Gene Annotation and Gene Set Analysis

After you obtain a short list of genes/clusters/classifiers—what next?

- For each gene, you may ask
  - What it is
  - What is does
  - What processes is it involved in
  - Which chromosome is it located in
  - Which are the genes that may involve in common pathways/functions or may be physically close to this gene
  - …

Useful Bioconductor packages

- AnnotationDbi: Provides user interface and database connection code for annotation data packages using SQLite data storage.
- biomaRt: enables retrieval of large amounts of data in a uniform way without the need to know the underlying database schemas or write complex SQL queries. Examples of BioMart databases are Ensembl,Uniprot and HapMap.
- GO.db: gene ontology database

Microarray example

From meaningless probeset IDs to something interpretable

```
> library("hgu133a.db")
> ls("package:hgu133a.db")
 [1] "hgu133a"               "hgu133a_dbconn"        "hgu133a_dbfile"
 [4] "hgu133a_dbInfo"        "hgu133a_dbschema"      "hgu133aACCNUM"
 [7] "hgu133aALIAS2PROBE"    "hgu133aCHR"            "hgu133aCHRLENGTHS"
[10] "hgu133aCHRLOC"         "hgu133aCHRLOCEND"      "hgu133aENSEMBL"
[13] "hgu133aENSEMBL2PROBE" "hgu133aENTREZID"       "hgu133aENZYME"
[16] "hgu133aENZYME2PROBE"  "hgu133aGENENAME"       "hgu133aGO"
[19] "hgu133aGO2ALLPROBES"  "hgu133aGO2PROBE"       "hgu133aMAP"
[22] "hgu133aMAPCOUNTS"     "hgu133aOMIM"           "hgu133aORGANISM"
[25] "hgu133aORGPKG"        "hgu133aPATH"           "hgu133aPATH2PROBE"
[28] "hgu133aPFAM"          "hgu133aPMID"           "hgu133aPMID2PROBE"
[31] "hgu133aPROSITE"       "hgu133aREFSEQ"         "hgu133aSYMBOL"
[34] "hgu133aUNIGENE"       "hgu133aUNIPROT"
> probeSet=ls(hgu133aSYMBOL)
>get(probeSet[100],hgu133aSYMBOL)
[1] "KARS"
>  get(probeSet[100],hgu133aGENENAME)
[1] "lysyl-tRNA synthetase"
```

Gene Ontology consortium

- http://www.geneontology.org/
- Gene names can be insufficient and unclear
  - The same name can be used to describe different concepts
  - One gene can have more than one functions
  - Different terms may refer to the same function
    - Glucose synthesis
    - Glucose biosynthesis
    - Glucose formation
    - Glucose anabolism
    - Gluconeogenesis
    - **All refer to the process of making glucose from simpler components**

- GO address the need for **consistent descriptions** of gene products in different databases
- three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

```
 tmp= get(probeSet[100],hgu133aGO)
> tmp[1]
$`GO:0006430`
$`GO:0006430`$GOID
[1] "GO:0006430"
$`GO:0006430`$Evidence
[1] "IEA“     (IEA: Inferred from Electronic Annotation )
$`GO:0006430`$Ontology
[1] "BP"
> Term("GO:0006430")
                 GO:0006430
"lysyl-tRNA aminoacylation"
> Definition("GO:0006430")
    GO:0006430
"The process of coupling lysine to lysyl-tRNA, catalyzed by lysyl-tRNA synthetase. In
    tRNA aminoacylation, the amino acid is first activated by linkage to AMP and then
    transferred to either the 2'- or the 3'-hydroxyl group of the 3'-adenosine residue
    of the tRNA."
>
```

Details see:

http://www.bioconductor.org/packages/release/data/annotation/manuals/GO.db/man/GO.db.pdf

The 3 Gene Ontologies

- Molecular function (elemental activity/task)
  - activities at the molecular level.
  - Examples: catalytic or binding activities
- Biological process (biological goal or objective)
  - series of events accomplished by one or more ordered assemblies of molecular functions.
  - Examples: cellular physiological process or signal transduction
- Cellular component (location or complex)
  - Where does gene product act
  - Examples: an anatomical structure (endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

