

Outline

Sequencing (NGS) in general

Sequencing data analysis in general

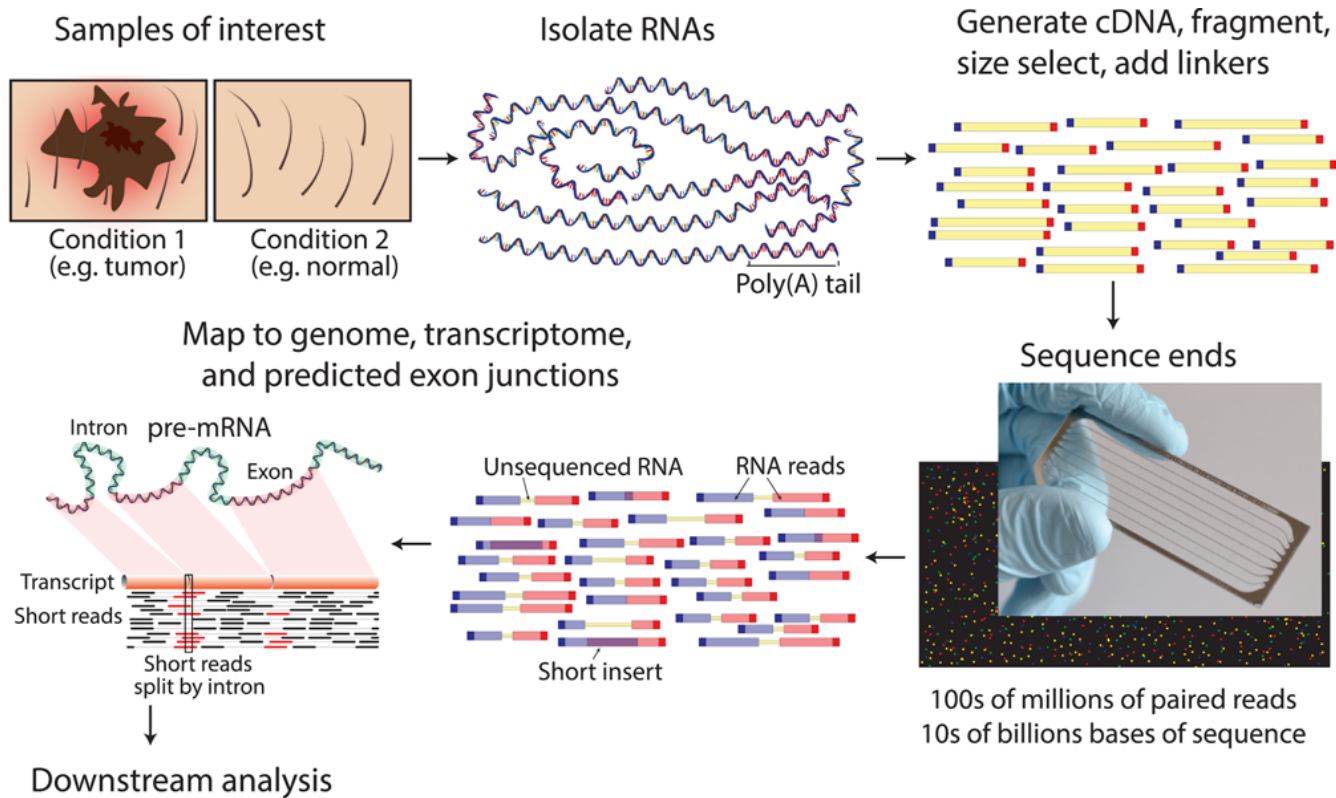
Kathi Zarnack and Julian König data

Results (selected)

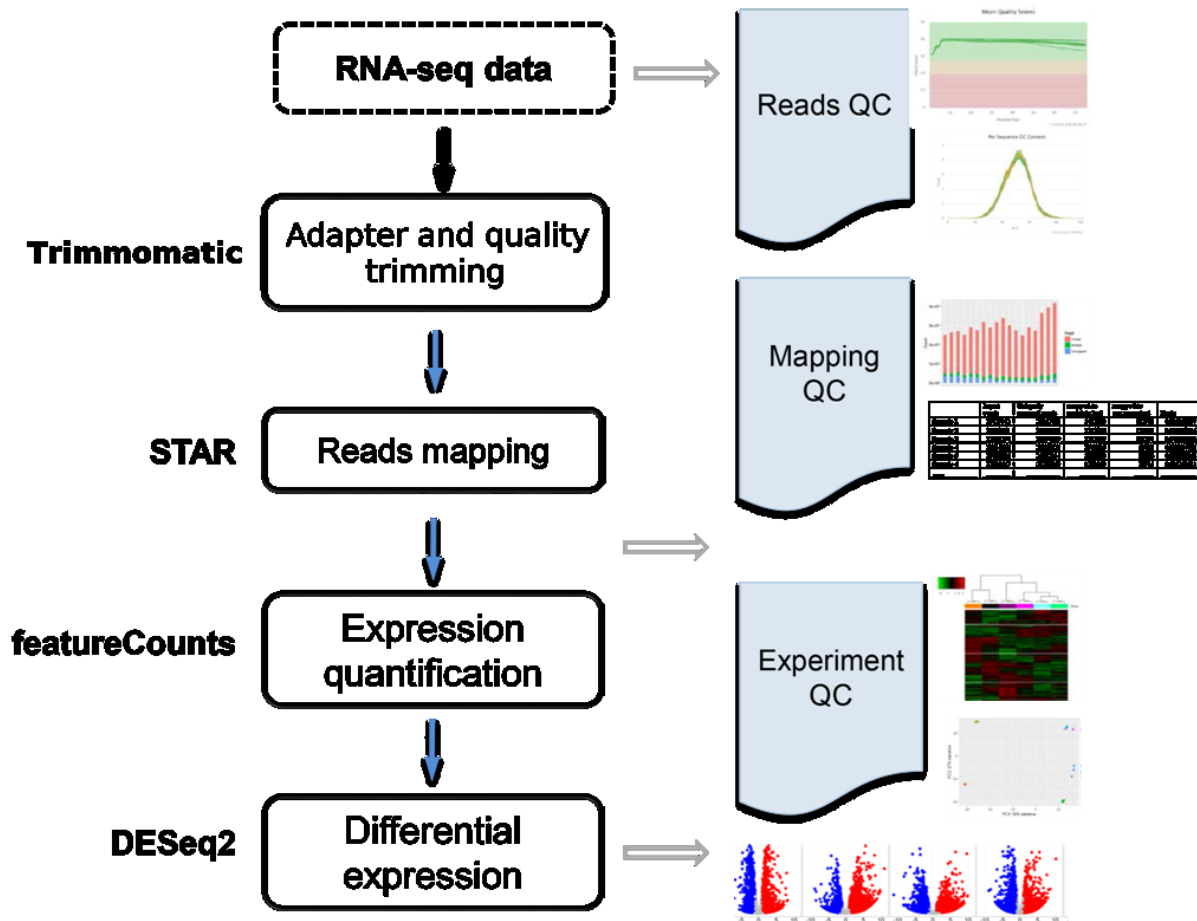
Galaxy

RNA-Seq data analysis

Sequencing (NGS) in general



Sequencing data analysis in general



Kathi Zarnack data

Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of *Alu* Elements

Kathi Zarnack,^{1,8} Julian König,^{2,8} Mojca Tajnik,^{2,3} Iñigo Martincorena,¹ Sebastian Eustermann,² Isabelle Stévant,¹ Alejandro Reyes,⁴ Simon Anders,⁴ Nicholas M. Luscombe,^{1,5,6,7,*} and Jernej Ule^{2,*}

¹European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

²MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

³Faculty of Medicine, University of Ljubljana, Vrazov trg 2, SI-1104 Ljubljana, Slovenia

⁴EMBL, Genome Biology Unit, Meyerhofstraße 1, 69117 Heidelberg, Germany

⁵UCL Genetics Institute, Department of Genetics, Environment and Evolution, University College London, Gower Street, London WC1E 6BT, UK

⁶Cancer Research UK London Research Institute, 44 Lincoln's Inn Fields, London WC2A 3LY, UK

⁷Okinawa Institute for Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa 904-0495, Japan

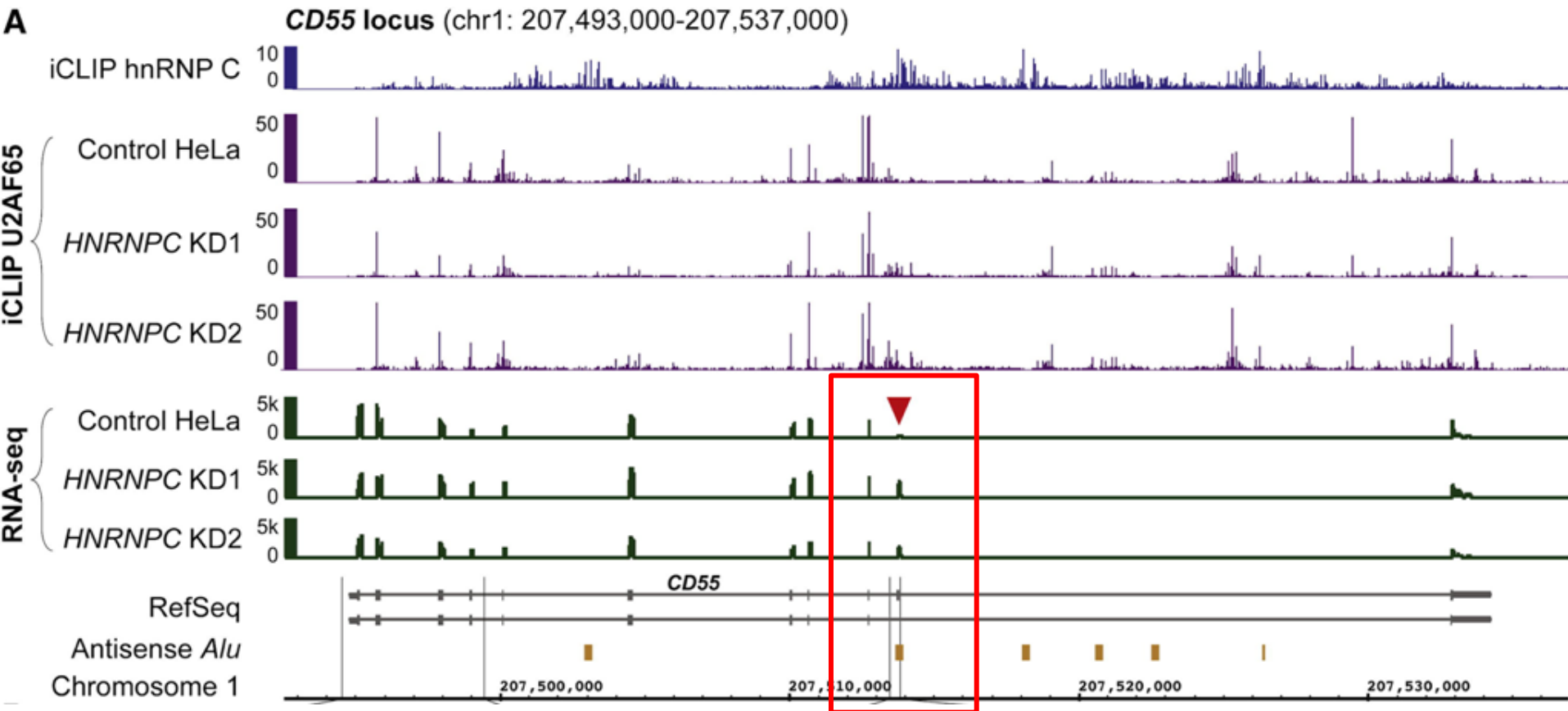
⁸These authors contributed equally to this work

*Correspondence: nicholas.luscombe@ucl.ac.uk (N.M.L.), jule@mrc-lmb.cam.ac.uk (J.U.)

<http://dx.doi.org/10.1016/j.cell.2012.12.023>

Summary of the results

There are ~650,000 ***Alu* elements** in transcribed regions of the human genome. These elements contain **cryptic splice sites**, so they are in constant **danger** of aberrant **incorporation into mature transcripts**. Despite posing a major threat to transcriptome integrity, **little is known** about the molecular **mechanisms preventing their inclusion**. Here, we present a mechanism for protecting the human transcriptome from the aberrant exonization of transposable elements. Quantitative **iCLIP data** show that the **RNA-binding protein hnRNP C competes** with the **splicing factor U2AF65** at many genuine and **cryptic splice sites**. **Loss of hnRNP C** leads to **formation** of previously **suppressed *Alu* exons**, which severely **disrupt transcript function**. Minigene experiments explain disease-associated mutations in *Alu* elements that hamper hnRNP C binding. Thus, by preventing U2AF65 binding to *Alu* elements, **hnRNP C** plays a **critical role** as a **genome-wide sentinel protecting the transcriptome**. The findings have important implications for human evolution and disease.



Goal of the practical

Get from the **raw sequencing data** to the **gene expression (RNA-Seq)**

Analyze **RNA-Seq** data and get **differential gene expression** and **expression** of individual **exons** (example at gene CD55 gene)

Show **coverage cryptic exon(s)** (example at gene CD55)

Do **everything** in **less than half a day**

Galaxy practical

Get the data

Or you just load the **preloaded data**

Shared Data -> Data Libraries -> Bi5444 -> RNA-Seq

Galaxy practical

1. Analyze Data Workflow **Shared Data** Visualization Help User

Data Libraries

Histories

Workflows

Visualizations

Pages



CEITEC

Welcome to the CEITEC MU private Galaxy s

2.

