Lecture Slides for Thursday April 2nd

11:05 AM EDT by Zoom

https://mit.zoom.us/j/348659452

For audio you can use your computer or call:

US : +1 646 558 8656 or +1 669 900 6833

Meeting ID: 348 659 452

International Numbers:
https://mit.zoom.us/u/adLEbsadSS

Note: class will be recorded and posted for later viewing.

My Revised Lecture Schedule

Date	Topic
March 31st	Cluster, PCA
April 2nd	RNA-Seq
April 7th	Transcriptional Regulation

Reminders on remote education:

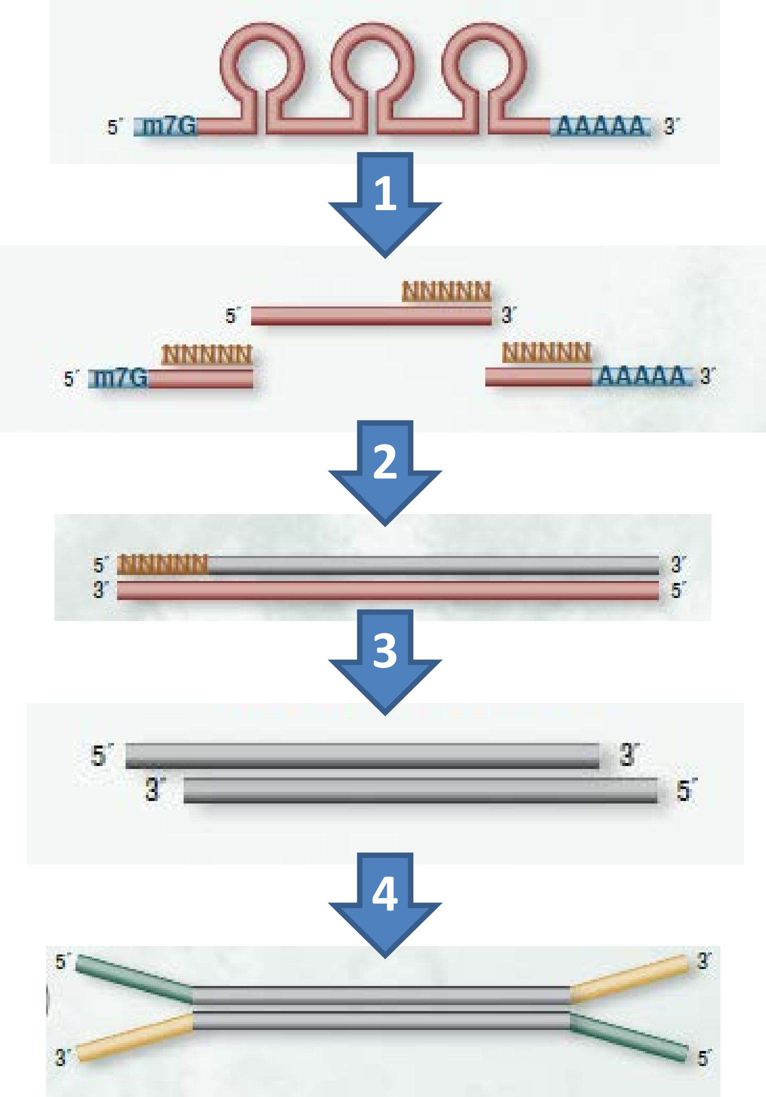
- This class is being recorded. We do not intend for anyone outside the class to access the recording, but …
  - If you are concerned, please turn off your video and send us an email.
- Please turn on your camera – and dress appropriately!
- Keep the session number handy in case you loose your connection: 348-659-452
- Remember you can join by phone for audio only if your computer malfunctions. +1 (646) 558-8656
- Feel free to use the chat function to talk to each other – but remember, all chats are recorded and will be posted with the lecture.

RNA-Seq Topics

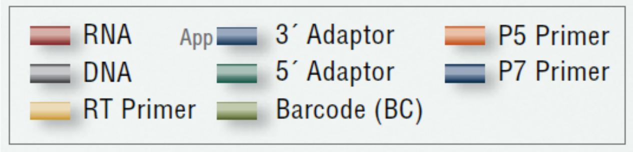
- Overview of experimental steps for RNA-Seq
- Deriving expression levels from sequence data
- Gene Ontology
- Statistical significance

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

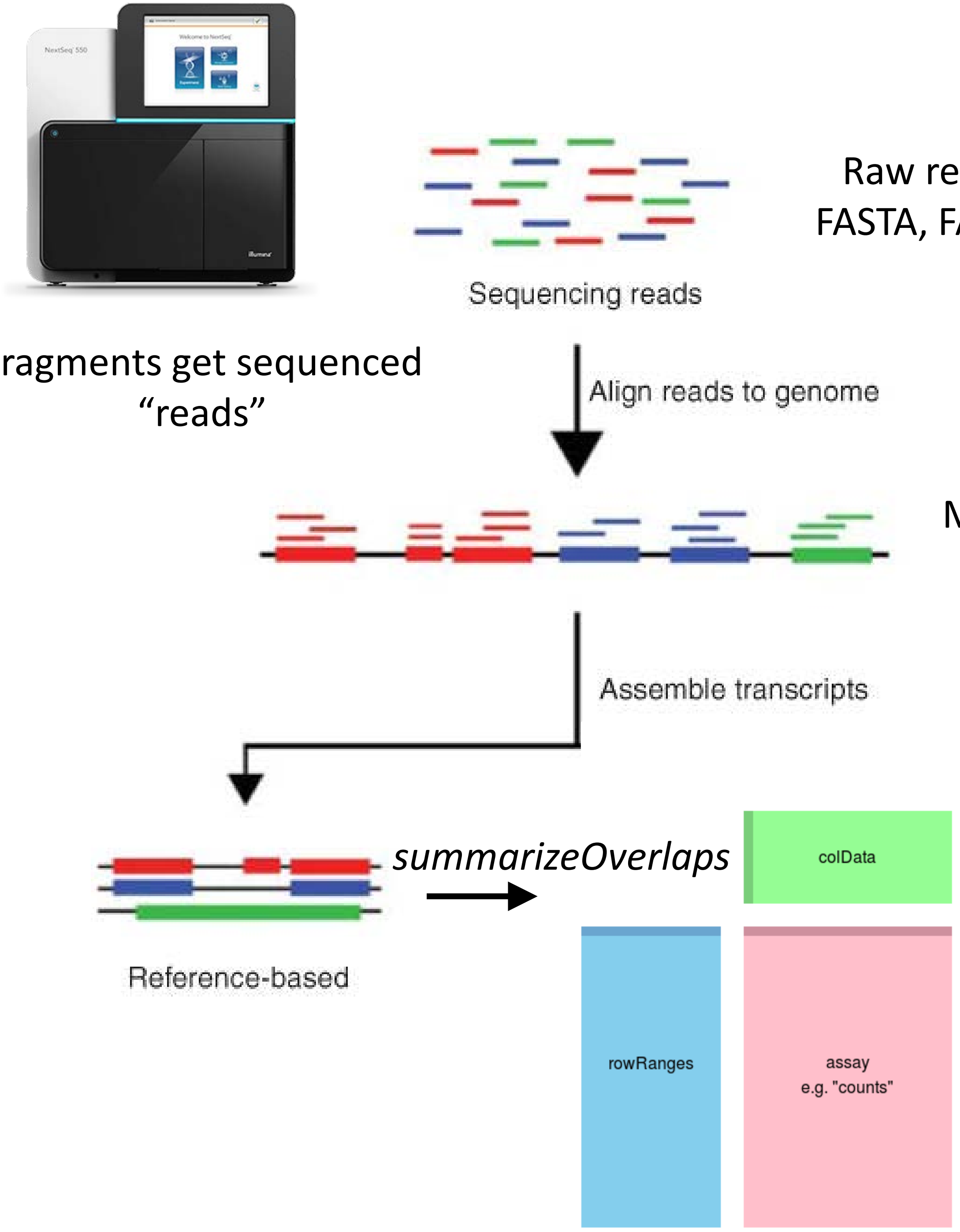


NEBNext® for Illumina®

NGS SAMPLE PREPARATION

Outline

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



Fragments get sequenced
“reads”

Raw reads
FASTA, FASTQ

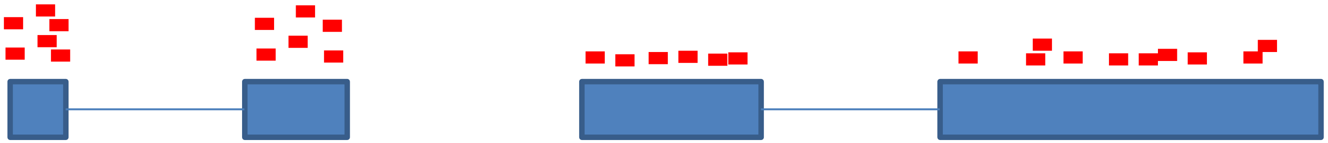
Align to genome
TopHat2

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.

Other ways to report transcripts

- **RPKM**: Reads Per Kilobase Million
  - This is what we just discussed
- **FPKM**: Fragments Per Kilobase Million
  - In “paired-end” sequencing, the fragments are sequenced from each end. Most of the time you detect both ends, but not always. FPKM reports results for the original DNA fragment regardless of whether you detected one or two ends

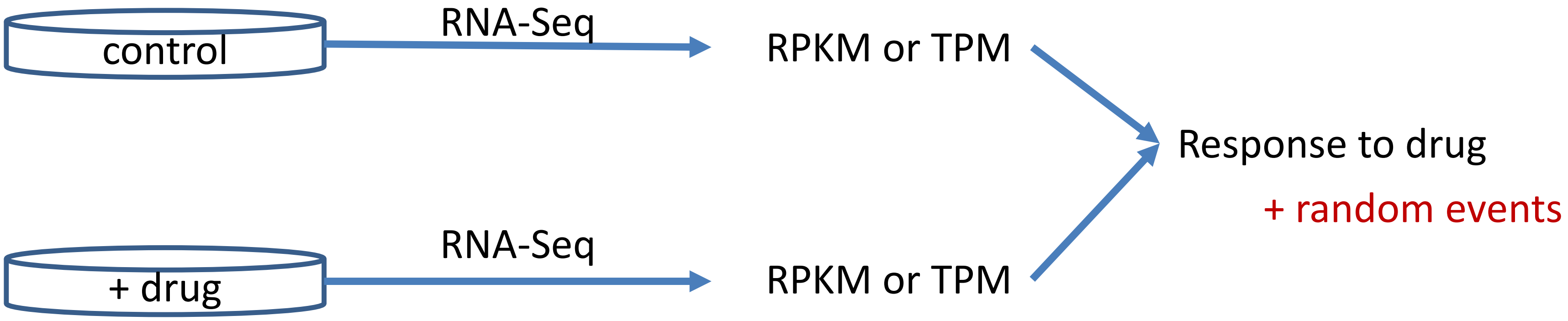
Other ways to report transcripts

- **TPM**: Transcript per million
  - – Provides a more accurate estimate of the relative molar concentration of transcripts
  - – Just as easy to compute
  - – Described in detail in the reference below:

Theory Biosci. 2012 Dec;131(4):281-5. doi: 10.1007/s12064-012-0162-3. Epub 2012 Aug 8.
Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.
Wagner GP[1], Kin K, Lynch VJ.

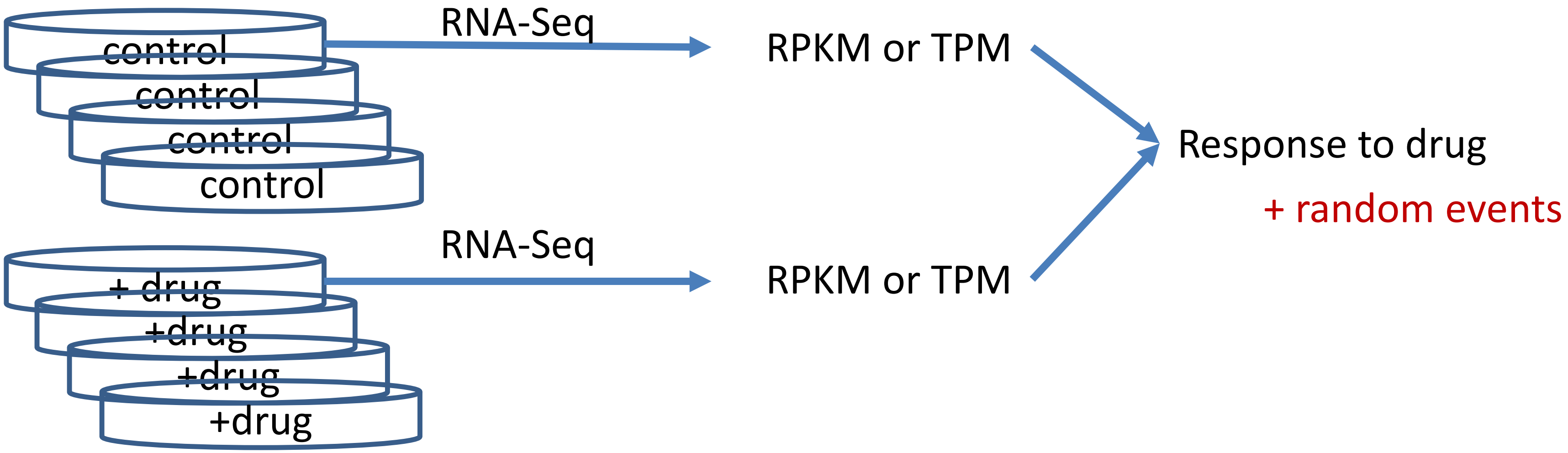
Differential expression

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



Differential expression

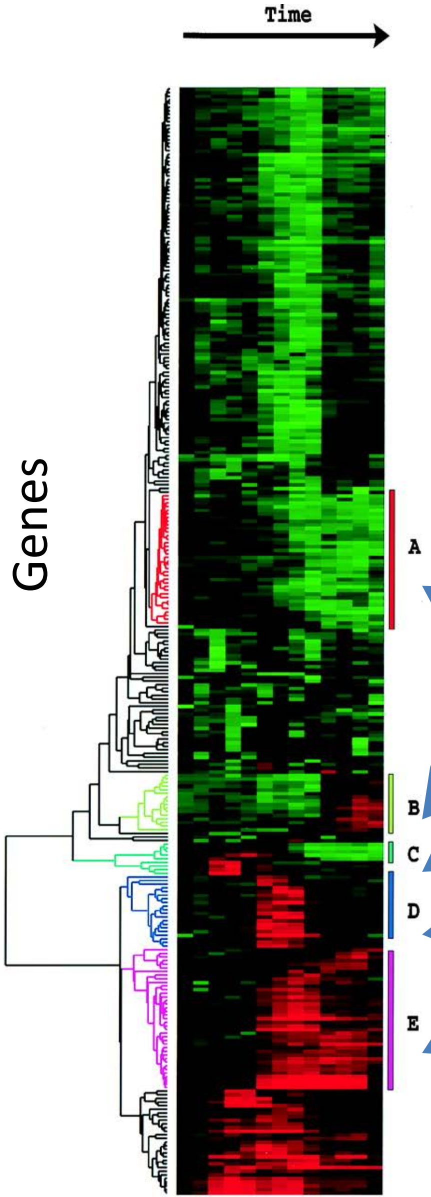
DESeq2: tests whether a difference in gene expression is a response to a change in condition vs. a random fluctuation



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

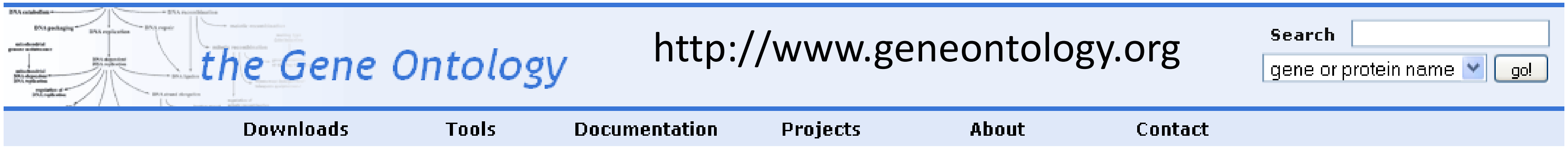
Iyer et al. *Science* 1999

Outline

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

Biological Insights

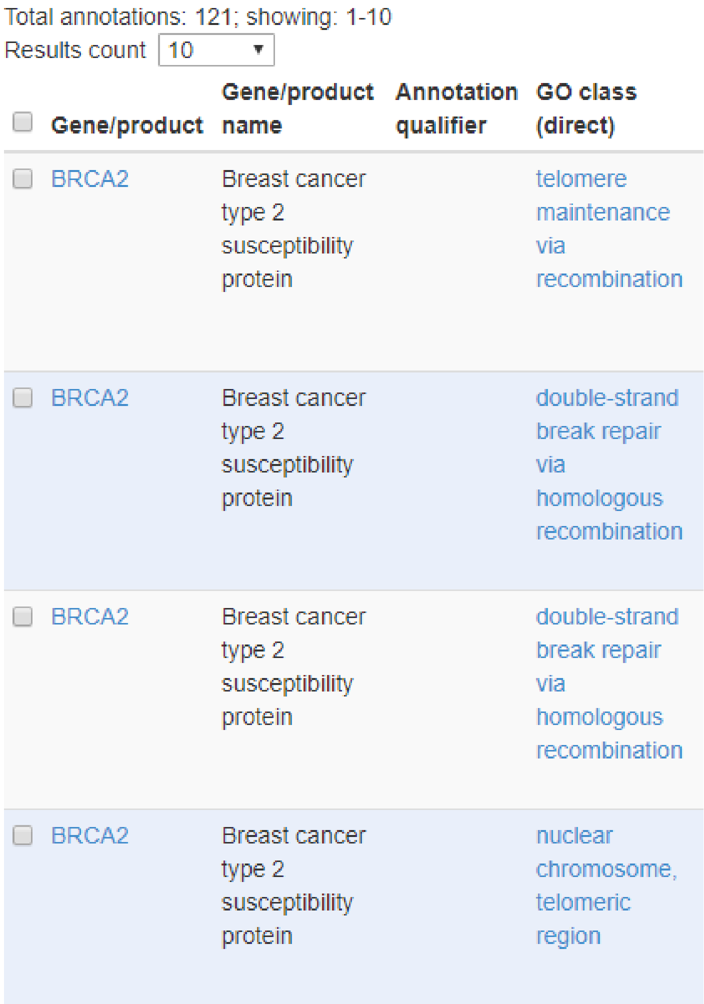
- What types of genes are being differentially expressed?



Controlled vocabulary to describe genes:

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

A gene often will have several annotations



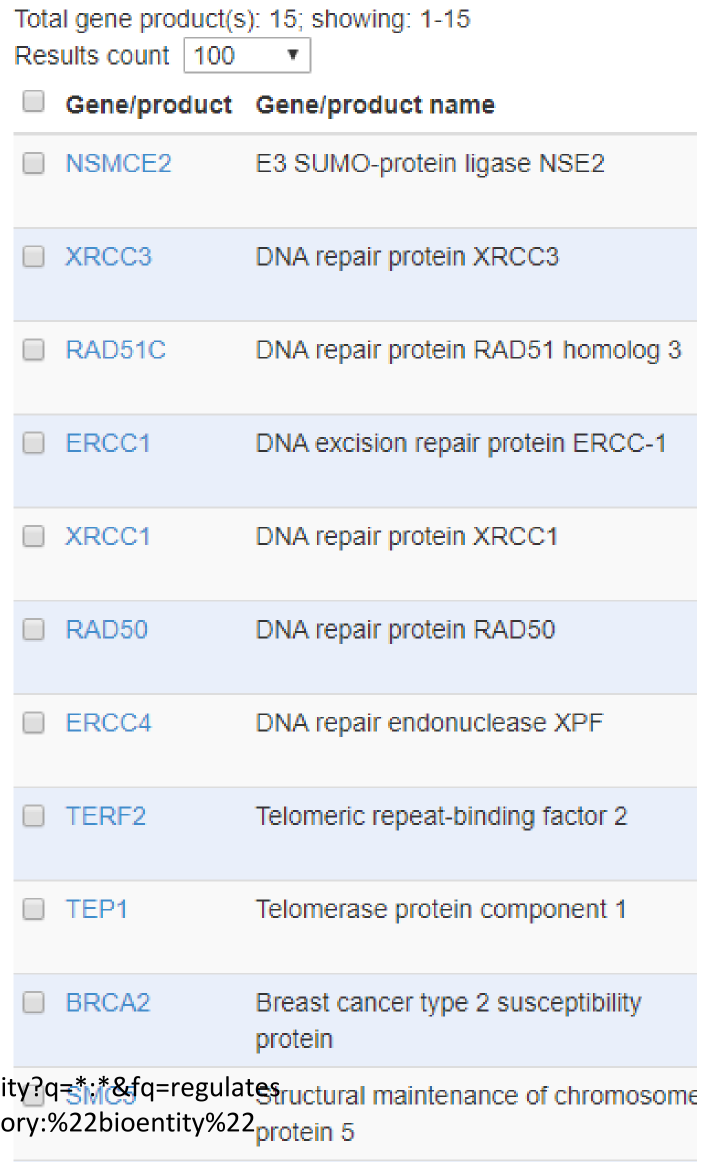
Total annotations: 121; showing: 1-10

Results count 10 ▾

□	Gene/product	Gene/product name	Annotation qualifier	GO class (direct)
□	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
□	BRCA2	Breast cancer type 2 susceptibility protein		nuclear chromosome, telomeric region

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

Annotations usually have many genes



Total gene product(s): 15; showing: 1-15

Results count 100 ▼

□	Gene/product	Gene/product name
□	NSMCE2	E3 SUMO-protein ligase NSE2
□	XRCC3	DNA repair protein XRCC3
□	RAD51C	DNA repair protein RAD51 homolog 3
□	ERCC1	DNA excision repair protein ERCC-1
□	XRCC1	DNA repair protein XRCC1
□	RAD50	DNA repair protein RAD50
□	ERCC4	DNA repair endonuclease XPF
□	TERF2	Telomeric repeat-binding factor 2
□	TEP1	Telomerase protein component 1
□	BRCA2	Breast cancer type 2 susceptibility protein
□	SMC5	Structural maintenance of chromosome protein 5

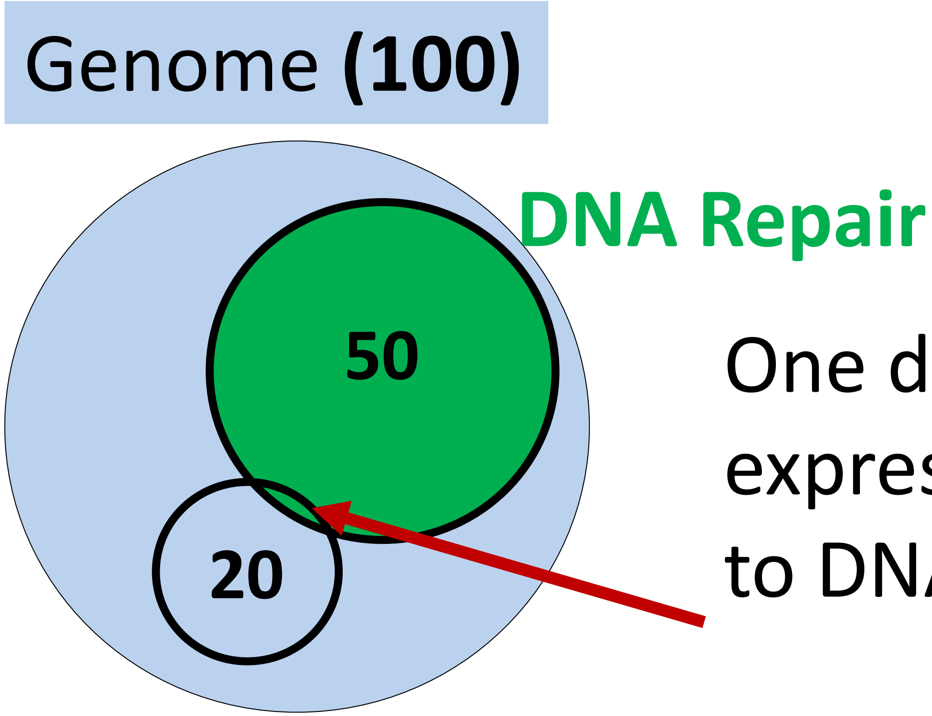
http://amigo.geneontology.org/amigo/search/bioentity?q=*:*&fq=regulates_closure:%22GO:0000722%22&sfq=document_category:%22bioentity%22

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?



Genome **(100)**

DNA Repair

50

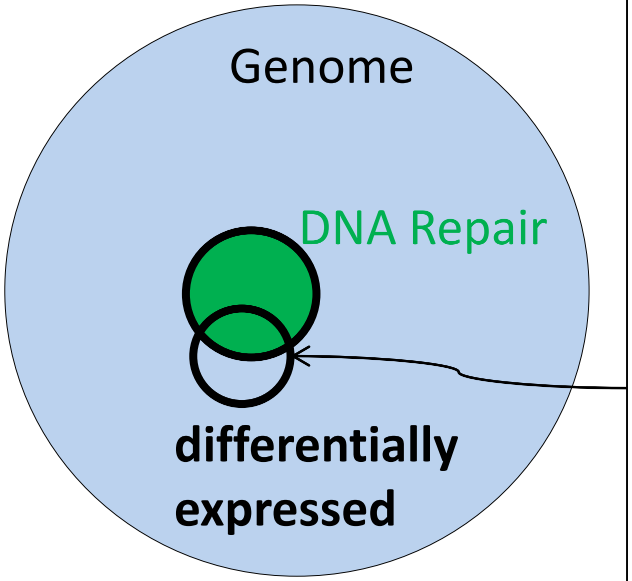
20

Differentially expressed

One differentially expressed gene is related to DNA repair.

Should I worry that our drug causes DNA damage?

Statistical significance

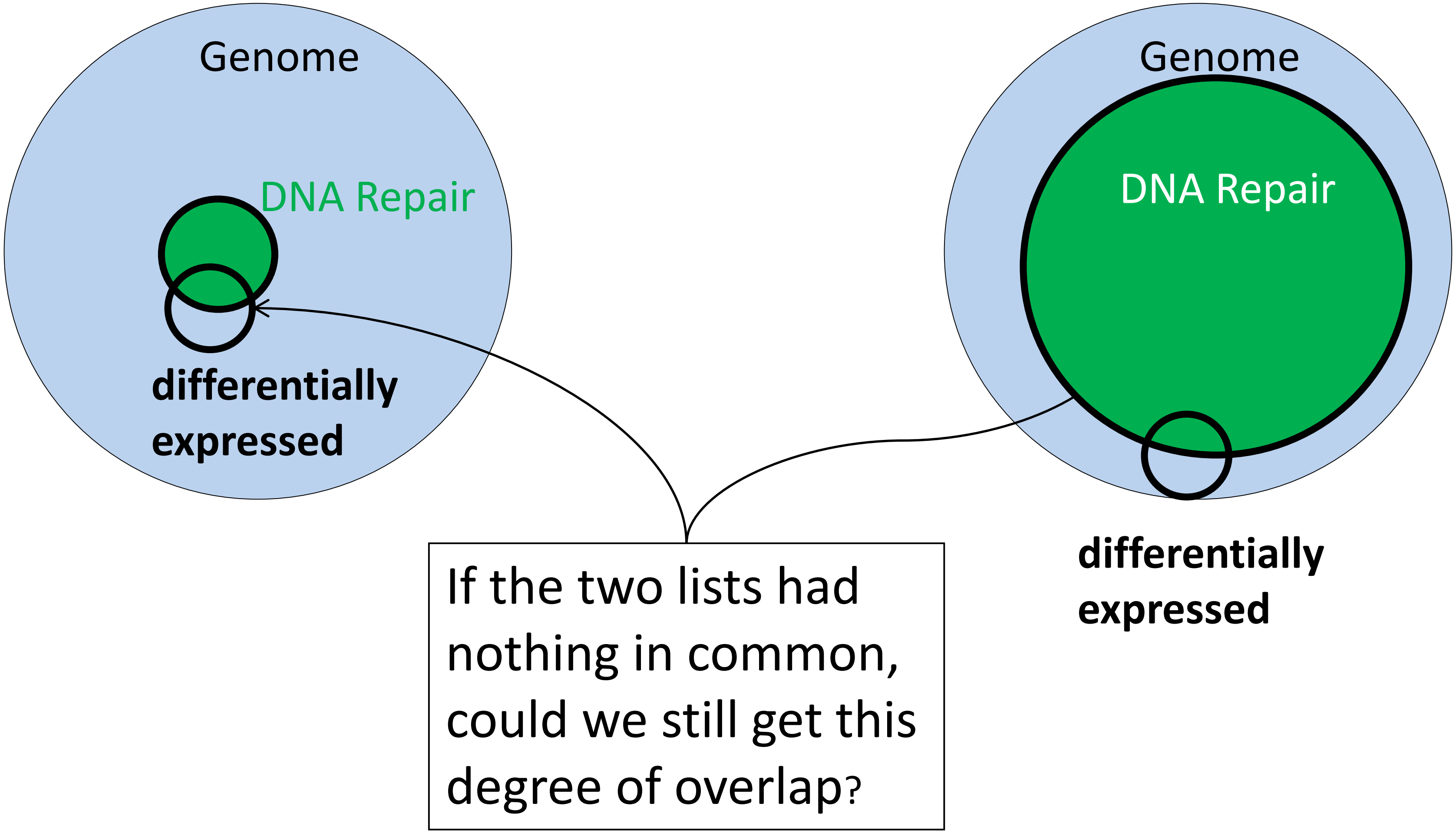


If I get many **more** repair genes than I would expect by chance, I need to find out if my drug is causing DNA damage.

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

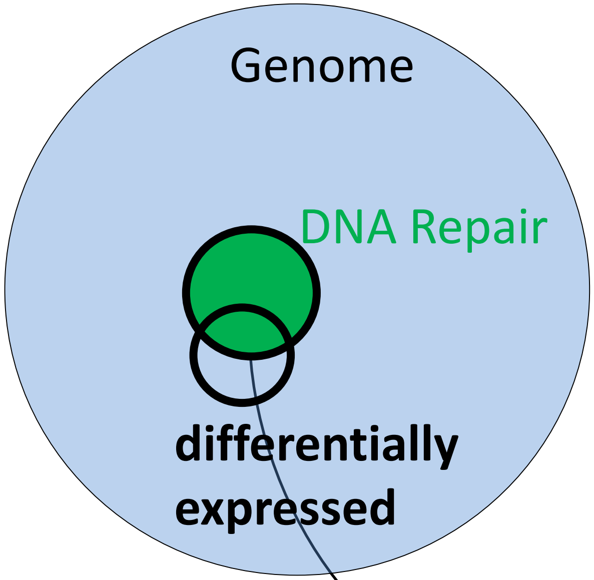
Statistical significance

The significance depends on the size of the lists.

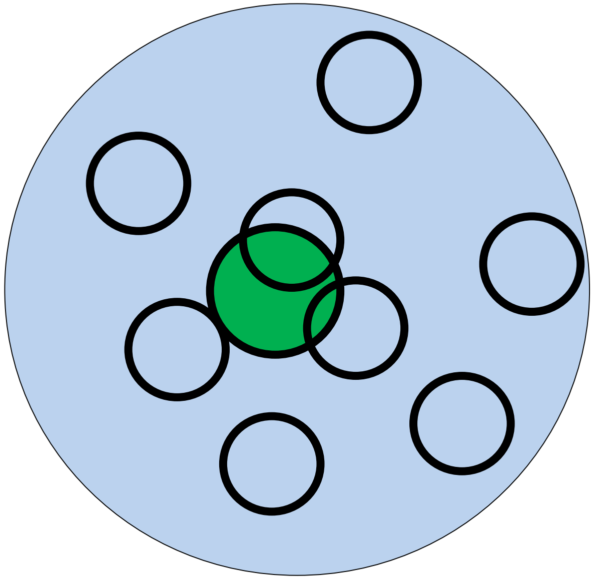


If the two lists had nothing in common, could we still get this degree of overlap?

Statistical significance

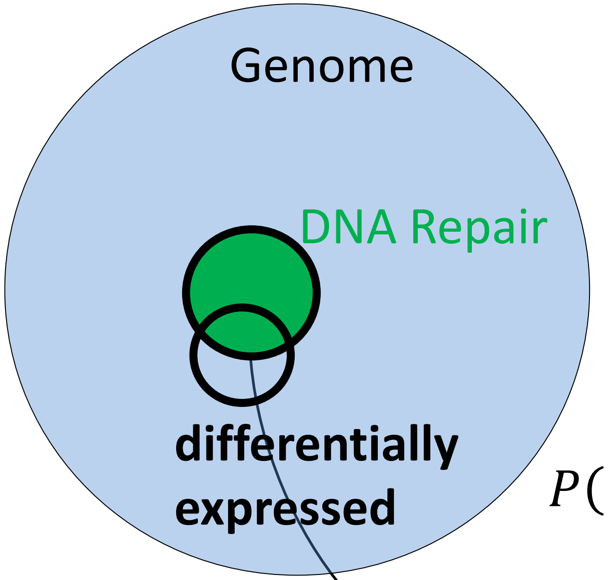


Empirical approach:
Find the distribution of observed “green genes” by random sampling



Is this overlap significant?

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:

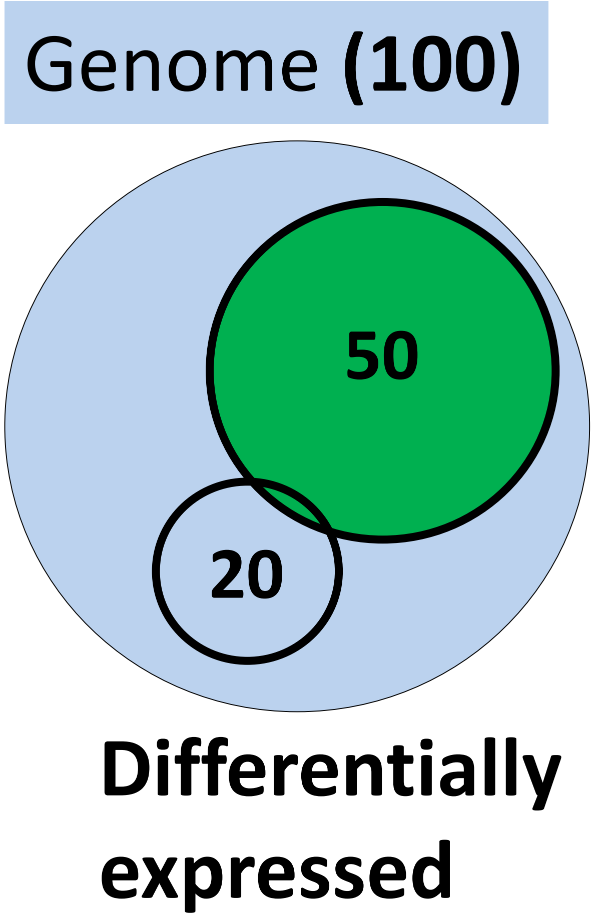
$$P(Overlap) = \frac{\binom{DNA\,repair}{Overlap}\binom{Genome - DNA\,repair}{DiffExp - Overlap}}{\binom{Genome}{DiffExp}}$$

Is this overlap significant?

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

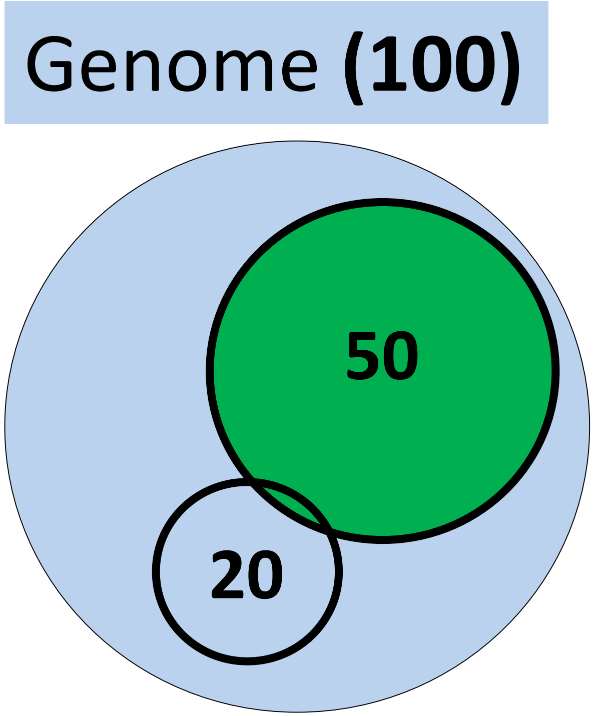
= the number of ways to choose k items from a set of n.

How you might use the HG test:



- 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

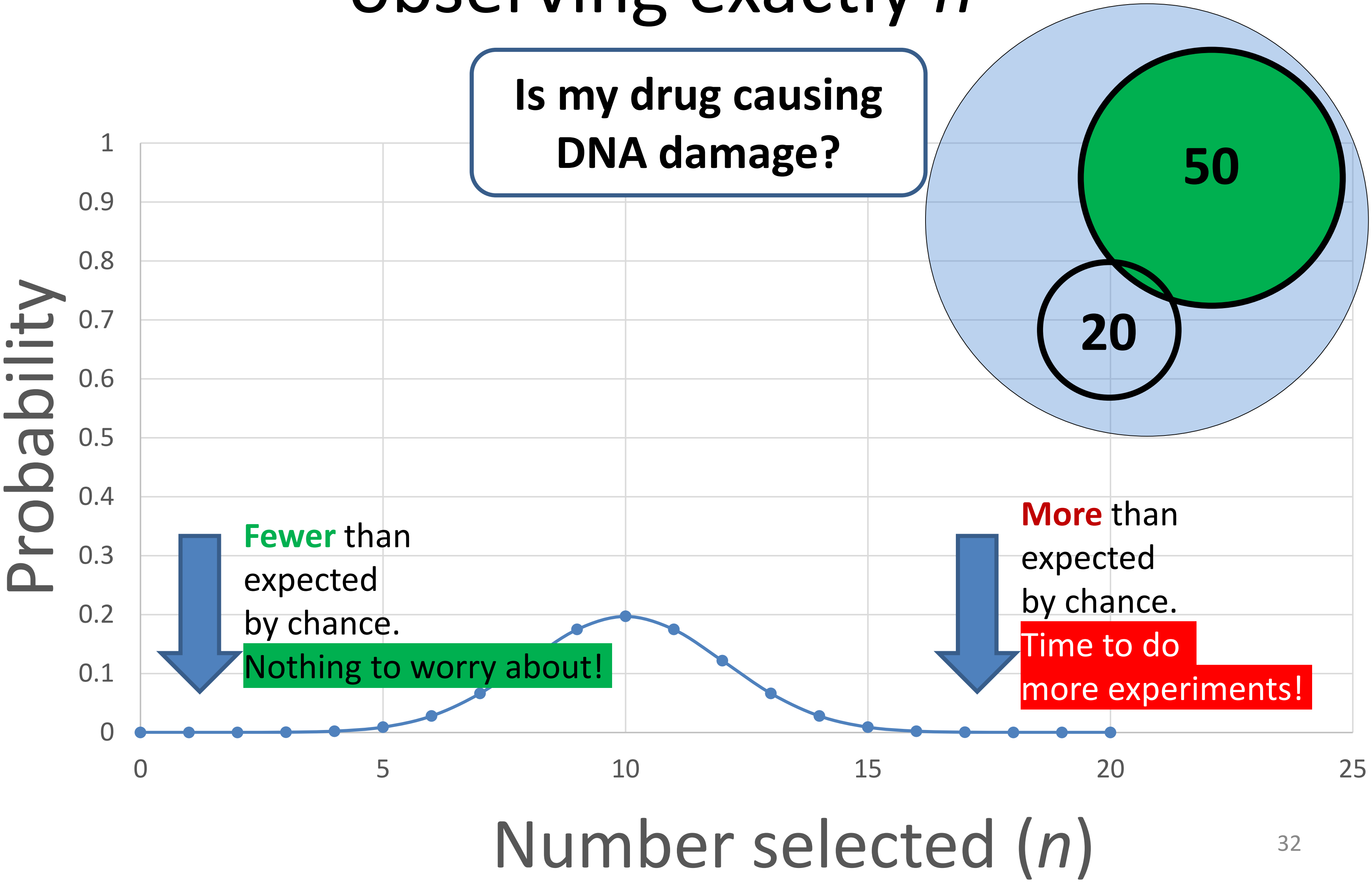


Differentially expressed

- 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.

HG measures the probability of observing exactly *n*

Is my drug causing DNA damage?

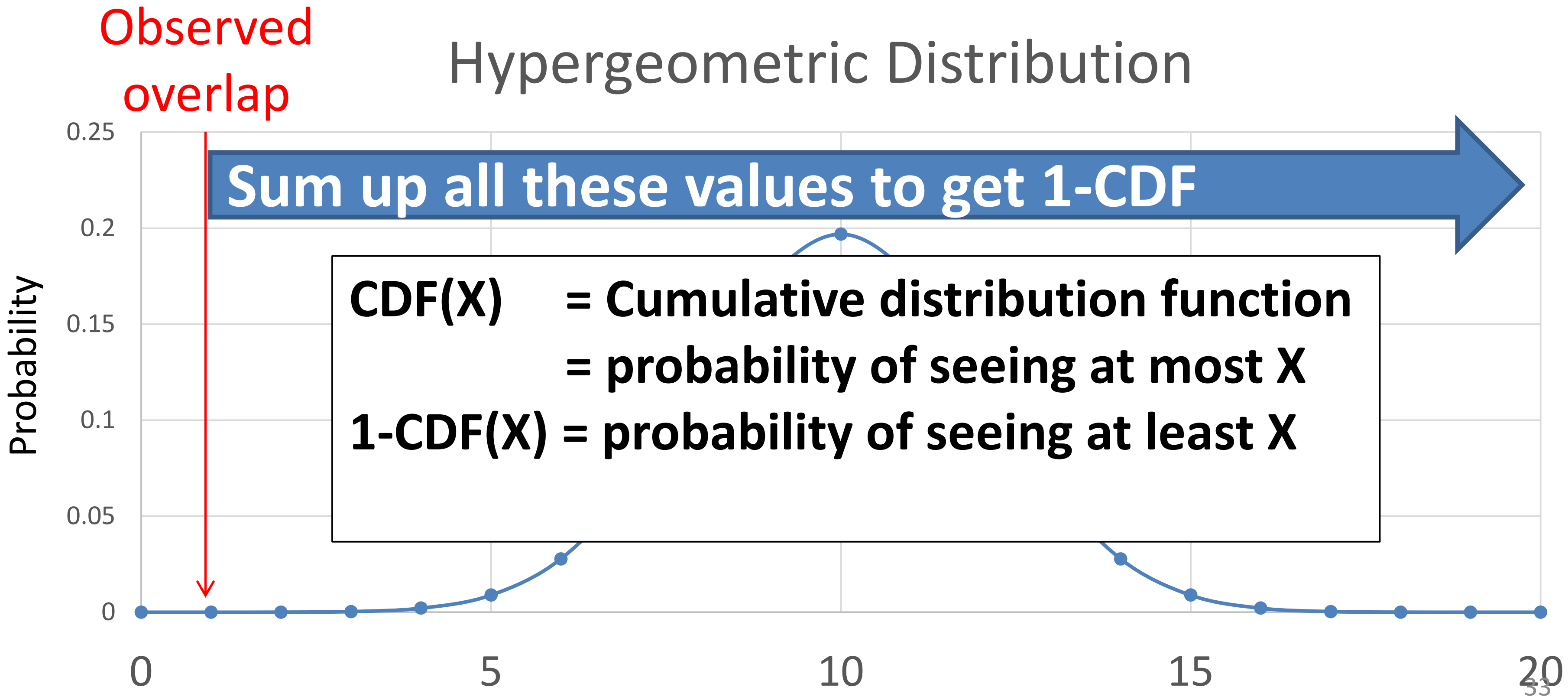


Fewer than expected by chance.
Nothing to worry about!

More than expected by chance.
Time to do more experiments!

The CDF helps us find enriched terms

$$1 - CDF(Overlap) = \sum_{n=overlap}^{\substack{Number\ of \\ genes\ in\ DNA\ Repair}} \frac{\binom{DNA\ repair}{n}\binom{Genome - DNA\ repair}{DiffExp - n}}{\binom{Genome}{DiffExp}}$$

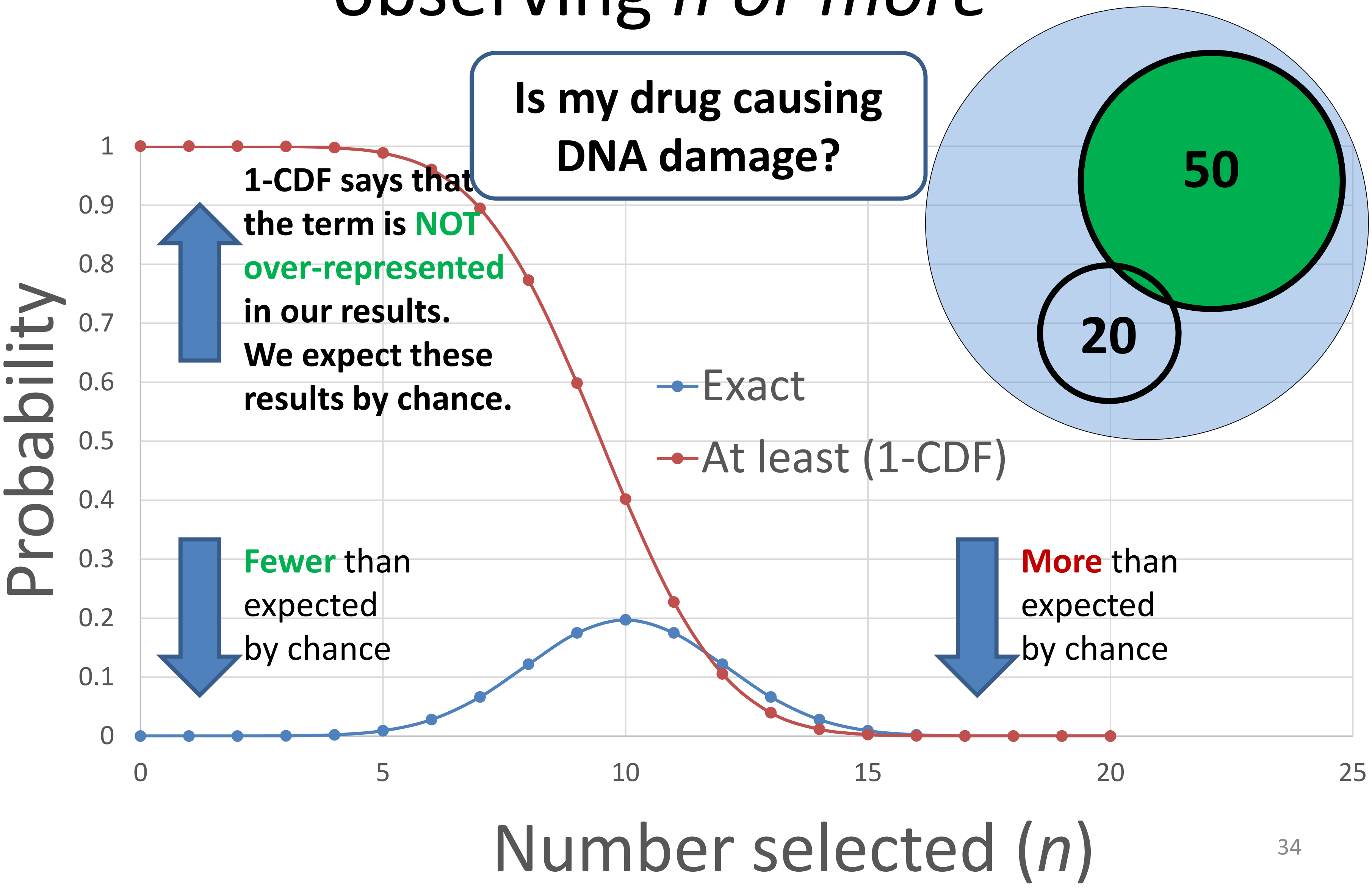


Observed overlap

Sum up all these values to get 1-CDF

CDF(X) = Cumulative distribution function
= probability of seeing at most X
1-CDF(X) = probability of seeing at least X

(1-CDF) measures the probability of observing *n or more*



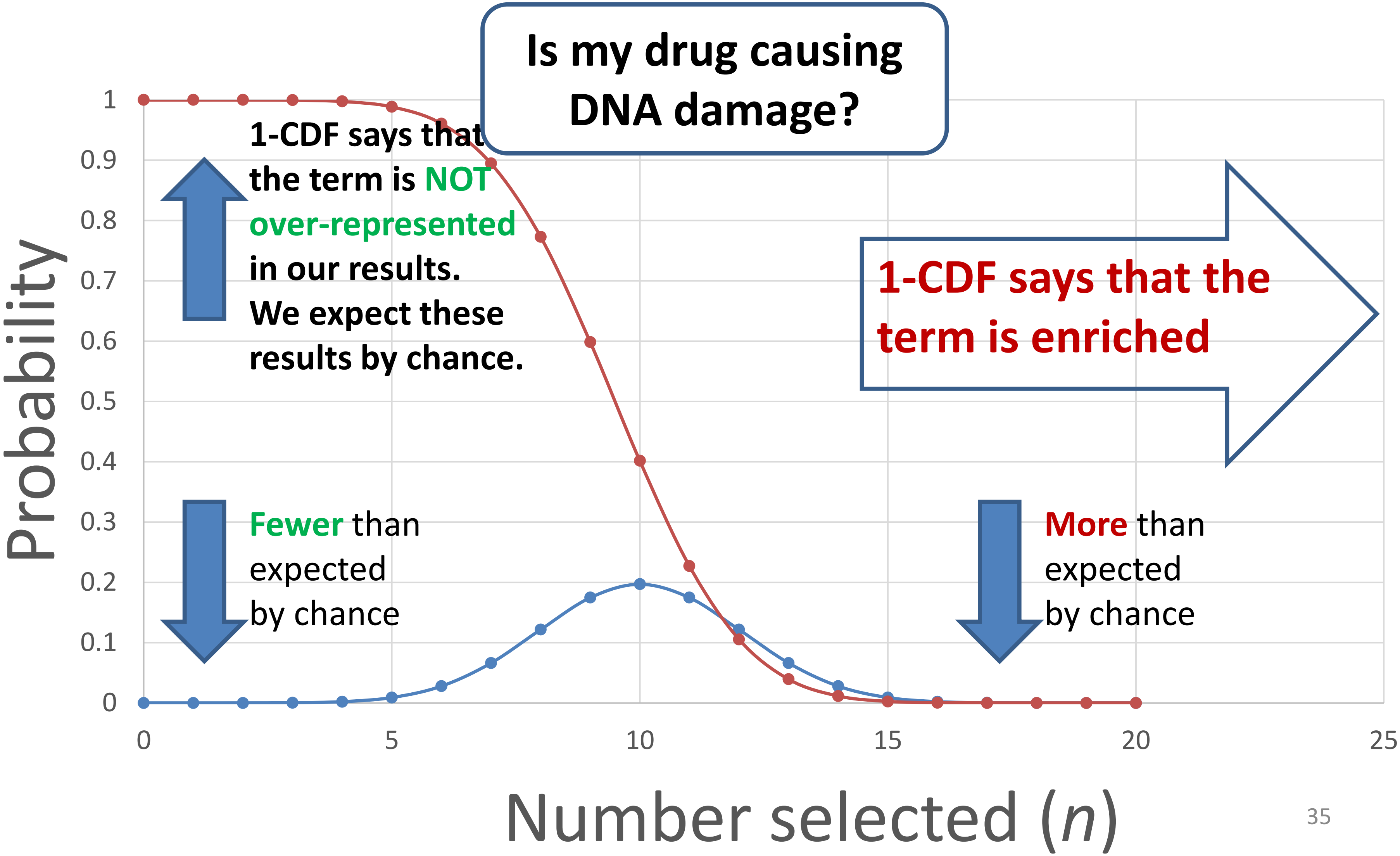
Is my drug causing DNA damage?

1-CDF says that the term is NOT over-represented in our results. We expect these results by chance.

Fewer than expected by chance

More than expected by chance

CDF measures the probability of observing at least *n*



Is my drug causing DNA damage?

1-CDF says that the term is NOT over-represented in our results. We expect these results by chance.

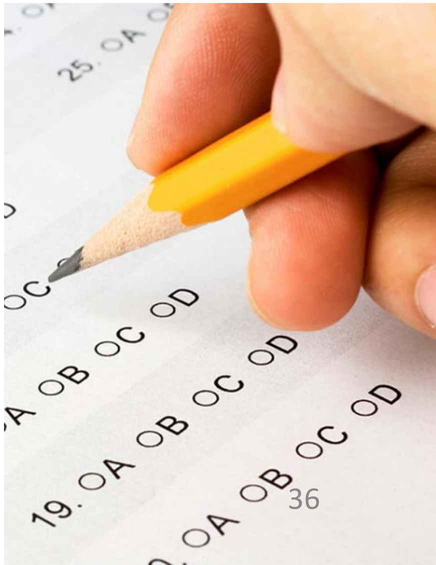
Fewer than expected by chance

1-CDF says that the term is enriched

More than expected by chance

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 ***<u>for each test.</u>***
- There are roughly 30,000 terms in GO.
- At this level, we expect roughly 300 false positives!



Multiple Hypotheses

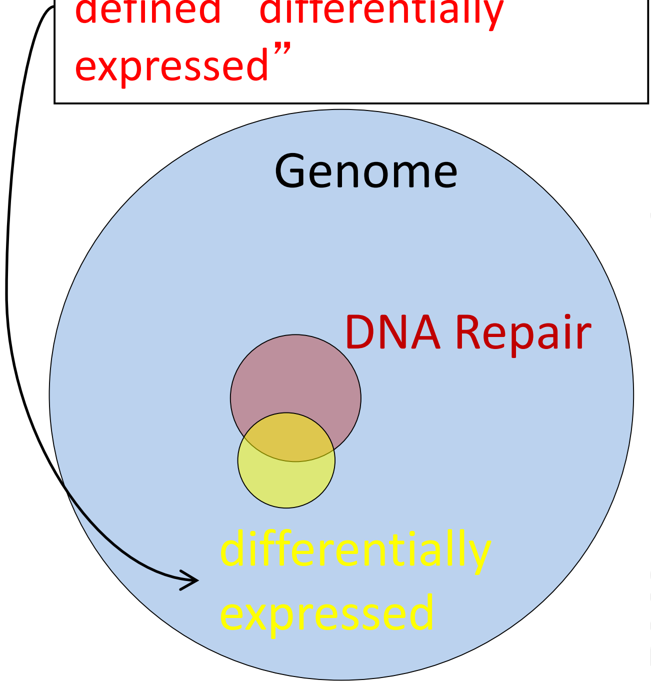
- A simple solution: require that the p-value be small enough to reduce the false positives to the desired level.
- This is called the Bonferroni correction.
- In our case, we would only accept terms with a

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

- 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”.

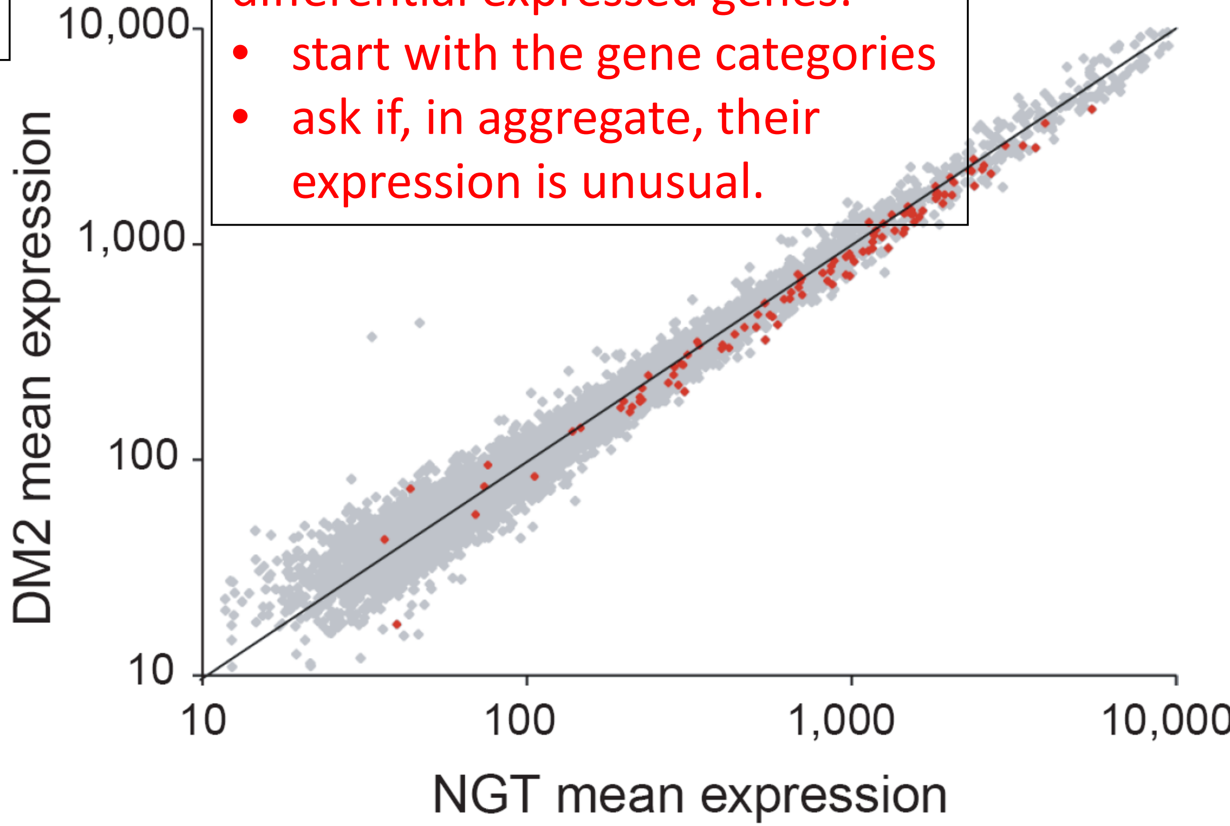
Aggregate score statistics

My results depend on how I defined "differentially expressed"



Instead of starting with differential expressed genes:

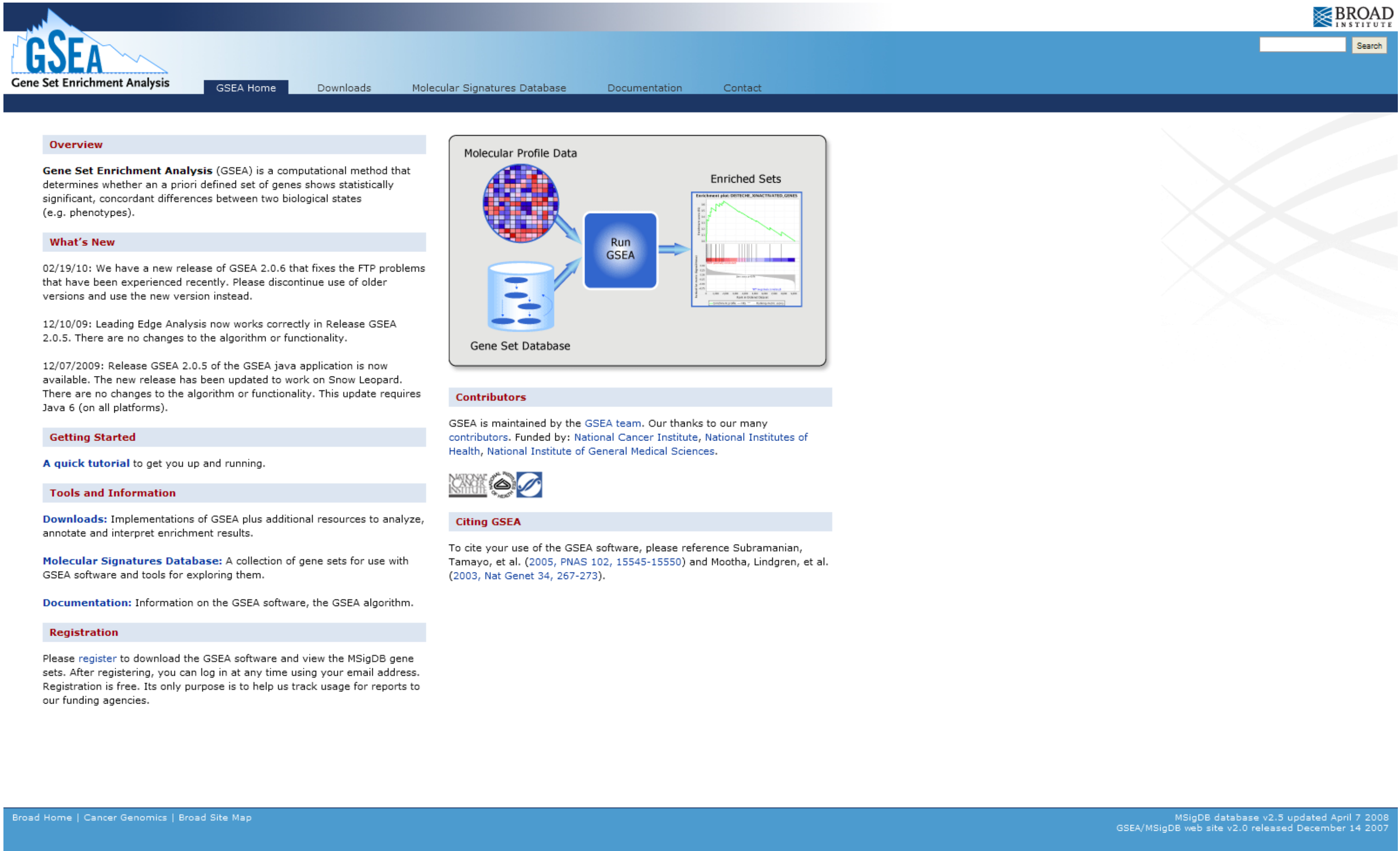
- start with the gene categories
- ask if, in aggregate, their expression is unusual.



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

Aggregate score statistics

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



Overview

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

What's New

02/19/10: We have a new release of GSEA 2.0.6 that fixes the FTP problems that have been experienced recently. Please discontinue use of older versions and use the new version instead.

12/10/09: Leading Edge Analysis now works correctly in Release GSEA 2.0.5. There are no changes to the algorithm or functionality.

12/07/2009: Release GSEA 2.0.5 of the GSEA java application is now available. The new release has been updated to work on Snow Leopard. There are no changes to the algorithm or functionality. This update requires Java 6 (on all platforms).

Getting Started

A quick tutorial to get you up and running.

Tools and Information

Downloads: Implementations of GSEA plus additional resources to analyze, annotate and interpret enrichment results.

Molecular Signatures Database: A collection of gene sets for use with GSEA software and tools for exploring them.

Documentation: Information on the GSEA software, the GSEA algorithm.

Registration

Please register to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Contributors

GSEA is maintained by the GSEA team. Our thanks to our many contributors. Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

Citing GSEA

To cite your use of the GSEA software, please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273).

Broad Home | Cancer Genomics | Broad Site Map

MSigDB database v2.5 updated April 7 2008
GSEA/MSigDB web site v2.0 released December 14 2007