How to effectively use the tools and resources at TAIR to enhance your research

Philippe Lamesch
Donghui Li
A.S. Karthikeyan

The Arabidopsis Information Resource
www.arabidopsis.org

contact us: curator@arabidopsis.org
The Arabidopsis Information Resource (TAIR) collects, curates, and distributes information on Arabidopsis species. Information is freely available from arabidopsis.org.
Outline

• Gene structure - Philippe
• Gene function - Donghui
• Metabolic pathway - Karthik

Goal of this workshop

• highlight important data sets and tools
• demonstrate search strategies
Slides available from TAIR

ICAR Workshop 2009

TAIR gave a workshop at the 2009 ICAR, ICAR is the International Congress of Arid Zone Research. ICAR is also the International Congress of Arid Zone Research.

Putting TAIR to work for you -
- Date: July 2, 2009
- Time: 4:30 - 6:00
- Location: Kitts
- Conference site:

Part I: Tips and strategies

1. Gene searches and the locus page (presented by Eva Huerta)
2. TAIR, GBrowse, and confidence scores (presented by Dave Swarbreck)
3. Generating and working with data sets (presented by Kate Brecher)
   - TAIR ICAR Workshop Resource Guide

Part II: Hands-on exercises and individual help (presenters and participants)

- TAIR ICAR Workshop Practice Questions
- TAIR workshop - part 2 - data file
  (Microarray / OMICs viewer data file for Question #1)
- TAIR ICAR Workshop Practice Questions with Answers
TAIR is used worldwide

Visits per month (source: Google Analytics)
What we do: (1) Arabidopsis genome annotation
What we do: (2) manual literature curation

- Controlled vocabulary annotations
  
  Gene Ontology (GO) http://www.geneontology.org/
  
  Plant Ontology (PO) http://www.plantontology.org/

- Gene name, symbol

- Allele, phenotype

- Summary statement composition
What we do: (3) metabolic pathway curation

AraCyc

A metabolic pathway database for Arabidopsis thaliana that contains information about both predicted and experimentally determined pathways, reactions, compounds, genes and enzymes.

PlantCyc and PMN (Plant Metabolic Network)
What we do: (4) work with ABRC to distribute research material
How do TAIR curators annotate gene structures? Where can gene structure related data be found at TAIR?

Part I: Tips for Accessing Gene Structure Data

Philippe Lamesch
Gene structure Annotation

- Arabidopsis genome sequenced almost 9 years ago
- High quality sequence with few gaps
- TIGR did initial genome annotation
- TAIR took over responsibility in 2005
- Current stats:
  27,379 protein coding genes
  4827 pseudogenes or transposable elements
  1312 ncRNAs
Responsibilities of a gene structure curator
Responsibilities of a gene structure curator

Add novel genes
Responsibilities of a gene structure curator

Delete wrongly predicted genes
Responsibilities of a gene structure curator

Update mispredicted exon-intron structure

cDNA
Responsibilities of a gene structure curator

Update mispredicted exon-intron structure

cDNA
Responsibilities of a gene structure curator

Annotate splice-variants
Responsibilities of a gene structure curator

Annotate ‘atypical’ gene classes

- Short protein-coding genes
- Transposable element genes
- Pseudogenes
- uORFs (genes within UTR of other genes)
Responsibilities of a gene structure curator

Define gene type

- Protein-coding
- tRNA
- snRNA
- snoRNA
- rRNA
...
Gene structure annotation in Arabidopsis

**TAIR 6**
- NEW: 681 genes; 828 exons
- UPDATED: 10,792 models and 14,050 exons

**TAIR 7**
- NEW: 1291 genes; 683 exons
- UPDATED: 3811 models; 4007 exons

**TAIR 8**
- NEW: 282 genes; 1056 exons
- UPDATED: 1254 models; 1144 exons

**TAIR 9**
How do MOD curators annotate genomes?

Experimental & Computational Evidence

Manual annotation → Automatic pipeline

Genome annotation
How do MOD curators annotate genomes?

Experimental & Computational Evidence

Manual annotation

Automatic pipeline

Genome annotation

ESTs
cDNAs
Automated pipeline at TAIR

Program for aligned sequence (PASA)
Automated pipeline at TAIR

Program for aligned sequence (PASA)

NCBI

Clustered transcripts

Resulting gene model

Previous gene model
Automated pipeline at TAIR

Program for aligned sequence (PASA)

NCBI

Clustered transcripts

Resulting gene model

Previous gene model

comparison
Automated pipeline at TAIR

Program for aligned sequence (PASA)

Based on a set of rules a decision is made
How do MOD curators annotate genomes?

Experimental & Computational Evidence

Manual annotation

Automatic pipeline

Genome annotation
How do MOD curators annotate genomes?

