PANTHER User Manual

For PANTHER 7.0

Date: September 19, 2011

Authors:

The PANTHER Team

Contents

Chapter 1

Welcome to PANTHER System

1.1 About this document

This manual provides a step-by-step instruction of how to use all the pages, functions and tools in the PANTHER Classification System website. The manual starts with a detailed description of different types of data pages (Chapter 2), followed by Chapter 3, which describes the basic functionalities of the system. Chapter 4 provides detailed description of our data analysis tools.

This manual does not serve as a scientific reference of the PANTHER System. It assumes that you are already familiar with the scientific background of the system. If not, we highly encourage you to read the papers listed in the next section to get familiar with our system.

1.2 How to cite PANTHER

Here is a list of PANTHER publications, based on the subject, that you can cite as references when you publish your results.

- **General PANTHER usage** [1]
  Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.*, 38(Database issue):D204-10
- **PANTHER pathway** [2]
  Mi H, Thomas P. (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol.*, 563:123-40.
- **PANTHER software and tools** [3]
  Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B. (2006) Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.*, 34(Web Server issue):W645-50.
- **PANTHER tree building algorithm (GIGA)** [4]
  Thomas PD. (2010) GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics.*, 11:312.

1.3 PANTHER help

This manual is not the only help you can get to learn how to use the PANTHER System. You can get additional help and support through the following:

- On most of our webpages, there are links to help tips that provide detailed information about the page, and how to use it. Simply click the  icon to get to the Tips page.
- You can always contact us through our feedback email at feedback@pantherdb.org.

1.4 Overview

OK, here we go. The PANTHER home page (Figure 1.1) provides an overview of the PANTHER functionality. The top menu bar (item 1-7 in Figure 1.1) contains links to individual subject main pages with links within that subject. The left panel (item 11-20 in Figure 1.1) contains links directly to the popular tools for quick access. The right panel contains PANTHER News and Publications (item 23 and 24, respectively, in Figure 1.1). The Keyword Search and Sequence Search (HMM scoring) (item 21 and 22, respectively, in Figure 1.1) on the home page allows you to retrieve quick results from our system (Section 3.1).

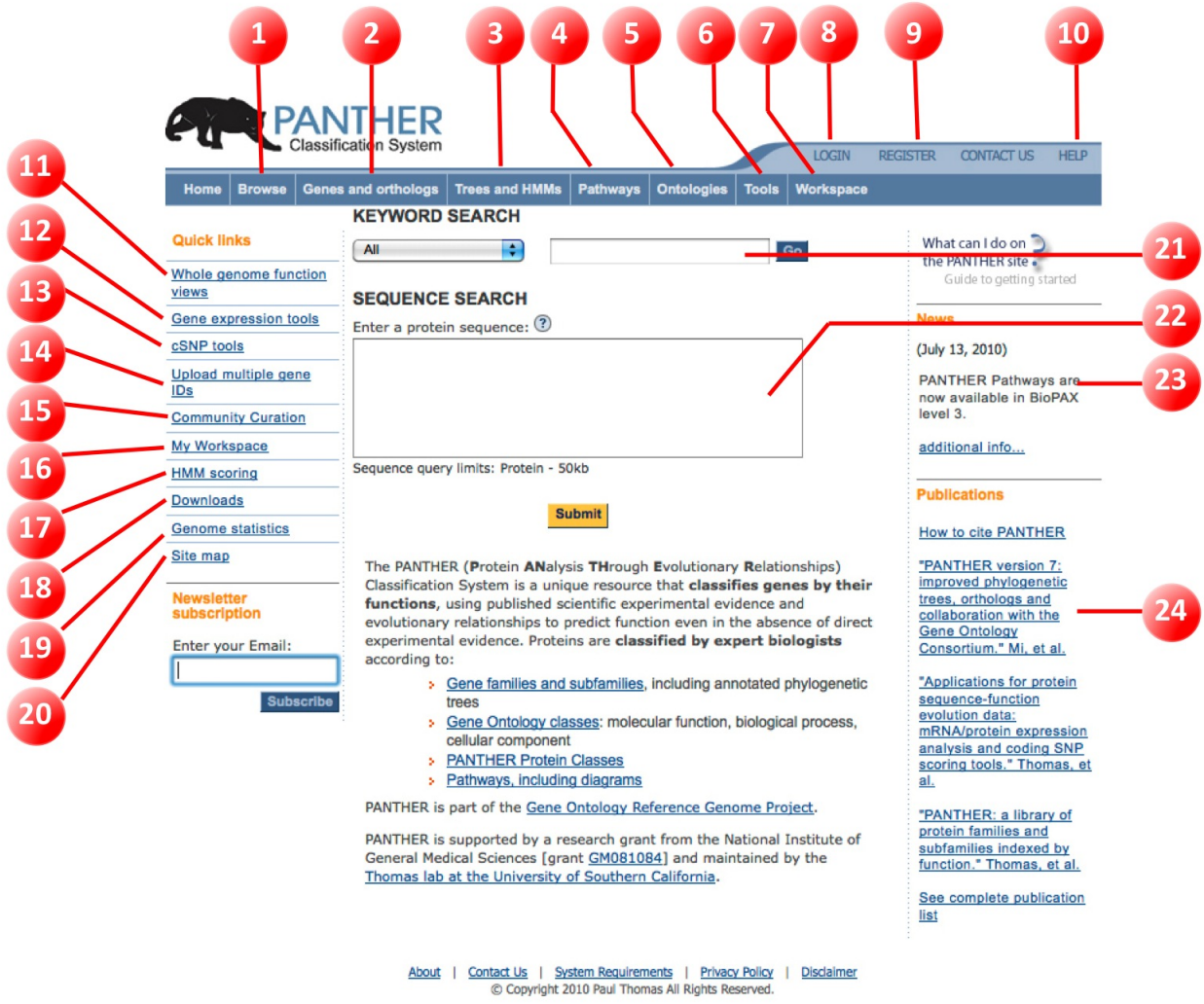


Figure 1.1: *The PANTHER home page*

Below is an overview of the PANTHER home page (the numbers correspond to those in Figure 1.1).

1. **Browse** – Links to the Prowler so that you can browse by functions, pathways and species (see Section 3.2).

2. **Genes and Orthologs** (Section 2.1.1) – Links to the Gene, Genome and Orthologs page

where you can perform simple Keyword Search (Section 3.1.1), find Whole Genome Pie Charts (Section 2.4.1), and upload batch IDs (Section 3.3).

3. **Trees and HMMs** – Links to the page that provides links to functions related to trees and HMMs, such as HMM scoring (Section 3.4), download PANTHER HMM scoring program (Section 3.5), etc.
4. **Pathways** – Links to the page that contains pathway related knowledge (Section 2.1.3).
5. **Ontologies** – Links to the page that provides information about the PANTHER GO Slim and PANTHER Protein Classes , and links to download them (Section 2.1.4).
6. **Tools** – Links to all the PANTHER tools page (Section 2.1.5. See details in Chapter 4).
7. **Workspace** – Links to the workspace available to the registered users (Section 2.1.6). Registration is free.
8. **Login** – Click here to login to the website and use the workspace. PANTHER website does not require a login, but the workspace is only available to registered users.
9. **Registration** – It is not required to register to use the PANTHER website. Registration allows you to create a workspace that you can store your analysis data on our site, and it is free.
10. **Help** – You can find more help tips at this link.
11. **Whole genome function views** – Links to the whole genome pie chart (Section 2.4.1).
12. **Gene expression tools** – Links directly to the Gene Expression Analysis Tools page (Section 4.1).
13. **cSNP tools** – Links directly to the cSNP Scoring tool page (Section 4.2).
14. **Upload multiple gene IDs** – Links directly to the Batch ID Upload page (Section 3.3).
15. **Community Curation** – This page is currently provides the link that allows you to download PAINT (Phylogenetic tree Annotation INference Tool) application.
16. **My Workspace** – Links to the workspace available to the registered users (Section 2.1.6). Registration is free.
17. **HMM scoring** – Links to the PANTHER HMM Scoring page (Section 3.4).
18. **Downloads** – Links to the PANTHER FTP site (Section 3.5).
19. **Genome statistics** – Links to page with statistics of 48 genomes in the PANTHER library.
20. **Site map** – Links to the site map.
21. **Keyword search** – Perform simple keywords search for genes, proteins, families, ontology and pathway terms. (Section 3.1.1)
22. **Sequence search** – Enter a protein sequence to score against PANTHER HMM library (Section 3.4).
23. **Newsletters** – The most recent newsletter is displayed. All archived newsletters can be retrieved by clicking the "additional info" link.

24. **Publications** – This section lists some of the PANTHER publications and information about citing PANTHER.

Chapter 2

PANTHER At A Glance

PANTHER data are stored in a database in the backend. PANTHER website presents the data to the users through a number of text pages and graph/diagram pages.

The text pages include:

- Subject main page
- List pages
- Information detail pages

The graph and diagram pages include:

- Pie charts
- Pathway diagrams
- Tree Viewer

2.1 Subject main page

2.1.1 Genes and orthologs

PANTHER now contains the complete sets of protein coding genes for 48 different organisms, obtained from various sources, including Model Organism Database, UniProt, RefSeq and ENSEMBL. This page serves as a portal for you to access information related to genes, genomes and orthologs (Figure 2.1). There are 4 links on this page that allows you to

- perform *simple keyword search* (Section 3.1.2);
- find *Whole Genome Pie Charts* (Section 2.4.1) from all 48 organisms;
- perform *batch ID search* (see Section 3.3);
- view statistics, such as, for each organism, number of total genes, number of genes in PANTHER families, number of genes with GO classifications, etc.

Figure 2.1: *Gene and orthologs home page.*

2.1.2 Trees and HMMs

PANTHER version 7.0 contains 6594 protein families, each with a phylogenetic tree relating modern-day genes in 48 organisms. Phylogenetic trees also infer the evolutionary events that occurred, such as speciation and gene duplication events. Expert biologists have divided each family into subfamilies, which are generally orthologous groups but may also contain recently duplicated paralogs. Each family and subfamily is also represented as a hidden Markov model (HMM), which can be used to classify new sequences to an existing subfamily.

This page serves as a portal for users to access information related to the trees and HMMs (Figure 2.2). There are 4 links on this page that allows you to

- search for PANTHER families, trees and HMMs. This basically takes you to the keyword search page (Section 3.1.2);
- browse families and subfamilies by classifications, which takes you to the prowler page (Section 3.2);
- score a new sequence against PANTHER HMMs, which takes you to the HMM score page (Section 3.4);
- download the PANTHER scoring tool to score multiple sequences against PANTHER HMMs (Section 3.5)

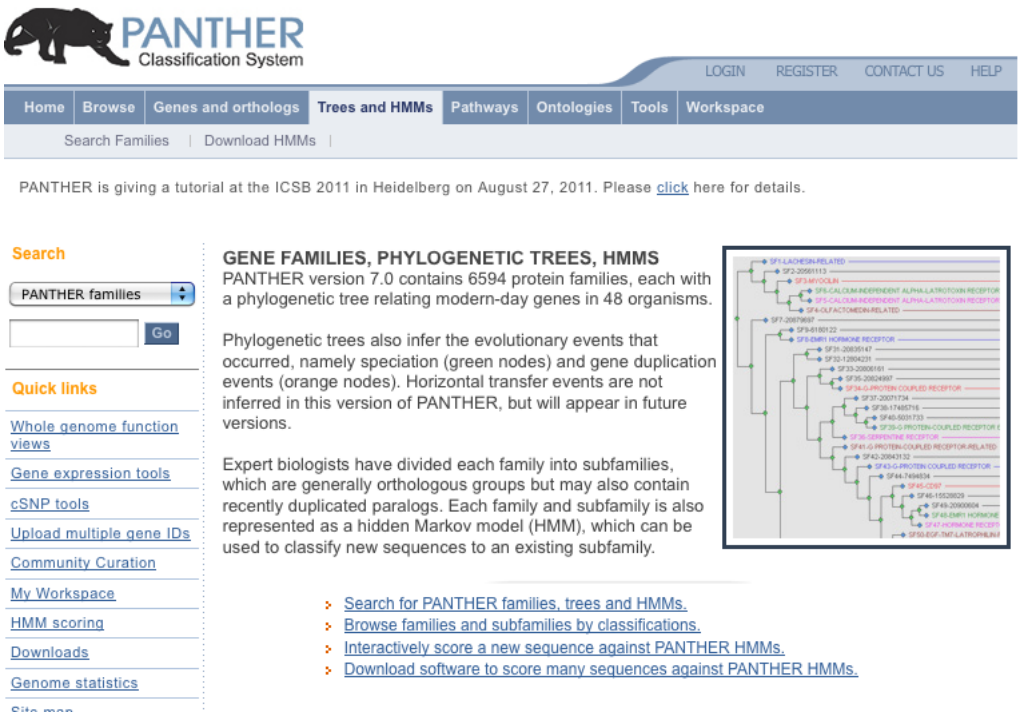


Figure 2.2: *Trees and HMMs home page.*

2.1.3 Pathways

The current version of PANTHER Pathway consists of 165 pathways, each with subfamilies and protein sequences mapped to individual pathway molecule class (aka component). All pathways are stored in both SBML and BioPAX format. Pathway diagrams are interactive and include tools for visualizing gene expression data in the context of the diagrams.

This page provides a port for users to access information related to PANTHER pathways (Figure 2.3), and provides the following links.

- Community Curation - this link doesn't work since the site is under modification at the moment.
- Browse pathways - this takes you to the prowler page (section 3.2).
- Search pathways - this takes you to the advanced keyword search page (section 3.1.2).
- Pathway resources - links to some useful pathway databases, tools and resources.
- List of pathways - lists all 165 PANTHER pathways, and links to view the interactive pathway diagram, download SBML and BioPAX files.
- Download pathway data - it links to the ftp site for all pathway data download (Section 3.5).

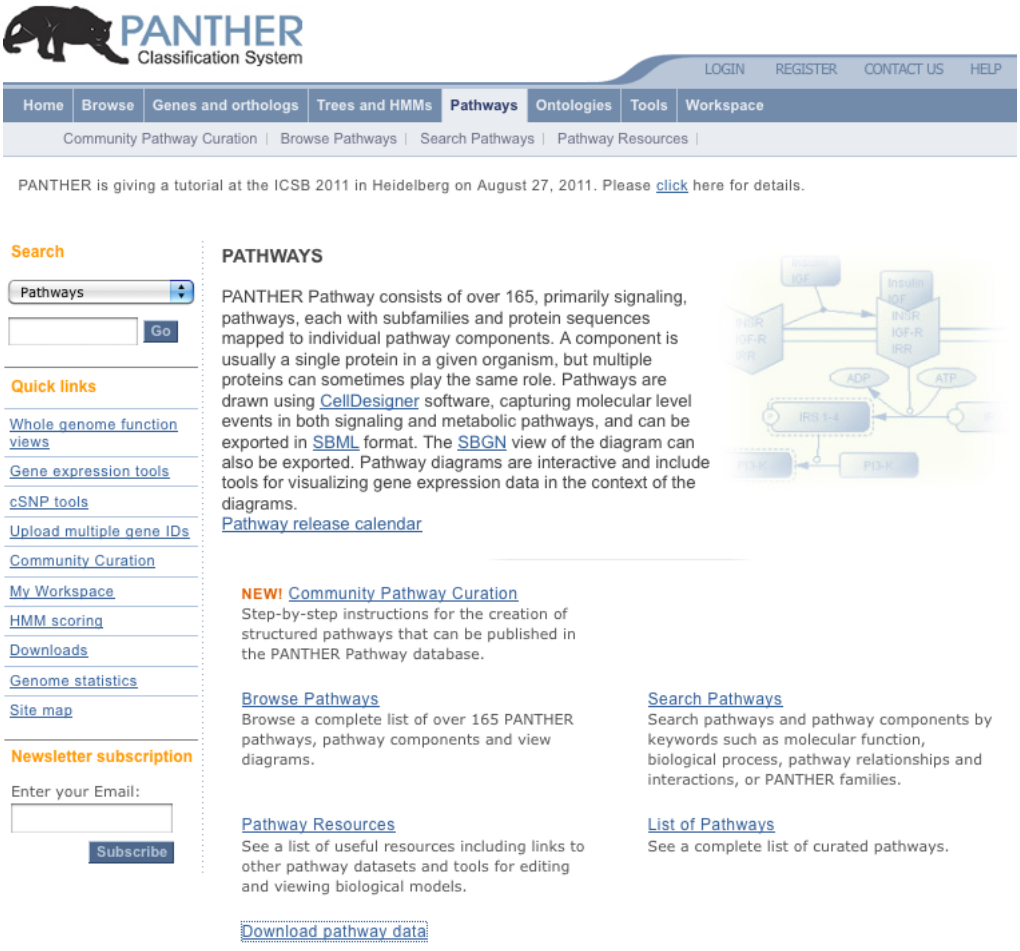


Figure 2.3: *PANTHER Pathway home page.*

2.1.4 Ontologies

This page provides a brief description of the ontolgoies used in PANTHER. It also provides links to browse in prowler (Section 3.2) and download the ontologies.

2.1.5 Tools

PANTHER System is not only a database, but also a data analysis platform that allows users to perform more advanced tasks, and analyze research results using this integrated system. This page provides links to the following tools:

- Gene expression data analysis tools (see detail in Section 4.1)
- Evolutionary analysis of coding SNPs (see detail in Section 4.2)
- PANTHER scoring (see detail in Section 3.4)
- Downloads (see Section 3.5)

2.1.6 Workspace

The Workspace is a unique feature in PANTHER that allows users to store the gene lists that they generate for future analysis. Although users do not have to register to use the PANTHER system, registration is required in order to user the workspace. Registration is free. This webpage allows registered users to login, or retrieve password, and for non-registered users to register.

A more detailed description of using *Workspace* can be found in Section 3.6.

2.2 List pages

2.2.1 Gene list page

Thge gene list page can be generated from the prowler, keyword search, pathway diagram, and other list and detail pages.
The gene list page contains the following information (Figure 2.4).

- **Gene ID** – This is the identifier for genes in the PANTHER library. The format is as follows: organism|gene database source=gene id|protein database source=protein id. For example, HUMAN|ENSEMBL=ENSG00000111262|UniProtKB=Q09470 is a human sequence, the gene sequence is from ENSEMBL with id ENSG00000111262, and the protein sequence is from UniProt with id Q09470. Clicking on the ID will lead to the Gene Detail page (section 2.3.1).
- **Gene Name/Gene Symbol** – The Entrez gene definition and gene symbol. Clicking on the gene symbol will lead to the Gene Detail page (section 2.3.1). Please note that some of the genes do not have the information in this column.
- **Protein ID** – The public protein identifier used in the GeneID above. Clicking on the ID will lead to the Gene Detail page (section 2.3.1).
- **PANTHER Family/Subfamily** – The family or subfamily name of the PANTHER model where the sequence is in. Clicking on the name will lead to the PANTHER family or subfamily detail page (section 2.3.2).
- **GO Molecular Function, Biological Process, Cellular Component**: These are Gene Ontology terms from PANTHER GO Slim describing the function of the gene product. Clicking on the ontology term will lead to the term detail page (section 2.3.3).

- **PANTHER Protein Class** – This is a PANTHER Index terms describing protein classes. Clicking on the term will lead to the term detail page (section 2.3.3).
- **Pathway** – Pathway and pathway component with which at least one training sequence within the PANTHER best hit family or subfamily were associated directly by manual curation. Clicking on the terms will lead to the pathway diagram page (section 2.4.2). If the pathway component is clicked, the corresponding component will be highlighted in yellow when the diagram is open.
- **Species** – The organism of the gene.

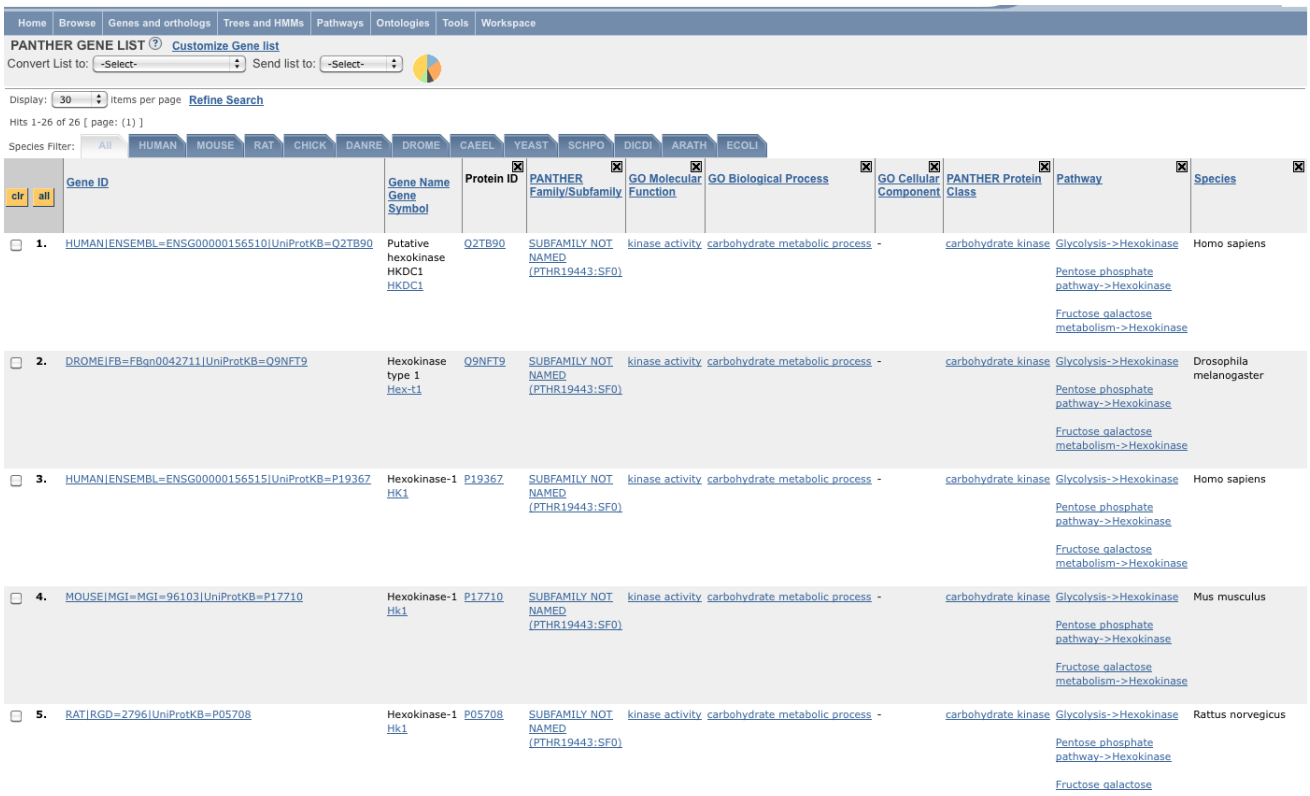


Figure 2.4: *Gene list page.*

Tip: Manipulate your list

Here is a description of what you can do to your list in any list pages in this section (2.2).

