Published online 28 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D183–D189***
doi: 10.1093/nar/gkw1138

PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements

Huaiyu Mi*, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang and Paul D. Thomas*

Division of Bioinformatics, Department of Preventive Medicine, Keck School of Medicine of USC, University of Southern California, Los Angeles, CA 90033, USA

Received October 05, 2016; Revised October 27, 2016; Editorial Decision October 28, 2016; Accepted November 16, 2016

ABSTRACT

The PANTHER database (Protein ANalysis THrough Evolutionary Relationships, http://pantherdb.org) contains comprehensive information on the evolution and function of protein-coding genes from 104 completely sequenced genomes. PANTHER software tools allow users to classify new protein sequences, and to analyze gene lists obtained from large-scale genomics experiments. In the past year, major improvements include a large expansion of classification information available in PANTHER, as well as significant enhancements to the analysis tools. Protein subfamily functional classifications have more than doubled due to progress of the Gene Ontology Phylogenetic Annotation Project. For human genes (as well as a few other organisms), PANTHER now also supports enrichment analysis using pathway classifications from the Reactome resource. The gene list enrichment tools include a new 'hierarchical view' of results, enabling users to leverage the structure of the classifications/ontologies; the tools also allow users to upload genetic variant data directly, rather than requiring prior conversion to a gene list. The updated coding single-nucleotide polymorphisms (SNP) scoring tool uses an improved algorithm. The hidden Markov model (HMM) search tools now use HMMER3, dramatically reducing search times and improving accuracy of E-value statistics. Finally, the PANTHER Tree-Attribute Viewer has been implemented in JavaScript, with new views for exploring protein sequence evolution.

INTRODUCTION

Protein ANalysis THrough Evolutionary Relationships (PANTHER) is a multifaceted data resource for classification of protein sequences by evolutionary history, and by function. Protein-coding genes from 104 organisms are classified by evolutionary relationships, and by structured representations of protein function including the Gene Ontology (GO) and biological pathways. The foundation of PANTHER is a comprehensive 'library' of phylogenetic trees of protein-coding gene families. These trees attempt to reconstruct the evolutionary events (speciation, gene duplication and horizontal gene transfer) that led to the modern-day family members. The trees are used to predict **orthologs** (genes that diverged via a speciation event), paralogs (genes that diverged via a duplication event) and xenologs (genes that diverged via horizontal transfer). Protein **subfamilies**, groups of proteins that are generally closely-related orthologs (see (1) for details) are also identified in the trees. A **hidden Markov model (HMM)** is constructed for each family and subfamily.

Perhaps most importantly, the trees enable inferences to be made about the **functions of genes**. Currently, these inferences are made by expert biocurators (2), as part of the Gene Ontology Phylogenetic Annotation project. In this project, a biocurator reviews the experimentally-supported annotations for all genes in a gene family, and constructs a parsimonious model of function (gene ontology (GO) term) gain and loss at specific branches in a phylogenetic tree (2). This model **predicts the functions of each uncharacterized gene** in the tree through inheritance from its ancestors. In addition to GO terms, PANTHER also provides curated associations of genes to biological pathways from the PANTHER Pathway resource (3).

The PANTHER website contains several data analysis tools that make use of the underlying PANTHER data (4). The **HMM search tool**, available both on the web (for small numbers of sequences) and as a downloadable soft-

[*]To whom correspondence should be addressed. Tel: +1 323 442 7975; Fax: +1 323 442 7995; Email: pdthomas@usc.edu
Correspondence may also be addressed to Huaiyu Mi. Tel: +1 323 442 7994; Fax: +1 323 442 7995; Email: huaiyumi@usc.edu

ware package (for large batches), rapidly compares new sequences to the existing families and subfamilies in PANTHER. If a new sequence is a statistical match to one of these models, it is classified as a member of that group and can be inferred to share the known properties of that group. The **coding single-nucleotide polymorphism (SNP) scoring tool**, also available on the web as well as for download, uses the family multiple sequence alignments and phylogenetic trees in PANTHER to identify whether a user-uploaded protein amino acid substitution will likely impact protein function (5–7). The **gene set overrepresentation and enrichment tools** compare a user-uploaded gene list to the functional classifications in PANTHER (GO and pathways), to help identify the shared biological functions among genes in the list (8).

PANTHER is tightly integrated with a number of other genomic resources. It is a member in the **InterPro Consortium** of protein classification resources (9); PANTHER HMMs can be searched from the InterPro website and using the InterProScan tool. It is a founding member of the **Quest for Orthologs Consortium** (10) that seeks to advance and evaluate ortholog prediction methods. PANTHER also makes use of the **UniProt Reference Proteome** data sets (11), the emerging standard for protein-coding gene sets across an increasing number of fully sequenced genomes. Recently, the PANTHER overrepresentation/enrichment tools were extended to include functional annotations directly downloaded from the **GO Consortium**, in addition to the phylogenetically inferred annotations (1,12).

