

A quick review

■ The clustering problem:

- partition genes into distinct sets with high homogeneity and high separation

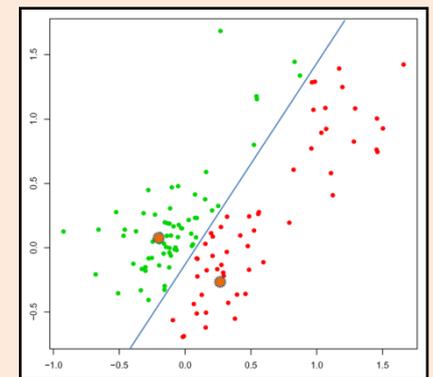
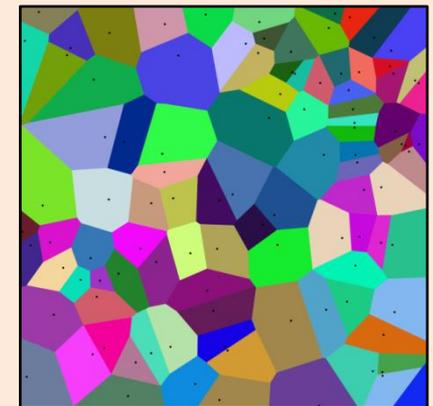
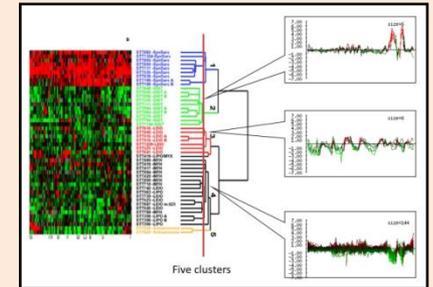
■ Hierarchical clustering algorithm:

1. Assign each object to a separate cluster.
2. Regroup the pair of clusters with shortest distance.
3. Repeat 2 until there is a single cluster.

■ Many possible distance metrics

■ K-mean clustering algorithm:

1. Arbitrarily select k initial centers
2. Assign each element to the closest center
 - **Voronoi** diagram
3. Re-calculate centers (i.e., means)
4. Repeat 2 and 3 until termination condition reached



Gene Ontology and Functional Enrichment

Genome 373

Genomic Informatics

Elhanan Borenstein

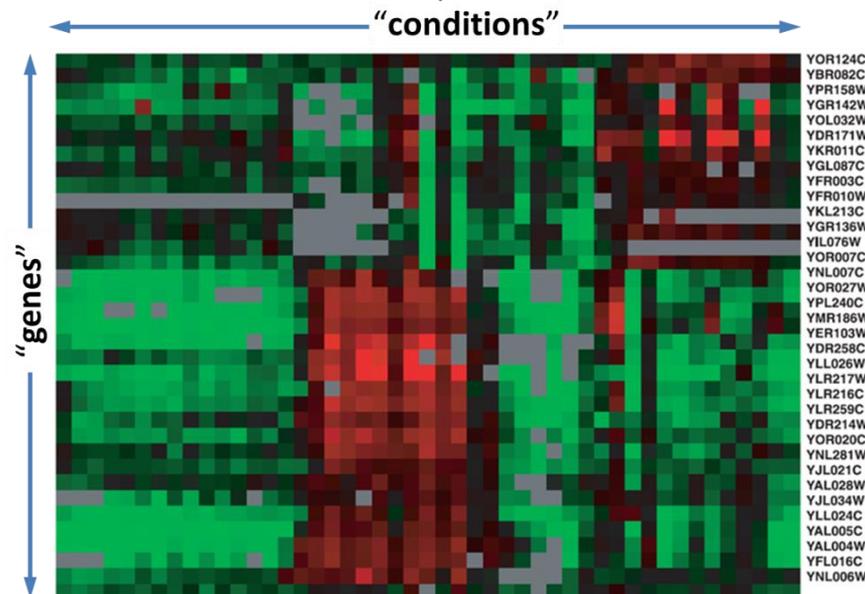
From sequence to function

Which molecular processes/functions are involved in a certain phenotype - disease, response, development, etc.

