

## Last time:

- Choose the right distance metric to compare the expression of two genes
- Describe why you would cluster expression by genes or experiments
- Manually cluster small vectors using hierarchical or k-means clustering
- Read a dendrogram
- Describe the results of Principal Component Analysis (PCA)

Last time

Find Similar  
Genes

Find Similar  
Conditions

Today

Find  
Changes in  
Expression

Find  
Functions of  
Genes of  
Interest

RNA-  
Seq

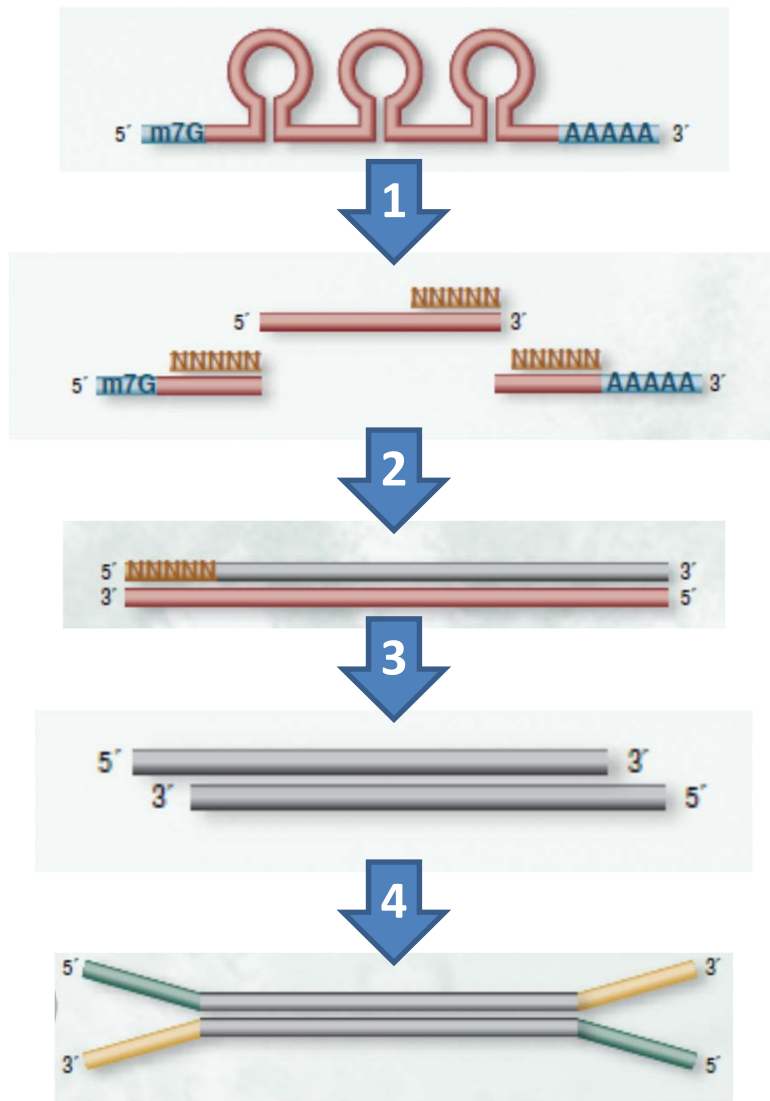
Expression of  
Each Gene



# Outline

- Overview of the steps of RNA-Seq
- Deriving expression levels from sequence data
- Gene Ontology
- Statistical significance

# 1. Fragment RNA and prime with random DNA primers



NEBNext® for Illumina®

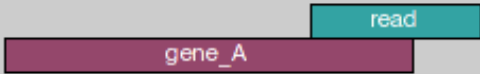

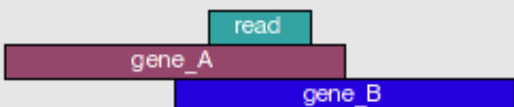
NGS SAMPLE PREPARATION

# Outline

- Overview of the steps of RNA-Seq
- Deriving expression levels from sequence data
- Gene Ontology
- Statistical significance

# When reads overlap more than one exon

“Union” is  
the default  
mode

	union	intersection_strict
	gene_A	gene_A
	gene_A	gene_A
	gene_A	gene_A
	gene_A	no_feature
	gene_A	no_feature
	ambiguous	gene_A
	ambiguous	ambiguous



Fragments get sequenced  
"reads"



Sequencing reads

Raw reads  
FASTA, FASTQ

Align reads to genome

Align to genome  
TopHat2



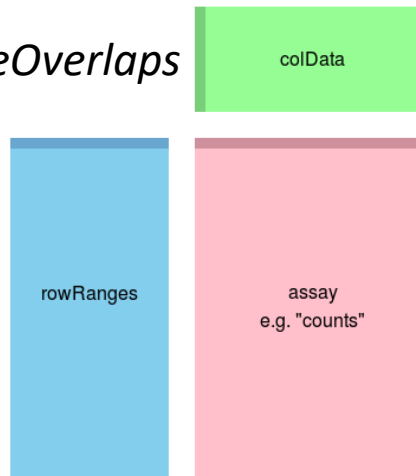
Mapped Reads  
SAM, BAM

Assemble transcripts



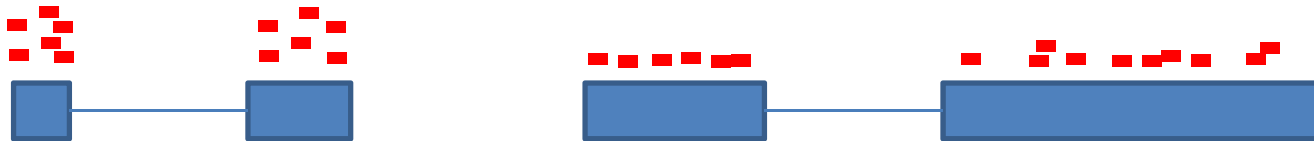
Reference-based

*summarizeOverlaps*



1. Find differentially expressed genes
2. Cluster
3. PCA

# Raw counts are misleading



1. A long transcript with a low level of expression will still produce more sequence reads than a short, highly expressed transcript.
2. An experiment that is sequenced more deeply will make all genes appear to be expressed at higher levels

To correct for this, we use “Reads per Kilobase Million (RPKM)”



Gene	Length in KB	Replicate 1	Replicate 2	Replicate 3
A	2	1.0E6	1.2E6	3.0E6
B	4	2.0E6	2.5E6	6.0E6
C	10	0	0	1.0E5

## Raw reads

- Count the number of reads in each sample in millions.