- Sort the list – You can always sort the list by clicking on any of the underlined column names. A yellow triangle appears in front of the column name that you choose to sort. The orientation of the triangle indicates the sort is ascending or descending.
- Customize columns: You can click on the "x" button next to the column names to collapse the column.
- Converting a list to another list type. Select the genes you want to convert by clicking the checkboxes. The default is for all genes in the list.
  - Click on the pull-down menu after *Convert list to*. The current list type is shown in the box.

  - Select the new list type from the pull-down menu (Figure 2.5.)
  - Each primary ID (first column of the list) is used to return the selected data type. Note that the mapping between different types is not necessarily one-to-one (e.g. a subfamily can map to more than one associated transcript).

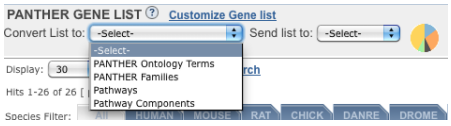


Figure 2.5: *Options to convert a list from one type to another.*

- Saving a list to your workspace. Select the genes you want to save by clicking the check-boxes. The default is for all genes in the list. You can select one of the followings from the pull-down menu as the destination (Figure 2.6).
  - Workspace – You need to register to save data to the workspace. The registration is free. When you make this selection, a pop-up window will ask you to name the list and add any comments. The name and comments can be edited at any time in the future from the Workspace page. Once the gene list is now at the site, it can be returned to at any time. Only the IDs are stored, and they are mapped to the internal PANTHER gene ids, so when you access a list in the future, all information will be updated and current.
  - Exporting a list to a file – The list will be exported as a tab-delimited file. You can now import the file into Excel or perform any post-processing you wish.
  - View the list as text on the website.
- Use the pie chart view by clicking the colorful pie chart icon. See Pie Charts section (2.4.1) for details.

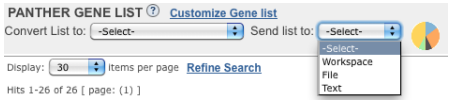


Figure 2.6: *Export list.*

2.2.2 Family/subfamily list page

The family/subfamily list page can be generated from prowler, keyword search, pathway diagram, and other list and detail pages.
This page provides a list of families and/or subfamilies with the following related information (Figure 2.7).

- **Family ID** – In this column, it usually lists either family or subfamily IDs. Click the ID will take you to the *Family Detail page* (see section 2.3.2).
- **Family Name** – This column lists the names of families or subfamilies. Click the name will take you to the *Family Detail page* (see section 2.3.2).

- **Genes** – This column lists the number of Genes in the family or subfamily. These are also the sequences used to build the model. Clicking on the number will take you to the *Gene List page* (see section 2.2.1).
- **GO Molecular Function, Biological Process, Cellular Component** – These are Gene Ontology terms from PANTHER GO Slim describing the function of the gene product. You can click on each term to view more information in the *Ontology Term Detail page* (see section 2.3.3).
- **PANTHER Protein Class** – This is a PANTHER Index terms describing protein classes. You can click on each term to view more information in the *Ontology Term Detail page* (see section 2.3.3).
- **Pathway** – This column lists the pathway and pathway component in which the families or subfamilies are involved. Clicking on the terms will lead to the pathway diagram page (section 2.4.2). If the pathway component is clicked, the corresponding component will be highlighted in yellow when the diagram is open.
- **Interpro** – If the PANTHER family or subfamily is integrated into Interpro, the Interpro accession is listed here. Clicking on the Interpro accession will lead you to the Interpro website.

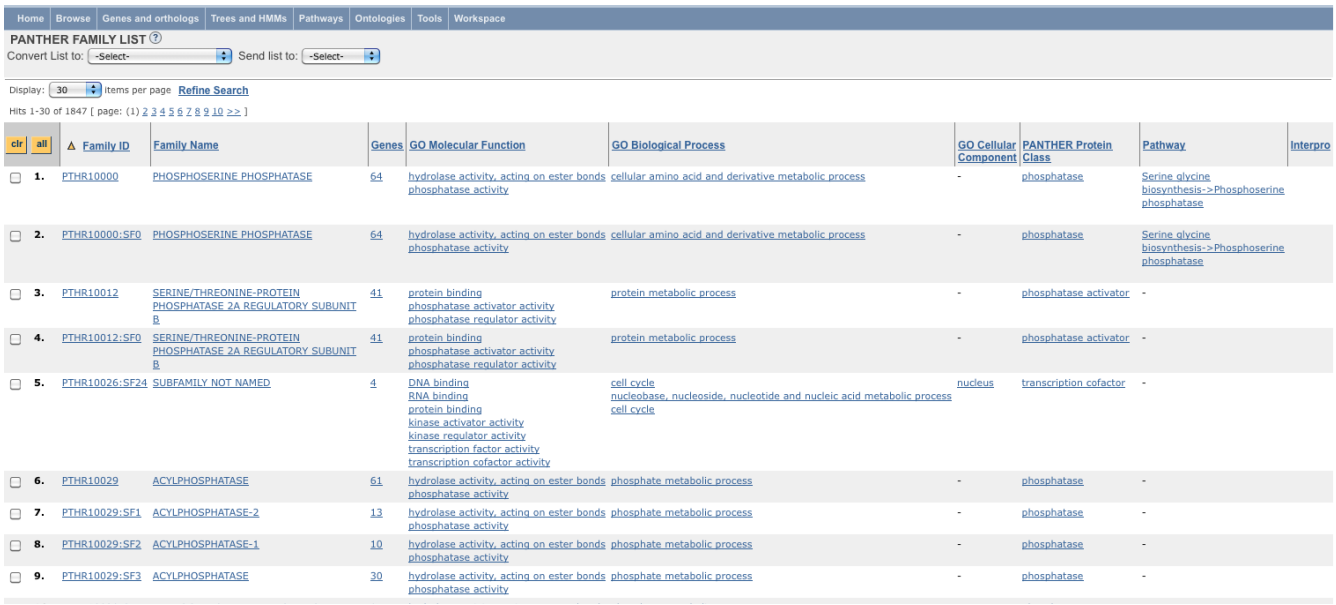


Figure 2.7: *Family/subfamily list page.*

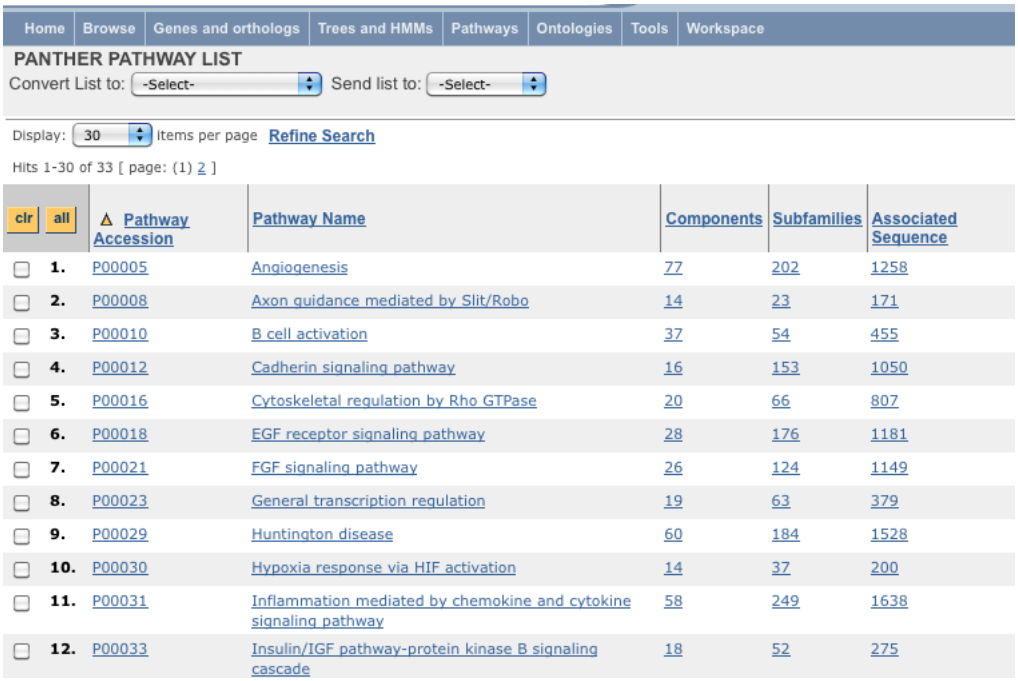
You can sort the list, convert the list to a different ID type and save the list as described in Gene List section (2.2.1).

2.2.3 Pathway list page

This page provides a list of pathways with the following information (Figure 2.8).

- Pathway Accession - This column lists the unique accession of each pathway. Clicking the accession will take you to the *Pathway Diagram* (section 2.4.2).
- Pathway Name - This column lists the pathway names in the list. Clicking the accession will take you to the *Pathway Diagram* (section 2.4.2).

- Components - This column lists the number of components (or molecule classes) in each pathway. Clicking the number will take you to the pathway component list page.
- Subfamilies - This column provides the number of subfamilies of all genes that are associated to each pathway. Clicking the number will take you to the *Subfamily list page* (section 2.2.2).
- Associated Sequences - This column lists the numbers of genes that are associated to the pathway. Clicking the number will take you to the *Gene list page* (section 2.2.1).



	Pathway Accession	Pathway Name	Components	Subfamilies	Associated Sequence
1.	P00005	Angiogenesis	77	202	1258
2.	P00008	Axon guidance mediated by Slit/Robo	14	23	171
3.	P00010	B cell activation	37	54	455
4.	P00012	Cadherin signaling pathway	16	153	1050
5.	P00016	Cytoskeletal regulation by Rho GTPase	20	66	807
6.	P00018	EGF receptor signaling pathway	28	176	1181
7.	P00021	FGF signaling pathway	26	124	1149
8.	P00023	General transcription regulation	19	63	379
9.	P00029	Huntington disease	60	184	1528
10.	P00030	Hypoxia response via HIF activation	14	37	200
11.	P00031	Inflammation mediated by chemokine and cytokine signaling pathway	58	249	1638
12.	P00033	Insulin/IGF pathway-protein kinase B signaling cascade	18	52	275

Figure 2.8: *Pathway list page.*

You can sort the list, convert the list to a different ID type and save the list as described in Gene List section (2.2.1).

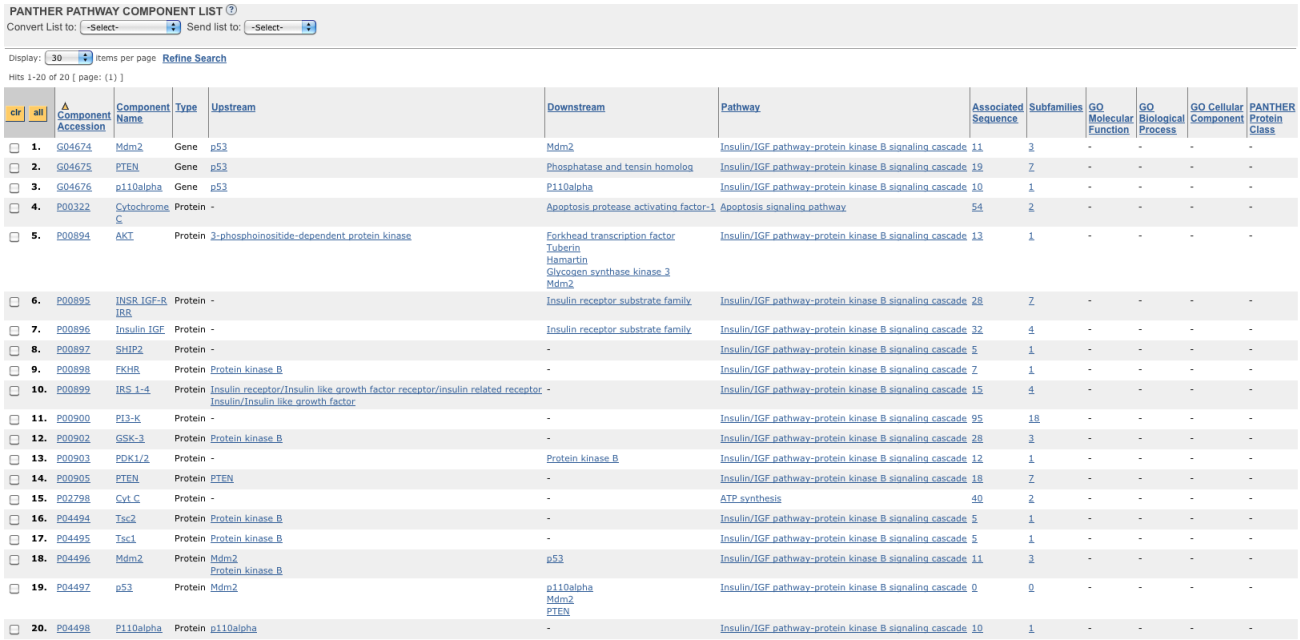
2.2.4 Pathway component list page

A pathway component (also called molecule class) represents a group of homologous proteins across various organisms that participate in the same specific biochemical reactions within the pathway. A pathway component is always associated with training sequences within the PANTHER libraries, thus, it is linked directly or indirectly to the PANTHER family/subfamily statistical models.

This page lists pathway components with the following information (Figure 2.9):

- Component Accession - The accession assigned to each pathway component. It is usually a 5 digit number proceeded by a letter to indicate the type of the component in the pathway (G for DNA or RNA, P for protein). Clicking it will lead to the *Pathway molecule class detail page* (section 2.3.5).
- Component Name - The display name of the component. Clicking it lead to the *Pathway molecule class detail page* (section 2.3.5).
- Type - Indicates the type of the component in the pathway, eg., gene, protein.

- Upstream and Downstream - These two columns list the pathway component(s) immediately upstream or downstream of the component. Upstream is defined as the component that controls or modulates the reaction that the current component is involved in. Downstream component(s) are those involved in the reaction that is controlled or modulated by the current component. Clicking it lead to the *Pathway molecule class detail page* (section 2.3.5).
- Pathway - The pathway that the component belongs to. Clicking it lead to the *Pathway molecule diagram page* (section 2.4.2).
- Associated Sequences - This column lists the numbers of genes that are associated to this pathway component. Clicking the number will take you to the *Gene list page* (section 2.2.1).
- Subfamilies - This column provides the number of subfamilies of all genes that are associated to the pathway component. Clicking the number will take you to the *Subfamily list page* (section 2.2.2).
- GO classifications and PANTHER protein class - These 4 columns lists all the categories classified to the genes that are associated to the pathway component. Clicking on an ontology terms will lead to the *Ontology Term Details page* (section 2.3.3)..



PANTHER PATHWAY COMPONENT LIST

Convert List to: -Select- Send list to: -Select-

Display: 30 items per page Refine Search

Hits 1-20 of 20 [page: (1)]

clr all		Component Accession	Component Name	Type	Upstream	Downstream	Pathway	Associated Sequence	Subfamilies	GO Molecular Function	GO Biological Process	GO Cellular Component	PANTHER Protein Class
☐	1.	G04674	Mdm2	Gene	p53	Mdm2	Insulin/IGF pathway-protein kinase B signaling cascade	11	3	-	-	-	-
☐	2.	G04675	PTEN	Gene	p53	Phosphatase and tensin homolog	Insulin/IGF pathway-protein kinase B signaling cascade	19	7	-	-	-	-
☐	3.	G04676	p110alpha	Gene	p53	P110alpha	Insulin/IGF pathway-protein kinase B signaling cascade	10	1	-	-	-	-
☐	4.	P00322	Cytochrome C	Protein	-	Apoptosis protease activating factor-1	Apoptosis signaling pathway	54	2	-	-	-	-
☐	5.	P00894	AKT	Protein	3-phosphoinositide-dependent protein kinase	Forkhead transcription factor Tuberin Hamartin Glycogen synthase kinase 3 Mdm2	Insulin/IGF pathway-protein kinase B signaling cascade	13	1	-	-	-	-
☐	6.	P00895	INSR IGF-R IRR	Protein	-	Insulin receptor substrate family	Insulin/IGF pathway-protein kinase B signaling cascade	28	7	-	-	-	-
☐	7.	P00896	Insulin IGF	Protein	-	Insulin receptor substrate family	Insulin/IGF pathway-protein kinase B signaling cascade	32	4	-	-	-	-
☐	8.	P00897	SHIP2	Protein	-	-	Insulin/IGF pathway-protein kinase B signaling cascade	5	1	-	-	-	-
☐	9.	P00898	FKHR	Protein	Protein kinase B	-	Insulin/IGF pathway-protein kinase B signaling cascade	7	1	-	-	-	-
☐	10.	P00899	IRS 1-4	Protein	Insulin receptor/Insulin like growth factor receptor/insulin related receptor Insulin/Insulin like growth factor	-	Insulin/IGF pathway-protein kinase B signaling cascade	15	4	-	-	-	-
☐	11.	P00900	PI3-K	Protein	-	-	Insulin/IGF pathway-protein kinase B signaling cascade	95	18	-	-	-	-
☐	12.	P00902	GSK-3	Protein	Protein kinase B	-	Insulin/IGF pathway-protein kinase B signaling cascade	28	3	-	-	-	-
☐	13.	P00903	PDK1/2	Protein	-	Protein kinase B	Insulin/IGF pathway-protein kinase B signaling cascade	12	1	-	-	-	-
☐	14.	P00905	PTEN	Protein	PTEN	-	Insulin/IGF pathway-protein kinase B signaling cascade	18	7	-	-	-	-
☐	15.	P02798	Cyt C	Protein	-	-	ATP synthesis	40	2	-	-	-	-
☐	16.	P04494	Tsc2	Protein	Protein kinase B	-	Insulin/IGF pathway-protein kinase B signaling cascade	5	1	-	-	-	-
☐	17.	P04495	Tsc1	Protein	Protein kinase B	-	Insulin/IGF pathway-protein kinase B signaling cascade	5	1	-	-	-	-
☐	18.	P04496	Mdm2	Protein	Mdm2 Protein kinase B	p53	Insulin/IGF pathway-protein kinase B signaling cascade	11	3	-	-	-	-
☐	19.	P04497	p53	Protein	Mdm2	p110alpha Mdm2 PTEN	Insulin/IGF pathway-protein kinase B signaling cascade	0	0	-	-	-	-
☐	20.	P04498	P110alpha	Protein	p110alpha	-	Insulin/IGF pathway-protein kinase B signaling cascade	10	1	-	-	-	-

Figure 2.9: *Pathway component list page.*

You can sort the list, convert the list to a different ID type and save the list as described in Gene List section (2.2.1).

2.2.5 Ontology term list page

This page provides a list of PANTHER ontology terms, including PANTHER GO Slim and PANTHER Protein Class, with the following related information.

- Category ID - A GO id (for PANTHER GO Slim) or a PANTHER Protein Class id, which is a 5 digit number preceded by letters PC. Clicking on the ID will lead to the *Ontology Term Detail page* (section 2.3.3).

- Name - Category names.
- Parent: The names of all parent categories. A parent category refers to a more general category of the current one, e.g., Receptor is a parent category of Protein kinase receptor. For GO terms, the parent is not the parent category in the full Gene Ontology, but the parent GO terms in the PANTHER GO Slim. Clicking on the category name will lead to the Category Detail page.
- Child: The names of all child categories. The names of all child categories. A child category refers to a more specific category of the current one, e.g., Serine/threonine protein kinase receptor is a child category of Protein kinase receptor. For GO terms, the child is not the parent category in the full Gene Ontology, but the child GO terms in the PANTHER GO Slim. Clicking on the category name will lead to the Category Detail page.
- Families and Subfamilies: These 2 columns list the number of families and subfamilies that are classified by the category.

The yellow triangle in front of the column name indicates that the table is currently sorted by the column.