Some other examples of evidence codes

Experimental Evidence Codes
- EXP: Inferred from Experiment
- IDA: Inferred from Direct Assay

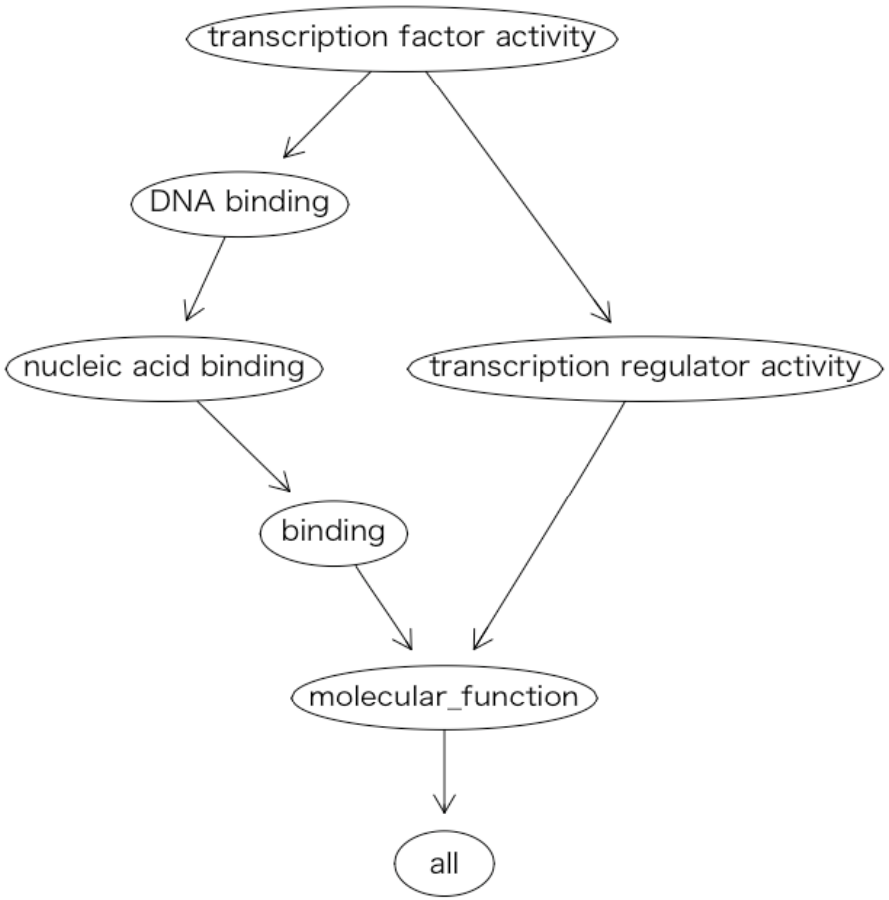
Computational Analysis Evidence Codes
- ISS: Inferred from Sequence or Structural Similarity
- ISA: Inferred from Sequence Alignment

Author Statement Evidence Codes
- TAS: Traceable Author Statement

Curator Statement Evidence Codes
- IC: Inferred by Curator

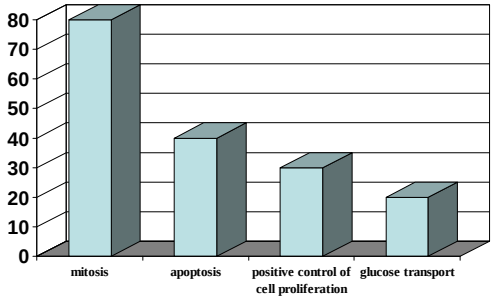
Automatically-assigned Evidence Codes
- IEA: Inferred from Electronic Annotation