Here, we describe the developments to the PANTHER database in the past year, building upon our previous work. We note that while the usage of the PANTHER website has steadily increased for over 10 years, we saw a surge in 2016. There was a 25% increase in users, an almost 50% increase in sessions and a 60% increase in page views through September 2016 compared to the same period during the previous year.

PANTHER DATA UPDATES

Updated sequences and families

PANTHER version 11 was released in July of 2016. The set of complete protein coding genes for 104 genomes was updated to the 2015 release of the UniProt Reference Proteomes. The coverage of protein coding genes varies from over 95% for most mammalian genomes (including human), to less than 40% for the parasitic protozoans (*Trichomonas vaginalis* and *Giardia intestinalis*). Full statistics are available at http://pantherdb.org/panther/summaryStats.jsp. In order to facilitate browsing and visualization of trees without negatively impacting the power to infer function from homology, some of the larger protein families that contain multiple ancestral genes dating back to the last universal common ancestor (LUCA) were divided into separate families each dating back to LUCA. A total of 542 families from PANTHER version 10 (∼5% of the total 11 819 families in version 10) were divided into 1819 families in PANTHER version 11. An example is the very diverse thiamine pyrophosphate (TPP) enzyme family (PTHR18968 in version 10). This family was divided into 8 families in version 11. The largest of these, acetolactate synthase large subunit, retains the identifier (PTHR18968 in version 11), while the other new families were given new family identifiers: 2-hydroxyacyl-coA lyase 1 (PTHR43710), pyruvate decarboxylase 1 (PTHR43452), protein Phyllo (PTHR42916), 2-oxoglutarate oxidoreductase beta subunit (PTHR43474), phosphonopyruvate decarboxylase (PTHR42818), 2-oxoacid:ferredoxin oxidoreductases (PTHR43838) and pyruvate dehydrogenase (ubiquinone) (PTHR42981). These families, while detectably homologous at the sequence level, likely diverged prior to LUCA and not surprisingly are also divergent in function.

Expanded homology-inferred gene functions from the GO phylogenetic annotation project

The Gene Ontology Phylogenetic Annotation Project has been annotating PANTHER trees with GO terms since 2011, and is now the single largest contributor of manually-reviewed GO annotations in the GO Consortium (12). The GO Phylogenetic Annotations for individual genes are also available directly from the GO website, but they are used differently on the PANTHER website, resulting in a larger number of inferred annotations. On the GO website, Phylogenetic Annotations are only included if they do not overlap with any experimental GO annotations for the same gene. At PANTHER, on the other hand, the GO Phylogenetic annotations are used to annotate PANTHER subfamilies and families. Thus they are **added to** the previously curated PANTHER (sub)family GO annotation sets (referred to as 'PANTHER GO-slim' sets on the website). As a result, the PANTHER GO-slim annotations include **all inferred annotations** from the GO Phylogenetic Annotation project; importantly, these annotations have passed an additional expert review process beyond the standard GO experimental annotation process. Furthermore, the application of these annotations to families and subfamilies enables functional annotation of any genome, by applying the PANTHER HMM scoring tools. Table 1 shows the number of genes (and total annotations) currently covered by PANTHER GO-slim annotations, for a sample of genomes. It also shows the number of new annotations added from the GO Phylogenetic Annotation project. The GO Phylogenetic Annotation project now accounts for about half of the homology-inferred annotations in PANTHER.

Family and subfamily HMMs constructed using HMMER3

HMMER is a software package for sequence analysis using profile hidden Markov models . HMMER is more accurate for detecting remote homologs than BLAST but the previous version (HMMER2) runs about 100X slower than comparable BLAST searches. HMMER3 is essentially as fast as BLAST for protein sequence searches (13). In PANTHER version 11, all family and subfamily HMMs have been updated to HMMER3. Specifically, the multiple sequence alignment (MSAs) for each family was first built with MAFFT (14), as in PANTHER versions 7 through 10. To support HMMER3, we added a step to replicate how the MSAs were 'trimmed' in previous versions of PANTHER. The trimming restricts the initial HMM topology to

Table 1. Coverage of version 11.1 PANTHER GO-slim functional annotations for selected genomes

