Biocuration:

Helping Researchers Harness the Data Explosion at TAIR and the Plant Metabolic Network

Kate Dreher

curator

TAIR/PMN

Department of Plant Biology
Carnegie Institution for Science
Stanford, California

kadreher@stanford.edu

Overview

- Biological data **explosion**
- Biocurators want to help!
- Biocuration practices and resources at two plant databases
  - The Arabidopsis Information Resource
  - The Plant Metabolic Network
- Request for *your* help!

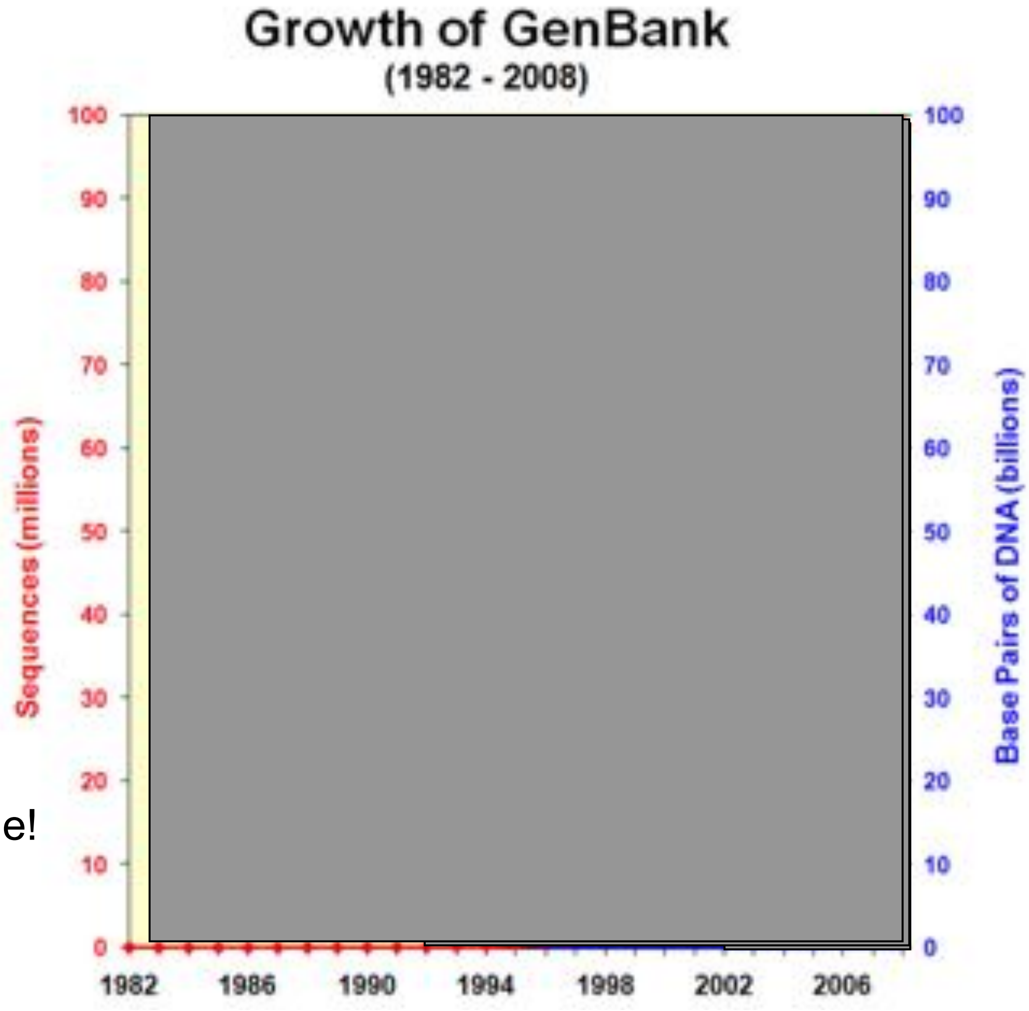
Growth of biological data

- Over time biological data increases in
  - Quantity
    - Methods improve
    - Costs decrease

Growth of biological data

- Nucleotide sequences
  - Number of sequences in 1982
    - **606**
  - Number of sequences in 1992:
    - **78,608**
  - Number of sequences in 2002:
    - **22,318,883**
  - Number of sequences in 2008:
    - **98,868,465**
  - And, the acceleration may continue!

Source: National Center for Biotechnology Information (NCBI)



http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html

Growth of biological data

- Over time biological data increases in
  - Complexity
    - Protein data
      - Primary sequence
      - 3D structure
      - Subcellular localization
      - Rate of degradation
      - Enzymatic activity properties
      - Post-translational modification
        - Phosphorylation
        - Prenylation
        - Methylation
        - Ubiquitination

Is it static or dynamic?

What stimuli cause it to change?

By how much?

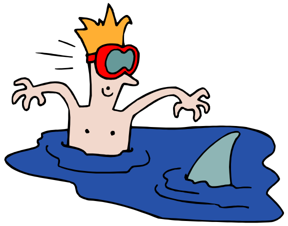
The waves of data keep mounting!

DNA, RNA
data

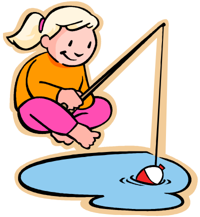
protein
data

metabolite
data

phenotype
data



Exploring the sea of biological data

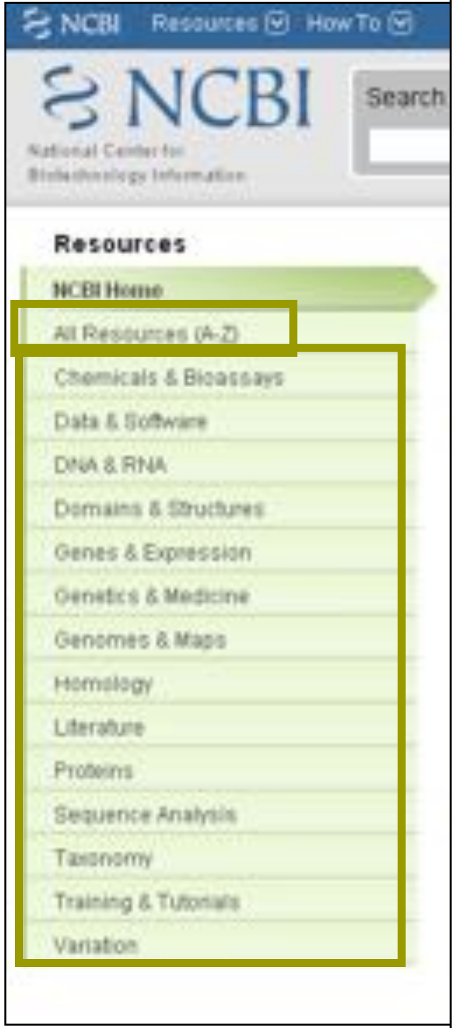


- Primary data source
  - Articles published in peer-reviewed journals
  - Over 18 million available through PubMed by 2008!
    - NOT a comprehensive set; many journals are missing
- Answering scientific questions
  - Specific focus:
    - Find ***every single piece of information ever discovered*** about my favorite gene – XYZ1 – to figure out exactly what it does
  - Broad search:
    - Compare the protein sequence of ***every single transcription factor ever discovered*** in a prokaryote or a eukaryote to study the evolution of nuclear-localization signals
  - ***How do you collect these data from ALL of the relevant research articles?***
    - Data repositories . . . staffed by **biocurators** . . . try to help!
      - Computer scientists and bioinformaticians contribute to these efforts as well!

Global data repositories / databases

- Centralized data hubs
    - Many data types
    - Many species
    - Asia
        - Several in Japan, e.g. RIKEN, China is adding new ones
    - Europe
        - European Bioinformatics Institute (EBI)
    - USA
        - National Center for Biotechnology Information (NCBI)

Global data re



NCBI Resources How To

NCBI
National Center for Biotechnology Information

Search

Resources

NCBI Home

All Resources (A-Z)

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

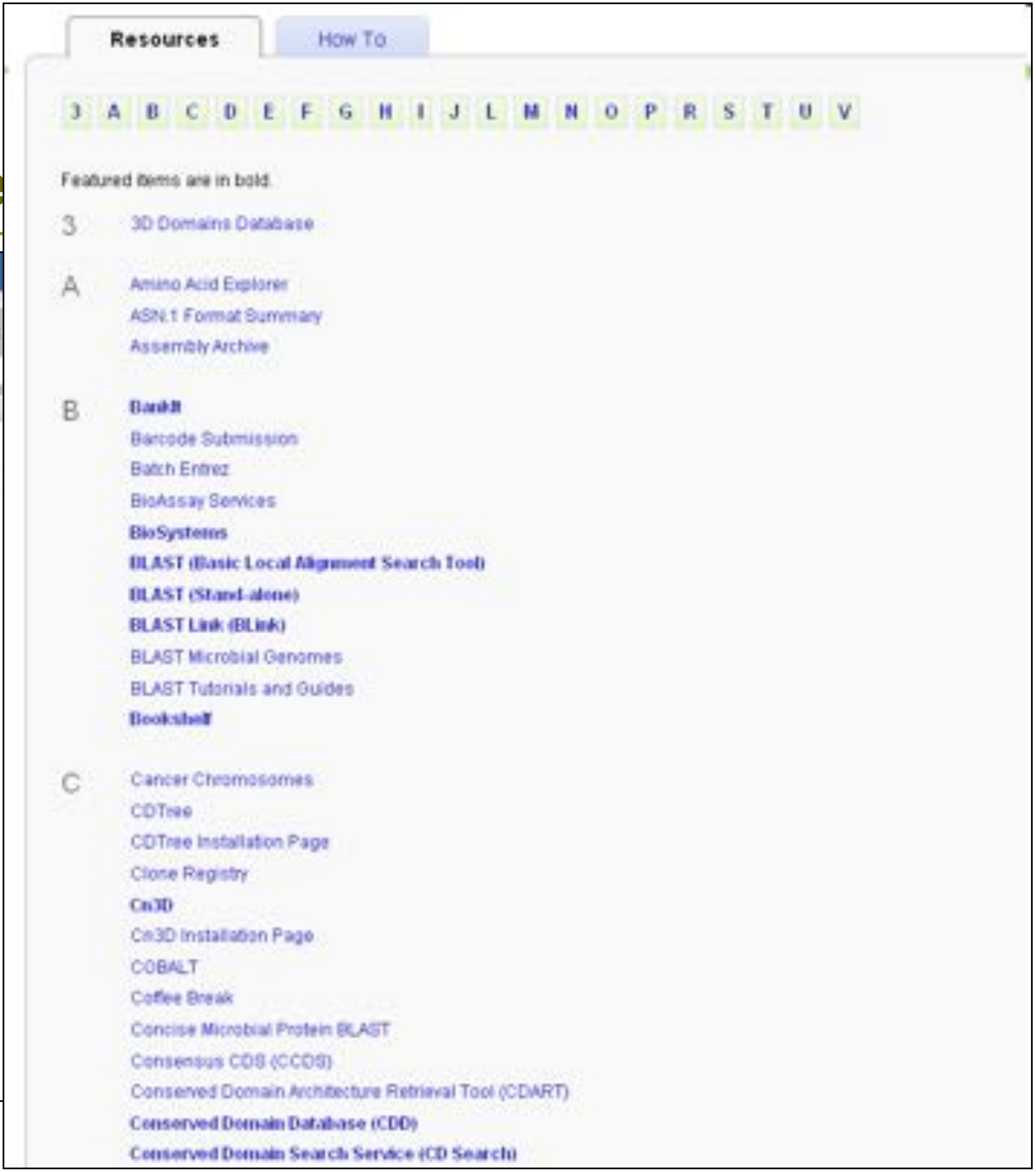
Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation



Resources How To

3 A B C D E F G H I J L M N O P R S T U V

Featured items are in bold.

3
- 3D Domains Database

A
- Amino Acid Explorer
- ASN.1 Format Summary
- Assembly Archive

B
- **BankIt**
- Barcode Submission
- Batch Entrez
- BioAssay Services
- **BioSystems**
- **BLAST (Basic Local Alignment Search Tool)**
- **BLAST (Stand-alone)**
- **BLAST Link (BLink)**
- BLAST Microbial Genomes
- BLAST Tutorials and Guides
- **Bookshelf**

C
- Cancer Chromosomes
- CDTree
- CDTree Installation Page
- Clone Registry
- **Cn3D**
- Cn3D Installation Page
- COBALT
- Coffee Break
- Concise Microbial Protein BLAST
- Consensus CDS (CCDS)
- Conserved Domain Architecture Retrieval Tool (CDART)
- **Conserved Domain Database (CDD)**
- **Conserved Domain Search Service (CD Search)**

Specialized data repositories / databases

- Model organism databases (MODs)
  - Mouse Genome Informatics (MGI)
  - Flybase (*Drosophila*)
  - Saccharomyces Genome Database (SGD) (yeast)
  - **The Arabidopsis Information Resource (TAIR)**

- Topical databases
  - Worldwide Protein Data Bank (3D structures)
  - miRbase (microRNAs)
  - **Plant Metabolic Network (PMN) (metabolic / biochemical pathways)**

Roles of biocurators at data repositories

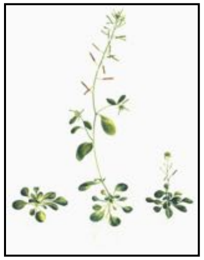
- Organize and process raw data
  - Assign unique stable identifiers for nucleotide sequences submitted by researchers
- Review and improve data to generate **curated** data sets
  - Manually correct errors in raw nucleotide sequences to make RefSeq gene structures
- Develop tools for accessing data
  - Provide a protein interaction viewer
- Train users
  - Present at conferences and universities
- ***Try to help researchers harness the data explosion!***
  - ***TAIR***
  - ***Plant Metabolic Network***

Introduction to TAIR

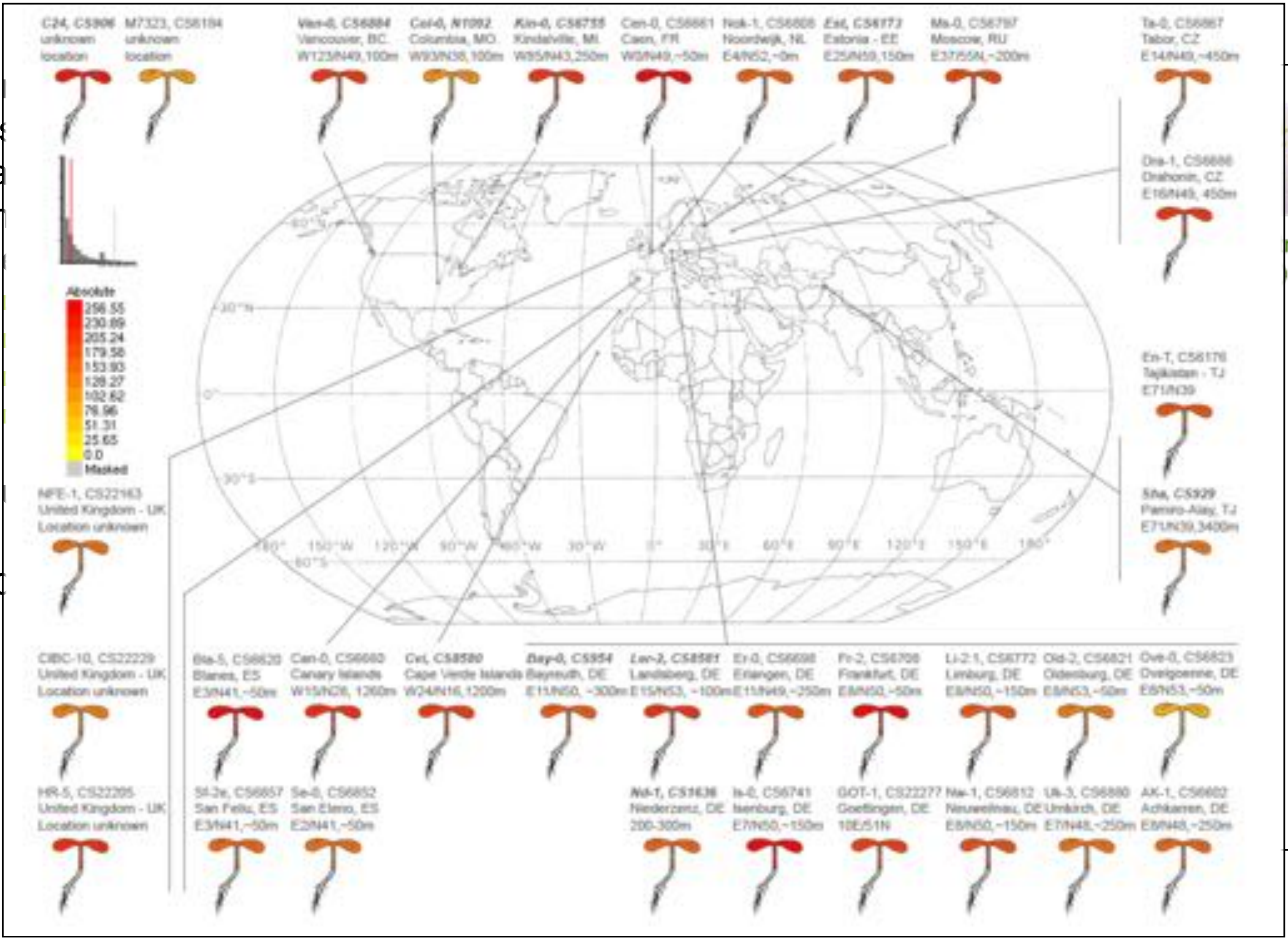
- TAIR = **T**he **A**rabidopsis **I**nformation **R**esource
- Why Arabidopsis?
- What does TAIR do?
- What can you do with TAIR?

Arabidopsis

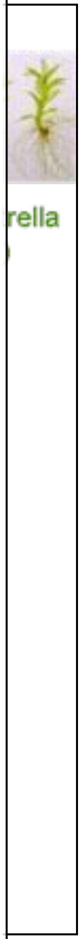
Introduction to Arabidopsis



- Basic
  - “s
  - als
  - ca
  - an
  - m
  - fo
- Why c



C24, CS906 unknown location
M7323, CS8184 unknown location
Van-0, CS6884 Vancouver, BC. W123/N49,100m
Col-0, N1092 Columbia, MO W93/N38,100m
Kin-0, CS6755 Kindalville, MI. W85/N43,250m
Cen-0, CS6661 Caen, FR W0/N49,~50m
Nok-1, CS6808 Noordwijk, NL E4/N52,~0m
Est, CS6173 Estonia - EE E25/N59,150m
Ms-0, CS6797 Moscow, RU E37/55N,~200m
Ta-0, CS6867 Tabor, CZ E14/N49,~450m
Dra-1, CS6686 Drahonin, CZ E16/N49, 450m
En-T, CS6176 Tajikistan - TJ E71/N39
Sha, CS929 Pamiro-Alay, TJ E71/N39,3400m
NFE-1, CS22163 United Kingdom - UK Location unknown
Absolute
256.55
230.89
205.24
179.58
153.93
128.27
102.62
76.96
51.31
25.65
0.0
Masked
60°N
30°N
0°
30°S
60°S
180°
150°W
120°W
90°W
60°W
30°W
0°
30°E
60°E
90°E
120°E
150°E
180°
CIBC-10, CS22229 United Kingdom - UK Location unknown
Bla-5, CS6620 Blanes, ES E3/N41,~50m
Can-0, CS6660 Canary Islands W15/N28, 1260m
Cvi, CS8580 Cape Verde Islands W24/N16,1200m
Bay-0, CS954 Bayreuth, DE E11/N50, ~300m
Ler-2, CS8581 Landsberg, DE E15/N53, ~100m
Er-0, CS6698 Erlangen, DE E11/N49,~250m
Fr-2, CS6708 Frankfurt, DE E8/N50,~50m
Li-2:1, CS6772 Limburg, DE E8/N50,~150m
Old-2, CS6821 Oldenburg, DE E8/N53,~50m
Ove-0, CS6823 Ovelgoenne, DE E8/N53,~50m
HR-5, CS22205 United Kingdom - UK Location unknown
St-2e, CS6857 San Feliu, ES E3/N41,~50m
Se-0, CS6852 San Elmo, ES E2/N41,~50m
Nd-1, CS1636 Niederzenz, DE 200-300m
Is-0, CS6741 Isenburg, DE E7/N50,~150m
GOT-1, CS22277 Goettingen, DE 10E/51N
Nw-1, CS6812 Neuweilnau, DE E8/N50,~150m
Uk-3, CS6880 Umkirch, DE E7/N48,~250m
AK-1, CS6602 Achkarren, DE E8/N48,~250m



Arabidopsis offers some advantages

- “Good” genome
  - very small: 125 Mb - ~27,000 genes
  - diploid
  - 5 haploid chromosomes
  - fewer/smaller regions of repetitive DNA than many plants
- Quite *easily* transformable with *Agrobacterium*
  - NO tissue culture required
- Inertia!
  - A group of scientists lobbied for Arabidopsis
  - The genome was sequenced (2000)
  - **MANY resources have been developed**

Arabidopsis research can be applied to “real plants”

- Over-expression of the *hardy* gene from Arabidopsis can improve water use efficiency in **rice** (*Karaba 2007)*

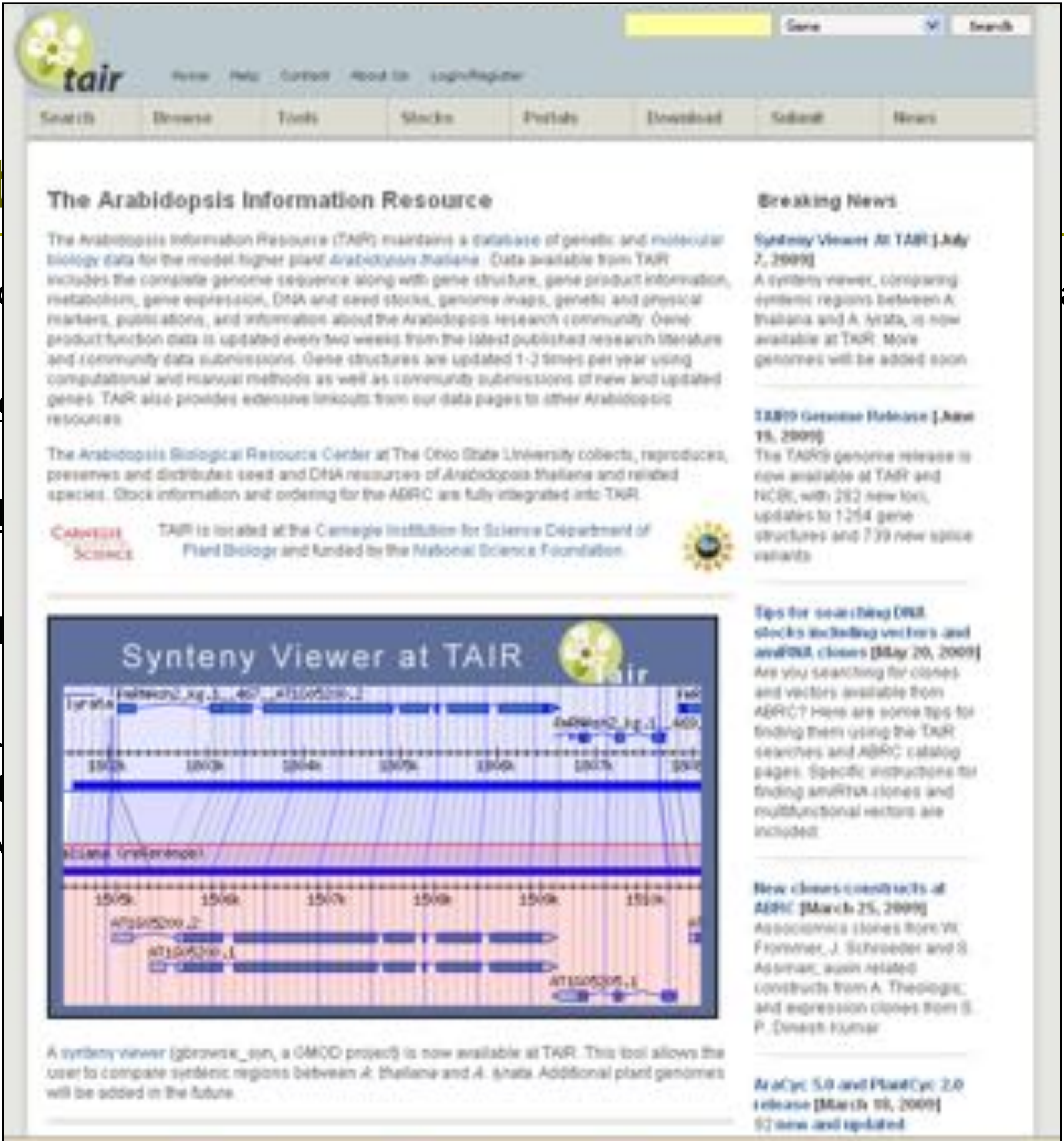
- A high throughput screen performed using **castor bean** cDNAs expressed in Arabidopsis found three cDNAs that increase hydroxy fatty acid levels in seeds (*Lu 2006*)

- These experiments and many more benefit from the work of curators trying to help harness the Arabidopsis data explosion . . .

