

# Introduction to Gene Set Enrichment

# Lecture Outline

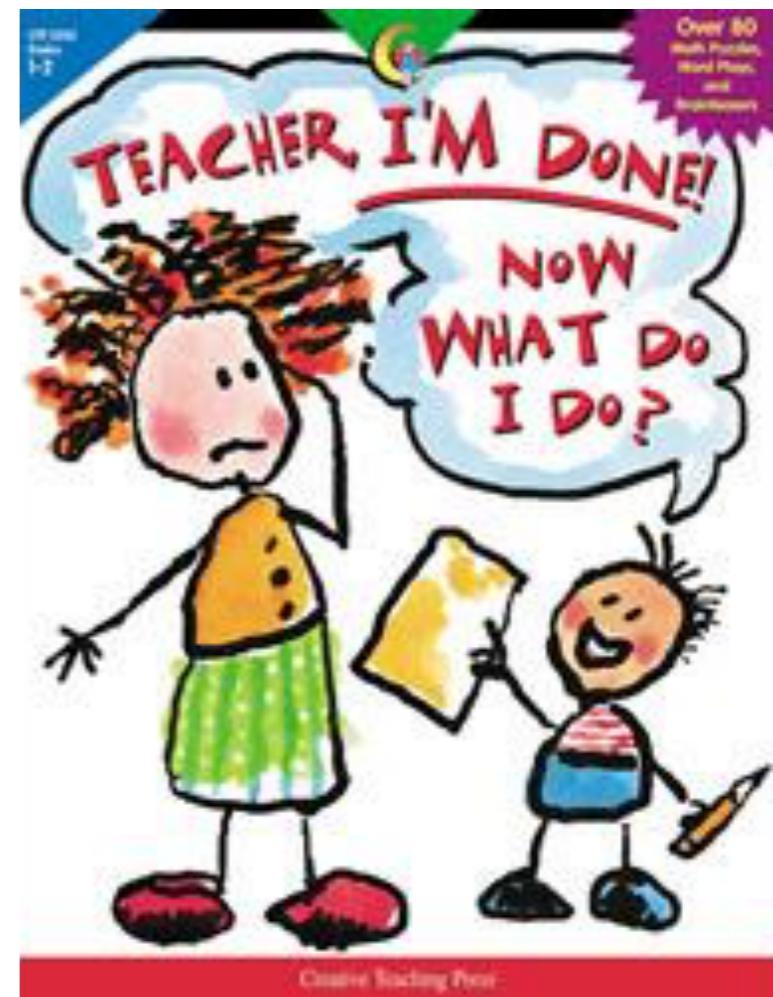
- Some theory:
  - How to analyse large gene lists.
  - The hypergeometric test and Fisher's Exact test
  - Gene set annotations.
  - Some online/R tools
- Exercise:
  - Enrichment of GO-term in FANTOM5 tissue specific genes
  - Some more advanced theory.