Galaxy

DATA LIBRARIES

< 0 1 2 >

6 librar

name↓

Bi5444

Bioda_group

CEITEC_Workshop

3.

DATA LIBRARIES include deleted

+ Create Folder

Libraries / Bi5444

name ↓



RNA-Seq

Galaxy practical

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with 'Galaxy' on the left and 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', and 'User' on the right. Below this is a toolbar with buttons for 'Create Folder', 'Add Datasets', 'To History', 'Download', 'Delete', 'Details', and 'Help'. The main content area shows a library named 'B15444 / RNA-Seq' with a table of datasets. A dialog box titled 'Import into History' is open in the foreground, allowing the user to select a history or create a new one. The 'To History' button in the toolbar and the 'name 1!' input field in the library list are highlighted with red boxes. The 'or create new:' input field in the dialog is highlighted with a blue box.

Galaxy

Analyze Data Workflow Visualize Shared Data Help User

DATA LIBRARIES include deleted Create Folder Add Datasets To History Download Delete Details Help

Libraries / B15444 / RNA-Seq

name 1!

Control_rep1_1.fastq
Control_rep1_2.fastq
Control_rep2_1.fastq
Control_rep2_2.fastq
Ensemble_Homo_sapiens.GRCh38.94.gtf
HUGO_Gene_information
KD1_rep1_1.fastq
KD1_rep1_2.fastq
KD1_rep2_1.fastq
KD1_rep2_2.fastq
KD2_rep1_1.fastq
KD2_rep1_2.fastq
KD2_rep2_1.fastq
KD2_rep2_2.fastq

description data_type size time updated (UTC) state

Import into History

Select history: Unnamed history

or create new: RNA-Seq data Library import

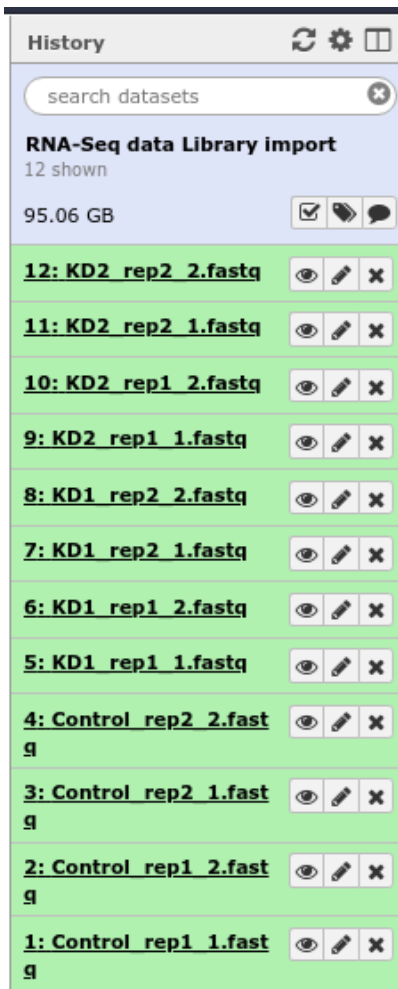
Import Close

fastqsanger 6.1 GB 2019-08-28 11:44 AM Manage
fastqsanger 6.1 GB 2019-08-28 11:44 AM Manage
fastqsanger 6.1 GB 2019-08-28 11:44 AM Manage

< 1 2 > 14 items shown (change) 14 total

Galaxy practical

Get the data



The screenshot shows the Galaxy History panel. At the top, there is a 'History' header with refresh, settings, and window icons. Below it is a search bar labeled 'search datasets'. The main content area is titled 'RNA-Seq data Library import' and shows '12 shown' items with a total size of '95.06 GB'. Each item is represented by a green row with a file name and three icons (eye, pencil, and X).

File Name	View	Edit	Delete
12: <u>KD2_rep2_2.fastq</u>			
11: <u>KD2_rep2_1.fastq</u>			
10: <u>KD2_rep1_2.fastq</u>			
9: <u>KD2_rep1_1.fastq</u>			
8: <u>KD1_rep2_2.fastq</u>			
7: <u>KD1_rep2_1.fastq</u>			
6: <u>KD1_rep1_2.fastq</u>			
5: <u>KD1_rep1_1.fastq</u>			
4: <u>Control_rep2_2.fastq</u>			
3: <u>Control_rep2_1.fastq</u>			
2: <u>Control_rep1_2.fastq</u>			
1: <u>Control_rep1_1.fastq</u>			

Galaxy practical

Initial quality check

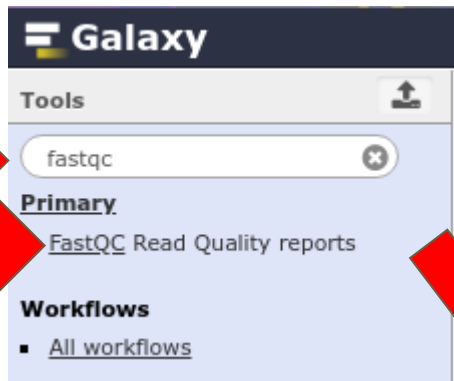
Check the **raw** reads **quality**

Using **F**ast**Q**C tool

Input **FASTQ**, output **HTML**

Galaxy practical

Initial quality check



Galaxy

Tools

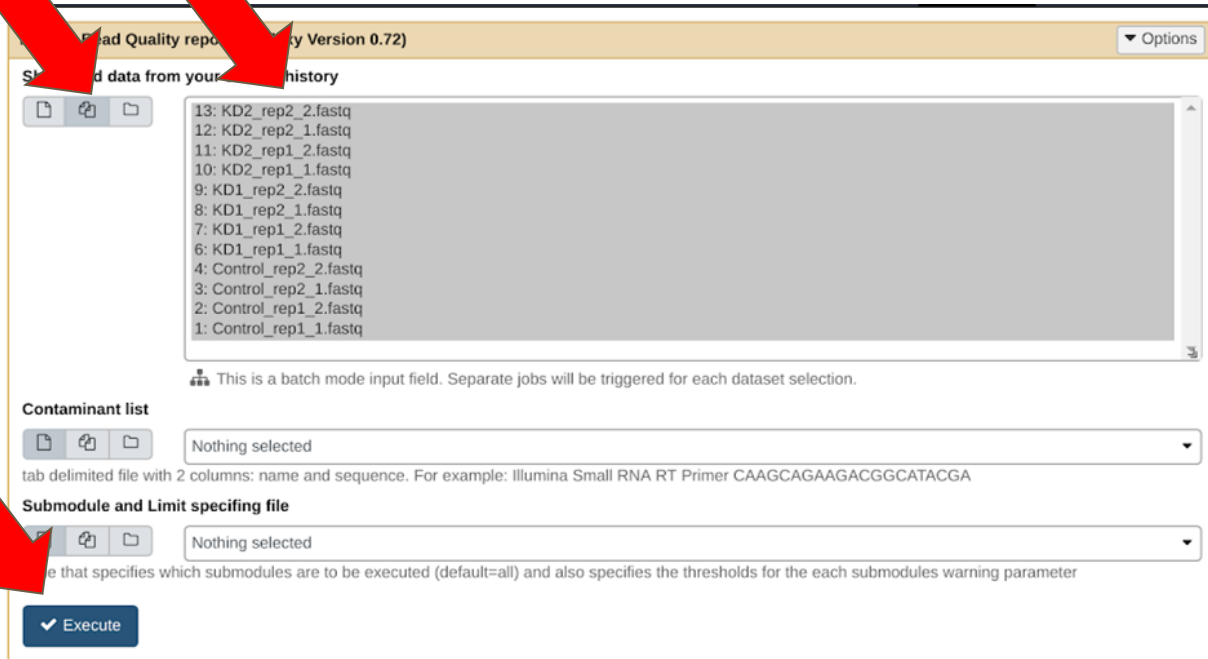
fastqc

Primary

FastQC Read Quality reports

Workflows

- All workflows



FastQC Read Quality reports (Galaxy Version 0.72) Options

Selected data from your history

```
13: KD2_rep2_2.fastq
12: KD2_rep2_1.fastq
11: KD2_rep1_2.fastq
10: KD2_rep1_1.fastq
9: KD1_rep2_2.fastq
8: KD1_rep2_1.fastq
7: KD1_rep1_2.fastq
6: KD1_rep1_1.fastq
4: Control_rep2_2.fastq
3: Control_rep2_1.fastq
2: Control_rep1_2.fastq
1: Control_rep1_1.fastq
```

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifying file

Nothing selected

File that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Execute

Galaxy practical

Initial quality check

It is still running, right?

But without that, we cannot proceed ☐

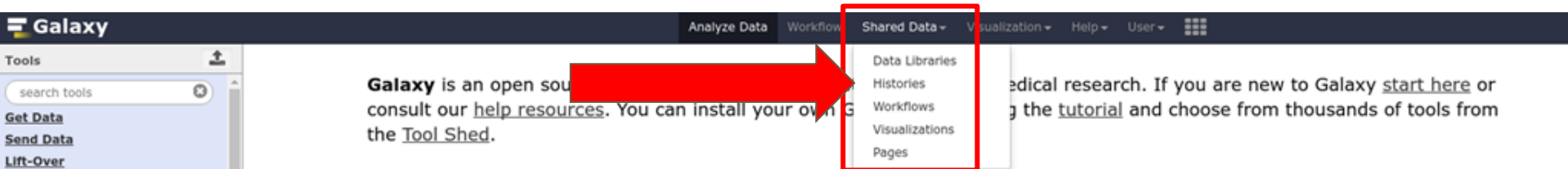
We have a solution! :)

Import **Galaxy history**

Galaxy practical

Initial quality check

Import Galaxy history



The screenshot displays the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Shared Data' dropdown menu is open, showing options: 'Data Libraries', 'Histories', 'Workflows', 'Visualizations', and 'Pages'. A red arrow points from the text 'Galaxy is an open source...' to the 'Histories' option in the dropdown menu. The left sidebar contains a 'Tools' section with a search bar and buttons for 'Get Data', 'Send Data', and 'Lift-Over'. The main content area contains text: 'Galaxy is an open source... consult our [help resources](#). You can install your own Galaxy on the [Tool Shed](#). ... medical research. If you are new to Galaxy [start here](#) or [try the tutorial](#) and choose from thousands of tools from...

Galaxy practical

Initial quality check

Import Galaxy history

Analyze Data Workflow

Published Histories

search name, annotation, own Q

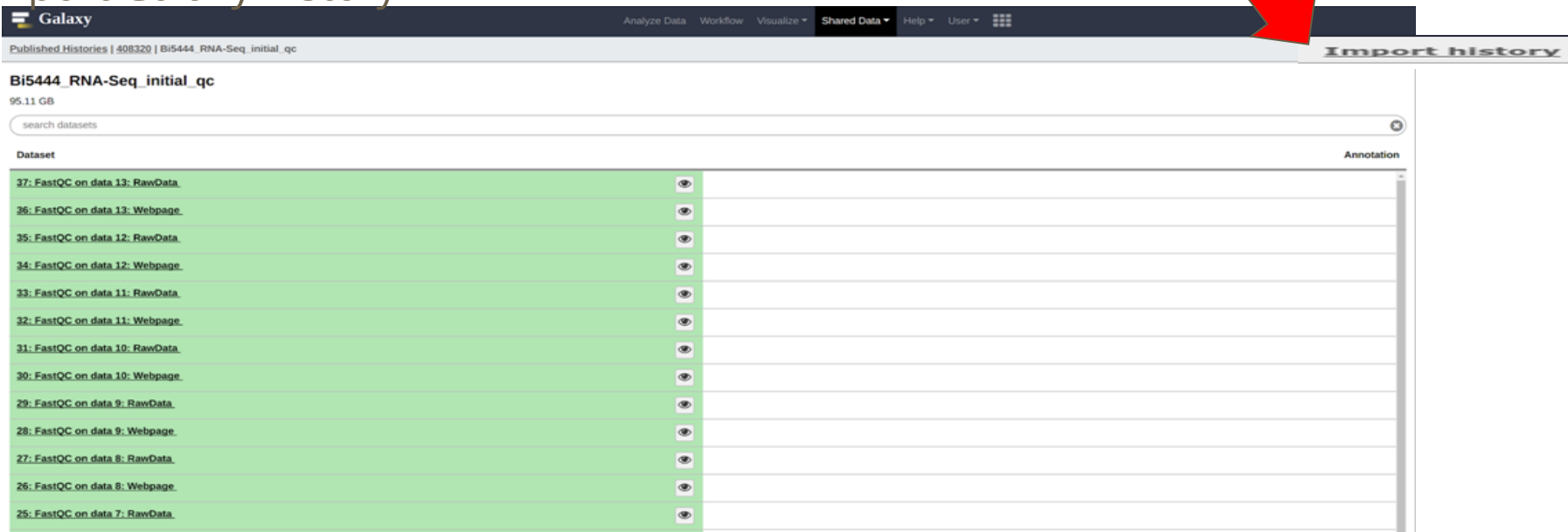
Advanced Search

Name	Annotation	Owner
BI5444_RNA-Seq_alignment		408320
BI5444_RNA-Seq_deprocess		408320
BI5444_RNA-Seq_initial_qc		408320
BI5444_RNA-Seq_qlc		408320
deseq2_calculation		323639
Lysak_group_tutorial - visualization		98640
Lysak_group_tutorial - Ks		98640
RNA-2018-03-Expression		323639

Galaxy practical

Initial quality check

Import Galaxy history



The screenshot displays the Galaxy web interface. At the top, there is a navigation bar with the Galaxy logo and menu items: Analyze Data, Workflow, Visualize, Shared Data, Help, and User. Below the navigation bar, the current history is identified as 'Published Histories | 408320 | Bi5444_RNA-Seq_initial_qc'. A red arrow points to the 'Import history' button in the top right corner. The main content area shows the details for the history 'Bi5444_RNA-Seq_initial_qc', which is 95.11 GB in size. There is a search bar for datasets. Below the search bar, a table lists the datasets in the history, with columns for 'Dataset' and 'Annotation'. The datasets are listed in descending order of ID, from 37 to 25.