- Experimental & Computational Evidence
- Manual annotation
- Automatic pipeline

Genome annotation
How do MOD curators annotate genomes?

Experimental & Computational Evidence

- Alternative gene models
- Short MS peptides
- Community submissions
- ...

Manual annotation

Automatic pipeline

Genome annotation
Manual annotation at different MODs

- Evidence set
- Genome editing tool
- Set of annotation rules
Manual annotation at different MODs

Evidence set
- Nucleotide sequence
- Short peptides
- Protein similarity
- Alternative predictions
...

Genome editing tool
- Apollo (Arabidopsis, Fly)
- Aceview (Worm)
- Zmap/Otterlace (Human)
- Artemis (Pathogen Project)
...

Set of annotation rules
- Exon size
- Intron size
- Number of UTRs
- Coding/Non-coding ratio
- Splice-junctions
...
Manual annotation at TAIR: Apollo

Radish sequence alignments

dicot sequence alignments

Aceview gene predictions

cDNAs

ESTs

2 gene isoforms

Short MS peptide

Eugene prediction

monocot sequence alignments
Examples of large-scale community datasets recently integrated into the Arabidopsis annotation

- Transposable elements (Quesneville Lab)
- Pseudogenes (Gerstein Lab)
- Short MS peptides (Baerenfaller et al, Castellana et al)
- Short genes (Hanada et al)
Where can you find gene structure data on TAIR?

- **ON GENE MODEL PAGE**
  - Graphic of exon-intron structure
  - Coordinates of each exon

- **ON GBROWSE**
  - Graphic display of structure and overlapping evidence data

- **ON FTP SITE**
  - GFF files with exact structures of each gene model
  - Files with gene confidence ranking information
Gene Locus Page

Locus: AT1G07780
Date last modified: 2007-04-17
TAIR Accession: Locus 2026569
Representative Gene Model: AT1G07780.3
Gene Model Type: protein_coding
Other names: AT1G07780.1, F24B9.11, F24B9.11, PAI1, PHOSPHOBOSYLANTRILATE ISOMERASE 1, TRANSIENT RECEPTOR POTENTIAL 6, TRP6
Description: Encodes phosphoribosylanthranilate isomerase which catalyzes the third step of the tryptophan biosynthetic pathway. Member of gene family
Other Gene Models: AT1G07780.4 (splice variant) AT1G07780.1 (splice variant)

Map Detail Image

Annotations
Category
GO Biological Process
GO Cellular Component
GO Molecular Function
Plant structure
Relationship Type
involved in
located in
has
expressed in
Keyword
tryptophan biosynthetic process
cloroplast
phosphoribosylanthranilate isomerase activity
guard cell
Annotation Detail
Where can you find gene structure data on TAIR?

- **ON GENE MODEL PAGE**
  - Graphic of exon-intron structure
  - Coordinates of each exon

- **ON GBROWSE**
  - Graphic display of structure and overlapping evidence data

- **ON FTP SITE**
  - GFF files with exact structures of each gene model
  - Files with gene confidence ranking information
Gbrowse
Where can you find gene structure data on TAIR?

- **ON GENE MODEL PAGE**
  - Graphic of exon-intron structure
  - Coordinates of each exon
- **ON GBROWSE**
  - Graphic display of structure and overlapping evidence data
- **ON FTP SITE**
  - GFF files with exact structures of each gene model
  - Files with gene confidence ranking information
The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and 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 data is updated every two weeks from the latest published research literature and from experimental data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR provides extensive links to our data pages from other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, preserves, and distributes seed and DNA resources of Arabidopsis thaliana and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.

TAIR is located at the Carnegie Institution for Science Department of Plant Biology and funded by the National Science Foundation.

TAIR at the 2009 ASPB Conference

Workshop I: How to effectively use the tools and resources at TAIR to enhance your research. July 20, 12:20-2pm Room 318 A&B

Workshop II: TAIR, PMN, Gramene and SGN workshop: focus on comparative genomics and new tools. July 20, 7:30-8:30pm Room 318 A&B

Annual Meeting of the American Society of Plant Biologists & the Physiological Society of America

Tips for searching DNA stocks including vectors and amiRNA clones [May 20, 2009]
Are you searching for clones and vectors available from ABRC? Here are some tips for finding them using the TAIR searches and ABRC catalog pages. Specific instructions for finding amiRNA clones and multifunctional vectors are included.
FTP site


- Assembly_GFF
- Community_annotation_GFF
- Expression_GFF
- Genomic_features_GFF
- Methylation_Phosphorylation_GFF
- Ortholog_GFF
- Sequence_similarity_GFF
- TAIR9_GFF3_genes.gff
- TAIR9_GFF3_genes_transposons.gff
- Variation_GFF
- arabidopsis.conf