(what is the cell doing vs. what it could possibly do)



Gene expression profiling



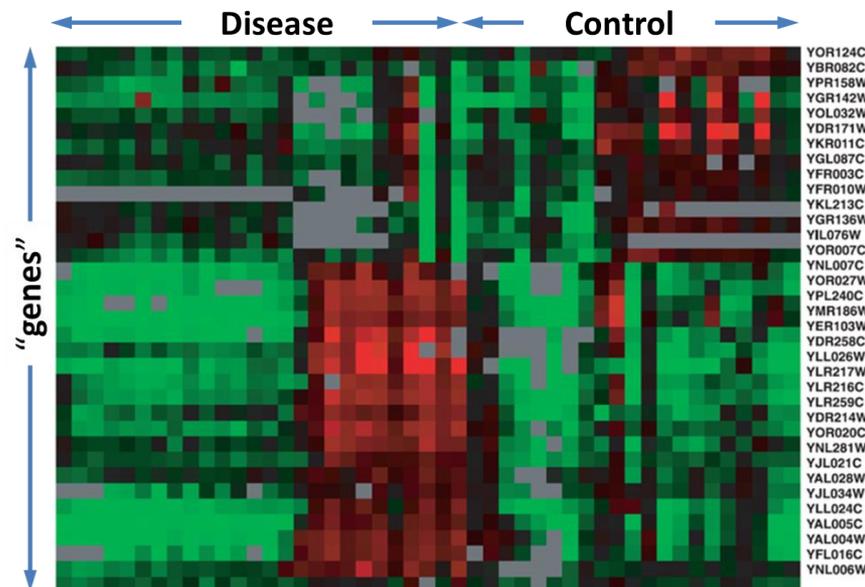
From sequence to function

Which molecular processes/functions are involved in a certain phenotype - disease, response, development, etc.

(what is the cell doing vs. what it could possibly do)

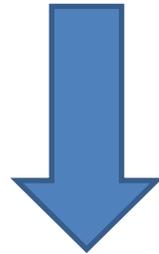


Gene expression profiling



Back in the good old days ...

1. Find the set of differentially expressed genes.
2. Survey the literature to obtain insights about the functions that differentially expressed genes are involved in.
3. Group together genes with similar functions.
4. Identify functional categories with many differentially expressed genes.



Conclude that these functions are important in disease/condition under study

The good old days were not so good!

Time-consuming

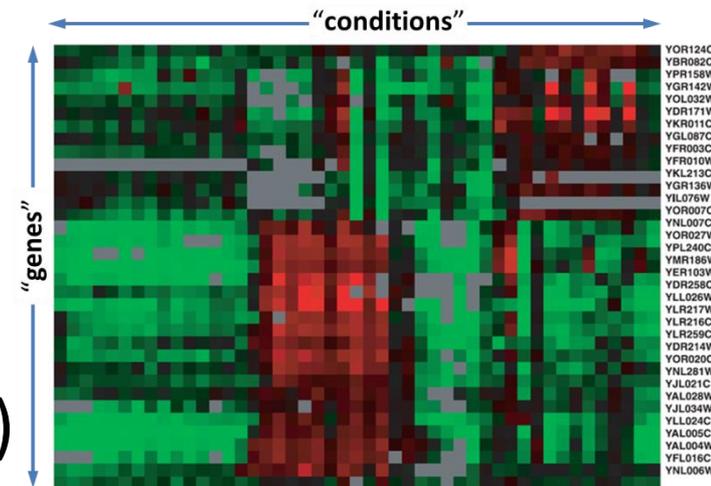
Not systematic

Extremely subjective

No statistical validation

What do we need?

- A shared functional vocabulary
- Systematic linkage between genes and functions
- A way to identify genes relevant to the condition under study
- Statistical analysis
(combining all of the above to identify cellular functions that contributed to the disease or condition under study)
- (A way to identify “related” genes)



What do we need?