Dataset	Annotation
37: FastQC on data 13: RawData	
36: FastQC on data 13: Webpage	
35: FastQC on data 12: RawData	
34: FastQC on data 12: Webpage	
33: FastQC on data 11: RawData	
32: FastQC on data 11: Webpage	
31: FastQC on data 10: RawData	
30: FastQC on data 10: Webpage	
29: FastQC on data 9: RawData	
28: FastQC on data 9: Webpage	
27: FastQC on data 8: RawData	
26: FastQC on data 8: Webpage	
25: FastQC on data 7: RawData	

Galaxy practical

Initial quality check

HTML report(s)



History

search datasets

Bi5444_RNA-Seq_initial_qc

37 shown

95.11 GB

- 37: FastQC on data 13: R
awData
- 36: FastQC on data 13:
Webpage
- 35: FastQC on data 12: R
awData
- 34: FastQC on data 12:
Webpage
- 33: FastQC on data 11: R
awData
- 32: FastQC on data 11:
Webpage
- 31: FastQC on data 10: R
awData
- 30: FastQC on data 10:
Webpage
- 29: FastQC on data 9: Ra
wData
- 28: FastQC on data 9: W
ebpage
- 27: FastQC on data 8: Ra
wData
- 26: FastQC on data 8: W
ebpage
- 25: FastQC on data 7: Ra
wData
- 24: FastQC on data 7: W
ebpage
- 23: FastQC on data 6: Ra
wData
- 22: FastQC on data 6: W
ebpage



Bi5444_RNA-Seq_initial_qc

37 shown

95.11 GB

- 37: FastQC on data 13: R
awData
- 36: FastQC on data 13:
Webpage
- 35: FastQC on data 12: R
awData
- 34: FastQC on data 12:
Webpage
- 33: FastQC on data 11: R
awData

Two large red arrows point from the top of the zoomed-in view back to the full history list on the left.

Galaxy practical

Initial quality check

HTML report(s)

But there is **too many of them**

MultiQC - makes you life simpler

This time, you can **try it on your own!**

Galaxy practical

Summary of the logs

Summarize the output **logs**

Using **MultiQC** tool

Input **LOG(s)**, output **HTML**

Galaxy practical

Summary of the logs

Galaxy

Tools

multiqc

Primary

MultiQC aggregate results from bioinformatics analyses into a single report

Workflows

- All workflows

MultiQC aggregate results from bioinformatics analyses into a single report (Galaxy Version 1.5.0) Versions Options

1: MultiQC

What was used generate logs?

FastQC

Software name

FastQC output

1: FastQC output

Type of FastQC output?

Raw data

FastQC output

- 37: FastQC on data 13: RawData
- 36: FastQC on data 13: Webpage
- 35: FastQC on data 12: RawData
- 34: FastQC on data 12: Webpage
- 33: FastQC on data 11: RawData
- 32: FastQC on data 11: Webpage
- 31: FastQC on data 10: RawData
- 30: FastQC on data 10: Webpage
- 29: FastQC on data 9: RawData

+ Insert FastQC output

+ Insert Results

Report title

It is printed as page header

Custom comment

It will be printed at the top of the report

Output the multiQC log file?

Yes No

mostly useful for debugging purposes

Execute

Bi5444_RNA-Seq...process

89 shown, 3 deleted, 3 hidden

3.7 GB

39: MultiQC on data 37, data 35, and others: Webpage

38: MultiQC on data 37, data 35, and others: Stats

a list with 3 items

37: FastQC on data 13: R

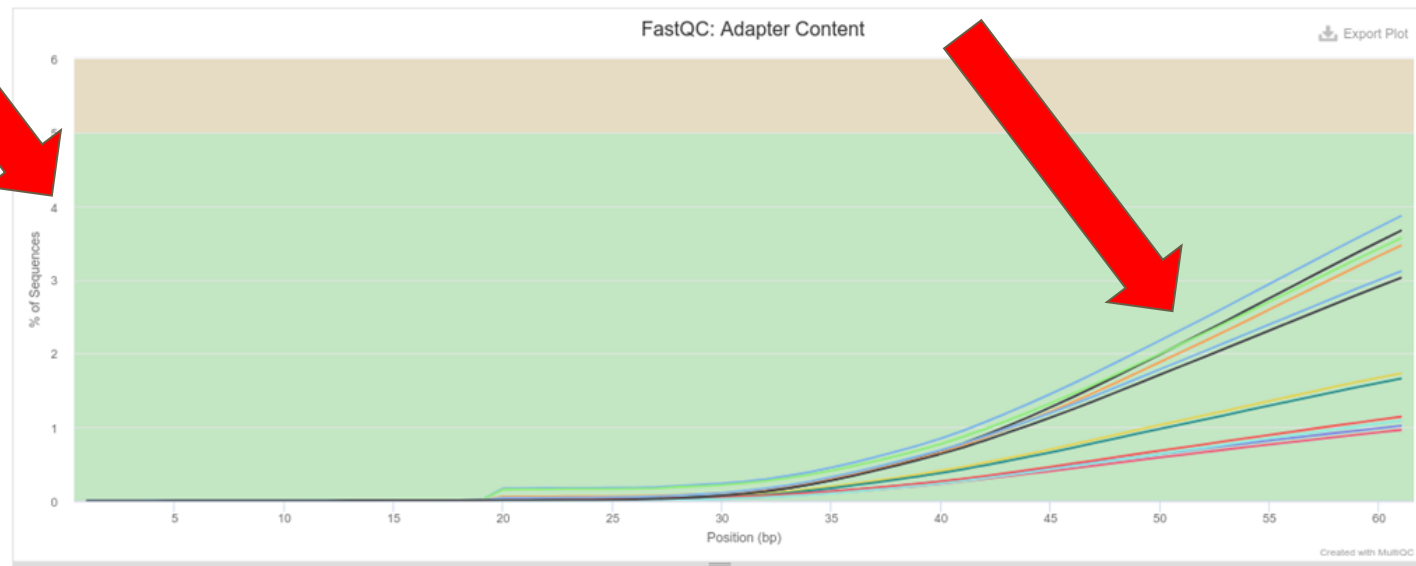
Galaxy practical

Initial quality check - Adapter content

MultiQC v1.5
Initial quality check
General Stats
FastQC
Sequence Quality Histograms
Per Sequence Quality Scores
Per Base Sequence Content
Per Sequence GC Content
Per Base N Content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content

Adapter Content 12

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. See the [FastQC help](#). Only samples with $\geq 0.1\%$ adapter contamination are shown. Y-Limits: on



Galaxy practical

Read preprocessing

Remove **adapters** and/or trim **low-quality** ends


Using **Trimmomatic** trimmer


Input **FASTQ**, output **FASTQ**

Galaxy practical

Read preprocessing

Galaxy

Tools 

trimmomatic 

Primary

- [Trimmomatic](#) flexible read trimming tool for Illumina NGS data

Genome assembly

- [Shovill](#) Faster SPAdes assembly of Illumina reads

Workflows

- [All workflows](#)


Automatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.5)

Trim paired-end or paired-end reads?

Paired-end (two separate input files)


Input FASTQ file (R1/first of pair)

- 19: Control_rep2_2.fastq
- 18: Control_rep2_1.fastq
- 17: Control_rep1_2.fastq
- 17: Control_rep1_1.fastq
- 16: Control_rep2_2.fastq

 This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Input FASTQ file (R2/second of pair)

- 13: KD2_rep1_1.fastq
- 12: KD1_rep2_2.fastq
- 11: KD1_rep2_1.fastq
- 10: KD1_rep1_2.fastq
- 9: KD1_rep1_1.fastq

 This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Perform initial ILLUMINACLIP step?

Cut adapter and other illumina-specific sequences from the read

Galaxy practical

Read preprocessing

Perform initial ILLUMINACLIP step?

Yes No

Remove adapter and other illumina-specific sequences from the read

Select standard adapter sequences or provide custom?

Custom

Custom adapter sequences in fasta format

```
>adapter  
AGATCGGAAGAGC
```

Write sequences in the fasta format.

Adapter sequence (partial):

```
>adapter
```

```
AGATCGGAAGAGC
```

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across

4

Average quality required

5

+ Insert Trimmomatic Operation

✓ Execute

Galaxy practical

Read preprocessing

The screenshot displays a Galaxy workflow titled "Bi5444_RNA-Seq_preprocess" with a total size of 183.7 GB. The workflow consists of 89 jobs, with 3 deleted and 3 hidden. The visible jobs are Trimmomatic operations on various fastq files, including unpaired and paired reads from two replicates (1 and 2) for both KD and Control samples. Each job includes icons for visibility, editing, and deletion. Two large red arrows point from the top right towards the bottom left, highlighting the sequence of jobs.

Bi5444_RNA-Seq_preprocess
89 shown, 3 deleted, 3 hidden
183.7 GB

- 66: Trimmomatic on KD 2_rep2_2.fastq (R2 unpaired).
- 65: Trimmomatic on KD 2_rep2_1.fastq (R1 unpaired).
- 64: Trimmomatic on KD 2_rep2_2.fastq (R2 paired).
- 63: Trimmomatic on KD 2_rep2_1.fastq (R1 paired).
- 62: Trimmomatic on KD 2_rep1_2.fastq (R2 unpaired).
- 61: Trimmomatic on KD 2_rep1_1.fastq (R1 unpaired).
- 60: Trimmomatic on KD 2_rep1_2.fastq (R2 paired).
- 59: Trimmomatic on KD 2_rep1_1.fastq (R1 paired).
- 58: Trimmomatic on KD 1_rep2_2.fastq (R2 unpaired).
- 57: Trimmomatic on KD 1_rep2_1.fastq (R1 unpaired).
- 56: Trimmomatic on KD 1_rep2_2.fastq (R2 paired).
- 55: Trimmomatic on KD 1_rep2_1.fastq (R1 paired).
- 54: Trimmomatic on KD 1_rep1_2.fastq (R2 unpaired).
- 53: Trimmomatic on KD 1_rep1_1.fastq (R1 unpaired).
- 52: Trimmomatic on KD 1_rep1_2.fastq (R2 paired).
- 51: Trimmomatic on KD 1_rep1_1.fastq (R1 paired).
- 50: Trimmomatic on Control_rep2_2.fastq (R2 unpaired).
- 49: Trimmomatic on Control_rep2_1.fastq (R1 unpaired).
- 48: Trimmomatic on Control_rep2_2.fastq (R2 paired).
- 47: Trimmomatic on Control_rep2_1.fastq (R1 paired).
- 46: Trimmomatic on Control_rep2_2.fastq (R2 unpaired).
- 45: Trimmomatic on Control_rep2_1.fastq (R1 unpaired).
- 44: Trimmomatic on Control_rep1_2.fastq (R2 paired).
- 43: Trimmomatic on Control_rep1_1.fastq (R1 paired).

Galaxy practical

Preprocessing quality check

Check the **preprocessed** reads **quality and summarize**

Using **FastQC** & **MultiQC** tools

Input **FASTQ/LOG**, output **HTML**

Galaxy practical

Preprocessing quality check

Please, run the `FastQC` and `MultiQC` on the preprocessed files and check the adapter content

Galaxy practical

Preprocessing quality check

Share Data -> Histories -> Bi5444 RNA-Seq preprocess

Published Histories

Q

[Advanced Search](#)

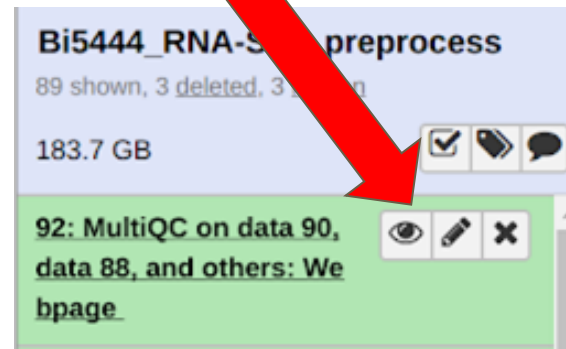
Name	Annotation
Bi5444_RNA-Seq_alignment	
Bi5444_RNA-Seq_preprocess	
Bi5444_RNA-Seq_initial_qc	
Bi5444 RNA-Seq - Clean	
deseq2 calculation	
Lysak group tutorial - visualization	
Lysak group tutorial - Ks	
RNA-2018-03-Expression	
RNA-2018-01a-Initial quality check	

Galaxy practical

Preprocessing quality check

Check the **preprocessed** reads **quality & summarize**

Are all the **bad** things **gone**?



Bi5444_RNA-Seq preprocess
89 shown, 3 deleted, 3 ...
183.7 GB

92: MultiQC on data 90, data 88, and others: We bpage_

Galaxy practical

Preprocessing quality check

Adapter Content

12

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. See the [FastQC help](#). Only samples with $\geq 0.1\%$ adapter contamination are shown.

No samples found with any adapter contamination > 0.1%

Galaxy practical

Preprocessing quality check

Check the **preprocessed** reads **quality & summarize**

Are all the **bad** things **gone**?

