ThaleMine: A Warehouse for Arabidopsis Data Integration and Discovery

Vivek Krishnakumar [1], Sergio Contrino [2], Chia-Yi Cheng [1], Irina Belyaeva [1], Erik S. Ferlanti [3], Jason R. Miller[1], Matthew W. Vaughn [3], Gos Micklem [2], Christopher D. Town [1], Agnes P. Chan [1]*

[1] Plant Genomics, J. Craig Venter Institute, 9714 Medical Center Dr, Rockville, MD 20850

[2] Department of Genetics, Cambridge Systems Biology Centre, Tennis Court Road, Cambridge, CB2 1QR

[3] Life Sciences Computing, Texas Advanced Computing Center, 10100 Burnet Rd, Austin, TX 78758

* Corresponding author

Abbreviations:

ABRC, Arabidopsis Biological Resource Center;

AGI, Arabidopsis Genome Initiative;

API, Application Programming Interface;

Araport, Arabidopsis Information Portal;

BAR, Bio Analytic Resource;

BED, Browser Extensible Data;

CDN, Content Delivery Network;

FAIR, Findability, Accessibility, Interoperability, and Reusability;

GeneRIF, Gene Reference Into Function;

GFF3, Generic Feature Format version 3;

GO, Gene Ontology;

HTTP, HyperText Transfer Protocol;

JCVI, J. Craig Venter Institute;

JSON, JavaScript Object Notation;

KEGG, Kyoto Encyclopedia of Genes and Genomes;

modENCODE, Model Organism ENCyclopedia Of DNA Elements;

NASC, Nottingham Arabidopsis Stock Center;

NCBI, National Center for Biotechnology Information;

PANTHER, Protein ANalysis THrough Evolutionary Relationships;

PO, Plant Ontology;

REST, REpresentational State Transfer;

TACC, Texas Advanced Computing Center;

TAIR, The Arabidopsis Information Resource;

TSV, Tab-Separated Values;

VM, Virtual Machine;

XML, eXtensible Markup Language

Abstract

ThaleMine (https://apps.araport.org/thalemine/) is a comprehensive data warehouse that integrates a wide array of genomic information of the model plant *Arabidopsis thaliana*. The data collection currently includes the latest structural and functional annotation from the Araport11 update, the Col-0 genome sequence, RNA-seq and array expression, co-expression, protein interactions, homologs, pathways, publications, alleles, germplasm and phenotypes. The data are collected from a wide variety of public resources. Users can browse gene-specific data through gene report pages, identify and create gene lists based on experiments or indexed keywords, and run GO enrichment analysis to investigate the biological significance of selected gene sets. Developed by the Arabidopsis Information Portal (Araport, https://www.araport.org/) project, ThaleMine uses the InterMine software framework, which builds well-structured data, and provides powerful data query and analysis functionality. The warehoused data can be accessed by users via graphical interfaces, as well as programmatically via web-services. Here we describe recent developments in ThaleMine including new features and extensions, and discuss future improvements. InterMine has been broadly adopted by the model organism research community including nematode, rat, mouse, zebrafish, budding yeast, the modENCODE project, as well as being used for human data. ThaleMine is the first InterMine developed for a plant model. As additional new plant InterMines are developed by the legume and other plant research communities, the potential of cross-organism integrative data analysis will be further enabled.

Keywords

Arabidopsis thaliana, InterMine, genomics, data warehouse, data integration, web services

Introduction

Arabidopsis thaliana (thale cress) has served as a model organism for understanding the complex processes required for plant growth and development. It was the first plant species to have its genome sequenced (Arabidopsis Genome Initiative 2000). The high quality Arabidopsis genome sequence and gene annotation have since served as a benchmark for other plant genome projects. For over a decade, the Arabidopsis Information Resource (TAIR) served as the model organism database for Arabidopsis, responsible for updating and hosting the structural and functional annotation of the Arabidopsis genome (Reiser *et al.* 2016). Additional online databases and resources such as NCBI, UniProt, and Ensembl are also key to Arabidopsis and plant researchers. Nevertheless, with the exponential increase in data volume in the present omics generation, a common issue facing researchers is the need to find and aggregate heterogeneous data types across multiple domain specific databases.

The Arabidopsis Information Portal (Araport, https://www.araport.org) uses advanced web technologies to present a new and modern online resource for data integration and interoperability across multiple data sources (Krishnakumar *et al.* 2015a). ThaleMine is part of the Araport project and its focus is on the integration of genomic data released by major centers into a single portal, with the goal of enhancing data accessibility and mining functionality (https://apps.araport.org/thalemine). ThaleMine is powered by InterMine, an open source data warehouse developed specifically for the integration and analysis of complex biological data (Smith *et al.* 2012). The InterMine framework builds structured data and provides powerful and flexible data query and analysis functions. The back end of InterMine is a relational database with a custom object-relational interface optimized for read-only performance. It provides high performance on large data sets without imposing constraints on the relational schema, by pre-computing table joins. Data within InterMine are structured on the basis of InterMine's core data model, developed based on ontologies such as Sequence Ontology (Eilbeck *et al.* 2005) to standardize biological features and relationships. InterMine data can be accessed via tabular and graphical user interfaces for interactive browsing. In addition, InterMine allows retrieval and query of stored data in the form of interactive data tables as well as analysis widgets that operate on lists of entities. Finally InterMine also provides an extensive set of RESTful (REpresentational State Transfer) web services to facilitate programmatic access and cross-resource sharing (Kalderimis *et al.* 2014)

The first InterMine instance was FlyMine, developed to support *Drosophila* and *Anopheles* genomics (Lyne *et al.* 2007), followed by other InterMine instances developed in collaboration with many of the major Model Organism Databases (YeastMine (SGD); RatMine (RGD); MouseMine (MGI); WormMine (WormBase); ZFINMine (ZFIN)) (Balakrishnan *et al.* 2012; Motenko *et al.* 2015; Ruzicka *et al.* 2015; Shimoyama *et al.* 2015; Howe *et al.* 2016). In addition, InterMine was adopted by the model organism ENCODE (modENCODE) project so producing modMine (Contrino *et al.* 2012). HumanMine, dedicated to human data, is also now available (Lyne *et al.* 2015). Among the Viridiplantae, several new InterMine instances for leguminous plants species have been implemented, which include *Medicago truncatula* (MedicMine, http://medicmine.jcvi.org) (Krishnakumar *et al.* 2015b) and the Legume Federation initiative (e.g. BeanMine, SoyMine, and PeanutMine; http://mines.legumeinfo.org). Additionally, the Phytozome plant comparative genomics portal at the Joint Genome Institute has developed a multi-species InterMine implementation called PhytoMine (Goodstein *et al.* 2012).

We describe below an expansive collection of biological data integrated and represented in the ThaleMine data warehouse, and the core functionalities for users to search, visualize, and analyze the broad range of Arabidopsis genome-centric data.

Data integration in ThaleMine

The ThaleMine and InterMine data model

InterMine uses an object-oriented data model where each element is represented by a class, described by attributes, and contains references to other classes. The data model is defined in an eXtensible

Markup Language (XML) file which is used to generate Java code for the various classes as well as the database and user interface. The generated Java classes automatically map to tables within the database schema. The terms and relationships used in the core InterMine data model adhere to the well-established Sequence Ontology, used to describe various biological sequence features. Figure 1A illustrates the representation of the core data model used to describe Gene entities. To enable integration with new data types, InterMine allows the addition of new classes and attributes, and logical extensions to the existing data model by re-using classes and adding additional attributes. Figure 1B illustrates one such case in ThaleMine, developed in order to accommodate the gene locus history. The ThaleMine data model is generated by merging the core InterMine data model with new models developed for Arabidopsis (i.e. gene locus history, GeneRIFs, array expression, RNA-seq expression, and stocks).

Gene models and functions

ThaleMine hosts gene structure and functional annotations for the Arabidopsis Col-0 reference genome. The Araport team has performed a comprehensive genome-wide annotation update referred to as Araport11 (Cheng et al, in press), which is available for download from both Araport and GenBank (accessions: CP002684-CP002688). The Araport11 annotation update pipeline utilized over 100 public RNA-seq datasets across 11 tissues/organs from NCBI Sequence Read Archive (SRA) to revise gene structures and add novel alternative splicing variants. Genome annotation was also updated for many non-coding gene subclasses and further expanded to long non-coding RNAs (lncRNAs), natural antisense RNAs (NATs), upstream open reading frames (uORFs), and novel transcribed regions. In ThaleMine, the Araport11 annotation is supplemented with accessory descriptors including gene symbols and synonyms, curator summaries and computational descriptions obtained from the latest public TAIR data release, albeit with a one-year delay due to the TAIR subscription model (Reiser *et al.* 2016).

On the ThaleMine Gene Report page, splicing variants and exon-intron structures are rendered within an embedded JBrowse displayer, with the underlying data retrieved from ThaleMine via built-in JBrowse-compatible REST APIs. To track the modification history of a particular gene locus for splits or merges, we have extended the ThaleMine data model to support the storage and retrieval of gene locus history information (see Figure 1B). The latest version of Gene Ontology (GO) and Plant Ontology (PO) annotation assignments from the Gene Ontology Consortium (GOC) and TAIR respectively are also available for each gene locus (Berardini *et al.* 2015; Gene Ontology Consortium 2015). GO term enrichment is one of the several analysis tools built into InterMine, and displays the GO terms which are statistically over-represented in a given list of genes.

Transcriptome profiles

Comprehensive Affymetrix microarray expression levels from over 80 Arabidopsis studies collected from multiple research groups were obtained from Bio-Analytic Resource (BAR) at the University of Toronto, Canada (Winter *et al.* 2007) . Additionally, we have extended the ThaleMine data model to include the expression profiles generated from RNA-seq datasets used in the Araport11 annotation. The expression profiles can be visualized on the Gene Report page using an embeddable heatmap displayer that we developed for this purpose. The displayer is an independent piece of software (JavaScript) made available to ThaleMine via its own Araport-hosted version of the InterMine Content Delivery Network (CDN). The displayer requests RNA-seq expression data via the ThaleMine application web services. The displayer widget was built using the D3.js library to ensure that the visualization tool is portable and can be adapted into other contexts (Bostock, Ogievetsky and Heer 2011).

