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!
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 repositories / databases

NCBI Resources

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 = The Arabidopsis Information Resource
- Why Arabidopsis?
- What does TAIR do?
- What can you do with TAIR?
Introduction to Arabidopsis

- Basic facts:
  - “small weed related to mustard”
  - also known as “mouse ear cress”
  - can grow to 20-25 cm tall
  - annual (or occasionally biennial) plant
  - member of the Brassicaceae family
    - broccoli
    - cauliflower
    - radish
    - cabbage
  - found around the northern hemisphere

- Why do so many people study THIS plant?
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 does TAIR do?

- Curators and computer tech team members work together under great directors.
- TAIR develops internal data sets and resources.
- TAIR links to external data sets and resources.
- TAIR provides free online access to everyone.

Funded by the National Science Foundation of the USA

Started in 1999

Available at www.arabidopsis.org

Dr. Eva Huala
Director

Dr. Sue Rhee
Co-PI

Curators

Computer tech team members
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
Functional curation at TAIR

- Functional curators try to answer the questions:
  - What does every gene/protein in Arabidopsis do?
  - When and where does it act?
  - We hope that this information can inform research in other plants

- Functional curation requires controlled vocabularies
  - Allow cross-species comparisons
  - TAIR curators work to develop and agree upon common terms

The seed-bearing structure in angiosperms, formed from the ovary after flowering

Plant Ontology:
Structure:
PO:0009001
Catalysis of the reaction:
IAA + UDP-D-glucose = indole-3-acetyl-beta-1-D-glucose + UDP

Functional curation at TAIR

IAA-Glu synthetase activity
IAA-glucose synthase activity
IAAGlu 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

Gene Ontology:
Molecular function:
GO:0047215

indole-3-acetate beta-glucosyltransferase activity
## Functional curation at TAIR

### Table: Gene Annotation Examples

<table>
<thead>
<tr>
<th>Gene</th>
<th>Molecular Function</th>
<th>Expression Pattern</th>
<th>Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT5G14570</td>
<td>nitrate transport</td>
<td>petal differentiation and expansion stage, 4 anthesis, 4 leaf senescence stage, C globular stage, D bilobal stage, E expanded cotyledon stage, F</td>
<td>The Arabidopsis Information Resource 2003-03-29</td>
</tr>
<tr>
<td>AT5G14570</td>
<td>nitrate transmembrane transporter activity</td>
<td>inferred from expression pattern</td>
<td>The Arabidopsis Information Resource 2007-09-13</td>
</tr>
</tbody>
</table>

### Categories
- GO Biological Process
- GO Cellular Component
- GO Molecular Function
- Growth and Developmental
**Functional curation at TAIR**

<table>
<thead>
<tr>
<th>Associated Loci</th>
<th>AT5G63980</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutagen</td>
<td>ethylmethane sulfonate</td>
</tr>
<tr>
<td>Inheritance</td>
<td>recessive</td>
</tr>
<tr>
<td>Allele Type</td>
<td>loss-of-function</td>
</tr>
</tbody>
</table>

**Description**
- Contains a substitution of G to A at the 1258th base pair of AT5g63980 DNA, resulting in a single amino acid change from glycine to aspartic acid at the 217th amino acid of the GAAT protein (GalR).<br>
- Published ID: T80591934.

**Phenotype**
- Loss-of-function seedlings are more tolerant to drought after 12 d of drought treatment than wild-type plants; increase carbohydrates and a sugar derivative accumulation; less difference in stomatal index on the adaxial side; content does not change under potentials; wild plants:
  - Dry growth and development;
  - Drought changes to the internal palisade structure, smaller type;
  - 1:1 hybrid is significantly due to their expression in Col-0; size, galactos, glucose, accumulation of unknown sugars and alcohols, very low starch.

**Image**
- Image provided by Barry Pogson
Providing access to external tools and data
Tech team members and curators

Load data sets into existing tools

BLAST
GBrowse

Providing Tools at TAIR
Providing Tools at TAIR
Providing Tools at TAIR
Providing Tools at TAIR

Tech team members and curators

Develop new tools

SeqViewer
Providing Tools at TAIR

- Tech team members and curators
  - Create quick search options
  - Create advanced search pages
Other Resources at TAIR

TAIR Community Detail [Help]

Name: Eleanor 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: 718-960-8843
Lab Phone: 718-960-4984
Mobile Phone: 516-381-5013
Affiliations: Organization: *Eleanor Wurtzel Laboratory
Job Title: "denotes Primary Investigator for this organization"
Record last updated: 10/31/2006
Keeping up with TAIR

