From FastQ to Counts

Workshop by Jared Bard
jbard@uchicago.edu
230914

Other Resources

https://hbctraining.github.io/main/

http://mcb112.org/

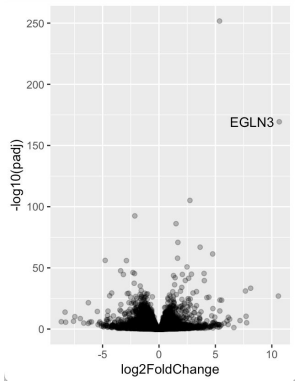
Goals for the Workshop

- Understand the logic of processing raw sequencing data
- Know the tools you can use
- Extract counts from sample data using galaxy

Goals for the Workshop

- Understand the logic of processing raw sequencing data
- Know the tools you can use
- Extract counts from sample data using galaxy

sample_1.fastq.gz
sample_2.fastq.gz



Download Class files

bit.ly/workshop230914

Files (106.2 MB)

Name	Size	
workshop.zip	106.2 MB	Preview Download

For real data, use the command line

- Goal is accuracy **and** reproducibility
- I recommend: **snakemake** or **nextflow**

Tell Snakemake what files you want to be created

```
rule:
  input: "A.txt", "B.txt", "C.txt"
```

Produce the files you want to have from some intermediate result

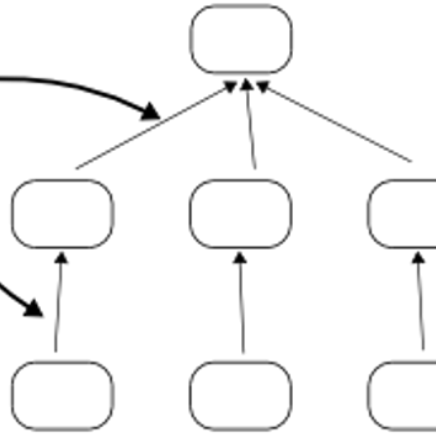
```
rule:
  input: "{sample}.inter"
  output: "{sample}.txt"
  shell: "somecommand {input} {output}"
```

Create a needed intermediate result

```
rule:
  input:  "{sample}.in"
  output: "{sample}.inter"
  run:
    somepythoncode()
```

Use wildcards to write general rules for all samples

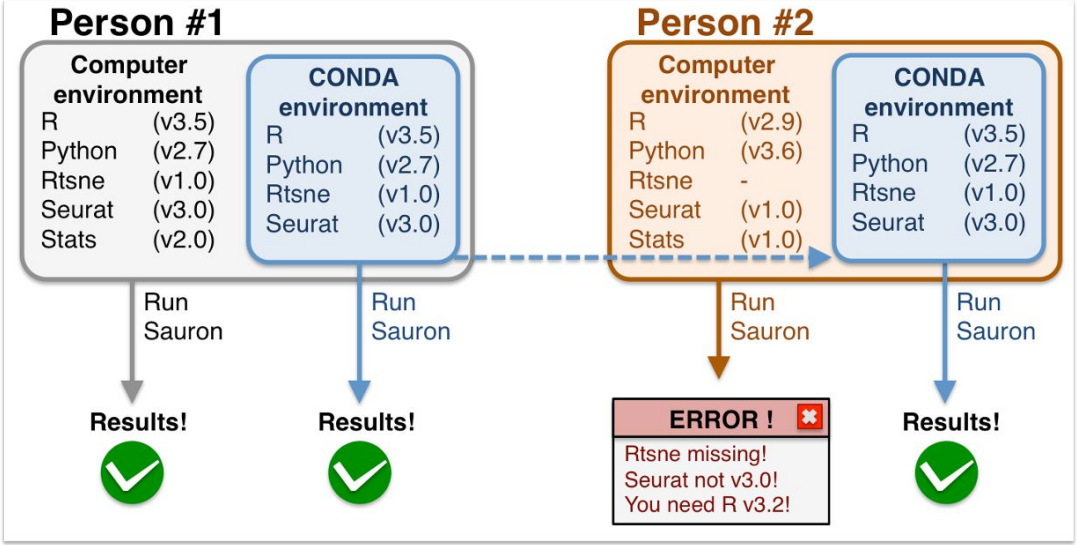
Snakemake determines the dependencies for you



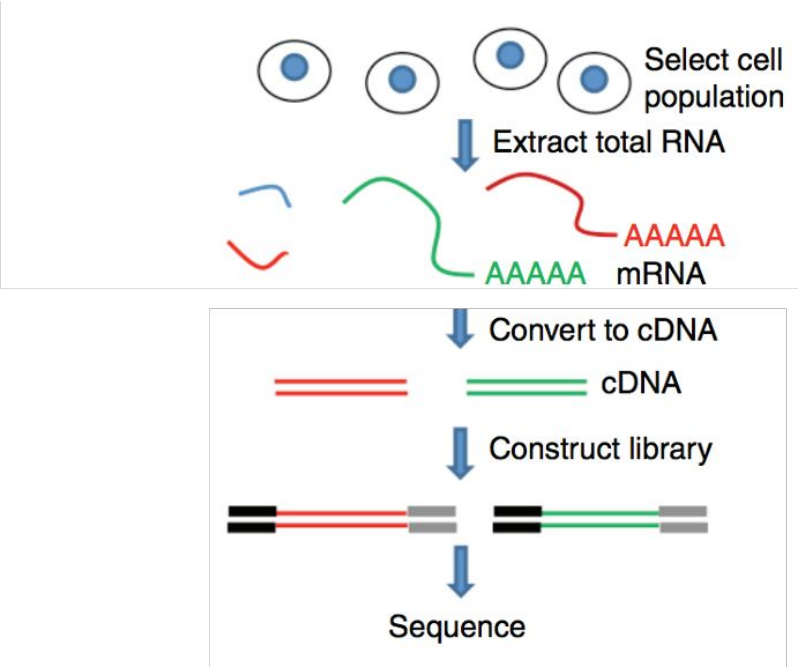
Documented, reproducible analysis. Easy to modify individual steps for troubleshooting.

