

An Introduction to Pathway Analysis

Alex Sánchez



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



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



Outline

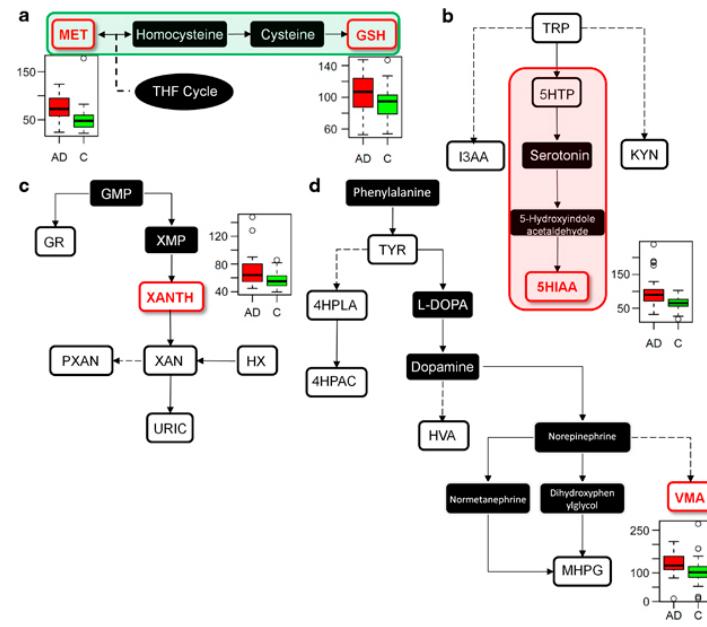
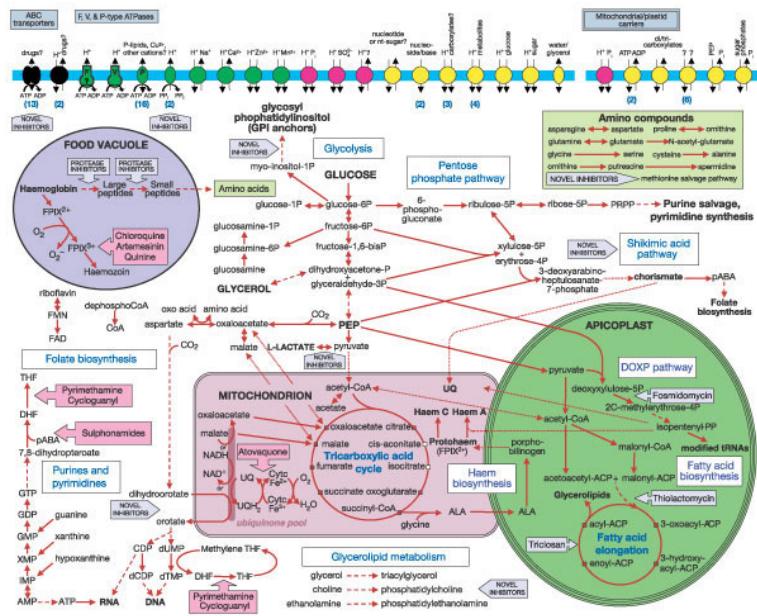
- Presentation
- Introduction and Background
- The problem: Interpreting gene lists
- Annotations and annotation databases
- The Gene Ontology Resource
- Gene list analysis using the GO and relatives
- Existing tools for pathway analysis

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

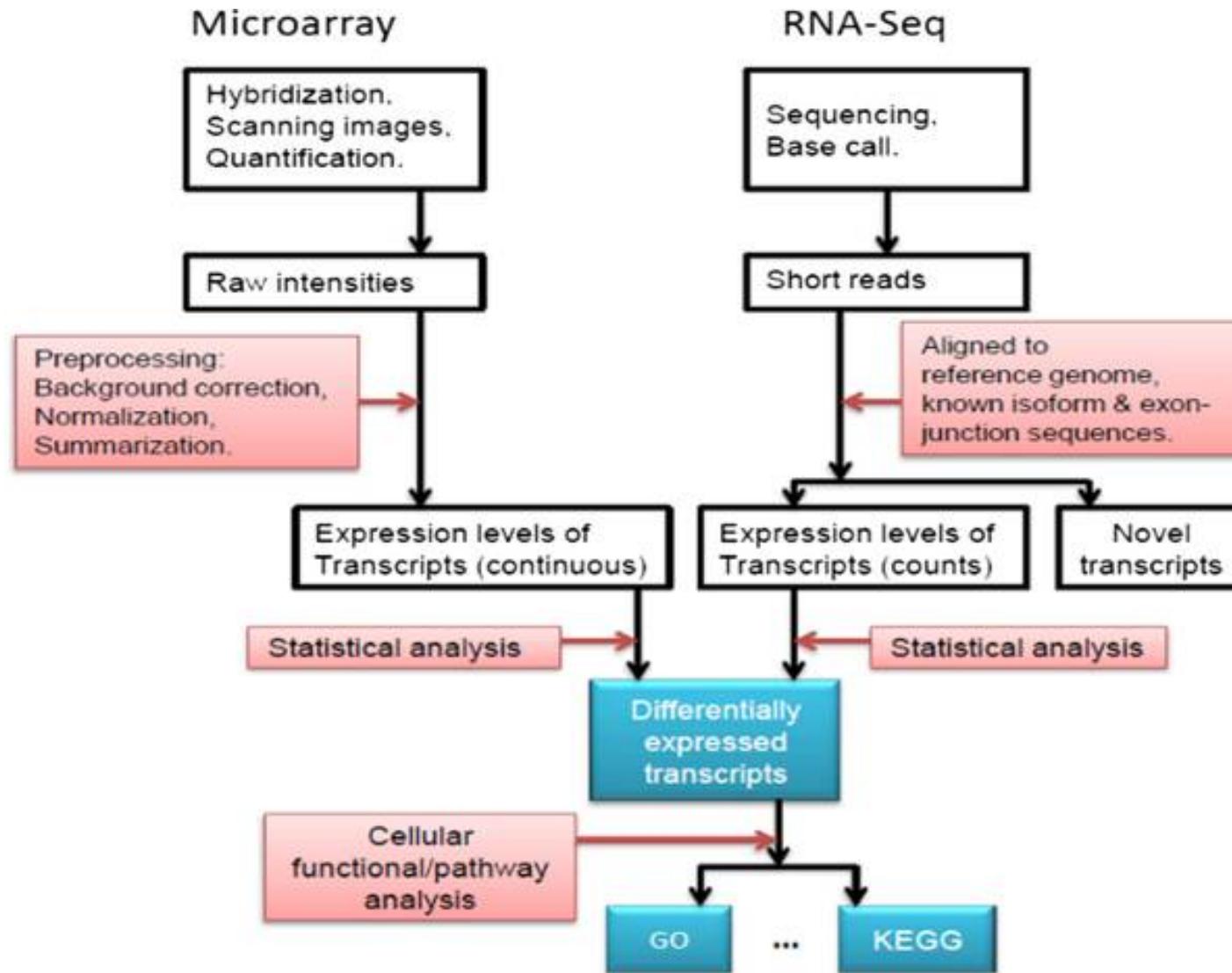
That's what Pathway Analysis is about (more or less)

Pathway Analysis

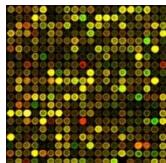
- The term Pathway or Network Analysis denotes *any analytic technique that benefits from biological pathway or molecular network information to gain insight into a biological system.* (Creixell et alt, Nature Methods 2015 (12 (7)))
- To be more specific, Pathway Analysis methods rely on high throughput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

Managing Gene Lists

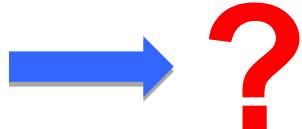
The life-cycle of an omics-based study



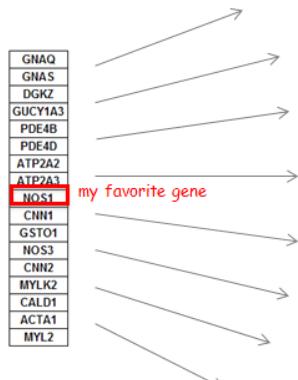
The (in)famous “*where to now?*” question



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



- 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 on all features in the list?



NCBI Resources How To

PubMed 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), 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]
Med - as supplied by publisher]

From gene lists to *Pathway Analysis*

- Gene lists contain useful information
 - This 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***

Case study 1

- Lists *AvsB*, *AvsL* and *BvsL* contain the IDs of genes selected by being differentially expressed between three types of breast cancer tumors.
 - Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M et al. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* 2005 Jul 7;24(29):4660-71. PMID: [15897907](#)
- See the analysis that generates the list in:
[https://github.com/alexsanchezpla/scripts
/tree/master/Exemple_Analysis_BioC](https://github.com/alexsanchezpla/scripts/tree/master/Exemple_Analysis_BioC)

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.

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

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 S000002187 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

Retraction: Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells

Hiroaki Kawasaki & Kazunari Taira

Nature 423, 838–842 (2003).

In this Article, the messenger RNA that is identified to be a target of microRNA-23 (miR-23) is from the gene termed human 'homolog of ES1' (HES1), accession number Y07572, and not from the gene encoding the transcriptional repressor 'Hairy enhancer of split' HES1 (accession number NM_00524) as stated in our paper. We incorrectly identified the gene because of the confusing nomenclature. The function of HES1 Y07572 is unknown but the encoded protein shares homology with a protein involved in isoprenoid biosynthesis. Our experiments in NT2 cells had revealed that the protein levels of the repressor Hes1 were diminished by miR-23. Although we have unpublished data that suggest the possibility that miR-23 might also interact with Hes1 repressor mRNA, the explanation for the finding that the level of repressor Hes1 protein decreases in response to miR-23 remains undefined with respect to mechanism and specificity. Given the interpretational difficulties resulting from our error, we respectfully retract the present paper. Further studies aimed at clarifying the physiological role of miR-23 will be submitted to a peer-reviewed journal subject to the outcome of our ongoing research.

Use ID converters to prepare list

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

Gene Accession Conversion Tool

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

Save the results
Submit the converted genes to DAVID for other analytical tools!!

Conversion Summary	Help	Genes that have been converted.	Right-click to Download the list	Help	Submit Converted List to DAVID
ID Count	In DAVID DB	Conversion	"1112_G_AT	5684	HOMO SAPIENS NEURAL CELL ADHESION MOLECULE 1
157 IDs	Yes	Successful	"1331_S_AT	8718	HOMO SAPIENS TUMOR NECROSIS FACTOR RECEPTOR SUPERFAMILY, MEMBER 25
0 IDs	Yes	None	"1355_G_AT	4915	HOMO SAPIENS NEUROTROPHIC TYROSINE KINASE, RECEPTOR, TYPE 2
1 IDs	No	NA	"1372_AT	2130	HOMO SAPIENS TUMOR NECROSIS FACTOR, ALPHA-INDUCED PROTEIN 6
Total Unique User IDs: 166			"1391_S_AT	1572	HOMO SAPIENS CYTOCHROME P450, FAMILY 4, SUBFAMILY A, POLYPEPTIDE 11
SUMMARY OF AMBIGUOUS GENE IDS					
ID Count	Possible Source	Convert All	"1403_S_AT	6332	HOMO SAPIENS CHEMOKINE (C-C MOTIF) LIGAND 5
1	ENTREZ_GENE_ID	<input type="checkbox"/>	"1419_G_AT	4843	HOMO SAPIENS NITRIC OXIDE SYNTHASE 2A (INDUCIBLE, HEPATOCYTES)
1	GI_ACCESSION	<input type="checkbox"/>	"1575_AT	5243	HOMO SAPIENS ATP-BINDING CASSETTE, SUB-FAMILY B (MDR/TAP), MEMBER 1
POSSIBLE SOURCES FOR AMBIGUOUS IDS					
Ambiguous ID	Possibility	Convert	"1645_AT	3814	HOMO SAPIENS KISS-1 METASTASIS-SUPPRESSOR
3558	ENTREZ_GENE_ID	<input type="checkbox"/>	"1786_AT	10461	HOMO SAPIENS C-MER PROTO-ONCOGENE TYROSINE KINASE
3558	GI_ACCESSION	<input type="checkbox"/>	"1855_AT	2748	HOMO SAPIENS FIBROBLAST GROWTH FACTOR 3 (MURINE MAMMARY TUMOR VIRUS INTEGRATION SITE V-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

[?] 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

ID Mapping Services

The screenshot shows the g:Profiler web interface. At the top, there's a navigation bar with links for Welcome, About, Contact, Beta, Archives, and R. Below the navigation is a sidebar with links for g:GOST, g:Cocoa, g:Convert, g:Sorter, and g:Orth. The main area has sections for Organism (Homo sapiens), Target database (UNIPROTSWISSPROT), and Output type (Table (HTML)). A search bar contains the query "TP53 MDM2 207105_S_AT P60484". To the right of the search bar are options to Interpret query as chromosome ranges and Numeric IDs treated as, with "AFFY_HUGENE_1.0_ST_V1" selected. Below the search bar is a "Query (genes, proteins, probes, term)" section. The main results table has columns for g#, initial alias, converted alias, name, description, and namespace. The table rows show mappings for TP53, MDM2, PIK3R2, and PTEN.

>> Static URL Come back later						
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	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

Input gene/protein/transcript IDs (mix)

Type of output ID

AFFY_HG_U95C
AFFY_HG_U95D
AFFY_HG_U95E
AFFY_HTA_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_NORM

- **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: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

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]

Options

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

[?] 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

[?] or Term ID:
g:Profile! Clear
Example or random query

>> 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

(0%) ENSG00000268173 (PIK3R2, 26 GO annot.) - phosphoinositide-3-kinase, regulatory subunit 2 (beta) [Source:HGNC Symbol;Acc:HGNC:8980]
(0%) ENSG00000105647 (PIK3R2, 26 GO annot.) - phosphoinositide-3-kinase, regulatory subunit 2 (beta) [Source:HGNC Symbol;Acc:HGNC:8980]

Resubmit query

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 R/Bioc, or a spreadsheet if no other option.
- 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

The Gene Ontology (at last)

Where is pathway information?