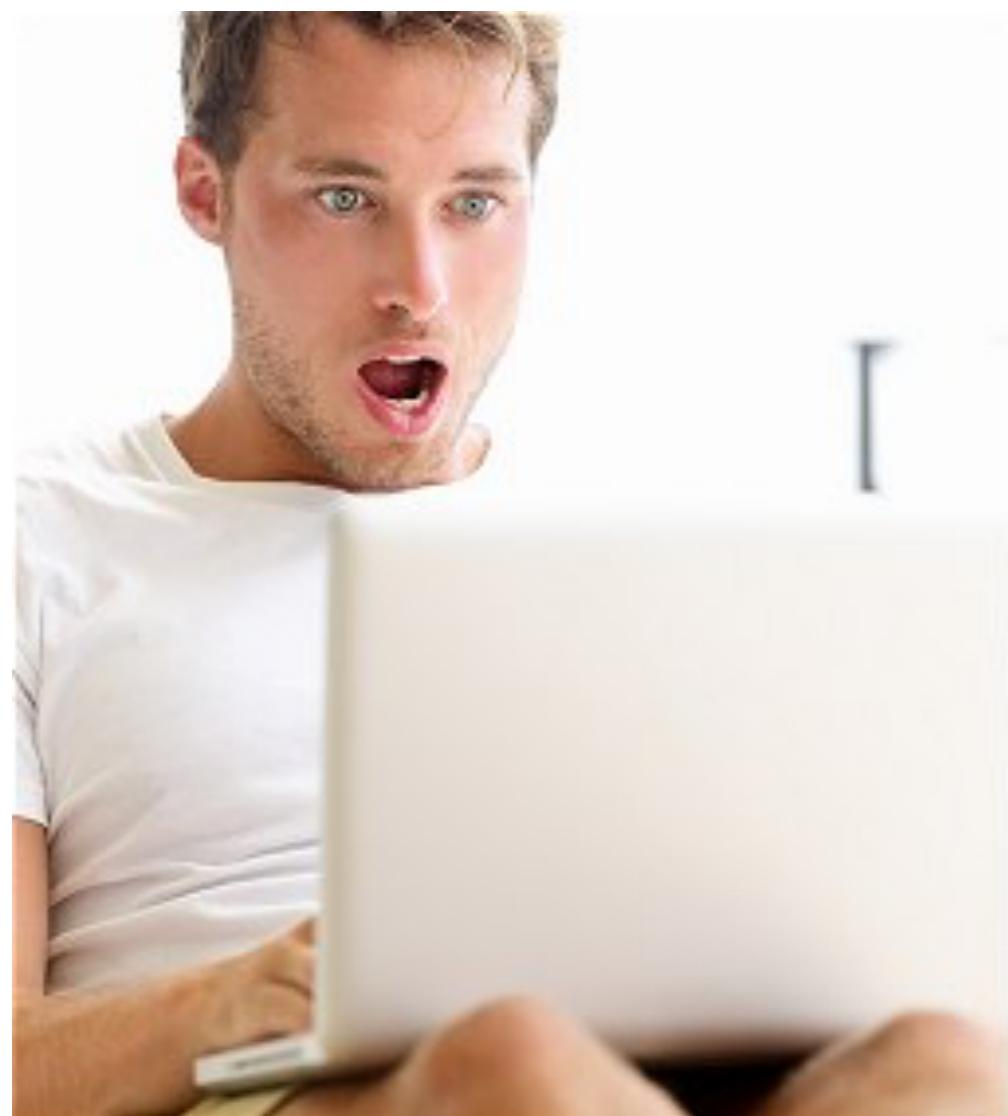
# Interpretation of large gene lists

# What to do when you are done?

- This lecture starts when all of the previous lectures have ended:
  - You have performed a high-throughput experiment (RNA-Seq, CAGE, microarray, etc.)
  - You have QC'ed and trimmed your data.
  - You have explored the data (for example using PCA and clustering)
  - You have performed a valid statistical test to obtain lists of “interesting” genes.
  - ... *What do you do now?*



# What you SHOULD NOT do



- It is common to see people to this:
  - *“I looked at my list of 924 genes and saw three genes from my favourite gene family (that I’m doing my thesis on). I therefore conclude that genes from my favourite gene family are the central regulators of the biological response, just as I suspected prior to the experiment.”*
  - What is wrong with this approach?

# It is extremely easy to over-interpret data such as this!

- There are several problems with this approach:
  - If your list of interesting is long enough, you are bound to choose at least some of your favourite genes by chance.
  - If the favourite gene family is big, the chance of finding these by chance is very big.
  - It is very biased to only look at a single gene family - what about categories of genes?
  - It is important not to be biased towards what you expect to see and accept the fact that the experiment might not show what you want/expect it to show.
  - It is easy to tell “stories” - but are they actually true (/ statistically valid)?
- In this lecture we will talk about how to analyse large gene lists, while keeping these important points in mind.



"Ralph's over-interpreting the data again."

