

An Introduction to Pathway Enrichment Analysis

Alex Sánchez



*Statistics and Bioinformatics Unit
Vall d'Hebron Institut de Recerca*



*Statistics and Bioinformatics Research Group
Statistics department, Universitat de Barcelona*



Outline

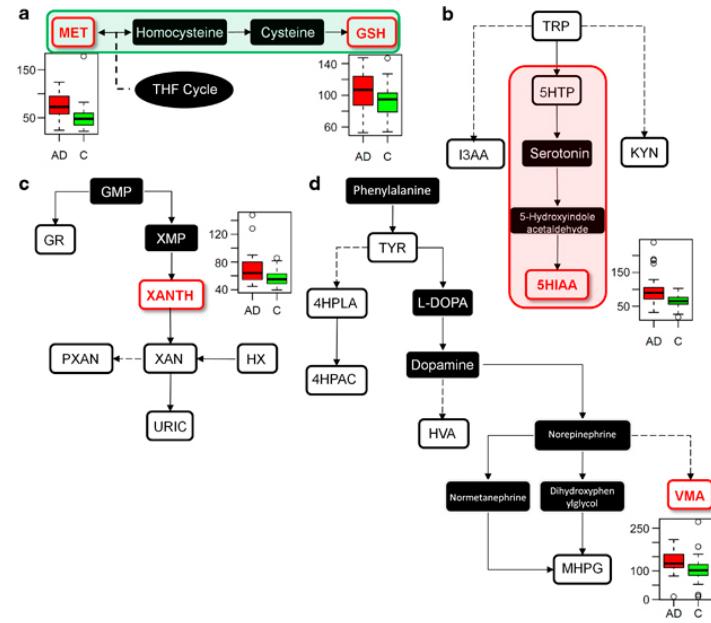
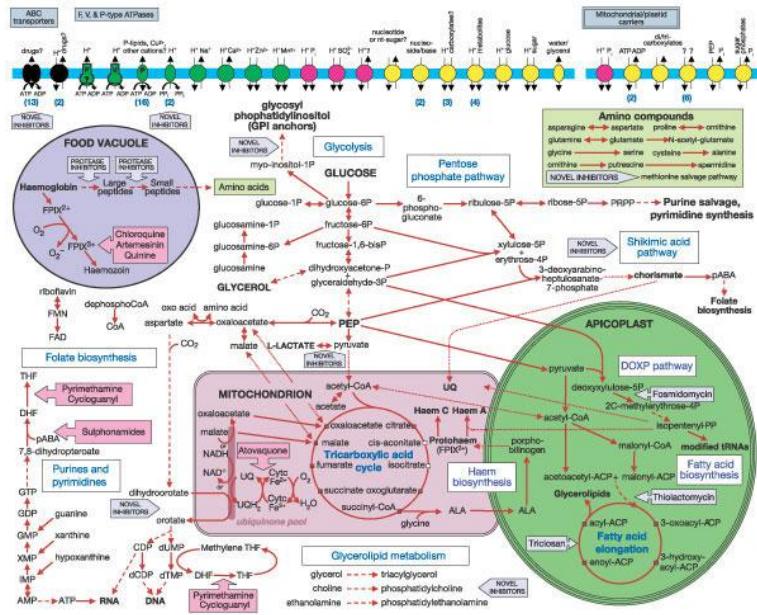
- Presentation
- Introduction and Background
 - Gene lists, Identifiers and Pathway databases
- Pathway Analysis: Methods and Tools
 - Overrepresentation analysis and GSEA
 - Multiple Testing Adjustments
 - Network Visualization and Enrichment Map
- A protocol for Pathway Enrichment Analysis
- A user experience

Introduction & Background

Health, disease and pathways

Metabolism is a complex network of chemical reactions within the confines of a cell that can be analyzed in self-contained parts called ***pathways***

One can generally assume that “normal” metabolism is what happens in healthy state or, reciprocally, that disease can *be associated with some type of alteration in metabolism*.



Pathways altered in ALZHEIMER disease

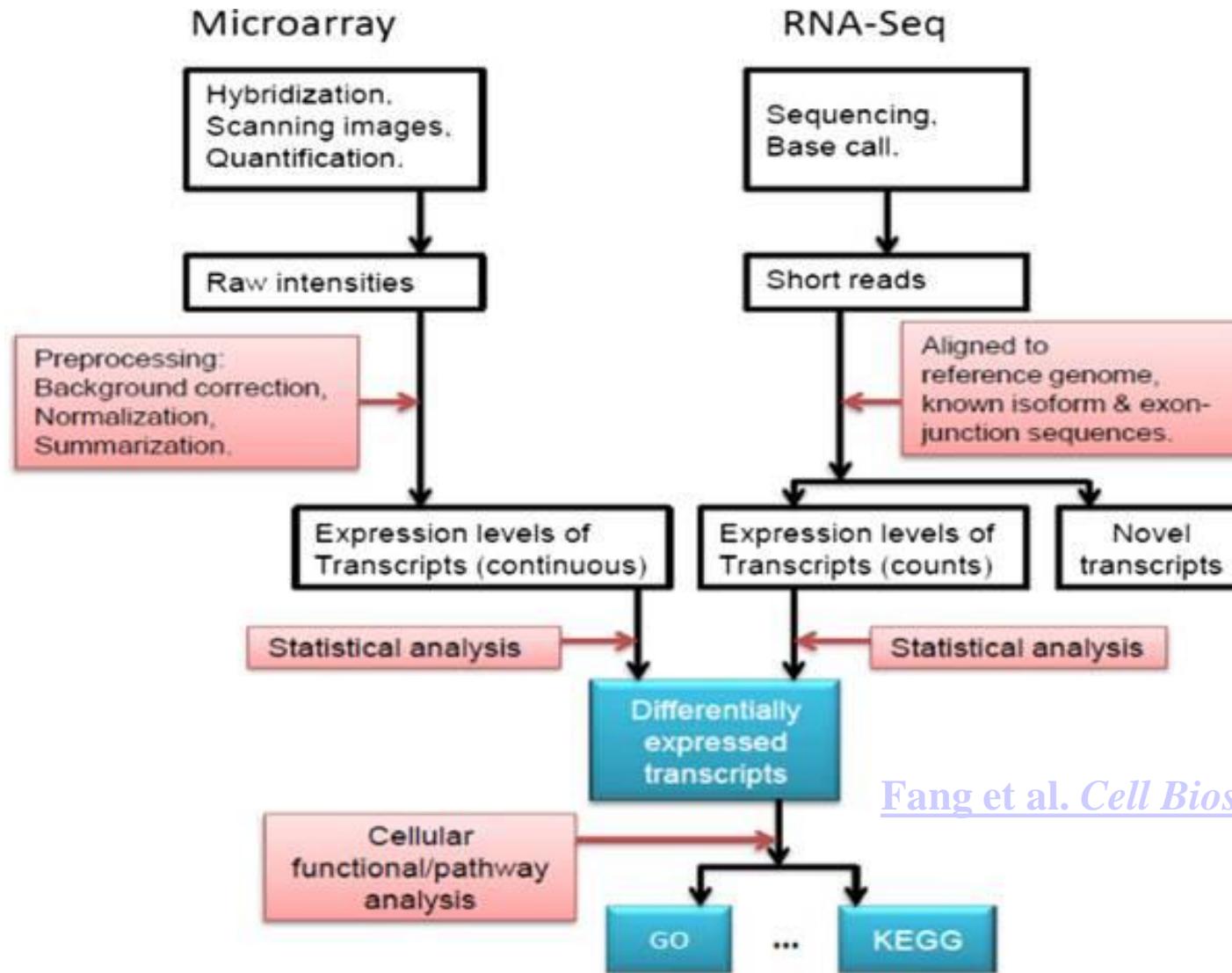
Characterization of disease can be attempted by studying how this affects or disrupts pathways
That's what Pathway Analysis is about (more or less)

Pathway Analysis

- The term Pathway Analysis denotes *any analytic technique that benefits from biological pathway or molecular network information to gain insight into a biological system.* (Creixell et al., Nature Methods 2015 (12 (7))
- To be more specific, Pathway Analysis methods rely on high throughput information provided by omics technologies to:
 - Contextualize findings to help understand the mechanism of disease
 - Identify genes/proteins associated with the aetiology of a disease
 - Predict drug targets
 - Understand how to therapeutically intervene in disease processes
 - Conduct target literature searches
 - Integrate diverse biological information

The beginning: *Gene Lists*

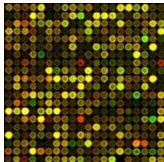
The life-cycle of an omics-based study



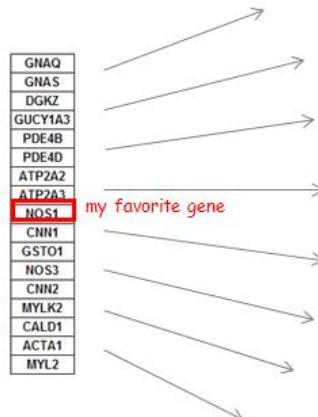
Fang et al. *Cell Biosci.* 2012; 2: 26.

