An Introduction to Pathway Analysis

Alex Sánchez

Statistics and Bioinformatics Research Group
Statistics department, Universitat de Barelona

Statistics and Bioinformatics Unit
Vall d'Hebron Institut de Recerca

Outline

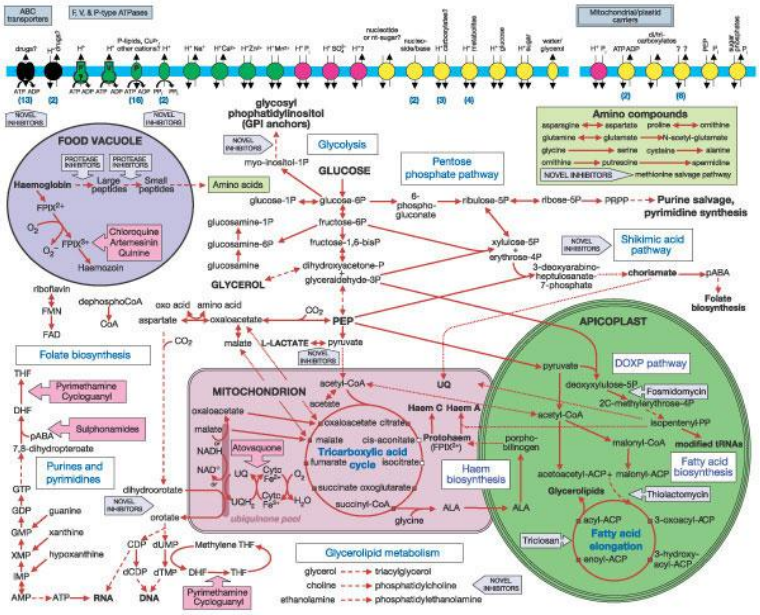
- Presentation
- Introduction and Background
- The problem: Interpreting gene lists
- Annotations and annotation databases
- The Gene Ontology Resource
- Gene list analysis using the GO and relatives
- Existing tools for pathway analysis

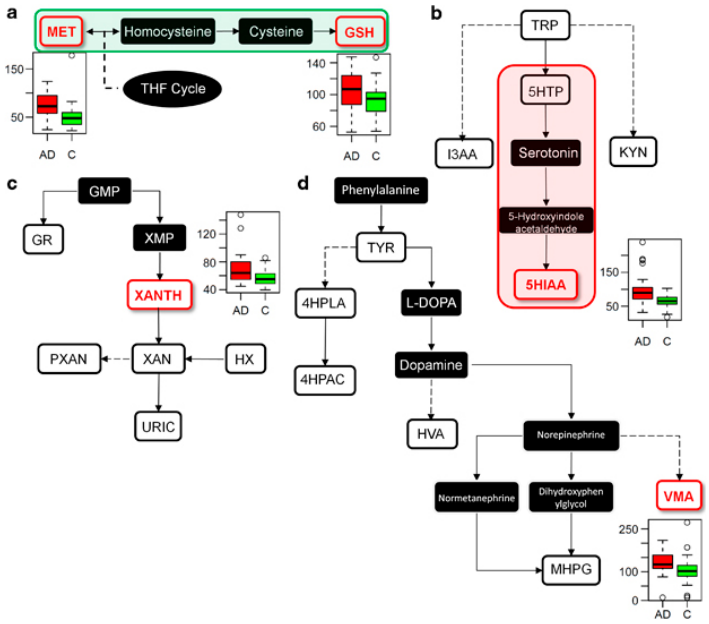
Introduction & Background

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 meta*bolism.





Pathways altered in ALZHEIMER disease

Characterization of disease can be attempted by studying how this affects or disrupts pathways

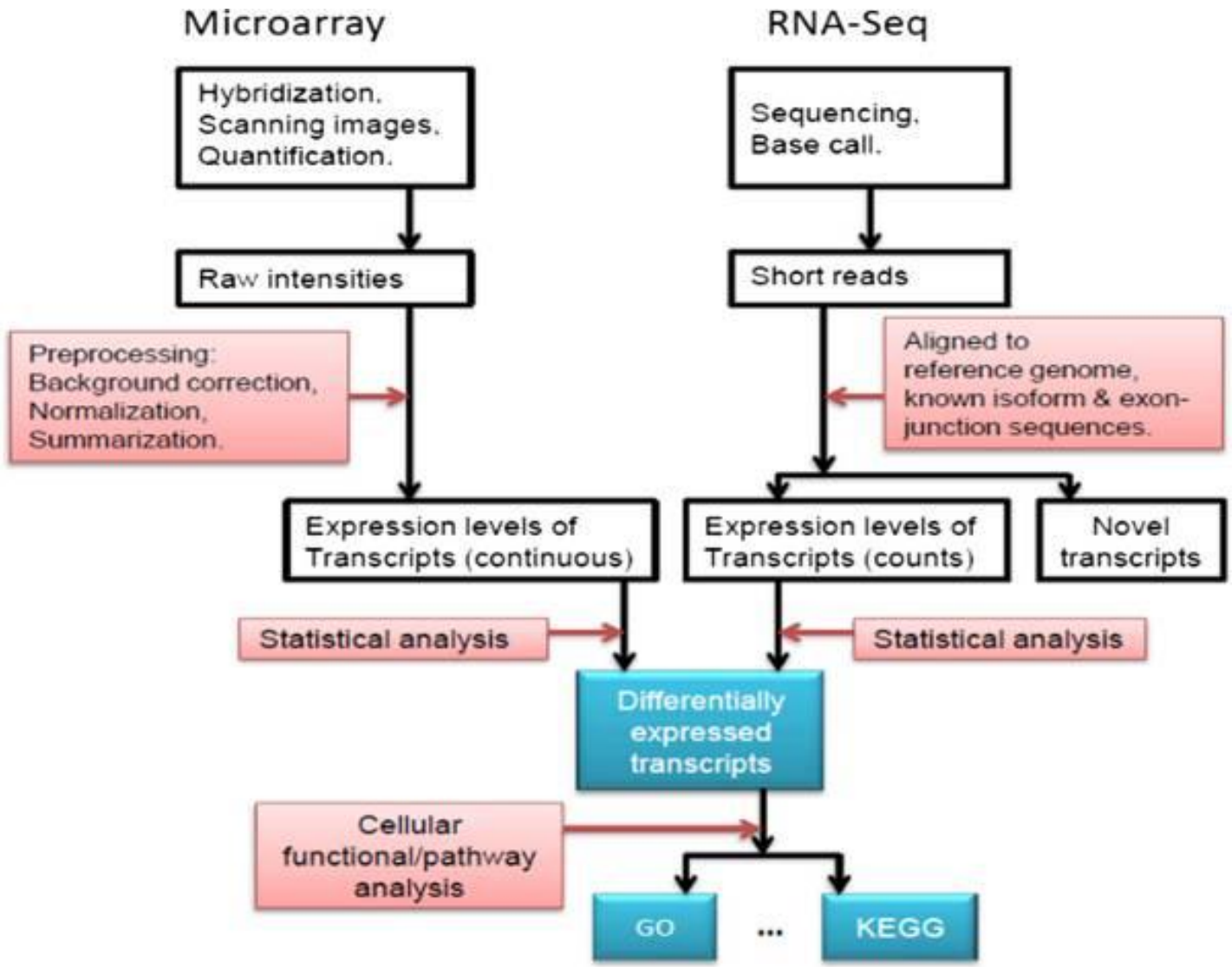
That’s what Pathway Analysis is about (more or less)

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 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 aetiology of a disease
  - Predict drug targets
  - Understand how to therapeutically intervene in disease processes
  - Conduct target literature searches
  - Integrate diverse biological information

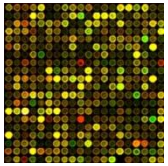
Managing Gene Lists

The life-cycle of an omics-based study



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

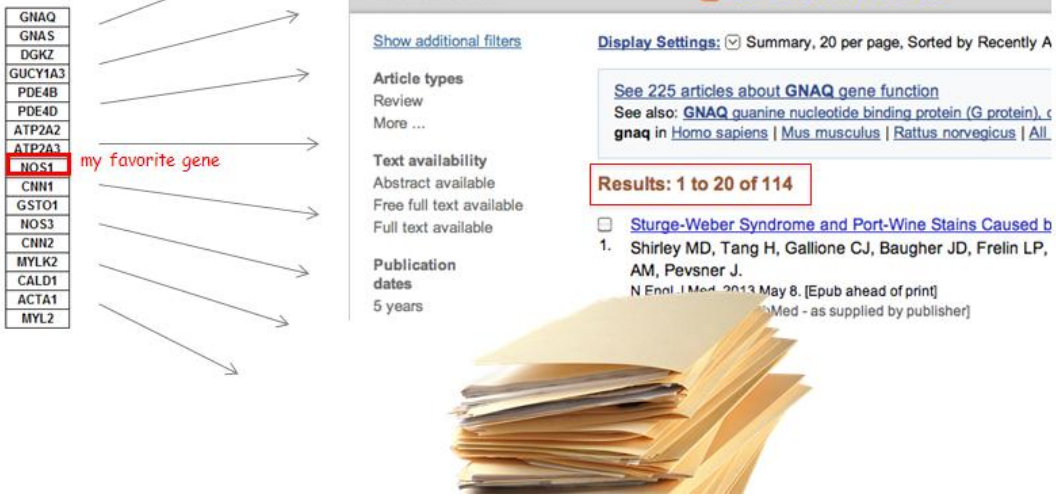
The (in)famous "*where to now?*" question



- 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

?



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***

Case study

- 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/Ejemplo_de_MDA_con_Bioconductor

