Functional analysis

Isabelle Dupanloup
BCF - Bioinformatics Core Facility
SIB - Swiss Institute of Bioinformatics
isabelle.dupanloup@sib.swiss

Swiss Institute of Bioinformatics

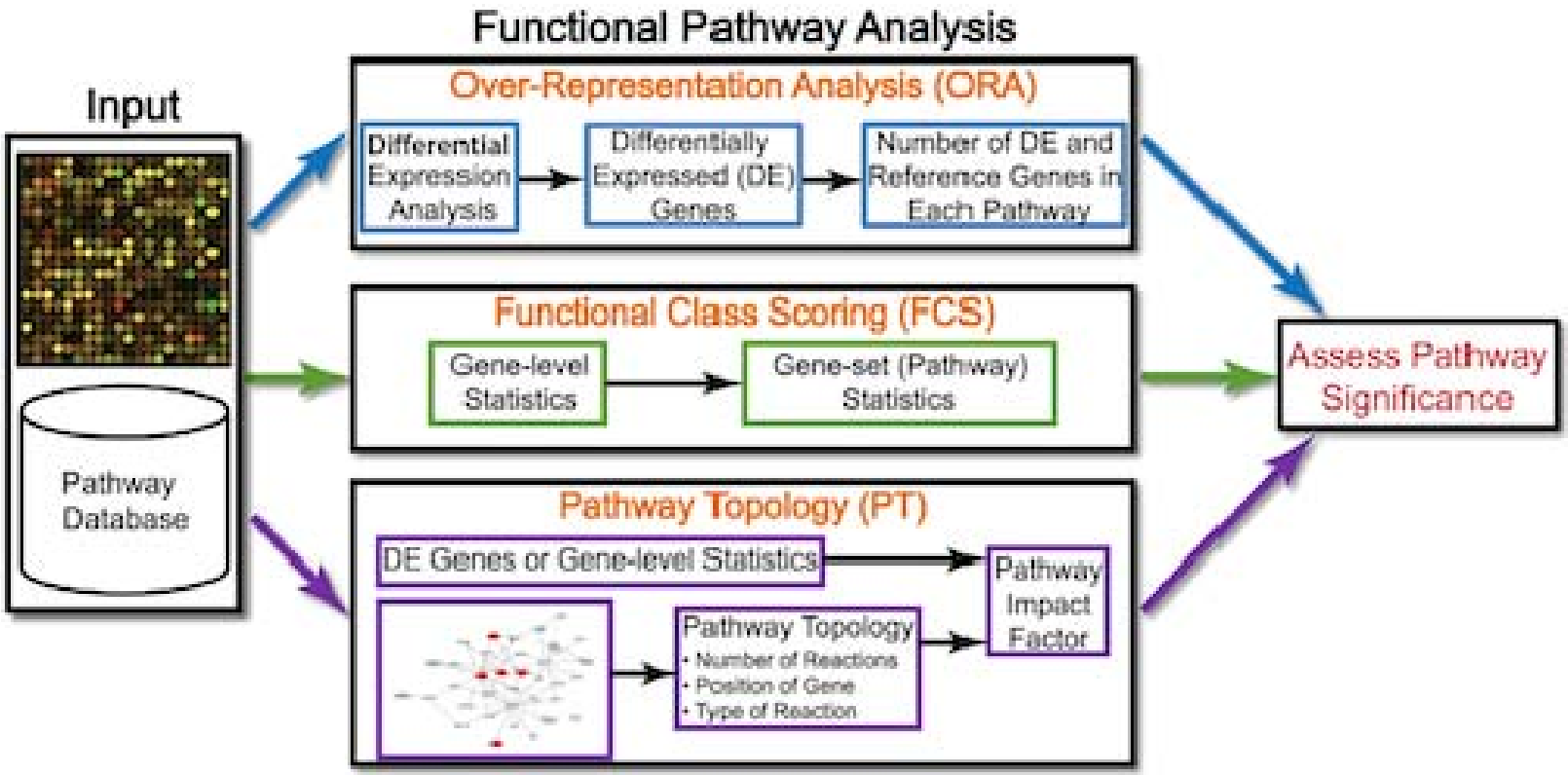
Teaching material from
Harvard Chan Bioinformatics Core training

Functional analysis

Learning Objectives

- Determine how functions are attributed to genes using Gene Ontology terms
- Understand the theory of how functional enrichment tools yield statistically enriched functions or interactions
- Discuss functional analysis using over-representation analysis, functional class scoring, and pathway topology methods
- Explore functional analysis tools

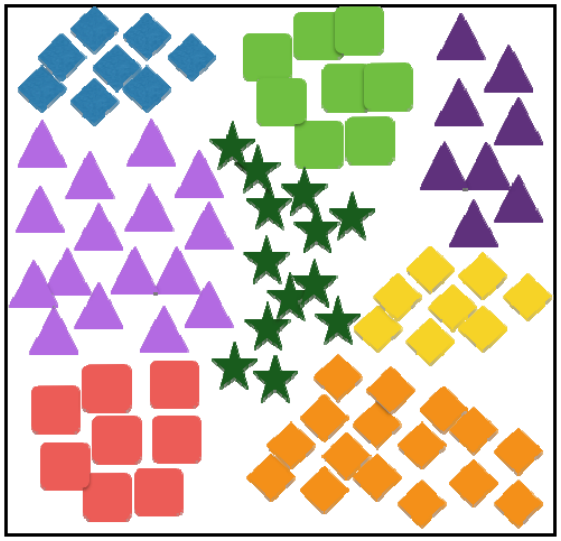
Functional analysis

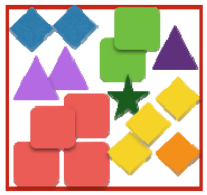


Functional analysis

Over-representation analysis

All known genes in a species
(categorized into groups)





DEGs

Genes categories	Organism-specific Background	DE results	Over-represented?
Functional category 1	35/13000	25/1000	Likely
Functional category 2	56/13000	4/1000	Unlikely
Functional category 3	90/13000	8/1000	Unlikely
Functional category 4	15/13000	10/1000	Likely
…			
…			

Hypergeometric test $$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Functional analysis

Gene Ontology project

- collaborative effort to address the need for consistent descriptions of gene products across databases
- GO Consortium: develop a comprehensive, computational model of biological systems, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life
- GO terms = GO categorizations
- GO term: each with a name (DNA repair) and a unique accession number (GO:0005125)

