

Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0)

Huaiyu Mi^{1*}, Anushya Muruganujan¹, Xiaosong Huang^{1,2}, Dustin Ebert¹, Caitlin Mills¹, Xinyu Guo¹ and Paul D. Thomas^{1*}

The PANTHER classification system (<http://www.pantherdb.org>) is a comprehensive system that combines genomes, gene function classifications, pathways and statistical analysis tools to enable biologists to analyze large-scale genome-wide experimental data. The current system (PANTHER v.14.0) covers 131 complete genomes organized into gene families and subfamilies; evolutionary relationships between genes are represented in phylogenetic trees, multiple sequence alignments and statistical models (hidden Markov models (HMMs)). The families and subfamilies are annotated with Gene Ontology (GO) terms, and sequences are assigned to PANTHER pathways. A suite of tools has been built to allow users to browse and query gene functions and analyze large-scale experimental data with a number of statistical tests. PANTHER is widely used by bench scientists, bioinformaticians, computer scientists and systems biologists. Since the protocol for using this tool (v.8.0) was originally published in 2013, there have been substantial improvements and updates in the areas of data quality, data coverage, statistical algorithms and user experience. This Protocol Update provides detailed instructions on how to analyze genome-wide experimental data in the PANTHER classification system.

This protocol is an update to *Nat. Protoc.* 8, 1551–1566 (2013): <https://doi.org/10.1038/nprot.2013.092>

Introduction

The PANTHER classification system is designed to be a comprehensive platform for the analysis of gene function on a genome-wide scale¹. Although its initial aim was to classify gene and protein functions^{2,3}, it has evolved through the years to also serve as an online resource for experimental data analysis^{4–6}. 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. Nearly 1,600 unique IP addresses access the PANTHER website with over 30,000 page views daily (please note that one IP address can have multiple users, so the actual number of users might be much larger). According to Google Scholar, over 11,000 publications have cited our work (5,000 of them since 2016). We believe that there are two main reasons that PANTHER is able to attract more users and may have an advantage compared to other tools in the field: first, as a GO consortium member, PANTHER is integrated into the GO curation process, especially the phylogenetic annotation effort⁷, and provides more up-to-date annotation data (updated monthly). Second, PANTHER provides support to more genomes than other tools do (~1,000) in collaboration with the Reference Proteome project⁸.

PANTHER was initially released publicly in 2003, and quickly became a popular online resource for genome-wide analysis of gene functions on various experimental data, including gene expression, proteomics and genome sequencing data. Our original protocol was published in this journal in 2013 with detailed background information⁶. Since then, a number of system-wide updates and improvements have been made in order to meet the needs of the user community (Fig. 1). The improvements have been made in four main areas, as described below.

¹Division of Bioinformatics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.

²Present address: School of Life Sciences, Guangzhou University, Guangzhou Higher Education Mega Center, Guangzhou, China.

*e-mail: huaiyumi@usc.edu; pdthomas@med.usc.edu

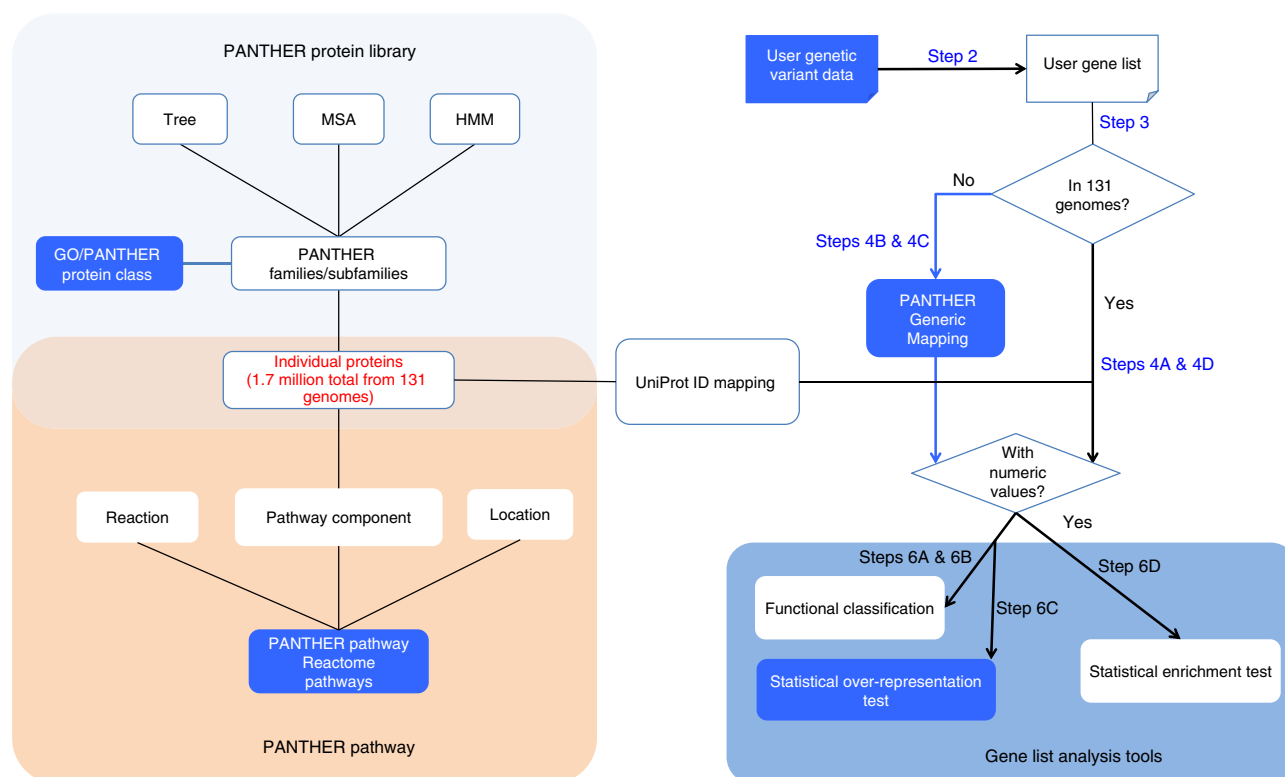


Fig. 1 | Overview of PANTHER infrastructure and recent improvements. PANTHER consists of three modules. The core module is the PANTHER protein library (left, light blue background) that contains a collection of PANTHER families and subfamilies, each of which is represented by a phylogenetic tree, a multiple sequence alignment (MSA) and an HMM. The second module is the pathway module that contains expert-curated pathways from both PANTHER and Reactome (orange background). The pathway components are associated with protein sequences that are also used to build the protein library (red text); in this way, 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 background). Major updates and improvements have been made to the components highlighted in royal blue. Blue arrows surrounding the PANTHER Generic Mapping file indicate a new workflow available for users, which dramatically expands the number of organisms that can be analyzed with PANTHER (Box 4). There are three types of analysis that can be performed: functional classification, statistical over-representation test and statistical enrichment test. Numeric values must be provided for the statistical enrichment test. The corresponding procedure steps are labeled next to the arrows.

Improved annotation data quality and coverage

All the analyses in PANTHER rely on the accuracy of the annotation datasets. During the past 5 years, improvements have focused on expanding the data coverage and accuracy in the following areas:

- 1 The PANTHER GO slim annotation datasets that were reported in the previous protocol now use the data from the GO phylogenetic annotation effort⁷. These are manually curated annotations to the ancestral nodes on PANTHER family trees based on the experimental annotations on their leaf descendants (extant genes). The annotation can be extrapolated to other leaf sequences (that have not been tested experimentally) under the annotated ancestral node.
- 2 The complete GO annotation datasets⁹ have been incorporated into the PANTHER tools for analysis. They include both experimental and electronic annotations. The data are updated monthly.
- 3 To expand the coverage of pathway data, we have added Reactome pathway¹⁰ datasets to the system. Our plan is to add more data from the Pathway Commons project¹¹ in the near future, including pathway databases (e.g., KEGG¹², HumanCyc¹³ and WikiPathway¹⁴) and protein–protein interaction databases (e.g., IntAct database¹⁵ and BioGRID¹⁶).

A detailed description of each of the above datasets available in PANTHER analysis tools can be found in Box 1.