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly_GFF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community_annotation_GFF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expression_GFF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genomic_features_GFF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methylation_Phosphorylation_GFF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ortholog_GFF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence_similarity_GFF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAIR9_GFF3_genes.gff</td>
<td>30298 KB</td>
<td>3/26/2009 12:12:00 AM</td>
</tr>
<tr>
<td>TAIR9_GFF3_genes_transposons.gff</td>
<td>44833 KB</td>
<td>6/19/2009 11:06:00 PM</td>
</tr>
<tr>
<td>Variation_GFF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>arabidopsis.conf</td>
<td>121 KB</td>
<td>6/29/2009 5:28:00 PM</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>chromosome</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>------------</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chr1</th>
<th>TAIR9</th>
<th>gene</th>
<th>3631</th>
<th>5899</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>mRNA</td>
<td>3631</td>
<td>5899</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>protein</td>
<td>3760</td>
<td>5630</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>3631</td>
<td>3913</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>five_prime_UTR</td>
<td>3631</td>
<td>3759</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>3760</td>
<td>3913</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>3996</td>
<td>4276</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>3996</td>
<td>4276</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>4486</td>
<td>4605</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>4486</td>
<td>4605</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>4706</td>
<td>5095</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>4706</td>
<td>5095</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>5174</td>
<td>5326</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>5174</td>
<td>5326</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>5439</td>
<td>5899</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>three_prime_UTR</td>
<td>5631</td>
<td>5899</td>
<td>+</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chr1</th>
<th>TAIR9</th>
<th>gene</th>
<th>5928</th>
<th>8737</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>mRNA</td>
<td>5928</td>
<td>8737</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>protein</td>
<td>6915</td>
<td>8666</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>five_prime_UTR</td>
<td>5667</td>
<td>8737</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>5871</td>
<td>8666</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>5871</td>
<td>8737</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>8417</td>
<td>8466</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>8417</td>
<td>8466</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>8236</td>
<td>8325</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>8236</td>
<td>8325</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>7942</td>
<td>7987</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>7942</td>
<td>7987</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>7762</td>
<td>7835</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>exon</td>
<td>7762</td>
<td>7835</td>
<td>+</td>
</tr>
<tr>
<td>Chr1</td>
<td>TAIR9</td>
<td>CDS</td>
<td>7544</td>
<td>7649</td>
<td>+</td>
</tr>
</tbody>
</table>

The table contains information about gene structures and annotations, such as chromosome, gene, mRNA, protein, and exon coordinates, as well as CDS and five_prime_UTR sequences.
Where can you find gene structure data on TAIR?

• **ON GENE MODEL PAGE**
  - Graphic of exon-intron structure
  - Coordinates of each exon

• **ON GBROWSE**
  - Graphic display of structure and overlapping evidence data

• **ON FTP SITE**
  - GFF files with exact structures of each gene model
  - Files with gene confidence ranking information
Gene Confidence Rank

- Attributes confidence scores to all exons and gene models based on different types of experimental and computational evidence

**TYPES OF EXPERIMENTAL & COMPUTATIONAL DATA USED**

We used 4 types of experimental evidence to generate confidence scores:

1. **Arabidopsis transcript data** (ESTs, cDNAs, fragmented mRNAs from Ecker lab: Lister et al Cell 2008)
2. **Proteomics data** (Castellana et al PNAS 2008; Baerenfeller et al Science 2008)
3. **X-species alignments** (Brassica, Radish, + 5 other dicot (Glycine max, Solanum tuberosum, Malus x domestica, Gossypium hirsutum and Vitis vinifera) and 5 other monocot (Triticum aestivum, Zea mays, Oryza sativa, Saccharum officinarum and Hordeum vulgare) sequences from JCVI http://plantta.jcvi.org/index.shtml)
4. **Vista genomic conservation** (Shah et al Bioinformatics 2004)
Assigning A Confidence Rank
Full support

No support
How to effectively use the tools and resources at TAIR to enhance your research

Part II: Tips for Accessing Gene Function Data

Donghui Li
Overview

Focus on how to use the tools, not on how we make functional annotation

1. Finding a specific gene of interest in TAIR
2. Getting lots of data about that gene of interest
3. Getting data sets on multiple genes
4. “Analyzing” your data sets
1. Finding a gene/protein

Case 1: You know exactly what gene you want

- You know the AGI locus code (e.g. At2g46990)
- You know the gene symbol (e.g. PHYA)
1. Finding a gene/protein: by name/AGI locus code

Use the quick search box!
PHYA resulted in 20 text matches with 28 distinct gene models.

Displaying 1 - 20.

