

RNA-seq to study HIV Infection in cells

Rebecca Batorsky
Pr Bioinformatics Scientist
Dec 2021

Research Technology Team



Delilah Maloney
High Performance Computing Specialist



Kyle Monahan
Senior Data Science Specialist



Shawn Doughty
Manager, Research Computing



Rebecca Batorsky
Senior 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

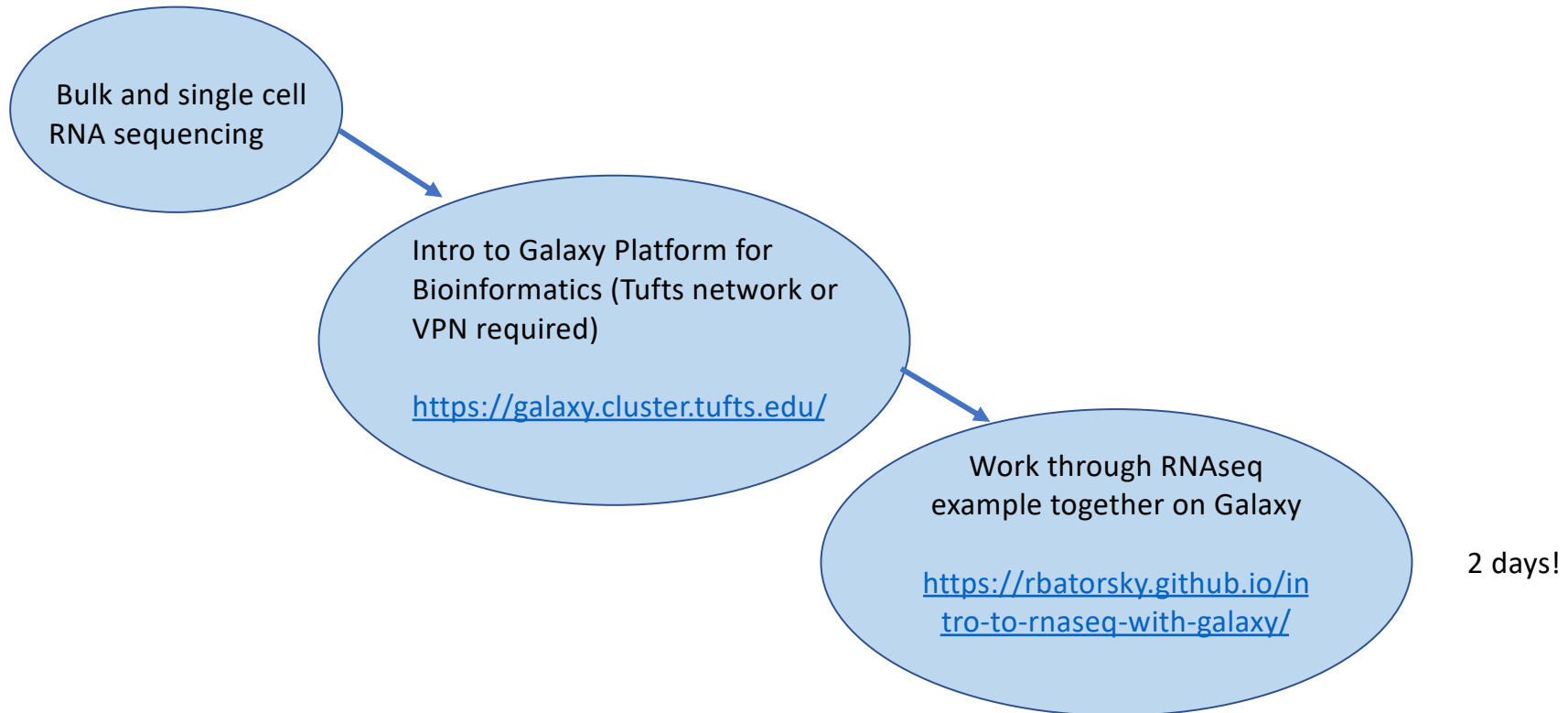


Uku-Kaspar Uustalu
Data Science Specialist

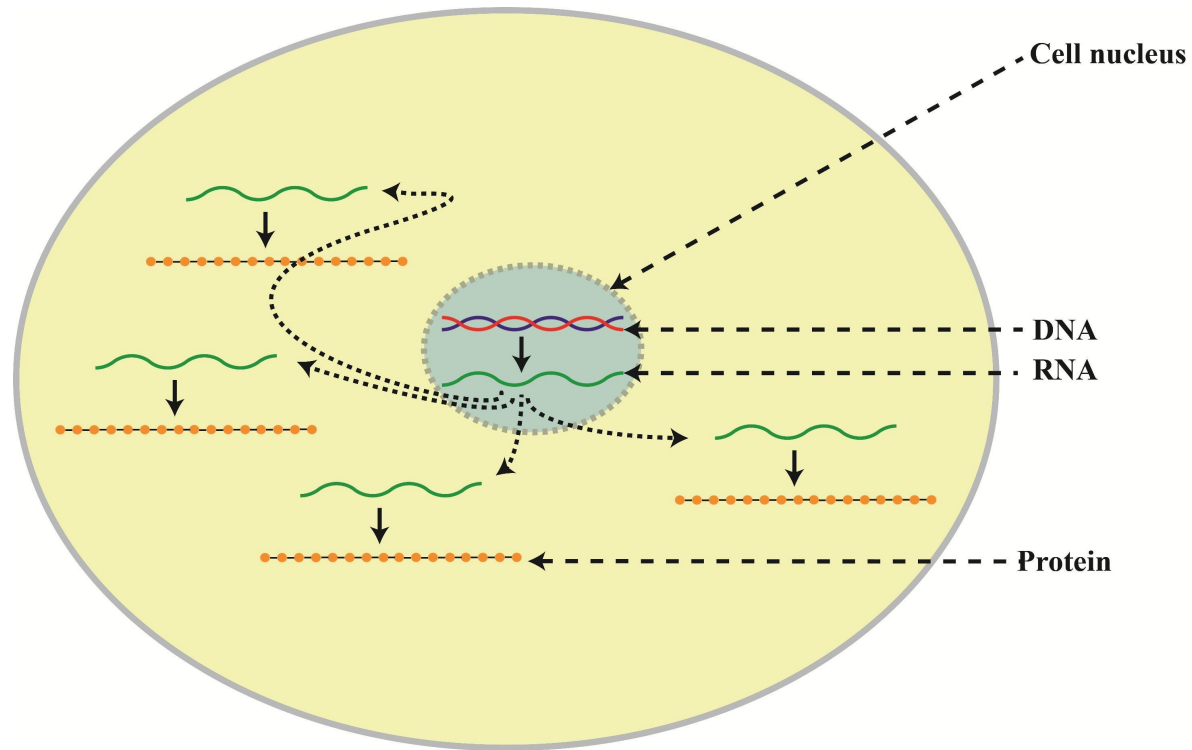
- ✓ Consultation on Projects and Grants
- ✓ High Performance Compute Cluster
- ✓ Workshops

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

Outline



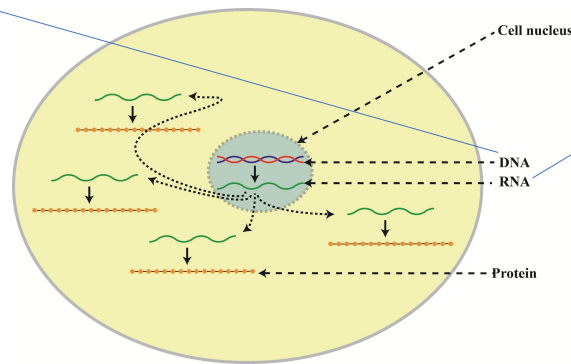
DNA and RNA in a cell



Two common analyses

DNA Sequencing

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



RNA Sequencing

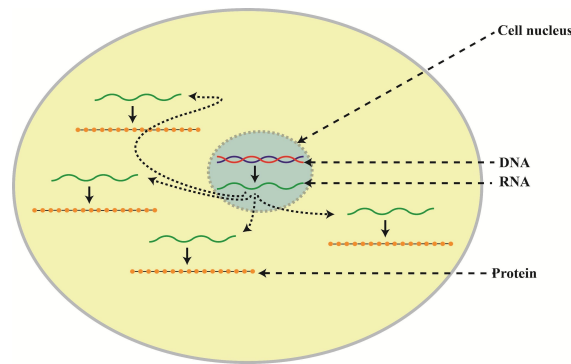
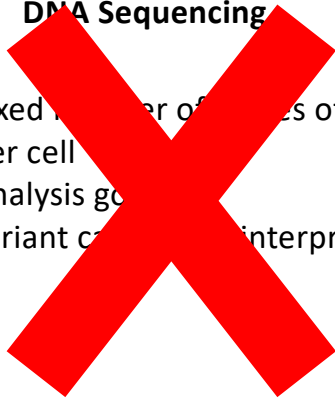
- Number of copies of a gene transcript per cell depends on gene expression
- Analysis goal:
 - Bulk : Differential expression
 - Single cell : Quantify different cell populations

<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

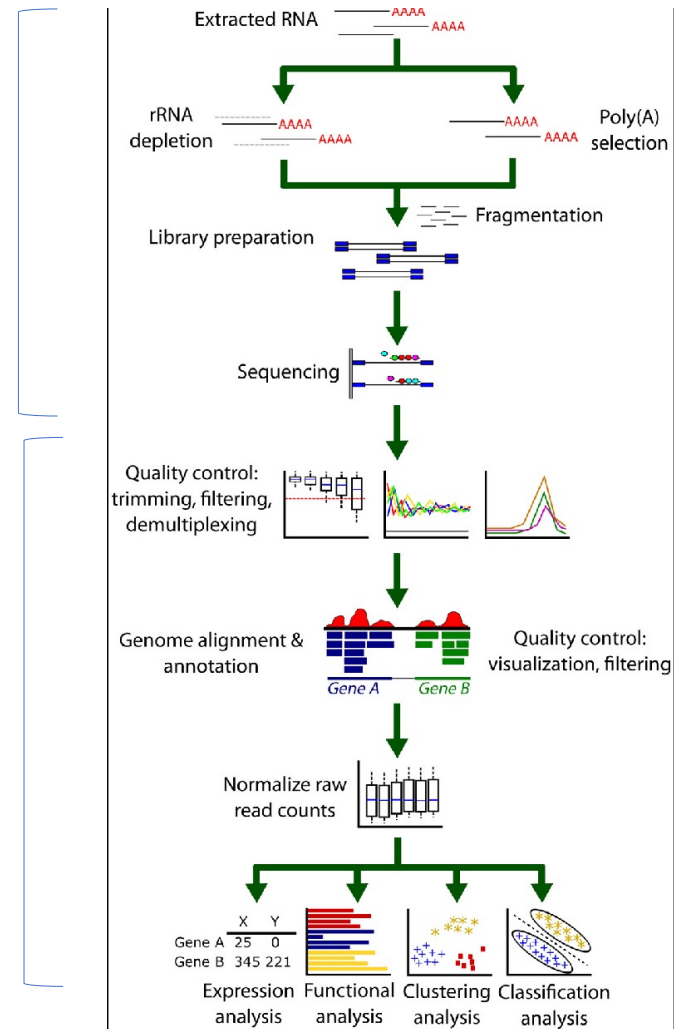
<https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg>

“Bulk” RNA seq workflow

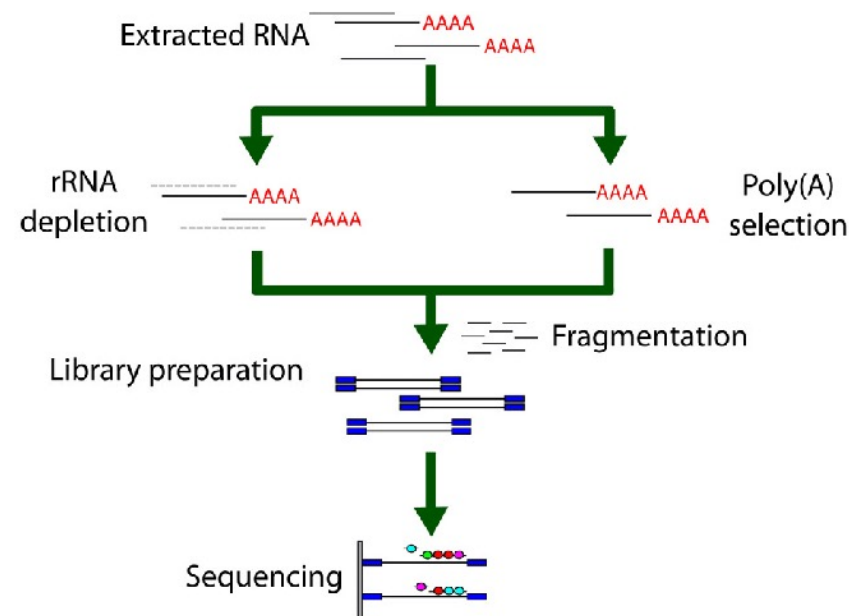
Library prep and sequencing

Bioinformatics

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



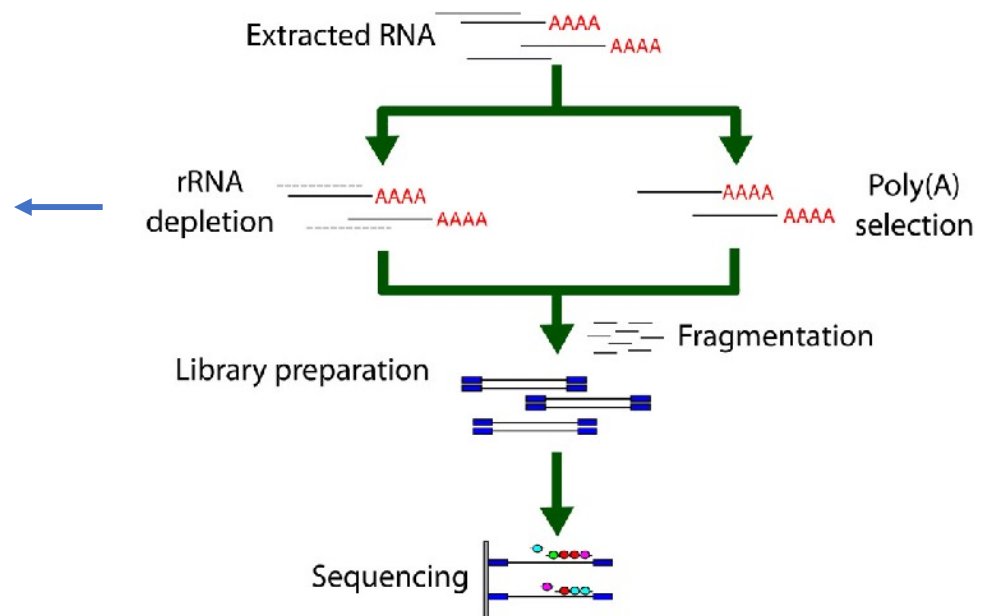
RNA seq library prep and sequencing



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

RNA seq library prep and sequencing

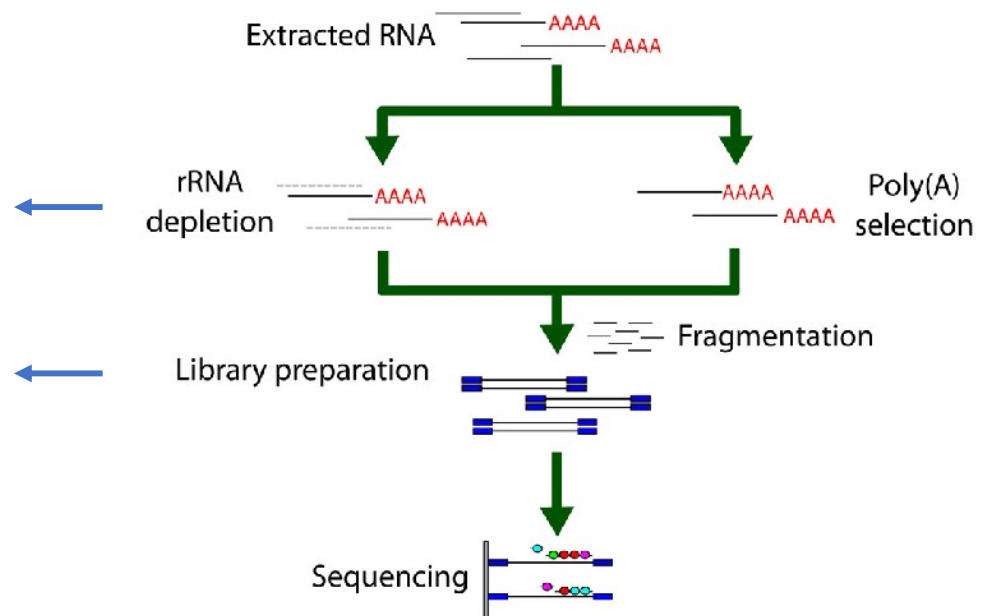
- Enrichment for mRNA, two options
- In humans, ~95%–98% of all RNA molecules are rRNAs



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:

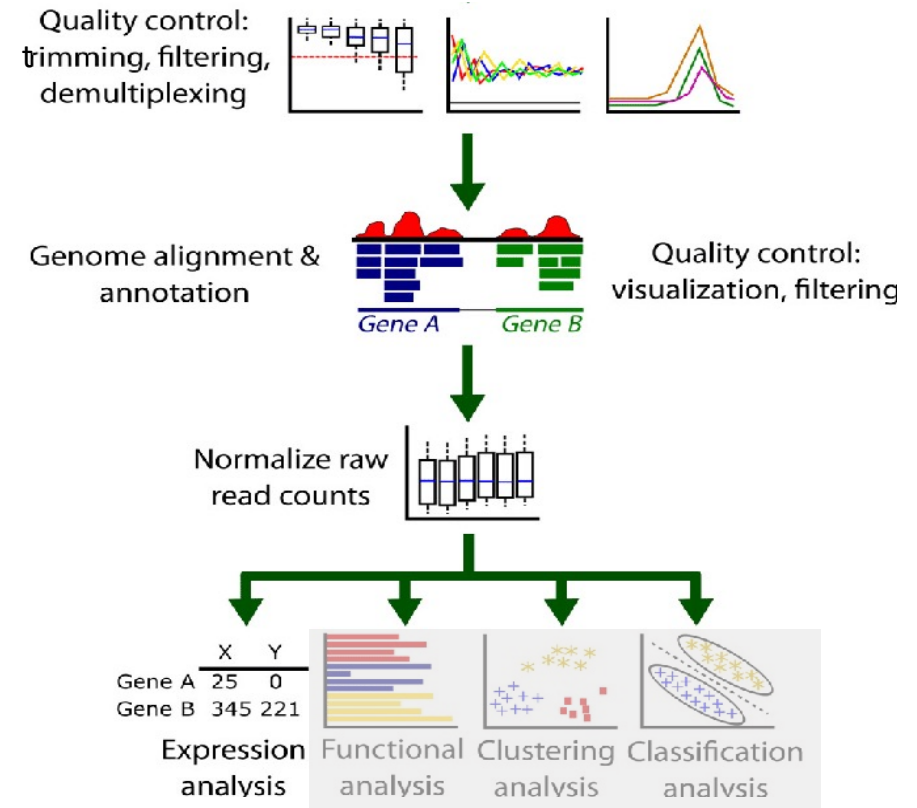
[Illumina Sequencing by Synthesis](#)

[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”

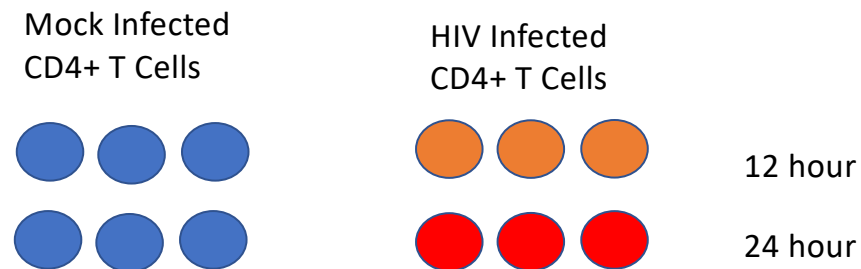


Oshlack et al. 2010. From RNA-seq reads to differential expression results. Genome Biology 2010, 11:220

Our dataset

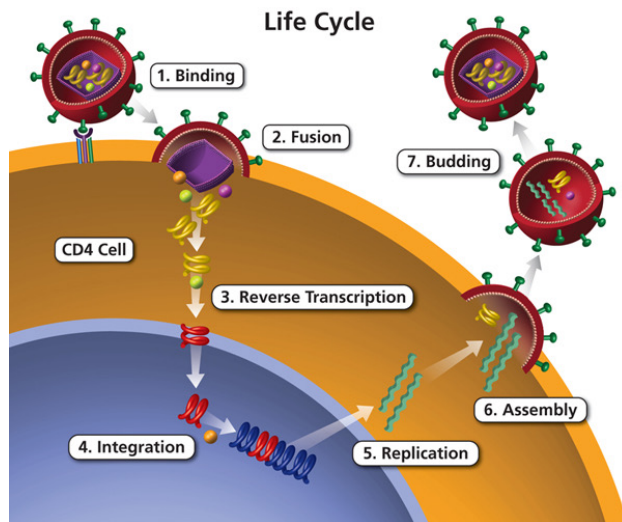
Next-Generation Sequencing Reveals HIV-1-Mediated Suppression of T Cell Activation and RNA Processing and Regulation of Noncoding RNA Expression in a CD4⁺ T Cell Line

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



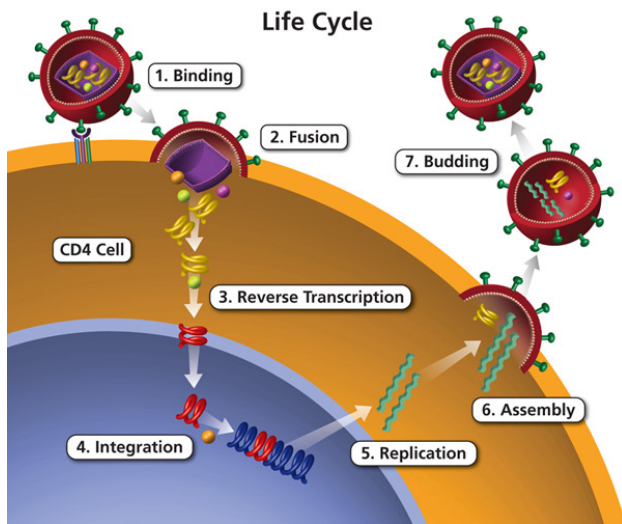
<https://www.ncbi.nlm.nih.gov/pubmed/21933919>

HIV lifecycle

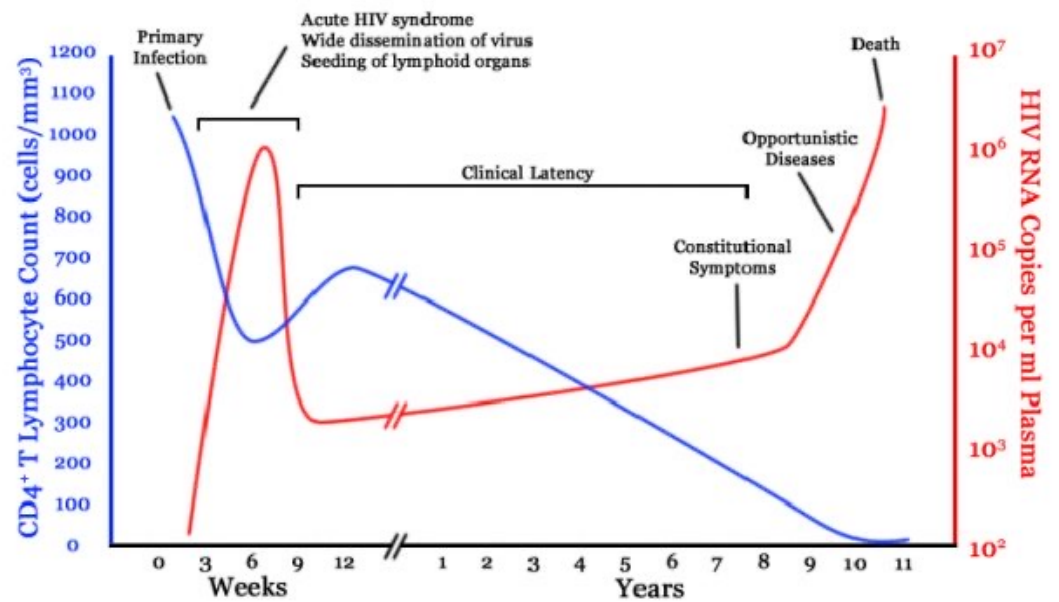


<https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle>

HIV lifecycle



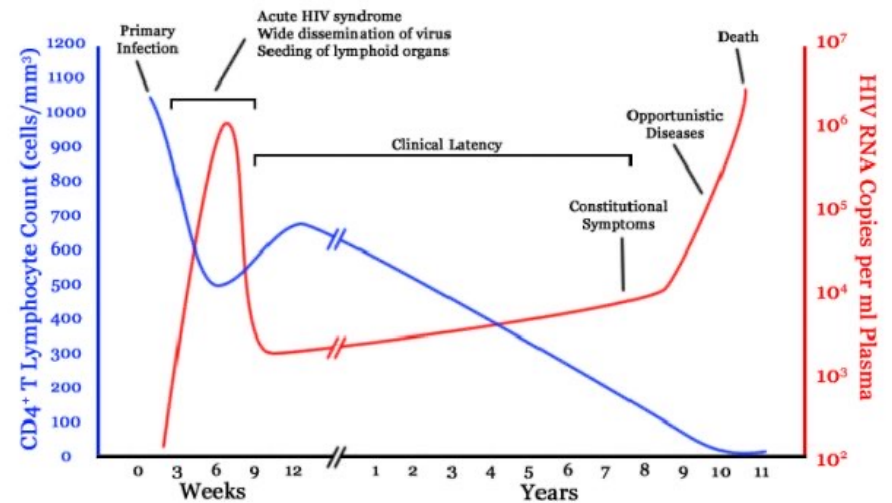
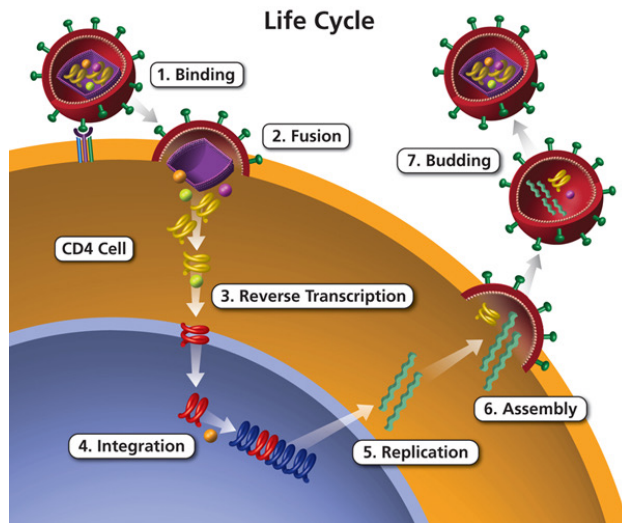
HIV infection in a human host



<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?

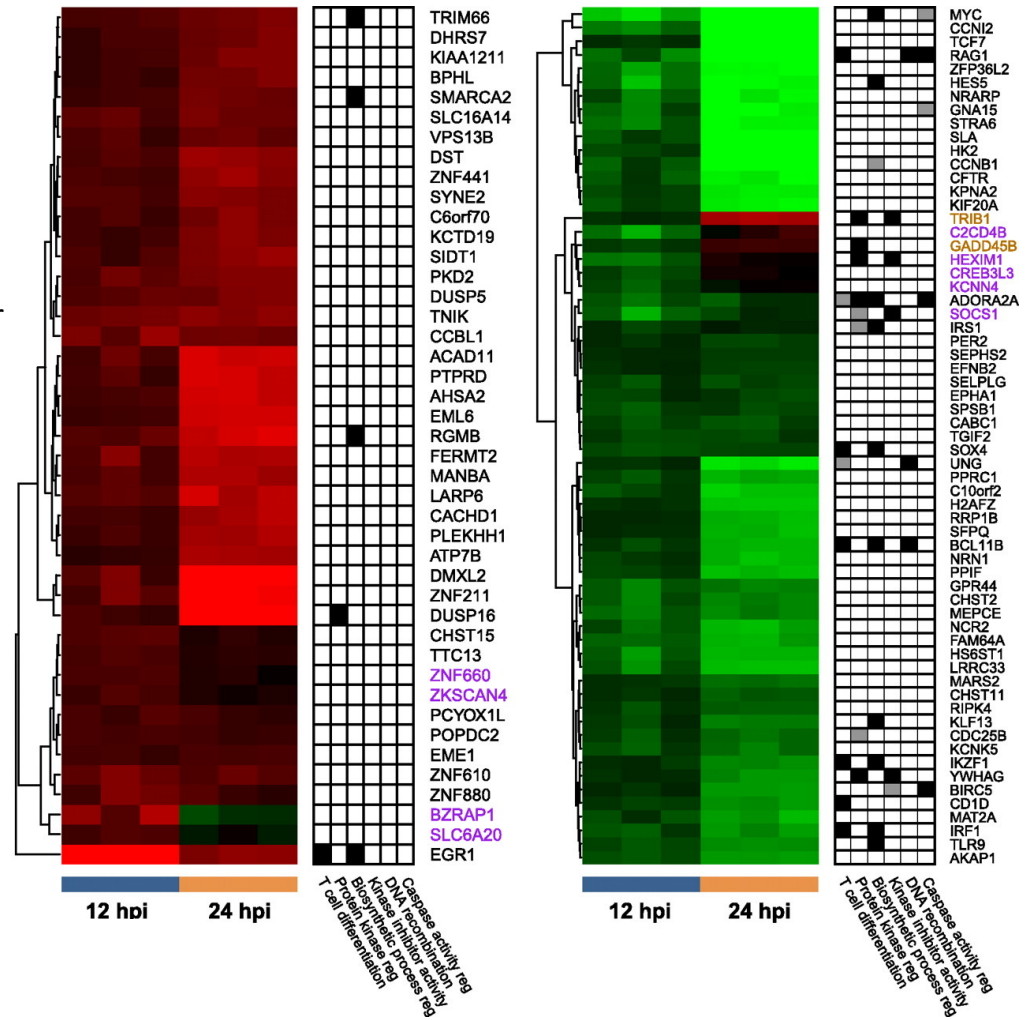
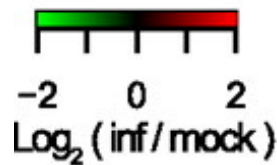


<https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle>

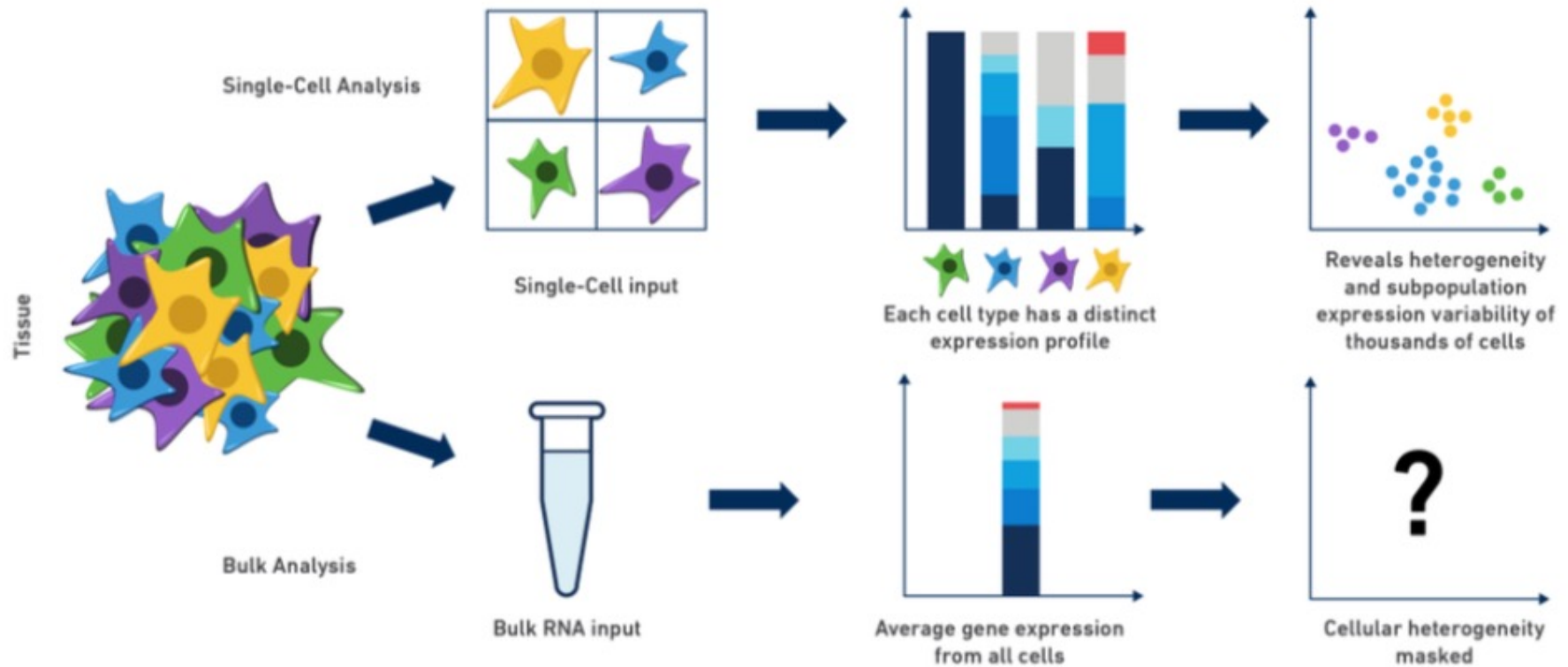
Study findings

Using RNAseq, authors demonstrate:

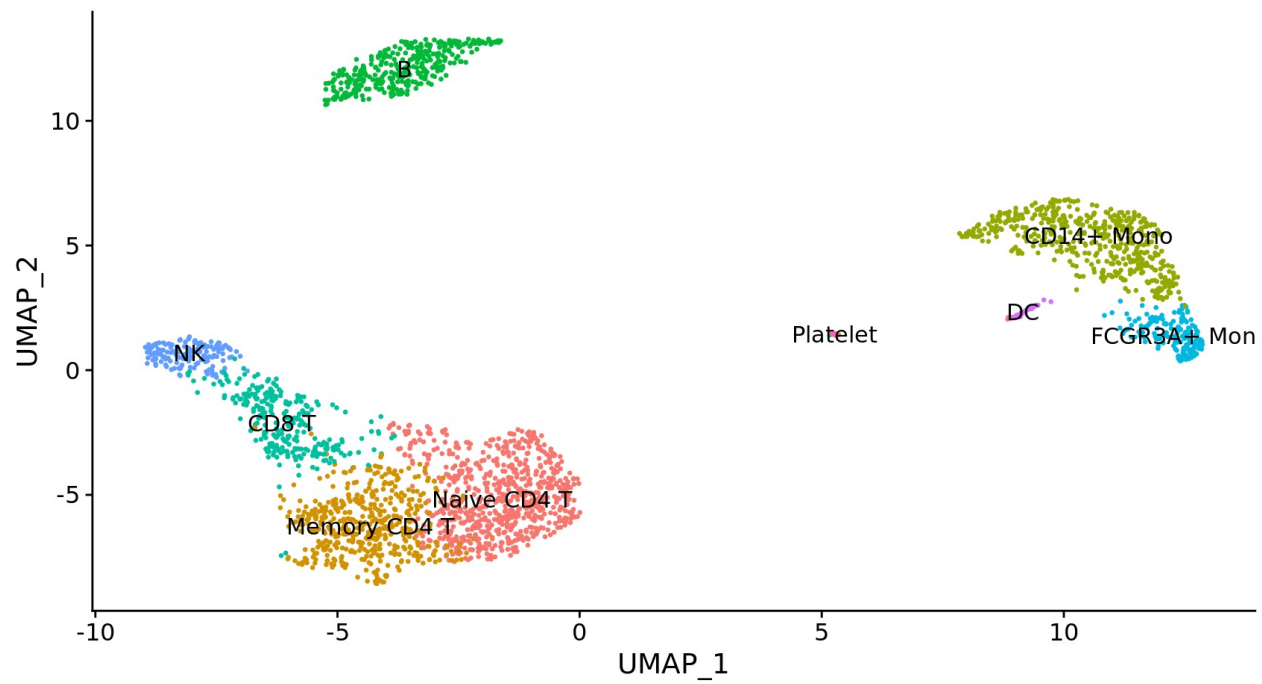
- 20% of reads mapped to HIV at 12 hr, 40% at 24hr
- Downregulation of T cell differentiation genes at 12hr
- 'Large-scale disruptions to host transcription' at 24hr



Bulk vs Single Cell RNA Sequencing

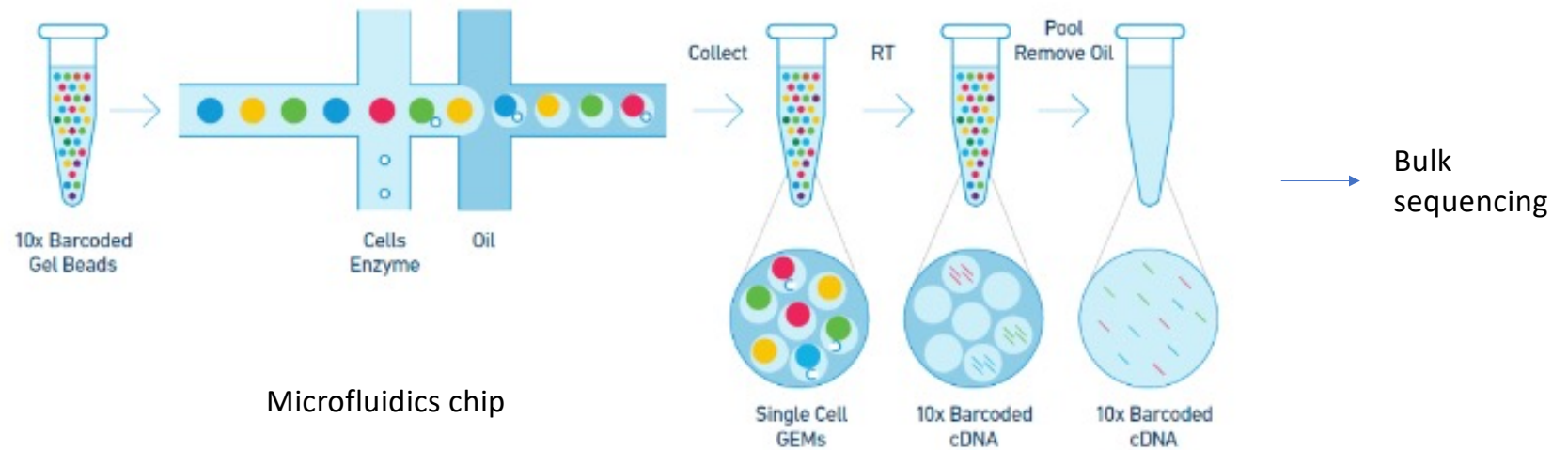


scRNA cell subsets in PBMC



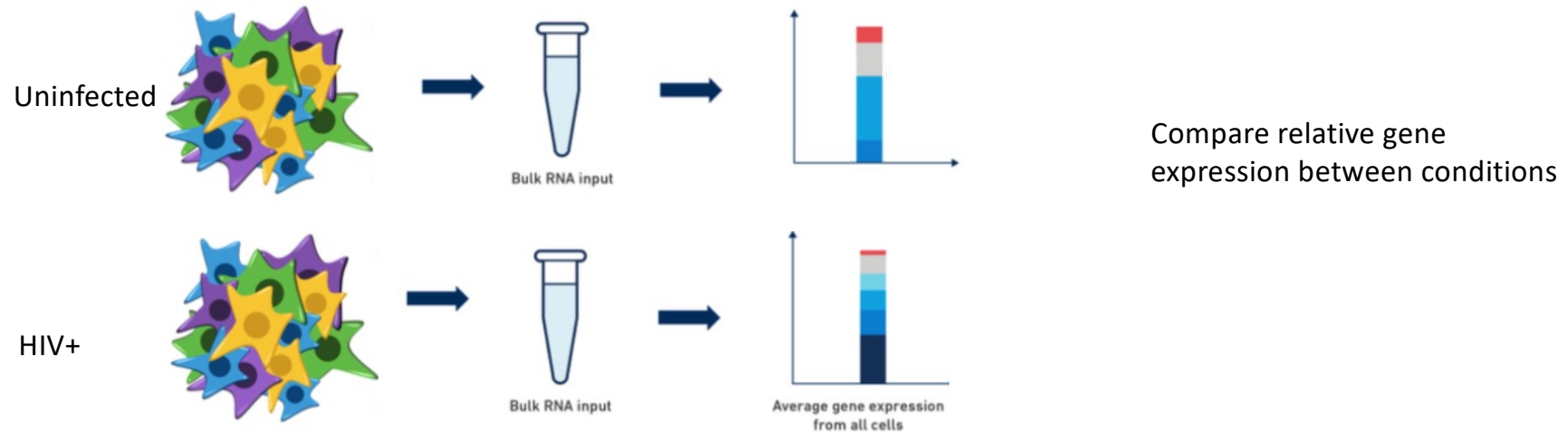
https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html

10x single cell technology

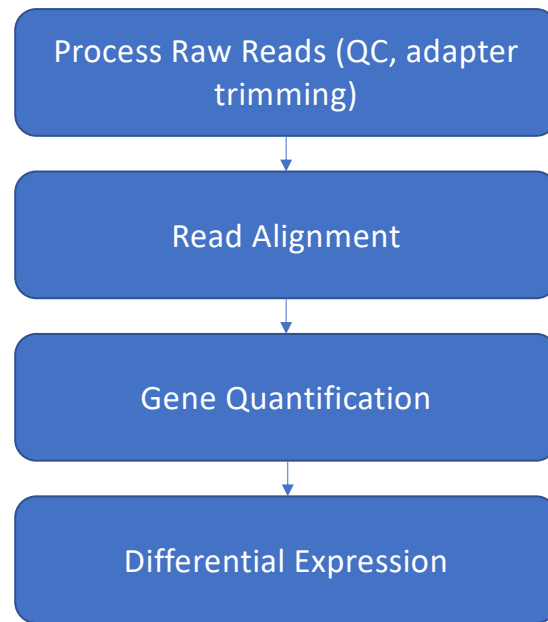


<https://github.com/hbctraining/scRNA-seq>

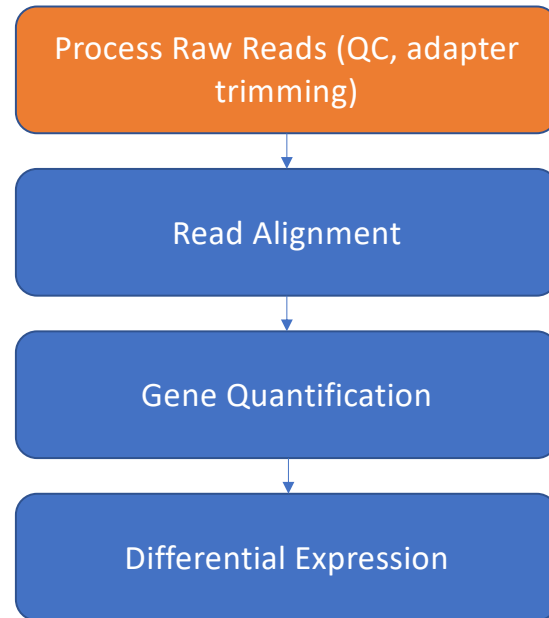
Bulk RNAseq for Differential Expression is OK!



Our (bulk) RNAseq Workflow



Quality control on Raw Reads



Raw reads in Fastq format

```
@SRR098401.109756285  
GACTCACGTAAC TTAAACTCTAACAGAAATATACTA...  
+  
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
GACTCACGTAACCTTAAACTCTAACAGAAATATACTA...
+
CAEFGDG?BCGGGEEDGGHGHGDFHEIEGGDDDD...
```

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

Raw read quality control

Fastq File

```
@SRR497699.30343179.1 HWI-EAS39X_10175_FC61MK0_4_117_4812_10346 length=75
CAGATGGCCGCAGAGGAAGCCATGAAGGCCCTGCATGGGGAGATCGGAAGAGCGTTTCAGCAGGAATGCCGAGAC
+
IIIIIGIIHFIIIBIIDII>IIDHIIHDIIIGIFIIEIGIBDDEFIG<EIEGEEG;<DB@A8CC7<><C@BBDDDB
@SRR497699.11626500.1 HWI-EAS39X_10175_FC61MK0_4_44_8384_16550 length=75
CGTACTGAACGTACAACGCTGATGCCATCCGCATATTTAAATTCGGCAGCGTTAATTAACCTCCCTGACCTCGGCG
+
HHHHHHHHHHHFFHHGHHHHHHB@HHHHHHHHFHHHHEHHHHHHHHHHHGEHDHHEHHHHBHHHGHHHHHHHHG
@SRR497699.29057557.1 HWI-EAS39X_10175_FC61MK0_4_112_12508_19308 length=75
CCGAGGCTTAGCTTTCATTATCACTGTCTCCAGGGTGTGCTTGTCAAAGAGATAAGATCGGAAGAGCGGTTTCAG
+
GGGBGGDGBHHDHHEGEGGHHHHHHGHGHHHHHHGBGGDGGEGDHHHHHHHHHHH@BHHGGHGHHHHHEEGHH
@SRR497699.1331889.1 HWI-EAS39X_10175_FC61MK0_4_5_4738_15920 length=75
CTTACTTTGTAGCCTTCATCAGGGTTTGTGAAGATGGCGGTATATAGGCTGAGCAAGAGGTTGAGGTTGATC
+
HHHHHHHHHGGGGHHGHGEBEEGGEDGGGGGHHHHHGGEGBDGGDDGBGGC<EADBEBE<GGGBEEDGD
```

...

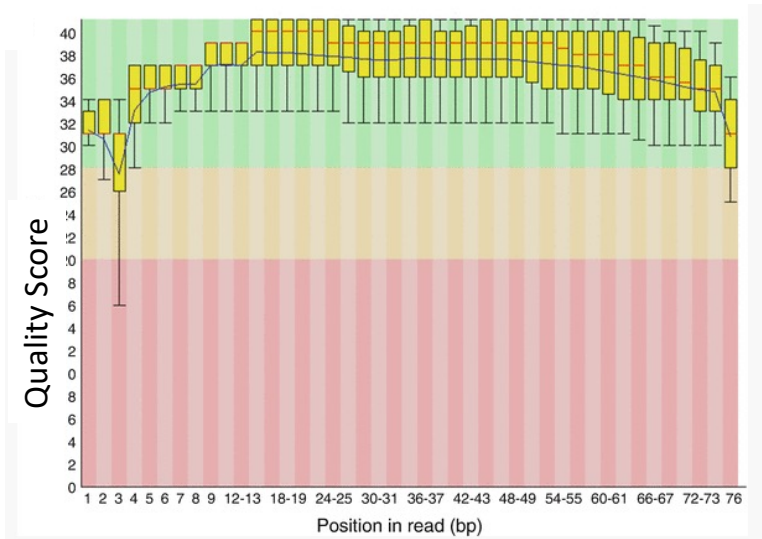
FastQC Tool



Metrics

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

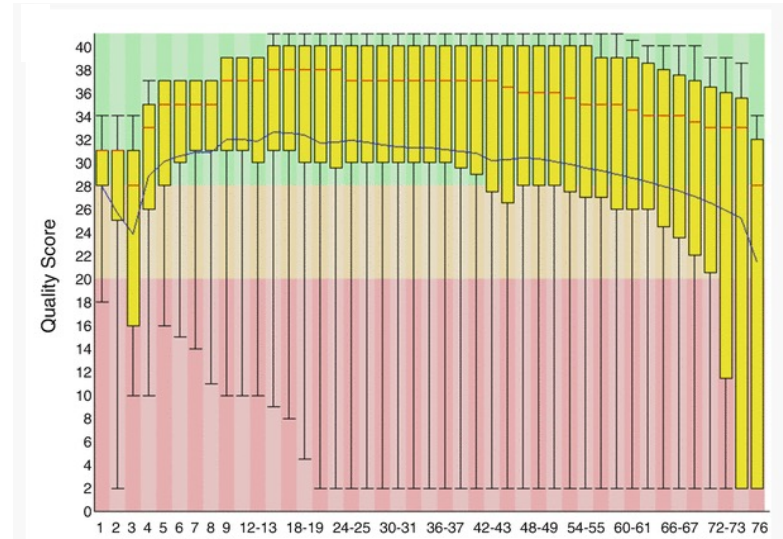
FastQC: Sequence Quality Histogram



Position in read (bp)

GOOD

High quality over the length of the read



Position in read (bp)

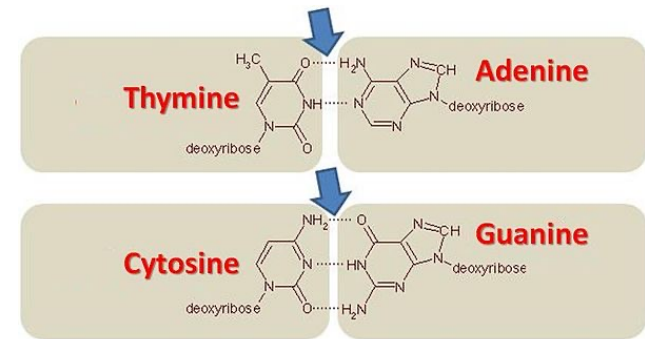
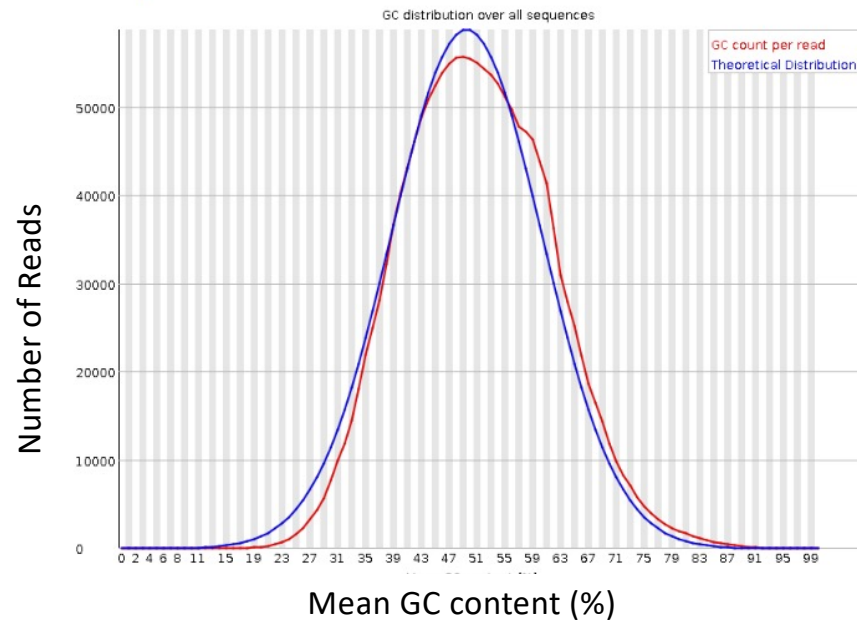
BAD

