PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions, and extensive API

Huaiyu Mi[1*], Dustin Ebert[1], Anushya Muruganujan[1], Caitlin Mills[1], Laurent-Philippe Albou[1], Tremayne Mushayamaha[1], Paul D. Thomas[1*]

[1]Division of Bioinformatics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089

*Corresponding authors:

Email to: huaiyumi@usc.edu and pdthomas@usc.edu

Abstract

PANTHER (Protein Analysis Through Evolutionary Relationships, http://www.pantherdb.org) is a resource for the evolutionary and functional classification of protein-coding genes from all domains of life. The evolutionary classification is based on a library of over 15,000 phylogenetic trees, and the functional classifications include Gene Ontology terms and pathways. Here, we analyze the current coverage of genes from genomes in different taxonomic groups, so that users can better understand what to expect when analyzing a gene list using PANTHER tools. We also describe extensive improvements to PANTHER made in the past two years. The PANTHER Protein Class ontology has been completely refactored, and 6101 PANTHER families have been manually assigned to a Protein Class, providing a high level classification of protein families and their genes. Users can access the TreeGrafter tool to add their own protein sequences to the reference phylogenetic trees in PANTHER, to infer evolutionary context as well as fine-grained annotations. We have added human enhancer-gene links that associate non-coding regions with the annotated human genes in PANTHER. We have also expanded the available services for programmatic access to PANTHER tools and data via application programming interfaces (APIs). Other improvements include additional plant genomes and an updated PANTHER GO-slim.

Introduction

PANTHER (Protein Analysis Through Evolutionary Relationships) is an integrated knowledgebase of evolutionary and functional relationships between protein-coding genes[1] (referred to hereafter as proteins), as well as tools for using the classifications to analyze large-scale genomics data[2,3]. Proteins are classified along two primary axes (Figure 1): evolutionary groupings (Protein Class, Protein Family, Subfamily), and functional groupings (Gene Ontology and pathways). **Evolutionary groupings** correspond to the "natural classification" of proteins according to their evolutionary histories. Protein families are groups of proteins that are homologous (share a common ancestor) and can be aligned to one another in a multiple sequence alignment that passes some basic quality checks. Subfamilies are orthologous groups of proteins identified in the phylogenetic tree inferred for a given protein family. There can be many subfamilies per family, and some families are very diverse, and the subfamilies identified by PANTHER are an important feature that give the annotations a high degree of functional specificity. At a more general level, protein families are grouped together into broad functional classes (Protein Class). **Functional groupings** classify proteins according to the functions of individual proteins, not families. Functions are classified in multiple ways: three different aspects defined by the Gene Ontology[4,5] (molecular function, cellular component, and biological process), and pathways (signaling and metabolic pathways). Functions are annotated using two different methods. The first method annotates clades (defined using internal tree nodes) of related proteins in a phylogenetic tree, resulting in annotations of all the proteins in the annotated clade. These annotations include Gene Ontology terms from PANTHER GO-Slim, a simplified or "slim" subset of the complete GO, as well as terms from the PANTHER Pathway[6] ontology. The PANTHER Pathway module[6] includes assignments of protein clades to pathway components, as well as a pathway model generated using CellDesigner[7], which can be displayed or downloaded. The second method annotates individual proteins to functional classes, either the "complete" Gene Ontology (imported from the GO resource), or the Reactome pathways[8] (imported from the Reactome resource).

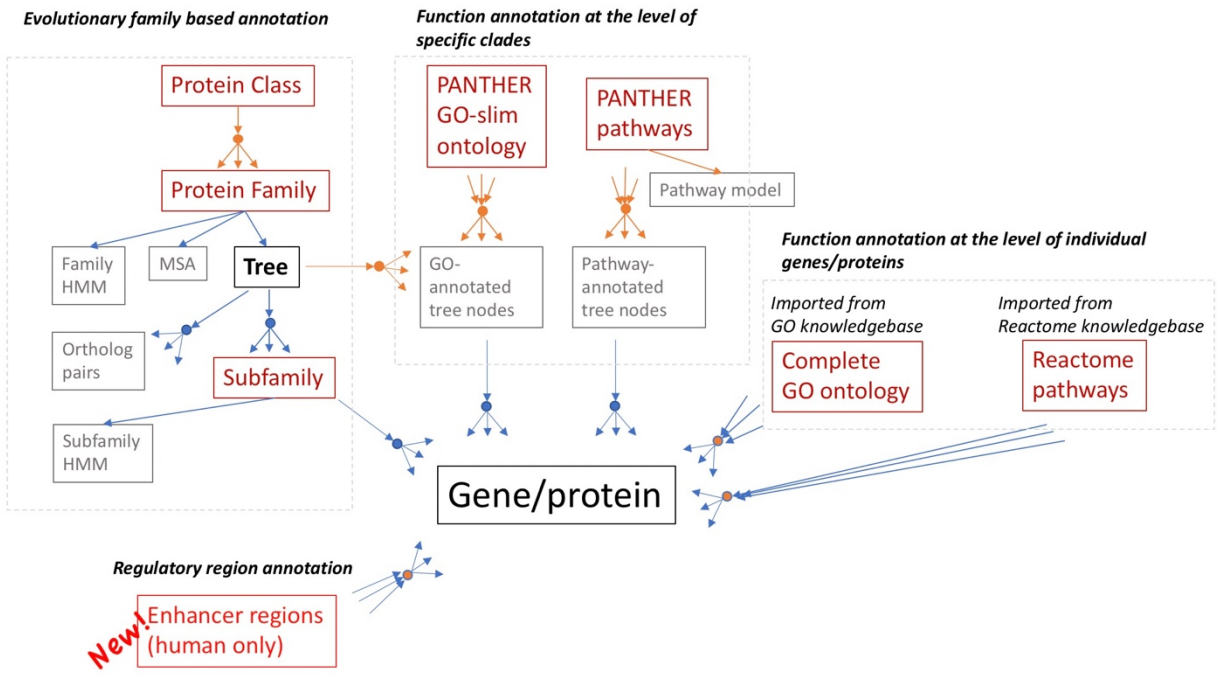


Figure 1. Relationships between classifications and data in PANTHER, showing the different ways in which a gene/protein (bottom) can be classified (red), as well as the other data available in PANTHER (gray). Classification types are shown in red. Arrows indicate the nature of the relationship (1:1, 1:many, many:many) and how the relationships are derived (blue=computationally, orange=expert curation, orange circle in blue=both methods). As a result, a given gene can be associated with only one subfamily, family and protein class, but can be associated with more than one GO term, pathway, or enhancer region.

