PROTOCOL

Large-scale gene function analysis with the PANTHER classification system

Huaiyu Mi, Anushya Muruganujan, John T Casagrande & Paul D Thomas

Department of Preventive Medicine, Division of Bioinformatics, Keck School of Medicine, University of Southern California, Los Angeles, California, USA.
Correspondence should be addressed to H.M. (huaiyumi@usc.edu).

Published online 18 July 2013; doi:10.1038/nprot.2013.092

The PANTHER (protein annotation through evolutionary relationship) classification system (http://www.pantherdb.org/) is a comprehensive system that combines gene function, ontology, pathways and statistical analysis tools that enable biologists to analyze large-scale, genome-wide data from sequencing, proteomics or gene expression experiments. The system is built with 82 complete genomes organized into gene families and subfamilies, and their evolutionary relationships are captured in phylogenetic trees, multiple sequence alignments and statistical models (hidden Markov models or HMMs). Genes are classified according to their function in several different ways: families and subfamilies are annotated with ontology terms (Gene Ontology (GO) and PANTHER protein class), and sequences are assigned to PANTHER pathways. The PANTHER website includes a suite of tools that enable users to browse and query gene functions, and to analyze large-scale experimental data with a number of statistical tests. It is widely used by bench scientists, bioinformaticians, computer scientists and systems biologists. In the 2013 release of PANTHER (v.8.0), in addition to an update of the data content, we redesigned the website interface to improve both user experience and the system's analytical capability. This protocol provides a detailed description of how to analyze genome-wide experimental data with the PANTHER classification system.

INTRODUCTION

The PANTHER classification system is designed to be a comprehensive platform for the analysis of gene function on a genome-wide scale[1]. Although it initially aimed to classify gene and protein functions[2,3], this system has evolved through the years to also serve as an online resource for analyzing experimental data[4]. The easy-to-use user interface and timely user support have made PANTHER one of the most widely used online resources for gene function classification and genome-wide data analysis.

PANTHER was initially released in 2003, and it was the first database to combine both phylogenetic and functional data to define protein subfamilies of shared function and sequences[3]. It was also the first database to associate ontology terms describing function to statistical models (HMMs), which can be used to assign genes—on the basis of sequence information alone—to subfamilies and functional classes. The novel approach that PANTHER introduced was to annotate subfamilies of related genes that are likely to share function rather than single genes one at a time. We demonstrated the accuracy and comprehensiveness of the classifications obtained via PANTHER on the *Drosophila melanogaster* genome[5]. In 2005, we began providing annotations of biochemical pathways, which can be viewed via the new PANTHER pathway applet[6,7]. These pathways were curated from the literature by expert biologists, and diagrams were drawn with CellDesigner, a pathway-editing tool that uses controlled graphical notations to represent pathway knowledge[8].

In 2006, the first version of gene analysis tools was released[4]. These tools were primarily aimed at the analysis of gene expression data. The tools were also designed to handle gene list data from any genome-wide experiments. In the 2013 release of PANTHER 8.0, we re-designed the web interface to integrate multiple tools into one user-friendly and flexible interface, so that users can easily access all the tools and choose among them to perform gene list analysis at the genome-wide level[1].

Since the initial release in 2003, PANTHER has become one of the most popular online resources for genome-wide data analysis. Currently, it has over 7,000 registered users, 600–800 daily users (weekdays), and over 14,000 monthly users worldwide. These counts are based on unique Internet Protocol (IP) addresses accessing the site, and thus the actual number of individual users is probably even greater. PANTHER has been cited in over 3,600 publications since 2003 (according to Google Scholar) and is growing steadily. Users have successfully analyzed data from gene expression[9,10], proteomics[11,12] and genome-wide association study (GWAS) experiments[13,14] in diverse areas of research, such as cancer research[10,11], neurological disorder studies[15,16], studies of autoimmune diseases[9] and cardiac disease studies[14,17].

PANTHER is currently included in a number of large international consortia. PANTHER has been a member of the InterPro Consortium of protein annotation resources since 2005, and thus PANTHER annotations can be automatically generated from the frequently used InterProScan software[18]. More recently, PANTHER also became part of the GO Consortium[19–21]. The phylogenetic-inferred curation paradigm has become part of the GO curation pipeline, and therefore PANTHER annotation reflects a more up-to-date GO curation[22]. Finally, the PANTHER pathway is in the process of being integrated into Pathway Commons[23].

The PANTHER system is composed of three functional modules (**Fig. 1**). The core module is a large protein library that contains all protein-coding genes from 82 organisms, organized first into families on the basis of sequence homology, and then into subfamilies on the basis of their shared functions (often including sets of orthologous genes). Each family or subfamily is represented by a statistical model (HMM) annotated with ontology terms (GO and PANTHER protein class terms). The second module is the PANTHER pathway module, which contains 176 expert-curated pathways. All pathways are connected, through manual curation,

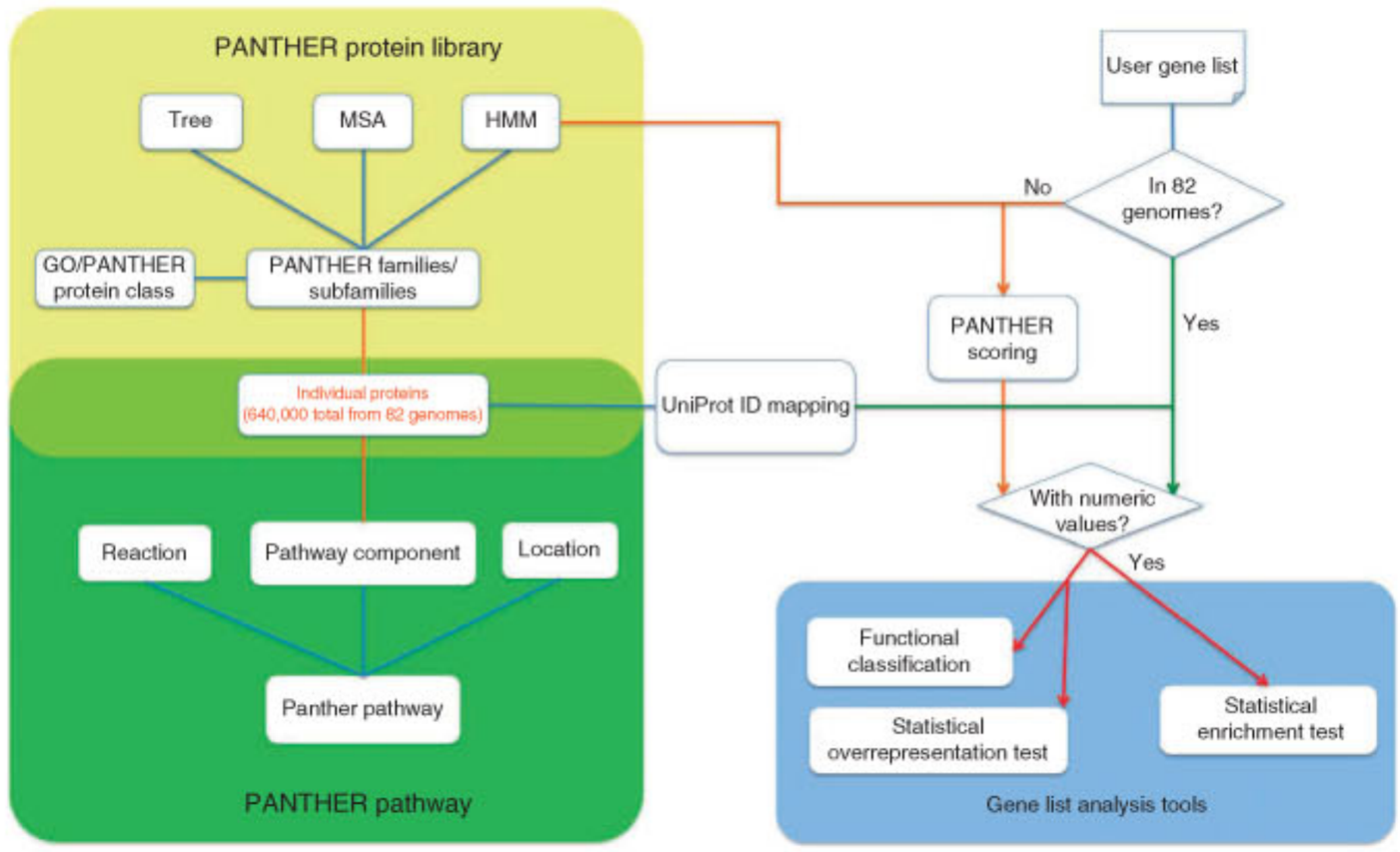

Figure 1 | Overview of PANTHER infrastructure. PANTHER consists of three modules. The core module is the PANTHER protein library (yellow), which contains genes from 82 complete genomes organized in PANTHER families and subfamilies, each of which is represented by a phylogenetic tree, an MSA (multiple sequence alignment) and an HMM. The second module is the pathway that contains 176 expert-curated pathways (dark green). The pathway components are associated with the protein sequences that are used to build the protein library (light green), and therefore the pathways are also linked to the subfamilies and HMMs. The third module is the tool suite. In this diagram, the gene list analysis tool is used as an example (blue). When the user uploads a gene list to the tool, and if the IDs in the list are from one of 82 organisms in PANTHER, the tool will map the IDs to the IDs in the PANTHER protein library (green arrows). If the uploaded IDs are not from one of the 82 organisms, the user can score the sequences against the PANTHER HMM library and generate the PANTHER Generic Mapping File (**Box 2**) (orange arrows). Three tests are included in the tool: functional classification, statistical overrepresentation test and statistical enrichment test. Numeric values from experimental results, such as raw readout, fold changes or *P* values, must be provided in order to implement the statistical enrichment test.

to individual proteins in the protein library through the pathway components, and therefore they are also linked to the phylogenetic information and statistical models. The last module is the website tool suite that contains a collection of bioinformatics tools and software, which enables users to not only query the data and classify genes and proteins but also to visualize, analyze and interpret genome-wide experimental data in the context of the enriched data content in the first two modules.