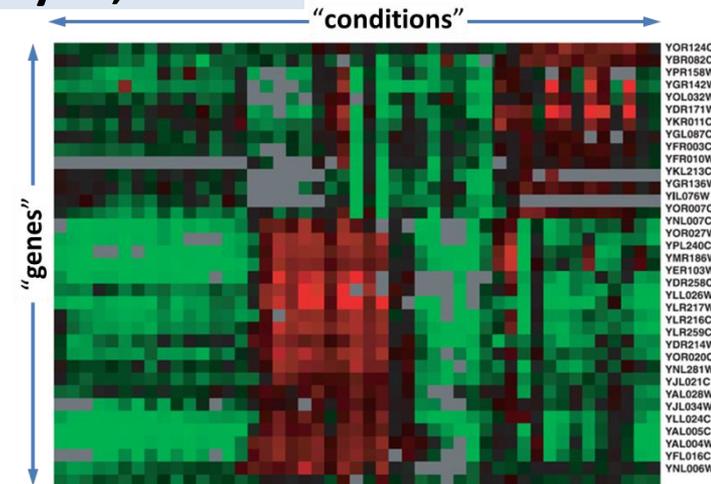
- A shared functional vocabulary
- Systematic linkage between genes and functions
- A way to identify genes relevant to the condition under study
- Statistical analysis
(combining all of the above to identify cellular functions that contributed to the disease or condition under study)

Gene Ontology

Annotation

Fold change,
Ranking, ANOVA

Enrichment
analysis, GSEA



The Gene Ontology (GO) Project

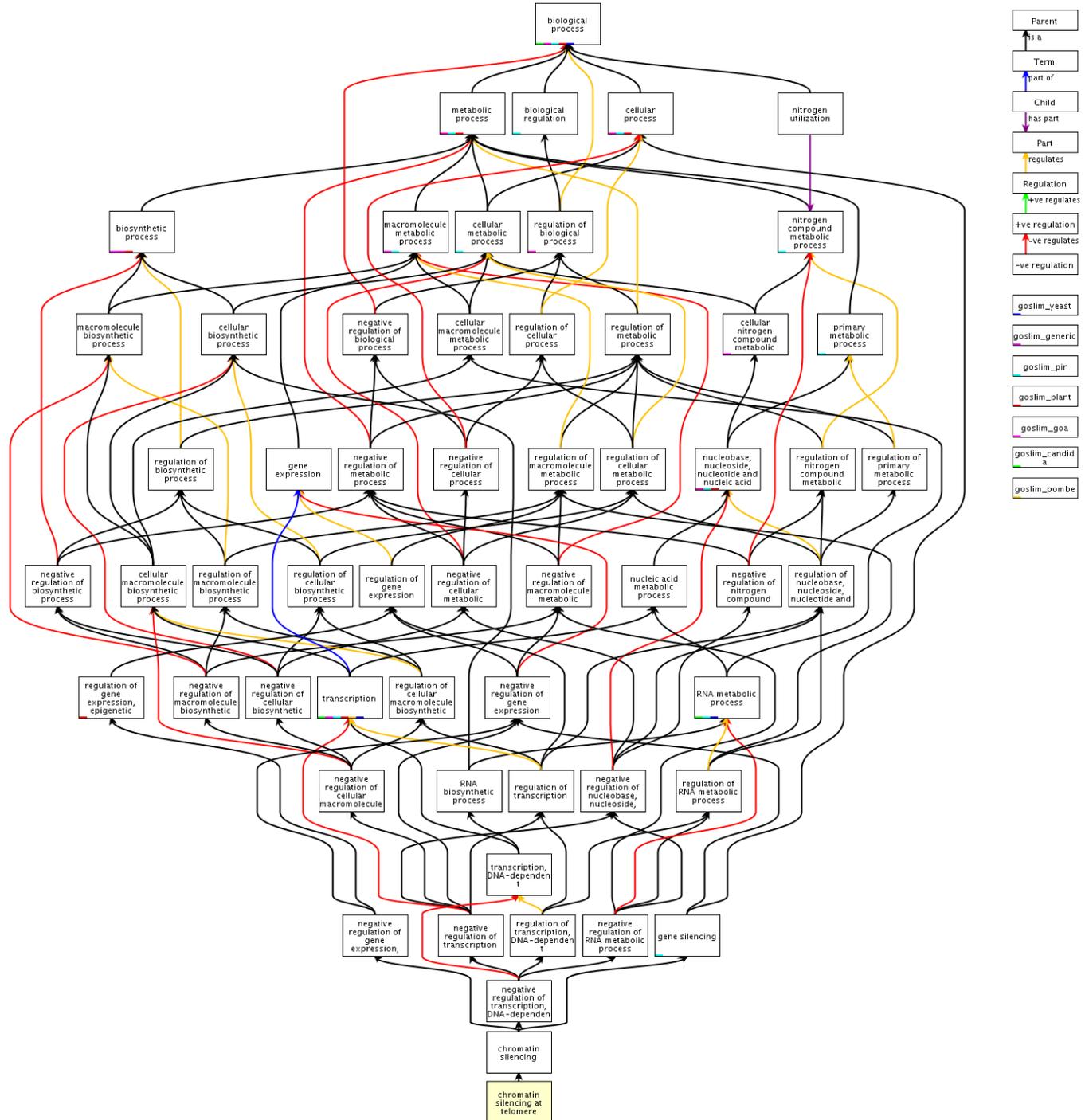
- A major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.
- Three goals:
 1. *Maintain and further develop its controlled **vocabulary** of gene and gene product attributes*
 2. ***Annotate** genes and gene products, and assimilate and disseminate annotation data*
 3. *Provide **tools** to facilitate access to all aspects of the data provided by the Gene Ontology project*