One of the primary features in PANTHER is a set of **reference phylogenetic trees**. PANTHER provides extensive data for each protein family:

- A phylogenetic tree
- GO annotations for internal nodes of the phylogenetic tree (available at pantree.org)
- Hidden Markov models (HMMs) for each family and subfamily
- Multiple sequence alignment (MSA) for all family members
- Pairwise orthologs, within-species paralogs, and xenologs

Reference trees are currently available for over 15,000 protein families, and include all members of each family in 142 fully sequenced genomes. These genomes sample the entire tree of life, though the selection of genomes is biased toward "model" organisms for which extensive experimentally-supported GO annotations have been captured by the GO Consortium: four vertebrates (human, mouse, rat, zebrafish), two invertebrates (fruit fly and the nematode worm *C. elegans*), two single-celled fungi ("yeasts": budding yeast and fission yeast), the slime mold *Dictyostelium*, the plant *Arabidopsis*, and the bacterium *E. coli*. Sampling of plant genomes has also recently been expanded. Table 1 shows the sample of organisms in PANTHER version 16 reference trees.

Table 1. Number of genomes from each kingdom/phylum in PANTHER 16.0

Type	Number

bacteria	35
archaea	8
fungi	14
plants	40
protista and alveolata	8
amoebozoa	3
invertebrates	15
vertebrates	19
TOTAL	**142**

As described in more detail below, we have also recently added the PEREGRINE database of gene-enhancer links [submitted]. Users can now view lists of enhancers that have been associated with the expression of each human protein-coding gene in PANTHER. Users can also upload a list of genomic regions to PANTHER and retrieve a list of genes linked to enhancers in those regions.

Finally, PANTHER provides three main types of software tools: 1) protein classification tools, 2) gene list analysis tools, and 3) a protein coding variant analysis tool. The protein classification tools enable users to apply the PANTHER classifications to their own protein sequences of interest. PANTHER now has two classification tools: PANTHER HMMs, and TreeGrafter. PANTHER HMMs use family and subfamily-level HMMs to classify user-specified protein sequences[2], and have also been integrated into InterProScan[10,11]. TreeGrafter is a new tool (described in more detail below) that utilizes the PANTHER trees for classification[12], localizing a user-specified sequence to its most likely branch in the tree. Because the PANTHER trees have been annotated with both subfamily names (subtrees) as well as GO terms describing functions, TreeGrafter can also assign subfamilies and GO functions. The gene list analysis tools enable users to upload lists of genes or proteins and perform statistical tests to find enriched functional classes in their list[3,13]. The over-representation tool takes an input list of genes and reports statistically over- and under-represented classes relative to a reference list. The enrichment tool takes an input list of genes and quantitative values (e.g. expression levels) and reports classes for which the gene values are distributed non-randomly (e.g. tend to be higher or lower than average). In addition to the 142 genomes in the reference trees, PANTHER has pre-computed annotations for hundreds of other genomes, which are also available for analysis on the PANTHER website.

For over 20 years, PANTHER has supported a large user community in various capacities. PANTHER maintains collaborations with core data resources and consortia. As a member of the Gene Ontology Consortium, PANTHER family trees are the backbone of the GO phylogenetic

annotation effort[14,15]. In addition, PANTHER gene list analysis tools serve as the backend tool for the GO enrichment tool. PANTHER also collaborates with UniProt[16], InterPro[10], and Phylogenes (http://www.phylogenes.org/). PANTHER tools are highly used by the scientific community, many of whom are bench scientists. According to Google Analytics, 170,838 unique users visited the PANTHER website this year (1/1/2020-9/8/2020), compared to 144,572 for the same period in 2018.

Here, we give an overview of the taxonomic coverage of PANTHER annotations, and report the improvements made to PANTHER since our last update two years ago[17]. Improvements have been made in several areas: the protein families, the classifications (particularly Protein Class, Protein Family, and PANTHER GO-slim), new web services via application programming interfaces (APIs), new software tools, and the addition of enhancer data.

Coverage of PANTHER annotations for different taxonomic groups

PANTHER has continued to expand coverage of the 142 genomes represented in the phylogenetic trees, and many clade-specific families have been added to PANTHER to increase coverage of these genomes. Nevertheless, as discussed above, the sampling is biased toward genomes that are relatively well-characterized experimentally. As a result of this uneven sampling, the current coverage of genes by PANTHER annotations varies somewhat for different genomes (Figure 2). PANTHER family and subfamily annotations are highest for vertebrates (over 90% on average), slightly lower for plants (over 80% on average), and slightly lower still for invertebrates, fungi and bacteria (over 70% on average). Not surprisingly the lowest coverage is for single-celled eukaryotic species and for archaea, although coverage is still fairly high (usually 50-70%). Other family- and clade-level annotation types have lower coverage than families and subfamilies (Figure 2). Thus, PANTHER analysis tools can be applied broadly to genomes of organisms across the tree of life, though annotations are generally more complete for those that are more closely related to a well-studied “model” organism (*viz*. vertebrates, insects, nematodes, ascomycete fungi, dicot plants, and gamma-proteobacteria; bacilli are also relatively well sampled among bacteria).

