

# Package ‘GOfuncR’

May 17, 2024

**Type** Package

**Title** Gene ontology enrichment using FUNC

**Version** 1.24.0

**Date** 2024-02-03

**Author** Steffi Grote

**Maintainer** Steffi Grote <grote.steffi@gmail.com>

**Description** GOfuncR performs a gene ontology enrichment analysis based on the ontology enrichment software FUNC. GO-annotations are obtained from OrganismDb or OrgDb packages ('Homo.sapiens' by default); the GO-graph is included in the package and updated regularly (01-May-2021). GOfuncR provides the standard candidate vs. background enrichment analysis using the hypergeometric test, as well as three additional tests:

- (i) the Wilcoxon rank-sum test that is used when genes are ranked,
- (ii) a binomial test that is used when genes are associated with two counts and
- (iii) a Chi-square or Fisher's exact test that is used in cases when genes are associated with four counts.

To correct for multiple testing and interdependency of the tests, family-wise error rates are computed based on random permutations of the gene-associated variables.

GOfuncR also provides tools for exploring the ontology graph and the annotations, and options to take gene-length or spatial clustering of genes into account. It is also possible to provide custom gene coordinates, annotations and ontologies.

**License** GPL (>= 2)

**Imports** Rcpp (>= 0.11.5), mapplots (>= 1.5), gtools (>= 3.5.0), GenomicRanges (>= 1.28.4), IRanges, AnnotationDbi, utils, grDevices, graphics, stats,

**Depends** R (>= 3.4), vioplot (>= 0.2),

**LinkingTo** Rcpp

**Suggests** Homo.sapiens, BiocStyle, knitr, markdown, rmarkdown, testthat

**VignetteBuilder** knitr

**biocViews** GeneSetEnrichment, GO

**NeedsCompilation** yes

**git\_url** <https://git.bioconductor.org/packages/GOfuncR>  
**git\_branch** RELEASE\_3\_19  
**git\_last\_commit** 74e3c1a  
**git\_last\_commit\_date** 2024-04-30  
**Repository** Bioconductor 3.19  
**Date/Publication** 2024-05-17

Contents

get_anno_categories . . . . .	2
get_anno_genes . . . . .	4
get_child_nodes . . . . .	6
get_ids . . . . .	7
get_names . . . . .	8
get_parent_nodes . . . . .	9
go_enrich . . . . .	11
plot_anno_scores . . . . .	14
refine . . . . .	16
<b>Index</b>	<b>19</b>

---

get_anno_categories	<i>Get all associated ontology categories for the input genes</i>
---------------------	---

---

Description

Returns all associated GO-categories given a vector of gene-symbols, e.g. `c('SPAG5', 'BTC')`.

Usage

```
get_anno_categories(genes, database = 'Homo.sapiens', annotations = NULL,  
term_df = NULL, godir = NULL, silent = FALSE)
```

Arguments

- `genes` a `character()` vector of gene-symbols, e.g. `c('SPAG5', 'BTC')`.
- `database` optional `character()` defining an `OrganismDb` or `OrgDb` annotation package from Bioconductor, like `'Mus.musculus'` (mouse) or `'org.Pt.eg.db'` (chimp).
- `annotations` optional `data.frame()` with two `character()` columns: gene-symbols and GO-categories. Alternative to `'database'`.
- `term_df` optional `data.frame()` with an ontology `'term'` table. Alternative to the default integrated GO-graph or `godir`.
- `godir` optional `character()` specifying a directory that contains the ontology table `'term.txt'`. Alternative to the default integrated GO-graph or `term_df`.
- `silent` logical. If `TRUE` all output to the screen except for warnings and errors is suppressed.

## Details

Besides the default 'Homo.sapiens', also other OrganismDb or OrgDb packages from Bioconductor, like 'Mus.musculus' (mouse) or 'org.Pt.eg.db' (chimp), can be used. It is also possible to directly provide a dataframe with annotations, which is then searched for the input genes and filtered for GO-categories that are present in the ontology.

By default the package's integrated ontology is used, but a custom ontology can be defined, too. For details on how to use a custom ontology with `term_df` or `godir` please refer to the package's vignette. The advantage of `term_df` over `godir` is that the latter reads the file 'term.txt' from disk and therefore takes longer.