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

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

Nature **426**, 100 (6 November 2003) | doi:10.1038/nature02141

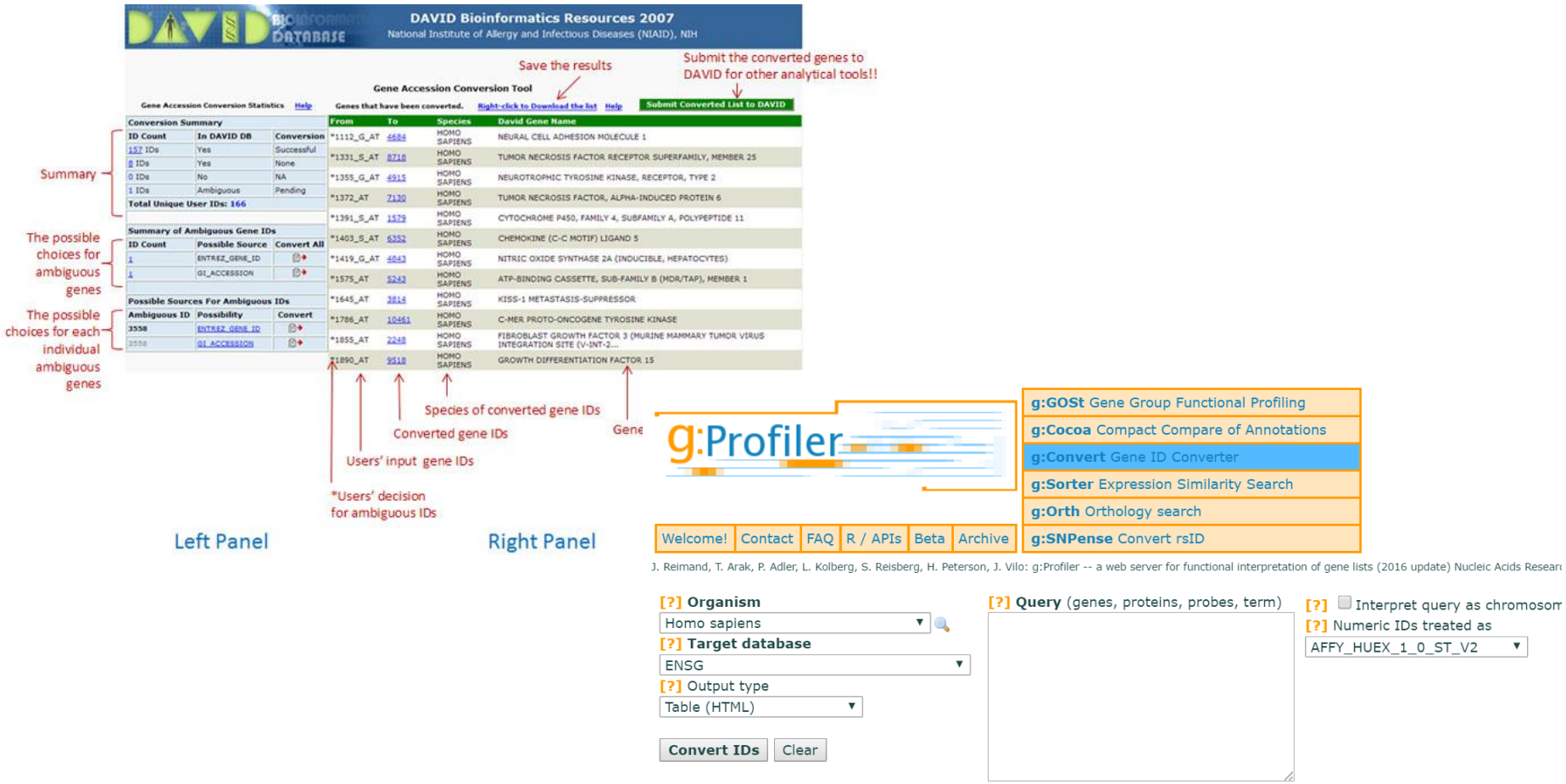
Retraction: Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells

Hiroaki Kawasaki & Kazunari Taira

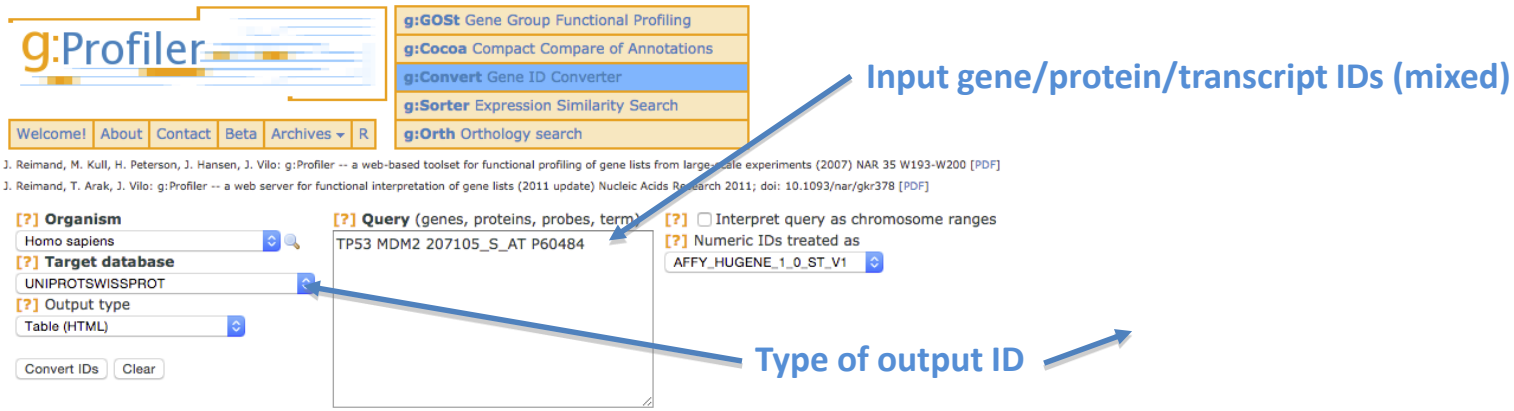
Nature **423**, 838–842 (2003).

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



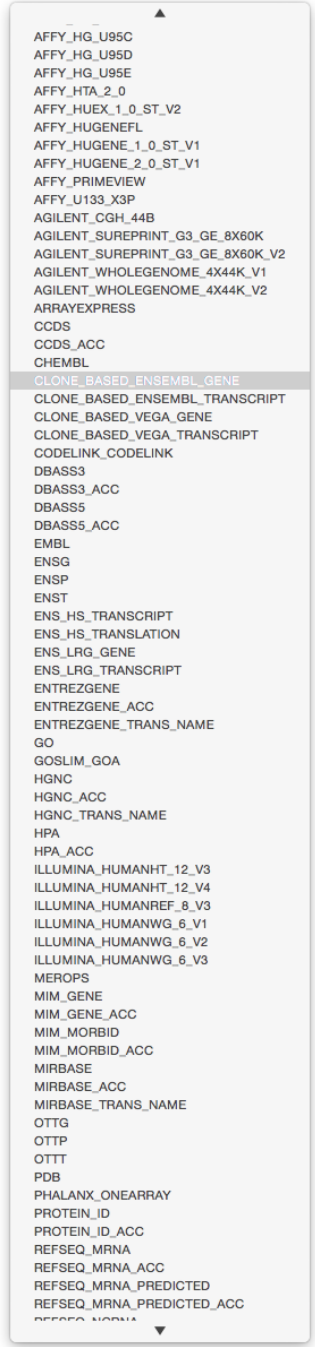
ID Mapping Services



>> Static URL

Come back later

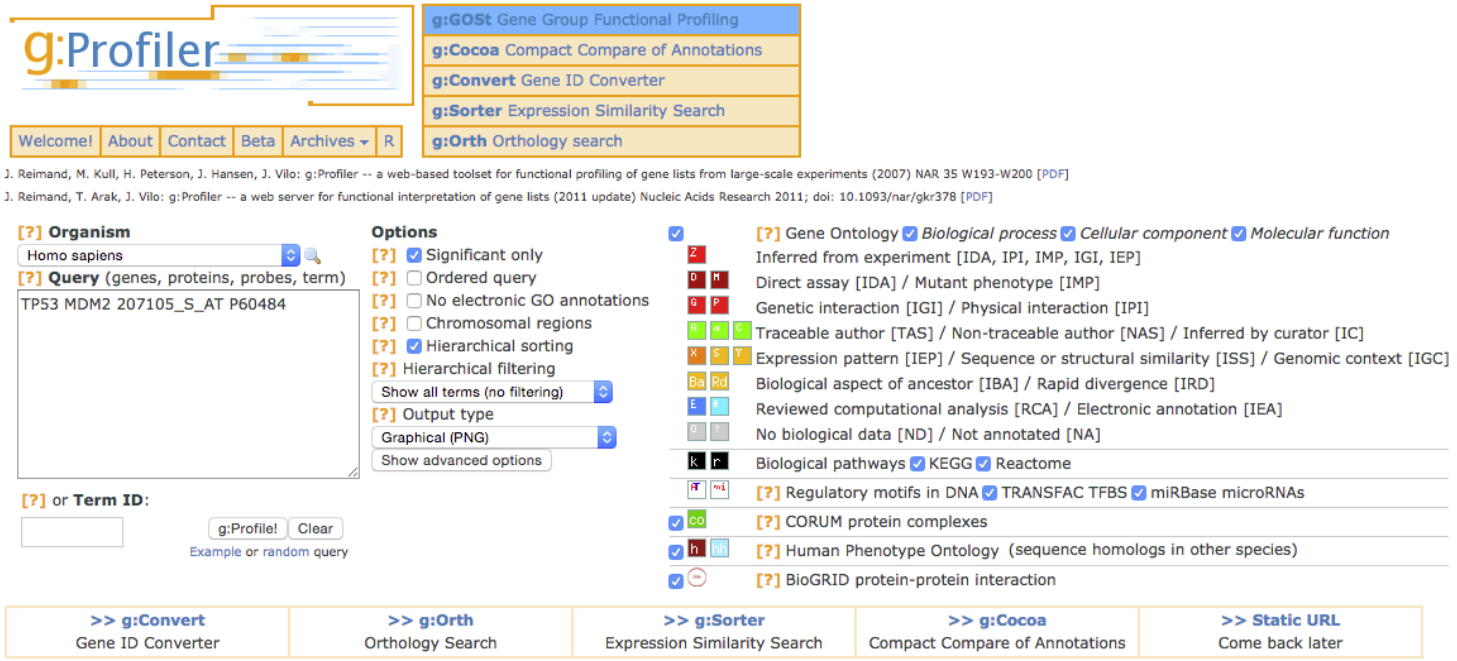
g#	initial alias >> g:GOSt >> g:Sorter >> g:Orth >> g:Cocoa	c#	converted alias >> g:GOSt >> g:Sorter >> g:Orth >> g:Cocoa >> Copy values	name >> g:GOSt >> g:Sorter >> g:Orth >> g:Cocoa >> Copy values	description	namespace
1	TP53	1.1	P04637	TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]	UNIPROT_GN, 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



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

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

Beware of ambiguous ID mappings



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? (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

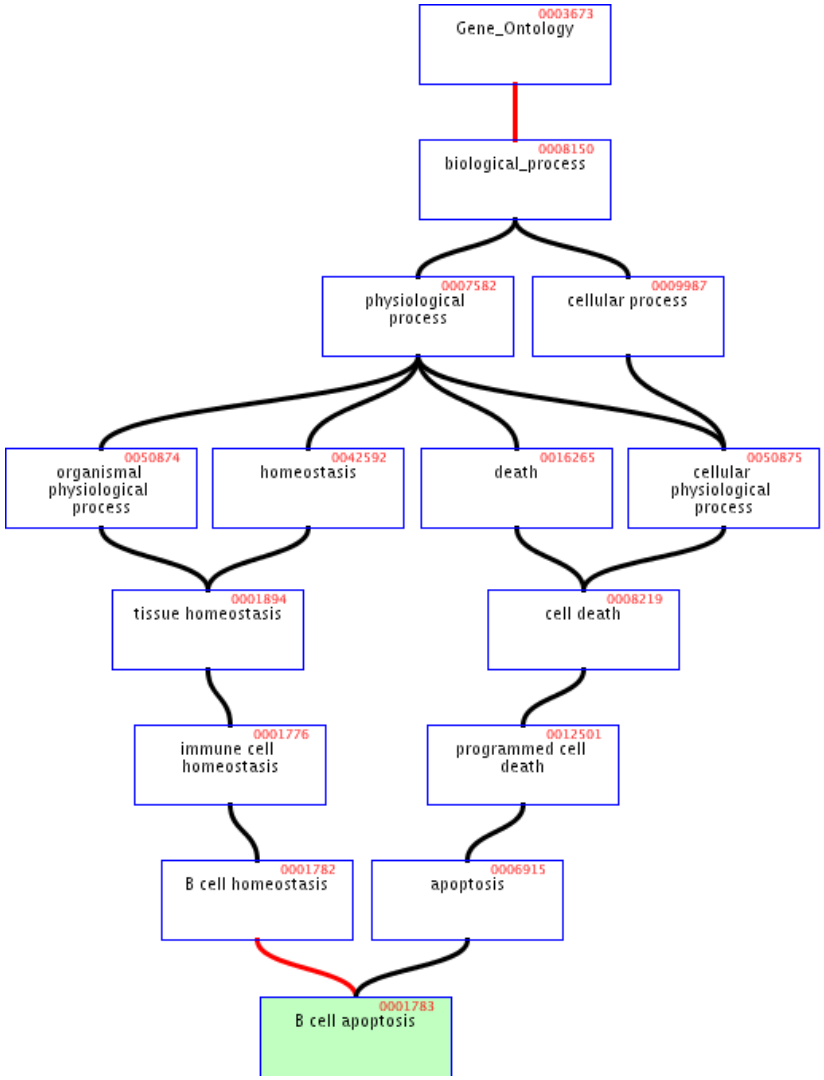
*Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges

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/

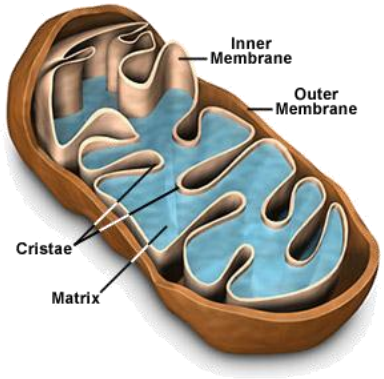
GO Structure

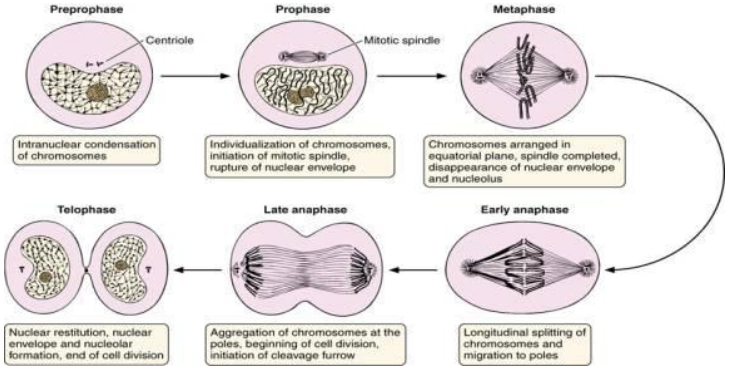
- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



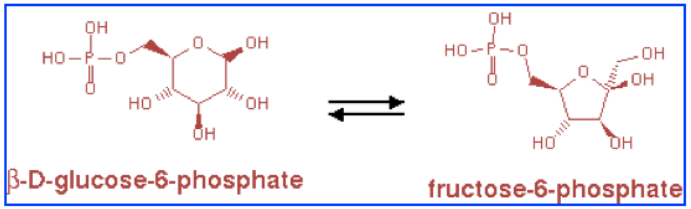
What is covered by the GO?

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





Cell division



glucose-6-phosphate isomerase activity

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

	Jun 2012	Apr 2015	increase
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

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)*

Pathway Analysis

Overrepresentation Analysis

Gene Set Enrichment Analysis

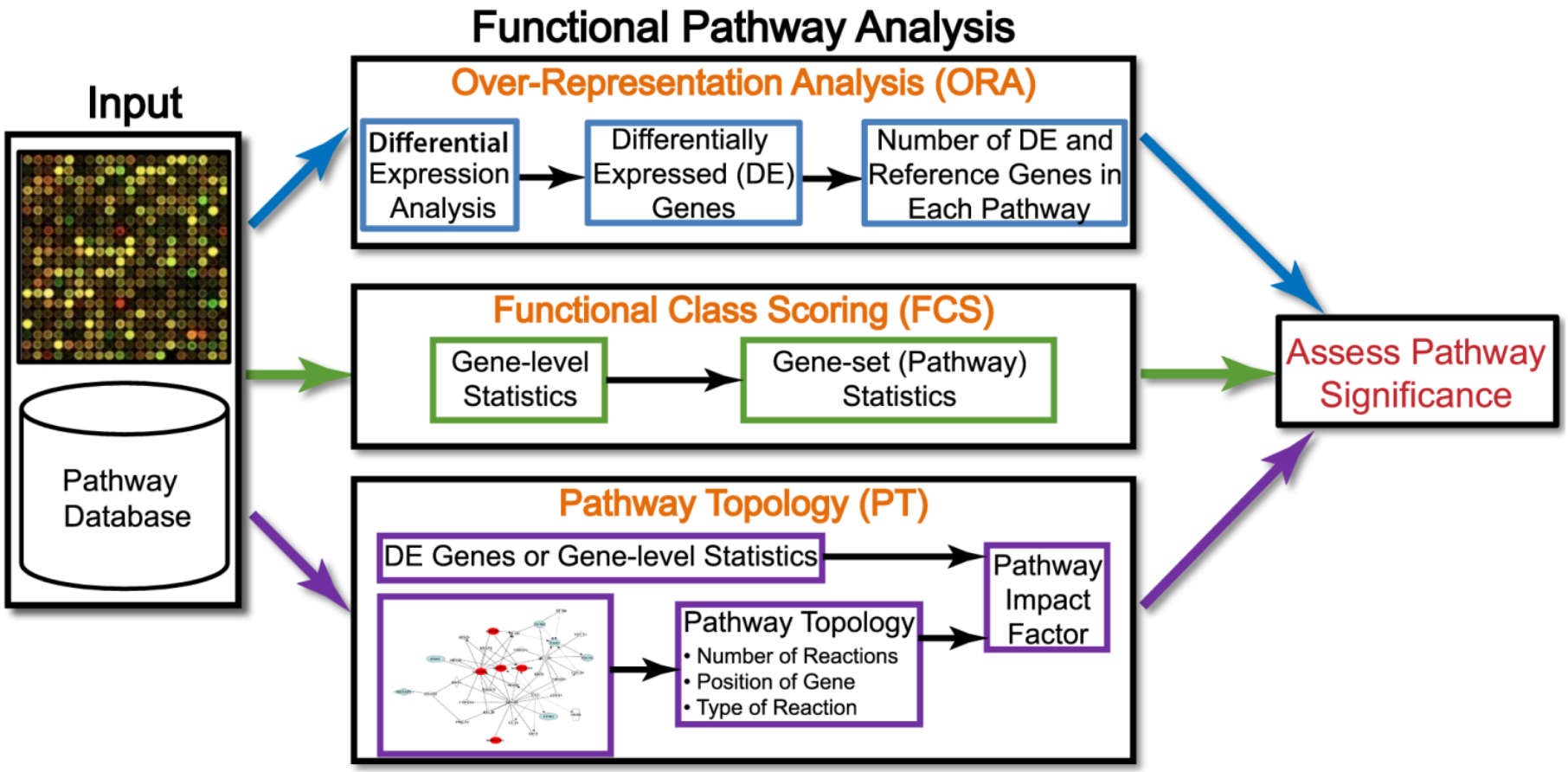
Pathway Analysis

- "*Any type of analysis that involves pathway or information*"
  - Most popular type is ***enrichment 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

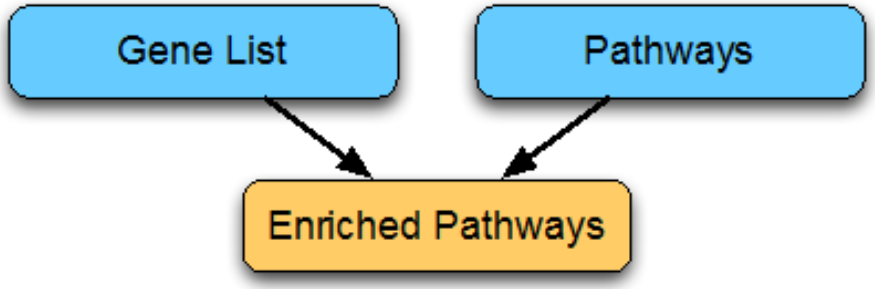


Khatri et alt. *10 years of Pathway Analysis*

Overrepresentation or *Enrichment* Analysis

Over-representation analysis

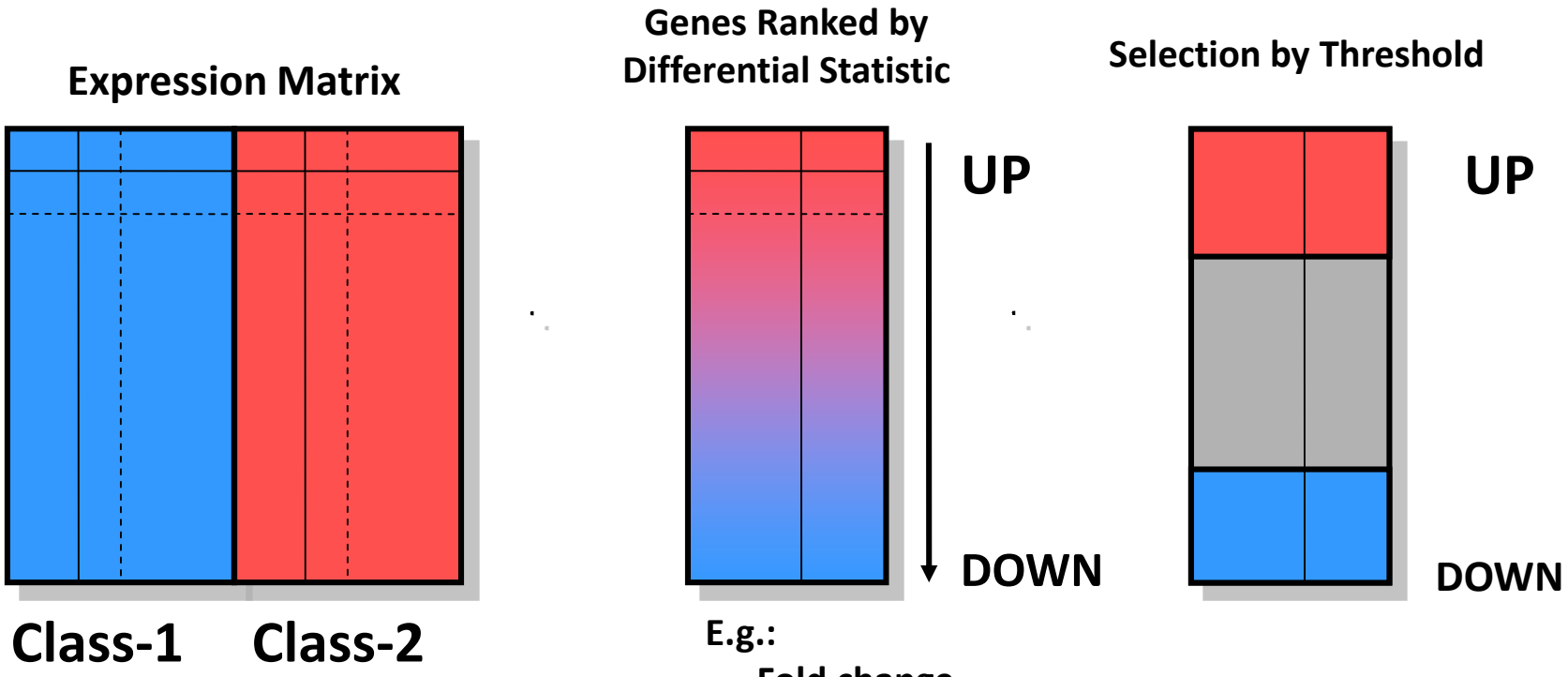
- Combines
  - Gene (feature) lists ← (Gen)omic experiment
  - Pathways and other gene annotations
    - Gene Ontology
    - Reactome
    - Pathway commons



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 correct for repeating the tests

1. Obtaining the gene lists



E.g.:

- Fold change
- Log (ratio)
- t-test
- Significance analysis of microarrays

2. 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 gl1 in GS1	**10/400=0.025**	**30/3600=0.00833**	

$0.025 >> 0.00833 \rightarrow$ "gl_1" is enriched in "GS_1"