# Gene set annotations

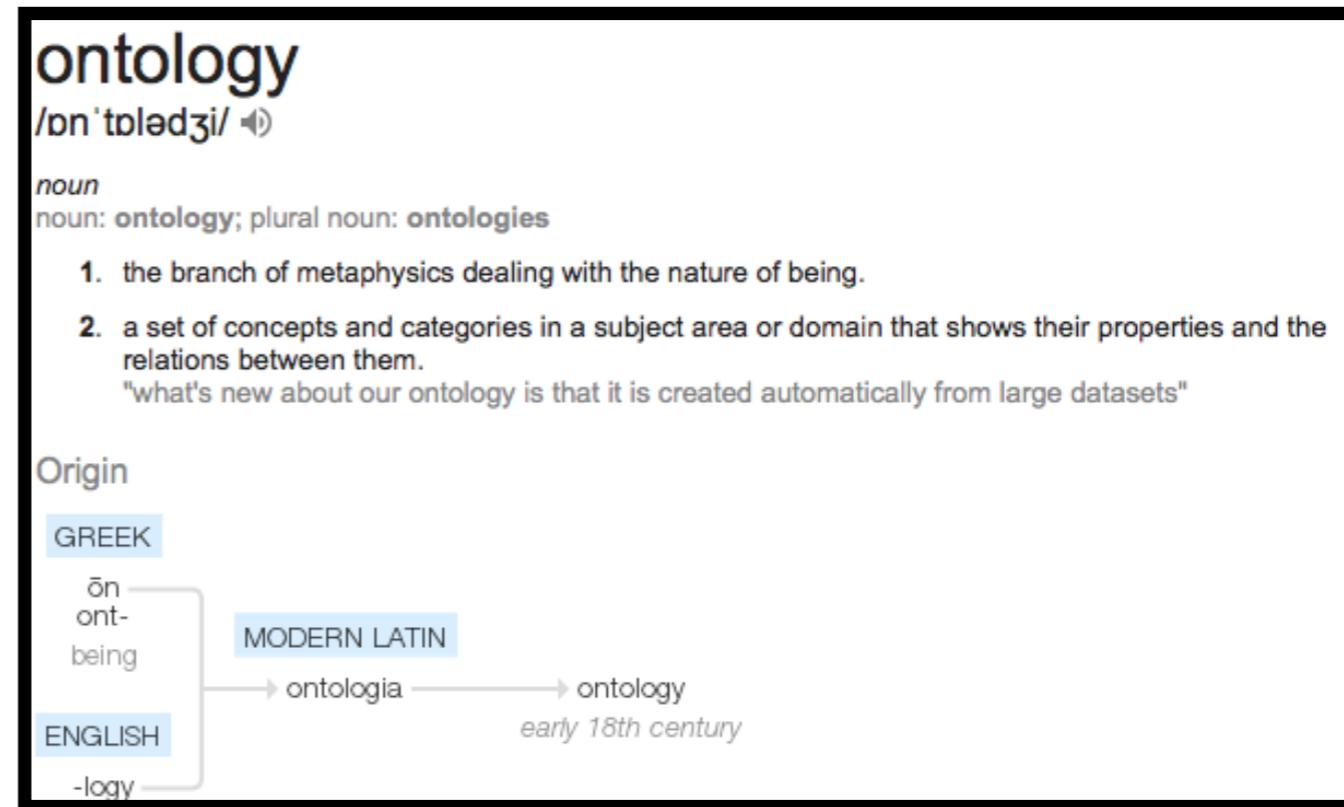
# Where to obtain annotations

- The first step in analysing gene tests is to obtain some information about the genes. For example looking at:
  - Gene families
  - Protein domains
  - KEGG pathways
  - Reactome
  - MySigDb
  - Disease-associated genes
  - Gene Ontology (GO) terms
- GO-terms are by far the most widely used, so we will focus at them in this lecture.