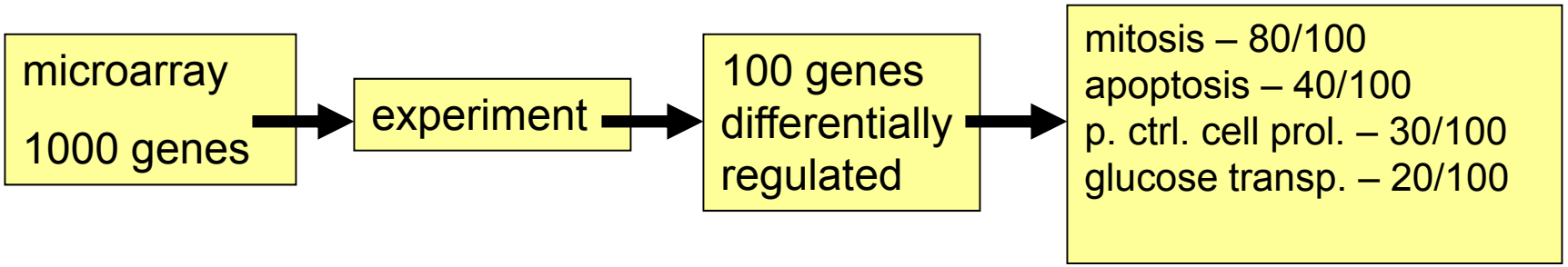
For a complete liste, see
http://www.geneontology.org/GO.evidence.shtml


transcription factor activity
DNA binding
nucleic acid binding
transcription regulator activity
binding
molecular_function
all

GOstat

- statistical measure
  - how likely your differentially regulated genes fall into that category by chance





Using GO in practice

- However, when you look at the distribution of all genes on the microarray:

Process	Genes on array	# genes expected in 100 random genes	occurred
mitosis	800/1000	80	80
apoptosis	400/1000	40	40
p. ctrl. cell prol.	100/1000	10	30
glucose transp.	50/1000	5	20

GOstats: finding overrepresented GO terms

Go term "**$`04910`**
[1] "Insulin signaling pathway"

X^2 test

	Short list	Reference list	
Appeared	25	100	125
Not	25	1900	1925
	50	2000	

Is this term over represented in your short list?

$$\chi^2 = \Sigma \, (O-E)^2/E$$

Null hypothesis H_0: equal representation in the short list and reference list.

GOstats: finding overrepresented GO terms

- Fisher's exact test

	Short list	Reference list	
Appeared	3	100	103
Not	17	1900	1917
	20	2000	

$$P(n_{11} = t) = \frac{\binom{103}{t}\binom{1917}{20-t}}{\binom{2200}{20}}$$

- GoStats sets a cutoff to obtain a subset of genes first (usually declared to be significantly different across conditions), and test if a GO term is over-represented in the subset
- Another angle to consider the problem is to start with predefined gene sets instead of data-determined short list.

Limitations of differential expression focusing on individual genes

- Sometimes too few genes are found to be significant after accounting for multiple testing. If the effect size is small, there may not be enough power
- Sometimes the opposite is true: too many genes appear to be different, without any unifying biological theme
- An important process may be different between the two conditions under comparison, yet each gene in this process is only moderately affected.
- An important process may be affected, yet not always the same gene in the process. Thus replication experiments may find little overlap in the actual gene list