To see ESTs associated with your gene of interest, click on the Locus link.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Description</th>
<th>Gene Model</th>
<th>Other Names</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G09530</td>
<td>Transcription factor interacting with photoreceptors phyA and phyB. Forms a ternary complex in vitro with G-box element of Fl.</td>
<td>AT1G09530.1</td>
<td>F14J9.18, F14J9.19, F6P3, PHOTOCURRENT 1, PHOTOCURRENT, PHOTORECEPTOR, PHOTORECEPTOR-INTERACTING-FACTOR 3, PHOTORECEPTOR-ASSOCIATED-PROTEIN 3, PIF3, POC1</td>
<td>4. anthetic, C. globular stage, D. bilateral stage, DNA binding, E. expanded cotyledon stage, F. mature embryo stage, L. S4 four leaves visible, carpell, cauline leaf, circadian rhythm, de-etiolation, embryo, flower, gibberellin acid mediated signaling, hypocotyl, inflorescence meristem, leaf, leaf apex, leaf lamina base, leaf phloem, nucleus, pedicel, petal, petal differentiation and expansion stage, positive regulation of anthocyanin metabolic processes, protein binding, red or far-red light signal transduction pathway, response to red or far-red light, seed, sepal, shoot, shoot apex, signal transduction, stem, transcription factor activity, transcription regulator activity</td>
</tr>
</tbody>
</table>

AT1G09530.2

2. AT1G09570 | Light-sensitive cryptochrome, red-flavorelated phytochrome involved in the regulation of photomorphogenesis. It exists in two isoforms. | AT1G09570.1 | ELONGATED HYPOCOTYL 8, F14J9.23, F14J9.23, FAR RED ELONGATED 1, FAR RED ELONGATED 2, HYPOCOTYL 2, FHY2, FIB1, HY1 | 4. anthetic, 4. leaf senescence stage, C. globular stage, D. bilateral stage, E. expanded cotyledon stage, F. mature embryo stage, G. protein-cooperated photoreceptor activity, L. S2 two leaves visible, L. S4 four leaves visible, L. S6 six leaves visible, L. S8 eight leaves visible, L. S10 ten leaves visible, L. S12 twelve leaves visible, carpell, cauline leaf, cotyledon, cytoplasm, embryo, flower, gravitropism, hypocotyl, inflorescence meristem, leaf, leaf apex, leaf lamina base, leaf phloem, maize gametophyte, nuclear body, nucleus, pedicel, petal, petal differentiation and expansion stage, petiole, phototropism, protein histidine kinase activity, red light signaling pathway, red or far-red light photoreceptor activity, regulation of transcription, DNA-dependent, response to anesthetic, response to continuous far-red light stimulus by the high-irradiance response system, response to far-red light, response to very low fluence red light |
1. Finding a gene/protein: Exact name search
1. Finding a gene/protein: by sequence

Case 2: You have a gene / protein sequence and want to find the most similar gene/protein(s) in TAIR

- You have the nucleotide sequence of a gene from another organism and want to find a homolog
- You have the amino acid sequence of one Arabidopsis protein family member and want to find related family members

Use a sequence similarity search tool
BLAST
1. Finding a gene/protein: by sequence
2. Getting data from locus page

- locus
- other names
- description
- map detail
- annotations
- RNA data
  - Nucleotide sequence
  - Protein data
- polymorphisms
- germplasms
- phenotype
- external links
- publication
2. Getting data from locus page: description

- curated description
- computational description

**Locus: AT1G09570**

- **Date last modified**: 2003-05-02
- **TAIR Accession**: Locus:2012200
- **Gene Model**: AT1G09570.1
- **Gene Model Type**: protein_coding
- **Other names**: ELONGATED-HYPOCOTYL 8, FHY3, FHY3-23, FHY3-2X, FAR RED SENSITIVE 1, FAR RED ELONGATED HYPOCOTYL 2, FHY3, FHY3, HY5, PHYA, PHYTOCHROME A
- **Description**: Light-harvesting chloroplast red/far-red light photoreceptor involved in the regulation of photomorphogenesis. It exists in two inter-convertible forms: F1 and F2 (active and inactive) and functions as a dimer. The N terminus carries a single tetrapyrole chromophore, and the C-terminus is involved in dimerization. It is the sole photoreceptor mediating the FR high irradiance response (HIR). Major regulator in red-light induction of phototropism enhancement. Involved in the regulation of de-etiolation. Involved in gravitropism and phototropism. Requires FHY1 for nuclear accumulation.
2. Getting data from locus page: description

