

GOToolBox

Functional analysis of gene sets based on Gene Ontology

David MARTIN



david.martin@crg.es

EMBRACE Workshop - *Applied Gene Ontology*
Bari, Italy - November 8th, 2007

Software **Open Access** **Highly accessed**

GOToolBox: functional analysis of gene datasets based on Gene Ontology

David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, Bernard Jacq

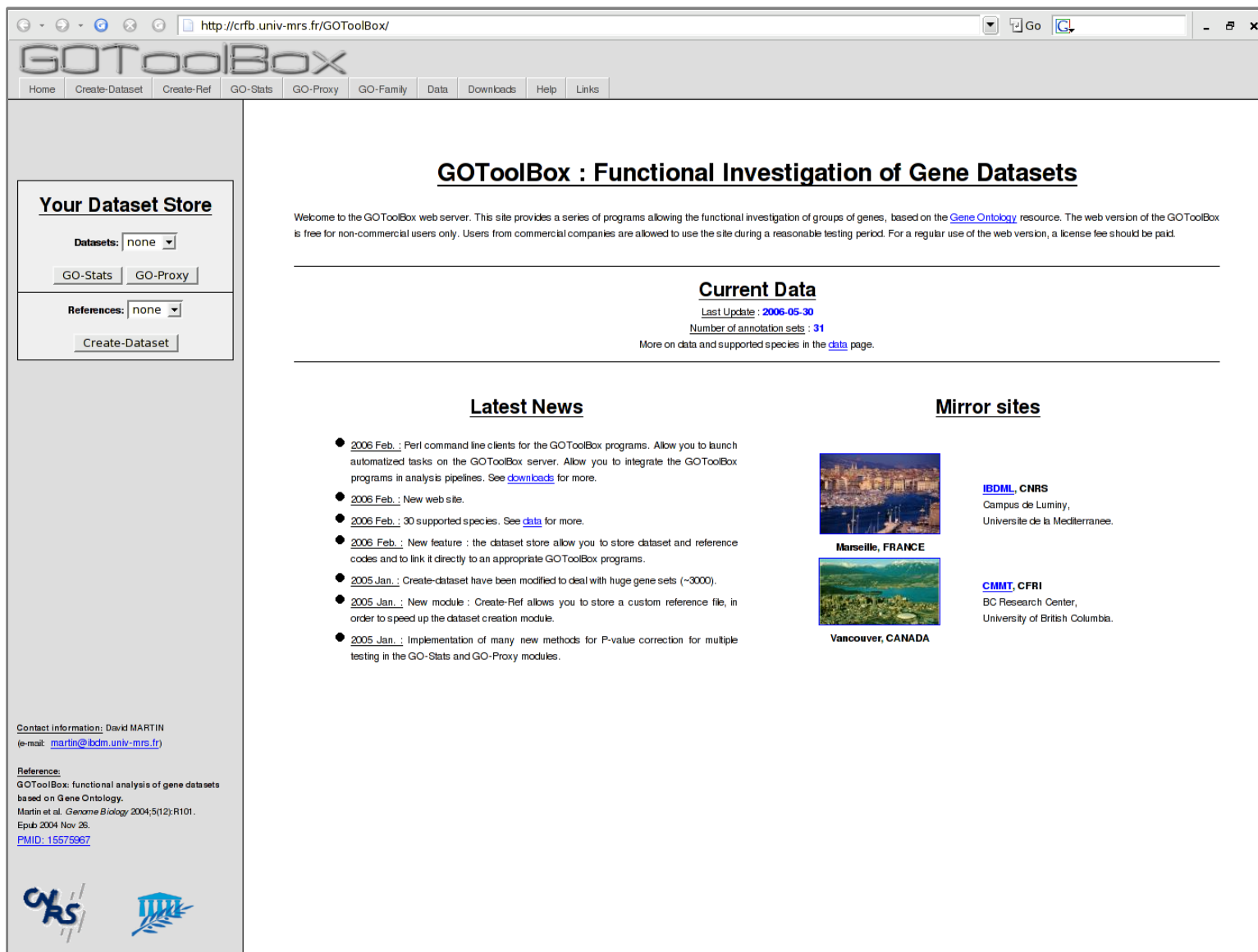
Genome Biology 2004, 5:R101 (26 November 2004)

 <http://crfb.univ-mrs.fr/GOToolBox/>

 <http://burgundy.cmmmt.ubc.ca/GOToolBox/>

 <http://crfb.univ-mrs.fr/GOToolBox/>
 <http://burgundy.cmmmt.ubc.ca/GOToolBox/>

GOToolBox website



The screenshot shows the GOToolBox website in a web browser window. The browser's address bar displays <http://crfb.univ-mrs.fr/GOToolBox/>. The website has a navigation menu with links: Home, Create-Dataset, Create-Ref, GO-Stats, GO-Proxy, GO-Family, Data, Downloads, Help, and Links. The main heading is "GOToolBox : Functional Investigation of Gene Datasets". Below this, a welcome message states that the site provides programs for functional investigation of gene groups based on Gene Ontology, and that the web version is free for non-commercial users. A "Current Data" section shows the last update as 2006-05-30 and 31 annotation sets. A "Latest News" section lists updates from 2005 and 2006. "Mirror sites" are listed for IBDML, CNRS in Marseille, France, and CMMT, CFRI in Vancouver, Canada. A left sidebar contains a "Your Dataset Store" with dropdown menus for Datasets and References, and buttons for GO-Stats, GO-Proxy, and Create-Dataset. At the bottom left, contact information for David MARTIN is provided, along with a reference to a paper in Genome Biology.

GOToolBox

Home Create-Dataset Create-Ref GO-Stats GO-Proxy GO-Family Data Downloads Help Links

Your Dataset Store

Datasets: none

GO-Stats GO-Proxy

References: none

Create-Dataset

GOToolBox : Functional Investigation of Gene Datasets

Welcome to the GOToolBox web server. This site provides a series of programs allowing the functional investigation of groups of genes, based on the [Gene Ontology](#) resource. The web version of the GOToolBox is free for non-commercial users only. Users from commercial companies are allowed to use the site during a reasonable testing period. For a regular use of the web version, a license fee should be paid.