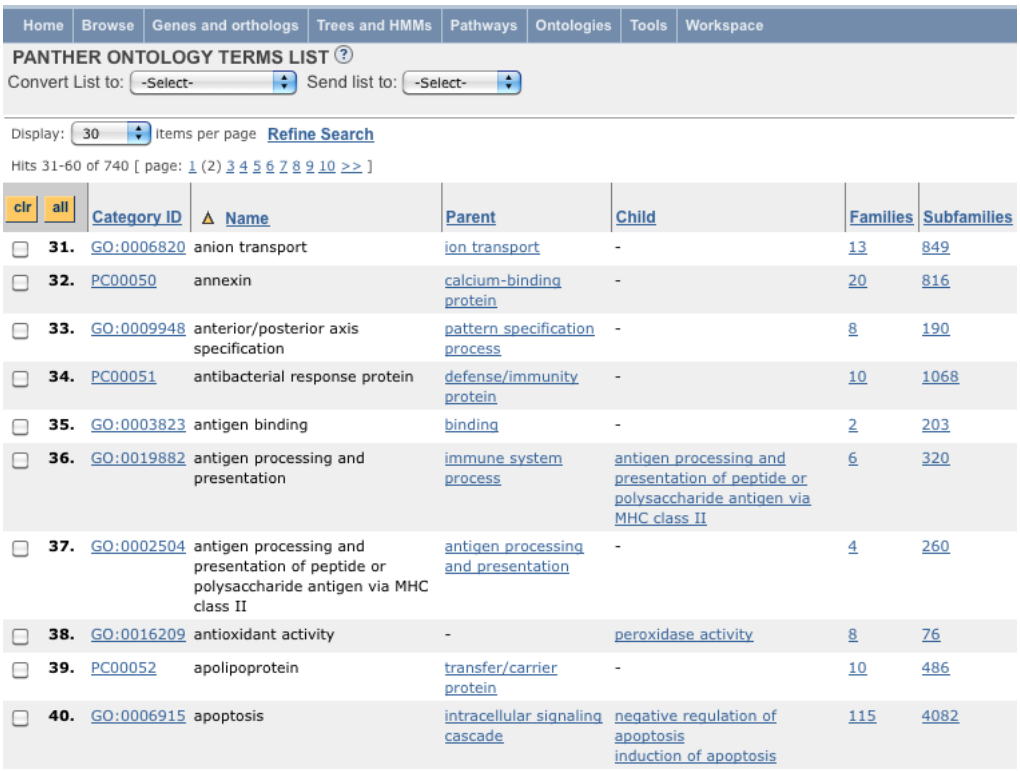


Figure 2.10: *Ontology term list page.*

You can sort the list, convert the list to a different ID type and save the list as described in Gene List section (2.2.1).

2.3 Information detail pages

2.3.1 Gene detail page

This page provides detailed information about a gene. The page contains 3 general sections.

1. **General information about the gene (Figure 2.11)**

   Gene ID From source database, either a model organism database (e.g. FlyBase), Ensembl, or Entrez Gene.

   Protein ID A UniProt (preferentially SwissProt) ID whenever possible, this is the "representative" protein sequence for the gene.

   Gene Name From source database, together with the Gene ID.

   Gene Symbol(s) Entrez gene symbol, if available.

   Organism The organism of the gene.

   Alternative IDs Identifiers in other databases for this gene or a protein product of the gene. This is mapped based on the UniProt ID Mapping tool.

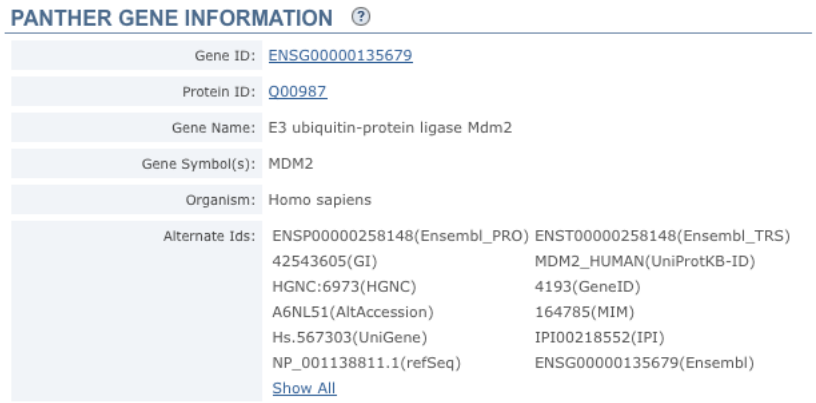


Figure 2.11: *Gene information section of the gene detail page.*

2. **PANTHER classification of the gene (Figure 2.12)**

   PANTHER family The family to which the gene belongs; there is a phylogenetic tree for this family. This link will bring up the PANTHER family page.

   PANTHER subfamily The subfamily to which the gene belongs; the subfamily is annotated on the phylogenetic tree. This link will bring up the PANTHER subfamily page.

   GO molecular function, biological process, cellular component These are Gene Ontology terms describing the function of the gene product.

   PANTHER Protein Class This is a PANTHER Index terms describing protein classes.

   Pathway This links to a diagram of the pathway(s) a gene product participates in.

Figure 2.12: *Gene classification section of the gene detail page.*

3. **Orthologs (Figure 2.13)**

   This section lists all the orthologs of this gene.

   - "LDO'' means least diverged ortholog, while other, more diverged orthologs (if any) are marked "O". Paralogs are marked "P".
   - For orthologs, the species column gives the modern-day species in which the ortholog is found. For paralogs, the species column gives the two speciation events between which the duplication occurred that generated the paralogous genes. "ND" means "not determined". Thus different paralogs can be distinguished by how long ago the relevant duplications occurred.
   - Orthologs are genes that can be traced to the same gene in the genome of their most recent common ancestor species. Paralogs are genes that are traced to related, but distinct, genes in the genome of their most recent common ancestor species.
   - The LDO can be loosely interpreted as the most nearly "equivalent" gene in another organism. If there were no gene duplication events following the speciation event that separated two species, there is only a single ortholog in each genome and they are therefore marked as "LDO." If there was a gene duplication event following speciation, there may be more than one ortholog and the least diverged ortholog pair is marked "LDO," while other ortholog pairs are marked "O."

ORTHOLOGS

ID	Organism	Type
HUMAN|ENSEMBL=ENSG00000198625|UniProtKB=O15151	Chordata-Osteichthyes	P
PANTR|ENSEMBL=ENSPTRG00000005203|ENSEMBL=ENSPTRP00000008832	Pan troglodytes	LDO
MACMU|ENSEMBL=ENSMMUG00000014193|ENSEMBL=ENSMMUP00000018653	Macaca mulatta	LDO
MOUSE|MGI=MGI=96952|UniProtKB=P23804	Mus musculus	LDO
RAT|RGD=1305332|NCBI=XP_235169	Rattus norvegicus	LDO
BOVIN|ENSEMBL=ENSBTAG00000026775|ENSEMBL=ENSBTAP00000013766	Bos taurus	O
BOVIN|ENSEMBL=ENSBTAG00000010422|UniProtKB=A5PJW5	Bos taurus	LDO
CANFA|ENSEMBL=ENSCAFG00000000418|UniProtKB=P56950	Canis familiaris	LDO
MONDO|ENSEMBL=ENSMODG00000007219|ENSEMBL=ENSMODP00000008952	Monodelphis domestica	LDO
ORNAN|ENSEMBL=ENSOANG00000005942|ENSEMBL=ENSOANP00000009460	Ornithorhynchus anatinus	LDO
CHICK|ENTREZ=395609|NCBI=XP_416084	Gallus gallus	LDO
XENTR|ENSEMBL=ENSXETG00000001434|UniProtKB=Q6P3Q9	Xenopus tropicalis	LDO
FUGRU|ENSEMBL=ENSTRUG00000008338|ENSEMBL=ENSTRUP00000020879	Fugu rubripes	LDO
DANRE|ZFIN=ZDB-GENE-990415-153|UniProtKB=Q561Z0	Danio rerio	LDO
CIOIN|ENSEMBL=ENSCING00000001188|ENSEMBL=ENSCINP00000002267	Ciona intestinalis	LDO
ORYSJ|ENTREZ=4343310|UniProtKB=Q0D693	Oryza sativa	LDO

Figure 2.13: *Ortholog table of the gene detail page.*

2.3.2 Family/subfamily detail page

This page provides details of a PANTHER family or subfamily model. It contains 2 sections.

1. Family/subfamily information

   The pages for family and subfamily information are slightly different. The family detail page contains the following information (Figure 2.14):

   - Family: The name of the family model. Family ID is in parentheses (e.g., PTHR10000).
   - Subfamilies: The number of subfamilies under the family. Clicking on the number will lead to the Subfamily List page (section 2.2.2).
   - PANTHER Links: Two links will lead to the Tree and MSA pages (section 2.4.3).
   - GO molecular function, biological process, cellular component: These are Gene Ontology terms from PANTHER GO Slim describing the function of the gene product.
   - PANTHER Protein Class: This is a PANTHER Index term describing protein classes.
   - PANTHER Pathway: This links to a diagram of the pathway(s) in which a gene product participates.
   - Genes: The number of genes that were used as training sequences to build the family model. Clicking on the number will lead to the Gene List Page (section 2.2.1).
   - HMM Length: The number of position in the hidden-Markov model.
   - Downloads: You can download the family HMM built with either HMMER or SAM method.

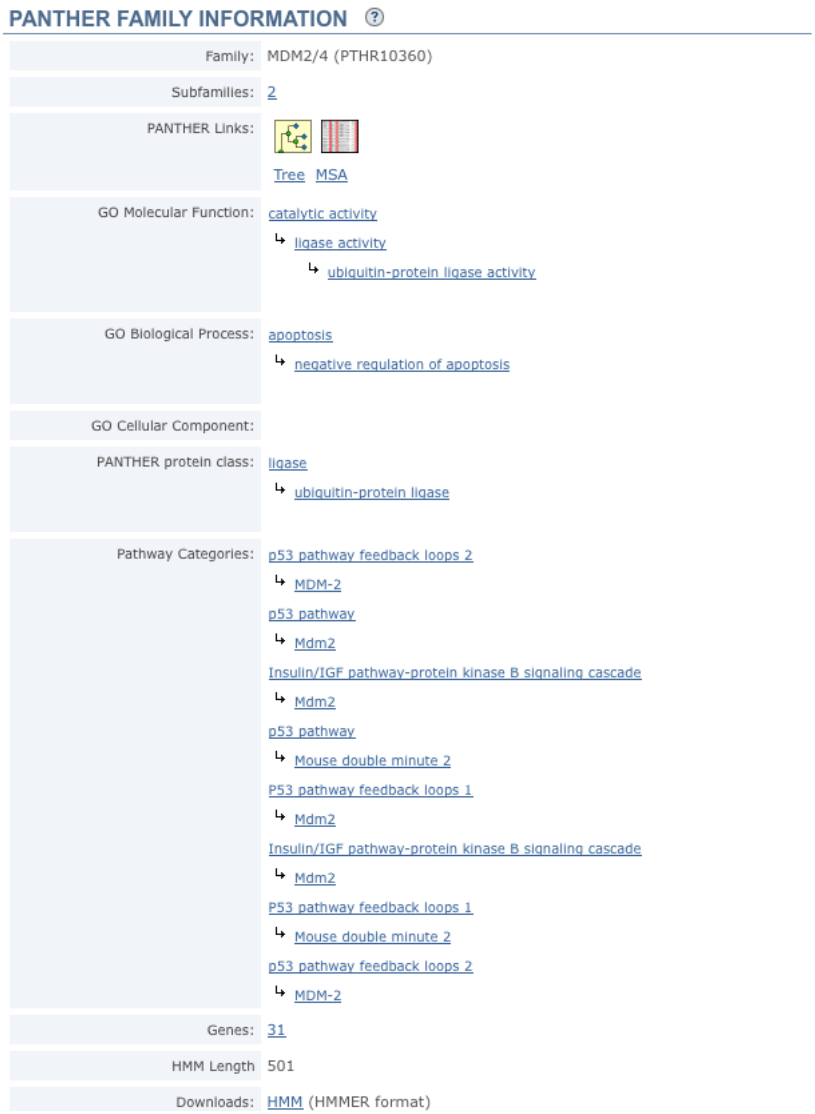


Figure 2.14: *Family information section of the family detail page.*

The subfamily detail page contains similar information as the family detail page except for the first two items (Figure 2.15):

- Subfamily: The name of the subfamily and the subfamily ID (eg., PTHR10000:SF1)
- Family: The family name that the subfamily belongs to. Clicking the name will lead you to the family detail page

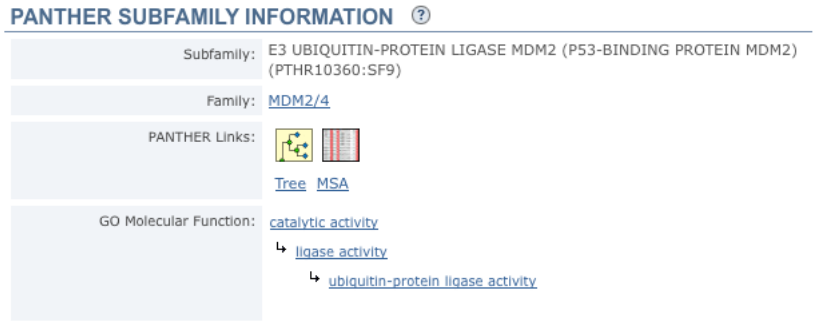


Figure 2.15: *Subfamily information section of the family detail page.*

2. Genes Assigned To This Family: A table listing the number of genes from various species that hit this model. Clicking on the number will lead to the Gene List page (section 2.2.1).

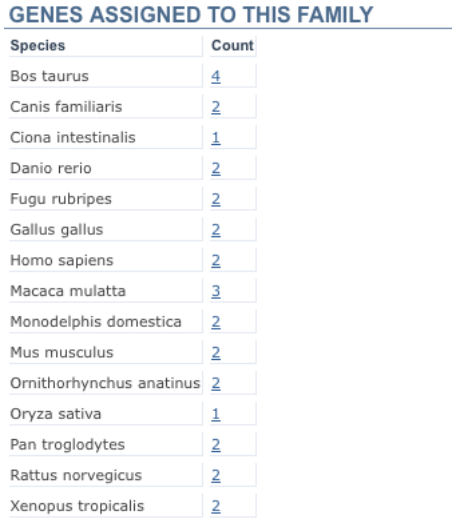


Figure 2.16: *List of genes assigned to the family in the family detail page.*

2.3.3 Ontology term detail page

This page provides the following information for a particular PANTHER GO Slim term or a PANTHER protein class category (Figure 2.17):

Name The name of the category.

Definition brief description of the category.

Class ID A GO id (for PANTHER GO Slim) or an internal tracking ID (for PANTHER Protein Class). The PANTHER id is usually a 5 digit number preceded by letters PC (for Protein Class).

Parent lineages The path of the category with all parent categories. For GO terms, the parent is not the parent category in the full Gene Ontology, but the parent GO category in the PANTHER GO Slim. Clicking the name will take you to the detail page of the ontology term.

Child The list of names of all child categories. For GO terms, the child is not the child category in the full Gene Ontology, but the child GO terms in the PANTHER GO Slim. Clicking the name will lead you to the PANTHER detail page of the child category.

Number of subfamilies/Families The number of subfamilies and families that are classified by the category. Clicking on the number will lead to the Family/Subfamily List page (section 2.3.2).

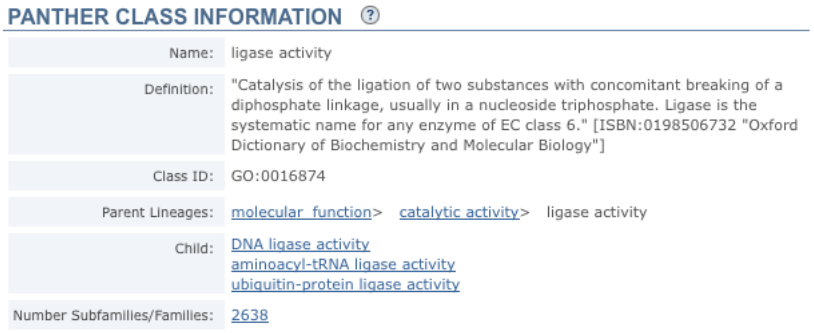


Figure 2.17: *Ontology term detail page.*

2.3.4 Pathway description page

This page provides details of the pathway. To reach this page, just click the pathway name link anywhere on the PANTHER website, and then click the "Pathway Description" tab on the page. You can toggle between this page and Pathway Diagram page.

This page provides following information about a pathway (Figure 2.18):

Definition A detailed description of the pathway.

Pathway Accession

Components The number of pathway components within the pathway. A pathway component is a group of homologous proteins across various organisms that participate in the same specific biochemical reactions within the pathway.

Clicking on the number will lead to a Pathway component list page (section 2.2.4).

Subfamilies The number of subfamilies that have at least one training sequence associated directly with the pathway. Clicking on the number will lead to a Subfamily list page (section 2.2.2).

Associated Sequences The number of training sequences that are directly associated with the pathway by manual curation. Clicking on the number will lead to the Gene list page (section 2.2.1).

References A list of literature references or websites used to generate the pathway diagram.

Author The author who generated the pathway diagram, and curated training sequences to pathway associations.

Released On The date the pathway was generated and released.

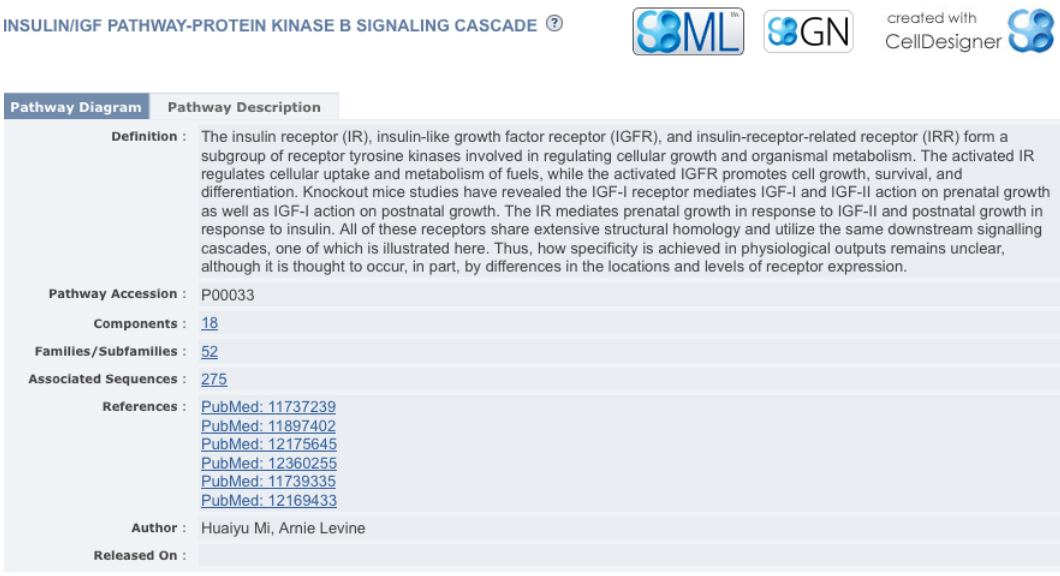


Figure 2.18: *Pathway Description page.*

2.3.5 Pathway molecular class (component) detail page

A pathway component represents a group of homologous proteins across various organisms that participate in the same specific biochemical reactions within the pathway. A pathway component is always associated with genes within the PANTHER libraries, thus, it is linked directly or indirectly to the PANTHER family/subfamily statistical models.

This page provides details of a pathway component in two separate parts (Figure 2.19):

1. Pathway component

Name The name that appears on the pathway diagram. It is usually an acronym or a short version of the full name.

Full name The complete, more descriptive version of the name.

Synonyms All other names used to describe the component.

Definition A short description of the component.

Category ID Each component is considered as a pathway ontology term, and is assigned with an ID. It is usually a 5 digit number preceded by letter(s) to indicate the type of category, e.g., G for DNA or RNA, and P for protein.

Upstream A list of component(s) that modifies the current component in the pathway. It usually implies that these components have protein-protein interaction.

Downstream A list of component(s) that is modified by the current component in the pathway. It usually implies that there are protein-protein interaction involved.

Number Subfamilies Number of subfamilies that have training sequences directly associated with the component. Clicking on the number will lead to the subfamily list page (section 2.2.2).

Number Associated Sequences Number of genes that are directly associated with the component. Clicking on the number will lead gene list page (section 2.2.1).

GO classifications and PANTHER protein class The ontology terms classified to the genes that are associated to the pathway component. Clicking on an ontology terms will lead to the *Ontology Term Details page* (section 2.3.3).

References A list of literature references or websites used to generate the pathway diagram.

2. Sequence Association A list of all training sequences that are associated with the component by manual curation. The list includes the following columns:

   PANTHER Subfamily PANTHER subfamily Ids. It links to PANTHER subfamily information page.

   Associated Sequences Training sequence ID. It links to NCBI Entrez page.

   Definition The gene definition for the training sequence

   Evidence Code* See below.

   Evidence A list papers with experimental results indicating the involvement of the sequence in the pathway. It is usually in the form of PubMed IDs.

*What is an Evidence Code?

Evidence Codes are used to denote the type of the evidence used for associating sequences to pathway components. The PANTHER Evidence Code is modeled after the GO Evidence Codes (http://www.geneontology.org/GO.evidence.shtml).

