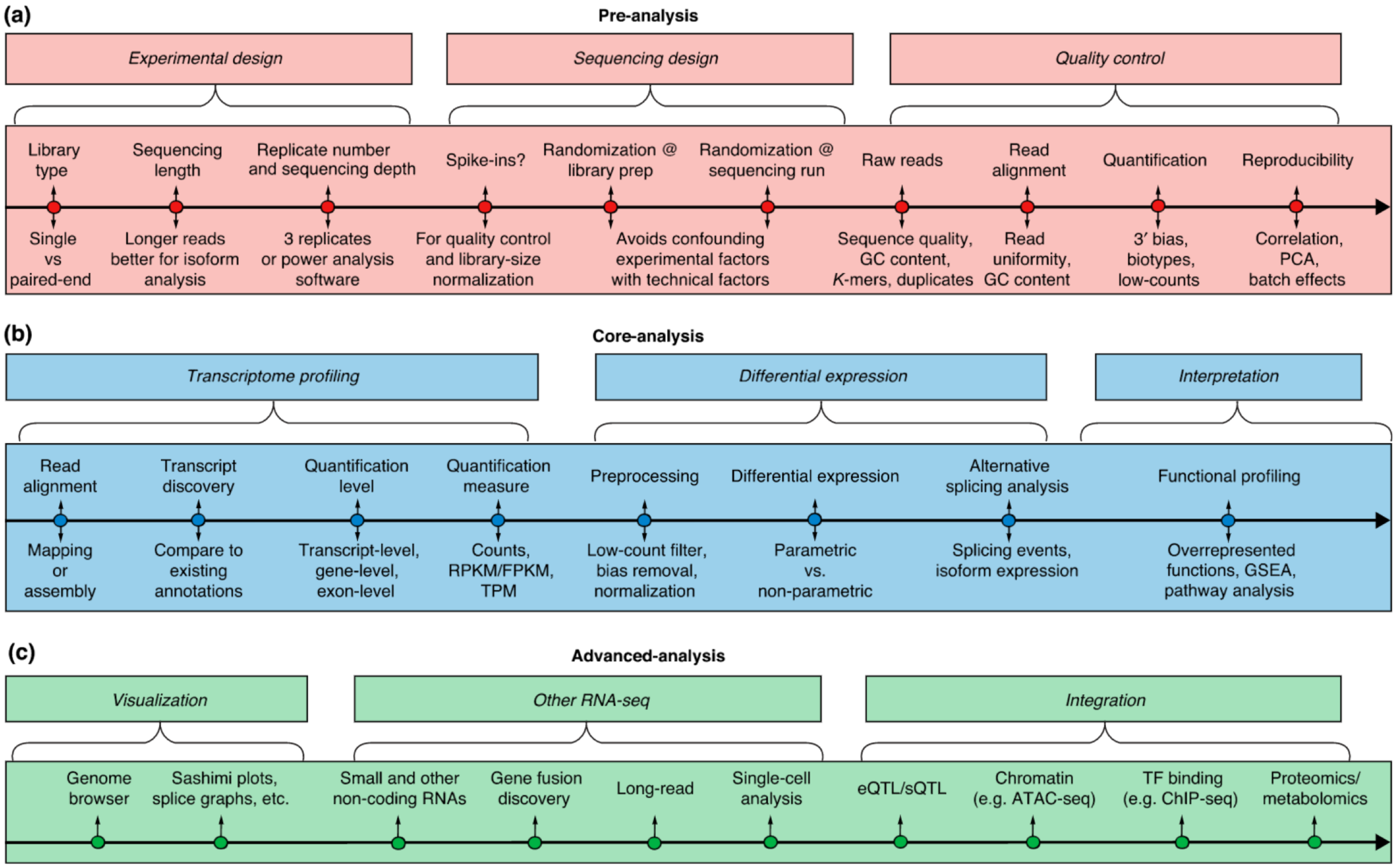
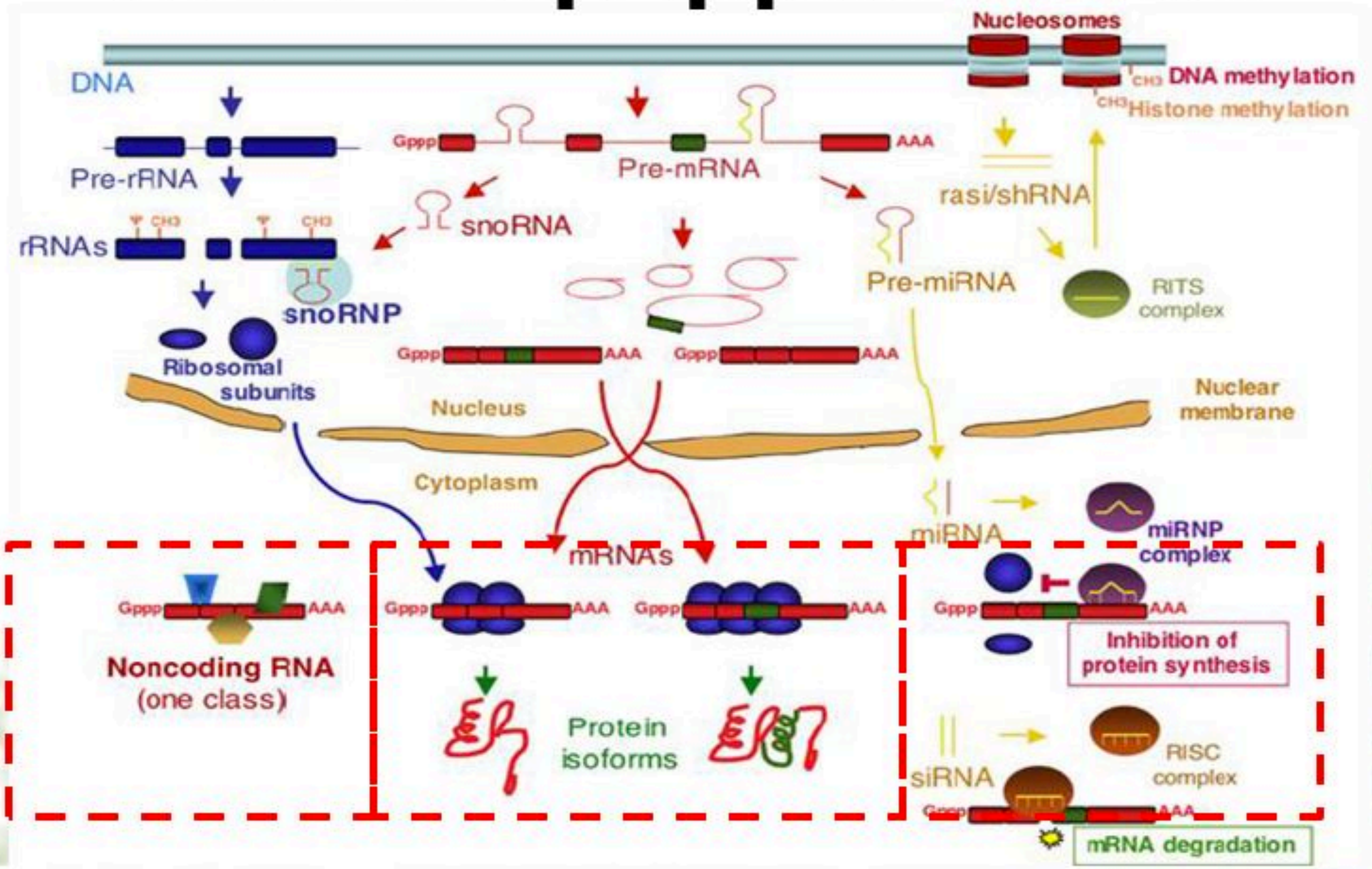

(a)
Pre-analysis
Experimental design
Sequencing design
Quality control
Library type
Single vs paired-end
Sequencing length
Longer reads better for isoform analysis
Replicate number and sequencing depth
3 replicates or power analysis software
Spike-ins?
For quality control and library-size normalization
Randomization @ library prep
Randomization @ sequencing run
Avoids confounding experimental factors with technical factors
Raw reads
Sequence quality, GC content, K-mers, duplicates
Read alignment
Read uniformity, GC content
Quantification
3′ bias, biotypes, low-counts
Reproducibility
Correlation, PCA, batch effects
(b)
Core-analysis
Transcriptome profiling
Differential expression
Interpretation
Read alignment
Mapping or assembly
Transcript discovery
Compare to existing annotations
Quantification level
Transcript-level, gene-level, exon-level
Quantification measure
Counts, RPKM/FPKM, TPM
Preprocessing
Low-count filter, bias removal, normalization
Differential expression
Parametric vs. non-parametric
Alternative splicing analysis
Splicing events, isoform expression
Functional profiling
Overrepresented functions, GSEA, pathway analysis
(c)
Advanced-analysis
Visualization
Other RNA-seq
Integration
Genome browser
Sashimi plots, splice graphs, etc.
Small and other non-coding RNAs
Gene fusion discovery
Long-read
Single-cell analysis
eQTL/sQTL
Chromatin (e.g. ATAC-seq)
TF binding (e.g. ChIP-seq)
Proteomics/ metabolomics

Fig. 1 A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

RNA-seq Applications



Long ncRNA

Expression level
Structure
SNP
Novel ncRNA

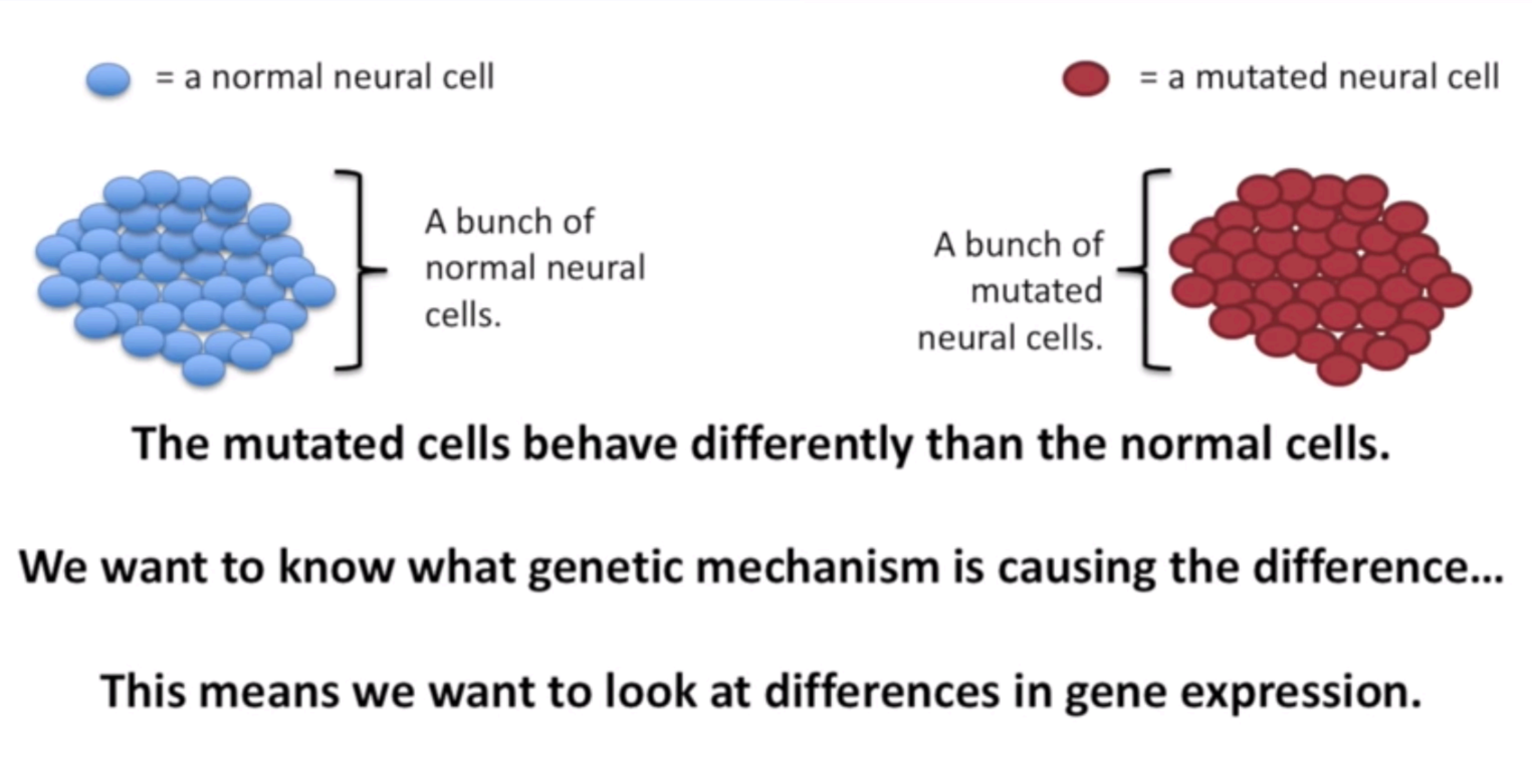
mRNA

Expression level
Alternative splicing
SNP
RNA editting
Gene fusion
Novel mRNA
Transcriptome assembly

Small RNA

Expression level
SNP
Novel small RNA

RNA-Seq experiments



🔵 = a normal neural cell

🔴 = a mutated neural cell

A bunch of normal neural cells.

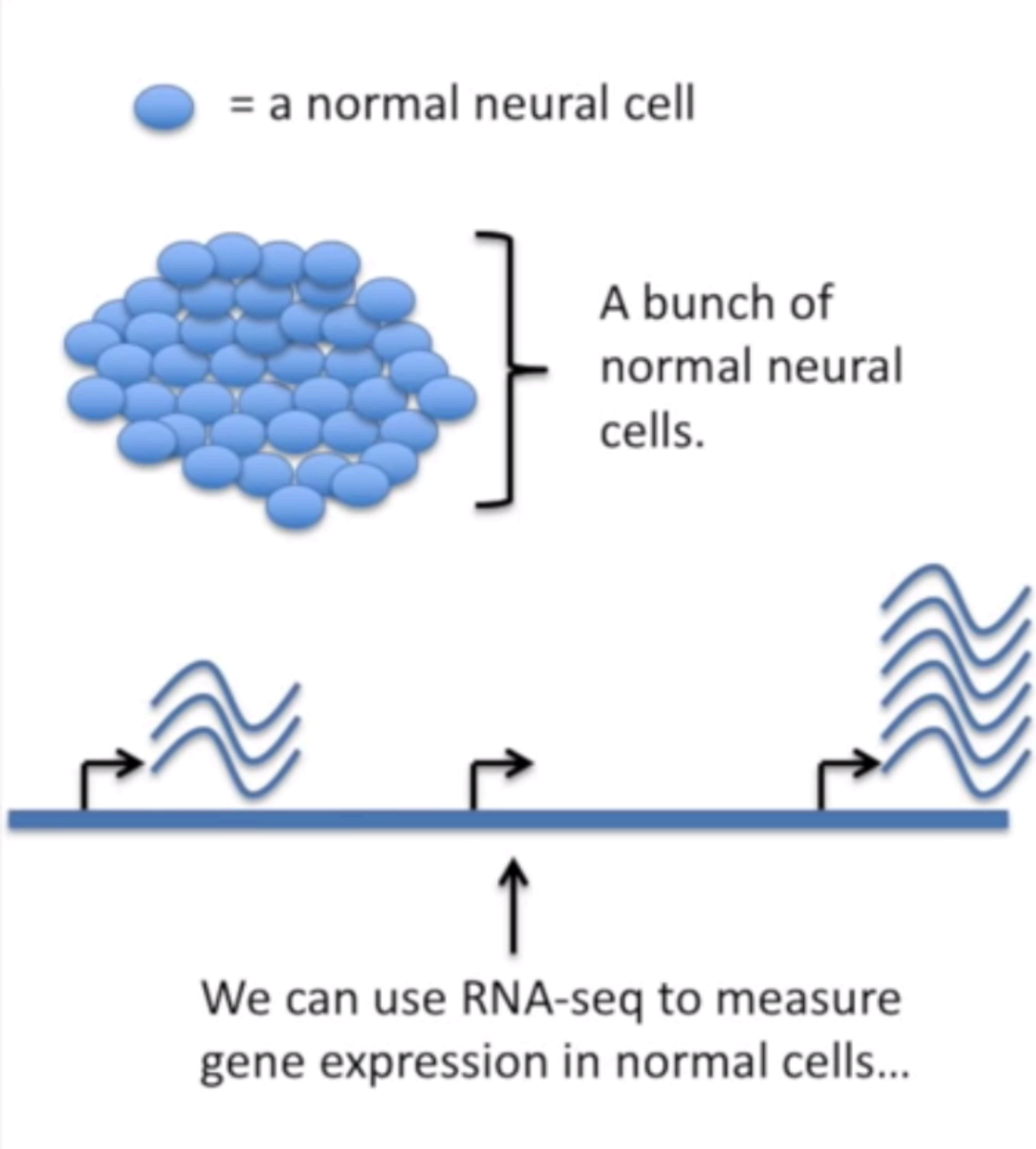
A bunch of mutated neural cells.

The mutated cells behave differently than the normal cells.

We want to know what genetic mechanism is causing the difference…

This means we want to look at differences in gene expression.

RNA-Seq experiments

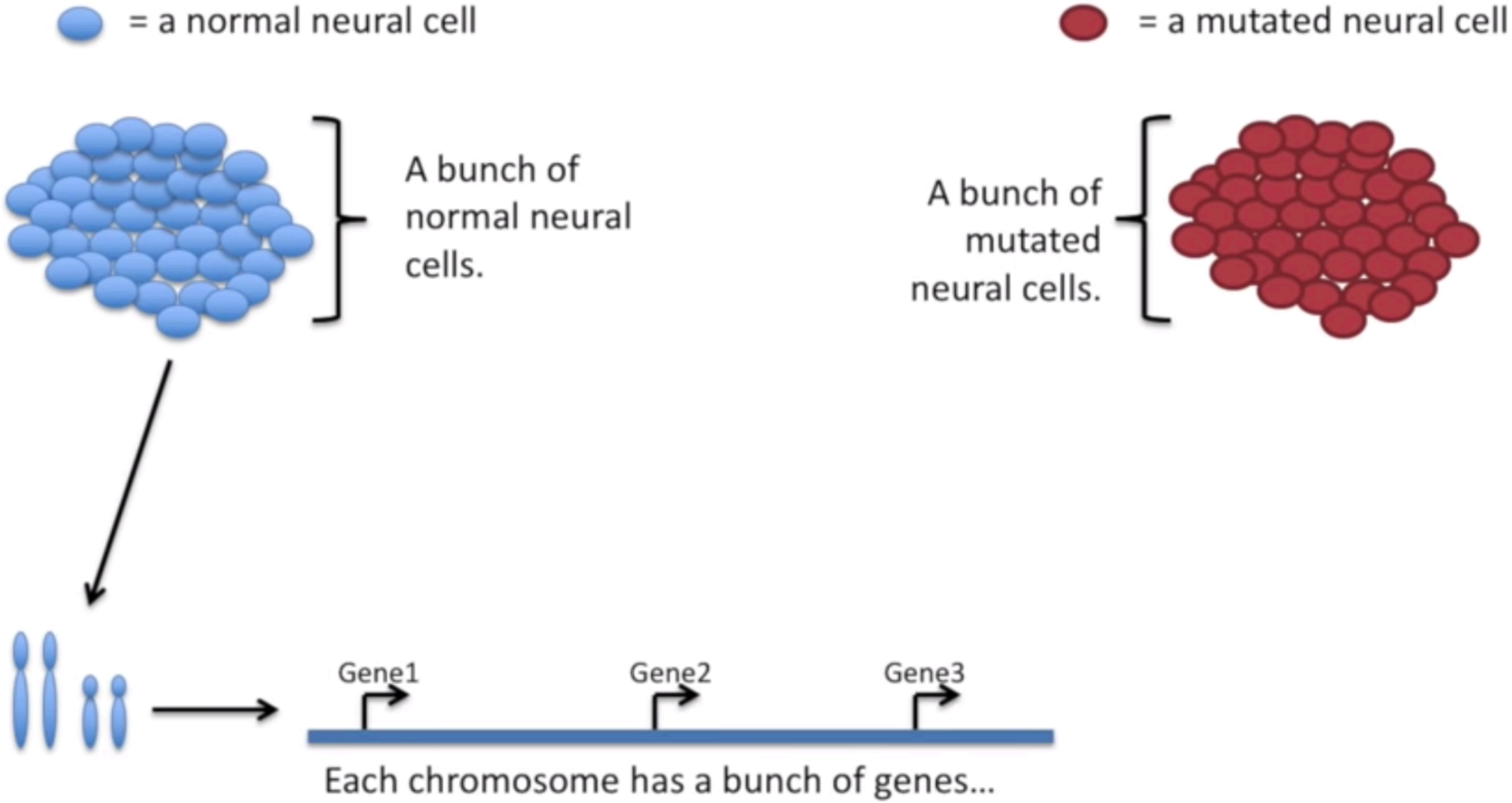


= a normal neural cell

A bunch of normal neural cells.

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

RNA-Seq experiments



= a normal neural cell

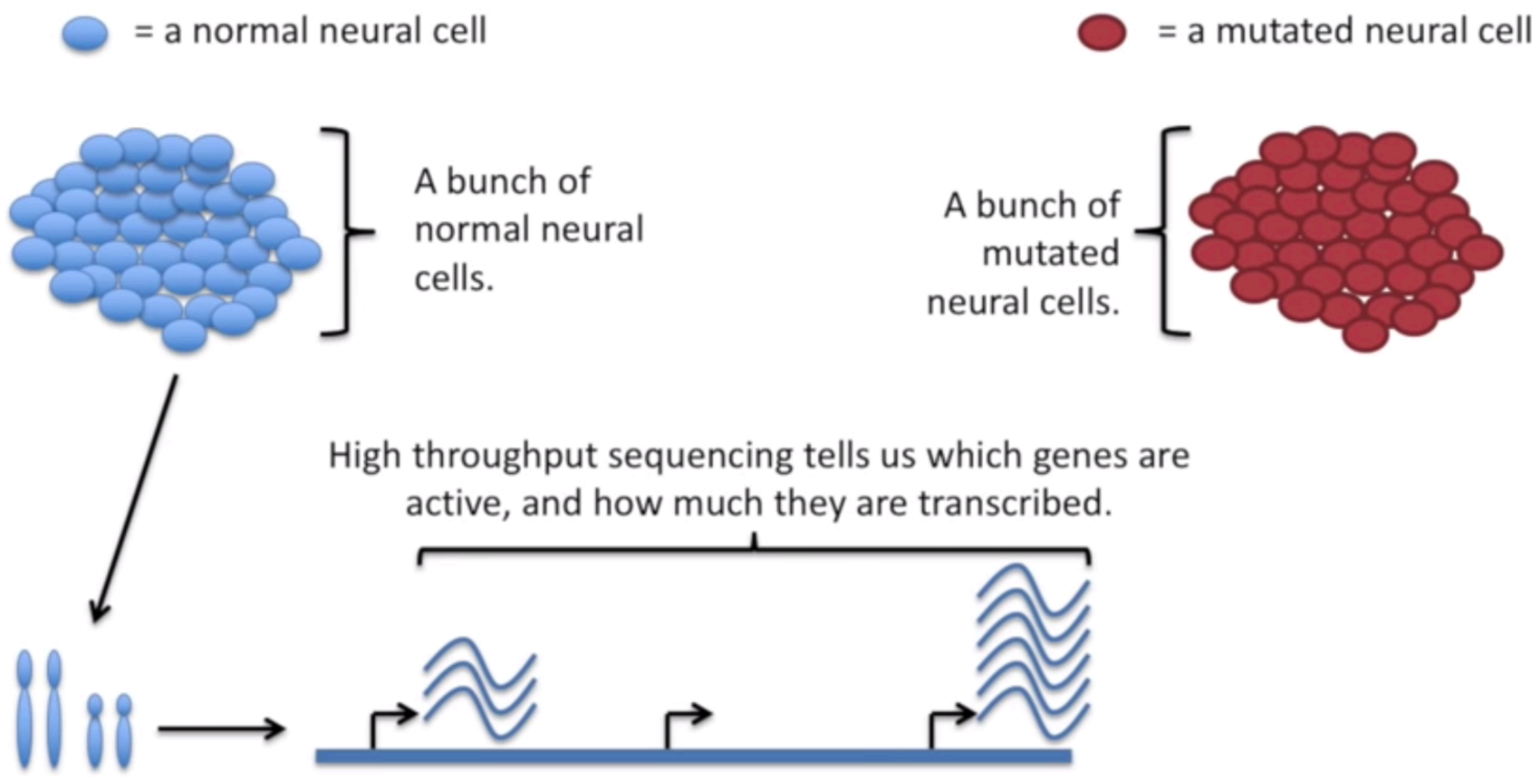
= a mutated neural cell

A bunch of normal neural cells.

A bunch of mutated neural cells.

Each chromosome has a bunch of genes...

RNA-Seq experiments



● = a normal neural cell

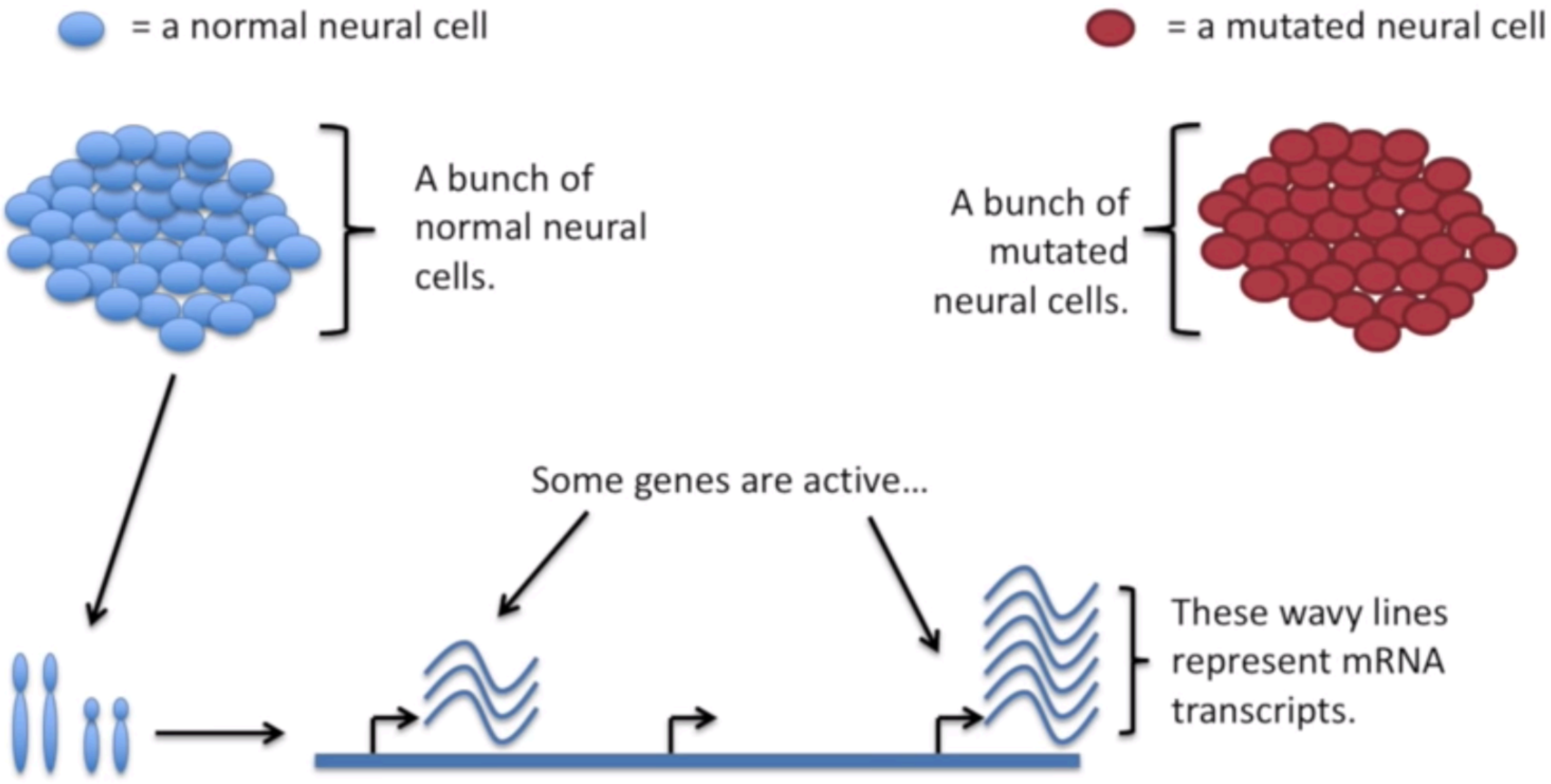
● = a mutated neural cell

A bunch of normal neural cells.

A bunch of mutated neural cells.

High throughput sequencing tells us which genes are active, and how much they are transcribed.