- computational description
### Locus: AT1G09570

**Date last modified:** 2003-05-02  
**TAIR Accession:** Locus:20123000  
**Representative Gene Model:** AT1G09570.1  
**Gene Model Type:** protein_coding  
**Other names:** ELONGATED-HYPOCOTYL 8, FHY3:23, FHY3:23, FAR RED (FR) RESPONSE FACTOR 1, HYPOCOTYL 2, PHY2, FRI, HY5, PHYA, PHOTOMORPHISM A  
**Description:** Light-table cytoplasmic red/far-red light photoreceptor involved in the regulation of photomorphogenesis. It exists in two photo-converting forms: Pr and Prh (active) and functions as a dimmer. The N terminus carries a single tetrapyrrole chromophore, and the C terminus is involved in dimerization. It is the sole photoreceptor mediating the FR high irradiance response (HIR). Major regulator in red-light induction of phototropic enhancement. Involved in the regulation of de-etiolation. Involved in gravitropism and phototropism. Requires FHY1 for nuclear accumulation.

**Other Gene Models:** AT1G09570.2  
(splice variant)

### Map Detail

<table>
<thead>
<tr>
<th>Gene Model</th>
<th>3096a</th>
<th>3097a</th>
<th>3098a</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G09570.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT1G09570.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Annotations

<table>
<thead>
<tr>
<th>Category</th>
<th>Relationship Type</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO Biological Process</td>
<td>involved in</td>
<td>gravitropism, phototropism, regulation of transcription, DNA-dependent, response to arsenic, red light signaling pathway, response to very low light, response to continuous far red light stimulus by the high-irradiance response system</td>
</tr>
<tr>
<td>GO Cellular Component</td>
<td>related to</td>
<td>cytoplasmic, nuclear body, nucleus</td>
</tr>
<tr>
<td>GO Molecular Function</td>
<td>has</td>
<td>G-protein coupled photoreceptor activity, protein histidine kinase activity, signal transduction activity, red or far-red light photoreceptor activity</td>
</tr>
<tr>
<td>Growth and Developmental Stages</td>
<td>expressed during</td>
<td>petal differentiation and expansion stage, 4 anthocyanin, 4 leaf senescence stage, C globular stage, E expanded cotyledon stage, F mature embryo stage, LP.03 two leaves visible, LP.04 four leaves visible, LP.08 six leaves visible, LP.09 eight leaves visible, LP.10 ten leaves visible, LP.12 twelve leaves visible</td>
</tr>
<tr>
<td>Plant structure</td>
<td>expressed in</td>
<td>male gametophyte, seed, embryo, cotyledon, hypocotyl, root, shoot, flower, stamen, carpel, sepal, petal, inflorescence meristem, pedicle, leaf, leaf apex, petiole, stem, cauline leaf, shoot apex, leaf lamina base, leaf lamina tip</td>
</tr>
</tbody>
</table>

**free text annotations**

**controlled vocabulary annotations**
2. Getting data: GO / PO annotations

<table>
<thead>
<tr>
<th>Locus</th>
<th>Gene or Locus</th>
<th>Relationship Type</th>
<th>Keyword</th>
<th>Keyword Category</th>
<th>Evidence Code</th>
<th>Evidence Description</th>
<th>Evidence URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G08670</td>
<td>AT1G08670</td>
<td>involved in:</td>
<td>proteasome</td>
<td>biological process</td>
<td>inferred from mutant phenotype: analysis of insertional T-DNA lines</td>
<td>The Arabidopsis Information Resource 2005-09-28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT1G08670</td>
<td>AT1G08670</td>
<td>involved in:</td>
<td>transcription, DNA-dependent</td>
<td>biological process</td>
<td>inferred from sequence of structural similarity alignment with other TFs</td>
<td>The Institute for Genomic Research 2003-09-15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Getting data: getting the plant lines you want

- Are there reported mutant phenotypes?
- Can I get the seeds?
2. Getting data: expression

<table>
<thead>
<tr>
<th>Category</th>
<th>Relationship Type</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological Process</td>
<td>involved in</td>
<td>gravitropism, phototropism, regulation of transcription, DNA-dependent, response to arsenic, red light signaling pathway, response to very low fluence red light stimulus, response to continuous far red light stimulus by the high-irradiance response system</td>
</tr>
<tr>
<td>Cellular Component</td>
<td>related to</td>
<td>response to far red light</td>
</tr>
<tr>
<td>Molecular Function</td>
<td>located in</td>
<td>cytoplasm, nuclear bodies, nucleus</td>
</tr>
<tr>
<td>Growth and Development Stages</td>
<td>expressed during</td>
<td>petal differentiation and expansion stage, 4 anther, 4 leaf senescence stage, C globular stage, E expanded cotyledon stage, F mature embryo stage, LP.02 two leaves visible, LP.04 four leaves visible, LP.06 six leaves visible, LP.08 eight leaves visible, LP.10 ten leaves visible, LP.12 twelve leaves visible</td>
</tr>
<tr>
<td>Plant structure</td>
<td>expressed in</td>
<td>D bilateral stage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>male gametophyte, seed, embryos, cotyledon, hypocotyl, root, shoot, flower, stamen, carpel, sepal, petal, inflorescence meristem, pedicel, leaf, leaf apex, petiole, stem, cauline leaf, shoot apex, leaf lamina base, leaf whorl</td>
</tr>
</tbody>
</table>
2. Getting data: more expression data via external links

What else is known about this gene’s expression pattern?