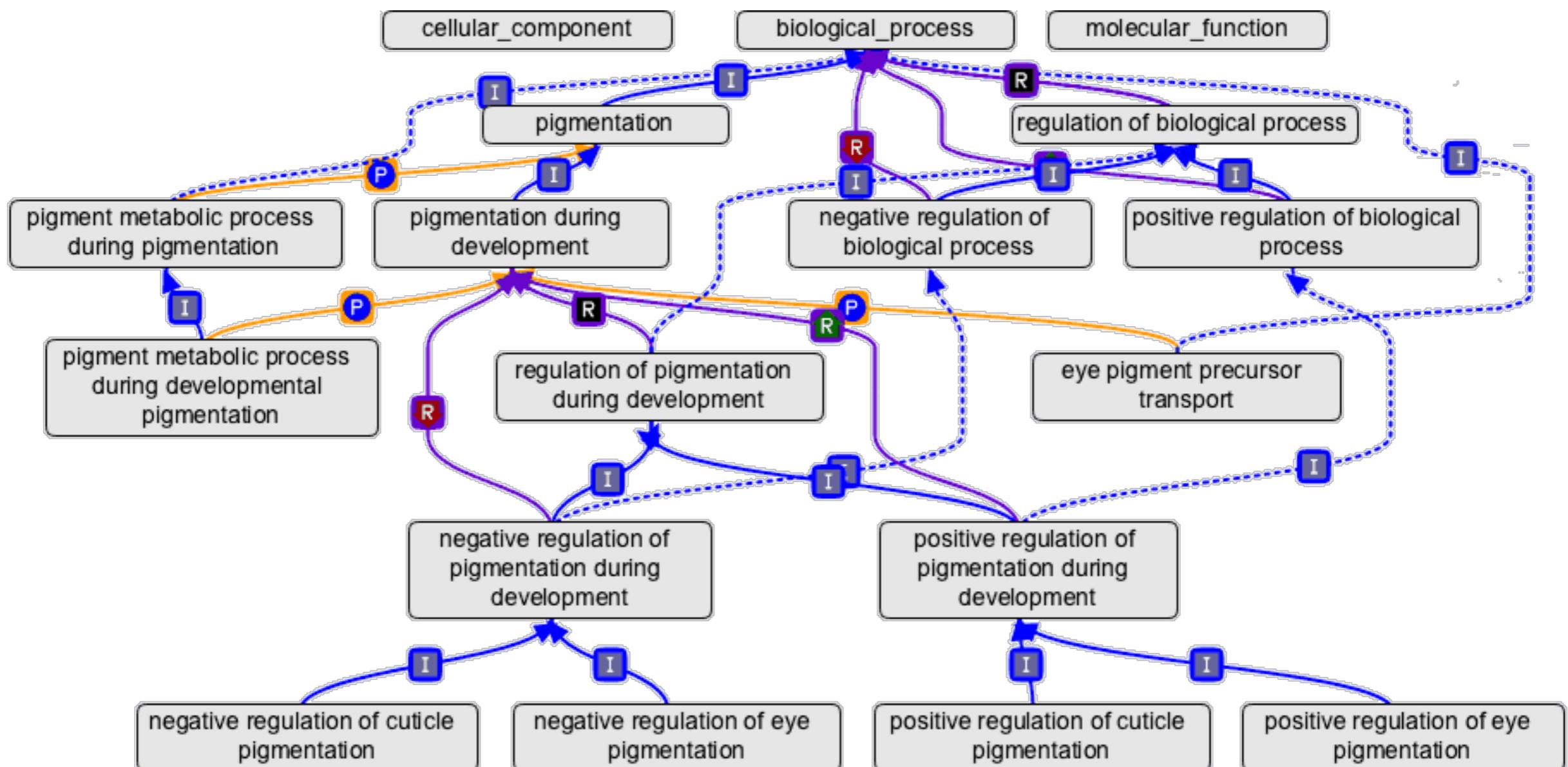


# GO-terms in a single slide

- Online database:
  - Gene Ontology is a collection of controlled vocabularies describing the biology of a gene product in any organism.
  - Includes both manually curated annotations, as well as electronically inferred.
  - Continuously updated.
- Three domains:
  - Cellular Component (CC)
  - Molecular Function (MF)
  - Biological Process (BP)



# GO-term hierarchy



# Exercise: 5 minutes

- Go to the GO-term consortium website: <http://geneontology.org/> and look up a gene:
  - Left half of class: **DEFA5**
  - Right half of class: **MUC5AC**
- What do the genes do? What GO-terms do they have associated?

