D336–D342 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 17 November 2015*
doi: 10.1093/nar/gkv1194

PANTHER version 10: expanded protein families and functions, and analysis tools

Huaiyu Mi*, Sagar Poudel, Anushya Muruganujan, John T. Casagrande and Paul D. Thomas*

Division of Bioinformatics, Department of Preventive Medicine, Keck School of Medicine of USC, University of Southern California, Los Angeles, CA 90089, USA

Received September 28, 2015; Revised October 22, 2015; Accepted October 23, 2015

ABSTRACT

PANTHER (Protein Analysis THrough Evolutionary Relationships, http://pantherdb.org) is a widely used online resource for comprehensive protein evolutionary and functional classification, and includes tools for large-scale biological data analysis. Recent development has been focused in three main areas: genome coverage, functional information ('annotation') coverage and accuracy, and improved genomic data analysis tools. The latest version of PANTHER, 10.0, includes almost 5000 new protein families (for a total of over 12 000 families), each with a reference phylogenetic tree including protein-coding genes from 104 fully sequenced genomes spanning all kingdoms of life. Phylogenetic trees now include inference of horizontal transfer events in addition to speciation and gene duplication events. Functional annotations are regularly updated using the models generated by the Gene Ontology Phylogenetic Annotation Project. For the data analysis tools, PANTHER has expanded the number of different 'functional annotation sets' available for functional enrichment testing, allowing analyses to access *all* Gene Ontology annotations—updated monthly from the Gene Ontology database—in addition to the annotations that have been inferred through evolutionary relationships. The Prowler (data browser) has been updated to enable users to more efficiently browse the entire database, and to create custom gene lists using the multiple axes of classification in PANTHER.

INTRODUCTION

PANTHER is a comprehensive database of evolutionary and functional information about protein-coding genes. Protein sequences (one representative sequence per gene) from 104 complete genomes (PANTHER uses the UniProt Reference Proteomes data set (1)) are organized into families of homologous genes. For each family, a phylogenetic tree is constructed, representing the evolutionary relationships between all genes in the family, and the processes by which these genes were first copied and then diverged from each other: speciation, gene duplication and horizontal transfer (2,3). In speciation, a gene is copied when an ancestral population separates into two or more isolated populations. In gene duplication, a gene is copied back into its genome of origin. In horizontal transfer—unlike speciation and duplication, which are 'vertical' processes of inheritance from parents to offspring—a gene is copied into an organism that is not a direct offspring. PANTHER trees attempt to infer all of these processes, for nearly all protein-coding genes across a broad phylogenetic range of organisms.

Grouping genes into families and reconstructing a family tree is only the starting point for classification. The phylogeny is used to identify *subfamilies* within each family—groups of genes that share a particularly high degree of similarity due to limited divergence from their common ancestor. It is also used to infer *functions* shared among related genes. Starting in 2011, the PANTHER families have been used in the Gene Ontology (GO) Phylogenetic Annotation project (4). This project uses the gene family phylogeny to integrate disparate experimentally-derived GO functional annotations across related genes. For each family, a curator creates an explicit model of gain and loss of functions over specific branches of the tree, such that the model matches the experimental data. The model can then be used to infer functions of uncharacterized genes. This representation is a strict evolutionary formalization of homology-based 'annotation transfer' (5).

These inferred functional annotations are of vast practical importance, particularly for interpretation of genomic data sets. For human genes, the GO Phylogenetic Annotation project has more than doubled the number of GO annotations. The annotations—computable functional information over thousands of genes in a typical genome—enable biological interpretation of large-scale experimental data sets, so-called 'omics' data. The PANTHER resource has provided online functional-class en-

*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

richment analysis tools since 2003 (6–9), and these tools continue to increase in usage.

PANTHER has been under continual development since 1998. The overarching structure and methods employed has been described recently (10), and here we focus on the improvements made to the resource since that time.