	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**	

$0.025 \approx 0.02 \rightarrow$ Can't say that "gl_2" is enriched in "GS_2"

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)

```
> GOnnnnCounts<-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.")))
> GOnnnnCounts
                 Test
GeneSet           Diff.expressed Not diff.expr.
  In Gene Set                 10             30
  Not in Gene Set            390           1220
> fisher.test(GOnnnnCounts, alternative="greater")

        Fisher's Exact Test for Count Data

data:  GOnnnnCounts
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

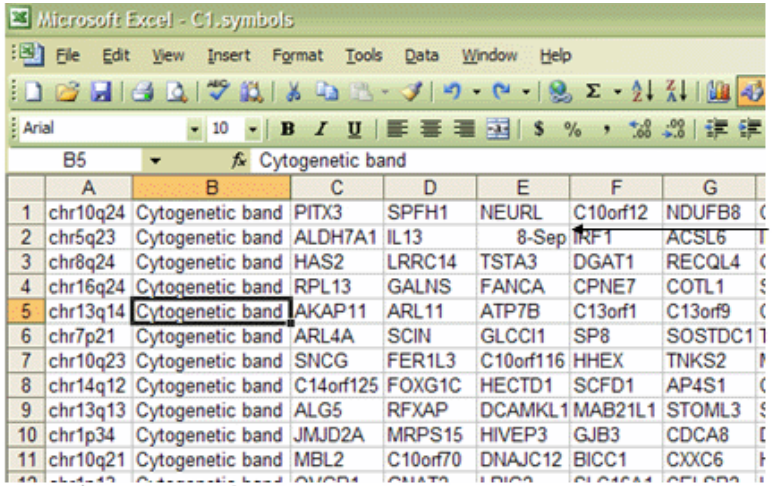
- No “natural” value for the threshold
- Possible loss of statistical power due to thresholding
  - No resolution between significant signals with different strengths
  - Weak signals neglected
- Different results at different threshold settings
- Based on the wrong assumption of independent gene (or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

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



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	3402
	CP: Canonical pathways KEGG/Biocarta/REACTOME	1320
C3: motif gene sets	MIR: microRNA targets	221
	TFT: transcription factor targets	615
C4: computational gene sets	CGN: cancer gene neighborhoods	427
	CM: cancer modules	431
C5: GO gene sets	BP: GO biological process	825
	CC: GO cellular component	233
	MF: GO molecular function	396
C6: oncogenic signatures		189
C7: immunologic signatures		1910
	Total	10295

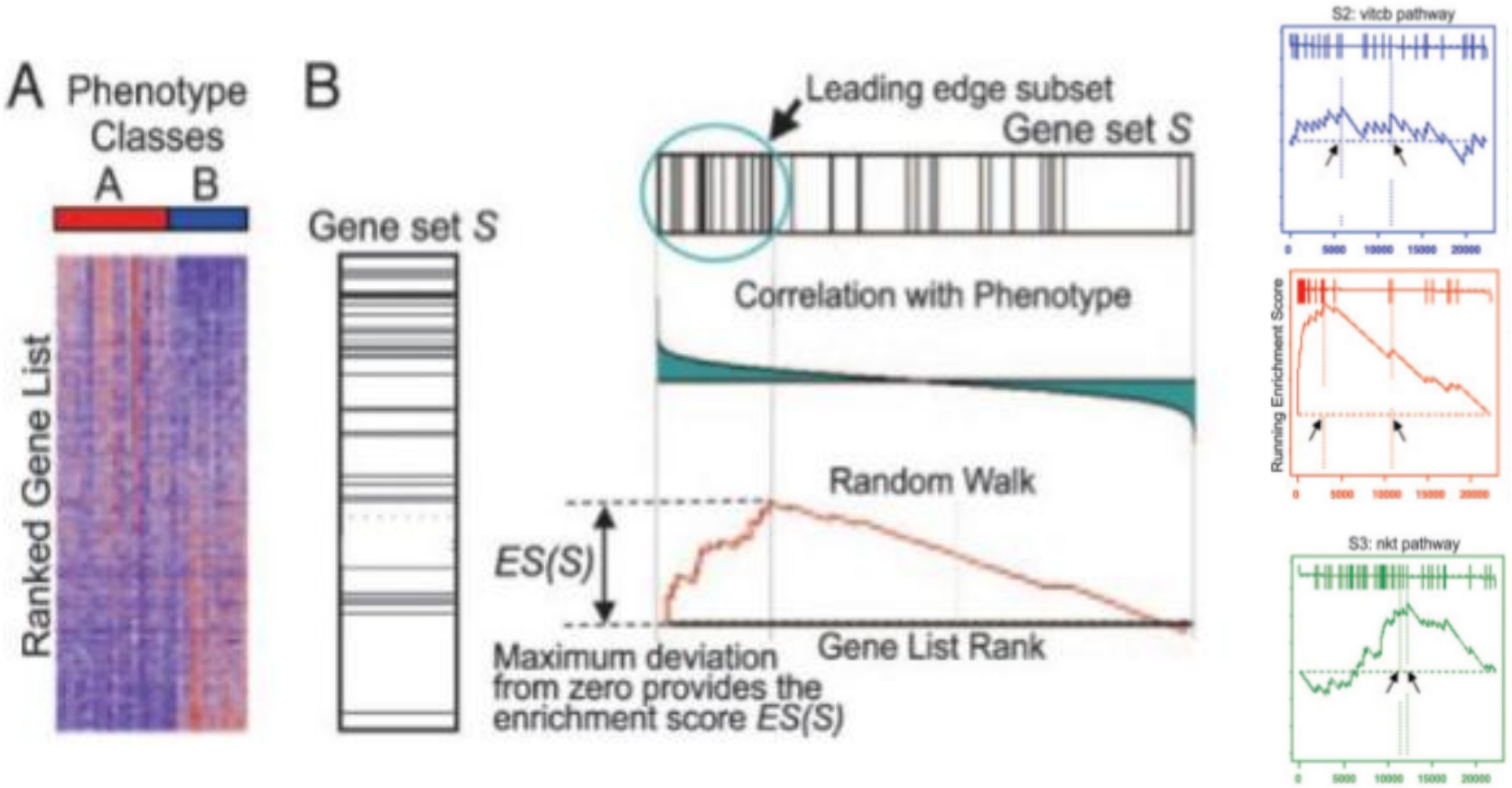
Gene Set (Enrichment) Analysis

- 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 affece 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 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 test (K-S test) statistic
- P-values are computed based on a randomization.

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

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

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

Testing repeatedly