  - ~2400 articles discussing Arabidopsis in PubMed per year!

What

- Curat at directors
- TAIR
- TAIR
- TAIR
  - Fu
  - St
  - Av



The Arabidopsis Information Resource

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

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, 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.

Carnegie Science

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

A synteny viewer (gbrowse_syn, a GMOD project) is now available at TAIR. This tool allows the user to compare syntenic regions between A. thaliana and A. lyrata. Additional plant genomes will be added in the future.

Breaking News

Synteny Viewer At TAIR [July 7, 2009]
A synteny viewer, comparing syntenic regions between A. thaliana and A. lyrata, is now available at TAIR. More genomes will be added soon.

TAIR9 Genome Release [June 19, 2009]
The TAIR9 genome release is now available at TAIR and NCBI, with 282 new loci, updates to 1254 gene structures and 739 new splice variants.

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.

New clones/constructs at ABRC [March 25, 2009]
Associomics clones from W. Frommer, J. Schroeder and S. Assman; auxin related constructs from A. Theologis; and expression clones from S. P. Dinesh Kumar.

AraCyc 5.0 and PlantCyc 2.0 release [March 18, 2009]
52 new and updated

Structural curation at TAIR

- Structural curators try to answer the question:

  What are ALL of the genes in Arabidopsis?

  - Use many types of data
    - ESTs
    - full-length cDNAs
    - peptides
    - orthology
    - RNASeq data**

  - Determine gene coordinates and features
    - Establish intron, exon, and UTR boundaries
    - Add alternative splice variants
    - Classify genes
      - protein coding
      - miRNA
      - pseudogene

Structural curation at TAIR

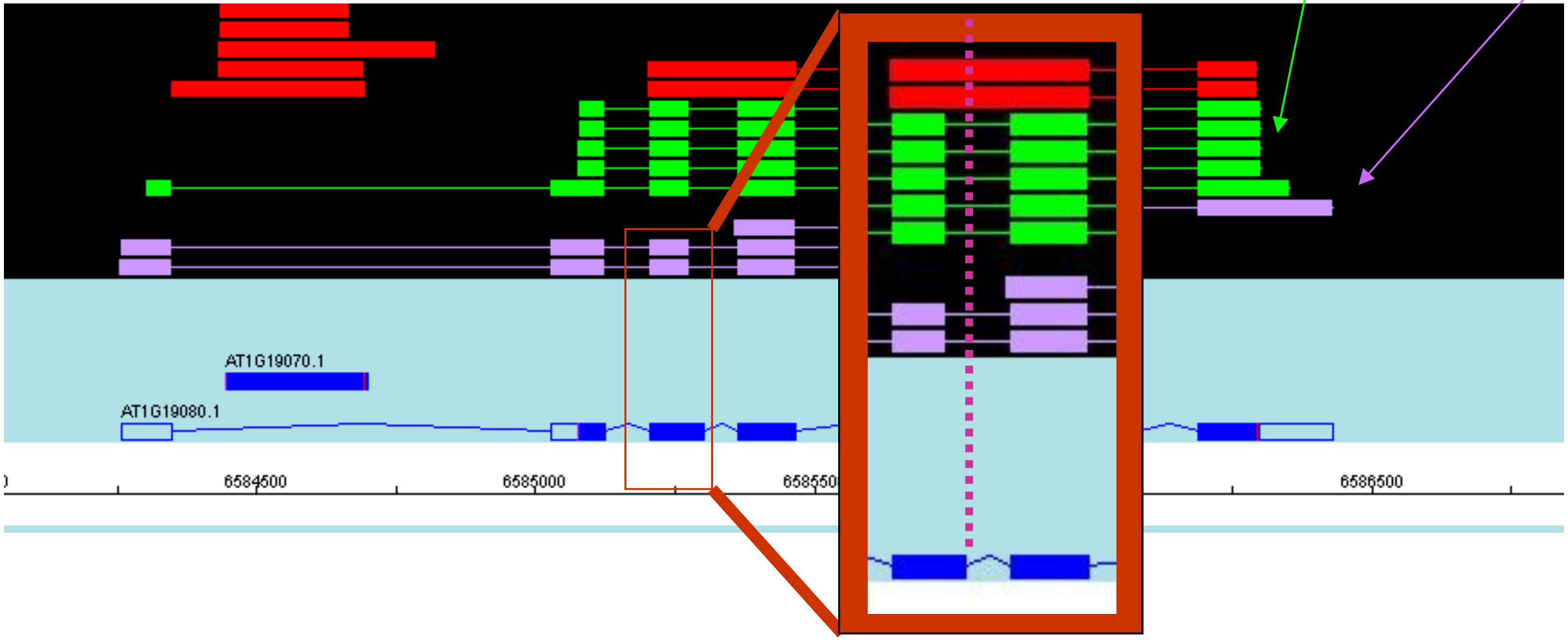
- Even though the genome was sequenced in 2000 . . .
- . . . the work goes on!

  - TAIR9 – released June 2009
  - 282 new loci and 739 new splice variants

  - TAIR10 – on its way
  - 126 novel genes
  - 1182 updated genes
  - 5885 new splice variants added (18% of all loci)

Structural curation at TAIR

- Apollo is a program to assist with structural curation



Protein similarity

cDNAs

ESTs

Functional AIR

- Functional cur questions:
  - *What does e bidopsis do?*
  - *When and w*
  - We hope tha orm research in other plants
- Functional cur *d vocabularies*
  - Allow cross
  - TAIR curat d agree upon common terms

achene
berry
capsule
caryopsis
circumcissile
capsule
cypsela
drupe
follicle
grain
kernel
legume
loculicidal capsule
lomentum
nut
pod
pome
poricidal capsule
schizocarp
septicidal capsule
septifragal capsule
silique

The seed-bearing st
formed from the ova

→ **FRUIT**

Plant Ontology:
Structure:
PO:0009001

Functional curation at TAIR

Catalysis of the reaction:
IAA + UDP-D-glucose = indole-3-acetyl-beta-1-D-glucose + UDP

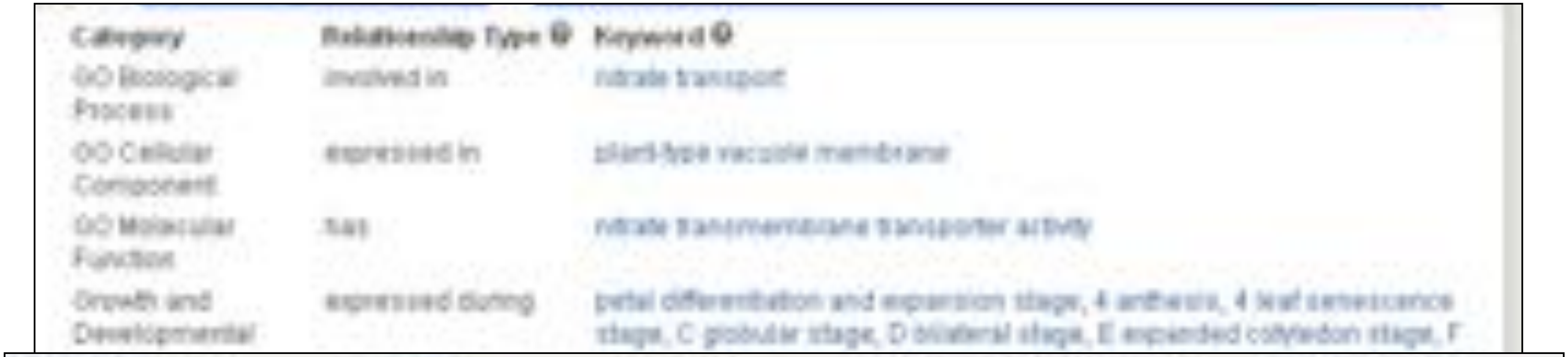
IAA-Glu synthetase activity
IAA-glucose synthase activity
IAGlu synthase activity
indol-3-ylacetylglucose synthase activity
UDP-glucose:(indol-3-yl)acetate beta-D-glucosyltransferase activity
UDP-glucose:indol-3-ylacetate glucosyl-transferase activity
UDP-glucose:indol-3-ylacetate glucosyltransferase activity
UDPG-indol-3-ylacetyl glucosyl transferase activity
UDPglucose:indole-3-acetate beta-D-glucosyltransferase activity
uridine diphosphoglucose-indoleacetate glucosyltransferase activity

↓

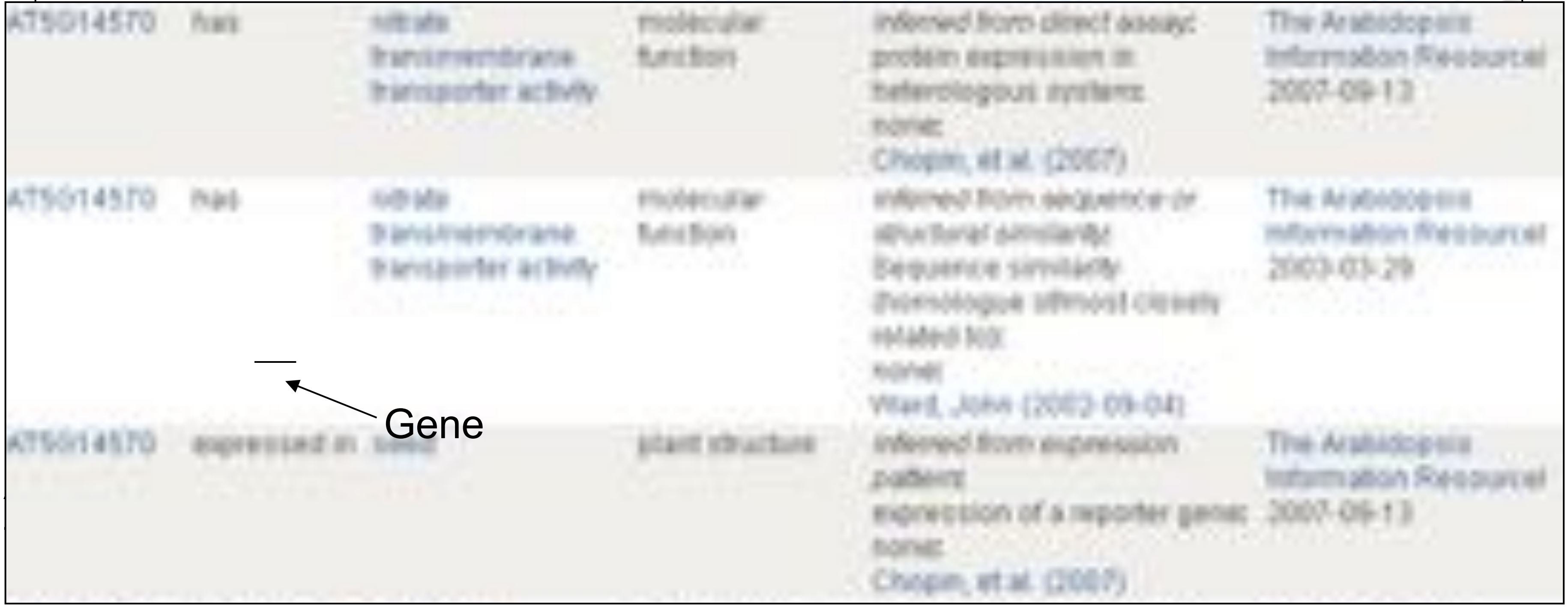
indole-3-acetate beta-glucosyltransferase activity

Gene Ontology:
Molecular function:
GO:0047215

Functional curation at TAIR



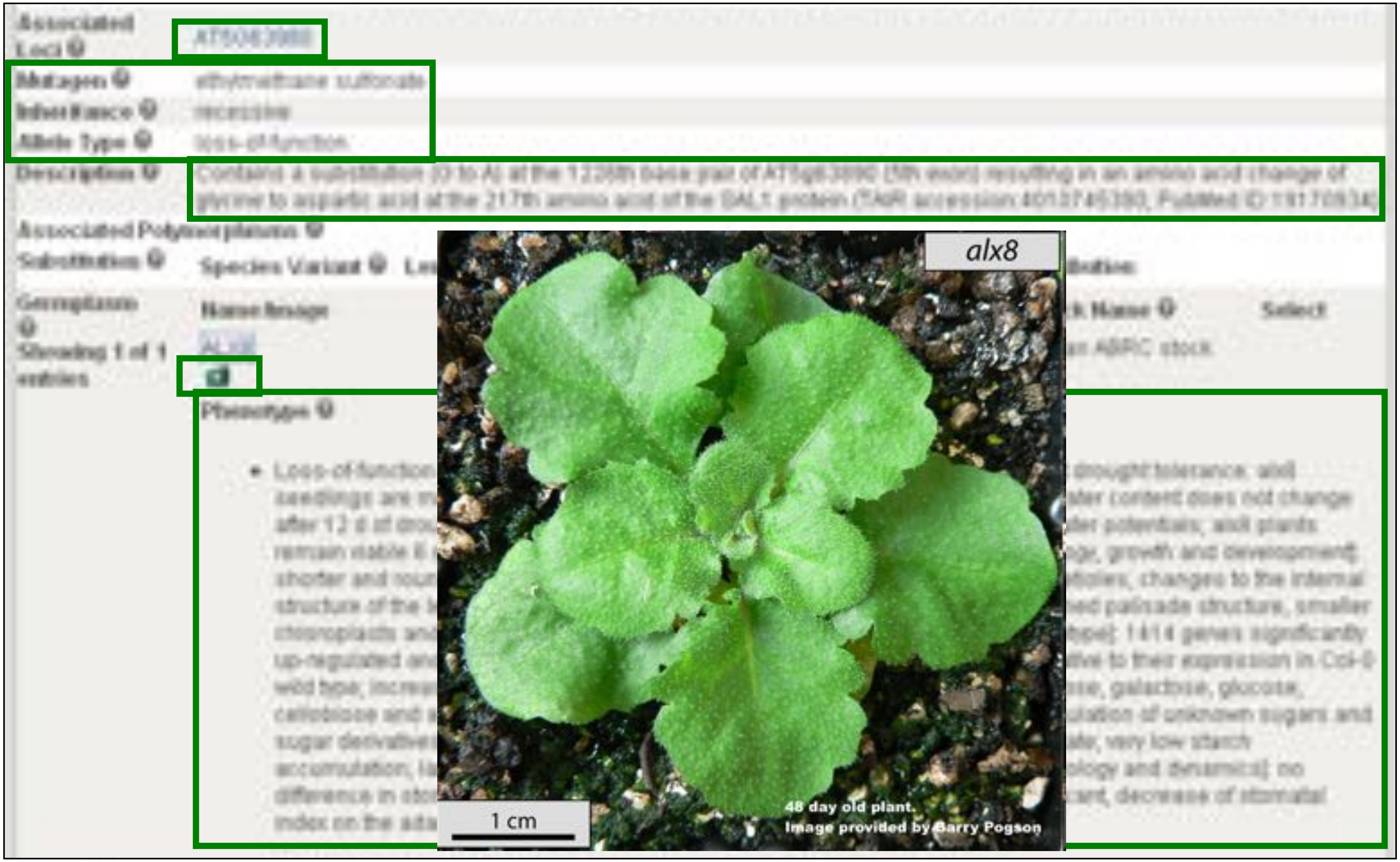
Category	Relationship Type	Keyword
GO Biological Process	involved in	nitrate transport
GO Cellular Component	expressed in	plant-type vacuole membrane
GO Molecular Function	has	nitrate transmembrane transporter activity
Growth and Developmental	expressed during	petal differentiation and expansion stage, 4 anthesis, 4 leaf senescence stage, C globular stage, D bilateral stage, E expanded cotyledon stage, F



AT5G14570	has	nitrate transmembrane transporter activity	molecular function	inferred from direct assay: protein expression in heterologous systems none: Chopin, et al. (2007)	The Arabidopsis Information Resource 2007-09-13
AT5G14570	has	nitrate transmembrane transporter activity	molecular function	inferred from sequence or structural similarity: Sequence similarity (homologue of/most closely related to): none: Ward, John (2002-09-04)	The Arabidopsis Information Resource 2003-03-29
AT5G14570	expressed in	seed	plant structure	inferred from expression pattern: expression of a reporter gene: none: Chopin, et al. (2007)	The Arabidopsis Information Resource 2007-09-13

Gene

Functional curation at TAIR



Associated Loci: AT5G63980

Mutagen: ethylmethane sulfonate

Inheritance: recessive

Allele Type: loss-of-function

Description: Contains a substitution (G to A) at the 1228th base pair of AT5g63980 (5th exon) resulting in an amino acid change of glycine to aspartic acid at the 217th amino acid of the SAL1 protein (TAIR accession:4013745380, PubMed ID:19170934)

Associated Polymorphisms

Substitution: Species Variant Le ibution

Germplasm: Name Image k Name Select

ALX8 an ABRC stock

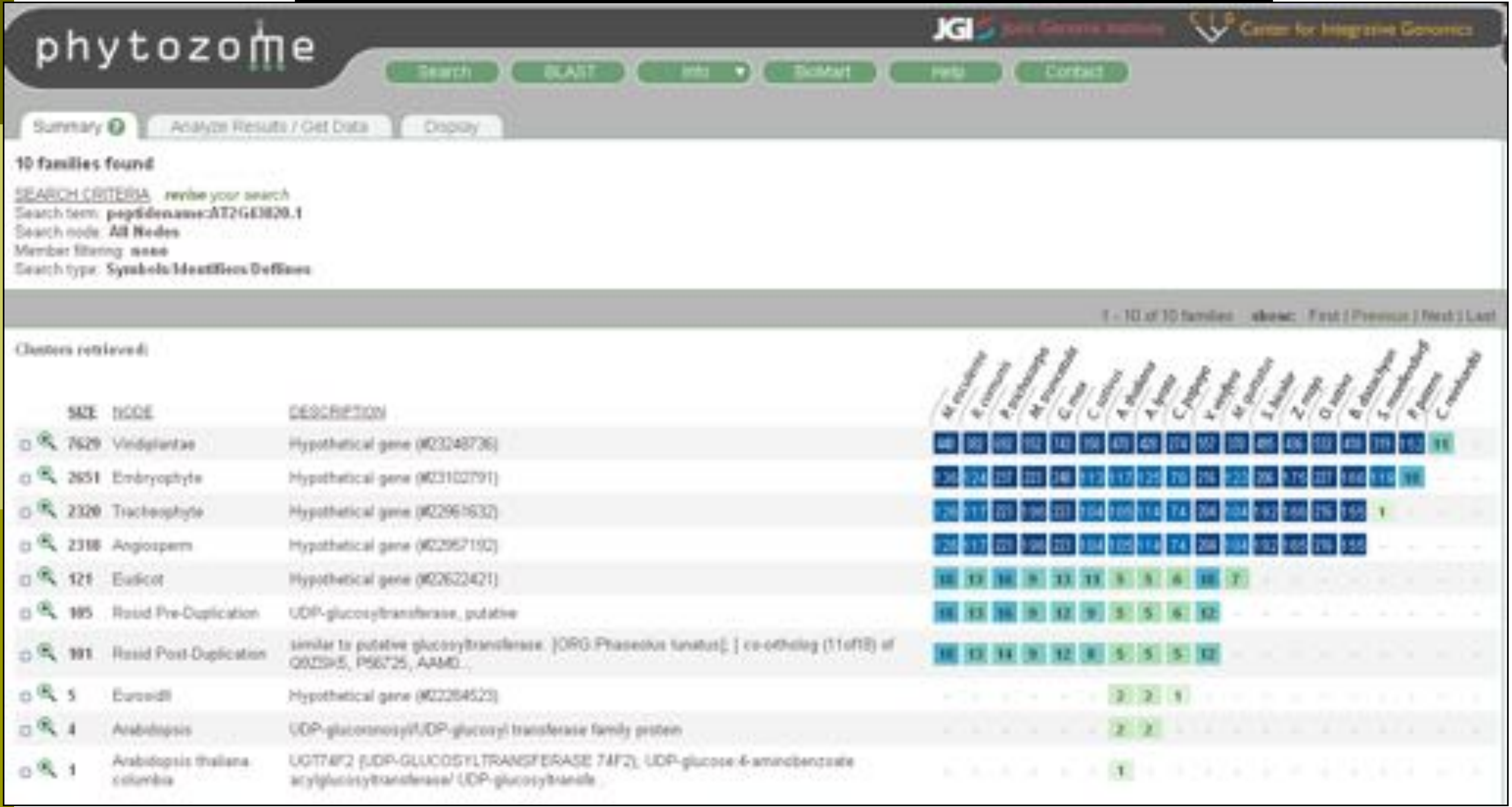
Showing 1 of 1 entries

Phenotype

- Loss-of-function drought tolerance; alx8
  seedlings are m ater content does not change
  after 12 d of drou ater potentials; alx8 plants
  remain viable 8 ogy, growth and development;
  shorter and roun etioles, changes to the internal
  structure of the le hed palisade structure, smaller
  chloroplasts and type]; 1414 genes significantly
  up-regulated an ative to their expression in Col-0
  wild type; increa ose, galactose, glucose,
  cellobiose and s ulation of unknown sugars and
  sugar derivatives ate, very low starch
  accumulation, la ology and dynamics]; no
  difference in sto cant, decrease of stomatal
  index on the ada