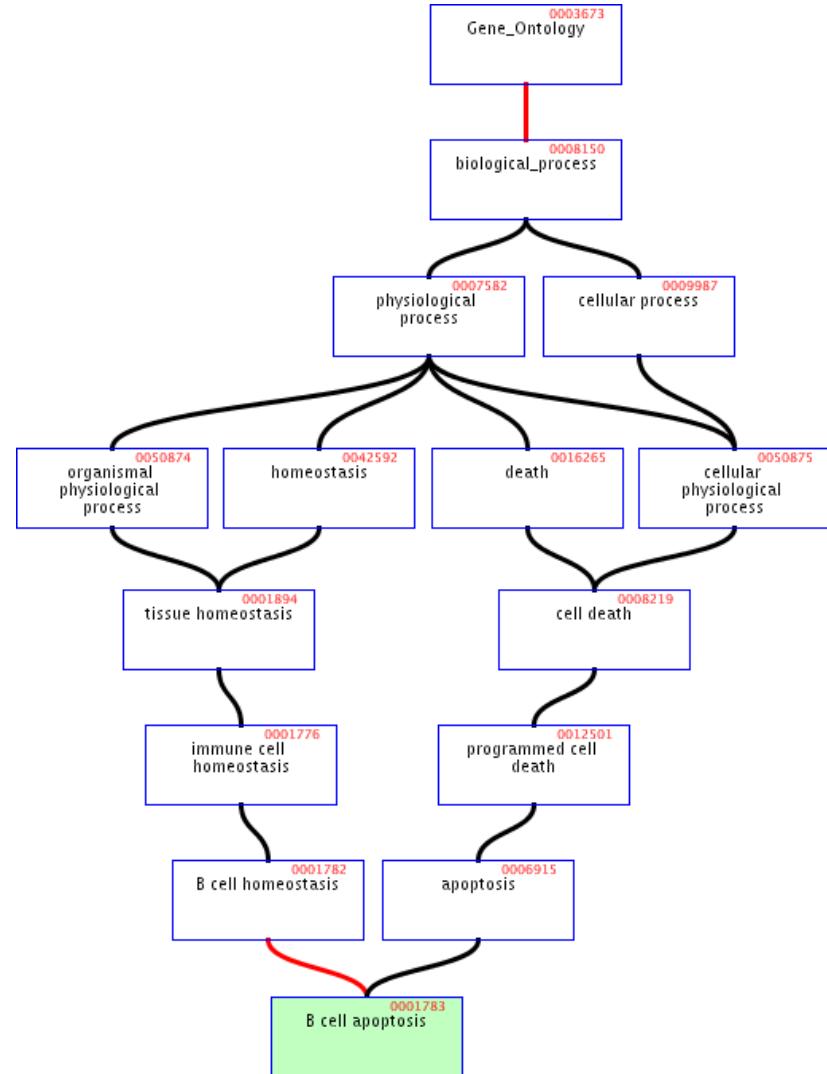
- Pathways
 - Gene Ontology biological process, pathway databases e.g. Reactome
- Other annotations
 - Gene Ontology molecular function, cell location
 - Chromosome position
 - Disease association
 - DNA properties (TF binding sites, gene structure (intron/exon), SNPs, ...)
 - Transcript properties (Splicing,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

What is the Gene Ontology (GO)?

- Set of biological phrases (terms) which are applied to genes:
 - protein kinase
 - apoptosis
 - membrane
- Dictionary: term definitions
- Ontology: A formal system for describing knowledge
- www.geneontology.org

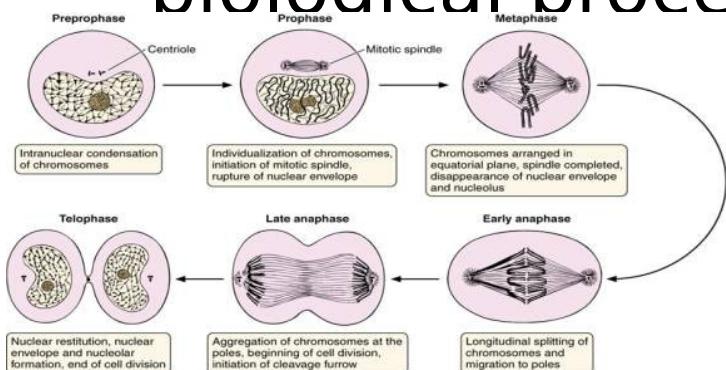
GO Structure

- 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

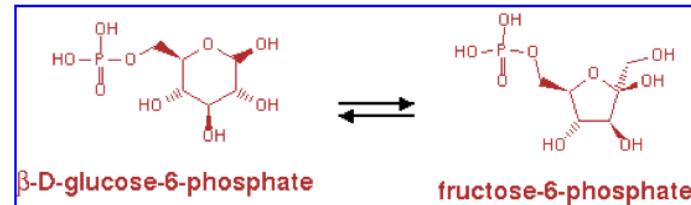
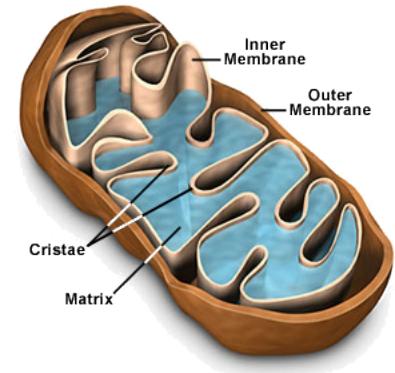


What is covered by the GO?

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



**Cell
division**



**glucose-6-phosphate
isomerase activity**

Part 1/2: Terms

- Where do GO terms come from?
 - GO terms are added by editors at EBI and gene annotation database groups
 - Terms added by request
 - Experts help with major development

	<u>Jun 2012</u>	<u>Apr 2015</u>	<u>increase</u>
Biological process	23,074	28,158	22%
Molecular function	9,392	10,835	15%
Cellular component	2,994	3,903	30%
total	37,104	42,896	16%

Part 2/2: Annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
 - Known as ‘gene associations’ or GO annotations
 - Multiple annotations per gene
- Some GO annotations created automatically (without human review)

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**

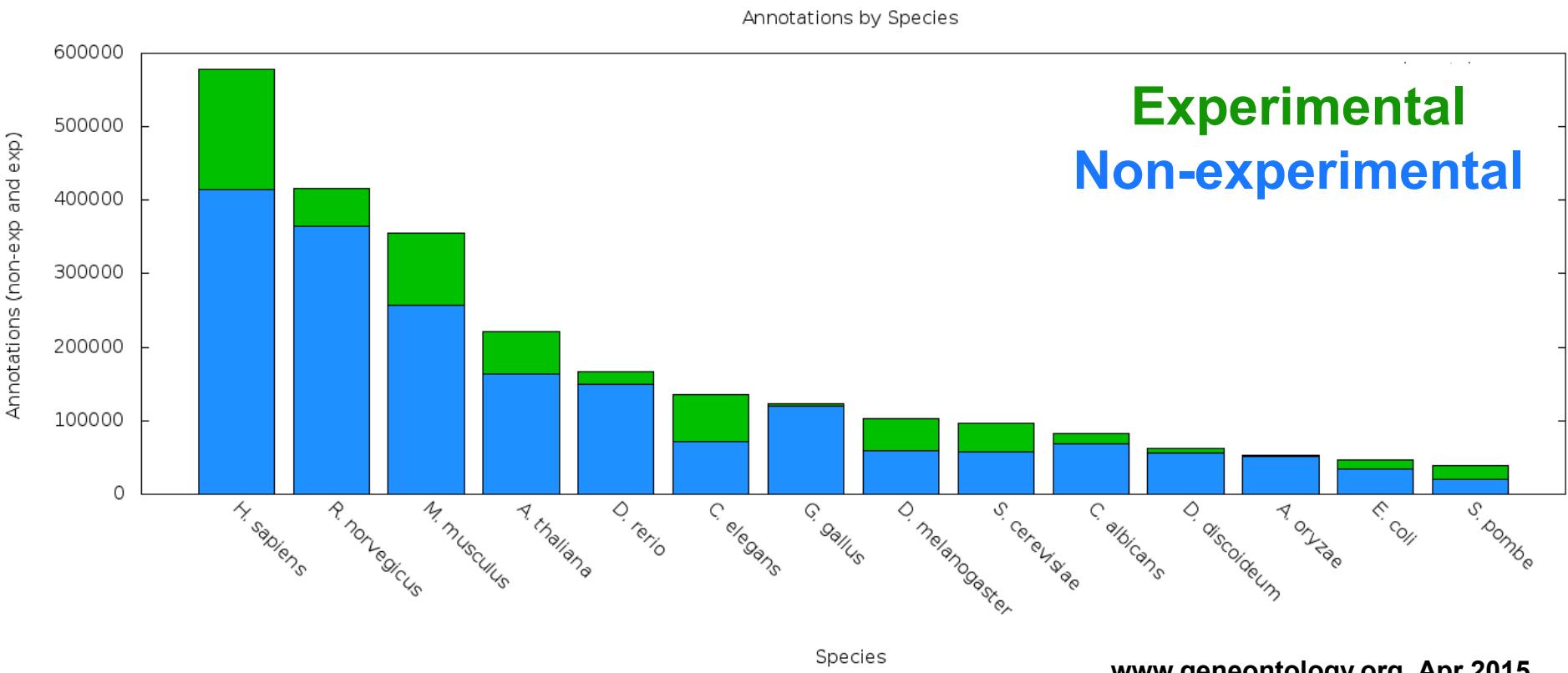


<http://www.geneontology.org/GO.evidence.shtml>

Species Coverage

- All major eukaryotic model organism species and human
- Several bacterial and parasite species through TIGR and GeneDB at Sanger
- New species annotations in development
- Current list:
 - <http://www.geneontology.org/GO.downloads.annotations.shtml>

Variable Coverage

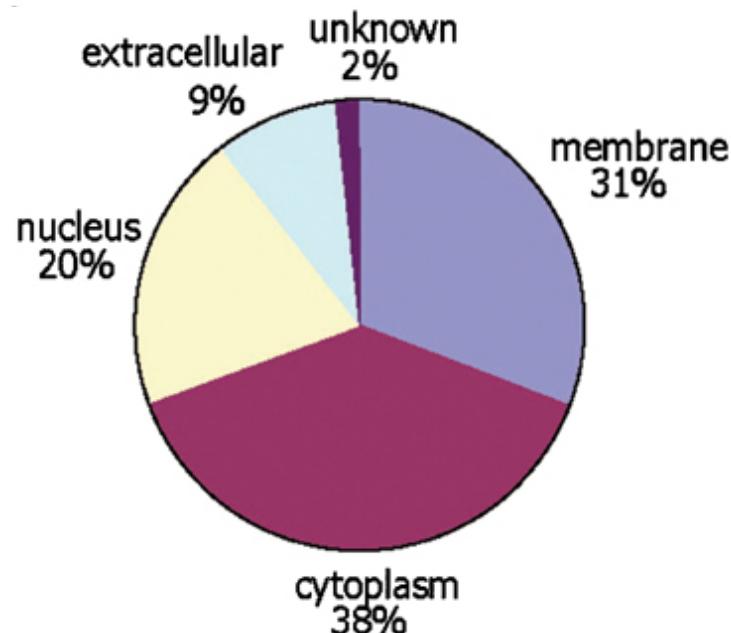


Contributing Databases

- [Berkeley *Drosophila* Genome Project \(BDGP\)](#)
- dictyBase (*Dictyostelium discoideum*)
- FlyBase (*Drosophila melanogaster*)
- GeneDB ([Schizosaccharomyces pombe](#), *Plasmodium falciparum*, *Leishmania major* and *Trypanosoma brucei*)
- [UniProt Knowledgebase](#) (Swiss-Prot/TrEMBL/PIR-PSD) and [InterPro](#) databases
- Gramene (grains, including rice, *Oryza*)
- Mouse Genome Database (MGD) and Gene Expression Database (GXD) (*Mus musculus*)
- Rat Genome Database (RGD) (*Rattus norvegicus*)
- Reactome
- [*Saccharomyces* Genome Database \(SGD\)](#) (*Saccharomyces cerevisiae*)
- The [*Arabidopsis* Information Resource \(TAIR\)](#) (*Arabidopsis thaliana*)
- The Institute for Genomic Research (TIGR): databases on several bacterial species
- WormBase (*Caenorhabditis elegans*)
- Zebrafish Information Network (ZFIN): (*Danio rerio*)

GO Slim Sets

- GO has too many terms for some uses
 - Summaries (e.g. Pie charts)
- GO Slim is an official reduced set of GO terms
 - Generic, plant, yeast

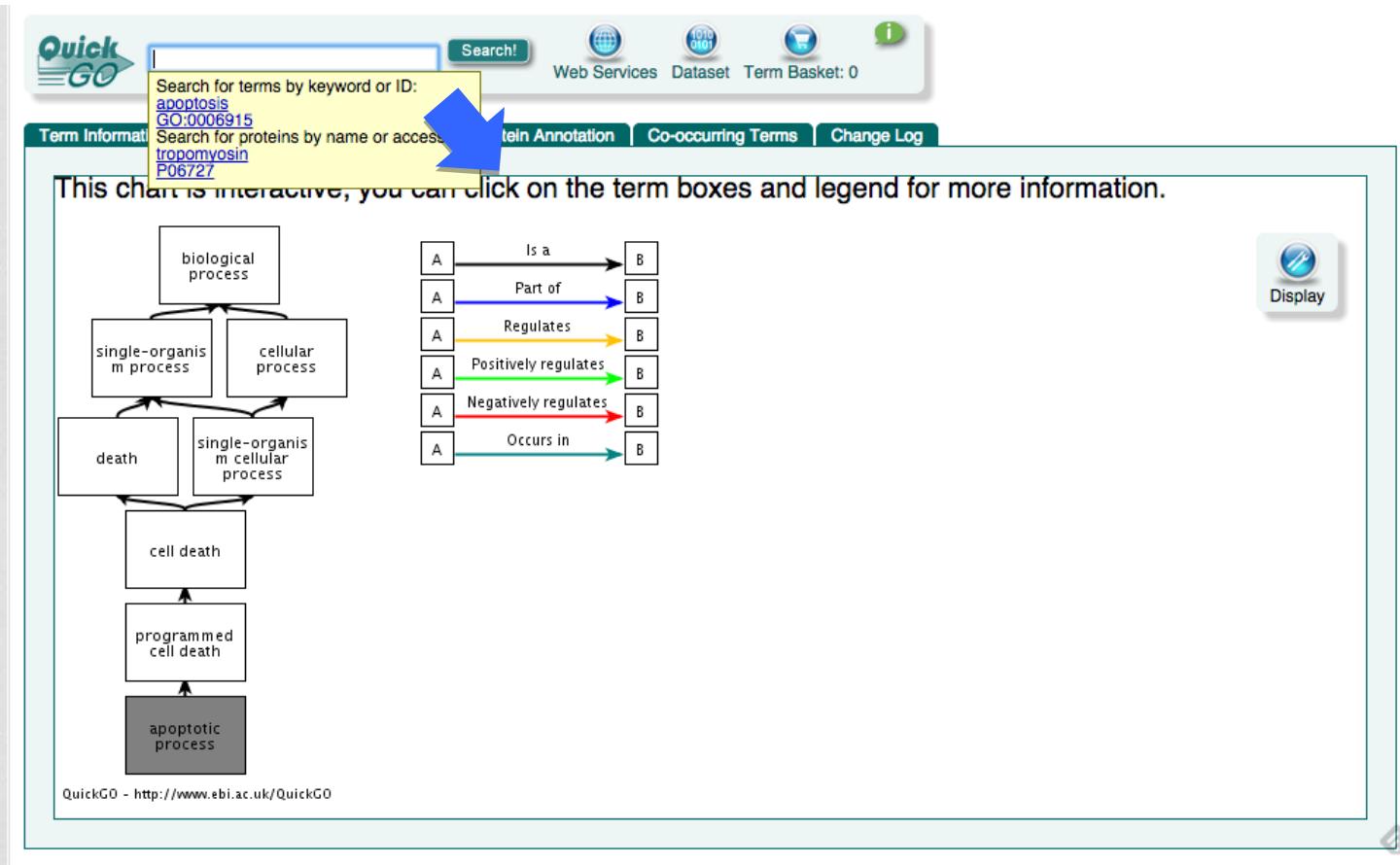


Crockett DK et al. Lab Invest. Nov. 2005;
85(11):1405-15

GO Software Tools

- GO resources are freely available to anyone without restriction
 - ontologies, gene associations and tools developed by GO
- Other have used GO to create versatile tools
 - <https://omictools.com/gene-ontologies-category>
 - <https://omictools.com/gene-set-analysis-category>