External Link

- NCBI Gene
- NCBI Exome Viewer
- miRNA Database
- T-DNA Express
- SNP Viewer
- ASTRAL (Protein Interaction Database at EBI)
- ASTRAL
- Plant Proteome
- Plant Proteome Database
- T-DNA Express
- SRA SRA Viewer
- Gene Expression Viewer
2. Getting data: more expression data via external links
2. Getting data: gene model page
2. Getting data: model-specific data
Sometimes, one gene isn’t enough

- Scientists often want to work with more than one gene or protein that are related through some common feature.
- TAIR offer some basic tools to create and/or enhance these customized data sets.
3. Customized data sets: case studies

You have mapped a mutation that disrupts flower development to a region of Chromosome 1.

Q: What genes in the mapping interval are good candidates?

- Get a list of all the genes in the mapping interval involved in “flower development”
- Download all the associated GO and PO terms for the candidate genes
3. Customized data sets: Search: Genes

- Get a list of all the genes involved in "flower development".
3. Customized data sets: Search: Genes

<table>
<thead>
<tr>
<th>TAIR Accession</th>
<th>Locus</th>
<th>Gene Model</th>
<th>Gene Type</th>
<th>Description</th>
<th>Other Name(s)</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene:20071482.1</td>
<td></td>
<td>AT1G69490</td>
<td>protein_coding</td>
<td>Encodes a member of the HAC transcriptional activator protein similar to G-coupled receptors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2007709.1</td>
<td></td>
<td>AT1G45270</td>
<td>protein_coding</td>
<td>Serin-threonine kinase targeted to chloroplasts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2008069.1</td>
<td></td>
<td>AT1G1870</td>
<td>protein_coding</td>
<td>ATRIA (ATRIAL SIGNALING 1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2008665.1</td>
<td></td>
<td>AT1G58630.1</td>
<td>protein_coding</td>
<td>Encodes ATP1, a histone H3A protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2008621.1</td>
<td></td>
<td>AT1G44600.1</td>
<td>protein_coding</td>
<td>Encodes a mitogen-activated protein kinase kinase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2003775.1</td>
<td></td>
<td>AT1G51660.1</td>
<td>protein_coding</td>
<td>ASK1.4 encodes a protein similar to JAR1, a blue light receptor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2004659.1</td>
<td></td>
<td>AT1G04400.1</td>
<td>protein_coding</td>
<td>One of the three genes encoding subunit A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2004650.1</td>
<td></td>
<td>AT1G62500.1</td>
<td>protein_coding</td>
<td>Encodes an ubiquitin-specific protease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2001998.1</td>
<td></td>
<td>AT1G50720.1</td>
<td>protein_coding</td>
<td>One of the three Arabidopsis genes encoding subunit A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2002460.1</td>
<td></td>
<td>AT1G70710.1</td>
<td>protein_coding</td>
<td>Encodes ATP5A-2, one of the three Arabidopsis genes encoding subunit A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2002460.1</td>
<td></td>
<td>AT1G32160.1</td>
<td>protein_coding</td>
<td>Encodes a member of class I knotted-like homebox gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene:2002460.1</td>
<td></td>
<td>AT1G32160.1</td>
<td>protein_coding</td>
<td>Encodes a member of class I knotted-like homeobox gene</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Customized data sets: Search: Genes

Gene search

Function: identify a list of genes (that have certain function) (located in a specific region) (additional parameters)
3. Customized data sets

You work on a transcription factor, you have a mutant of this TF gene, you have obtained a list of genes whose expression is altered in this mutant

Q: What’s the cis-element for this TF?

- Get upstream “promoter” sequences for these genes
- Search for over-represented DNA sequences in “promoters”

Q: What are the potential target genes that might be regulated by this TF?

