Functional analysis of gene lists
...WORK IN PROGRESS...

Ferran Briansó and Alex Sánchez-Pla.
Statistics department. UB
& Statistics and Bioinformatics Unit (UEB). VHIR.

October 25, 2016

Contents

1 Introduction 1
1.1 Input Data for Functional Analysis 2
1.2 Input data preprocessing 3

2 Analysis methods and tools 4
2.1 Enrichment Analysis with `GOstats` and `DAVID` 5
2.1.1 Enrichment Analysis using `GOstats` 5
2.1.2 Enrichment Analysis using `topGO` 5
2.1.3 Enrichment Analysis (?) using DAVID 5
2.2 Gene Set Enrichment Analysis 5
2.2.1 Gene Set Enrichment Analysis with `GAGE` 5
2.2.2 Gene Set Enrichment Analysis with `GSA` 6
2.3 Network Analysis 6
2.3.1 Analysis with `FGNet` 6
2.3.2 Analysis with `IPA` 6

3 Results 7
3.1 Gene enrichment analysis using `GOstats` 7
3.2 Gene enrichment analysis using `topGO` 9
3.2.1 Comparison between `GOstats` and `topGO` output 12

4 Results presentation 12

1 Introduction

This document provides some information on the different analyses perfomed on several gene lists to help gain biological insight on the results of a differential expression analysis. Overall these analyses are known as *Functional Analysis*.

Functional analysis can be performed in many different ways that lead to similar (or not-so-similar) results. Because there is not a universal acceptance of what is a `complete, well done functional analysis` some different approaches will be shown and an arbitrary selection based on ”doability”, ”scriptability” and of course ”interpretability” will be performed.

Complementary information to this document can be found at `http://eib.stat.ub.edu/GO2016`

1.1 Input Data for Functional Analysis

Functional analysis can be made, on a first approach on

- A list of genes selected by being differentially expressed in a given experimental setting.
- The whole list of genes -or even the whole expression matrix- that has been used in the analysis.

Most tools require that gene list consist of gene identifiers in some standard notation such as `Entrez`, `ENSEMBL` or other related to these.

These gene lists can be easily extracted from output tables provided by microarrays or RNA-seq data analysis tools.

The analysis below is applied on a set of three gene lists obtained from a cancer study, but it can be easily extended to more lists or other studies.

```
##
## Header of top Table for comparison AvsB
```

	SymbolsA	EntrezsA	logFC	AveExpr	t
204667_at	FOXA1	3169	-3.038344	8.651157	-14.362164
215729_s_at	VGLL1	51442	3.452290	6.137595	12.814829
220192_x_at	SPDEF	25803	-3.016315	9.521883	-10.859194
214451_at	TFAP2B	7021	-5.665059	7.432823	-10.829548
217528_at	CLCA2	9635	-5.622086	6.763101	-9.666128
217284_x_at	SERHL2	253190	-4.313116	9.133307	-9.528373

```
##
## Header of top Table for comparison AvsL
```

	SymbolsA	EntrezsA	logFC	AveExpr	t
205009_at	TFF1	7031	4.735050	8.692478	10.564670
205862_at	GREB1	9687	3.958563	6.082835	9.983993
205225_at	ESR1	2099	3.964939	9.300546	9.849787
209443_at	SERPINA5	5104	2.198392	7.586226	8.531873
217528_at	CLCA2	9635	-4.429254	6.763101	-7.615275
205696_s_at	GFRA1	2674	2.333785	6.239876	7.600491

```
##
## Header of top Table for comparison BvsL
```

	SymbolsA	EntrezsA	logFC	AveExpr	t
204667_at	FOXA1	3169	2.961042	8.651157	13.996760
215729_s_at	VGLL1	51442	-3.744599	6.137595	-13.899875
205009_at	TFF1	7031	5.729322	8.692478	12.783054
205225_at	ESR1	2099	3.939276	9.300546	9.786035
205862_at	GREB1	9687	3.774303	6.082835	9.519268
218211_s_at	MLPH	79083	2.808408	10.932769	8.813968