Actually, for **modern aligners** such as **STAR**, it **doesn't** matter that **much**

They can perform **soft-clipping**

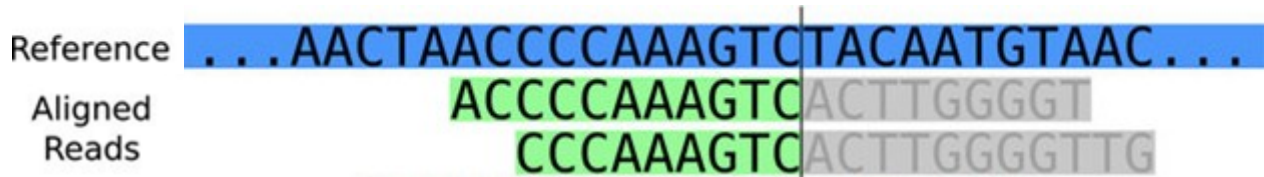
Galaxy practical

Soft-clipping in alignment

Hiding of **non-matching** parts of the **reads**

Can overcome **remaining adapter** or **low-quality** sequences

Only to **specified limits** (in **STAR** the default is max. **33%** of the read length)



Soft-clipped part

Galaxy practical

Alignment to genome

Align **RNA-Seq** data to a genome

Using **STAR** aligner

Input **FASTQ**, output **BAM**

Galaxy practical

Alignment to genome

Galaxy

Tools

star

Text Manipulation

- Text reformatting with awk

Convert Formats

- MAF to BED Converts a MAF formatted file to the BED format
- MAF to FASTA Converts a MAF formatted file to FASTA format

NGS: RNA Analysis

- RNA STAR** Gapped-read mapper for RNA-seq data new

NGS: SAMtools

RNA STAR Gapped-read mapper for RNA-seq data (Galaxy Version 2.6.0b-1) Options

Single-end or paired-end reads

Paired-end (as individual datasets)

RNA-Seq FASTQ/FASTA file, forward read

- 66: Trimmomatic on KD2_rep1_2.fastq (R2 unpaired)
- 65: Trimmomatic on KD2_rep1_1.fastq (R1 unpaired)
- 64: Trimmomatic on KD2_rep2_2.fastq (R2 paired)
- 63: Trimmomatic on KD2_rep2_1.fastq (R1 paired)
- 62: Trimmomatic on KD2_rep1_2.fastq (R2 unpaired)
- 61: Trimmomatic on KD2_rep1_1.fastq (R1 unpaired)
- 60: Trimmomatic on KD2_rep2_2.fastq (R2 paired)
- 59: Trimmomatic on KD2_rep2_1.fastq (R1 paired)
- 58: Trimmomatic on KD1_rep2_2.fastq (R2 unpaired)
- 57: Trimmomatic on KD1_rep2_1.fastq (R1 unpaired)

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

RNA-Seq FASTQ/FASTA file, reverse read

- 66: Trimmomatic on KD2_rep2_2.fastq (R2 paired)
- 65: Trimmomatic on KD2_rep1_1.fastq (R1 paired)
- 64: Trimmomatic on KD2_rep2_2.fastq (R2 paired)
- 63: Trimmomatic on KD2_rep2_1.fastq (R1 paired)
- 62: Trimmomatic on KD2_rep1_2.fastq (R2 unpaired)
- 61: Trimmomatic on KD2_rep1_1.fastq (R1 unpaired)
- 60: Trimmomatic on KD2_rep1_2.fastq (R2 paired)
- 59: Trimmomatic on KD2_rep1_1.fastq (R1 paired)
- 58: Trimmomatic on KD1_rep2_2.fastq (R2 unpaired)
- 57: Trimmomatic on KD1_rep2_1.fastq (R1 unpaired)

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Custom or built-in reference genome

Use a built-in index

Built-ins were indexed using default options

Reference genome with or without an annotation

use genome reference without builtin gene-model

Must the index have been created WITH a GTF file (if not you can specify one afterward).

Select reference genome

Human (Homo sapiens) (Ens94): GRCh38

If your genome of interest is not listed, contact the Galaxy team (--genomeDir)

Execute

Sit back, wait and relax



You don't need to know what's happening or how long it's going to take.



Galaxy practical

Alignment to genome

Share Data -> Histories -> Bi5444 RNA-Seq alignment

Galaxy practical

Alignment to genome

STAR performs well even **with defaults**

Main output is the **BAM** file

This is one of the few files worth to **keep** and **save**

Bi5444_CLIP-Seq_alignment
6 shown
27.7 GB

6: KD2_rep2.bam
KD2_rep2

5: KD2_rep1.bam
KD2_rep1

4: KD1_rep2.bam
KD1_rep2

3: KD1_rep1.bam
KD1_rep1

2: Control_rep2.bam
Control_rep2

1: Control_rep1.bam
Control_rep1

7.6 GB
format: **bam**, database: **GRCh38**

Sep 16 12:00:14 started STAR run
Sep 16 12:00:14 loading genome
Sep 16 12:04:51 started mapping
Sep 16 12:16:06 started sorting BAM
Sep 16 12:24:18 finished successfully

Binary bam alignments file

Galaxy practical

Quality control of alignment

Run **MultiQC** to assess the alignment

MultiQC aggregate results from bioinformatics analyses into a single report (Galaxy Version 1.5.0) Versions Options

Results

Which tool was used generate logs?

STAR

Software name

STAR output

1: STAR output

Type of STAR output?

Log

STAR log output

```
11: RNA STAR on data 64 and data 63: log
9: RNA STAR on data 60 and data 59: log
7: RNA STAR on data 56 and data 55: log
5: RNA STAR on data 52 and data 51: log
3: RNA STAR on data 48 and data 47: log
1: RNA STAR on data 44 and data 43: log
```

+ Insert STAR output

+ Insert Results

Report title

It is printed as page header

Custom comment

It will be printed at the top of the report

Output the multiQC log file?

Yes No

This is mostly useful for debugging purposes

Execute

Galaxy practical

Quality control of alignment

Seq. ...
14 shown, ...
27.7 GB

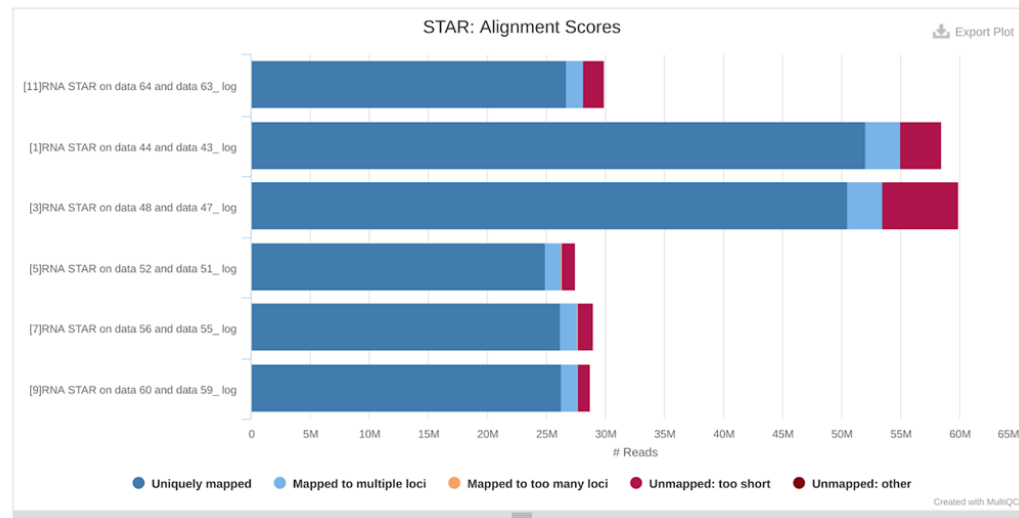
- 14: MultiQC on data 44 and data 9, and others: Webpage
- 13: MultiQC on data 11, data 9, and others: Stats
- 12: KD2_rep2.bam
- 11: RNA STAR on data 64 and data 63: log
- 10: KD2_rep1.bam
- 9: RNA STAR on data 60 and data 59: log

STAR

STAR is an ultrafast universal RNA-seq aligner.

Alignment Scores

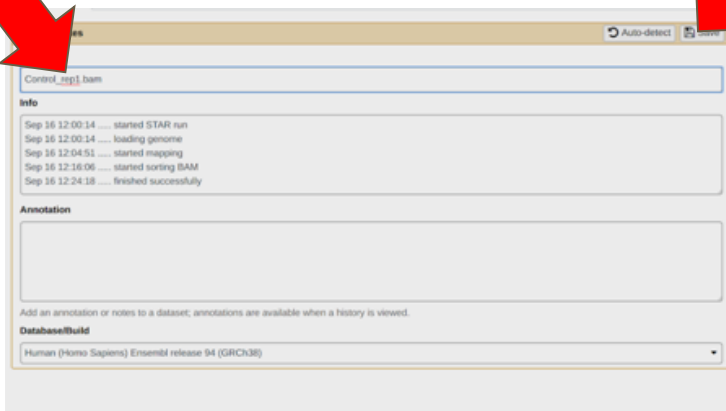
Number of Reads Percentages



Galaxy practical

Rename and tags

Better names comprehensibility



Control_rep1.bam

Info

Sep 16 12:00:14 started STAR run
Sep 16 12:00:14 loading genome
Sep 16 12:04:51 started mapping
Sep 16 12:16:06 started sorting BAM
Sep 16 12:24:18 finished successfully

Annotation

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build

Human (Homo Sapiens) Ensembl release 94 (GRCh38)



RNA-Seq alignment

211.4

100: RNA STAR on data 44 and data 43: mapped.bam

99: RNA STAR on data 44 and data 43: log

98: RNA STAR on data 44 and data 43: mapped.bam

97: RNA STAR on data 44 and data 43: splice junctions.bed

96: RNA STAR on data 44 and data 43: log

92: MultiQC on data 90, data 88, and others: Wc bpage

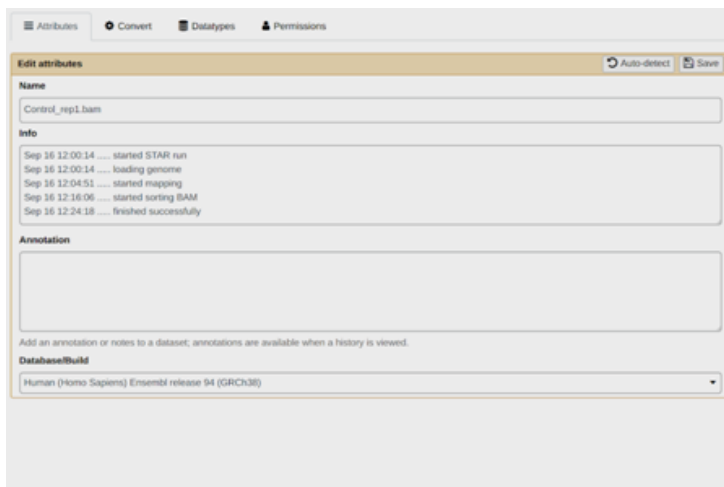
91: MultiQC on data 90, data 88, and others: Stats

3 of 3 items

Galaxy practical

Rename and tags

Better names comprehensibility



The screenshot shows the 'Edit attributes' panel in Galaxy. The 'Name' field contains 'Control_rep1.bam'. The 'Info' section lists the workflow steps: 'started STAR run', 'loading genome', 'started mapping', 'started sorting BAM', and 'finished successfully'. The 'Database/Build' dropdown is set to 'Human (Homo Sapiens) Ensembl release 94 (GRCh38)'.



The screenshot shows a dataset list for 'Bi5444 RNA-Seq_alignment'. Three red arrows point to specific items in the list: the top item '100: RNA STAR on data 48 and data 47, splice junctions.bed', the middle item '97: RNA STAR on data 4 and data 47, log', and the bottom item '97: RNA STAR on data 4 and data 42, splice junctions.bed'. The 'Control_rep1.bam' item is also visible in the list.

Galaxy practical

Gene counts

For the **raw gene counts** (expression) you need to have a list of genes and their positions in the genome - gene annotation

Using **UCSC Main table browser**

Input nothing, output **GTF**

Galaxy practical

Gene counts

Share Data -> Data Libraries -> Bi5444

-> RNA-Seq -> Ensemble_Homo_sapiens.GRCh38.94.gtf.gz

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with the Galaxy logo and the word 'Analyz'. Below this, there is a toolbar with buttons for 'Create Folder', 'Add Datasets', 'To History', 'Download', 'Delete', 'Details', and 'Help'. The main content area shows a data library named 'Bi5444' containing RNA-Seq data. The library is organized into a table with columns for 'name' and 'size'. The table contains several files, including 'Control_rep1_1.fastq', 'Control_rep1_2.fastq', 'Control_rep2_1.fastq', 'Control_rep2_2.fastq', 'Ensemble_Homo_sapiens.GRCh38.94.gtf.gz', 'HUGO.Gene information', and 'KD1_rep1_1.fastq'. The file 'Ensemble_Homo_sapiens.GRCh38.94.gtf.gz' is highlighted in orange. A red arrow points to the 'To History' button in the toolbar, and another red arrow points to the highlighted file in the table.

