An Introduction to Biological Significance Analysis

Alex Sánchez

Vall d'Hebron Institut de Recerca

Statistics and Bioinformatics Unit
Vall d'Hebron Institut de Recerca

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

UNIVERSITAT DE BARCELONA

Statistics and Bioinformatics Research Group
Statistics department, Universitat de Barelona

GRBIO

Outline

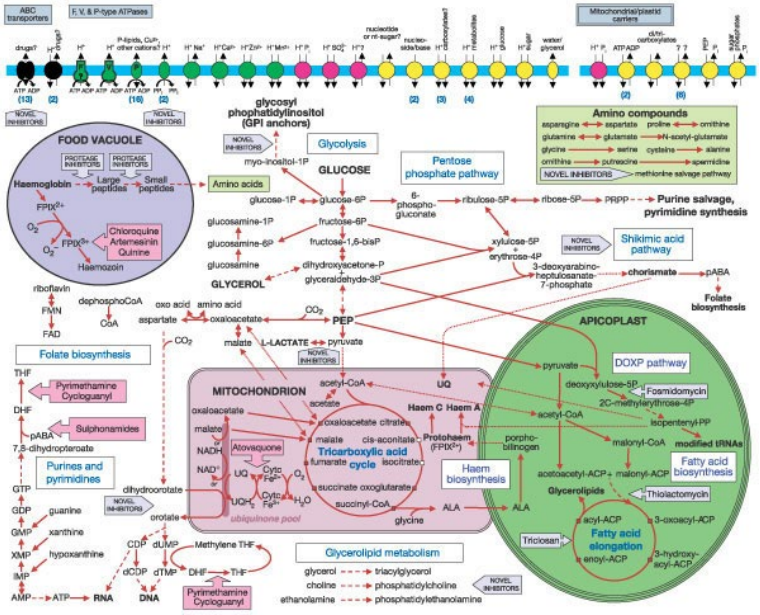
- Presentation
- Introduction and Background
  - Gene lists, Identifiers and Pathway databases
- Pathway Analysis: Methods and Tools
  - Overrepresentation analysis
  - GSEA: Gene Set Enrichment Analysis
  - Multiple Testing Adjustments
- Examples with R and Bioconductor

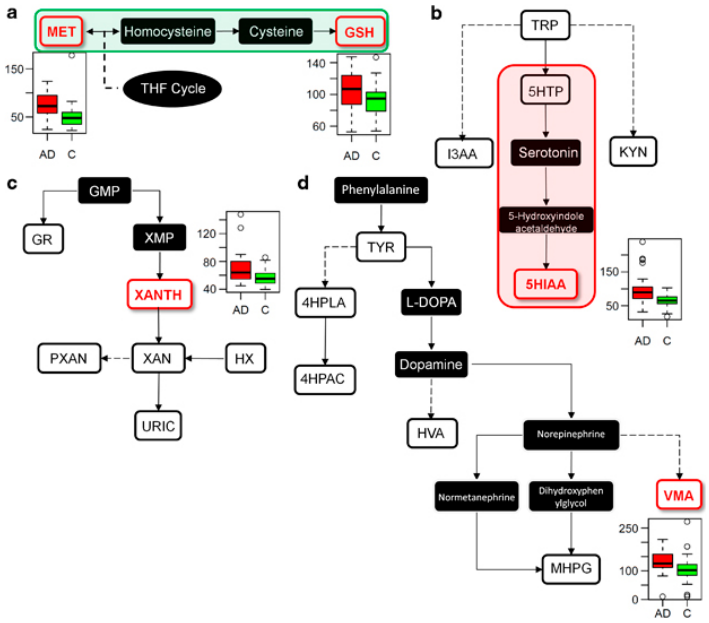
Introduction & Background

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

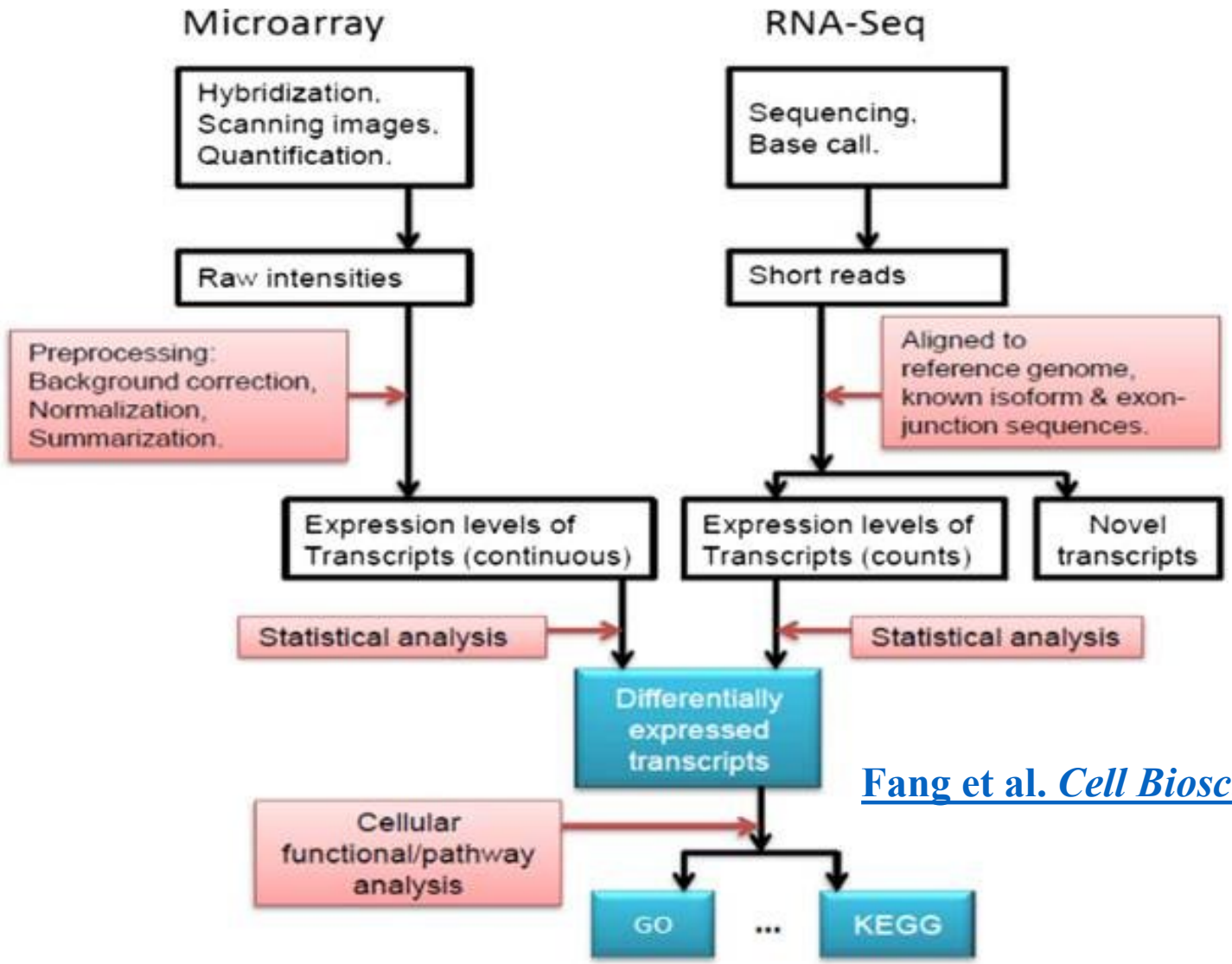
That’s what Pathway Analysis is about (more or less)