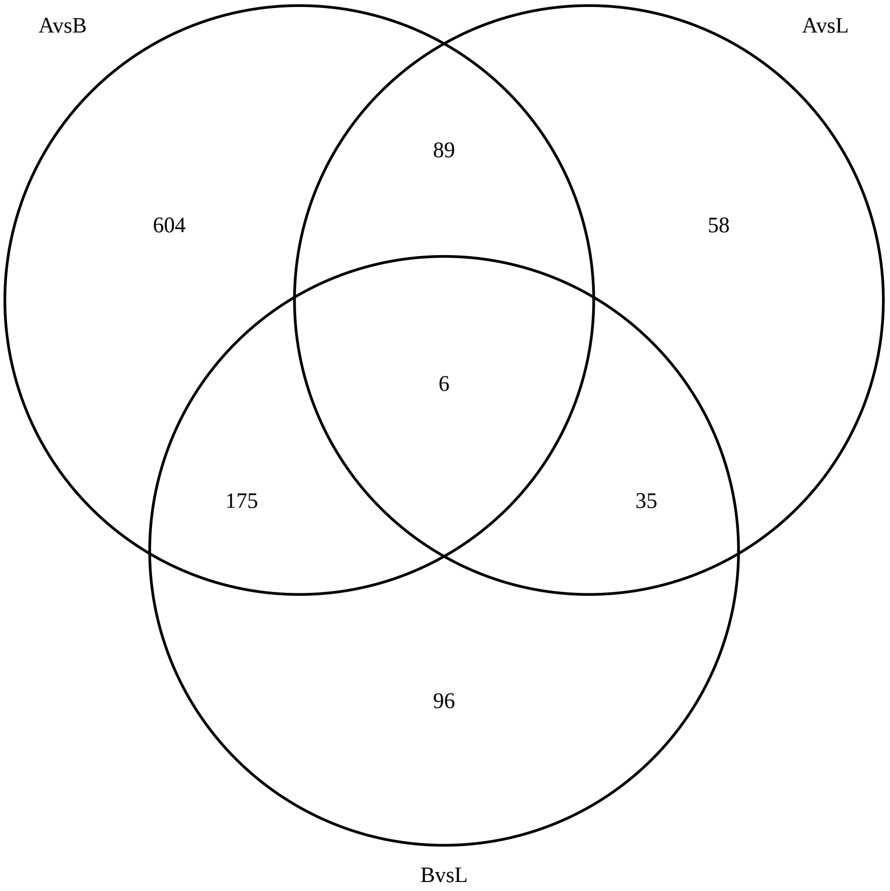
1.2 Input data preprocessing

Sometimes lists may need some preprocessing (e.g. in this example the gene list has multiple transcripts per gene identifier that have to be unitized previous to the analysis).

This is done using two ad-hoc functions created specifically for this aim and available from github.

```
##
## Number of genes selectable (AvsB) with adjusted p-value < 0.1 and logFC > 0.75:
874
##
## Number of genes selectable (AvsL) with adjusted p-value < 0.1 and logFC > 0.75:
188
##
## Number of genes selectable (BvsL) with adjusted p-value < 0.1 and logFC > 0.75:
312
```

The following diagram shows which genes there are in common (or not) between the three lists.



```
## null device
##           1
```

2 Analysis methods and tools

Following [4] three different approaches can be applied on these data:

- **Classical enrichment or Overrepresentation Analysis**. This has been done using ad-hoc functions included in our analysis pipelines and also using DAVID a public software tool available at `https://david.ncifcrf.gov/`.
- **Gene Set Expression Analysis**. This can be done using the Bioconductor package `gage` which provides nice representations of genes overexpressed/downregulated in the context of KEGG pathways. A simpler but powerful version of this type of analysis can be applied with package `GSA`.