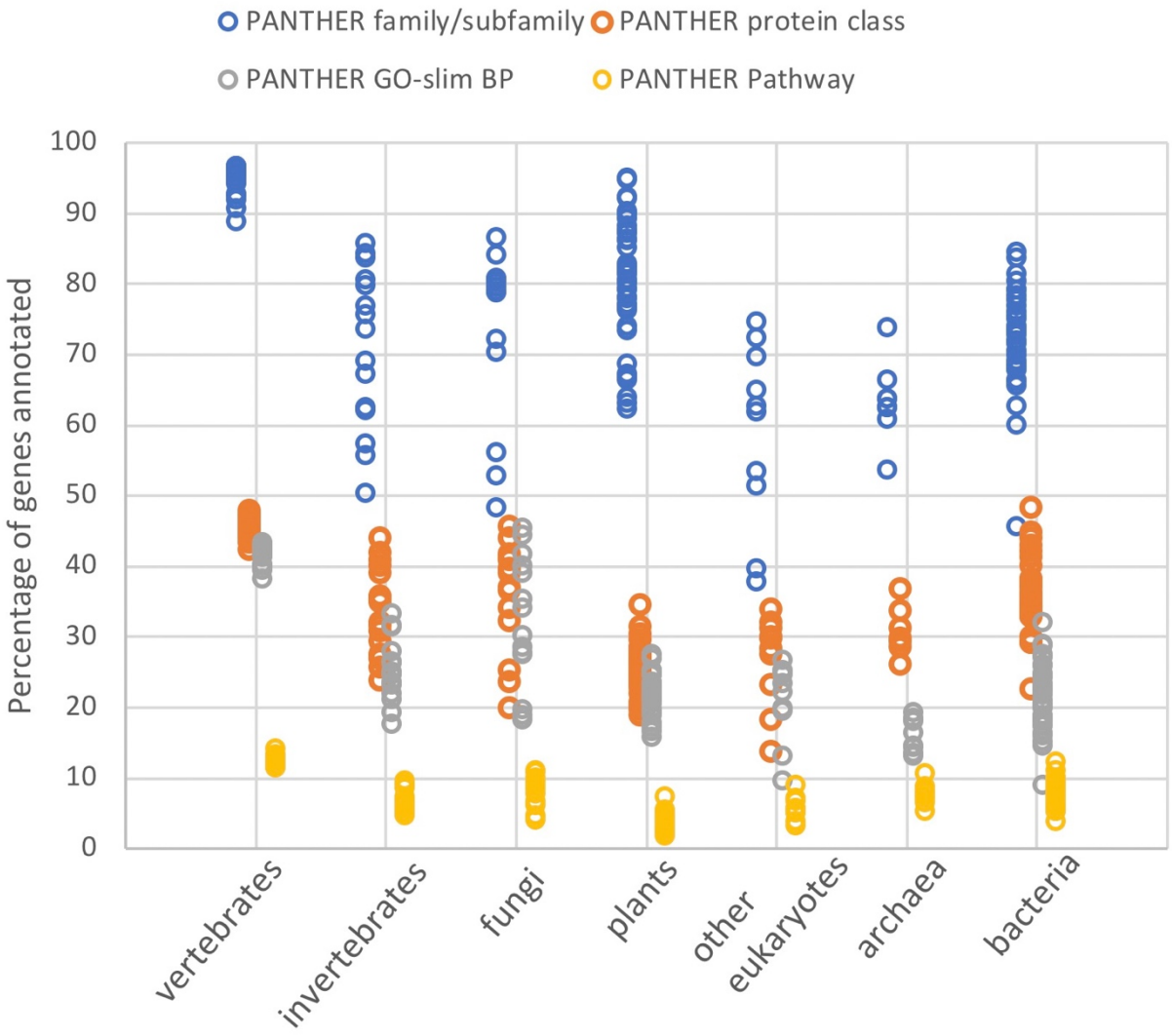


Figure 2. Coverage of protein-coding genes by PANTHER evolutionary and clade-level annotations. Each circle represents the coverage of one of the 142 PANTHER reference genomes, for different annotation types. Vertebrate genomes have the highest coverage for all annotation types, but other genomes are still covered to an appreciable extent.

The coverage of individual gene-level annotations depends even more strongly on the organism being analyzed (Figure 3). GO "complete" annotations, imported from the GO database, are assigned by many different methods including manual annotation from the literature, manually reviewed homology inferences (including the PANTHER GO-slim annotations) and computational methods such as InterPro2GO[9]. Reactome pathway annotations are imported as well, and cover genes in humans and a few other selected organisms based on orthology relationships to human genes.

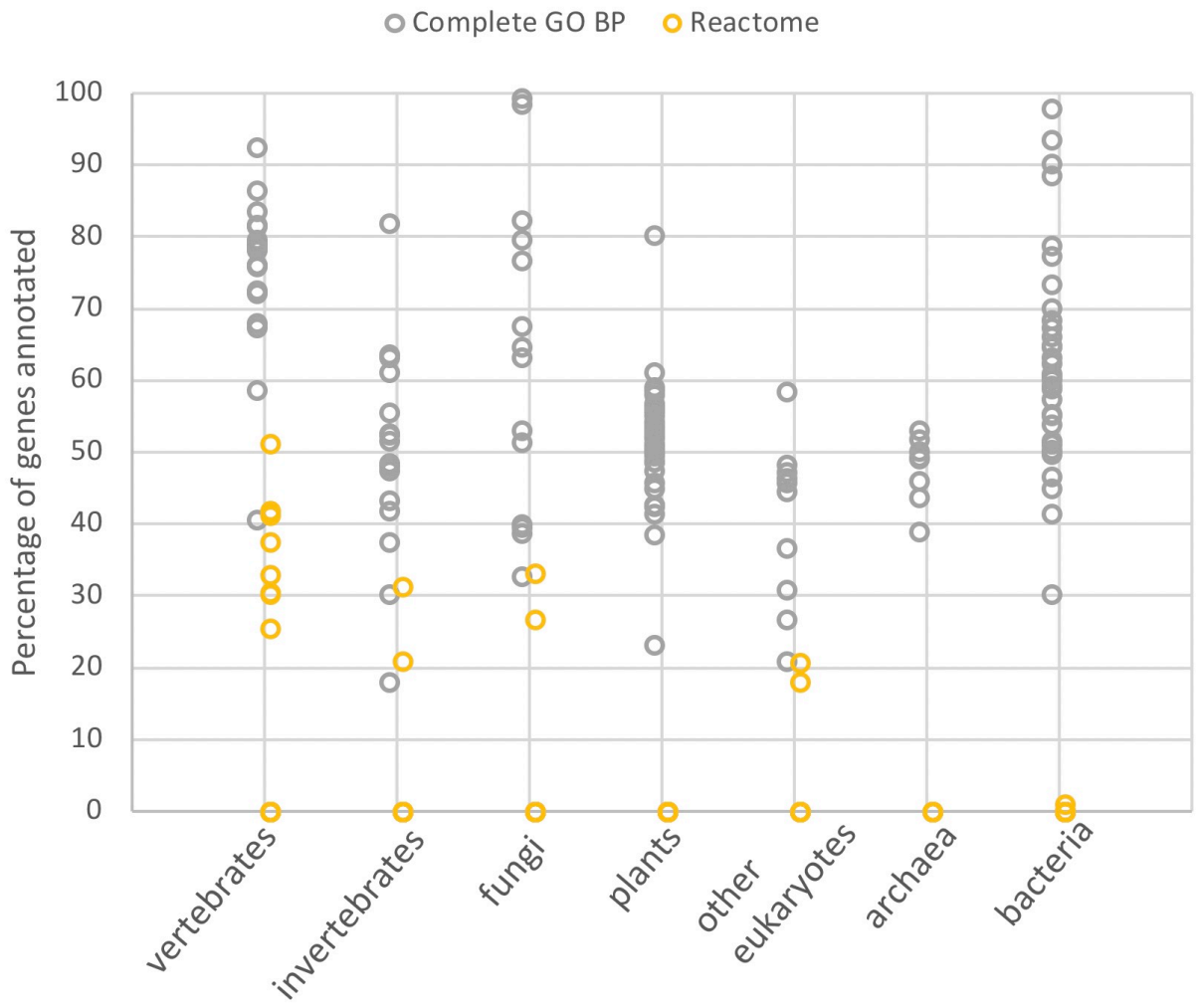


Figure 3. Coverage of protein-coding genes by individual protein-level annotations imported from GO and Reactome. Each circle represents the coverage of one of the 142 PANTHER reference genomes, for different annotation types. Coverage is more variable than for evolutionary and clade-level annotations, and most genomes are not covered by Reactome annotations (overlapping points at 0).