PANTHER protein library

The core of the PANTHER system is a collection of phylogenetically defined protein families and subfamilies generated by computational algorithms and curated by expert biologists using an extensive software system for associating ontology terms[1,3]. The current release contains over 640,000 proteins from 82 genomes, 79 of which are from the Reference Proteome Project (http://www.ebi.ac.uk/reference_proteomes/; see also **Fig. 1**). UniProt identifiers (IDs) are used as primary protein identifiers. These proteins are representatives of their respective genes. Therefore, each gene is represented by only one protein. In addition, UniProt IDmapping (http://www.uniprot.org/mapping/) is used to map the primary protein IDs to other IDs from different databases and resources, an approach that expands the capability of PANTHER to support a wider range of ID types (see Supported IDs in **Box 1**). The proteins are divided into 7,729 families, each of which is represented by a phylogenetic tree, an HMM and a multiple sequence alignment (**Fig. 1**).

Protein family trees are constructed computationally from sequence data using a phylogenetic tree inference algorithm

Box 1 | Input file format

The input file is a tab-delimited text file (.txt or .tab). Only the data in the columns specified below will be used in the analyses. Data in additional columns are ignored. Microsoft Excel files are not accepted by the tool. Given below are three file types that can be used.

ID list

The first column must be the gene or protein identifiers. See below for the supported IDs. A second column of numerical values is required if a user wants to run the statistical enrichment test.

Previously exported text search results

A text search result can be viewed as a gene list, which can be saved as a text file. This file contains the gene or protein identifiers in the first column. This file type is not associated with numeric values, and thus it cannot be used for the statistical enrichment test.

PANTHER Generic Mapping File

For IDs from organisms other than the 82 organisms in the PANTHER database, user-generated data containing mappings between those IDs and their corresponding PANTHER IDs can be used (see **Box 2**). The file must be tab-delimited and must contain the following columns: the first column can contain a list of unique IDs from the user; the second column should be the corresponding PANTHER family or subfamily ID (e.g., PTHR10078 or PTHR10078:SF6), and is used to look up the association with GO and PANTHER terms (molecular function, biological process and pathway). Note that if you are uploading data for the statistical enrichment test tool, a third column is required that contains the numeric value of the experiment.

(continued)

Box 1 | (continued)

Supported IDs

If the 'ID list' file type is used, the IDs in the first column of the file must be from one of the following databases that are supported in the PANTHER system.

Ensembl: Ensembl gene identifier. Example: 'ENSG00000126243'

Ensembl_PRO: Ensembl protein identifier. Example: 'ENSP00000337383'

Ensembl_TRS: Ensembl transcript identifier. Example: 'ENST00000391828'

Gene ID: EntrezGene IDs. Example: '10203' (for Entrez gene GeneID:10203)

Gene symbol: for example, 'CALCA'

GI: NCBI GI numbers. Example: '16033597'

HGNC: HUGO Gene Nomenclature ids. Example: 'HGNC:16673'

IPI: International Protein Index ids. Example: 'IPI00740702'

UniGene: NCBI UniGene ids. Examples: 'Hs.654587', 'At.36040'

UniProtKB: UniProt accession. Example: 'O80536'

UniProtKB-ID: UniProt ID. Example: 'AGAP3_HUMAN'

If you are not certain about the ID type in your uploaded gene list, or when you find that your IDs are not mapped to any PANTHER IDs in the result page, you can simply search your ID at NCBI (http://www.ncbi.nlm.nih.gov/) or a search engine website such as Google. You can find the ID type based on the database source on the result page.

called GIGA[24]. Nodes in the tree, corresponding to common ancestors of extant family members, are annotated by expert biologists with their inferred GO terms and PANTHER protein class terms, which are based on experiments performed on extant proteins. Subfamilies are determined on the basis of the tree structure and often define the orthologous group, especially in organisms in the Deuterostomia superphylum. The functional annotations are propagated from the annotated ancestral nodes to the subfamily nodes, each of which is also represented by an HMM to enable classification of newly discovered protein sequences (**Fig. 2**).

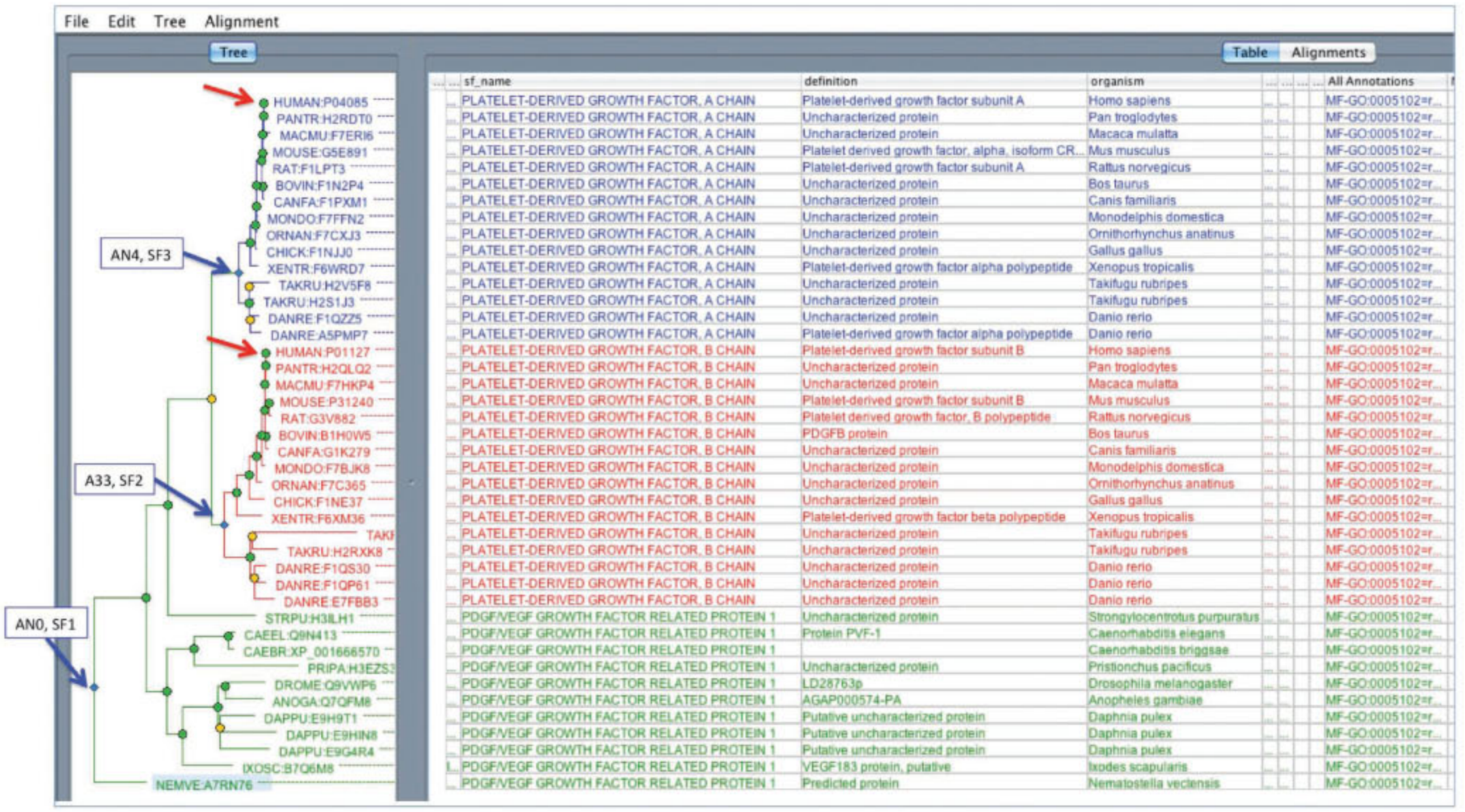

Figure 2 | Sample phylogenetic tree from PANTHER. Shown is PANTHER accession code PTHR11633, platelet-derived growth factor. The family contains three subfamilies (blue arrows). SF1 is annotated as 'PDGF/VEGF growth factor–related protein 1' based on the annotation in the *Drosophila* and *Caenorhabditis elegans* sequences (Uniprot accession codes Q9VWP6 and Q9N413, respectively). A recent duplication generates the PDGF A chain (SF3) and the PDGF B chain (SF2). Ontology terms are annotated to the node that represents the common ancestor in the extant family, in this case, AN0 (SF1). The classifications are propagated to all the descent nodes, including the AN4 (SF3) and AN33 (SF2). Please note that PDGF stands for platelet-driven growth factor and VEGF stands for vascular endothelial growth factor.