Current Data


Last Update : **2006-05-30**
Number of annotation sets : **31**
More on data and supported species in the [data](#) page.

Latest News

- **2006 Feb.** : Perl command line clients for the GOToolBox programs. Allow you to launch automatized tasks on the GOToolBox server. Allow you to integrate the GOToolBox programs in analysis pipelines. See [downloads](#) for more.
- **2006 Feb.** : New web site.
- **2006 Feb.** : 30 supported species. See [data](#) for more.
- **2006 Feb.** : New feature : the dataset store allow you to store dataset and reference codes and to link it directly to an appropriate GOToolBox programs.
- **2005 Jan.** : Create-dataset have been modified to deal with huge gene sets (~3000).
- **2005 Jan.** : New module : Create-Ref allows you to store a custom reference file, in order to speed up the dataset creation module.
- **2005 Jan.** : Implementation of many new methods for P-value correction for multiple testing in the GO-Stats and GO-Proxy modules.

Mirror sites


Marseille, FRANCE




Vancouver, CANADA

IBDML, CNRS
Campus de Luminy,
Universite de la Mediterranee.

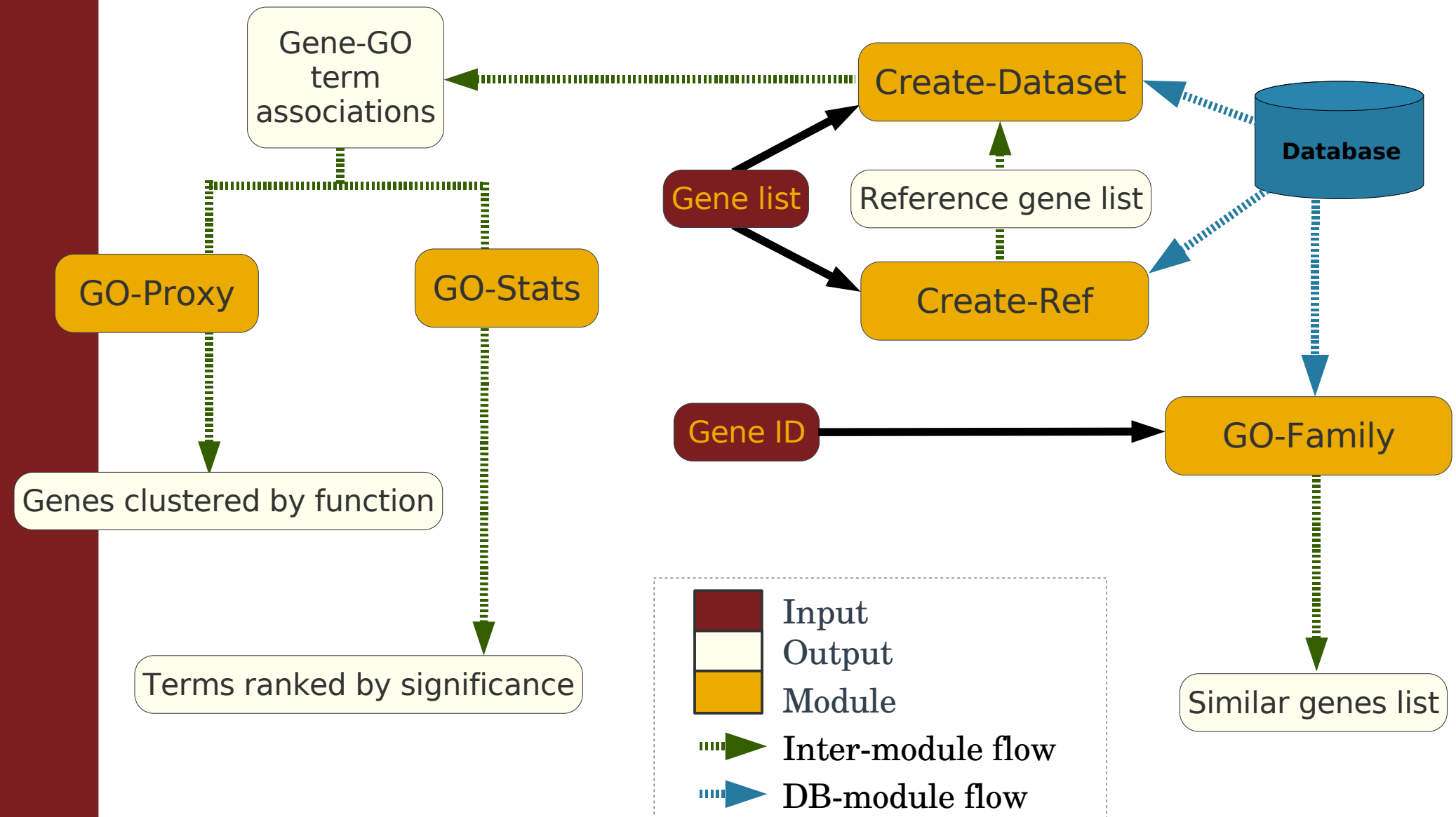
CMMT, CFRI
BC Research Center,
University of British Columbia.

Contact information: David MARTIN
(e-mail: martin@ibdm.univ-mrs.fr)