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

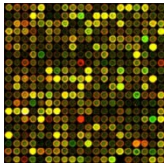
The beginning: *Gene Lists*

The life-cycle of an omics-based study



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

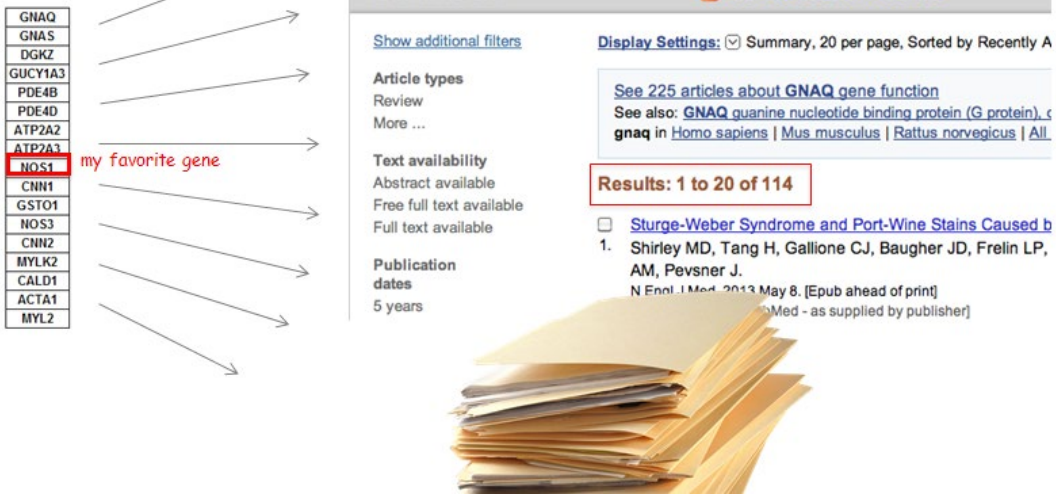
The (in)famous "*where to now?*" question



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

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

?

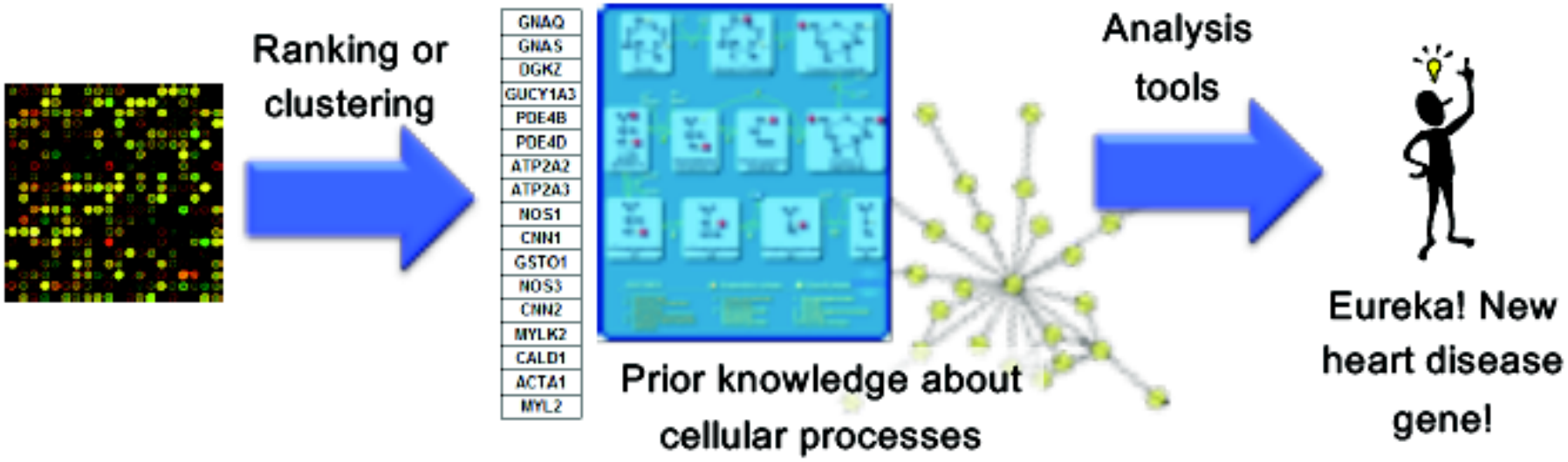


From gene lists to *Pathway Analysis*

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

Pathway Analysis Wishlist

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



Example 1

- 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

Example 2

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

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

Example 3

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

rank	GeneName	test statistic
1	IGDCC3	35.5553322839225
2	ANTXR1	35.3770766531836
3	AEBP1	33.0690543534961
4	FBN1	32.1199562790897
5	ANGPTL2	31.8605806216522
6	COL16A1	31.7641267462069
7	BGN	31.533826423921
…	…	…
15201	IRF1	-14.7629673442493
15202	CXCL10	-14.982733665643
15203	TAP2	-15.1488606179238
15204	UBE2L6	-15.7162058907796
15205	KIAA0319	-15.7796986548781
15206	PSMB8	-15.7846188665582
15207	PSME1	-16.4510045533584
15208	CSAG3	-16.8014265945244
15209	OVGP1	-17.6903158148446
15210	GBP4	-17.9447602030134
15211	TAP1	-18.0549262210415
15212	PSME2	-18.3639448844986
15213	PSMB9	-18.6614452029879

Gene Lists and Annotations

Gene and Protein Identifiers

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

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

Common Identifiers

Gene
Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript
GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein
Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

Species-specific
HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD 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 favourite gene name
  - Link to related resources
  - Identifier translation
    - E.g. Proteins to genes, Affy ID to Entrez Gene
  - Merging data from different sources
    - Find equivalent records

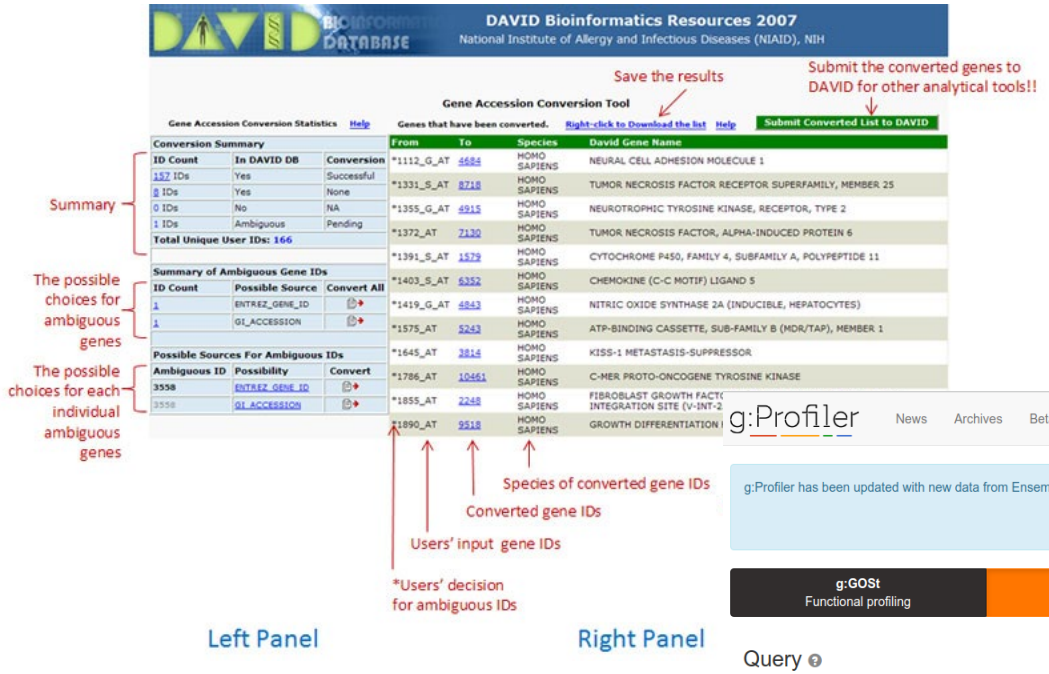
ID Challenges

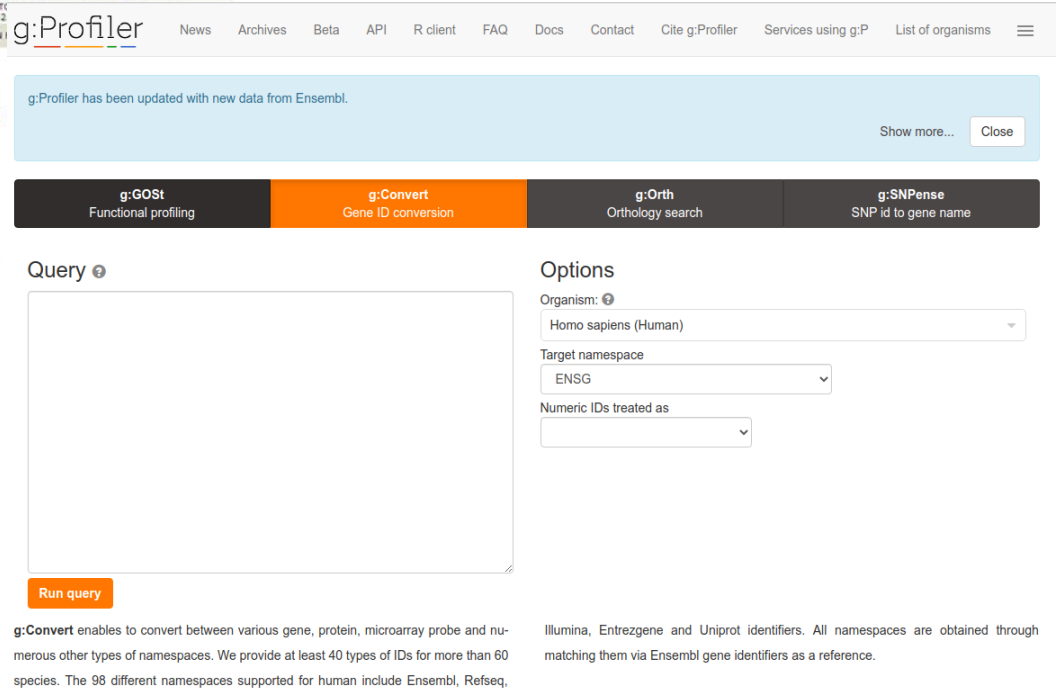
- Avoid errors: map IDs correctly
  - Beware of 1-to-many mappings
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol: TP53
- Excel error-introduction
  - OCT4 is changed to October-4 (paste as text)
- Problems reaching 100% coverage
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

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

Use ID converters to prepare list

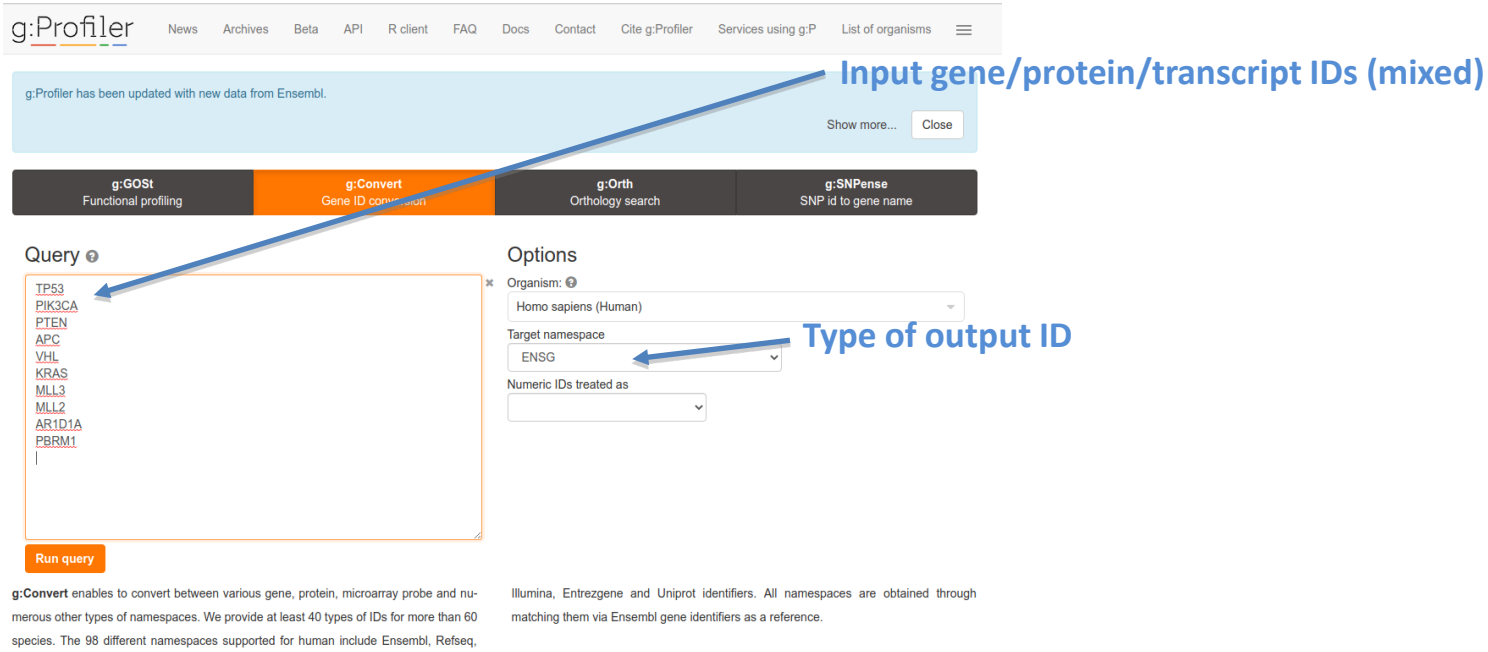
Option 1: DAVID



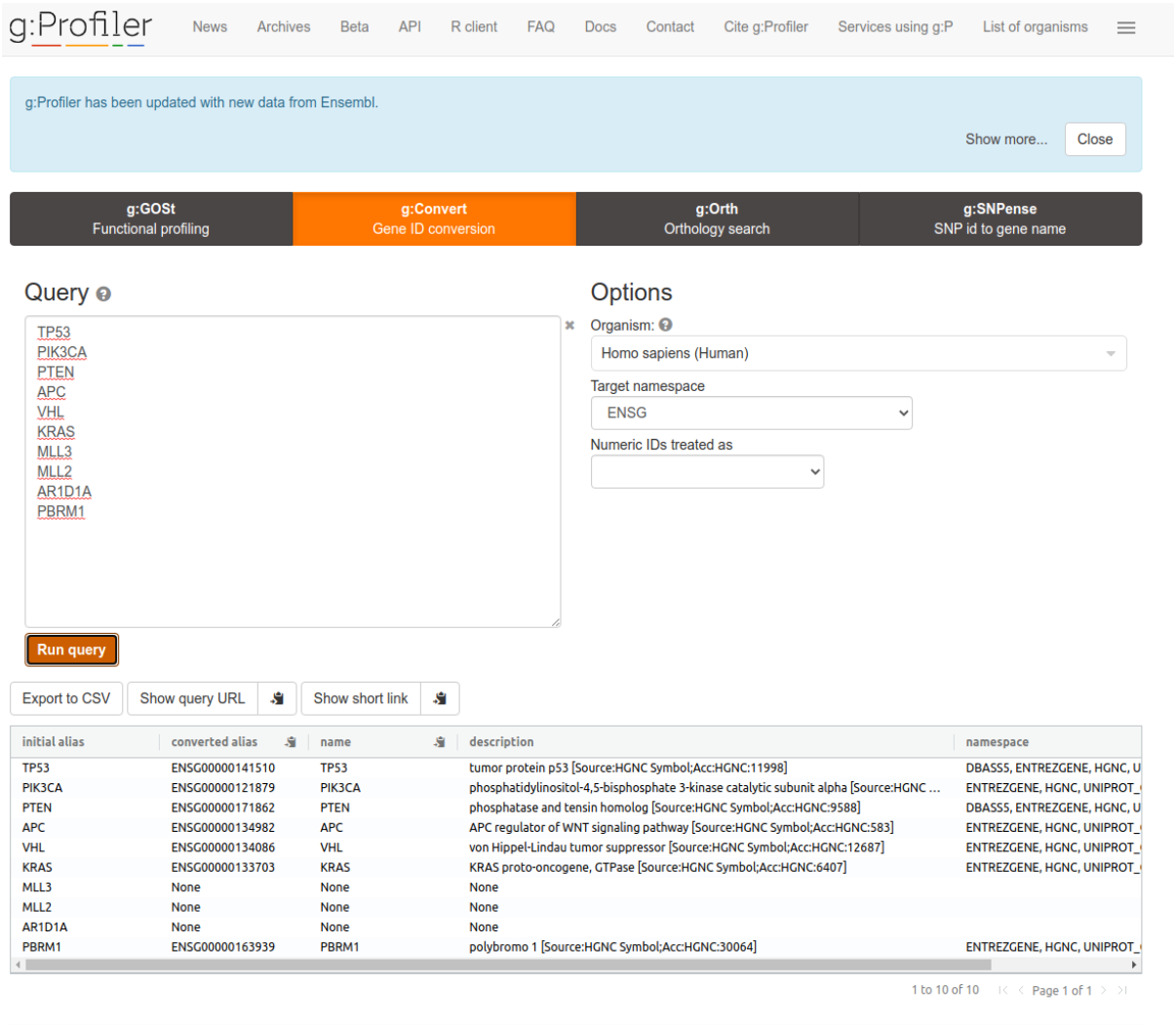


Example 1: Gene ID con version with g:Converter

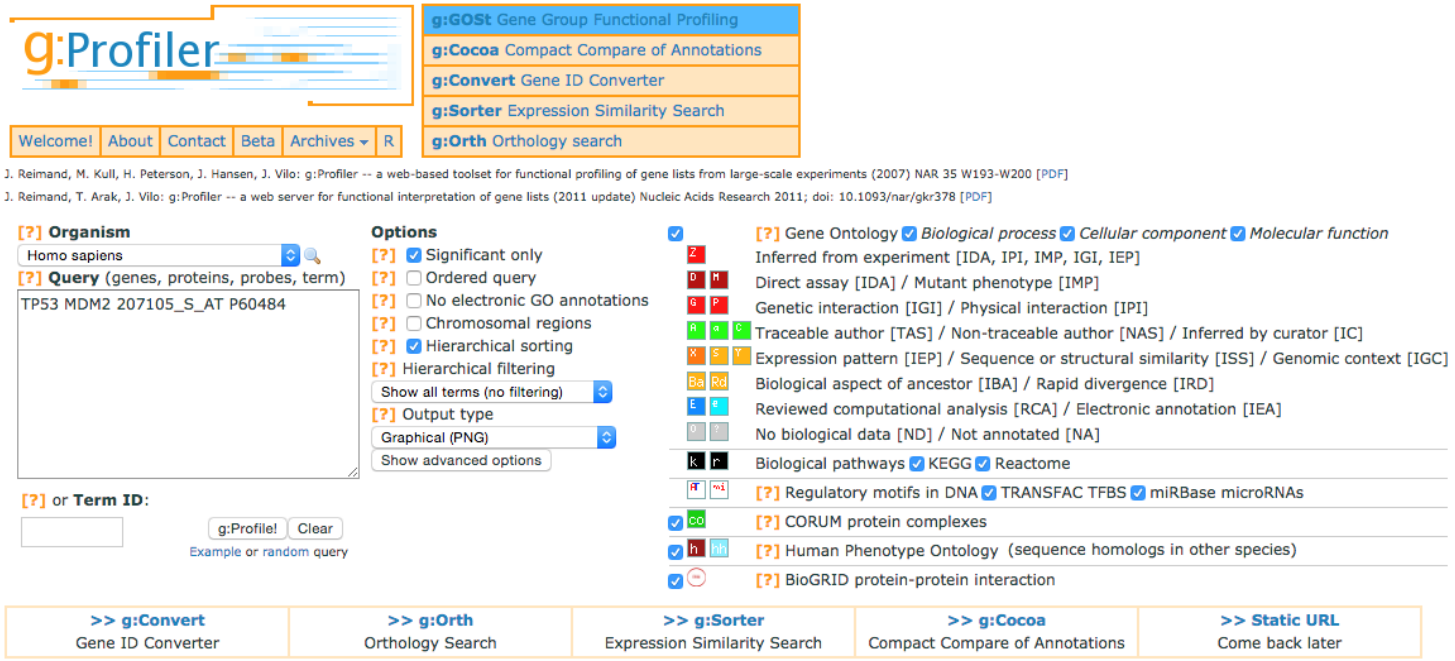
Example: Gene ID conversion with *g:Convert*



Example: Gene ID conversion with *g:Convert*



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 gProfiler, DAVID or Biomart.
- If 100% coverage desired, manually curate missing mappings using multiple resources
- Be careful of Excel auto conversions – especially when pasting large gene lists!
  - Remember to format cells as 'text' before pasting

Pathway and *Gene Sets* Databases

Where is pathway information? (1)

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

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

Where is pathway information? (2)

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

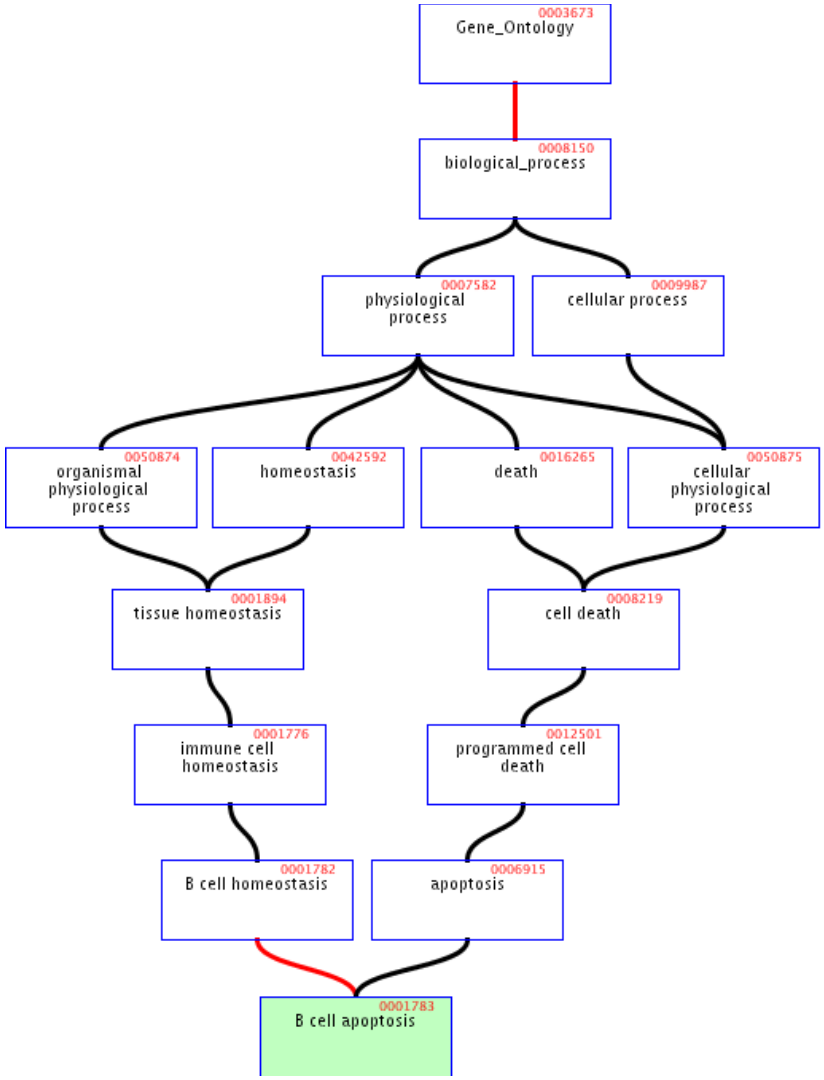
*Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges

What is the Gene Ontology (GO)?

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

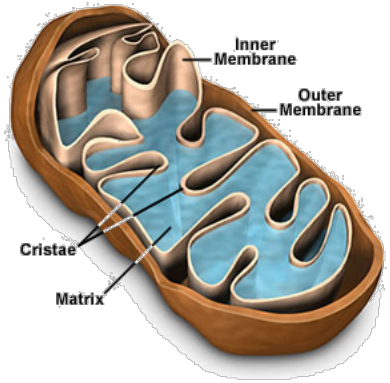
GO Structure

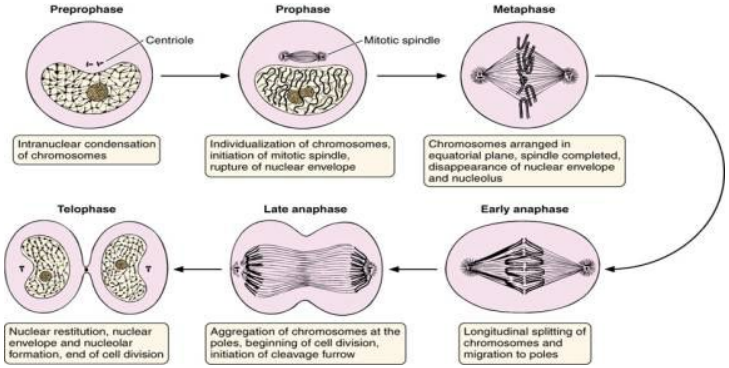
- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



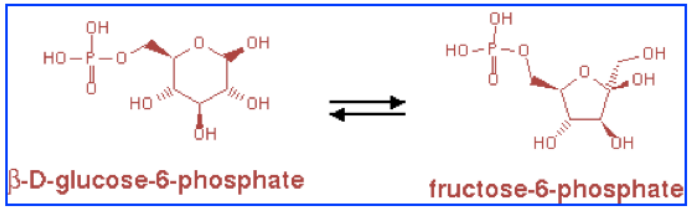
What is covered by the GO?

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





Cell division



glucose-6-phosphate isomerase activity

Annotation Sources

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

Evidence Types

- Experimental Evidence Codes
  - EXP: Inferred from Experiment
  - IDA: Inferred from Direct Assay
  - IPI: Inferred from Physical Interaction
  - IMP: Inferred from Mutant Phenotype
  - IGI: Inferred from Genetic Interaction
  - IEP: Inferred from Expression Pattern

- Computational Analysis Evidence Codes
  - ISS: Inferred from Sequence or Structural Similarity
  - ISO: Inferred from Sequence Orthology
  - ISA: Inferred from Sequence Alignment
  - ISM: Inferred from Sequence Model
  - IGC: Inferred from Genomic Context
  - RCA: inferred from Reviewed Computational Analysis

- Author Statement Evidence Codes
  - TAS: Traceable Author Statement
  - NAS: Non-traceable Author Statement
- Curator Statement Evidence Codes
  - IC: Inferred by Curator
  - ND: No biological Data available

- IEA: Inferred from electronic annotation

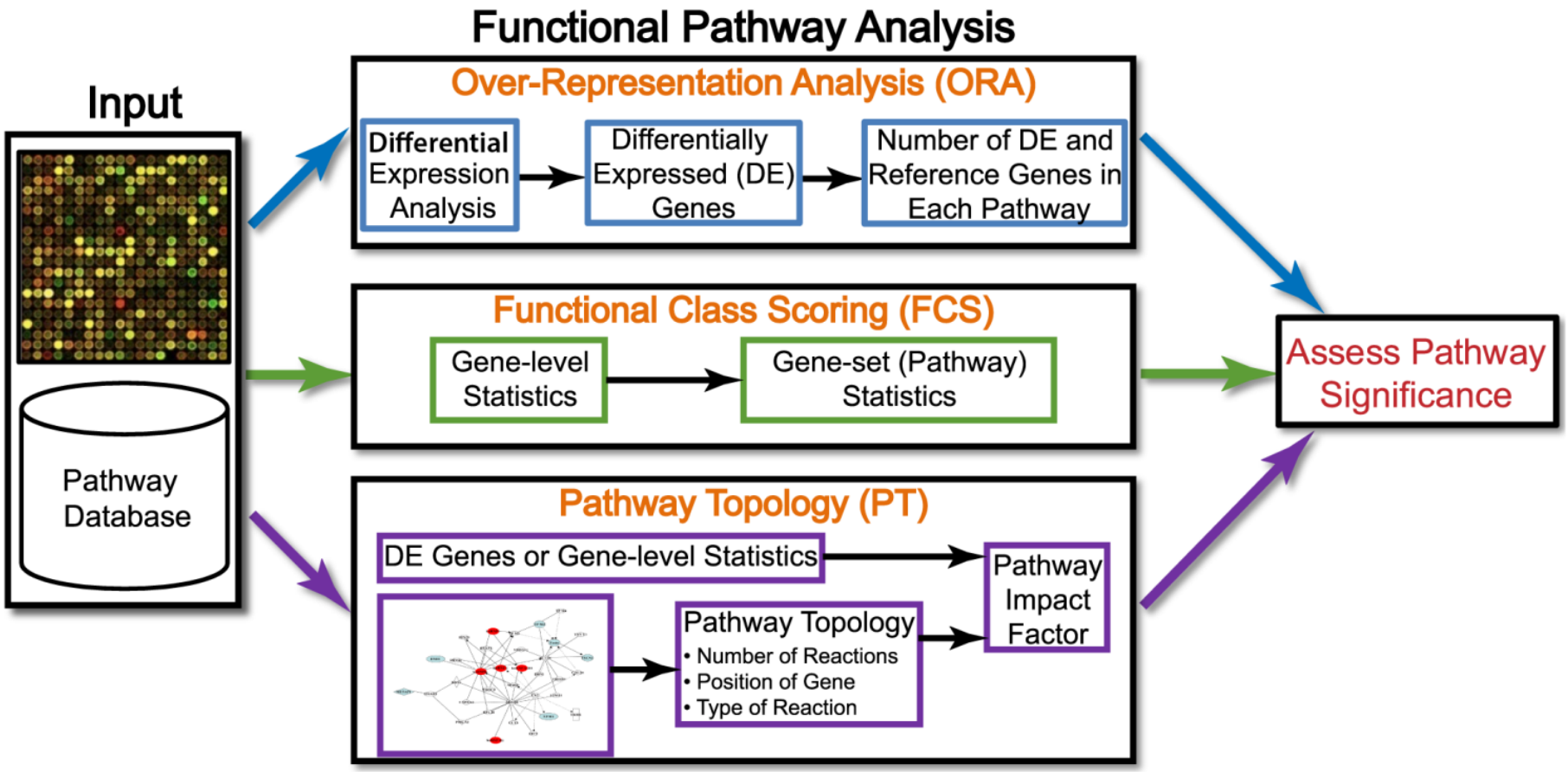
http://www.geneontology.org/GO.evidence.shtml

Pathway Analysis Methods

Overrepresentation Analysis

Gene Set Enrichment Analysis

Types of Pathway Analysis



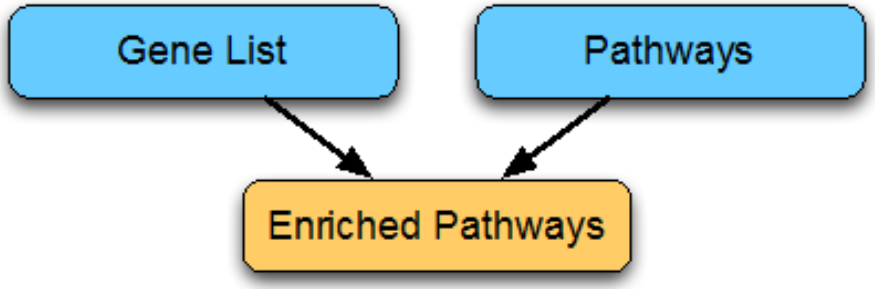
Khatri et alt. *10 years of Pathway Analysis*

Analysis of *thresholded* lists with *Enrichment Analysis*

(also called Overrepresentation A.)

Over-representation analysis

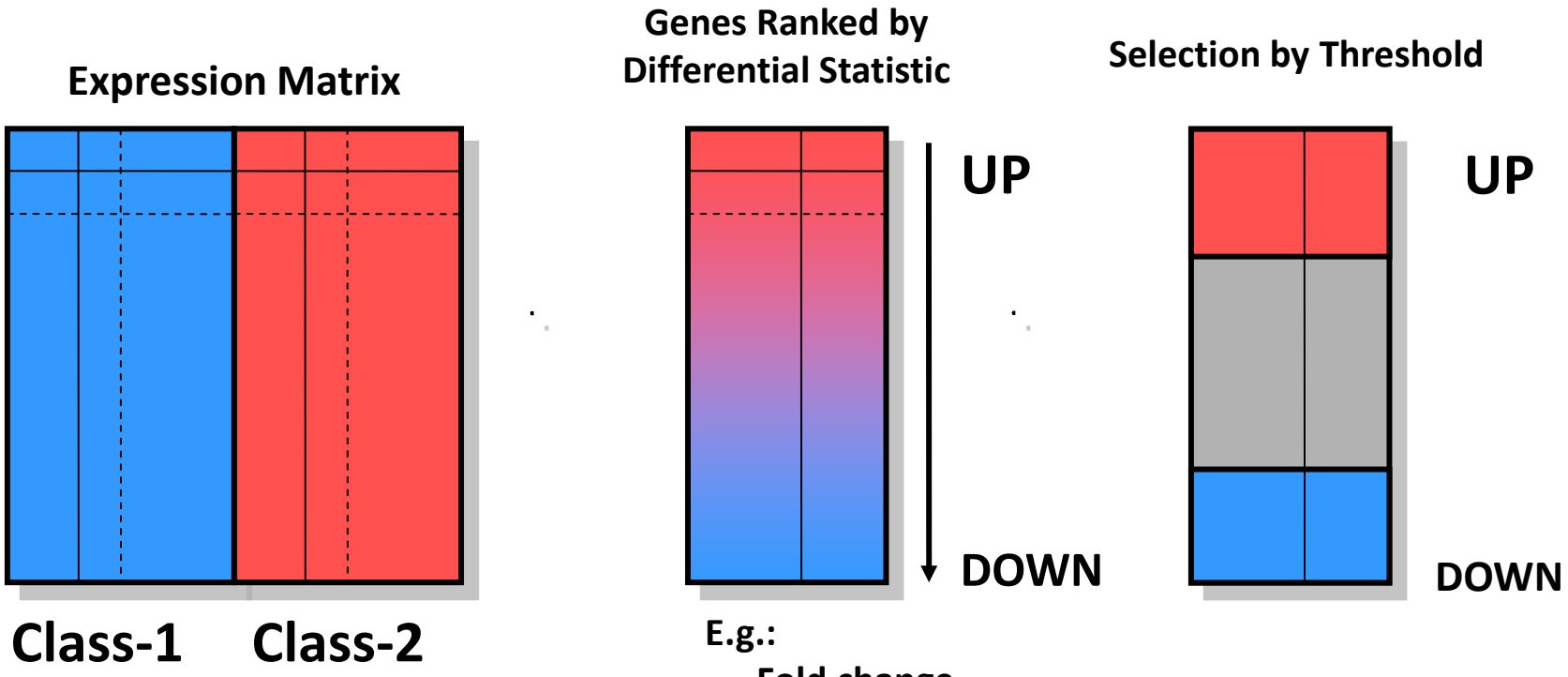
- Combines
  - Gene (feature) lists ← (Gen)omic experiment
  - Pathways and other gene annotations
    - Gene Ontology
    - Reactome
    - Pathway commons



Over-representation analysis

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

Obtaining the gene lists

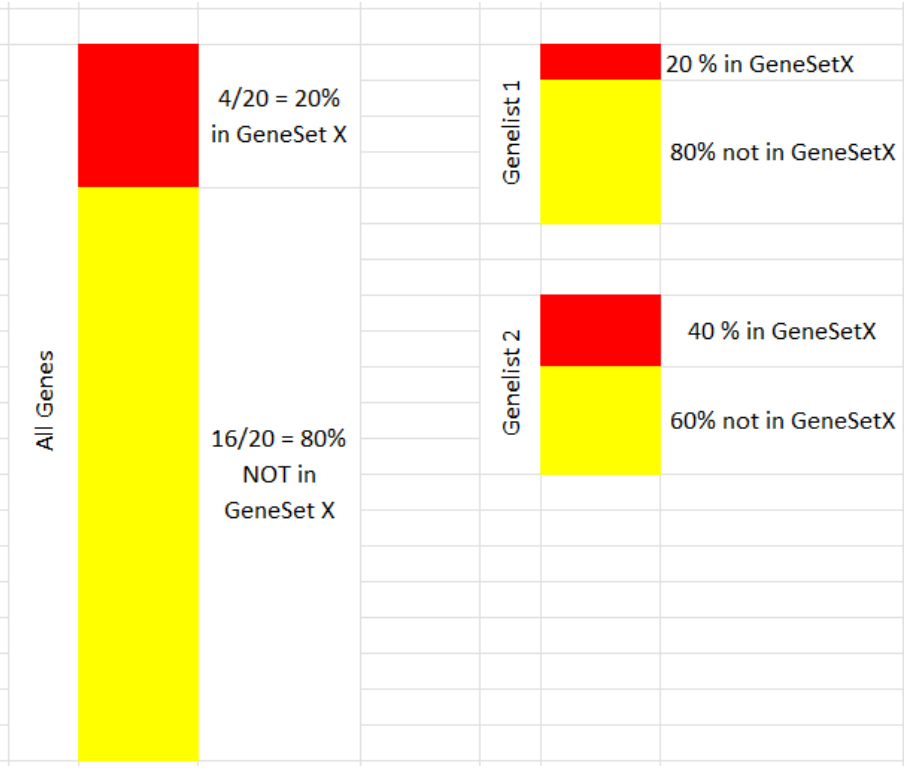


E.g.:

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

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 : "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=0.02**	

0.025 ≈ 0.02 : 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 tests

- 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

An example of ORA using Bioconductor (1)

```
# Get Genelist and expression marix
topTabAvsB <- read.table ("datasets/Top_AvsB.csv2",  head=T, sep=";", dec=",", row.names=1)
expresAvsB <- read.table ("datasets/expres_AvsB.csv2", head=T, sep=";", dec=",", row.names=1)

# Define Gene list using arbitrary though reasonable cutoffs
probesUniverse <- rownames(topTabAvsB)
whichGenesInTop<- topTab["adj.P.Val"]<0.05 & topTab["logFC"] > 1
# Annotate Gene Universe and Gene list
entrezUniverse<- select(hgu133a.db, probesUniverse, "ENTREZID")
entrezUniverse <- entrezUniverse$ENTREZID
topGenes <-   entrezUniverse[whichGenesInTop]
entrezUniverse <- entrezUniverse[!duplicated(entrezUniverse)]
topGenes <- topGenes[!duplicated(topGenes)]
```

An example of ORA using Bioconductor (2)

```
library(GOstats)
GOparams = new("GOHyperGParams",
        geneIds=topGenes, universeGeneIds=entrezUniverse,
        annotation="hgu133a.db", ontology="BP",  pvalueCutoff=0.001)
GOhyper = hyperGTest(GOparams)
head(summary(GOhyper))
```

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0019370	0.0000917494	31.294444	0.2904762	4	7	leukotriene biosynthetic process
2	GO:0046395	0.0004796855	3.282433	4.4816327	13	108	carboxylic acid catabolic process
3	GO:0016054	0.0005736863	3.213594	4.5646259	13	110	organic acid catabolic process
4	GO:0072329	0.0007287196	4.193992	2.4897959	9	60	monocarboxylic acid catabolic process
5	GO:0006691	0.0007580159	13.402381	0.4564626	4	11	leukotriene metabolic process
6	GO:0045109	0.0008510094	8.401076	0.7884354	5	19	intermediate filament organization

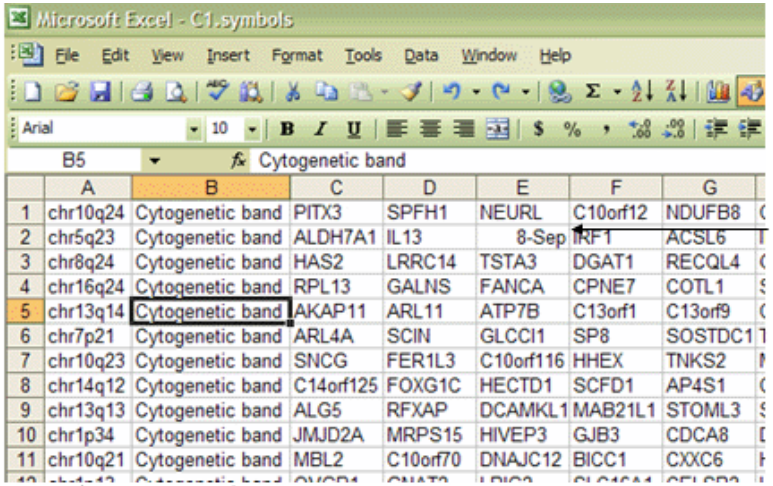
Analysis of ranked gene lists with *Gene Set Enrichment Analysis* (also called Functional Class Scoring)

Gene Sets

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

Gene Sets

Each row represents one gene set



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

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

The GSEA method

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

Calculating enrichment score (ES)

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

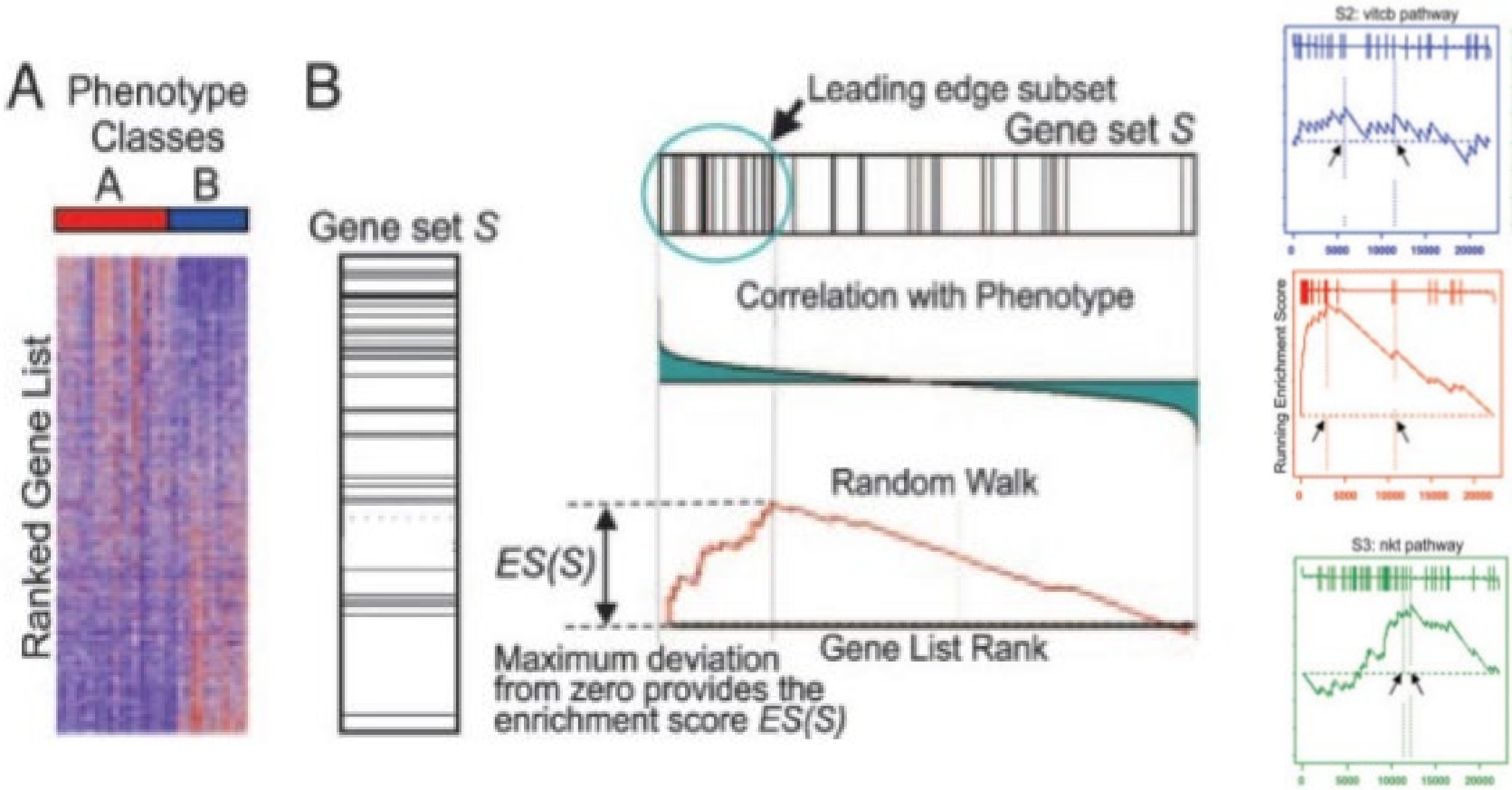
$$X_i = -\sqrt{\frac{N_S}{N - N_S}}$$

If gene p is in set S, then add

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

This creates a running sum
The maximum sum over the whole list L is the Enrichment Score MES

The GSEA method



Recipe for ranked list enrichment test

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

GSEA variants

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

Multiple test adjustments

Why we need to “adjust”

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

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

	Actual Situation "Truth"	
Decision	**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

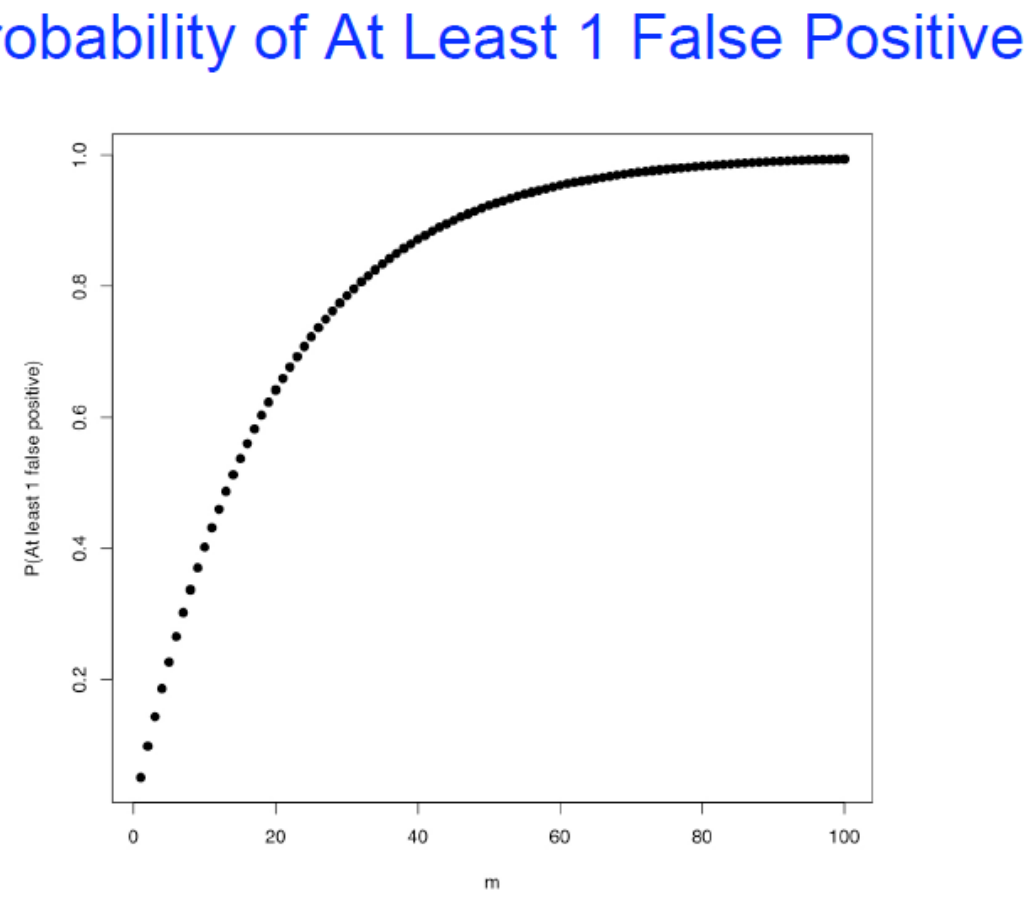
- Omics studies are “high throughput”
  - Selecting genes: One test per each gene
  - Finding enriched gene sets: One test per each gene set
- Doing many tests means facing repeatedly the probability of making one false positive.
  - As the number of tests increases →
  - The chance of observing at least 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 $< \alpha$
  2. The expected number of false positives is below a fixed threshold.

  …

Controlling the FWER: *Bonferroni*

If *M* = # of annotations tested:

Adjusted P-value = *M* x original P-value

Adjusted P-value is greater than or equal to the probability that ***one or more of the observed enrichments*** are 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 adjustments 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: *AdjP-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

Adjusted P-value is "nominal" P-value times # of tests divided by the rank of the P-value in sorted list: *AdjP-value = P-value X [# of tests] / Rank*

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

Adjusted P-value is "nominal" P-value times # of tests divided by the rank of the P-value in sorted list: *AdjP-value = P-value X [# of tests] / Rank*

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.

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

Reducing adjustment stringency

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

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.