An Introduction to Pathway Analysis

Alex Sánchez

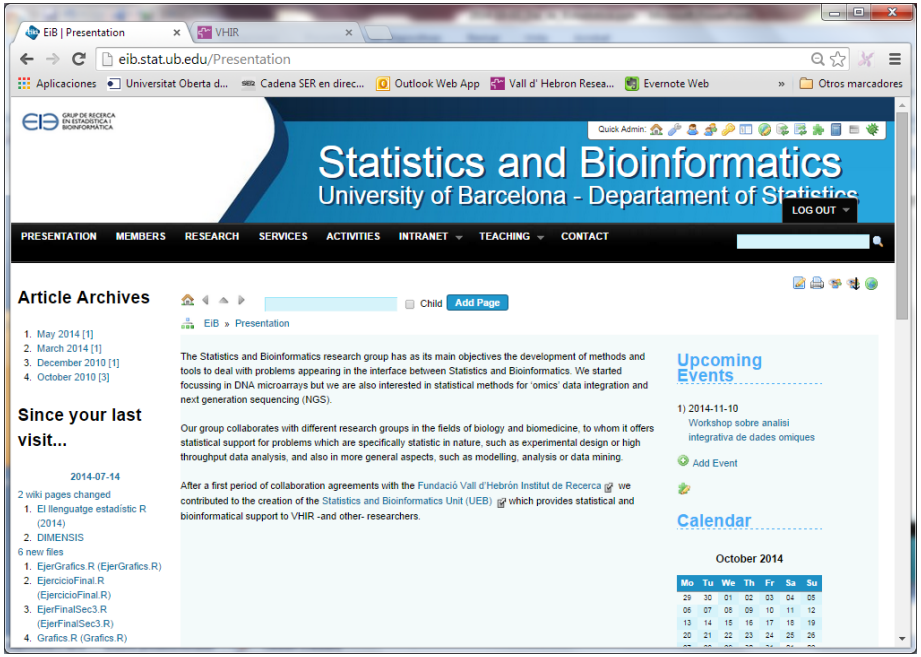
Statistics and Bioinformatics Research Group
Statistics department, Universitat de Barelona

Statistics and Bioinformatics Unit
Vall d'Hebron Institut de Recerca

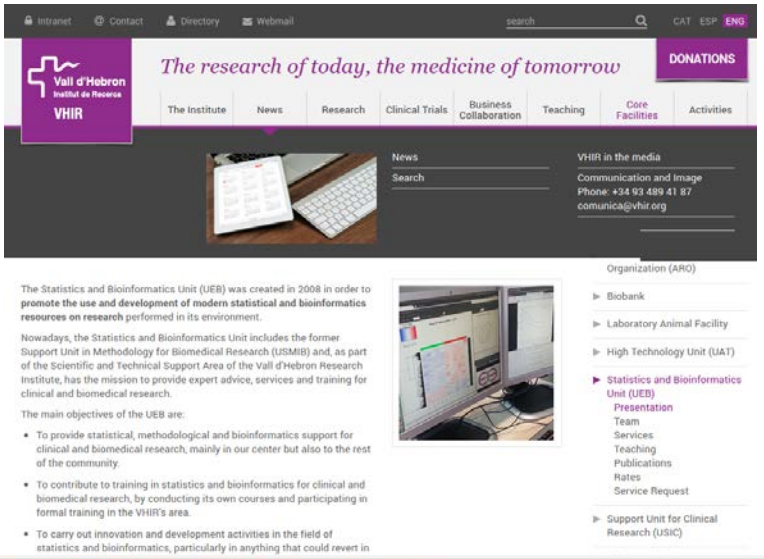
Outline

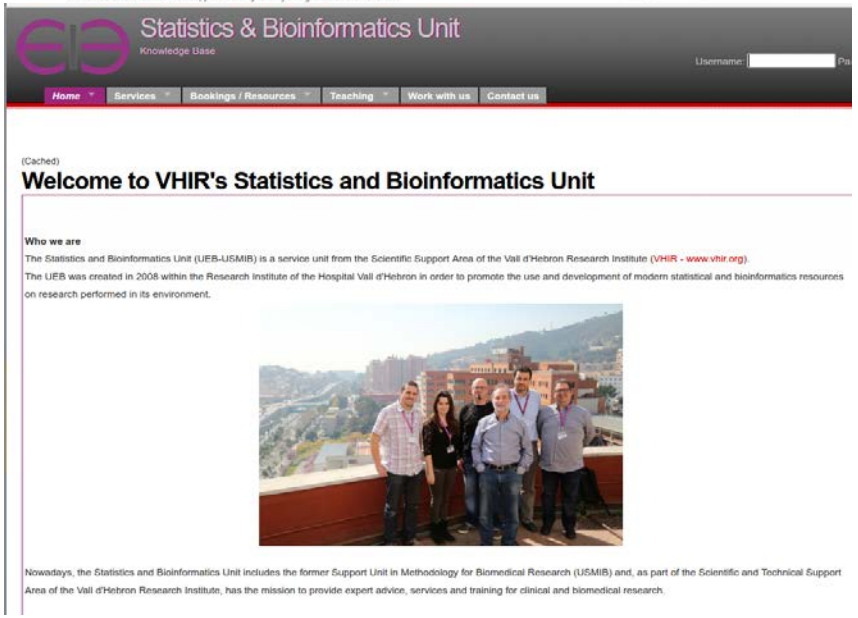
- Presentation
- Introduction and Background
- The problem: Interpreting gene lists
- Annotations and annotation databases
- The Gene Ontology Resource
- Gene list analysis using the GO and relatives
- Existing tools for pathway analysis

(Bio)Statistics and Bioinformatics Research groups



The Statistics and Bioinformatics Unit





Introduction & Background

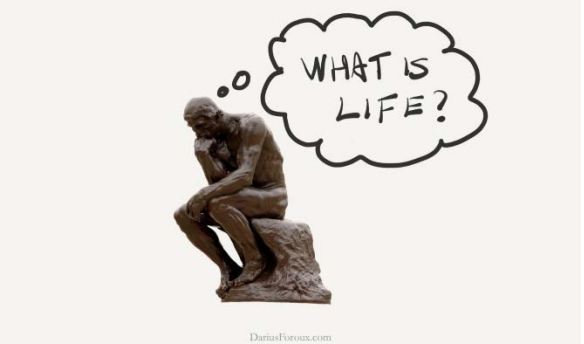
The main question

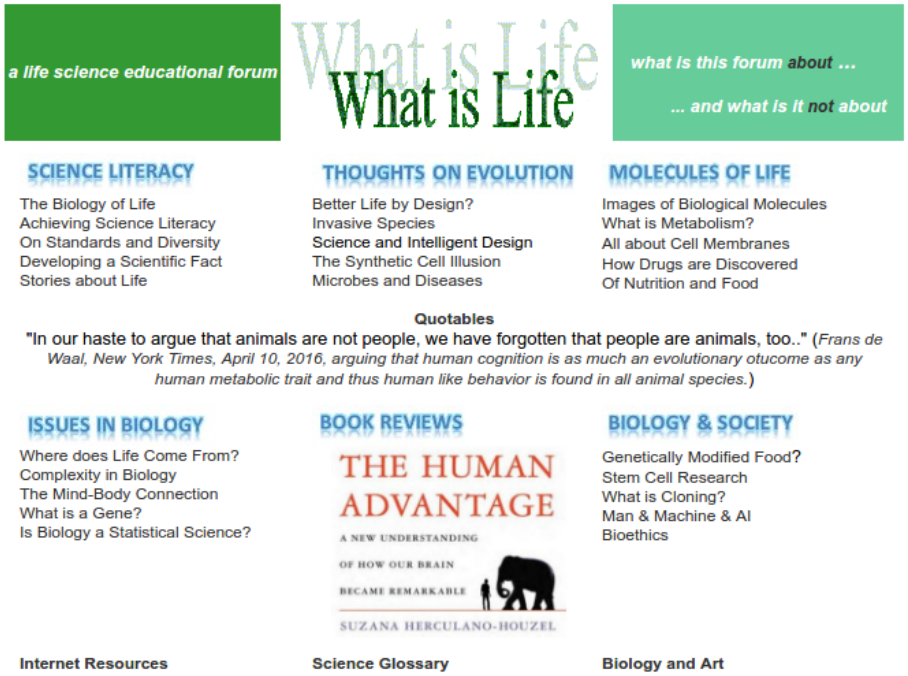
- We might sart by asking

  What is *LIFE*?

- A hard question to deal with
- What if we ask:

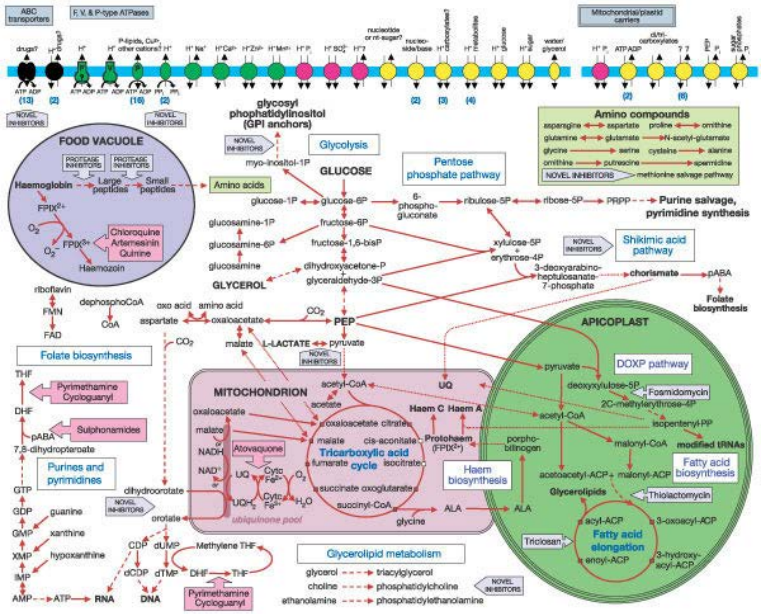
  What is life?