Accessing GO: QuickGO



<http://www.ebi.ac.uk/QuickGO/>

Pathway Databases

- <http://www.pathguide.org/> lists ~550 pathway related databases
- MSigDB:
<http://www.broadinstitute.org/gsea/msigdb/>
- <http://www.pathwaycommons.org/> collects major

The screenshot shows the homepage of Pathway Commons. At the top, there is a dark blue header bar with the text "Pathway Commons" on the left and "Download F.A.Q. Publications Contact" on the right. Below the header, the main title "Pathway Commons" is displayed in a large, bold, dark blue font. Underneath the title, a subtitle reads "Search and visualize public biological pathway information. Single point of access." In the center, there is a search bar containing the text "BRCA1, BRCA2, MDM2" and a green button labeled "Start exploring »". At the bottom, a descriptive paragraph explains the purpose of the site: "Pathway Commons is a network biology resource and acts as a convenient point of access to biological pathway information collected from public pathway databases, which you can search, visualize and download. All data is freely available, under the license terms of each contributing database."

Pathway Analysis

*Overrepresentation Analysis
Gene Set Enrichment Analysis*

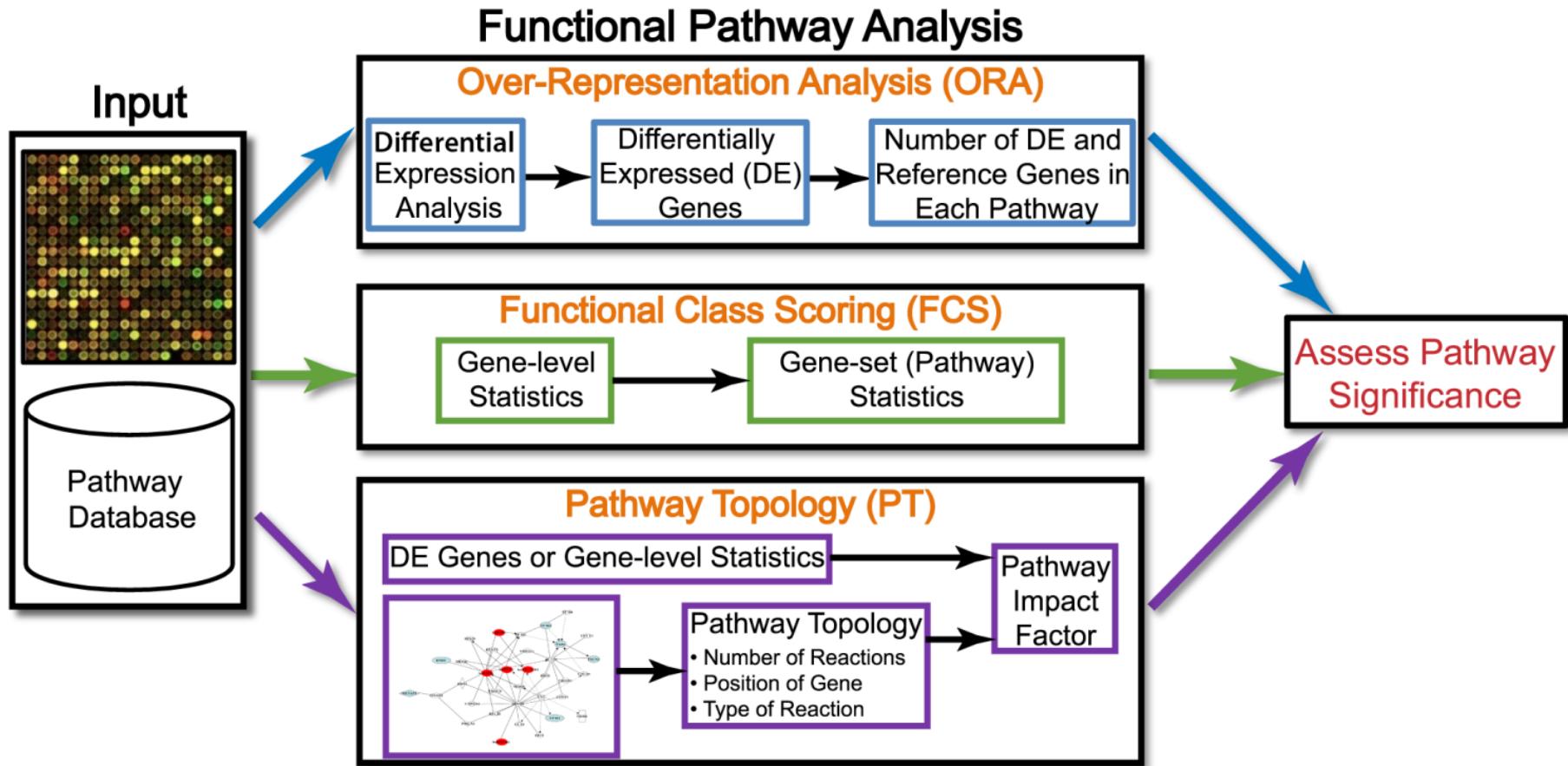
Pathway and Network Analysis

- Any type of analysis that involves pathway or network information
- Helps gain mechanistic insight into ‘omics’ data
 - Identifying a master regulator, drug targets, characterizing pathways active in a sample
- Most commonly applied to help interpret lists of genes
- Most popular type is pathway ***enrichment analysis***, but many others are useful

Benefits of Pathway Analysis

- Easier 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

Types of input for Pathway Analysis

- **Gene list** (e.g. expression change > 2-fold)
 - Answers the question: **Are any gene sets surprisingly enriched (or depleted) in my gene list?**
 - Statistical test: Fisher's Exact Test (aka Hypergeometric test)
- **Ranked list** (e.g. by differential expression)
 - Answers the question: **Are any gene set ranked surprisingly high or low in my ranked list of genes?**
 - Statistical test: minimum hypergeometric test (+ others we won't discuss)

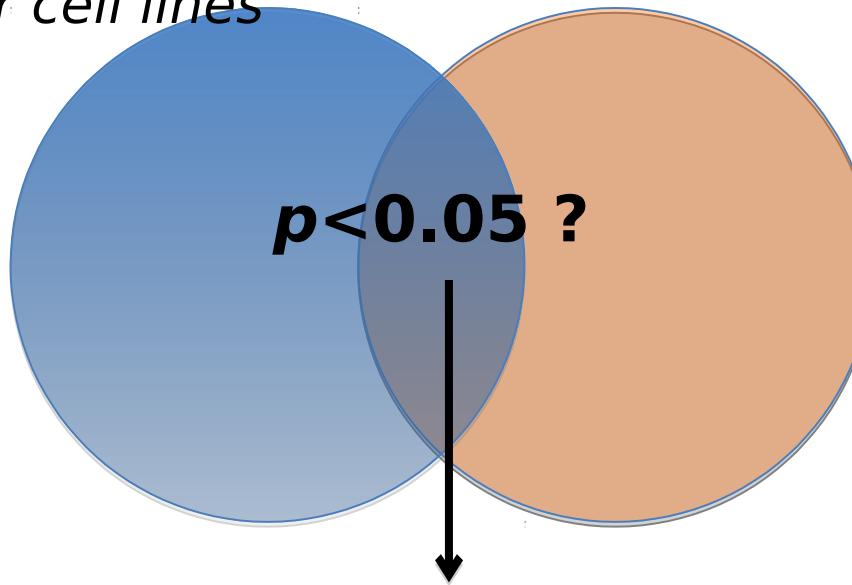
Types of Pathway Analysis depending on *input and test* performed

- Over-representation Analysis
 - Takes a list of differentially expressed genes
 - Looks for “categories” that appear more often in the gene list than would be expected by chance,
 - Uses a hypergeometric /Fisher test approach
 - Intuitive and simple with many limitations
- Gene Set Enrichment Analysis
 - Uses all genes, not only differentially expressed
 - Tests selected gene sets to see if number of genes on top of list of ranked comparisons is more abundant in gene set than would be expected by chance.
 - Uses K-S or wilcoxon test to compare distributions of test statistics.

Over-representation analysis

Gene list from experiment:

Genes down-regulated in drug-sensitive brain cancer cell lines



$p < 0.05 ?$

Pathway
information:

All genes known
to be involved in
Neurotransmitter
signaling



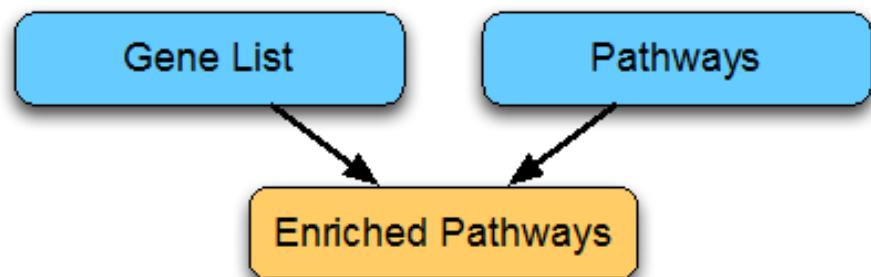
Test many
pathways

Statistical test: are there more annotations in gene list than expected?

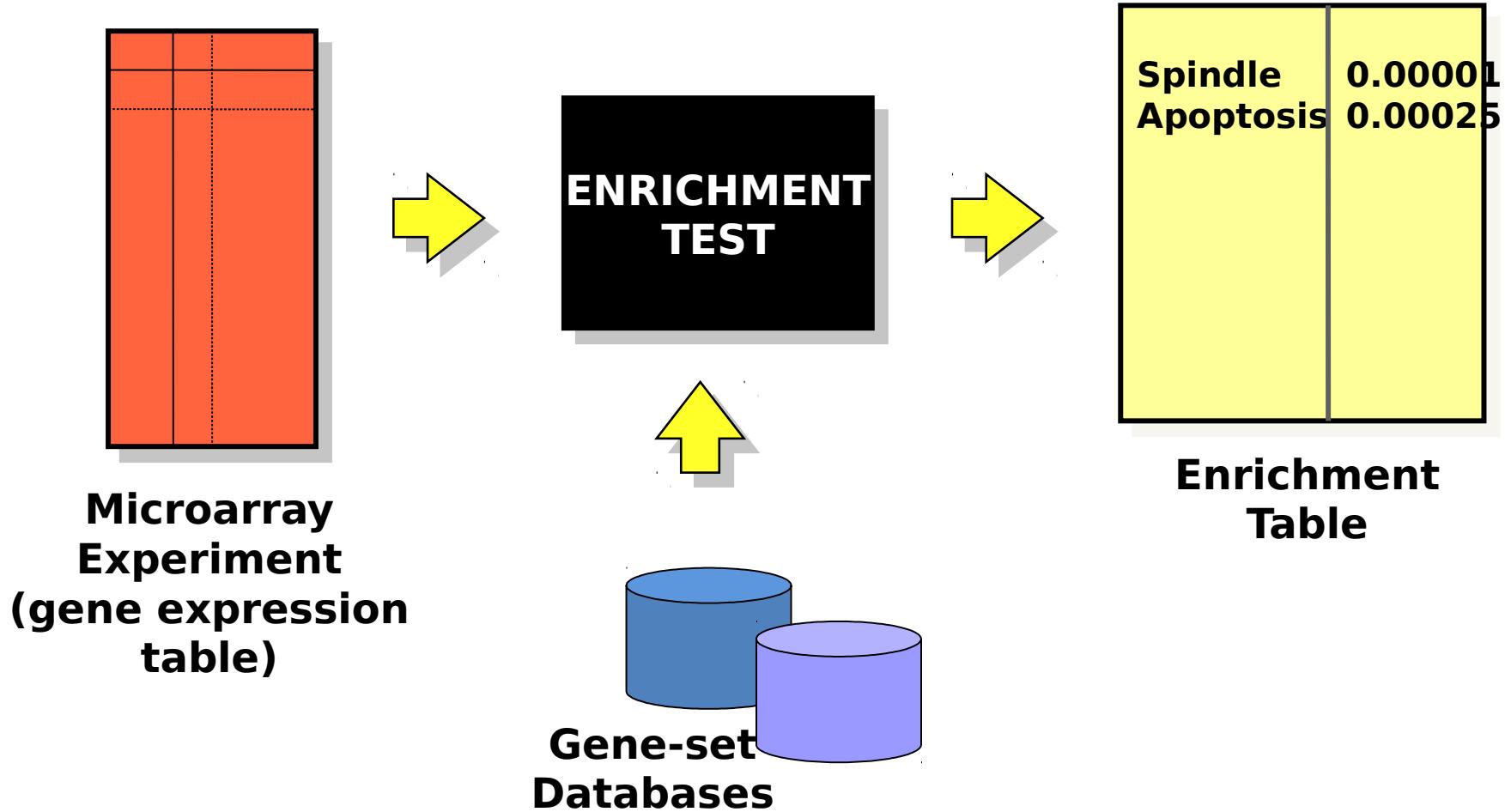
Hypothesis: drug sensitivity in brain cancer is related to reduced neurotransmitter signaling

Over-representation analysis

- Combines
 - Gene(feature) lists ≡ (Gen)omic experiment
 - Pathways and other gene annotations
 - Gene Ontology
 - Ontology Structure
 - Annotation
 - BioMart
 - Other resources