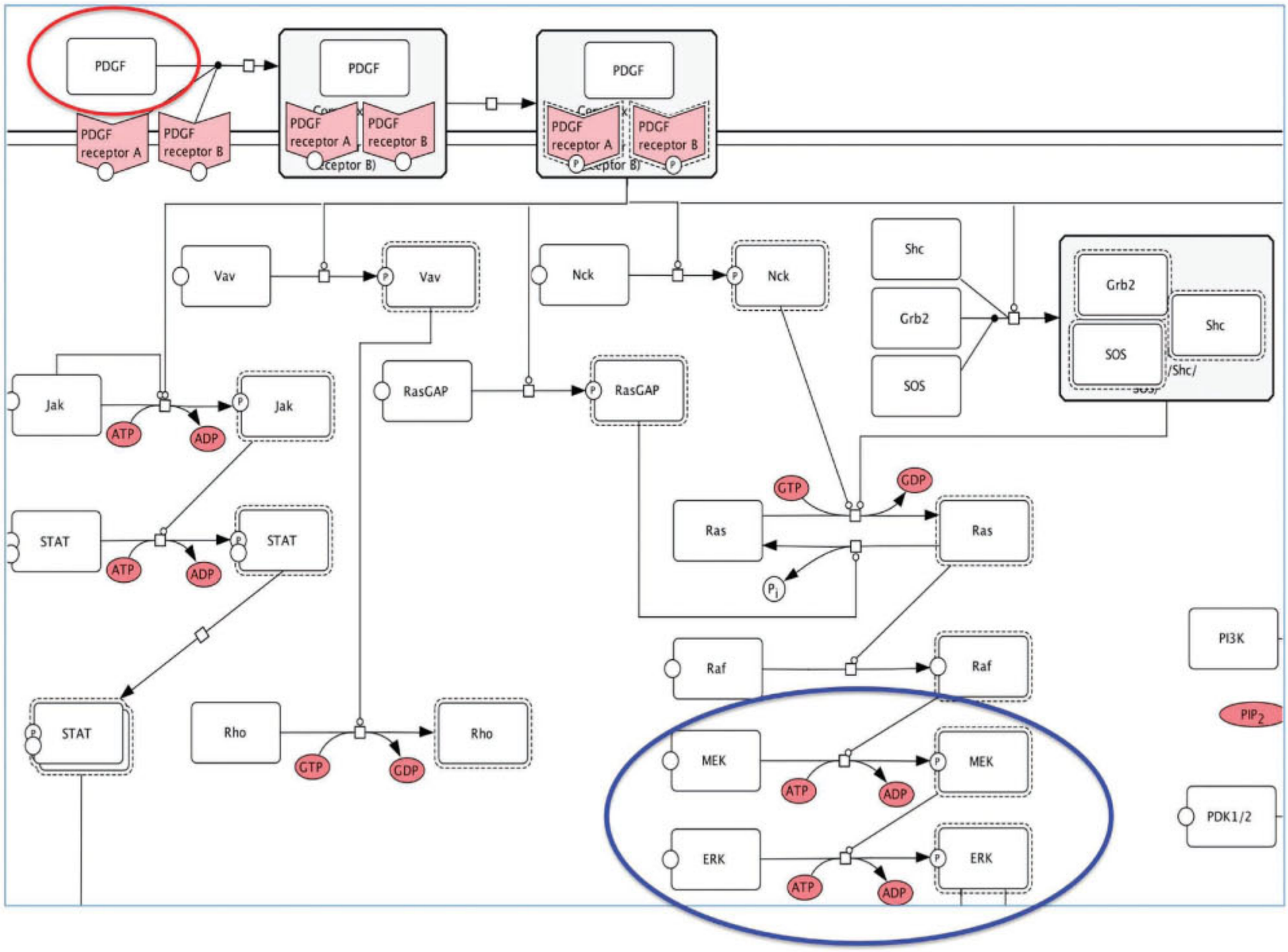

Figure 3 | Example of a PANTHER pathway diagram. The diagram is shown in a CellDesigner process diagram (PANTHER accession code P00047, PDGF signaling pathway), which is similar to the SBGN-PD format. For example (blue circle), a transition of an input (e.g., ERK) to an output (e.g., phosphorylated ERK) catalyzed by a modifier (e.g., phosphorylated MEK). A pathway component (e.g., PDGF in the red circle) is associated with the human protein sequences in the protein library (red arrows in **Fig. 2**) through expert curation. This association is supported by literature evidence. As a result, the pathway component of PDGF can be inferred to other orthologous protein sequences in the subfamilies in the library (SF2 and SF3 in **Fig. 2**).

In addition, the HMM library and the scoring pipeline enable users to analyze genome-wide data from any organisms outside of the 82 organisms within the PANTHER system. Phylogenetic trees and functional annotations at the ancestral nodes offer a great advantage in the annotation of previously unclassified genes. The PANTHER phylogenetic trees are currently used by the GO in its curation pipeline[22].

The PANTHER pathway

The PANTHER pathway data set uses controlled vocabulary and graphical notation to describe pathway knowledge[7]. The graphical representation of pathways is generated by CellDesigner, a pathway-editing tool[8], and is compliant with the Systems Biology Graphical Notation Process Description standard (SBGN-PD)[25] (**Fig. 3**). Currently, 176 expert-curated pathways are present in PANTHER. The scope of the pathways is similar to what has been described in textbooks or review articles; such pathways include, for instance, glycolysis, the platelet-derived growth factor (PDGF) signaling pathway or the p53 pathway. Each pathway contains three key classes of data (**Fig. 1**). First is the pathway component (or molecule class), which represents a specific class of molecules that has the same mechanistic role within a pathway. It can be a protein (e.g., PDGF, Jak), a DNA region or a simple molecule (e.g., ATP or glucose). If a pathway component is a protein, gene or transcribed RNA, it is associated with the protein sequences in the PANTHER protein library through manual curation (**Figs. 1** and **2**). The individual protein sequences are instances of the pathway component. In these cases, a pathway component is typically a group of homologous or orthologous proteins across various organisms, which is involved in the same biochemical reaction within the pathway. The second class of data relates to reaction, which represents biochemical relationships among different pathway components. As all PANTHER pathways are in compliance with SBGN-PD, a typical reaction is a transition of an input (or reactant) to an output (a product) controlled by a modifier (**Fig. 3**). The last class is the cell or tissue type and cellular component, which provides the location where the reaction occurs.

Two features are key to the PANTHER pathway data model. First, the association between the molecule classes and protein sequences links the pathway to the phylogenetic relationships and statistical

Box 2 | Construction of input files with data from organisms whose genome is not in the PANTHER data set

The PANTHER HMM scoring tool is the downloadable version of the PANTHER scoring tool, which enables users to submit a large number of protein sequences in .fasta file format and score them against the PANTHER HMM library, so that the sequence identifiers can be mapped to PANTHER HMM IDs and used in the gene list analysis tools.

▲ **CRITICAL** UNIX and Perl are required on your computer in order to use the tool. The user needs to have a basic knowledge of using UNIX and Perl in order to complete the procedures described herein. If you do not feel you have adequate knowledge in these areas, you may want to get help from a colleague who has the technical expertise and knowledge, such as a bioinformatics support person. You can also send an e-mail to feedback@pantherdb.org for help.

Software downloads ● TIMING ~30 min

1. Download the the pantherScore script (ftp://ftp.pantherdb.org//hmm_scoring/current_release/).
2. Download the PANTHER HMM library (ftp://ftp.pantherdb.org/panther_library/current_release/)
3. Download HMMER2 (ftp://selab.janelia.org/pub/software/hmmer/2.3.2/hmmer-2.3.2.tar.gz). Please note that this is an archived version of HMMER2. The current release is HGMMER3. The panther scoring script does not support HMMER3.
4. Download BLAST (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.24/ncbi-blast-2.2.24+-x64-linux.tar.gz)

Software decompression ● TIMING 30 min

5. Decompress all four downloads.

Installation of HMMER2 and BLAST ● TIMING ~30 min for experts, ~2 h for everyone else

6. Install HMMER2 and BLAST according to the installation instructions that come with the two algorithms.
7. Define the location of the HMMER and BLAST binaries in the $PATH variables on your computer. How these locations are defined usually differs quite substantially depending on the UNIX shell environment on the computer being used. Basically, experimenters need to update the $PATH on the UNIX shell files, such as .cshrc (for C shell) or .profile (for Bourne shell). If you are not familiar with $PATH, consult someone with IT knowledge to help you.

Scoring sequences on the PANTHER HMM library ● TIMING each sequence takes 20 s to score (i.e., 180 sequences per hour)

8. Run the script on UNIX command line.

```
cd pantherScore
source panther.cshrc
./pantherScore.pl -l <panther_hmm_library> -D B -V -i <fasta file> -o <output file> -n,
```

where '-l' is the path to the PANTHER HMM library downloaded above; '-D' displays type for results (Options: B, best hit: A, all hits); and '-I' is the input fasta file to score. A sample .fasta file is included in the downloaded called test.fasta; '-o' is the output file and '-n' serves to display family and subfamily names in the output file. Please note that if many sequences need to be processed, the .fasta file can be split up into smaller files so that the script can be run on multiple computers. Please also note that the output file is a tab-delimited file in the following format:

col 1 - sequence ID
col 2 - PANTHER accession (if contains :SF, is a subfamily HMM)
col 3 - PANTHER family or subfamily name
col 4 - HMM e-value score, as reported by HMMER
col 5 - HMM score, as reported by HMMER (not used by PANTHER)
col 6 - alignment range of protein for this particular HMM

9. Use this file as PANTHER Generic Mapping File for the gene list analysis tool. If the statistical enrichment test is used, please note that the numeric experimental values need to be inserted in the third column.

models (HMMs) of the protein families. As a result, if a protein is associated with a pathway component via experimental evidence, its orthologs can be inferred to the same component on the basis of the PANTHER phylogenetic tree. Second, the PANTHER pathway supports community standards, and all pathways are available in SBML (systems biology markup language[26], BioPAX (biological pathway exchange format)[27] and SBGN[25] formats.

PANTHER tools

PANTHER provides a number of useful research tools, including the PANTHER HMM scoring tool, coding single-nucleotide polymorphism (cSNP) analysis tool and gene list analysis tool. Both the PANTHER HMM scoring tool and the cSNP analysis tool have online and downloadable versions. The online versions allow only one protein sequence to be analyzed at a time, whereas the downloadable versions enable the analysis of large sequence batches. A brief description of the downloadable version of PANTHER HMM scoring tool can be found in **Box 2**. The details of how to use these tools can be found in the PANTHER help page (http://www.pantherdb.org/help/PANTHERhelp.jsp) and the user manual. In this protocol, we will focus on the gene list analysis tool. Sample gene list files can be found in **Supplementary Data 1** and **Supplementary Data 2** for testing.

The gene list analysis tool can be accessed directly from the PANTHER home page (**Fig. 4**). Several options exist for users to input a list of genes (and optionally quantitative data) for analysis.