For real-time data integration, we have implemented proof-of-concept data integration of several remotely located data sources. For example, electronic pictograph images that display expression levels across developmental stages and treatment conditions for a gene of interest are retrieved through web services provided by the BAR web site (Winter *et al.* 2007). Similarly, co-expression gene lists for a gene of interest are retrieved dynamically through the web services provided by the ATTED-II site at Tohoku University, Japan. (Aoki *et al.* 2016).

Proteins and domains

For the annotated proteins in Araport11, we have established links to the existing UniProt protein entries based on sequence identity. This information is used to cross-reference the protein information from Araport11 and UniProt (UniProt Consortium 2014) within Araport and support querying either by the UniProt (e.g. DCL1_ARATH or Q9SP32) or the customary AGI identifiers (e.g. AT1G01040.1) that are now maintained and curated by Araport. Accessory protein descriptors such as curator comments (e.g. function, subcellular location, developmental expression, etc.) and domain compositions were obtained from UniProt and InterPro respectively. We have developed an embeddable protein domain displayer using the D3.js library and ThaleMine web services, for ThaleMine Protein Report pages. The displayer provides a color-coded visualization of protein domain locations along the length of the protein.

Protein Interactions and pathways

ThaleMine incorporates Arabidopsis protein-protein interactions from the IntAct database at the European Bioinformatics Institute, UK (EBI) (Orchard *et al.* 2013), and the BioGrid database in the US/Canada (Chatr-Aryamontri *et al.* 2015), and uses a recently updated interaction displayer from InterMine to show proteins as nodes and the types of interactions (i.e. genetic or physical) as edges. The assignments of Arabidopsis genes to metabolic pathways follows the Arabidopsis catalog of the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway database (Kanehisa *et al.* 2016).

Publications and GeneRIFs

Scientific publications that describe Arabidopsis gene and proteins were collected from NCBI PubMed and UniProt. ThaleMine consolidates publications from both data sources to provide a comprehensive view of literature associated with any particular *Arabidopsis* gene locus (Maglott *et al.* 2011; UniProt Consortium 2015). GeneRIFs, a project initiated at NCBI, refers to *Gene Reference into Function* and collects community-contributed data. Each GeneRIF is comprised of a gene, a short description of a gene function, and a publication that supports the described gene function. Arabidopsis GeneRIFs were obtained from NCBI and integrated into ThaleMine Gene Report pages. We have also implemented functionality on the ThaleMine Gene Report page for users to directly submit new GeneRIFs to NCBI.

Orthologs and paralogs

For a given Arabidopsis gene of interest, human and yeast orthologs, as well as Arabidopsis paralogs, are sourced from the PANTHER database from the University of Southern California (Mi *et al.* 2016) and integrated into ThaleMine. Within the context of the ThaleMine Gene Report page, we have also developed a tool based on the InterMine interactive data table library (im-tables) to retrieve and integrate in real-time orthologous and paralogous relationships inferred from the JGI Phytozome gene families. ThaleMine retrieves the Phytozome gene family information across sequenced plant genomes provided by the PhytoMine web services (Goodstein *et al.* 2012).

Several InterMine instances have recently been developed for legume species including MedicMine (Krishnakumar *et al.* 2015b), BeanMine (http://mines.legumeinfo.org/beanmine), SoyMine (http://mines.legumeinfo.org/soymine), and PeanutMine (http://mines.legumeinfo.org/peanutmine), which store legume gene families (including Arabidopsis as outgroup) and their homology, as computed by the Legume Information Systems (LIS) (Dash *et al.* 2016). We establish links from ThaleMine to the legume InterMine instances using the core InterMine functionality, FriendlyMines (FriendlyMineManager (intermine-javadoc)), which can provide built-in interconnectivity across a pair of InterMine instances on the basis of homology or similar types of biological relationships (e.g. synteny). The tool requires the relationship information to be stored in only one of the InterMine pairs (e.g. only SoyMine stores the soybean vs. Arabidopsis homology relationships).

Seed stocks and germplasms

The Arabidopsis research community has collectively produced a wide array of genetic resources (such as knock-downs, insertion lines, point mutations), and donated them to stock centers (Sessions *et al.* 2002; Alonso *et al.* 2003; Rosso *et al.* 2003; Woody *et al.* 2007). We collected Arabidopsis stock data from TAIR, the Arabidopsis Biological Resource Center (ABRC) in the US (http://abrc.osu.edu/), and the Nottingham Arabidopsis Stock Center (NASC) in the UK (http://arabidopsis.info/). We have extended the core InterMine data model to accommodate the stock data, with the catalogs of mutant alleles, germplasm, genotypes, phenotypes and seed availabilities linked to a given gene and presented on the Gene Report page.

ThaleMine core functions

Data tables: Interactive and web-services-enabled

ThaleMine data are presented via data tables and/or visualization widgets. The data tables are based on the InterMine interactive data table library (im-tables) and provide dynamic functions for spreadsheet-type operations including sorting, filtering, hiding, and summarizing (e.g. providing simple statistics such as row counts for distinct attributes) of individual data columns. The data tables can be exported in common formats such as tab/comma-delimited values, GFF3, BED, JSON, XML, for downstream processing. Nucleotide and protein sequence data are also retrievable in FASTA format. Each data table is also associated with a code-generation function to automatically build a code snippet in several supported languages (Perl, Python, Java and JavaScript) for programmatic access via ThaleMine web services.

Data queries: pre-defined, user-defined and keyword search

Data hosted in the ThaleMine warehouse are fully accessible via the flexible data query system of InterMine. A pre-defined set of data queries available as simple web forms covering most common data questions can be found under the “Templates” tab. For example, the “Gene -> RNA-seq expression” query allows users to retrieve expression values across 113 RNA-seq datasets and tissue types for one or more genes or interest. The “Gene -> Interacting Genes” query retrieves interacting partners for one or more genes of interest. Users can use the predefined data queries as the basis to further customize data attributes or constraints, using the intuitive point-and-click graphical interface provided. For more advanced data mining and exploration of the warehouse, the “Query Builder” function tab provides a hierarchical display across all data objects stored in the warehouse. It allows building complex data queries *de novo*, and returns data in the form of interactive data tables through which the original query can be iteratively modified. Data queries can also be saved into a user's account and then become available under the "MyMine" tab as well as shared with other users.

In addition, InterMine provides a simple keyword search based on a Lucene index (http://lucene.apache.org), available through the upper right search box on all pages. The results of the Lucene-index-based keyword searches are categorized and allow easy navigation to the relevant ThaleMine page.

List functions: Enrichment tests and list operations

A common usage of the Lists function is gene list enrichment analyses. For example, a list of differentially expressed genes was obtained from an expression study and one would like to run a quick check on any candidate pathways or biological functions involved. Under the "Lists" tab, users can input a list of Arabidopsis gene identifiers using AGI locus ID and/or gene symbols (e.g. At3G24650 or RBL6). Alternatively, users can also create lists from specific columns of the InterMine data tables (see Use cases section below). For all identifiers that match gene entities in ThaleMine, a gene list will be created in ThaleMine. A set of basic analyses will be automatically performed, including a chromosomal distribution plot, an RNA-seq expression heatmap, and statistical tests for the enrichment of Gene Ontology terms or KEGG pathways. User-created lists can be saved into a user's account which will allow reuse of the lists by the user, sharing with other users, and performing set operations such as intersection, union, etc. These operations allow users to refine a list of genes of interest by intersection with another of their lists or with one of a number of provided reference lists such as known transcription factors.

Genomic Intervals: Feature extraction

ThaleMine also provides easy access to extracting features from the Arabidopsis Col-0 genome reference sequence using the "Regions" tab. The extracted features can be exported in a variety of data formats including tab/comma-delimited values, GFF3, BED and FASTA. For example, this powerful tool allows users to obtain a list of gene identifiers residing in a selected genomic region, retrieve the 1 kb upstream and downstream sequences for a group of genes in FASTA format, or extract all the T-DNA insertion sites reported by TDNA-seq in a given region (O'Malley *et al.* unpublished).

ThaleMine statistics

ThaleMine includes a wide variety of large-scale functional genomic datasets for the model organism *Arabidopsis thaliana*. The datasets used for building the warehouse are regularly updated through ThaleMine data releases. The "Data Sources" function tab provides an up-to-date summary of all data sources used to build the ThaleMine warehouse (https://www.araport.org/thalemine/data-statistics). It shows a dynamic catalog of the various data sources and datasets, the number of associated genes and features for each data type, a brief description of the data, a version number or date stamp, as well as literature references. The major data types are also summarized in Table 1.

Since the inception of the Araport project in 2013, 10 major ThaleMine releases (from 1.0.0 to the present release 1.10.0 as of Aug 2016) have been generated and made available to community. The size of the database for the present release is around 120 GB and the updates for each release are documented in the Release Notes section of the Araport website (https://www.araport.org/release-notes).

ThaleMine is deployed onto a CentOS 6.5 based virtual machine (VM), configured under an Apache/Tomcat web server and uses PostgreSQL as the underlying database management system. All pages are encrypted and served via secure HTTP (i.e. HTTPS). To handle user logins, ThaleMine uses federated authentication provided by Araport, which was built using an industry-standard technology called OAuth2 to enable single-sign-on across all cooperating web sites.

ThaleMine use cases

Below are examples of ThaleMine data search queries and the information retrieved from the current ThaleMine release v1.10.0 of Aug 2016. ThaleMine functions are freely accessible and do not require user login. Note that the advantage of performing ThaleMine analysis with user login is that it will allow gene list and data queries to be automatically stored in the user account for subsequent reuse or sharing with other ThaleMine users.

1) Look up Gene Report for a gene of interest (Figure 2).
   a) On the ThaleMine home page, https://apps.araport.org/thalemine/, enter a gene symbol or AGI locus identifier in the "Search" box (e.g. ABI3)
   b) On the search result page, select the "AT3G24650 | ABI3" link under the "Gene" category.
   c) On the "Gene Report" page, you can browse or download most of the following data features as flat files for downstream analysis, some of which are described below (see blue bounded boxes in Fig 2).
      i) At the top of the report page, the "Summary" section offers Gene functional annotation information such as "Brief Description", "Curator Summary".
      ii) Under the "Genomics" section, information pertaining to the gene structure and sequences of alternative splicing isoforms, upstream/downstream sequences, T-DNA insertion sites reported by TDNA-seq are made available (O'Malley *et al.* unpublished).
      iii) Under the "Expression" section, data pertaining to array expression in log fold-change values, RNA-seq expression in TPM values computed using Salmon (Patro, Duggal and Kingsford 2015), and co-expressed gene lists can be perused.
      iv) Under the "Interactions" section, a list of interacting partners are visualized as an interactive graph.
      v) Under the "Links to other Mines" section, contextual links to entities in remote InterMine instances (human, yeast, and legumes) are made available, based on homology information from Panther, and a collection of legume InterMine warehouses.

vi) Under the the “Stocks” section, a list of genetic variants and seed stock availability are displayed.

2) Run functional enrichment analysis for a gene list of interest (Figure 3).

a) On the ThaleMine home page, https://apps.araport.org/thalemine/, select the “Lists” tab in the top menu bar.

b) Paste in a list of AGI locus identifiers or use the example provided to create a gene list in ThaleMine (refer to top panel of Fig 3). Name the gene list as “Gene List - Demo - 2016”. Click on “Save a list” to create the list. Without user login, the gene list will be accessible under the functional tab “MyMine” only during your current working session. With user login, the gene list will be stored in your user account indefinitely and is available for reuse and sharing with other users.

c) A “List Analysis” page will be displayed. You can browse or download statistical test results from enrichment widgets for GO terms, protein domains, and other features of the gene list. Under the “Function” section, you can expand the GO terms and pathway data tables (by clicking on the green arrow icon of the section header) for interactive browsing or download. Under the “Expression” section, you can also browse the gene expression heatmap, or expand the array and RNA-seq expression data tables (by clicking on the green arrow icon of the section header) for interactive browsing or download.

3) Identify interacting protein partners for a gene list, and filter by RNA-seq expression levels (Figure 4).

a) To collect interacting protein partners for the gene list:

i) Select the “Templates” tab on the top menu bar

ii) Select the predefined query Gene → Interacting Genes (Note: Type a keyword “interact” in the “Filter” box to quickly narrow down to the group of relevant templates)

iii) Check the box “constrain to be IN”

iv) Select from the drop down menu “Gene List - Demo - 2016”

   v) A data table will be returned showing interacting partners for the input gene list.

   vi) Save the list of interacting genes by clicking on "Save as List", i.e. "Interaction > Participant2". Name the new gene list as "Interacting Gene List - Demo - 2016".

   vii) Your new gene list is now stored and accessible via the "List" tab > View.

b) To filter gene list by RNA-seq expression values

   i) Select the "Templates" tab on the top menu bar

   ii) Select the predefined query Gene → RNA-seq Expression (Note: Type a keyword "RNA" in the "Filter" box to quickly narrow down to the group of relevant templates)

   iii) Check the box "constrain to be IN"

   iv) Select from the drop down menu "Interacting Gene List - Demo - 2016"

   v) A data table will be returned showing RNA-seq expression levels for the input gene list.

   vi) You can filter by tissue types and the range of expression levels via the column header "Filter" icon and "View column summary" icon.

   vii) To download the gene expression data table for downstream analysis, use the "Export" button on the right. You can preview the output format via the "Preview" function on the left menu.

4) Run a data query to collect genes related to a keyword and export their translated sequences in FASTA format (Figure 5).

   a) Go to the ThaleMine home page, https://apps.araport.org/thalemine/

   b) Select the "Templates" tab on the top menu bar.

   c) Select the predefined query "Gene → Protein sequence"

   d) In the LOOKUP search box, enter the keyword "flower" flanked by asterisks (*) on both sides as wild cards like so "*flower*". The lookup function will search across multiple functional attributes (e.g. gene names, curator summary, etc.) to identify matches for the provided keyword.

e) In the returned data table showing 122 rows, locate the “View column summary” icon available in each column header. Click on the icon to run a column summary for the “Protein” column and the “Genes” column. It will show that the data table contains 122 rows of protein isoforms originating from 54 distinct gene loci.

f) To download the protein sequences for downstream analysis, use the “Export” button on the right and select the “FASTA” option from the drop-down menu.

5) Collect Panther homolog sequences for a gene of interest.

a) To collect homologs for a gene of interest (refer to Figure 4a for steps):

i) Go to the ThaleMine home page, https://apps.araport.org/thalemine/

ii) Select the “Templates” tab on the top menu bar.

iii) Select the predefined query “Gene → Homologs”.

iv) Enter an AGI locus ID “AT3G24650” (for the ABI3 gene) to search for homologs from the Panther database. (Note: ThaleMine currently provides Arabidopsis paralogs, and human and yeast orthologs from the Panther database.)

v) Click on “Show Results” to run the query. A data table with 16 rows of Panther homologs for ABI3 will be displayed.

vi) To create a list of gene identifiers for the ABI3 homologs, click on “Save as List” and select the third column, i.e. “Gene > Homologues > Homologue”. Name the new gene list as “Panther homolog test”.

b) To export translated sequences in FASTA format of homologs (refer to Figure 5 for steps):

i) Now navigate to the “Templates” tab on the top menu bar and select the “Gene -> Protein sequence” template. Check the box “constrain to be IN”, and select your new gene list “Panther homolog test” from the drop down menu. A data table with 28 rows of ABI3 protein homologs will be displayed. Use the “Export” button and select “FASTA” from the drop-down menu to download the protein sequences in FASTA format. (Note: Use the "Gene -> CDS sequence" template instead if you were looking for the nucleotide coding sequences for the genes).

Discussion/Perspectives

Our aim is to integrate and present extensive and ready-to-use datasets for biologists and bioinformaticians to analyze and interpret to advance their research. We have implemented ThaleMine, an advanced biological data warehouse for *Arabidopsis thaliana* based on the InterMine framework. We have expanded InterMine core data model to accommodate the specifics of the Arabidopsis genomic data. We have also introduced proof-of-concept data display modules to demonstrate the capability of real-time integration of remote data originating from third party web-services.

A major feature of the InterMine warehouse (including ThaleMine) is that in addition to a set of intuitive graphical user-interfaces for human interactions, data loaded in the warehouse are automatically set up for machine access through web services. Actively exposing biological data in the form of web services ensures the discoverability and accessibility of the data befitting the FAIR principles for scientific data management which emphasizes Findability, Accessibility, Interoperability, and Reusability (Wilkinson *et al.* 2016). An example of a third-party biological database making use of ThaleMine web services is a new and upgraded version of ePlant (http://bar.utoronto.ca/eplant/, (Fucile *et al.* 2011)) being developed by BAR. The new ePlant uses ThaleMine web services to dynamically retrieve and integrate the latest Araport11 gene models, their attributes, and associated sequences with other ePlant omics data types (Waese et al., personal communication). We believe biological data presented as web services will be a prerequisite to broad data sharing and efficient integration, and provides the flexibility to leverage the power of advanced web technologies.

For future ThaleMine development, content will be further enriched with large-scale Arabidopsis datasets from the research community such as population variants (e.g. 1001 Genomes Project, http://1001genomes.org/, (1001 Genomes Consortium 2016)), biochemical pathways (e.g. AraCyc of the

Plant Metabolic Network (PMN), http://www.plantcyc.org, (Chae *et al.* 2014)) and gene regulatory networks (e.g. HRGRN, http://plantgrn.noble.org/hrgrn/, (Dai *et al.* 2016)). Most importantly, dataset suggestions and active participation including feedback from the research community will play an important part in defining the future role of ThaleMine as a modern generation model organism data repository for Arabidopsis.

In addition to major animal model organisms that have adopted InterMine as a data warehouse of choice, the Legume Federation project (http://www.legumefederation.org), encompassing creators of leguminous plant databases, has recently spearheaded efforts among the plant research community and set up a collection of legume plant InterMine warehouses. Building organism-specific InterMine warehouses for individual organisms will provide a way to unify and homogenize genomic data representation across plant species. This will be one of the essential first steps towards cross-organism interoperability and comparative analysis in future development.

Funding

This work was supported by the National Science Foundation (NSF) [DBI-1262414] and the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/L027151/1].

Disclosures

Conflicts of interest: No conflicts of interest declared.

Acknowledgements

We would like to thank the InterMine team at the University of Cambridge for their valuable input and continuous support over the course of development of ThaleMine, the TACC Technology Infrastructure (TI) team for their support of the ThaleMine web and database servers, JCVI DBA Michael Heaney for assistance with database maintenance, the Nick Provart group (BAR, University of Toronto) for help with testing ThaleMine web services and providing feedback, and the InterMine developer community and the ThaleMine user community for feedback, discussion and support.

References

1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 2016;**166**:481–91.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P. *et al.* Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science* 2003;**301**:653–7.

Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K., and Obayashi, T. ATTED-II in 2016: A Plant Coexpression Database Towards Lineage-Specific Coexpression. *Plant Cell Physiol* 2016;**57**:e5.

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 2000;**408**:796–815.

Balakrishnan, R., Park, J., Karra, K., Hitz, B.C., Binkley, G., Hong, E.L. *et al.* YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database* 2012;**2012**, DOI: 10.1093/database/bar062.

Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. *et al.* The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis* 2015;**53**:474–85.

Bostock, M., Ogievetsky, V., and Heer, J. D3 Data-Driven Documents. *IEEE Trans Vis Comput Graph* 2011;**17**:2301–9.

Chae, L., Kim, T., Nilo-Poyanco, R., and Rhee, S.Y. Genomic signatures of specialized metabolism in plants. *Science* 2014;**344**:510–3.

Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 2015;**43**:D470–8.

Contrino, S., Smith, R.N., Butano, D., Carr, A., Hu, F., Lyne, R. *et al.* modMine: flexible access to modENCODE data. *Nucleic Acids Res* 2012;**40**:D1082–8.

Dai, X., Li, J., Liu, T., and Zhao, P.X. HRGRN: A Graph Search-Empowered Integrative Database of Arabidopsis Signaling Transduction, Metabolism and Gene Regulation Networks. *Plant Cell Physiol* 2016;**57**:e12.

Dash, S., Campbell, J.D., Cannon, E.K.S., Cleary, A.M., Huang, W., Kalberer, S.R. *et al.* Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family. *Nucleic Acids Res* 2016;**44**:D1181–8.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol* 2005;**6**:R44.

FriendlyMineManager (intermine-javadoc).

Fucile, G., Di Biase, D., Nahal, H., La, G., Khodabandeh, S., Chen, Y. *et al.* ePlant and the 3D data display initiative: integrative systems biology on the world wide web. *PLoS One* 2011;**6**:e15237.

Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res* 2015;**43**:D1049–56.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 2012;**40**:D1178–86.

Howe, K.L., Bolt, B.J., Cain, S., Chan, J., Chen, W.J., Davis, P. *et al.* WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res* 2016;**44**:D774–80.

Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J. *et al.* InterMine: extensive web services for modern biology. *Nucleic Acids Res* 2014;**42**:W468–72.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**:D457–62.

Krishnakumar, V., Hanlon, M.R., Contrino, S., Ferlanti, E.S., Karamycheva, S., Kim, M. *et al.* Araport: the Arabidopsis information portal. *Nucleic Acids Res* 2015a;**43**:D1003–9.

Krishnakumar, V., Kim, M., Rosen, B.D., Karamycheva, S., Bidwell, S.L., Tang, H. *et al.* MTGD: The Medicago truncatula genome database. *Plant Cell Physiol* 2015b;**56**:e1.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 2012;**40**:D1202–10.

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F. *et al.* FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol* 2007;**8**:R129.

Lyne, R., Sullivan, J., Butano, D., Contrino, S., and Heimbach, J. Cross□organism analysis using InterMine. *Genesis* 2015.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2011;**39**:D52–7.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 2016;**44**:D336–42.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* 2015;**43**:D213–21.

Motenko, H., Neuhauser, S.B., O'Keefe, M., and Richardson, J.E. MouseMine: a new data warehouse for MGI. *Mamm Genome* 2015;**26**:325–30.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 2013, DOI: 10.1093/nar/gkt1115.

Patro, R., Duggal, G., and Kingsford, C. Accurate, fast, and model-aware transcript expression quantification with Salmon. *bioRxiv* 2015:021592.

Reiser, L., Berardini, T.Z., Li, D., Muller, R., Strait, E.M., Li, Q. *et al.* Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. *Database* 2016;**2016**, DOI: 10.1093/database/baw018.

Rosso, M.G., Li, Y., Strizhov, N., Reiss, B., Dekker, K., and Weisshaar, B. An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol Biol*

2003;**53**:247–59.

Ruzicka, L., Bradford, Y.M., Frazer, K., Howe, D.G., Paddock, H., Ramachandran, S. *et al.* ZFIN, The zebrafish model organism database: Updates and new directions. *Genesis* 2015;**53**:498–509.

Sessions, A., Burke, E., Presting, G., Aux, G., McElver, J., Patton, D. *et al.* A high-throughput Arabidopsis reverse genetics system. *Plant Cell* 2002;**14**:2985–94.

Shimoyama, M., De Pons, J., Hayman, G.T., Laulederkind, S.J.F., Liu, W., Nigam, R. *et al.* The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res* 2015;**43**:D743–50.

Smith, R.N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F. *et al.* InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics* 2012;**28**:3163–5.

UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 2014;**42**:D191–8.

UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., and Provart, N.J. An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets. *PLoS One* 2007;**2**:e718.

Woody, S.T., Austin-Phillips, S., Amasino, R.M., and Krysan, P.J. The WiscDsLox T-DNA collection: an arabidopsis community resource generated by using an improved high-throughput T-DNA sequencing pipeline. *J Plant Res* 2007;**120**:157–65.

Tables

Table 1. A summary of data sources hosted in ThaleMine release 1.10.0 (Aug 2016).

Data Types	Data Sources	Data Contents	References
Genes	Araport: Arabidopsis Information Portal	Araport11 genome annotation, Coding sequence FASTA, Protein Sequence FASTA, RNA-seq expression counts	(Krishnakumar *et al.* 2015a)
	TAIR: The Arabidopsis Information Resource	Genome assembly, gene summary, TAIR/ABRC ecotypes, germplasm, phenotypes, polymorphisms	(Lamesch *et al.* 2012)
	NCBI: National Center for Biotechnology Information	GeneRIF, Publications to gene mapping, Sequence	(Maglott *et al.* 2011)

		Read Archive	
Proteins	UniProt: Universal Protein Resource	Swiss-Prot and TrEMBL protein annotation, UniProt FASTA sequences, UniProt keywords	(UniProt Consortium 2014)
	InterPro: Protein sequence analysis & classification	InterPro protein domains	(Mitchell *et al.* 2015)
Homology	PANTHER: Protein ANalysis THrough Evolutionary Relationships	PANTHER human and yeast orthologs	(Mi *et al.* 2016)
	Phytozome: Plant Comparative Genomics portal	*Phytozome plant orthologs	(Goodstein *et al.* 2012)
Ontology	GO: Gene Ontology Consortium (GOC)	GO annotations from GOC, IntAct, RefGenome, TAIR, TIGR, InterPro and UniProt	(Gene Ontology Consortium 2015)
Interactions	BioGRID: The Biological General Repository for Interaction Datasets	BioGRID protein-protein interactions	(Chatr-Aryamontri *et al.* 2015)
	IntAct: Molecular Interaction Database	IntAct protein-protein interactions	(Orchard *et al.* 2013)
Expression & Co-Expression	BAR: The Bio-Analytic Resource for Plant Biology	*Arabidopsis electronic fluorescent pictograph (eFP), Gene to probe lookup, Affymetrix array expression	(Winter *et al.* 2007)
	ATTED-II: Arabidopsis thaliana trans-factor and cis-element prediction database	*Co-expression gene lists	(Aoki *et al.* 2016)
	TAIR: The Arabidopsis Information Resource	Plant Ontology (PO) annotations	(Berardini *et al.* 2015)
Pathways	KEGG: Kyoto Encyclopedia of Genes and Genomes	Arabidopsis metabolic pathways	(Kanehisa *et al.* 2016)

* : Real-time integration of remote data sources using web services exposed by 3rd party data providers.

Legends to Figures

Figure 1. An example of a core InterMine data model entity and extensions/additions made to it in the context of ThaleMine. Class hierarchy is illustrated by the overlapping boxes. Connecting lines signify the relationships between the various entities.

(A) Illustrates a part of the core InterMine data model used to describe the Gene class and its attributes. In this example, the Gene class inherits attributes from the SequenceFeature class, which in turn inherits from BioEntity (the root of the data model tree); (B) Illustrates the data model additions made to the Gene class in ThaleMine. In this example, the Gene entity has a new boolean attribute “Is Obsolete?” tracking whether a particular gene locus is active or not. The Locus History entity of the data model maintains a collection of genes involved in a particular operation (e.g. a merge involves 2 or more gene loci).

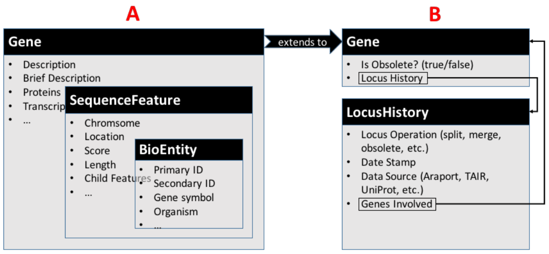
Figure 2. ThaleMine use case: Look up Gene Report for a gene of interest.

Figure 3. ThaleMine use case: Run functional enrichment analysis for a gene list of interest.

Figure 4. ThaleMine use case: Identify interacting protein partners for a gene list, and filter by RNA-seq expression levels.

Figure 5. ThaleMine use case: Run a data query to collect genes related to a keyword and export their translated sequences in FASTA format.