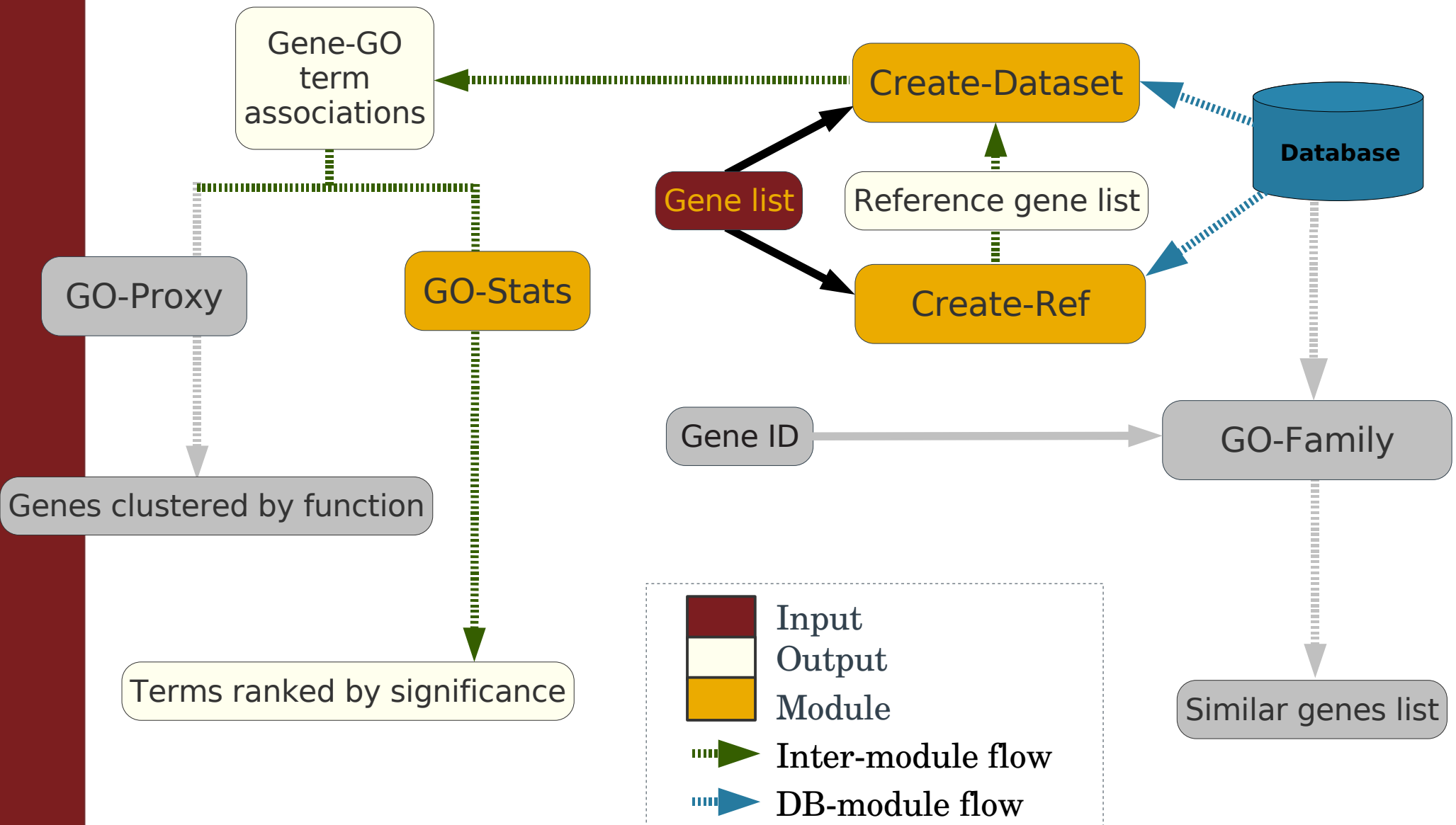
Reference:
GOToolBox: functional analysis of gene datasets
based on Gene Ontology.
Martin et al. *Genome Biology* 2004;5(12):R101.
Epub 2004 Nov 26.
[PMID: 15573967](#)

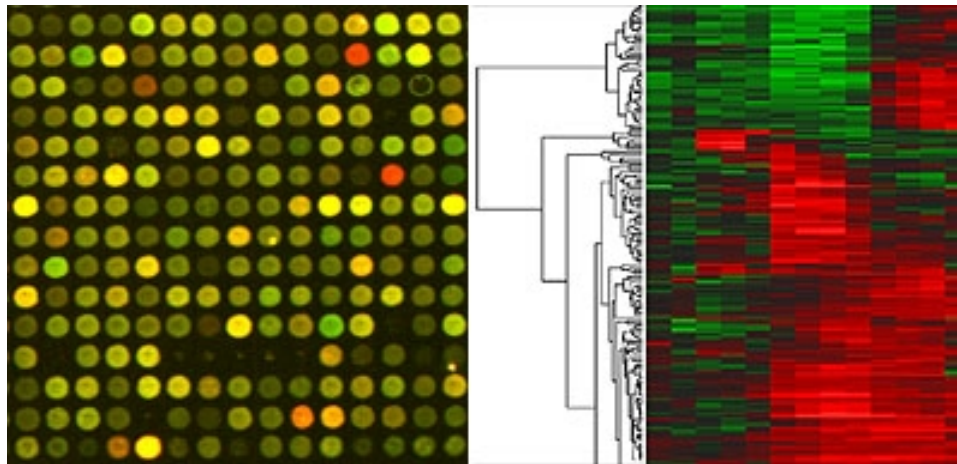
GOToolBox Flowchart



GOToolBox: microarray analysis components



Microarray analysis in GOToolBox: a two step process



Cluster of co-expressed genes [gene list]

Create-Dataset

Associated GO terms

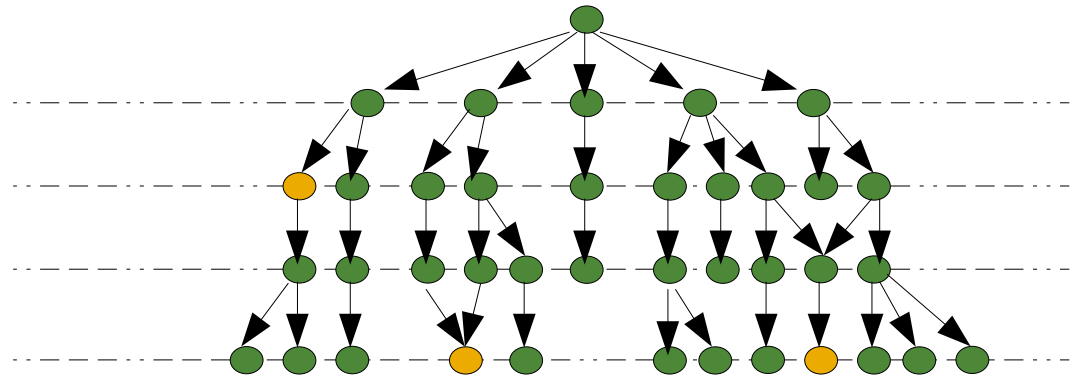
GO-Stats

Relevant GO terms

Step 1: term retrieval

Gene A annotations:

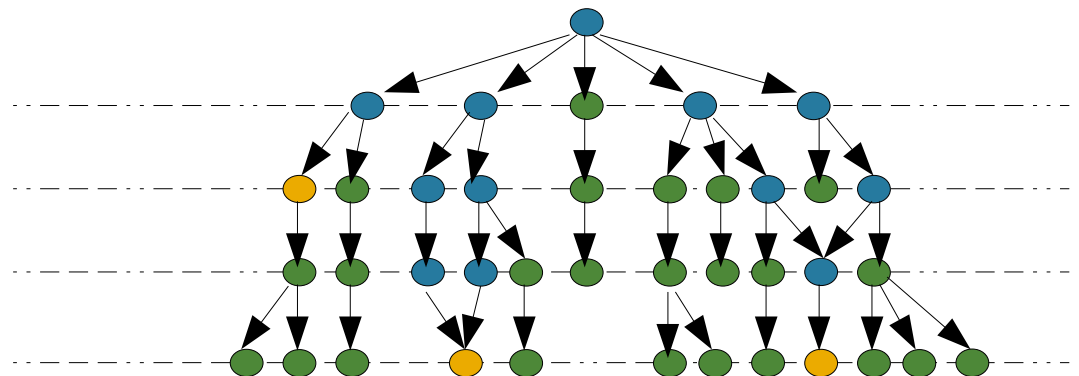
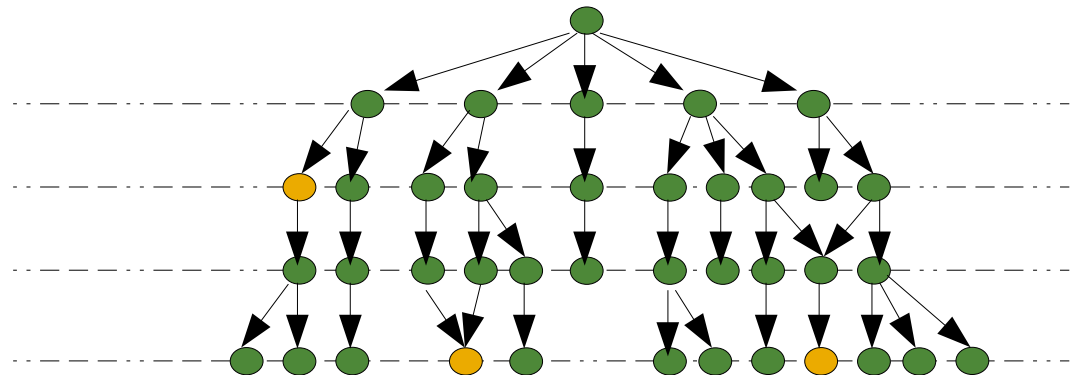
- term1
- term2
- term3



Step 1: term retrieval

Gene A annotations:

- term1
- term2
- term3
- (parent) term 4
- (parent) term 5
- (parent) term 6
- (parent) term 7
- (parent) term 8
- (parent) term 9
- ...

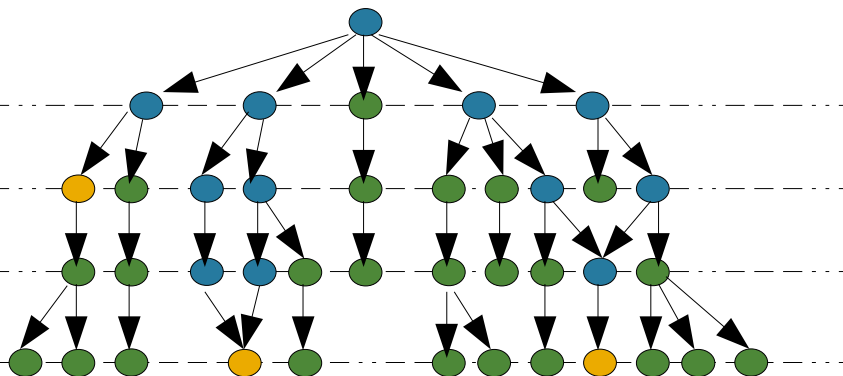


Advanced term retrieval

- Take into account only those terms that are the most reliable on the basis of the **GO evidence codes**: terms inferred on the basis by curators or from direct assays are more reliable than terms inferred from sequence similarity for example.

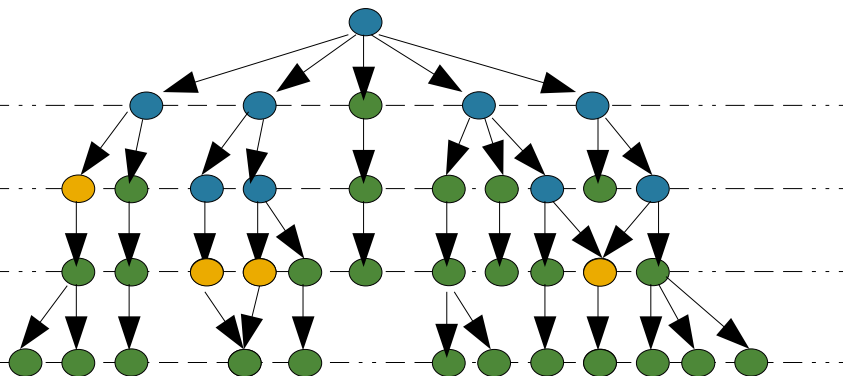
Advanced term retrieval

- Take into account only those terms that are the most reliable on the basis of the **GO evidence codes**: terms inferred on the basis by curators or from direct assays are more reliable than terms inferred from sequence similarity for example.
- Decrease the size of the functional vocabulary:
 - **Depth** (level) limitation