RNA-Seq experiments



🔵 = a normal neural cell

🔴 = a mutated neural cell

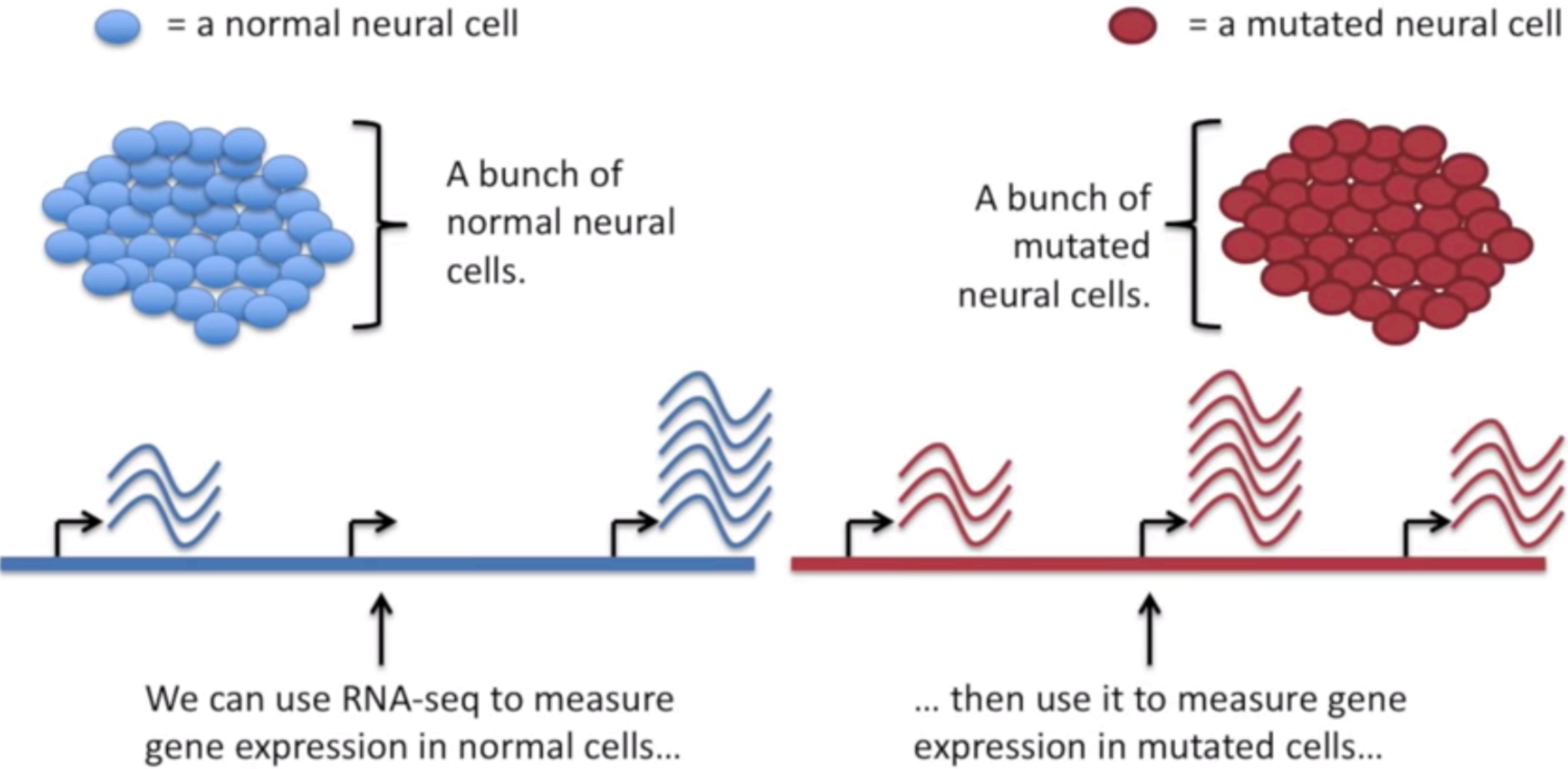
A bunch of normal neural cells.

A bunch of mutated neural cells.

Some genes are active...

These wavy lines represent mRNA transcripts.

RNA-Seq experiments



🔵 = a normal neural cell

🔴 = a mutated neural cell

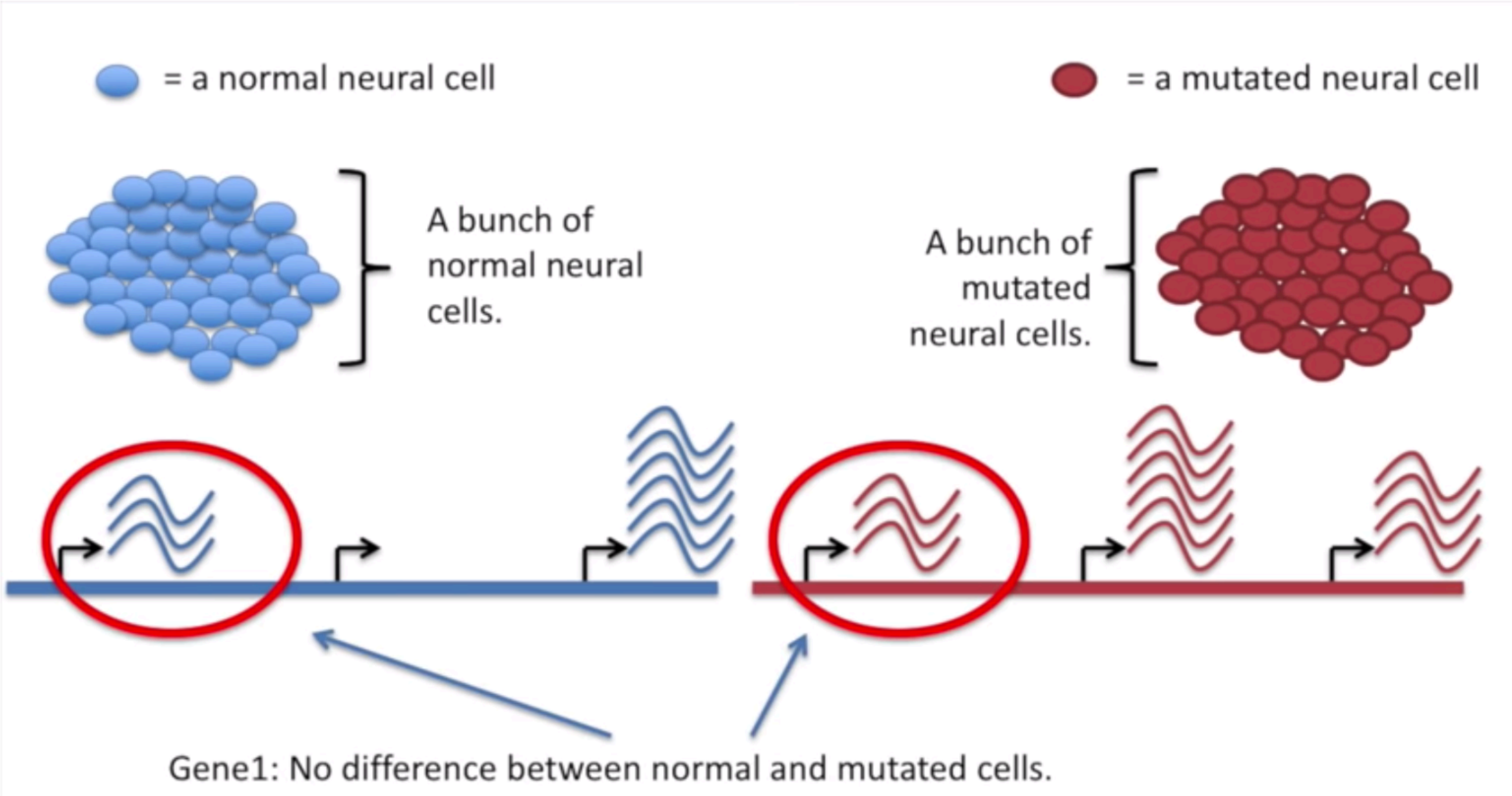
A bunch of normal neural cells.

A bunch of mutated neural cells.

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

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

RNA-Seq experiments



● = a normal neural cell

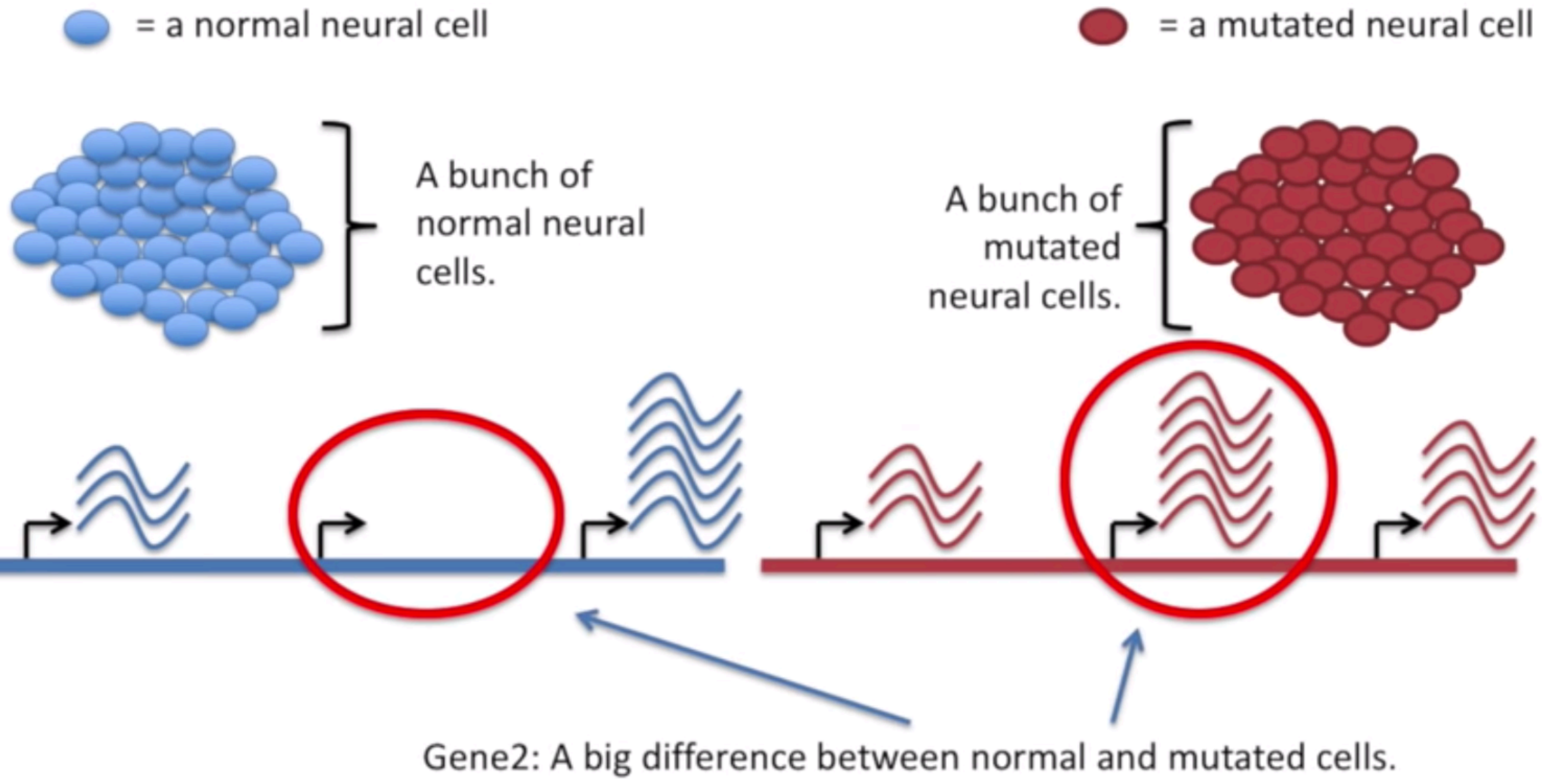
● = a mutated neural cell

A bunch of normal neural cells.

A bunch of mutated neural cells.

Gene1: No difference between normal and mutated cells.

RNA-Seq experiments



● = a normal neural cell

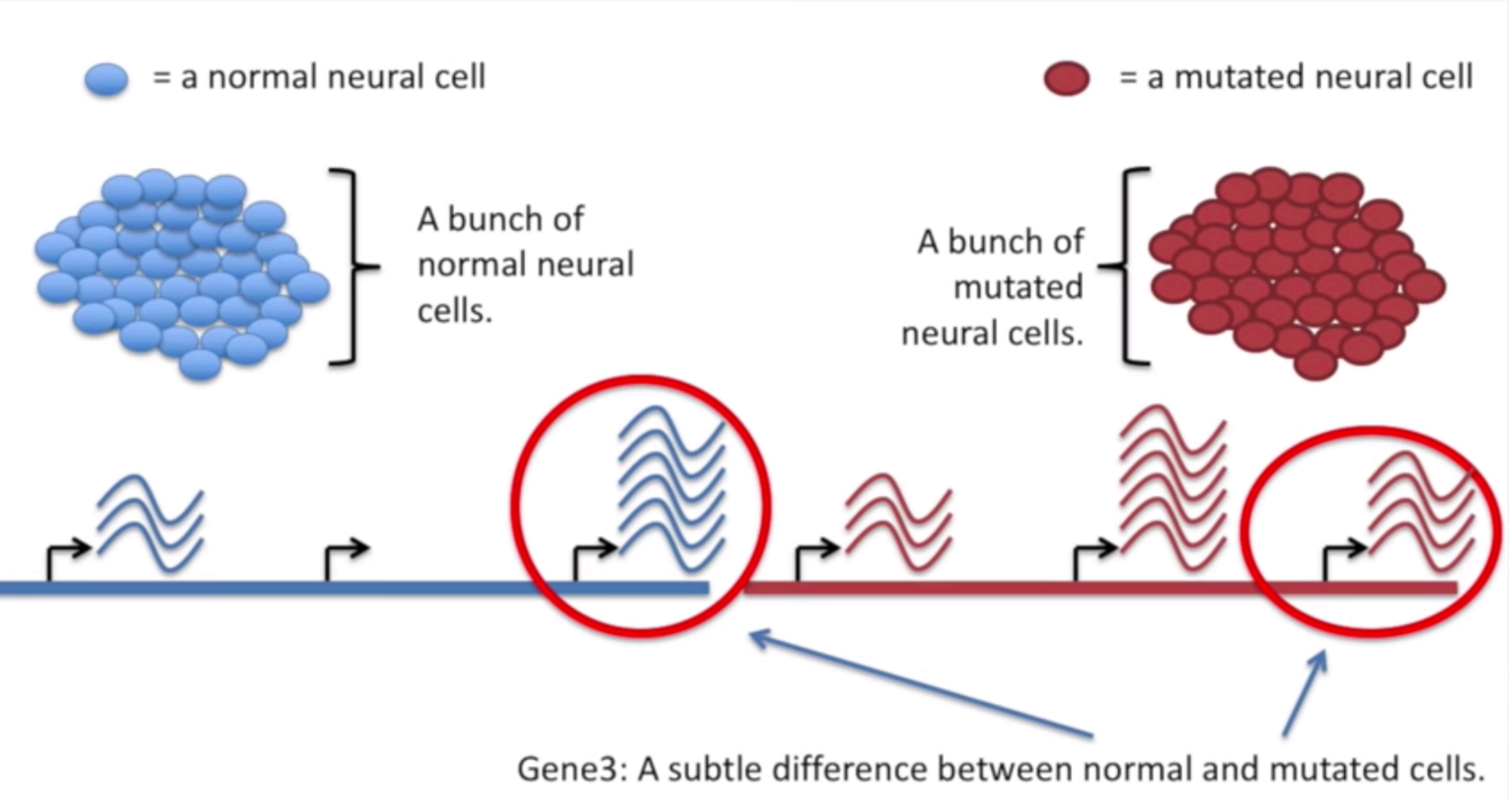
● = a mutated neural cell

A bunch of normal neural cells.

A bunch of mutated neural cells.

Gene2: A big difference between normal and mutated cells.

RNA-Seq experiments



● = a normal neural cell

● = a mutated neural cell

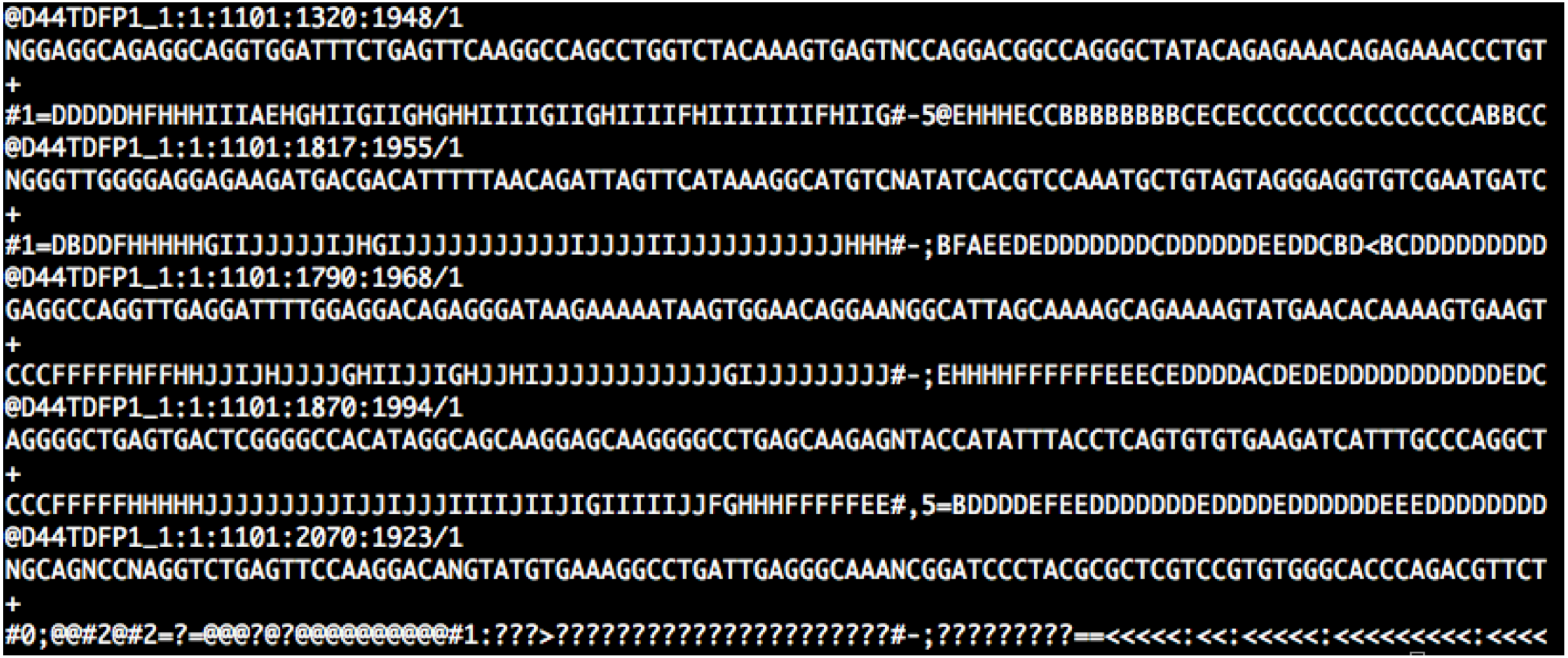
A bunch of normal neural cells.

A bunch of mutated neural cells.

Gene3: A subtle difference between normal and mutated cells.

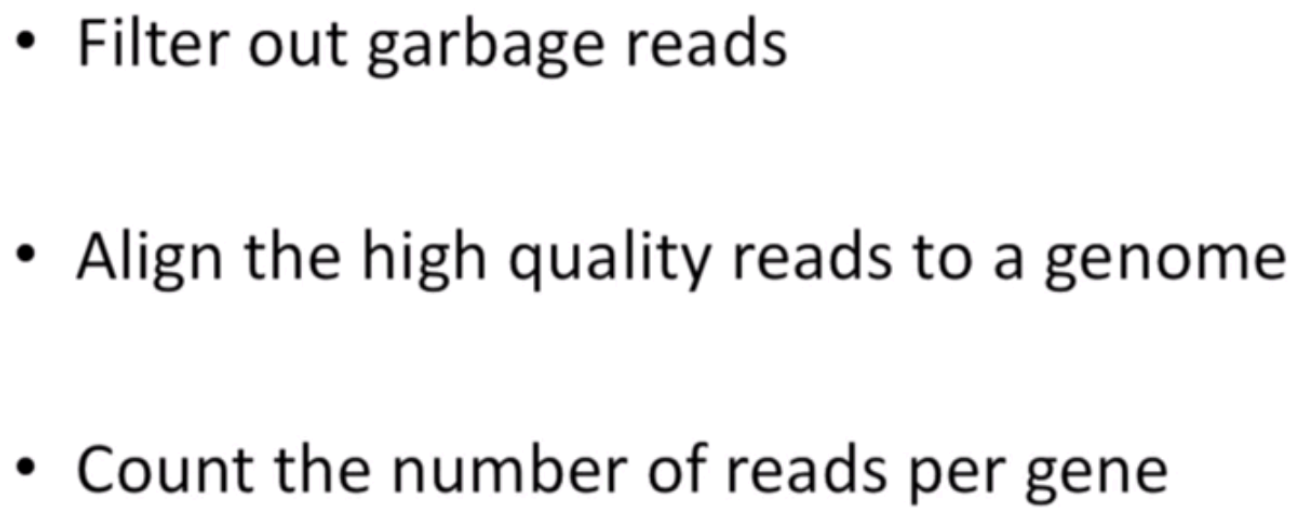
Input Data Structure

Fastq File



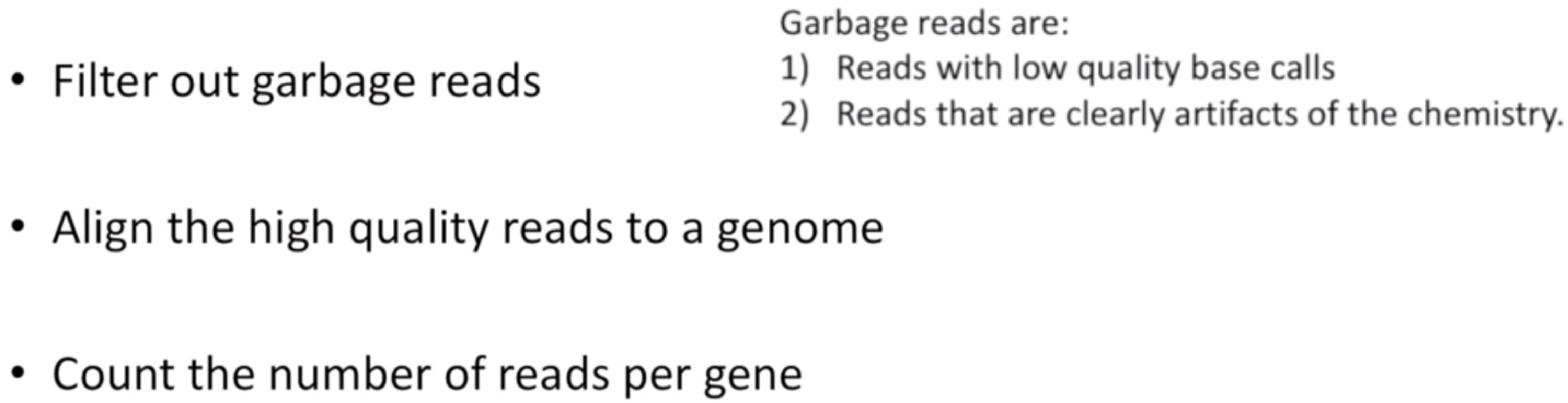
```
@D44TDFP1_1:1:1101:1320:1948/1
NGGAGGCAGAGGCAGGTGGATTTCTGAGTTCAAGGCCAGCCTGGTCTACAAAGTGAGTNCCAGGACGGCCAGGGCTATACAGAGAAACAGAGAAACCCTGT
+
#1=DDDDDHFHHHIIIAEHGHIIGIIGHGHHIIIIGIIGHIIIIFHIIIIIIIIFHIIG#-5@EHHHECCBBBBBBBBCECECCCCCCCCCCCCCCCCABBCC
@D44TDFP1_1:1:1101:1817:1955/1
NGGGTTGGGGAGGAGAAGATGACGACATTTTTAACAGATTAGTTCATAAAGGCATGTCNATATCACGTCCAAATGCTGTAGTAGGGAGGTGTCGAATGATC
+
#1=DBDDFHHHHHGIIJJJJJIJHGIJJJJJJJJJJJIJJJJIIJJJJJJJJJJJHHH#-;BFAEEDEDDDDDDDCDDDDDDEEDDCBD<BCDDDDDDDDD
@D44TDFP1_1:1:1101:1790:1968/1
GAGGCCAGGTTGAGGATTTTGGAGGACAGAGGGATAAGAAAAATAAGTGGAACAGGAANGGCATTAGCAAAAGCAGAAAAGTATGAACACAAAAGTGAAGT
+
CCCFFFFFHFFHHJJIJHJJJJGHIIJJIGHJJHIJJJJJJJJJJJJGIJJJJJJJJJ#-;EHHHHFFFFFFEEECEDDDDACDEDEDDDDDDDDDDDEDC
@D44TDFP1_1:1:1101:1870:1994/1
AGGGGCTGAGTGACTCGGGGCCACATAGGCAGCAAGGAGCAAGGGGCCTGAGCAAGAGNTACCATATTTACCTCAGTGTGTGAAGATCATTTGCCCAGGCT
+
CCCFFFFFHHHHHJJJJJJJJJIJJIJJJIIIIJIIJIGIIIIIJJFGHHHFFFFFEE#,5=BDDDDEFEEDDDDDDDEDDDDEDDDDDDEEEDDDDDDDD
@D44TDFP1_1:1:1101:2070:1923/1
NGCAGNCCNAGGTCTGAGTTCCAAGGACANGTATGTGAAAGGCCTGATTGAGGGCAAANCGGATCCCTACGCGCTCGTCCGTGTGGGCACCCAGACGTTCT
+
#0;@@#2@#2=?=@@@?@?@@@@@@@@@@@#1:???>??????????????????????#-;?????????==<<<<<:<<:<<<<<:<<<<<<<<<:<<<<
```

With the set of reads obtained from the sequencing we need to:



- Filter out garbage reads
- Align the high quality reads to a genome
- Count the number of reads per gene

With the set of reads obtained from the sequencing we need to:



- Filter out garbage reads

Garbage reads are:
1) Reads with low quality base calls
2) Reads that are clearly artifacts of the chemistry.

- Align the high quality reads to a genome

- Count the number of reads per gene

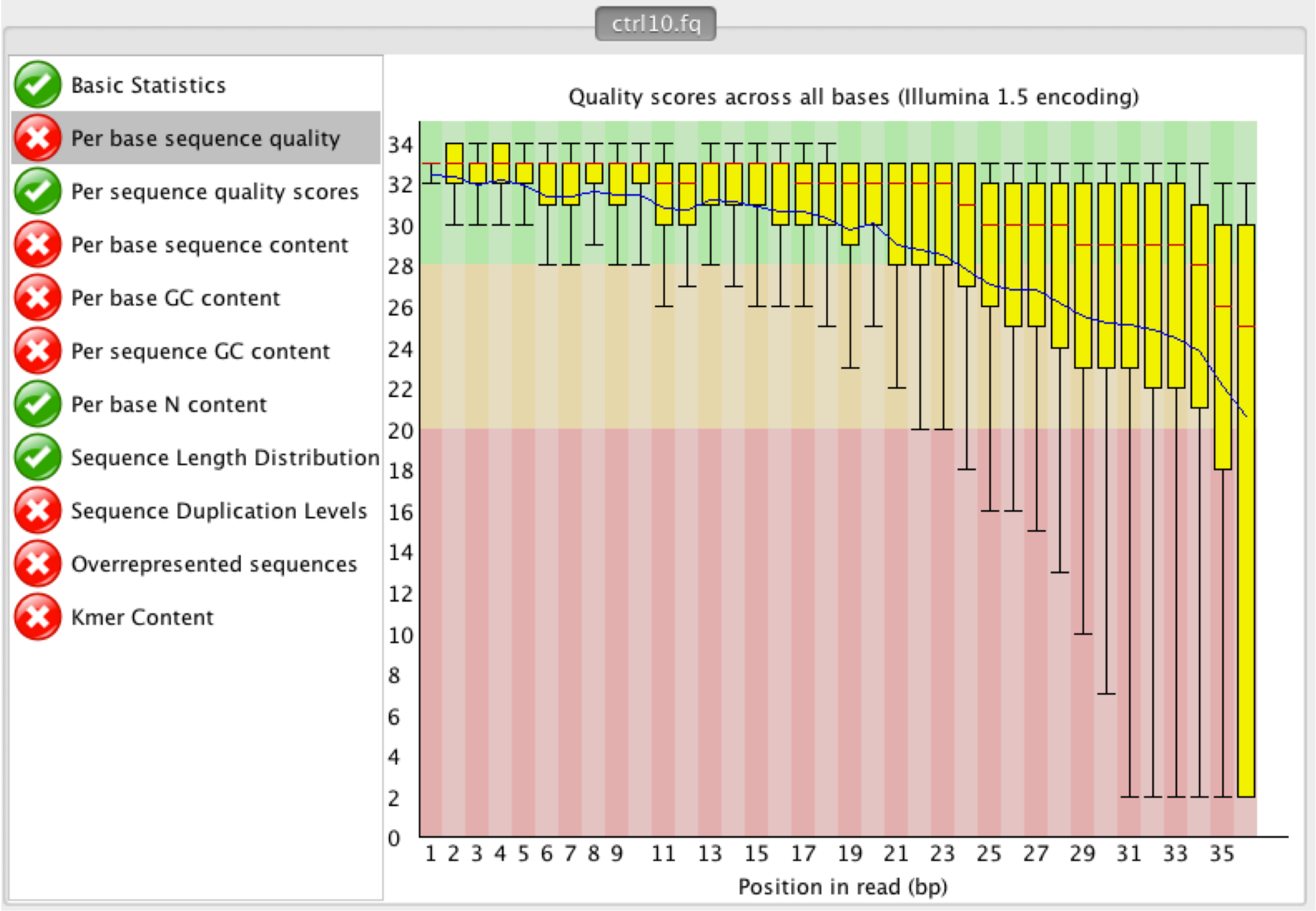
Inspecting raw data

Fastq QC

- Before starting a RNA-seq analysis it is better to have a look at the overall quality of raw data.
- FastQC is a java tool that allows quality controls at the level of various type of sequencing files.

Low quality base calls

Inspecting raw data



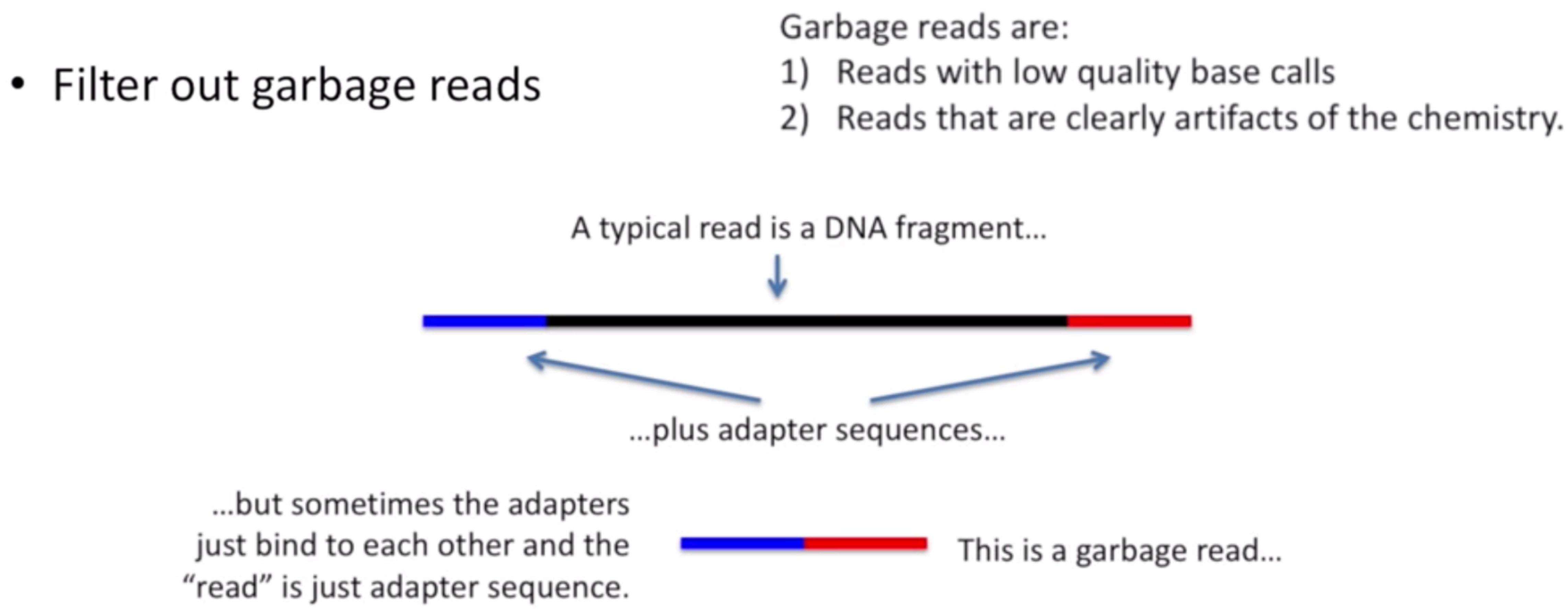
ctrl10.fq

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✗ Kmer Content

Red median value
Blue mean value

Background code:
Green: good
Orange: reasonable
Red: poor

With the set of reads obtained from the sequencing we need to:



- Filter out garbage reads

Garbage reads are:

1) Reads with low quality base calls
2) Reads that are clearly artifacts of the chemistry.

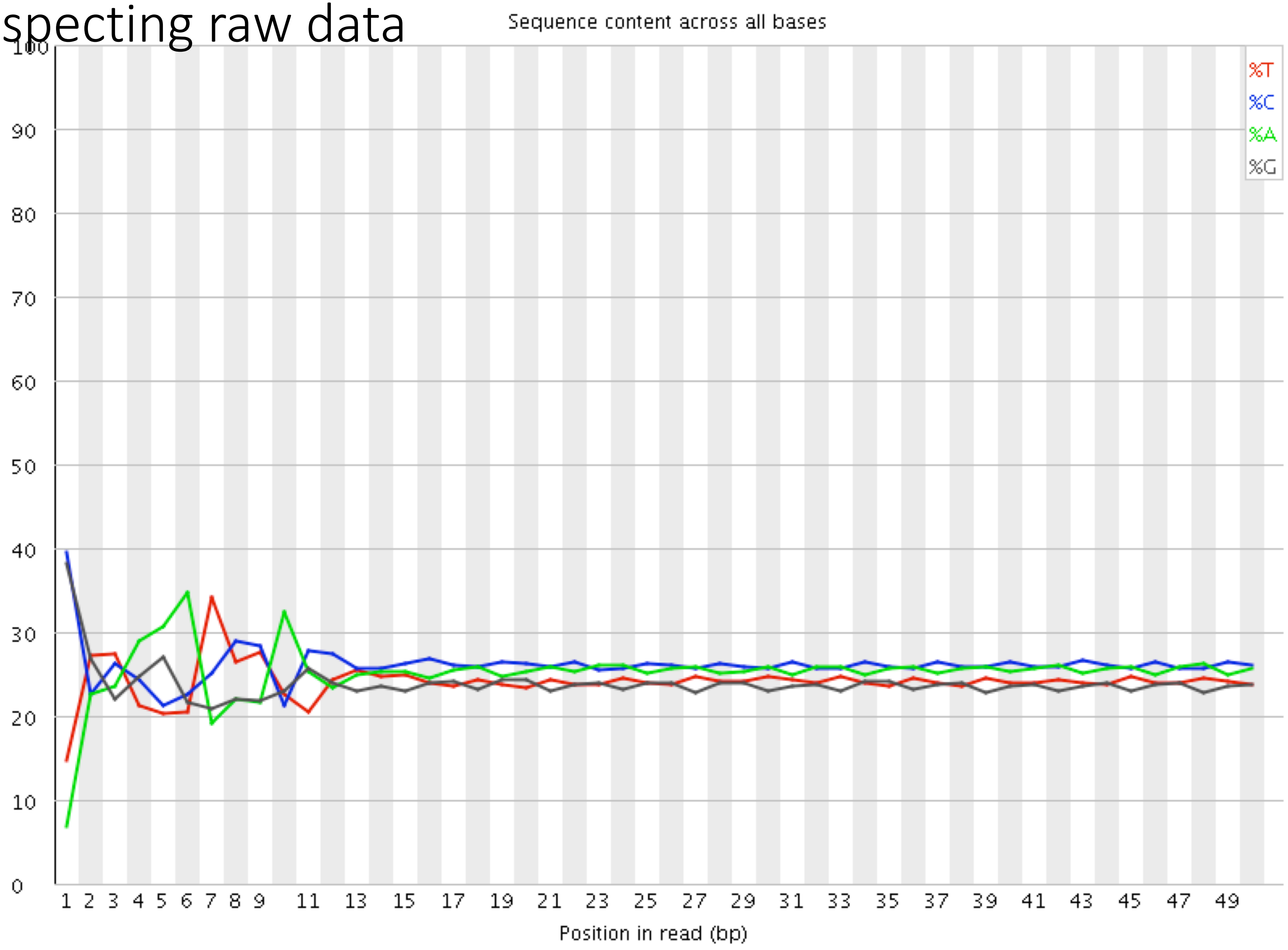
A typical read is a DNA fragment…

…plus adapter sequences…

…but sometimes the adapters just bind to each other and the “read” is just adapter sequence.

This is a garbage read…

Inspecting raw data



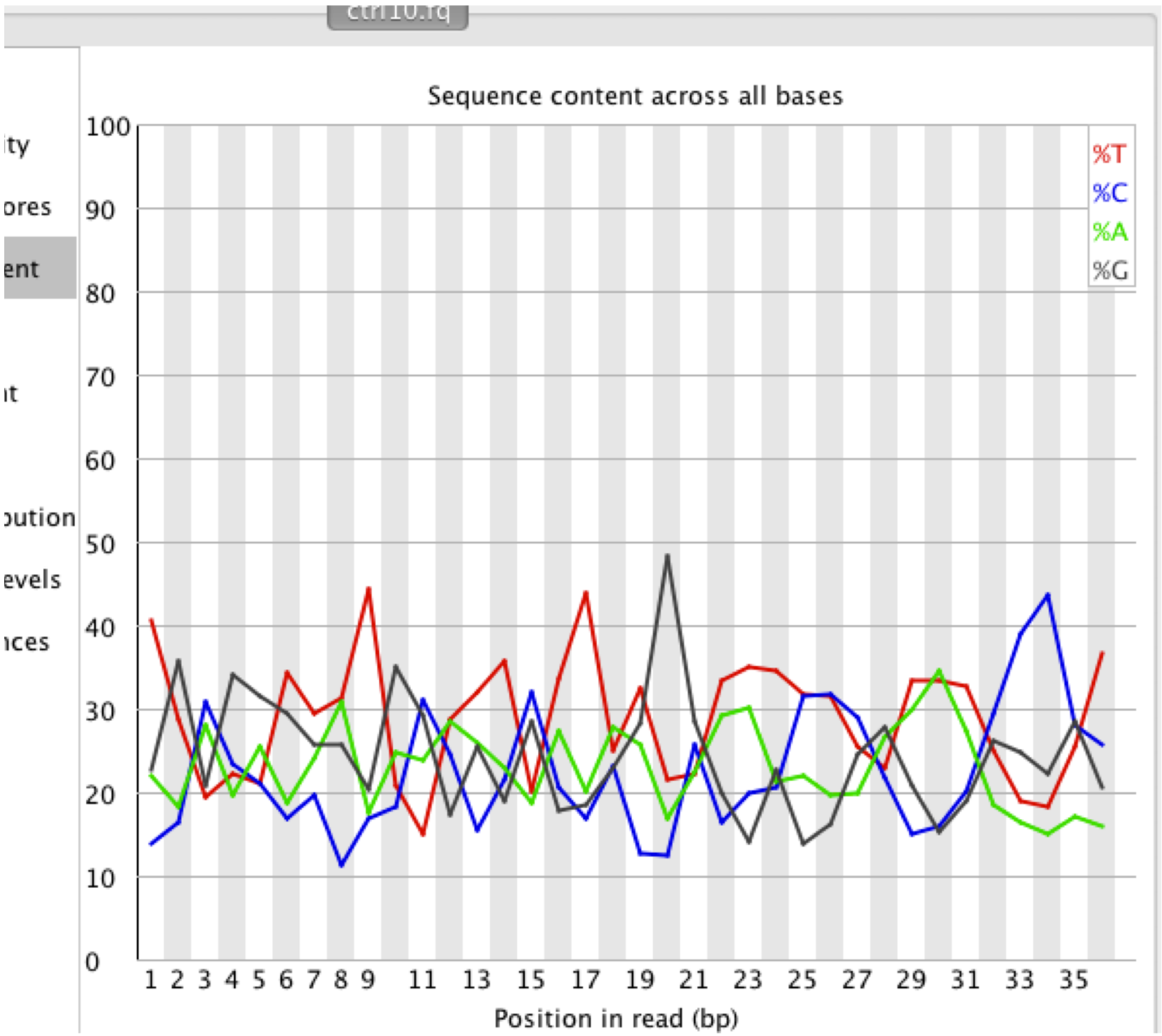
RNAseq

The random hexamer primers have been shown to cause mismatches in the begining of the Illumina RNA-seq redas.

The quality associated to these positions are good.

The first bases can be trimmed by a dedicated software

Inspecting raw data



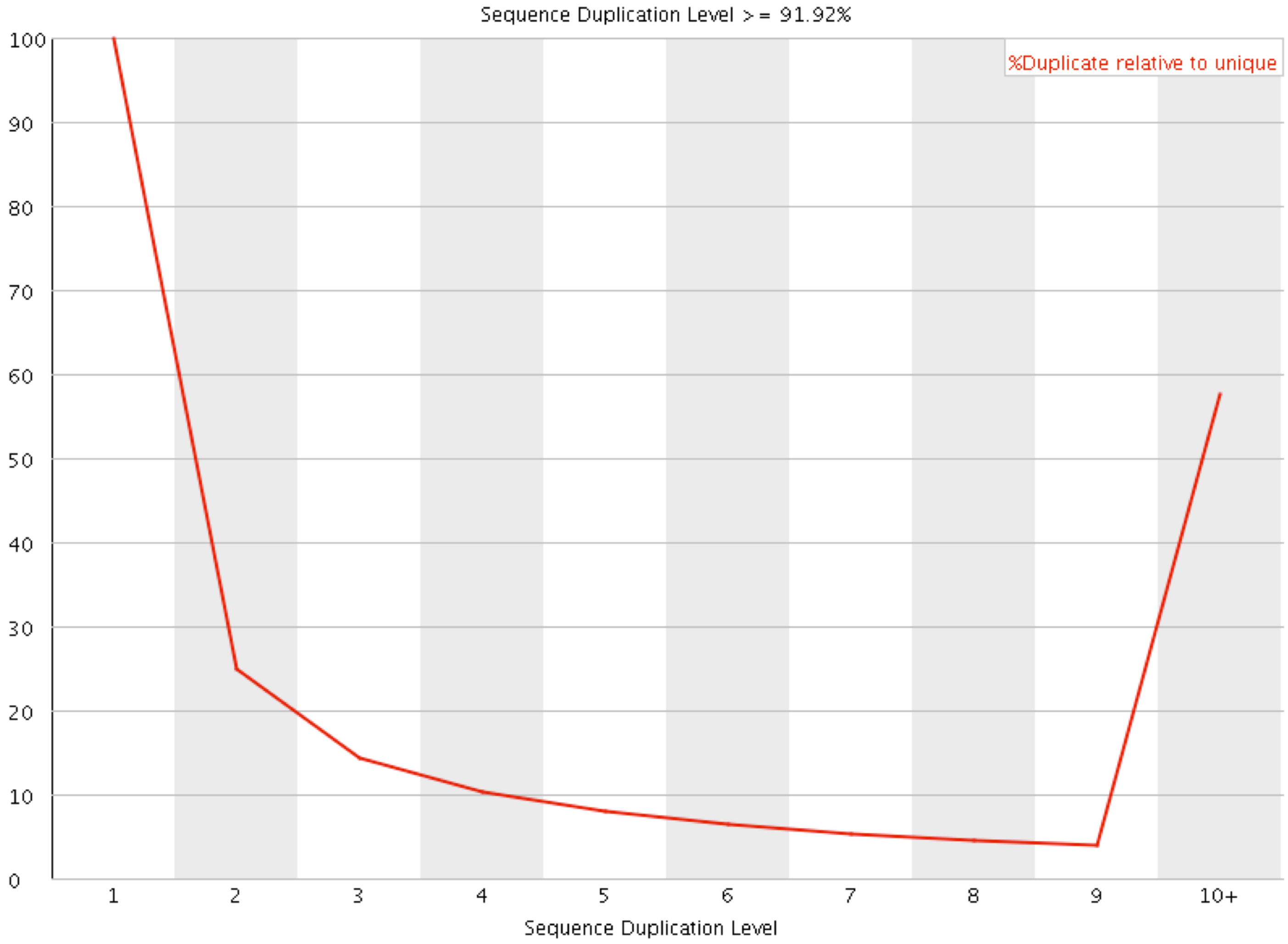
miRNAseq

FastQC plots base compositions along the reads which shuold produce flat line where the amount of each base resembles that of the organism.

If the difference between A and T or G and C is bigger than 10% at any read position, a warning is reported.

Inspecting raw data

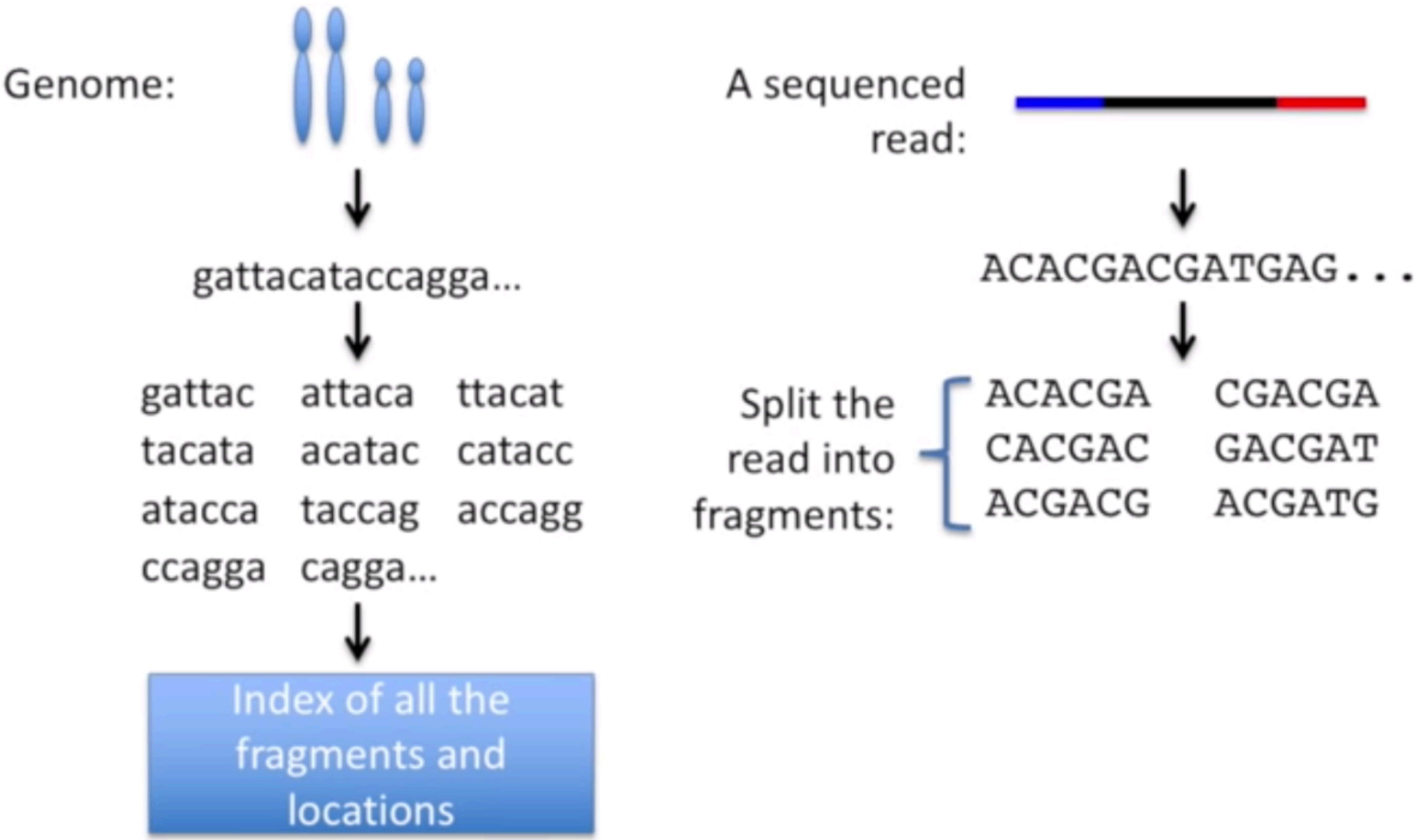
miRNAseq



Most of the reads should be unique.

High level of identical reads can indicate PCR overamplification but in the context of RNA-seq the duplicates are the natural conseguence of sequencing highly expressed transcripts.

Align the reads with respect to the genome sequence



Genome:

gattacataccagga...

gattac attaca ttacat
tacata acatac catacc
atacca taccag accagg
ccagga cagga...

Index of all the fragments and locations

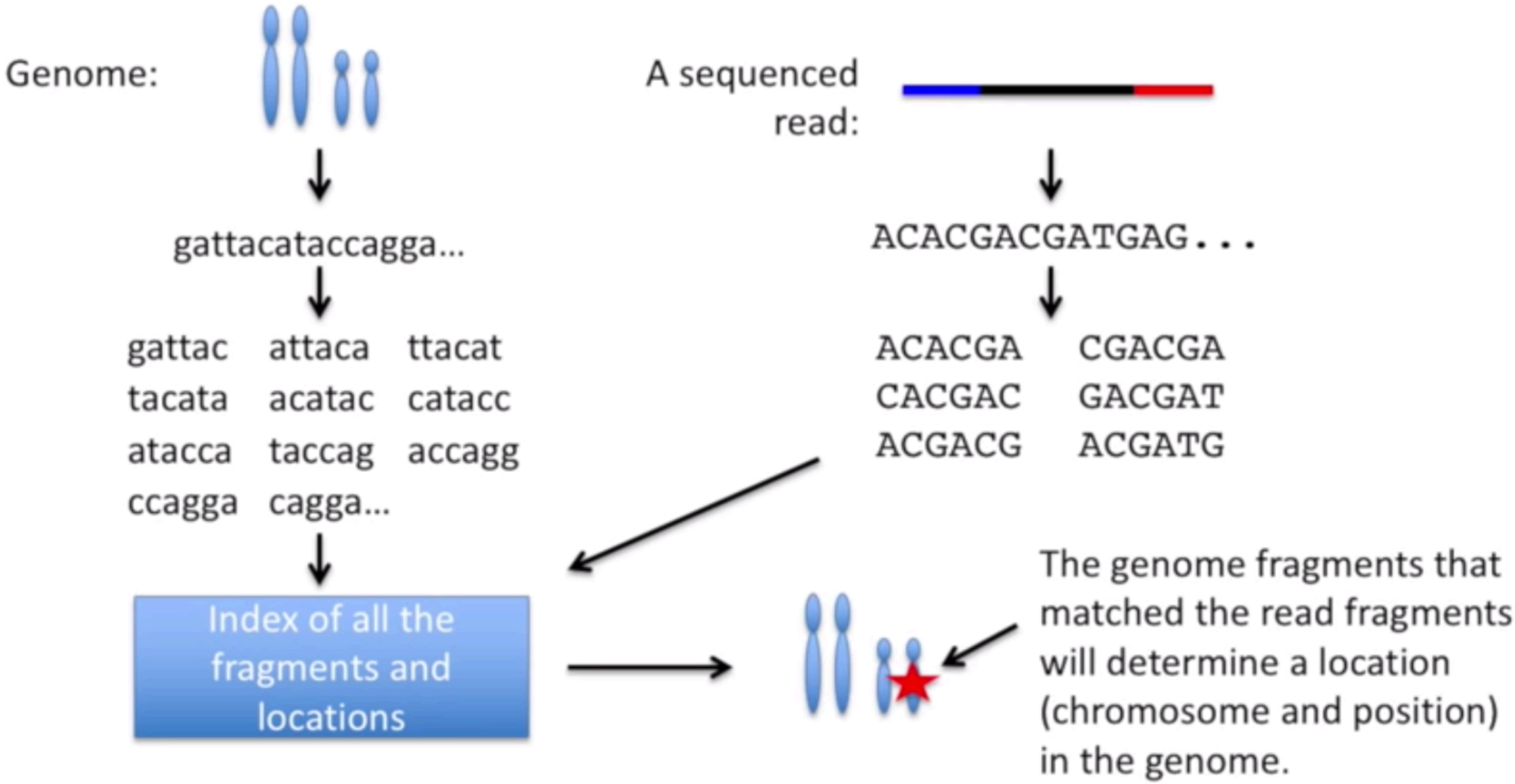
A sequenced read:

ACACGACGATGAG...

Split the read into fragments:

ACACGA CGACGA
CACGAC GACGAT
ACGACG ACGATG

Align the reads with respect to the genome sequence



Genome:

gattacataccagga...

gattac attaca ttacat
tacata acatac catacc
atacca taccag accagg
ccagga cagga...

Index of all the fragments and locations

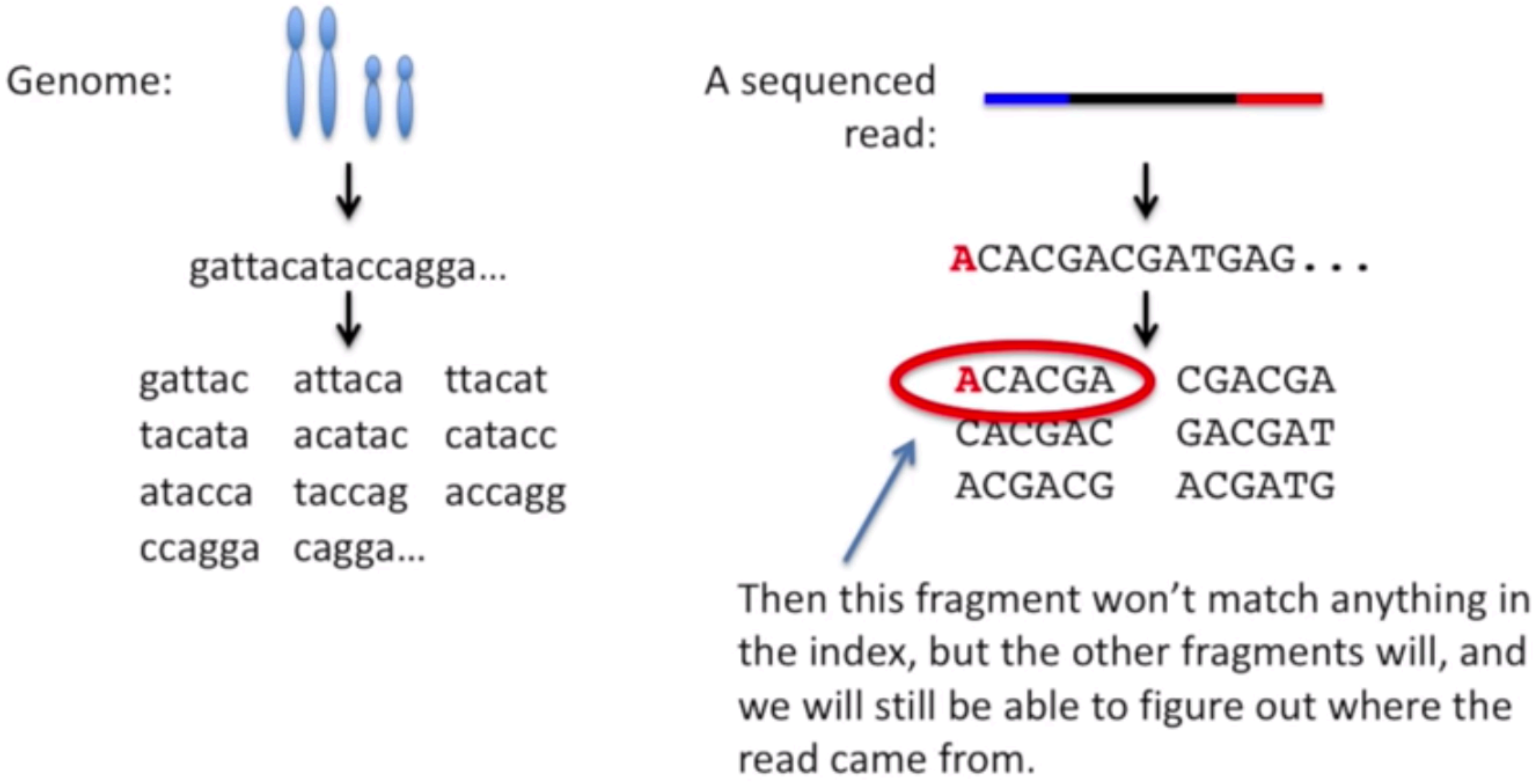
A sequenced read:

ACACGACGATGAG...

ACACGA CGACGA
CACGAC GACGAT
ACGACG ACGATG

The genome fragments that matched the read fragments will determine a location (chromosome and position) in the genome.

Align the reads with respect to the genome sequence



Genome:

gattacataccagga...

gattac attaca ttacat
tacata acatac catacc
atacca taccag accagg
ccagga cagga...

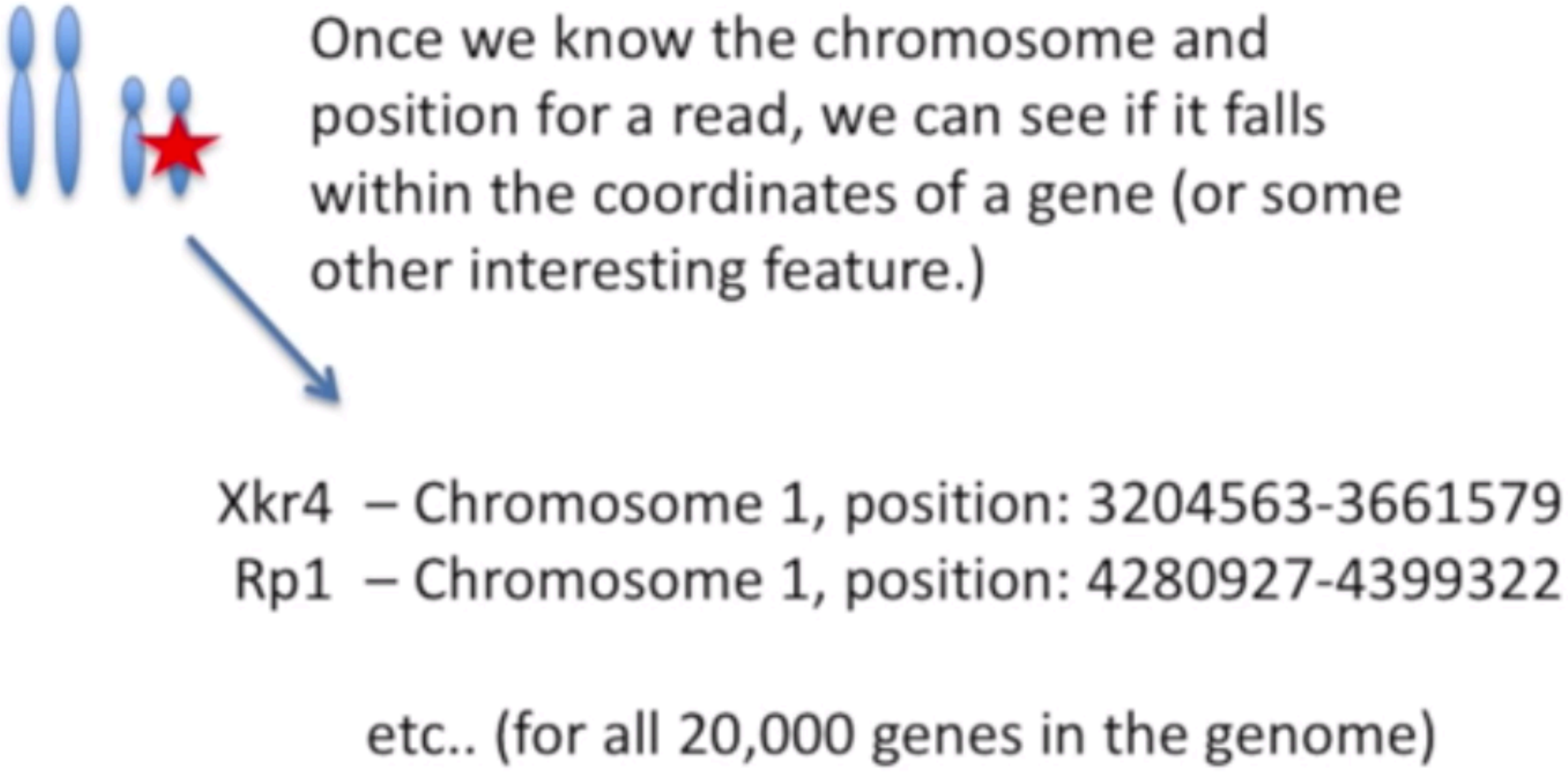
A sequenced read:

ACACGACGATGAG...

ACACGA CGACGA
CACGAC GACGAT
ACGACG ACGATG

Then this fragment won't match anything in the index, but the other fragments will, and we will still be able to figure out where the read came from.

Count the reads per gene

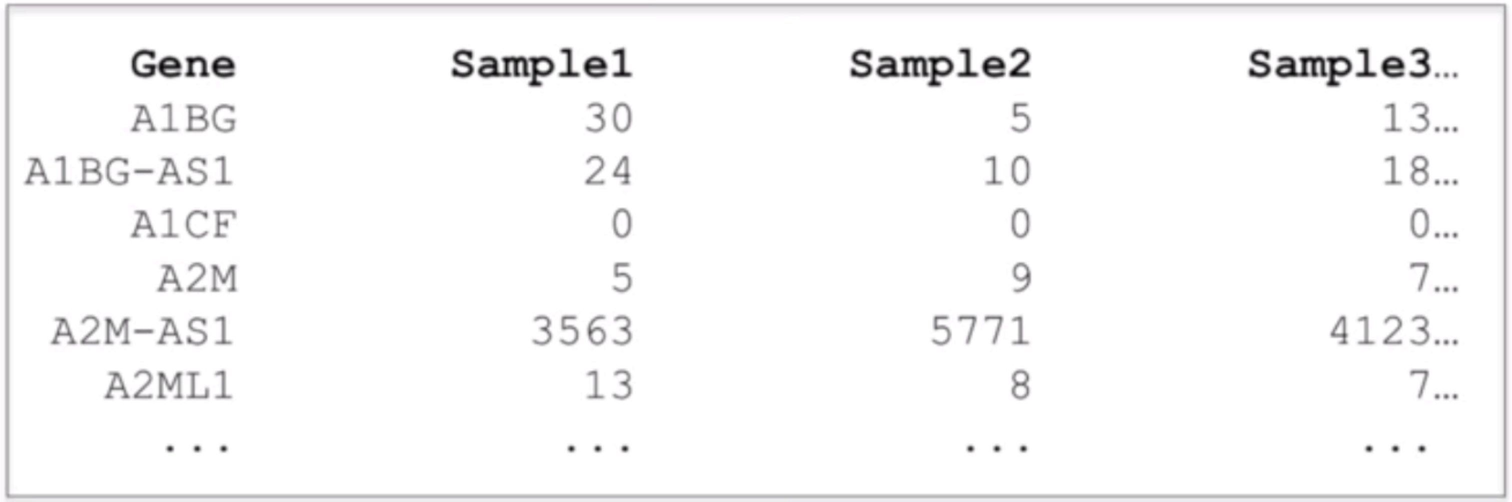


Once we know the chromosome and position for a read, we can see if it falls within the coordinates of a gene (or some other interesting feature.)

Xkr4	– Chromosome 1, position: 3204563-3661579
Rp1	– Chromosome 1, position: 4280927-4399322

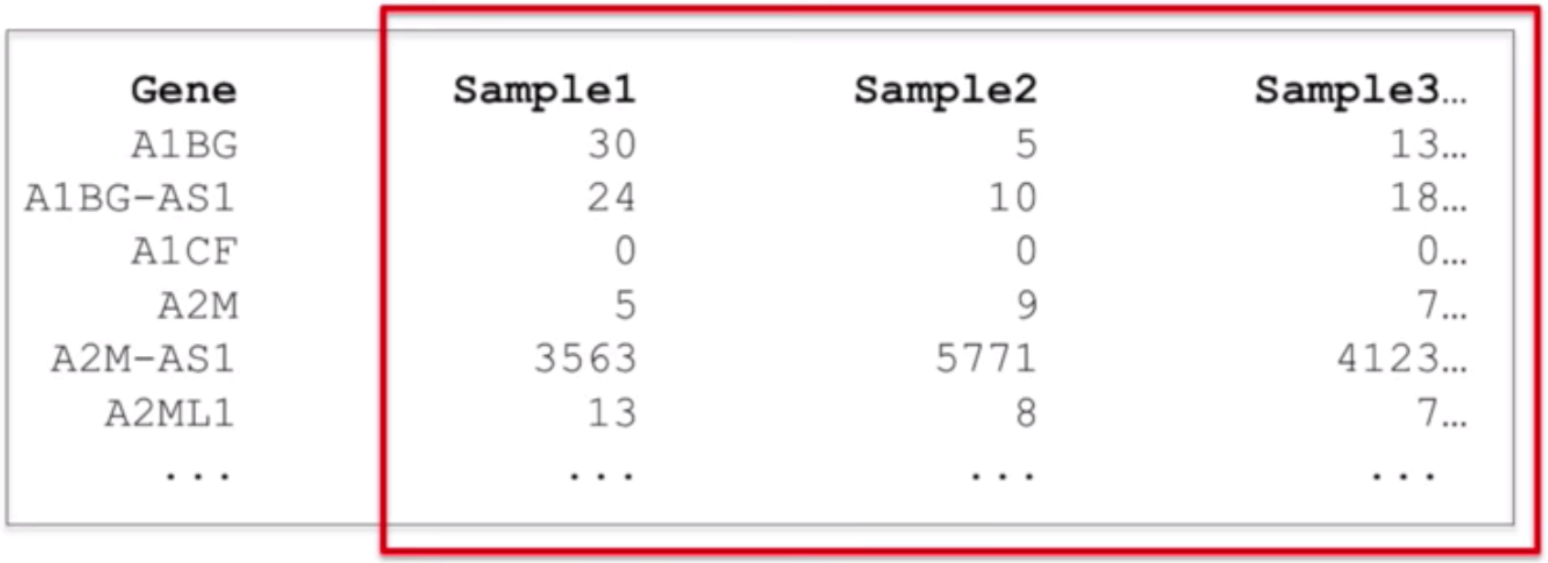
etc.. (for all 20,000 genes in the genome)

Count the reads per gene



Gene	Sample1	Sample2	Sample3...
A1BG	30	5	13...
A1BG-AS1	24	10	18...
A1CF	0	0	0...
A2M	5	9	7...
A2M-AS1	3563	5771	4123...
A2ML1	13	8	7...
...	...	...	...

After you count the reads per gene, you end up with a matrix of numbers like this...

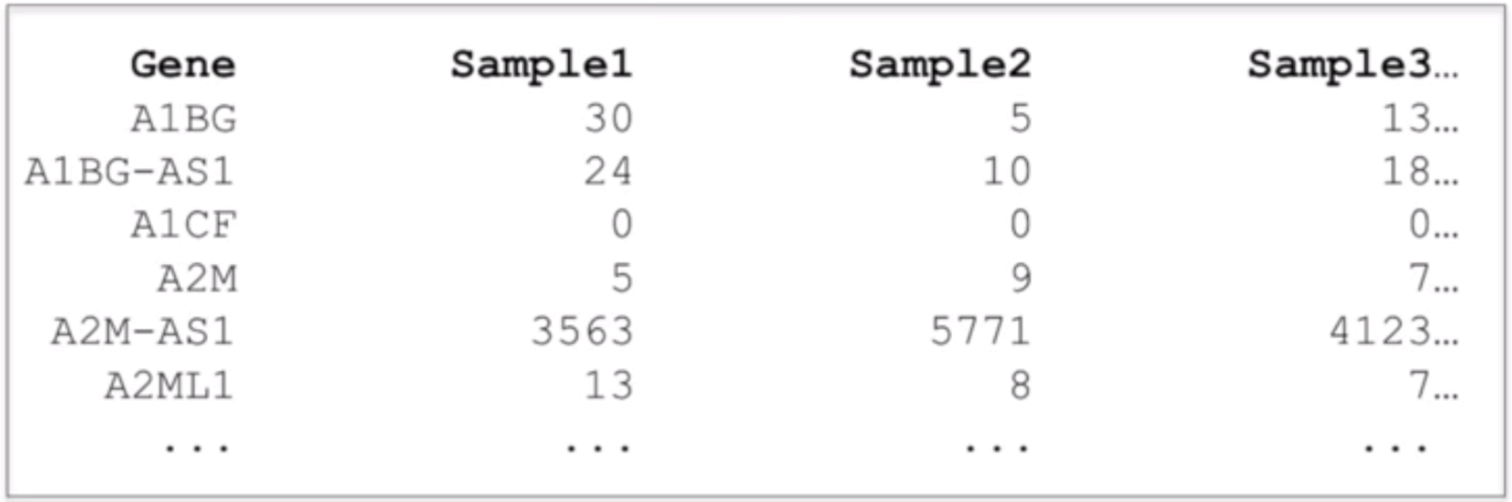


Gene	Sample1	Sample2	Sample3...
A1BG	30	5	13...
A1BG-AS1	24	10	18...
A1CF	0	0	0...
A2M	5	9	7...
A2M-AS1	3563	5771	4123...
A2ML1	13	8	7...
...	...	...	...

"Bulk" RNA-seq, where a "sample" is the average of a pool of cells (usually 6 million cells), might have 3 "normal" samples and 3 "disease state" samples, or 6 total.

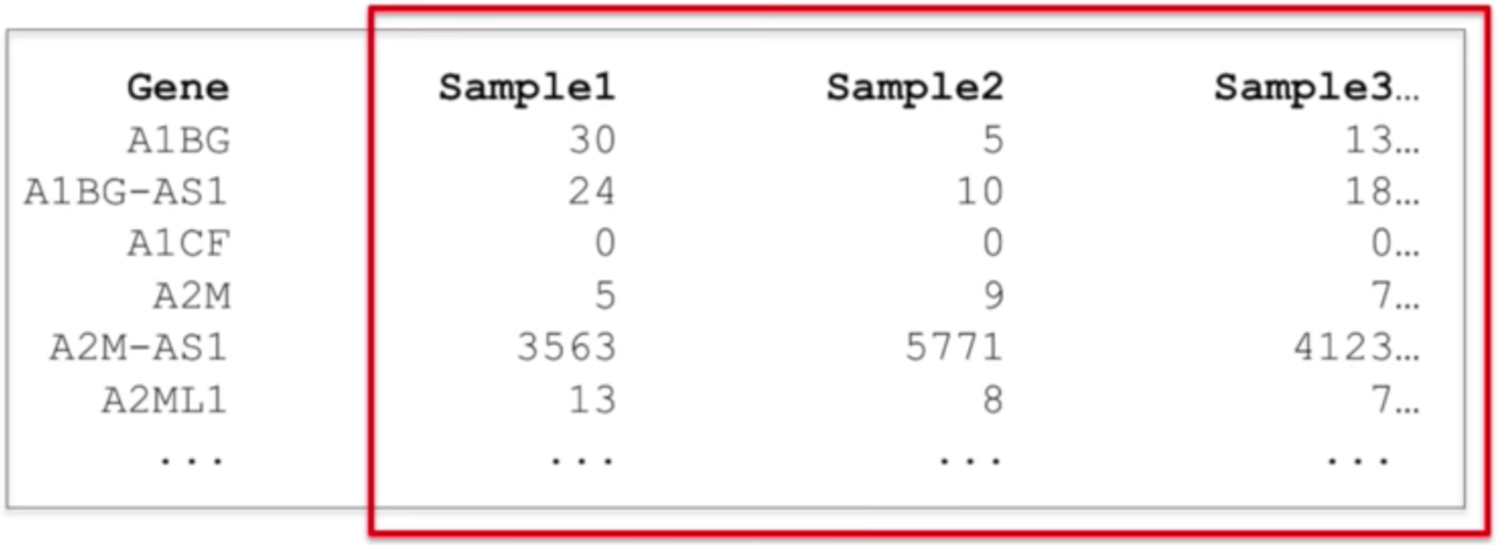
There are usually between 6 and 800+ samples.

Count the reads per gene



Gene	Sample1	Sample2	Sample3…
A1BG	30	5	13…
A1BG-AS1	24	10	18…
A1CF	0	0	0…
A2M	5	9	7…
A2M-AS1	3563	5771	4123…
A2ML1	13	8	7…
...	...	...	...

After you count the reads per gene, you end up with a matrix of numbers like this…



Gene	Sample1	Sample2	Sample3…
A1BG	30	5	13…
A1BG-AS1	24	10	18…
A1CF	0	0	0…
A2M	5	9	7…
A2M-AS1	3563	5771	4123…
A2ML1	13	8	7…
...	...	...	...

"Single-cell" RNA-seq treats each cell like an individual sample, so it can generate a lot of samples.

There are usually between 6 and 800+ samples.

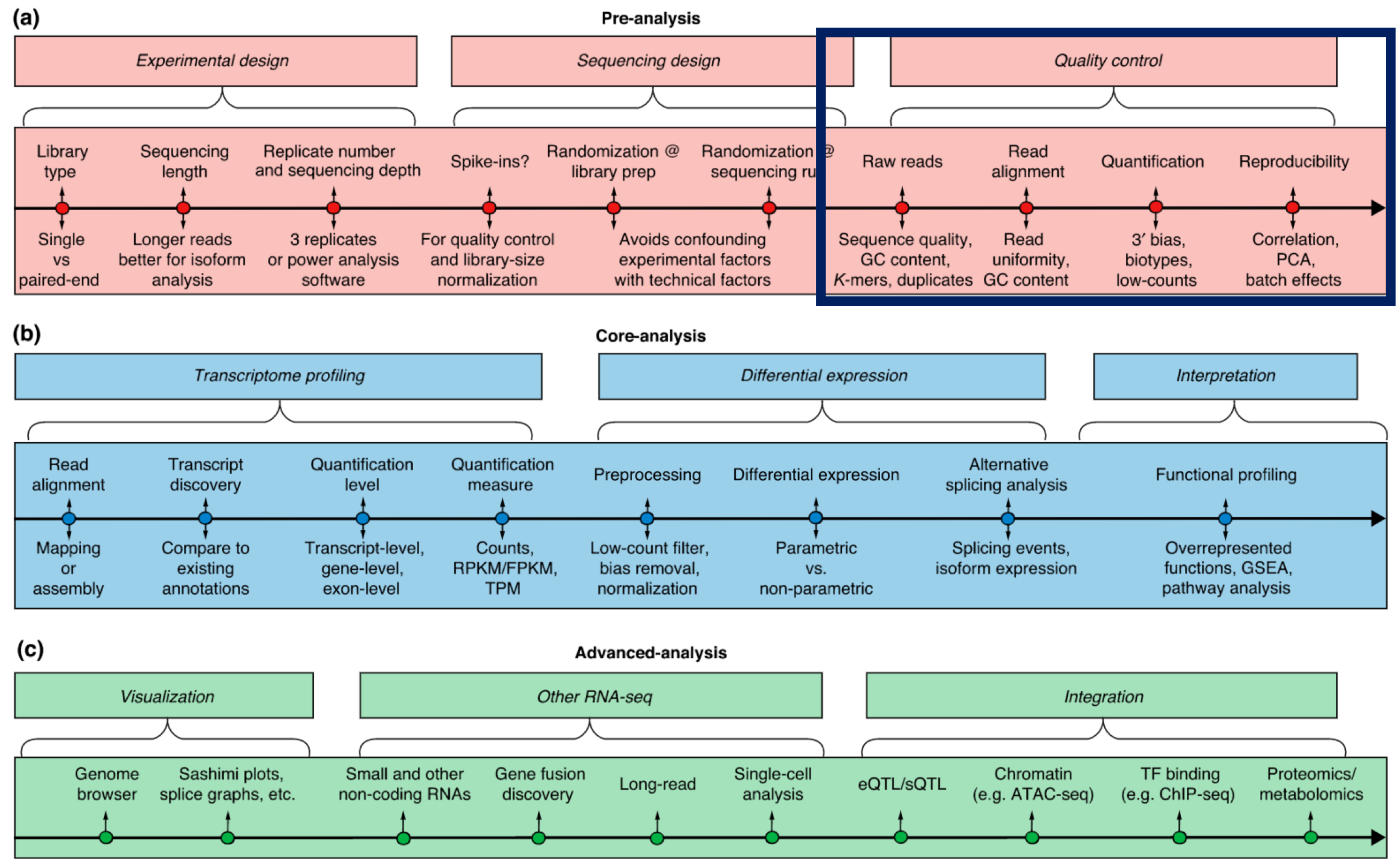
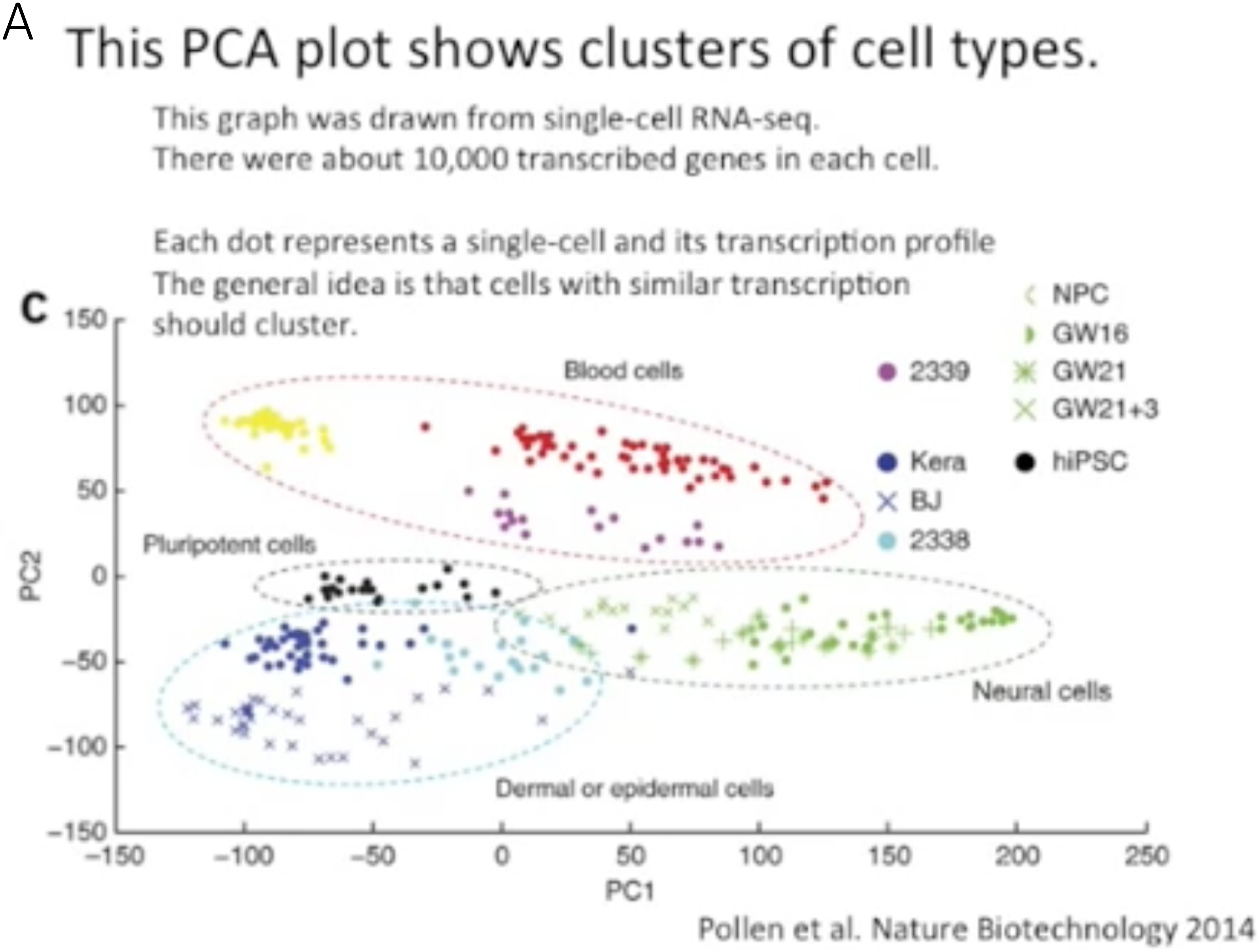

(a)
Pre-analysis
Experimental design
Sequencing design
Quality control
Library type
Single vs paired-end
Sequencing length
Longer reads better for isoform analysis
Replicate number and sequencing depth
3 replicates or power analysis software
Spike-ins?
For quality control and library-size normalization
Randomization @ library prep
Avoids confounding experimental factors with technical factors
Randomization @ sequencing ru
Raw reads
Sequence quality, GC content, K-mers, duplicates
Read alignment
Read uniformity, GC content
Quantification
3′ bias, biotypes, low-counts
Reproducibility
Correlation, PCA, batch effects
(b)
Core-analysis
Transcriptome profiling
Differential expression
Interpretation
Read alignment
Mapping or assembly
Transcript discovery
Compare to existing annotations
Quantification level
Transcript-level, gene-level, exon-level
Quantification measure
Counts, RPKM/FPKM, TPM
Preprocessing
Low-count filter, bias removal, normalization
Differential expression
Parametric vs. non-parametric
Alternative splicing analysis
Splicing events, isoform expression
Functional profiling
Overrepresented functions, GSEA, pathway analysis
(c)
Advanced-analysis
Visualization
Other RNA-seq
Integration
Genome browser
Sashimi plots, splice graphs, etc.
Small and other non-coding RNAs
Gene fusion discovery
Long-read
Single-cell analysis
eQTL/sQTL
Chromatin (e.g. ATAC-seq)
TF binding (e.g. ChIP-seq)
Proteomics/ metabolomics

Fig. 1 A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

Reproducibility - PCA



This PCA plot shows clusters of cell types.

This graph was drawn from single-cell RNA-seq.
There were about 10,000 transcribed genes in each cell.

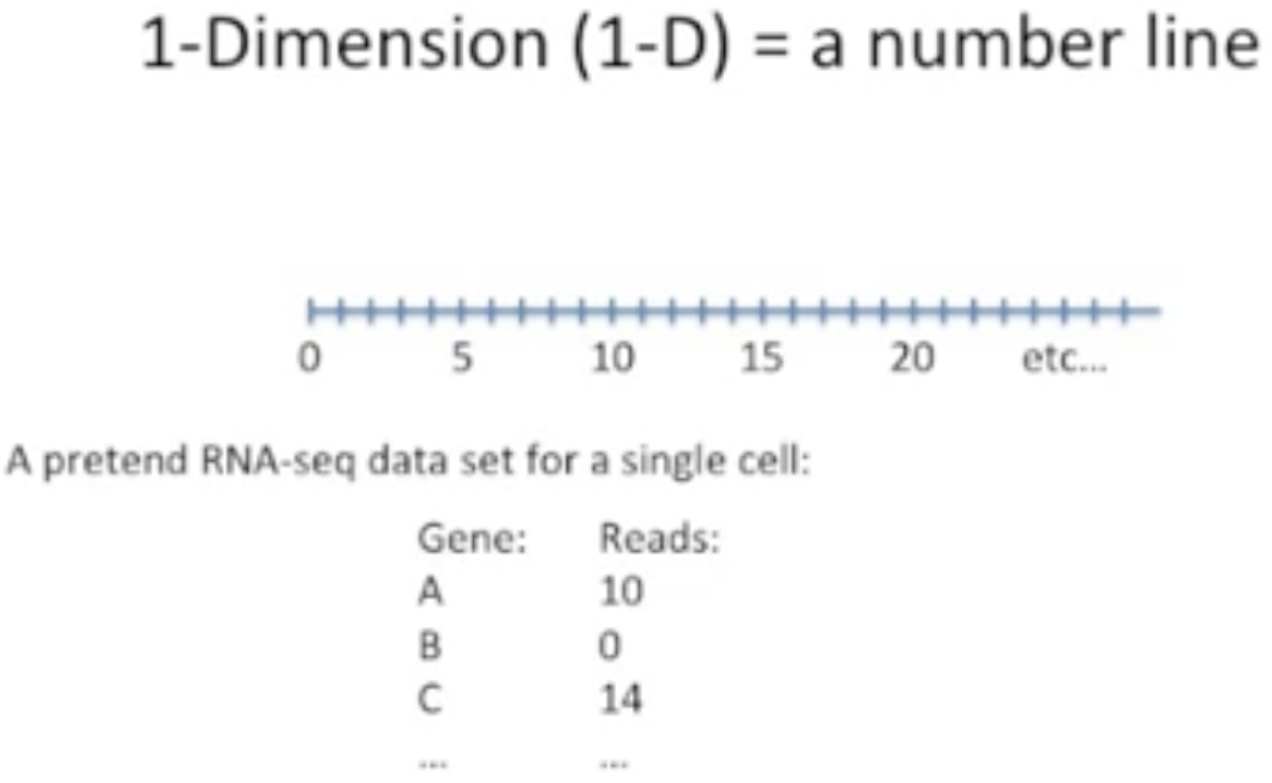
Each dot represents a single-cell and its transcription profile
The general idea is that cells with similar transcription should cluster.

Pollen et al. Nature Biotechnology 2014

Reproducibility - PCA

How does transcription from 10,000 genes get compressed to a single dot on a graph?

PCA is a method for compressinf a loto fo data into somenthing that captures the essence of the original data.

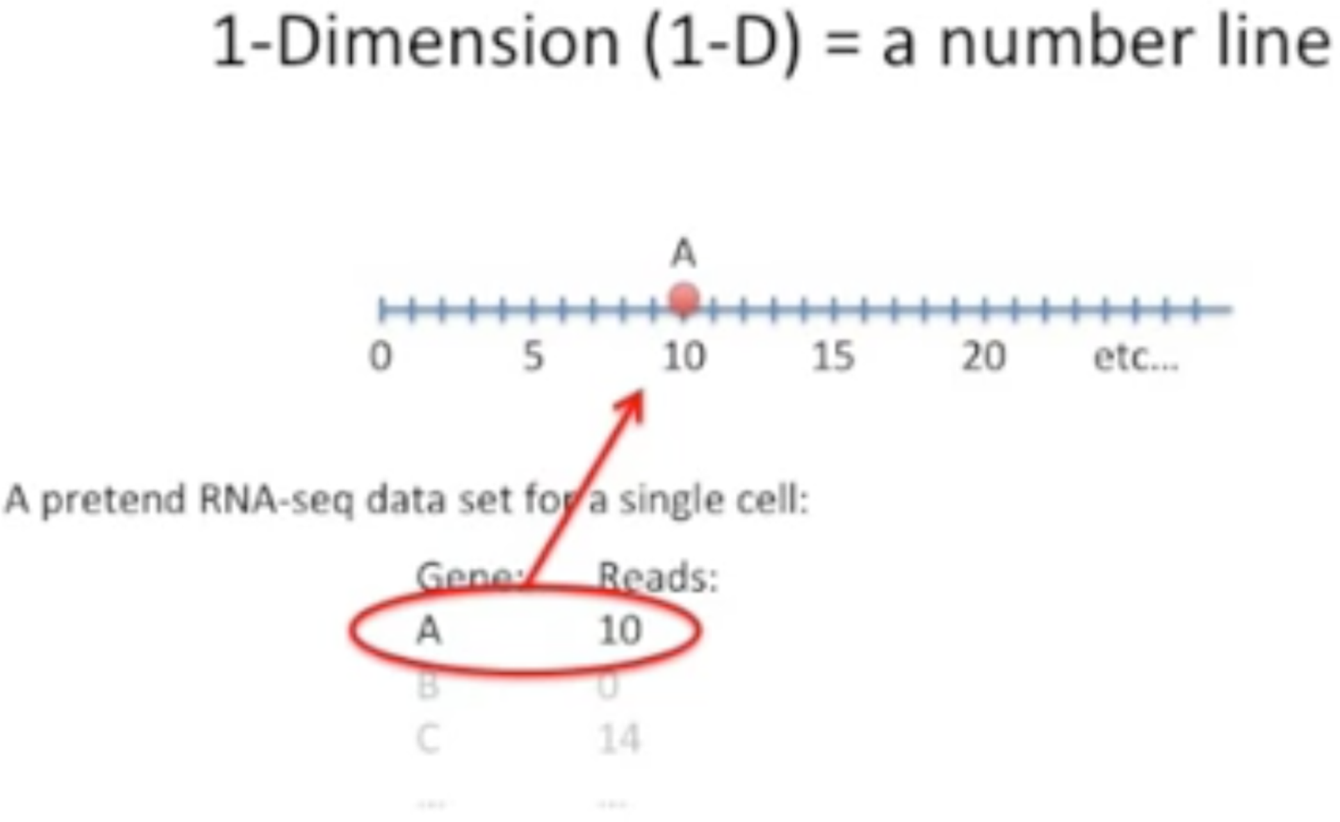


1-Dimension (1-D) = a number line

A pretend RNA-seq data set for a single cell:

Gene:	Reads:
A	10
B	0
C	14
...	...

Reproducibility - PCA

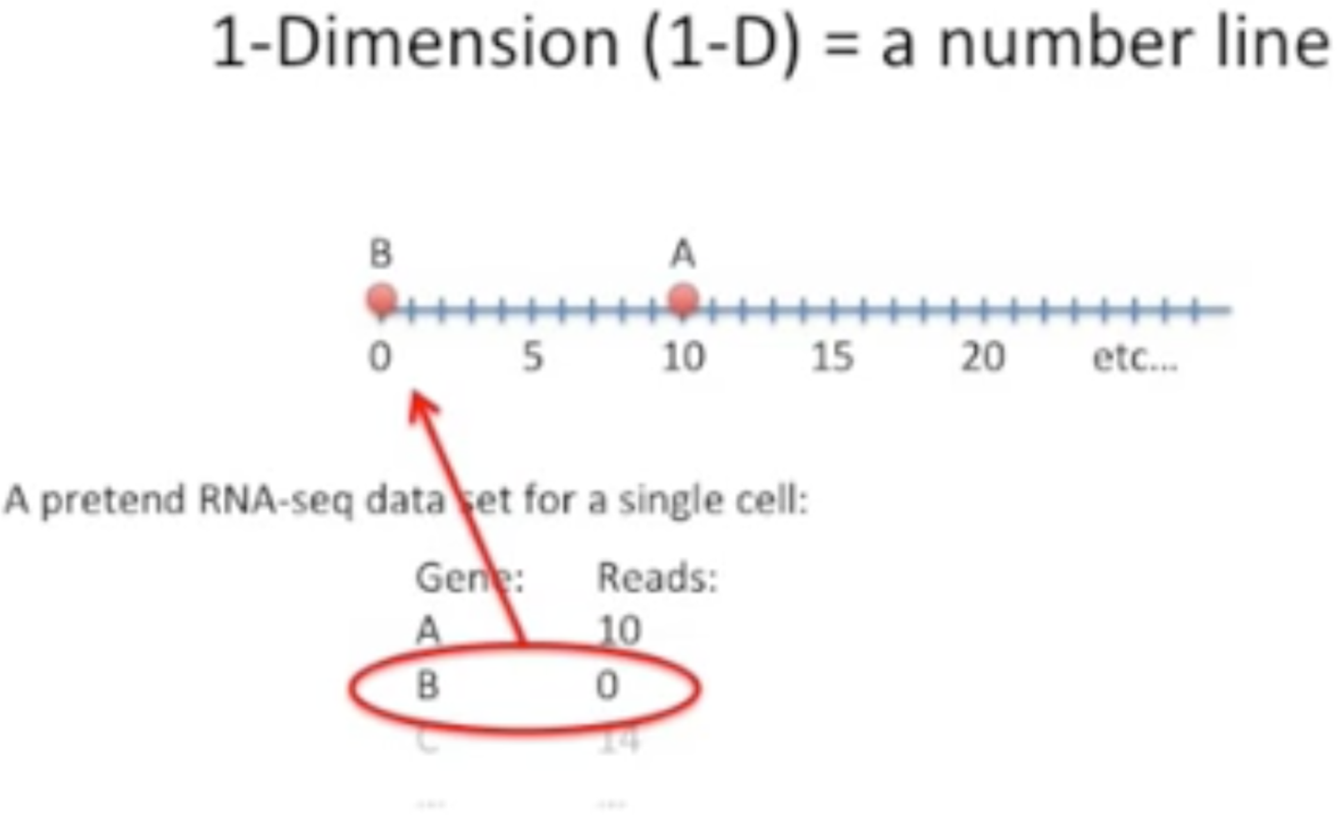


1-Dimension (1-D) = a number line

A pretend RNA-seq data set for a single cell:

Gene:	Reads:
A	10
B	0
C	14
...	...

Reproducibility - PCA

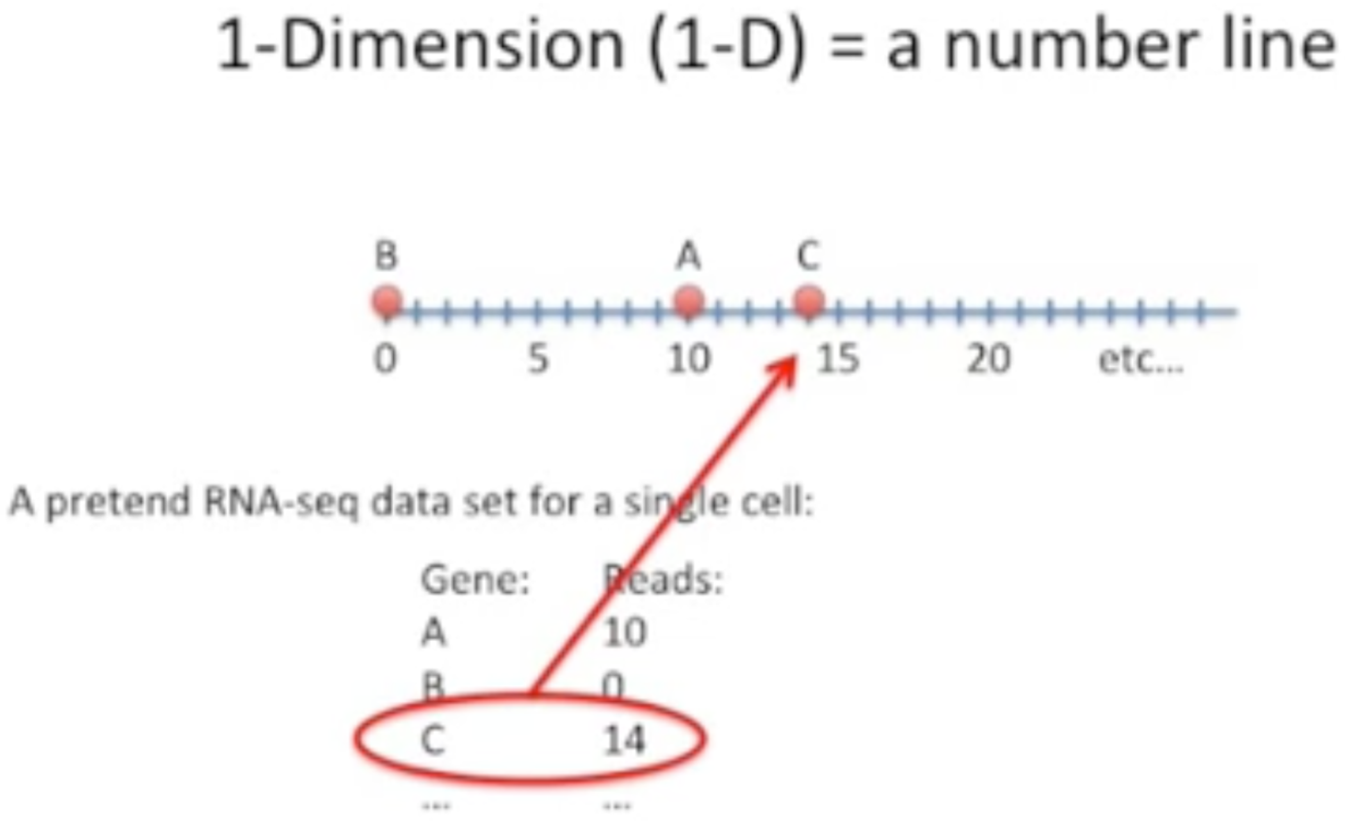


1-Dimension (1-D) = a number line

A pretend RNA-seq data set for a single cell:

Gene:	Reads:
A	10
B	0
C	14
...	...

Reproducibility - PCA

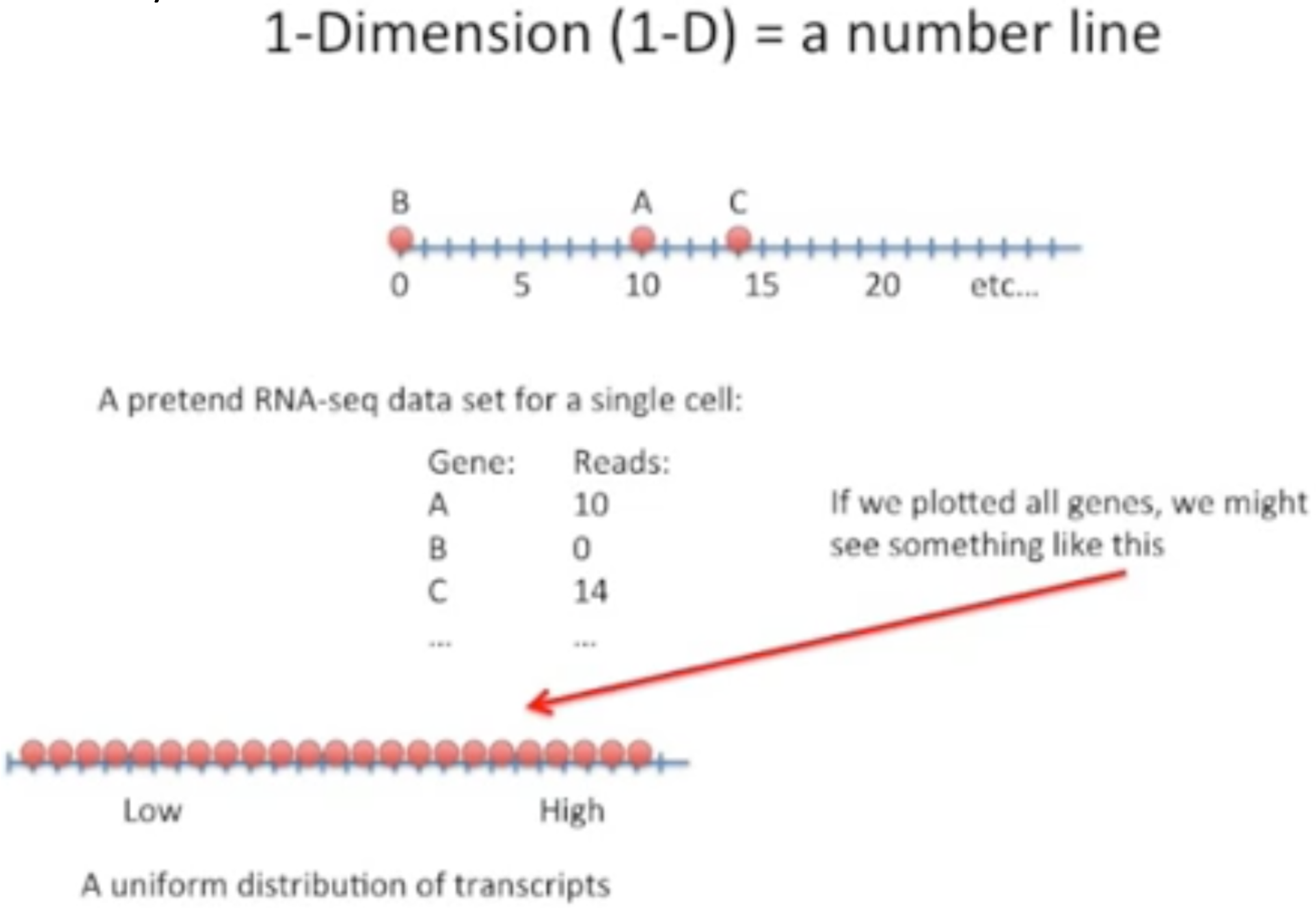


1-Dimension (1-D) = a number line

A pretend RNA-seq data set for a single cell:

Gene:	Reads:
A	10
B	0
C	14
...	...

Reproducibility - PCA



1-Dimension (1-D) = a number line

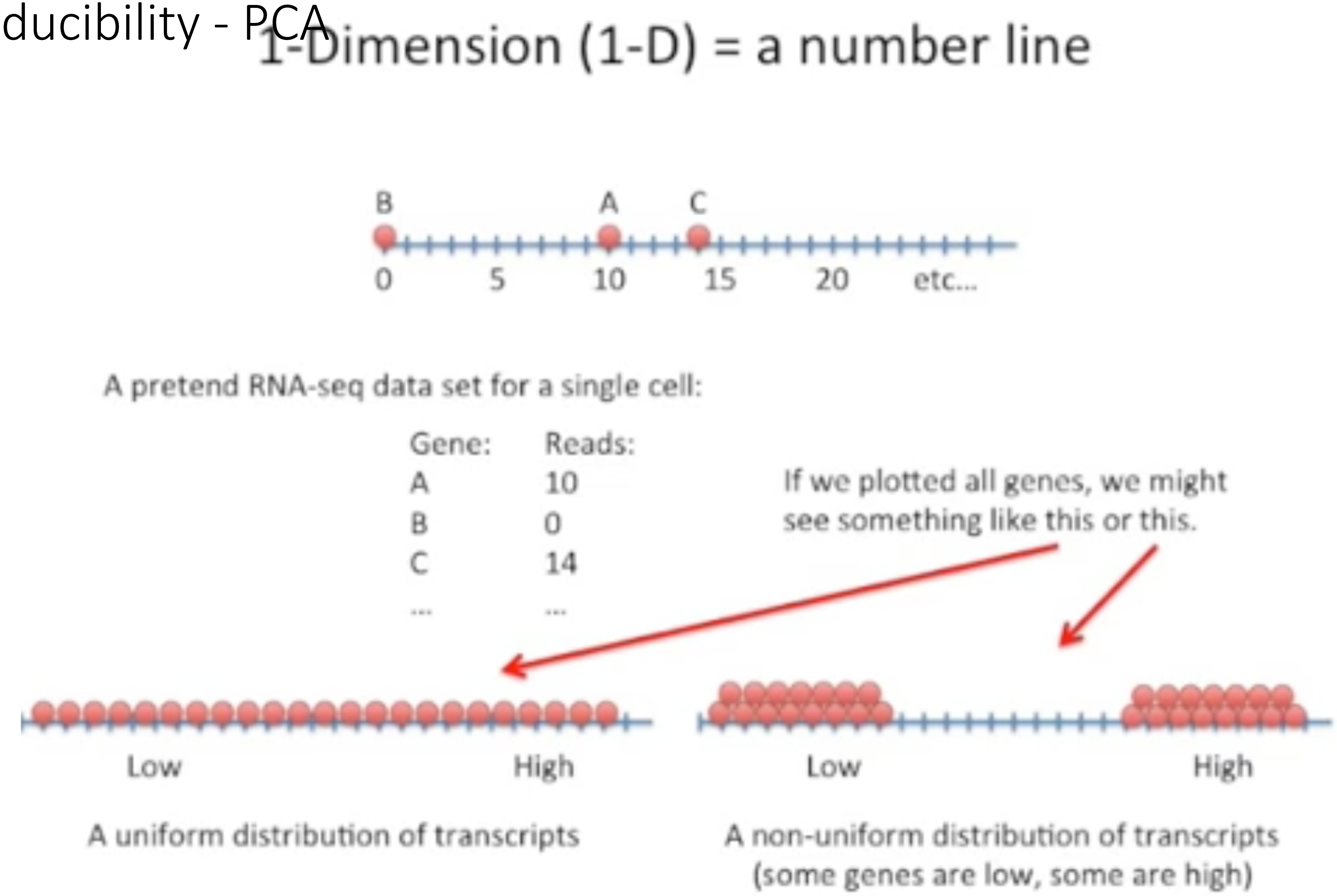
A pretend RNA-seq data set for a single cell:

Gene:	Reads:
A	10
B	0
C	14
...	...

If we plotted all genes, we might see something like this

A uniform distribution of transcripts

Reproducibility - PCA



1-Dimension (1-D) = a number line

A pretend RNA-seq data set for a single cell:

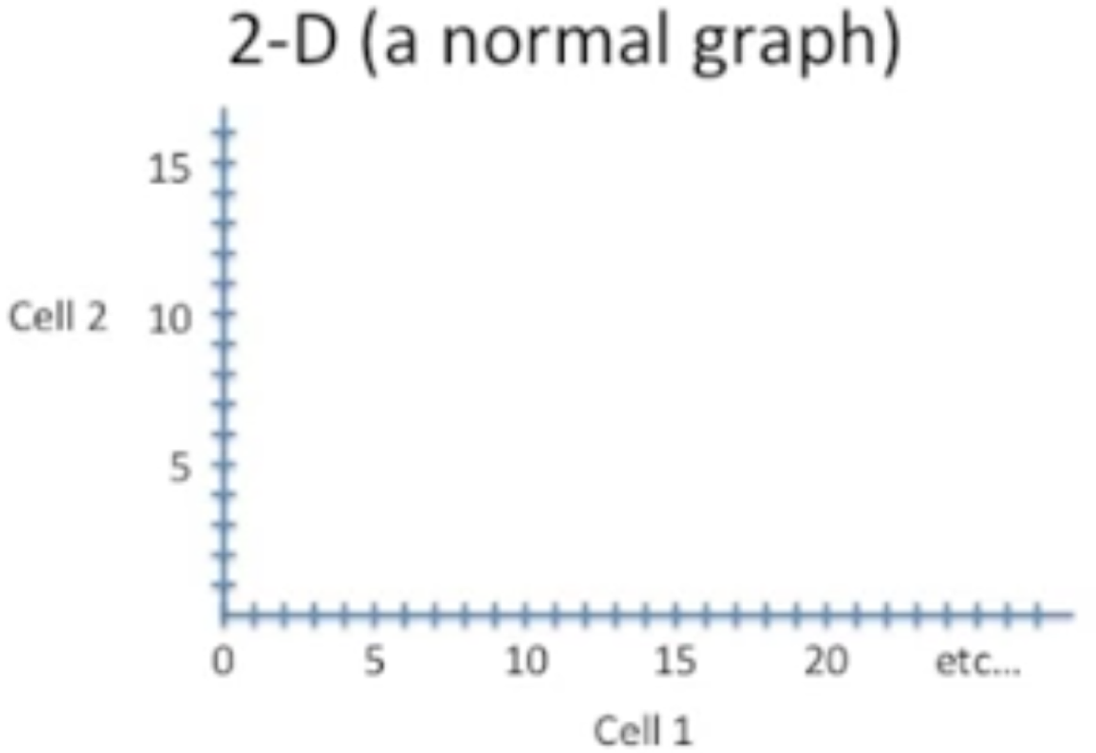
Gene:	Reads:
A	10
B	0
C	14
...	...

If we plotted all genes, we might see something like this or this.

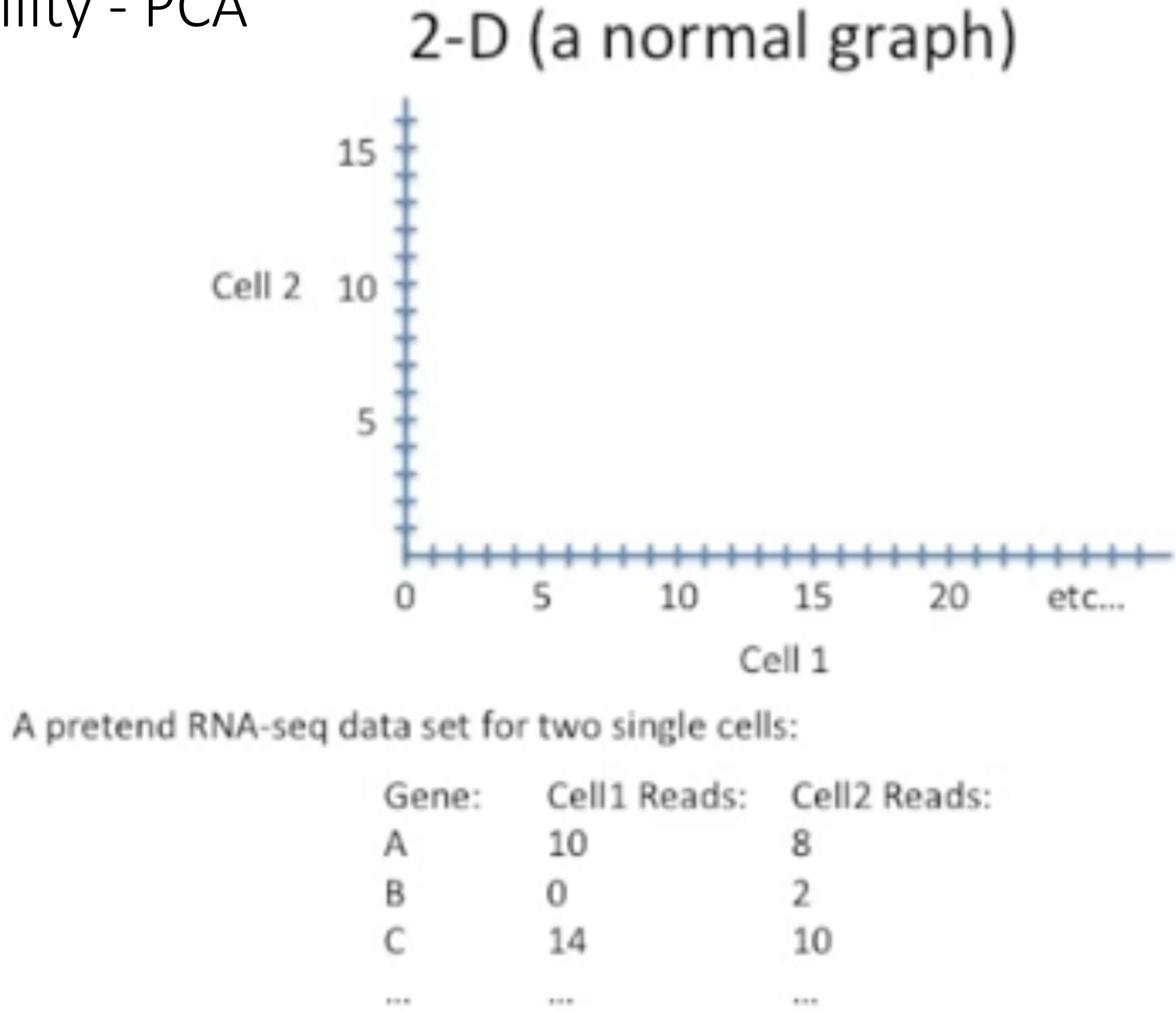
A uniform distribution of transcripts

A non-uniform distribution of transcripts (some genes are low, some are high)

Reproducibility - PCA



Reproducibility - PCA

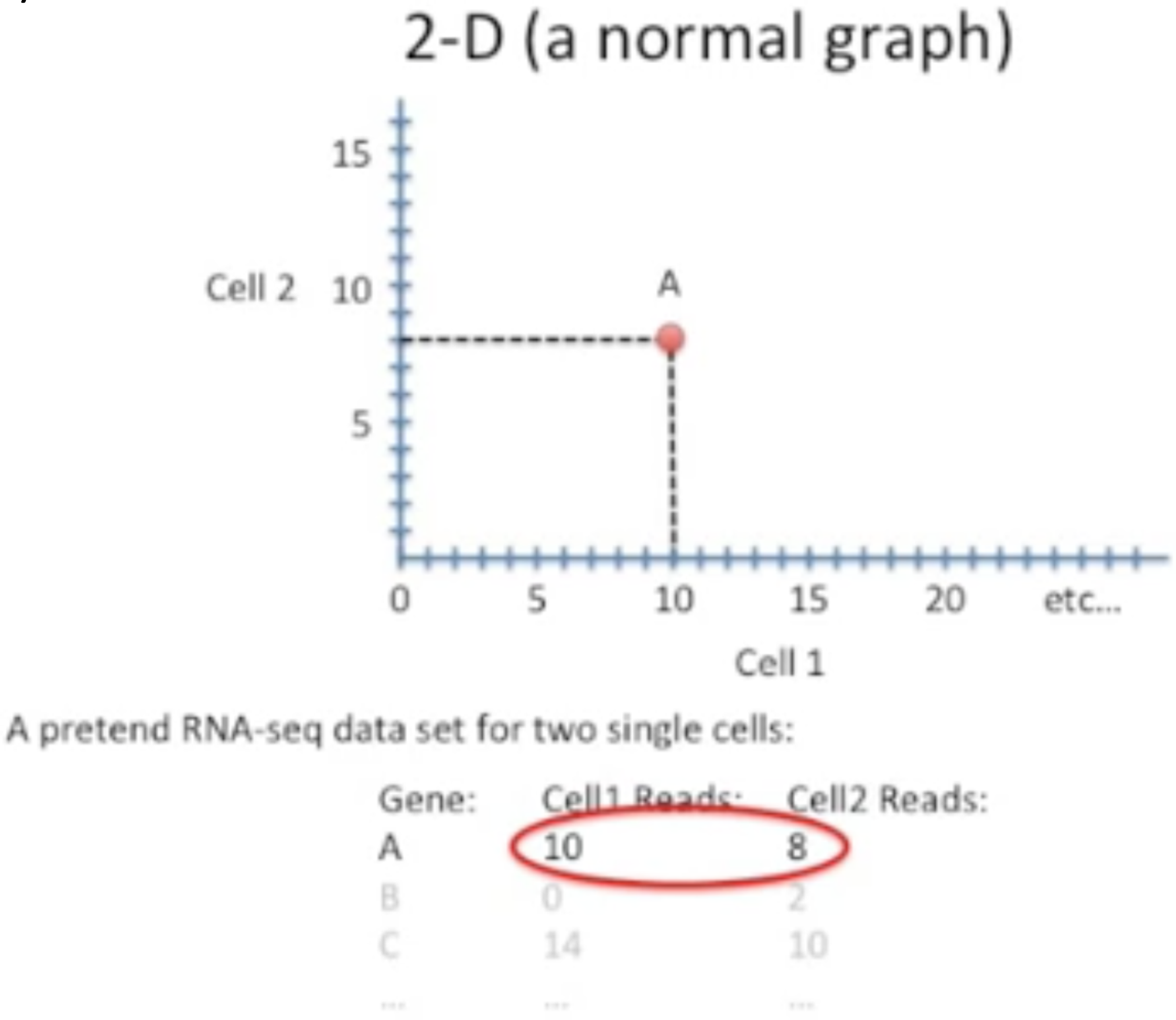


2-D (a normal graph)

A pretend RNA-seq data set for two single cells:

Gene:	Cell1 Reads:	Cell2 Reads:
A	10	8
B	0	2
C	14	10
...	...	...

Reproducibility - PCA

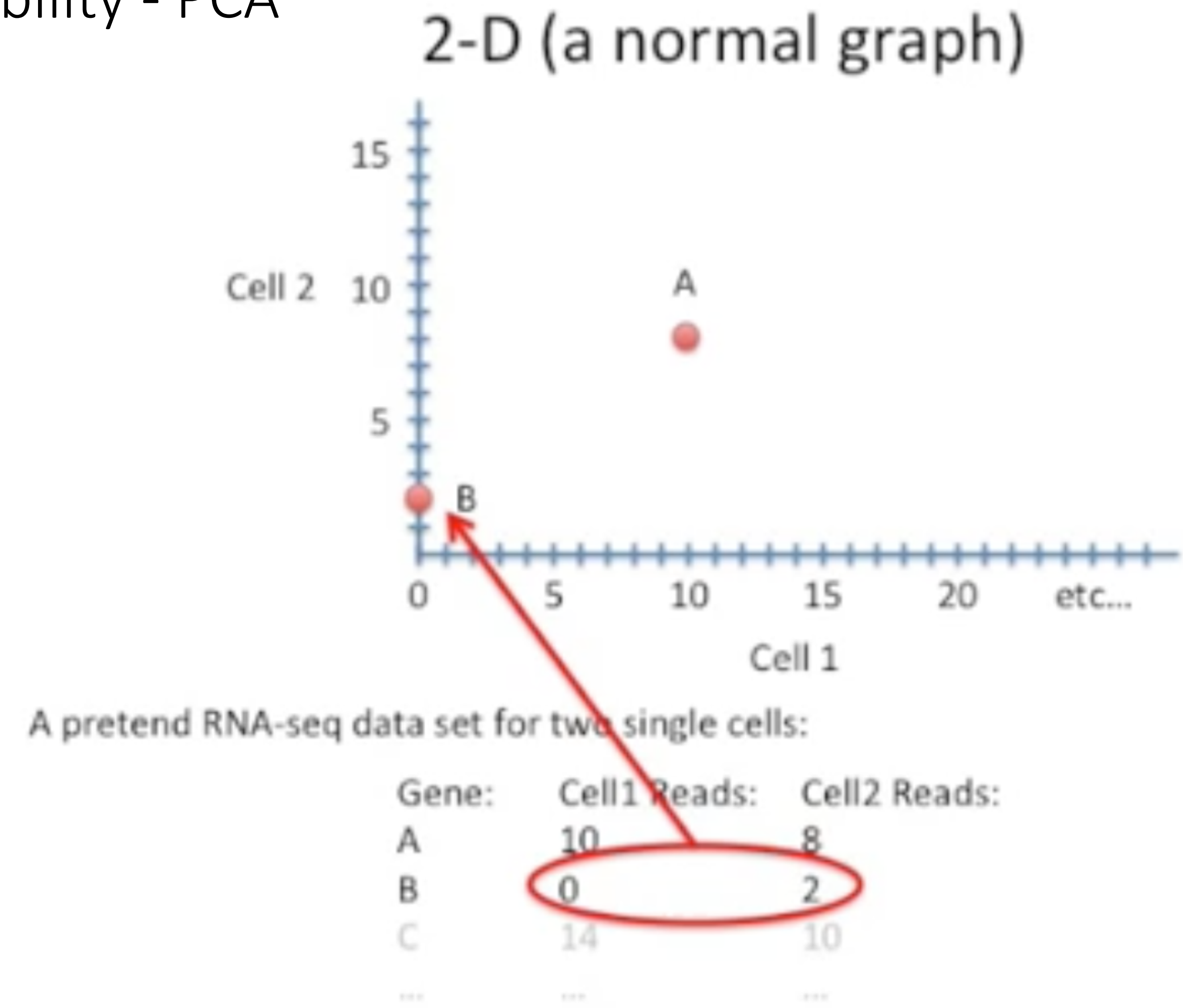


2-D (a normal graph)

A pretend RNA-seq data set for two single cells:

Gene:	Cell1 Reads:	Cell2 Reads:
A	10	8
B	0	2
C	14	10
...	...	...

Reproducibility - PCA

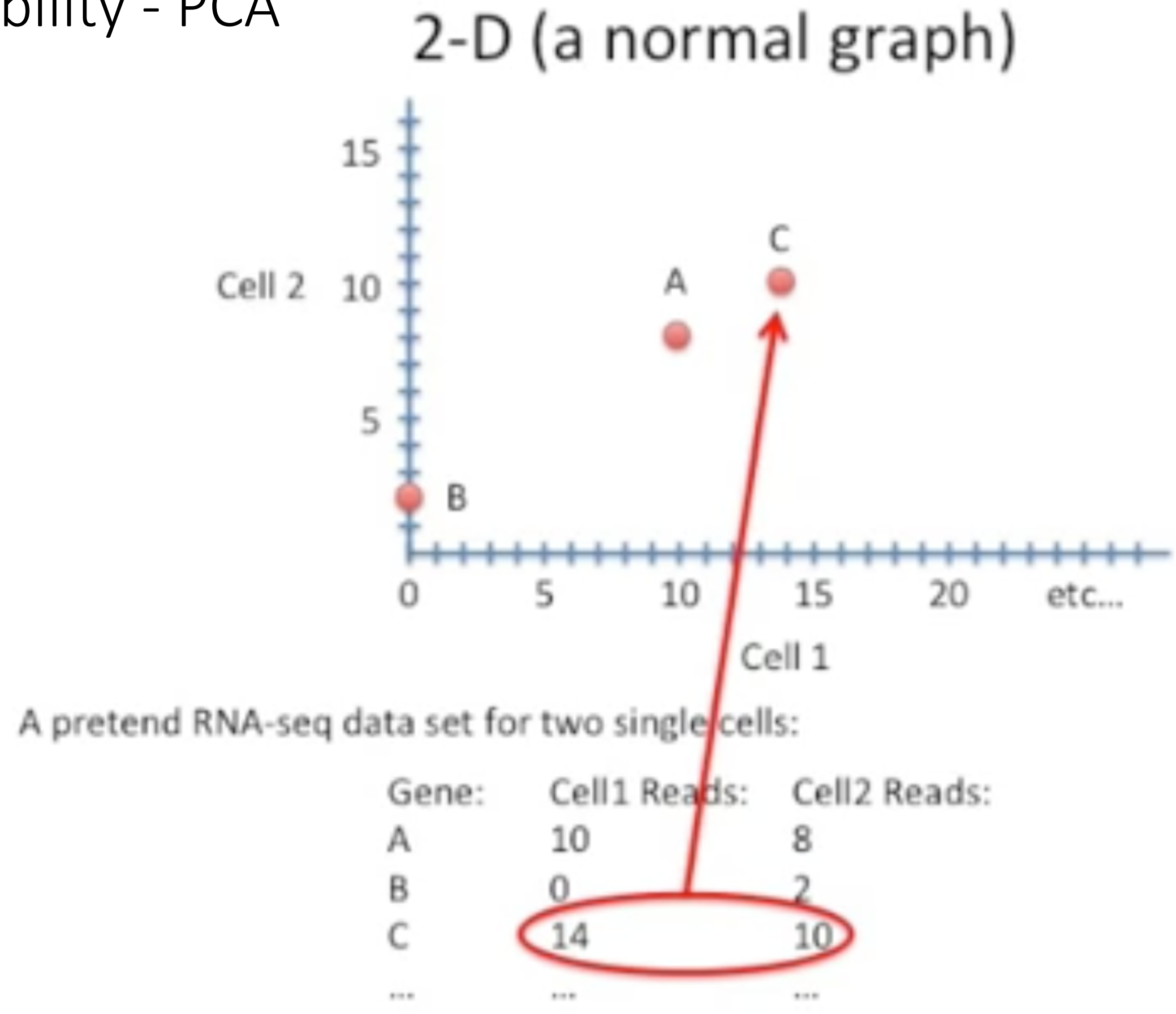


2-D (a normal graph)

A pretend RNA-seq data set for two single cells:

Gene:	Cell1 Reads:	Cell2 Reads:
A	10	8
B	0	2
C	14	10
...	...	...

Reproducibility - PCA

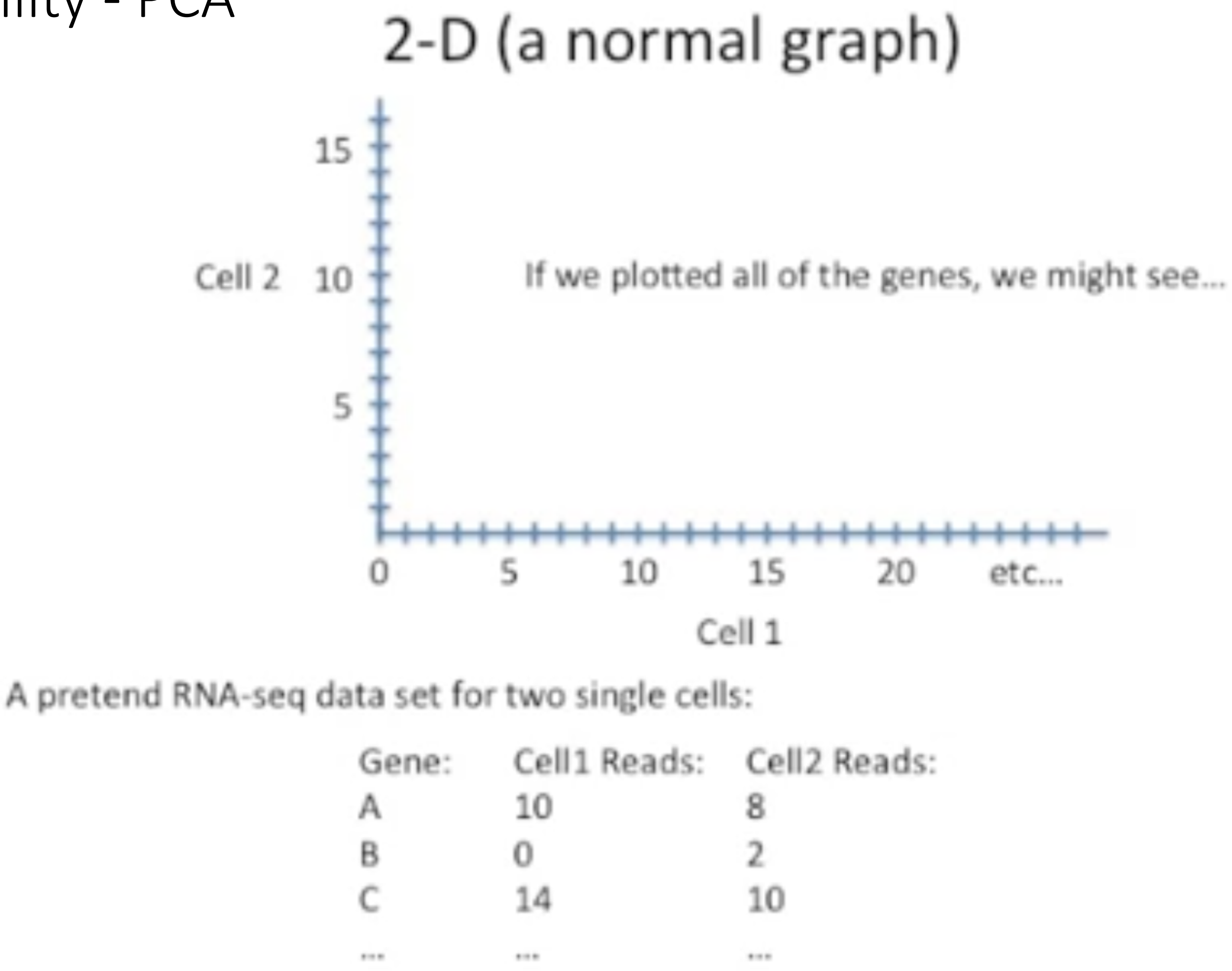


2-D (a normal graph)

A pretend RNA-seq data set for two single cells:

Gene:	Cell1 Reads:	Cell2 Reads:
A	10	8
B	0	2
C	14	10
...	...	...

Reproducibility - PCA



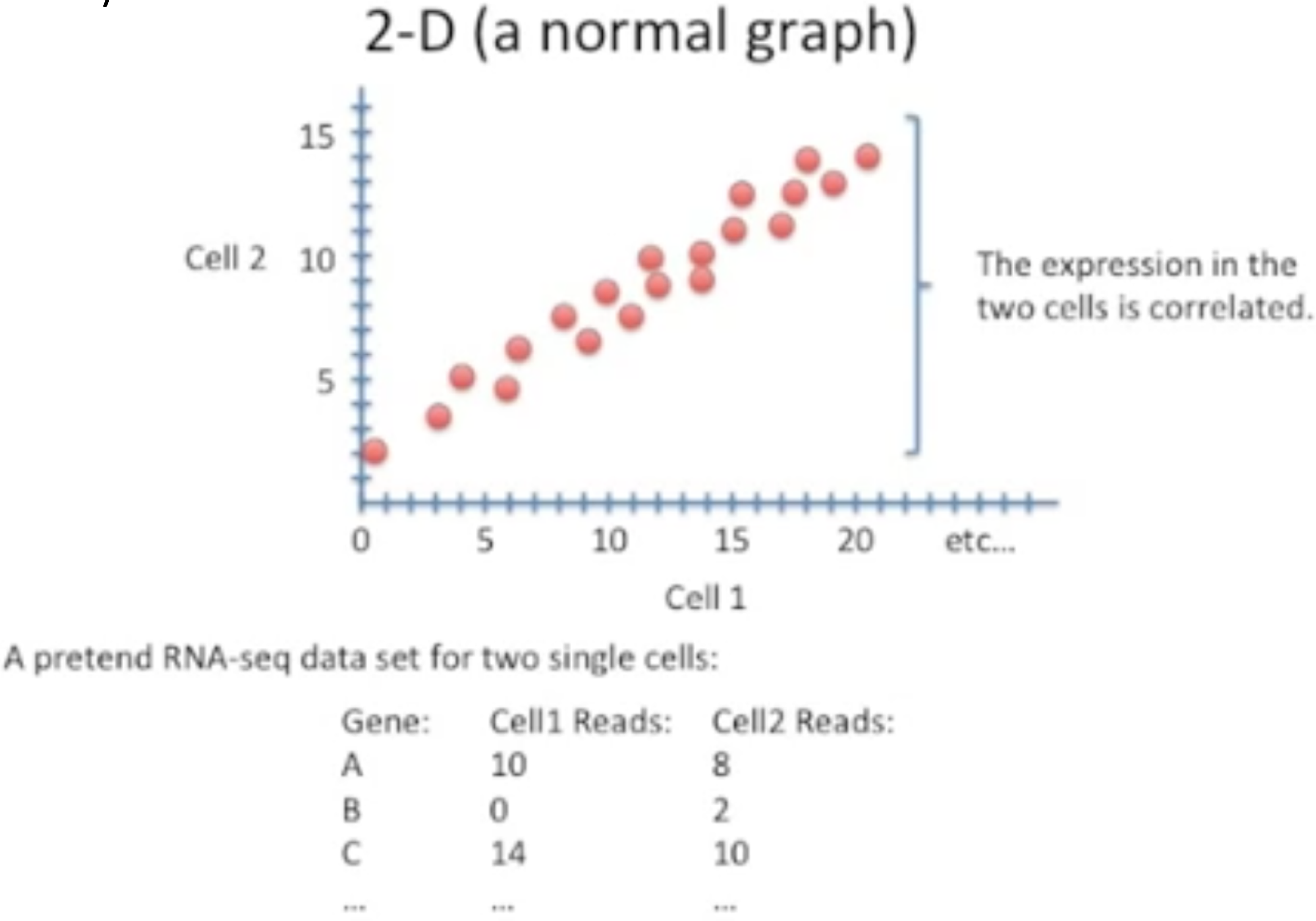
2-D (a normal graph)

If we plotted all of the genes, we might see...

A pretend RNA-seq data set for two single cells:

Gene:	Cell1 Reads:	Cell2 Reads:
A	10	8
B	0	2
C	14	10
...	...	...

Reproducibility - PCA



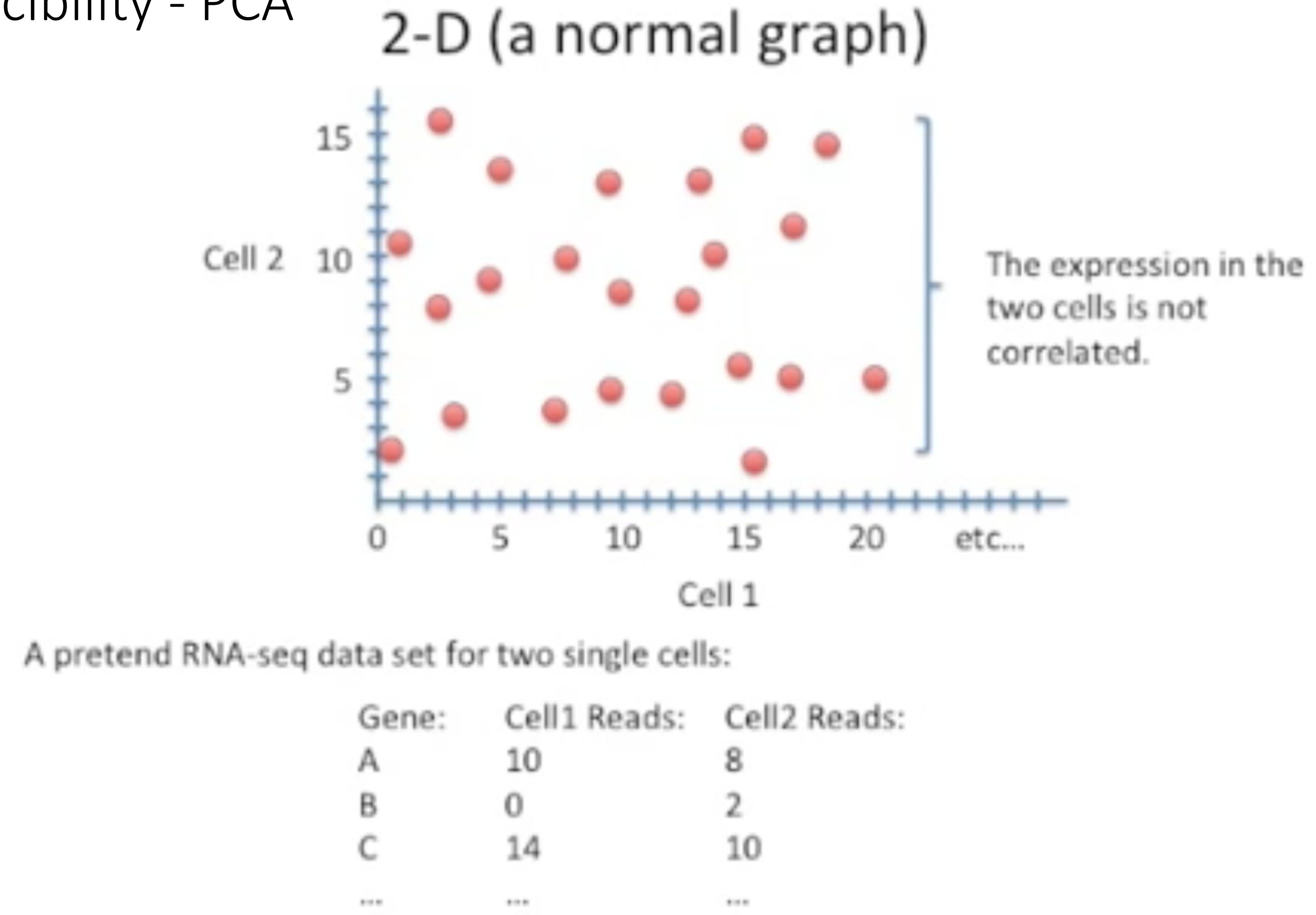
2-D (a normal graph)

The expression in the two cells is correlated.

A pretend RNA-seq data set for two single cells:

Gene:	Cell1 Reads:	Cell2 Reads:
A	10	8
B	0	2
C	14	10
...	...	...

Reproducibility - PCA



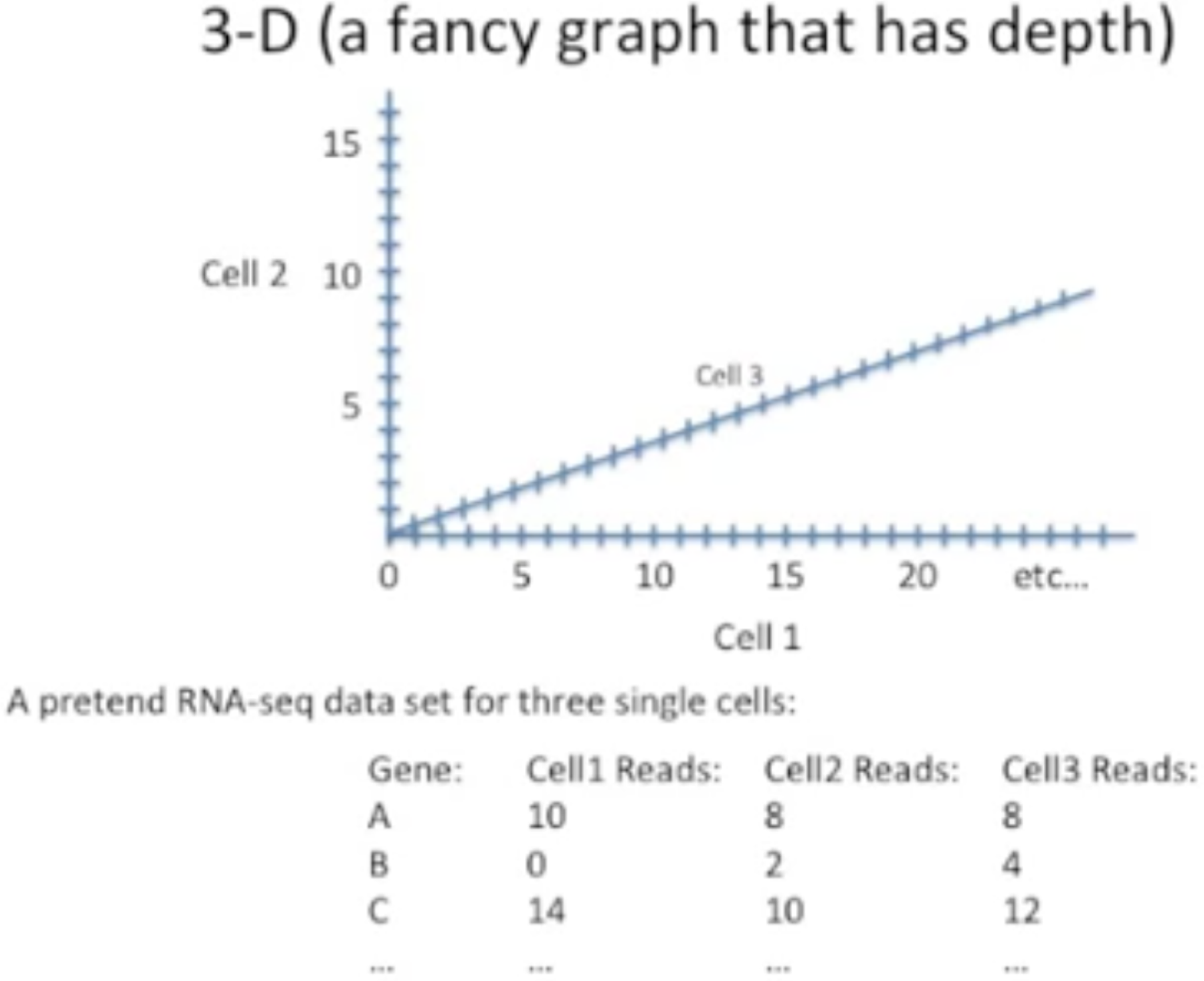
2-D (a normal graph)

The expression in the two cells is not correlated.

A pretend RNA-seq data set for two single cells:

Gene:	Cell1 Reads:	Cell2 Reads:
A	10	8
B	0	2
C	14	10
...	...	...

Reproducibility - PCA

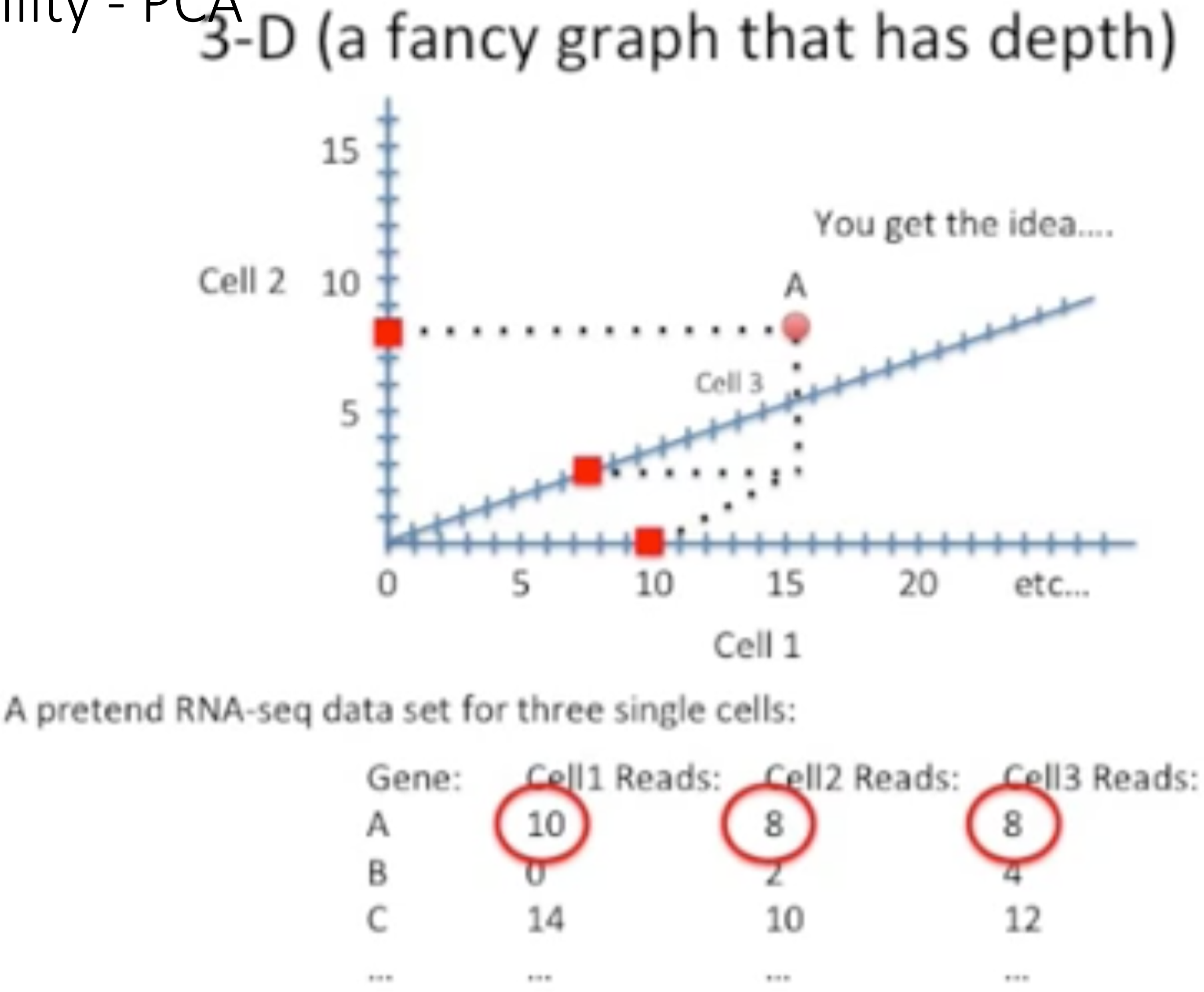


3-D (a fancy graph that has depth)

A pretend RNA-seq data set for three single cells:

Gene:	Cell1 Reads:	Cell2 Reads:	Cell3 Reads:
A	10	8	8
B	0	2	4
C	14	10	12
...	...	...	...

Reproducibility - PCA



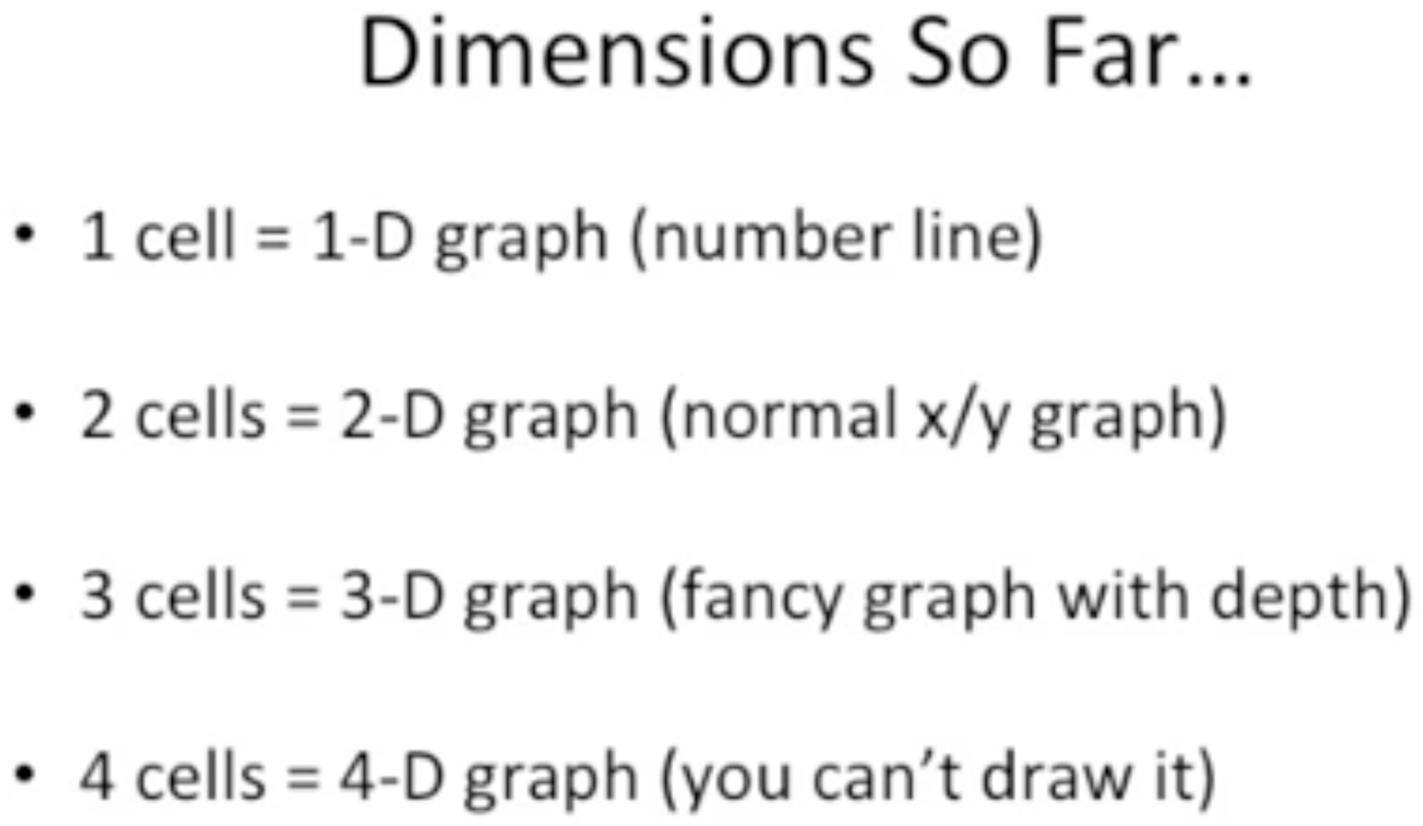
3-D (a fancy graph that has depth)

You get the idea....

A pretend RNA-seq data set for three single cells:

Gene:	Cell1 Reads:	Cell2 Reads:	Cell3 Reads:
A	10	8	8
B	0	2	4
C	14	10	12
...	...	...	...

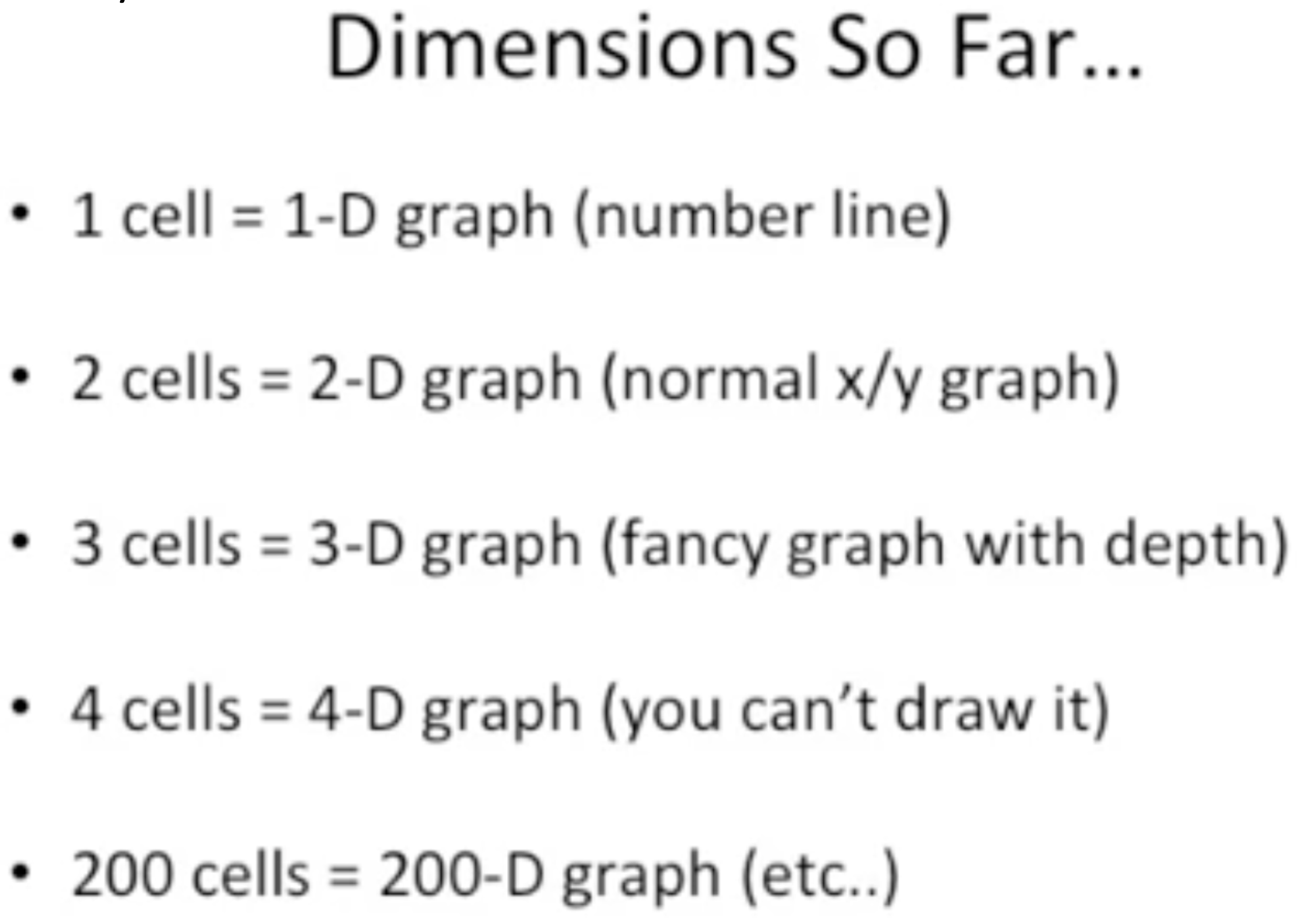
Reproducibility - PCA



Dimensions So Far…

- 1 cell = 1-D graph (number line)
- 2 cells = 2-D graph (normal x/y graph)
- 3 cells = 3-D graph (fancy graph with depth)
- 4 cells = 4-D graph (you can’t draw it)

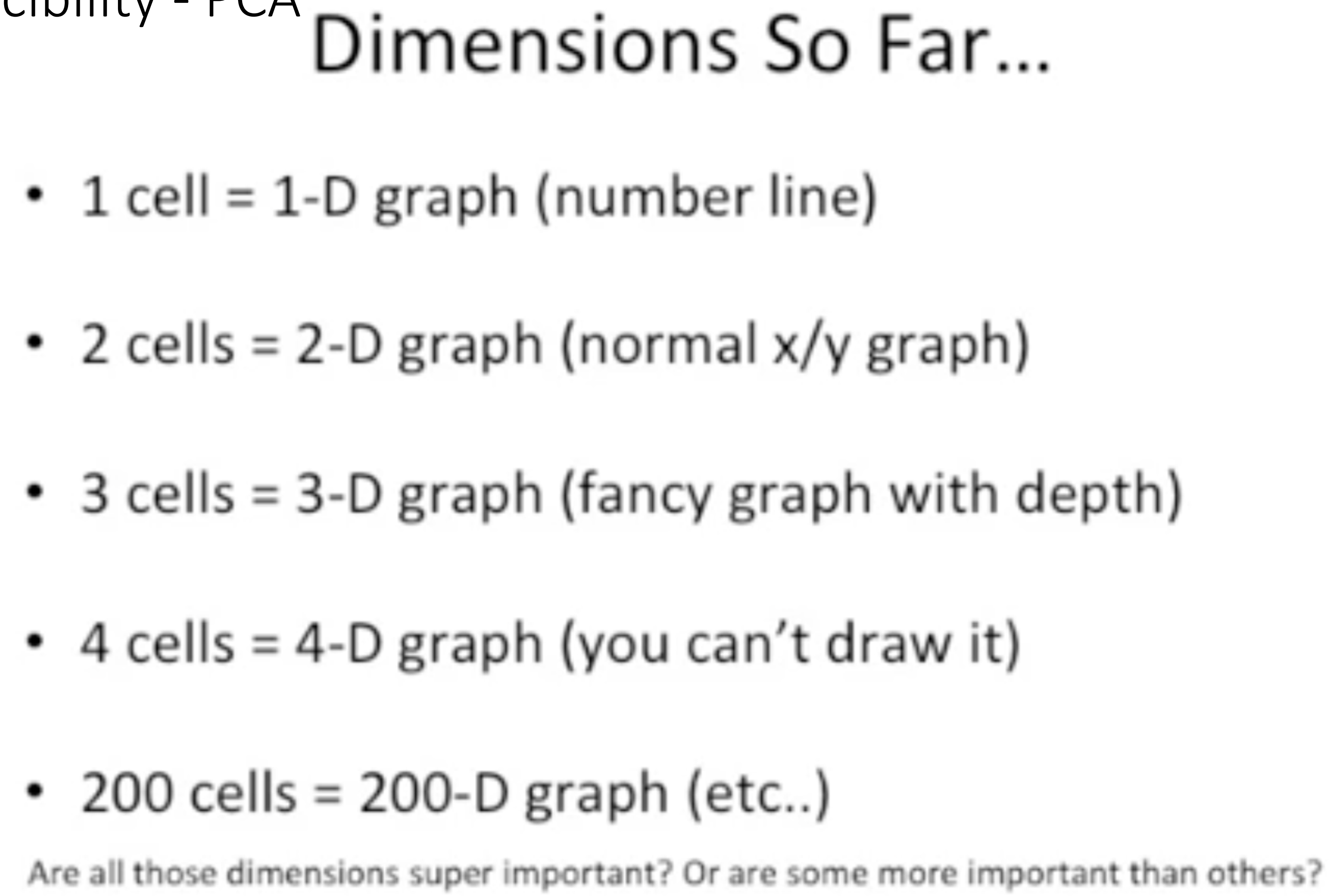
Reproducibility - PCA



Dimensions So Far…

- 1 cell = 1-D graph (number line)
- 2 cells = 2-D graph (normal x/y graph)
- 3 cells = 3-D graph (fancy graph with depth)
- 4 cells = 4-D graph (you can’t draw it)
- 200 cells = 200-D graph (etc..)

Reproducibility - PCA

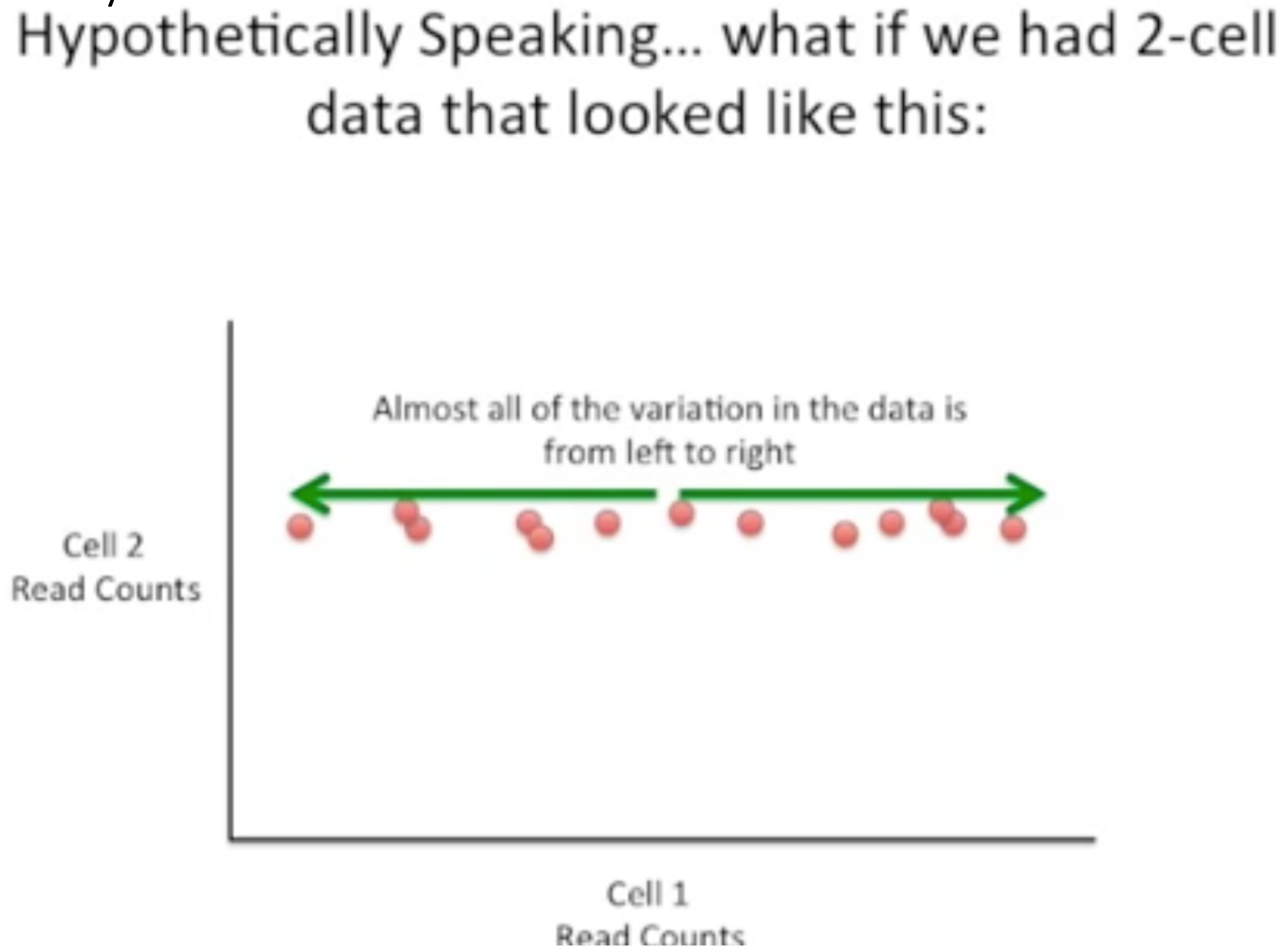


Dimensions So Far…

- 1 cell = 1-D graph (number line)
- 2 cells = 2-D graph (normal x/y graph)
- 3 cells = 3-D graph (fancy graph with depth)
- 4 cells = 4-D graph (you can't draw it)
- 200 cells = 200-D graph (etc..)

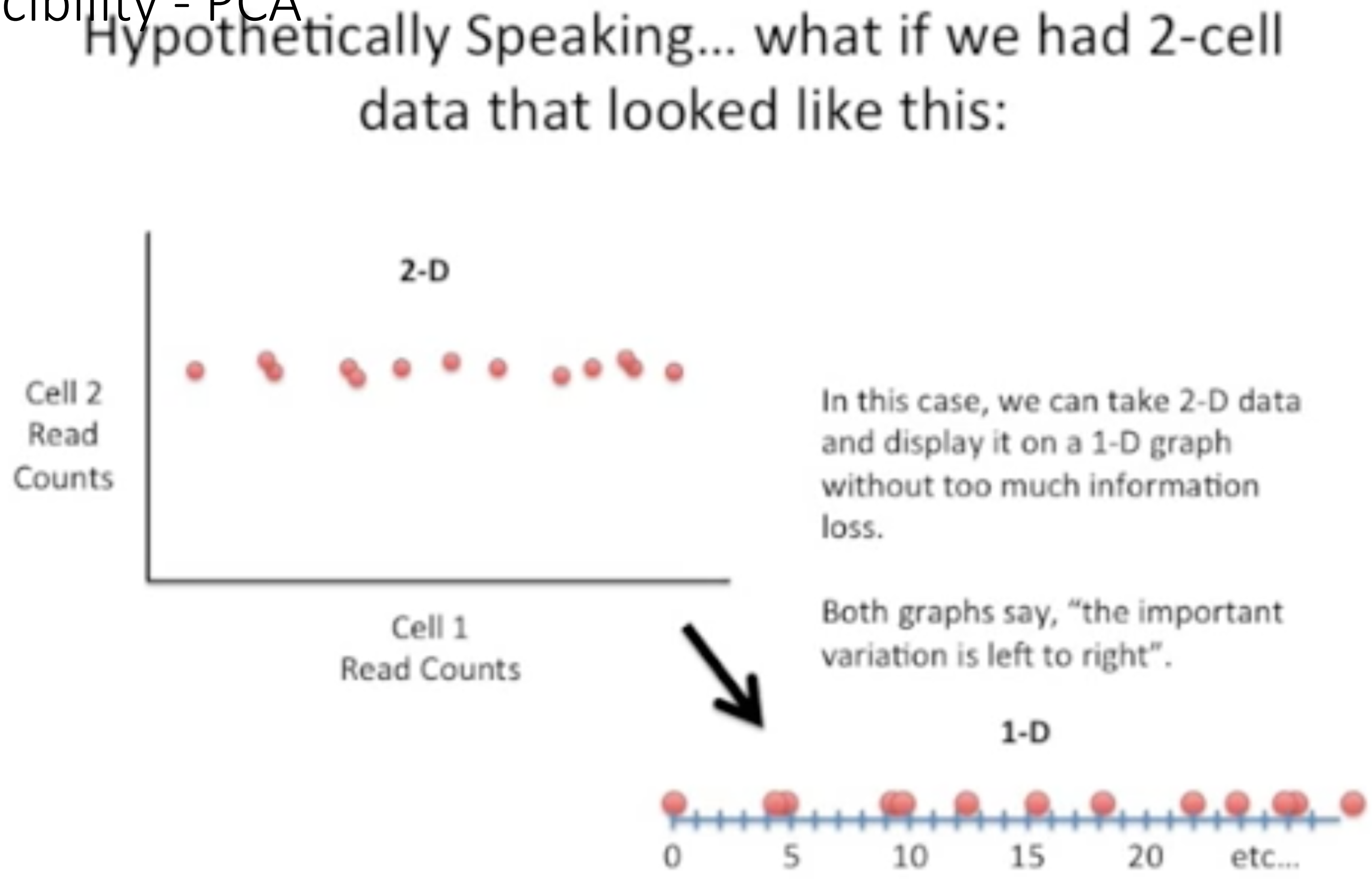
Are all those dimensions super important? Or are some more important than others?

Reproducibility - PCA



Hypothetically Speaking… what if we had 2-cell data that looked like this:

Reproducibility - PCA



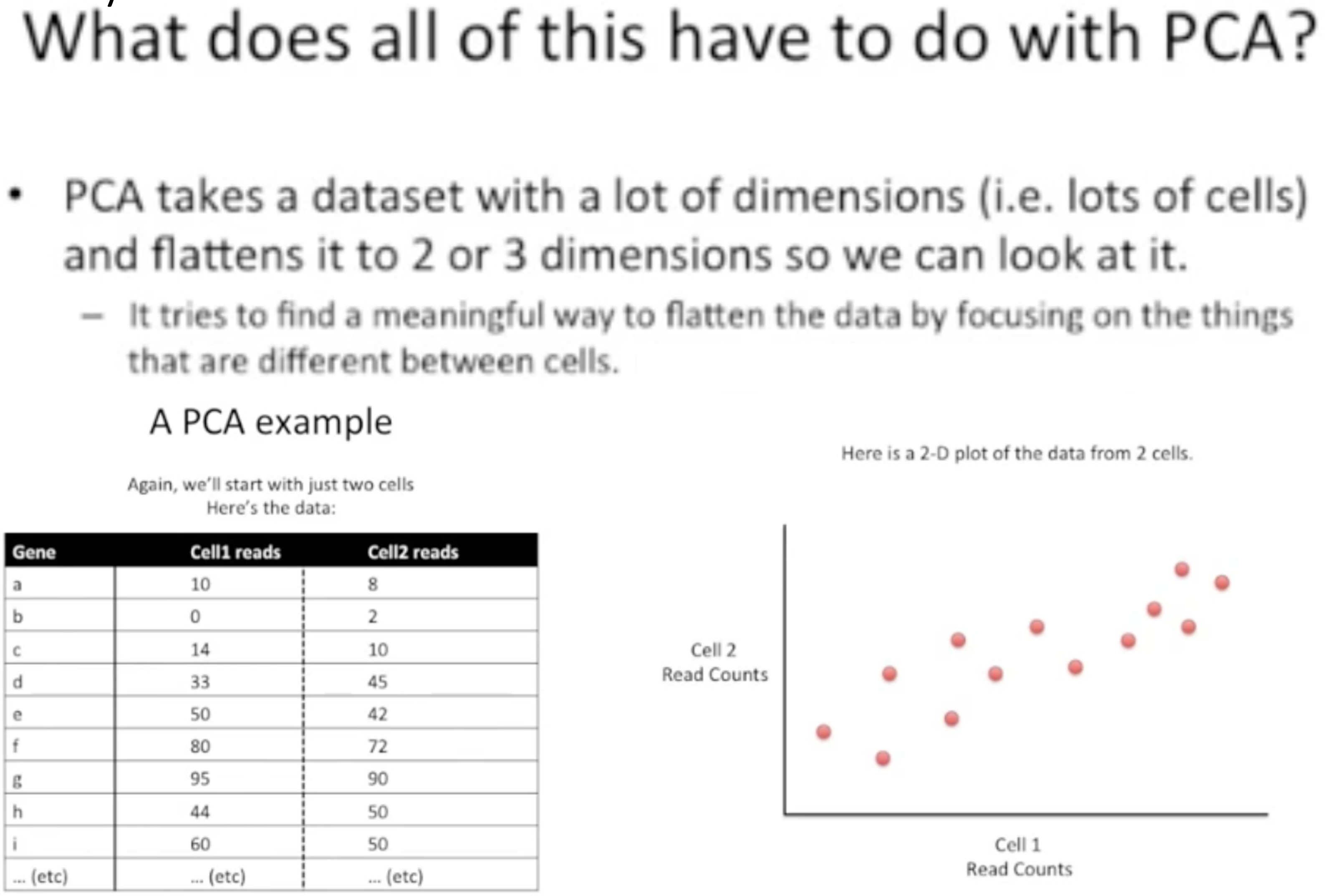
Hypothetically Speaking… what if we had 2-cell data that looked like this:

In this case, we can take 2-D data and display it on a 1-D graph without too much information loss.

Both graphs say, "the important variation is left to right".

Some dimensions are more important the others

Reproducibility - PCA



What does all of this have to do with PCA?

- PCA takes a dataset with a lot of dimensions (i.e. lots of cells) and flattens it to 2 or 3 dimensions so we can look at it.
  - It tries to find a meaningful way to flatten the data by focusing on the things that are different between cells.

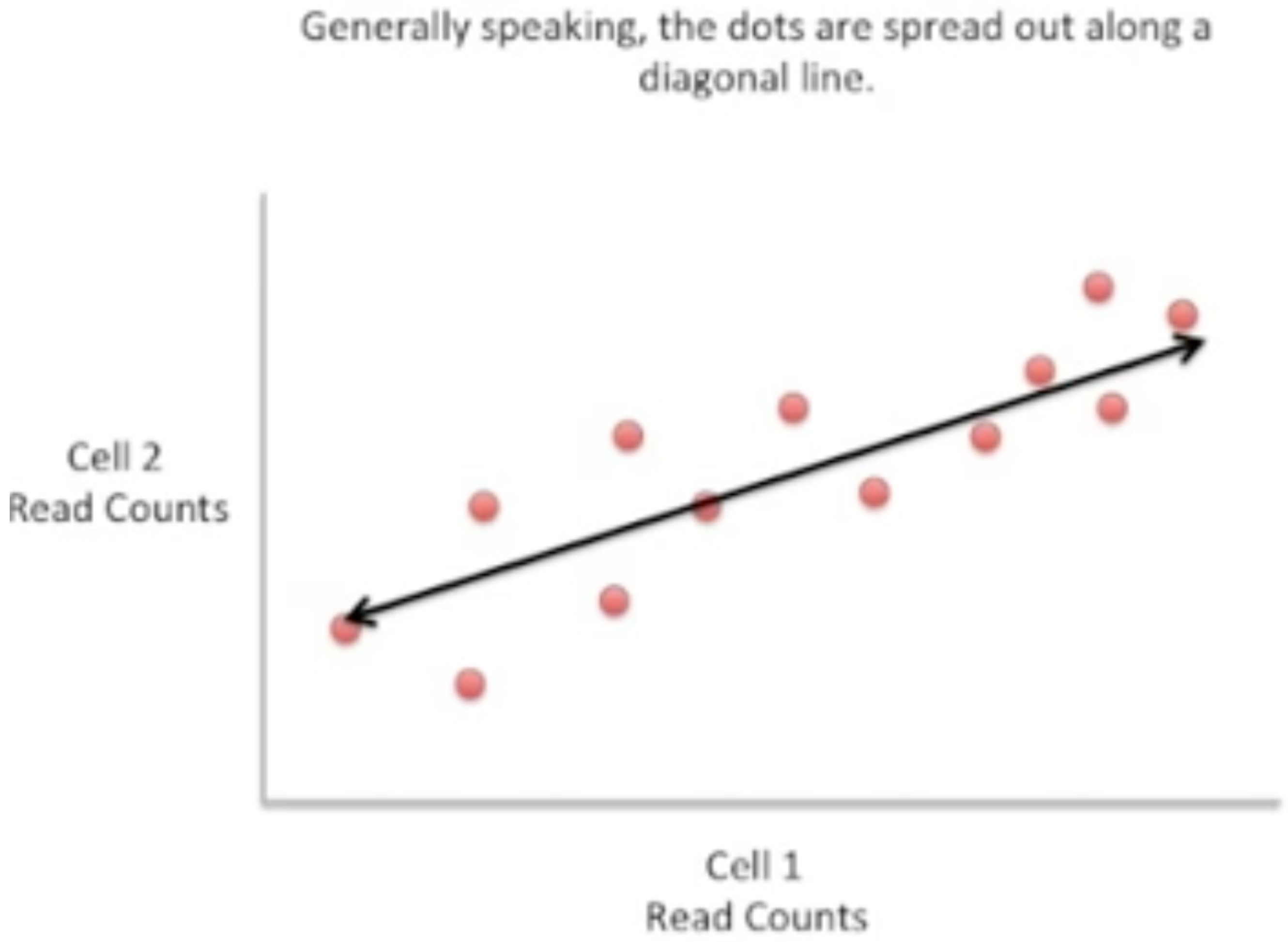
A PCA example

Again, we'll start with just two cells
Here's the data:

Gene	Cell1 reads	Cell2 reads
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
... (etc)	... (etc)	... (etc)

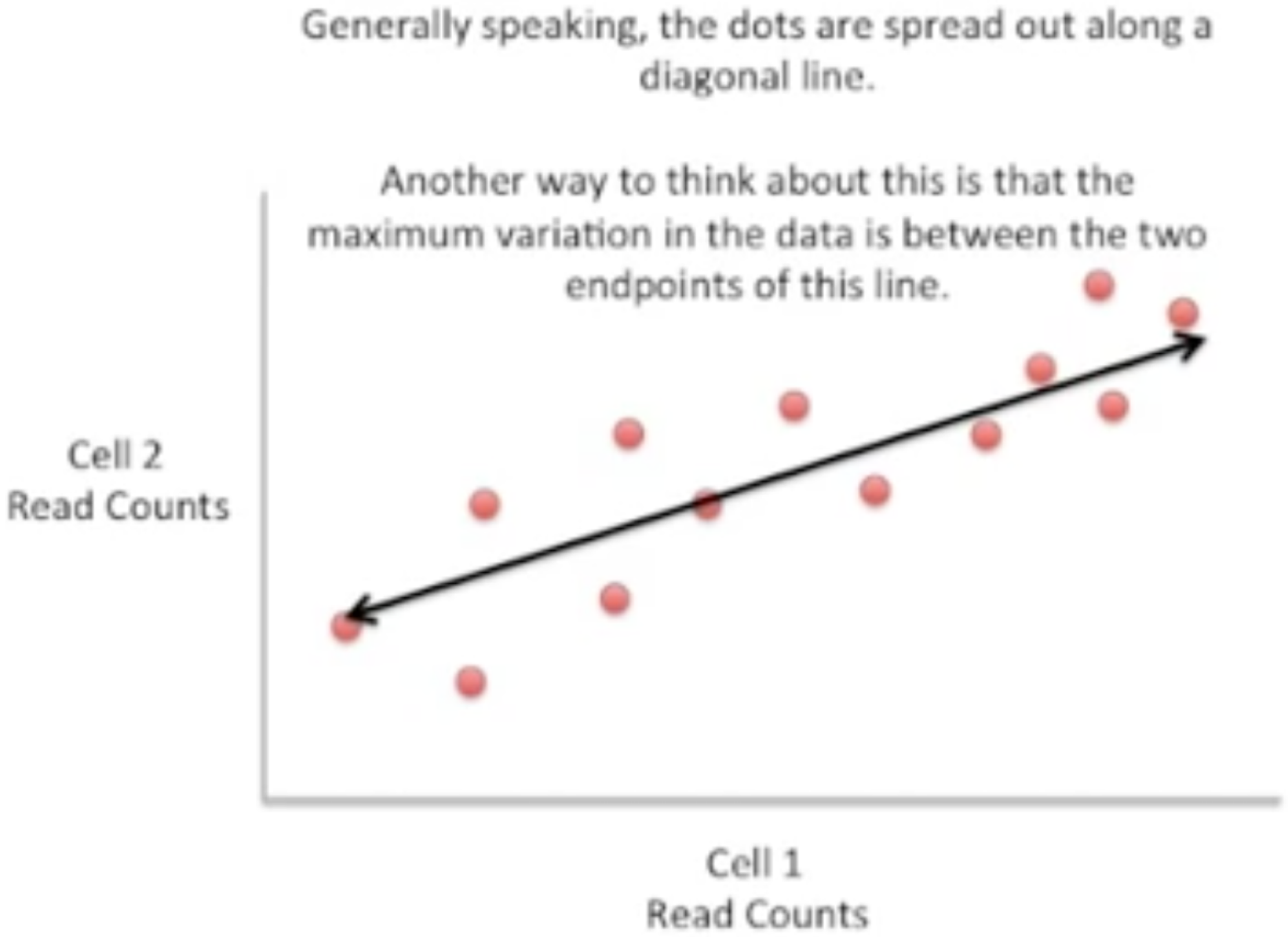
Here is a 2-D plot of the data from 2 cells.

Reproducibility - PCA



Generally speaking, the dots are spread out along a diagonal line.

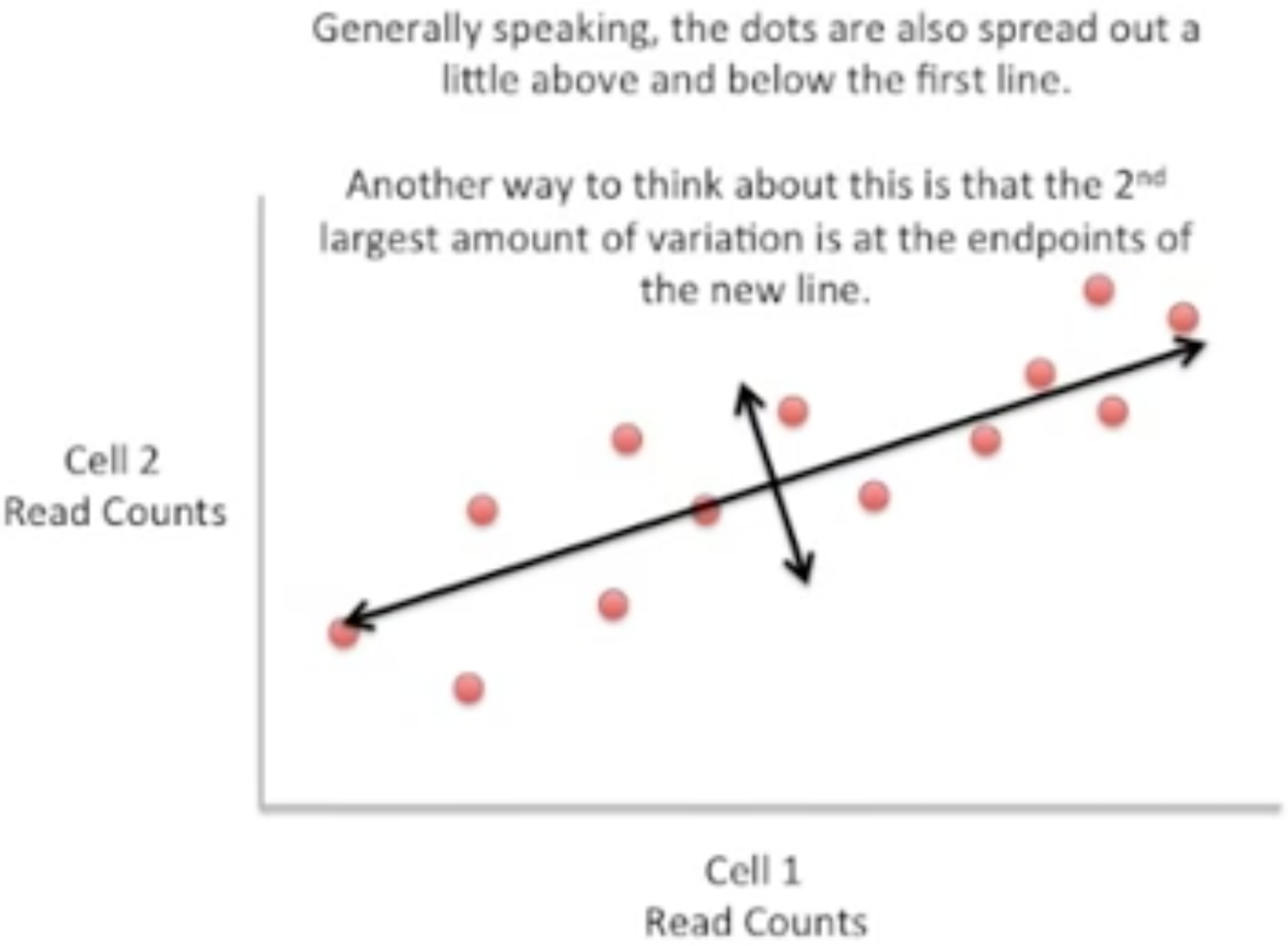
Reproducibility - PCA



Generally speaking, the dots are spread out along a diagonal line.

Another way to think about this is that the maximum variation in the data is between the two endpoints of this line.

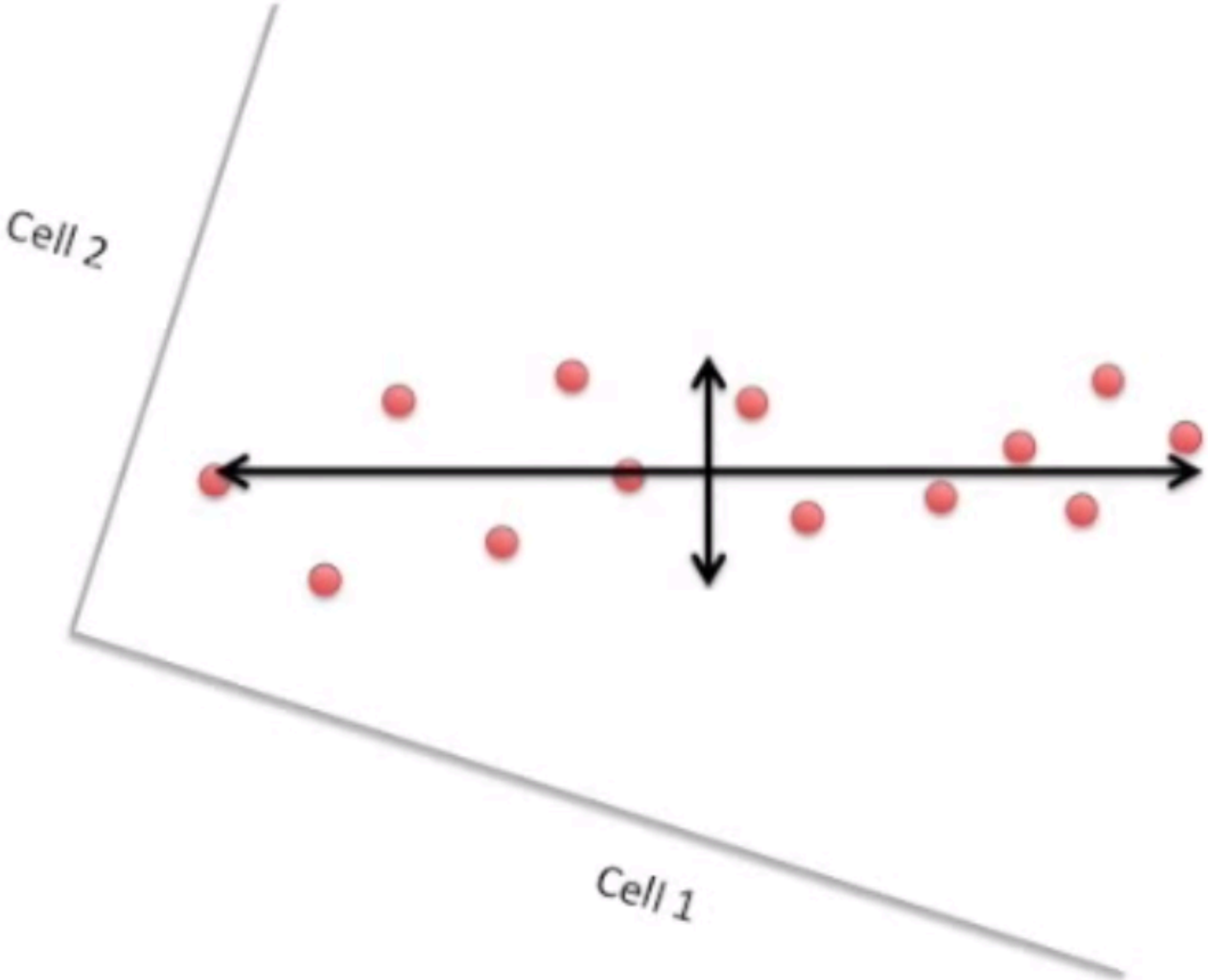
Reproducibility - PCA



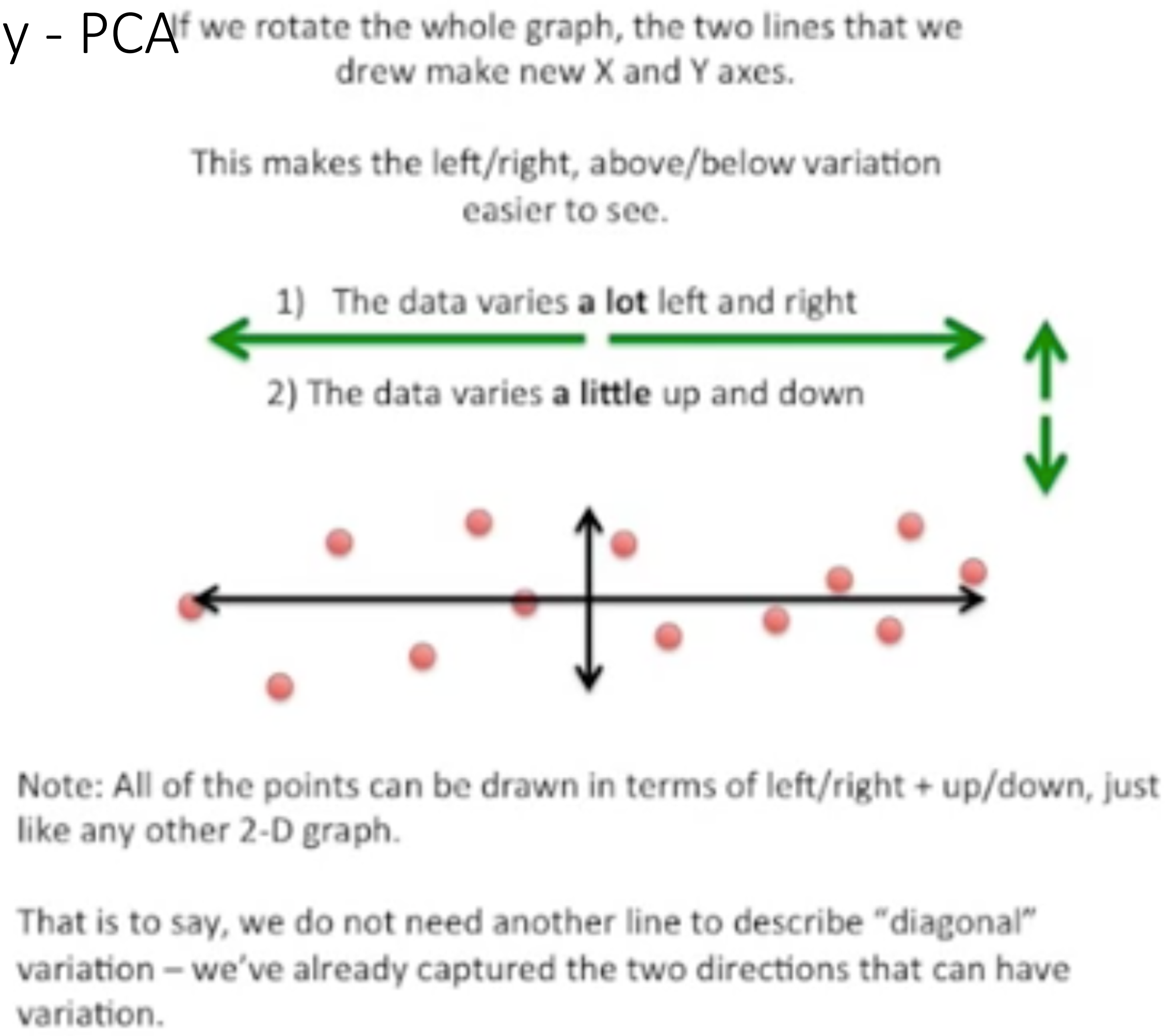
Generally speaking, the dots are also spread out a little above and below the first line.

Another way to think about this is that the 2nd largest amount of variation is at the endpoints of the new line.

If we rotate the whole graph, the two lines that we drew make new X and Y axes.



Reproducibility - PCA



If we rotate the whole graph, the two lines that we drew make new X and Y axes.

This makes the left/right, above/below variation easier to see.

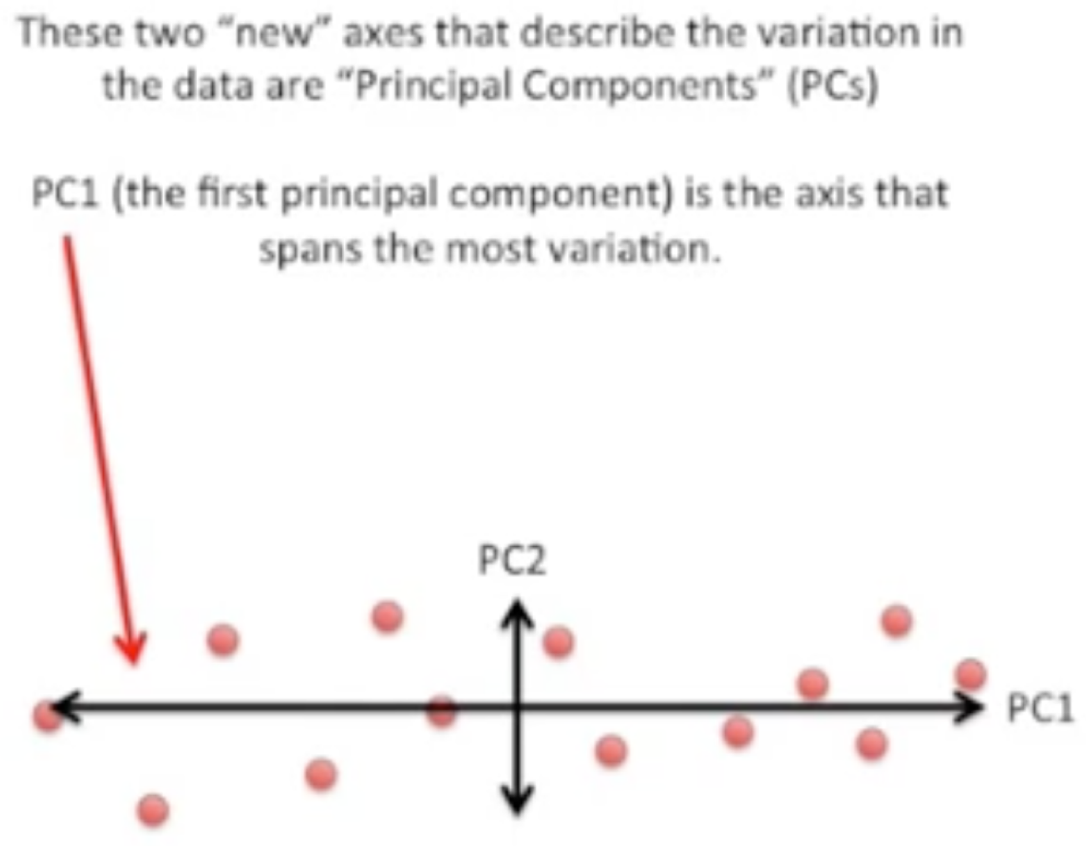
1) The data varies **a lot** left and right

2) The data varies **a little** up and down

Note: All of the points can be drawn in terms of left/right + up/down, just like any other 2-D graph.

That is to say, we do not need another line to describe "diagonal" variation – we've already captured the two directions that can have variation.

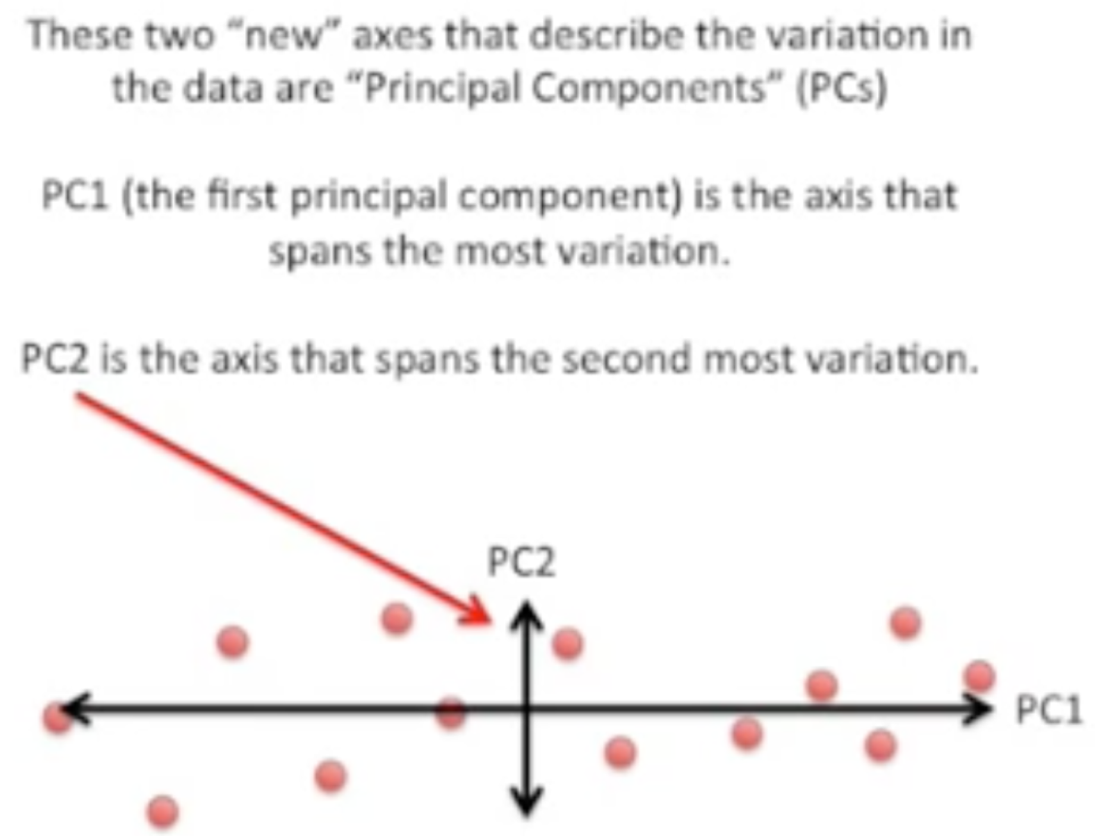
Reproducibility - PCA



These two "new" axes that describe the variation in the data are "Principal Components" (PCs)

PC1 (the first principal component) is the axis that spans the most variation.

Reproducibility - PCA

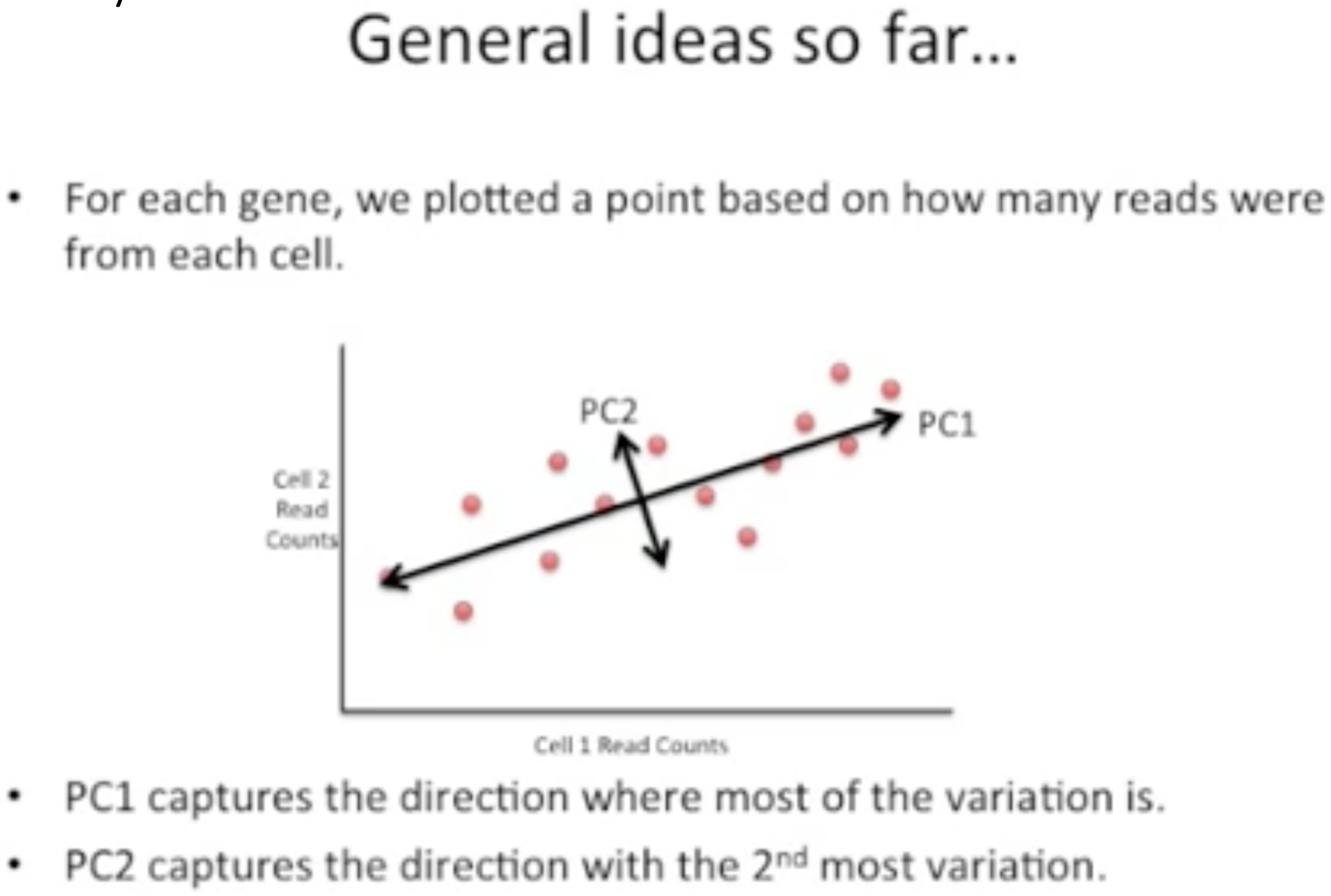


These two "new" axes that describe the variation in the data are "Principal Components" (PCs)

PC1 (the first principal component) is the axis that spans the most variation.

PC2 is the axis that spans the second most variation.

Reproducibility - PCA

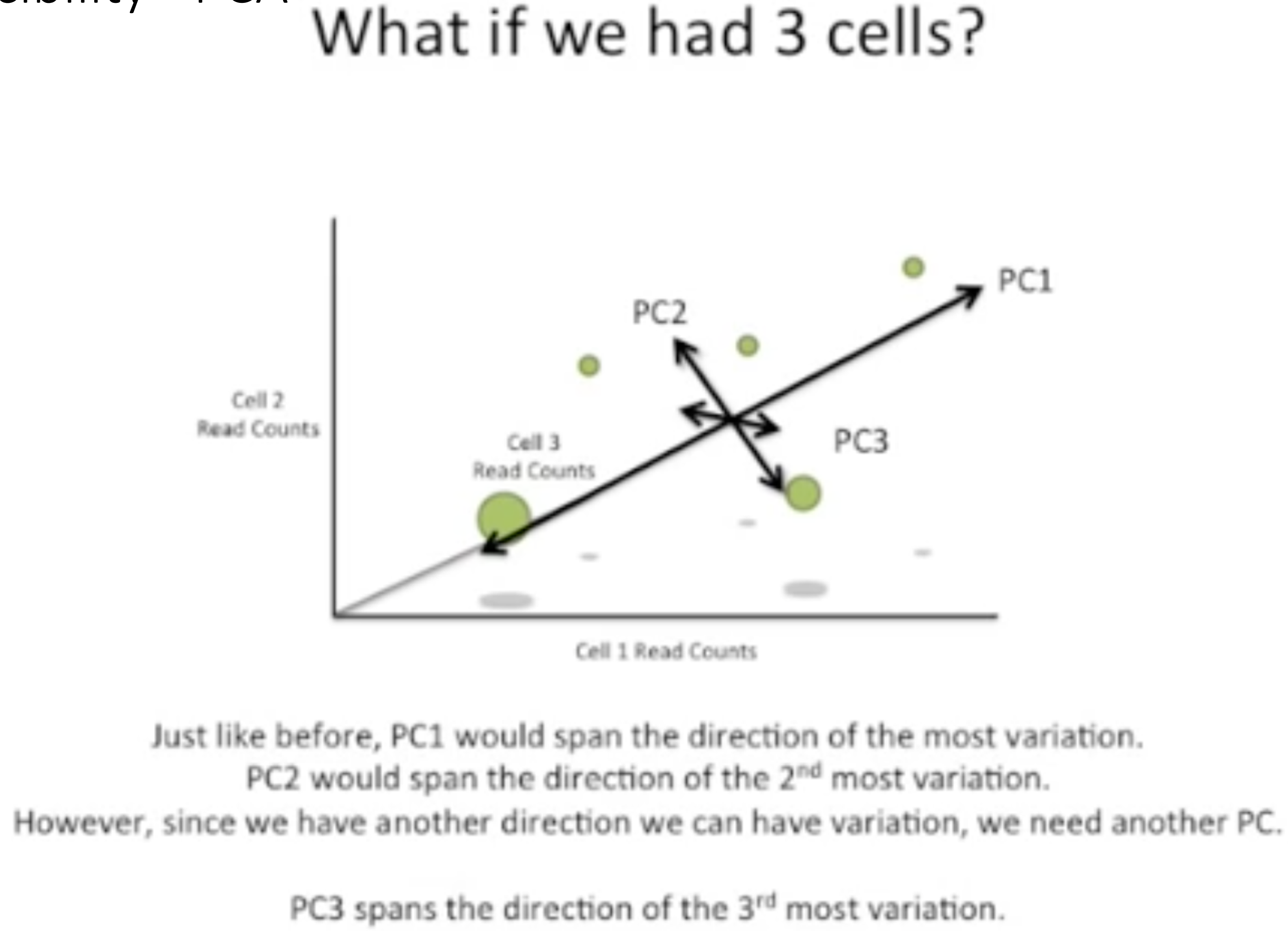


General ideas so far…

- For each gene, we plotted a point based on how many reads were from each cell.

- PC1 captures the direction where most of the variation is.
- PC2 captures the direction with the 2nd most variation.

Reproducibility - PCA

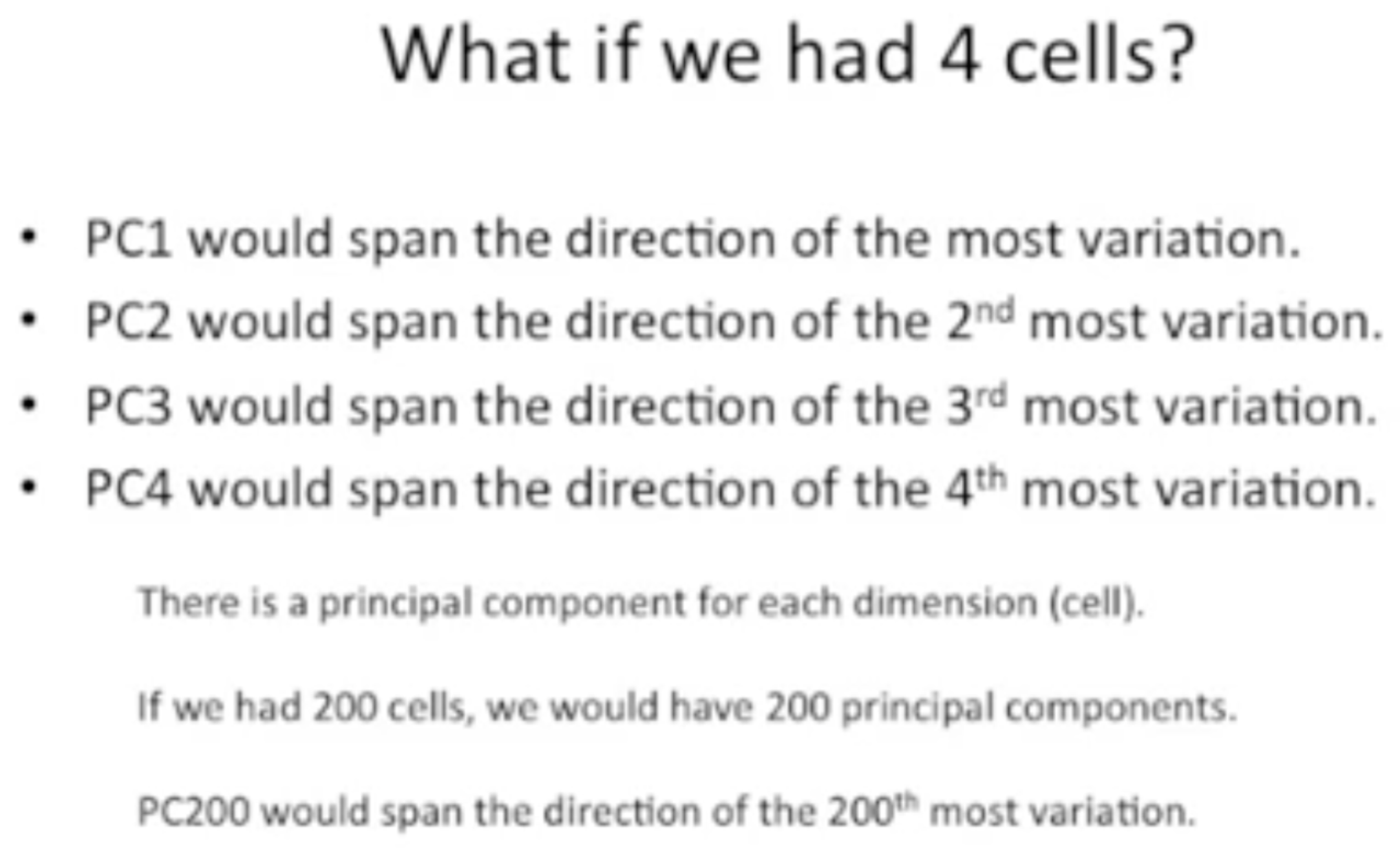


What if we had 3 cells?

Just like before, PC1 would span the direction of the most variation.
PC2 would span the direction of the 2nd most variation.
However, since we have another direction we can have variation, we need another PC.

PC3 spans the direction of the 3rd most variation.

Reproducibility - PCA



What if we had 4 cells?

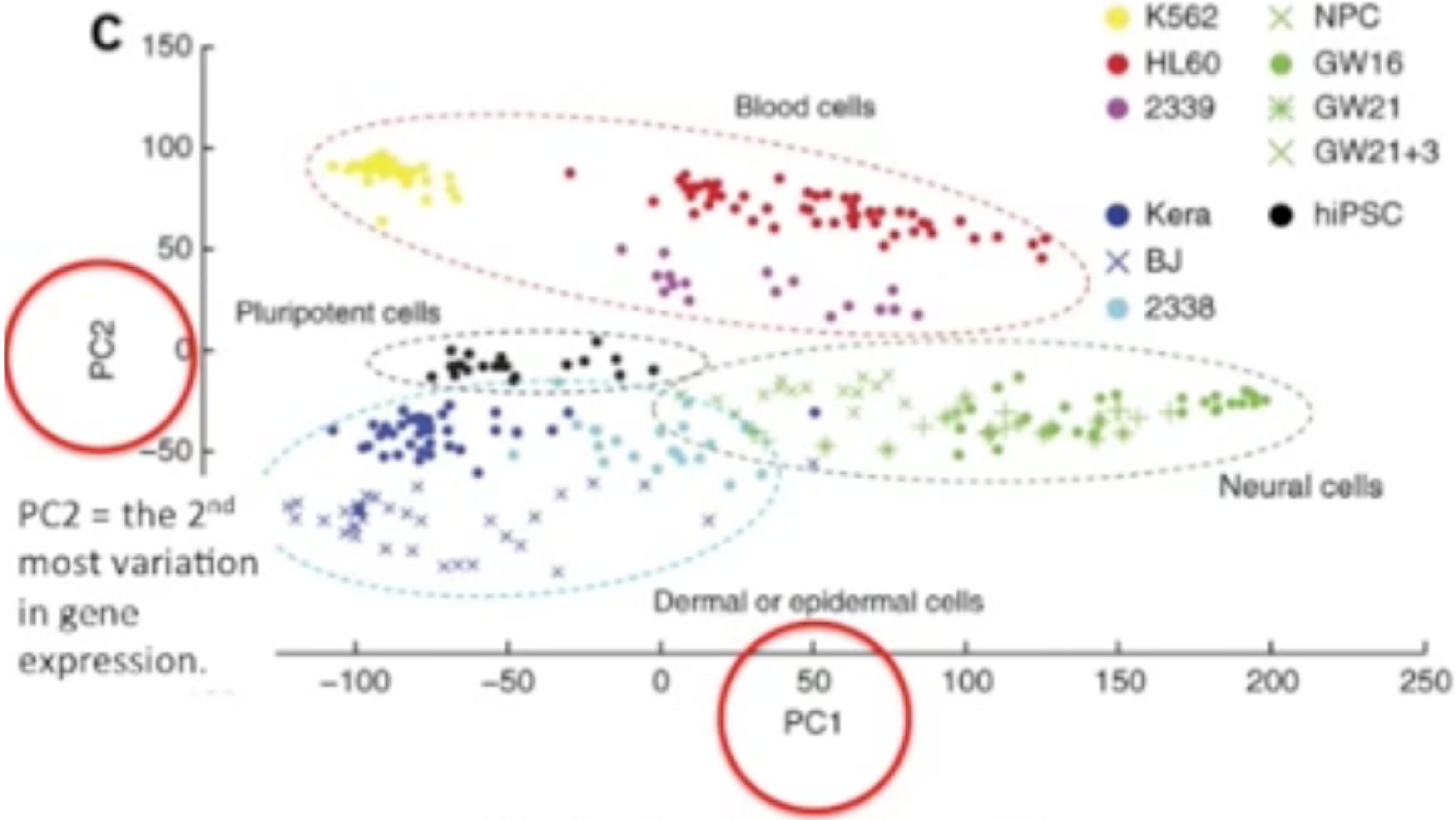
- PC1 would span the direction of the most variation.
- PC2 would span the direction of the 2nd most variation.
- PC3 would span the direction of the 3rd most variation.
- PC4 would span the direction of the 4th most variation.

There is a principal component for each dimension (cell).

If we had 200 cells, we would have 200 principal components.

PC200 would span the direction of the 200th most variation.

Reproducibility - PCA



PC2 = the 2nd most variation in gene expression.

PC1 – the direction of the most variation in gene expression.