- A **Network analysis** can be done using Bioconductor packages such as FGNet. It is also possible to use commercial software such as the Ingenuity Pathways Software.

2.1 Enrichment Analysis with GOstats and DAVID

Given a list of (potentially) differentially expressed genes Enrichment Analysis or Overrepresentation Analysis (**ORA**) seeks to select functions and biological processes that characterize this list, this meaning that these functions appear more often in the list than in the remaining set of analyzed genes.

There have been developped many variations of this type of analysis ([3]). We have applied "classical" enrichment analysis ([1]) implemented in the GOstats and topGO Bioconductor packages. We also present the improved version of ORA implemented in the DAVID software package, which is powerful but harder to use because it is based on an online tool difficult to "script".

2.1.1 Enrichment Analysis using GOstats

2.1.2 Enrichment Analysis using topGO

2.1.3 Enrichment Analysis (?) using DAVID

DAVID (the Database for Annotation, Visualization and Integrated Discovery, [2]) is a free online bioinformatics resource developed by the Laboratory of Immunopathogenesis and Bioinformatics at NIH (*National Institute of Allergy and Infectious Diseases (NIAID)*).

DAVID provides a set of tools to help the functional interpretation of lists of genes derived from genomic studies. DAVID can be found at http://david.niaid.nih.gov or http://david.abcc.ncifcrf.gov

Given an uploaded gene list, the DAVID Resources provides classical gene-term enrichment analysis, but also new tools and functions that allow users to condense large gene lists into gene functional groups or cluster redundant and heterogeneous terms into groups. This grouping jointly with the dynamical access from results to Biological knowledge database provides a clearer and easy to interpret output than classical enrichment analysis.

2.2 Gene Set Enrichment Analysis

Gene set analysis (GSA) is a widely used strategy for gene expression data analysis based on pathway knowledge. GSA focuses on sets of related genes and has established major advantages over individual gene analyses, including greater robustness, sensitivity and biological relevance.

2.2.1 Gene Set Enrichment Analysis with GAGE

We have applied a recent version of GSA called Generally Applicable Gene-set Enrichment (GAGE, [5]). This method is more robust than other existing related approaches and has been seen to work well with different sample sizes, experimental designs and profiling techniques. GAGE has been shown to provide significantly good results results in the following three aspects: (1) consistency across repeated studies/experiments; (2) sensitivity and specificity; (3) biological relevance of the regulatory mechanisms inferred.

Essentially what GAGE -as most GSEA methods- does is to test if a given gene set is *associated* with a gene list. That is for each gene set tested it selects those gene sets whose expression tends to be higher or lower than the expression of the genes in the genelist (in which case the gene set is called to be upregulated or downregulated).

Gene Sets can be GO categories or KEGG pathways and each is shown in a different format.

- GO categories over or underrepresented are shown in a heatmap
- KEGG pathways are shown on a picture of the pathway with the genes that belong to the list marked in red or green depending on if they are up or down-regulated.

2.2.2 Gene Set Enrichment Analysis with `GSA`

The `GSA` package offers a simple approach to gene set analysis which has two main advantages:

- It is very intuitive and simple to use
- It can be applied in situations more complex than the usual ”two sample test” that many GSEA methods assume.

An obvious drawback is that the output is also simple and consists of a list of differentially expressed gene sets, without any links to databases or pathway visualizations.

2.3 Network Analysis

Network analysis is a generic expression to describe distinct analysis that use some type of network or graph representation of the data or the results. Related with what we are describing here one can find different methods for *Module enrichment analysis* aimig at finding subnetworks with a particular biological meaning. We consider two such programs here, `FGNet` and `Inegnuity Pathway Analysis`.

2.3.1 Analysis with `FGNet`

2.3.2 Analysis with `IPA`

Ingenuity Pathway Analysis is a commercial tool allows searching and using information extracted from public databases and from full text article extraction. It categorizes its findings based on a in-house curated ontology, covering entities (proteins and other molecules), relationships between the entities (inhibites, activates,...) and functional information (pathways, biological process, disease,...).