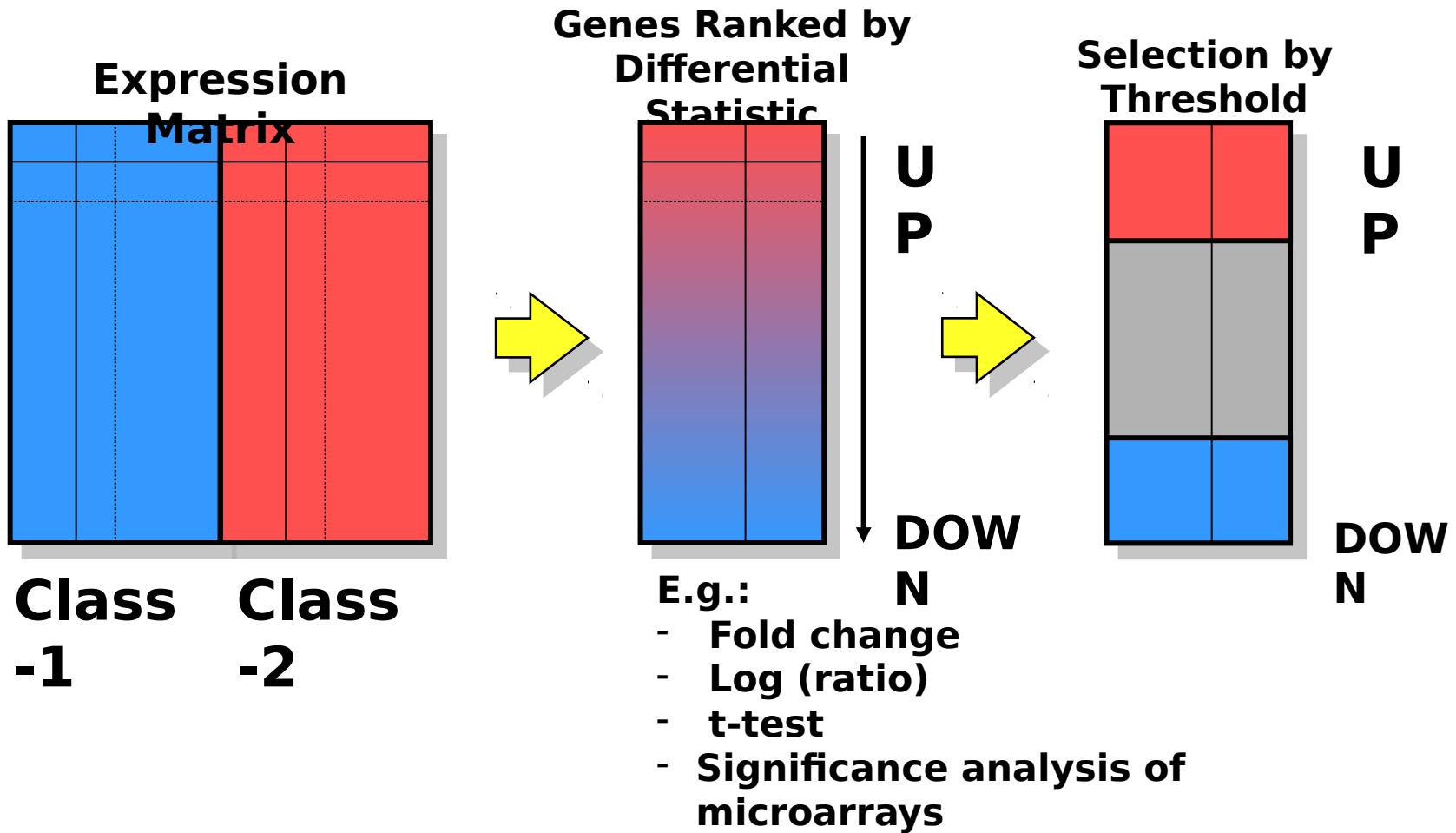
Enrichment Test



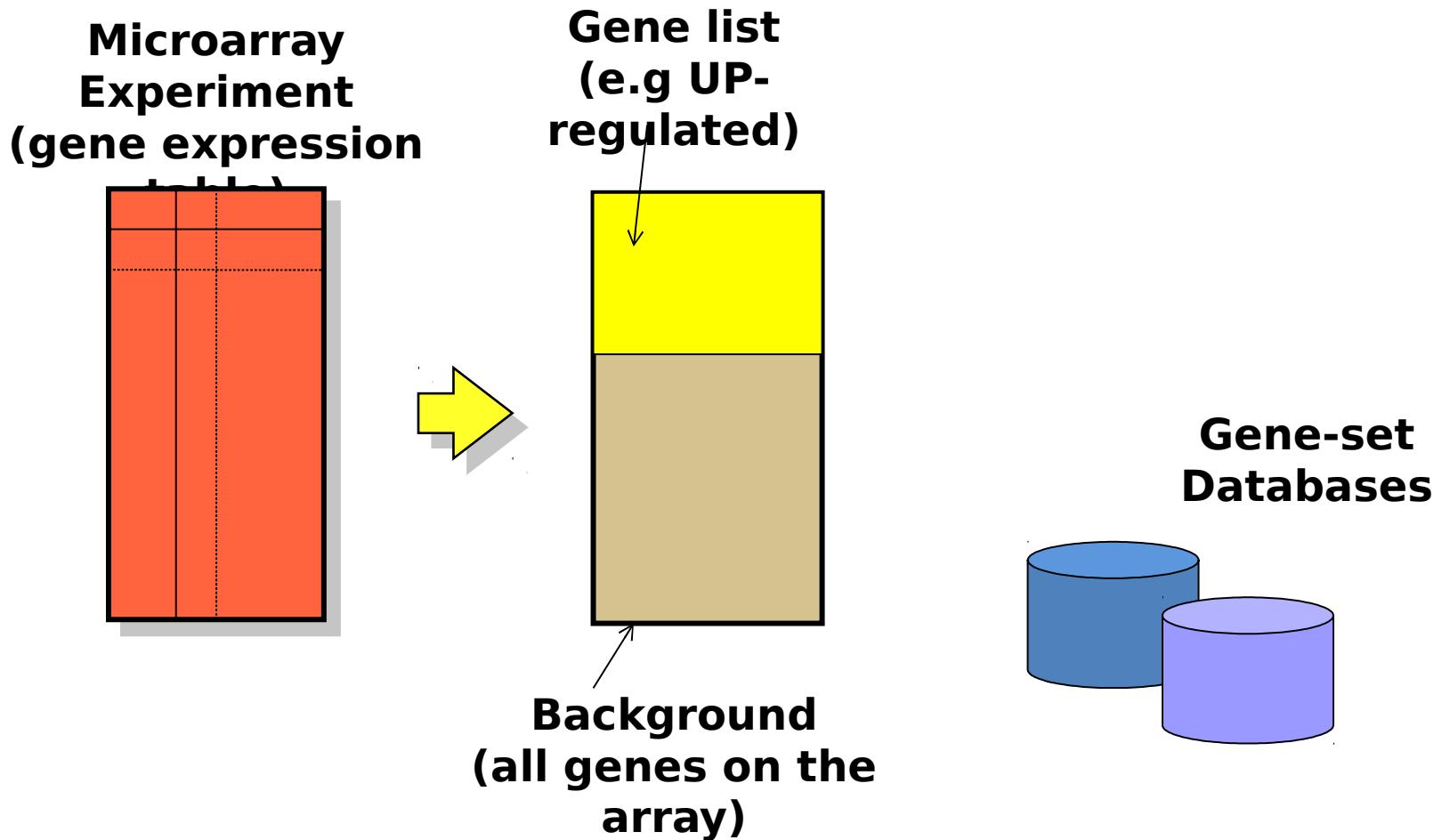
Gene list enrichment analysis

- 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:
 - Where do the gene lists come from?
 - How to assess “surprisingly” (statistics)
 - How to correct for repeating the tests

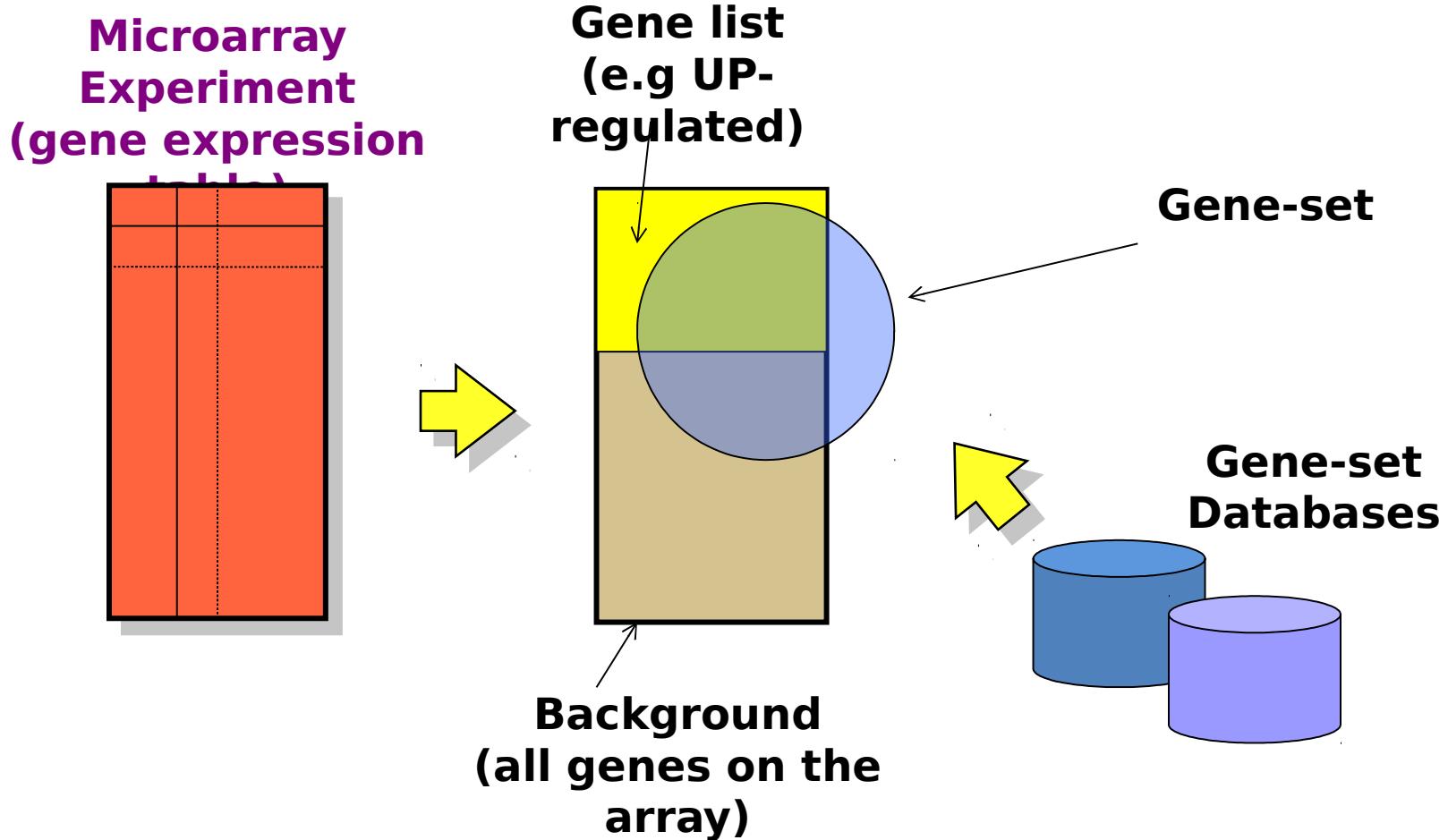
Two-class design for gene lists



Example gene list enrichment test



Example gene list enrichment test



Example gene list enrichment test

- Given a list of differentially expressed genes and a collection of gene sets apply the following strategy
 - For each gene set fill a 2x2 contingency table
 - Calculate p-value using Fisher test
 - Compute FDR to adjust p-values for doing many tests

	Differently expressed	Not differentially expressed	TOTAL
In Gene Set	10	30	40
Not In Gene Set	390	3570	3960
TOTAL	400	3600	4000

Warning: Background must be carefully chosen!

Naive Analysis: A Fisher test using R

```
> 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
```

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

Alternative: Gene Set Testing

- 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	chr1p12	Cytogenetic band	QUB1	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	1320
C4: computational gene sets	CGN: cancer gene neighborhoods CM: cancer modules	221
C5: GO gene sets	BP: GO biological process CC: GO cellular component MF: GO molecular function	615
C6: oncogenic signatures		427
C7: immunologic signatures		431
		825
		233
		396
		189
		1910
	Total	10295

Gene Set (Enrichment) Analysis

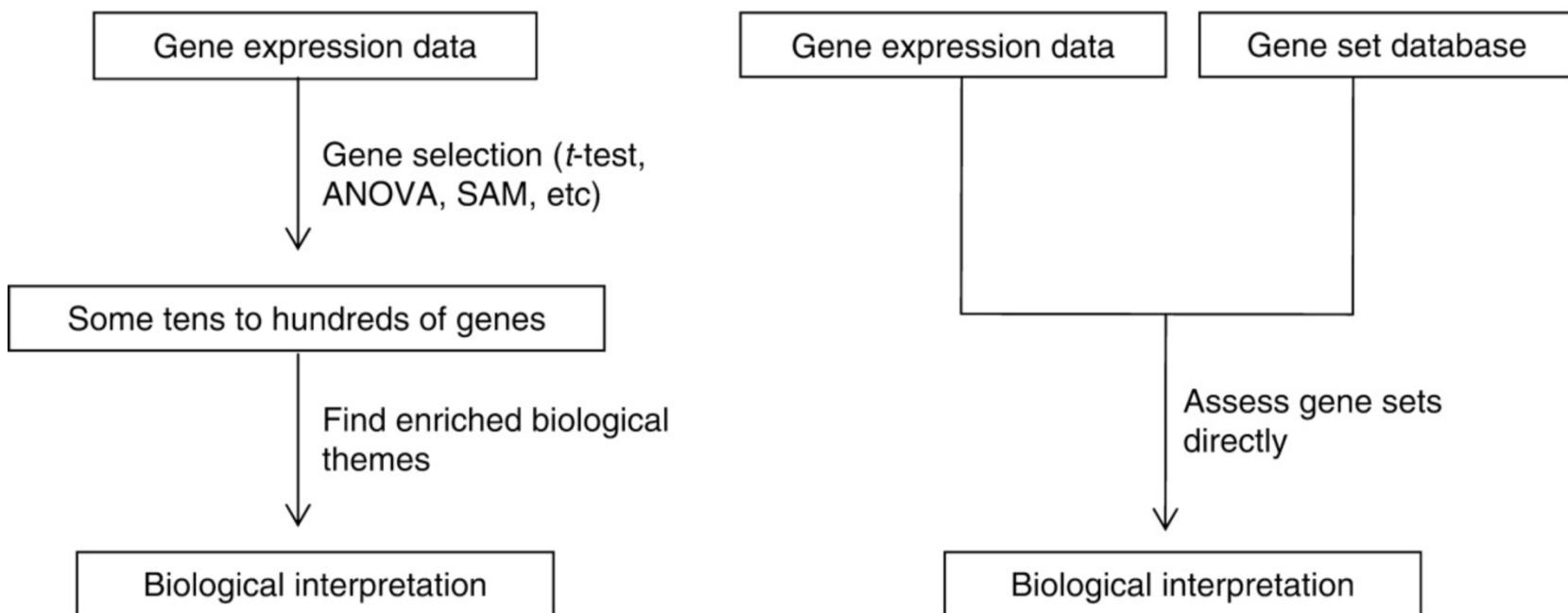
- Introduced by Mootha (2003) as an alternative to ORA.
- It aims to identify gene sets with *subtle but coordinated expression changes* that cannot be detected by ORA methods.
 - Weak changes in individual genes gathered to large gene sets can show a significant pattern.
- Results of GSA are not affected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