Organism	Total number of genes in genome	Total number of genes annotated	Total number of annotations	Total number of genes with new annotations from GO Phylo-project	Total number of new annotations from GO Phylo-project
Human	20 972	13 387	57 271	5925	28 996
Mouse	22 322	14 331	61 828	6804	33 318
Rat	23 781	15 536	67 112	7642	37 378
Pig	21 398	13 469	58 475	6266	31 080
Dog	19 692	12 715	54 307	5727	27 629
Chicken	15 789	10 056	43 605	4114	20 524
Zebrafish	27 187	17 113	73 161	6904	33 578
Fruit fly (*D. melanogaster*)	14 217	6942	26 408	3128	13 452
Nematode (*C. elegans*)	20 472	8308	31 072	3781	16 052
Slime mold (*D.discoideum*)	13 044	5133	19 582	2267	10 121
Budding yeast (*S. cerevisiae*)	6728	3372	13 364	1796	8363
Fission yeast (*S. pombe*)	5140	3110	12 306	1610	7533
C. albicans, fungus	9329	3489	13 455	1780	8146
A. thaliana, plant	27 352	12 130	44 243	6119	24 204
Rice	62 904	17 873	61 811	9307	34 673
Tomato	34 783	13 630	48 464	6820	26 737
E. coli	4262	1990	6267	705	2249
B. subtilis	4199	1702	5452	546	1733

The total number of genes with at least one functional annotation (across molecular function, biological process and cellular component aspects) is listed, followed by the number of genes/annotations added by the GO Phylogenetic Annotation project. Detailed statistics, such as coverage of each aspect, for all genomes in PANTHER can be found at http://pantherdb.org/panther/summaryStats.jsp.

include as match states only those alignment columns that have only a small proportion of sequences in the alignment showing deletions. In this step, each MSA was converted to Stockholm format, and then the sequence masking feature of HMMER3 was used to exclude highly deleted columns. It was found that the masking of the MSA significantly improved the HMM accuracy as judged by self-scoring, described below. The masked MSAs were then input into the *hmmbuild* program with default settings to build HMMs in HMMER3 format.

In order to assess the accuracy and performance of the models, a self-scoring procedure was performed. Briefly, all sequences in PANTHER were scored against the entire HMM library, and the percentage of sequences that hit their own HMM as the best hit (in competition with other HMM models) was assessed. Approximately 90% family models and ∼80% subfamily models have at least 95% of their sequences with their own HMMs as the best hit. These results are similar to those obtained with the HMM models built in previous versions of PANTHER with SAM program (15) and then converted to HMMER2 format, but with a significant gain in computational speed with HMMER3. The remaining 10% of families (those with relatively poor self-scoring performance) are generally large, divergent families with only a small shared, homologous domain, such as leucine-rich repeat domains or zinc-finger domains. We plan to divide these divergent families into smaller families of more closely related genes, in our next release of PANTHER.

ANALYSIS AND VISUALIZATION TOOL ENHANCEMENTS

PANTHER HMM scoring tool in HMMER3

The PANTHER HMM scoring tool (pantherScore) was updated to support HMMER3. It provides options for users to choose either *hmmscan* or *hmmsearch* programs that are preferred to score small and large number of sequences, respectively, against the PANTHER library. Because of the time required for HMMER2 searches, the previous version of pantherScore required a heuristic preprocessing step that searched HMM-derived consensus sequences using BLAST, to select a subset of HMMs for HMMER2 scoring. Because of the increased computational speed of HMMER3 programs, the new pantherScore tool directly scores each query sequence against all family and subfamily HMMs, preventing any loss of sensitivity due to preprocessing. The software package can be downloaded from ftp://ftp.pantherdb.org//hmm_scoring/current_release. HMMER3 models and scoring are now integrated into the InterPro resource as well (9).