GO terms

- The Gene Ontology (GO) is a **controlled vocabulary**, a set of standard **terms** (words and phrases) used for indexing and retrieving information.

| Term Information | |
|-------------------|--|
| Accession | GO:0006348 |
| Ontology | Biological Process |
| Synonyms | exact: heterochromatic silencing at telomere exact: telomere chromatin silencing exact: telomeric silencing |
| Definition | Repression of transcription of telomeric DNA by altering the structure of chromatin. <i>Source:</i> PMID:10219245 |
| Comment | None |
| Subset | None |
| Community | Add usage comments for this term at GONUTS . |

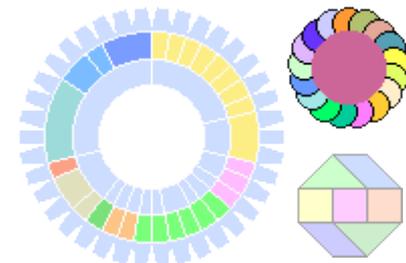
Back to top



Ontology and annotation databases



eggNOG



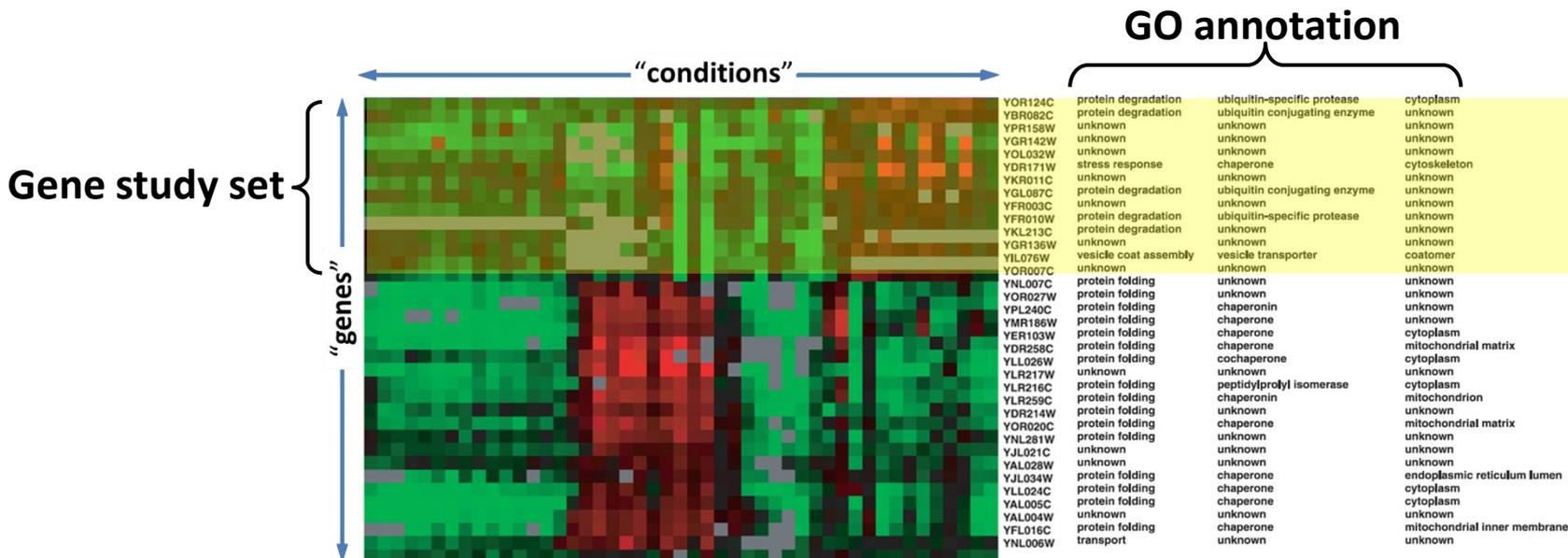
Clusters of Orthologous
Groups (COG)