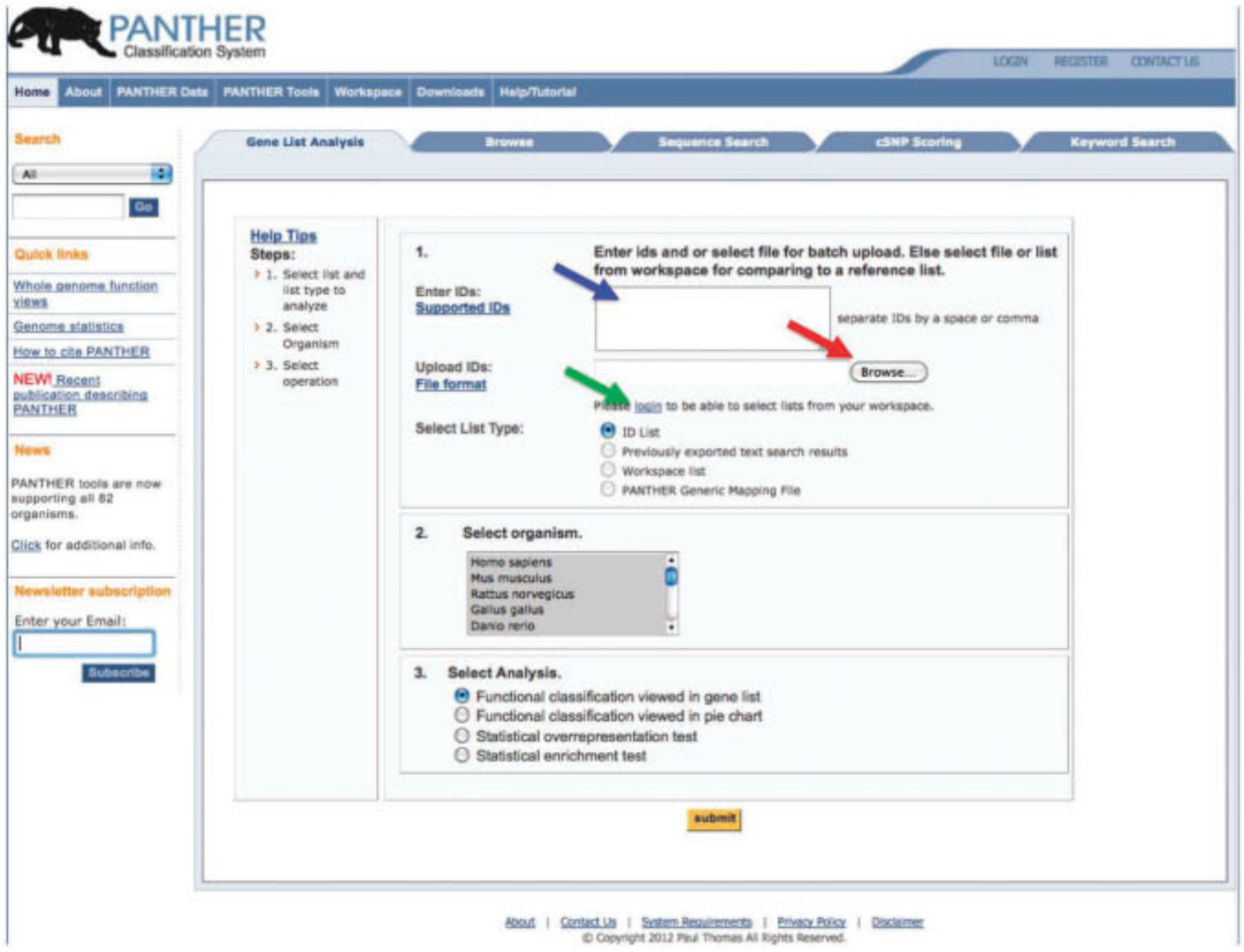

Figure 4 | The PANTHER homepage with the gene list analysis tools.

The PANTHER database uses some database identifiers as the primary IDs for each gene, and they are the most common input file options (e.g., a UniProt identifier, an Entrez Gene identifier, or even a gene symbol). The PANTHER database also maps identifiers from a number of different sources for 82 different organisms using the UniProt IDmapping mechanism, and the identifiers in the user's list are automatically mapped to the primary IDs in the PANTHER database (**Fig. 1**). A total list of the supported IDs can be found in **Box 1**. A report is generated that lists not only the mapping of each gene but also the identifiers that could not be mapped, if any. As each gene in the PANTHER database belongs to a phylogenetic tree that is annotated with GO and PANTHER ontology terms and pathways, the mapped IDs from the gene list will inherit these annotations. It needs to be pointed out that, in very rare instances, a single gene symbol or its synonym can be mapped to more than one gene from the IDmapping. We are currently working with UniProt to improve such mappings. However, this occurrence is so rare that we do not think it would significantly affect statistical results. If a gene list is not from the 82 organisms, the PANTHER tools can still be used, but the identifiers need to be mapped to PANTHER identifiers first, by using the downloadable PANTHER HMM scoring tool (**Fig. 1** and **Box 2**) and creating the PANTHER Generic Mapping File with two columns (the user's ID and the ID of the best PANTHER HMM hit; see **Box 1**). Each gene in the list will inherit the same annotations as those stored in the PANTHER database for the given HMM.

Box 3 | Statistical overrepresentation test

The input (or test) list is usually a list of genes that the researcher intends to analyze. It may be a list of genes that are upregulated in the gene expression experiment or have significant *P* values from your GWAS experiment. The list is divided into groups on the basis of GO or PANTHER classification (molecular function, biological process, cellular component, PANTHER protein class or PANTHER pathway). As many as four test lists can be uploaded for each analysis. A reference list, which usually contains all the genes and/or proteins from which the list was drawn, is divided into groups in the same way. PANTHER provides reference proteome data set as default reference lists for all 82 genomes, and thus uploading a reference list is optional. For each functional category, then (for instance, protein kinase for molecular function, cell proliferation for biological process or apoptosis signaling pathway for pathway), the binomial test is applied to determine whether there is a statistical overrepresentation or underrepresentation of the genes and/or proteins in the test list relative to the reference list.

***P* value calculation in the overrepresentation test**

The 'expected value' is the number of genes that would be expected to be present in the test list for a particular PANTHER category on the basis of the reference list. For example, out of a total of 20,000 genes in the human genome, 440 map to the GO term 'induction of apoptosis'. Therefore, 2.2% (440 divided by 20,000) of the genes in the reference list are involved in the induction of apoptosis. If a test list that contains 500 genes is uploaded to PANTHER, after analysis, 11 genes (500 multiplied by 2.2%) would be expected to be involved in induction of apoptosis.

If, for the biological process under investigation, more genes are observed in the test list than expected, there is an 'overrepresentation' (+) of genes involved in the induction of apoptosis. If fewer genes are observed than expected, there is an underrepresentation (–). A *P* value is calculated then to determine whether the over- or underrepresentation is significant or not. For example, if 21 genes involved in induction of apoptosis were observed in the test list, although this number is almost twice as big as the expected value, the *P* value for it would be large and statistically not significant (0.722). Alternatively, if 35 such genes were observed, the associated *P* value would be small and thus significant ($6.21e^{-7}$). This small *P* value indicates that the result is nonrandom and potentially interesting, and that it is worth looking at in closer detail. A *P* value cutoff of 0.05 is recommended as a starting point.

(continued)

Box 3 | (continued)

The statistical method used to conduct the overrepresentation test is the binomial test. The underlying assumption in this method is the 'null' hypothesis, that is, genes in the test list are sampled from the same general population as genes from the reference set, and thus the probability $P(C)$ of observing a gene from a particular category C in the test list is the same as in the reference list. We first estimate the probability $P(C)$ from the reference set by assuming that it is large and representative:

$$P(C) = n(C)/N$$

where $n(C)$ is the number of genes mapped to category C, and N is the total number of genes in the reference set. The above estimate is then used to calculate the P value: the probability of observing $k(C)$ genes in the uploaded list of size K. Under the null hypothesis, the number of genes mapped to C is distributed binomially with probability parameter $P(C)$, and thus the P value would be

$$P \text{ value} = \sum \binom{K}{k} P(C)^k (1 - P(C))^{K-k}$$

where the sum runs from $k(C)$ to K in the case of overrepresentation (i.e., when the number of observed genes $k(C)$ is greater than expected $P(C) \times K$ under the null hypothesis), and 0 to $k(C)$, in the case of underrepresentation (i.e., when $k(C)$ is smaller than $P(C) \times K$).

When developing this analysis tool, we also tested other statistical methods. We decided to adopt the binomial test because other methods tend to be less accurate when the population size or the expected number is small.

Once the gene list is classified with those functional terms, it can be analyzed in three different ways:

- *Functional classification.* This tool provides the functional classification results of the uploaded list and displays them in either a gene list page or pie chart.
- *Statistical overrepresentation test.* This tool is based conceptually on the simple binomial test described previously[28]. It compares a test gene list uploaded by the user to a reference gene list, and it determines whether a particular class (e.g., a GO biological process or the PANTHER pathway) of genes is overrepresented or underrepresented. A more detailed description of the tool can be found in **Box 3**.
- *Statistical enrichment test.* This tool, based on the work by the PANTHER group[29] and Lander's group[30] for two different genomic data analyses at about the same time, uses the Mann-Whitney test[31] to determine whether any ontology class or pathway has numeric values that are nonrandomly distributed with respect to the entire list of values. The numerical data can be normalized raw readouts from the microarray experiments, or, more commonly, the fold-change value for each gene in a differential expression experiment. Normalized or calculated P values from a GWAS experiment. A more detailed description can be found in **Box 4**.

It is worth pointing out that other similar tools exist to perform overrepresentation and enrichment tests, most notably GSEA[32]

Box 4 | Mann-Whitney rank-sum test (*U* test)

The statistical test is general enough to handle any numerical data, continuous or discontinuous, generated by experiments such as gene expression, proteomics or GWAS. First, a reference distribution is generated using all values from the input data. Next, the entire list is divided into groups according to GO or PANTHER classification (molecular function, biological process, cellular component, PANTHER protein class or the PANTHER pathway), and the distributions for each group are generated. The probability that the functional category distribution was drawn randomly from the reference distribution is estimated using the Mann-Whitney rank-sum test (U test)[29].

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

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

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

$$U1 = n1 \times n2 + (n1 \times (n1+1))/2 - R1$$
$$U2 = n2 \times n2 + (n1 \times (n2+1))/2 - R2$$

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

$$Z\text{-score} = (U - (n1 \times n2)/2)/\text{sqrt}(n1 \times n2 \times (n1 + n2 + 1)/12).$$

From this equation we can derive that the P value is the integral under the standard normal density.

and DAVID[33]. Although all three tools use very similar statistical algorithms in the back end, there are some differences among them. GSEA requires download and installation, and thus its target users are computer-savvy bioinformaticians. Both DAVID and PANTHER are online tools and are more appealing to bench biologists. Compared with DAVID, PANTHER has a few advantages. First, PANTHER is currently part of GO, and it integrates more updated GO curation data with the tools. In fact, DAVID downloads PANTHER data and integrates them in its analysis tools. Second, PANTHER enables users to analyze genome data from 82 organisms using the online tool, and to analyze data from any other organisms using the PANTHER scoring tool. Third, the phylogenetic trees in PANTHER protein library enable users to make more accurate ortholog prediction and thus greatly enhance the capability of the tools. One weakness of PANTHER is that it cannot analyze data against resources outside of PANTHER, such as KEGG (Kyoto Encyclopedia of Genes and Genomes) and Reactome.

Workspace

The Workspace is a unique feature in PANTHER that allows users to store the gene lists that they generate for future analysis. Although users do not have to register to use the PANTHER system, registration is required in order to use the Workspace. Registration is free. In any PANTHER gene list display page, the user can easily send the list to the Workspace.

MATERIALS

REAGENTS

- Gene list data set from genome-wide experiments. Please refer to **Box 1** for supported file formats

EQUIPMENT

- Laptop or desktop computer with internet connection. High-speed internet connection is highly recommended

EQUIPMENT SETUP

Operating system requirements Windows XP or Windows 7 for PC users and MacOS 10.5 or later for Mac users are required. A minimum of 2 GB RAM is recommended.