The (in)famous “*where to now?*” question

- You obtained a list of features. What's next?
 - Select some genes for validation?
 - Follow up experiments on some genes/proteins/...?
 - Publish a huge table with all results?
 - Try to learn about **all** features in the list?



GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2



NCBI Resources How To

PubMed.gov US National Library of Medicine National Institutes of Health

GNAQ RSS Save search Advanced

Show additional filters Article types Review More ...

Text availability Abstract available Free full text available Full text available

Publication dates 5 years

Display Settings: Summary, 20 per page, Sorted by Recently A

See 225 articles about **GNAQ** gene function
See also: **GNAQ** guanine nucleotide binding protein (G protein), c
gnaq in *Homo sapiens* | *Mus musculus* | *Rattus norvegicus* | All

Results: 1 to 20 of 114

[Sturge-Weber Syndrome and Port-Wine Stains Caused b](#)

1. Shirley MD, Tang H, Gallione CJ, Baugher JD, Frelin LP, AM, Pevsner J.
N Engl J Med. 2013 May 8. [Epub ahead of print]

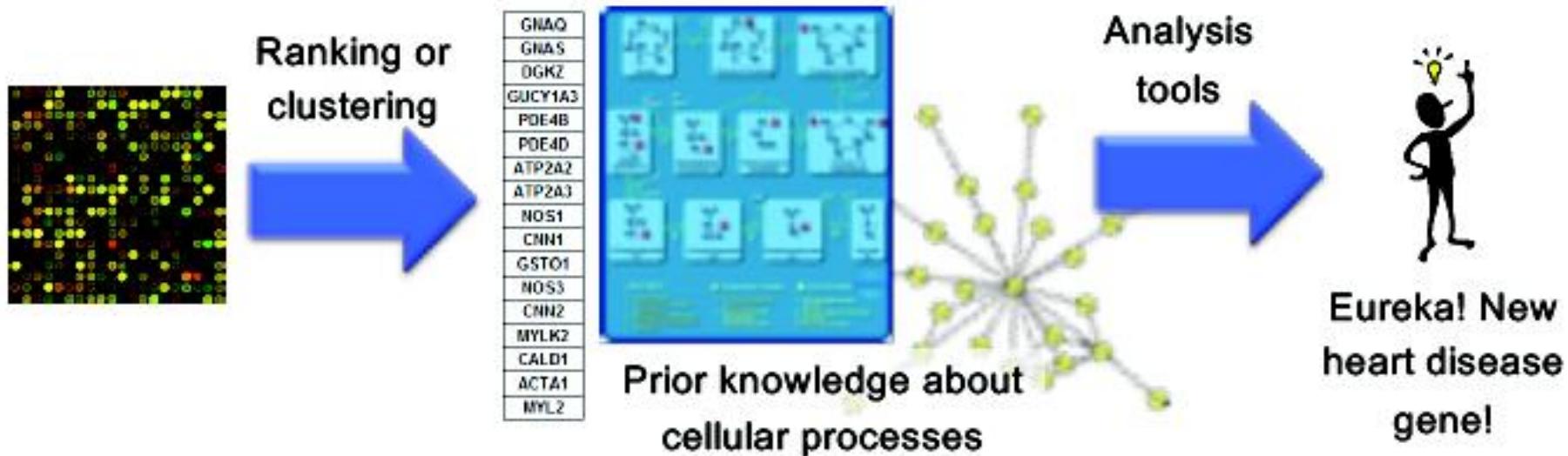
PubMed - as supplied by publisher

From gene lists to *Pathway Analysis*

- Gene lists are made of individual genes
 - Information about each gene can be extracted from databases.
 - Generically described as ***Gene Annotation***
- Besides, we may obtain information from the analysis of *gene sets*
 - Genes don't act individually, rather in groups
More ***realistic*** approach
 - There are less gene sets than individual genes
Relatively ***simpler*** to manage.
 - Generically described as ***Pathway Analysis***

Pathway Analysis Wishlist

- Tell me what's interesting about these genes
 - Are they enriched in known pathways, complexes, functions



Example 1

- Genes with frequent somatic SNVs identified in TCGA exome sequencing data of 3,200 tumors of 12 types
- 127 cancer driver genes displaying higher than expected mutation frequencies were detected using the MuSiC software.
- Genes are ranked in decreasing order of significance and mutation frequency

TP53
PIK3CA
PTEN
APC
VHL
KRAS
MLL3
MLL2
ARID1A
PBRM1
NAV3
EGFR
NF1
PIK3R1
CDKN2A
GATA3
RB1
NOTCH1
FBXW7
CTNNB1
DNMT3A
MAP3K1
FLT3
MALAT1
TSHZ3

Example 2

- Second example is a ranked list of genes obtained from TCGA ovarian cancer dataset.
- Two subgroups - immunoreactive and mesenchymal- were compared.
- The list contains **all genes, not only differentially expressed**, ranked by the value of statistic.

rank	GeneName	test statistic
1	IGDCC3	35.5553322839225
2	ANTXR1	35.3770766531836
3	AEBP1	33.0690543534961
4	FBN1	32.1199562790897
5	ANGPTL2	31.8605806216522
6	COL16A1	31.7641267462069
7	BGN	31.533826423921
...
15201	IRF1	-14.7629673442493
15202	CXCL10	-14.9827363665643
15203	TAP2	-15.1488606179238
15204	UBE2L6	-15.7162058907796
15205	KIAA0319	-15.7796986548781
15206	PSMB8	-15.7846188665582
15207	PSME1	-16.4510045533584
15208	CSAG3	-16.8014265945244
15209	OVGP1	-17.6903158148446
15210	GBP4	-17.9447602030134
15211	TAP1	-18.0549262210415
15212	PSME2	-18.3639448844986
15213	PSMB9	-18.6614452029879

Gene Lists and Annotations

Gene and Protein Identifiers

- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
 - E.g. Social Insurance Number, Entrez Gene ID 41232
- But, information on features is stored in many databases.
 - Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
 - Important to recognize the correct record type
 - E.g. Entrez Gene records don't store sequence. They link to DNA regions, RNA transcripts and proteins e.g. in RefSeq, which stores sequence.

TP53
PIK3CA
PTEN
APC
VHL
KRAS
MLL3
MLL2
ARID1A
PBRM1
NAV3
EGFR
NF1
PIK3R1
CDKN2A
GATA3
RB1
NOTCH1
FBXW7
CTNNB1
DNMT3A
MAP3K1
FLT3
MALAT1
TSHZ3

Common Identifiers

Gene

Ensembl ENSG00000139618

Entrez Gene 675

Unigene Hs.34012

RNA transcript

GenBank BC026160.1

RefSeq NM_000059

Ensembl ENST00000380152

Protein

Ensembl ENSP00000369497

RefSeq NP_000050.2

UniProt BRCA2_HUMAN or

A1YBP1_HUMAN

IPI IPI00412408.1

EMBL AF309413

PDB 1MIU

Species-specific

HUGO HGNC BRCA2

MGI MGI:109337

RGD 2219

ZFIN ZDB-GENE-060510-3

FlyBase CG9097

WormBase WBGene00002299 or ZK1067.1

SGD S00002187 or YDL029W

Annotations

InterPro IPR015252

OMIM 600185

Pfam PF09104

Gene Ontology GO:0000724

SNPs rs28897757

Experimental Platform

Affymetrix 208368_3p_s_at

Agilent A_23_P99452

CodeLink GE60169

Illumina GI_4502450-S

Red =

Recommended

Identifier Mapping

- There are many IDs!
 - Software tools recognize only a handful
 - May need to map from your gene list IDs to standard IDs
- Four main uses
 - Searching for a favorite gene name
 - Link to related resources
 - Identifier translation
 - E.g. Proteins to genes, Affy ID to Entrez Gene
 - Merging data from different sources
 - Find equivalent records

ID Challenges

- Avoid errors: map IDs correctly
 - Beware of 1-to-many mappings
- Gene name ambiguity – not a good ID
 - e.g. FLJ92943, LFS1, TRP53, p53
 - Better to use the standard gene symbol: TP53
- Excel error-introduction
 - OCT4 is changed to October-4 (paste as text)
- Problems reaching 100% coverage
 - E.g. due to version issues
 - Use multiple sources to increase coverage

Zeeberg BR et al. *Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics*
BMC Bioinformatics. 2004 Jun 23;5:80

Use ID converters to prepare list

DAVID Bioinformatics Resources 2007
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Gene Accession Conversion Tool

Summary
The possible choices for ambiguous genes
The possible choices for each individual ambiguous genes

Conversion Summary
ID Count In DAVID DB Conversion
157 IDs Yes Successful
0 IDs Yes None
0 IDs No NA
1 IDs Ambiguous Pending
Total Unique User IDs: 166

Genes that have been converted. Right-click to Download the list Help Save the results Submit the converted genes to DAVID for other analytical tools!!