48 day old plant.
Image provided by Barry Pogson

Providing access to external tools and data


phytozome
JGI Joint Genome Institute
Center for Integrative Genomics
Search
BLAST
Info
BioMart
Help
Contact
Summary
Analyze Results / Get Data
Display

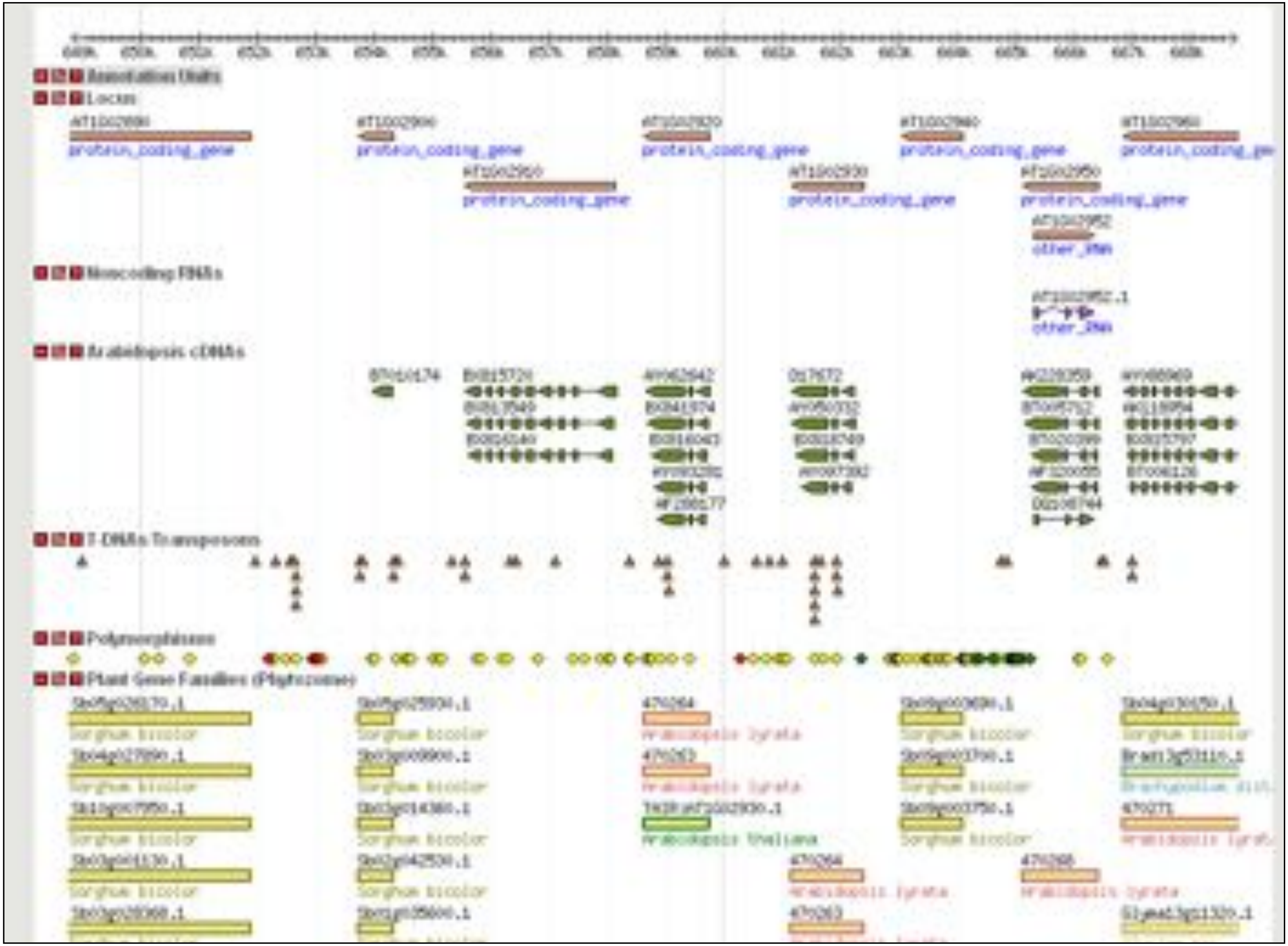
10 families found

SEARCH CRITERIA revise your search
Search term: **peptidename:AT2G43820.1**
Search node: **All Nodes**
Member filtering: **none**
Search type: **Symbols/Identifiers/Deflines**

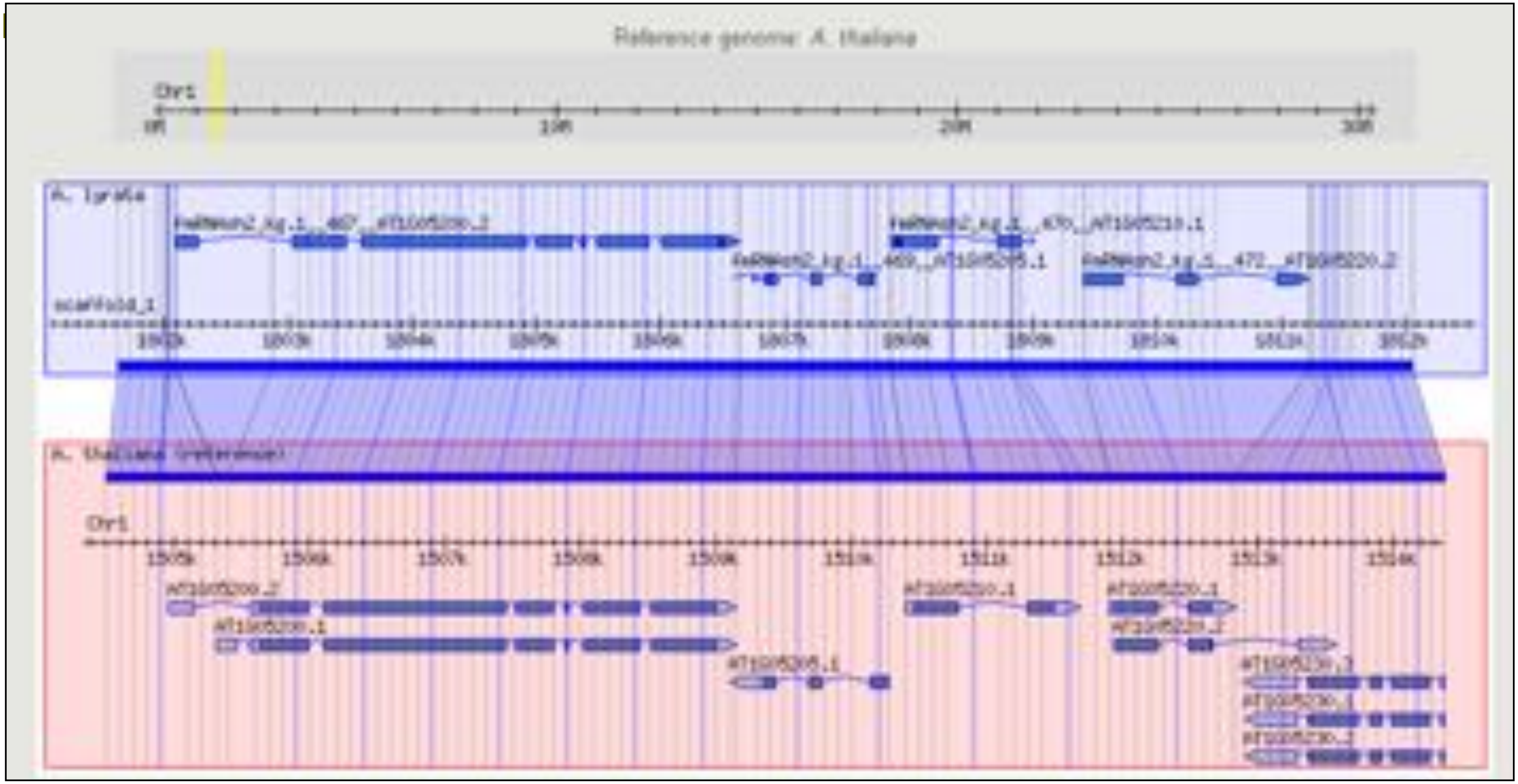
1 - 10 of 10 families show: First | Previous | Next | Last

Clusters retrieved:

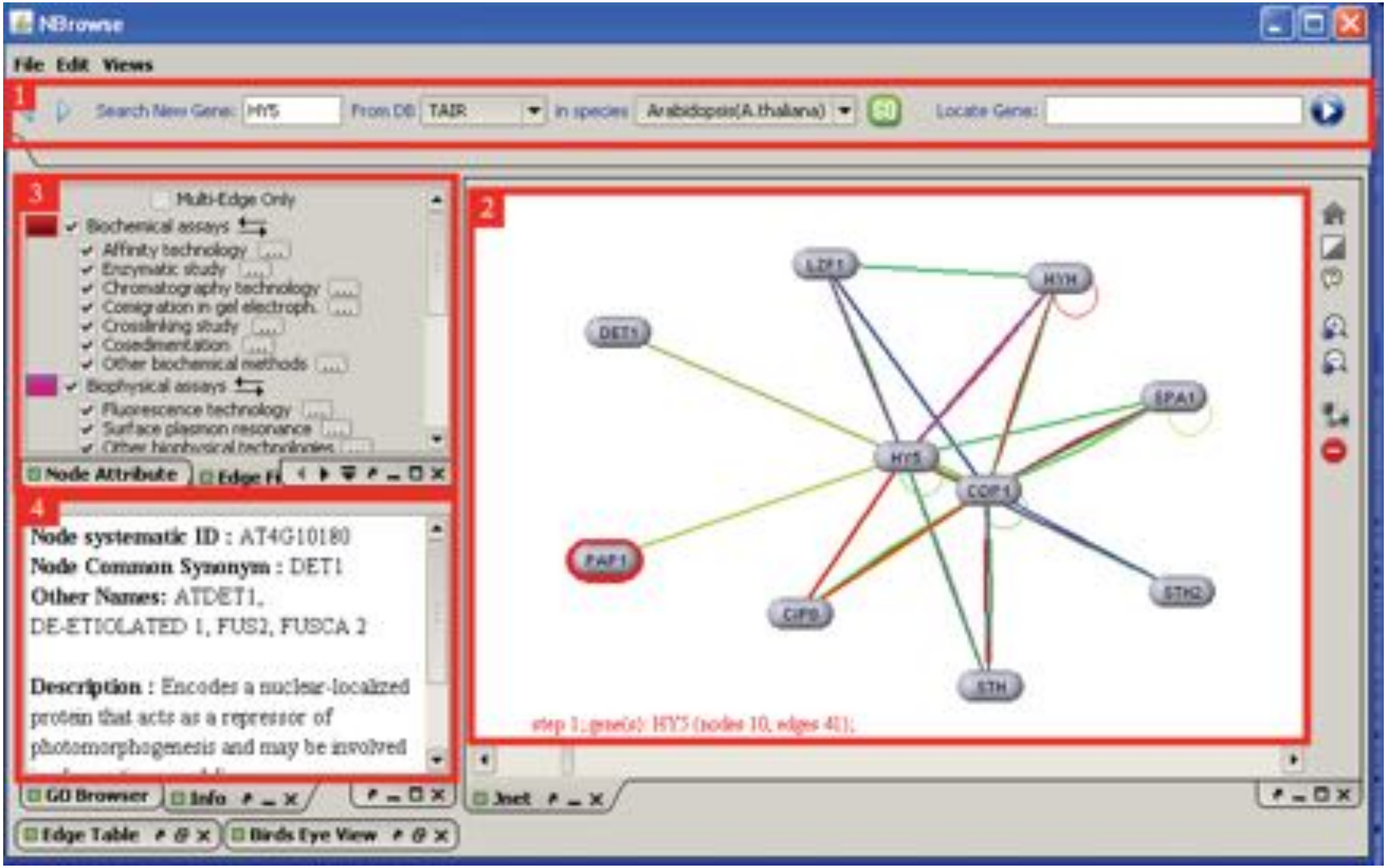
SIZE	NODE	DESCRIPTION
7629	Viridiplantae	Hypothetical gene (#23248736)
2651	Embryophyte	Hypothetical gene (#23102791)
2320	Tracheophyte	Hypothetical gene (#22961632)
2318	Angiosperm	Hypothetical gene (#22967192)
121	Eudicot	Hypothetical gene (#22622421)
105	Rosid Pre-Duplication	UDP-glucosyltransferase, putative
101	Rosid Post-Duplication	similar to putative glucosyltransferase. [ORG:Phaseolus lunatus]; [co-ortholog (11of18) of Q9ZSK5, P56725, AAM0...
5	Eurosid II	Hypothetical gene (#22264523)
4	Arabidopsis	UDP-glucoronosyl/UDP-glucosyl transferase family protein
1	Arabidopsis thaliana columbia	UGT74F2 (UDP-GLUCOSYLTRANSFERASE 74F2); UDP-glucose:4-aminobenzoate acylglucosyltransferase/ UDP-glucosyltransfe...


Annotation Units
Locus
protein_coding_gene
other_RNA
Noncoding RNAs
Arabidopsis cDNAs
T-DNAs Transposons
Polymorphisms
Plant Gene Families (Phytozome)
Sorghum bicolor
Arabidopsis lyrata
Arabidopsis thaliana
Brachypodium dist...

Providing Tools at TAIR



Providing Tools at TAIR



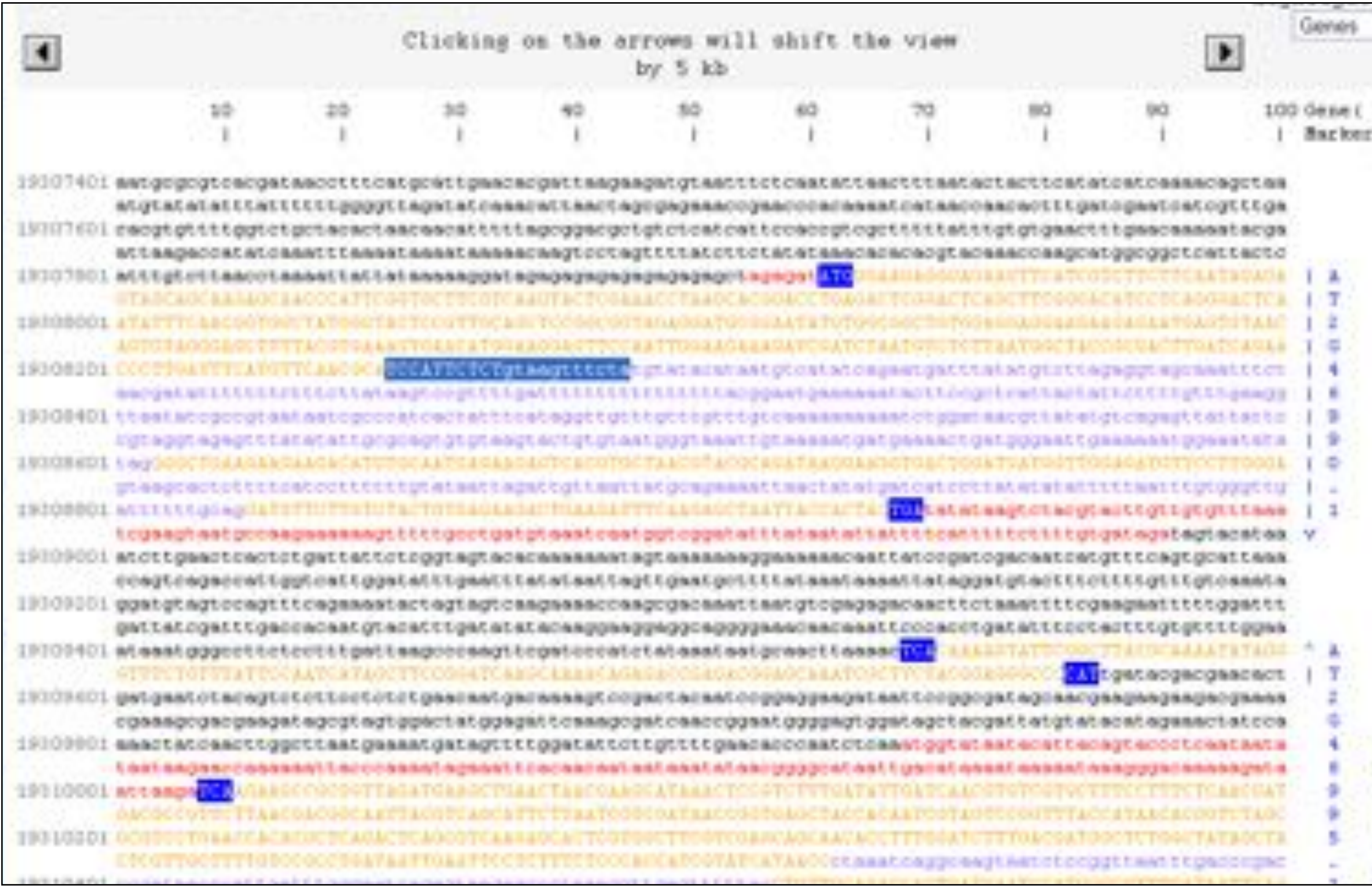
Search New Gene: HY5 From DB: TAIR In species: Arabidopsis(A.thaliana) GO Locate Gene:

□ Multi-Edge Only
✓ Biochemical assays
✓ Affinity technology
✓ Enzymatic study
✓ Chromatography technology
✓ Comigration in gel electroph.
✓ Crosslinking study
✓ Cosedimentation
✓ Other biochemical methods
✓ Biophysical assays
✓ Fluorescence technology
✓ Surface plasmon resonance
✓ Other biophysical technologies

Node systematic ID : AT4G10180
Node Common Synonym : DET1
Other Names: ATDET1, DE-ETIOLATED 1, FUS2, FUSCA 2

Description : Encodes a nuclear-localized protein that acts as a repressor of photomorphogenesis and may be involved

step 1:gene(s): HY5 (nodes 10, edges 41);

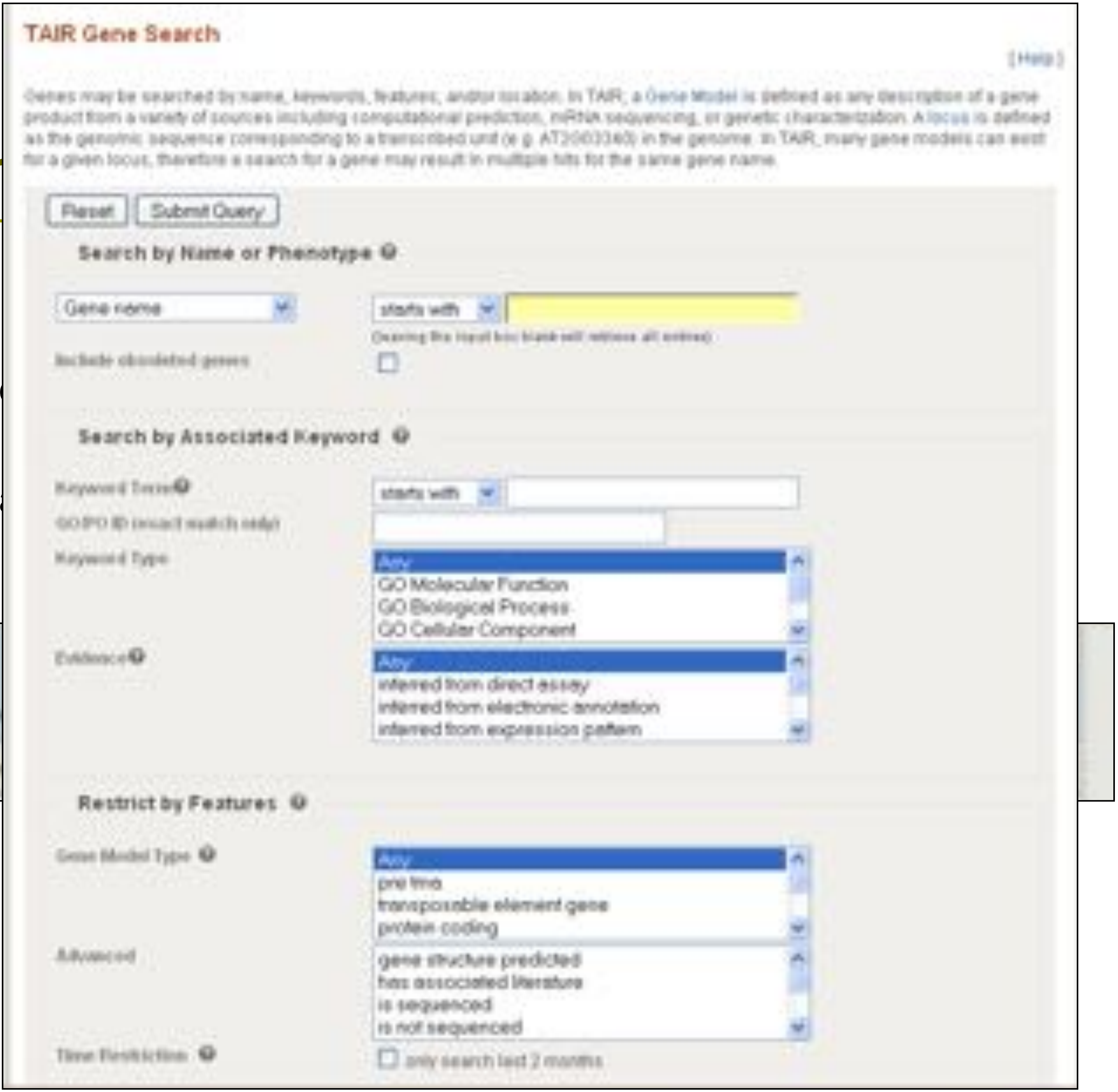


Clicking on the arrows will shift the view by 5 kb

```
          10        20        30        40        50        60        70        80        90       100 Gene (
           |         |         |         |         |         |         |         |         |         |  Marker
19307401 [illegible]
         [illegible]
19307601 [illegible]
         [illegible]
19307801 [illegible]
         [illegible]
19308001 [illegible]
         [illegible]
19308201 [illegible]
         [illegible]
19308401 [illegible]
         [illegible]
19308601 [illegible]
         [illegible]
19308801 [illegible]
         [illegible]
19309001 [illegible]
         [illegible]
19309201 [illegible]
         [illegible]
19309401 [illegible]
         [illegible]
19309601 [illegible]
         [illegible]
19309801 [illegible]
         [illegible]
19310001 [illegible]
         [illegible]
19310201 [illegible]
         [illegible]
```

Providin

- Tech team
  - Create
  - Create



TAIR Gene Search

[Help]

Genes may be searched by name, keywords, features, and/or location. In TAIR, a Gene Model is defined as any description of a gene product from a variety of sources including computational prediction, mRNA sequencing, or genetic characterization. A locus is defined as the genomic sequence corresponding to a transcribed unit (e.g. AT3G03340) in the genome. In TAIR, many gene models can exist for a given locus, therefore a search for a gene may result in multiple hits for the same gene name.