PROTEIN FAMILY AND PHYLOGENETIC TREE IMPROVEMENTS

More genomes

PANTHER version 8.0 included 82 genomes in the phylogenetic trees. Based on feedback from the user community, PANTHER 10.0 has been expanded to 104 genomes. All genomes are listed at the PANTHER statistics page (http://www.pantherdb.org/panther/summaryStats.jsp). The 22 new genomes are listed in Table 1. The current coverage of different phylogenetic clades is shown in Figure 1. All kingdoms of life are represented, but deuterostome animals (including mammals) are overrepresented, as the highest usage of PANTHER remains in human biomedical research. The increased coverage of plants in PANTHER 10.0, performed in collaboration with the Phytozome resource (11) is intended to better support genomic analysis in agricultural applications.

Improved families

The PANTHER families have been improved in both quality and coverage. In order to improve the quality of the trees, two separate processes were performed. First, families containing primarily sequence fragments were identified systematically, and removed, resulting in the removal of ∼550 families between PANTHER versions 8 and 9. Non-fragments in these families were merged into existing families, or into new families as described below. Second, between versions 9 and 10, some families were identified as homologous to another family based on feedback from the PANTHER user community, including the GO Phylogenetic Annotation Project (see below), resulting in ∼250 families that were merged into other, larger families.

In order to improve the coverage of each of the proteomes, new families were added. In order to define new families, protein sequences (from the 104 proteomes) that did not match any of the existing PANTHER family or subfamily hidden Markov models (HMMs) were clustered using the TRIBE-MCL algorithm (12). The TRIBE-MCL inflation parameter was set so as to maximize family size (-I 5.0). A new family was created for each cluster that met the following two conditions: (i) it contains more than five genes in total, (ii) it contains genes from more than one organism and (iii) at least one of the genes is from 30 relatively well-studied organisms (Supplemental Table S1). A total of 4981 new families were added, comprising over 106 000 genes. On average, the new families increase proteome coverage by about 11.5%, with plants and bacteria having the largest increase (*A. thaliana* coverage was increased by 20%, and *E. coli* by 23%). Table 2 lists the coverage before and after adding the new families, for the 12 model organism species. While space restrictions limit Table 2 to only 12 genomes, the new families increase coverage of all 104 proteomes in PANTHER (full statistics are available at http://www.pantherdb.org/panther/summaryStats.jsp).

Improved phylogenetic trees

For PANTHER version 10, the GIGA algorithm (13) has been updated to include inference of horizontal transfer events. The details of the methodology will be described in detail elsewhere (manuscript in prep.). In brief, at each step in tree building the algorithm considers the number of gene deletions that would be implied by a history of vertical inheritance, and if that number is too large, a horizontal transfer event is considered to be the more likely interpretation. Estimation of gene loss is only possible because the algorithm uses a known species tree (Figure 1). The basic idea was proposed by Mirkin *et al.* (14), although the specific implementation in GIGA is novel in that it is assessed at each step of the tree reconstruction process.

The addition of horizontal transfer was necessary given the dramatic expansion in the number of bacterial species in PANTHER version 10 (Table 1), as horizontal transfer is common among bacteria and neglecting it would lead to incorrect evolutionary inferences. We also note that events such as the engulfment of the mitochondrion by the proto-eukaryotic cell will be represented at the level of gene family trees as horizontal transfer from a proteobacterial ancestor to the eukaryotic common ancestor. Indeed, this is the single most common pattern of transfer observed in the PANTHER trees. Overall, there are 3317 horizontal transfer events in 1785 families. PTHR24220, an ATP-transporter found predominantly in prokaryotes, has the highest number (36) of inferred horizontal transfer events.

Consistent subfamily identification

Prior to version 9.0, PANTHER subfamilies were based on review and curation by expert biologists. Subfamily definition has historically, both in PANTHER and in the broader scientific literature, been a subjective procedure. However, now that we can reconstruct the evolutionary events that created a gene family (speciation, gene duplication, horizontal transfer) with reasonable accuracy, it is possible to define subfamilies consistently across families, in a way that still reflects the underlying intuition that has guided subfamily identification both in the scientific literature and in earlier versions of PANTHER. The basic idea is simple: subfamilies can be defined by their evolutionary histories. In the current version of PANTHER, subfamilies are, in general, closely-related orthologs. The precise definition is that a new subfamily is created within a family after every gene duplication event, or horizontal transfer event (i.e. *the creation of a new gene locus*). After horizontal transfer, the transferred copy becomes the founder of a new subfamily; the vertically inherited copy remains in the original subfamily. After gene duplication, the copy that changes faster in sequence immediately following the duplication becomes the founder of a new subfamily; the slower-evolving copy remains in the same subfamily. There are two exceptions to this rule: (i) because of the high frequency of gene duplication prior to the vertebrate common ancestor, each vertebrate copy following a gene duplication event founds a new