Browser requirements PANTHER requires Microsoft Internet Explorer 8 or later (recommended for PC users), Safari version 5 (recommended for Mac users), Firefox version 19 or Google Chrome version 26.

Java requirements The latest version of Java is required (can be downloaded from http://www.java.com/en/download/); JavaScript, Java applets and cookies must be enabled in your browser; Java applet runtime parameters should be set to -ms128m -mx512m -Xss16m.

PROCEDURE

Accessing the PANTHER website ● TIMING instantaneous

1| Access the PANTHER website by entering http://www.pantherdb.org in your web browser.

2| Prepare input file(s) according to option A, if you are working with one of the 82 genomes in the PANTHER database (see Step 5 to choose one of the 82 organisms), or option B, if you are working with an organism whose genome is not one of the 82 present in the database. Please note that sample gene list files can be found in **Supplementary Data 1** and **2**. Please do consider trying this protocol on those data sets to familiarize yourself with the tools included in the PANTHER system.
▲ **CRITICAL STEP** The input file must be in simple text file format (.txt or .tab). It also must use IDs supported in the PANTHER system and the correct tab-delimited format as described in **Box 1**.

(A) Input file preparation for data from organisms included in the PANTHER genomic database ● TIMING 15–30 min

(i) Prepare input file(s) in simple text file (.txt or .tab) with gene or protein identifiers as the first column and numeric values as the second column if you want to use the statistical enrichment test. Detailed instructions on file format and supported IDs can be found in **Box 1**.

(B) Input file preparation for data from organisms not included in the PANTHER genomic database ● TIMING variable

(i) Prepare input file(s) by mapping your sequence identifiers to the PANTHER HMM IDs (PANTHER Generic Mapping File) by using the procedure described in **Box 2**.

3| Upload the gene list to the PANTHER tool system using option A, if you have a small list for functional classification tools. Use option B if you have a large list and want to analyze it by using the entire gene list analysis tool, or use option C if you previously saved the list in the Workspace.

? TROUBLESHOOTING

(A) Uploading an ID list by entering the IDs to the Enter ID box ● TIMING 1 min

(i) Paste the ID list prepared in Step 2 into the Enter ID box. Alternatively, you can also type IDs, one per line, into the box (blue arrow in **Fig. 4**). Please note that the ID upload to the Enter ID box only supports 'functional classification tools'.

Figure 5 | Results of functional classification displayed as a gene list page. The results follow the upload of **Supplementary Data 1** as the gene list file.

(B) Uploading an ID list from a list file on your computer ● TIMING 1 min

(i) Upload the list file prepared in Step 2 by clicking the 'Browse' button (red arrow in **Fig. 4**), and then follow the online instruction to locate the file.

(C) Uploading an ID list from the Workspace ● TIMING 3 min

(i) If you have previously saved your list into the Workspace, you can use it by clicking the 'login' link (green arrow in **Fig. 4**), and follow the online instruction to locate the file in the Workspace. Please note that numeric values cannot be saved in the Workspace, and therefore this type of upload does not support the statistical enrichment test.

Selecting a list type ● TIMING instantaneous

4| Select a corresponding list type in order for the tool to work properly. Three list types are supported by the tools: ID List, Previously exported text search results and PANTHER Generic Mapping File. See **Box 1** for details of the list types.

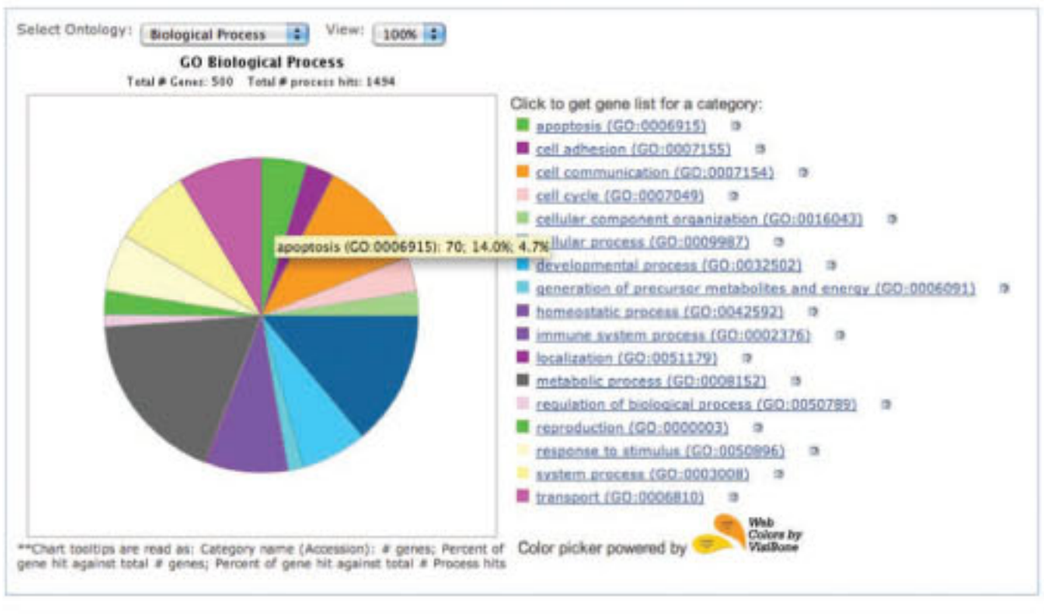

Figure 6 | PANTHER pie chart results using **Supplementary Data 1** as the input gene list file. You can use the 'Select Ontology' drop-down menu to switch to the pie chart of different ontologies. Click on the pie chart section to display the child categories. Click on the legends on the right side to retrieve the list of genes for that category.

Selecting an organism ● TIMING instantaneous

5| (Optional) If data are not from the 82 organisms and the researcher is uploading a PANTHER Generic Mapping File prepared from Step 2B, this step can be skipped. If your data are from one of those 82 organisms, select the relevant organism from the drop-down menu, which lists the

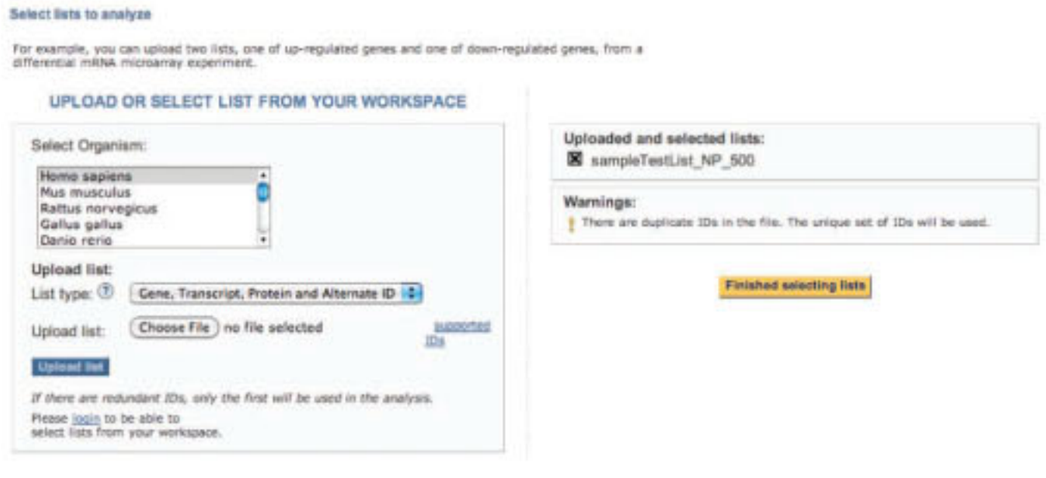

Figure 7 | User interface of the statistical overrepresentation test, which enables users to select additional test gene lists.

