Gene Ontology 2 – Function/Pathway Enrichment

Biol4559 Thurs, April 12, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

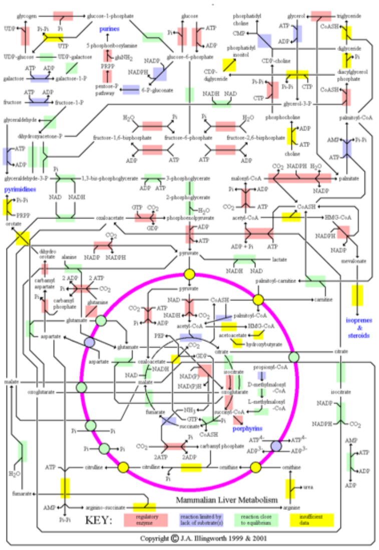
- Function/Pathway enrichment analysis
  - do sets (subsets) of differentially expressed genes reflect a pathway?
- Over Representation Analyis (ORA)
  - Fisher exact test, hypergeometric
  - competitive vs self-contained tests
- Functional Class Scoring (FTS)
  - GSEA : Gene Set Enrichment Analysis
- Pathway Topology (PT)
  - SPIA : Signaling Pathway Impact Analysis
- What are the right "controls"?

To learn more:

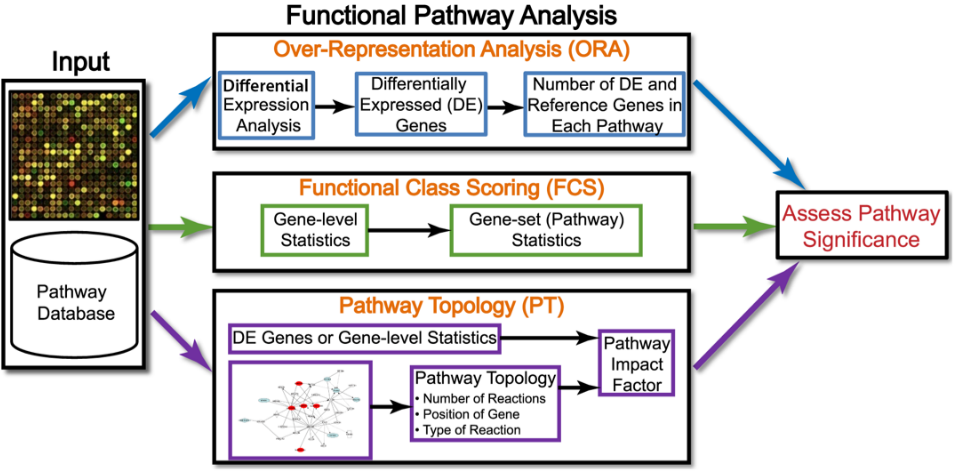
1. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* **8,** e1002375 (2012).
2. Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nat Rev Genet* **9,** 509–515 (2008).
3. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102,** 15545–15550 (2005).

What is happening in the cell?

- Cellular functions are chemical
- Fundamental biochemical processes are lined chemical reactions: pathways
  - cell division: DNA replication, mitosis, segregation
  - metabolism: energy, amino-acids, detoxification
  - response to stimuli: signaling
- Some pathways are better understood than others


Mammalian Liver Metabolism
KEY:
regulatory enzyme
reaction limited by lack of substrate(s)
reaction close to equilibrium
insufficient data
Copyright © J.A. Illingworth 1999 & 2001

from genes to pathways: enrichment analysis



Functional Pathway Analysis

Input

Pathway Database

Over-Representation Analysis (ORA)

Differential Expression Analysis → Differentially Expressed (DE) Genes → Number of DE and Reference Genes in Each Pathway

Functional Class Scoring (FCS)

Gene-level Statistics → Gene-set (Pathway) Statistics

Pathway Topology (PT)

DE Genes or Gene-level Statistics → Pathway Impact Factor

Pathway Topology
- Number of Reactions
- Position of Gene
- Type of Reaction

Assess Pathway Significance

Khatri, *et al. PLoS Comput Biol* **8,** e1002375 (2012).

Enrichment analysis

- Given a set of differentially expressed (up/down) genes
- And a set of Gene Ontology or Pathway relationships
- Can we use the differentially expressed genes to identify the biological process/pathway involved

GO/KEGG/PFAM enrichment

- are my 100's of candidates involved in similar process/pathways/functions?
- hypergeometric test for independence:

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

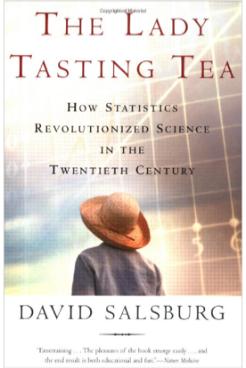
	significant	insignificant	total
in group:	k	m-k	m
not in group:	n-k	N+k-n-m	N-m
total:	n	N-n	N

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}$$

What should 'N' be?

- Total number of genes?
- Number of genes expressed?
- Number of genes up? down?