Read quality drops at the beginning and end



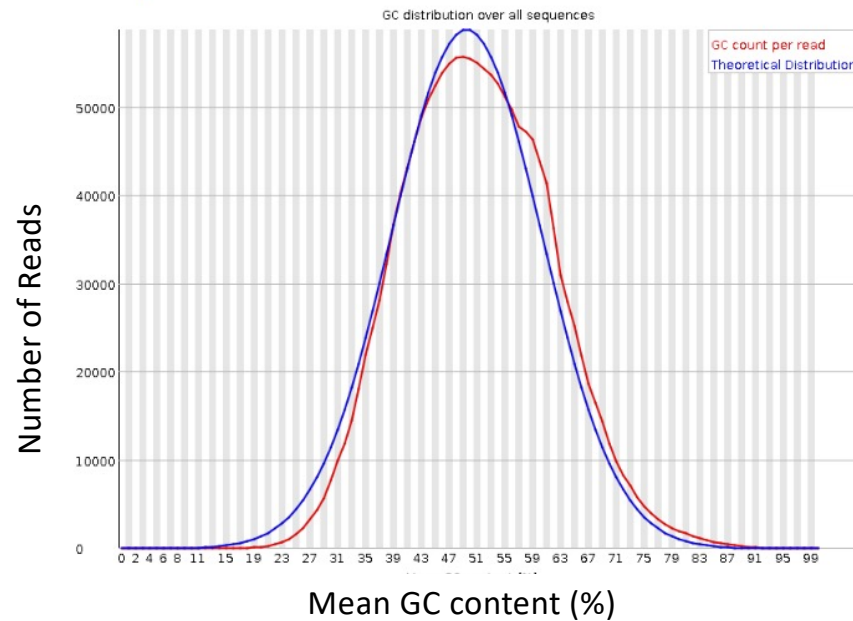
FastQC: Per sequence GC content

✔ Per sequence GC content



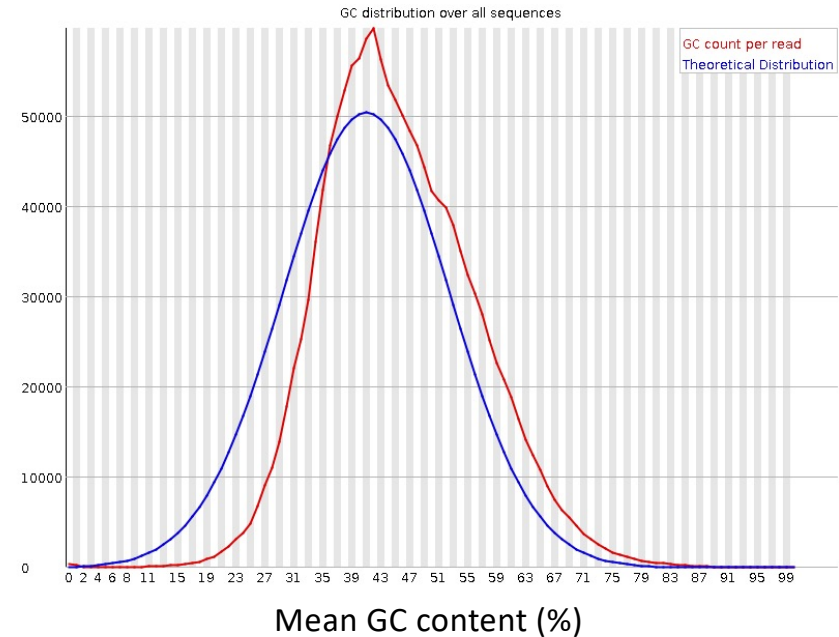
FastQC: Per sequence GC content

✔ Per sequence GC content



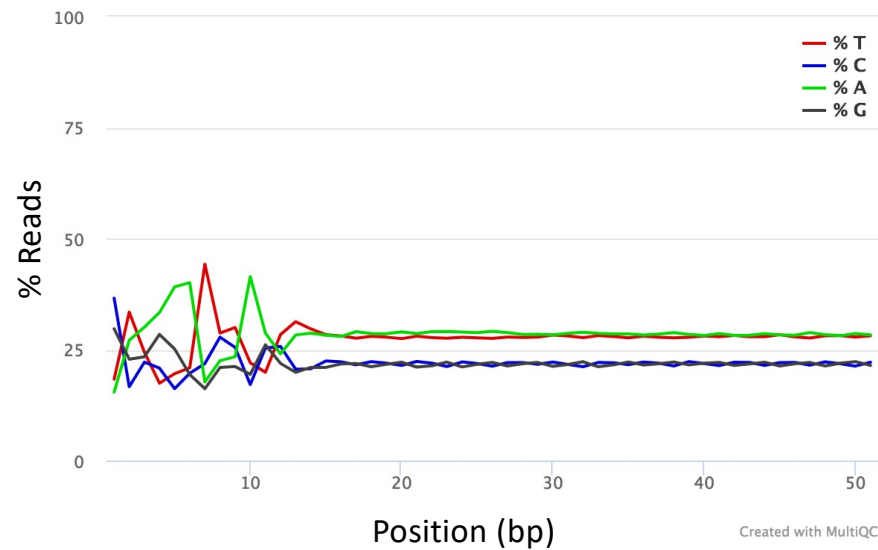
GOOD: follows normal distribution (sum of deviations is < 15% of reads)

✘ Per sequence GC content



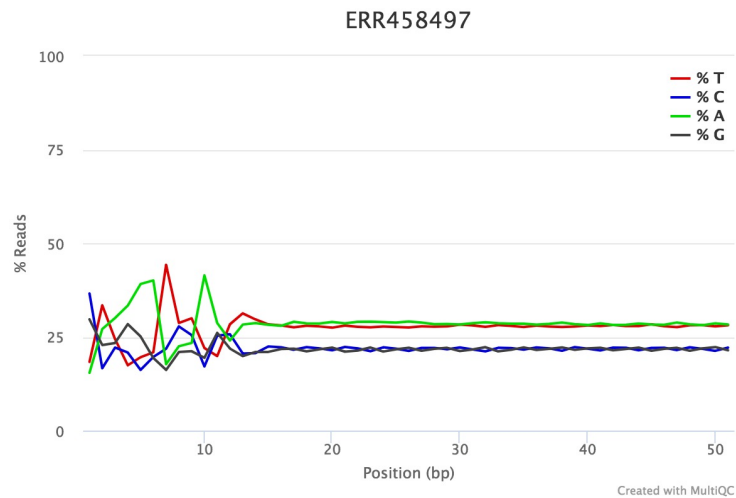
BAD: can indicate contamination with adapter dimers, or another species

FastQC: Per Base Sequence Content

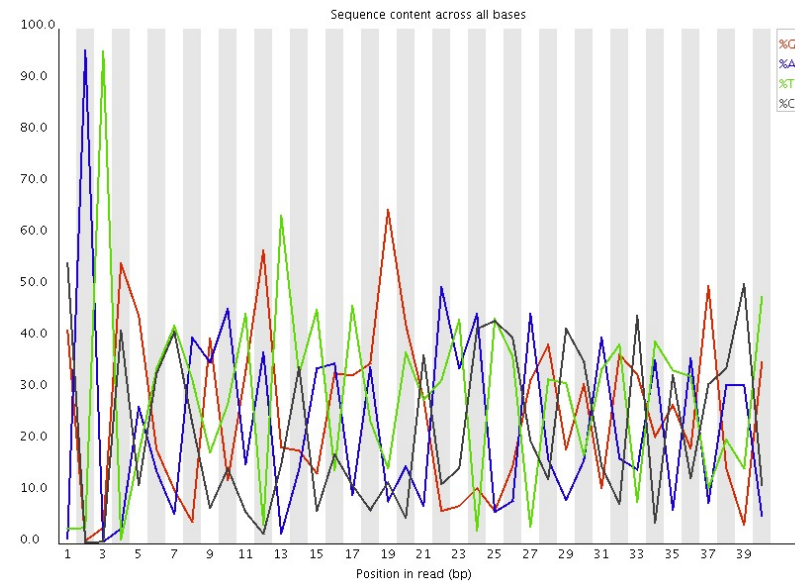


- Proportion of each position for which each DNA base has been called
- RNAseq data tends to show a positional sequence bias in the first ~12 bases
- The "random" priming step during library construction is not truly random and certain hexamers are more prevalent than others

FastQC: Per Base Sequence Content



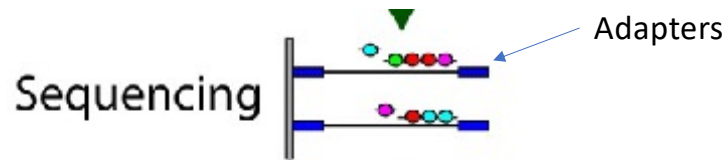
EXPECTED for RNAseq



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

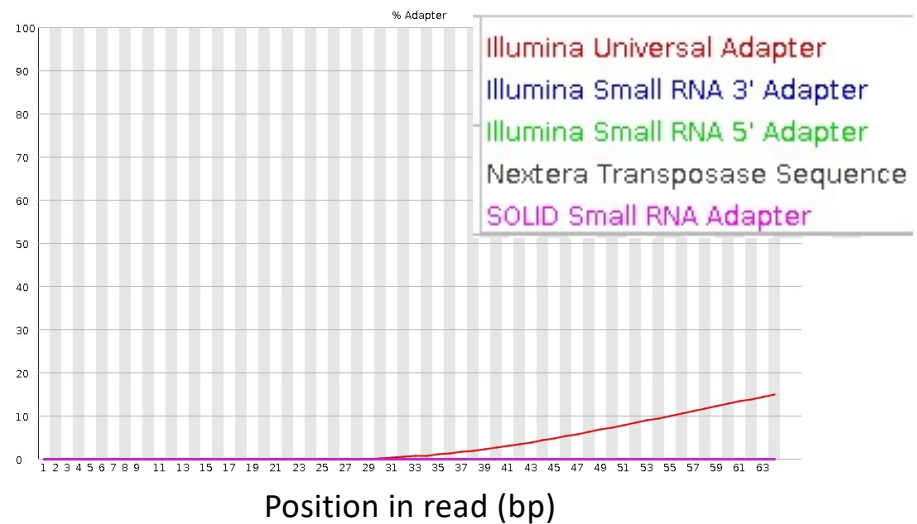
FastQC: Adapter content



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

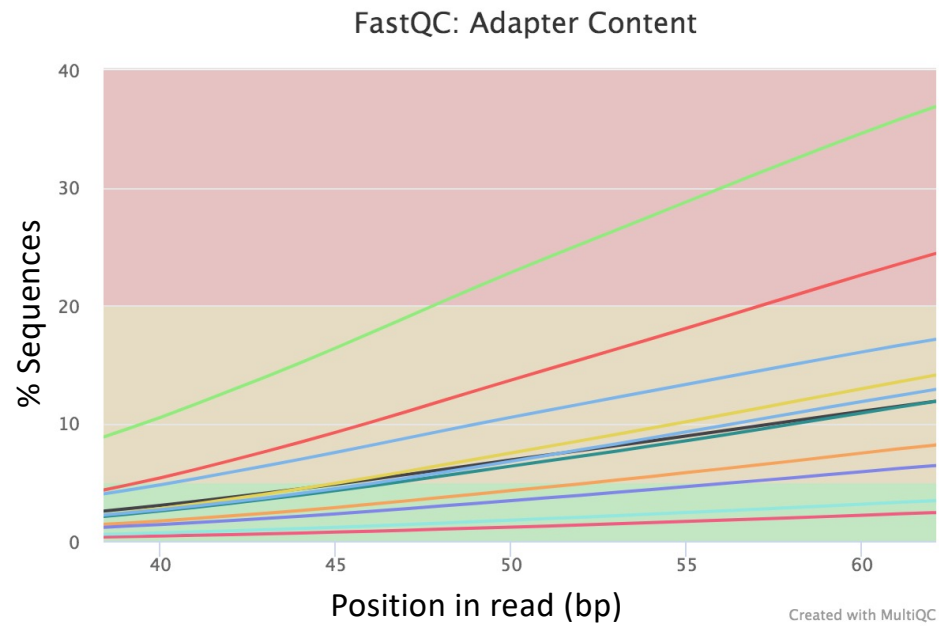
Solution – Adapter trimming!



sequencing.qcfail.com

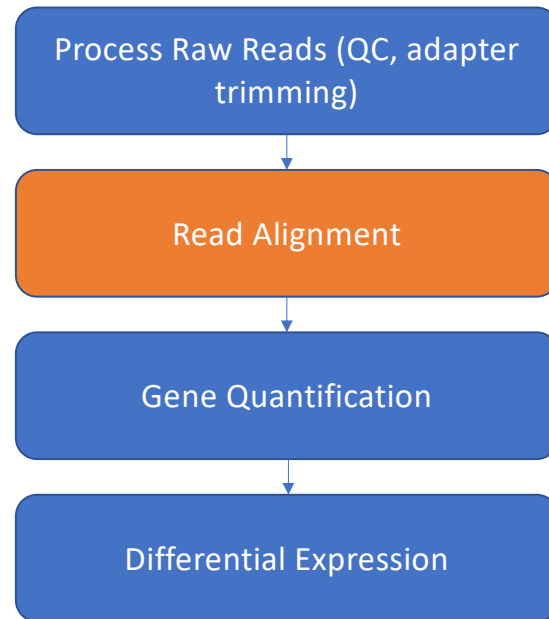
FastQC -> MultiQC

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



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 an exon-aware mapping algorithm.

Reference sequence



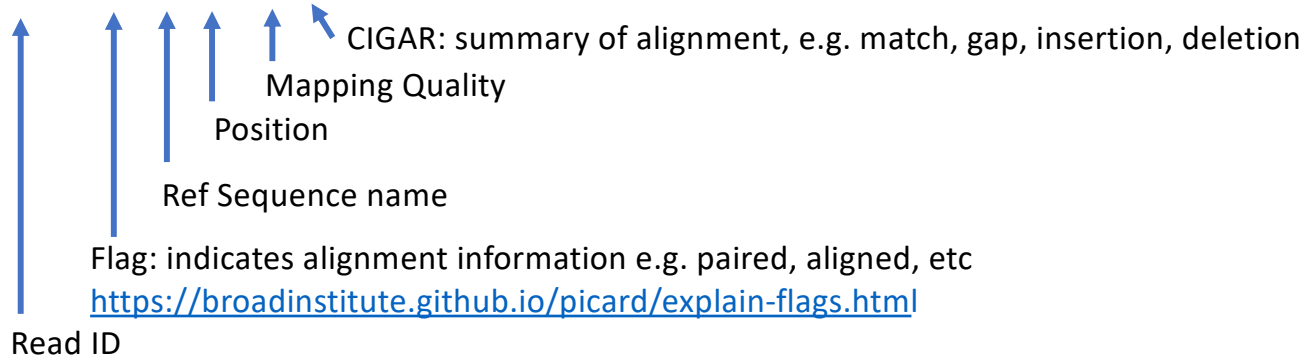
Sequence Alignment Map (SAM)



Header section										
@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



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

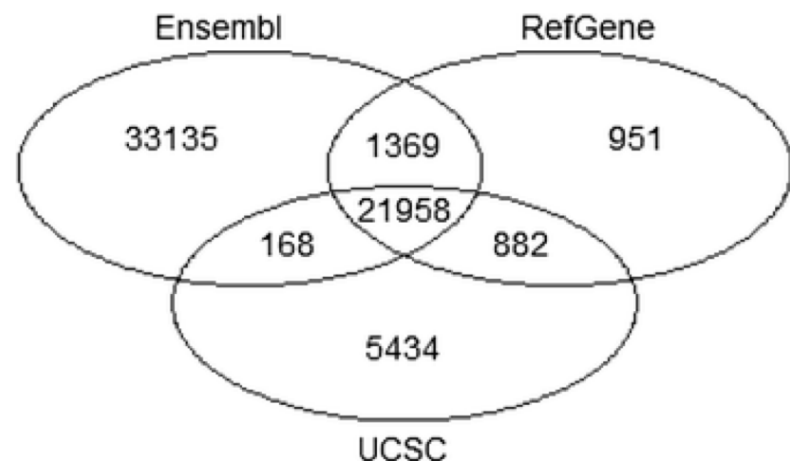
Sequence

Quality Score

Optional Fields

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!



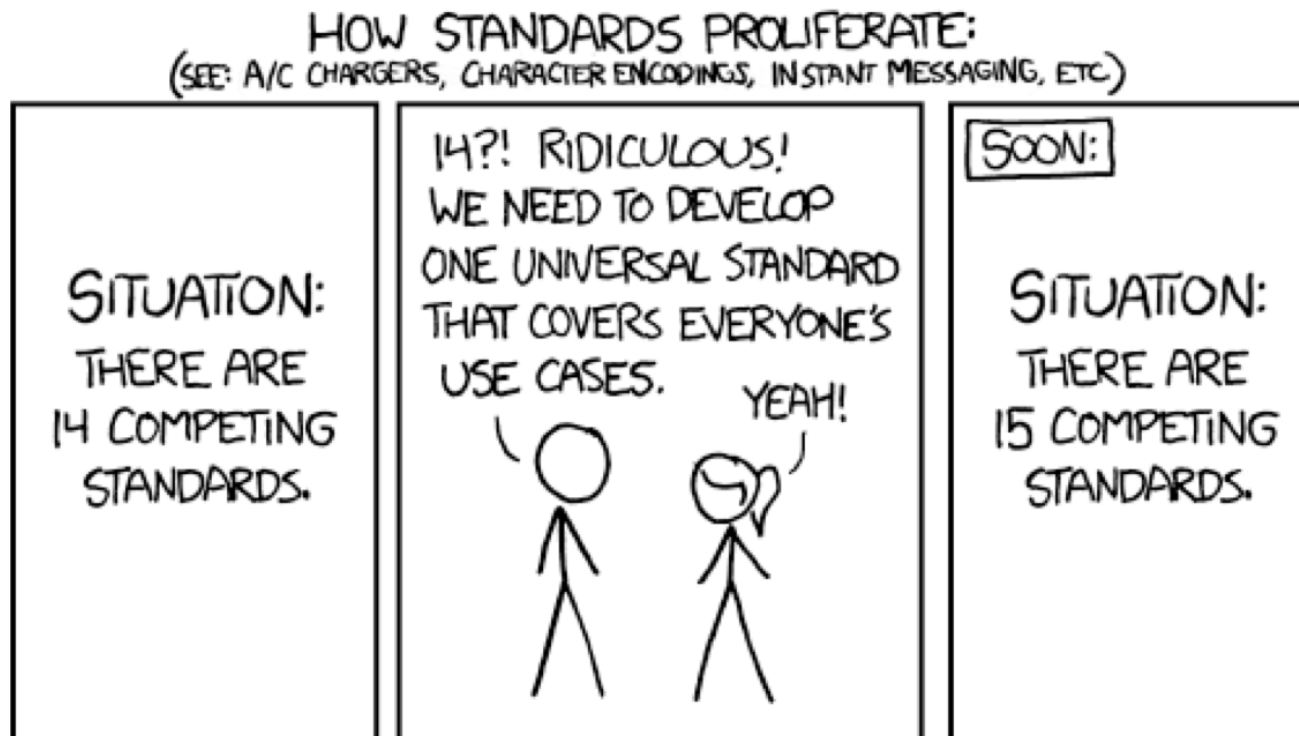
Gene Annotation Format (GTF)

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

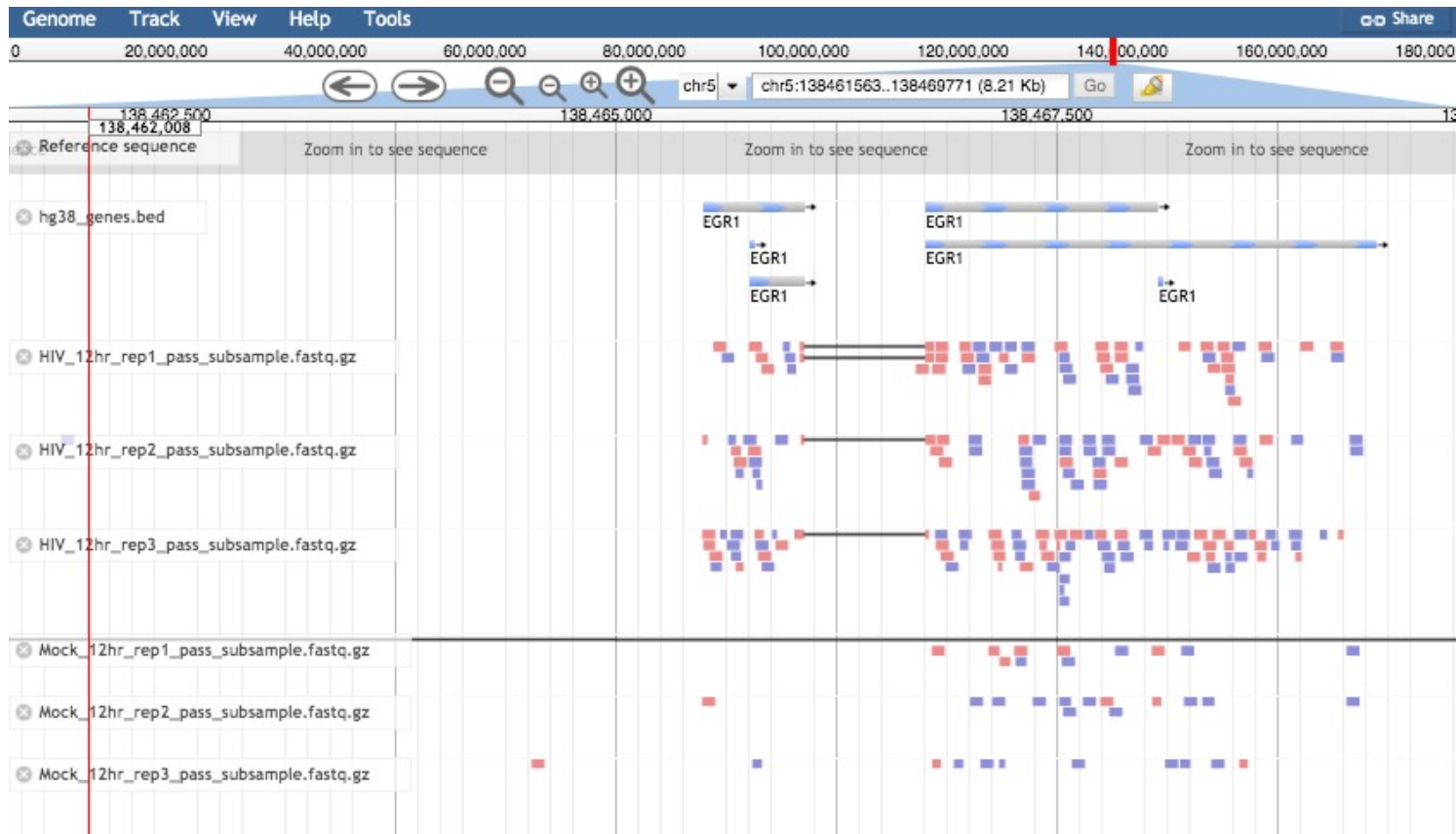
Chrom	Source	Feature type	Start	Stop	Frame Strand (Score)			Attribute
chr5	hg38_refGene	exon	138465492	138466068	.	+	.	gene_id "EGR1";
chr5	hg38_refGene	CDS	138465762	138466068	.	+	0	gene_id "EGR1";
chr5	hg38_refGene	start_codon	138465762	138465764	.	+	.	gene_id "EGR1";
chr5	hg38_refGene	CDS	138466757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

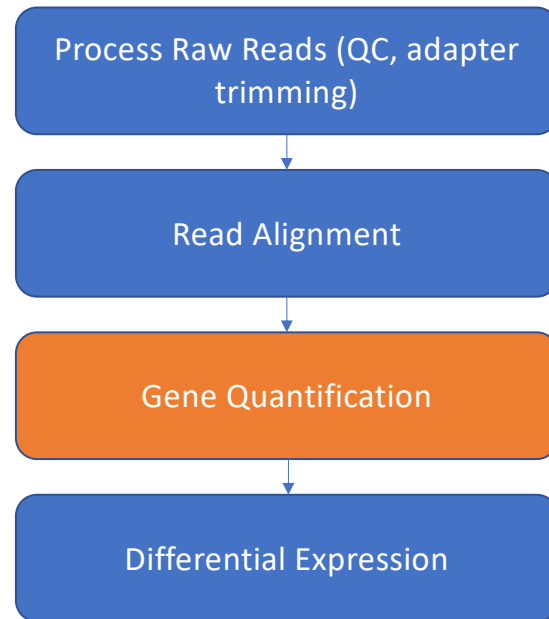
A note on standards



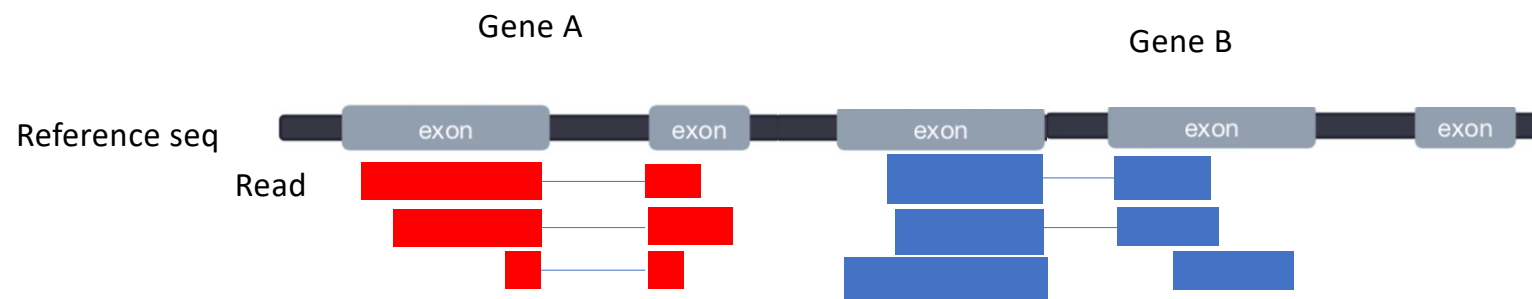
Visualizing reads with JBrowse



Workflow

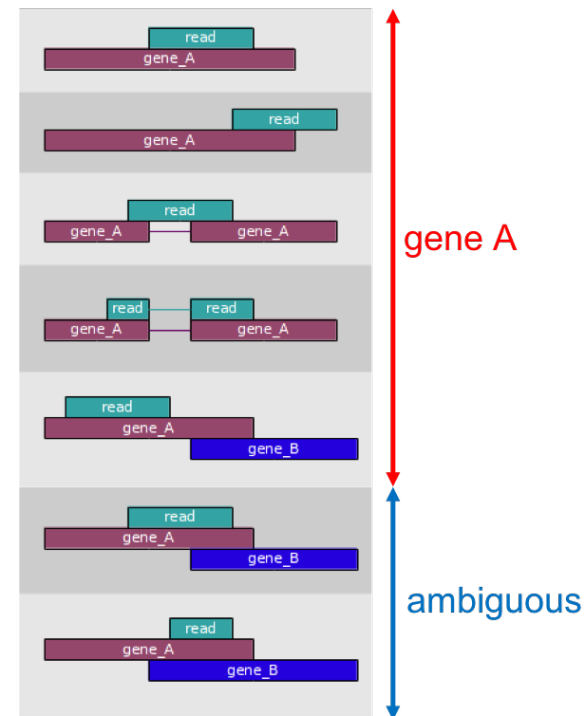


Counting reads for each gene



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 ≥ 1 bp are counted as belonging to that feature
- Ambiguous reads will be discarded

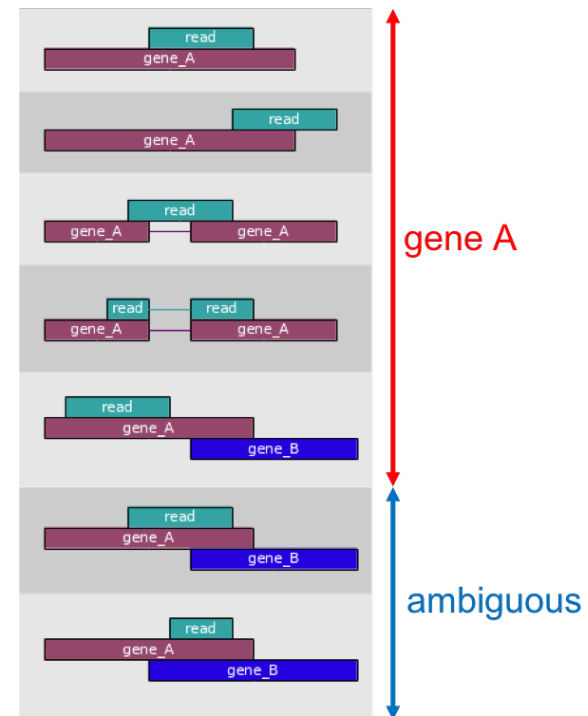


Counting reads: featurecounts

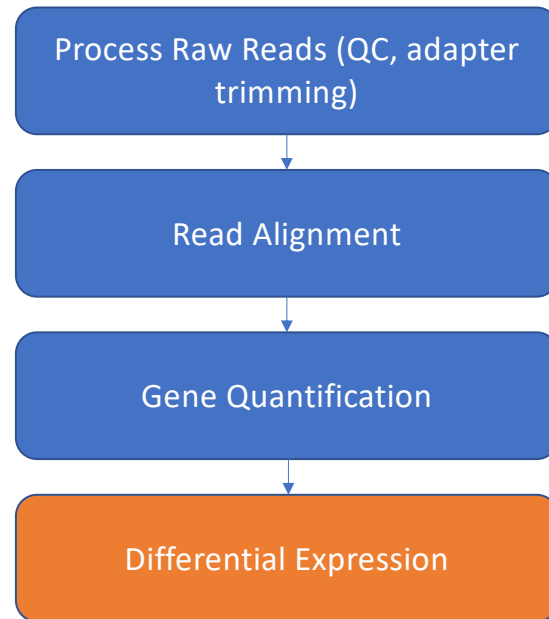
- The mapped coordinates of each read are compared with the features in the GTF file
- Reads that overlap with a gene by ≥ 1 bp are counted as belonging to that feature
- Ambiguous reads will be discarded

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

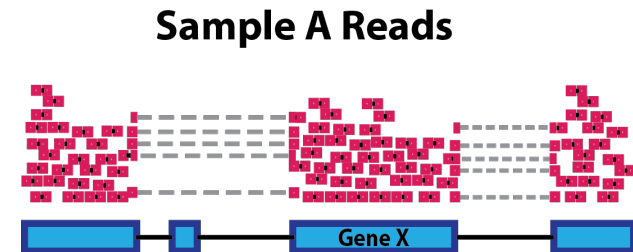


Workflow



Normalization

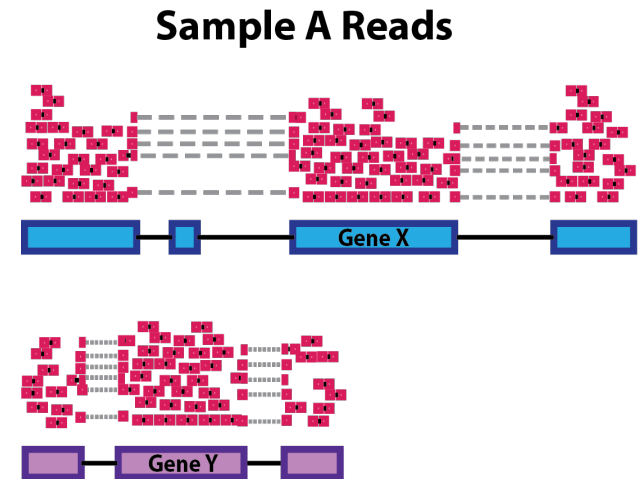
- Raw Count != Expression strength
- Normalization:
 - Eliminates factors that are not of interest for our experiment
 - Enables accurate comparison between samples or genes



Normalization

The number of reads mapped to a gene depends on

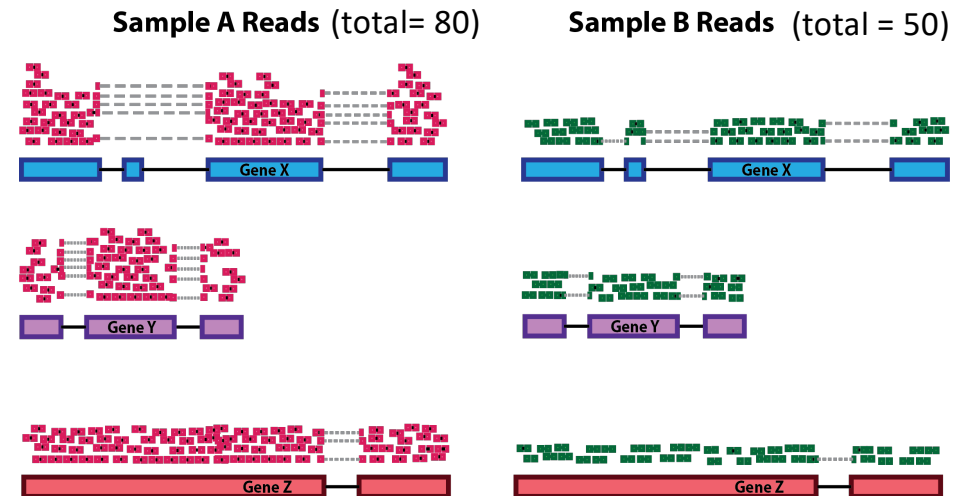
- **Gene Length**



Normalization

The number of reads mapped to a gene depends on

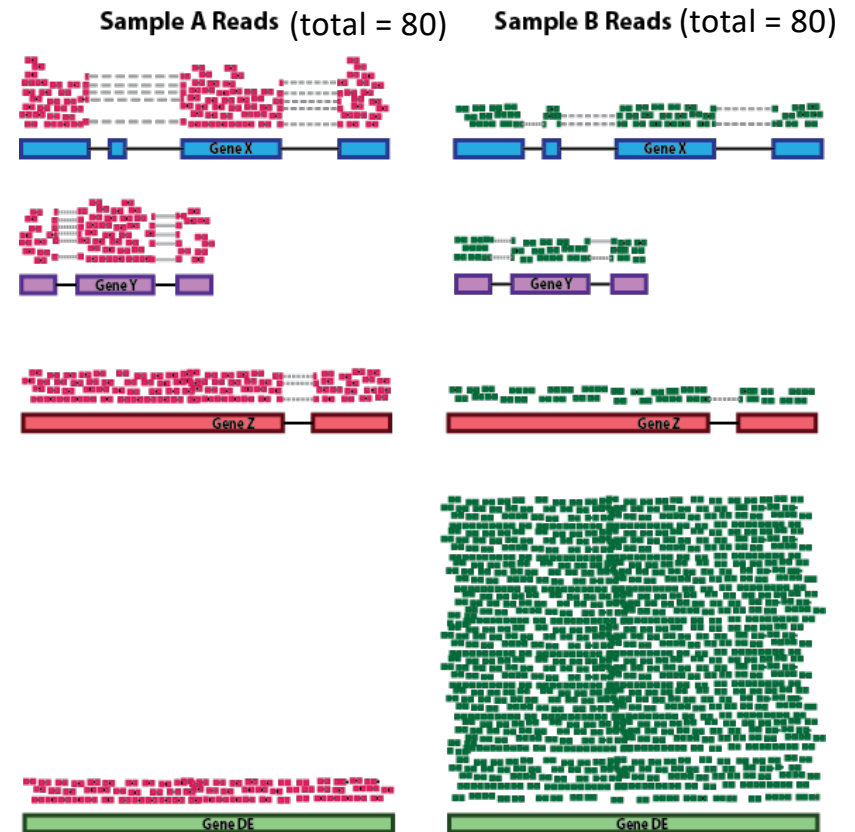
- Gene Length
- **Sequencing depth**



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)**



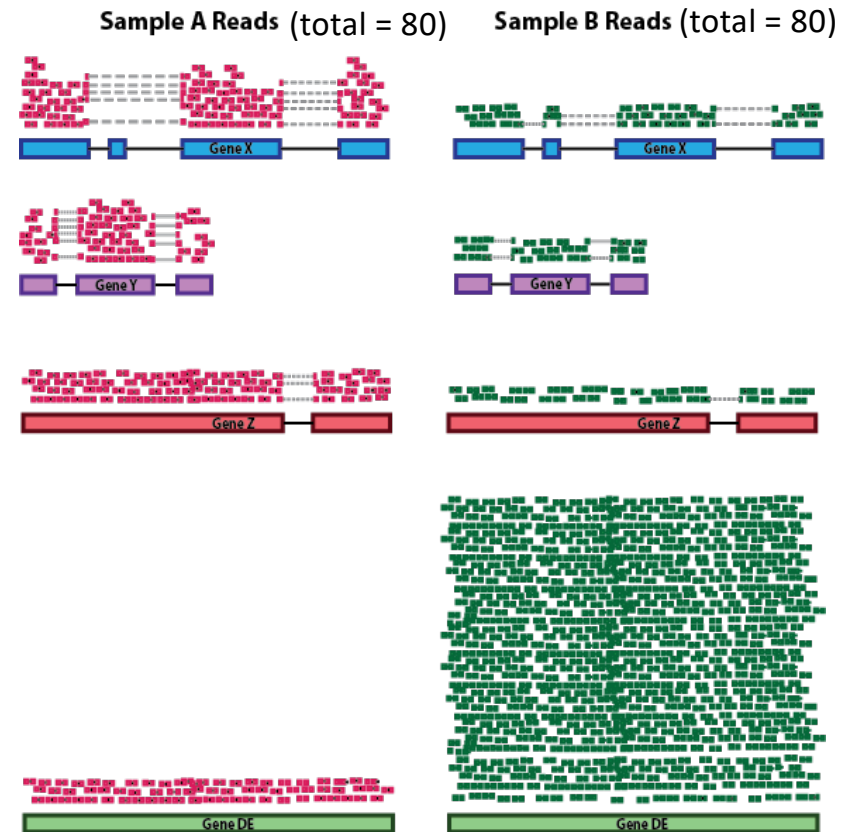
Adapted from https://hbctraining.github.io/DGE_workshop

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)

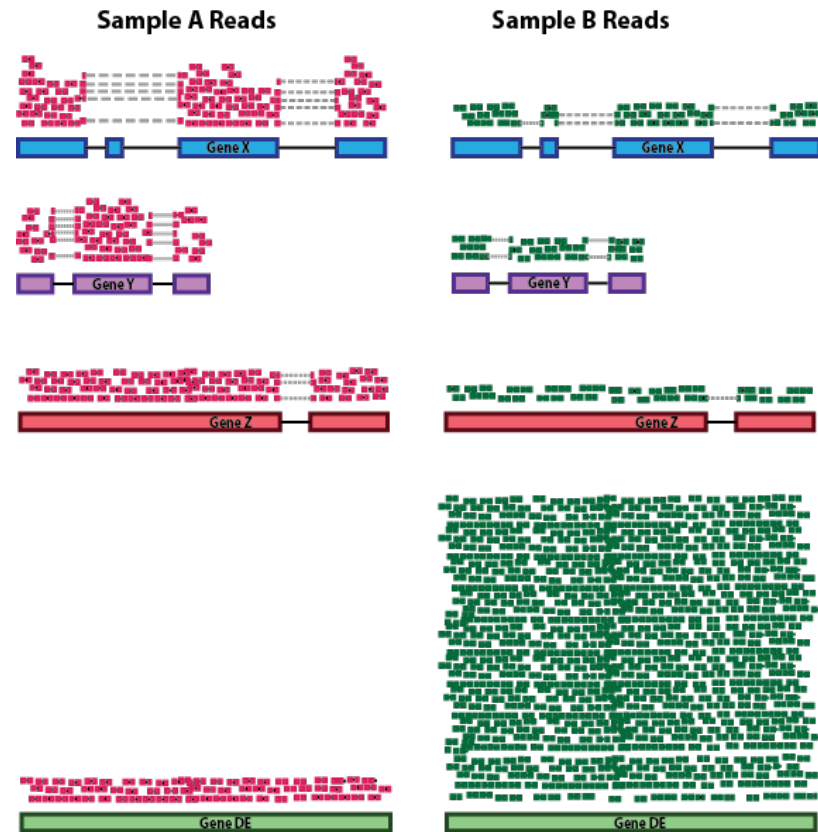
DESeq2 Median of Ratios



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	Avg. Sample
X	26	10	16
Y	26	10	16
Z	26	10	16
DE	2	50	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	Avg. Sample
X	26	10	16
Y	26	10	16
Z	26	10	16
DE	2	50	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) $\sqrt[n]{x_1 x_2 \cdots x_n}$

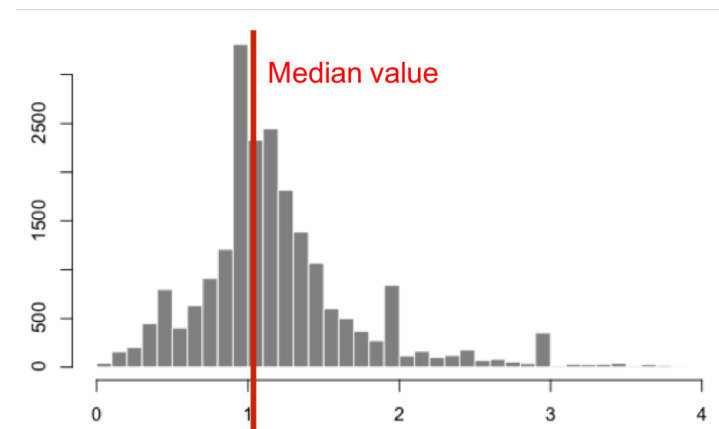
Gene	Sample A	Sample B	Avg. Sample
X	26	10	16
Y	26	10	16
Z	26	10	16
DE	2	50	16

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	Avg. Sample
X	26	10	16
Y	26	10	16
Z	26	10	16
DE	2	50	16

2. Divide all rows by the Average Sample for that gene (**Ratio**)
4. Divide all counts by sample specific size factor

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

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!

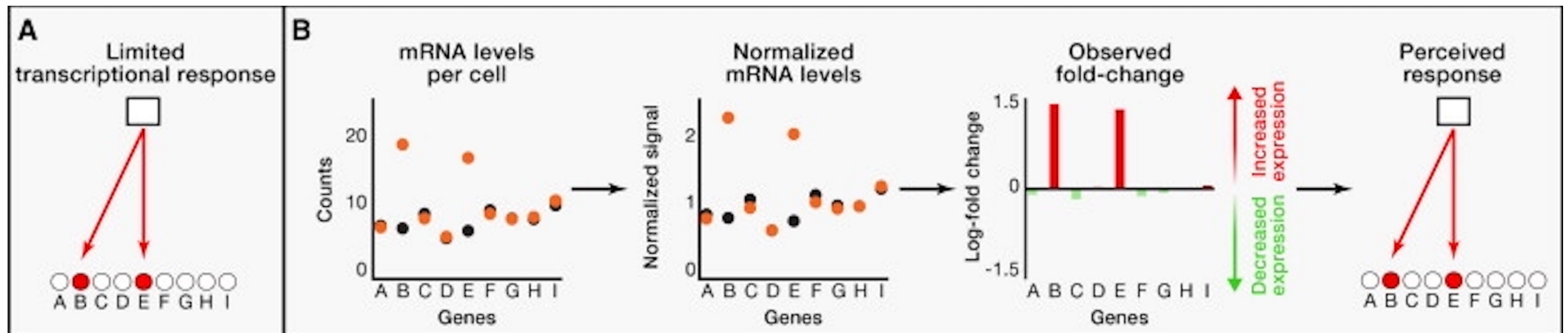
3. Take the **median** of each column. Should be ~ 1 for all

Size factor	1.6	0.6
-------------	-----	-----

`estimateSizeFactors(dds)`

Assumption of DESeq2 Median of Ratios

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

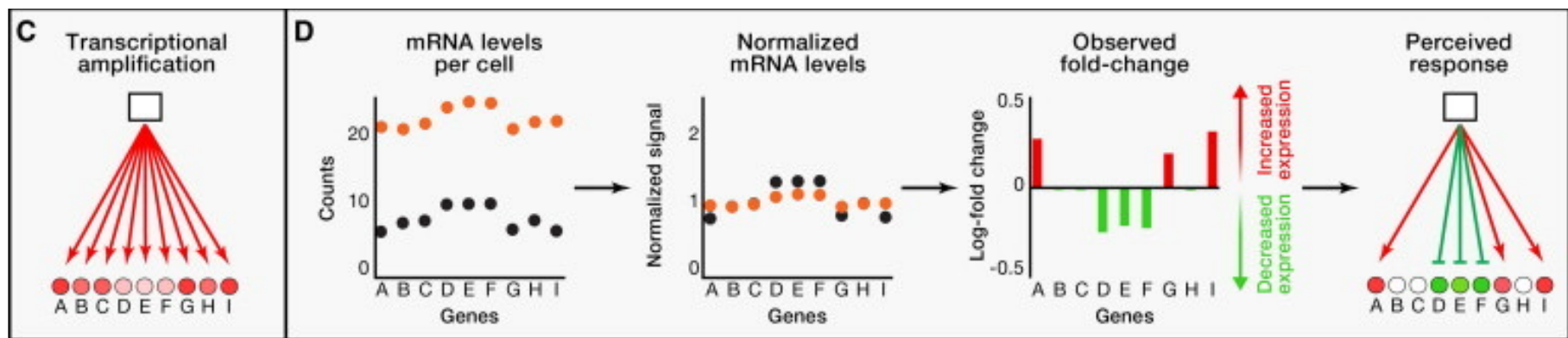


Loven et al "Revisiting Global Gene Expression Analysis" Cell 2012 <https://doi.org/10.1016/j.cell.2012.10.012>

Assumption of DESeq2 Median of Ratios

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

COUNTER EXAMPLE



- 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.

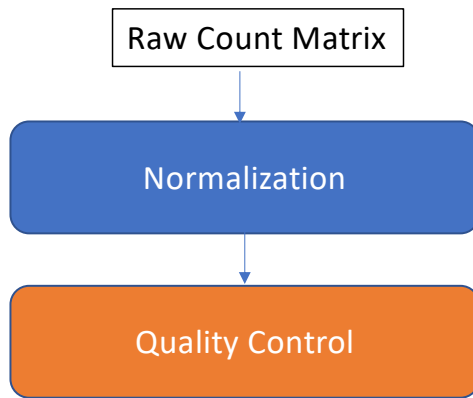
Loven et al "Revisiting Global Gene Expression Analysis" Cell 2012 <https://doi.org/10.1016/j.cell.2012.10.012>

Normalization methods

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

Similar to DESeq2: EdgeR, limma-voom

Quality Control Visualizations



Examine sources of variation in the data

- Principal Component Analysis
- Hierarchical Clustering

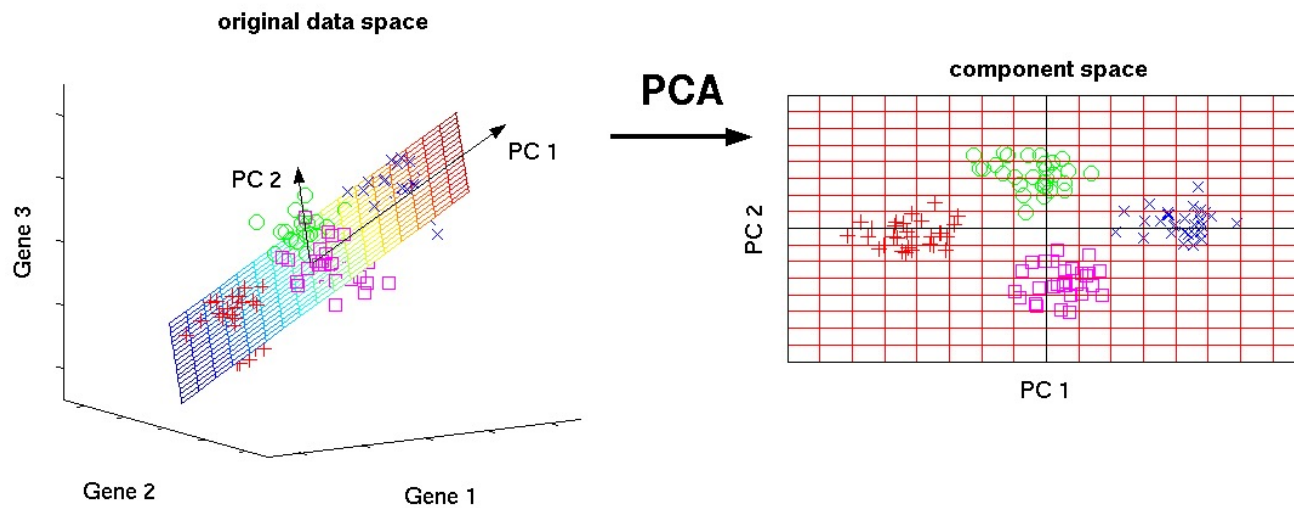
(Log₂ + 1) Transformed, Normalized Count Table

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

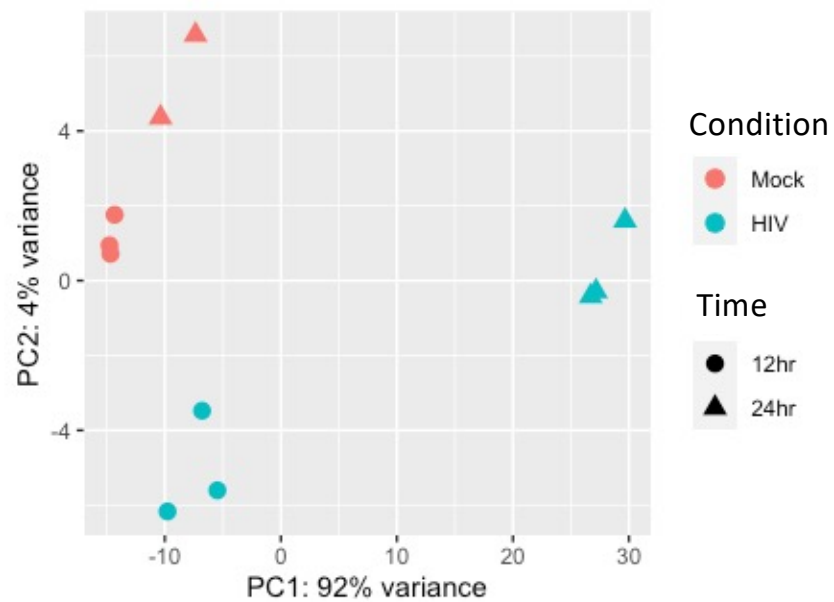
Gene	Mock_12h	Mock_12h	Mock_24h	Mock_24h	HIV_12h	HIV_12h	HIV_24h	HIV_24h
Gene 1	8.9	8.9	8.9	9.0	8.9	8.9	9.0	6.8
Gene 2	0.6	-1.0	0.6	-1.0	0.6	-1.0	0.6	3.8
Gene 3	4.1	11.9	4.1	-0.5	4.1	8.7	4.0	4.4



Do your samples cluster as expected?

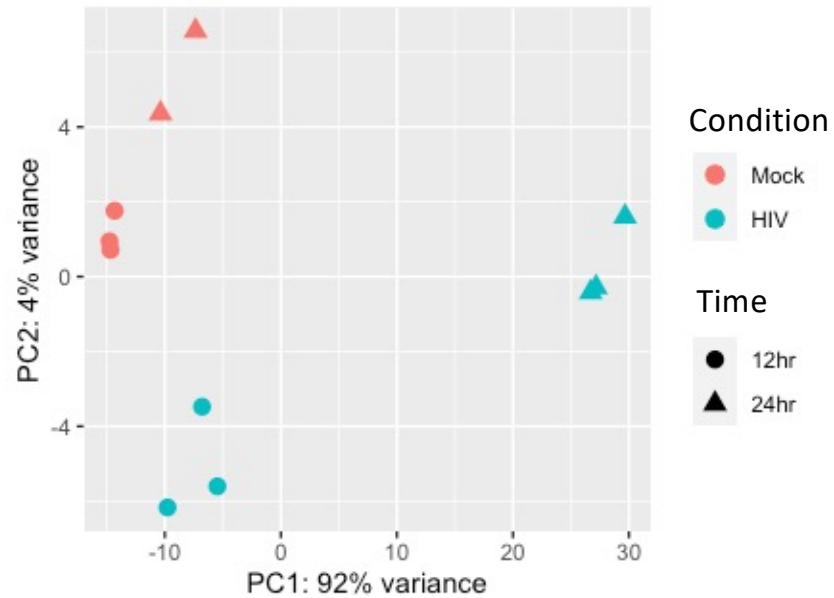
What are the major sources of variation in the data?

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?

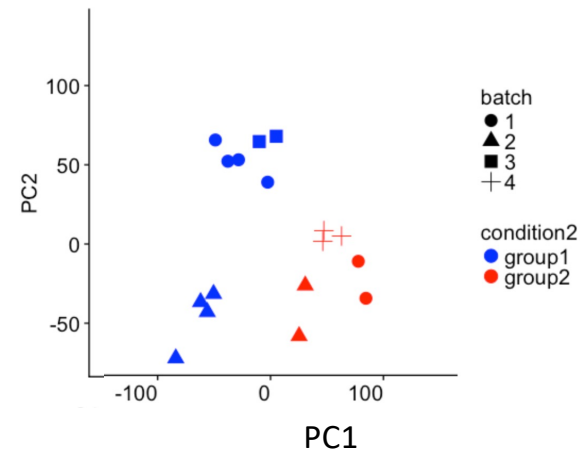
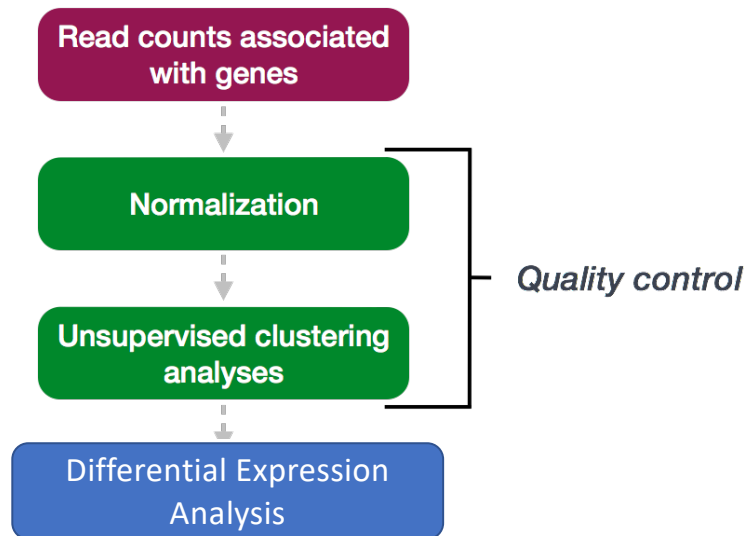
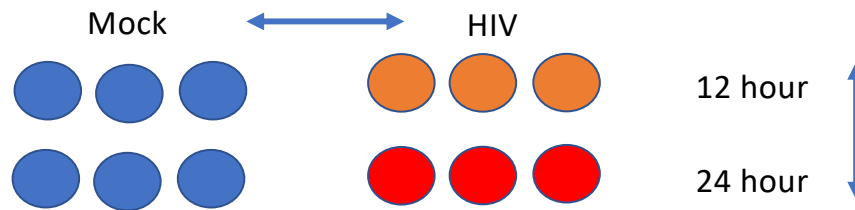


Image <https://support.bioconductor.org/p/111491/>

Differential Expression with DESeq2



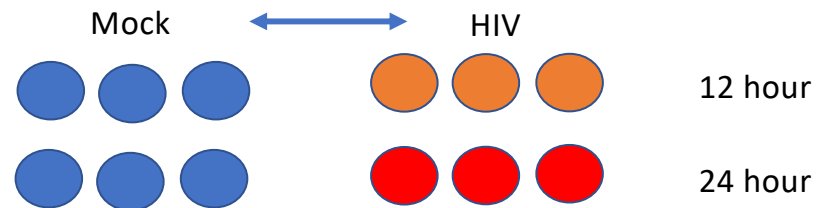
Multi-factor experiment design



Factor 1:
Infection status (Mock or HIV)

Factor 2:
Time (12 or 24 hr)

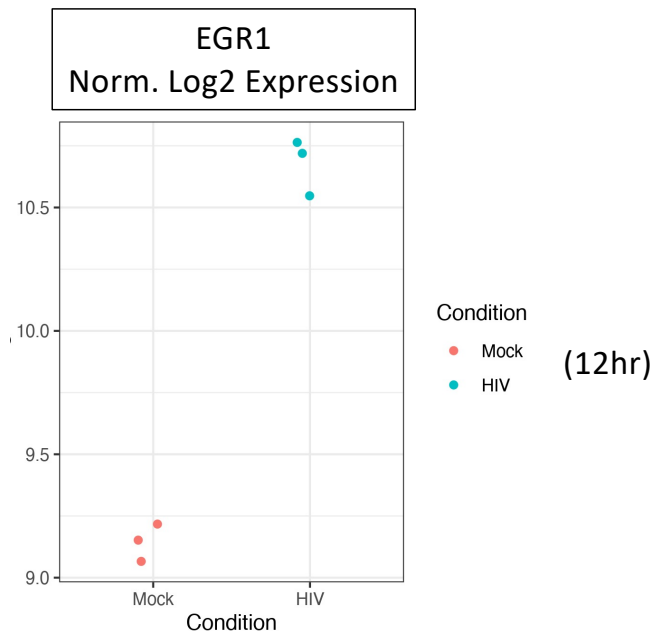
Multi-factor experiment design



- 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

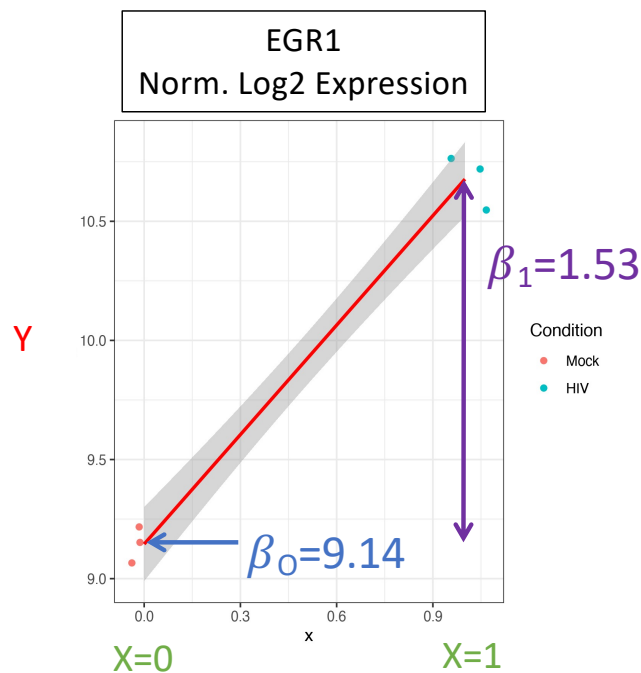
Step 1: Modeling gene expression values

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



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:



$$Y = \beta_0 + \beta_1 X + e$$

Log2 Expression Values

Intercept

Condition (0-Mock, 1-HIV)

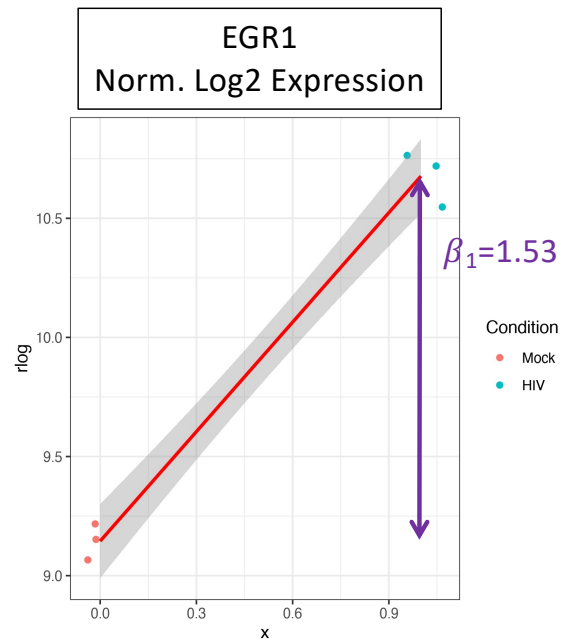
Slope: difference between Mock /HIV

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



$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0$$

H_0 : 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

Is EGR1 differentially expressed?

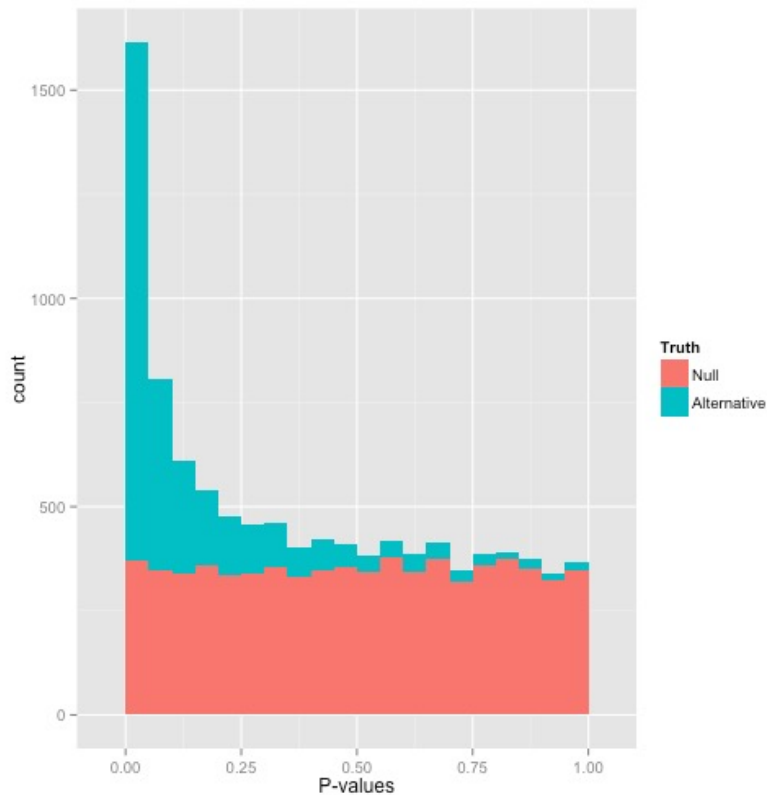
Yes! $p \ll 0.05$

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_0 were true
- P-adj – accounting for multiple testing correction

DESeq2 P-value histogram



- Histogram of raw p-values for all genes examined
- P-value: Probability of getting a log2FoldChange as extreme as observed if the true log2FoldChange = 0 for that gene (null hypothesis)

How to interpret:

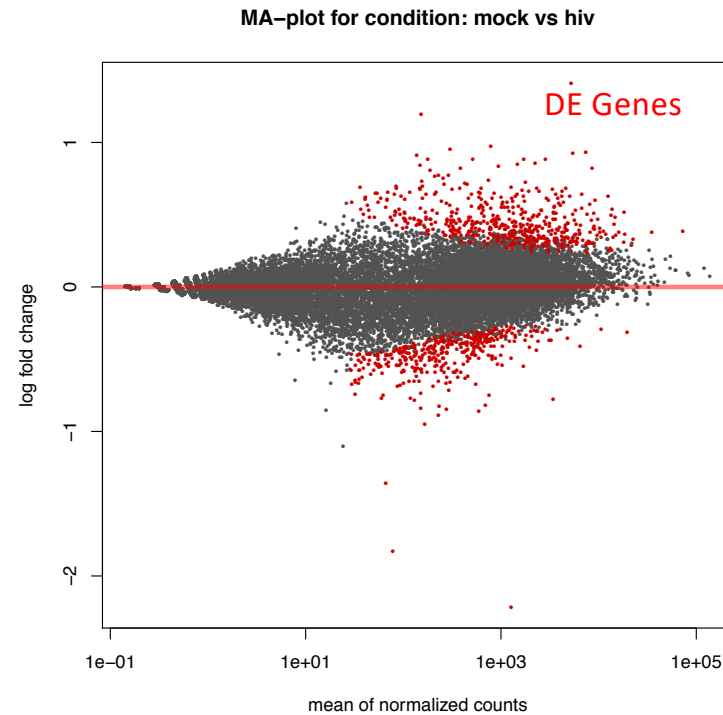
- Random P-values are expected to be uniform, if you have true positives you should see a peak close to zero

<http://varianceexplained.org/statistics/interpreting-pvalue-histogram/>

DESeq2 MA plot

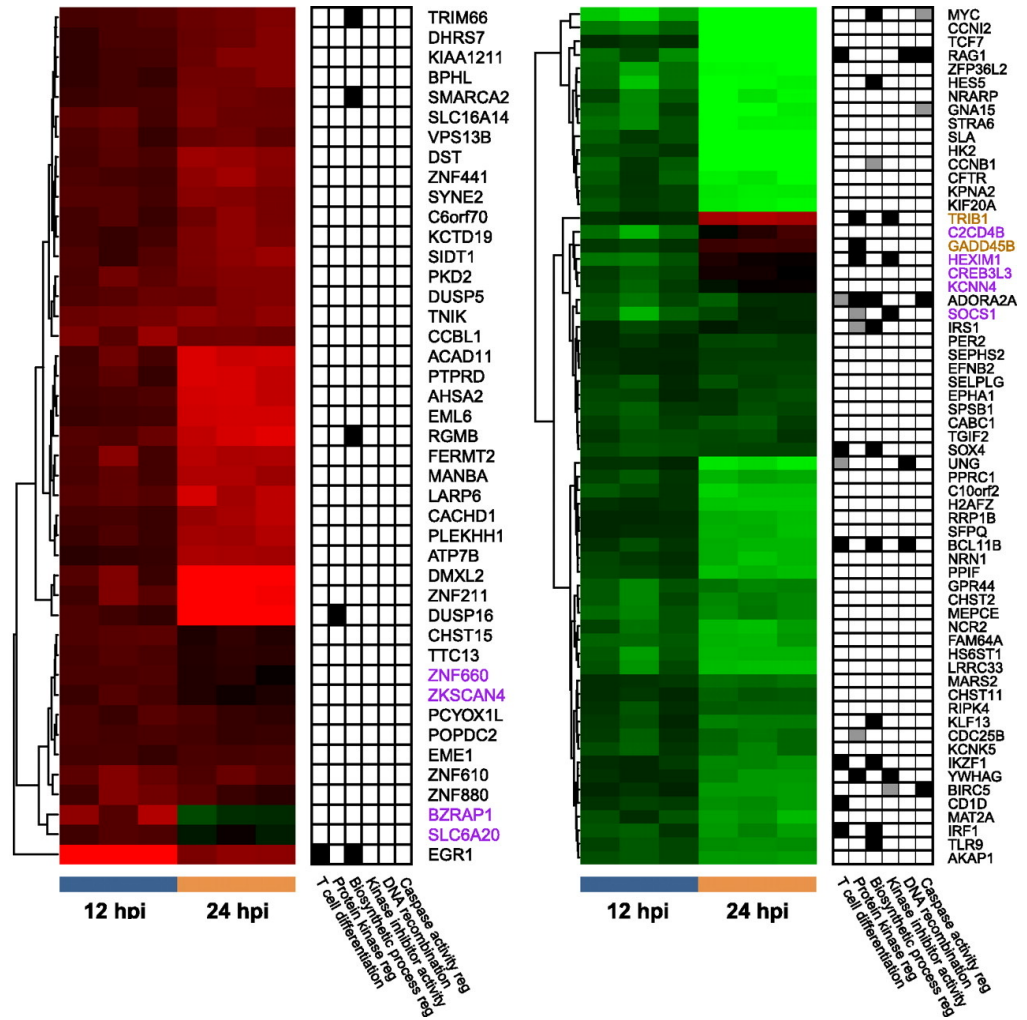
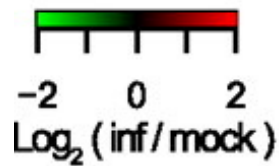
Shows the relationship between

- 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
- Gives an overview of your results

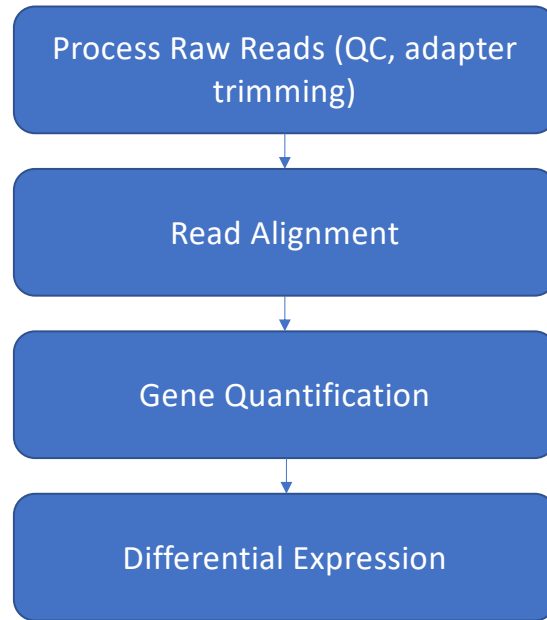


Study findings

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



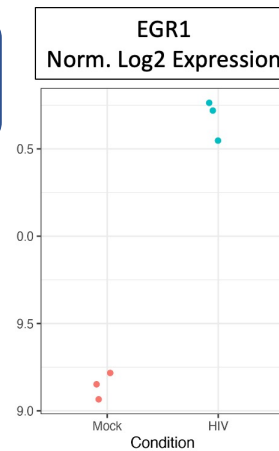
Conclusions



```
@SRR497699.30343179.1 HWI-EAS39X_10175_FC61MK0_4_117_4812_10346 length=75  
CAGATGGCCCGAGGAAGCCATGAAGGCCCTGCATGGGAGATCGGAAGCGGTTTCAGCAGGAATGCCGAGAC  
+  
IIIIIGIIFIIIBIIIDII>IIDHIIHDIIIGIFIIEIGIBDDEFIG<EIEGEEG;<DB@A8CC7<<C@BBDD8
```



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



$\log_2\text{FoldChange} = 1.55$
 $\text{Adjusted } p\text{-val} \ll 0.05$

References

DESeq2 vignette (R/Rstudio):

<http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#differential-expression-analysis>

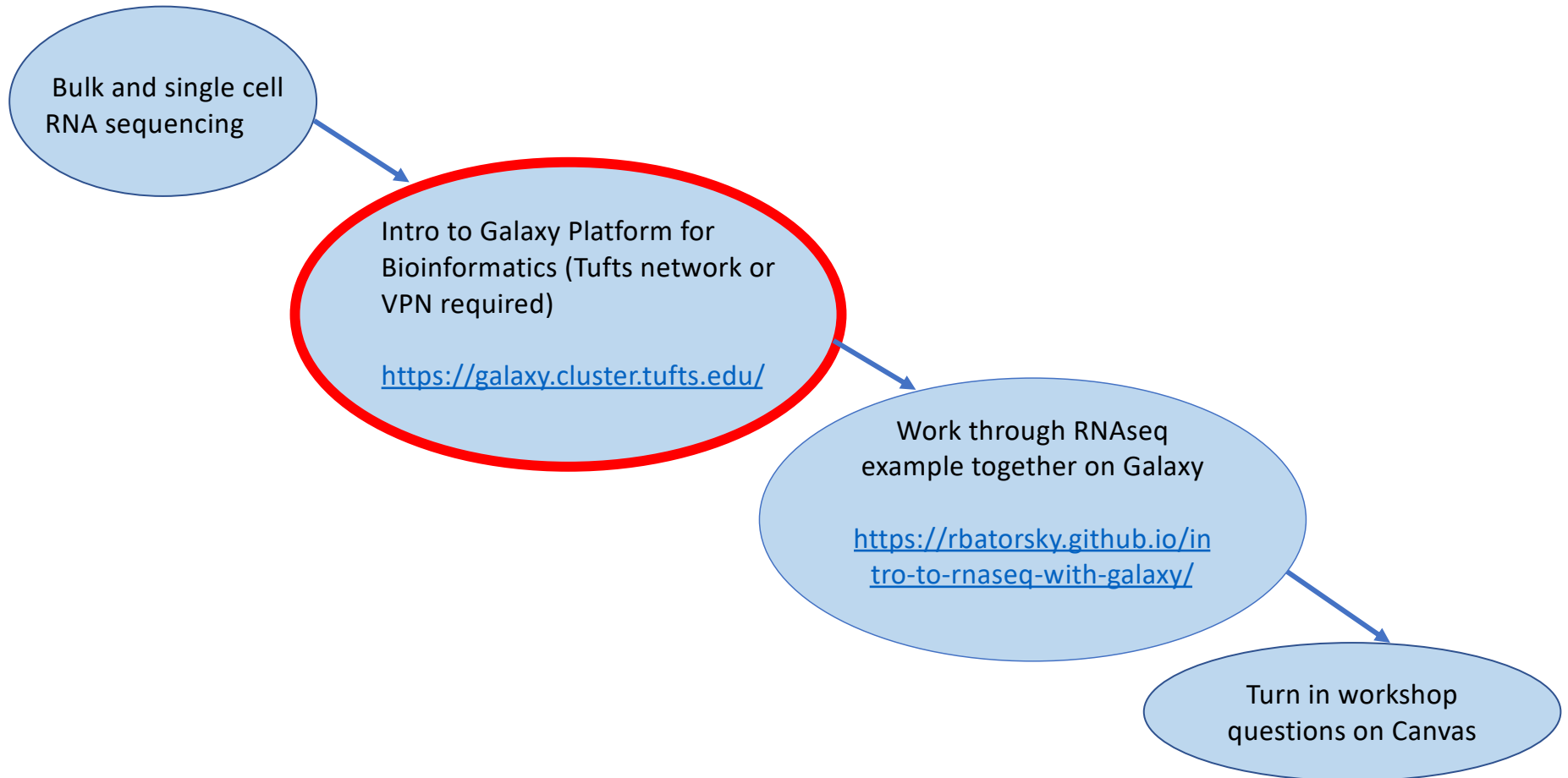
HBC Training (Command line/R):

https://hbctraining.github.io/DGE_workshop

Galaxy Training:

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

Outline



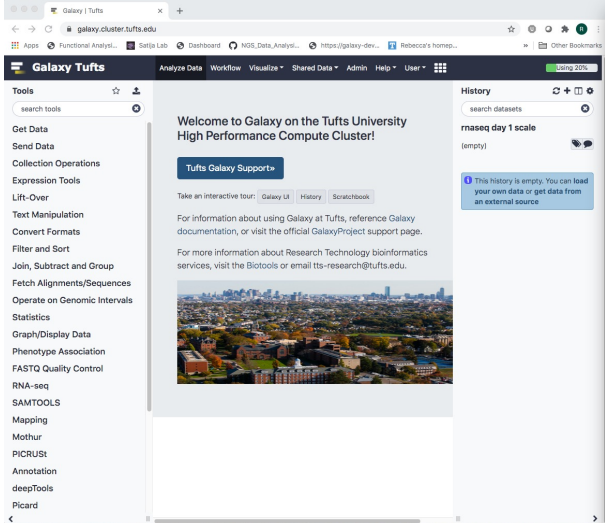


- ❖ **Web-based** platform for running data analysis and integration, geared towards bioinformatics
 - Open-source
 - Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic with many more outside contributions
 - Large and extremely responsive community

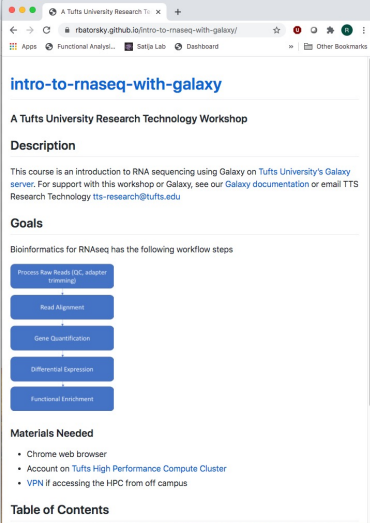
Access Galaxy

1. Connect to Tufts Network, either on campus or via [VPN](#)
2. Visit <https://galaxy.cluster.tufts.edu/>
3. Log in with you cluster username and password
4. In another browser window go to course workflow: <https://rbatorsky.github.io/intro-to-rnaseq-with-galaxy/>

Suggested screen layout



The screenshot shows the Galaxy Tufts web interface. The main content area displays a welcome message: "Welcome to Galaxy on the Tufts University High Performance Compute Cluster!". Below this, there is a "Tufts Galaxy Support" button and a link to an interactive tour. The sidebar on the left lists various tool categories such as "Tools", "Get Data", "Send Data", "Collection Operations", "Expression Tools", "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", "Mothur", "PICRUST", "Annotation", "deepTools", and "Picard". The top navigation bar includes "Analyze Data", "Workflow", "Visualize", "Shared Data", "Admin", "Help", and "User".



The screenshot shows the course workflow page titled "intro-to-rnaseq-with-galaxy". The page is a Tufts University Research Technology Workshop. It includes a "Description" section stating: "This course is an introduction to RNA sequencing using Galaxy on Tufts University's Galaxy server. For support with this workshop or Galaxy, see our Galaxy documentation or email TTS Research Technology ts-research@tufts.edu". Below the description, there is a "Goals" section with a list of workflow steps: "Process Raw Reads (QC, adapter trimming)", "Read Alignment", "Gene Quantification", "Differential Expression", and "Functional Enrichment". Each step is represented by a blue button. At the bottom, there is a "Materials Needed" section with a list: "Chrome web browser", "Account on Tufts High Performance Compute Cluster", and "VPN if accessing the HPC from off campus". A "Table of Contents" section is also visible at the bottom.

User Interface

The screenshot displays the Galaxy Tufts user interface. At the top, a dark navigation bar contains the 'Galaxy Tufts' logo, a search bar for tools, and a menu with options: Analyze Data, Workflow, Visualize, Shared Data, Admin, Help, and User. A 'Using 20%' indicator is visible in the top right corner.

The left sidebar features a 'Tools' section with a search bar and a list of tool categories: Get Data, Send Data, Collection Operations, Expression Tools, 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, and SAMTOOLS.

The main content area displays a welcome message: 'Welcome to Galaxy on the Tufts University High Performance Compute Cluster!'. Below this is a 'Tufts Galaxy Support' button and a section for an interactive tour with buttons for 'Galaxy UI', 'History', and 'Scratchbook'. Further down, there are two paragraphs of text providing information about using Galaxy at Tufts and Research Technology bioinformatics services, along with an email address: tts-research@tufts.edu. An aerial photograph of the Tufts University campus is shown at the bottom of the main content area.

The right sidebar shows a 'History' section with a search bar for datasets and an 'Unnamed history' section that is currently empty. A blue information box states: 'This history is empty. You can load your own data or get data from an external source'.

At the bottom left of the interface, a small status bar shows 'javascript:void(0)'. A right arrow icon is visible at the bottom right of the main content area.

User Interface

TOP MENU BAR

The screenshot displays the Galaxy user interface with three main sections highlighted by colored boxes:

- TOOLS (Green box):** A sidebar on the left containing a search bar and a list of tool categories: Get Data, Send Data, Collection Operations, Expression Tools, 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, and SAMTOOLS. The 'RNA-seq' and 'SAMTOOLS' categories are highlighted with a green border.
- MAIN (Purple box):** The central content area with the heading "Welcome to Galaxy on the Tufts University High Performance Compute Cluster!". It includes a "Tufts Galaxy Support»" button, a "Take an interactive tour:" section with links for "Galaxy UI", "History", and "Scratchbook", and two paragraphs of introductory text. A large image of the Tufts University campus is shown at the bottom.
- HISTORY (Red box):** A sidebar on the right titled "History" with a search bar and the text "Unnamed history (empty)". A blue information box states: "This history is empty. You can load your own data or get data from an external source".

The top navigation bar includes links for "Analyze Data", "Workflow", "Visualize", "Shared Data", "Admin", "Help", and "User".

Galaxy User Interface

To return to home screen

The screenshot shows the Galaxy Tufts user interface. The top navigation bar includes 'Galaxy Tufts', 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Admin', 'Help', and 'User'. A 'Using 30%' indicator is in the top right. The left sidebar contains a 'Tools' section with a search bar and a list of tool categories: Get Data, Send Data, Collection Operations, Expression Tools, 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, Mothur, and PICRUST. The main content area displays a welcome message: 'Welcome to Galaxy on the Tufts University High Performance Compute Cluster!' with a 'Tufts Galaxy Support»' button. Below this, there are links for 'Galaxy UI', 'History', and 'Scratchbook', and text providing information about using Galaxy at Tufts and contacting support. A cityscape image is also present. The right sidebar shows a 'History' section with a search bar and an 'Unnamed history' section that is currently empty. A blue notification box states: 'This history is empty. You can load your own data or get data from an external source'. At the bottom, there are four red circles highlighting the minimize/adjust toolbars: a left arrow, a square, a square, and a right arrow.

Minimize/Adjust toolbars

History

Create New History

View all Histories

History

search datasets

Unnamed history

(empty)

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

History

Create New History

View all Histories

History

search datasets

Unnamed history

(empty)

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

Galaxy

Analyze Data Workflow Visualize Shared Data Admin Help User

Using 33.9 GB

search histories search all datasets

Current History Unnamed history Unnamed history Unnamed history Unnamed history

RNA-seq 6.8 GB 122: WT_3_collection 134: WT_3_collection 106: WT_1_collection 98: SNF2_3_collection 90: SNF2_2_collection 82: SNF2_1_collection 76: Concatenate datasets on d ata 68, data 67, and others 73: Concatenate datasets on d ata 61, data 60, and others 72: Concatenate datasets on d ata 64, data 63, and others 71: Concatenate datasets on d ata 47, data 46, and others

5: Concatenate datasets on dat a 1 and data 2 2: mov10_on1/SRR060460_pas s.fastq.gz 1: mov10_on1/SRR060459_pas s.fastq.gz

181: RNA STAR on data 86: ma pped.bam 180: MultiQC on data 86, d ata 85, and others: Log 179: MultiQC on data 86, da ta 85, and others: Webpage 178: MultiQC on data 86, data 85, and others: Stats

177: featureCounts on data 163 and data 158: Summary 176: featureCounts on data 163 and data 158: Counts 175: featureCounts on data 163 and data 155: Summary 174: featureCounts on data 163 and data 155: Counts 173: featureCounts on data 163 and data 152: Summary 172: featureCounts on data 163 and data 152: Counts

3: Create DBKey and Referen ce Genome 2: Create DBKey and Referen ce Genome 1: Create DBKey and Referen ce Genome

1: bed_file.bed

Tools

The screenshot displays the Galaxy web interface. The top navigation bar includes the Galaxy logo, a search bar for tools, and various menu items like 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Admin', 'Help', and 'User'. The left sidebar contains a 'Tools' menu with a search bar and a list of tool categories. The 'RNA-seq' category is highlighted with a red circle and a green arrow. The main content area shows a 'Welcome to Galaxy on the Tufts cluster' message with a 'Bioinformatics @ Tufts' button and links for 'Galaxy UI', 'History', and 'Scratchbook'. The right sidebar shows the 'History' panel, which is currently empty.

Tools

search tools

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

DESeq2 Determines differentially expressed features from count tables

featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files.

RNA STAR Gapped-read mapper for RNA-seq data

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.

History

search datasets

Unnamed history

(empty)

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

Using 14.7 GB

Tools

Click on the name of the tool to open it in the main panel

The screenshot displays the Galaxy web interface. On the left, the 'Tools' sidebar lists various categories, with 'featureCounts' highlighted in a green box. A green arrow points from this box to the main panel. The main panel shows the configuration for the 'featureCounts' tool (Version 1.6.4). The configuration includes sections for 'Alignment file', 'Specify strand information' (set to 'Unstranded'), 'Gene annotation file' (set to 'locally cached'), 'Using locally cached annotation' (set to 'No options available'), 'Output format' (set to 'Gene-ID "\t" read-count'), and 'Create gene-length file' (set to 'Yes'). An 'Execute' button is visible at the bottom of the configuration panel. The right sidebar shows the 'History' section, which is currently empty.

Importing data

The screenshot displays the Galaxy web interface. At the top, a navigation bar includes 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Admin', 'Help', and 'User'. The 'Shared Data' menu is highlighted with a red box and a red arrow pointing to it from the text 'Import shared data libraries'. On the left sidebar, the 'Tools' section has a search bar and a list of tool categories. A red box highlights the 'Upload data from local storage or from the cluster' icon, with a red arrow pointing to it from the text 'Upload data from local storage or from the cluster'. The main content area shows a 'Welcome to Galaxy on the Tufts cluster' message with a 'Bioinformatics @ Tufts' button and links for 'Galaxy UI', 'History', and 'Scratchbook'. The right sidebar shows 'History' and 'Unnamed history' sections. A blue information box at the bottom right of the history section states: 'This history is empty. You can load your own data or get data from an external source'. The page number '17' is located in the bottom right corner.

Access Galaxy

1. Connect to Tufts Network, either on campus or via [VPN](#)
2. Visit <https://galaxy.cluster.tufts.edu/>
3. Log in with you cluster username and password
4. In another browser window go to course workflow: <https://rbatorsky.github.io/intro-to-rnaseq-with-galaxy/>
5. Under Table of Contents click on **“Introduction and Setup”**

Suggested screen layout

The image displays two browser windows side-by-side. The left window shows the Galaxy Tufts dashboard at galaxy.cluster.tufts.edu. The dashboard features a top navigation bar with 'Galaxy Tufts' and various menu items like 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Admin', 'Help', and 'User'. A left sidebar lists tool categories such as 'Tools', 'Get Data', 'Send Data', 'Collection Operations', 'Expression Tools', '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', 'Mothur', 'PICRUST', 'Annotation', 'deepTools', and 'Picard'. The main content area displays a welcome message: 'Welcome to Galaxy on the Tufts University High Performance Compute Cluster!' with a 'Tufts Galaxy Support' button and links for an interactive tour, history, and scratchbook. Below this is a cityscape image. The right window shows a course workflow page titled 'intro-to-rnaseq-with-galaxy' from a Tufts University Research Technology Workshop. It includes a 'Description' section, a 'Goals' section listing workflow steps like 'Process Raw Reads (QC, adapter trimming)', 'Read Alignment', 'Gene Quantification', 'Differential Expression', and 'Functional Enrichment'. It also has a 'Materials Needed' section listing 'Chrome web browser', 'Account on Tufts High Performance Compute Cluster', and 'VPN if accessing the HPC from off campus'. A 'Table of Contents' section is visible at the bottom.