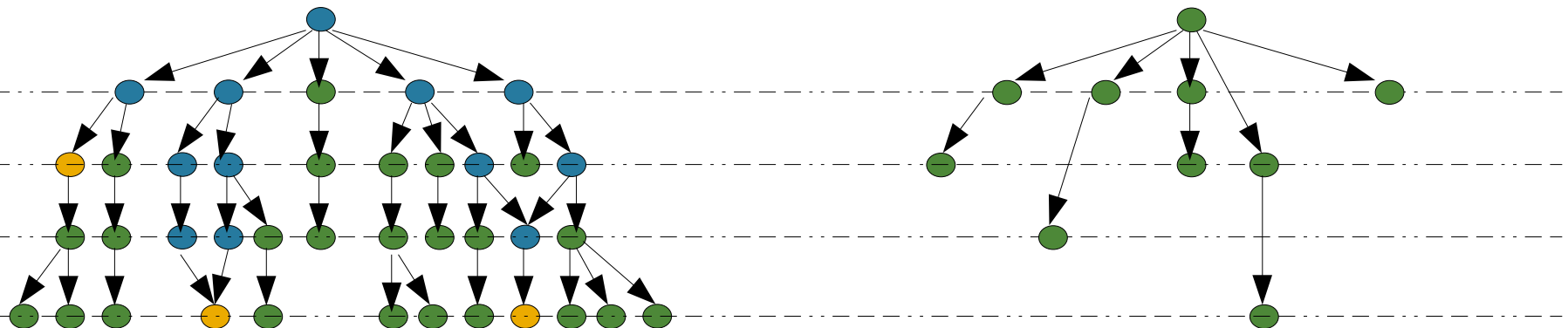
Advanced term retrieval

- Take into account only those terms that are the most reliable on the basis of the **GO evidence codes**: terms inferred on the basis by curators or from direct assays are more reliable than terms inferred from sequence similarity for example.
- Decrease the size of the functional vocabulary:
 - **Depth** (level) limitation



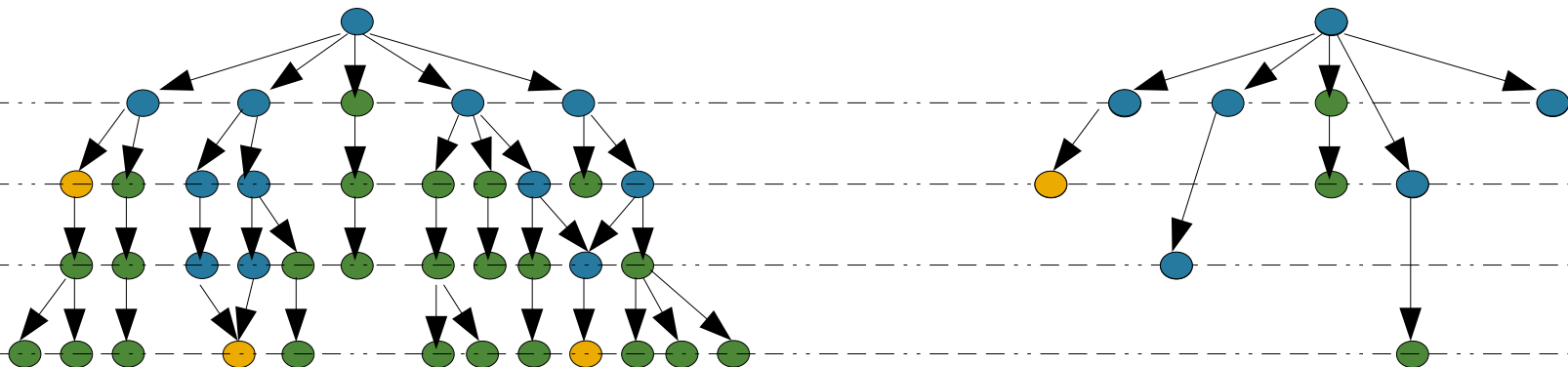
Advanced term retrieval

- Take into account only those terms that are the most reliable on the basis of the **GO evidence codes**: terms inferred on the basis by curators or from direct assays are more reliable than terms inferred from sequence similarity for example.
- Decrease the size of the functional vocabulary:
 - **Depth** (level) limitation
 - **Slim ontology** (cut-down version of the ontologies, gives a broad overview of the ontology content. <http://www.geneontology.org/GO.slims.shtml>)



Advanced term retrieval

- Take into account only those terms that are the most reliable on the basis of the **GO evidence codes**: terms inferred on the basis by curators or from direct assays are more reliable than terms inferred from sequence similarity for example.
- Decrease the size of the functional vocabulary:
 - **Depth** (level) limitation
 - **Slim ontology** (cut-down version of the ontologies, gives a broad overview of the ontology content. <http://www.geneontology.org/GO.slims.shtml>)



Term retrieval output

	Count in the input set	Number of genes in the set
Term 1	5	50
Term 2	8	50
Term 3	1	50
Term 4	25	50
Term 5	42	50
Term 6	2	50
Term 7	12	50
Term 8	7	50
...

GO-Stats: Identify the relevant terms

Given a list of genes and associated GO terms, which terms are significantly over-represented? ... but by the way, as compared to what?

The input list of genes comes from a bigger set of genes, called the **reference** gene set.

- In microarray experiments, the reference set represents those genes having probes spotted on the array.
- For a given organism, Gene Ontology provides annotations for the whole set of gene products encoded by the genome.
- Then, the user might have to create his own reference corresponding to the list of genes spotted on the array he is using (which could differ significantly from the whole genome set) !!!