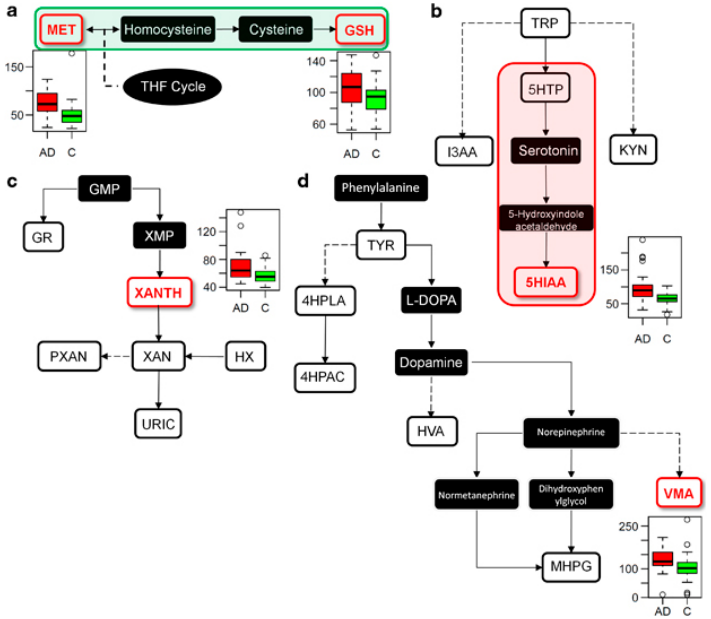


Life, disease and pathways

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



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



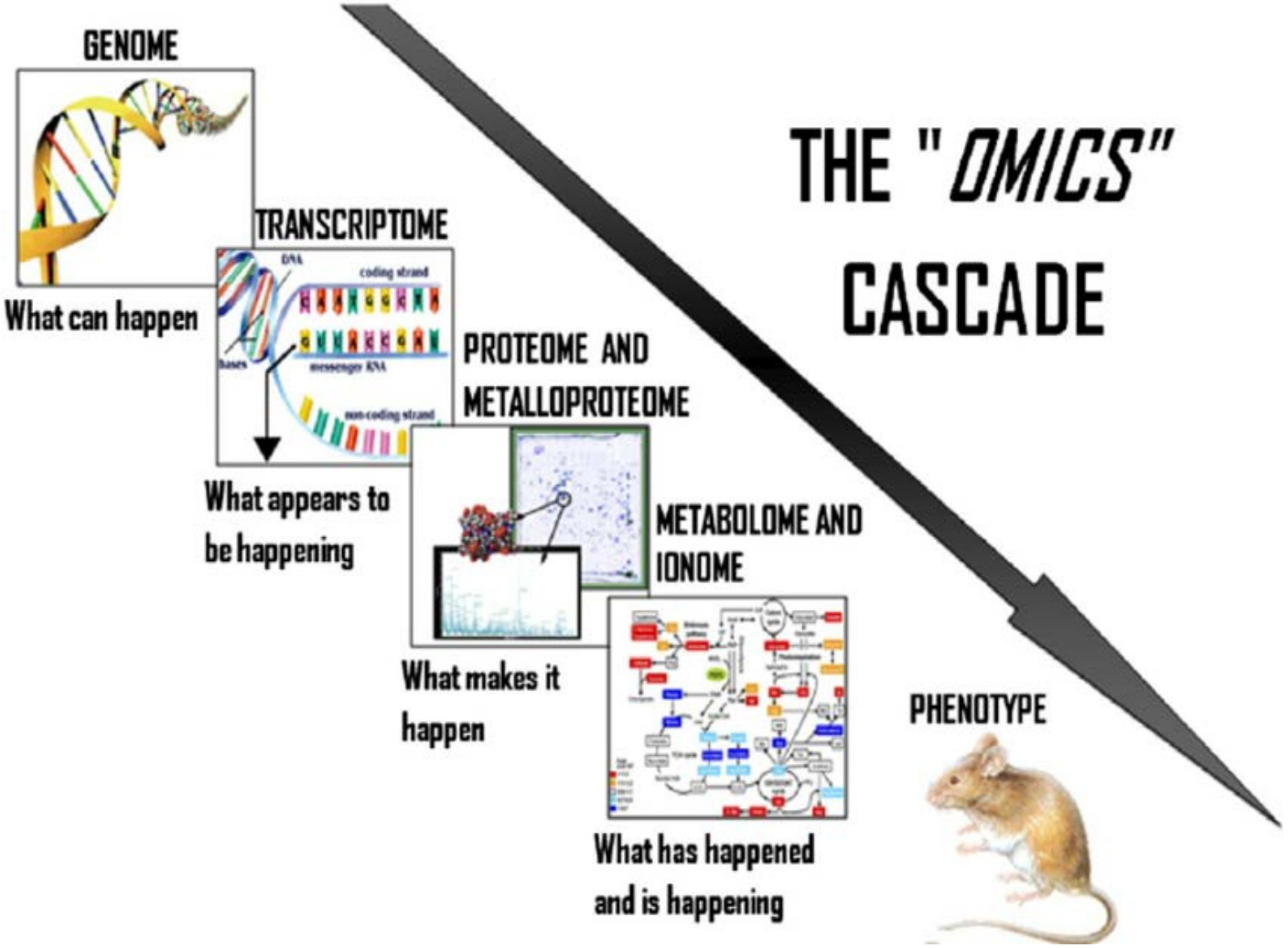
Pathways altered in ALZHEIMER disease

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

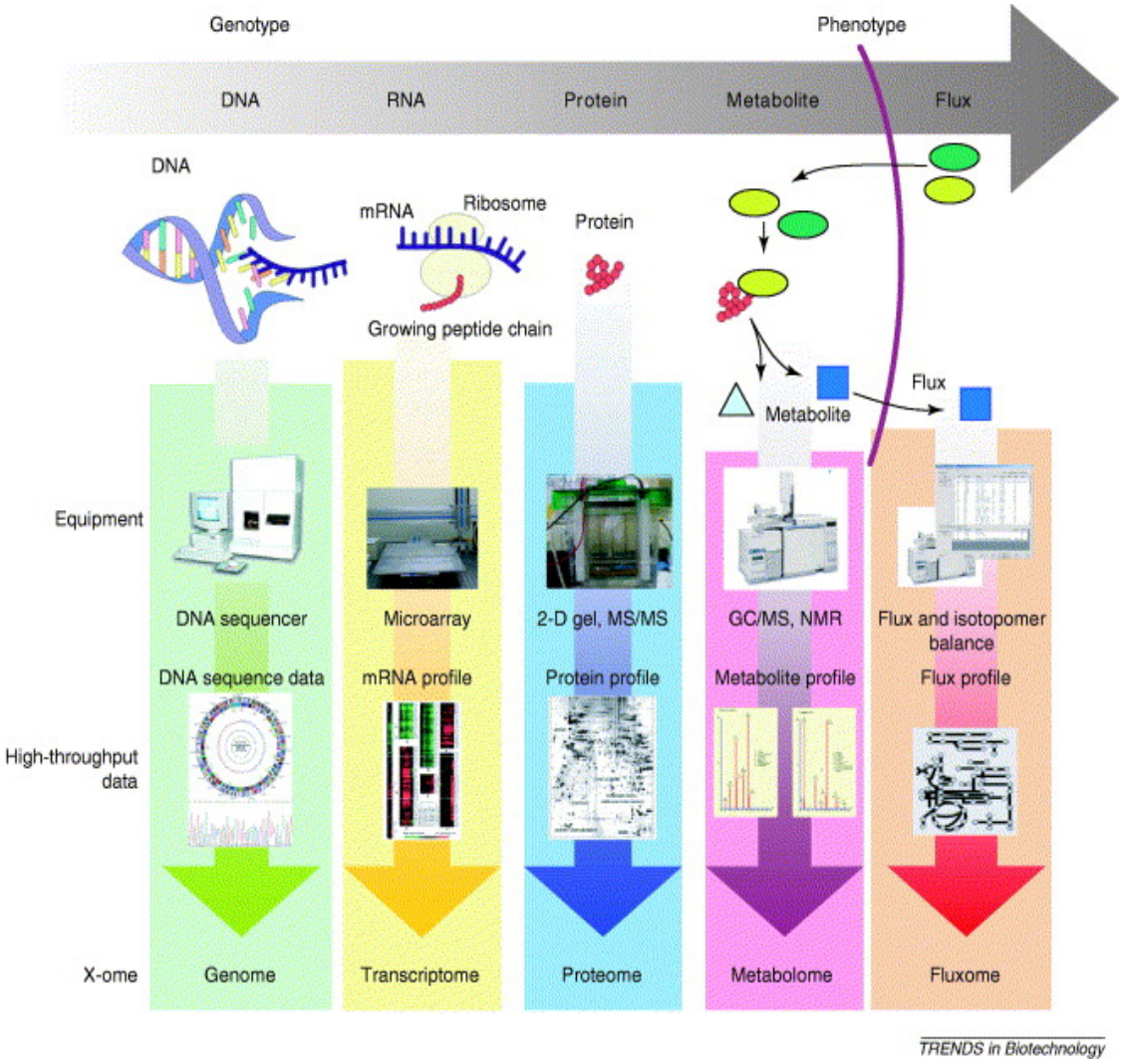
Pathway Analysis

- In spite of our introduction the study of biochemical pathways is not what we understand by 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 different omics technologies.

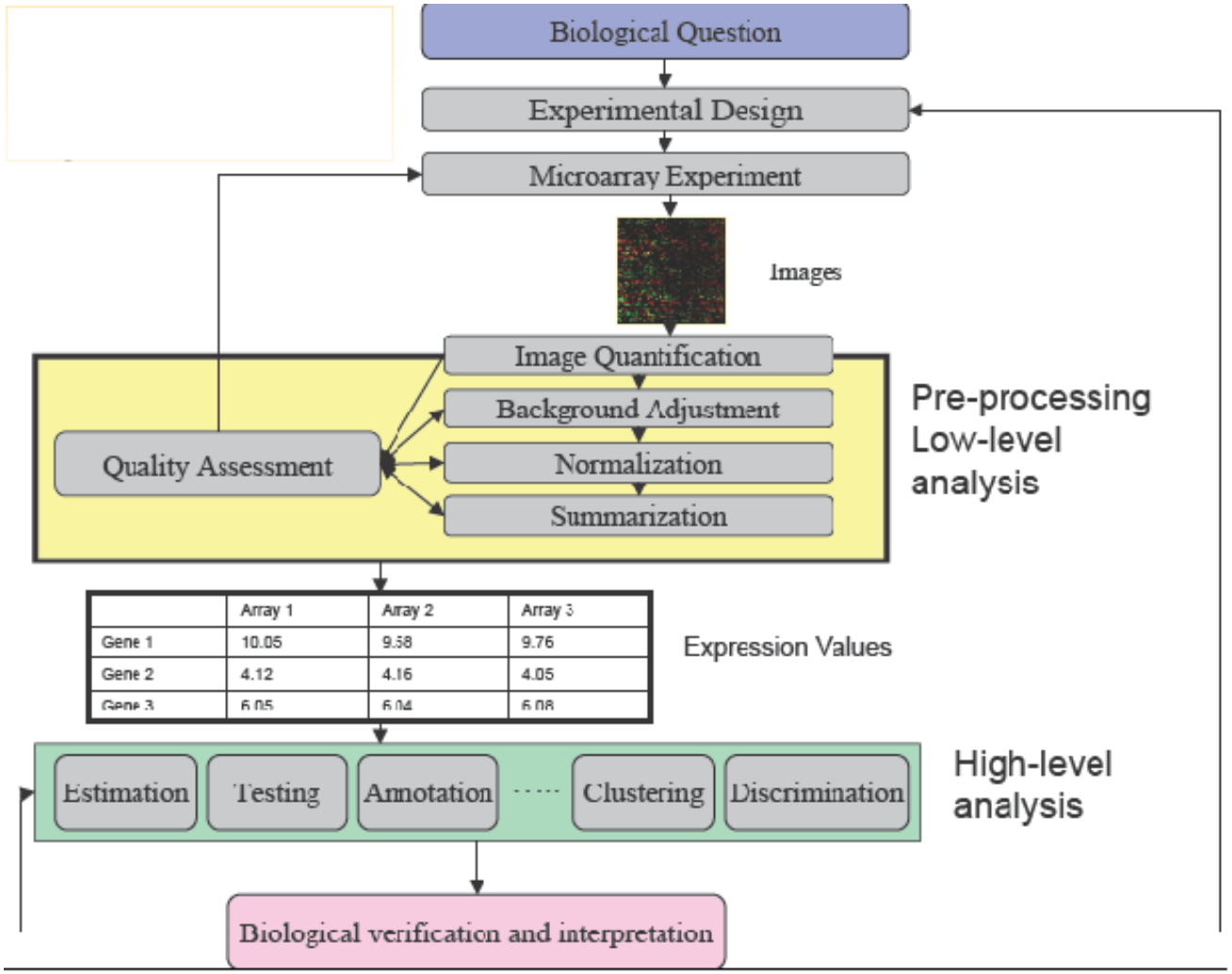
Biology: The “omes”



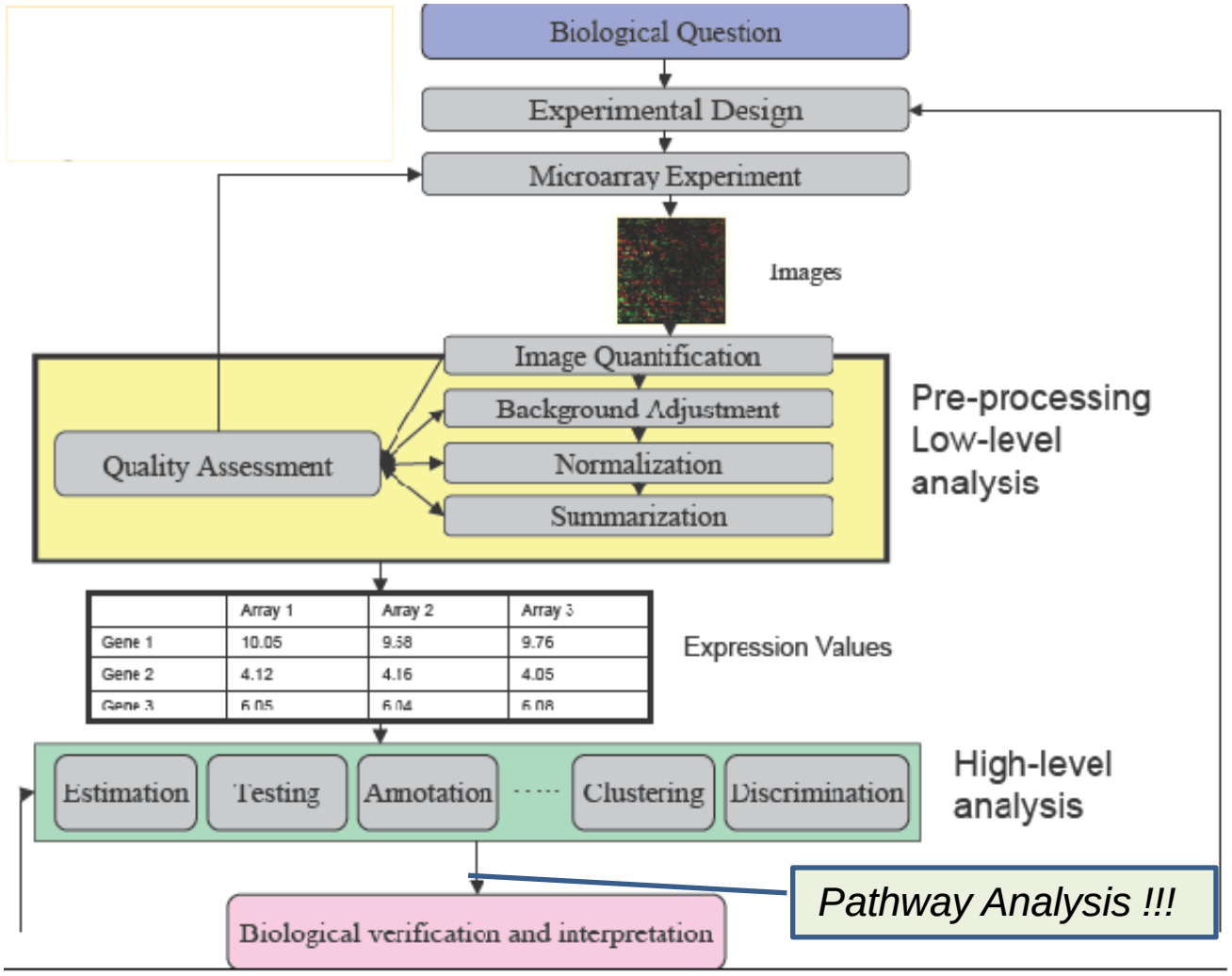
(Bio)technology: The “omics”



The life-cycle of an omics-based study

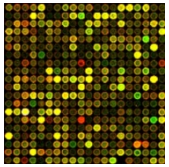


The life-cycle of an omics-based study



Managing Gene Lists

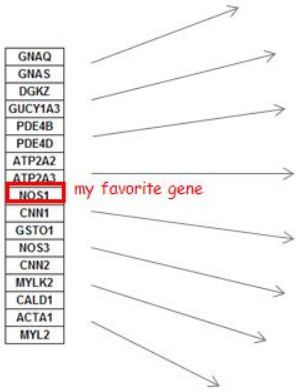
The (in)famous "*where to now?*" question

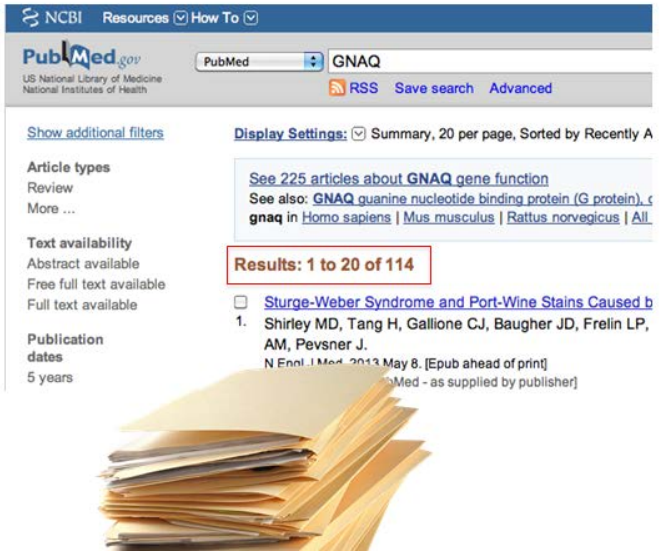


GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

?

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





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***

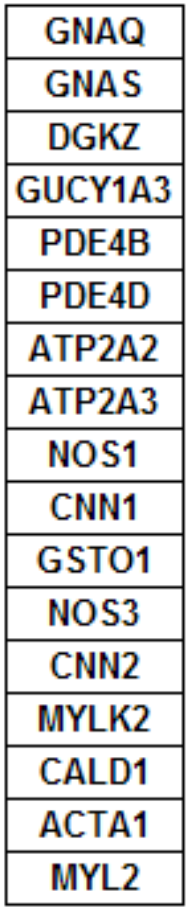
Case study 1

- 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

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.



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

Letters to Nature

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

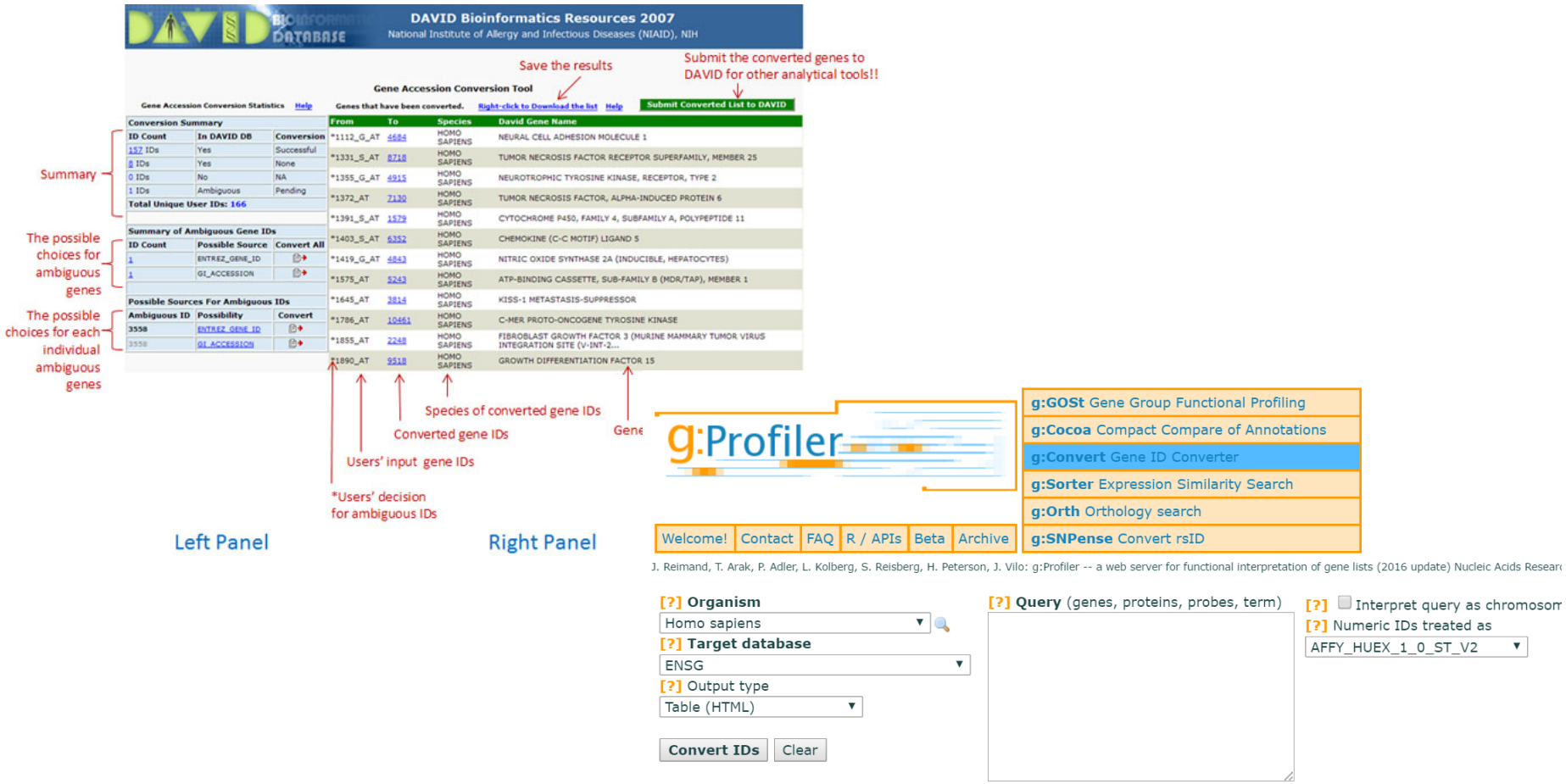
Retraction: Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells

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



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?