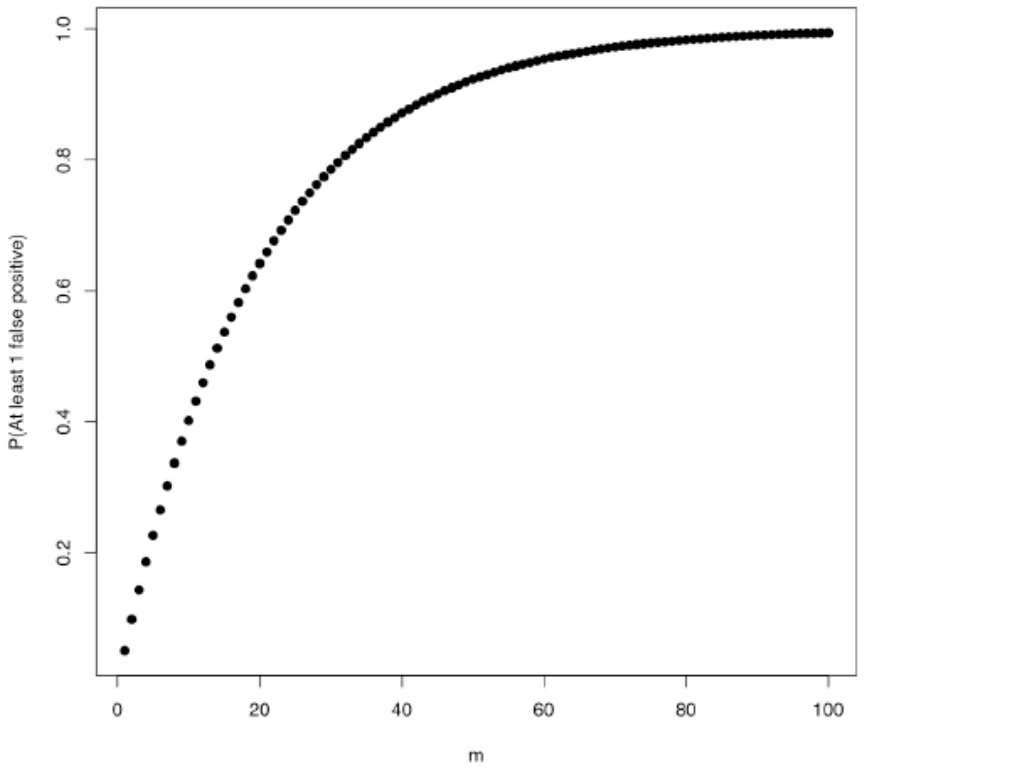
- Omics studies are “high throughput”
  - Selecting genes: One test per each gene
  - Finding enriched gene sets: One test per each gene set
- Doing many tests means facing repeatedly the probability of making one false positive.
  - As the number of tests increases →
  - The chance of observing at least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

Why multiple testing matters

- The probability of observing one false positive if testing once is:
  - P(Making a type I error) = α
  - P(not making a type I error) = 1- α
- Now imagine we perform *m* tests independently
  - P(not making a type I error in *m* tests) = $(1-\alpha)^m$
  - P(making at least a type I error in *m* 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

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 < α
  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 x 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	*Transcriptional*	0.001
2	*regulation*	0.002
3	*Transcription factor*	0.003
4	*Initiation of*	0.0031
5	*transcription*	0.005
…	*Nuclear localization*	…
	Chromatin modification	
52	…	0.97
53	*Cytoplasmic localization*	0.99
	Translation	

Sort P-values of all tests in decreasing order

Benjamini-Hochberg example II

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

Adjusted P-value = P-value X [# of tests] / Rank

Benjamini-Hochberg example III

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

P-value threshold for FDR < 0.05

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

Red: non-significant

Green: significant at FDR < 0.05

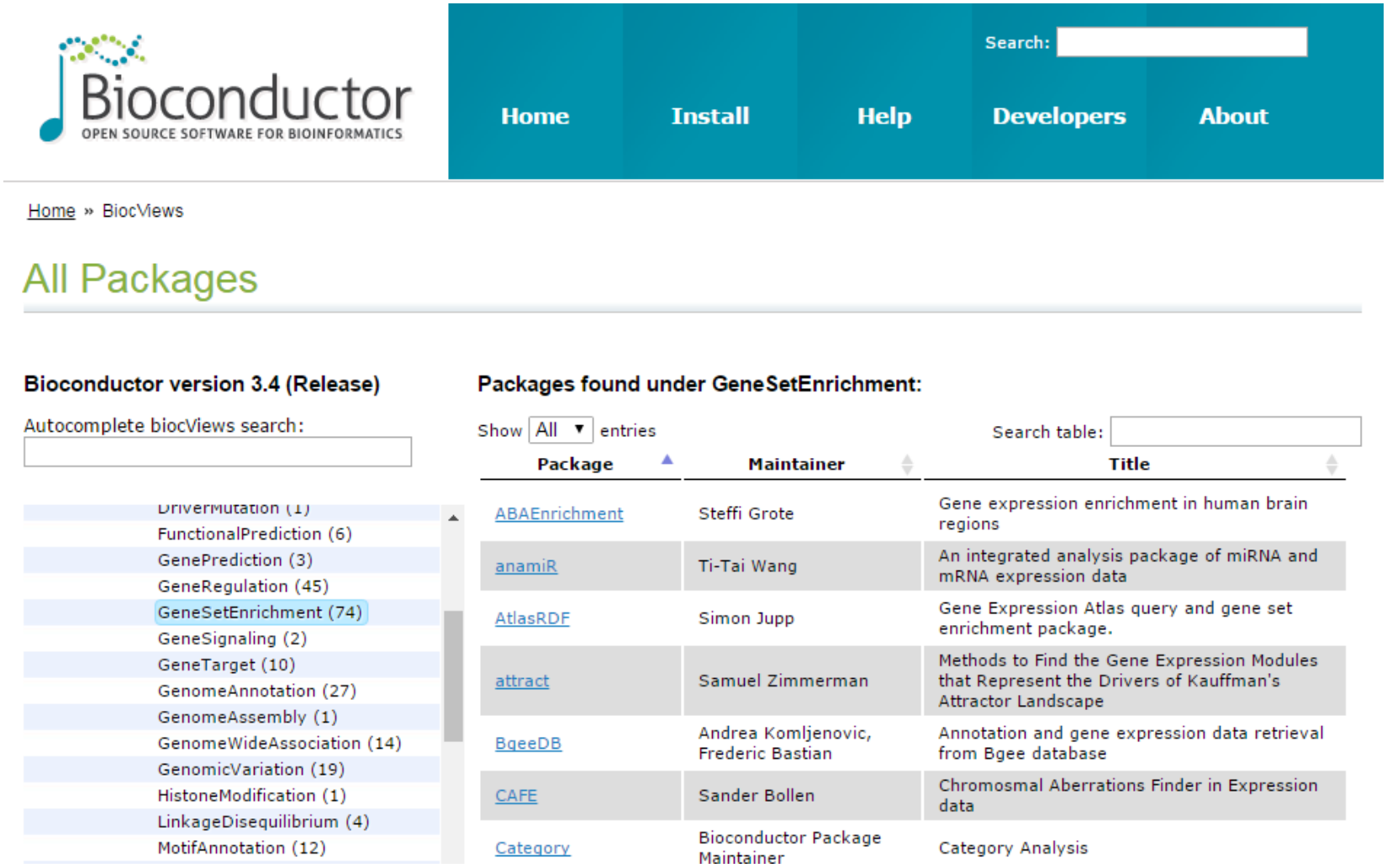
P-value threshold is highest ranking P-value for which corresponding Q-value is below desired significance threshold

Reducing 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



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/)
- GeneGo/MetaCore (www.genego.com)
- Ingenuity Pathway Analysis (www.ingenuity.com)
- Pathway Studio (www. ariadnegenomics.com)

BABELOMICS (FATIGO et alt.)


