



BGGN 213

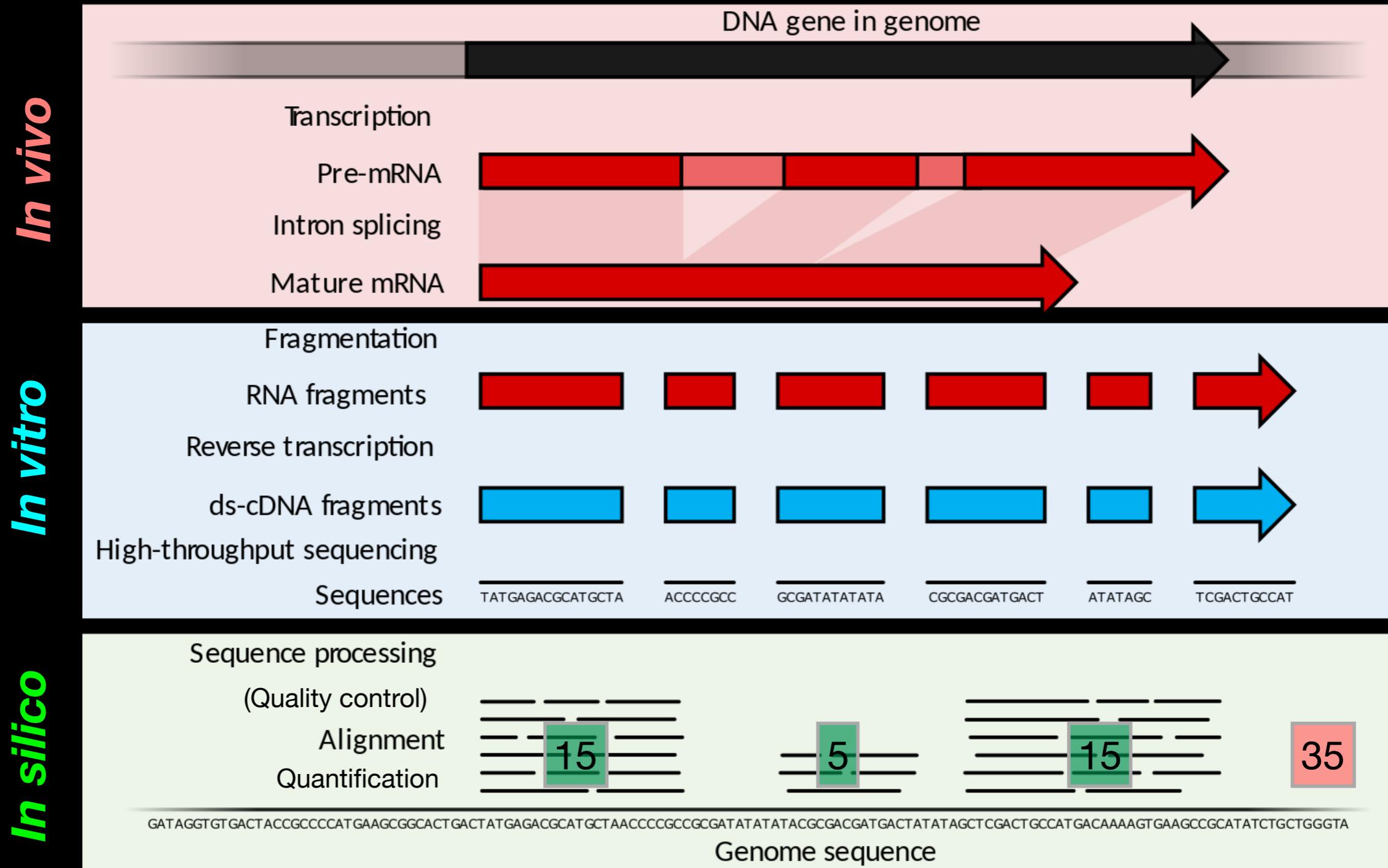
Genome Informatics II

Lecture 14

Barry Grant
UC San Diego

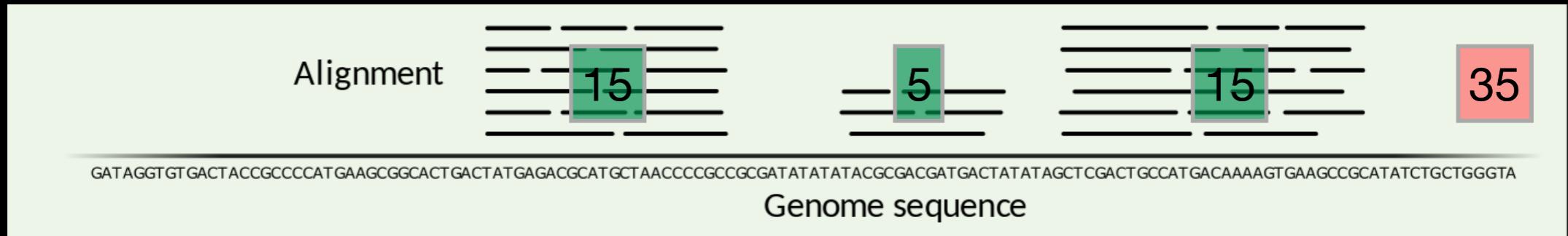
<http://thegrantlab.org/bggn213>

RNA sequencing overview



Goal: RNA quantification, transcript discovery, variant identification

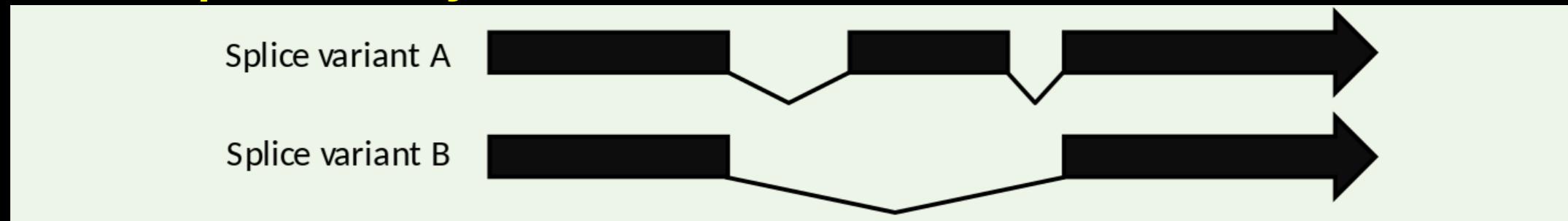
Mapping/Alignment



Quantification

Absolute read counts	15	5	15	(35)
Normalized read counts	$RPKM = \frac{totalTranscriptReads}{mappedReads(millions) \times transcriptLength(Kb)}$			(0.7)

Transcript discovery



Variant discovery



Discovering Variation: SNPs and INDELs

SNP (C/T)

ATCCTGATTGGTGAACGTTATCGACGATCCGATCGA
ATCCTGATTGGTGAACGTTATCGACGATCCGATCGA
CGGTGAACGTTATCGACGATTGATCGAACTGTCAGC
GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
TGAACGTTATCGACGTTCCGATCGAACTGTCATCGGC
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT

sequencing error or genetic variant?

ATCCTGATTGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATGCTAGTG

reference genome

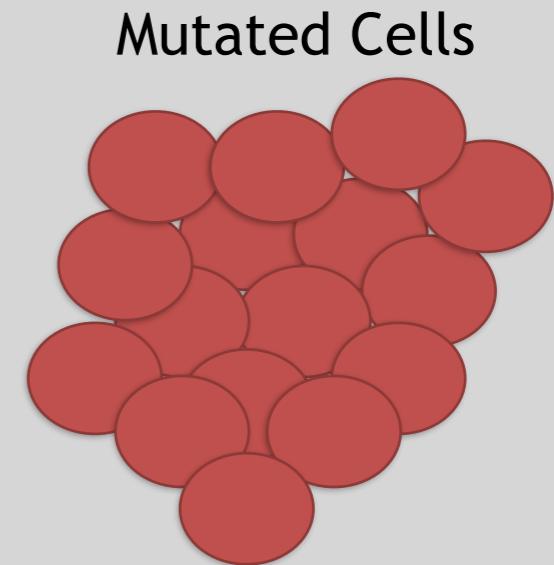
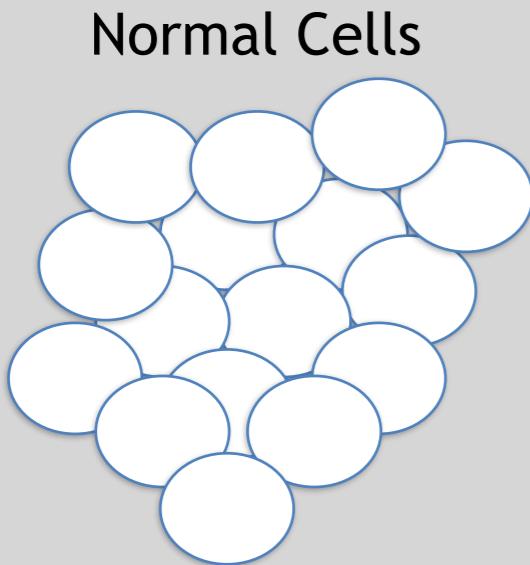
TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
TCGACGATCCGATCGAACTGTCAGCGGCAAGCTGAT
ATCCGATCGAACTGTCAGCGGCAAGCTGATCG CGAT
TCCGAGCGAACTGTCAGCGGCAAGCTGATCG CGATC
TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGA
GATCGAACTGTCAGCGGCAAGCTGATCG CGATCGA
AACTGTCAGCGGCAAGCTGATCG CGATCGATGCTA
TGTCAGCGGCAAGCTGATCGATCGATCGATGCTAG
TCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG

sequencing error or genetic variant?

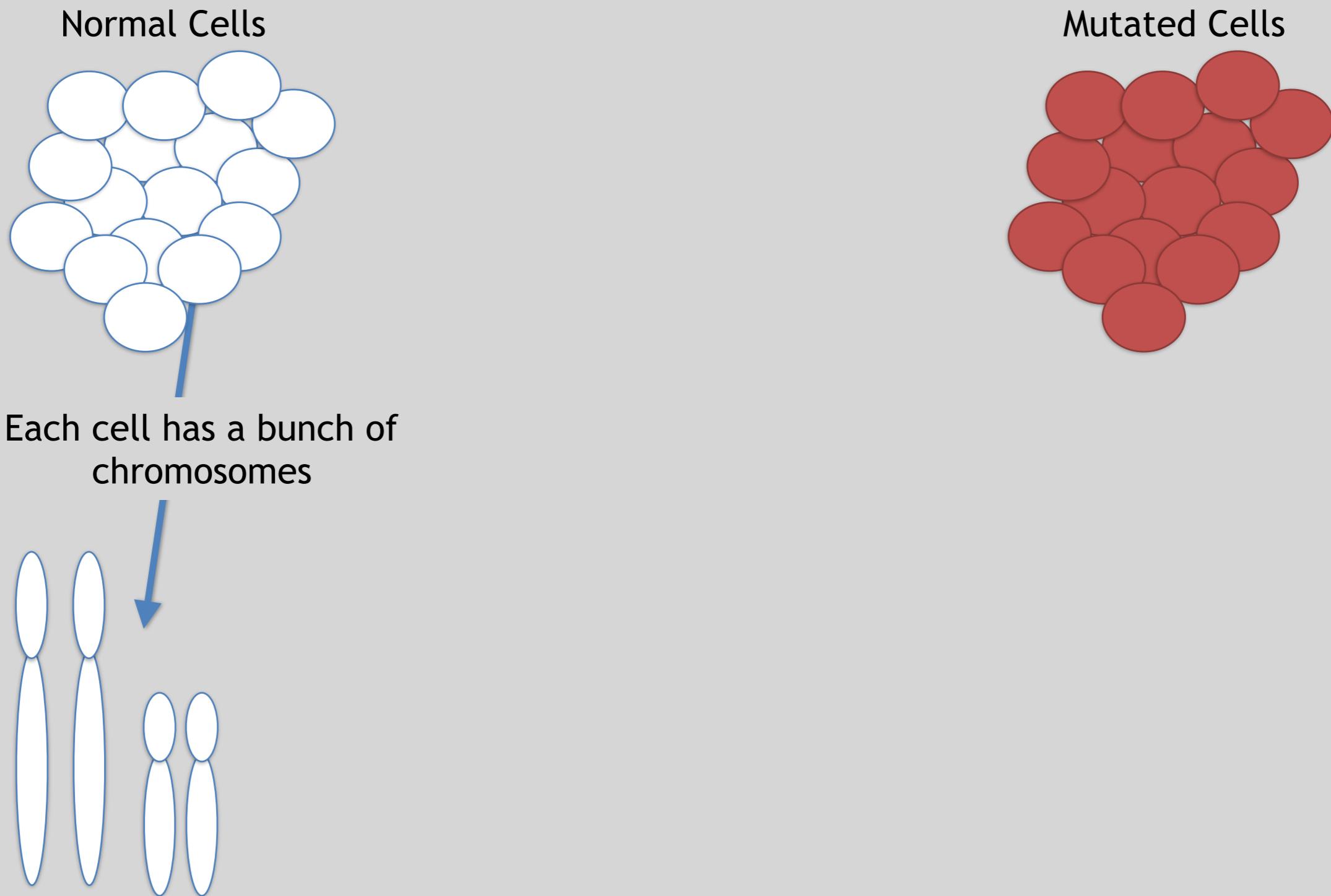
INDEL

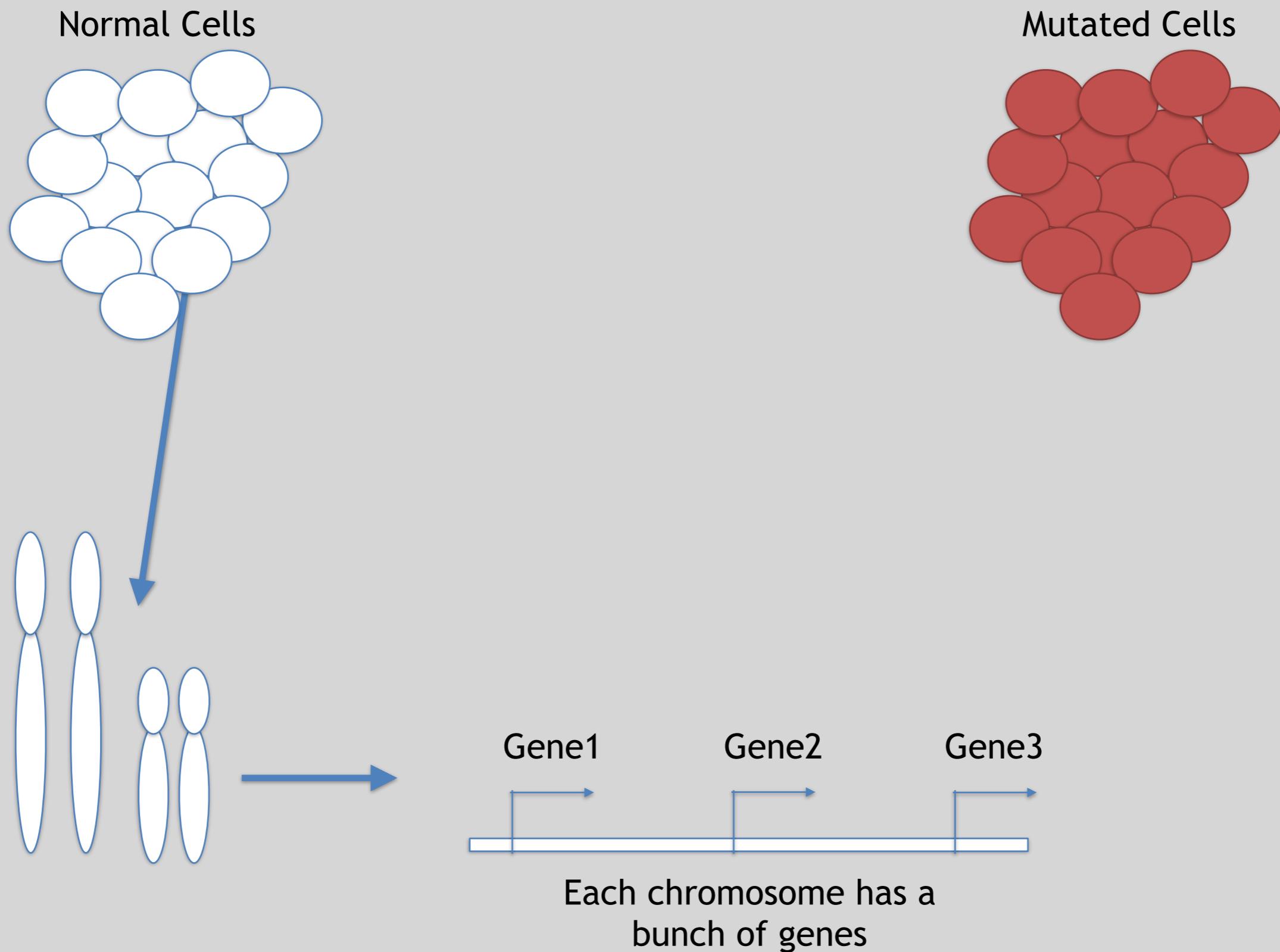
RNA Sequencing

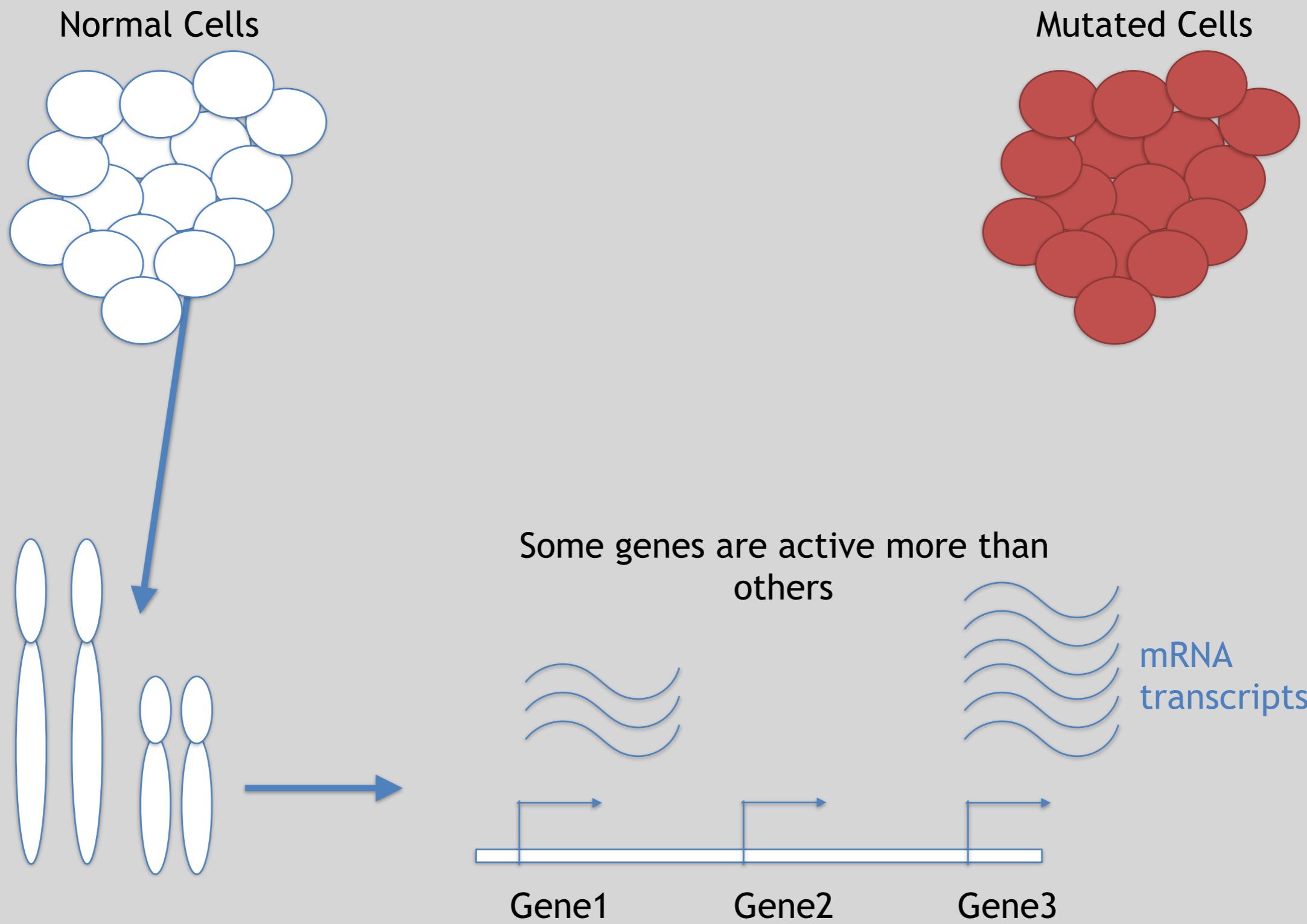
The absolute basics

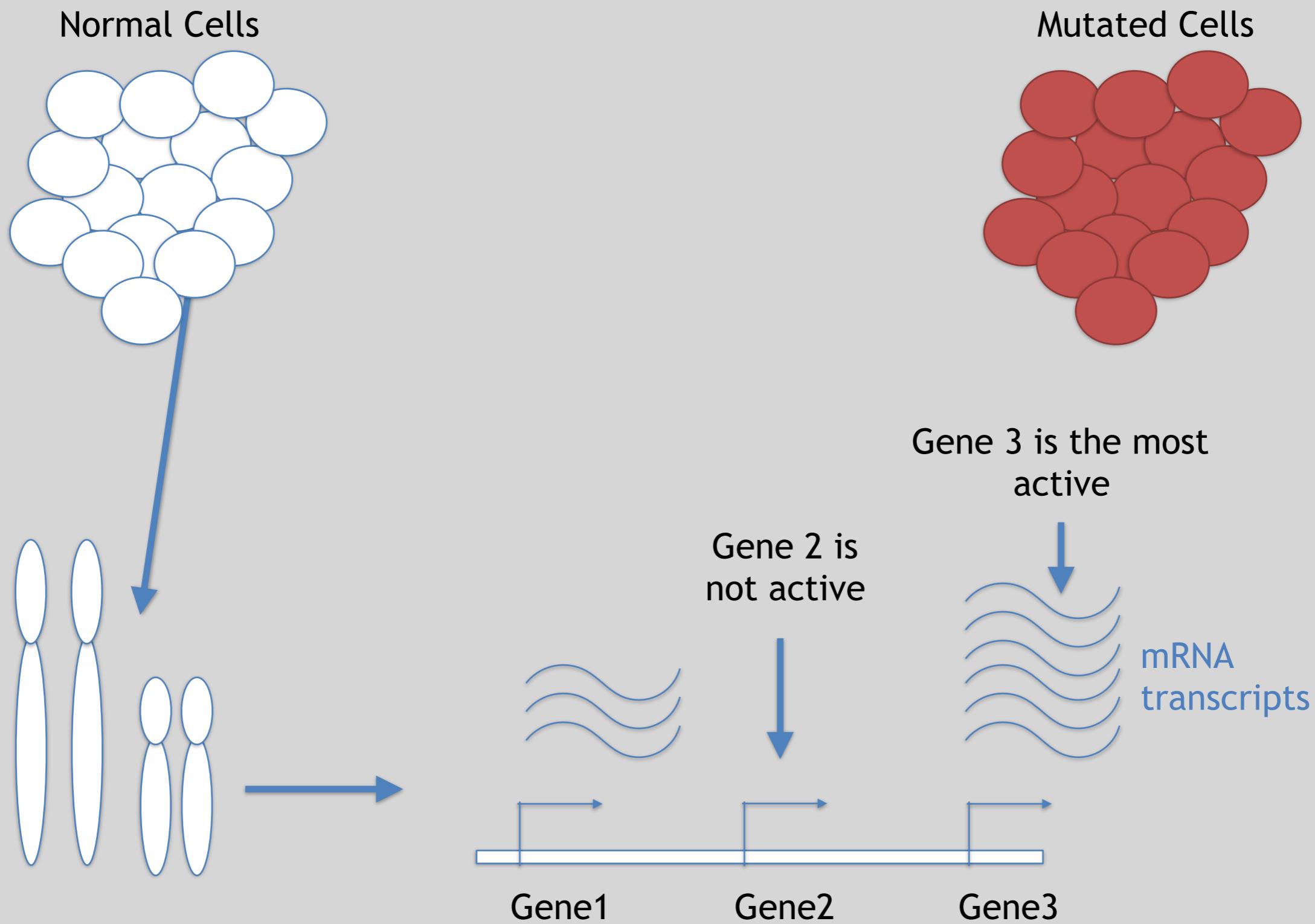


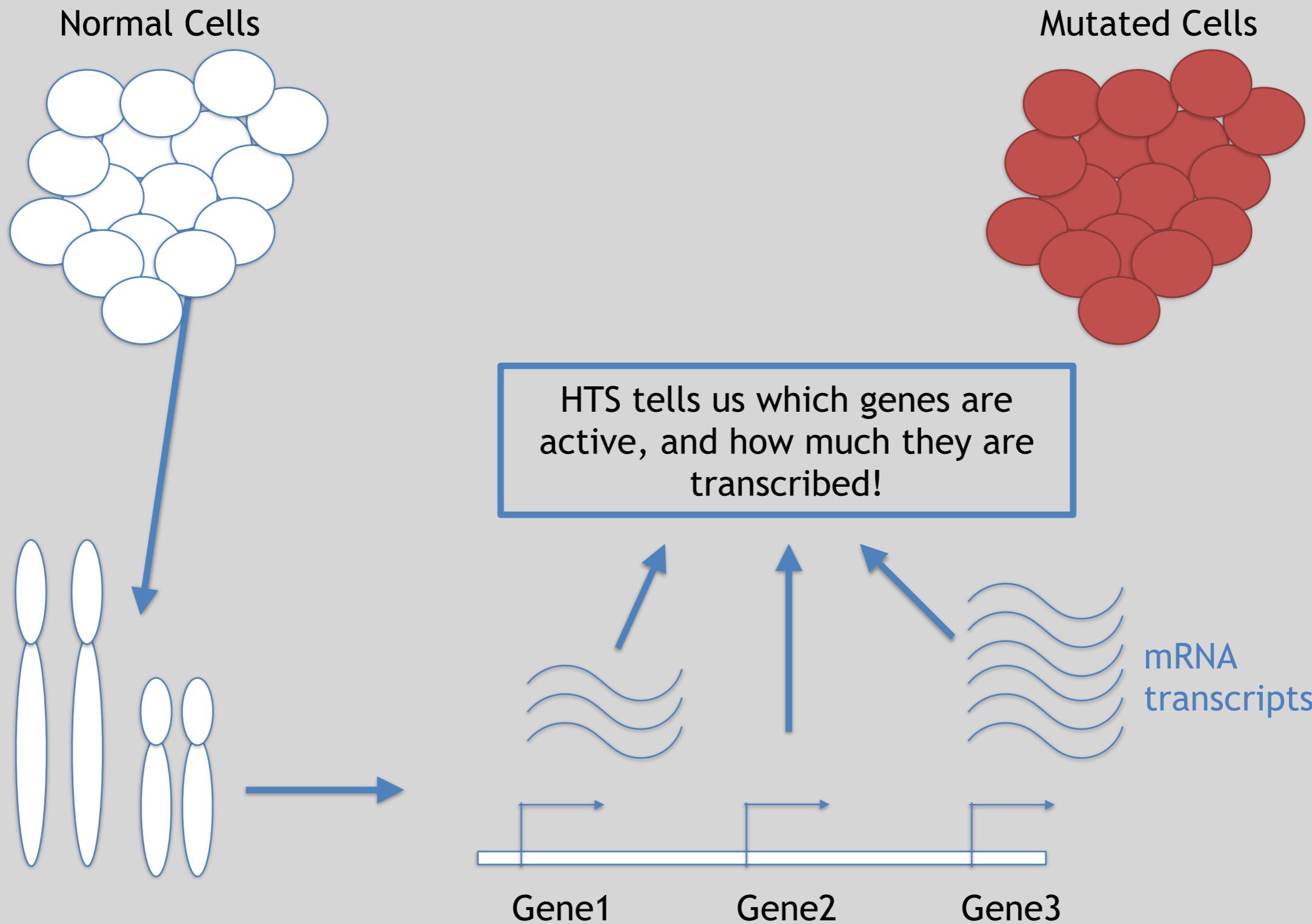
- The **mutated cells** behave differently than the **normal cells**
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing...

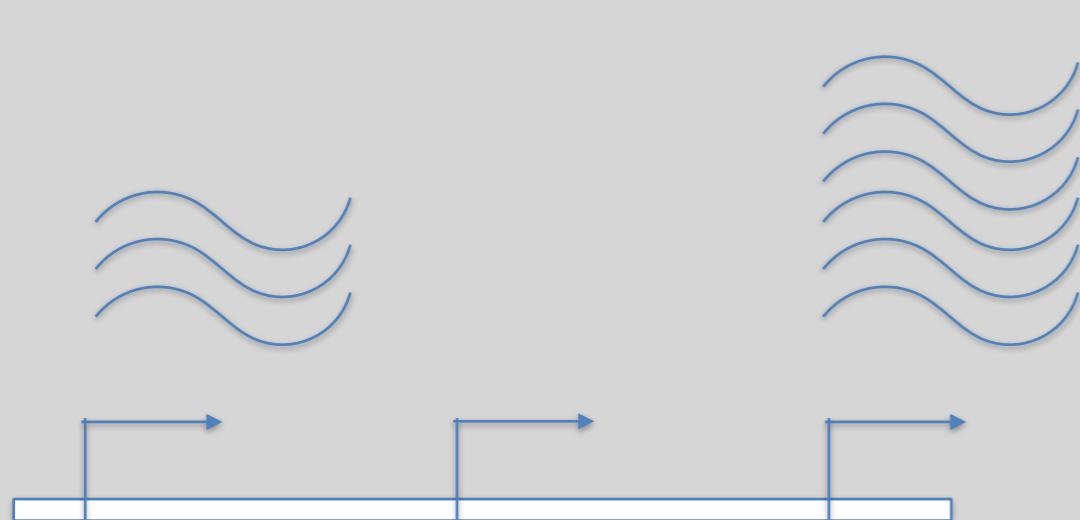
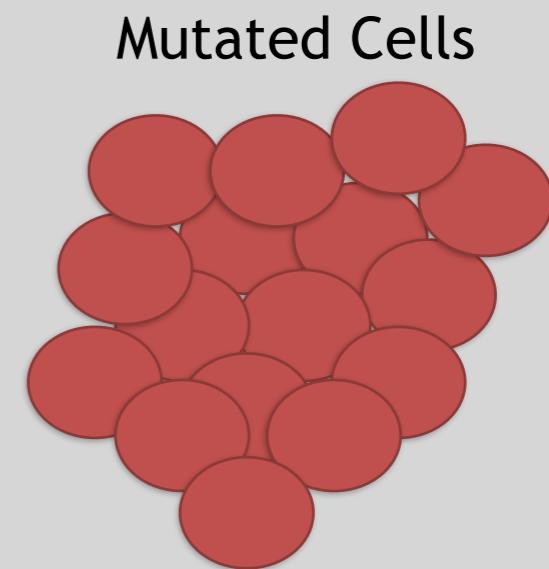
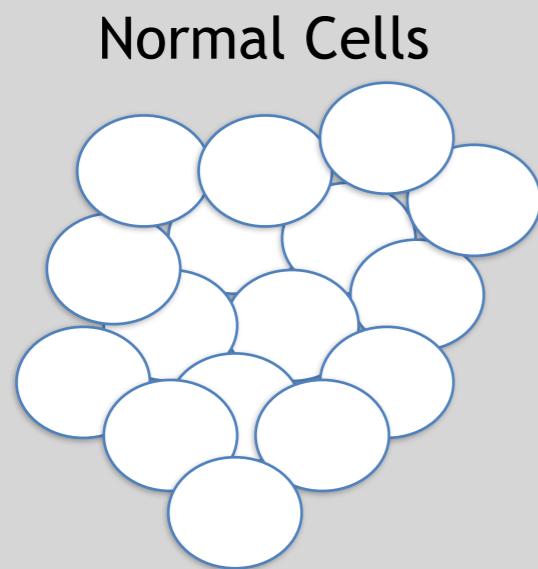




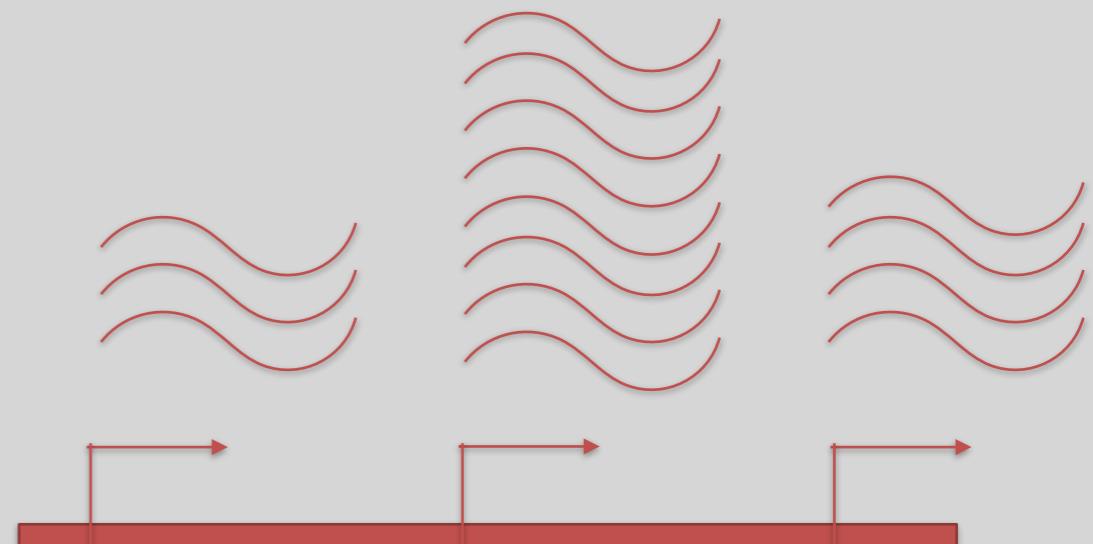




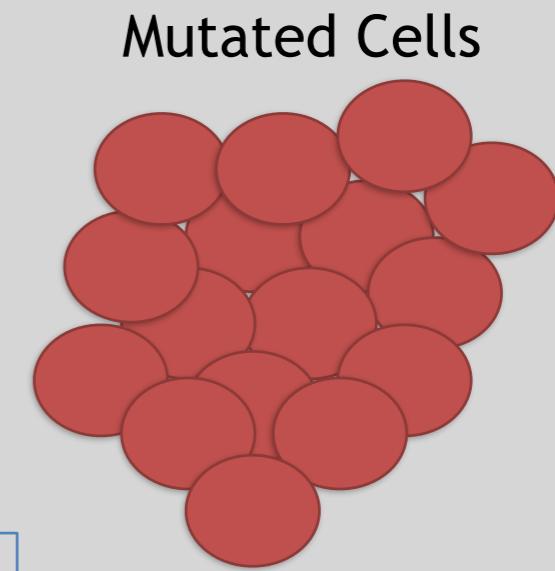
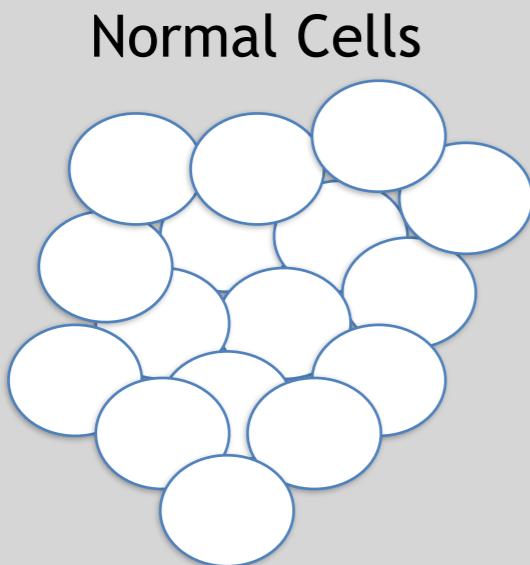




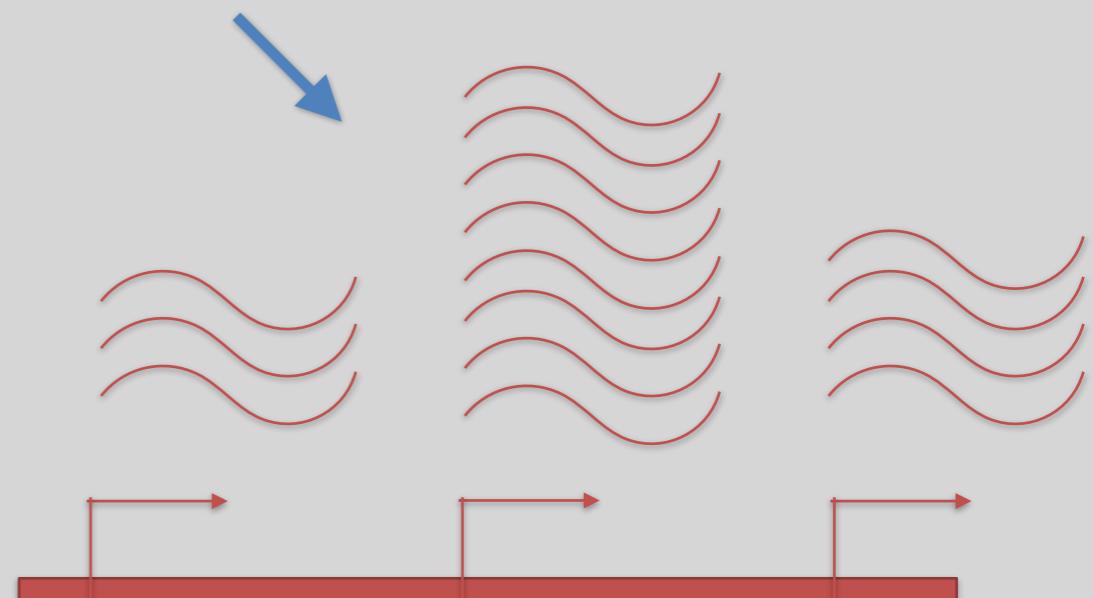
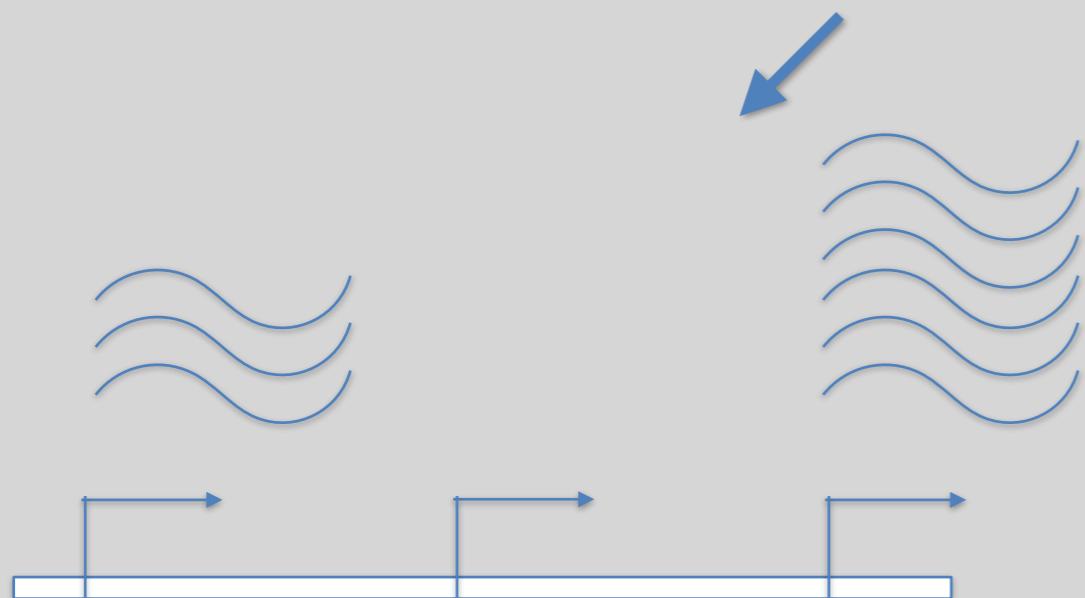
We use RNA-Seq to measure gene expression in normal cells ...

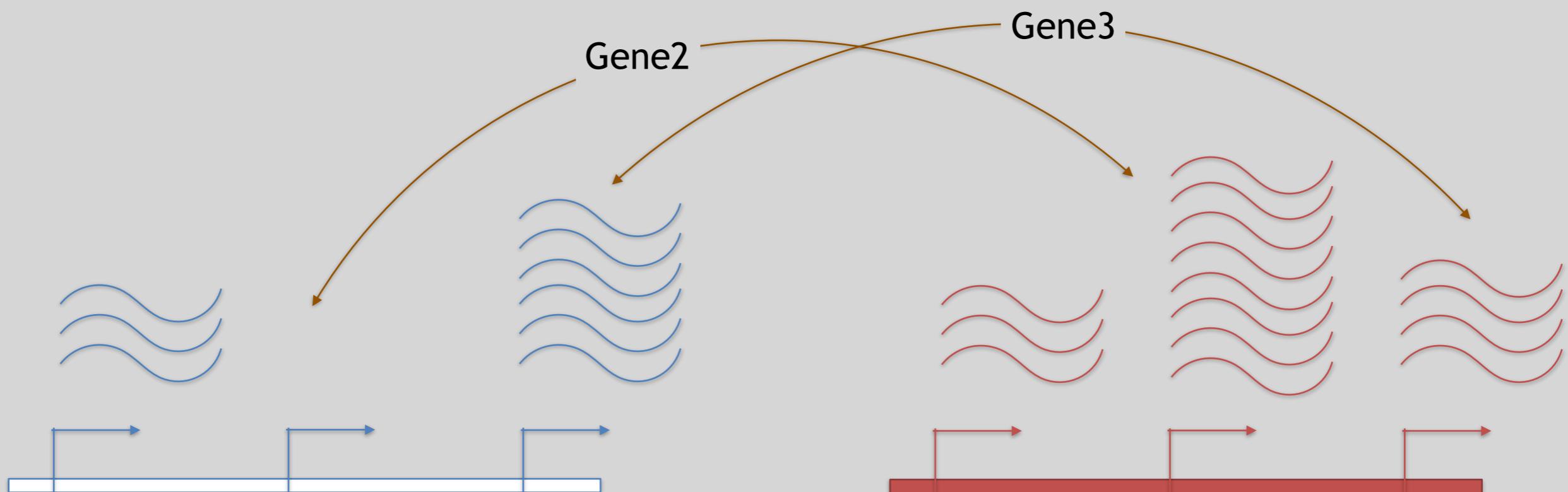
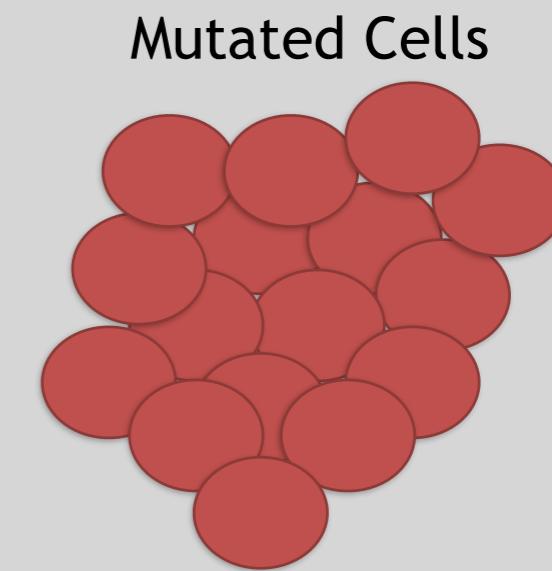
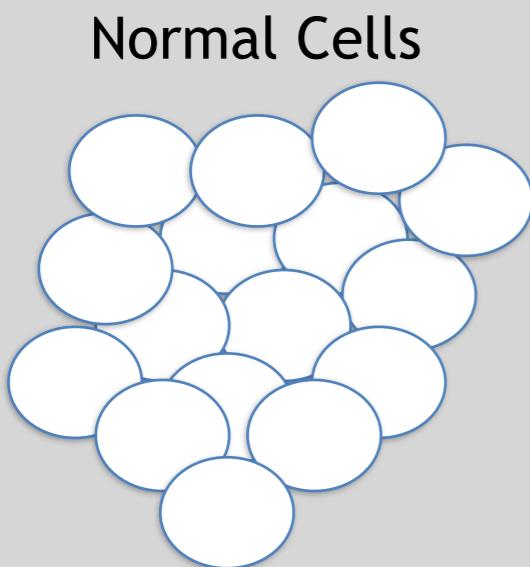


... then use it to measure gene expression in mutated cells



Then we can compare the two cell types to figure out what is different in the mutated cells!





Differences apparent for Gene 2 and
to a lesser extent Gene 3

3 Main Steps for RNA-Seq:

1) Prepare a sequencing library

(RNA to cDNA conversion via reverse transcription)

2) Sequence

(Using the same technologies as DNA sequencing)

3) Data analysis

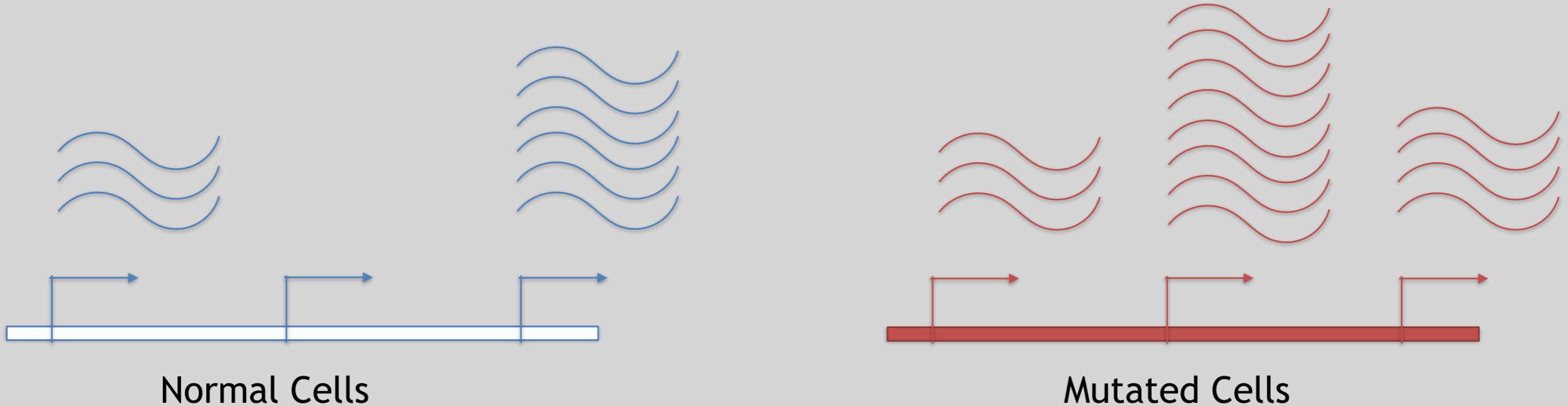
(Often the major bottleneck to overall success!)

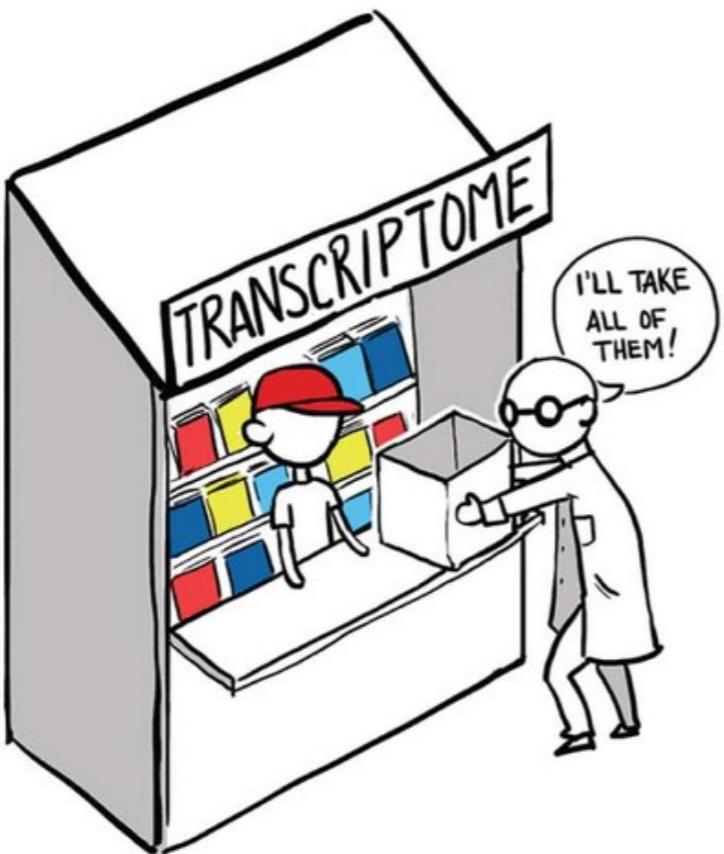
We will discuss each of these steps - but we will focus on step 3 today!

Our last class got us to the start of step 3!

Gene	WT-1	WT-2	WT-3	...
A1BG	30	5	13	...
AS1	24	10	18	...
...

We sequenced, aligned, counted the reads per gene in each sample to arrive at our count table/matrix





Inputs

Control

Reads R1

FastQ

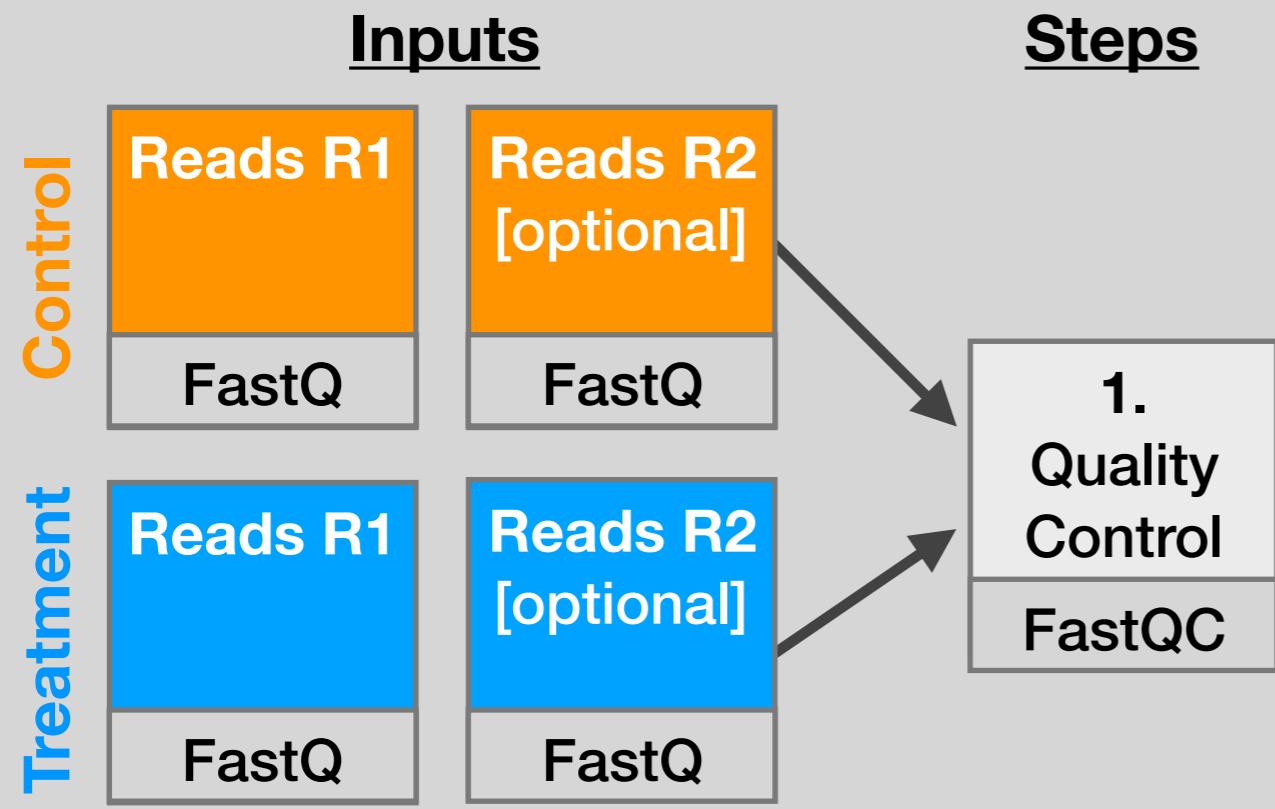
Treatment

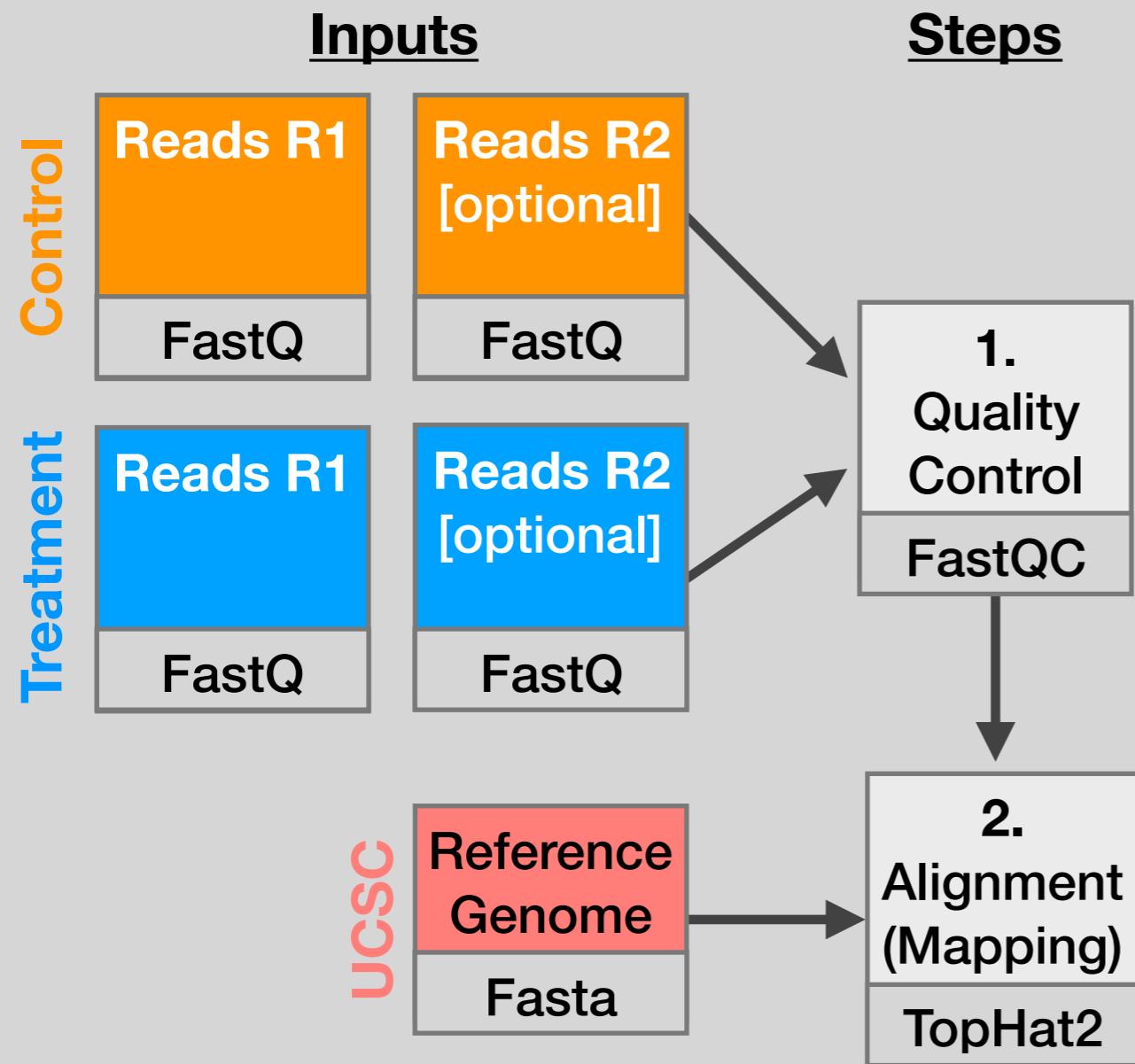
Reads R1

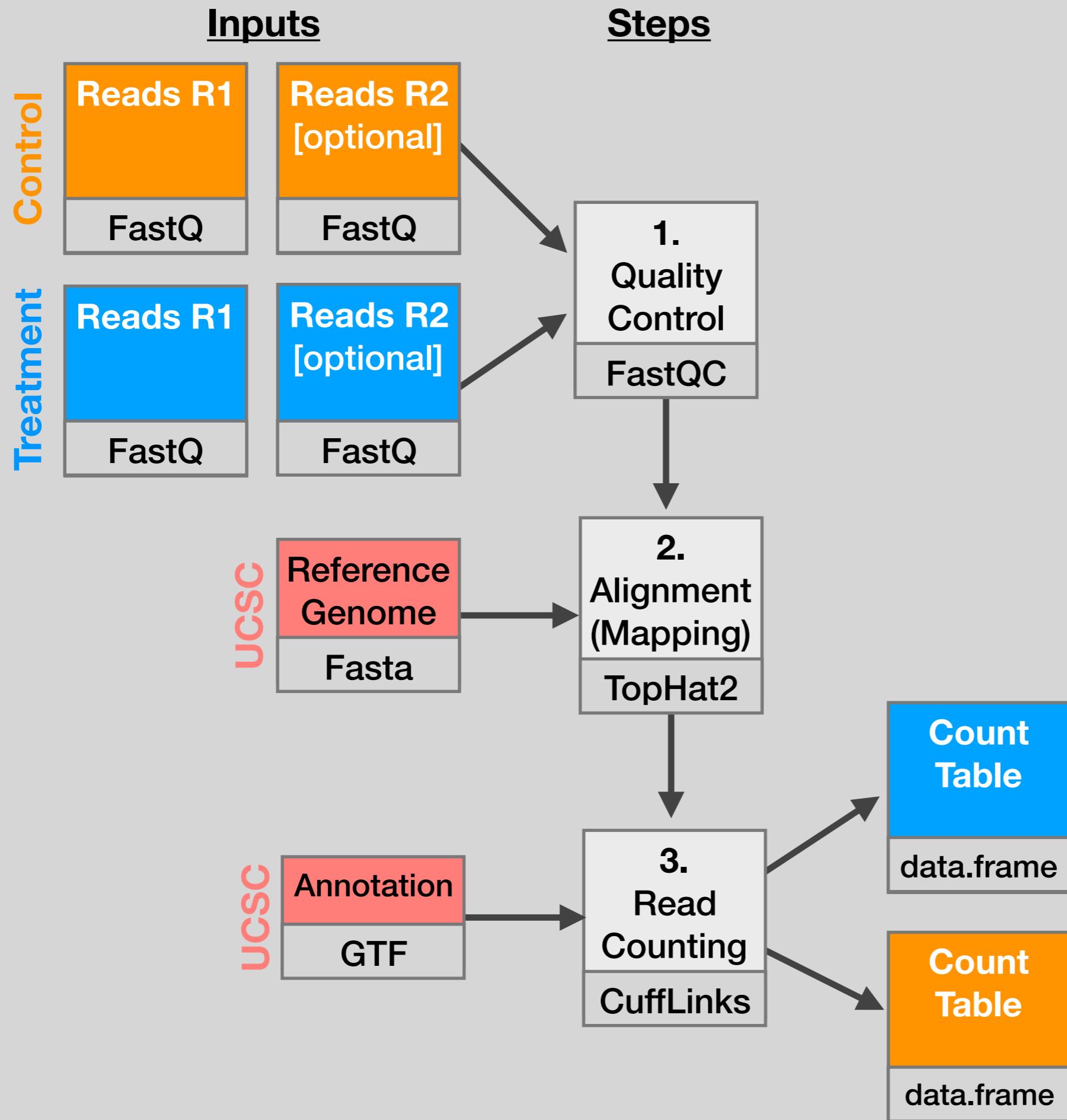
FastQ

Inputs



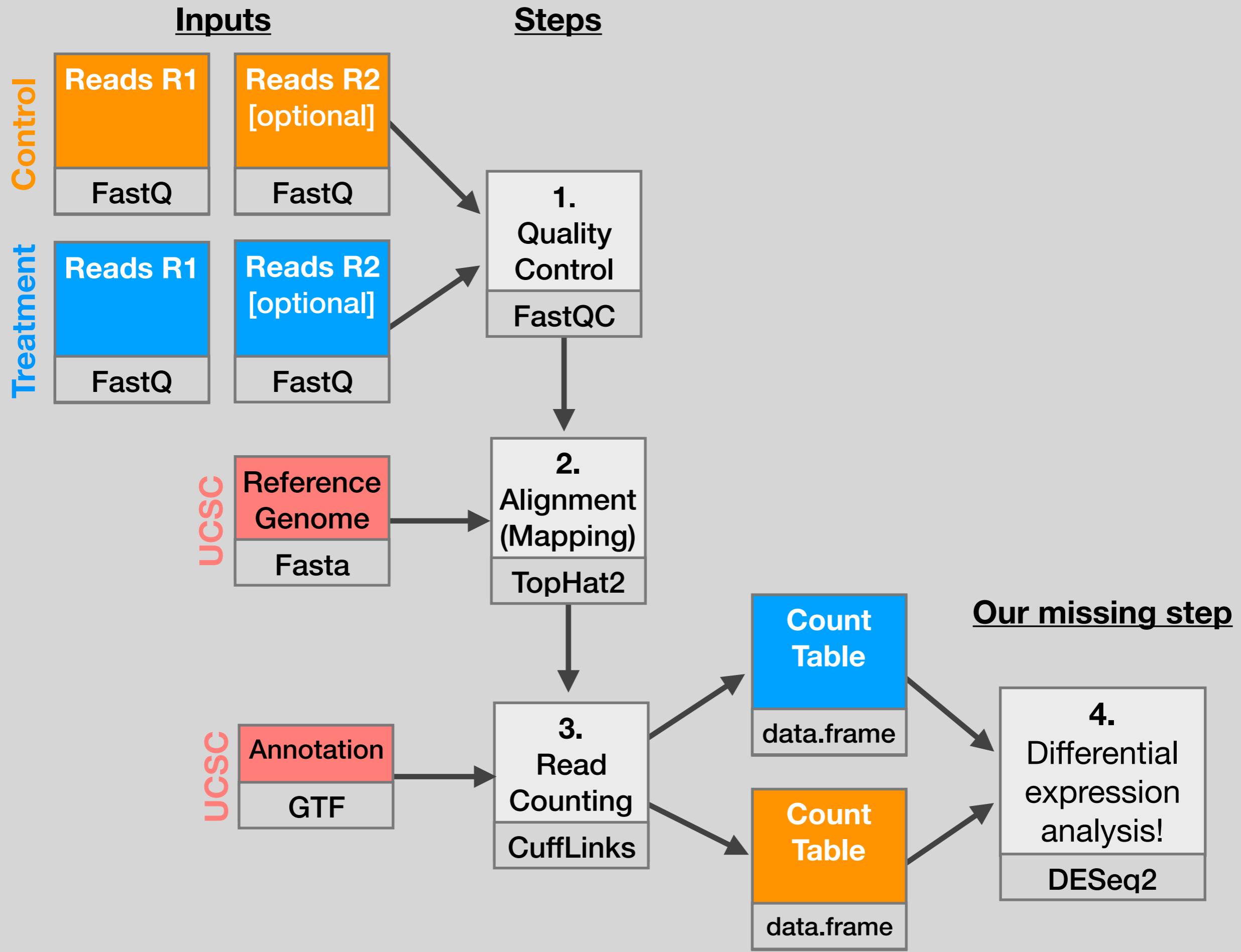






...Now what?

This is where we stoped last day



Install DESeq2

[Bioconductor Setup Link](#)

```
install.packages("BiocManager")
BiocManager::install()

# For this class, you'll also need DESeq2:
BiocManager::install("DESeq2")
```

Note: Answer **NO** to prompts to install from source or update...

```
Old packages: 'devtools', 'dplyr', 'DT', 'ggplot2', 'ggpubr',
'lattice', 'MASS', 'Matrix', 'mclust', 'mgcv', 'openssl',
'packrat', 'pkgload', 'ps', 'psych', 'raster', 'rcmdcheck',
'Rcpp', 'remotes', 'rsconnect', 'sessioninfo', 'shiny',
'shinystyles', 'survival', 'tidyverse', 'tinytex', 'xfun'
```

Update all/some/none? [a/s/**n**]:

n

Install DESeq2

[Bioconductor Setup Link](#)

```
install.packages("BiocManager")
BiocManager::install()
```

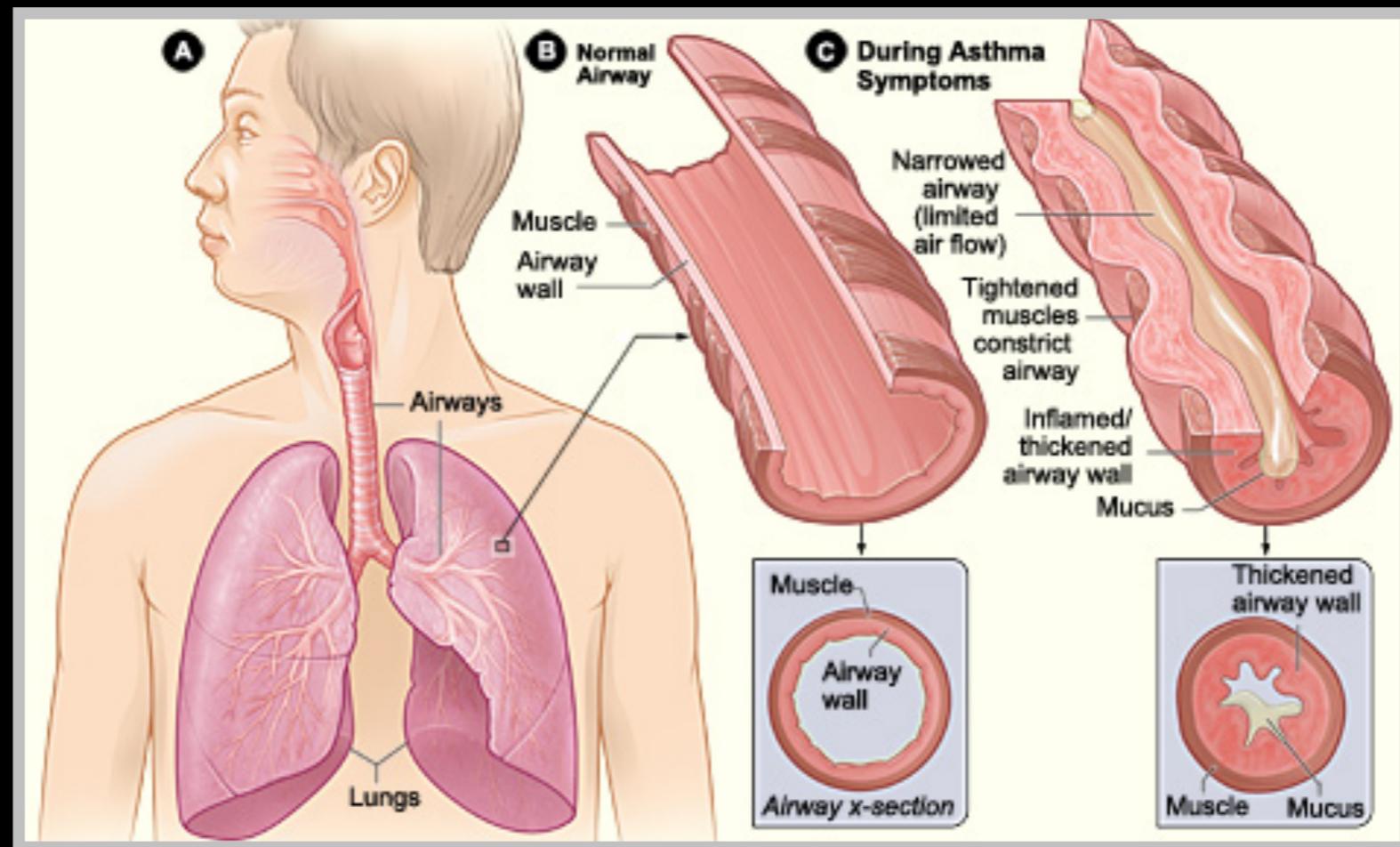
For this class, you'll also need DESeq2:

```
BiocManager::install("DESeq2")
```

Note: Answer **NO** to prompts to install from source or update...

Background to Today's Data

Glucocorticoids inhibit inflammatory processes and are often used to treat **asthma** because of their anti-inflammatory effects on airway smooth muscle (ASM) cells.



- Data from: Himes *et al.* "[RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells](#)." PLoS ONE. 2014 Jun 13;9(6):e99625.

Background to Today's Data

- The anti-inflammatory effects of glucocorticoids on airway smooth muscle (ASM) cells has been known for some time but the underlying molecular mechanisms are unclear.
- Himes *et al.* used **RNA-seq** to profile gene expression changes in 4 ASM cell lines treated with **dexamethasone** (a common synthetic glucocorticoid).
- Used Tophat and Cufflinks and found many differentially expressed genes. Focus on CRISPLD2 that encodes a secreted protein involved in lung development
- SNPs in CRISPLD2 in previous GWAS associated with inhaled corticosteroid resistance and bronchodilator response in asthma patients.
- Confirmed the upregulated CRISPLD2 with qPCR and increased protein expression with Western blotting.

Data pre-processing

- Analyzing RNA-seq data starts with sequencing reads.
- Many different approaches, see references on class website.
- Our workflow (previously done):
 - Reads downloaded from GEO ([GSE:GSE52778](#))
 - Quantify transcript abundance ([kallisto](#)).
 - Summarize to gene-level abundance ([txImport](#))
- Our starting point is a **count matrix**: each cell indicates the number of reads originating from a particular **gene** (in rows) for each **sample** (in columns).

counts + metadata

1

countData

gene	ctrl_1	ctrl_2	exp_1	exp_2
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
...

2

colData

id	treatment	sex	...
ctrl_1	control	male	...
ctrl_2	control	female	...
exp_1	treated	male	...
exp_2	treated	female	...

colData describes metadata about
the *columns* of countData

countData is the count matrix
(Number of reads coming from each
gene for each sample)

N.B. First column of colData must match column names (i.e. sample names) of countData (-1st)

Hands-on time!

https://bioboot.github.io/bggn213_F19/lectures/#14

Step 1 in any analysis is always the same:

Step 1 in any analysis is always the same:

PLOT THE DATA!!

Step 1 in any analysis is always the same:

PLOT THE DATA!!

- If there were only two genes, then plotting the data would be easy

Gene	WT-1	WT-2	WT-3
A1BG	30	5	13
AS1	24	10	18

Step 1 in any analysis is always the same:

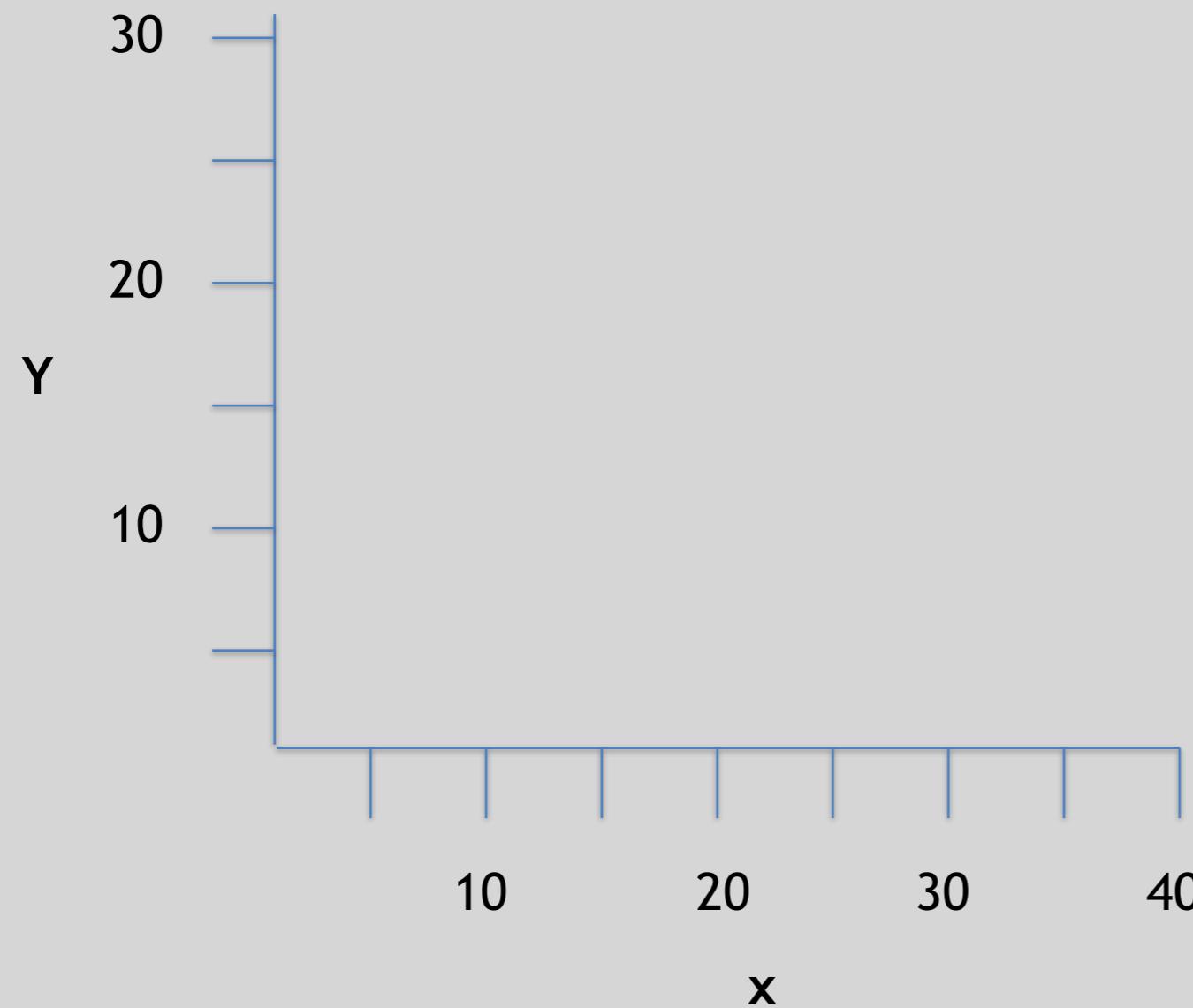
PLOT THE DATA!!

- If there were only two genes, then plotting the data would be easy

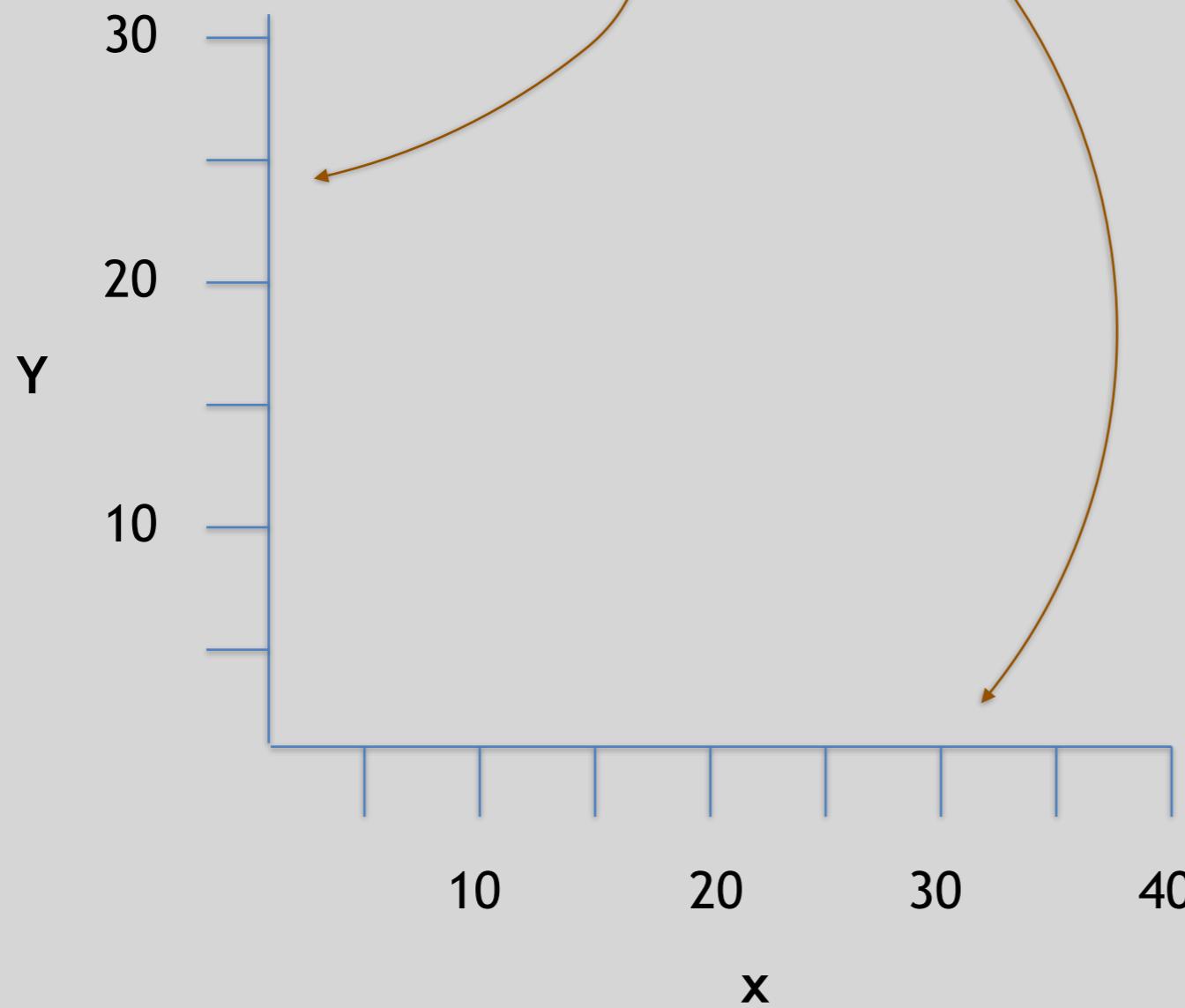
Gene	WT-1	WT-2	WT-3
x	30	5	13
y	24	10	18

Just replace the gene names with “x” and “y” and plot!

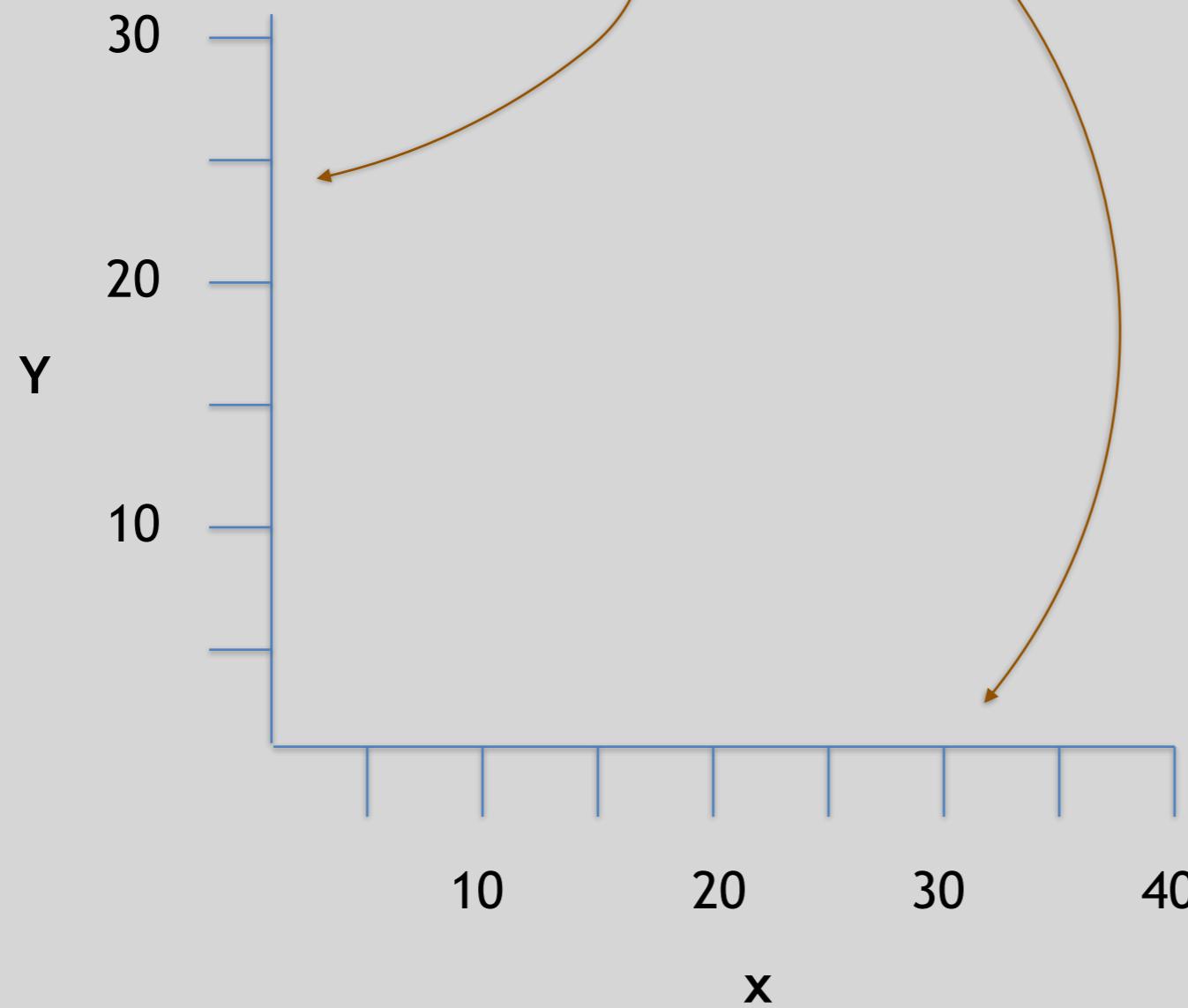
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



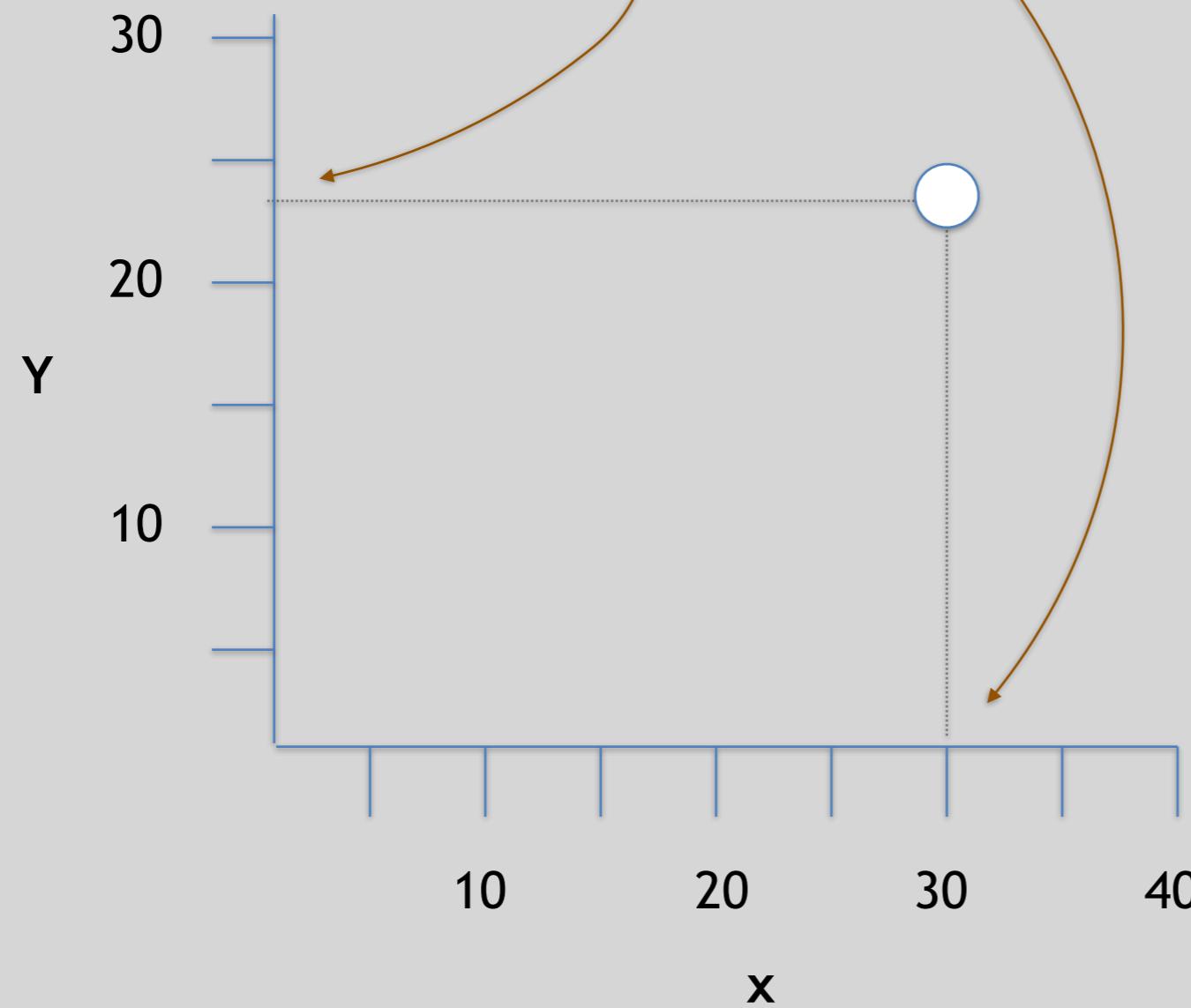
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



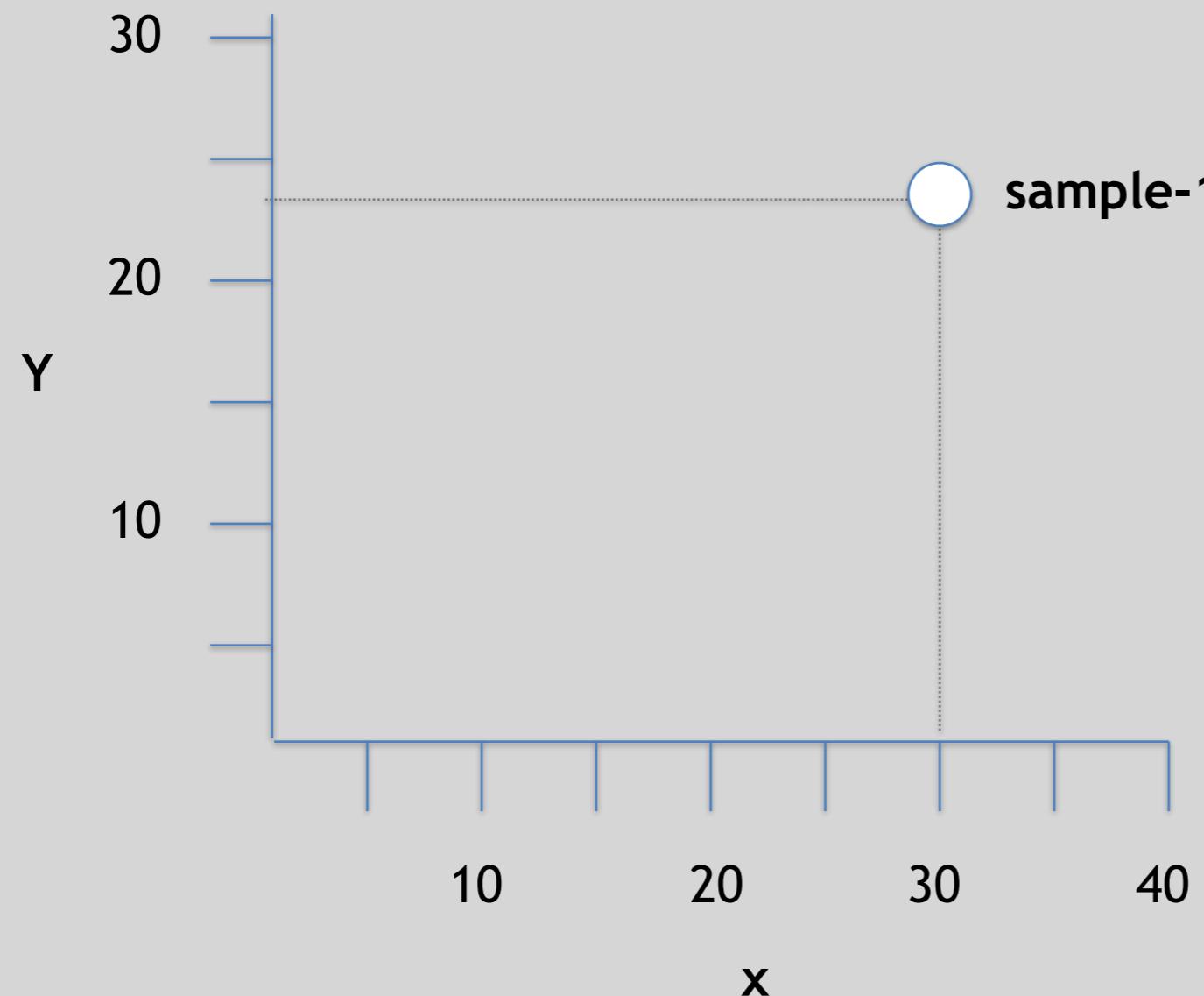
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



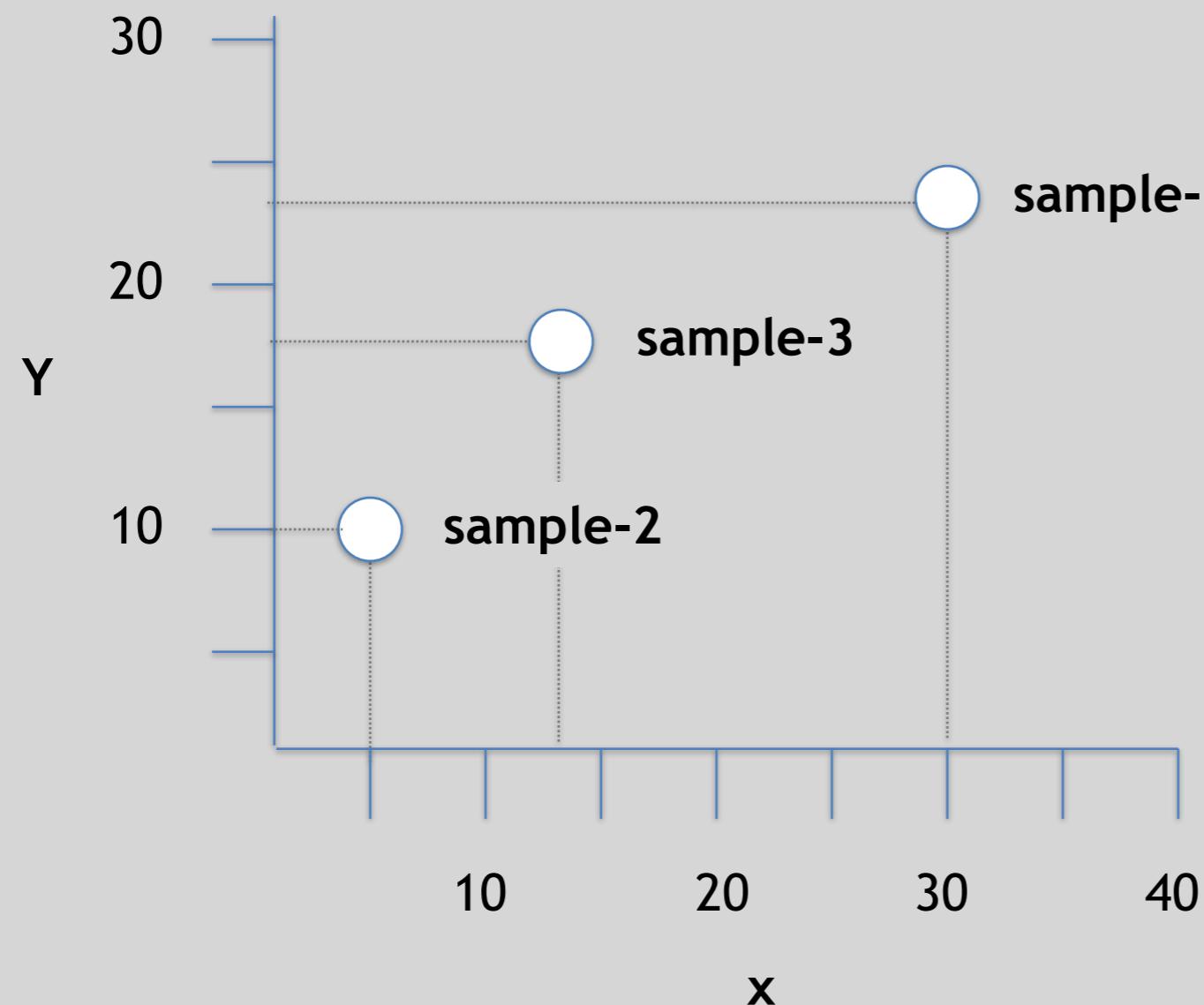
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



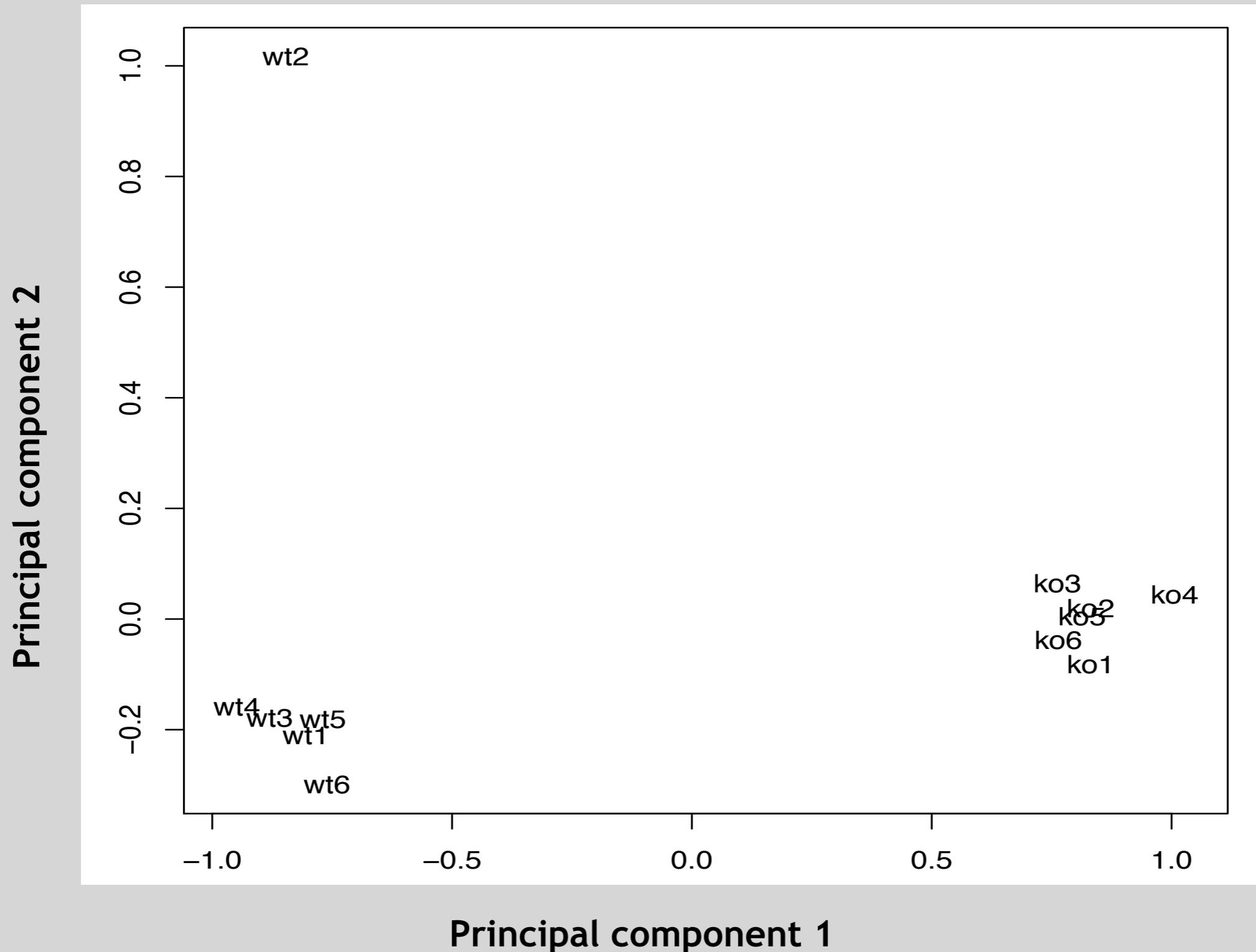
But we have 20,000+ genes...

So we would need a graph with 20,000+ axes to plot
the data!

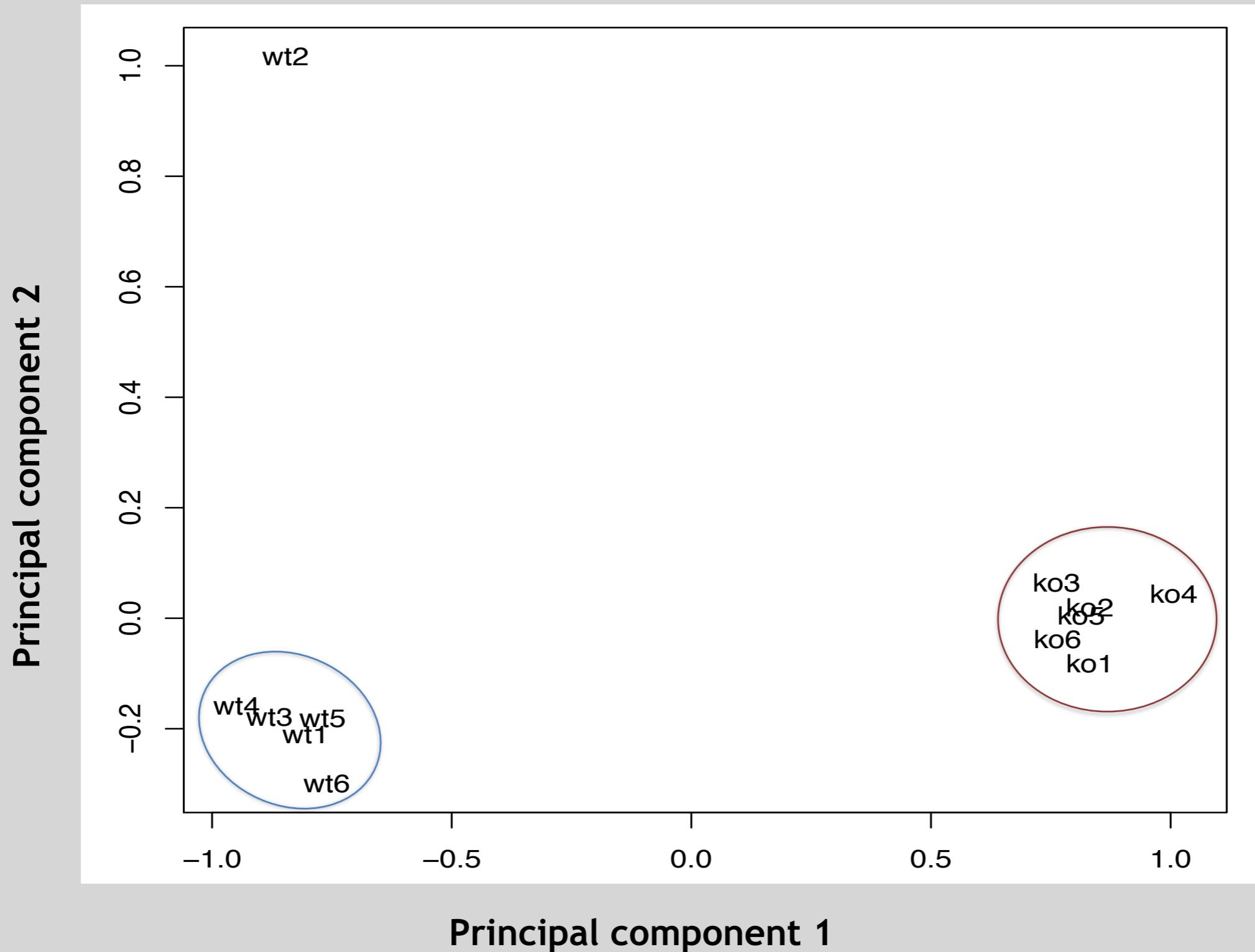
So we use PCA (principal component analysis) or something like it to plot this data.

PCA reduces the number of axes you need to display the important aspects of the data.

This is a PCA plot from a real RNA-seq experiment done on neural cells. The “wt” samples are “normal”. The “ko” samples are samples that were mutated.



This is a PCA plot from a real RNA-seq experiment done on neural cells. The “wt” samples are “normal”. The “ko” samples are samples that were mutated.

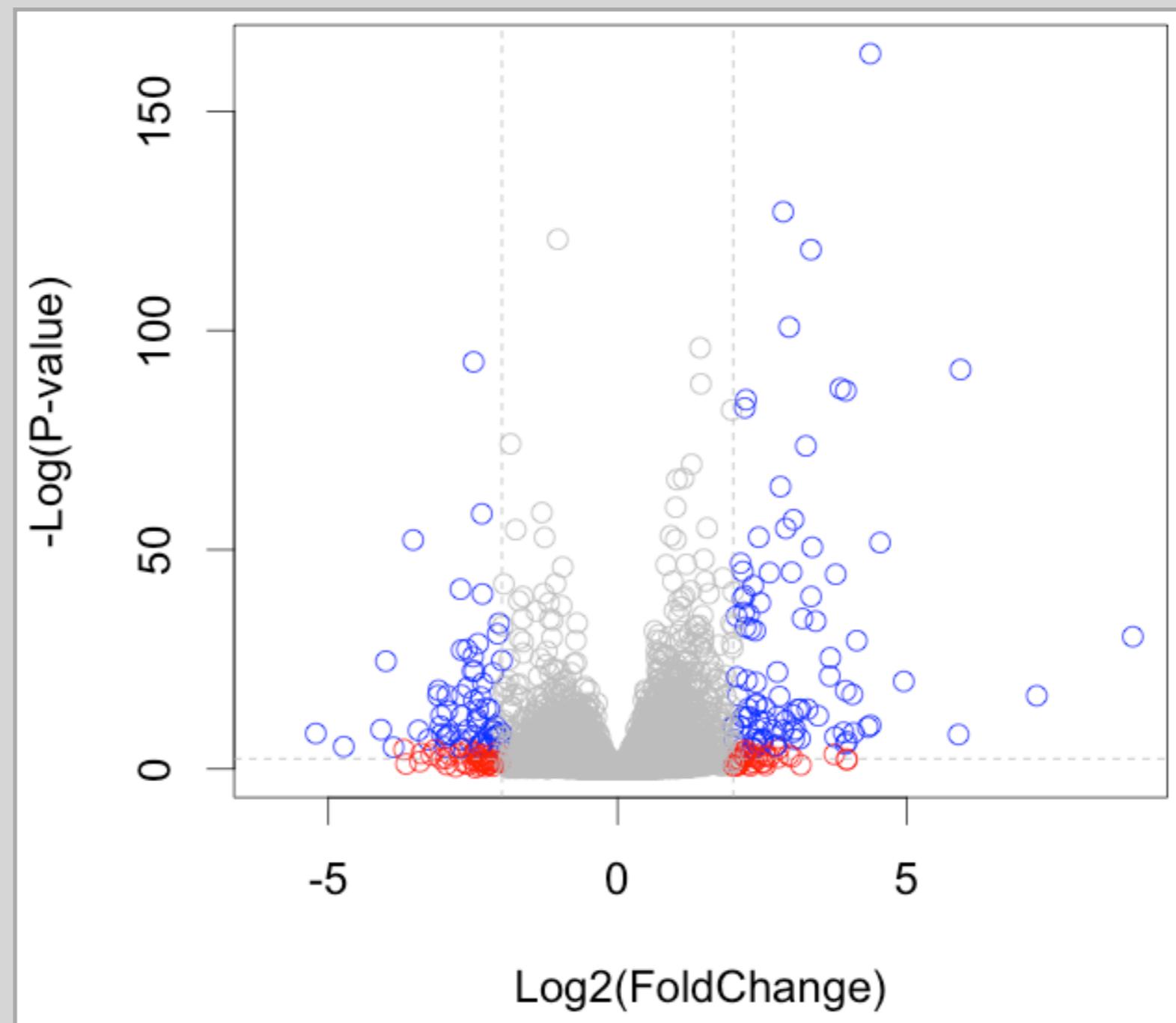


Plotting the data:

- (1) Tells us if we can expect to find some interesting differences**
- (2) Tells us if we should exclude some samples from any down stream analysis.**

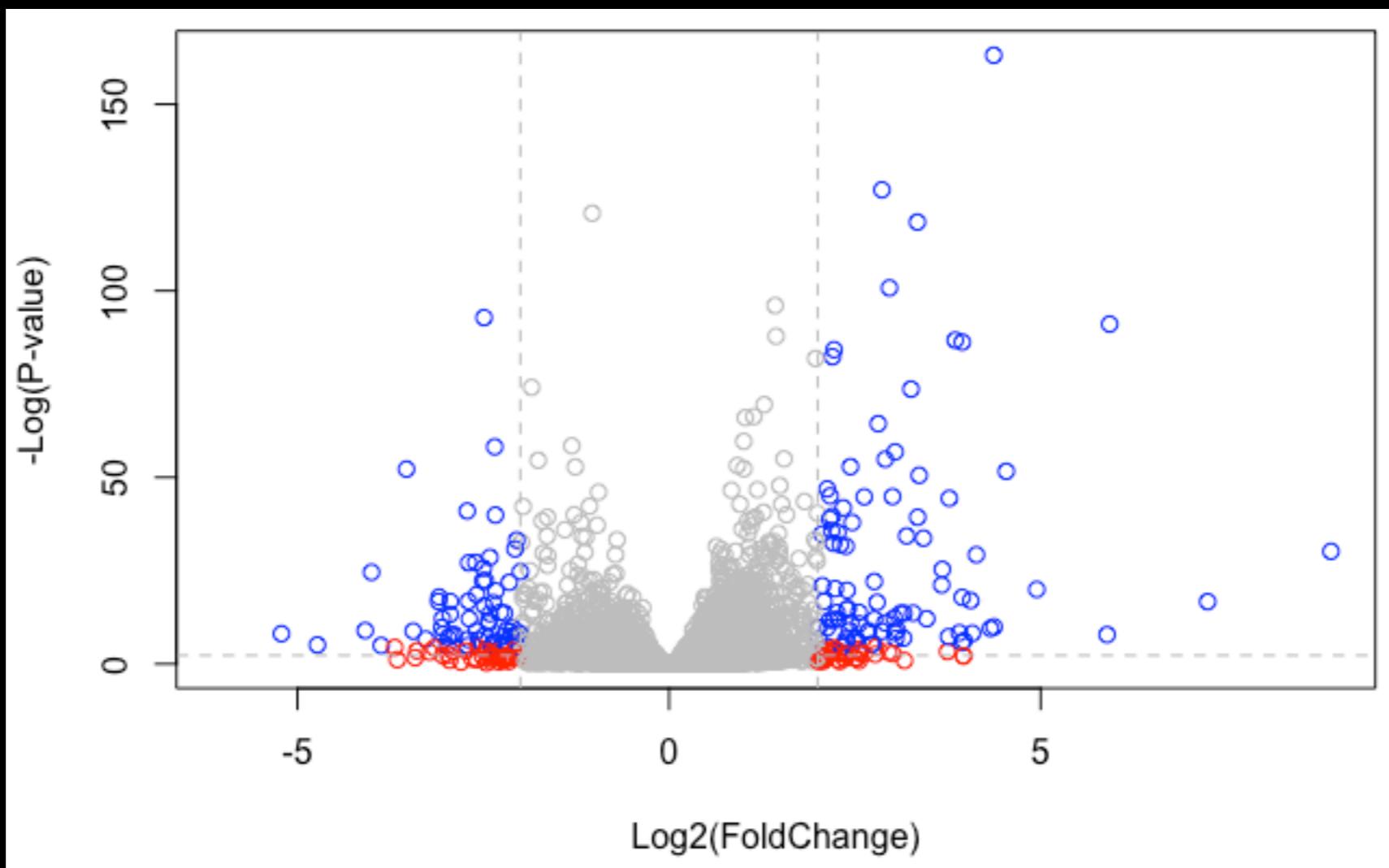
Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples

This is typically done using R with either the `edgeR` or `DESeq2` packages and the results are generally displayed using graphs like this one



Volcano plot

A common summary figure used to highlight genes that are both significantly regulated and display a high fold change



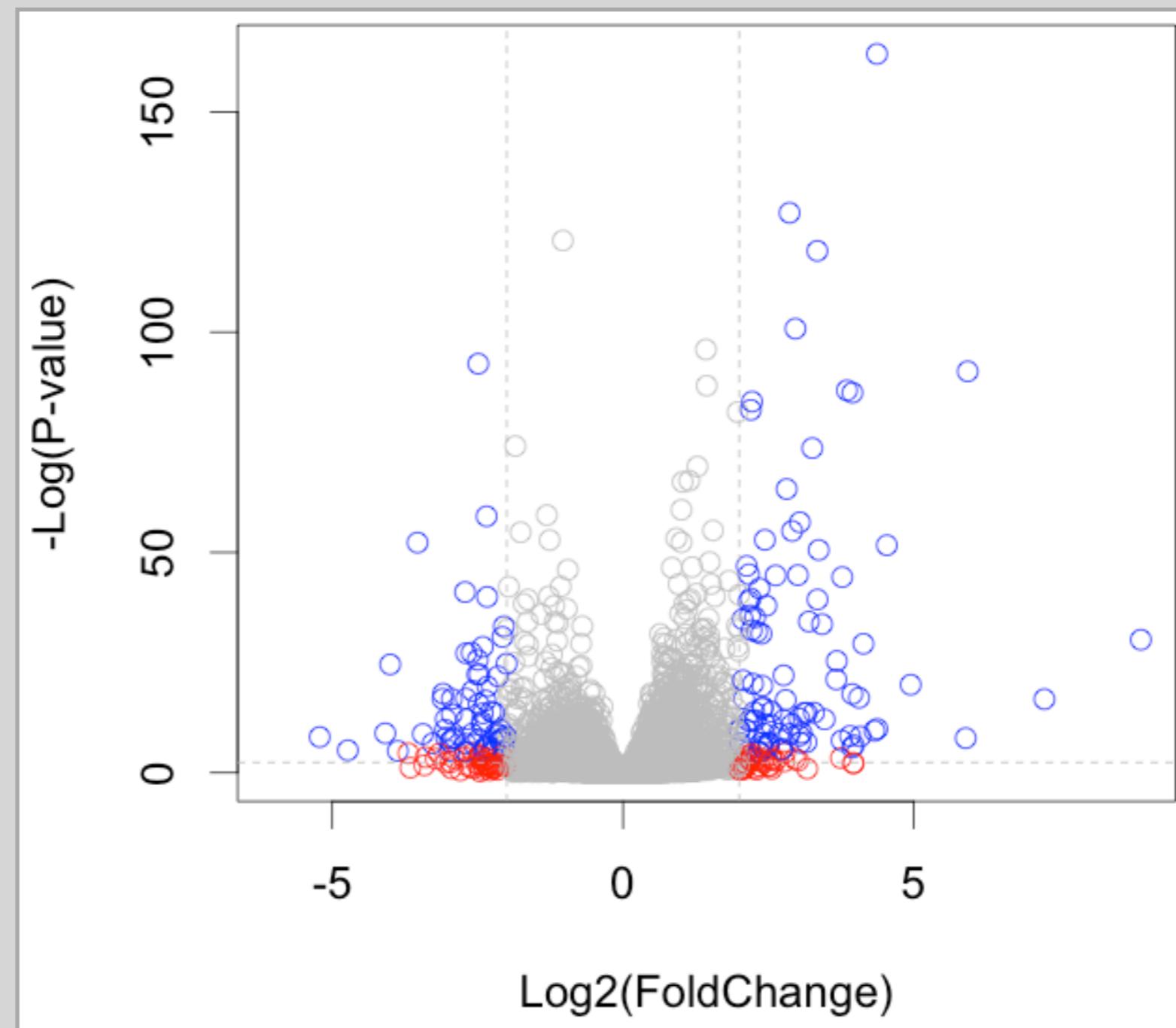
A volcano plot shows fold change (x-axis) versus -log of the *p-value* (y-axis) for a given transcript. The more significant the *p-value*, the larger the -log of that value will be. Therefore we often focus on 'higher up' points.

Fold change (log ratios)

- To a **statistician** fold change is sometimes considered **meaningless**. Fold change can be large (e.g. >>two-fold up- or down-regulation) without being statistically significant (e.g. based on probability values from a t-test or ANOVA).
- To a **biologist** fold change is almost always considered **important** for two reasons. First, a very small but statistically significant fold change might not be relevant to a cell's function. Second, it is of interest to know which genes are most dramatically regulated, as these are often thought to reflect changes in biologically meaningful transcripts and/or pathways.

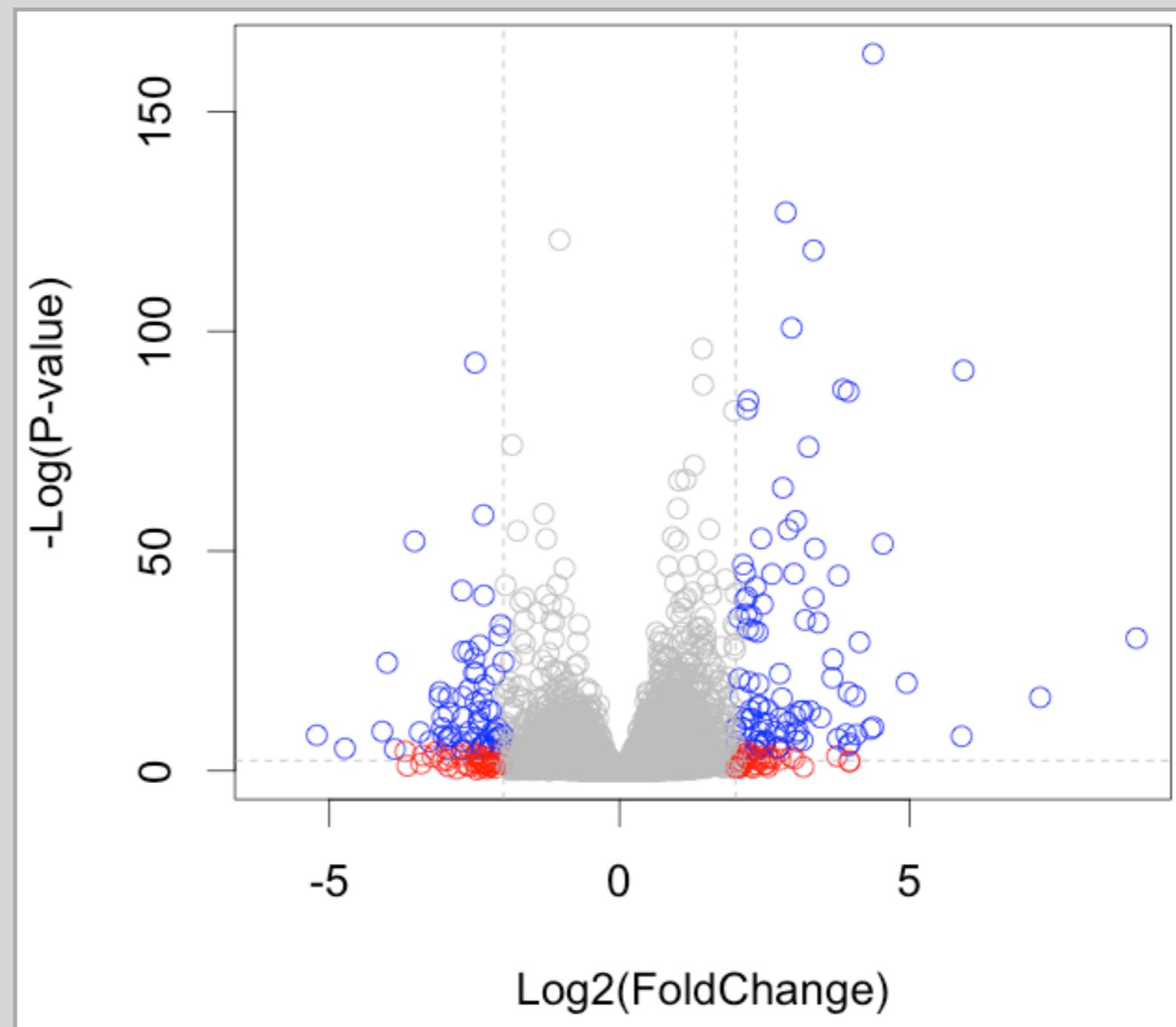
Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples

A **Blue** dot is a gene that is different between “normal” and “mutant” samples (gray dots are the same, **red** dots are different but this difference is not significant).



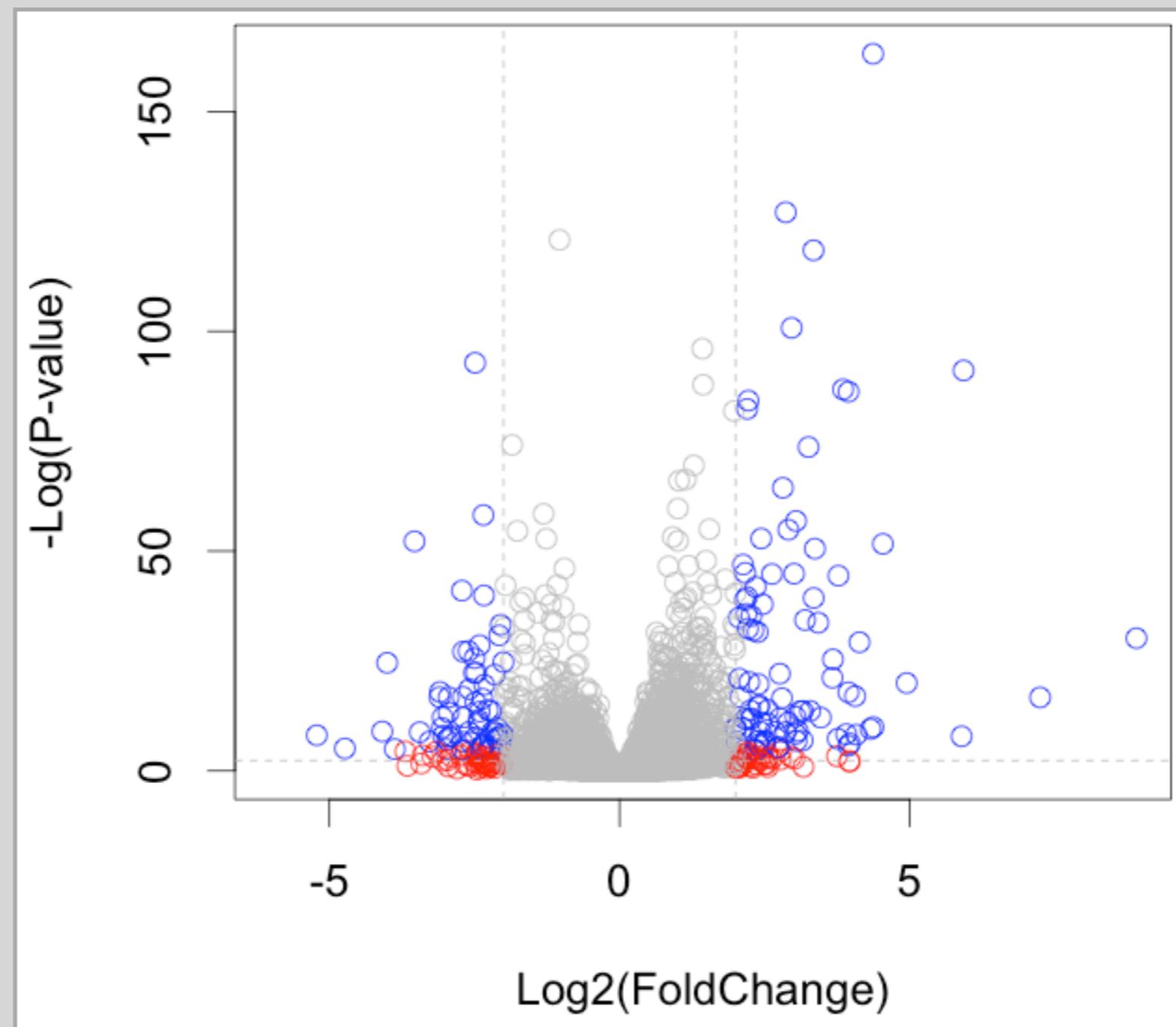
Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples

The x axis tells you how big the relative difference is between “normal” and “mutant” (FC stands for Fold change)

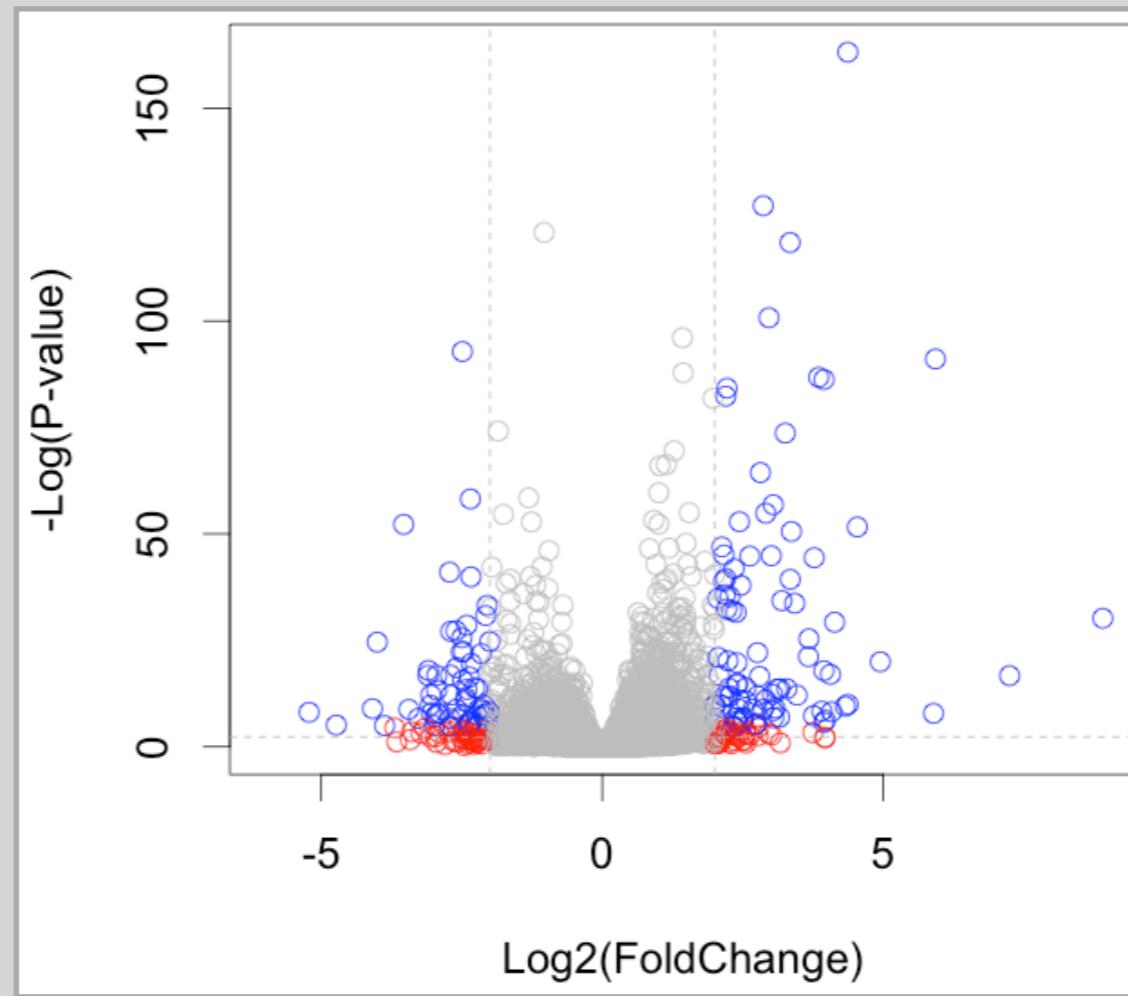


Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples

The y axis tells how significant the difference is given replicate samples (note we are plotting the minus log of the p-value, so large values on this axis represent smaller p-values)



Step 3 and beyond: We've identified interesting genes, now what?



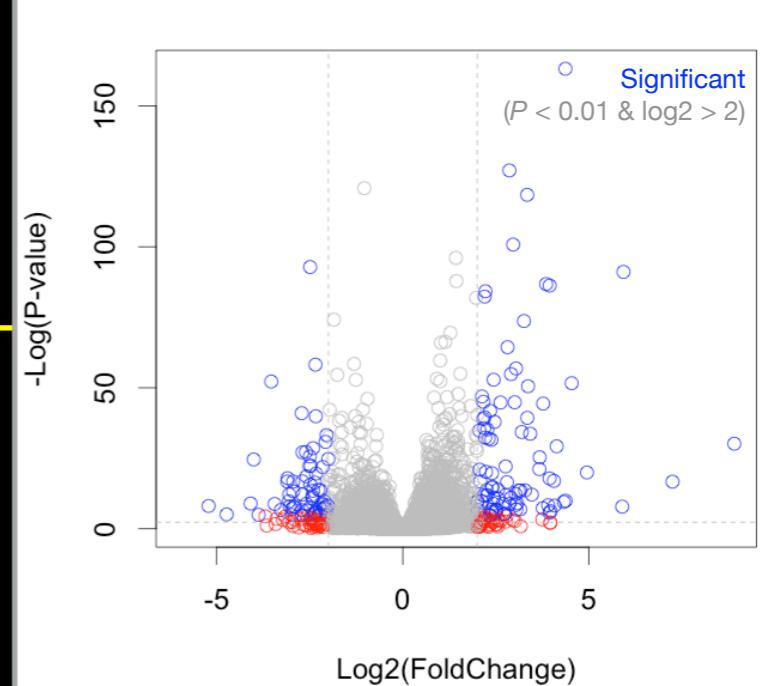
1. If you know what you're looking for, you can see if the experiment validated your hypothesis.
2. If you don't know what you're looking for, you can see if certain pathways are enriched in either the normal or mutant gene sets.



```
# Setup your point color vector  
> mycols <- rep("gray", nrow(res01))  
> mycols[ abs(res01$log2FoldChange) > 2 ] <- "red"  
  
> inds <- (res01$padj < 0.01) & (abs(res01$log2FoldChange) > 2 )  
> mycols[ inds ] <- "blue"
```

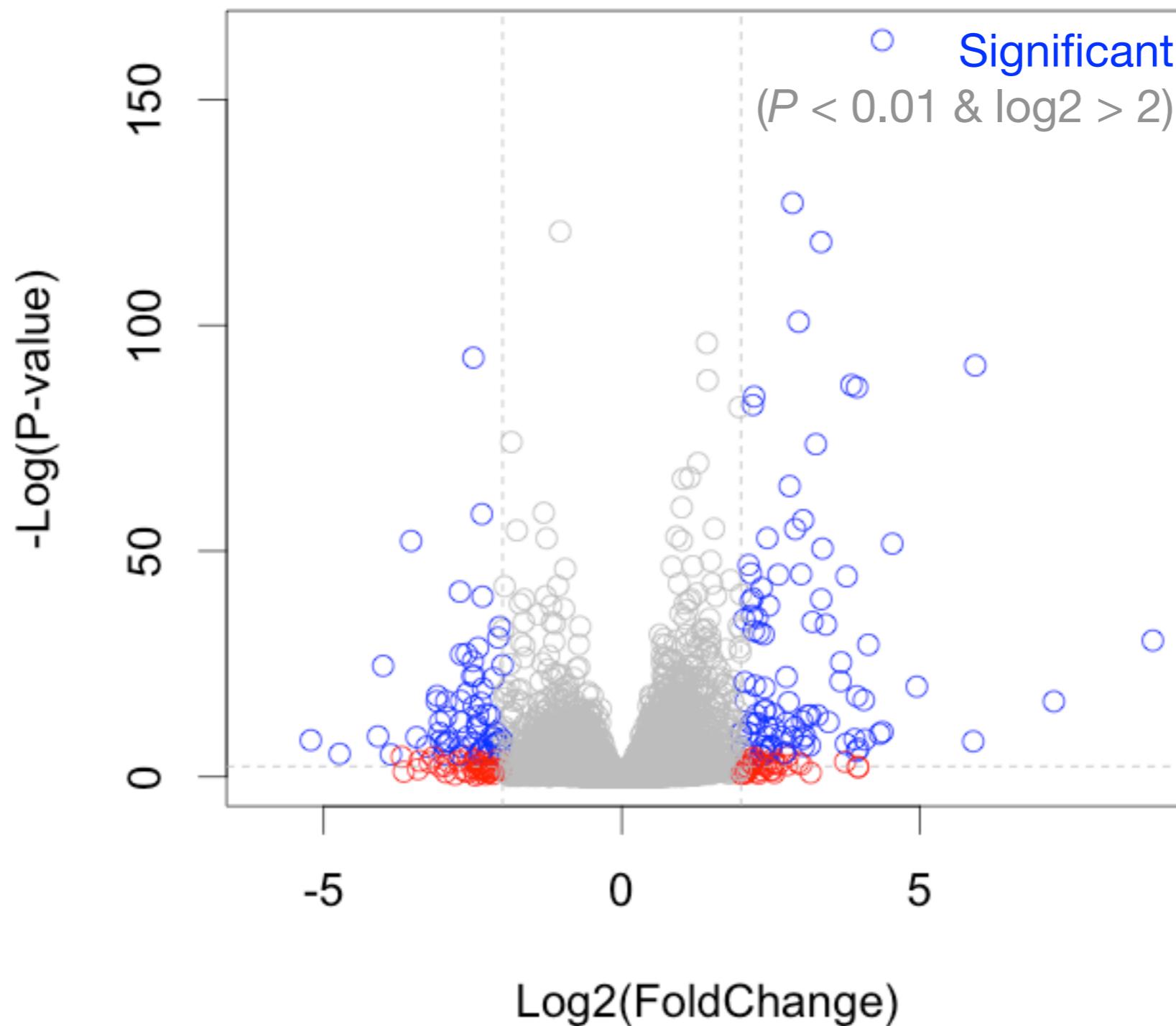
```
# Volcano plot with custom colors  
> plot( res01$log2FoldChange, -log(res01$padj),  
       col=mycols, ylab="-Log(P-value)",  
       xlab="Log2(FoldChange)" )
```

```
> abline(v=c(-2,2), col="gray", lty=2)  
> abline(h=-log(0.1), col="gray", lty=2)
```



Volcano Plot

Fold change vs P-value



Recent developments in RNA-Seq

- **Long read sequences:**
 - ➔ PacBio and Oxford Nanopore [[Recent Paper](#)]
- **Single-cell RNA-Seq:** [[Review article](#)]
 - ➔ Observe heterogeneity of cell populations
 - ➔ Detect sub-population
- **Alignment-free quantification:**
 - ➔ Kallisto [[Software link](#)]
 - ➔ Salmon [[Software link](#), [Blog post](#)]

Additional Reference Slides

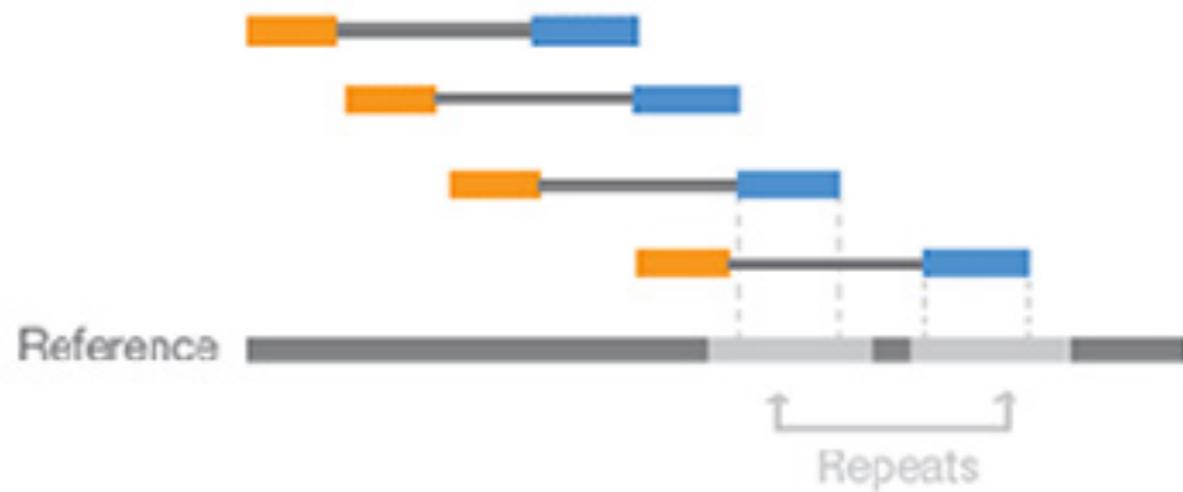
Public RNA-Seq data sources

- **Gene Expression Omnibus** (GEO):
 - ▶ <http://www.ncbi.nlm.nih.gov/geo/>
 - ▶ Both microarray and sequencing data
- **Sequence Read Archive** (SRA):
 - ▶ <http://www.ncbi.nlm.nih.gov/sra>
 - ▶ All sequencing data (not necessarily RNA-Seq)
- **ArrayExpress**:
 - ▶ <https://www.ebi.ac.uk/arrayexpress/>
 - ▶ European version of GEO
- All of these have links between them

Paired-End Reads



Alignment to the Reference Sequence



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Taken From: <https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html>

Count Normalization

- Normalization is required to make comparisons in gene expression
 - Between 2+ genes in one sample
 - Between genes in 2+ samples
- Genes will have more reads mapped in a sample with high coverage than one with low coverage
 - $2x \text{ depth} \approx 2x \text{ expression}$
- Longer genes will have more reads mapped than shorter genes
 - $2x \text{ length} \approx 2x \text{ more reads}$

Normalization: RPKM, FPKM & TPM

- N.B. Some tools for differential expression analysis such as edgeR and DESeq2 want raw read counts - i.e. non normalized input!
- However, often for your manuscripts and reports you will want to report normalized counts
- RPKM, FPKM and TPM all aim to normalize for sequencing depth and gene length. For the former:
 - Count up the total reads in a sample and divide that number by 1,000,000 - this is our “per million” scaling.
 - Divide the read counts by the “per million” scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
 - Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

- FPKM was made for paired-end RNA-seq
- With paired-end RNA-seq, two reads can correspond to a single fragment
- The only difference between RPKM and FPKM is that FPKM takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice).

- TPM is very similar to RPKM and FPKM. The only difference is the order of operations:
 - First divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
 - Count up all the RPK values in a sample and divide this number by 1,000,000. This is your “per million” scaling factor.
 - Divide the RPK values by the “per million” scaling factor. This gives you TPM.
- Note, the only difference is that you normalize for gene length first, and then normalize for sequencing depth second.

- When you use TPM, the sum of all TPMs in each sample are the same.
- This makes it easier to compare the proportion of reads that mapped to a gene in each sample.
- In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.