For real data, use package managers

- Rules use conda/mamba to create environments
- https://github.com/mamba-org/mamba

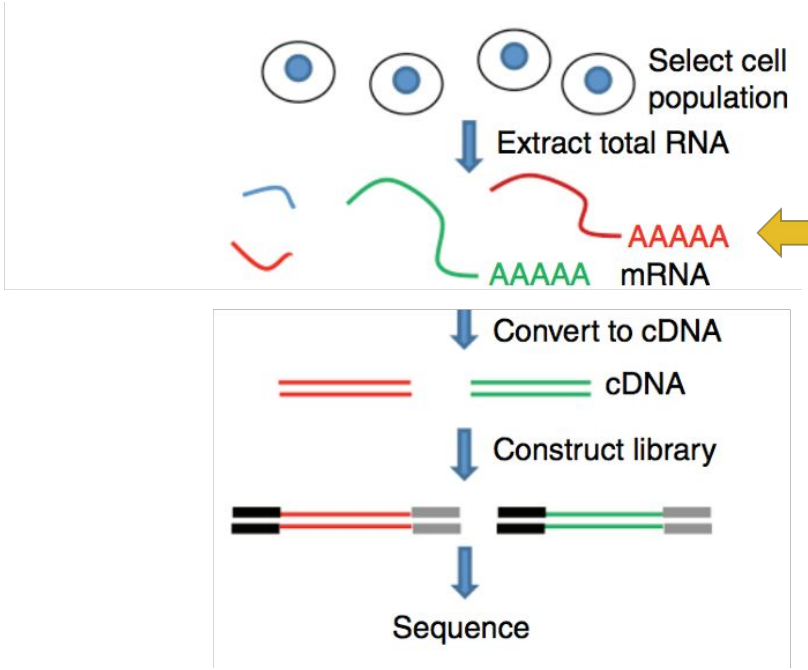


What is RNA-seq?



https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/01a_RNAseq_processing_workflow.html

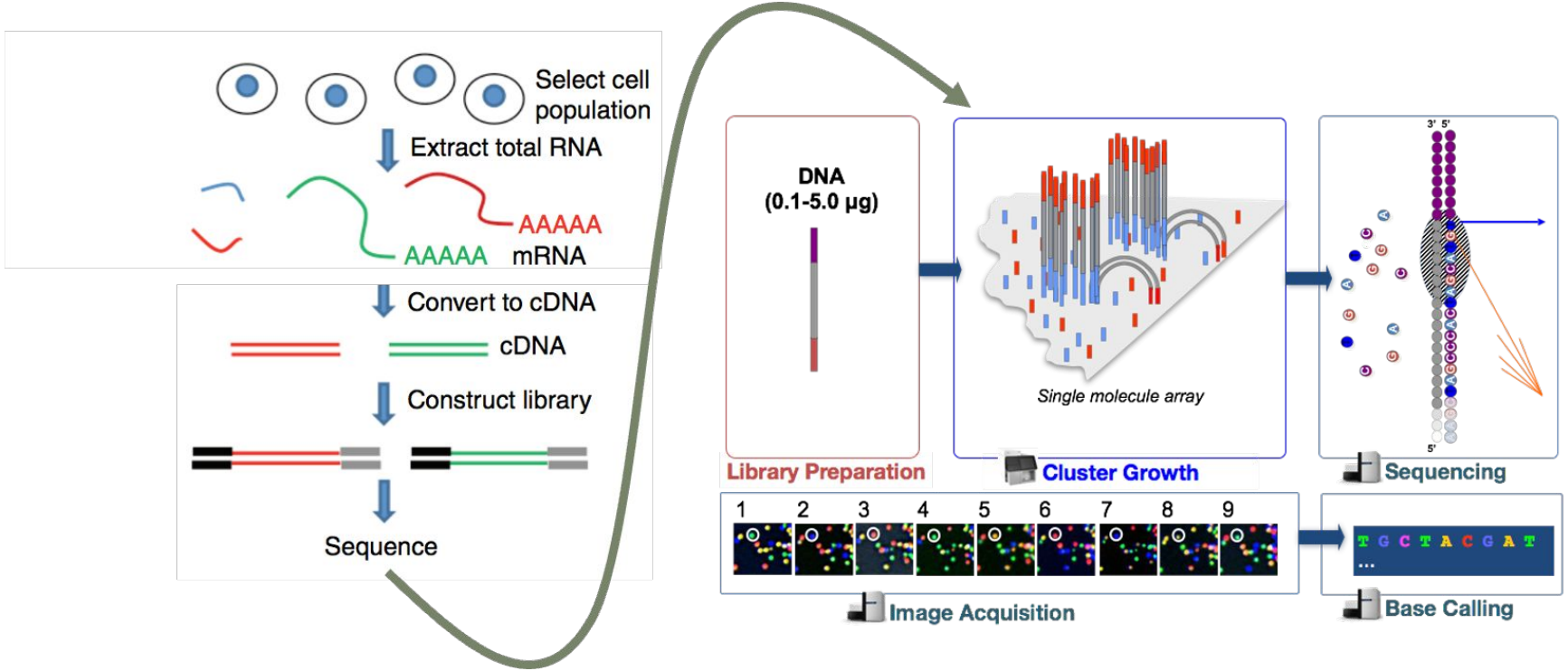
What is RNA-seq?



95% of total RNA is rRNA:
Use rRNA depletion or poly(A) enrichment

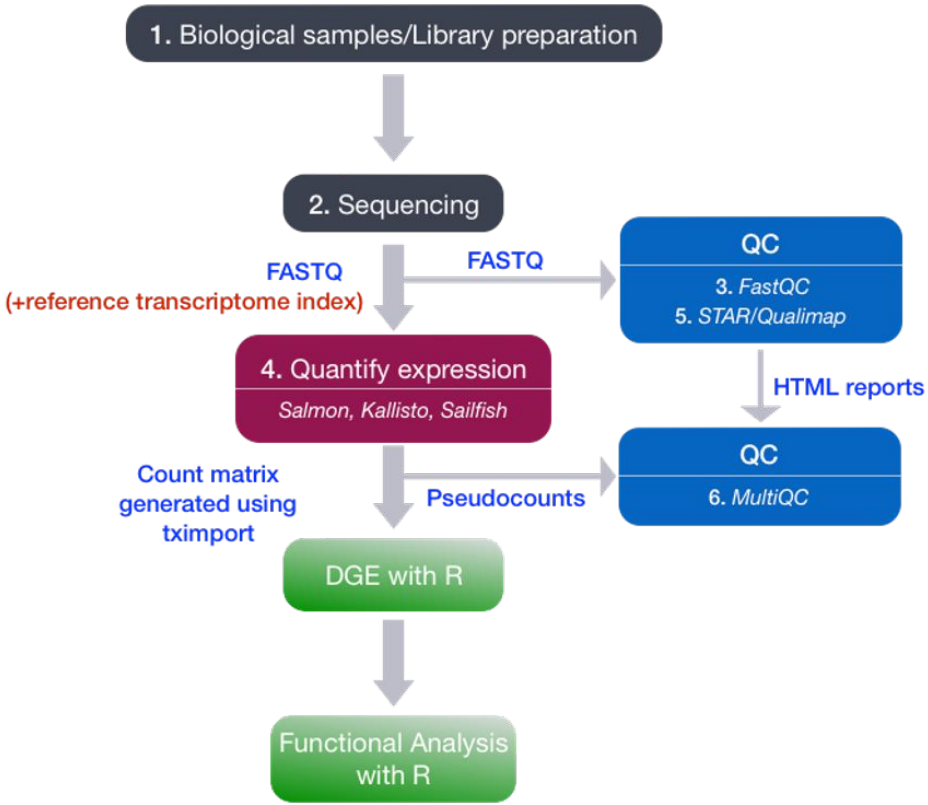
https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/01a_RNAseq_processing_workflow.html