Figure 2.19: *Pathway Molecule Class (component) detail page.*

2.4 Graph and diagram pages

2.4.1 Pie charts

A pie chart can be generated from a gene list (section 2.2.1). Whole genome pie charts for all 48 organisms in PANTHER have been pregenerated and can be accessed from the *Genes and Orthologs* subject home page (section 2.1.1). Here is what you can do on this page (Figure 2.20).

- If it is a whole genome pie chart, you can choose the ontology you want to display from the *Select ontology* drop-down menu.
- From the *Select species* drop-down menu, you can choose one of the 48 organism you want to display.
- You can see the category names and statistics by mouse-over the pie chart selection. See detailed explanation of the statistics below.
- You can drill down to pie chart of child categories by clicking on a pie chart selection.
- You can retrieve a gene list by clicking on the chart legend link on the right side.
- The page also provides links to allow you to convert the pie chart to bar chart or to export the data as .txt file, so you can use your own program to create charts.

How to read the pie chart statistics?

When you mouse-over the pie chart selection, the category name and a series of numbers are displayed. These numbers are the statistics for the category you selected. Here is how the numbers are from.

1. The first number is the number of genes that are classified to this category. In our example in figure 2.20, it is 2067
2. The second number is the percent of genes classified to this category over the total number of genes. Right above the pie chart, it shows that the total number of genes is 17181. Remember this is the total number of human genes in the PANTHER library.
3. The third number is the percent of genes classified to this category over total number of class hits, in this case 13148. Class hit means independent ontology terms. If a gene is classified to 2 ontology terms that are not parent or child to each other, it counts as 2 class hits.

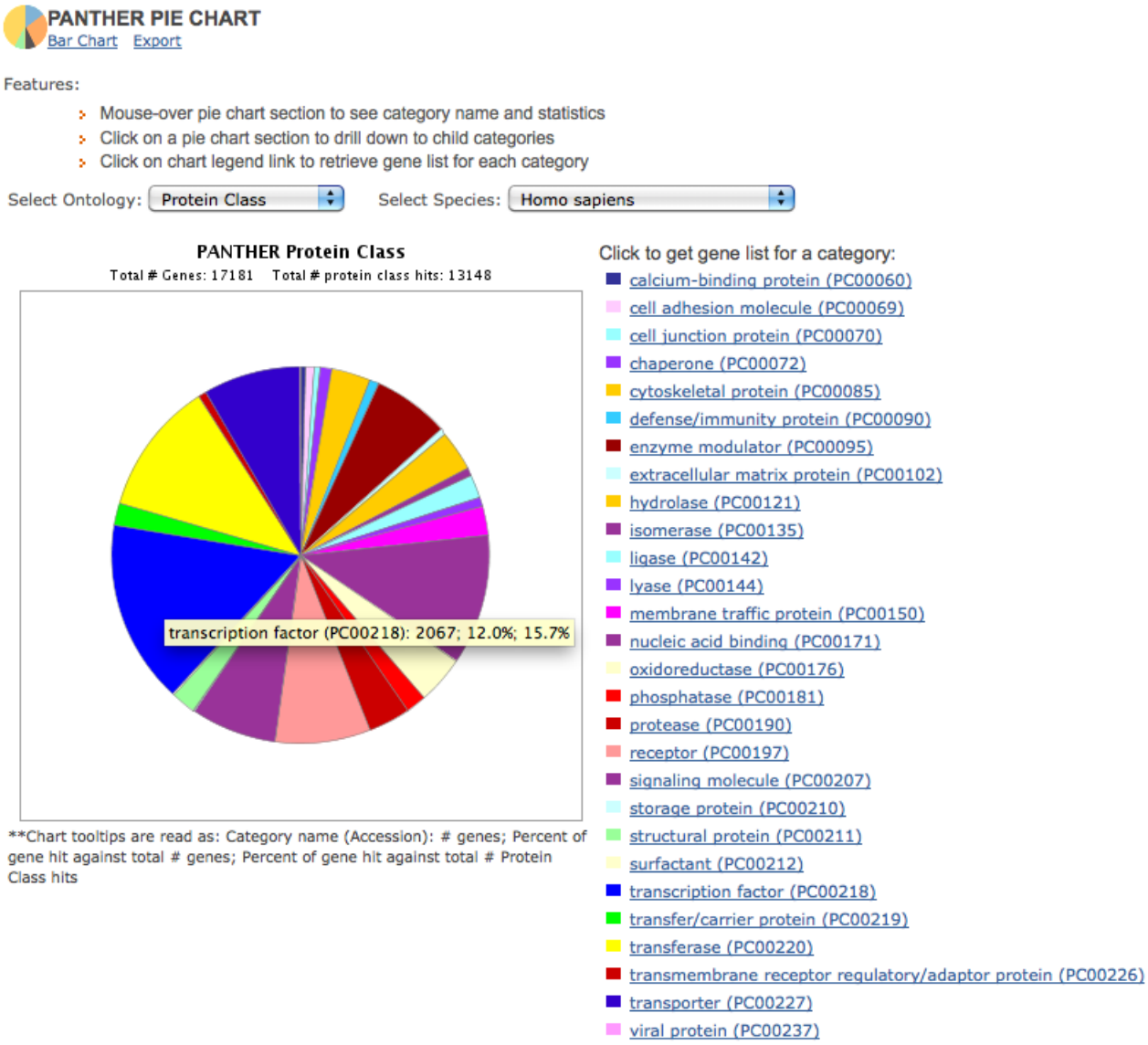


Figure 2.20: *The pid chart of the entire human genome in PANTHER Protein Class.*

2.4.2 Pathway diagram

2.4.2.1 PNATHER Pathway Applet Basics

The pathway diagram on this page is interactive, and is displayed in an applet window. The window is divided into 2 panels (Figure 2.21).

- The left panel shows a list of molecules shown in the diagram. The types of the molecules are in parentheses, e.g., PROTEIN, SIMPLE_MOLECULE, GENE, etc.
- The right panel is the diagram. All objects on the diagram are interactive, so you can customize the diagram according to your preferences. The modified diagram can be saved as png file as described in section 2.4.2.2.

There are three views for the diagram.

- Interactive Standard/Activity Flow - When this tab is selected, the applet allows you to view two views that are not SBGN compliant:

– Standard View – This view faithfully reflects the original diagram generated by the CellDesigner software. It captures detailed molecular events or biochemical reactions within the pathway. It is a little different from the conventional way of illustrating pathways. For example, if a protein A activates protein B, it will be illustrated as protein A catalyzing a transition reaction of protein B from a non-active to an active form.

– Activity Flow – This view uses simpler or more conventional relationship notations to illustrate information flowchart of the pathway. It captures mainly 4 relationships between 2 proteins: activation, inhibition, neutral, and unknown. It is derived based on the Standard View. In cases where a reaction can not be simply illustrated as one of the above four, it will keep it in the original Standard view. This view is not compliant with SBGN Activity Flow Diagram.

The Activity Flow view is the default view of the pathways. You can toggle the two view by clicking the tab on the top of the applet window.

- SBGN Image View - When this tab is selected, a static (non-interactive) image of SBGN Process Description Diagram is displayed.

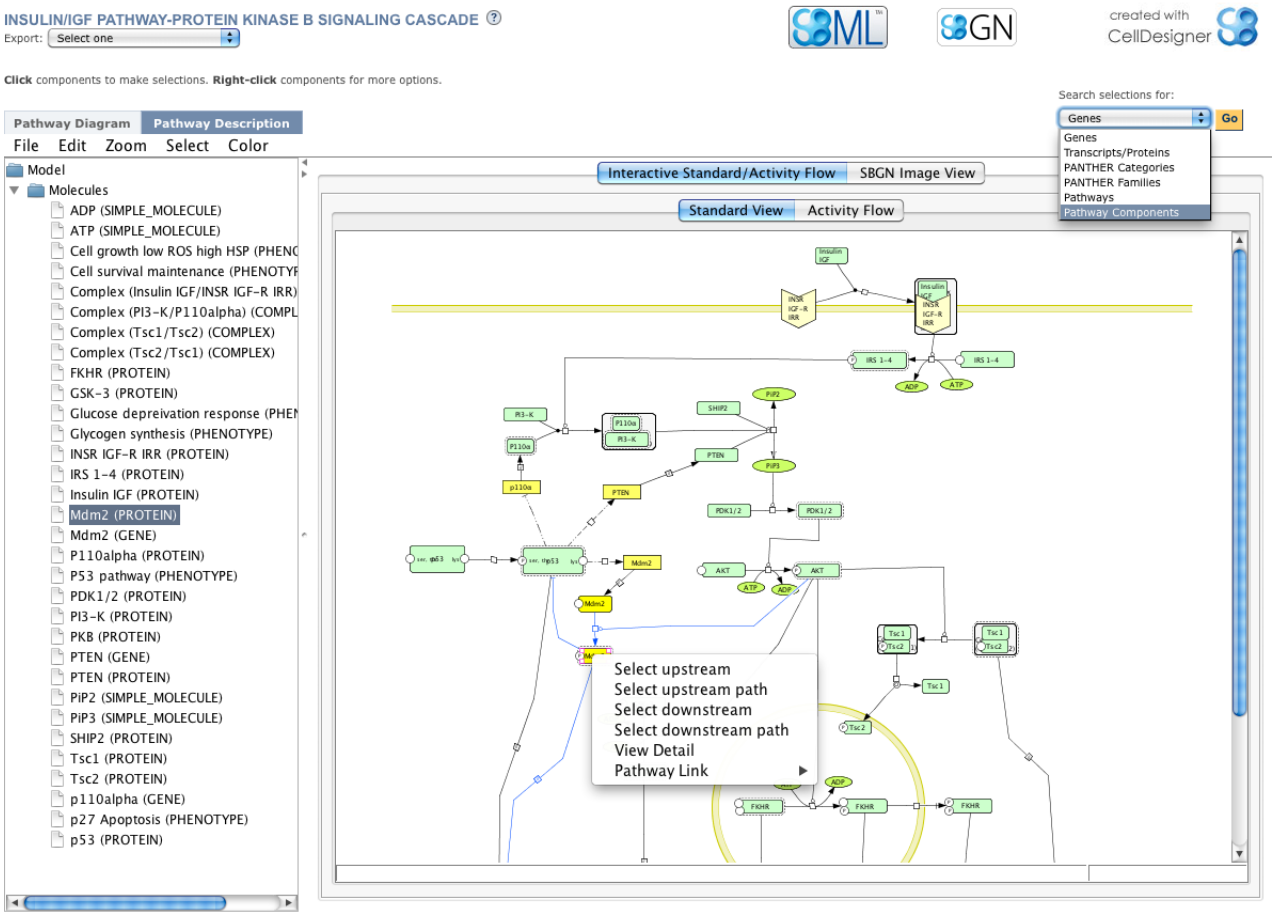


Figure 2.21: *Pathway Diagram page.*

2.4.2.2 Export data from this page

Users can export pathway related data using the following options.

- From the *Export* drop-down menu, all the pre-generated images and SBML file of the pathway can be exported
- From the applet panel, under "File", select "Export image" to export the current pathway diagram. This function allows you to export the modified image.

2.4.2.3 Using the pathway diagram

The diagram page allows you to make selections on pathway components, and then generate gene lists, transcript lists, or family lists.
You can use any of the following ways to select a single or multiple components.

1. You can select a component by clicking on the name of the component in the left (list) panel of the applet. If you want to select more than one component, press and hold the control key while you select. You can clear your selection by going to "Select -¿ Clear all" on the applet menu bar.
2. You can select components from the right diagram panel by clicking on the component symbol. You can clear your selection by clicking on the selected item again. You don't need to hold the control key for multiple selections.
3. Under "Select" applet menu bar, you can do the following.
   - Select all proteins, genes and RNA. Since only proteins, genes and RNA in PANTHER pathway are associated with PANTHER family genes, you have an option to select all of them by one click.
   - Select all. This function allows you to select all items, including simple molecules, ions, phenotypes, etc.
   - Clear all. You can use this function to clear all your selections.
4. By right-clicking on the component, you can make the following selections.
   - Upstream or Downstream component. This function will select the immediate or downstream component. If the selected component appears in multiple places in the same pathway, they will be highlighted also.
   - Upstream or Downstream path. This function will select all components in the entire upstream or downstream path.

Once the selections are made, you can generate any of the following lists by making the selection to the *Search selection for* drop-down menu on the right side of the page.

- Genes: A list if genes whose proteins hit PANTHER family and subfamily models with at least one training sequence associated with the selected pathway component.
- Transcripts/proteins: A list of proteins that hit PANTHER family and subfamily models with at least one training sequence associated with the selected pathway component.
- PANTHER Ontology Terms: A list of PANTHER Molecular Function and Biological Process ontology terms that are classified to PANTHER family or subfamily models with at least one training sequence associated with the selected pathway component.
- PANTHER Families: A list of PANTHER families or subfamilies with at least one training sequence associated with the selected pathway component.
- Pathway
- Pathway Components

2.4.3 TreeViewer

PANTHER trees allow you to explore the relationships between sequences in a particular family, as well as view some of the key information that was used to annotate the families and subfamilies. The current PANTHER release (version 7.0) contain publicly-available protein sequences from UniProt, Refseq, ENSEMBL and Model Organism Databases.

The Tree Viewer has two panels that are mapped to each other (Figure 2.22). The left panel displays the relationship between the different sequences, and is labeled as *Tree*. The right panel has two views. The first is the "attribute table", which is labeled as *Grid* and contains information for the each corresponding sequence in the tree panel. The second view is the "multiple sequence alignment" view, which is labeled as *MSA* (Figure 2.24). You can click the the button to toggle between the two views.

2.4.3.1 *Trees*

In the tree panel on the left of the Tree Viewer, you can explore that phylogenetic relationships among different genes within the family. The longer the (horizontal) branch length, the more distant the groups joined by those branches. Vertical branch length is fixed for ease of viewing together with the information on the panel.
There are 4 types of nodes in the phylogenetic tree:

- Blue diamonds - Subfamily nodes. Subfamilies are usually colored to help distinguish between different subfamilies. Aside from this, the subfamily color does not have any special significance.
- Orange circles - Gene duplication nodes.
- Green circles - Speciation nodes.
- End of the horizontal branch - Leaf nodes. They usually point to the sequence IDs

The following are a few things that you can do to these nodes:

1. When mouse-over these nodes, the subfamily annotation information is displayed. For the the subfamily node (blue diamond), it displays the subfamily id and name. For all other node types, it displays which subfamily it belongs to.
2. When a node is clicked (except for the leaf node), it collapses.
    - A collapsed subfamily view provides a high-level view of the tree, in which subfamilies are the most specific "leaves" of the tree. The subfamily name given by curators appears in the "sf_name" column of the Collapsed view. The GO and PANTHER protein class terms are shown in the appropriate columns also.
    - A collapsed non-subfamily node is shown as a red circle. The annotation node (AN) id is shown in the "Protein id" column, and the common GO and protein class annotation for all sequence below the collapsed node is shown in the appropriate columns.

Under the "Tree" menu, you can do the followings (Figure 2.22):

- Collapse lowest level of Subfamilies - A PANTHER subfamily node can have other subfamily node(s) as descendants. This option collapses all subfamily nodes that do not contain another subfamily node. All subfamily nodes that nest another subfamily node remain expanded.
- Expand all Subfamily nodes - All subfamily nodes are expanded. All non-subfamily nodes remain unchanged.

- Expand all Nodes - All nodes, including non-subfamily nodes can be expanded
- Use distances - If this option is selected (a check mark shows in front of the option), the tree branches are shown based on the distances calculated in tree building algorithm (GIGA) [4]. If this option is unchecked, the tree branch will be shown with fixed length.
- Ladder Top - Re-arrange the tree so that the node with more descendants are always above the nodes with less.
- Ladder Bottom - Opposite to the above. Re-arrange the tree so that the node with more descendants are always below the nodes with less.
- Latter Original - Set the order of the tree to what was loaded from the server. It is different from either Ladder Top or Ladder Bottom.
- Colore Subfamilies

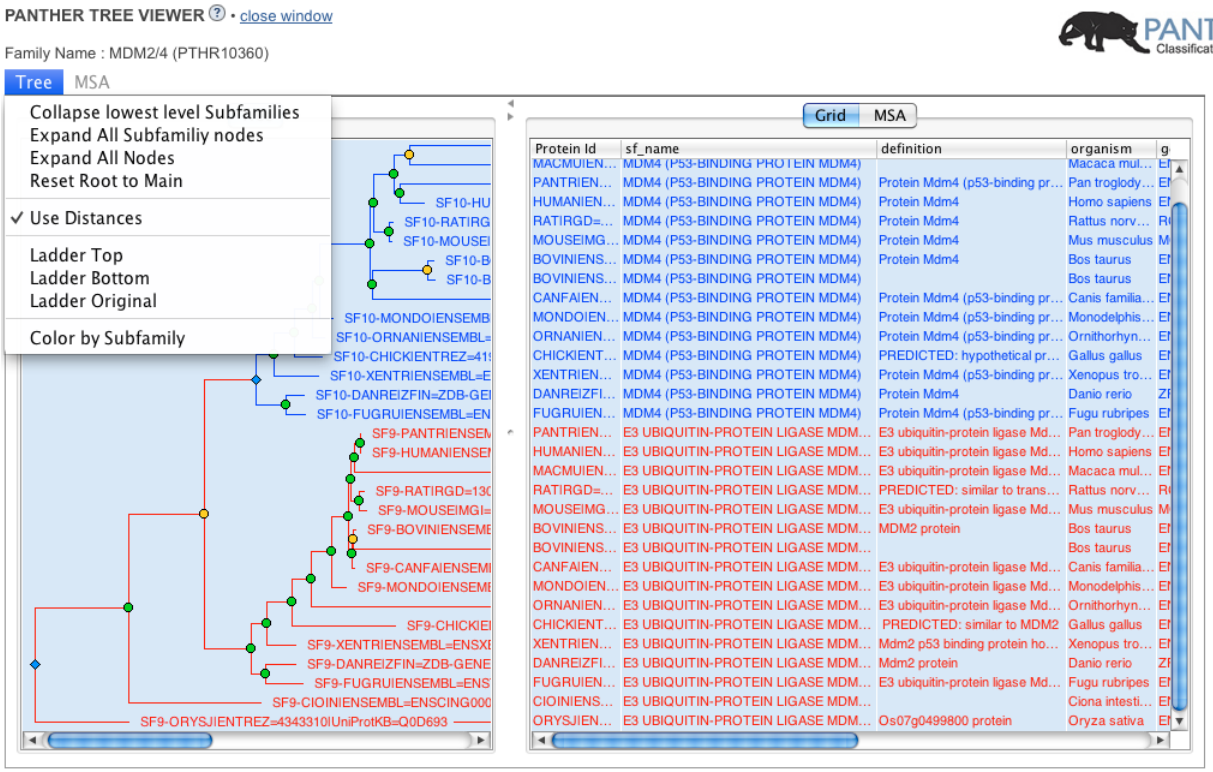


Figure 2.22: *A phylogenetic tree of PANTHER family PTHR10360 viewed in the tree viewer.*

2.4.3.2 Attribute table

The attribute table contains one row for each sequence in the tree. Each column displays a different attribute of the sequences as follows (Figure 2.23).:

- Protein id - This is the long id for each sequence. The format of the id is as below: organism|gene DB=gene id|protein DB=protein id. When the tree is collapsed, this column shows the annotation node (AN) ids.
- sf_name - The name annotated for the subfamily.
- definition - This is the brief definition line parsed out from the protein database where the protein sequence is from.

- organism - This is the organism from which the sequence was derived. Click on the organism name to open the full taxonomy record for that organism.
- gene id
- gene symbol
- OrthoMCL - The OrthoMCL id that the gene is predicted to belong to.
- Molecular function, Biological Process and Cellular component - These are Gene Ontology terms from PANTHER GO Slim describing the function of the gene product.
- Protein Class - This is a PANTHER Index term describing protein classes.