Galaxy practical

Gene counts

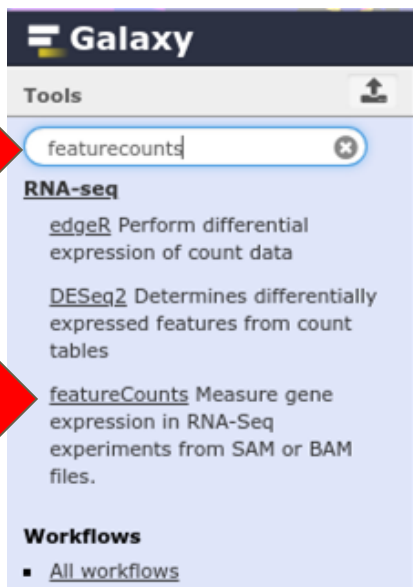
Get the **raw gene counts**

Using **featureCounts** tool

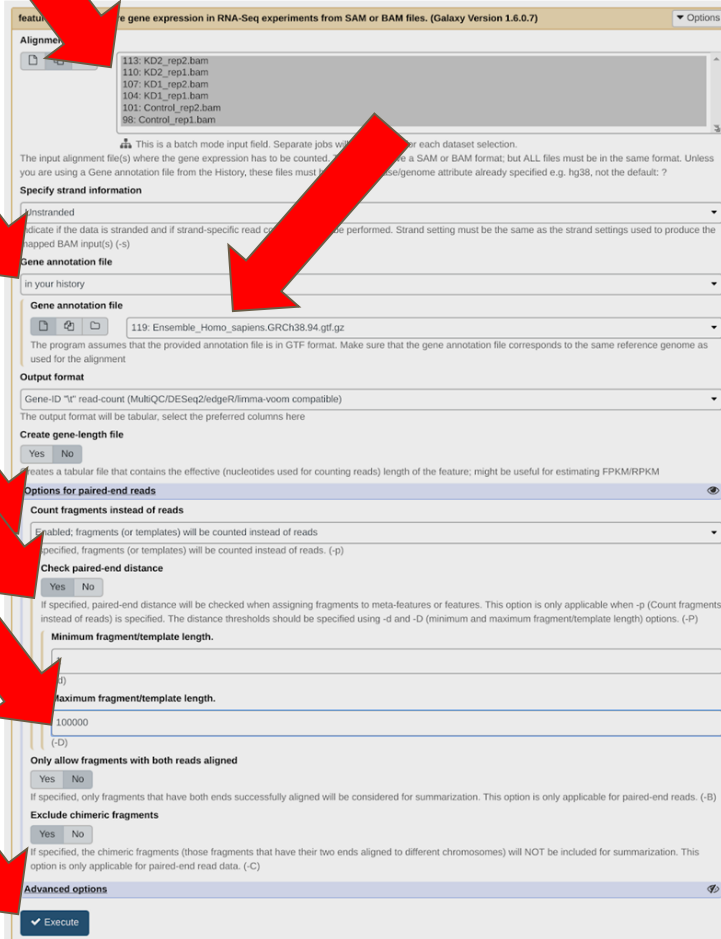
Input **BAM** and annotation **GTF**, output **TXT** (raw gene counts)

Galaxy practical

Gene counts



The screenshot shows the Galaxy interface with a search bar containing 'featurecounts'. Below the search bar, the 'RNA-seq' category is expanded, showing several tools. The 'featureCounts' tool is highlighted with a red arrow. The tool description reads: 'featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files.' Other tools listed include 'edgeR Perform differential expression of count data' and 'DESeq2 Determines differentially expressed features from count tables'. The 'Workflows' section is also visible at the bottom.



The screenshot shows the configuration page for the 'featureCounts' tool. The title is 'featureCounts: Measure gene expression in RNA-Seq experiments from SAM or BAM files. (Galaxy Version 1.6.0.7)'. The 'Alignments' section contains a list of input files: 113: KD2_rep2.bam, 110: KD2_rep1.bam, 107: KD1_rep2.bam, 104: KD1_rep1.bam, 101: Control_rep2.bam, and 98: Control_rep1.bam. The 'Gene annotation file' is set to '119: Ensemble_Homo_sapiens.GRCh38.94.gtf.gz'. The 'Output format' is 'Gene-ID "1" read-count (MultiQC/DESeq2/edgeR/lmma-voom compatible)'. The 'Create gene-length file' option is set to 'No'. The 'Options for paired-end reads' section includes 'Count fragments instead of reads' (set to 'Enabled'), 'Check paired-end distance' (set to 'No'), 'Minimum fragment/template length' (set to '100000'), and 'Maximum fragment/template length' (set to '100000'). The 'Only allow fragments with both reads aligned' option is set to 'No', and the 'Exclude chimeric fragments' option is set to 'No'. The 'Execute' button is at the bottom.

Galaxy practical

Gene counts

Quality control of gene counts

Again MultiQC

MultiQC aggregate results from bioinformatics analyses into a single report (Galaxy Version 1.5.0) Versions Options

Results

Results

Which tool was used generate logs?

featureCounts

Software name

Output of FeatureCounts

✕ 131: featureCounts on data 119 and data 113: summary ✕ 129: featureCounts on data 119 and data 110: summary

✕ 127: featureCounts on data 119 and data 107: summary ✕ 125: featureCounts on data 119 and data 104: summary

✕ 123: featureCounts on data 119 and data 101: summary ✕ 121: featureCounts on data 119 and data 98: summary

+ Insert Results

Report title

It is printed as page header

Custom comment

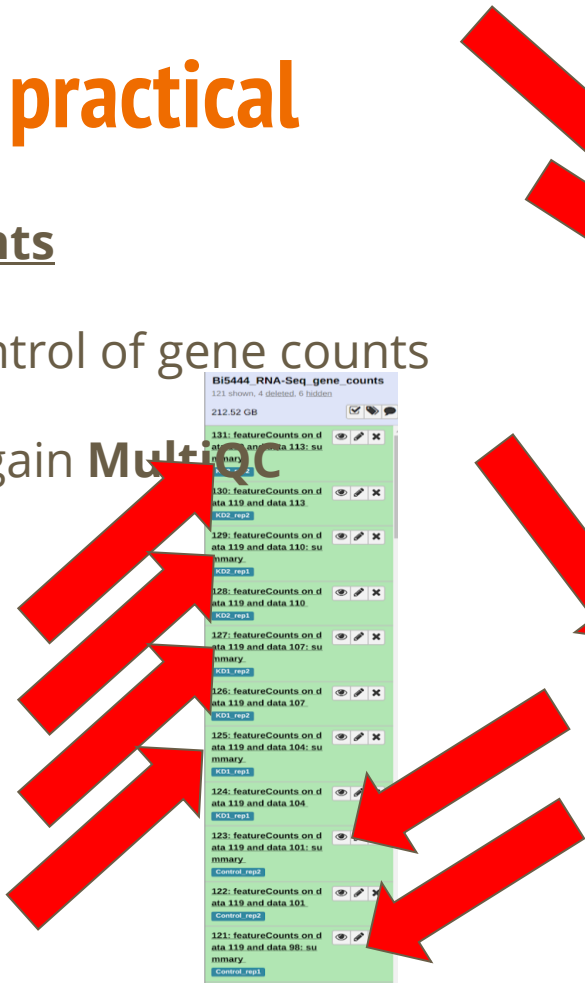
It will be printed at the top of the report

Output the multiQC log file?

Yes No

This is mostly useful for debugging purposes

Execute



Galaxy practical

MultiQC v1.5

General Stats
featureCounts

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2019-09-18, 09:54 based on data in: /home/galaxyuser/galaxy/jobs/818/18323/working/multiqc_501r

Welcome! Not sure where to start? [Watch a tutorial video](#) (6:06) [Start over](#)

General Statistics

Copy table | Configure Columns | Plot | Showing % rows and % columns.

Sample Name	% Assigned	M Assigned
Control_rep1	76.1%	47.3
Control_rep2	75.9%	45.9
KD1_rep1	75.6%	22.5
KD1_rep2	75.8%	23.6
KD2_rep1	76.6%	23.6
KD2_rep2	76.7%	24.2

featureCounts

Subread featureCounts is a highly efficient general-purpose read summarization program that counts mapped reads for genomic features such as genes, exons, promoter, gene bodies, genomic bins and chromosomal locations.

Number of Reads | Percentages

featureCounts: Assignments

Control_rep1
Control_rep2
KD1_rep1
KD1_rep2
KD2_rep1
KD2_rep2

Assigned | Unassigned FragmentLength | Unassigned MultiMapping | Unassigned NoFeatures | Unassigned Ambiguity

BI5444 | counts
212.53 GB

113: MultiQC on data 1 31, data 129, and other 8: Webpage
132: MultiQC on data 111, data 128, and others: Stats
131: featureCounts on data 119 and data 113: summary
130: featureCounts on data 119 and data 113: summary
129: featureCounts on data 119 and data 110: summary
128: featureCounts on data 119 and data 110: summary
127: featureCounts on data 119 and data 107: summary
126: featureCounts on data 119 and data 107: summary
125: featureCounts on data 119 and data 104: summary
124: featureCounts on data 119 and data 104: summary
123: featureCounts on data 119 and data 101: summary
122: featureCounts on data 119 and data 101: summary
121: featureCounts on data 119 and data 98: summary

Galaxy practical

Differential gene expression

Get **differential gene expression** from the raw counts

Using `edgeR` and `DESeq2` tools

Galaxy practical

Differential gene expression - note

Optimally, the experiment **should** be **designed** with at least **three biological replicates**

However, if the **data** are only “**supportive**” **two replicates** is enough

Bioinformatics. 2013

Liu Y, Zhou J, White KP. *RNA-seq differential expression studies: more sequence or more replication?*

Rna. 2016

Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M. *How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?.*

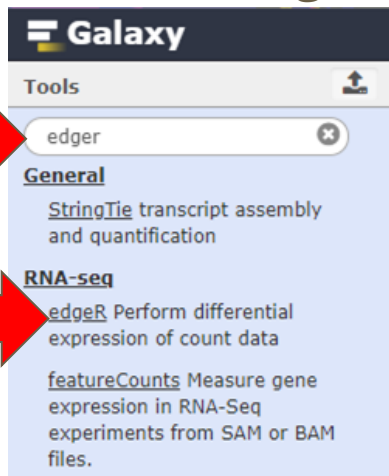
Galaxy practical

Differential gene expression

Shared Data -> Histories -> Bi5444 RNA-Seq DE start

Galaxy practical

Differential gene expression



Galaxy

Tools

edgeR

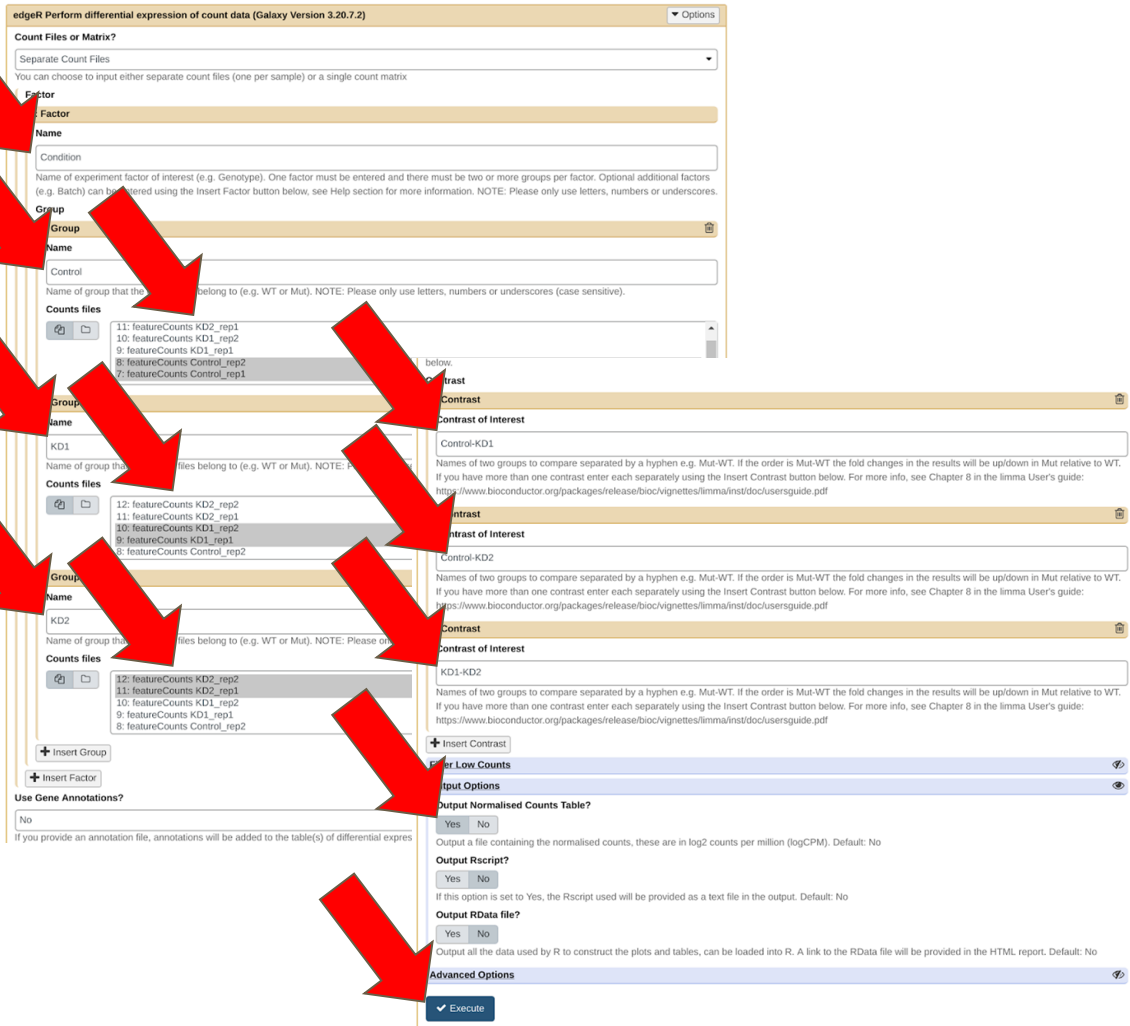
General

[StringTie](#) transcript assembly and quantification

RNA-seq

[edgeR](#) Perform differential expression of count data

[featureCounts](#) Measure gene expression in RNA-Seq experiments from SAM or BAM files.



edgeR Perform differential expression of count data (Galaxy Version 3.20.7.2)

Count Files or Matrix?
Separate Count Files

You can choose to input either separate count files (one per sample) or a single count matrix

Factor

Group

Group

Name

Condition

Name of experiment factor of interest (e.g. Genotype). One factor must be entered and there must be two or more groups per factor. Optional additional factors (e.g. Batch) can be entered using the Insert Factor button below. see Help section for more information. NOTE: Please only use letters, numbers or underscores.