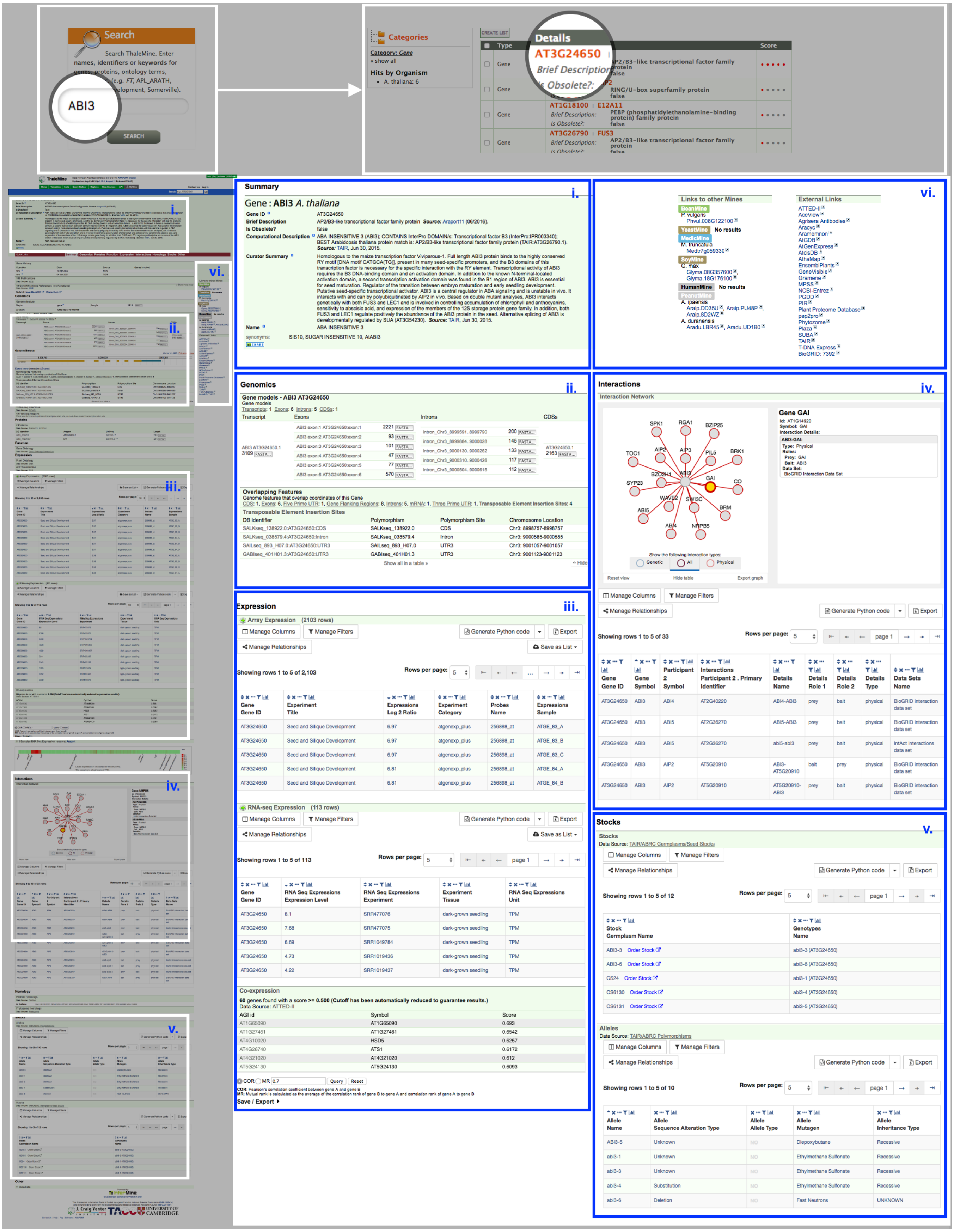
Figure 1



InterMine core data model
describing the **Gene** class

ThaleMine data model extension
describing **Gene** Locus History

Figure 2



Search

Search ThaleMine. Enter names, identifiers or keywords for genes, proteins, ontology terms, (e.g. FT, APL_ARATH, ...velopment, Somerville).

ABI3

SEARCH

Categories

Category: Gene

« show all

Hits by Organism

- A. thaliana: 6

CREATE LIST

☐ Type	Details	Score
☐ Gene	AT3G24650 ︱ ... Brief Description: AP2/B3-like transcriptional factor family protein Is Obsolete?: false	●●●●●
☐ Gene	...P2 RING/U-box superfamily protein false	●●●●●
☐ Gene	AT1G18100 ︱ E12A11 Brief Description: PEBP (phosphatidylethanolamine-binding protein) family protein Is Obsolete?: false	●●●●●
☐ Gene	AT3G26790 ︱ FUS3 Brief Description: AP2/B3-like transcriptional factor family protein Is Obsolete?: false	●●●●●

ThaleMine

Summary

i.

Gene : ABI3 *A. thaliana*

Gene ID	AT3G24650
Brief Description	AP2/B3-like transcriptional factor family protein *Source:* Araport11 (06/2016).
Is Obsolete?	false
Computational Description	ABA INSENSITIVE 3 (ABI3); CONTAINS InterPro DOMAIN/s: Transcriptional factor B3 (InterPro:IPR003340); BEST Arabidopsis thaliana protein match is: AP2/B3-like transcriptional factor family protein (TAIR:AT3G26790.1). *Source:* TAIR, Jun 30, 2015.
Curator Summary	Homologous to the maize transcription factor Viviparous-1. Full length ABI3 protein binds to the highly conserved RY motif [DNA motif CATGCA(TG)], present in many seed-specific promoters, and the B3 domains of this transcription factor is necessary for the specific interaction with the RY element. Transcriptional activity of ABI3 requires the B3 DNA-binding domain and an activation domain. In addition to the known N-terminal-located activation domain, a second transcription activation domain was found in the B1 region of ABI3. ABI3 is essential for seed maturation. Regulator of the transition between embryo maturation and early seedling development. Putative seed-specific transcriptional activator. ABI3 is a central regulator in ABA signaling and is unstable in vivo. It interacts with and can by polyubiquitinated by AIP2 in vivo. Based on double mutant analyses, ABI3 interacts genetically with both FUS3 and LEC1 and is involved in controlling accumulation of chlorophyll and anthocyanins, sensitivity to abscisic acid, and expression of the members of the 12S storage protein gene family. In addition, both FUS3 and LEC1 regulate positively the abundance of the ABI3 protein in the seed. Alternative splicing of ABI3 is developmentally regulated by SUA (AT3G54230). *Source:* TAIR, Jun 30, 2015.
Name	ABA INSENSITIVE 3
synonyms:	SIS10, SUGAR INSENSITIVE 10, AtABI3

SHARE

vi.

Links to other Mines

BeanMine
P. vulgaris
Phvul.008G122100

YeastMine No results

MedicMine
M. truncatula
Medtr7g059330

SoyMine
G. max
Glyma.08G357600, Glyma.18G176100

HumanMine No results

PeanutMine
A. ipaensis
Araip.DD35U, Araip.PU48P, Araip.8D2WZ
A. duranensis
Aradu.LBR45, Aradu.UD1B0

External Links

ATTED-II
AceView
Agrisera Antibodies
Aracyc
Aramemnon
AtGDB
AtGenExpress
AtcisDB
AthaMap
EnsemblPlants
GeneVisible
Gramene
MPSS
NCBI-Entrez
PGDD
PIR
Plant Proteome Database
pep2pro
Phytozome
Plaza
SUBA
TAIR
T-DNA Express
BioGRID: 7392

Genomics

ii.

Gene models - ABI3 AT3G24650

Gene models

Transcripts: 1 Exons: 6 Introns: 5 CDSs: 1

Transcript	Exons		Introns		CDSs
ABI3 AT3G24650.1 3109 FASTA...	ABI3:exon:1 AT3G24650:exon:1	2221 FASTA...	intron_Chr3_8999591..8999790	200 FASTA...	AT3G24650.1 2163 FASTA...
	ABI3:exon:2 AT3G24650:exon:2	93 FASTA...	intron_Chr3_8999884..9000028	145 FASTA...	
	ABI3:exon:3 AT3G24650:exon:3	101 FASTA...	intron_Chr3_9000130..9000262	133 FASTA...	
	ABI3:exon:4 AT3G24650:exon:4	47 FASTA...	intron_Chr3_9000310..9000426	117 FASTA...	
	ABI3:exon:5 AT3G24650:exon:5	77 FASTA...	intron_Chr3_9000504..9000615	112 FASTA...	
	ABI3:exon:6 AT3G24650:exon:6	570 FASTA...			

Overlapping Features

Genome features that overlap coordinates of this Gene

CDS: 1, Exons: 6, Five Prime UTR: 1, Gene Flanking Regions: 8, Introns: 5, mRNA: 1, Three Prime UTR: 1, Transposable Element Insertion Sites: 4

Transposable Element Insertion Sites

DB identifier	Polymorphism	Polymorphism Site	Chromosome Location
SALKseq_138922.0:AT3G24650:CDS	SALKseq_138922.0	CDS	Chr3: 8998757-8998757
SALKseq_038579.4:AT3G24650:Intron	SALKseq_038579.4	Intron	Chr3: 9000585-9000585
SAILseq_893_H07.0:AT3G24650:UTR3	SAILseq_893_H07.0	UTR3	Chr3: 9001057-9001057
GABIseq_401H01.3:AT3G24650:UTR3	GABIseq_401H01.3	UTR3	Chr3: 9001123-9001123

Show all in a table »

Hide

Expression

iii.

Array Expression (2103 rows)

Manage Columns | Manage Filters | Generate Python code | Export | Manage Relationships | Save as List

Showing rows 1 to 5 of 2,103

Rows per page: 5

Gene Gene ID	Experiment Title	Expressions Log 2 Ratio	Experiment Category	Probes Name	Expressions Sample
AT3G24650	Seed and Silique Development	6.97	atgenexp_plus	256898_at	ATGE_83_A
AT3G24650	Seed and Silique Development	6.97	atgenexp_plus	256898_at	ATGE_83_B
AT3G24650	Seed and Silique Development	6.97	atgenexp_plus	256898_at	ATGE_83_C
AT3G24650	Seed and Silique Development	6.81	atgenexp_plus	256898_at	ATGE_84_A
AT3G24650	Seed and Silique Development	6.81	atgenexp_plus	256898_at	ATGE_84_B

RNA-seq Expression (113 rows)

Manage Columns | Manage Filters | Generate Python code | Export | Manage Relationships | Save as List

Showing rows 1 to 5 of 113

Rows per page: 5 page 1

Gene Gene ID	RNA Seq Expressions Expression Level	RNA Seq Expressions Experiment	Experiment Tissue	RNA Seq Expressions Unit
AT3G24650	8.1	SRR477076	dark-grown seedling	TPM
AT3G24650	7.68	SRR477075	dark-grown seedling	TPM
AT3G24650	6.69	SRR1049784	dark-grown seedling	TPM
AT3G24650	4.73	SRR1019436	dark-grown seedling	TPM
AT3G24650	4.22	SRR1019437	dark-grown seedling	TPM

Co-expression

60 genes found with a score **>= 0.500 (Cutoff has been automatically reduced to guarantee results.)**

Data Source: ATTED-II

AGI id	Symbol	Score
AT1G65090	AT1G65090	0.693
AT1G27461	AT1G27461	0.6542
AT4G10020	HSD5	0.6257
AT4G26740	ATS1	0.6172
AT4G21020	AT4G21020	0.612
AT5G24130	AT5G24130	0.6093

● COR ○ MR 0.7 Query Reset

COR: Pearson's correlation coefficient between gene A and gene B

MR: Mutual rank is calculated as the average of the correlation rank of gene B to gene A and correlation rank of gene A to gene B

Save / Export ▸

Interactions

iv.