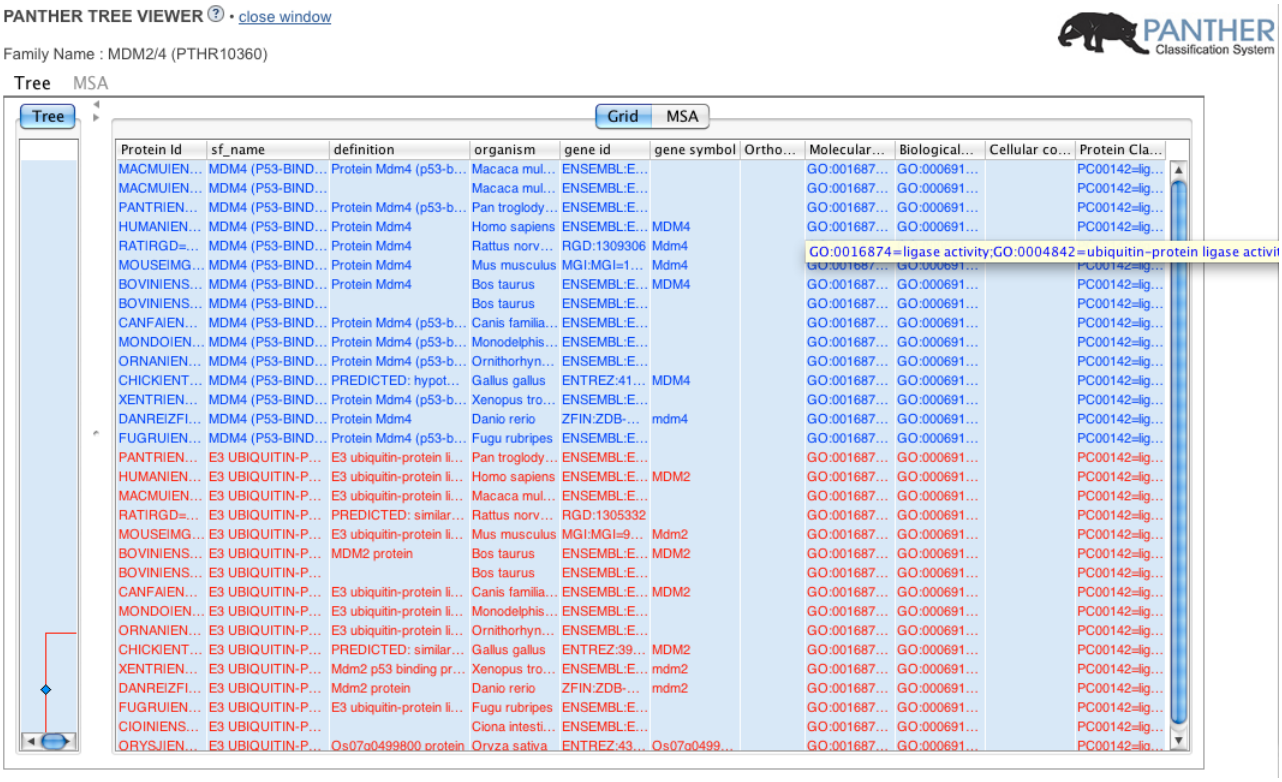


Figure 2.23: *A tree viewer with expanded attribute table.*

2.4.3.3 MSA

When the "MSA" button is clicked on the right panel, the multiple sequence alignment view is shown. You can toggle back to the attribute table view by clicking the "Grid" button.

PANTHER multiple sequence alignments (MSAs) are the basis for the PANTHER distance trees, and therefore of the family/subfamily classification. MSA is built using the MAFFT software [5]. MSA is aligned across the entire length of all sequences in the family. The evolutionarily conserved part of the alignment is used to build HMM, and it is called the *match state*, which is designated with *uppercase letters*. The other less conserved region is designated by *lowercase letters*, and is called *non-match state*. If a sequence misses a position in the match state, it is called a *delete state* and is designated by a *dash*. If a sequence misses a position in the non-match state, it is called an *insert state* and is designated by a *dot*.

In the MSA menu of the MSA view, you can make the following selections.

- Percentage Identity for Entire Alignment - This is the default selection and shows the alignment across the entire length of the all sequences.
- Percentage Identity for Match States - This only shows the alignment for the match states

The MSA uses uppercase letters to indicate match state positions, lowercase letters to indicate non-match state positions, dots to indicate insert state positions, and dashes to indicate delete state positions.

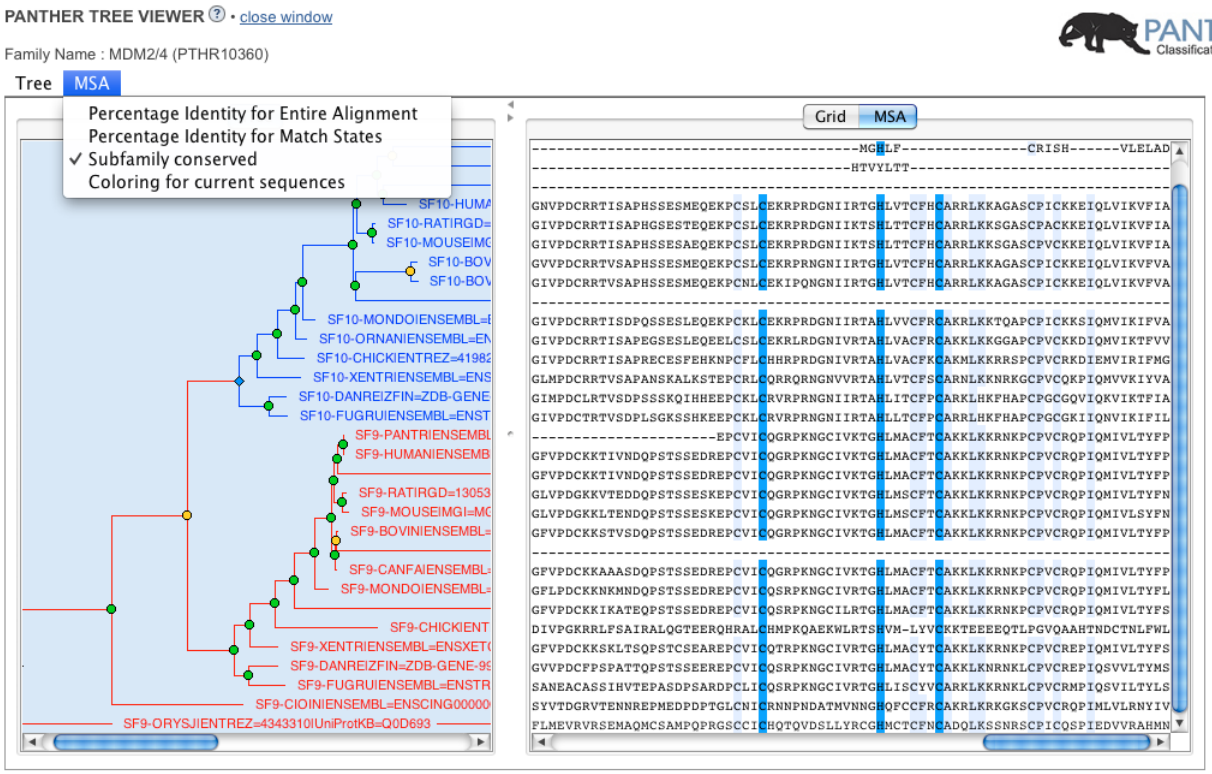


Figure 2.24: *The multiple sequence alignment of PANTHER family PTHR10360 viewed in the tree viewer.*

Chapter 3

PANTHER Basics

3.1 Keyword Search

There are two types of keyword search pages, the **simple keyword search** and **advanced keyword search**.

3.1.1 Simple keyword search

The **simple keyword search** function can be found on the PANTHER home page as well as all the subject main page (section 2.1). It allows users to do a quick search of the PANTHER database. Tips about the search:

- The search term can be an identifier, a word or a phrase (multiple terms).
- The search looks for the exact match of the entire word. For example, if you enter the term "hexo", the search will not find "hexokinase". Therefore, if you are not sure about the word, a wild card (*) can be used. In this case, if you enter "hexo*", the search will find "hexokinase".
- You can specify a particular subject to search from the drop-down menu, such as *Genes and orthologs* or *Pathways*. The default on the home page is *All*, and the default on each subject main page is the subject itself.
- The search looks for the search term(s) in all fields of the selected subject. Some of these fields may not be visible to you on the result list page, such as synonyms, definitions.

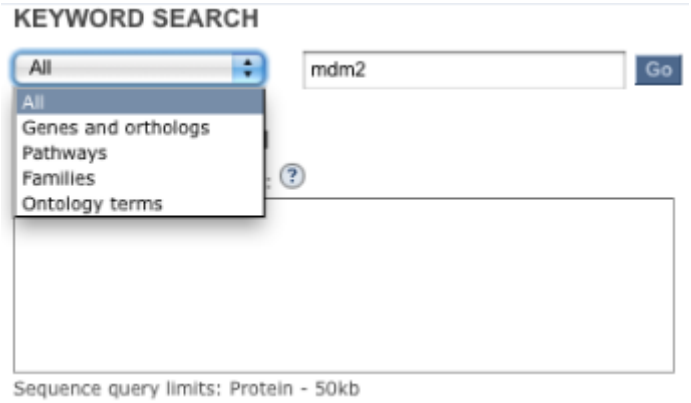


Figure 3.1: *Simple keyword search*

3.1.2 Advanced keyword search

The result of the *simple keyword search* will lead to the *advanced keyword search* page, where you can refine your search, or you can display the results (list pages). This search page can also be accessed from the keyword search link in the *Genes and orthologs* main page (section 2.1.1).

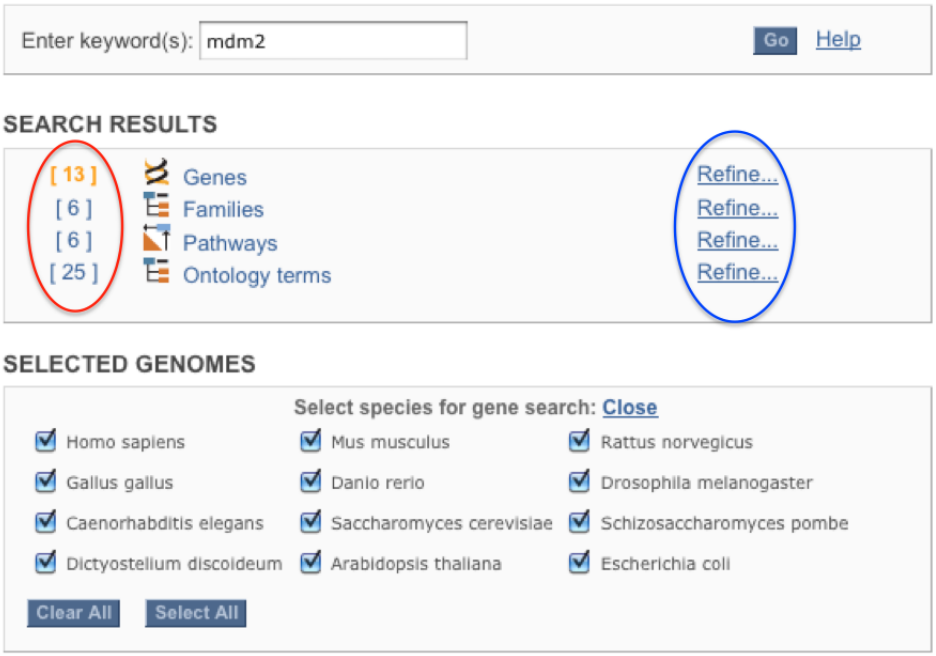


Figure 3.2: *Advanced keyword search*

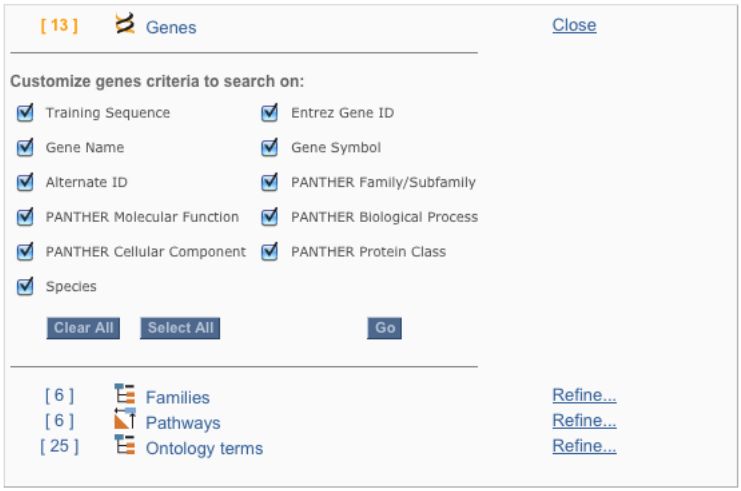


Figure 3.3: *The search criteria can be refined by clicking the* refine *link for each subject.*

The page is divided into 3 parts (Fig 3.2).

- The top part is the search box to enter the search term(s).
- The middle part is the search results section. It displays results from 4 subjects. The number of results for each subject is on the left side (red circle). In the example in Figure

3.2, the search term of *mdm2* yielded results of 13 genes, 6 families, 6 pathways and 25 ontology terms (If you access this page from the keyword search link in *Genes and orthologs* main page (section 2.1.1), it shows "none" in these brackets). You can refine the search criteria by clicking the "Refine" link on the right side (blue circle), and make selections on the expanded panel (Figure 3.3).

- The bottom part allows you to select genome(s) from the 12 Model Organisms that you want to search. The default is all.

3.2 Prowler

The PANTHER Ontology Browser (or PANTHER Prowler) is a highly interactive Java applet used to browse the PANTHER ontologies, make selection(s) , and retrieve results for different data associated with the ontology terms, such as individual genes or families and subfamilies of proteins.

The Ontology Browser is composed of 4 panels (Figure 3.4).

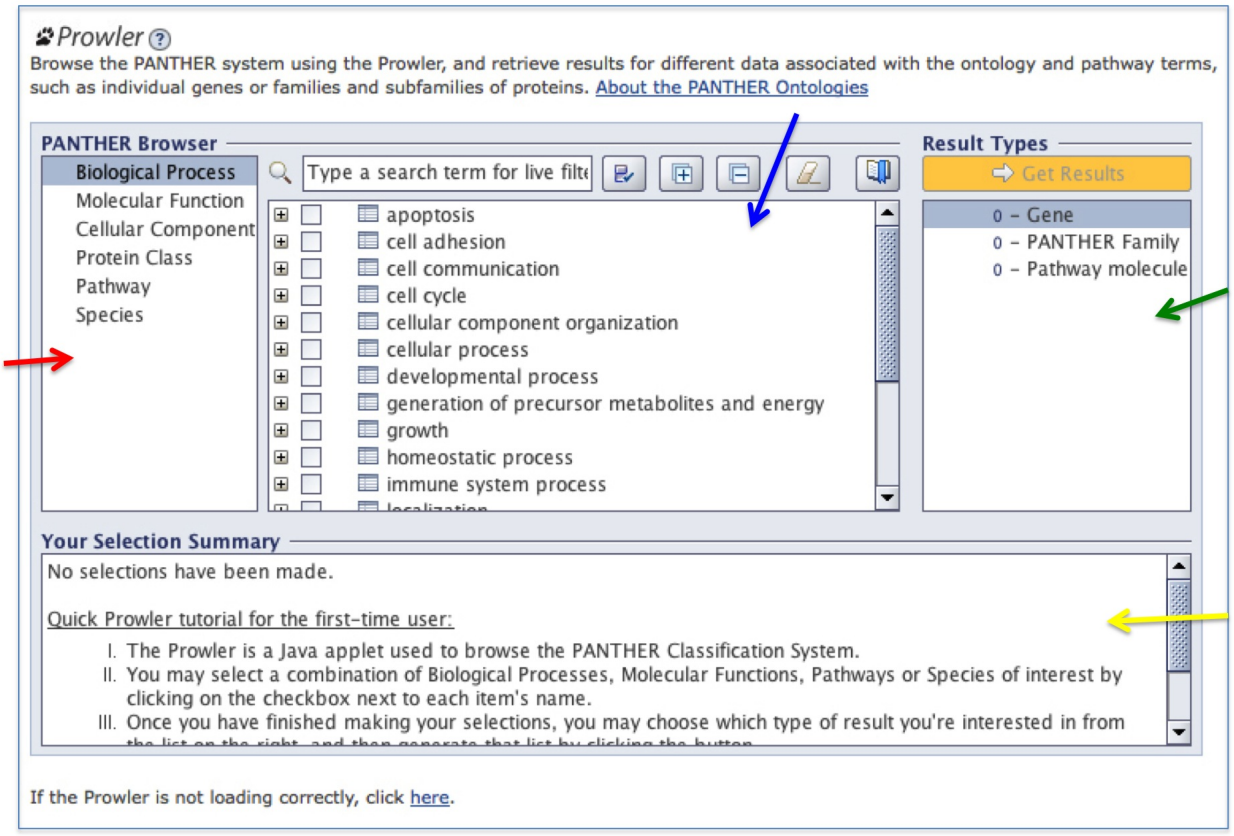


Figure 3.4: *The PANTHER Prowler*

- Ontology Type (red arrow) — On the left side of the Browser is the Ontology Type panel. There are **five** ontology types you can choose.
  - Biological Process – this is a GO Slim *biological process* ontology
  - Molecular Function – this is a GO Slim *molecular function* ontology

  - Cellular Component – this is a GO Slim *cellular component* ontology
  - Protein Class – this is PANTHER ontology to categorize protein families
  - Pathway

  The last choice is *species*. You can choose the species when querying over certain result sets such as genes, to narrow your search. There are 48 total species that you can choose, with 12 model organism species listed first, and the 36 remaining in alphabetical order. The default is all species.

- Ontology (blue arrow) — This panel is in the middle of the prowler, and it changes based on the selection of the ontology type. You can click on the + to expand the next level of the ontology terms. The details of each ontology term can be viewed by mouse-over the name, or by clicking the little icon right in front of the name.

  On the top of the panel, there are a few special utilities:

  - Search. You can type the term in the *Type the search term for live filtering* box, and only the ontology terms containing the search will be displayed.
  - Expand to selections. Clicking this button will expand only the paths to the selected ontology terms.
  - Expand all. Clicking this button will expand the ontology so that all child terms (subcategories) are shown.
  - Collapse all. Clicking this button will collapse the ontology so that only the topmost (general) terms are shown.
  - Clear all selections. Clicking this button will clear all selections of all types of ontologies.
  - Bookmark. Clicking this button will cause the page to reload with a special address (URL). This address will have all your selection encoded. If you save this address to your bookmark or favorites, you can return to your current selection at any time. This address is also suitable for cut and paste into your documents, email messages, or reference documentation.

- Results Type (green arrow) — This panel is on the right side, and it displays the number of genes, PANTHER families, and pathway molecule classes (components) that matches your selections. You can choose one of the above result types, and click the *Get Results* button. The results will be displayed as a list page (Section 2.2).

- Selection Summary (yellow arrow) — This panel is in the bottom of the Prowler. Since you are allowed to make selections from multiple ontologies, a summary of the query results from your selections is displayed in this panel.

Browse the PANTHER ontologies
Figure 3.5 shows an example of making selections using the PANTHER Prowler.

1. Select an ontology on the *Ontology type* panel.

2. Select the ontology you are interested in (left panel). Typing in the text field will filter the list of possible selections. Click on a + (red arrow) to expand a category to view subcategories, and select categories by clicking on the checkbox next to the name. Brief details about each category can be seen by mousing-over a name, or full details by clicking on the icon to the left of a name. For pathways, full details include a detailed diagram of the molecular interactions and reactions in the pathway.

3. More than one ontology can be selected. When multiple selections are made, it means that results will meet the criteria from all the selections. In this example, selections are made in both *Molecular Function* and *Pathway*, and the selected ontology can be seen as marked in the Ontology panel (red oval in Figure 3.5).

4. Review your selections in the Selection Summary panel (green arrow). For example, in this example, one selects *Human* from the species ontology as well as *Protein kinase activity* from the molecular function ontology and *Apoptosis signaling pathway* in pathway to retrieve all genes that have protein kinase activity that involve in apoptosis pathway in the human genome.

5. Human curators have associated ontology terms with genes, PANTHER families, subfamilies and pathways To see these associations, select the result type in the right panel and press the orange *Get Results* button (blue arrow). It will take you to the list pages.

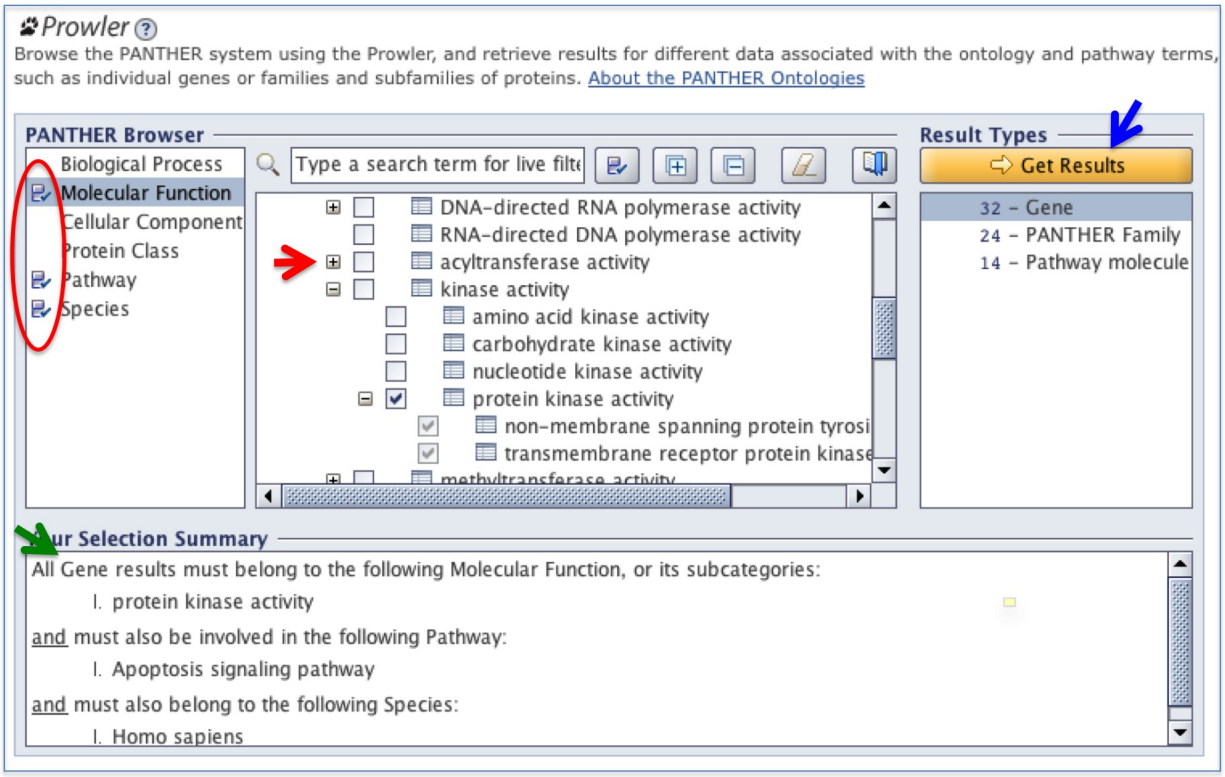


Figure 3.5: *An example showing selections using Prowler*

3.3 Batch ID Search

Batch ID search is a unique functionality that you can leverage expert curated PANTHER knowledgebase to classify multiple genes of your interest. This is how it works. PANTHER library contains genes and proteins from UniProt, Refseq, ENSEMBL, and Model Organism Databases. Each gene belongs to a protein subfamily. Many subfamilies have been manually curated with ontology terms, and therefore, the sequence will have the same classifications

as the subfamily that it belongs to. Using the IDmapping mechanism developed by UniProt, each gene is mapped to multiple commonly used IDs, including Entrez GeneID, and GI, Gene Symbol, etc (referred to as PANTHER Supported IDs). When you submit a list of IDs, the system will automatically search through all the supported IDs. The mapped one will carry over all the PANTHER classifications.