Group

Group

Name

Name of group that the files belong to (e.g. WT or Mut). NOTE: Please only use letters, numbers or underscores (case sensitive).

Counts files

11: featureCounts KD2_rep1
10: featureCounts KD1_rep2
9: featureCounts KD1_rep1
8: featureCounts Control_rep2
7: featureCounts Control_rep1

Group

Group

Name

Condition

Name of experiment factor of interest (e.g. Genotype). One factor must be entered and there must be two or more groups per factor. Optional additional factors (e.g. Batch) can be entered using the Insert Factor button below. see Help section for more information. NOTE: Please only use letters, numbers or underscores.

Group

Group

Name

Name of group that the files belong to (e.g. WT or Mut). NOTE: Please only use letters, numbers or underscores (case sensitive).

Counts files

12: featureCounts KD2_rep2
11: featureCounts KD2_rep1
10: featureCounts KD1_rep2
9: featureCounts KD1_rep1
8: featureCounts Control_rep2

Group

Group

Name

Condition

Name of experiment factor of interest (e.g. Genotype). One factor must be entered and there must be two or more groups per factor. Optional additional factors (e.g. Batch) can be entered using the Insert Factor button below. see Help section for more information. NOTE: Please only use letters, numbers or underscores.

Group

Group

Name

Name of group that the files belong to (e.g. WT or Mut). NOTE: Please only use letters, numbers or underscores (case sensitive).

Counts files

12: featureCounts KD2_rep2
11: featureCounts KD2_rep1
10: featureCounts KD1_rep2
9: featureCounts KD1_rep1
8: featureCounts Control_rep2

Contrast

Contrast of Interest

Control-KD1

Names of two groups to compare separated by a hyphen e.g. Mut-WT. If the order is Mut-WT the fold changes in the results will be up/down in Mut relative to WT. If you have more than one contrast enter each separately using the Insert Contrast button below. For more info, see Chapter 8 in the limma User's guide: <https://www.bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf>

Contrast

Contrast of Interest

Control-KD2

Names of two groups to compare separated by a hyphen e.g. Mut-WT. If the order is Mut-WT the fold changes in the results will be up/down in Mut relative to WT. If you have more than one contrast enter each separately using the Insert Contrast button below. For more info, see Chapter 8 in the limma User's guide: <https://www.bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf>

Contrast

Contrast of Interest

KD1-KD2

Names of two groups to compare separated by a hyphen e.g. Mut-WT. If the order is Mut-WT the fold changes in the results will be up/down in Mut relative to WT. If you have more than one contrast enter each separately using the Insert Contrast button below. For more info, see Chapter 8 in the limma User's guide: <https://www.bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf>

Insert Contrast

Factor Low Counts

Output Options

Output Normalised Counts Table?

Yes No

Output a file containing the normalised counts, these are in log2 counts per million (logCPM). Default: No

Output Rscript?

Yes No

If this option is set to Yes, the Rscript used will be provided as a text file in the output. Default: No

Output RData file?

Yes No

Output all the data used by R to construct the plots and tables, can be loaded into R. A link to the RData file will be provided in the HTML report. Default: No

Advanced Options

Execute

Galaxy practical

Differential gene expression

Galaxy

Tools

deseq2

General

[StringTie](#) transcript assembly and quantification

RNA-seq

[edgeR](#) Perform differential expression of count data

[DESeq2](#) Determines differentially expressed features from count tables

[featureCounts](#) Measure gene expression in RNA-Seq experiments from SAM or BAM files.

Workflows

- All workflows

DESeq2 Determines differentially expressed features from count tables (Galaxy Version 2.11.40.2)

Factor

Specify a factor name, e.g. effects_drug_x or cancer_markers

Condition

Only letters, numbers and underscores will be retained in this field

Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

Control

Only letters, numbers and underscores will be retained in this field

Counts file(s)

11: featureCounts KD2_rep1
10: featureCounts KD1_rep2
9: featureCounts KD1_rep1
8: featureCounts Control_rep2
7: featureCounts Control_rep1

Factor

Specify a factor name, typical values could be 'tumor', 'normal', 'treated' or 'control'

KD1

Only letters, numbers and underscores will be retained in this field

Counts file(s)

12: featureCounts KD2_rep2
11: featureCounts KD2_rep1
10: featureCounts KD1_rep2
9: featureCounts KD1_rep1
8: featureCounts Control_rep2

Factor

Specify a factor name, typical values could be 'tumor', 'normal', 'treated' or 'control'

KD2

Only letters, numbers and underscores will be retained in this field

Counts file(s)

12: featureCounts KD2_rep2
11: featureCounts KD2_rep1
10: featureCounts KD1_rep2
9: featureCounts KD1_rep1
8: featureCounts Control_rep2

Files have header?

Yes No

If this option is set to Yes, the tool will assume that the count files have column headers in the first

Choice of Input data

Count data (e.g. from HTSeq-count, featureCounts or StringTie)

Count data (e.g. from HTSeq-count, featureCounts or StringTie)

Visualising the analysis results

Yes No

Output additional PDF files

Output normalized counts table

Yes No

Output all levels vs all levels of primary factor (use when you have >2 levels for primary factor)

Yes No

DESeq2 performs independent filtering by default using the mean of normalized counts as a filter statistic

Fit type

parametric

Turn off outliers replacement (only affects with >6 replicates)

Yes No

When there are more than 6 replicates for a given sample, the DESeq2 will automatically replace counts with large Cook's distance with the trimmed mean over all samples, scaled up by the size factor or normalization factor for that sample

Turn off outliers filtering (only affects with >2 replicates)

Yes No

When there are more than 2 replicates for a given sample, the DESeq2 will automatically filter genes which contain a Cook's distance above a cutoff

Turn off independent filtering

Yes No

DESeq2 performs independent filtering by default using the mean of normalized counts as a filter statistic

Execute

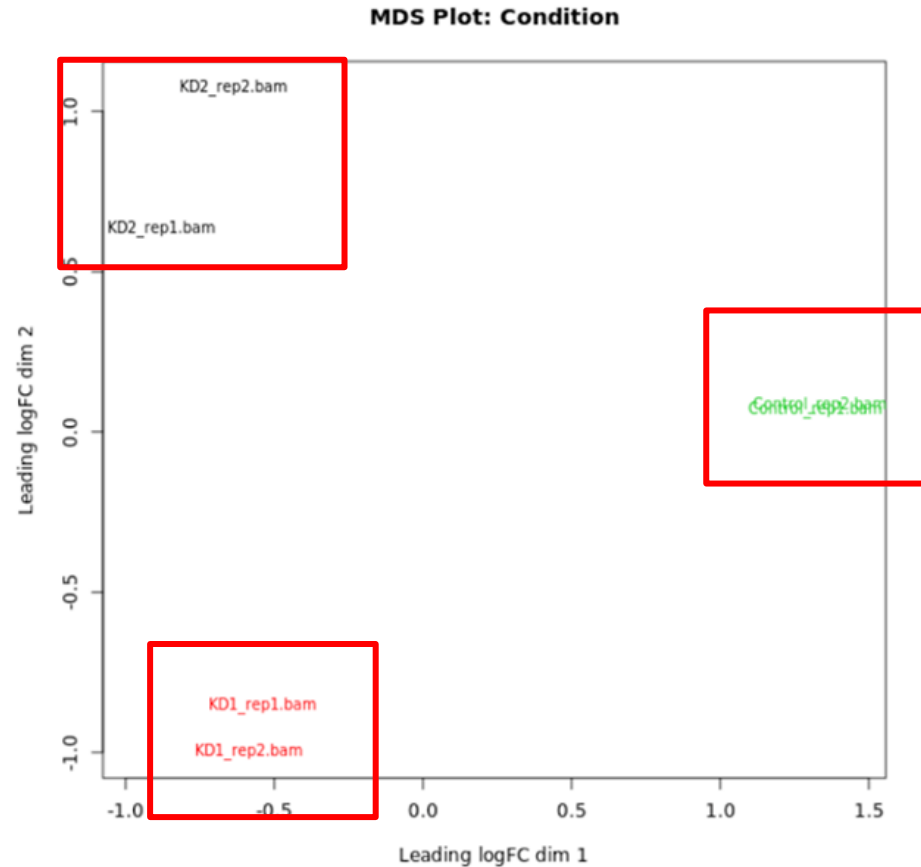
Galaxy practical

Differential gene expression



edgeR Analysis Output:

Links to PDF copies of plots are in 'Plots' section below.



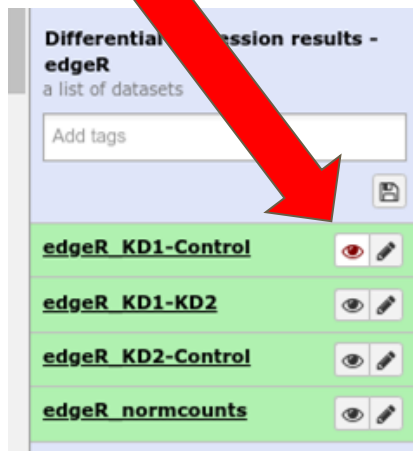
Galaxy practical

Differential gene expression



17 shown, 7 hidden
27.76 GB

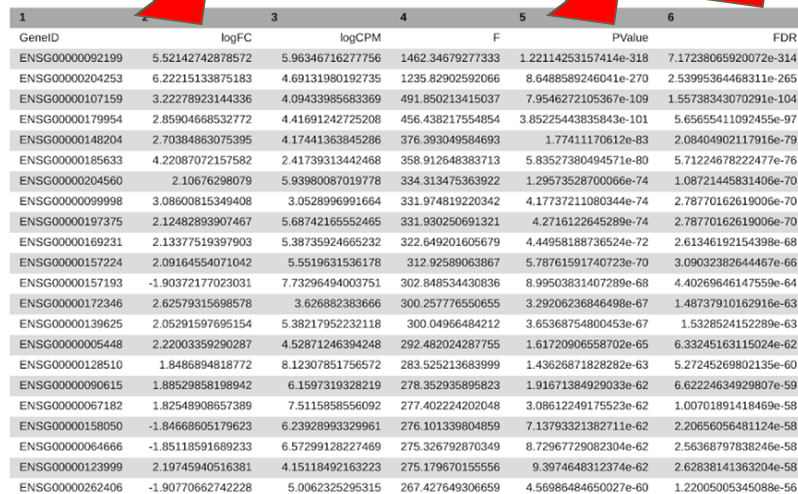
- 21: Normalized counts file on data 12, data 11, and others.
- 20: DESeq2 plots on data 12, data 11, and others.
- 19: DESeq2 result files on data 12, data 11, and others.
- 14: edgeR on data 12, data 11, and others: Repo
- 13: edgeR on data 12, data 11, and others: Tables
- 12: featureCounts KD2_rep2
- 11: featureCounts KD2_rep1



Differential gene expression results - edgeR
a list of datasets

Add tags

- edgeR_KD1-Control
- edgeR_KD1-KD2
- edgeR_KD2-Control
- edgeR_normcounts



1	2	3	4	5	6
GeneID	logFC	logCPM	F	PValue	FDR
ENSG00000092199	5.52142742878572	5.96346716277756	1462.34679277333	1.22114253157414e-318	7.17238065920072e-314
ENSG00000204253	6.22215133875183	4.69131980192735	1235.82902592066	8.6488589246041e-270	2.53995364468311e-265
ENSG00000107159	3.22278923144336	4.09433985683369	491.850213415037	7.9546272105367e-109	1.55738343070291e-104
ENSG00000179954	2.85904668532772	4.41691242725208	456.438217554854	3.85225443835843e-101	5.65655411092455e-97
ENSG00000148204	2.70384863075395	4.17441363845286	376.393049584693	1.77411170612e-83	2.08404902117916e-79
ENSG00000185633	4.22087072157582	2.41739313442468	358.912648383713	5.83527380494571e-80	5.71224678222477e-76
ENSG00000204560	2.10676298079	5.93980087019778	334.313475363922	1.29573528700066e-74	1.08721445831406e-70
ENSG00000099998	3.08600815349408	3.052896991664	331.974819220342	4.17737211080344e-74	2.78770162619006e-70
ENSG00000197375	2.12482893907467	5.68742165552465	331.930250691321	4.2716122645289e-74	2.78770162619006e-70
ENSG00000169231	2.13377519397903	5.38735924665232	322.649201605679	4.44958188736524e-72	2.61346192154398e-68
ENSG00000157224	2.09164554071042	5.5519631536178	312.92589063867	5.78761591740723e-70	3.09032382644467e-66
ENSG00000157193	-1.90372177023031	7.73296494003751	302.848534430836	8.99503831407289e-68	4.40269646147559e-64
ENSG00000172346	2.62579315695878	3.626882383666	300.257776550655	3.29206236846498e-67	1.48737910162916e-63
ENSG00000139625	2.05291597695154	5.38217952232118	300.04966484212	3.65368754800453e-67	1.5328524152289e-63
ENSG00000054448	2.22003359290287	4.52871246394248	292.482024287755	1.61720906558702e-65	6.33245163115024e-62
ENSG00000128510	1.8486894818772	8.12307851756572	283.525213683999	1.43626871828282e-63	5.27245269802135e-60
ENSG00000090615	1.88529858198942	6.1597319328219	278.352935895823	1.91671384929033e-62	6.62224634929807e-59
ENSG00000067182	1.82548908657389	7.5115858556092	277.402224202048	3.08612249175523e-62	1.00701891418469e-58
ENSG00000158050	-1.846686605179623	6.23928993329961	276.011339804859	7.13793321382711e-62	2.20656056481124e-58
ENSG00000064666	-1.85118591869233	6.57299128227469	275.326792870349	8.72967729082304e-62	2.56368797838246e-58
ENSG00000123999	2.19745940516381	4.15118492163223	275.179670155556	9.3974648312374e-62	2.62838141363204e-58
ENSG00000262406	-1.90770662742228	5.0062325295315	267.427649306659	4.56986484650027e-60	1.22005005345088e-56