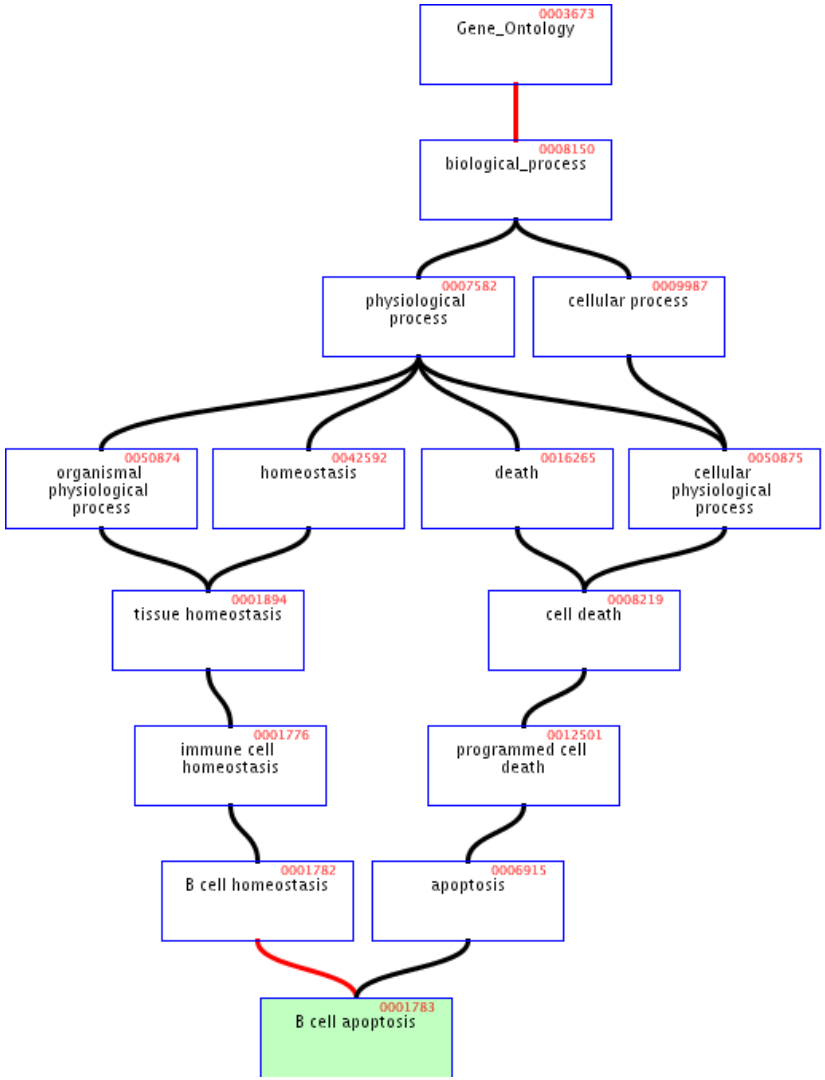
- Pathways
  - Gene Ontology biological process, pathway databases e.g. Reactome
- Other annotations
  - Gene Ontology molecular function, cell location
  - Chromosome position
  - Disease association
  - DNA properties (TF binding sites, gene structure (intron/exon), SNPs, …)
  - Transcript properties (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

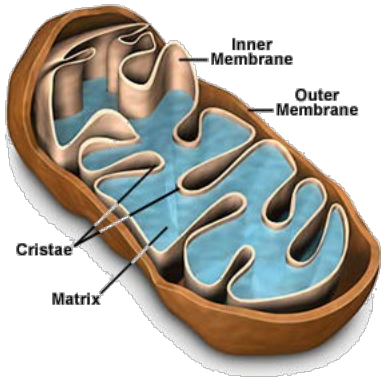
GO Structure

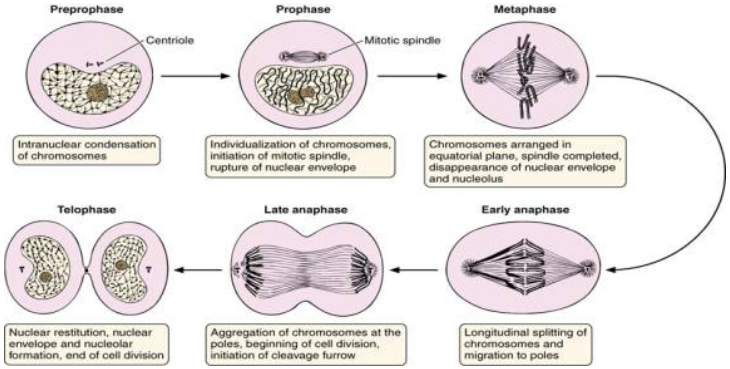
- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



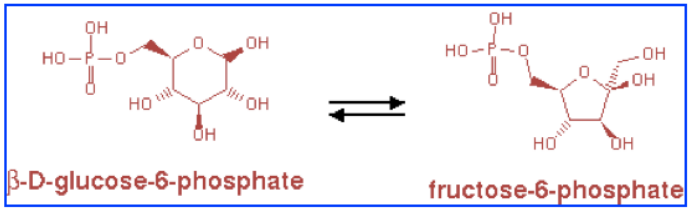
What is covered by the GO?

- 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

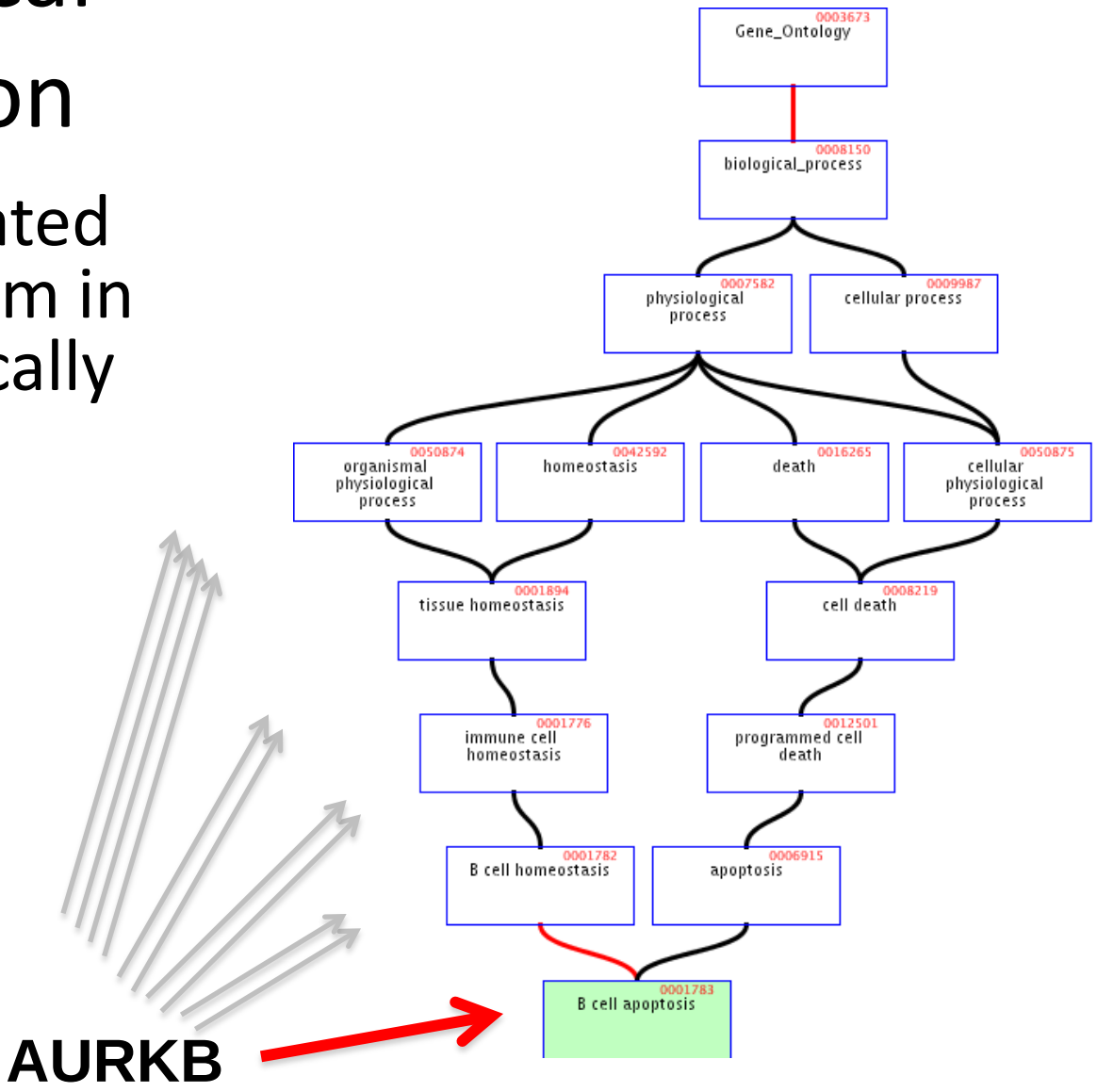
	Jun 2012	Apr 2015	increase
Biological process	23,074	28,158	22%
Molecular function	9,392	10,835	15%
Cellular component	2,994	3,903	30%
total	**37,104**	**42,896**	**16%**

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)

Hierarchical annotation

- Genes annotated to specific term in GO automatically added to *all parents* of that term



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

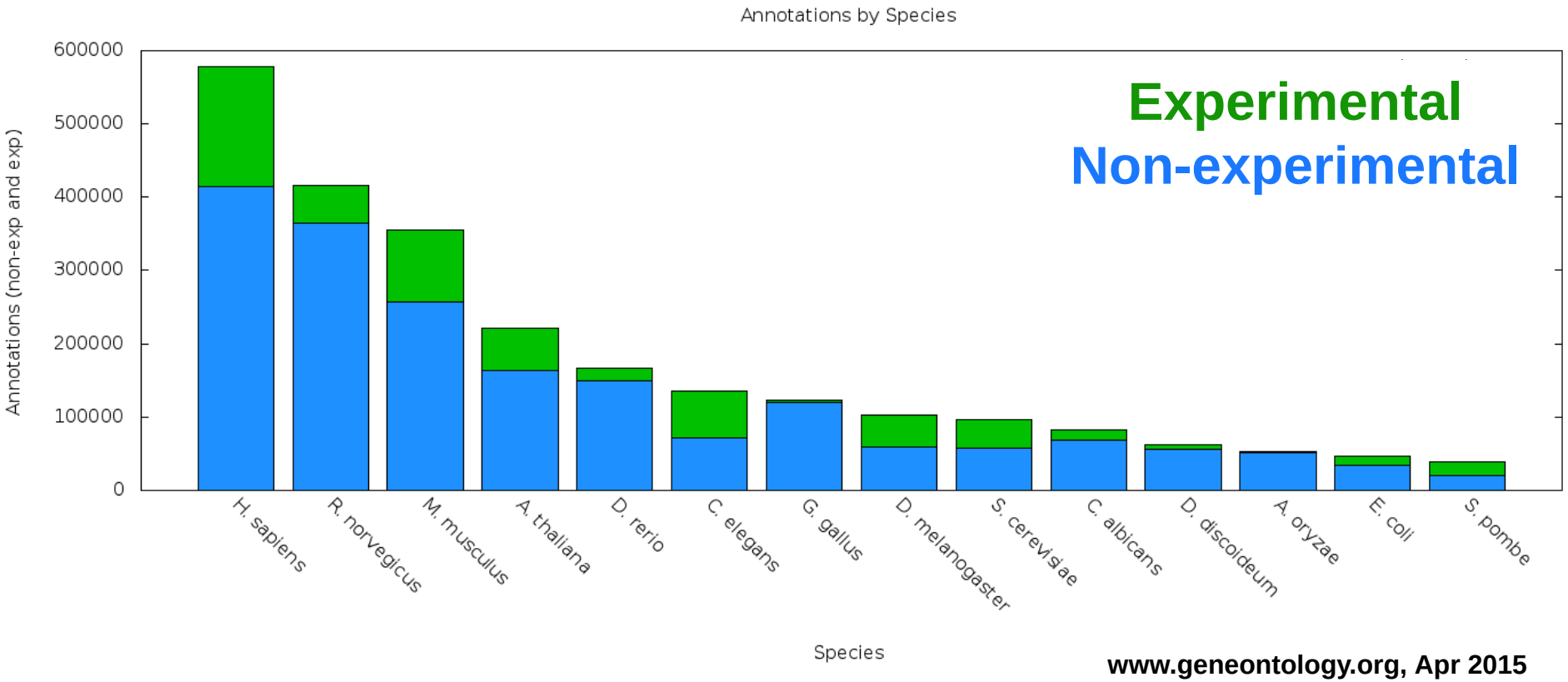
- IEA: Inferred from electronic annotation

