Lecture 16

Gene Ontology Analysis

MCB 416A/516A
Statistical Bioinformatics and Genomic Analysis

Prof. Lingling An
Univ of Arizona

Outline

- Introduction to Gene Ontology
- Gene set enrichment
- R code

After detect significant genes in differential expression analysis…

Data on *n* genes for *m* samples results in a
nxm gene-by-sample data matrix

	sample1	sample2	sample3	sample4	sample5 …
Gene1	0.46	0.30	0.80	1.51	0.90 ...
Gene2	-0.10	0.49	0.24	0.06	0.46 ...
Gene3	0.15	0.74	0.04	0.10	0.20 ...
Gene4	-0.45	-1.03	-0.79	-0.56	-0.32 ...
Gene5	-0.06	1.06	1.35	1.09	-1.09 ...
…	…	…	…	…	… ...

Preprocessing->normalization->summarization->testing=>

List of differentially expressed genes

How to interprete the data?

- Driven by experimental questions, but with a long list of siginficant genes
  - — which genes are of interest?
  - — what's special about the differentially expressed genes?
- Solution: pooling of genes into functional classes
  - — provides a general overview
- Gene Ontology database provides such a functional classification

After clustering genes, then what?

- Assign (or hypothesize about) biological meanings to clusters
  - Identify over-represented functional categories in the clusters (i.e., cluster may contain more genes known to be involved in a specific biological process than would be expected by chance)
  - If you find a gene of unknown function in a cluster of genes in which a known function is overrepresented, your gene of unknown function may have the same or a related function!
- Requirements for systematic analysis:
  - Standard assignment of genes into functional categories
  - Controlled vocabulary for describing biological meanings
    - Gene Ontology or GO project at NCBI

Gene Ontology

Ontology: A structured vocabulary

- describes concepts that exist in an area of knowledge
- describes relationships that exist between them

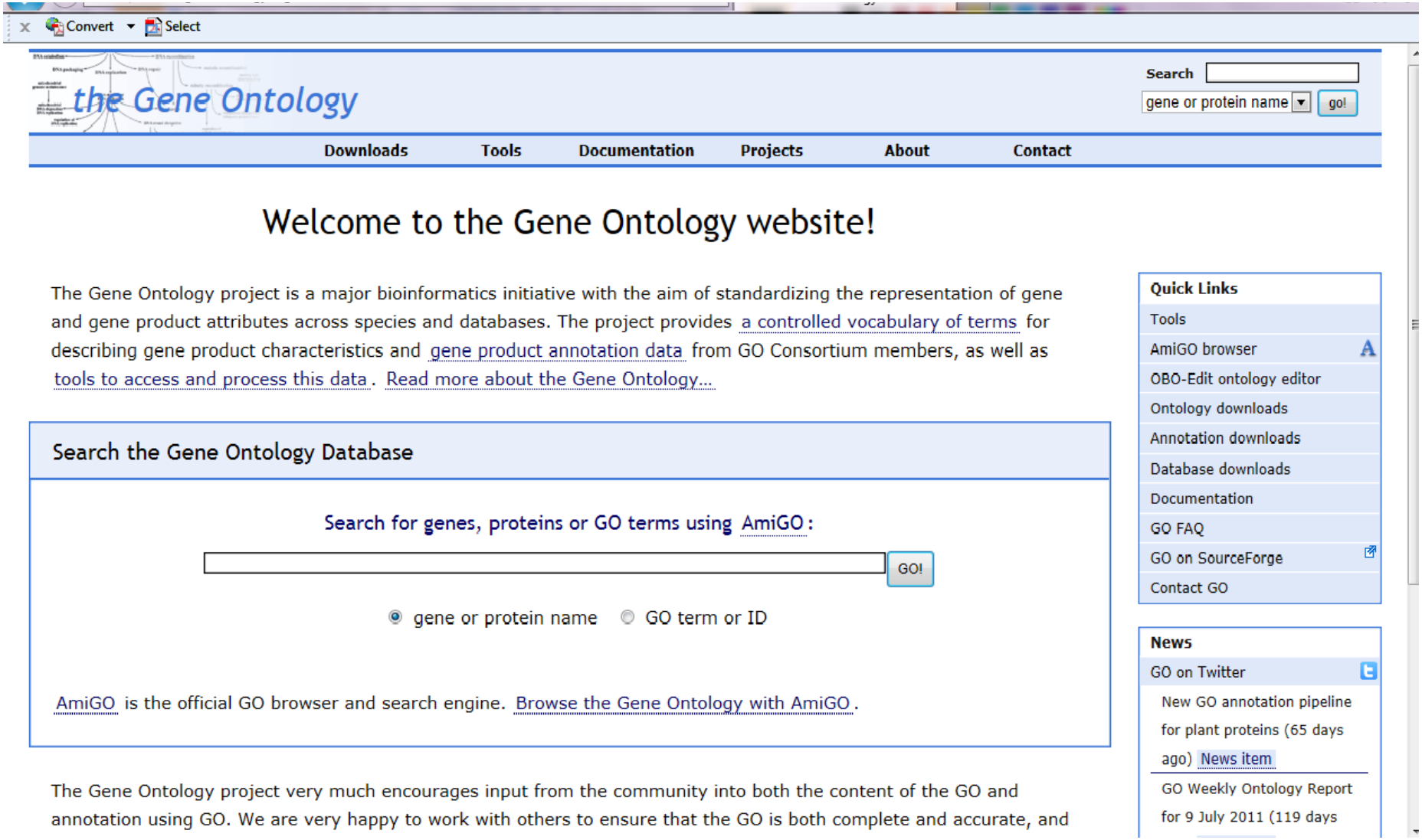
Gene Ontology (GO):