- Find all genes with similar DNA sequences in their upstream regions
- Identify all domains found in the encoded proteins
3. Customized data sets:
Download: Bulk Data Retrieval: Sequences

- Get upstream “promoter” sequences
3. Customized data sets: Tools: Motif Analysis

- Search for over-represented or prevalent DNA sequences in “promoters”
  - Use the Motif Analyzer in TAIR to identify common 6-mers
3. Customized data sets: Tools: Patmatch
3. Customized data sets:
Download: Bulk Data Retrieval: Proteins

- Identify all of the other domains found in those proteins.
3. Customized data sets: summary

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Dataset Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>bulk data retrieval: sequence</td>
<td>list of AGI codes</td>
<td>upstream sequence</td>
</tr>
<tr>
<td>motif analyzer</td>
<td>list of AGI codes</td>
<td>DNA motif in promoter</td>
</tr>
<tr>
<td>Patmatch</td>
<td>motif (DNA or protein)</td>
<td>list of sequences</td>
</tr>
<tr>
<td>Bulk data retrieval: protein</td>
<td>list of AGI codes</td>
<td>protein features</td>
</tr>
</tbody>
</table>
4. Analyze data sets: Functional categorization

- Classify your list of genes/proteins using GO annotations.

[Image of a webpage with options for GO annotation search, functional categorization, and download.]

- Functional categorization allows you to classify gene expression profiles within broad functional categories based on the high level terms in the GO hierarchy.

[Another image showing a search interface with uploaded identifiers and options for whole genome categorization.]
GO annotation search, functional categorization and download

Paste locus identifiers (such as Atg01230) into the textbox and press one of the submit buttons below. The identifiers have to be separated by tabs, commas, carriage returns or spaces. Alternatively, you can upload a file, same formatting as for the textbox.

Clicking on Get all GO annotations will display in detail all the GO annotations done to your set of genes. Clicking on Functional categorization will group the genes into broad functional categories based on the high level terms in GO hierarchy.

You may download the whole genome GO annotations from TAIR FTP site.

Whole Genome Categorization

Get:
Whole Genome Categorization

GO Annotations

Locus identifiers:

Upload file:
Browse...

Output Options:

Select output type:
HTML, Text
(Please note that if more than 1000 loci are entered, only text output will be given)
How to effectively use the tools and resources at TAIR to enhance your research

Part III: Tools for analyzing the results of experiments on an Arabidopsis metabolic map

A.S.Karthikeyan
Everyone is studying metabolism

- Many biological processes connect to metabolism
  - Drought tolerance – changes in osmolyte concentrations
  - Hormone signaling – biosynthesis and degradation of hormones
  - Photosynthesis – chlorophyll production and ROS scavenging
  - Translation – amino acid biosynthesis and riboswitching
  - Plant defense – phytoalexin synthesis
  - Your favorite process . . .

- AraCyc can help you find these connections!
  - Arabidopsis Metabolic Encyclopedia
  - Database of metabolic pathways found in Arabidopsis
    - www.arabidopsis.org/biocyc/ (TAIR)
    - www.plantcyc.org/ARA (Plant Metabolic Network)
How to Access AraCyc

- AraCyc – **Arabidopsis Metabolic Encyclopedia**
  - Database of metabolic pathways found in Arabidopsis

- Accessible from:
  - TAIR – The Arabidopsis Information Resource
    - www.arabidopsis.org
How to Access AraCyc

- AraCyc – Arabidopsis Metabolic Encyclopedia
  - Database of metabolic pathways found in Arabidopsis

- Accessible from:
  - PMN – Plant Metabolic Network
    - www.plantcyc.org
## AraCyc 5.0

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathways</td>
<td>332</td>
</tr>
<tr>
<td>Compounds</td>
<td>2312</td>
</tr>
<tr>
<td>Reactions</td>
<td>1928</td>
</tr>
<tr>
<td>Genes (in pathways)</td>
<td>1970</td>
</tr>
<tr>
<td>Citations</td>
<td>2543</td>
</tr>
</tbody>
</table>
AraCyc Pathway pages

All items can be clicked on to obtain more information

+ Additional curated information
AraCyc Pathway pages

Classification

Summary:

Pathway information:

- The first step in choline biosynthesis is the direct decarboxylation of serine to ethanolamine, which is catalyzed by a serine decarboxylase unique to plants.

- Ethanolamine is widely recognized as the entrance compound to choline biosynthesis.

- The pathway variant displayed (nucleotide pathway) represents the biosynthetic route as found in diverse plant families. The synthesis of choline from ethanolamine may take place at three parallel pathways, where three consecutive N-methylation steps are carried out either on free-bases or phospho-bases.

- The synthesis of intermediates on both the phospho-base and phosphatidyl-base level includes the nucleotide pathway via CDP-phosphoethanolamine and the methylation pathway. However, it has been pointed out that the synthesis of phosphatidylethanolamine and phosphatidylcholine is characterized by a high degree of interaction and function on the various levels of arising intermediates. Consequently, it has been assumed that the reactions embedded in the nucleotide and methylation pathways may be two characteristics of one overall phosphoethanolamine pathway for the formation of phosphatidylcholine in plants.

