RNA-seq to study HIV Infection in cells

Rebecca Batorsky

Sr Bioinformatics Specialist

Oct 2020

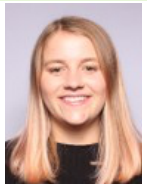
Research Technology Team

Delilah Maloney
High Performance Computing Specialist

Kyle Monahan
Senior Data Science Specialist

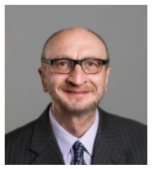
Shawn Doughty
Manager, Research Computing

Rebecca Batorsky
Senior Bioinformatics Specialist



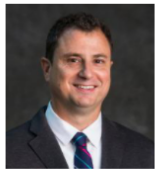
Meg Farley
Bioinformatics Intern

Chris Barnett
Senior Geospatial Analyst



Lionel H. Zupan
Director, Research Technology

Tom Phimmasen
Senior Data Consultant



Patrick Florance
Director, Academic Data Services

Jake Perl
Digital Humanities NLP Specialist

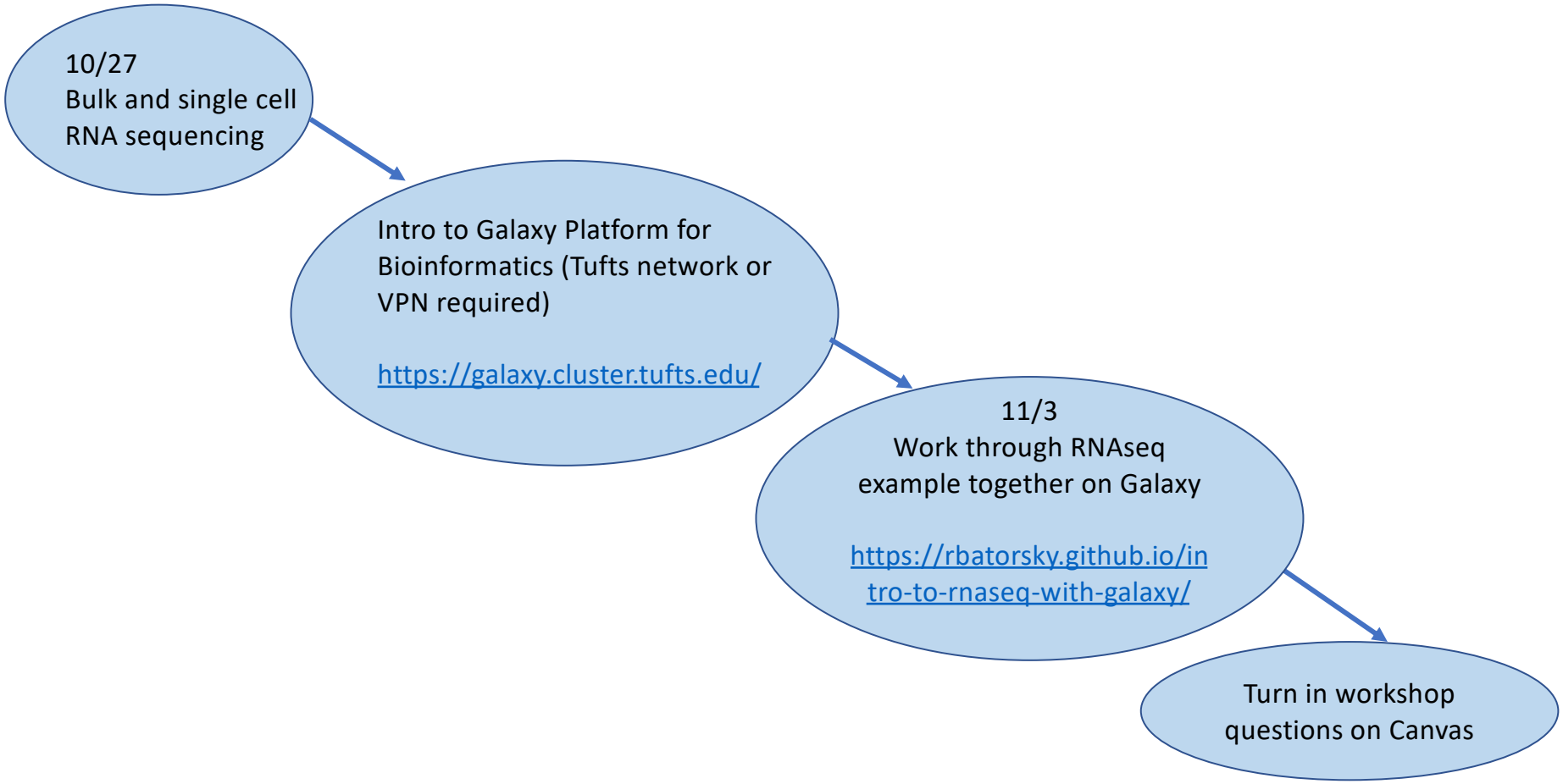
Carolyn Talmadge
Senior GIS Specialist

Uku-Kaspar Uustalu
Data Science Specialist

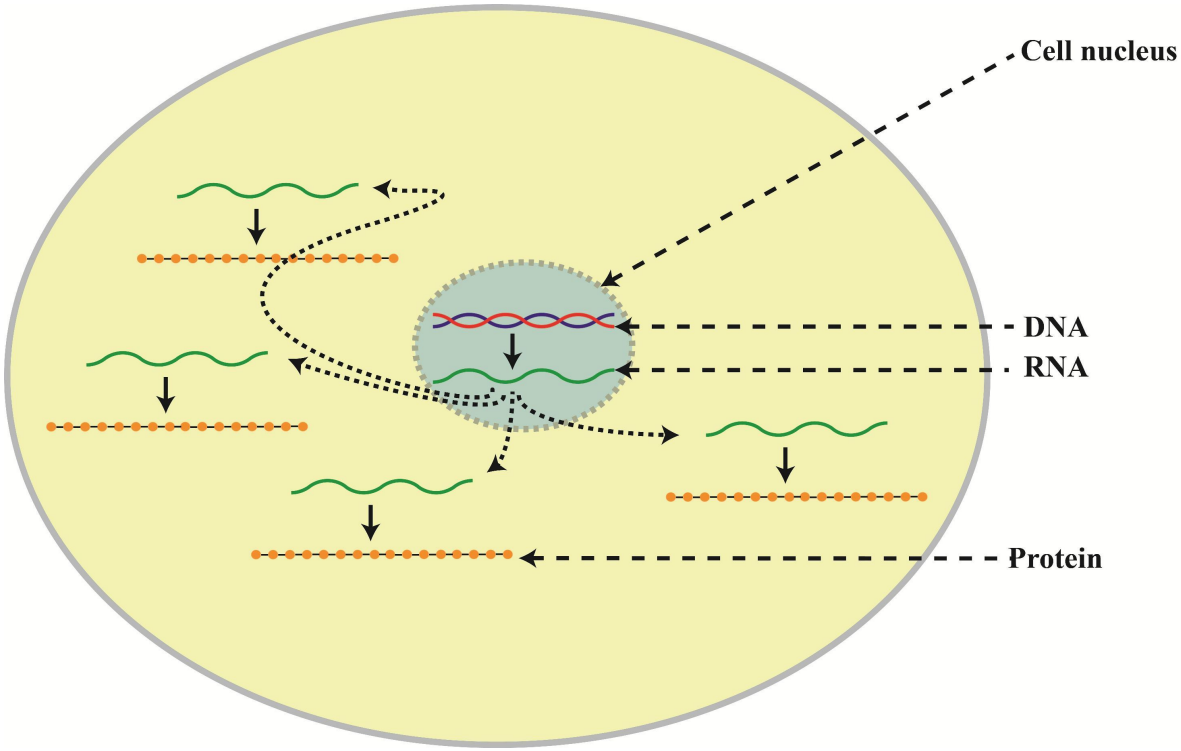
- ✓ Consultation on Projects and Grants
- ✓ High Performance Compute Cluster
- ✓ Workshops

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

Outline


10/27
Bulk and single cell RNA sequencing
Intro to Galaxy Platform for Bioinformatics (Tufts network or VPN required)
https://galaxy.cluster.tufts.edu/
11/3
Work through RNAseq example together on Galaxy
https://rbatorsky.github.io/intro-to-rnaseq-with-galaxy/
Turn in workshop questions on Canvas

DNA and RNA in a cell

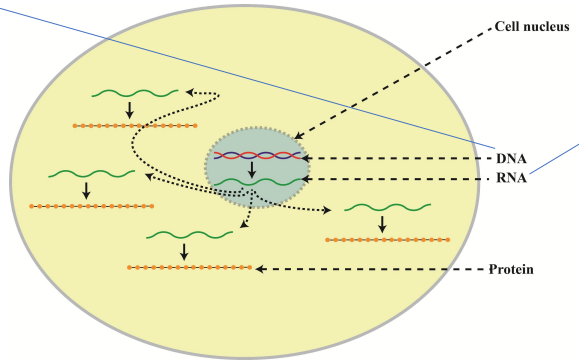


https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

Two common analyses

DNA Sequencing

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



RNA Sequencing

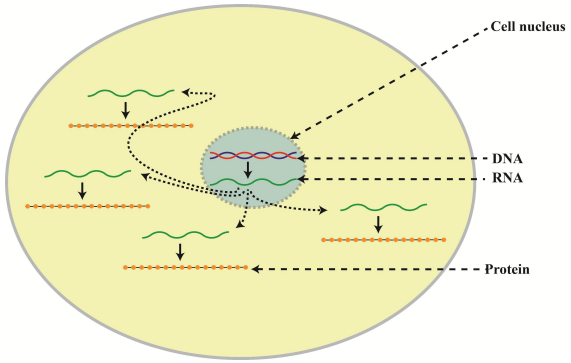
- Copy of a gene (mRNA) per cell depends on gene expression
- Analysis goal:
  - Bulk : Differential expression
  - Single cell : Quantify different cell populations

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

Today we will cover RNA sequencing

DNA Sequencing

- Fixed [illegible] of a [illegible] per cell
- Analysis go[illegible]
  Variant c[illegible] interpretation



RNA Sequencing

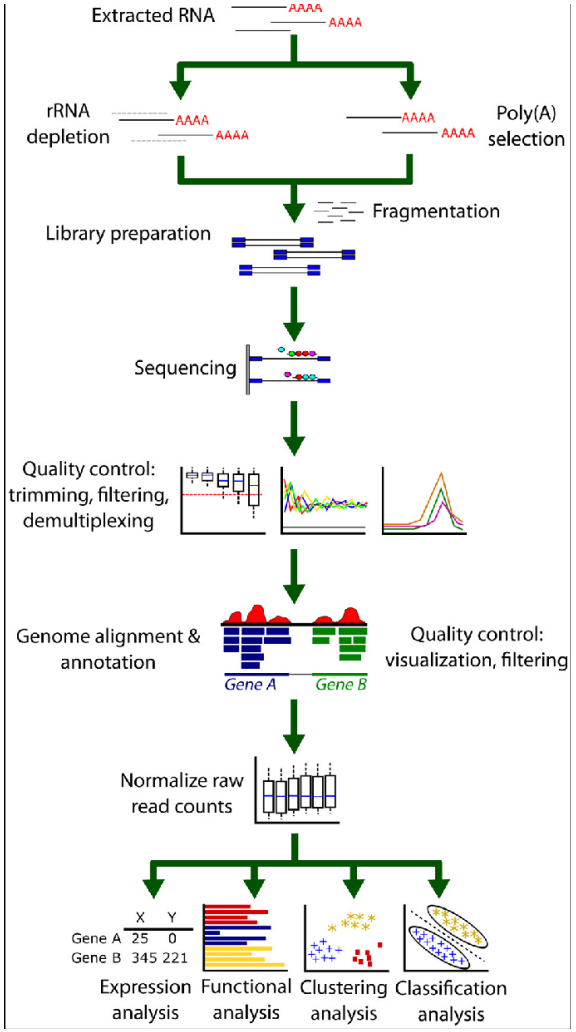
- Copy of a gene (mRNA) per cell depends on gene expression
- Analysis goal:
  - Bulk : Differential expression
  - Single cell : Quantify different cell populations

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

“Bulk” RNA seq workflow

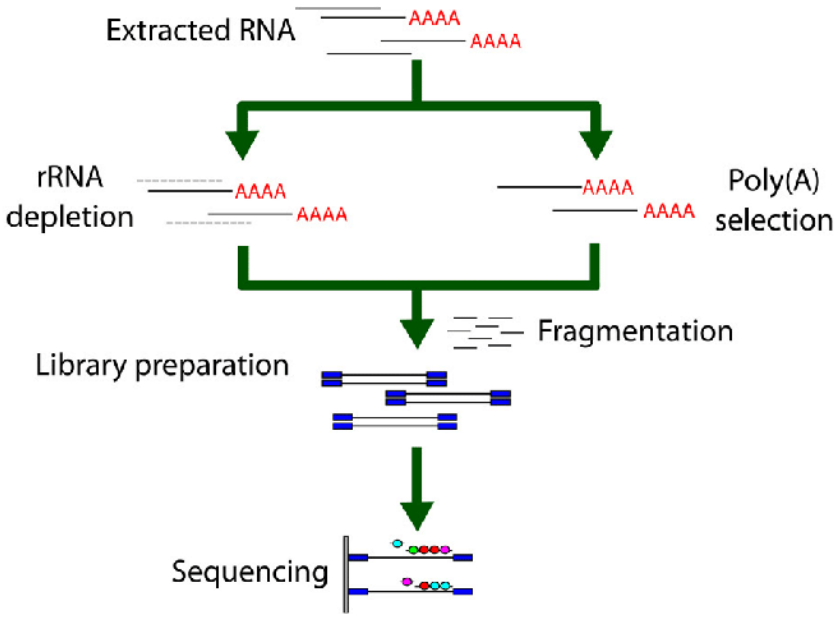
Library prep and sequencing

Bioinformatics



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

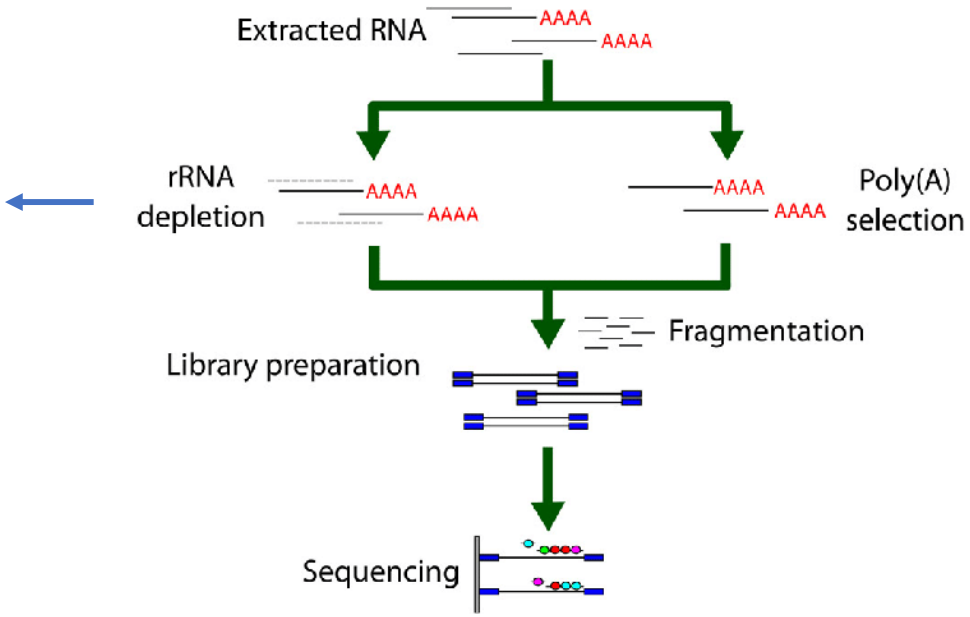
RNA seq library prep and sequencing



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

RNA seq library prep and sequencing

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

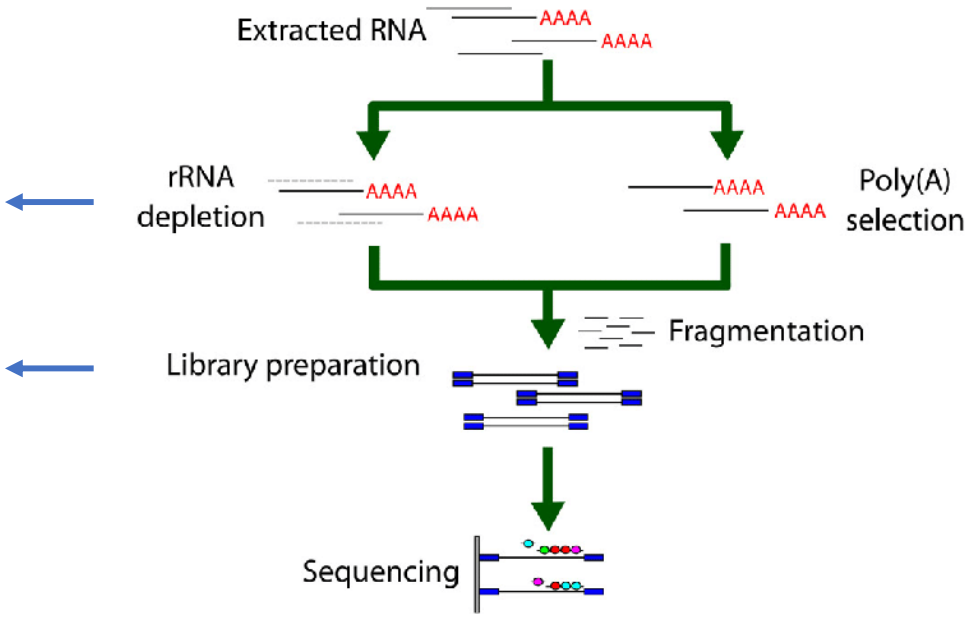


Good resource: [Griffiths et al Plos Comp Bio 2015]

RNA seq library prep and sequencing

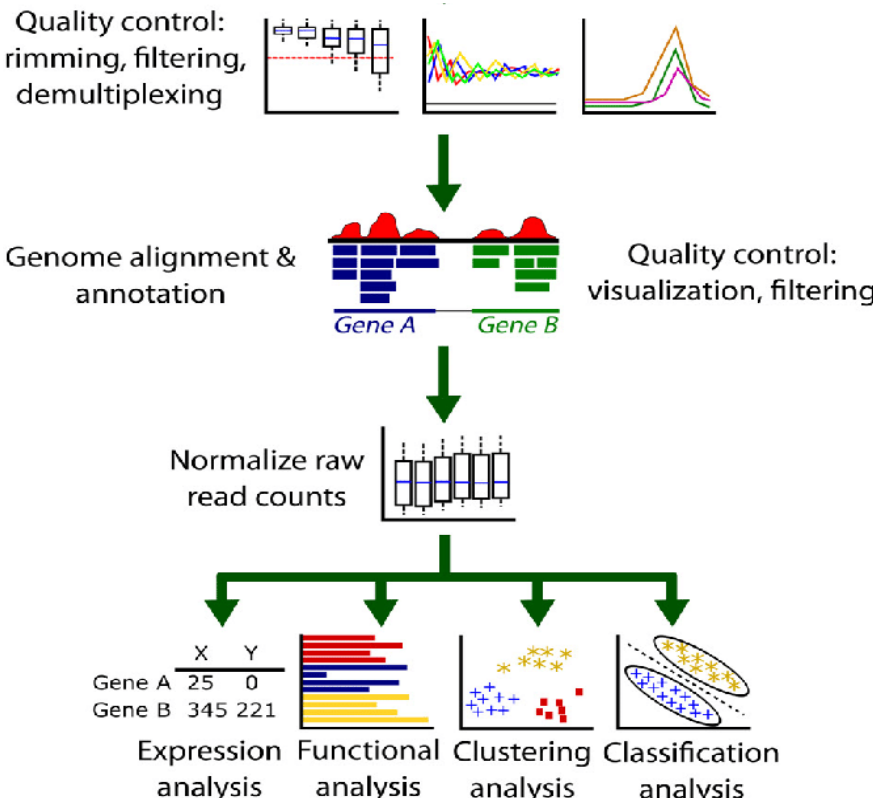
- Enrichment for mRNA
- In humans, ~95%–98% of all RNA molecules are rRNAs

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



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

RNA seq bioinformatics



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

Goal of Differential Expression in RNAseq

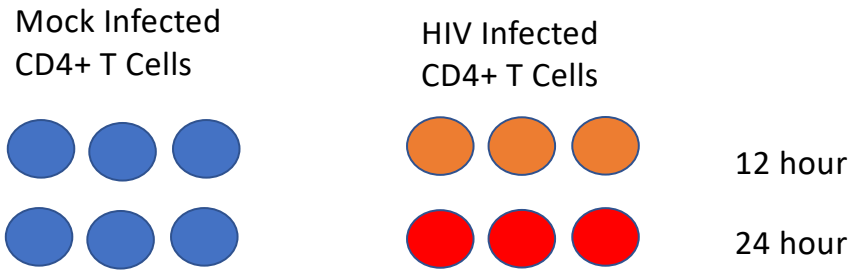
"How can we detect genes for which the counts of reads change between conditions **more systematically** than as expected by chance"

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

Our dataset for next week

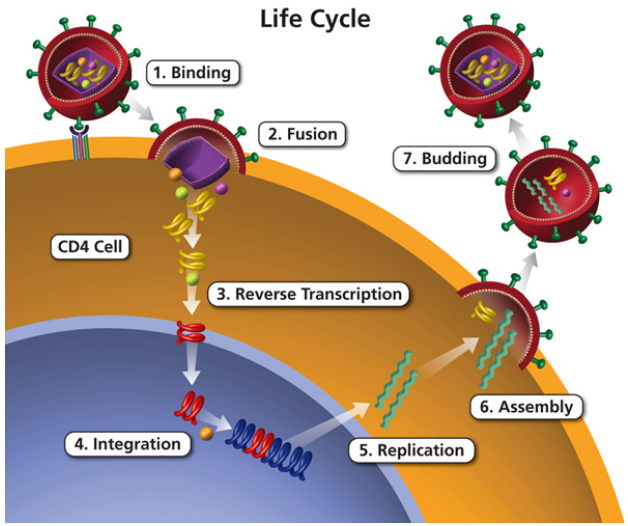
Next-Generation Sequencing Reveals HIV-1-Mediated Suppression of T Cell Activation and RNA Processing and Regulation of Noncoding RNA Expression in a CD4$^+$ T Cell Line

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



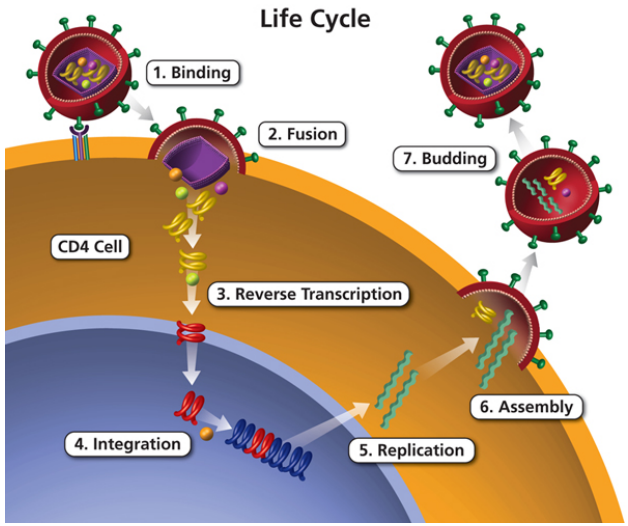
https://www.ncbi.nlm.nih.gov/pubmed/21933919

HIV lifecycle

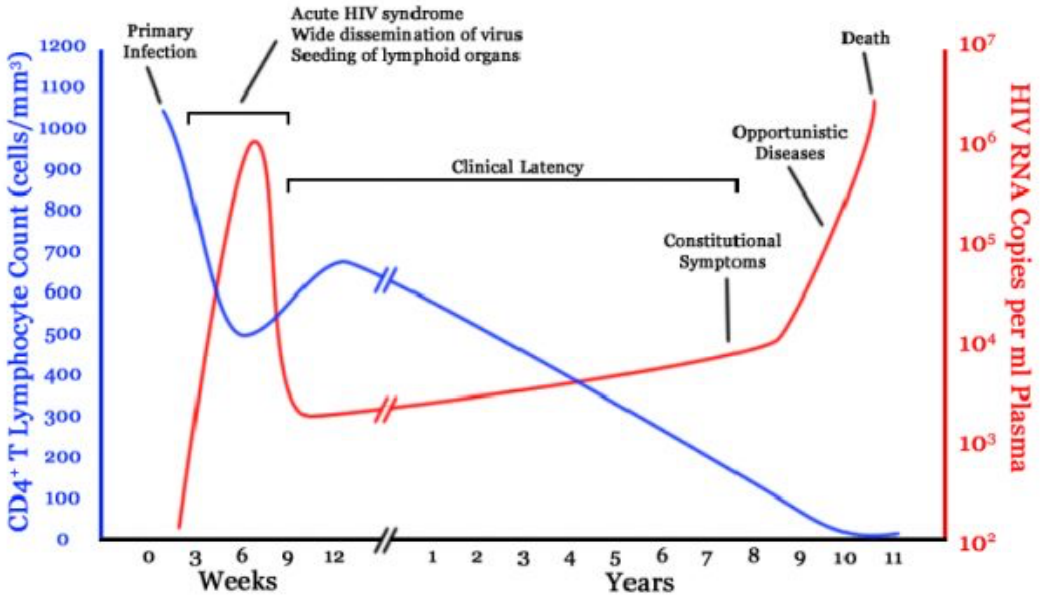


https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle

HIV lifecycle



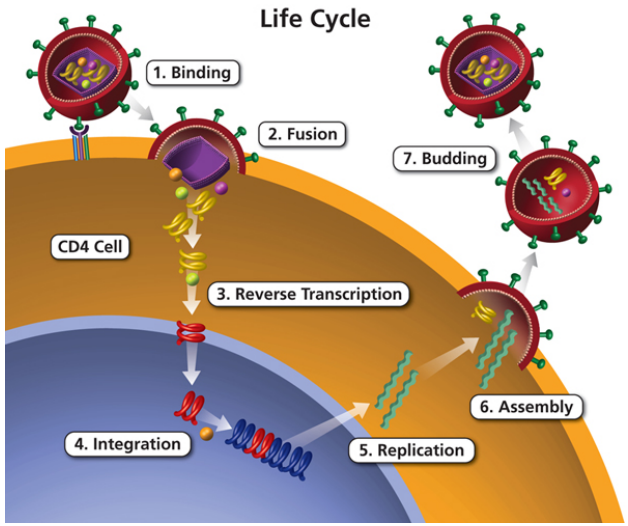
HIV infection in a human host

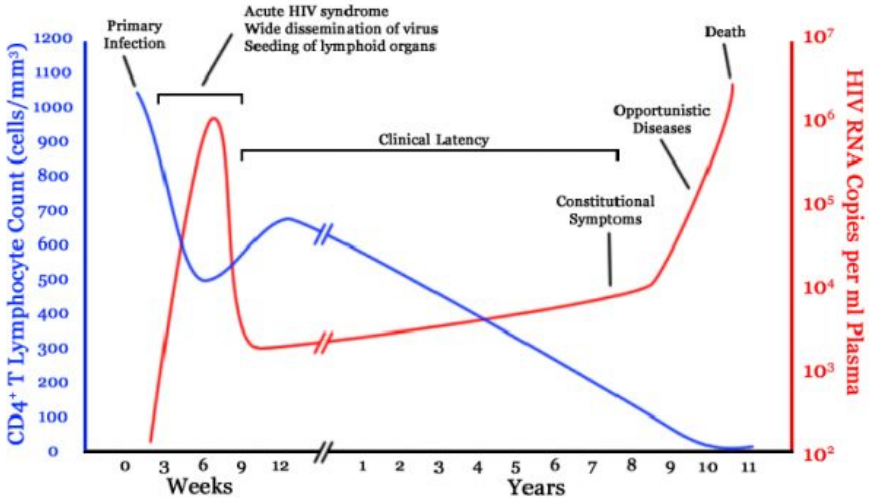


https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle

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?



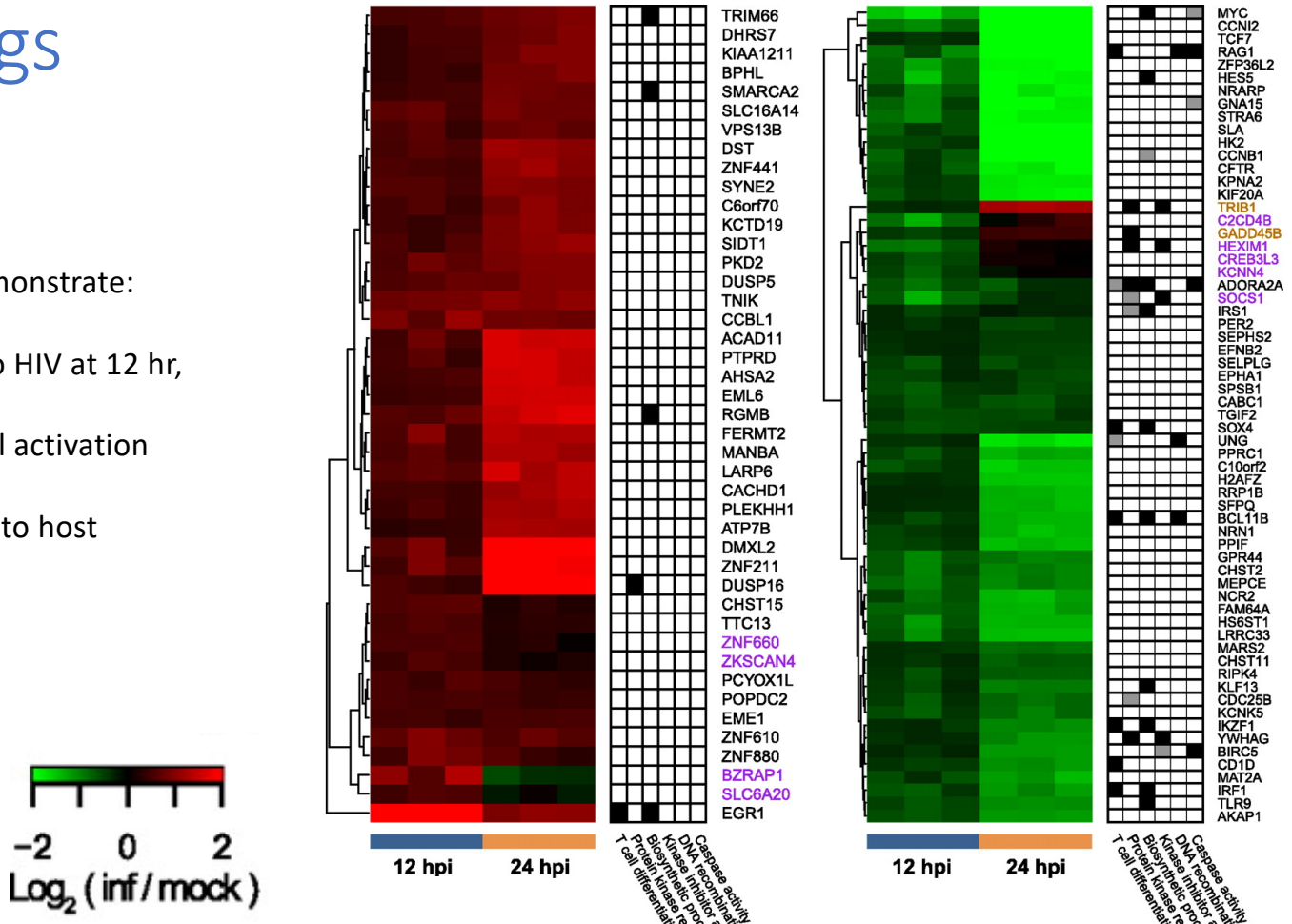


https://aidsinfo.nih.gov/understanding-hiv-aids/glossary/1596/life-cycle

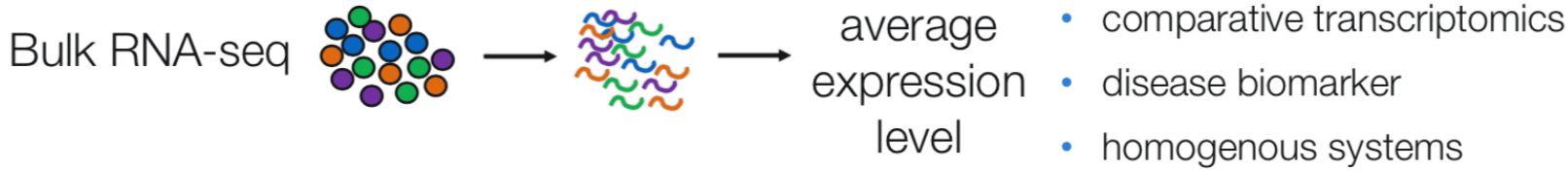
Study findings

Using RNAseq, authors demonstrate:

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

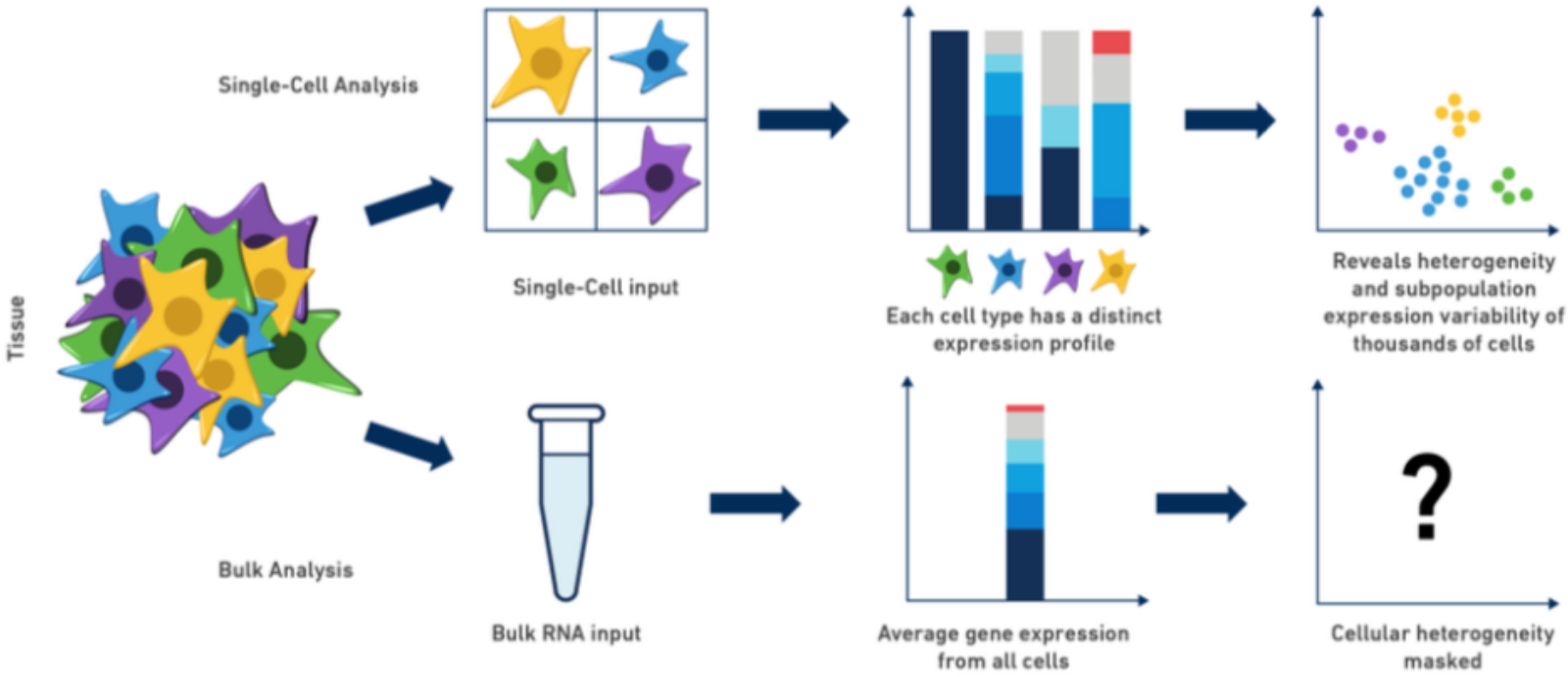


Bulk vs Single Cell RNA Sequencing



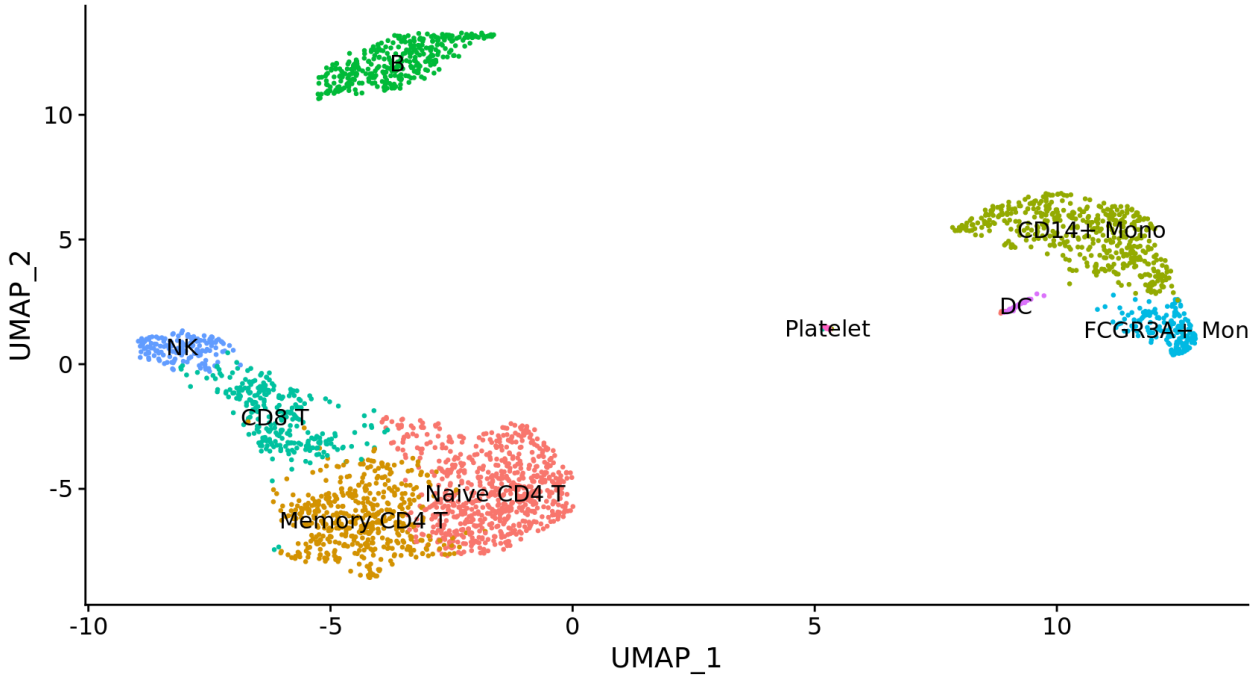
https://github.com/hbctraining/scRNA-seq

scRNA to study Heterogeneity in cell populations



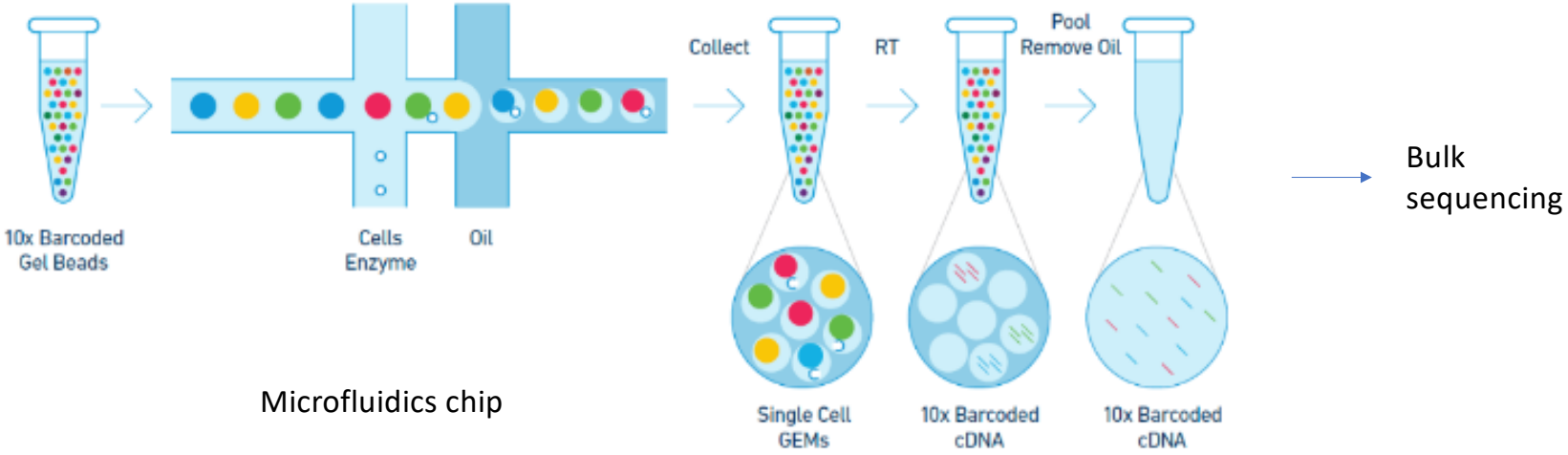
https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

scRNA cell subsets in PBMC



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

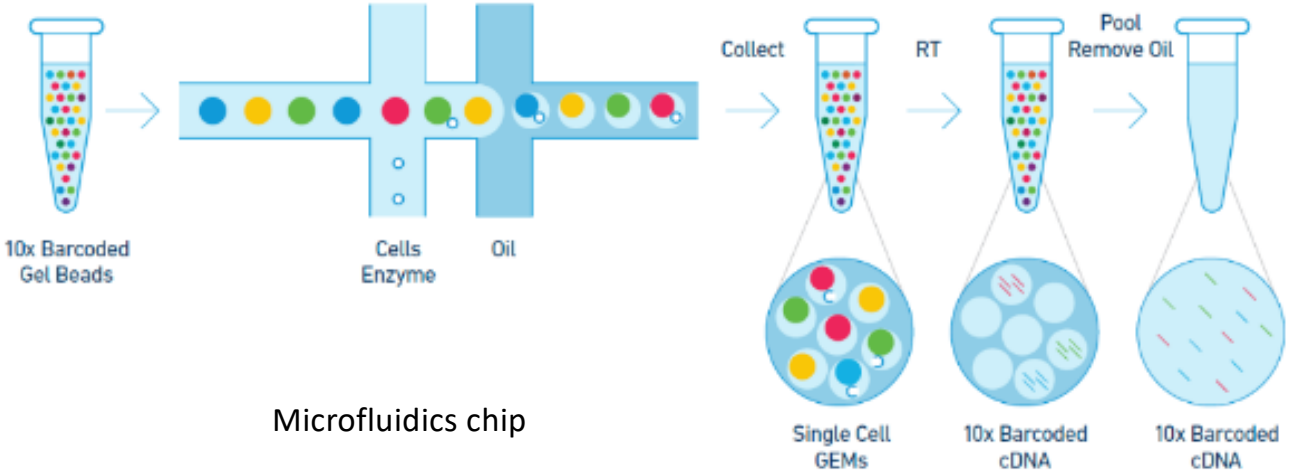
10x single cell technology



Microfluidics chip

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

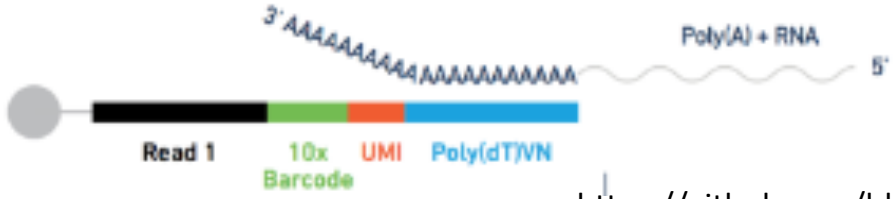
10x single cell technology



Microfluidics chip

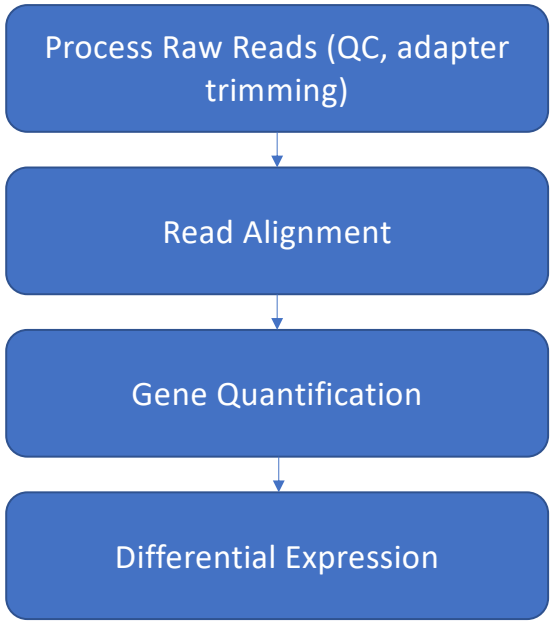
Bulk sequencing

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

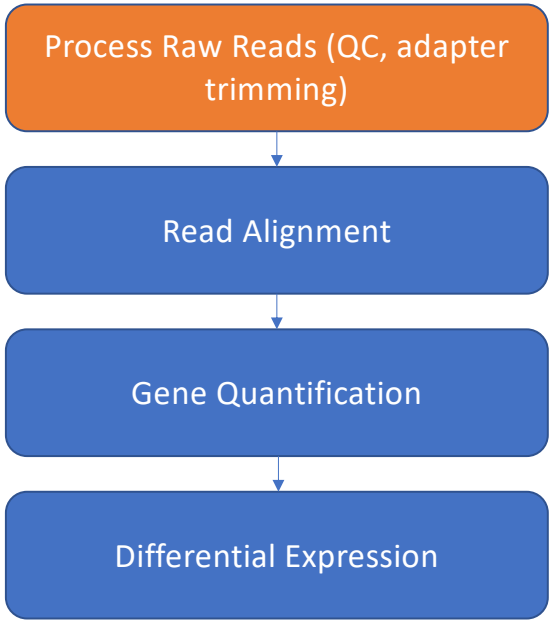


https://github.com/hbctraining/scRNA-seq

Our (bulk) RNAseq Workflow



Quality control on Raw Reads



Raw reads in Fastq format

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

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

Base Quality Scores

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

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
   Quality score: 0........10........20........30........40
```

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

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

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

Base Quality Scores

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

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
   Quality score: 0........10........20........30........40
```

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

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

Back to our read:

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

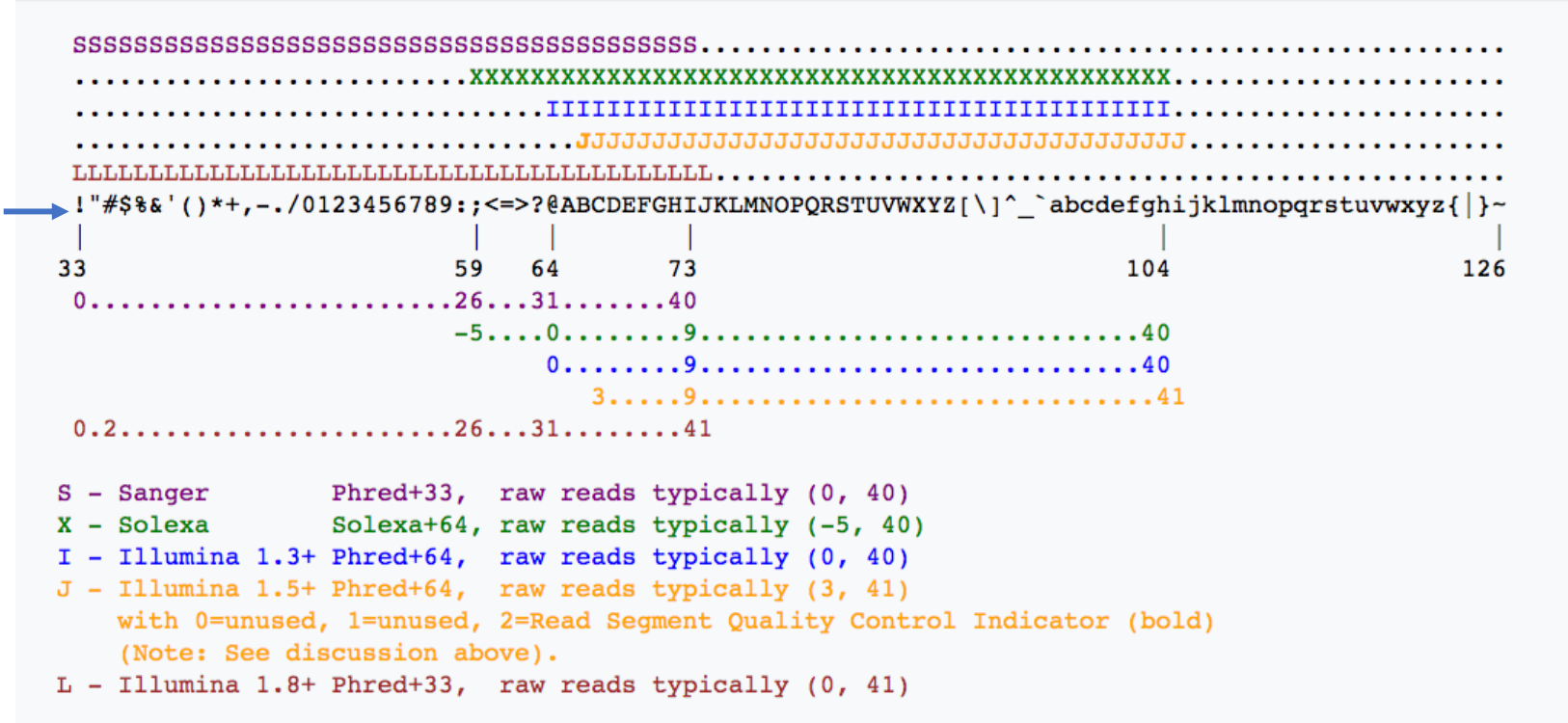
C –> Q = 34 -> Probability < 1/1000 of an error

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

Base Quality Scores

Quality encoding

Quality scores

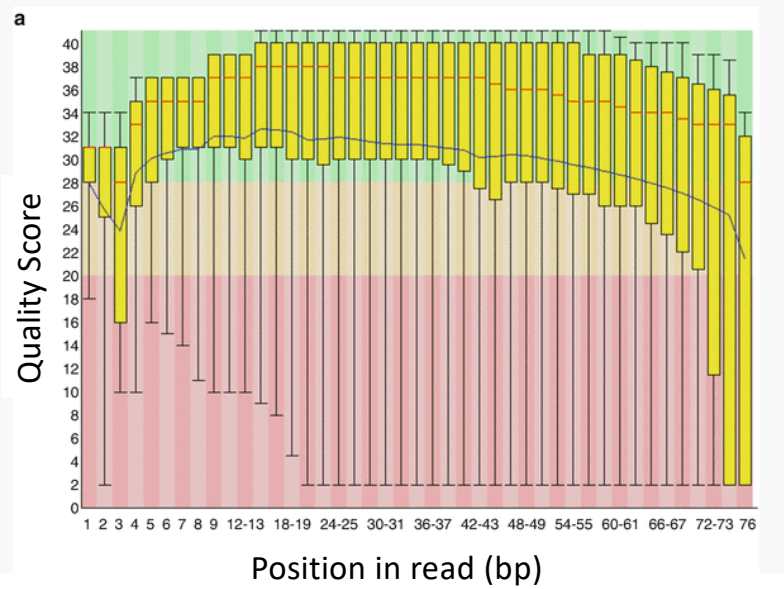


https://en.wikipedia.org/wiki/FASTQ_format

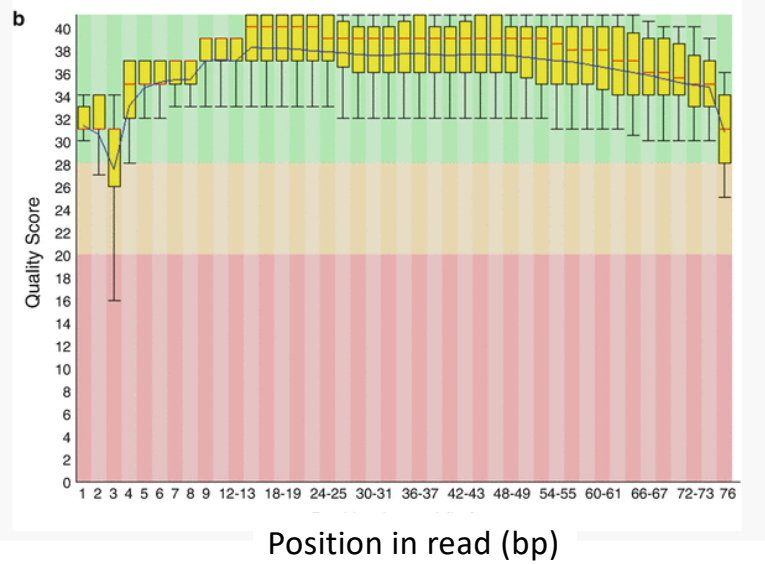
Raw read quality control

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

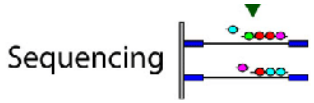
FastQC: Sequence Quality Histogram



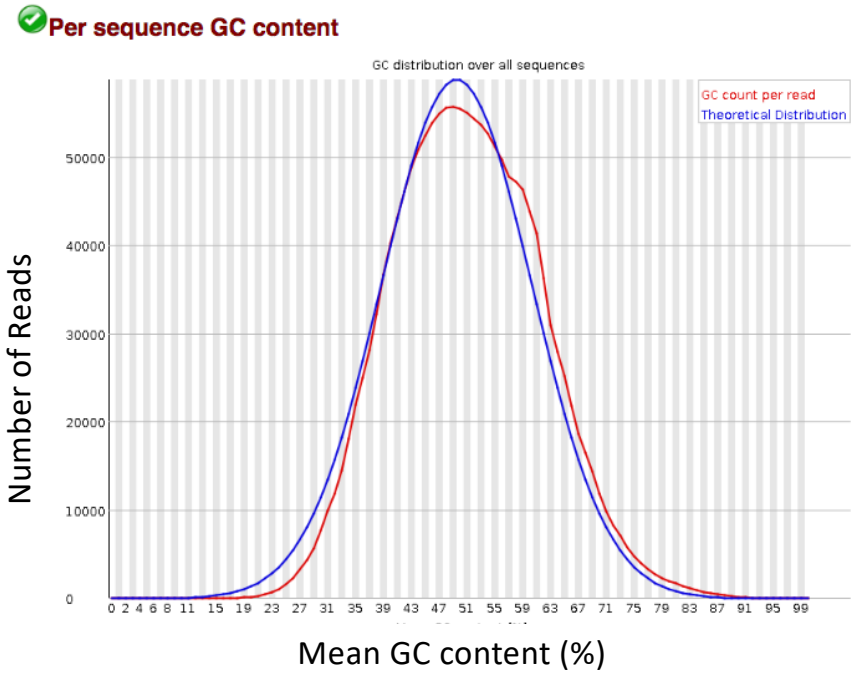
GOOD
High quality over the length of the read

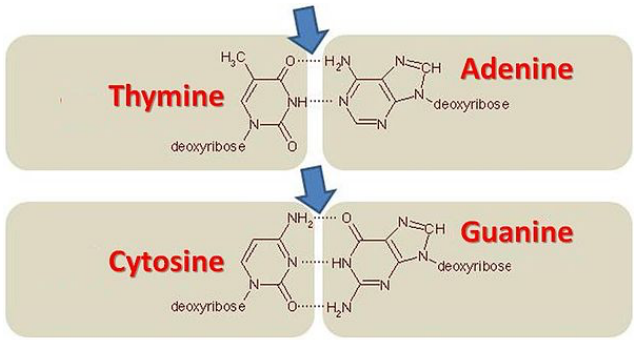


BAD
Read quality drops at the beginning and end

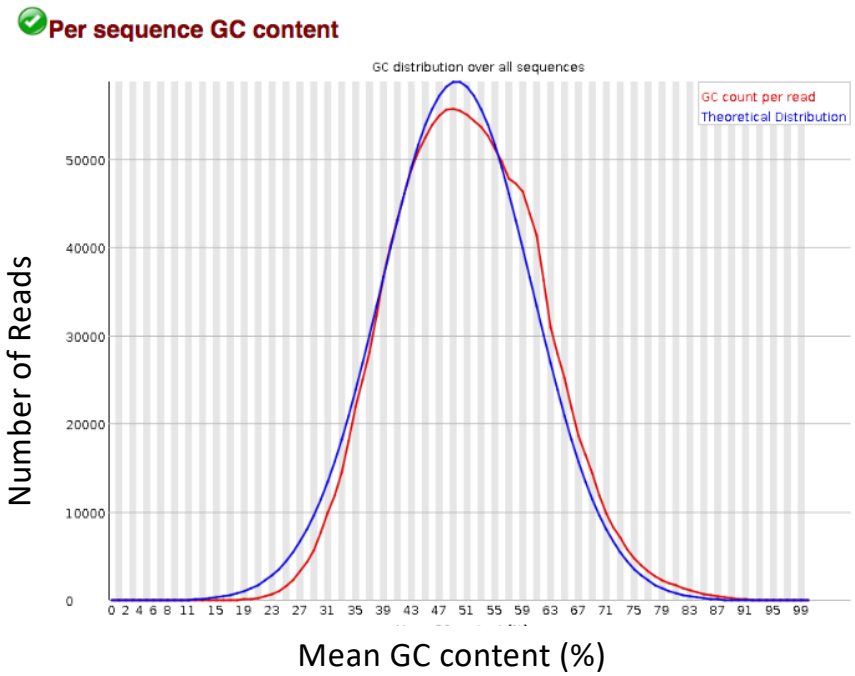


FastQC: Per sequence GC content

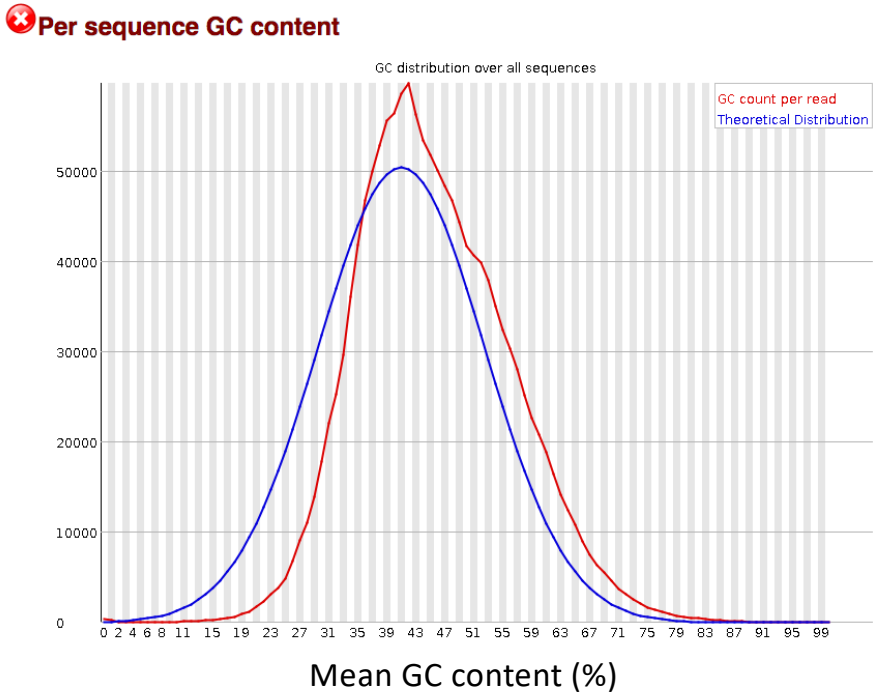




FastQC: Per sequence GC content

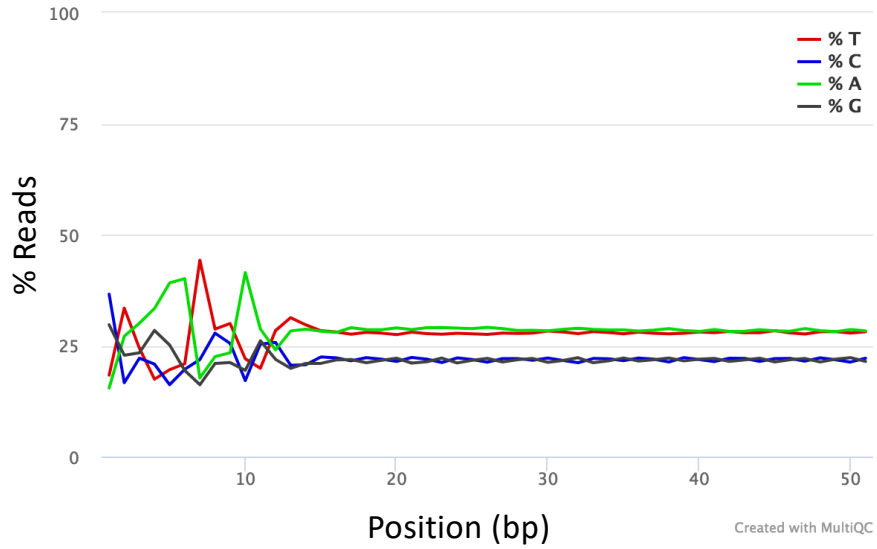


GOOD: follows normal distribution (sum of deviations is < 15% of reads)



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

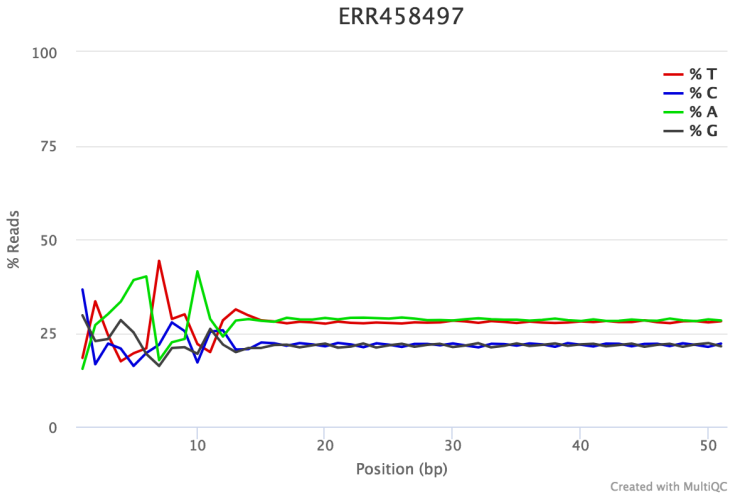
FastQC: Per Base Sequence Content



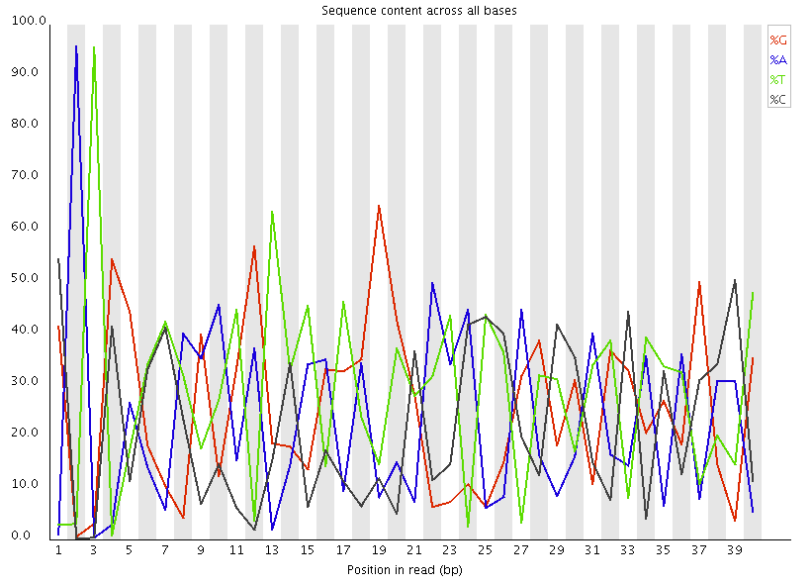
- Proportion of each position for which each DNA base has been called
- RNAseq data tends to show a positional sequence bias in the first ~12 bases
- The "random" priming step during library construction is not truly random and certain hexamers are more prevalent than others

sequencing.qcfail.com

FastQC: Per Base Sequence Content

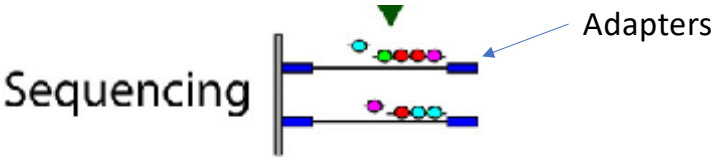


EXPECTED



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

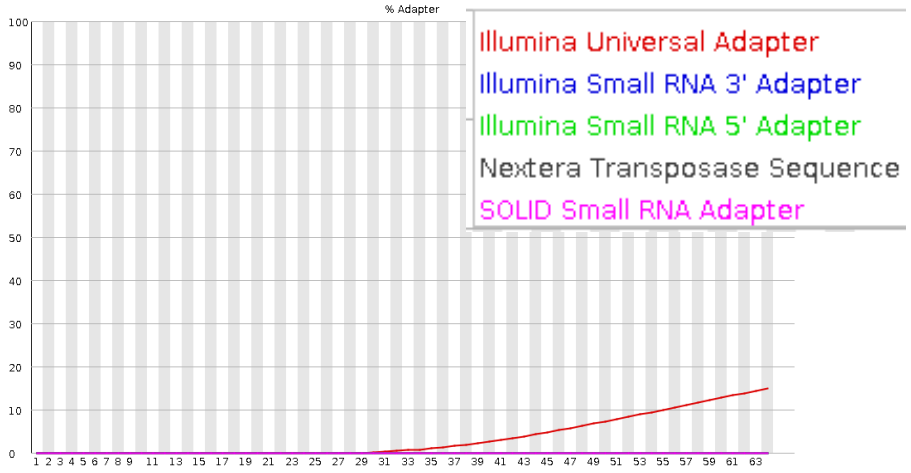
FastQC: Adapter content



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!

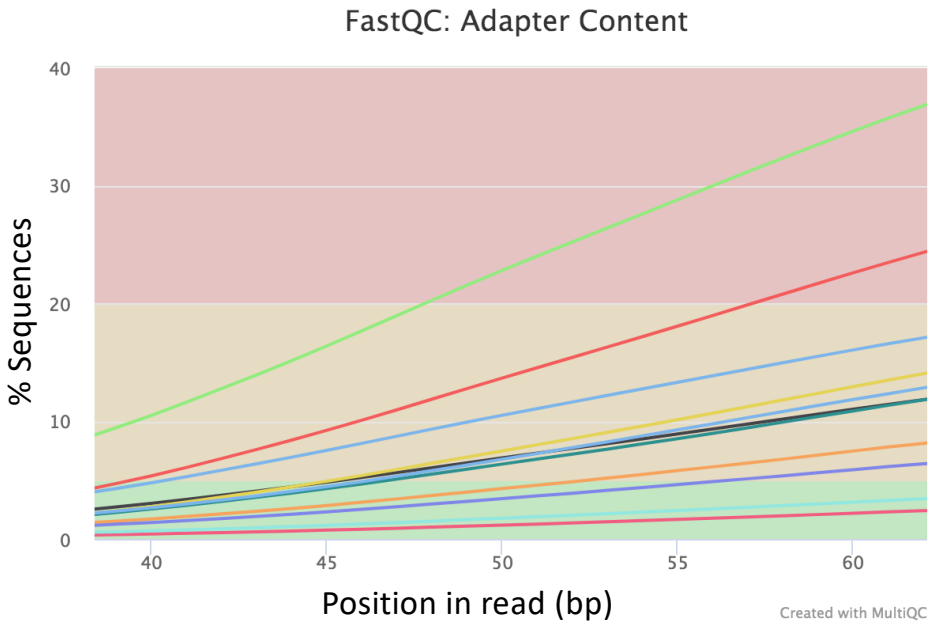


Position in read (bp)

sequencing.qcfail.com

FastQC -> MultiQC

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

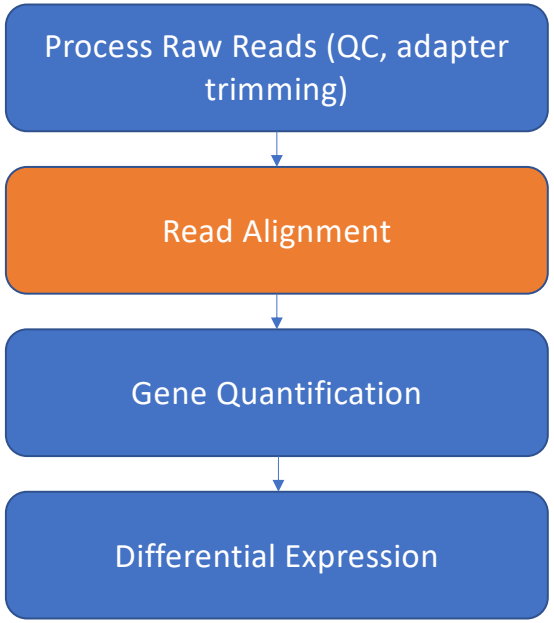


Adapter trimming

Trim Galore! is a tool that:

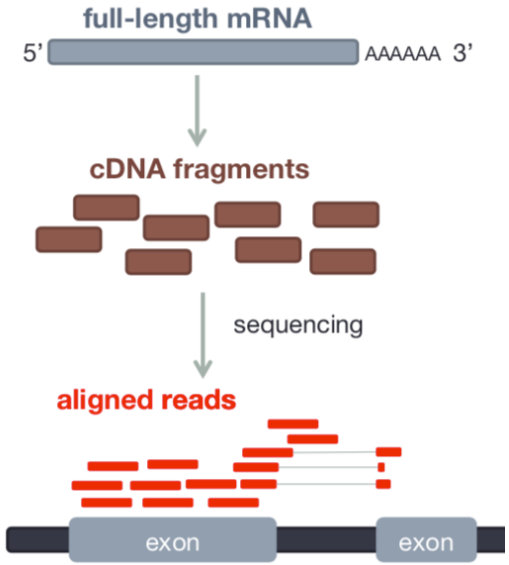
- Scans and removes known Illumina or custom adapters
- Performs read trimming for low quality regions at the end of reads
- Removes reads that become too short in the trimming process

Workflow


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

Read Alignment

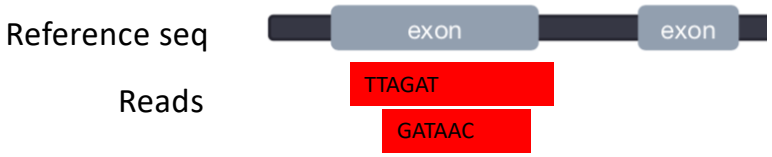
- RNAseq data originates from spliced mRNA (no introns)
- When aligning to the genome, our aligner must find a spliced alignment for reads
- We use a tool called STAR (Spliced Transcripts Alignment to a Reference) that has a exon-aware mapping algorithm.



Reference sequence

Dobin et al Bioinformatics 2013

Sequence Alignment Map (SAM)




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

Header section

Alignment section

CIGAR: summary of alignment, e.g. match, gap, insertion, deletion

Mapping Quality

Position

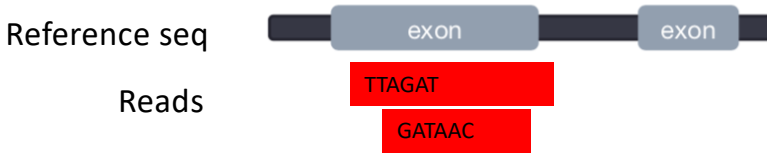
Ref Sequence name

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

Read ID

www.samformat.info

Sequence Alignment Map (SAM)




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

Header section

Alignment section

Paired end info

Sequence

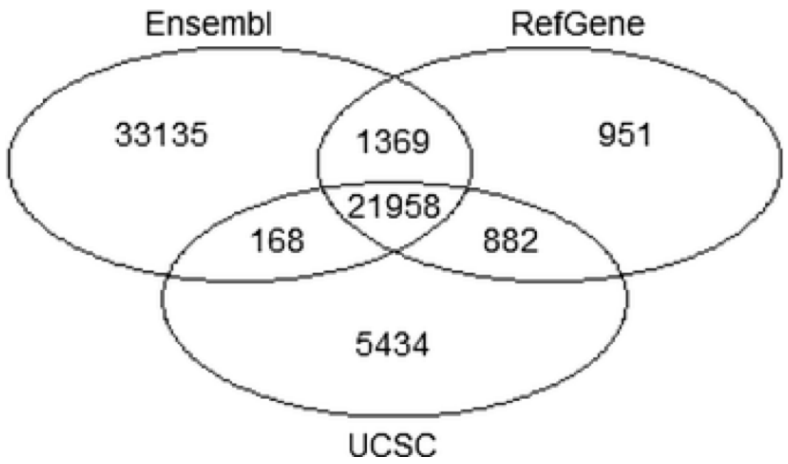
Quality Score

Optional Fields

www.samformat.info

Genome Annotation Standards

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



Zhao et al Bioinformatics 2015

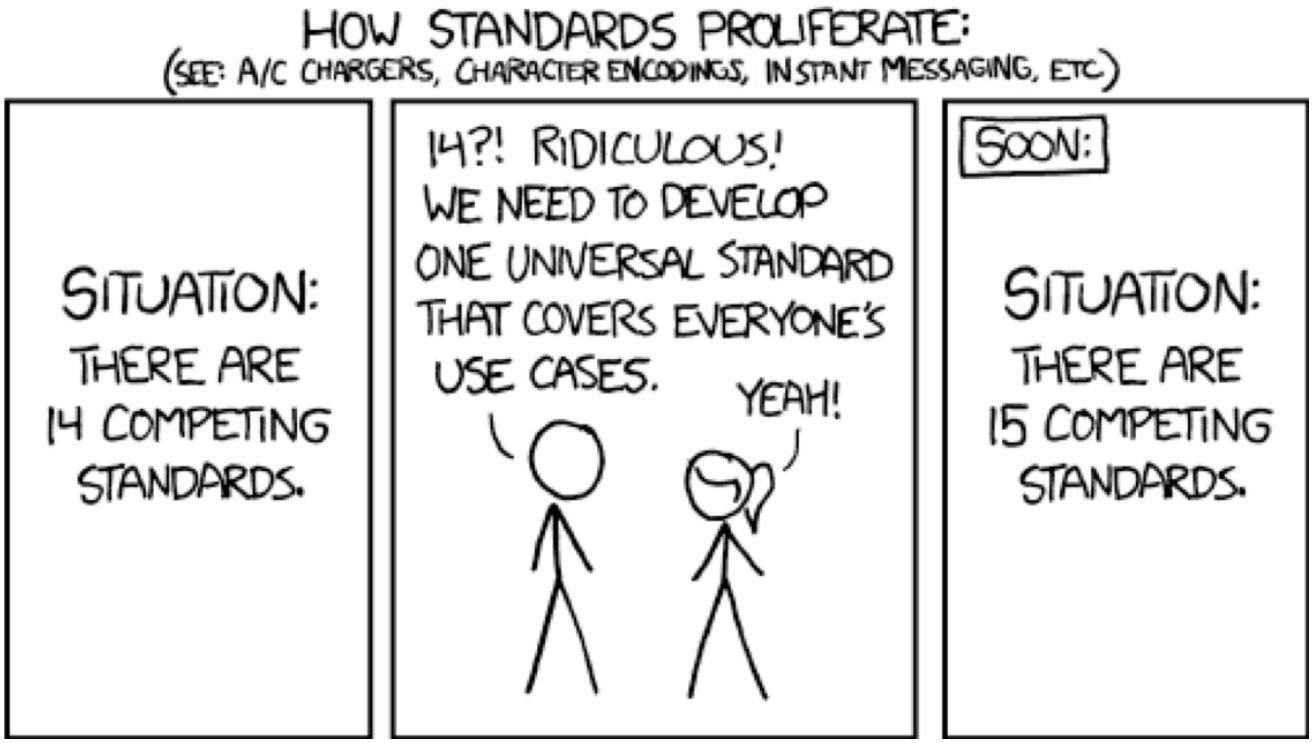
Gene Annotation Format (GTF)

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

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

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

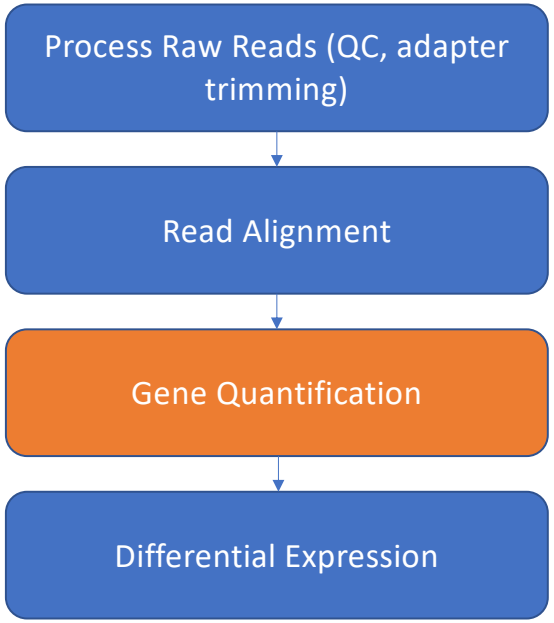
A note on standards



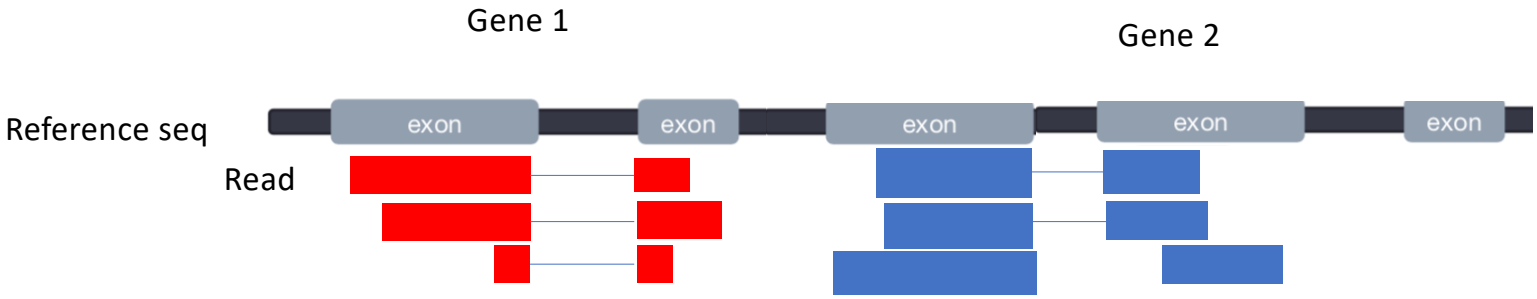
https://xkcd.com/927/

Visualizing reads with JBrowse

Workflow

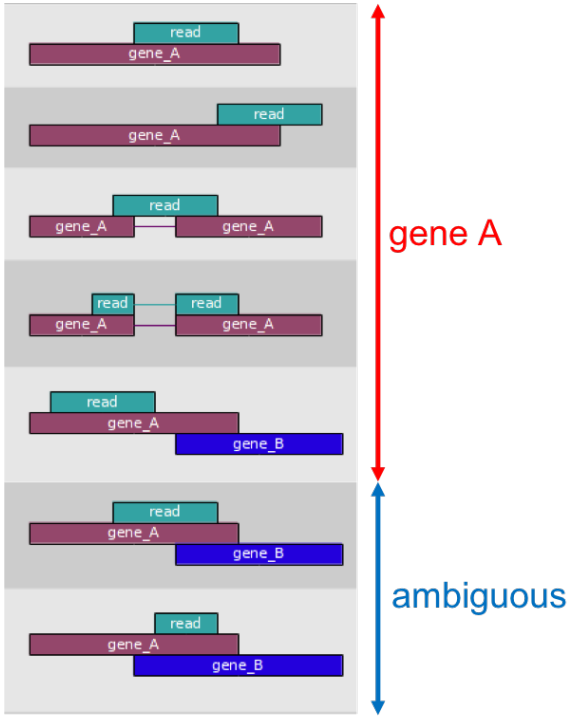

Process Raw Reads (QC, adapter trimming)
Read Alignment
Gene Quantification
Differential Expression

Counting reads for each gene



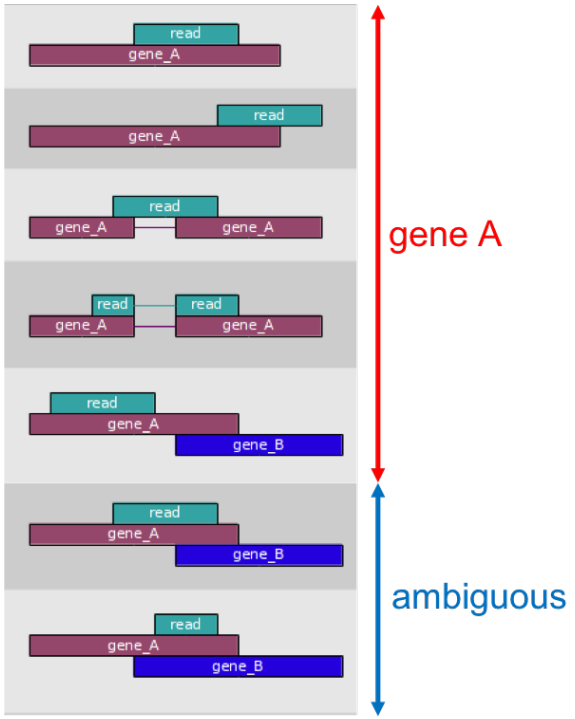
Counting reads: featurecounts

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



Counting reads: featurecounts

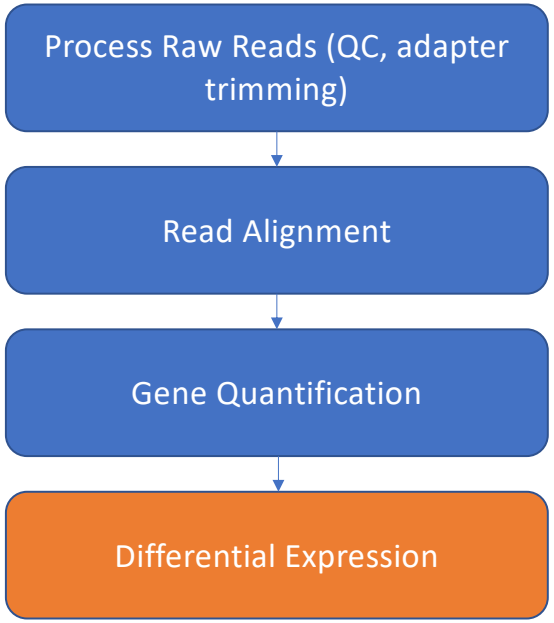
- The mapped coordinates of each read are compared with the features in the GTF file
- Reads that overlap with a gene by >=1 bp are counted as belonging to that feature
- Ambiguous reads will be discarded



Result is a gene count matrix:

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

Workflow

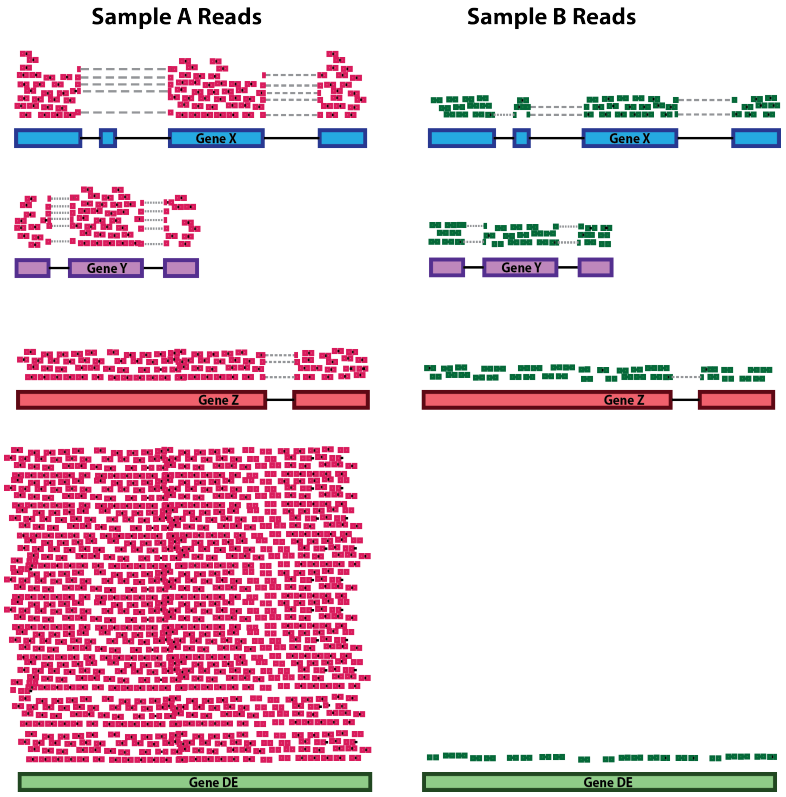

Process Raw Reads (QC, adapter trimming)
Read Alignment
Gene Quantification
Differential Expression

Testing for Differential Expression

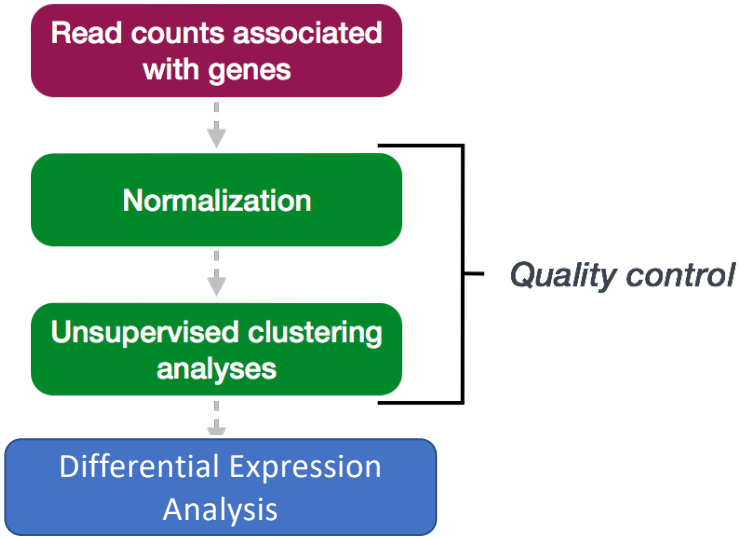
The goal of differential expression analysis (DE) is to find gene differences between conditions, developmental stages, treatments etc.

In particular DE has two goals:

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



Differential Expression with DESeq2



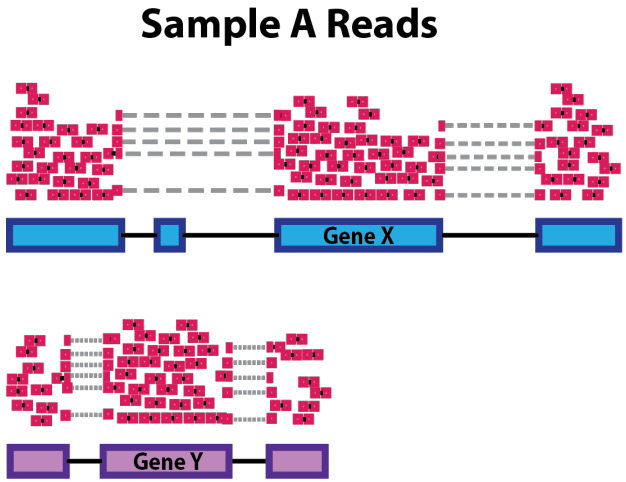
All steps are done with one click in Galaxy!

https://hbctraining.github.io/DGE_workshop

Normalization

The number of sequenced reads mapped to a gene depends on

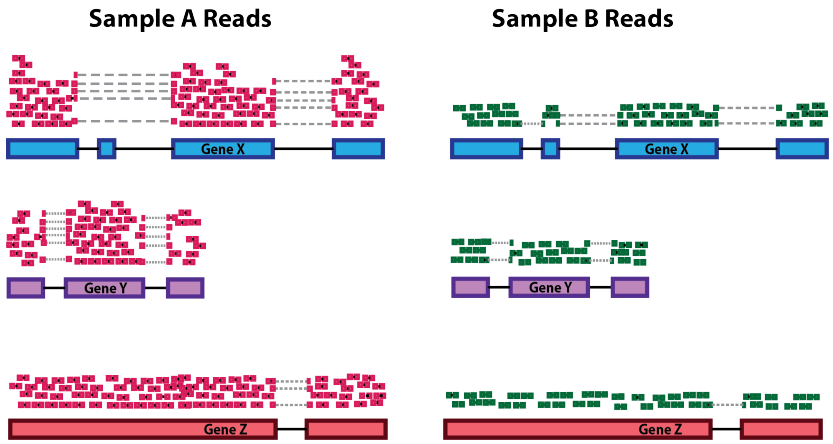
- Gene Length



Normalization

The number of sequenced reads mapped to a gene depends on

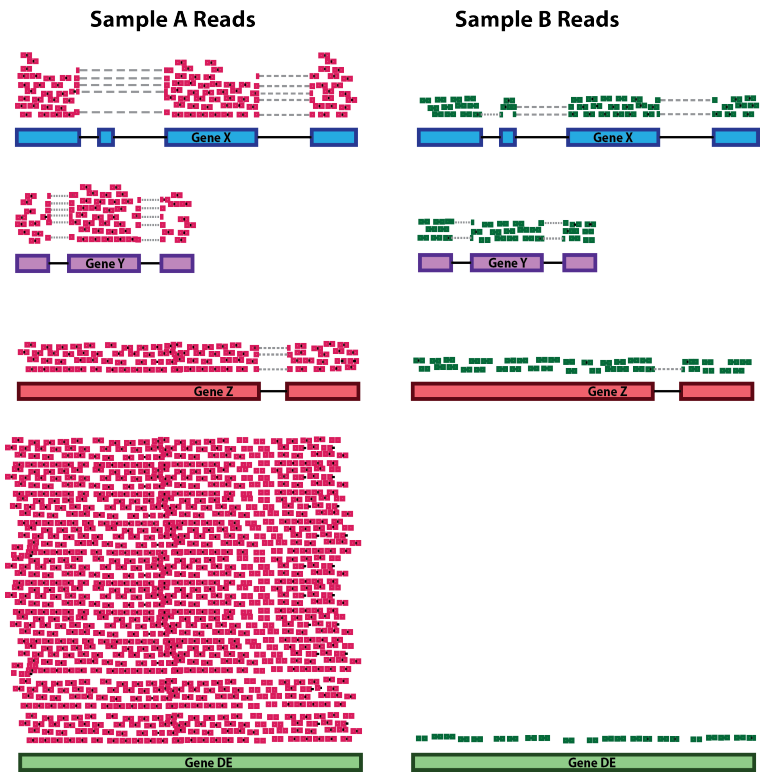
- Gene Length
- Sequencing depth



Normalization

The number of sequenced reads mapped to a gene depends on

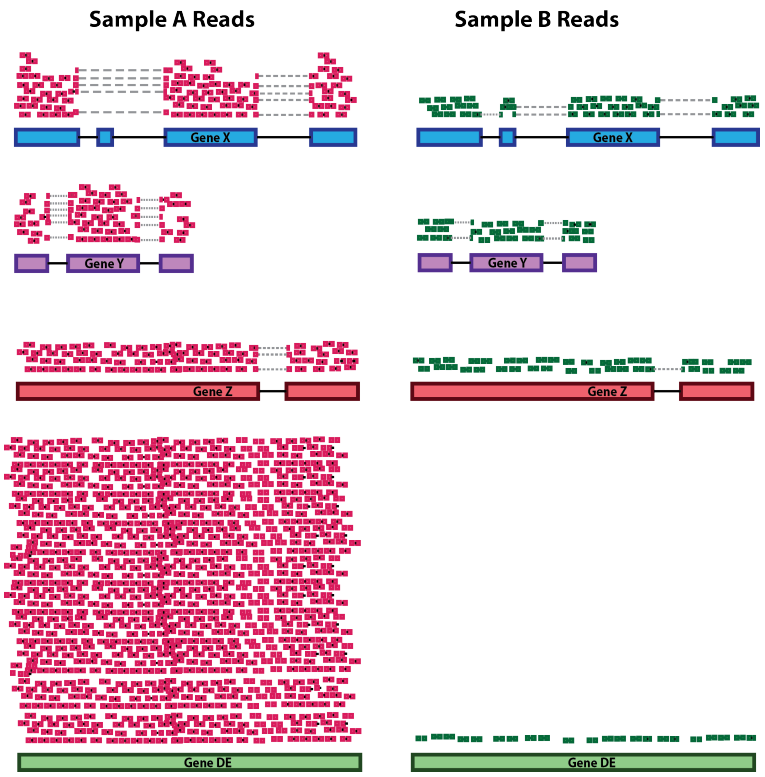
- Gene Length
- Sequencing depth
- The expression level of other genes in the sample



Normalization

The number of sequenced reads mapped to a gene depends on

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



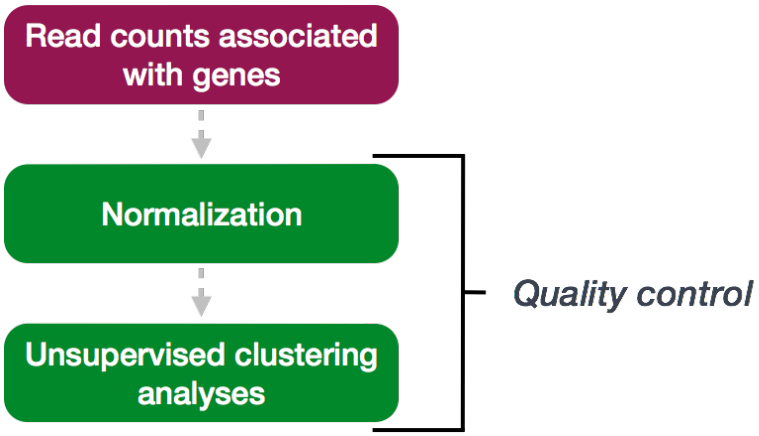
Normalization eliminates the factors that are not of interest!

Normalization methods

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

https://hbctraining.github.io/DGE_workshop

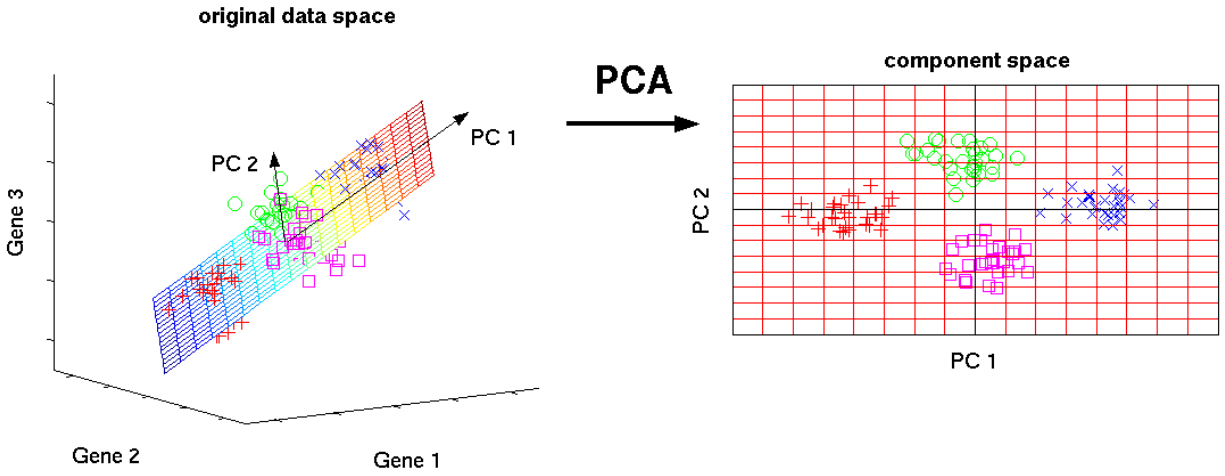
Unsupervised Clustering



Principle Component Analysis

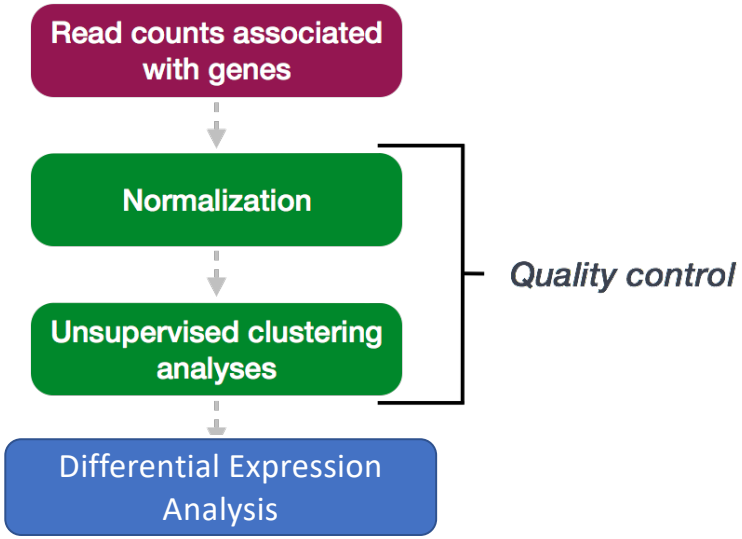
Here is an example with three genes measured in many samples:

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



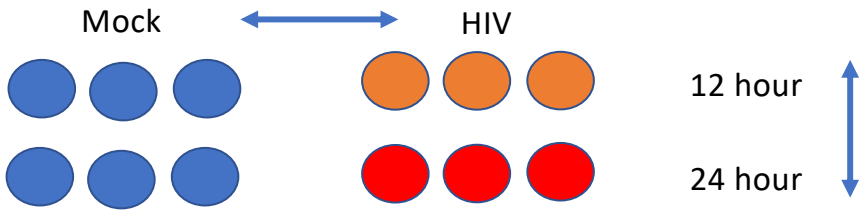
Do your samples cluster as expected?

Differential Expression with DESeq2



https://hbctraining.github.io/DGE_workshop

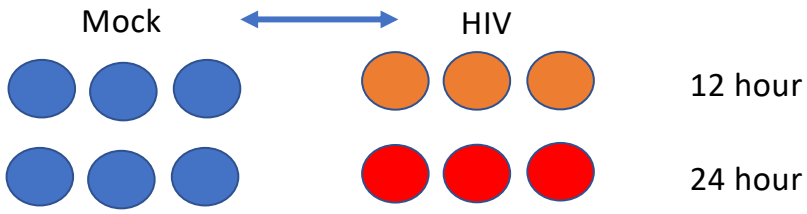
Multi-factor experiment design



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

DESeq2 Test for Differential Expression

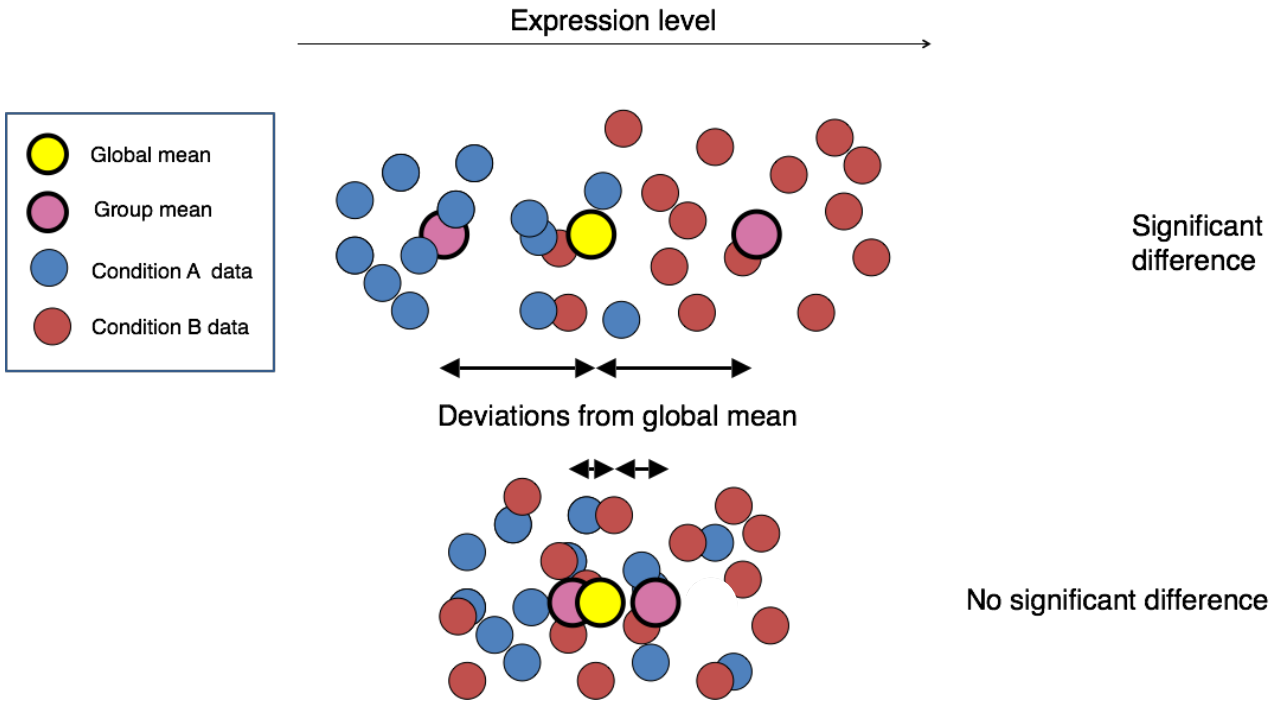


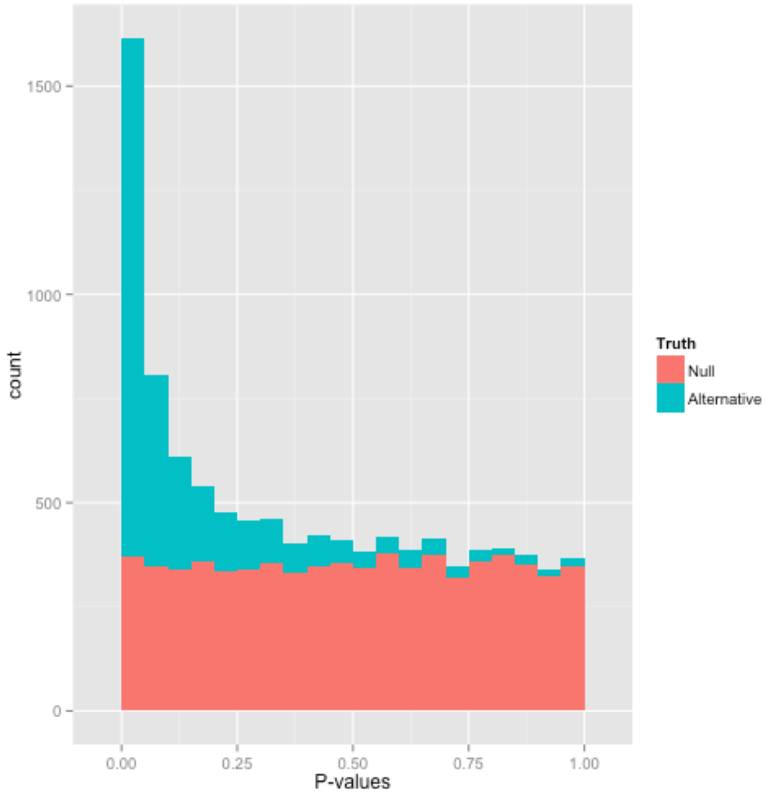
Image credit: Paul Pavlidis, UBC
https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html

DESeq2 Results table

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

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

DESeq2 P-value histogram



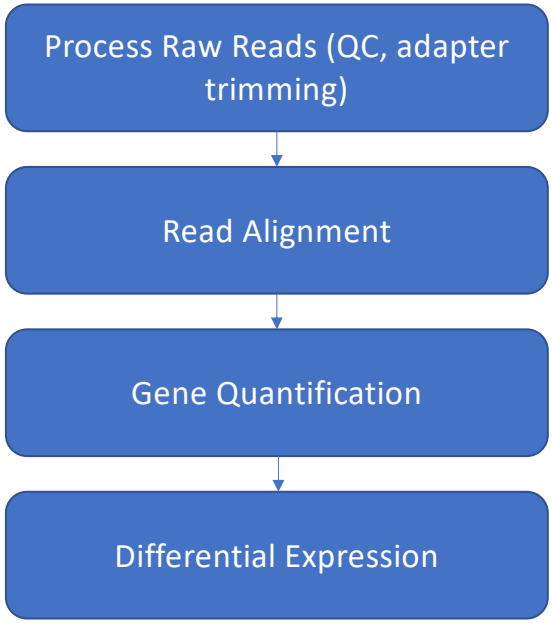
- Histogram of raw p-values for all genes examined
- P-value: Probability of getting a logFC as extreme as observed if the true logFC = 0 for that gene (null hypothesis)

How to interpret:

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

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

Conclusions



References

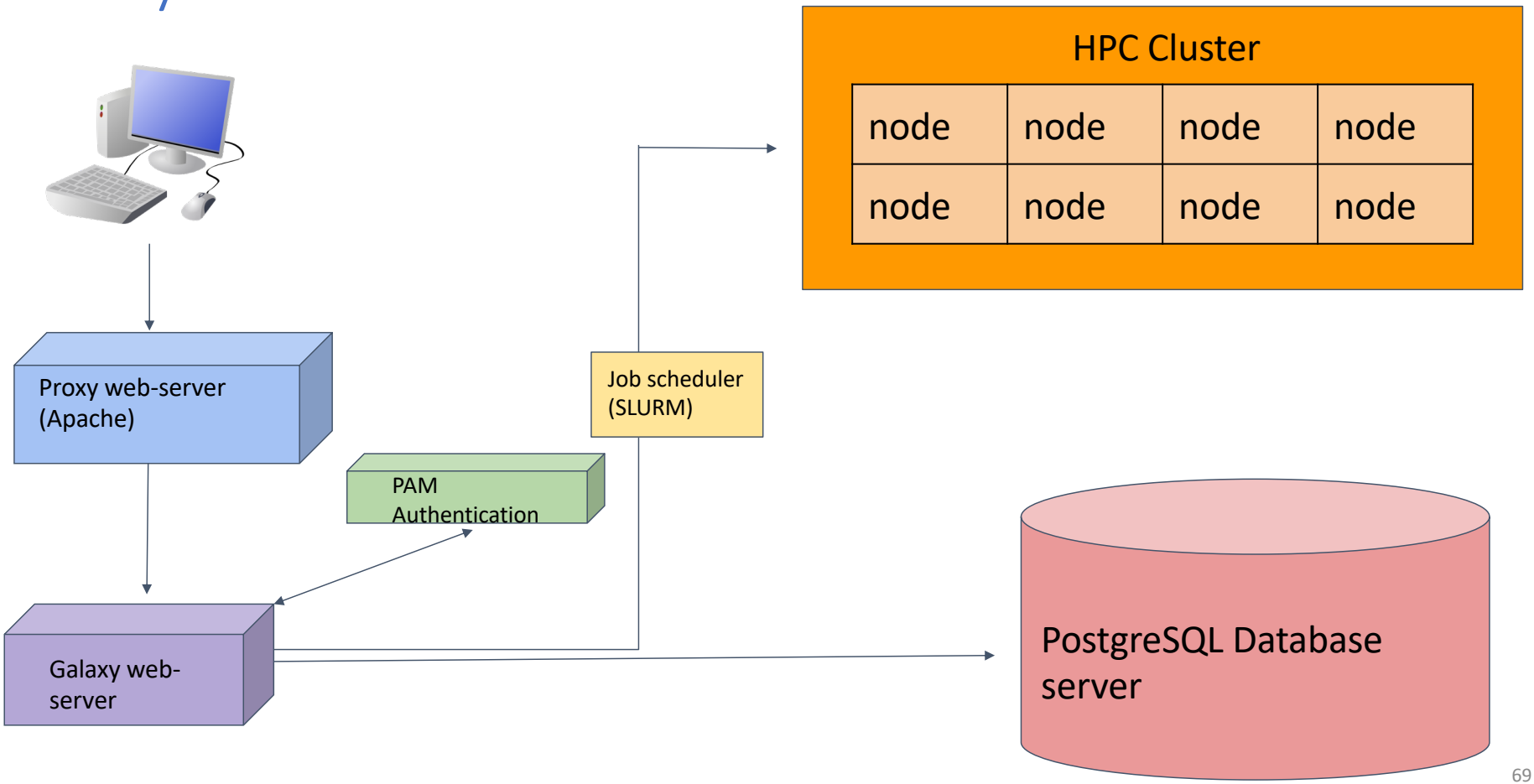
https://www.bioconductor.org/packages/3.3/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

https://hbctraining.github.io/DGE_workshop

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

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

Galaxy on the Tufts HPC



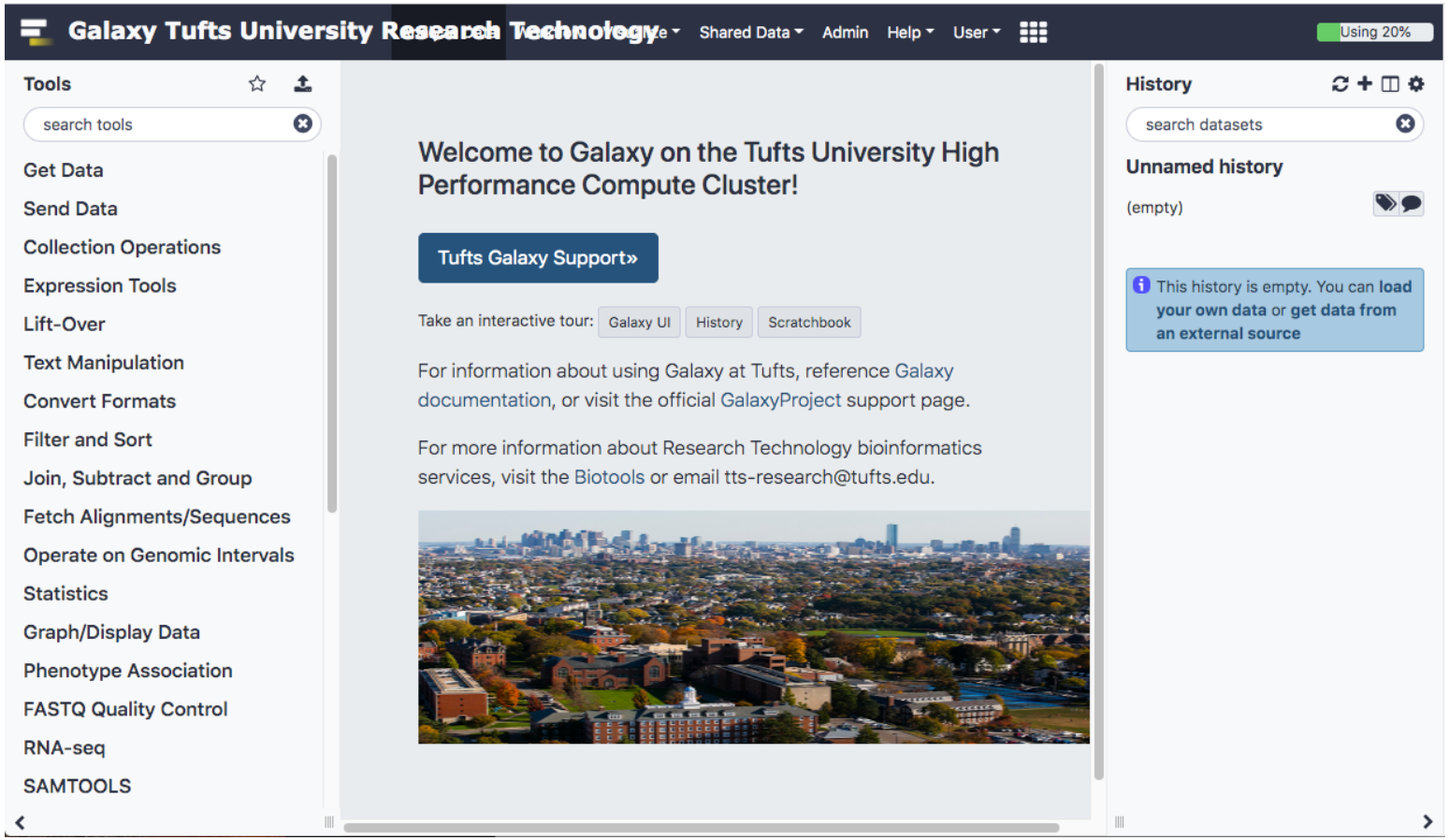
Access Galaxy

You must be connected to Tufts Network, either on campud or via VPN

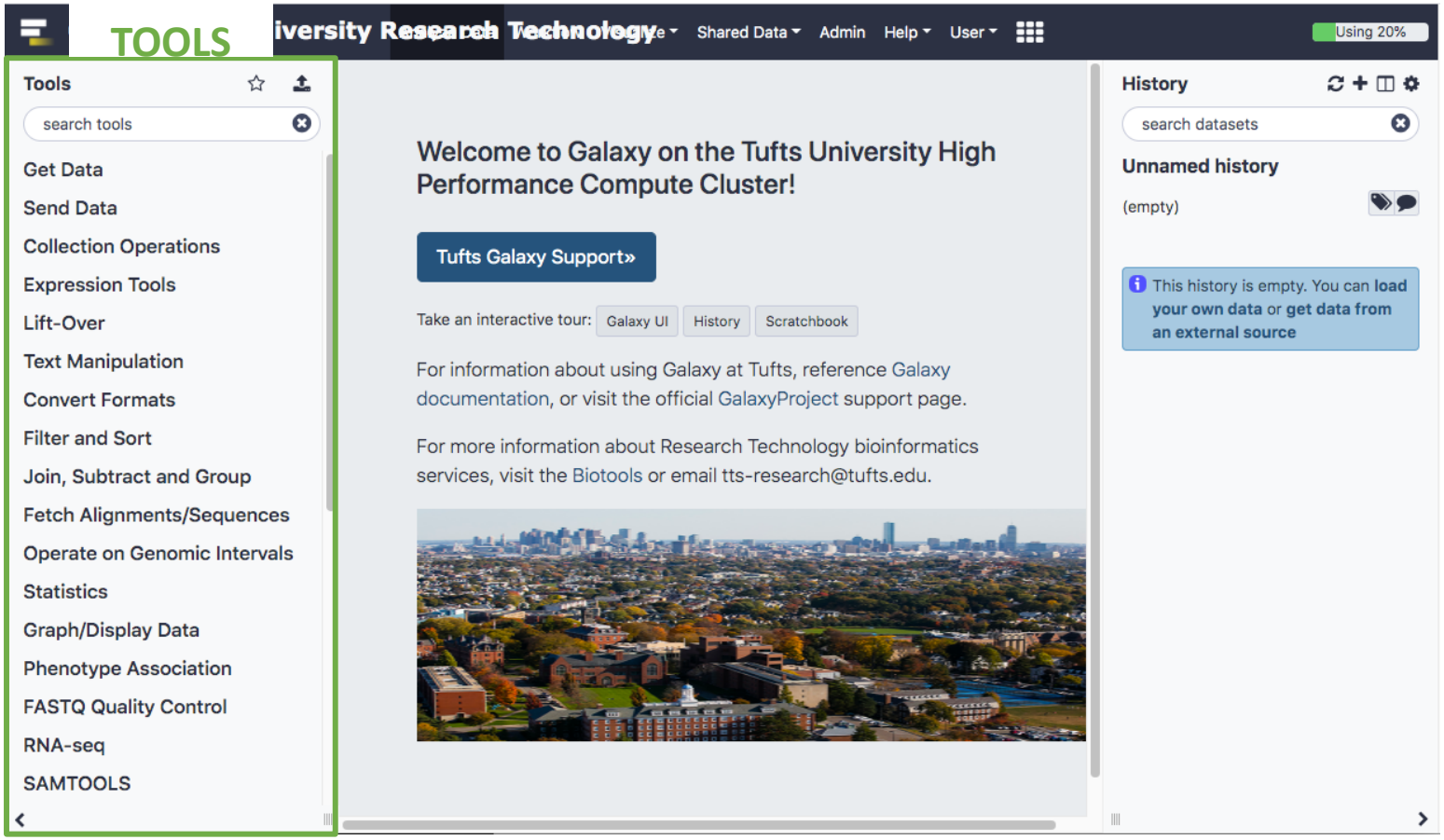
https://galaxy.cluster.tufts.edu/

Login with you cluster username and password

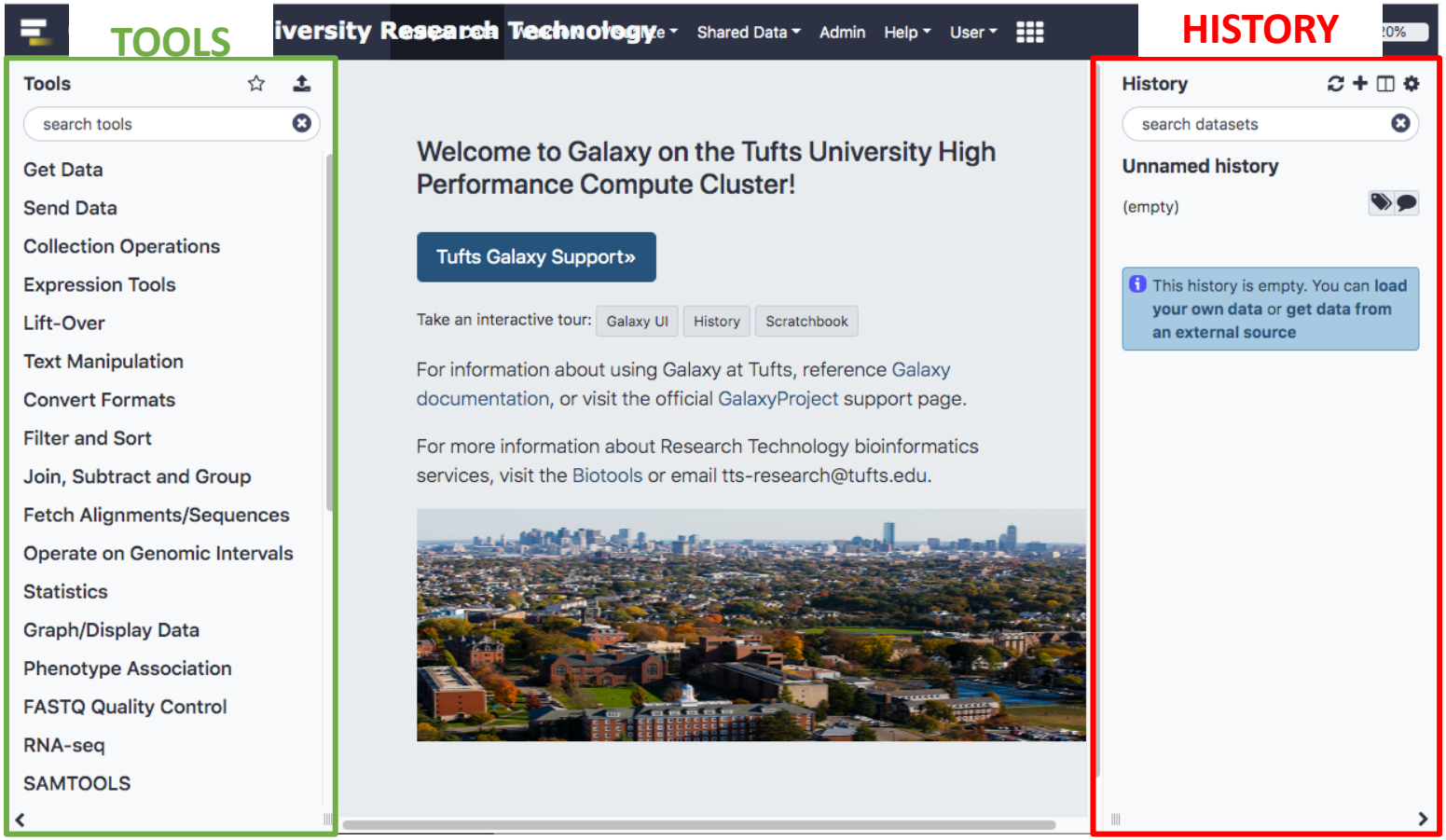
User Interface



User Interface

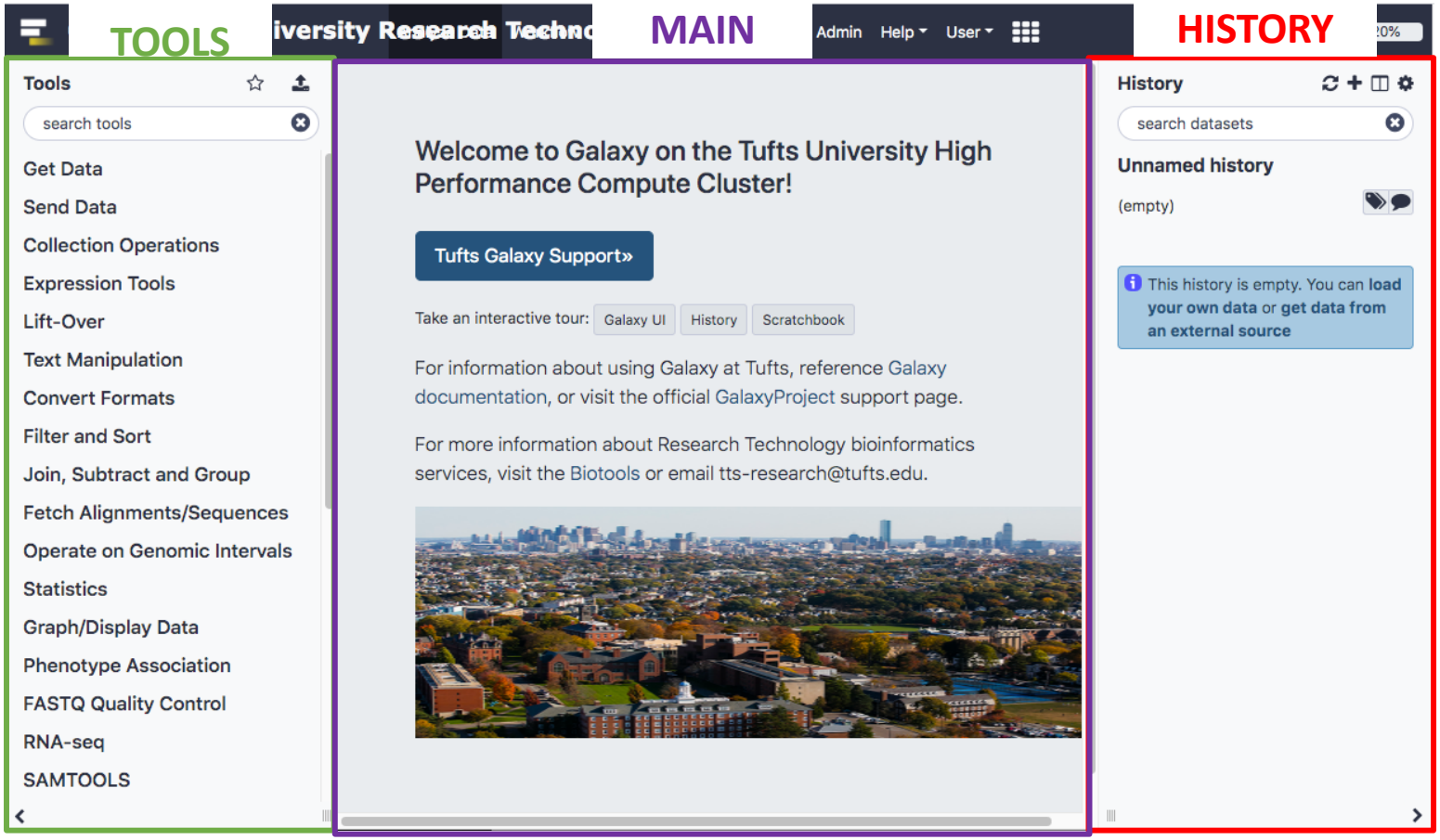


User Interface



User Interface

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



Start the workflow

Visit the workshop page at
https://rbatorsky.github.io/intro-to-rnaseq-with-galaxy/