- describes possible functions of genes
- describes relationships between genes

GO is independent of the organism

- some functions are common to many organisms

GO is a public resource



Gene Ontology (GO) project

http://www.geneontology.org/

- Purpose:

  1) Establish a unified framework for organism-independent gene annotation

  2) Define controlled terms (ontologies) for description of gene products from 3 aspects:

     - Biological process (DNA repair, mitosis)
     - Molecular function (protein serine/threonine kinase activity, transcription factor activity)
     - Cellular component (nucleus, ribosome)

- Characteristics:

  1) A gene can have multiple associations in each ontology

  2) GO terms are organized in hierarchical structures called directed acyclic graphs (DAGs)

     - The most general classifications are at top levels of the graph

     - More specialized classifications at lower levels


all : all (183091)
GO:0008150 : biological_process (116737)
GO:0007610 : behavior (1929)
GO:0000004 : biological_process unknown (30095)
GO:0009987 : cellular process (72817)
GO:0007154 : cell communication (13030)
GO:0030154 : cell differentiation (2601)
GO:0050875 : cellular physiological process (64591)
GO:0006944 : membrane fusion (247)
GO:0050794 : regulation of cellular process (3758)
GO:0048523 : negative regulation of cellular process (1520)
GO:0048522 : positive regulation of cellular process (1483)
GO:0030155 : regulation of cell adhesion (116)
GO:0045595 : regulation of cell differentiation (497)
GO:0001558 : regulation of cell growth (305)
GO:0051244 : regulation of cellular physiological process (2793)
GO:0043012 : regulation of fusion of sperm to egg plasma membrane (8)
GO:0009966 : regulation of signal transduction (773)
GO:0007275 : development (15345)
GO:0007582 : physiological process (76830)
GO:0050789 : regulation of biological process (14938)
GO:0050794 : regulation of cellular process (3758)
GO:0048523 : negative regulation of cellular process (1520)
GO:0048522 : positive regulation of cellular process (1483)
GO:0030155 : regulation of cell adhesion (116)
GO:0045595 : regulation of cell differentiation (497)
GO:0001558 : regulation of cell growth (305)
GO:0051244 : regulation of cellular physiological process (2793)
GO:0043012 : regulation of fusion of sperm to egg plasma membrane (8)
GO:0009966 : regulation of signal transduction (773)

Three ontologies

Biological Process

— a recognized series of events (more than one step)

- cell cycle, development, metabolism, signal transduction

Molecular Function

— what does a gene's product do?