Improved Protein Families

PANTHER families are constructed each year, using the latest protein sequences available from UniProt Reference Proteomes[18]. These proteomes are "gene-centric," meaning that there is a single representative protein sequence for each protein-coding gene in each of the selected genomes. Consequently, the protein families can also be interpreted as protein-coding gene families. Proteins (tree leaves, or extant genes) and internal tree nodes (ancestral genes) are given stable identifiers that are forward-tracked between releases. Since our last update, we have added 10 more plant genomes to PANTHER, in order to improve the phylogenetic representation of plant gene phylogeny. The complete list of genomes in PANTHER trees can be found at http://pantherdb.org/panther/summaryStats.jsp.

The addition of these plant genomes resulted in a significant increase in the number of genes for some families that have been highly duplicated in plants. As a result, some of these families became much more diverse, affecting the alignment quality and therefore the phylogenetic trees, as well as difficulties in visualizing and navigating them. We therefore performed manual curation to split these trees into smaller and more well-defined families. As a result, there is a slight increase in the number of families (Table 2).

The other major improvement was to provide a meaningful name to the families. This was done through both manual curation and computational assignment. Most of the new family name assignments utilize the PANTHER automatic subfamily name assignments, which in turn rely on protein names from UniProt[16] as described previously[19]. We increased the number of families with a meaningful name from 4905 (32%) two years ago to 14849 (95%) now (Table 2).

Table 2 Comparison of Numbers of Genomes, Genes and Families between PANTHER 14 and 16

	PANTHER 14.1 (2018)	PANTHER16.0 (2020)
Genomes	132	142
Genes in PANTHER families (trees)	1,750,742	2,065,831
Number of families	15524	15702
Number of families with a family name	4905	14849

Ontology Updates

PANTHER Protein Class Refactoring

The PANTHER Protein Class (PC) ontology has been completely refactored since our last update article in 2019. PANTHER PC (originally called PANTHER/X) was designed as a simple, high-level classification of the functions of proteins and protein families[1]. On the PANTHER website, PANTHER PC can be used for browsing the protein families, or to visualize and compare the protein-coding gene content of different genomes.

Starting with PANTHER version 15 (released in February 2020), the PANTHER PC classification has been refactored to adhere to several principles:

1. The PANTHER PC ontology is a strict hierarchy, i.e. a functional class cannot have multiple parent classes.
2. Protein families, and not individual subfamilies or proteins, are each assigned to a PC class. PANTHER PC is therefore a classification at the level of protein families.
3. Each protein family is associated with only one PC class. The only exception to this rule is for proteins that contain multiple functional domains that belong to different classes. In other

words, multiple classifications are allowed only in the case where there are two distinct, modular functions encoded by the same protein-coding gene.

4. The functional class of a protein family usually applies to all members of the family, but in some cases there may be a few family members that do not possess that function themselves. In other words, the functional class reflects the ancestral, and highly conserved, function of a protein family, but there may be a few functionally divergent members.
5. Classes are designed to be as biologically informative as possible, while remaining accurate when applied to whole protein families. Relatively uninformative terms in previous versions of PANTHER PC like *nucleic acid binding* have been obsoleted. New terms have been added or modified to make important distinctions between biological functions (see below).

These properties make the PC classifications easier to navigate and interpret than the other classifications in PANTHER such as Gene Ontology. Navigation is strictly up and down the hierarchy and not across categories, and genes will not be double-counted because of their association with multiple distinct functional classes. The trade-off for this simplicity is that the functional classes in PC are relatively broad and general, and will not provide a detailed description of protein function. In addition, a few subclasses (enumerated below) could arguably appear under more than one top level class, but we have retained only the more informative one. In adopting these simplifications, the PANTHER Protein Class is complementary to the detailed Gene Ontology and Pathway classifications in PANTHER, and users can apply these different classifications in various combinations to explore sets of protein-coding genes.

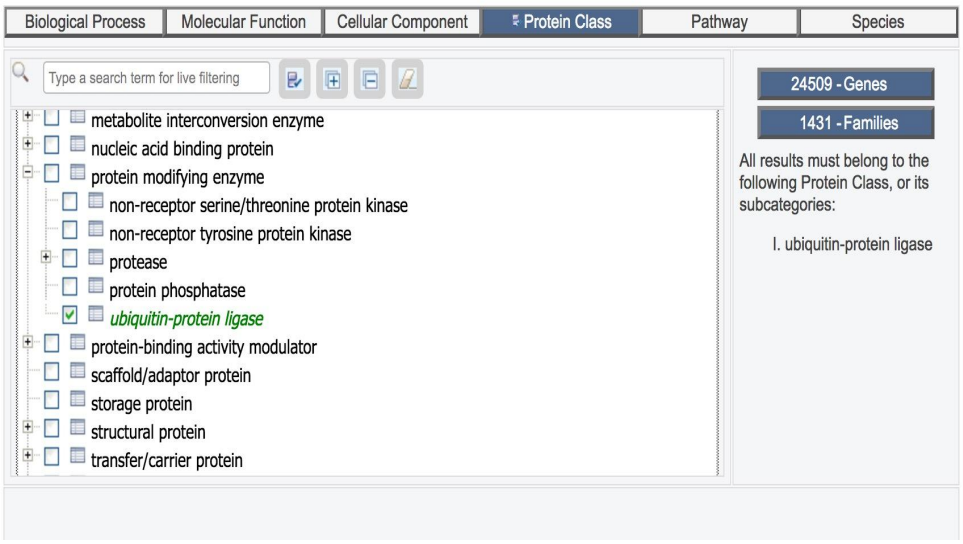
In version 16, there are 210 classes in PANTHER PC. The hierarchy is no more than four levels deep. The largest number of classes at any level is at the top level, where there are 24 classes. Compared with previous versions, we have introduced 9 new classes, renamed and redefined 28 classes, and moved some of the classes within the hierarchy. The most prominent changes are:

1. Distinguishing classes of enzymes by the substrate they act on. This arrangement is more biologically informative than our previous classification by enzyme mechanism. Mechanism-based subclasses are still used underneath, but not above, the top-level classes that distinguish the following substrate types:
    a. *Metabolite interconversion enzyme* (PC00262) for enzymes that act to convert small molecules into different small molecules
    b. *Protein modifying enzyme* (PC00260) for enzymes that covalently modify proteins. This class excludes enzymes that function in the context of protein folding (protein-disulfide isomerases are classified under *chaperone*) and chromatin modification (histone modifying enzymes), for which there is a more biologically informative class in PC.
    c. *DNA metabolism protein* (PC00009, formerly *DNA binding*) for proteins that are involved in the synthesis, modification or breakdown of DNA. Other proteins that bind DNA are now classified elsewhere, including DNA binding transcription factors, chromatin-related proteins and RNA polymerase subunits.
    d. *RNA metabolism protein* (PC00031, formerly *RNA binding*) for proteins involved specifically in the metabolism of RNA. Some RNA binding proteins formerly in the more general RNA binding class have been moved to a more appropriate parent class, e.g. DNA primase is now under *DNA metabolic enzyme*.