Interaction Network

Gene GAI

Id: AT1G14920

Symbol: GAI

Interaction Details:

ABI3-GAI:

Type: Physical

Roles:

Prey: GAI

Bait: ABI3

Data Set:

BioGRID Interaction Data Set

Manage Columns | Manage Filters | Manage Relationships | Generate Python code | Export

Showing rows 1 to 5 of 33

Rows per page: 5 page 1

Gene Gene ID	Gene Symbol	Participant 2 Symbol	Interactions Participant 2 . Primary Identifier	Details Name	Details Role 1	Details Role 2	Details Type	Data Sets Name
AT3G24650	ABI3	ABI4	AT2G40220	ABI4-ABI3	prey	bait	physical	BioGRID interaction data set
AT3G24650	ABI3	ABI5	AT2G36270	ABI5-ABI3	prey	bait	physical	BioGRID interaction data set
AT3G24650	ABI3	ABI5	AT2G36270	abi5-abi3	prey	bait	physical	IntAct interactions data set
AT3G24650	ABI3	AIP2	AT5G20910	ABI3-AT5G20910	bait	prey	physical	BioGRID interaction data set
AT3G24650	ABI3	AIP2	AT5G20910	AT5G20910-ABI3	prey	bait	physical	BioGRID interaction data set

Stocks

v.

Stocks

Data Source: TAIR/ABRC Germplasms/Seed Stocks

Manage Columns | Manage Filters | Manage Relationships | Generate Python code | Export

Showing rows 1 to 5 of 12

Rows per page: 5 page 1

Stock Germplasm Name	Genotypes Name
ABI3-3 Order Stock	abi3-3 (AT3G24650)
ABI3-6 Order Stock	abi3-6 (AT3G24650)
CS24 Order Stock	abi3-1 (AT3G24650)
CS6130 Order Stock	abi3-4 (AT3G24650)
CS6131 Order Stock	abi3-5 (AT3G24650)

Alleles

Data Source: TAIR/ABRC Polymorphisms

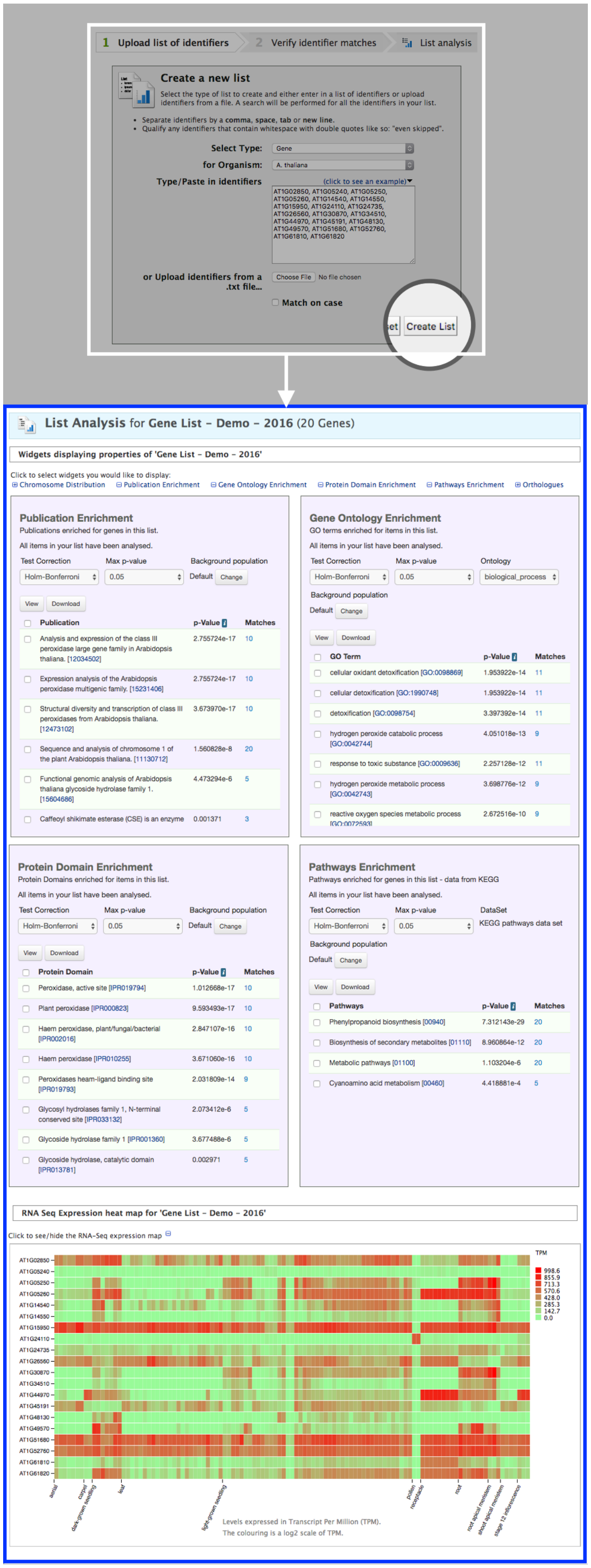
Manage Columns | Manage Filters | Manage Relationships | Generate Python code | Export

Showing rows 1 to 5 of 10

Rows per page: 5 page 1

Allele Name	Allele Sequence Alteration Type	Allele Allele Type	Allele Mutagen	Allele Inheritance Type
ABI3-5	Unknown	NO	Diepoxybutane	Recessive
abi3-1	Unknown	NO	Ethylmethane Sulfonate	Recessive
abi3-3	Unknown	NO	Ethylmethane Sulfonate	Recessive
abi3-4	Substitution	NO	Ethylmethane Sulfonate	Recessive
abi3-6	Deletion	NO	Fast Neutrons	UNKNOWN

Figure 3



1 Upload list of identifiers | 2 Verify identifier matches | List analysis

Create a new list

Select the type of list to create and either enter in a list of identifiers or upload identifiers from a file. A search will be performed for all the identifiers in your list.

- Separate identifiers by a **comma**, **space**, **tab** or **new line**.
- Qualify any identifiers that contain whitespace with double quotes like so: "even skipped".

Select Type: Gene

for Organism: A. thaliana

Type/Paste in identifiers (click to see an example)

AT1G02850, AT1G05240, AT1G05250,
AT1G05260, AT1G14540, AT1G14550,
AT1G15950, AT1G24110, AT1G24735,
AT1G26560, AT1G30870, AT1G34510,
AT1G44970, AT1G45191, AT1G48130,
AT1G49570, AT1G51680, AT1G52760,
AT1G61810, AT1G61820

or Upload identifiers from a .txt file... Choose File No file chosen

☐ **Match on case**

List Analysis for Gene List – Demo – 2016 (20 Genes)

Widgets displaying properties of 'Gene List – Demo – 2016'

Click to select widgets you would like to display:
⊞ Chromosome Distribution ⊟ Publication Enrichment ⊟ Gene Ontology Enrichment ⊟ Protein Domain Enrichment ⊟ Pathways Enrichment ⊞ Orthologues

Publication Enrichment

Publications enriched for genes in this list.

All items in your list have been analysed.

Test Correction: Holm-Bonferroni | Max p-value: 0.05 | Background population: Default Change

View Download

☐	Publication	p-Value	Matches
☐	Analysis and expression of the class III peroxidase large gene family in Arabidopsis thaliana. [12034502]	2.755724e-17	10
☐	Expression analysis of the Arabidopsis peroxidase multigenic family. [15231406]	2.755724e-17	10
☐	Structural diversity and transcription of class III peroxidases from Arabidopsis thaliana. [12473102]	3.673970e-17	10
☐	Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana. [11130712]	1.560828e-8	20
☐	Functional genomic analysis of Arabidopsis thaliana glycoside hydrolase family 1. [15604686]	4.473294e-6	5
☐	Caffeoyl shikimate esterase (CSE) is an enzyme	0.001371	3

Gene Ontology Enrichment

GO terms enriched for items in this list.

All items in your list have been analysed.

Test Correction: Holm-Bonferroni | Max p-value: 0.05 | Ontology: biological_process

Background population: Default Change

View Download

☐	GO Term	p-Value	Matches
☐	cellular oxidant detoxification [GO:0098869]	1.953922e-14	11
☐	cellular detoxification [GO:1990748]	1.953922e-14	11
☐	detoxification [GO:0098754]	3.397392e-14	11
☐	hydrogen peroxide catabolic process [GO:0042744]	4.051018e-13	9
☐	response to toxic substance [GO:0009636]	2.257128e-12	11
☐	hydrogen peroxide metabolic process [GO:0042743]	3.698776e-12	9
☐	reactive oxygen species metabolic process [GO:0072593]	2.672516e-10	9

Protein Domain Enrichment

Protein Domains enriched for items in this list.

All items in your list have been analysed.

Test Correction: Holm-Bonferroni | Max p-value: 0.05 | Background population: Default Change

View Download

☐	Protein Domain	p-Value	Matches
☐	Peroxidase, active site [IPR019794]	1.012668e-17	10
☐	Plant peroxidase [IPR000823]	9.593493e-17	10
☐	Haem peroxidase, plant/fungal/bacterial [IPR002016]	2.847107e-16	10
☐	Haem peroxidase [IPR010255]	3.671060e-16	10
☐	Peroxidases heam-ligand binding site [IPR019793]	2.031809e-14	9
☐	Glycosyl hydrolases family 1, N-terminal conserved site [IPR033132]	2.073412e-6	5
☐	Glycoside hydrolase family 1 [IPR001360]	3.677488e-6	5
☐	Glycoside hydrolase, catalytic domain [IPR013781]	0.002971	5

Pathways Enrichment

Pathways enriched for genes in this list - data from KEGG

All items in your list have been analysed.