Below is a brief description of how you can use the tool (Figure 3.6).

1. Enter your IDs. There are two ways to enter your IDs.
   - Type or paste a list of IDs directly to the box.
   - Use the Browse button to upload a list file.

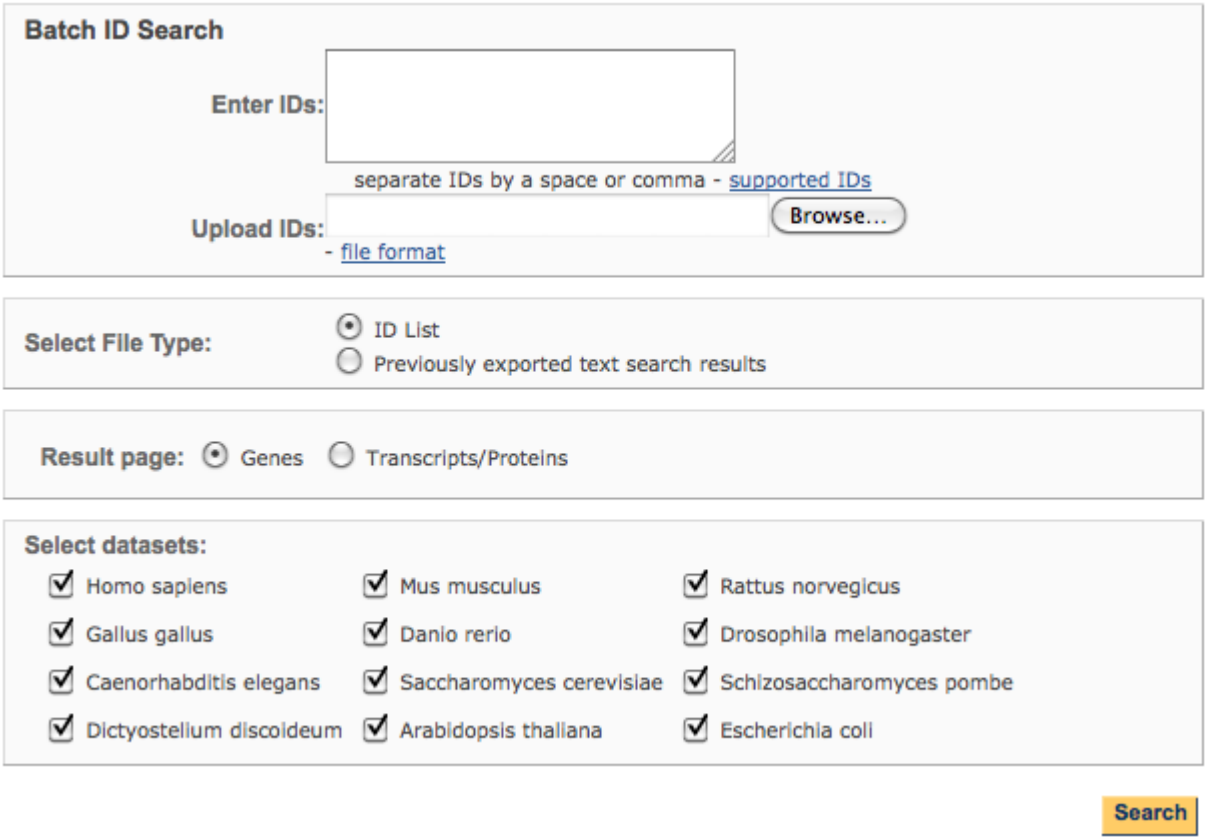


Figure 3.6: *Batch ID upload.*

Please pay attention to the following.

- Supported IDs - As mentioned earlier, not all IDs are supported by the PANTHER. We are working hard to improve this, but at the moment, only the following IDs are supported (You can also click the "supported IDs" link to get the following list). Please note that the system is not able to do partial match so you need to input the full IDs.
  - Ensembl: Ensembl gene identifier. Example: ENSG00000126243
  - Ensembl_PRO: Ensembl protein identifier. Example: ENSP00000337383
  - Ensembl_TRS: Ensembl transcript identifier. Example: ENST00000391828

  - Gene ID: EntrezGene IDs. examples include, GeneID:10203, 10203 (for Entrez gene GeneID:10203)
  - Gene symbol: for example, CALCA
  - GI: NCBI GI numbers. Example: 16033597
  - HGNC: HUGO Gene Nomenclature ids. Example: HGNC:16673
  - IPI: International Protein Index ids. Example: IPI00740702
  - UniGene: NCBI UniGene ids. Examples: Hs.654587, At.36040
  - UniProtKB:UniProt accession. Example: O80536
  - UniProtKB-ID: UniProt ID. Example: AGAP3_HUMAN
- File format - The tool only accepts simple text (.txt) in the following format.
  - If your file contains a list of IDs, separate the IDs by *tab, carriage return, line return, comma, pipe* or *space*.
  - If your file contains previously exported genelist (see section 2.2.1 about how to export a gene list), use the tab-delimited file resulting from the export feature. The system searches on the Gene ID column.

  Microsoft Excel file format is not accepted by the tool. To convert an Excel format (.xls or .xlsx) to a simple text format (.txt) - Open the file in Excel, under *File* menu, choose *Save As*, and under the *Format* option, choose *Tab Delimited Text*, and then click the *Save* button.

2. Select File Type - As mentioned above, you can select either *ID list* or *Previously exported gene list*.

3. Results page - You can choose what to display on the result page. It can be either a Gene list page (section 2.2.1 or protein list page.

4. Select datasets - This allows you to select what organism(s) you are searching. Default is 12 model organisms.

3.4 PANTHER HMM Score

The function can be accessed from the PANTHER home page, or the "interactive score of new sequences against PANTHER HMMs" link on the *Trees and HMMs* subject home page (section 2.1.2). This online functionality only allows you to submit one sequence at a time. For multiple sequence scoring, you need to download the PANTHER scoring tool from the PANTHER ftp site at http://www.pantherdb.org/downloads/.

To score, just enter the amino acid sequence into the box, and click the "Submit" button. Here is how the result page is read (Figure 3.7).

- PANTHER hit: Only the top hit HMM is reported here. The family or subfamily id and name of the top hit are displayed on the page. Clicking the name will take you to the family/subfamily detail page (section 2.3.2).

- PANTHER score: The score of the PANTHER hit. The green dots next to the score indicates how closely related the protein is to the model. There are three categories:

  - closely related (indicated by 3 greet dots): if the score is better than E-23 (very likely to be a correct functional assignment)
  - related (indicated by 2 green dots): if the score is better than E-11, but worse than E-23 (molecular function likely to be the correct but biological process/pathway less certain)

– distantly related (indicated by 1 green dot): if the score is better than E-3, but worse than E-11 (protein is evolutionarily related but function may have diverged)

PANTHER HMM SEQUENCE SCORING RESULTS

The top scoring HMM is reported, along with the E-value (the number of expected false-positive hits expected). If the E-value is less than 1e-3, no hits are reported.

PANTHER Hit: APOLIPOPROTEIN E (PTHR18976:SF2)
HMM E-value score: 2.2e-155

```
Sequence Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
sequence   1/1       1   317 []     1   331 [.   527.0 2.2e-155

Alignments of top-scoring domains:
sequence: domain 1 of 1, from 1 to 317: score 527.0, E = 2.2e-155
                     *->lmkllwalmkvlllallvslLagcqakmkglavllalalltgcqars
                                mkvl++allv++Lagcqak++++++++++++l+ +q++
    sequence     1    --------MKVLWAALLVTFLAGCQAKVEQAVETEPEPELR-QQTE- 37

                     lpladepkqrWEealdrFwdYlrelqtladdvqeelkssQiskELdtLie
                         +++ qrWE+al+rFwdYlr++qtl+++vqeel+ssQ+++EL++L++
    sequence    38 ----WQSGQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELRALMD 83

                     DTMtELkaYkeeLeeqLtPvaeetrarlskelqalqarLgaDMeDakeRl
                     +TM+ELkaYk+eLeeqLtPvaeetrarlskelqa+qarLgaDMeD+  Rl
    sequence    84 ETMKELKAYKSELEEQLTPVAEETRARLSKELQAAQARLGADMEDVCGRL 133

                     tqYrgElqamleQnteevRaRvssylrKLrKRLlkDaeeLqkrlavYqag
                     +qYrgE+qaml+Q+tee+R R++s+lrKLrKRLl+Da++LqkrlavYqag
    sequence   134 VQYRGEVQAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQKRLAVYQAG 183

                     aregaergvealRerlePlleqgqdraqaklgslaealksQakelkerle
                     aregaerg++a+Rerl+Pl+eqg+ r +a++gsla +
    sequence   184 AREGAERGLSAIRERLGPLVEQGRVR-AATVGSLAGQ------------- 219

                     sqaedlkerleataeelrgrLeeagerlrdrLdevreqieevkakvee..
                         +l+er++a++e+lr+r+ee+g+r+rdrLdev+eq++ev+ak+ee++
    sequence   220 ----PLQERAQAWGERLRARMEEMGSRTRDRLDEVKEQVAEVRAKLEEqa 265

                     ..................FePLvedlqrQWagLvEKvQaavkpsaaetPas
                     ++ ++++++ ++++   FePLved+qrQWagLvEKvQaav +saa++P s
    sequence   266 qqirlqaeafqarlkswFEPLVEDMQRQWAGLVEKVQAAVGTSAAPVP-S 314

                     dnh<-*
                     dnh
    sequence   315 DNH    317
```

Figure 3.7: *The score result file from the PANTHER HMM score page.*

3.5 Download

This is the PANTHER FTP site. You can reach this page by clicking the "Downloads" link on the *Quick links* panel on the left side of the Home page. You can download the following PANTHER data and tools.

- PANTHER HMM library - The current version is 7.0 that contains the PANTHER HMMs for 6594 protein family HMMs, divided into 62,972 functionally distinct protein subfamily HMMs. The PANTHER HMMs are in the HMMER format. There is also a version of the library available that can be used for the Coding SNP Analysis tool.
- PANTHER HMM scoring tool - Download the PANTHER scoring tool, to score protein

sequences against the entire PANTHER HMM library and analyze your sequences. The HMM library should be downloaded separately.

- PANTHER cSNP scoring tool - Download the PANTHER Coding SNP Analysis tool to estimate the likelihood that a particular nonsynonymous (amino-acid changing) coding SNP will cause a functional impact to the protein. A special version of HMM library for this tool should be downloaded separate.
- GIGA - GIGA is an efficient tree building program that allows phylogenetic reconstruction of very large gene families and determination of orthologs on a large scale. It is used to generate trees in the PANTHER Classification System.
- PANTHER Pathway - Contains 165 regulatory and metabolic pathways, each with subfamilies and protein sequences mapped to individual pathway components. All pathway diagrams are abailable in both SBML and BioPAX level 3 format. SBML files, BioPAX files and Protein Sequence Association data are available for download.
- PANTHER Sequence Classification file - Contains the PANTHER family, subfamily, GO slim, PANTHER protein class, and pathway classifications for the the complete genomes derived from the 48 organisms used in the PANTER database.
- PANTHER HMM Classification file - Contains the PANTHER family/subfamily name, and the GO slim, PANTHER protein class, and pathway classifications for every PANTHER protein family and subfamily in PANTHER HMM library.
- PAINT - PAINT is a java application for viewing tree files. Minimum Java version 1.4 required. To install, uncompress the download file and execute launchTAV.bat for windows or sh launchTAV.sh for MAC.

3.6 Workspace

If you have registered and have a *Workspace* account in PANTHER, here is a description about how you can use it.

You can export any list from the list page to the *Workspace*. Typically, after a keyword search, prowler search, or batch search, you would like to save the resulting list so that you do not have to perform such search over and over. *Workspace* allows you to do so. When you use the export to workspace function on any of the list page (see description is section 2.2.1), you will be asked to provide a name (required) and a description (optional) to the list.

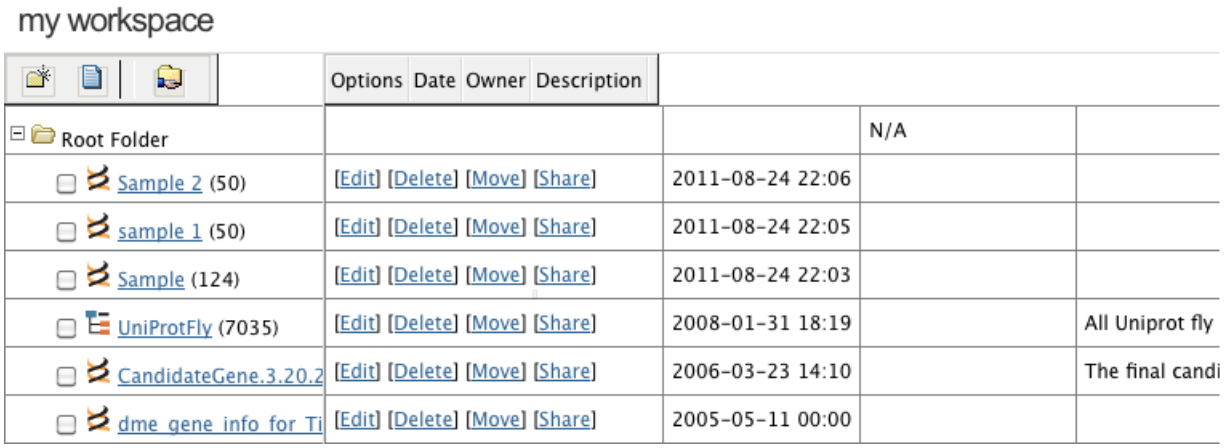


Figure 3.8: *Screenshot of a sample workspace page.*

To access your *Workspace*, you just need to click the "Workspace" link at the PANTHER Home page (Section]refsec:ov) and login. You will then reach the main Workspace page (Figure 3.8).

The exported list is in the "root folder" by default. You can organize the lists by using the "Move" link in the second column of the table to move the list to the appropriate folder. You can also create a new folder by clicking the "create new folder" icon on the top left part of the table.

Here are some other things you can do to your lists.

- Go to the list by clicking the list name in the first column. The number in the parenthesis indicates the number of entries (e.g., genes) in the list.
- Share the list with other registered PANTHER users by clicking the *Share* link or the icon. You then need to enter the username of the *Workspace* that you would like to share.
- Modify the name and description by clicking the *Edit* button.
- Delete the list.

Chapter 4

PANTHER Tools

In this chapter, we will describe the PANTHER tools for experimental data analysis [3]. As mentioned in the toool home page section (2.1.5). there are three categories of tools, gene expression analysis tools, coding SNP analysis tools, PANTHER scoring tool. PANTHER scoring tool has already been mentioned in section 3.4. This chapter will describe the rest of the tools.

4.1 Gene expression data analysis tools

4.1.1 Compare Gene List

Use the binomial statistics tool to compare classifications of multiple clusters of lists to a reference list to statistically determine over- or under- representation of PANTHER classification categories. Each list is compared to the reference list using the binomial distribution test (Cho and Campbell, TIGs 2000) [6] for each molecular function, biological process, cellular component, PANTHER protein class, or pathway term in PANTHER.

Here is how it works.

- PANTHER select specific data source(s) for the genome of each organism. It uses UniProt idmapping mechanism to map to a number of other IDs (see supported IDs).
- The tool maps the IDs from your uploaded file to the IDs in the PANTHER database.
- Since all IDs in the PANTHER database map to the IDs we use to build PANTHER protein families, thus your IDs are mapped to our PANTHER families and subfamilies, and carry over all the ontology and pathway classifications.
- The tool will then count the number of your IDs in each of the category, compare with the reference list, and use binomial distribution to calculate the p value.

Below is a step-by-step description of how to use the tool.

4.1.1.1 Select lists to analyze

When you enter the tool, the first step is to click "Select file(s)" button to upload the list(s) you would like to analyze (Figure 4.1). You may upload up to *four* lists. There are two ways you can upload the list(s).

1. Browse and upload from your computer (Figure 4.2). This is similar to Batch ID Search described earlier (section 3.3). Please pay attention to the following.
   - Supported IDs - As mentioned earlier, not all IDs are supported by the PANTHER. We are working hard to improve this, but at the moment, only the following IDs are supported (You can also click the "supported IDs" link to get the following list).

Please note that the system is not able to do partial match so you need to input the full IDs.

- Ensembl: Ensembl gene identifier. Example: ENSG00000126243
- Ensembl_PRO: Ensembl protein identifier. Example: ENSP00000337383
- Ensembl_TRS: Ensembl transcript identifier. Example: ENST00000391828
- Gene ID: EntrezGene IDs. examples include, GeneID:10203, 10203 (for Entrez gene GeneID:10203)
- Gene symbol: for example, CALCA
- GI: NCBI GI numbers. Example: 16033597
- HGNC: HUGO Gene Nomenclature ids. Example: HGNC:16673
- IPI: International Protein Index ids. Example: IPI00740702
- UniGene: NCBI UniGene ids. Examples: Hs.654587, At.36040
- UniProtKB:UniProt accession. Example: O80536
- UniProtKB-ID: UniProt ID. Example: AGAP3_HUMAN

- **PANTHER Generic Mapping File** list type: If your id type is not listed in the above supported ID list, select this list type, and you can analyze IDs not in the PANTHER system, by uploading user-generated data containing mappings between any arbitrary identifier and its corresponding PANTHER ID (see below for details about mapping). You can map your sequences to PANTHER, by downloading the PANTHER HMM scoring tools (available at the downloads section on the PANTHER site) and scoring your sequences against the PANTHER HMM library. For sequences that do not match a PANTHER HMM, please assign a PANTHER ID of NOHIT. This is very important, since this information will be used to assign sequences to the "unclassified" category. This will ensure that the statistics can be calculated correctly.
- File format - The tool only accepts simple text (.txt) in the following format.
  - If your file contains a list of IDs supported by PANTHER, separate the IDs by *tab, carriage return, line return, comma, pipe* or *space*.
  - If you are uploading a **PANTHER Generic Mapping File**, the file must be tab-delimited and contain the following columns:
    - The first column can contain any arbitrary ID (ex: a probe ID), but must be unique, since this allows the user to uniquely specify each record in the dataset, so they can track the identifier on the PANTHER website.
    - The second column should be corresponding PANTHER ID (ex: PTHR10078 or PTHR10078:SF1), and is used to look up the molecular function, biological process, and pathway associations.

  Microsoft Excel file format is not accepted by the tool. To convert an Excel format (.xls or .xlsx) to a simple text format (.txt) - Open the file in Excel, under *File* menu, choose *Save As*, and under the *Format* option, choose *Tab Delimited Text*, and then click the *Save* button.

2. Choose from your Workspace (Figure 4.3). If you registered and have a Workspace, you can upload lists that saved in the Workspace directly for analysis. See Workspace section () for its details.