- The release of choline from the different pathway levels is also species-specific. Phosphocholine can either be directly dephosphorylated to release choline as observed in spinach or incorporated into phosphatidylcholine with the subsequent release of choline, as is observed in tobacco. The latter reaction has been shown to be specifically catalyzed by phospholipase D in certain beans. Although a well-defined physiological role of phospholipase D for lipid metabolism, its substrates are often hard to distinguish from other cellular functions.

- The remaining enzymes involved in this pathway, phosphoethanolamine cytidylyltransferase and CDP-ethanolamine phosphotransferase, cover a broader spectrum of substrates. This may be beneficial to process the heterogeneous mixture of possible substrates but it also indicates that the pathway flux is probably controlled more upstream.

Superpathways:

- superpathway of choline biosynthesis

Variants:

- choline biosynthesis [choline biosynthesis]

Unification Links:

- MetaCyc: P0015

Legend for Pathway Diagram:

If an enzyme name is shown in bold, there is experimental evidence for this enzymatic activity.

References:

OMICS Viewer:

Analyzing the results of experiments
on an Arabidopsis metabolic map
OMICS Viewer to the rescue . . .

- Overlay experimental data on a metabolic map
  - Transcript data (for enzymes)
  - Proteomic data (for enzymes)
  - Metabolomic data

- Use multiple data sources
  - Your data
  - Publicly available data
    - Gene Expression Omnibus
    - NASC Proteomics Database
    - NSF2010 Metabolomics
    - Many more . . .

- Generate new testable hypotheses about your favorite
  - gene
  - metabolite
  - biological process
  - etc.
Addressing common research conundrums . . .

– Compare transcript levels / protein levels / metabolite levels in wild-type and mutant plants using the OMICS viewer

– Look for “hidden” perturbations in metabolism

– Use the OMICS viewer to compare data from WT and mutant plants

– Download publicly available data sets related to the biological process

– Scan for affected areas of metabolism

– Use the OMICS Viewer to quickly identify metabolic pathways that are up- or down-regulated

– And many more . . .
related pathways are grouped together
OMICS Viewer Overview

- COMPOUNDS (icons)
- TRANSCRIPTS or PROTEINS (lines)
Viewing gene/protein families

Expression levels for individual genes/proteins can be displayed.

Caution: Only the highest expression level is shown using the single line on the diagram.

Changes in the levels of homologs are hidden in the low resolution map.
Getting more information

Clicking on a pathway brings you to a pathway detail page
AraCyc Pathway pages

All items can be clicked on to obtain more information

+ Additional curated information
OMICS Viewer in action

- Multiple data sets can be entered using the same input file

- An animation can show changes in data sets:
  - wild type / mutant a / mutant b / mutant c / . . .
  - time points 0, 1, 2, 3, . . .
  - compound concentration x, y, z, . . .

Suberin Biosynthesis

Wild type  mutant A  mutant B
OMICS Viewer data inputs

- **Input file**
  - tab-delimited
  - pre-analyzed/cleaned data
  - Identifiers in first column
- **Genes**
  - AGI locus codes (e.g. At2g46990)
- **Enzymes**
  - AGI locus codes (e.g. At2g46990)
  - EC number (e.g. 2.4.2.18)
- **Compounds**
  - Enzyme name (e.g. arogenate dehydrogenase)
  - Compound name
  - Data in additional column(s)
    - absolute or relative values
- Sample file and tutorial available at website
OMICS Viewer data upload

TAIR - www.arabidopsis.org

AraCyc - www.arabidopsis.org/biocyc
OMICS Viewer data upload

PMN – www.plantcyc.org
OMICS Viewer data upload

• Select species

• Upload data file

• Enter appropriate parameters

• GO!
New ideas to pursue . . .

Wild type

Mutant with “no phenotype”

Mutant with a metabolic phenotype

The transcript levels of several enzymes in the *lysine biosynthesis pathway* are reduced ~4-fold!

New hypotheses and experiments
Acknowledgements

AraCyc and TAIR

PI and Co-PI
Eva Huala
Sue Rhee

Curators:
- Peifen Zhang
- Tanya Berardini
- David Swarbreck
- Kate Dreher
- Rajkumar Sasidharan

Tech Team:
- Bob Muller
- Larry Ploetz
- Raymond Chetty
- Anjo Chi
- Vanessa Kirkup
- Cynthia Lee
- Tom Meyer
- Shanker Singh
- Chris Wilks

Metabolic Pathway Software:
- Peter Karp and SRI group
TAIR Workshop II

TAIR, PMN, SGN & Gramene Joint Workshop

When: **Tonight at 7:00 PM**

Where: **318 A & B**