## Value

a `data.frame()` with four columns: `gene` (`character()`), `GO-ID` (`character()`), `GO-name` (`character()`) and `GO-domain` (`character()`).

## Note

This gives only direct annotations of genes to GO-categories. By definition genes are also indirectly annotated to all parent nodes of those categories. Use [get\\_parent\\_nodes](#) to get the higher level categories of the directly annotated GO-categories.

Also note that GO-categories which are not represented or obsolete in the internal GO-graph of GOfuncR or the custom ontology provided through `term_df` or `godir` are removed to be consistent with the annotations used in [go\\_enrich](#).

## Author(s)

Steffi Grote

## References

[1] Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25-29.

## See Also

[get\\_anno\\_genes](#)  
[get\\_parent\\_nodes](#)  
[get\\_names](#)

## Examples

```
## get the GO-annotations for two random genes
anno1 = get_anno_categories(c('BTC', 'SPAG5'))
head(anno1)
```

get\_anno\_genes

*Get genes that are annotated to GO-categories***Description**

Given a vector of GO-IDs, e.g. `c('GO:0072025','GO:0072221')` this function returns all genes that are annotated to those GO-categories. This includes genes that are annotated to any of the child nodes of a GO-category.

**Usage**

```
get_anno_genes(go_ids, database = 'Homo.sapiens', genes = NULL, annotations = NULL,
               term_df = NULL, graph_path_df = NULL, godir = NULL)
```

**Arguments**

<code>go_ids</code>	character() vector of GO-IDs, e.g. <code>c('GO:0051082', 'GO:0042254')</code> .
<code>database</code>	optional character() defining an OrganismDb or OrgDb annotation package from Bioconductor, like <code>'Mus.musculus'</code> (mouse) or <code>'org.Pt.eg.db'</code> (chimp).
<code>genes</code>	optional character() vector of gene-symbols. If defined, only annotations of those genes are returned.
<code>annotations</code>	optional data.frame() with two character() columns: gene-symbols and GO-categories. Alternative to <code>'database'</code> .
<code>term_df</code>	optional data.frame() with an ontology <code>'term'</code> table. Alternative to the default integrated GO-graph or <code>godir</code> . Also needs <code>graph_path_df</code> .
<code>graph_path_df</code>	optional data.frame() with an ontology <code>'graph_path'</code> table. Alternative to the default integrated GO-graph or <code>godir</code> . Also needs <code>term_df</code> .
<code>godir</code>	optional character() specifying a directory that contains the ontology tables <code>'term.txt'</code> and <code>'graph_path.txt'</code> . Alternative to the default integrated GO-graph or <code>term_df</code> + <code>graph_path_df</code> .

**Details**

Besides the default `'Homo.sapiens'`, also other OrganismDb or OrgDb packages from Bioconductor, like `'Mus.musculus'` (mouse) or `'org.Pt.eg.db'` (chimp), can be used. It is also possible to directly provide a data.frame() with annotations, which is then searched for the input GO-categories and their child nodes.

By default the package's integrated GO-graph is used to find child nodes, but a custom ontology can be defined, too. For details on how to use a custom ontology with `term_df` + `graph_path_df` or `godir` please refer to the package's vignette. The advantage of `term_df` + `graph_path_df` over `godir` is that the latter reads the files `'term.txt'` and `'graph_path.txt'` from disk and therefore takes longer.

**Value**

A data.frame() with two columns: GO-IDs (character()) and the annotated genes (character()). The output is ordered by GO-ID and gene-symbol.

**Author(s)**

Steffi Grote

**References**

[1] Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25-29.

