

RNA-seq to study HIV Infection in cells

Rebecca Batorsky
Pr Bioinformatics Scientist
Dec 2021

Research Technology Team



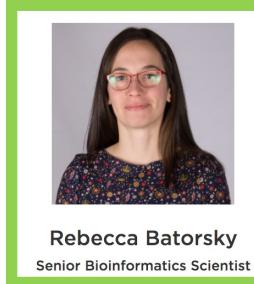
Delilah Maloney
High Performance Computing Specialist



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

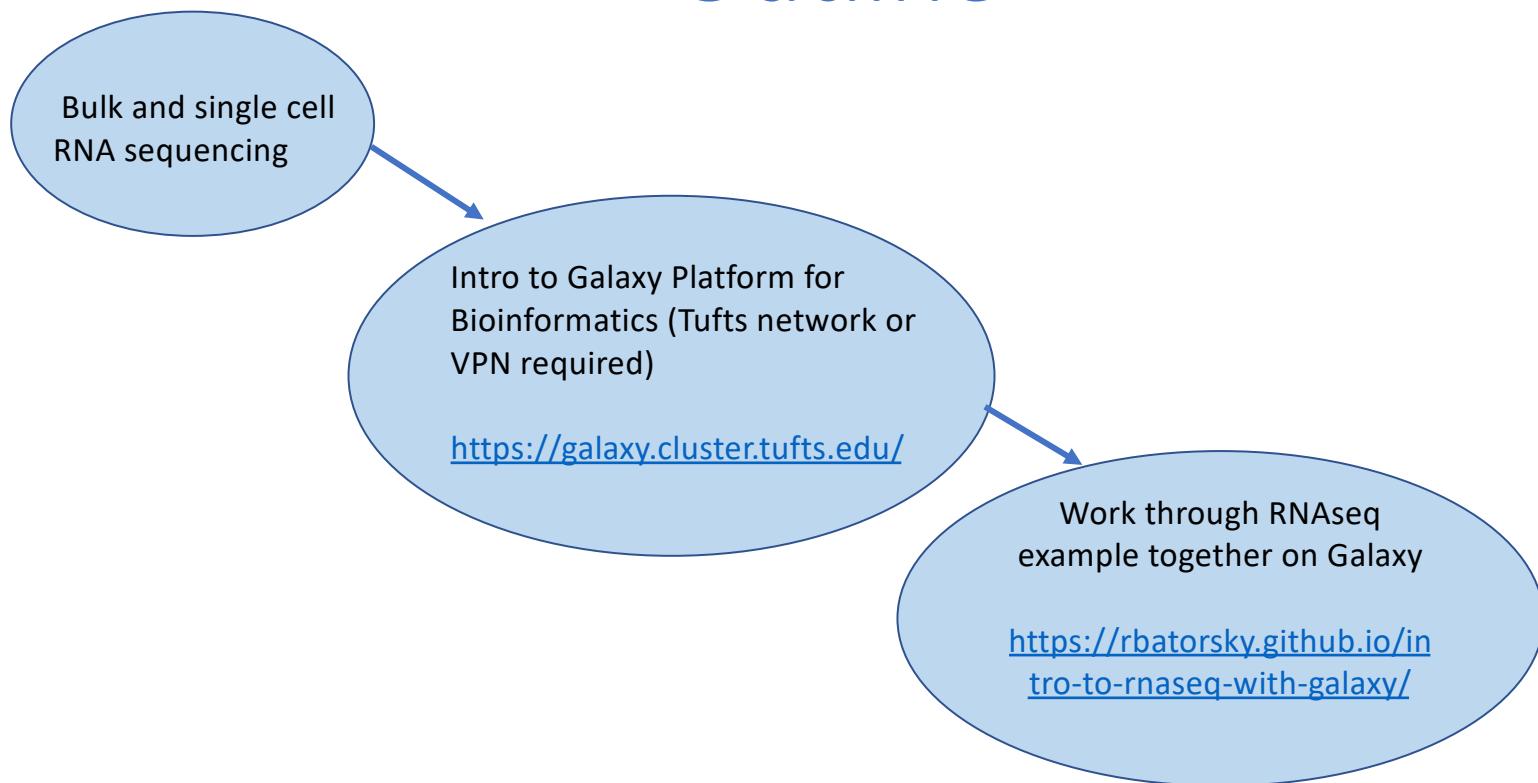


Uku-Kaspar Uustalu
Data Science Specialist

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

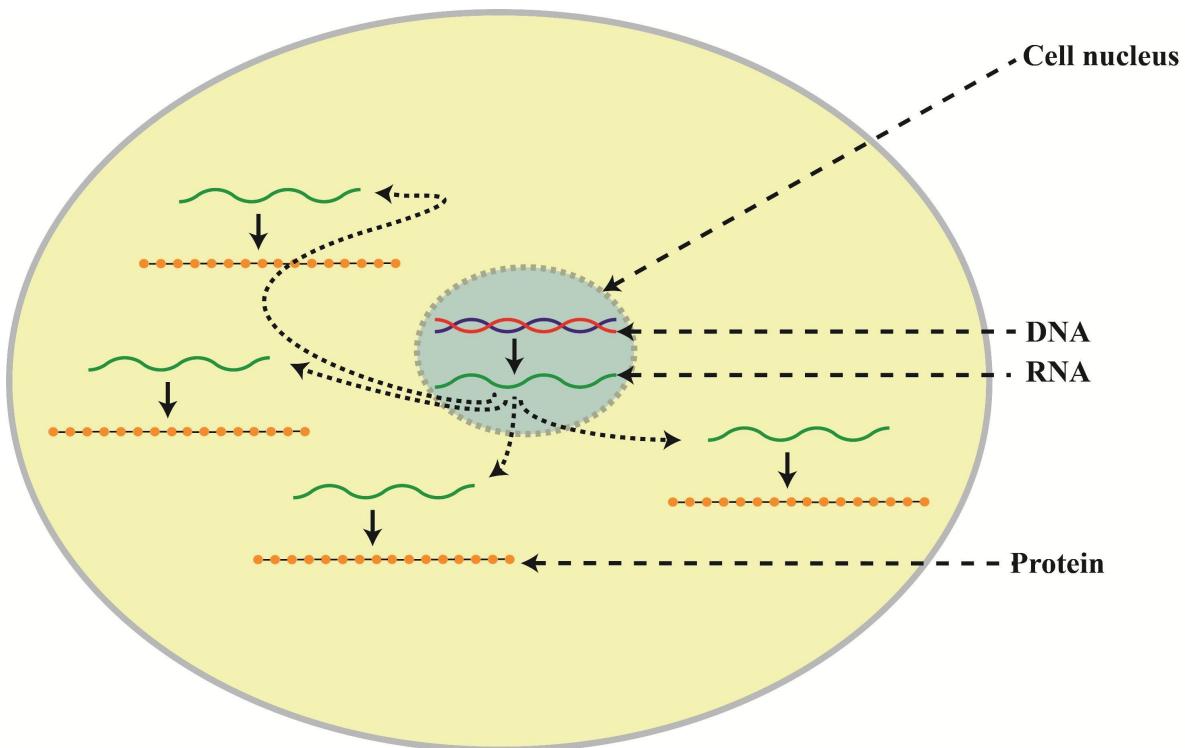
<https://it.tufts.edu/research-technology>

Outline



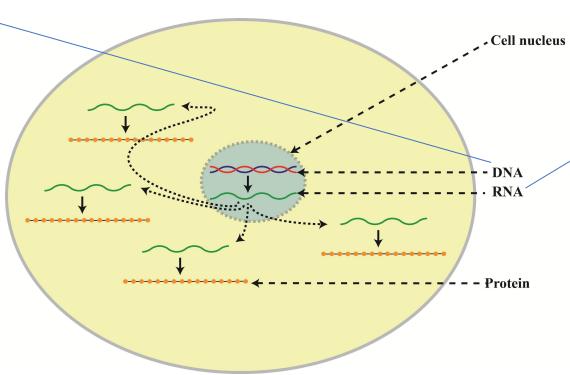
2 days!

DNA and RNA in a cell



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

Two common analyses

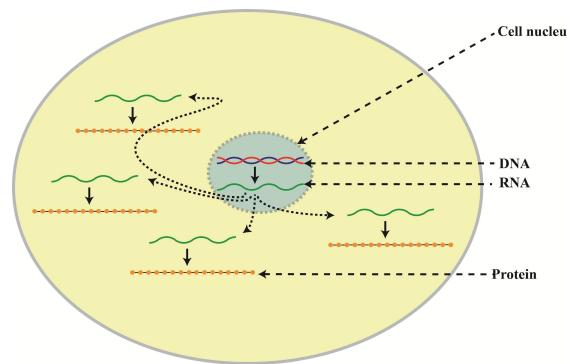
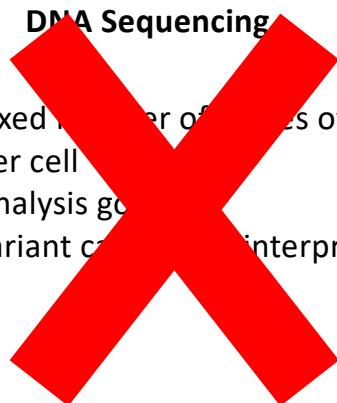
- 
- The diagram illustrates a cell with a yellow cytoplasm and a grey nucleus. Inside the nucleus, a red double helix represents DNA. A green wavy line represents RNA being transcribed from DNA. A blue wavy line represents proteins being translated from RNA. Arrows indicate the flow of information from DNA to RNA and RNA to proteins.
- Fixed number of copies of a gene per cell
 - Analysis goal:
Variant calling and interpretation
 - Number of copies of a gene transcript per cell depends on gene expression
 - Analysis goal:
 - Bulk : Differential expression
 - Single cell : Quantify different cell populations
- DNA Sequencing** ←
- **RNA Sequencing**

<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, 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

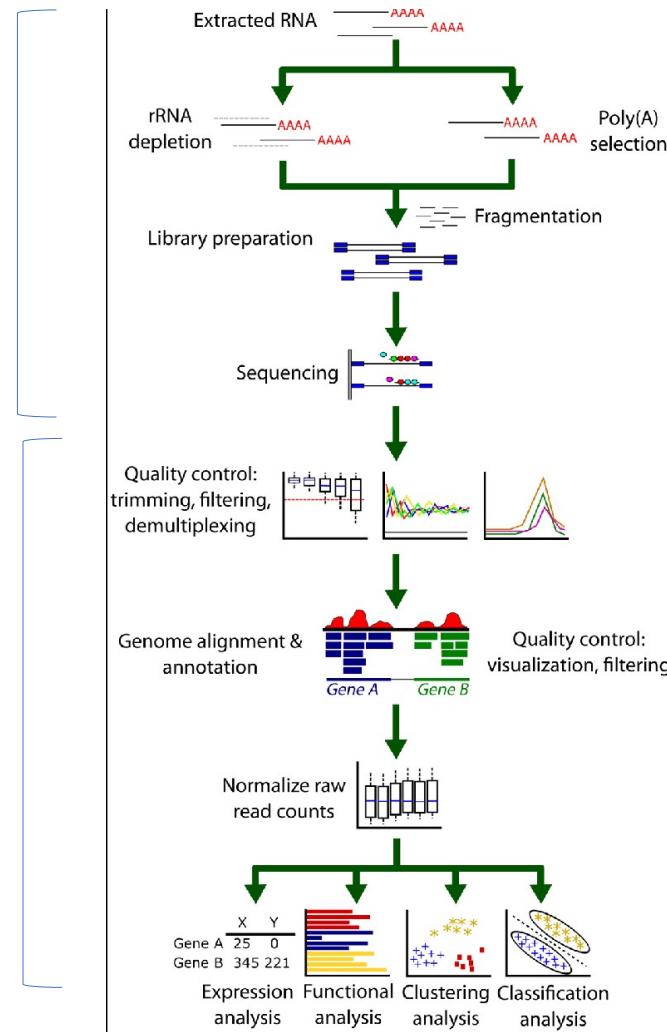
<https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg>

“Bulk” RNA seq workflow

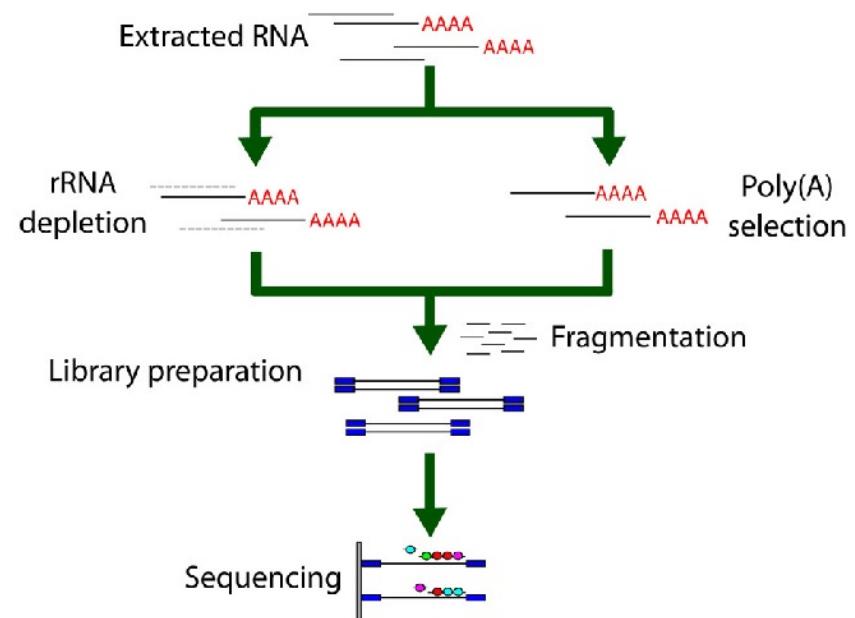
Library prep and sequencing

Bioinformatics

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



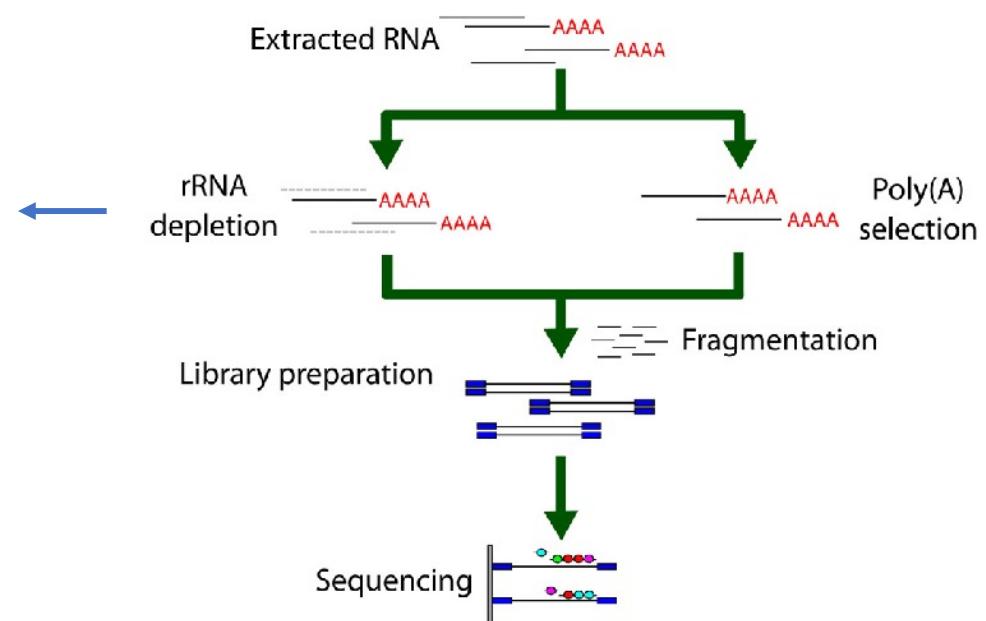
RNA seq library prep and sequencing



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

RNA seq library prep and sequencing

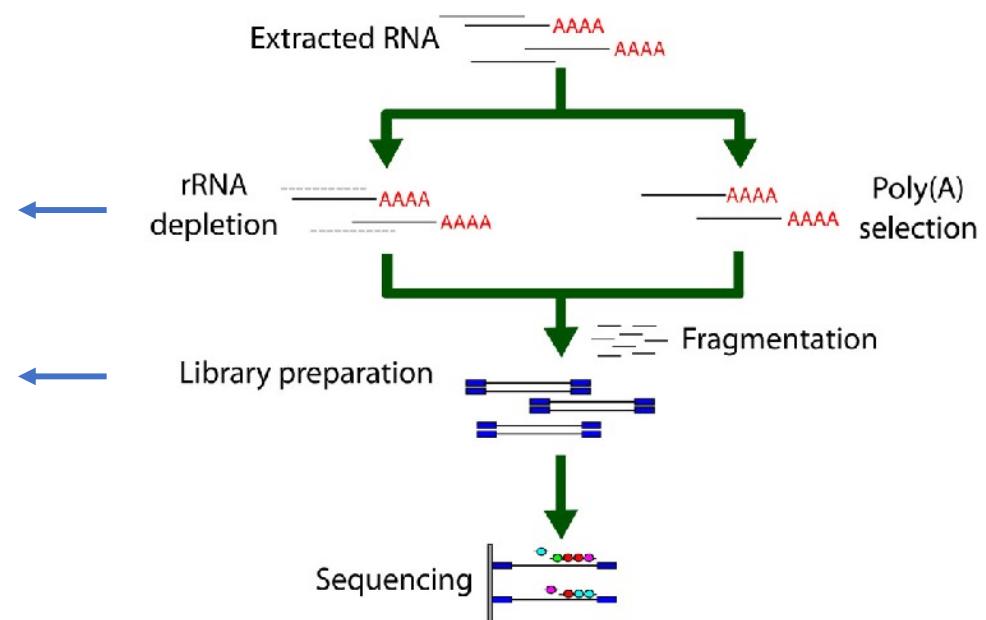
- Enrichment for mRNA, two options
- In humans, ~95%–98% of all RNA molecules are rRNAs



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



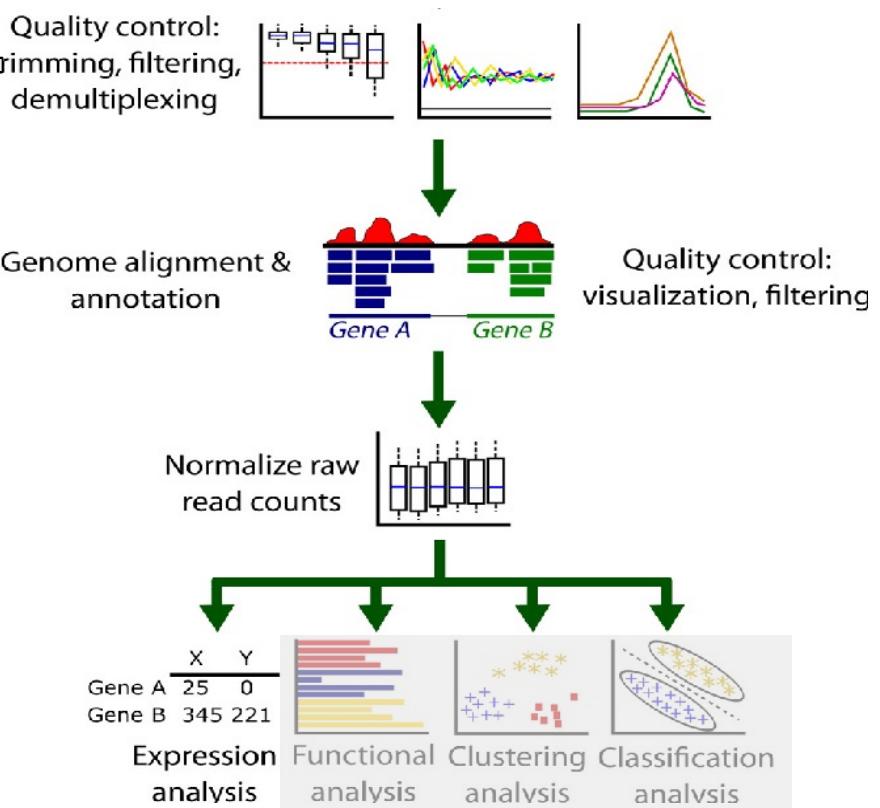
Resources:

[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”

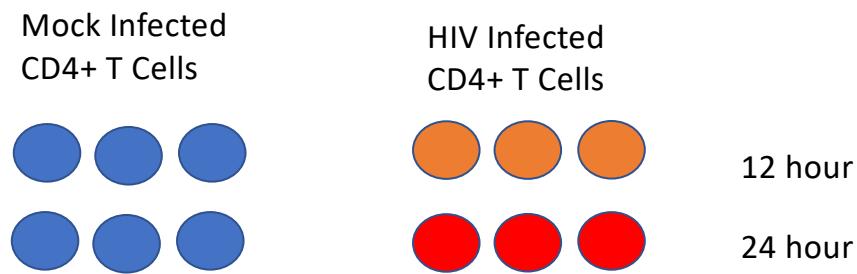


Oshlack et al. 2010. From RNA-seq reads to differential expression results. Genome Biology 2010, 11:220

Our dataset

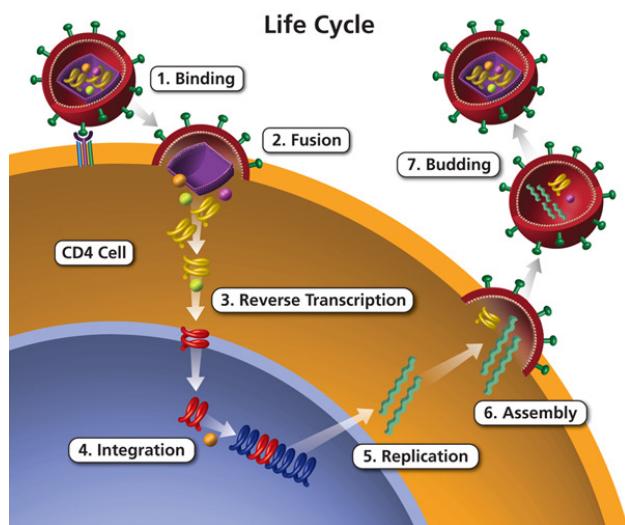
Next-Generation Sequencing Reveals HIV-1-Mediated Suppression of T Cell Activation and RNA Processing and Regulation of Noncoding RNA Expression in a CD4⁺ T Cell Line

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



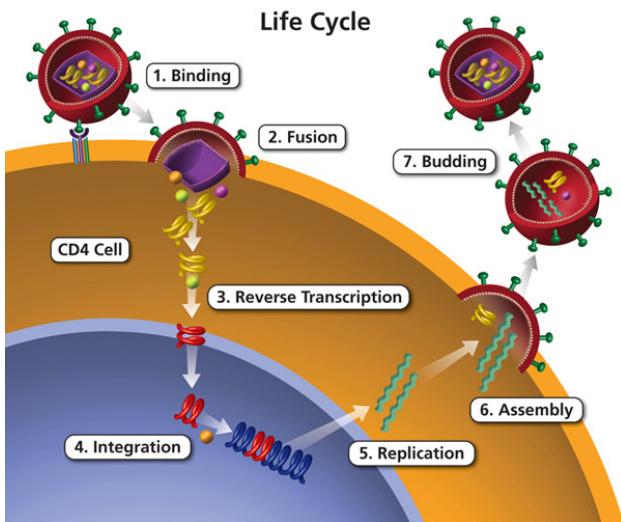
<https://www.ncbi.nlm.nih.gov/pubmed/21933919>

HIV lifecycle

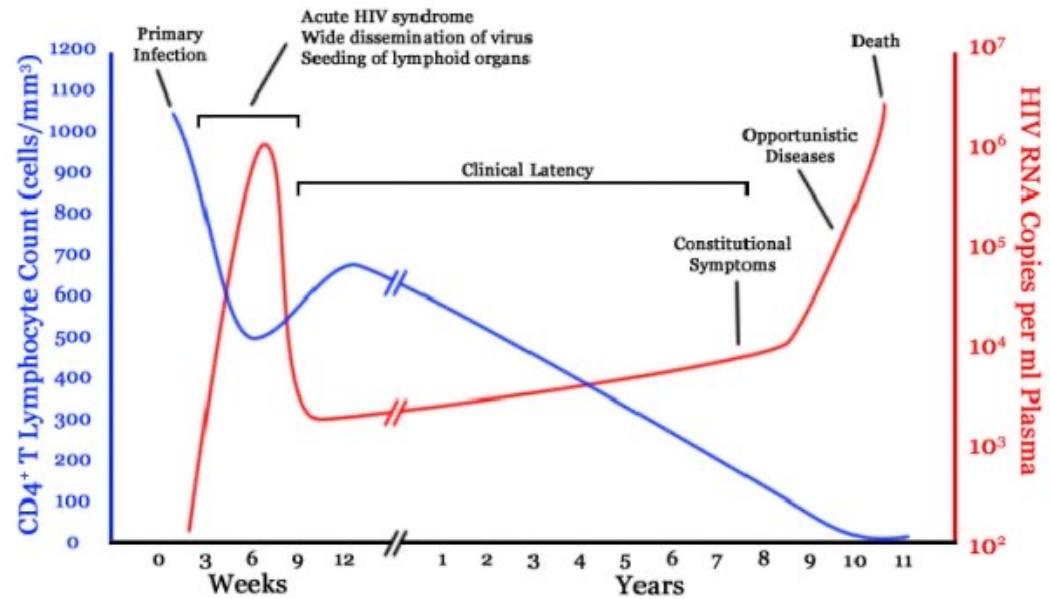


<https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle>

HIV lifecycle



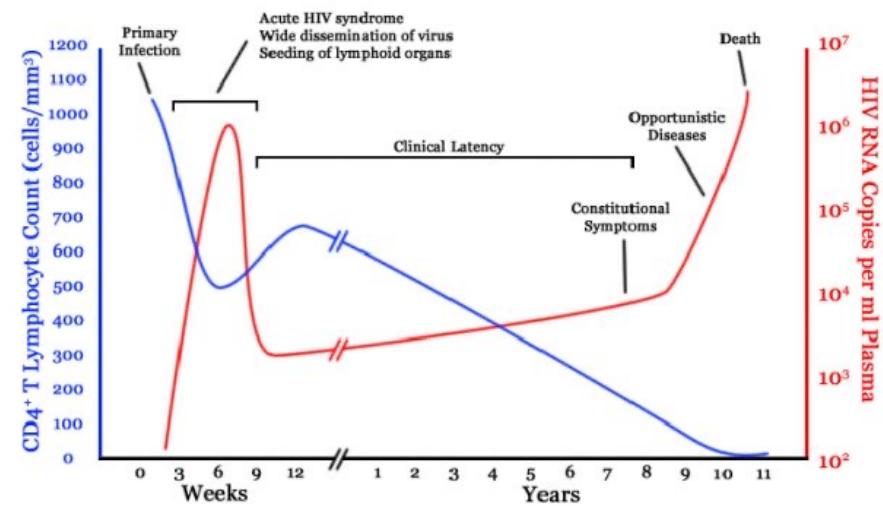
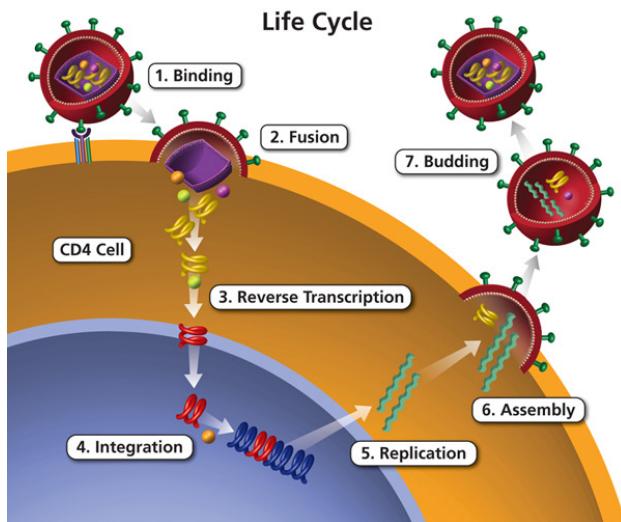
HIV infection in a human host



<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?

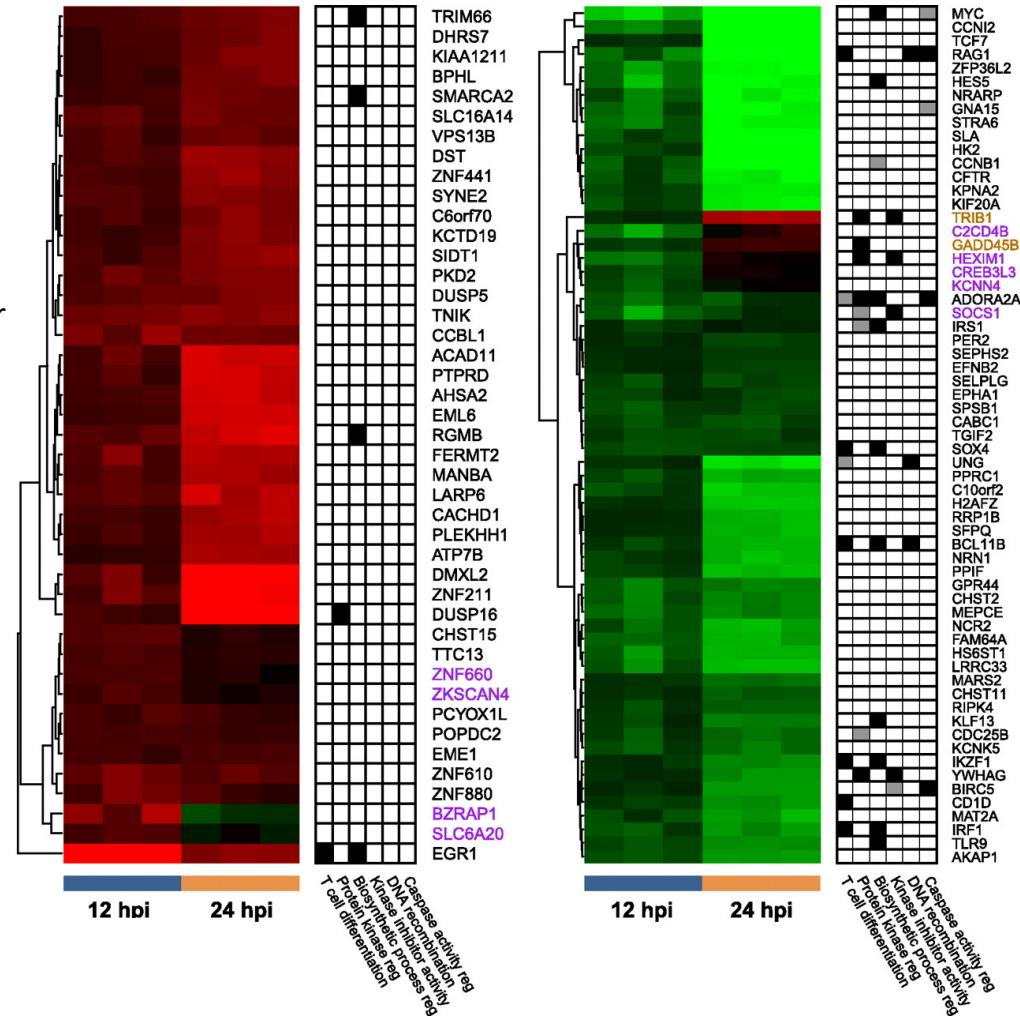
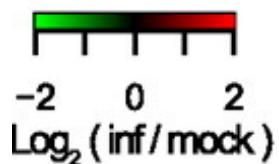


<https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle>

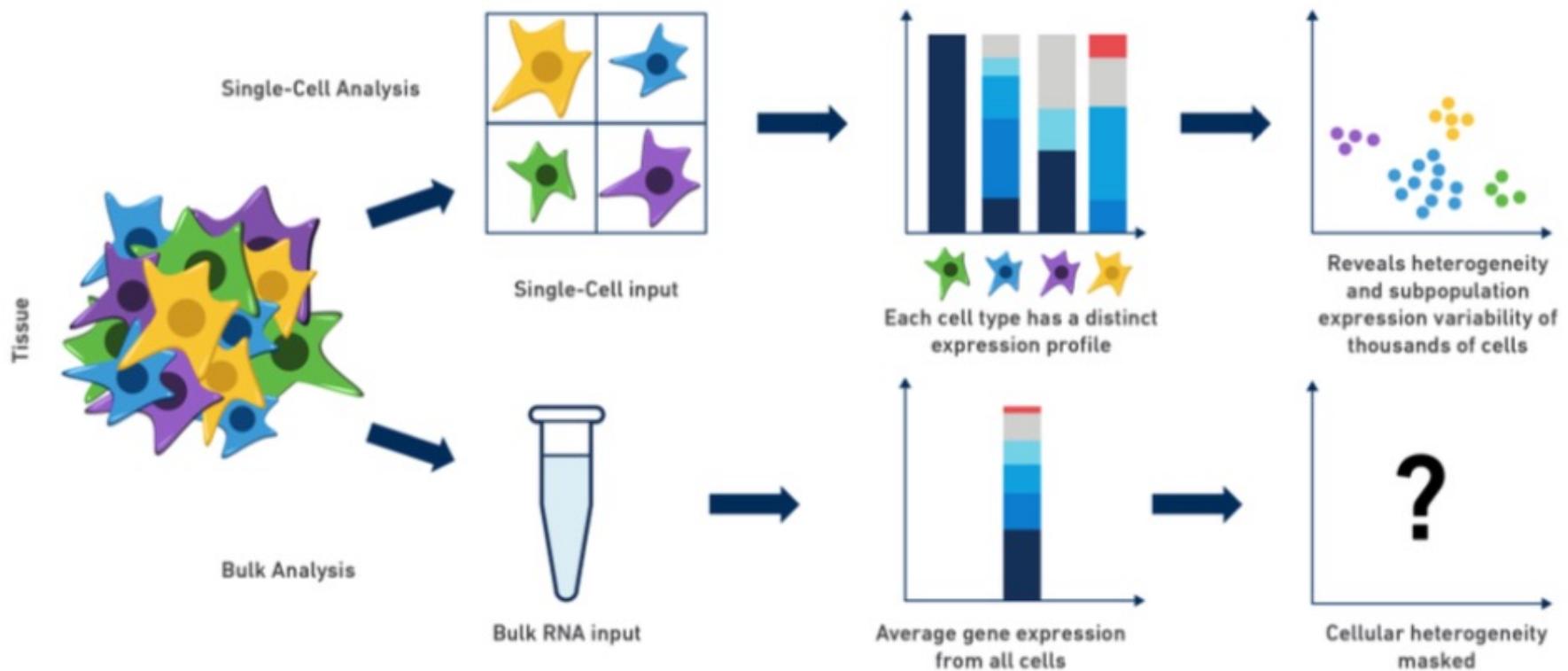
Study findings

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

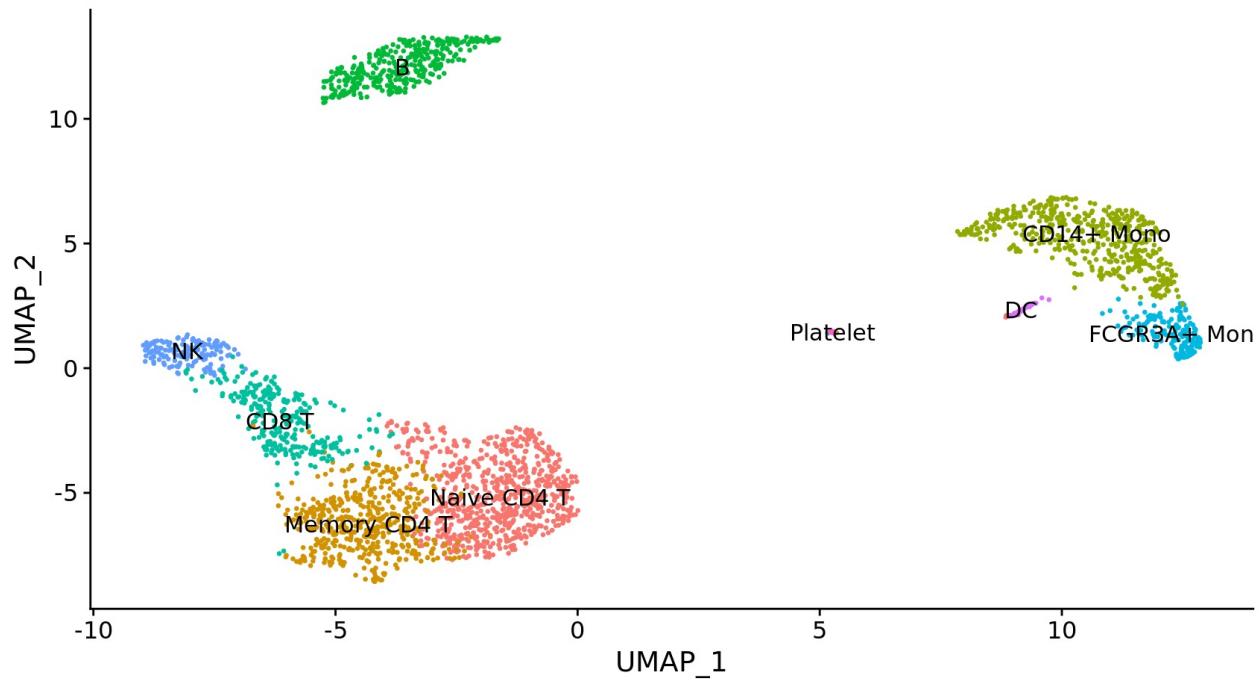


Bulk vs Single Cell RNA Sequencing



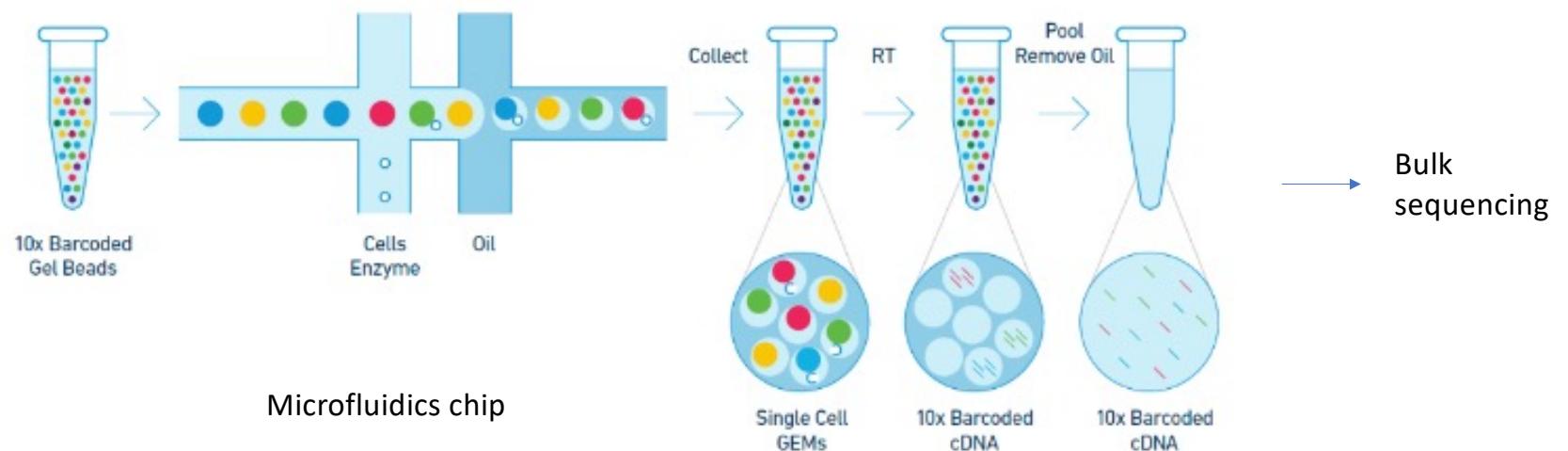
<https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started>

scRNA cell subsets in PBMC



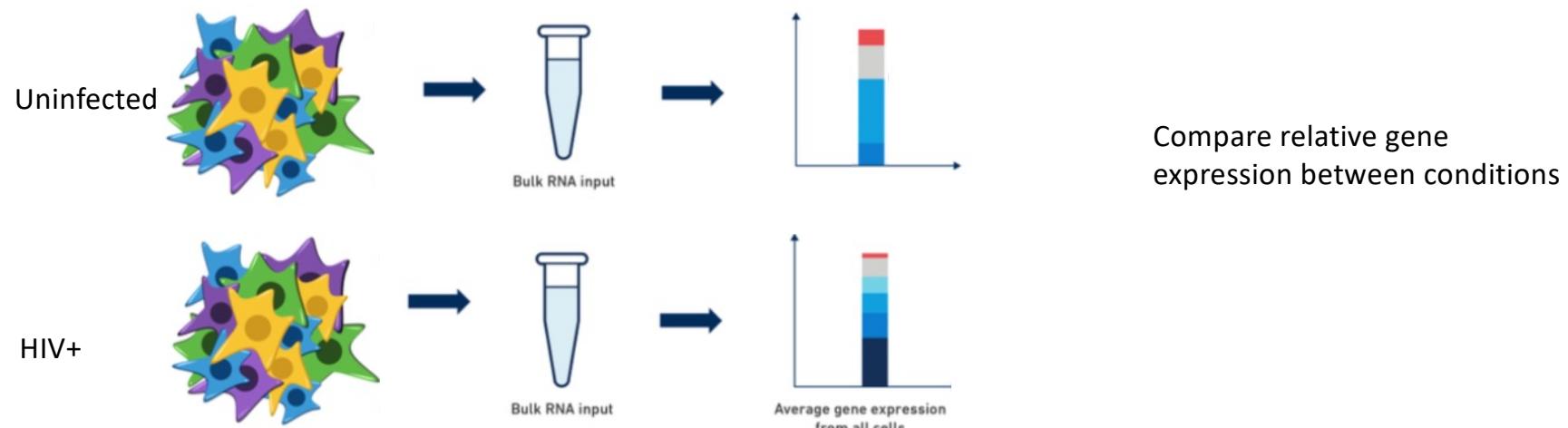
https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html

10x single cell technology

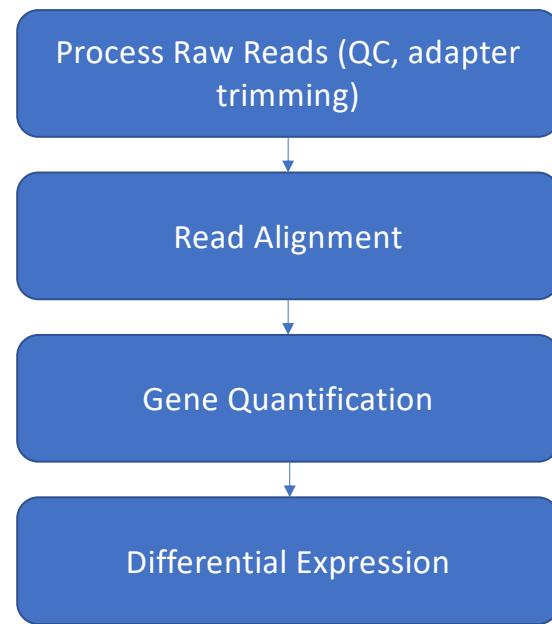


<https://github.com/hbctraining/scRNA-seq>

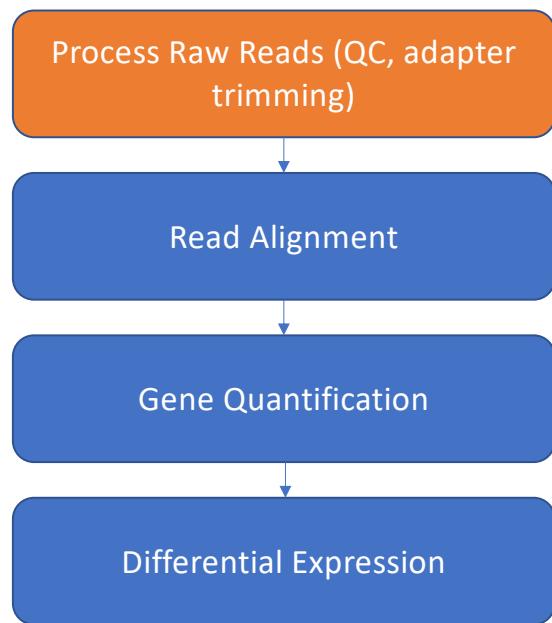
Bulk RNAseq for Differential Expression is OK!



Our (bulk) RNAseq Workflow



Quality control on Raw Reads



Raw reads in Fastq format

```
@SRR098401.109756285
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

<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
CAGATGGCCGAGAGGAAGCCATGAAGGCCCTGCATGGGGAGATCGGAAGAGCGGTTACGCAGGAATGCCGAGAC
+
IIIIIIHIIIIIBIIDII>IIDHIIHIDIIGIFIIIEGIBDFIG<EIEGEEG;<DB@A8CC7<><C@BBDBB
@SRR497699.11626500.1 HWI-EAS39X_10175_FC61MK0_4_44_8384_16550 length=75
CGTACTGAACGTACAACGCTGATGCCATCCGCATATTAAATTCCGGCAGCGTTAACCTCCGTACCTCGCG
+
HHHHHHHHHHHHFHHGHHHHHHB@HHHHHHHHFHHHHHEHHHHHHHHHHGEHDHHEHHHHBHHGHHHHHHHG
@SRR497699.29057557.1 HWI-EAS39X_10175_FC61MK0_4_112_12508_19308 length=75
CCGAGGCTTAGCTTCATTACTGCTCCCAGGGTGTGCTGTCAAAGAGATAAGATCGGAAGAGCGGTTCAG
+
GGGBGGGDGBHHDHGEGGGHHHHGHHGHHHHGBGGDGGEGDHHHHHHHHHH@BHHGGHGHHHHEEGHH
@SRR497699.1331889.1 HWI-EAS39X_10175_FC61MK0_4_5_4738_15920 length=75
CTTACTTTGTAGCCTCATCAGGGTTGCTGAAGATGGCGGTATATAGGCTGAGCAAGAGGTGGTGAGGTTGATC
+
HHHHHHHHHHGGGGHHHGEBEEGGEDGGGGGGHHHHHGEGBDGGDDGBGGC<EADBEBE<GGGBEEDGD
...
...
```

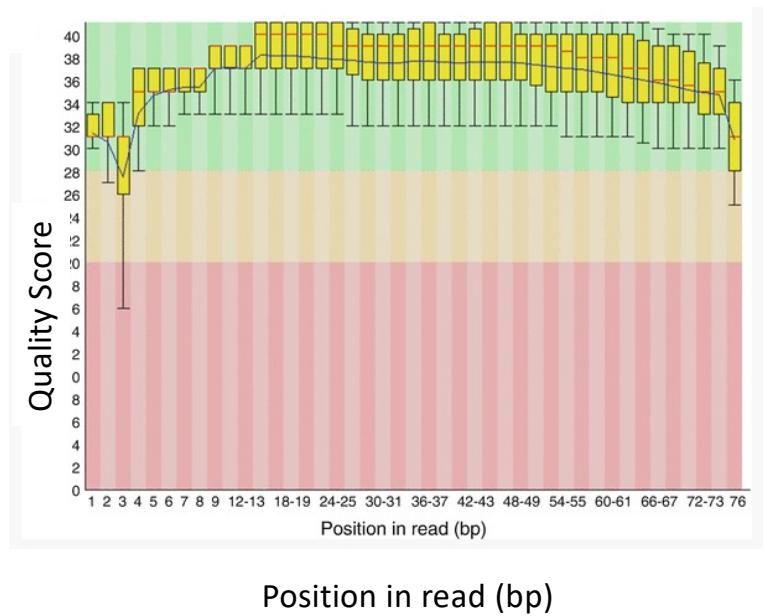
FastQC Tool



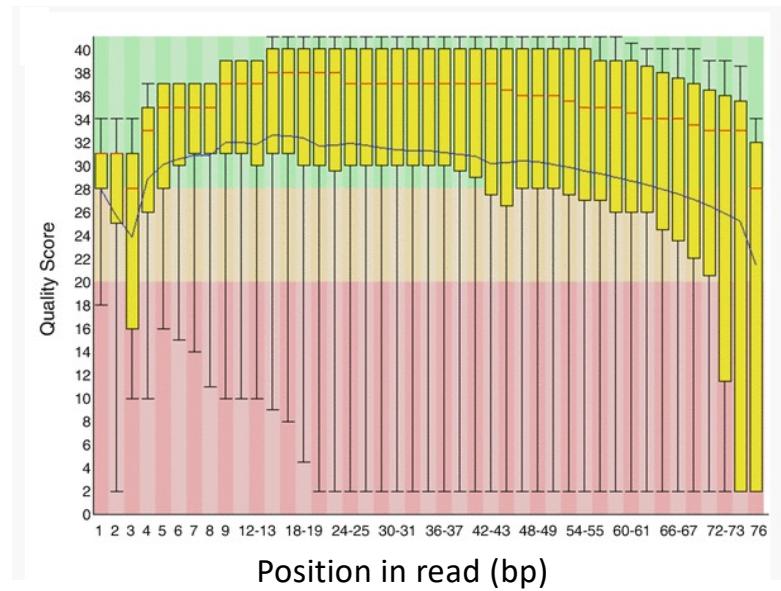
Metrics

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

FastQC: Sequence Quality Histogram



GOOD
High quality over the length of the read

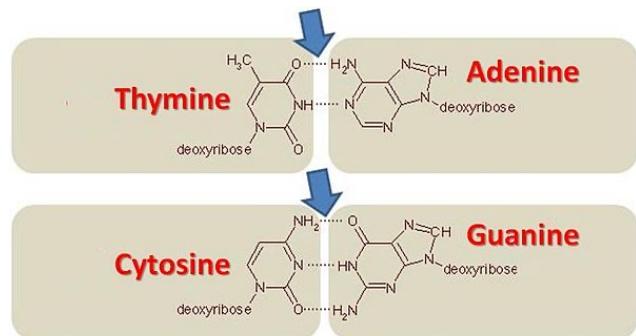
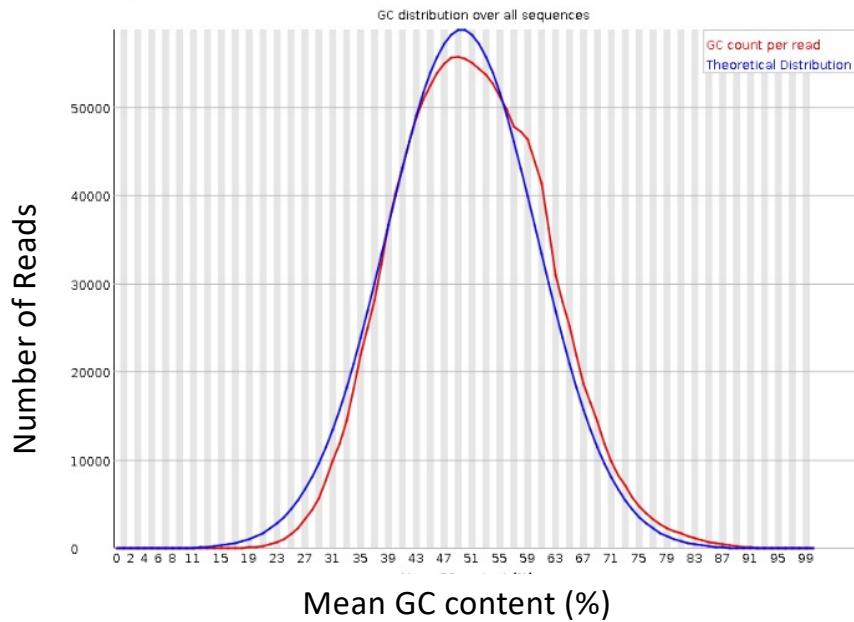


BAD
Read quality drops at the beginning and end



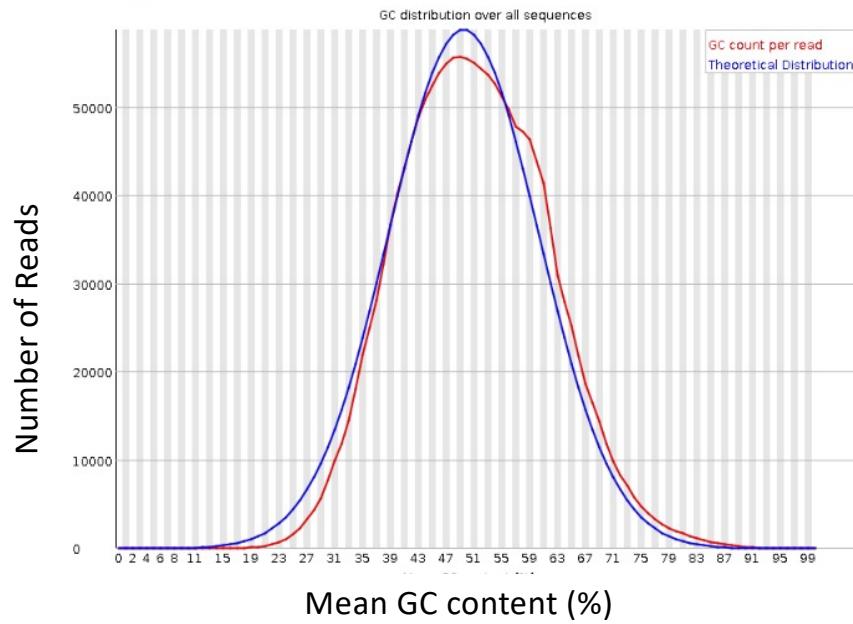
FastQC: Per sequence GC content

Per sequence GC content



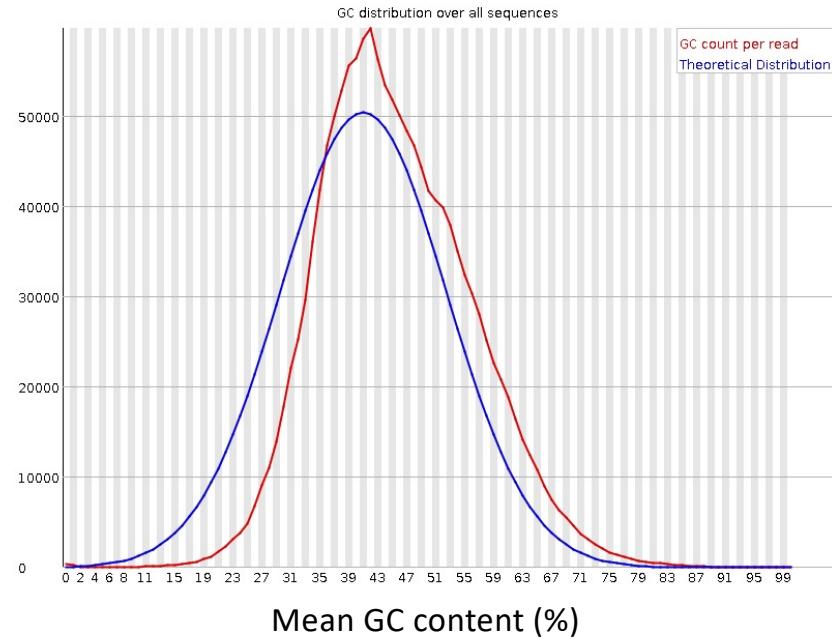
FastQC: Per sequence GC content

✓ Per sequence GC content



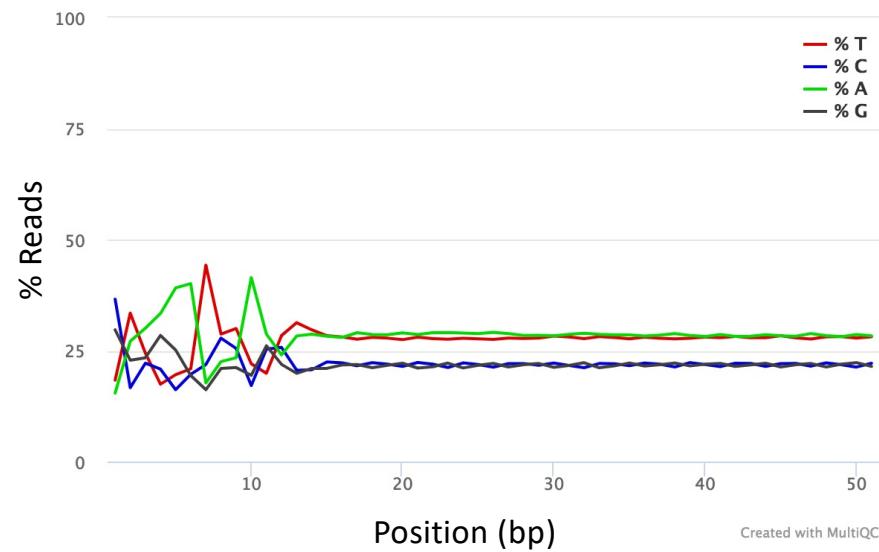
GOOD: follows normal distribution (sum of deviations is < 15% of reads)

✗ Per sequence GC content



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

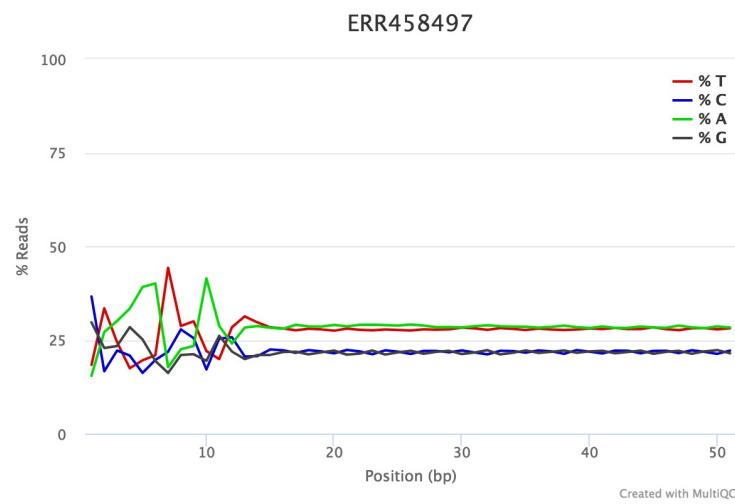
FastQC: Per Base Sequence Content



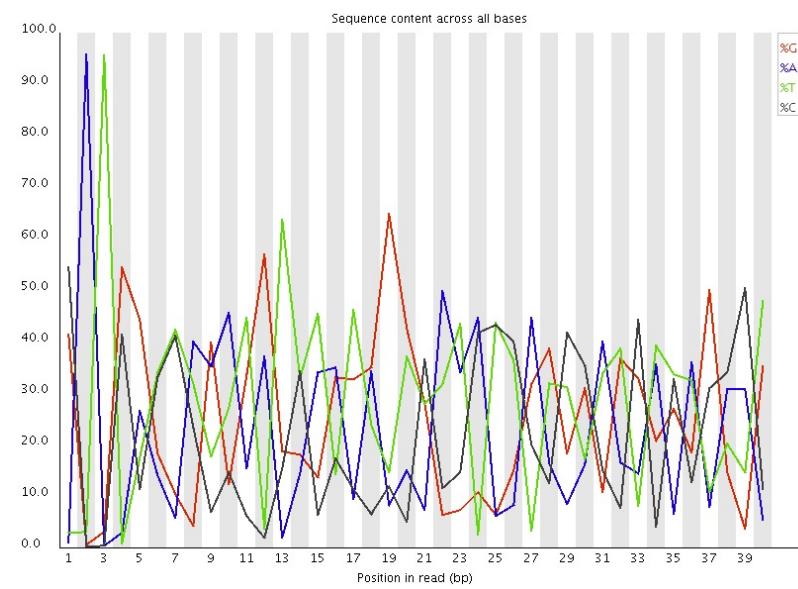
- Proportion of each position for which each DNA base has been called
- RNAseq data tends to show a positional sequence bias in the first ~12 bases
- The "random" priming step during library construction is not truly random and certain hexamers are more prevalent than others

sequencing.qcfail.com

FastQC: Per Base Sequence Content



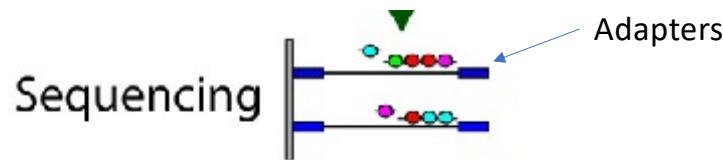
EXPECTED for RNAseq



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

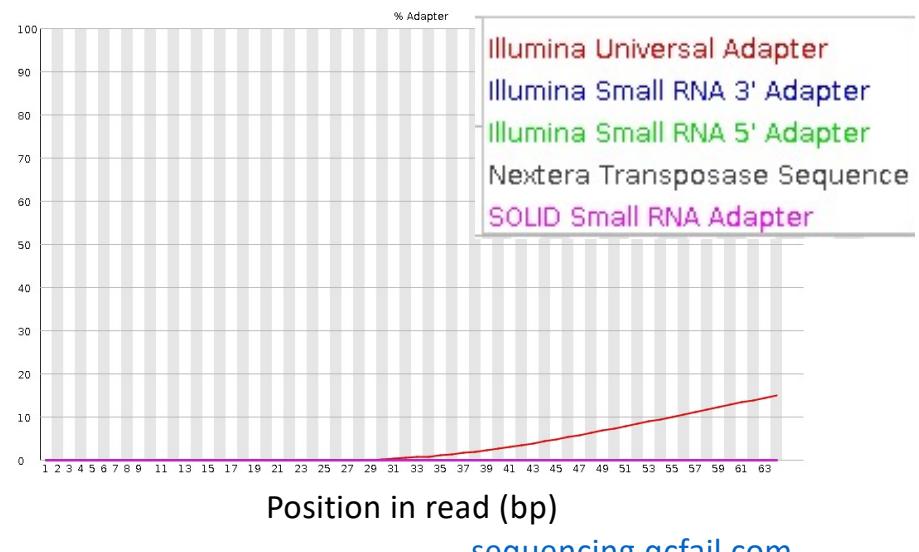
FastQC: Adapter content



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

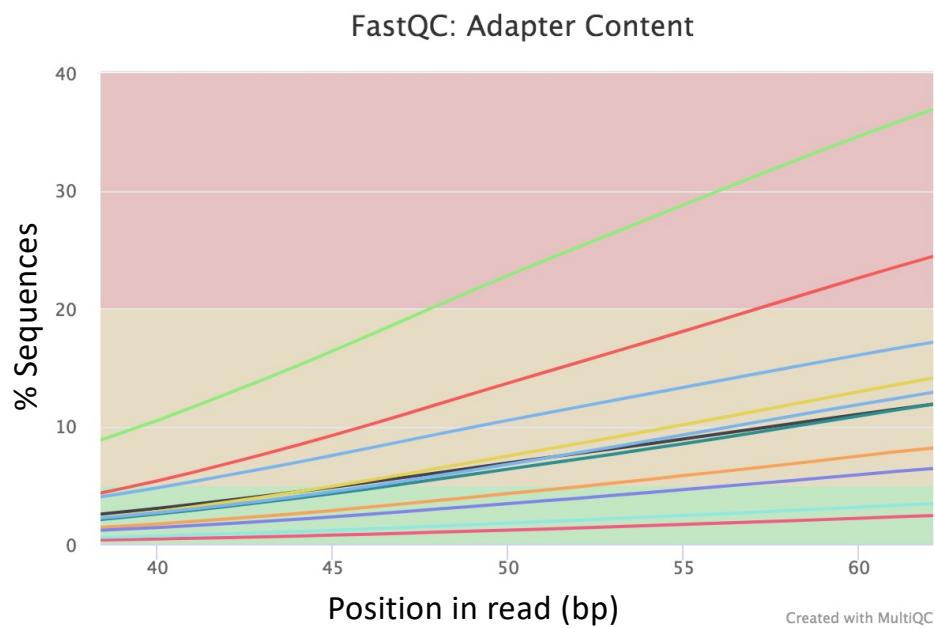
The plot shows that the adapter content rises over the course of the read

Solution – Adapter trimming!



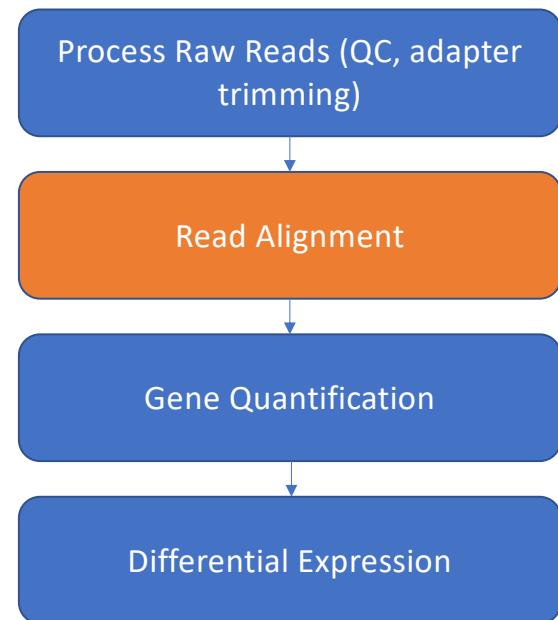
FastQC -> MultiQC

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



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

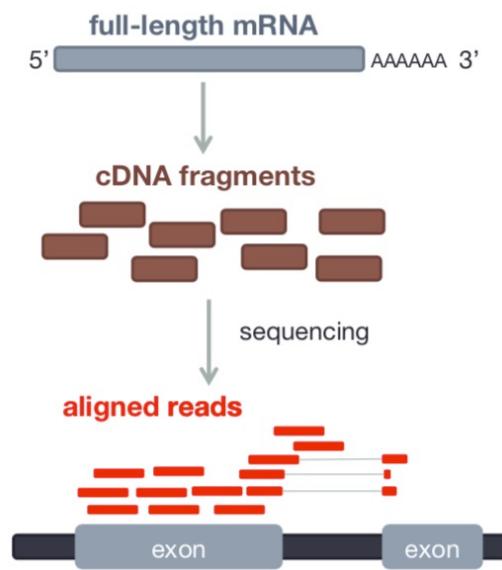
Workflow



Read Alignment

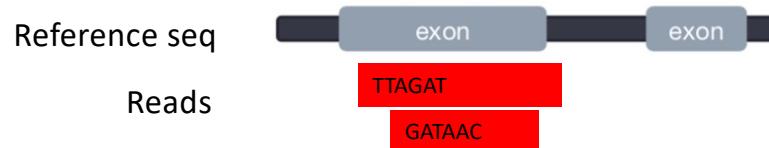
- RNAseq data originates from spliced mRNA (no introns)
- When aligning to the genome, our aligner must find a spliced alignment for reads
- We use a tool called STAR (Spliced Transcripts Alignment to a Reference) that has a exon-aware mapping algorithm.

Reference sequence



[Dobin et al Bioinformatics 2013](#)

Sequence Alignment Map (SAM)



QHD VN:1.5 SO:coordinate	Header section
CSQ 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	

↑ 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

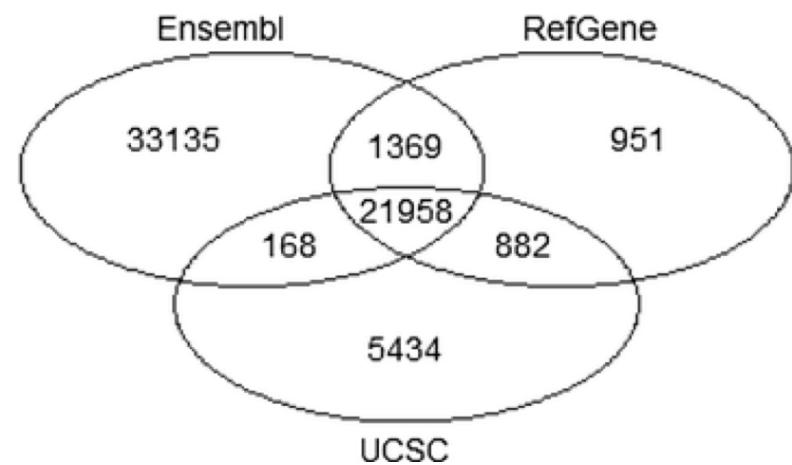
www.samformat.info

Sequence Alignment Map (SAM)



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](#)

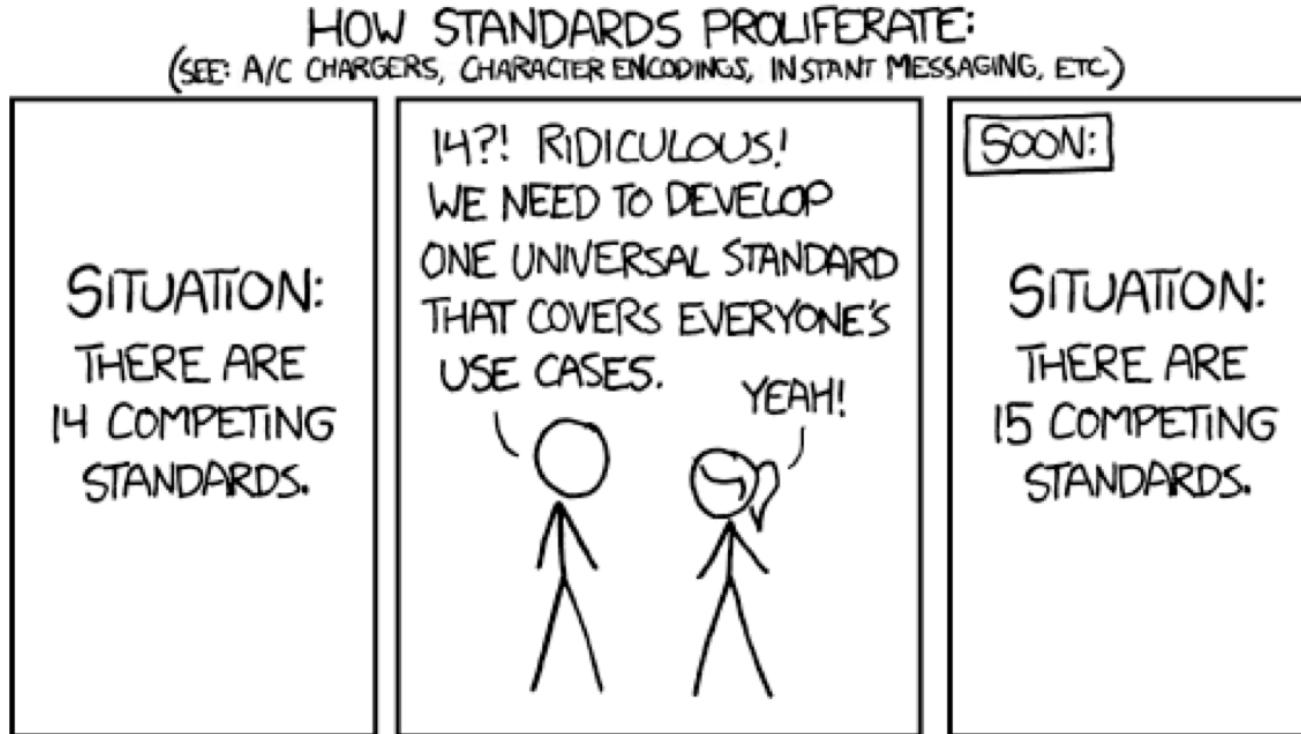
Gene Annotation Format (GTF)

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)	Frame	Strand	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

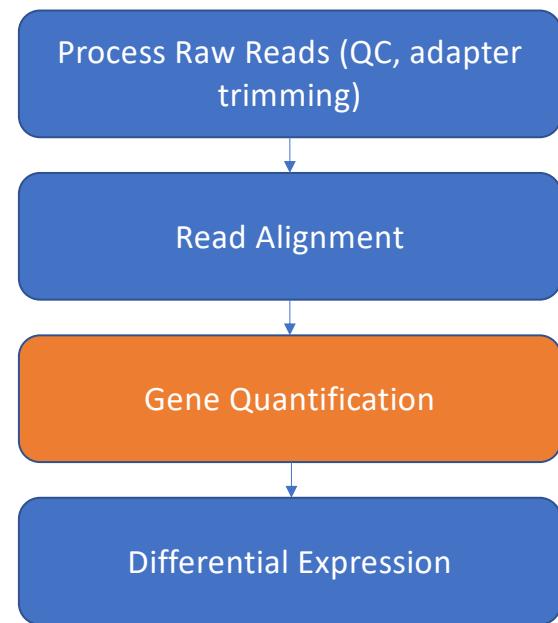


<https://xkcd.com/927/>

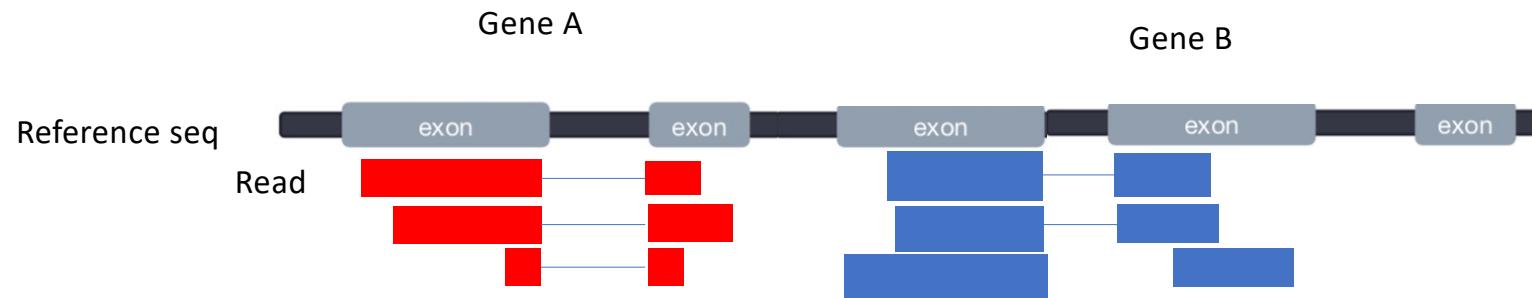
Visualizing reads with JBrowse



Workflow

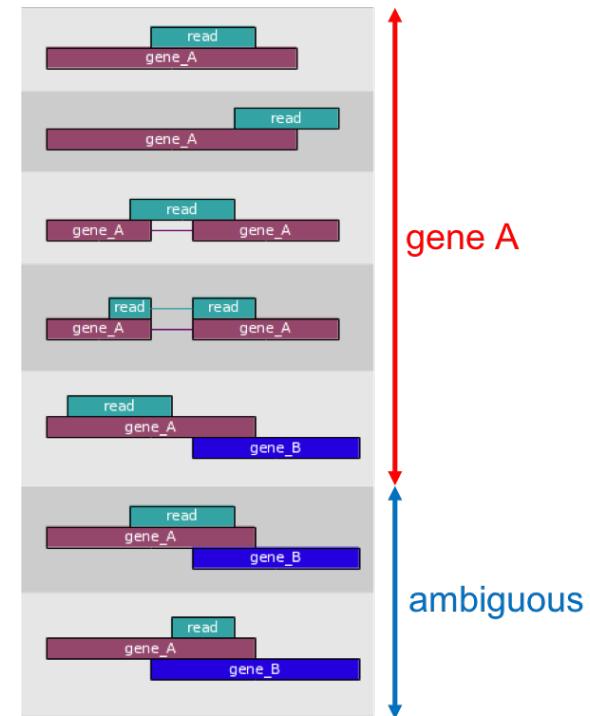


Counting reads for each gene



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

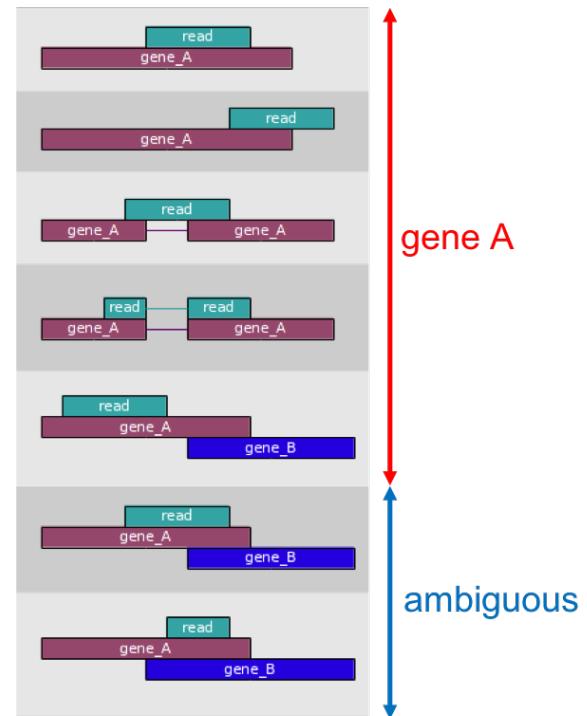


Counting reads: featurecounts

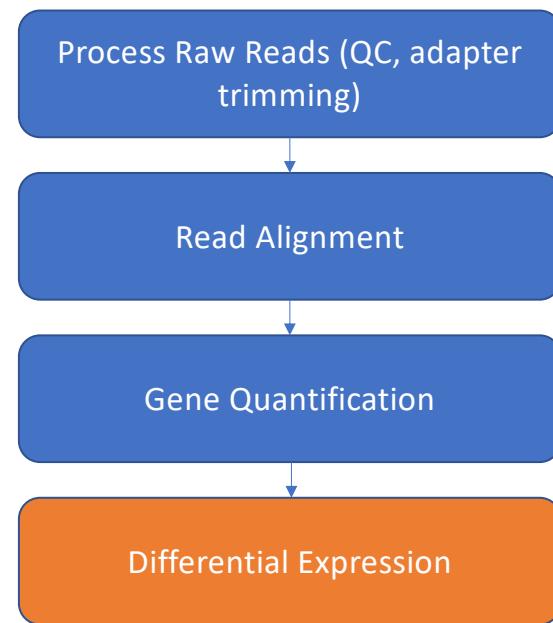
- The mapped coordinates of each read are compared with the features in the GTF file
- Reads that overlap with a gene by ≥ 1 bp are counted as belonging to that feature
- Ambiguous reads will be discarded

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



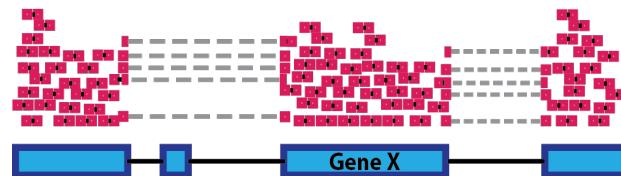
Workflow



Normalization

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

Sample A Reads



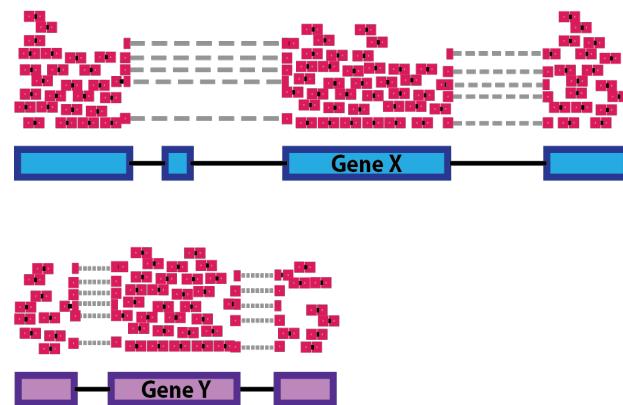
Adapted from https://hbctraining.github.io/DGE_workshop

Normalization

The number of reads mapped to a gene depends on

- **Gene Length**

Sample A Reads

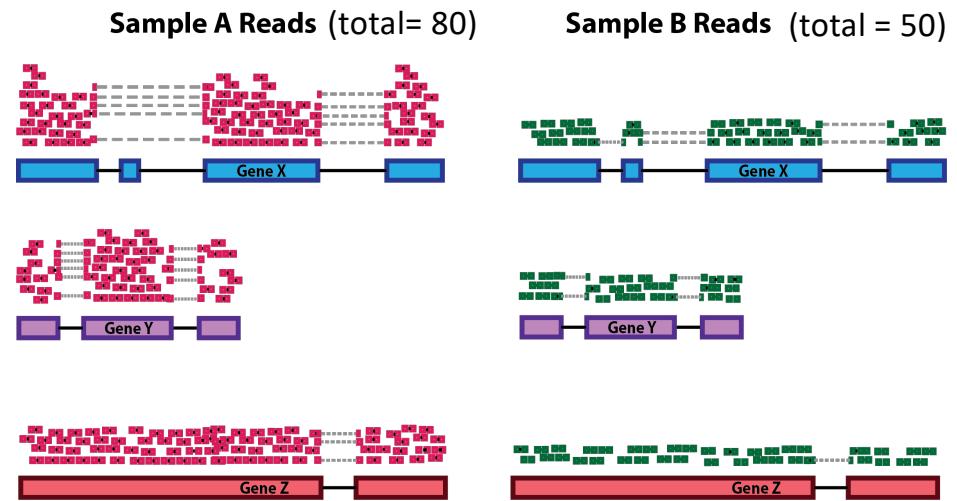


Adapted from https://hbctraining.github.io/DGE_workshop

Normalization

The number of reads mapped to a gene depends on

- Gene Length
- **Sequencing depth**

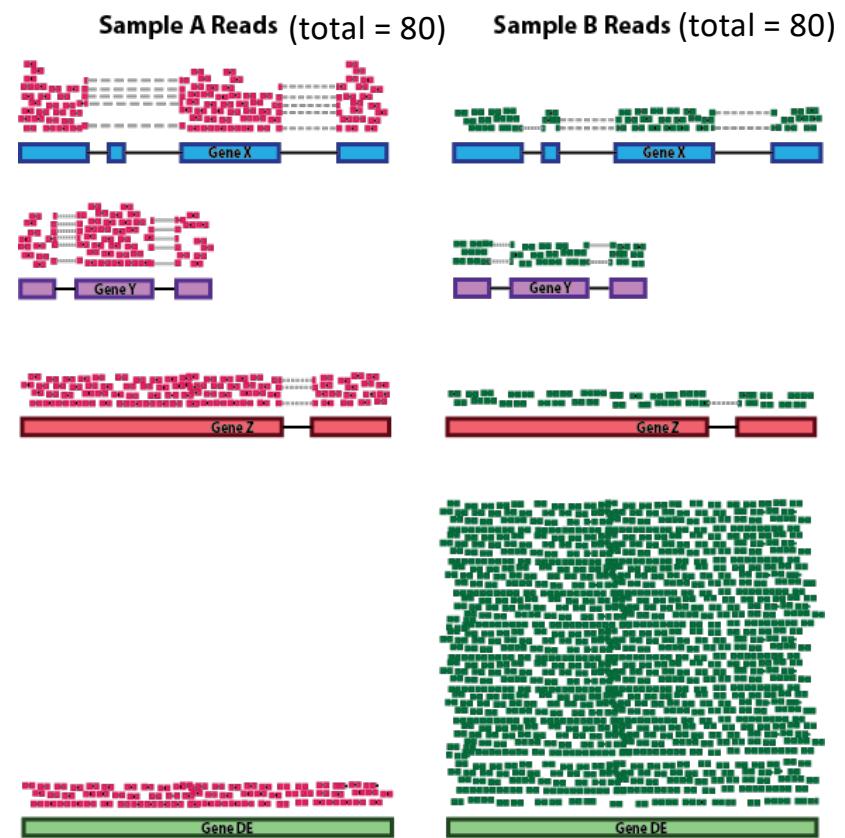


Adapted from https://hbctraining.github.io/DGE_workshop

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)**



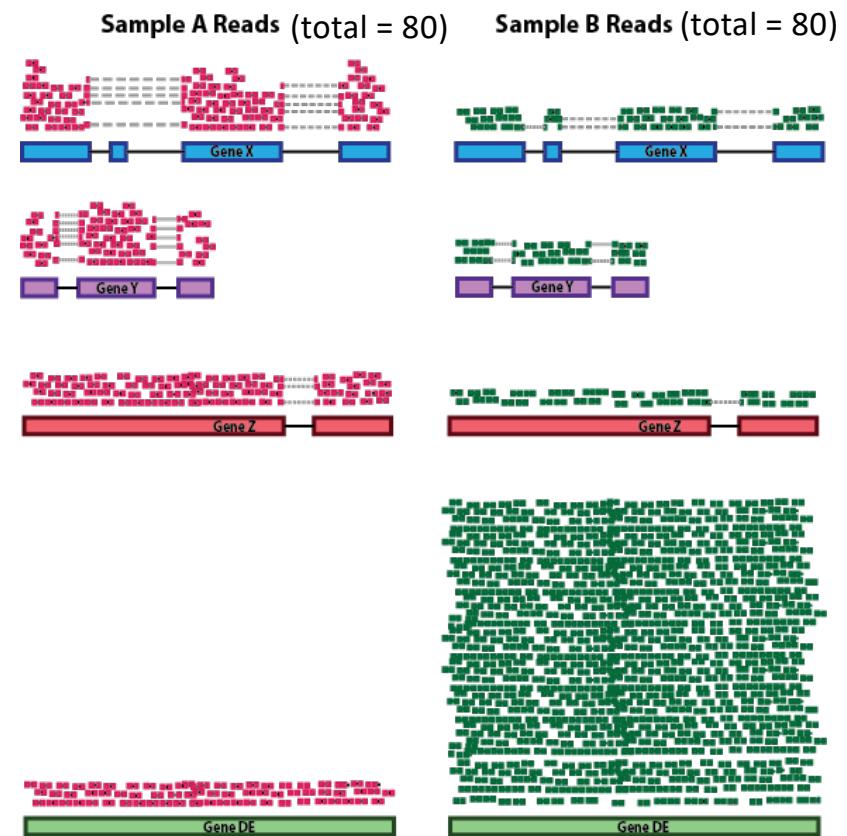
Adapted from https://hbctraining.github.io/DGE_workshop

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)

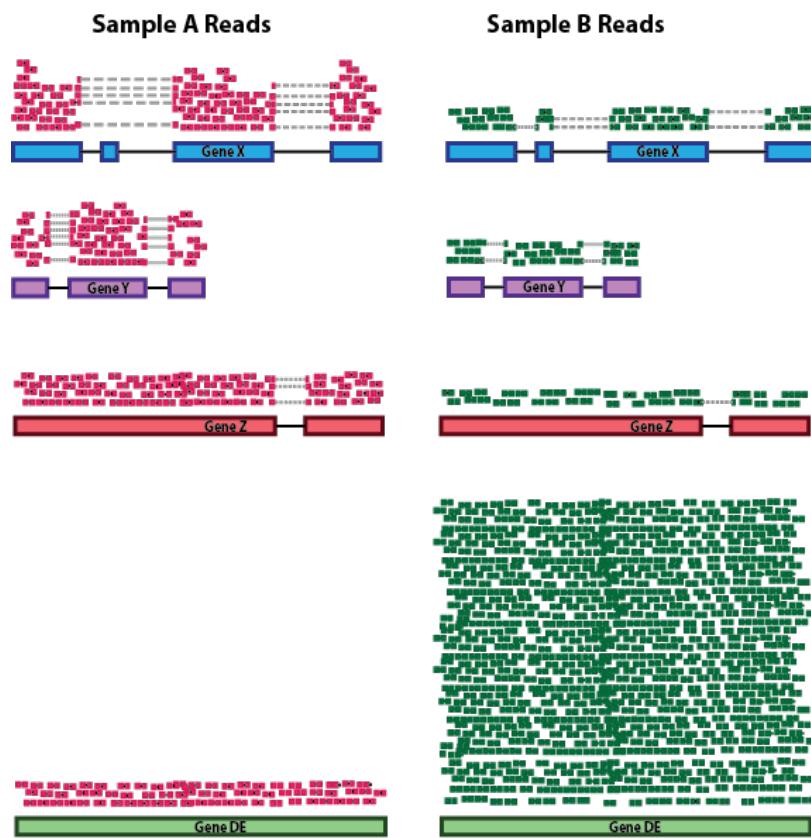
DESeq2 Median of Ratios



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	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) $\sqrt[n]{x_1 x_2 \cdots x_n}$

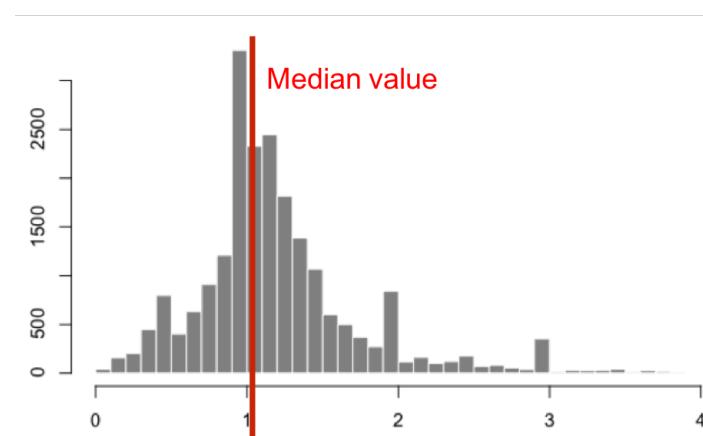
Gene	Sample A	Sample B	Avg. Sample
X	26	10	16
Y	26	10	16
Z	26	10	16
DE	2	50	16

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**)

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

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!

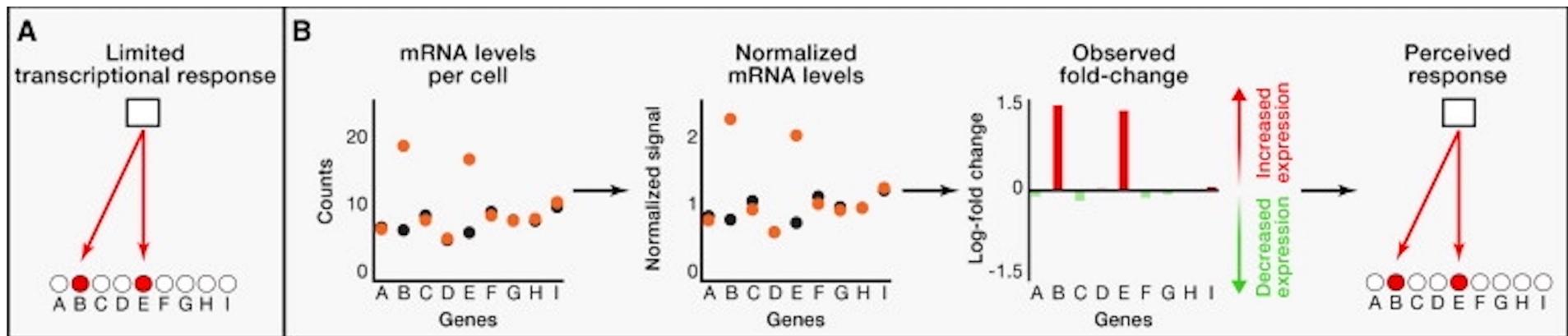
3. Take the **median** of each column. Should be ~1 for all

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

`estimateSizeFactors(dds)`

Assumption of DESeq2 Median of Ratios

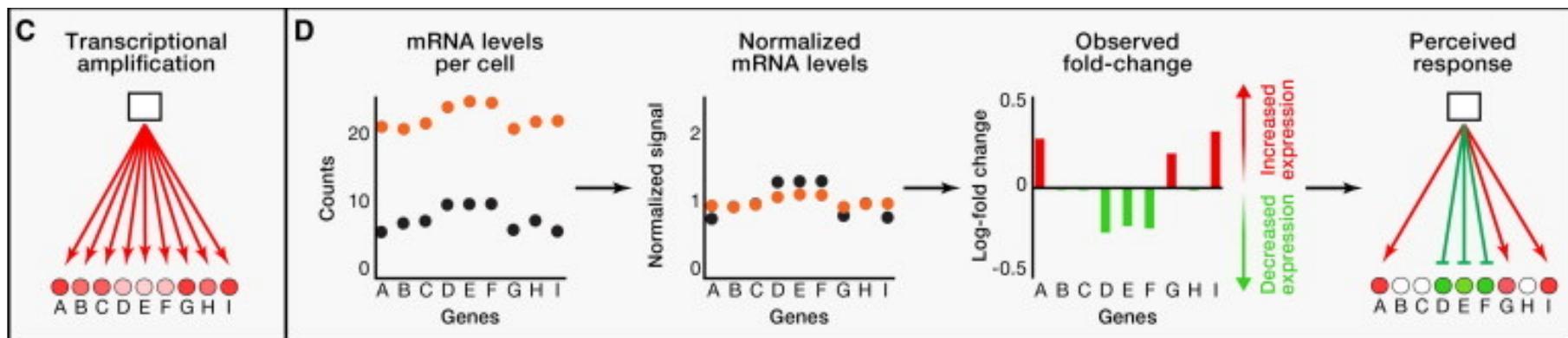
Median of Ratios method assumes that most genes are not Differentially Expressed between samples.



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>

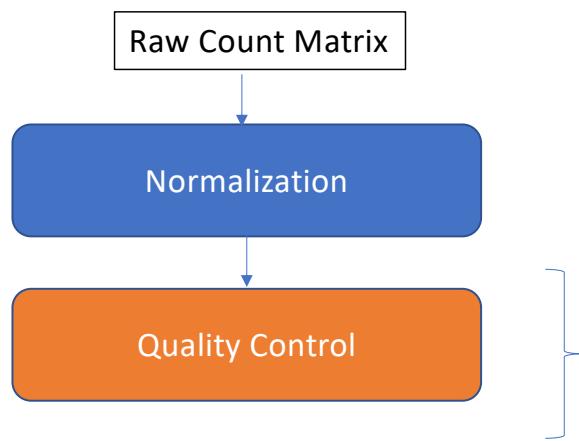
Normalization methods

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

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

Quality Control Visualizations



Examine sources of variation in the data

- Principal Component Analysis
- Hierarchical Clustering

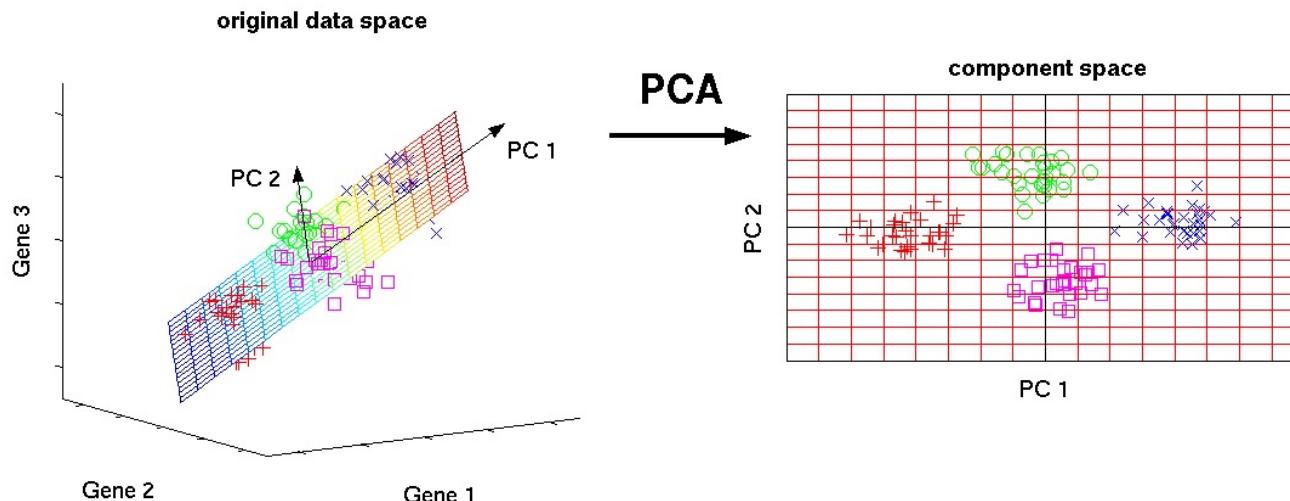
(Log2 + 1) Transformed, Normalized Count Table

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

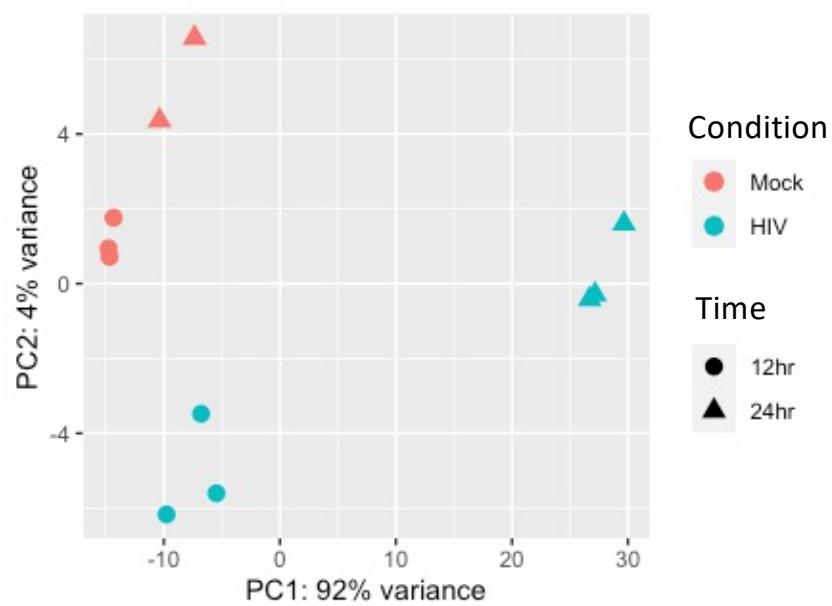
Gene	Mock_12h	Mock_12h	Mock_24h	Mock_24h	HIV_12h	HIV_12h	HIV_24h	HIV_24h
Gene 1	8.9	8.9	8.9	9.0	8.9	8.9	9.0	6.8
Gene 2	0.6	-1.0	0.6	-1.0	0.6	-1.0	0.6	3.8
Gene 3	4.1	11.9	4.1	-0.5	4.1	8.7	4.0	4.4



Do your samples cluster as expected?

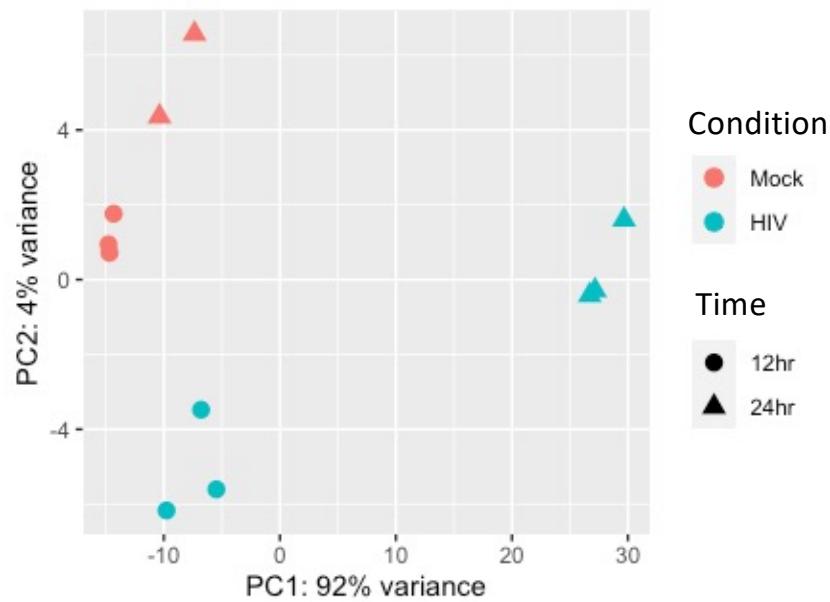
What are the major sources of variation in the data?

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?

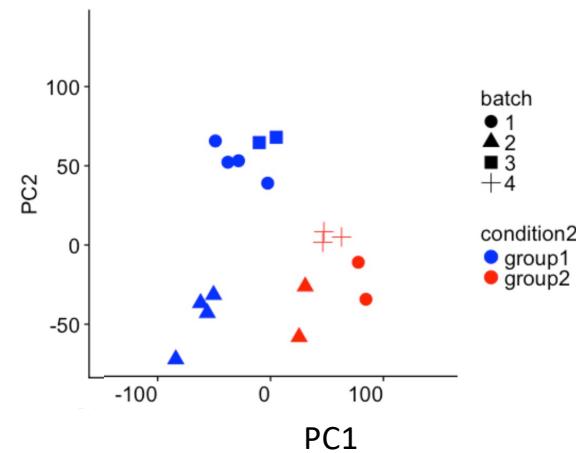
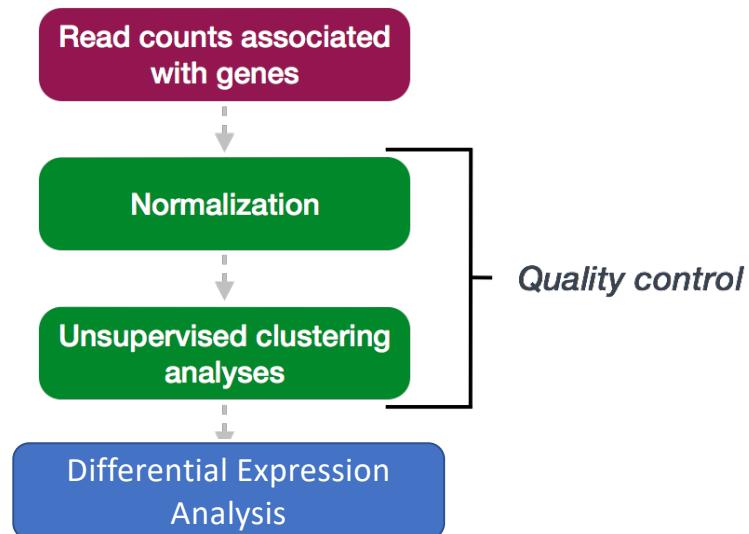


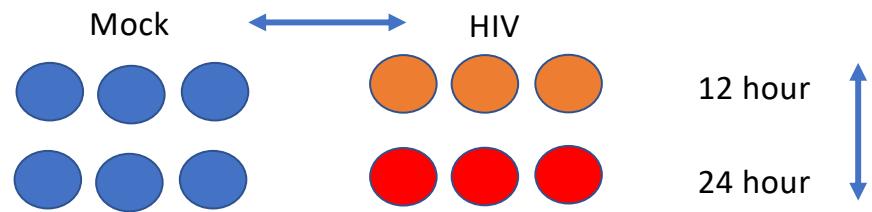
Image <https://support.bioconductor.org/p/111491/>

Differential Expression with DESeq2



https://hbctraining.github.io/DGE_workshop

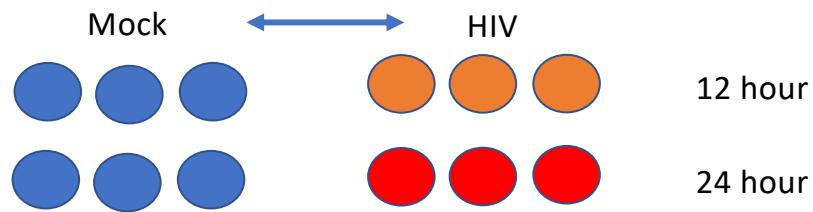
Multi-factor experiment design



Factor 1:
Infection status (Mock or HIV)

Factor 2:
Time (12 or 24 hr)

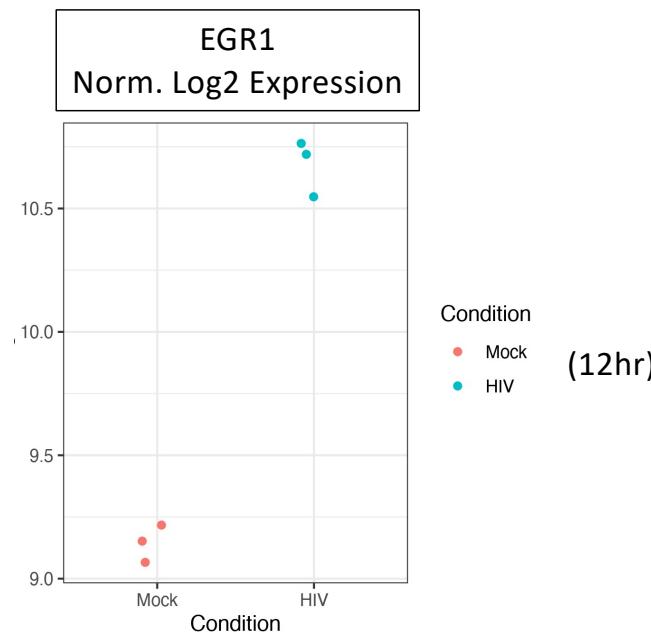
Multi-factor experiment design



- 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

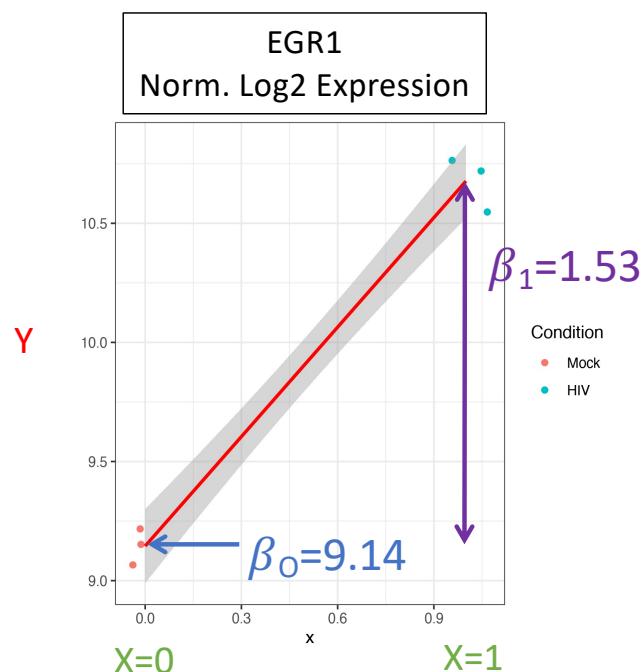
Step 1: Modeling gene expression values

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



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:



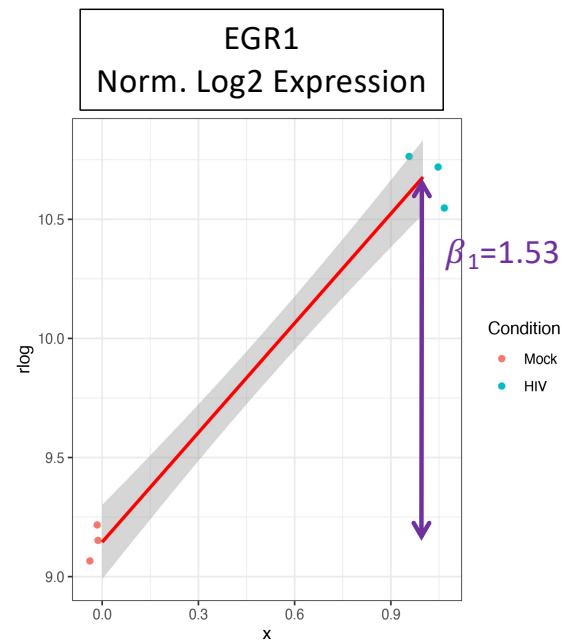
$$Y = \beta_0 + \beta_1 X + e$$

Intercept Condition (0-Mock, 1-HIV)
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



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

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?

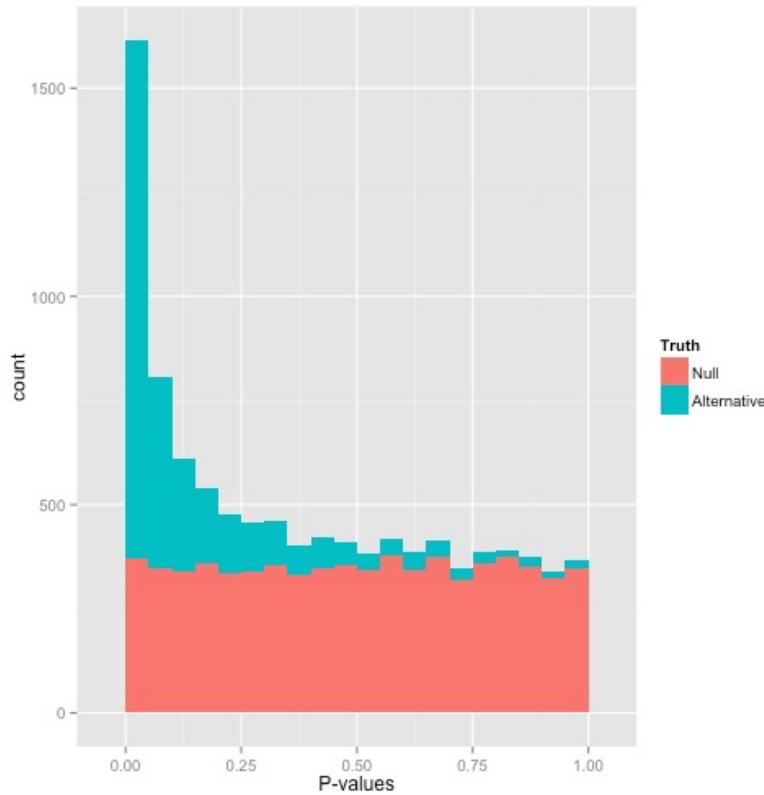
Yes! $p < 0.05$

DESeq2 Results table

GenelD	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:

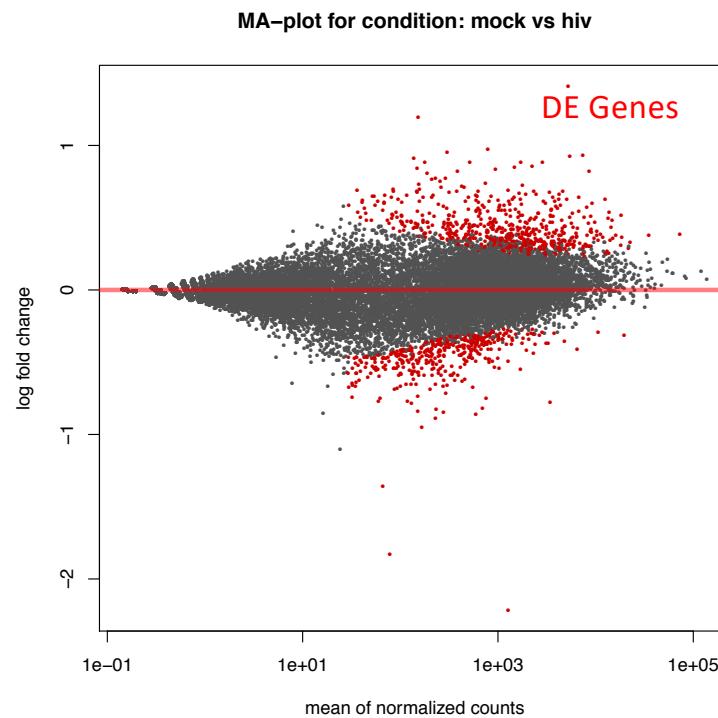
- Random P-values are expected to be uniform, if you have true positives you should see a peak close to zero

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

DESeq2 MA plot

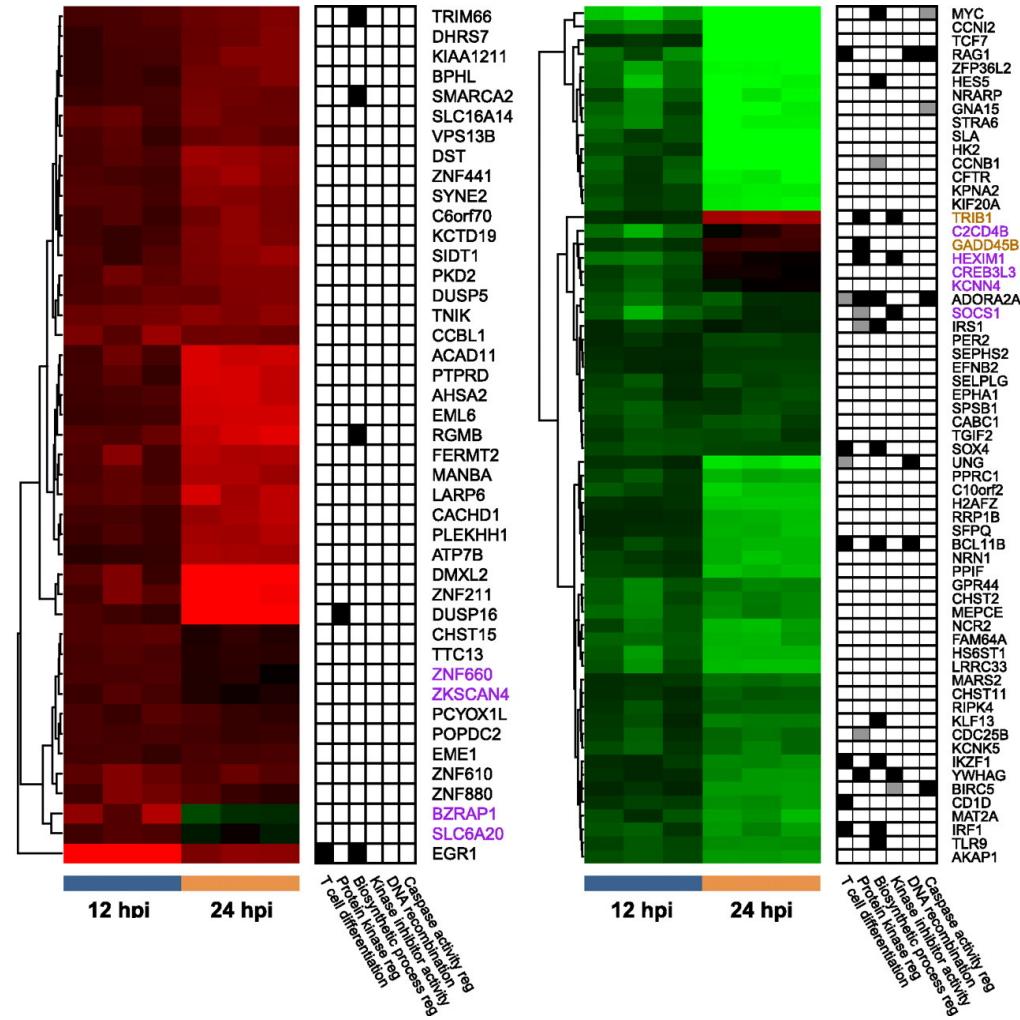
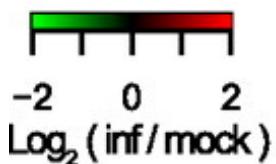
Shows the relationship between

- 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

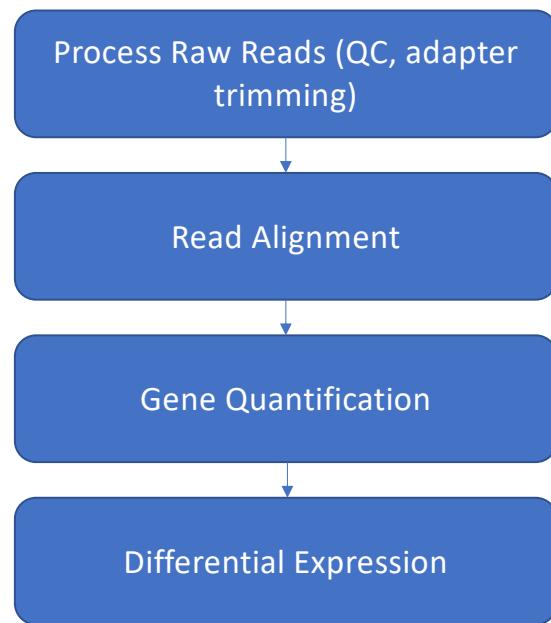


Study findings

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



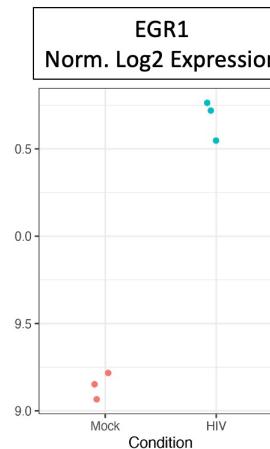
Conclusions



@SRR497699_30343179_1 HWI-EAS39X_10175_FC61MK0_4_117_4812_10346 length=75
CAGATGGCCGAGAGGAAGCCATGAAGGCCCTGCATGGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCAGAC
+
IIIIIGIIHFFFFFFIIDII>IIDHHIHDIIGIFIEIGIBDDEFIG<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



log2FoldChange = 1.55
Adjusted p-val <<0.05

References

DESeq2 vignette (R/Rstudio):

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

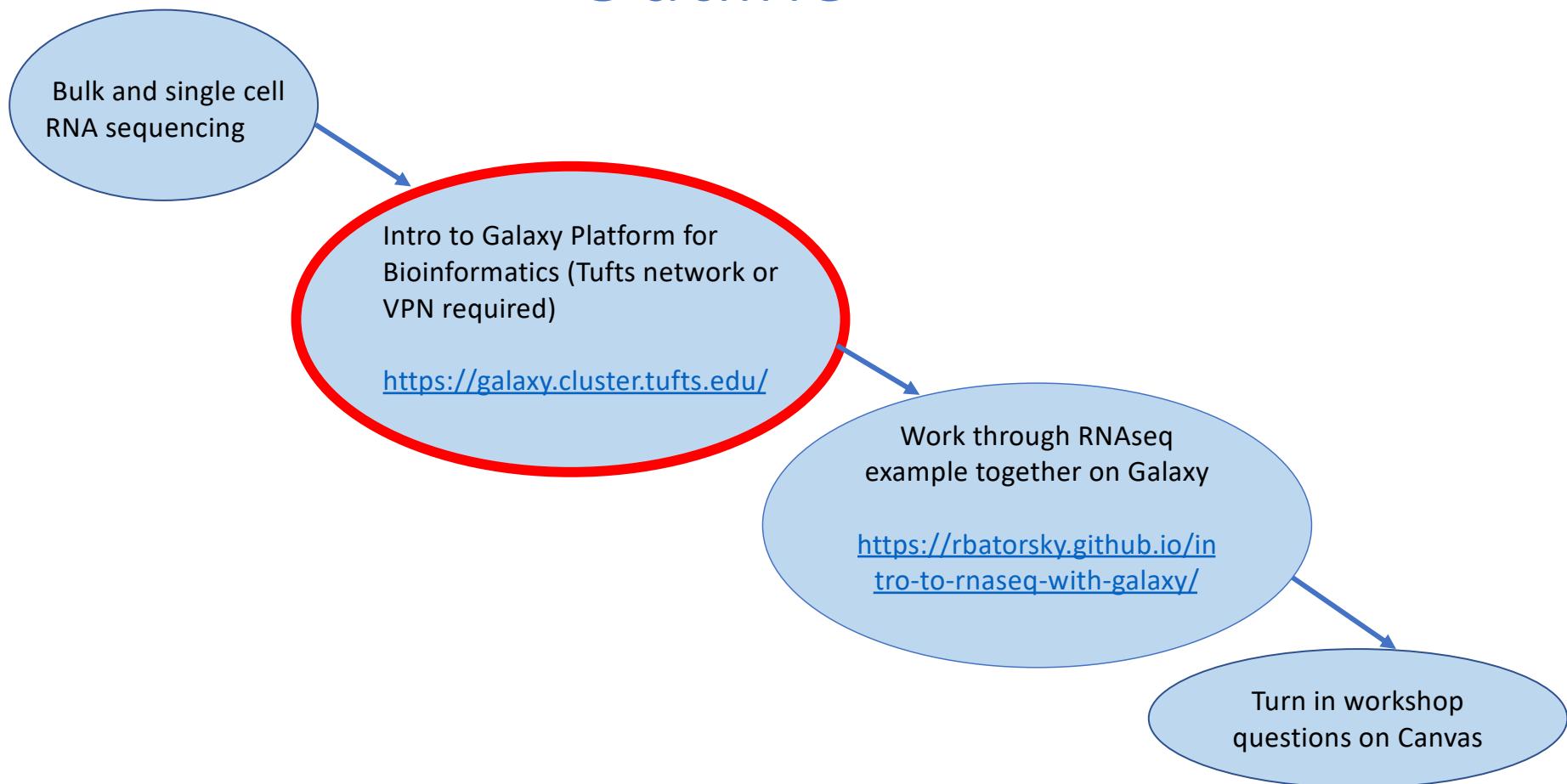
HBC Training (Command line/R):

https://hbctraining.github.io/DGE_workshop

Galaxy Training:

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

Outline



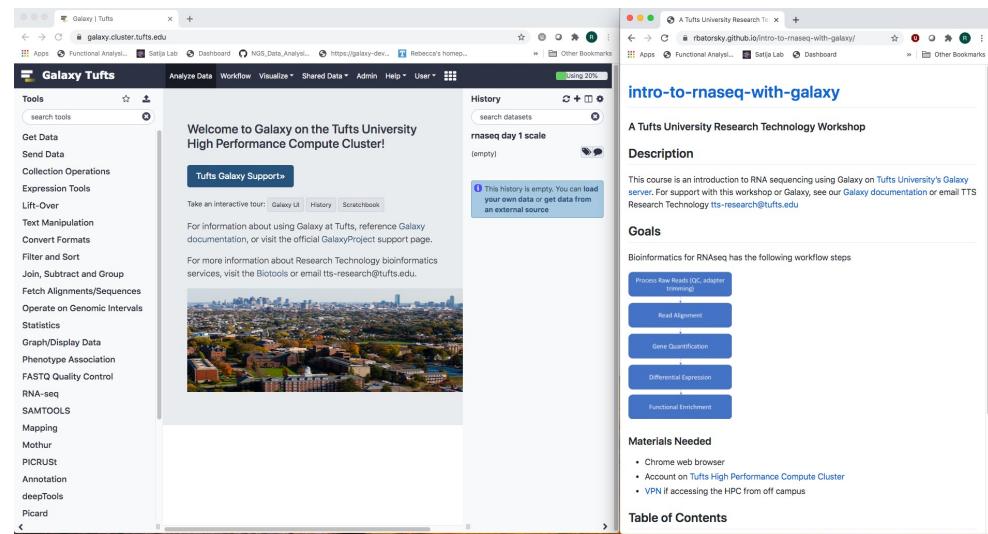


- ❖ **Web-based** platform for running data analysis and integration, geared towards bioinformatics
 - Open-source
 - Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic with many more outside contributions
 - Large and extremely responsive community

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://rbatortsy.github.io/intro-to-rnaseq-with-galaxy/>

Suggested screen layout



User Interface

Galaxy Tufts

Analyze Data Workflow Visualize Shared Data Admin Help User Using 20%

Tools search 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

search datasets

Unnamed history

(empty)

Tufts Galaxy Support»

Welcome to Galaxy on the Tufts University High Performance Compute Cluster!

Take an interactive tour: Galaxy UI History Scratchbook

For information about using Galaxy at Tufts, reference Galaxy documentation, or visit the official GalaxyProject support page.

For more information about Research Technology bioinformatics services, visit the Biotools or email tts-research@tufts.edu.



javascript:void(0)

77

User Interface

TOP MENU BAR

Analyze Data Workflow Visualize Shared Data Admin Help User

MAIN

Welcome to Galaxy on the Tufts University High Performance Compute Cluster!

Tufts Galaxy Support»

Take an interactive tour: Galaxy UI History Scratchbook

For information about using Galaxy at Tufts, reference Galaxy documentation, or visit the official GalaxyProject support page.

For more information about Research Technology bioinformatics services, visit the Biotools or email tts-research@tufts.edu.

TOOLS

Tools

search 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

javascript:void(0)

HISTORY

History

search datasets

Unnamed history

(empty)

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

Galaxy User Interface

To return to home screen

The screenshot shows the Galaxy User Interface (UI) for the Tufts University High Performance Compute Cluster. The UI is a web-based application with a dark header bar and a light gray content area.

Header Bar: The header bar includes the "Galaxy Tufts" logo, a search bar, and various navigation links: Analyze Data, Workflow, Visualize, Shared Data, Admin, Help, User, and a grid icon. A green progress bar at the top right indicates "Using 30%".

Left Sidebar (Tools): A sidebar titled "Tools" contains a search bar and a list of tool categories. The categories are grouped into sections: General 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, and PICRUSt.

Main Content Area: The main content area features a welcome message: "Welcome to Galaxy on the Tufts University High Performance Compute Cluster!" followed by a "Tufts Galaxy Support»" button. Below this, there are links for "Take an interactive tour: Galaxy UI, History, Scratchbook". It also provides information about using Galaxy at Tufts, including links to Galaxy documentation and the official GalaxyProject support page. A photograph of the Boston skyline is displayed.

Right Sidebar (History): The history sidebar is titled "History" and shows an "Unnamed history" section which is currently empty. It includes a search bar, a message indicating it's empty, and options to load data or get data from an external source.

Bottom Navigation: At the bottom of the interface, there are several small icons with red circles around them, likely for minimizing or adjusting toolbars. These icons include arrows for navigating between panels and a square icon.

History

The screenshot shows a user interface for managing histories. At the top left is a "History" button. To its right is a toolbar with four icons: a circular arrow, a plus sign, a square with a minus sign, and a gear. A red arrow points from the text "Create New History" to the plus sign icon. Another red arrow points from the text "View all Histories" to the square minus sign icon. Below the toolbar is a search bar labeled "search datasets" with a clear button. Underneath the search bar is the title "Unnamed history". Below the title is the text "(empty)". To the right of "(empty)" are two small icons: a tag and a speech bubble. At the bottom of the interface is a blue info box containing the text: "This history is empty. You can load your own data or get data from an external source".

History

Create New History

View all Histories

search datasets

Unnamed history

(empty)

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

History

History

search datasets x

Unnamed history

(empty)

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

Create New History 

View all Histories 

Create New History

View all Histories

Galaxy

Analyze Data Workflow Visualize Shared Data Admin Help User

Using 30.9 GB

Current History

72 shown, 1 deleted, 49 hidden
6.8 GB

Drag datasets here to copy them to the current history

122: WT_3_collection

114: WT_2_collection

106: WT_1_collection

98: SNPF_2_collection

90: SNPF_2_collection

82: SNPF_1_collection

74: Concatenate datasets on data 68, data 67, and others

73: Concatenate datasets on data 61, data 60, and others

72: Concatenate datasets on data 54, data 53, and others

71: Concatenate datasets on data 47, data 46, and others

Switch to: search datasets search all datasets

Unnamed history

3 shown, 5 deleted
7.56 GB

5: Concatenate datasets on data 1 and data 2

3: merge2_ext/SRR960440_pss & fastq.gz

1: merge2_ext/SRR960450_pss & fastq.gz

181: RNA_STAR on data 86, data 85, and others: Log

180: MultiQC on data 86, data 85, and others: Log

179: MultiQC on data 86, data 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

Switch to: search datasets search all datasets

Unnamed history

3 shown, 9 deleted, 19 hidden
14.01 GB

3: Create DBKey and Reference Genome

2: Create DBKey and Reference Genome

1: Create DBKey and Reference Genome

Switch to: search datasets search all datasets

Unnamed history

3 shown
9.04 kB

1: bed_file.bed

Switch to: search datasets search all datasets

Unnamed history

1 shown
(empty)

Switch to: search datasets search all datasets

Tools

The screenshot shows the Galaxy web interface on a Tufts cluster. The left sidebar is titled "Tools" and contains a search bar labeled "search tools". A red circle highlights this search bar. Below it is a list of tool categories: 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, featureCounts, RNA STAR, SAMTOOLS, Mapping, and Workflows. A green arrow points to the "RNA-seq" entry. Another red circle highlights the "RNA-seq" entry. The main content area displays a welcome message: "Welcome to Galaxy on the Tufts cluster" and "Bioinformatics @ Tufts". It includes links for "Take an interactive tour: Galaxy UI, History, Scratchbook". Below this is a paragraph about Galaxy's purpose and support. The right sidebar is titled "History" and shows an "Unnamed history" section which is currently empty. A blue box contains the message: "This history is empty. You can load your own data or get data from an external source".

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.

21

Tools

The screenshot shows the Galaxy web interface. On the left, the 'Tools' panel is open, listing various bioinformatics tools. A green box highlights the 'featureCounts' tool under the 'RNA-seq' section. An arrow points from this box to the main tool configuration panel on the right. The main panel displays the 'featureCounts' tool details, including its version (1.6.4), input requirements (Alignment file, Gene annotation file), output format (Gene-ID "-t" read-count), and execution options. The 'History' panel on the right shows an empty history list.

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

featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files. (Galaxy Version 1.6.4)

Alignment file

No bam or sam dataset available.

The input alignment file(s) where the gene expression has to be counted. The file can have a SAM or BAM format; but ALL files must be in the same format. Unless you are using a Gene annotation file from the History, these files must have the database/genome attribute already specified e.g. hg38, not the default: ?

Specify strand information

Unstranded

Indicate if the data is stranded and if strand-specific read counting should be performed. Strand setting must be the same as the strand settings used to produce the mapped BAM input(s) (-s)

Gene annotation file

locally cached

Using locally cached annotation

No options available

If the annotation file you require is not listed here, please contact the Galaxy administrator

Output format

Gene-ID "-t" read-count (MultiQC/DESeq2/edgeR/limma-voom compatible)

The output format will be tabular, select the preferred columns here

Create gene-length file

Yes No

Creates a tabular file that contains the effective (nucleotides used for counting reads) length of the feature; might be useful for estimating FPKM/RPKM

Options for paired-end reads

Advanced options

Execute

History

search datasets

Unnamed history

(empty)

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

Importing data

The screenshot shows the Galaxy web interface. On the left, there's a sidebar titled "Tools" with various options like "Get Data", "Send Data", and "Workflows". In the center, it says "Welcome to Galaxy on the Tufts cluster" and "Bioinformatics @ Tufts". At the top, there's a navigation bar with "Analyze Data", "Workflow", "Visualize", "Shared Data" (which is highlighted with a red box and has a red arrow pointing to it from the text "Import shared data libraries"), "Admin", "Help", "User", and "Using 14.7 GB". On the right, there's a "History" section showing an "Unnamed history" with "(empty)" datasets. A message box says "This history is empty. You can load your own data or get data from an external source".

Import shared data libraries

Upload data from local storage or from the cluster

Shared Data

Analyze Data Workflow Visualize Shared Data Admin Help User Using 14.7 GB

Galaxy

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

History

Unnamed history

(empty)

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

17

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://rbatarsky.github.io/intro-to-rnaseq-with-galaxy/>

5. Under Table of Contents click on “Introduction and Setup”

Suggested screen layout