Test Correction: Holm-Bonferroni | Max p-value: 0.05 | DataSet: KEGG pathways data set

Background population: Default Change

View Download

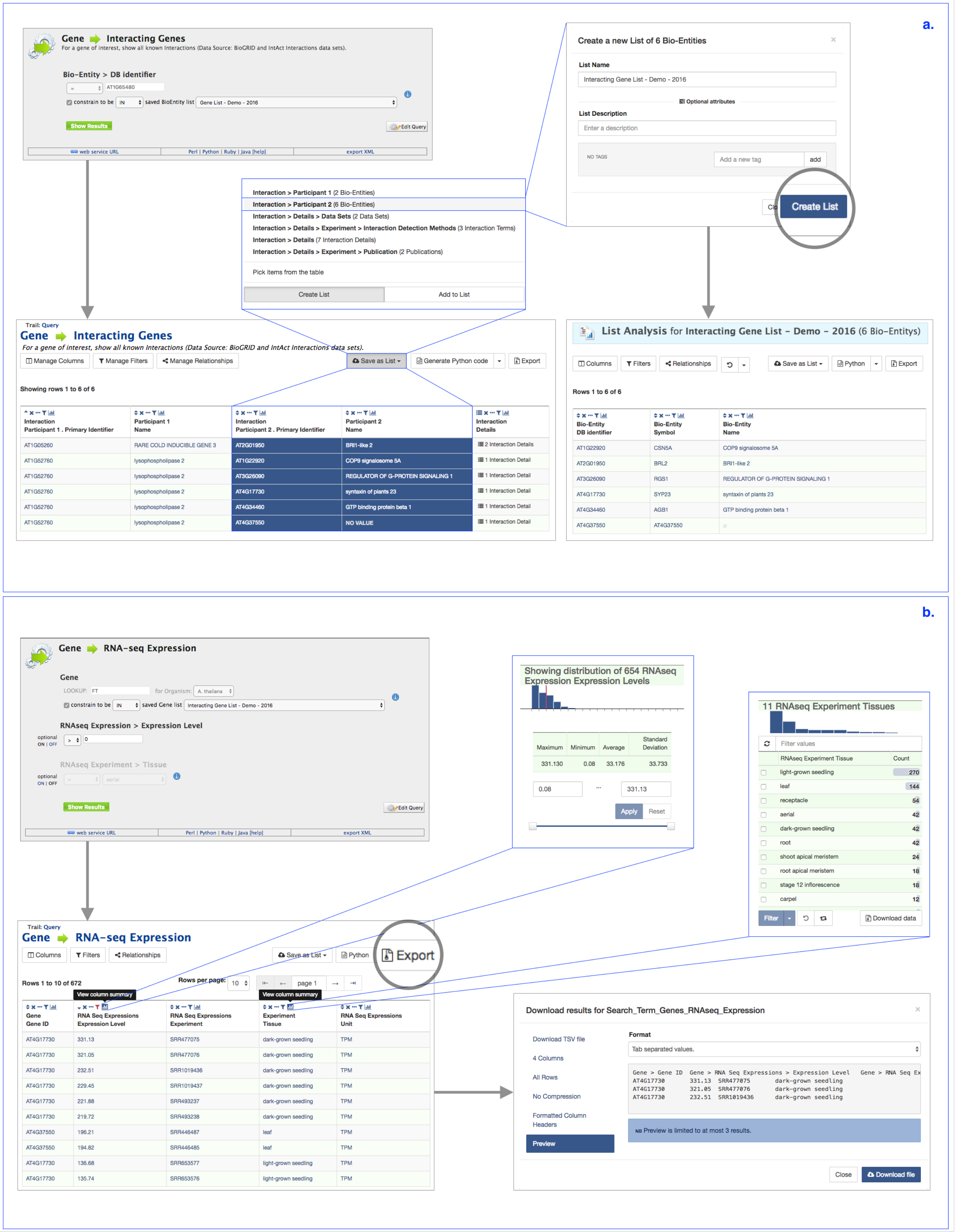
☐	Pathways	p-Value	Matches
☐	Phenylpropanoid biosynthesis [00940]	7.312143e-29	20
☐	Biosynthesis of secondary metabolites [01110]	8.960864e-12	20
☐	Metabolic pathways [01100]	1.103204e-6	20
☐	Cyanoamino acid metabolism [00460]	4.418881e-4	5

RNA Seq Expression heat map for 'Gene List – Demo – 2016'

Click to see/hide the RNA-Seq expression map ⊟

Levels expressed in Transcript Per Million (TPM).
The colouring is a log2 scale of TPM.

Figure 4



Gene → Interacting Genes
For a gene of interest, show all known Interactions (Data Source: BioGRID and IntAct Interactions data sets).

Bio-Entity > DB identifier
= AT1G65480
☑ constrain to be IN saved BioEntity list Gene List - Demo - 2016

Interaction > Participant 1 (2 Bio-Entities)
Interaction > Participant 2 (6 Bio-Entities)
Interaction > Details > Data Sets (2 Data Sets)
Interaction > Details > Experiment > Interaction Detection Methods (3 Interaction Terms)
Interaction > Details (7 Interaction Details)
Interaction > Details > Experiment > Publication (2 Publications)
Pick items from the table

Create a new List of 6 Bio-Entities

List Name
Interacting Gene List - Demo - 2016

Optional attributes

List Description
Enter a description

NO TAGS
Add a new tag

Trail: Query
Gene → Interacting Genes
For a gene of interest, show all known Interactions (Data Source: BioGRID and IntAct Interactions data sets).

Showing rows 1 to 6 of 6

Interaction Participant 1 . Primary Identifier	Participant 1 Name	Interaction Participant 2 . Primary Identifier	Participant 2 Name	Interaction Details
AT1G05260	RARE COLD INDUCIBLE GENE 3	AT2G01950	BRI1-like 2	2 Interaction Details
AT1G52760	lysophospholipase 2	AT1G22920	COP9 signalosome 5A	1 Interaction Detail
AT1G52760	lysophospholipase 2	AT3G26090	REGULATOR OF G-PROTEIN SIGNALING 1	1 Interaction Detail
AT1G52760	lysophospholipase 2	AT4G17730	syntaxin of plants 23	1 Interaction Detail
AT1G52760	lysophospholipase 2	AT4G34460	GTP binding protein beta 1	1 Interaction Detail
AT1G52760	lysophospholipase 2	AT4G37550	NO VALUE	1 Interaction Detail

List Analysis for **Interacting Gene List – Demo – 2016** (6 Bio–Entitys)

Rows 1 to 6 of 6

Bio-Entity DB identifier	Bio-Entity Symbol	Bio-Entity Name
AT1G22920	CSN5A	COP9 signalosome 5A
AT2G01950	BRL2	BRI1-like 2
AT3G26090	RGS1	REGULATOR OF G-PROTEIN SIGNALING 1
AT4G17730	SYP23	syntaxin of plants 23
AT4G34460	AGB1	GTP binding protein beta 1
AT4G37550	AT4G37550	∅

Gene → RNA-seq Expression

Gene
LOOKUP: FT for Organism: A. thaliana
☑ constrain to be IN saved Gene list Interacting Gene List - Demo - 2016

RNAseq Expression > Expression Level
optional ON | OFF > 0

RNAseq Experiment > Tissue
optional ON | OFF = aerial

Trail: Query
Gene → RNA-seq Expression

Rows 1 to 10 of 672   Rows per page: 10   page 1

Gene Gene ID	RNA Seq Expressions Expression Level	RNA Seq Expressions Experiment	Experiment Tissue	RNA Seq Expressions Unit
AT4G17730	331.13	SRR477075	dark-grown seedling	TPM
AT4G17730	321.05	SRR477076	dark-grown seedling	TPM
AT4G17730	232.51	SRR1019436	dark-grown seedling	TPM
AT4G17730	229.45	SRR1019437	dark-grown seedling	TPM
AT4G17730	221.88	SRR493237	dark-grown seedling	TPM
AT4G17730	219.72	SRR493238	dark-grown seedling	TPM
AT4G37550	196.21	SRR446487	leaf	TPM
AT4G37550	194.82	SRR446485	leaf	TPM
AT4G17730	136.68	SRR653577	light-grown seedling	TPM
AT4G17730	135.74	SRR653576	light-grown seedling	TPM

Showing distribution of 654 RNAseq Expression Expression Levels

Maximum	Minimum	Average	Standard Deviation
331.130	0.08	33.176	33.733

0.08 ··· 331.13

11 RNAseq Experiment Tissues

Filter values

RNAseq Experiment Tissue	Count
□ light-grown seedling	270
□ leaf	144
□ receptacle	54
□ aerial	42
□ dark-grown seedling	42
□ root	42
□ shoot apical meristem	24
□ root apical meristem	18
□ stage 12 inflorescence	18
□ carpel	12

Download results for Search_Term_Genes_RNAseq_Expression

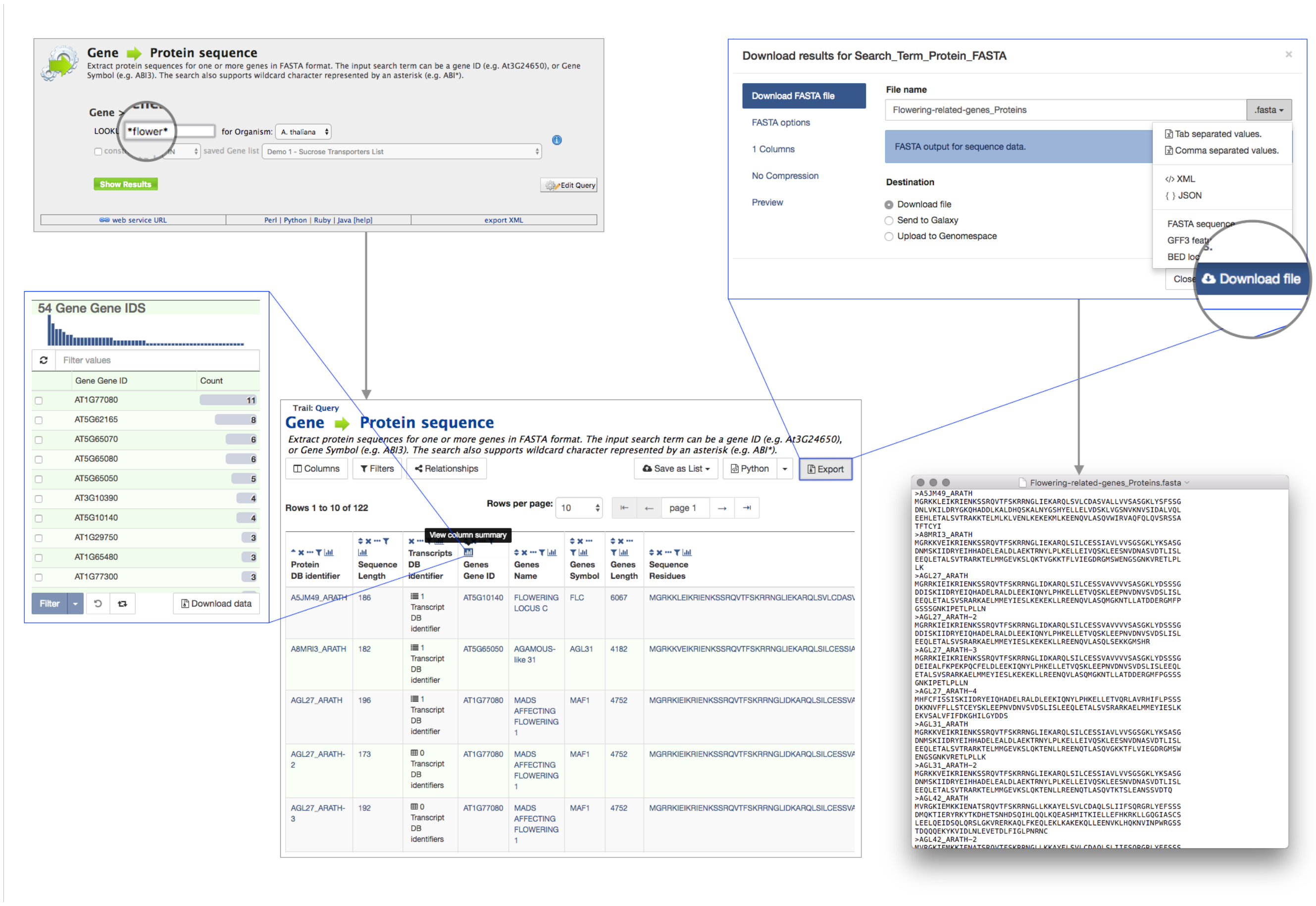
Download TSV file
4 Columns
All Rows
No Compression
Formatted Column Headers

Format
Tab separated values.

```
Gene > Gene ID  Gene > RNA Seq Expressions > Expression Level   Gene > RNA Seq Ex
AT4G17730       331.13  SRR477075       dark-grown seedling
AT4G17730       321.05  SRR477076       dark-grown seedling
AT4G17730       232.51  SRR1019436      dark-grown seedling
```

NB Preview is limited to at most 3 results.

Figure 5



Gene ➡ Protein sequence

Extract protein sequences for one or more genes in FASTA format. The input search term can be a gene ID (e.g. At3G24650), or Gene Symbol (e.g. ABI3). The search also supports wildcard character represented by an asterisk (e.g. ABI*).

Gene >

LOOKU *flower* for Organism: A. thaliana

□ const IN saved Gene list Demo 1 - Sucrose Transporters List

54 Gene Gene IDS

Filter values

	Gene Gene ID	Count
□	AT1G77080	11
□	AT5G62165	8
□	AT5G65070	6
□	AT5G65080	6
□	AT5G65050	5
□	AT3G10390	4
□	AT5G10140	4
□	AT1G29750	3
□	AT1G65480	3
□	AT1G77300	3

Trail: Query

Gene ➡ Protein sequence

Extract protein sequences for one or more genes in FASTA format. The input search term can be a gene ID (e.g. At3G24650), or Gene Symbol (e.g. ABI3). The search also supports wildcard character represented by an asterisk (e.g. ABI).*

Rows 1 to 10 of 122

Rows per page: 10 page 1

Protein DB identifier	Sequence Length	Transcripts DB identifier	Genes Gene ID	Genes Name	Genes Symbol	Genes Length	Sequence Residues
A5JM49_ARATH	186	1 Transcript DB identifier	AT5G10140	FLOWERING LOCUS C	FLC	6067	MGRKKLEIKRIENKSSRQVTFSKRRNGLIEKARQLSVLCDAS\
A8MRI3_ARATH	182	1 Transcript DB identifier	AT5G65050	AGAMOUS-like 31	AGL31	4182	MGRKKVEIKRIENKSSRQVTFSKRRNGLIEKARQLSILCESSIA
AGL27_ARATH	196	1 Transcript DB identifier	AT1G77080	MADS AFFECTING FLOWERING 1	MAF1	4752	MGRRKIEIKRIENKSSRQVTFSKRRNGLIDKARQLSILCESSVA
AGL27_ARATH-2	173	0 Transcript DB identifiers	AT1G77080	MADS AFFECTING FLOWERING 1	MAF1	4752	MGRRKIEIKRIENKSSRQVTFSKRRNGLIDKARQLSILCESSVA
AGL27_ARATH-3	192	0 Transcript DB identifiers	AT1G77080	MADS AFFECTING FLOWERING 1	MAF1	4752	MGRRKIEIKRIENKSSRQVTFSKRRNGLIDKARQLSILCESSVA

Download results for Search_Term_Protein_FASTA

File name: Flowering-related-genes_Proteins .fasta

FASTA output for sequence data.

Destination

● Download file
○ Send to Galaxy
○ Upload to Genomespace

Flowering-related-genes_Proteins.fasta

```
>A5JM49_ARATH
MGRKKLEIKRIENKSSRQVTFSKRRNGLIEKARQLSVLCDASVALLVVSASGKLYSFSSG
DNLVKILDRYGKQHADDLKALDHQSKALNYGSHYELLELVDSKLVGSNVKNVSIDALVQL
EEHLETALSVTRAKKTELMLKLVENLKEKEKMLKEENQVLASQVWIRVAQFQLQVSRSSA
TFTCYI
>A8MRI3_ARATH
MGRKKVEIKRIENKSSRQVTFSKRRNGLIEKARQLSILCESSIAVLVVSGSGKLYKSASG
DNMSKIIDRYEIHHADELEALDLAEKTRNYLPLKELLEIVQSKLEESNVDNASVDTLISL
EEQLETALSVTRARKTELMMGEVKSLQKTVGKKTFLVIEGDRGMSWENGSGNKVRETLPL
LK
>AGL27_ARATH
MGRRKIEIKRIENKSSRQVTFSKRRNGLIDKARQLSILCESSVAVVVVSASGKLYDSSSG
DDISKIIDRYEIQHADELRALDLEEKIQNYLPHKELLETVQSKLEEPNVDNVSVDSLISL
EEQLETALSVSRARKAELMMEYIESLKEKEKLLREENQVLASQMGKNTLLATDDERGMFP
GSSSGNKIPETLPLLN
>AGL27_ARATH-2
MGRRKIEIKRIENKSSRQVTFSKRRNGLIDKARQLSILCESSVAVVVVSASGKLYDSSSG
DDISKIIDRYEIQHADELRALDLEEKIQNYLPHKELLETVQSKLEEPNVDNVSVDSLISL
EEQLETALSVSRARKAELMMEYIESLKEKEKLLREENQVLASQLSEKKGMSHR
>AGL27_ARATH-3
MGRRKIEIKRIENKSSRQVTFSKRRNGLIDKARQLSILCESSVAVVVVSASGKLYDSSSG
DEIEALFKPEKPQCFELDLEEKIQNYLPHKELLETVQSKLEEPNVDNVSVDSLISLEEQL
ETALSVSRARKAELMMEYIESLKEKEKLLREENQVLASQMGKNTLLATDDERGMFPGSSS
GNKIPETLPLLN
>AGL27_ARATH-4
MHFCFISSISKIIDRYEIQHADELRALDLEEKIQNYLPHKELLETVQRLAVRHIFLPSSS
DKKNVFFLLSTCEYSKLEEPNVDNVSVDSLISLEEQLETALSVSRARKAELMMEYIESLK
EKVSALVFIFDKGHILGYDDS
>AGL31_ARATH
MGRKKVEIKRIENKSSRQVTFSKRRNGLIEKARQLSILCESSIAVLVVSGSGKLYKSASG
DNMSKIIDRYEIHHADELEALDLAEKTRNYLPLKELLEIVQSKLEESNVDNASVDTLISL
EEQLETALSVTRARKTELMMGEVKSLQKTENLLREENQTLASQVGKKTFLVIEGDRGMSW
ENGSGNKVRETLPLLK
>AGL31_ARATH-2
MGRKKVEIKRIENKSSRQVTFSKRRNGLIEKARQLSILCESSIAVLVVSGSGKLYKSASG
DNMSKIIDRYEIHHADELEALDLAEKTRNYLPLKELLEIVQSKLEESNVDNASVDTLISL
EEQLETALSVTRARKTELMMGEVKSLQKTENLLREENQTLASQVTKTSLEANSSVDTQ
>AGL42_ARATH
MVRGKIEMKKIENATSRQVTFSKRRNGLLKKAYELSVLCDAQLSLIIFSQRGRLYEFSSS
DMQKTIERYRKYTKDHETSNHDSQIHLQQLKQEASHMITKIELLEFHKRKLLGQGIASCS
LEELQEIDSQLQRSLGKVRERKAQLFKEQLEKLKAKEKQLLEENVKLHQKNVINPWRGSS
TDQQQEKYKVIDLNLEVETDLFIGLPNRNC
>AGL42_ARATH-2
MVRGKIEMKKIENATSRQVTFSKRRNGLLKKAYELSVLCDAQLSLIIFSQRGRLYEFSSS
```