**See Also**

[get\\_anno\\_categories](#)  
[get\\_ids](#)  
[get\\_names](#)  
[get\\_child\\_nodes](#)  
[get\\_parent\\_nodes](#)

**Examples**

```
## find all genes that are annotated to GO:0000109
## ("nucleotide-excision repair complex")
get_anno_genes(go_ids='GO:0000109')

## find out wich genes from a set of genes
## are annotated to some GO-categories
genes = c('AGTR1', 'AN01', 'CALB1', 'GYG1', 'PAX2')
gos = c('GO:0001558', 'GO:0005536', 'GO:0072205', 'GO:0006821')
anno_genes = get_anno_genes(go_ids=gos, genes=genes)
# add the names and domains of the GO-categories
cbind(anno_genes ,get_names(anno_genes$go_id)[,2:3])

## find all annotations to GO-categories containing 'serotonin receptor'
sero_ids = get_ids('serotonin receptor')
sero_anno = get_anno_genes(go_ids=sero_ids$go_id)
# merge with names of GO-categories
head(merge(sero_ids, sero_anno))
```

---

get_child_nodes	<i>Get all child nodes of gene ontology categories</i>
-----------------	--

---

### Description

Returns all child nodes (sub-categories) of GO-categories given their GO-IDs, e.g. `c('GO:0042254', 'GO:0000109')`. The output also states the shortest distance to the child node. Note that a GO-ID itself is also considered as child with distance 0.

### Usage

```
get_child_nodes(go_ids, term_df = NULL, graph_path_df = NULL, godir = NULL)
```

### Arguments

<code>go_ids</code>	a <code>character()</code> vector of GO-IDs, e.g. <code>c('GO:0051082', 'GO:0042254')</code> .
<code>term_df</code>	optional <code>data.frame()</code> with an ontology 'term' table. Alternative to the default integrated GO-graph or <code>godir</code> . Also needs <code>graph_path_df</code> .
<code>graph_path_df</code>	optional <code>data.frame()</code> with an ontology 'graph_path' table. Alternative to the default integrated GO-graph or <code>godir</code> . Also needs <code>term_df</code> .
<code>godir</code>	optional <code>character()</code> specifying a directory that contains the ontology tables 'term.txt' and 'graph_path.txt'. Alternative to the default integrated GO-graph or <code>term_df</code> + <code>graph_path_df</code> .

### Details

By default the package's integrated GO-graph is used, but a custom ontology can be defined, too. For details on how to use a custom ontology with `term_df` + `graph_path_df` or `godir` please refer to the package's vignette. The advantage of `term_df` + `graph_path_df` over `godir` is that the latter reads the files 'term.txt' and 'graph\_path.txt' from disk and therefore takes longer.

### Value

a `data.frame()` with four columns: parent GO-ID (`character()`), child GO-ID (`character()`), child GO-name (`character()`) and distance (`numeric()`).

### Author(s)

Steffi Grote

### References

[1] Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, 25-29.

**See Also**

[get\\_names](#)  
[get\\_parent\\_nodes](#)

**Examples**

```
## get the child nodes (sub-categories) of two random GO-IDs
child_nodes = get_child_nodes(c('GO:0090070', 'GO:0000112'))
child_nodes
```

---

get_ids	<i>Get the ID of a GO-category given its name</i>
---------	---

---

**Description**

Returns GO-categories given (part of) their name. Matching is not case-sensitive.

**Usage**

```
get_ids(go_name, term_df = NULL, godir = NULL)
```

**Arguments**

go_name	character(). (partial) name of a GO-category
term_df	optional data.frame() with an ontology 'term' table. Alternative to the default integrated GO-graph or godir.
godir	optional character() specifying a directory that contains the ontology table 'term.txt'. Alternative to the default integrated GO-graph or term_df.

**Details**

For details on how to use a custom ontology with term\_df or godir please refer to the package's vignette. The advantage of term\_df over godir is that the latter reads the file 'term.txt' from disk and therefore takes longer.

**Value**

a data.frame() with three columns: the full names (character()) of the GO-categories that contain go\_name; together with the GO-domain ('cellular\_component', 'biological\_process' or 'molecular\_function') and the GO-category IDs (character()).

**Note**

This is just a `grep(..., ignore.case=TRUE)` on the node names of the ontology. More sophisticated searches, e.g. with regular expressions, could be performed on the table returned by `get_ids('')` which lists all non-obsolete GO-categories.

**Author(s)**

Steffi Grote

**References**

[1] Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25-29.