# Hypergeometric testing

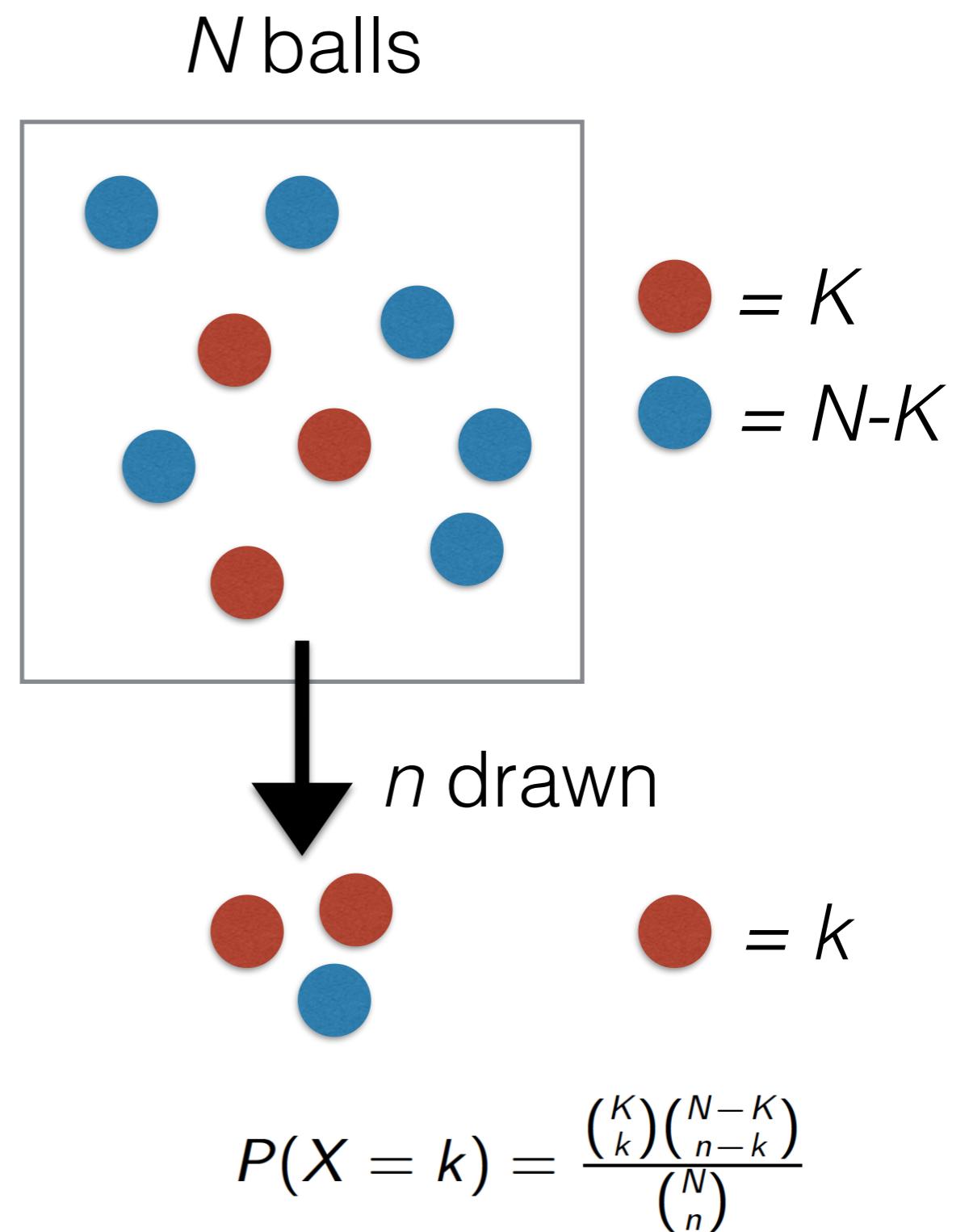
have whole DB with pool of GO terms ---> how probable by chance for our genes set to have that specific subset of GO pool (similar to how probable to draw 3 mirror force in first turn from hand of 4 cards)

# Intuition

- The question we want to ask is:
  - Are there GO-terms that appear more often in my list of interesting genes than what would be expected by chance (If genes were picked at random)?
  - If you remember back to your course(s) in statistics, this is a problem of *sampling without replacement*, which can be described with a *hypergeometric distribution*.

# The hypergeometric distribution

- Consider the following example:
  - You have a box containing  $N$  balls, of which  $K$  are red and  $N - K$  are blue
  - You draw  $n$  balls from the box without putting balls back after each draw
  - The hypergeometric distribution describes the probability of drawing  $k$  red balls.



# R's hyper

R has excellent built in functions for working with a wide range of probability distributions, including the hypergeometric:

```
# Setting all the variables with some random numbers
N <- 100 # The total number of balls
K <- 25 # The number of red balls, aka the "successes"
n <- 10 # The number of draws from the box, aka. the "trials"
k <- 4 # The number of observed red balls when drawing
```

```
# Using the hypergeometric distribution:
phyper(q=k-1, # We want equal or bigger
       m=K,
       n=N-K,
       k=n,
       lower.tail=FALSE) # P[X > x]
## [1] 0.2146156
```

This means that the chance of drawing 4 (out of 10) or more red balls from a box of 25 (out of 100) is 0.21.