IPA has several differences with the previous tools.

- It relies not only on public databases but on their own commercial databases and ontology, that they claim to be better anotated and more intuitive.
- It performs enrichment analysis but besides it does different types of network analysis and allows to view selected genes in:
    1. Association networks, based on co-citation of genes in the literature
    2. Canonical pathways based on known biological pathways.
- It provides other types of information for analyzed genes such as their association with disease or their potential toxicogenomic properties.

3 Results

This section contains the results of applying *some of* the analyses described above to the examples lists.

3.1 Gene enrichment analysis using GOstats

We present below two ways to do enrichment analysis with GOStats. The first one makes direct use of functions available in the GOstats package, while the second uses some home-encapsulated functions.

```
##
## Comparison:  AvsB
## GO
##       GOBPID       Pvalue OddsRatio  ExpCount Count Size
## 1 GO:0006629 2.588340e-06  1.707120 79.057118   117  567
## 2 GO:0044255 1.034654e-05  1.744666 61.210009    93  439
## 3 GO:0032787 1.275704e-05  1.985539 35.973080    61  258
## 4 GO:0008610 2.121849e-05  1.891353 40.434858    66  290
## 5 GO:0002682 2.469534e-05  1.551362 99.971700   137  717
## 6 GO:0007259 3.052054e-05  3.385397  8.365833    21   60
##                                    Term
## 1               lipid metabolic process
## 2      cellular lipid metabolic process
## 3 monocarboxylic acid metabolic process
## 4            lipid biosynthetic process
## 5   regulation of immune system process
## 6                      JAK-STAT cascade
## KEGG
##   KEGGID       Pvalue OddsRatio  ExpCount Count Size
## 1  05213 0.0003981226  3.901831  4.820961    13   33
## 2  05223 0.0008045551  3.782377  4.528781    12   31
## 3  04010 0.0024236800  2.001661 16.508138    28  113
## 4  03320 0.0024503737  2.992958  5.697499    13   39
## 5  04914 0.0054220691  2.538264  6.866217    14   47
## 6  01040 0.0064550128  5.076559  1.899166     6   13
##                                       Term
## 1                    Endometrial cancer
## 2             Non-small cell lung cancer
## 3                 MAPK signaling pathway
## 4                 PPAR signaling pathway
## 5 Progesterone-mediated oocyte maturation
## 6 Biosynthesis of unsaturated fatty acids
##
## Comparison:  AvsL
## GO
##       GOBPID       Pvalue OddsRatio   ExpCount Count Size
## 1 GO:0044281 5.440527e-06  2.111702 36.6372735    62 1173
## 2 GO:0005996 4.001358e-05  3.811863  4.5289042    15  145
## 3 GO:0044282 5.275232e-05  3.532283  5.1848145    16  166
## 4 GO:0006570 1.171659e-04 94.601124  0.1249353     3    4
## 5 GO:0043436 1.678394e-04  2.216064 16.4914582    32  528
## 6 GO:0006082 1.931328e-04  2.196617 16.6163934    32  532
##                                Term
## 1 small molecule metabolic process
## 2 monosaccharide metabolic process
## 3 small molecule catabolic process
## 4       tyrosine metabolic process
## 5        oxoacid metabolic process
## 6   organic acid metabolic process
## KEGG
##   KEGGID       Pvalue OddsRatio   ExpCount Count Size
## 1  00983 0.0002266541 12.047952  0.6097658     5   16
## 2  00350 0.0005507651  9.454474  0.7240969     5   19
## 3  01100 0.0017676527  1.982550 21.3036919    34  559
## 4  00360 0.0038130654 12.994624  0.3429933     3    9
## 5  00010 0.0055977524  5.065511  1.1814212     5   31
## 6  00982 0.0071986206  6.153453  0.8003176     4   21
##                                 Term
## 1   Drug metabolism - other enzymes
## 2               Tyrosine metabolism
## 3                Metabolic pathways
## 4          Phenylalanine metabolism
## 5      Glycolysis / Gluconeogenesis
## 6 Drug metabolism - cytochrome P450
##
## Comparison:  BvsL
## GO
##       GOBPID       Pvalue OddsRatio    ExpCount Count Size
## 1 GO:0036499 9.461963e-06 19.802158   0.5880932     6   12
## 2 GO:0032501 1.017167e-05  1.694021 121.6372735   157 2482
## 3 GO:0044707 1.104044e-05  1.690072 118.8438309   154 2425
## 4 GO:0006984 1.571641e-05  9.367434   1.2251941     8   25
## 5 GO:0048732 2.050040e-05  2.776800  10.7327006    26  219
## 6 GO:0008285 2.216348e-05  2.404664  16.1235548    34  329
##                                         Term
## 1   PERK-mediated unfolded protein response
## 2          multicellular organismal process
## 3     single-multicellular organism process
## 4              ER-nucleus signaling pathway
## 5                         gland development
## 6 negative regulation of cell proliferation
## KEGG
##   KEGGID       Pvalue OddsRatio ExpCount Count Size
## 1  05219 0.0006285001  5.871449 1.554188     7   29
## 2  05216 0.0033135057  6.074359 1.071854     5   20
## 3  00010 0.0051396857  4.388837 1.661374     6   31
## 4  05222 0.0070941074  3.201643 2.894006     8   54
## 5  04974 0.0094987904  3.777065 1.875744     6   35
##                                Term
## 1                    Bladder cancer
## 2                    Thyroid cancer
## 3      Glycolysis / Gluconeogenesis
## 4            Small cell lung cancer
## 5  Protein digestion and absorption

## Error in ‘*tmp*‘[[i]]:  subscript out of bounds
```