From: Gene-set approach for expression pattern analysis

Brief Bioinform. 2008;9(3):189-197. doi:10.1093/bib/bbn001

IGA

GSA

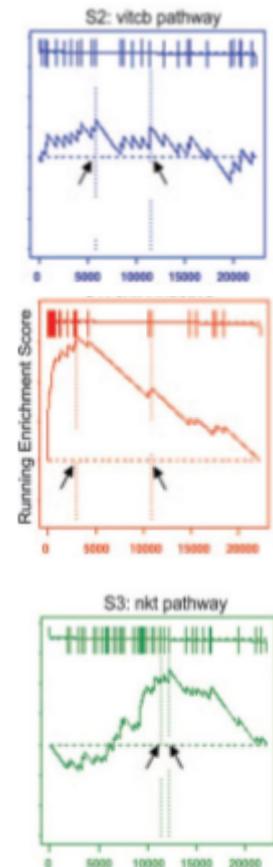
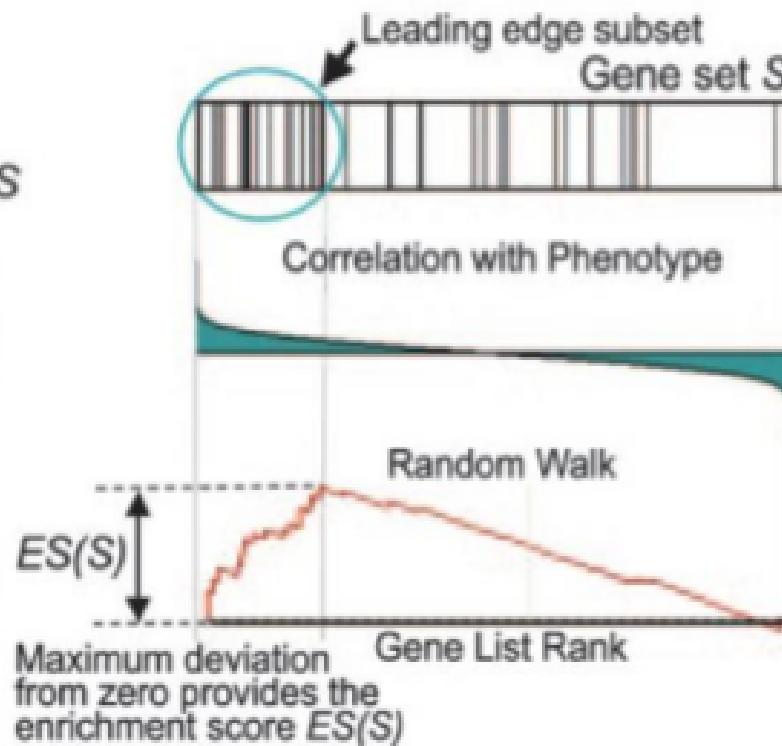
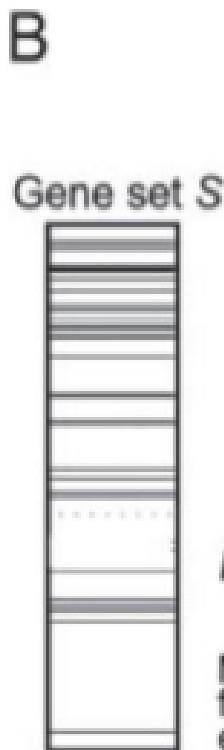
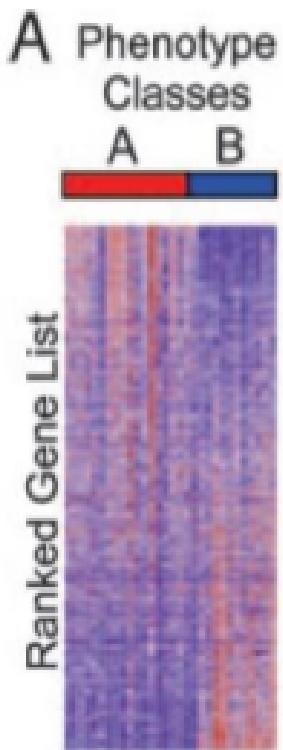


A schematic diagram comparing gene set analysis (GSA) with individual gene analysis (IGA). IGA is a two-step process which first selects some tens to hundreds of genes by an arbitrarily chosen cutoff and then, from the selected genes, infers the biological meaning of the gene expression data. In contrast, GSA is single-step process which in advance prepares gene sets from diverse sources as a testable hypothesis and then directly infers the biological meaning of gene expression data by applying either a sample or a gene randomization test.

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 a running sum is computed.
 - Let $N = \#$ genes in the array, $G = \#$ genes in the gene set.
 - If a gene belongs to the gene set a quantity is added
 - If a gene does not belong to gene set a quantity is subtracted
 - If there is no concentration of genes belonging to the gene set (this appear at random) the random sum behaves as a random walk
 - If, instead, genes in the gene set tend to be more abundant in the top part of the list the running sum will tend to increase deviating from the random walk distribution.
- The distribution of the running sum is compared with that of the random walk using a Kolmogorov-Smirnov

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! ;)

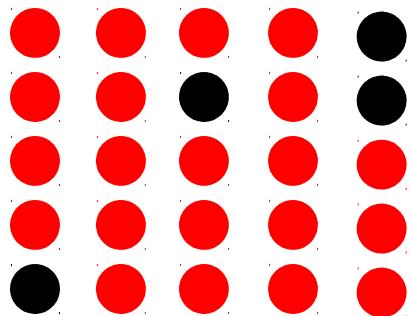
GSEA variants

- 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 corrections

How to win the P-value lottery

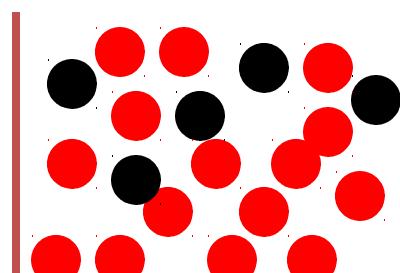
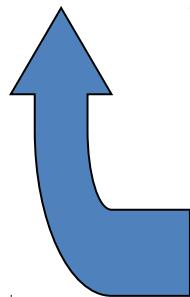
Random draws



... 7,834
draws later ...



Expect a random draw with observed enrichment once every : “1 / P-value” draws



Background population:
500 black genes,
4500 red genes

Simple P-value correction: *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 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</i>	0.001
2	<i>regulation</i>	0.002
3	<i>Transcription</i>	0.003
4	<i>factor</i>	0.0031
5	<i>Initiation of transcription</i>	0.005
...		...
	<i>Nuclear</i>	
52	<i>localization</i>	0.97
53	<i>Chromatin modification</i>	0.99
...		
	<i>Cytoplasmic localization</i>	
	<i>Translation</i>	

Benjamini-Hochberg example II

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

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

$$\text{Adjusted P-value} = \text{P-value} \times [\# \text{ of tests}] / \text{Rank}$$
$$0.99 \times 53/53 = 0.99$$

Benjamini-Hochberg example III

Rank	Category	(Nominal) P-value	Adjusted P-value	FDR / Q-value
1	<i>Transcriptional regulation</i>	0.001	0.001	$0.001 \times 53/1 = 0.040$
2	<i>Transcription factor</i>	0.002	0.053	0.040
3	<i>Initiation of transcription</i>	0.003	0.002	$0.002 \times 53/2 = 0.040$
4	<i>Nuclear localization</i>	0.0031	0.053	0.040
5	<i>Chromatin modification</i>	0.005	0.003	$0.003 \times 53/3 = 0.053$
...	<i>Cytoplasmic localization</i>	...	0.053	...
52	<i>Translation</i>	0.97	0.040	0.99
53	...	0.99	$0.005 \times 53/5 = 0.053$	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 (Nominal)	Adjusted P-value	FDR / Q-value
1	<i>Transcriptional regulation</i>	0.001	0.001 $\times 53/1 =$	0.040
2	<i>Transcription factor</i>	0.002	0.053	0.040
4	<i>Initiation of transcription</i>	0.0031	0.002 $\times 53/2 =$	0.040
5	<i>Nuclear localization</i>	0.005	0.053	0.053
...	<i>Chromatin modification</i>	...	0.0031 $\times 53/3 =$...
52	<i>Cytoplasmic localization</i>	0.97	0.040	0.99
53	<i>Translation</i>	0.99	0.005 $\times 53/5 =$ 0.053	0.99
			...	Red: non-significant
			0.985 $\times FDR/520.05$	Green: significant at
			1.004	$\times FDR/520.05$
			0.99 $\times 53/53 =$	0.99
			0.99	0.99

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

Reducing correction stringency

- The correction 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:***
 - e.g. use GO slim;
 - restrict testing to the appropriate GO annotations;
 - or 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 3.4 (Release)

Autocomplete biocViews search:

DriverMutation (1)
FunctionalPrediction (6)
GenePrediction (3)
GeneRegulation (45)
GeneSetEnrichment (74)
GeneSignaling (2)
GeneTarget (10)
GenomeAnnotation (27)
GenomeAssembly (1)
GenomeWideAssociation (14)
GenomicVariation (19)
HistoneModification (1)
LinkageDisequilibrium (4)
MotifAnnotation (12)

Packages found under GeneSetEnrichment:

Show	All	▼ entries	Search table:	<input type="text"/>
Package	▲	Maintainer	Title	◀ ▶
ABAEnrichment		Steffi Grote	Gene expression enrichment in human brain regions	
anamiR		Ti-Tai Wang	An integrated analysis package of miRNA and mRNA expression data	
AtlasRDF		Simon Jupp	Gene Expression Atlas query and gene set enrichment package.	
attract		Samuel Zimmerman	Methods to Find the Gene Expression Modules that Represent the Drivers of Kauffman's Attractor Landscape	
BgeeDB		Andrea Komljenovic, Frederic Bastian	Annotation and gene expression data retrieval from Bgee database	
CAFE		Sander Bollen	Chromosomal Aberrations Finder in Expression data	
Category		Bioconductor Package Maintainer	Category Analysis	

As of March 2017 there are 74 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)

BABELOMICS (FATIGO et alt.)

PPR Network enrichment analysis. Estimates, on the one hand, the role of the list within the interactions and, on the other hand, the comparative behaviour of the list as a subnetwork module.

- Set enrichment analysis**
 - Gene set analysis
 - PPR gene sets with common annotations overrepresented associated to high and/or low values in a ranked gene list by gene-set analysis
 - MannetreeScan
 - Implements gene-set analysis in the context of associations obtained by interactome
 - GelGOAP
 - Gene set analysis (also known as Pathway Based Analysis) in the context of genome-wide association analysis (GWAS) with SNPs or CNVs
 - NetworKINer
 - PPR Subnetwork enrichment analysis. Finds significant subnetworks of protein-protein interactions within a list of ranked gene/proteins
- Module enrichment analysis**
 - Genecos
 - Find annotations that frequently co-occur in a list of genes and rank them by statistical significance
- Tissue phenotype based profiling**
 - Affymetrix
 - Single enrichment analysis using expression values from public domain experiments (Affymetrix chips)
 - SAGE
 - Single enrichment analysis using expression values from public domain experiments of SAGE
- Functional annotation**
 - BiNGO
 - It's a tool for functional annotation of gene lists and the analysis of association data

BIOINFORMATICS APPLICATIONS NOTE

FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes

Fátima Al-Shabani, Ramón Díez-Úbeda and Joaquín Dopazo¹

¹Centromix SA, Centro de Investigación en Oncología (CINO), Instituto de Biología (IB) (UPM-CSIC), Madrid, Spain

Received 14 August 2006; revised 10 September 2006; accepted 10 October 2006

ABSTRACT

Summary: We present a simple but powerful web-based tool to assist Gene Ontology (GO) terms that are overrepresented in a group of genes. This tool uses expression data from public domain experiments (Affymetrix microarrays, etc.). Gene expression has been increasing as a web application, FatiGO, allowing for easy and interactive gene expression analysis. In this paper we describe how FatiGO performs GO enrichment analysis. We also describe how FatiGO can be used to identify significantly overrepresented GO terms for various expression datasets, namely, the Affymetrix and TIGR datasets (GO enrichment). FatiGO is available at http://www.centromex.com/~fati/go_enrichment.html. It is a user friendly, intuitive and fast tool for GO enrichment analysis. © 2006 Bioinformatics Research Group (IB) (UPM-CSIC), Madrid, Spain. Published online in *bioRxiv* on 10 October 2006. This version of the article may be used for reference and quoting purposes only. The final published version (in the journal *bioinformatics*) is available at <http://bioinformatics.oxfordjournals.org/>.

Gene list

Are targets for a specific regulator over-represented in my gene list with respect to the normal regulation in the genome?

Genome

● GATA1 = 40%	● GATA1 = 10%
● SP1 = 20%	● SP1 = 20%

The gene list is over-enriched for GATA1 targets.

	Gene list	Genome
Regulated by GATA1	4	2
Not regulated by GATA1	6	18

- Takes two lists of genes (ideally a group of interest and the rest of the genes in the experiment, although any two groups, formed in any way, can be tested against each other)
- Convert them into two lists of gene sets using the corresponding gene-gen set association table. The gene sets are functional termed grouped by Gene Ontology, KEGG pathways, InterPro motifs, Swissprot keywords, microRNA, TFBSs, BioCarta pathways and cisRED motifs.
- Then a Fisher's exact test for 2x2 contingency tables is used to check for significant overrepresentation of gene sets in one of the sets with respect to the other one.
- Multiple test correction to account for the multiple hypothesis tested (one for each term) is applied as previously described.

Upload gene list, set parameters

Define your comparison

Id list vs Id list
 Id List vs Rest of genome
 Id List vs Rest of ids contained in your annotations (complementary list)

Select your data

List 1: [browse server](#) Entrez from CinRn_vs_Cinlni

List 2: Rest of genome

Options

Filter exact test: [etailed](#)

Remove duplicates? [Remove on each list separately](#)

Databases

Organism: Human (homo sapiens)

GO biological process [\[options\]](#)
 GO molecular function [\[options\]](#)
 GO cellular component [\[options\]](#)
 GOSlim GOA [\[options\]](#)
 Interpro [\[options\]](#)
 KEGG pathways [\[options\]](#)
 Reactome [\[options\]](#)
 Biocarta [\[options\]](#)
 miRNA targets [\[options\]](#)
 Jaspar TFBS [\[options\]](#)

Your annotations [browse server](#) no data selected. This data format is:
Or go to Upload Data form: [Upload \[annotation\]](#)

Job

job name: [CinRn_vs_Cinlni from Entrez](#)

job description: [Full_topTable_entrezIds vs human genome](#)

Run

Obtain “significantly enriched” sets

Define your comparison

Id list vs Id list
 Id List vs Rest of genome
 Id List vs Rest of ids contained in your annotations (complementary list)

Select your data

List 1: [browse server](#) Entrez from CinRn_vs_Cinlni

List 2: Rest of genome

Options

Fisher exact test: [Two tailed](#)

Remove duplicates? [Remove on each list separately](#)

Databases

Organism: Human (homo sapiens)

GO biological process [\[options\]](#)
 GO molecular function [\[options\]](#)
 GO cellular component [\[options\]](#)
 GO Slim GOA [\[options\]](#)
 Interpro [\[options\]](#)
 KEGG pathways [\[options\]](#)
 Reactome [\[options\]](#)
 Biocarta [\[options\]](#)
 miRNA targets [\[options\]](#)
 Jaspar TFBS [\[options\]](#)

[browse server](#) no data selected. This data format is:
Or go to Upload Data form: [Upload \[annotation\]](#)

Job

job name: [CinRn_vs_Cinlni from Entrez](#)
job description: [Full topTable entrezIds vs human genome](#)

Input data

Species: hsa
Duplicates management: [Remove list 1 ids from genome](#)
Fisher exact test: [Two tailed](#)
List 1 (after duplicates managing): [clean_list1.txt](#)
Genome (after duplicates managing): [clean_list2.txt](#)

Summary

Id annotations per DB:

DB	Count	Description
GO biological process (levels from 3 to 9)	3581	of 6813 (52.56%) 9.88 annotations/id
GO cellular component (levels from 3 to 9)	2399	of 6813 (35.21%) 1.38 annotations/id
GO molecular function (levels from 3 to 9)	3446	of 6813 (50.58%) 3.28 annotations/id
KEGG	1761	of 6813 (25.85%) 0.79 annotations/id
miRNA target	4805	of 6813 (70.53%) 23.3 annotations/id
Genome	8902	of 16805 (52.97%) 6.31 annotations/id
	5237	of 16805 (31.16%) 0.94 annotations/id
	8690	of 16805 (51.71%) 2.46 annotations/id
	3405	of 16805 (20.26%) 0.51 annotations/id
	12339	of 16805 (73.42%) 19.17 annotations/id

Duplicates management:

Detail	List 1	Genome
Number of duplicates	870 of 7683 (0.11%)	6816 of 23621 (0.29%)
Number of finally used ids	6813	16805

Significant Results

Number of significant terms per DB:

DB	Number of significant terms
GO biological process (levels from 3 to 9)	641
GO cellular component (levels from 3 to 9)	56
GO molecular function (levels from 3 to 9)	126
KEGG	76
miRNA target	455

Visualize results

KEGG

KEGG significant terms (pvalue<0.05) : [significant_kegg_0.05.txt](#) ([download as EXCEL](#))

Term	Term size	Term size (in genome)	Term annotation % per list	Annotated ids	Odds ratio (log _e)	pvalue	Adjusted pvalue
Oxidative phosphorylation (hsa00190)	139	134	list 1: 1% list 2: 0.42%	list 1: 245973,1349,106... list 2: ENSG000000006695,ENSG...	0.8655	4.414e-7	0.000009516
Epithelial cell signaling in Helicobacter pylori infection (hsa05120)	77	74	list 1: 0.53% list 2: 0.24%	list 1: 245973,4602,323... list 2: ENSG00000070831,ENSG...	0.7756	0.0006437	0.003122
MAPK signaling pathway (hsa04010)	300	289	list 1: 1.95% list 2: 0.99%	list 1: 5578,11184,5159... list 2: ENSG000000006283,ENSG...	0.6849	7.031e-9	2.273e-7
ErbB signaling pathway (hsa04012)	91	88	list 1: 0.53% list 2: 0.30%	list 1: 5578,53358,3236... list 2: ENSG000000051382,ENSG...	0.481	0.018	0.04655
Calcium signaling pathway (hsa04020)	190	176	list 1: 1.28% list 2: 0.61%	list 1: 5578,5159,89832... list 2: ENSG00000004468,ENSG...	0.7407	4.971e-7	0.000009643
Wnt signaling pathway (hsa04310)	163	156	list 1: 0.92% list 2: 0.6%	list 1: 5578,1144,7473... list 2: ENSG00000002745,ENSG...	0.4441	0.00435	0.01507
VEGF signaling pathway (hsa04370)	78	74	list 1: 0.54% list 2: 0.24%	list 1: 5578,5534,9317... list 2: ENSG000000051382,ENSG...	0.8032	0.0003834	0.002188
Focal adhesion (hsa04510)	110	106	list 1: 1.54% list 2: 0.68%	list 1: 5578,3371,55742... list 2: ENSG00000017427,ENSG...	0.8205	2.188e-9	1.415e-7
Gap junction (hsa04540)	62	59	list 1: 0.69% list 2: 0.31%	list 1: 5578,5159,114,5... list 2: ENSG00000061918,ENSG...	0.8056	0.00006547	0.0005522
Long-term potentiation (hsa04424)	100	96	list 1: 0.54% list 2: 0.23%	list 1: 5578,114,5534,2... list 2: ENSG000000005339,ENSG...	0.8533	0.0001996	0.001249

Search the term hsa04510 Focal adhesion

General databases
[Ensembl](#) [novoseek](#)

Functional databases
[KEGG](#)

Other info
hsa04510 pathway description

hsa04510 pathway description

GO molecular function (levels from 3 to 9) significant terms (pvalue<0.05) : [significant_go_molecular_function_3_9_0.05.txt](#) ([download as EXCEL](#))

Visualize results

KEGG

KEGG significant terms (pvalue<0.05) : [significant_kegg_0.05.txt](#) ([download as EXCEL](#))

Term	Term size
Oxidative phosphorylation (hsa00190)	139
Epithelial cell signaling in Helicobacter pylori infection (hsa05120)	77
MAPK signaling pathway (hsa04010)	300
ErbB signaling pathway (hsa04012)	91
Calcium signaling pathway (hsa04020)	190
Wnt signaling pathway (hsa04310)	163
VEGF signaling pathway (hsa04370)	78
Focal adhesion (hsa04510)	Search the term hsa04510
Gap junction (hsa04540)	General databases Ensembl novoseek
Long-term potentiation (hsa04420)	Functional databases KEGG Other info hsa04510 pathway diagram

Pathway menu | Organism menu | Pathway entry | [Download GML](#) | Show description | User data mapping | Homo sapiens (human) | Go | 100% |

FOCAL ADHESION

The diagram illustrates the Focal Adhesion pathway (hsa04510) in Homo sapiens. Key components include:

- ECM-receptor interaction:** ECM binds to ITGB, which activates integrins (ITGA, ITGB). Integrins activate FAK, PI3K, and Rac.
- RhoA signaling:** RhoA is activated by RhoGEF and inhibited by RhoGAP. RhoA activates ROCK, which in turn activates MLC2 and mDia1.
- PI3K-Akt signaling:** PI3K is activated by FAK, PTEN, and PI3KCA. Akt/PKB activates GSK-3β and Cdk4/Cdk2.
- MAPK signaling:** FAK activates MEK1/2, which activates ERK1/2. ERK1/2 activates Elk1, which in turn activates c-Jun and JNK.
- Actin dynamics:** Actin is polymerized by MLCP and depolymerized by MLCK. It is also regulated by mDia1 and ROCK.
- Wnt signaling:** GSK-3β inhibits β-Catenin, which is part of the Wnt signaling pathway.
- Cell cycle and survival:** Cdk4/Cdk2, c-Jun, and Elk1 regulate cell cycle genes (CycD, Cyclin E) and survival genes (c-Myc, Bcl-2, Bad).

GO molecular function (levels from 3 to 9) significant terms (p < 0.05):

04510 8/13/14
(c) Kanehisa Laboratories

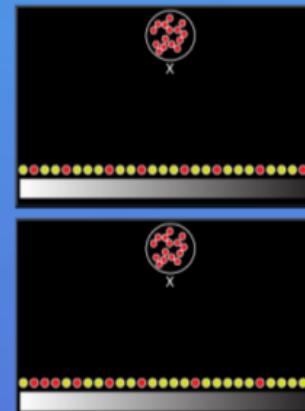
“Official” GSEA. BROAD Institute

The screenshot shows the GSEA homepage. It features a main navigation bar with links to "GSEA Home", "Downloads", "Molecular Signatures Database", "Documentation", and "Contact". The "GSEA Home" link has a small "register" button next to it. Below the navigation is a large diagram illustrating the GSEA process: "Molecular Profile Data" (represented by a scatter plot) is input into the "Run GSEA" step, which then produces "Enriched Sets" (represented by a bar chart). A "Gene Set Database" icon is also shown. To the left, there's a "Overview" section with a brief description of GSEA, a "What's New" section with a note about version 2.0.14, and a "Registration" section with instructions for creating an account.

This is a screenshot of a news article from the journal "nature genetics". The title of the article is "MSigDB: Molecular Signatures Database". The article discusses the creation of MSigDB, which contains gene sets derived from various sources like pathway databases and publications. It highlights the use of MSigDB in a study of gene expression profiles in human disease.

The screenshot shows the MSigDB homepage. At the top, it says "Molecular Signatures Database v4.1". Below that is an "Overview" section with a brief description of MSigDB as a collection of annotated gene sets. The page then lists seven major collections: **c1 positional gene sets** (for each human chromosome and cytobands), **c2 curated gene sets** (from online pathway databases, publications in BioMart, and knowledge of domain experts), **c3 multi-gene sets** (based on conserved cis regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes), **c4 computational gene sets** (defined by mining large collections of cancer-associated transcriptome data), **c5 50 gene sets** (consist of genes annotated by the same GO term), **c6 oncogenic signatures** (defined directly from transcriptome gene expression data from cancer gene participants), and **c7 inactivating signatures** (defined directly from transcriptome gene expression data from inactivating mutants).

- No cut-off, uses "all" genes ranked
- For each functional annotation
 - Are genes randomly distributed in ranked list?
or
 - Are genes distributed towards the top/bottom?
- Calculate enrichment score (ES)
- Calculate significance of ES
- Correct for multiple testing



Upload data matrix (not gene list!)

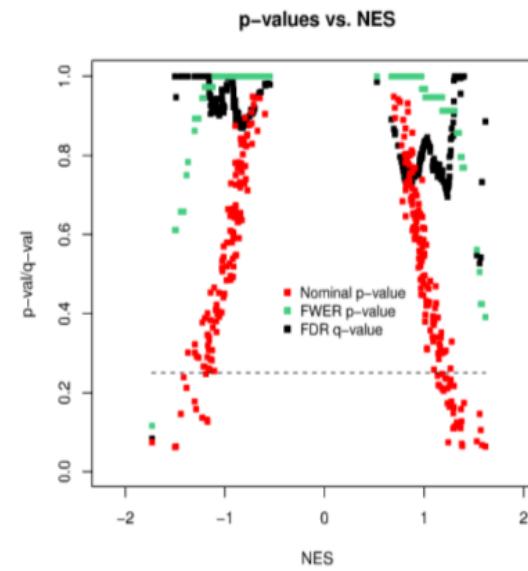
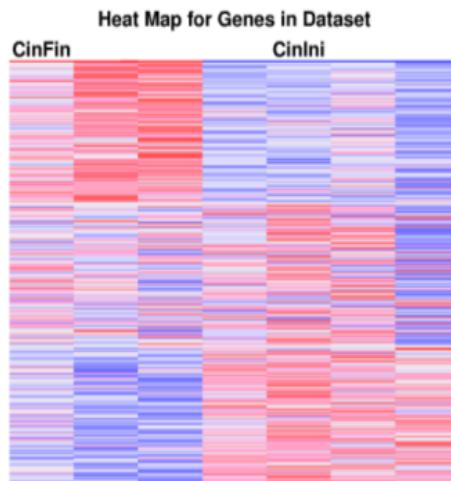
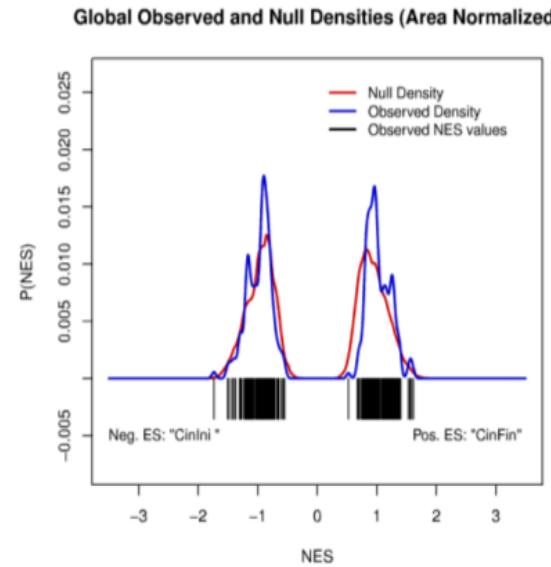
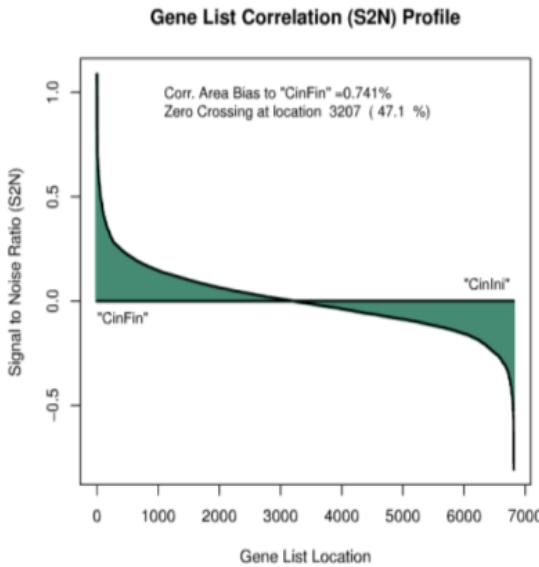
#1.2											
NAME	Description	Cf9	Cf10	Cf11	Cf4	Cf5	Cf7	Cf8			
18 ABAT		6.7100903124	6.2010968359	6.597489765	6.5013934729	6.8451356349	7.5118595998	6.4110123122			
19 ABCA1		5.4530779893	5.4355752984	5.4610085993	5.5476548331	5.4520121732	6.0489388288	5.0407680213			
22 ABCB7		7.3694752885	6.8977680549	7.3073544278	7.3465162227	7.5633266456	7.7642116686	7.1601794575			
30 ACAA1//OTTHUMG000001559		4.7248754828	5.0149574378	4.5595306969	4.706261494	4.8956576824	5.2365107743	4.257629554			
34 ACADM		8.9297666695	8.7971612382	8.5892949897	9.0426688459	9.6077263301	9.1942861042	8.6120161117			
36 ACADSB		5.7822772441	5.2725699164	6.0307174265	5.8207351756	6.0645458856	6.4805674854	5.7841453389			
39 ACAT2//LOC100129518//SOD2		6.9231386649	6.5549055906	6.4733739317	7.0506799232	7.4855891211	6.8006708513	5.9320368941			
41 ASIC1		4.7737457913	4.6841654133	5.1766921232	4.7347939833	4.4541148569	5.1399967599	4.7122823994			
43 ACHE		4.1624435064	3.6882333994	3.9935732941	3.7993435183	2.8043924676	3.6560631404	3.7848718218			
51 ACOX1		6.7826908205	7.0496507119	7.0241582248	6.4457072968	6.5778707815	6.8922885681	6.3313127929			
52 ACP1		7.4211781076	7.097706037	6.9938476336	7.5431622044	7.746785656	7.6525873676	7.0653335002			
53 ACP2		5.059193215	4.8278872873	5.1630183499	4.7642311305	4.3365637313	4.8455273799	4.4079333028			
54 ACP5		3.37202491	3.7503157578	3.7381985644	4.6264847891	3.4352961728	3.843873691	3.5304751138			
58 ACTA1		3.5406402898	3.9470695083	2.5025194518	2.7164043575	3.3941808636	3.7235832021	3.0603353631			
60 ACTB		6.2064344458	5.6953236783	5.7415926085	6.8800198082	5.7209218115	5.5140222137	6.9150639376			
71 ACTG1		10.9730301369	11.0722992954	10.7531966455	10.7612095627	10.6072998343	11.01349861	10.3146970978			
86 ACTL6A		7.306880239	7.3581695226	7.0414218102	7.3981514689	7.9485237615	7.4044889147	6.869460735			
88 ACTN2		4.8202948791	4.9049509308	4.7554192235	4.9223164025	5.2463745747	5.11490026	4.630485298			
90 ACVR1		5.0631062492	5.2052068044	4.6386208381	4.7515376273	4.9948847171	4.7537831842	4.3205454954			
94 ACVRL1		4.0012691816	3.5307672326	3.5967210171	4.4299320674	3.4079464211	4.335039134	3.5205318489			
97 ACYP1		5.7893977955	6.0223275581	6.3080750412	6.0767124941	6.1296880315	5.3047725532	5.8596962219			
100 ADA		3.283936647	3.5490351114	3.5513605652	3.5467692743	3.2251494559	3.5644893261	3.2357167769			
107 ADCY1		7.9872080552	7.5776132191	7.2672454373	7.3952269045	7.1706457374	7.7049262511	7.3400486978			
111 ADCY5		5.1798093059	5.0993243605	5.6976263009	5.3114479241	5.3256353917	5.6350976706	4.9818460015			
112 ADCY6//MIR4701		5.3929438612	5.4234156153	5.7267878509	5.8827262514	5.2677475971	5.4827914174	5.0904796544			

Set analysis parameters and gene sets

#1.2	NAME	Description	Cf9	Cf10	Cf11	Cf4	Cf5	Cf7	Cf8
6811			6.7100903124	6.2010968359	6.597489765	6.5013934729			
18 ABAT			5.4530779893	5.4355752984	5.4610085993	5.5475648331			
19 ABCA1			7.3694752885	6.8977680549	7.3073544278	7.3465162227			
22 ABCB7			4.7248754828	5.0149574378	4.5595306969	4.706261494			
30 ACAA1//OTTHUMG000001559			8.9297666695	8.7971612382	8.5892949897	9.0426688459			
34 ACADM			5.7822772441	5.2725699164	6.0307174265	5.8207351756			
36 ACADSB			6.9231386649	6.5549055906	6.4733739317	7.0506799232			
39 ACAT2//LOC100129518//SOD2			4.7737457913	4.6841654133	5.1766921232	4.7347939833			
41 ASIC1			4.1624435064	3.6882333994	3.9935732941	3.7993435183			
43 ACHE			6.7826908205	7.0496507119	7.0241582248	6.4457072968			
51 ACOX1			7.4211781076	7.097706037	6.9938476336	7.5431622044			
52 ACP1			5.059193215	4.8278872873	5.1630183499	4.7642311305			
53 ACP2			3.37202491	3.7503157578	3.7381985644	4.6264847891			
54 ACP5			3.5406402898	3.9470695083	2.5025194518	2.7164043575			
58 ACTA1			6.2064344458	5.6953236783	5.7415926085	6.8800198082			
60 ACTB			10.9730301369	11.0722992954	10.7531966455	10.7612095627			
71 ACTG1			7.306880239	7.3581695226	7.0414218102	7.3981514689			
86 ACTL6A			4.8202948791	4.9049509308	4.7554192235	4.9223164025			
88 ACTN2			5.0631062492	5.2052068044	4.6386208381	4.7515376273			
90 ACVR1			4.0012691816	3.5307672326	3.5967210171	4.4299320674			
94 ACVRL1			5.7893977955	6.0223275581	6.3080750412	6.0767124941			
97 ACYP1			3.283936647	3.5490351114	3.5513605652	3.5467692743			
100 ADA			7.9872080552	7.5776132191	7.2672454373	7.3952269045			
107 ADCY1			5.1798093059	5.0993243605	5.6976263009	5.3114479241			
111 ADCY5			5.3929438612	5.4234156153	5.7267878509	5.8827262514			

```
Run.A276bis.R x | Source on Save Run Source
1 # GSEA 1.0 -- Gene Set Enrichment Analysis / Broad Institute
2 #
3 # R script to run custom GSEA analysis of the UEB study ID #A276,
4 # based on the R script to run GSEA Analysis of the Leukemia ALL/AML vs C1 example
5
6 GSEA.program.location <- "/home/ferran/gsea_home/GSEA-P-R/GSEA.1.0.R" # R source program
7 source(GSEA.program.location, verbose=T, max.deparse.length=9999)
8
9 GSEA(                                     # Input/Output Files :
10 # input.ds = "/home/ferran/estudios/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSEA/Human"
11 input.ds = "/home/ferran/estudios/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSEA/Human"
12 input.cls = "/home/ferran/estudios/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSEA/VehFi"
13 # gs.db = "/home/ferran/estudios/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSEA/geneSets"
14 gs.db = "/home/ferran/gsea_home/GSEA-P-R/GeneSetDatabases/c2.all.v4.0.entrez.gmt",      # Gene set
15 # output.directory = "/home/ferran/estudios/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GS"
16 output.directory = "/home/ferran/estudios/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSE
17 # Program parameters :-
18 doc.string          = "A276.VehFin.vs.VehInit",    # Documentation string used as a prefix to name res
19 non.interactive.run = F,                            # Run in interactive (i.e. R GUI) or batch (R command line) m
20 reshuffling.type   = "sample.labels",             # Type of permutation reshuffling: "sample.labels" or "gene.l
21 nperm               = 1000,                         # Number of random permutations (default: 1000)
22 weighted.score.type = 1,                           # Enrichment correlation-based weighting: 0=no weight (KS), 1
23 nom.p.val.threshold = -1,                          # Significance threshold for nominal p-vals for gene sets (de
24 fwer.p.val.threshold = -1,                         # Significance threshold for FWER p-vals for gene sets (defau
25 fdr.q.val.threshold = 0.25,                        # Significance threshold for FDR q-vals for gene sets (defau
26 topgs               = 10,                           # Besides those passing test, number of top scoring gene sets
27 adjust.FDR.q.val   = F,                           # Adjust the FDR q-vals (default: F)
28 gs.size.threshold.min = 25,                         # Minimum size (in genes) for database gene sets to be consid
29 gs.size.threshold.max = 1500,                       # Maximum size (in genes) for database gene sets to be consid
30 reverse.sign        = F,                           # Reverse direction of gene list (pos. enrichment becomes neg
31 preproc.type        = 0,                           # Preproc.normalization: 0=none, 1=col(z-score).., 2=col(rank)
32 random.seed         = 123456,                      # Random number generator seed. (default: 123456)
33 perm.type           = 0,                           # For experts only. Permutation type: 0 = unbalanced, 1 = bal
34 fraction            = 1.0,                         # For experts only. Subsampling fraction. Set to 1.0 (no resa
35 replace              = F,                           # For experts only. Resampling mode (replacement or not repla
36 save.intermediate.results = F,                  # For experts only, save intermediate results (e.g. matrix of
37 OLD.GSEA             = F,                           # Use original (old) version of GSEA (default: F)
38 use.fast.enrichment.routine = T,                 # Use faster routine to compute enrichment for random permuta
39 )
40 #
41
42
1:1 (Top Level) R Script
```

Get results. Interpret output



DAVID

Gene-annotation enrichment:
typical batch annotation and gene-GO term
enrichment analysis to highlight the most relevant
GO terms associated with a given gene list.

- Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* 2009;4(1):44-57. [PubMed]
- Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009;37(1):1-13. [PubMed]

Pathway mapping / Pathway Viewer:
can display genes from a user's list on KEGG
and BioCarta pathway maps to facilitate
biological interpretation in a network context.

Functional Annotation Clustering:
measures relationships among the annotation
terms based on the degrees of their co-association
genes to group the similar, redundant, and
heterogeneous annotation contents from the same
or different resources into annotation groups.

The screenshot shows the DAVID Bioinformatics Resources 6.7 homepage. The top navigation bar includes links for Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, Why DAVID?, and About. A sidebar on the left titled "Shortcut to DAVID Tools" lists four options: Functional Annotation, Gene Functional Classification, Gene ID Conversion, and Gene Name Batch Viewer. The main content area features a red banner stating "Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!" Below this is a section titled "Welcome to DAVID 6.7" with a timeline from 2003 to 2014. To the right, there are two sections: "What's Important in DAVID" (listing features like current release, new requirements, and novel classification algorithms) and "Statistics of DAVID" (a bar chart showing citation counts from 2004 to 2010, with a note that DAVID has over 10,000 citations).

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About

Shortcut to DAVID Tools

- Functional Annotation
- Gene Functional Classification
- Gene ID Conversion
- Gene Name Batch Viewer

Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2014

What's Important in DAVID

- Current (v 6.7) release now
- New requirement to cite DAVID
- IDs of Affy Exon and Germline supported
- Novel Classification Algorithms
- Pre-built Affymetrix and RNA-seq backgrounds
- User's customized gene lists
- Enhanced calculating speed

Statistics of DAVID

DAVID Bioinformatic Resource

3182

0 2004 05 06 07 08 09 2010

> 10,000 Citations

Upload gene lists. Define background

 DAVID BIOINFORMATICS DATABASE

Functional Annotation Tool
DAVID Bioinformatics Resources 6.7, NIAID/NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

Upload List Background

Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -
Homo sapiens(159)
Unknown(5)

Select Species

List Manager [Help](#)

demolist1

Select List to:

[View Unmapped Ids](#)

Annotation Summary Results

Current Gene List: demolist1 155 DAVID IDs [Help and Tool Manual](#)

Current Background: Homo sapiens Check Defaults

Disease (1 selected)
 Functional_Categories (3 selected)
 Gene_Ontology (3 selected)
 General Annotations (0 selected)
 Literature (0 selected)
 Main_Accessions (0 selected)
 Pathways (3 selected)
 Protein_Domains (3 selected)
 Protein_Interactions (0 selected)
 Tissue_Expression (0 selected)

Red annotation categories denote DAVID defined defaults

Combined View for Selected Annotation

Results

Functional Annotation Clustering

[Help and Manual](#)

Current Gene List: demolist1

Current Background: Homo sapiens

155 DAVID IDs

Options Classification Stringency Medium

[Rerun using options](#) [Create Sublist](#)

72 Cluster(s)

[Download File](#)

Annotation Cluster 1		Enrichment Score: 4.81	G		Count	P Value	Benjamini
<input type="checkbox"/>	UP_SEQ_FEATURE	signal peptide	RT		50	6.5E-7	4.2E-4
<input type="checkbox"/>	SP_PIR_KEYWORDS	signal	RT		50	8.6E-7	2.8E-4
<input type="checkbox"/>	UP_SEQ_FEATURE	disulfide bond	RT		45	1.2E-6	4.0E-4
<input type="checkbox"/>	SP_PIR_KEYWORDS	disulfide bond	RT		46	1.7E-6	2.7E-4
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular region	RT		40	6.9E-6	1.5E-3
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular region part	RT		24	3.8E-5	4.0E-3
<input type="checkbox"/>	SP_PIR_KEYWORDS	Secreted	RT		29	7.2E-5	4.6E-3
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular space	RT		19	9.4E-5	6.5E-3
<input type="checkbox"/>	SP_PIR_KEYWORDS	glycoprotein	RT		53	2.3E-4	7.5E-3
<input type="checkbox"/>	UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc..)	RT		48	1.6E-3	1.6E-1
Annotation Cluster 2		Enrichment Score: 2.64	G		Count	P Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	response to bacterium	RT		10	1.4E-4	9.1E-2
<input type="checkbox"/>	SP_PIR_KEYWORDS	antibiotic	RT		6	1.7E-4	7.9E-3
<input type="checkbox"/>	SP_PIR_KEYWORDS	Antimicrobial	RT		6	2.1E-4	8.5E-3
<input type="checkbox"/>	INTERPRO	Defensin_propeptide	RT		3	7.2E-4	2.5E-1
<input type="checkbox"/>	INTERPRO	Alpha_defensin	RT		3	7.2E-4	2.5E-1
<input type="checkbox"/>	INTERPRO	Alpha-defensin	RT		3	7.2E-4	2.5E-1
<input type="checkbox"/>	GOTERM_BP_FAT	defense response to bacterium	RT		7	8.9E-4	3.4E-1
<input type="checkbox"/>	PIR_SUPERFAMILY	PIRSF001875:alpha-defensin	RT		3	1.2E-3	1.2E-1
<input type="checkbox"/>	INTERPRO	Mammalian defensin	RT		3	2.0E-3	2.3E-1
<input type="checkbox"/>	SMART	DEFSN	RT		3	2.8E-3	2.5E-1
<input type="checkbox"/>	SP_PIR_KEYWORDS	fungicide	RT		3	3.0E-3	6.0E-2