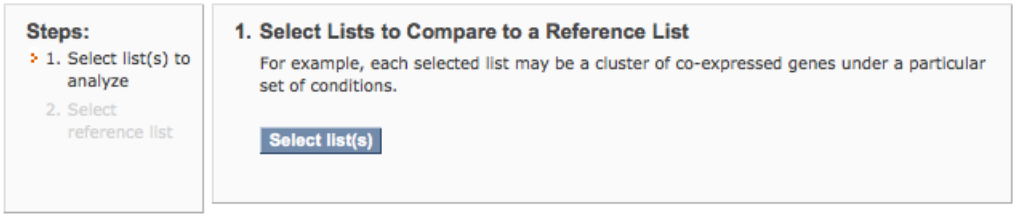


Figure 4.1: *Select a list to analyze.*

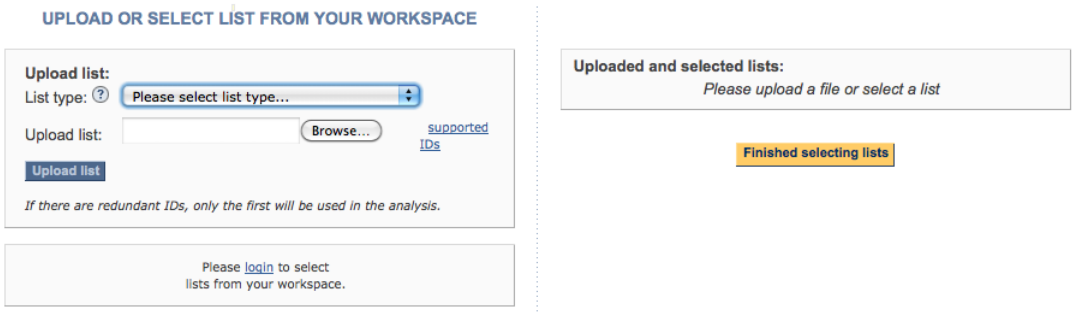


Figure 4.2: *Upload a gene list.*

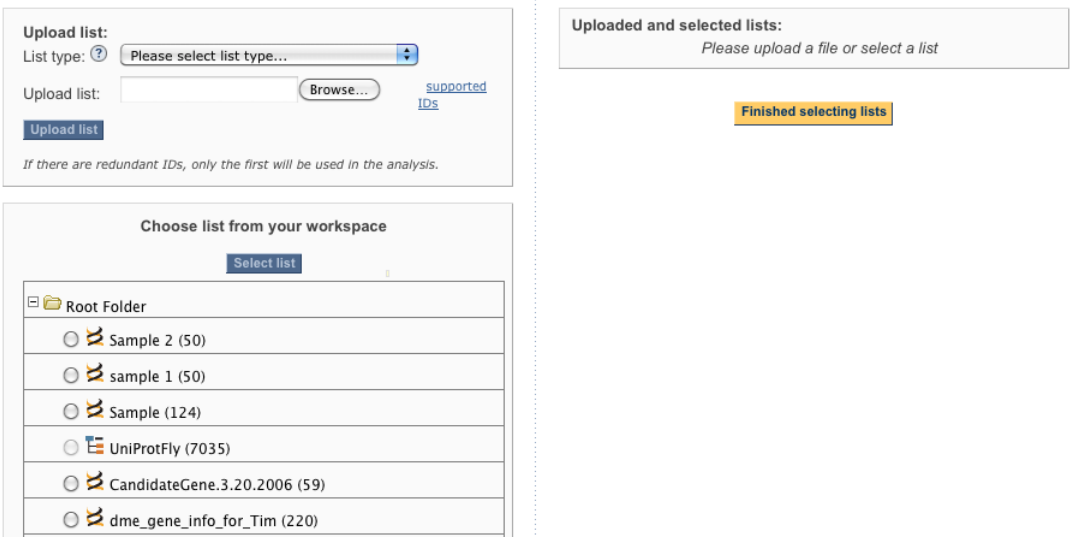


Figure 4.3: *Choose a list from the workspace.*

4.1.1.2 *Modify reference gene list and proceed*

Once a list is uploaded, it appears on the right side of the page. You can delete a list by clicking the "x" icon in front of the list name (Figure 4.4).

If multiple IDs are mapped to the same PANTHER ID, the tool will only choose one unique ID.

When all list(s) are uploaded, click the "Finished selecting lists" button (Figure 4.4).

The next page allows you to do the following(Figure 4.5).

- Modify your selected gene by clicking the "Select list(s)" button.
- Select a reference gene list . There are 3 default reference list (human, mouse, rat whole genome list). You can click the "Select reference list" button to upload your own list.
- Select a type of classification (molecular function, biological process, cellular component, protein, and pathway) you would like to analyze.

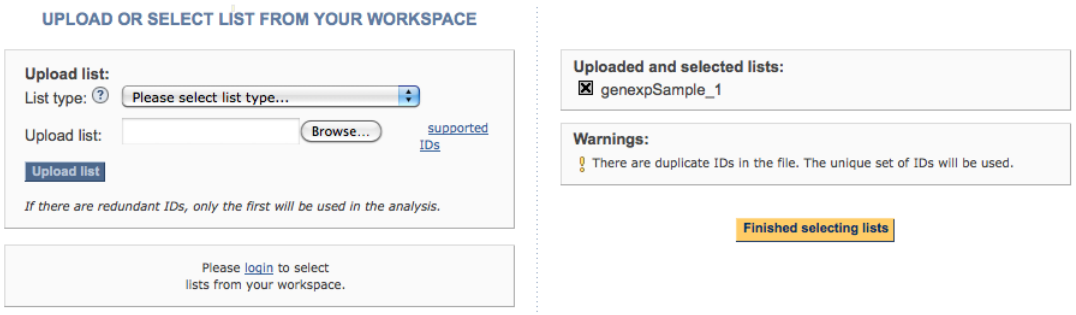


Figure 4.4: *Upload is completed.*

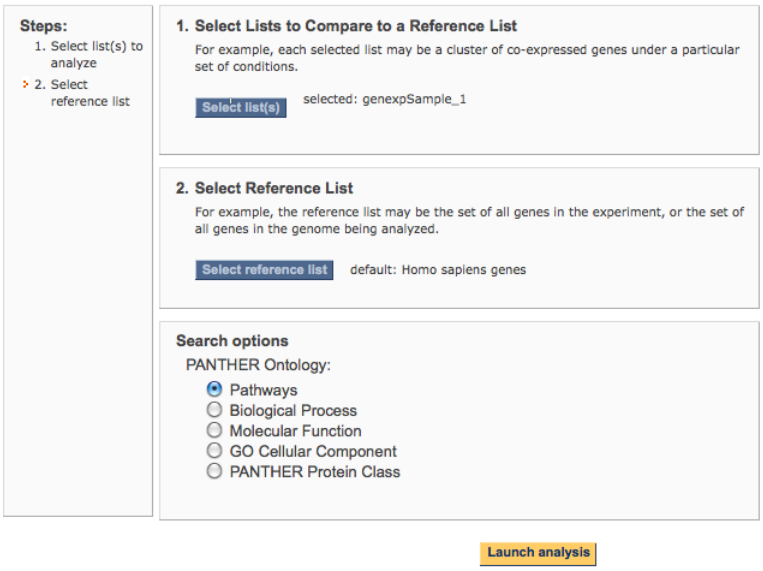


Figure 4.5: *Modify reference list if needed, and then start the analysis.*

4.1.1.3 Analysis results

On the results page, there is a table with six essential columns of data:

1. The first column contains the name of the PANTHER classification category. If you are doing this analysis in terms of pathways, you can click on the pathway name to view the corresponding pathway diagram (see section 4.1.1.4 below).
2. The second column contains the number of genes in the reference list that map to this particular PANTHER classification category.

3. The third column contains the number of genes in your uploaded list that map to this PANTHER classification category.

4. The fourth column contains the expected value, which is the number of genes you would expect in your list for this PANTHER category, based on the reference list. See below for more detailed explanation.

5. The fifth column has either a + or -. A plus sign indicates over-representation of this category in your experiment: you observed more genes than expected based on the reference list (for this category, the number of genes in your list is greater than the expected value). Conversely, a negative sign indicates under-representation.

6. The sixth column is the p-value as determined by the binomial statistic. This is the probability that the number of genes you observed in this category occurred by chance (randomly), as determined by your reference list. A small p-value indicates that the number you observed is significant and potentially interesting. A cutoff of 0.05 is recommended as a starting point. See below for more explanation.

If you upload more than one list, the 3-6 columns are repeated for each list. The result is sorted based on p-values of the first list, but you can click on any column header to sort based on that column. The yellow triangle in front of the header indicates which column is sorted, and the orientation of the triangle indicates whether the sort is ascending or descending.

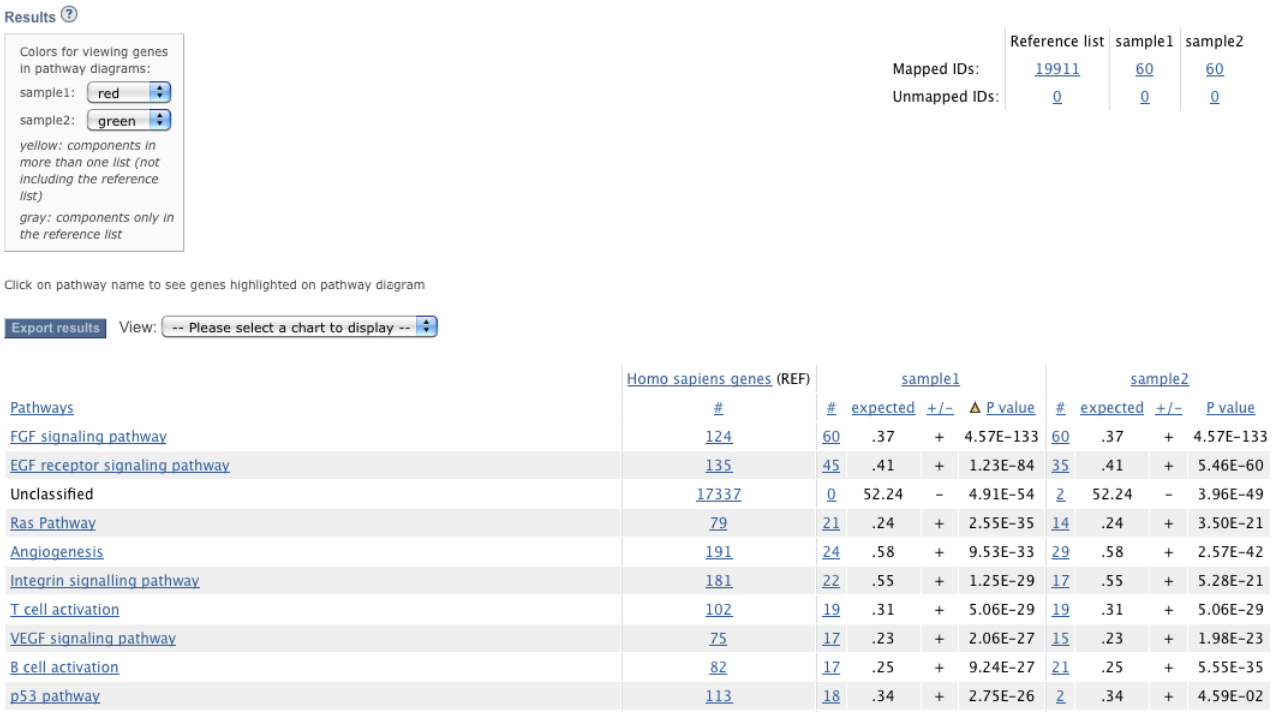


Figure 4.6: *Results of the binomial distribution analysis.*

Tip: Binomial distribution test

Expected Value

The expected value is the number of genes you would expect in your list for a particular PANTHER category, based on the reference list. As an example, say we do do the analysis in terms of biological processes, and we are interested in determining if there is over- or under-representation of chromatin packaging and remodeling genes in your uploaded list.

As an example, there are 23,481 genes in your reference list (ex: the human genome). 125 of these genes map to the 'Chromatin packaging and remodeling' category. Based on this, 0.532% (125 divided by 23481) of the genes in the reference list are involved in chromatin packaging and remodeling.

Now your upload list contains 978 genes. Based on the reference list, we would expect that 5.21 genes (978 * 0.532%) in your uploaded list would be involved in chromatin packaging and remodeling.

If for this biological process you observe more genes in your uploaded list than expected, you have an over-representation (+) of genes involved in chromatin packaging and remodeling. If you observe fewer genes than expected, you have an under-representation (-).

P-Value calculated by the Binomial statistic

In the example above we expected 5.21 genes in your uploaded list to be involved in chromatin packaging and remodeling. Say that we actually observe 7 genes involved in this biological process. This is very similar to 5.21, so you would expect a large p-value (the p-value would be 0.27). Alternatively, if 15 of your uploaded genes map to this process, this is very different than the expected p-value, so you would expect a small, significant p-value (the p-value would be 3.3E-04). This small p-value indicates that the result is non-random and potentially interesting, and worth looking at in closer detail. A p-value cutoff of 0.05 is recommended as a start point. The binomial statistic is a commonly used statistic (it can be found in any statistics book). In the binomial test we assume that under the NULL hypothesis, genes in the uploaded list are sampled from the same general population as genes from the reference set, i.e. the probability p(C) of observing a gene from a particular category C in the uploaded list is the same as in the reference list. We first estimate the probability p(C) from the reference set assuming that it is large and representative:

p(C)=n(C)/N,

where n(C) is the number of genes mapped to category C, and N is the total number of genes in the reference set.

We then use the above estimate to find the p-value: the probability of observing k(C) genes (or a more extreme number) in the uploaded list of size K. Under the NULL hypothesis, the number of genes of mapped to C is distributed binomially with probability parameter p(C) and thus the p-value would be

$$p-value = \sum \binom{K}{k} p(c)^k (1-p(c))^{K-k}$$

where the sum runs from k(C) to K in the case of over-representation (i.e. when the number of observed genes k(C) is greater than expected p(C)*K under the NULL hypothesis), and 0 to k(C), in the case of under-representation (i.e. when k(C) is smaller than p(C)*K).

When developing this analysis tool, we tested using both the Chi-Square and Binomial statistical tests. We decided to use the Binomial, since the Chi-Square is not as accurate when the population sizes or the expect number is small.

4.1.1.4 View pathway diagram

If you are analyzing pathways, you can click the pathway name in the first column and view pathway diagram with your IDs overlay on the diagram. Each list has an assigned color as shown on the top of the result page (Figure 4.6). The drop-down menu allows you to select a color of choice for each list.

Figure 4.7 shows an example of such a diagram. Red and green indicate the genes appearing in each of your list. Yellow indicates that the pathway component is in more than one of the uploaded list (not including the reference list). Gray indicates the components not in your list, but only in the reference list.

The diagram can be manipulated as described in section 2.4.2. The colored image can only be exported as a png file using the "File-¿Export image" function in the applet menu.

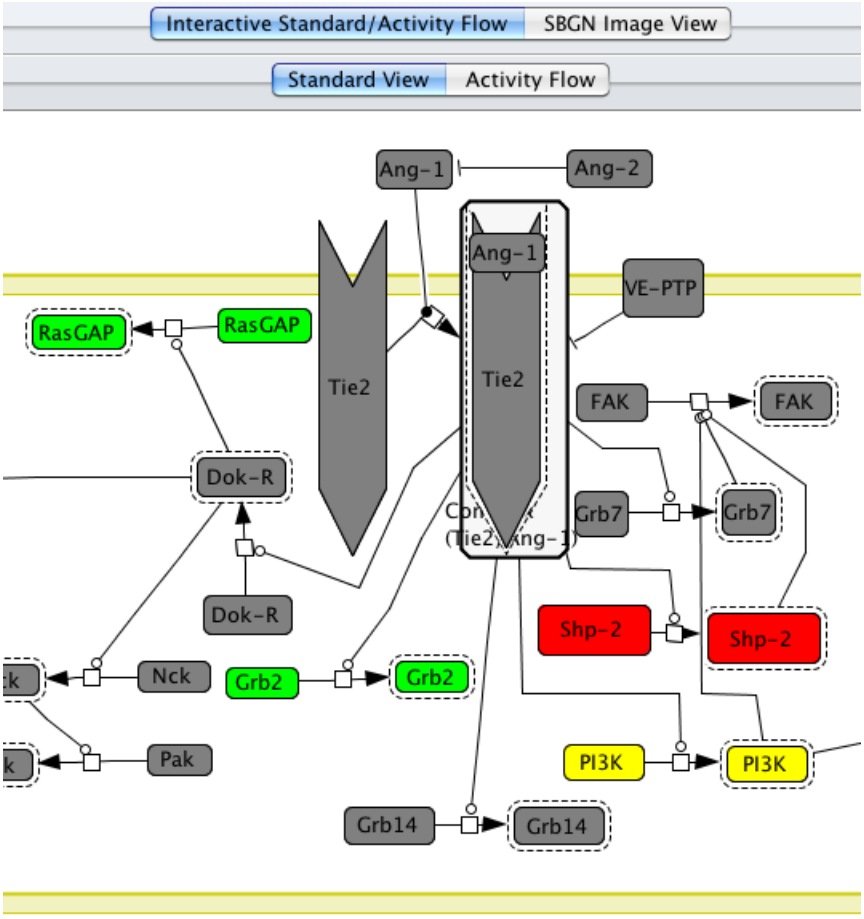


Figure 4.7: *View results in pathways. Each list id colored by a different color. A total of 4 lists can be uploaded.*

4.1.2 Analyze gene list with expression values

For each molecular function, biological process, cellular component, PANTHER protein class, or pathway term in PANTHER, the genes associated with that term are evaluated according to the likelihood that their numerical values were drawn randomly from the overall distribution of values. The Mann-Whitney U Test (Wilcoxon Rank-Sum Test) is used to determine the P-value that, say, the chromatin packaging and remodeling genes have random values relative to overall list of values that were input.

This approach has been used by our group (Clark et al., 2003) [7] and is similar to a method from Eric Lander's group (Mootha et al., 2003) [8], to find weakly coordinated shifts that elude methods based on defining strict cutoffs in the data, e.g. only focusing on genes whose expression has changed by over 1.5- or 2-fold.

For the rank-sum test, it is important to provide values for as many genes as possible (subject to noise level and reliability) so that randomness can be properly assessed across the experiment. In most cases, you should upload the entire list of the expression experiment.

4.1.2.1 Upload gene expression file

When you enter the tool, you can upload your gene expression file from your computer to the system using the interface as shown in Figure 4.8. The uploaded file must be a tab-delimited text file, and must contain an identifier to be analyzed, and the corresponding numerical value. You must select the file type, to specify what type of identifier (ID) you are uploading.

If your file uses one of the supported IDs, you select "Gene, transcript, protein and alternative ID" option, and the file must contain tow columns. The first column is the ID and the second is the numerical value.

If you use the **PANTHER Generic Mapping File**, the file should contain 3 columns.

- The first column can contain any arbitrary ID, but most be unique, since this allows the user to uniquely specify each record in the dataset, so they can track the identifier on the PANTHER website.
- The second column should be corresponding PANTHER ID (ex: PTHR10078 or PTHR10078:SF1), and is used to look up the molecular function, biological process, and pathway associations.
- The third column must be the corresponding numerical value

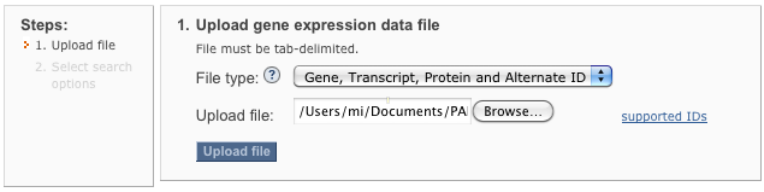


Figure 4.8: *Upload a gene list with expression values.*

After you select the file, click the "Upload file" button. On the next page, you can select the ontology or pathway to analyze (Figure 4.9).

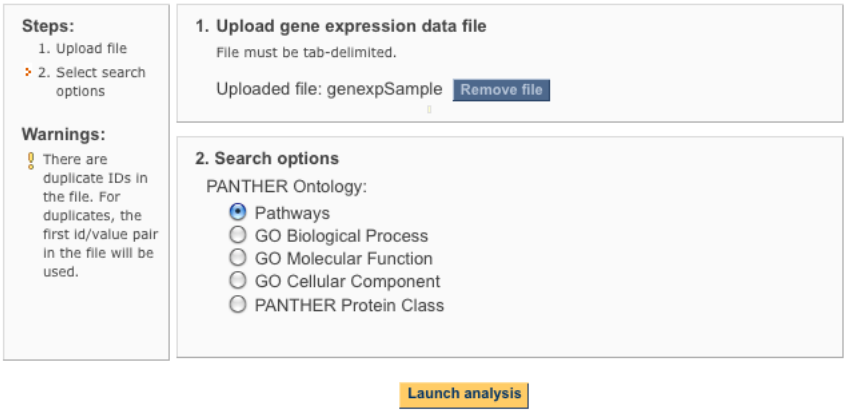


Figure 4.9: *Select a PANTHER ontology or pathway to analyze.*

4.1.2.2 Results

On the results page, there is a table with four essential columns of data:

1. The first column contains the name of the PANTHER classification category. If you are doing this analysis in terms of pathways, you can click on the pathway name to view the pathway diagram (see below). The genes in the pathway diagram are colored according to the gene expression value, and the rules for this can be specified by clicking on the 'Specify color ranges' button.

2. The second column contains the number of genes that map to this particular PANTHER classification category.

3. The third column has either a + or -. A plus sign indicates that for this category, the distribution of values for your uploaded list is shifted towards greater values than the overall distribution of all genes that were uploaded. A negative sign indicates that the uploaded list is shifted towards smaller values than the overall list.

4. The fourth column contains the p-value as calculated from the Mann-Whitney U Test (Wilcoxon Rank-Sum test). A large p-value indicates that the genes for this category have a distribution that is similar to randomly choosing genes from the overall distribution. In other words, the values of the uploaded genes for this category have a similar distribution to the overall list of values that were input. A small, significant p-value indicates that the distribution for this category is non-random and different than the overall distribution. A cutoff of 0.05 is recommended as a starting point.