# Exercise: 5 minutes

For the following three cases, calculate which was the least likely outcome under random sampling without replacement:

$N=100, K=25, n=20, k=15$

$N=100, K=25, n=30, k=20$

$N=100, K=25, n=2, k=2$

# Hypergeometric test

- We can see this probability as a p-value for the following hypotheses:
  - Null hypothesis ( $H_0$ ): We are sampling red balls at random
  - Alternative hypothesis ( $H_1$ ): We are sampling *more* red balls than at random.
- Note, that since we have been using the upper tail of the distribution, we are in effect performing a one-sided test. This is often called a test for *enrichment*. Similarly, we could have used the lower tail of the distribution to perform a test for *depletion*.

# How would this work with genes?

The above example can be easily translated to genes:

```
# Continuing the example from previously:  
N <- 10000 # The total number of genes  
K <- 50 # The number of gene belonging to a gene family  
n <- 942 # The list of interesting genes  
k <- 3 # The number of gene family members in the interesting genes
```

```
# Using the hypergeometric distribution:  
phyper(q=k-1, m=K, n=N-K, k=n, lower.tail=FALSE) # P[X > x]  
## [1] 0.8624552
```

Not very significant!

# Fisher's Exact Test

A popular way of doing a hypergeometric test is to do a Fisher's Exact test. This tests starts from a 2-by-2 contingency table:

```
# Continuing the example from previous:  
N <- 10000 # The total number of genes  
K <- 50 # The number of gene belonging to a gene family  
n <- 942 # The list of interesting genes  
k <- 3 # The number of gene family members in the interesting genes
```

```
# Rearrange the data as a 2x2 table  
m <- matrix(c(k, K - k, n - k, N - K - n + k),  
            2, 2, dimnames = list(c("In List", "Rest"),  
                                  c("In Family", "Rest"))))
```

```
m  
##          In Family Rest  
## In List           3  939  
## Rest            47 9011
```

# R's fisher.test

Fisher's exact test is included in base R:

```
res <- fisher.test(x=m, alternative="greater")

res
##
## Fisher's Exact Test for Count Data
##
## data: m
## p-value = 0.8625
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
## 0.1612245 Inf
## sample estimates:
## odds ratio
## 0.6125624

res$p.value
## [1] 0.8624552
```

# Hypergeometric Test vs Fisher's Exact Test

- What test to use?
  - For single-sided tests, gives identical results.
  - `fisher.test` allows for easier calculation of two-sided p-values.
  - `fisher.test` also calculates the odds ratio.
  - It is very simple to make 2-by-2 contingency tables in R using the `table()`-function, making it easy to apply `fisher.test` to big datasets.

# Online tools

# Tools for GO-term enrichment

- While the underlying statistics for GO-term enrichment is rather simple, implementing it in code is a bit tricky.
- There are some issues with matching gene names across different databases.
- Using an online tool can solve both of these issues, at the cost of resorting to “copy-paste” analysis, prone to manual errors.



# Popular tools

- Some popular (stand-alone) online tools are:
  - DAVID: <https://david.ncifcrf.gov/>
  - PANTHER: <http://www.pantherdb.org/>
  - GOrilla: <http://cbl-gorilla.cs.technion.ac.il/>
  - gProfiler: <http://biit.cs.ut.ee/gprofiler/> ==> My favourite!
  - ... and many more!
- Most of these tools will, in addition to the actual enrichment analysis also help with annotating genes with GO-terms.
- **Small warning:** Be wary of online tools that are no longer kept updated. For example, the popular GOstats tools (<http://gostats.com/>) is actually no longer being updated