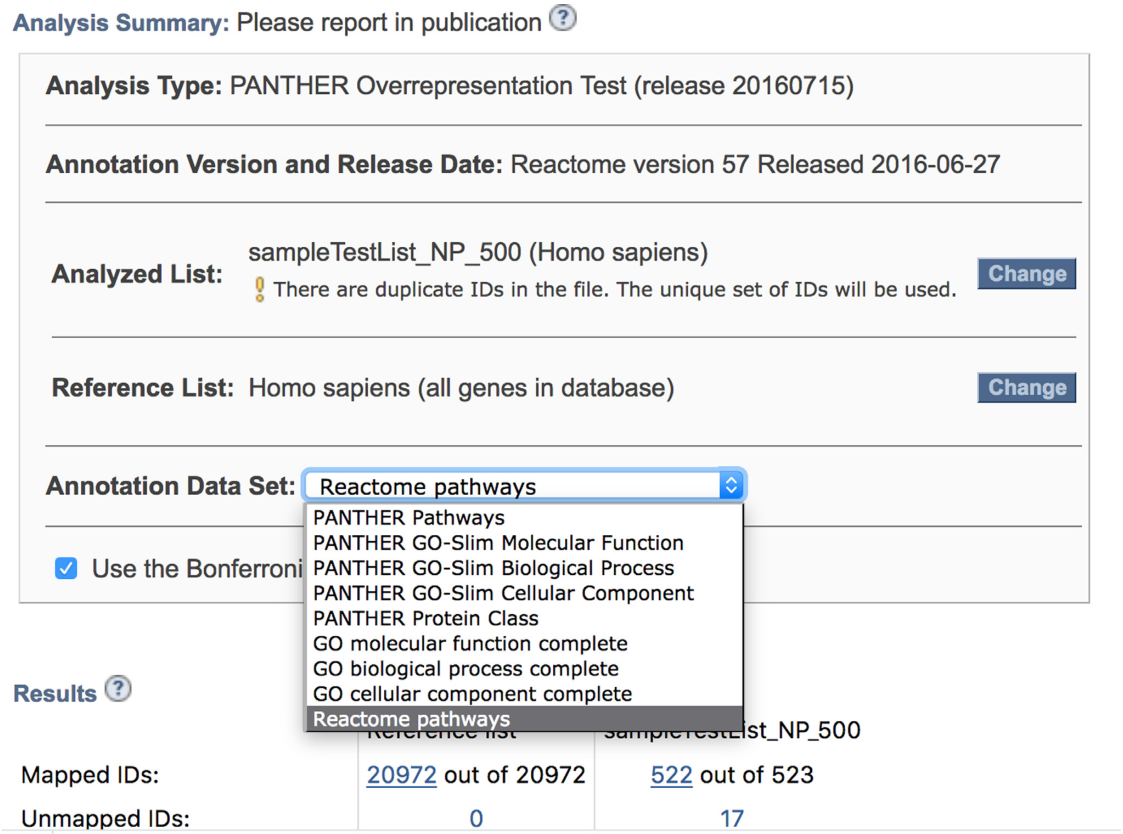
New coding SNP analysis tool: PSEP

The PANTHER website has offered a coding SNP analysis tool since 2003 using the ‘subPSEC’ (substitution position-specific evolutionary conservation) score (5,6). This tool uses the match between an uploaded missense variant and the aligned HMM match state probabilities, to estimate whether a substitution is likely to affect protein function: The worse the match, the higher the probability of a functional effect. However, this tool has not been updated since 2009 due to some large differences between PANTHER version 6.1 and version 7.0. In the interim, we have developed a new coding variant tool, that we call PSEP (position-specific evolutionary preservation). The algorithm is described in detail elsewhere (7). Briefly, it reconstructs probabilistic ancestral sequences for all nodes in a phylogenetic tree, and then traces back from the modern-day sequence to determine how long the modern amino acid has been ‘preserved’ in its ancestors. The longer the preservation time, the greater chance that a coding variant at that position will impact protein function. We have shown that PSEP outperforms subPSEC, as well as a variety of other methods, on standard benchmarks (7). The PANTHER website now has a PSEP scoring tool available either from the web interface or as a downloadable software package. Notably, the PSEP method is able to score a much larger number of variants than the older, subPSEC method (Table 2). This is

Table 2. Coverage of non-synonymous variants scored by PANTHER PSEP, compared to the older subPSEC method

	Varibench (non-pathologic)	variBench (pathologic)	swissVar (non-pathologic)	swissVar (pathologic)
Total number of variants	23 683	19 335	41 165	30 074
Variants scored by PSEP (current version)	18 492	18 655	38 307	28 814
Variants scored by PSEC (previous version)	13 910	15 710	27 340	24 405

Coverage is improved for all categories of variants, for two standard variant benchmarking sets, VariBench (17) and SwissVar (18).



Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (release 20160715)

Annotation Version and Release Date: Reactome version 57 Released 2016-06-27

Analyzed List: sampleTestList_NP_500 (Homo sapiens)

There are duplicate IDs in the file. The unique set of IDs will be used.

Reference List: Homo sapiens (all genes in database)

Annotation Data Set: Reactome pathways

☑ Use the Bonferroni

?
Change
Change
PANTHER Pathways
PANTHER GO-Slim Molecular Function
PANTHER GO-Slim Biological Process
PANTHER GO-Slim Cellular Component
PANTHER Protein Class
GO molecular function complete
GO biological process complete
GO cellular component complete
Reactome pathways
Results

	Reference list	sampleTestList_NP_500
Mapped IDs:	20972 out of 20972	522 out of 523
Unmapped IDs:	0	17