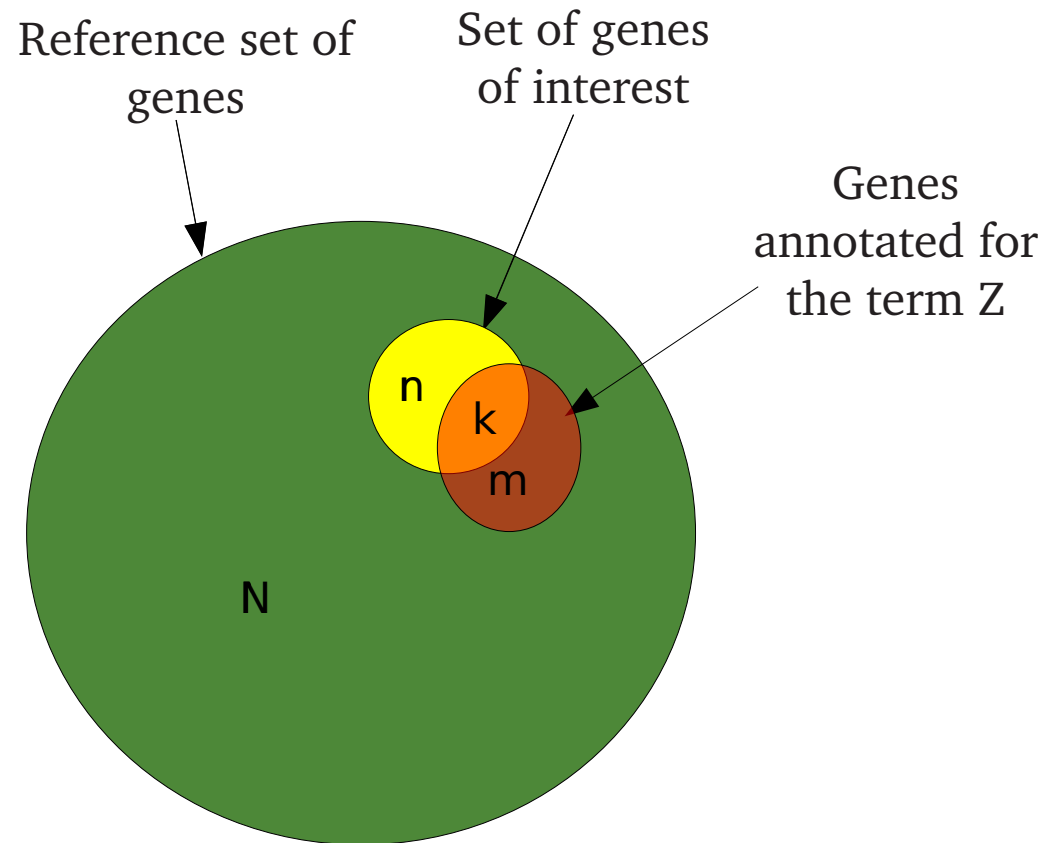
Term retrieval output: adding the reference

	Count in the input set	Number of genes in the input set	Count in the reference set	Number of genes in the reference set
Term 1	5	50	6	5000
Term 2	8	50	560	5000
Term 3	1	50	48	5000
Term 4	25	50	465	5000
Term 5	42	50	1500	5000
Term 6	2	50	49	5000
Term 7	12	50	15	5000
Term 8	7	50	73	5000
...

Statistical tests available in GOToolBox

Three tests are implemented in GO-Stats to measure enrichment-depletion of the terms as compared to a reference gene set:

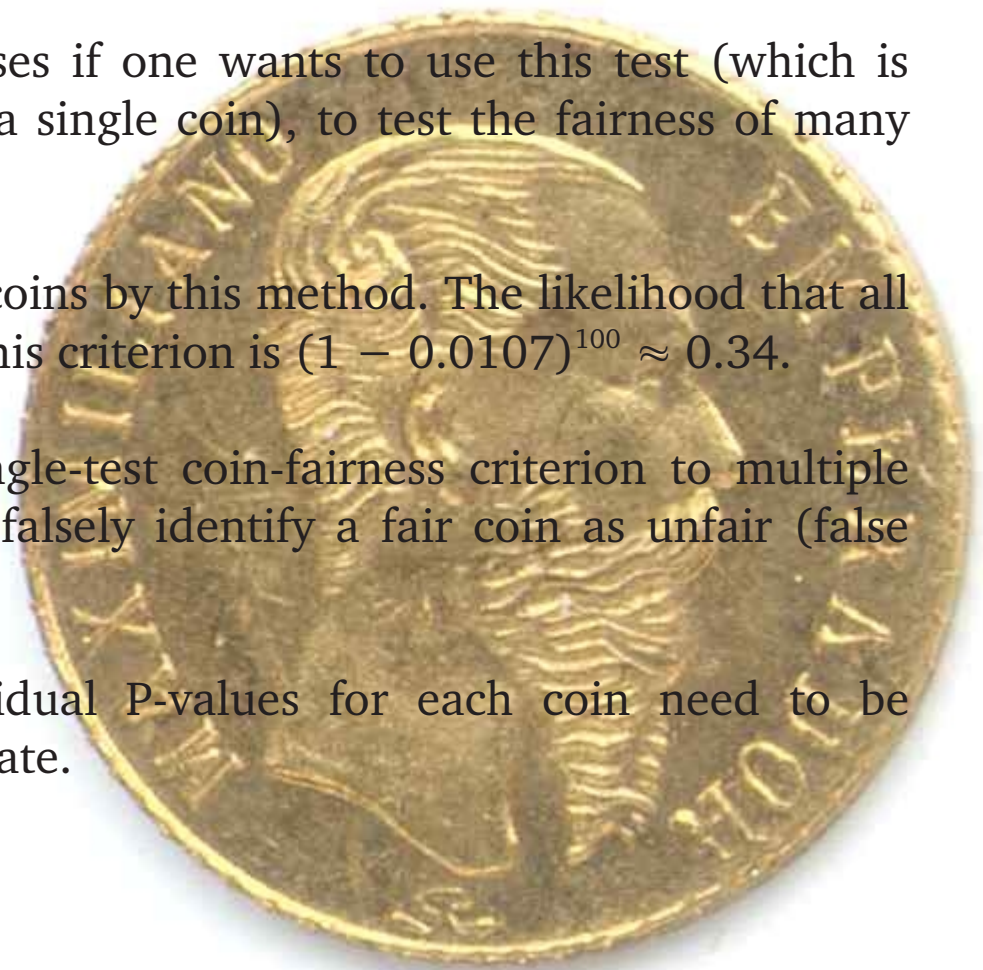
- [Hypergeometric-based test](#)
- [Binomial-based test](#)
- [Fisher's exact test](#)



Let consider a pool of N genes (reference), out of which m are annotated for the term Z . When picking n genes out of N , is it exceptional to have k genes annotated for the term Z ?