The analysis below is done using the functions defined in file `hyperGeometricAnalysis.R`

3.2 Gene enrichment analysis using `topGO`

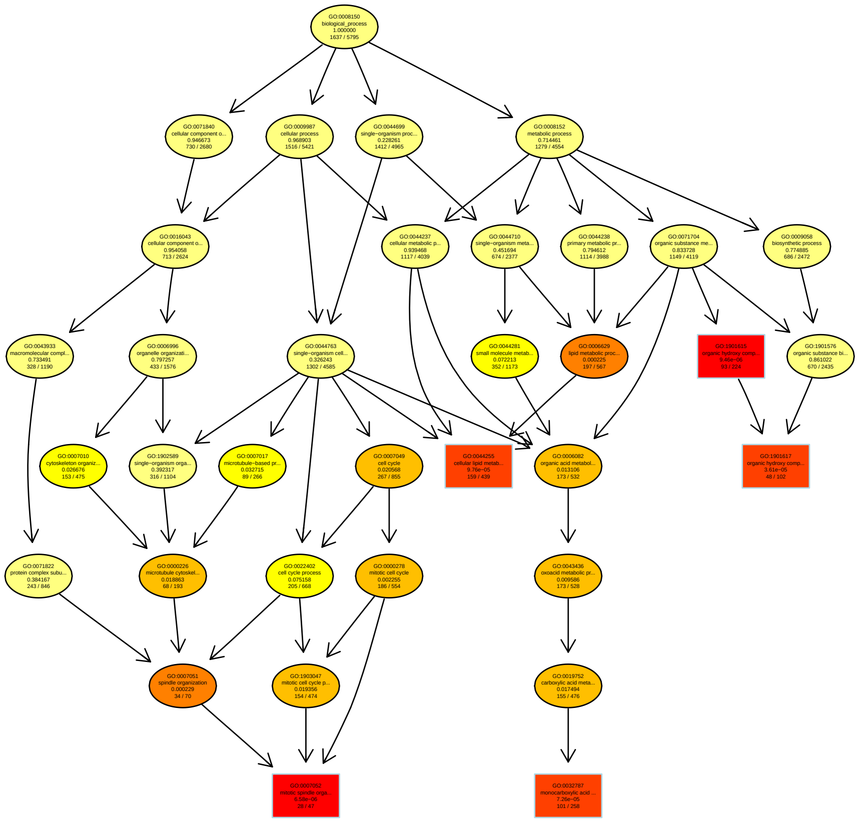
The analysis using `topGO` is relatively similar to the one done with `GOstats`. One creates a data structure that is used to call the main analysis function. However it is different from that from `GOstats` in that, instead of two lists of genes (Universe/Differentially expressed) it is provided with a list of scores, *whose names are the gene names* and the name of a function whos output (TRUE/FALSE) helps selecting which genes are differentially expressed in the list.