From To Species David Gene Name

*1112_G_AT 5684 HOMO SAPIENS NEURAL CELL ADHESION MOLECULE 1
*1331_S_AT 8718 HOMO SAPIENS TUMOR NECROSIS FACTOR RECEPTOR SUPERFAMILY, MEMBER 25
*1355_G_AT 4915 HOMO SAPIENS NEUROTROPHIC TYROSINE KINASE, RECEPTOR, TYPE 2
*1372_AT 7120 HOMO SAPIENS TUMOR NECROSIS FACTOR, ALPHA-INDUCED PROTEIN 6
*1391_S_AT 1572 HOMO SAPIENS CYTOCHROME P450, FAMILY 4, SUBFAMILY A, POLYPEPTIDE 11
*1403_S_AT 6322 HOMO SAPIENS CHEMOKINE (C-C MOTIF) LIGAND 5
*1419_G_AT 4843 HOMO SAPIENS NITRIC OXIDE SYNTHASE 2A (INDUCIBLE, HEPATOCYTES)
*1575_AT 5243 HOMO SAPIENS ATP-BINDING CASSETTE, SUB-FAMILY B (MDR/TAP), MEMBER 1
*1645_AT 3814 HOMO SAPIENS KISS-1 METASTASIS-SUPPRESSOR
*1786_AT 10461 HOMO SAPIENS C-MER PROTO-ONCOGENE TYROSINE KINASE
*1855_AT 2248 HOMO SAPIENS FIBROBLAST GROWTH FACTOR 3 (MURINE MAMMARY TUMOR VIRUS INTEGRATION SITE 1V-INT-2...)
*1890_AT 9518 HOMO SAPIENS GROWTH DIFFERENTIATION FACTOR 15

Species of converted gene IDs
Converted gene IDs
Users' input gene IDs
*Users' decision for ambiguous IDs

g:Profiler

Welcome! Contact FAQ R / APIs Beta Archive

g:GOST Gene Group Functional Profiling
g:Cocoa Compact Compare of Annotations
g:Convert Gene ID Converter
g:Sorter Expression Similarity Search
g:Orth Orthology search
g:SNPense Convert rsID

[?] Organism: Homo sapiens
[?] Target database: ENSG
[?] Output type: Table (HTML)
Convert IDs Clear

[?] Query (genes, proteins, probes, term)
[?] Interpret query as chromosome
[?] Numeric IDs treated as
AFFY_HUEX_1_0_ST_V2

Example 1: Gene ID conversion with g:Profiler

ID Mapping Services

Input gene/protein/transcript IDs (mixed)

Type of output ID

g#	initial alias >> g:GOST >> g:Sorter >> g:Orth >> g:Cocoa	c#	converted alias >> g:GOST >> g:Sorter >> g:Orth >> g:Cocoa >> Copy values	name >> g:GOST >> g:Sorter >> g:Orth >> g:Cocoa >> Copy values	description	namespace
1	TP53	1.1	P04637	TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]	UNIPROT_GN, ENTREZGENE, VEGA_GENE, DBASS5, DBASS3, HGNC, WIKIGENE
2	MDM2	2.1	Q00987	MDM2	MDM2 proto-oncogene, E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:HGNC:6973]	UNIPROT_GN, ENTREZGENE, VEGA_GENE, HGNC, WIKIGENE
3	207105_S_AT	3.1	O00459	PIK3R2	phosphoinositide-3-kinase, regulatory subunit 2 (beta) [Source:HGNC Symbol;Acc:HGNC:8980]	AFFY_HG_U133_PLUS_2, AFFY_HG_FOCUS, AFFY_HG_U133A_2, AFFY_HG_U133A
4	P60484	4.1	P60484	PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]	UNIPROTSWISSPROT

AFFY_HG_U95C
AFFY_HG_U95D
AFFY_HG_U95E
AFFYHTA_2_0
AFFY_HUEX_1_0_ST_V2
AFFY_HUGENEFL
AFFY_HUGENE_1_0_ST_V1
AFFY_HUGENE_2_0_ST_V1
AFFY_PRIMEVIEW
AFFY_U133_X3P
AGILENT_CGH_44B
AGILENT_SUREPRINT_G3_GE_8X60K
AGILENT_SUREPRINT_G3_GE_8X60K_V2
AGILENT_WHOLEGENOME_4X44K_V1
AGILENT_WHOLEGENOME_4X44K_V2
ARRAYEXPRESS
CCDS
CCDS_ACC
CHEMBL
CLONE_BASED_ENSEMBL_GENE
CLONE_BASED_ENSEMBL_TRANSCRIPT
CLONE_BASED_VEGA_GENE
CLONE_BASED_VEGA_TRANSCRIPT
CODELINK_CODELINK
DBASS3
DBASS3_ACC
DBASS5
DBASS5_ACC
EMBL
ENSG
ENSP
ENST
ENS_HS_TRANSCRIPT
ENS_HS_TRANSLATION
ENS_LRG_GENE
ENS_LRG_TRANSCRIPT
ENTREZGENE
ENTREZGENE_ACC
ENTREZGENE_TRANS_NAME
GO
GOSLIM_GOA
HGNC
HGNC_ACC
HGNC_TRANS_NAME
HPA
HPA_ACC
ILLUMINA_HUMANHT_12_V3
ILLUMINA_HUMANHT_12_V4
ILLUMINA_HUMANREF_8_V3
ILLUMINA_HUMANWG_6_V1
ILLUMINA_HUMANWG_6_V2
ILLUMINA_HUMANWG_6_V3
MEROPS
MIM_GENE
MIM_GENE_ACC
MIM_MORBID
MIM_MORBID_ACC
MIRBASE
MIRBASE_ACC
MIRBASE_TRANS_NAME
OTTG
OTTP
OTTT
PDB
PHALANX_ONEARAY
PROTEIN_ID
PROTEIN_ID_ACC
REFSEQ_MRNA
REFSEQ_MRNA_ACC
REFSEQ_MRNA_PREDICTED
REFSEQ_MRNA_PREDICTED_ACC
REFSEQ_MRNA_NORMA

- **g:Convert**
- <http://biit.cs.ut.ee/gprofiler/gconvert.cgi>

- **Ensembl Biomart**
- <http://www.ensembl.org>

Beware of ambiguous ID mappings

g:Profiler

Welcome! About Contact Beta Archives ▾ R

g:GOSet Gene Group Functional Profiling
g:Cocoa Compact Compare of Annotations
g:Convert Gene ID Converter
g:Sorter Expression Similarity Search
g:Orth Orthology search

J. Reimand, M. Kull, H. Peterson, J. Hansen, J. Vilo: g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments (2007) NAR 35 W193-W200 [PDF]
J. Reimand, T. Arak, J. Vilo: g:Profiler -- a web server for functional interpretation of gene lists (2011 update) Nucleic Acids Research 2011; doi: 10.1093/nar/gkr378 [PDF]

[?] Organism
Homo sapiens

[?] Query (genes, proteins, probes, term)
TP53 MDM2 207105_S_AT P60484

Options

Significant only
 Ordered query
 No electronic GO annotations
 Chromosomal regions
 Hierarchical sorting
 Hierarchical filtering
Show all terms (no filtering)
[?] Output type
Graphical (PNG)
Show advanced options

[?] Gene Ontology Biological process Cellular component Molecular function
Inferred from experiment [IDA, IPI, IMP, IGI, IEP]
[?] Direct assay [IDA] / Mutant phenotype [IMP]
[?] Genetic interaction [IGI] / Physical interaction [IPI]
[?] Traceable author [TAS] / Non-traceable author [NAS] / Inferred by curator [IC]
[?] Expression pattern [IEP] / Sequence or structural similarity [ISS] / Genomic context [IGC]
[?] Biological aspect of ancestor [IBA] / Rapid divergence [IRD]
[?] Reviewed computational analysis [RCA] / Electronic annotation [IEA]
[?] No biological data [ND] / Not annotated [NA]
[?] Biological pathways KEGG Reactome
[?] Regulatory motifs in DNA TRANSFAC TFBS miRBase microRNAs
[?] CORUM protein complexes
[?] Human Phenotype Ontology (sequence homologs in other species)
[?] BioGRID protein-protein interaction

>> g:Convert Gene ID Converter **>> g:Orth** Orthology Search **>> g:Sorter** Expression Similarity Search **>> g:Cocoa** Compact Compare of Annotations **>> Static URL** Come back later

Warning: Some gene identifiers are ambiguous. Resolve these manually?