RSS feeds

Breaking news

Plant biology jobs, graduate, and post-doc opportunities

Twitter

www.twitter.com/tair_news

Facebook

http://www.facebook.com/tairnews
<table>
<thead>
<tr>
<th>City</th>
<th>Visits</th>
<th>Pages/Visit</th>
<th>Avg. Time on Site</th>
<th>% New Visits</th>
<th>Bounce Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ithaca</td>
<td>675</td>
<td>6.30</td>
<td>00:07:13</td>
<td>10.22%</td>
<td>25.19%</td>
</tr>
<tr>
<td>New York</td>
<td>527</td>
<td>5.75</td>
<td>00:08:14</td>
<td>17.08%</td>
<td>33.02%</td>
</tr>
<tr>
<td>Cold Spring Harbor</td>
<td>248</td>
<td>6.48</td>
<td>00:10:01</td>
<td>14.94%</td>
<td>19.09%</td>
</tr>
<tr>
<td>Syracuse</td>
<td>199</td>
<td>7.01</td>
<td>00:08:48</td>
<td>14.00%</td>
<td>29.00%</td>
</tr>
<tr>
<td>Upton</td>
<td>99</td>
<td>4.66</td>
<td>00:05:43</td>
<td>24.44%</td>
<td>21.11%</td>
</tr>
<tr>
<td>Stony Brook</td>
<td>71</td>
<td>2.26</td>
<td>00:01:07</td>
<td>4.23%</td>
<td>16.90%</td>
</tr>
<tr>
<td>Bronx</td>
<td>67</td>
<td>7.39</td>
<td>00:07:22</td>
<td>29.03%</td>
<td>41.94%</td>
</tr>
<tr>
<td>Cortland</td>
<td>38</td>
<td>2.53</td>
<td>00:02:30</td>
<td>0.00%</td>
<td>84.21%</td>
</tr>
<tr>
<td>Huntington Station</td>
<td>36</td>
<td>9.69</td>
<td>00:22:34</td>
<td>0.00%</td>
<td>19.44%</td>
</tr>
<tr>
<td>Briarcliff Manor</td>
<td>18</td>
<td>1.22</td>
<td>00:00:04</td>
<td>0.00%</td>
<td>94.44%</td>
</tr>
</tbody>
</table>
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
  - Identify 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 = The Plant Metabolic Network
  - Created in 2008
  - Funded by the National Science Foundation

- What is

- What data

- How do

- How can

- How can you help the PMN to grow?
What is the PMN?

- A Network of Plant Metabolic Pathway Databases and Communities

**Major goals:**

- Create metabolic pathway databases to catalog all of the biochemical pathways present in specific species
- Create PlantCyc - a comprehensive multi-plant pathway database
- Create an automated pathway prediction “pipeline”
- Create a website to bring together researchers working on plant metabolism

PMN website: [www.plantcyc.org](http://www.plantcyc.org)

- 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 . . .
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:**

<table>
<thead>
<tr>
<th>PGDB</th>
<th>Plant</th>
<th>Source</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>RiceCyc **</td>
<td>Rice</td>
<td>Gramene</td>
<td>some curation</td>
</tr>
<tr>
<td>SorghumCyc</td>
<td>Sorghum</td>
<td>Gramene</td>
<td>no curation</td>
</tr>
<tr>
<td>MedicCyc **</td>
<td>Medicago</td>
<td>Noble Foundation</td>
<td>some curation</td>
</tr>
<tr>
<td>LycoCyc **</td>
<td>Tomato</td>
<td>Sol Genomics Network</td>
<td>some curation</td>
</tr>
<tr>
<td>PotatoCyc</td>
<td>Potato</td>
<td>Sol Genomics Network</td>
<td>no curation</td>
</tr>
<tr>
<td>CapCyc</td>
<td>Pepper</td>
<td>Sol Genomics Network</td>
<td>no curation</td>
</tr>
<tr>
<td>NicotianaCyc</td>
<td>Tobacco</td>
<td>Sol Genomics Network</td>
<td>no curation</td>
</tr>
<tr>
<td>PetuniaCyc</td>
<td>Petunia</td>
<td>Sol Genomics Network</td>
<td>no curation</td>
</tr>
<tr>
<td>CoffeaCyc</td>
<td>Coffee</td>
<td>Sol Genomics Network</td>
<td>no curation</td>
</tr>
</tbody>
</table>

** Significant numbers of genes from these databases have been integrated into PlantCyc **
## PMN database content statistics

<table>
<thead>
<tr>
<th>Category</th>
<th>PlantCyc 4.0</th>
<th>AraCyc 7.0</th>
<th>PoplarCyc 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathways</td>
<td>685</td>
<td>369</td>
<td>288</td>
</tr>
<tr>
<td>Enzymes</td>
<td>11058</td>
<td>5506</td>
<td>3420</td>
</tr>
<tr>
<td>Reactions</td>
<td>2929</td>
<td>2418</td>
<td>1707</td>
</tr>
<tr>
<td>Compounds</td>
<td>2966</td>
<td>2719</td>
<td>1397</td>
</tr>
<tr>
<td>Organisms</td>
<td>343</td>
<td>1</td>
<td>1*</td>
</tr>
</tbody>
</table>