Analysis of genotype data

To respond to the increasing requests from PANTHER users who apply the tool to analyze genetic variation data, including single-nucleotide polymorphism (SNP) data, from experiments such as

Box 1 | Annotation datasets

There are nine annotation datasets, in four general data types, available in PANTHER for users to choose in their analysis. Below is a brief description of each of them.

PANTHER GO-Slim

GO annotations from the phylogenetic curation effort are captured in 3,039 GO Slim terms. The annotation datasets use the data from the GO phylogenetic annotation effort⁷. These are manually curated annotations to the ancestral nodes on PANTHER family trees based on the experimental annotations on their leaf descendants (extant genes). The annotation can be extrapolated to other leaf sequences (that have not been tested experimentally) under the annotated ancestral node. There are three annotation datasets corresponding to three GO aspects in this data type:

- Biological process (default)
- Molecular function
- Cellular component

Complete GO annotation datasets

These datasets include complete GO annotations, both manually curated and electronic. Electronic annotations are generated by computer algorithm on the basis of sequence similarity; they are usually not reviewed by curators, and thus are less reliable. There are three datasets corresponding to three GO aspects in this data type:

- GO molecular function complete
- GO biological process complete
- GO cellular component complete

Pathways

The current PANTHER database includes two pathway datasets from the following two resources:

- PANTHER Pathway—PANTHER Pathway consists of over 177, primarily signaling, pathways, each with PANTHER subfamilies and protein sequences mapped to individual pathway components¹⁹.
- Reactome Pathways—Reactome is a freely available, open-source relational database of signaling and metabolic molecules and their relations organized into biological pathways and processes. The core unit of the Reactome data model is the reaction. Entities (nucleic acids, proteins, complexes, vaccines, anticancer therapeutics and small molecules) that participate in reactions form a network of biological interactions and are grouped into pathways¹⁰.

PANTHER Protein Class

The PANTHER Protein Class ontology was adapted from the PANTHER/X molecular function ontology³ and includes commonly used classes of protein families, many of which are not covered by GO molecular function. There is one corresponding dataset in this data type.

genome-wide association studies (GWASs) and genome sequencing, we have added a new feature to support genetic variant data in variant call format (VCF), which is a text file format for storing sequence variation data. Its specification can be found on GitHub at <https://samtools.github.io/hts-specs/VCFv4.3.pdf>. A sample VCF file can be found in Supplementary Data 1 for test purposes. Currently, only the over-representation test supports the analysis using VCF file format, and the human reference genome release GRCh38/hg38 is supported. Each variant is mapped to a gene if it is within the gene region or the flanking region specified by the user. To avoid artifacts, multiple variants in the same gene are counted only once, and a given variant can be associated with only a single gene. The statistical analysis is performed on the converted gene list as described in Box 2. Users are able to upload both a test list (or lists) and a reference list.

Improved statistical tests

To meet the current standard in the field, we have made the following improvements:

- 1 Fisher's exact test, with the Benjamini–Hochberg false discovery rate (FDR) correction¹⁷ for multiple testing, has been added as the default algorithm for the over-representation test.
- 2 FDR correction has been added to the binomial distribution test. The option to use the original settings (binomial distribution test with Bonferroni correction) is still available in the configuration panel (Step 6C(ii–iv)).
- 3 FDR correction has been added as one of the multiple testing correction methods to the enrichment test.

More information about these tests can be found in Box 2.

Box 2 | Statistical tests

Statistical over-representation test

The input (or test) list is usually a list of genes or variants of interest. It can be a list of genes that are upregulated in the gene expression experiment, or a list of significant SNPs from a GWAS experiment, for example. The list is divided into groups based on annotation classification (molecular function, biological process, cellular component, PANTHER Protein Class or pathways). As many as four test lists can be uploaded for each analysis. A reference list, which usually contains all the genes/proteins from which the list was drawn, is divided into groups in the same way. PANTHER provides reference proteome datasets as default reference lists for all 131 genomes, so uploading a reference list is optional. If you work with genomes other than those 131, then you must prepare and upload a reference list. For each functional category (e.g., 'protein kinase' for GO Molecular Function, 'cell proliferation' for GO Biological Process or 'apoptosis signaling pathway' for PANTHER Pathway), the statistical test is applied to determine whether there is statistical over- or under-representation of genes/proteins in the test list relative to the reference list.

For *P* value calculation in the over-representation test, the 'expected' value is the number of genes you would expect in the test list for a particular PANTHER category, based on the reference list. For example, there are 20,000 genes in the reference list (e.g., the entire human genome); 440 of these genes map to the GO term 'induction of apoptosis'. Based on this, 2.2% (440 divided by 20,000) of the genes in the reference list are involved in induction of apoptosis. Now a test list that contains 500 genes is uploaded. Based on the reference list, it is expected that 11 genes (500 multiplied by 2.2%) in the test list would be involved in induction of apoptosis.

If for this biological process more genes are observed in the test list than expected, you have an over-representation (+) of genes involved in induction of apoptosis. If fewer genes are observed than expected, you have an under-representation (-). A *P* value is calculated then to determine whether the over- or under-representation is significant. For example, let us assume that 21 genes are observed in the test list and are involved in induction of apoptosis. Although this is almost twice the expected value, the *P* value is large and not significant (0.722). Alternatively, 35 observed genes would be very different from the expected value, so you would expect a small, significant *P* value (here it would be 6.21×10^{-7}). This small *P* value indicates that the result is nonrandom and potentially interesting, and thus worth looking at in closer detail. A *P* value cutoff of 0.05 is recommended as a starting point.

There are two statistical methods used in this test: Fisher's exact test and binomial test. They are both standard statistical methods commonly used in the field. Detailed descriptions of the methods can be found at https://en.wikipedia.org/wiki/Fisher%27s_exact_test (for Fisher's exact test) and https://en.wikipedia.org/wiki/Binomial_distribution (for the binomial distribution test).

Statistical enrichment test

The algorithm used in this test is the Mann-Whitney rank-sum test (*U* test)²⁰. A detailed description of the method can be found at https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test.

The statistical enrichment test is general enough to handle any numerical data, continuous or discontinuous, generated by experiments such as gene expression, proteomics or GWASs. First, a reference distribution is generated using all values from the input data (blue curve in Fig. 8). Although the test would work with any number of input genes, it is statistically meaningful to input all genes from the experiment, which could be the entire genome, or all genes on an array chip, for example. Then the entire list is divided into groups based on annotation classification (molecular function, biological process, cellular component, PANTHER Protein Class or pathways), and the distributions for each group are generated (red curve in Fig. 8). For each category in the classification, the probability that the distribution of input numerical values was drawn randomly from the reference distribution is calculated using the Mann-Whitney rank-sum test (*U* test)²⁰. A *P* value cutoff of 0.05 is recommended as a starting point. If the test returns a result with a *P* value of <0.05, it means that the distribution of numeric values from the functional category was not drawn randomly from the reference distribution—in other words, the distribution is significant.

An easier protocol to analyze genomes that are not in the PANTHER database

The current PANTHER (v.14.0) contains 131 of the most commonly researched genomes. One of the most frequent requests from PANTHER users is to include additional genomes beyond these 131. We have now dramatically simplified the steps for analyzing additional genomes. Before, users had to download a scoring tool to map their genes of interest to the PANTHER HMMs. This proved to be a major obstacle for researchers who were not highly trained in bioinformatics applications. To address this problem, we, in collaboration with InterPro¹⁸ and UniProt Reference Proteomes⁸, implemented an easier process to support 877 additional Reference Proteome genomes. We have precalculated the PANTHER HMM hits for all of the genes in each Reference Proteome (with UniProtKB identifiers), and stored the results in PANTHER Generic Mapping file format. Users just need to convert their gene list to UniProtKB IDs, and upload them to the website. The list of these supported genomes can be found in Supplementary Table 1.

In the following Procedure, we provide an updated step-by-step protocol for using the PANTHER tool, and subsequent sections present a detailed description of the anticipated results.

Materials

Equipment

A laptop or desktop computer with a high-speed Internet connection is highly recommended.

System requirements

- Operating system: Windows XP, Windows 7 or Windows 10 is recommended for PC users; MacOS 10.12 or higher is recommended for Mac users; minimum of 2 GB RAM recommended
- Browser: Firefox version 59 or higher; Google Chrome version 67 or higher; Microsoft Internet Explorer 11 or higher; Safari version 11.1 or higher. JavaScript and cookies must be enabled in your browser

Procedure

▲ CRITICAL Note that sample gene list files can be found in Supplementary Data 2–5 if you want to try the tools.