Reset | Submit Query

Search by Name or Phenotype

Gene name | starts with

(leaving the input box blank will retrieve all entries)

Include obsoleted genes ☐

Search by Associated Keyword

Keyword Term: starts with

GO/PO ID (exact match only):

Keyword Type: Any | GO Molecular Function | GO Biological Process | GO Cellular Component

Evidence: Any | inferred from direct assay | inferred from electronic annotation | inferred from expression pattern

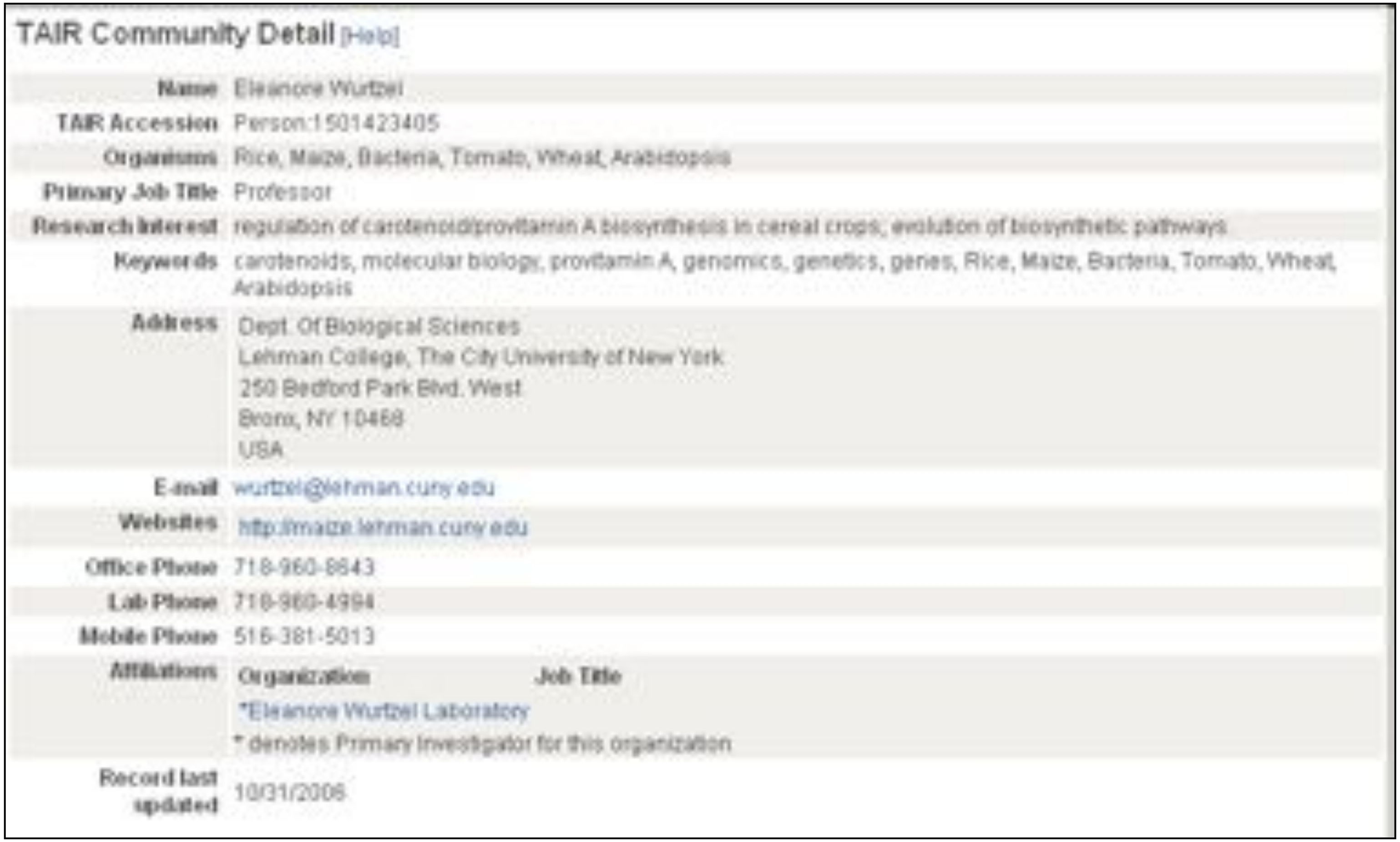
Restrict by Features

Gene Model Type: Any | pre trna | transposable element gene | protein coding

Advanced: gene structure predicted | has associated literature | is sequenced | is not sequenced

Time Restriction: ☐ only search last 2 months

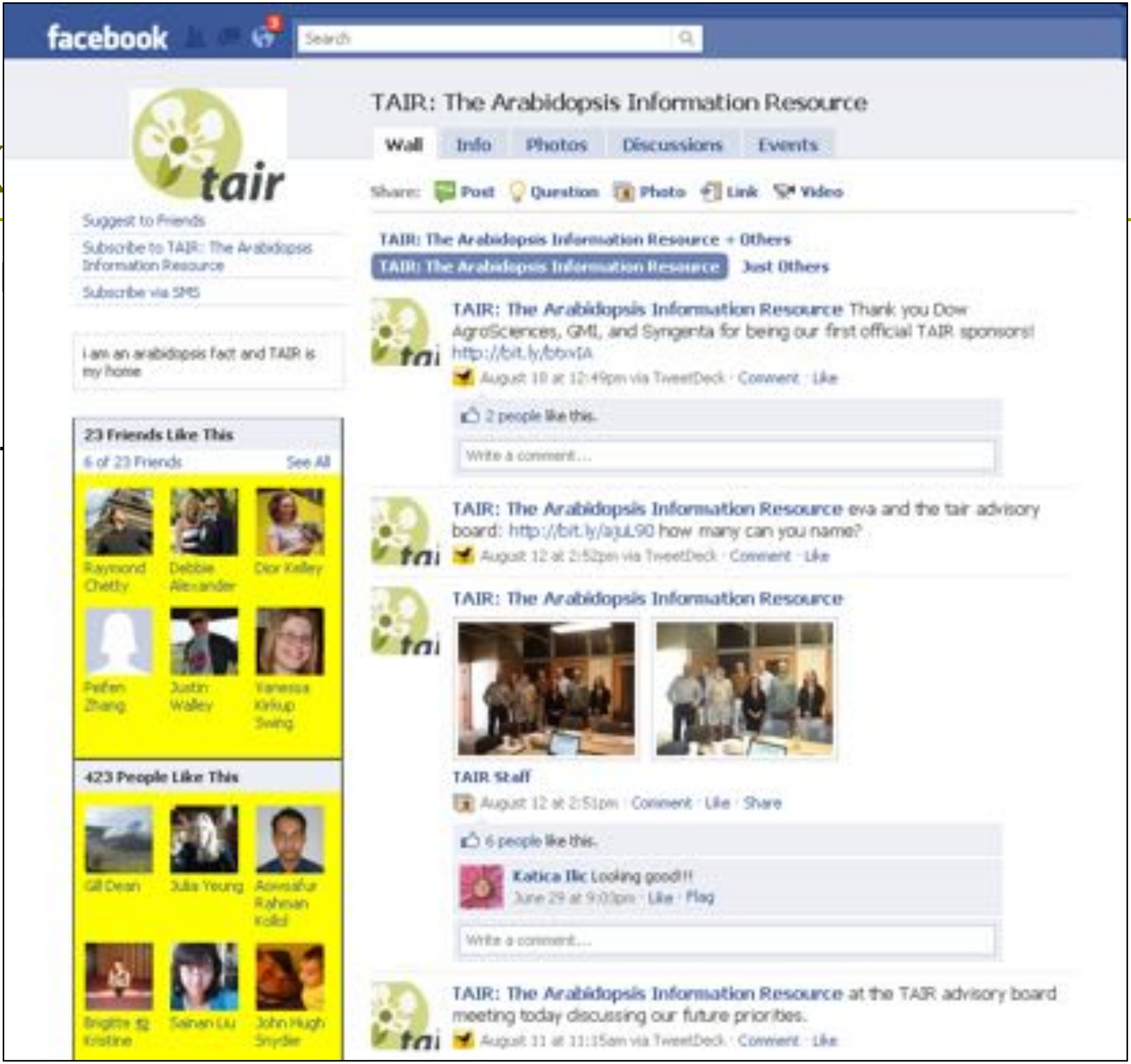
Other Resources at TAIR



TAIR Community Detail [Help]

Name	Eleanore Wurtzel
TAIR Accession	Person:1501423405
Organisms	Rice, Maize, Bacteria, Tomato, Wheat, Arabidopsis
Primary Job Title	Professor
Research Interest	regulation of carotenoid/provitamin A biosynthesis in cereal crops; evolution of biosynthetic pathways
Keywords	carotenoids, molecular biology, provitamin A, genomics, genetics, genes, Rice, Maize, Bacteria, Tomato, Wheat, Arabidopsis
Address	Dept. Of Biological Sciences Lehman College, The City University of New York 250 Bedford Park Blvd. West Bronx, NY 10468 USA
E-mail	wurtzel@lehman.cuny.edu
Websites	http://maize.lehman.cuny.edu
Office Phone	718-960-8643
Lab Phone	718-960-4994
Mobile Phone	516-381-5013
Affiliations	**Organization** **Job Title** *Eleanore Wurtzel Laboratory * denotes Primary Investigator for this organization
Record last updated	10/31/2006

K



TAIR: The Arabidopsis Information Resource

Suggest to Friends

Subscribe to TAIR: The Arabidopsis Information Resource

Subscribe via SMS

i am an arabidopsis fact and TAIR is my home

23 Friends Like This

6 of 23 Friends See All

Raymond Chetty, Debbie Alexander, Dior Kelley, Peifen Zhang, Justin Walley, Vanessa Kirkup Swing

423 People Like This

Gil Dean, Julia Young, Aowsafur Rahman Kollol, Brigitte & Kristine, Sainan Liu, John Hugh Snyder

TAIR: The Arabidopsis Information Resource + Others

TAIR: The Arabidopsis Information Resource

Just Others

TAIR: The Arabidopsis Information Resource Thank you Dow AgroSciences, GMI, and Syngenta for being our first official TAIR sponsors! http://bit.ly/bbxitA

August 18 at 12:49pm via TweetDeck · Comment · Like

2 people like this.

Write a comment...

TAIR: The Arabidopsis Information Resource eva and the tair advisory board: http://bit.ly/ajuL90 how many can you name?

August 12 at 2:52pm via TweetDeck · Comment · Like

TAIR: The Arabidopsis Information Resource

TAIR Staff

August 12 at 2:51pm · Comment · Like · Share

6 people like this.

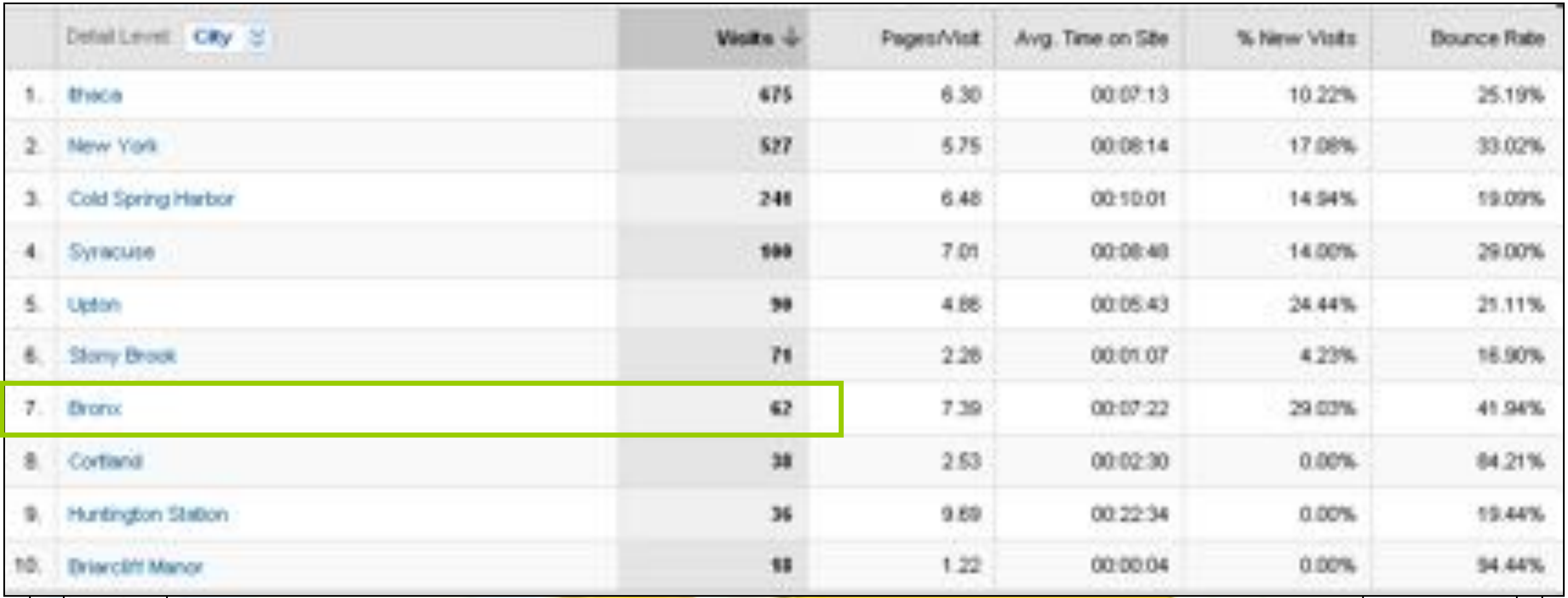
Katica Ilic Looking good!!!

June 29 at 9:03pm · Like · Flag

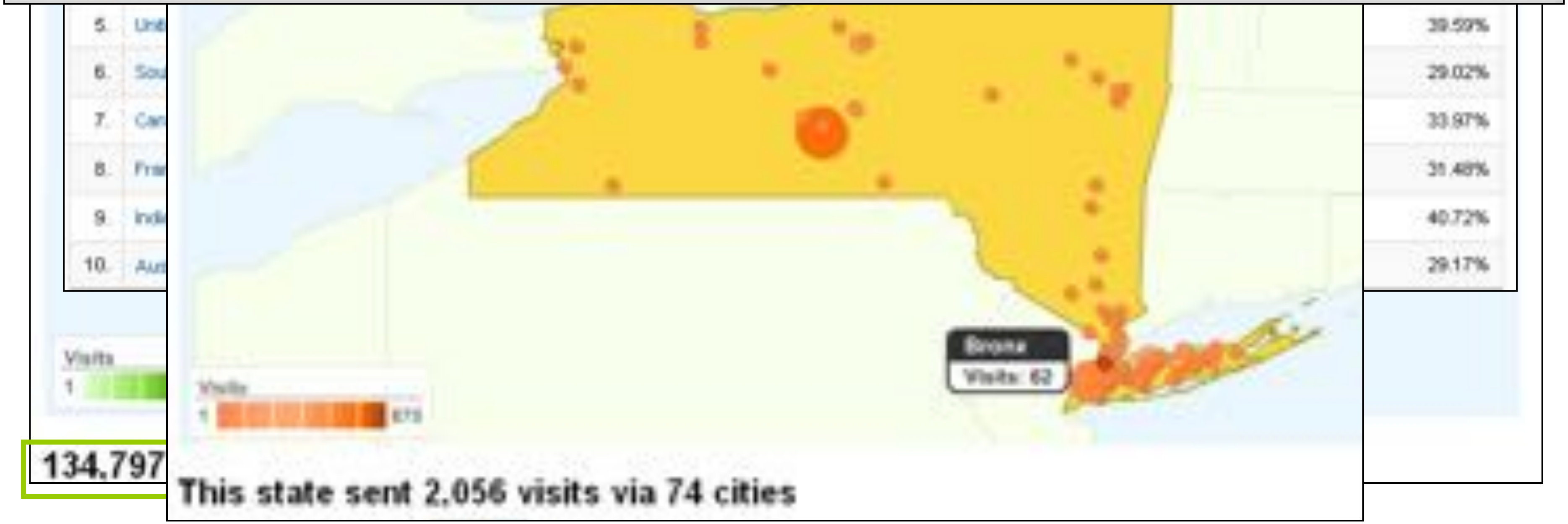
Write a comment...

TAIR: The Arabidopsis Information Resource at the TAIR advisory board meeting today discussing our future priorities.

August 11 at 11:15am via TweetDeck · Comment · Like



Detail Level: City	Visits ↓	Pages/Visit	Avg. Time on Site	% New Visits	Bounce Rate
1. Ithaca	675	6.30	00:07:13	10.22%	25.19%
2. New York	527	5.75	00:08:14	17.08%	33.02%
3. Cold Spring Harbor	248	6.48	00:10:01	14.94%	19.09%
4. Syracuse	100	7.01	00:08:48	14.00%	29.00%
5. Upton	90	4.86	00:05:43	24.44%	21.11%
6. Stony Brook	71	2.28	00:01:07	4.23%	16.90%
7. Bronx	62	7.39	00:07:22	29.03%	41.94%
8. Cortland	38	2.53	00:02:30	0.00%	84.21%
9. Huntington Station	36	9.69	00:22:34	0.00%	19.44%
10. Briarcliff Manor	18	1.22	00:00:04	0.00%	94.44%



5.	Uni	39.59%
6.	Sou	29.02%
7.	Can	33.97%
8.	Fra	31.48%
9.	Indi	40.72%
10.	Aus	29.17%

134,797

This state sent 2,056 visits via 74 cities

How can TAIR contribute to your work?

- If you work on Arabidopsis . . .
  - Find specific information about individual genes and proteins
  - Access large Arabidopsis-specific data sets

- If you work on another species . . .
  - Take your gene / protein of interest and find all the data TAIR contains for its ortholog
  - Look up your favorite:
    - biological process
    - molecular function
    - subcellular compartment
    - organ or tissue
    - developmental stage
    - mutant phenotype

    - Indentify many related genes in TAIR and then find orthologs in your species

- But . . . if you want more on plant metabolism . . . .

Welcome to the PMN!