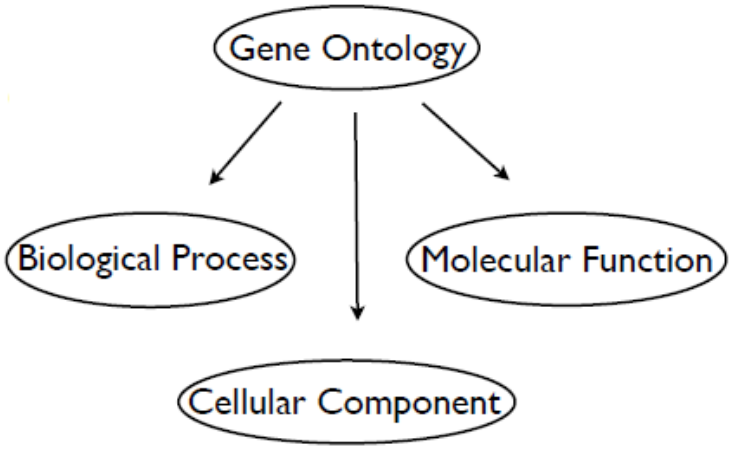
- binding, enzyme, ligand

Cellular Component

— localization within a cell

- membrane, nucleus, complex

It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

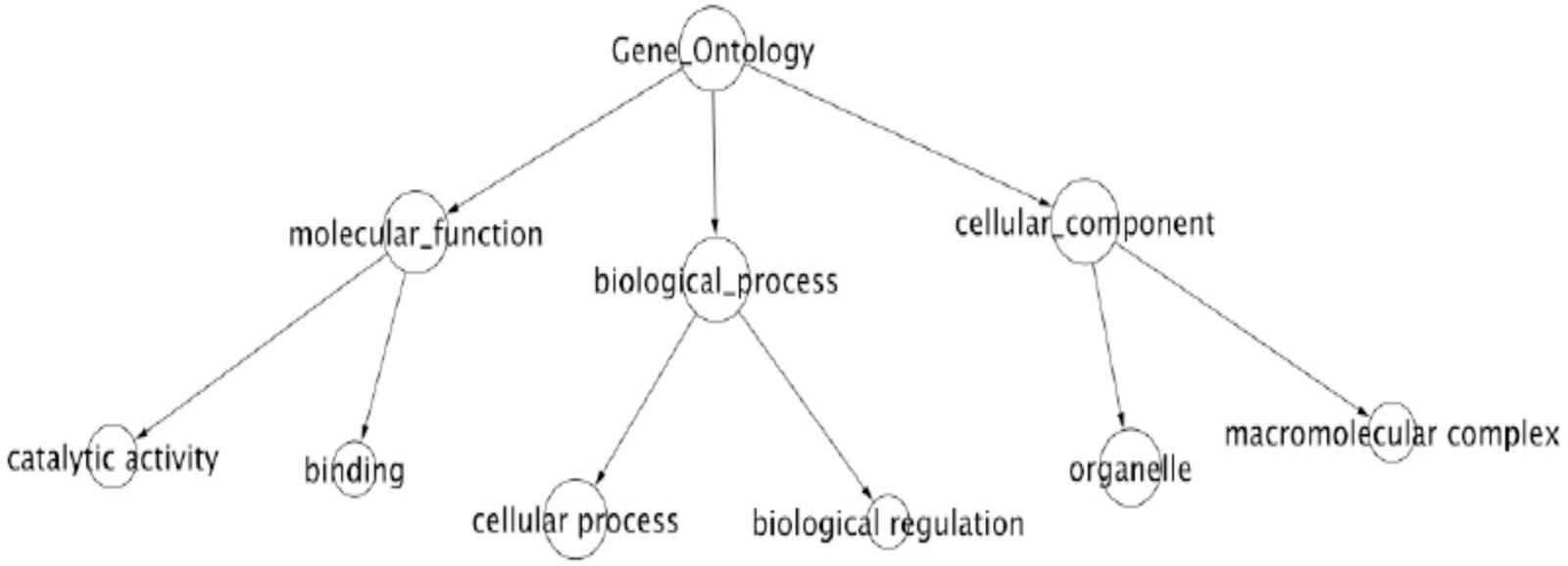


Example: Gene Product = hammer

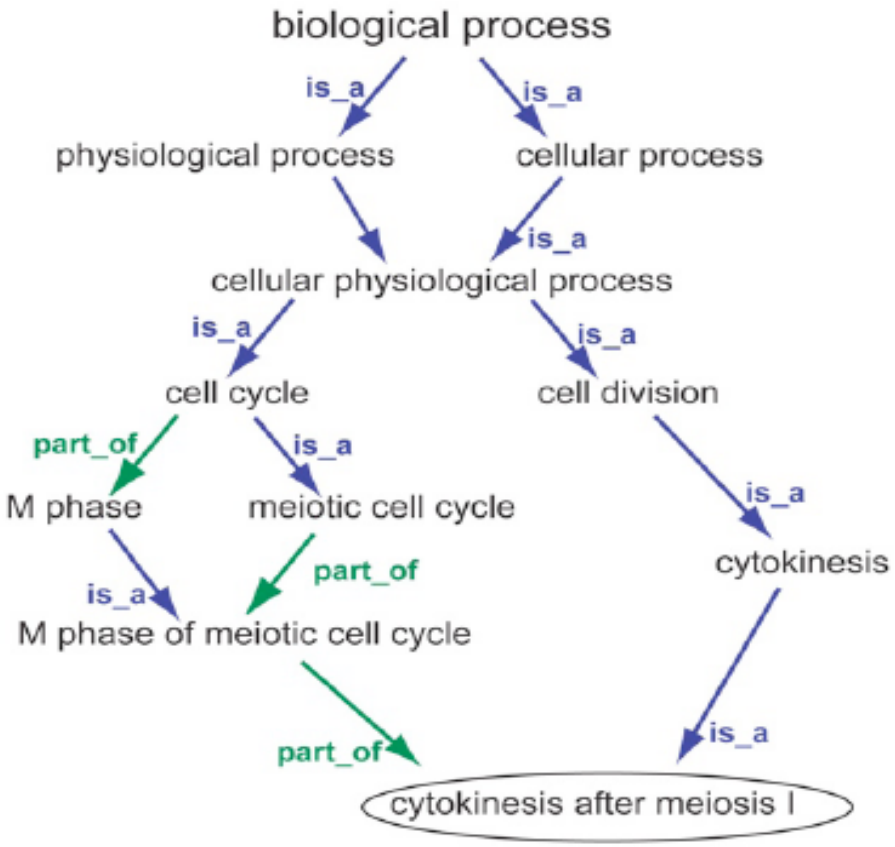
Function (what)		Process (why)
Drive a nail - into wood		Carpentry
Drive stake - into soil		Gardening
Smash a bug		Pest Control
A performer's juggling object		Entertainment

Tree Structure of GO

- Controlled networked terms
  - Parent / child network organized as a tree
  - Terms get more detailed as you move down the tree