What is RNA-seq?



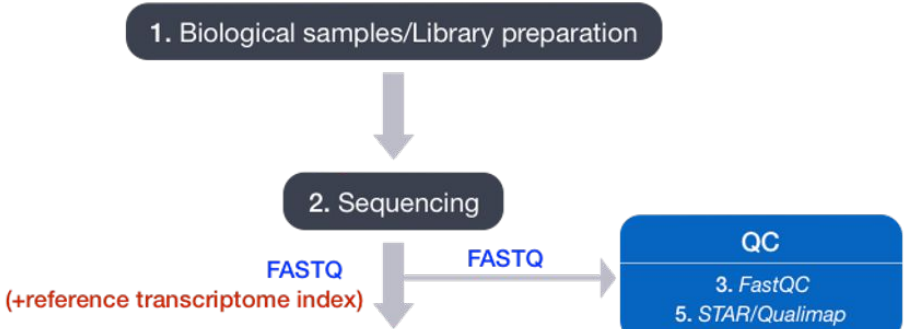
https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/01a_RNAseq_processing_workflow.html

RNA-Seq Workflow



https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/01a_RNAseq_processing_workflow.html

RNA-Seq Workflow

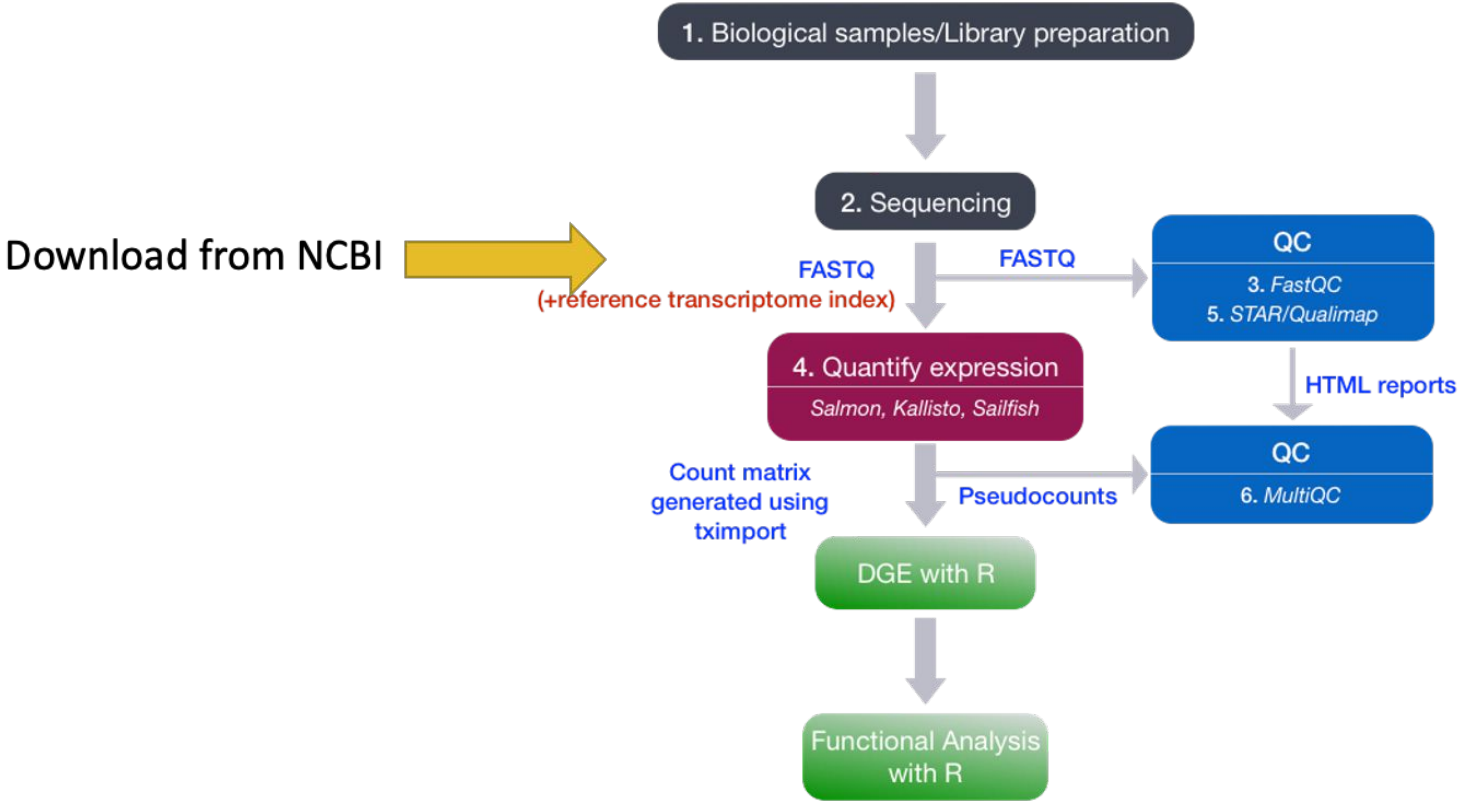


fastq:

```
@HWI-ST330:304:H045HADXX:1:1101:1111:61397
CACTTGTAAGGGCAGGCCCCCTTCACCCTCCCGCTCCTGGGGGANNNNNNNNNNANNNCGAGGCCCTGGGGTAGAGGGNNNNNNNNNNNNNNGATCTTGG
+
@?@DDDDDDHHH?GH:?FCBGGB@C?DBEGIIIIAEF;FCGGI##################################################
```

https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/01a_RNAseq_processing_workflow.html