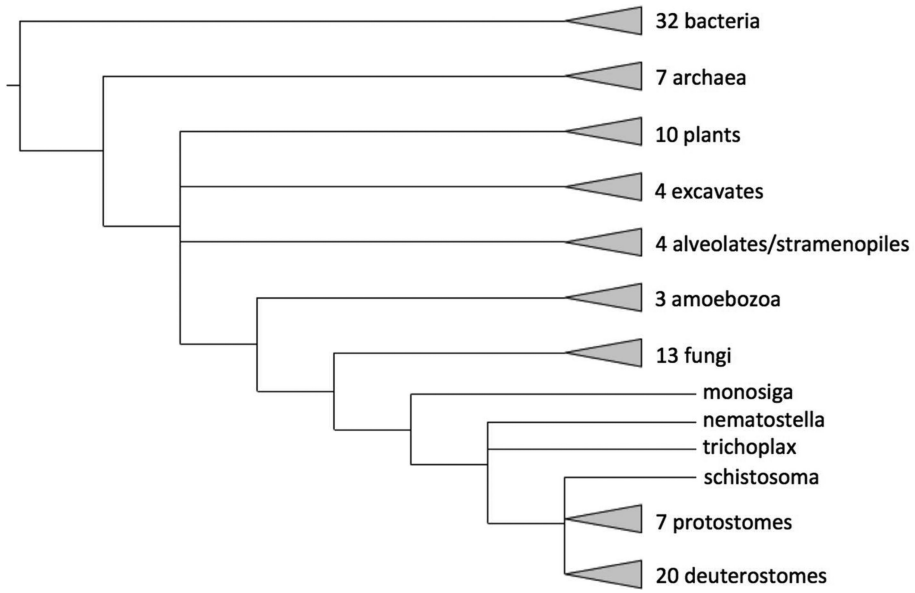

Figure 1. Phylogenetic distribution of genomes available in PANTHER 10. The tree is derived from information in the NCBI Taxonomy resource (17), with some multifurcating nodes resolved using the sTOL resource (18).

Table 1. Additional genomes added to PANTHER 9.0 and 10.0

Bacteria	Plants	Animals (Tetrapods)
Bacillus cereus	Purple false brome (*Brachypodium distachyon*)	Lizard (*Anolis carolinensis*)
Clostridium botulinum	Soybean (*Glycine max*)	Horse (*Equus caballus*)
Coxiella burnetii	Black cottonwood (*Populus trichocarpa*)	Cat (*Felis catus*)
Haemophilus influenzae	Tomato (*Solanum lycopersicum*)	Pig (*Sus scrofa*)
Listeria monocytogenes	Sorghum (*Sorghum bicolor*)	
Salmonella typhimurium	Grape (*Vitis vinifera*)	
Shewanella oneidensis		
Staphylococcus aureus		
Streptococcus pneumoniae		
Vibrio cholerae		
Xanthomonas campestris		
Yersinia pestis		

Table 2. Increase in PANTHER version 10 coverage of protein-coding genes in major model organisms

Organism	# genes before new families	# genes in new families	Total # genes in PANTHER 10 families	Total # in proteome	Percent coverage
Human	18390	794	19184	20814	92%
Mouse	20085	722	20807	22320	93%
Rat	20622	749	21371	22656	94%
Chicken	13854	404	14258	15696	91%
Zebrafish (*D. rerio*)	21836	639	22475	25357	89%
Fruit fly (*D. melanogaster*)	10447	340	10787	13690	79%
Nematode (*C. elegans*)	14004	764	14768	20490	72%
Budding yeast (*S. cerevisiae*)	4370	384	4754	6597	72%
Fission yeast (*S. pombe*)	4038	323	4361	5101	85%
Slime mold (*D. discoideum*)	8665	576	9241	13051	71%
Arabidopsis (*A. thaliana*)	19396	3858	23254	26684	87%
E. coli	2695	628	3323	4260	78%
Total (104 genomes)	920552	105869	1026421	1424953	72%

subfamily, and (ii) because a subfamily should contain at least two genes, duplicated genes do not found a subfamily if they did not lead to orthologs in at least two extant species.