“The nice thing about standards is that there are so many to choose from”

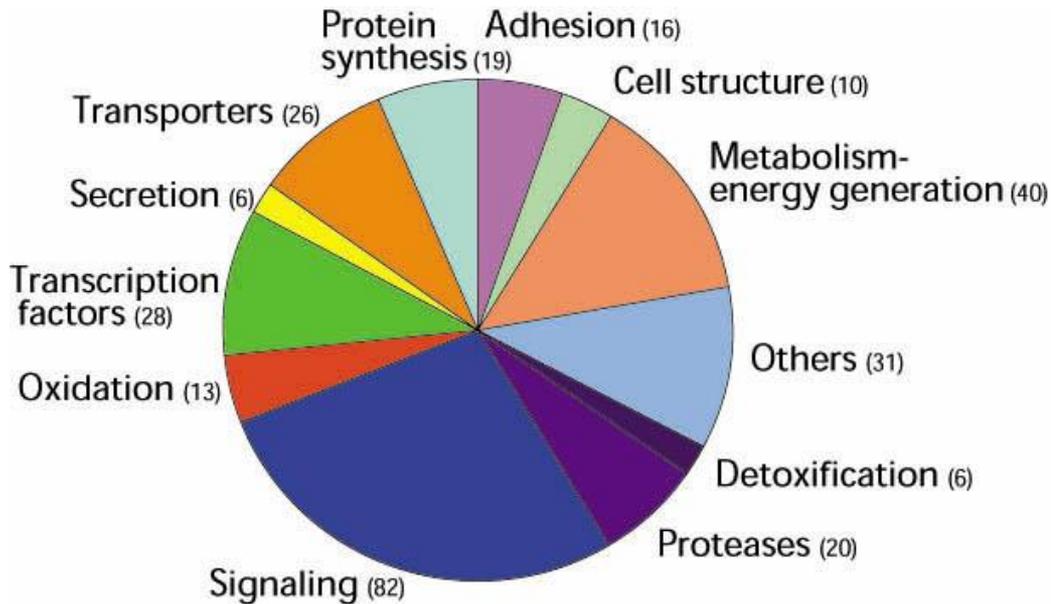
Andrew S. Tanenbaum

Picking “relevant” genes

- In most cases, we will consider differential expression as a marker:
 - Fold change cutoff (e.g., > two fold change)
 - Fold change rank (e.g., top 10%)
 - Significant differential expression (e.g., ANOVA) (don't forget to correct for multiple testing, e.g., Bonferroni or FDR)