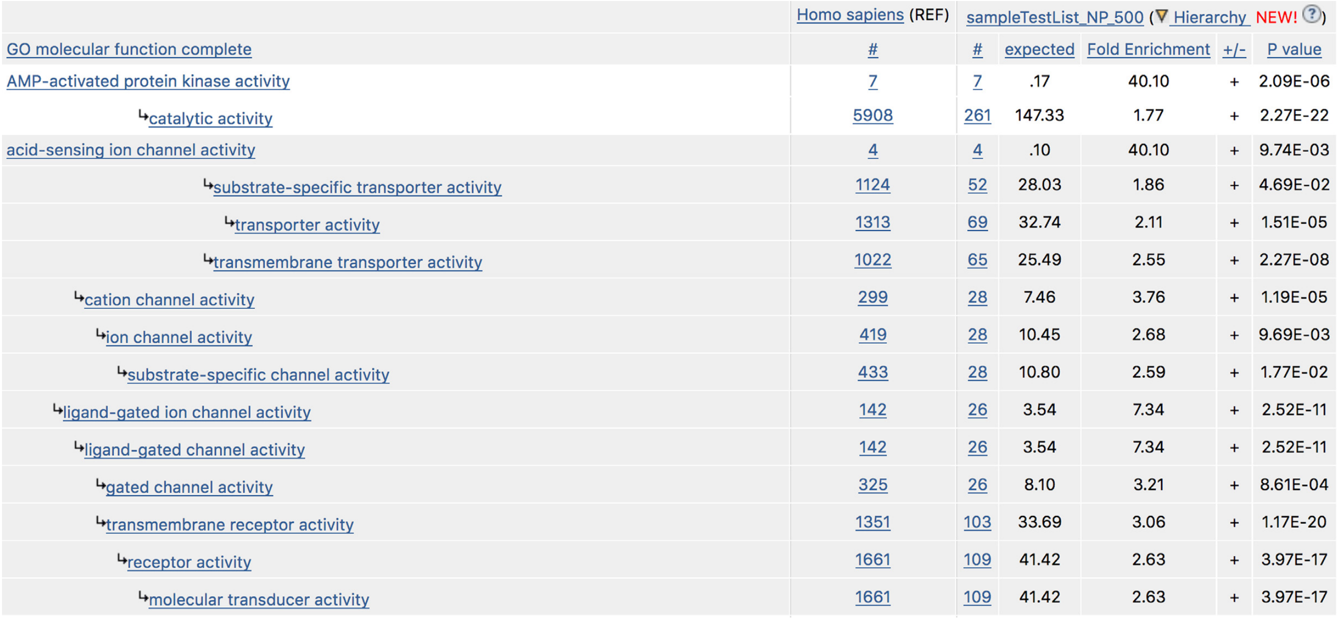
Figure 1. Reactome pathways in the PANTHER gene list analysis tool. Users can select Reactome pathways from a drop-down list of annotation sets. They can also upload a custom reference list for calculating overrepresentation by clicking on the 'Change' button. The Reactome pathways are accessible to both statistical test types in PANTHER: the overrepresentation tool (shown here) that takes a user-uploaded gene list as input, and the enrichment tool (not shown), that takes a user-uploaded list of genes, each of which is associated with a value, such as log-fold-change in expression or association *P*-value.

because PSEP considers the entire multiple sequence alignment, rather than only those sites that align to match states of the HMM.

Gene list analysis tool enhancements

Support for Reactome pathways. Reactome is a manually curated pathway database that is freely available to all users (16). All pathways have been authored by expert biologists and peer-reviewed. It is widely used by the research community. Reactome has now been added to the suite of data sets available for use with the PANTHER overrepresentation and enrichment tests. It can be selected from the 'Annotation Data Set' drop-down in the Analysis Summary Box on and PANTHER analysis results page (Figure 1). In the results page, all links to Reactome pathway classes link directly back to the Reactome resource for further information. The tests at the PANTHER website differ from that currently available on the Reactome site, in two ways: (i) PANTHER supports both overrepresentation and enrichment tests (the Reactome site only supports overrepresentation), and (ii) PANTHER allows users to upload a custom reference list, against which to calculate over/under-representation. The latter is an important feature, as it is recommended that the reference list include all genes that could have been observed, which can be very different than considering all genes in a genome (the default reference at PANTHER and most other tools).