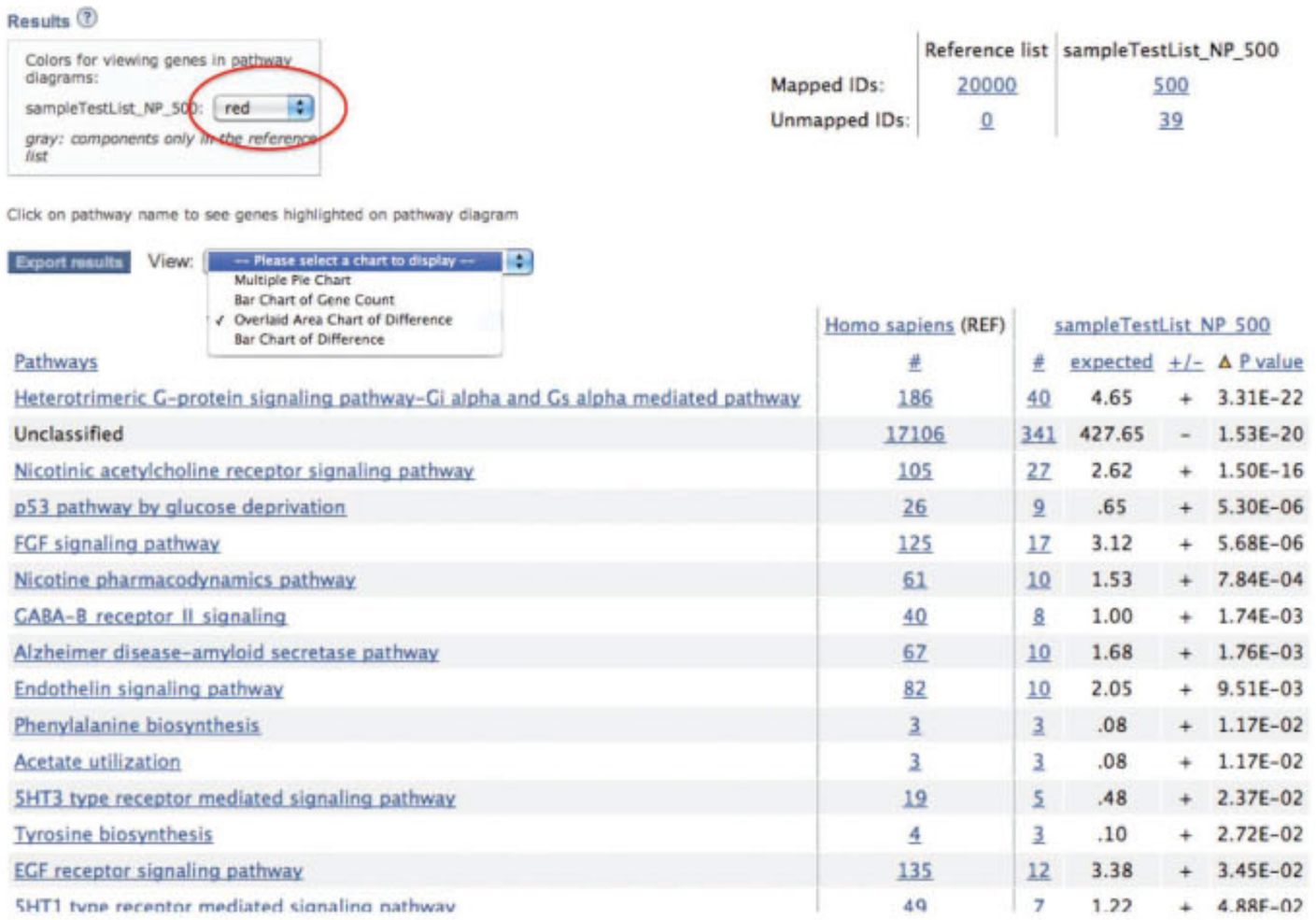

Figure 8 | Summary of the results from the statistical overrepresentation test displayed in a table. The results are based on using **Supplementary Data 1** as the input gene list file. The table can be exported as a tab-delimited file by clicking the 'Export results' button. Other views of the results are available by using the 'View' drop-down menu. If the analysis is done in the pathway as shown here, the pathway name can be clicked and the pathway diagram will be displayed. The pathway components that have genes in your test list will be highlighted. The color of the highlighted component can be defined at the top of the page (red circle). A total of four test lists can be analyzed and viewed at the same time.

12 model organisms first, followed by the remaining 70 organisms ordered alphabetically. Please note that there are two purposes for selecting an organism at this point. First, some identifiers, such as gene symbols, are not organism specific. By selecting an organism here, the IDs are mapped to those in the organism you are interested in. Second, if the statistical overrepresentation test is selected, the default reference gene list is based on the selected organism**.**

6| To find out the functional classification of the genes in your list (**Fig. 5**), analyze the data uploaded according to option A. Use option B to obtain the functional classification of the genes in your list displayed as a pie chart (**Fig. 6**). Use option C to find out which functional classes are over- or underrepresented in the list. Finally, use option D to perform a gene set enrichment test. See ANTICIPATED RESULTS for details on how to interpret the results of the present operation.

(A) Classifying the uploaded list and viewing the results in a gene list page ● TIMING instantaneous

(i) Select 'Functional classification viewed in gene list' by clicking the radial button in the Select Analysis box, and then click the 'submit' button.

? TROUBLESHOOTING

(B) Classifying the uploaded list and viewing the results in a pie chart ● TIMING instantaneous

(i) Select 'Functional classification viewed in pie chart' by clicking the radial button in the Select Analysis box, and then click the 'submit' button.

(C) Analyzing the uploaded list with the overrepresentation test ● TIMING 5–10 min

▲ CRITICAL STEP Please note that Step 6C(ii–viii) is optional and should be implemented if more than one list is going to be analyzed. Please note that a total of four gene lists can be uploaded and analyzed.

(i) Select 'Statistical over-representation test' by clicking the radial button in the Select Analysis box, and then click the 'submit' button.

Figure 9 | Results from the statistical enrichment test. The results are derived from the use of **Supplementary Data 2** as the input file. The output of the tool provides a list of *P* values for each comparison between a functional category distribution and the reference distribution.

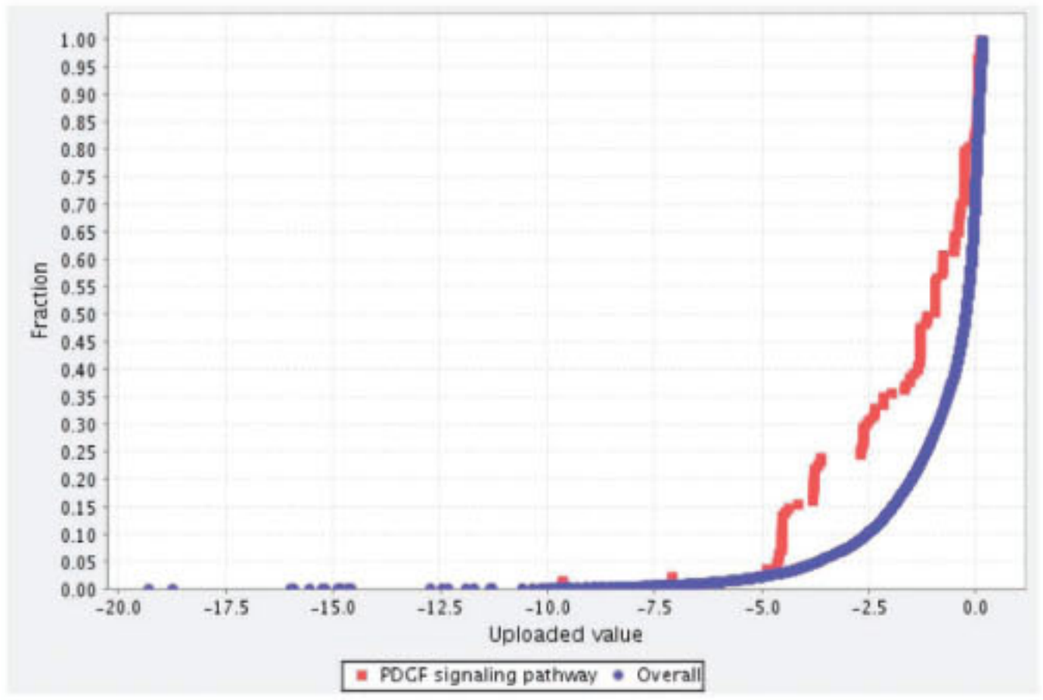

Figure 10 | Graph view of results from the statistical enrichment test. Comparison of the distributions from the PDGF signaling pathway (red) and the reference (blue) in graph view.

(ii) In the subsequent page (**Fig. 7**), select additional gene lists for the analysis. Note that a total of four test gene lists can be uploaded for this tool.
(iii) Click the 'Browse' button.
(iv) Select the gene list from your computer.
(v) Select the organism. The default organism is the one selected when the first gene list is uploaded.
(vi) Click 'Upload list'.
(vii) Select the 'List type'. The default is the one selected when the first gene list is uploaded.
(viii) After all lists are uploaded, click the 'Finish selecting lists' button. The tool will take you to the next page, which allows you to make the selections described in the following steps.

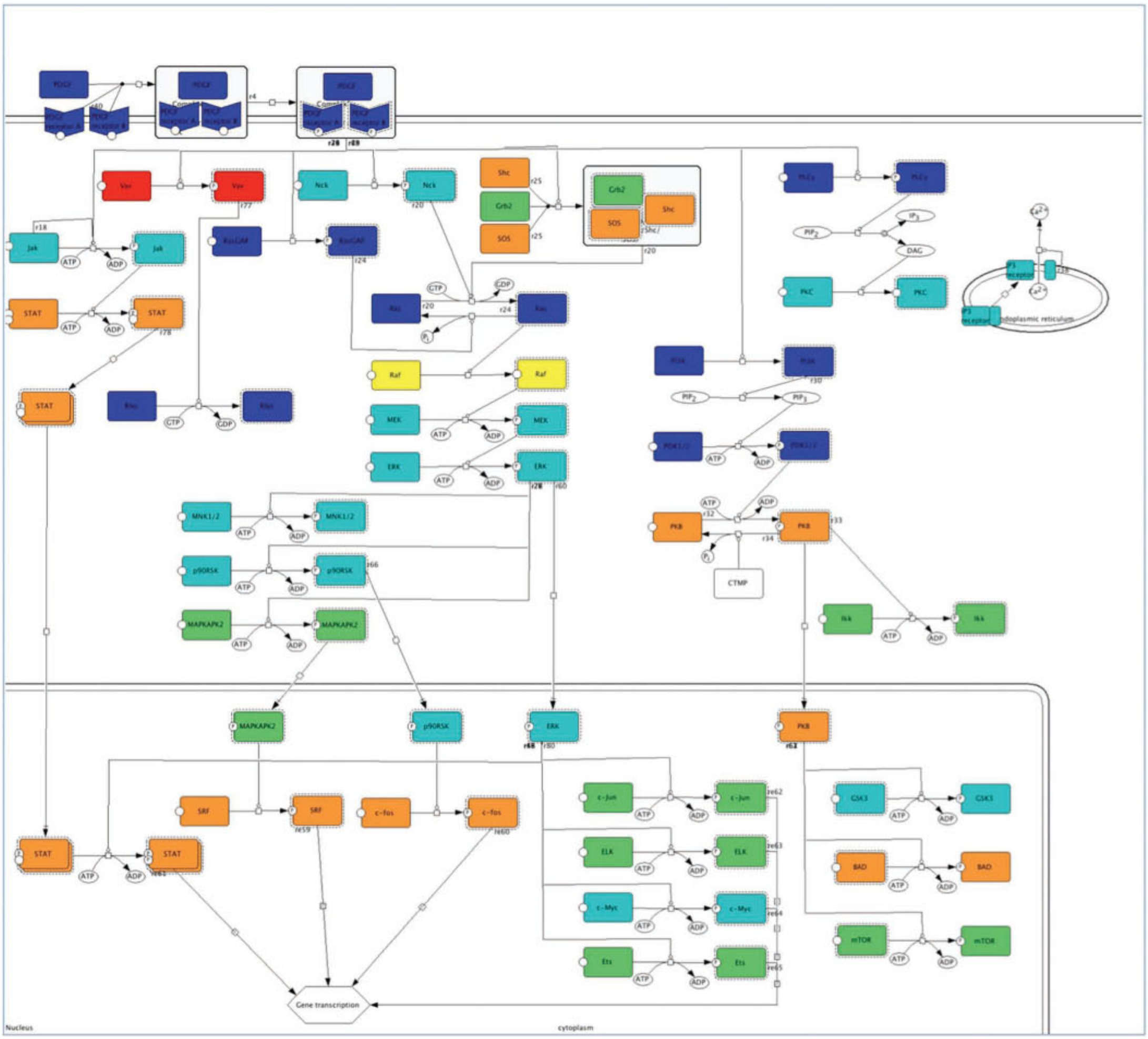

Figure 11 | Results from the statistical enrichment test as visualized in the pathway diagram. A pathway diagram of the PDGF signaling pathway is visualized with an interactive pathway Java applet that colors the pathway using a heat map derived from the input values.