PMN

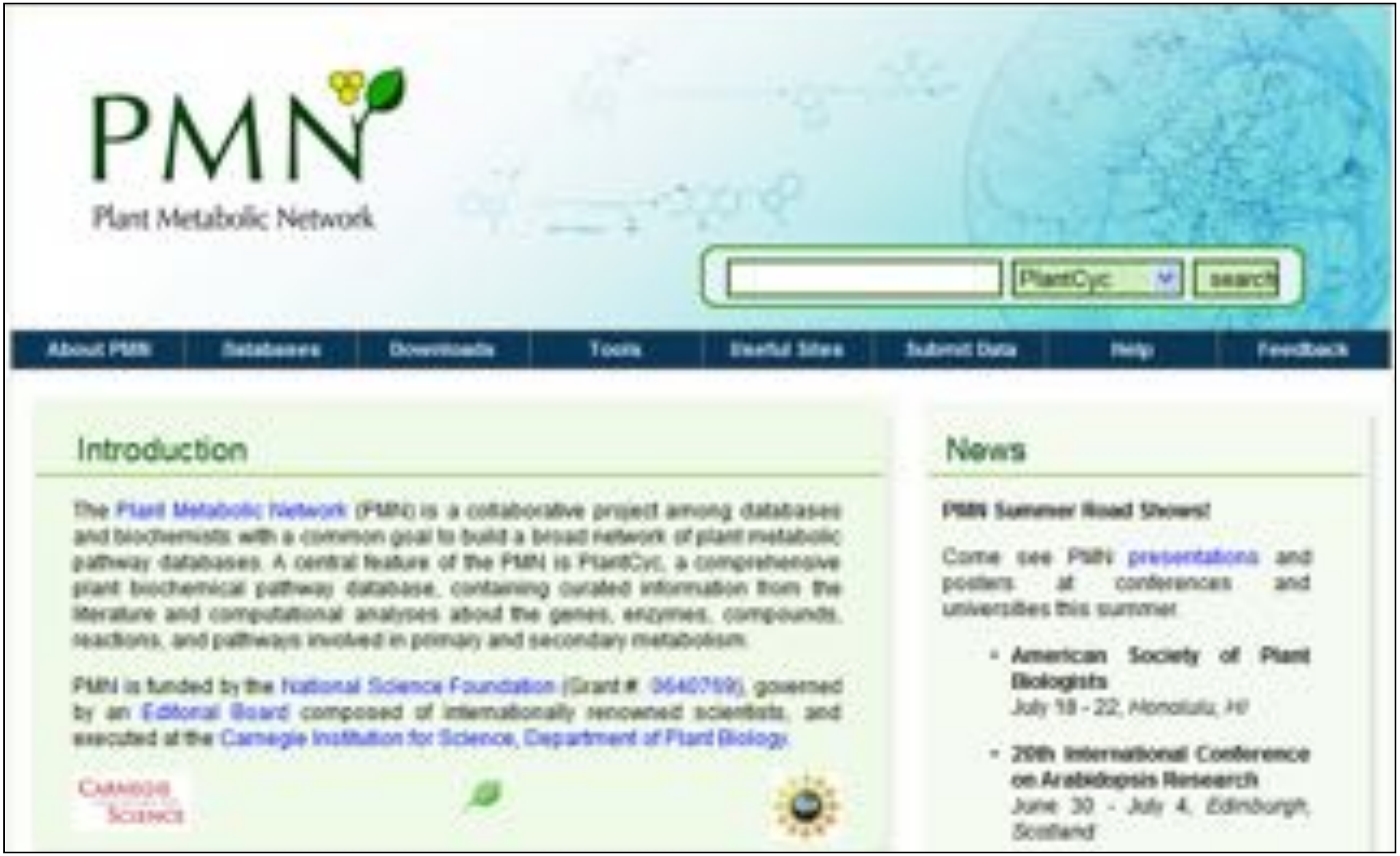
Plant Metabolic Network

- PMN = The **P**lant **M**etabolic **N**etwork
  - Created in 2008
  - Funded by the National Science Foundation
- What is
- What da
- How do
- How car
- How can you help the PMN to grow?

Sue Rhee (PI)

Peifen Zhang (Director)

What is the PMN?



Introduction

The Plant Metabolic Network (PMN) is a collaborative project among databases and biochemists with a common goal to build a broad network of plant metabolic pathway databases. A central feature of the PMN is PlantCyc, a comprehensive plant biochemical pathway database, containing curated information from the literature and computational analyses about the genes, enzymes, compounds, reactions, and pathways involved in primary and secondary metabolism.

PMN is funded by the National Science Foundation (Grant #: 0640769), governed by an Editorial Board composed of internationally renowned scientists, and executed at the Carnegie Institution for Science, Department of Plant Biology.

News

PMN Summer Road Shows!

Come see PMN presentations and posters at conferences and universities this summer

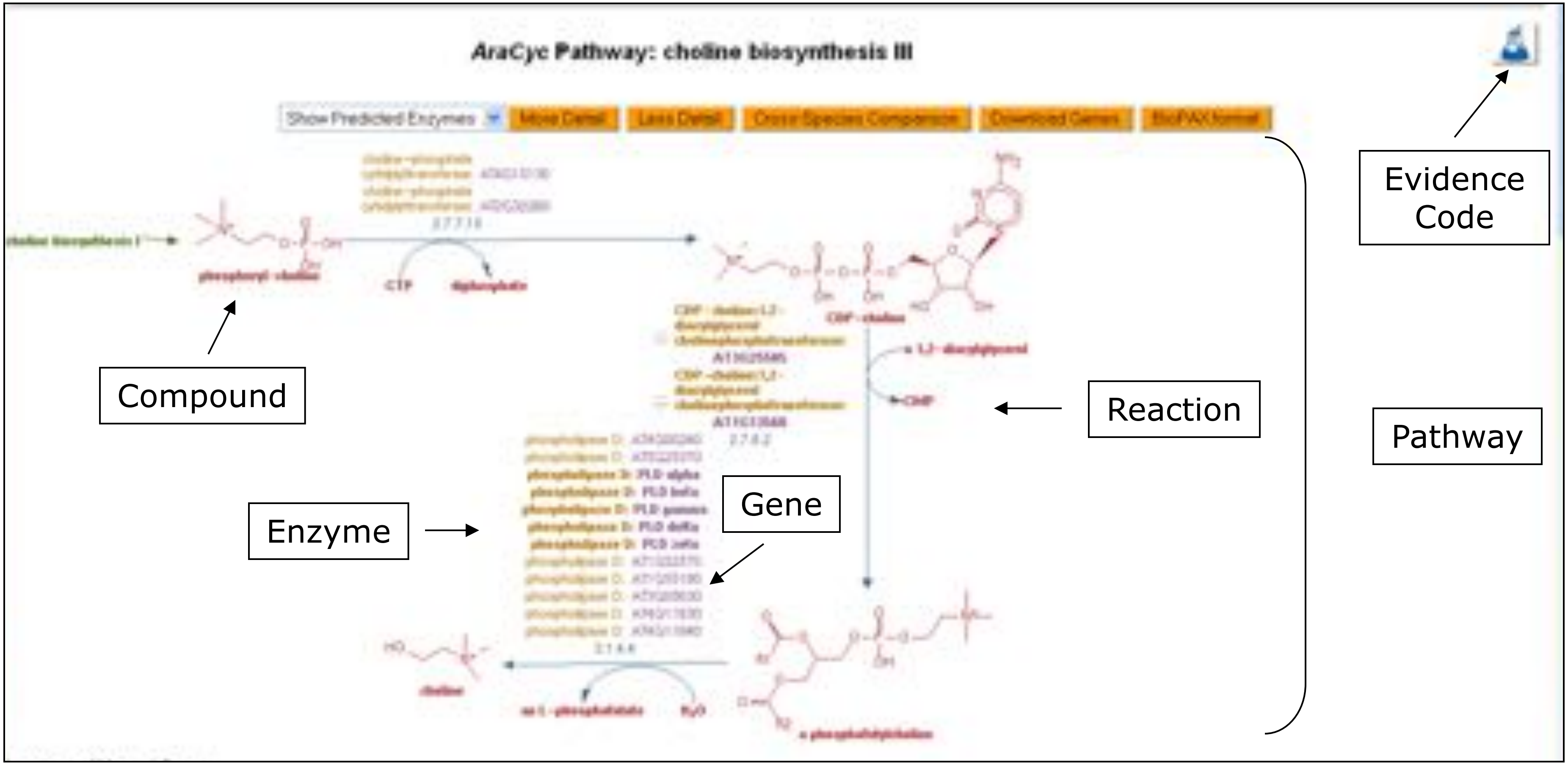
- **American Society of Plant Biologists**
  July 18 - 22, *Honolulu, HI*
- **20th International Conference on Arabidopsis Research**
  June 30 - July 4, *Edinburgh, Scotland*

- Facilitate research that benefits society

Connecting the PMN to important research efforts

- **More nutritious foods**
  - vitamin A biosynthesis, folate biosynthesis . . .
- **Medicines**
  - morphine biosynthesis, taxol biosynthesis . . .
- **More pest-resistant plants**
  - maackiain biosynthesis, capsidiol biosynthesis . . .
- **Higher photosynthetic capacity and yield in crops**
  - chlorophyll biosynthesis, Calvin cycle . . .
- **Better biofuel feedstocks**
  - cellulose biosynthesis, lignin biosynthesis . . .
- **Many additional applications relevant to rational metabolic engineering**
  - ethylene biosynthesis, resveratrol biosynthesis . . .

What data are in the PMN?


AraCyc Pathway: choline biosynthesis III
Show Predicted Enzymes
More Detail
Less Detail
Cross Species Comparison
Download Genes
BioPAX format
choline biosynthesis I
phosphoryl-choline
choline-phosphate cytidylyltransferase
choline-phosphate cytidylyltransferase
2.7.7.15
CTP
diphosphate
CDP-choline
CDP-choline:1,2-diacylglycerol cholinephosphotransferase
CDP-choline:1,2-diacylglycerol cholinephosphotransferase
2.7.8.2
a 1,2-diacylglycerol
CMP
phospholipase D: PLD alpha
phospholipase D: PLD beta
phospholipase D: PLD gamma
phospholipase D: PLD delta
phospholipase D: PLD zeta
3.1.4.4
choline
an L-phosphatidate
H₂O
a phosphatidylcholine

Evidence Code

Compound

Reaction

Pathway

Enzyme

Gene

Pathway Tools software provided by collaborators at SRI International

PMN databases

- Current PMN databases: **PlantCyc, AraCyc, PoplarCyc**
  - Coming soon: databases for wine grape, maize, cassava, Selaginella, and more . . .

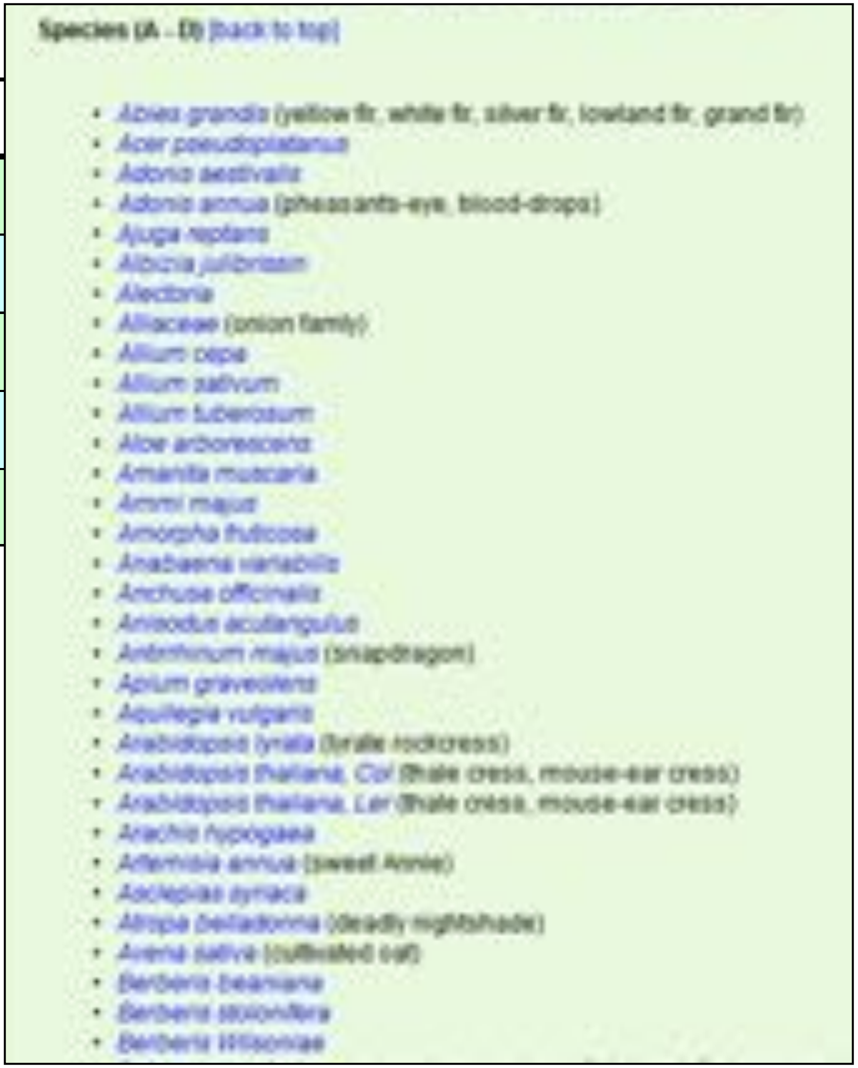
- Other plant databases accessible from the PMN:

PGDB	Plant	Source	Status
RiceCyc **	Rice	Gramene	some curation
SorghumCyc	Sorghum	Gramene	no curation
MedicCyc **	Medicago	Noble Foundation	some curation
LycoCyc **	Tomato	Sol Genomics Network	some curation
PotatoCyc	Potato	Sol Genomics Network	no curation
CapCyc	Pepper	Sol Genomics Network	no curation
NicotianaCyc	Tobacco	Sol Genomics Network	no curation
PetuniaCyc	Petunia	Sol Genomics Network	no curation
CoffeaCyc	Coffee	Sol Genomics Network	no curation

** Significant numbers of genes from these databases have been integrated into PlantCyc

PMN database content statistics

	PlantCyc 4.0
Pathways	685
Enzymes	11058
Reactions	2929
Compounds	2966
Organisms	343



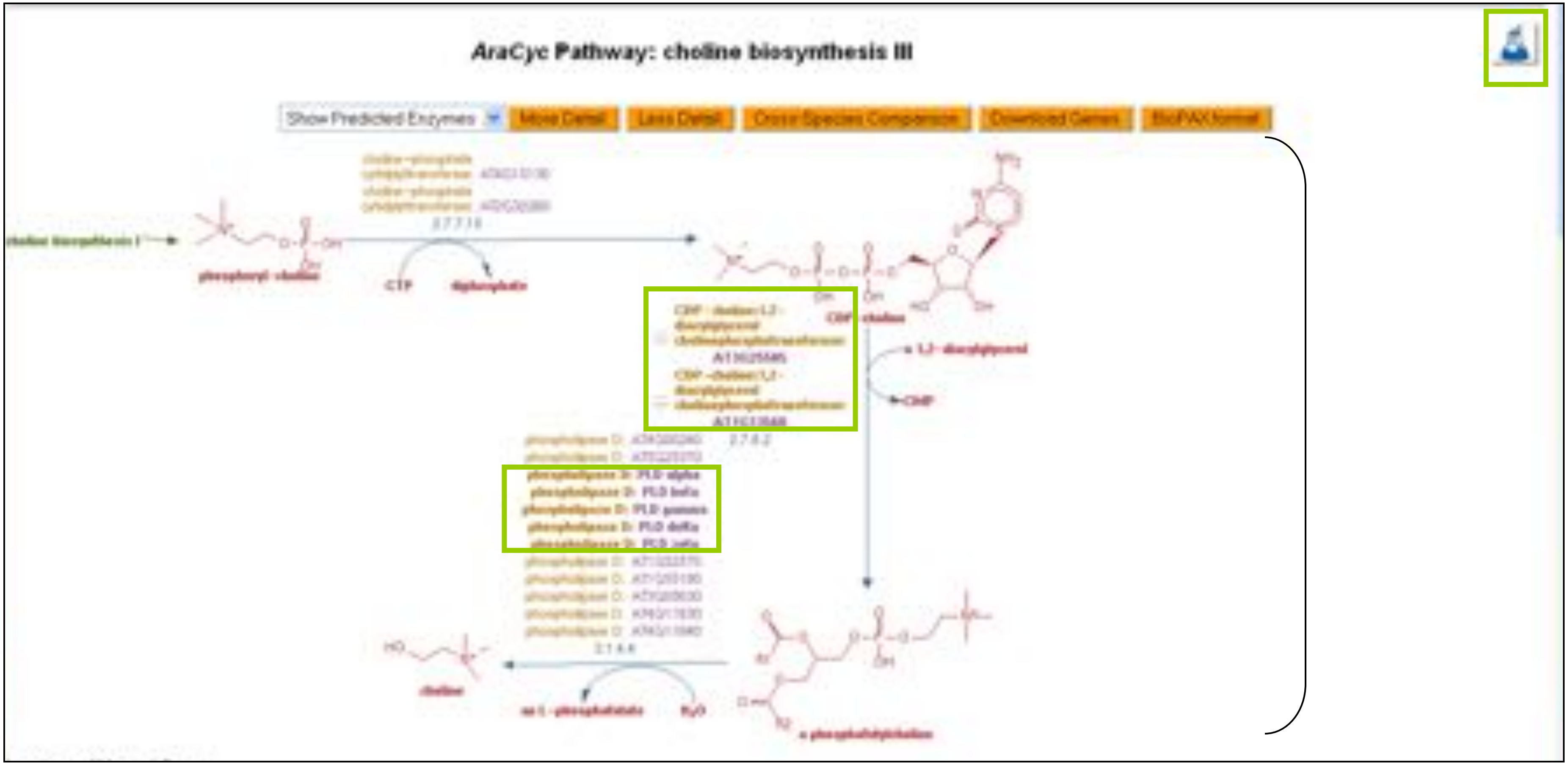
Species (A - D) [back to top]

- *Abies grandis* (yellow fir, white fir, silver fir, lowland fir, grand fir)
- *Acer pseudoplatanus*
- *Adonis aestivalis*
- *Adonis annua* (pheasants-eye, blood-drops)
- *Ajuga reptans*
- *Albizia julibrissin*
- *Alectoria*
- *Alliaceae* (onion family)
- *Allium cepa*
- *Allium sativum*
- *Allium tuberosum*
- *Aloe arborescens*
- *Amanita muscaria*
- *Ammi majus*
- *Amorpha fruticosa*
- *Anabaena variabilis*
- *Anchusa officinalis*
- *Anisodus acutangulus*
- *Antirrhinum majus* (snapdragon)
- *Apium graveolens*
- *Aquilegia vulgaris*
- *Arabidopsis lyrata* (lyrate rockcress)
- *Arabidopsis thaliana, Col* (thale cress, mouse-ear cress)
- *Arabidopsis thaliana, Ler* (thale cress, mouse-ear cress)
- *Arachis hypogaea*
- *Artemisia annua* (sweet Annie)
- *Asclepias syriaca*
- *Atropa belladonna* (deadly nightshade)
- *Avena sativa* (cultivated oat)
- *Berberis beaniana*
- *Berberis stolonifera*
- *Berberis Wilsoniae*

How does experimentally verified data enter the PMN?

- Biocurators perform manual curation
  - Use journal articles to enter information
  - Receive helpful messages from researchers
  - Request specific data from experts
  - Invite editorial board members to review metabolic domains

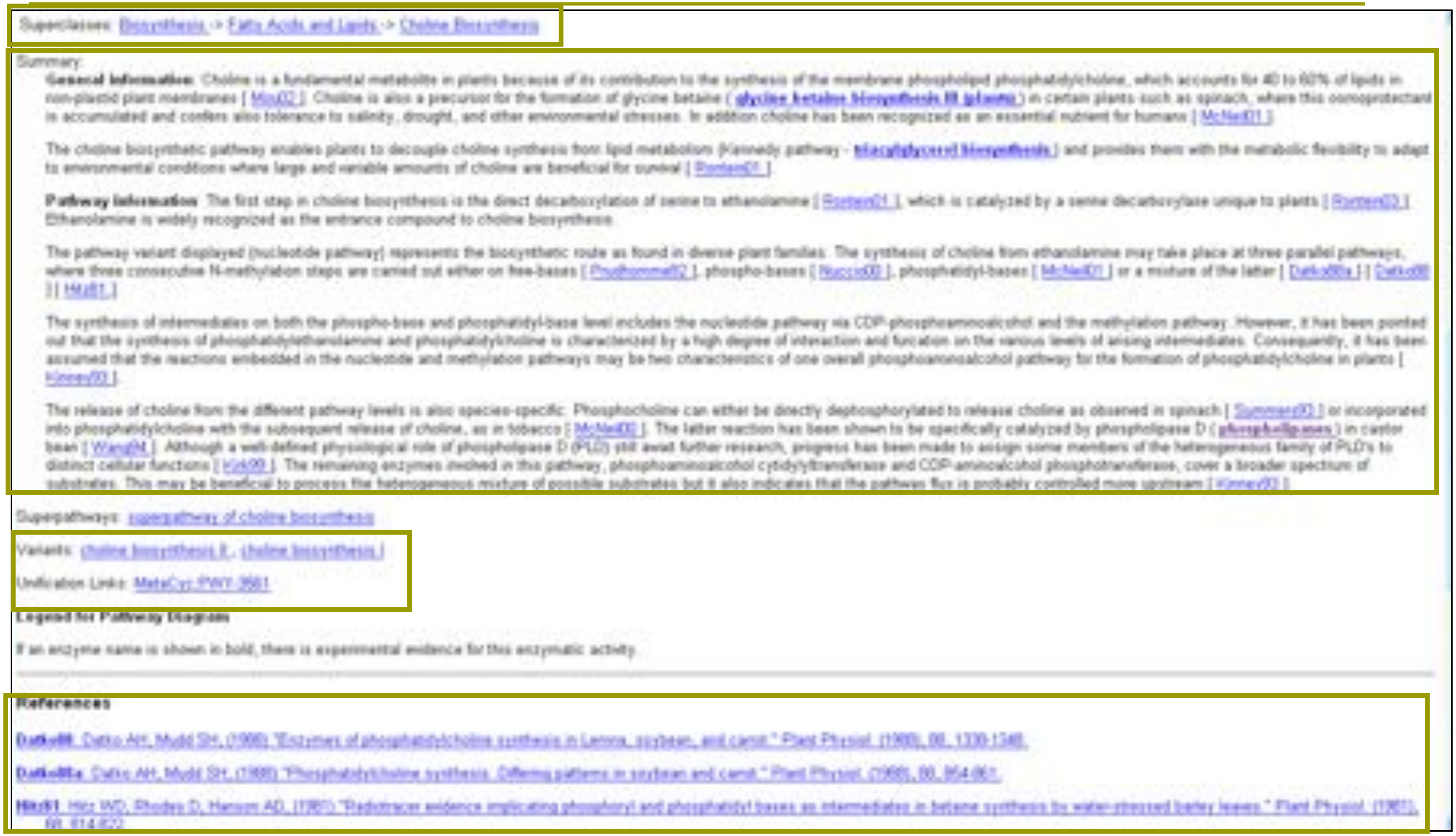
Pathway information



AraCyc Pathway: choline biosynthesis III

Show Predicted Enzymes
More Detail
Less Detail
Cross Species Comparison
Download Genes
BioPAX format
choline-phosphate cytidylyltransferase AT4G15130
choline-phosphate cytidylyltransferase AT2G32260
2.7.7.15
choline biosynthesis I
phosphoryl-choline
CTP
diphosphate
CDP-choline
CDP-choline:1,2-diacylglycerol cholinephosphotransferase AT3G25585
CDP-choline:1,2-diacylglycerol cholinephosphotransferase AT1G13560
2.7.8.2
a 1,2-diacylglycerol
CMP
phospholipase D: AT4G00240
phospholipase D: AT5G25370
phospholipase D: PLD alpha
phospholipase D: PLD beta
phospholipase D: PLD gamma
phospholipase D: PLD delta
phospholipase D: PLD zeta
phospholipase D: AT1G52570
phospholipase D: AT1G55180
phospholipase D: AT3G05630
phospholipase D: AT4G11830
phospholipase D: AT4G11840
3.1.4.4
choline
an L-phosphatidate
H₂O
a phosphatidylcholine

Pathway information



Superclasses: Biosynthesis -> Fatty Acids and Lipids -> Choline Biosynthesis

Summary:

General information: Choline is a fundamental metabolite in plants because of its contribution to the synthesis of the membrane phospholipid phosphatidylcholine, which accounts for 40 to 60% of lipids in non-plastid plant membranes [Mou02]. Choline is also a precursor for the formation of glycine betaine (**glycine betaine biosynthesis III (plants)**) in certain plants such as spinach, where this osmoprotectant is accumulated and confers also tolerance to salinity, drought, and other environmental stresses. In addition choline has been recognized as an essential nutrient for humans [McNeil01].