### Species (A - D) [back to top]
- Abies grandis (yellow fir, white fir, silver fir, lowland fir, grand fir)
- Acer pseudoplatanus
- Aesculus hippocastanum
- Adonis annua (pheasants-eye, blood-drop)
- Ajuga reptans
- Alchemilla umbellata
- Allostephanus roseus
- Aloe arborescens
- Anemone nemorosa
- Anemone pulsatilla
- Anemone ranunculoides
- Anemone virgata
- Anemone x hybrida
- Anaphalis margaritacea
- Anaphalis nemorosa
- Anaphalis quinquefolia
- Anaphalis tenuifolia
- Anaphalis vulgaris
- Anaphalis x hybrida
- Anaphalis x nemorosa
- Anaphalis x quinquefolia
- Anaphalis x tenuifolia
- Anaphalis x vulgaris
- Anaphalis x x hybrida
- Anaphalis x nemorosa
- Anaphalis x quinquefolia
- Anaphalis x tenuifolia
- Anaphalis x vulgaris
- Anaphalis x x hybrida
- Anaphalis x nemorosa
- Anaphalis x quinquefolia
- Anaphalis x tenuifolia
- Anaphalis x vulgaris
- Anaphalis x x hybrida
- Anaphalis x nemorosa
- Anaphalis x quinquefolia
- Anaphalis x tenuifolia
- Anaphalis x vulgaris
- Anaphalis x x hybrida
- Anaphalis x nemorosa
- Anaphalis x quinquefolia
- Anaphalis x tenuifolia
- Anaphalis x vulgaris
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
Pathway information

Summary:
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-plastidic plant membranes. Choline is also a precursor for the formation of glycine betaine, a major osmoprotectant in certain plants such as spinach, where this osmoprotectant is accumulated and confers tolerance to salinity, drought, and other environmental stresses. In addition, choline has been recognized as an essential nutrient for humans.

Pathway information:
The first step in choline biosynthesis is the direct decarboxylation of ornithine to ethanolamine, which is catalyzed by a specific decarboxylase unique to plants. Ethanolamine is widely recognized as the intermediate 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, phospho-bases, phosphatidyl-bases, or a mixture of the latter.

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 phosphoethanolamine and phosphatidylcholine is characterized by a high degree of interaction and funneling on the various levels of combining intermediates. Consequently, it has been assumed that the interconversions in these 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 in tobacco.

References:
**Compound: CDP-choline**

**Synonyms:**
citobolin, citoboline, citobolin, cytidine 5'-diphosphocholine, cytidine diphosphate choline

**Classification(s):**
a nucleic acid component -> a base derivative
a nucleic acid component -> a pyrimidine-related compound

**Empirical Formula:** $C_{14}H_{22}N_{4}O_{6}P_{2}$

**Molecular Weight:** 489.34 daltons

**Smiles:** C(OP)(O)(=O)OP(O)(=O)OC(C(N+)(C)(C)(C)C1)OC(C(0)(C)C(O)(C(=O)O)OCCN(C(=O)NC(=O))N)

**Unification Links:** CAS: 907-75-0

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

**In Pathway Reactions as a Reactant:**
- phospholipid biosynthesis
  - a, 1,2-diacetylglcerolphosphate $\rightarrow$ CDP-choline + a phosphatidylcholine + CMP

**In Pathway Reactions as a Product:**
- phospholipid biosynthesis
  - phosphorylcholine + CTP $\rightarrow$ CDP-choline + diphosphate
  - choline biosynthesis III
  - phosphorylcholine + CTP $\rightarrow$ CDP-choline + diphosphate

**Appears as Reactant**

**Appears as Product**
Enzyme information

**Arabidopsis** Enzyme: phosphatidyltransferase

---

**Reaction**

\[ \text{a}_1,2-	ext{diacylglycerol} + \text{CDP}-\text{choline} \rightarrow \text{phosphatidylcholine} + \text{CMP} \]

The reaction is shown as, that is, \( A + B \rightarrow C + D \) versus \( C + D \rightarrow A + B \). This is in accordance with the Enzyme Commission system.

Reversibility of this reaction is unspecified.

---

**Pathway(s)**

- Choline biosynthesis
- Phospholipid biosynthesis

---

**Inhibitors, Kinetic Parameters, etc.**

---

**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 aminoalcohol phosphotransferases 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 phosphatidylalcohols 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 Ca²⁺ and Cytidine monophosphate (CMP) than AtAAPT1. Both enzymes (AtAAPT1, AtAAPT2) were able to catalyze the reverse reaction supporting the proposal that diacylglycerol, involved as substrate in both PC and triglyceride biosynthesis (triacylglycerol biosynthesis), is in equilibrium with PC and maintains this equilibrium via the reversibility of the cholinephosphate transfer reaction.