**See Also**

[get\\_names](#)  
[get\\_parent\\_nodes](#)  
[get\\_child\\_nodes](#)

**Examples**

```
## get GO-IDs of categories that contain 'gabaergic' in their names
get_ids('gabaergic')

## get GO-IDs of categories that contain 'blood-brain barrier' in their names
get_ids('blood-brain barrier')

## get all valid GO-categories
all_nodes = get_ids('')
head(all_nodes)
```

---

get\_names

---

*Get the full names of gene ontology categories given the IDs*


---

**Description**

Returns the full names and the domains of GO-categories given the GO-IDs, e.g. 'GO:0042254'. By default the package's integrated GO-graph is used, but a custom ontology can be defined, too.

**Usage**

```
get_names(go_ids, term_df = NULL, godir = NULL)
```

**Arguments**

go_ids	a character() vector of GO-IDs, e.g. c('GO:0051082', 'GO:0042254').
term_df	optional data.frame() with an ontology 'term' table. Alternative to the default integrated GO-graph or godir.
godir	optional character() specifying a directory that contains the ontology table 'term.txt'. Alternative to the default integrated GO-graph or term_df.



## Details

For details on how to use a custom ontology with `term_df` or `godir` please refer to the package's vignette. The advantage of `term_df` over `godir` is that the latter reads the file 'term.txt' from disk and therefore takes longer.

## Value

a `data.frame()` with three columns: `go_id` (`character()`), `go_name` (`character()`) and `root_node` (`domain, character()`).

## Author(s)

Steffi Grote

## References

[1] Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, 25-29.

## See Also

[get\\_ids](#)  
[get\\_child\\_nodes](#)  
[get\\_parent\\_nodes](#)

## Examples

```
## get the full names of three random GO-IDs  
get_names(c('GO:0051082', 'GO:0042254', 'GO:0000109'))
```

---

<code>get_parent_nodes</code>	<i>Get all parent nodes of gene ontology categories</i>
-------------------------------	---

---

## Description

Returns all parent nodes (higher level categories) of GO-categories given their GO-IDs, e.g. `c('GO:0042254', 'GO:0000109')`. The output also states the shortest distance to the parent node. Note that a GO-ID itself is also considered as parent with distance 0.

## Usage

```
get_parent_nodes(go_ids, term_df = NULL, graph_path_df = NULL, godir = NULL)
```

**Arguments**

<code>go_ids</code>	a <code>character()</code> vector of GO-IDs, e.g. <code>c('GO:0051082', 'GO:0042254')</code> .
<code>term_df</code>	optional <code>data.frame()</code> with an ontology 'term' table. Alternative to the default integrated GO-graph or <code>godir</code> . Also needs <code>graph_path_df</code> .
<code>graph_path_df</code>	optional <code>data.frame()</code> with an ontology 'graph_path' table. Alternative to the default integrated GO-graph or <code>godir</code> . Also needs <code>term_df</code> .
<code>godir</code>	optional <code>character()</code> specifying a directory that contains the ontology tables 'term.txt' and 'graph_path.txt'. Alternative to the default integrated GO-graph or <code>term_df</code> + <code>graph_path_df</code> .

**Details**

By default the package's integrated GO-graph is used, but a custom ontology can be defined, too. For details on how to use a custom ontology with `term_df` + `graph_path_df` or `godir` please refer to the package's vignette. The advantage of `term_df` + `graph_path_df` over `godir` is that the latter reads the files 'term.txt' and 'graph\_path.txt' from disk and therefore takes longer.

**Value**

a `data.frame()` with four columns: child GO-ID (`character()`), parent GO-ID (`character()`), parent GO-name (`character()`) and distance (`numeric()`).

**Author(s)**

Steffi Grote

**References**

[1] Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, 25-29.

**See Also**

[get\\_names](#)  
[get\\_child\\_nodes](#)

**Examples**

```
## get the parent nodes (higher level GO-categories) of two random GO-IDs
parents = get_parent_nodes(c('GO:0051082', 'GO:0042254'))
parents
```

go\_enrich

*Test gene sets for enrichment in GO-categories***Description**

Tests GO-categories for enrichment of user defined gene sets, using either the hypergeometric (default), Wilcoxon rank-sum, binomial or 2x2 contingency table test.