Once a subfamily has been defined, it is given a stable identifier and name. Subfamily identifiers are tracked forward from previous versions of PANTHER based on the stable node identifier of the ancestral 'founding member' of the subfamily (i.e. a speciation node immediately following a gene duplication or horizontal transfer event). In keeping with the original tradition of naming PANTHER subfamilies after a representative subfamily member from the Swiss-Prot database (15), we have implemented the following procedure for choosing the representative subfamily member. If a subfamily contains an annotated Swiss-Prot entry from any of the 12 model organisms, then the curated 'protein name' is used to name the subfamily (if there is more than one Swiss-Prot entry in the subfamily, the one with highest precedence, as listed in Table 2, is selected). If a subfamily does not contain any Swiss-Prot entries from the 12 model organisms, but contains Swiss-Prot entries from other organisms, the most common Swiss-Prot protein name is used as the subfamily name. If a subfamily does not contain any members from the Swiss-Prot database, a protein name from the TrEMBL entry is automatically selected as the subfamily name. It should be noted that these TrEMBL names are not curated so they may not necessarily be meaningful. An attempt is made to select a meaningful name by excluding names with terms like 'unknown,' 'uncharacterized,' etc., and if no such name exists the subfamily is labeled with 'unnamed.'

ANNOTATION DATA IMPROVEMENTS

Updated homology-inferred function annotations

The motivation behind PANTHER was prediction of gene function using homology, and a large number of new and revised function annotations have been added to PANTHER version 10. These updated annotations are taken from the Gene Ontology Phylogenetic Annotation Project, described in Gaudet *et al.* (4). In brief, the project combines GO experimental annotations and a PANTHER phylogenetic tree to infer gene functions. For each gene family, a biologist curator reviews all experimentally-supported GO annotations, and then builds a model of gain and loss for each relevant, distinct function (represented by a GO term) over evolutionary time. In this way, the sparse experimental information about a few genes in a family is used to create a family-wide model of function evolution that is consistent with the available data. The model of gain and loss events allows inference (by inheritance through a tree) of functions for uncharacterized genes.

In the current PANTHER 10.0 release, GO annotations have been updated for 795 families (including 9457 subfamilies). These have resulted in almost 640 000 annotations total for approximately 130 000 protein-coding genes.

Table 3 shows the number of genes from different model organism genomes that received an updated phylogenetic-based annotation. In PANTHER version 10.0, ∼14.5% of human genes have received at least one annotation through this process, reflecting the progress of the GO Phylogenetic Annotation Project as of early 2015. Based on the latest statistics from the GO Phylogenetic Annotation Project (25% of human genes covered as of September 2015) and current annotation rate, we expect that at least 40% of human genes will have received updated and expanded function annotations in PANTHER version 11, which is planned for release in 2016.

WEBSITE AND ANALYSIS TOOL IMPROVEMENTS

Additional GO annotation sets

PANTHER has long provided predicted GO function annotations, and, as described above, many of these have been updated in PANTHER version 10. On the PANTHER website, there are *gene list analysis tools* that statistically analyze any user-specified gene list (optionally including quantitative measurements) using the predicted GO function annotations in the database. Initially released publicly in 2003 (6,15), these tools have been widely used and cited in the scientific literature, and are the most commonly used pages within PANTHER. One of the major advantages of our tools, compared to other similar resources, is close collaboration with the GO Consortium to ensure that the GO annotation data is correct, and kept up to date with the constantly expanding corpus of annotations generated by the Consortium. Also, the user interface was designed specifically to enable researchers with little computer background to use these tools, and to accurately report information necessary for other groups to reproduce the analysis.