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



www.geneontology.org, Apr 2015

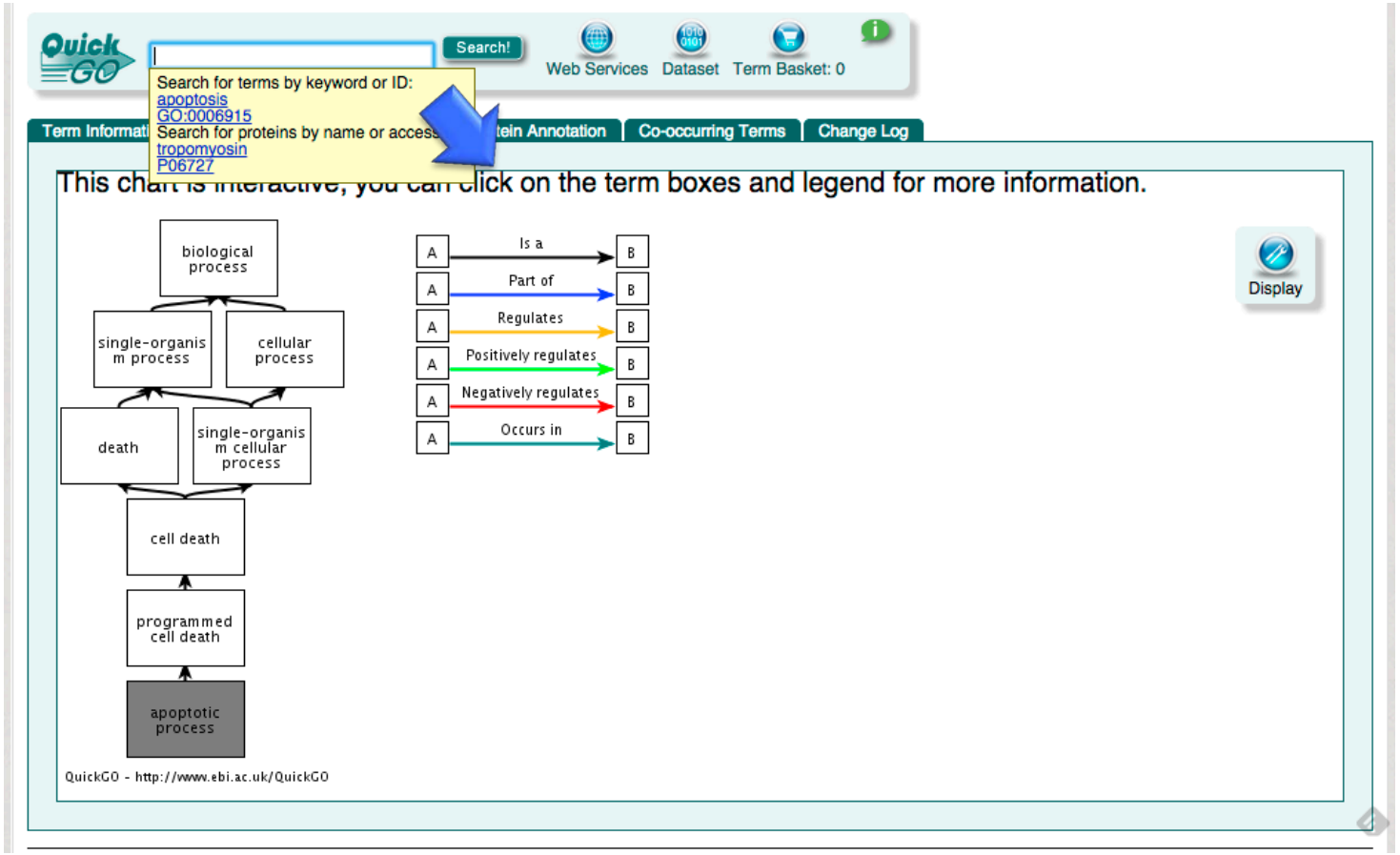
Contributing Databases

- 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 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

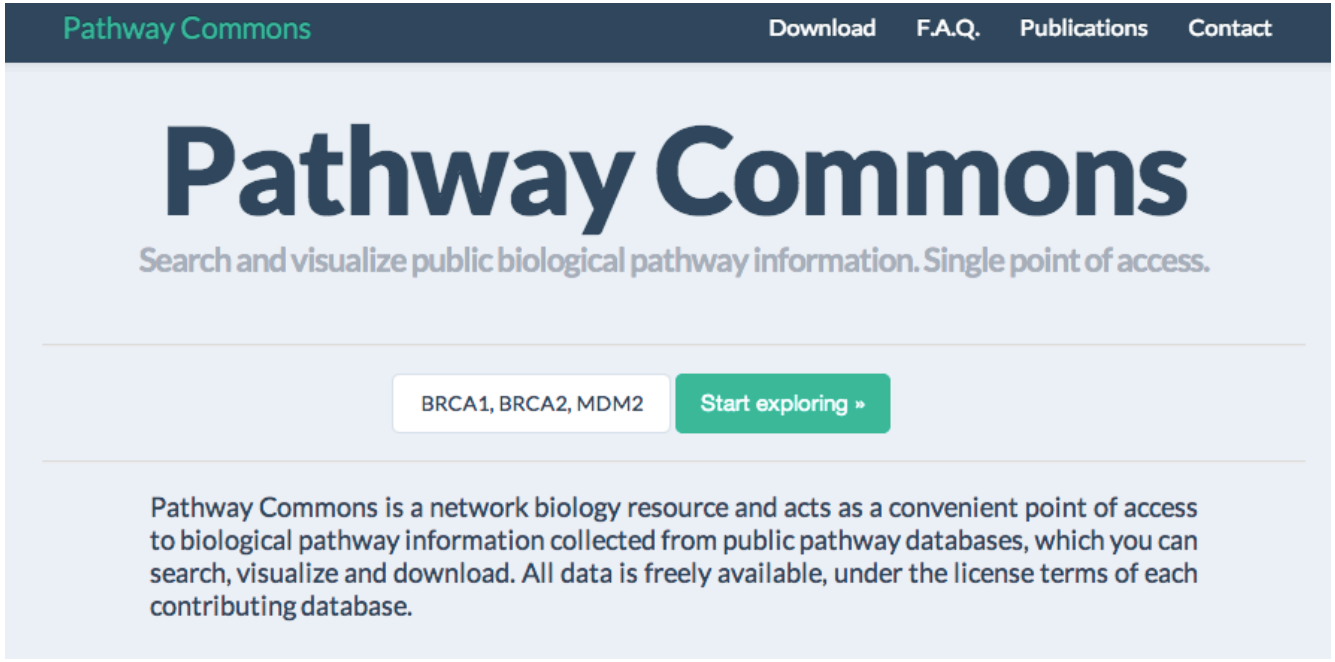
Accessing GO: QuickGO



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 ones

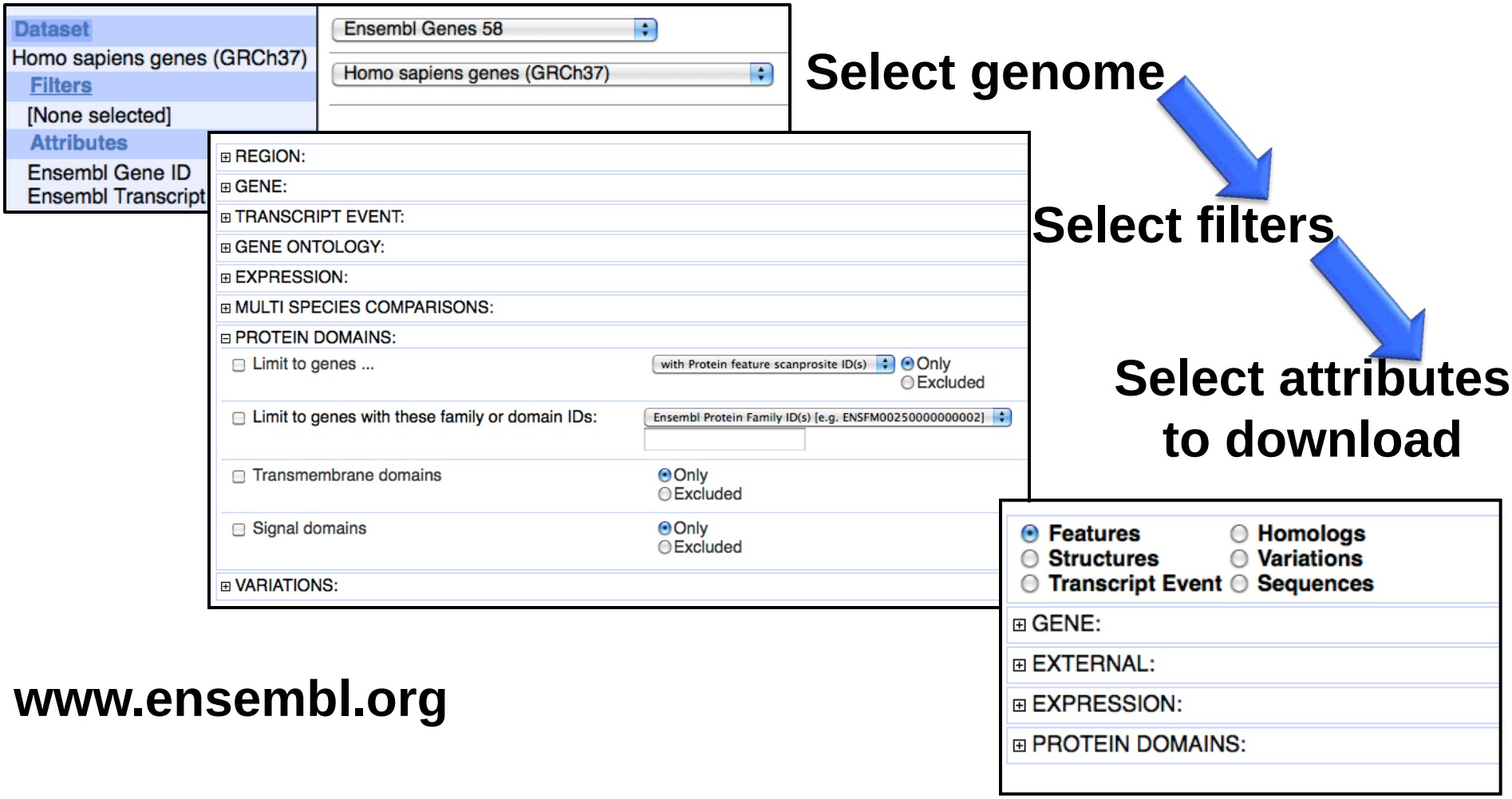


Sources of Gene Attributes

- Ensembl BioMart (general)
  - http://www.ensembl.org
- Entrez Gene (general)
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
- Model organism databases
  - E.g. SGD: http://www.yeastgenome.org/

Ensembl BioMart

- Convenient access to gene list annotation



www.ensembl.org

Pathway Analysis

Overrepresentation Analysis

Gene Set Enrichment Analysis

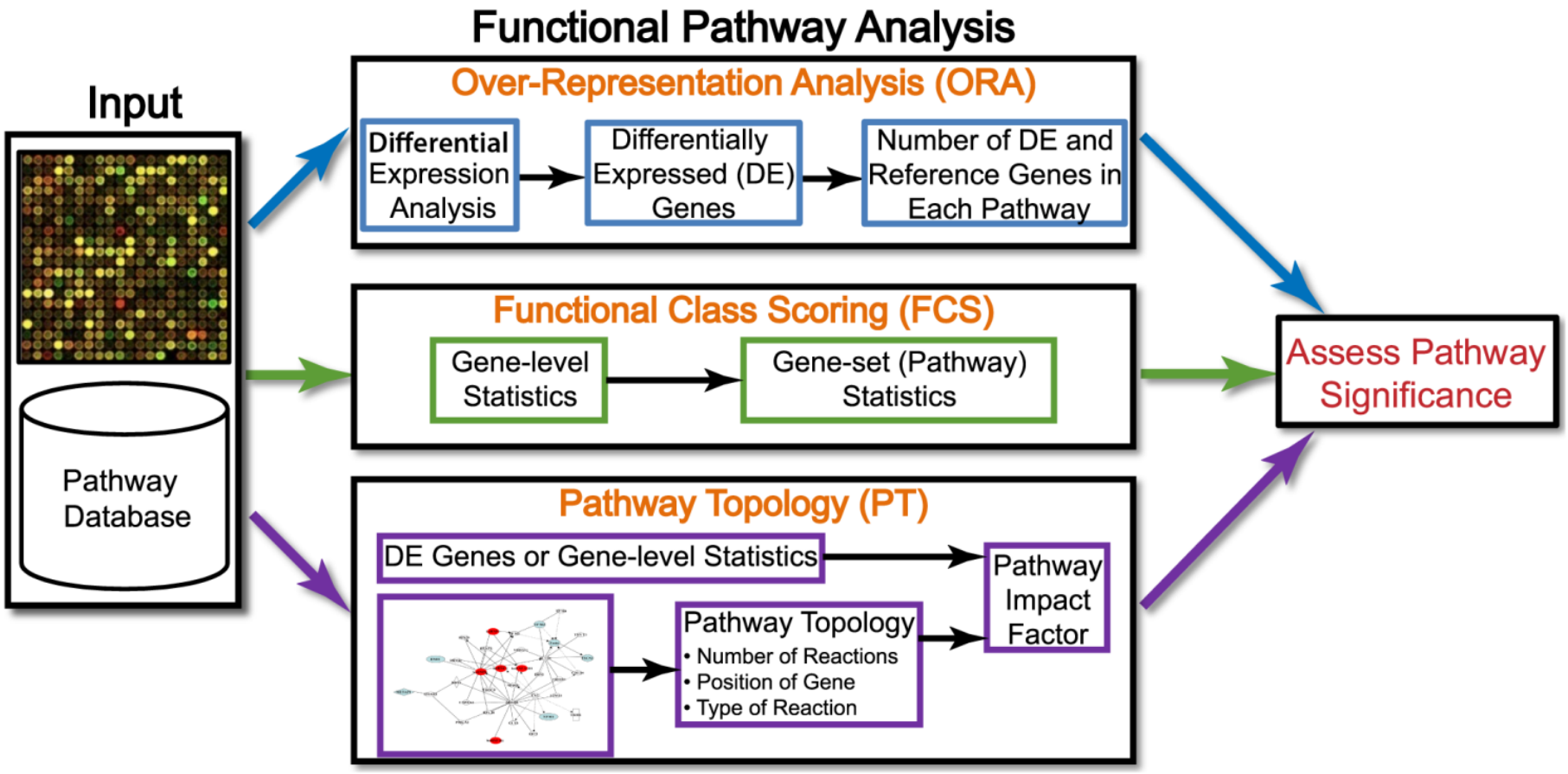
Pathway and Network Analysis

- 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*

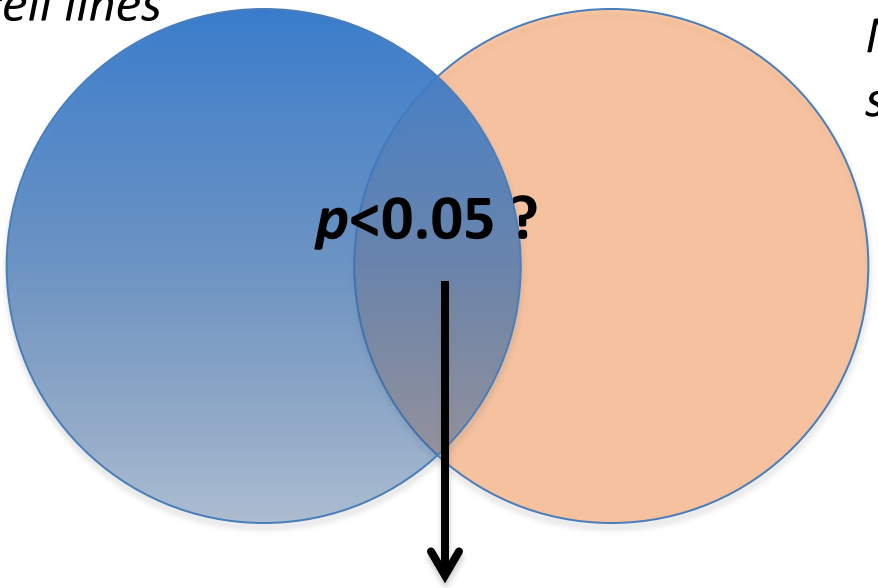
Types of Enrichment Analysis

- **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)

Pathway enrichment analysis

Gene list from experiment:
Genes down-regulated in drug-sensitive brain cancer cell lines

Pathway information:
All genes known to be involved in Neurotransmitter signaling



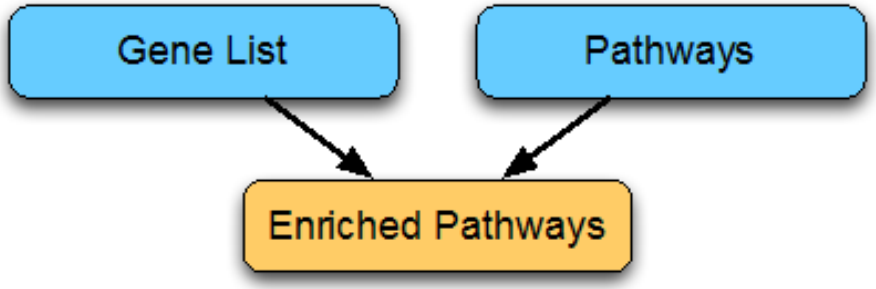
Test many pathways

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

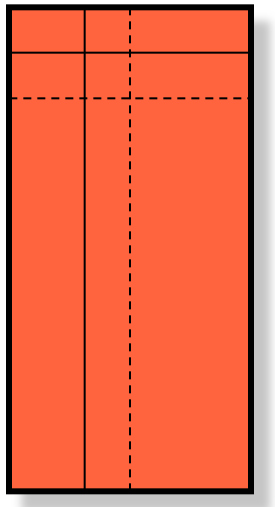
Hypothesis: drug sensitivity in brain cancer is related to reduced neurotransmitter signaling