**Usage**

```
go_enrich(genes, test = 'hyper', n_randsets = 1000, organismDb = 'Homo.sapiens', gene_len = FALSE,
          regions = FALSE, circ_chrom = FALSE, silent = FALSE, domains = NULL, orgDb = NULL,
          txDb = NULL, annotations = NULL, gene_coords = NULL, godir = NULL)
```

**Arguments**

genes	a data.frame() with gene-symbols (character()) in the first column and test-dependent additional numeric() columns: If test='hyper' (default) a second column with 1 for candidate genes and 0 for background genes. If no background genes are defined, all remaining genes from the internal dataset are used as background. All candidate genes are implicitly part of the background gene set. If test='wilcoxon' a second column with the score that is associated with each gene. If test='binomial' two additional columns with two gene-associated integers. If test='contingency' four additional columns with four gene-associated integers. For test='hyper' with regions=TRUE the first column describes chromosomal regions ('chr:start-stop', e.g. '9:0-39200000'). Note that each gene has to be unique in the data.frame; e.g. for test='wilcoxon' a gene cannot have two different scores assigned.
test	character(). 'hyper' (default) for the hypergeometric test, 'wilcoxon' for the Wilcoxon rank-sum test, 'binomial' for the binomial test and 'contingency' for the 2x2-contingency table test (fisher's exact test or chi-square).
n_randsets	integer defining the number of random sets for computing the FWER.
organismDb	character(). Annotation package for GO-annotations and gene coordinates. Besides the default 'Homo.sapiens' also 'Mus.musculus' and 'Rattus.norvegicus' are currently available on Bioconductor.
gene_len	logical. Correct for gene length. If TRUE the probability of a background gene to be chosen as a candidate gene in a random set is dependent on the gene length. If FALSE genes are chosen randomly with equal probability each. Only for test='hyper'.
regions	logical. If TRUE chromosomal regions are analyzed, and genes are automatically extracted from the regions defined in genes[,1] as e.g. '9:0-39200000'. Note that this option requires the input of background regions.

<code>circ_chrom</code>	logical. When <code>genes</code> defines chromosomal regions, <code>circ_chrom=TRUE</code> uses background regions from the same chromosome and allows randomly chosen blocks to overlap multiple background regions. Only if <code>test='hyper'</code> .
<code>silent</code>	logical. If <code>TRUE</code> all output to the screen except for warnings and errors is suppressed.
<code>domains</code>	optional <code>character()</code> vector containing one or more of the three GO-domains <code>'cellular_component'</code> , <code>'biological_process'</code> and <code>'molecular_function'</code> . If defined, the analysis will be reduced to those domains which saves time.\ or If a custom ontology is specified in <code>godir</code> it might have a different domains.
<code>orgDb</code>	optional <code>character()</code> naming an <code>OrgDb</code> annotation package from Bioconductor. If <code>orgDb</code> is set, <code>organismDb</code> is not used. <code>OrgDb</code> annotations are available for a wider range of species, e.g. <code>'org.Pt.eg.db'</code> for chimp and <code>'org.Gg.eg.db'</code> for chicken. Note that options <code>gene_len</code> and <code>regions</code> also need a <code>txDb</code> for the gene coordinates (which are integrated in <code>OrganismDb</code> ).
<code>txDb</code>	optional <code>character()</code> naming an <code>TxDb</code> annotation package from Bioconductor (e.g. <code>'TxDb.Ptroglyotes.UCSC.panTro4.refGene'</code> ) for the gene coordinates. Only needed when <code>orgDb</code> is specified, and options <code>gene_len</code> or <code>regions</code> are set. Note that <code>orgDb</code> is needed whenever <code>txDb</code> is defined, even when custom annotations are provided, because the <code>orgDb</code> package is used for Entrez-ID to gene-symbol conversions.
<code>annotations</code>	optional <code>data.frame()</code> for custom annotations, with two <code>character()</code> columns: (1) gene-symbols and (2) GO-categories. Note that options <code>gene_len</code> and <code>regions</code> also need an <code>organismDb</code> or <code>txDb + orgDb</code> .
<code>gene_coords</code>	optional <code>data.frame()</code> for custom gene coordinates, with four columns: gene-symbols ( <code>character</code> ), chromosome ( <code>character</code> ), start ( <code>integer</code> ), end ( <code>integer</code> ). When <code>gene_len=TRUE</code> or <code>regions=TRUE</code> these custom gene coordinates are used instead of the ones obtained from <code>organismDb</code> or <code>txDb</code> .
<code>godir</code>	optional <code>character()</code> specifying a directory <code>()</code> that contains a custom GO-graph (files <code>'term.txt'</code> , <code>'term2term.txt'</code> and <code>'graph_path.txt'</code> ). Alternative to the default integrated GO-graph. For details please refer to the package's vignette.