In order to better support our users in their gene set analysis, the PANTHER website now allows users to access all GO annotations generated by the GO Consortium, in addition to the annotations predicted from phylogenetic inference over the trees in PANTHER. Users can now select from a number of 'annotation sets' when using either the PANTHER overrepresentation tool or the PANTHER enrichment tool (see Mi *et al.* (9) for more details on these tools). Put simply, annotation sets differ in the method by which they were generated. Currently, users can choose between 'PANTHER' sets (inferred by homology to experimentally characterized genes), 'GO Experimental' (taken directly from published papers), or 'GO Complete' (including experimental as well as computationally inferred by several methods). Table 4 lists the distinct sets, along with a brief description of each set. The additional sets are downloaded monthly directly from the Gene Ontology database (http://www.geneontology.org).

The user interface for the PANTHER gene list analysis tools has been improved to facilitate use of the different GO annotation sets, and to help users to correctly report the tool settings in any publications (Figure 2). At the top of the analysis results page, all tool settings (including the annotation set) are listed, and users can make changes directly from that page and launch a new analysis, rather than having to return to the homepage. Reporting of tool settings is extremely important in order to ensure reproducibility of results, as the gene list analysis results can depend on the annotation set that is used. As these sets are constantly being expanded and improved, each set name also specifies the version number and/or release date.

Table 3. Number of new GO annotations added to PANTHER 10 by the GO Phylogenetic Annotation Project, for selected genomes

Organism	Number of genes with new annotations	Number of annotations[a]
Human	2936	15731
Mouse	3171	16938
Rat	3242	17351
Chicken	2157	11855
Zebrafish (*D. rerio*)	3505	19003
Fruit fly (*D. melanogaster*)	1358	7024
Nematode (*C. elegans*,)	1840	9445
Budding yeast (*S. cerevisiae*)	718	3561
Fission yeast (*S. pombe*)	700	3548
Slime mold *(D. discoideum*)	999	4923
Arabidopsis *(A. thaliana)*	2263	10644
E. coli	298	1055

[a]An annotation is the combination of a gene and a GO term (counting only those with the most specific term in the ontology graph)

Table 4. Annotation sets available for PANTHER gene list analysis tools

Annotation gene set	Description
PANTHER Pathway	Signaling and metabolic pathways manually curated by experts (16)
PANTHER GO-slim Molecular Function	Phylogenetically inferred annotations, to a reduced set of GO molecular function terms (GO-slim).
PANTHER GO-slim Biological Process	Phylogenetically inferred annotations, to a reduced set of GO biological process terms (GO-slim).
PANTHER GO-slim Cellular Component	Phylogenetically inferred annotations, to a reduced set of GO cellular component terms (GO-slim).
PANTHER Protein Class	PANTHER/X ontology terms to classify protein families and subfamilies. These classes are grouping terms commonly used by biologists for families and subfamilies, that are sometimes but not always related to molecular function.
GO molecular function experimental only	GO molecular function annotations, from the GO database, based on experimental evidence published in scientific publications. These annotations are usually manually curated by GO curators.
GO biological process experimental only	GO biological process annotations, from the GO database, based on experimental evidence published in scientific publications. These annotations are usually manually curated by GO curators.
GO cellular component experimental only	GO cellular component annotations, from the GO database, based on experimental evidence published in scientific publications. These annotations are usually manually curated by GO curators.
GO molecular function complete	Complete GO molecular function annotations including both manually curated (above) and electronic annotations. Electronic annotations are generated by a variety of different computer algorithms.
GO biological process complete	Complete GO biological process annotations including both manually curated (above) and electronic annotations. Electronic annotations are generated by a variety of different computer algorithms.
GO cellular component complete	Complete GO cellular component annotations including both manually curated (above) and electronic annotations. Electronic annotations are generated by a variety of different computer algorithms.

Prowler in JavaScript

The PANTHER Prowler, a web interface for rapidly browsing the families and genes in the database (by any combination of the classification axes—protein class, pathway, species, molecular function, cellular component, and biological process) has been updated to HTML5 compliant JavaScript (Figure 3). This is a critical update, as the previous implementation was as a client Java applet, which is no longer properly supported by many web browsers.