Warning: We are using the R2HTML package to write the output. This package doesn't admit whitespaces in the output path so files are written to another directory and recovered later.

```
## Loading required package:  R2HTML

##
## Building most specific GOs .....

## Loading required package:  hgu133a.db
##

##  ( 8208 GO terms found. )
##
## Build GO DAG topology .......... ( 12028 GO terms and 28831 relations. )
##
## Annotating nodes ............... ( 5795 genes annotated to the GO terms. )
##
##   -- Classic Algorithm --
##
##   the algorithm is scoring 4249 nontrivial nodes
##   parameters:
##   test statistic:  fisher
##
##  *** Output redirected to directory:  ~
##  *** Use HTMLStop() to end redirection.
## Building most specific GOs ..... ( 8208 GO terms found. )
##
## Build GO DAG topology .......... ( 12028 GO terms and 28831 relations. )
##
## Annotating nodes ............... ( 5795 genes annotated to the GO terms. )
##
##   -- Classic Algorithm --
##
##   the algorithm is scoring 4073 nontrivial nodes
##   parameters:
##   test statistic:  fisher
##
##  *** Output redirected to directory:  ~
##  *** Use HTMLStop() to end redirection.
## Building most specific GOs ..... ( 8208 GO terms found. )
##
## Build GO DAG topology .......... ( 12028 GO terms and 28831 relations. )
##
## Annotating nodes ............... ( 5795 genes annotated to the GO terms. )
##
##   -- Classic Algorithm --
```

```
##
##   the algorithm is scoring 4183 nontrivial nodes
##   parameters:
##   test statistic:  fisher
##
##  *** Output redirected to directory:  ~
##  *** Use HTMLStop() to end redirection.
```

With topGO it is possible to extract a graph with nodes colored according to the significance. Below is shown the graph associated with the first comparison. Additionally all graphs are written to .pdf files