Pathway Enrichment Analysis

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

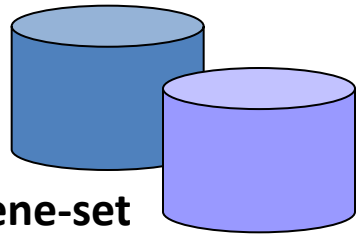


Enrichment Test



Microarray
Experiment
(gene expression table)

ENRICHMENT
TEST



Gene-set
Databases

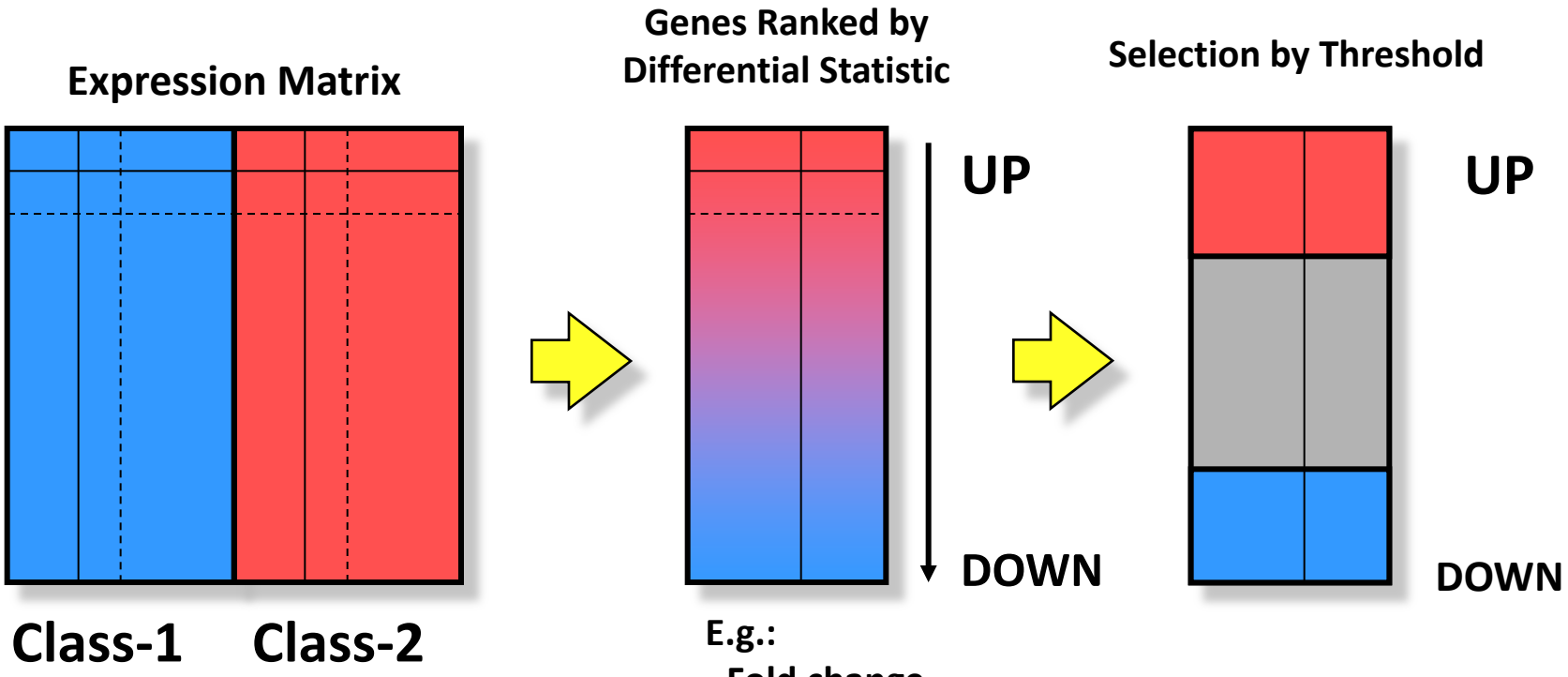
Spindle	0.00001
Apoptosis	0.00025

Enrichment Table

Gene list enrichment analysis

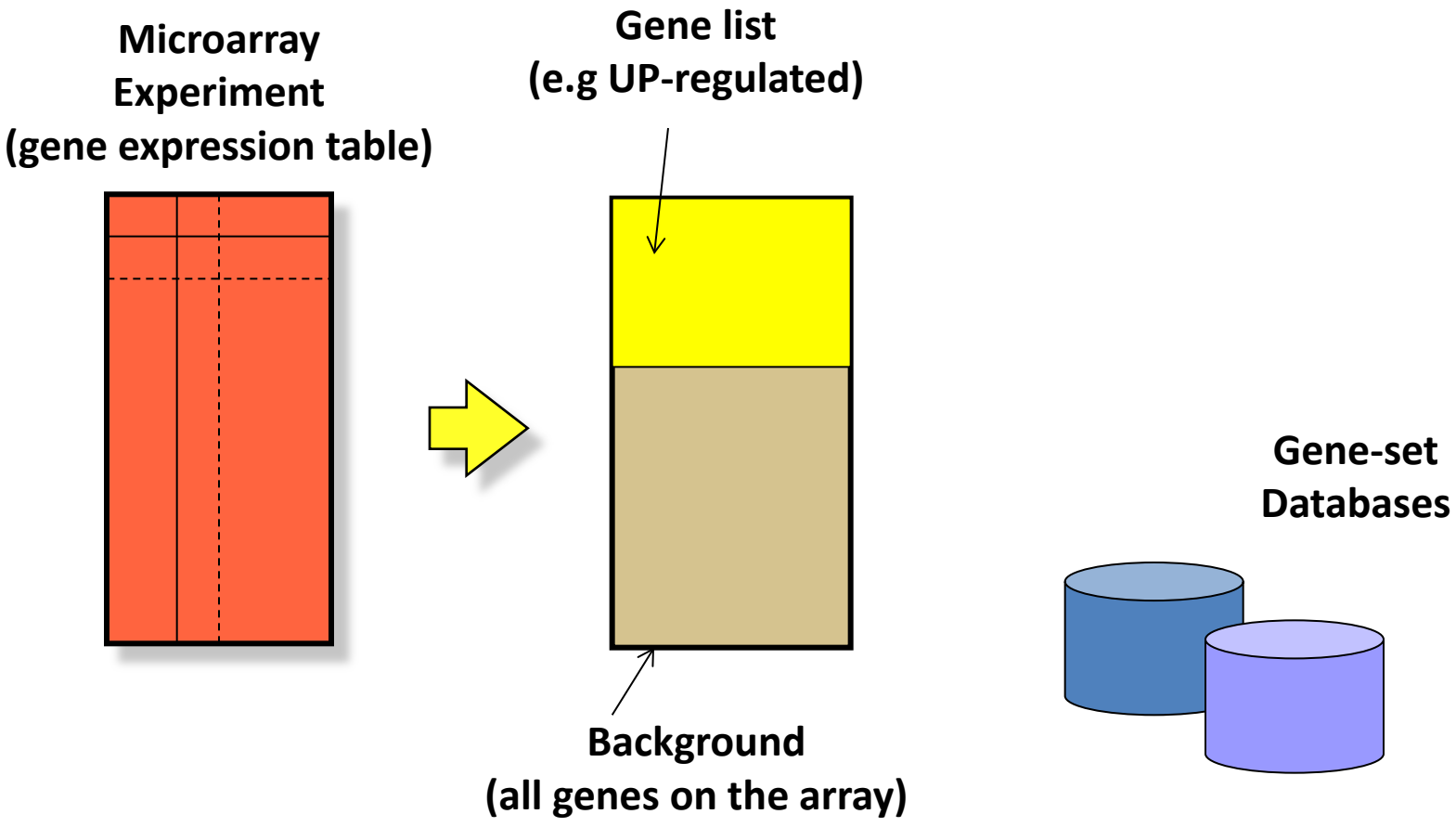
- Given:
  1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast)
  2. Gene sets or annotations: e.g. Gene ontology, transcription factor binding sites in promoter
- Question: *Are any of the gene annotations surprisingly enriched in the gene list?*
- Details:
  - Where do the gene lists come from?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

Two-class design for gene lists

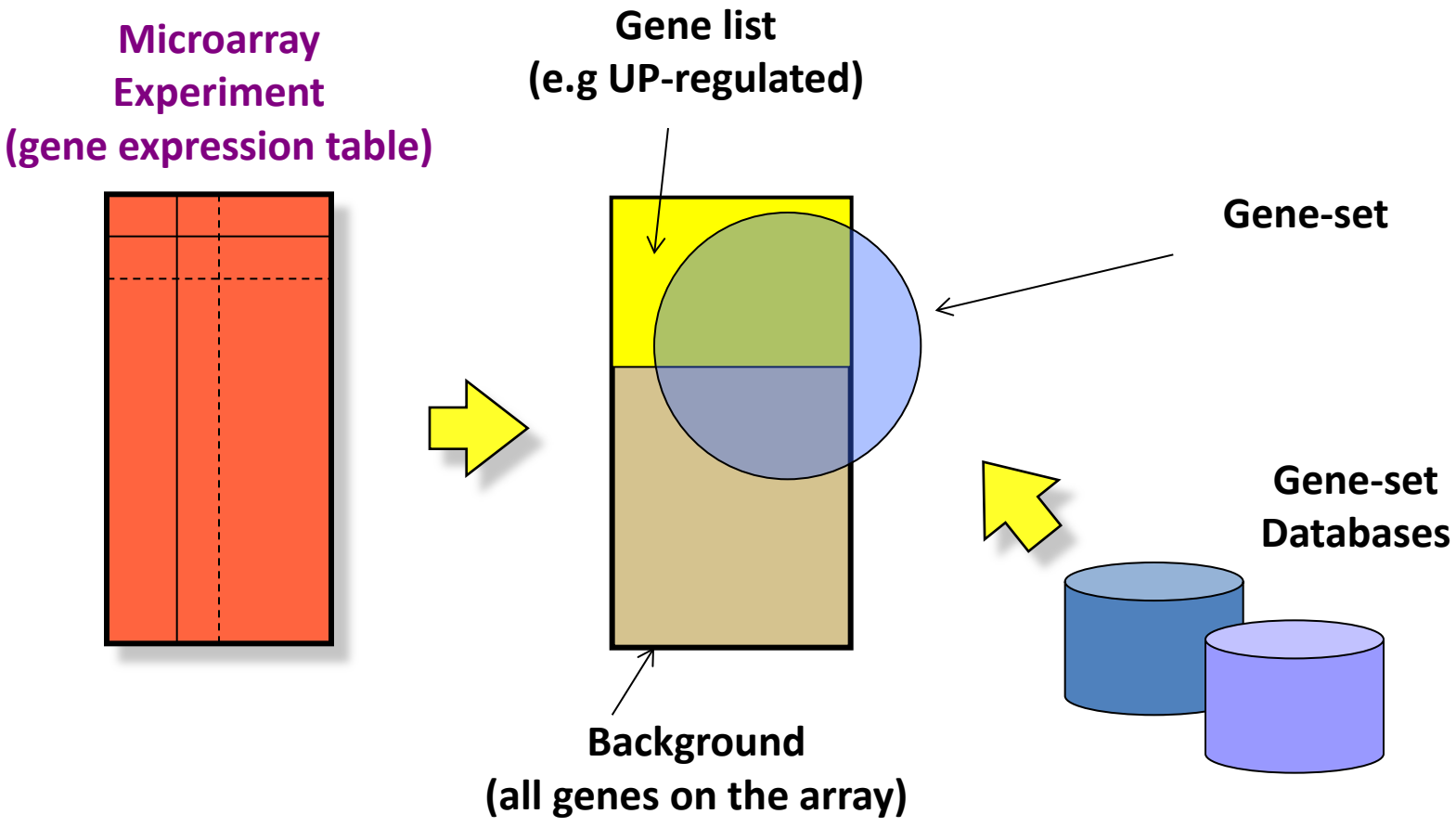


E.g.:
- **Fold change**
- **Log (ratio)**
- **t-test**
-**Significance analysis of microarrays**

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 by hypergeometric testing
  - Compute FDR to adjust p-values for doing many tests

	Differentially 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!

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

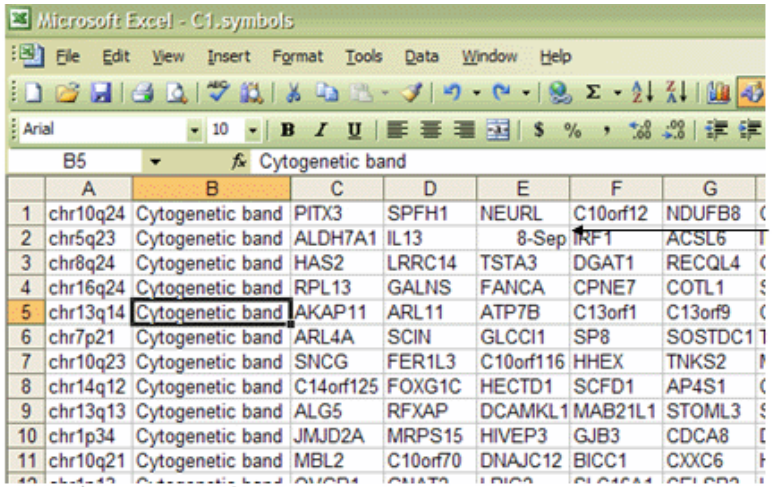
- No “natural” value for the threshold
- Possible loss of statistical power due to thresholding
  - No resolution between significant signals with different strengths
  - Weak signals neglected
- Different results at different threshold settings
- Based on the wrong assumption of independent gene (or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

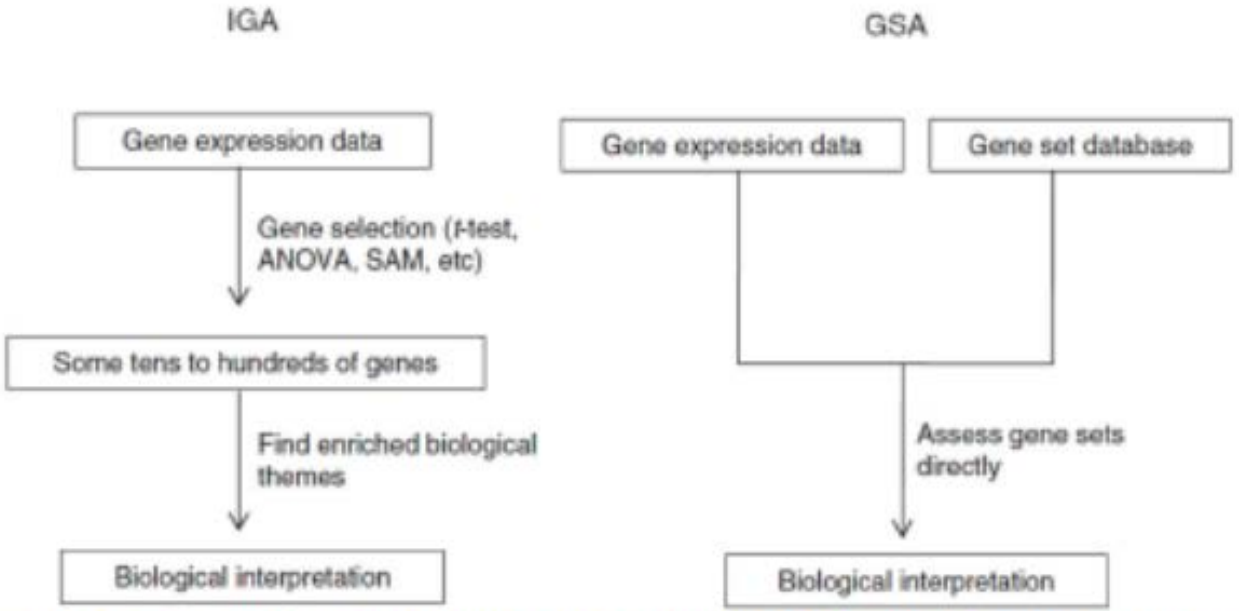
Second column contains a brief description. Its optional – you can fill in a dummy field (e.g. "na")