Functional analysis

Gene Ontology

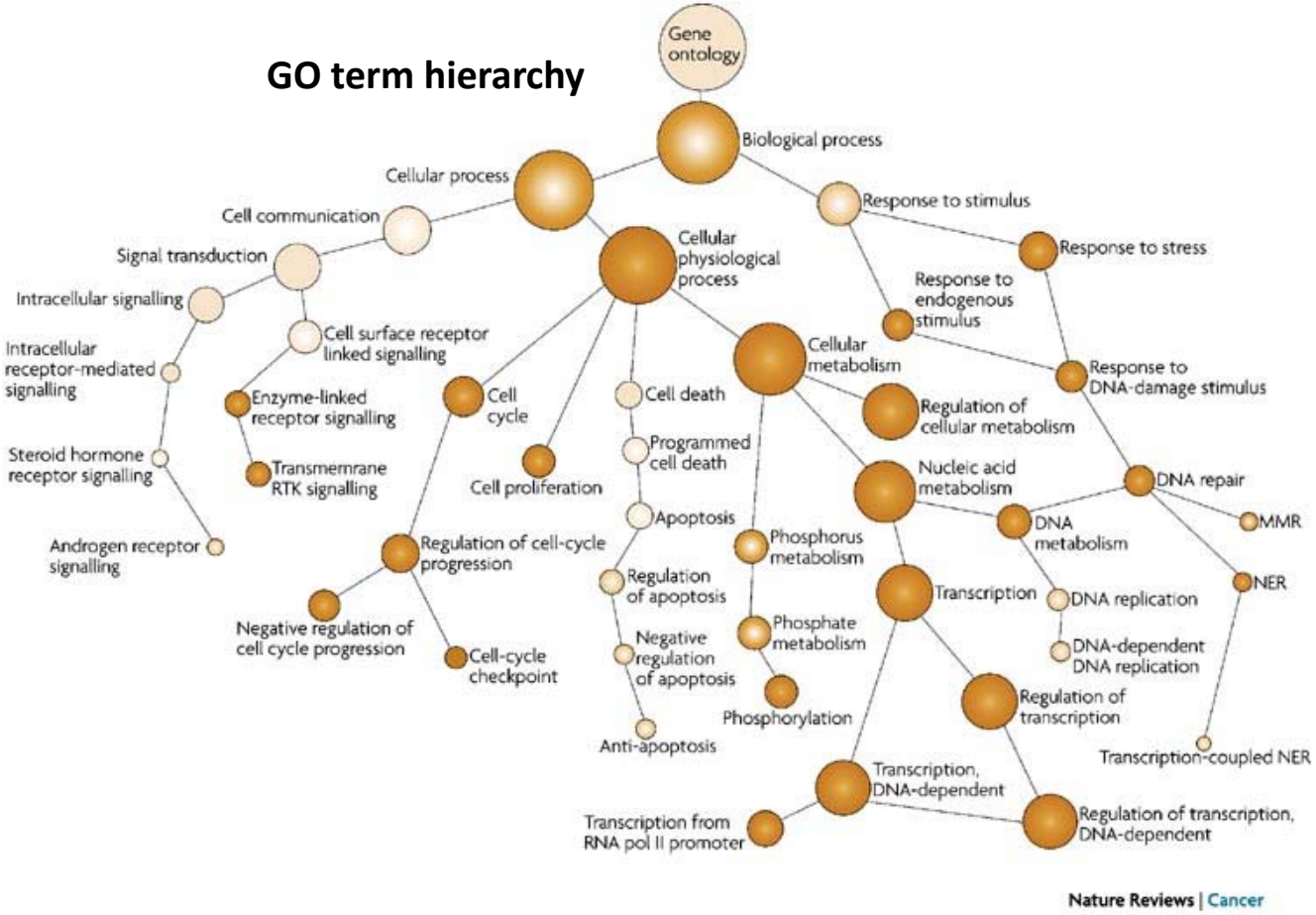
GO ontologies: GO terms organized in 3 independent controlled vocabularies

- **Biological process**: refers to the biological role involving the gene or gene product, and could include "transcription", "signal transduction", and "apoptosis". A biological process generally involves a chemical or physical change of the starting material or input.
- **Molecular function**: represents the biochemical activity of the gene product, such activities could include "ligand", "GTPase", and "transporter".
- **Cellular component**: refers to the location in the cell of the gene product. Cellular components could include "nucleus", "lysosome", and "plasma membrane".

Functional analysis

Gene Ontology

GO term hierarchy

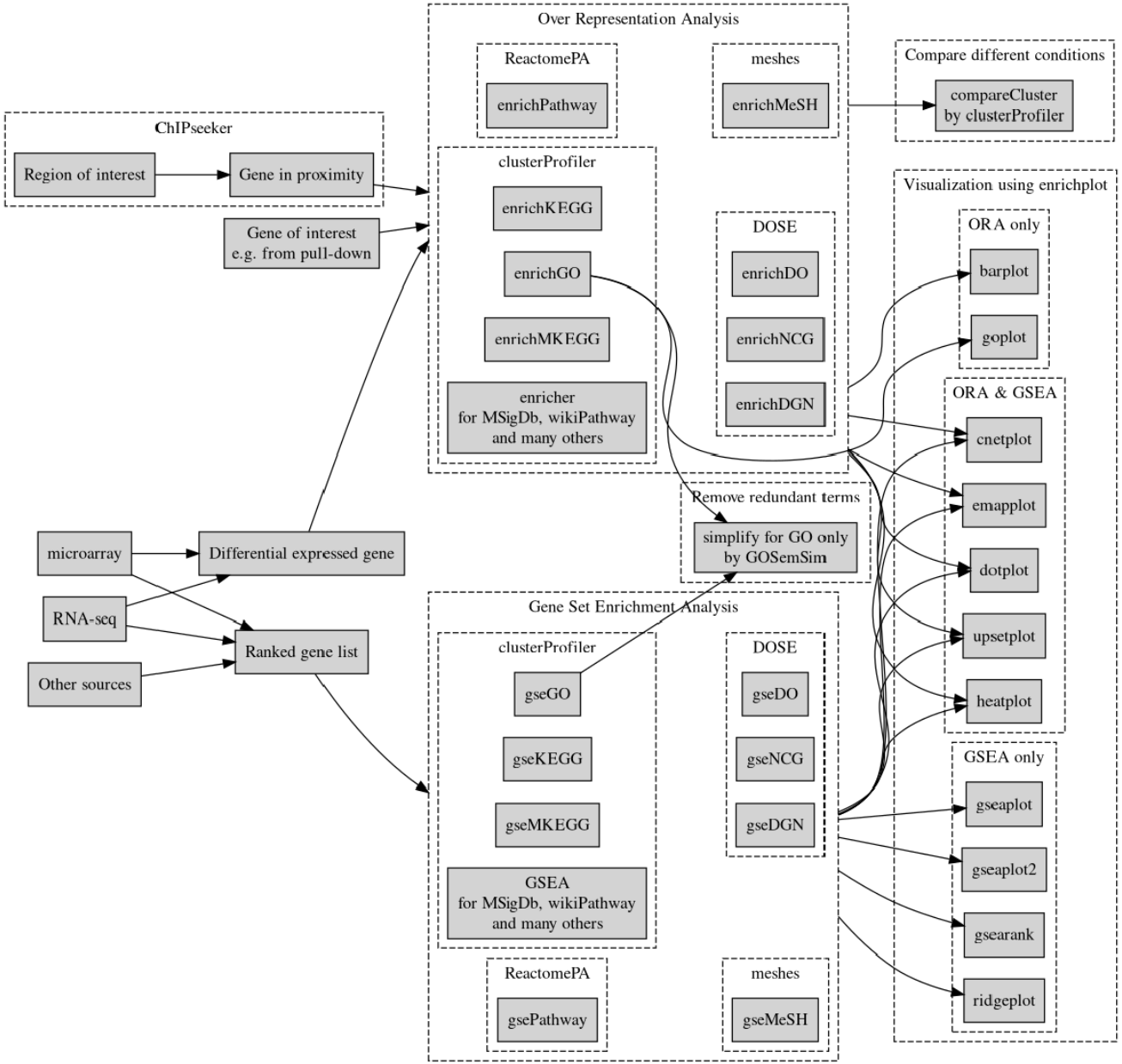


Sources of gene sets

- Online:
- MSigDB: database containing several types of gene set lists
  - https://www.gsea-msigdb.org/gsea/msigdb/index.jsp
  - GO
  - hallmark
  - published gene sets
- KEGG (bi-directional eg mTOR signaling): https://www.kegg.jp/kegg/pathway.html
- Reactome https://reactome.org/
- WikiPathways https://www.wikipathways.org/index.php/WikiPathways

This package implements methods to analyze and visualize functional profiles of genomic coordinates (supported by [ChIPseeker](), gene and gene clusters. https://yulab-smu.github.io/clusterProfiler-book/.

Authors
Guangchuang YU
School of Basic Medical Sciences, Southern Medical University



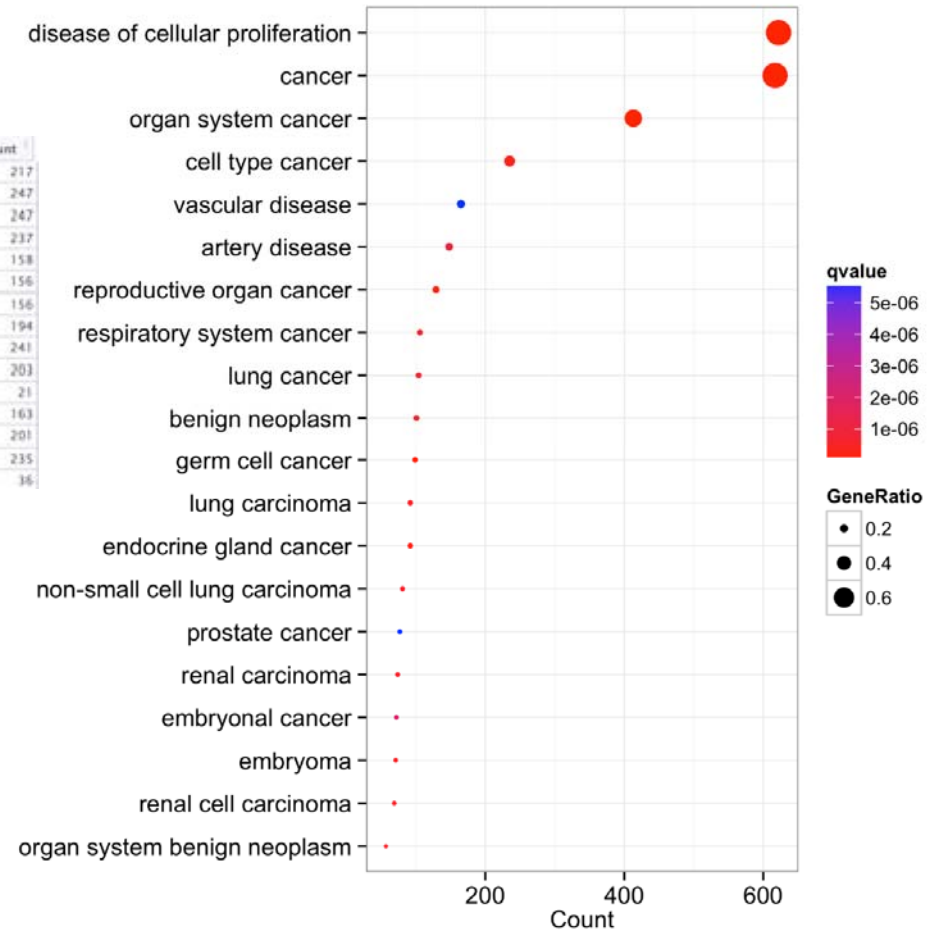
Functional analysis

Over-representation analysis

GO enrichment with clusterProfiler

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0008380	GO:0008380	RNA splicing	217/5660	393/16649	2.299032e-18	1.015934e-14	8.427986e-15	RBM11/RBM15B/RBM38/SNRPD1/SNRPD3/PP...	217
GO:0006397	GO:0006397	mRNA processing	247/5660	463/16649	3.630282e-18	1.015934e-14	8.427986e-15	RBM11/RBM15B/APP/RBM38/SNRPD1/FASTK...	247
GO:0010608	GO:0010608	posttranscriptional regulation of gene expres...	247/5660	481/16649	1.544298e-15	2.881145e-12	2.390140e-12	RPS27L/SMAD2/APP/RBM38/PSMB7/LSM14A...	247
GO:0034660	GO:0034660	ncRNA metabolic process	237/5660	473/16649	1.884052e-13	2.636260e-10	2.186988e-10	SMAD2/RAE1/SNRPD1/SMARCB1/RRP7BP/RP...	237
GO:0000375	GO:0000375	RNA splicing, via transesterification reactions	158/5660	292/16649	9.323903e-13	9.270219e-10	7.690386e-10	RBM11/RBM15B/SNRPD1/SNRPD3/SNRPC/LS...	158
GO:0000377	GO:0000377	RNA splicing, via transesterification reactions...	156/5660	288/16649	1.159398e-12	9.270219e-10	7.690386e-10	RBM11/RBM15B/SNRPD1/SNRPD3/SNRPC/LS...	156
GO:0000398	GO:0000398	mRNA splicing, via spliceosome	156/5660	288/16649	1.159398e-12	9.270219e-10	7.690386e-10	RBM11/RBM15B/SNRPD1/SNRPD3/SNRPC/LS...	156
GO:0022613	GO:0022613	ribonucleoprotein complex biogenesis	194/5660	379/16649	2.554931e-12	1.787494e-09	1.482869e-09	RPS27L/LSM14A/SNRPD1/RRP7BP/RPSA/WDR...	194
GO:0044772	GO:0044772	mitotic cell cycle phase transition	241/5660	499/16649	1.576324e-11	9.802982e-09	8.132356e-09	RPS27L/UBE2C/APP/INO80/RBM38/PSMB7/S...	241
GO:0018205	GO:0018205	peptidyl-lysine modification	203/5660	411/16649	5.414440e-11	3.030462e-08	2.514010e-08	RAE1/CTCFL/SMARCB1/RTF1/BAG6/EEF1AK...	203
GO:0002486	GO:0002486	antigen processing and presentation of endo...	21/5660	21/16649	1.410396e-10	7.176352e-08	5.953356e-08	HLA-C/HLA-A/HLA-B/HLA-A/HLA-C/HLA-A/HL...	21
GO:0034470	GO:0034470	ncRNA processing	163/5660	320/16649	2.315204e-10	1.079850e-07	8.958215e-08	SMAD2/RAE1/RRP7BP/RPSA/WDR46/TBL/NUP...	163
GO:0016570	GO:0016570	histone modification	201/5660	412/16649	2.625099e-10	1.130206e-07	9.375959e-08	CTCFL/DAXX/SMARCB1/RTF1/ZNF335/ATG7...	201
GO:0006281	GO:0006281	DNA repair	235/5660	496/16649	2.860491e-10	1.143583e-07	9.486937e-08	RPS27L/INO80/SMARCB1/BACH1/RBBP8/IER3...	235
GO:0002480	GO:0002480	antigen processing and presentation of exog...	36/5660	45/16649	3.073965e-10	1.146999e-07	9.515270e-08	HLA-E/HLA-F/HLA-C/HLA-A/HLA-B/HLA-E/HL...	36

dotplot



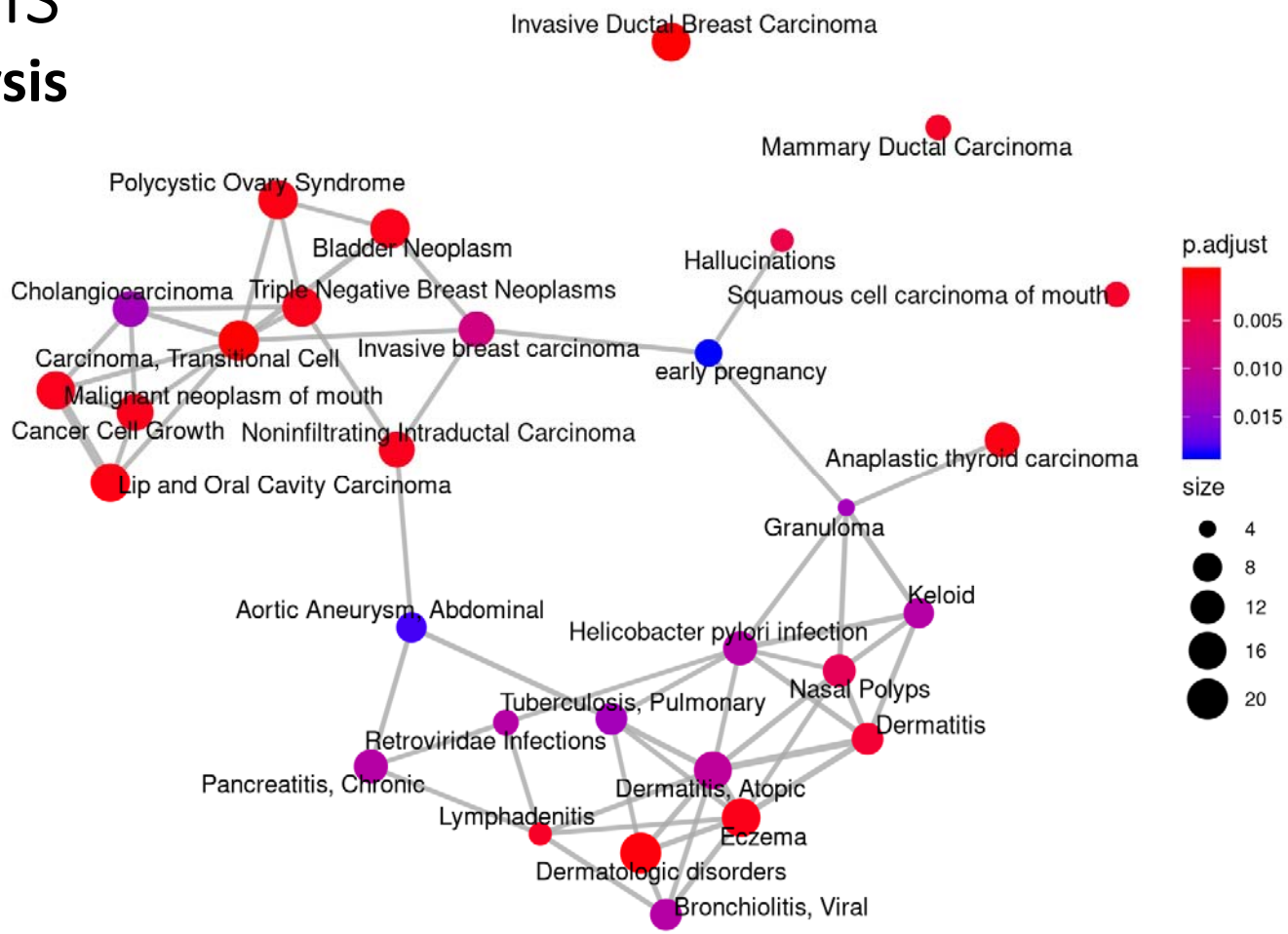
Functional analysis

Over-representation analysis

GO enrichment with clusterProfiler

enrichment map (emapplot): organizes enriched terms into a network with edges connecting overlapping gene sets

mutually overlapping gene sets tend to cluster together, making it easy to identify functional module

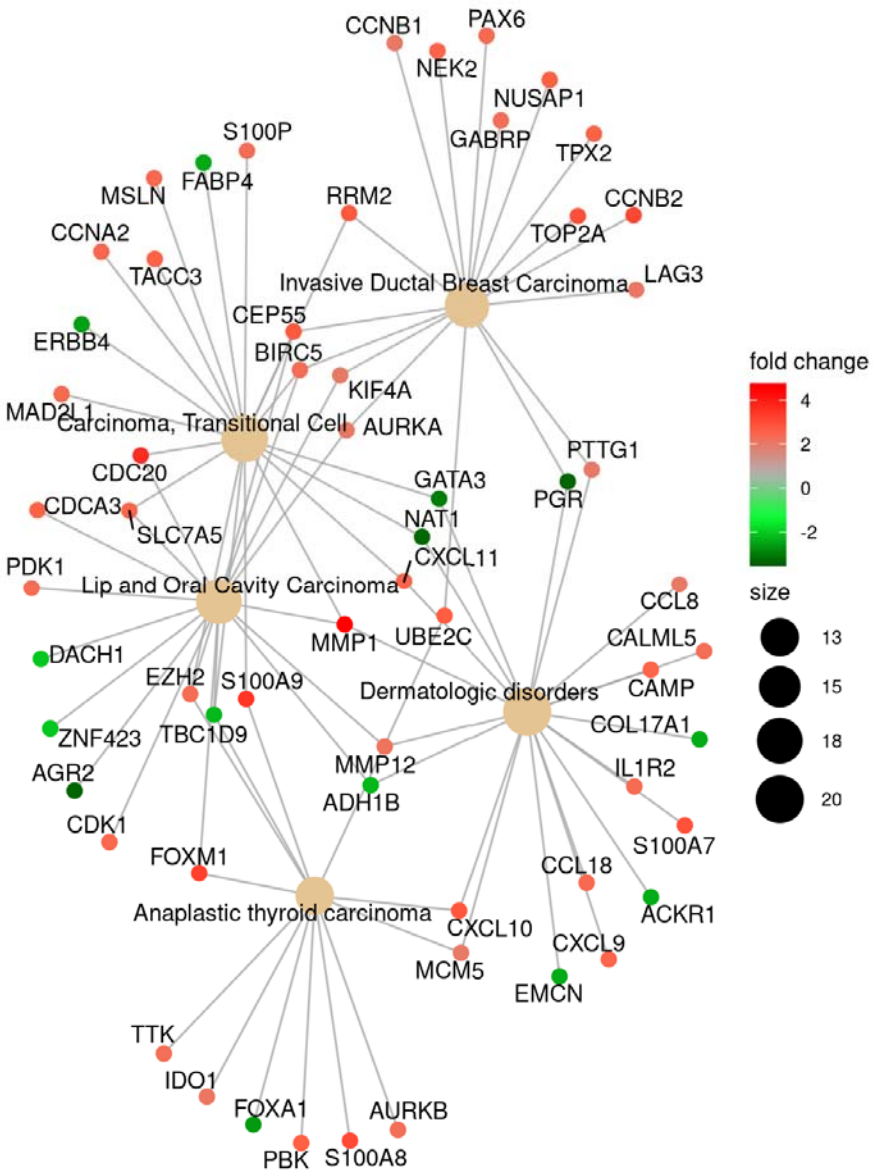


Functional analysis

Over-representation analysis

GO enrichment with clusterProfiler

Gene-Concept Network (cnetplot):
network with genes and GO terms
or KEGG pathways

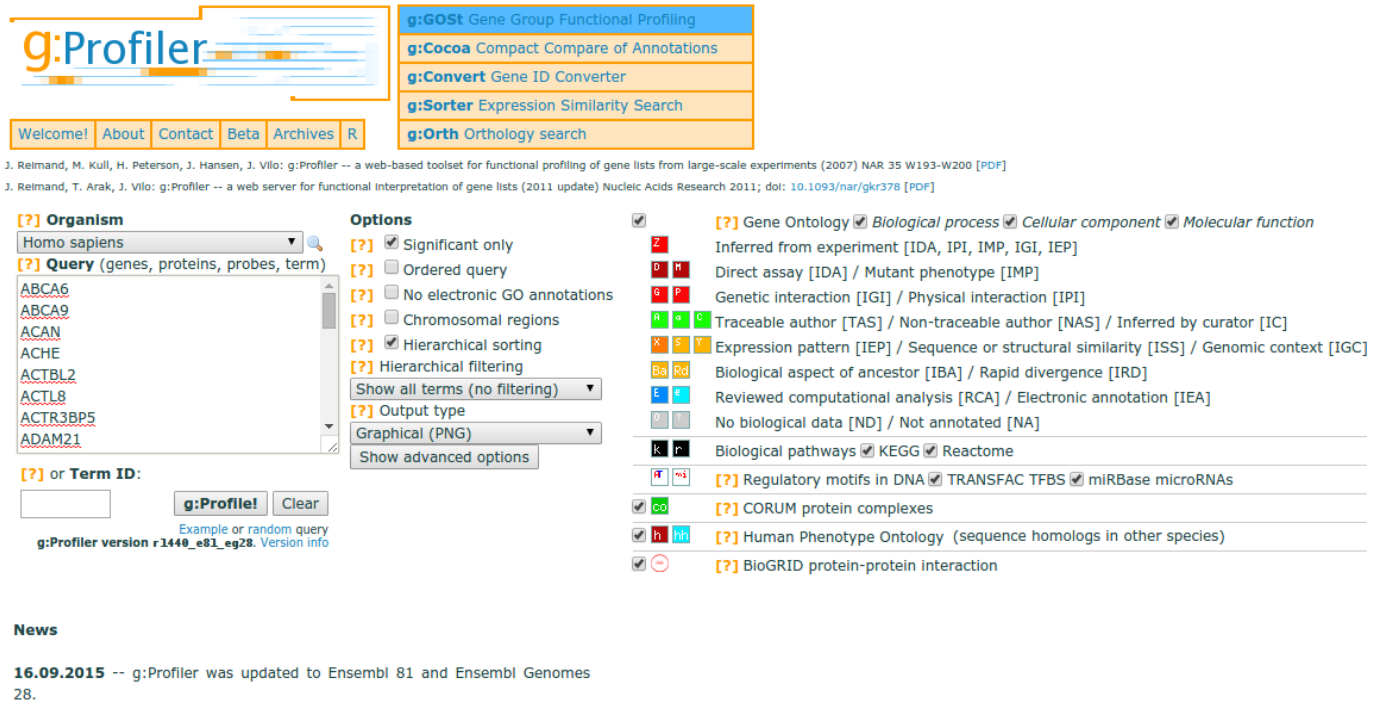


Functional analysis

Over-representation analysis

gProfileR

- Another tool for performing ORA
- considers multiple sources of functional evidence: Gene Ontology terms, biological pathways, regulatory motifs of transcription factors and microRNAs, human disease annotations, protein-protein interactions

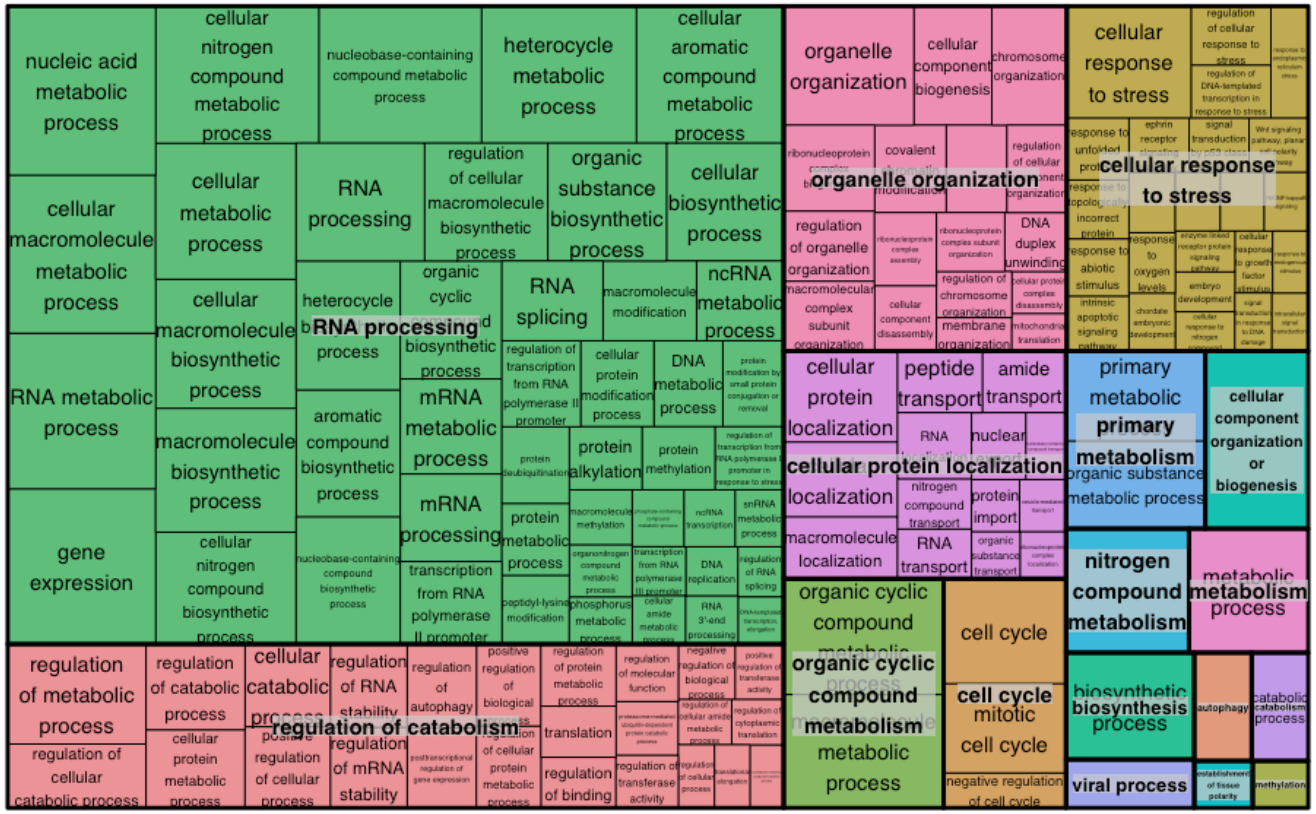


Functional analysis

REVIGO

- web-based tool that can take a list of GO terms, collapse redundant terms by semantic similarity, and summarize them graphically

REVIGO Gene Ontology treemap



Functional analysis

Alternative to REVIGO : GO-Figure!

•Article : https://www.biorxiv.org/content/10.1101/2020.12.02.408534v1

•Github : https://gitlab.com/evogenlab/GO-Figure

Summary Visualisations of Gene Ontology Terms with GO-Figure!

Maarten JMF Reijnders, Robert M Waterhouse

doi: https://doi.org/10.1101/2020.12.02.408534

Abstract Full Text Info/History Metrics Preview PDF

Abstract

The Gene Ontology (GO) is a cornerstone of functional genomics research that drives discoveries through knowledge-informed computational analysis of biological data from large- scale assays. Key to this success is how the GO can be used to support hypotheses or conclusions about the biology or evolution of a study system by identifying annotated functions that are overrepresented in subsets of genes of interest. Graphical visualisations of such GO term enrichment results are critical to aid interpretation and avoid biases by presenting researchers with intuitive visual data summaries. Amongst current visualisation tools and resources there is a lack of standalone open-source software solutions that facilitate systematic comparisons of multiple lists of GO terms. To address this we developed GO-Figure!, an open-source Python software for producing user-customisable semantic similarity scatterplots of redundancy-reduced GO term lists. The lists are simplified by grouping together GO terms with similar functions using their quantified information contents and semantic similarities, with user-control over grouping thresholds. Representatives are then selected for plotting in two-dimensional semantic space where similar GO terms are placed closer to each other on the scatterplot, with an array of user-customisable graphical attributes. GO-Figure! offers a simple solution for command-line plotting of informative summary visualisations of lists of GO terms, designed to support exploratory data analyses and multiple dataset comparisons.