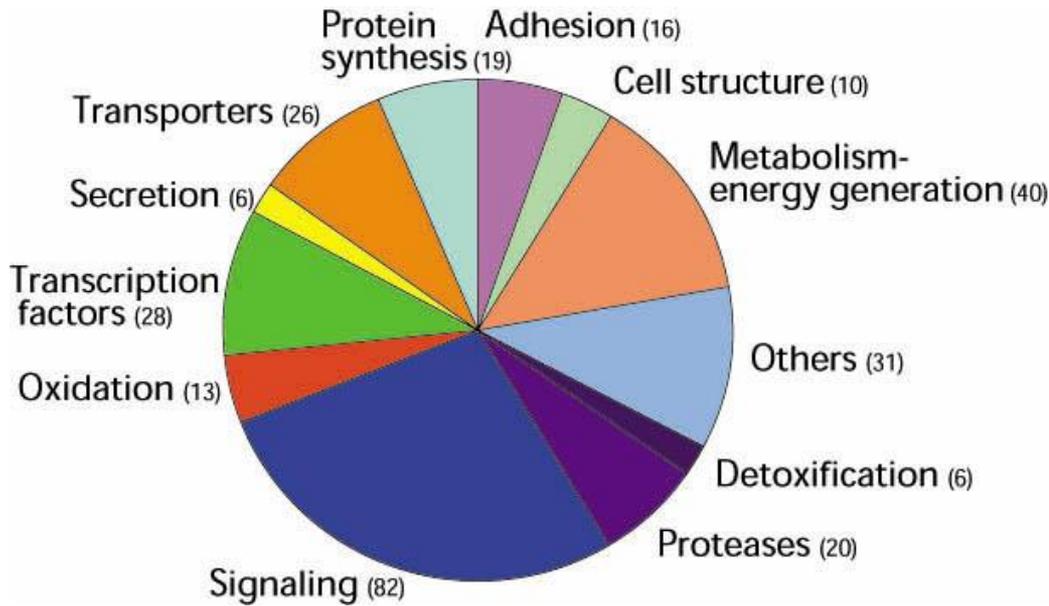
Enrichment analysis



Signalling category contains 27.6% of all genes in the study set - **by far the largest category**. Reasonable to conclude that signaling may be important in the condition under study

| Functional category | # of genes in the study set | % |
|---------------------|-----------------------------|------|
| Signaling | 82 | 27.6 |
| Metabolism | 40 | 13.5 |
| Others | 31 | 10.4 |
| Trans factors | 28 | 9.4 |
| Transporters | 26 | 8.8 |
| Proteases | 20 | 6.7 |
| Protein synthesis | 19 | 6.4 |
| Adhesion | 16 | 5.4 |
| Oxidation | 13 | 4.4 |
| Cell structure | 10 | 3.4 |
| Secretion | 6 | 2.0 |
| Detoxification | 6 | 2.0 |

Enrichment analysis – the wrong way



| Functional category | # of genes in the study set | % |
|---------------------|-----------------------------|------|
| Signaling | 82 | 27.6 |
| Metabolism | 40 | 13.5 |
| Others | 31 | 10.4 |
| Trans factors | 28 | 9.4 |
| Transporters | 26 | 8.8 |
| Proteases | 20 | 6.7 |
| Protein synthesis | 19 | 6.4 |
| Adhesion | 16 | 5.4 |
| Oxidation | 13 | 4.4 |
| Cell structure | 10 | 3.4 |
| Secretion | 6 | 2.0 |
| Detoxification | 6 | 2.0 |

Signaling category contains 27.6% of all genes in the study set - **by far the largest category.** Reasonable to conclude that signaling may be important in the condition under study.

Enrichment analysis – the wrong way

- What if ~27% of the genes on the array are involved in signaling?
 - The number of signaling genes in the set is what expected by chance.
 - We need to consider not only the number of genes in the set for each category, but also the total number on the array.

- We want to know which category is **over-represented** (occurs more times than expected by chance).

| Functional category | # of genes in the study set | % | % on array |
|-----------------------|-----------------------------|-------------|------------|
| Signaling | 82 | 27.6% | 26% |
| Metabolism | 40 | 13.5% | 15% |
| Others | 31 | 10.4% | 11% |
| Trans factors | 28 | 9.4% | 10% |
| Transporters | 26 | 8.8% | 2% |
| Proteases | 20 | 6.7% | 7% |
| Protein synthesis | 19 | 6.4% | 7% |
| Adhesion | 16 | 5.4% | 6% |
| Oxidation | 13 | 4.4% | 4% |
| Cell structure | 10 | 3.4% | 8% |
| Secretion | 6 | 2.0% | 2% |
| Detoxification | 6 | 2.0% | 2% |