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

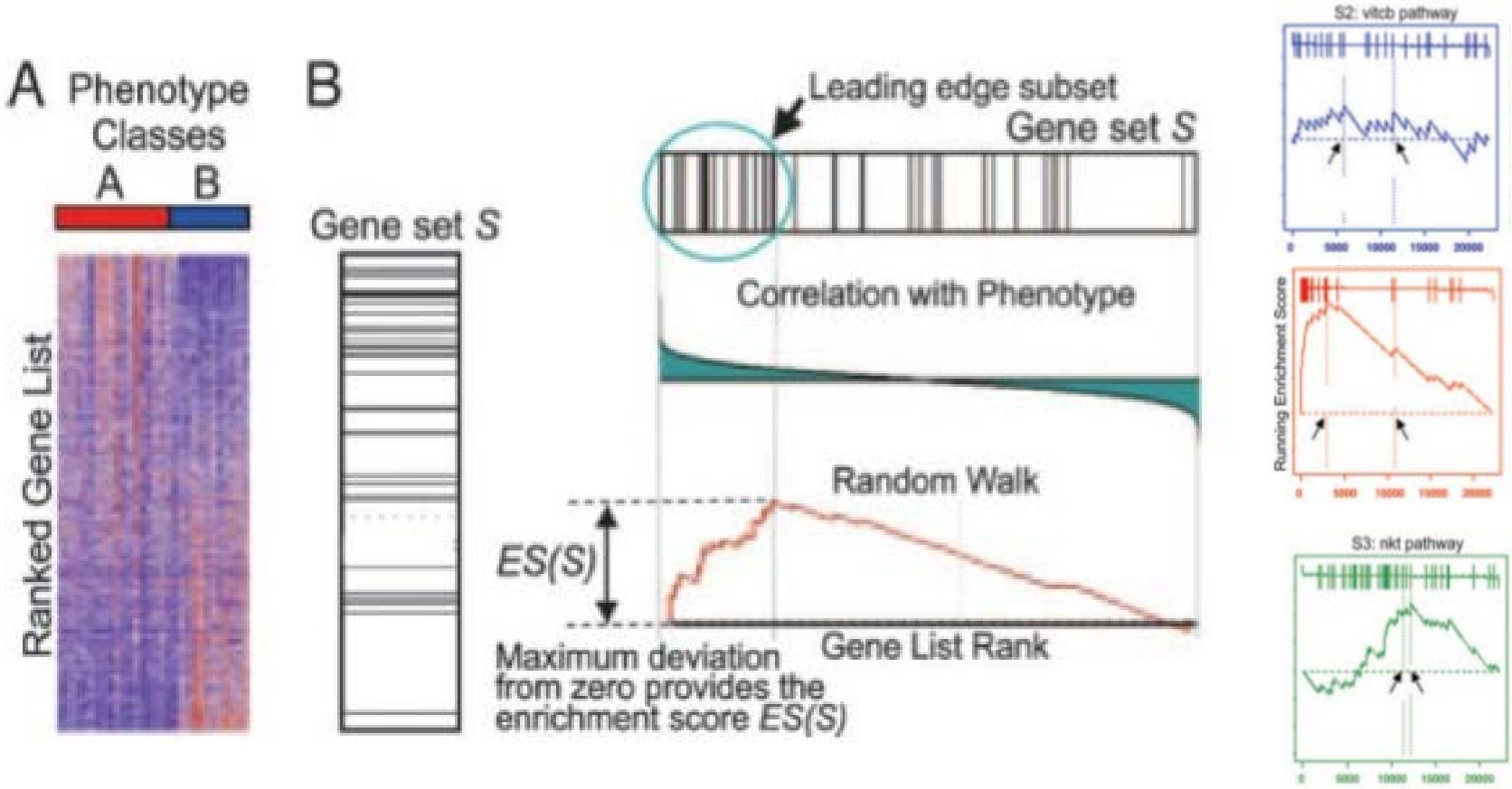


Source: Dougu Nam and Seon-Young Kim, Gene-set approach for expression pattern analysis, Briefings in Bioinformatics. 2008, 9(3). 189-197.

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 $\sqrt{\frac{N-G}{G}}$ is added
  - If a gene does not belong to gene set a quantity $\sqrt{\frac{G}{N-G}}$ 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…

Tools for Pathway Analysis

R/Bioconductor

BABELOMICS (FATIGO et alt.)

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%
SP1 = 20%

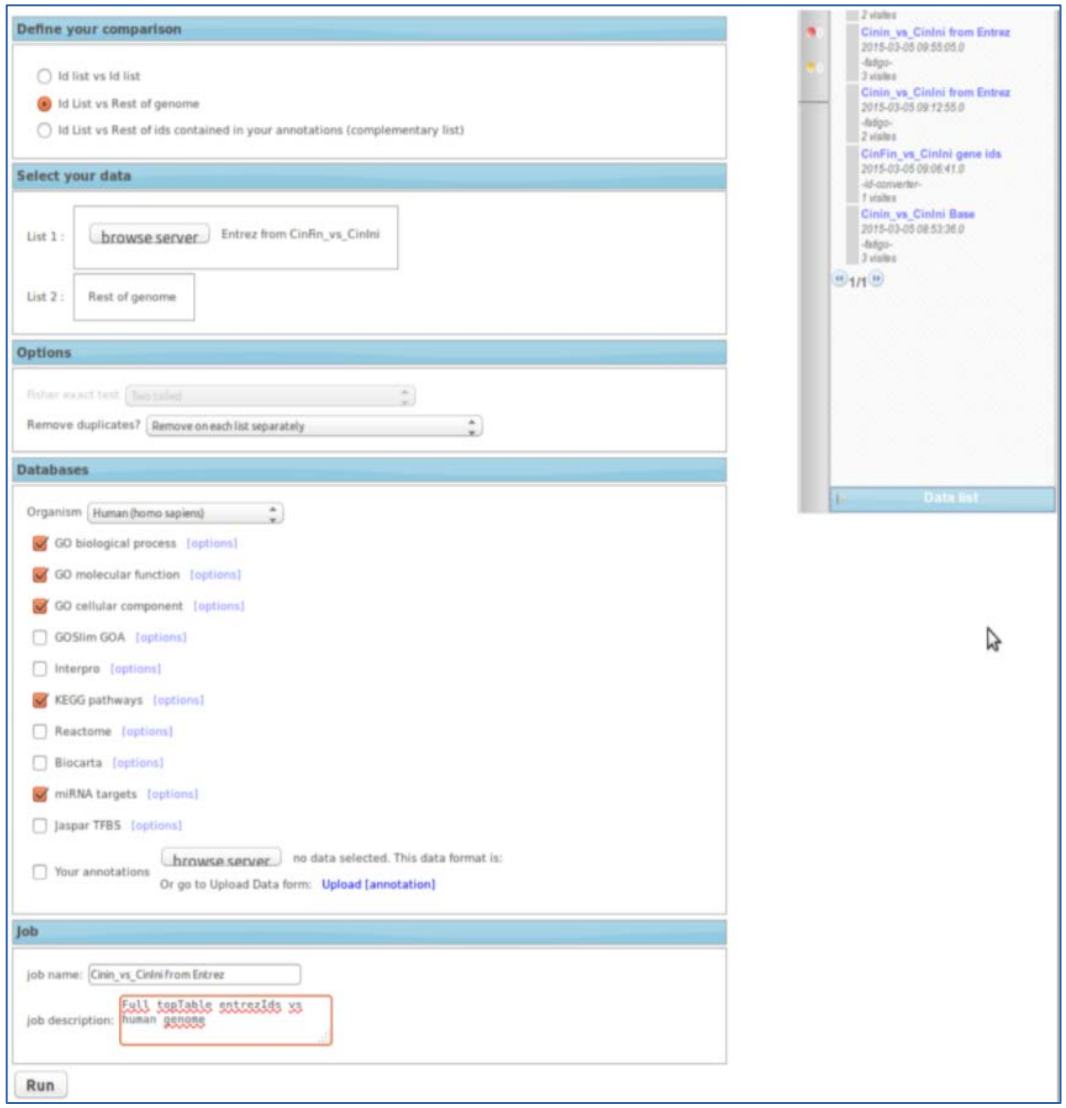
GATA1 = 10%
SP1 = 20%

The gene list is over-enriched for GATA1 targets.

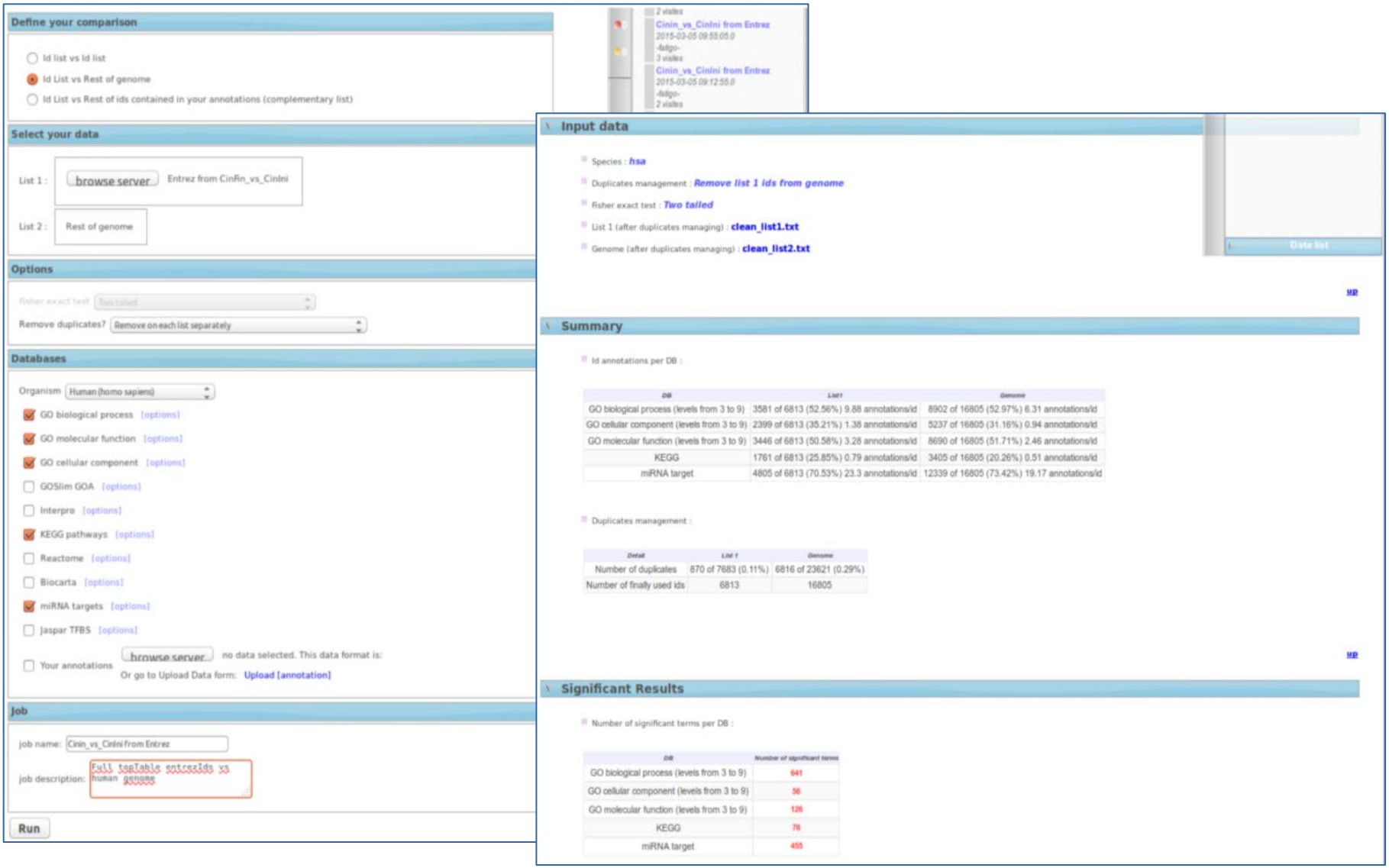
	Gene list	Genome
Regulated by GATA1	4	2
Not regulated by GATA1	6	18

- 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.

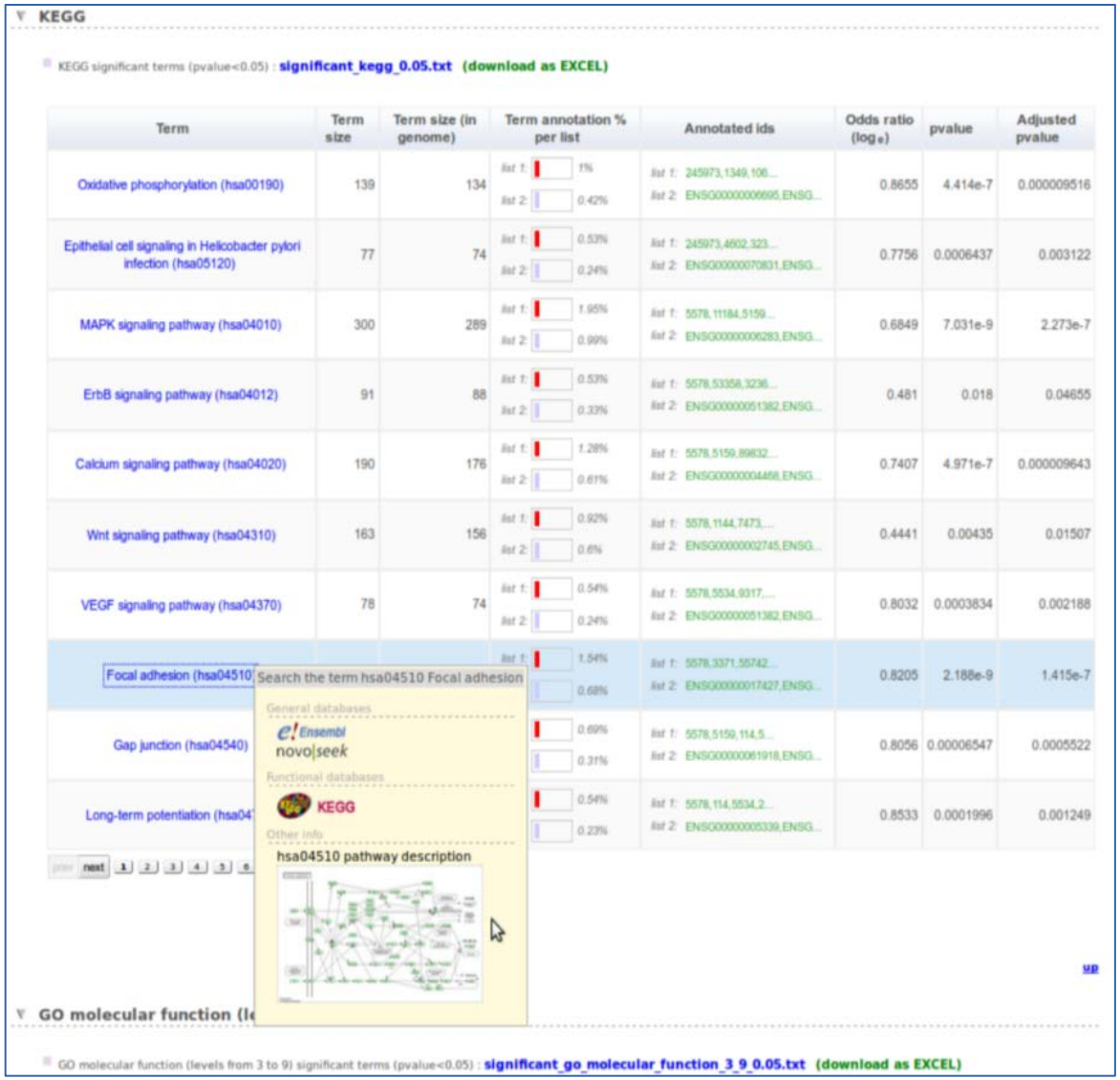
Upload gene list, set parameters



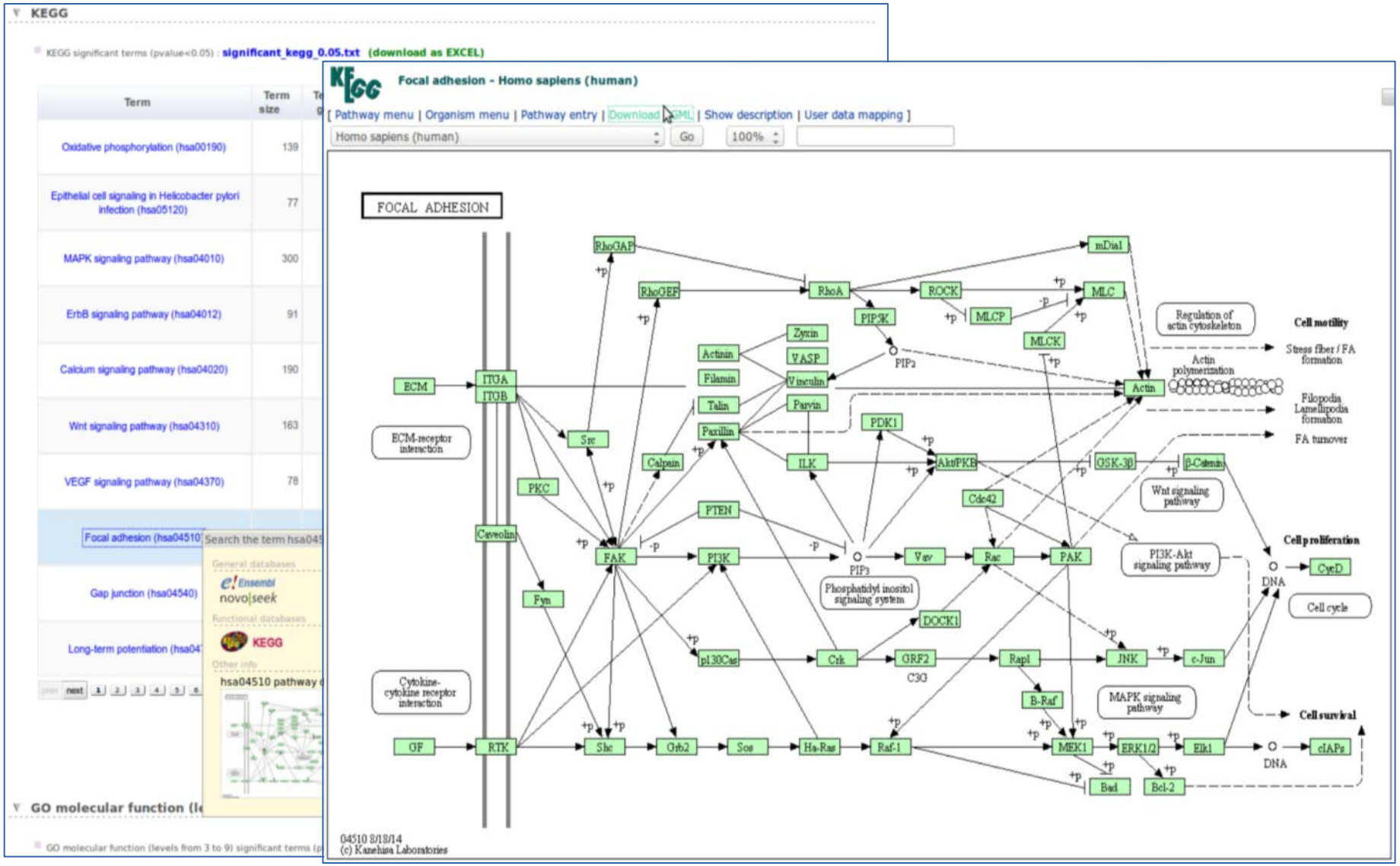
Obtain “significantly enriched” sets



Visualize results

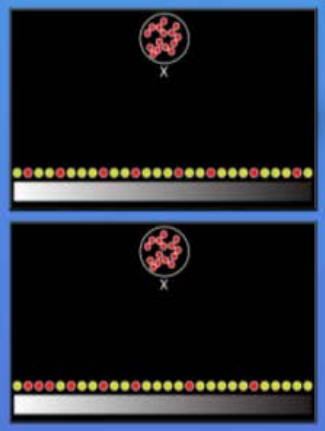


Visualize results



“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			

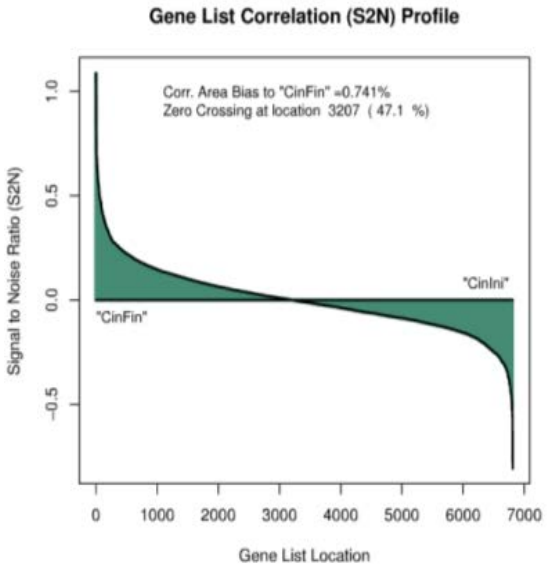
Run.A276bis.R

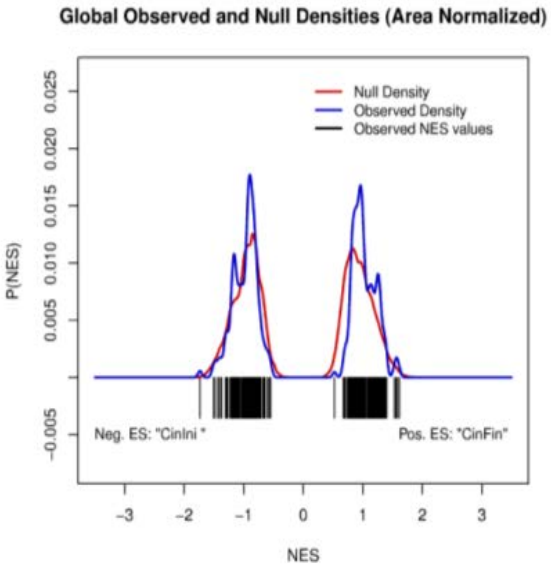
Source on Save | Run | Source

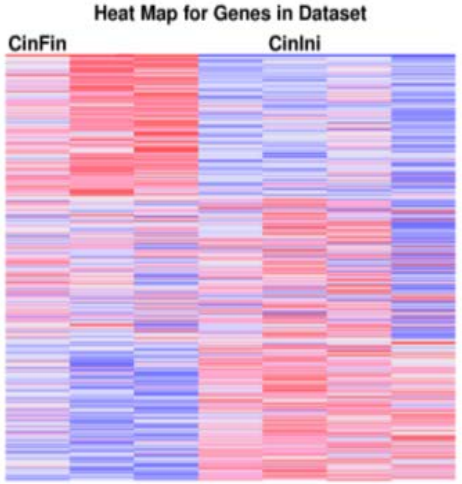
```
# GSEA 1.0 -- Gene Set Enrichment Analysis / Broad Institute
#
# R script to run custom GSEA analysis of the UEB study ID #A276,
# based on the R script to run GSEA Analysis of the Leukemia ALL/AML vs C1 example

GSEA.program.location <- "/home/ferran/gsea_home/GSEA-P-R/GSEA.1.0.R"   #  R source program
source(GSEA.program.location, verbose=T, max.deparse.length=9999)

GSEA(                                  # Input/Output Files :-------------------------------
# input.ds = "/home/ferran/estudis/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSEA/Human
 input.ds = "/home/ferran/estudis/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSEA/Human.
 input.cls = "/home/ferran/estudis/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSEA/VehFi
# gs.db = "/home/ferran/estudis/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSEA/geneSets
 gs.db = "/home/ferran/gsea_home/GSEA-P-R/GeneSetDatabases/c2.all.v4.0.entrez.gmt",         # Gene set
# output.directory = "/home/ferran/estudis/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GS
 output.directory = "/home/ferran/estudis/microarrays/2015-01-MartaGarcia-StJdDeu-A161-A276/results/GSE
# Program parameters :-------------------------------------------------------------------
 doc.string            = "A276.VehFin.vs.VehIni",   # Documentation string used as a prefix to name res
 non.interactive.run   = F,                 # Run in interactive (i.e. R GUI) or batch (R command line) m
 reshuffling.type      = "sample.labels", # Type of permutation reshuffling: "sample.labels" or "gene.l
 nperm                 = 1000,              # Number of random permutations (default: 1000)
 weighted.score.type   =  1,                # Enrichment correlation-based weighting: 0=no weight (KS), 1
 non.p.val.threshold   = -1,                # Significance threshold for nominal p-vals for gene sets (de
 fwer.p.val.threshold  = -1,                # Significance threshold for FWER p-vals for gene sets (defau
 fdr.q.val.threshold   = 0.25,              # Significance threshold for FDR q-vals for gene sets (defaul
 topgs                 = 10,                # Besides those passing test, number of top scoring gene sets
 adjust.FDR.q.val      = F,                 # Adjust the FDR q-vals (default: F)
 gs.size.threshold.min = 25,                # Minimum size (in genes) for database gene sets to be consid
 gs.size.threshold.max = 1500,              # Maximum size (in genes) for database gene sets to be consid
 reverse.sign          = F,                 # Reverse direction of gene list (pos. enrichment becomes neg
 preproc.type          = 0,                 # Preproc.normalization: 0=none, 1=col(z-score)., 2=col(rank)
 random.seed           = 123456,              # Random number generator seed. (default: 123456)
 perm.type             = 0,                 # For experts only. Permutation type: 0 = unbalanced, 1 = bal
 fraction              = 1.0,               # For experts only. Subsampling fraction. Set to 1.0 (no resa
 replace               = F,                 # For experts only, Resampling mode (replacement or not repla
 save.intermediate.results = F,             # For experts only, save intermediate results (e.g. matrix of
 OLD.GSEA              = F,                 # Use original (old) version of GSEA (default: F)
 use.fast.enrichment.routine = T            # Use faster routine to compute enrichment for random permuta
)
#-------------------------------------------------------------------------------------------


```

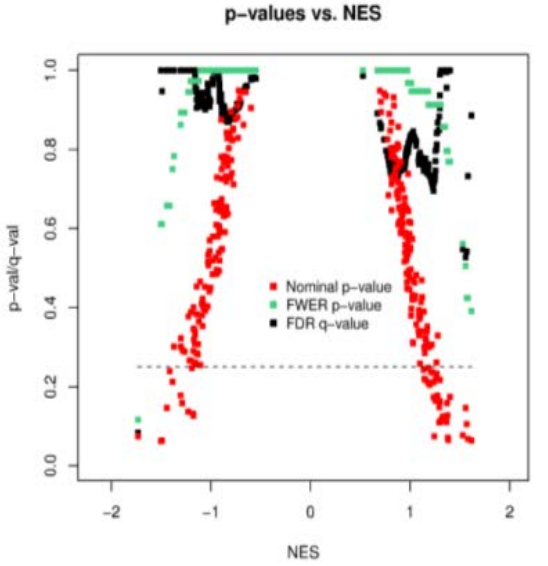
1:1 | (Top Level) | R Script

Get results. Interpret output









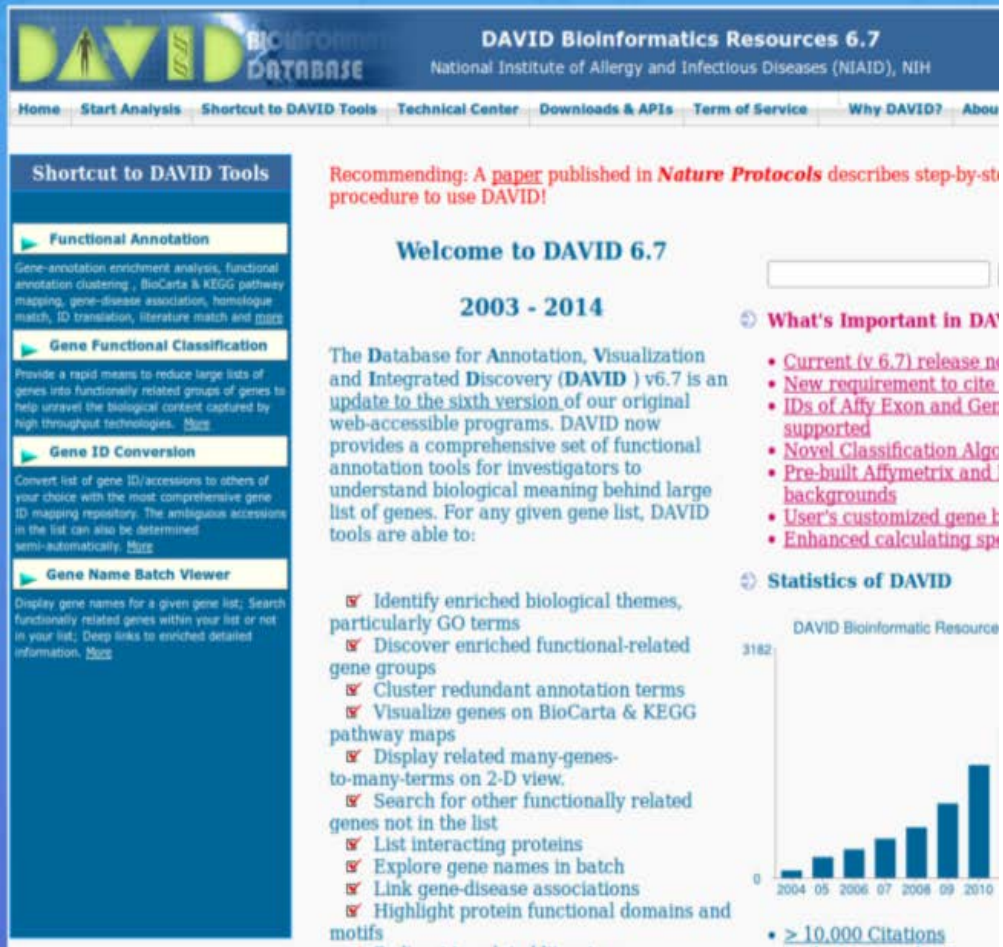
DAVID

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.

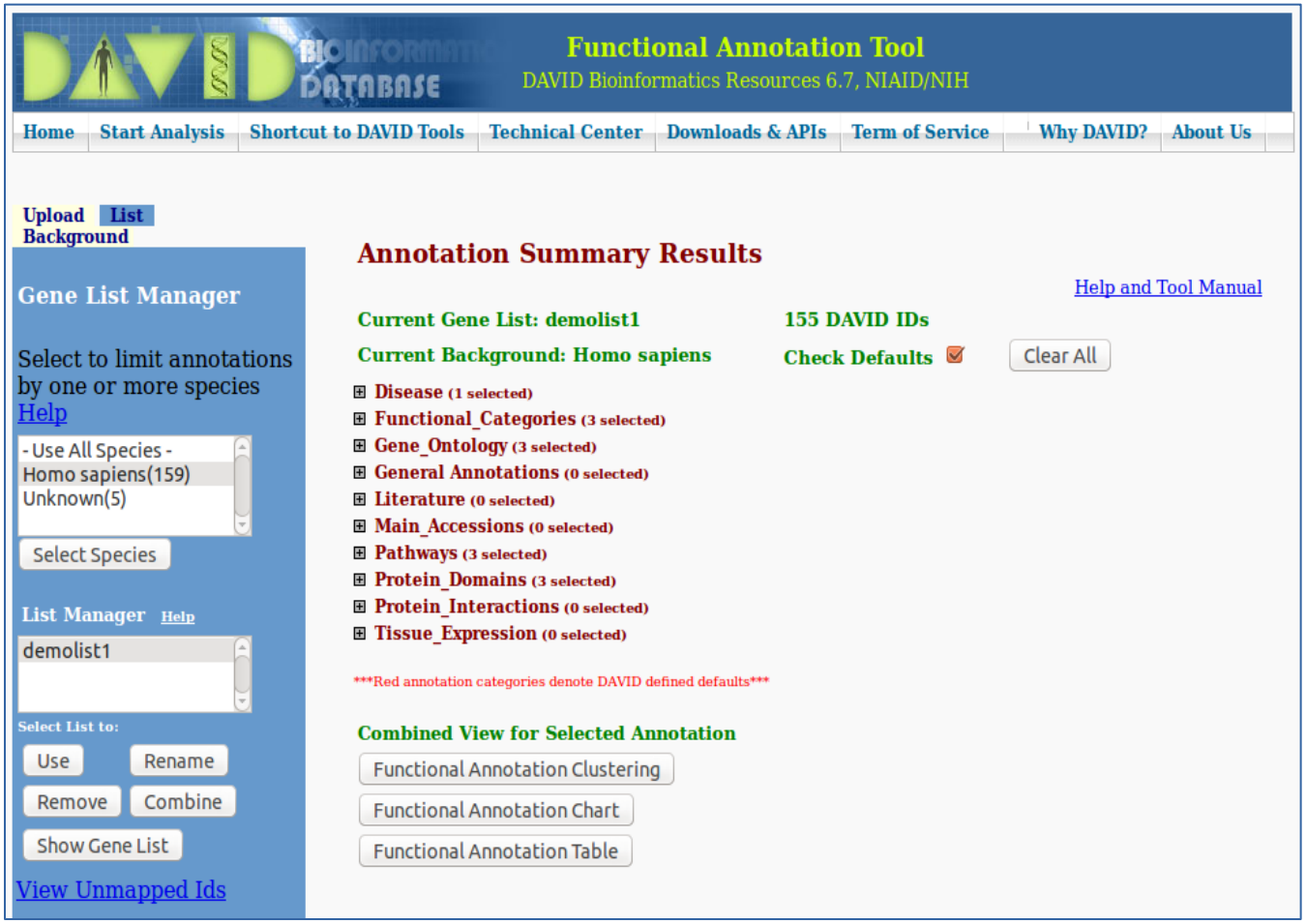
Functional Annotation Clustering:
measures relationships among the annotation terms based on the degrees of their co-association genes to group the similar, redundant, and heterogeneous annotation contents from the same or different resources into annotation groups.

- 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

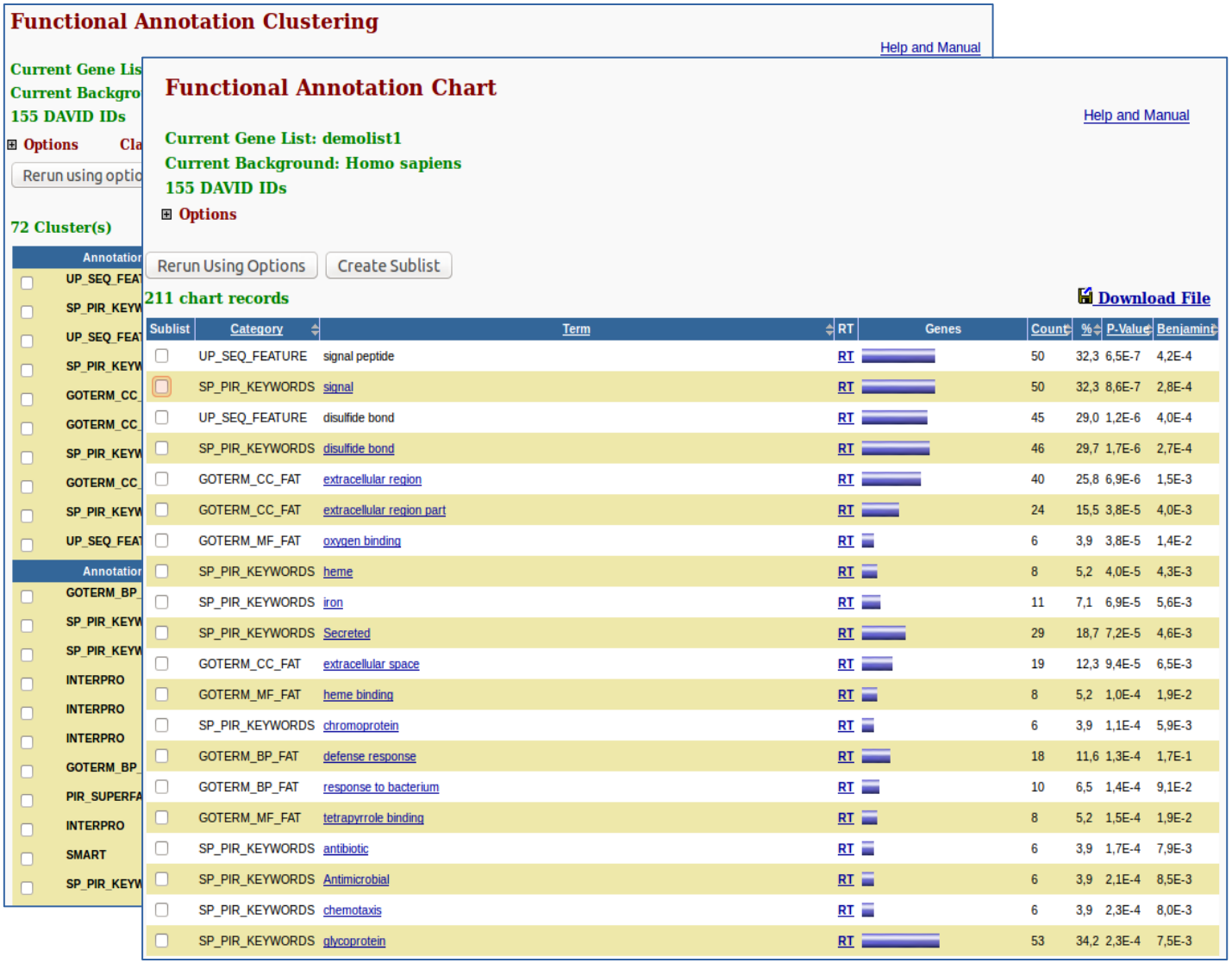
⊞ **Options** **Classification Stringency** Medium

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 optic

72 Cluster(s)

Annotation
UP_SEQ_FEA
SP_PIR_KEYW
UP_SEQ_FEA
SP_PIR_KEYW
GOTERM_CC_
GOTERM_CC_
SP_PIR_KEYW
GOTERM_CC_
SP_PIR_KEYW
UP_SEQ_FEA
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

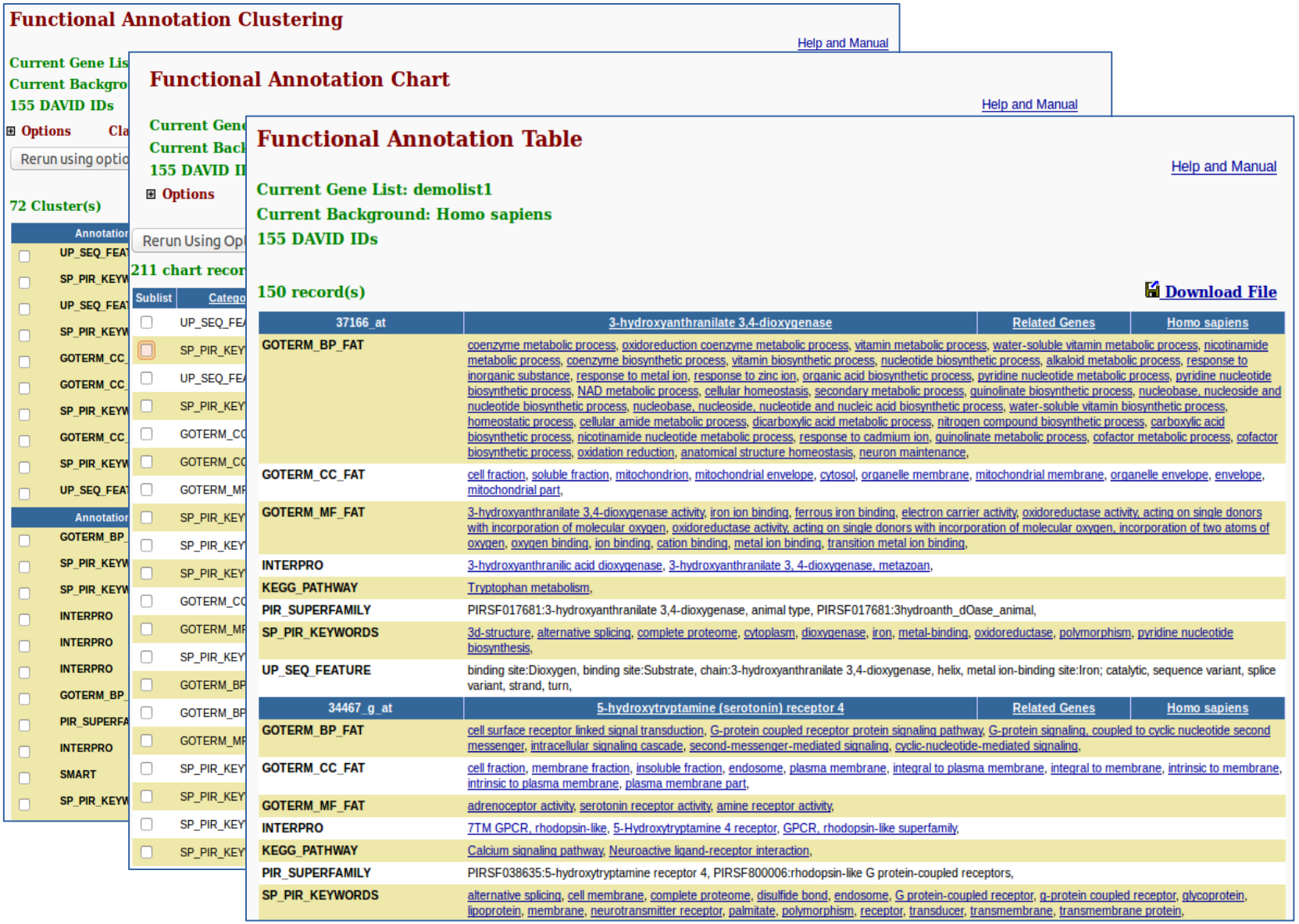
Rerun Using Options | Create Sublist

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 optic

72 Cluster(s)

Annotation
UP_SEQ_FEA
SP_PIR_KEYW
UP_SEQ_FEA
SP_PIR_KEYW
GOTERM_CC_
GOTERM_CC_
SP_PIR_KEYW
GOTERM_CC_
SP_PIR_KEYW
UP_SEQ_FEA
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_FE
	SP_PIR_KEY
	UP_SEQ_FE
	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

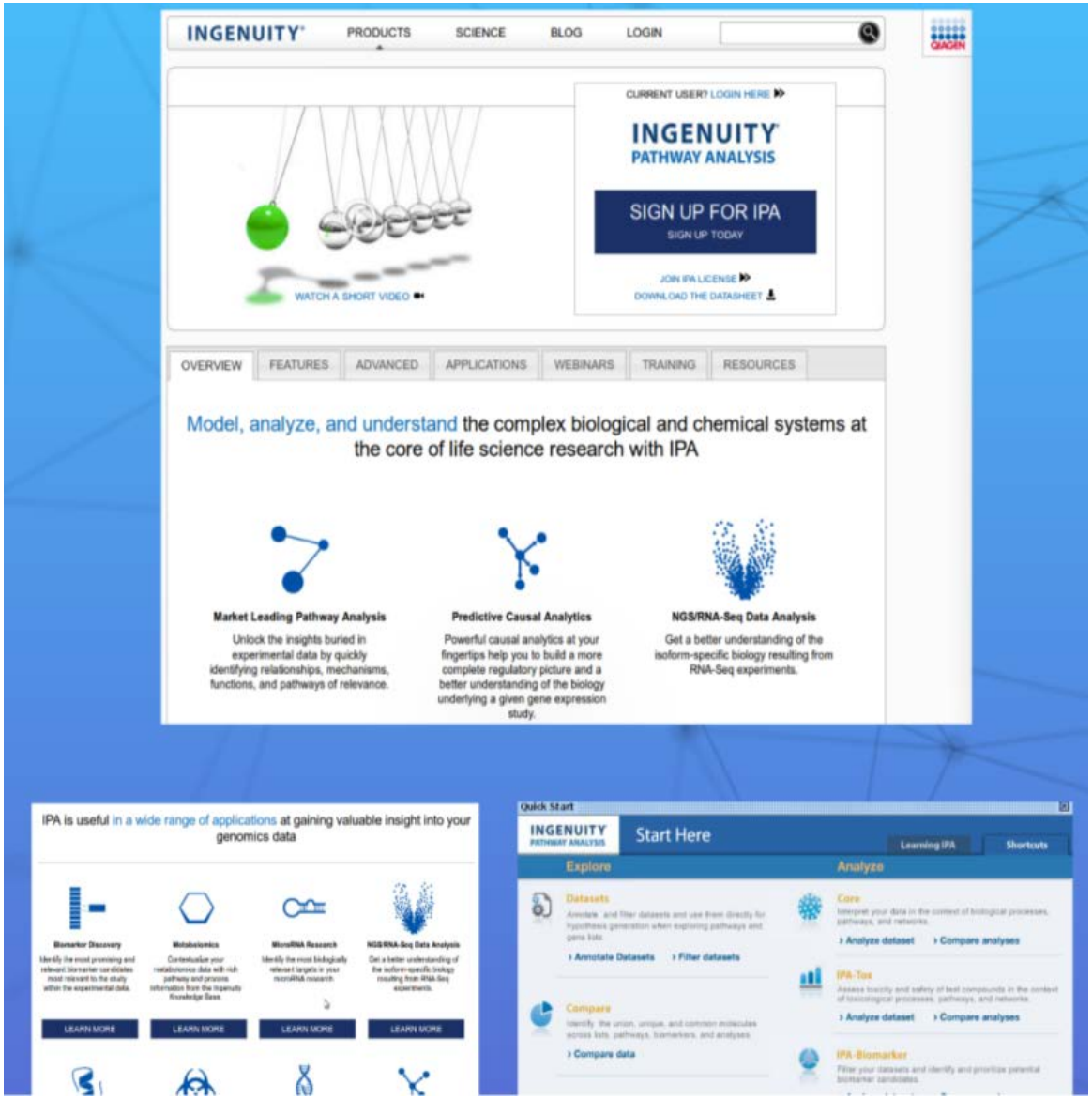
150 record(s)

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



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-1α-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

UNIVERSITAT DE BARCELONA

GRBIO

Vall d'Hebron
Institut de Recerca