BIOINFORMATICS APPLICATIONS NOTE

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

Fátima Al-Shahrour, Ramón Díaz-Uriarte and Joaquín Dopazo*

ABSTRACT

Gene list

Genome

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

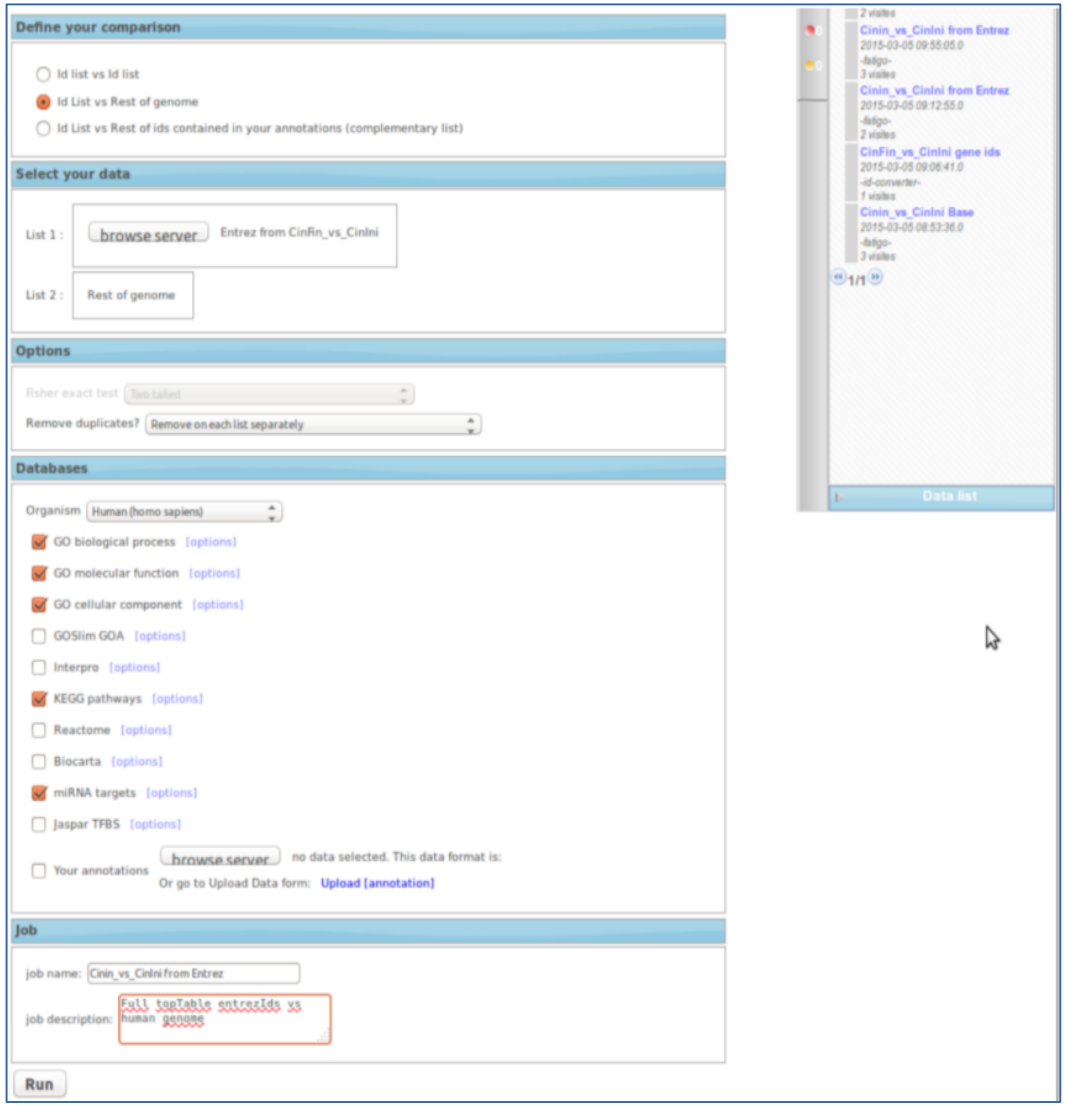
Gene list	Genome
GATA1 = 40%	GATA1 = 10%
SP1 = 20%	SP1 = 20%

The gene list is over-enriched for GATA1 targets.

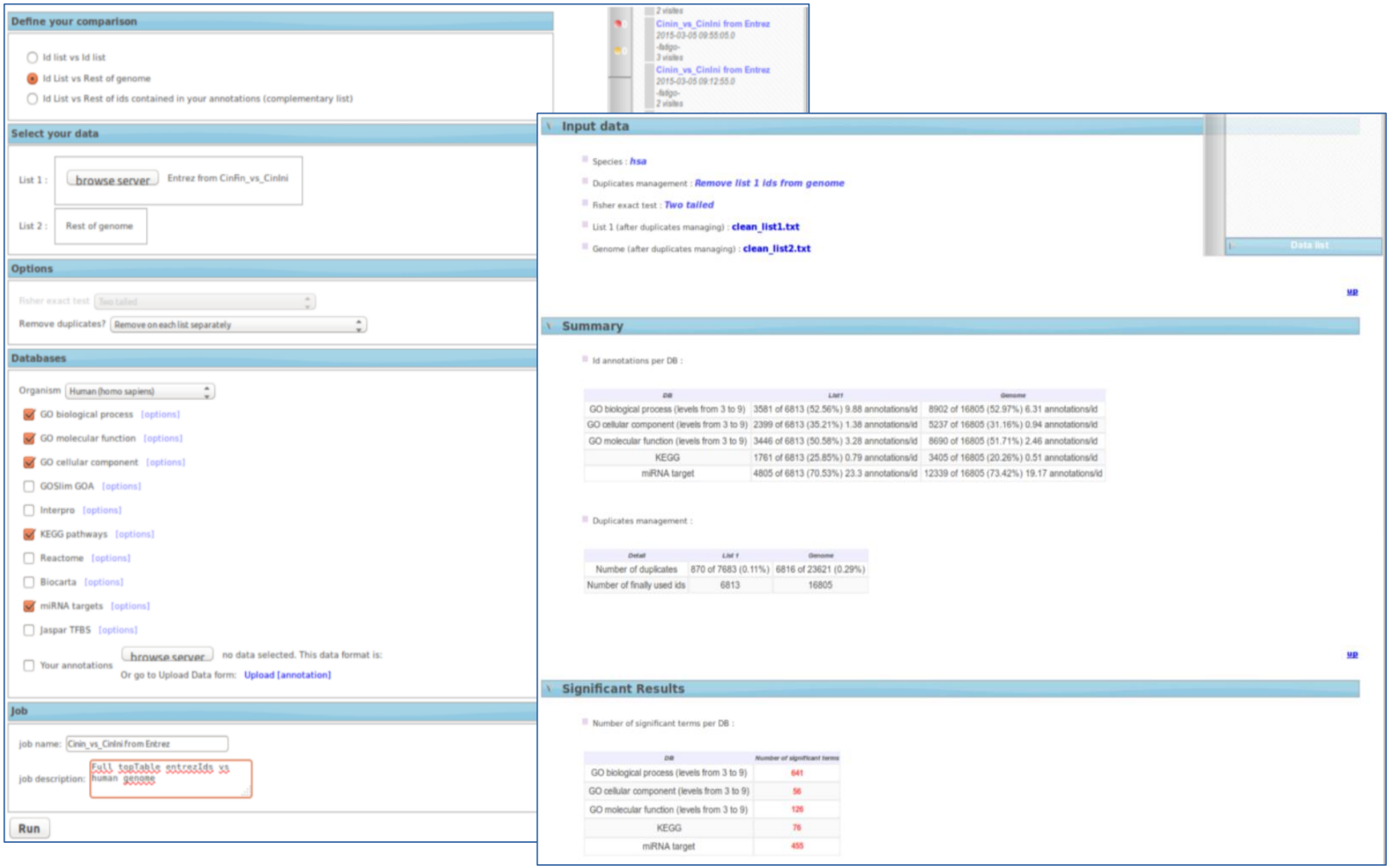
	Gene list	Genome
Regulated by GATA1	4	2
Not regulated by GATA1	6	18

- 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 2×2 contingency tables is used to check for significant over-representation 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.

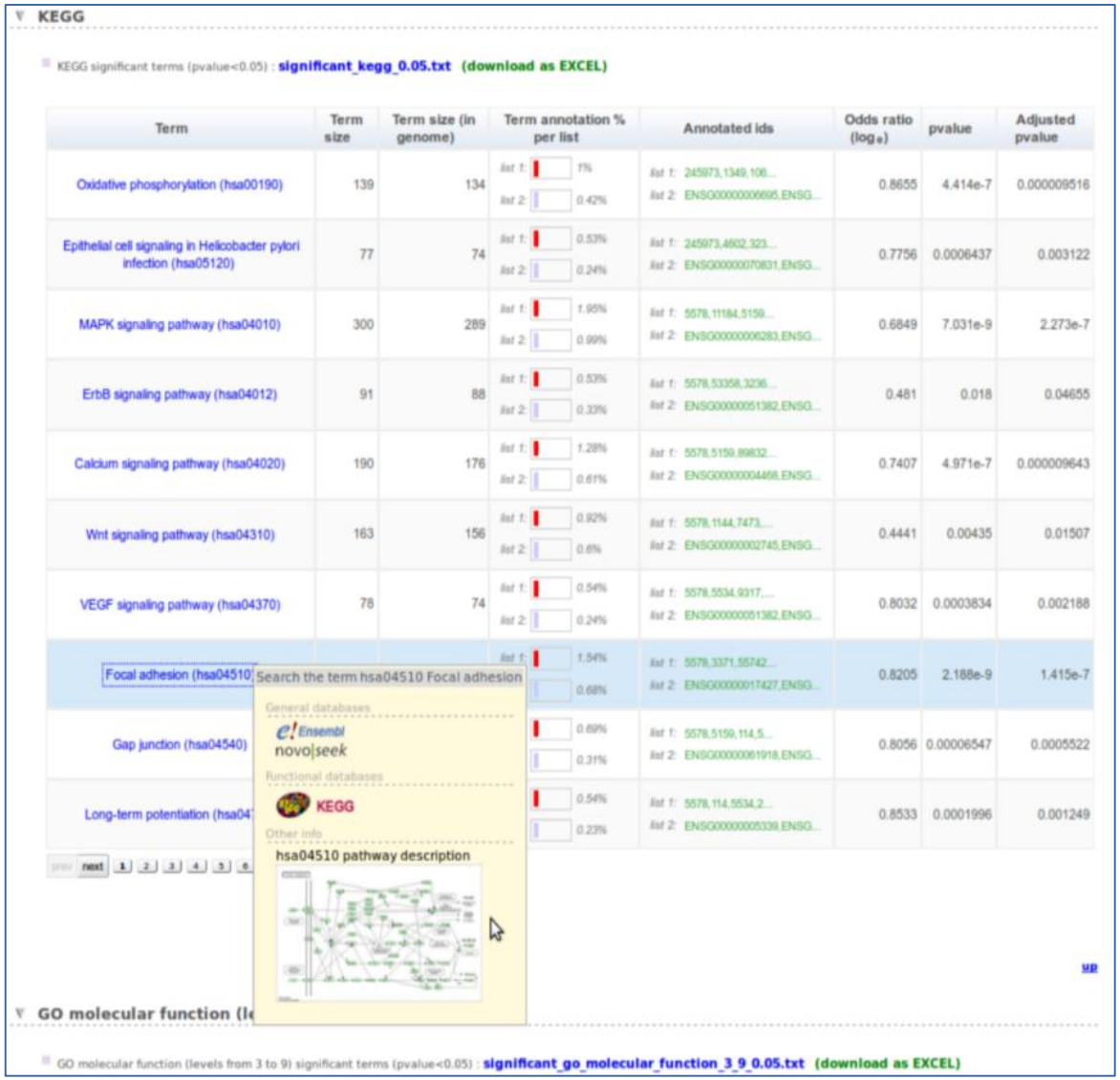
Upload gene list, set parameters



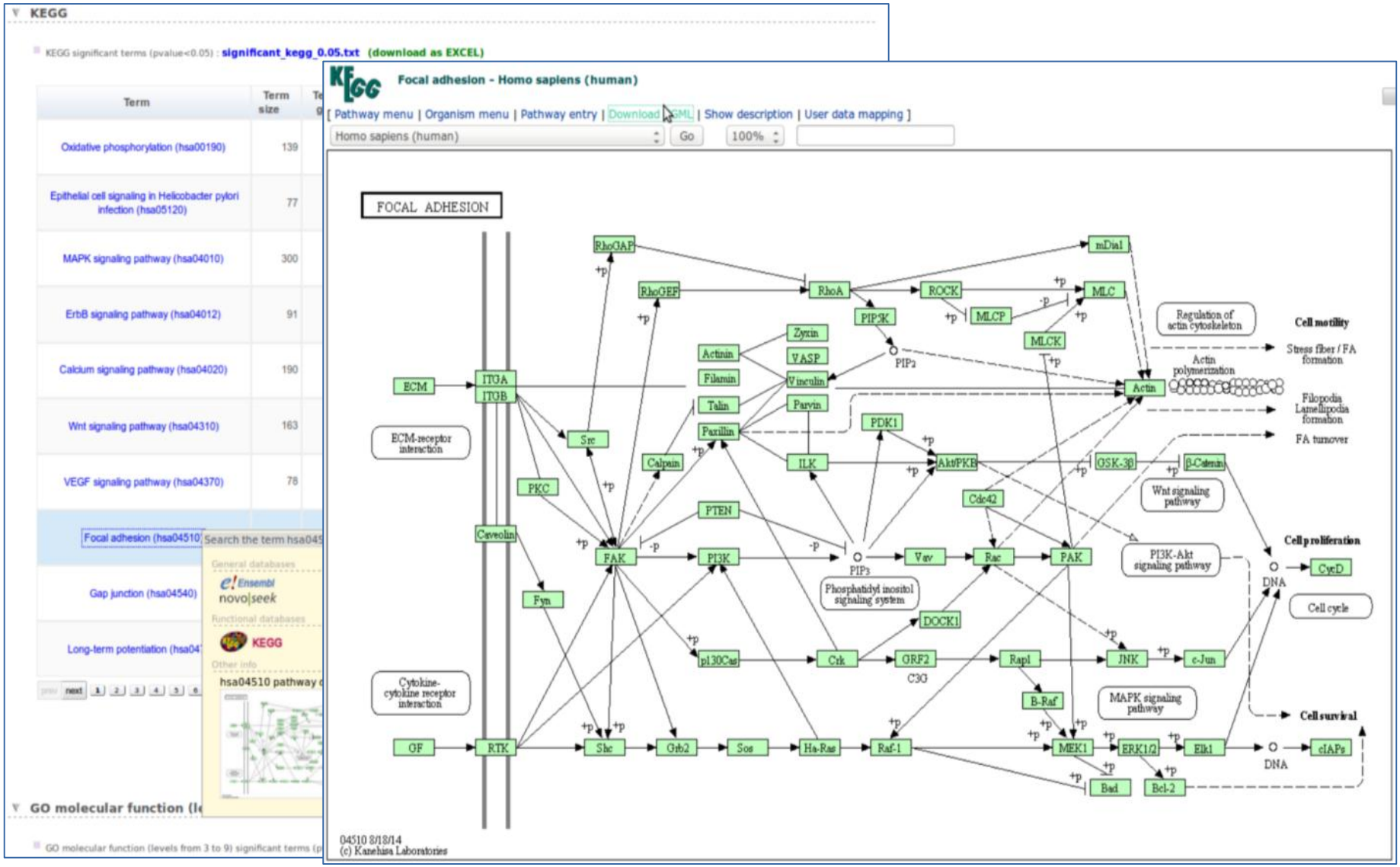
Obtain “significantly enriched” sets



Visualize results

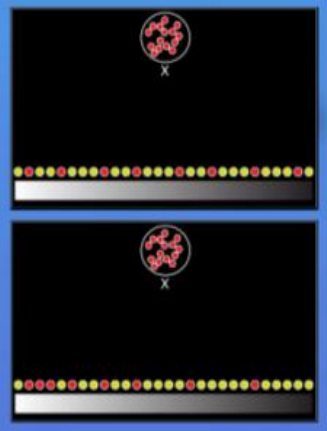


Visualize results



“Official” GSEA. BROAD Institute

- 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?



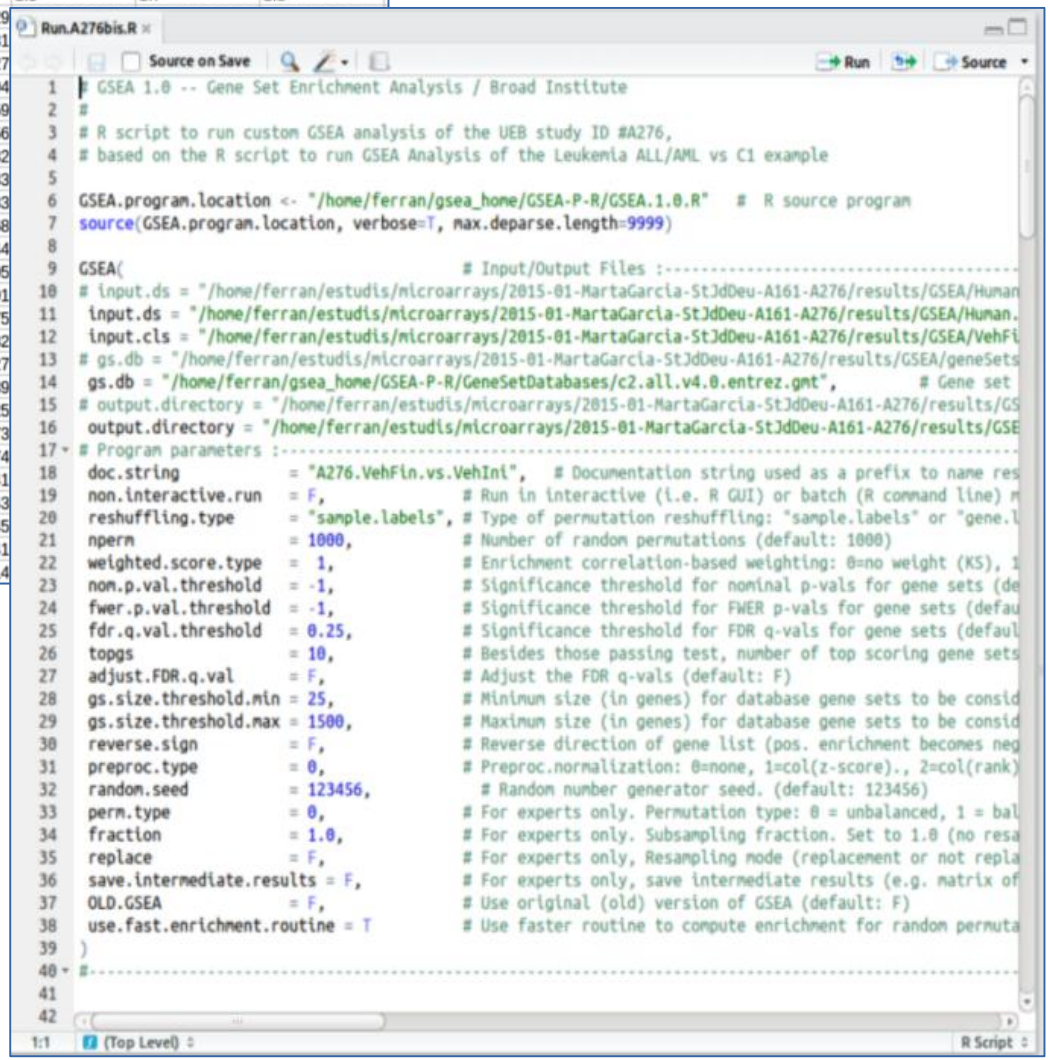
- Calculate enrichment score (ES)
- Calculate significance of ES
- Correct for multiple testing

Upload data matrix (not gene list!)

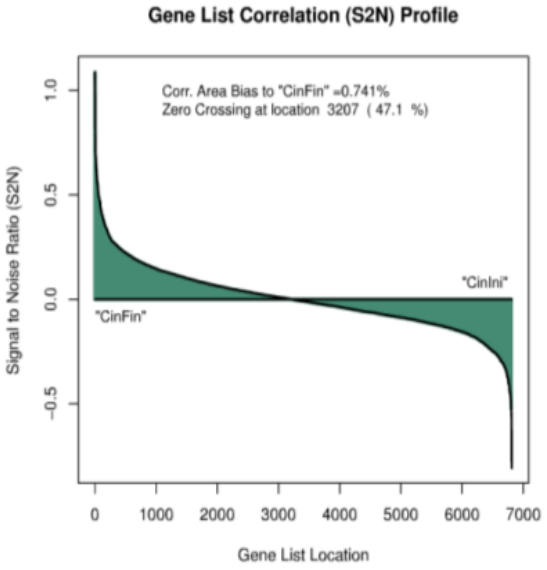
#1.2								
6811	7							
NAME	Description	Cf9	Cf10	Cf11	Ci4	Ci5	Ci7	Ci8
18	ABAT	6.7100903124	6.2010968359	6.597489765	6.5013934729	6.8451356349	7.5118595998	6.4110123122
19	ABCA1	5.4530779893	5.4355752984	5.4610085993	5.5476548331	5.4520121732	6.0489388288	5.0407680213
22	ABCB7	7.3694752885	6.8977680549	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//SOD	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	4.3941808636	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

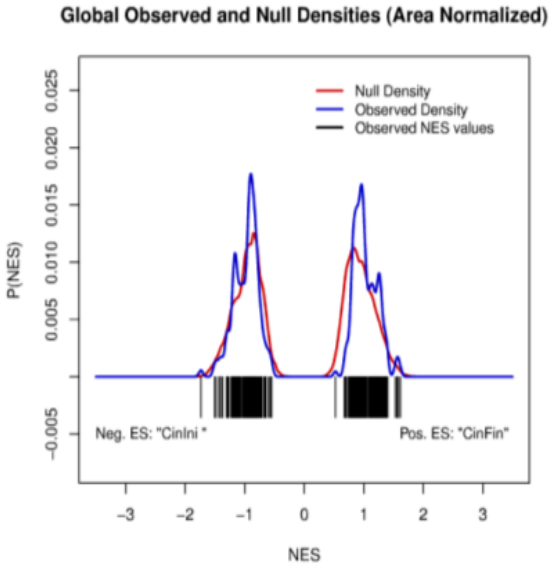
Set analysis parameters and gene sets

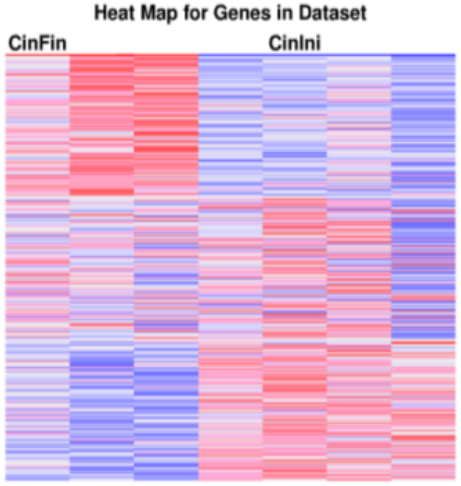
#1.2								
6811	7							
NAME	Description	Cf9	Cf10	Cf11	Ci4	Ci5	Ci7	Ci8
18	ABAT	6.7100903124	6.2010968359	6.597489765	6.5013934729			
19	ABCA1	5.4530779893	5.4355752984	5.4610085993	5.5476548331			