## Details

Please also refer to the package's vignette.

GO-annotations are taken from a Bioconductor annotation package (`OrganismDb` package `'Homo.sapiens'` by default), but also other `'OrganismDb'` or `'OrgDb'` packages can be used. It is also possible to provide custom annotations as a `data.frame()`.

The ontology graph is integrated, but a custom version can be defined as well with parameter `'godir'`. As long as the ontology tables are in the right format (see link to description in vignette), any ontology can be used in `GOfuncR`, it is not restricted to the gene ontology.

The statistical analysis is based on the ontology enrichment software `FUNC` [2]. `go_enrich` offers four different statistical tests: (1) the hypergeometric test for a candidate and a background gene set; (2) the Wilcoxon rank-sum test for genes that are ranked by scores (e.g. p-value for differential expression); (3) the binomial test for genes that have two associated counts (e.g. amino-acid changes

on the human and the chimp lineage); and (4) a 2x2-contingency table test for genes that have four associated counts (e.g. for a McDonald-Kreitman test).

To account for multiple testing family-wise error rates are computed using randomsets. Besides naming candidate genes explicitly, for the hypergeometric test it is also possible to provide entire genomic regions as input. The enrichment analysis is then performed for all genes located in or overlapping these regions and the multiple testing correction accounts for the spatial clustering of genes.

## Value

A list with components

results	a data.frame() with the FWERs from the enrichment analyses per ontology category, ordered by 'FWER_overrep', 'raw_p_overrep', '-FWER_underrep', '-raw_p_underrep', 'ontology' and 'node_id', or the corresponding columns if another test than the hypergeometric test is used. This table contains the combined results for all three ontology domains. Note that GO-categories without any annotations of candidate or background genes are not listed.
genes	the input genes data.frame(), excluding those genes for which no GO-annotations are available and which therefore were not included in the enrichment analysis. If gene_len=TRUE, also genes without gene coordinates are excluded.
databases	a data.frame() with the used annotation packages and GO-graph.
min_p	a data.frame() with the minimum p-values from the randomsets that are used to compute the FWER.

## Author(s)

Steffi Grote

## References

- [1] Ashburner, M. et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29. doi:[10.1038/75556](https://doi.org/10.1038/75556)
- [2] Pruefer, K. et al. (2007). FUNC: A package for detecting significant associations between gene sets and ontological. BMC Bioinformatics 8: 41. doi:[10.1186/14712105841](https://doi.org/10.1186/14712105841)

## See Also

[get\\_parent\\_nodes](#)  
[get\\_child\\_nodes](#)  
[get\\_anno\\_categories](#)  
[get\\_anno\\_genes](#)  
[plot\\_anno\\_scores](#)  
[get\\_names](#)  
[get\\_ids](#)

## Examples

```
#### Note that argument 'n_randsets' is reduced
#### to lower computational time in the following examples.
#### Using the default value is recommended.

#### Perform a GO-enrichment analysis for some human genes
#### with a defined background set
# create input dataframe that defines the candidate and background genes
candi_gene_ids = c('NCAPG', 'APOL4', 'NGFR', 'NXPH4', 'C21orf59', 'CACNG2',
  'AGTR1', 'ANO1', 'BTBD3', 'MTUS1', 'CALB1', 'GYG1', 'PAX2')
bg_gene_ids = c('FGR', 'NPHP1', 'DRD2', 'ABCC10', 'PTBP2', 'JPH4', 'SMARCC2',
  'FN1', 'NODAL', 'CYP1A2', 'ACSS1', 'CDHR1', 'SLC25A36', 'LEPR', 'PRPS2',
  'TNFAIP3', 'NKX3-1', 'LPAR2', 'PGAM2')
is_candidate = c(rep(1,length(candi_gene_ids)), rep(0,length(bg_gene_ids)))
genes = data.frame(gene_ids=c(candi_gene_ids, bg_gene_ids), is_candidate)
genes

# run enrichment analysis
go_res = go_enrich(genes, n_randset=100)

# go_enrich returns a list with 4 elements:
# 1) results from the analysis
# (ordered by FWER for overrepresentation of candidate genes)
head(go_res[[1]])
# see the top GOs from every GO-domain
by(go_res[[1]], go_res[[1]][,'ontology'], head)
# 2) all valid input genes
go_res[[2]]
# 3) annotation databases used
go_res[[3]]
# 4) minimum p-values from randomsets
head(go_res[[4]])

#### see the package's vignette for more examples
```

---

plot\_anno\_scores

*Plot distribution of scores of genes annotated to GO-categories*

---

## Description

Uses the result of a GO-enrichment analysis performed with [go\\_enrich](#) and a vector of GO-IDs and plots for each of these GO-IDs the scores of the annotated genes. This refers to the scores that were provided as user-input in the [go\\_enrich](#) analysis.

plot\_anno\_scores works with all four tests implemented in [go\\_enrich](#) (hypergeometric, Wilcoxon rank-sum, binomial and 2x2 contingency table test), with test-specific output (see details).

## Usage

```
plot_anno_scores(res, go_ids, annotations = NULL)
```

**Arguments**

<code>res</code>	an object returned from <code>go_enrich</code> (list() of 4 elements).
<code>go_ids</code>	character() vector of GO-IDs, e.g. <code>c('GO:0005737','GO:0071495')</code> . This specifies the GO-categories that are plotted.
<code>annotations</code>	optional data.frame() for custom annotations, with two character() columns: (1) gene-symbols and (2) GO-categories. This is needed if <code>go_enrich</code> was run with custom annotations to generate <code>res</code> , too.

**Details**

The plot depends on the statistical test that was specified in the `go_enrich` call.

For the hypergeometric test pie charts show the amounts of candidate and background genes that are annotated to the GO-categories and the root nodes (candidate genes in the colour of the corresponding root node). The top panel shows the odds-ratio and 95%-CI from fisher's exact test (two-sided) comparing the GO-categories with their root nodes. Note that `go_enrich` reports the the hypergeometric tests for over- and under-representation of candidate genes which correspond to the one-sided fisher's exact tests.

For the Wilcoxon rank-sum test violin plots show the distribution of the scores of genes that are annotated to each GO-category and the root nodes. Horizontal lines in the left panel indicate the median of the scores that are annotated to the root nodes. The Wilcoxon rank-sum test reported in the `go_enrich` result compares the scores annotated to a GO-category with the scores annotated to the corresponding root node.

For the binomial test pie charts show the amounts of A and B counts associated with each GO-category and root node, (A in the colour of the corresponding root node). The top-panel shows point estimates and the 95%-CI of  $p(A)$  in the nodes, as well as horizontal lines that correspond to  $p(A)$  in the root nodes. The p-value in the returned object is based on the null hypothesis that  $p(A)$  in a node equals  $p(A)$  in the corresponding root node. Note that `go_enrich` reports that value for one-sided binomial tests.

For the 2x2 contingency table test pie charts show the proportions of A and B, as well as C and D counts associated with a GO-category. Root nodes are not shown, because this test is independent of the root category. The top panel shows the odds ratio and 95%-CI from Fisher's exact test (two-sided) comparing A/B and C/D inside one node. Note that in `go_enrich`, if all four values are  $\geq 10$ , a chi-square test is performed instead of fisher's exact test.

**Value**

For the hypergeometric, binomial and 2x2 contingency table test, a data.frame() with the statistics that are used in the plots.

For the Wilcoxon rank-sum test no statistical results are plotted, just the distribution of annotated scores. The returned element in this case is a list() with three data frames: annotations of genes to the GO-categories, annotations of genes to the root nodes and a table which contains for every GO-ID the corresponding root node.

**Author(s)**

Steffi Grote

**See Also**

[go\\_enrich](#)  
[get\\_anno\\_genes](#)  
[get\\_names](#)  
[vioplot](#)

**Examples**

```
#### see the package's vignette for more examples

#### Note that argument 'n_randsets' is reduced
#### to lower computational time in the example.

## Assign two random counts to some genes to create example input
set.seed(123)
high_A_genes = c('G6PD', 'GCK', 'GYS1', 'HK2', 'PYGL', 'SLC2A8',
  'UGP2', 'ZWINT', 'ENGASE')
low_A_genes = c('CACNG2', 'AGTR1', 'ANO1', 'BTBD3', 'MTUS1', 'CALB1',
  'GYG1', 'PAX2')
A_counts = c(sample(15:25, length(high_A_genes)),
  sample(5:15, length(low_A_genes)))
B_counts = c(sample(5:15, length(high_A_genes)),
  sample(15:25, length(low_A_genes)))
genes = data.frame(gene=c(high_A_genes, low_A_genes), A_counts, B_counts)

## perform enrichment analysis to find GO-categories with high fraction of A
go_binom = go_enrich(genes, test='binomial', n_randsets=20)

## plot sums of A and B counts associated with the top GO-categories
top_gos = head(go_binom[[1]]$node_id)
stats = plot_anno_scores(go_binom, go_ids=top_gos)

## look at the results of binomial test used for plotting
## (this is two-sided, go_enrich reports one-sided tests)
head(stats)
```

---

refine

---

*Refine results given a FWER threshold*


---

**Description**

Given a FWER threshold, this function refines the results from `go_enrich()` like described in the elim algorithm of [1].

This algorithm removes genes from significant child-categories and then checks whether a category is still significant.

This way significant results are restricted to more specific categories.



**Usage**

```
refine(res, fwer = 0.05, fwer_col = 7, annotations = NULL)
```

**Arguments**

res	list() returned from go_enrich()
fwer	numeric() FWER-threshold. Categories with a FWER < fwer will be labeled significant.
fwer_col	numeric() or character() specifying the column of go_enrich()[[1]] that is to be filtered. E.g. 6 for under-representation or 7 over-representation of candidate genes in the hypergeometric test.
annotations	optional data.frame() with custom annotations. Only needed if go_enrich() was run with custom annotations in the first place.

**Details**

For each domain a p-value is found by interpolation, that corresponds to the input FWER threshold. Since GO-domains are independent graphs, the same FWER will correspond to different p-values, e.g. in 'molecular\_function' and 'biological\_process'.

**Value**

a data.frame() with p-values after refinement for categories that were significant in go\_enrich()[[1]] given the FWER-threshold.

**Author(s)**

Steffi Grote

**References**

[1] Alexa, A. et al. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22, 1600-1607.

**See Also**

[go\\_enrich](#)

**Examples**

```
## perform enrichment analysis for some genes
gene_ids = c('NCAPG', 'APOL4', 'NGFR', 'NXPH4', 'C21orf59', 'CACNG2', 'AGTR1',
             'ANO1', 'BTBD3', 'MTUS1', 'CALB1', 'GYG1', 'PAX2')
input_hyper = data.frame(gene_ids, is_candidate=1)
res_hyper = go_enrich(input_hyper, n_randset=100, silent=TRUE)
head(res_hyper[[1]])
## perform refinement for categories with FWER < 0.1
refined = refine(res_hyper, fwer=0.1)
```

refined

# Index

## \* **htest**

go\_enrich, 11

get\_anno\_categories, 2, 5, 13

get\_anno\_genes, 3, 4, 13, 16

get\_child\_nodes, 5, 6, 8–10, 13

get\_ids, 5, 7, 9, 13

get\_names, 3, 5, 7, 8, 8, 10, 13, 16

get\_parent\_nodes, 3, 5, 7–9, 9, 13

go\_enrich, 3, 11, 14–17

plot\_anno\_scores, 13, 14

refine, 16

vioplot, 16