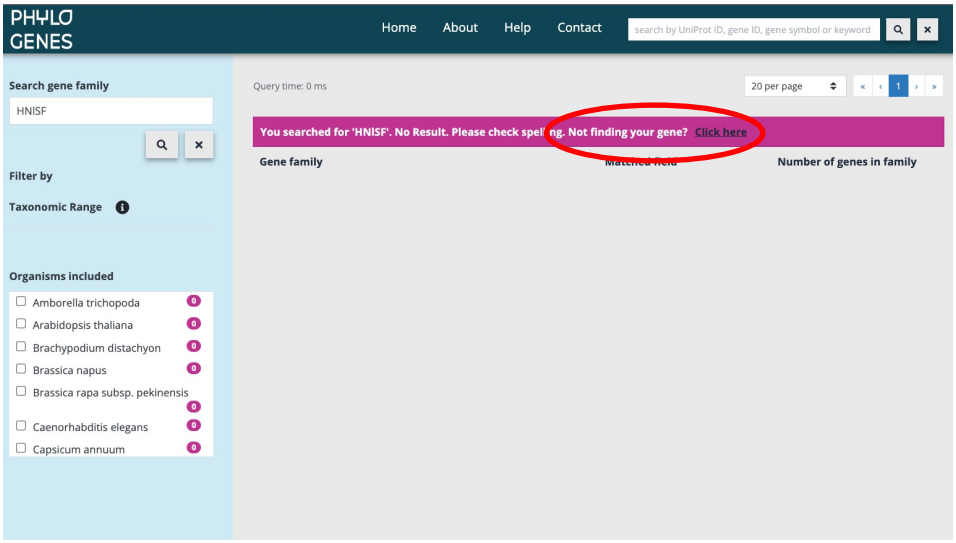

New Annotation Submission
Publication ID
10.1007/s11103-022-01275-8
Genes
AT1G28580, ACETYL XYLAN ESTERASE 1, AXE1
Annotations
AT1G28580 functions in acetylxylan esterase activity GO:0046555 inferred from direct assay (IDA), inferred from enzymatic activity assay evidence used in manual assertion ECO:0005801
AT1G28580 involved in (biological process) polysaccharide metabolic process GO:0005976 Inferred from Mutant Phenotype (IMP), inferred from mutant visible phenotype evidence used in manual assertion ECO:0007118
AT1G28580 located in plasma membrane, GO:0005886 inferred from direct assay (IDA), inferred from green fluorescent protein fusion protein localization evidence used in manual assertion ECO:0007106
Edit Form
a
b
Submit Annotations


A
PHYLO GENES
Home
About
Help
Contact
search by UniProt ID, gene ID, gene symbol or keyword
FLIPPASE KINASE 1-RELATED (PTHR45637), 251 genes, 9 Organisms (pruned view) , spanning Eukaryota
a
b
c
d
e
f
g
Show MSA >
3 Cols Hidden
1/2
PHOT1
Gene
protein deneddylat...
protein phosphory...
regulation of estab...
regulation of proto...
regulation of stom...
response to auxin
response to blue li...
MUD12.10 (A. thaliana)
2 Genes (Gossypium hirsutum)
2 Genes (Helianthus annuus)
2 Genes (Eucalyptus grandis)
2 Genes (Brassica)
D6PKL3 (A. thaliana)
8 Genes (Gossypium hirsutum)
Csa_3G840350 (cucumber)
2 Genes (Pooideae)
PHOT1 (sunflower)
HannXRQ_Chr04g0103261 (sunflower)
EUGRSUZ_J02109 (flooded gum)
2 Genes (Brassica)
PHOT1 (A. thaliana)
3 Genes (Gossypium hirsutum)
Csa_6G301020 (cucumber)
BRADI_5g07360v3 (purple false brome)
PHOT2 (sunflower)
EUGRSUZ_I01551 (flooded gum)
7 Genes (Brassica)
PHOT2 (A. thaliana)
5 Genes (Gossypium hirsutum)
MUD12.10
D6PKL3
Csa_3G840350
PHOT1
HannXRQ_Chr04g0103261
EUGRSUZ_J02109
PHOT1
Csa_6G30102
100844338
PHOT2
EUGRSUZ_IC
PHOT2
Uniprot ID: O48963
GO term
Evidence description
Reference
More
response to blue light
genetic interaction
1
QuickGO
OK
Tree panel
Data panel
B
Organisms (uncheck an organism to remove from tree)
Organism
Number of genes
Amborella trichopoda (A. trichopoda) 13
Arabidopsis thaliana (A. thaliana) 23
Brachypodium distachyon (purple false brome) 21
Brassica napus (rapeseed) 35
Brassica rapa subsp. pekinensis (Chinese cabbage) 32
Capsicum annuum (pepper) 20
Chlamydomonas reinhardtii (C. reinhardtii) 1
Citrus sinensis (orange) 12
Cucumis sativus (cucumber) 16
Erythranthe guttata (yellow monkeyflower) 22
Eucalyptus grandis (flooded gum) 17
Glycine max (soybean) 43
Gossypium hirsutum (cotton) 51
Helianthus annuus (sunflower) 35
Update tree
Close
C
Customize gene info table (3 Cols Hidden)
Organism
Molecular function
blue light photoreceptor activity
DNA-directed 5'-3' RNA polymerase activity
FMN binding
identical protein binding
metallopeptidase activity
NEDD8-specific protease activity
Update Table
Close