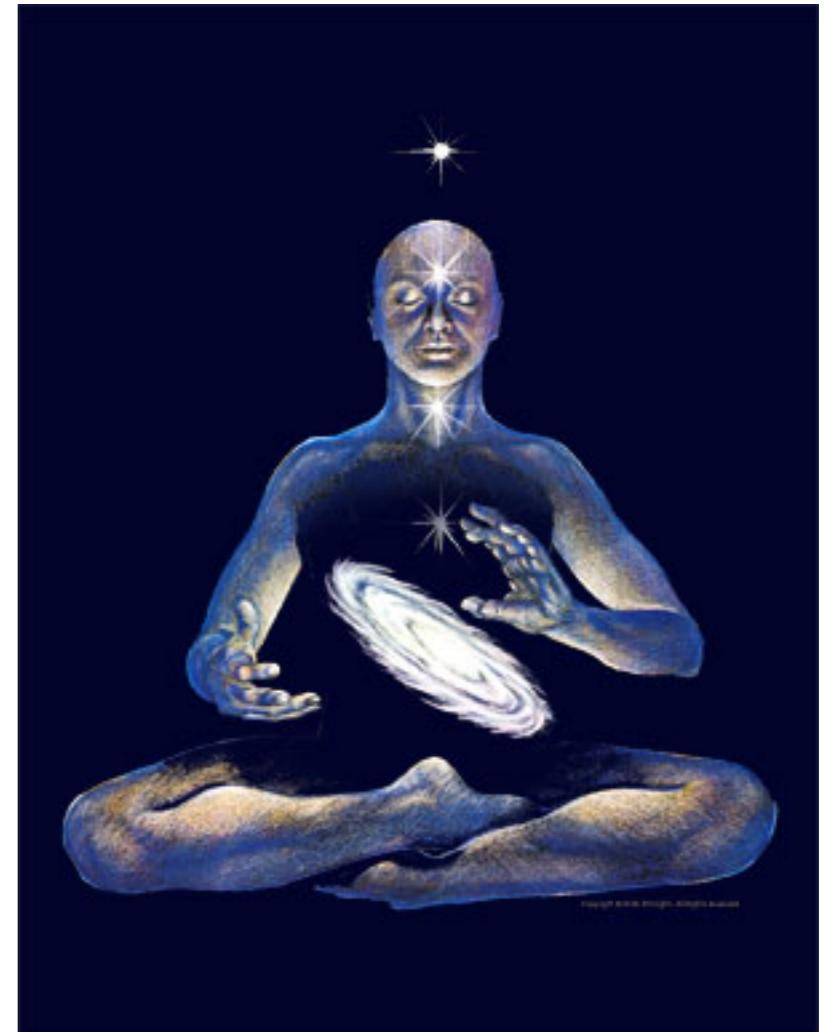


# What should be uploaded to an online tool?

- List of interesting genes or *foreground set*, all with the same id-type (i.e. gene symbol, Entrez ID, etc.)
- A *background set* or *universe*.
- **WARNING:** Many tools will run without a universe input, and instead use the all genes in the genome as background - why is this often a bad idea?

sth to compare against that isn't whole genome bc the context

eg. RNA from some brain tiss should use genes express in brain tiss as background not whole genome



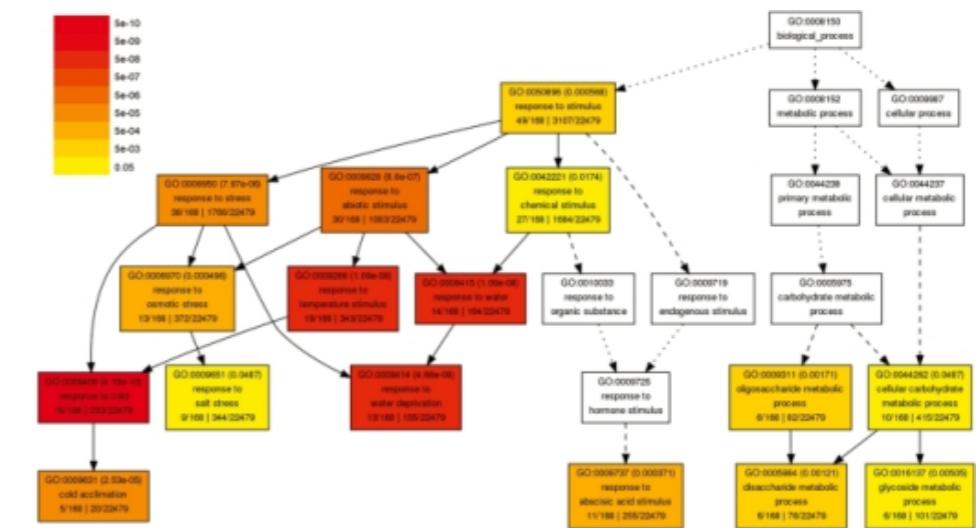
FANTOM5 tissue specific  
genes - a GO exercise  
See course web page

# Take home messages

- Converting between IDs can be very difficult and can affect results.
- p-values depends a lot on sample sizes.
- GO-enrichments are often ambiguous!
- Be very careful about overinterpretation!