Results

Functional Annotation Clustering

[Help and Manual](#)

Current Gene List

Current Background

155 DAVID IDs

Options

Class

Rerun using options

72 Cluster(s)

Annotation

UP_SEQ_FEAT

SP_PIR_KEYWORDS

UP_SEQ_FEAT

SP_PIR_KEYWORDS

GOTERM_CC_FAT

GOTERM_CC_FAT

SP_PIR_KEYWORDS

GOTERM_CC_FAT

SP_PIR_KEYWORDS

UP_SEQ_FEAT

Annotation

GOTERM_BP_FAT

SP_PIR_KEYWORDS

SP_PIR_KEYWORDS

INTERPRO

INTERPRO

INTERPRO

GOTERM_BP_FAT

PIR_SUPERFAMILY

INTERPRO

SMART

SP_PIR_KEYWORDS

Functional Annotation Chart

[Help and Manual](#)

Current Gene List: demolist1

Current Background: Homo sapiens

155 DAVID IDs

Options

Rerun Using Options Create Sublist

211 chart records

[Download File](#)

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	UP_SEQ_FEATURE	signal peptide	RT	50	32,3	6,5E-7	4,2E-4	
	SP_PIR_KEYWORDS	signal	RT	50	32,3	8,6E-7	2,8E-4	
	UP_SEQ_FEATURE	disulfide bond	RT	45	29,0	1,2E-6	4,0E-4	
	SP_PIR_KEYWORDS	disulfide bond	RT	46	29,7	1,7E-6	2,7E-4	
	GOTERM_CC_FAT	extracellular region	RT	40	25,8	6,9E-6	1,5E-3	
	GOTERM_CC_FAT	extracellular region part	RT	24	15,5	3,8E-5	4,0E-3	
	GOTERM_MF_FAT	oxygen binding	RT	6	3,9	3,8E-5	1,4E-2	
	SP_PIR_KEYWORDS	heme	RT	8	5,2	4,0E-5	4,3E-3	
	GOTERM_BP_FAT	iron	RT	11	7,1	6,9E-5	5,6E-3	
	SP_PIR_KEYWORDS	Secreted	RT	29	18,7	7,2E-5	4,6E-3	
	SP_PIR_KEYWORDS	extracellular space	RT	19	12,3	9,4E-5	6,5E-3	
	GOTERM_MF_FAT	heme binding	RT	8	5,2	1,0E-4	1,9E-2	
	SP_PIR_KEYWORDS	chromoprotein	RT	6	3,9	1,1E-4	5,9E-3	
	GOTERM_BP_FAT	defense response	RT	18	11,6	1,3E-4	1,7E-1	
	GOTERM_BP_FAT	response to bacterium	RT	10	6,5	1,4E-4	9,1E-2	
	GOTERM_MF_FAT	tetrapyrrole binding	RT	8	5,2	1,5E-4	1,9E-2	
	SP_PIR_KEYWORDS	antibiotic	RT	6	3,9	1,7E-4	7,9E-3	
	SP_PIR_KEYWORDS	Antimicrobial	RT	6	3,9	2,1E-4	8,5E-3	
	SP_PIR_KEYWORDS	chemotaxis	RT	6	3,9	2,3E-4	8,0E-3	
	SP_PIR_KEYWORDS	glycoprotein	RT	53	34,2	2,3E-4	7,5E-3	

Results

Functional Annotation Clustering

Current Gene List

Current Background

155 DAVID IDs

Options

Clustering

Rerun using option

72 Cluster(s)

Annotation

UP_SEQ_FEAT

SP_PIR_KEYW

UP_SEQ_FEAT

SP_PIR_KEYW

GOTERM_CC

GOTERM_CC

SP_PIR_KEYW

GOTERM_CC

SP_PIR_KEYW

UP_SEQ_FEAT

Functional Annotation Chart

[Help and Manual](#)

[Help and Manual](#)

Current Gene List

Current Background

155 DAVID IDs

Options

Clustering

Rerun Using Option

211 chart records

Sublist

Category

Functional Annotation Table

[Help and Manual](#)

Current Gene List: demolist1

Current Background: Homo sapiens

155 DAVID IDs

150 record(s)

[Download File](#)

	37166_at	3-hydroxyanthranilate 3,4-dioxygenase	Related Genes	Homo sapiens
GOTERM_BP_FAT	coenzyme metabolic process, oxidoreduction coenzyme metabolic process, vitamin metabolic process, water-soluble vitamin metabolic process, nicotinamide metabolic process, coenzyme biosynthetic process, vitamin biosynthetic process, nucleotide biosynthetic process, alkaloid metabolic process, response to inorganic substance, response to metal ion, response to zinc ion, organic acid biosynthetic process, pyridine nucleotide biosynthetic process, NAD metabolic process, cellular homeostasis, secondary metabolic process, quinolinate biosynthetic process, nucleobase, nucleoside and nucleotide biosynthetic process, nucleobase, nucleoside, nucleotide and nucleic acid biosynthetic process, water-soluble vitamin biosynthetic process, homeostatic process, cellular amide metabolic process, dicarboxylic acid metabolic process, nitrogen compound biosynthetic process, carboxylic acid biosynthetic process, nicotinamide nucleotide metabolic process, response to cadmium ion, quinolinate metabolic process, cofactor metabolic process, cofactor biosynthetic process, oxidation reduction, anatomical structure homeostasis, neuron maintenance,			
GOTERM_CC_FAT	cell fraction, soluble fraction, mitochondrion, mitochondrial envelope, cytosol, organelle membrane, mitochondrial membrane, organelle envelope, envelope, mitochondrial part,			
GOTERM_MF_FAT	3-hydroxyanthranilate 3,4-dioxygenase activity, iron ion binding, ferrous iron binding, electron carrier activity, oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen, oxygen binding, ion binding, cation binding, metal ion binding, transition metal ion binding,			
INTERPRO	3-hydroxyanthranilic acid dioxygenase, 3-hydroxyanthranilate 3,4-dioxygenase, metazoan,			
KEGG_PATHWAY	Tryptophan metabolism,			
PIR_SUPERFAMILY	PIRSF017681:3-hydroxyanthranilate 3,4-dioxygenase, animal type, PIRSF017681:3hydroanth_dOase_animal,			
SP_PIR_KEYWORDS	3d-structure, alternative splicing, complete proteome, cytoplasm, dioxygenase, iron, metal-binding, oxidoreductase, polymorphism, pyridine nucleotide biosynthesis,			
UP_SEQ_FEATURE	binding site:Dioxygen, binding site:Substrate, chain:3-hydroxyanthranilate 3,4-dioxygenase, helix, metal ion-binding site:iron; catalytic, sequence variant, splice variant, strand, turn,			
	34467_g_at	5-hydroxytryptamine (serotonin) receptor 4	Related Genes	Homo sapiens
GOTERM_BP_FAT	cell surface receptor linked signal transduction, G-protein coupled receptor protein signaling pathway, G-protein signaling, coupled to cyclic nucleotide second messenger, intracellular signaling cascade, second-messenger-mediated signaling, cyclic-nucleotide-mediated signaling,			
GOTERM_CC_FAT	cell fraction, membrane fraction, insoluble fraction, endosome, plasma membrane, integral to plasma membrane, integral to membrane, intrinsic to membrane, intrinsic to plasma membrane, plasma membrane part,			
GOTERM_MF_FAT	adrenoceptor activity, serotonin receptor activity, amine receptor activity,			
INTERPRO	7TM GPCR, rhodopsin-like, 5-Hydroxytryptamine 4 receptor, GPCR, rhodopsin-like superfamily,			
KEGG_PATHWAY	Calcium signaling pathway, Neuroactive ligand-receptor interaction,			
PIR_SUPERFAMILY	PIRSF038635:5-hydroxytryptamine receptor 4, PIRSF800006:rhodopsin-like G protein-coupled receptors,			
SP_PIR_KEYWORDS	alternative splicing, cell membrane, complete proteome, disulfide bond, endosome, G protein-coupled receptor, g-protein coupled receptor, glycoprotein, lipoprotein, membrane, neurotransmitter receptor, palmitate, polymorphism, receptor, transducer, transmembrane, transmembrane protein,			

Ingenuity Pathways

The screenshot displays the Ingenuity Pathway Analysis (IPA) website and its software interface.

Top Navigation: The top navigation bar includes links for **INGENUITY**, **PRODUCTS**, **SCIENCE**, **BLOG**, **LOGIN**, and a search icon. The QIAGEN logo is located in the top right corner.

Hero Section: A large image of a Newton's cradle with one green ball and several silver balls. Below it is a button labeled **WATCH A SHORT VIDEO**. To the right, there is a box for **CURRENT USER? LOGIN HERE**, the **INGENUITY PATHWAY ANALYSIS** logo, a **SIGN UP FOR IPA** button, and links for **SIGN UP TODAY**, **JOIN IPA LICENSE**, and **DOWNLOAD THE DATASHEET**.

Navigation Bar: Below the hero section is a horizontal navigation bar with links for **OVERVIEW**, **FEATURES**, **ADVANCED**, **APPLICATIONS**, **WEBINARS**, **TRAINING**, and **RESOURCES**.

Section: Model, analyze, and understand the complex biological and chemical systems at the core of life science research with IPA

Market Leading Pathway Analysis: Features a network graph icon and text explaining how it helps users unlock insights from experimental data by identifying relationships, mechanisms, functions, and pathways of relevance.

Predictive Causal Analytics: Features a network graph icon and text explaining how it helps users build a more complete regulatory picture and a better understanding of the biology underlying a given gene expression study.

NGS/RNA-Seq Data Analysis: Features a scatter plot icon and text explaining how it provides a better understanding of isoform-specific biology resulting from RNA-Seq experiments.

Bottom Left: A section titled "IPA is useful in a wide range of applications" featuring icons for **Biomarker Discovery**, **Metabolomics**, **MicroRNA Research**, and **NGS/RNA-Seq Data Analysis**, each with a brief description and a "LEARN MORE" button.

Bottom Right: A screenshot of the IPA software interface titled "Start Here". It shows sections for **Explore** (Datasets, Compare), **Analyze** (Core, IPA-Tox, IPA-Biomarker), and **Shortcuts**.

Exercise

- Obtain a **gene list** and a **background list** from a differential expression analysis (background may be the list of **all** genes analyzed) (see next slide)
- Convert identifiers into “Entrez” ids if they are not already converted.
- Select two pathway analysis tools e.g. DAVID and Babelomics
- Do a Gene Enrichment Analysis with each tool.
- Compare the 5-10 top enriched categories and comment about the differences.
- Alternatively do it with R/Bioconductor with the code from the following slides.

R code to prepare the data

```
topTab <- read.table("https://raw.githubusercontent.com/  
alexсанchezпла/scripts/master/Exemple_Analisis_BioC/results  
/ExpressAndTop_AvsB.csv2", head=TRUE, sep=";", dec=",")  
colnames(topTab)  
head(topTab)  
geneListUp <- unique(  
  topTab$EntrezsA [topTab$adj.P.Val<0.05 & topTab$logFC > 0] )  
length(geneListUp)  
geneListDown <- unique(  
  topTab$EntrezsA [topTab$adj.P.Val<0.05 & topTab$logFC < 0] )  
length(geneListDown)  
geneUniverse <- unique(topTab$EntrezsA)  
length(geneUniverse)  
write.csv(geneListUp, file="selectedAvsB.up.csv")  
write.csv(geneListDown, file="selectedAvsB.down.csv")  
write.csv(geneUniverse, file="geneUniverse.csv")
```

A quick ORA analysis with R

```
# GOAnalysis
require(GOstats)
## Creamos los "hiperparametros" en que se basa el analisis
GParams = new("GOHyperGParams",
              geneIds=geneListUp, universeGeneIds= geneUniverse,
              annotation="org.Hs.eg.db", ontology="BP",
              pvalueCutoff=0.001, conditional=FALSE, testDirection="over")
KEGGparams = new("KEGGHyperGParams",
                 geneIds=geneListUp, universeGeneIds= geneUniverse,
                 annotation="org.Hs.eg.db",      pvalueCutoff=0.01, testDirection="over")
## Ejecutamos los analisis
GOhyper = hyperGTest(GParams)
KEGGhyper = hyperGTest(KEGGparams)
cat("GO\n"); print(head(summary(GOhyper)))
cat("KEGG\n"); print(head(summary(KEGGhyper)))
# Creamos un informe html con los resultados
GOfilename =file.path(paste("GOResults.",".html", sep=""))
KEGGfilename =file.path(paste("KEGGResults.",".html", sep=""))
htmlReport(GOhyper, file = GOfilename, summary.args=list("htmlLinks"=TRUE))
htmlReport(KEGGhyper, file = KEGGfilename, summary.args=list("htmlLinks"=TRUE))
```

Expected output

Gene to GO BP test for over-representation

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0043436	0.000	1.947	42	73	889	oxoacid metabolic process
GO:0019752	0.000	2.003	37	67	792	carboxylic acid metabolic process
GO:0006082	0.000	1.909	42	73	904	organic acid metabolic process
GO:0044710	0.000	1.503	196	245	4164	single-organism metabolic process
GO:0006629	0.000	1.842	47	78	1000	lipid metabolic process
GO:1900101	0.000	10.270	1	7	21	regulation of endoplasmic reticulum unfolded protein response
GO:0044255	0.000	1.843	36	60	757	cellular lipid metabolic process
GO:0006631	0.000	2.443	13	29	278	fatty acid metabolic process

Gene to KEGG test for over-representation

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04914	0.000	3.963	4	13	74	Progesterone-mediated oocyte maturation
04146	0.001	3.439	4	11	70	Peroxisome
05221	0.002	3.589	3	9	55	Acute myeloid leukemia
04910	0.003	2.460	7	15	128	Insulin signaling pathway
01100	0.003	1.538	48	66	912	Metabolic pathways
04114	0.009	2.492	5	11	92	Oocyte meiosis

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.

Acknowledgements

