



# RNA-seq to study HIV Infection in cells

Rebecca Batorsky  
Sr Bioinformatics  
Specialist  
Oct 2020

# Research Technology Team



**Delilah Maloney**  
High Performance Computing Specialist



**Kyle Monahan**  
Senior Data Science Specialist



**Shawn Doughty**  
Manager, Research Computing



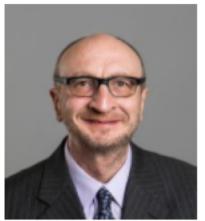
**Rebecca Batorsky**  
Senior Bioinformatics Specialist



**Meg Farley**  
Bioinformatics Intern



**Chris Barnett**  
Senior Geospatial Analyst



**Lionel H. Zupan**  
Director, Research Technology



**Tom Phimmasen**  
Senior Data Consultant



**Patrick Florance**  
Director, Academic Data Services



**Jake Perl**  
Digital Humanities NLP Specialist



**Carolyn Talmadge**  
Senior GIS Specialist



**Uku-Kaspar Uustalu**  
Data Science Specialist

- ✓ Consultation on Projects and Grants
- ✓ High Performance Compute Cluster
- ✓ Workshops

<https://it.tufts.edu/research-technology>

# Overview

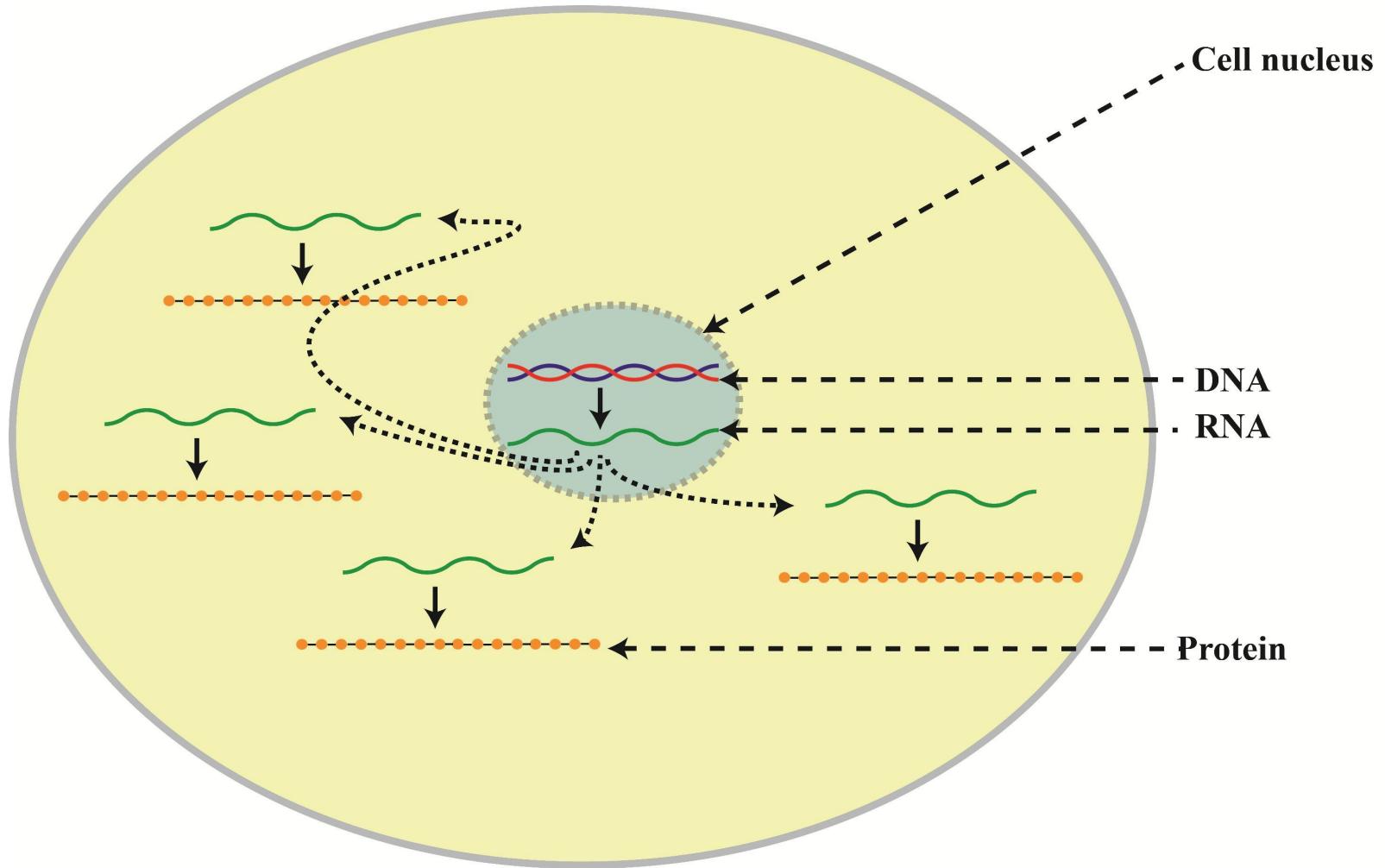
10/27

1. RNA sequencing
  - Bulk RNA sequencing
  - Single cell RNA sequencing
2. Intro to Galaxy platform for bioinformatics

11/3

1. Work through RNAseq example together on Galaxy

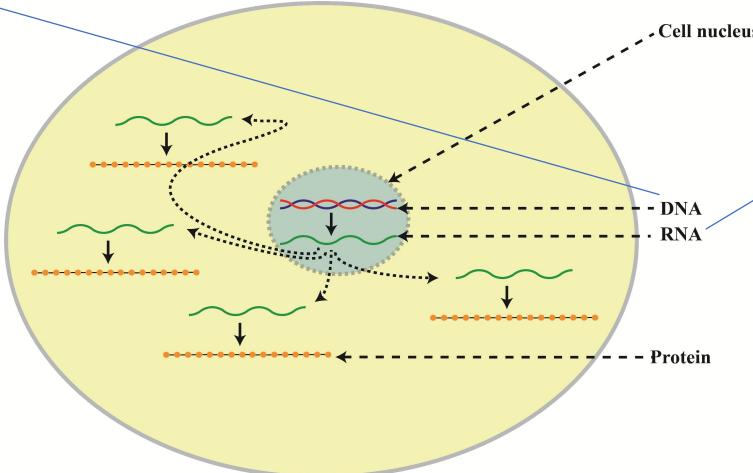
# DNA and RNA in a cell



# Two common analyses

## DNA Sequencing

- Fixed copy of a gene per cell
- Analysis goal:  
Variant calling and interpretation



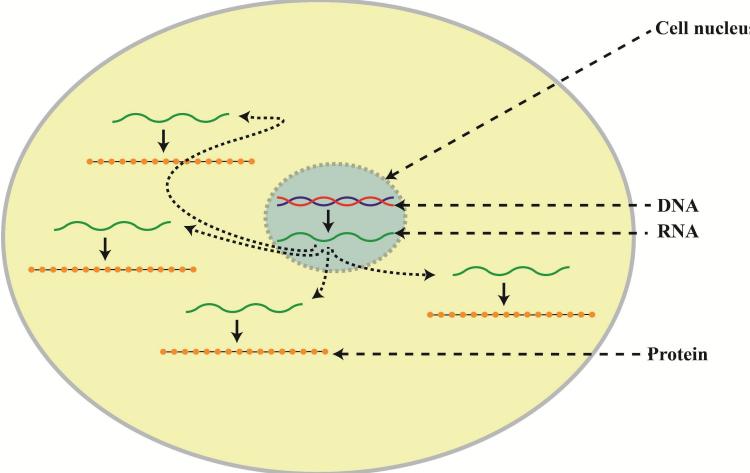
## RNA Sequencing

- Copy of a gene (mRNA) per cell depends on gene expression
- Analysis goal:
  - Bulk : Differential expression
  - Single cell : Quantify different cell populations

# Today we will cover RNA sequencing

## DNA Sequencing

- Fixed copy of a genome per cell
- Analysis goal:  
Variant calling, interpretation



## RNA Sequencing

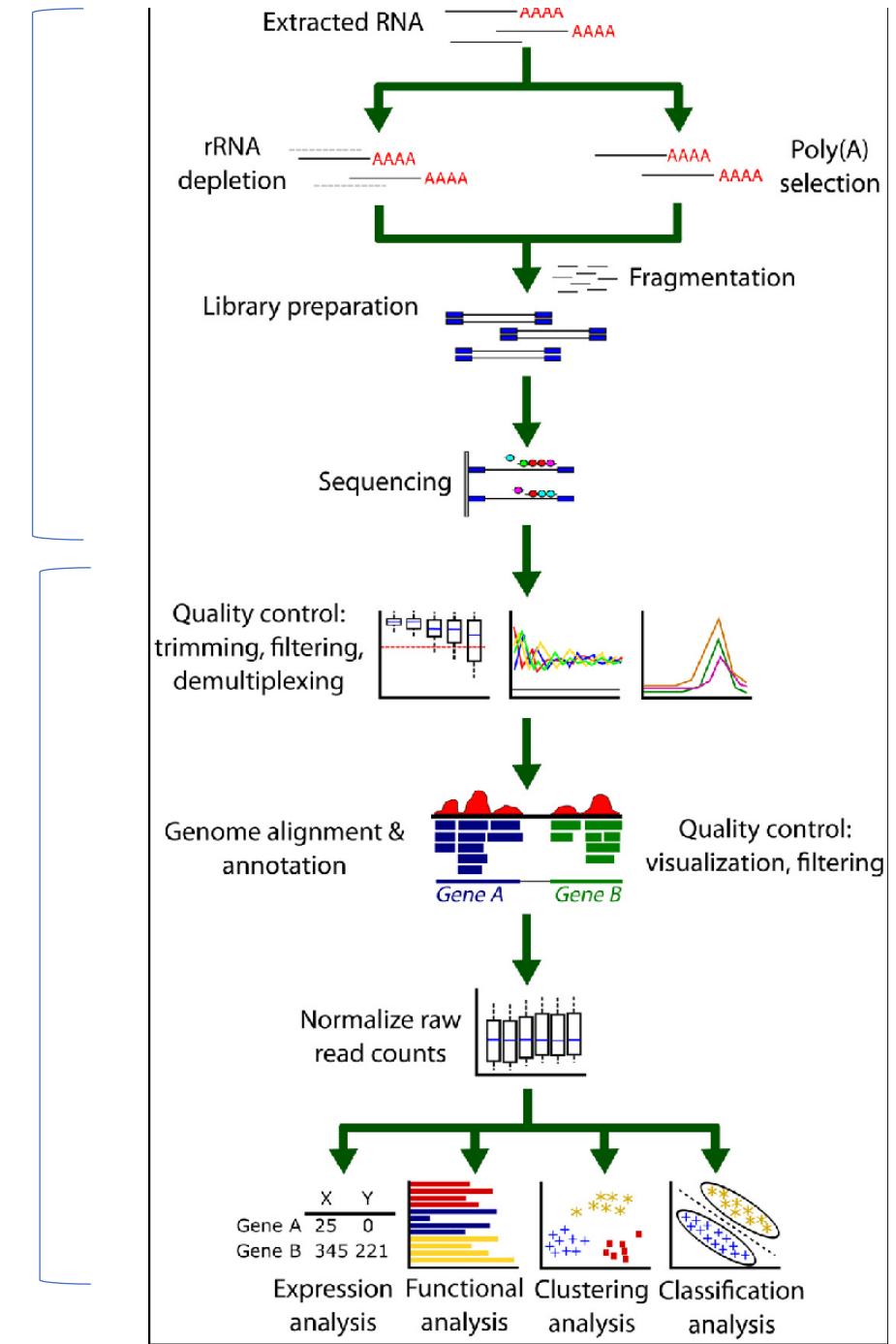
- Copy of a gene (mRNA) per cell depends on gene expression
- Analysis goal:
  - Bulk : Differential expression
  - Single cell : Quantify different cell populations

# “Bulk” RNA seq workflow

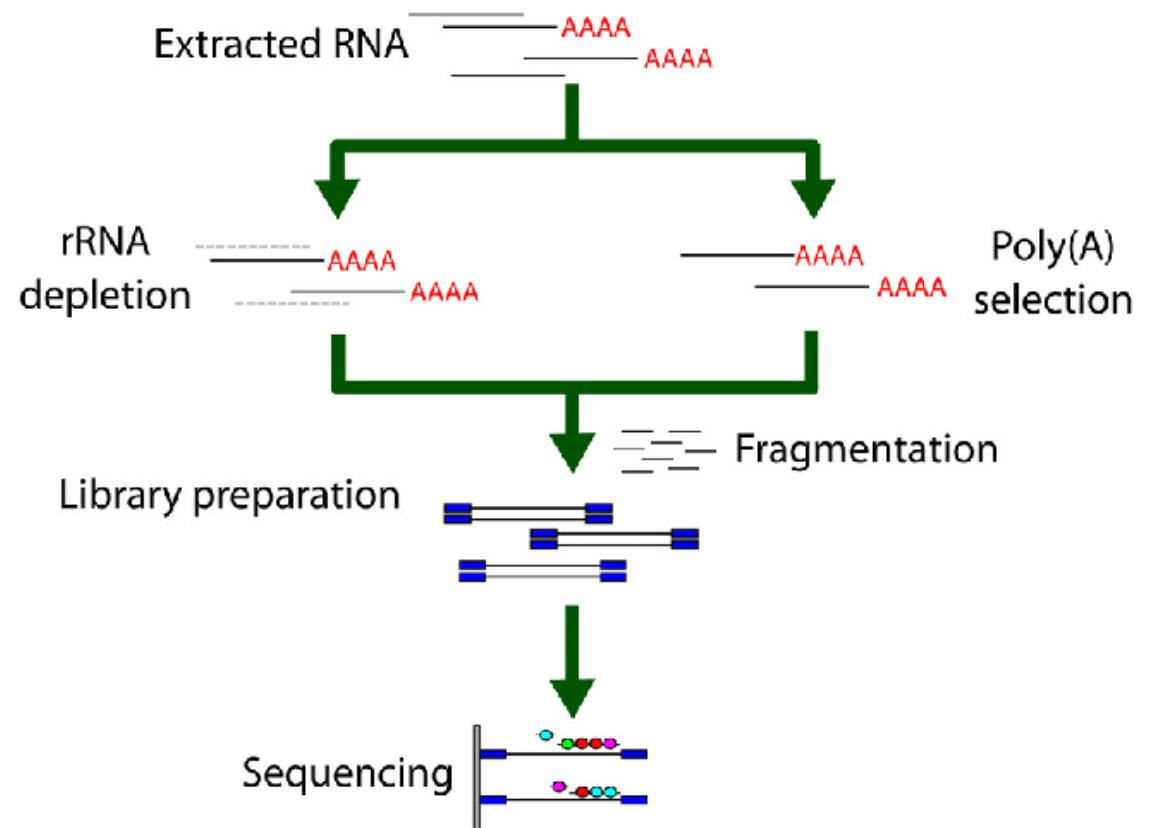
Library prep and sequencing

Bioinformatics

Good resource: [Griffiths et al Plos Comp Bio 2015](#)

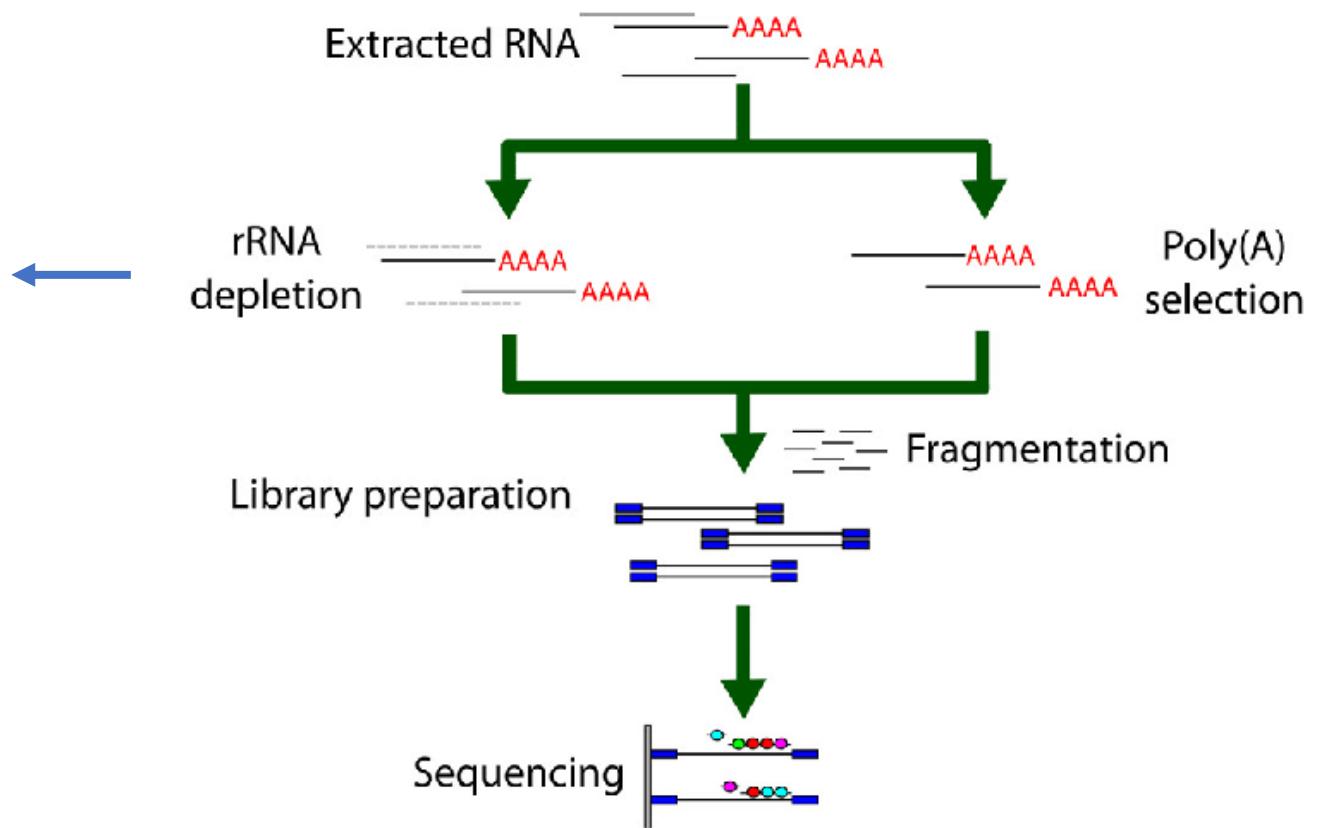


# RNA seq library prep and sequencing



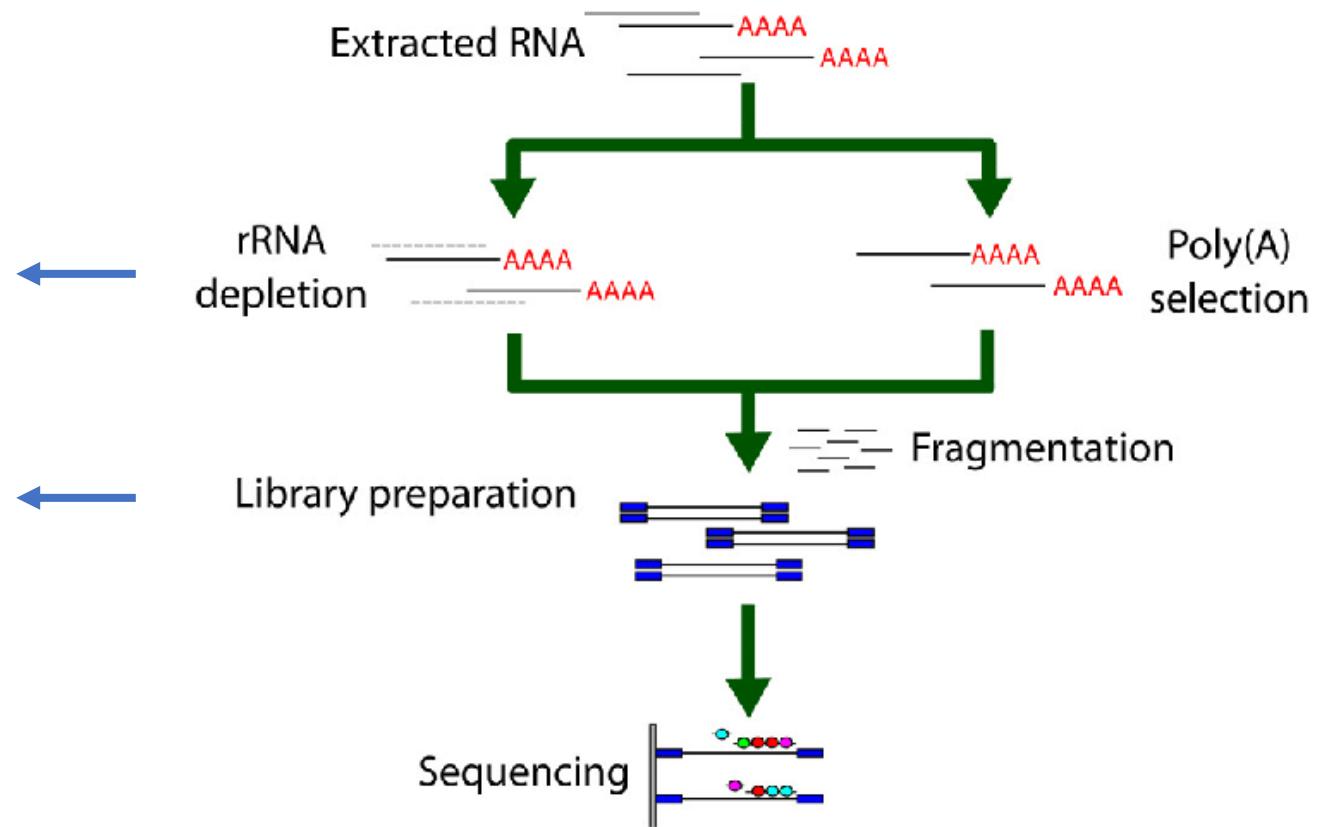
# RNA seq library prep and sequencing

- Enrichment for mRNA
- In humans, ~95%–98% of all RNA molecules are rRNAs

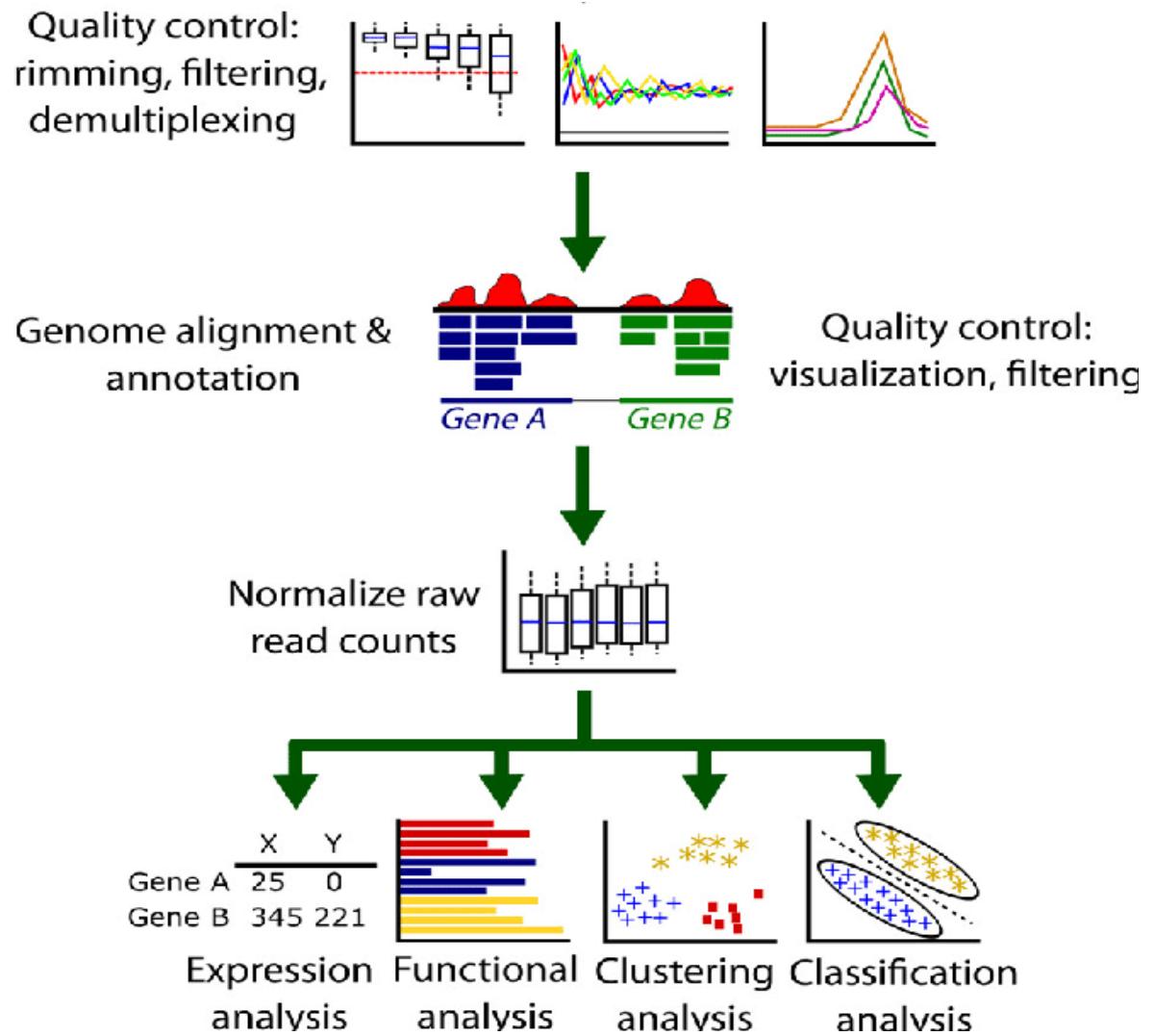


# RNA seq library prep and sequencing

- Enrichment for mRNA
- In humans, ~95%–98% of all RNA molecules are rRNAs
- Random priming and reverse transcription
- Double stranded cDNA synthesis
- Sequencing adapter ligation



# RNA seq bioinformatics



# Goal of RNAseq

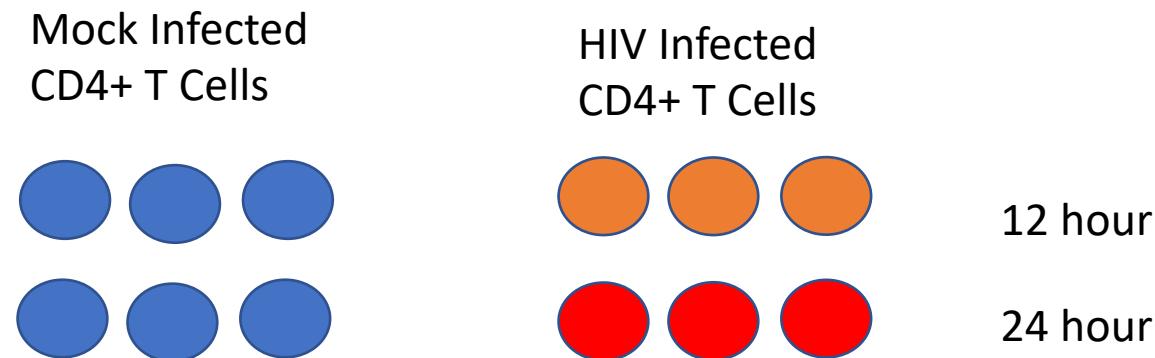
“How can we detect genes for which the counts of reads change between conditions **more systematically** than as expected by chance”

Oshlack et al. 2010. From RNA-seq reads to differential expression results. *Genome Biology* 2010, 11:220  
<http://genomebiology.com/2010/11/12/220>

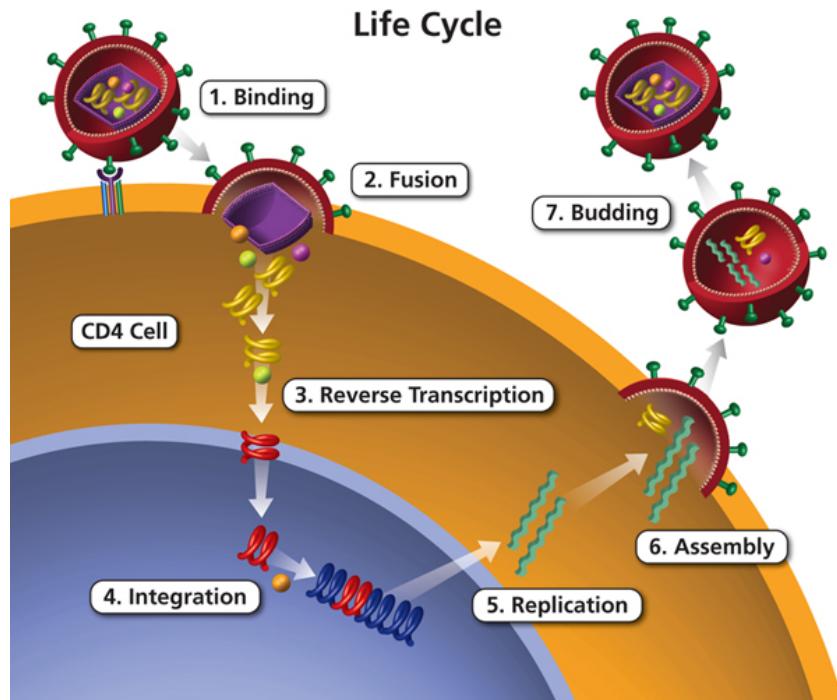
# Our dataset for next week

## Next-Generation Sequencing Reveals HIV-1-Mediated Suppression of T Cell Activation and RNA Processing and Regulation of Noncoding RNA Expression in a CD4<sup>+</sup> T Cell Line

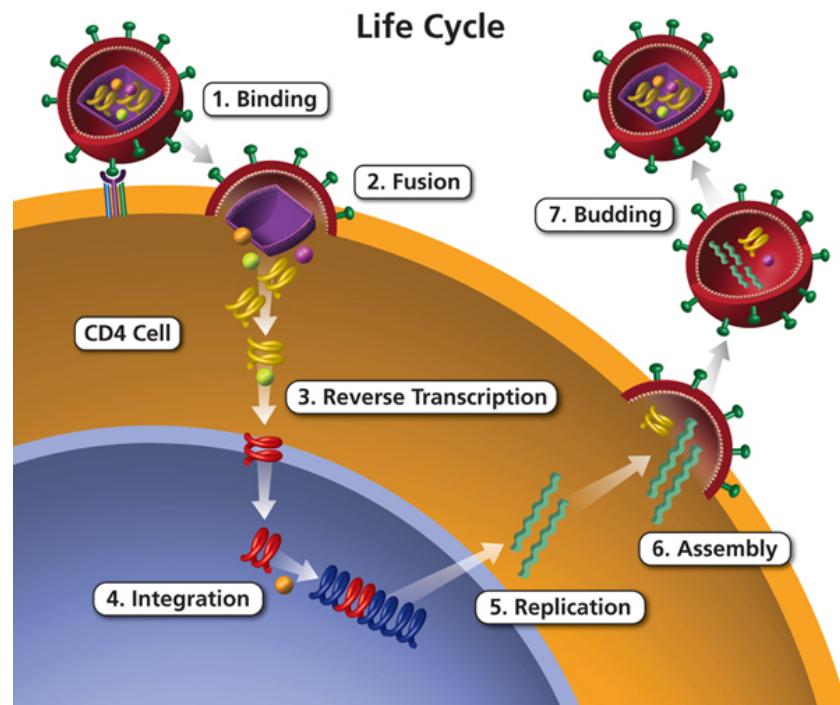
Stewart T. Chang, Pavel Sova, Xinxia Peng, Jeffrey Weiss, G. Lynn Law, Robert E. Palermo, Michael G. Katze



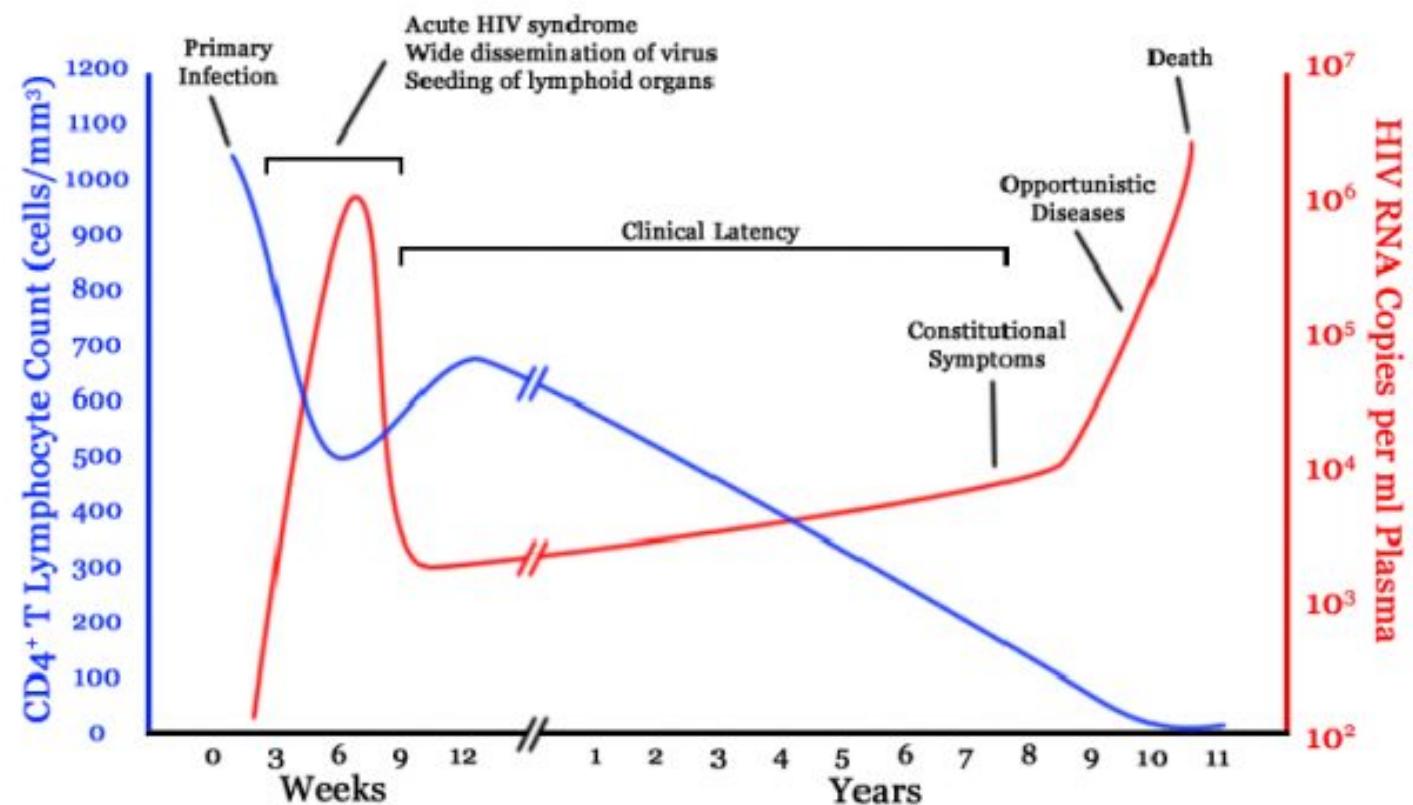
# HIV lifecycle



# HIV lifecycle

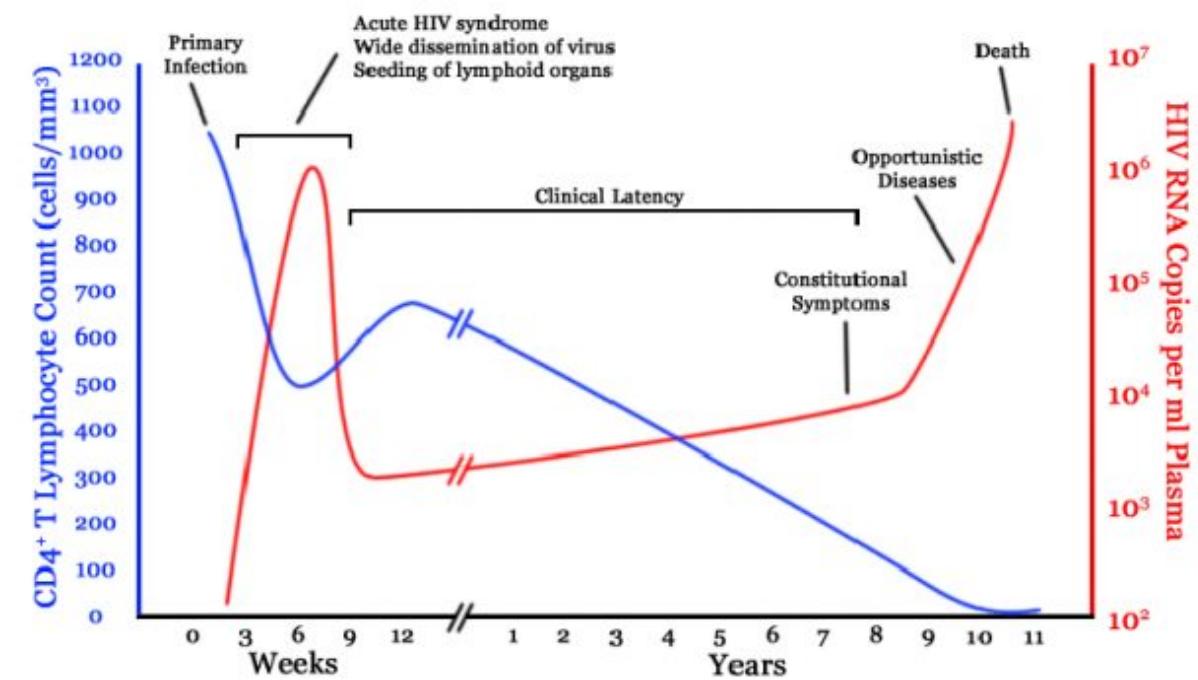
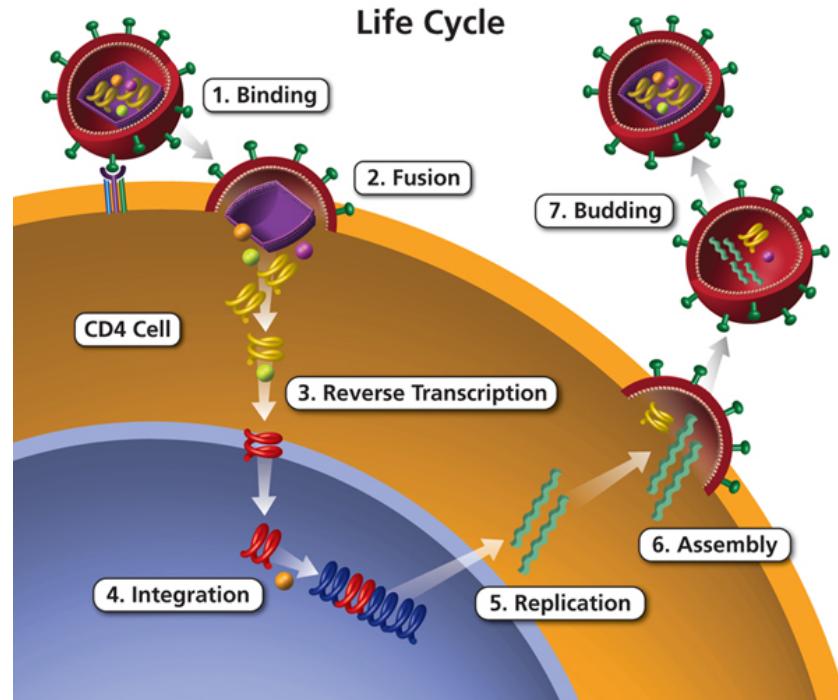


## HIV infection in a human host



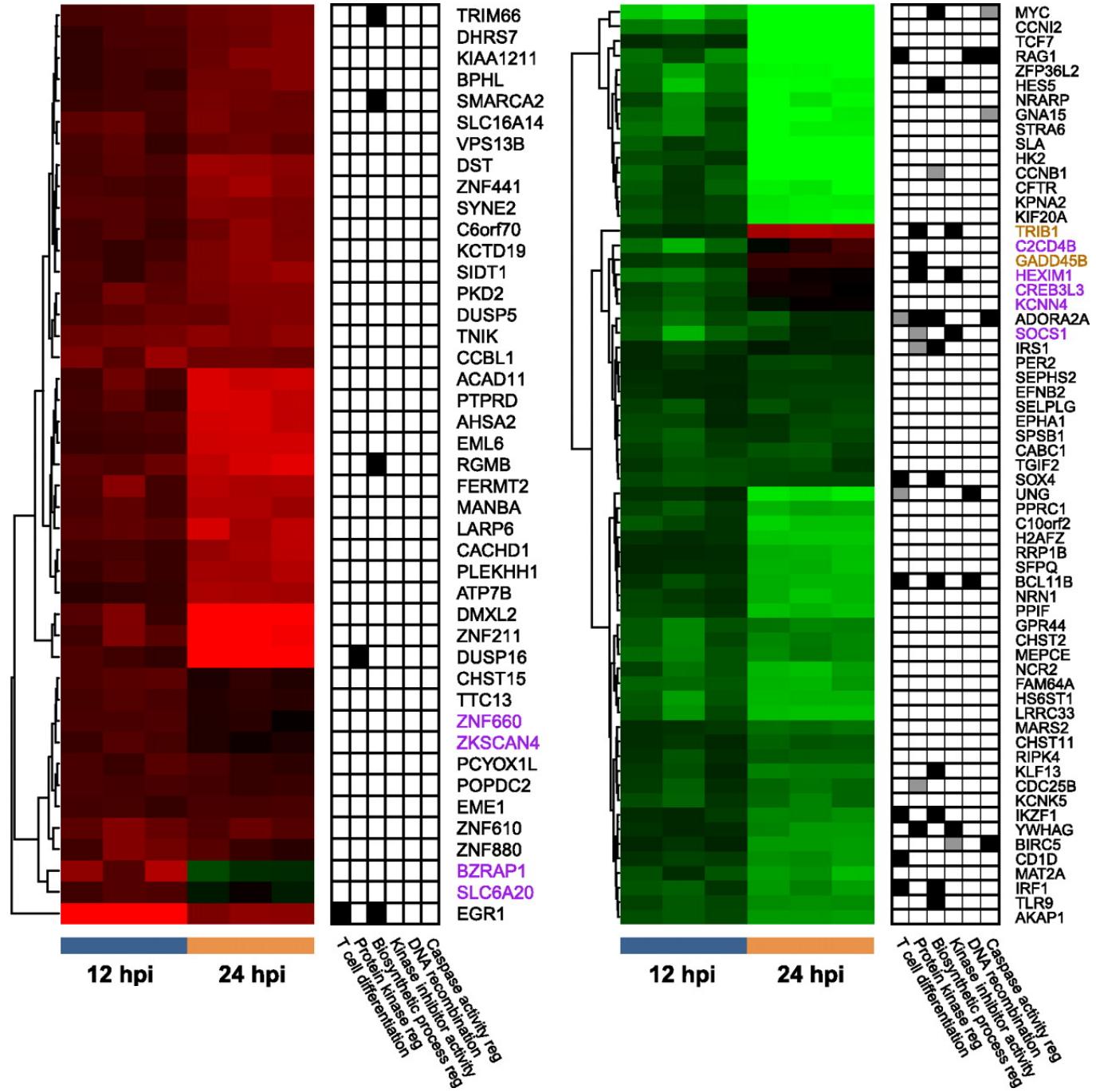
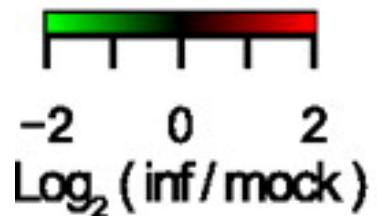
# The study question

What changes take place in the first 12-24 hours of HIV infection in terms of gene expression of host cell and viral replication levels?

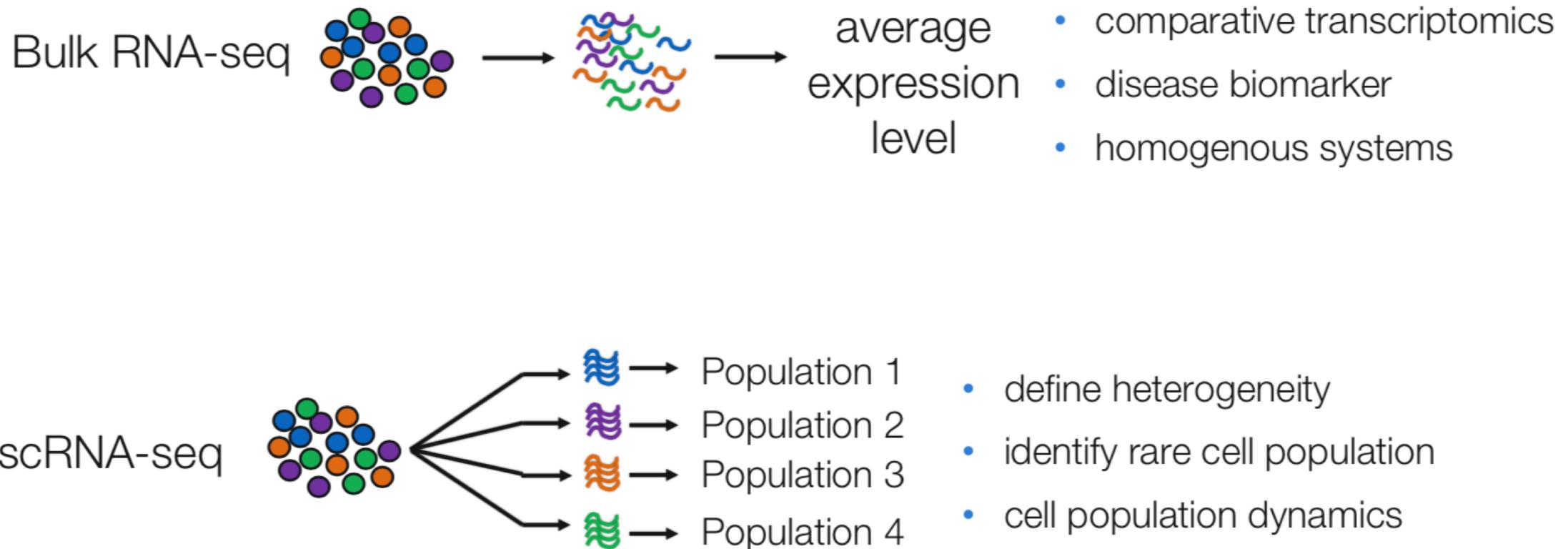


# Study findings

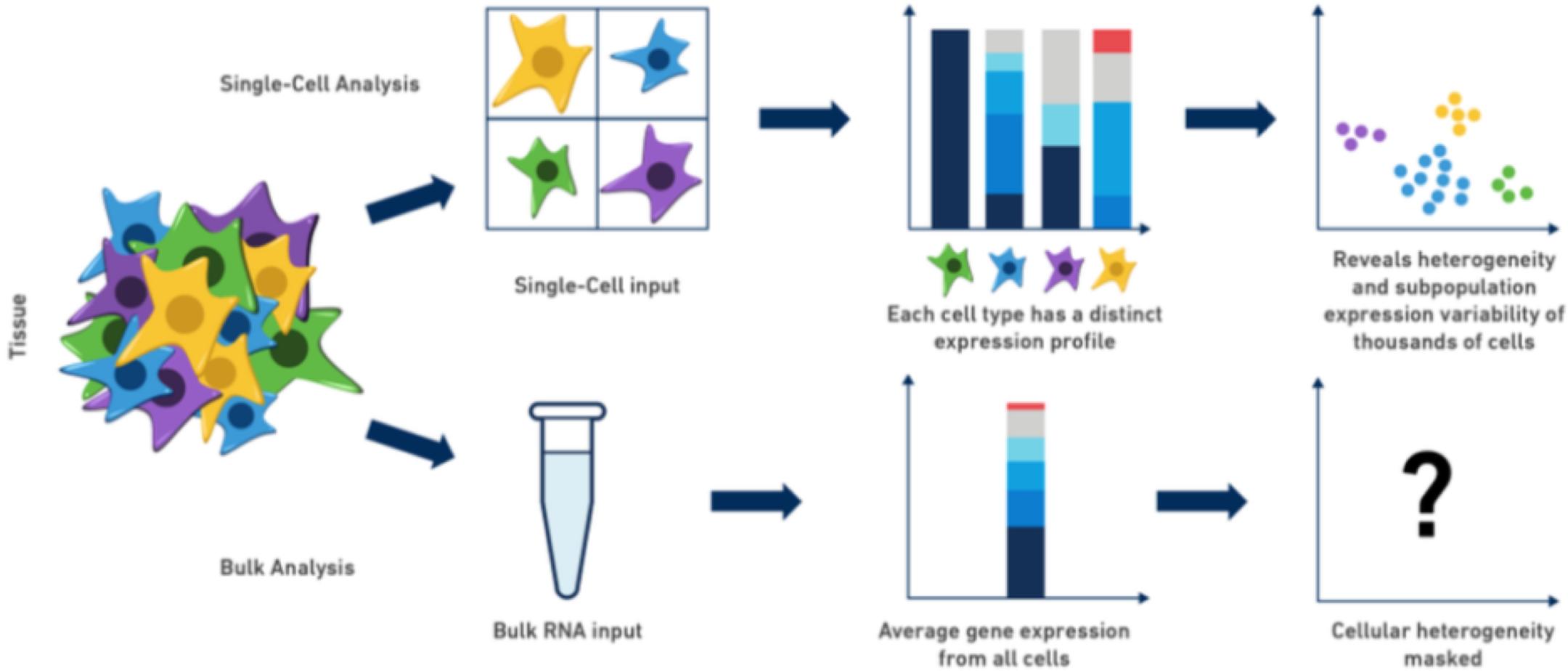
- 20% of reads mapped to HIV at 12 hr,  
40% at 24 hr
- Downregulation of T cell activation  
genes at 12 hr
- ‘Large-scale disruptions to host  
transcription’ at 24 hr



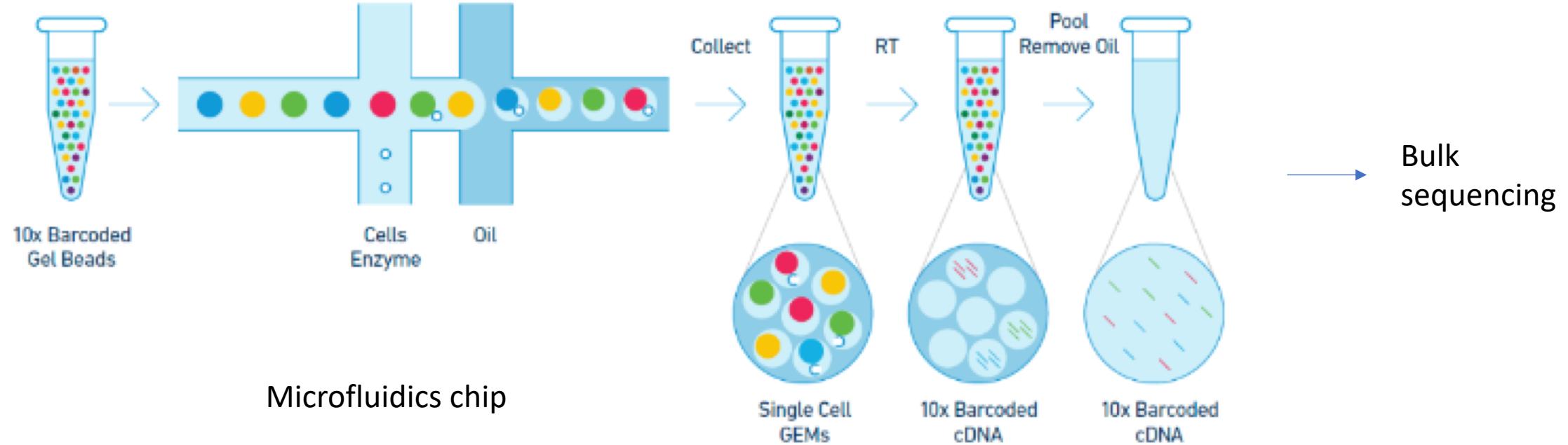
# Bulk vs Single Cell RNA Sequencing



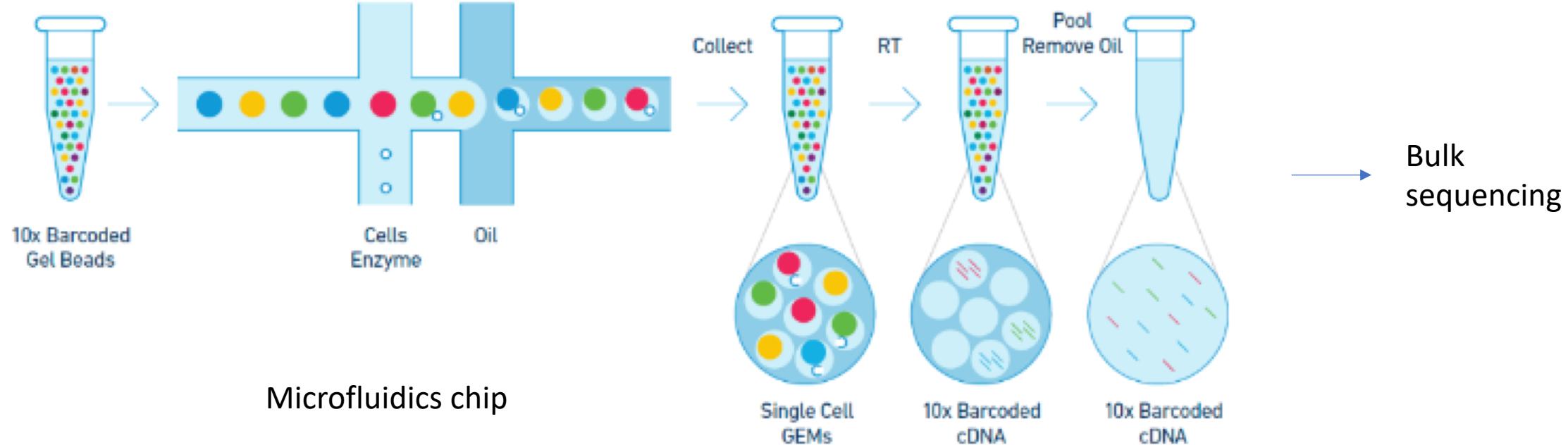
# scRNA to study Heterogeneity in cell populations



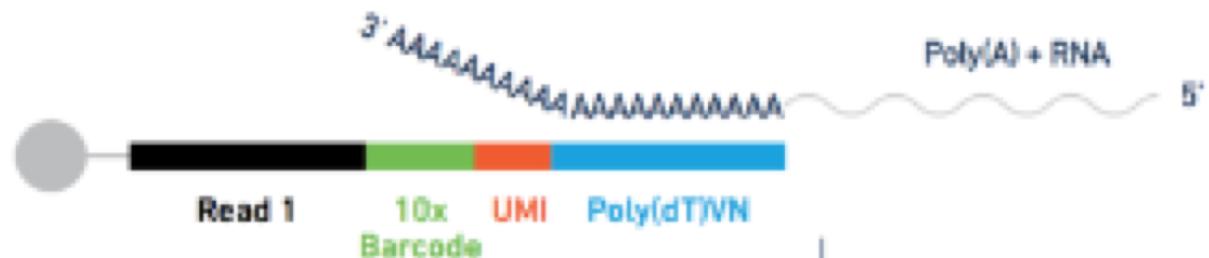
# 10x single cell technology



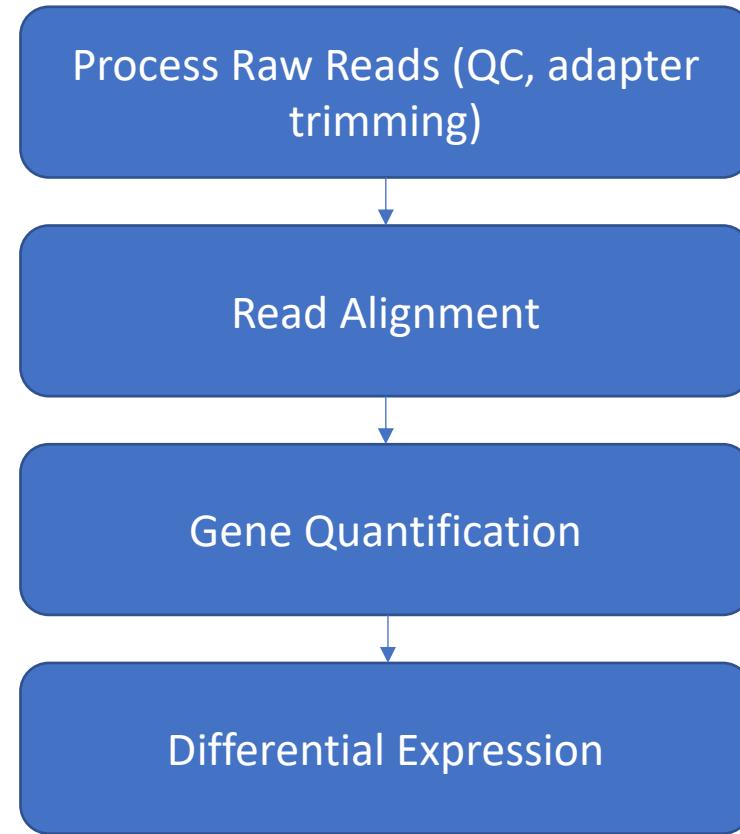
# 10x single cell technology



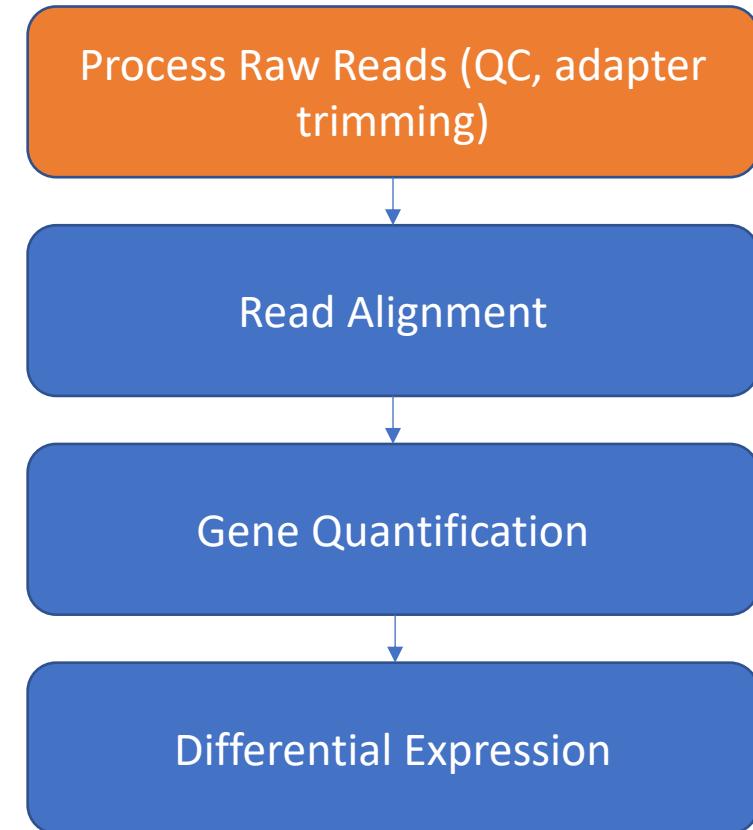
Beads are barcoded, and RT occurs inside the GEM -> all reads from a given will have the same barcode



# Our (bulk) RNAseq Workflow



# Workflow



# Raw reads in Fastq format

```
@SRR098401.109756285
GACTCACGTAACTTAAACTCTAACAGAAATATACTA...
+
CAEFGDG?BCGGGEEDGGHGHGDFHEIEGGDDDD...
```

1. Sequence identifier
2. Sequence
3. + (optionally lists the sequence identifier again)
4. Quality string

# Base Quality Scores

The symbols we see in the read quality string are an encoding of the quality score:

Quality encoding: !#\$%&'()\*,./0123456789:;<=>?@ABCDEFGHI

| | | | |

Quality score: 0.....10.....20.....30.....40

A quality score is a prediction of the probability of an error in base calling:

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

# Base Quality Scores

The symbols we see in the read quality string are an encoding of the quality score:

Quality encoding: !#\$%&'()\*, -./0123456789:;=>?@ABCDEFGHI

| | | | | |

Quality score: 0.....10.....20.....30.....40

A quality score is a prediction of the probability of an error in base calling:

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

Back to our read:

```
@SRR098401.109756285  
GACTCACGTAACTTAAACTCTAACAGAAATATACTA...  
+  
CAEFGDG?BCGGGEEDGGHGHGDFHEIEGGDDDD...
```

C → Q = 34 → Probability < 1/1000 of an error

# Base Quality Scores

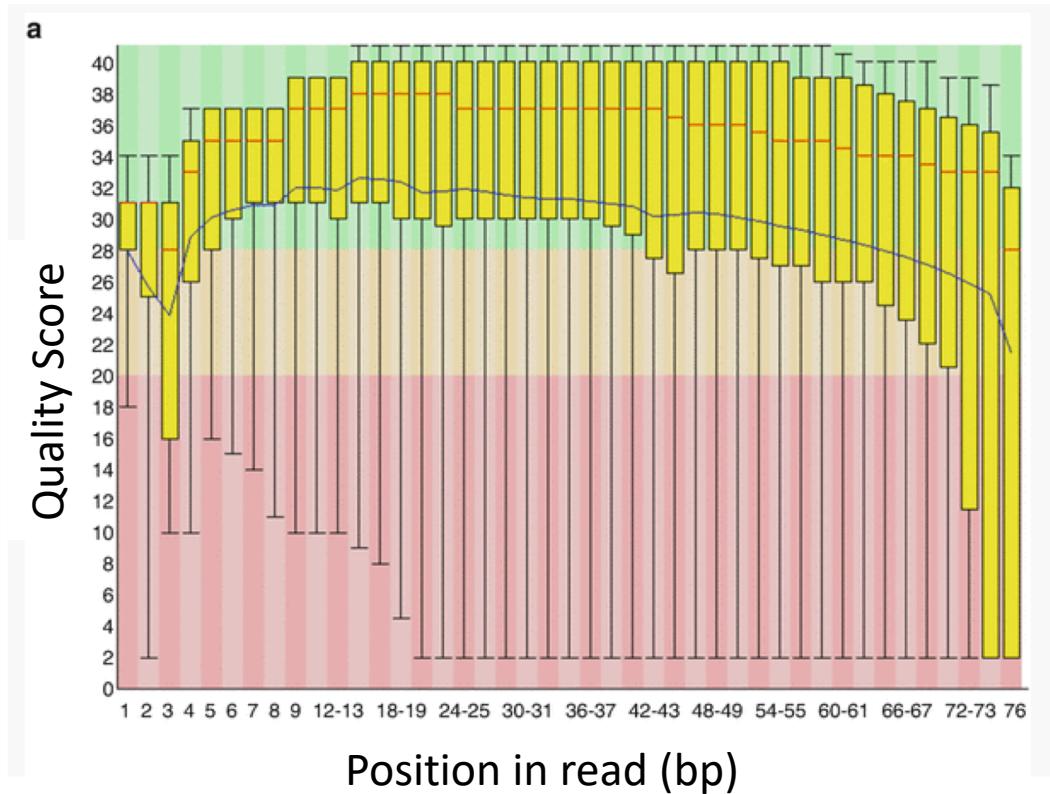
Character	Score	Raw Read Range
S	33	0 to 40
X	64	-5 to 40
I	73	0 to 40
J	104	3 to 41
L	126	0.2 to 41

**S - Sanger** Phred+33, raw reads typically (0, 40)  
**X - Solexa** Solexa+64, raw reads typically (-5, 40)  
**I - Illumina 1.3+** Phred+64, raw reads typically (0, 40)  
**J - Illumina 1.5+** Phred+64, raw reads typically (3, 41)  
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (**bold**)  
(Note: See discussion above).  
**L - Illumina 1.8+** Phred+33, raw reads typically (0, 41)

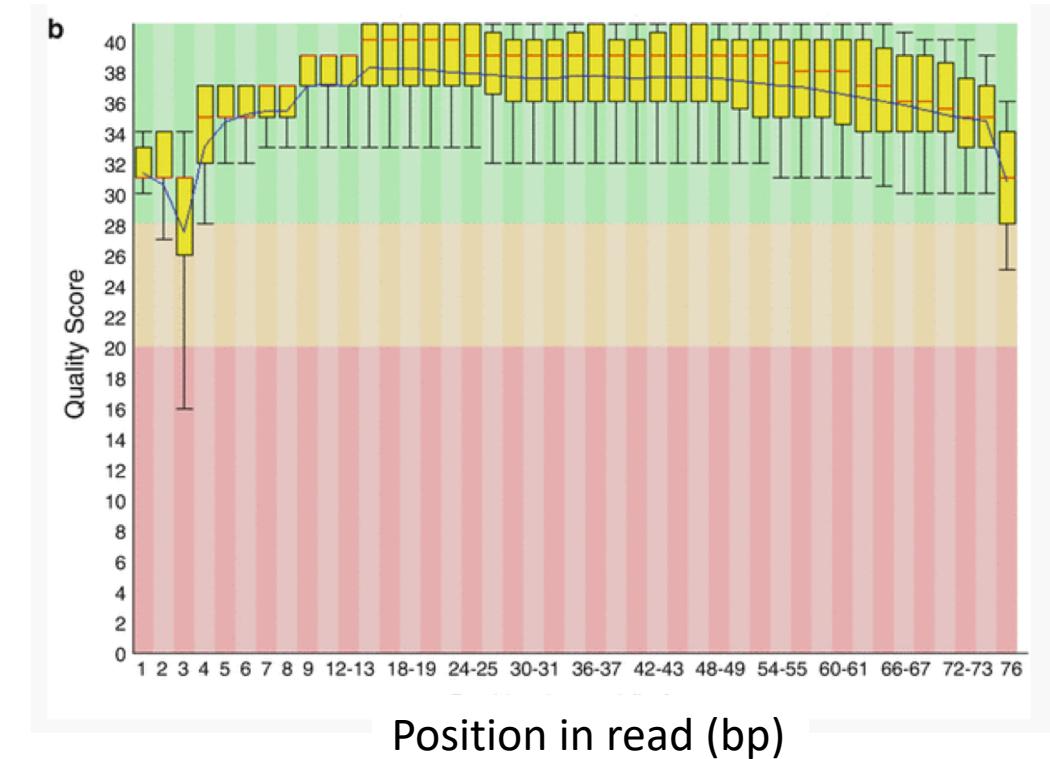
# Raw read quality control

- Quality distribution over the length of the read
- GC content
- Per base sequence content
- Adapters in Sequence

# FastQC: Sequence Quality Histogram



**GOOD**  
High quality over the length of the read

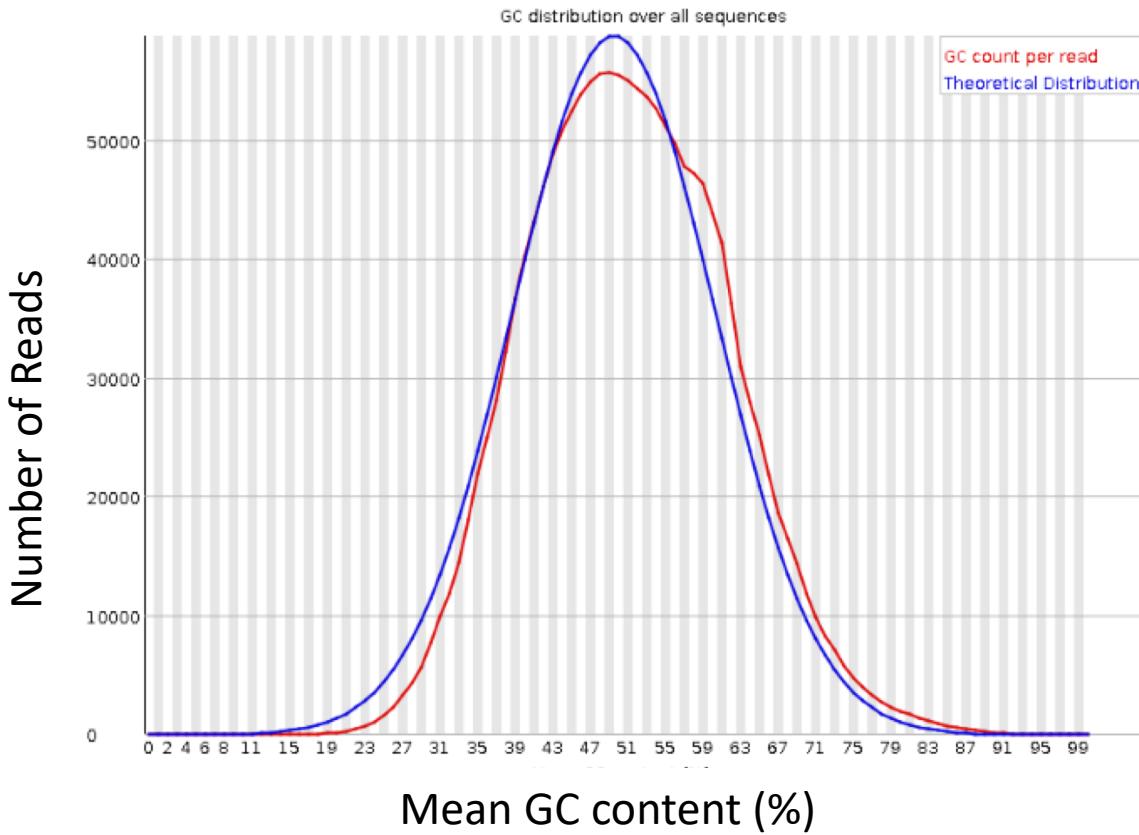


**BAD**  
Read quality drops at the beginning and end



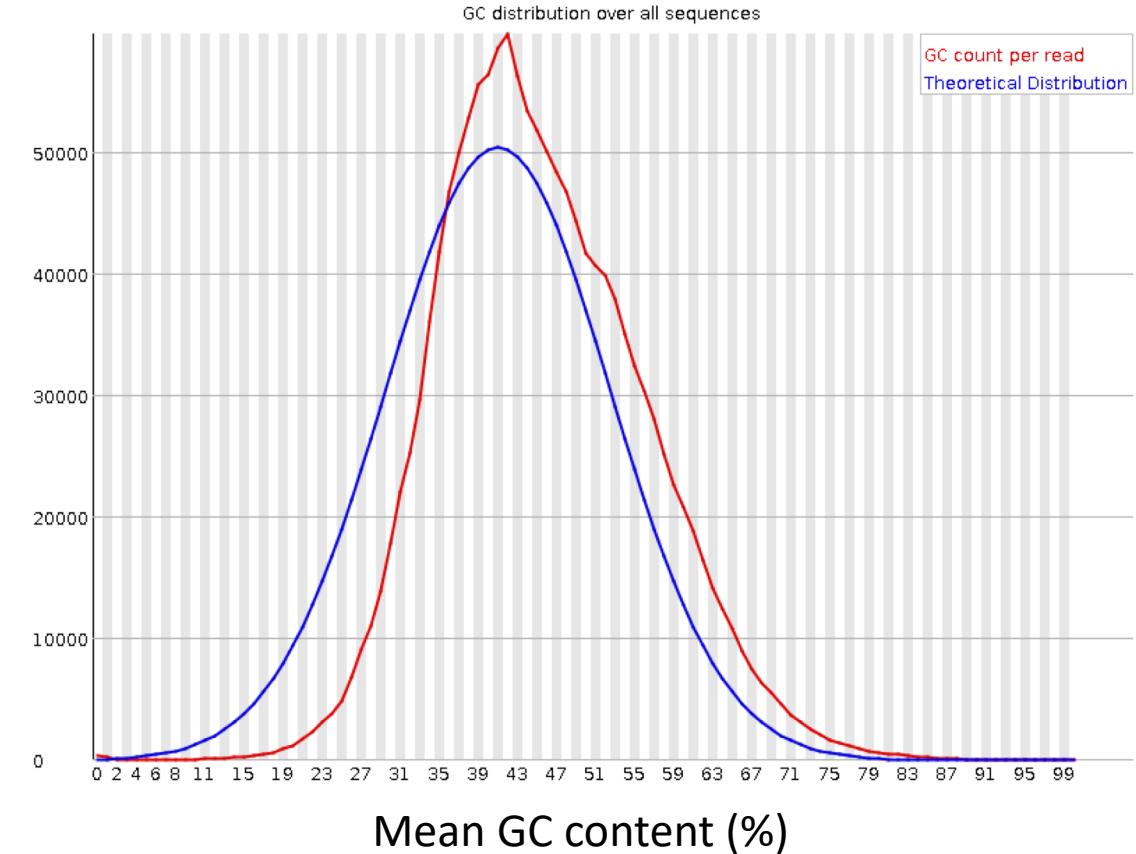
# FastQC: Per sequence GC content

## Per sequence GC content



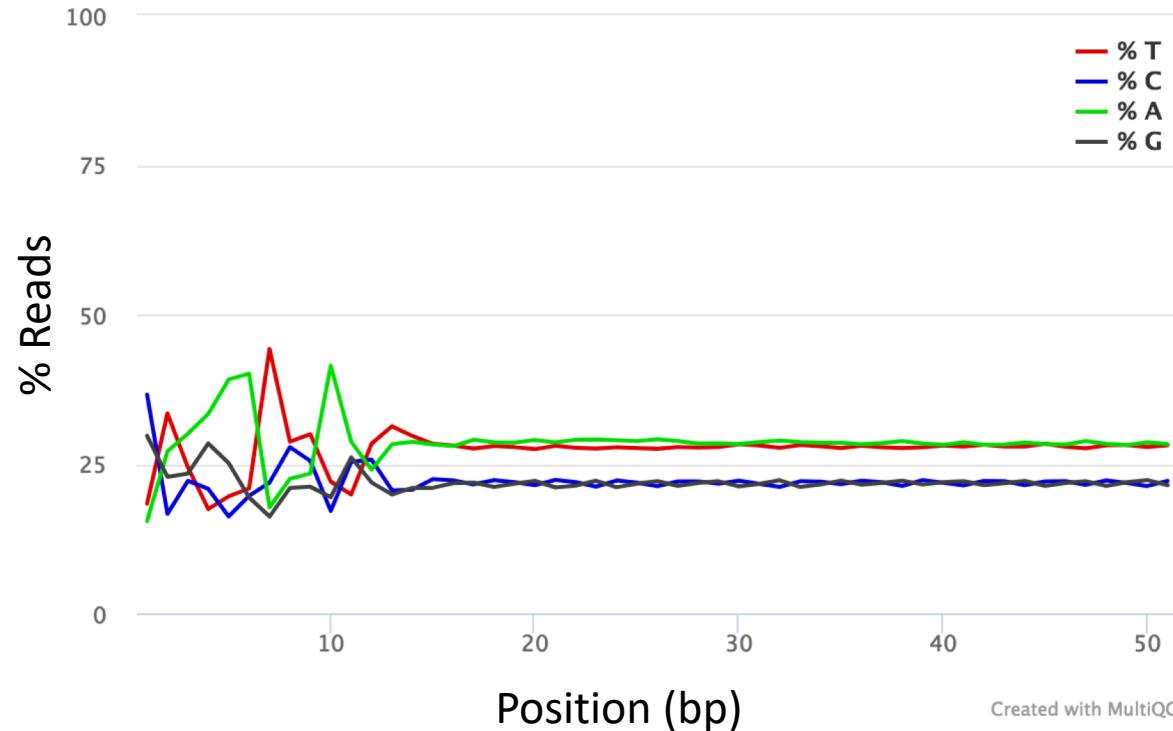
GOOD: follows normal distribution (sum of deviations is < 15% of reads)

## Per sequence GC content



BAD: can indicate contamination with adapter dimers, or another species

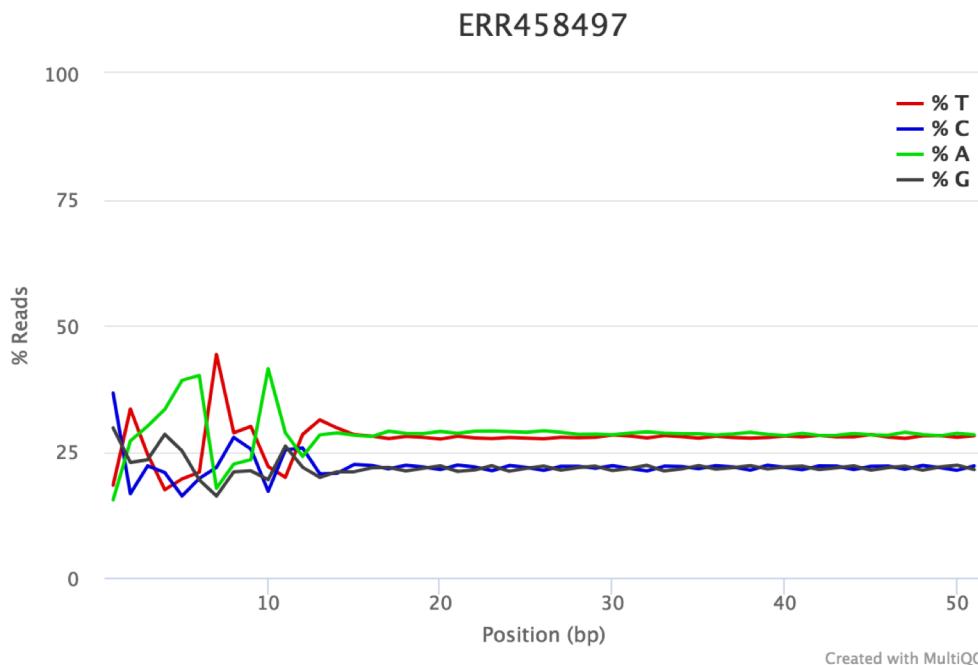
# FastQC: Per Base Sequence Content



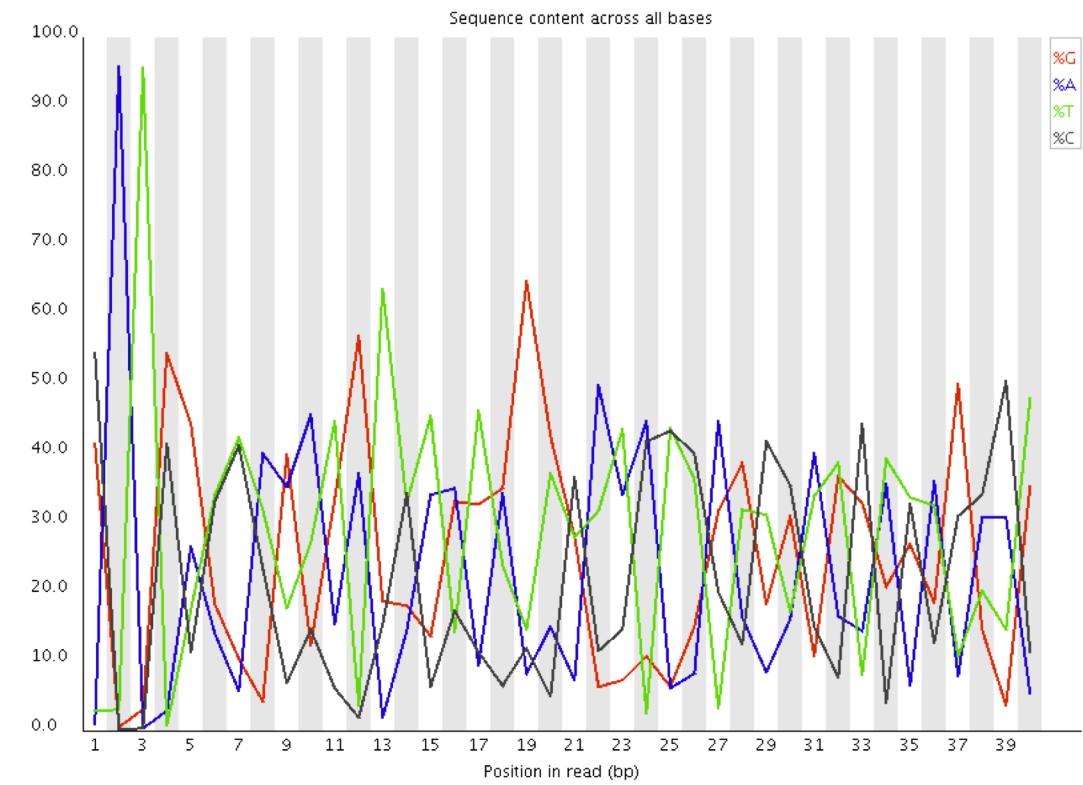
## EXPECTED for RNAseq

- Proportion of each position for which each DNA base has been called
  - RNAseq data tends to show a positional sequence bias in the first ~12 bases
  - The "random" priming step during library construction is not truly random and certain hexamers are more prevalent than others
  - Studies have shown that this does NOT cause mis-called bases or drastic bias in sequenced fragments
- Read quality drops at the beginning and end

# FastQC: Per Base Sequence Content



EXPECTED



BAD:

Shows a strong positional bias throughout the reads, which in this case is due to the library having a certain sequence that is overrepresented

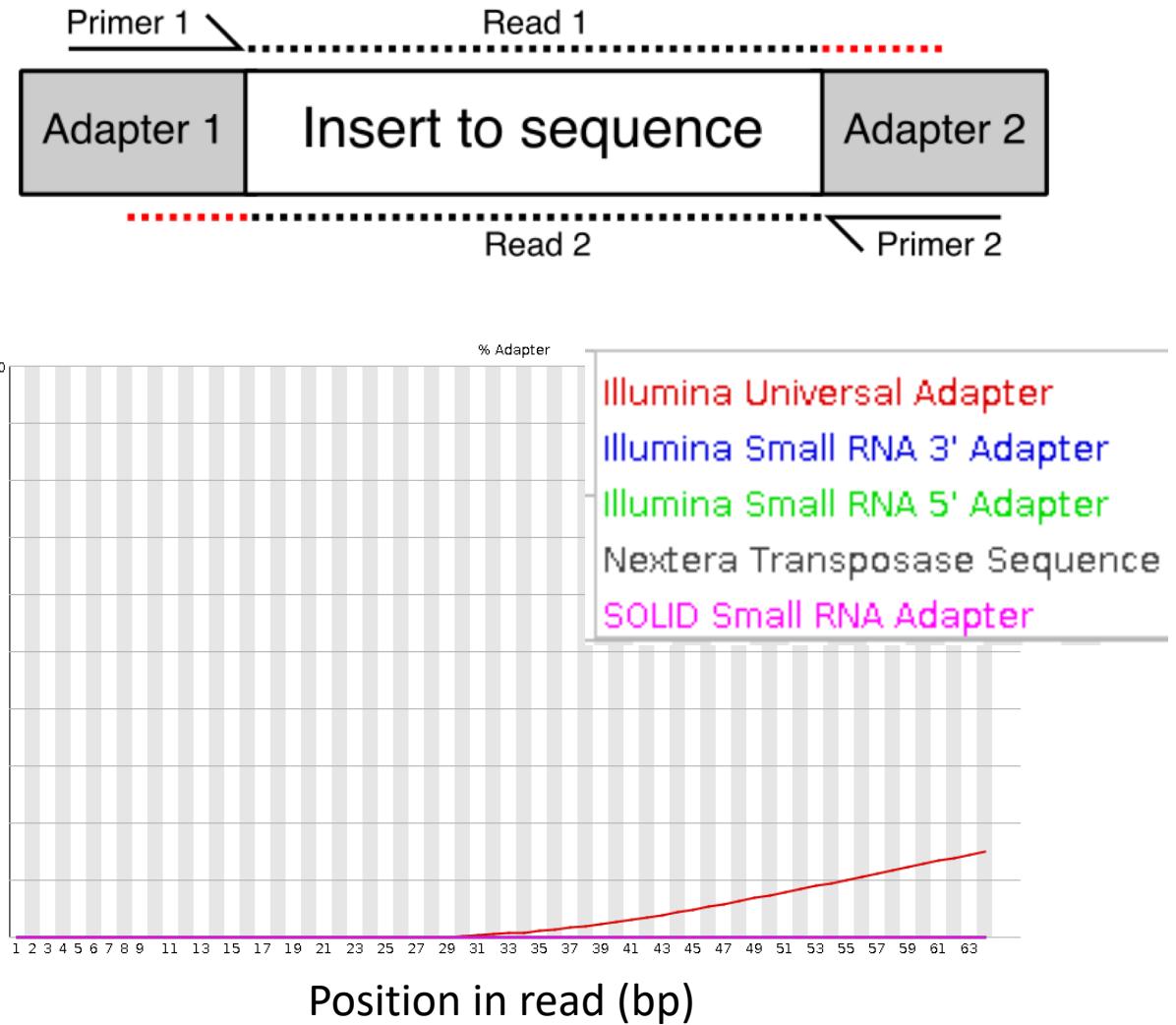
# FastQC: Adapter content

The cause: The “insert” sequence is shorter than the read, and the read contains part of the adapter sequence

FastQC will scan each read for the presence of known adapter sequences

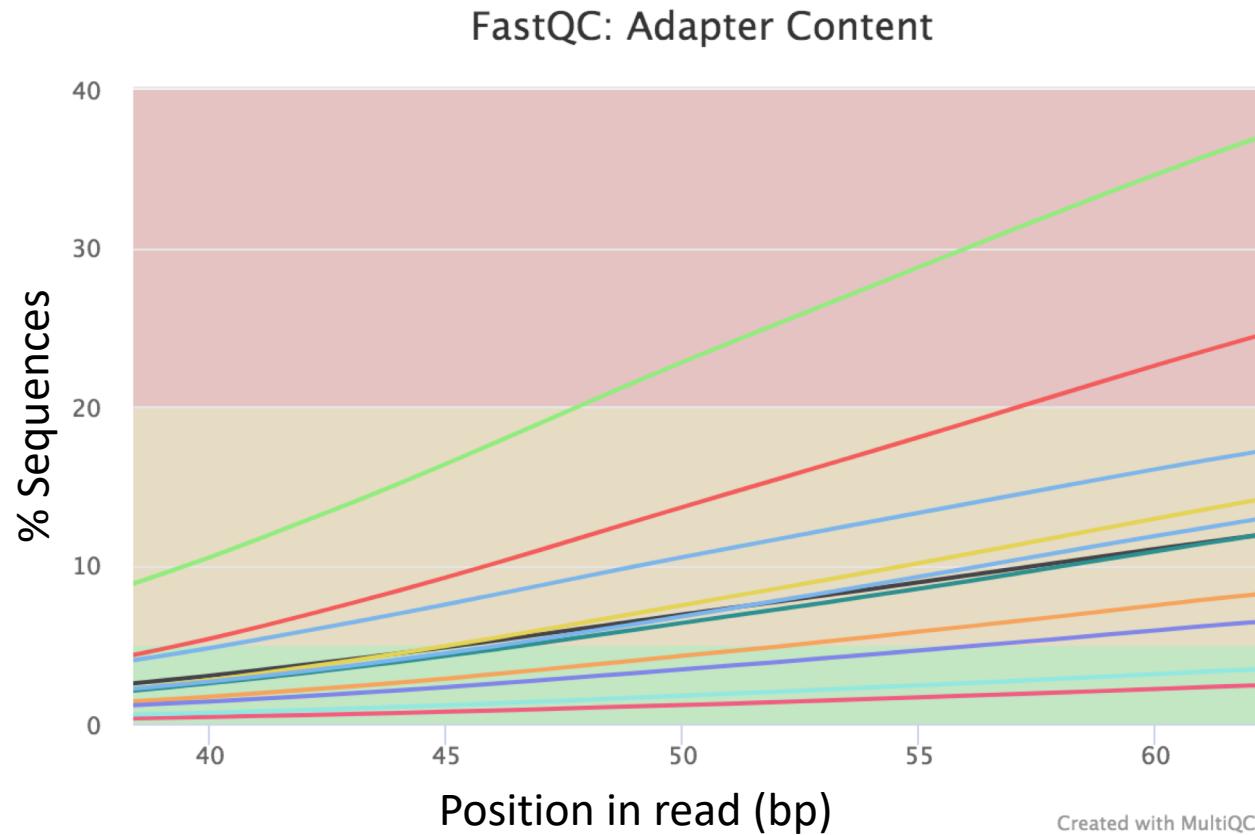
The plot shows that the adapter content rises over the course of the read

Solution – Adapter trimming!



# FastQC -> MultiQC

Should view all samples at once to notice abnormalities for our dataset.

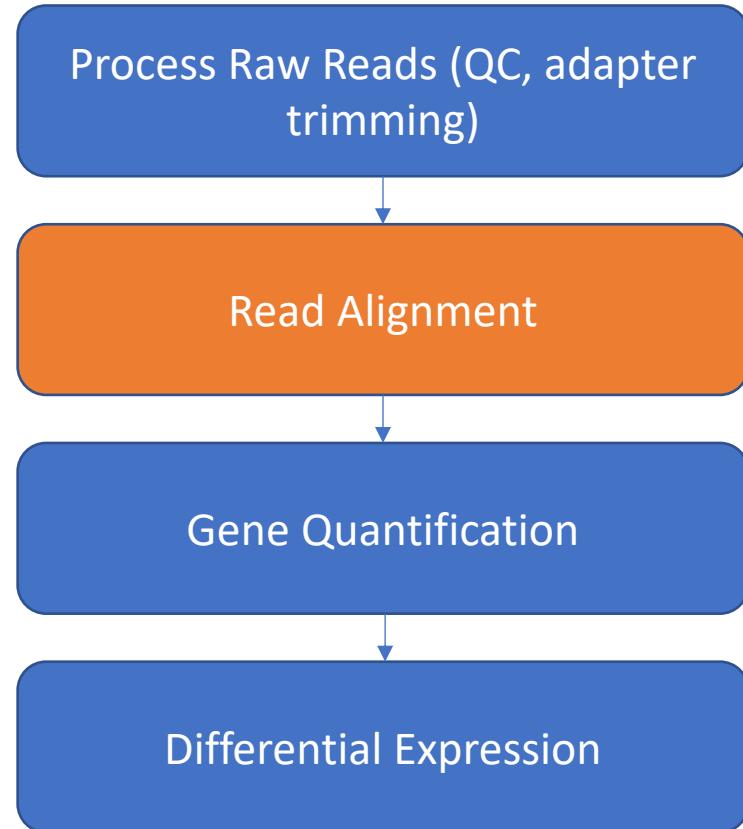


# Adapter trimming

Trim Galore! is a tool that:

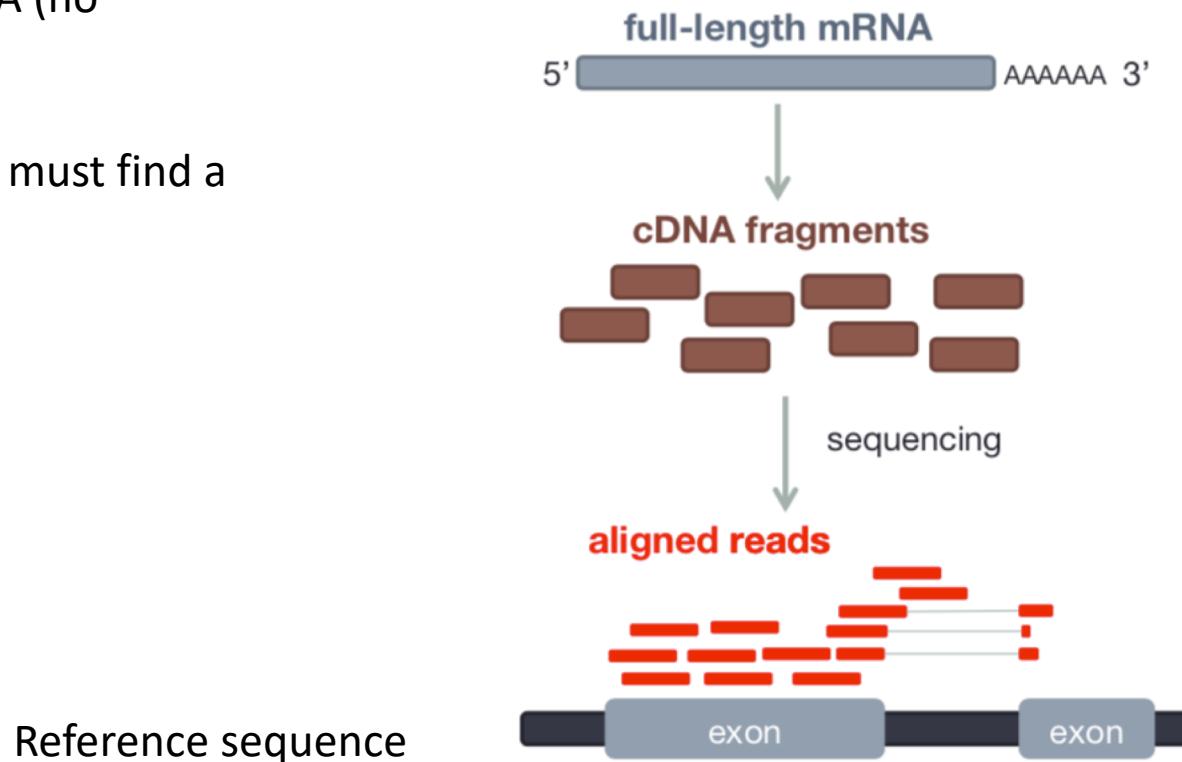
- Scans and removes known Illumina or custom adapters
- Performs read trimming for low quality regions at the end of reads
- Removes reads that become too short in the trimming process

# Workflow



# Read Alignment

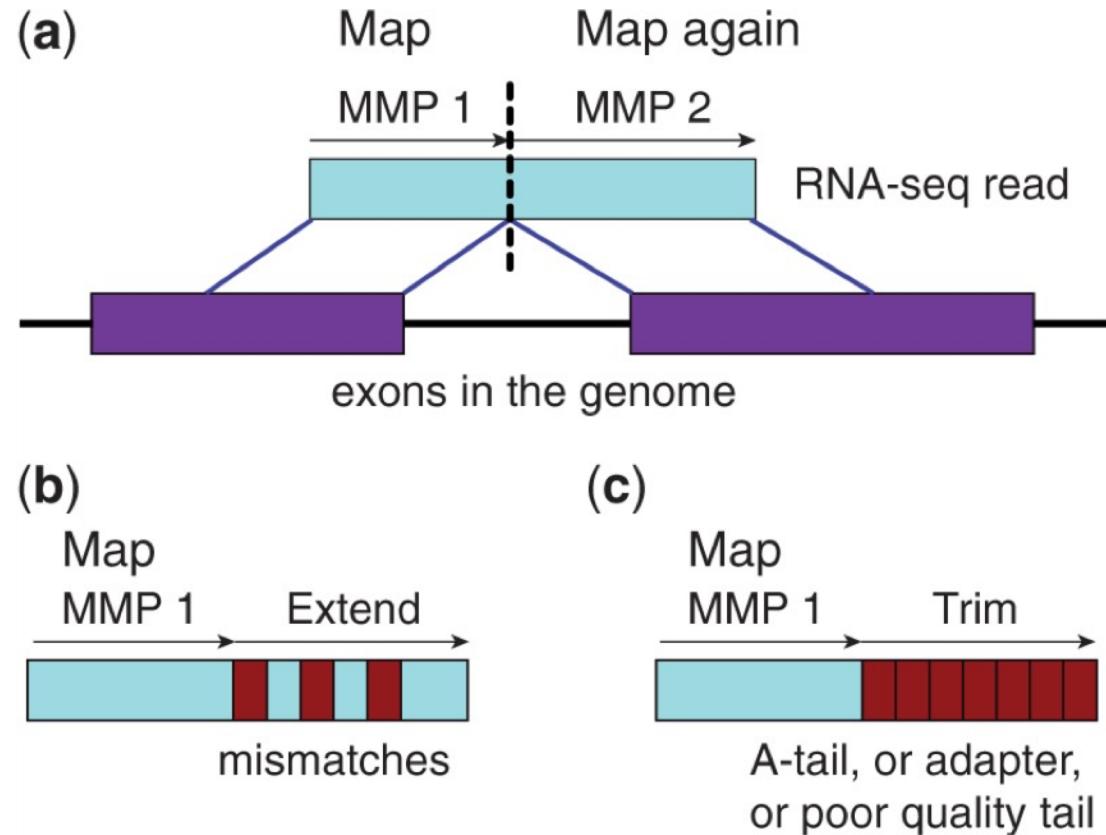
- RNAseq data originates from spliced mRNA (no introns)
- When aligning to the genome, our aligner must find a spliced alignment for reads



# STAR Aligner (Spliced Transcripts Alignment to a Reference)

Highly accurate, memory intensive aligner  
Two phase mapping process

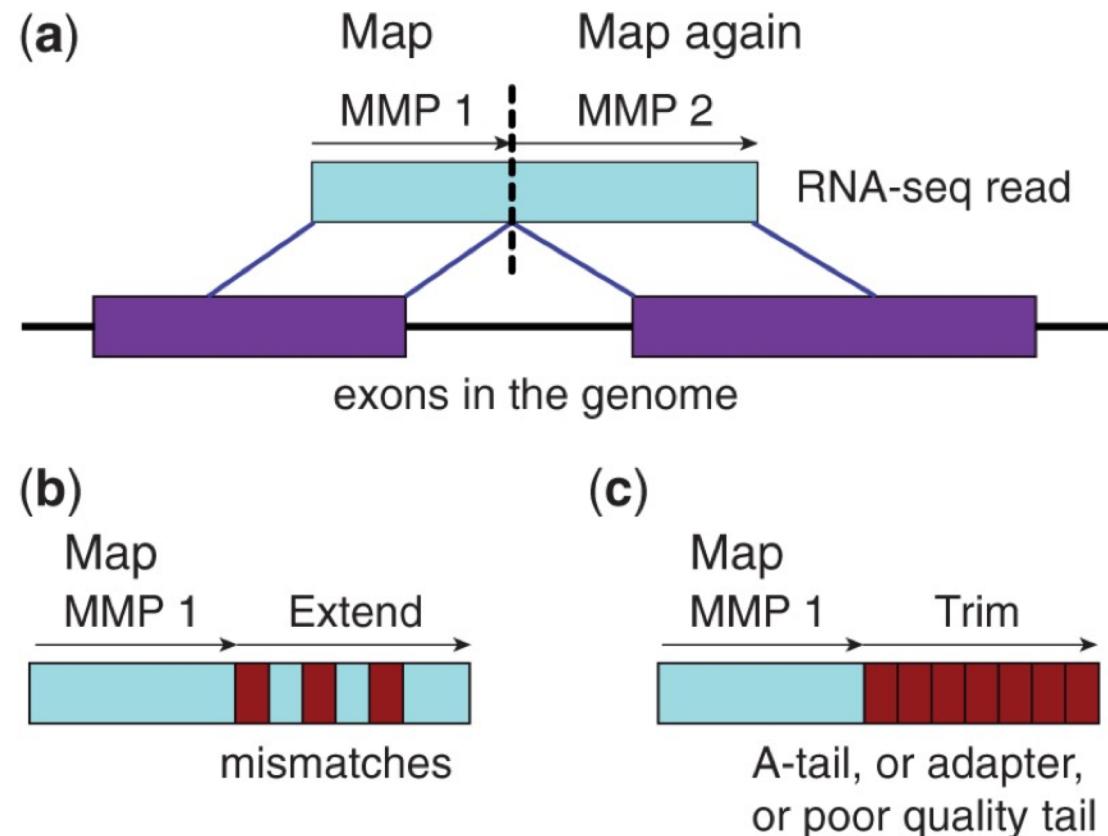
1. Find Maximum Mappable Prefixes (MMP) in a read. MMP can be extended by
  - mismatches
  - Indels
  - soft-clipping



# STAR Aligner (Spliced Transcripts Alignment to a Reference)

Highly accurate, memory intensive **aligner**  
Two phase mapping process

1. Find Maximum Mappable Prefixes (MMP) in a read. MMP can be extended by
  - mismatches
  - Indels
  - soft-clipping
2. Clustering MMP, stitching and scoring to determine final read location



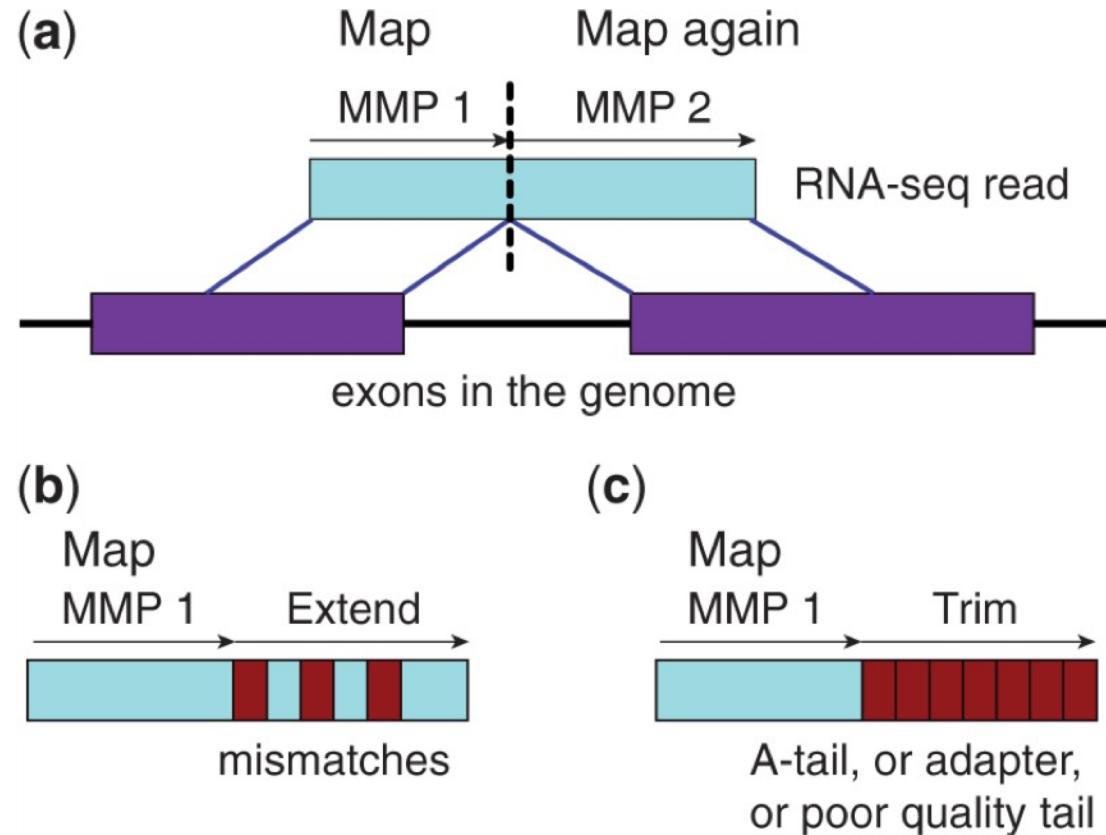
# STAR Aligner (Spliced Transcripts Alignment to a Reference)

Highly accurate, memory intensive **aligner**  
Two phase mapping process

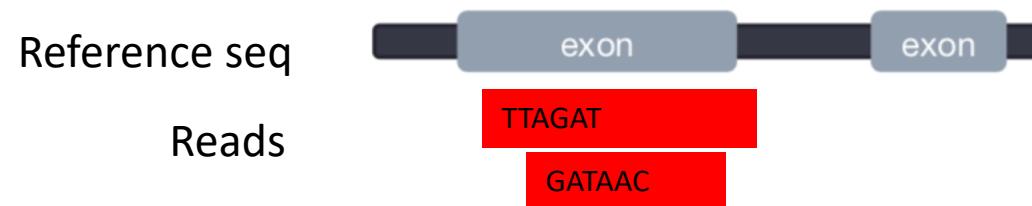
1. Find Maximum Mappable Prefixes (MMP) in a read. MMP can be extended by
  - mismatches
  - Indels
  - soft-clipping

2. Clustering MMP, stitching and scoring to determine final read location

Output is a Sequence Alignment Map (SAM) file



# Sequence Alignment Map (SAM)



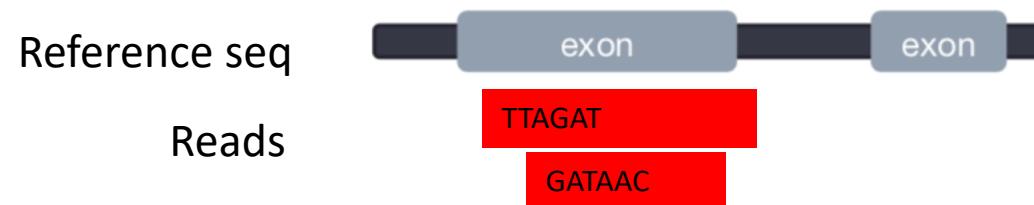
```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAACGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

Header section

Alignment section

↑ CIGAR: summary of alignment, e.g. match, gap, insertion, deletion  
↑ Mapping Quality  
↑ Position  
↑ Ref Sequence name  
↑ Flag: indicates alignment information e.g. paired, aligned, etc  
<https://broadinstitute.github.io/picard/explain-flags.html>

# Sequence Alignment Map (SAM)



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAACGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

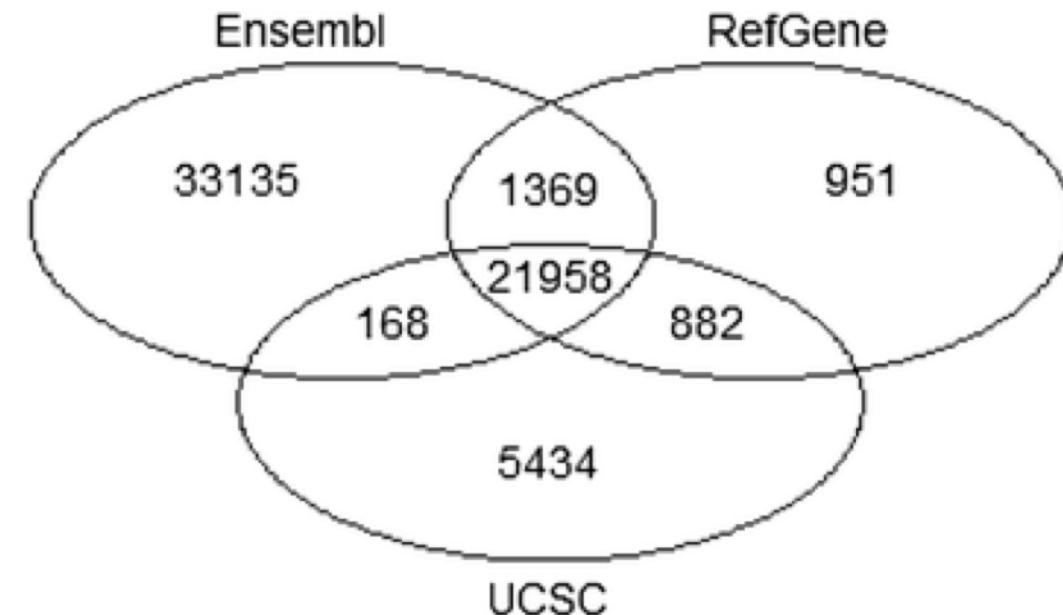
Header section

Alignment section

Paired end info  
↑  
Sequence  
↑  
Quality Score  
↑  
Optional Fields

# Genome Annotation Standards

- STAR can use an annotation file gives the location and structure of genes in order to improve alignment in known splice junctions
- Annotation is dynamic and there are at least three major sources of annotation
- The intersection among RefGene, UCSC, and Ensembl annotations shows high overlap. RefGene has the fewest unique genes, while more than 50% of genes in Ensembl are unique
- Be consistent with your choice of annotation source!

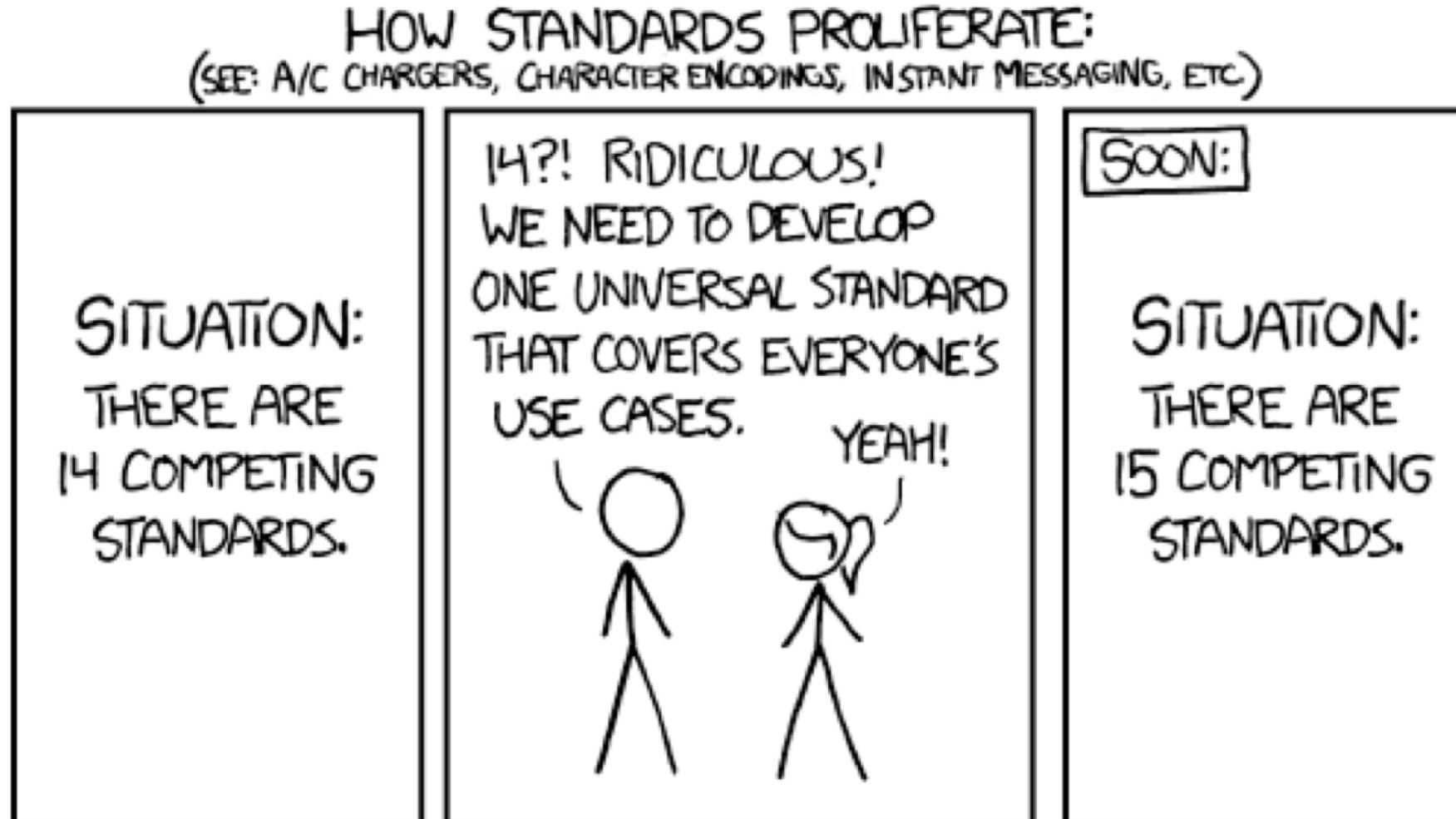


# Gene Annotation Format (GTF)

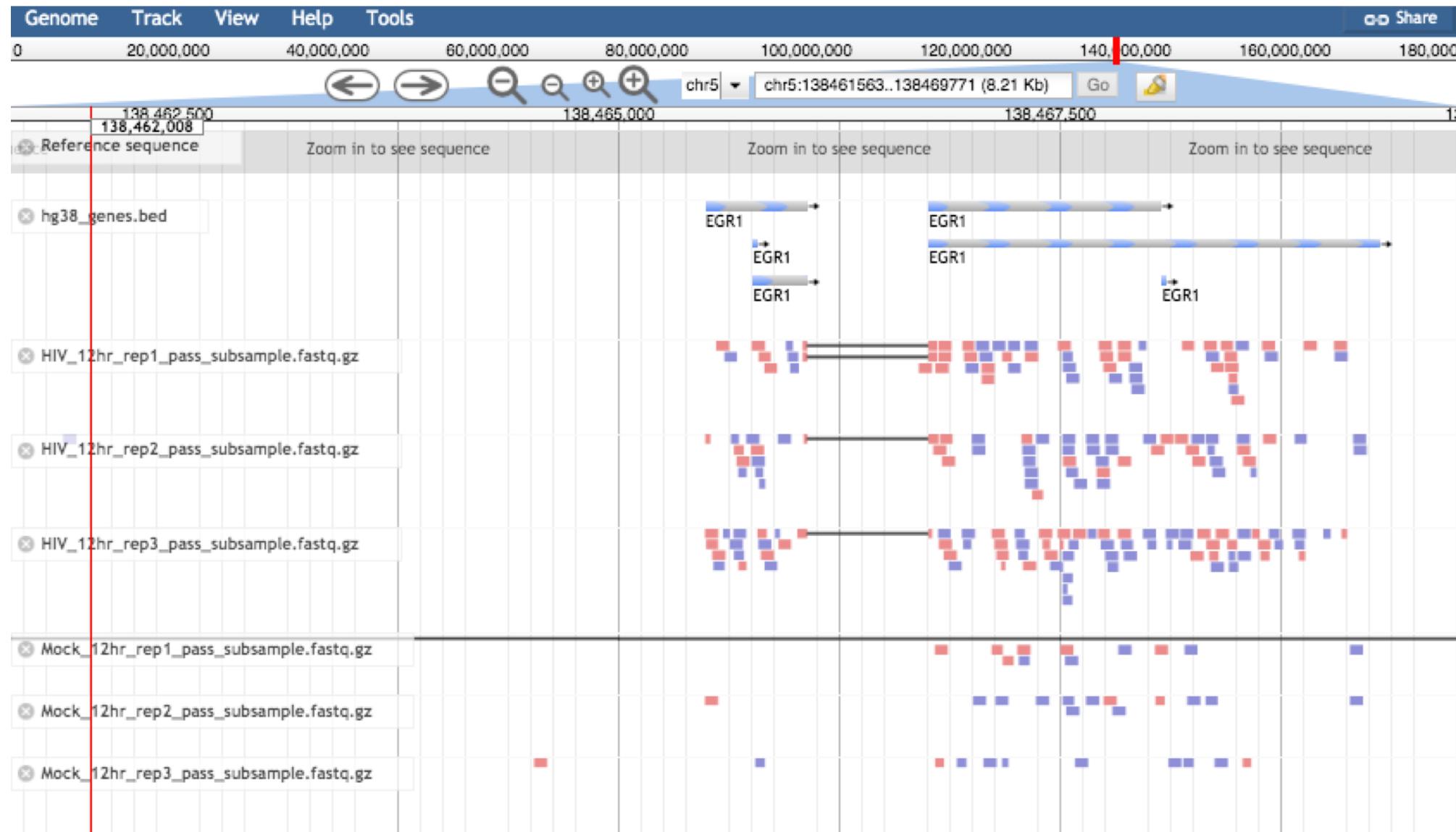
In order to count genes, we need to know where they are located in the reference sequence  
STAR uses a Gene Transfer Format (GTF) file for gene annotation

Chrom	Source	Feature type	Start	Stop	(Score)	Strand	Frame	Attribute
chr5	hg38_refGene	exon	138465492	138466068	.	+	.	gene_id "EGR1";
chr5	hg38_refGene	CDS	138465762	138466068	.	+	0	gene_id "EGR1";
chr5	hg38_refGene	start_codon	138465762	138465764	.	+	.	gene_id "EGR1";
chr5	hg38_refGene	CDS	138466757	138468078	.	+	2	gene_id "EGR1";
chr5	hg38_refGene	exon	138466757	138469315	.	+	.	gene_id "EGR1";
chr5	hg38_refGene	stop_codon	138468079	138468081	.	+	.	gene_id "EGR1";

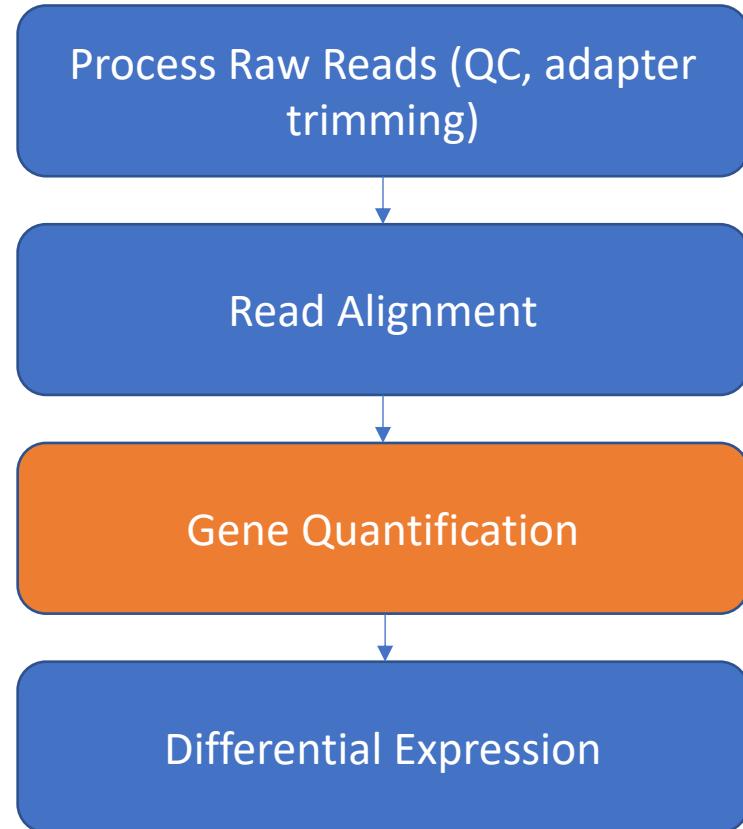
# A note on standards



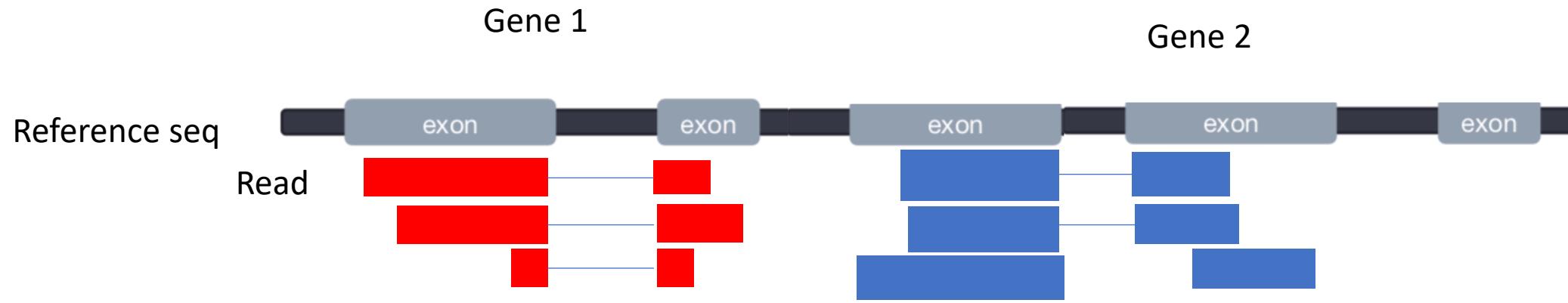
# Visualizing reads with JBrowse



# Workflow

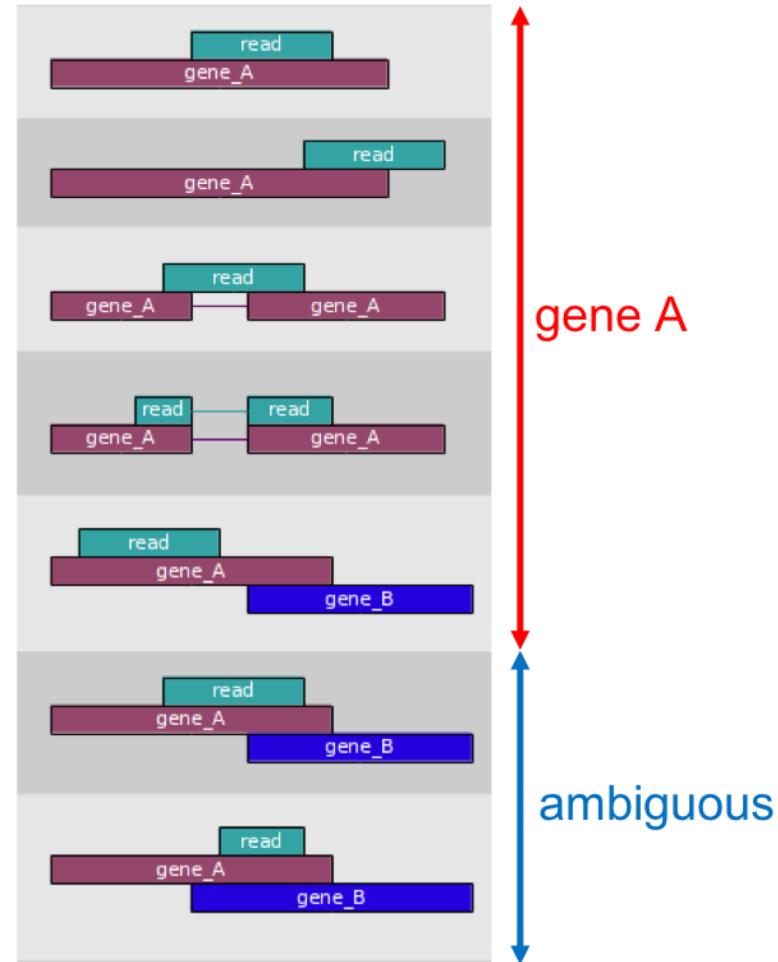


# Counting reads for each gene



# Counting reads: featurecounts

- The mapped coordinates of each read are compared with the features in the GTF file
- Reads that overlap with a gene by  $\geq 1$  bp are counted as belonging to that feature
- Ambiguous reads will be discarded
- Output will be a matrix of genes and samples

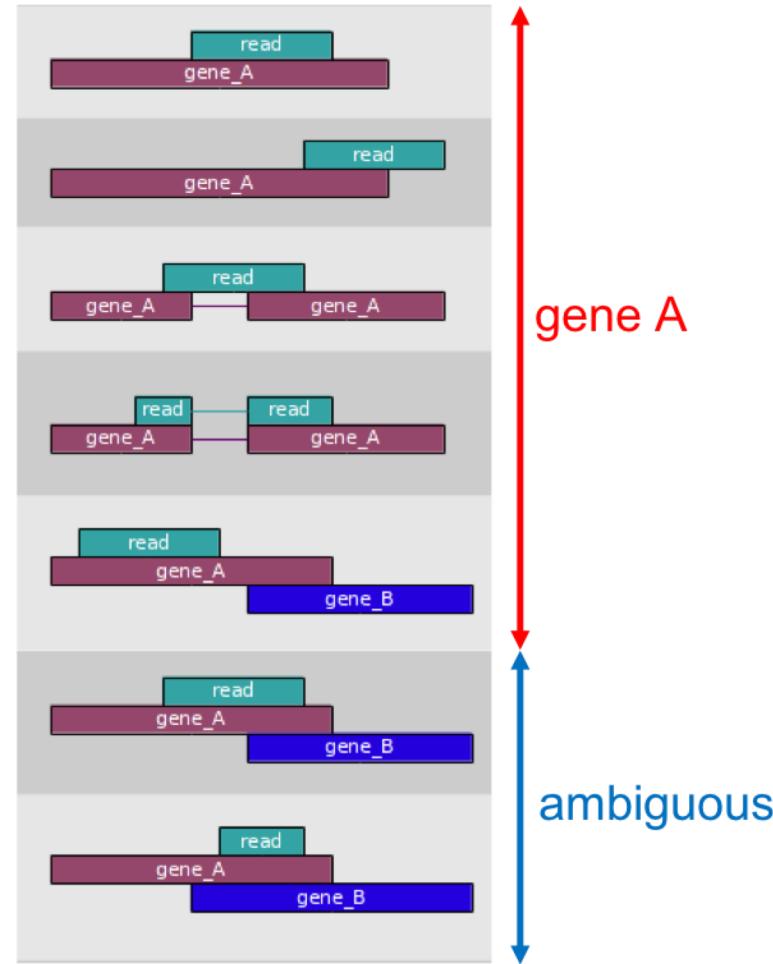


# Counting reads: featurecounts

- The mapped coordinates of each read are compared with the features in the GTF file
- Reads that overlap with a gene by  $\geq 1$  bp are counted as belonging to that feature
- Ambiguous reads will be discarded
- Output will be a matrix of genes and samples

Result is a gene count matrix:

Gene	Sample 1	Sample 2	Sample 3	Sample 4
A	1000	1000	100	10
B	10	1	5	6
C	10	1	10	20

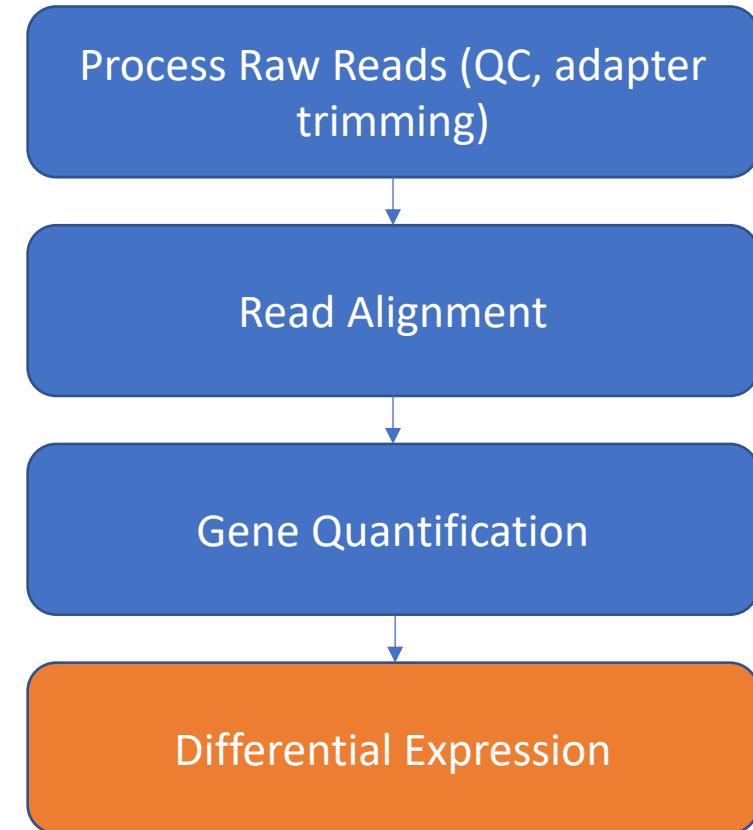


# Tracking read numbers

Revisit quality control after each processing step!

Number of Reads	Source	Result
Raw reads	FastQC run 1	8 M
After Trimming	FastQC run 2	7.1 M
Aligned to genome	STAR log	6 M
Associated with genes	FeatureCounts log	5.4 M

# Workflow

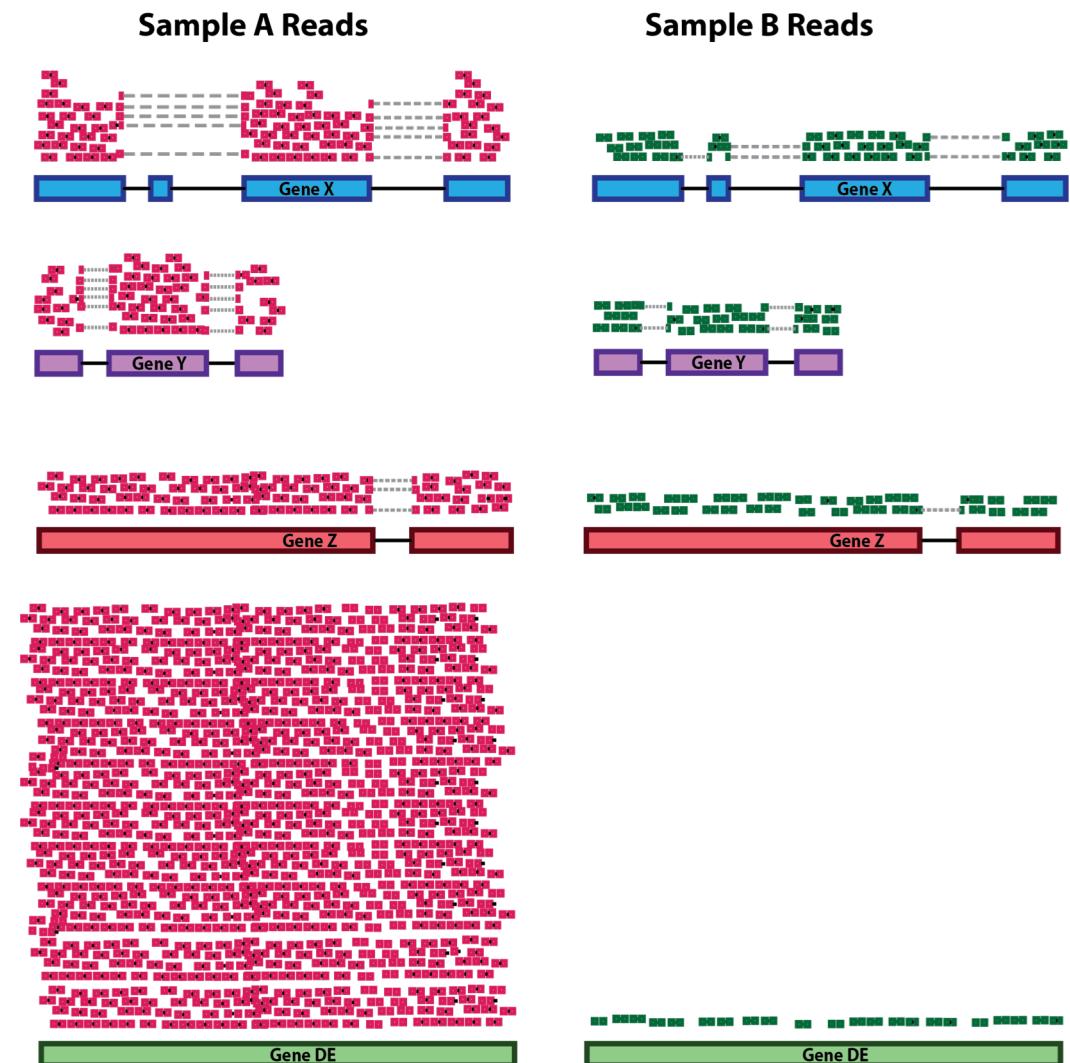


# Testing for Differential Expression

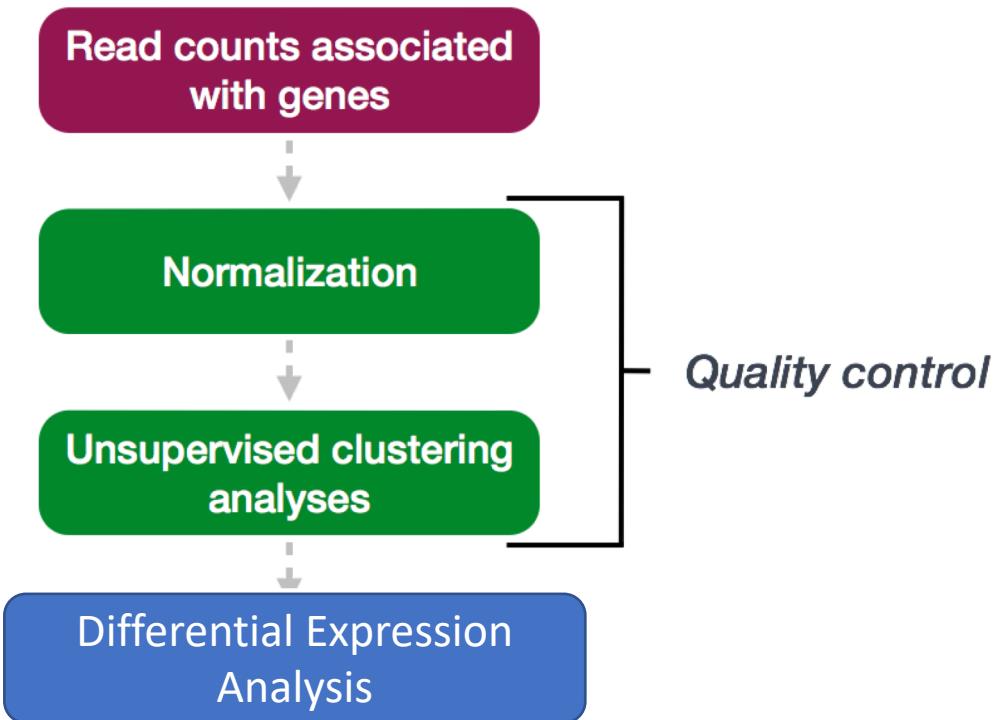
The goal of differential expression analysis (DE) is to find gene differences between conditions, developmental stages, treatments etc.

In particular DE has two goals:

- Estimate the *magnitude* of expression differences;
- Estimate the *significance* of expression differences.



# Differential Expression with DESeq2



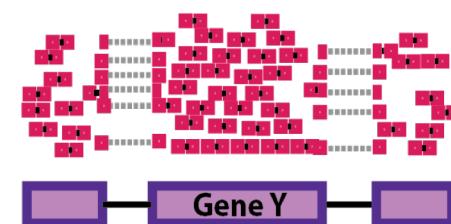
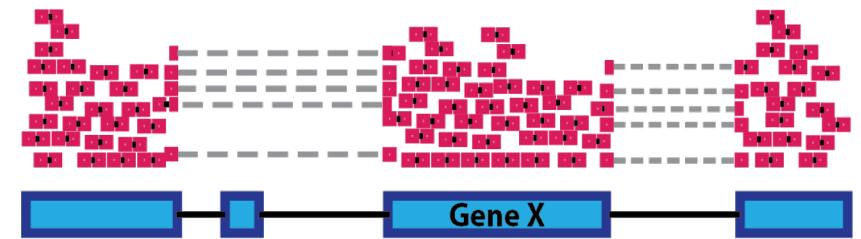
All steps are done with one click in Galaxy!

# Normalization

The number of sequenced reads mapped to a gene depends on

- Gene Length

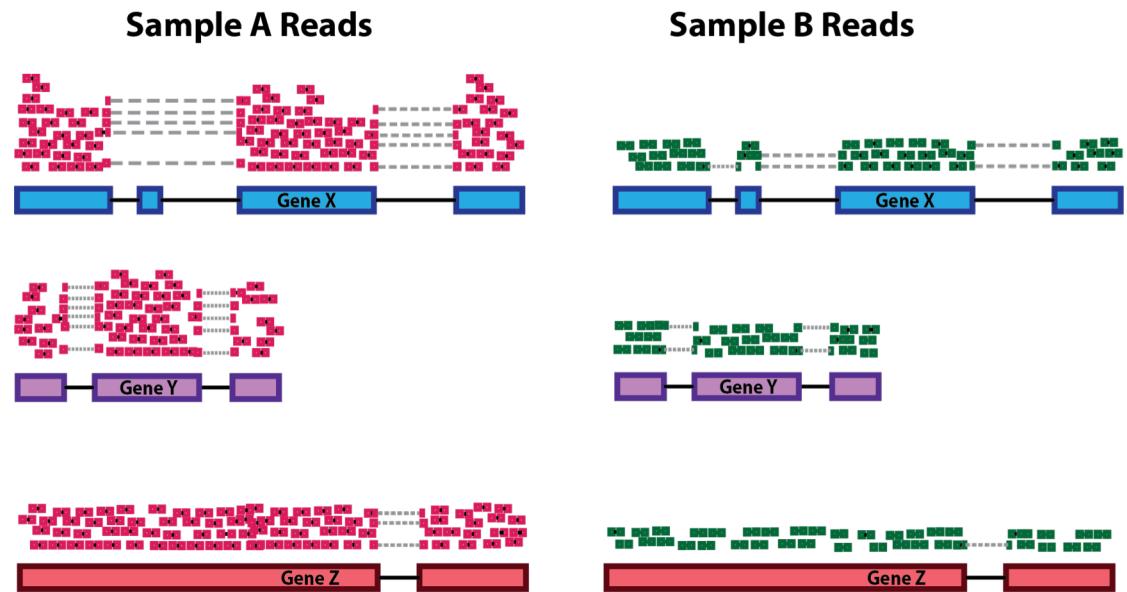
**Sample A Reads**



# Normalization

The number of sequenced reads mapped to a gene depends on

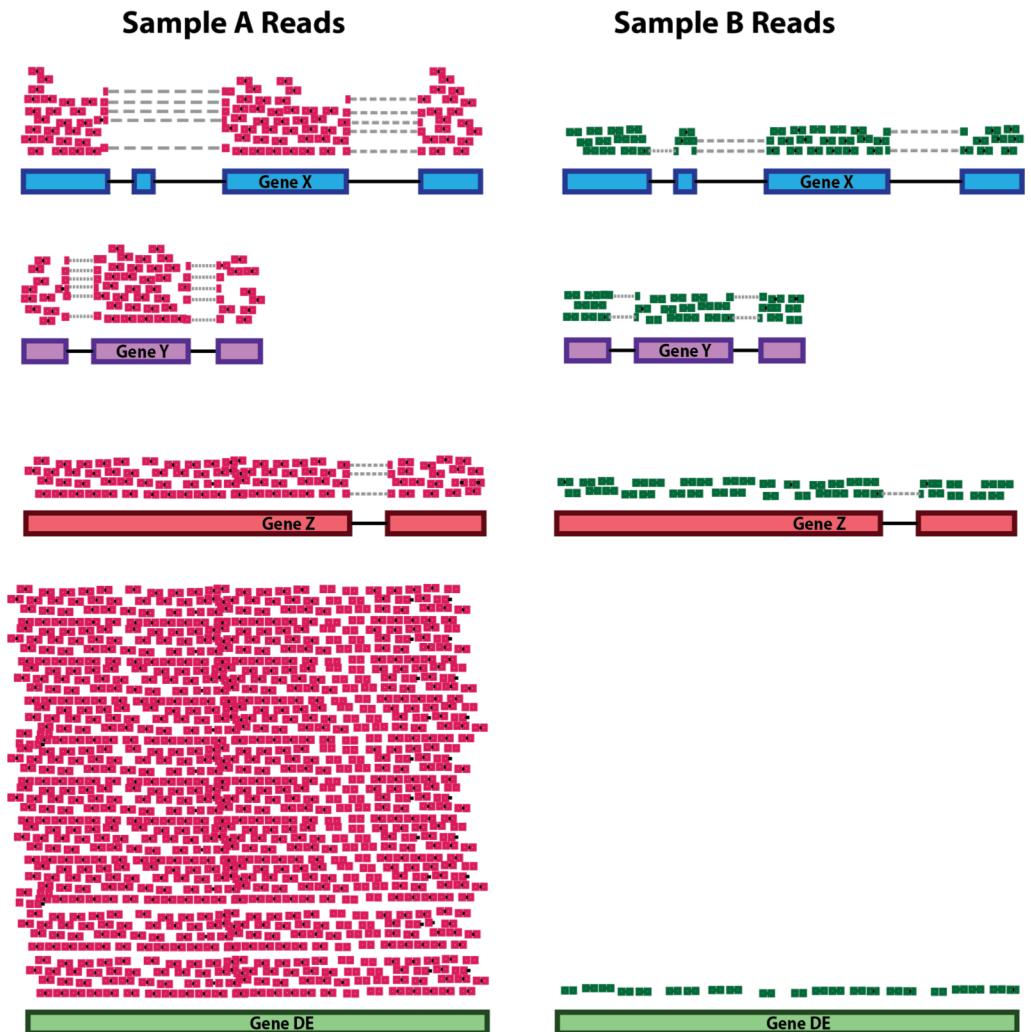
- Gene Length
- Sequencing depth



# Normalization

The number of sequenced reads mapped to a gene depends on

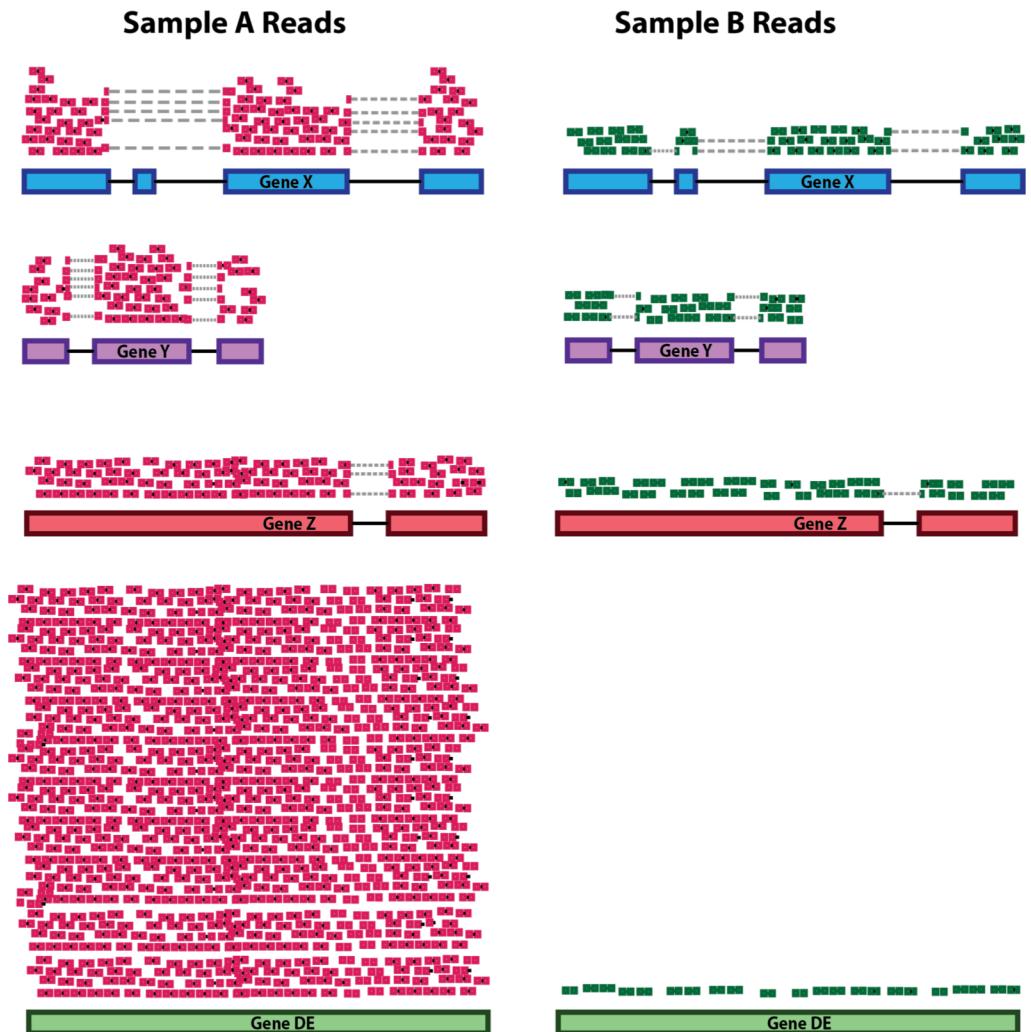
- Gene Length
- Sequencing depth
- The expression level of other genes in the sample



# Normalization

The number of sequenced reads mapped to a gene depends on

- Gene Length
- Sequencing depth
- The expression level of other genes in the sample
- **It's own expression level**

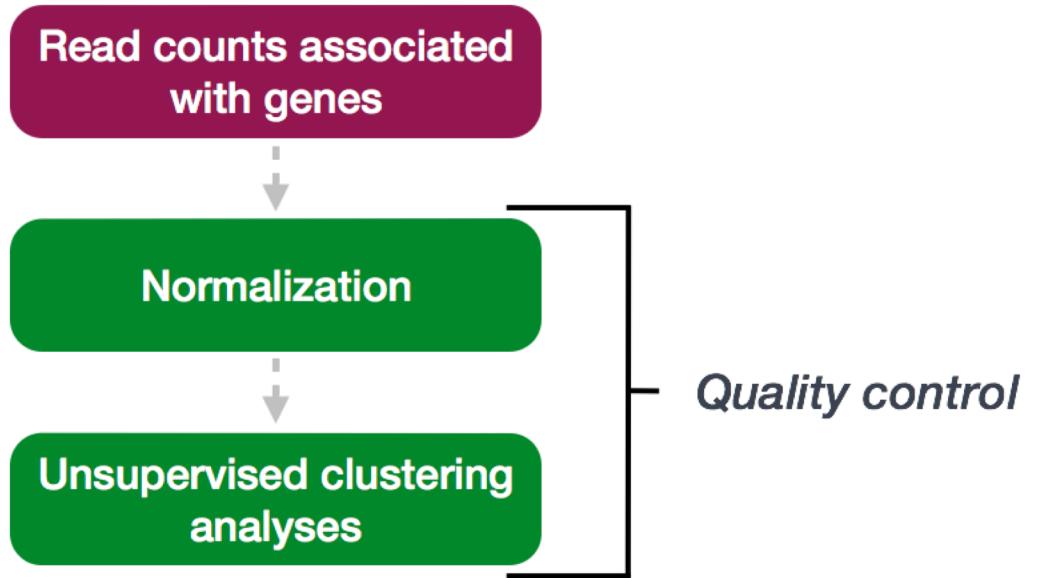


Normalization eliminates the factors that are not of interest!

# Normalization methods

Normalization method	Description	Accounted factors	For Differential Expression?
CPM (counts per million)	counts scaled by total number of reads in a sample	sequencing depth	NO
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	NO
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	NO
DESeq2's median of ratios [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	YES

# Unsupervised Clustering

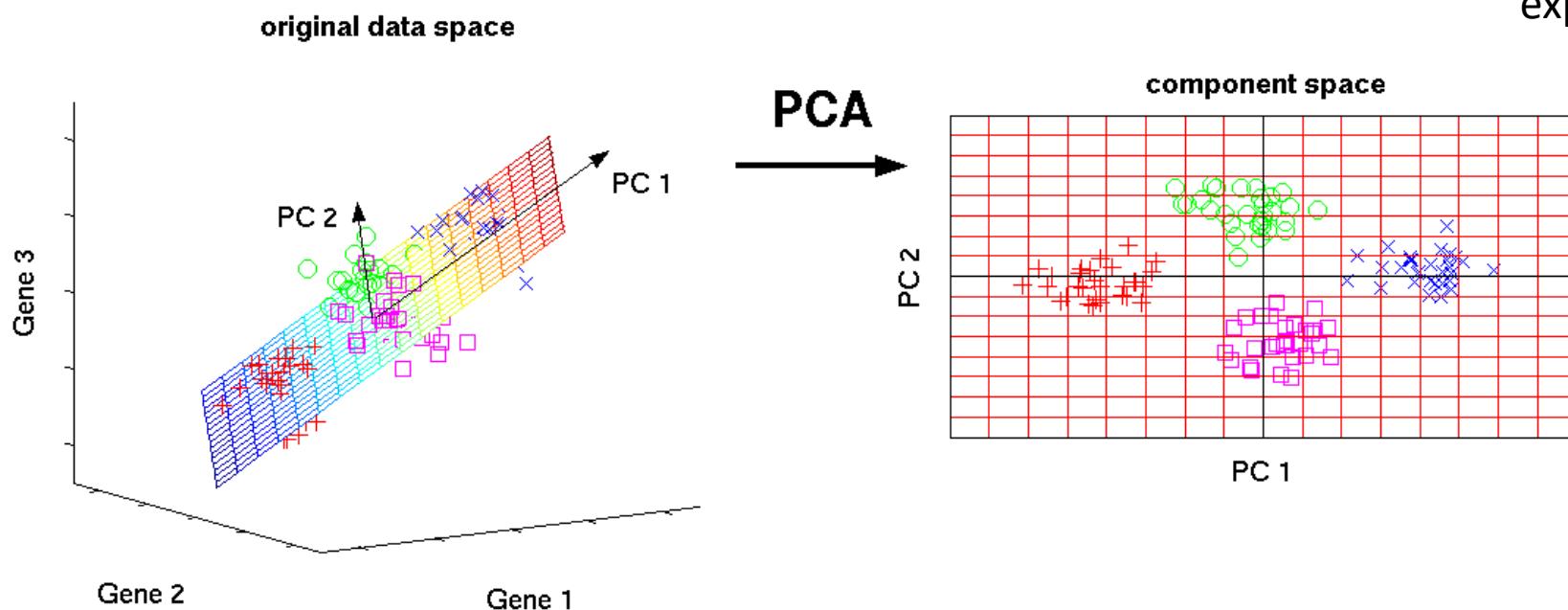


# Principle Component Analysis

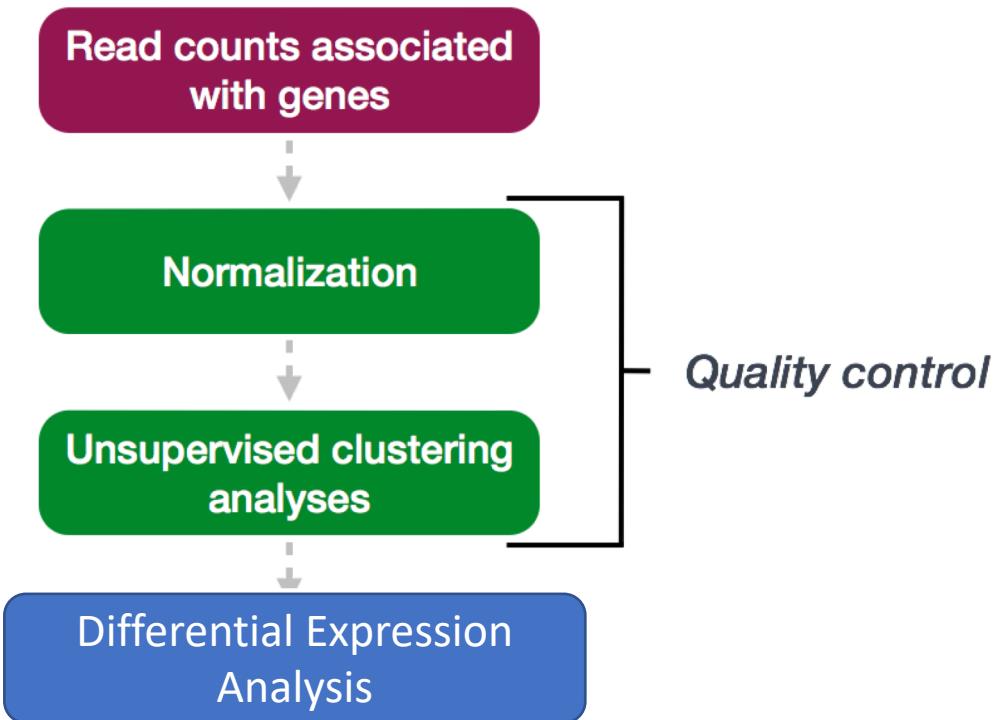
Here is an example with three genes measured in many samples:

Gene	Sample 1	Sample 2	Sample 3	Sample 4
Gene 1	1000	1000	100	10
Gene 2	10	1	5	6
Gene 3	10	1	10	20

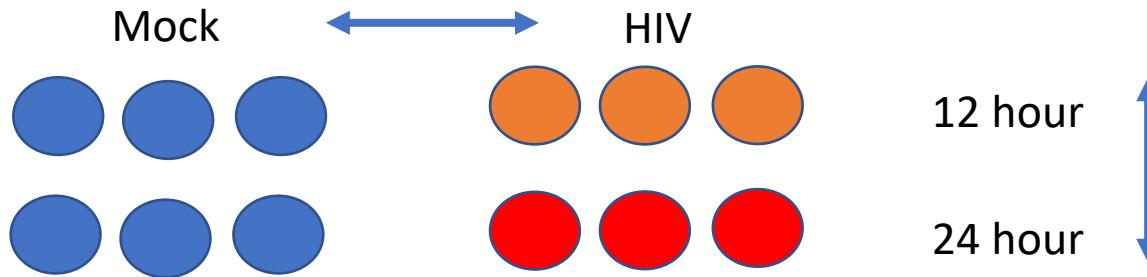
- Each gene that we measure is a "dimension" and we can visualize up to 3
- PCA can help us visualize relationships in our data in a lower number of dimensions
- PCA is an important QC step! Do your samples cluster as expected?



# Differential Expression with DESeq2



# Multi-factor experiment design

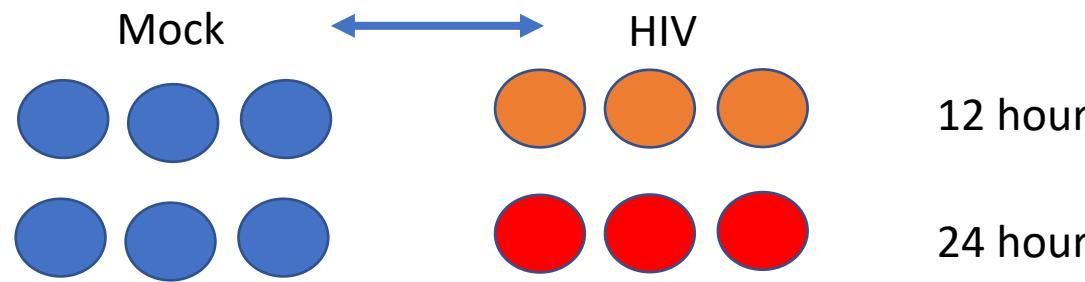


Factor 1:  
Infection status (Mock or HIV)

Factor 2:  
Time (12 or 24 hr)

Sample	Condition	Time
1	Mock	12
2	Mock	12
3	Mock	12
4	Mock	24
5	Mock	24
6	Mock	24
7	HIV	12
8	HIV	12
9	HIV	12
10	HIV	24
11	HIV	24
12	HIV	24

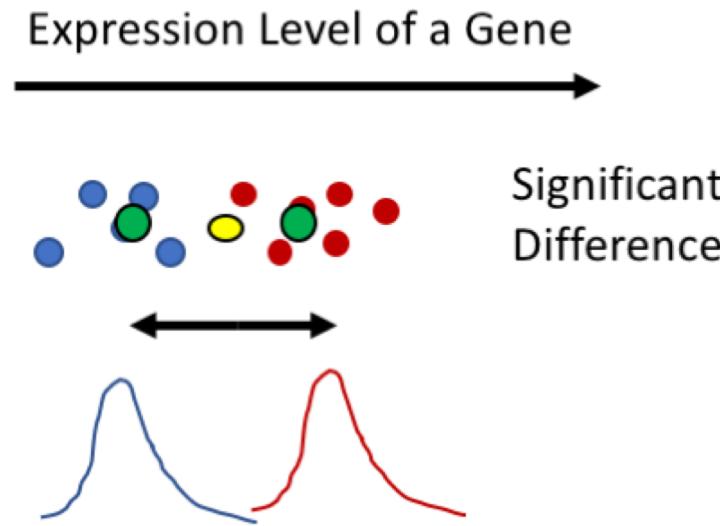
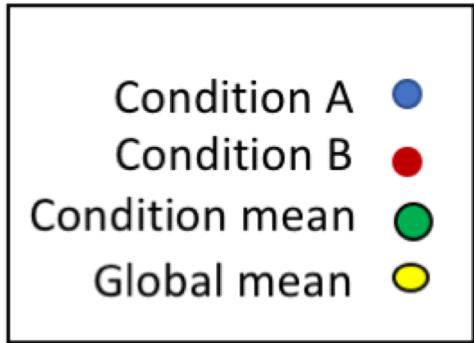
# Multi-factor experiment design



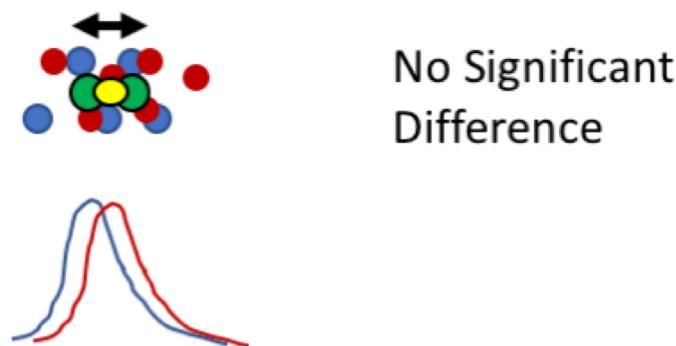
Sample	Condition	Time
1	Mock	12
2	Mock	12
3	Mock	12
4	Mock	24
5	Mock	24
6	Mock	24
7	HIV	12
8	HIV	12
9	HIV	12
10	HIV	24
11	HIV	24
12	HIV	24

- Differential Expression compares two conditions
- We'll choose a primary factor = **Infection status** for comparison
- More complex statistics can include multiple factors

# DESeq2 Test for Differential Expression



Deviation from global mean

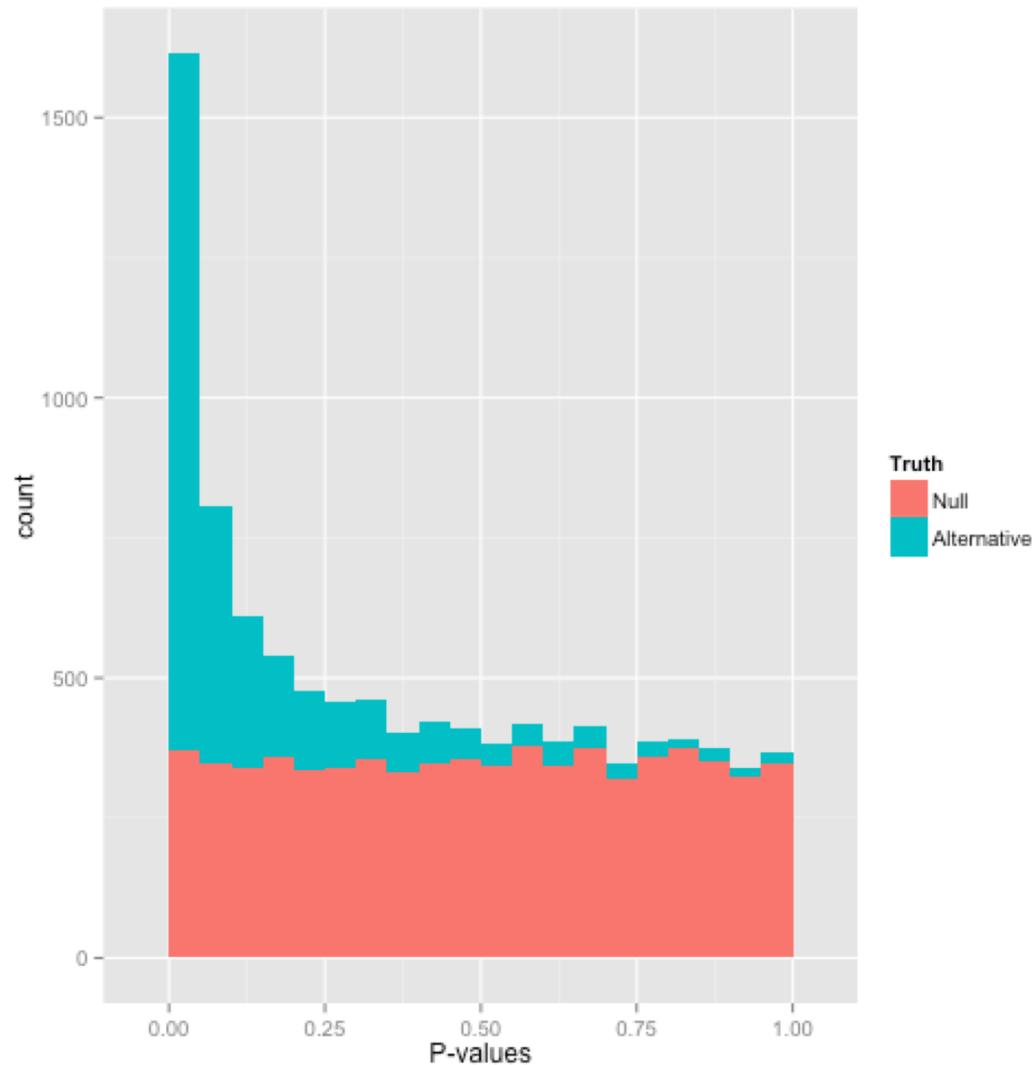


# DESeq2 Results table

GenelD	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
EGR1	1273.65	-2.22	0.12	-18.65	1.25E-77	1.44E-73
MYC	5226.12	1.41	0.11	12.53	4.95E-36	2.87E-32
OPRK1	78.35	-1.83	0.17	-10.57	4.11E-26	1.59E-22
CCNI2	7427.12	0.93	0.10	9.43	4.27E-21	1.24E-17
STRA6	785.78	0.97	0.11	8.61	7.29E-18	1.69E-14

- Mean of normalized counts – averaged over all samples from two conditions
- Log of the fold change between two conditions
- Standard Error of Log FC estimate – will reflect the “noisiness” of the gene
- P-value – the probability that the log2FoldChange is not zero
- Adjusted P value – accounting for multiple testing correction

# DESeq2 P-value histogram



- Plot of raw p-values
- P-value: Probability of getting a logFC as extreme as observed if the true logFC = 0 for that gene (null hypothesis)

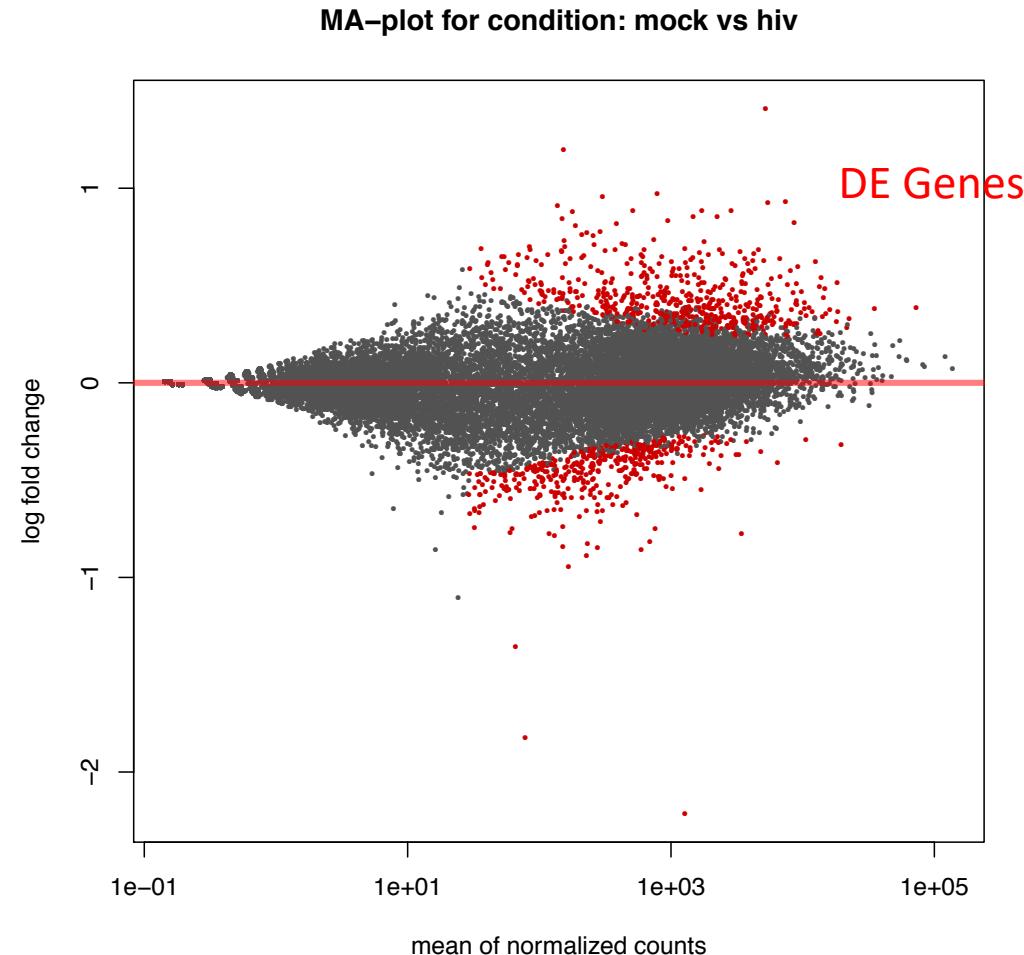
How to interpret:

- Random P-values are expected to be uniform, if you have true positives you should see a peak close to zero

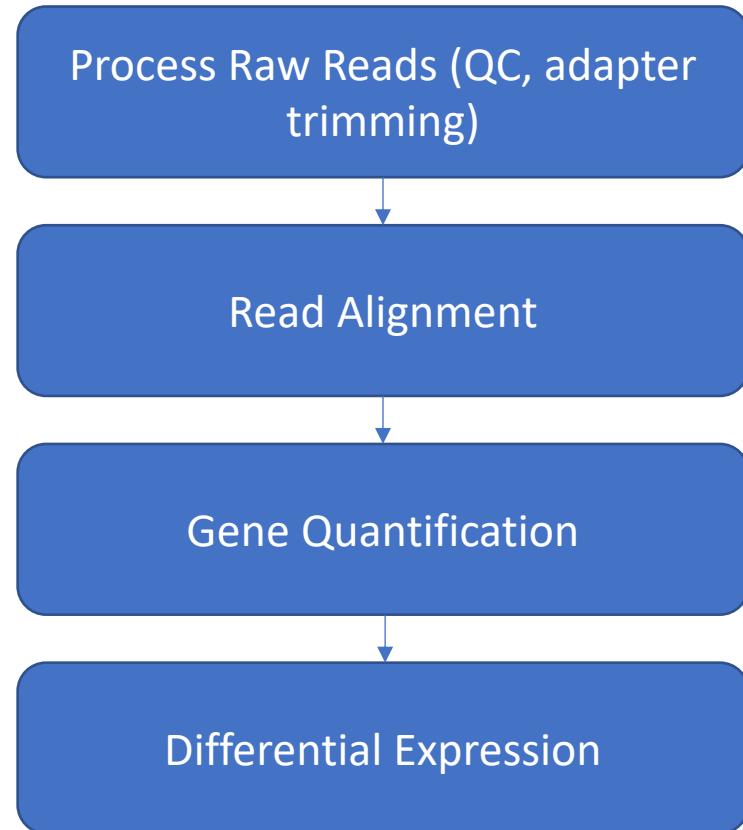
# DESeq2 MA plot

Shows the relationship between

- M: The difference in expression  
 $\text{Log}(\text{HIV}) - \text{Log}(\text{Mock}) = \text{Log}(\text{HIV}/\text{Mock})$
- A: Average expression strength  
 $\text{Average}(\text{Mock}, \text{HIV})$
- Genes with adjusted  $p$ -value  $< 0.1$  are in red
- Can be used as an overview or to diagnose problems



# Conclusions



# References

<https://www.bioconductor.org/packages/3.3/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf>

[https://hbctraining.github.io/DGE\\_workshop](https://hbctraining.github.io/DGE_workshop)

[https://galaxyproject.org/tutorials/rb\\_rnaseq/](https://galaxyproject.org/tutorials/rb_rnaseq/)

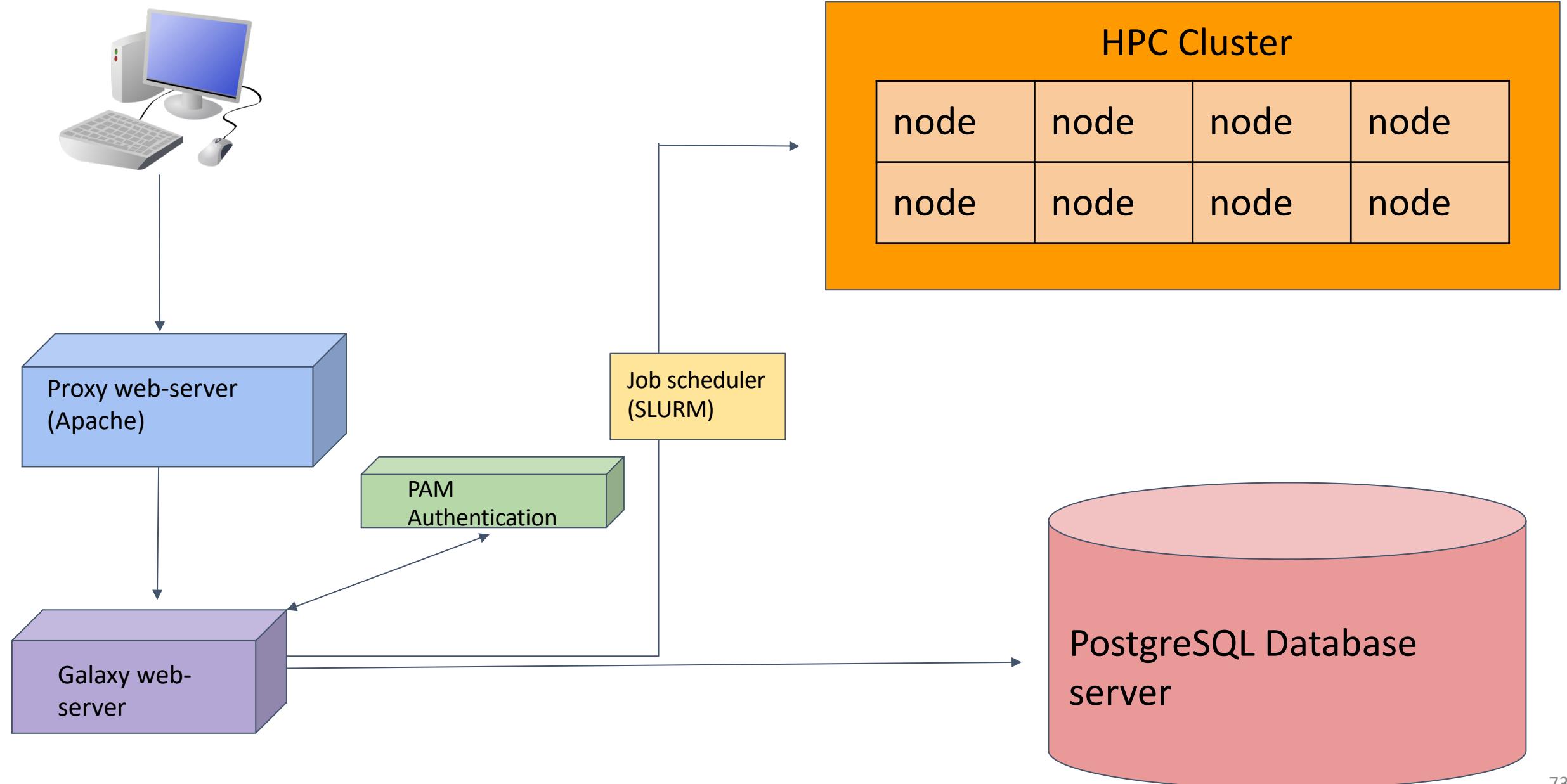


- ❖ **Web-based** platform for running data analysis and integration, geared towards bioinformatics
  - Open-source
  - Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic with many more outside contributions
  - Large and extremely responsive community

# Benefits of Galaxy

- **Accessible** for inexperienced programmers
- Workflow system
- Lots of tools available
- **Reproducibility** of results through history
- Easy to create new tools
- **Transparency**

# Production Environment



# Access Galaxy

**<https://galaxy.cluster.tufts.edu/>**

Login with you cluster username and password!

## Tools



search tools



Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

RNA-seq

SAMTOOLS

Mapping

Workflows

All workflows

## Welcome to Galaxy on the Tufts cluster

Bioinformatics @ Tufts

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by [The Galaxy Team](#) with the support of [many contributors](#).

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and [Johns Hopkins University](#).

Galaxy Interface

## History



search datasets



## Unnamed history

(empty)



This history is empty. You can load your own data or get data from an external source



## Tools

search tools X

Get Data

Send Data

Collection Operations

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

FASTQ Quality Control

RNA-seq

SAMTOOLS

Mapping

Workflows

All workflows

TOOLS

## Welcome to Galaxy on the Tufts cluster

Bioinformatics @ Tufts

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by [The Galaxy Team](#) with the support of [many contributors](#).

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and [Johns Hopkins University](#).

## History

search datasets X

## Unnamed history

(empty)



This history is empty. You can load your own data or get data from an external source

Tools

search tools

---

- Get Data
- Send Data
- Collection Operations
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Phenotype Association
- FASTQ Quality Control
- RNA-seq
- SAMTOOLS
- Mapping
- Workflows

All workflows

**TOOLS**

## Welcome to Galaxy on the Tufts cluster

Bioinformatics @ Tufts

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by [The Galaxy Team](#) with the support of [many contributors](#).

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and [Johns Hopkins University](#).

History

search datasets

---

Unnamed history

(empty)

This history is empty. You can load your own data or get data from an external source

**HISTORY**

**Tools**

search tools

Get Data  
Send Data  
Collection Operations  
Lift-Over  
Text Manipulation  
Convert Formats  
Filter and Sort  
Join, Subtract and Group  
Fetch Alignments/Sequences  
Operate on Genomic Intervals  
Statistics  
Graph/Display Data  
Phenotype Association  
FASTQ Quality Control  
RNA-seq  
SAMTOOLS  
Mapping  
Workflows  
All workflows

**TOOLS**

# Welcome to Galaxy on the Tufts cluster

Bioinformatics @ Tufts

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by [The Galaxy Team](#) with the support of [many contributors](#).

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and [Johns Hopkins University](#).

**MAIN**

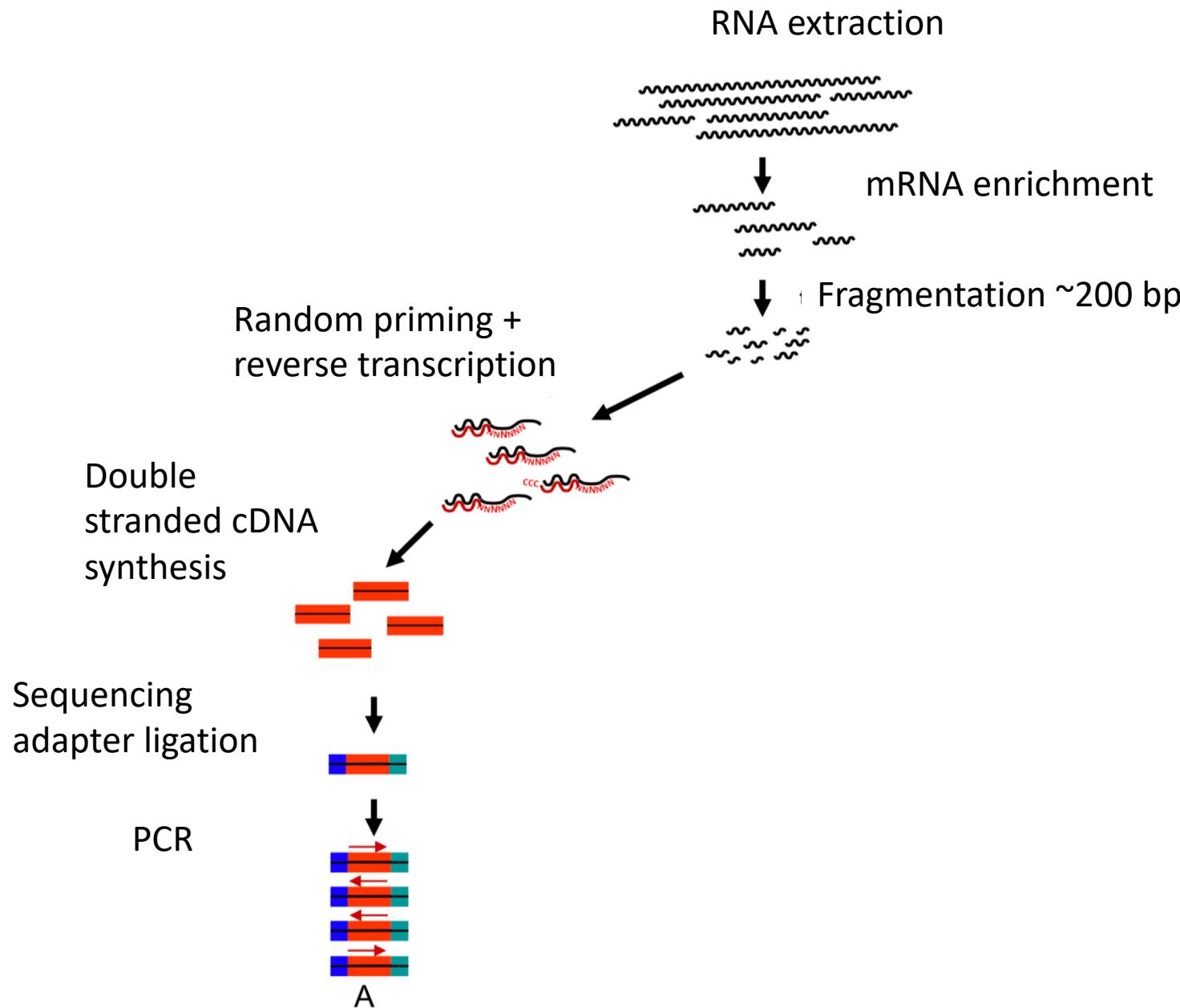
**History**

search datasets

Unnamed history  
(empty)

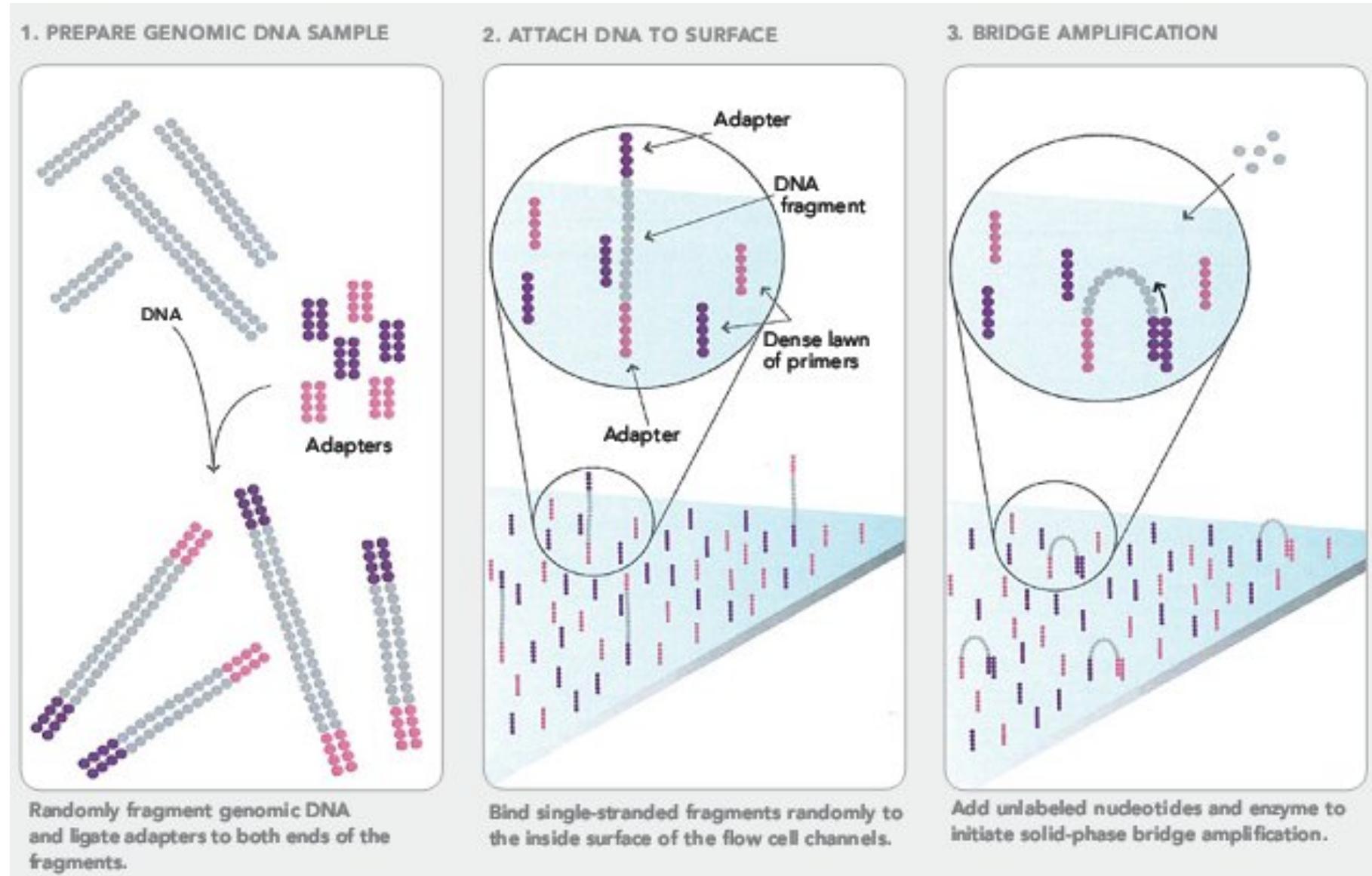
This history is empty. You can load your own data or get data from an external source

**HISTORY**

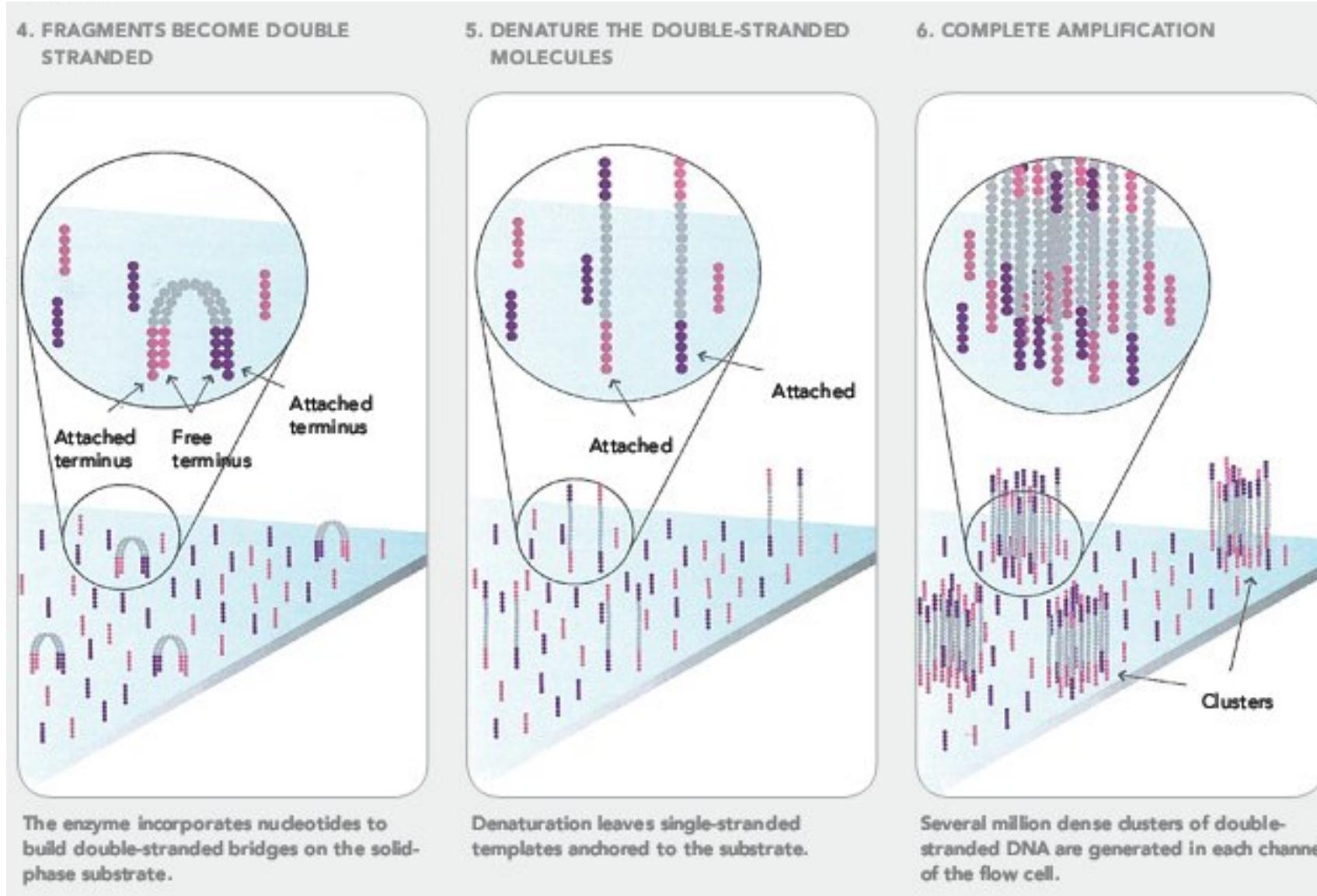


Good resource: [Griffiths et al Plos Comp Bio 2015](#)

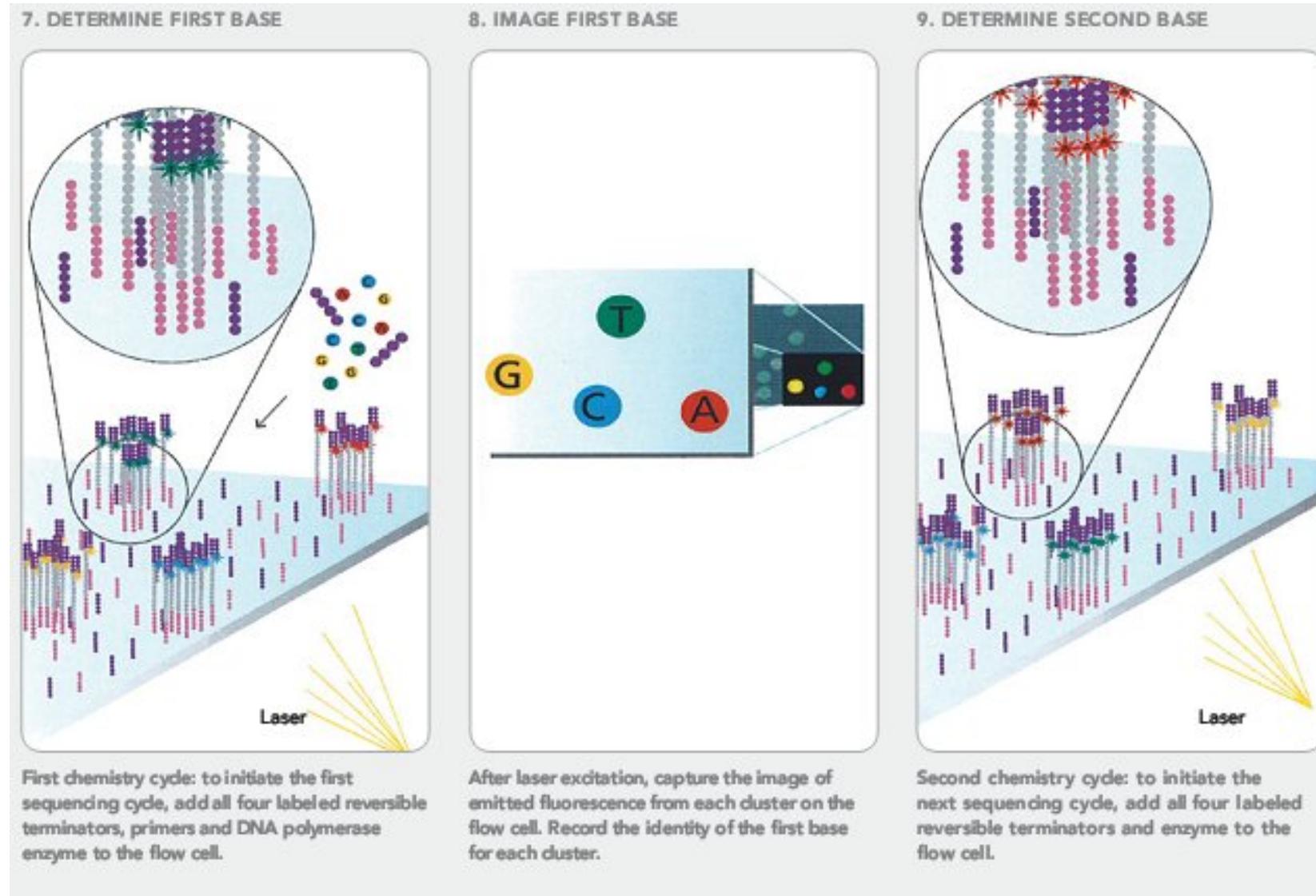
# Next Generation Sequencing (NGS)



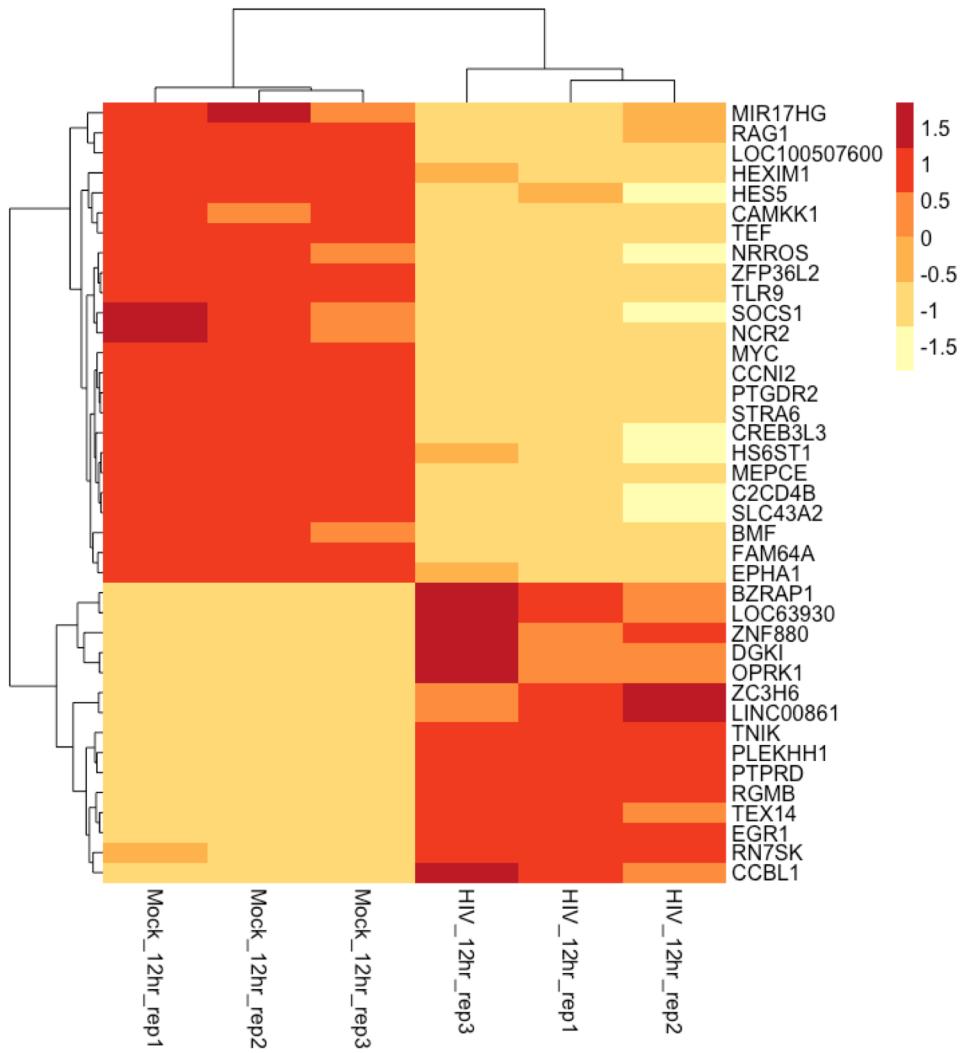
# Next Generation Sequencing (NGS)



# Next Generation Sequencing (NGS)



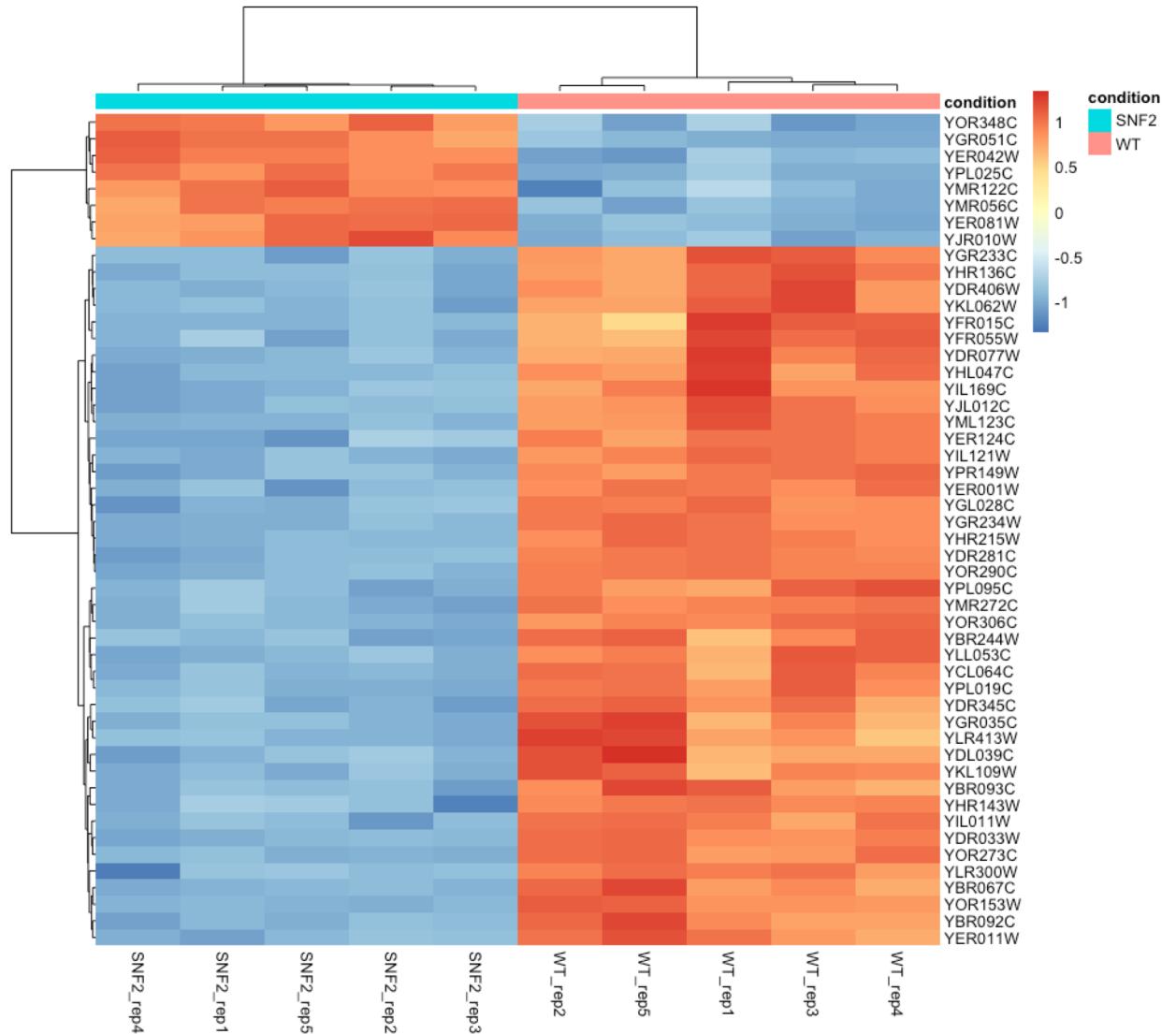
# Final Heatmap – not part of DESeq2 output



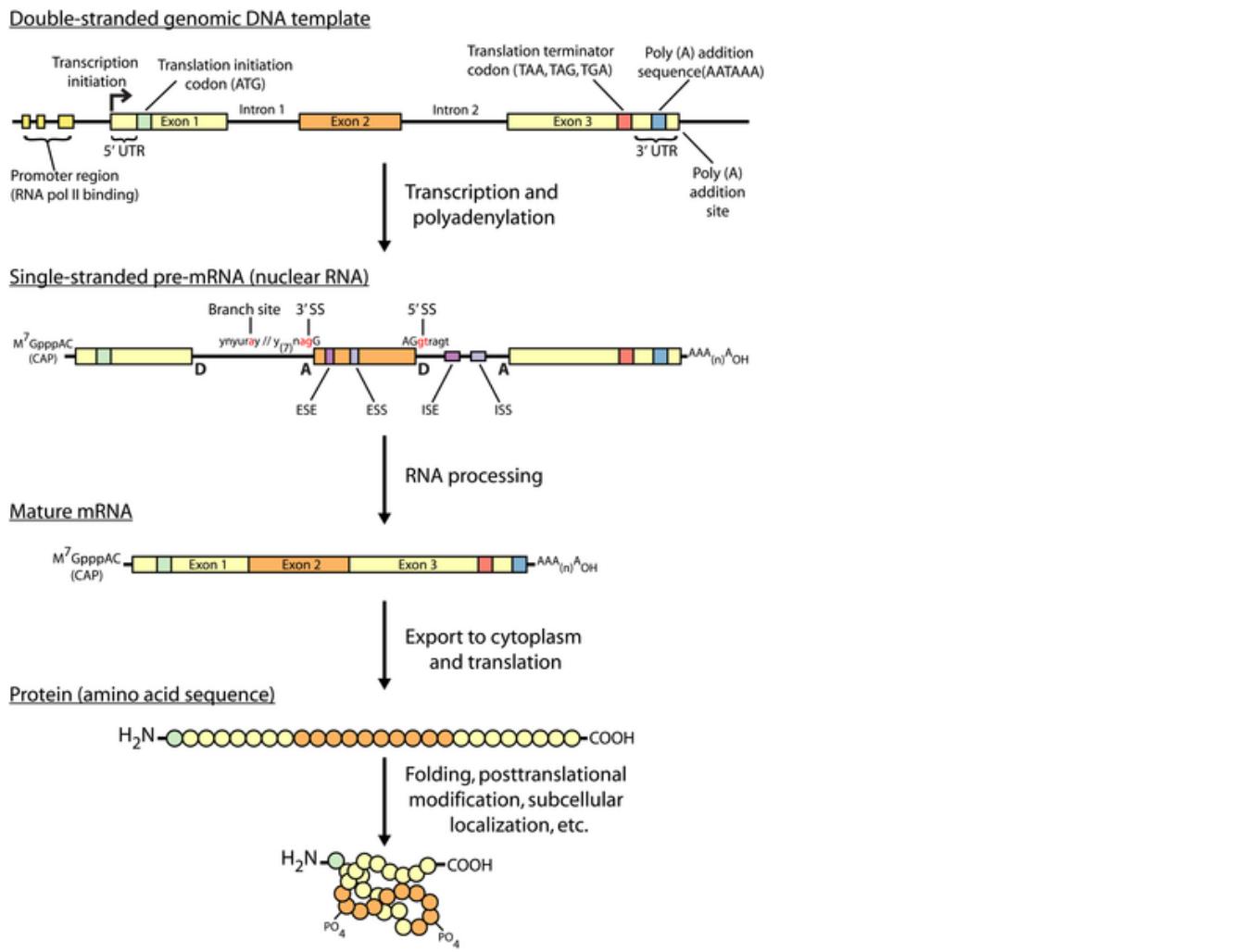
## Common RNAseq analysis goals

- Novel transcript discovery
- Transcriptome assembly
- Single cell analysis
- Quantify alternative splicing
- **Differential Expression**

Replace with actual heatmap

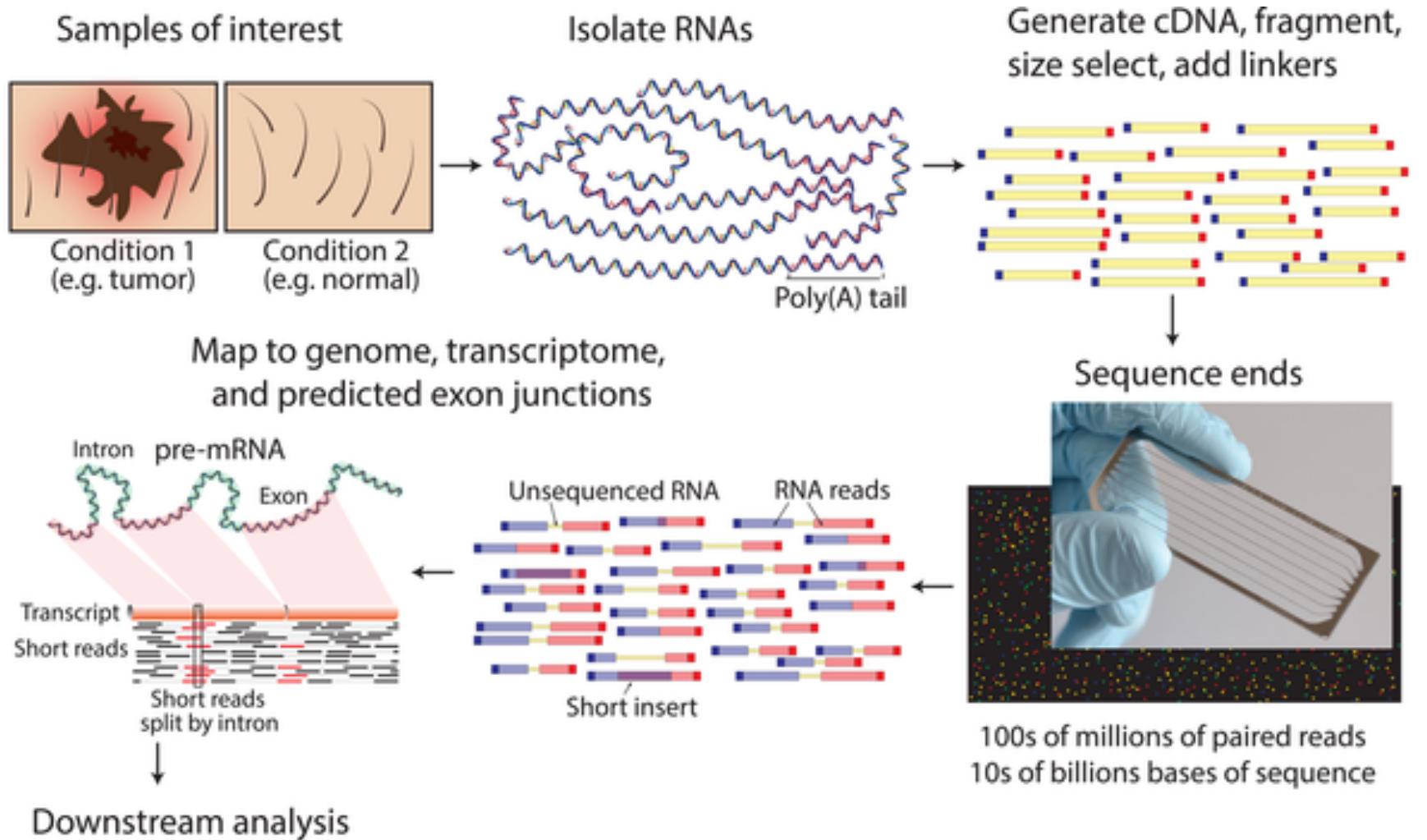


**Fig 1. An overview of the central dogma of molecular biology.**



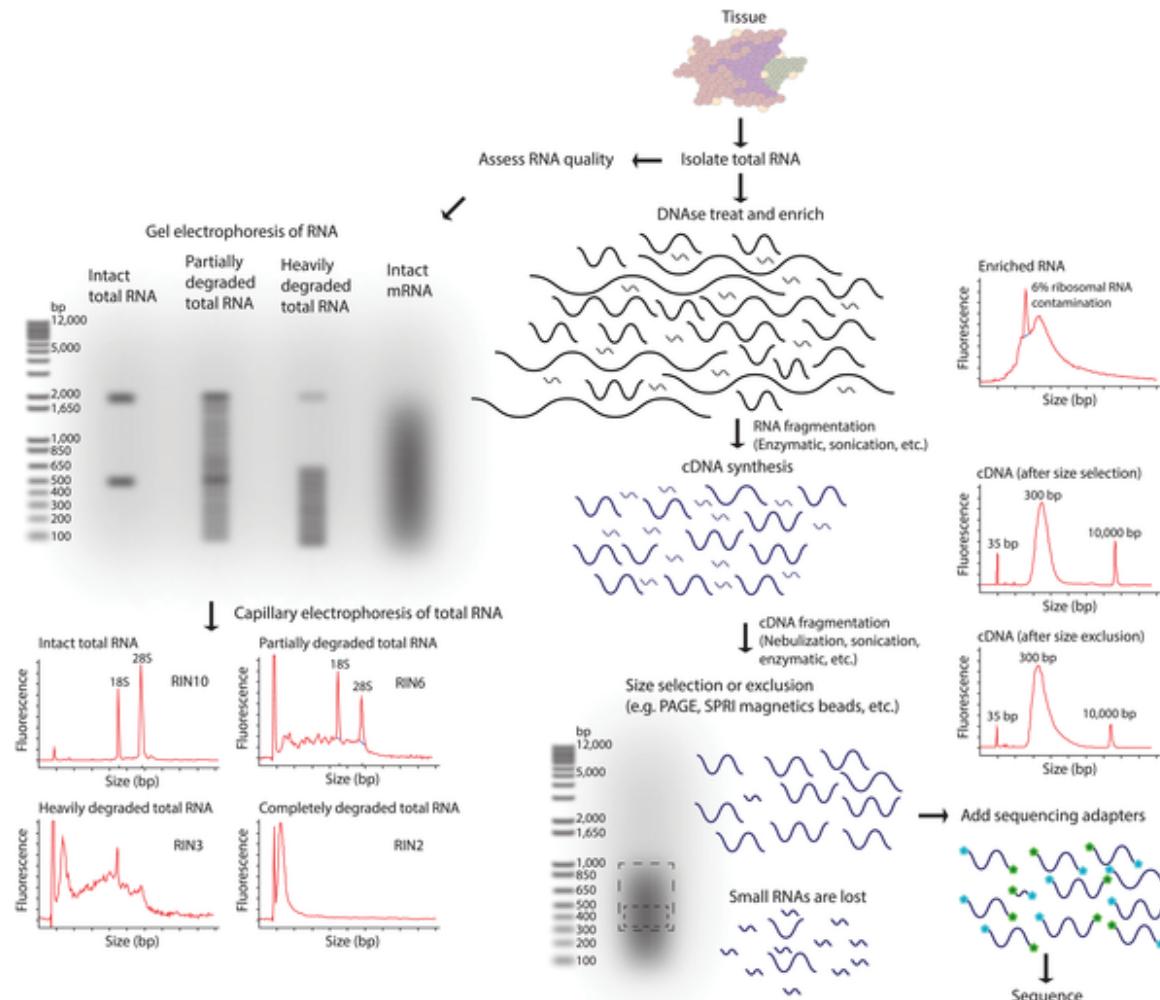
Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL (2015) Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLOS Computational Biology 11(8): e1004393. <https://doi.org/10.1371/journal.pcbi.1004393>

**Fig 2. RNA-seq data generation.**



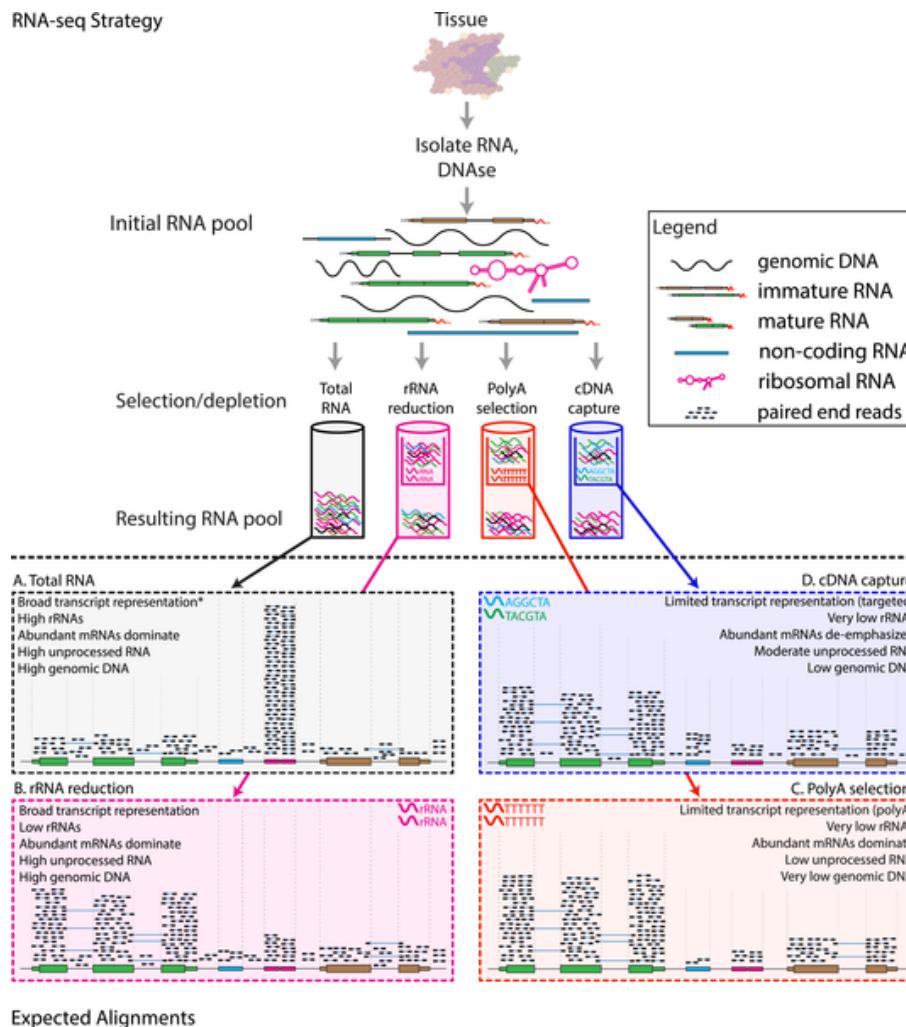
Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL (2015) Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLOS Computational Biology 11(8): e1004393. <https://doi.org/10.1371/journal.pcbi.1004393>  
<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393>

**Fig 3. RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.**



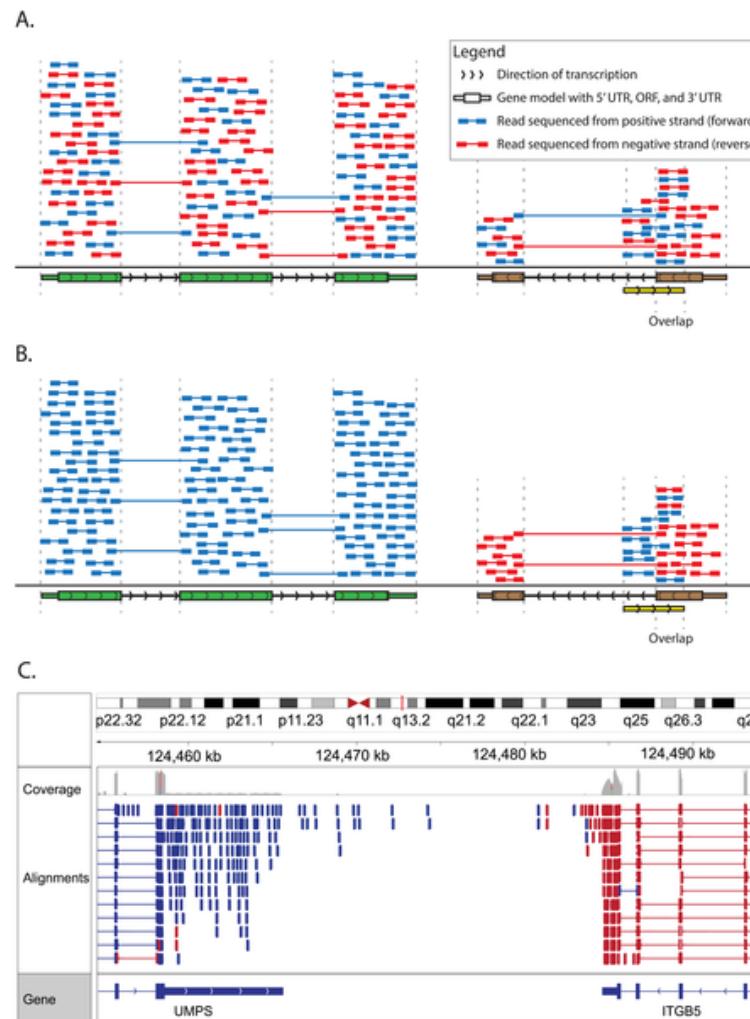
Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL (2015) Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLOS Computational Biology 11(8): e1004393. <https://doi.org/10.1371/journal.pcbi.1004393>  
<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393>

**Fig 4. RNA-seq library enrichment strategies that influence interpretation and analysis.**



Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL (2015) Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLOS Computational Biology 11(8): e1004393. <https://doi.org/10.1371/journal.pcbi.1004393>  
<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393>

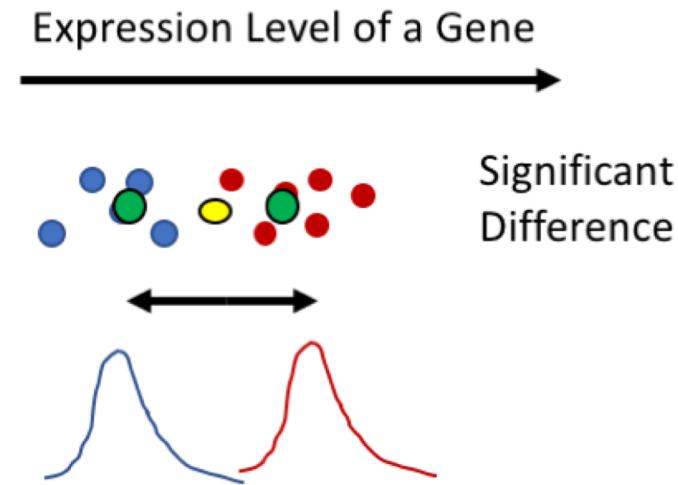
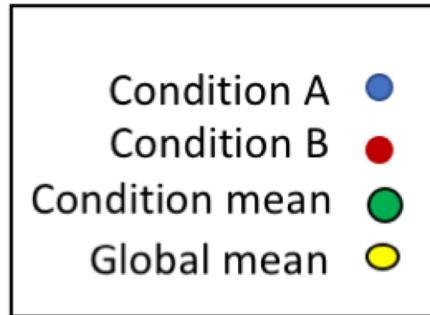
**Fig 6. Comparison of stranded and unstranded RNA-seq library methods and their influence on interpretation and analysis.**



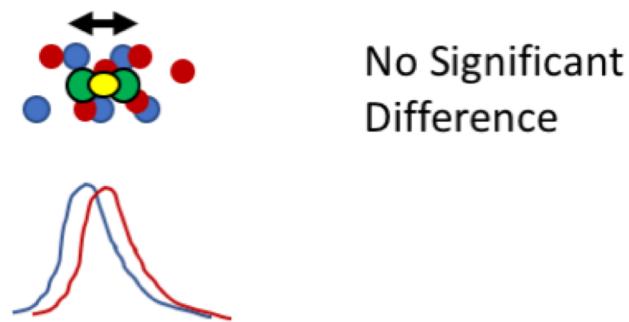
Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL (2015) Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLOS Computational Biology 11(8): e1004393. <https://doi.org/10.1371/journal.pcbi.1004393>  
<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393>

# Test for Differential Expression

DESeq2 will seek to fit a probability distribution to each gene we measured and perform a statistical test to determine whether there is a difference between conditions



Deviation from global mean



# Tufts Galaxy Server for Bioinformatics

# Normalization: DESeq2 Median of Ratios

Accounts for both sequencing depth and composition

## Step 1: creates a pseudo-reference sample (row-wise geometric mean)

For each gene, a pseudo-reference sample is created that is equal to the geometric mean across all samples.

gene	sampleA	sampleB	pseudo-reference sample
1	1000	1000	$\sqrt{(1000 * 1000)} = 1000$
2	10	1	$\sqrt{(10 * 1)} = 3.16$
...	...	...	...

# Normalization: DESeq2 Median of Ratios

## Step 2: calculates ratio of each sample to the reference

Calculate the ratio of each sample to the pseudo-reference. Since most genes aren't differentially expressed, ratios should be similar.

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
1	1000	1000	1000	$1000/1000 = 1.00$	$1000/1000 = 1.00$
2	10	1	3.16	$10/3.16 = 3.16$	$1/3.16 = 0.32$
...	...	...	...		

# Normalization: DESeq2 Median of Ratios

## Step 2: calculates ratio of each sample to the reference

Calculate the ratio of each sample to the pseudo-reference.

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
1	1000	1000	1000	$1000/1000 = 1.00$	$1000/1000 = 1.00$
2	10	1	3.16	$10/3.16 = 3.16$	$1/3.16 = 0.32$
...	...	...	...	...	...

Median = 2.08      Median = 0.66

## Step 3: calculate the normalization factor for each sample (size factor)

The median value of all ratios for a given sample is taken as the normalization factor (size factor) for that sample:

# Normalization: DESeq2 Median of Ratios

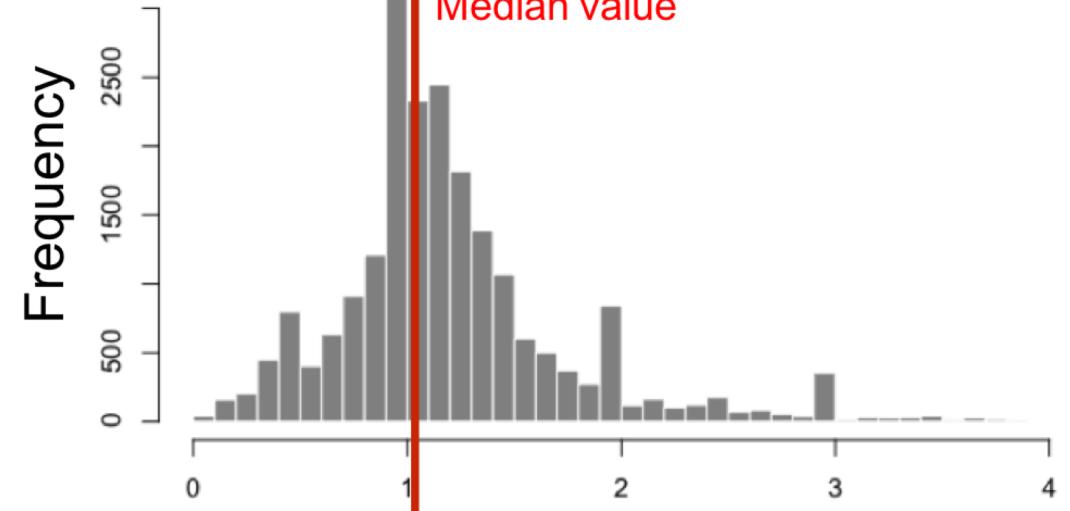
Visualization of normalization factor for a sample:

- Median should be ~1 for each sample, otherwise data should be examined for the presence of large outliers
- This method is robust to imbalance in up-/down- regulation and large numbers of differentially expressed genes

Assumptions of this method:

Not all genes are differentially expressed

sample 1 / pseudo-reference sample



# Normalization: DESeq2 Median of Ratios

## Step 4: calculate the normalized count values using the normalization factor

This is performed by dividing each raw count value in a given sample by that sample's size factor to generate normalized count values.

SampleA normalization factor = 2.08

SampleB normalization factor = 0.66

**Raw Counts**

gene	sampleA	sampleB
1	1000	1000
2	10	1

**Normalized Counts**

gene	sampleA	sampleB
1	$1000/2.08 = 480.77$	$1000 / 0.66 = 1515.16$
2	$10/2.08 = 4.81$	$1 / 0.66 = 1.52$

# Reference-based vs Reference-free RNAseq

RNAseq can be roughly divided into two "types":

- **Reference genome-based** - an assembled genome exists for a species for which an RNAseq experiment is performed. It allows reads to be aligned against the reference genome and significantly improves our ability to reconstruct transcripts. This category would obviously include humans and most model organisms
- **Reference genome-free** - no genome assembly for the species of interest is available. In this case one would need to assemble the reads into transcripts using *de novo* approaches. This type of RNAseq is as much of an art as well as science because assembly is heavily parameter-dependent and difficult to do well.

In this lesson we will focus on the **Reference genome-based** type of RNA seq.