Galaxy practical

Gene symbol annotation

But we do **not see any gene symbol/names** which **we all like**

Merge with **HUGO information** (<https://www.genenames.org/>)

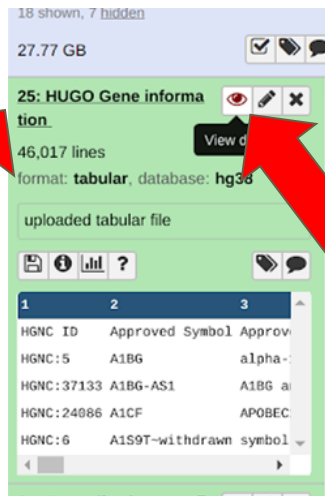
Share Data -> Data Libraries -> Bi5444

-> RNA-Seq -> HUGO Gene information

Galaxy practical

Gene symbol annotation

Merge with **HUGO** information



18 shown, 7 hidden

27.77 GB

25: HUGO Gene information

46,017 lines

format: tabular, database: hg38

uploaded tabular file

1	2	3
HGNC ID	Approved Symbol	Approved
HGNC:5	A1BG	alpha-
HGNC:37133	A1BG-AS1	A1BG a
HGNC:24886	A1CF	APOBEC
HGNC:6	A1S9T-withdrawn	symbol

Galaxy practical

Gene symbol annotation



1	2	3
HGNC ID	Approved Symbol	Approved Name
HGNC:5	A1BG	alpha-1-B glycoprotein
HGNC:37133	A1BG-AS1	A1BG antisense RNA 1
HGNC:24086	A1CF	APOBEC1 complementation factor
HGNC:6	A1S9T~withdrawn	symbol withdrawn, see UBA1
HGNC:7	A2M	alpha-2-macroglobulin
HGNC:27057	A2M-AS1	A2M antisense RNA 1
HGNC:23336	A2ML1	alpha-2-macroglobulin like 1
HGNC:41022	A2ML1-AS1	A2ML1 antisense RNA 1
HGNC:41523	A2ML1-AS2	A2ML1 antisense RNA 2
HGNC:8	A2MP1	alpha-2-macroglobulin pseudogene 1
HGNC:9	A2MR~withdrawn	symbol withdrawn, see LRP1
HGNC:10	A2MRAP~withdrawn	symbol withdrawn, see LRPAP1
HGNC:30005	A3GALT2	alpha 1,3-galactosyltransferase 2
HGNC:18149	A4GALT	alpha 1,4-galactosyltransferase (P blood group)
HGNC:17968	A4GNT	alpha-1,4-N-acetylglucosaminyltransferase

9	10	11	12	13
RefSeq IDs	Entrez Gene ID(supplied by NCBI)	RefSeq(supplied by NCBI)	Ensembl ID(supplied by Ensembl)	UCSC ID(supplied by UCSC)
NM_130786	1	NM_130786	ENSG00000121410	uc002qsd.5
NR_015380	503538	NR_015380	ENSG00000268895	uc002qse.3
NM_014576	29974	NM_001198818	ENSG00000148584	uc057tgv.1
NM_000014	2	NM_000014	ENSG00000175899	uc001qvk.2
NR_026971	144571	NR_026971	ENSG00000245105	uc009zgj.2
NM_144670	144568	NM_001282424	ENSG00000166535	uc001quz.6
	100874108		ENSG00000256661	uc058kxy.1
	106478979		ENSG00000256904	uc058kyb.1
NG_001067	3	NR_040112	ENSG00000256069	
NM_001080438	127550	NM_001080438	ENSG00000184389	uc031plq.1
NM_017436	53947	NM_001318038	ENSG00000128274	uc062ewl.1
NM_016161	51146	NM_016161	ENSG00000118017	uc003ers.2

Galaxy practical

Gene symbol annotation

Tools

join

Join, Subtract and Group

- CD-HIT-EST Cluster a nucleotide dataset into representative sequences
- CD-HIT_PROTEIN Cluster a protein dataset into representative sequences
- Join two Datasets side by side on a specified field
- Compare two Datasets to find common or distinct rows
- Group data by a column and perform aggregate operation on other columns.

Fetch Alignments/Sequences

- Join MAF blocks by Species

RNA-seq

Join two Datasets side by side on a specified field (Galaxy Version 2.1.1)

Join

13: edgeR on data 12, data 11, and others: Tables

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Joining column

Column: 1

with

22: HUGO Gene information

Joining column

Column: 12

Keep lines of first input that do not join with second input

Yes

Keep lines of first input that are incomplete

Yes

Fill empty columns

Yes

Only fill unjoined rows

Yes

Fill Columns by

Single fill value

Fill value

Keep the header lines

Yes

Execute

Galaxy practical

Differential gene expression



search datasets

Bi5444_RNA-Seq_DE
19 shown, 11 hidden
27.81 GB

26: Join two Datasets on collection 13

a list with 4 items

Control_rep1 Control_rep2 KD1_rep1

22: HUGO Gene information

46,017 lines
format: **tabular**, database: **hg38**
uploaded tabular file

1	2	3
HGNC ID	Approved Symbol	Approved
HGNC:5	A1B6	alpha-1-B
HGNC:37133	A1B6-AS1	A1B6 ant1
HGNC:24086	A1CF	APOBEC1 c
HGNC:6	A1S9T-withdrawn	symbol w1

21: Normalized counts file on data 12, data 11, and others.

Control_rep1 Control_rep2 KD1_rep1

History

< Back Bi5444_RNA-Seq_DE

Join two Datasets on collection 13
a list with 4 items

#Control_rep1
#Control_rep2
#KD1_rep1
#KD1_rep2

edgeR_Control-KD1

58,739 lines
format: **tabular**, database: **GRCn38**

1	2	3
GeneID	logFC	logCP
ENSG00000092199	5.52142742878572	5.963
ENSG00000204253	6.22215133875183	4.691
ENSG00000107159	3.22278923144336	4.094
ENSG00000179954	2.85904668532772	4.416

edgeR_Control-KD2

edgeR_KD1-KD2

edgeR_normcounts

Galaxy practical

Differential gene expression

1	2	3	4	5	6	7	8	9	10	11
GeneID	logFC	logCPM	F	PValue	FDR	HGNC ID	Approved Symbol	Approved Name	Status	Previous Symbols
ENSG00000092199	5.52142742878572	5.96346716277756	1462.346792777333	1.22114253157414e-318	7.17238065920072e-314	HGNC:5035	HNRNPC	heterogeneous nuclear ribonucleoprotein C (C1/C2)	Approved	HNRPC
ENSG00000204233	8.22219133879183	4.89131980292733	1235.82902392068	8.0485582400418e-270	2.329953649663218e-265	HGNC:48849	HNRNPC2	heterogeneous nuclear ribonucleoprotein C pseudogene 2	Approved	
ENSG00000107159	3.22278923144336	4.0943396583369	491.850213415037	7.9546272105367e-109	1.55738343070291e-104	HGNC:1383	CA9	carbonic anhydrase 9	Approved	
ENSG00000179954	2.85904668532772	4.41691242725208	456.438217554854	3.85225443835843e-101	5.65655411092455e-97	HGNC:26641	SSC5D	scavenger receptor cysteine rich family member with 5 domains	Approved	
ENSG00000148204	2.70384863075395	4.17441363845286	376.393049584693	1.77411170612e-83	2.08404902117916e-79	HGNC:18688	CRB2	crumbs 2, cell polarity complex component	Approved	
ENSG00000185633	4.22087072157582	2.41739313442468	358.912648383713	5.83527380494571e-80	5.71224678222477e-76	HGNC:29836	NDUFA4L2	NDUFA4, mitochondrial complex associated like 2	Approved	
ENSG00000204560	2.10676298079	5.93960087019778	334.313475363922	1.29573528700066e-74	1.08721445831406e-70	HGNC:2739	DHX16	DEAH-box helicase 16	Approved	DDX16
ENSG00000099998	3.08600815349408	3.0528969911664	331.974819220342	4.17737211080344e-74	2.78770162619005e-70	HGNC:4260	GGT5	gamma-glutamyltransferase 5	Approved	GGT1A1
ENSG00000197375	2.12482893907467	5.68742165552465	331.930250691321	4.2716122645289e-74	2.78770162619005e-70	HGNC:10989	SLC22A5	solute carrier family 22 member 5	Approved	CDSP
ENSG00000169231	2.13377519397903	5.38735924665232	322.649201605679	4.44958188736524e-72	2.61346192154399e-68	HGNC:11787	THBS3	thrombospondin 3	Approved	
ENSG00000157224	2.09164554071042	5.5519631536178	312.925896063867	5.78761591740723e-70	3.09032382644467e-66	HGNC:2034	CLDN12	claudin 12	Approved	
ENSG00000157193	-1.90372177023031	7.73296494003751	302.848534430836	8.99503831407289e-68	4.40269646147559e-64	HGNC:6700	LRP8	LDL receptor related protein 8	Approved	
ENSG00000172346	2.62579315696578	3.626882383666	300.257776550655	3.29206236846498e-67	1.48737910162916e-63	HGNC:30359	CSDC2	cold shock domain containing C2	Approved	
ENSG00000139625	2.05291597695154	5.38217952232118	300.04966484212	3.65368754800453e-67	1.5328524152289e-63	HGNC:6851	MAP3K12	mitogen-activated protein kinase kinase kinase 12	Approved	ZPK
ENSG00000005448	2.22003359290287	4.52871246394248	292.482024287755	1.61720900558702e-65	6.33245163115024e-62	HGNC:25770	WDR54	WD repeat domain 54	Approved	
ENSG00000128510	1.8486894818772	8.12307851756572	283.525213683999	1.43626871828282e-63	5.27245269802135e-60	HGNC:15740	CPA4	carboxypeptidase A4	Approved	
ENSG00000096115	1.88529858198942	6.1597319328219	278.352935895823	1.91671384929033e-62	6.62224634929807e-59	HGNC:4426	GOLGA3	golgin A3	Approved	
ENSG00000067182	1.82548908657389	7.51158585566092	277.402224202048	3.08612249175523e-62	1.00701891418469e-58	HGNC:11916	TNFRSF1A	TNF receptor superfamily member 1A	Approved	TNFR1
ENSG00000158050	-1.84668605179623	6.23928993329961	276.101339804859	7.13793321382711e-62	2.20656056481124e-58	HGNC:3068	DUSP2	dual specificity phosphatase 2	Approved	
ENSG00000064666	-1.85118591689233	6.57299128227469	275.326792870349	8.72967729082304e-62	2.56368797838245e-58	HGNC:2156	CNN2	calponin 2	Approved	
ENSG00000123999	2.19745940516381	4.15118492163223	275.179670155556	9.3974648312374e-62	2.62838141363204e-58	HGNC:6065	INH1A	inhibin subunit alpha	Approved	
ENSG00000262406	-1.90770662742228	5.0062325295315	267.427649306659	4.56986484650027e-60	1.22005005345088e-56	HGNC:7158	MMP12	matrix metalloproteinase 12	Approved	
ENSG00000159792	1.87315272517581	5.81432694173605	265.276959847098	1.34261834082599e-59	3.42863861949629e-56	HGNC:9529	PSKH1	protein serine kinase H1	Approved	

Galaxy practical

Alignment coverage

Visualization of **coverage** of **aligned** data (and expressed exons)

Using `deepTools` -> `bamCoverage`

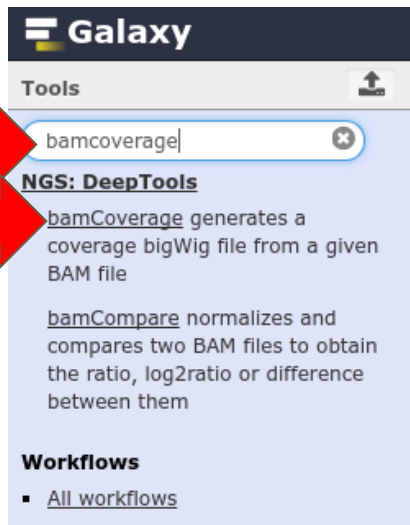
Input **BAM**, output **BIGWIG**

Effective genome size (hg38): **2913022398**

Galaxy practical

<https://deeptools.readthedocs.io/en/develop/content/feature/effectiveGenomeSize.html>

