# 基因富集分析在线软件(g:Profiler)

**Citation**

Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, Jaak Vilo: **g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)** Nucleic Acids Research 2019

文章链接，doi:10.1093/nar/gkz369 [PDF].

## 1．基本介绍

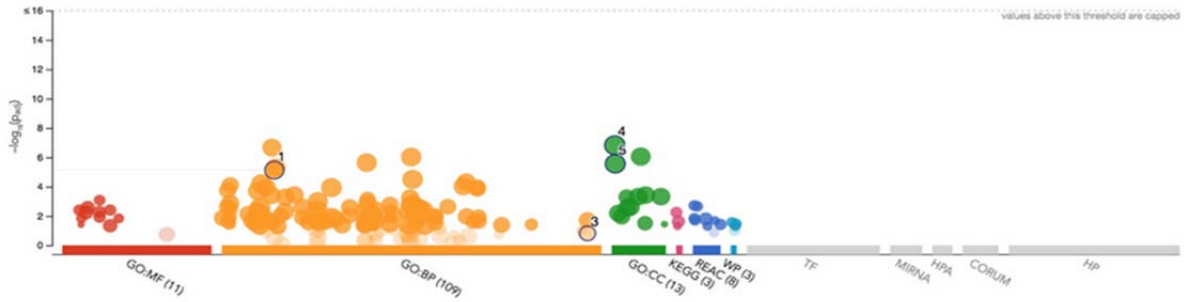
g:Profiler is a public web server for characterising and manipulating gene lists. g:Profiler has a simple user-friendly web interface with powerful visualisations and is currently available for 400+ species, including mammals, plants, fungi, insects from Ensembl and Ensembl Genomes. g:Profiler is updated approximately in every three months and follows quarterly releases of Ensembl databases. g:Profiler tool set consists of the following tools:

- **g:GOSt**, the core of the g:Profiler, **performs statistical enrichment analysis to provide interpretation to user-provided gene lists.** The gene lists can be either flat or ordered gene lists. We accept majority of the identifier types, chromosomal regions and term IDs as input. We provide data from multiple sources of functional evidence, including Gene Ontology terms, biological pathways, regulatory motifs of transcription factors and microRNAs, human disease annotations and protein-protein interactions.
- **g:Convert** is a gene identifier conversion tool. It uses information in Ensembl databases to handle hundreds of types of IDs for genes, proteins, transcripts, microarray probesets, etc, for many species, experimental platforms and biological databases. g:Convert is flexible: it accepts a mixed list of IDs and recognises their types automatically. It can also serve as a service to get all genes belonging to a particular functional category.
- **g:Orth** is a tool for mapping homologous genes across related organisms based on Ensembl data. Given a selected target organism, g:Orth retrieves the genes of the target organism that are similar in sequence to the initial genes in the input.
- **g:SNPense** is a tool for mapping human single nucleotide polymorphisms (SNP) to gene names, chromosomal locations and variant consequence terms from Sequence Ontology.

## 2. Using g:GOSt

Visual reference for g:GOSt [Results] -- Manhattan-like-plot:

### （1）图形表达形式的基因富集结果

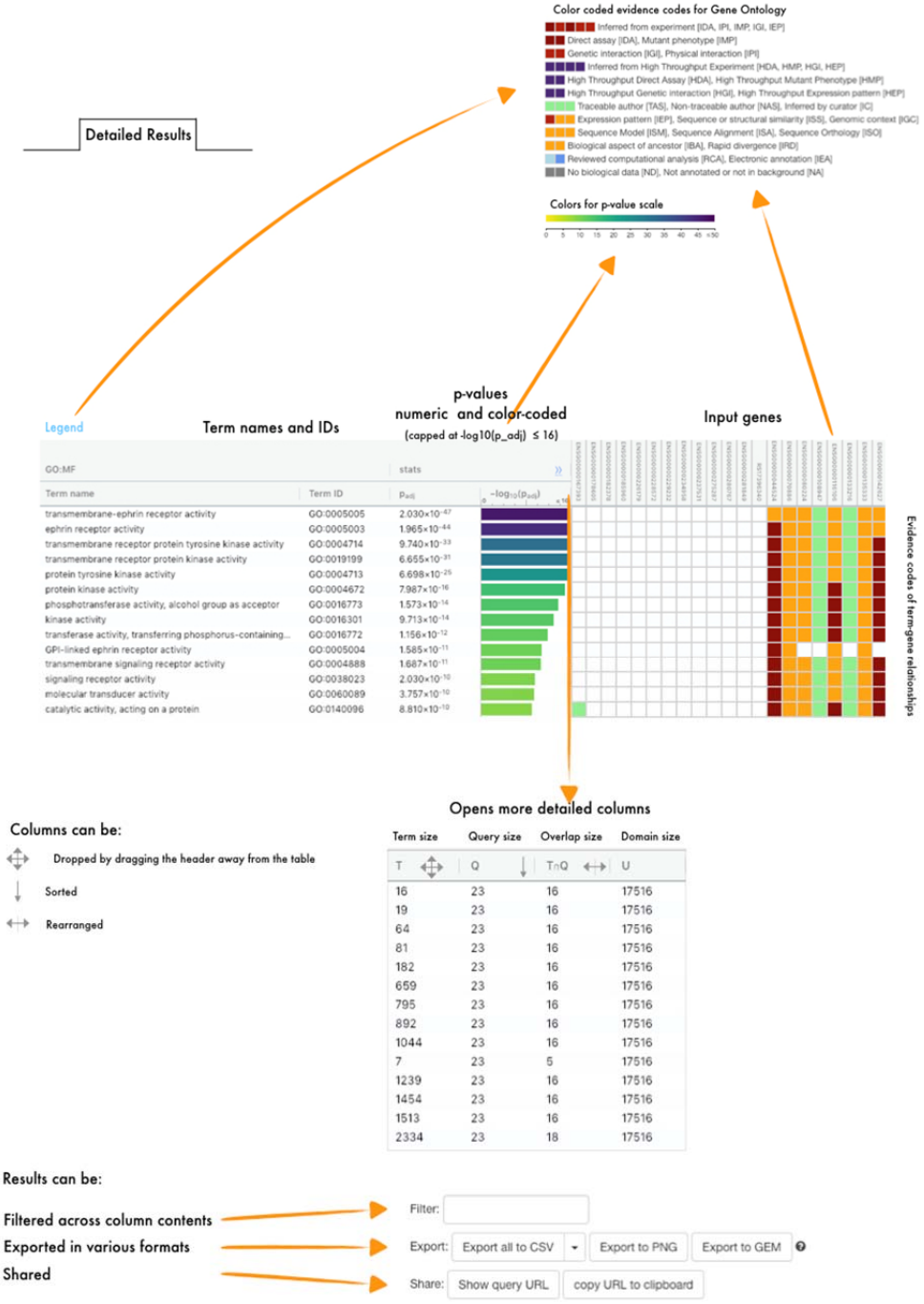



The first visible output in g:GOSt is an interactive Manhattan plot that illustrates the enrichment analysis results (see the figure above). This example figure represents the result of a multiquery with two input gene lists. In case of a single query, only one of such Manhattan plot is generated. All the features remain the same. The x-axis represents functional terms that are grouped and colour-coded by data sources (e.g. Molecular Function from GO is red; the sources that were not included in the analysis are shown in grey). The y-axis shows the adjusted enrichment p-values in negative log10 scale. The light circles represent insignificant terms (if available). For example, the circle number 1 in the figure illustrates the enrichment of the term GO:0009888 (tissue development; includes 1980 annotated genes) and shows that the p-value in the corresponding query is 2.584x10^-10. This information is shown if hovering over the circle. The circle sizes are in accordance with the corresponding term size, i.e. larger terms have larger circles. The term location on the x-axis is fixed and terms from the same GO subtree are located closer to each other (circles number 4 and 5). This is helpful for faster and more intuitive interpretation of the results. The number behind the source name in the x-axis labels shows how many significantly enriched terms there were from this source. Clicking on a circle will pin the circle and create a result table below the figure. The circle will be highlighted with an identifier which is then also referred to in the table showing the detailed information such as data source, id and name of the term together with corresponding p-value. The selection can be removed by clicking again on the circle. There are two views for Manhattan plot, capped and uncapped, which can be selected from the dedicated selection box in the upper left corner. By default, we show the capped version which collects the term circles with p-values less than 10^-16. This fixes the scale of y-axis to keep Manhattan plots from different queries comparable and is also intuitive as, statistically, p-values smaller than that can all be summarised as highly significant. The same threshold is also used in the statistical tests in R. This selection can also be switched off to show the p-values in a wider scale range.In case of multiquery, a separate Manhattan plot for every input query is shown with corresponding query names on top of the image (>polysomal and >total in the figure). These images are interlinked. This allows to highlight the same term across multiple queries by hovering over the term circle in any of the individual plot. Clicking will pin the corresponding circle in every plot. This allows to easily compare the results from multiple queries, both in the image and in the results table.

（2）表格表达形式的结果

| id | source | term id | term name | p_adj (polysomal) | p_adj (total) |
|---|---|---|---|---|---|
| 1 | GO:BP | GO:0009888 | tissue development | $2.584\times10^{-10}$ | $7.503\times10^{-8}$ |
| 2 | GO:BP | GO:1904018 | positive regulation of vasculature development | $3.609\times10^{-7}$ | 1.000 |
| 3 | GO:BP | GO:2000146 | negative regulation of cell motility | $1.443\times10^{-6}$ | $1.523\times10^{-1}$ |
| 4 | GO:CC | GO:0005576 | extracellular region | $6.793\times10^{-4}$ | $1.511\times10^{-7}$ |
| 5 | GO:CC | GO:0005615 | extracellular space | $3.045\times10^{-3}$ | $2.755\times10^{-6}$ |

Visual reference for [Detailed results] output:



（3）Query 参数说明

## Query

**The default input** of g:GOSt is a list of genes/proteins. g:GOSt accepts a simple whitespace-separated gene lists that can consist of mixed types of **gene IDs** (proteins, transcripts, microarray IDs, etc), **SNP IDs**, **chromosomal intervals** or **term IDs**.

In case of chormosomal regions ENSG genes from these regions are retrieved automatically. Genes need not fit the region fully, and hence one may even study single nucleotides (SNPs). g:Profiler uses a chromosome:start:end format for chromosomal regions (e.g. X:1:2000000).

In case of term ID, g:GOSt retrieves all genes of a given organism associated to the given term, and analyses this set of genes as a query. For example, when queried for GO:0007507 (heart development) with organism H. sapiens, g:GOSt retrieves about a hundred human genes associated to heart development, and performs an analysis for this gene set. One may then observe statistically significant related pathways from KEGG and REACTOME, or putative regulatory elements from TRANSFAC.

g:Profiler tools accept mixed types of gene IDs as input. These are automatically converted via g:Convert to an internal format based on Ensembl genes (ENSG). Note that chromosomal regions and termIDs can also be mixed in with regular gene list query.

All g:Profiler tools assume space-separated queries, i.e. spaces, tabs, and newlines are used to distinguish gene IDs. All queries are case insensitive.

Fully numeric IDs need to prefixed prior to gene list submission. All encountered numeric IDs will be prefixed by g:Profiler automatically, using the prefix determined by the Numeric IDs treated as dropdown menu. See advanced options to change the default type of prefixes.

A random query may also be submitted (button just below query box). This is constructed by randomly picking 50% of the query symbols over all GO terms and 2*25% from two mid-sized GO terms, hence usually generating a statistically significant result.

**Multi-queries** consist of several gene lists that can be submitted to g:GOSt for comparative enrichment analysis. Each individual gene list needs a FASTA style header line (>QueryTitle) that can include **a query title.**

**Organism**

Organism is one of the most important input parameters to g:Profiler tools. This parameter defines the organism where to the input genes, proteins and probes belong to.

g:Profiler supports organism specific queries. Any identifiers not recognised within a given organism are marked as unknown ('?' or 'N/A').

Default organism in g:Profiler is human (Homo sapiens). The organism drop-down list has the 10 most frequently queried organisms at the top and all the other supported organisms grouped by sources (ensembl, fungi, metazoa, parasites and plants). The organism list has flexible search functionality for easier species

selection. Search functionality matches across latin and common names (e.g. both human and Homo sapiens can be searched for).

A rule of thumb: If your query gives you no results, you have most probably chosen an incorrect organism.

A full list of available organisms supported by g:Profiler can be found here.

**Ordered query**

g:Profiler gene lists may be interpreted as ordered lists where elements (genes, proteins, probesets) are in the order of decreasing importance. The ordered query option is useful when the genes can be placed in some biologically meaningful order. For instance, according to differential expression in a given RNA-seq experiment; ranked by the number of protein-protein interactions; absolute expression values etc. **g:Profiler then performs incremental enrichment analysis with increasingly larger numbers of genes starting from the top of the list. This procedure identifies specific functional terms that associate to most dramatic changes in given experimental setup, as well as broader terms that characterise the gene set as a whole.**

**Run as multiquery**

Queries are separated by lines that start with a > symbol and optionally contain a title for particular gene list. Any number of following lines until the next > belong to the query and should contain gene names, chromosomal regions, etc. Note that the query header is not needed in case of a single list of genes.

## All results

By default g:GOSt shows out only statistically significant results after multiple correction is applied. However, when a user wants to obtain the full list of gene-term relationships or check particular terms then "All results" option can be applied.

**Measure underrepresentation** -- Check this option if you wish to determine significantly under-represented functional terms. By default over-representation is measured, i.e. the probability that the intersection of query and a functional category has arisen by chance. In contrast, measuring under-representation tests whether a particular term has less genes overlaping with the query genes than expected by chance.

**Statistical domain scope** -- Statistical domain size N describes the total number of genes used for random selection and is one of the four parameters for the hypergeometric probability function of statistical significance used in g:GOSt.

**Only annotated genes** -- The default behaviour of g:GOSt considers only genes with at least one annotation to be part of the domain. Less-studied mammalian genomes involve a large fraction of genes that lack even "unknown" tags and the most generic annotations (biological process, molecular function, cell component). These unverified genes are often spurious and in most cases could be omitted from statistics. This results in a smaller amount of significant results everywhere; very large terms should almost never appear statistically highlighted.

**All known genes** -- g:GOSt also allows to calculate statistical significance considering all genes of the given organism in the Ensembl database. Since the number of annotated genes may be much smaller than the number of all genes, g:GOSt always delivers more results that appear as significant. In larger queries, topmost GO terms such as Biological Process, Molecular Function, Cell Component tend to come out as highly significant using all known genes. This is misleading in queries with few unannotated genes. Statistical domain size of all known genes makes sense in situations where the input includes a lot of unknown and unannotated genes.

**Custom** -- In order to compute functional enrichments of gene lists, g:GOSt uses the background set of all organism-specific genes annotated in the Ensembl database. In several occasions, it is advisable to limit the background set for more accurate statistics. For instance, one may use a custom background when the number of genes and corresponding probesets of a microarray platform is considerably smaller than the number of known genes, or only genes of a specific chromosome are considered. g:GOSt provides means to define the custom background as a mixed list of gene, probeset and protein IDs in the corresponding form field. It is also possible to select a predefined custom background from a list of popular microarray platforms.

**Custom over annotated genes** -- Use the set of all genes that are annotated in the data source and also included in the custom background. This mimics the default setting ("Only annotated genes") while limiting the background genes to the ones provided.

**Custom over all known genes** -- Use the set of all known genes in the custom background as the statistical domain.

## Significance threshold

Multiple testing problem is a statistical concept that relates to the increased chance of getting significant-looking false positive results, when evaluating a large number of alternative hypotheses simultaneously. It is an important issue in functional enrichment analysis, since each input query is compared against thousands of Gene Ontology terms, pathways, regulatory motifs, etc. Multiple testing correction systematically reduces the significance of detected p-values to discard false positives.

g:GOSt uses multiple testing correction by default and applies our tailor-made algorithm g:SCS for reducing significance scores. Alternatively, one may select Bonferroni correction (BC) or Benjamini-Hochberg FDR (False Discovery Rate) -- these two standard solutions to the multiple testing problem are available under Advanced options. In comparison to the latter two, our algorithm takes into account the unevenly distributed structure of functionally annotated gene sets. Our simulations in (Nucleic Acids Research, 2007) show that g:SCS provides a better threshold between significant and non-significant results than FDR or BC.

**g:SCS algorithm** -- g:SCS method is the default method for computing multiple testing correction for p-values gained from GO and pathway enrichment analysis. It corresponds to an experiment-wide threshold of a=0.05, i.e. at least 95% of matches above threshold are statistically significant.

This approach is based on the idea that standard multiple testing corrections such as Bonferroni correction and Benjamini-Hochberg FDR are designed for multiple tests that are independent of each other. This is not correct for the analysis in g:GOSt, since GO consists of hierarchically related general and specific terms. The True Path Rule of GO states that genes associated to a given GO term are implicitly associated to all more general parents of this term.

g:SCS threshold is a value pre-calculated for query list sizes up to 1000 genes. Given a fixed input query size, g:SCS analytically approximates a threshold t corresponding to the 5% upper quantile of randomly generated queries of that size. All actual p-values resulting from the query are transformed to corrected p-values by multiplying these to the ratio of the approximate threshold t and the initial experiment-wide threshold a=0.05.

The algorithm considers the set structure underlying gene sets annotated to terms of each organism, and should therefore give a tighter threshold to significant results. g:SCS thresholds perfectly agreed in simulations with randomly generated gene sets of fixed input query sizes.

**Bonferroni correction** -- Bonferroni correction [1] is a simple and well-known Family Wise Error Rate p-value correction for multiple testing. Family Wise Error Rate measures the probability of at least one random result considered significant (Type I error) within experiment. Given an experiment-wide significance threshold a, Bonferroni correction takes into account only number n of performed independent or dependent tests n in given experiment, and defines individual significance level as a* = a/n . Every match with p-value below a* is considered insignificant.

In g:GOSt analysis, the expected experiment-wide significance level is a=0.05. The variable n denotes the number of independent tests, i.e. the number of GO terms, KEGG pathways, etc, compared against the given input query. Two different approaches for n are mentioned in literature. A more common approach considers only terms that have some genes common with input query. Other approach suggests that n should be considered equal to the number of all annotated terms in the given genome. The first case would involve correction for a few hundred tests, while the second case observes tests with several thousand terms. g:GOSt follows the first approach for calculating the correction.

Bonferroni correction is considered rather conservative in the sense that it increases the amount of discarded significant results. Matches with smaller annotation sets never appear significant, as even a 100% overlap of input query and given term will not result in a sufficiently low p-value.

[1] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8:3 62, 1936.

**False Discovery Rate** -- False Discovery Rate (FDR)[1] is a simple multiple testing correction, that measures expected proportion of false significant matches (Type I errors) within results. Benjamini-Hochberg FDR approach takes into account p-values observed in the experiment. The latter is often considered more applicable for Gene Ontology analysis than Bonferroni correction.

Let a be defined significance level for the whole experiment of n statistical tests. FDR method sorts n p-values from tests in increasing order, and picks a* to be the largest p-value, p_j, that is smaller than its proportional significance level a_j. Proportional significance level a_i for ith result, is calculated as a_i = i*a/n. Every match with p-value above Benjamini-Hochberg corrected significance a* level is discarded as insignificant.

In g:GOSt analysis, the expected experiment-wide significance level is a=0.05. The variable n denotes the number of independent tests, i.e. the number of GO terms, KEGG pathways, etc, compared against the given input query. Two different approaches for n are mentioned in literature. A more common approach considers only terms that have some genes common with input query. Other approach suggests that n should be considered equal to the number of all annotated terms the given genome. The first case would involve correction for a few hundred tests, while the second case observes tests with several thousand terms. g:GOSt follows the first approach for calculating the correction.

FDR method is argued improper for analysis of gene queries, as GO terms are hierarchically related, gene annotation sets are highly intersecting because of the True Path Rule, and therefore, statistical tests with many matching terms should not be considered entirely independent of each other. It is not yet clear whether Gene Ontology hierarchy complies with variants of FDR designed for dependent testing.

[1] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society B, 57(1):289-300, 1995.

**User threshold**

User-defined p-value threshold provides a possibility to additionally filter results. The threshold defaults to p=0.05, meaning that all significant results are shown. If the threshold is set to less than 0.05, matches with p-values above threshold are not shown.

User-defined threshold differs from significance thresholds, since it does not attempt to decide over significance, i.e. change the colour of results in output.

**Numeric IDs treated as**

Fully numeric gene IDs need to get an appropriate prefix prior to gene list submission. All encountered numeric IDs are automatically detected and a prefix is added by g:Profiler based on the Numeric IDs defined in the corresponding drop-down menu. For all namespaces supported by g:Profiler, see the list of namespaces.

## Data sources

**Gene Ontology**

Gene Ontology(GO) is a directed acyclic graph structure that defines biological functions and their relationships to one another. The GO annotations, accompanied by evidence-based statements describe specific gene product and specific ontology term (biological function) relationship. The GO has three major subontologies - **Molecular Functions (MF)** that describe gene product's functions; **Biological Process (BP)** that describe in which biological process the gene product participates; and **Cellular Component (CC)** that describe in which part of the cell the particular gene product is physically located.

**Electronic GO annotations [IEA]**

A large proportion of functional annotations of Gene Ontology are assigned to genes using in silico curation methods and have the IEA evidence code (Inferred from Electronic Annotation). While IEA annotations are an invaluable resource in mapping gene functions, manually curated annotations of experimental and computational studies are generally of higher confidence. Therefore it might be advisable to exclude electronically inferred annotations from enrichment analysis and focus on annotations with stronger evidence when higher confidence is wanted for the enrichment results. Excluding IEA annotations may also help to reduce bias towards abundant and ubiquitous housekeeping genes like the ribosomes. These annotations can be excluded under the Data sources selection.

**Biological Pathway databases**

In g:GOSt we support **Reactome**, **KEGG** and **WikiPathways**. Reactome pathways are centrally curated by selected domain experts. KEGG pathways are a mixture of species specific curation and orthologous knowlege mapping between species. WikiPathways is a community driven platform where pathways are curated by anyone interested.

**Regulatory motifs in DNA**

**Transfac** putative transcription factor binding sites (TFBSs) from TRANSFAC database are retrieved into g:GOSt through a special prediction pipeline. First, TFBSs are found by matching TRANSFAC position specific matrices using the program Match on range +/-1kb from TSS as provided by APPRIS (Annotating principal splice isoforms) via Ensembl biomart. For genes with multiple primary TSS annotations we selected one with most TF matches. The matching range for C. elegans, D. melanogaster and S. cerevisiae is 1kb upstream from ATG (translation start site). A cut-off value to minimize the number of false positive matches (provided by TRANSFAC) is then applied to remove spurious motifs. Remaining matches are split into two inclusive groups based on the amount of matches, i.e TFBSs that have at least 1 match are classified as match class 0 and TFBSs that have at least 2 matches per gene are classified as match class 1.

**mirTarBase** is a database that holds experimentally validated information about genes that are targetted by miRNAs. We include all the organisms that are covered by mirTarBase.

**Protein databases**

**CORUM** is a database of manually annotated protein complexes from mammalian organisms. g:Profiler includes gene annotations for all CORUM protein complexes in human, mouse and rat.

**Human Phenotype Ontology**

We include gene annotations from the Human Phenotype Ontology (HPO), a standardized vocabulary of phenotypic abnormalities encountered in human disease. Due to ethical constraints, a significant portion of research on human disease is conducted in model organisms like mouse and rat. However HPO only provides annotations to human genes at the time. To complement HPO annotations of human genes, we have extended their gene database to genes of other organisms in g:Profiler using gene orthology information from Ensembl. As a result, researchers can directly see enrichments of human disease associations in their gene lists of model organisms.