Attempt to automatically resolve symbols using a namespace (percentage of ambiguous symbols resolved in brackets):

207105_S_AT

ENSG00000268173 (PIK3R2, 26 GO annot.) - phosphoinositide-3-kinase, regulatory subunit 2 (beta) [Source:HGNC Symbol;Acc:HGNC:8980]
 ENSG00000105647 (PIK3R2, 26 GO annot.) - phosphoinositide-3-kinase, regulatory subunit 2 (beta) [Source:HGNC Symbol;Acc:HGNC:8980]

Recommendations

- For proteins and genes
 - (doesn't consider splice forms)
 - Map everything to Entrez Gene IDs or Official Gene Symbols using an appropriate tool, such as gProfiler, DAVID or Biomart.
- If 100% coverage desired, manually curate missing mappings using multiple resources
- Be careful of Excel auto conversions – especially when pasting large gene lists!
 - Remember to format cells as ‘text’ before pasting

Pathway and Gene Sets databases

Where is pathway information? (1)

- Most common sources*
 - Gene Ontology: Biological process,
 - Pathway databases:
 - Reactome : <http://reactome.org>
 - <http://www.pathguide.org>
 - MSigDB:
<http://www.broadinstitute.org/gsea/msigdb/>
 - <http://www.pathwaycommons.org/>

*[Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges](#)

Where is pathway information? (2)

- Other annotations
 - Gene Ontology molecular function, cell location
 - Chromosome position
 - Disease association
 - DNA properties (TF binding sites, gene structure (intron/exon), SNPs, ...)
 - Transcript properties (Splicing, 3' UTR, microRNA binding sites, ...)
 - Protein properties (Domains, 2ry and 3ry structure, PTM sites)
 - Interactions with other genes

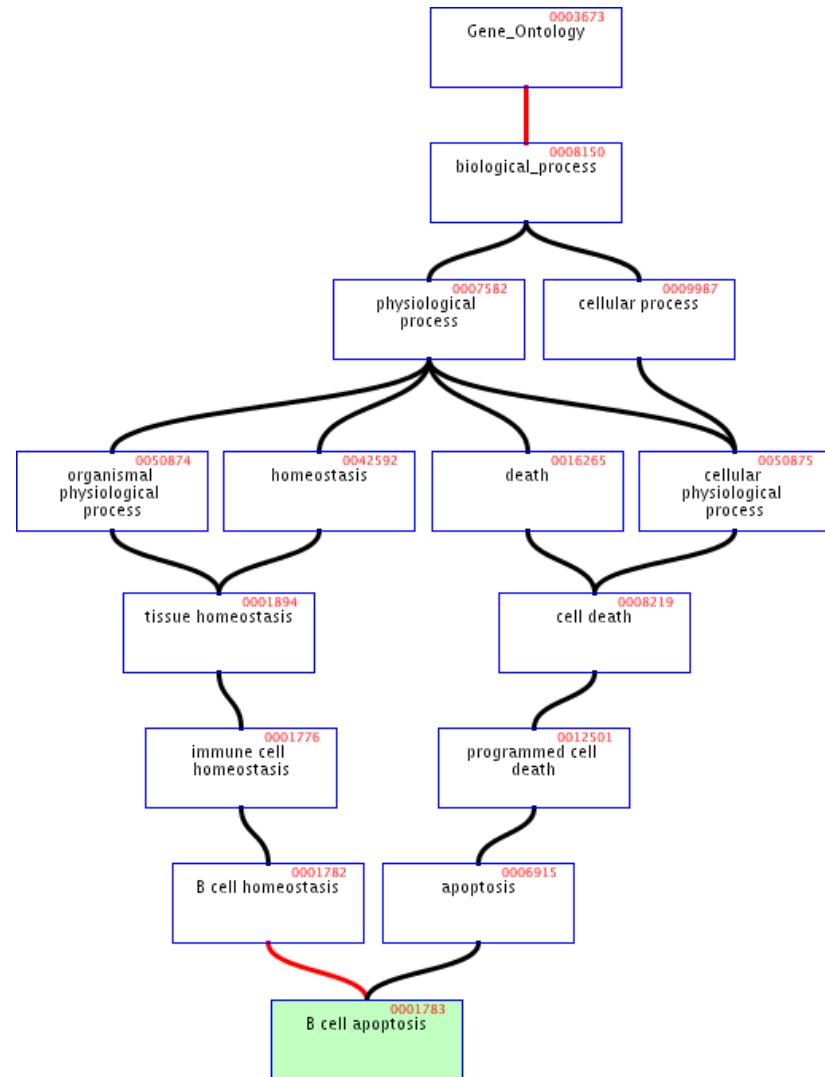
*Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges

What is the Gene Ontology (GO)?

- Set of biological phrases (terms) which are applied to genes:
 - protein kinase, apoptosis, membrane
- An ontology is not a dictionary
 - Dictionary: A collection of term definitions,
 - Alphabetic organization
 - Ontology: A formal system for describing knowledge
 - Hierarchical organization
- <http://geneontology.org/>

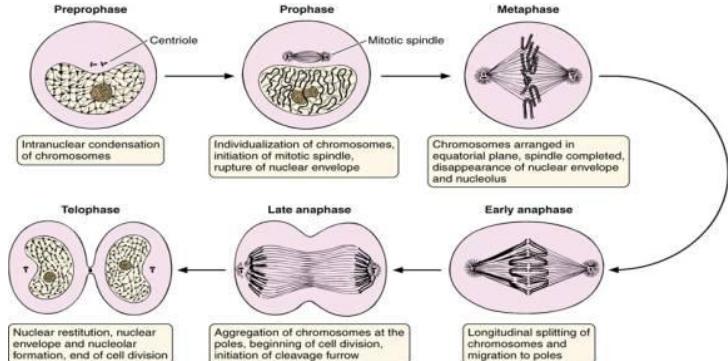
GO Structure

- Terms are related within a hierarchy
 - is-a
 - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child

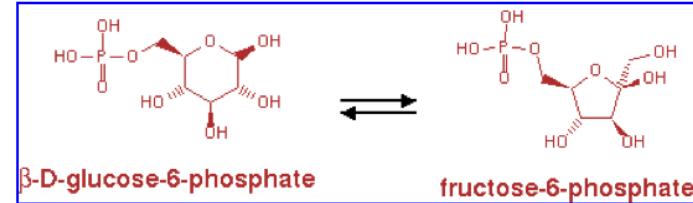
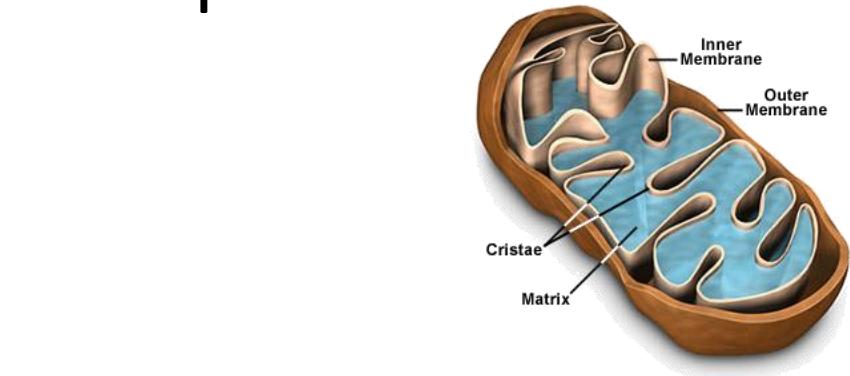


What is covered by the GO?

- GO terms divided into three aspects:
 - cellular component
 - molecular function
 - biological process



**Cell
division**



**glucose-6-phosphate
isomerase activity**

Annotation Sources

- Manual annotation
 - Curated by scientists
 - High quality
 - Small number (time-consuming to create)
 - Reviewed computational analysis
- Electronic annotation
 - Annotation derived without human validation
 - Computational predictions (accuracy varies)
 - Lower ‘quality’ than manual codes
- Key point: be aware of annotation origin

Evidence Types

- Experimental Evidence Codes
 - EXP: Inferred from Experiment
 - IDA: Inferred from Direct Assay
 - IPI: Inferred from Physical Interaction
 - IMP: Inferred from Mutant Phenotype
 - IGI: Inferred from Genetic Interaction
 - IEP: Inferred from Expression Pattern
- Computational Analysis Evidence Codes
 - ISS: Inferred from Sequence or Structural Similarity
 - ISO: Inferred from Sequence Orthology
 - ISA: Inferred from Sequence Alignment
 - ISM: Inferred from Sequence Model
 - IGC: Inferred from Genomic Context
 - RCA: inferred from Reviewed Computational Analysis
- Author Statement Evidence Codes
 - TAS: Traceable Author Statement
 - NAS: Non-traceable Author Statement
- Curator Statement Evidence Codes
 - IC: Inferred by Curator
 - ND: No biological Data available
- IEA: Inferred from electronic annotation