The Reactome pathway data were downloaded from the Reactome website (http://www.reactome.org/pages/download-data/). There are two main types of data that were processed and integrated in PANTHER. The first is the Reactome pathway hierarchy. Reactome pathways are organized in a hierarchical structure in which a particular pathway can belong to a larger pathway, or set of pathways. For example, a RAF/MAP kinase cascade' (R-HSA-5673001.2) belongs to a more general 'MAPK1/MAPK3 signaling' (R-HSA-5684996.1) pathway. To capture and



	Homo sapiens (REF)	sampleTestList_NP_500 (▼ Hierarchy NEW! ?)				
GO molecular function complete	#	#	expected	Fold Enrichment	+/-	P value
AMP-activated protein kinase activity	7	7	.17	40.10	+	2.09E-06
↳catalytic activity	5908	261	147.33	1.77	+	2.27E-22
acid-sensing ion channel activity	4	4	.10	40.10	+	9.74E-03
↳substrate-specific transporter activity	1124	52	28.03	1.86	+	4.69E-02
↳transporter activity	1313	69	32.74	2.11	+	1.51E-05
↳transmembrane transporter activity	1022	65	25.49	2.55	+	2.27E-08
↳cation channel activity	299	28	7.46	3.76	+	1.19E-05
↳ion channel activity	419	28	10.45	2.68	+	9.69E-03
↳substrate-specific channel activity	433	28	10.80	2.59	+	1.77E-02
↳ligand-gated ion channel activity	142	26	3.54	7.34	+	2.52E-11
↳ligand-gated channel activity	142	26	3.54	7.34	+	2.52E-11
↳gated channel activity	325	26	8.10	3.21	+	8.61E-04
↳transmembrane receptor activity	1351	103	33.69	3.06	+	1.17E-20
↳receptor activity	1661	109	41.42	2.63	+	3.97E-17
↳molecular transducer activity	1661	109	41.42	2.63	+	3.97E-17

Figure 2. Hierarchical view of GO (molecular function) in gene list analysis result page. Each block of related function classes (the first block starts with *AMP-activated protein kinase activity*, the second with *acid-sensing ion channel activity*) is arranged with the most specific class at the top, with less specific classes indented below it.

reconstruct the hierarchical relationships, the 'Complete list of pathways' and 'Pathway hierarchy relationship' files were used. The second type of data is the gene to pathway association. In order to integrate these associations into PANTHER, we needed to map the Reactome protein identifiers to the UniProt Reference Proteome identifiers used in PANTHER. The 'UniProt to All pathway mapping' file was used from the Reactome site. In most cases these matched a Reference Proteome entry (77 528), but for those that did not, we used BLAST scoring with 90% or more length coverage and 90% or more sequence identity (10 149). In total, of the 111 627 protein identifiers in Reactome, 87 677 could be mapped to PANTHER using the above process. The major reasons for a lack of mapping from Reactome to PANTHER were the following: (i) the protein is not from one of the 104 genomes in PANTHER; (ii) the protein is a non-canonical isoform of a UniProt reference protein (Reactome considers isoforms and partial sequencing products from TrEMBL to be distinct entities, while PANTHER considers each protein-coding gene to be a distinct entity); (iii) the UniProt identifier from Reactome has been obsoleted.

Enabling direct upload of variant data. During the past 10 years, there has been an increase in studies of disease risk by discovering the association between diseases and genetic variants. With the advancement in genome sequencing technologies, such studies have become more routine. Currently, a growing number of PANTHER users are using our tools to analyze such data. They have to first map the variants to the genes themselves, and then input the gene list to PANTHER for enrichment or overrepresentation tests. In order to better support our users, we improved our gene list analysis tools to allow users to submit genetic variants directly to the statistical tests. Users upload a VCF-formatted file directly. The PANTHER tools first map the variant to a gene when it is either within a gene or within a flanking region of the gene. The default is 20 kb but users can define their own flanking regions. The variants that cannot be mapped to a gene are ignored in the analysis. If a single variant is mapped to more than one gene, the nearest gene is selected. If more than one variant is mapped to the same gene, the gene is only counted once. Note that the overrepresentation test takes as input only a list of variants, while the enrichment test also takes a numerical value associated with each variant, in this case the 'variant-trait association *P*-value.' To generate a gene-based *P*-value, each gene is assigned a *P*-value equal to the smallest *P*-value of any variant mapped to the gene, multiplied by the total number of variants mapped to the gene (i.e. a Bonferroni correction).

Hierarchical view of gene list analysis results. All the classifications in PANTHER have a hierarchical structure, or, more generally, that of a directed acyclic graph (DAG). While a hierarchy requires a subclass to have at most one parent class, a DAG allows for multiple parentages. In both cases, understanding the class-subclass relations is critical for data interpretation, as all genes in a subclass (e.g. mismatch repair) are also members of its parent (and grandparent, etc.) classes (e.g. DNA repair). As a result, positive overrepresentation/enrichment test results for related classes should not be interpreted as independent from each other, and may even reflect elements of the same underlying biology. In order to help users utilize the relationships between classes, we have now implemented a 'hierarchical view' as the default for analysis results. This view groups together related classes into a 'block,' and orders the blocks by the largest enrichment value of any class in the block (Figure 2). The most specific class is shown at the top of each block (these tend to be the most informative), with its more