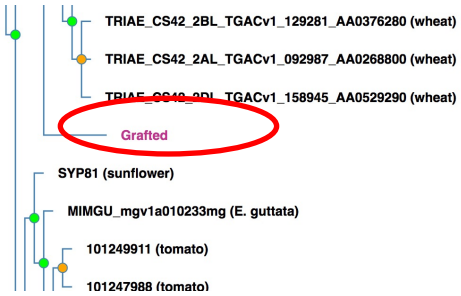
A


PHYLO GENES
Home
About
Help
Contact
PhyloGenes displays pre-computed phylogenetic trees of gene families alongside experimental gene function data to facilitate inference of unknown gene function in plants.
8521 trees (gene families)
1194693 proteins
40 plant species
10 non-plant model organisms
Search gene family
search by UniProt ID, gene ID, gene symbol or keyword
Explore a sample tree
Introduction to PhyloGenes (updated)
Getting started
Watch on YouTube
More in a Webinar
Protein sequences from these plant species are included in the current PhyloGenes release (version 4.0):
View species tree
Not seeing your species?
Amborella trichopoda
Arabidopsis thaliana
Brachypodium distachyon (purple false brome)
Brassica napus (rapeseed)
Brassica rapa subsp. Pekinensis (Chinese cabbage)
Capsicum annuum (pepper)
Chlamydomonas reinhardtii
Citrus sinensis (orange)
Cucumis sativus (cucumber)
Erythranthe guttata (yellow monkey flower)
Eucalyptus grandis (flooded gum)
Glycine max (soybean)
Gossypium hirsutum (cotton)
Helianthus annuus (sunflower)
Hordeum vulgare (barley)
Juglans regia (walnut)
Klebsormidium nitens
Lactuca sativa (lettuce)
Manihot esculenta (cassava)
Medicago truncatula (barrelclover)
Marchantia polymorpha
Musa acuminata (banana)
Nelumbo nucifera (sacred lotus)
Nicotiana tabacum (tobacco)
Oryza sativa (rice)
Physcomitrella patens
Populus trichocarpa (black cottonwood)
Prunus persica (peach)
Ricinus communis (castor bean)
Selaginella moellendorffii
Setaria italica (foxtail millet)
Solanum lycopersicum (tomato)
Solanum tuberosum (potato)
Sorghum bicolor (sorghum)
Spinacia oleracea (spinach)
Theobroma cacao (cocoa)
Triticum aestivum (wheat)
Vitis vinifera (grape)
Zea mays (corn)
Zostera marina (eelgrass)
Protein sequences from the following non-plant model organisms are include to provide functional information that can be useful for when no experimental plant data is available:
Caenorhabditis elegans (nematode worm)
Danio rerio (zebrafish)
Dictyostelium discoideum
Drosophila melanogaster (fruit fly)
Escherichia coli
Homo sapiens (human)
Rattus norvegicus (rat)
Saccharomyces cerevisiae (budding yeast)
Schizosaccharomyces pombe (fission yeast)

B


PHYLO GENES
Home
About
Help
Contact
search by UniProt ID, gene ID, gene symbol or keyword
Search gene family
HNISF
Filter by
Taxonomic Range
Organisms included
Amborella trichopoda
Arabidopsis thaliana
Brachypodium distachyon
Brassica napus
Brassica rapa subsp. pekinensis
Caenorhabditis elegans
Capsicum annuum
Query time: 0 ms
20 per page
You searched for 'HNISF'. No Result. Please check spelling. Not finding your gene? Click here
Gene family
Number of genes in family

C


TRIAE_CS42_2BL_TGACv1_129281_AA0376280 (wheat)
TRIAE_CS42_2AL_TGACv1_092987_AA0268800 (wheat)
Grafted
SYP81 (sunflower)
MIMGU_mgv1a010233mg (E. guttata)
101249911 (tomato)
101247988 (tomato)

Query Gene/Region	Brassica rapa Syntelog(s)	GeVo Link
AT1G01010.1 a	Brara.I05584.1.v1.3 Brara.J00084.1.v1.3 b	GeVo c

d

High score Segment Pairs (HSPs) indicating regions of high sequence similarity are drawn as colored rectagles above each gene model. Click
syntenic pairs. Press Shift amd click to connect HSPs of all syntenic pairs between two tracks. If you have questions about the results, check o

c