(ix) Select a reference list. Please note that the default list is always the entire proteome of the organism selected above.
(x) (Optional) If you choose to change the reference list, click the 'Select reference list' button and then select a reference list from another organism or upload your own using an interface similar to the one used to upload the test gene list.
(xi) On the basis of the type of function you want to analyze, select one of the following five ontologies: PANTHER pathway, GO molecular function, GO biological process, GO cellular location or PANTHER protein class. Please note that multiple test correction (Bonferroni correction) is selected by default at this stage.
(xii) Click the 'Launch analysis' button.
? TROUBLESHOOTING
(xiii) On the results page (**Fig. 8**), visualize the results by implementing one of the following three options: export the result table in a tab-delimited file by clicking the 'Export results' button; view the results in graphs by using the 'View' drop-down menu; or, if your analysis is done in the pathway as shown here, click the pathway name and display the pathway diagram. The pathway components that have genes in your test list will be highlighted. The color of the highlighted component can be defined at the top of the page (**Fig. 8**, red circle). Please note that a total of four test lists can be analyzed and viewed at the same time.
? TROUBLESHOOTING

(D) Analyzing the uploaded list with the enrichment test ● TIMING 2–5 min

(i) Select 'Statistical enrichment test' by clicking the radial button in the Select Analysis box.
(ii) After clicking the 'submit' button, on the next page, select an ontology.
(iii) Click the 'Launch analysis' button.
? TROUBLESHOOTING
(iv) On the results page (**Fig. 9**), visualize the results by implementing one of the following three options: Export the result table in a tab-delimited file by clicking the 'Export results' button; compare the distribution curve in graph view by checking the box in front of the category or pathway of your interest, and by clicking the 'Graph selected categories' button (**Fig. 10**); or, if your analysis is done in the pathway as shown here, click the pathway name and display the pathway diagram. The pathway components are colored in a heat map on the basis of the input numeric values (**Fig. 11**). Click the 'Specify color ranges of pathway diagrams' button on the result page to view or specify the color ranges. Please note that, in order to use this last tool, you need to ensure that the uploaded gene list contains a second column with numerical values.
? TROUBLESHOOTING

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
3	Failure to upload the file	The input file is probably in the wrong file format	Make sure that your file is in simple text format (.txt or .tab) If you are uploading a file with numeric values for the enrichment test, make sure that the second column contains only numeric numbers. Any rows with no values should be removed instead of leaving them blank or marking them as 'n/a' Make sure that the first column is not blank
6A(i), 6C(xii), 6D(iii)	IDs in the uploaded file do not have a mapped ID in PANTHER	The current PANTHER data is based on the April 2012 release of Reference Proteome Project and its ID mapping. It is possible that a small fraction of the IDs may not map because of outdated data either in the PANTHER database or your uploaded file	There is no solution to this problem. If you believe that you are using the current IDs, please do the following: Make sure that the IDs in the uploaded file are supported by PANTHER. Refer to **Box 1** for 'Supported IDs' IDs from certain databases may contain a version number at the end of them, e.g., NP_000242.1. Do not include the version number ".1" in the ID, and just use NP_000242 Send feedback to feedback@pantherdb.org
6C(xiii), 6D(iv)	Pathway applet cannot be launched to view the pathway diagrams	Your Java plug-in may be outdated	Read the system requirements and make sure that your computer has the most updated Java version

● TIMING

Step 1, accessing the PANTHER website: instantaneous
Step 2A, input file preparation for data from organisms included in the PANTHER genomic database: 15–30 min
Step 2B, input file preparation for data from organisms not included in the PANTHER genomic database: variable; on average, 180 sequences per hour, plus 1.5–3.5 h for tool installation
Step 3A, uploading an ID list by entering the IDs to the Enter ID box: 1 min
Step 3B, uploading an ID list from a list file on your computer: 1 min
Step 3C, uploading an ID list from workspace: 3 min
Step 4, selecting a list type: instantaneous
Step 5, selecting an organism: instantaneous
Step 6A, classifying the uploaded list and viewing the results in a gene list page: instantaneous
Step 6B, classifying the uploaded list and viewing the results in a pie chart: instantaneous
Step 6C, analyzing the uploaded list with the overrepresentation test: 5–10 min
Step 6D, analyzing the uploaded list with the enrichment test: 2–5 min
Box 2, software downloads: ~30 min
Box 2, software decompression: 30 min
Box 2, installation of HMMER2 and BLAST: ~30 min for experts, ~2 h for everyone else
Box 2, scoring sequences on the PANTHER HMM library: each sequence takes 20 s to score (i.e., 180 sequences per hour)

ANTICIPATED RESULTS

Functional classification tool viewed in gene list page

This tool returns the results as a gene list webpage (**Fig. 5**). The page displays all the IDs from the uploaded gene list and their mapped PANTHER sequence IDs, as well as ontology and pathway terms. The page contains the following information (**Fig. 5**).

- *Gene ID*. This is the identifier for genes in the PANTHER library. The format is as follows: organism|gene database source = gene id|protein database source = protein id. For example, HUMAN|ENSEMBL = ENSG00000111262|UniProtKB = Q09470 is a human sequence; the gene sequence is from ENSEMBL with id ENSG00000111262 and the protein sequence is from UniProt with id Q09470.
- *Mapped IDs*. IDs from the uploaded gene list that are mapped to the gene ids in the first column.
- *Gene Name/Gene Symbol*. The Entrez gene definition and gene symbol.
- *PANTHER Family/Subfamily*. The name and identifier of the PANTHER family or subfamily where the gene in the first column is in.
- *GO Molecular Function, Biological Process and Cellular Component*. These are GO terms from PANTHER GO slim (GO slim is a subset of GO terms that gives a broad overview of the GO ontology) describing the function of the gene product.
- *PANTHER Protein Class*. This is a PANTHER index term describing protein classes.
- *Pathway*. Pathway and pathway components that are linked to the PANTHER subfamily in column 4. The subfamily is linked to the pathway component when at least one of its member genes is associated with the component directly by manual curation.
- *Species*. The organism of the gene in column 1.

Once you reach the gene list page, you can make the following changes to view the results.

- *Sort the list*. The list can always be sorted by clicking on any of the underlined column names. A yellow triangle appears in front of the column name that you choose to sort. The orientation of the triangle indicates whether the sort is ascending or descending.
- *Customize columns*. You can click on the 'x' button next to the column names to collapse the column.
- *Converting a list to another list type*. Select the genes you want to convert by clicking the checkboxes. The default is for all genes in the list.
- *Saving the list*. Select the genes you want to save by clicking the checkboxes. The default is for all genes in the list. You can select one of the following from the pull-down menu as the destination. (i) Workspace; you need to register to save data to the workspace. The registration is free. When you make this selection, a pop-up window will prompt you to name the list and add any comments. The name and comments can then be edited from the Workspace page. Once the gene list is saved in your workspace, it can be returned to at any time. Only the IDs are stored, and they are mapped to the internal PANTHER gene ids, and thus whenever you access a list again all information will have been updated and current. (ii) Exporting a list to a file; by making this selection, you will export the list as a tab-delimited file. In this manner, you can import the file to Excel or postprocess it as you wish. (iii) View the list as text on the website.
- *Pie chart view*. Use the pie chart view by clicking the colorful pie chart icon (see Glossary for the meaning of the abbreviations). See the next section for details about how to interpret the pie chart.

Functional classification tool viewed in pie chart

This tool returns the results in a pie chart, which displays an overview of all ontology terms at the first (or most general) level within the same ontology (**Fig. 6**). When a slice of the pie chart, which represents an ontology term, is clicked, a new pie chart will appear that contains its child ontology terms. As a gene can be classified according to more than one term, the pie chart is calculated according to the number of 'hits' to the terms over the total number of 'class hits.' A class hit indicates independent ontology terms. For example, if a gene is classified according to 2 ontology terms that are not parent or child to each other, this classification counts as 2 class hits.

When you place the computer mouse over a slice, the category name and a series of counts are displayed. In our example in **Figure 6**, the name is a GO term 'apoptosis' (GO:0006915) followed by

(i) the number of genes (70) from the uploaded list that are classified to the term apoptosis;
(ii) the percentage (14%) of genes classified to apoptosis (70) over the total number of genes (500);
(iii) the percentage (4.7%) of genes classified to this apoptosis (70) over the total number of class hits (1494).

Statistical overrepresentation test

The results of the implementation of this analysis tool are displayed in a table (**Fig. 8**). If a test gene list is uploaded, the table contains six columns of data:

In the first column the name of the PANTHER classification category is reported. If you are implementing this analysis in pathways, the corresponding pathway diagram can be viewed by clicking on the pathway name.

In the second column, the number of genes in the reference list that map to the specific PANTHER classification category is reported.

In the third column, the number of genes in the list uploaded that map to the PANTHER classification category is reported.

In the fourth column, the number of genes to be expected in your list for the PANTHER category, on the basis of the reference list (**Box 3**), is reported.

In the fifth column, a plus or a minus sign appears. '+' indicates overrepresentation of this category in the test list: more genes than expected were observed, based on the reference list; in other words, for this category, the number of genes in your list is higher than expected. Conversely, '−' indicates underrepresentation, that is, when fewer genes than expected are present in the list.

In the sixth column, the *P* value is reported, as determined by the binomial statistic (**Box 3**). This parameter represents the probability that the number of genes observed in this category occurred by chance, as determined by your reference list. A small *P* value indicates that the observed number is significant and potentially interesting. We recommend a cutoff value of 0.05 for *P* as a starting point.

If more than one test list is uploaded, columns 3–6 are repeated for each list.

From this result page, various statistics can be exported by using the drop-down menu next to the 'Export results' button. The list of genes or proteins in any functional group can be viewed by clicking on the listed counts. When pathways are chosen as the functional categories, clicking on the pathway name brings up pathway diagrams colored according to preferences specified by the user (**Fig. 12**). The resulting pathway diagram can be exported as an image file (.png) by choosing *File* → *Export image* from the applet menu.

Statistical enrichment test

The results are displayed in a table with four essential columns of data (**Fig. 9**):

In the first column, the name of the PANTHER classification category is reported. If this analysis is being performed in terms of pathways, the pathway name can be clicked to view the pathway diagram. The genes in the pathway diagram are colored according to the numeric value provided in the uploaded gene list file, and the rules for this can be specified by clicking on the 'Specify color ranges' button.

In the second column, the number of genes that map to this particular PANTHER classification category is reported.