- 1 Access the PANTHER website by entering the URL (<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 131 genomes in the PANTHER database (see list of genomes at <http://pantherdb.org/panther/summaryStats.jsp>), option B if you are working with one of the Reference Proteome genomes other than the 131 in the database, option C if you are working with a genome that is not one of the Reference Proteome genomes or one of the 131 genomes in the database, or option D if you are working with genetic variant/SNP data. All input files must be in a simple text format (such as .txt or .tab format).

(A) Working with one of the 131 genomes in the PANTHER database ● Timing 15–30 min

- (i) Prepare input file(s) as a simple text file with gene or protein identifiers as the first column. If you want to use the ‘statistical enrichment’ test, you will also need a second column with a numerical value for each gene. Sample files in Supplementary Data 2 and 3 can be used for test purposes. Detailed instructions for the file format and supported IDs can be found in Box 3.

(B) Working with one of the Reference Proteome genomes now also supported in addition to the 131 genomes in the database ● Timing <15 min

- (i) Prepare input file(s) with UniProt IDs in the first column. If you want to use the statistical enrichment test, you will also need a second column with a numerical value for each gene. For more information about mapping other IDs to UniProt ones, please refer to Box 4. Sample files for this type of input list are available in Supplementary Data 4 and 5.

(C) Working with a genome that is not one of the 131 genomes in the database or one of the additional Reference Proteome genomes ● Timing Variable; usually 15 min to download the data, and 10 min to run the script

- (i) Prepare input file(s) in the PANTHER Generic Mapping format by mapping your IDs to the PANTHER HMM IDs via the procedure described in Box 4.

(D) Working with genetic variant/SNP data

- (i) Gather input file(s) in VCF file format. These files are usually generated by other tools used for genome sequencing and GWAS experiments. See Box 3 for more details about this file format. A sample VCF file (sample_vcf.txt) is available for test purposes in Supplementary Data 1.

▲ CRITICAL STEP The input file must be in a simple text file format (.txt or .tab). It must also use IDs supported in the PANTHER system and the correct tab-delimited format as described in Box 3.

Using online tools ● Timing 10 min

- 3 Upload the gene list to the PANTHER tool system using option A if you have a small list (<200 IDs), option B if you have a large list or VCF file, or option C if you previously saved the list in the Workspace.

(A) Uploading a small gene list (<200 IDs)

- (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 (solid blue arrow in Fig. 2).

(B) Uploading a large gene list

- (i) Upload the list file prepared in Step 2 by clicking on the ‘Browse’ button (open blue arrow in Fig. 2), and follow the online instructions to locate the file.

Box 3 | Input files

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 the additional columns are ignored. Microsoft Excel format is not accepted by the tool. Below are four file types that can be used.

- 1 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.
- 2 Previously exported text search results: any gene list on the PANTHER site (e.g., generated by a text search, or from a set of uploaded identifiers) can be saved as a text file (see Step 6A in the 'Anticipated results' section). This file contains the gene or protein identifiers in the first column. This file type is not associated with numeric values, so it cannot be used for the statistical enrichment test.
- 3 PANTHER Generic Mapping file: for IDs from organisms other than the 131 organisms in the PANTHER database, user-generated data containing mappings between those IDs and their corresponding PANTHER IDs can be used (see Box 2 for details about mapping). The file must be tab-delimited and contain the following columns:
 - (i) The first column can contain a list of unique IDs from the user.
 - (ii) The second column should be the corresponding PANTHER family or subfamily ID (e.g., PTHR10078 or PTHR10078:SF1) and is used to look up the association to GO and PANTHER terms (molecular function, biological process and pathway).
 - (iii) If you are uploading data for the statistical enrichment test tool, a third column is required that contains the numeric value of the experiment.
- 4 VCF: this is a text file format for storing sequence variation data. It was previously maintained by the 1000 Genomes Project. Currently, the group leading the management and expansion of the format is the Global Alliance for Genomics and Health Data Working group file format team. The VCF specification can be found on GitHub at <https://samtools.github.io/hts-specs/VCFv4.3.pdf>.

▲ **CRITICAL** If you are using the statistical enrichment test, numeric values should be provided to a designated column as described above. No blank or letter-based entries (e.g., N/A) are allowed in that column.

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'
- Model Organism Database (MOD). Examples:
 - MGI: 'MGI:2444594'
 - RGD: '1583843'
 - ZFIN: 'ZDB-GENE-060519-23'
 - FlyBase: 'FBgn0039374'
 - WormBase: 'WBGene00001400'
 - SGD: 'S000005863'
 - PomBase: 'SPBC947.14c'
 - dictyBase: 'DDB_G0291918'
 - TAIR: example: 'AT1G58450'
 - Ecoli (EcoGene ID): 'EG11161'
- UniGene: NCBI UniGene IDs. Examples: 'Hs.654587', 'At.36040'
- UniProtKB: UniProt accession. Example: 'O80536'
- UniProtKB-ID: UniProt ID. Example: 'AGAP3_HUMAN'

The primary IDs used for gene annotations in the PANTHER system are Ensembl gene ID or MOD IDs for genes, and UniProtKB IDs for proteins. All other supported IDs are mapped to those primary IDs using the UniProt 'ID mapping' mechanism (<https://www.uniprot.org/uploadlists/>).

If you are not certain about the ID type in your uploaded gene list, or if 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.

(C) Using a gene list previously saved in the Workspace

- (i) If you have previously saved your list in the Workspace, you can use it by clicking on the 'login' link (orange arrow in Fig. 2) and following the online instructions 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.

? TROUBLESHOOTING

Box 4 | PANTHER Generic Mapping

If you are working with a genome that is not one of the 131 in the PANTHER database, you can still use the tool. The back-end mechanism for such analysis is to convert your input list into a PANTHER Generic Mapping file, and then analyze. However, depending on the type of genome you are working on, there are two different approaches. We have precalculated the PANTHER Generic Mapping for all the Reference Proteome genomes. Therefore, if you are working with one of them, you can submit your list with UniProt IDs and the tool will take care of the rest. If you are working with a genome that is not in the Reference Proteome Project either, you can generate the mapping file using the PANTHER HMM scoring tool. The details of both approaches are described below.

A simple web interface

If you are working with one of the Reference Proteome genomes, you can use this interface to analyze your data with our tools. There are 877 genomes supported by this protocol. The list can be found in the 'Organism for ID list' drop-down menu right below the 'IDs from Reference Proteome Genome' option on the home page (Fig. 2). The details are described in the main text (Steps 2B and 4B).

One crucial requirement is that UniProt IDs must be used in the uploaded list. We recommend that you use the UniProt ID-mapping tool to convert other IDs to UniProt ones. The tool can be found at <https://www.uniprot.org/mapping/>.

Score sequences using the PANTHER HMM library

If the genome you are working with is not one of the Reference Proteome genomes, you will need to score your proteins against the PANTHER HMM library using the PANTHER Scoring Tool in order to generate the PANTHER Generic Mapping file.

The PANTHER HMM library is a library of HMMER3 models²¹ from over 1.7 million training sequences in 131 genomes. There are a total of over 120,000 models, of which 15,500 are family models and 104,500 are subfamily models. A subfamily model is built with a subset of genes in a family, often orthologs to each other, that carry out more specific biological functions. Each HMM model is annotated with a name, functions (GO terms) and pathways. The PANTHER Scoring Tool allows users to submit a large number of protein sequences in FASTA file format, score them against the PANTHER HMM library so that the sequence identifiers can be mapped to PANTHER HMM IDs and the functional groups can be annotated to them, and use them with the gene list analysis tools.

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

The location to Perl must be defined in your \$PATH variable or specified by the users in the arguments. If you have any questions on how to set up \$PATH, please contact your UNIX system administrator.