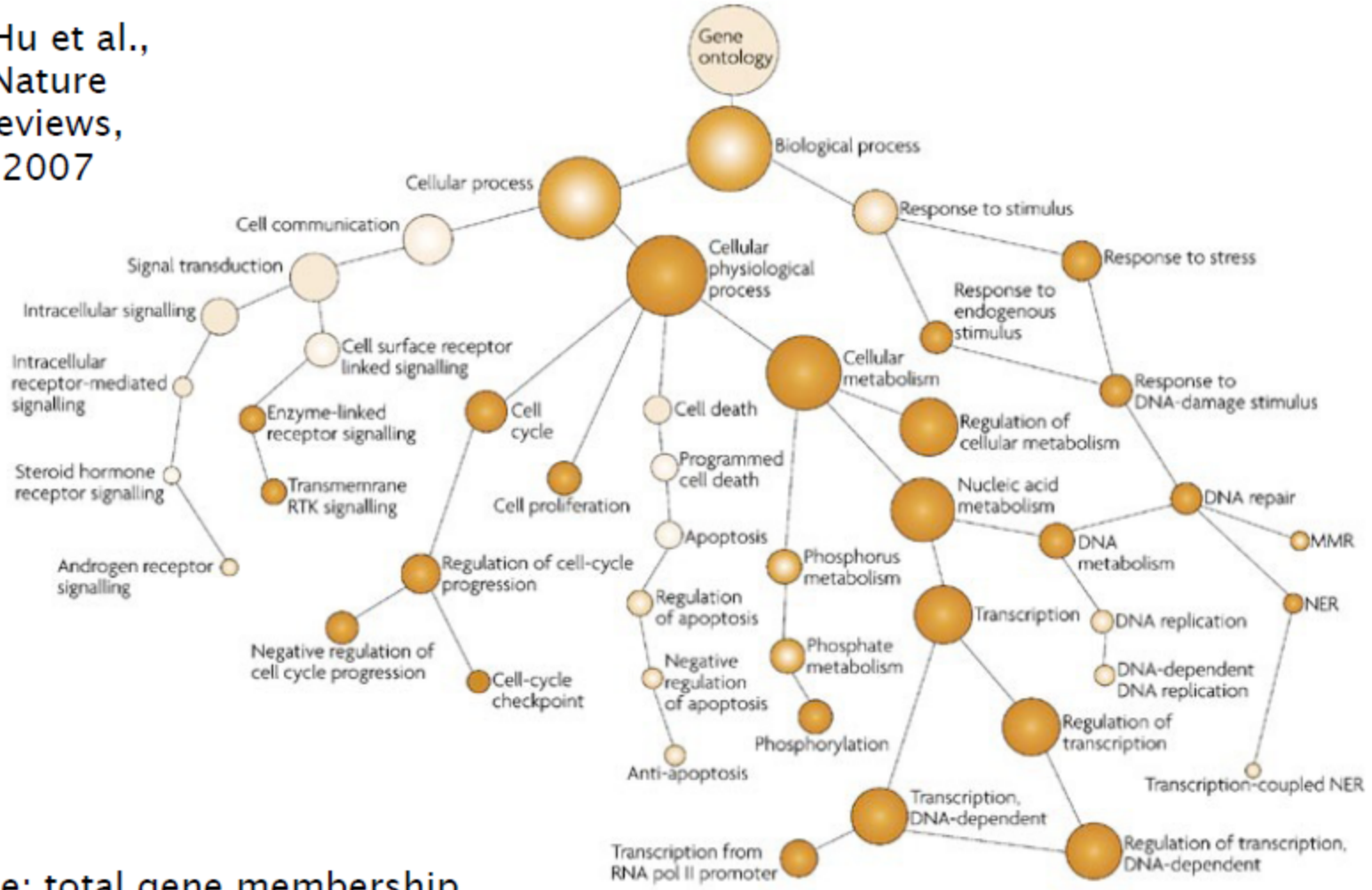
GO tree example



- Any one gene can be a member of more than one GO classification
- GO tree: A child can have more than one parent

Number of genes associated with the term decreases with specificity

P. Hu et al., Nature Reviews, 2007



Size: total gene membership
Color: statistical significance of cancer genes

Nature Reviews | Cancer

Finding Enriched GO terms

- Using Gene Ontology, question we ask:

  — Are any GO terms overrepresented in the gene list, compared to what would happen by chance?

  e.g., is a GO term over-represented? Ex: sample:10/100 from pop1: 600/6000 or from pop2: 15/6000

- A comparison between two gene set

  — Looking for the GO terms that are enriched in one of the gene sets and relatively depleted in the other

- For every GO term, a p-value can be calculated by using hypergeometric test (i.e., Fisher exact test)

Finding Enriched GO terms

- For every GO term, a p-value can be calculated from the following table

	# of genes associated with this GO term	# of genes not associated
Gene List 1	Obs_{11}	Obs_{12}
Gene List 2	Obs_{21}	Obs_{22}

--- Using hypergeometric test

The null hypothesis:

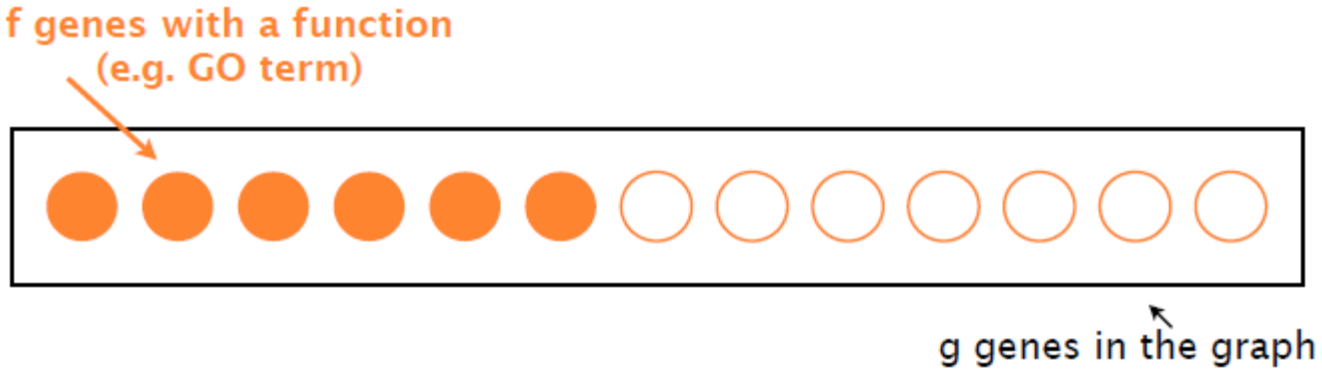
$$\frac{Obs_{11}}{Obs_{21}} = \frac{Obs_{12}}{Obs_{22}}$$

The alternative:

$$\frac{Obs_{11}}{Obs_{21}} \neq \frac{Obs_{12}}{Obs_{22}}$$

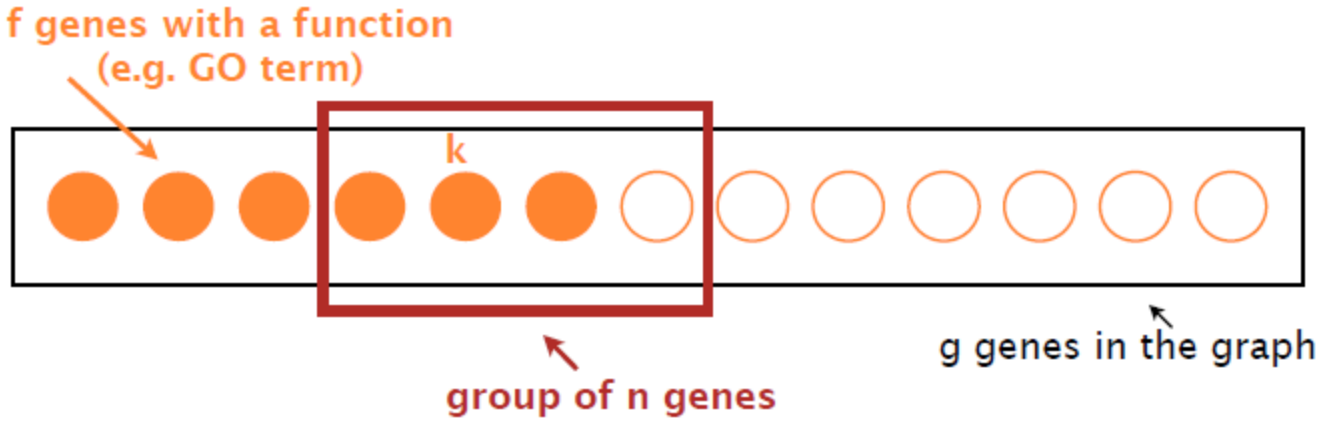
Hypergeometric testing

compare the overlap between two sets



Hypergeometric testing

compare the overlap between two sets

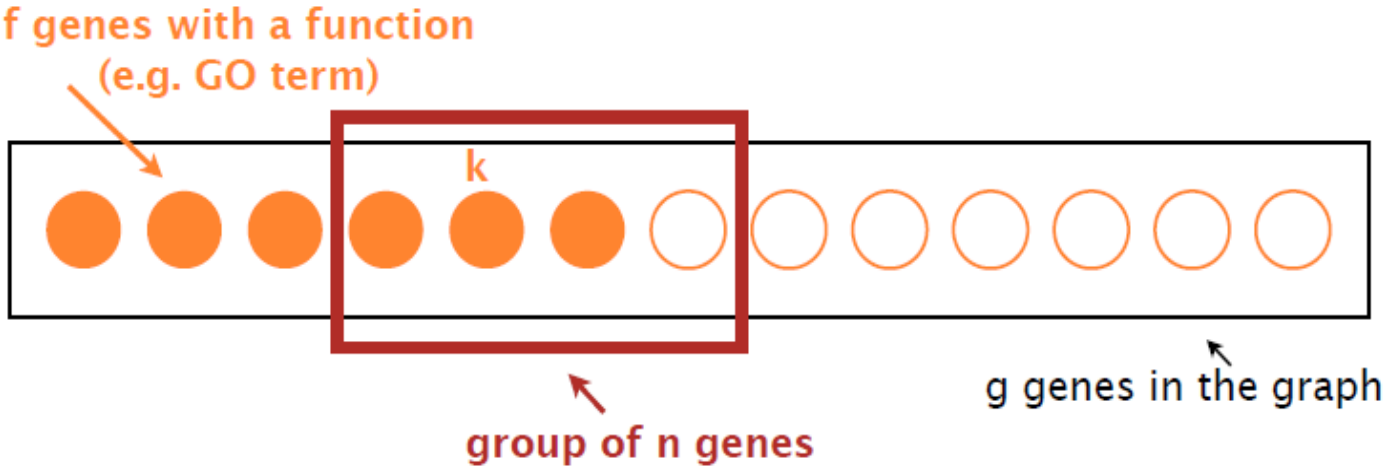


Ho: k genes with the GO term among n genes is consistent with random sampling
Ha: k is larger than what we'd expect by random chance

- The probability of selecting k genes with this function in a cluster of n genes follows hypergeometric distribution with parameters (f, g, n)

Hypergeometric test

- P-value = probability of observing k or more genes with this function in a cluster of n genes



$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}}$$

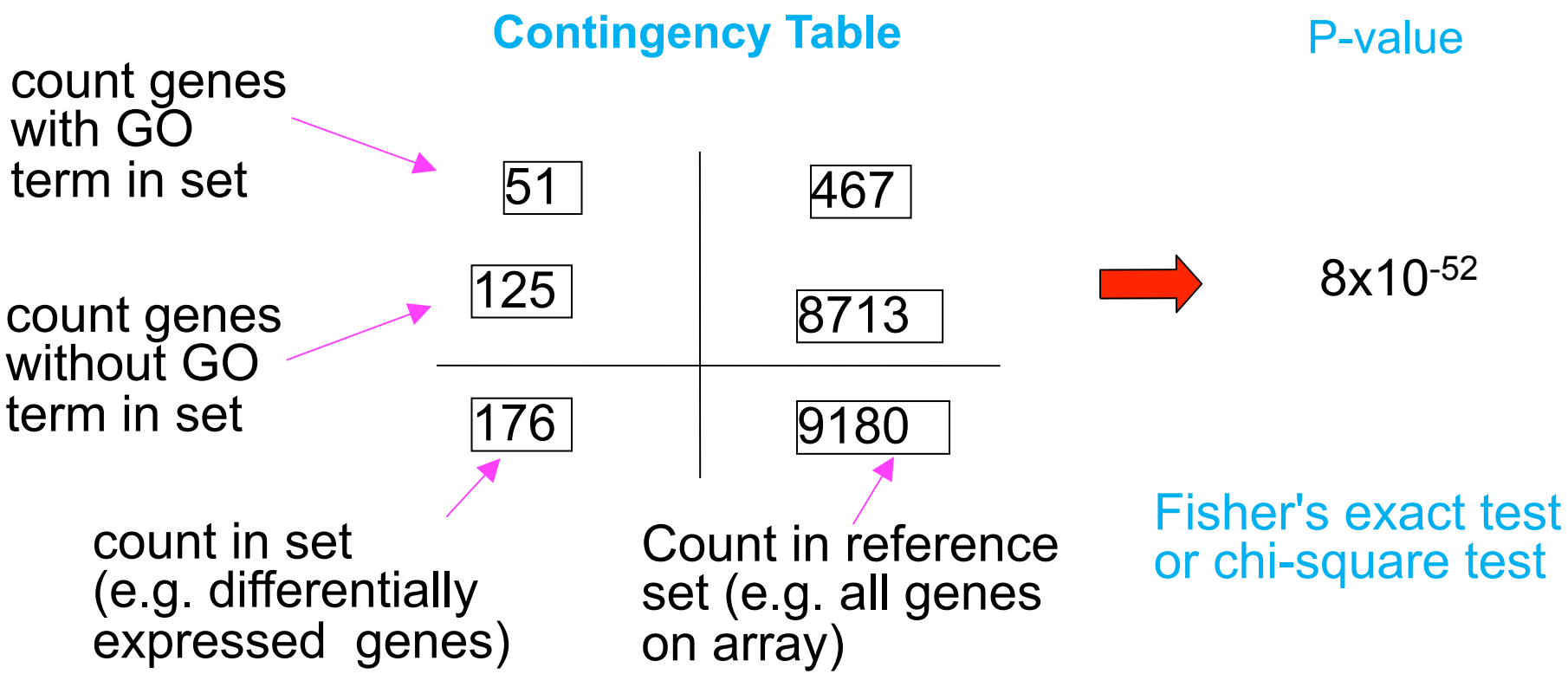
- Small p-value = the term is "significantly enriched" in this group

	List 1			List 2		
	G_1	…	G_{n1}	G_{n1+1}	…	G_{n1+n2}
GO_1	$go_1(G_1)$	…	$go_1(G_{n1})$	$go_1(G_{n1+1})$	…	$go_1(G_{n1+n2})$
GO_2	$go_2(G_1)$	…	$go_2(G_{n1})$	$go_2(G_{n1+1})$	…	$go_2(G_{n1+n2})$
…						

	# of genes associated with GO_1	# of genes not associated with GO_1	
List 1	$\text{Obs}_{11} = \sum_{j=1}^{n1} go(G_j)$	$\text{Obs}_{12} = n1 - \text{Obs}_{11}$	$\Rightarrow P_1$
List 2	$\text{Obs}_{21} = \sum_{j=n1+1}^{n1+n2} go(G_j)$	$\text{Obs}_{22} = n2 - \text{Obs}_{21}$	$\Rightarrow P_2$

