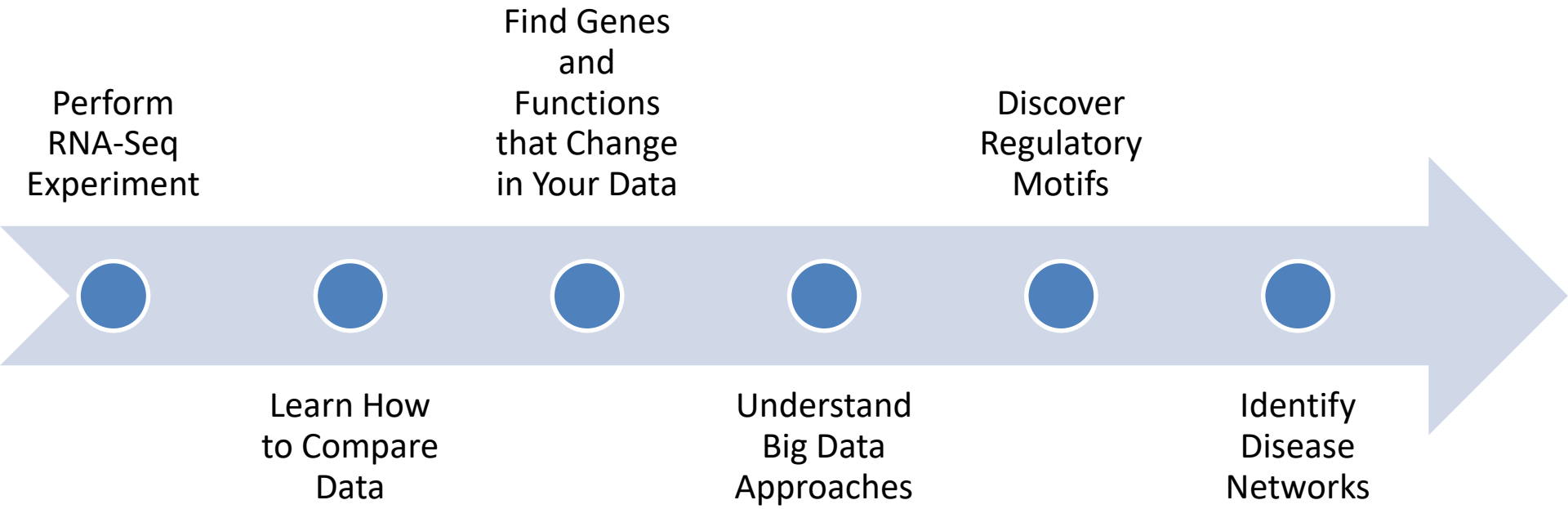
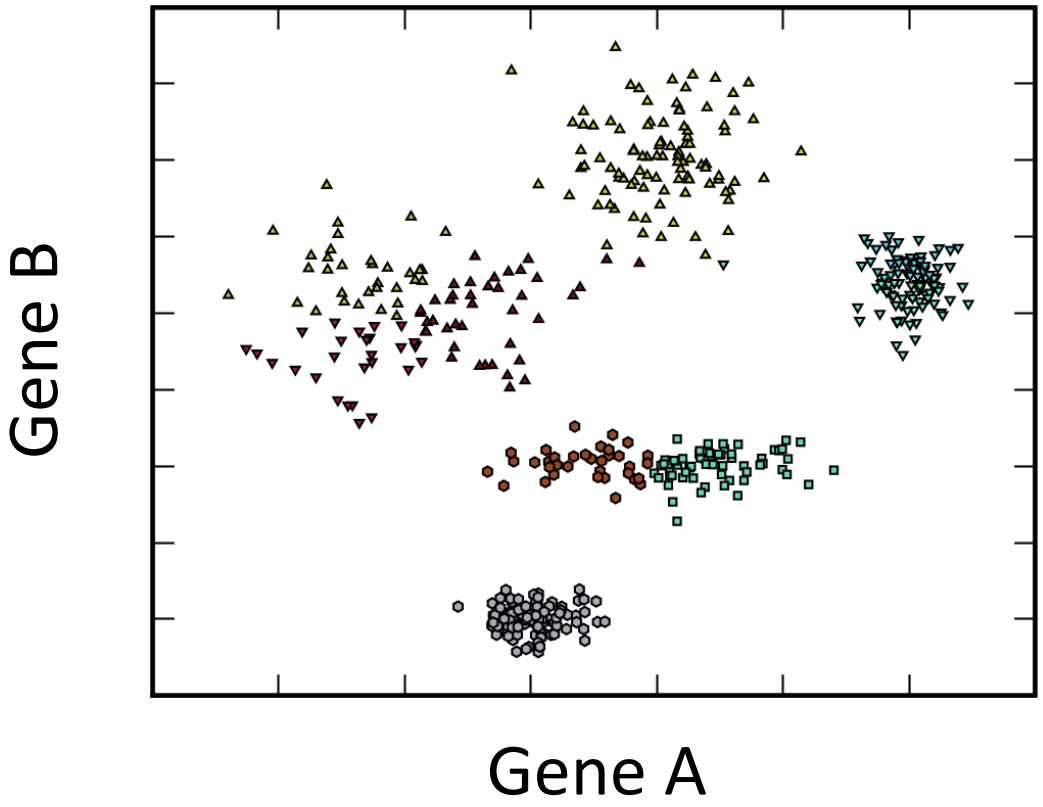

Perform RNA-Seq Experiment
Learn How to Compare Data
Find Genes and Functions that Change in Your Data
Understand Big Data Approaches
Discover Regulatory Motifs
Identify Disease Networks

Write on Board Before Class:

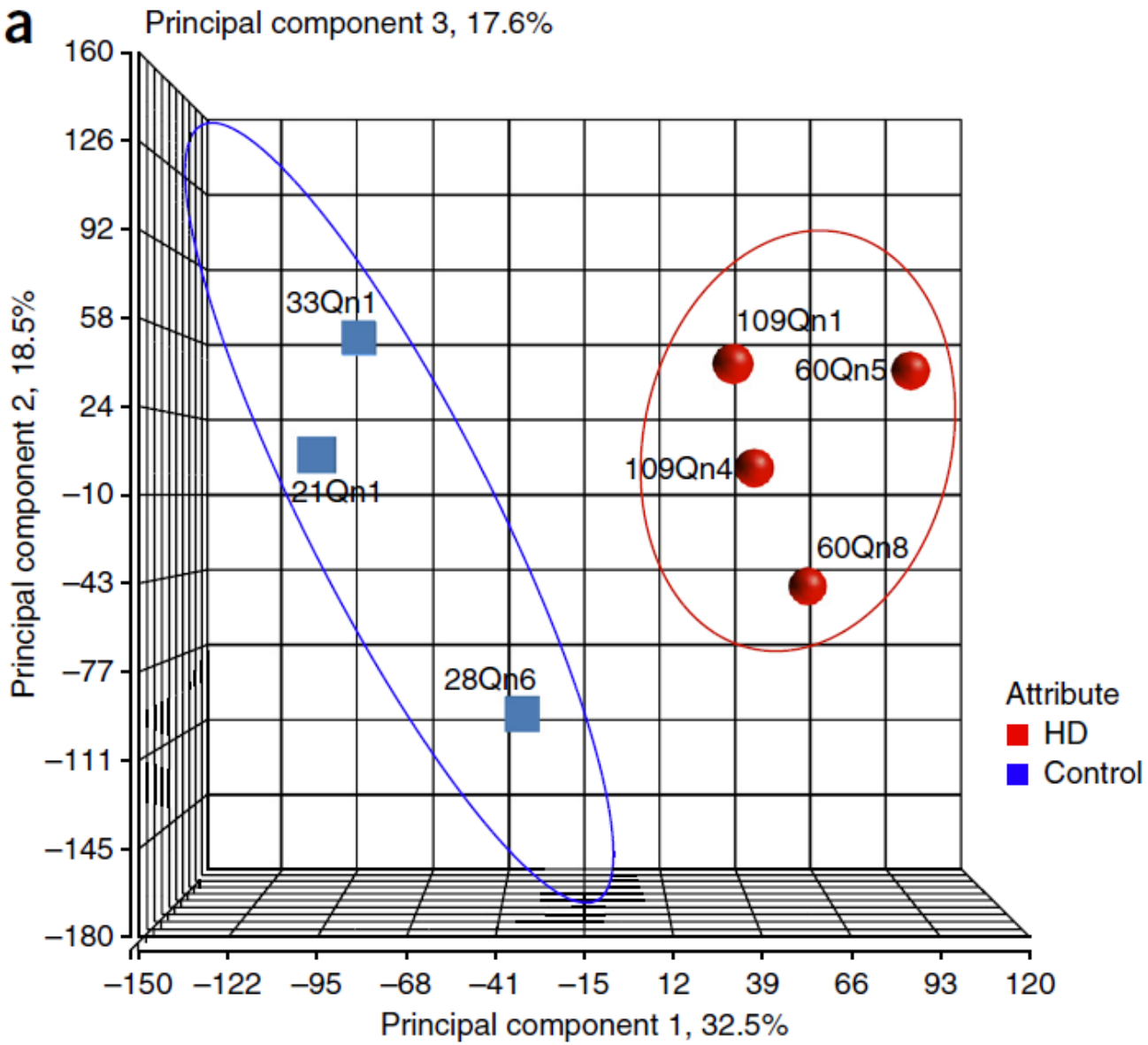
Learning Objectives

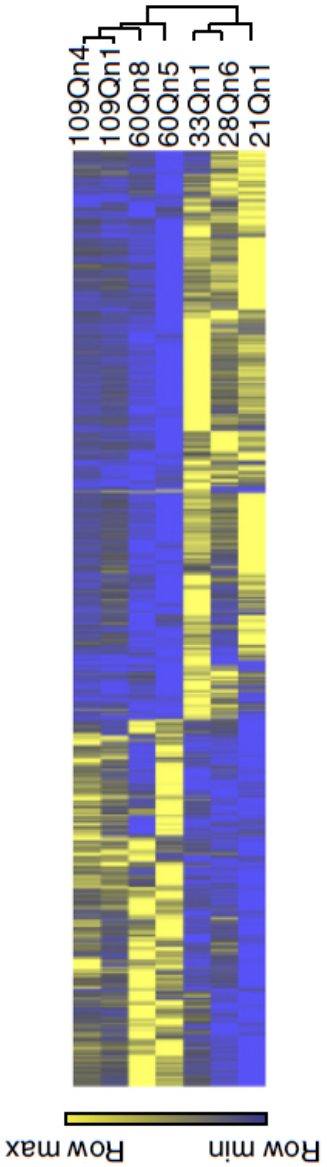
- Describe the results of Principal Component Analysis (PCA)
- Understand the basis of an RNA-Seq experiment
- Describe the steps from “raw reads” to gene counts
- Calculate RPKM values
- Explain the role of DESeq2
- Interpret Gene Ontology
- Evaluate statistical significance of GO terms for sets of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

Clustering works well for 2D data. But how could you visualize clusters in 20,000D?



The basics of PCA

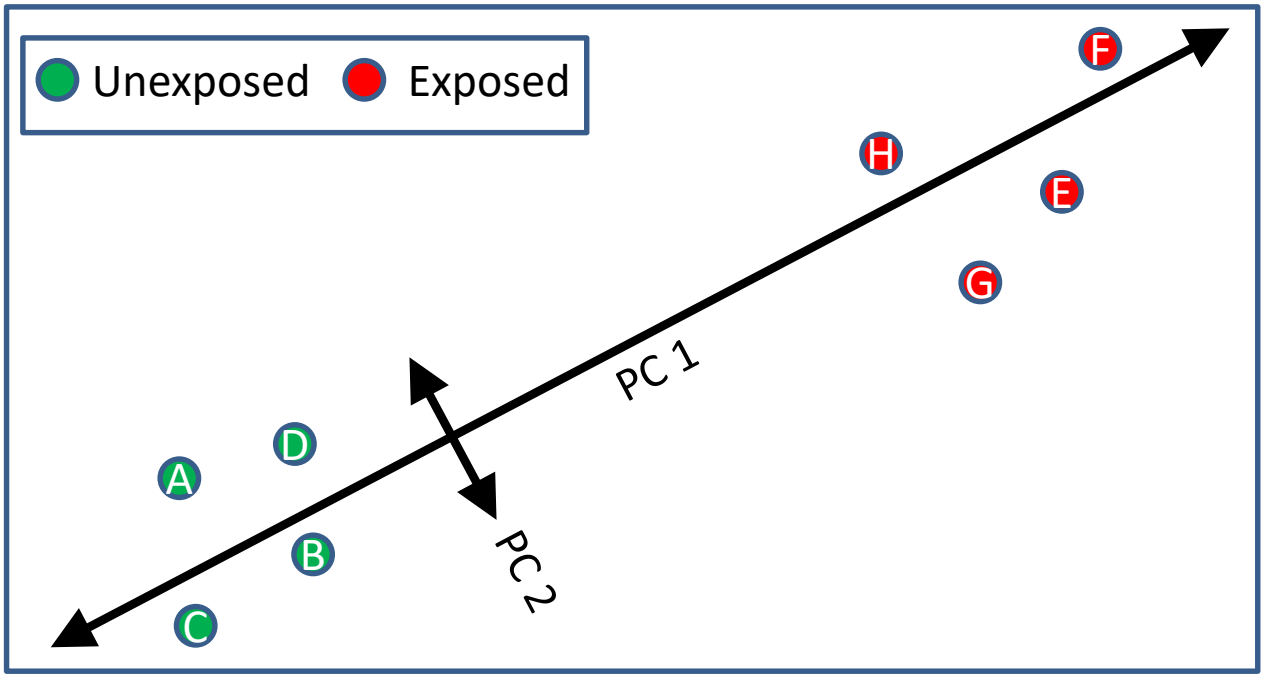




Principal Component Analysis

- 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?

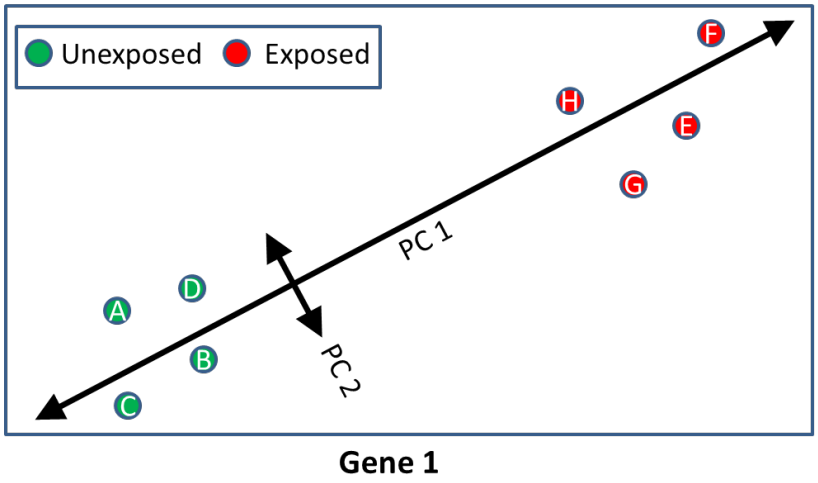
Principal Component Analysis

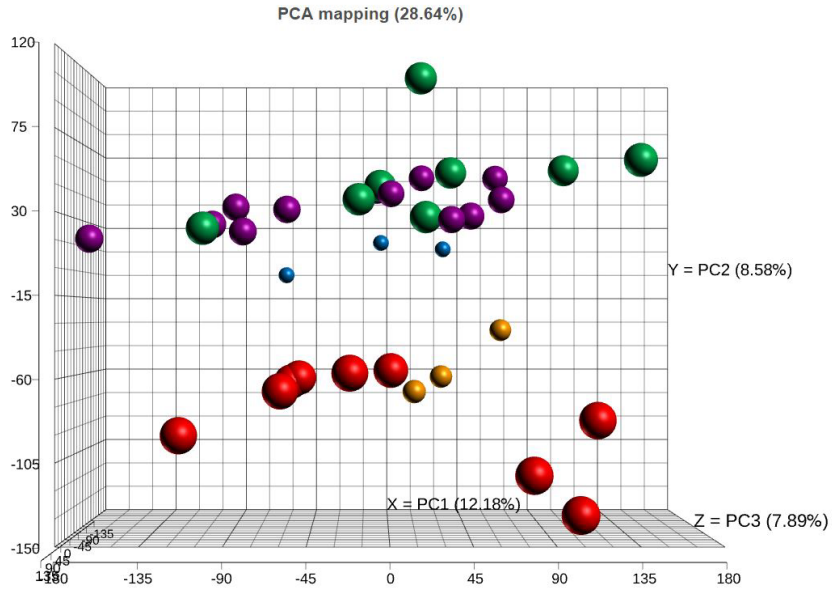


Here's an example where one dimension is almost as good as two.

We can generalize this approach so that one dimension is almost as good as N, where N is large.

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.



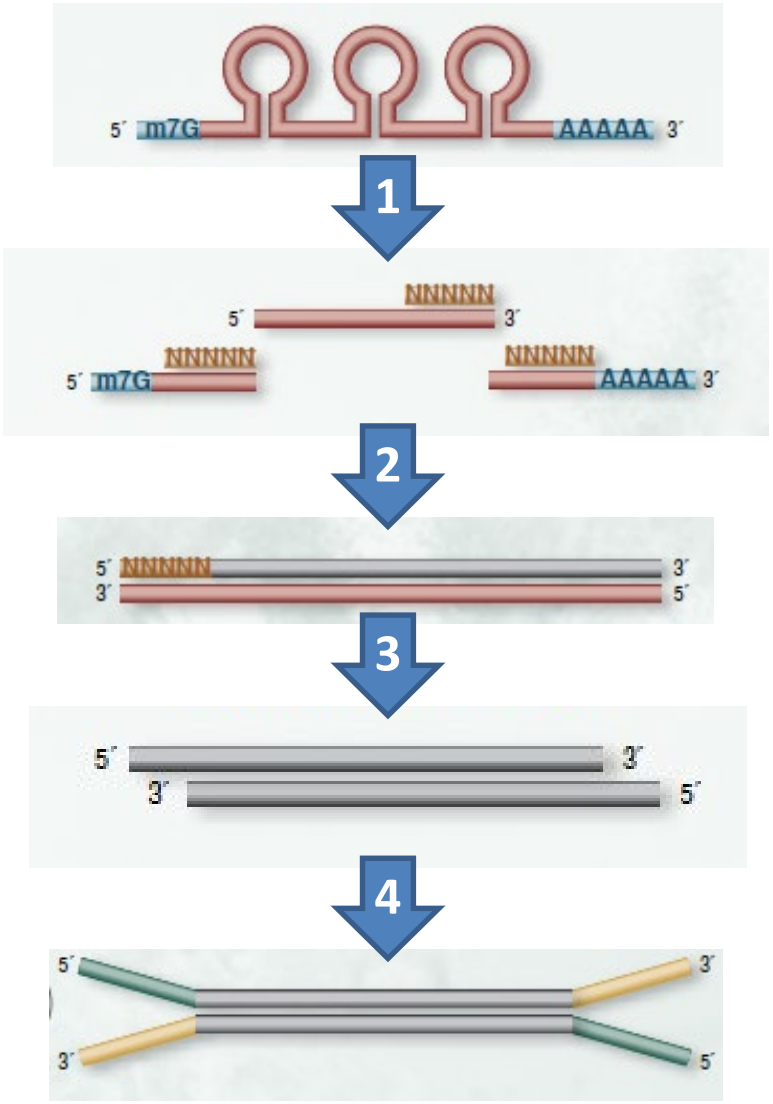


Outline

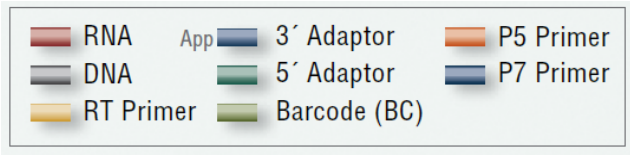
- Overview of the steps of RNA-Seq
- Deriving expression levels from sequence data
- Gene Ontology
- Statistical significance

Experimental Design for RNA-Seq

- Goal of RNA-Seq:
  - To measure the expression of all genes in a sample
- Sequencing machines are great but have limitations:
  - They work on DNA, not RNA
  - They 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



Outline

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

Fragments get sequenced
"reads"

Sequencing reads

Raw reads
FASTA, FASTQ

Align to genome
TopHat2

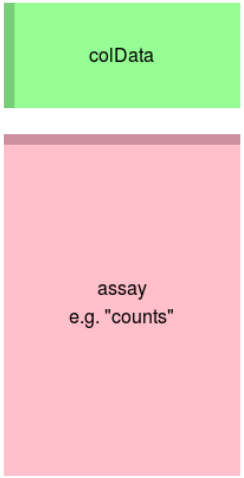
Align reads to genome

Mapped Reads
SAM, BAM

Assemble transcripts

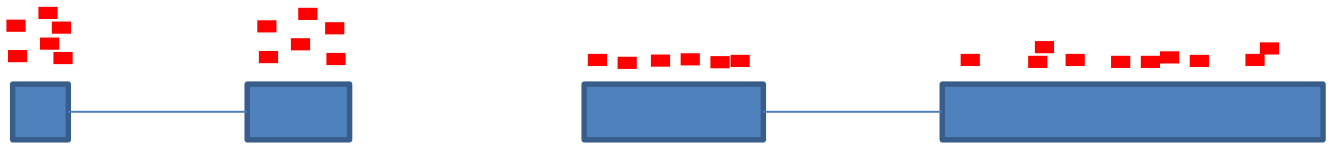
Reference-based

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

Raw reads

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)

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

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

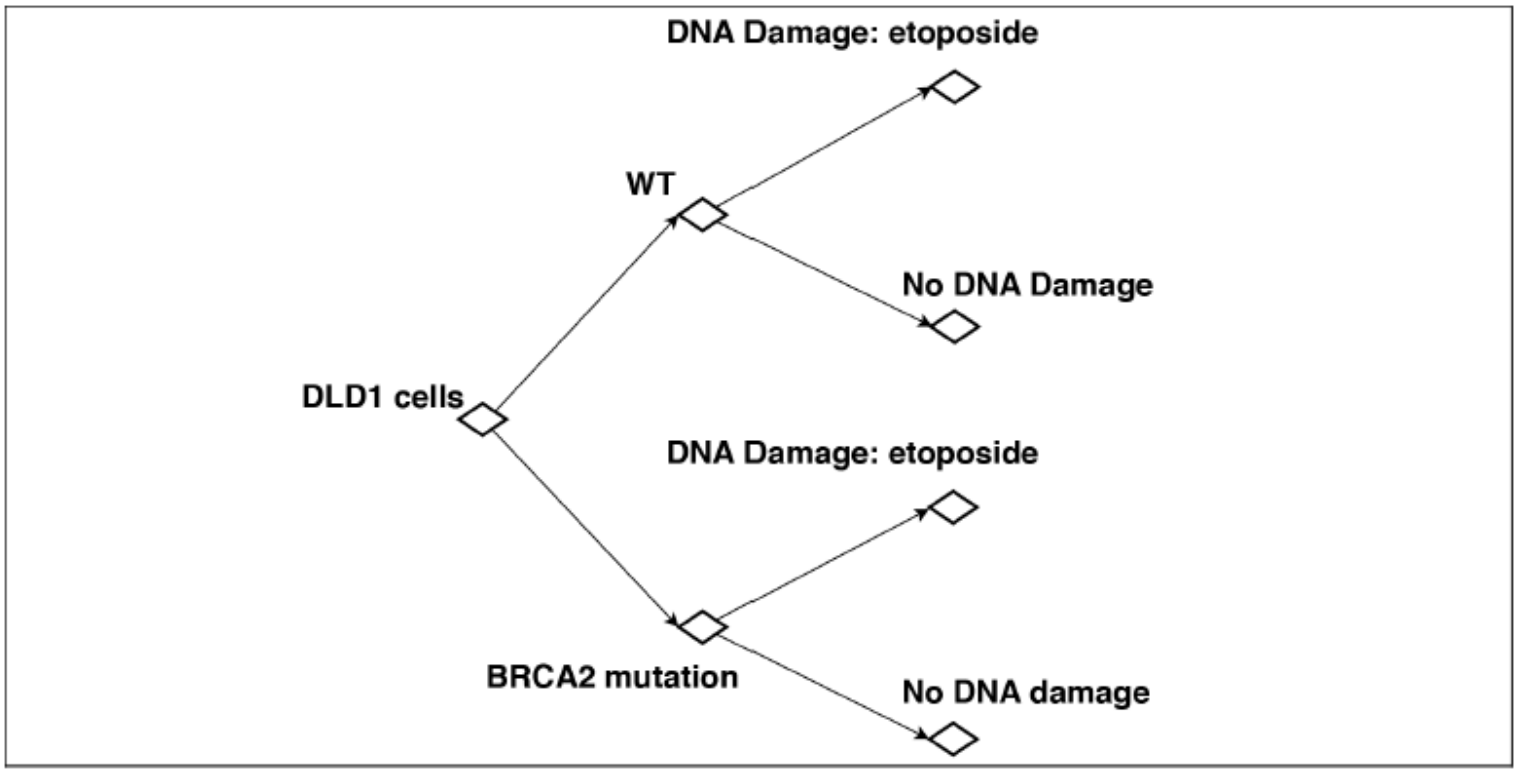
Reads per kilobase million RPKM

	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

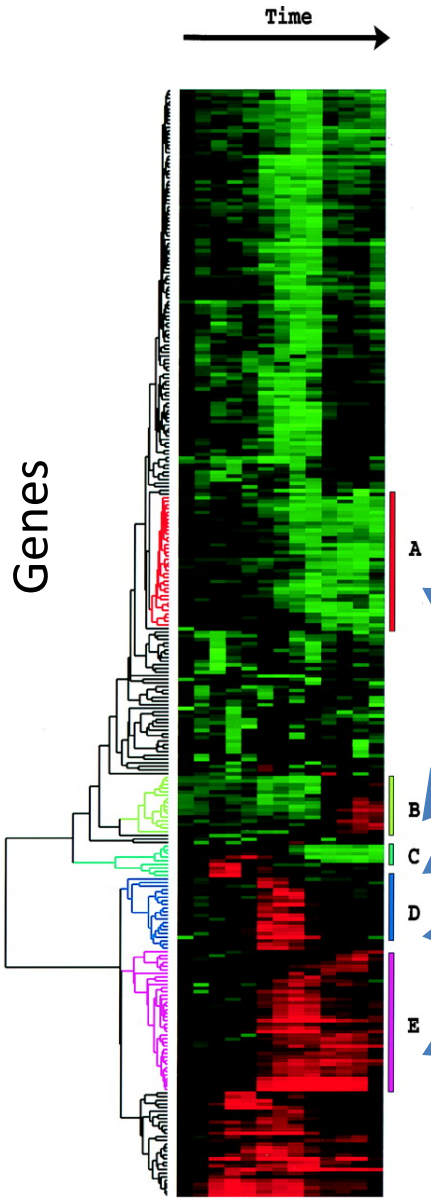
DESeq2: tests whether a difference in gene expression is a response to a change in condition vs. a random fluctuation



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?

Controlled vocabulary to describe genes:

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

http://amigo.geneontology.org/amigo/gene_product/UniProtKB:P51587

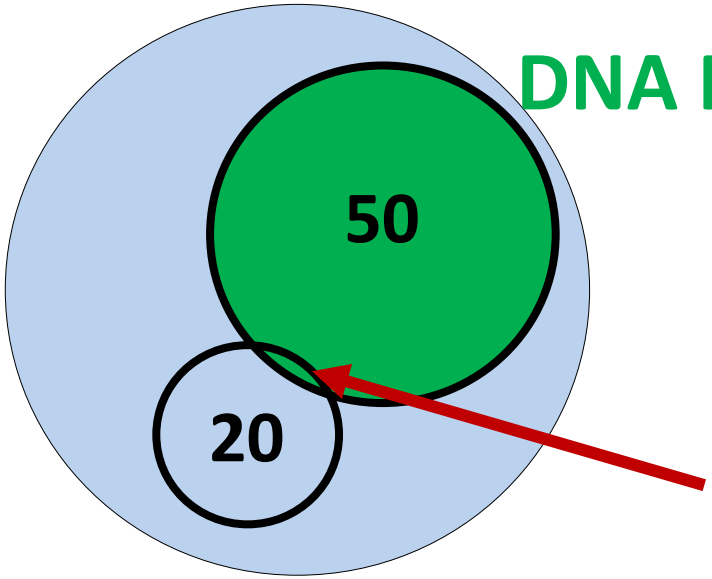
Gene/product	Gene/product name	Annotation qualifier	GO class (direct)
BRCA2	Breast cancer type 2 susceptibility protein		mitotic cytokinesis
BRCA2	Breast cancer type 2 susceptibility protein		telomere maintenance via recombination
BRCA2	Breast cancer type 2 susceptibility protein		double-strand break repair via homologous recombination
BRCA2	Breast cancer type 2 susceptibility protein		double-strand break repair via homologous recombination

Outline

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

Statistical Significance

- Your startup just developed a new drug, but related compounds cause cancer
- You want to know if it's safe
- Your idea: test it on cell lines and see what genes change in expression
- You find that it activates some genes involved in DNA Repair
- Could it be causing DNA damage?

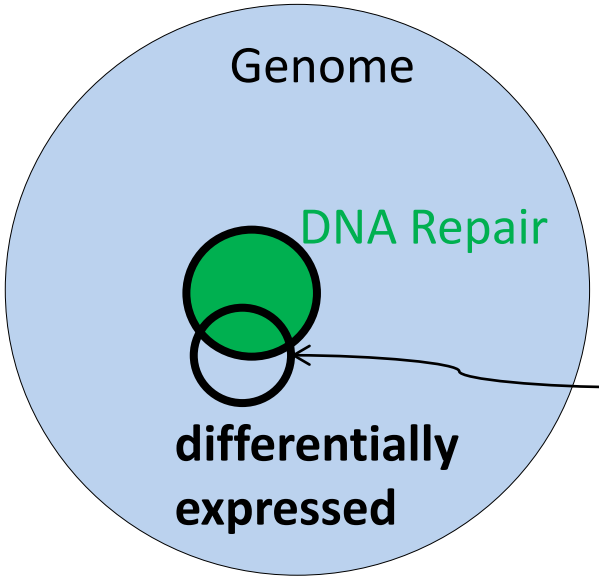


Differentially expressed

One gene differentially expressed gene is related to DNA repair.

Should I worry that our drug causes DNA damage?

Statistical significance

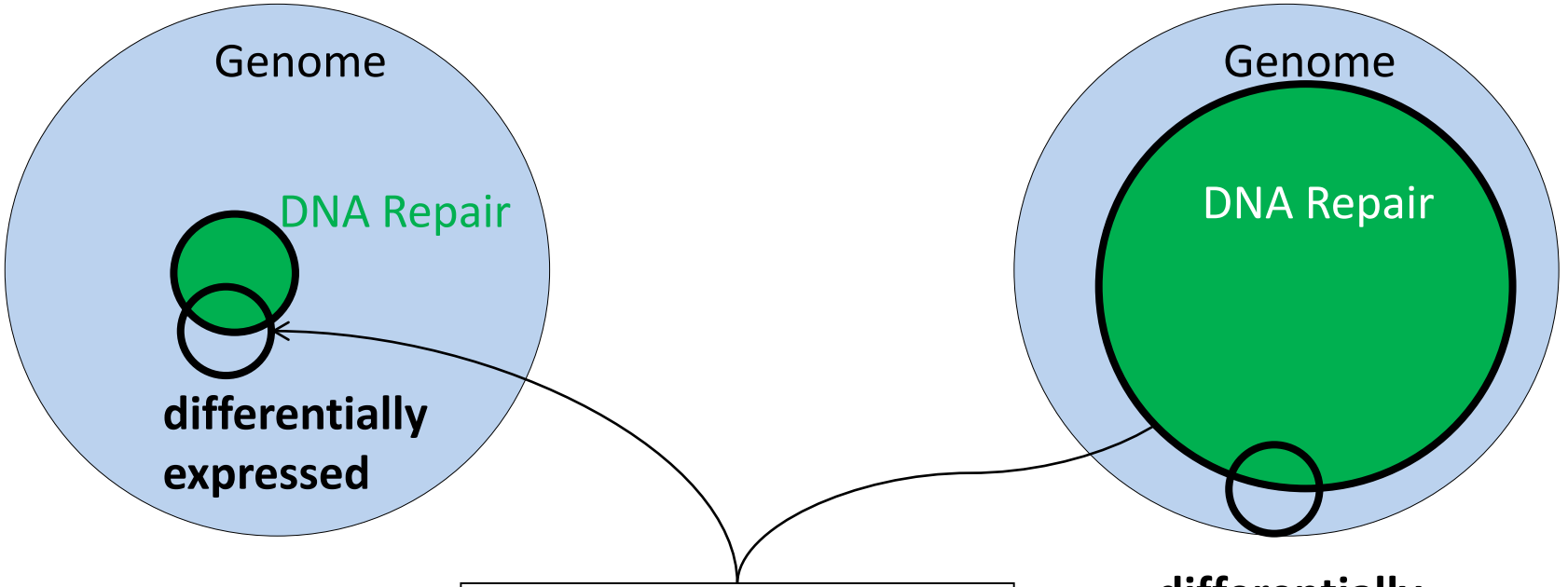


If I get many **more** repair genes than I would expect by chance, I need to find out if my drug is causing DNA damage.

In other words: are the differentially expressed genes **enriched** for ones involved in DNA repair?

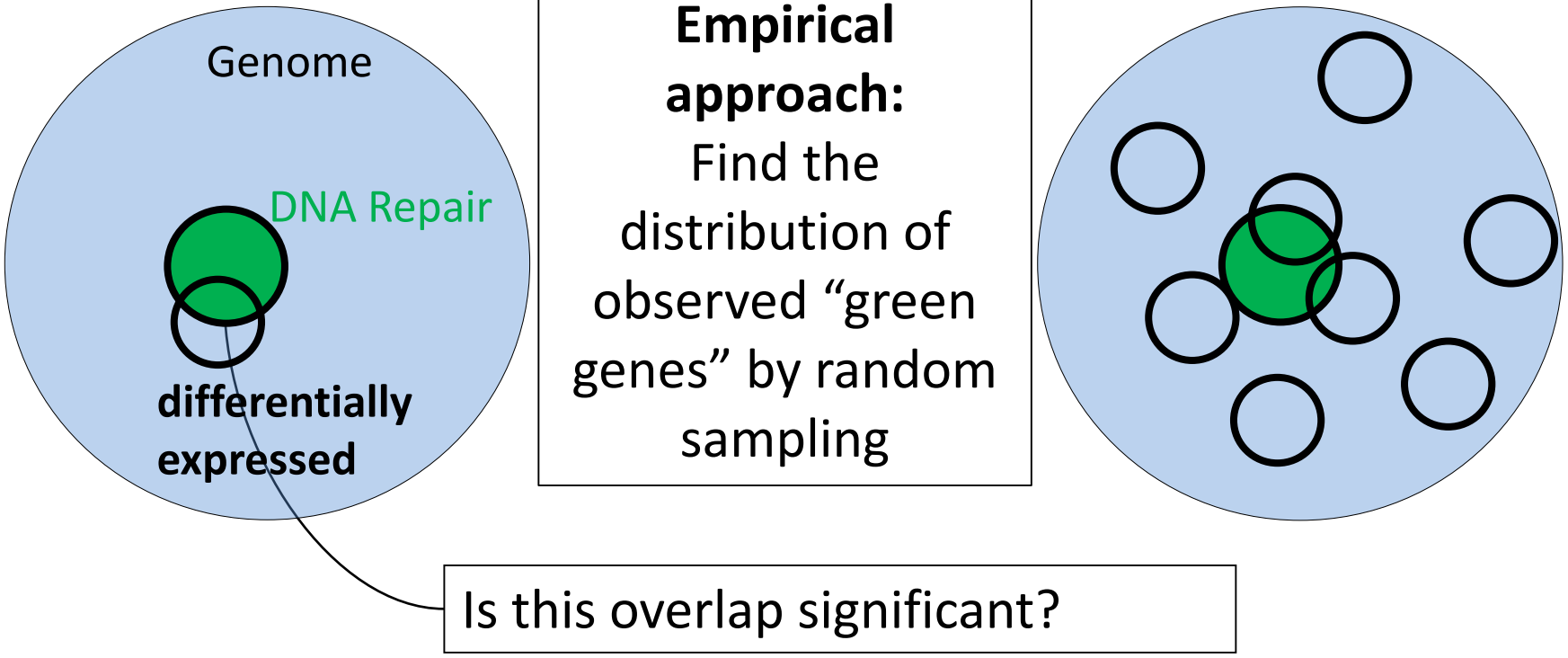
Statistical significance

The significance depends on the size of the lists.

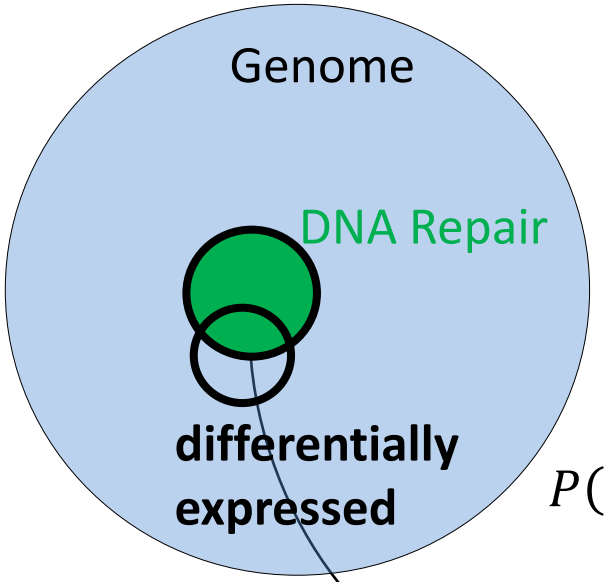


If the two lists had nothing in common, could we still get this degree of overlap?

Statistical significance



Statistical significance



Analytical Approach:
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:

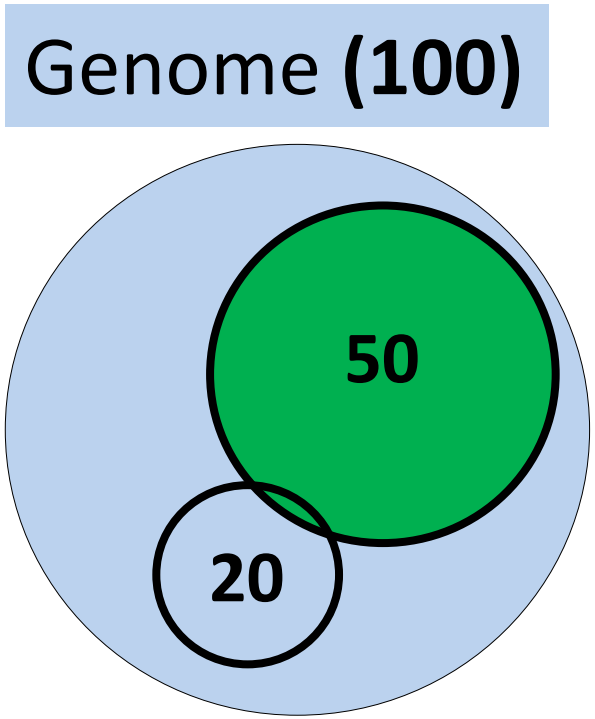
$$P(Overlap) = \frac{\binom{DNA\ repair}{Overlap}\binom{Genome - DNA\ repair}{DiffExp - Overlap}}{\binom{Genome}{DiffExp}}$$

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.

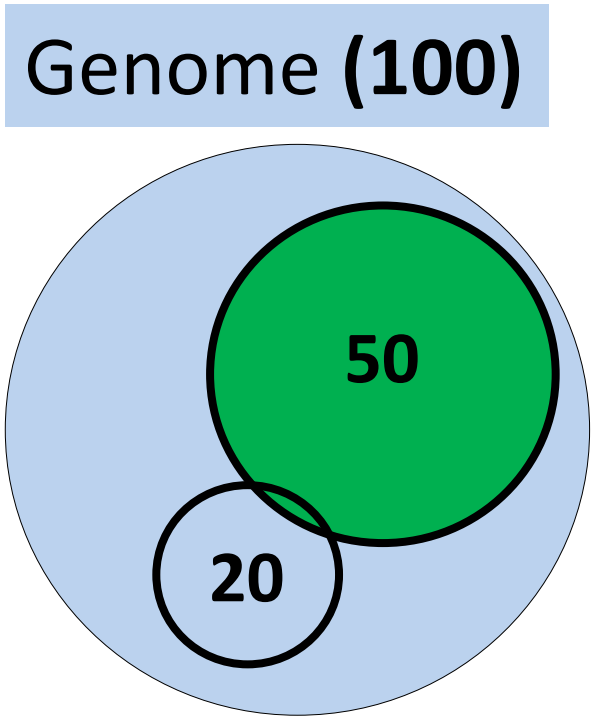
How you might use the HG test:



Differentially expressed

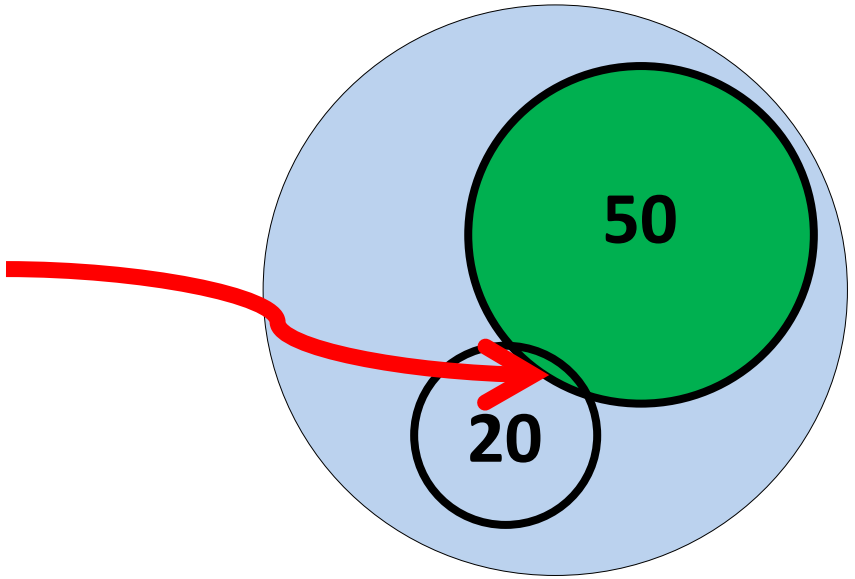
- Your startup just developed a new drug, but related compounds cause cancer
- You want to know if it's safe
- Your idea: test it on cell lines and see what genes change in expression
- You find that it activates some genes involved in DNA Repair
- Could it be causing DNA damage?

Statistical significance



- Usually, 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.

One gene overlaps.
Should I worry that our drug causes DNA damage?



Observed overlap

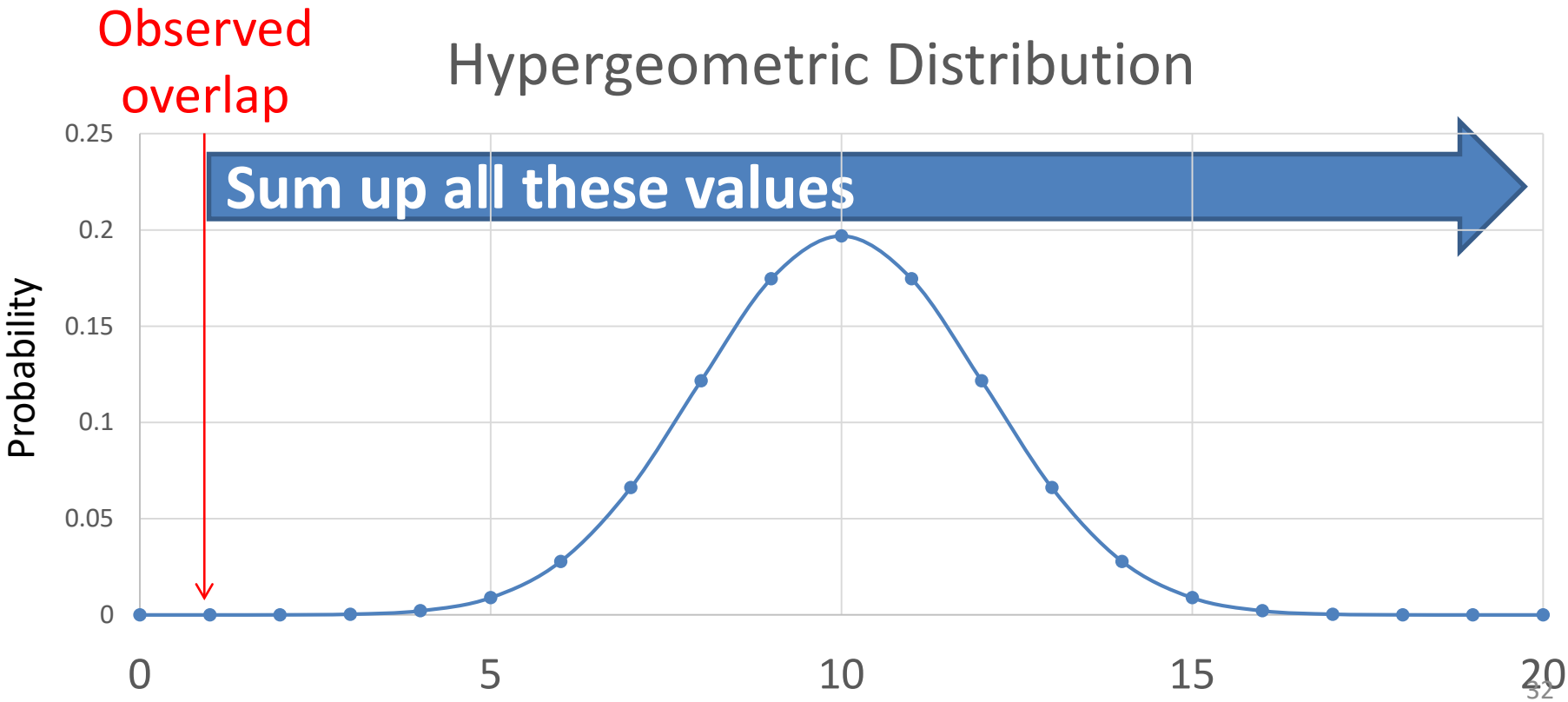
Hypergeometric Distribution

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

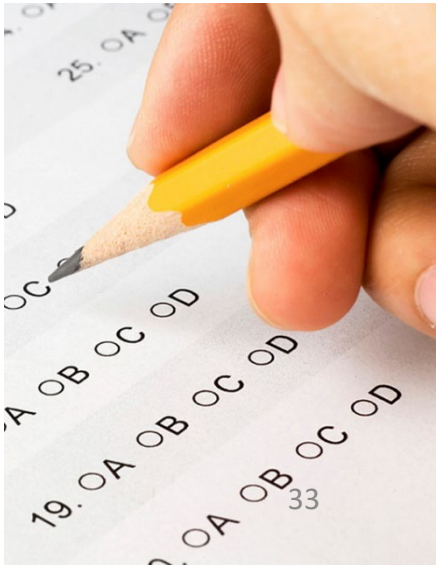
$$CDF(Overlap) = \sum_{n=overlap}^{Number\ of\ genes\ in\ DNA\ Repair} \frac{\binom{DNA\ repair}{n}\binom{Genome - DNA\ repair}{DiffExp - n}}{\binom{Genome}{DiffExp}}$$



CDF=Cumulative distribution function

Testing Multiple Hypotheses

- Example: Filter GO terms using a p-value threshold of 0.01
- By definition, 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

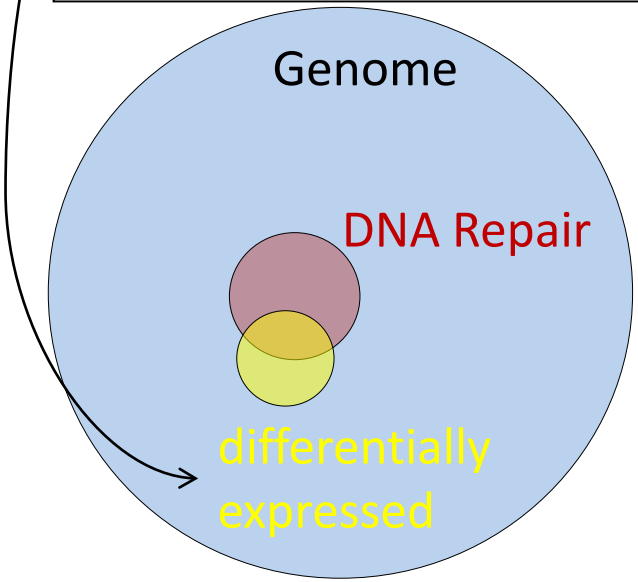
- A simple solution: require that the p-value be small enough to reduce the false positives to the desired level.
- This is called the Bonferroni correction.
- In our case, we would only accept terms with a

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

- 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”.

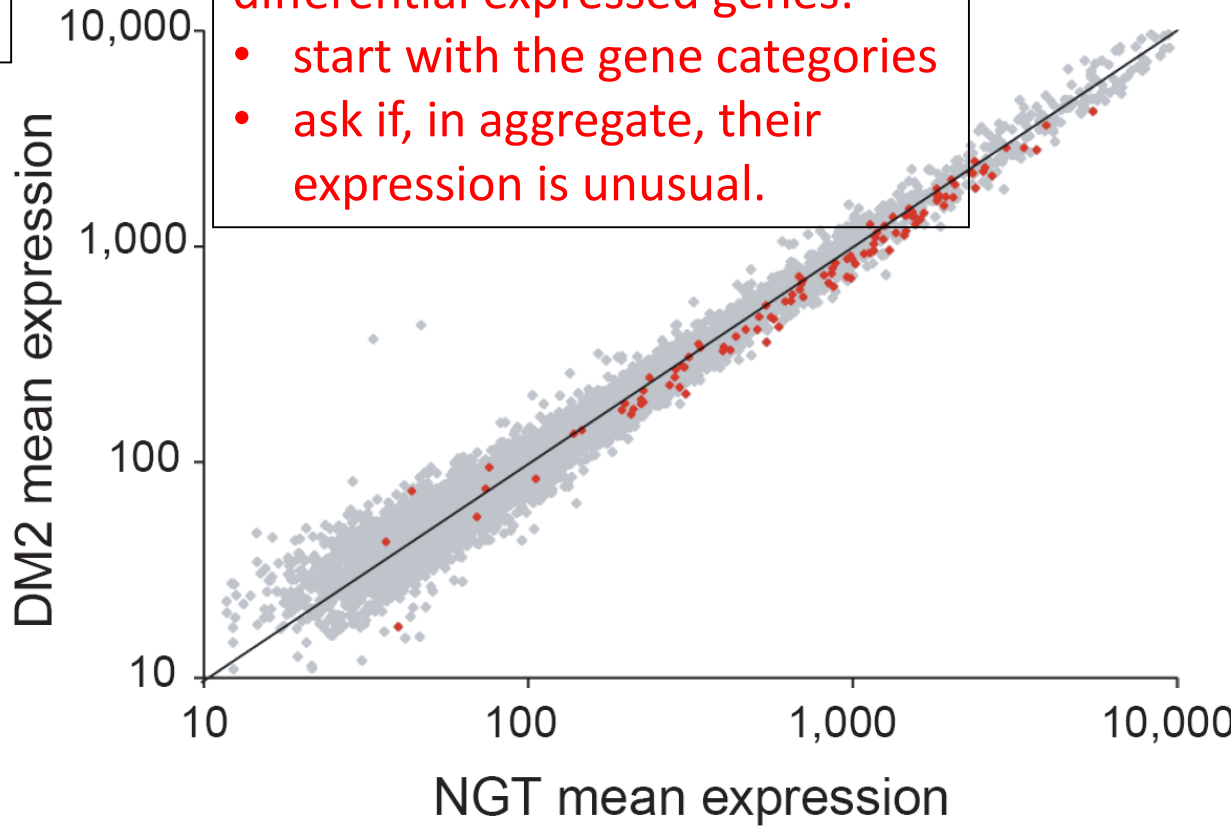
Aggregate score statistics

My results depend on how I defined "differentially expressed"



Instead of starting with differential expressed genes:

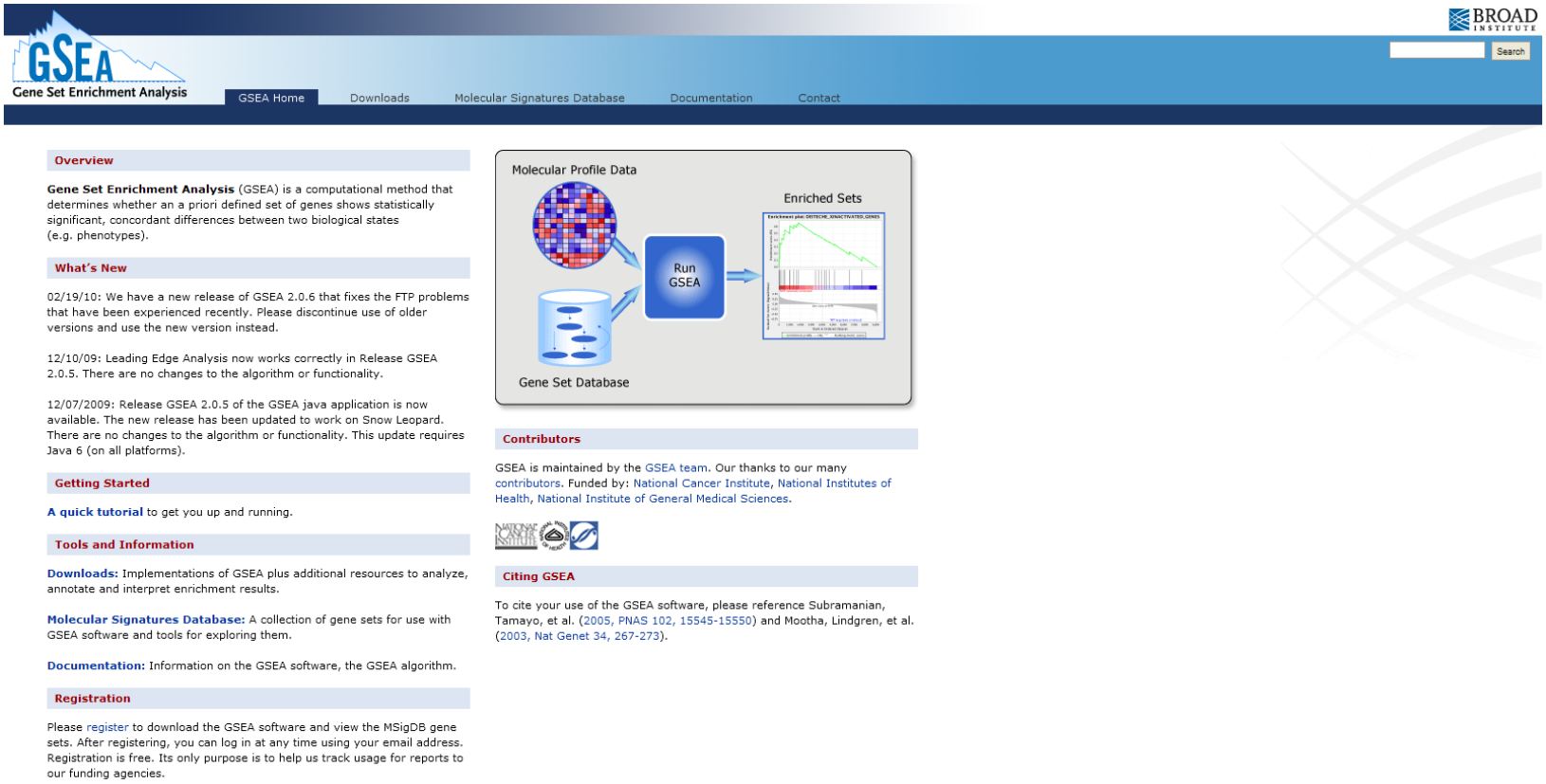
- start with the gene categories
- ask if, in aggregate, their expression is unusual.



Mootha *et al.* (2003). *Nature Genetics* **34**, 267 – 273. doi:10.1038/ng1180

Aggregate score statistics

http://www.broadinstitute.org/gsea/

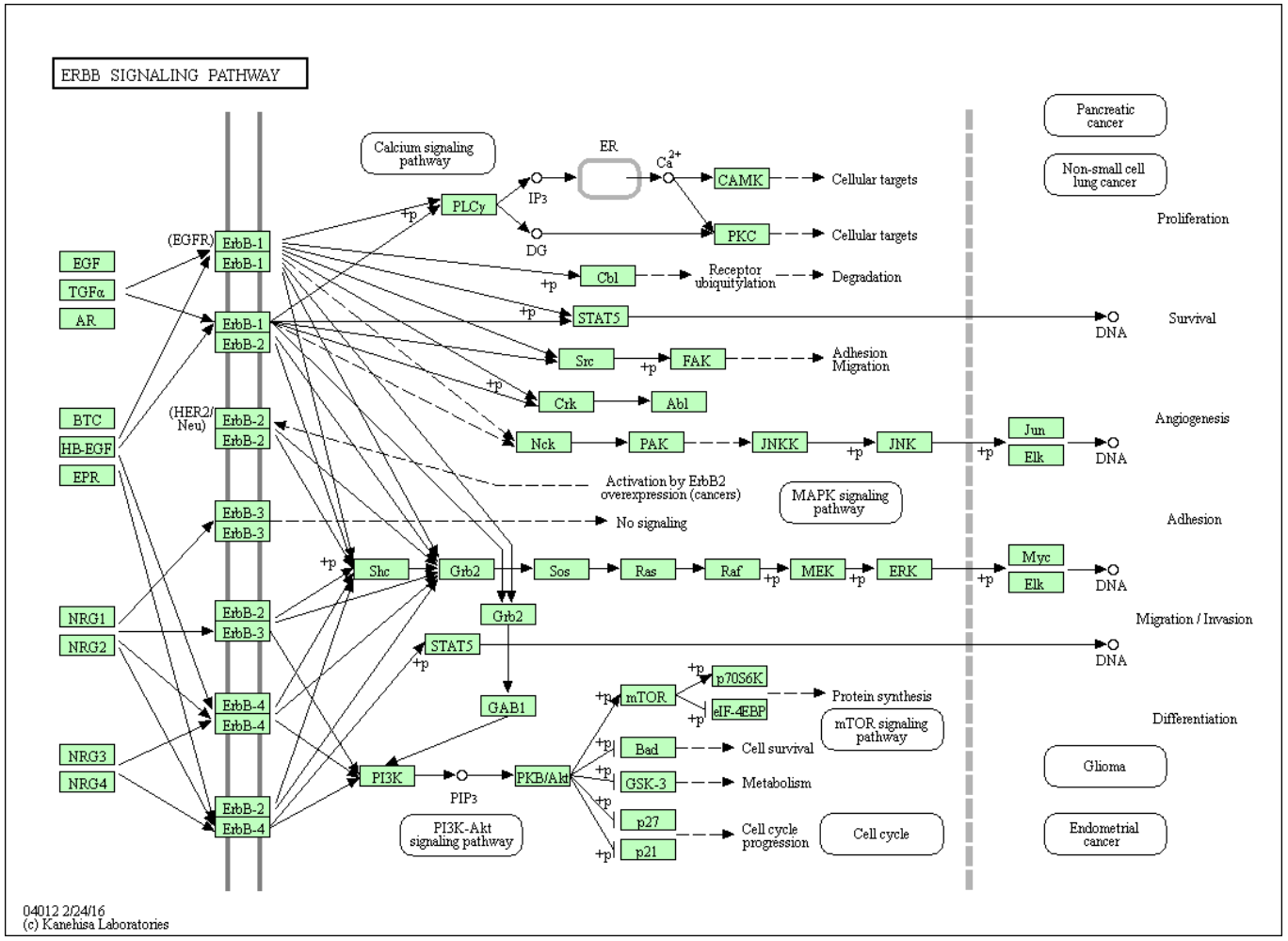


- **Biological process**
- A biological process is not equivalent to a pathway.
  - Does not represent the dynamics or dependencies of a pathway.

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

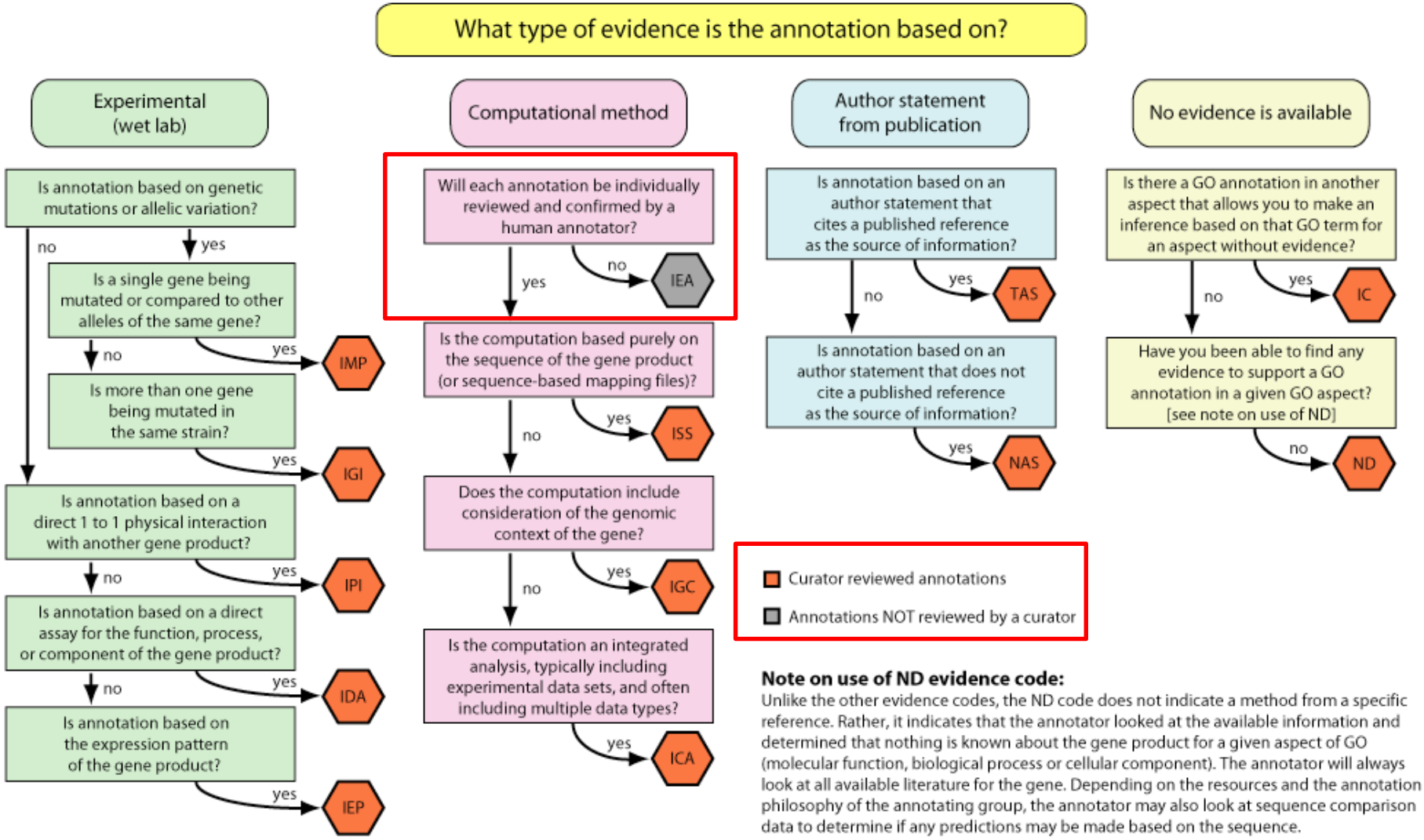


Select all | Clear all | Perform an action with this page's selected terms... | Go!

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

Not just the obvious categories

GO Evidence Code Decision Tree



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.