Procedure

- Download the following scripts and data.
 - pantherScore script (ftp://ftp.pantherdb.org/hmm_scoring/current_release/) ● **Timing 2 min**
 - PANTHER HMM library (ftp://ftp.pantherdb.org/panther_library/current_release/) ● **Timing 15 min**
 - HMMER3 (<http://eddylab.org/software/hmmer3/3.1b2/hmmer-3.1b2.tar.gz>) ● **Timing 5 min**
- Decompress all three downloads. ● **Timing 30 min**
- Define the location of the HMMER binaries in the \$PATH variables on your computer. This is usually done quite differently depending on the UNIX shell environment on your computer. Basically, you must 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, you must consult someone with IT knowledge to help you. ● **Timing 15 min for an expert with bioinformatics training, up to 1–2 h for others**

Use the following commands:

```
% tcsh
% cd pantherScore2.1
% source panther.cshrc
% ./pantherScore2.1.pl -l <panther_hmm_library> -D B -V -i <fasta file> -o <output file> -n
or
% ./pantherScore2.1.pl -l <panther_hmm_library> -D B -V -i <fasta file> -o <output file> -n -s
where
```

-l The path to the PANTHER HMM library downloaded above

-D display type for results

Options: B (best hit), A (all hits)

-i input FASTA file to score. A sample FASTA file is included in the downloaded data, called test.fasta

-o the output file

-n to display family and subfamily names in the output file

-s specify using hmmsearch program instead of hmmscan program (default) for scoring a large number of input sequences.

! CAUTION If you have a lot of sequences, you can split the FASTA file and run the script on multiple computers.

The output file is a tab-delimited file in the following format:

```
col 1: sequence ID
col 2: PANTHER accession (PTHRnnnnn for family HMMs, PTHRnnnnn:SFnn for subfamilies)
col 3: PANTHER family or subfamily name
col 4: HMM E-value score, as reported by HMMER tool
col 5: HMM bitscore, as reported by HMMER tool (not used by PANTHER)
col 6: alignment range of protein for this particular HMM
```

This file can be used as a PANTHER Generic Mapping file for the gene list analysis tool. By default, all results with E -value $< 10^{-3}$ are included in the file. However, the classification confidence is considered high when the E -value is $< 10^{-23}$, and medium when the E -value is $< 10^{-11}$. Users should feel free to filter the results depending on the confidence level of their choice.

If the statistical enrichment test is used, the numeric values must be inserted in the third column.

Gene List Analysis

Please refer to our article in [Nature Protocols](#) for detailed instructions on how to use this page.

Help Tips

Steps:

1. Select list and list type to analyze
2. Select Organism
3. Select operation

1. Enter IDs and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.

Enter IDs: [Supported IDs](#)

Upload IDs: [File format](#)

Select List Type:

- ☒ ID List
- ☐ Previously exported text search results
- ☐ Workspace list
- ☐ PANTHER Generic Mapping
- ☐ ID's from Reference Proteome Genome
- ☐ VCF File

Organism for id list: [Absidia glauca \(ABSG\)](#)

Flanking region: [20 Kb](#)

2. Select organism.

- Homo sapiens
- Mus musculus
- Rattus norvegicus
- Gallus gallus
- Danio rerio

3. Select Analysis.

- ☒ Functional classification viewed in gene list
- ☐ Functional classification viewed in graphic charts
- ☐ Statistical overrepresentation test
- ☐ Statistical enrichment test

Bar chart ☐ Pie chart

Use default settings ☐ Use default settings ☐

submit

Fig. 2 | The PANTHER home page with the gene list analysis tools. The solid blue arrow points to the 'Enter IDs' box, where the user can paste the ID list or type IDs, one per line, to upload the gene list. The open blue arrow points to the 'Browse' button, which the user can use to upload the list file from the computer. The orange arrow points to the 'login' link through which the user can access a saved list in the Workspace.

4. Select a corresponding list type in order for the tool to work properly. Four list types are supported by the tools: ID list, previously exported text search results, PANTHER generic mapping file, and VCF file format. Box 3 provides details about the list types.
 - (A) **File prepared using Step 2A**
 - (i) If you prepared your file using Step 2A, select 'ID List'.
 - (B) **File prepared using Step 2B**
 - (i) If you prepared your file using Step 2B, select 'IDs from Reference Proteome Genomes'. You must also select an organism from the 'Organism for ID list' drop-down menu. The genome name used here is a combination of the species name and the organism mnemonic (a five-letter symbol) used to specify the strain. For details on organism mnemonics, please visit <https://www.uniprot.org/taxonomy/>. The list of organisms is also presented in Supplementary Table 1.
 - (C) **File prepared using Step 2C**
 - (i) If you prepared your file using Step 2C, select 'PANTHER Generic Mapping'.
 - (D) **File prepared using Step 2D**
 - (i) If you prepared your file using Step 2D, select the VCF file format and specify a flanking region. A flanking region is the number of base pairs on either side of the gene on the chromosome, and is usually considered to be associated with the gene function by serving as the regulatory region of the gene. When an SNP is located in the flanking region, it will be mapped to the gene. By default, the tool uses 20 kb on either side of the gene as the flanking region for the analysis.

Table 1 | Summary of ID type, list type and organism selection for various types of genomes

Genome to analyze	One of 131 genomes in PANTHER	One of additional Reference Proteomes	Any other genome
ID type to use in the input file	IDs supported by PANTHER	UniProt	Any
Select list type	ID list	IDs from Reference Proteome genome	PANTHER Generic Mapping
Select organism	Required	Required	Not required
Reference list for over-representation test	Default or user upload (with IDs supported by PANTHER)	Default or user upload (PANTHER Generic Mapping)	User upload (PANTHER Generic Mapping)

- 5 Select an organism from the drop-down menu, which lists the 12 ‘model organisms’ first, followed by the remaining organisms ordered alphabetically. Note that organism selection was done for the ‘IDs from Reference Proteome Genomes’ option (Step 4B) and is not required for the ‘PANTHER Generic Mapping’ option (Step 4C). Table 1 summarizes the properties of input lists for various types of genomes.

! CAUTION There are two reasons to select an organism at this point. First, some identifiers, such as gene symbols, are not organism specific. Selecting an organism here ensures that the IDs are mapped to those in the organism you are interested in. Second, if the statistical over-representation test is selected, the default reference gene list is based on the selected organism.

- 6 In the ‘Select Analysis’ box, select one of the following four options (option A for functional classification viewed in gene list, option B for functional classification viewed in graphic chart, option C for statistical over-representation test or option D for statistical enrichment test) by clicking on the radio button, and then click on the ‘submit’ button.

(A) Functional classification viewed in gene list

- (i) Select ‘Functional classification viewed in gene list’ if you want to view the functional classifications of the genes in your list (Fig. 3). Once you reach the results page, you can make modifications, as described in the steps below, to view, customize and export the results.

? TROUBLESHOOTING

- (ii) Sort the list: you can always sort the list by clicking on any of the underlined column names. A yellow triangle appears in front of the column name that you choose to sort. The orientation of the triangle indicates whether the sort is ascending or descending.
- (iii) Customize columns: you can click on the ‘x’ button next to the column names to remove the column.
- (iv) Customize the gene list: users can customize the annotation data that they see on the results page. By default, only PANTHER Protein Class data are displayed.
- (v) Convert a list to another list type: select the genes you want to convert by clicking on the checkboxes. The default is for all genes in the list. Then choose the list type from the drop-down menu after ‘Convert List to:’ at the top of the page.
- (vi) Save the list: select the genes you want to save by clicking on the checkboxes. The default is for all genes in the list. The pull-down menu provides options for the save destination. One is the Workspace; you must register to save data to the Workspace, but the registration is free. When you make this selection, a pop-up window will ask you to name the list and add any comments. The name and comments can be edited at any time in the future from the Workspace page. Once the gene list is 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, so when you access a list in the future, all the information will be updated and current. You also can export a list to a file; the list will be exported as a tab-delimited file. You can then import the file into Excel or perform any desired post-processing, or you can view the list as text on the website.
- (vii) Use the pie chart view by clicking on the colorful pie chart icon (the glossary provides the meaning of the abbreviations). See Step 6B(i) for details about how to interpret the pie chart.