CONCLUSIONS

PANTHER continues to improve in all facets of the resource. On the data side, >20 genomes have been added, as well as almost 5000 additional protein families, greatly expanding the representation of plant and bacterial genomes. The phylogenetic trees have been improved by adding inference of horizontal transfer events. Gene Ontology function annotations have been expanded and revised.

On the online tools side, the gene list analysis tools have been expanded to access GO annotations not only from PANTHER (phylogenetic inference), but also from the GO database (experimental annotations, as well as computational inferences by numerous methods). These annotations are updated every month, to keep pace with the constant addition of new information by the multiple annotation projects within the Gene Ontology Consortium. Users can perform genome analysis, including gene list analysis, on all 104 genomes in the database. The Prowler, the PANTHER data browser, has also been updated to be HTML5 compliant, enabling users to efficiently browse, and create their own custom gene lists from, the over 2 million genes now classified in PANTHER.

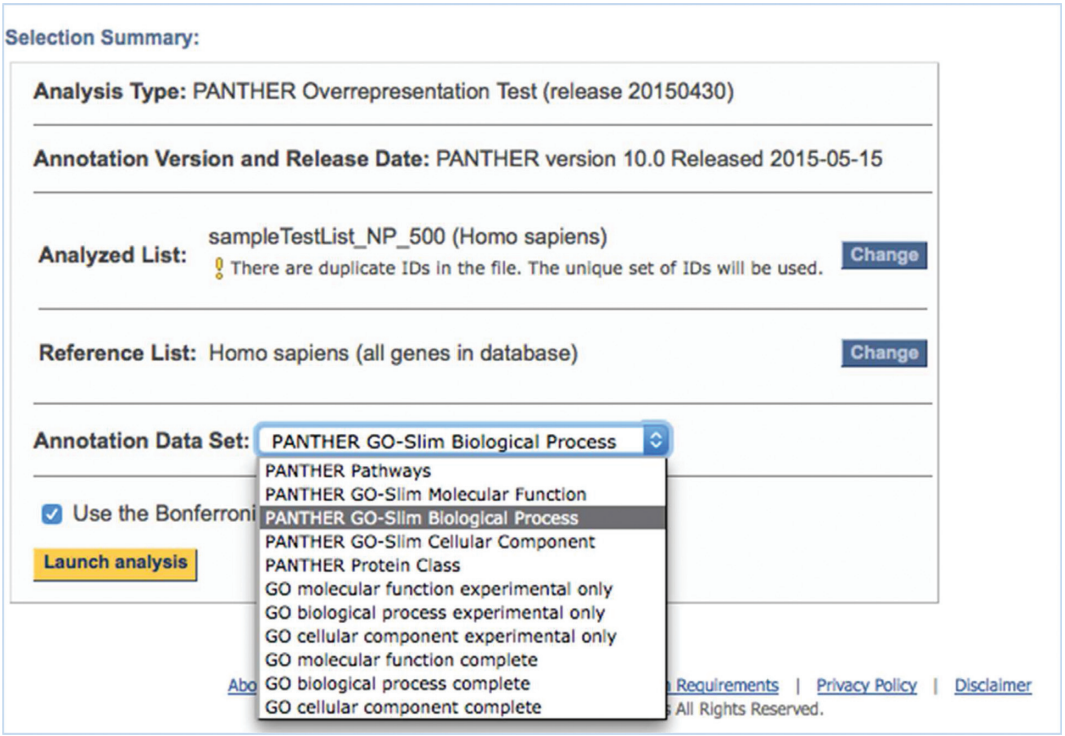

Figure 2. Screenshot of the settings summary of the gene list analysis tool. Settings can be easily changed from this summary table, and results rapidly regenerated using the new settings.