GO:0008150 biological_process 1.000000 1637 / 5795
GO:0071840 cellular component o... 0.946673 730 / 2680
GO:0009987 cellular process 0.968903 1516 / 5421
GO:0044699 single−organism proc... 0.228261 1412 / 4965
GO:0008152 metabolic process 0.714461 1279 / 4554
GO:0016043 cellular component o... 0.954058 713 / 2624
GO:0044237 cellular metabolic p... 0.939468 1117 / 4039
GO:0044710 single−organism meta... 0.451694 674 / 2377
GO:0044238 primary metabolic pr... 0.794612 1114 / 3988
GO:0071704 organic substance me... 0.833728 1149 / 4119
GO:0009058 biosynthetic process 0.774885 686 / 2472
GO:0043933 macromolecular compl... 0.733491 328 / 1190
GO:0006996 organelle organizati... 0.797257 433 / 1576
GO:0044763 single−organism cell... 0.326243 1302 / 4585
GO:0044281 small molecule metab... 0.072213 352 / 1173
GO:0006629 lipid metabolic proc... 0.000225 197 / 567
GO:1901615 organic hydroxy comp... 9.46e−06 93 / 224
GO:1901576 organic substance bi... 0.861022 670 / 2435
GO:0007010 cytoskeleton organiz... 0.026676 153 / 475
GO:1902589 single−organism orga... 0.392317 316 / 1104
GO:0007017 microtubule−based pr... 0.032715 89 / 266
GO:0007049 cell cycle 0.020568 267 / 855
GO:0044255 cellular lipid metab... 9.76e−05 159 / 439
GO:0006082 organic acid metabol... 0.013106 173 / 532
GO:1901617 organic hydroxy comp... 3.61e−05 48 / 102
GO:0071822 protein complex subu... 0.384167 243 / 846
GO:0000226 microtubule cytoskel... 0.018863 68 / 193
GO:0022402 cell cycle process 0.075158 205 / 668
GO:0000278 mitotic cell cycle 0.002255 186 / 554
GO:0043436 oxoacid metabolic pr... 0.009586 173 / 528
GO:0007051 spindle organization 0.000229 34 / 70
GO:1903047 mitotic cell cycle p... 0.019356 154 / 474
GO:0019752 carboxylic acid meta... 0.017494 155 / 476
GO:0007052 mitotic spindle orga... 6.58e−06 28 / 47
GO:0032787 monocarboxylic acid ... 7.26e−05 101 / 258

```
## $dag
## A graphNEL graph with directed edges
## Number of Nodes = 35
## Number of Edges = 56
```

```
##
## $complete.dag
## [1] "A graph with 35 nodes."
```

3.2.1 Comparison between GOstats and topGO output

4 Results presentation

The results of the analyses performed following the methodology described in the previous sections consist of a high number of tables and figures.

In order to facilitate its organization and review a web page (an html file indeed) is prepared that allows to access each file.

This page consists consists of a list of links organized by topics which correspond to the files outputted as result of the different steps undertaken in the analysis.

A brief description of each group of results and its contents follows below.

1. Section **Reports and results summaries** gives access to the report describing the goals and methods applied in the analysis.

2. Section **Input Files for Biological Significance** gives access to different files from where the genes to be included in the analysis have been obtained.

3. Section **Gene Set Expression Analysis (GAGE)** gives access to the lists of results found, for each comparison of interest, in the Gene Set Enrichment Analysis against the Kyoto Encyclopedia of Genes and the Gene Ontology database, considering up- and down-regulated elements separately.

   - For the analysis against the Gene Ontology database. GO Molecular Functions (MFs), GO Biological Processes (BPs), and GO Cellular Components (CCs) classifications have been tested, considering up- and down-regulated elements separately. In this case, only GO terms with an enrichment test q-value (adjusted p-value) below 0.05 have been considered as relevant and included in the tables.
   - For each KEGG analyis, a zipped folder with details (plots and xml or txt files) is provided. In this case, only pathways with an enrichment test q-value (adjusted p-value) below 0.15 have been considered as relevant and included in the tables.

4. Section **DAVID Analysis Results**

   This section gives access to two types of files generated by DAVID

   - Functional annotation chart with the main results of enrichment analysis

- Functional annotation clustering where the resulting categories obtained from enrichment analysis are grouped by similarity of functions.

5. Section **Ingenuity Pathway Analysis**

   IPA generates many types of results and is intended to be used interactively exploring them on their web site (which requires a password protected access). In order to have an overview of what has been obtaineda pdf with a summary of these resuts can be generated.

   The links in this section give access to the analysis summary for each gene list.

References

[1] R. Gentleman. Using go for statistical analysis. *Bioconductor's Compendiums (www.bioconductor.org)*, 2004.

[2] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. 37(1):1–13.

[3] P. Khatri and S. Drăghici. Ontological analysis of gene expression data: current tools, limitations, and problems. *Bioinformatics*, 18:3587–3595, 2005.

[4] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375, 02 2012.

[5] Weijun Luo, Michael S. Friedman, Kerby Shedden, Kurt D. Hankenson, and Peter J. Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. 10:161.