The choline biosynthetic pathway enables plants to decouple choline synthesis from lipid metabolism (Kennedy pathway - **triacylglycerol biosynthesis**) and provides them with the metabolic flexibility to adapt to environmental conditions where large and variable amounts of choline are beneficial for survival [Rontein01].

Pathway information: The first step in choline biosynthesis is the direct decarboxylation of serine to ethanolamine [Rontein01], which is catalyzed by a serine decarboxylase unique to plants [Rontein03]. 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 [Prudhomme92], phospho-bases [Nuccio00], phosphatidyl-bases [McNeil01] or a mixture of the latter [Datko88a] [Datko88] [Hitz81]

The synthesis of intermediates on both the phospho-base and phosphatidyl-base level includes the nucleotide pathway via CDP-phosphoaminoalcohol 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 furcation 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 phosphoaminoalcohol pathway for the formation of phosphatidylcholine in plants [Kinney93].

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 [Summers93] or incorporated into phosphatidylcholine with the subsequent release of choline, as in tobacco [McNeil00]. The latter reaction has been shown to be specifically catalyzed by phospholipase D (**phospholipases**) in castor bean [Wang94]. Although a well-defined physiological role of phospholipase D (PLD) still await further research, progress has been made to assign some members of the heterogeneous family of PLD's to distinct cellular functions [Kirk99]. The remaining enzymes involved in this pathway, phosphoaminoalcohol cytidylyltransferase and CDP-aminoalcohol 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 pathways flux is probably controlled more upstream [Kinney93].

Superpathways: superpathway of choline biosynthesis

Variants: choline biosynthesis II , choline biosynthesis I

Unification Links: MetaCyc:PWY-3561

Legend for Pathway Diagram

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

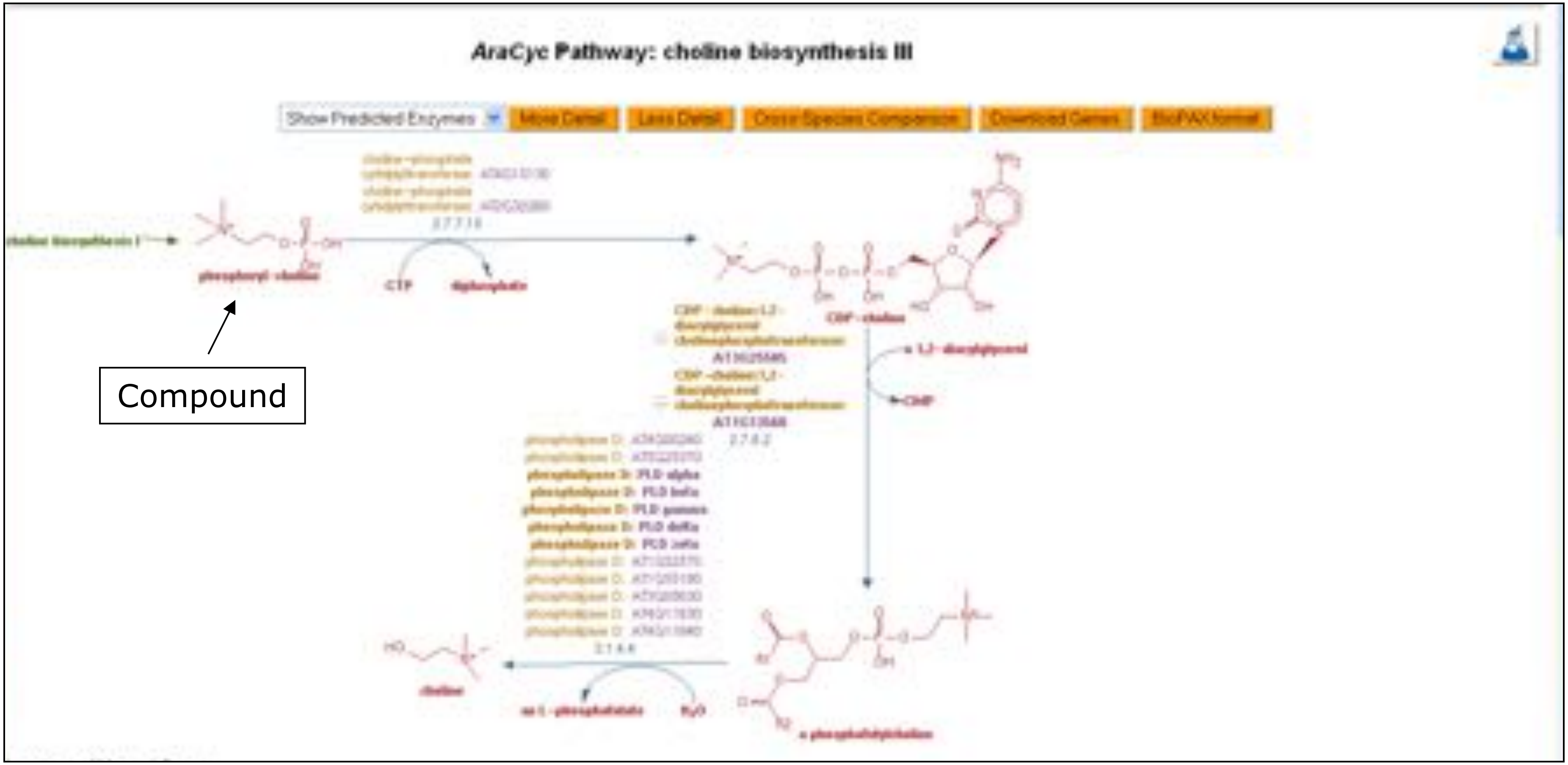
References

Datko88: Datko AH, Mudd SH, (1988) "Enzymes of phosphatidylcholine synthesis in Lemna, soybean, and carrot." Plant Physiol. (1988), 88, 1338-1348.

Datko88a: Datko AH, Mudd SH, (1988) "Phosphatidylcholine synthesis. Differing patterns in soybean and carrot." Plant Physiol. (1988), 88, 854-861.

Hitz81: Hitz WD, Rhodes D, Hanson AD, (1981) "Radiotracer evidence implicating phosphoryl and phosphatidyl bases as intermediates in betaine synthesis by water-stressed barley leaves." Plant Physiol. (1981), 68, 814-822.

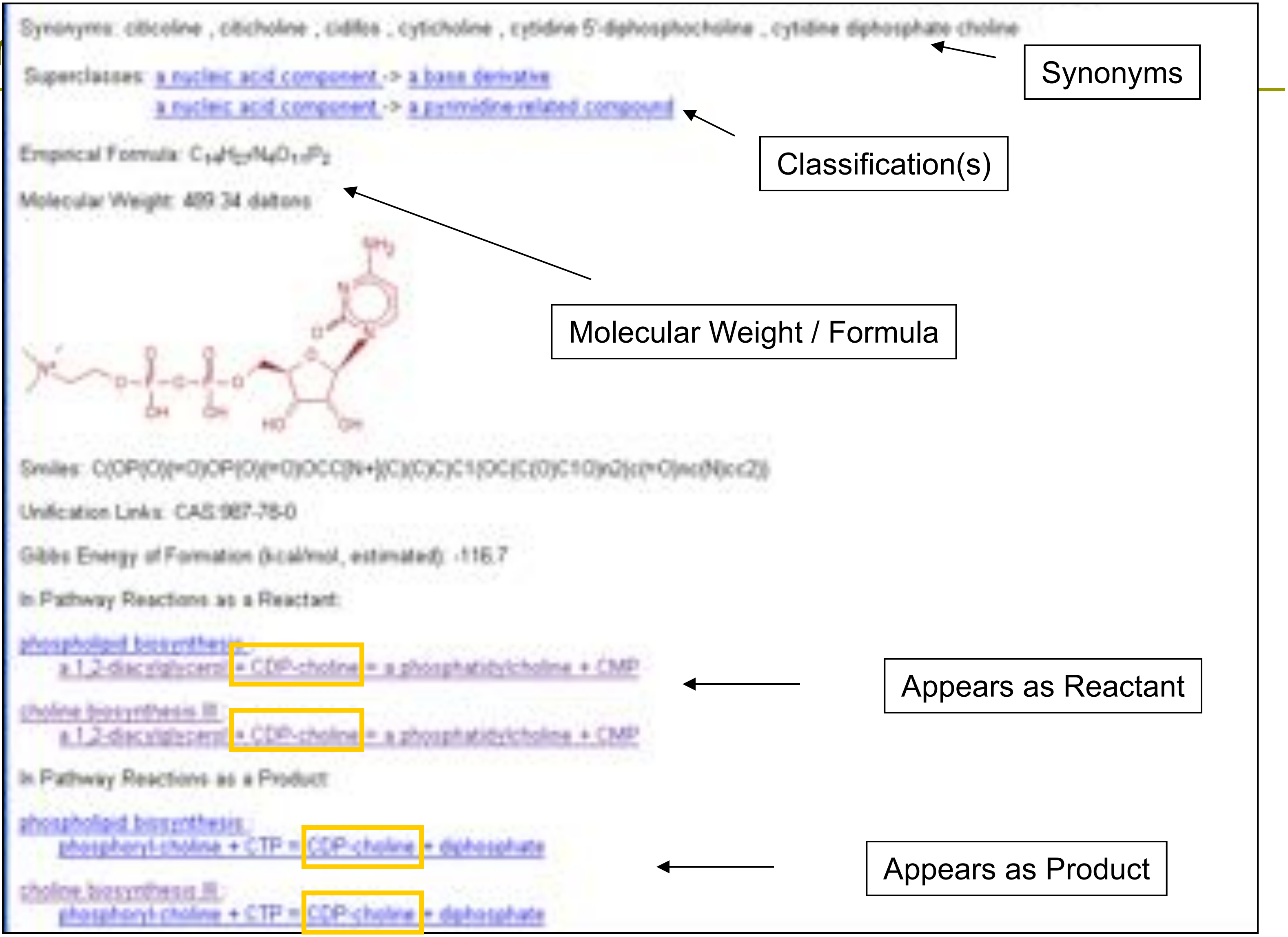
Compound information



Compound

Com

Compound: CDP-choline



Synonyms: citicoline , citicholine , cidifos , cyticholine , cytidine 5'-diphosphocholine , cytidine diphosphate choline

Superclasses: a nucleic acid component -> a base derivative
a nucleic acid component -> a pyrimidine-related compound

Empirical Formula: $C_{14}H_{27}N_4O_{11}P_2$

Molecular Weight: 489.34 daltons

Smiles: C(OP(O)(=O)OP(O)(=O)OCC[N+](C)(C)C)C1(OC(C(O)C1O)n2)c(=O)nc(N)cc2)

Unification Links: CAS:987-78-0

Gibbs Energy of Formation (kcal/mol, estimated): -116.7

In Pathway Reactions as a Reactant:

phospholipid biosynthesis :
a 1,2-diacylglycerol + CDP-choline = a phosphatidylcholine + CMP

choline biosynthesis III :
a 1,2-diacylglycerol + CDP-choline = a phosphatidylcholine + CMP

In Pathway Reactions as a Product:

phospholipid biosynthesis :
phosphoryl-choline + CTP = CDP-choline + diphosphate

choline biosynthesis III :
phosphoryl-choline + CTP = CDP-choline + diphosphate

Synonyms

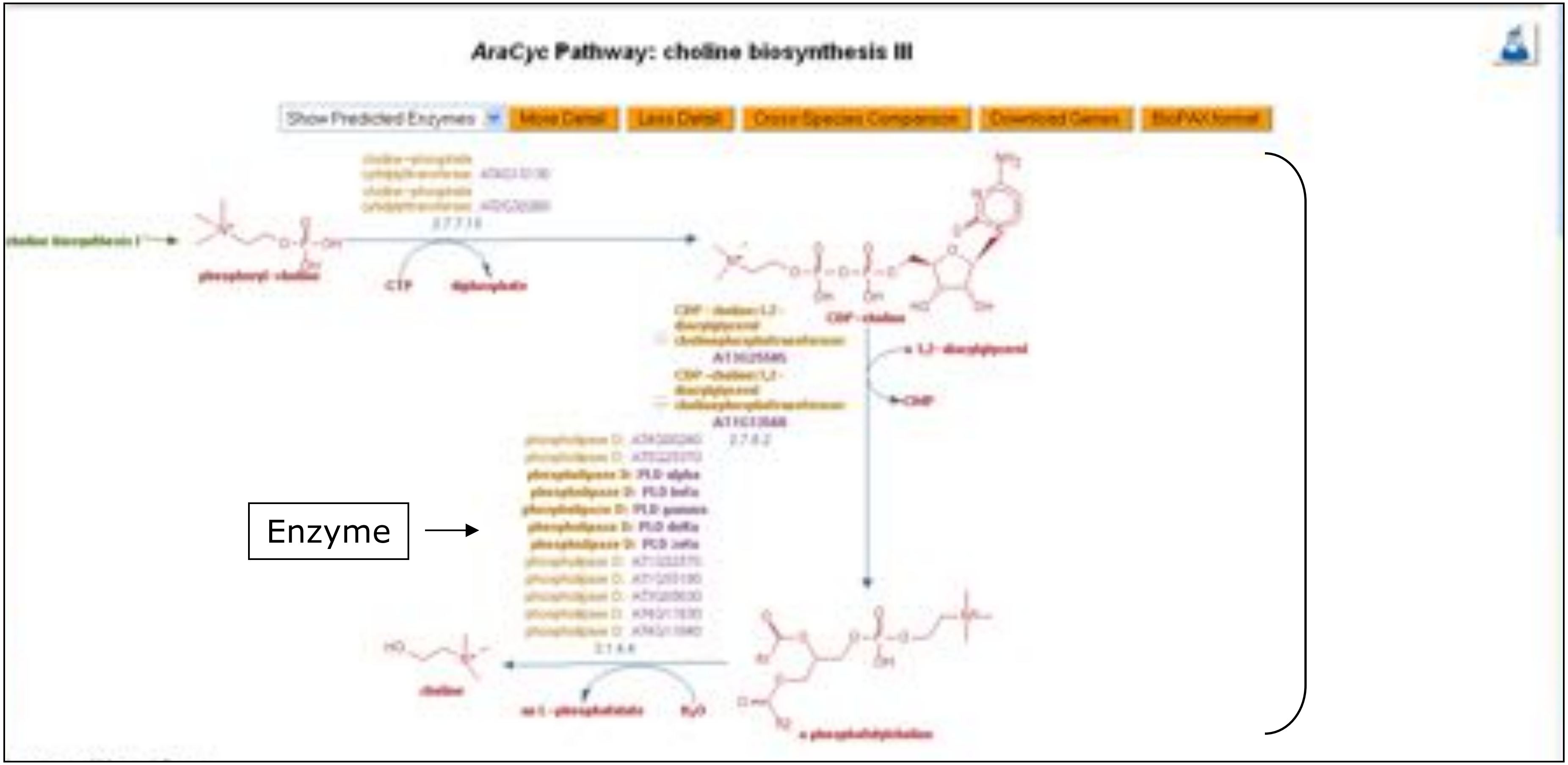
Classification(s)

Molecular Weight / Formula

Appears as Reactant

Appears as Product

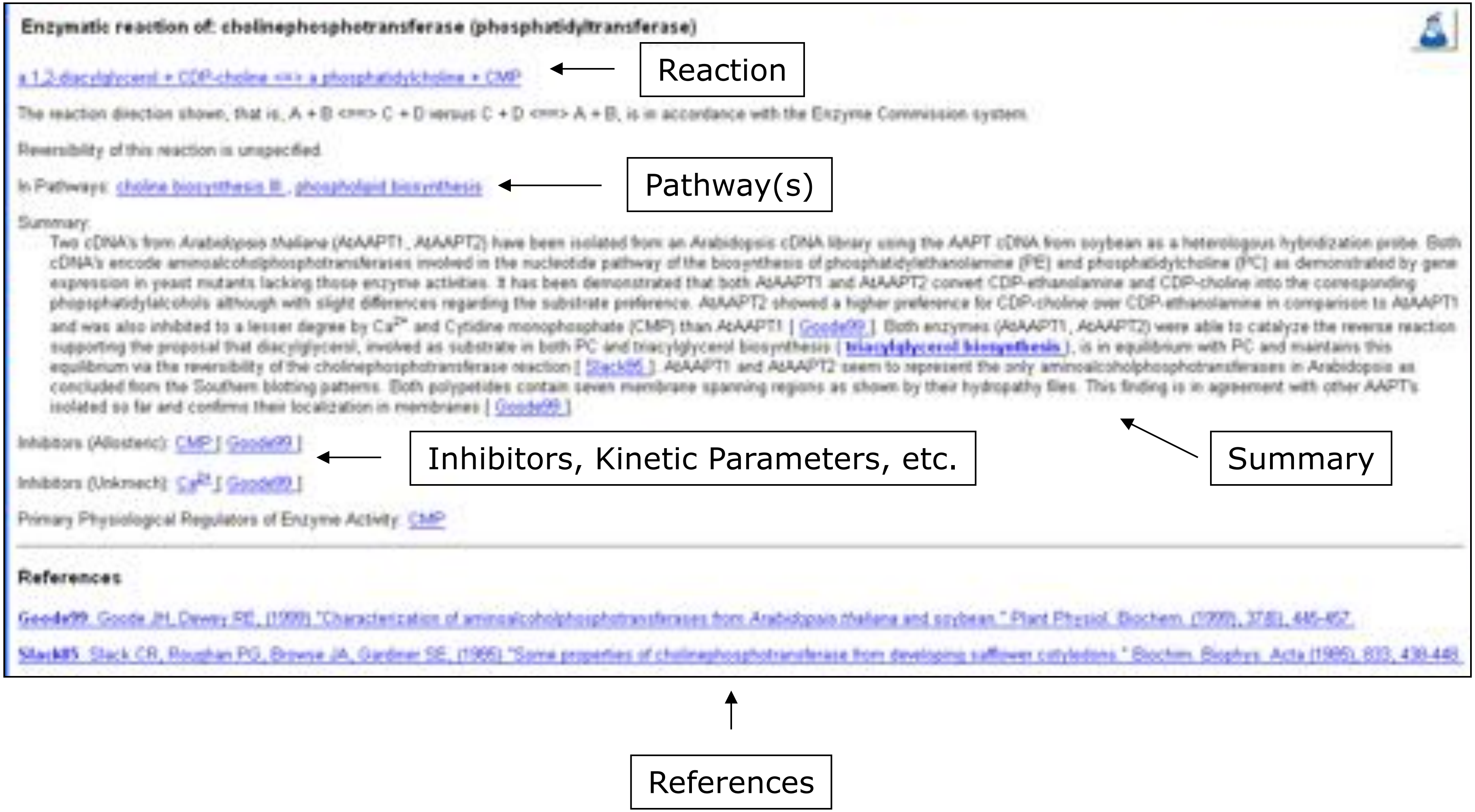
Enzyme information


AraCyc Pathway: choline biosynthesis III
Show Predicted Enzymes
More Detail
Less Detail
Cross Species Comparison
Download Genes
BioPAX format
choline-phosphate cytidylyltransferase AT4G15130
choline-phosphate cytidylyltransferase AT2G32260
2.7.7.15
choline biosynthesis I
phosphoryl-choline
CTP
diphosphate
CDP-choline
CDP-choline:1,2-diacylglycerol cholinephosphotransferase
AT3G25585
CDP-choline:1,2-diacylglycerol cholinephosphotransferase
AT1G13560
2.7.8.2
a 1,2-diacylglycerol
CMP
phospholipase D: AT4G00240
phospholipase D: AT5G25370
phospholipase D: PLD alpha
phospholipase D: PLD beta
phospholipase D: PLD gamma
phospholipase D: PLD delta
phospholipase D: PLD zeta
phospholipase D: AT1G52570
phospholipase D: AT1G55180
phospholipase D: AT3G05630
phospholipase D: AT4G11830
phospholipase D: AT4G11840
3.1.4.4
choline
an L-phosphatidate
H_2O
a phosphatidylcholine

Enzyme

Enzyme information

Arabidopsis **Enzyme: phosphatidyltransferase**



Enzymatic reaction of: cholinephosphotransferase (phosphatidyltransferase)

a 1,2-diacylglycerol + CDP-choline <==> a phosphatidylcholine + CMP

The reaction direction shown, that is, A + B <==> C + D versus C + D <==> A + B, is in accordance with the Enzyme Commission system.

Reversibility of this reaction is unspecified.