<http://www.geneontology.org/GO.evidence.shtml>

Pathway Analysis

Overrepresentation Analysis

Gene Set Enrichment Analysis

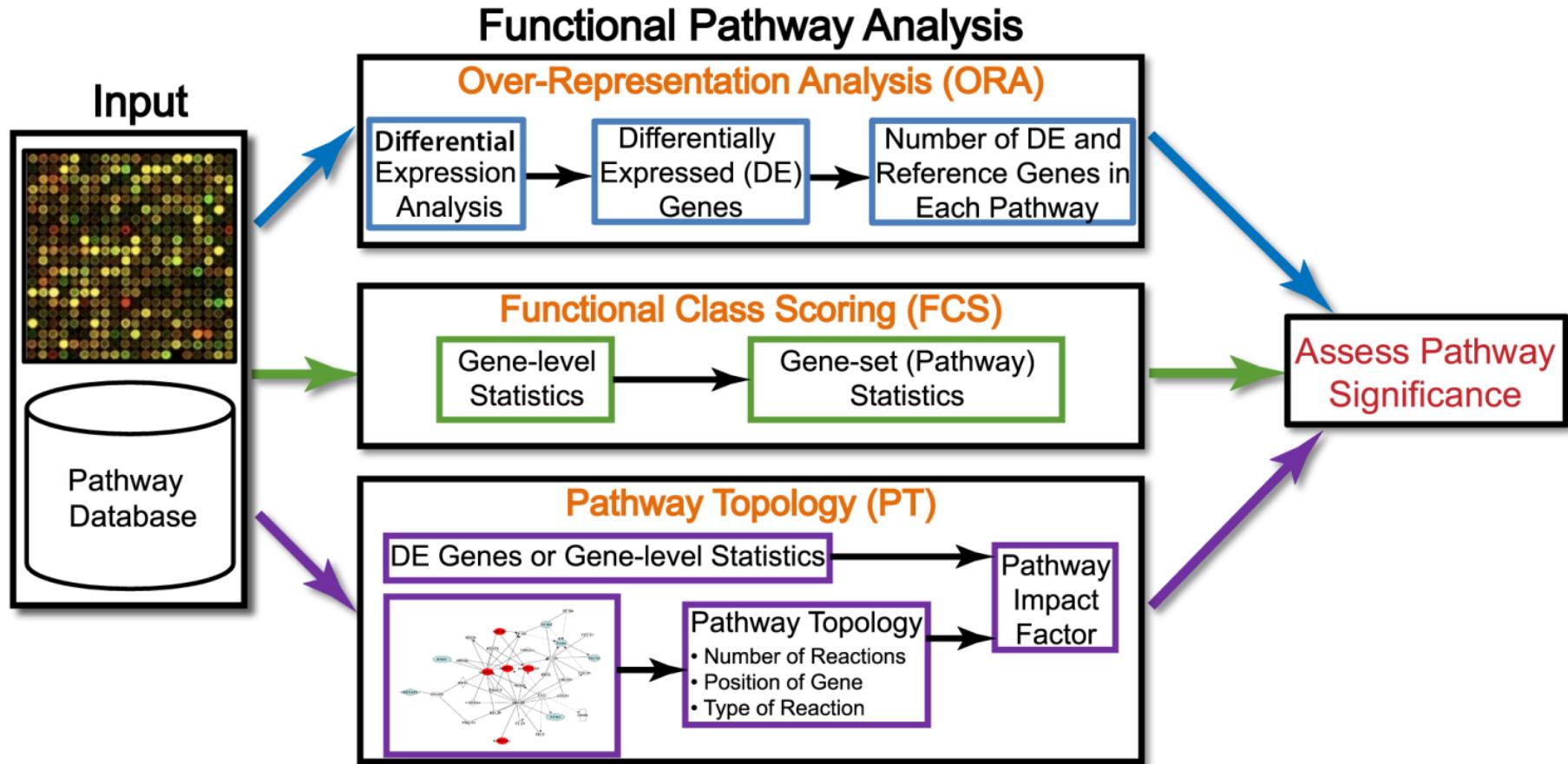
Pathway Analysis

- “*Any type of analysis that involves pathway or information*”
 - Most popular type is ***over-representation analysis***, but many others exist.
- Intended to gain insight into ‘omics’ data. E.g:
 - Identifying a master regulator gene,
 - Finding drug targets,
 - Characterizing pathways active in a sample.

Benefits of Pathway Analysis

- Relatively easy to interpret
 - Familiar concepts e.g. cell cycle
- Identifies possible causal mechanisms
- Predicts new roles for genes
- Improves statistical power
 - Fewer tests, aggregates data from multiple genes into one pathway
- More reproducible
 - E.g. gene expression signatures
- Facilitates integration of multiple data types

Types of Pathway Analysis

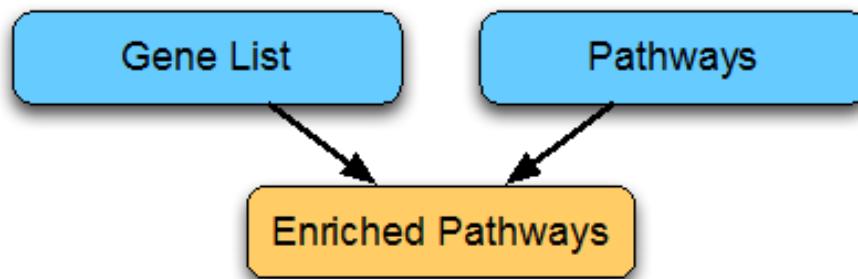


Khatri et alt. 10 years of Pathway Analysis

Analysis of *thresholded* lists
with *Enrichment Analysis*
(also called Overrepresentation A.)

Over-representation analysis

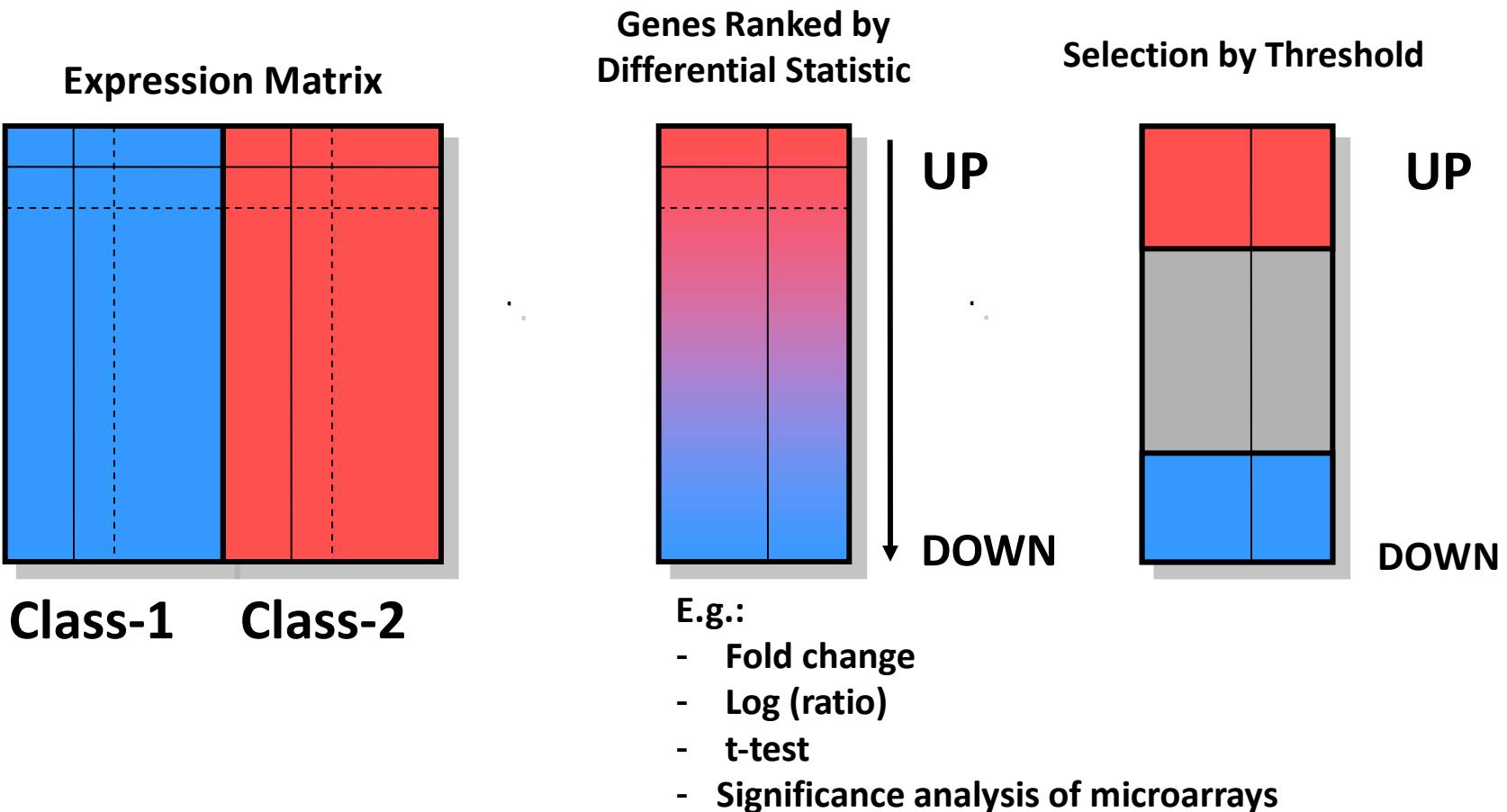
- Combines
 - Gene (feature) lists ← (Gen)omic experiment
 - Pathways and other gene annotations
 - Gene Ontology
 - Reactome
 - Pathway commons



Over-representation analysis

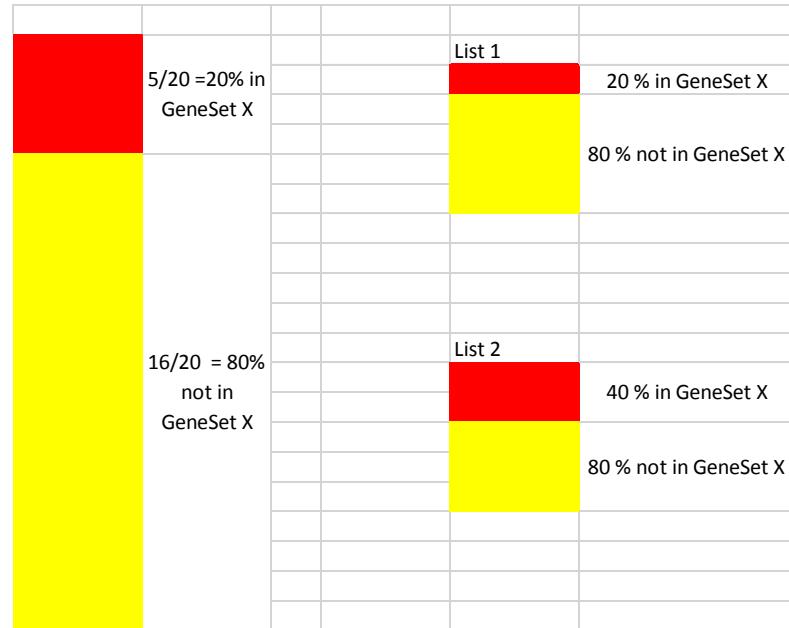
- Given:
 1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast)
 2. Gene sets or annotations: e.g. Gene ontology, transcription factor binding sites in promoter
- Question: *Are any of the gene annotations surprisingly enriched in the gene list?*
- Details:
 1. Where do the gene lists come from?
 2. How to assess “surprisingly” (statistics)
 3. How to adjust for test multiplicity?

Obtaining the gene lists



Assessing “surprisingly”

- Given a gene list, “gl”, and a gene set, “GS”, check:
- Is the % of genes in “gl” annotated in “GS” the same as the % of genes globally annotated in “GS”?
 - If both percentages are similar → *No Enrichment*
 - If the % of genes annotated in “GS” is greater in “gl” than in the rest of genes → “gl” is *enriched in “GS”*



Examples

	Differentially expressed (gl_1)	Not differentially expressed	TOTAL
In Gene Set (GS1)	10	30	40
Not In Gene Set	390	3570	3960
TOTAL	400	3600	4000
% of gl_1 in GS1	$10/400=0.025$	$30/3600=0.00833$	

$0.025 >> 0.00833 \rightarrow "gl_1" \text{ is enriched in "GS}_1\text{"}$

	Differentially expressed (gl_2)	Not differentially expressed	TOTAL
In Gene Set (GS2)	10	30	40
Not In Gene Set	390	1220	1610
TOTAL	400	1500	1650
% of gl_2 in GS ₂	$10/400=0.025$	$30/1500=0.2$	

$0.025 \approx 0.02 \rightarrow \text{Can't say that "gl}_2\text{" is enriched in "GS}_2\text{"}$

Assessing significance: Fisher test

- The examples shows two cases
 - One where percentages are quite different
 - Another where percentages are similar
- How can we set a threshold to decide that the difference is “big enough” to call it “Enriched”
 - Use Fisher Test or, equivalently,
 - a test to compare proportions or
 - a hypergeometric test.

Assessing significance: Fisher test (1)

```
> GOnnnnCounts<- matrix(c(10, 30, 390, 3570),  
+ nrow = 2, byrow=TRUE,  
+ dimnames = list(GeneSet = c("In Gene Set", "Not in Gene Set"),  
+                 Test =c("Differentially expressed", "Not Dif. Expr.")))  
> GOnnnnCounts  
          Test  
GeneSet      Differentially expressed Not Dif. Expr.  
  In Gene Set                      10            30  
  Not in Gene Set                  390           3570  
> fisher.test(GOnnnnCounts, alternative = "greater")  
  
  Fisher's Exact Test for Count Data  
  
data: GOnnnnCounts  
p-value = 0.004836  
alternative hypothesis: true odds ratio is greater than 1  
95 percent confidence interval:  
 1.508343      Inf  
sample estimates:  
odds ratio  
 3.049831
```

P-value small, odds-ratio high → List is *surprisingly* enriched in Gene Set

Assessing significance: Fisher test (2)

```
> G0nnnnCounts<-matrix(c(10,30,390,1220), nrow=2, byrow=TRUE,
+                         dimnames=list(
+                           GeneSet=c("In Gene Set", "Not in Gene Set"),
+                           Test=c("Diff. expressed", "Not diff. expr.")))
> G0nnnnCounts
      Test
GeneSet        Diff. expressed Not diff. expr.
  In Gene Set              10          30
  Not in Gene Set           390         1220
> fisher.test(G0nnnnCounts, alternative="greater")

  Fisher's Exact Test for Count Data

data: G0nnnnCounts
p-value = 0.517
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.5149828      Inf
sample estimates:
odds ratio
 1.042711
```

P-value not small, odds-ratio approx. 1 → List is not *surprisingly* enriched in Gene Set

Recipe for gene list enrichment test

- **Step 1:** Define **gene list** (e.g. thresholding analyzed list) and **background list**,
- **Step 2:** Select gene sets to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

Possible problems with gene list test

- No “natural” value for the threshold
- Possible loss of statistical power due to thresholding
 - No resolution between significant signals with different strengths
 - Weak signals neglected
- Different results at different threshold settings
- Based on the wrong assumption of independent gene (or gene group) sampling, which increases false positive predictions

Analysis of ranked gene lists with
Gene Set Enrichment Analysis
(also called Functional Class Scoring)

Gene Sets

- A gene set
 - a group of genes with related functions.
 - sets of genes or pathways, for their association with a phenotype.
 - Examples: metabolic pathway, protein complex, or GO (gene ontology) category.
- Identified from a prior biological knowledge.
- May better reflect the true underlying biology.
- May be more appropriate units for analysis.

Gene Sets

Each row represents one gene set →

	A	B	C	D	E	F	G
1	chr10q24	Cytogenetic band	PITX3	SPFH1	NEURL	C10orf12	NDUFB8
2	chr5q23	Cytogenetic band	ALDH7A1	IL13	8-Sep	INF1	ACSL6
3	chr8q24	Cytogenetic band	HAS2	LRRC14	TSTA3	DGAT1	RECQL4
4	chr16q24	Cytogenetic band	RPL13	GALNS	FANCA	CPNE7	COTL1
5	chr13q14	Cytogenetic band	AKAP11	ARL11	ATP7B	C13orf1	C13orf9
6	chr7p21	Cytogenetic band	ARL4A	SCIN	GLCCI1	SP8	SOSTDC1
7	chr10q23	Cytogenetic band	SNCG	FER1L3	C10orf116	HHEX	TNKS2
8	chr14q12	Cytogenetic band	C14orf125	FOXG1C	HECTD1	SCFD1	AP4S1
9	chr13q13	Cytogenetic band	ALG5	RFXAP	DCAMKL1	MAB2L1	STOML3
10	chr1p34	Cytogenetic band	JMJD2A	MRPS15	HIVEP3	GJB3	CDCA8
11	chr10q21	Cytogenetic band	MBL2	C10orf70	DNAJC12	BICC1	CXXC6
12			QVCP1	QHAT2	LPIC2	QLC4CA1	QELQD2

If editing in excel, watch out for its tendency to auto-format gene sets (SEP8 becomes 8-Sep)

First column are gene set names. Duplicates are not allowed

Second column contains a brief description. Its optional – you can fill in a dummy field (e.g. "na")

Unequal lengths (i.e # of genes) is allowed

MSigDB Collection	Subcollection	No. Gene Sets
C1: positional gene sets		326
C2: curated gene sets	CGP: chemical and genetic perturbations CP: Canonical pathways KEGG/Biocarta/REACTOME	3402
C3: motif gene sets	MIR: microRNA targets TFT: transcription factor targets	221 615
C4: computational gene sets	CGN: cancer gene neighborhoods CM: cancer modules	427 431
C5: GO gene sets	BP: GO biological process CC: GO cellular component MF: GO molecular function	825 233 396
C6: oncogenic signatures		189
C7: immunologic signatures		1910
	Total	10295

Gene Set (Enrichment) Analysis

- Mootha (2003) as an alternative to ORA.
- It aims to identify gene sets with *subtle but coordinated expression changes* that cannot be detected by ORA methods.
 - Weak changes in individual genes gathered to large gene sets can show a significant pattern.
- Results not affected by arbitrarily chosen cutoffs.
- It does not provide information as detailed as ORA

The GSEA method

- Original GSEA method is based on comparing, for each gene group, the distribution of the test statistic within the group with the overall distribution of those statistics, i.e. the calculated for all genes.
- To do this, test statistics are ranked (from biggest to smallest) and **for each gene set** a running sum is computed such that
 - If a gene is in the gene set add a certain quantity (moderate)
 - If a gene is not in the gene set, subtract a (small) quantity
- The distribution of the running sum is compared with that of the random walk using a Kolmogorov-Smirnov test (K-S test) statistic
- P-values are computed based on a randomization.

Calculating enrichment score (ES)

Create a running sum statistic based on the following
If gene p is not in set S, then add

$$X_i = -\sqrt{\frac{N_S}{N - N_S}}$$

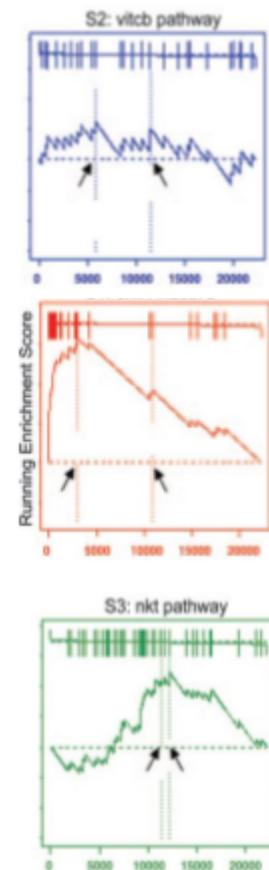
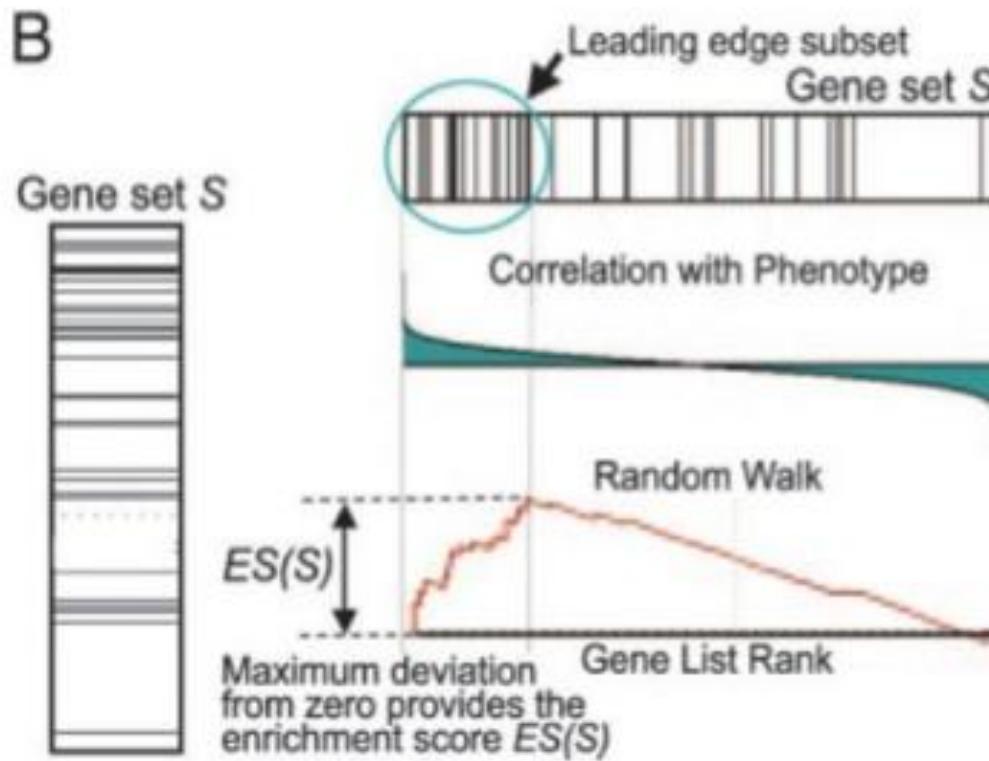
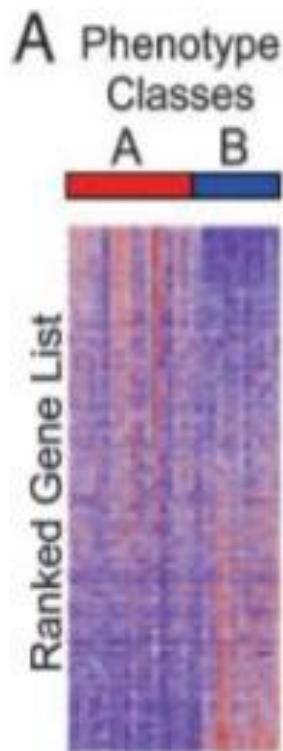
If gene p is in set S, then add

$$X_i = \sqrt{\frac{N - N_S}{N_S}}$$

This creates a running sum

The maximum sum over the whole list L is the Enrichment Score
MES

The GSEA method



Recipe for **ranked** list enrichment test

- **Step 1:** Rank ALL your genes,
- **Step 2:** Select gene sets to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

GSEA variants

- GSEA is not free from criticisms
 - Use of KS test
 - Null hypothesis is not clear
- Many alternative available
 - Efron's GSA
 - Limma's ROAST
 - Irizarry's simple GSA based on Wilcoxon...

Multiple test adjustments

Why we need to “adjust”

- We use a statistical test to decide if a gene list is “surprisingly” enriched in a Gene Set.
 - We use “surprisingly” instead of “significantly”
- Remember that when doing statistical tests one can be right or wrong differently.
 - Right
 - Rejecting the null hypothesis (H_0) when it is false
 - Not rejecting H_0 when it is true
 - Wrong
 - Rejecting the null hypothesis (H_0) when it is true
 - Not rejecting H_0 when it is false

Errors and Successes in tests: Type I and type II errors

		Actual Situation “Truth”	
		H_0 True	H_0 False
Decision	Do Not Reject H_0	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
	Reject H_0	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

$$\alpha = P(\text{Type I Error}) \quad \beta = P(\text{Type II Error})$$

Testing repeatedly

- Omics studies are “high throughput”
 - Selecting genes: One test per each gene
 - Finding enriched gene sets: One test per each gene set
- Doing many tests means facing repeatedly the probability of making one false positive.
 - As the number of tests increases →
 - The chance of observing at least one false positive is going to increase too.

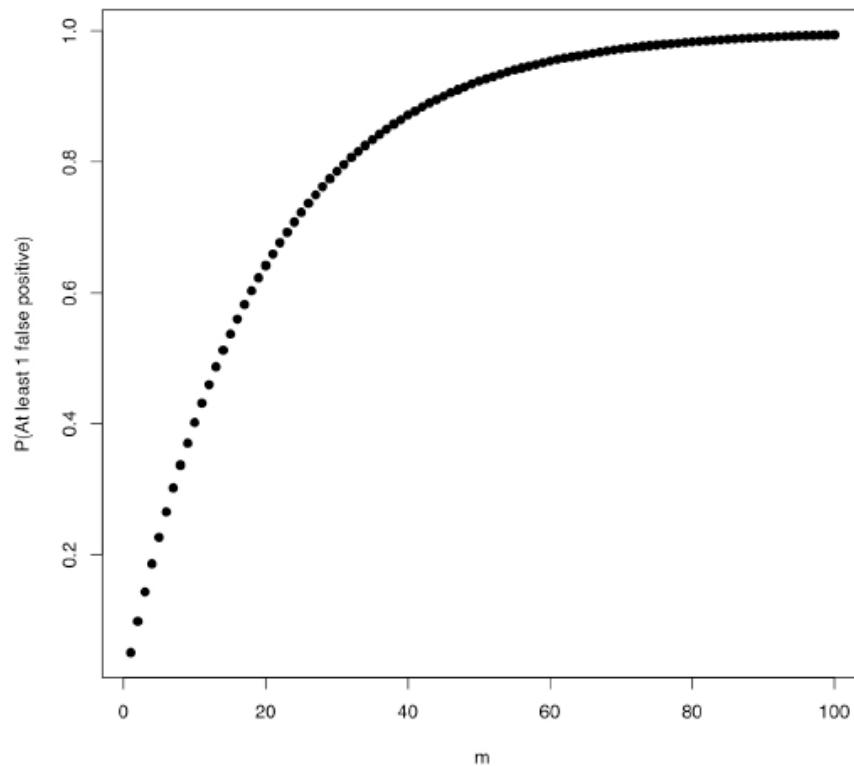
Why multiple testing matters

- The probability of observing one false positive if testing once is:
 - $P(\text{Making a type I error}) = \alpha$
 - $P(\text{not making a type I error}) = 1 - \alpha$
- Now imagine we perform m tests independently
 - $P(\text{not making a type I error in } m \text{ tests}) = (1 - \alpha)^m$
 - $P(\text{making at least a type I error in } m \text{ tests}) = 1 - (1 - \alpha)^m$
- As m increases the probability of having at least one type error tends to increase

Type I error not useful in multiple testing

Probability of At Least 1 False Positive

Number of tests: m	P(making at least a type I error) = $1-(1-a)^m$
1	0.050000
2	0.097500
3	0.142625
4	0.185494
5	0.226219
6	0.264908
7	0.301663
8	0.336580



How can we deal with this issue?

- Controlling for type I error is not feasible if many tests.
- Idea: Modify α (or alternatively the p-value) so the error probability is ***controlled overall***
- This may mean different things:
 1. The probability of at least one error in m tests is $< \alpha$
 2. The expected number of false positives is below a fixed threshold.

...

Controlling the FWER: *Bonferroni*

If $M = \#$ of annotations tested:

Corrected P-value = $M \times$ original P-value

Corrected P-value is greater than or equal to the probability that ***one or more of the observed enrichments*** could be due to random draws.

The jargon for this correction is “controlling for the *Family-Wise Error Rate (FWER)*”

Bonferroni correction caveats

- Bonferroni correction is very stringent and can “wash away” real enrichments leading to false negatives,
- Often one is willing to accept a less stringent condition, the “false discovery rate” (FDR), which leads to a gentler correction when there are real enrichments.