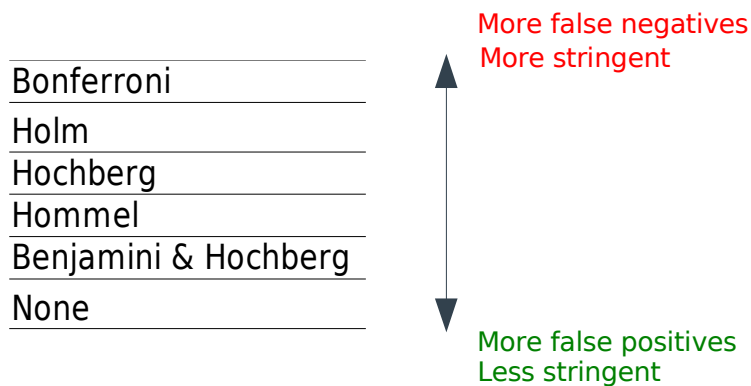
Correction for multiple testing: introduction

- For example, one might declare that a coin is biased if in 10 flips it landed heads at least 9 times ($P=0.0107$).
- A multiple comparisons problem arises if one wants to use this test (which is appropriate for testing the fairness of a single coin), to test the fairness of many coins.
- Imagine if one wants to test 100 fair coins by this method. The likelihood that all 100 fair coins are identified as fair by this criterion is $(1 - 0.0107)^{100} \approx 0.34$.
- Therefore the application of our single-test coin-fairness criterion to multiple comparisons would be more likely to falsely identify a fair coin as unfair (false positives).
- In order to compensate, the individual P-values for each coin need to be 'corrected' to restrict the false positive rate.



Correction for multiple testing in GOToolBox

Multiple testing corrections adjust p-values derived from multiple statistical tests to correct for occurrence of false positives. In GO annotations analysis, false positives are terms that are found to be statistically enriched in your gene set as compared to the reference set, but are not in reality.



GO-Stats output

GOToolBox - functional analysis of gene datasets - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://burgundy.cmmmt.ubc.ca/GOToolBox/

Entrez PubMed CRG CRG webmail

GOToolBox

Home Create-Dataset Create-Ref GO-Stats GO-Proxy GO-Family Data Downloads Help Links

Your Dataset Store

Datasets: none ▼

GO-Stats GO-Proxy

References: none ▼

Create-Dataset

GO-Stats Results

Your dataset contains **54** annotated gene products.
You can download a tabulated text output file [here](#), for use with a spreadsheet application.


Color codes	
E	The term is enriched in your gene set.
P-value < 0.01	
RO	Number of genes annotated for this term in the reference set.
DO	Number of genes annotated for this term in your gene set.
D	The term is depleted in your gene set.
0.01 < P-value < 0.05	
RF	Frequency of genes annotated for this term in the reference set.
DF	Frequency of genes annotated for this term in your gene set.

The terms are ranked according to the P-value column representing their statistical relevance. The 'level' column describes the depth at which a given GO term is found in the GO hierarchy (note that some terms can be found at several levels simultaneously). The last column indicates whether a given GO term is enriched (E) or depleted (D), based on the term frequency ratio (DF/RF). To visualize the hierarchy between these terms, you can click on the radio button close to a GO ID, and all its parent terms in the list are highlighted in orange. When moving the mouse pointer on the GO ID column will make all the genes associated with a given GO term appear in a box. The terms are hyperlinked to the QuickGO EBI browser.

GO ID	Level	GO Term	RO	RF	DO	DF	P-value	ED
<input type="radio"/> GO:0007154	3	cell communication	1620	0.1599	21	0.3889	0.0142360	E
<input type="radio"/> GO:0007424	4	tracheal system development (sensu Insecta)	119	0.0117	6	0.1111	0.0164582	E
<input type="radio"/> GO:0007165	4	signal transduction	1355	0.1338	18	0.3333	0.0468596	E
<input type="radio"/> GO:0050801	4	ion homeostasis	28	0.0028	3	0.0556	0.1952095	E
<input type="radio"/> GO:0035239	4	tube morphogenesis	75	0.0074	4	0.0741	0.2908931	E
<input type="radio"/> GO:0007166	5	cell surface receptor linked signal transduction	706	0.0697	11	0.2037	0.3690890	E
<input type="radio"/> GO:0008360	6,5	regulation of cell shape	87	0.0086	4	0.0741	0.5017696	E
<input type="radio"/> GO:0035295	3	tube development	88	0.0087	4	0.0741	0.5230461	E
<input type="radio"/> GO:0000902	5,4	cellular morphogenesis	417	0.0412	8	0.1481	0.5670668	E
<input type="radio"/> GO:0048731	3	system development	752	0.0742	11	0.2037	0.6014142	E
<input type="radio"/> GO:0007430	6,5	terminal branching of trachea, cytoplasmic projection extension (sensu Insecta)	11	0.0011	2	0.0370	0.6914397	E
<input type="radio"/> GO:0051234	4	establishment of localization	2162	0.2134	21	0.3889	0.7155814	E
<input type="radio"/> GO:0008045	10,7,11,12,8	motor axon guidance	12	0.0012	2	0.0370	0.8254633	E
<input type="radio"/> GO:0007442	8,7	hindgut morphogenesis	40	0.0039	2	0.0370	1.0000000	E
<input type="radio"/> GO:0006793	5	phosphorus metabolism	651	0.0643	8	0.1481	1.0000000	E
<input type="radio"/> GO:0008610	6,5,7	lipid biosynthesis	86	0.0085	1	0.0185	1.0000000	E
<input type="radio"/> GO:0046621	5	negative regulation of organ size	10	0.0010	1	0.0185	1.0000000	E
<input type="radio"/> GO:0006041	7,8	glucosamine metabolism	75	0.0074	1	0.0185	1.0000000	E
<input type="radio"/> GO:0009416	5	response to light stimulus	53	0.0052	2	0.0370	1.0000000	E
<input type="radio"/> GO:0008654	8,7,9	phospholipid biosynthesis	26	0.0026	1	0.0185	1.0000000	E
<input type="radio"/> GO:0040011	3	locomotion	310	0.0306	5	0.0926	1.0000000	E
<input type="radio"/> GO:0001654	4	eye development	350	0.0346	2	0.0370	1.0000000	E
<input type="radio"/> GO:0045034	7,6	neuroblast division	14	0.0014	1	0.0185	1.0000000	E
<input type="radio"/> GO:0019941	9,10	modification-dependent protein catabolism	78	0.0077	1	0.0185	1.0000000	E
<input type="radio"/> GO:0009966	5,4	regulation of signal transduction	112	0.0111	1	0.0185	1.0000000	E
<input type="radio"/> GO:0043067	6,5	regulation of programmed cell death	125	0.0123	1	0.0185	1.0000000	E
<input type="radio"/> GO:0019933	8	cAMP-mediated signaling	13	0.0013	1	0.0185	1.0000000	E
<input type="radio"/> GO:0016079	9,7,8	synaptic vesicle exocytosis	58	0.0057	1	0.0185	1.0000000	E
<input type="radio"/> GO:0030832	9	regulation of actin filament length	13	0.0013	1	0.0185	1.0000000	E
<input type="radio"/> GO:0007267	4	cell-cell signaling	487	0.0481	4	0.0741	1.0000000	E
<input type="radio"/> GO:0007460	5	photoreceptor fate commitment (sensu Embryoniscordi)	42	0.0042	1	0.0185	1.0000000	E