In Pathways: choline biosynthesis III , phospholipid biosynthesis

Summary:

Two cDNA's from *Arabidopsis thaliana* (AtAAPT1 , AtAAPT2) have been isolated from an Arabidopsis cDNA library using the AAPT cDNA from soybean as a heterologous hybridization probe. Both cDNA's encode aminoalcoholphosphotransferases involved in the nucleotide pathway of the biosynthesis of phosphatidylethanolamine (PE) and phosphatidylcholine (PC) as demonstrated by gene expression in yeast mutants lacking those enzyme activities. It has been demonstrated that both AtAAPT1 and AtAAPT2 convert CDP-ethanolamine and CDP-choline into the corresponding phospsphatidylalcohols although with slight differences regarding the substrate preference. AtAAPT2 showed a higher preference for CDP-choline over CDP-ethanolamine in comparison to AtAAPT1 and was also inhibited to a lesser degree by Ca^{2+} and Cytidine monophosphate (CMP) than AtAAPT1 [Goode99]. Both enzymes (AtAAPT1 , AtAAPT2) were able to catalyze the reverse reaction supporting the proposal that diacylglycerol, involved as substrate in both PC and triacylglycerol biosynthesis [**triacylglycerol biosynthesis**], is in equilibrium with PC and maintains this equilibrum via the reversibility of the cholinephosphotransferase reaction [Stack85]. AtAAPT1 and AtAAPT2 seem to represent the only aminoalcoholphosphotransferases in Arabidopsis as concluded from the Southern blotting patterns. Both polypetides contain seven membrane spanning regions as shown by their hydropathy files. This finding is in agreement with other AAPT's isolated so far and confirms their localization in membranes [Goode99]

Inhibitors (Allosteric): CMP [Goode99]

Inhibitors (Unkmech): Ca^{2+} [Goode99]

Primary Physiological Regulators of Enzyme Activity: CMP

References

Goode99: Goode JH, Dewey RE, (1999) "Characterization of aminoalcoholphosphotransferases from Arabidopsis thaliana and soybean." Plant Physiol. Biochem. (1999), 37(6), 445-457.

Stack85: Stack CR, Roughan PG, Browse JA, Gardiner SE, (1985) "Some properties of cholinephosphotransferase from developing safflower cotyledons." Biochim. Biophys. Acta (1985), 833, 438-448

Reaction

Pathway(s)

Inhibitors, Kinetic Parameters, etc.

Summary

References

How does computationally predicted data enter the PMN?

ANNOTATED GENOME

DNA sequences

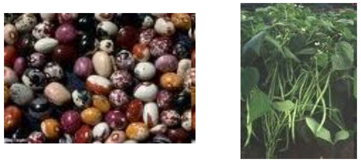
Predicted proteins

Pv1234.56.a

Predicted functions

chorismate mutase

Phaseolus vulgaris



PlantCyc / MetaCyc

PathoLogic

Single species database

chorismate mutase

prephenate aminotransferase

arogenate dehydratase

chorismate 5.4.99.5 → prephenate … 91 → L-phenylalanine

chorismate mutase

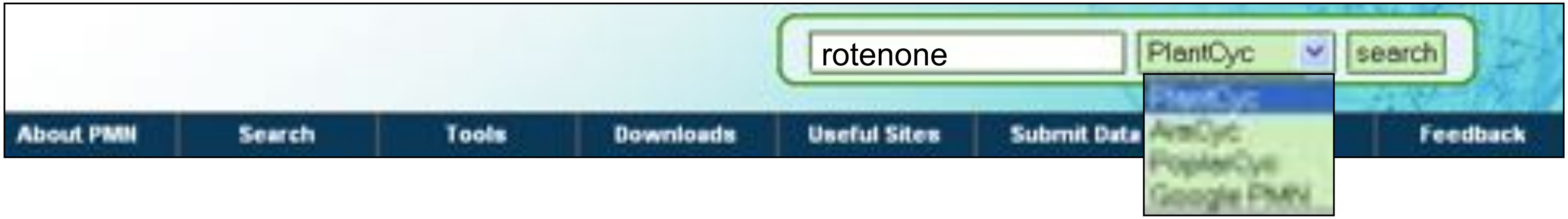
Pv1234.56.a

PhaseolusCyc

+ validation

How can researchers use the PMN?

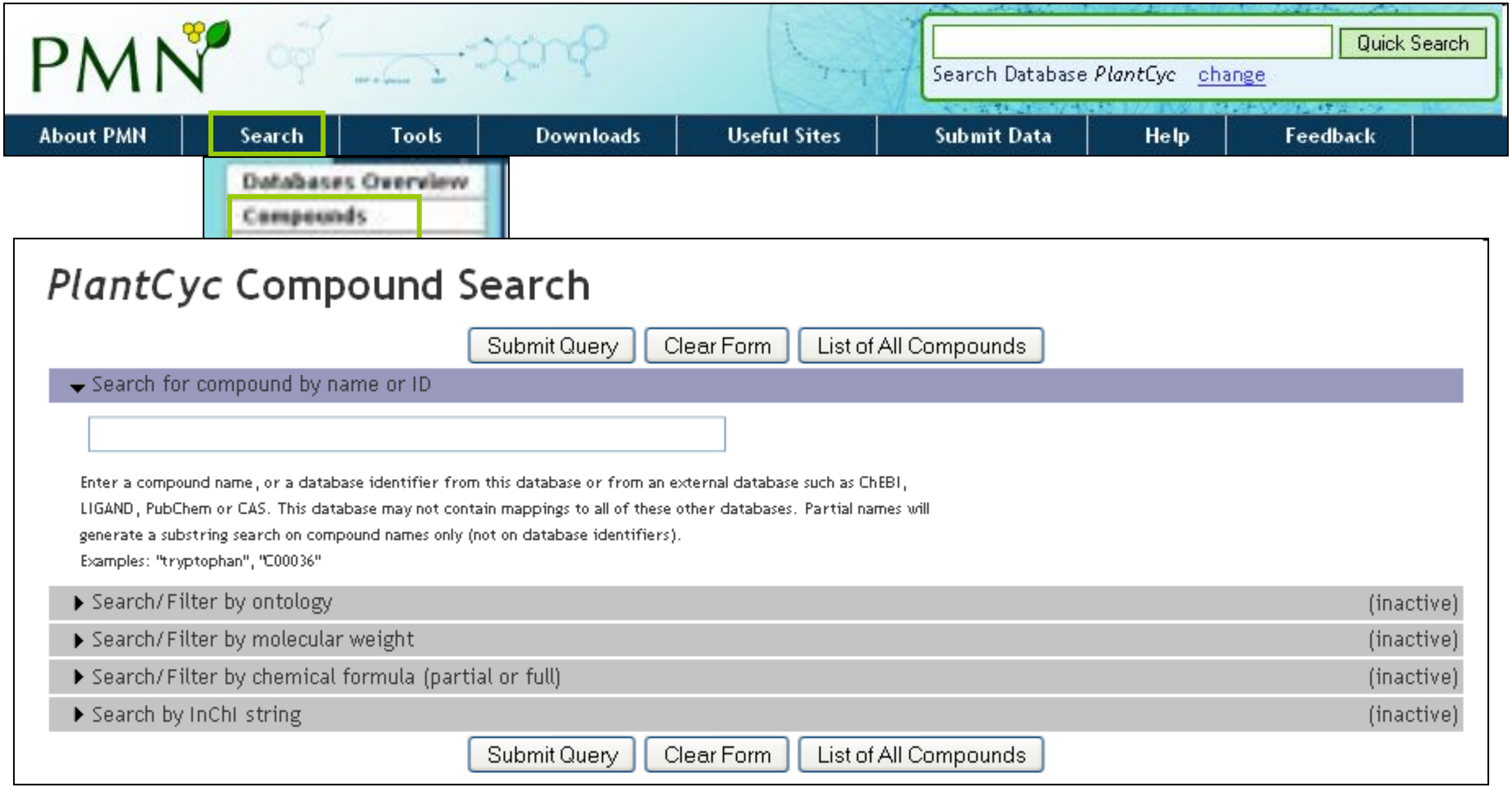
- Learn background information about particular metabolic pathways
  - Utilize simple and advanced search tools
  - Quick search bar



rotenone

  - Specific search menus

How can researchers use the PMN?



PlantCyc Compound Search

▾ Search for compound by name or ID

Enter a compound name, or a database identifier from this database or from an external database such as ChEBI, LIGAND, PubChem or CAS. This database may not contain mappings to all of these other databases. Partial names will generate a substring search on compound names only (not on database identifiers).

Examples: "tryptophan", "C00036"

▸ Search/Filter by ontology (inactive)

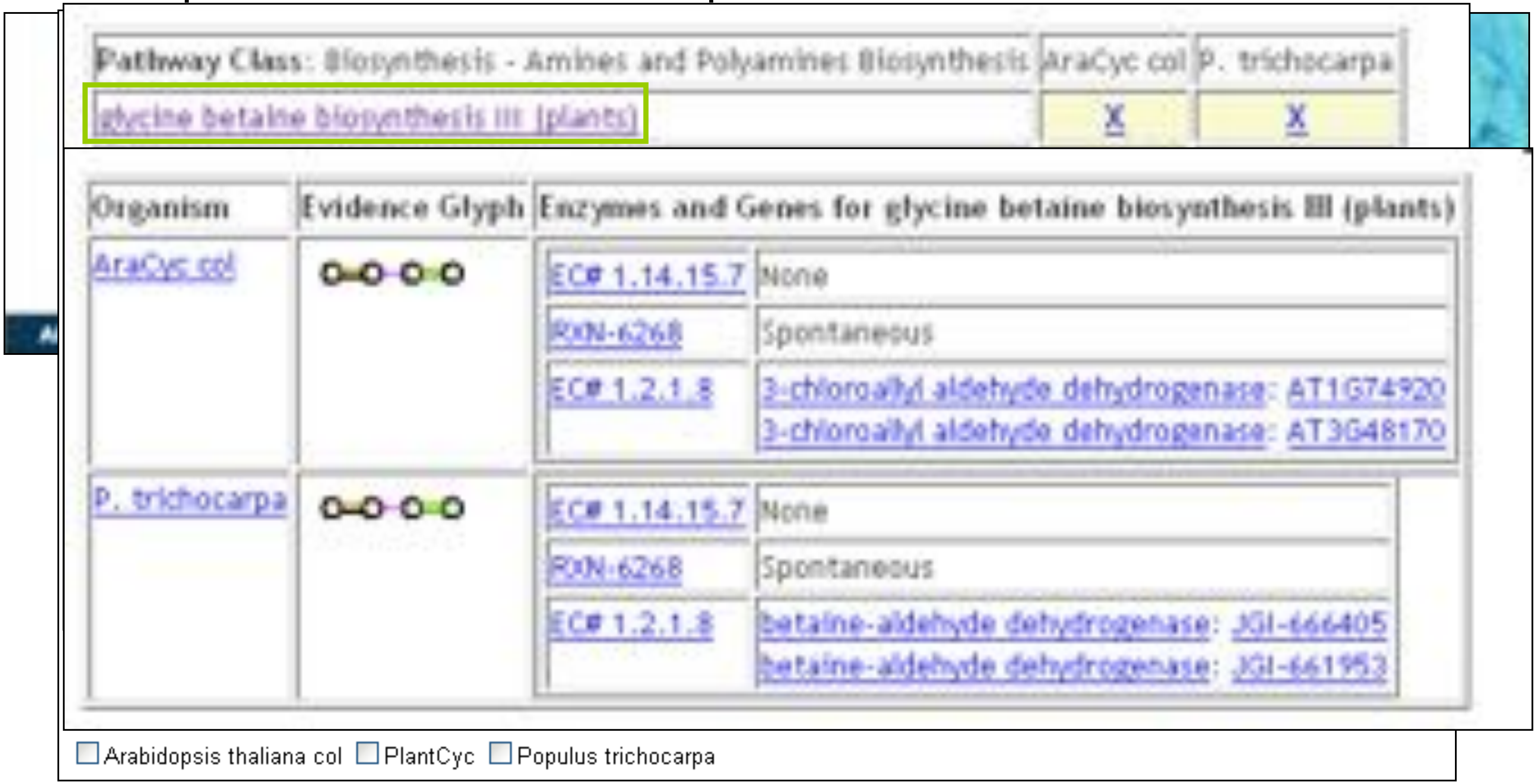
▸ Search/Filter by molecular weight (inactive)

▸ Search/Filter by chemical formula (partial or full) (inactive)

▸ Search by InChI string (inactive)

How can researchers use the PMN?

- Compare metabolism across species



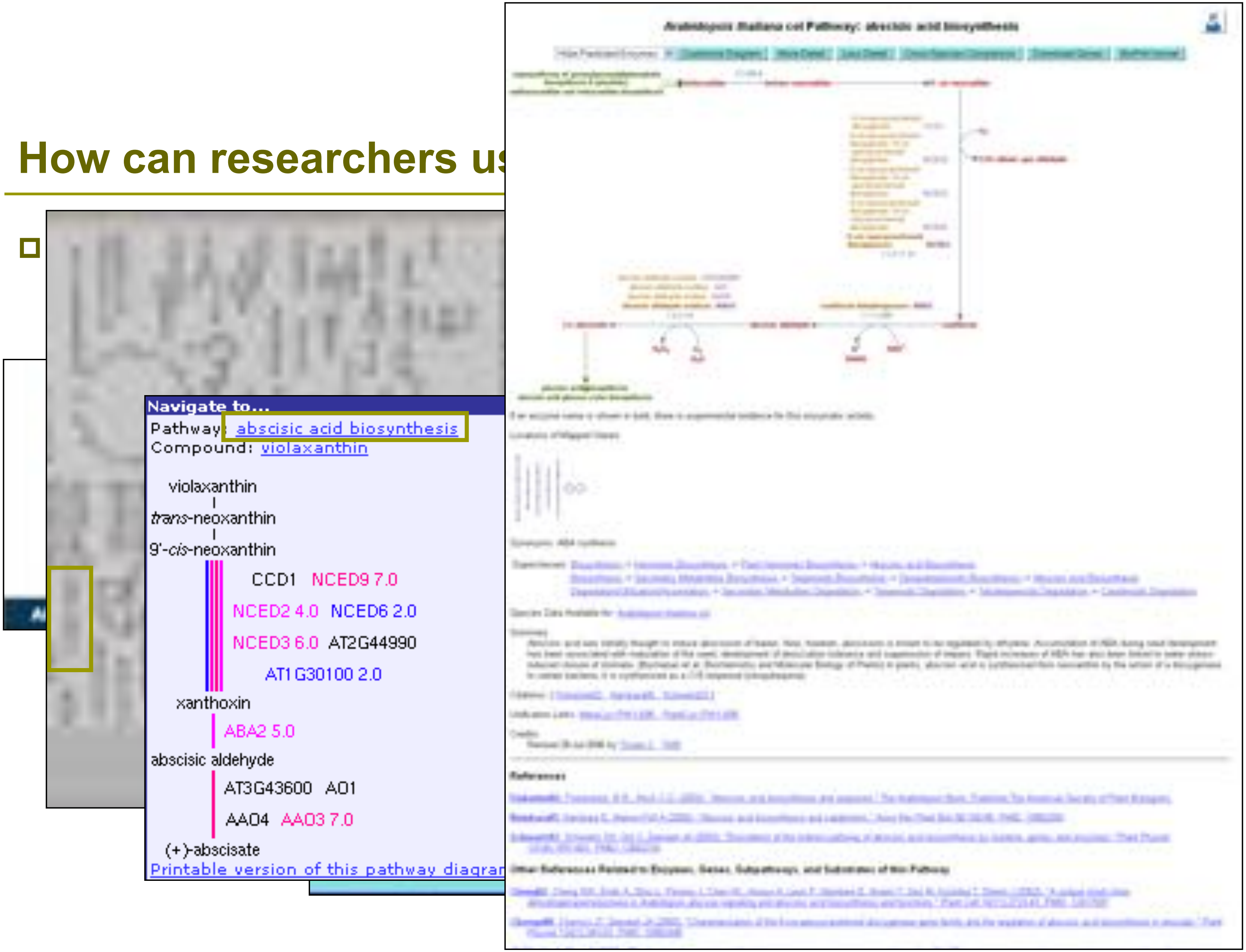
Pathway Class: Biosynthesis - Amines and Polyamines Biosynthesis	AraCyc col	P. trichocarpa
glycine betaine biosynthesis III (plants)	X	X

Organism	Evidence Glyph	Enzymes and Genes for glycine betaine biosynthesis III (plants)	
AraCyc col	○–○ ○–○	EC# 1.14.15.7	None
		RXN-6268	Spontaneous
		EC# 1.2.1.8	3-chloroallyl aldehyde dehydrogenase: AT1G74920 3-chloroallyl aldehyde dehydrogenase: AT3G48170
P. trichocarpa	○–○ ○–○	EC# 1.14.15.7	None
		RXN-6268	Spontaneous
		EC# 1.2.1.8	betaine-aldehyde dehydrogenase: JGI-666405 betaine-aldehyde dehydrogenase: JGI-661953

□ Arabidopsis thaliana col □ PlantCyc □ Populus trichocarpa

How can researchers us

□



Arabidopsis thaliana col Pathway: abscisic acid biosynthesis

Navigate to...

Pathway: abscisic acid biosynthesis

Compound: violaxanthin

violaxanthin

trans-neoxanthin

9'-*cis*-neoxanthin

CCD1 NCED9 7.0

NCED2 4.0 NCED6 2.0

NCED3 6.0 AT2G44990

AT1G30100 2.0

xanthoxin

ABA2 5.0

abscisic aldehyde

AT3G43600 AO1

AAO4 AAO3 7.0

(+)-abscisate

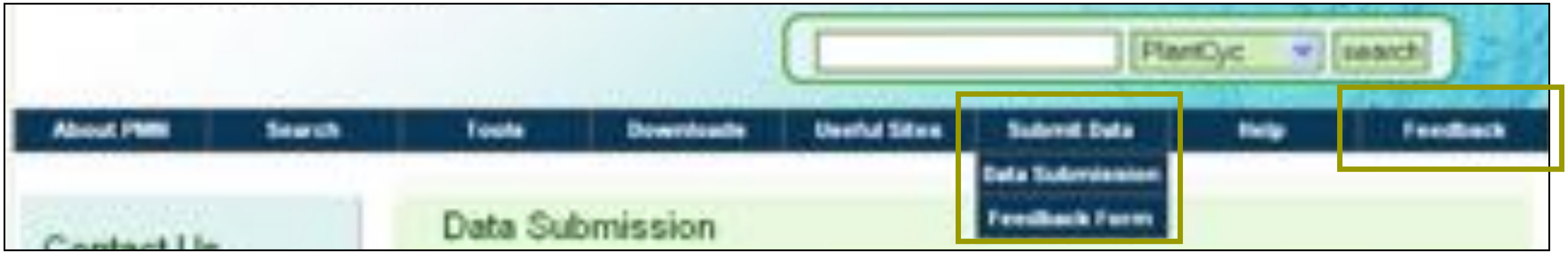
Printable version of this pathway diagram

How will the PMN grow in the future?

- Help from the research community!!!
- You are the experts with great knowledge to share!

Building better databases together

- To submit data, report an error, or volunteer to help validate . . .
  - Send an e-mail: **curator@plantcyc.org**
  - Use data submission “tools”



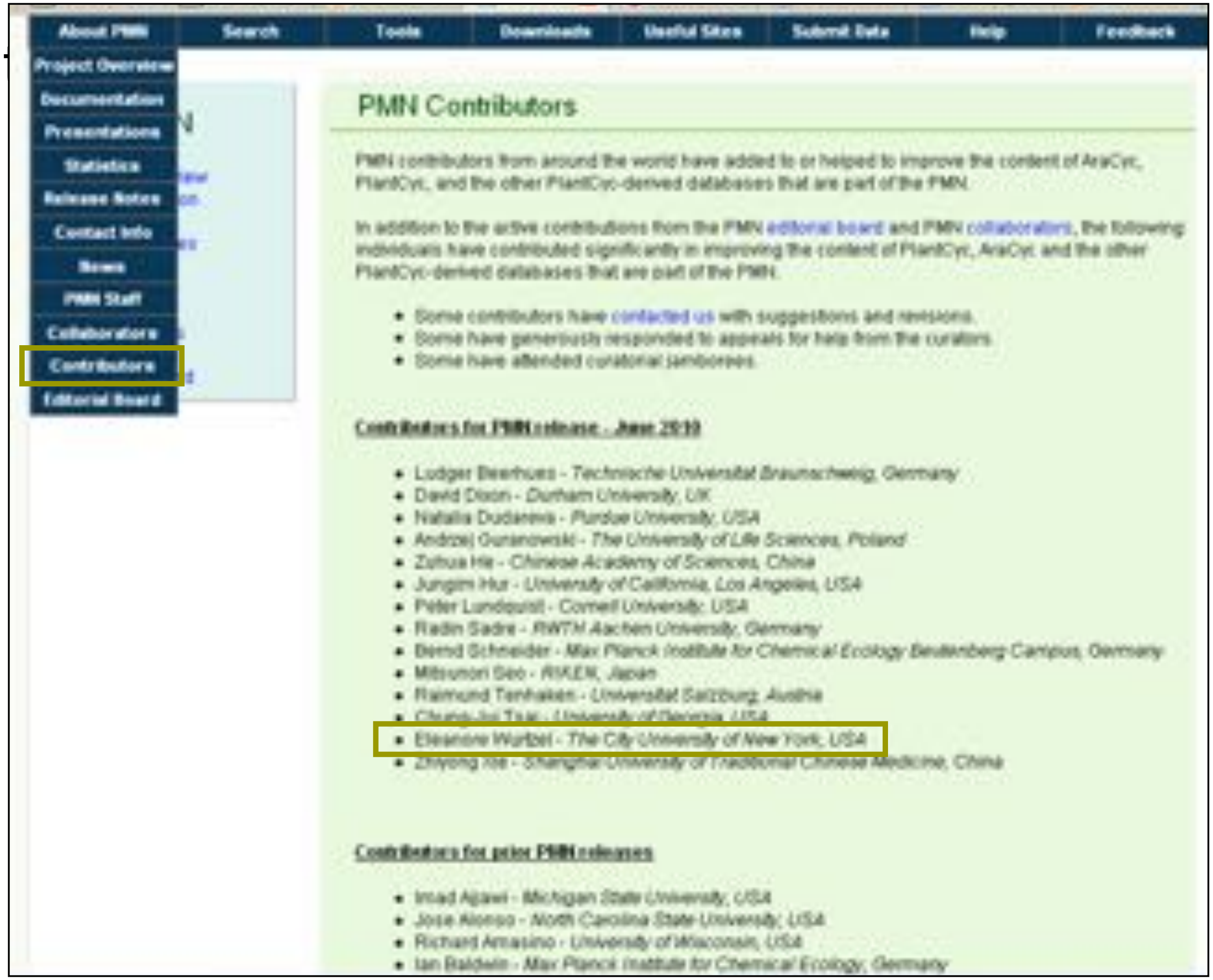
  - **Meet with me this afternoon**
    - . . . or later this week
    - . . . or later this year

Building better databases together

- **Details are very, very welcome!!**
  - Reactions:
    - All co-factors, co-substrates, etc.
    - EC suggestions – partial or full
  - Compounds
    - Structure – visual representation / compound file (e.g. mol file)
    - Synonyms
    - Unique IDs (e.g. ChEBI, CAS, KEGG)
  - Enzymes
    - Unique IDs (e.g. At2g46480, UniProt, Genbank)
    - Specific reactions catalyzed

Community gratitude

- We t



PMN Contributors

PMN contributors from around the world have added to or helped to improve the content of AraCyc, PlantCyc, and the other PlantCyc-derived databases that are part of the PMN.