LOGIN REGISTER CONTACT US

Home About PANTHER Data PANTHER Tools Workspace Downloads Help/Tutorial

UPL14.0 New! PANTHER14.0 is generated from the 2018_04 release of ReferenceProteome dataset

PANTHER GENE LIST ? Customize Gene list

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

Display: 30 items per page Refine Search

Hits 1-30 of 523 [page: (1) 2 3 4 5 6 7 8 9 10 >>] Number of mapped ids found 523 IDs not found (16)

	Gene ID	Mapped IDs	Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class	Species
<input type="checkbox"/>	1. HUMAN HGNC=7135 UniProtKB=P51825	P51825	AF4/FMR2 family member 1 AFF1 ortholog	AF4/FMR2 FAMILY MEMBER 1 (PTHR10528:SF6)	transcription factor	Homo sapiens
<input type="checkbox"/>	2. HUMAN HGNC=143 UniProtKB=P68032	P68032	Actin, alpha cardiac muscle 1 ACTC1 ortholog	ACTIN, ALPHA CARDIAC MUSCLE 1 (PTHR11937:SF176)	actin and actin related protein	Homo sapiens
<input type="checkbox"/>	3. HUMAN HGNC=23989 UniProtKB=Q96QU6	Q96QU6	1-aminocyclopropane-1-carboxylate synthase-like protein 1 ACCS ortholog	1-AMINOCYCLOPROPANE-1-CARBOXYLATE SYNTHASE-LIKE PROTEIN 1 (PTHR43795:SF17)	transaminase	Homo sapiens
<input type="checkbox"/>	4. HUMAN HGNC=13800 UniProtKB=Q9GZZ6	Q9GZZ6	Neuronal acetylcholine receptor subunit alpha-10 CHRNA10 ortholog	NEURONAL ACETYLCHOLINE RECEPTOR SUBUNIT ALPHA-10 (PTHR18945:SF752)	GABA receptor acetylcholine receptor	Homo sapiens
<input type="checkbox"/>	5. HUMAN HGNC=5292 UniProtKB=P30939	P30939	5-hydroxytryptamine receptor 1F HTR1F ortholog	5-HYDROXYTRYPTAMINE RECEPTOR 1F (PTHR24247:SF34)	G-protein coupled receptor	Homo sapiens
<input type="checkbox"/>	6. HUMAN HGNC=165 UniProtKB=Q08043	Q08043	Alpha-actinin-3 ACTN3 ortholog	ALPHA-ACTININ-3 (PTHR11915:SF432)	-	Homo sapiens
<input type="checkbox"/>	7. HUMAN HGNC=30546 UniProtKB=Q6P4F2	Q6P4F2	Ferredoxin-2, mitochondrial FDX2 ortholog	FERREDOXIN-2, MITOCHONDRIAL (PTHR23426:SF25)	-	Homo sapiens
<input type="checkbox"/>	8. HUMAN HGNC=177 UniProtKB=Q03154	Q03154	Aminoacylase-1 ACY1 ortholog	AMINOACYLASE-1 (PTHR45892:SF1)	-	Homo sapiens

Fig. 3 | Results of functional classification displayed as a gene list page. The results are based on Supplementary Data 3.

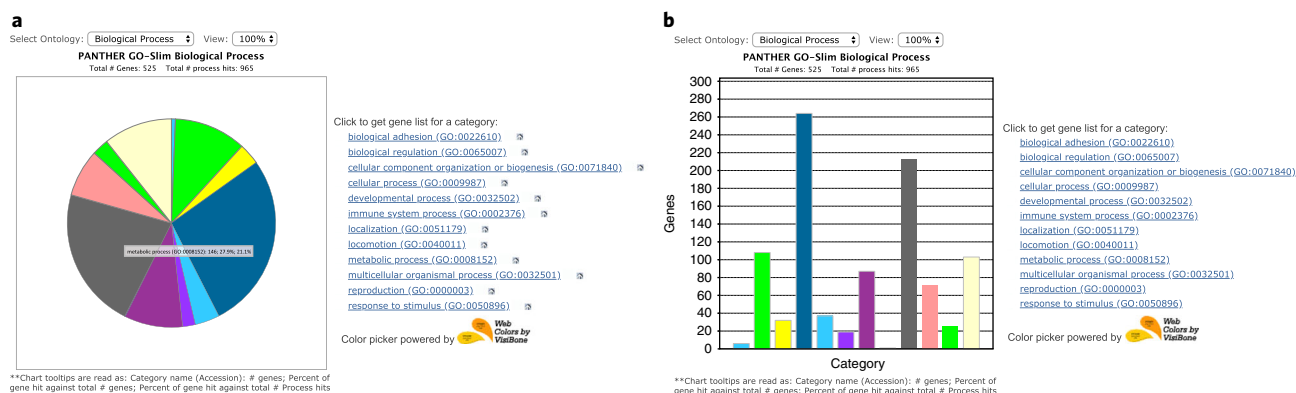



Fig. 4 | Graphical PANTHER results. a,b, PANTHER results from Supplementary Data 3 shown in pie chart (a) and bar chart (b) form. You can use the 'Select Ontology' drop-down menu to switch to the charts of different ontologies. Click on the chart section to display the child categories. Click on the legends on the right-hand side to retrieve the list of the genes for that category.

a Selection Summary:

Analysis Type: PANTHER Overrepresentation Test (Released 20181113)


Annotation Version and Release Date: PANTHER version 14.0 Released 2018-12-03

Analyzed List: sampleTestList_human_500 (Homo sapiens)
 There are duplicate IDs in the file. The unique set of IDs will be used. [Change](#)

Reference List: Homo sapiens (all genes in database) [Change](#)

Annotation Data Set: [PANTHER GO-Slim Biological Processes](#)

Test Type: ☒ Fisher's Exact ☐ Binomial

Correction: ☒ Calculate False Discovery Rate ☐ Use the Bonferroni correction for multiple testing  ☐ No correction

[Launch analysis](#)

b


Select lists to analyze

For example, you can upload two lists, one of up-regulated genes and one of down-regulated genes, from a differential mRNA microarray experiment.

UPLOAD OR SELECT LIST FROM YOUR WORKSPACE

Select Organism: (Not applicable for Generic mapping file or Reference Proteome ids)

Homo sapiens
 Mus musculus
 Rattus norvegicus
 Gallus gallus
 Danio rerio

Upload list:
 Please select list type...
☒ Gene, Transcript, Protein and Alternate ID
☐ PANTHER Generic Mapping File
☐ ID's from Reference Proteome Genome
 Organism for id list: [Absidia glauca \(ABSG\)](#)
☐ VCF file Flanking region: 20 Kb 

Upload list: [Browse...](#) No file selected. [supported IDs](#)

[Upload list](#)

If there are redundant IDs, only the first will be used in the analysis.
 Please [login](#) to be able to select lists from your workspace.

Uploaded and selected lists:
☒ sampleTestList_human_500

[Finished selecting lists](#)

Fig. 5 | User interface of the statistical over-representation test, allowing the user to configure the analysis criteria. a, The configuration page where the user can view the versions of the tool and annotation dataset, change/add test lists, change the reference list and annotation dataset, and select the test type. **b,** The user interface for users to change/add gene lists. Reproduced with permission from ref. ⁶, Springer Nature.

(B) Functional classification viewed in graphic chart

- (i) Select 'Functional classification viewed in graphic chart' to get the functional classification of the genes in your list displayed as either a pie chart or a bar chart (Fig. 4).

(C) Statistical over-representation test

- (i) Select 'Statistical overrepresentation test' to find functional classes that are statistically over- or under-represented in the input list compared with randomly selected genes.
- (ii) The default option uses the PANTHER reference list for the genome as the reference list, and the 'PANTHER GO-Slim Biological Process' annotation gene set for the analysis. You can change these (highly recommended) by deselecting the default option and updating the form on the 'Analysis Summary' panel. The user can repeat the tests with different annotation datasets and on different user data by updating the parameters at the top of the result page.
- (iii) Highly recommended: on the configuration page (Fig. 5a), you can change/add the 'Analyzed' list or change the 'Reference' list by clicking on the 'Change' button. A new

Box 5 | Change/add gene list or reference list in the over-representation test

The over-representation test allows you to analyze up to four gene lists at a time. You can also upload your own reference list instead of using the default one. To do so, make sure the 'default' checkbox is unselected before you click on the 'Submit' button.