Reads per million	A	0.333	0.324	0.330
	B	0.667	0.676	0.659
	C	0	0	0.011

Reads per kilobase million RPKM		Replicate 1	Replicate 2	Replicate 3
	A	0.167	0.162	0.165
	B	0.167	0.169	0.165
	C	0.00	0.00	0.001

Gene	Length in KB	Replicate 1	Replicate 2	Replicate 3
A	2	1.0E6	1.2E6	3.0E6
B	4	2.0E6	2.5E6	6.0E6
C	10	0	0	1.0E5

Reads  
per  
million

A	0.333	0.324	0.330
B	0.667	0.676	0.659
C	0	0	0.011

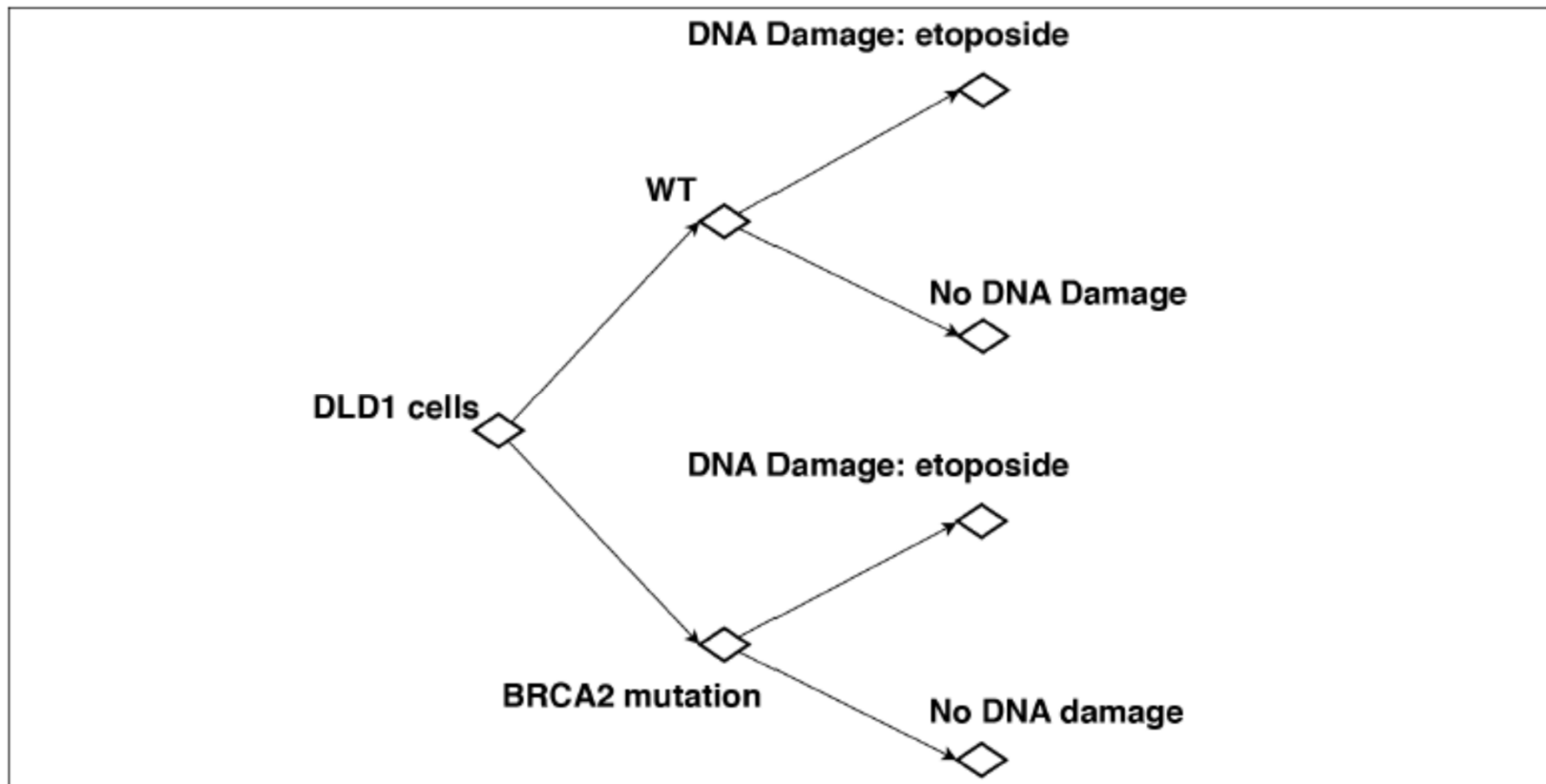
This step corrects for sequencing depth. Note that numbers are now more consistent across replicates

Reads  
per  
kilobase  
million  
RPKM

	Replicate 1	Replicate 2	Replicate 3
A	0.167	0.162	0.165
B	0.167	0.169	0.165
C	0.00	0.00	0.001

This step corrects for gene length. Note that genes A and B have similar RPKMs but very different raw read counts.

# Differential expression

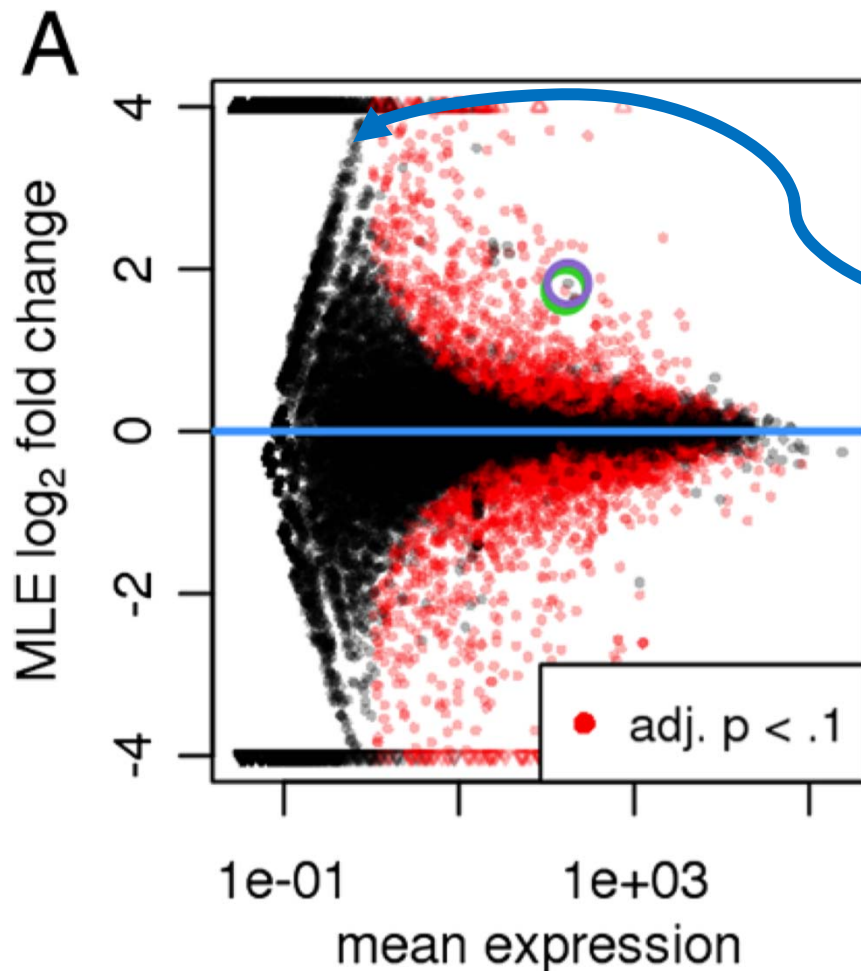


Unfortunately, we can't just compare RPKM values across conditions.

Random sampling errors will produce different values even for genes that are expressed at a constant level.