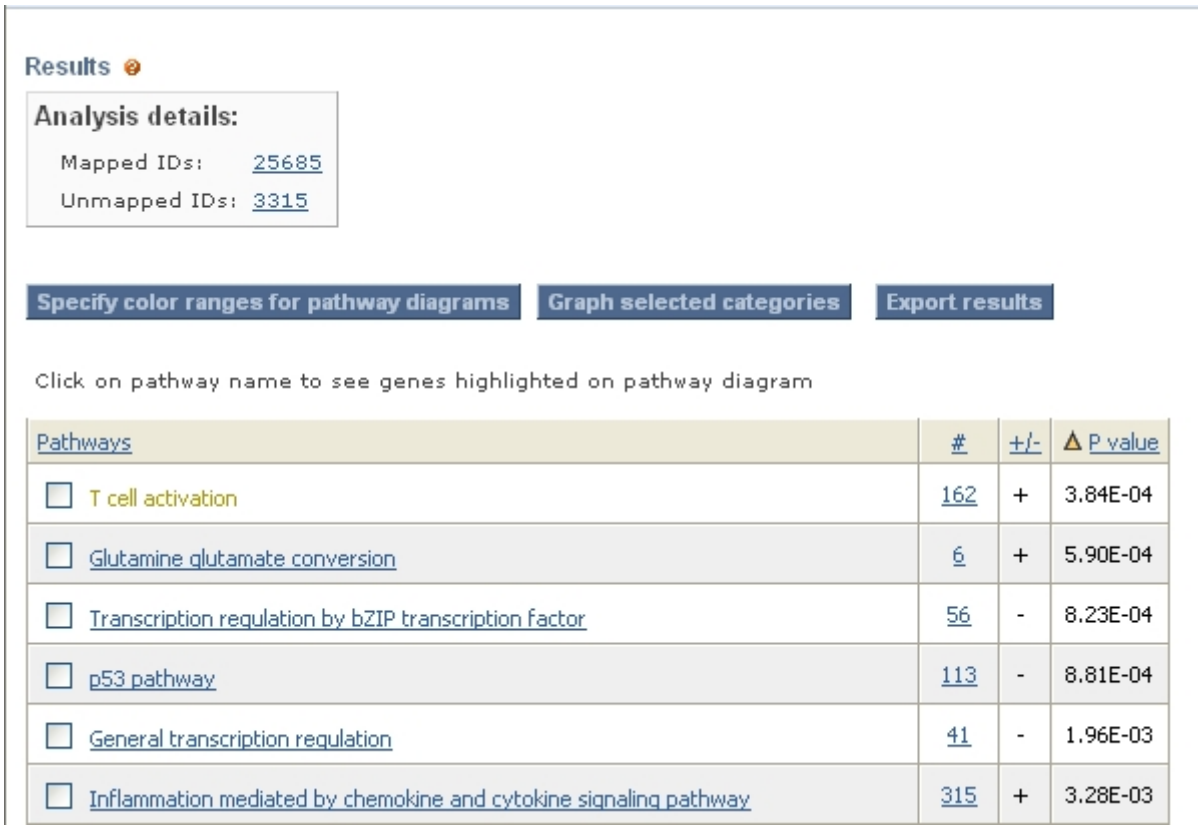


Figure 4.10: *The Mann-Whitney test result page..*

To have a visual representation of these distributions, select the checkboxes of the categories of interest, and click on the "Graph selected categories" button near the top of the page. Figure 4.11 shows an example of the graph. The x-axis is your uploaded value. The y-axis is the

cumulative fraction. In other words, if you look at the data point x=1.4, y=.75 this means that 75% of your uploaded values have a value of 1.4 or smaller.

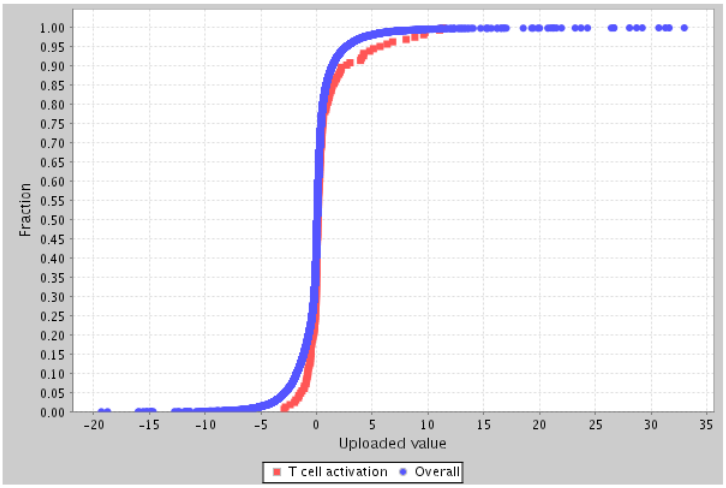


Figure 4.11: *Mann-Whitney test results shown in graph view.*

The Mann-Whitney U Test (Wilcoxon Rank-Sum test) Statistic

To perform the rank sum test, first the values of the genes that map to a given category are combined with the overall list of values that were input. Then, all the values are ranked from smallest to largest, with the smallest value getting a rank of 1. If multiple values are identical, the average of the ranks for these values is used.

Then the rank sum is calculated for this category, by summing up the ranks for all of the genes that map to this category. The average rank, R1 is then calculated by dividing the rank sum by the number of genes, n1, that map to the category. Likewise, the rank sum is calculated for the list of all IDs uploaded, and the average rank, R2, is calculated by dividing the rank sum by the total number of genes uploaded, n2.

Next, the Mann Whitney U statistic is calculated for both populations:

U1 = n1* n2 + (n1 * (n1 + 1)) / 2 - R1
U2 = n2* n2 + (n1 * (n2 + 1)) / 2 - R2

The larger of these two values is the Mann Whitney U-statistic, U, whose distribution for small sample sizes can be found in most statistic books. In our case, our application is for large sample sizes, so we use the normal approximation:

Z-score = (U- (n1* n2)/2)/sqrt(n1*n2*(n1+n2+1)/12).

It follows that the p-value is the integral under the standard normal density.

4.1.2.3 View results in pathway diagram

If you select to analyze the data in PANTHER pathway, you can click the pathway name from the column 1 of the result page, and view the results in pathway diagram in "head map" (Figure 4.12).

The color range reflects the expression value of your genes mapped to the pathway component. You can define the color range by selecting "Specify color ranges for pathway diagram" button on the result page (Figure 4.10). First there are two color ranges to select. Select equal

distributions (default), and the ranges will be determined by sorting your gene values, and then splitting the values into six equal distributions. Select fold changes, and standard gene expression fold change ranges will be used.

Some pathway components have multiple genes mapped to them. You can she color ranges for these components that are determined by

- median of the gene values *(default)*
- average of the gene values
- maximum gene value
- minimum gene value

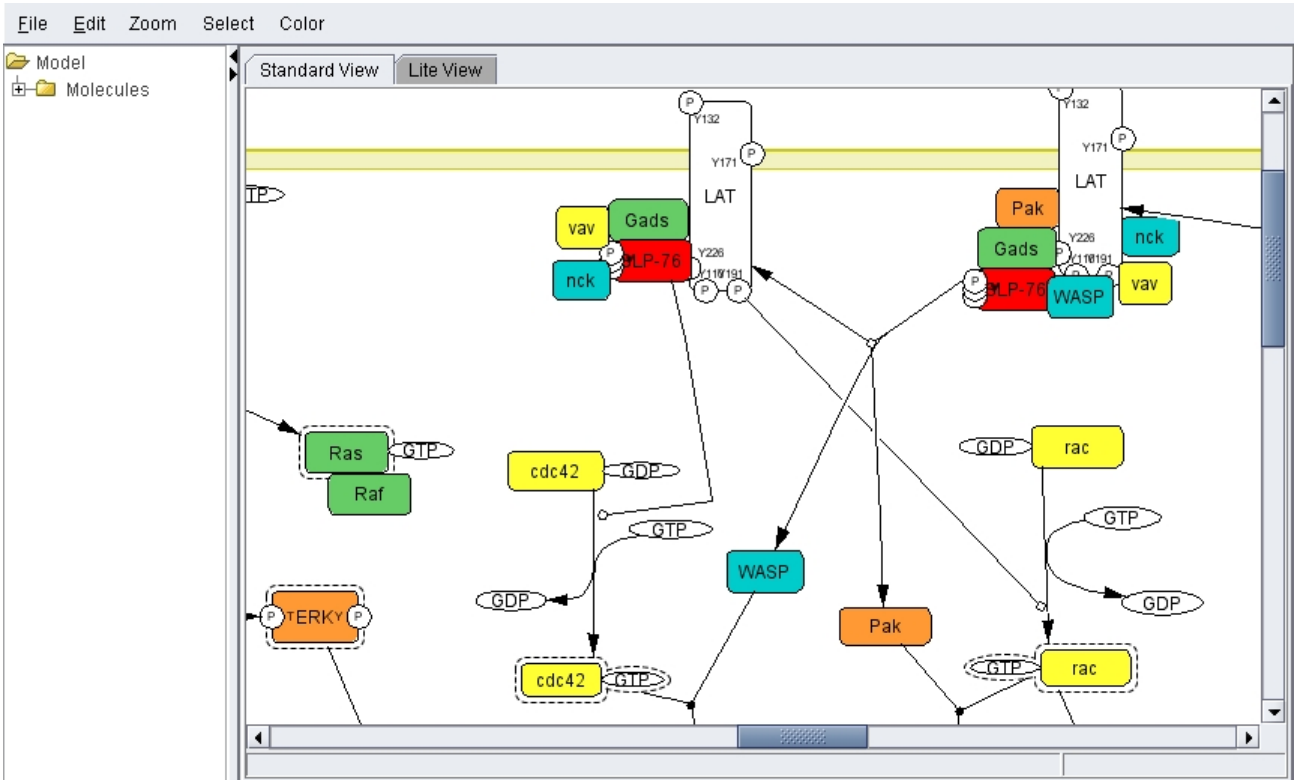


Figure 4.12: *Mann-Whitney test results shown in pathway diagram in a "head map".*

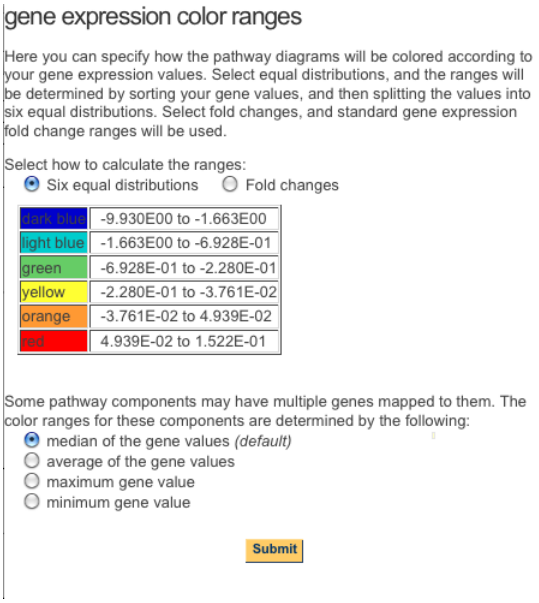


Figure 4.13: *Color code for the head map.*

4.2 Evolutionary Analysis of Coding SNPs

This tool estimates the likelihood of a particular nonsynonymous (amino-acid changing) coding SNP to cause a functional impact on the protein. It calculates the subPSEC (substitution position-specific evolutionary conservation) score based on an alignment of evolutionarily related proteins, as described in Brunham et al., 2005 [9], Thomas et al., 2003 [10] and Thomas and Kejariwal, 2004 [11].

4.2.1 Input protein and substitution data

Paste the wildtype protein sequence in the first box (FASTA format, or just the protein sequence itself), and then enter the substitution(s) relative to this input sequence in the standard amino acid substitution format, e.g. A265V. Multiple substitutions should be separated by a tab, space, or return (Figure 4.14).

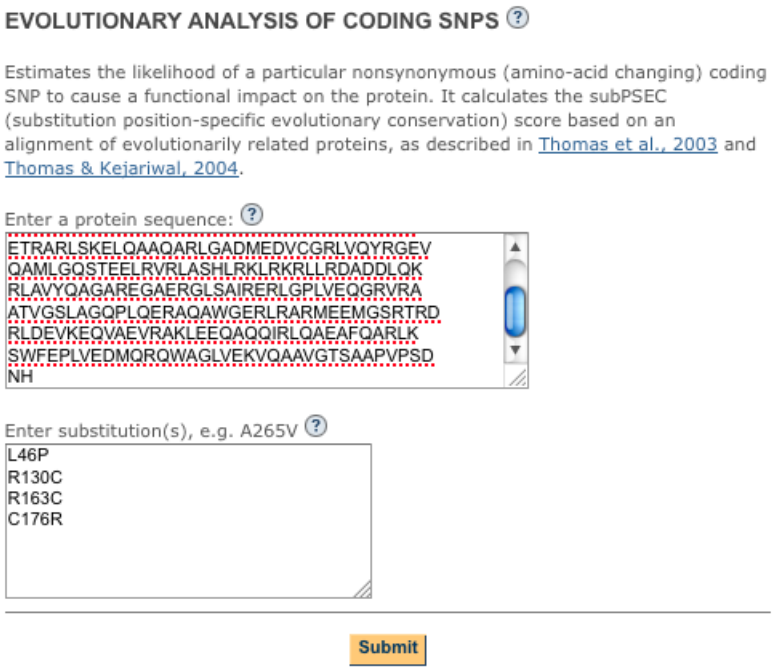


Figure 4.14: *Enter sequence and SNP substitutions.*

Note that the sequence numbering must be correct relative to the sequence pasted in the box and not necessarily the numbering found in the literature. For example, APOE4 is listed as containing C112R in the literature, but this requires removing the signal peptide from the SWISS-PROT entry for APE_HUMAN, or entering the substitution as C130R.

4.2.2 Results of cSNP analysis tool

4.2.2.1 Protein score

The input protein sequence is scored against the HMMs in the PANTHER library. The alignment to the HMM with the most significant score is used for the analysis. Proteins that scored greater than 1e-23 are excluded from the analysis, as the alignments are less reliable.

4.2.2.2 subPSEC score and Pdeleterious

The subPSEC (substitution position-specific evolutionary conservation) score estimates the likelihood of a functional effect from a single amino acid substitution. It is the negative logarithm of the probability ratio of the wild-type and mutant amino acids at a particular position. PANTHER subPSEC scores are continuous values from 0 (neutral) to about -10 (most likely to be deleterious). -3 is the previously identified cutoff point for functional significance.

A cutoff of -3 corresponds to a 50% probability that a score is deleterious. From this, the probability that a given variant will cause a deleterious effect on protein function is estimated by Pdeleterious, such that a subPSEC score of -3 corresponds to a Pdeleterious of 0.5.

For more details and to see the corresponding equations, please see our most recent publication Brunham et al., 2005 [9]. Additional information is in Thomas et al., 2003 [10] and Thomas and Kejariwal, 2004 [11].

4.2.2.3 Multiple sequence alignment

Click on the link on the number of the multiple sequence alignment (MSA) position to view the column in the MSA where the substitution occurs. The column is highlighted in red, and the subfamilies (rows in the MSA) that were used to gather the statistics for amino acid probabilities

are highlighted in blue. In the MSA view, click on a blue diamond (a subfamily node) in the tree to hide or show sequences in the subfamily.

4.2.2.4 Number of independent counts (NIC)

NIC (number of independent counts) is an estimate of the number of independent observations used to calculate the amino acid probabilities. The probabilities are calculated from a combination of prior knowledge (e.g. that isoleucine often substitutes for valine) and observations, so the larger NIC, the more the probabilities rely on the amino acids observed in the multiple sequence alignment. position does not align to the HMM If the substitution occurs at a position that does not appear in the multiple sequence alignment; a subPSEC score cannot be generated and the output will return the text string "position does not align to the HMM", indicating that the substitution occurs at a position that is inserted relative to the consensus HMM for the given HMM. In most cases, these positions are not modeled by the HMMs simply because they do not appear in most of the related sequences; as a result, substitutions at inserted positions are not generally likely to be deleterious.

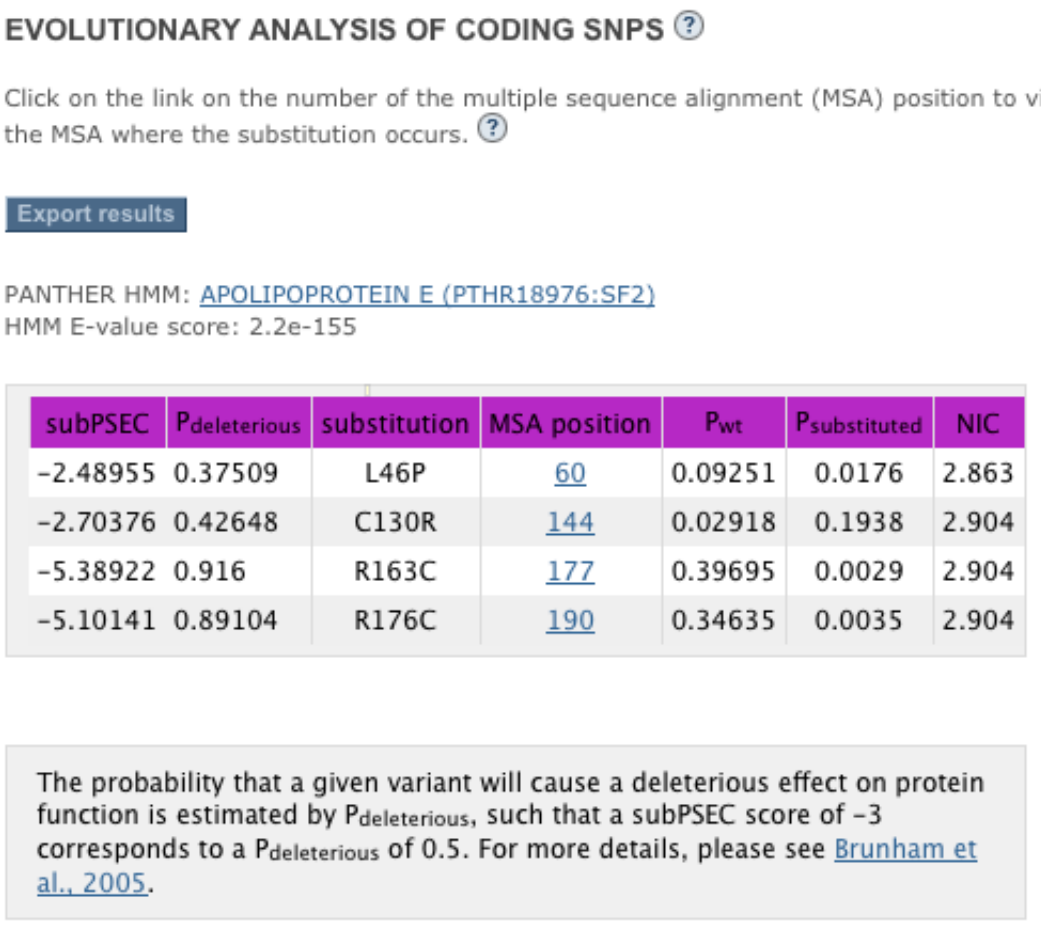


Figure 4.15: *cSNP analysis results page.*

Appendix A

Systems requirements

- PC Users (Recommended):
  - Windows 2000, or Windows XP
  - Microsoft Internet Explorer 6.x
  - Java version 1.4.2 is required
- Macintosh Users:
  - A G4 Macinosh with MacOS 10.3.7 is required
  - Safari version 1.2.4
  - Java version 1.4.2 is required
- For both:
  - Minimum of 128 MB RAM, 256MB RAM recommended
  - JavaScript, Java applets and cookies must be enabled in your browser
  - Java applet runtime parameters set to -ms128m -mx512m -Xss16m
  - Screen resolution of at least 800x600 dpi is strongly recommended

Bibliography

[1] Huaiyu Mi, Qing Dong, Anushya Muruganujan, Pascale Gaudet, Suzanna Lewis, and Paul D. Thomas. Panther version 7: improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium. *Nucleic Acids Res*, 38(Database issue):D204–10, 1 2010.

[2] Huaiyu Mi and Paul Thomas. Panther pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol*, 563:123–40, 2009.

[3] Paul D. Thomas, Anish Kejariwal, Nan Guo, Huaiyu Mi, Michael J. Campbell, Anushya Muruganujan, and Betty Lazareva-Ulitsky. Applications for protein sequence-function evolution data: mrna/protein expression analysis and coding snp scoring tools. *Nucleic Acids Res*, 34(Web Server issue):W645–50, 7 2006.

[4] Paul D. Thomas. Giga: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics*, 11:312, 2010.

[5] K. Katoh, K. Kuma, H. Toh, and T. Miyata. Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–8, 2005.

[6] Raymond J. Cho and Michael J. Campbell. Transcription, genomes, function. *Trends in Genetics*, 16(9):409 – 415, 2000.

[7] A. G. Clark, S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, 302(5652):1960–3, 12 2003.

[8] Vamsi K. Mootha, Cecilia M. Lindgren, Karl-Fredrik F. Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J. Daly, Nick Patterson, Jill P. Mesirov, Todd R. Golub, Pablo Tamayo, Bruce Spiegelman, Eric S. Lander, Joel N. Hirschhorn, David Altshuler, and Leif C. Groop. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–73, 7 2003.

[9] L. R. Brunham, R. R. Singaraja, T. D. Pape, A. Kejariwal, P. D. Thomas, and M. R. Hayden. Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the abca1 gene. *PLoS Genet*, 1(6):e83, 12 2005.

[10] Paul D. Thomas, Michael J. Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–41, 9 2003.

[11] Paul D. Thomas and Anish Kejariwal. Coding single-nucleotide polymorphisms associated with complex vs. mendelian disease: evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci U S A*, 101(43):15398–403, 10 2004.

Index