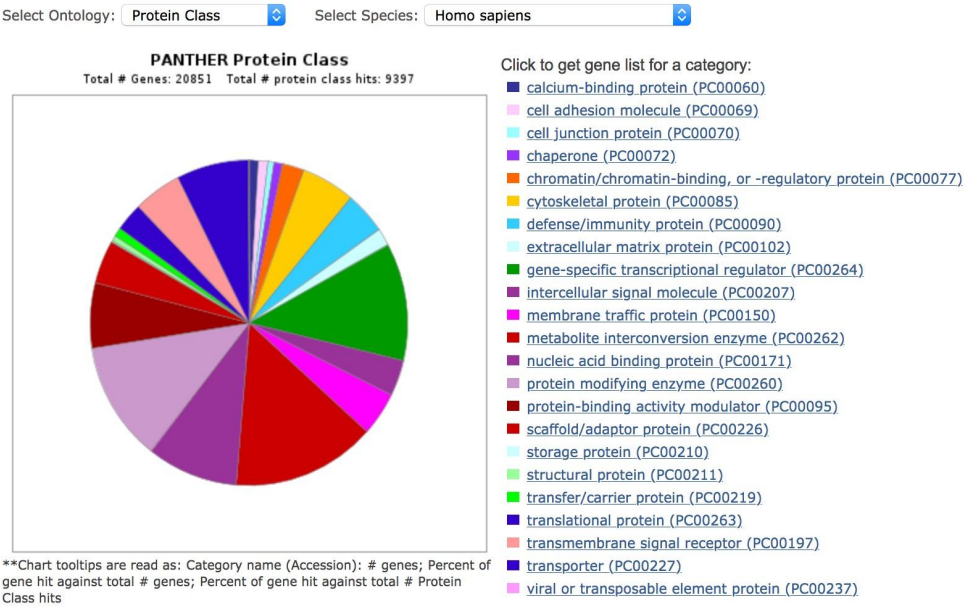
2. Distinguishing between general and specific transcription factors:
   a. New class for *general transcription factor* (PC00259), under *RNA metabolism protein*.
   b. New class for *gene-specific transcription regulator* (including both DNA binding transcription factors and protein-binding co-factors).
   c. Many additional new classes for subtypes of specific transcription factors (PC00244 through PC00256) to more closely match the classes in TFClass[20].
3. New top-level class for *chromatin, chromatin-binding or -regulatory protein* (PC00077, formerly *chromatin binding protein*), which now includes subclasses for histones and chromatin modifying proteins. Note that *histone modifying enzyme* (PC00261) is a new subclass here rather than under *protein modifying enzyme*, as their primary function is to modify chromatin structure, not protein structure.
4. Clear definition of *receptor* (PC00197) as *transmembrane signal receptor*. In previous versions this class was a mix of distinct functions for which the term *receptor* is also applied in biology, such as vesicle cargo receptors (now moved under *vesicle traffic protein*) and ligand-gated ion channels (now only under *transporter*).
5. Distinguishing between protein kinases according to whether they act intracellularly (under *protein modifying enzyme*), or as transmembrane receptors (under *transmembrane signal receptor*). In previous versions, protein kinases were classified by catalytic mechanism (*transferase* > *kinase* > *protein kinase*), with transmembrane receptor kinase classes having a second parentage under *receptor* (PC00197).
6. New top level class for proteins involved in translation (*translational protein*, PC00263) which groups together ribosomal proteins, translation factors and aminoacyl-tRNA synthetases.
7. Closer alignment with the Transporter Classification Database (TCDB)[21] classes for transporters. PANTHER PC now distinguishes between primary active transporters (P00068) and secondary carrier transporters (PC00258) and classifies proteins like nuclear pore subunits and ATP synthase subunits under the transporter branch.

The main use cases for the PC ontology are in **browsing the collection of gene families** in PANTHER, and in getting a clear **overview of an entire protein-coding genome** (or comparing more than one genome). Figure 4 illustrates these use cases, with screen shots of the actual tools available on the PANTHER website. Users can also use PC classifications to visualize or subset any gene list, including those created from GO or pathway classifications, or a user-specified gene list that was uploaded to the PANTHER website.

Figure 4. Using the Protein Class ontology. a) Browsing the PANTHER data using Protein Class. Note that number of families also includes subfamilies. b) Overview of entire set of protein-coding genes in a genome.
(a)



(b)



In the past two years, we have completely re-curated the associations between PANTHER families and PANTHER PC classes, including reviewing all classifications from previous versions. In PANTHER version 16.0, 6101 of the 15702 families (39%) are associated with a Protein Class. These tend to be the larger families, resulting in a large fraction of protein-coding genes with a PC assignment. In PANTHER version 16, 58% of human genes (12043) are in families with a PC assignment. Users should note that a family can be associated with a PANTHER PC class at any level, and while we strive to make the classification as specific as possible, many protein families may appear only in a more general class. If users see a classification they think might be incorrect or too general, we encourage them to notify us so we can prioritize it for review.

PANTHER GO-Slim Update

The new PANTHER GO-slim was updated from the latest phylogenetic annotations provided by the GO Phylogenetic Annotation project[15]. It contains 3336 terms, with 2234 biological process, 557 molecular function and 542 cellular component terms. The number of classes has increased by 289 terms compared to the one reported in our last update in 2019[17], due to the number of additional GO terms that have been used by the GO Phylogenetic Annotation project. The ontology can be downloaded from http://data.pantherdb.org/PANTHER16.0/ontology/panther_slim.obo.

New Services and Tools

New PANTHER services

The era of big data has been driven by increases in the speed and resolution of data acquisition methods, and PANTHER is no exception. Automated methods to analyze data has become a necessity for researchers and biologists for comprehensive studies of biological systems. In addition to data analysis, scientists are no longer interested in raw data, but data that has been categorized and curated. PANTHER datasets have also grown and there is an ever-increasing number of users who are interested in accessing it to further their research. PANTHER services have been created to address the exponential growth in data by providing tools to retrieve and analyze data. It gives users programmatic access to PANTHER's expert-curated data and analytical tools via RESTful Web Services[22]. Supporting programmatic access allows researchers to easily incorporate PANTHER into existing or new workflows for scientific discoveries. The services are described using OpenAPI Specification (link http://api.pantherdb.org) which is a broadly adopted industry standard.