…

An example


Contingency Table
P-value
count genes with GO term in set
count genes without GO term in set
51
467
125
8713
176
9180
8x10-52
count in set (e.g. differentially expressed genes)
Count in reference set (e.g. all genes on array)
Fisher's exact test or chi-square test

Another example

- List of 80 significant genes from a microarray experiment of yeast (~6000 genes)
- 10 of the 80 genes are in BP-GO term: DNA replication – Total number of yeast genes in the GO terms is100

Fisher exact test:
p-value: $6.6\ 10^{-07}$

The GO term *DNA replication is overrepresented in our list*

Population size…

But if the population genes g=1000 => p-value=0.98

- Need unspecific prefiltering (remove genes not expressed in any sample)
- Remove genes not present in any GO terms

to reduce false positive rate.

Multiple hypothesis testing

- Suppose there are totally *N* GO terms being tested.
  - The *N* GO terms are tested simultaneously
  - The GO terms are not independent to each other
    - the hierarchical structure of the GO
    - the usage of multiple GO terms in the annotation of one gene.
  - We do not even know the distribution of the p-values when the null is true for all the GO terms

  Use False discovery rate

- Online softwares
  - DAVID (http://david.abcc.ncifcrf.gov/)
  - Gostat (www.***gostat***.*wehi.edu.au/)*
  - …
- R packages
  - topGO
  - Gostat
  - …

Step1: Preparation

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("topGO")
biocLite("ALL")

library(topGO)
library(ALL)

data(ALL)
```

When the topGO package is loaded three new environments GOBPTerm, GOMFTerm and GOMFTerm are created and binded to the package environment. These environments are built based on the GOTERM environment from package GO. They are used for fast recovering of the information specific to each ontology. In order to access all GO groups that belong to a specific ontology, e.g. Biological Process (BP), one can type:

```
BPterms <- ls(GOBPTerm)


MFterms <- ls(GOMFTerm)    ### for MF GO terms
CCterms <- ls(GOCCTerm)    ### for CC GO terms
```

load the annotation data

##The chip used for the experiment is HGU95aV2 Affymetrix

```
biocLite("hgu95av2.db")

library(hgu95av2.db)
```

Gene filtering

remove genes with low expression value and genes which might have very small variability across the samples using package "genefilter"

```
biocLite("genefilter")
library(genefilter)
f1 <- pOverA(0.25, log2(100))
f2 <- function(x) (IQR(x) > 0.5)
ff <- filterfun(f1, f2)
eset <- ALL[genefilter(ALL, ff), ]
```

The filter selects only 2400 probesets out of 12625 probesets available on the hgu95av2 array

Step 2: Creating a topGOdata object

This object will contain all information necessary for the GO analysis, namely the gene list, the list of interesting genes, the scores of genes (if available) and the part of the GO ontology (the GO graph) which needs to be used in the analysis.

need to define the set of genes that are to be annotated with GO terms. Usually, one starts with all genes present on the array. In our case we start with 2400 genes, genes that were not removed by the filter.

```
geneNames <- featureNames(eset)
length(geneNames)
```

2.1: Using score to determine a list of interesting gene

P-value for the t-test to discriminate between ALL cells delivered from either B-cell or T-cell precursors (95 B-cell ALL samples and 33 T-cell ALL samples).

```
>y = as.integer(sapply(eset$BT, function(x)
  return(substr(x, 1, 1) == "T")))
>table(y)
 0  1
95 33
>library(multtest)
>geneList = getPvalues(exprs(eset), classlabel = y,
  alternative = "greater", correction="BH")
> hist(geneList, br = 50)
```

2.1: Using score to determine a list of interesting gene -2

```
topDiffGenes <- function(allScore) {
  return(allScore < 0.0001)
 }
```

This function selects genes based on their scores (in our case the adjusted p-values) and returns a logical vector specifying which gene is selected and which not.

```
 x <- topDiffGenes(geneList)
sum(x)
```

2.2: build topGOdata object

```
>sampleGOdata <- new("topGOdata", ontology = "BP",
  allGenes = geneList, geneSel = topDiffGenes,
  description = "GO analysis of ALL data based on
  diff. expression.",  annot = annFUN.db, affyLib =
  "hgu95av2.db")


> sampleGOdata
```

Step 3: Running the desired tests

#Once we have an object of class topGOdata we can start with the enrichment analysis. Since for each gene we have a score and there is also a procedure to select interesting genes based on the scores we will use two types of test statistics:

#Fisher exact test and Kolmogorov-Smirnov (KS) test

#For KS, there are the classic and the elim methods.

```
>resultFisher <- runTest(sampleGOdata, algorithm =
  "classic", statistic = "fisher")
>resultKS <- runTest(sampleGOdata, algorithm =
  "classic", statistic = "ks")
>resultKS.elim <- runTest(sampleGOdata, algorithm =
  "elim", statistic = "ks")
```

To look at the results of significant GO terms, put all resulting p-values into a list. Then we can use the GenTable function to generate a table with the results.