Enrichment analysis – the right way

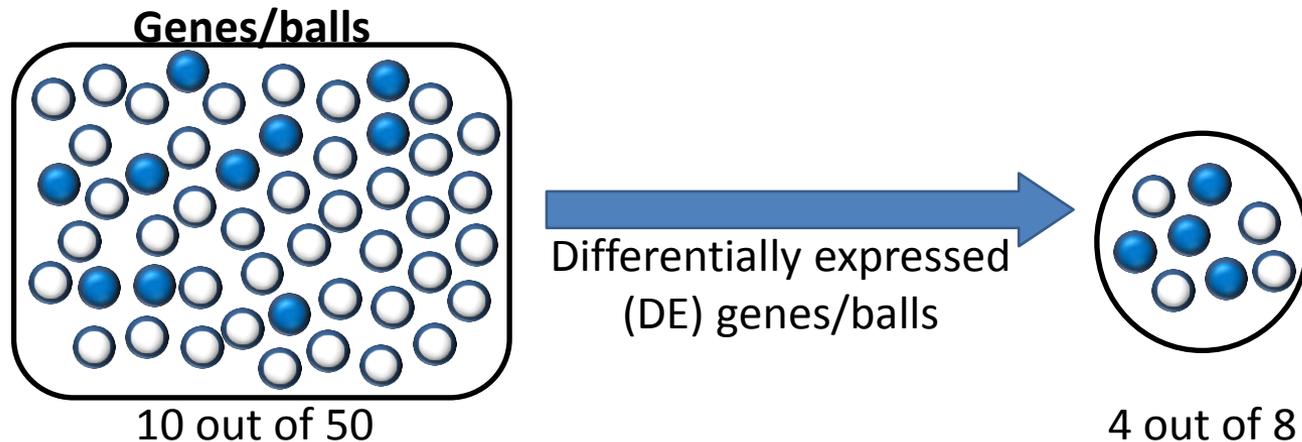
*Say, the microarray contains 50 genes, 10 of which are annotated as ‘signaling’. Your expression analysis reveals 8 differentially expressed genes, 4 of which are annotated as ‘signaling’.
Is this significant?*

A statistical test, based on a null model

Assume the study set has nothing to do with the specific function at hand and was selected randomly, would we be surprised to see this number of genes annotated with this function in the study set?

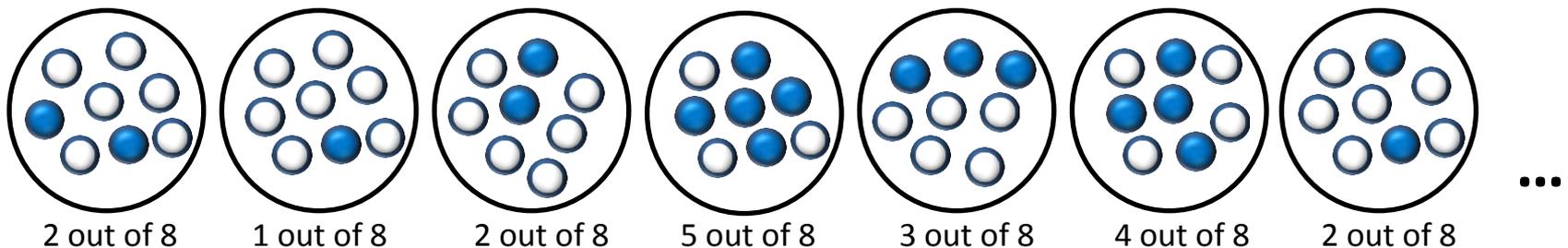
The “urn” version: You pick a random set of 8 balls from an urn that contains 50 balls: 40 white and 10 blue. How surprised will you be to find that 4 of the balls you picked are blue?

Enrichment analysis – the right way



Do I have a surprisingly high number of blue genes?

Null model: the 8 genes/balls are selected randomly



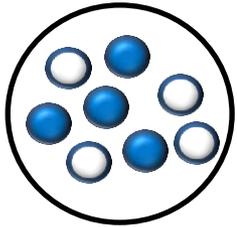
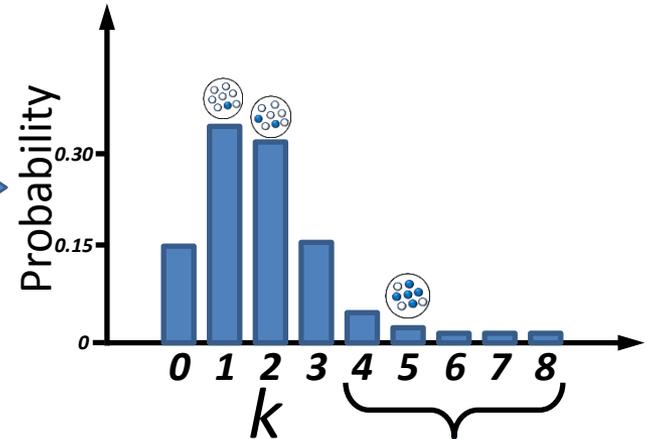
So, if you have 50 balls, 10 of them are blue, and you pick 8 balls randomly, what is the probability that k of them are blue?