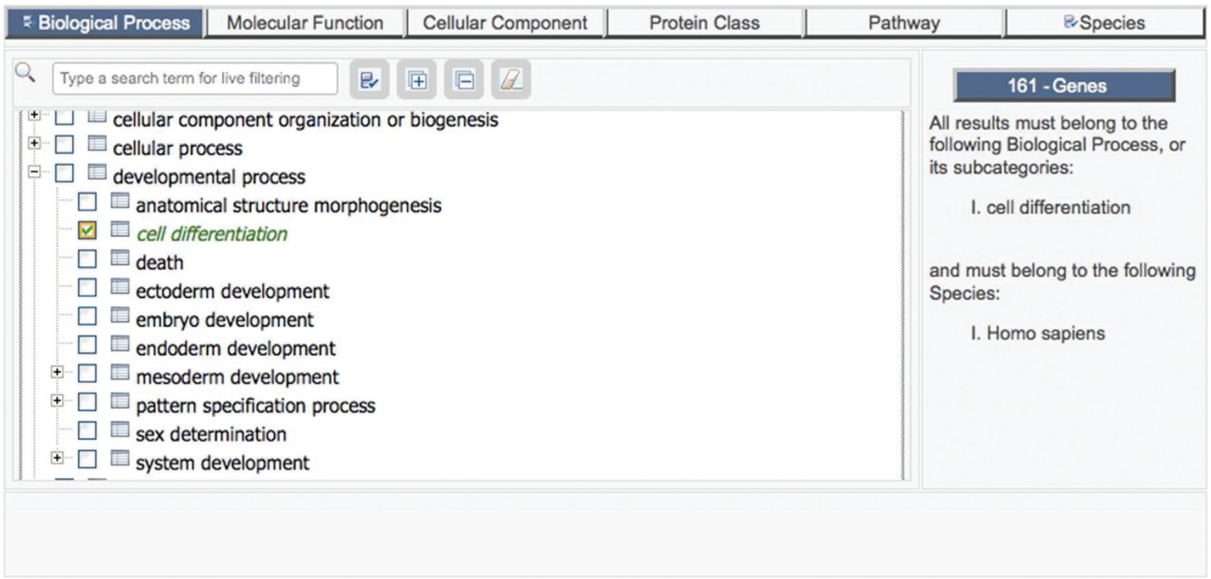

Figure 3. New Prowler for browsing the different classification axes, and retrieving genes from class intersections. Here, a user first selected a species (Homo sapiens) and then the biological process cell differentiation, and pressing the button on the right will retrieve from the database all 161 human genes annotated as involved in cell differentiation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors want to acknowledge the contributions of the GO Phylogenetic Annotation curators: Marc Feuermann, Pascale Gaudet, Karen Christie, Donghui Li.

FUNDING

NHGRI (NIH) [U41HG002273], and NIGMS (NIH) [U24GM088849]. Funding for open access charge: NHGRI (NIH) [U41HG002273].
Conflict of interest statement. None declared.

REFERENCES

1. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
2. Koonin,E.V. and Galperin,M.Y. (2003) Evolutionary Concept in Genetics and Genomics. In: *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Kluwer Academic, Boston.
3. Thomas,P.D. (2006) PANTHER: Protein families and subfamilies modeled on the divergence of function. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, Wiley, Chichester, pp. 1–5.
4. 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.
5. Eisen,J.A. (1998) A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res.*, **26**, 4291–4300.
6. Clark,A.G., Glanowski,S., Nielsen,R., Thomas,P.D., Kejariwal,A., Todd,M.A., Tanenbaum,D.M., Civello,D., Lu,F. *et al.* (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–1963.
7. Mi,H., Lazareva-Ulitsky,B., Loo,R., Kejariwal,A., Vandergriff,J., Rabkin,S., Guo,N., Muruganujan,A., Doremieux,O. *et al.* (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, **33**, D284–D288.
8. 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.
9. 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.
10. Mi,H., Muruganujan,A. and Thomas,P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.
11. Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
12. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
13. Thomas,P.D. (2010) GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics*, **11**, 312.
14. Mirkin,B.G., Fenner,T.I., Galperin,M.Y. and Koonin,E.V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.*, **3**, 2.
15. 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.
16. Mi,H. and Thomas,P. (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.*, **563**, 123–140.
17. NCBI Resource Coordinators (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **42**, D7–D17.
18. Fang,H., Oates,ME., Pethica,R.B., Greenwood,J.M., Sardar,A.J., Rackham,O.J., Donoghue,P.C., Stamatakis,A., de Lima Morais,D.A. and Gough,J. (2013) A daily-updated tree of (sequenced) life as a reference for genome research. *Sci. Rep.*, doi:10.1038/srep02015.