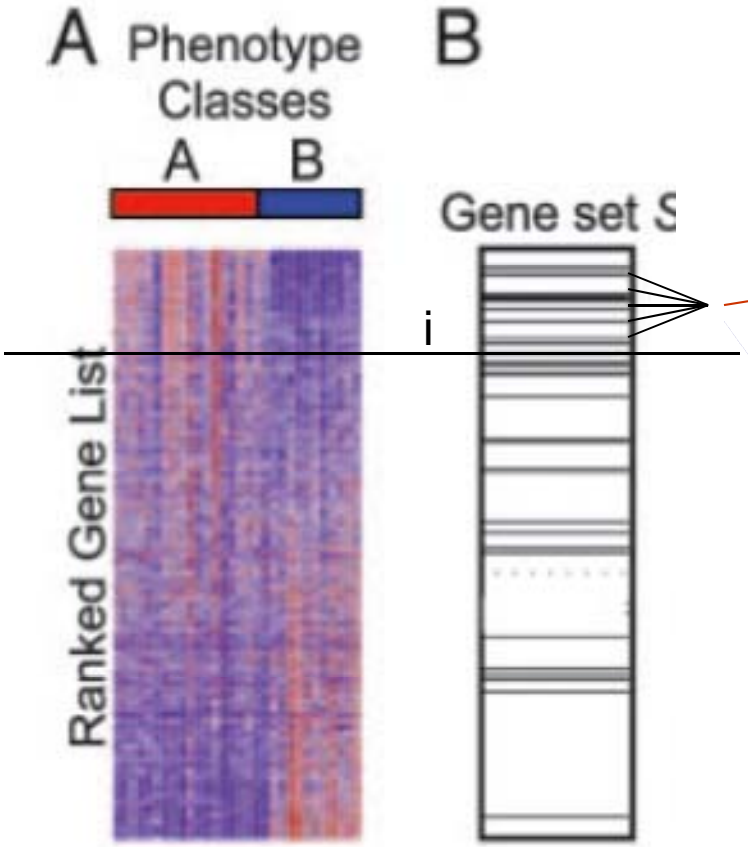
Subramanian et al, 2005 PNAS 102(43)15545

Consider a group of a gene instead

- Consider gene sets defined a priori
  - Metabolic pathway
  - Cellular component
  - GO category
- The complete list of genes, ordered by a statistics representing differential expression : L
- Null: Gene Set S is not associated with the conditions being compared
  - Genes in a gene set S are randomly distributed throughout L
- Alternative: Genes in S would be *enriched* at the tails of L

Enrichment

- Sort the N genes into an ordered list *L* by a metric *r* representing the association with phenotype
  - $r(g_j)=r_j$
  - $L=\{g_1,\ldots g_N\}$
- Walking down the list *L*, at position *i*: for the top *i* genes, some of them belongs to set *S* (hits) and some do not (miss)



$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

hits

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}$$

The rest are misses

ES: max(P_{hit}-P_{Miss})

$N_{H:}$The size of set S

p: weight

p=0: ES reduces to Kolmogorov statistic

Statistical significance

- Permuting the sample labels
- Compute enrichment score for each permuted dataset
- Establish ES_{NULL} from permutation
- Computing p-value for actual ES

Kolmogorov–Smirnov statistic

Empirical CDF of n observations

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$$

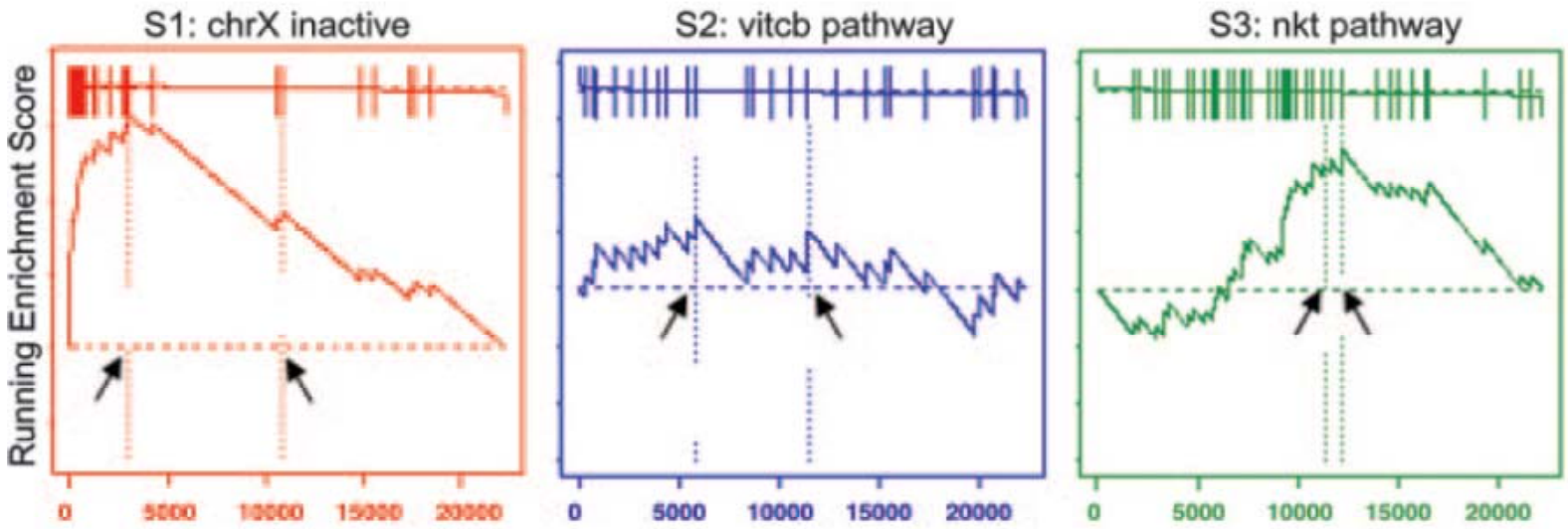
For two samples, the Kolmogorov–Smirnov statistic is