# Heteroskedasticity

variance of LFCs depends on the mean



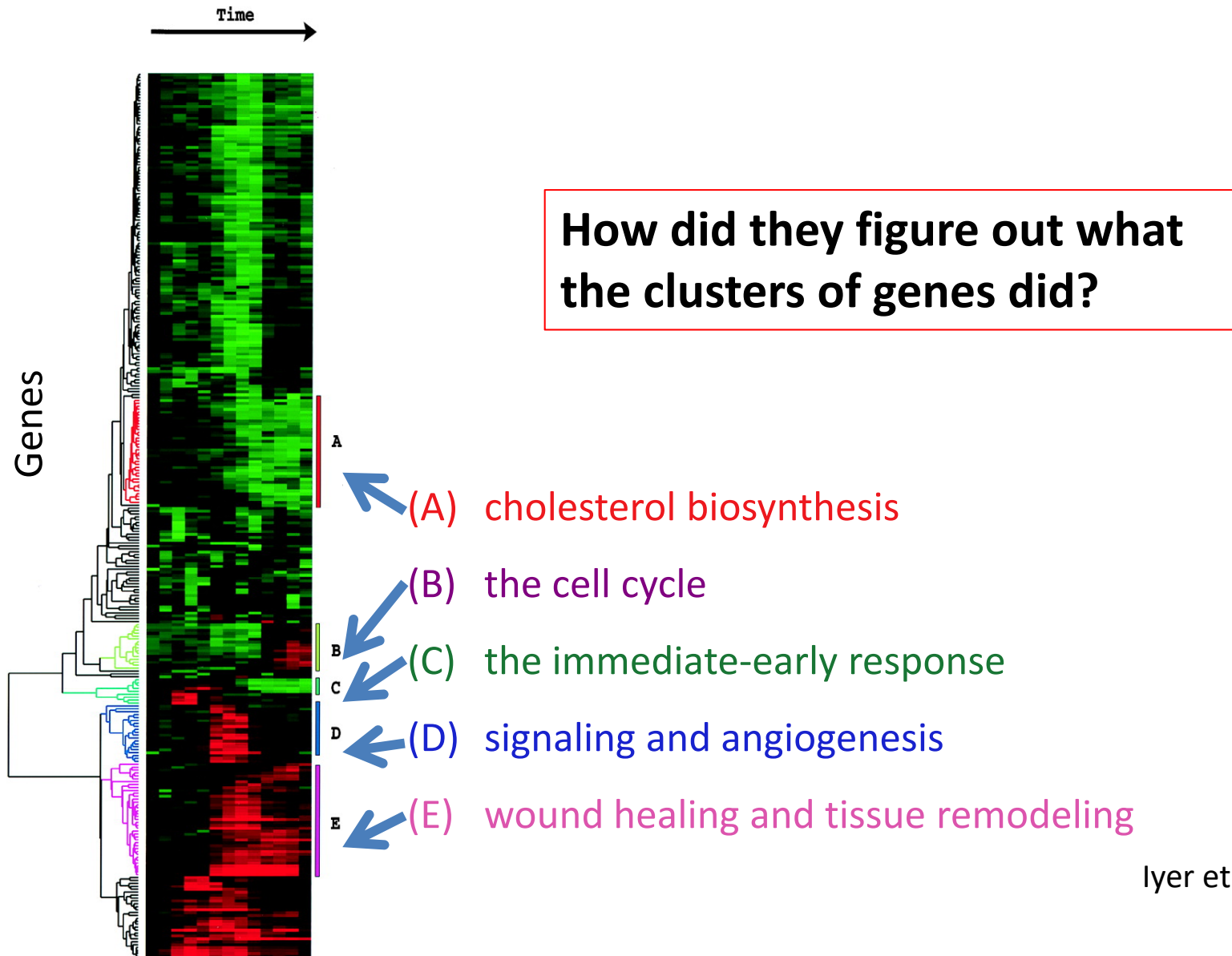
Love et al. *Genome Biology* (2014) 15:550

- Why are large fold-changes so common for poorly expressed genes?
- Ratios with small numbers are always more noisy.
- Transforming the data can reduce this bias.
- DESeq2 uses something called a *regularized logarithm* transformation (rlog).

# Do your data make sense?

- Technical replicates should be very similar ( $R^2 > .9$ )
- Biological replicates should cluster together

# Interpreting your results



Iyer et al. *Science* 1999

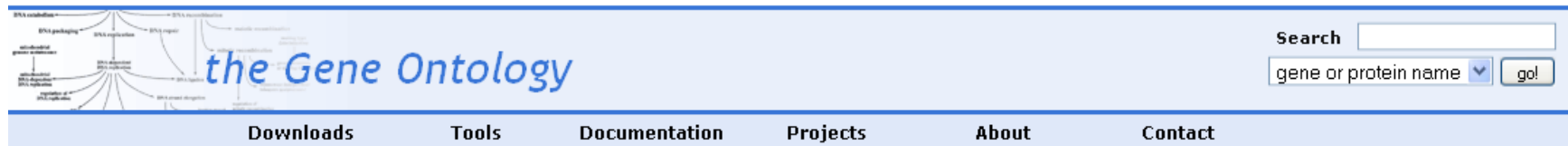
# Outline

- Overview of the steps of RNA-Seq
- Deriving expression levels from sequence data
- Gene Ontology
- Statistical significance

# Biological Insights

- What types of genes are being differentially expressed?

<http://www.geneontology.org>

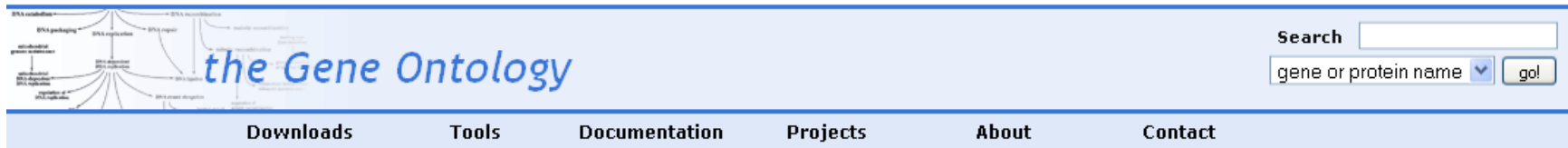


Controlled vocabulary to describe genes:

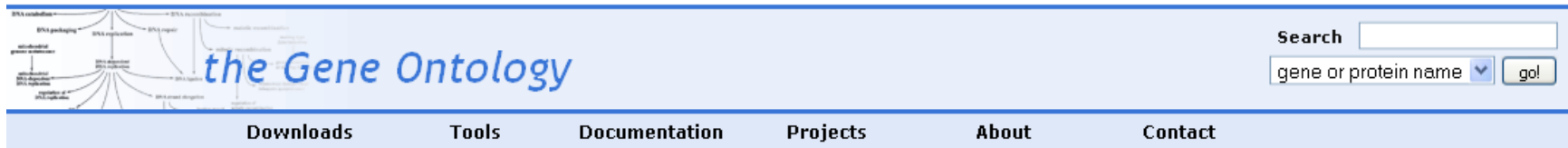
- Biological process
- Cellular component
- Molecular function



- Biological process
  - signal transduction; glucose transport
- Cellular component
  - nucleus; ribosome; protein dimer
- Molecular function
  - binding; transporter



- **Biological process**
- A series of events accomplished by one or more ordered assemblies of molecular functions.
- Examples of broad biological process terms are **cellular physiological process** or **signal transduction**.
- A process should have at least two distinct steps.