Functional analysis

Functional class scoring tools

- use the gene-level statistics from the DEA
- see whether gene sets for particular biological pathways are enriched among the large positive or negative fold changes
- example: GSEA
- hypotheses:
  - **large changes in individual genes can have significant effects on pathways, and will be detected via ORA methods**
  - **weaker but coordinated changes in sets of functionally related genes can also have significant effects, and will be detected with FCS methods**

Functional analysis

Functional class scoring tools

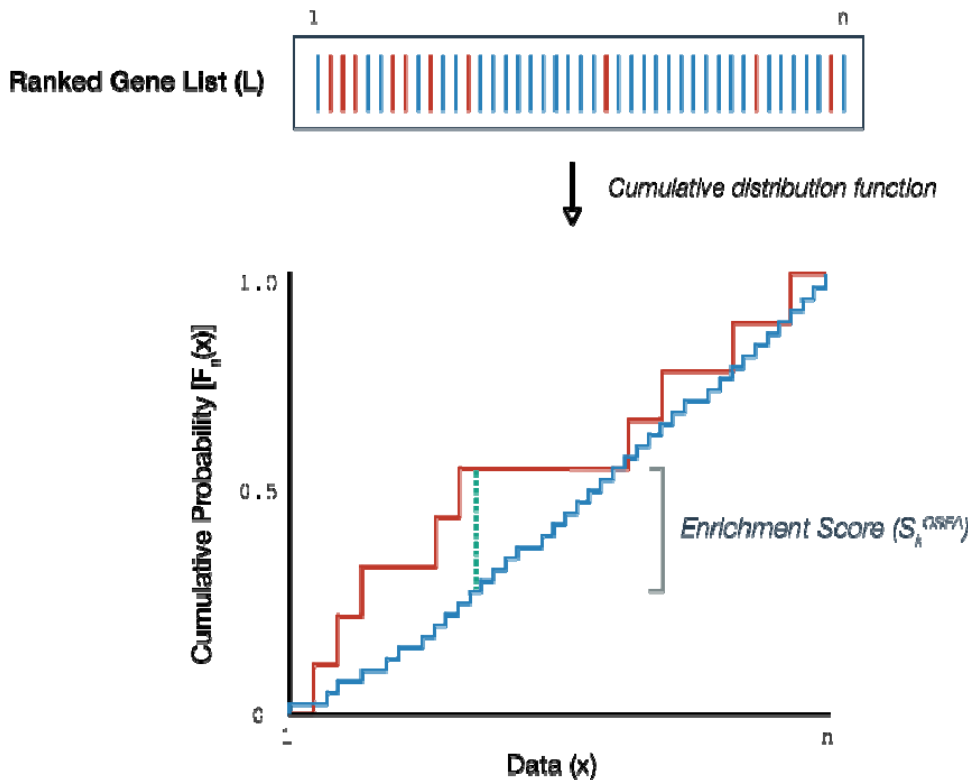
- rather than setting an arbitrary threshold to identify 'significant genes', **all genes are considered in the analysis**
- aggregation of gene-level statistics to generate **a single pathway-level statistic** (+ significance)
- particularly helpful if the differential expression analysis only outputs a small list of significant DE genes

Functional analysis

Functional class scoring tools

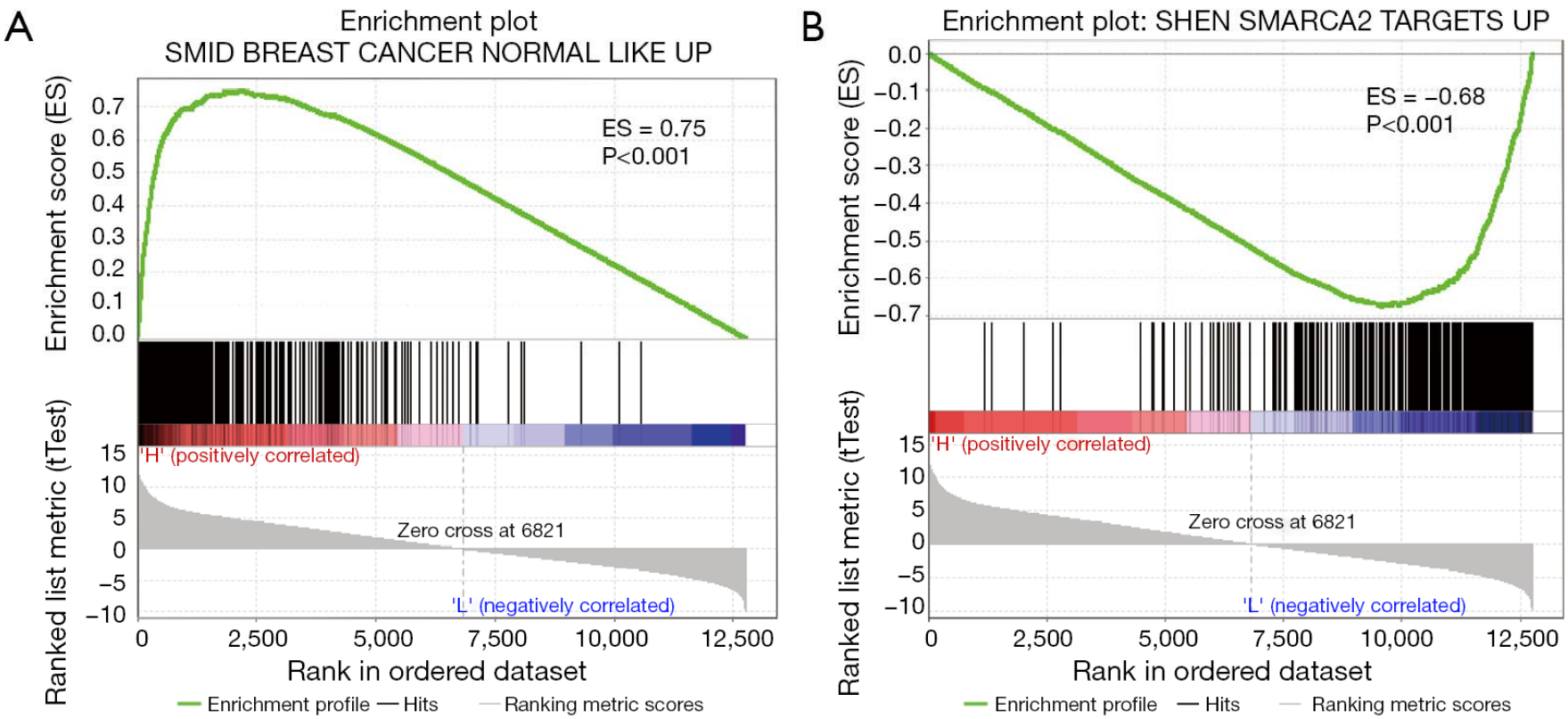
GSEA

- Goal: determine whether the members of a gene set *S* are randomly distributed throughout the ranked gene list (*L*) or primarily found at the top or bottom
- Enrichment score: calculated by walking down the list *L*, increasing a running-sum statistic when we encounter a gene in *S* and decreasing when it is not
- p-value: estimated by permutations