22	ABCB7	7.3694752885	6.8977680549	7.3073544278	7.3465162227			
30	ACAA1//OTTHUMG000001559	4.7248754828	5.0149574378	4.5595306969	4.706261494			
34	ACADM	8.9297666695	8.7971612382	8.5892949897	9.0426688459			
36	ACADSB	5.7822772441	5.2725699164	6.0307174265	5.8207351756			
39	ACAT2//LOC100129518//SOD	6.9231386649	6.5549055906	6.4733739317	7.0506799232			
41	ASIC1	4.7737457913	4.6841654133	5.1766921232	4.7347939833			
43	ACHE	4.1624435064	3.6882333994	3.9935732941	3.7993435183			
51	ACOX1	6.7826908205	7.0496507119	7.0241582248	6.4457072968			
52	ACP1	7.4211781076	7.097706037	6.9938476336	7.5431622044			
53	ACP2	5.059193215	4.8278872873	5.1630183499	4.7642311305			
54	ACP5	3.37202491	3.7503157578	3.7381985644	4.6264847891			
58	ACTA1	3.5406402898	3.9470695083	2.5025194518	2.7164043575			
60	ACTB	6.2064344458	5.6953236783	5.7415926085	6.8800198082			
71	ACTG1	10.9730301369	11.0722992954	10.7531966455	10.7612095627			
86	ACTL6A	7.306880239	7.3581695226	7.0414218102	7.3981514689			
88	ACTN2	4.8202948791	4.9049509308	4.7554192235	4.9223164025			
90	ACVR1	5.0631062492	5.2052068044	4.6386208381	4.7515376273			
94	ACVRL1	4.0012691816	3.5307672326	3.5967210171	4.4299320674			
97	ACYP1	5.7893977955	6.0223275581	6.3080750412	6.0767124941			
100	ADA	3.283936647	3.5490351114	3.5513605652	3.5467692743			
107	ADCY1	7.9872080552	7.5776132191	7.2672454373	7.3952269045			
111	ADCY5	5.1798093059	5.0993243605	5.6976263009	5.3114479241			
112	ADCY6//MIR4701	5.3929438612	5.4234156153	5.7267878509	5.8827262514			

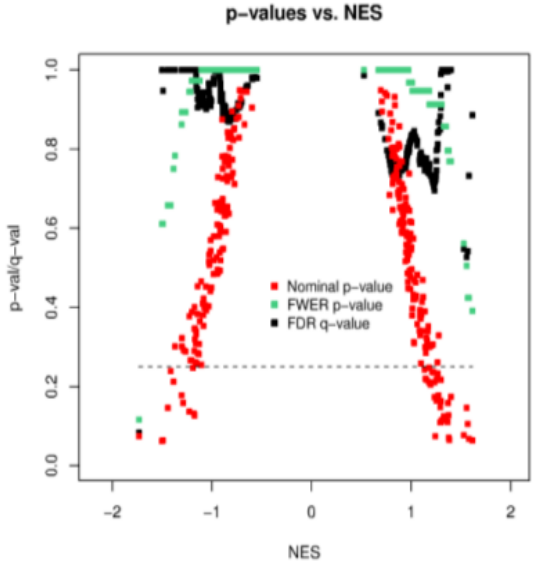


Get results. Interpret output









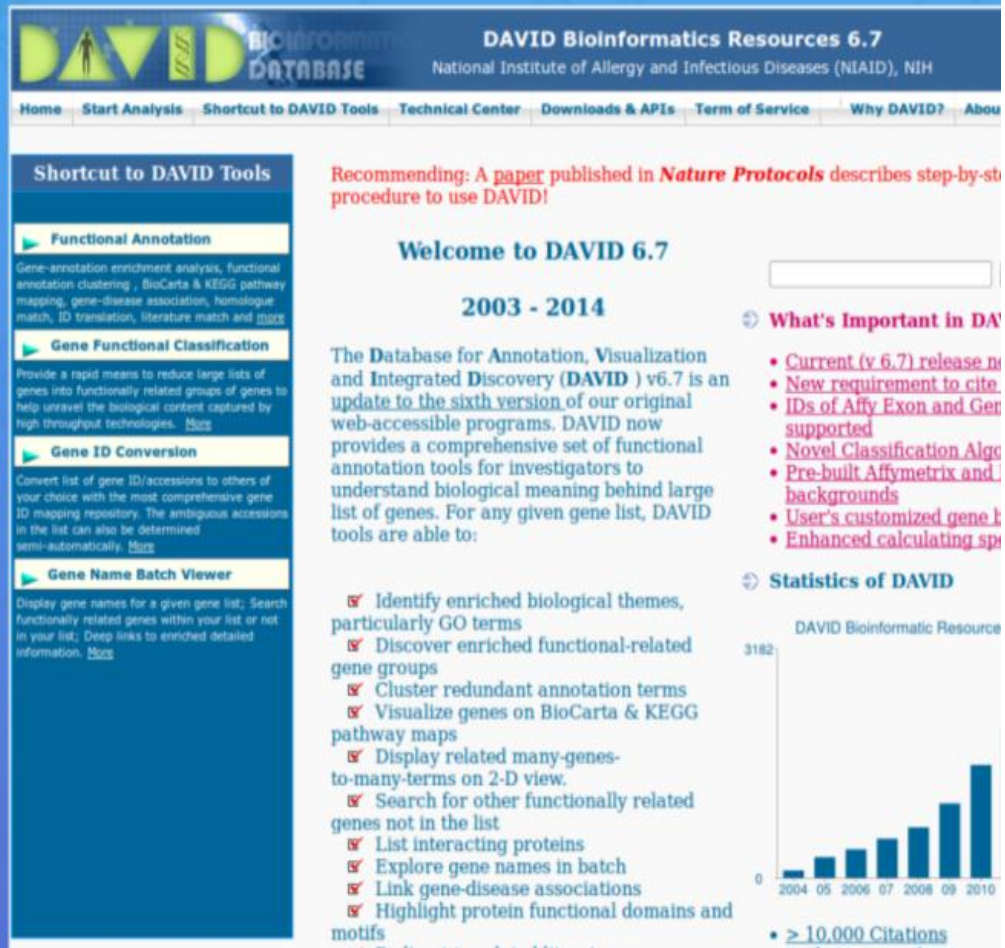
DAVID

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.

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.

- 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.



Upload gene lists. Define background

Results

Functional Annotation Clustering

Help and Manual

Current Gene List: demolist1
Current Background: Homo sapiens
155 DAVID IDs

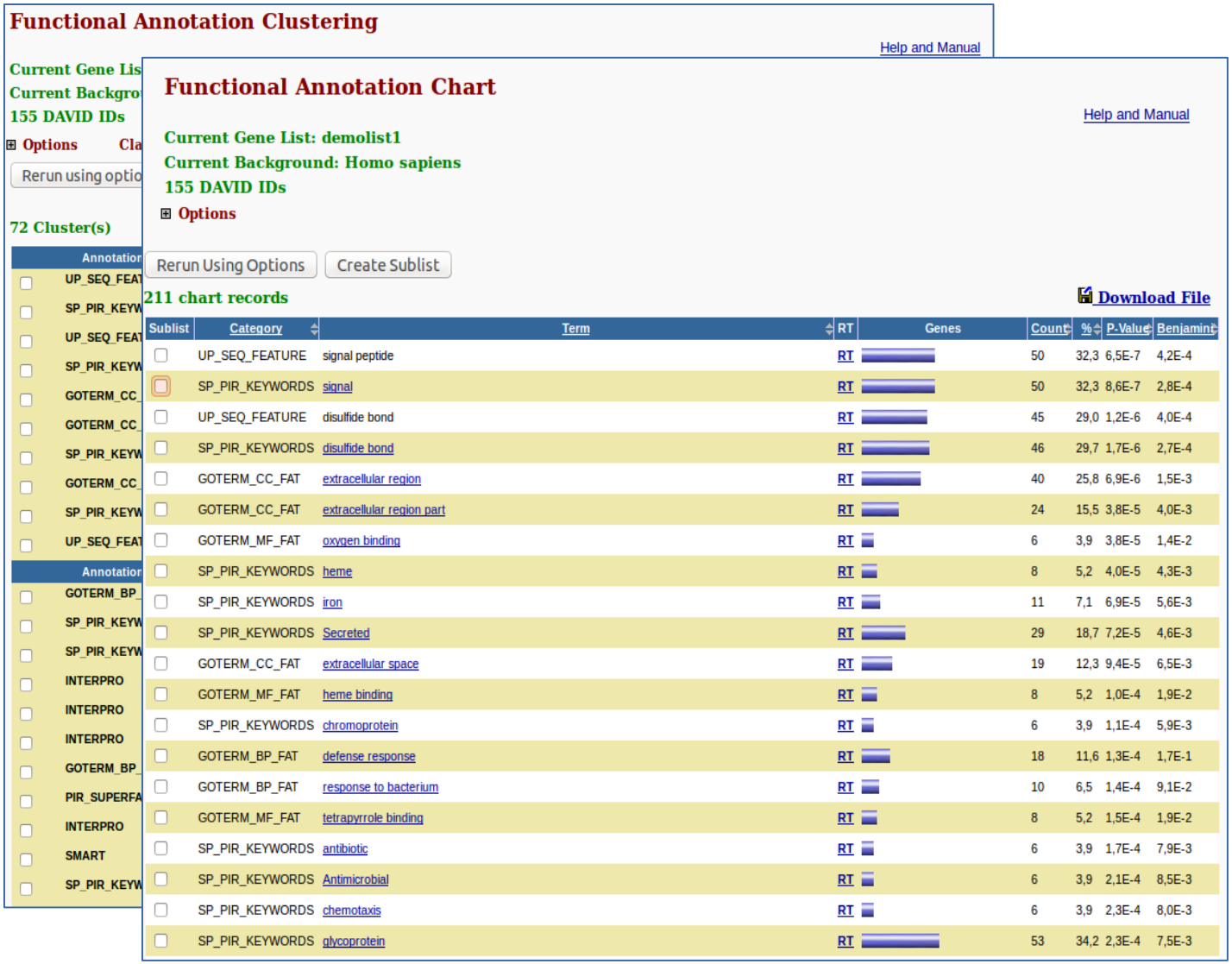
⊞ Options Classification Stringency: Medium