Contact information: David MARTIN
(e-mail: martin@bcm.umr-rms.fr)

Reference:
GOToolBox: functional analysis of gene datasets
based on Gene Ontology.
Martin et al. Genome Biology 2004;5(12):R101.
Epub 2004 Nov 26.
[PMID: 15575967](https://doi.org/10.1186/gb-2004-5-12-r101)



Done

Large-scale microarray analysis with GOToolBox

The new GOToolBox suite provides Perl scripts to automatize the tasks (check the download section)!

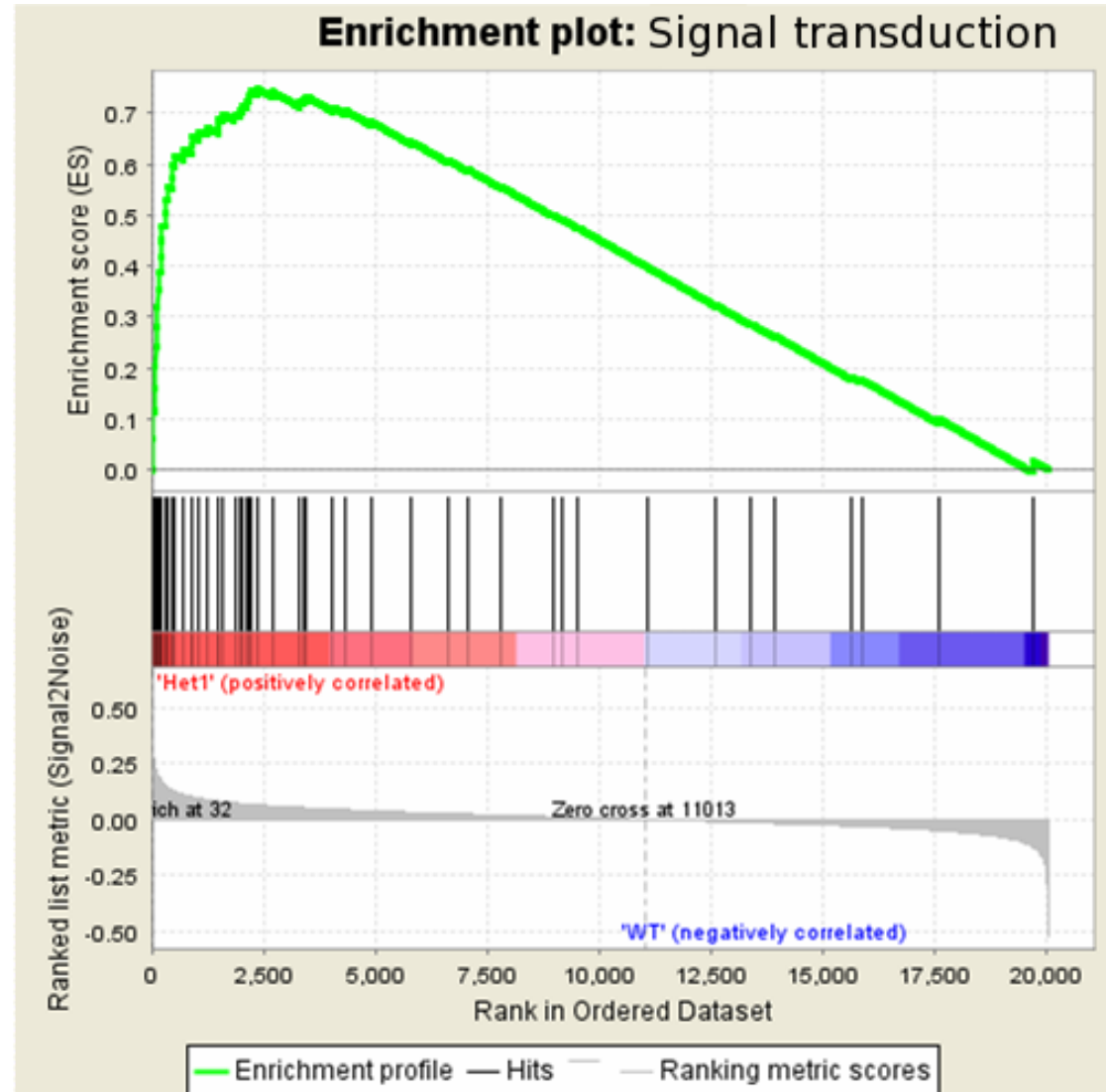
These scripts allow you to launch as many GO analysis as you want without having to stay in front of the computer.

These scripts can be easily modified to fit your needs.

<http://www.geneontology.org/GO.tools.microarray.shtml>

Other GO-based microarray analysis tools

- Most of them are functioning on the same principle than GOTOolBox.
- Some are additionally integrating other sources of annotations.
- An interesting recent one is GSEA (not in the GO tools list).



<http://www.broad.mit.edu/gsea/>

Grazie mille!

Thanks for your attention and interest.

Thanks a lot to the organizers of the workshop!!!