False discovery rate (FDR)

- FDR is *the expected proportion of “False Positives” that is of the observed enrichments due to random chance.*
- Compare to Bonferroni correction which is a bound on *the probability that any one of the observed enrichments could be due to random chance.*
- Typically FDR corrections are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the “q-value”

Benjamini-Hochberg example I

Rank	Category	(Nominal) P-value
1	<i>Transcriptional regulation</i>	0.001
2	<i>Transcription factor Initiation of transcription</i>	0.002
3	<i>Nuclear localization</i>	0.003
4	<i>Chromatin modification</i>	0.0031
5	<i>...</i>	0.005
52	<i>Cytoplasmic localization</i>	...
53	<i>Translation</i>	0.97
		0.99

Sort P-values of all tests in decreasing order

Benjamini-Hochberg example II

Rank	Category	(Nominal) P-value	Adjusted P-value
1	<i>Transcriptional regulation</i>	0.001	$0.001 \times 53/1 = 0.053$
2	<i>Transcription factor</i>	0.002	$0.002 \times 53/2 = 0.053$
3	<i>Initiation of transcription</i>	0.003	$0.003 \times 53/3 = 0.053$
4	<i>Nuclear localization</i>	0.0031	$0.0031 \times 53/4 = 0.040$
5	<i>Chromatin modification</i>	0.005	$0.005 \times 53/5 = 0.053$
...
52	<i>Cytoplasmic localization</i>	0.97	$0.985 \times 53/52 = 1.004$
53	<i>Translation</i>	0.99	$0.99 \times 53/53 = 0.99$

Adjusted P-value is “nominal” P-value times # of tests divided by the rank of the P-value in sorted list

$$\text{Adjusted P-value} = \text{P-value} \times [\# \text{ of tests}] / \text{Rank}$$

Benjamini-Hochberg example III

Rank	Category	(Nominal) P-value	Adjusted P-value	FDR / Q-value
1	<i>Transcriptional regulation</i>	0.001	$0.001 \times 53/1 = 0.053$	0.040
2	<i>Transcription factor</i>	0.002	$0.002 \times 53/2 = 0.053$	0.040
3	<i>Initiation of transcription</i>	0.003	$0.003 \times 53/3 = 0.053$	0.040
4	<i>Nuclear localization</i>	0.0031	$0.0031 \times 53/4 = 0.040$	0.040
5	<i>Chromatin modification</i>	0.005	$0.005 \times 53/5 = 0.053$	0.053
...
52	<i>Cytoplasmic localization</i>	0.97	$0.985 \times 53/52 = 1.004$	0.99
53	<i>Translation</i>	0.99	$0.99 \times 53/53 = 0.99$	0.99

Q-value (or FDR) corresponding to a nominal P-value is the smallest adjusted P-value assigned to P-values with the same or larger ranks.

Benjamini-Hochberg example III

Rank	Category	P-value threshold for FDR < 0.05 (Nominal)	Adjusted P-value	FDR / Q-value
		P-value		
1	<i>Transcriptional regulation</i>	0.001	$0.001 \times 53/1 = 0.053$	0.040
2	<i>Transcription factor</i>	0.002	$0.002 \times 53/2 = 0.053$	0.040
3	<i>Initiation of transcription</i>	0.003	$0.003 \times 53/3 = 0.053$	0.040
4	<i>Nuclear localization</i>	0.0031	$0.0031 \times 53/4 = 0.040$	0.040
5	<i>Chromatin modification</i>	0.005	$0.005 \times 53/5 = 0.053$	0.053
...
52	<i>Cytoplasmic localization</i>	0.97	$0.985 \times 53/52 = 1.004$	0.99
53	<i>Translation</i>	0.99	$0.99 \times 53/53 = 0.99$	0.99

Red: non-significant

Green: significant at FDR < 0.05

P-value threshold is highest ranking P-value for which corresponding Q-value is below desired significance threshold

Reducing adjustment stringency

- The adjustment to the P-value threshold depends on the # of tests that you do,
- So, no matter what, *the more tests you do, the more sensitive the test needs to be*
- Can control the stringency by ***reducing the number of tests:***
 - Don't use all collections of Gene Sets available
 - Restrict testing to the appropriate GO annotations;
 - Filter gene sets by size

Tools for Pathway Analysis

R/Bioconductor



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Search:

[Home](#) » [BiocViews](#)

All Packages

Bioconductor version 3.10 (Release)

Autocomplete biocViews search:

Software (1823)

- ▶ AssayDomain (732)
- ▼ BiologicalQuestion (756)
 - AlternativeSplicing (38)
 - Coverage (117)
 - DifferentialExpression (315)
 - DifferentialMethylation (42)
 - DifferentialPeakCalling (12)
 - DifferentialSplicing (30)
 - DNA3DStructure (5)
 - DriverMutation (1)
 - FunctionalPrediction (18)
 - GeneFusionDetection (1)
 - GenePrediction (17)
 - GeneRegulation (88)
 - GeneSetEnrichment (120)**
 - GeneSignaling (6)

Packages found under GeneSetEnrichment:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All ▾ entries Search table:

Package	Maintainer	Title	Rank
limma	Gordon Smyth	Linear Models for Microarray Data	13
edgeR	Yunshun Chen, Aaron Lun, Mark Robinson, Gordon Smyth	Empirical Analysis of Digital Gene Expression Data in R	24
fgsea	Alexey Sergushichev	Fast Gene Set Enrichment Analysis	41
DOSE	Guangchuang Yu	Disease Ontology Semantic and Enrichment analysis	42
clusterProfiler	Guangchuang Yu	statistical analysis and visualization of functional profiles for genes and gene clusters	44
enrichplot	Guangchuang Yu	Visualization of Functional Enrichment Result	49
GSEABase	Bioconductor Package Maintainer	Gene set enrichment data structures and methods	53
pathview	Weijun Luo	a tool set for pathway based data integration and visualization	64
Category	Bioconductor Package Maintainer	Category Analysis	68
GOSTats	Bioconductor Package Maintainer	Tools for manipulating GO and microarrays	79

Gene Set Variation Analysis for microarray and

March 2017:

February 2020:

74 packages under the view “Gene Set Enrichment”

120 packages under the view “Gene Set Enrichment”

Other (non-R) pathway analysis tools

- DAVID
- Pathway Painter
- Babelomics
- GenMAPP (www.genmapp.com)
- WikiPathways (www.wikipathways.org)
- cPath (cbio.mskcc.org/cpath)
- BioCyc (www.biocyc.org)
- Pubgene (www.pubgene.org)
- PANTHER (www.pantherdb.org)
- WebGestalt (bioinfo.vanderbilt.edu/webgestalt/)
- ToppGeneSuite ([/toppgene.cchmc.org/](http://toppgene.cchmc.org/))
- GeneGo/MetaCore (www.genego.com)
- Ingenuity Pathway Analysis (www.ingenuity.com)
- Pathway Studio (www.riadnegenomics.com)

Summary

- Pathway Analysis is a useful approach to help gain biological understanding from omics-based studies.
- There are many ways, many methods, many tools
- Choice of the method should be guided by
 - a combination of availability, ease of use and usefulness ,
 - Usually obtained from a good understanding of how it
- Different methods may yield different results
 - Worth checking!

References

- Efron, Bradley, and Robert Tibshirani. 2007. “On Testing the Significance of Sets of Genes.” *The Annals of Applied Statistics* 1 (1): 107–29. doi:10.1214/07-AOAS101.
- Irizarry, Rafael A., Chi Wang, Yun Zhou, and Terence P. Speed. 2009. “Gene Set Enrichment Analysis Made Simple.” *Statistical Methods in Medical Research* 18 (6): 565–75. doi:10.1177/0962280209351908.
- Khatri, Purvesh, and Sorin Drăghici. 2005. “Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems.” *Bioinformatics (Oxford, England)* 21 (18): 3587–95. doi:10.1093/bioinformatics/bti565.
- Khatri, Purvesh, Marina Sirota, and Atul J. Butte. 2012. “Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges.” *PLOS Computational Biology* 8 (2): e1002375. doi:10.1371/journal.pcbi.1002375.
- Maciejewski, Henryk. 2014. “Gene Set Analysis Methods: Statistical Models and Methodological Differences.” *Briefings in Bioinformatics* 15 (4): 504–18. doi:10.1093/bib/bbt002.
- Mootha, Vamsi K., Cecilia M. Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, et al. 2003. “PGC-1 α -Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes.” *Nature Genetics* 34 (3): 267–73. doi:10.1038/ng1180.
- Pan, Kuang-Hung, Chih-Jian Lih, and Stanley N. Cohen. 2005. “Effects of Threshold Choice on Biological Conclusions Reached during Analysis of Gene Expression by DNA Microarrays.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (25): 8961–65. doi:10.1073/pnas.0502674102.
- Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. doi:10.1073/pnas.0506580102.
- the Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium. 2015. “Pathway and Network Analysis of Cancer Genomes.” *Nature Methods* 12 (7): 615–21. doi:10.1038/nmeth.3440.