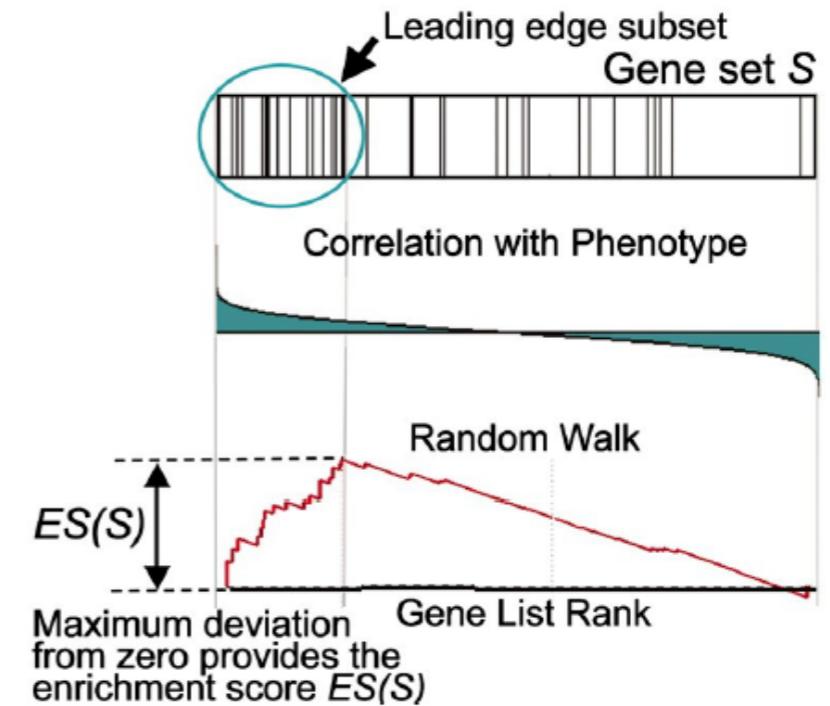
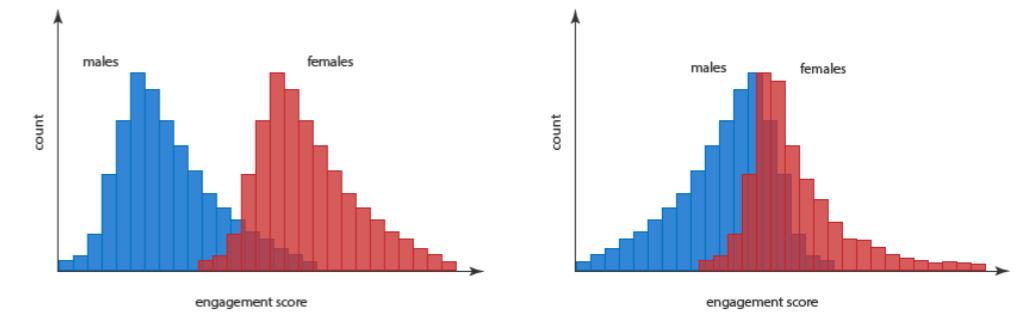
# Extensions of the hypergeometric test

- Use a “biased urn” model:
  - Some genes have a higher chance of being samples
  - Tools: goSeq, limma’s goana
- Use a more sophisticated FDR correction:
  - GO-term enrichments p-values are not really independent, since the same gene can be in multiple of the same groups.
  - Take the GO-term hierarchy into accounts when accounting for multiple testing



# Threshold-free approaches

- In many cases the “interesting vs rest” is artificial, (i.e. unregulated vs non-regulated).
- Real signal is smooth across genes - i.e. some genes are VERY upregulated, some only a little.
- Calculate some statistics for each gene (logFC, t-statistics, etc), and perform a test (i.e. Wilcox or t-test) for difference between genes in and outside of a geneset.
- Some statistical issues arise when there is intergene correlation within gene sets. This can be corrected using Variance Inflation Factors (i.e. limma’s CAMERA)
- The popular GSEA method uses a running-sum maximum instead of simple test.



Subramanian et al., 2005, fig 1.

# More detailed statistics

- Goeman *et al* 2008 (<http://www.ncbi.nlm.nih.gov/pubmed/17303618>) introduced the statistical foundation for gene set testing:
  - **Competitive tests:** Is a gene set enriched relative to other genes?
    - *Permutation over genes*
    - Hypergeometric and friends + Foreground/background methods
    - Issues with intergene correlations
  - **Self-contained tests:** Is a gene set associated with the outcome in an absolute sense (without looking at other genes)?
    - *Permutation over samples.*
    - More advanced statistics: globalTest and limma's ROAST.
    - Preserves intergene correlation, but suffers from issues with low sample size.