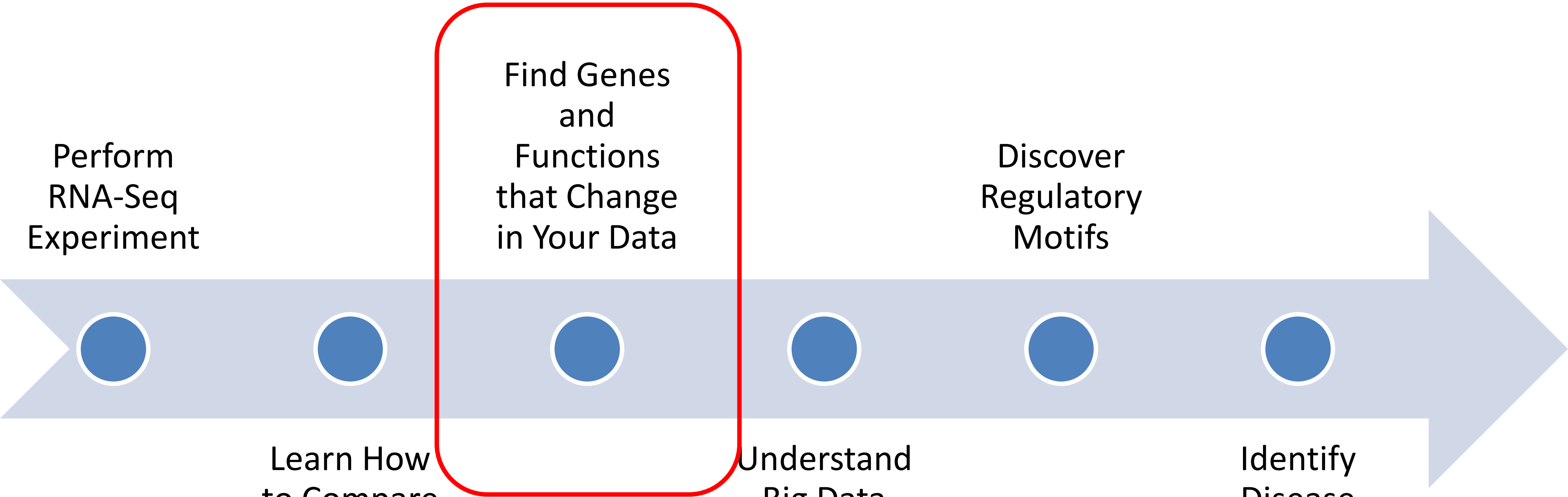


Perform RNA-Seq Experiment

Learn How to Compare Data

Find Genes and Functions that Change in Your Data

Understand Big Data Approaches

Discover Regulatory Motifs

Identify Disease Networks

Learning Objectives

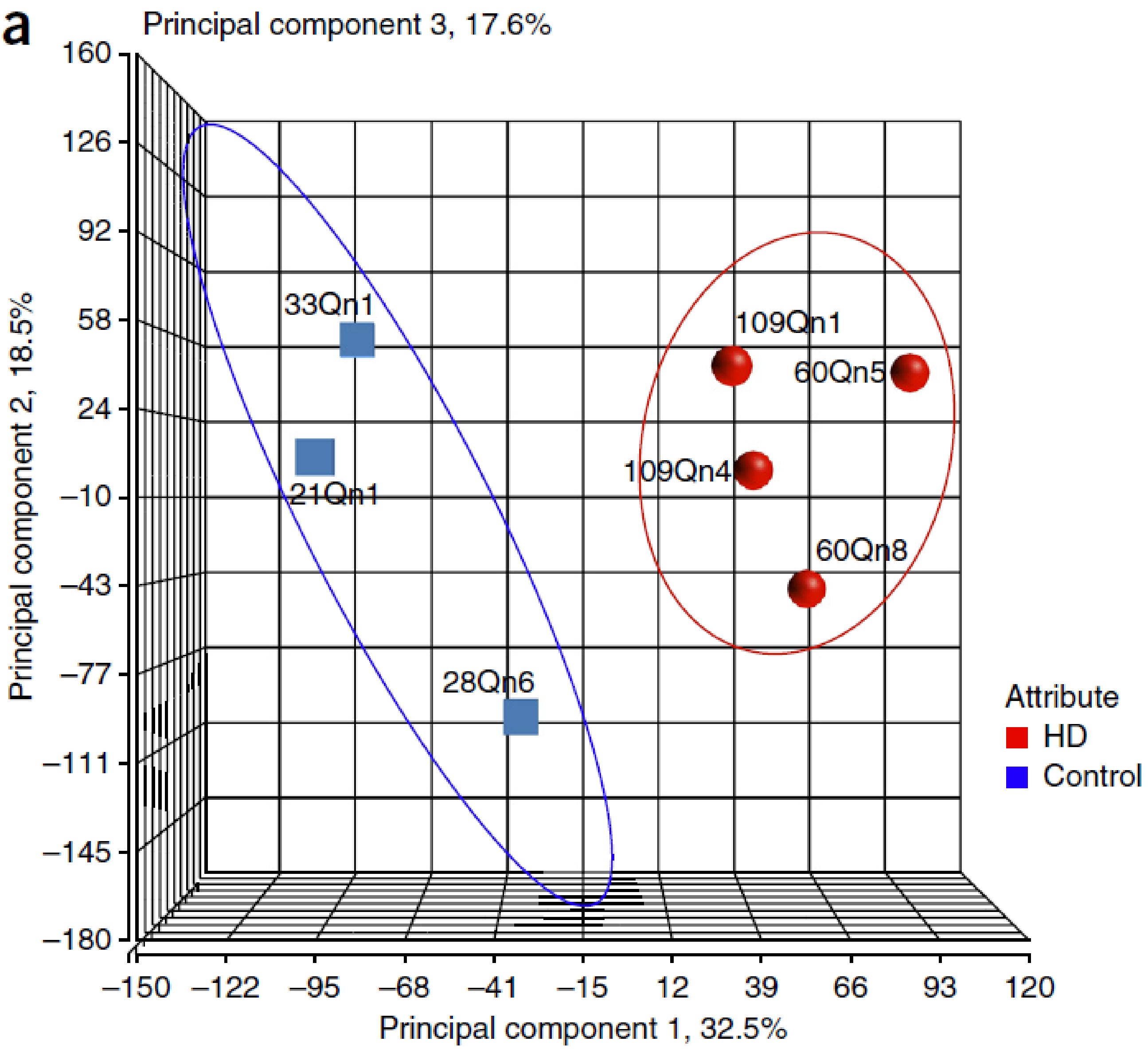
- Understand the basis of an RNA-Seq experiment
- Describe the steps from “raw reads” to gene counts
- Calculate RPKM values
- Explain the role of DESeq2
- Interpret Gene Ontology
- Evaluate statistical significance of GO terms for sets of genes

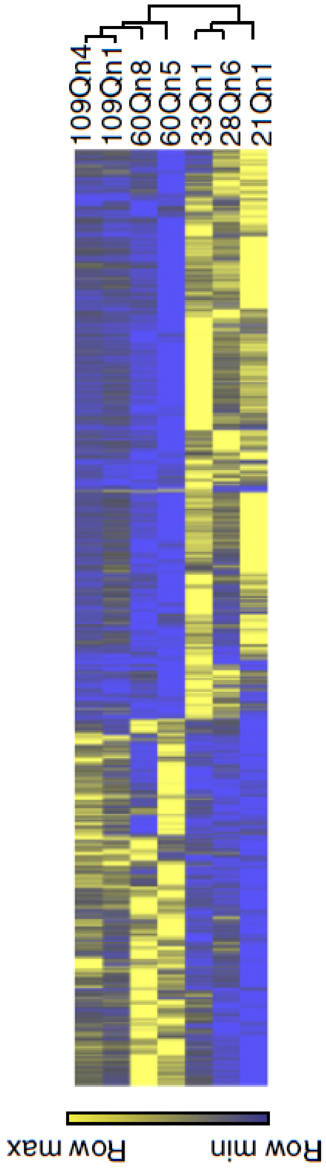
Last time:

- 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 dendogram
- Describe the results of Principal Component Analysis (PCA)

How could you visualize clusters in 20,000D instead of 2D?

Principal Component Analysis

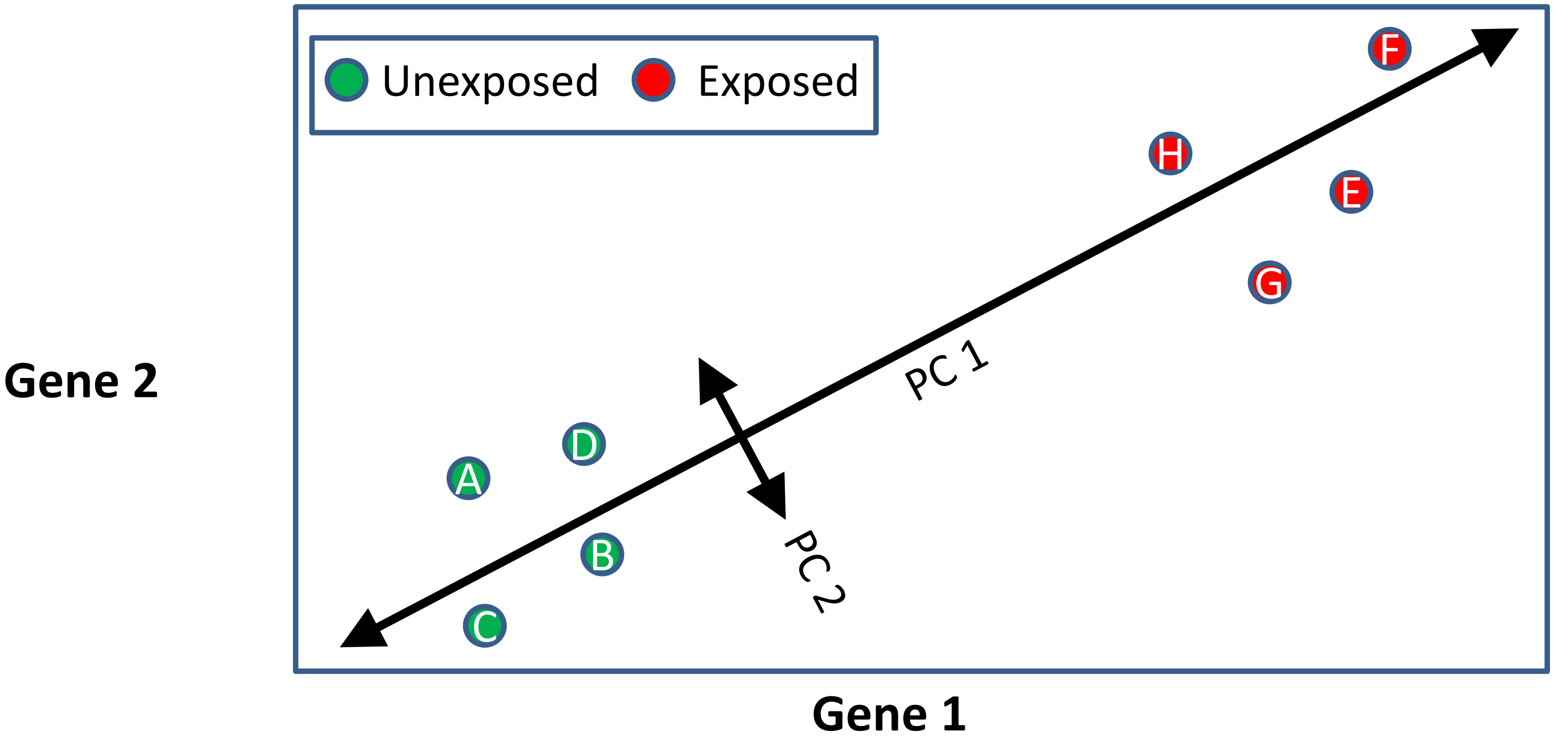




Principal Component Analysis

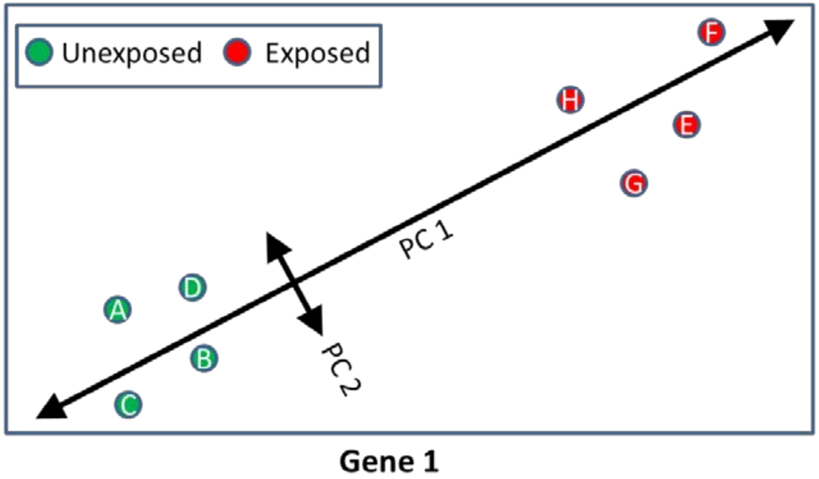
- Each sample is currently described by the expression of roughly 20,000 genes.
- Our goal:
  to find a 2-D or 3-D way to present the data that captures the greatest variance
  - Obviously, I could select any two genes, but they might be the wrong ones.
  - Can we find “interesting” linear combinations of genes?

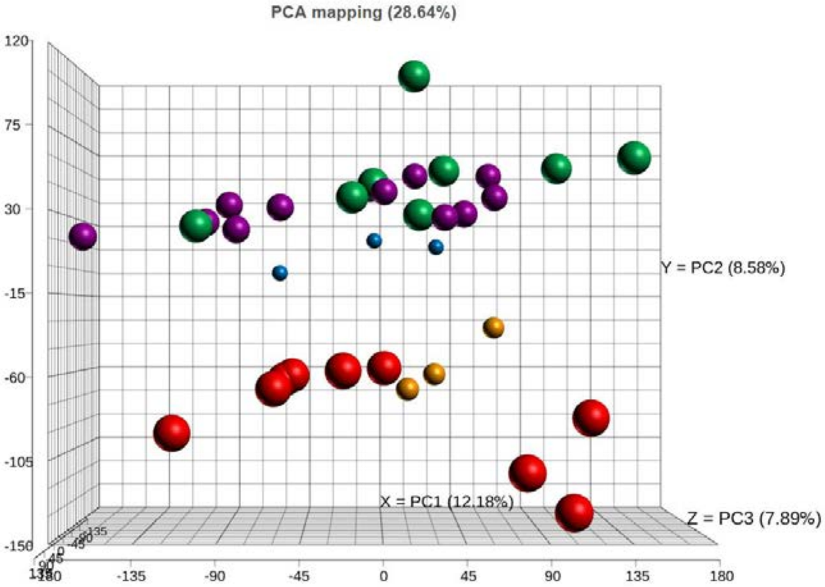
Principal Component Analysis



Goal: find a linear combination of the axes that captures most of the variation

1. PCA finds useful linear combinations of thousands of variables.
2. There are as many PCs as there were dimensions in the original data.
3. The PCs are orthogonal.
4. Often, a few PCs will capture most of the variance.

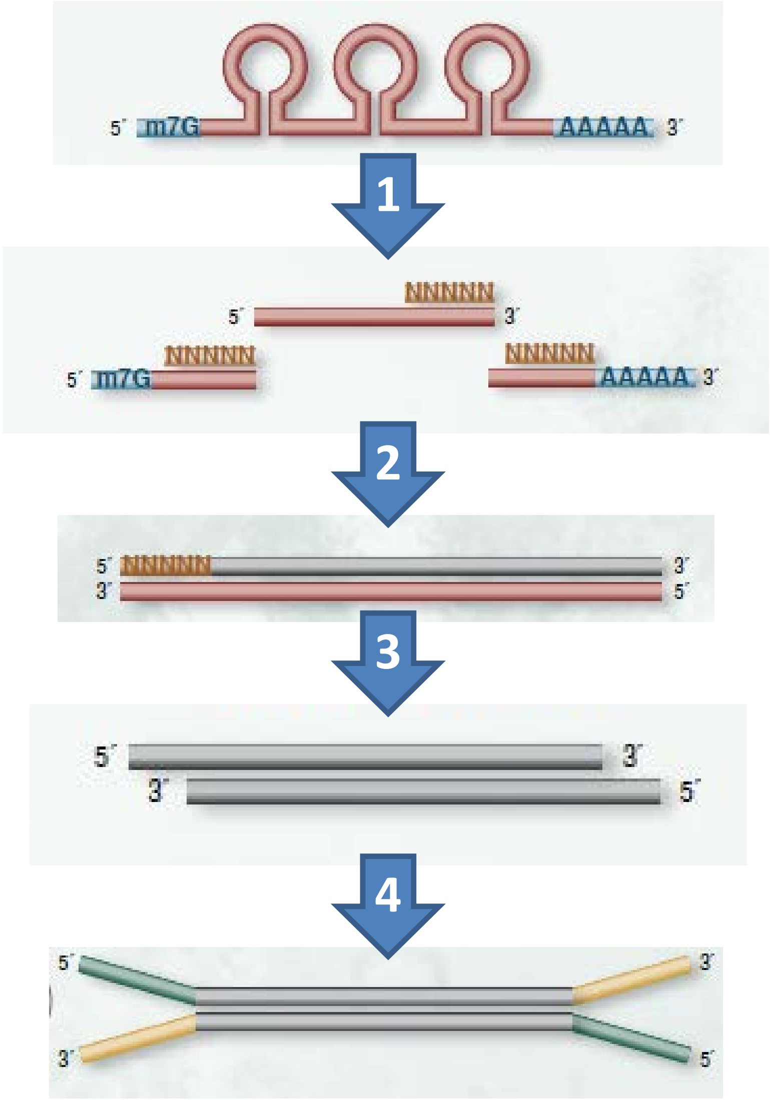




Outline

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

- Use column purification to separate DNA/RNA
- Need to be very careful to avoid RNAse
- We must separate ribosomal RNA: rRNA >90% of cellular RNA. mRNA ~2%
- Sequencing machines
  - work on DNA, not RNA
  - Are best for short fragments

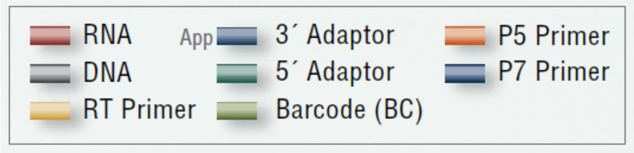


1. Fragment RNA and prime with random DNA primers

2. Synthesize second strand with Reverse Transcriptase

3. Remove RNA and synthesize second strand of DNA

4. Ligate adaptors for sequencing



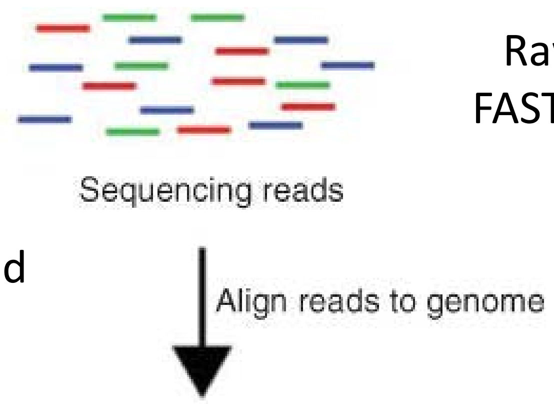
NEBNext® for Illumina®

NGS SAMPLE PREPARATION

Outline

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

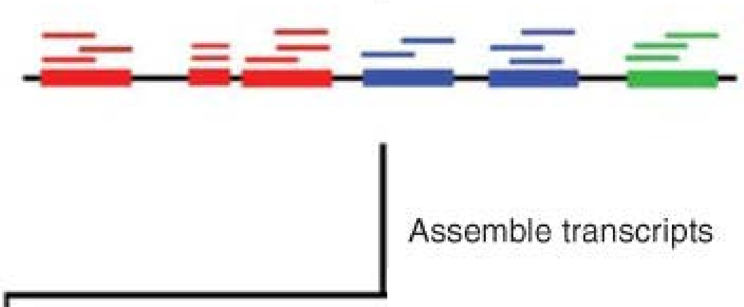
From Raw Sequence to Expression Levels



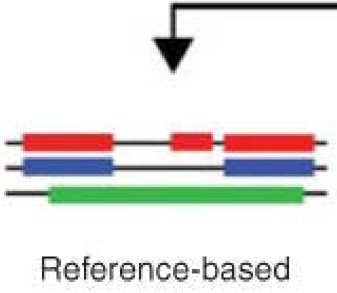
Raw reads
FASTA, FASTQ

Fragments get sequenced
“reads”

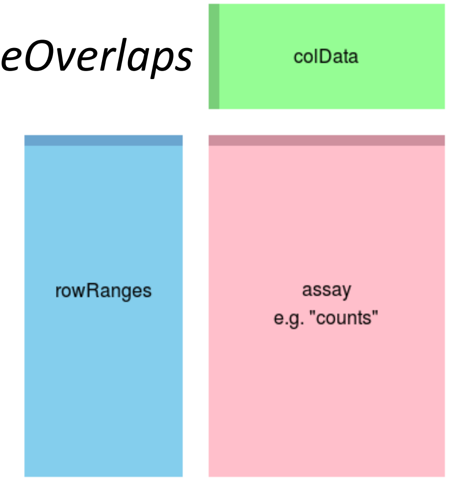
Align to genome
TopHat2



Mapped Reads
SAM, BAM

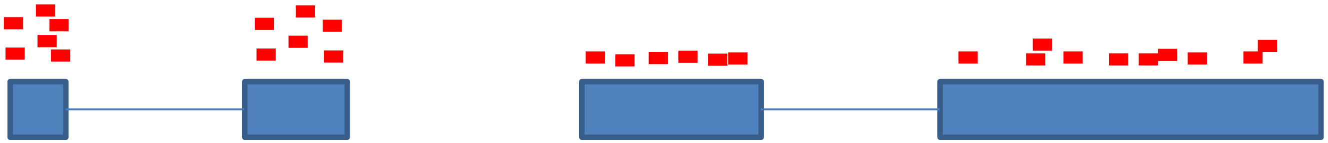


summarizeOverlaps



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)”

Raw reads

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
Total reads		3.0E6	3.7E6	9.1E6
Reads/1,000,000		3	3.7	9.1

1. 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

2. Divide reads for a gene by the number of reads in the replicate (in millions)

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

3. Divide by gene length in kilobases

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
Total reads		3.0E6	3.7E6	9.1E6
Reads/1,000,000		3	3.7	9.1

Reads per million

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

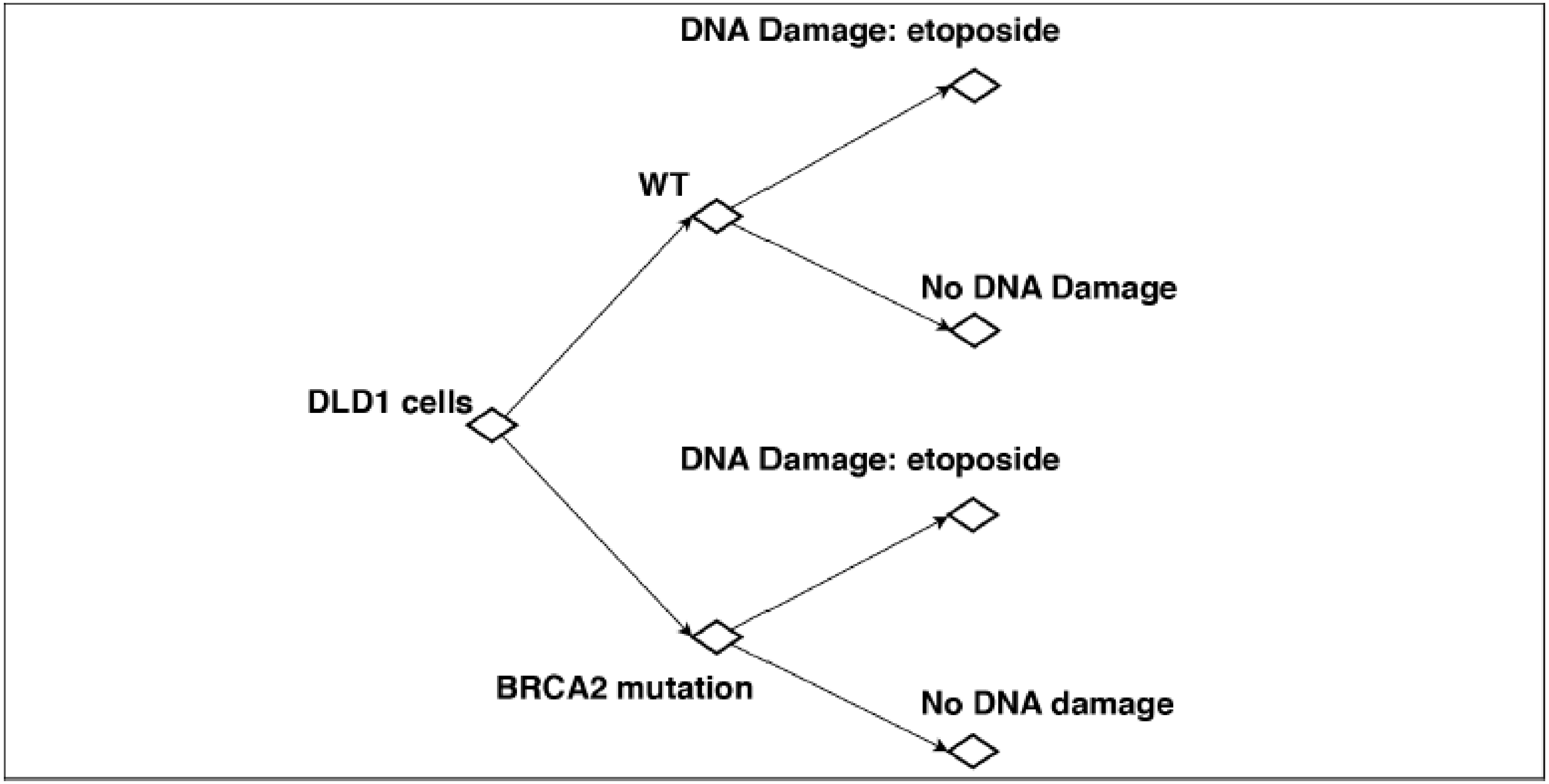
This step corrects for sequencing depth.
Note that numbers are now more consistent across replicates

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



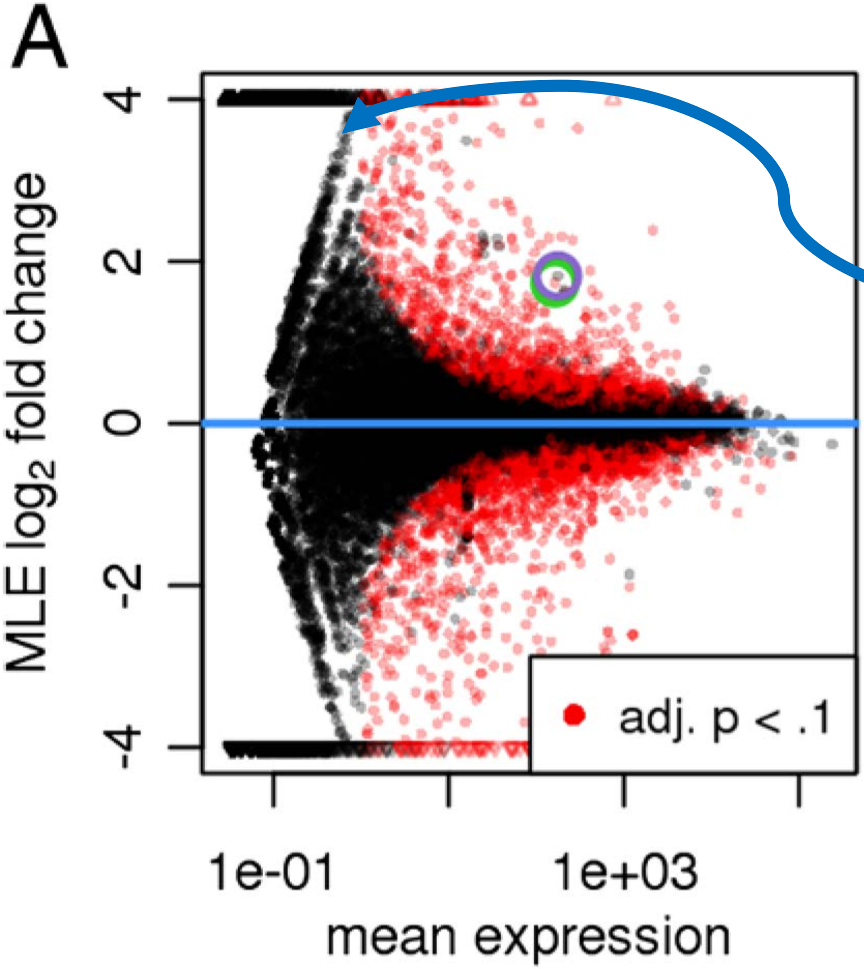
Unfortunately, we can’t just compare RPKM values across conditions.
Random sampling errors will produce different values even for genes that are expressed at a constant level.

Differential Expression

- We want to test the null hypothesis that the log-fold-change is zero.
- We also want to be careful not to over-interpret very small changes that are statistically significant.
- In the lab, you will use DESeq2 to address these questions

Heteroskedasticity

variance of LFCs depends on the mean



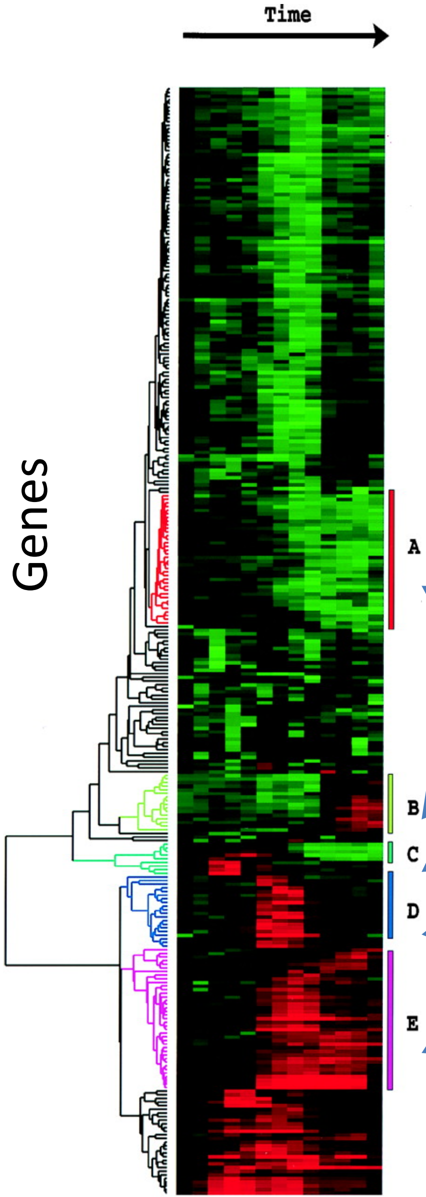
Love *et al. Genome Biology* (2014) 15:550

- 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



How did they figure out what the clusters of genes did?

(A) cholesterol biosynthesis

(B) the cell cycle

(C) the immediate-early response

(D) signaling and angiogenesis

(E) wound healing and tissue remodeling

Iyer et al. *Science* 1999

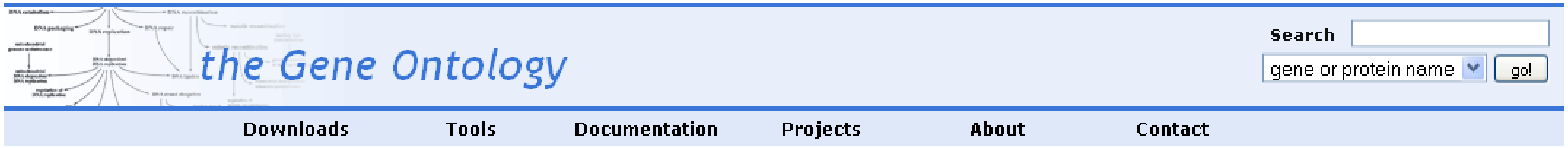
Outline

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

Biological Insights

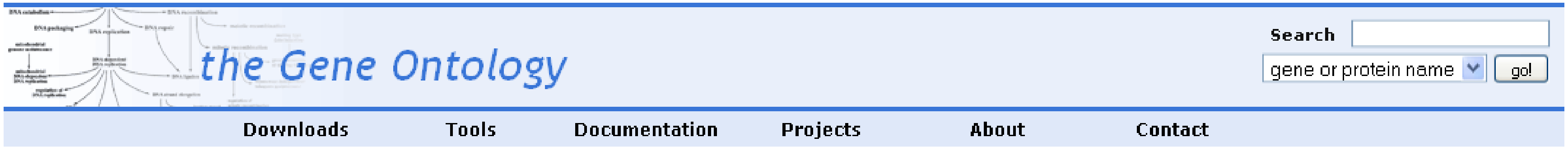
- What types of genes are being differentially expressed?

http://www.geneontology.org

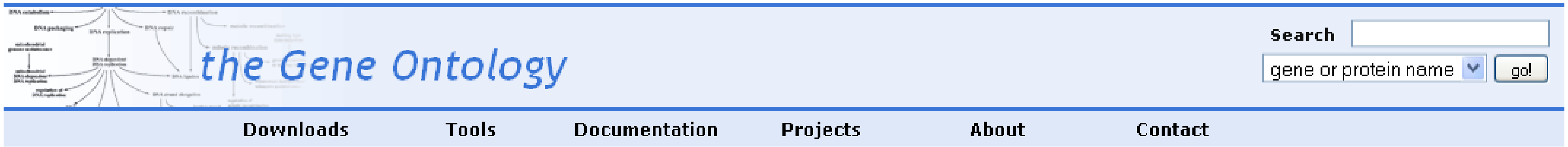


Controlled vocabulary to describe genes:

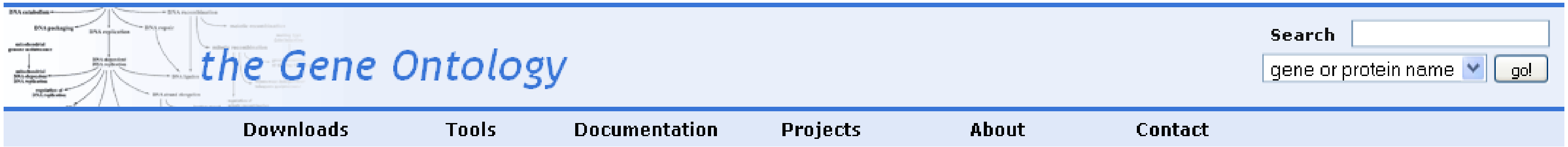
- Biological process
- Cellular component
- Molecular function



- Biological process
  - signal transduction; glucose tranport
- 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.



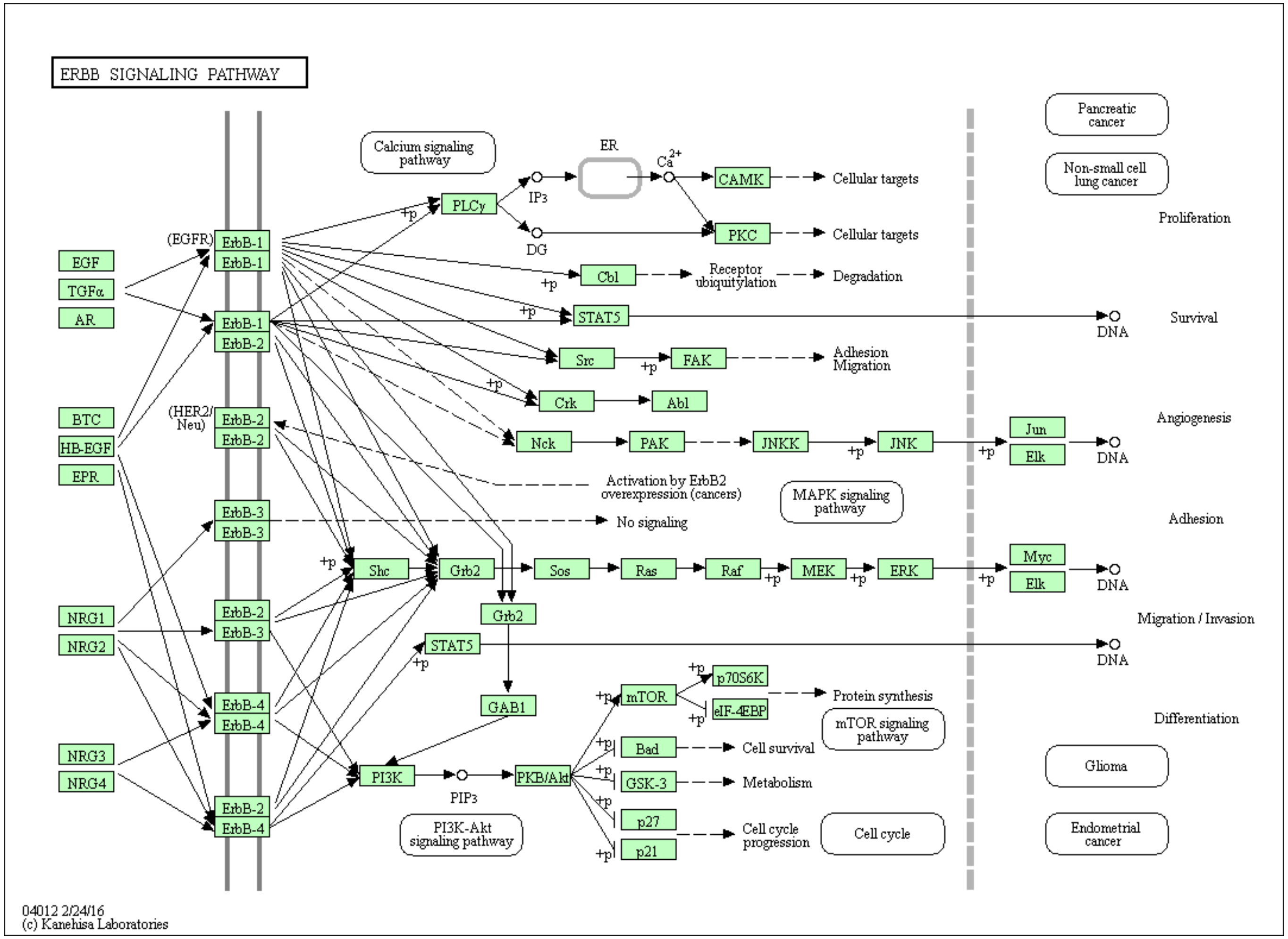
- **Biological process**
- A biological process is not equivalent to a pathway.
  - Does not represent the dynamics or dependencies of a pathway.

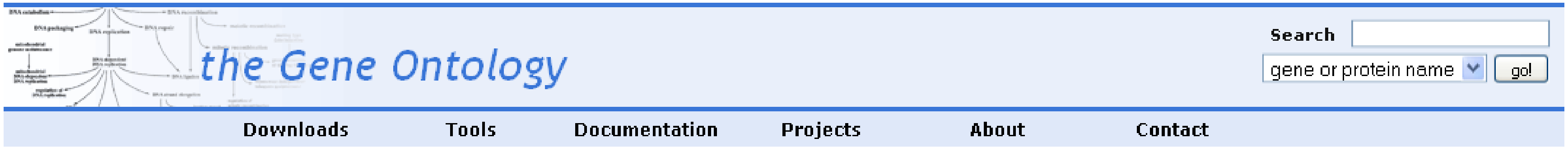
GO

BTC
CDC37
Cpne3
CPNE3
CUL5
EGF
EGFR
ERBB2
ERBB3
ERBB4
ERBIN
EREG
GAB1
GRB2
GRB7
HBEGF
HRAS
HSP90AA1
KRAS
MATK
Myoc
MYOC

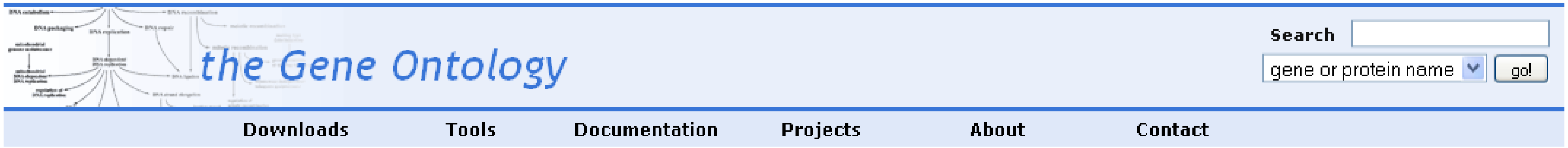
NRAS
NRG1
NRG2
NRG4
PIK3CA
PIK3R1
PRKCA
PTK6
PTPN12
PTPN18
Ptprr
PTPRR
RPS27A
SHC1
SOS1
SRC
STUB1
Symbol
UBA52
UBB
UBC

KEGG Pathway

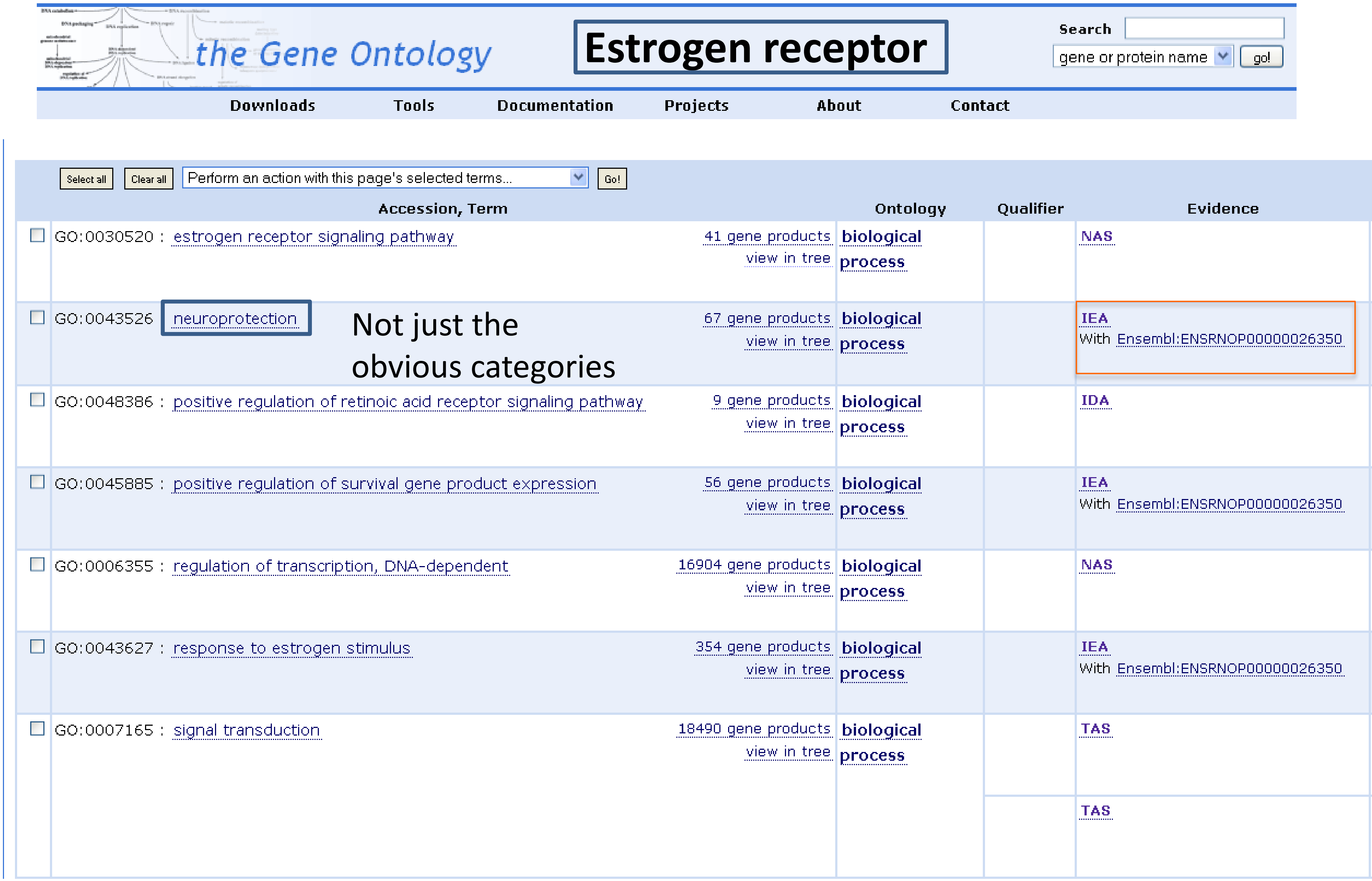




- **Cellular component**
- Part of a
  - anatomical structure (e.g. **rough endoplasmic reticulum or nucleus**) or a
  - gene product group (e.g. **ribosome, proteasome or a protein dimer**).



- **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**.

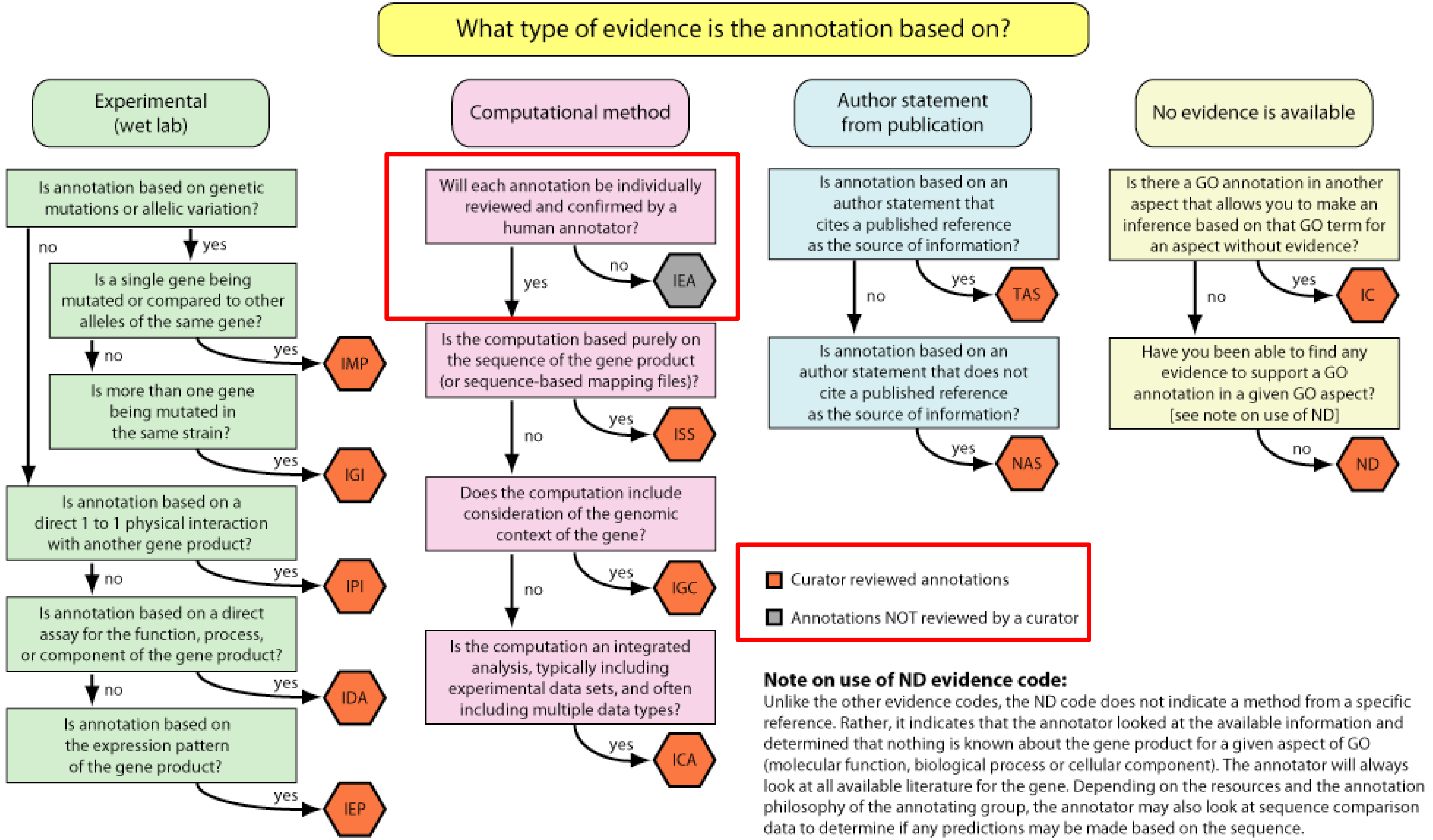


Estrogen receptor

Not just the obvious categories

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

GO Evidence Code Decision Tree



What type of evidence is the annotation based on?

Experimental (wet lab)

Is annotation based on genetic mutations or allelic variation?

no

yes

Is a single gene being mutated or compared to other alleles of the same gene?

no

yes

IMP

Is more than one gene being mutated in the same strain?

yes

IGI

Is annotation based on a direct 1 to 1 physical interaction with another gene product?

no

yes

IPI

Is annotation based on a direct assay for the function, process, or component of the gene product?

no

yes

IDA

Is annotation based on the expression pattern of the gene product?

yes

IEP

Computational method

Will each annotation be individually reviewed and confirmed by a human annotator?

yes

no

IEA

Is the computation based purely on the sequence of the gene product (or sequence-based mapping files)?

no

yes

ISS

Does the computation include consideration of the genomic context of the gene?

no

yes

IGC

Is the computation an integrated analysis, typically including experimental data sets, and often including multiple data types?

yes

ICA

Author statement from publication

Is annotation based on an author statement that cites a published reference as the source of information?

no

yes

TAS

Is annotation based on an author statement that does not cite a published reference as the source of information?

yes

NAS

No evidence is available

Is there a GO annotation in another aspect that allows you to make an inference based on that GO term for an aspect without evidence?

no

yes

IC

Have you been able to find any evidence to support a GO annotation in a given GO aspect? [see note on use of ND]

no

ND

🟧 Curator reviewed annotations

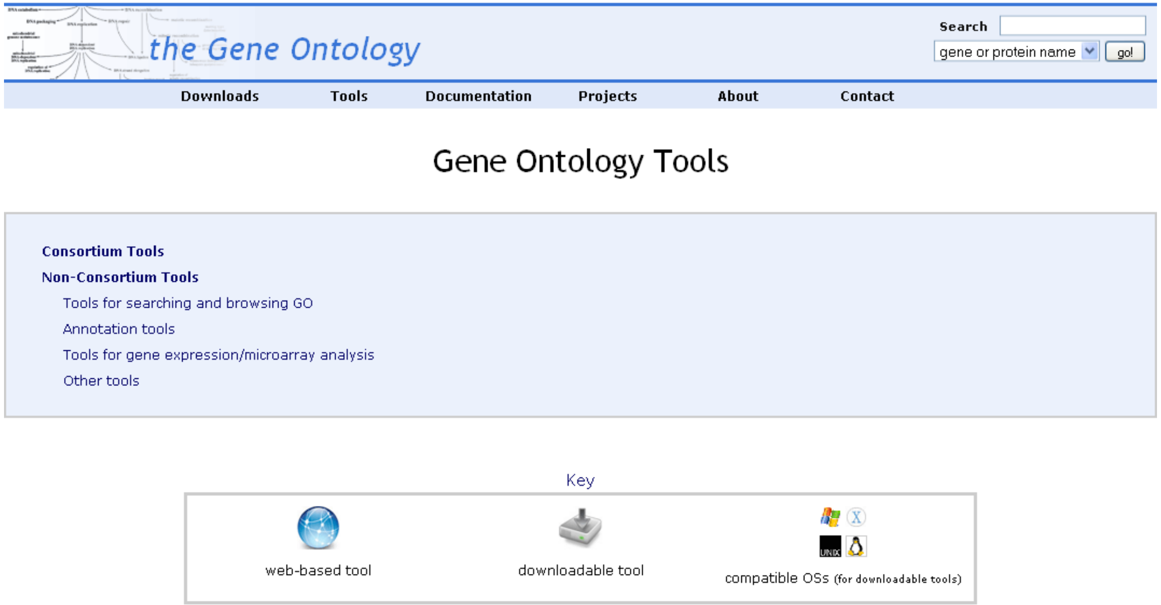
▨ Annotations NOT reviewed by a curator

Note on use of ND evidence code:

Unlike the other evidence codes, the ND code does not indicate a method from a specific reference. Rather, it indicates that the annotator looked at the available information and determined that nothing is known about the gene product for a given aspect of GO (molecular function, biological process or cellular component). The annotator will always look at all available literature for the gene. Depending on the resources and the annotation philosophy of the annotating group, the annotator may also look at sequence comparison data to determine if any predictions may be made based on the sequence.

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

web-based tool

downloadable tool

compatible OSs (for downloadable tools)

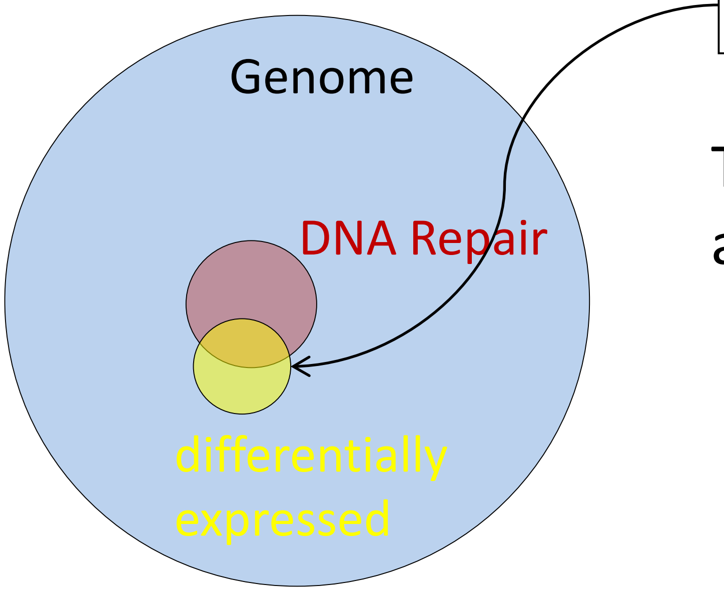
Outline

- 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

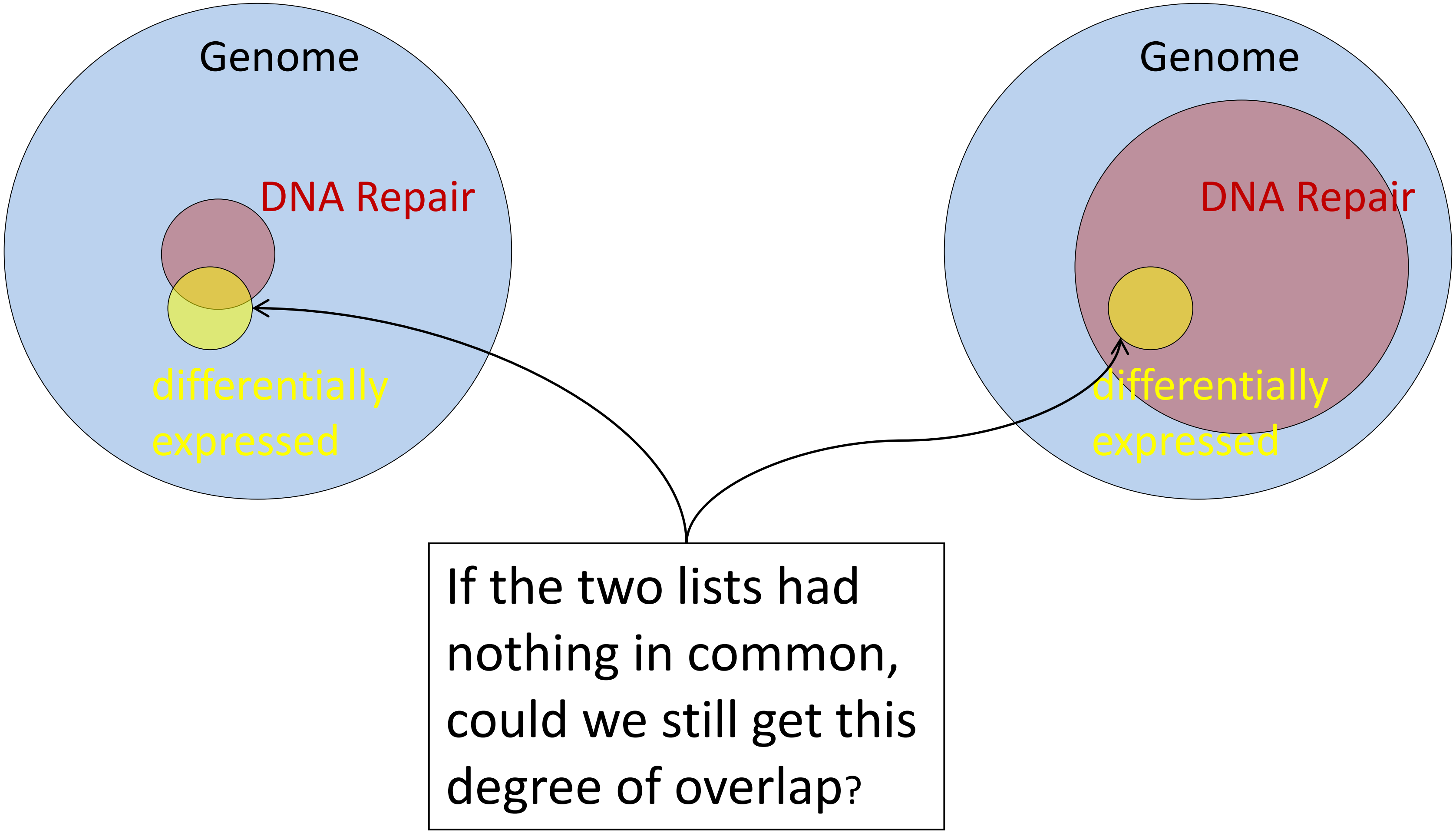


Is this overlap significant?

To answer this question we need a null model.

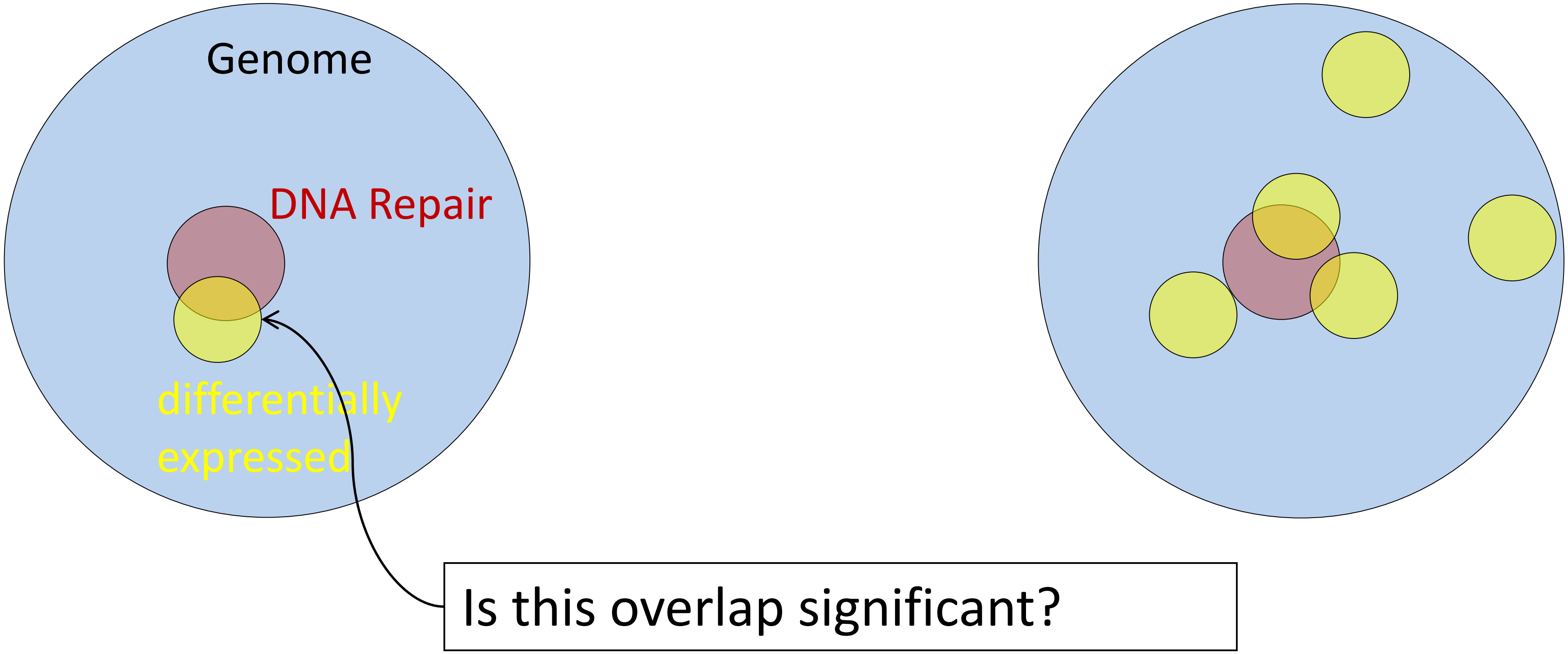
Statistical significance

The significance depends on the size of the lists.



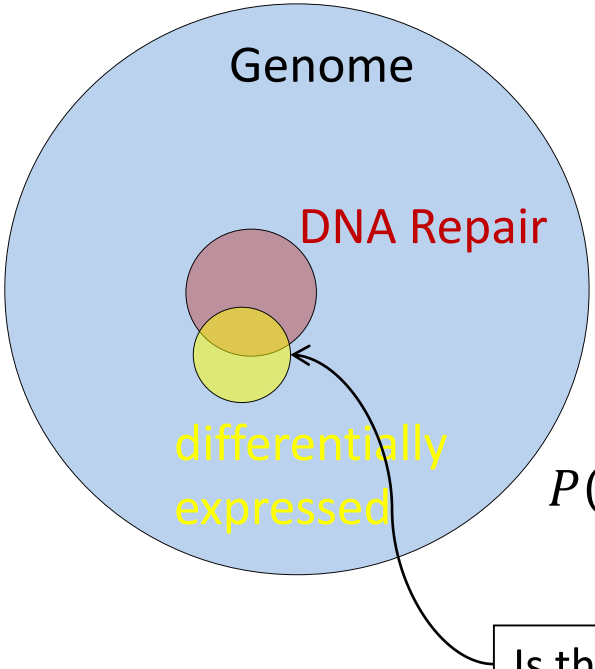
If the two lists had nothing in common, could we still get this degree of overlap?

Statistical significance



Is this overlap significant?

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:

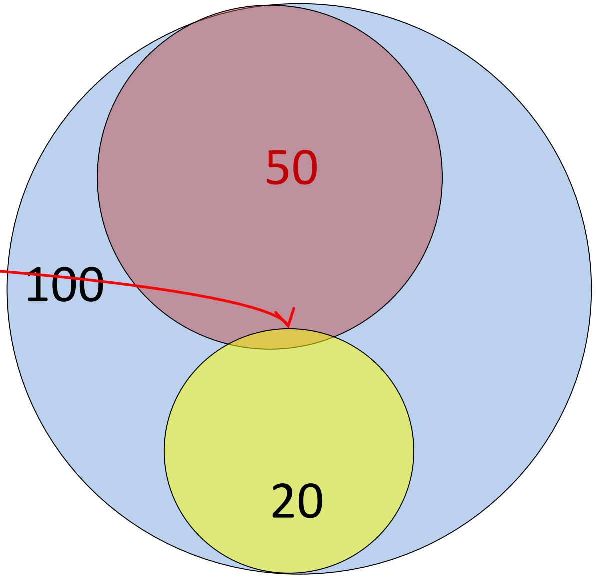
$$P(Overlap) = \frac{\binom{DNA\,repair}{Overlap}\binom{Genome - DNA\,repair}{DiffExp - Overlap}}{\binom{Genome}{DiffExp}}$$

Is this overlap significant?

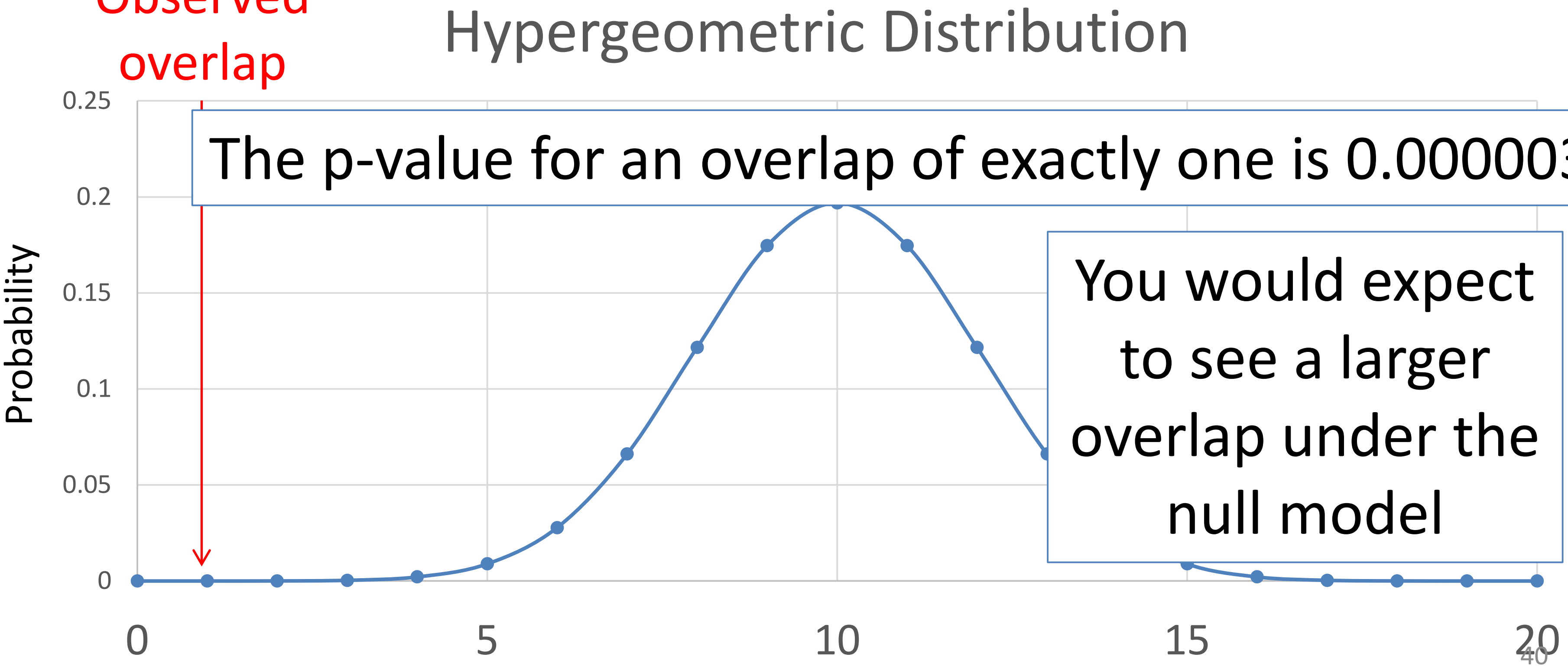
Recall that $\left(\frac{n}{k}\right)$ ("n choose k") is the binomial coefficient.

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

There is only one overlapping gene.
Is that surprising?



Observed overlap

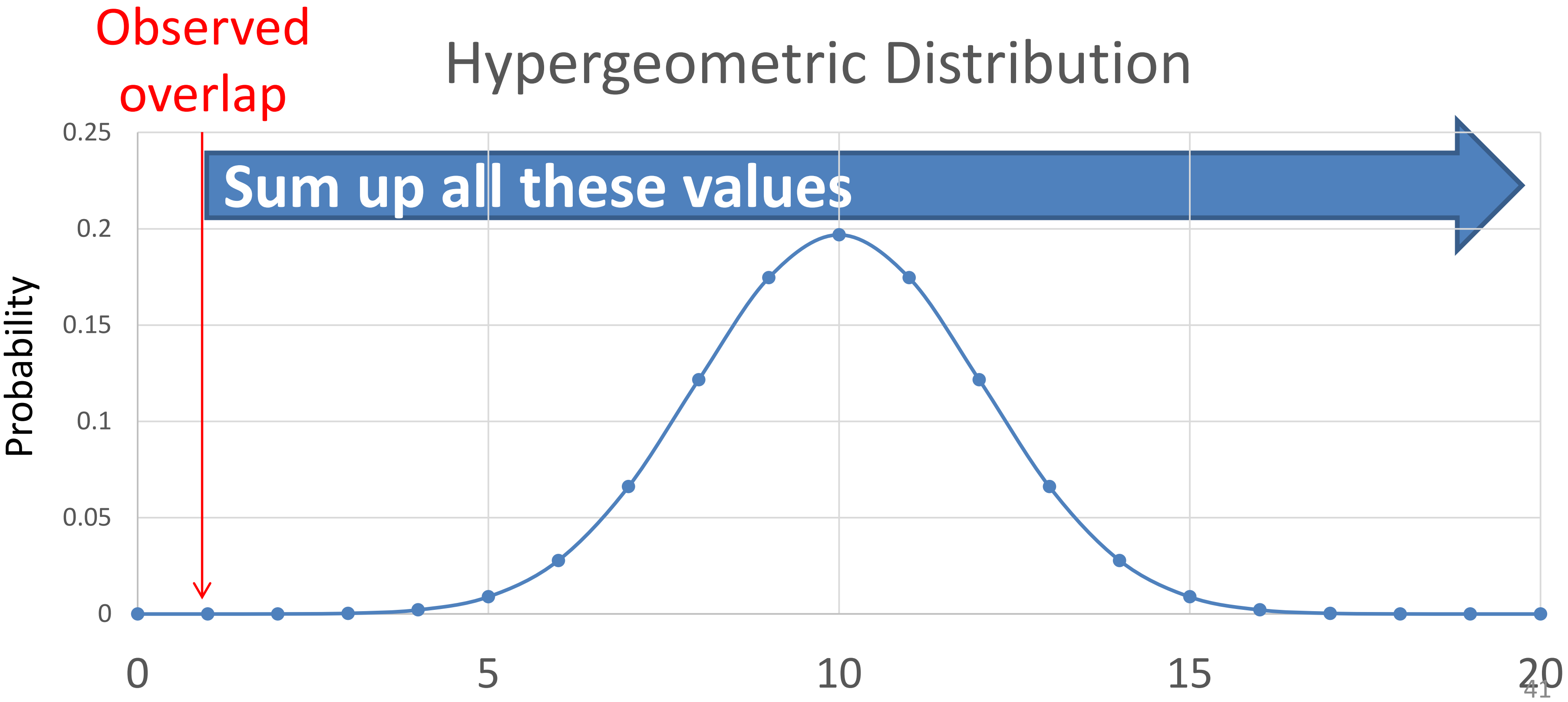


The p-value for an overlap of exactly one is 0.000003.

You would expect to see a larger overlap under the null model

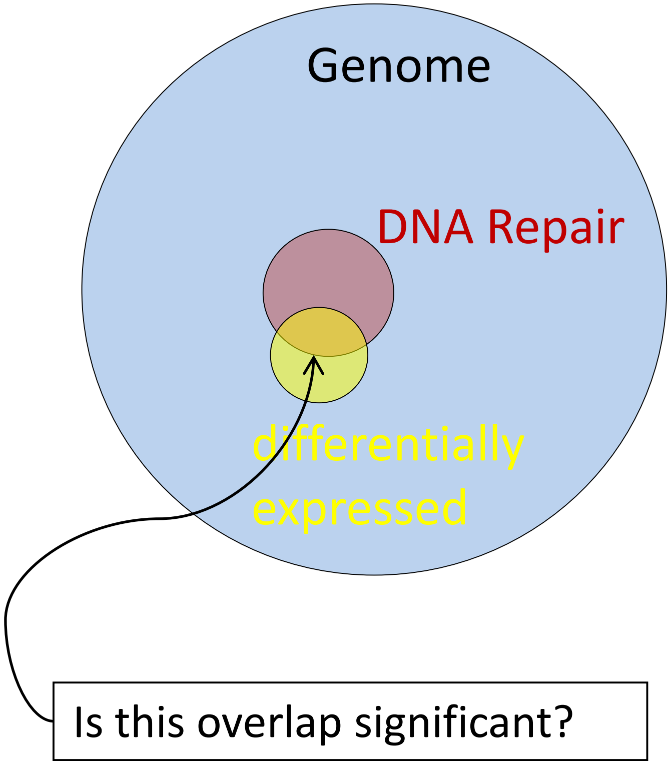
The CDF helps us find enriched terms

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



CDF=Cumulative distribution function

Statistical significance



- We wish to test if a term is “enriched” in our data.
- But the hypergeometric gives the probability of getting **exactly** this amount of overlap for two randomly chosen sets of genes of the same size.
- Using the CDF, we can ask if we see ***more*** of a term than we would expect under the null model.