Hypergeometric Distribution

Hypergeometric distribution

$$\mathbb{P}(\sigma_t = k) = \frac{\binom{m_t}{k} \binom{m-m_t}{n-k}}{\binom{m}{n}}$$

$$m=50, m_t=10, n=8$$



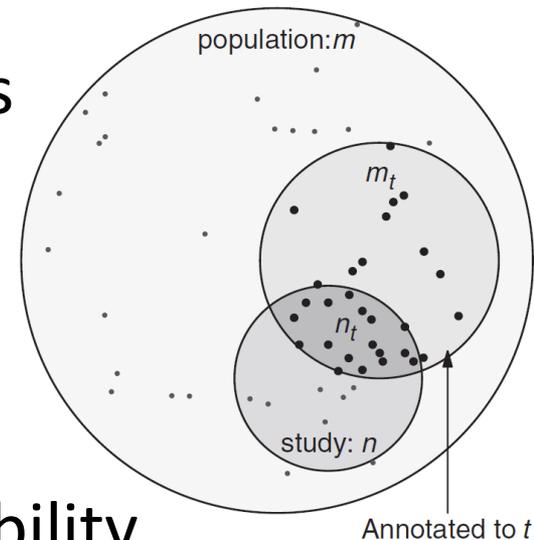
So ... do I have a surprisingly high number of blue genes?

What is the probability of getting at least 4 blue genes in the null model?

$$P(\sigma_t \geq 4)$$

Modified Fisher's Exact Test

- Let m denote the total number of genes in the array and n the number of genes in the study set.
- Let m_t denote the total number of genes annotated with function t and n_t the number of genes in the study set annotated with this function.
- We are interested in knowing the probability of seeing n_t or more annotated genes!

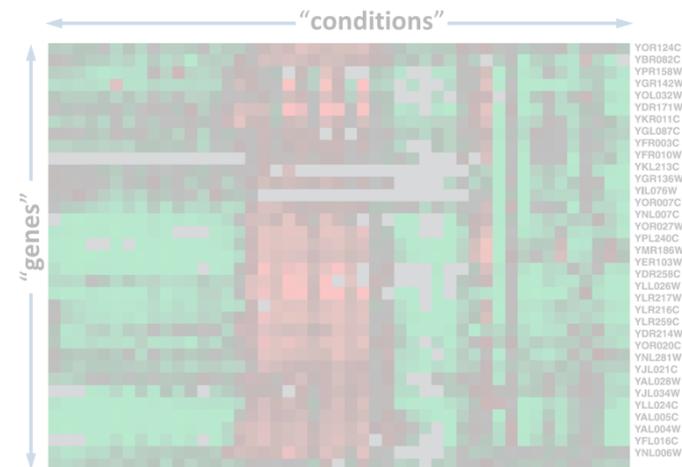


$$\mathbb{P}(\sigma_t \geq n_t) = \sum_{k=n_t}^{\min(m_t, n)} \frac{\binom{m_t}{k} \binom{m-m_t}{n-k}}{\binom{m}{n}}$$

(This is equivalent to a one-sided Fisher exact test)

So ... what do we have so far?

- A shared functional vocabulary ✓
- Systematic linkage between genes and functions ✓
- A way to identify genes relevant to the condition under study ✓
- Statistical analysis ✓
(combining all of the above to identify cellular functions that contributed to the disease or condition under study)
- A way to identify “related” genes



Still far from being perfect!

- A shared functional vocabulary
- Systematic linkage between genes and functions
 - Arbitrary!
 - Considers only a few genes
- A way to identify genes relevant to the condition under study
 - Limited hypotheses
 - Simplistic null model!
- Statistical analysis
(combining all of the above to identify cellular functions that contributed to the disease or condition under study)
- A way to identify “related” genes