In addition to the active contributions from the PMN editorial board and PMN collaborators, the following individuals have contributed significantly in improving the content of PlantCyc, AraCyc and the other PlantCyc-derived databases that are part of the PMN.

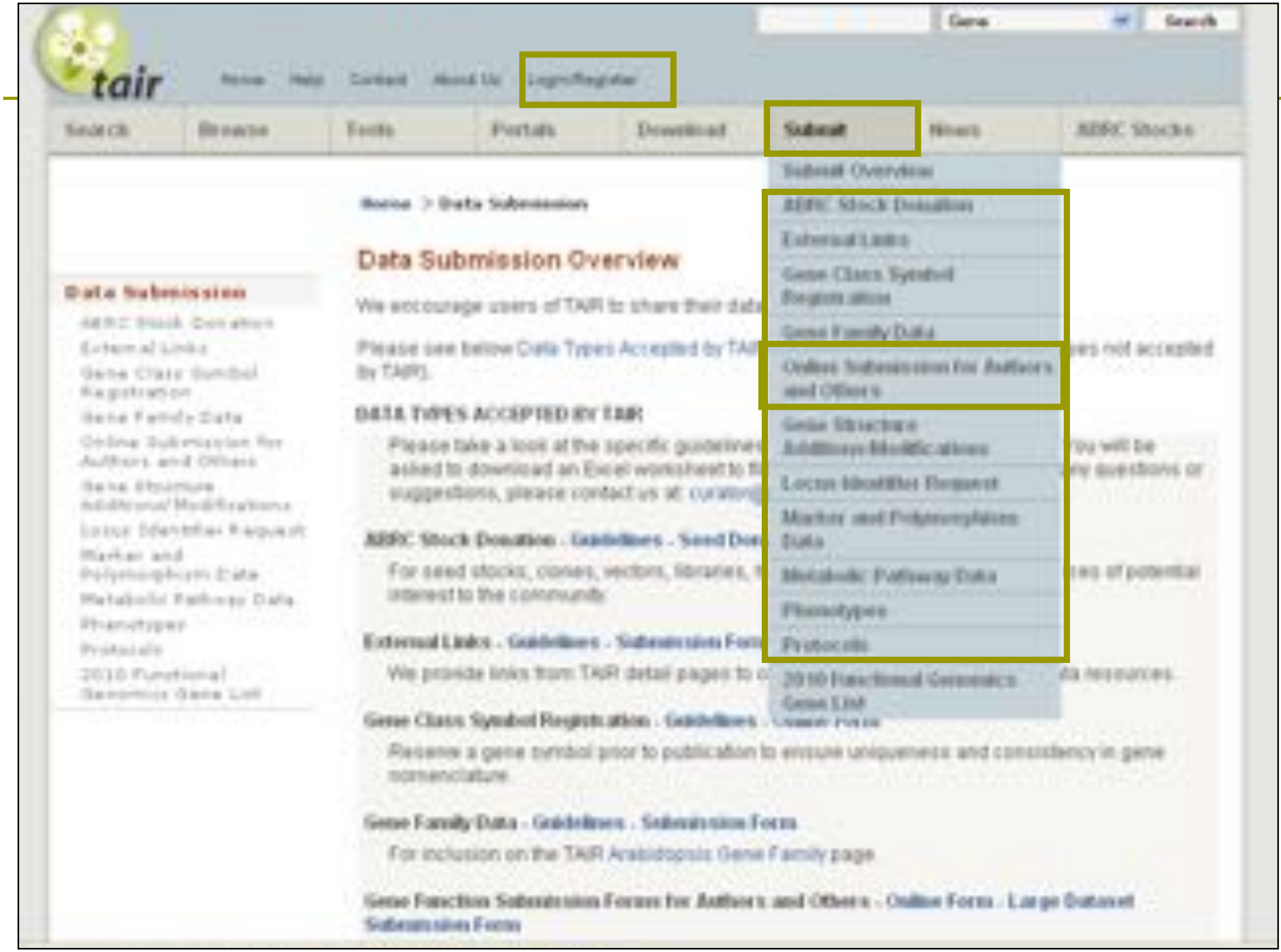
- Some contributors have contacted us with suggestions and revisions.
- Some have generously responded to appeals for help from the curators.
- Some have attended curatorial jamborees.

Contributors for PMN release - June 2010

- Ludger Beerhues - *Technische Universitat Braunschweig, Germany*
- David Dixon - *Durham University, UK*
- Natalia Dudareva - *Purdue University, USA*
- Andrzej Guranowski - *The University of Life Sciences, Poland*
- Zuhua He - *Chinese Academy of Sciences, China*
- Jungim Hur - *University of California, Los Angeles, USA*
- Peter Lundquist - *Cornell University, USA*
- Radin Sadre - *RWTH Aachen University, Germany*
- Bernd Schneider - *Max Planck Institute for Chemical Ecology Beutenberg Campus, Germany*
- Mitsunori Seo - *RIKEN, Japan*
- Raimund Tenhaken - *Universitat Salzburg, Austria*
- Chung-Jui Tsai - *University of Georgia, USA*
- Eleanore Wurtzel - *The City University of New York, USA*
- Zhiyong Xie - *Shanghai University of Traditional Chinese Medicine, China*

Contributors for prior PMN releases

- Imad Ajjawi - *Michigan State University, USA*
- Jose Alonso - *North Carolina State University, USA*
- Richard Amasino - *University of Wisconsin, USA*
- Ian Baldwin - *Max Planck Institute for Chemical Ecology, Germany*



Home > Data Submission

Data Submission Overview

We encourage users of TAIR to share their data

Please see below Data Types Accepted by TAI … pes not accepted by TAIR).

DATA TYPES ACCEPTED BY TAIR

Please take a look at the specific guideline … You will be asked to download an Excel worksheet to fi … any questions or suggestions, please contact us at: curator

ABRC Stock Donation - Guidelines - Seed Don

For seed stocks, clones, vectors, libraries, t … es of potential interest to the community

External Links - Guidelines - Submission Form

We provide links from TAIR detail pages to c … ta resources.

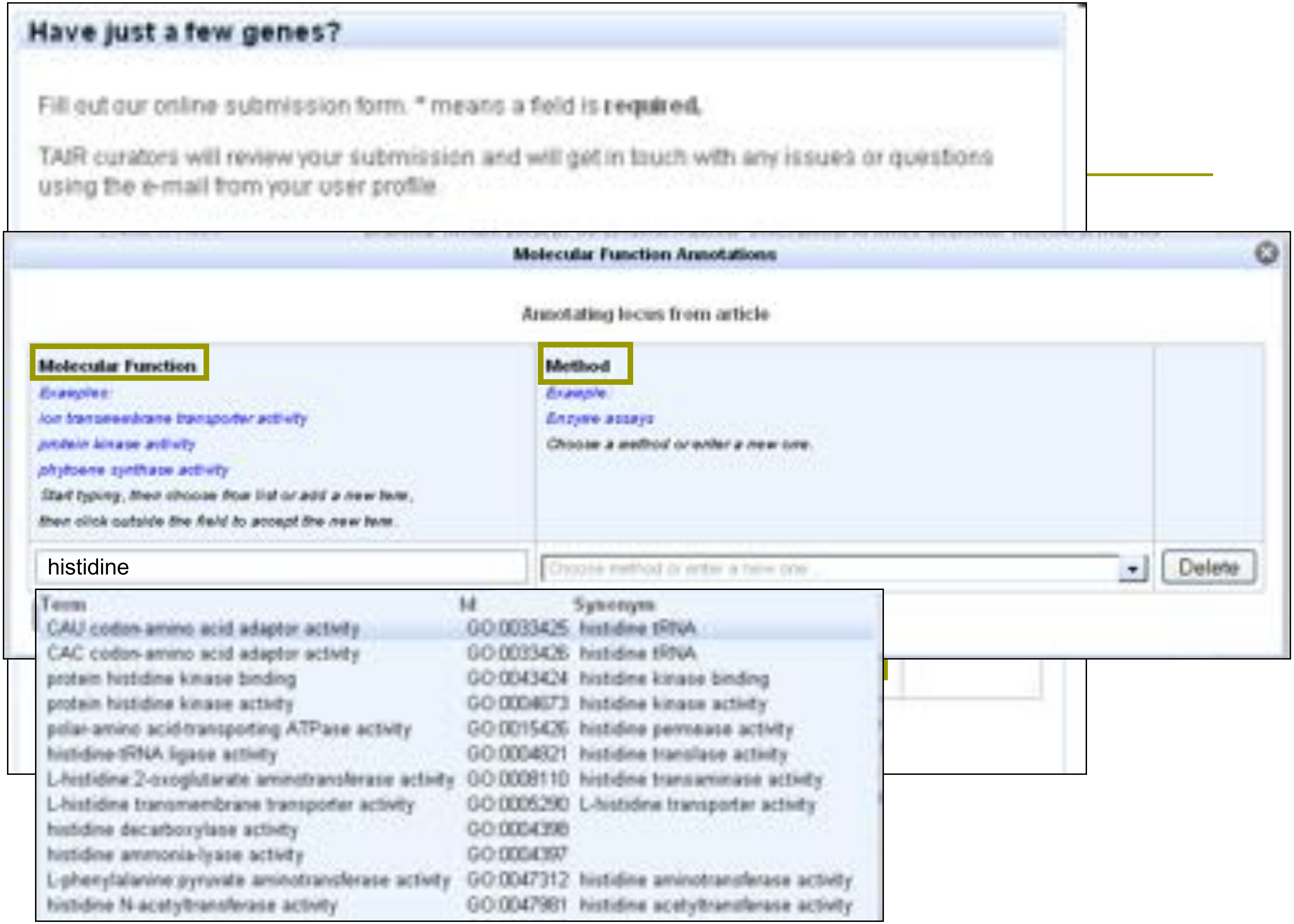
Gene Class Symbol Registration - Guidelines - Create Form

Reserve a gene symbol prior to publication to ensure uniqueness and consistency in gene nomenclature

Gene Family Data - Guidelines - Submission Form

For inclusion on the TAIR Arabidopsis Gene Family page

Gene Function Submission Forms for Authors and Others - Online Form - Large Dataset Submission Form



Have just a few genes?

Fill out our online submission form. * means a field is **required.**

TAIR curators will review your submission and will get in touch with any issues or questions using the e-mail from your user profile

Molecular Function Annotations

Annotating locus from article

Molecular Function	**Method**	
Examples: *ion transmembrane transporter activity* *protein kinase activity* *phytoene synthase activity* *Start typing, then choose from list or add a new term,* *then click outside the field to accept the new term.*	*Example:* *Enzyme assays* *Choose a method or enter a new one.*	
histidine	Choose method or enter a new one ...	Delete

Term	Id	Synonym
CAU codon-amino acid adaptor activity	GO:0033425	histidine tRNA
CAC codon-amino acid adaptor activity	GO:0033426	histidine tRNA
protein histidine kinase binding	GO:0043424	histidine kinase binding
protein histidine kinase activity	GO:0004673	histidine kinase activity
polar-amino acid-transporting ATPase activity	GO:0015426	histidine permease activity
histidine-tRNA ligase activity	GO:0004821	histidine translase activity
L-histidine:2-oxoglutarate aminotransferase activity	GO:0008110	histidine transaminase activity
L-histidine transmembrane transporter activity	GO:0005290	L-histidine transporter activity
histidine decarboxylase activity	GO:0004398	
histidine ammonia-lyase activity	GO:0004397	
L-phenylalanine:pyruvate aminotransferase activity	GO:0047312	histidine aminotransferase activity
histidine N-acetyltransferase activity	GO:0047981	histidine acetyltransferase activity

Biological networking . . .

- Please use our data
- Please use our tools
- Please help us to improve our databases!
- Please contact us if we can be of any help!

tair

PMN
Plant Metabolic Network

AraCyc

curator@arabidopsis.org

curator@plantcyc.org

www.arabidopsis.org

www.plantcyc.org

TAIR and PMN Acknowledgements

Sue Rhee *(PI - PMN)*
Eva Huala (*PI-TAIR*)
Peifen Zhang *(Director-PMN)*

Current Curators:

-Tanya Berardini (*lead curator*)
- Philippe Lamesch (*lead curator)*
- Donghui Li (*curator*)
- Dave Swarbreck (*former lead curator*)

- Debbie Alexander (*curator*)
- A. S. Karthikeyan (*curator*)
- Marga Garcia (*curator)*
- Leonore Reiser

PMN Collaborators:

- Peter Karp (SRI)
- Ron Caspi (SRI)
- Suzanne Paley (SRI)
- SRI Tech Team
- Lukas Mueller (SGN)
- Anuradha Pujar (SGN)
- Gramene and MedicCyc

Current Tech Team Members:

- Bob Muller (*Manager*)
- Larry Ploetz (*Sys. Administrator*)
- Anjo Chi
- Raymond Chetty
- Cynthia Lee
- Shanker Singh
- Chris Wilks

PMN project post-doc

- Lee Chae

National Science Foundation
WHERE DISCOVERIES BEGIN

CARNEGIE
INSTITUTION FOR
SCIENCE

Biological networking . . .

- Please use our data
- Please use our tools
- Please help us to improve our databases!
- Please contact us if we can be of any help!

tair

PMN
Plant Metabolic Network

AraCyc

curator@arabidopsis.org

curator@plantcyc.org

www.arabidopsis.org

www.plantcyc.org

Out-takes

- The following slides are relevant but were removed from the presentation due to time constraints

tair

PMN
Plant Metabolic Network

AraCyc

curator@arabidopsis.org

curator@plantcyc.org

www.arabidopsis.org

www.plantcyc.org

Arabidopsis has good model organism traits

- Fast life cycle (6 weeks)
- Thousands of plants fit in a small space
- Fairly easy to grow
- Thousands of seeds produced by each plant
- Self-fertile (in-breeding)
- Many different subspecies/ecotypes
- Serves as a good model for **crop plants**

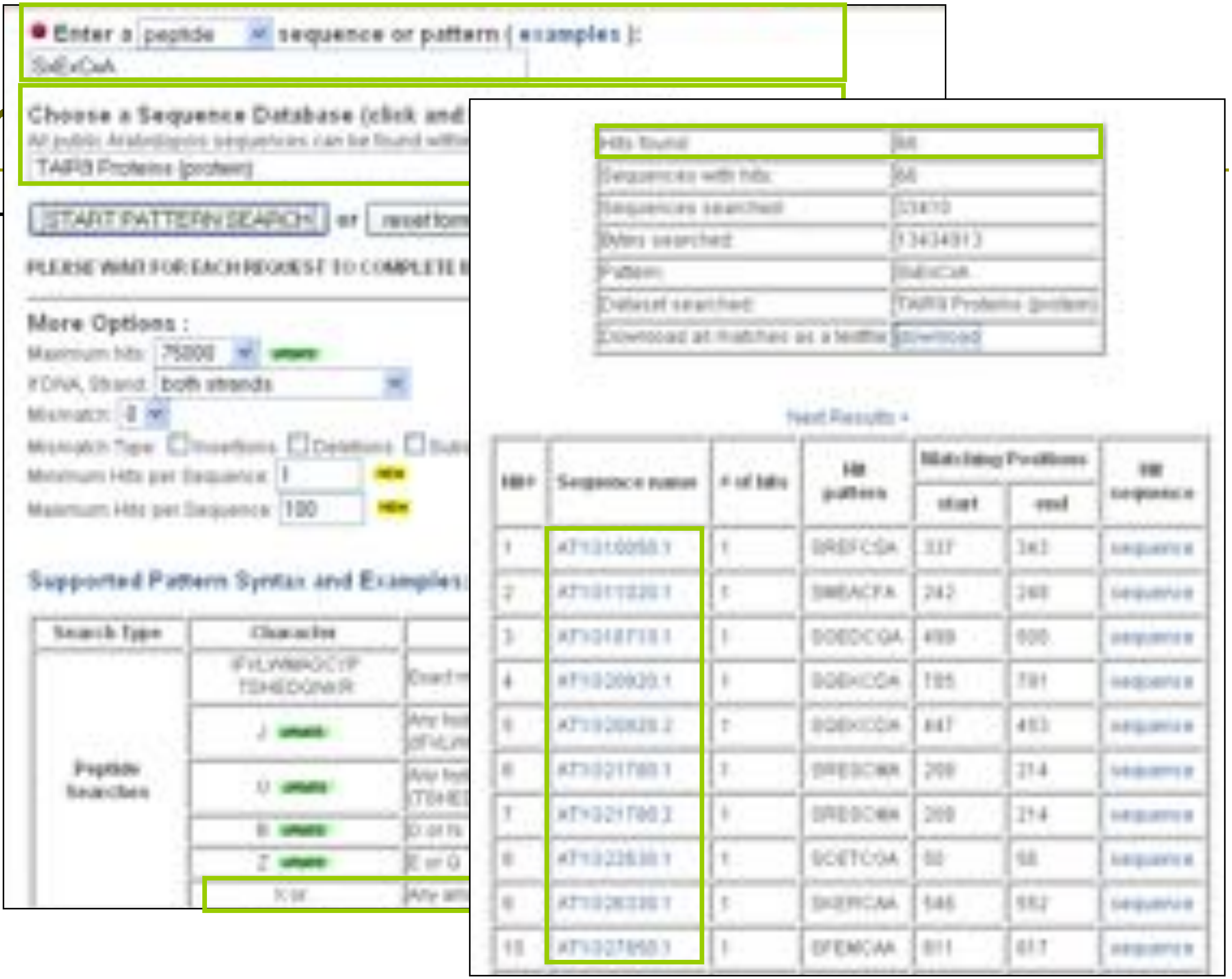
- But why Arabidopsis instead of **other plants**?

Arabidopsis data explosion

- TONS of data are generated about Arabidopsis
  - Over 2400 “Arabidopsis” articles published each year are indexed in PubMed
  - Tens of thousands of mutants have been generated
  - Hundreds of microarray experiments have been performed
  - Proteomics and metabolomics studies are becoming popular
  - “1001” Arabidopsis genomes are being sequenced
  - Large-scale phenotypic studies are scheduled to start soon
- TAIR tries to bring data together to benefit scientists and society

Pr

- T



● Enter a peptide ▾ sequence or pattern (examples):
SxExCxA

Choose a Sequence Database (click and
All public Arabidopsis sequences can be found within
TAIR9 Proteins [protein]

START PATTERN SEARCH or reset form

PLEASE WAIT FOR EACH REQUEST TO COMPLETE B

More Options :

Maximum hits: 75000 ▾ update
If DNA, Strand: both strands ▾
Mismatch: 0 ▾
Mismatch Type: □ Insertions □ Deletions □ Subs
Minimum Hits per Sequence: 1 new
Maximum Hits per Sequence: 100 new

Supported Pattern Syntax and Examples:

Search Type	Character	
Peptide Searches	IFVLWMAGCYP TSHEDQNKR	Exact m
	J update	Any hyd (IFVLW
	O update	Any hyd (TSHED
	B update	D or N
	Z update	E or Q
	X or	Any am

Hits found	66
Sequences with hits	66
Sequences searched	33410
Bytes searched	13434913
Pattern	SxExCxA
Dataset searched	TAIR9 Proteins (protein)
Download all matches as a textfile	download

Next Results »

Hit #	Sequence name	# of hits	Hit pattern	Matching Positions		Hit sequence
				start	end	
1	AT1G10050.1	1	SREFCSA	337	343	sequence
2	AT1G11220.1	1	SMEACFA	242	248	sequence
3	AT1G18710.1	1	SQEDCQA	499	505	sequence
4	AT1G20920.1	1	SQEKCQA	195	201	sequence
5	AT1G20920.2	1	SQEKCQA	447	453	sequence
6	AT1G21780.1	1	SRESCWA	208	214	sequence
7	AT1G21780.2	1	SRESCWA	208	214	sequence
8	AT1G22630.1	1	SQETCQA	52	58	sequence
9	AT1G26330.1	1	SKERCAA	546	552	sequence
10	AT1G27850.1	1	SFEMCAA	611	617	sequence

What data are in the PMN?

- Plants provide crucial benefits to the ecosystem and humanity

- A better understanding of plant metabolism may contribute to:

  - More nutritious foods
  - New medicines
  - More pest-resistant plants
  - Higher photosynthetic capacity and yield in crops
  - Better biofuel feedstocks
  - Improved industrial inputs (e.g. oils, fibers, etc.)
  - Enhanced ability to do rational metabolic engineering
  - . . . many more applications

- How can the PMN help?

What metabolites are in the PMN?

- “Primary” metabolites (“essential”)
  - sugars
    - glucose, fructose, . . .
  - amino acids
    - tryptophan, glutamine, . . .
  - lipids
    - waxes, phosphatidylcholine , . . .
  - vitamins
    - A, E, K, C, thiamine, niacin, . . .
  - hormones
    - auxin, brassinosteroids, ethylene . . .

What metabolites are in the PMN?

- “Secondary” metabolites (important, but not “essential”)
  - terpenoids
    - orzyalexin, menthol, . . .
  - organosulfur compounds
    - glucosinolates, camalexin . . .
  - isoflavonoids
    - glyceollin, daidzein. . .
  - alkaloids
    - caffeine, capsaicin, . . .
  - polyketides
    - aloesone, . . .
  - many more . . .

How do computational predictions enter the PMN?

- **New sets of DNA sequences -> predicted proteome**
  - Genomes are sequenced
  - Large RNAseq or EST data sets are created
- **Predicted proteome -> set of predicted enzyme functions**
  - Performed using computer algorithms
  - The PMN is working to develop better algorithms to increase the accuracy of the predictions
- **Set of predicted enzyme functions -> set of predicted metabolic pathways**
  - The PathoLogic program uses a reference database to predict the metabolic pathways for the enzyme sets
- **Set of predicted metabolic pathways -> set of “validated” metabolic pathways**
  - Curators remove incorrect information and add additional data