On the configuration page (Fig. 5a), click on the 'Change' button. A new webpage will open (Fig. 5b), and you can do the following:

- 1 Click on the 'Browse' button.
- 2 Select the gene list from your computer.
- 3 Select the organism. The default is the organism selected when the first gene list is uploaded.
- 4 Select the list type. The default is the one selected when the first gene list is uploaded.
- 5 Click on 'Upload list'.
- 6 Steps 1–5 above can be repeated to upload up to four analyzed gene lists.
- 7 It is highly recommended to upload your own reference list file. The reference list should be the list of all the genes from which your smaller analysis list was selected. For example, in a list of differentially expressed genes, the reference list should contain only genes that were detected at all in the experiment, and thus potentially could have been on a list of genes derived from the experiment.
- 8 If the 'IDs from Reference Proteome Genome' option is selected, a default reference list is uploaded. If you decide to upload your own reference list, it has to be in the PANTHER Generic Mapping file format.
- 9 If the 'PANTHER Generic Mapping file' option is selected, a reference list in the same format must be uploaded here.
- 10 After all lists have been uploaded, click on the 'Finish selecting lists' button. The tool will return you to the selection summary panel page displaying the uploaded lists.

webpage will open, and you can upload those lists there (Fig. 5b). The detailed step-by-step procedure can be found in Box 5. On the configuration page, you can also change the 'Annotation' dataset from the drop-down menu. There are nine datasets to select. See Box 1 for the list and description of the datasets.

- (iv) On the configuration page, you can also select the statistical test for the over-representation analysis. The default is Fisher's exact test with FDR multiple test correction. You can also select the binomial test with FDR multiple test correction or Bonferroni correction.
- (v) Click on the 'Launch analysis' button.
- (vi) On the results page (Fig. 6), you can export the result table as a tab-delimited file by clicking on the 'Export results' button, or view the results graphically by using the 'View' drop-down menu and choosing one of the annotation classes (PANTHER Pathway, PANTHER GO-Slim or PANTHER Protein Class).

? TROUBLESHOOTING

(D) Statistical enrichment test

- (i) Select 'Statistical enrichment test'.
- (ii) The default setting is for analysis using the 'PANTHER GO-Slim Biological Process' annotation gene set. The user can change the settings by deselecting the default option and modifying on the 'Analysis Summary' panel.
- (iii) On the configuration page, you can also change the Annotation dataset from the drop-down menu. There are nine datasets available for analysis. See Box 1 for the list and description of the list.
- (iv) Select either FDR or Bonferroni multiple test for the analysis.
- (v) Click on the 'Submit' button if you choose to use the default setting from Step 6D(i), or the 'Launch Analysis' button after you modify the settings on the configuration page as in Step 6D(ii–iv).

? TROUBLESHOOTING

- (vi) On the results page (Fig. 7), you can export the result table in a tab-delimited file by clicking on the 'Export results' button or compare the distribution curve in graph view. To do so, check the box in front of the category or pathway of your interest, and then click on the 'Graph selected categories' button (Fig. 8).

! CAUTION To use the 'Statistical enrichment' tool, make sure that the uploaded gene list contains a second column with numerical values.

Troubleshooting

Troubleshooting advice can be found in Table 2.

Table 2 | Troubleshooting table

Step	Problem	Possible reasons	Solutions
3	Failed to upload the file	This is usually because the input file is 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 being left blank or marked as 'n/a', etc. Make sure that there are no blank rows in the first column
6A(i), 6C(vi), 6D(v)	IDs in the uploaded file do not have a mapped ID in PANTHER	The current PANTHER data are based on the April 2017 release of Reference Proteome Project and its ID mapping. It is possible that a small fraction of the IDs might not map because of outdated data in either the PANTHER database or your uploaded file.	There is no solution for outdated data. If you believe that you are using the current IDs, please do the following: (i) make sure that the IDs in the uploaded file are supported by PANTHER. Refer to Box 3 for Supported IDs. (ii) IDs from certain databases may contain a version number at the end (e.g., ".1" in NP_000242.1). Do not include the version number in the ID, and use just the main base number. Send feedback to feedback@pantherdb.org
6C(vi), 6D(v)	No results returned	By default, only the rows with significant P or Q values are returned	Click on the 'Click to display all results' link to see all the results

Timing

Step 1, launch website: instant
 Step 2A, prepare input file: 15–30 min
 Step 2B, prepare UniProt ID input file: <15 min
 Step 2C, prepare the PANTHER Generic Mapping file: depends on the speed of the Internet; usually 15 min to download the data, and 10 min to run the script
 Step 2D, gather the VCF files: instant
 Steps 3–6, using online tools: 10 min
 Box 4, step 1, downloading of scripts and data: 22 min
 Box 4, step 2, decompressing the downloaded scripts and data: 30 min
 Box 4, step 3, defining the location of the HMMER binaries: 15 min for an expert, up to 1–2 h for others

Anticipated results

As illustrated in Fig. 1, the tools use the UniProt ID Mapping to map the uploaded IDs to the IDs in the PANTHER annotation dataset. Not all IDs can be mapped, mainly because of outdated IDs used in the uploaded list. The result of the mapping is summarized at the top of the results page (Figs. 6 and 7). The user can access the list of genes by clicking on the counts.

Step 6A: functional classification tool viewed in gene list page

This tool returns the results as a gene list webpage (Fig. 3). The page displays all the IDs from the uploaded gene list and their mapped PANTHER sequence IDs, as well as annotation data. The page contains the following information:

- Gene ID—this is the identifier for the genes in the PANTHER library. The format is as follows: organism|gene database source=gene ID|protein database source=protein ID. The organism is denoted by the five-letter organism mnemonic code as listed at <https://www.uniprot.org/taxonomy/>. For example, HUMAN|HGNC=6218|UniProtKB=Q09470 is a human sequence; the gene sequence is from HGNC (Human Gene Nomenclature Committee; <https://www.genenames.org/>) with ID HGNC:6218, 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.

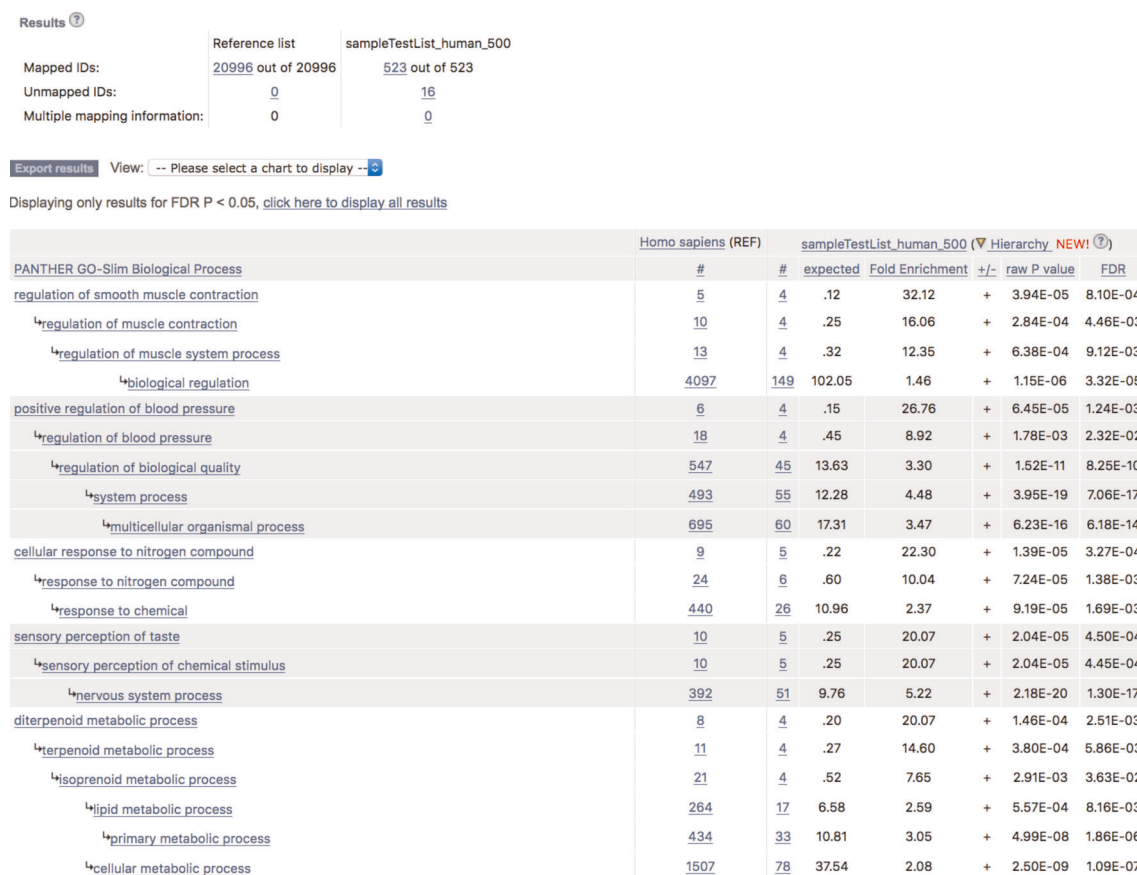


Fig. 6 | Result from the statistical over-representation test. The results are based on Supplementary Data 3. The summary of the results is displayed in a table. You can export the table as a tab-delimited file by clicking on the 'Export results' button. You can also view the results in other views by using the 'View' drop-down menu. If your analysis is done in a pathway as shown here, you can click on the pathway name and display the pathway diagram. A total of four test lists can be analyzed and viewed at the same time.