The significance of differences: Fisher's Exact Test

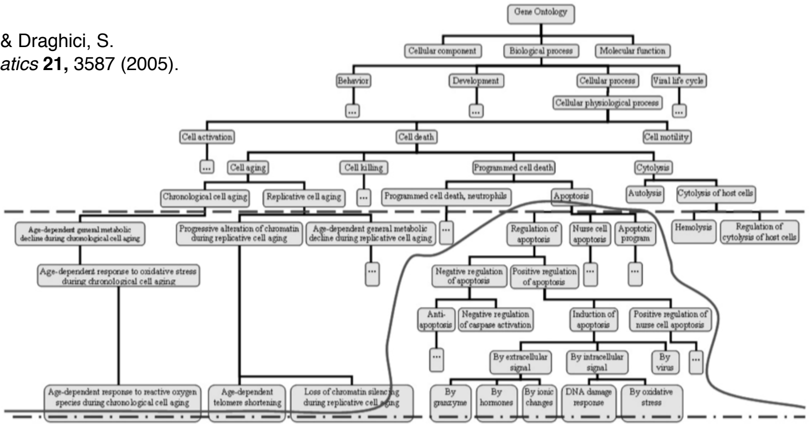

THE LADY TASTING TEA
HOW STATISTICS REVOLUTIONIZED SCIENCE IN THE TWENTIETH CENTURY
DAVID SALSBURG

1. Around 1930, Muriel Bristol claimed, in a conversation with R. A. Fisher, that she could tell when milk was poured into tea, which was much preferable to tea being poured into milk.
2. Fisher choose to test this hypothesis by preparing 8 cups of tea, 4 tea first, 4 milk first, and asking Ms. Bristol to identify the 4 cups with tea first.
3. If she has no ability to identify milk first/tea first, then one expects her to be right 50% of the time (2 cups). But what if she was right for 3 of the 4 cups?

```
> fisher.test(matrix(c(4,0,0,4),nrow=2),
+             alternative='greater')
         Fisher's Exact Test for Count Data
data:  matrix(c(4, 0, 0, 4), nrow = 2)
p-value = 0.01427
alternative hypothesis: true odds ratio is not equal to 1
```

Enrichment: In group / Not in group

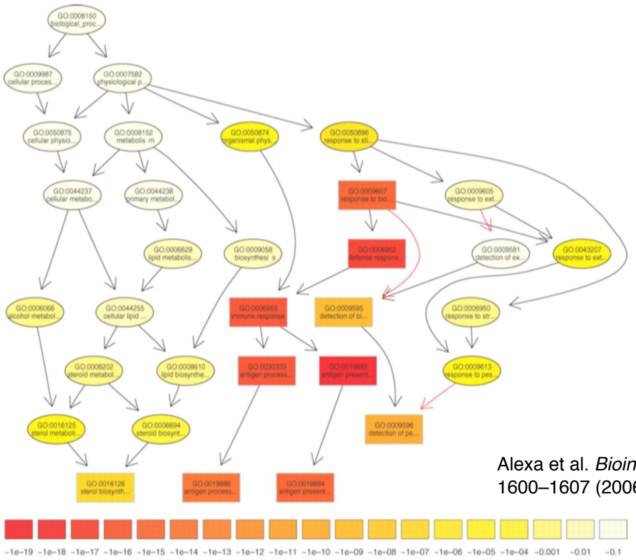
Khatri, P. & Draghici, S.
Bioinformatics **21,** 3587 (2005).


Gene Ontology
Cellular component
Biological process
Molecular function
Behavior
Development
Cellular process
Viral life cycle
Cellular physiological process
...
Cell activation
Cell death
Cell motility
Cell aging
Cell killing
Programmed cell death
Cytolysis
Chronological cell aging
Replicative cell aging
Programmed cell death, neutrophils
Apoptosis
Autolysis
Cytolysis of host cells
Age-dependent general metabolic decline during chronological cell aging
Progressive alteration of chromatin during replicative cell aging
Age-dependent general metabolic decline during replicative cell aging
Regulation of apoptosis
Nurse cell apoptosis
Apoptotic program
Hemolysis
Regulation of cytolysis of host cells
Age-dependent response to oxidative stress during chronological cell aging
Negative regulation of apoptosis
Positive regulation of apoptosis
Anti-apoptosis
Negative regulation of caspase activation
Induction of apoptosis
Positive regulation of nurse cell apoptosis
By extracellular signal
By intracellular signal
By virus
Age-dependent response to reactive oxygen species during chronological cell aging
Age-dependent telomere shortening
Loss of chromatin silencing during replicative cell aging
By granzyme
By hormones
By ionic changes
DNA damage response
By oxidative stress

	significant	insignificant	total
in group:	k	m-k	m
not in group:	n-k	N+k-n-m	N-m
total:	n	N-n	N

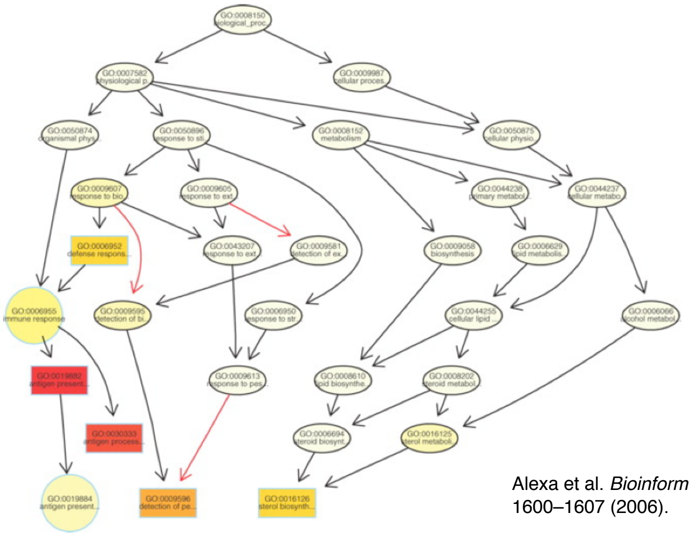
What should 'N' be?
- Total number of genes?
- Number of genes expressed?
- Number of genes up? down?

Many levels of GO annotation:

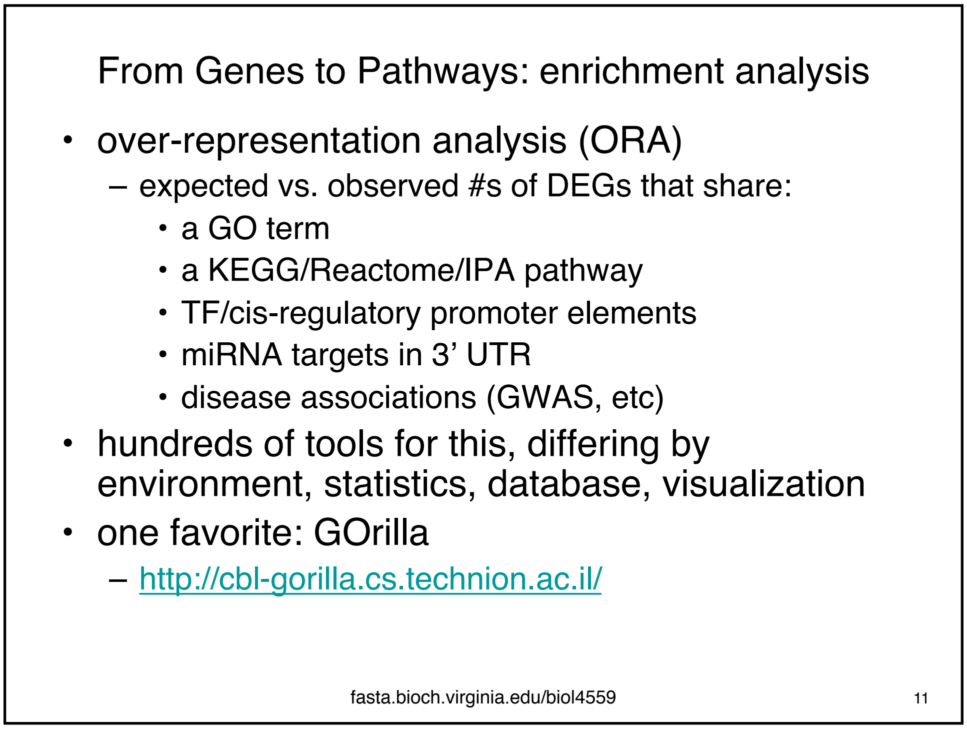

GO:0008150 biological_proc...
GO:0009987 cellular proces...
GO:0007582 physiological p...
GO:0050875 cellular physio...
GO:0008152 metabolis m
GO:0050874 organismal phys...
GO:0050896 response to sti...
GO:0044237 cellular metabo...
GO:0044238 primary metabol...
GO:0009607 response to bio...
GO:0009605 response to ext...
GO:0006629 lipid metabolis...
GO:0009058 biosynthesi s
GO:0006952 defense respons...
GO:0009581 detection of ex...
GO:0043207 response to ext...
GO:0006066 alcohol metabol...
GO:0044255 cellular lipid ...
GO:0006955 immune response
GO:0009595 detection of bi...
GO:0006950 response to str...
GO:0008202 steroid metabol...
GO:0008610 lipid biosynthe...
GO:0030333 antigen process...
GO:0019882 antigen present...
GO:0009613 response to pes...
GO:0016125 sterol metaboli...
GO:0006694 steroid biosynt...
GO:0009596 detection of pe...
GO:0016126 sterol biosynth...
GO:0019886 antigen process...
GO:0019884 antigen present...
~1e−19
~1e−18
~1e−17
~1e−16
~1e−15
~1e−14
~1e−13
~1e−12
~1e−11
~1e−10
~1e−09
~1e−08
~1e−07
~1e−06
~1e−05
~1e−04
~0.001
~0.01
~0.1

Alexa et al. *Bioinformatics* **22,** 1600–1607 (2006).

Correcting for multiple inheritance

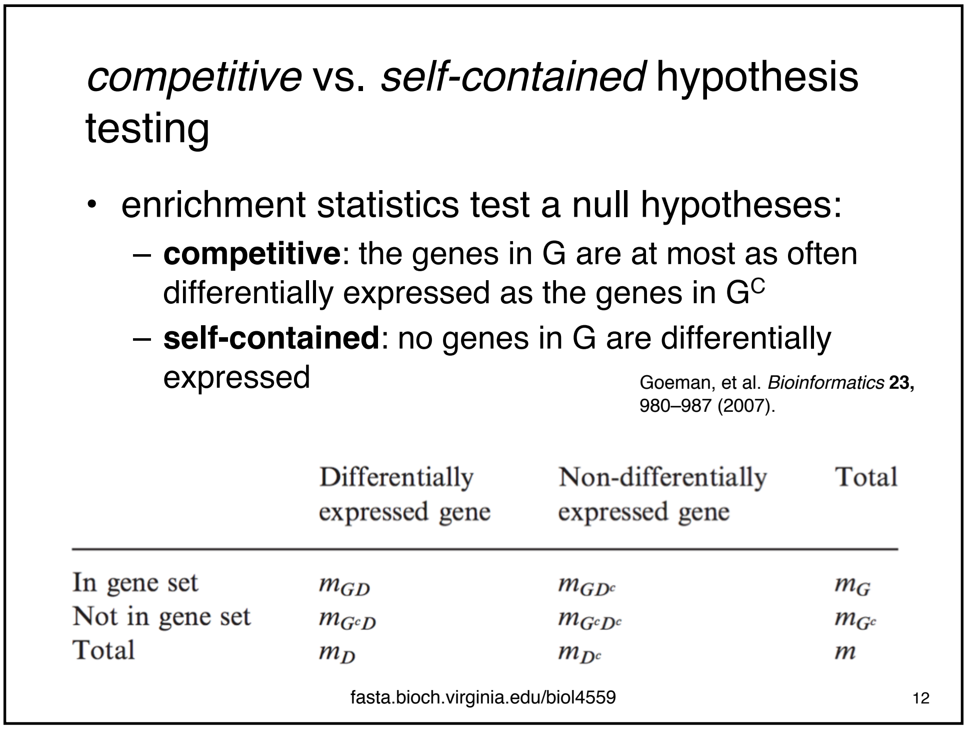

GO:0008150 iological_proc.
GO:0007582 hysiological p.
GO:0009987 cellular proces.
GO:0050874 organismal phys..
GO:0050896 esponse to sti.
GO:0008152 metabolism
GO:0050875 cellular physio.
GO:0009607 esponse to bio.
GO:0009605 esponse to ext.
GO:0044238 primary metabol.
GO:0044237 cellular metabo.
GO:0006952 defense respons...
GO:0043207 esponse to ext.
GO:0009581 detection of ex.
GO:0009058 biosynthesis
GO:0006629 lipid metabolis.
GO:0006955 immune response
GO:0009595 detection of bi.
GO:0006950 esponse to str.
GO:0044255 cellular lipid ...
GO:0006066 alcohol metabol.
GO:0019882 antigen present.
GO:0009613 esponse to pes.
GO:0008610 lipid biosynthe.
GO:0008202 steroid metabol.
GO:0030333 antigen process.
GO:0006694 steroid biosynt.
GO:0016125 sterol metaboli.
GO:0019884 antigen present..
GO:0009596 detection of pe...
GO:0016126 sterol biosynth...

Alexa et al. *Bioinformatics* **22,** 1600–1607 (2006).



From Genes to Pathways: enrichment analysis

- over-representation analysis (ORA)
  - expected vs. observed #s of DEGs that share:
    - a GO term
    - a KEGG/Reactome/IPA pathway
    - TF/cis-regulatory promoter elements
    - miRNA targets in 3’ UTR
    - disease associations (GWAS, etc)
- hundreds of tools for this, differing by environment, statistics, database, visualization
- one favorite: GOrilla
  - http://cbl-gorilla.cs.technion.ac.il/



competitive vs. *self-contained* hypothesis testing

- enrichment statistics test a null hypotheses:
  - **competitive**: the genes in G are at most as often differentially expressed as the genes in G^C
  - **self-contained**: no genes in G are differentially expressed

Goeman, et al. *Bioinformatics* **23,** 980–987 (2007).

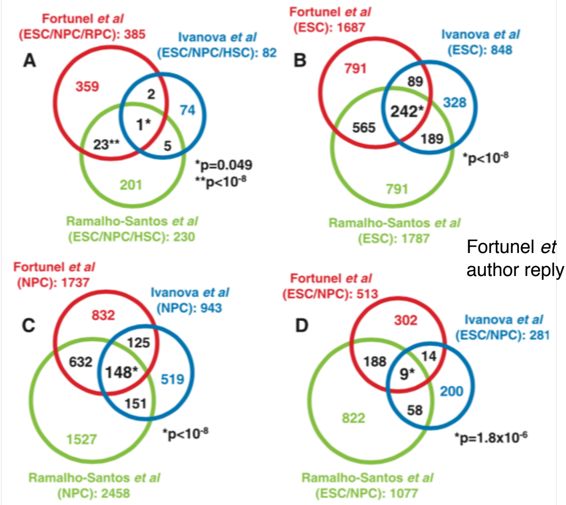
	Differentially expressed gene	Non-differentially expressed gene	Total
In gene set	m_{GD}	m_{GD^c}	m_G
Not in gene set	m_{G^cD}	$m_{G^cD^c}$	m_{G^c}
Total	m_D	m_{D^c}	m

competitive vs. *self-contained* hypothesis testing

- **competitive**: the genes in G are at most as often differentially expressed as the genes in G^C
  - testing for *excess* of differential expression across genes in G, relative to genes not in G
  - depends strongly on G^C distribution/universe
- **self-contained**: no genes in G are differentially expressed
  - testing for *presence* of any differential expression somewhere within G, across all genes in G
  - stronger, more powerful testing (more false positives)

Goeman, et al. *Bioinformatics* **23,** 980–987 (2007).

Over Representation Analysis - Reproducibility


A
Fortunel et al (ESC/NPC/RPC): 385
Ivanova et al (ESC/NPC/HSC): 82
Ramalho-Santos et al (ESC/NPC/HSC): 230
359
2
74
1*
23**
5
201
*p=0.049
**p<10-8
B
Fortunel et al (ESC): 1687
Ivanova et al (ESC): 848
Ramalho-Santos et al (ESC): 1787
791
89
328
242*
565
189
791
*p<10-8
C
Fortunel et al (NPC): 1737
Ivanova et al (NPC): 943
Ramalho-Santos et al (NPC): 2458
832
125
632
148*
519
151
1527
*p<10-8
D
Fortunel et al (ESC/NPC): 513
Ivanova et al (ESC/NPC): 281
Ramalho-Santos et al (ESC/NPC): 1077
302
14
188
9*
200
58
822
*p=1.8x10-6

Fortunel *et* al *Science* **302,** 393–author reply 393 (2003).

(**A**) “Stemness” genes. (**B**) ESC-enriched genes (**C**) NPC-enriched genes. (**D**) Overlap of “stemness” genes—two types of stem cell (ESC/NPC)-enriched genes

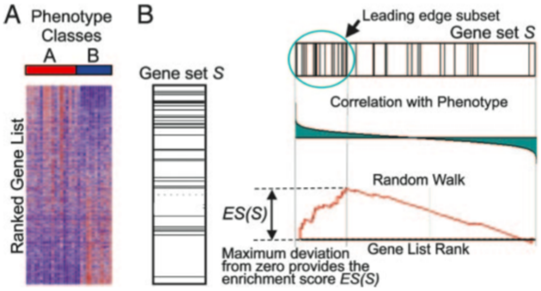
Issues with ORA

1. arbitrary significance thresholds for inclusion
2. Differential Expression magnitude/directionality not considered
3. sensitive to choice of background "universe"
   - all genes, genes on chip, or genes with sufficient signal that could possibly be called DEG?
4. correlation between genes ignored
5. correlation/cross-talk between pathways

Functional Class Scoring (FCS) methods fix #1-3

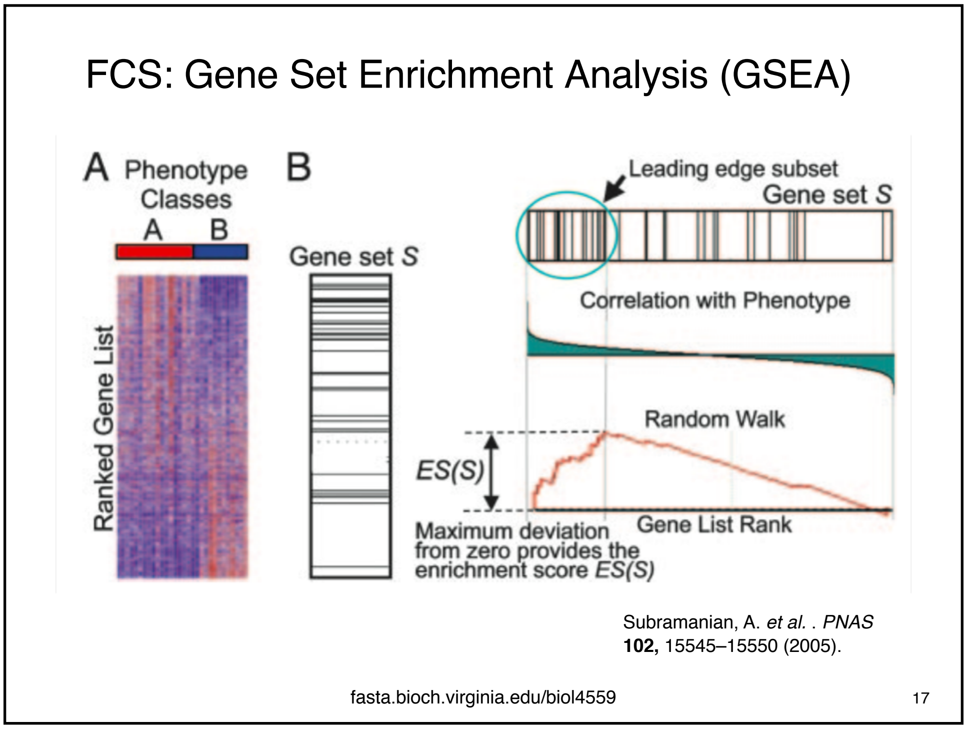
FCS: Gene Set Enrichment Analysis (GSEA)

Given an *a priori* defined set of genes *S* (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout list *L* or primarily found at the top or bottom.



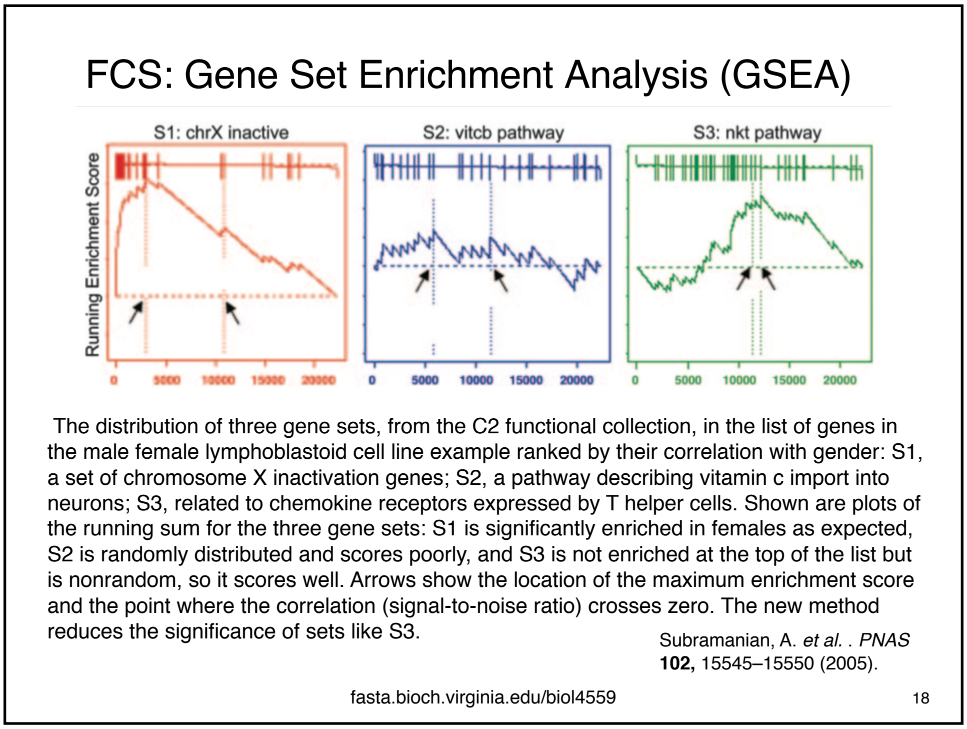
Subramanian, A. *et al.* . *PNAS* **102,** 15545–15550 (2005).

- no P value/FDR threshold
- more sensitive than hypergeometric tests
- statistics calculated by permutation testing



FCS: Gene Set Enrichment Analysis (GSEA)

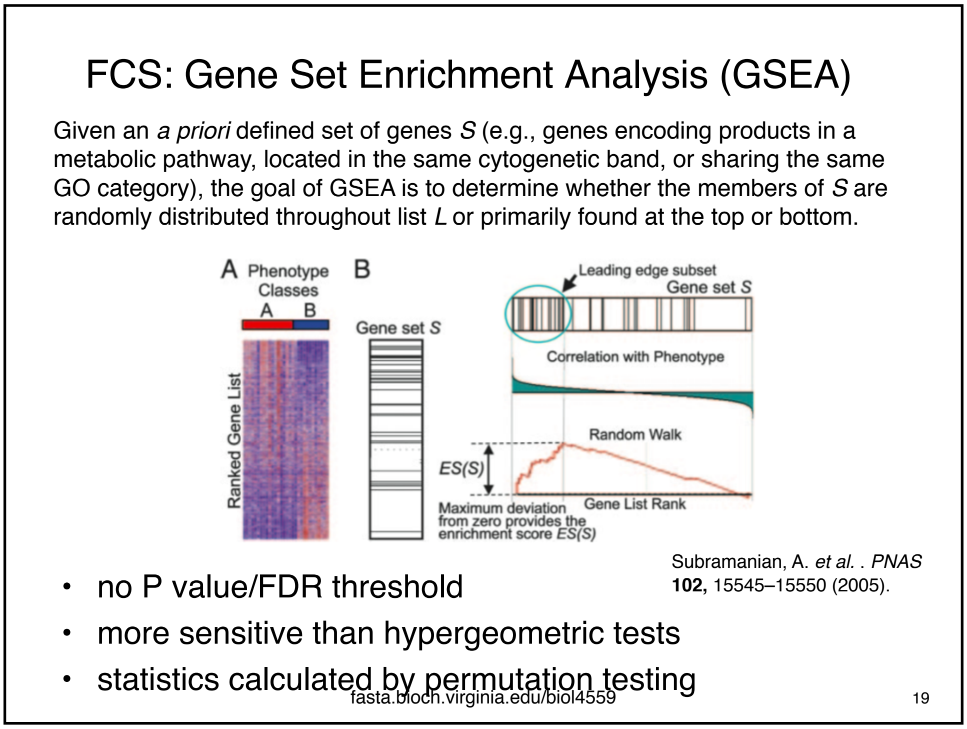
Subramanian, A. *et al.* . *PNAS* **102,** 15545–15550 (2005).



FCS: Gene Set Enrichment Analysis (GSEA)

The distribution of three gene sets, from the C2 functional collection, in the list of genes in the male female lymphoblastoid cell line example ranked by their correlation with gender: S1, a set of chromosome X inactivation genes; S2, a pathway describing vitamin c import into neurons; S3, related to chemokine receptors expressed by T helper cells. Shown are plots of the running sum for the three gene sets: S1 is significantly enriched in females as expected, S2 is randomly distributed and scores poorly, and S3 is not enriched at the top of the list but is nonrandom, so it scores well. Arrows show the location of the maximum enrichment score and the point where the correlation (signal-to-noise ratio) crosses zero. The new method reduces the significance of sets like S3.

Subramanian, A. *et al.* . *PNAS* **102,** 15545–15550 (2005).

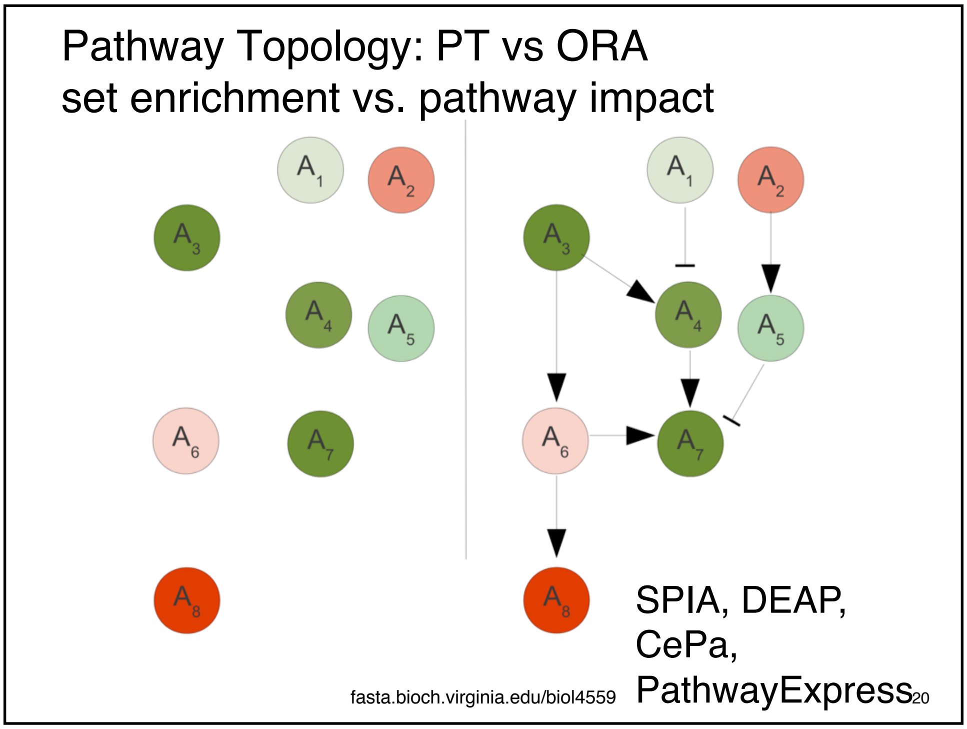


FCS: Gene Set Enrichment Analysis (GSEA)

Given an *a priori* defined set of genes *S* (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout list *L* or primarily found at the top or bottom.

Subramanian, A. *et al.* . *PNAS* **102,** 15545–15550 (2005).

- no P value/FDR threshold
- more sensitive than hypergeometric tests
- statistics calculated by permutation testing



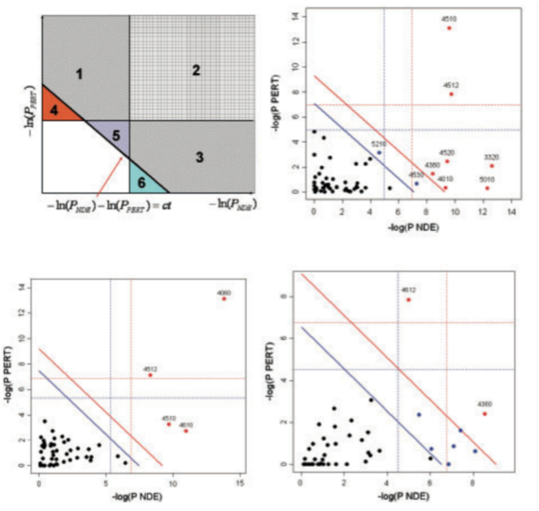
Pathway Topology: PT vs ORA set enrichment vs. pathway impact

SPIA, DEAP, CePa, PathwayExpress

SPIA – Signaling Pathway Impact Analysis

The X-axis shows the over-representation evidence, while the Y-axis shows the perturbation evidence. In the top-left plot, areas 2, 3 and 6 together will include pathways that meet the over-representation criterion (PNDE <α). Areas 1, 2 and 4 together will include pathways that meet the perturbation criterion (PPERT <α). Areas 1, 2, 3 and 5 will include the pathways that meet the combined SPIA criteria (PG <α). Note how SPIA results are different from a mere logical operation between the two criteria (OR would be areas 1, 2, 3, 4 and 6; AND would be area 2).

Pathway analysis results on the Colorectal cancer (top right), LaborC (bottom left) and Vessels (bottom right) datasets. Each pathway is represented by a point. Pathways above the oblique red line are significant at 5% after Bonferroni correction, while those above the oblique blue line are significant at 5% after FDR correction. The vertical and horizontal thresholds represent the same corrections for the two types of evidence considered individually. Note that for the Colorectal cancer dataset (top right), the colorectal cancer pathway (ID = 5210) is only significant according to the combined evidence but not so according to any individual evidence PNDE or PPERT.

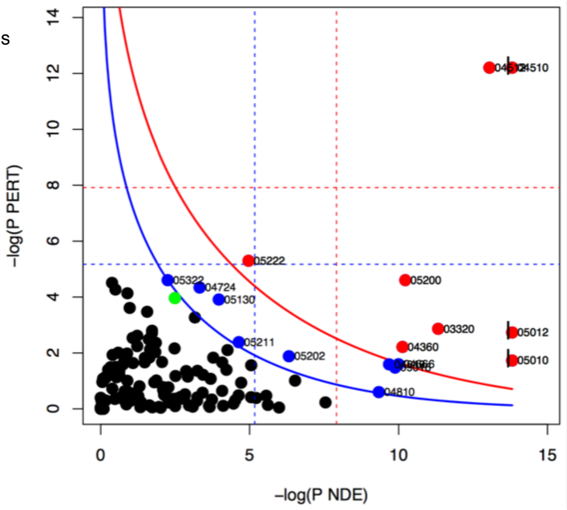


Tarca *et al. Bioinformatics* **25,** 75–82 (2009).

SPIA – Signaling Pathway Impact Analysis

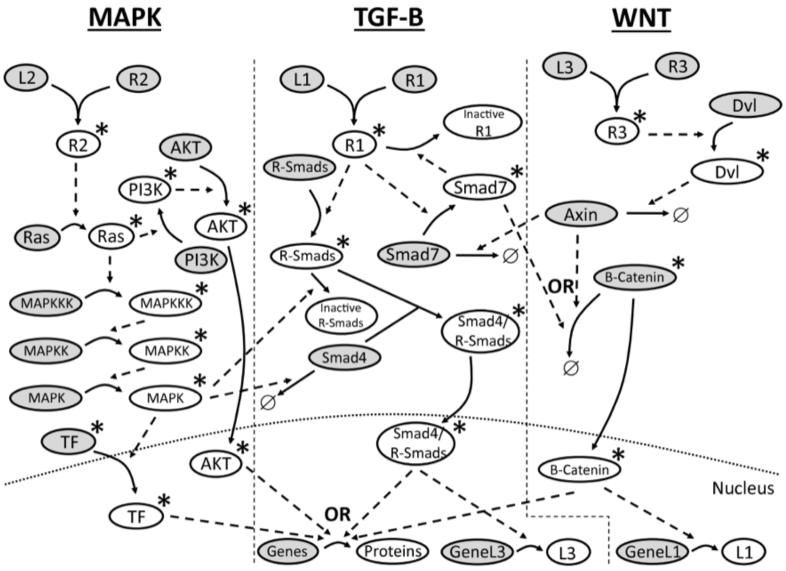
Figure 3: SPIA evidence plot for the colorectal cancer dataset. Each pathway is represented by one dot. The pathways at the right of the red curve are significant after Bonferroni correction of the global p-values, pG, obtained by combining the pPERT and pNDE using the normal inversion method. The pathways at the right of the blue curve line are significant after a FDR correction of the global p-values, pG.

The green dot shows the KEGG:05210 colon cancer pathway. This pathway is marginally significant (RDR < 0.05) with "normal inversion" combination of PERT and NDE, but not significant with Fisher's method.

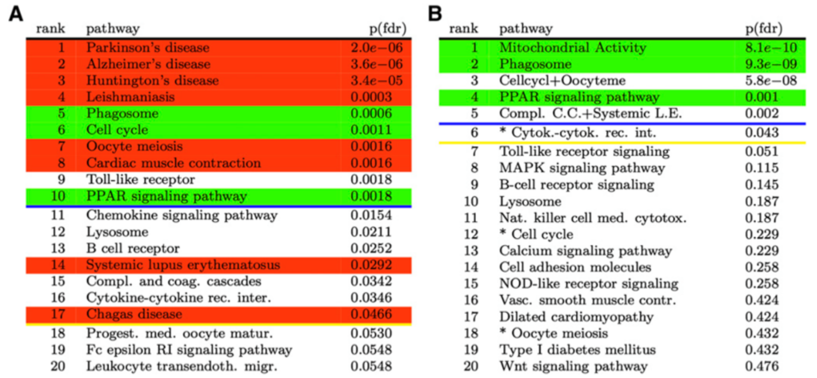


http://www.bioconductor.org/packages/release/bioc/vignettes/SPIA/inst/doc/SPIA.pdf

many pathways exhibit “crosstalk”



pathway crosstalk yields false positives:



A

rank	pathway	p(fdr)
1	Parkinson’s disease	2.0e−06
2	Alzheimer’s disease	3.6e−06
3	Huntington’s disease	3.4e−05
4	Leishmaniasis	0.0003
5	Phagosome	0.0006
6	Cell cycle	0.0011
7	Oocyte meiosis	0.0016
8	Cardiac muscle contraction	0.0016
9	Toll-like receptor	0.0018
10	PPAR signaling pathway	0.0018
11	Chemokine signaling pathway	0.0154
12	Lysosome	0.0211
13	B cell receptor	0.0252
14	Systemic lupus erythematosus	0.0292
15	Compl. and coag. cascades	0.0342
16	Cytokine-cytokine rec. inter.	0.0346
17	Chagas disease	0.0466
18	Progest. med. oocyte matur.	0.0530
19	Fc epsilon RI signaling pathway	0.0548
20	Leukocyte transendoth. migr.	0.0548

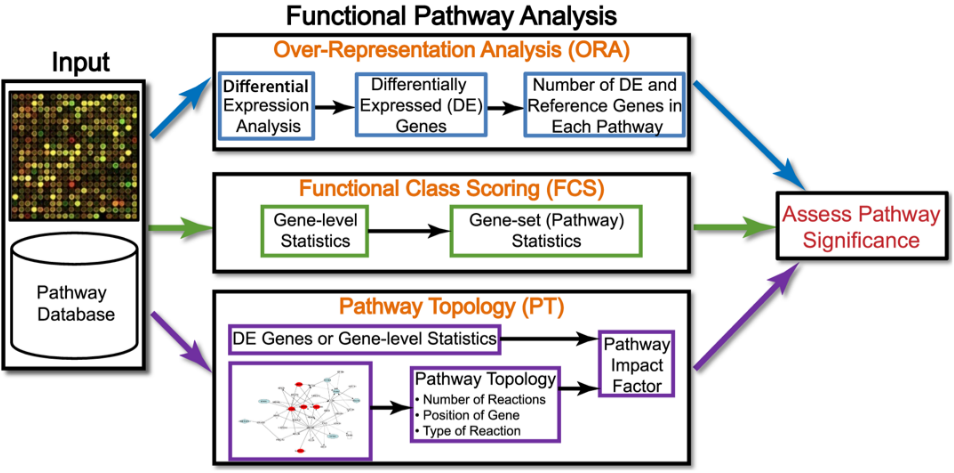
B

rank	pathway	p(fdr)
1	Mitochondrial Activity	8.1e−10
2	Phagosome	9.3e−09
3	Cellcycl+Oocyteme	5.8e−08
4	PPAR signaling pathway	0.001
5	Compl. C.C.+Systemic L.E.	0.002
6	* Cytok.-cytok. rec. int.	0.043
7	Toll-like receptor signaling	0.051
8	MAPK signaling pathway	0.115
9	B-cell receptor signaling	0.145
10	Lysosome	0.187
11	Nat. killer cell med. cytotox.	0.187
12	* Cell cycle	0.229
13	Calcium signaling pathway	0.229
14	Cell adhesion molecules	0.258
15	NOD-like receptor signaling	0.258
16	Vasc. smooth muscle contr.	0.424
17	Dilated cardiomyopathy	0.424
18	* Oocyte meiosis	0.432
19	Type I diabetes mellitus	0.432
20	Wnt signaling pathway	0.476

The results of the ORA analysis in the fat remodeling experiment for the comparison between days 3 and 0, before (A) and after (B) correction for crosstalk effects. All P-values are FDR corrected. The lines show the significance thresholds: (blue) 0.01, (yellow) 0.05. Pathways highlighted in red represent pathways not related to the phenomenon in analysis, while pathways highlighted in green are those for which we know, with reasonable confidence, are involved in the given phenomenon. The white background indicates pathways for which we do not have conclusive information on their involvement (or lack of) with the phenomenon in analysis. (A) The top 20 pathways resulting from classical ORA before correction for crosstalk. The top four pathways are not related to fat remodeling. (B) The top 20 pathways after correction for crosstalk. Pathways ranked 1, 3, and 5 are modules that are functioning independently of the rest of their pathways in this particular condition. Starred pathways are pathways edited by removing such modules. Note the lack of any obvious false positive above the significance threshold(s).

Donato, M. *et al. Genome Res* **23,** 1885–1893 (2013)

from genes to pathways: enrichment analysis



Functional Pathway Analysis

Input

Pathway Database

Over-Representation Analysis (ORA)

Differential Expression Analysis → Differentially Expressed (DE) Genes → Number of DE and Reference Genes in Each Pathway

Functional Class Scoring (FCS)

Gene-level Statistics → Gene-set (Pathway) Statistics

Pathway Topology (PT)

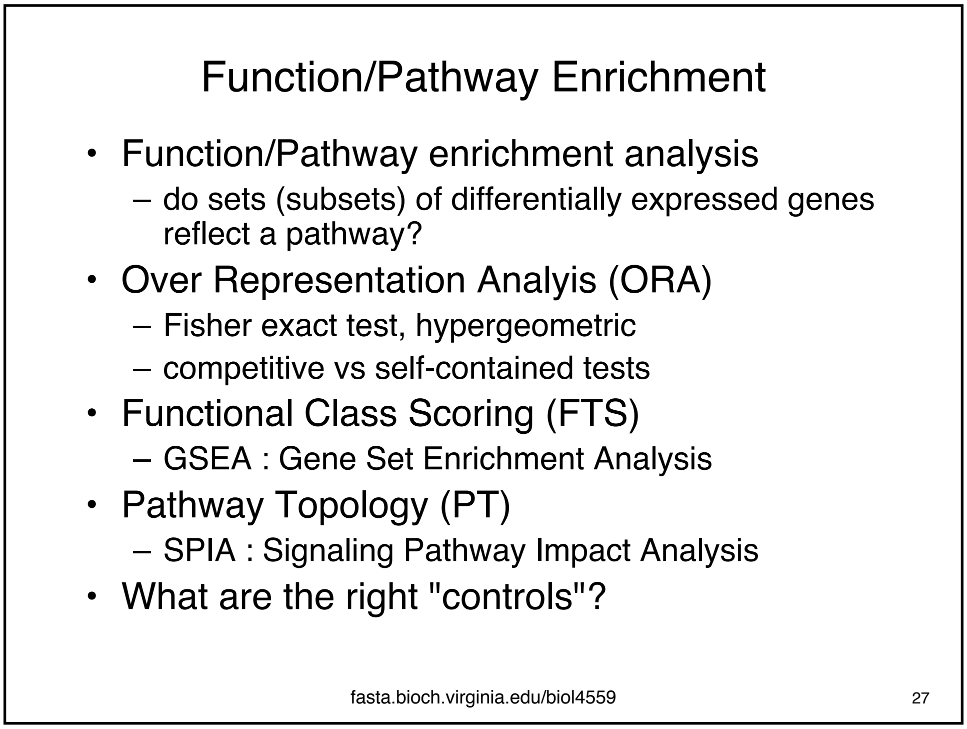
DE Genes or Gene-level Statistics → Pathway Impact Factor

Pathway Topology
- Number of Reactions
- Position of Gene
- Type of Reaction

Assess Pathway Significance

Functional analyis: ORA, FC, PT

- Methods assume independence, but pathways and GO DAGs are anything but independent
  – statistics may be too generous (false positives)
  – statistics may be too strict (false negatives)
- What is the right control?
  – try different approaches?
  – compare to other published datasets?
  – do "positive control" on well understood pathways
- All methods need experimental confirmation
  – find a drug that blocks the pathway
  – ablate a gene (or genes) in the pathway



Function/Pathway Enrichment

- Function/Pathway enrichment analysis
  - do sets (subsets) of differentially expressed genes reflect a pathway?
- Over Representation Analyis (ORA)
  - Fisher exact test, hypergeometric
  - competitive vs self-contained tests
- Functional Class Scoring (FTS)
  - GSEA : Gene Set Enrichment Analysis
- Pathway Topology (PT)
  - SPIA : Signaling Pathway Impact Analysis
- What are the right "controls"?