Rerun using options | Create Sublist

72 Cluster(s) Download File

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

Results



Functional Annotation Clustering

Help and Manual

Current Gene Lis
Current Backgro
155 DAVID IDs

Options Cla

Rerun using optio

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
Annotation
GOTERM_BP_
SP_PIR_KEYW
SP_PIR_KEYW
INTERPRO
INTERPRO
INTERPRO
GOTERM_BP_
PIR_SUPERFA
INTERPRO
SMART
SP_PIR_KEYW

Functional Annotation Chart

Help and Manual

Current Gene List: demolist1
Current Background: Homo sapiens
155 DAVID IDs

Options

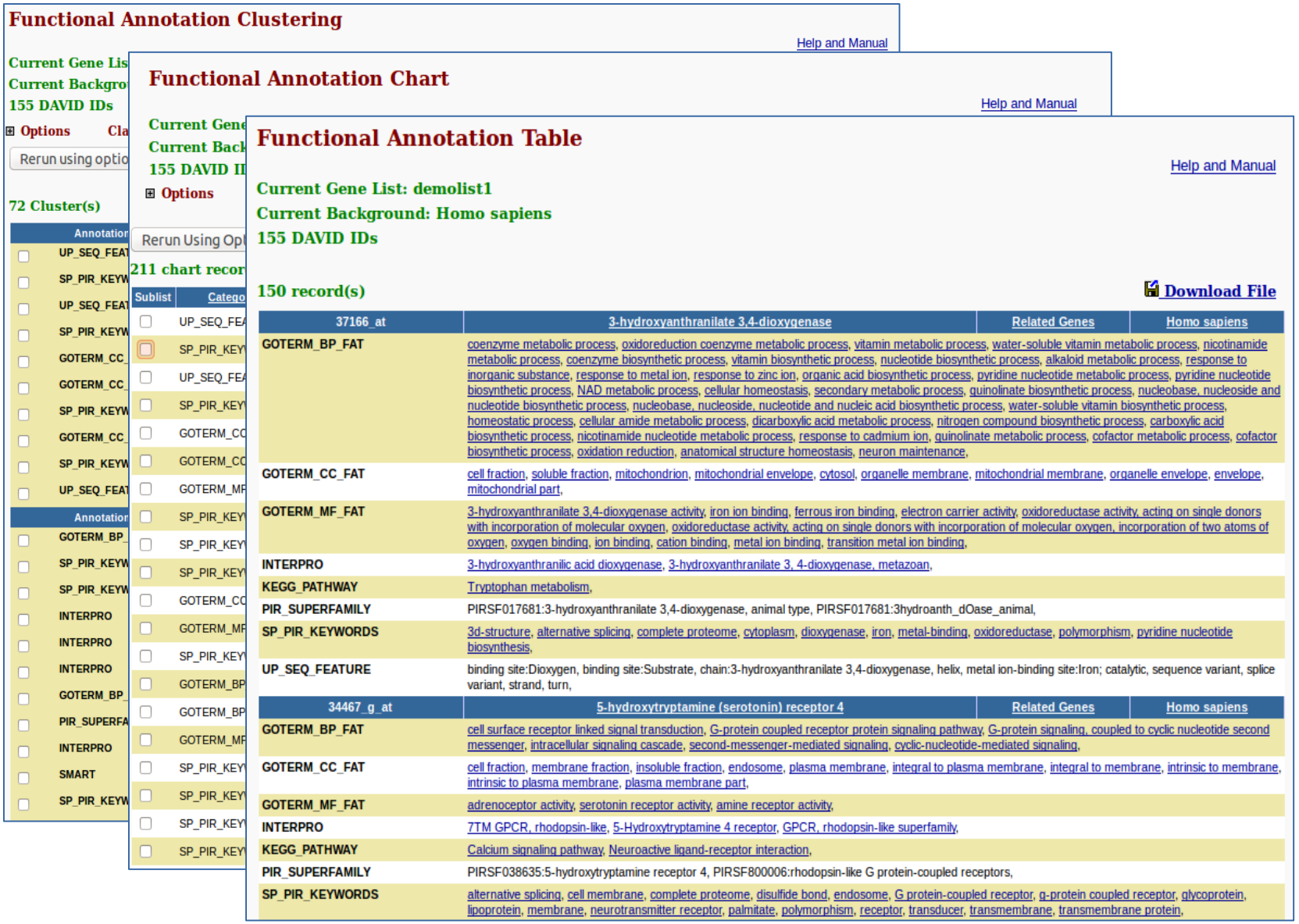
Rerun Using Options | Create Sublist

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
	SP_PIR_KEYWORDS	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
	GOTERM_CC_FAT	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

Help and Manual

Current Gene Lis
Current Backgro
155 DAVID IDs

Options Cla

Rerun using optio

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
Annotation
GOTERM_BP_
SP_PIR_KEYW
SP_PIR_KEYW
INTERPRO
INTERPRO
INTERPRO
GOTERM_BP_
PIR_SUPERFA
INTERPRO
SMART
SP_PIR_KEYW

Functional Annotation Chart

Help and Manual

Current Gene
Current Back
155 DAVID II

Options

Rerun Using Opt

211 chart recor

Sublist	Catego
	UP_SEQ_FEA
	SP_PIR_KEY
	UP_SEQ_FEA
	SP_PIR_KEY
	GOTERM_CC
	GOTERM_CC
	GOTERM_MF
	SP_PIR_KEY
	SP_PIR_KEY
	SP_PIR_KEY
	GOTERM_CC
	GOTERM_MF
	SP_PIR_KEY
	GOTERM_BP
	GOTERM_BP
	GOTERM_MF
	SP_PIR_KEY
	SP_PIR_KEY
	SP_PIR_KEY
	SP_PIR_KEY

Functional Annotation Table

Help and Manual

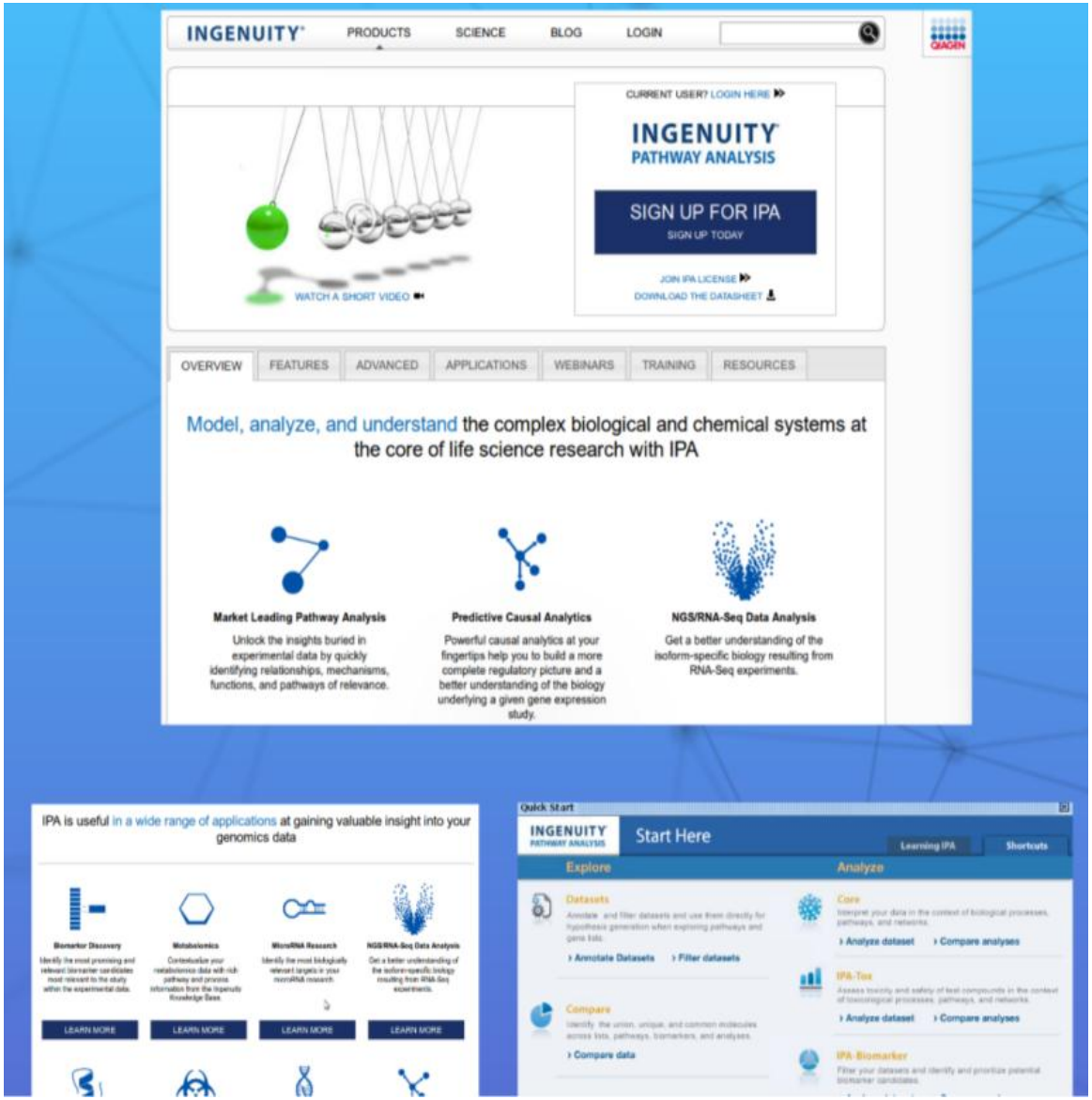
Current Gene List: demolist1
Current Background: Homo sapiens
155 DAVID IDs

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 metabolic 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



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/alexsanchezpla/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
GOparams = 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(GOparams)
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: