

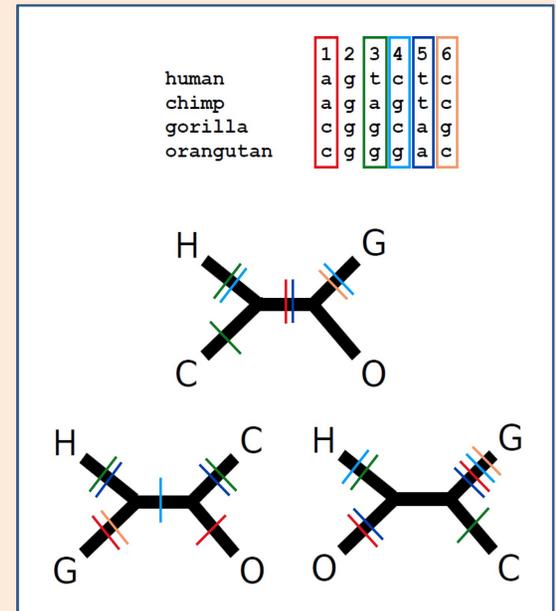
# Gene Ontology and Functional Enrichment

Genome 559: Introduction to Statistical and  
Computational Genomics

**Elhanan Borenstein**

# A quick review

- The parsimony principle:
  - Find the tree that requires the fewest evolutionary changes!
- A fundamentally different method:
  - Search rather than reconstruct
- Parsimony algorithm
  1. Construct all possible trees
  2. For each site in the alignment and for each tree count the minimal number of changes required
  3. Add sites to obtain the total number of changes required for each tree
  4. Pick the tree with the lowest score

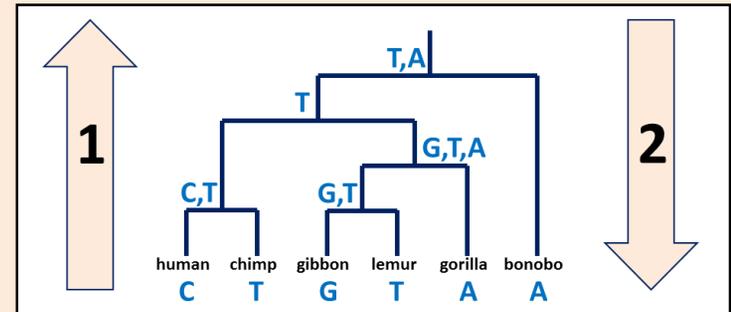


# A quick review – cont'

- Small vs. large parsimony

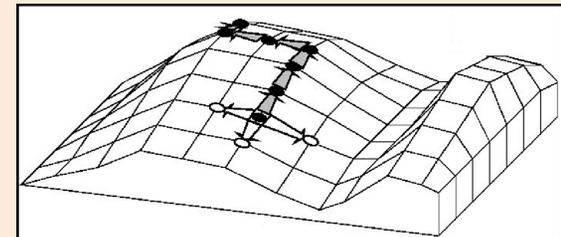
- Fitch's algorithm:

1. **Bottom-up phase:** Determine the set of possible states
2. **Top-down phase:** Pick a state for each internal node



- Searching the tree space:

- Exhaustive search, branch and bound
- Hill climbing with Nearest-Neighbor Interchange



- Branch confidence and bootstrap support

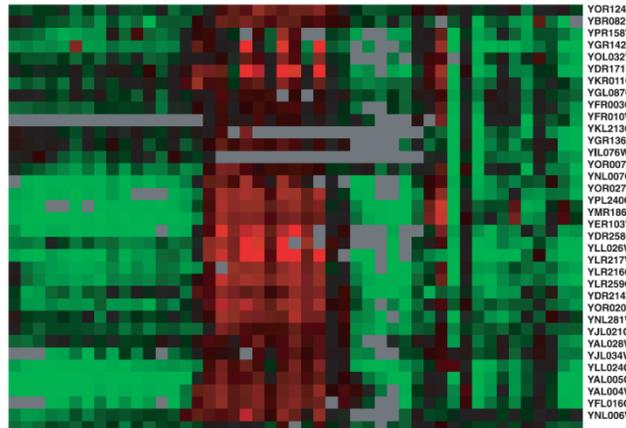
# From sequence to function

**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)

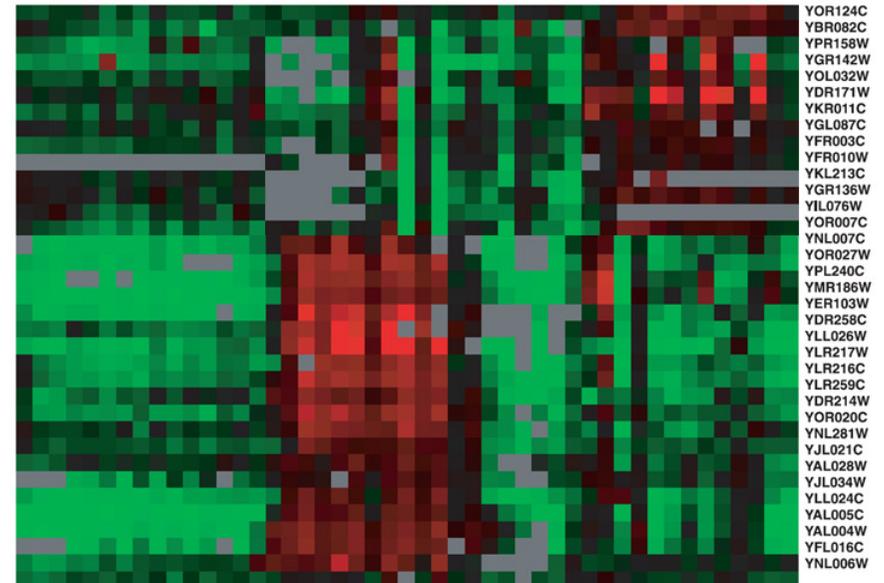


**Gene expression profiling**

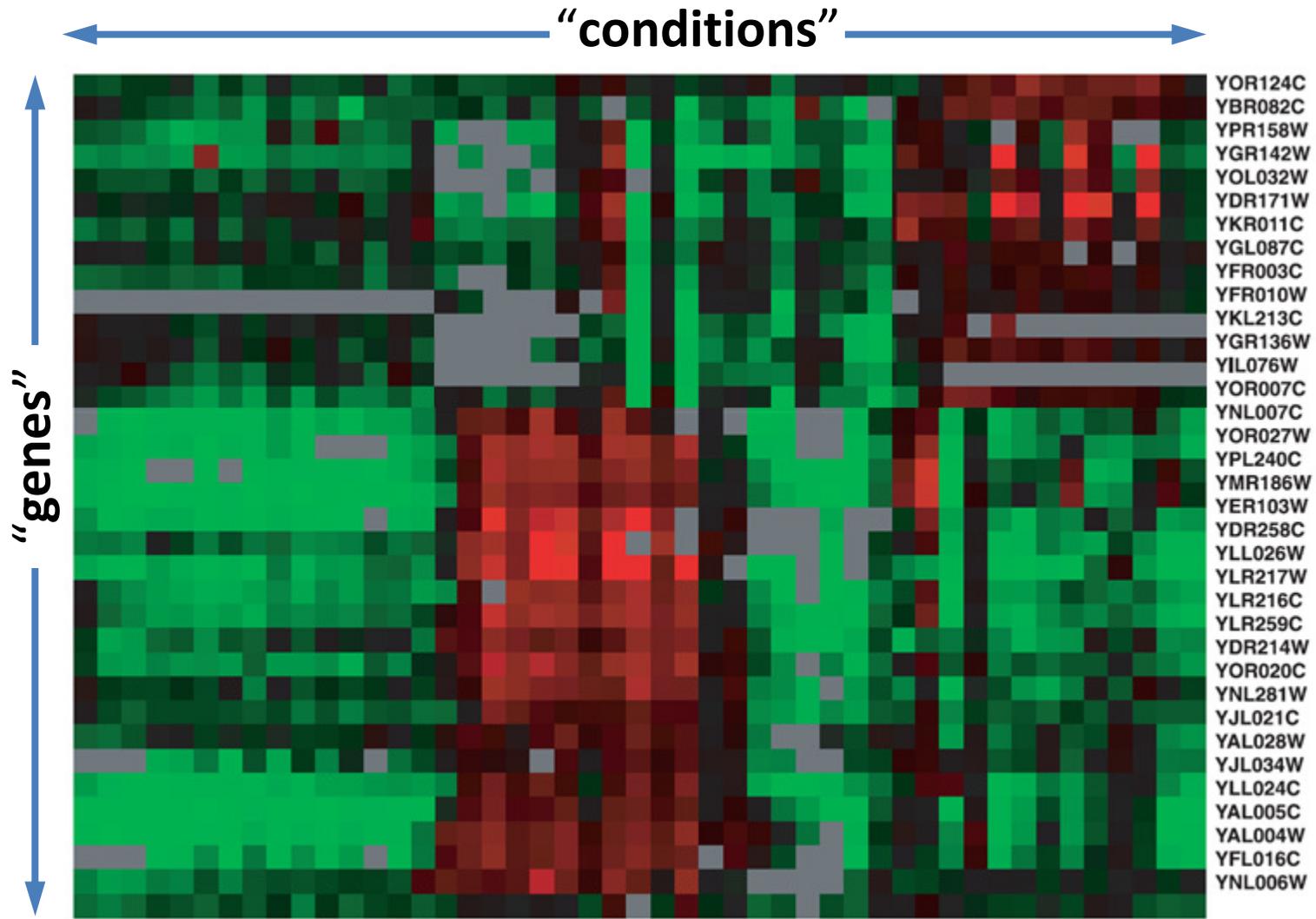


# Gene expression profiling

- Measuring gene expression:
  - (Northern blots and RT-qPCR)
  - **Microarray**
  - RNA-Seq
- Experimental conditions:
  - Disease vs. control
  - Across tissues
  - Across time
  - Across environments
  - Many more ...



# Different techniques, same structure



# 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*

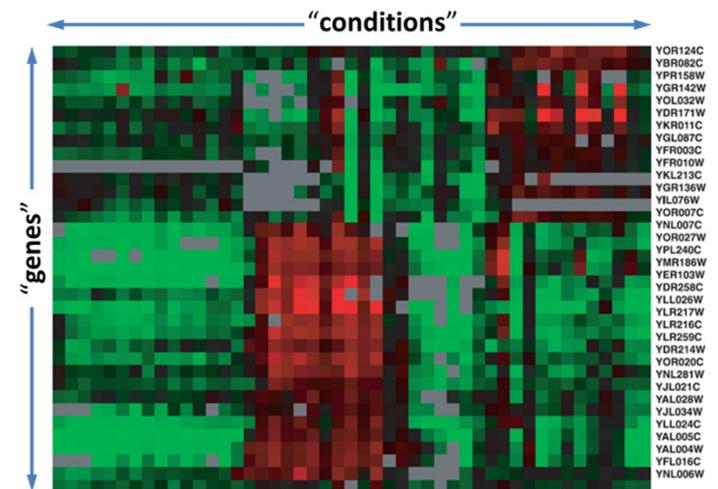
*Not systematic*

*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?

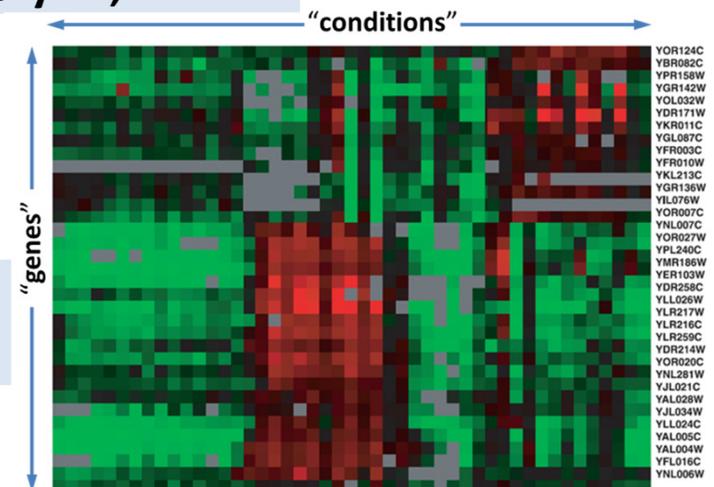
- A shared functional vocabulary
- Systematic linkage between genes and functions
- A way to identify genes relevant to the condition under study
- Statistical analysis (combining all of the above to identify cellular functions that contributed to the disease or condition under study)
- A way to identify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

Gene Ontology  
Annotation

Fold change,  
Ranking, ANOVA

Enrichment  
analysis, GSEA

Clustering,  
classification



# 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.

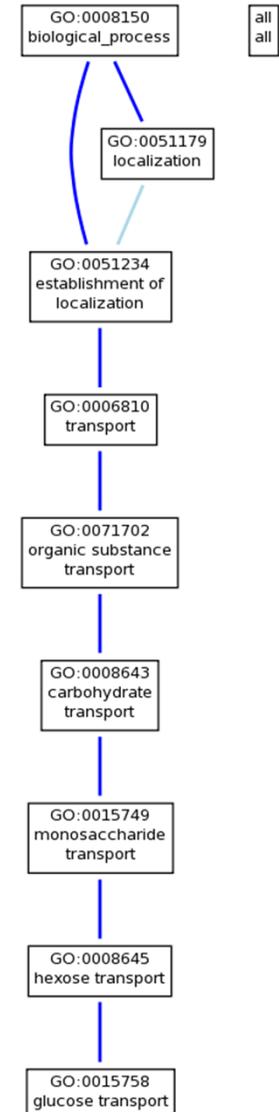
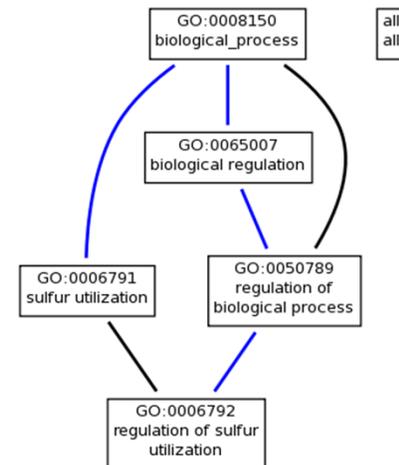
Term Information	
<b>Accession</b>	GO:0006348
<b>Ontology</b>	<b>Biological Process</b>
<b>Synonyms</b>	<b>exact:</b> heterochromatic silencing at telomere <b>exact:</b> telomere chromatin silencing <b>exact:</b> telomeric silencing
<b>Definition</b>	Repression of transcription of telomeric DNA by altering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

Back to top

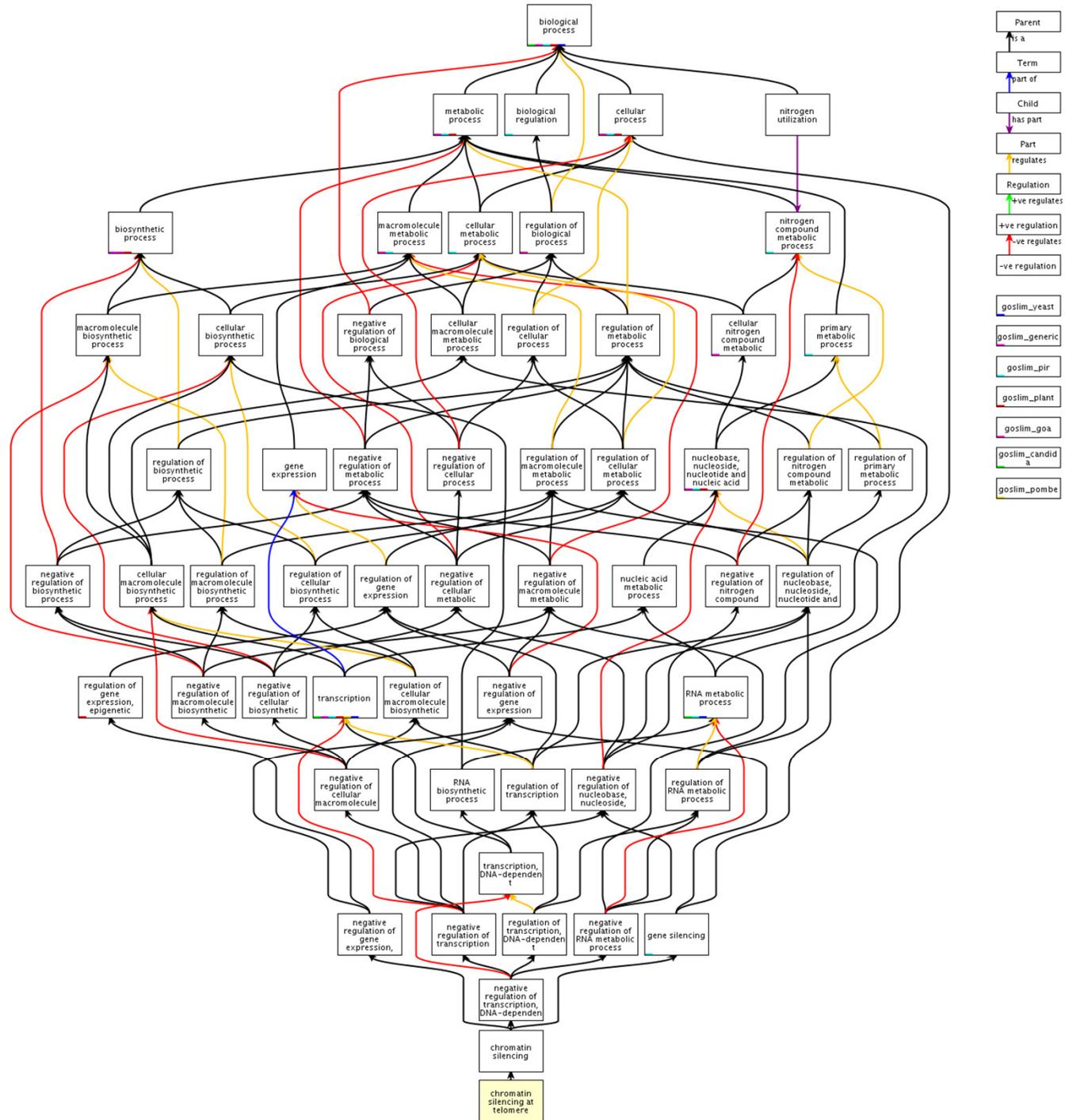
# Ontology structure

- GO also defines the **relationships** between the terms, making it a **structured** vocabulary.
- GO is structured as a **directed acyclic graph**, and each term has defined relationships to one or more other terms.

The screenshot displays the Gene Ontology (GO) web interface. At the top, there are filter options for 'Filter by ontology' (with a dropdown menu showing 'biological process', 'cellular component', and 'molecular function'), 'Filter Gene Product Counts' (with dropdowns for 'Data source' and 'Species'), and 'View Options' (with radio buttons for 'Tree view', 'Full', and 'Compact'). Below these filters, a tree view shows the hierarchy of gene products. The root is 'all : all [462390 gene products]'. It branches into 'GO:0008150 : biological\_process [358011 gene products]', 'GO:0051234 : establishment of localization [41128 gene products]', and 'GO:0006810 : transport [40715 gene products]'. The 'biological\_process' term further branches into 'GO:0071702 : organic substance transport [5258 gene products]', 'GO:0008643 : carbohydrate transport [1810 gene products]', 'GO:0015749 : monosaccharide transport [582 gene products]', and 'GO:0008645 : hexose transport [527 gene products]'. The 'establishment of localization' term branches into 'GO:0051179 : localization [46032 gene products]' and 'GO:0006810 : transport [40715 gene products]'. The 'transport' term branches into 'GO:0071702 : organic substance transport [5258 gene products]', 'GO:0008643 : carbohydrate transport [1810 gene products]', 'GO:0015749 : monosaccharide transport [582 gene products]', and 'GO:0008645 : hexose transport [527 gene products]'. The 'monosaccharide transport' term branches into 'GO:0015749 : monosaccharide transport [582 gene products]' and 'GO:0008645 : hexose transport [527 gene products]'. The 'hexose transport' term branches into 'GO:0015758 : glucose transport [417 gene products]'.







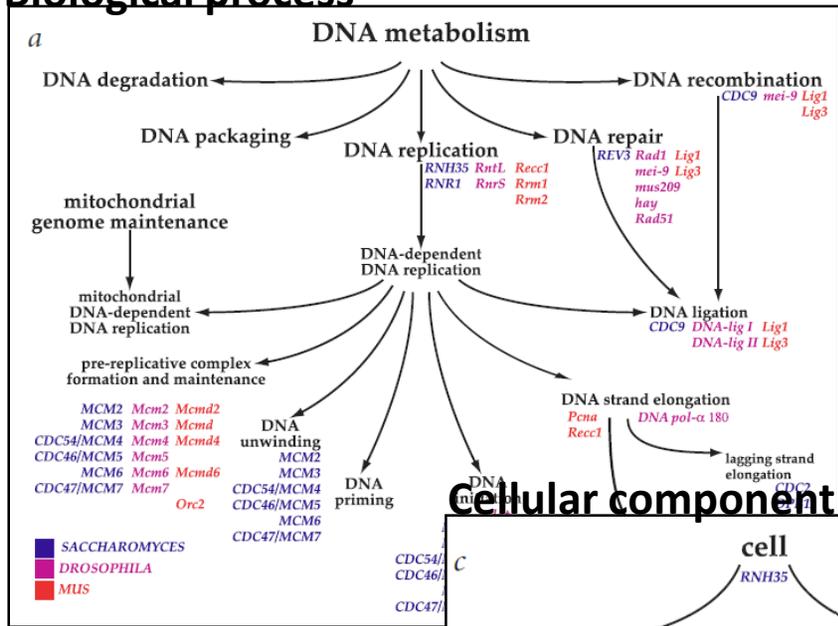
# GO domains

- Three ontology domains:
  - 1. Molecular function:** basic activity or task  
*e.g. catalytic activity, calcium ion binding*
  - 2. Biological process:** broad objective or goal  
*e.g. signal transduction, immune response*
  - 3. Cellular component:** location or complex  
*e.g. nucleus, mitochondrion*
- Genes can have multiple annotations:

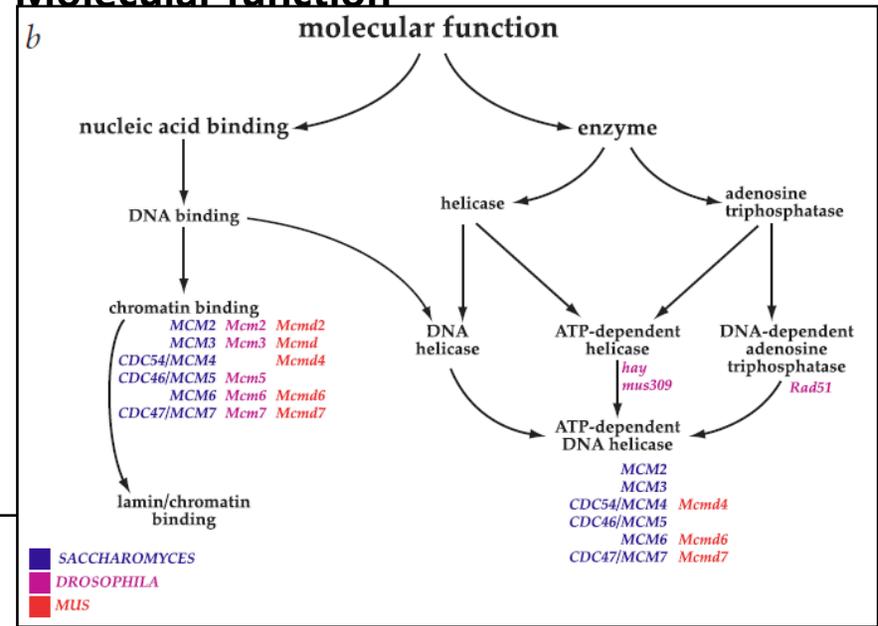
*For example, the gene product cytochrome c can be described by the molecular function term oxidoreductase activity, the biological process terms oxidative phosphorylation and induction of cell death, and the cellular component terms mitochondrial matrix and mitochondrial inner membrane.*

# Go domains

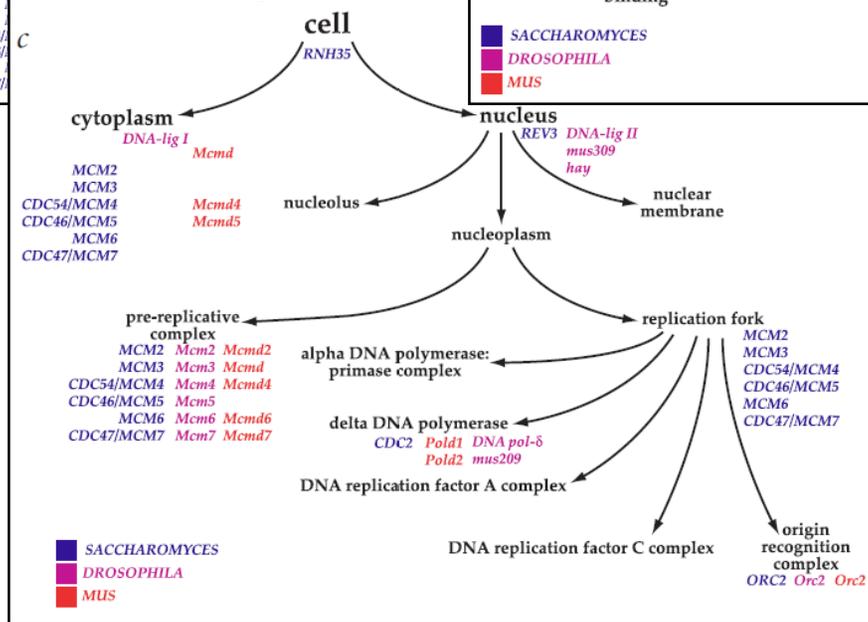
## Biological process



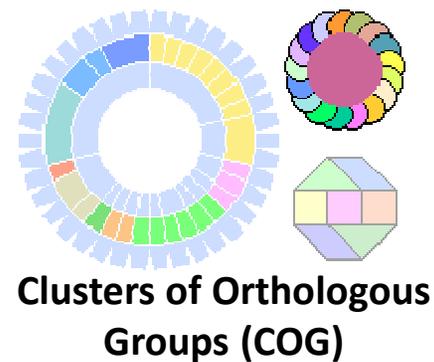
## Molecular function



## Cellular component



# Ontology and annotation databases



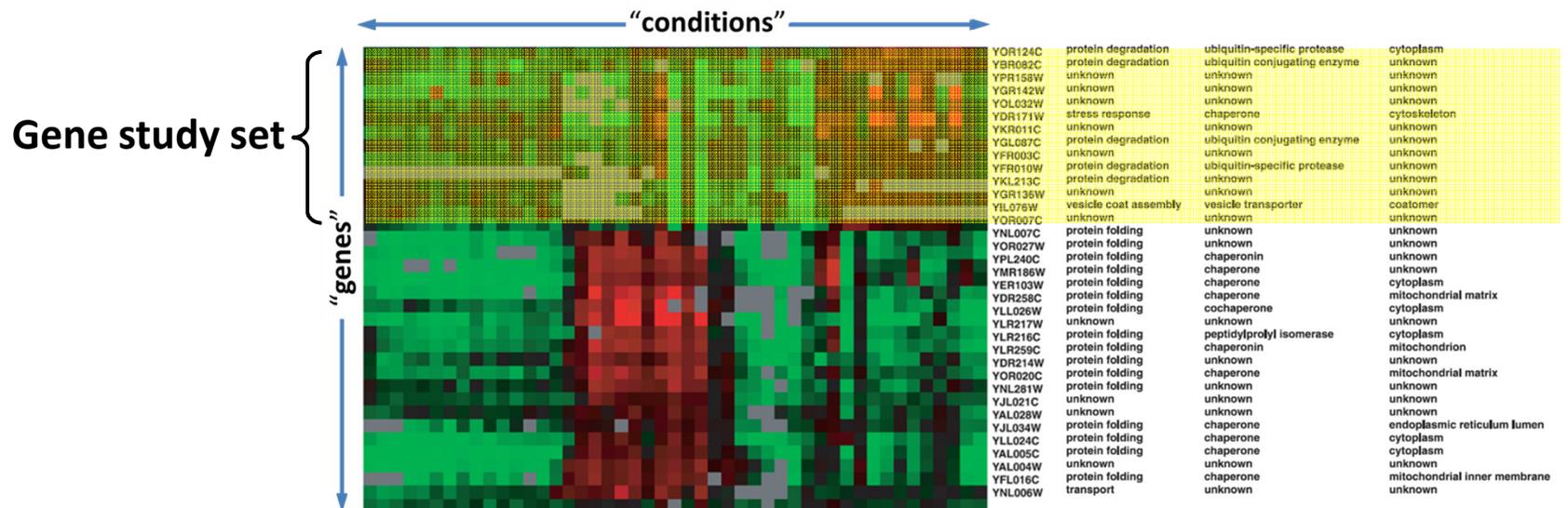
**“The nice thing about standards is that there are so many to choose from”**

Andrew S. Tanenbaum

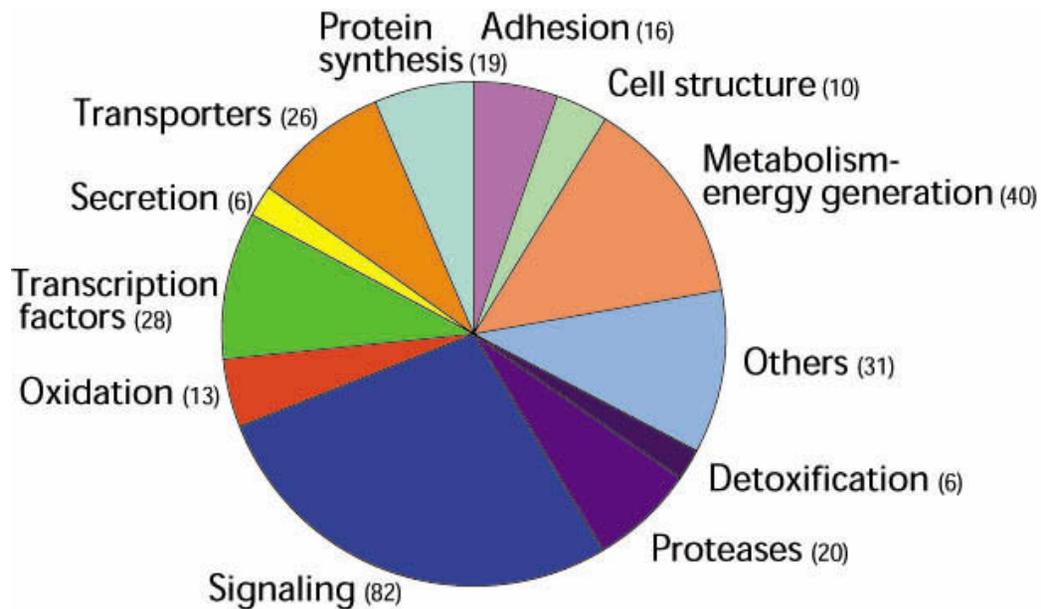
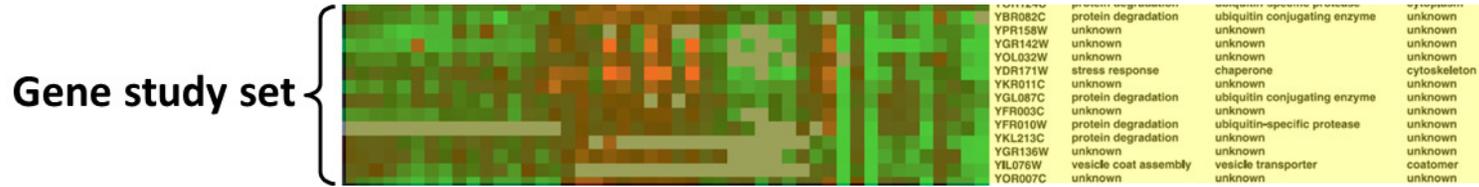


# Picking “relevant” genes

- 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)



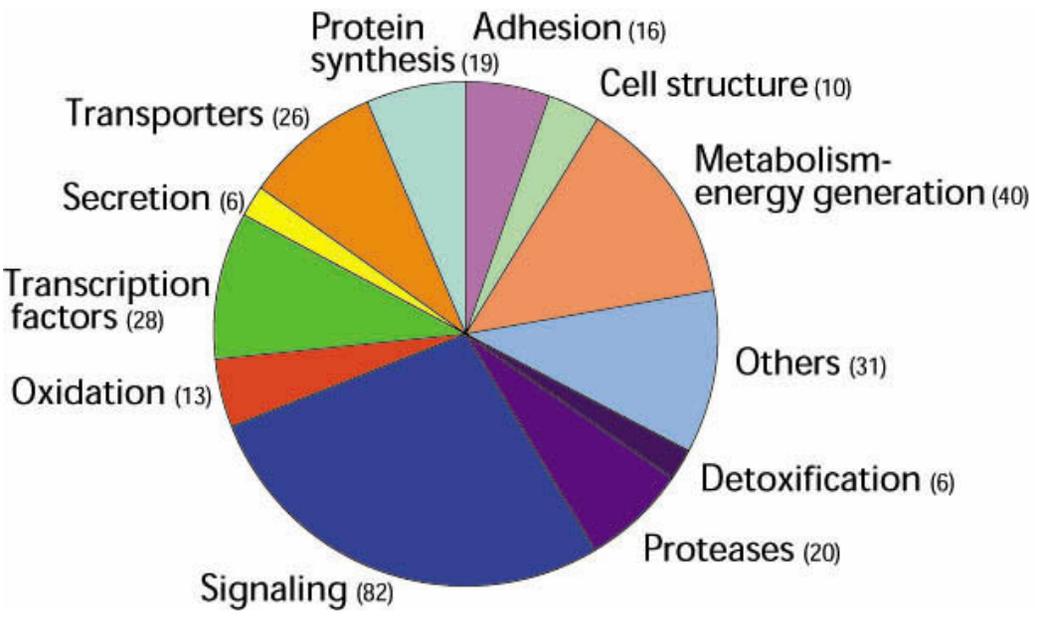
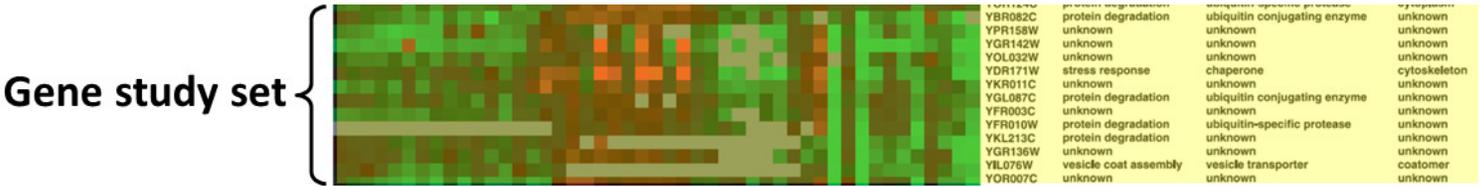
# Enrichment analysis



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

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

# 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%
<b>Transporters</b>	<b>26</b>	<b>8.8%</b>	<b>2%</b>
Proteases	20	6.7%	7%
Protein synthesis	19	6.4%	7%
Adhesion	16	5.4%	6%
Oxidation	13	4.4%	4%
<b>Cell structure</b>	<b>10</b>	<b>3.4%</b>	<b>8%</b>
Secretion	6	2.0%	2%
Detoxification	6	2.0%	2%

# Enrichment analysis – the right way

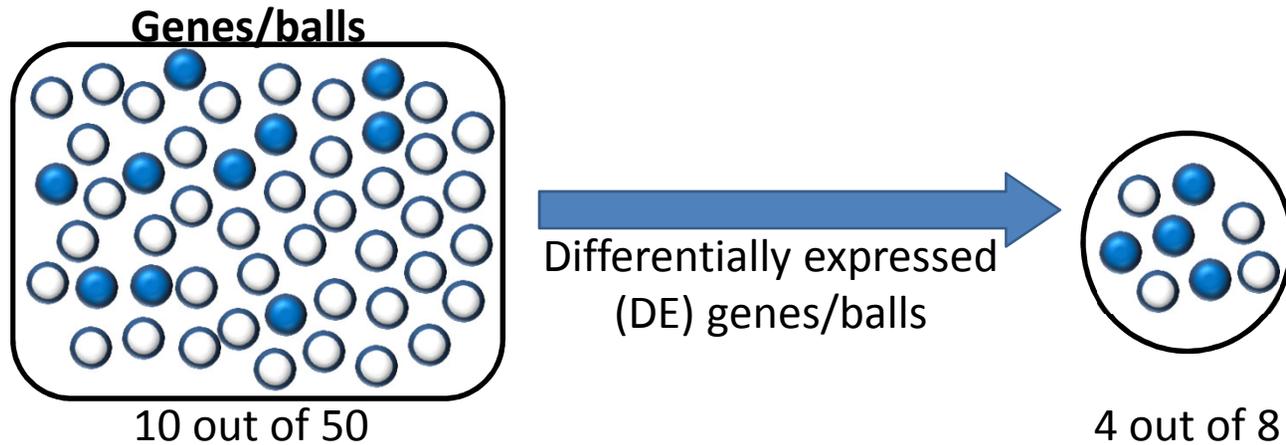
*Say, the microarray contains 50 genes, 10 of which are annotated as ‘signaling’. Your expression analysis reveals 8 differentially expressed genes, 4 of which are annotated as ‘signaling’.  
Is this significant?*