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|$$

The null hypothesis, that the two samples are drawn from the same distribution, is rejected if

$$\sqrt{\frac{nn'}{n+n'}} D_{n,n'} > K_\alpha$$

where k is the Kolmogorov distribution.



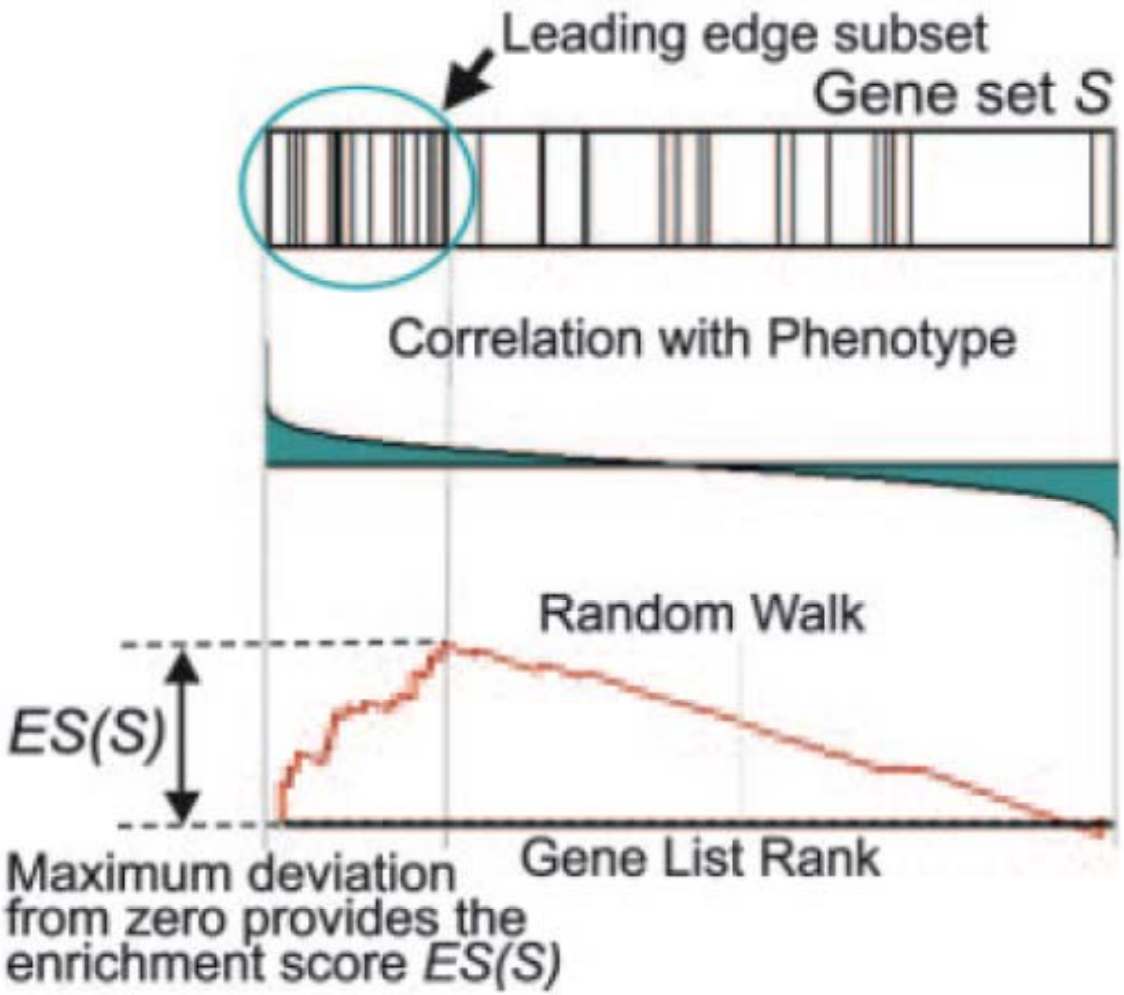
Enrichment scores using p=0

A.Enriched at top

B.Random

C.Sets clustered near the middle are not null (Random), but not biologically interesting.

--Solution: using weight p>0, typically p=1

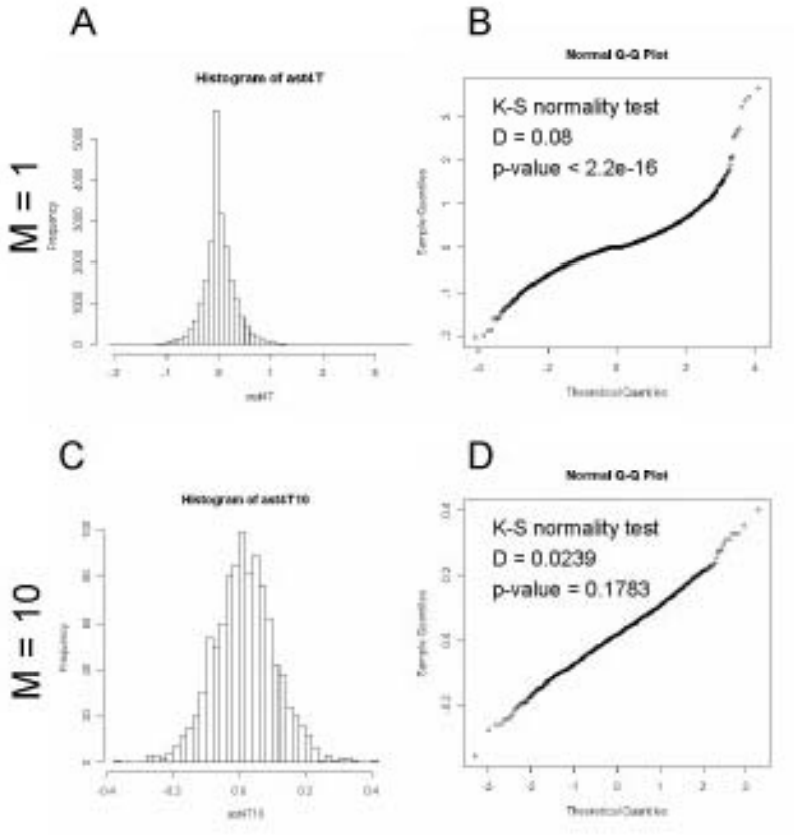

Leading edge subset
Gene set S
Correlation with Phenotype
Random Walk
ES(S)
Gene List Rank
Maximum deviation from zero provides the enrichment score ES(S)

PAGE: Parametric Analysis of Gene Set Enrichment

Seon-Young Kim and **David J Volsky** BMC Bioinformatics

- From the entire list of genes
  - mean of total fold change values (μ)
  - standard deviation of total fold change (δ)
- For a given gene set of size *m*
  - mean of fold change values of genes in the set : S_m
  - Z score = $\dfrac{Sm - \mu}{\delta / \sqrt{m}}$

- CLT: if m is large, S_m is a sample mean. The distribution of sample means would be close to a normal distribution
- How big does m have to be?



Questions for you

- When do you get to reject the null?
- Which gene sets may be interesting, but will not be detected?
- Multiple testing
- “Reference distribution”