In the third column, a '+' or '−' appears. A plus sign indicates that, for this category, the distribution of values for your uploaded list is shifted toward greater values than the overall distribution of all genes that were uploaded. A negative sign indicates that the uploaded list is shifted toward smaller values than the overall list.

In the fourth column, the *P* value, as calculated from the Mann-Whitney *U* test (Wilcoxon rank-sum test) (**Box 4**), is reported. A large *P* value indicates that the genes for this category have a distribution that is similar to randomly choosing genes from the overall distribution. In other words, the values of the uploaded genes for this category have a distribution similar to the overall list of values that were input. A small, significant *P* value indicates that the distribution for this category is nonrandom and different from the overall distribution. We recommend a cutoff value of 0.05 for *P* as a starting point.

For a visual representation of these distributions, select the checkboxes of the categories of interest, and then click on the 'Graph selected categories' button. The graph will be displayed in a new window (**Fig. 10**). The *x* axis is your uploaded value.

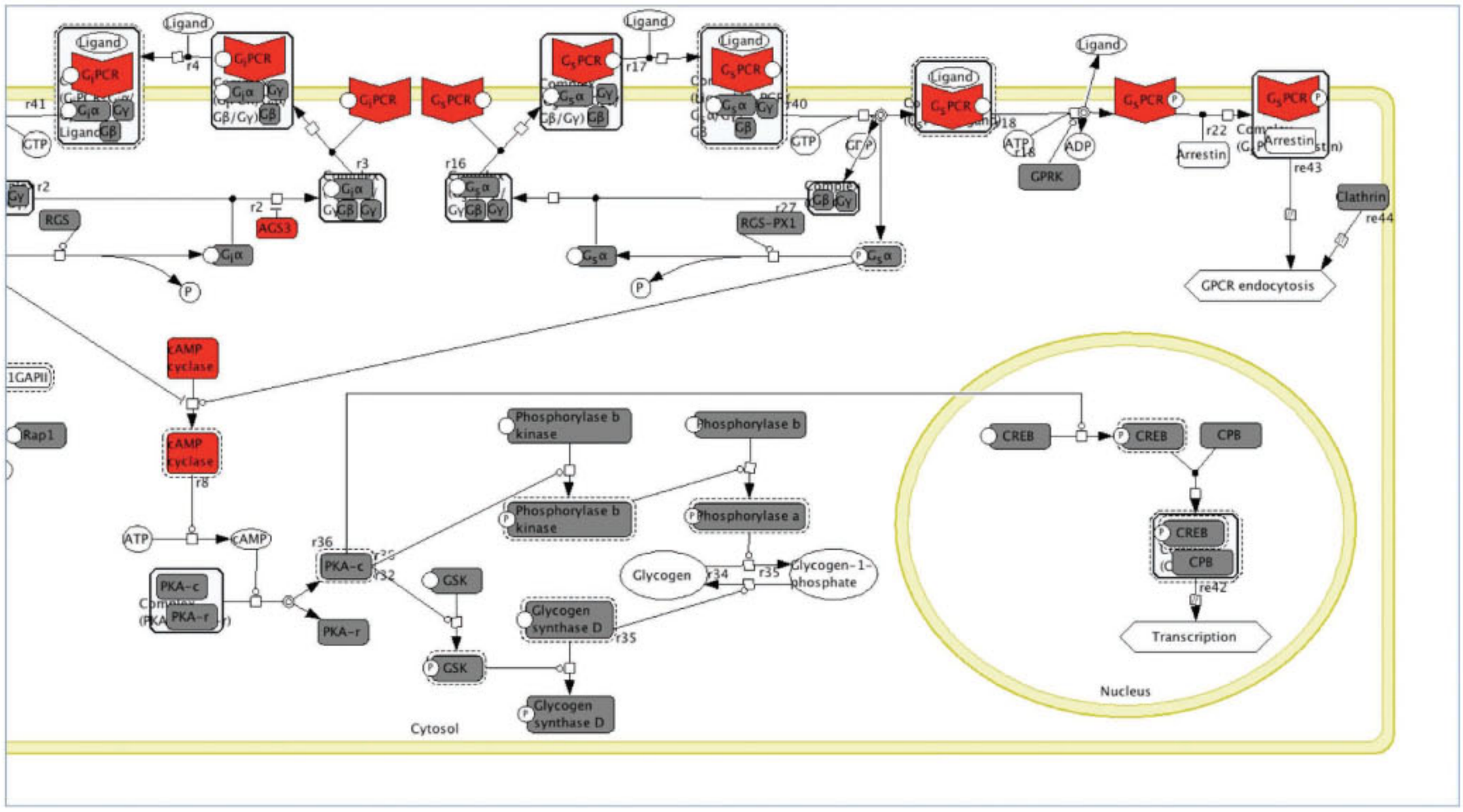

Figure 12 | Results of the statistical overrepresentation test viewed in the PANTHER pathway, 'Heterotrimeric G protein signaling pathway – G_i-α and G_s-α–mediated pathway (PANTHER accession code P00026)'. The components that contain the genes in the test gene list are reported in red.

The *y* axis is the cumulative fraction. The blue curve is the overall distribution for all genes. The red curve is the selected functional category. In this case, it is the PDGF signaling pathway. For the data point $x = -2.5$, $y = 0.3$ for the red curve and $y = 0.1$ for the blue curve. This means that 30% of the uploaded genes have a value of –0.25 or smaller, but only 10% of the overall genes have a value of –0.25 or smaller. In other words, this set of values shows that the distribution of the category tends to be smaller than the overall distribution. We find that this information is critical to interpret any deviation between the functional category distribution and the overall distribution.

Clicking on the listed counts enables the user to view the genes or proteins in each category from the output page. In addition, in the case of pathways, clicking on the name of the pathway brings up an interactive Java applet that colors the pathway using a heat map derived from the input values (**Fig. 11**). The resulting pathway diagram can be exported as an image file (.png) by choosing *File* → *Export image* from the applet menu.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS We thank the Reference Proteome team, especially C. McAnulla and M. Martin, for their support in providing up-to-date Reference Proteome data set, and we thank Y. Matsuoka and K. Manami from the Systems Biology Institute Japan for their support on CellDesigner and pathway file update. This work is supported by the US National Institutes of Health (NIH)/National Institute of General Medical Sciences (NIGMS) grant no. GM081084 to P.D.T. Funding for open access was provided by the University of Southern California.

AUTHOR CONTRIBUTIONS A.M. developed the software code for the website. J.T.C. maintained the database and web servers. H.M. generated the content of the system and supervised the project. P.D.T. provided the funding and supervised the project. H.M. wrote the manuscript with contributions from all the authors.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

1. Mi, H., Muruganujan, A. & Thomas, P.D. 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 (2013).
2. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Thomas, P.D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).
4. Thomas, P.D. *et al.* Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.* **34**, W645–W650 (2006).
5. Mi, H. *et al.* Assessment of genome-wide protein function classification for *Drosophila melanogaster*. *Genome Res.* **13**, 2118–2128 (2003).
6. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33**, D284–D288 (2005).
7. Mi, H. & Thomas, P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.* **563**, 123–140 (2009).
8. Funahashi, A. *et al.* CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc. IEEE* **96**, 1254–1265 (2008).
9. van Baarsen, L.G.M. *et al.* Gene expression profiling in autoantibody-positive patients with arthralgia predicts development of arthritis. *Arthritis Rheum.* **62**, 694–704 (2010).

10. Verma, G., Bhatia, H. & Datta, M. Gene expression profiling and pathway analysis identify the integrin signaling pathway to be altered by IL-1β in human pancreatic cancer cells: role of JNK. *Cancer Lett.* **320**, 86–95 (2012).
11. Boyer, A.P., Collier, T.S., Vidavsky, I. & Bose, R. Quantitative proteomics with siRNA screening identifies novel mechanisms of trastuzumab resistance in HER2-amplified breast cancers. *Mol. Cell Proteomics* **12**, 180–193 (2013).
12. Stützer, I. *et al.* Systematic proteomic analysis identifies β-site amyloid precursor protein cleaving enzyme 2 and 1 (BACE2 and BACE1) substrates in pancreatic beta cells. *J. Biol. Chem.* **288**, 10536–10547 (2013).
13. Shi, Y. *et al.* Genome-wide association study identified eight new risk loci for polycystic ovary syndrome. *Nat. Genet.* **44**, 1020–1025 (2012).
14. den Hoed, M. *et al.* Identification of heart rate–associated loci and their effects on cardiac conduction and rhythm disorders. *Nat. Genet.* **45**, 621–631 (2013).
15. Feng, J. *et al.* Dnmt1 and Dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons. *Nat. Neurosci.* **13**, 423–430 (2010).
16. Hek, K. *et al.* A genome-wide association study of depressive symptoms. *Biol. Psychiatr.* **73**, 667–678 (2013).
17. Neely, G.G. *et al.* A global *in vivo Drosophila* RNAi screen identifies *NOT3* as a conserved regulator of heart function. *Cell* **141**, 142–153 (2010).
18. McDowall, J. & Hunter, S. InterPro protein classification. *Methods Mol. Biol.* **694**, 37–47 (2011).
19. Gene Ontology Consortium. The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.* **40**, D559–D564 (2012).
20. Mi, H. *et al.* PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* **38**, D204–D210 (2010).
21. Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput. Biol.* **5**, e1000431 (2009).
22. Gaudet, P., Livstone, M.S., Lewis, S.E. & Thomas, P.D. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinform.* **12**, 449–462 (2011).
23. Cerami, E.G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2010).
24. Thomas, P.D. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics* **11**, 312 (2010).
25. Le Novere, N. BioModels Database—a database of annotated published models http://www.ebi.ac.uk/biomodels-main/static-pages.do?page=home (2011).
26. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
27. Demir, E. *et al.* The BioPAX community standard for pathway data sharing. *Nat. Biotech.* **28**, 935–942 (2010).
28. Cho, R.J. & Campbell, M.J. Transcription, genomes, function. *Trends Genet.* **16**, 409–415 (2000).
29. Clark, A.G. *et al.* Inferring non-neutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
30. Mootha, V.K. *et al.* Identification of a gene causing human cytochrome *c* oxidase deficiency by integrative genomics. *Proc. Natl. Acad. Sci. USA* **100**, 605–610 (2003).
31. Mann, H.B. & Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* **18**, 50–60 (1947).
32. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
33. Sherman, B. *et al.* DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics* **8**, 426 (2007).