```
>allRes <- GenTable(sampleGOdata, classicFisher =
  resultFisher, classicKS = resultKS, elimKS =
  resultKS.elim, orderBy = "elimKS", topNodes = 10)
> allRes
```

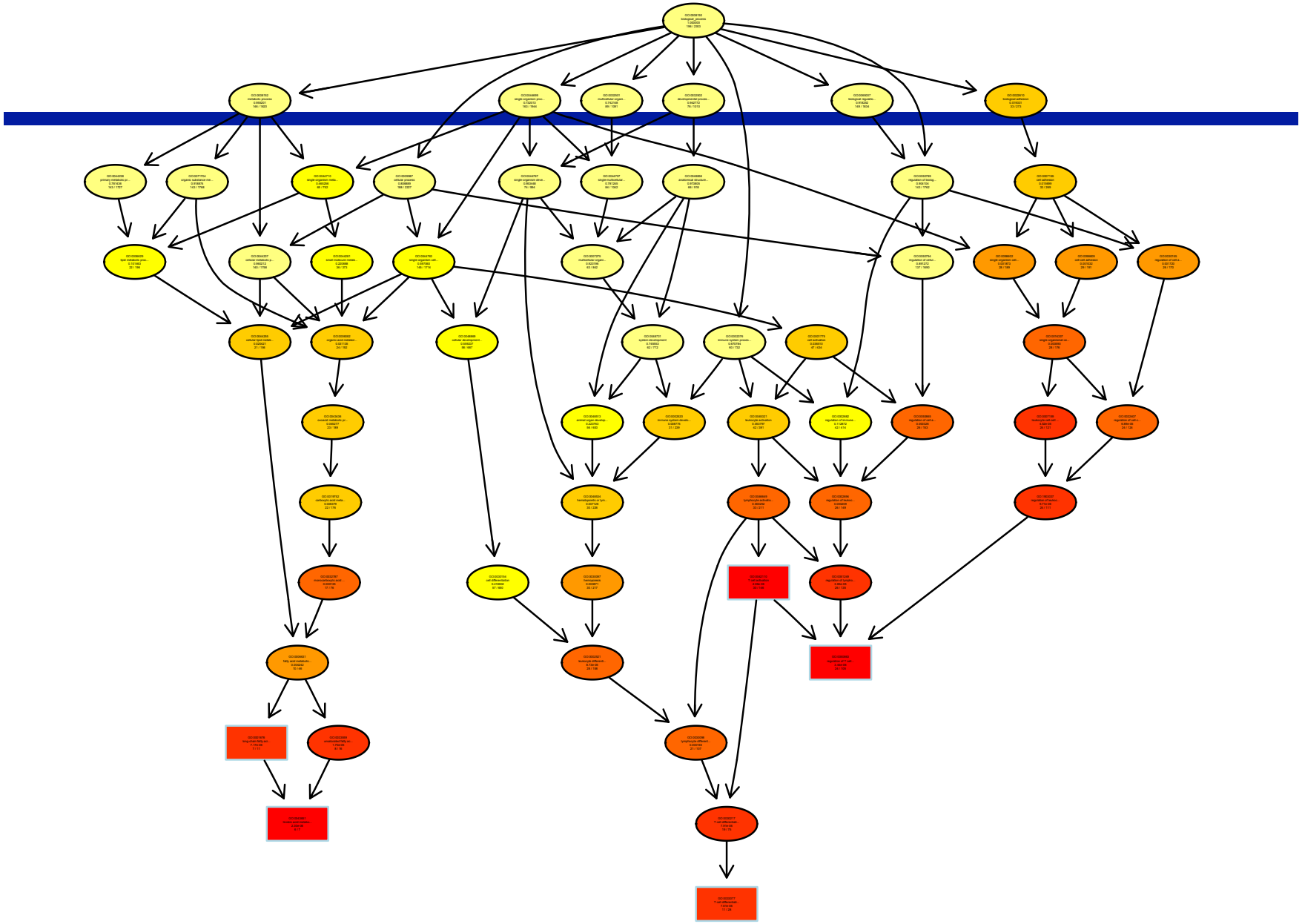
Optional part: graphs by Rgraphviz

Investigating how the significant GO terms are distributed over the GO graph

```
> biocLite("Rgraphviz")
```

```
>library(Rgraphviz)
```
install carefully (as it need us to change the path in the system settings)

```
>showSigOfNodes(sampleGOdata, score(resultFisher),
  firstSigNodes = 5, useInfo = "all")
```



In this plot,

- The subgraph induced by the top 5 GO terms identified by the classic algorithm for scoring GO terms for enrichment.
- Boxes indicate the 5 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant).
- Black arrows indicate is-a relationships and red arrows part-of relationships.

Check which genes in which top terms

For the selected GO terms (e.g., top 10), count the number of annotated genes and obtain their annotation.

```
topRes <- GenTable(sampleGOdata, classicFisher = resultFisher, orderBy = "classicFisher", topNodes = 20)
topRes
sel.terms=topRes$GO.ID[1:10]
sel.terms
num.ann.genes <- countGenesInTerm(sampleGOdata, sel.terms)
num.ann.genes
ann.genes <- genesInTerm(sampleGOdata, sel.terms)
ann.genes
```