- Gene Name/Gene Symbol—the Entrez gene definition and gene symbol.
- PANTHER Family/Subfamily—the name and identifier of the PANTHER family or subfamily that the gene in the first column is in.
- PANTHER Protein Class—this is a PANTHER Index term describing protein classes. The default view shows only PANTHER Protein Class. You can view other annotation data by customizing the columns (see below).
- Species—the organism of the gene in column 1.

Step 6B: functional classification tool viewed in graphical chart

There are two types of charts in this option: pie chart (Fig. 4a) and bar chart (Fig. 4b). In either chart, the default view displays an overview of all ontology terms at the first (or most general) level within the same ontology. When a slice of the pie chart or a bar in the bar chart, which represents an ontology term, is clicked on, a new chart will appear that contains its child ontology terms.

Because one gene can be classified under more than one term, the pie chart is calculated on the basis of the number of 'hits' to the terms over the total number of class hits. A class hit means independent ontology terms. For example, if a gene is classified under two ontology terms that are not parent or child to each other, it counts as two class hits.

For the bar chart, the *x*-axis represents the classification terms (GO terms or pathway terms), and the *y*-axis indicates the number of genes annotated to each term.

When you place the computer mouse pointer over a slice or bar, the category name and a series of counts are displayed. In our example in Fig. 4a, the name is the GO term "metabolic process (GO:0008152)" followed by

- 1 the number of genes (212) from the uploaded list that are classified under the term 'metabolic process',

Results ?

Analysis details:

Mapped IDs: [18386](#)Unmapped IDs: [987](#)Multiple mapping information: [0](#)[Graph selected categories](#)[Export results](#)Displaying only results with $P < 0.05$; [click here to display all results](#)

PANTHER Pathways	#	+/-	P value	FDR
<input type="checkbox"/> PDGF signaling pathway (P00047)	141	-	2.34E-10	3.84E-08
<input type="checkbox"/> Ras Pathway (P04393)	72	-	1.61E-09	1.32E-07
<input type="checkbox"/> Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	251	-	1.81E-09	9.87E-08
<input type="checkbox"/> Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway (P00027)	123	-	1.25E-07	5.11E-06
<input type="checkbox"/> De novo purine biosynthesis (P02738)	27	-	1.38E-07	4.52E-06
<input type="checkbox"/> VEGF signaling pathway (P00056)	65	-	2.50E-07	6.83E-06
<input type="checkbox"/> Histamine H1 receptor mediated signaling pathway (P04385)	42	-	5.91E-07	1.38E-05
<input type="checkbox"/> Integrin signalling pathway (P00034)	187	-	1.23E-06	2.52E-05
<input type="checkbox"/> B cell activation (P00010)	66	-	1.28E-06	2.33E-05
<input type="checkbox"/> EGF receptor signaling pathway (P00018)	132	-	1.57E-06	2.58E-05
<input type="checkbox"/> Gonadotropin-releasing hormone receptor pathway (P06664)	230	-	2.11E-06	3.15E-05
<input type="checkbox"/> FGF signaling pathway (P00021)	118	-	3.02E-06	4.13E-05
<input type="checkbox"/> p53 pathway feedback loops 2 (P04398)	50	-	4.44E-06	5.61E-05
<input type="checkbox"/> Endothelin signaling pathway (P00019)	80	-	1.08E-05	1.27E-04

Fig. 7 | The results from the statistical enrichment test. The results are based on Supplementary Data 2. The output of the tool is shown with a list of P values for each comparison between a functional category distribution and the reference distribution.

- the percentage (40.4%) of genes classified under ‘metabolic process’ (212) among the total number of genes (525), and
- the percentage (22%) of genes classified under this metabolic process (212) out of the total number of annotations (965).

The results can be exported as a tab-delimited file via the ‘Export’ link on the page.

Step 6C: statistical over-representation test

The results obtained with Fisher’s exact test are displayed in a table (Fig. 6). If one test gene list is uploaded, the table contains eight columns of data (seven for the binomial distribution test).

- The first column contains the name of the PANTHER classification category. If you are doing this analysis using the PANTHER Pathway annotation set, you can click on the pathway name to view the corresponding pathway diagram. For other analyses, the link will take you to more information about that category.
- The second column contains the number of genes in the reference list that map to this particular PANTHER classification category.
- The third column contains the observed number of genes in your uploaded list that map to this PANTHER classification category.
- The fourth column contains the expected value (Box 2), which is the number of genes you would expect in your list for this PANTHER category, based on the reference list.
- The fifth column shows the fold enrichment, which is the ratio of the value of column 3 (observed number) over that of column 4 (expected number).
- The sixth column shows either + or -. A plus sign indicates over-representation of this category in the analyzed list: you observed more genes than expected on the basis of the reference list (for this

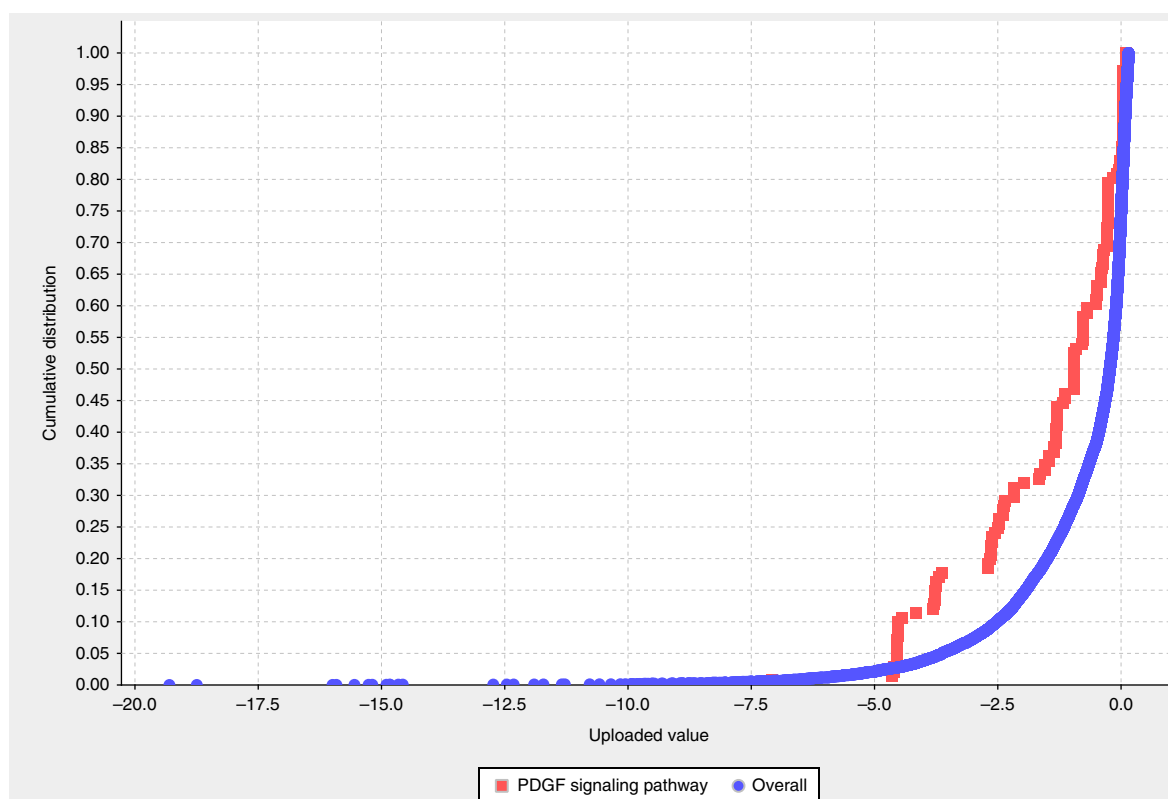


Fig. 8 | Graph of the results from the enrichment test. A comparison of the distributions from the PDGF signaling pathway and the reference. Reproduced with permission from ref. ⁶, Springer Nature.

category, the number of genes in your list is greater than the expected value). Conversely, a negative sign indicates under-representation, that is, fewer genes than expected.