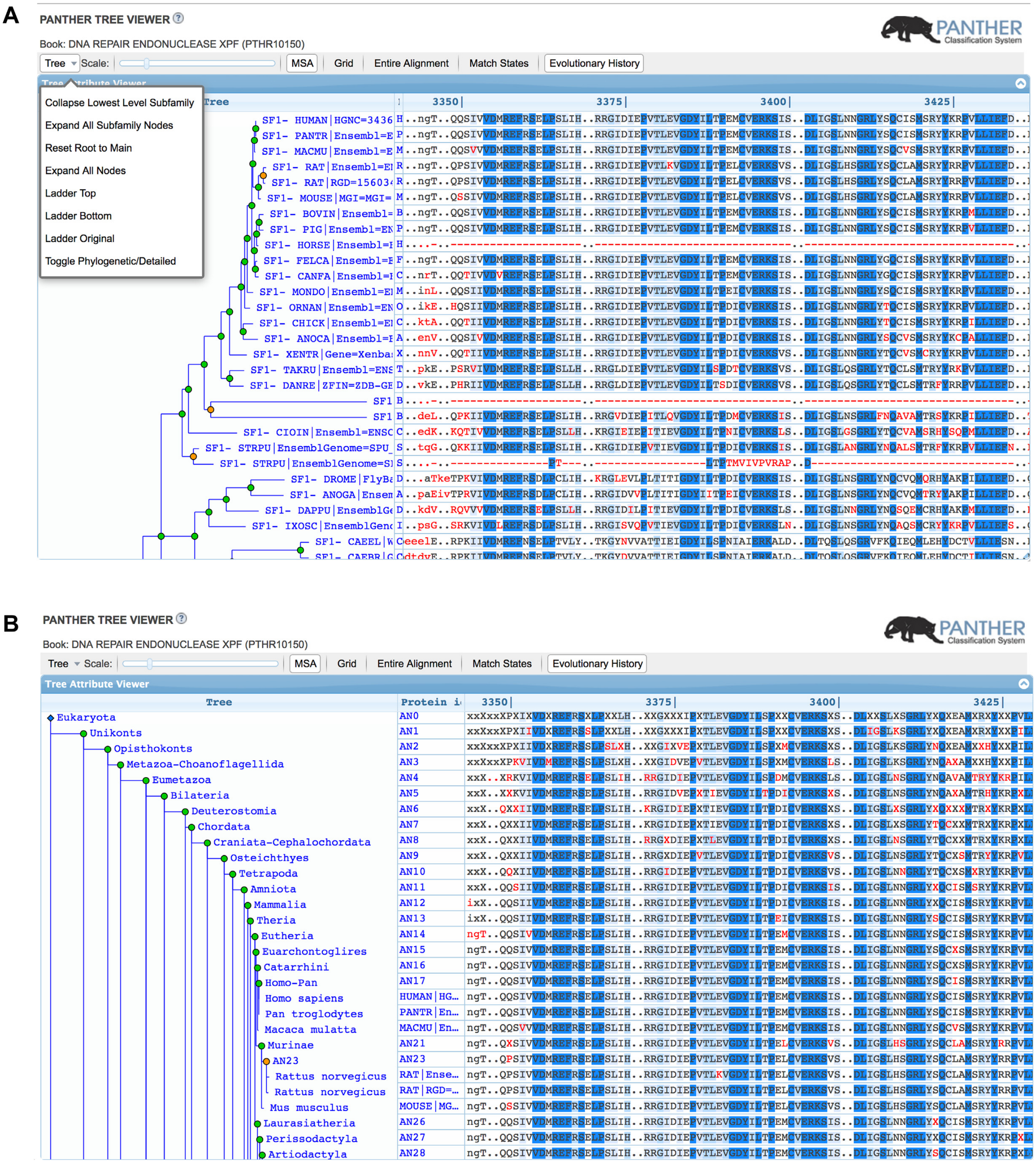


Figure 3. Tree Viewer, showing the tree panel on the left, and multiple sequence alignment panel on the right with the 'Evolutionary History' view selected. This view shows in red text the amino acids that have changed in the tree branch immediately leading to each sequence. (**A**) The standard 'phylogenetic view' of the tree, showing only the sequences of the leaves (extant genes) of the tree. (**B**) The new 'hierarchical view' of the tree, showing the sequences of both internal nodes (inferred ancestral sequences) and leaves.

general classes indented below it. As has always been the case, users can also sort by any other value in the results table, by clicking on the column header. Thus, the older default setting can be restored with a single click. Note that in the hierarchical view, each ontology class is displayed only once to avoid unnecessary redundancy. When a class has more than one descendant, it is shown only in the same block as its first descendant starting at the top of the results list.