**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?*

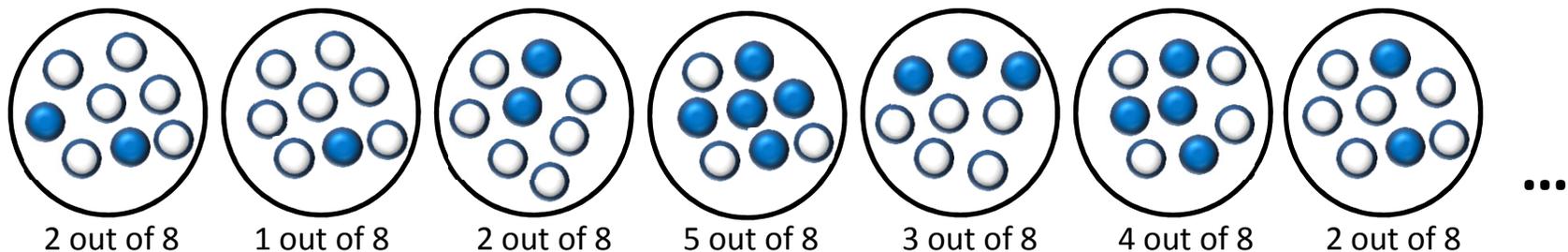
*The “urn” version: You pick a random set of 8 balls from an urn that contains 50 balls: 40 white and 10 blue. How surprised will you be to find that 4 of the balls you picked are blue?*

# A quick review: Modified Fisher's exact test



Do I have a surprisingly high number of blue genes?

Null model: the 8 genes/balls are selected randomly



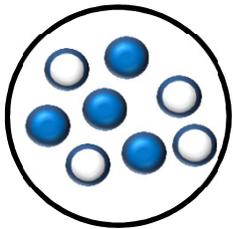
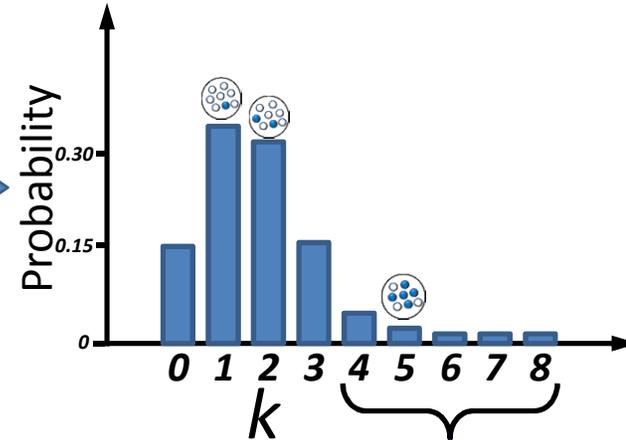
So, if you have 50 balls, 10 of them are blue, and you pick 8 balls randomly, what is the probability that  $k$  of them are blue?

# A quick review: Modified Fisher's exact test

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?

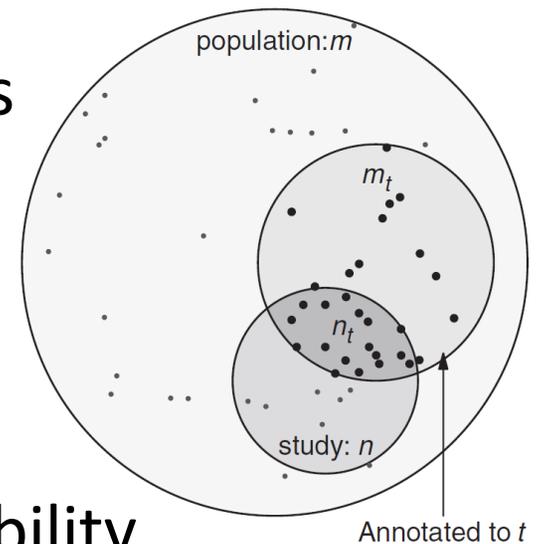
What is the probability of getting at least 4 blue genes in the null model?



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

# 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!

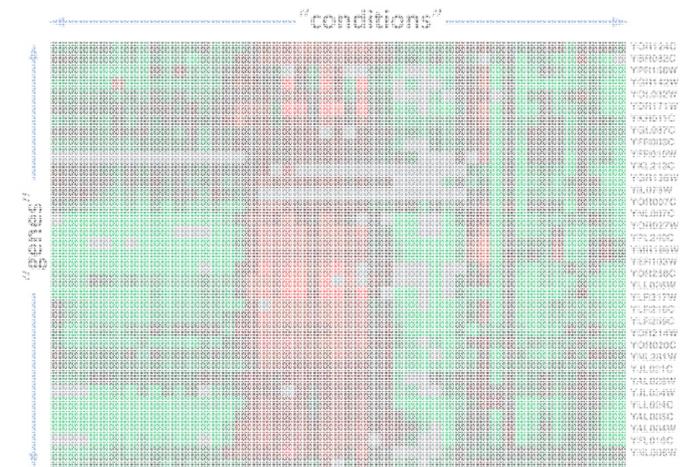


$$\mathbb{P}(\sigma_t \geq n_t) = \sum_{k=n_t}^{\min(m_t, n)} \frac{\binom{m_t}{k} \binom{m-m_t}{n-k}}{\binom{m}{n}}$$

(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!

Ignores links between GO categories

- **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