RNA-Seq Workflow



https://hbctraining.github.io/DGE_workshop_salmon_online/lessons/01a_RNAseq_processing_workflow.html

Sci Adv. 2022 Nov; 8(47): eade1942.
Published online 2022 Nov 23. doi: 10.1126/sciadv.ade1942

PMCID: PMC9683707
PMID: 36417539

A pleiotropic hypoxia-sensitive *EPAS1* enhancer is disrupted by adaptive alleles in Tibetans

OLIVIA A. GRAY, JENNIFER YOO, DÉBORA R. SOBREIRA, JORDAN JOUSMA, DAVID WITONSKY, NOBORU J. SAKABE, YING-JIE PENG, NANDURI R. PRABHAKAR, YUN FANG, MARCELO A. NOBRÉGA, AND ANNA DI RIENZO

fewer Authors Info & Affiliations

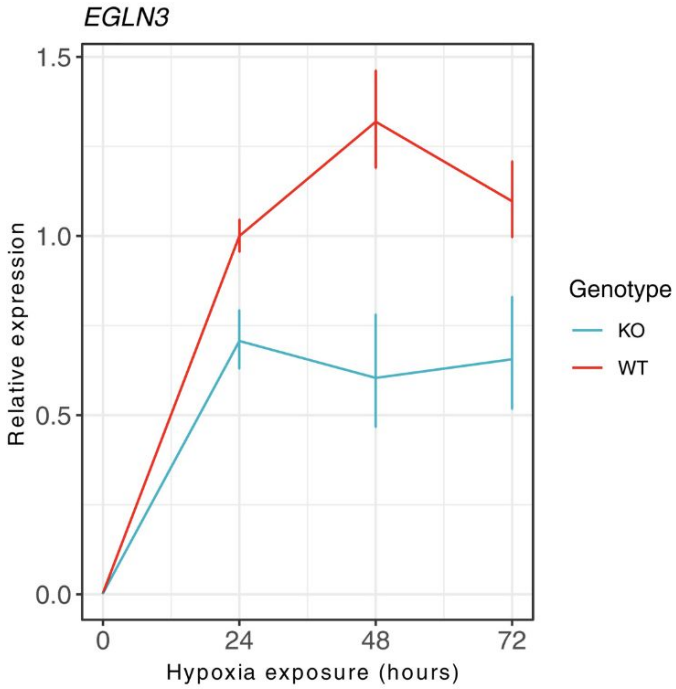
https://pubmed.ncbi.nlm.nih.gov/36417539/

Data and materials availability: Data described in this paper can be found in the Supplementary Materials or in publicly available databases including the UCSC genome browser and Epigenome Roadmap. All sequencing data are deposited in the GEO accession database (www.ncbi.nlm.nih.gov/geo/); accession no. GSE197527 (HAEC ATAC-seq/RNA-seq: GSE197523; teloHAEC RNA-seq: GSE197525; Capture Hi-C: GSE197526, mouse RNA-seq: GSE197524).

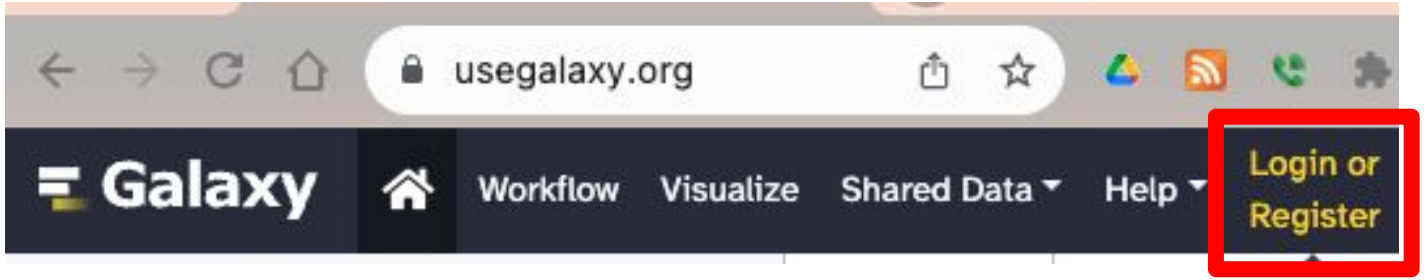
Samples (12)
Less...

GSM5919529	KO3.N
GSM5919530	KO4.N
GSM5919531	KO5.N
GSM5919532	WT4.N
GSM5919533	WT5.N
GSM5919534	WT6.N

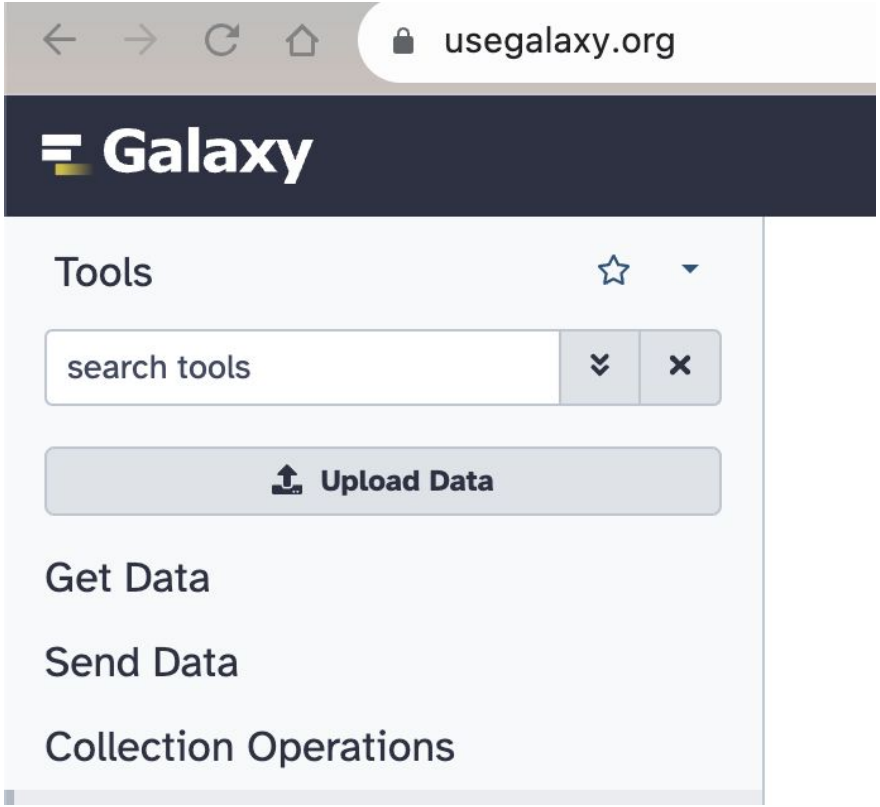
Fig. 3. Deletion of ENH5 in endothelial cells results in dampening of transcriptional responses to acute and sustained hypoxia.



Make a galaxy account



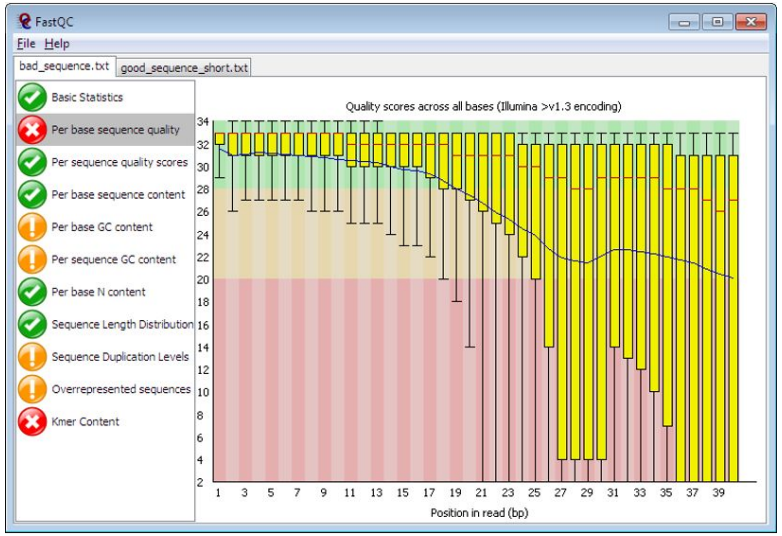
Upload data (everything in the zip)



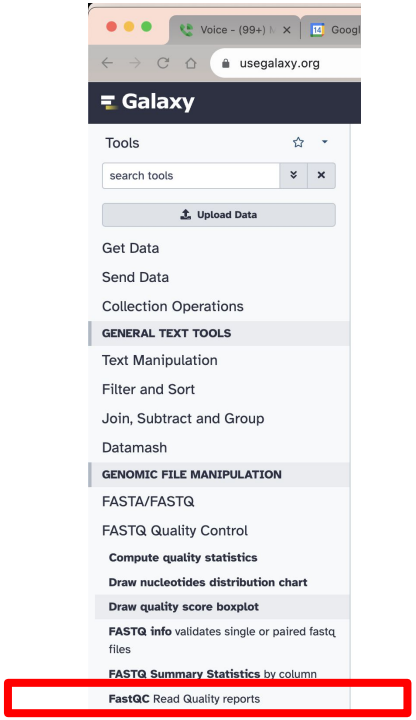
Quality control

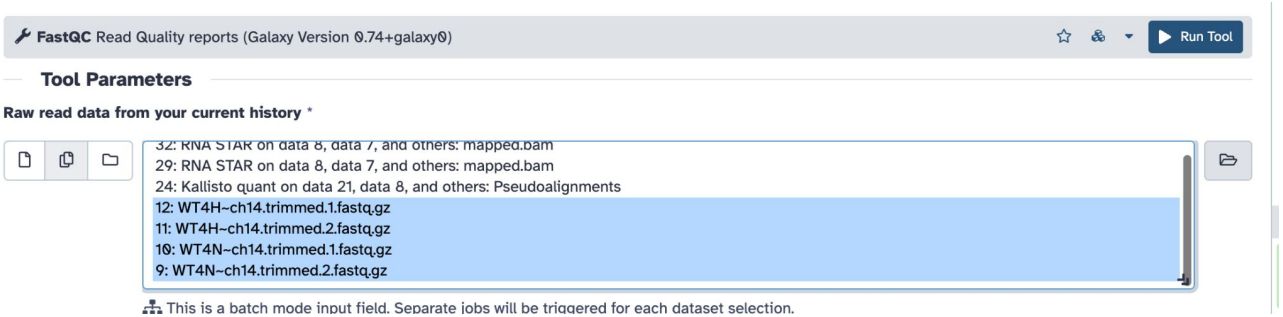
fastqc

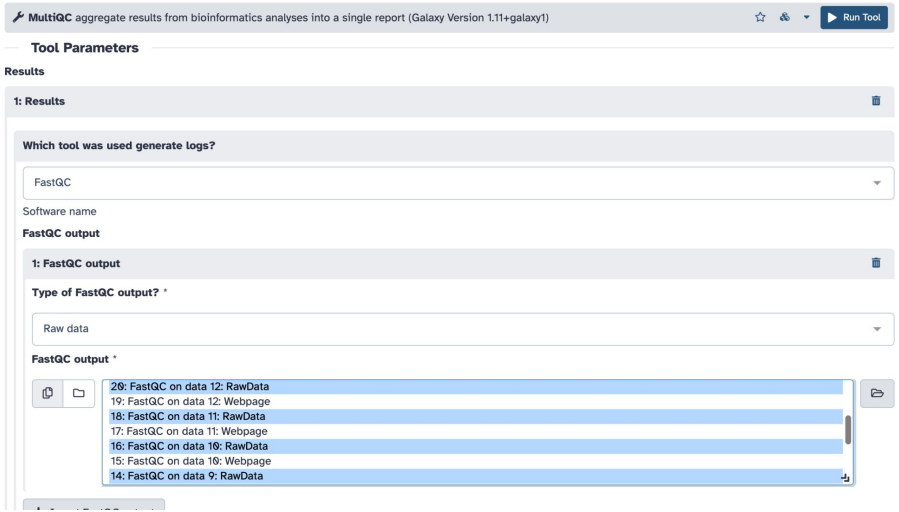
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/



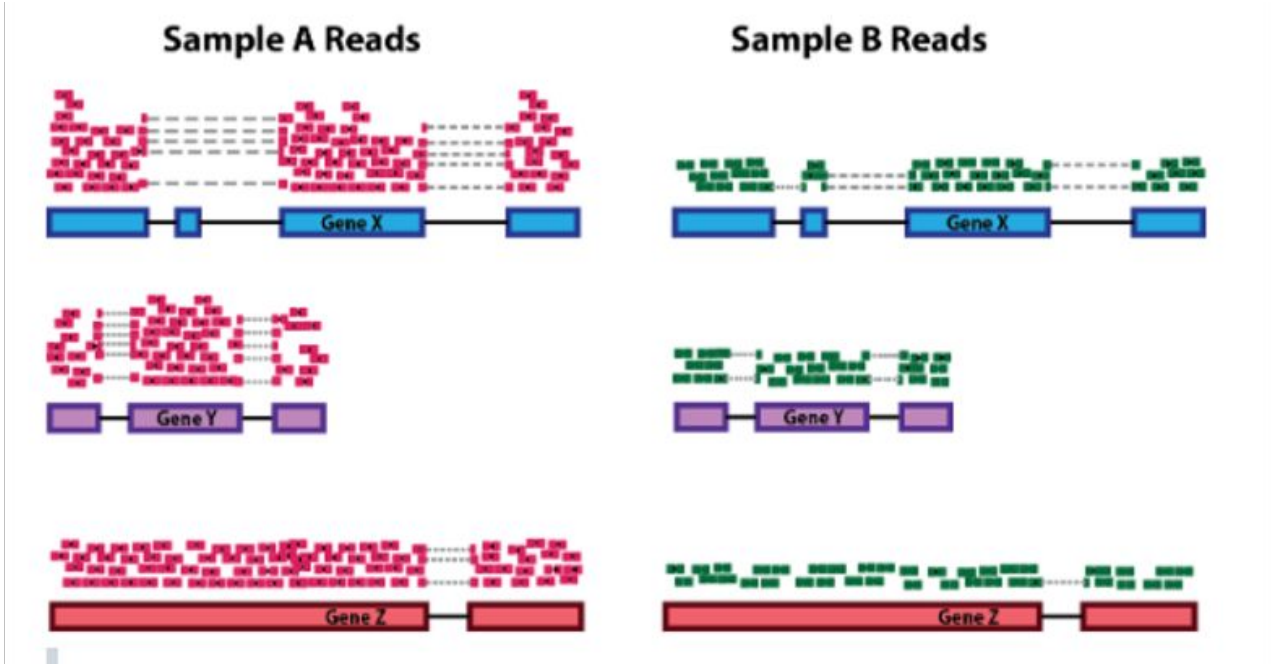
Run FastQC


usegalaxy.org
Galaxy
Tools
search tools
Upload Data
Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Datamash
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
Compute quality statistics
Draw nucleotides distribution chart
Draw quality score boxplot
FASTQ info validates single or paired fastq files
FASTQ Summary Statistics by column
FastQC Read Quality reports


FastQC Read Quality reports (Galaxy Version 0.74+galaxy0)
Run Tool
Tool Parameters
Raw read data from your current history *
29: RNA STAR on data 8, data 7, and others: mapped.bam
24: Kallisto quant on data 21, data 8, and others: Pseudoalignments
12: WT4H~ch14.trimmed.1.fastq.gz
11: WT4H~ch14.trimmed.2.fastq.gz
10: WT4N~ch14.trimmed.1.fastq.gz
9: WT4N~ch14.trimmed.2.fastq.gz
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.


MultiQC aggregate results from bioinformatics analyses into a single report (Galaxy Version 1.11+galaxy1)
Run Tool
Tool Parameters
Results
1: Results
Which tool was used generate logs?
FastQC
Software name
FastQC output
1: FastQC output
Type of FastQC output? *
Raw data
FastQC output *
20: FastQC on data 12: RawData
19: FastQC on data 12: Webpage
18: FastQC on data 11: RawData
17: FastQC on data 11: Webpage
16: FastQC on data 10: RawData
15: FastQC on data 10: Webpage
14: FastQC on data 9: RawData

Mapping:



https://github.com/hbctraining/DGE_workshop_salmon_online/blob/master/lessons/02_DGE_count_normalization.md

Mapping: pseudoalignment vs genomic alignment

kallisto/salmon https://pachterlab.github.io/kallisto/
-aligns to transcripts (not genome)
-better at isoforms

STAR/hisat2
-aligns to genome (splice-aware)

Annotations: https://useast.ensembl.org/index.html

https://github.com/hbctraining/Training-modules/blob/master/DGE-functional-analysis/lessons/AnnotationHub.md

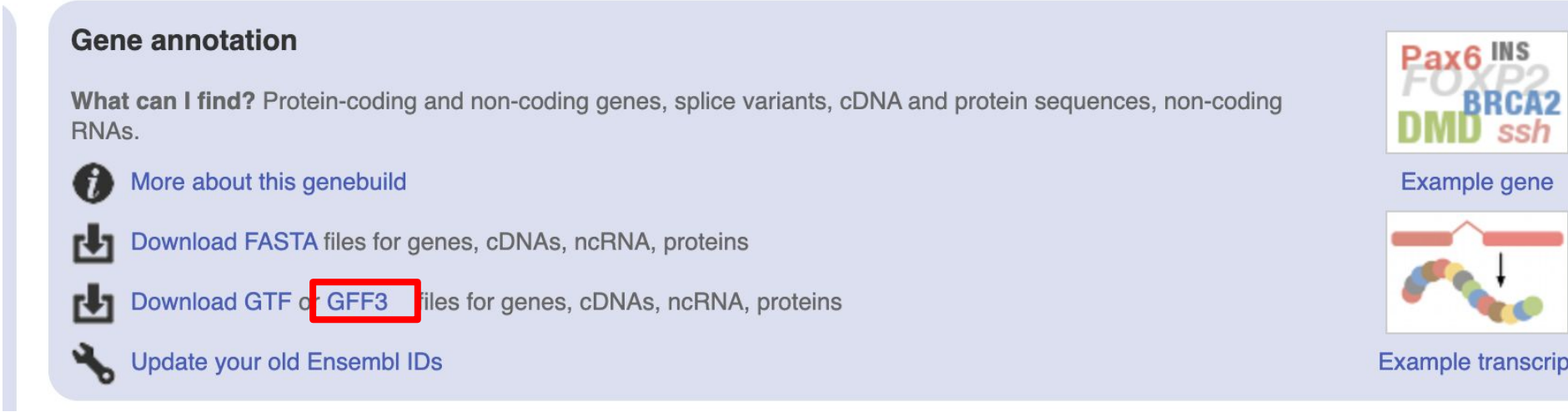
- Genome:
-

https://hbctraining.github.io/Accessing_public_genomic_data/lessons/accessing_genome_reference_data.html

Annotations: https://useast.ensembl.org/index.html

https://github.com/hbctraining/Training-modules/blob/master/DGE-functional-analysis/lessons/AnnotationHub.md

- Genome: Homo_sapiens.GRCh38.dna.chr14.fa
- GFF: Homo_sapiens.GRCh38.110~chr14.gff3
-



https://hbctraining.github.io/Accessing_public_genomic_data/lessons/accessing_genome_reference_data.html

Annotations: https://useast.ensembl.org/index.html

https://github.com/hbctraining/Training-modules/blob/master/DGE-functional-analysis/lessons/AnnotationHub.md

- Genome: Homo_sapiens.GRCh38.dna.chr14.fa
- GFF: Homo_sapiens.GRCh38.110~chr14.gff3
- Transcripts: Homo_sapiens.GRCh38.cdna.chr14.fa
  - cdna = entier transcripts
  - cds = no UTRs

Index of /pub/release-110/fast

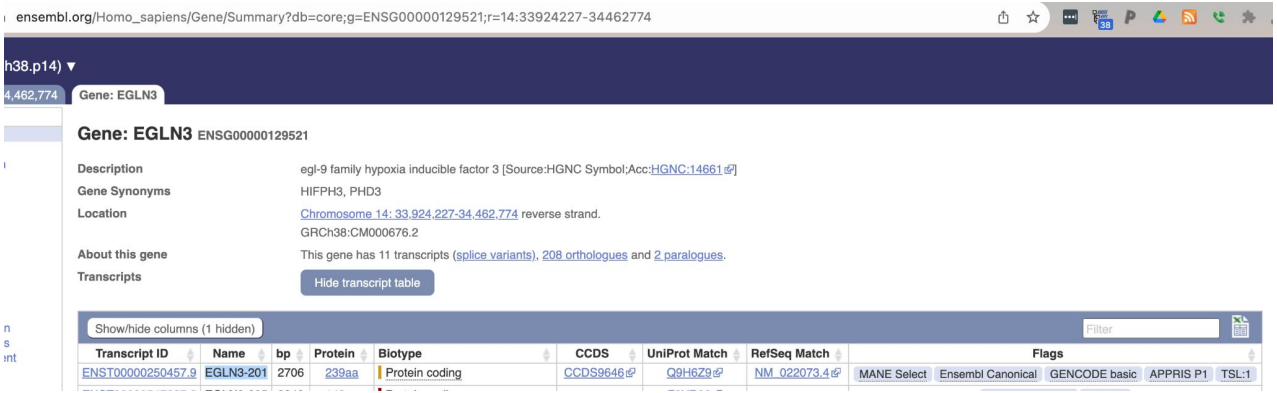
Name	Last modified	Size	Description
Parent Directory		-	
cdna/	2023-05-04 05:26	-	
cds/	2023-05-04 05:26	-	
dna/	2023-05-04 05:27	-	
dna_index/	2023-05-04 05:27	-	
ncrna/	2023-05-04 05:27	-	
pep/	2023-05-04 05:27	-	

https://hbctraining.github.io/Accessing_public_genomic_data/lessons/accessing_genome_reference_data.html

Annotations: https://useast.ensembl.org/index.html

https://github.com/hbctraining/Training-modules/blob/master/DGE-functional-analysis/lessons/AnnotationHub.md

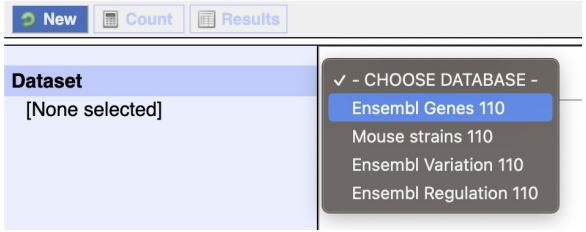
- Genome: Homo_sapiens.GRCh38.dna.chr14.fa
- GFF: Homo_sapiens.GRCh38.110~chr14.gff3
- Transcripts: Homo_sapiens.GRCh38.cdna.chr14.fa
  - cdna = entier transcripts
  - cds = no UTRs

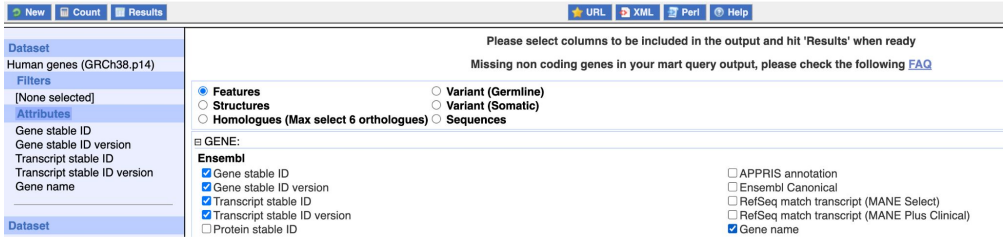


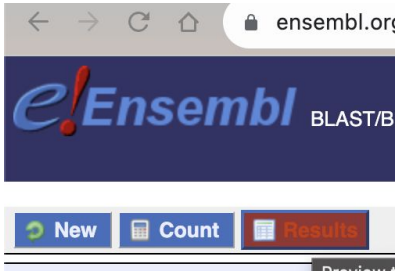
https://hbctraining.github.io/Accessing_public_genomic_data/lessons/accessing_genome_reference_data.html

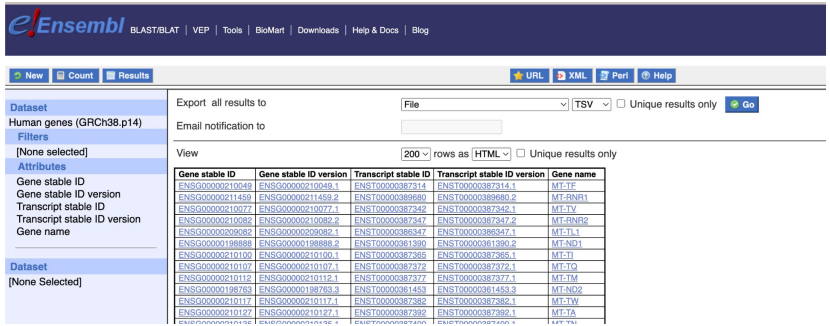
Annotations (mapping transcript IDs to useful names):

https://www.ensembl.org/info/data/biomart/index.html









Try this link for humans:

http://www.ensembl.org/biomart/martview/a89f2939f3bb17d1b733b99c97fb5ff2?VIRTUALSCHEMANAME=default&ATTRIBUTES=hsapiens_gene_ensembl.default.feature_page.ensembl_gene_id|hsapiens_gene_ensembl.default.feature_page.ensembl_gene_id_version|hsapiens_gene_ensembl.default.feature_page.ensembl_transcript_id|hsapiens_gene_ensembl.default.feature_page.ensembl_transcript_id_version|hsapiens_gene_ensembl.default.feature_page.external_gene_name&FILTERS=&VISIBLEPANEL=resultspanel

Running STAR for QC

Use genome file (.dna) from ensembl

RNA STAR Gapped-read mapper for RNA-seq data (Galaxy Version 2.7.10b+galaxy4)

Run Tool

Paired-end (as individual datasets)

RNA-Seq FASTQ/FASTA file, forward reads *

10: WT4N~ch14.trimmed.1.fastq.gz

RNA-Seq FASTQ/FASTA file, reverse reads *

9: WT4N~ch14.trimmed.2.fastq.gz

Custom or built-in reference genome

Use reference genome from history and create temporary index

Built-ins were indexed using default options

Select a reference genome *

7: Homo_sapiens.GRCh38.dna.chr14.fa

(--genomeFastaFiles)

Length of the SA pre-indexing string *

14

Typically between 10 and 15. Longer strings will use much more memory, but allow faster searches. For small genomes, the parameter --genomeSAindexNbases must be scaled down to min(14, log2(GenomeLength)/2 - 1) (--genomeSAindexNbases)

Build index with or without known splice junctions annotation

build index with gene-model

To build an index with known splice junctions annotated, you will have to provide a GTF or GFF3 dataset that describes the gene models (the location of genes, transcripts and exons) known for the reference genome.

Gene model (gff3,gtf) file for splice junctions *

8: (unavailable) Homo_sapiens.GRCh38.110~chr14.gff3

32: RNA STAR on data 8, data 7, and others: mapped.bam

Add Tags

25.7 MB

format **bam**, database **?**

```
STAR --runMode genomeGenerate --
genomeDir tempstargenomedir --
genomeFastaFiles refgenome.fa --
```

Download

Binary bam alignments file

Visualize

https://igv.org

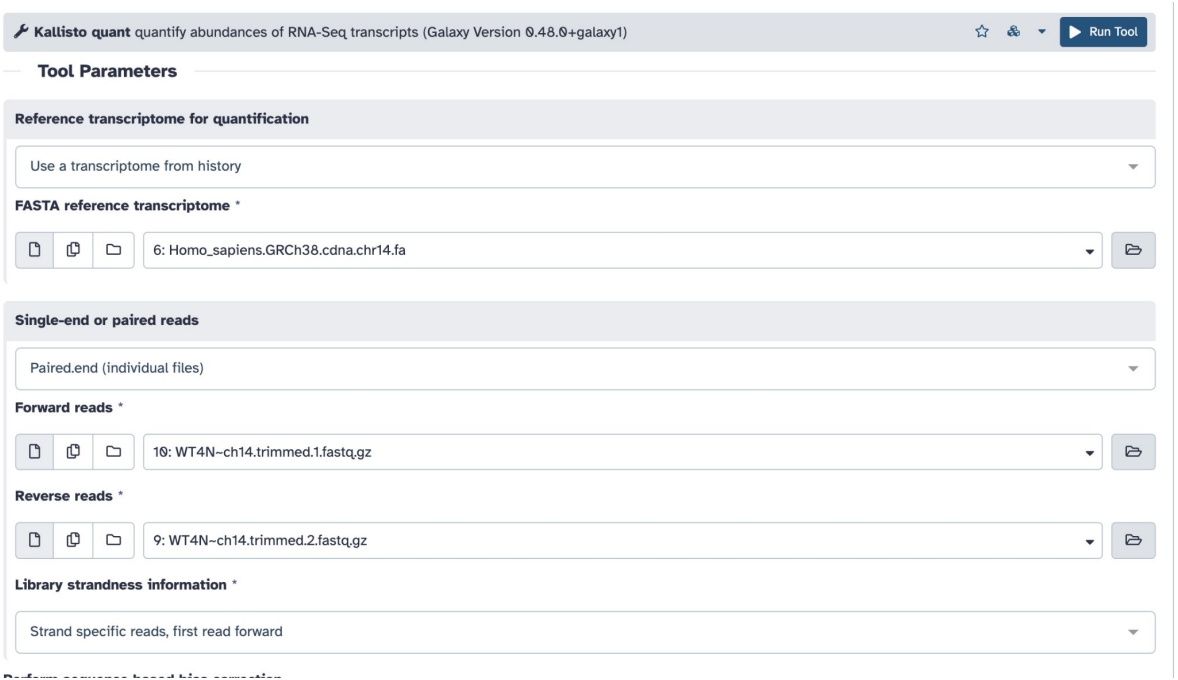
Visualize

Reads in exons = good

Kallisto mapping

Use cdna file (transcriptome)

Strandedness: https://rnabio.org/module-09-appendix/0009/12/01/StrandSettings/



Column 1	Column 2	Column 3	Column 4	Column 5
target_id	length	eff_length	est_counts	tpm
ENST00000250457.9	2706	2470.76	998.086	2202.38
ENST00000553215.5	1004	768.764	161.159	1142.93
ENST00000487915.6	848	612.764	0	0
ENST00000547327.2	2849	2613.76	21.755	45.3784

Length: actual length of transcript

Eff_length: “effective length” accounting for read length

Est_counts: “estimated counts” how many reads mapped to that transcript

Tpm: “transcripts per million” *http://mcb112.org/w01/w01-lecture.html*

Next Time

DESeq2 identifies genes that have statistically significant changes between conditions