Functional analysis

Functional class scoring tools



Functional analysis

Pathway topology tools

- identification of dysregulated pathways: taking into account gene interaction information + fold changes and adjusted p-values from DEA
- example: SPIA (Signaling Pathway Impact Analysis)

KEGG pathway	P_{NDE}	P_{PERT}	P_G	P_{FDR}	P_{FWER}	Status
Focal adhe..4510	0.0001	0.0000	0.0000	0.00000	0.00000	Act.
ECM-recept..4512	0.0001	0.0004	0.0000	0.00001	0.00002	Act.
PPAR signa..3320	0.0000	0.1240	0.0000	0.00011	0.00034	Inh.
Alzheimers..5010	0.0000	0.7260	0.0001	0.00059	0.00235	Act.
Adherens j..4520	0.0001	0.0852	0.0001	0.00090	0.00452	Act.
Axon guida..4360	0.0002	0.2324	0.0006	0.00487	0.02922	Act.
MAPK signa..4010	0.0001	0.7112	0.0007	0.00504	0.03527	Inh.
Tight junc..4530	0.0007	0.5156	0.0032	0.02073	0.16585	Act.

$P_{NDE} = P(X \geq N_{DE} \mid H_0)$

P_{PERT}: probability to observe a larger perturbation than observed

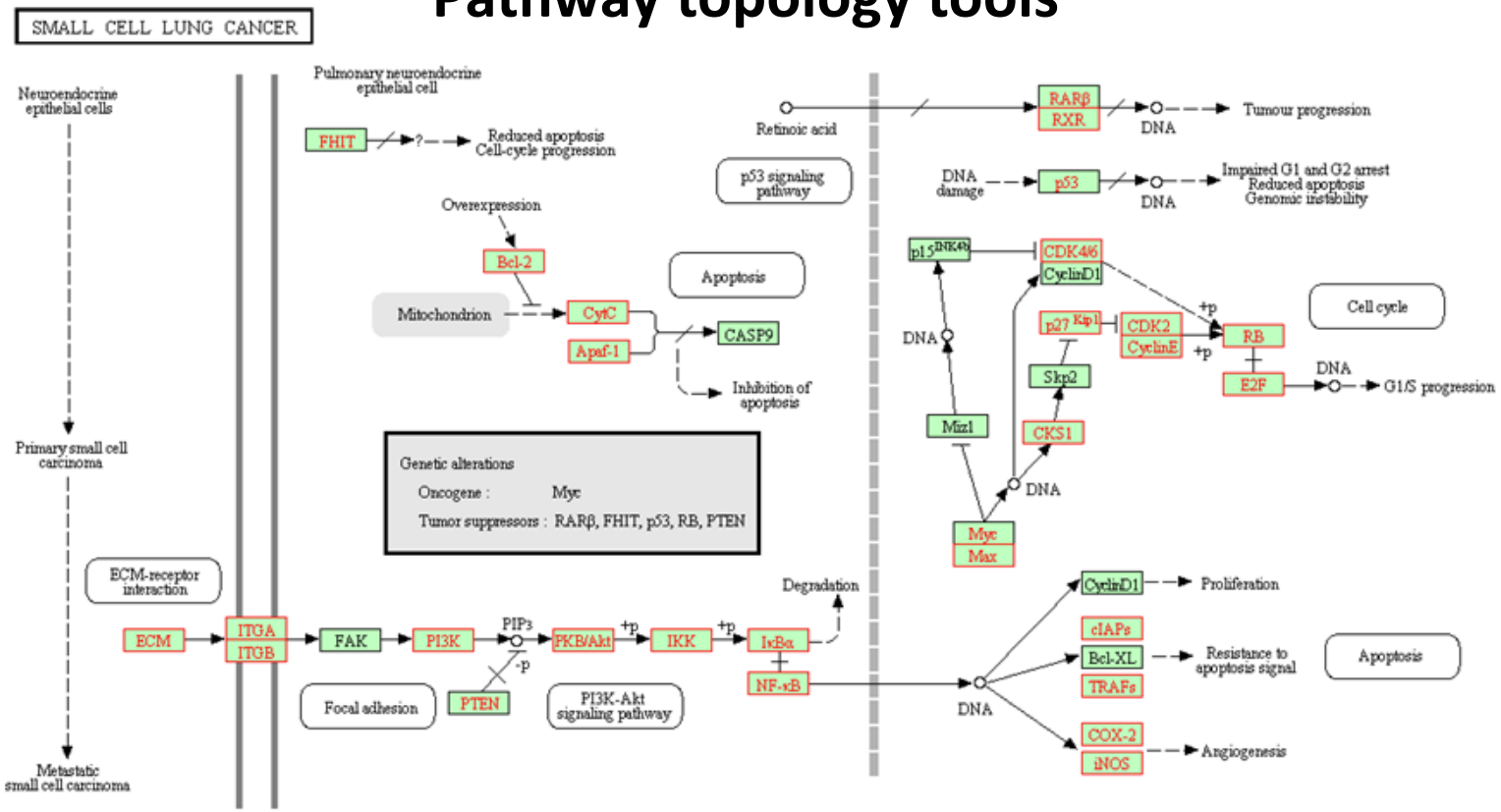
P_G: combination of P_{NDE} and P_{PERT}

P_{FDR}: adjusted FDR p-value

P_{FWER}: adjusted FDR p-value (more conservative)

Functional analysis

Pathway topology tools



Resources for functional analysis

- g:Profiler - http://biit.cs.ut.ee/gprofiler/index.cgi
- DAVID - http://david.abcc.ncifcrf.gov/tools.jsp
- clusterProfiler - http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
- GeneMANIA - http://www.genemania.org/
- GenePattern - http://www.broadinstitute.org/cancer/software/genepattern/ (need to register)
- WebGestalt - http://bioinfo.vanderbilt.edu/webgestalt/ (need to register)
- AmiGO - http://amigo.geneontology.org/amigo
- ReviGO (visualizing GO analysis, input is GO terms) - http://revigo.irb.hr/
- WGCNA - http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork
- GSEA - http://software.broadinstitute.org/gsea/index.jsp
- SPIA - https://www.bioconductor.org/packages/release/bioc/html/SPIA.html
- GAGE/Pathview - http://www.bioconductor.org/packages/release/bioc/html/gage.html