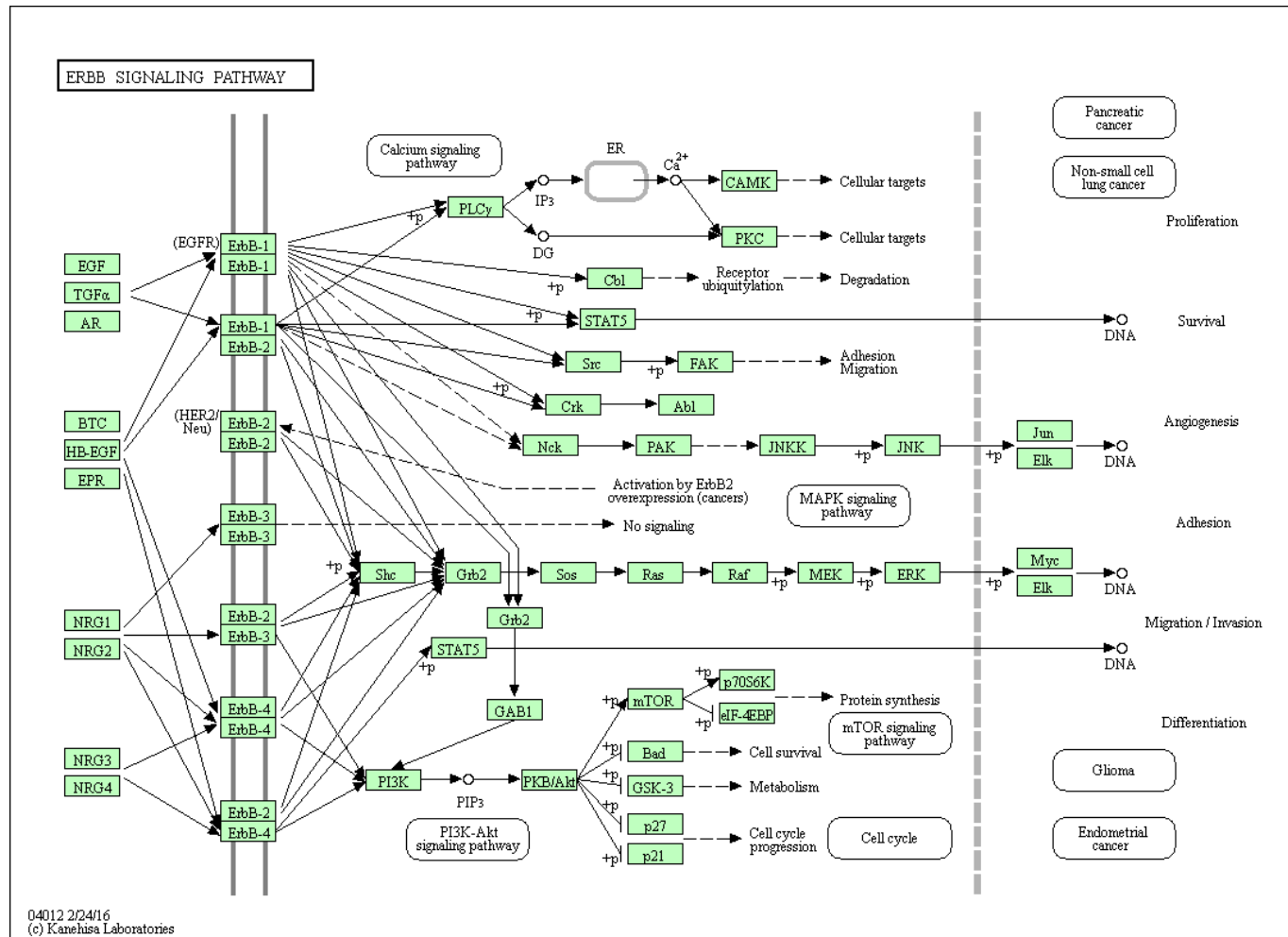


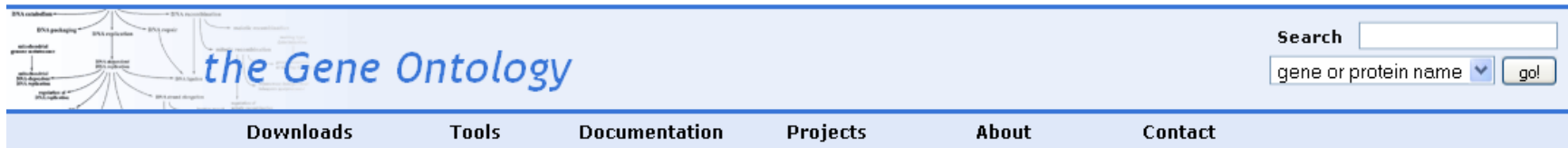
- **Biological process**
- A biological process is not equivalent to a pathway.
  - Does not represent the dynamics or dependencies of a pathway.

GO

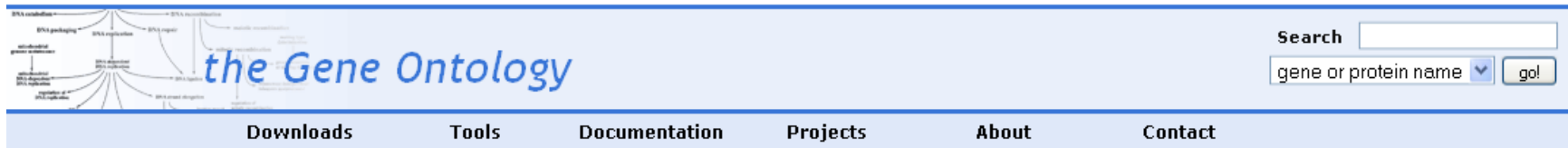
BTC	NRAS
CDC37	NRG1
Cpne3	NRG2
CPNE3	NRG4
CUL5	PIK3CA
EGF	PIK3R1
EGFR	PRKCA
ERBB2	PTK6
ERBB3	PTPN12
ERBB4	PTPN18
ERBIN	Ptprr
EREG	PTPRR
GAB1	RPS27A
GRB2	SHC1
GRB7	SOS1
HBEGF	SRC
HRAS	STUB1
HSP90AA1	Symbol
KRAS	UBA52
MATK	UBB
Myoc	UBC
MYOC	

## KEGG Pathway

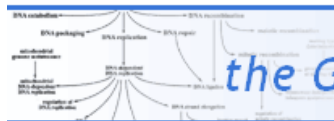




- **Cellular component**
- Part of a
  - anatomical structure (e.g. **rough endoplasmic reticulum or nucleus**) or a
  - gene product group (e.g. **ribosome, proteasome or a protein dimer**).



- **Molecular function**
- Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level.
- Examples:
  - **Broad: catalytic activity, transporter activity, or binding**
  - **Narrow: adenylate cyclase activity or Toll receptor binding.**



the Gene Ontology

# Estrogen receptor

Search   
gene or protein name

[Downloads](#)

[Tools](#)

[Documentation](#)

[Projects](#)

[About](#)

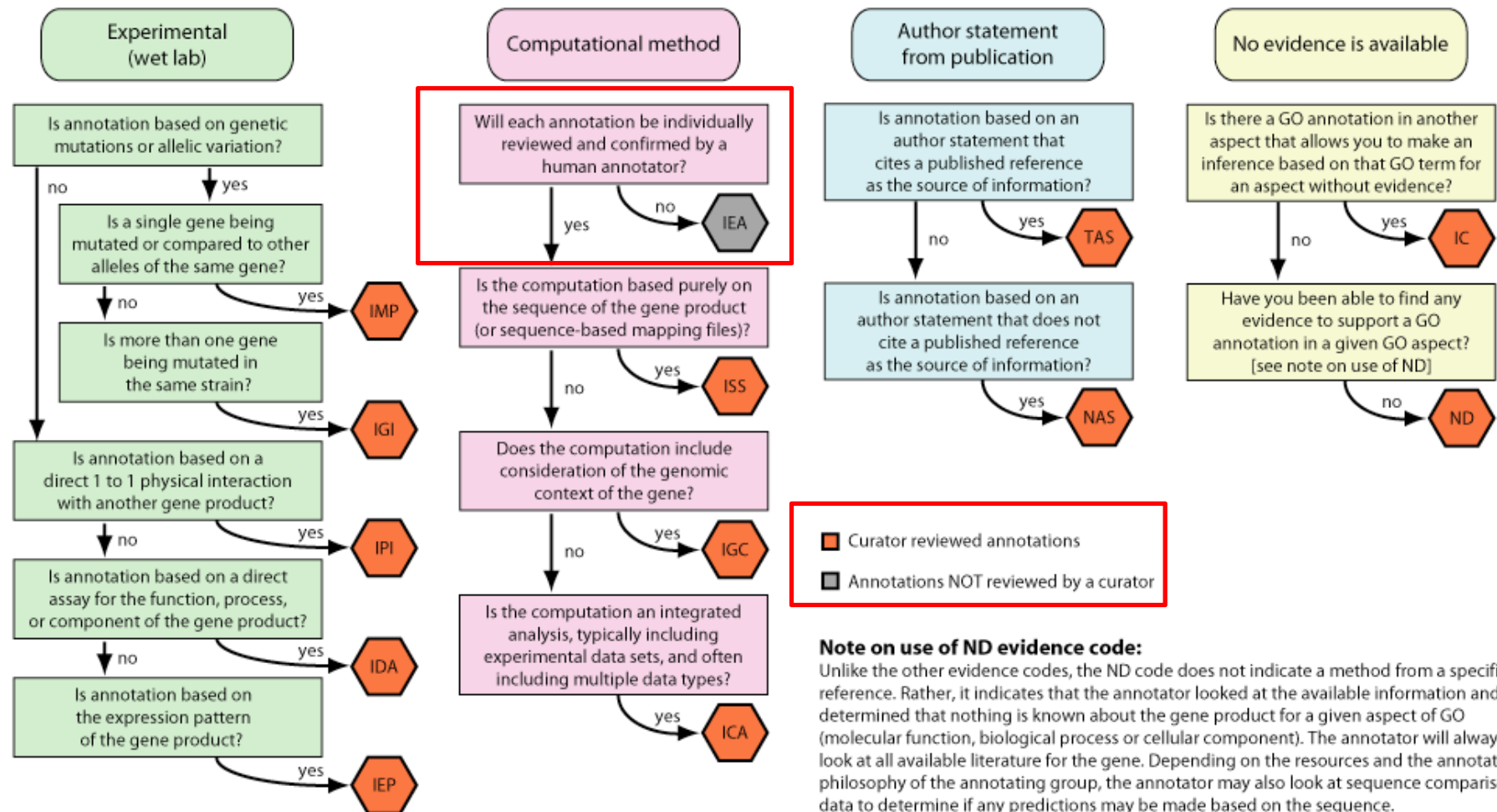
[Contact](#)

Accession, Term		Ontology	Qualifier	Evidence
<input type="checkbox"/> GO:0030520 : <a href="#">estrogen receptor signaling pathway</a>	<a href="#">41 gene products</a> <a href="#">view in tree</a>	<a href="#">biological process</a>		<a href="#">NAS</a>
<input type="checkbox"/> GO:0043526 : <a href="#">neuroprotection</a>	<a href="#">67 gene products</a> <a href="#">view in tree</a>	<a href="#">biological process</a>		<a href="#">IEA</a> With <a href="#">Ensembl:ENSRNOP00000026350</a>
<input type="checkbox"/> GO:0048386 : <a href="#">positive regulation of retinoic acid receptor signaling pathway</a>	<a href="#">9 gene products</a> <a href="#">view in tree</a>	<a href="#">biological process</a>		<a href="#">IDA</a>
<input type="checkbox"/> GO:0045885 : <a href="#">positive regulation of survival gene product expression</a>	<a href="#">56 gene products</a> <a href="#">view in tree</a>	<a href="#">biological process</a>		<a href="#">IEA</a> With <a href="#">Ensembl:ENSRNOP00000026350</a>
<input type="checkbox"/> GO:0006355 : <a href="#">regulation of transcription, DNA-dependent</a>	<a href="#">16904 gene products</a> <a href="#">view in tree</a>	<a href="#">biological process</a>		<a href="#">NAS</a>
<input type="checkbox"/> GO:0043627 : <a href="#">response to estrogen stimulus</a>	<a href="#">354 gene products</a> <a href="#">view in tree</a>	<a href="#">biological process</a>		<a href="#">IEA</a> With <a href="#">Ensembl:ENSRNOP00000026350</a>
<input type="checkbox"/> GO:0007165 : <a href="#">signal transduction</a>	<a href="#">18490 gene products</a> <a href="#">view in tree</a>	<a href="#">biological process</a>		<a href="#">TAS</a>
				<a href="#">TAS</a>

Not just the obvious categories

## GO Evidence Code Decision Tree

What type of evidence is the annotation based on?





# Tools

<http://www.geneontology.org/GO.tools.shtml>



## Gene Ontology Tools

### Consortium Tools

#### Non-Consortium Tools

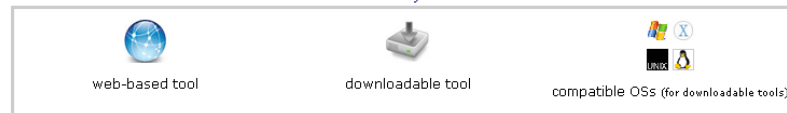
Tools for searching and browsing GO

Annotation tools

Tools for gene expression/microarray analysis

Other tools

### Key



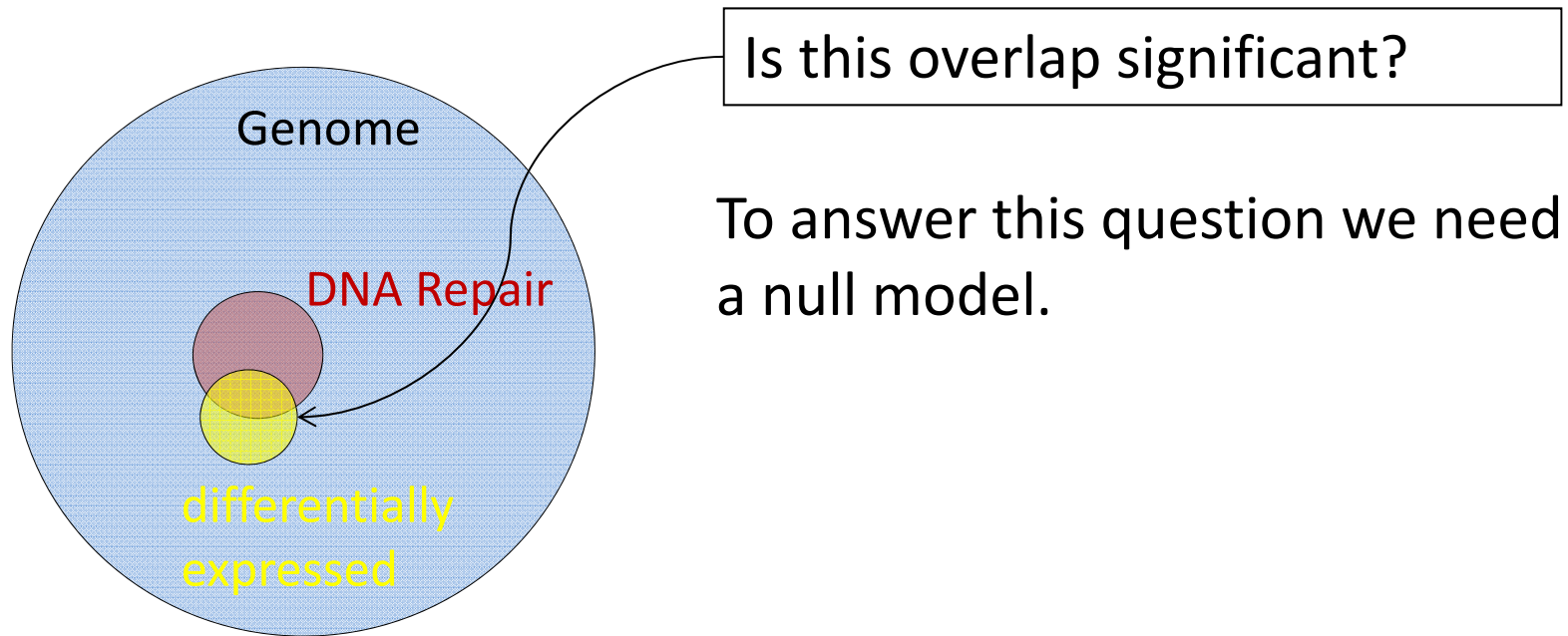
# Outline

- Overview of the steps of RNA-Seq
- Deriving expression levels from sequence data
- Gene Ontology
- Statistical significance

# Statistical significance

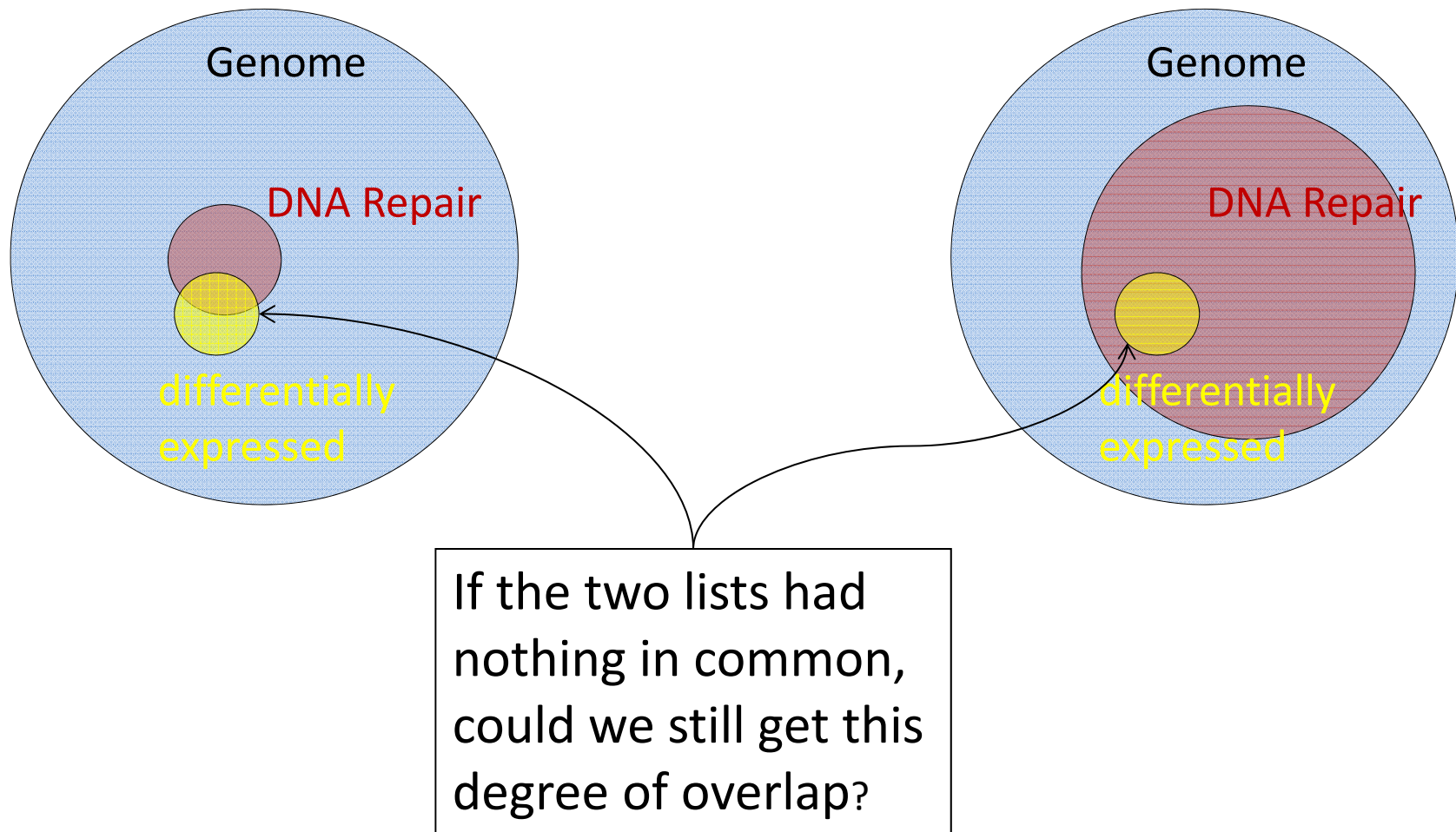
- I found that ten of the upregulated genes in my dataset are annotated as “DNA Repair” ...

# Statistical significance

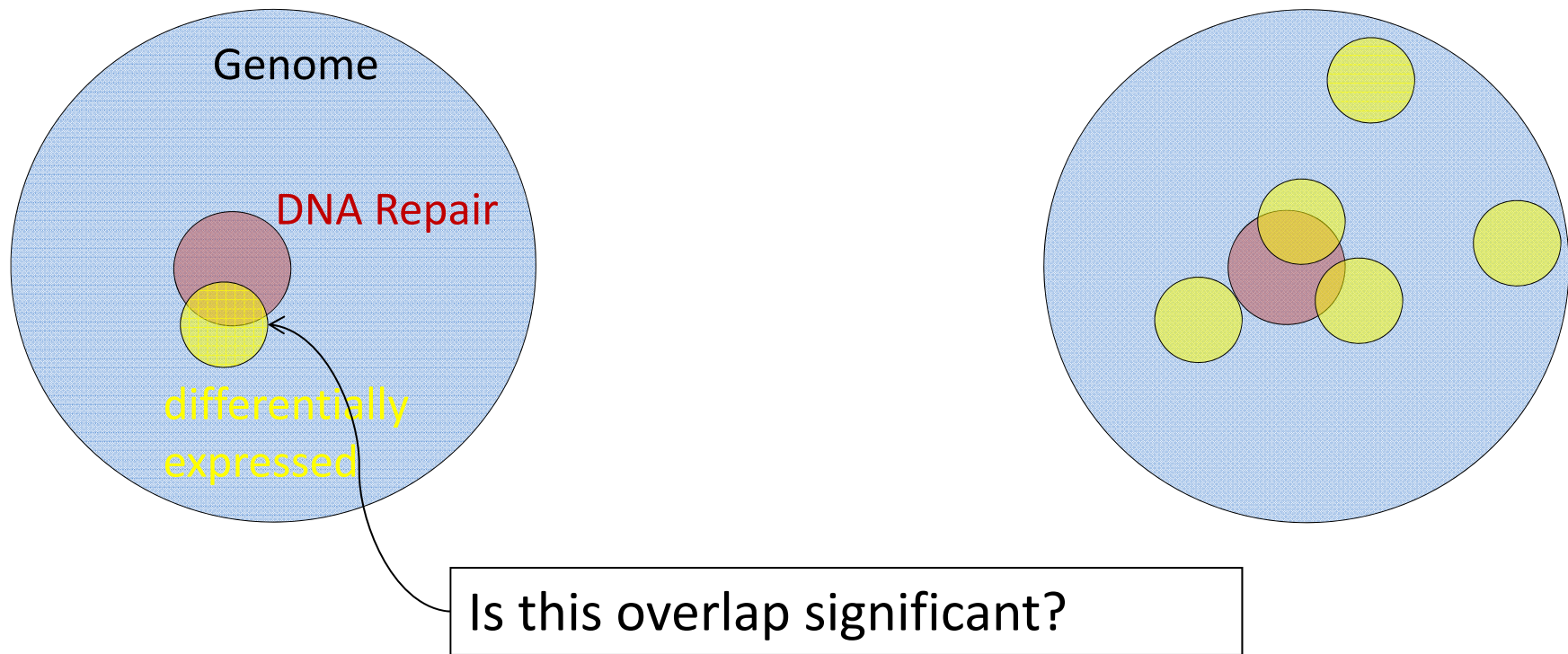


# Statistical significance

The significance depends on the size of the lists.

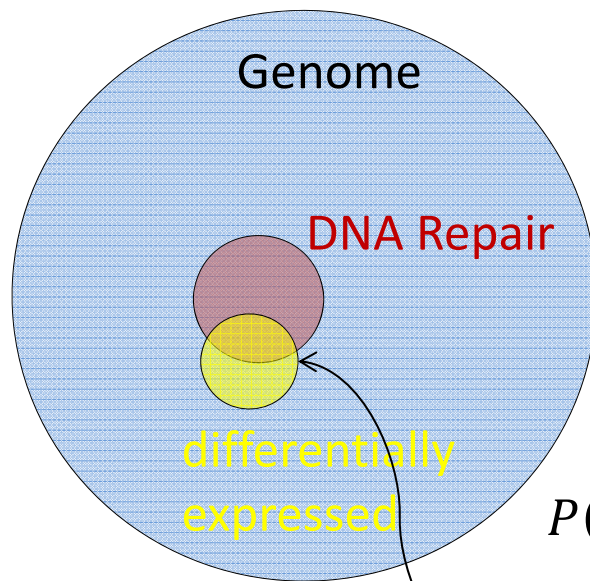


# Statistical significance



# Statistical significance

The probability of getting **exactly** this amount of overlap for two randomly chosen sets of genes of the same size is given by the hypergeometric distribution:



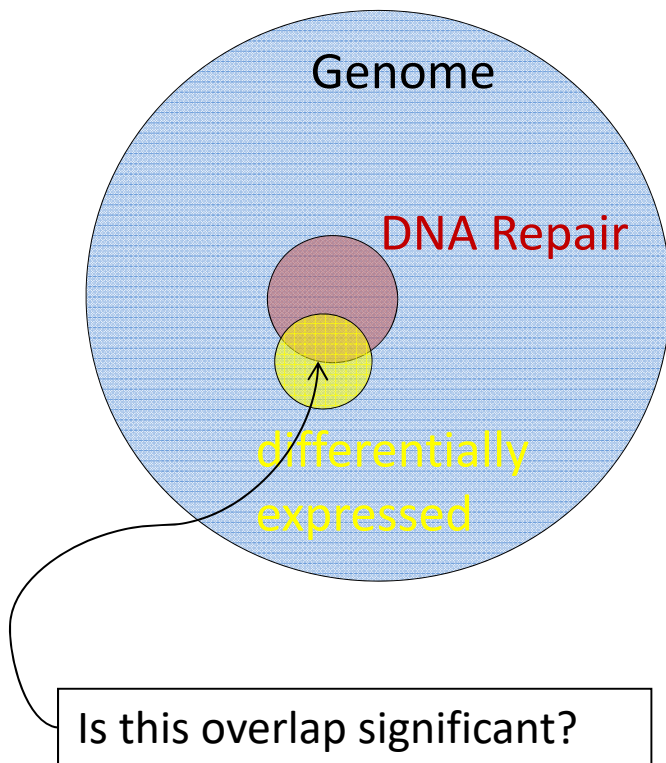
$$P(Overlap) = \frac{\binom{DNA\ repair}{Overlap} \binom{Genome - DNA\ repair}{DiffExp - Overlap}}{\binom{Genome}{DiffExp}}$$

Is this overlap significant?

Recall that  $\binom{n}{k}$  ("n choose k") is the binomial coefficient.

= the number of ways to choose k items from a set of n.

# Statistical significance

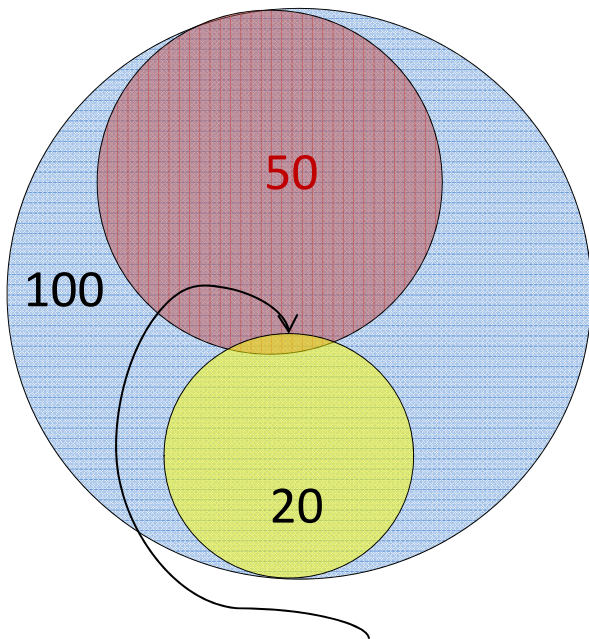


- The hypergeometric gives the probability of getting **exactly** this amount of overlap for two randomly chosen sets of genes of the same size.
- But that's not exactly what we need to know.
- We wish to test if a term is "enriched" in our data.
- Do we see **more** of a term than we would expect in the null model?



# Statistical significance

In this case at left, the p-value for an overlap of exactly one is 0.000003.

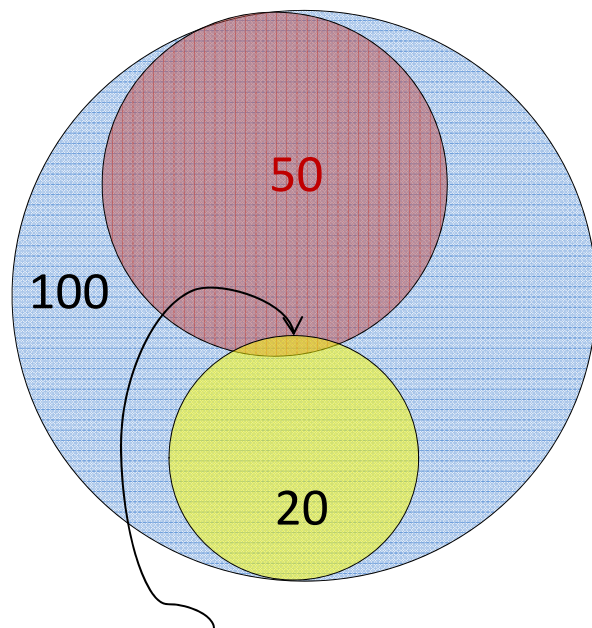


Only one overlapping gene.

In fact, you would expect to see a larger overlap under the null model.

Overlap	P-value
1	0.000003
2	0.00004
3	0.0004
4	0.002
5	0.009
6	0.02
7	0.07
8	0.12
9	0.17
10	0.2

# Statistical significance



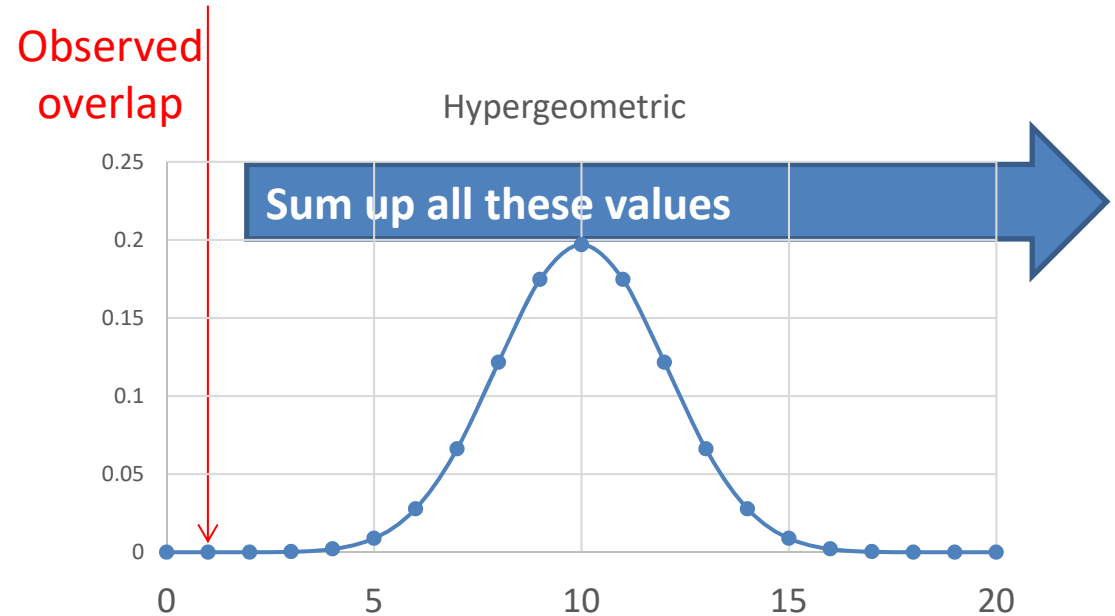
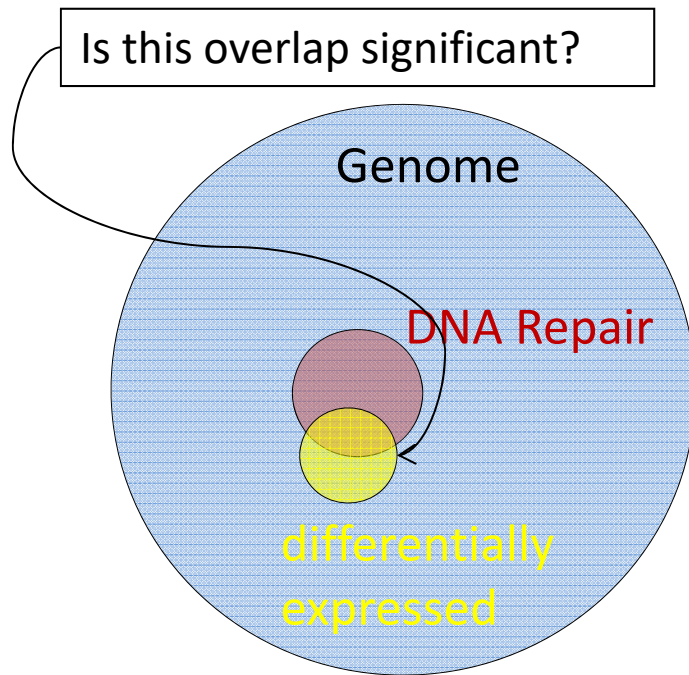
Only one  
overlapping  
gene.

To determine if we see **more** of a term than we would expect in the null model, we compute the cumulative distribution function.

$$\begin{aligned} P(\text{overlap} = 1) &= 0.000003 \\ P(\text{overlap} \geq 1) &= 0.999997 \end{aligned}$$

Overlap	P-value
0	9E-8
1	0.000003
2	0.00004
3	0.0004
4	0.002
5	0.009
6	0.02
7	0.07
8	0.12
9	0.17
10	0.2

# Statistical significance

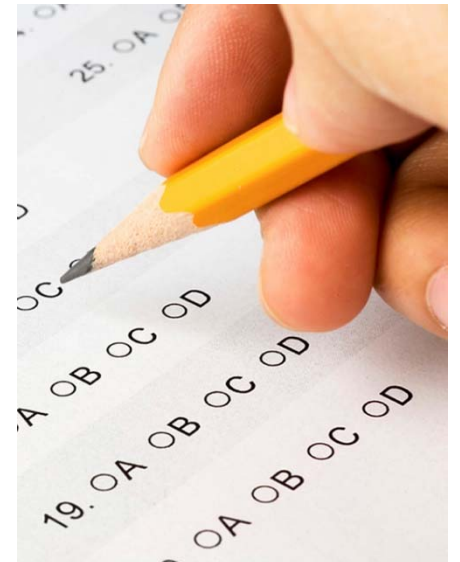


$$CDF(Overlap) = \sum_{n=overlap}^{\text{Number of genes in DNA Repair}} \frac{\binom{DNA\ repair}{n} \binom{Genome - DNA\ repair}{DiffExp - n}}{\binom{Genome}{DiffExp}}$$

CDF=Cumulative distribution function

# Multiple Hypotheses

- Let's imagine we test each GO term using the hypergeometric distribution, each time filtering with a p-value of 0.01
- From the definition of the p-value, we expect that the null-hypothesis has a 1% probability of being correct **for each test.**
- There are roughly 30,000 terms in GO.
- At this level, we expect roughly 300 false positives!



# Multiple Hypotheses

- A simple solution: require that the p-value be small enough to reduce the false positives to the desired level.
- This is called the Bonferroni correction.
- In our case, we would only accept terms with a

$$p \leq \frac{0.01}{30,000} = \frac{\textit{desired threshold}}{\textit{number of tests}}$$

- Since our tests are not all independent, this is very conservative, and will miss many true positives
- More sophisticated approaches exist, such as controlling the “false discovery rate”.

# P-values don't mean what you think!

## **The ASA's Statement on p-Values:**

### **Context, Process, and Purpose**

The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.