- 7 For the results from Fisher's exact test, this column shows the raw *P* values. For the binomial distribution test, this column is the *P* value as determined by the binomial statistic. In either case, this is the probability that the number of genes you observed in this category occurred by chance (randomly), as determined by your reference list.
- 8 The eighth column is the *Q* value (the adjusted *P* value, reflecting the FDR) as calculated via the Benjamini–Hochberg procedure¹⁷. By default, a critical value of 0.05 is used to filter results, so all results shown are valid for an overall FDR < 0.05 even if the FDR for an individual comparison is greater than that value. This value is output when the Fisher's exact test option is selected.

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

The default filter will already have limited your results to those with statistical significance (overall FDR < 0.05, meaning that you can expect that <95% of the enriched classes are true associations). You can view all results regardless of FDR by clicking on the 'click here to view all results' link above the table. While this can be useful for troubleshooting, we caution that the additional results are not statistically significant.

By default, the results are sorted 'hierarchically' to help users understand the hierarchical relations between over-represented or enriched functional classes. Sorting is done by only the most specific subclass first, with its parent terms indented directly below it. These are all related classes in an ontology, and are often interpretable as a group rather than individually. If a term is a parent of more than one term in the results table, it is shown only under its first descendant. You can still sort by a single column (e.g., fold change or *P* value) by clicking on that column header.

From this result page, the user can export various statistics via the drop-down menu next to the 'Export results' button, or view the list of genes/proteins in any functional group by clicking on the listed counts. When PANTHER Pathways is chosen as the annotation set, clicking on the pathway name brings up pathway diagrams. The resulting pathway diagram can be exported as an image file (.png) through the 'Export' function on the page.

Step 6D: statistical enrichment test

The returned results are displayed in a table with four essential columns of data (Fig. 7).

- 1 If you are doing this analysis with the PANTHER Pathway annotation set, you can click on the pathway name to view the corresponding pathway diagram. For other analyses, the link will take you to more information about that category.
- 2 The second column contains the number of genes that map to this particular PANTHER classification category.
- 3 The third column shows either + or -. 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.
- 4 The fourth column contains the P value as calculated from the Mann–Whitney U test (Wilcoxon rank-sum test) (Box 2). A large P value indicates that the genes for this category have a distribution that is similar to that obtained by randomly choosing genes from the overall distribution. In other words, the values of the uploaded genes for this category have a distribution similar to that of 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. A cutoff of 0.05 is recommended as a starting point (note that these are already adjusted for multiple testing using the Bonferroni correction).
- 5 If FDR is selected as the multiple test, a fifth column is returned to display the Q value (the adjusted P value, reflecting the FDR), as calculated by the Benjamini–Hochberg procedure¹⁷. By default, a critical value of 0.05 is used to filter results, so all results shown are valid for an overall FDR < 0.05 even if the FDR for an individual comparison is greater than that value.

By default, only the results with statistical significance are displayed. You can view all results regardless of the P value by clicking on the ‘click here to view all results’ link above the table. While this can be useful for troubleshooting, we caution that the additional results are not statistically significant. Again, the results are sorted hierarchically by default (see details in the previous section).

To have a visual representation of these distributions, select the checkboxes of the categories of interest, and click on the ‘Graph selected categories’ button. The graph will be displayed in a new window (Fig. 8). The x -axis is your uploaded value. 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 is 0.3 for the red curve and 0.1 for the blue curve. This means that 30% of the uploaded genes have a value of -0.25 or less, but only 10% of the overall genes have a value of -0.25 or less. In other words, it shows that the distribution of the category tends to be smaller than the overall distribution. We find that visualization is essential for interpreting any deviation between the functional category distribution and the overall distribution.

The user can also view genes/proteins in each category from the output page by clicking on the listed counts. In addition, for pathways, clicking on the pathway name will bring up the pathway diagram (Fig. 8), which can be exported as an image file (.png) via the ‘Export’ function from the page.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

Source codes for various PANTHER software, including the PANTHER scoring tool, and the tree-building tool (GIGA), can be downloaded at <http://www.pantherdb.org/downloads/index.jsp>.

Data availability

All PANTHER data are publicly available and can be downloaded at <http://www.pantherdb.org/downloads/index.jsp>.

References

1. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim, and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
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. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.* **31**, 334–341 (2003).
5. 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).
6. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).
7. 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).
8. UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
9. The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
10. Fabregat, A. et al. The Reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
11. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
12. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **45**, D353–D361 (2018).
13. Caspi, R. et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).
14. Slenter, D. N. et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46**, D661–D667 (2018).
15. Kerrien, S. et al. The intact molecular interaction database in 2012. *Nucleic Acids Res.* **40**, D841–D846 (2012).
16. Chatr-Aryamontri, A. et al. The biogrid interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379 (2017).
17. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* **57**, 289–300 (1995).
18. Finn, R. D. et al. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
19. Mi, H. & Thomas, P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.* **563**, 123–140 (2009).
20. Clark, A. G. et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
21. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).

Acknowledgements

The authors thank the InterPro team, especially L. Richardson and R. Finn, for their support in providing PANTHER matching files used to generate PANTHER Generic Mapping files; the Reference Proteome team, especially M. Martin, for their support in providing up-to-date Reference Proteome datasets; the Gene Ontology Consortium, especially P. Gaudet, M. Feuerhahn and S. Lewis, for their support in providing GO phylogenetic annotation data; and the Reactome team, especially R. Haw, for their support in providing the Reactome dataset. The authors also thank A. Toga and the BDOS (Big Data for Discovery Science) Project for providing the funding to develop the software for supporting genetic variant analysis. This work is supported by NIH/NHGRI U41HG002273 and NSF 1458808 to P.D.T., and NIH U54EB020406 to A. Toga. Funding for open-access charge: University of Southern California.

Author contributions

A.M. developed the software code for the website. X.H. generated the content of the PANTHER database and provided administrative support to maintain the database. D.E. built the current PANTHER release (v.14.0). C.M. developed the workflow to integrate Reactome data into the analysis. X.G. helped in the development and implementation of Fisher's exact test in PANTHER. H.M. developed the new process to generate the PANTHER Generic Mapping file, and supervised the project. P.D.T. provided funding and supervised the project. H.M. wrote the manuscript, with contributions from all other authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41596-019-0128-8>.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to H.M.P.D.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 18 August 2018; Accepted: 3 January 2019;

Published online: 25 February 2019

Related link**Key references using this protocol**

Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. *Nucleic Acids Res.* **47**, D419–D426 (2019): <https://doi.org/10.1093/nar/gky1038>

The Gene Ontology Consortium. *Nucleic Acids Res.* **45**, D331–D338 (2017): <https://doi.org/10.1093/nar/gkw1108>

Mi, H. & Thomas, P. *Methods Mol. Biol.* **563**, 123–140 (2009): https://doi.org/10.1007/978-1-60761-175-2_7

Protocol update to:

Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. *Nat. Protoc.* **8**, 1551–1566 (2013): <https://doi.org/10.1038/nprot.2013.092>

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

n/a

Data analysis

n/a

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Provide your data availability statement here.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<input type="text" value="n/a"/>
Data exclusions	<input type="text" value="n/a"/>
Replication	<input type="text" value="n/a"/>
Randomization	<input type="text" value="n/a"/>
Blinding	<input type="text" value="n/a"/>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging