RNA-seq to study HIV Infection in cells

Jason Laird

Bioinformatics Scientist

Feb 2022

Research Technology Team

Delilah Maloney
High Performance Computing Specialist

Kyle Monahan
Senior Data Science Specialist

Shawn Doughty
Manager, Research Computing

Jason Laird
Bioinformatics Scientist

Chris Barnett
Senior Geospatial Analyst

Tom Phimmasen
Senior Data Consultant

Patrick Florance
Director, Academic Data Services

Jake Perl
Digital Humanities NLP Specialist

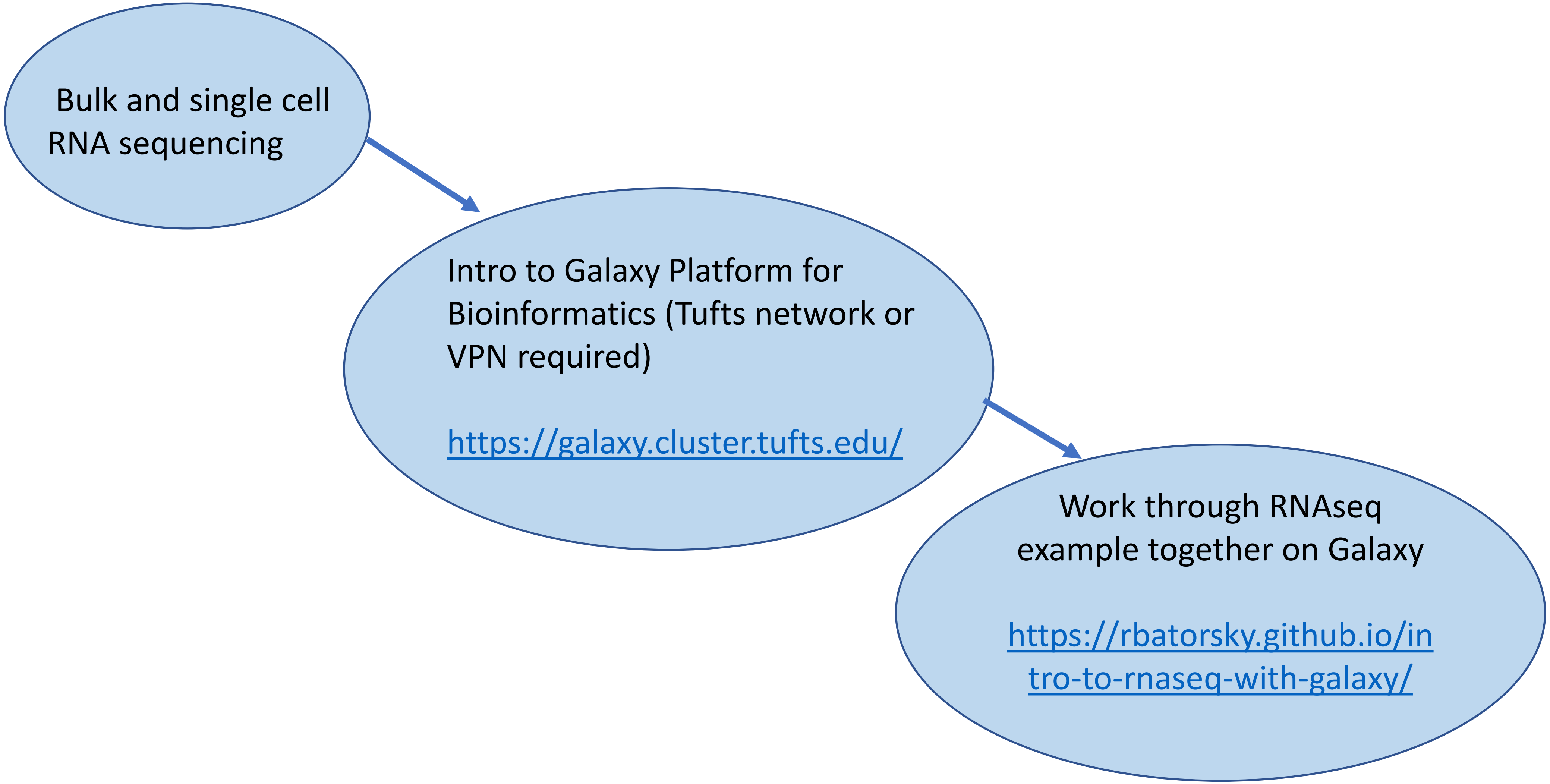
Carolyn Talmadge
Senior GIS Specialist

Uku-Kaspar Uustalu
Data Science Specialist

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

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

Outline



Bulk and single cell RNA sequencing

Intro to Galaxy Platform for Bioinformatics (Tufts network or VPN required)

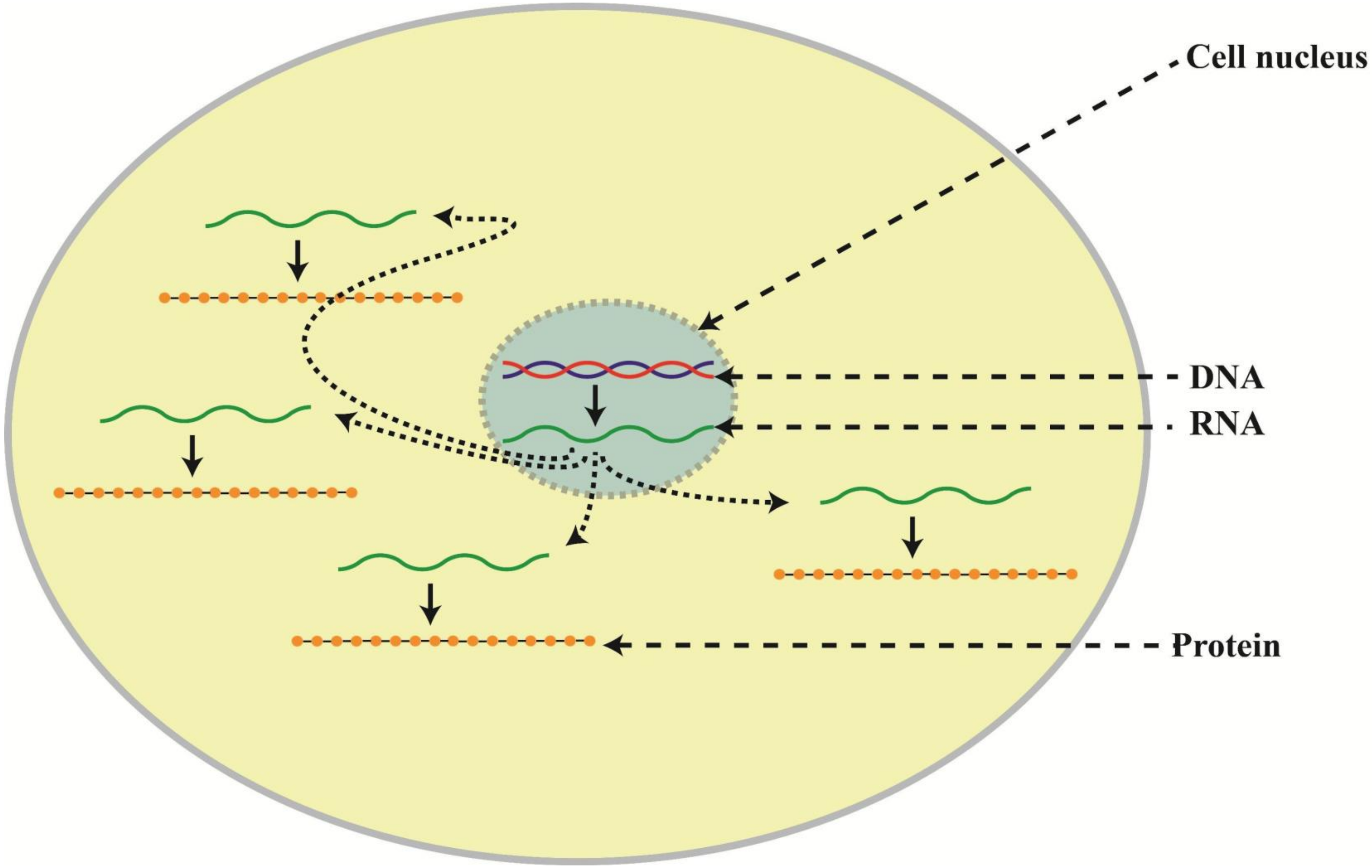
https://galaxy.cluster.tufts.edu/

Work through RNAseq example together on Galaxy

https://rbatorsky.github.io/intro-to-rnaseq-with-galaxy/

2 days!

DNA and RNA in a cell

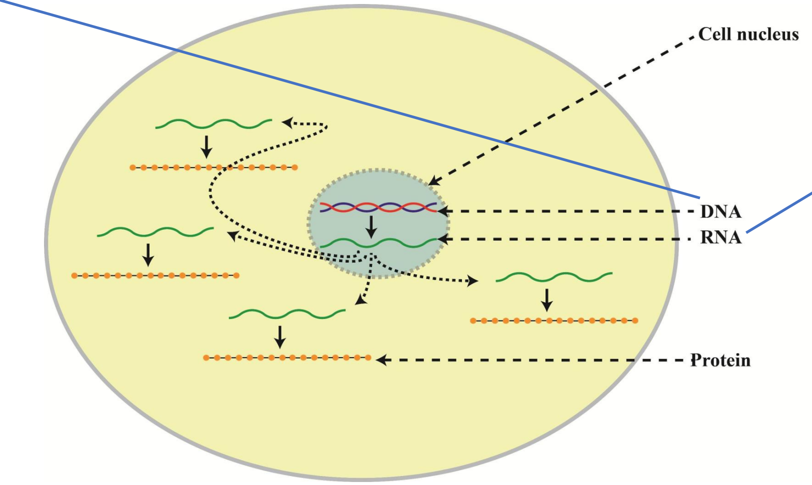


https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

Two common analyses

DNA Sequencing

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



RNA Sequencing

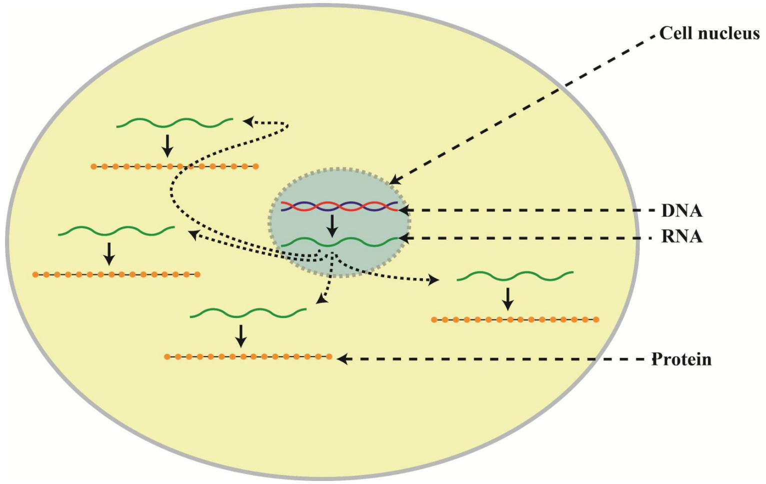
- Number of copies of a gene transcript per cell depends on gene expression
- Analysis goal:
  - Bulk : Differential expression
  - Single cell : Quantify different cell populations

https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

Today we will cover RNA sequencing

~~DNA Sequencing~~

- ~~Fixed number of copies of a gene per cell~~
- ~~Analysis goal: Variant calling and interpretation~~



RNA Sequencing

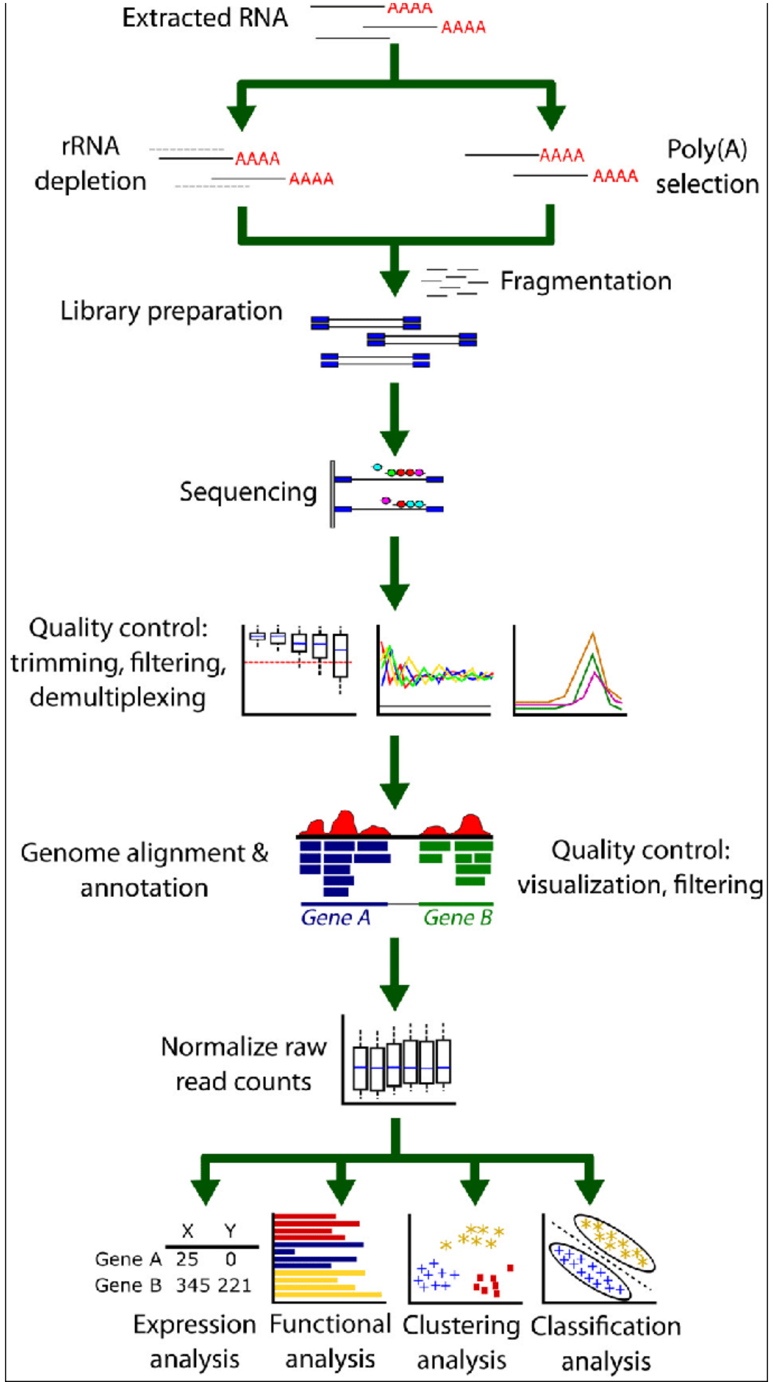
- Number of copies of a gene transcript per cell depends on gene expression
- Analysis goal:
  - Bulk : Differential expression
  - Single cell : Quantify different cell populations

https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

“Bulk” RNA seq workflow

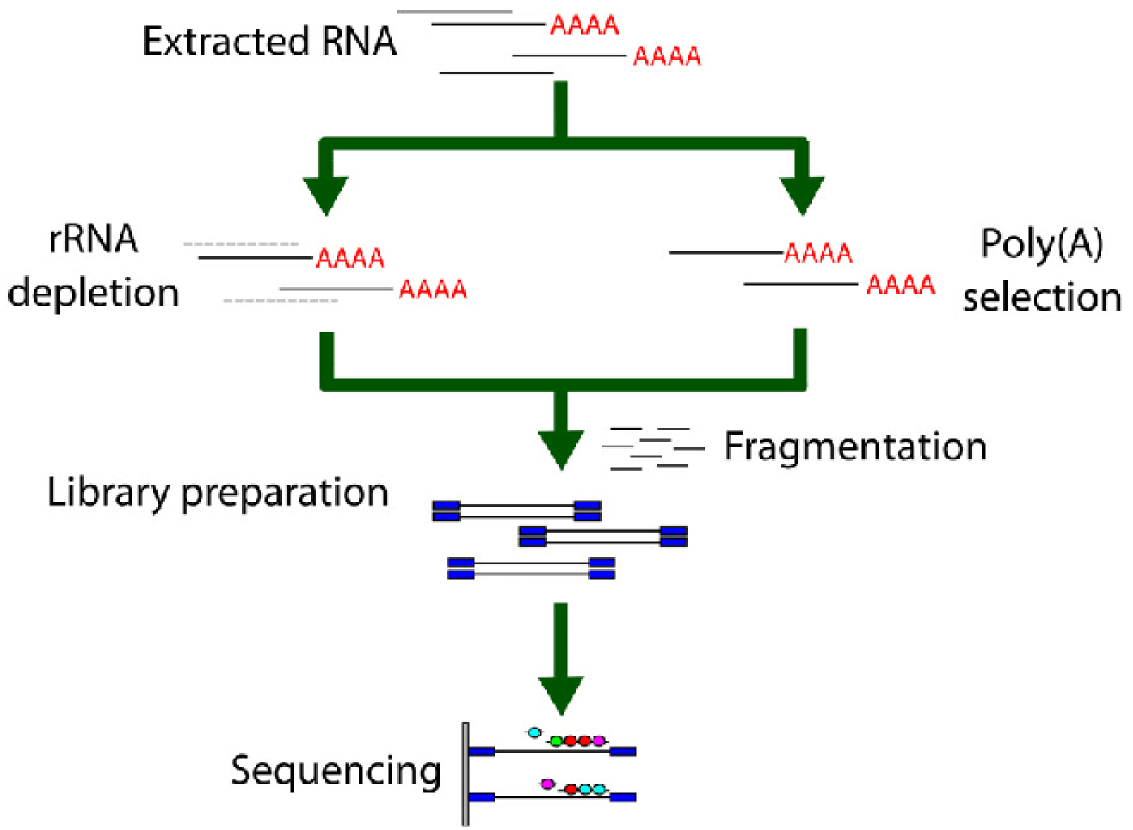
Library prep and sequencing

Bioinformatics



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

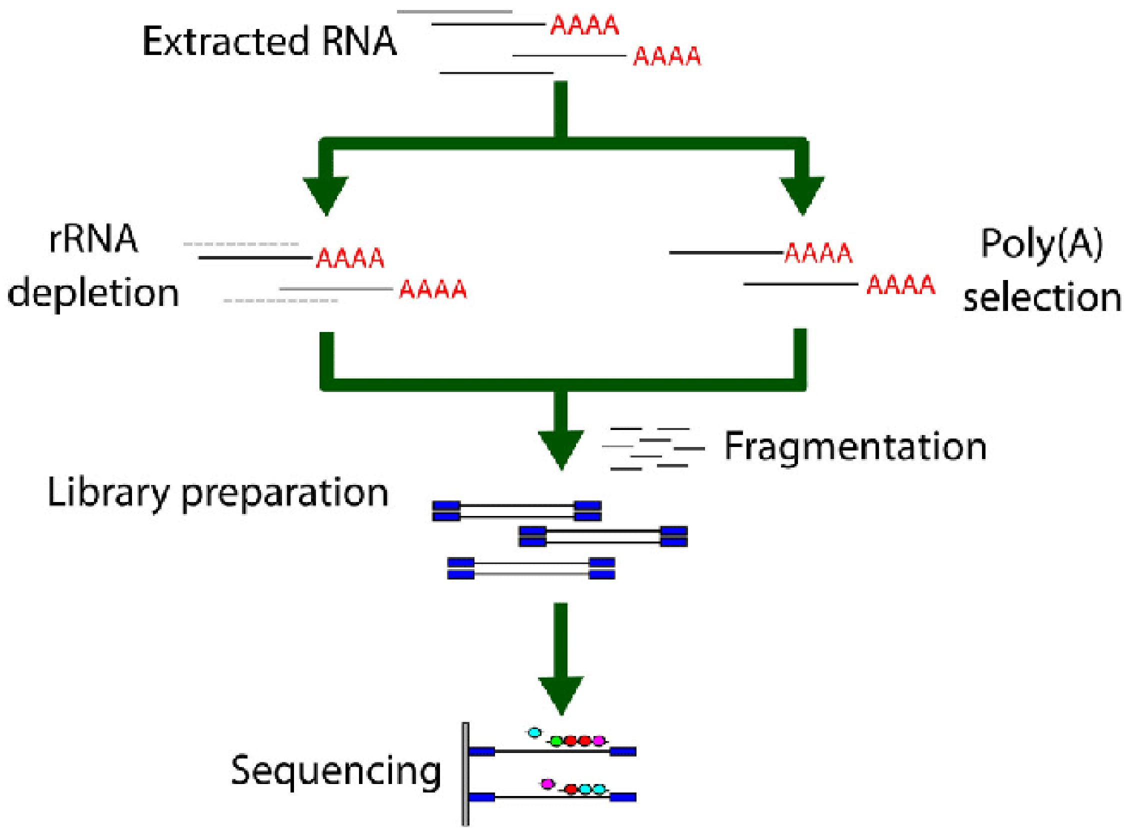
RNA seq library prep and sequencing



Good resource: Griffiths et al Plos Comp Bio 2015

RNA seq library prep and sequencing

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

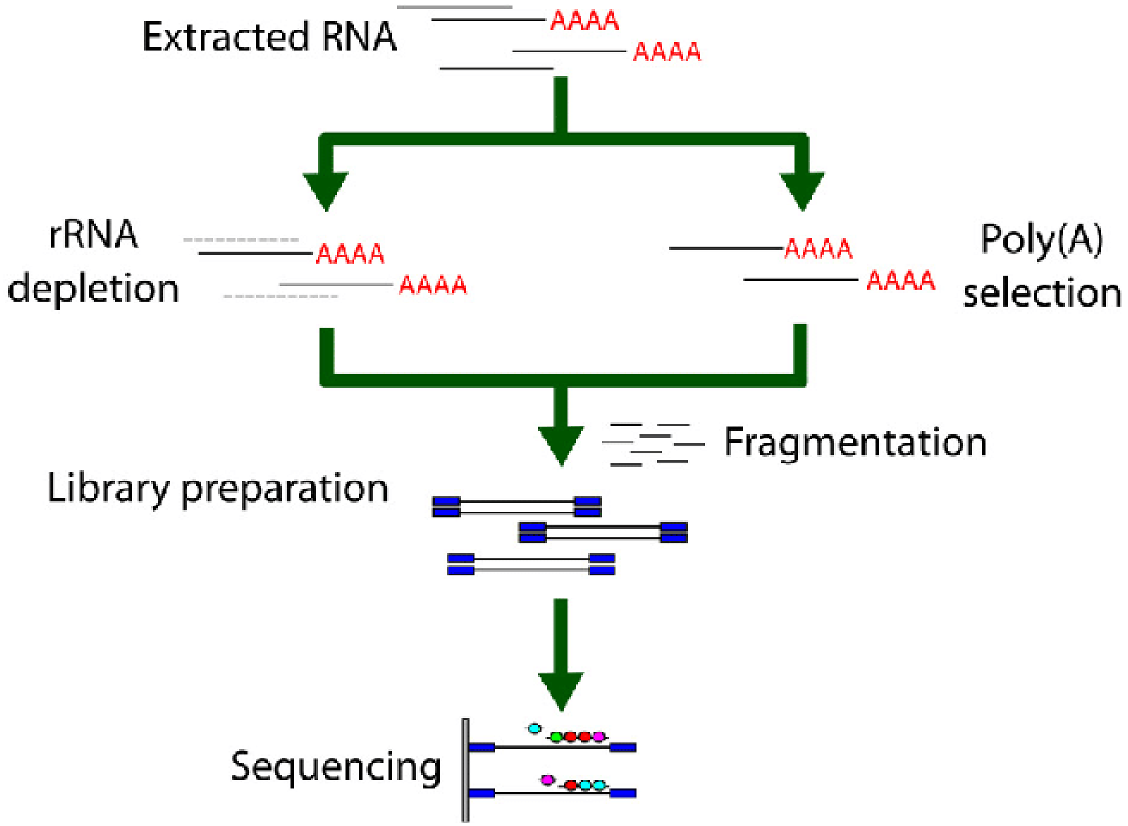


Good resource: Griffiths et al Plos Comp Bio 2015

RNA seq library prep and sequencing

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

- Random priming and reverse transcription
- Double stranded cDNA synthesis
- Sequencing adapter ligation

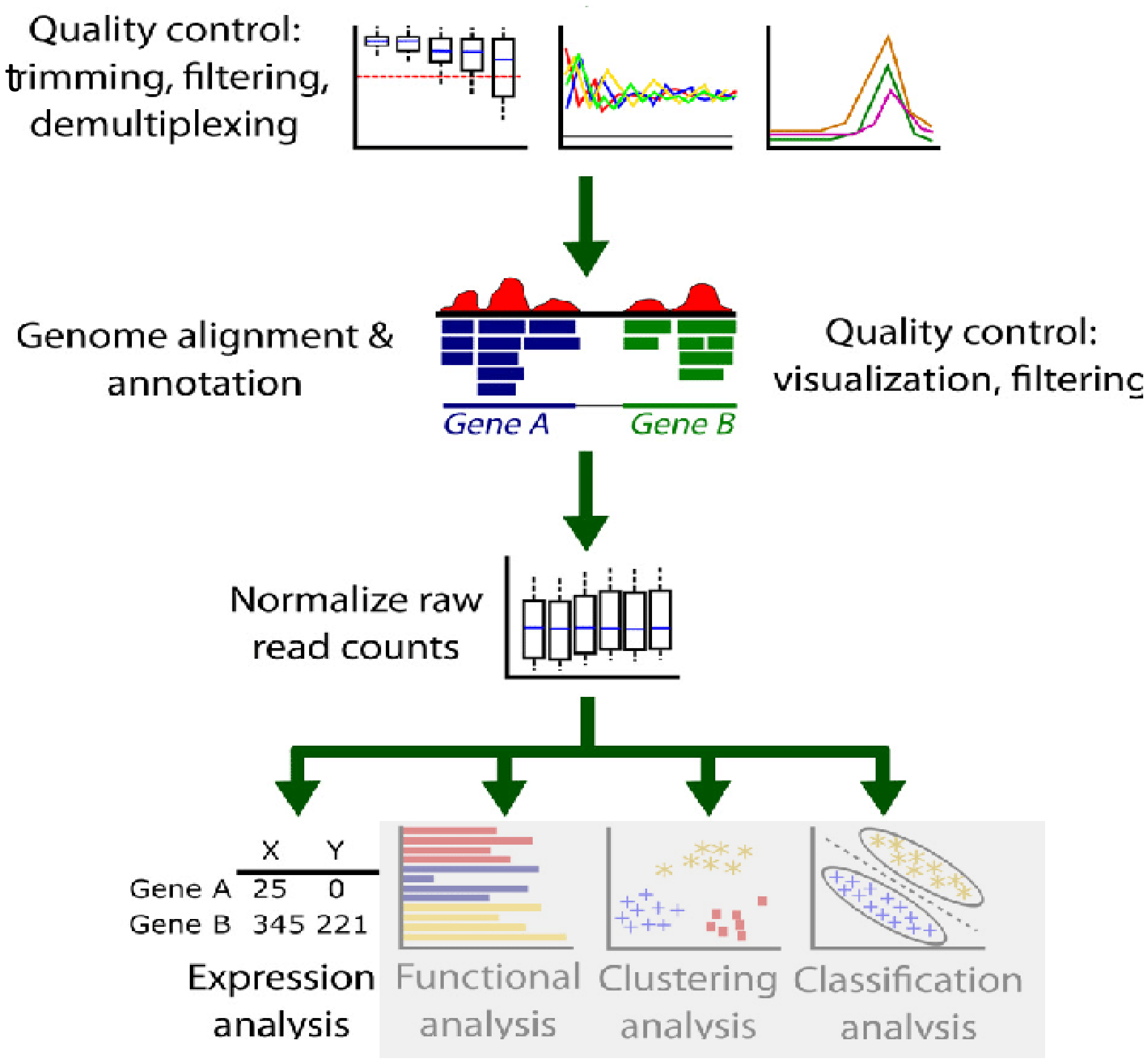


Resources:
[Illumina Sequencing by Synthesis]
[Griffiths et al Plos Comp Bio 2015]

RNA seq bioinformatics

Goal of Differential Expression

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

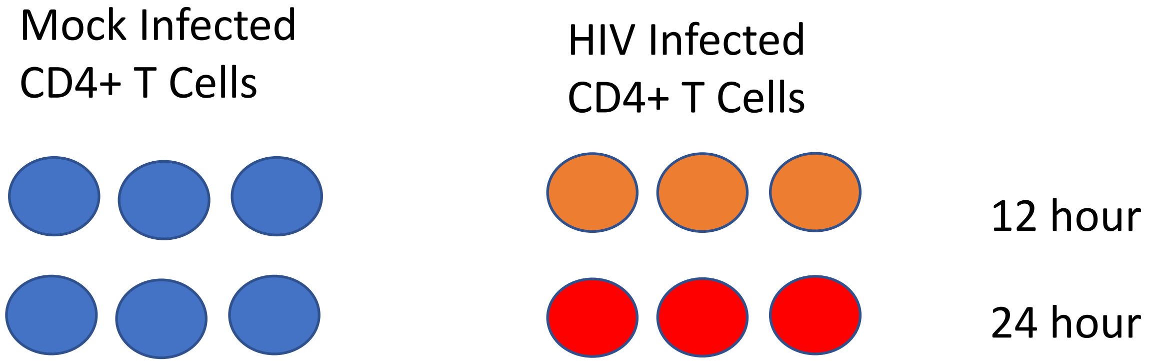


Oshlack et al. 2010. From RNA-seq reads to differential expression results. Genome Biology 2010, 11:220

Our dataset

Next-Generation Sequencing Reveals HIV-1-Mediated Suppression of T Cell Activation and RNA Processing and Regulation of Noncoding RNA Expression in a CD4$^+$ T Cell Line

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



Mock Infected
CD4+ T Cells

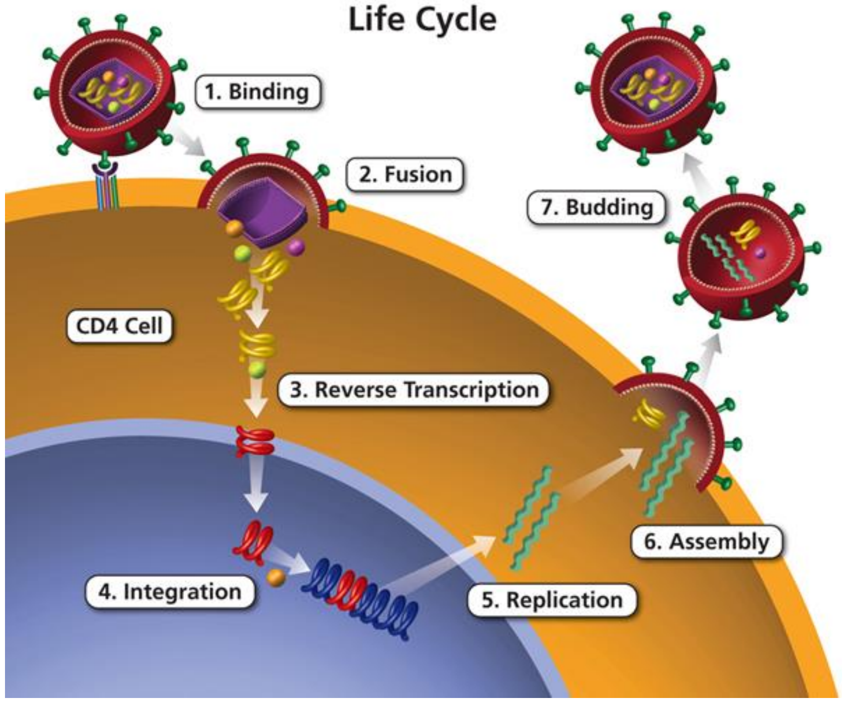
HIV Infected
CD4+ T Cells

12 hour

24 hour

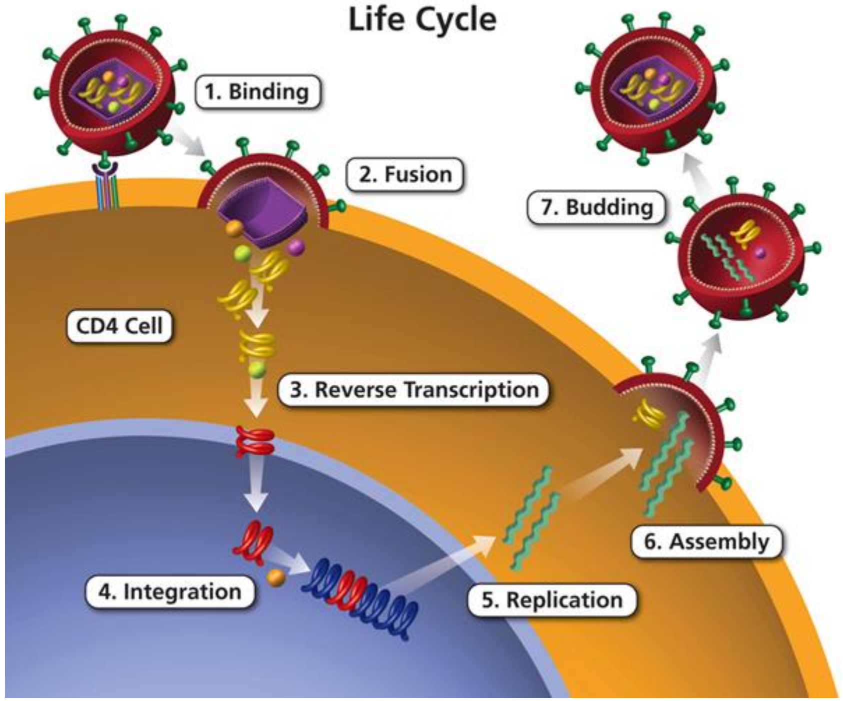
https://www.ncbi.nlm.nih.gov/pubmed/21933919

HIV lifecycle

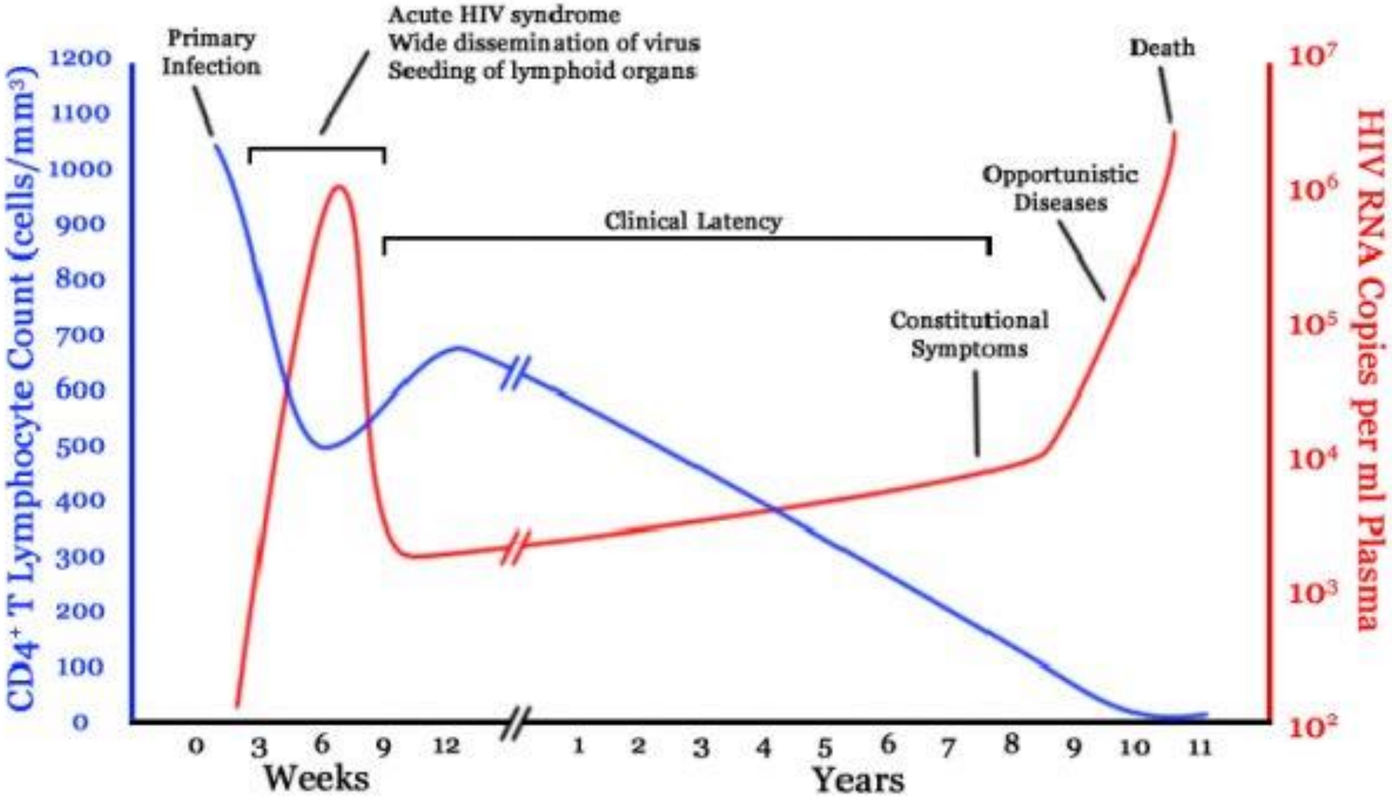


https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle

HIV lifecycle



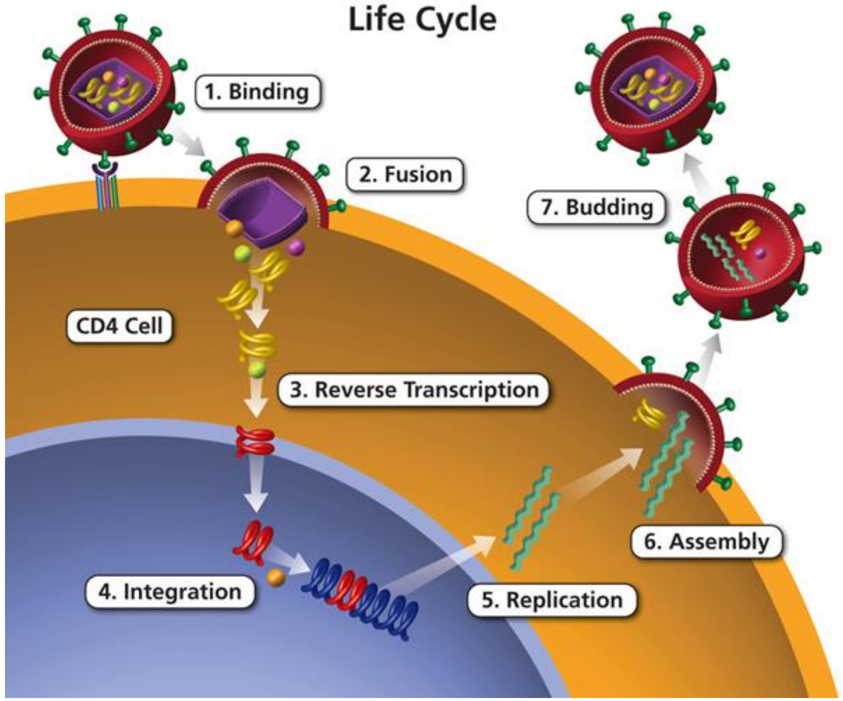
HIV infection in a human host

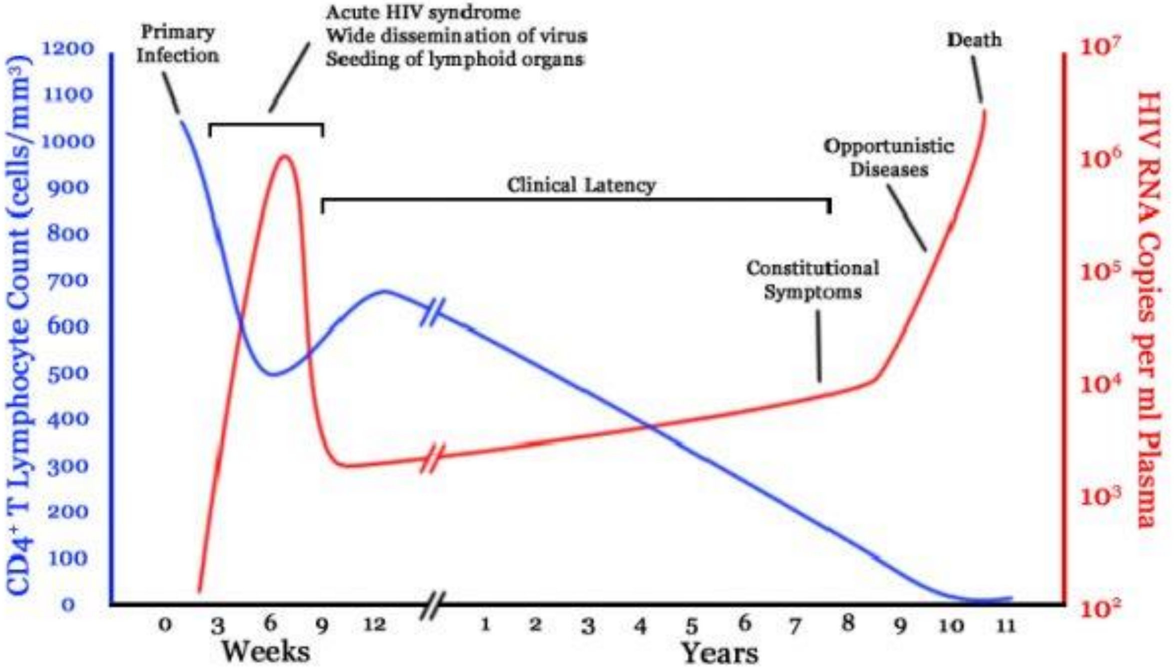


https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle

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?



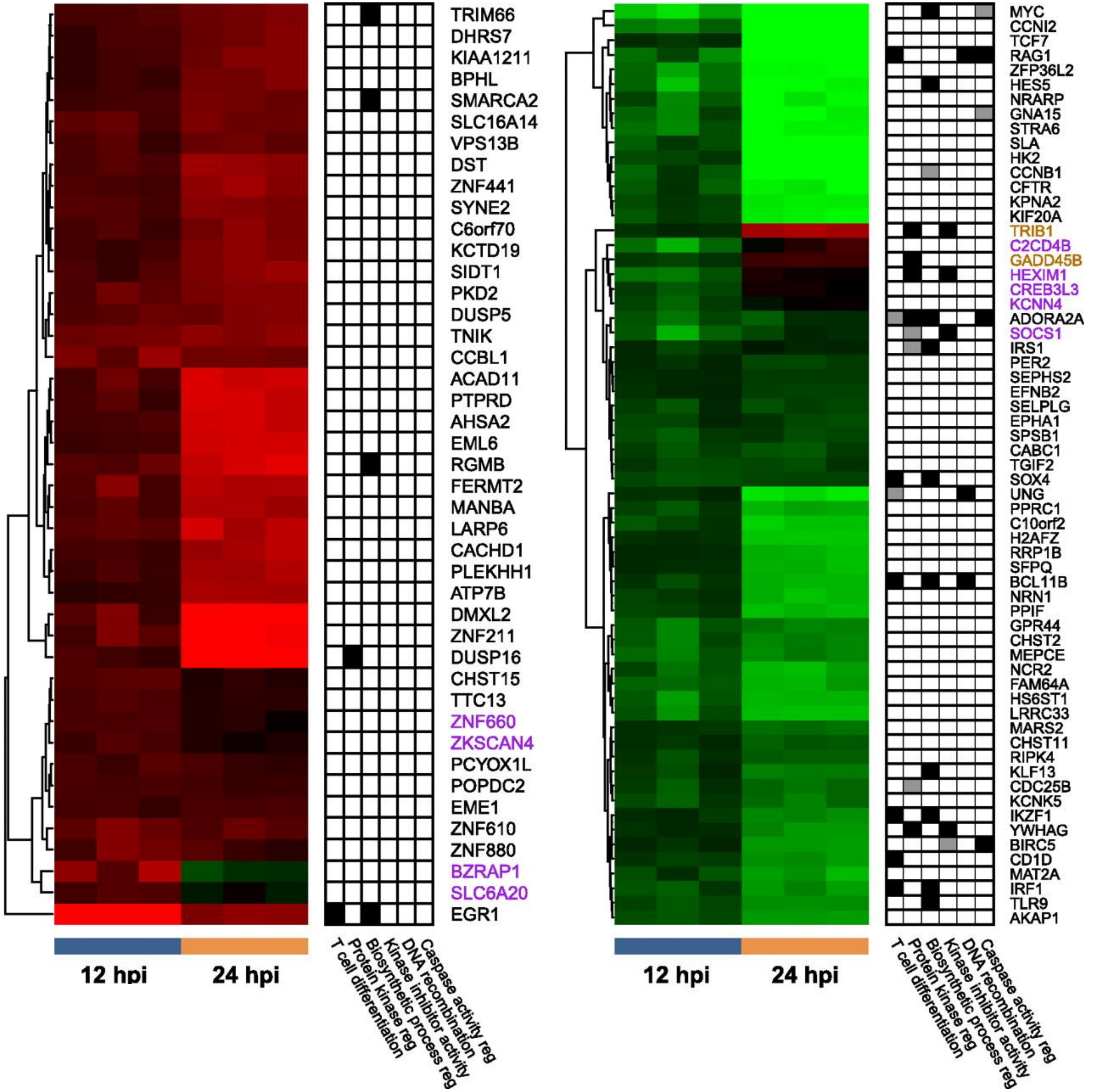


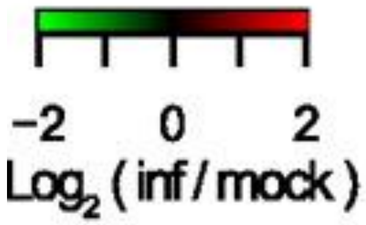
https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle

Study findings

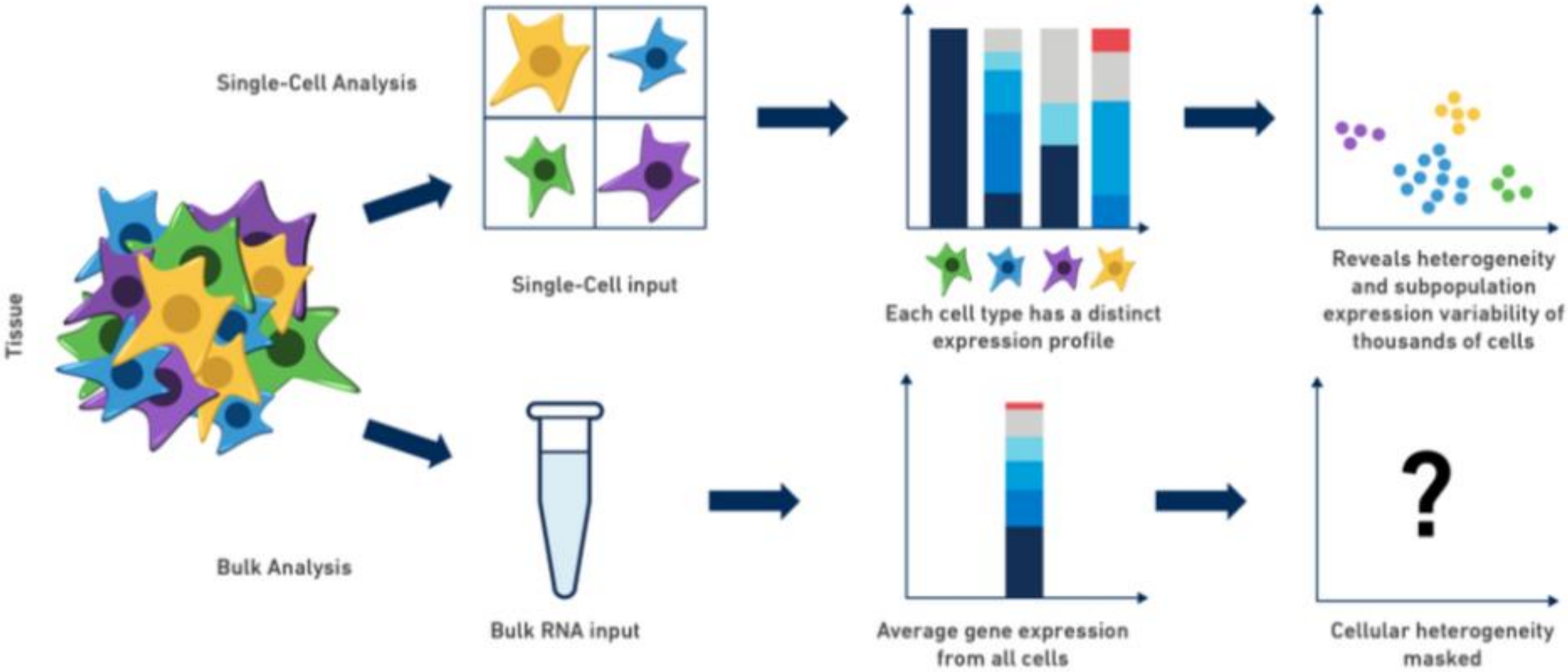
Using RNAseq, authors demonstrate:

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



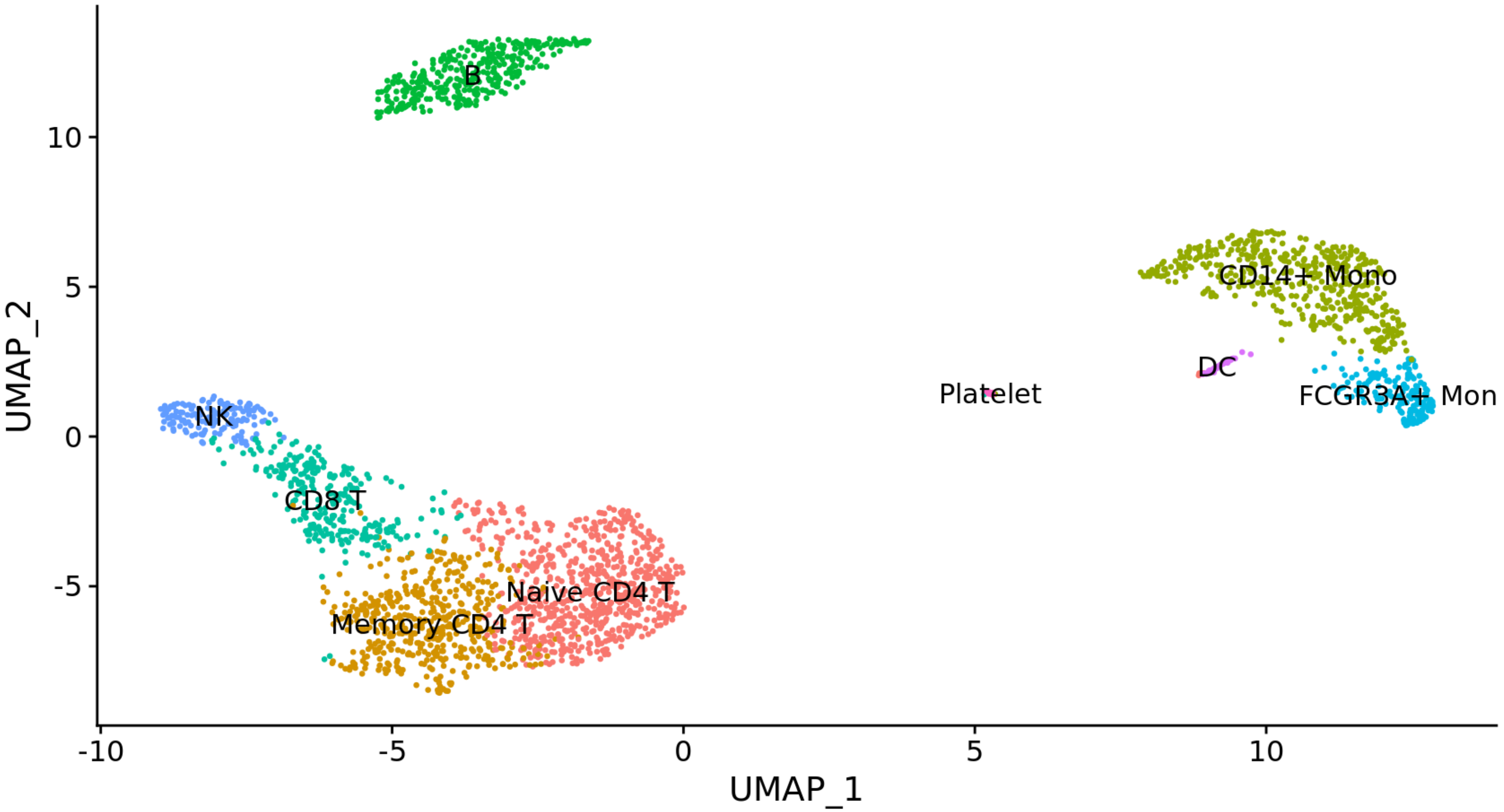


Bulk vs Single Cell RNA Sequencing


Single-Cell Analysis
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Tissue
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked

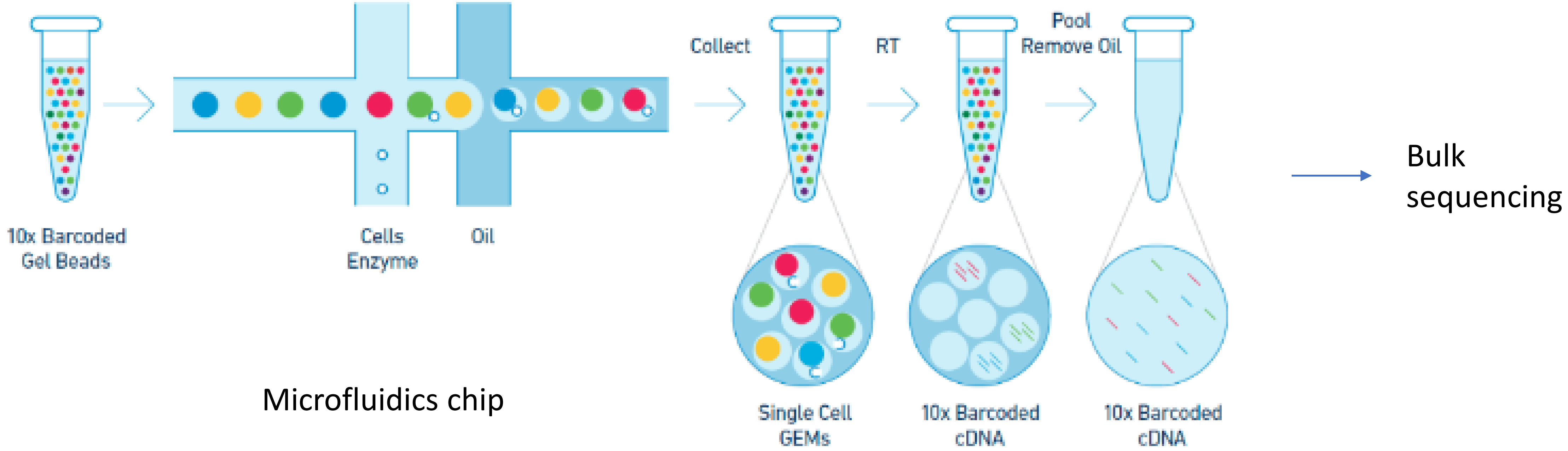
https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

scRNA cell subsets in PBMC



https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html

10x single cell technology



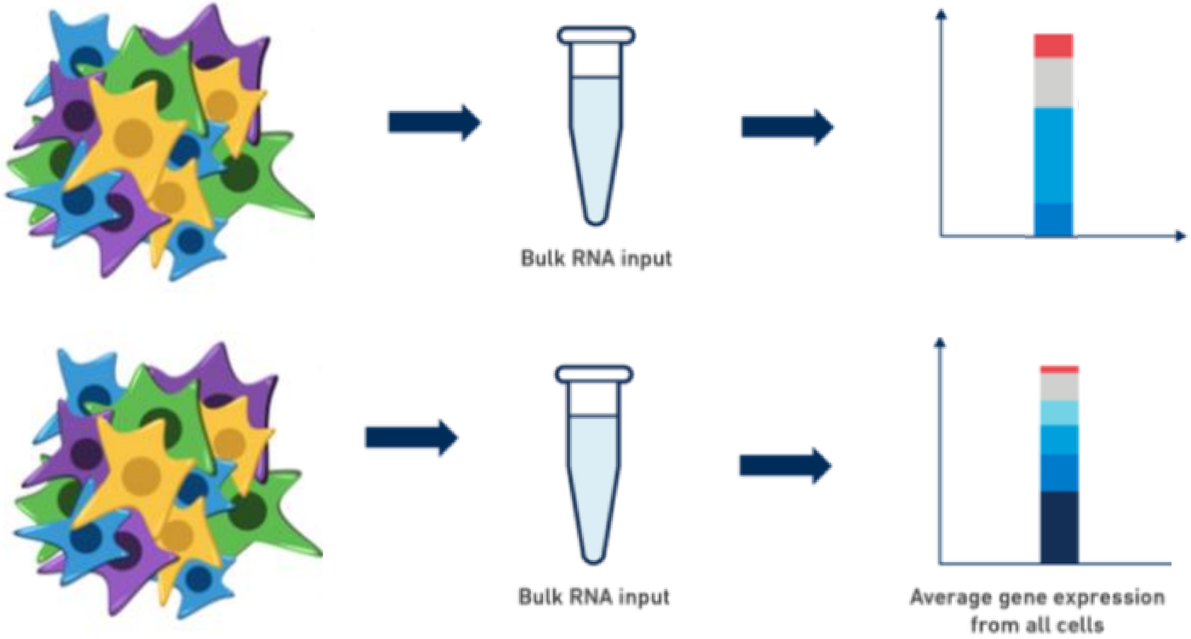
Microfluidics chip

https://github.com/hbctraining/scRNA-seq

Bulk RNAseq for Differential Expression is OK!

Uninfected

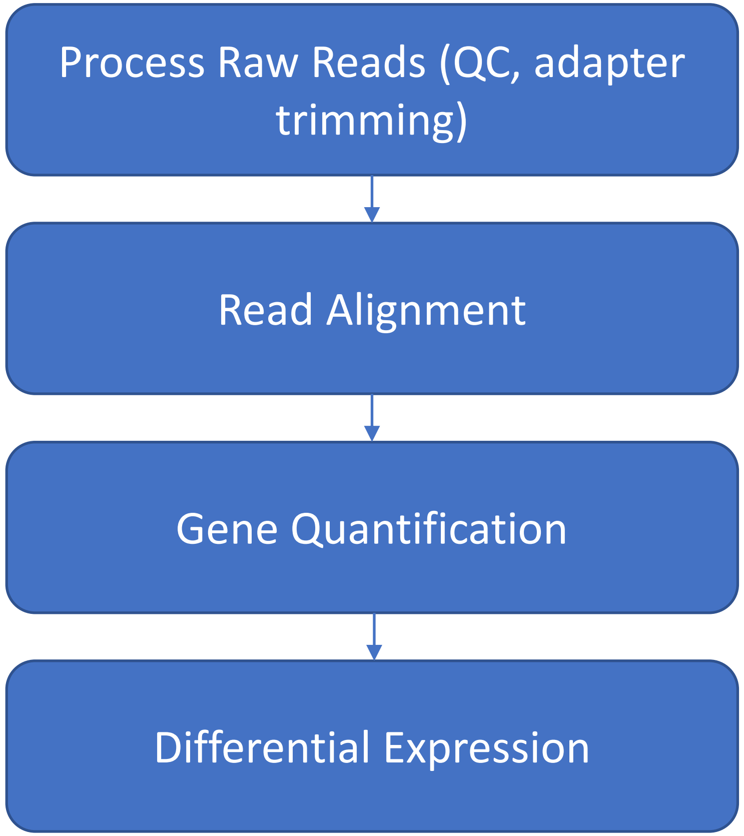
HIV+



Compare relative gene expression between conditions

https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

Our (bulk) RNAseq Workflow



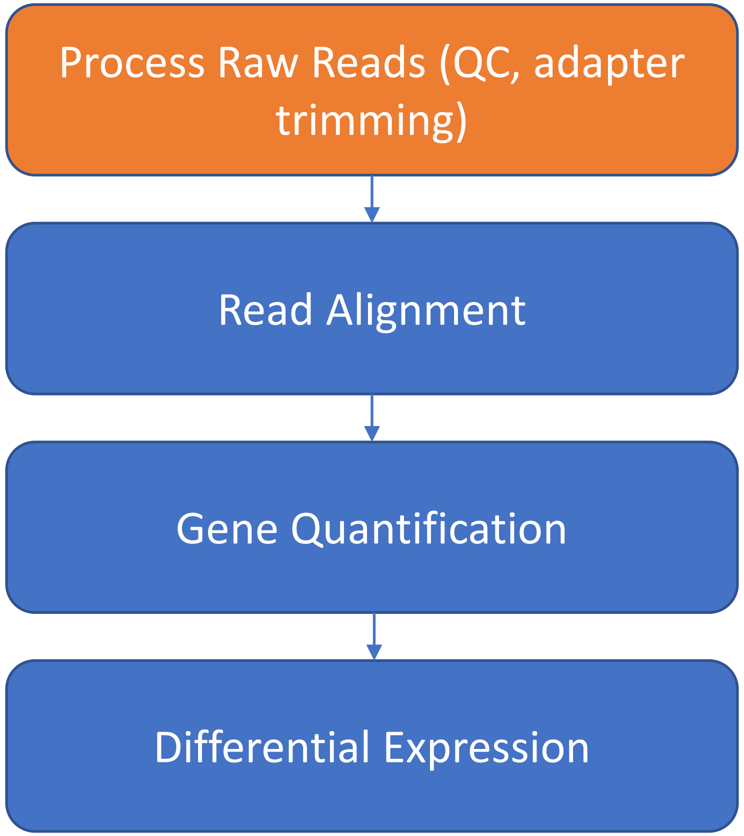
Process Raw Reads (QC, adapter trimming)

Read Alignment

Gene Quantification

Differential Expression

Quality control on Raw Reads



Process Raw Reads (QC, adapter trimming)

Read Alignment

Gene Quantification

Differential Expression

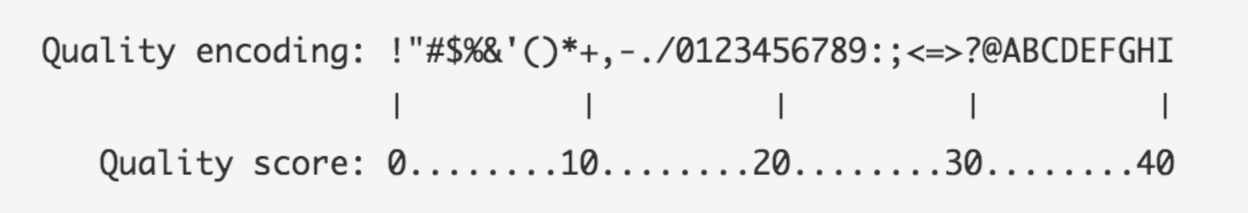
Raw reads in Fastq format

```
@SRR098401.109756285
GACTCACGTAACTTTAAACTCTAACAGAAATATACTA…
+
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
```

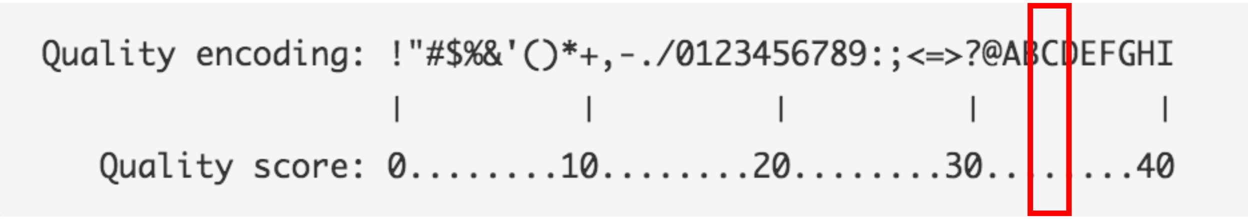
A quality score is a prediction of the probability of an error in base calling:

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%

https://www.illumina.com/science/education/sequencing-quality-scores.html

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:

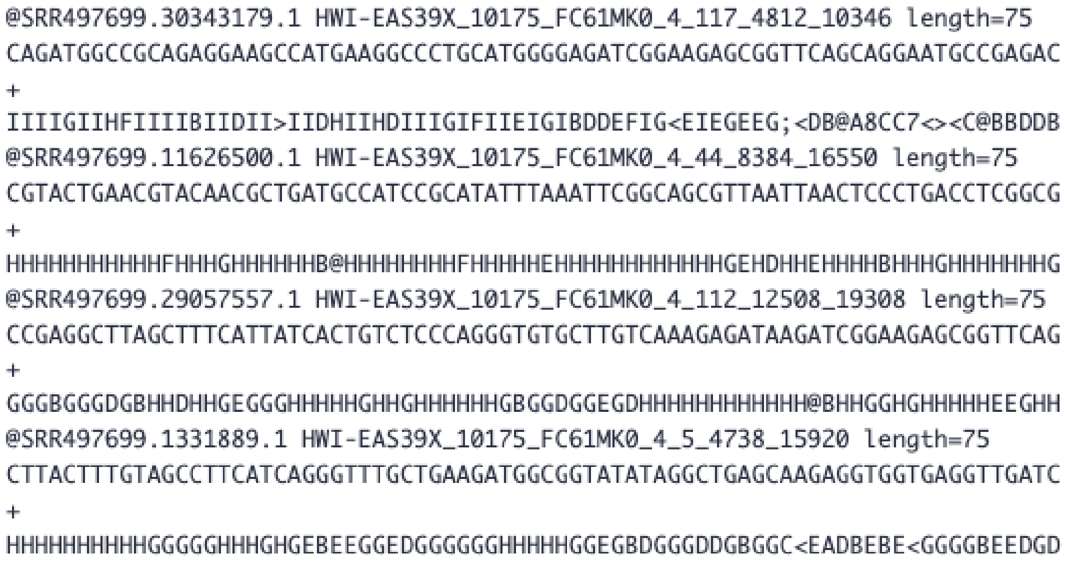
```
@SRR098401.109756285
GACTCACGTAACTTTAAACTCTAACAGAAATATACTA…
+
CAEFGDG?BCGGGEEDGGHGHGDFHEIEGGDDDD…
```

C –> Q = 34 -> Probability < 1/1000 of an error

https://www.illumina.com/science/education/sequencing-quality-scores.html

Raw read quality control

Fastq File



```
@SRR497699.30343179.1 HWI-EAS39X_10175_FC61MK0_4_117_4812_10346 length=75
CAGATGGCCGCAGAGGAAGCCATGAAGGCCCTGCATGGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGAC
+
IIIIGIIHFIIIIBIIDII>IIDHIIHDIIIGIFIIEIGIBDDEFIG<EIEGEEG;<DB@A8CC7<><C@BBDDB
@SRR497699.11626500.1 HWI-EAS39X_10175_FC61MK0_4_44_8384_16550 length=75
CGTACTGAACGTACAACGCTGATGCCATCCGCATATTTAAATTCGGCAGCGTTAATTAACTCCCTGACCTCGGCG
+
HHHHHHHHHHHFHHHGHHHHHHB@HHHHHHHHFHHHHHEHHHHHHHHHHHHGEHDHHEHHHHBHHHGHHHHHHHG
@SRR497699.29057557.1 HWI-EAS39X_10175_FC61MK0_4_112_12508_19308 length=75
CCGAGGCTTAGCTTTCATTATCACTGTCTCCCAGGGTGTGCTTGTCAAAGAGATAAGATCGGAAGAGCGGTTCAG
+
GGGBGGGDGBHHDHHGEGGGHHHHHGHHGHHHHHHGBGGDGGEGDHHHHHHHHHHH@BHHGGHGHHHHHEEGHH
@SRR497699.1331889.1 HWI-EAS39X_10175_FC61MK0_4_5_4738_15920 length=75
CTTACTTTGTAGCCTTCATCAGGGTTTGCTGAAGATGGCGGTATATAGGCTGAGCAAGAGGTGGTGAGGTTGATC
+
HHHHHHHHHHGGGGGHHHGHGEBEEGGEDGGGGGGHHHHHGGEGBDGGGDDGBGGC<EADBEBE<GGGGBEEDGD
```

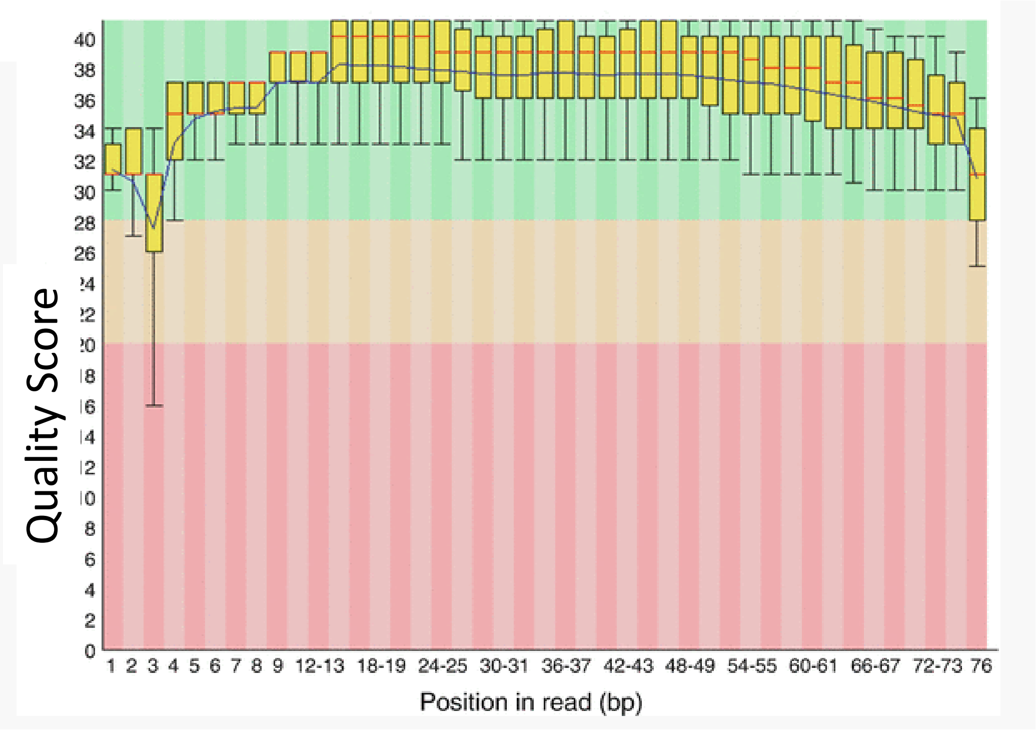
…

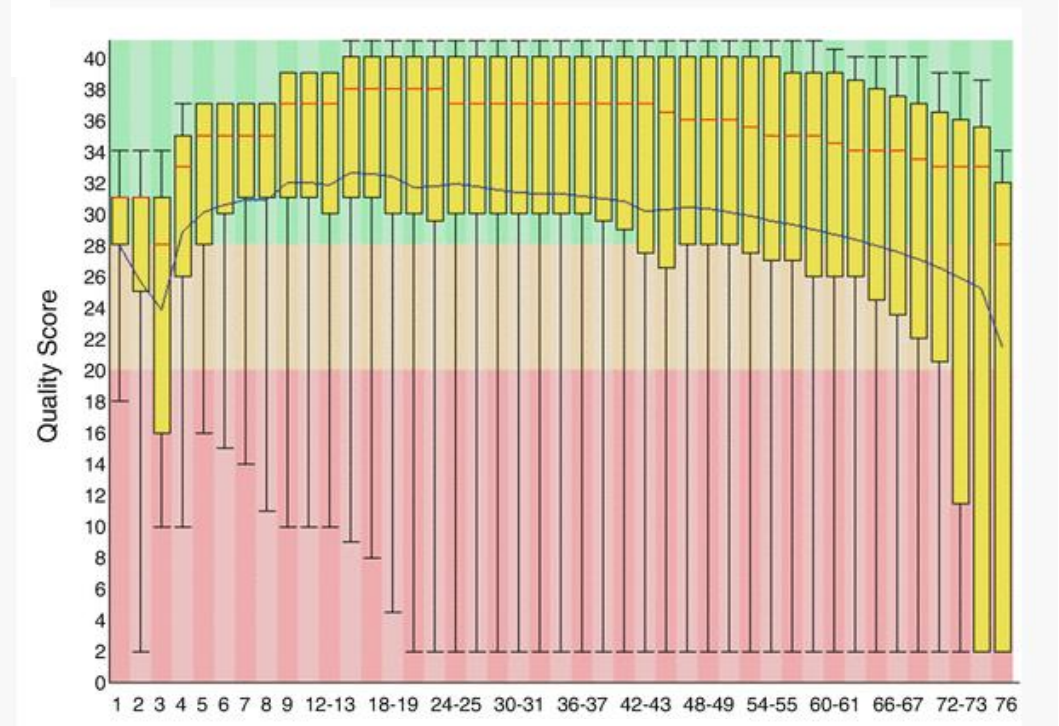
FastQC Tool →

Metrics

- Sequence Quality
- GC content
- Per base sequence content
- Adapters in Sequence

FastQC: Sequence Quality Histogram





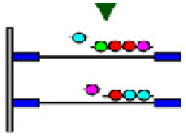
Position in read (bp)

Position in read (bp)

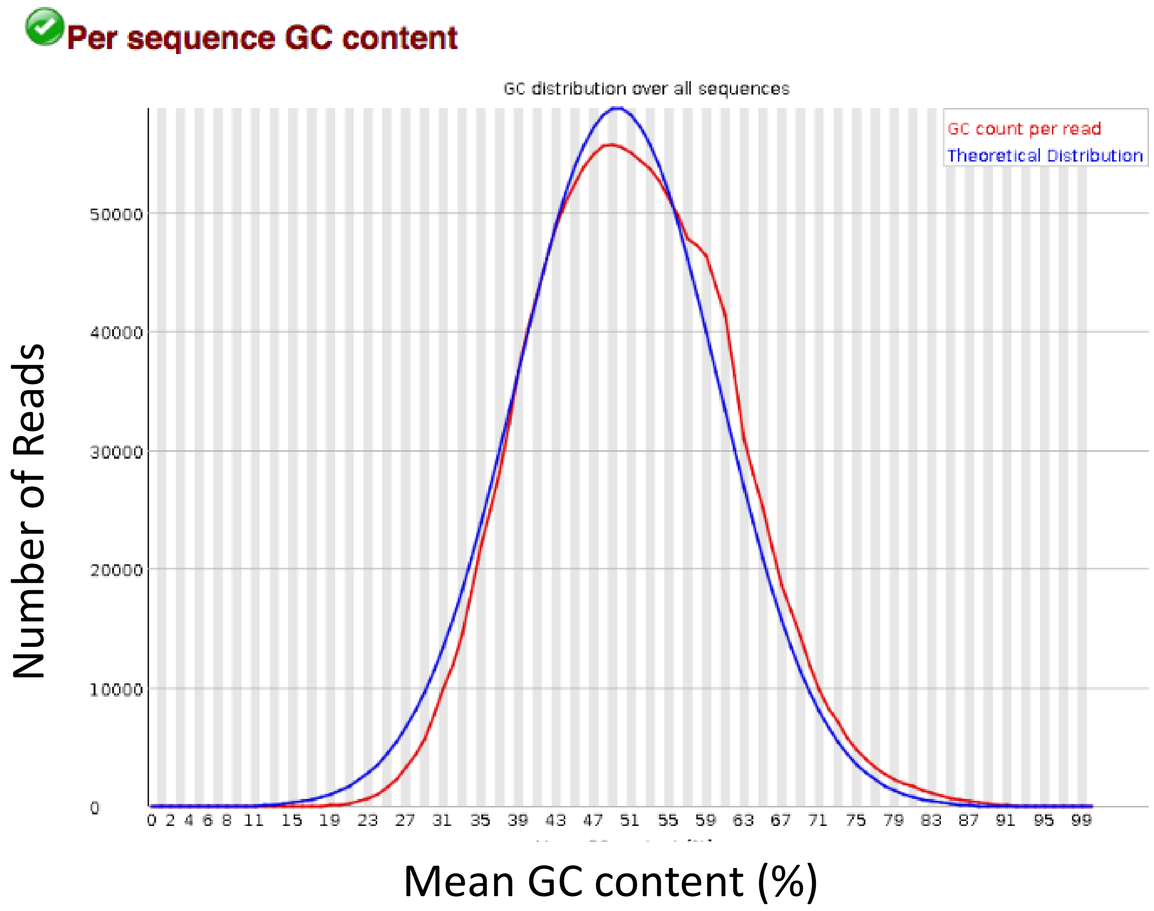
GOOD
High quality over the length of the read

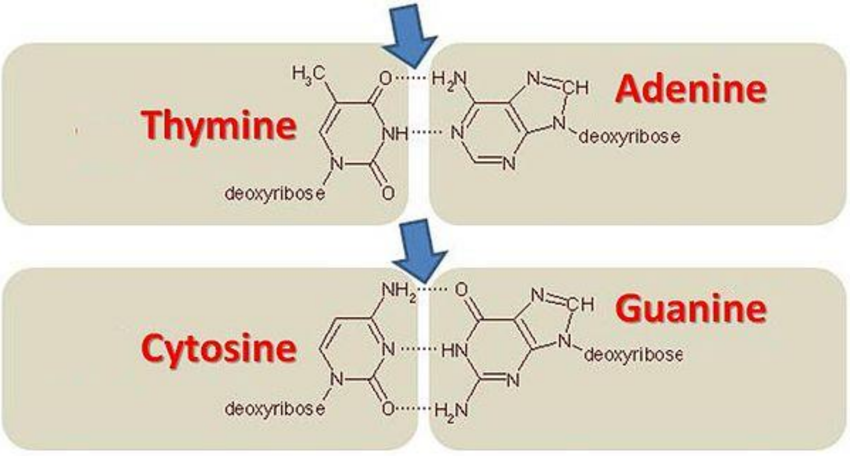
BAD
Read quality drops at the beginning and end

Sequencing



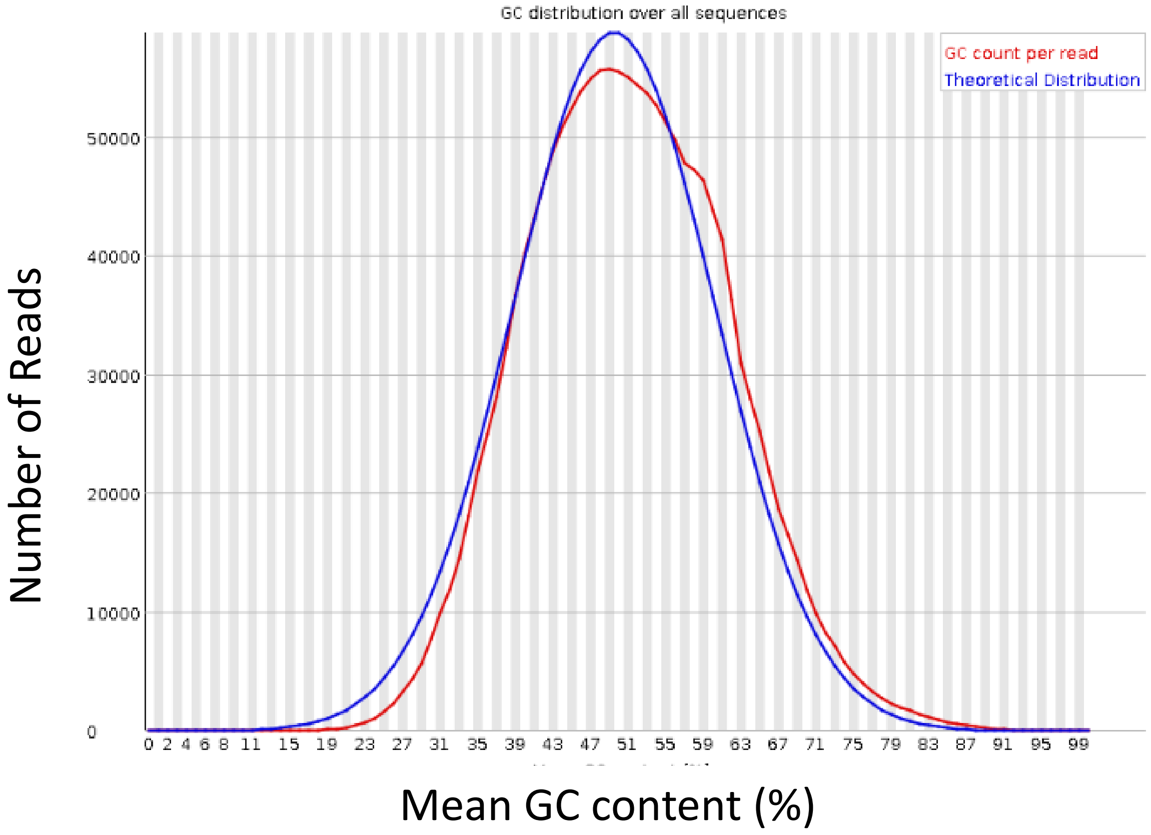
FastQC: Per sequence GC content





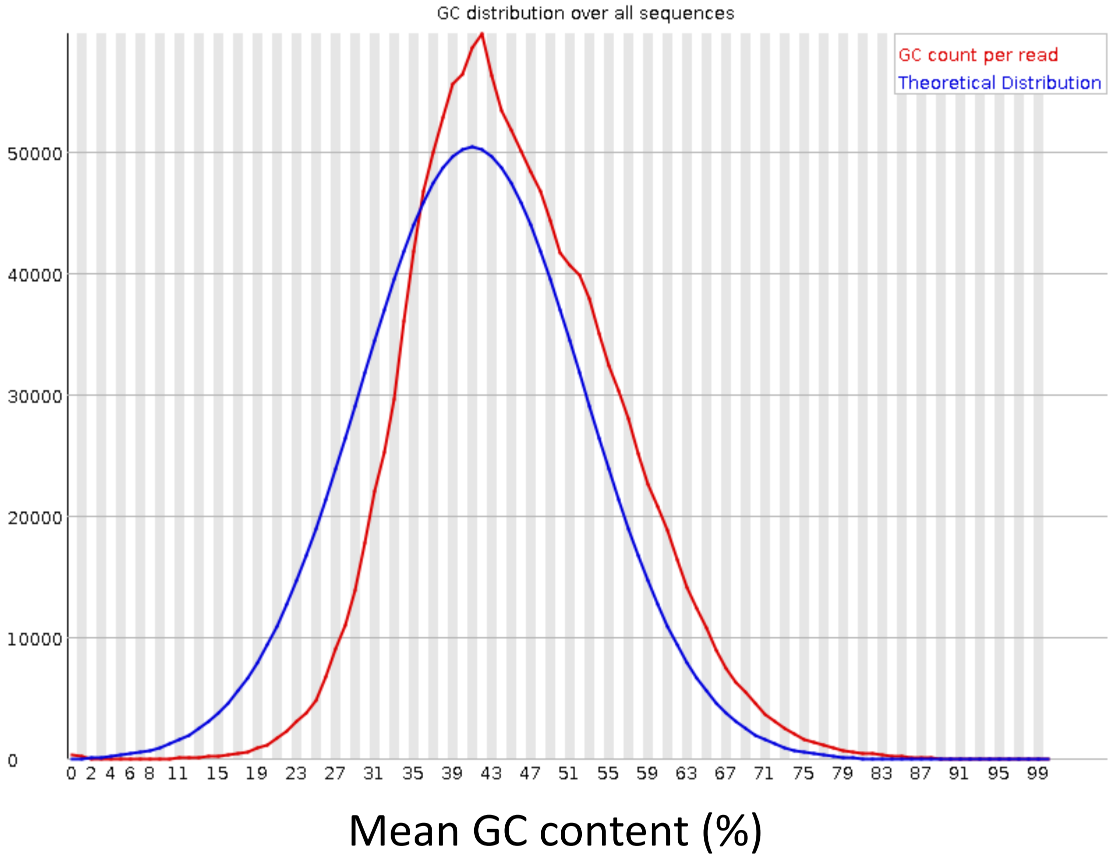
FastQC: Per sequence GC content

✓ **Per sequence GC content**



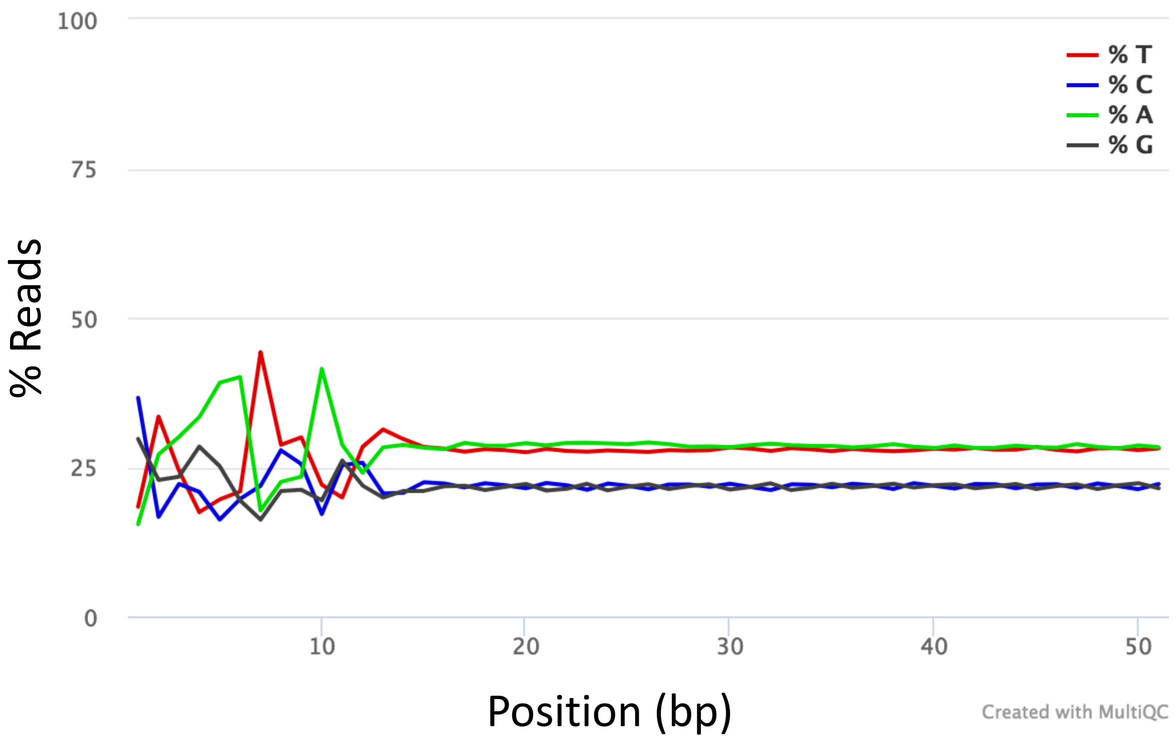
GOOD: follows normal distribution (sum of deviations is < 15% of reads)

✗ **Per sequence GC content**



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

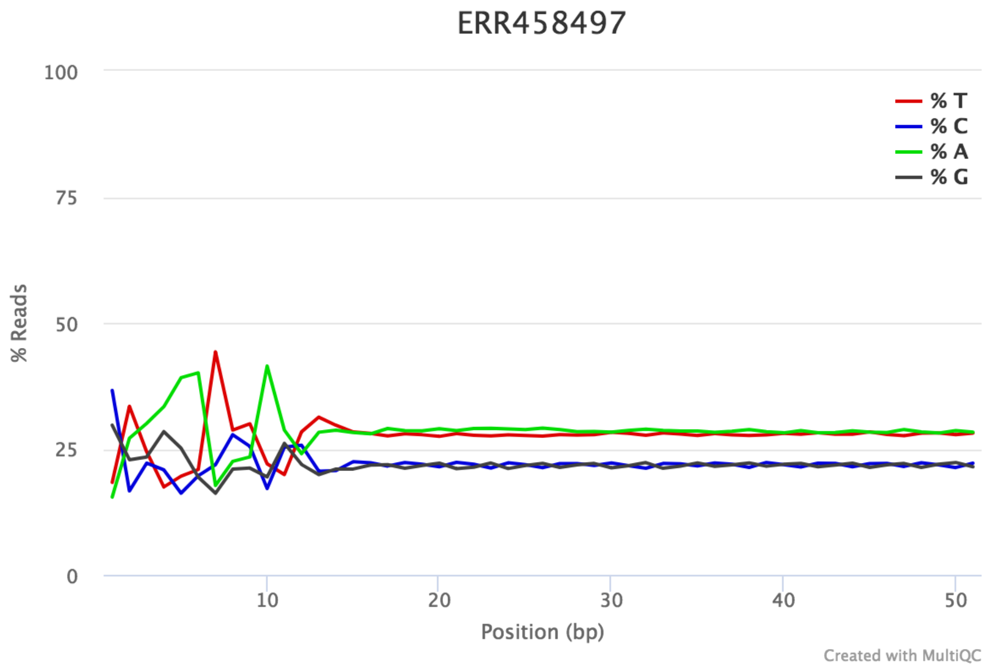
FastQC: Per Base Sequence Content



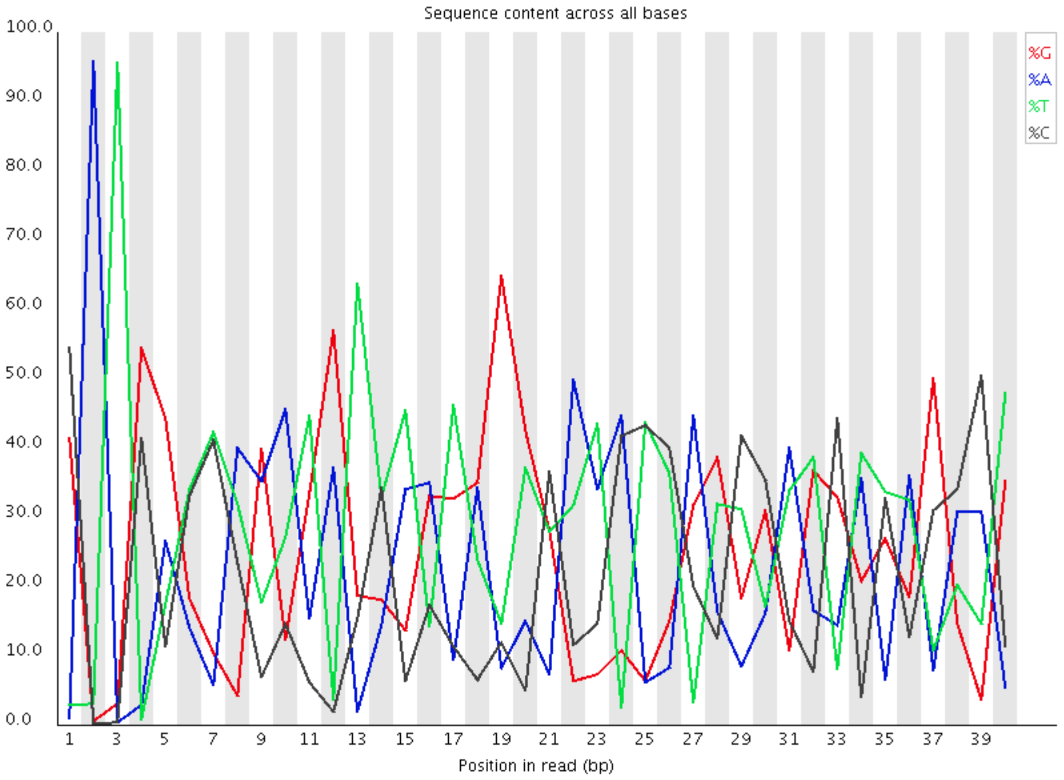
- Proportion of each position for which each DNA base has been called
- RNAseq data tends to show a positional sequence bias in the first ~12 bases
- The "random" priming step during library construction is not truly random and certain hexamers are more prevalent than others

sequencing.qcfail.com

FastQC: Per Base Sequence Content

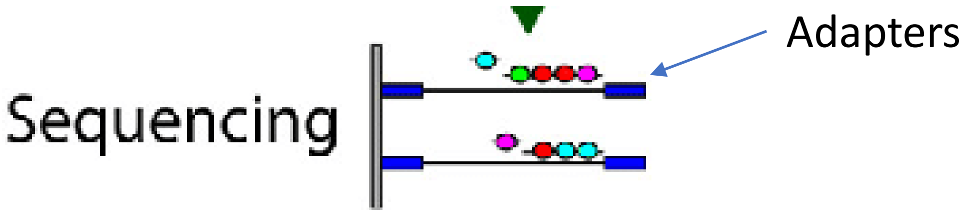


EXPECTED for RNAseq



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



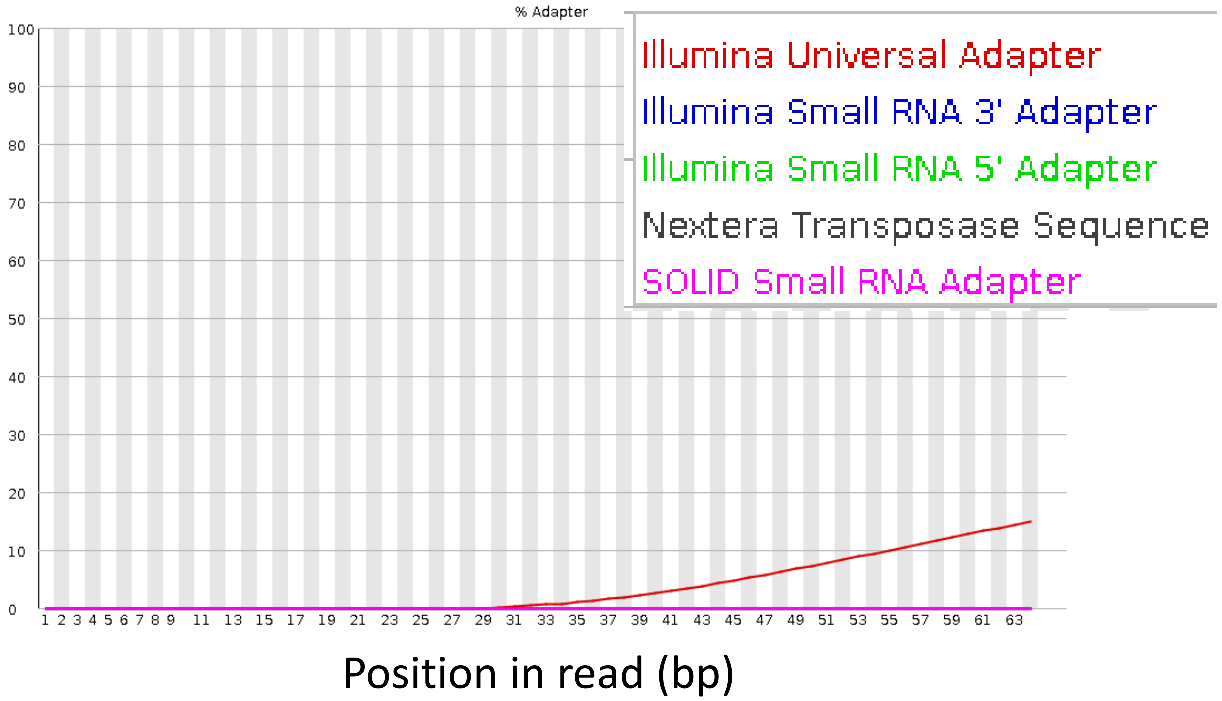
Sequencing

Adapters

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

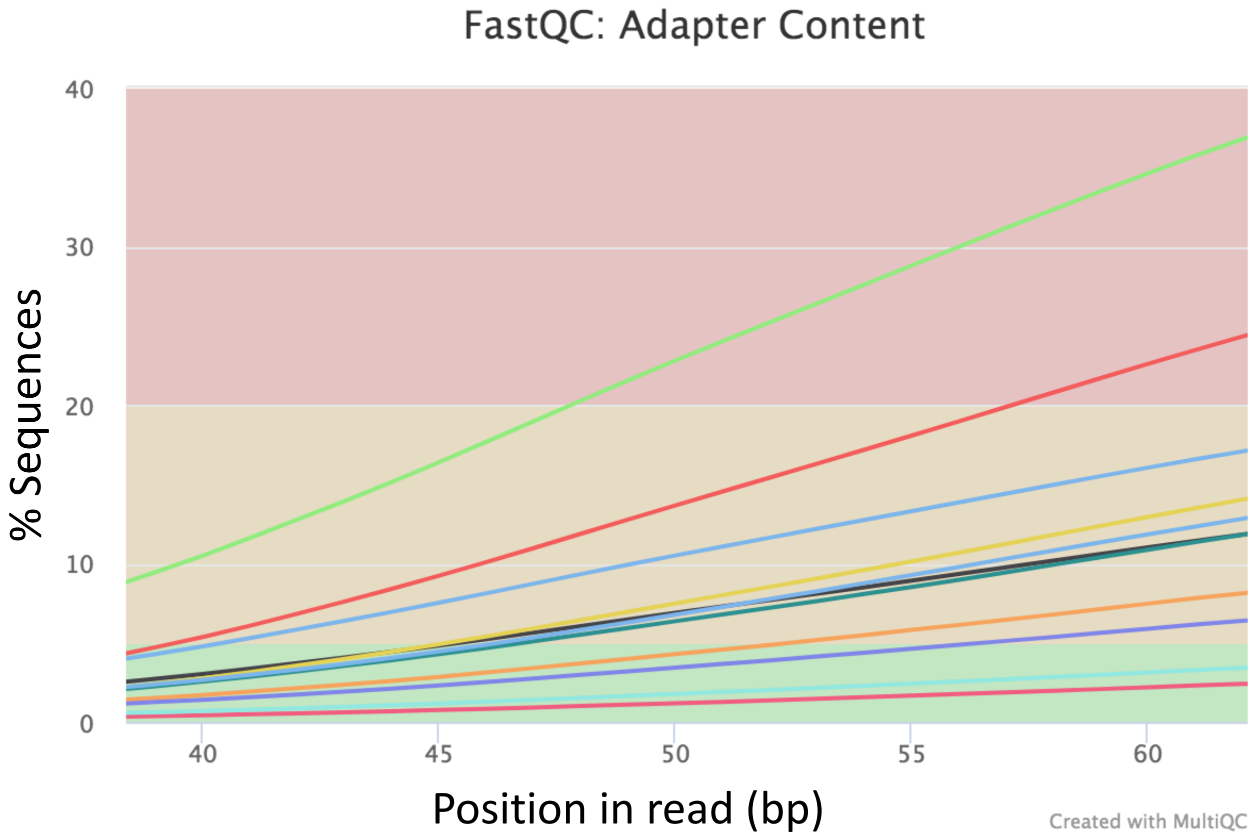
Solution – Adapter trimming!



sequencing.qcfail.com

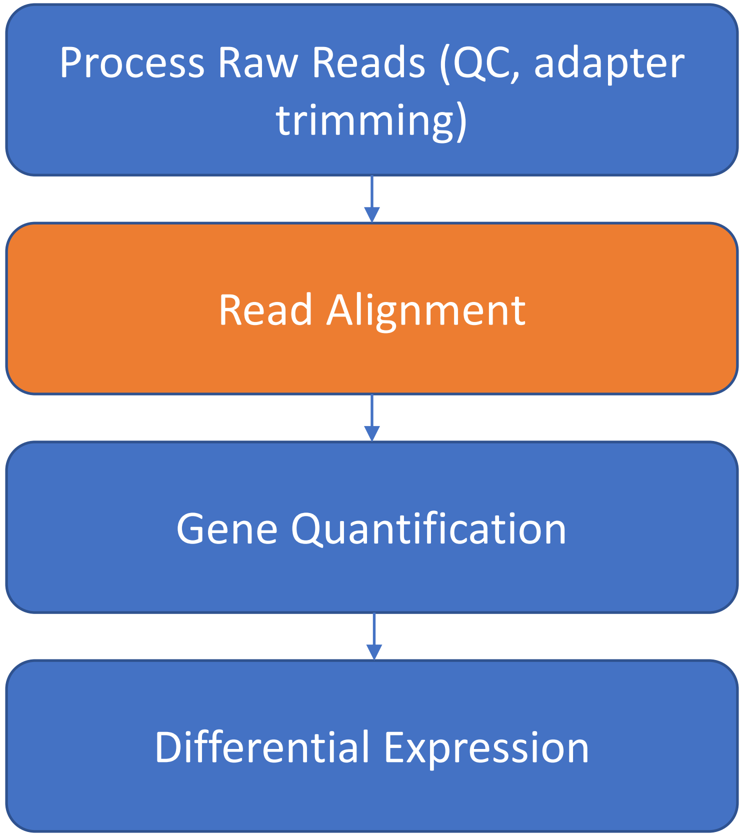
FastQC -> MultiQC

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



We'll use a tool called "Trim Galore!" to trim adapters and remove low quality bases/reads.

Workflow



Process Raw Reads (QC, adapter trimming)

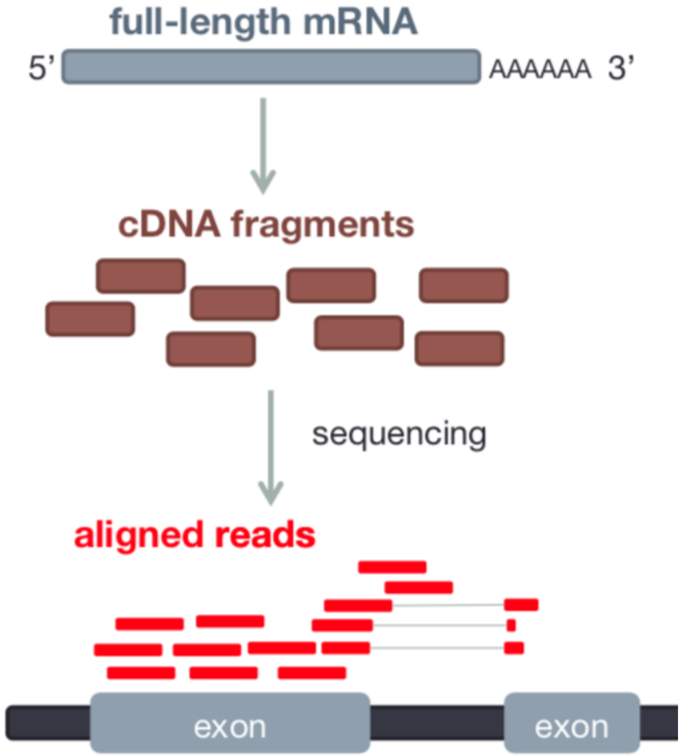
Read Alignment

Gene Quantification

Differential Expression

Read Alignment

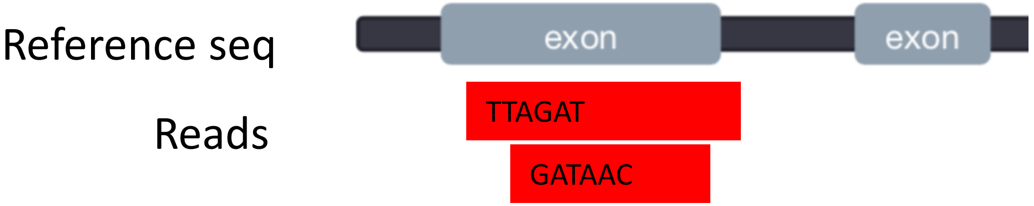
- RNAseq data originates from spliced mRNA (no introns)
- When aligning to the genome, our aligner must find a spliced alignment for reads
- We use a tool called STAR (Spliced Transcripts Alignment to a Reference) that has a exon-aware mapping algorithm.

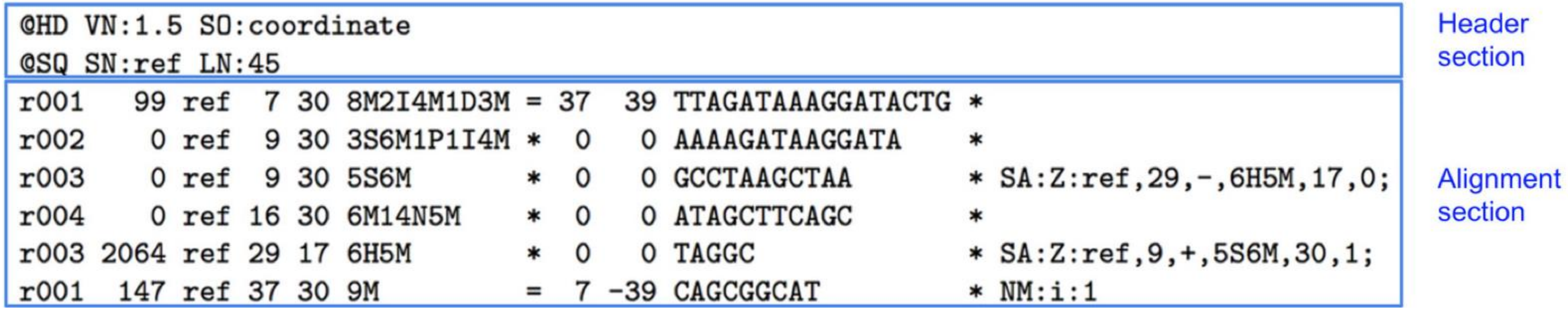


Reference sequence

Dobin et al Bioinformatics 2013

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 GCCTAAGCTAA       * 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

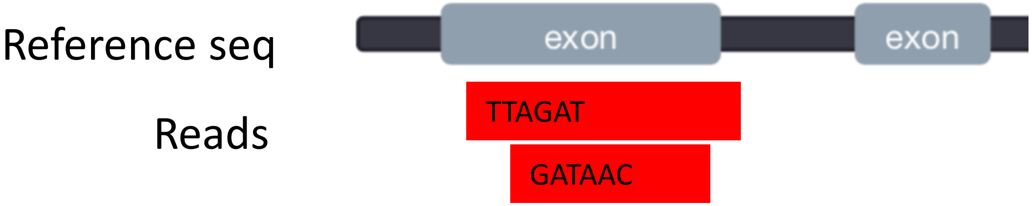
Ref Sequence name

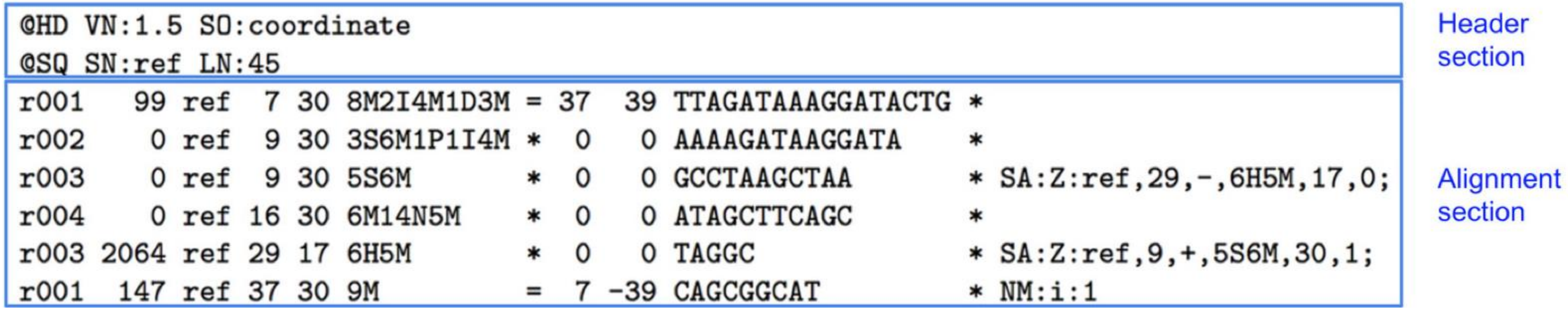
Flag: indicates alignment information e.g. paired, aligned, etc
https://broadinstitute.github.io/picard/explain-flags.html

Read ID

www.samformat.info

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 GCCTAAGCTAA       * 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

www.samformat.info

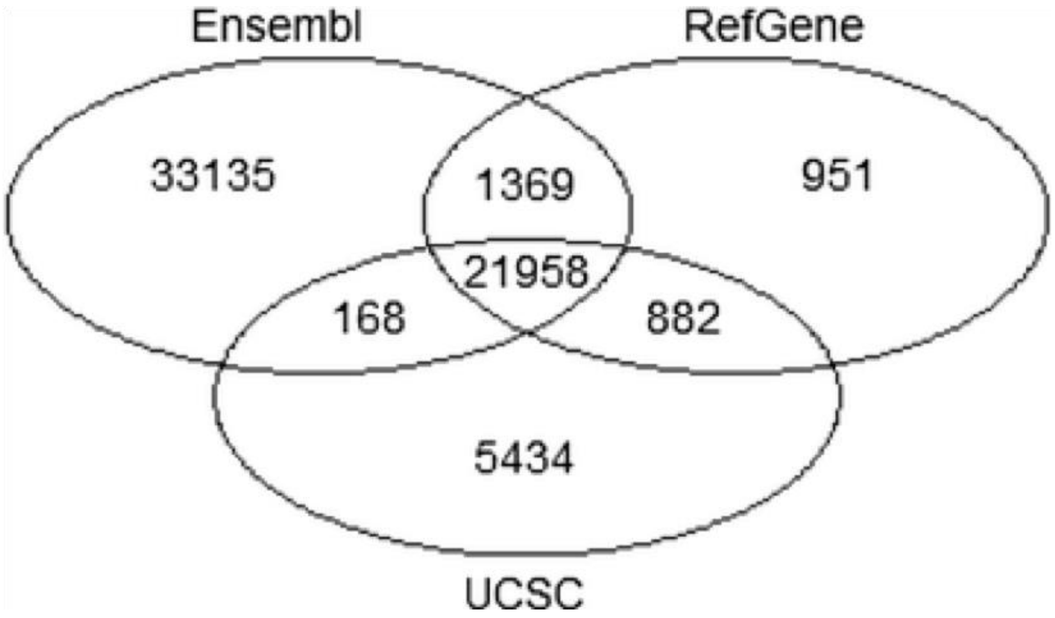
Genome Annotation Standards

- 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!



Zhao et al Bioinformatics 2015

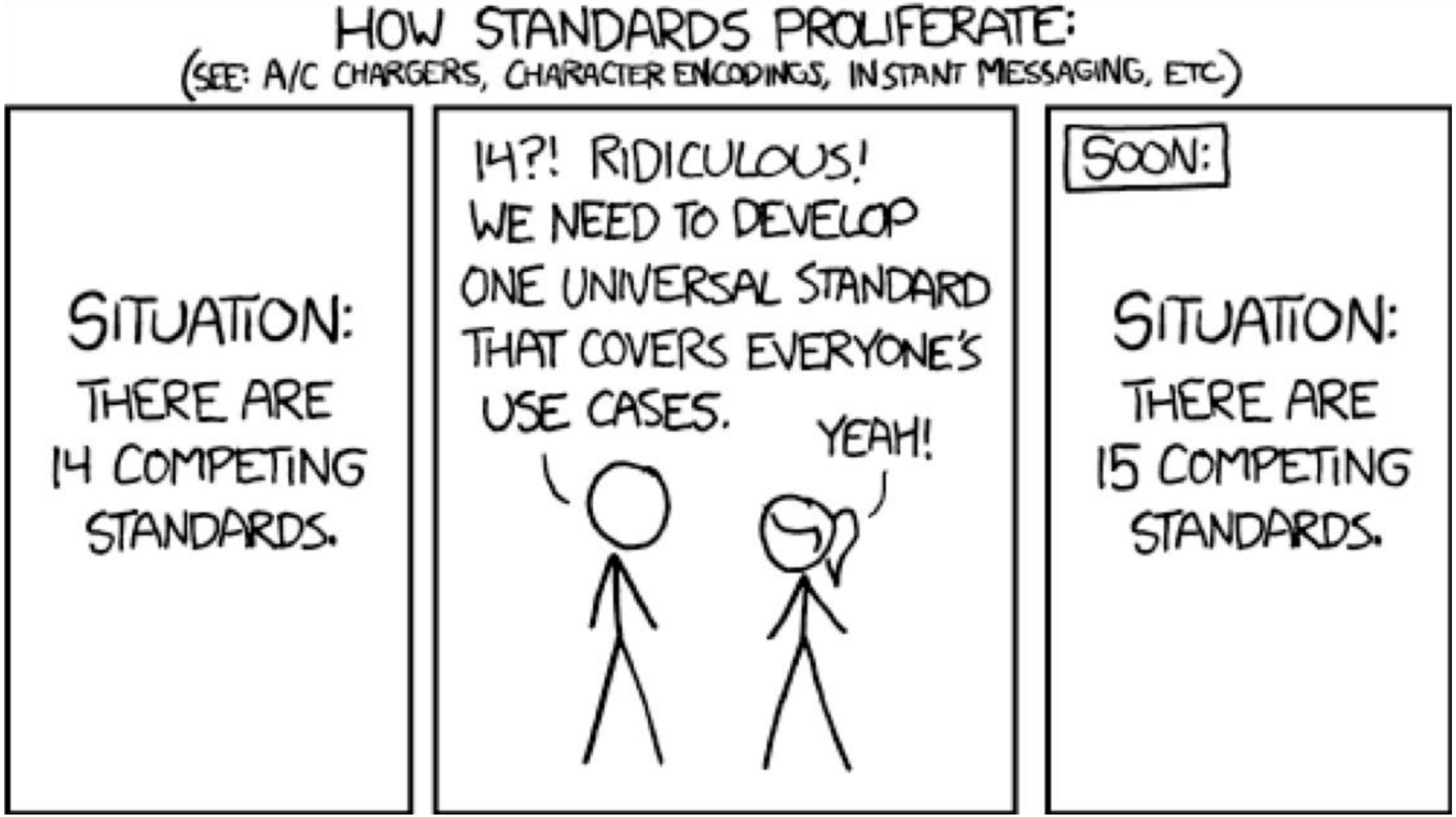
Gene Annotation Format (GTF)

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";

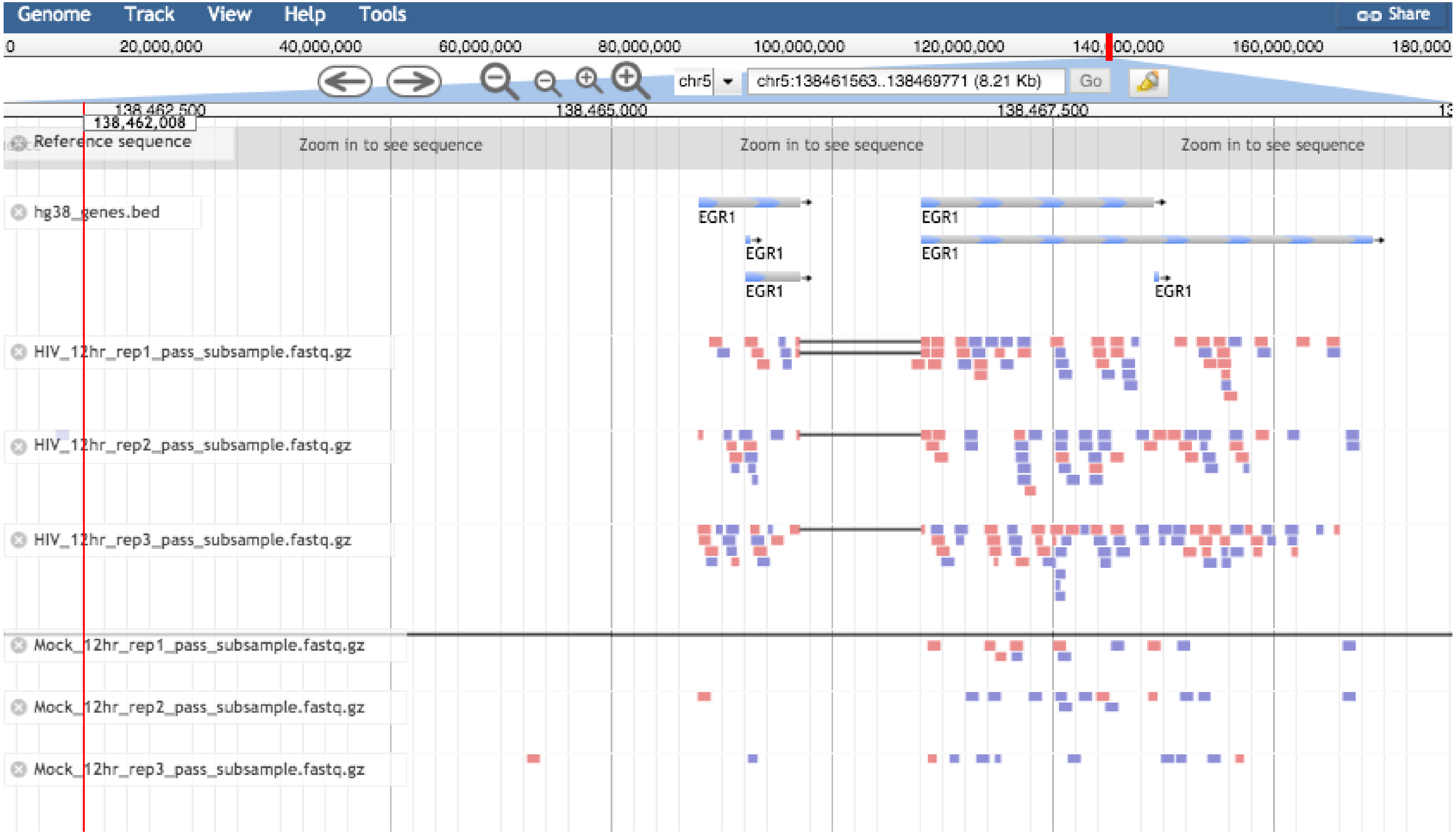
https://useast.ensembl.org/info/website/upload/gff.html

A note on standards

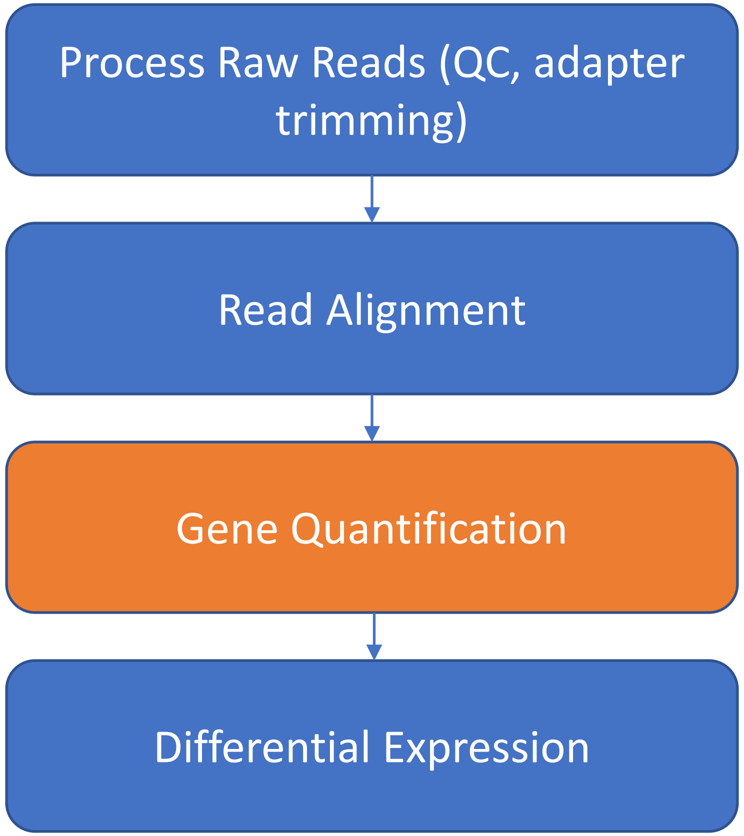

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

https://xkcd.com/927/

Visualizing reads with JBrowse

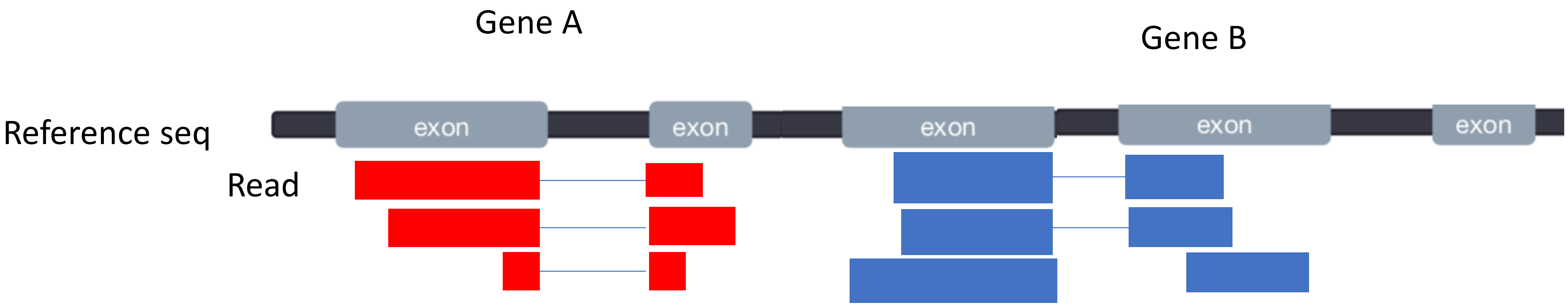


Workflow



Process Raw Reads (QC, adapter trimming)
Read Alignment
Gene Quantification
Differential Expression

Counting reads for each gene

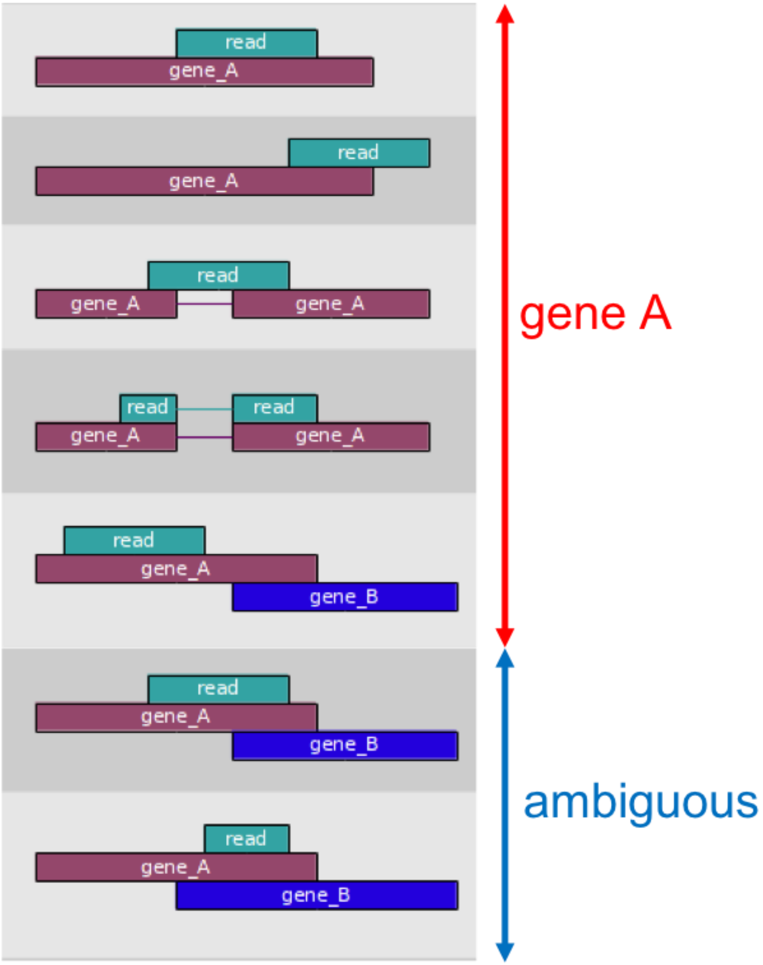


Counting reads: featurecounts

- The mapped coordinates of each read are compared with the features in the GTF file

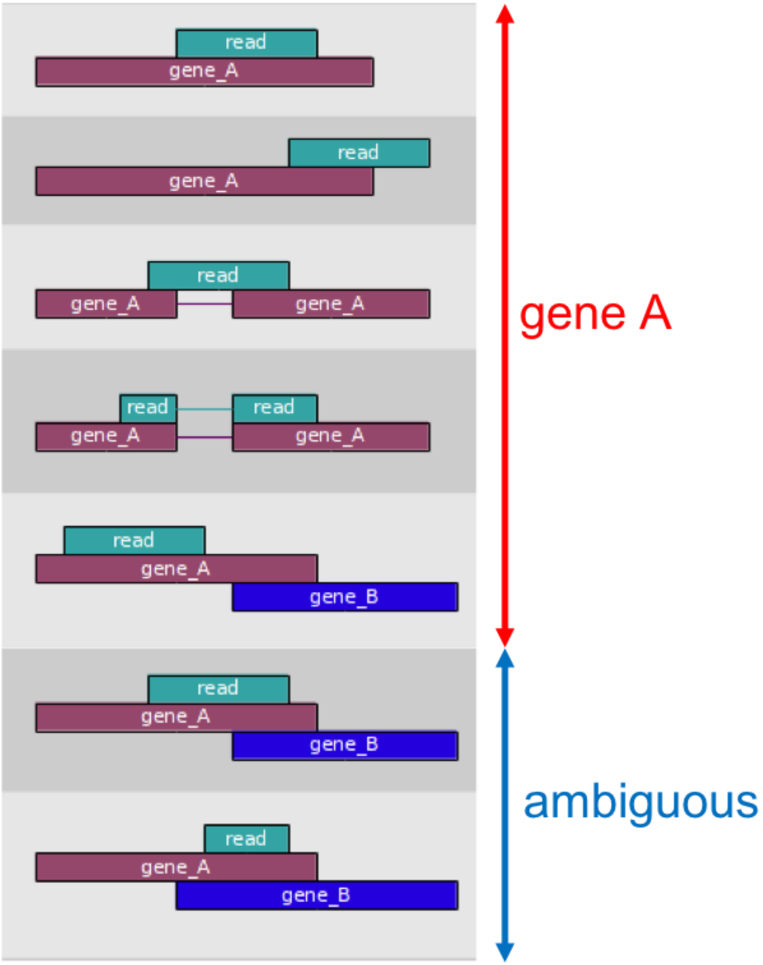
- Reads that overlap with a gene by >=1 bp are counted as belonging to that feature

- Ambiguous reads will be discarded



Counting reads: featurecounts

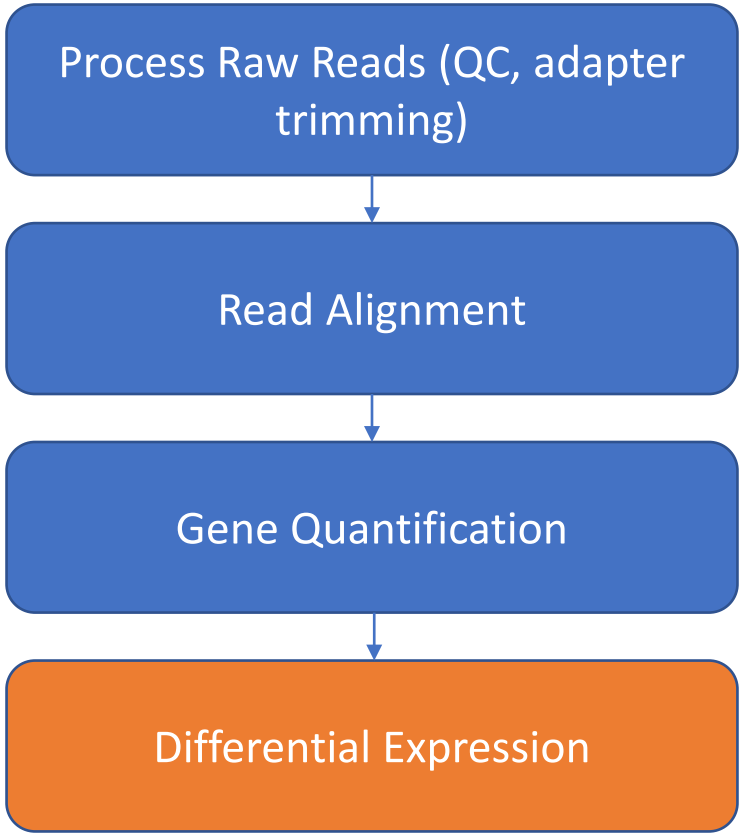
- The mapped coordinates of each read are compared with the features in the GTF file
- Reads that overlap with a gene by >=1 bp are counted as belonging to that feature
- Ambiguous reads will be discarded



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

Workflow



Process Raw Reads (QC, adapter trimming)

Read Alignment

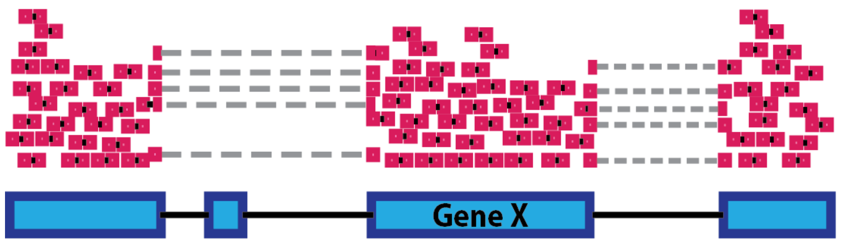
Gene Quantification

Differential Expression

Normalization

- Raw Count != Expression strength
- Normalization:
  - Eliminates factors that are not of interest for our experiment
  - Enables accurate comparison between samples or genes

Sample A Reads

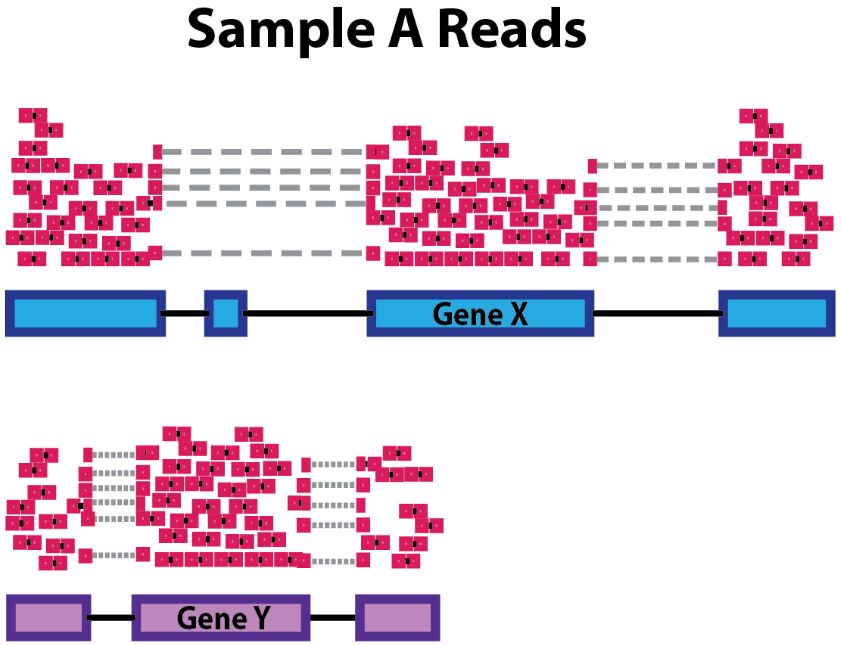


Adapted from https://hbctraining.github.io/DGE_workshop

Normalization

The number of reads mapped to a gene depends on

- **Gene Length**

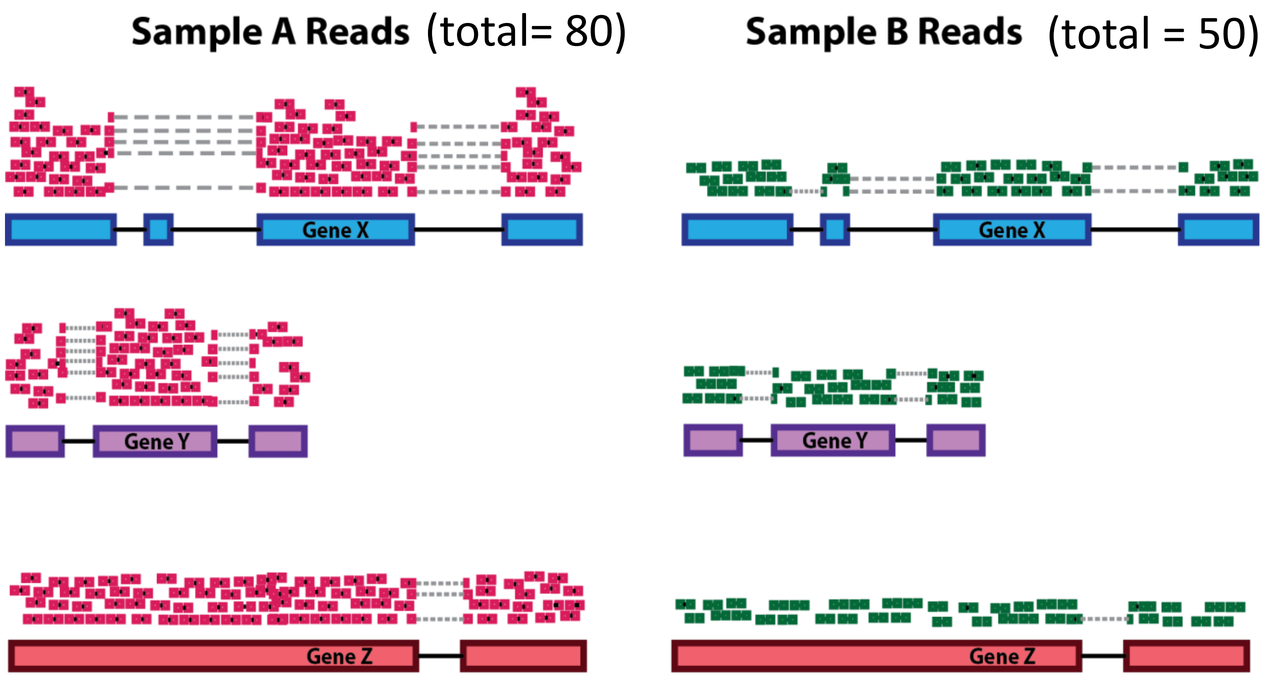


Adapted from https://hbctraining.github.io/DGE_workshop

Normalization

The number of reads mapped to a gene depends on

- Gene Length
- **Sequencing depth**

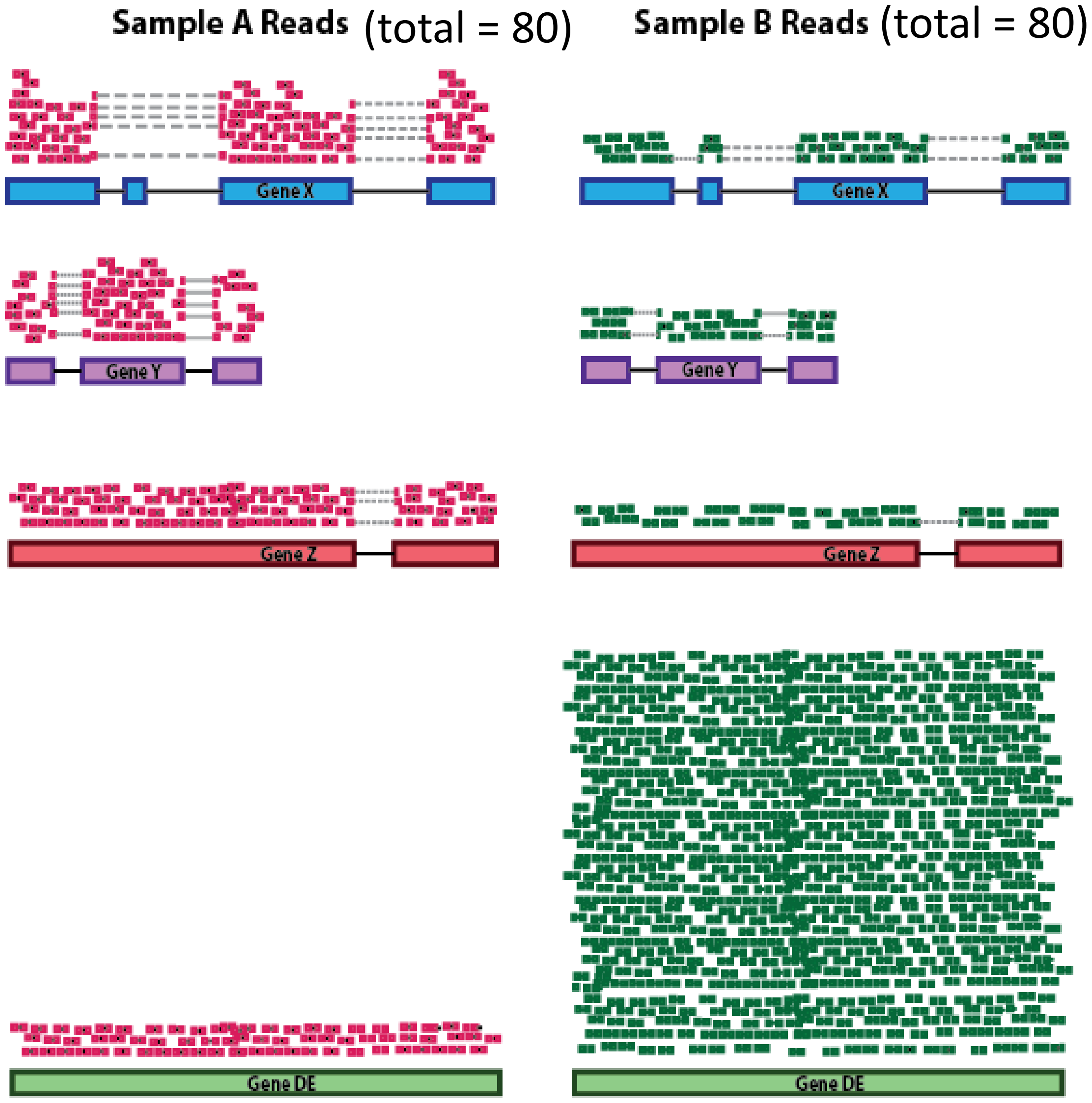


Adapted from https://hbctraining.github.io/DGE_workshop

Normalization

The number of reads mapped to a gene depends on

- Gene Length
- Sequencing depth
- **The expression level of other genes in the sample (RNA Composition)**



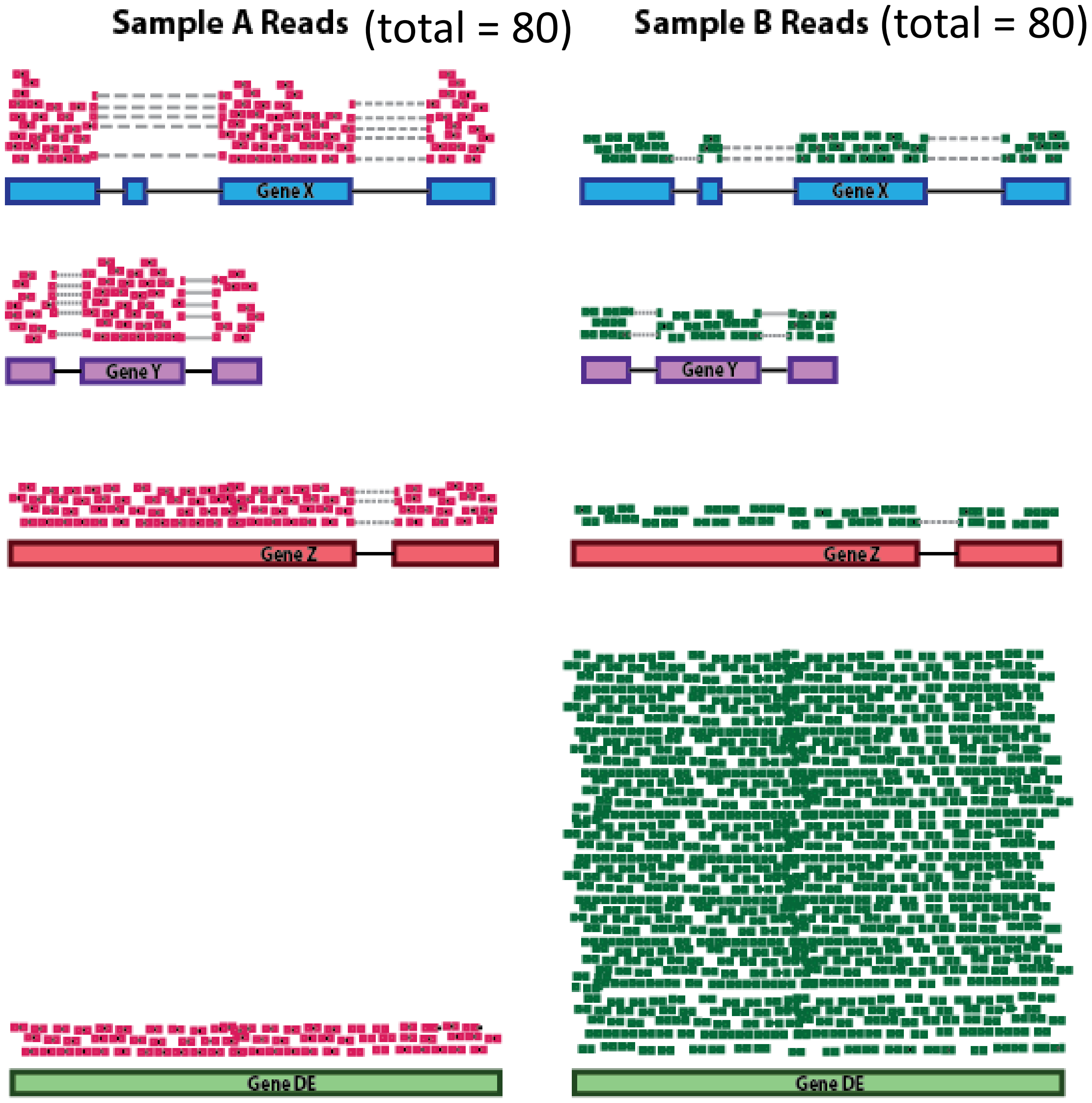
Adapted from https://hbctraining.github.io/DGE_workshop

Normalization

The number of reads mapped to a gene depends on

- Gene Length
- **Sequencing depth**
- **The expression level of other genes in the sample (RNA Composition)**

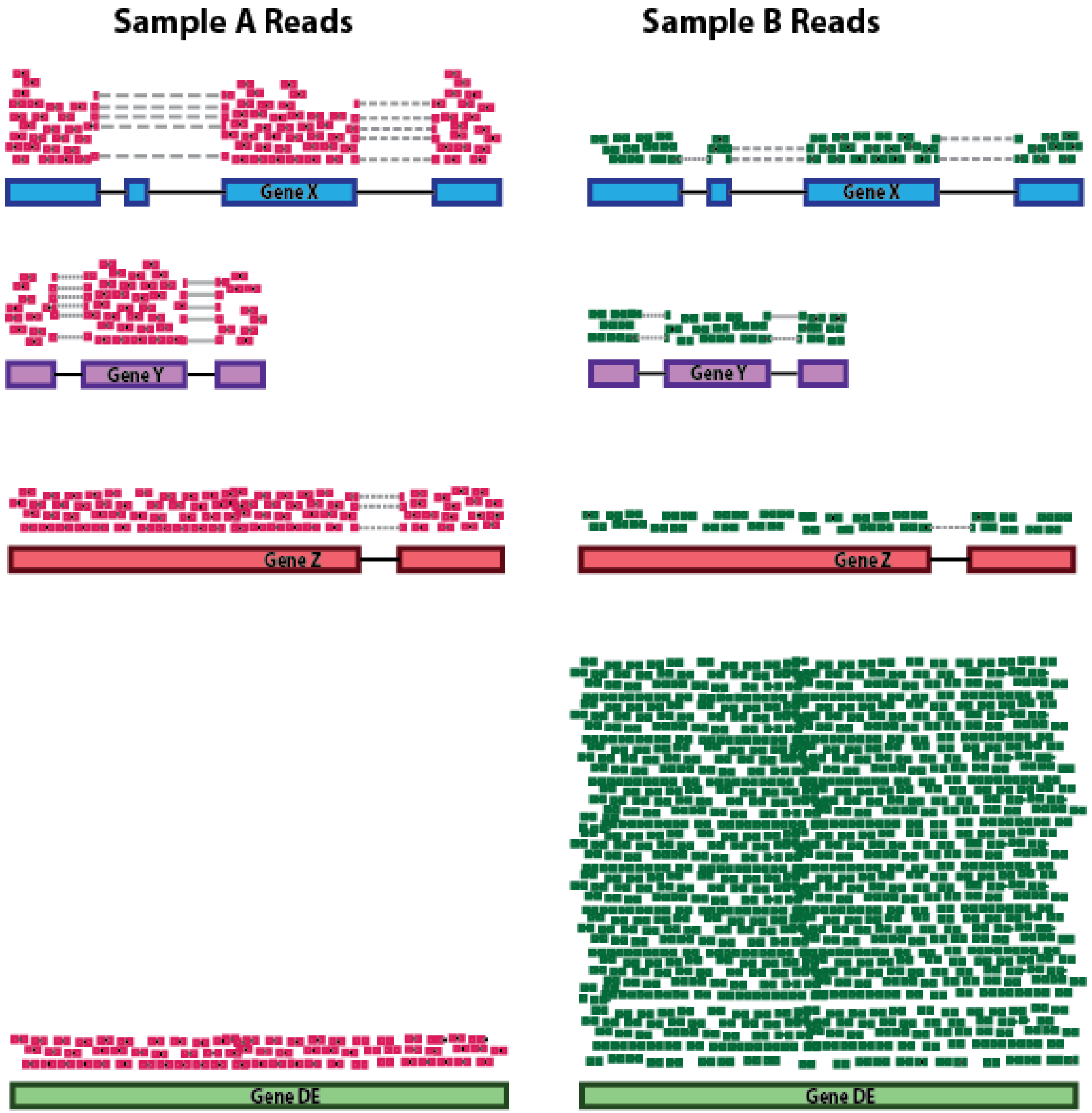
DESeq2 Median of Ratios



Adapted from https://hbctraining.github.io/DGE_workshop

Normalization: DESeq2 Median of Ratios

Gene	Sample A	Sample B
X	26	10
Y	26	10
Z	26	10
DE	2	50
Total =	80	80



Normalization: DESeq2 Median of Ratios

1. Take a row-wise average to produce an average sample (geometric mean) $\sqrt[n]{x_1 x_2 \cdots x_n}$

Gene	Sample A	Sample B
X	26	10
Y	26	10
Z	26	10
DE	2	50

Avg. Sample
16
16
16
10

Normalization: DESeq2 Median of Ratios

1. Take a row-wise average to produce an average sample (geometric mean) $\sqrt[n]{x_1 x_2 \cdots x_n}$

Gene	Sample A	Sample B
X	26	10
Y	26	10
Z	26	10
DE	2	50

Avg. Sample
16
16
16
10

2. Divide all rows by the Average Sample for that gene (**Ratio**)

Gene	Sample A/Avg.	Sample B /Avg.
X	26/16 = 1.6	10/16 = 0.6
Y	1.6	0.6
Z	1.6	0.6
DE	0.2	5

Normalization: DESeq2 Median of Ratios

1. Take a row-wise average to produce an average sample (geometric mean) $\sqrt[n]{x_1 x_2 \cdots x_n}$

Gene	Sample A	Sample B
X	26	10
Y	26	10
Z	26	10
DE	2	50

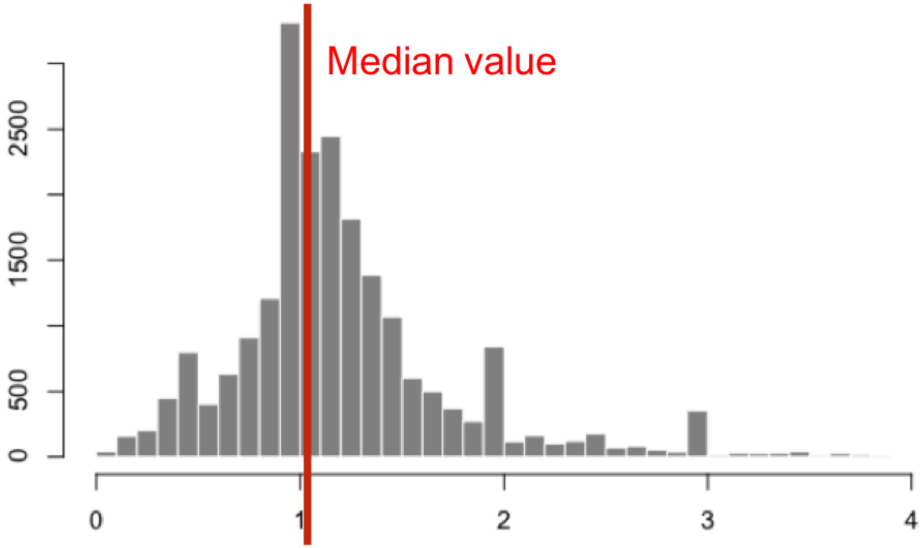
Avg. Sample
16
16
16
16

2. Divide all rows by the Average Sample for that gene (**Ratio**)

Gene	Sample A/Avg.	Sample B /Avg.
X	26/16 = 1.6	10/16 = 0.6
Y	1.6	0.6
Z	1.6	0.6
DE	0.2	5

3. Take the **median** of each column. Should be ~1 for all

Size factor	1.6	0.6



Normalization: DESeq2 Median of Ratios

1. Take a row-wise average to produce an average sample (geometric mean) $\sqrt[n]{x_1 x_2 \cdots x_n}$

Gene	Sample A	Sample B
X	26	10
Y	26	10
Z	26	10
DE	2	50

→

Avg. Sample
16
16
16
16

2. Divide all rows by the Average Sample for that gene (**Ratio**)

Gene	Sample A/Avg.	Sample B /Avg.
X	26/16 = 1.6	10/16 = 0.6
Y	1.6	0.6
Z	1.6	0.6
DE	0.2	5

3. Take the **median** of each column. Should be ~1 for all

Size factor	1.6	0.6

4. Divide all counts by sample specific size factor

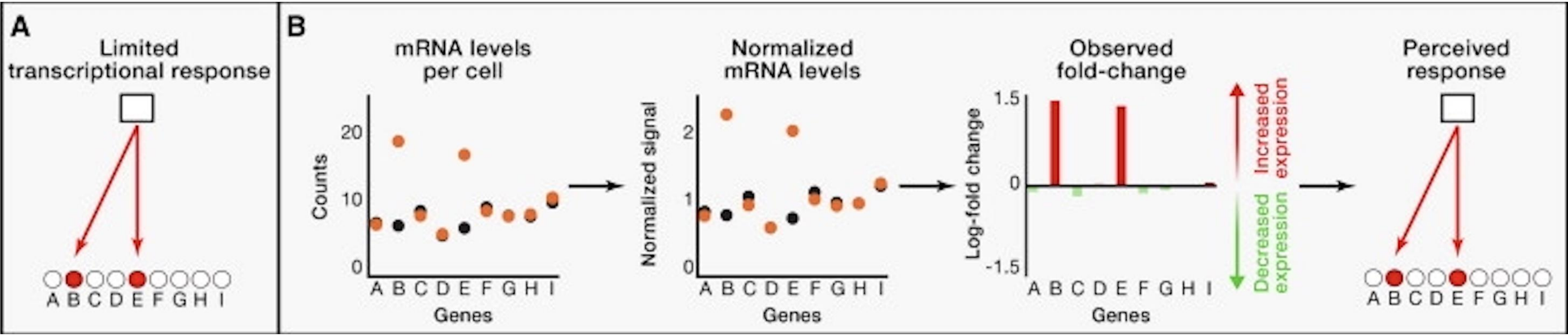
Gene	Sample A / S_A	Sample B / S_B
X	16.3	16.7
Y	16.3	16.7
Z	16.3	16.7
DE	1.3	83.3

Normalized counts for non-DE genes are similar!

```
estimateSizeFactors(dds)
```

Assumption of DESeq2 Median of Ratios

Median of Ratios method assumes that most genes are not Differentially Expressed between samples.

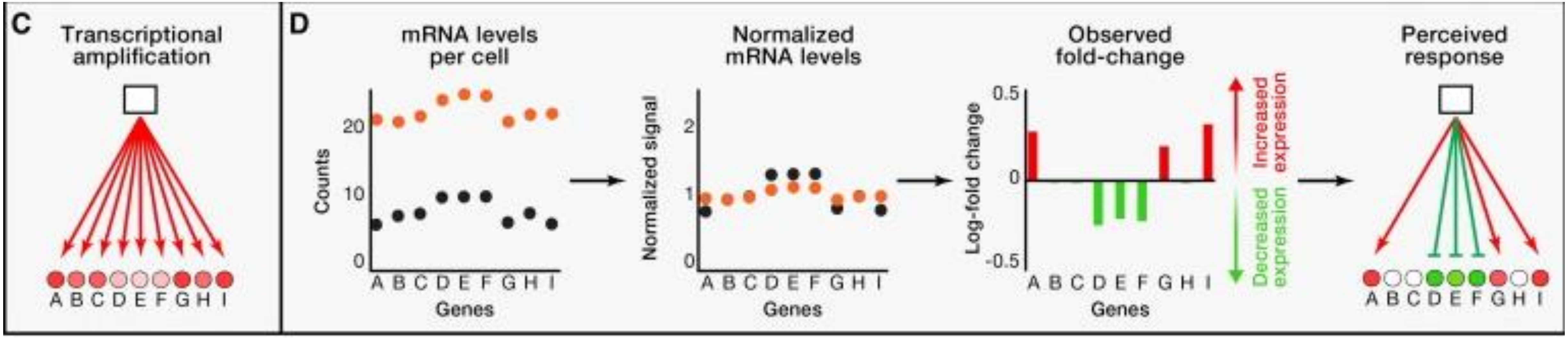


Loven et al “Revisiting Global Gene Expression Analysis” Cell 2012 https://doi.org/10.1016/j.cell.2012.10.012

Assumption of DESeq2 Median of Ratios

Median of Ratios method assumes that most genes are not Differentially Expressed between samples.

<u>COUNTER EXAMPLE</u>



NOTE: add back full picture or remove

- Late stage cell death (total RNA DOWN)
- High c-Myc cells (total RNA UP)

Known quantity spike-in transcripts (ERCC) can be used to normalize in these cases.

Loven et al “Revisiting Global Gene Expression Analysis” Cell 2012 https://doi.org/10.1016/j.cell.2012.10.012

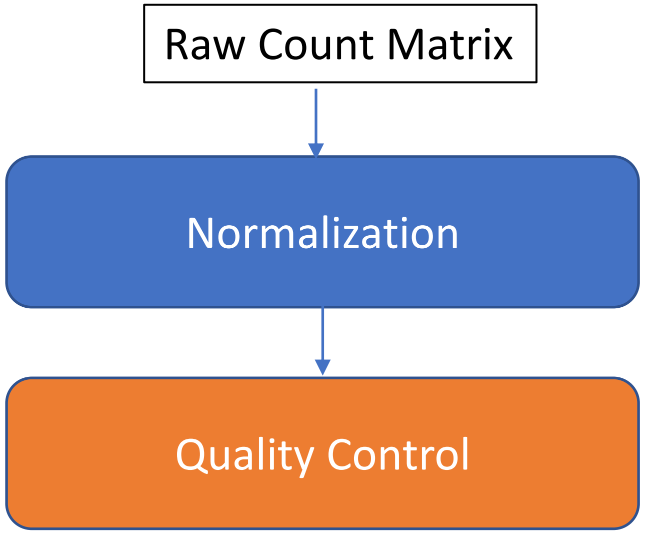
Normalization methods

Normalization method	Description	Accounted factors	Recommended use
CPM (counts per million)	$\frac{\mathrm{K_i}}{Total\ Reads\ per\ Sample/10^6}$	sequencing depth	Comparison between replicates of the sample group
R/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	$\frac{\mathrm{K_i}}{Gene\ Length/10^3 *\ Total\ Reads\ per\ Sample/10^6}$	sequencing depth and gene length	Comparison between genes in a sample
DESeq2's **median of ratios** [1]	$\mathrm{K_i}$ divided by sample-specific size factors	sequencing depth and RNA composition	**Differential Expression** between samples

Similar to DESeq2: EdgeR, limma-voom

Adapted from https://hbctraining.github.io/DGE_workshop

Quality Control Visualizations



Raw Count Matrix

Normalization

Quality Control

Examine sources of variation in the data

- Principal Component Analysis
- Hierarchical Clustering

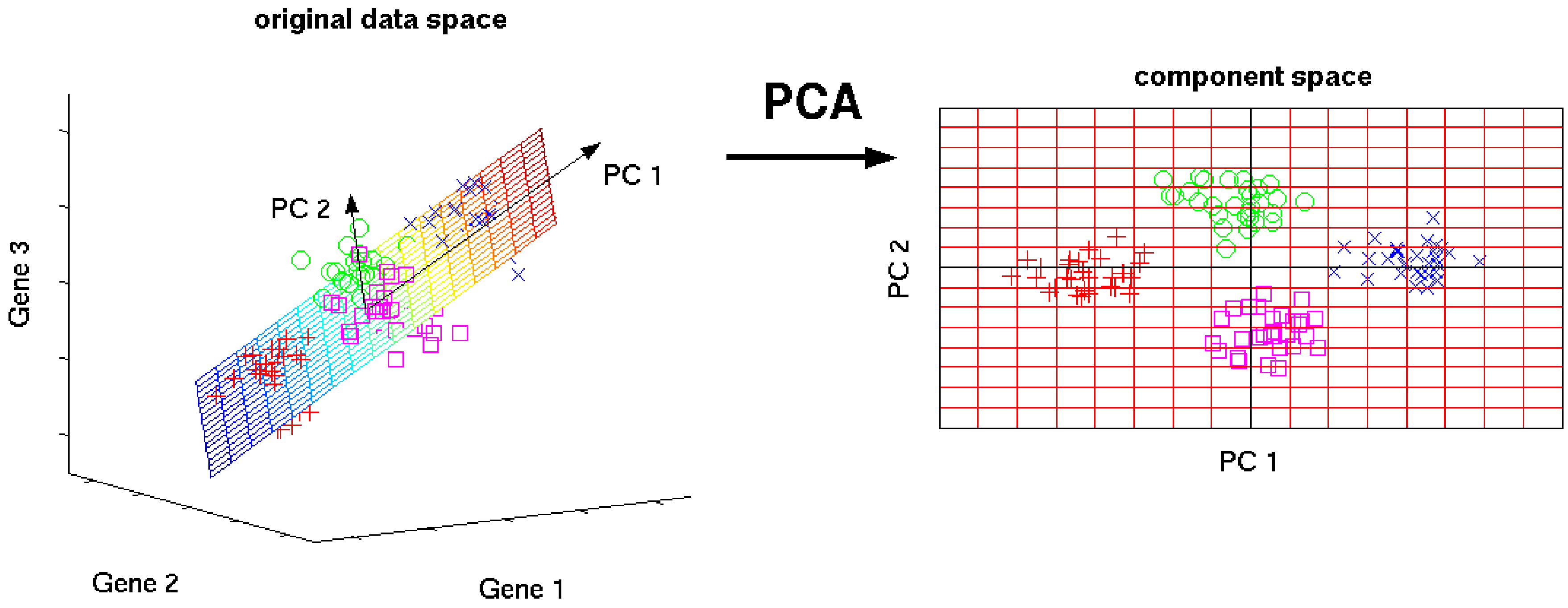
(Log2 + 1) Transformed, Normalized Count Table

Gene	Sample A	Sample B	Sample C
1	1	1.6	0.5
2	2.2	-0.2	1
3	-1	1	3.1

Principle Component Analysis

Dimension reduction technique
Example: 3 gene dimensions -> 2 PC

Gene	Mock_12h	Mock_12h	Mock_24h	Mock_24h	HIV_12h	HIV_12h	HIV_24h	HIV_24h
Gene 1	8.9	8.9	8.9	9.0	8.9	8.9	9.0	6.8
Gene 2	0.6	-1.0	0.6	-1.0	0.6	-1.0	0.6	3.8
Gene 3	4.1	11.9	4.1	-0.5	4.1	8.7	4.0	4.4



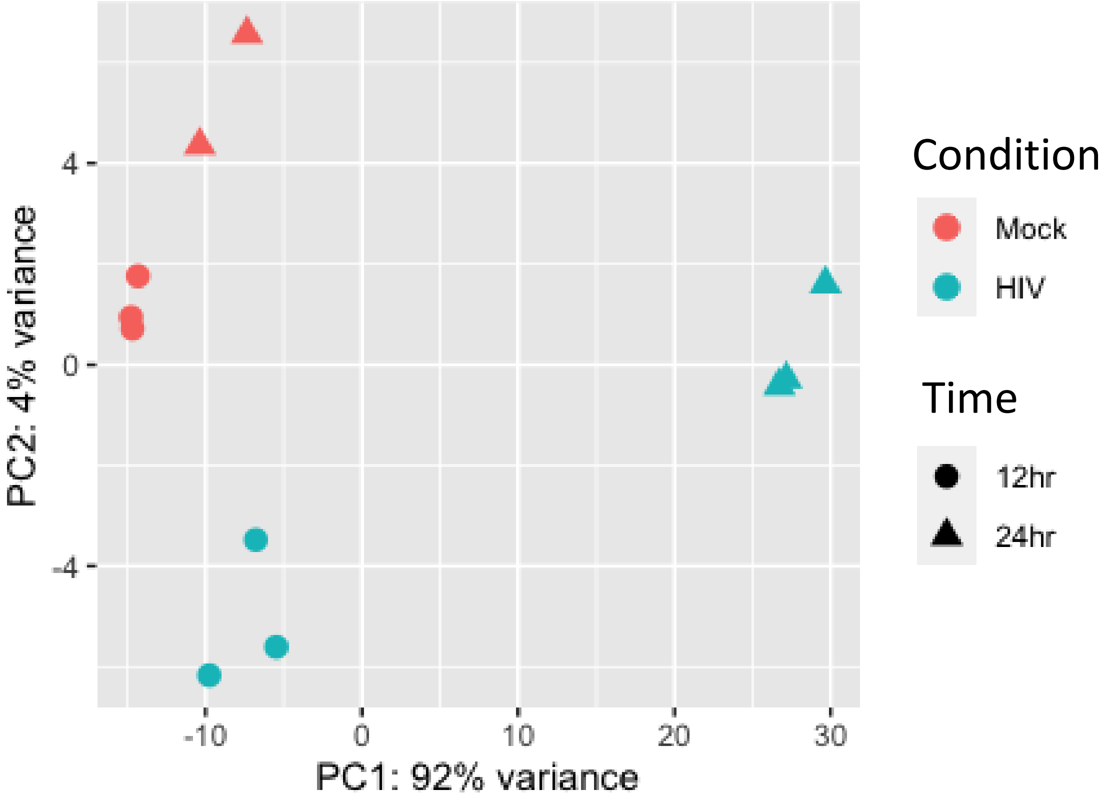
original data space

component space

Do your samples cluster as expected?

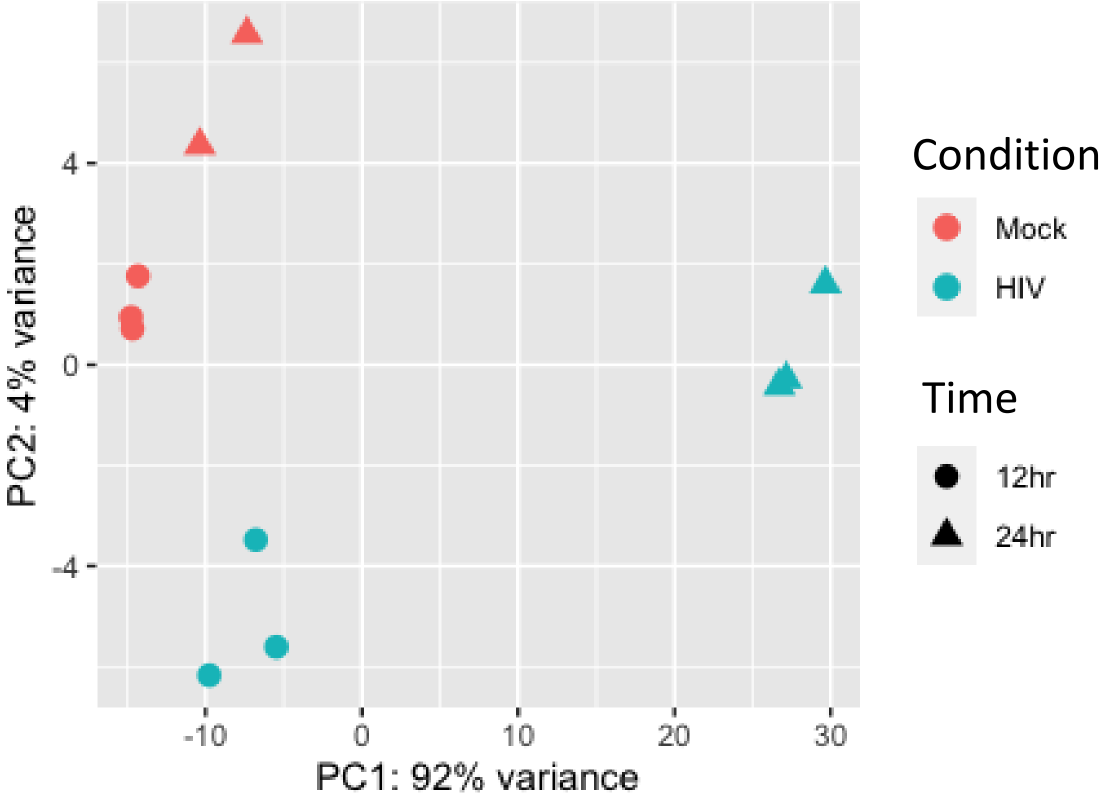
What are the major sources of variation in the data?

Principle Component Analysis



- ✓ Do your samples cluster as expected?
- ✓ What are the major sources of variation in the data?

Principle Component Analysis



- ✓ Do your samples cluster as expected?
- ✓ What are the major sources of variation in the data?
- ✓ Is there a batch effect?

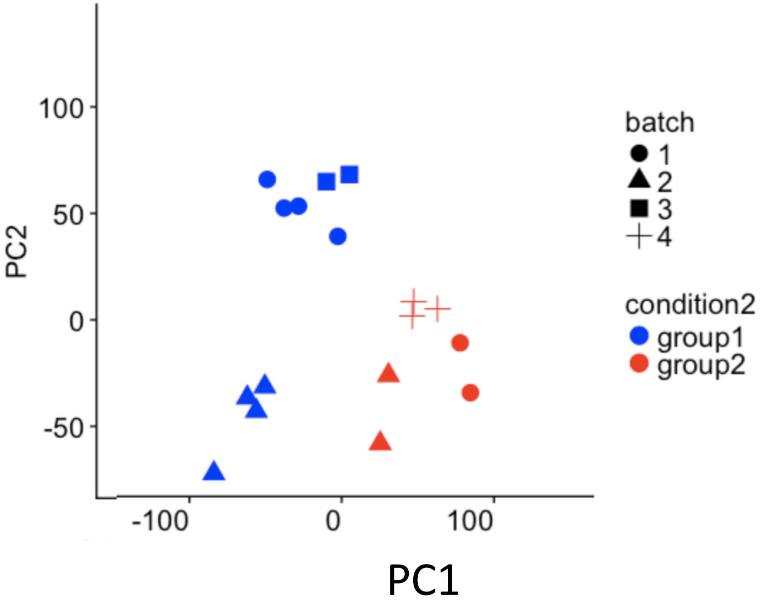
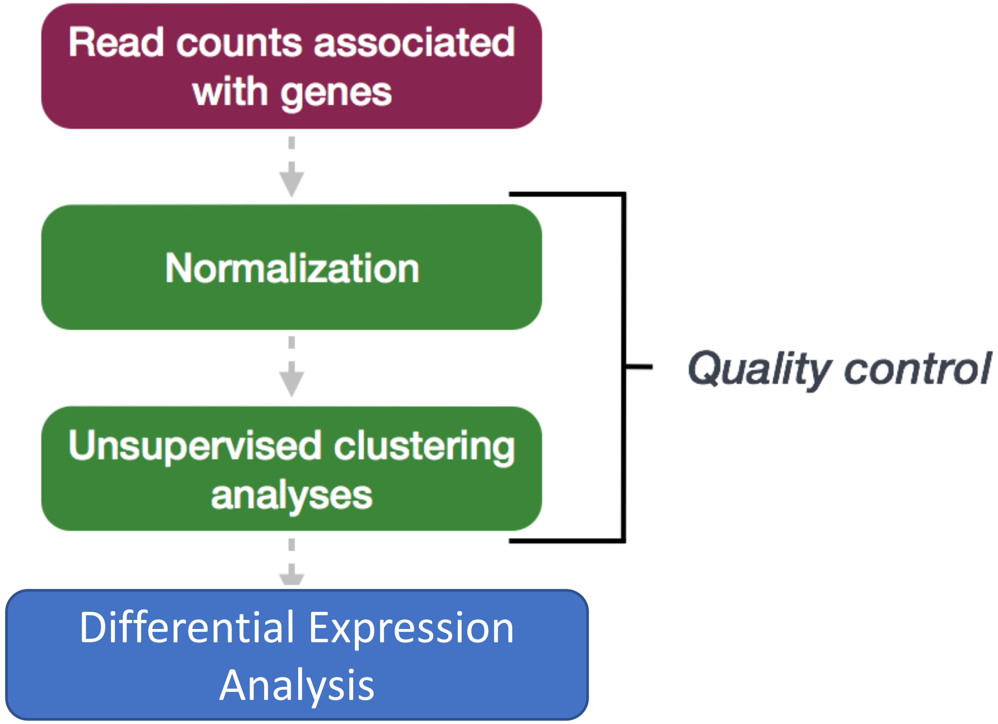


Image https://support.bioconductor.org/p/111491/

Differential Expression with DESeq2



Read counts associated with genes

Normalization

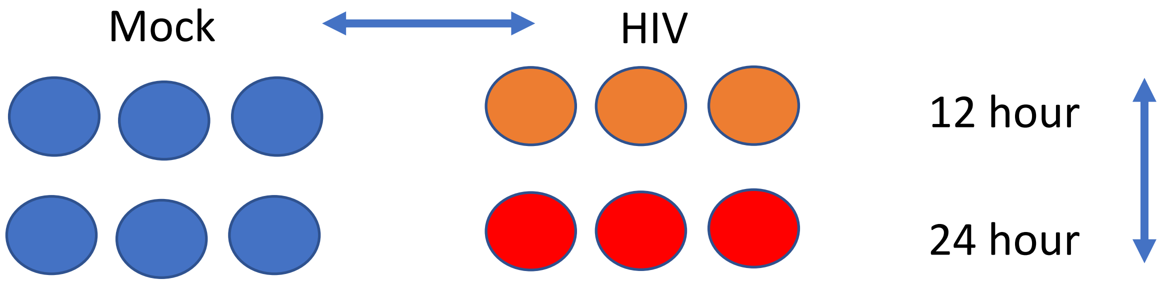
Unsupervised clustering analyses

Quality control

Differential Expression Analysis

https://hbctraining.github.io/DGE_workshop

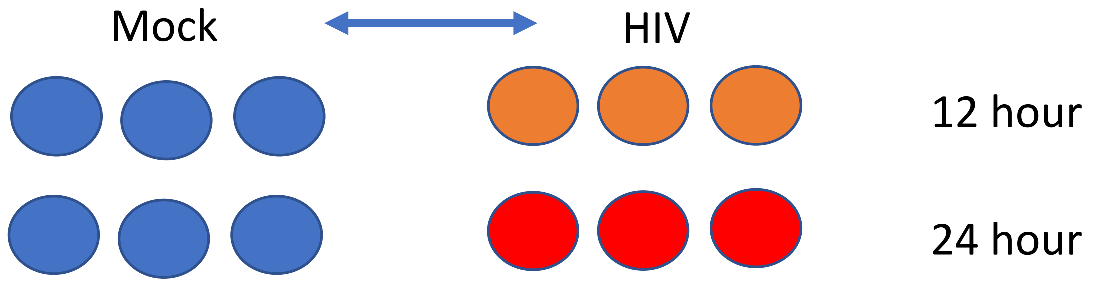
Multi-factor experiment design



Factor 1:
Infection status (Mock or HIV)

Factor 2:
Time (12 or 24 hr)

Multi-factor experiment design

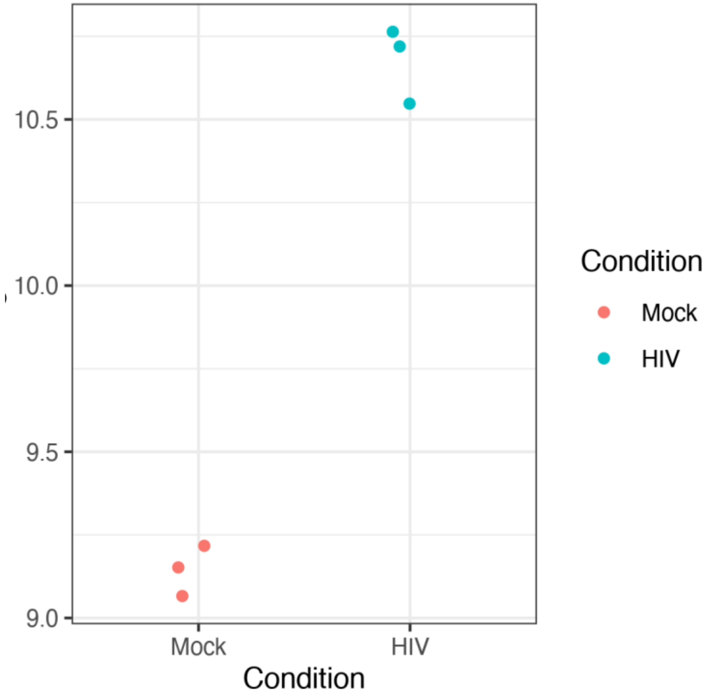


- Differential Expression compares two conditions
- We’ll choose Infection status at 12 hr (Mock or HIV) for comparison
- We could also choose time, or a combination of multiple factors

Step 1: Modeling gene expression values

All leading DE tools use **regression models** to estimate the fold change between conditions for **each gene**

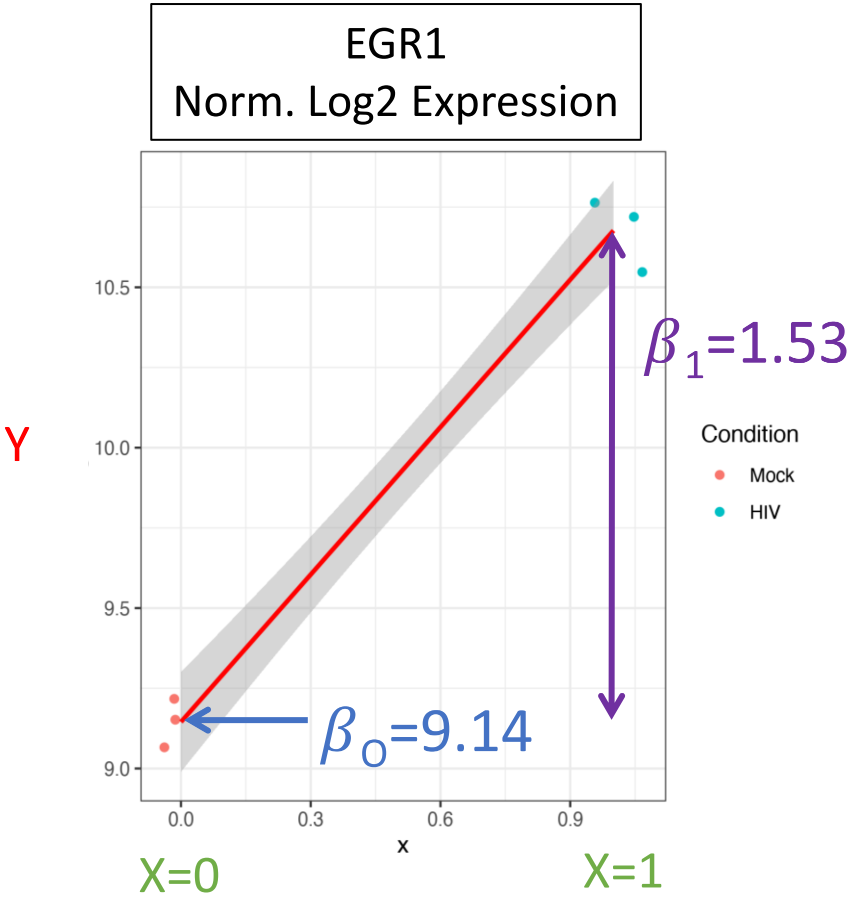
EGR1
Norm. Log2 Expression



(12hr)

Step 1: Modeling gene expression values

All leading DE tools use **regression models** to estimate the fold change between conditions for **each gene**
Example, simple linear regression:



Intercept Condition (0-Mock, 1-HIV)

$$Y = \beta_o + \beta_1 X + e$$

Log2 Expression Values

Slope: difference between Mock /HIV

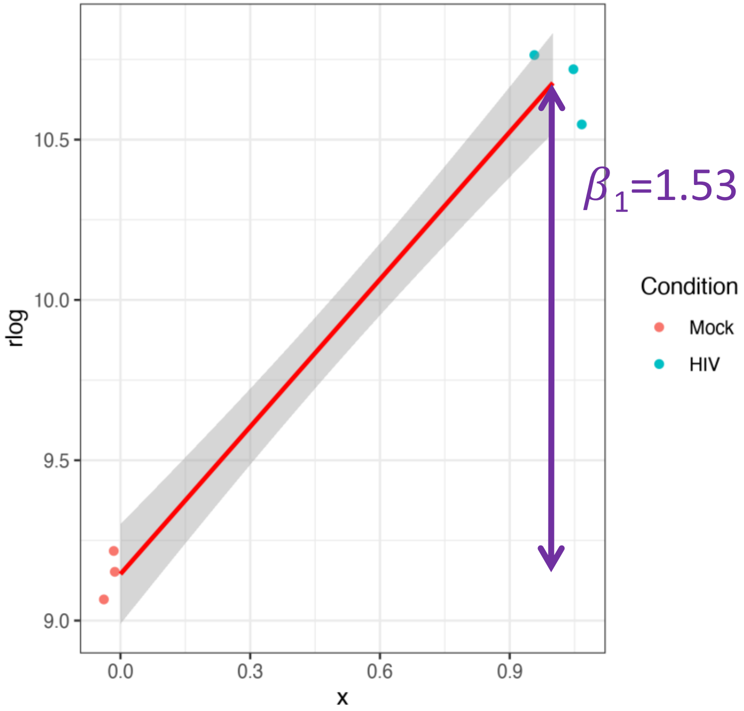
Error

DESeq2 uses a Generalized Linear Model with a Negative Binomial error Distribution, which has been shown to be best fit for RNAseq data.

Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 (2014).

Step 2: Hypothesis Testing

EGR1
Norm. Log2 Expression



Is EGR1 differentially expressed?

Yes! $p << 0.05$

$$H_O : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0$$

H_o: there is no systematic difference between the average read count values for Mock vs. HIV

- Statistical test – Wald test (similar to t-test) on β_1
- $Z = \beta_1 / SE_{\beta 1}$
- Z-statistic is compared to the normal distribution and probability of getting a statistic at least as extreme is computed

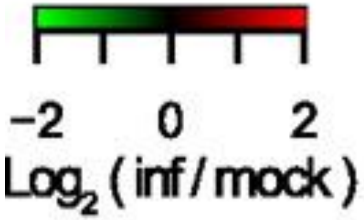
DESeq2 Results table

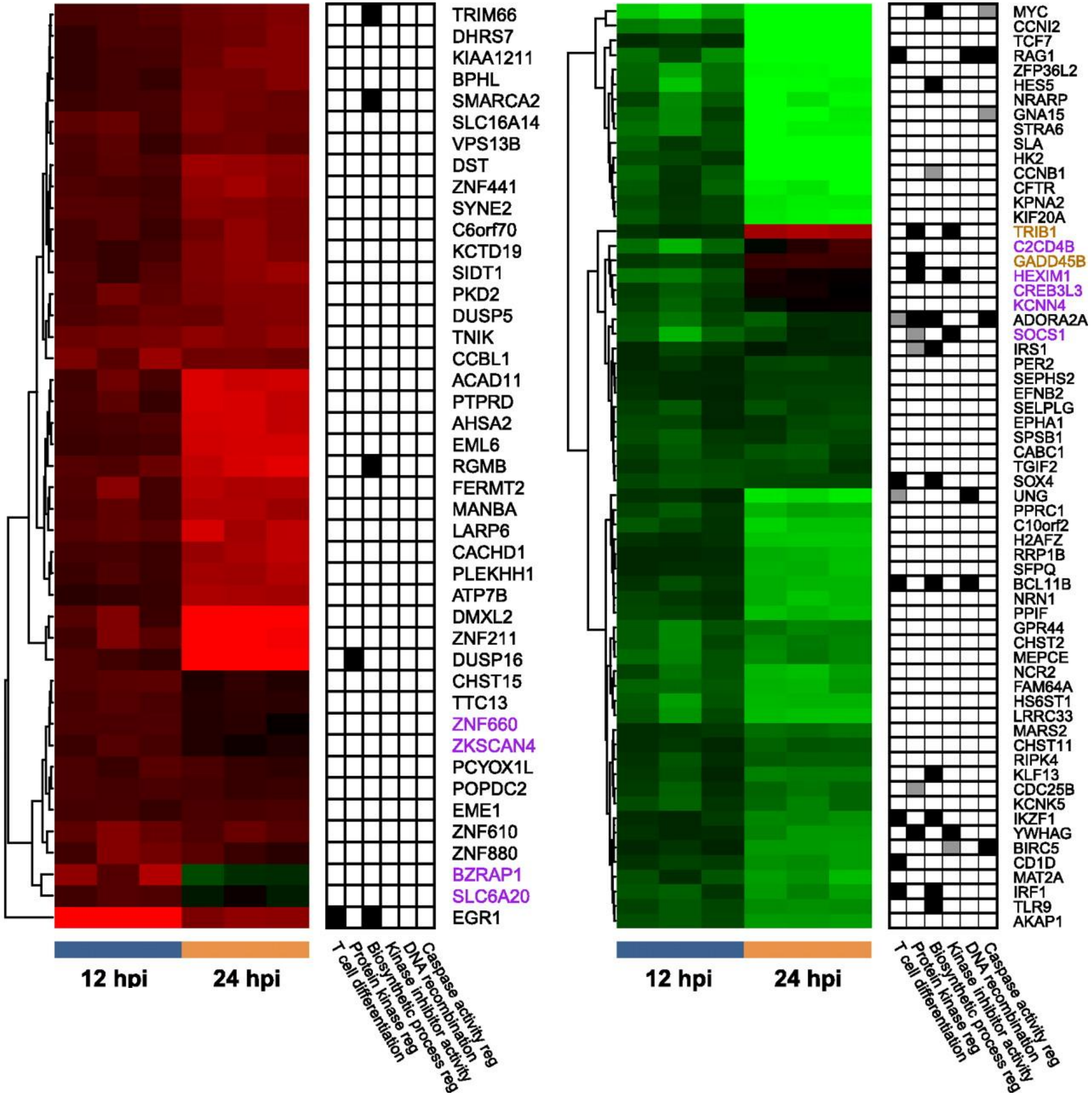
GeneID	Base mean	log2FoldChange	StdErr	P-value	P-adj
EGR1	1273	1.55	0.13	1.19e-77	1.52e-73
MYC	5226	-1.53	0.14	1.63e-36	1.03e-32

- Mean of normalized counts – averaged over all samples from two conditions (HIV, Mock)
- Log of the fold change between two conditions
- StdErr – Standard error of coefficient (e.g. b_1)
- P-value – the probability that the Wald statistic is as extreme as observed if H_O were true
- P-adj – accounting for multiple testing correction

Study findings

- T cell differentiation-related genes were overrepresented in the DEG at 24hr
- ‘Large-scale disruptions to host transcription’ at 24hr





References

DESeq2 vignette (R/Rstudio):
http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#differential-expression-analysis

HBC Training (Command line/R):
https://hbctraining.github.io/DGE_workshop

Galaxy Training:
https://galaxyproject.org/tutorials/rb_rnaseq/

Next: Introduction To Galaxy