Final Exam

- Tuesday, May 13 from 8 – 10 am
- The exam will be on material from the “GWAS-3” lecture and later lectures. This is the material that was **not** on the midterm exam
- Next lecture (Thurs., May 1) we will review for the final exam

Questions?

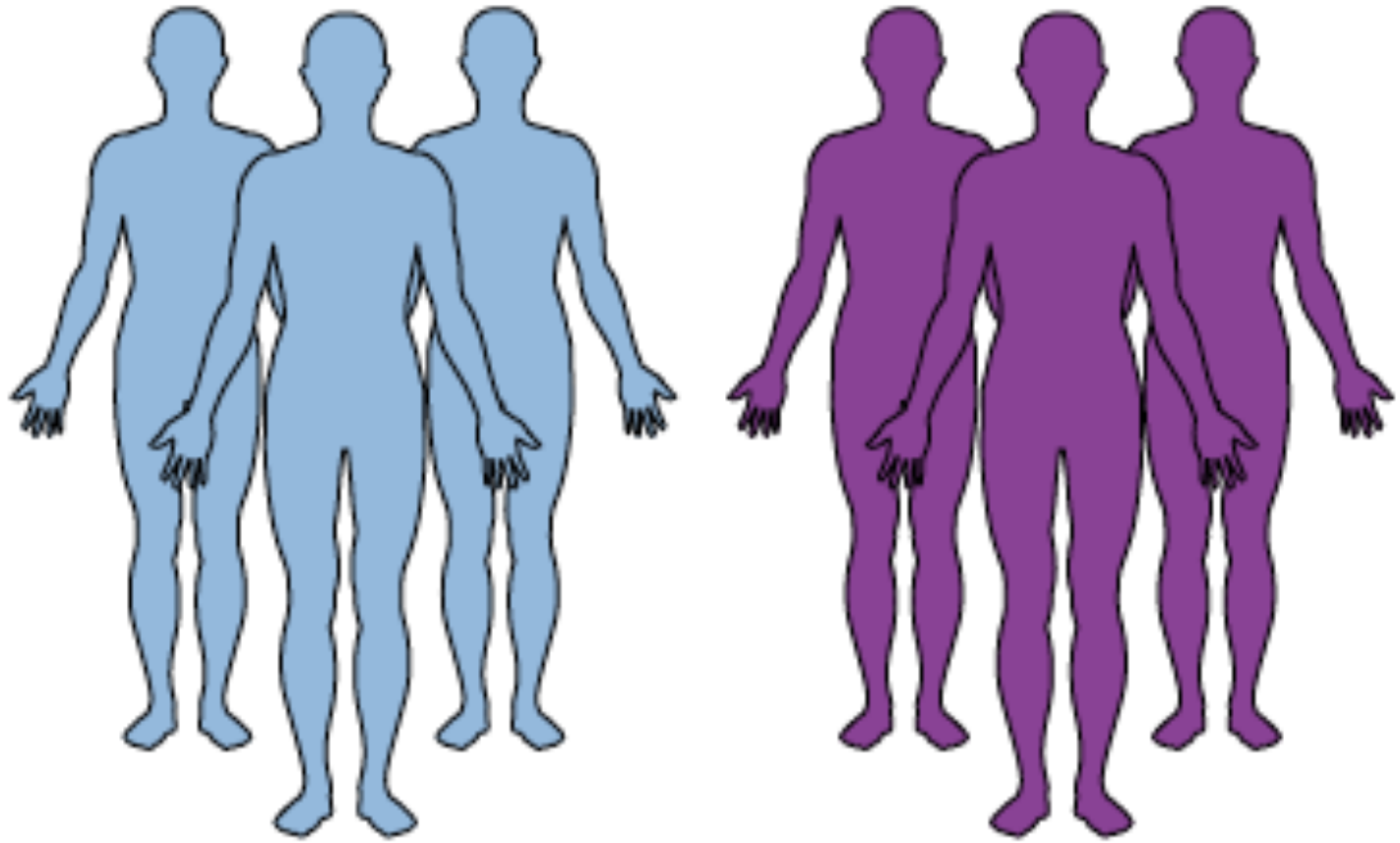
GWAS-3 and GWAS-4

Genome Wide Association Studies (GWAS)

- Use SNP chips to genotype ~1,000 to ~1,000,000 million unrelated individuals at ~1,000,000 SNPs

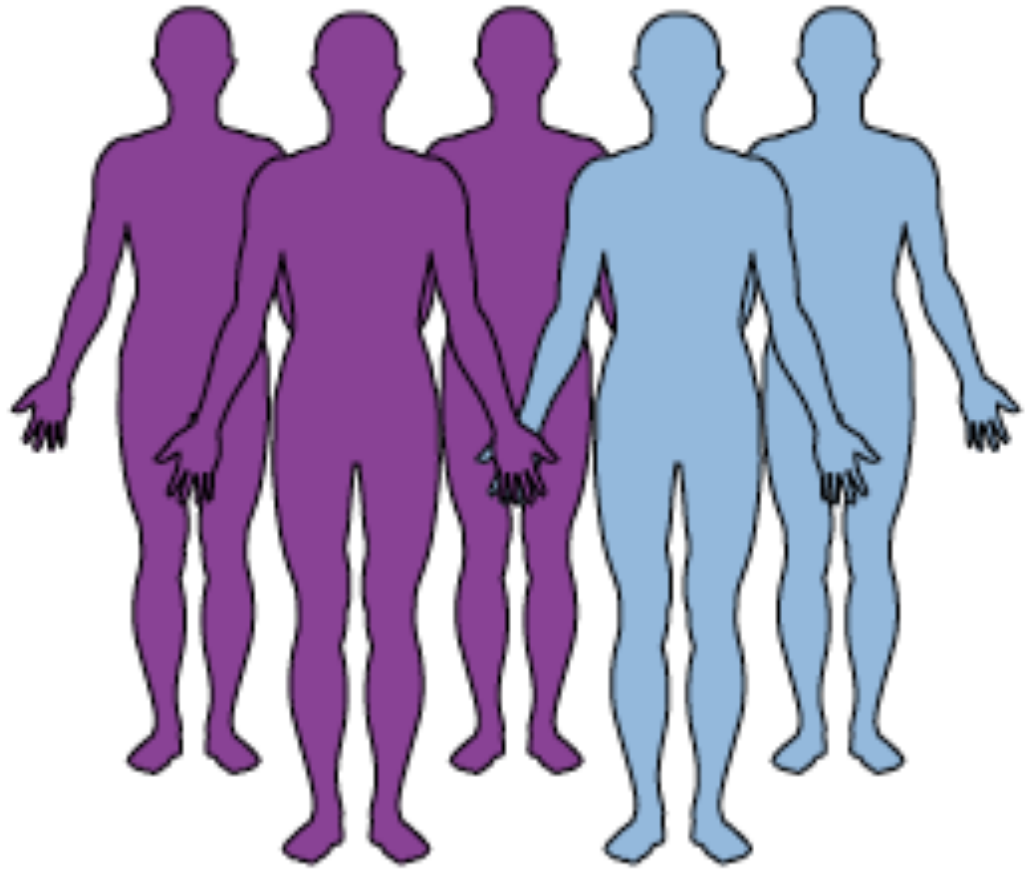
- Goal: find regions of the genome that are associated with different traits and diseases

Disease



Controls Cases

Trait



Unselected sample

- We discussed population structure, polygenic risk scores, and pleiotropy

- We discussed 10 different variations on the standard GWAS study design

We also discussed the missing heritability problem

The case of the missing heritability

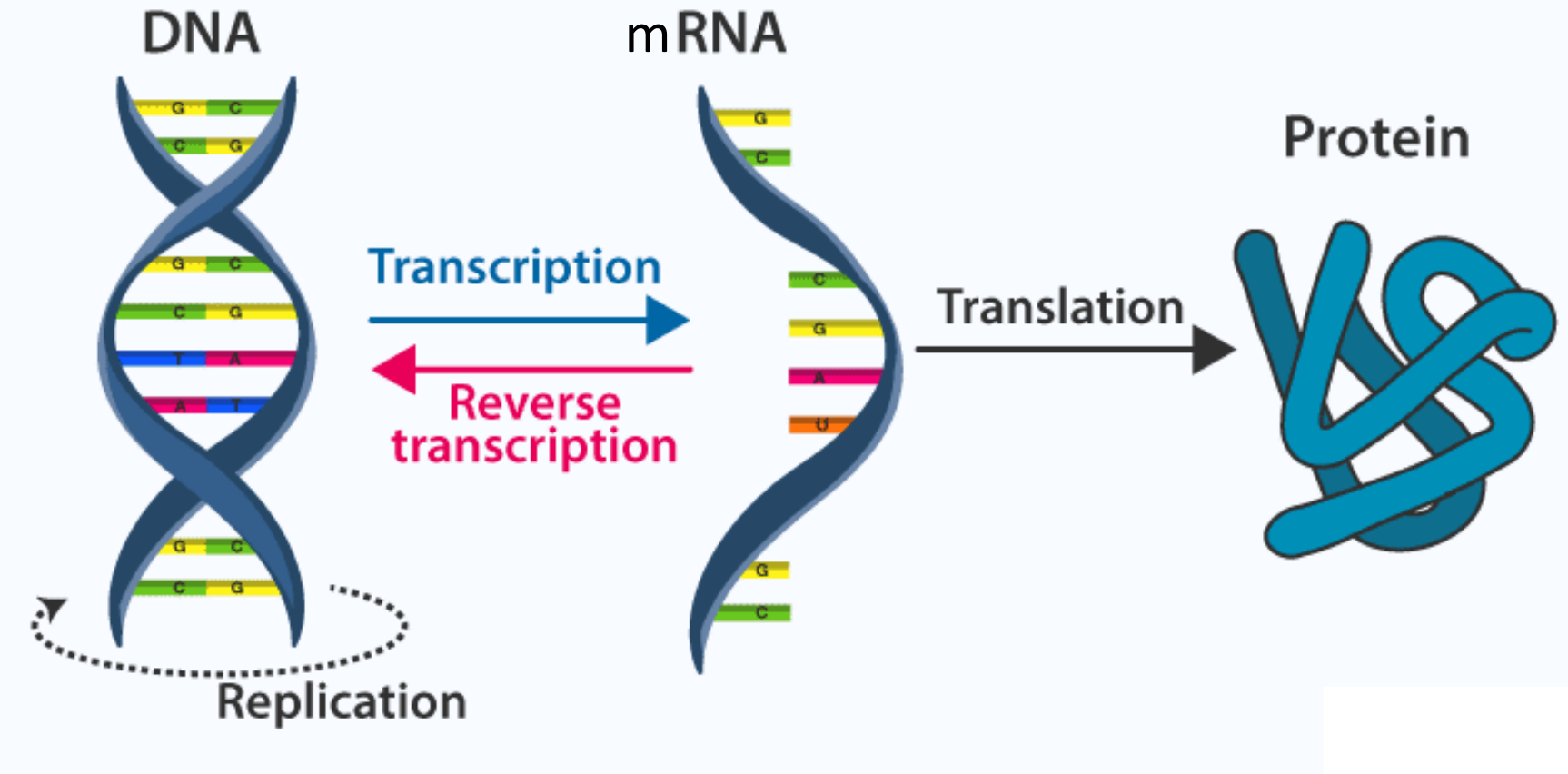
When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Mark's lectures

- Mark gave the following lectures:
  - Storing petabytes of DNA and other databases
  - Sequence evolution and function (2 lectures)
  - LCP
  - Comparative genomics and multiple sequence alignment
- I am not reviewing Mark's lectures today, but the material in these lectures is fair game for the exam

RNA-seq (4 lectures)

Central Dogma of Molecular Biology

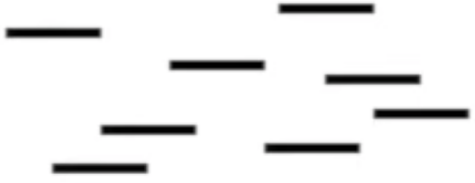


Preparing an RNA-seq library

Step 1: Isolate the RNA

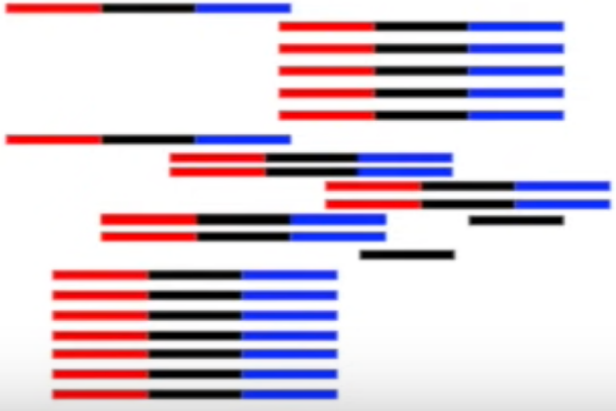
Step 2: Break the RNA into small fragments.

Step 3: Convert the RNA fragments into double stranded DNA.



Step 4: Add sequencing adaptors.

Step 5: PCR amplify.



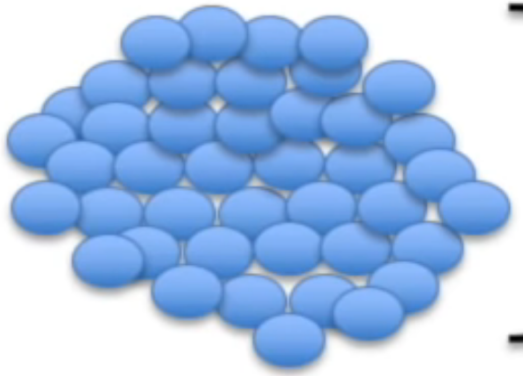
Step 6: QC

1) Verify library concentration

2) Verify library fragment lengths

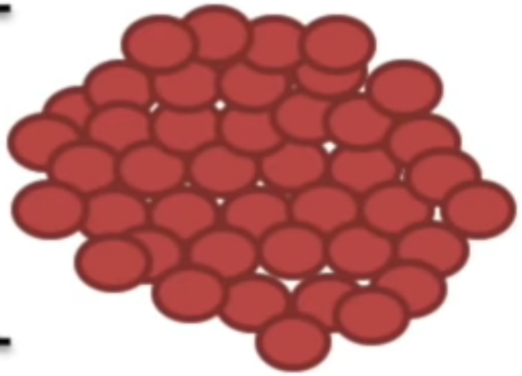
= a normal neural cell

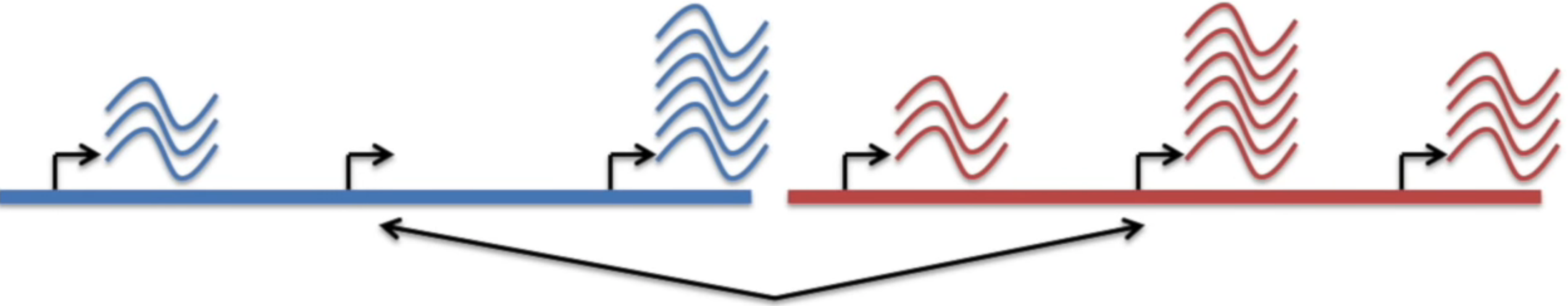
= a mutated neural cell



A bunch of normal neural cells.

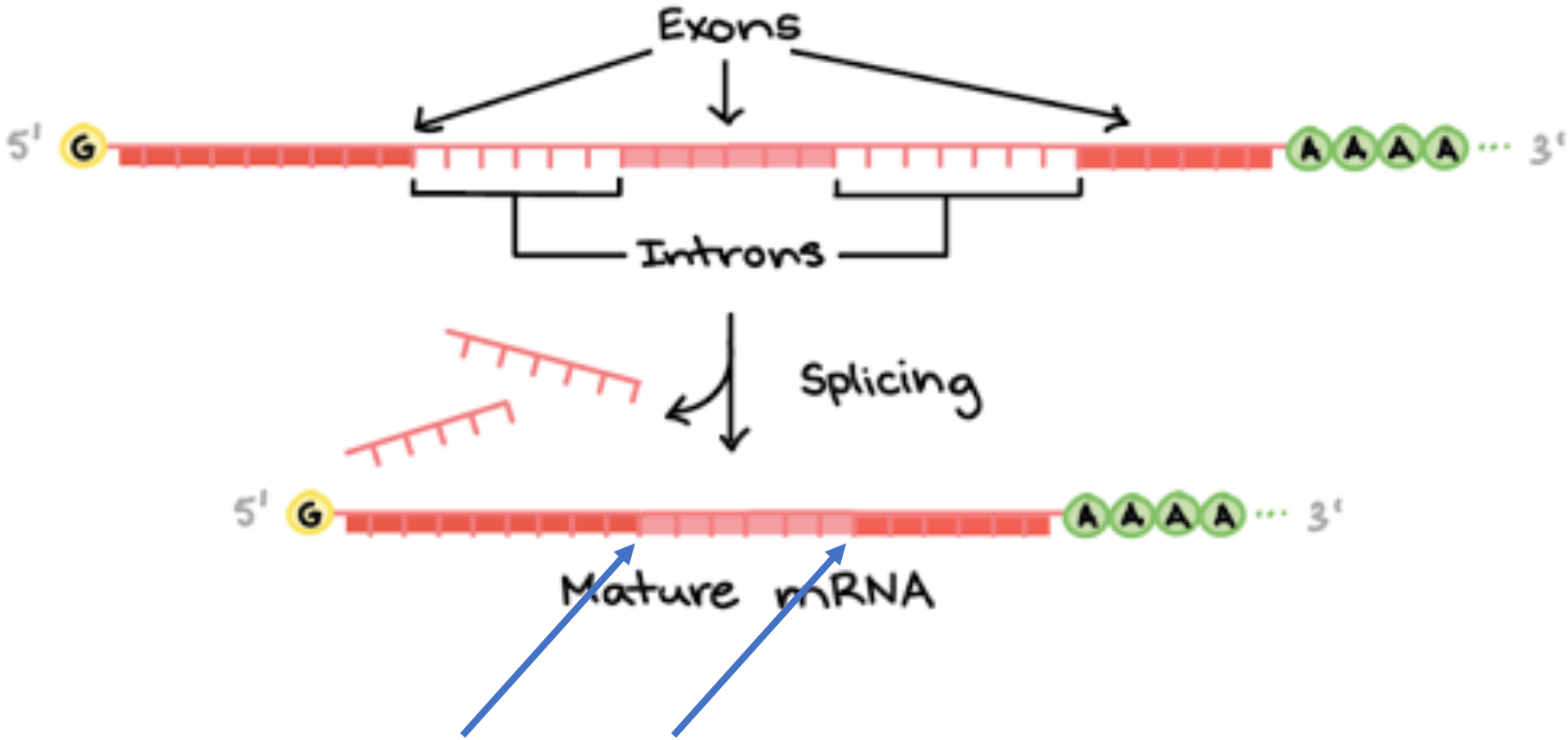
A bunch of mutated neural cells.

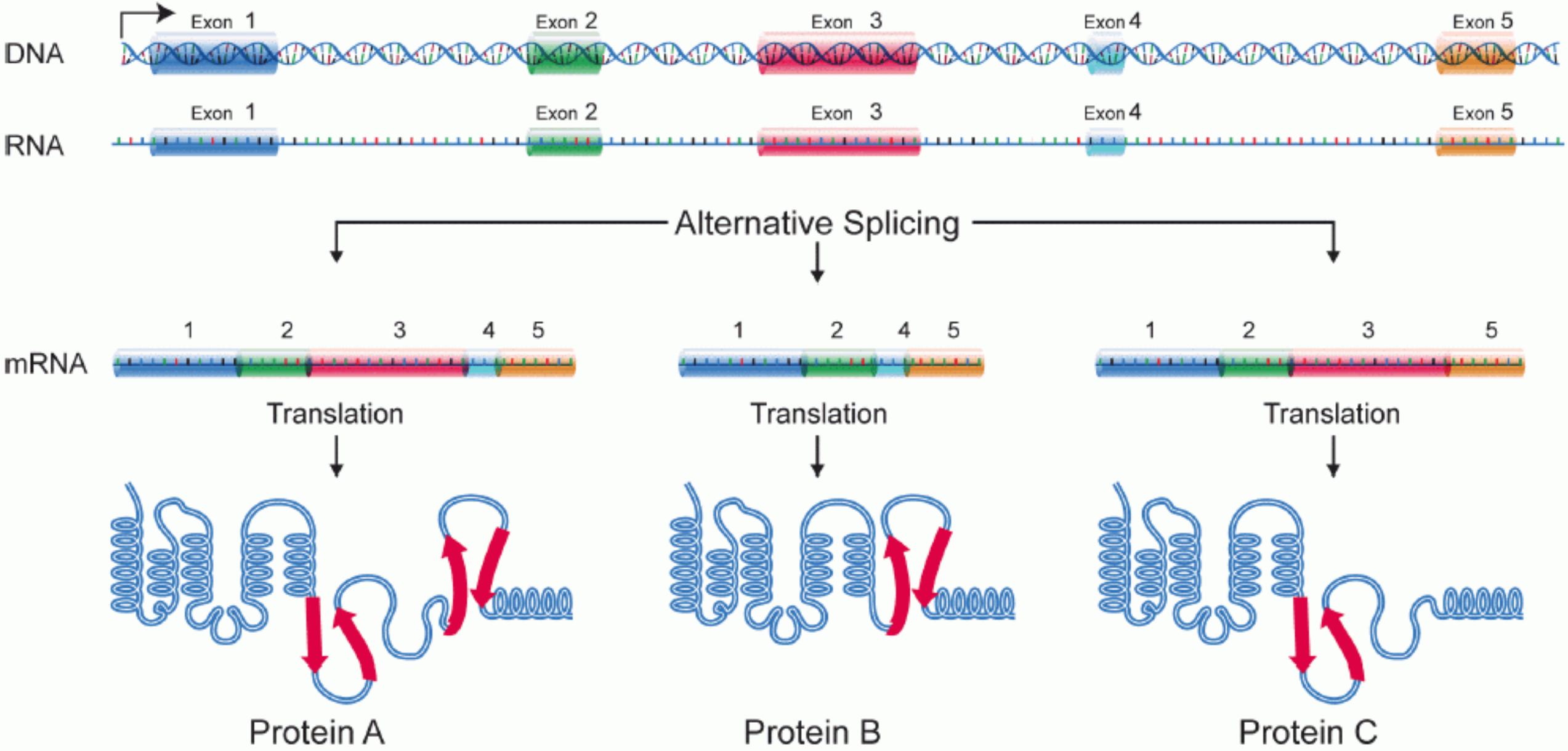


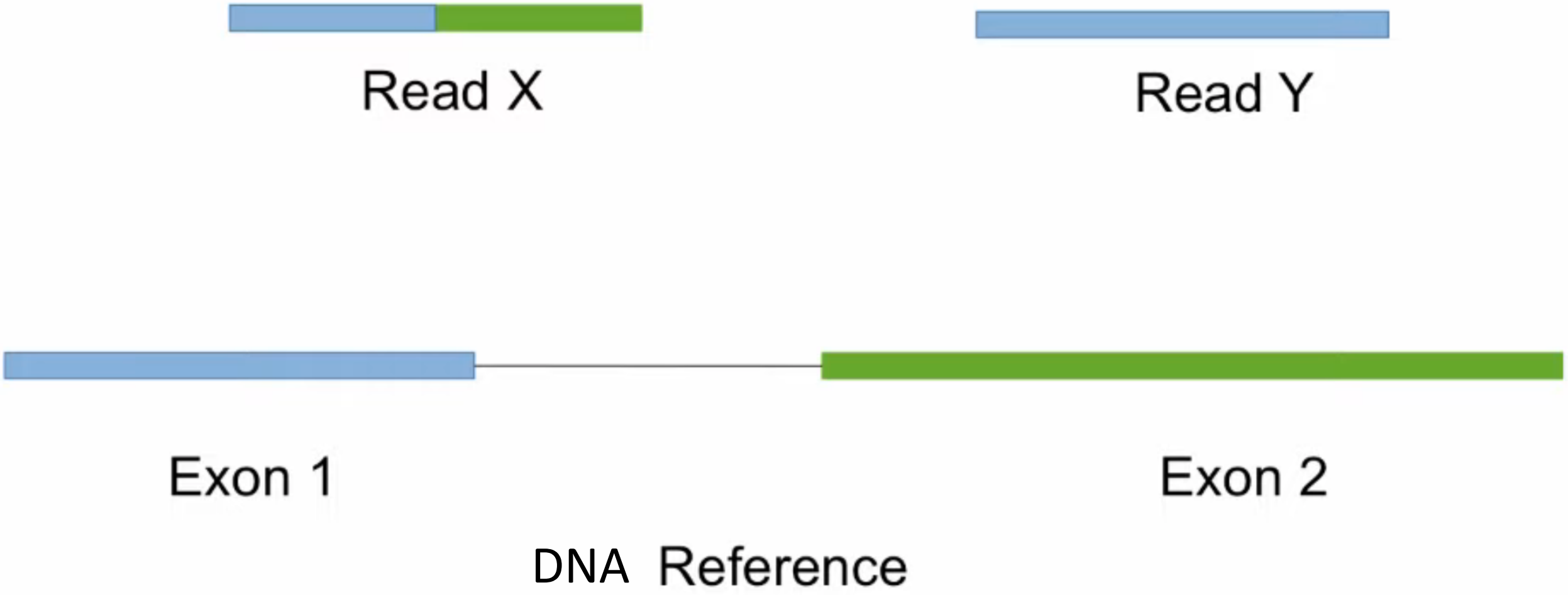


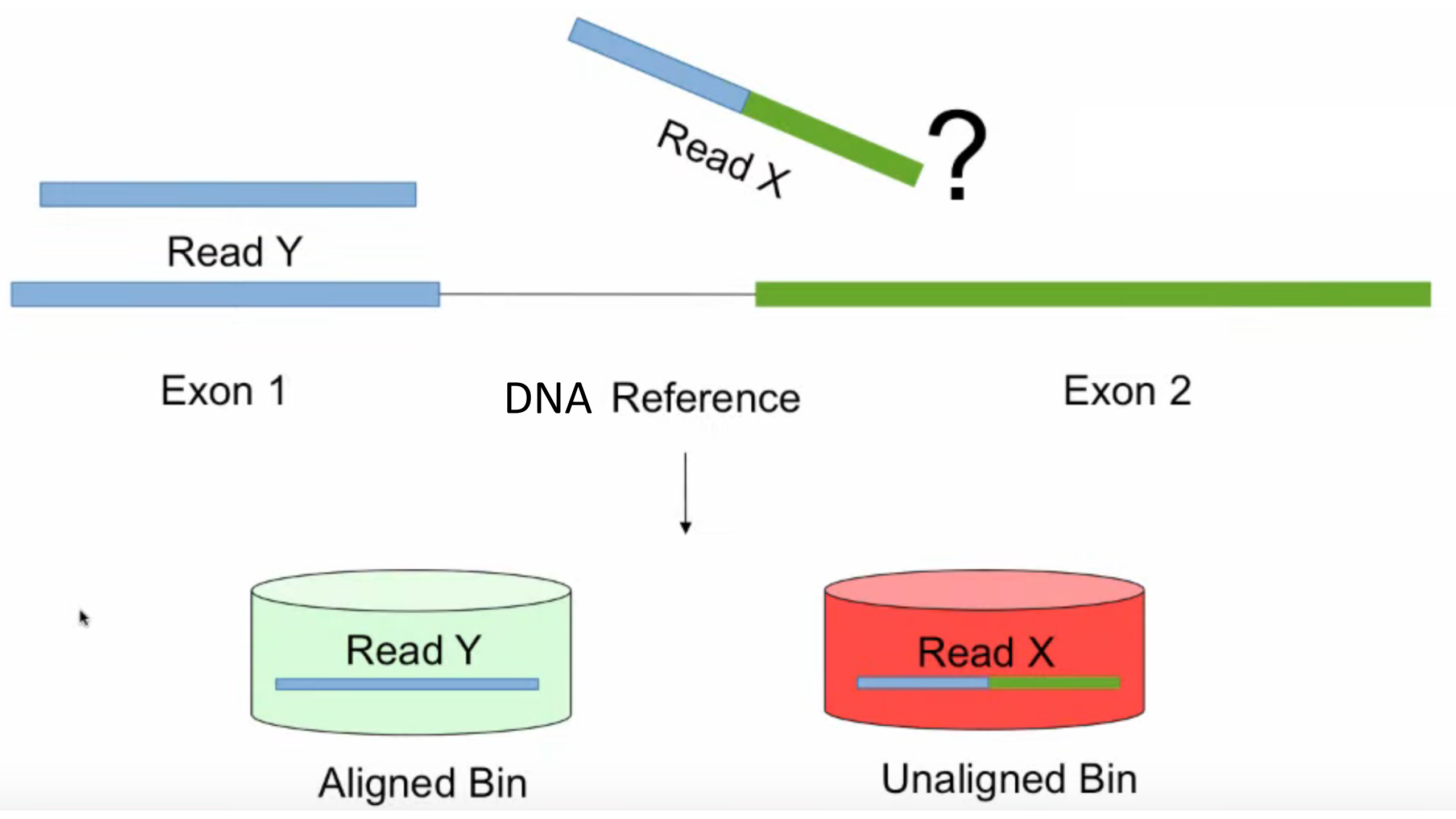
Then we can compare the two cell types and figure out what's different in the mutated cells.

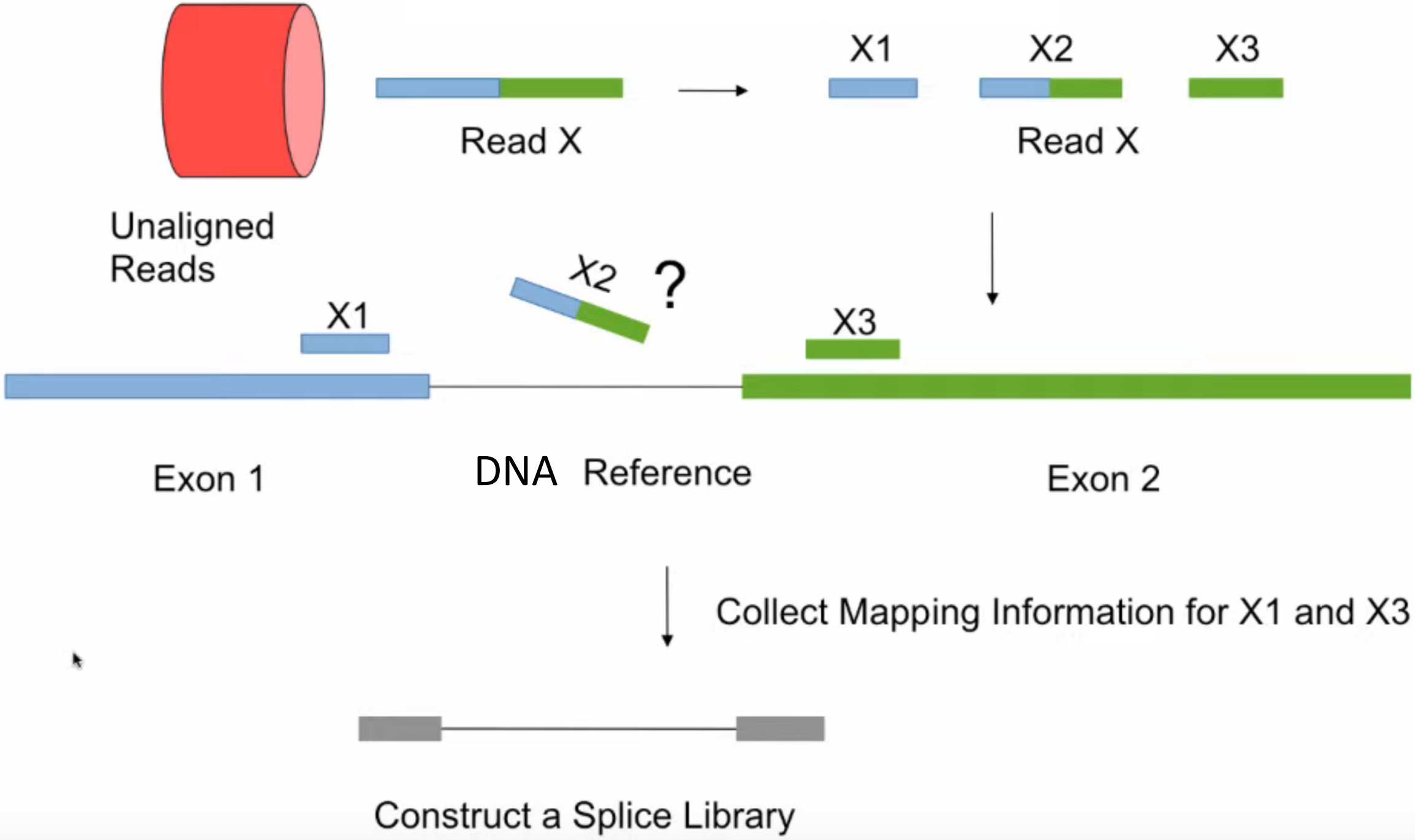
Splice junctions




DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C


Read X
Read Y
Exon 1
Exon 2
DNA Reference

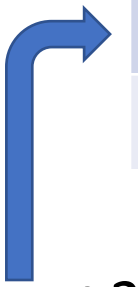

Read X
?
Read Y
Exon 1
DNA Reference
Exon 2
Read Y
Aligned Bin
Read X
Unaligned Bin


X1
X2
X3
Read X
Read X
Unaligned
Reads
X2
?
X1
X3
Exon 1
DNA Reference
Exon 2
Collect Mapping Information for X1 and X3
Construct a Splice Library

- We discussed using paired-end sequencing and longer-read sequencing technologies

Normalization

Gene Name	Sample 1 counts	Sample 2 counts	Sample 3 counts
A (2 kb)	10,000	12,000	30,000
B (4 kb)	20,000	25,000	60,000
C (1 kb)	5,000	8,000	15,000
D (10 kb)	10,000	0	1,000
…	…	…	…
Total mapped reads	35,000,000	45,000,000	106,000,000



~20,000 human genes

Genes have different sizes

Gene Name	Sample 1 counts	Sample 2 counts	Sample 3 counts
A (2 kb)	10,000	12,000	30,000
B (4 kb)	20,000	25,000	60,000
C (1 kb)	5,000	8,000	15,000
D (10 kb)	10,000	0	1,000
…	…	…	…
Total mapped reads	35,000,000	45,000,000	106,000,000

Samples have different numbers of mapped reads

Gene Name	Sample 1 counts	Sample 2 counts	Sample 3 counts
A (2 kb)	10,000	12,000	30,000
B (4 kb)	20,000	25,000	60,000
C (1 kb)	5,000	8,000	15,000
D (10 kb)	10,000	0	1,000
…	…	…	…
Total mapped reads	35,000,000	45,000,000	106,000,000

It could be that Sample 2 had poorer sequence quality so fewer reads were aligned
It could be that we used fewer sequencing resources for Sample 2 (shared a lane, smaller run, etc.)

RPKM

- Reads Per Kilobase Million
  - Kilobase refers to gene length
  - Million refers to sequencing depth

RPKM normalizes for read depth **first**

Gene Name	Sample 1	Sample 2	Sample 3
A (2 kb)	10,000 / 35 = 285.7	12,000 / 45 = 266.7	30,000 / 106 = 283.0
B (4 kb)	20,000 / 35 = 571.4	25,000 / 45 = 555.6	60,000 / 106 = 566.0
C (1 kb)	5,000 / 35 = 142.9	8,000 / 45 = 177.8	15,000 / 106 = 141.5
D (10 kb)	10,000 / 35 = 285.7	0 / 45 = 0	1,000 / 106 = 9.4
…	…	…	…
Total mapped reads	35,000,000	45,000,000	106,000,000
Scaling factor	35,000,000/1,000,000 = 35	45,000,000/1,000,000 = 45	106,000,000/1,000,000 = 106

RPKM normalizes for gene length **second**

Gene Name	Sample 1	Sample 2	Sample 3
A (2 kb)	285.7 / 2 = 142.9	266.7 / 2 = 133.4	283.0 / 2 = 141.5
B (4 kb)	571.4 / 4 = 142.9	555.6 / 4 = 138.9	566.0 / 4 = 141.5
C (1 kb)	142.9 / 1 = 142.9	177.8 / 1 = 177.8	141.5 / 1 = 141.5
D (10 kb)	285.7 / 10 = 28.6	0 / 10 = 0	9.4 / 10 = 0.94
…	…	…	…
Total mapped reads	35,000,000	45,000,000	106,000,000

RPKM

Gene Name	Sample 1 normalized counts	Sample 2 normalized counts	Sample 3 normalized counts
A (2 kb)	142.9	133.4	141.5
B (4 kb)	142.9	138.9	141.5
C (1 kb)	142.9	177.8	141.5
D (10 kb)	28.6	0	0.94
…	…	…	…

FPKM

- Fragments Per Kilobase Million. Very similar to RPKM
- RPKM is for single-end reads
  - For single-end, a read maps to a gene or it doesn’t
- FPKM is for paired-end reads
  - For paired-end, both reads might map to a gene
  - It is also possible one read maps to a gene and the other read has poor sequence quality and doesn’t map
  - In both cases, count **once**

TPM

- Transcripts Per Million
- The order of operations is switched

TPM normalizes for gene length **first**

Gene Name	Sample 1	Sample 2	Sample 3
A (2 kb)	10,000 / 2 = 5,000	12,000 / 2 = 6,000	30,000 / 2 = 15,000
B (4 kb)	20,000 / 4 = 5,000	25,000 / 4 = 6,250	60,000 / 4 = 15,000
C (1 kb)	5,000 / 1 = 5,000	8,000 / 1 = 8,000	15,000 / 1 = 15,000
D (10 kb)	10,000 / 10 = 1,000	0 / 10 = 0	1,000 / 10 = 100
…	…	…	…
~~Total mapped reads~~ Column sum For this example, we don't know "…" so we can't compute. For illustration assume the sum of normalized numbers is:	~~35,000,000~~ 5,000 +5,000 +5,000 +1,000 +… = 8,000,000	~~45,000,000~~ 6,000 +6,250 +8,000 +0 +…= 11,000,000	~~106,000,000~~ 15,000 +15,000 +15,000 +100 +…= 25,000,000

TPM normalizes for sequencing depth **second**

Gene Name	Sample 1	Sample 2	Sample 3
A (2 kb)	5,000 / 8 = 625	6,000 / 11 = 545.5	15,000 / 25 = 600
B (4 kb)	5,000 / 8 = 625	6,250 / 11 = 568.2	15,000 / 25 = 600
C (1 kb)	5,000 / 8 = 625	8,000 / 11 = 727.3	15,000 / 25 = 600
D (10 kb)	1,000 / 8 = 125	0 / 11 = 0	100 / 25 = 4
…	…	…	…
Column sum	8,000,000	11,000,000	25,000,000
Scaling factor	8,000,000 / 1,000,000 = 8	11,000,000 / 1,000,000 = 11	25,000,000 / 1,000,000 = 25

TPM

Gene Name	Sample 1 normalized counts	Sample 2 normalized counts	Sample 3 normalized counts
A (2 kb)	625	545.5	600
B (4 kb)	625	568.2	600
C (1 kb)	625	727.3	600
D (10 kb)	125	0	4
…	…	…	…

RPKM

Gene Name	Sample 1	Sample 2	Sample 3
A (2 kb)	142.9	133.4	141.5
B (4 kb)	142.9	138.9	141.5
C (1 kb)	142.9	177.8	141.5
D (10 kb)	28.6	0	0.94
…	…	…	…

TPM

Gene Name	Sample 1	Sample 2	Sample 3
A (2 kb)	625	545.5	600
B (4 kb)	625	568.2	600
C (1 kb)	625	727.3	600
D (10 kb)	125	0	4
…	…	…	…

What's the difference?

- For the final normalized counts for TPM, each of the columns sums to **one million**

- This makes it easier to compare different samples

- For example,
  - in Sample 1, 625 out of one million reads map to Gene A
  - in Sample 3, 600 out of one million reads map to Gene A

- For the final normalized counts for RPKM, all of the column sums are **different**
- So it is harder to compare samples for RPKM

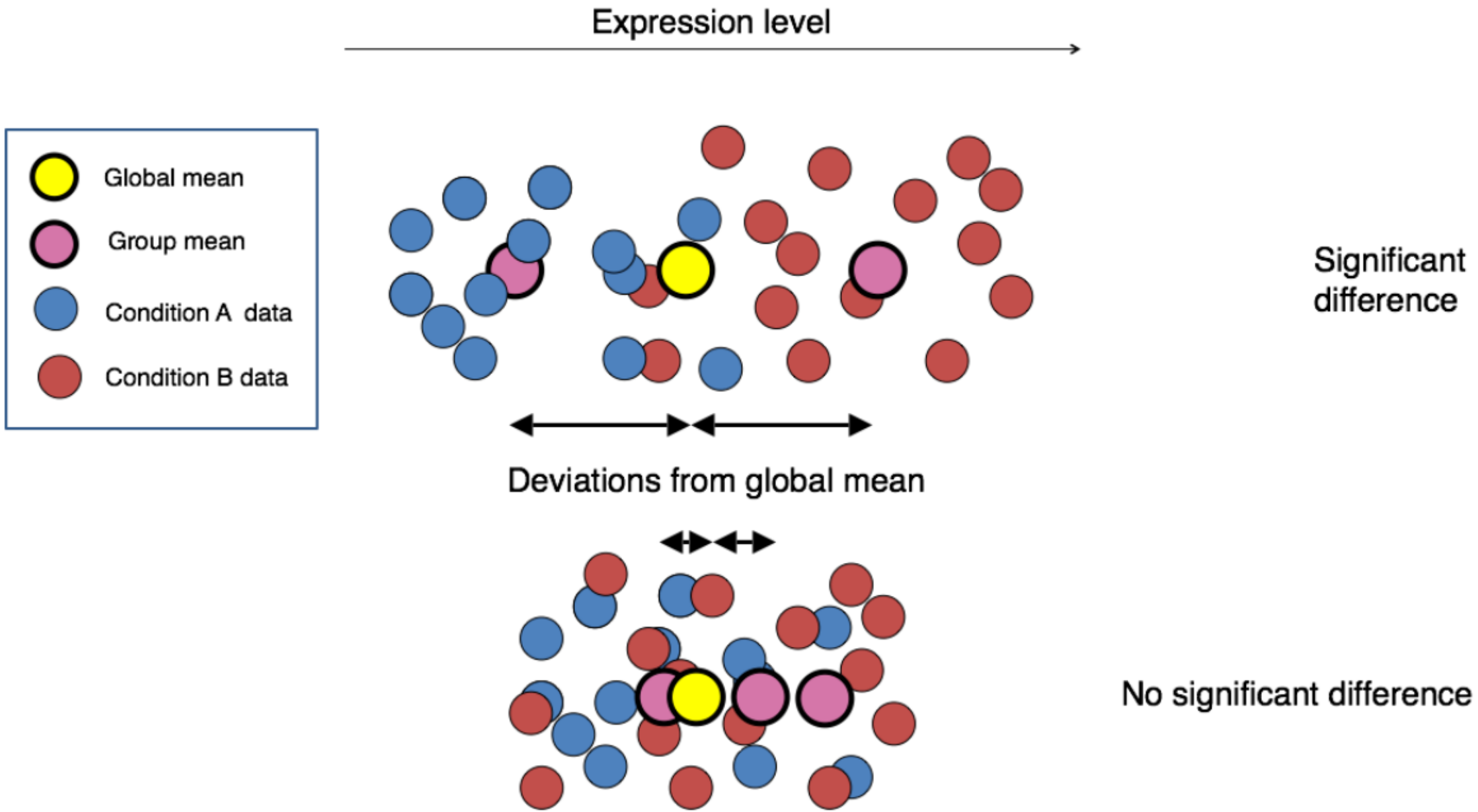
- You will still see both TPM and RPKM in research papers

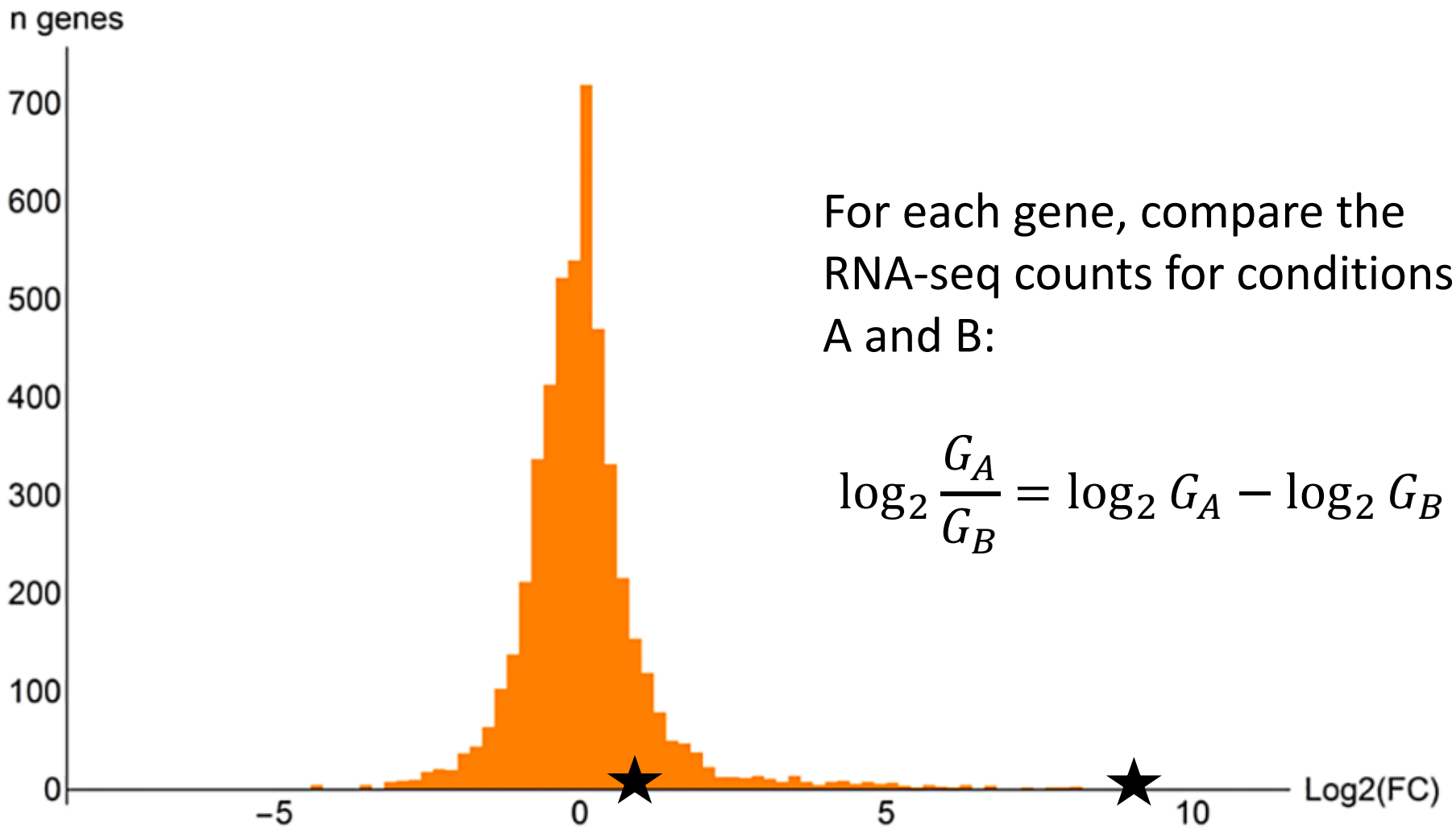
TMM

- There is yet another normalization acronym: TMM (Trimmed Mean of M-values)
- This one has many more steps to calculate and we will not do the details
- It corrects for sequencing depth and **library composition**
- It also assumes most genes will not change their expression much, and tries to limit the impact of outliers

Differential Expression

- RNA-seq under different conditions
- For example, mutant cells and wild-type cells
- Determine which (if any) genes are expressed differently under the different conditions


Expression level
Global mean
Group mean
Condition A data
Condition B data
Deviations from global mean
Significant difference
No significant difference



For each gene, compare the RNA-seq counts for conditions A and B:

$$\log_2 \frac{G_A}{G_B} = \log_2 G_A - \log_2 G_B$$

Fig 2. Histogram of RNA-seq $\log_2$(fold change) values. Statistical procedures

doi:10.1371/journal.pone.0156242.g002

- Many algorithms

- EdgeR and DESeq are two popular algorithms

- While RPKM, FPKM, TPM are often used in figures. They are **not** used in these differential expression algorithms

- These algorithms start with the raw RNA-seq counts. They then normalize similar to TMM for sequencing depth and library composition

After normalization

	Wild-type 1	Wild-type 2	Wild-type 3	Mutant 1	Mutant 2	Mutant 3
Gene A	119	95	109	79	98	112
Gene B	98	109	118	109	111	94
Gene C	101	78	116	118	116	89
Gene D	110	104	96	142	147	152
…						
Total	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000	1,000,000

Plan

- We are going to do a **separate** hypothesis test for each of the ~20,000 genes to see which genes (if any) are significantly differently expressed under the two conditions

Hypothesis Testing

- **Null Hypothesis** Gene A is **not** differentially expressed between the two conditions

- **Alternative Hypothesis** Gene A is differentially expressed between the two conditions

- **p-value** Assuming the null hypothesis is true, the probability a statistic is as extreme, or more extreme, than was observed in the data

- The idea is that if we get a small p-value then we have doubts that the assumption that the null hypothesis is true was correct

- How small is small? We select a significance level α (often = 0.05) before we see the data

- If our p-value < α then we reject the null hypothesis (**<u>significant result</u>**)
- If our p-value > α then we cannot reject the null hypothesis

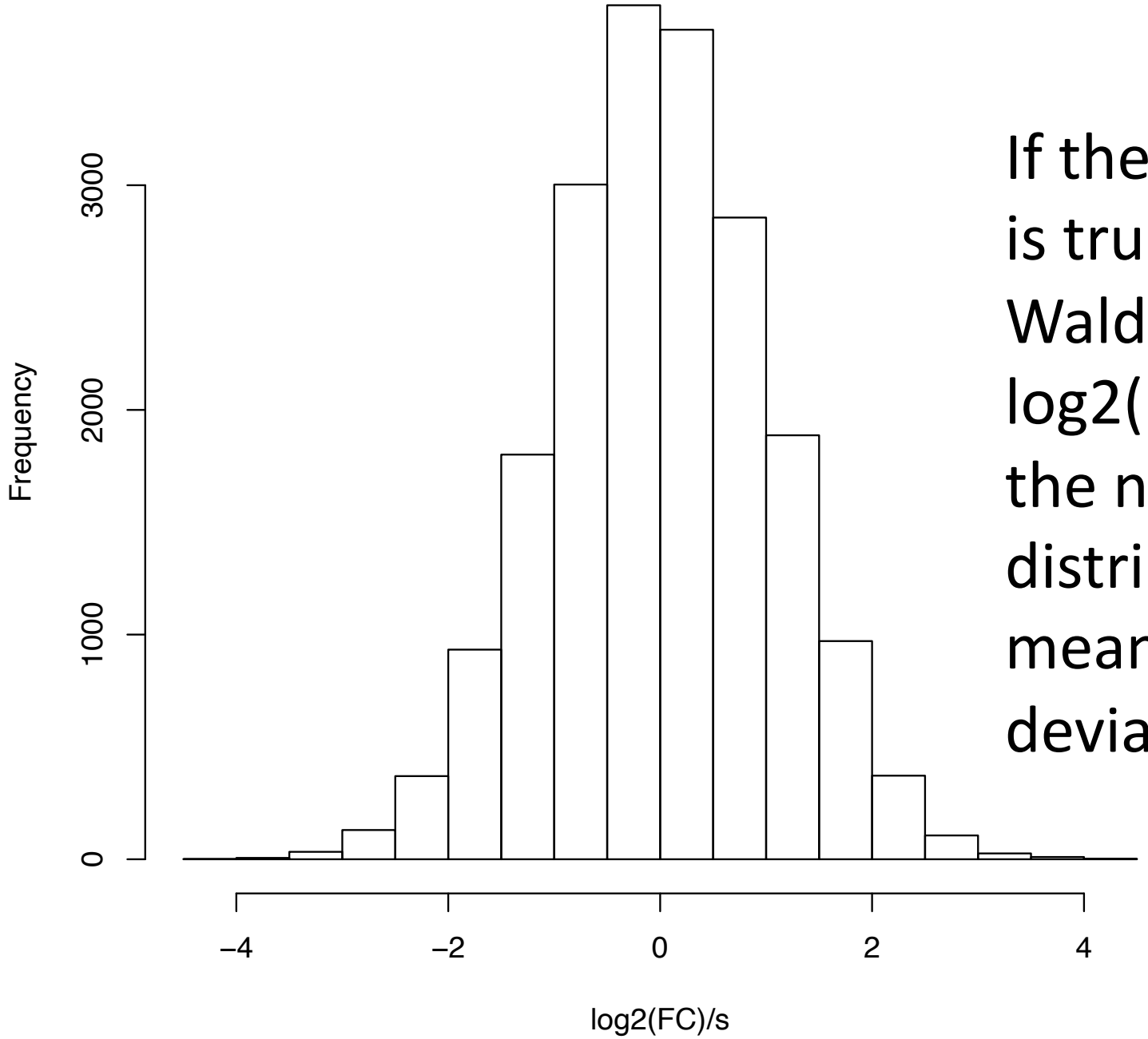
- There are many possible tests, e.g., chi-squared or t-test

- The test that is most commonly used for RNA-seq is to compute the log2 of the fold change (this is then normalized by a negative binomial fit) and then to use the Wald test

	Wild-type 1	Wild-type 2	Wild-type 3	Mutant 1	Mutant 2	Mutant 3	log2(FC)/s	p-value
Gene A	119	95	109	79	98	112	-1.00	
Gene B	98	109	118	109	111	94	-0.31	
Gene C	101	78	116	118	116	89	0.81	
Gene D	110	104	96	142	147	152	3.2	
…								

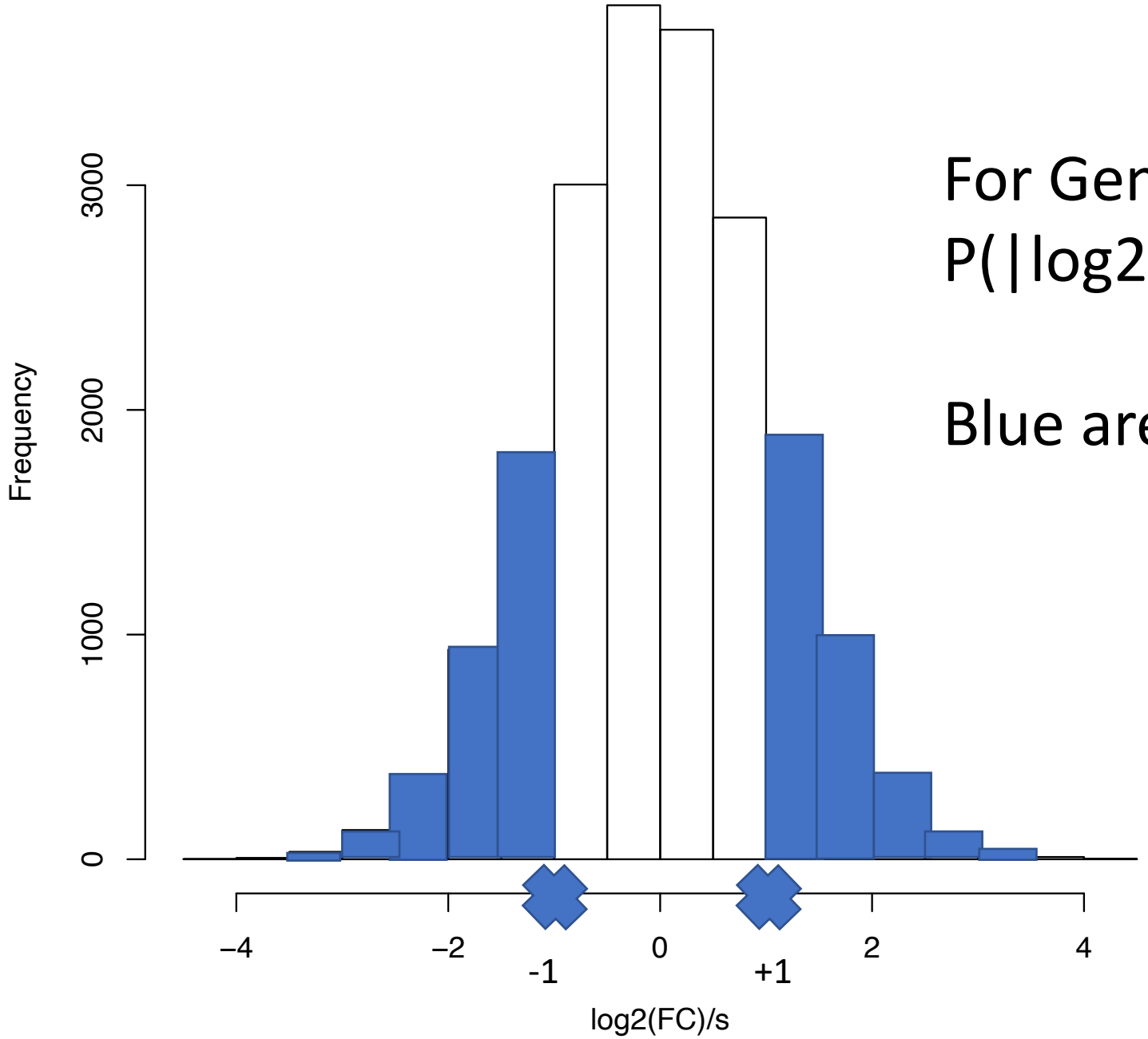
e.g., for Gene A, fold change = $\frac{79+98+112}{119+95+109}$

Then take the log2 and normalize (the s term depends on the negative binomial fit)



If the null hypothesis is true, based on the Wald test we expect log2(FC)/s will have the normal distribution (with mean=0, standard deviation=1)

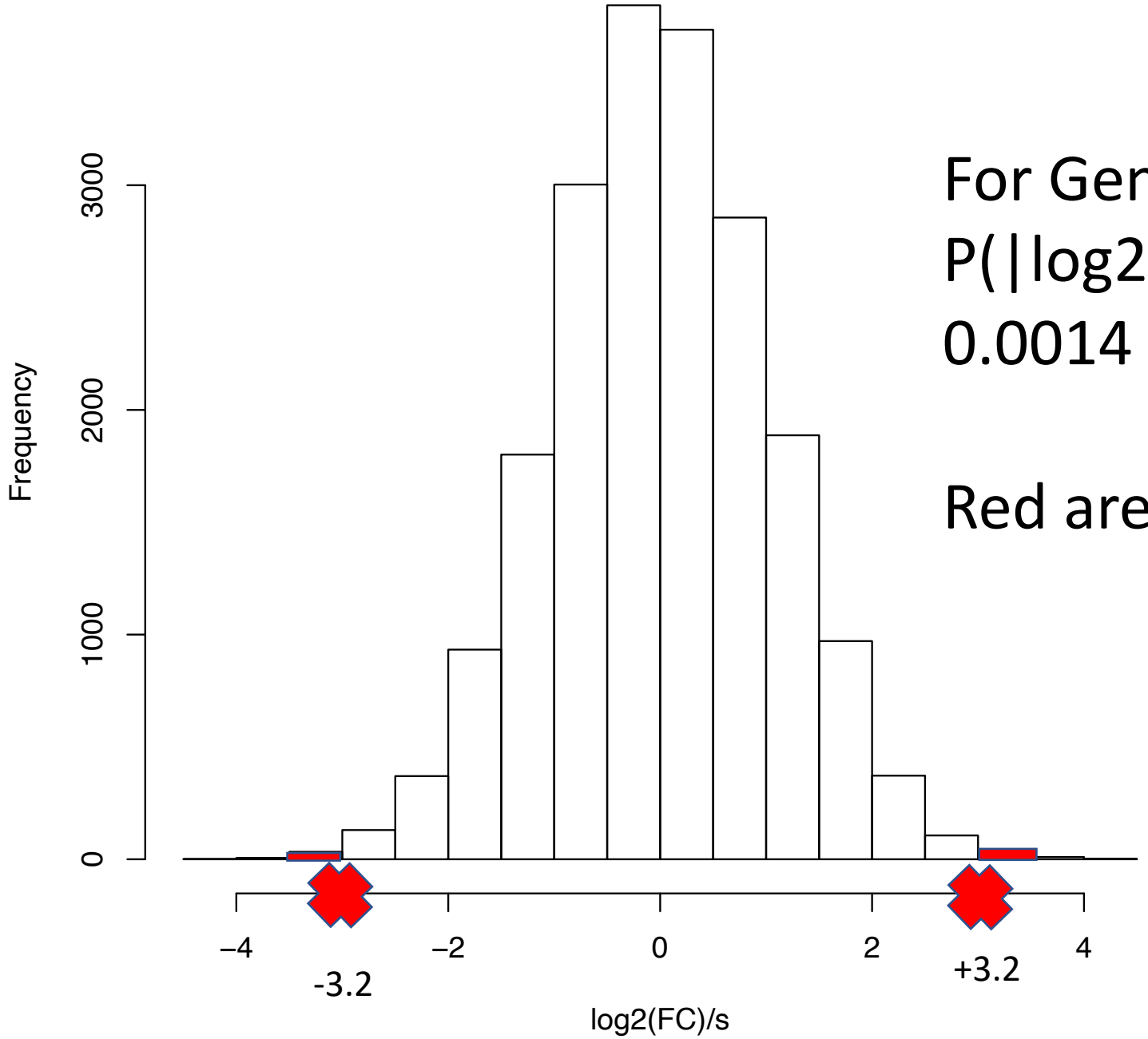
	Wild-type 1	Wild-type 2	Wild-type 3	Mutant 1	Mutant 2	Mutant 3	log2(FC)/s	p-value
Gene A	119	95	109	79	98	112	-1.00	0.31
Gene B	98	109	118	109	111	94	-0.31	0.76
Gene C	101	78	116	118	116	89	0.81	0.42
Gene D	110	104	96	142	147	152	3.2	0.0014
…								



For Gene A,
P(|log2(FC)/s|>1) = 0.31

Blue area

	Wild-type 1	Wild-type 2	Wild-type 3	Mutant 1	Mutant 2	Mutant 3	log2(FC)/s	p-value
Gene A	119	95	109	79	98	112	-1.00	0.31
Gene B	98	109	118	109	111	94	-0.31	0.76
Gene C	101	78	116	118	116	89	0.81	0.42
Gene D	110	104	96	142	147	152	3.2	0.0014
…								



For Gene D,
P(|log2(FC)/s|>3.2) = 0.0014

Red area

Multiple tests

- Since we're doing a separate test for each gene, we have a multiple tests problem
- In the GWAS lectures we discussed the Bonferroni correction
- In the RNA-seq lectures we discussed a less conservative approach called the false discovery rate (FDR)

	Null true	Null false	Total
Reject Null	V (False Positive)	S (Correct)	R
Don’t Reject Null	U (Correct)	T (False Negative)	M - R
Total	M_0	$M - M_0$	M

Suppose we do M tests
We reject the null for R tests
V of these are false positives
The **False Discovery Rate** equals V/R, the fraction of significant results that are false positives
The method we are about to show limits this rate

Benjamini-Hochberg Procedure

unadjusted sorted p-values	10-6	.01	.11	.30	.31	.41	.61	.71	.81	.91
rank	1	2	3	4	5	6	7	8	9	10
pre	10-5	.05	.37	.75	.62	.68	.87	.89	.90	.91
adjusted p-values	10-5	.05	.37	.62	.62	.68	.87	.89	.90	.91

Adjusted p-values less than the significance level α are significant

FDR vs. Bonferroni

- If we just take the “pre” value (and forget about taking the min to preserve the order of the p-values) then the adjusted p-value is approximately,

  (unadjusted p-value) x (number tests) / rank

- Another way to think about the Bonferroni correction is instead of adjusting the significance threshold we adjust the p-value (and compare to significance level α as in FDR). Viewed this way the Bonferroni adjusted p-value is,

  (unadjusted p-value) x (number tests)

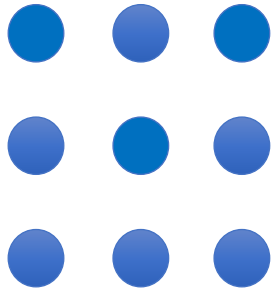
Single cell RNA-seq

- Instead of measuring the transcriptome for a collection of cells (bulk), measure it for a **single cell**

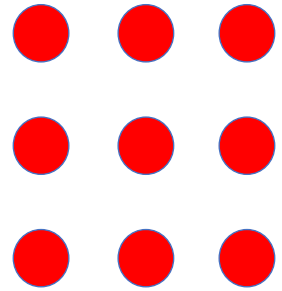
Colors indicate expression of Gene A

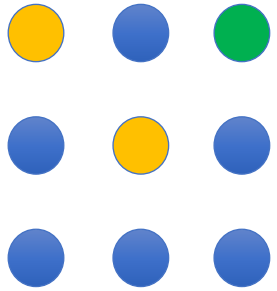
Condition #1

Condition #2

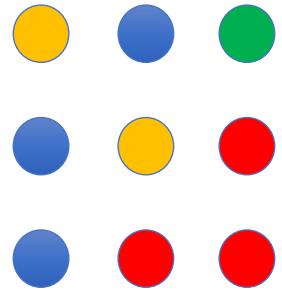


Scenario #1

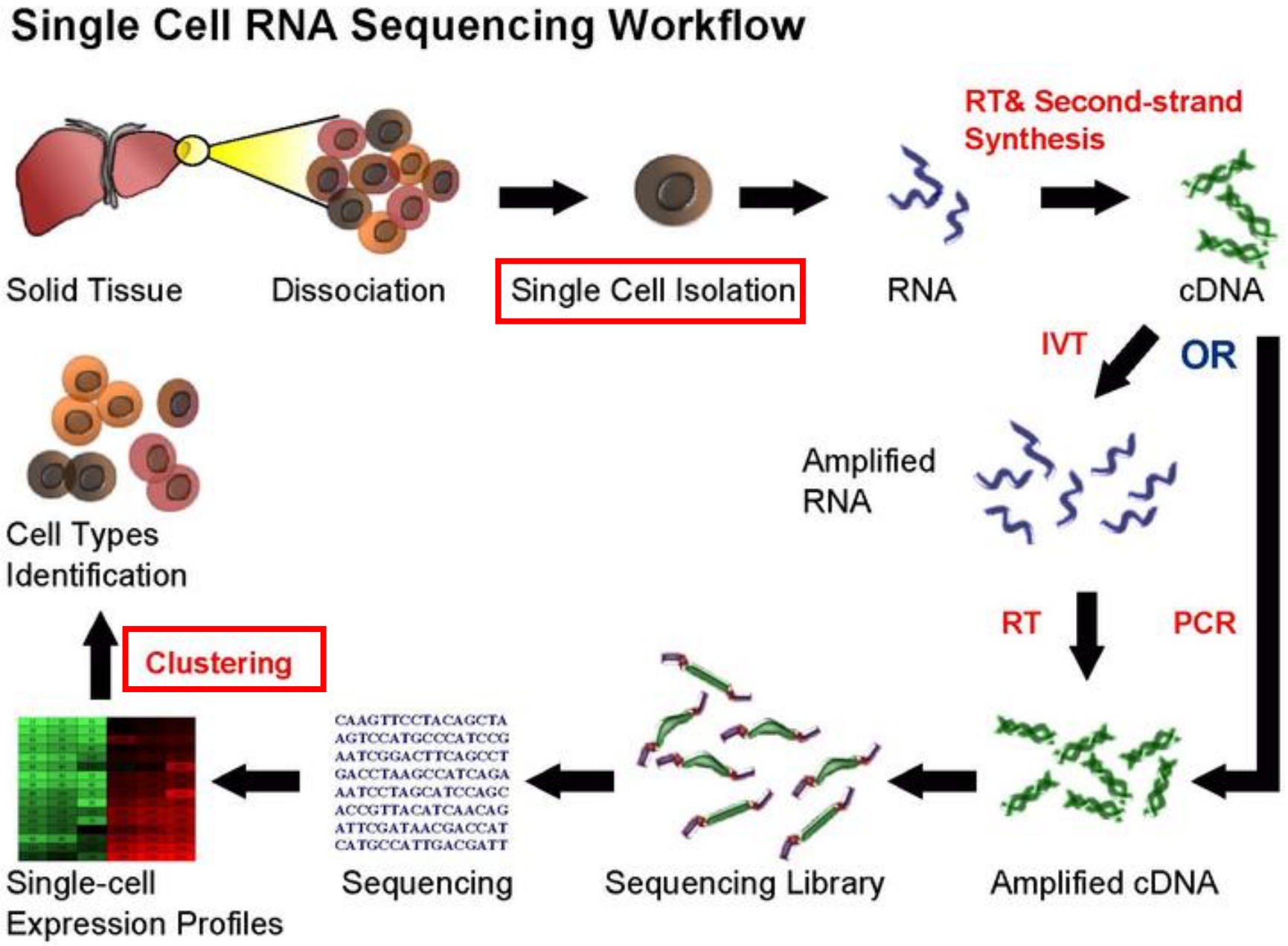




Scenario #2



Science 2018 Breakthrough of the Year

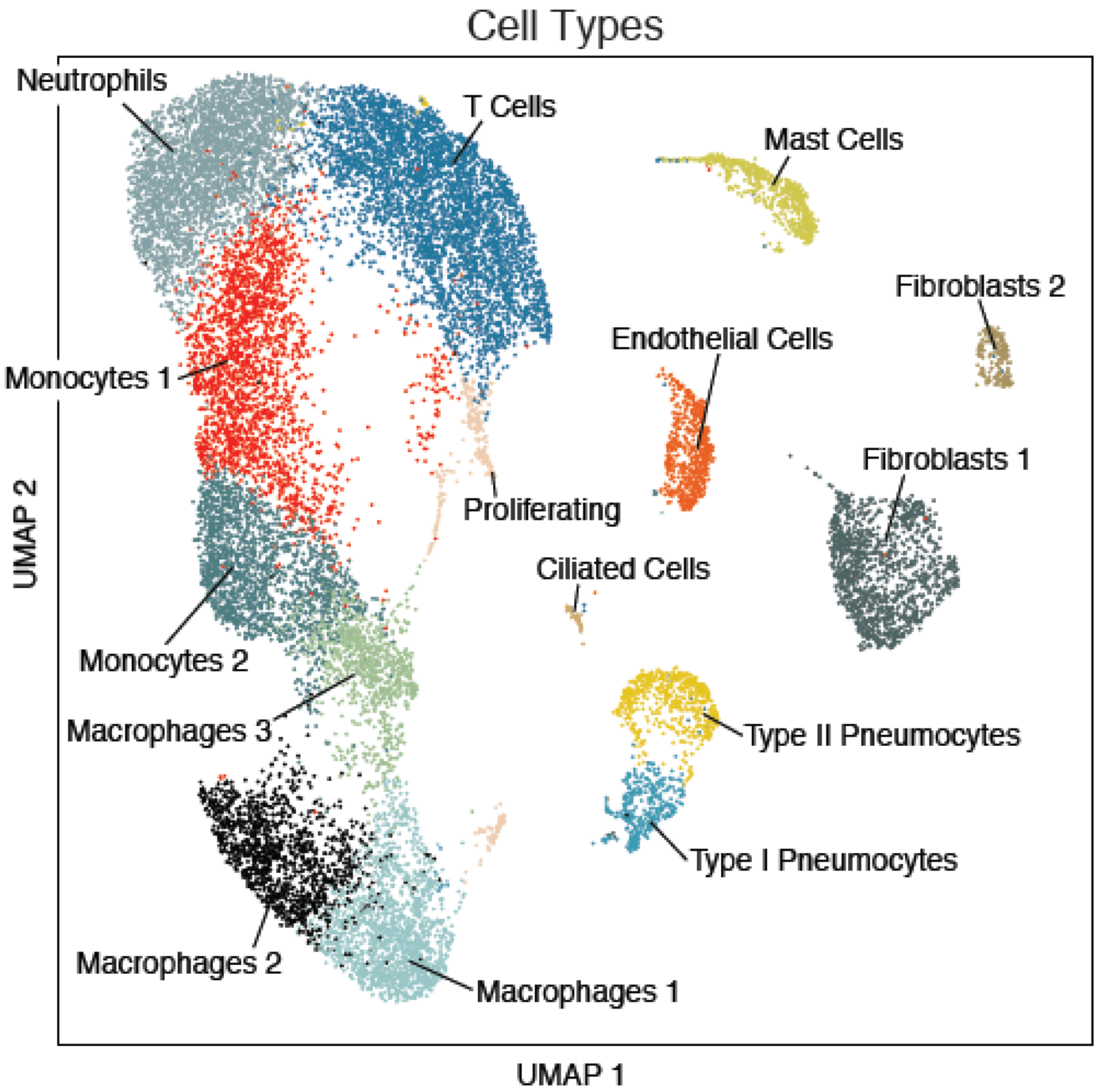


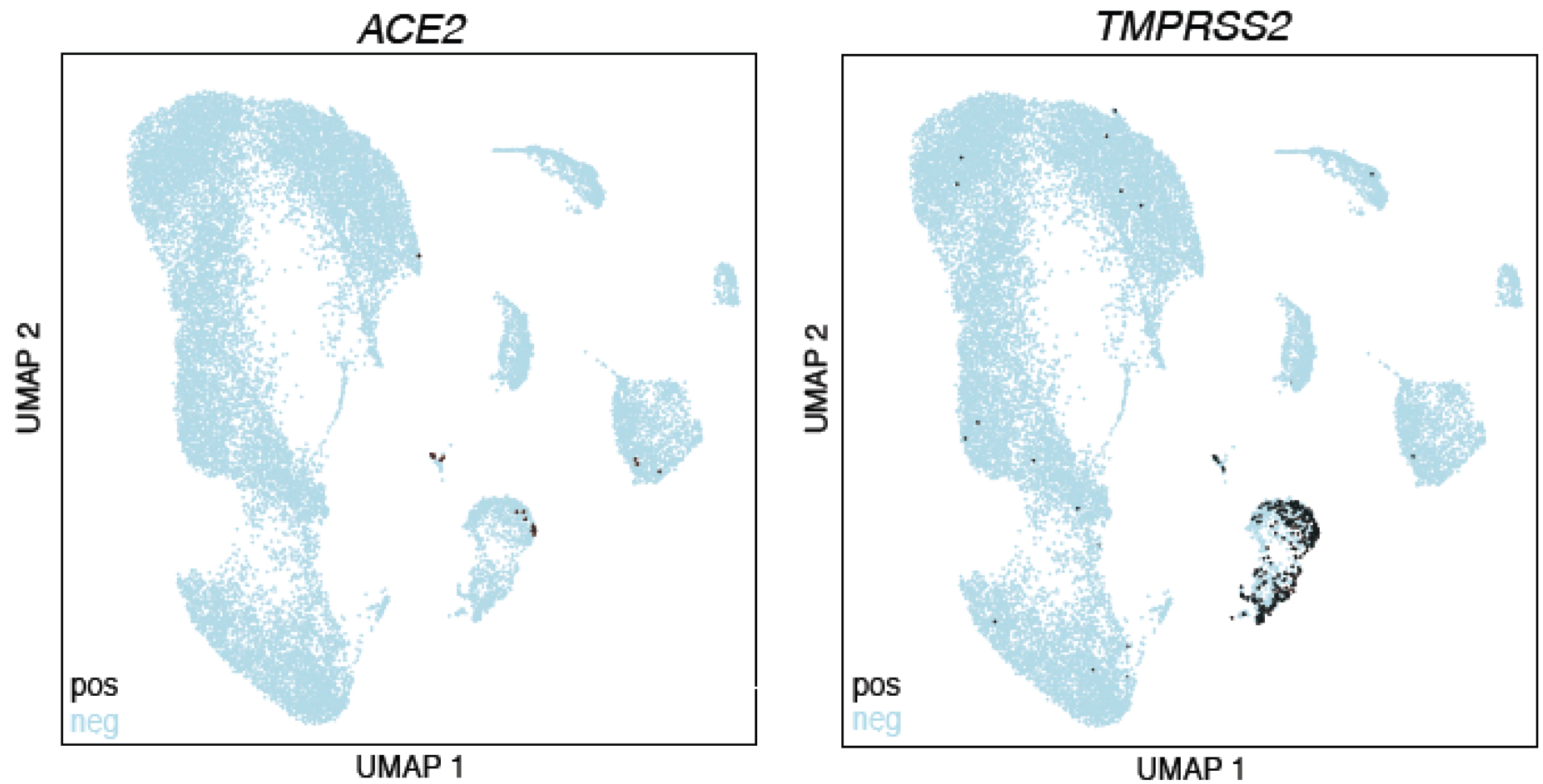
Dimension reduction

- We discussed the t-SNE dimension reduction algorithm
- The UMAP algorithm is similar

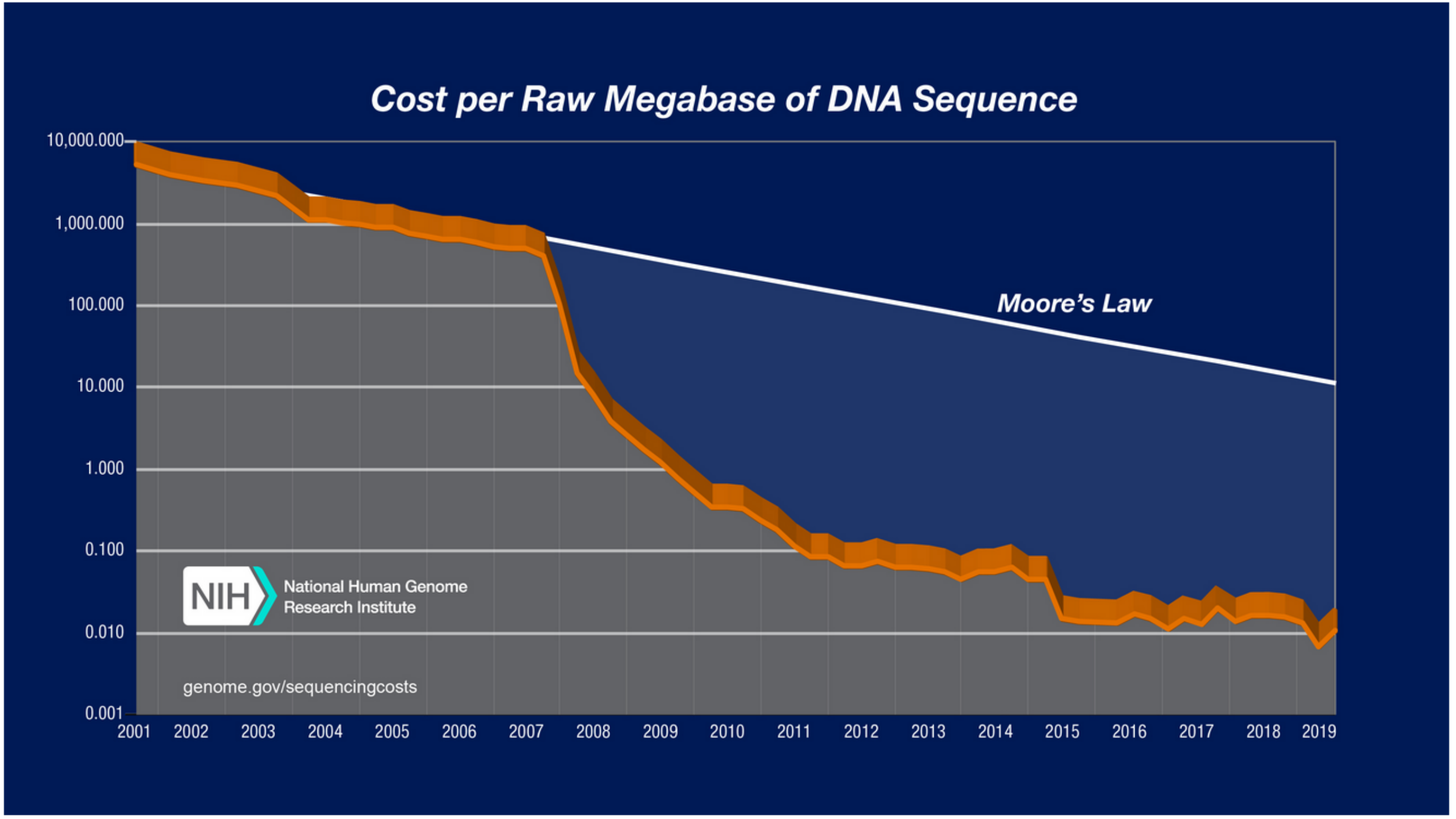
Each dot is RNA-seq data from a single cell (from human lungs)

For each cell, the gene expression for all ~20,000 genes is summarized in just **two dimensions**




ACE2
UMAP 2
pos
neg
UMAP 1
TMPRSS2
UMAP 2
pos
neg
UMAP 1

ChIP-seq (4 lectures)


Cost per Raw Megabase of DNA Sequence
10,000.000
1,000.000
100.000
10.000
1.000
0.100
0.010
0.001
Moore's Law
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

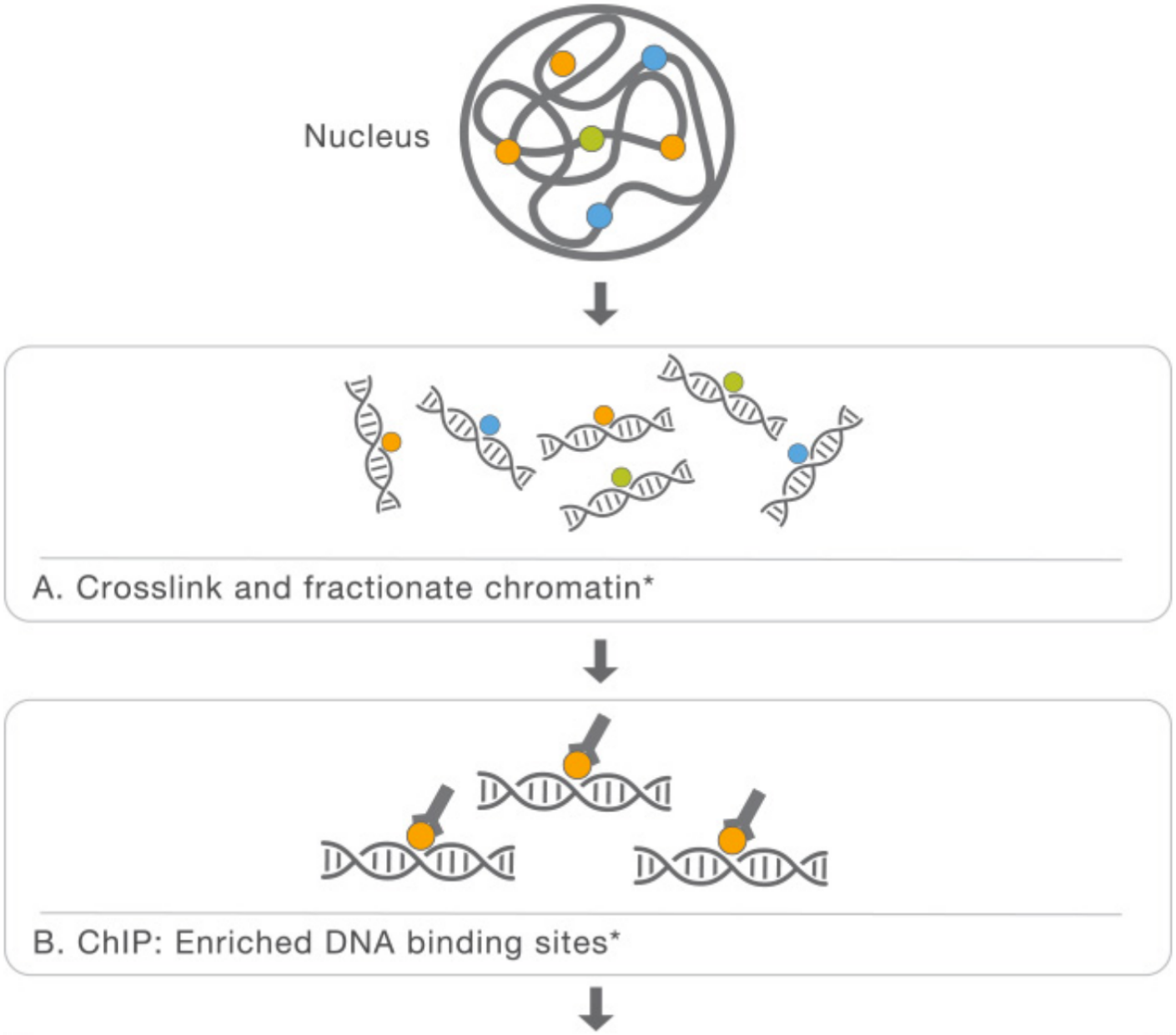
- We discussed 10 different experimental assays that take advantage of the price drop in DNA sequencing

- In most of these assays: something sticks to the DNA, something else cuts the DNA, then we just sequence some desired subset of the DNA

- The two assays we spent most of the next two lectures discussing were ChIP-seq and ATAC-seq

ChIP-seq

- Combine **ch**romatin **i**mmunoprecip**p**itation with massively parallel DNA **seq**uencing to analyze protein interactions with DNA



Sequence DNA of **just** these fragments

- The protein of interest interacts with the “pulled-down” DNA that we sequence

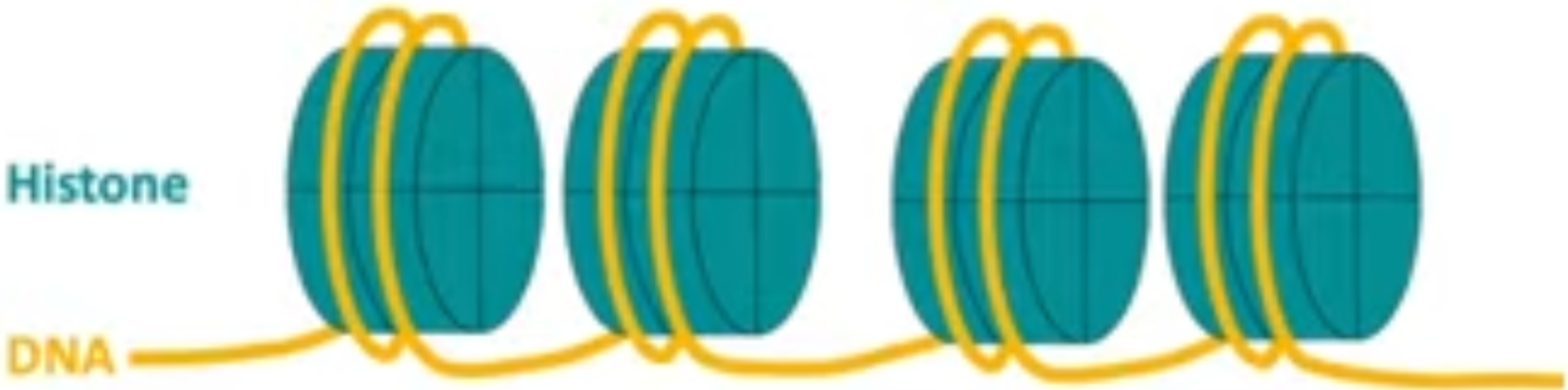
- For example, ChIP-seq is used to study transcription factors

ATAC-seq

- **A**ssay for **T**ransposase-**A**ccessible **C**hromatin

- Used to assess genome-wide chromatin accessibility and how it varies between different cell types

Chromatin Accessibility



inactive

active

- The pattern of which parts of the chromatin are accessible (also called active or “open”) and which parts are not accessible (also called inactive or “closed”) affects how the genome is regulated. Basically the genes in the open regions are expressed more than the genes in the closed regions

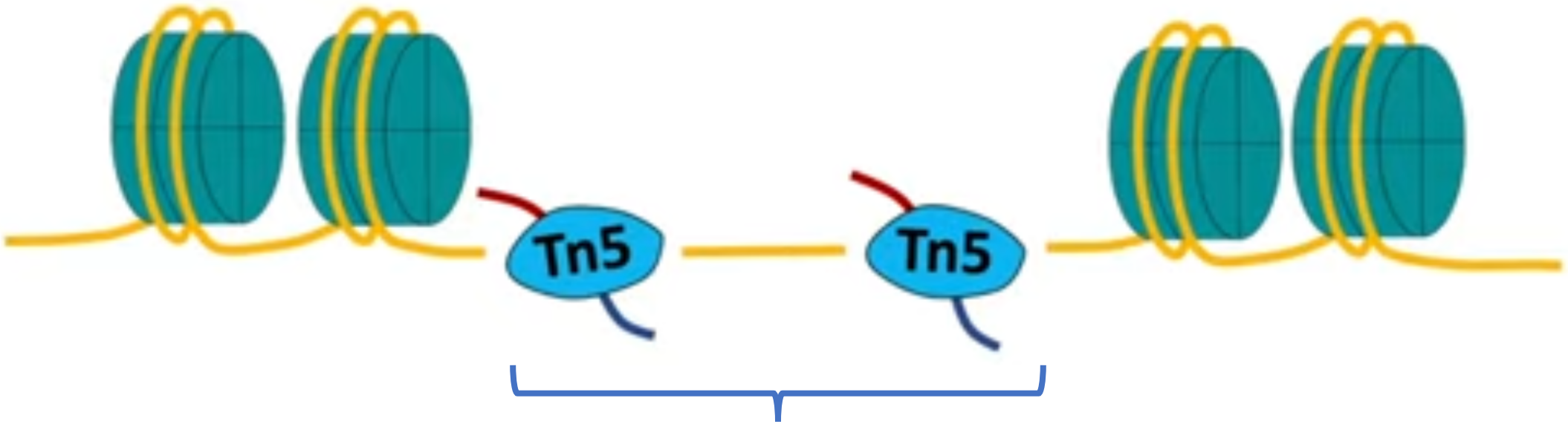
- This pattern is different in different cell types, and is believed to be the main reason why different cell types are different

- A transposase is an enzyme capable of binding to the end of a transposon and catalyzing its movement to another part of the genome

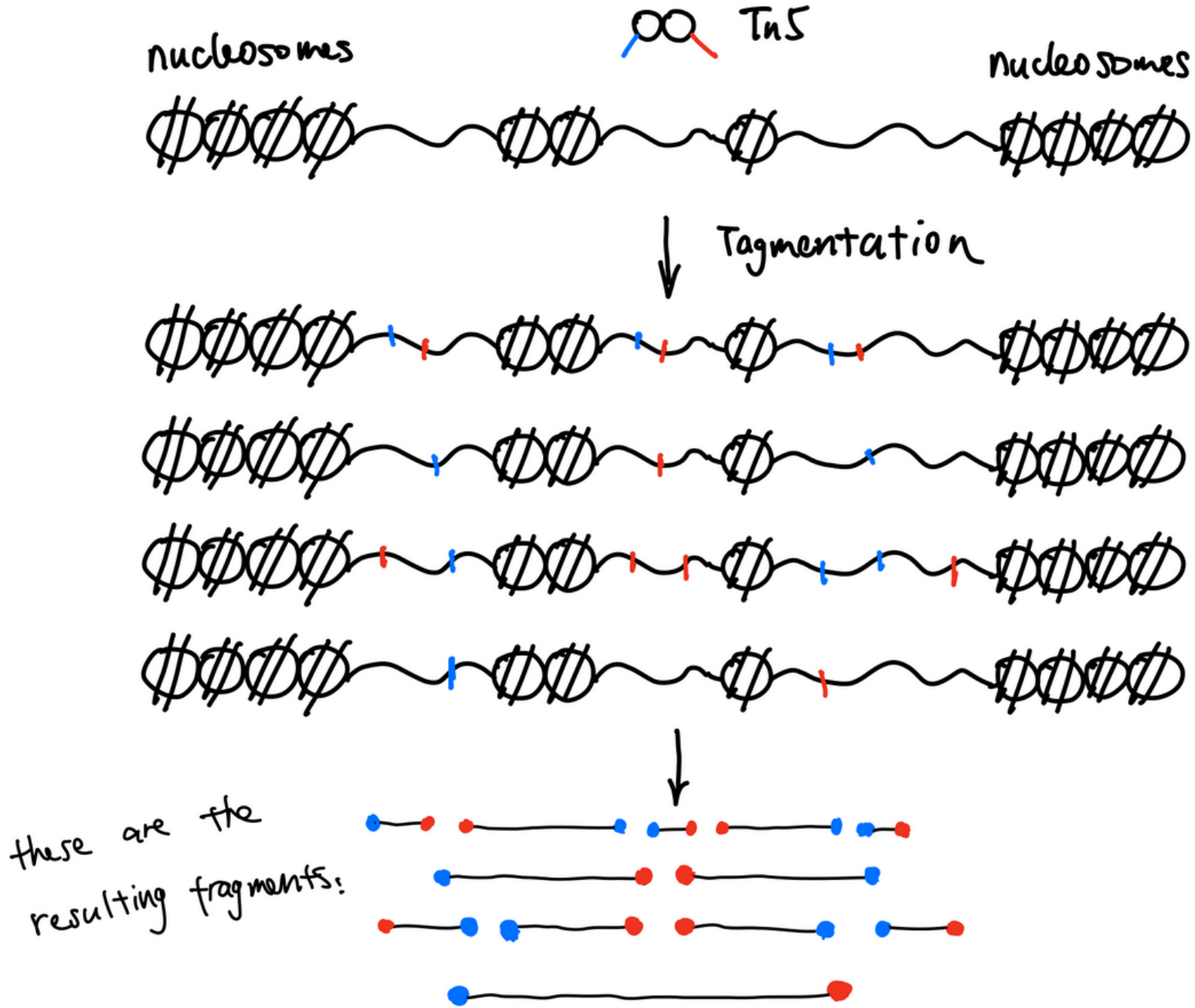
- The transposase Tn5 (like most other transposases) is notably inactive. But there is a mutant form that is hyperactive, and cleaves and tags double-stranded accessible DNA with sequencing adaptors in a single enzymatic step

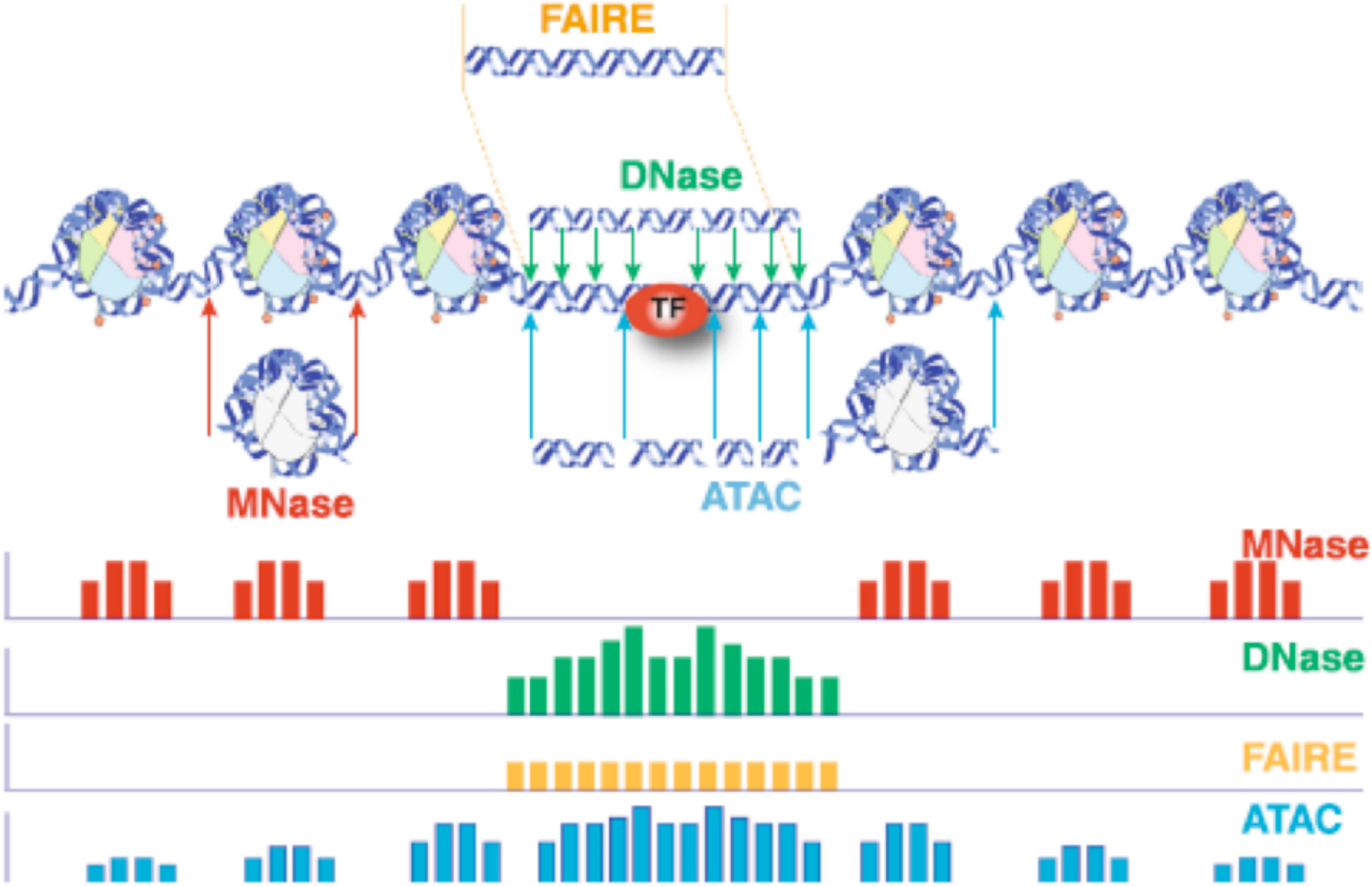
ATAC seq

Assay for Transposase-Accessible Chromatin



Just sequence the accessible part


nucleosomes
Tn5
nucleosomes
Tagmentation
these are the resulting fragments:


FAIRE
DNase
TF
MNase
ATAC
MNase
DNase
FAIRE
ATAC

- ATAC-seq sequences where the Tn5 transposase cuts, so it reveals **both** the large open regions and simultaneously the additional chromatin structures (nucleosome positioning and packing, transcription factor occupancy, etc.)

- ATAC-seq preparation can be completed in under 3 hours (FAIRE, DNase, and MNase each take over 24 hours)

- For both ChIP-seq and ATAC-seq we can do the assay either in bulk or on single cells

- One difference between RNA-seq and these methods, is that in RNA-seq we knew where the genes were before we did the experiment

- In ChIP-seq, we don’t know where a given protein will bind before we do the experiment. And in ATAC-seq, we don’t know where the regions of accessible chromatin are before we do the experiment

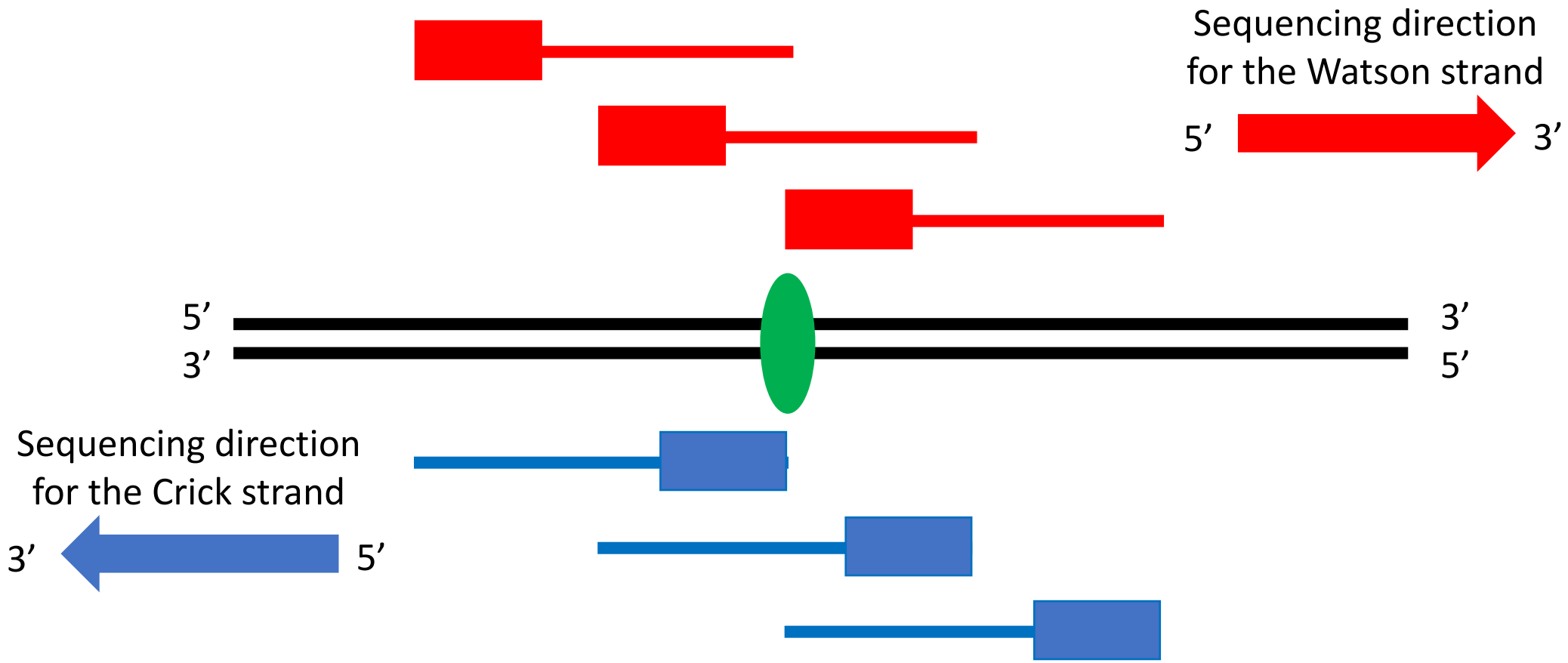
Peak Detection

- There are many algorithms for peak detection
- We discussed the MACS (Model-based Analysis of ChIP-seq) algorithm
- This algorithm can work with and without control samples. There are four main steps

Step 1. Removing redundancy

- If multiple reads start at the exact same genomic position, the default is to only count one (because it assumes the others are due to unequal PCR amplification)

Step 2. Modeling the shift size

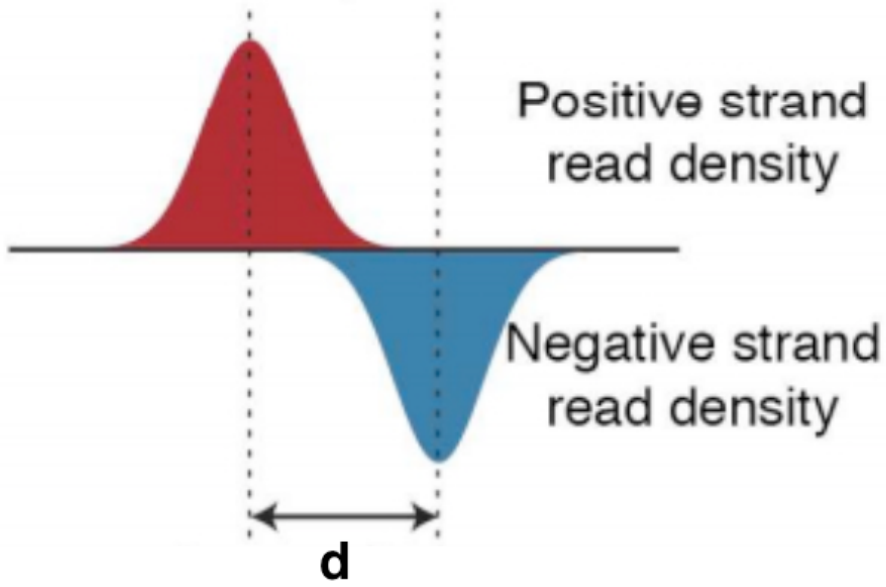


We don’t know the fragment size. The fragments overlap the DNA where the protein of interest (shown in green) is bound. For half of the fragments we sequence the Watson strand, and for half of the fragments we sequence the Crick strand. We only sequence part of the fragment. I have shown the case where the fragments are ~300 bp and we do single-end sequencing of 100 bp on the 5’ end (the sequenced part of the fragment is shown as a box).

The beginning (5’ end) of the Watson reads (shown in red) are to the “left” of the beginning (5’ end) of the Crick reads (shown in blue).

The distance between the modes of the the two peaks is the parameter “d”. d estimates the fragment size. We shift the reads d/2 towards the 3’ end. This combines the two peaks into one peak, and better represents the precise location of the protein-DNA interaction site.

The figure on the previous slide showed single-end sequencing but we similarly model the shift for paired-end sequencing.



Step 3. Peak detection

- For each peak, we compare the number of reads to the genomic background
- This genomic background is dynamic allowing for local variation. If we have controls we use them to estimate the background, if we don’t have controls we have another method to estimate the background
- We model the number of reads at a genomic position as Poisson distributed, and then perform the Poisson test to see if the number of reads is greater than expected based on the genomic background

Step 4. Estimating false discovery rate (FDR)

- There are many peaks so we must do a multiple tests correction
- If there are no controls, we do the Benjamini-Hochberg correction, exactly as we did in RNA-seq
- If there are controls, we swap the control and ChIP-seq samples and compute the empirical FDR

ATAC-seq

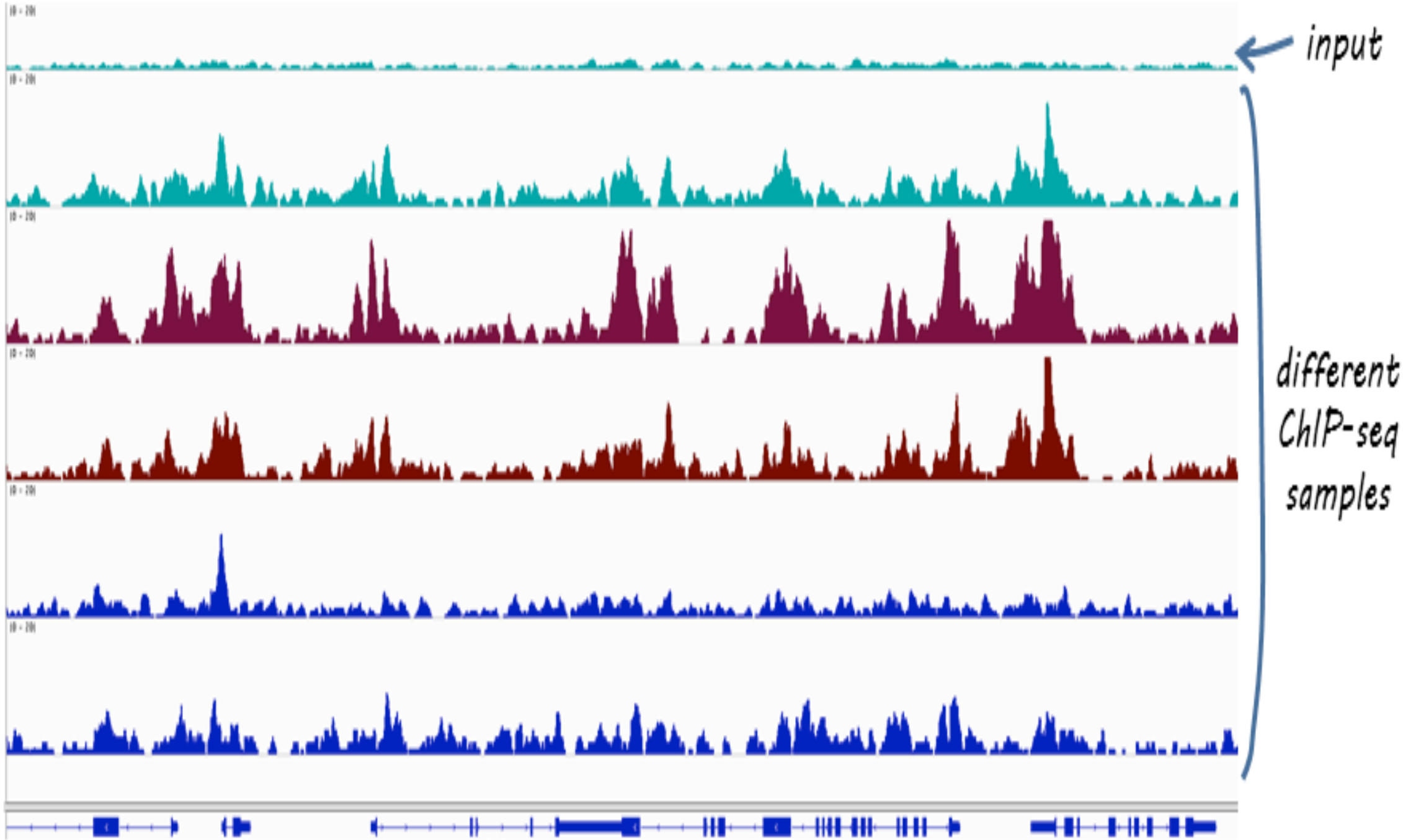
- The MACS algorithm was developed for ChIP-seq data, but it can be used with slight modifications for ATAC-seq data

- Modification #1 There are usually no controls in ATAC-seq experiments

- Modification #2 We do not have to model the shift size

Differential Peak Analysis

- There are also many algorithms for differential peak analysis
- We discussed the DiffBind algorithm


input
different ChIP-seq samples

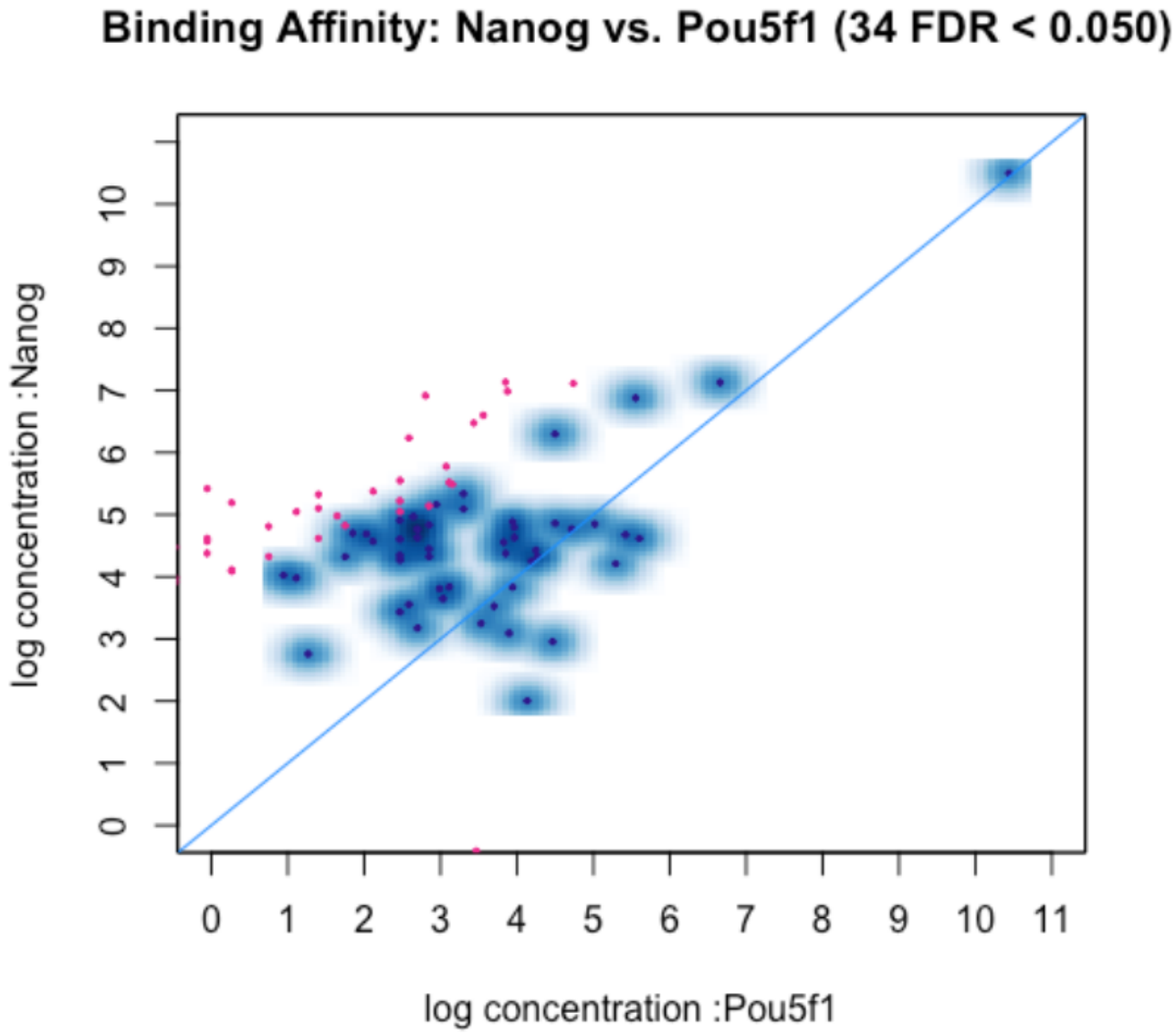
DiffBind

1. DiffBind starts with a set of peaksets that were determined for each ChIP-seq sample separately (by MACS or another algorithm)

2. DiffBind then merges overlapping peaks to derive a consensus peakset. This consensus set is all of the candidate binding sites considered for further analysis

3. DiffBind then runs EdgeR and DiffSeq. The data is the number of reads at each of the consensus peak sites

Each dot is a peak. The red dots are those that DESeq found to be significantly preferentially enriched. We see that all of the red dots are enriched for Nanog



```
> res_deseq

GRanges object with 85 ranges and 6 metadata columns:
     seqnames            ranges strand |      Conc Conc_Nanog Conc_Pou5f1      Fold
        <Rle>         <IRanges>  <Rle> | <numeric>  <numeric>   <numeric> <numeric>
  64    chr12 25644533-25644920      * |      4.45       5.42       -0.06      5.48
  28    chr12 12162157-12162655      * |         6       6.92         2.8      4.12
  35    chr12 13898546-13898828      * |      4.24       5.19        0.26      4.93
  46    chr12 16266148-16266518      * |      4.42       5.32         1.4      3.92
  36    chr12 13902296-13902867      * |      5.35       6.24        2.59      3.65
     p-value       FDR
     <numeric> <numeric>
  64  7.73e-06  0.000657
  28  3.14e-05   0.00133
  35  0.000129   0.00366
  46  0.000281   0.00597
  36  0.000594   0.00894
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

- We can also use DiffBind for ATAC-seq
- There are many other algorithms for differential peak analysis for ATAC-seq, most of these algorithms were originally developed for ChIP-seq
- We also discussed using artificial intelligence

Precision or Personalized Medicine (1 lecture)

- Rather than a “one-size-fits-all” approach, subgroups of patients can be defined, often by genomics, and targeted in more specific ways
- Three examples: cystic fibrosis, precision oncology, and pharmacogenomics
- The challenges of getting such large data sets
- UK Biobank

Questions?