Both polypeptides contain seven membrane spanning regions as shown by their hydrophathy files. This finding is in agreement with other AAPT's isolated so far and confirm their localization in membranes.

---

**References**

How does computationally predicted data enter the PMN?

**ANNOTATED GENOME**

- DNA sequences
- Predicted proteins
  - *Pv1234.56.a*
- Predicted functions
  - chorismate mutase

**PlantCyc / MetaCyc**

**Phaseolus vulgaris**

**PathoLogic**

- chorismate mutase
  - 5.4.99.5
- prephenate aminotransferase
- arogenate dehydratase

**Single species database**

- chorismate
- prephenate
- L-phenylalanine

**PhaseolusCyc**

+ validation
How can researchers use the PMN?

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

- Quick search bar

- Specific search menus
How can researchers use the PMN?
How can researchers use the PMN?

- Compare metabolism across species
How can researchers use the PMN?

- Examine OMICs data in a metabolic context
- Look at changes in transcript expression following 2 days of drought stress
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 thank you publicly!
TAIR would like help, too!
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 Annotating loci from article

<table>
<thead>
<tr>
<th>Molecular Function</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ion transmembrane transporter activity</td>
<td>Example:</td>
</tr>
<tr>
<td>protein kinase activity</td>
<td>Enzyme assays</td>
</tr>
<tr>
<td>phytochrome synthase activity</td>
<td>Choose a method or enter a new one.</td>
</tr>
</tbody>
</table>

Start typing, then choose from list or add a new line, then click outside the field to accept the new text.

histidine
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!

curator@arabidopsis.org  curator@plantcyc.org

www.arabidopsis.org  www.plantcyc.org
TAIR and PMN Acknowledgements

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

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

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

curator@arabidopsis.org  
www.arabidopsis.org

curator@plantcyc.org  
www.plantcyc.org
Out-takes

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

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
Providing Tools at TAIR

- Tech team members and curators
  - Develop new tools and modify existing tools

SeqViewer

Patmatch

---

**SeqViewer**

Choose a Sequence Database (click and drag to select)

- All public Arabidopsis sequences can be found within TAIR3 Proteins [protein]

START PATTERN SEARCH

PLEASE WAIT FOR EACH REQUEST TO COMPLETE FIRST

More Options:
- Maximum hits: 75000
- If DNA, Strand: both strands
- Mismatch: 8

Supported Pattern Syntax and Examples:

<table>
<thead>
<tr>
<th>Search Type</th>
<th>Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide Searches</td>
<td>Any</td>
</tr>
<tr>
<td></td>
<td>E or Q</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HITS</th>
<th>Sequence name</th>
<th># of hits</th>
<th>Hit pattern</th>
<th>Matching Positions</th>
<th>Hit sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AT1G00050.1</td>
<td>1</td>
<td>DREFCDA</td>
<td>342, 363</td>
<td>sequence</td>
</tr>
<tr>
<td>2</td>
<td>AT1G12220.1</td>
<td>1</td>
<td>SMEACFA</td>
<td>242, 248</td>
<td>sequence</td>
</tr>
<tr>
<td>3</td>
<td>AT1G18110.1</td>
<td>1</td>
<td>DREDTCQA</td>
<td>483, 485</td>
<td>sequence</td>
</tr>
<tr>
<td>4</td>
<td>AT1G20029.1</td>
<td>1</td>
<td>B3HSCQA</td>
<td>- 397</td>
<td>sequence</td>
</tr>
<tr>
<td>5</td>
<td>AT1G20602.2</td>
<td>1</td>
<td>B3HOCQA</td>
<td>209, 213</td>
<td>sequence</td>
</tr>
<tr>
<td>6</td>
<td>AT1G21705.1</td>
<td>1</td>
<td>DREDOCMA</td>
<td>209, 212</td>
<td>sequence</td>
</tr>
<tr>
<td>7</td>
<td>AT1G22556.1</td>
<td>1</td>
<td>DREDOCMA</td>
<td>209, 212</td>
<td>sequence</td>
</tr>
<tr>
<td>8</td>
<td>AT1G23850.1</td>
<td>1</td>
<td>SOTGCLA</td>
<td>206, 207</td>
<td>sequence</td>
</tr>
<tr>
<td>9</td>
<td>AT1G26320.1</td>
<td>1</td>
<td>SHERCLA</td>
<td>204, 206</td>
<td>sequence</td>
</tr>
</tbody>
</table>
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