JavaScript tree viewer with new 'evolutionary history' view of reconstructed ancestral sequences

The Java Tree Viewer applet has been an integral part of the PANTHER website since 2003. It not only allows users to view sequences in an evolutionary context, but also displays the attributes associated with the sequences as well as the sequence alignment such that they are aligned with the nodes in the tree. Recently, many popular browsers have stopped supporting Java applets. We have now developed an HTML5 compliant JavaScript web application with all the features that were available on the Tree Viewer applet, and more. Although there are various web-based phylogenetic tree viewers, they do not have functionality to display sequence attributes aligned with the nodes of the tree in a tabular fashion.

Two novel views of protein family data have been added to help users view the information about not only modern-day genes, but also ancestral genes. First, a 'hierarchical view' has been added to the tree panel, so that the tree can be viewed alternatively as a hierarchy, starting with the root node at the top (Figure 3B). The hierarchical view can be toggled with the standard 'phylogenetic view' (Figure 3A) by a selection under the 'Tree' menu item. The hierarchical view includes a line for every node in the tree, not just leaf nodes, so the multiple sequence alignment panel includes inferred ancestral sequences for those nodes. Note that if a site in an ancestral sequence cannot be confidently reconstructed, it is represented by 'X' (the symbol for an unknown amino acid). In addition, with either the hierarchical or phylogenetic view of the tree, users can select the 'evolutionary history' view of the multiple sequence alignment (Figure 3, right panels) that highlights the amino acids where sequence changes occurred along the tree branch immediately leading to that node or leaf. Thus users can more easily explore how protein sequences evolved during different periods and lineages in the family's evolutionary history.

ACKNOWLEDGEMENTS

The authors want to acknowledge the contributions of the GO Phylogenetic Annotation curators: Marc Feuermann, Pascale Gaudet, Karen Christie, Donghui Li. The authors would also like to acknowledge the Reactome developers for their excellent work on the resource, and Robin Haw for valuable input on integrating the Reactome data.

FUNDING

National Science Foundation [1458808]; National Human Genome Research Institute of the National Institutes of Health [award number U41HG002273]; National Institutes of Health [award number U54EB020406]. Funding for open access charge: National Institutes of Health [award number U54EB020406]; National Science Foundation [1458808].
Conflict of interest statement. None declared.

REFERENCES

1. Mi,H., Poudel,S., Muruganujan,A., Casagrande,J.T. and Thomas,P.D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.*, **44**, D336–D342.
2. Gaudet,P., Livstone,M.S., Lewis,S.E. and Thomas,P.D. (2011) Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Brief. Bioinform.*, **12**, 449–462.
3. Mi,H. and Thomas,P. (2009) PANTHER pathway: An ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.*, **563**, 123–140.
4. Thomas,P.D, Kejariwal,A., Guo,N., Mi,H., Campbell,M.J., Muruganujan,A. and Lazareva-Ulitsky,B. (2006) Applications for protein sequence-function evolution data: MRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.*, **34**, W645–W650.
5. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.
6. Thomas,P.D. and Kejariwal,A. (2004) 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**, 15398–15403.
7. Tang,H. and Thomas,P.D. (2016) PANTHER-PSEP: Predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*, **32**, 2230–2232.
8. Mi,H., Muruganujan,A., Casagrande,J.T. and Thomas,P.D. (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.*, **8**, 1551–1566.
9. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R. *et al.* (2015) The interpro protein families database: The classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
10. Sonnhammer,E.L., Gabaldón,T., Sousa da Silva,A.W., Martin,M., Robinson-Rechavi,M., Boeckmann,B., Thomas,P.D., Dessimoz,C. and Quest for Orthologs consortium. (2014) Big data and other challenges in the quest for orthologs. *Bioinformatics*, **30**, 2993–2998.
11. UniProt Consortium. (2015) UniProt: A hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
12. Gene Ontology Consortium. (2017) Expansion of the gene ontology knowledgebase and resources. *Nucleic Acid Res.*, **45**, doi:10.1093/nar/gkw1108.
13. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
14. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
15. Hughey,R., Karplus,K. and Krough,A. (2003) SAM: Sequence alignment and modeling software system. *Tech. Rep.*,
16. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S. *et al.* (2016) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
17. Sasidharan,Nair P. and Vihinen,M. (2013) VariBench: A benchmark database for variations. *Hum. Mutat.*, **34**, 42–49.
18. Mottaz,A., David,F.P., Veuthey,A.L. and Yip,Y.L. (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using swissvar. *Bioinformatics*, **26**, 851–852.