As indicated by system usage statistics, PANTHER is primarily used in three ways: 1) to compare lists of genes to determine if any biological processes are under or over represented in a list, 2) to retrieve annotation information for a given set of genes, and 3) to retrieve information about related genes based on evolutionary history.

Careful consideration was given to creating service endpoints that could be incorporated into meaningful workflows. In order to support the various use cases, three main categories of services have been created: Utilities, Data Mapping and Phylogenetic Tree services (Table 3). The Utilities module contains the widely used overrepresentation service, which compare a list of genes to a reference list to statistically determine over- or under-representation of a specified ontology class. By default, the system uses Fisher's exact test with FDR correction, but there is also an option to use the Binomial test. Both methods have options for FDR correction, Bonferroni correction or no correction. This service supports various annotation data sets: Complete GO (three aspects, listed above), PANTHER GO-slim (3 aspects), PANTHER Protein Class, PANTHER Pathway, and Reactome Pathways.

Annotations associated with “complete” Gene Ontology classes are updated monthly from the official Gene Ontology release. This group of PANTHER web services also includes modules to search for orthologs and other homologs of genes.

Table 3. Summary of PANTHER Services

Services	Tools	Description
Utility	statistical overrepresentation test	Fisher's Exact test with option of FDR or Bonferroni correction
	ortholog/paralog service	Returns orthologs from a user specified organisms or paralogs for an input gene list.
	homolog position service	Given a gene and an amino acid position, returns a list of homologs for the gene from a user specified organisms and corresponding amino acid positions
Data Mapping	gene list classification	Returns functional classification of an input gene list.
Phylogenetic Tree	tree service	Returns the tree topology and its attributes of an input of PANTHER family ID.
	MSA service	Returns the multiple sequence alignment of an input PANTHER family ID. The alignment includes both leaf sequences and internal nodes.
	family ortholog	Returns all orthologs for sequences for an input PANTHER family ID.
	TreeGrafting	Returns a tree topology of a family tree with grafted protein sequence.

The Data Mapping service allows users to upload a list of genes and get back information about expert curated annotations and pathways that have been assigned to the input list of genes.

PANTHER trees contain groups of genes based on evolutionary history, and given that PANTHER currently supports over 142 genomes, it was imperative that users were equipped with tools to focus on specific taxa based on their field of research. The PANTHER species tree (http://pantherdb.org/panther/speciesTree.jsp) allows users to select higher order taxonomic groups or lower level species of interest and obtain associated lists of families and taxonomies. The list of families can be used as input with the Phylogenetic Tree services module to only retrieve trees containing species of interest. Furthermore, the NCBI taxonomy identifiers can be specified with the Phylogenetic Tree services to include only proteins from selected taxa. A service to determine the list of orthologous genes in a family is also available as part of this module.

A grafting module (see TreeGrafter description below) is also available that allows users to specify a protein sequence and retrieve the best matching homologous family and graft in the best location in the tree. Additionally, it also returns the associated annotations associated with the specified sequence.

Due to the nature of constantly changing data, annotations and tools, all services include versioning and date information for ensuring that analysis results can be replicated by researchers, peers and reviewers.

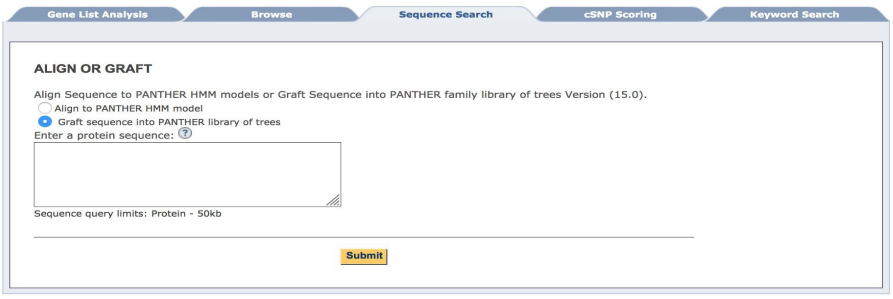
TreeGrafter software for classifying user-specified protein sequences

TreeGrafter is a software package for classifying a protein sequence by "grafting" it onto a reference protein family tree[12]. The algorithm proceeds in three steps: 1) scoring a sequence against a library of family HMMs, 2) adding the new sequence to the reference multiple sequence alignment for the highest-scoring family, and 3) determining the most parsimonious graft point for the new sequence, relative to other sequences in the reference tree. TreeGrafter has been shown to be more accurate than the subfamily HMM scoring that users of PANTHER have been applying for over 20 years.

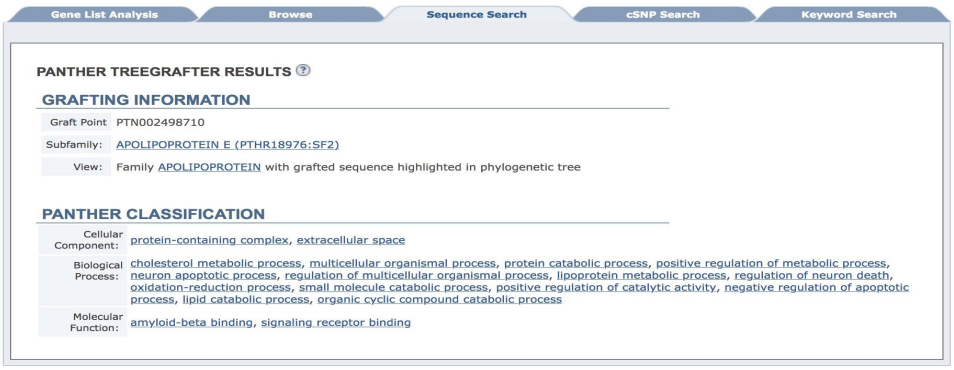
Starting with PANTHER version 15 (released in early 2020), users can now apply TreeGrafter for classification of new sequences. This is an additional option for users; the older subfamily HMM searching is still available. TreeGrafter can be accessed interactively on the PANTHER website, from the "Sequence Search" tab on the homepage (Figure 5a). Annotations for a new sequence depend on the graft point on the reference tree, and include PANTHER subfamily, and manually-annotated GO terms on tree branches made by the Gene Ontology Phylogenetic Annotation project[15] (Figure 5b). In addition, the phylogenetic tree showing the graft point of the sequence can be visualized using the PANTHER TreeViewer tool (Figure 5c). TreeGrafter for classifying a single sequence can also be accessed via web services (see below), or downloaded as a command-line tool for use at large-scale (https://github.com/pantherdb/TreeGrafter). Note that TreeGrafter annotations are updated with every PANTHER release.

Figure 5. Using TreeGrafter to classify a new protein sequence. a) Users can classify new sequences using HMMs, or TreeGrafter. b) TreeGrafter results show subfamily assignment, and GO terms that depend on the graft point in the tree. c) Users can view the grafted sequence in the context of the reference phylogenetic tree. The new sequence is an additional leaf in the tree (marked 'ANGRAFTED' and highlighted in blue), with the new internal node (pointed by a red arrow) induced by the grafting (marked 'ANINDUCED').

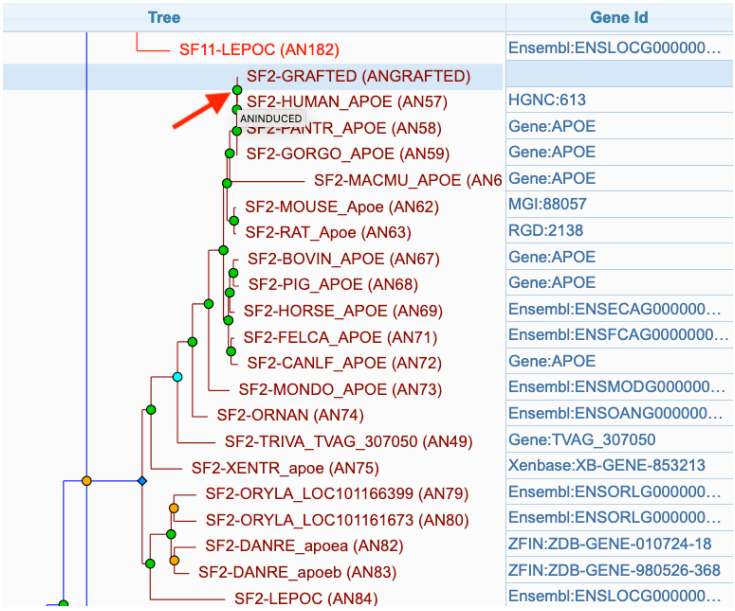
a)



b)



c)



Enhancer-gene links

Enabling annotation of genetic variants in non-coding enhancer regions

In our 2017 update[23], we described the extension of the gene list analysis tool to allow direct upload of human genetic variants. The tool initially considered only variants in the protein-coding regions of the human genome. However, the majority of the genetic variants are in the non-coding regions of the genome. Enhancers are short DNA elements in those regions that function primarily as clusters of transcription factor binding sites that are spatially coordinated to regulate expression of one or more specific target genes. PEREGRINE (**P**redicted by **E**xperimental **R**esults: **E**nhancer-**G**ene **R**elationships **I**llustrated by a **N**exus of **E**vidence) is an enhancer to gene link database (www.peregrineproj.org) with genome-wide characterization of regulatory connections between enhancers and target genes (submitted). It incorporates publicly available experimental data from ChIA-PET, eQTL, and Hi-C assays across 78 cell and tissue types to link 449,627 enhancers to 17,643 protein-coding genes. The PEREGRINE human enhancer-gene links have

now been added to the PANTHER database. There are two ways for users to access the data. The first is to find the enhancers that are associated to a particular gene by querying the target gene using the existing gene search function in PANTHER (this is currently on a test site: http://panthertest6.med.usc.edu:8086/. It will be moved to the production PANTHER website before this paper is published). An enhancer module can be enabled to retrieve the list of enhancers (Figure 6a). The second is to submit a list of genetic variants (in VCF format) and retrieve a list of enhancers these variants are located in, and the target gene(s) each enhancer may interact with (Figure 6b). The details about the enhancer-gene links, including tissue information and assays that support the links, can be viewed by clicking each IDs in the Enhancer column.

Figure 6. PANTHER now includes PEREGRINE enhancer-gene links data. a) A screenshot of a gene list page after the user queries the PANTHER system with a list of genes (UniProt IDs in Mapped IDs column). The enhancers associated to each gene are listed in Enhancer column (red arrow). b) A screenshot of the PANTHER gene list page when a user submits a VCF file. The coordinates of the variants are listed in the Mapped IDs column. PEREGRINE will map the variants to the enhancers (Enhancer column) as well as the gene(s) that is regulated by the enhancers (Gene ID column).

a)

clr all		Gene ID	Mapped IDs	Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class	Species	Enhancer
	1.	HUMAN|HGNC=7763|UniProtKB=Q15784	Q15784	Neurogenic differentiation factor 2 NEUROD2 ortholog	NEUROGENIC DIFFERENTIATION FACTOR 2 (PTHR19290:SF83)	basic helix-loop-helix transcription factor nuclease	Homo sapiens	24370 EH37E0432254 EH37E0432603 EH37E0432796
	2.	HUMAN|HGNC=8546|UniProtKB=P13674	P13674	Prolyl 4-hydroxylase subunit alpha-1 P4HA1 ortholog	PROLYL 4-HYDROXYLASE SUBUNIT ALPHA-1 (PTHR10869:SF101)	-	Homo sapiens	70386 EH37E0179053 EH37E0179158 EH37E0179441
	3.	HUMAN|HGNC=8283|UniProtKB=P47893	P47893	Olfactory receptor 3A2 OR3A2 ortholog	OLFACTORY RECEPTOR 3A2 (PTHR26451:SF349)	-	Homo sapiens	24241 EH37E0420457 EH37E0420458 EH37E0420463
	4.	HUMAN|HGNC=8477|UniProtKB=Q15620	Q15620	Olfactory receptor 8B8 OR8B8 ortholog	OLFACTORY RECEPTOR 8B8 (PTHR26452:SF214)	-	Homo sapiens	EH37E0245422 EH37E0245423 EH37E0245424 EH37E0245425
	5.	HUMAN|HGNC=7758|UniProtKB=Q99519	Q99519	Sialidase-1 NEU1 ortholog	SIALIDASE-1 (PTHR10628:SF21)	hydrolase	Homo sapiens	EH37E0835732 EH37E0835733 EH37E0835734 EH37E0835735
	6.	HUMAN|HGNC=847|UniProtKB=P18859	P18859	ATP synthase-coupling factor 6, mitochondrial ATP5J ortholog	ATP SYNTHASE-COUPLING FACTOR 6, MITOCHONDRIAL (PTHR12441:SF10)	-	Homo sapiens	37724 EH37E0612229 EH37E0612232 EH37E0612632

b)

clr all		Gene ID	Mapped IDs	Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class	Species	Enhancer
	1.	HUMAN|HGNC=1192|UniProtKB=Q9H2W6	15:88661739	39S ribosomal protein L46, mitochondrial MRPL46 ortholog	39S RIBOSOMAL PROTEIN L46, MITOCHONDRIAL (PTHR13124:SF12)	-	Homo sapiens	EH37E0384154
	2.	HUMAN|HGNC=3575|UniProtKB=O95864	11:61746291	Fatty acid desaturase 2 FADS2 ortholog	FATTY ACID DESATURASE 2 (PTHR19353:SF12)	-	Homo sapiens	EH37E0221484
	3.	HUMAN|HGNC=10963|UniProtKB=O15245	6:160071159	Solute carrier family 22 member 1 SLC22A1 ortholog	SOLUTE CARRIER FAMILY 22 MEMBER 1 (PTHR24064:SF291)	-	Homo sapiens	EH37E0879369
	4.	HUMAN|HGNC=25477|UniProtKB=Q7L5Y6	15:88661739	DET1 homolog DET1 ortholog	DET1 HOMOLOG (PTHR13374:SF3)	ubiquitin-protein ligase	Homo sapiens	EH37E0384154
	5.	HUMAN|HGNC=13771|UniProtKB=Q9BQB4	17:43714850	Sclerostin SOST ortholog	SCLEROSTIN (PTHR14903:SF4)	-	Homo sapiens	EH37E0434745
	6.	HUMAN|HGNC=189|UniProtKB=O75078	17:44865603	Disintegrin and metalloproteinase domain-containing protein 11 ADAM11 ortholog	DISINTEGRIN AND METALLOPROTEINASE DOMAIN-CONTAINING PROTEIN 11 (PTHR11905:SF114)	metalloprotease	Homo sapiens	EH37E0435207

Conclusions

PANTHER development has been focused in two main directions, and we expect these trends to continue in the future. The first direction is the continual increase in the coverage of protein-coding gene diversity across the tree of life. In the past decade, the set of PANTHER reference trees has expanded from ~6500 families covering 48 fully-sequenced genomes in PANTHER version 7[24], to ~15,500 families covering 142 genomes in the current version. At the same time, we have continued to increase the quality of families, refining family boundaries and improving the phylogenetic trees based on ongoing review and semi-automated curation. PANTHER has become a suitable tool for analysis of lists of genes, and full genomes, from organisms across the tree of life.

The second direction of development is the increase in coverage and specificity of the annotations. PANTHER has expanded from family, subfamily and Protein Class annotations, to including GO terms and pathways, and now human enhancer regions. Specificity of annotation has dramatically improved. PANTHER was initially based on identifying subtrees containing genes of clearly diverged functions, and constructing HMMs that distinguished different subfamilies[1]. However, this approach becomes difficult to scale as annotations become more specific, and are confined to ever-smaller subtrees. Since 2010, individual nodes in PANTHER trees have been annotated with gain and loss of function annotations[14], enabling functional conservation and divergence to be represented at whatever level of detail is required, on a case-by-case basis. Functions are annotated using GO terms, as well as signaling and metabolic pathways. Using TreeGrafter, these fine-grained annotations can be transferred to sequences that are not in the reference tree by grafting directly onto a specific branch in the tree, rather than using subfamily HMMs to identify an approximate clade within the tree. Currently, TreeGrafter uses the tree graft point to infer the PANTHER subfamily, and PANTHER GO-slim annotations. In the future, we expect that tree grafting will also enable additional applications of the reference tree, including identifying the taxonomic group to which a sequence may belong, and inferring its orthologs in other organisms.

Acknowledgements

The authors want to thank the GO Phylogenetic Annotation curators for QA of the phylogenetic trees and family members: Marc Feuermann, Michael Kesling, Pascale Gaudet, Karen Christie, Donghui Li. We thank the UniProt team for guidance on using the Reference Proteomes as well as for QA of family members: Maria Martin, Dushyanth Jyothi, ThankGod Ebenezer, Herrmann Zellner, Alistair MacDougall, Gonzalo Nunez. We thank the Phylogenes team at Phoenix Bioinformatics for selecting plant genomes and for help with requirements and testing of web services: Eva Huala, Peifen Zhang, Tanya Berardini, Leonore Reiser.

Funding

National Science Foundation [1458808; 1661543; 1917302]; National Human Genome Research Institute of the National Institutes of Health [U41HG002273]; National Cancer Institute of the National Institutes of Health [P01CA196569]. Funding for open access charge: National Institutes of Health and National Science Foundation.

References

1. 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.
2. 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**, W645-650.
3. Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X. and Thomas, P.D. (2019) Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc*, **14**, 703-721.
4. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.
5. The Gene Ontology Consortium. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*, **47**, D330-D338.
6. Mi, H. and Thomas, P. (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol*, **563**, 123-140.
7. Funahashi, A., Tanimura, N., Morohashi, M. and Kitano, H. (2003) **CellDesigner: a process diagram editor for gene-regulatory and biochemical networks**. *BIOSILICO*, **1**, 159-162.
8. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res*, **48**, D498-D503.
9. Mitchell, A.L., Attwood, T.K., Babbitt, P.C., Blum, M., Bork, P., Bridge, A., Brown, S.D., Chang, H.Y., El-Gebali, S., Fraser, M.I. *et al.* (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res*, **47**, D351-D360.
10. Zdobnov, E.M. and Apweiler, R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847-848.
11. Tang, H., Finn, R.D. and Thomas, P.D. (2019) TreeGrafter: phylogenetic tree-based annotation of proteins with Gene Ontology terms and other annotations. *Bioinformatics*, **35**, 518-520.
12. 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.
13. Feuermann, M., Gaudet, P., Mi, H., Lewis, S.E. and Thomas, P.D. (2016) Large-scale inference of gene function through phylogenetic annotation of Gene Ontology terms: case study of the apoptosis and autophagy cellular processes. *Database (Oxford)*, **2016**.

14. 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.
15. Consortium, U. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, **47**, D506-D515.
16. Mi, H., Muruganujan, A., Ebert, D., Huang, X. and Thomas, P.D. (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*, **47**, D419-D426.
17. Burge, S., Kelly, E., Lonsdale, D., Mutowo-Muellenet, P., McAnulla, C., Mitchell, A., Sangrador-Vegas, A., Yong, S.Y., Mulder, N. and Hunter, S. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database (Oxford)*, **2012**, bar068.
18. Chen, C., Natale, D.A., Finn, R.D., Huang, H., Zhang, J., Wu, C.H. and Mazumder, R. (2011) Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One*, **6**, e18910.
19. 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**, D336-342.
20. Wingender, E., Schoeps, T., Haubrock, M., Krull, M. and Dönitz, J. (2018) TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res*, **46**, D343-D347.
21. Saier, M.H., Reddy, V.S., Tsu, B.V., Ahmed, M.S., Li, C. and Moreno-Hagelsieb, G. (2016) The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Res*, **44**, D372-379.
22. C, P. (2013) In A., B., Q., S. and F, D. (eds.), *Web Services Foundations*. Springer, New York, NY.
23. Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D. and Thomas, P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*, **45**, D183-D189.
24. Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S. and Thomas, P.D. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res*, **38**, D204-210.