Alignment coverage



Galaxy

Tools

bamcoverage

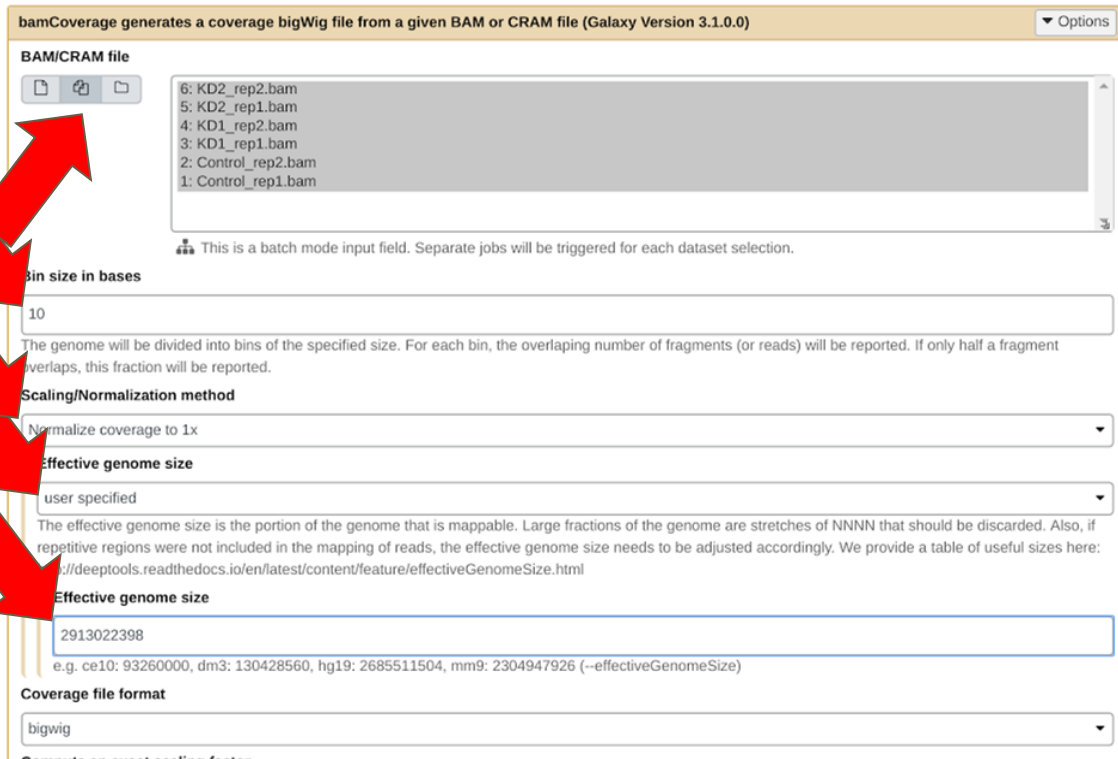
NGS: DeepTools

bamCoverage generates a coverage bigWig file from a given BAM file

bamCompare normalizes and compares two BAM files to obtain the ratio, log2ratio or difference between them

Workflows

- All workflows



bamCoverage generates a coverage bigWig file from a given BAM or CRAM file (Galaxy Version 3.1.0.0)

BAM/CRAM file

6: KD2_rep2.bam
5: KD2_rep1.bam
4: KD1_rep2.bam
3: KD1_rep1.bam
2: Control_rep2.bam
1: Control_rep1.bam

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Bin size in bases

10

The genome will be divided into bins of the specified size. For each bin, the overlapping number of fragments (or reads) will be reported. If only half a fragment overlaps, this fraction will be reported.

Scaling/Normalization method

Normalize coverage to 1x

Effective genome size

user specified

The effective genome size is the portion of the genome that is mappable. Large fractions of the genome are stretches of NNNN that should be discarded. Also, if repetitive regions were not included in the mapping of reads, the effective genome size needs to be adjusted accordingly. We provide a table of useful sizes here: <https://deeptools.readthedocs.io/en/latest/content/feature/effectiveGenomeSize.html>

Effective genome size

2913022398

e.g. ce10: 93260000, dm3: 130428560, hg19: 2685511504, mm9: 2304947926 (--effectiveGenomeSize)

Coverage file format

bigwig

Galaxy practical

Alignment coverage

BIGWIG coverage

The screenshot shows the 'History' panel in Galaxy. At the top, there is a search bar labeled 'search datasets'. Below it, the dataset group 'RNA-2018-03-Expression' is shown with '18 shown' items and a size of '148.17 MB'. The list of datasets is as follows:

- 18: KD1_rep1.bigWig (highlighted in green)
- 17: Control_rep1.bigWig (highlighted in green)
- 16: KD2_rep1.bigWig (highlighted in green)
- 15: Normalized counts with annotation - DES eq2 (highlighted in green)

Three red arrows point from the right side of the image to the dataset names: one to '18: KD1_rep1.bigWig', one to '17: Control_rep1.bigWig', and one to '16: KD2_rep1.bigWig'. The '15: Normalized counts with annotation - DES eq2' dataset has icons for visibility, edit, and delete.

Galaxy practical

Alignment coverage

BIGWIG visualization

The screenshot shows the Galaxy History panel. At the top, there is a search bar labeled "search datasets". Below it, a dataset entry for "RNA-2018-03-Expression" is shown with "18 shown" and a size of "148.17 MB". A red arrow points to the eye icon next to the entry "18: KD1_rep1.bigWig". Below this, the file size is "38.0 MB" and the format is "bigwig, database: hg38". A green box contains configuration details: "normalization: 1x (effective genome size 2913022398)", "defaultFragmentLength: read length", "blackListFileName: None", "minMappingQuality: None", "maxPairedFragmentLength: 1000", "bedFile: None", "chrsToSkip: []", "binLength: 10", "save_data: False", "maxFragmentLength: 0", and "numberO". A red arrow points to the "bigwig, database: hg38" text. At the bottom, there are icons for file operations and a red arrow points to the text "display at UCSC [main](#)". Other options include "display in IGB [View](#)" and "display with IGV [web](#) [current](#) [local](#)". A text input field at the bottom is labeled "Binary UCSC BigWig file".

Galaxy practical

Alignment coverage

CD55 region

Downloads My Data View Help About Us

UCSC Genome Browser

Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr1:207,321,508-207,360,966 39,499 bp. CD55 go

CD55 (Homo sapiens CD55 molecule (Cromer blood group) (CD55), transcript variant 7, mRNA. (from RefSeq NM_001300904))

Scale chr1: 207,325,000| 207,330,000| 207,335,000| 207,340,000| 207,345,000| 207,350,000| 207,355,000| 207,360,000 hg38

-knockdown-highlig

GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)

CD55

RefSeq Curated

OMIM Alleles

Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (578 donors)

CD55

Layered H3K27ac

100

0

4.00

DNase Clusters

DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)

100 vertebrates Basewise Conservation by PhyloP

Cons 100 Verts

-4.5

Multiz Alignments of 100 Vertebrates

Rhesus

Mouse

Dog

Elephant

Chicken

X_tropicalis

Zenaidura

Lamprey

Common SNPs (150)

SINE

LINE

LTR

Other

Simple

Low Complexity

Satellite

RNA

Other

Unknown

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

move start < 2.0 > move end < 2.0 >

track search default tracks default order hide all manage custom tracks track hubs configure multi-region reverse resize refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. expand all

Galaxy practical

Alignment coverage

BIGWIG size

Downloads My Data View Help About Us

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr1:207,321,508-207,360,966 39,459 bp. enter position, gene symbol, HGVS or search terms go

chr1 (q32.2) 1p31.1 1q12 1q41 [44]

Scale chr1: 207,325,000 207,330,000 207,335,000 10 kb hg38 207,340,000 207,345,000 207,350,000 207,355,000 207,360,000

-knockdown-bigmig RINseq-HNRHPC-knockdown-bigmig

GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)

RefSeq Curated RefSeq gene predictions from NCBI

OMIM Alleles OMIM Allelic Variants

Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (576 donors)

CDSS

Layered HSK279c HSK279c Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE

DNase Clusters DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)

Cons 100 Verts 100 vertebrates Basewise Conservation by PhyloP

Multiz Alignments of 100 Vertebrates

Rhesus Mouse Dog Elephant Chicken X_tropicalis Zebrafish Lambrey

Common SNPs(158) Simple Nucleotide Polymorphisms (dbSNP 158) Found in 100 Repeating Elements by RepeatMasker

SINE LINE LTR DNA Simple Low Complexity Satellite RNA Other Unknown

move start < 2.0 > move end < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click on bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

track search default tracks default order hide all manage custom tracks track hubs configure multi-region reverse refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes. expand all

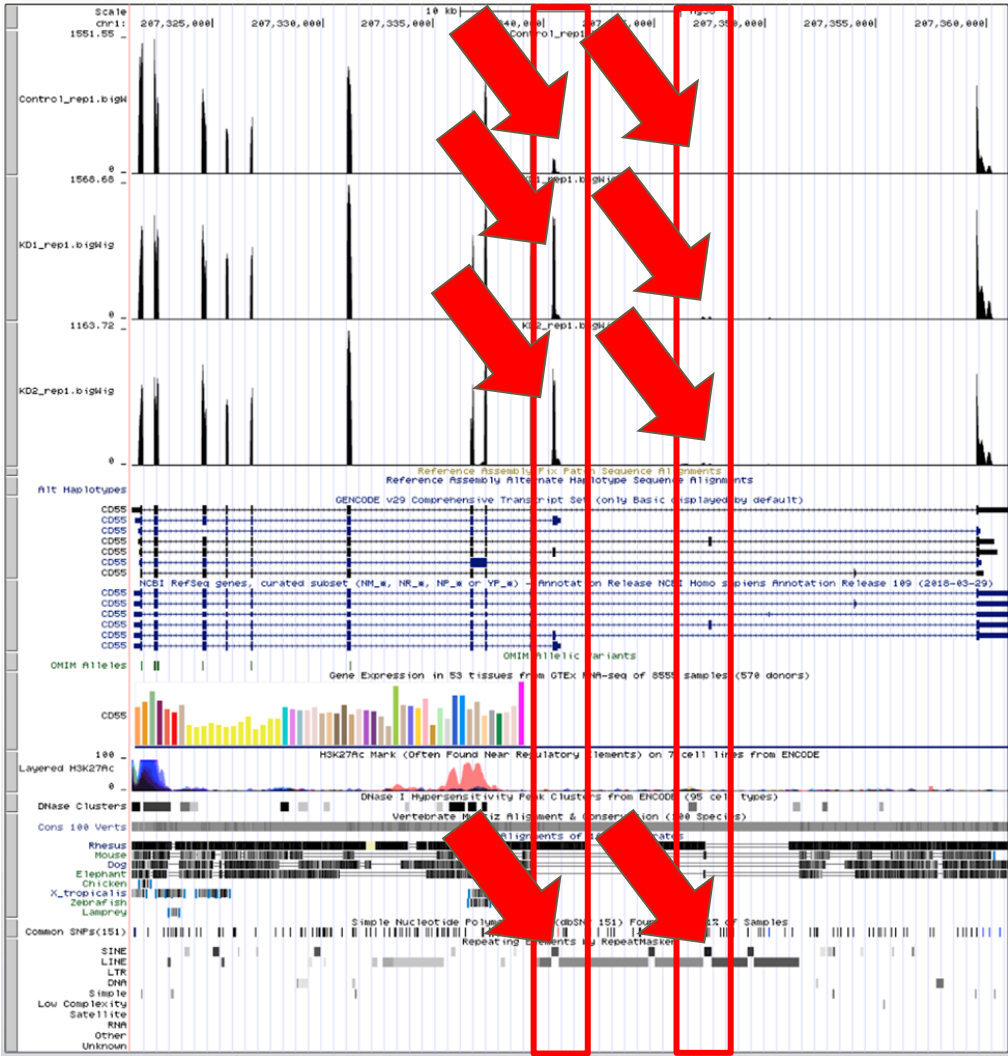
Custom Tracks refresh

Control_rep1.bigWig	KD1_rep1.bigWig	KD2_rep1.bigWig
full ▾	full ▾	full ▾

Galaxy practical

Alignment coverage

Now you do it for the **other two BIGWIG files**



Galaxy practical

Alignment coverage

If something went wrong, history of DE and coverage visualization

Shared Data -> Histories -> Bi5444 RNA-Seq DE full history

RNA-Seq data analysis - pipeline in Galaxy

1. **Initial quality check** - `FastQC`
 - Check for overall quality of the data, number of reads, read length distribution, ...
2. **Preprocessing** - `Trimmomatic`
 - Remove adapters, low quality ends, unwanted sequences, ...
3. **Alignment** - `STAR`
 - Map reads to the reference genome
4. **Alignment quality check** - `STAR log`, `featureCounts`
 - Check overall alignment statistics
5. **Genome coverage (peaks)** - `bamCoverage`
 - Get overview of mapped positions in the genome
6. **Gene annotation** - `UCSC Main table browser`
 - Get gene annotations for reference genome
7. **Quantification** - `featureCounts`
 - Get gene read counts
8. **Differential gene expression** - `edgeR`, `DESeq2` (genes)
 - Differences between conditions

RNA-Seq data analysis - other possibilities

- 1. Initial quality check** - FastQC
 - Check for overall quality of the data, number of reads, read length distribution, ...
- 2. Preprocessing** - Cutadapt, BBTools, seqtk
 - Remove adapters, low quality ends, unwanted sequences, ...
- 3. Alignment** - GSNAP, Bowtie2
 - Map reads to the reference genome
- 4. Alignment quality check** - Picard tools, RSeQC, Qualimap
 - Check overall alignment statistics
- 5. Genome coverage (peaks)** - STAR + bedGraphToBigWig, Bedtools
 - Get overview of mapped positions in the genome
- 6. Gene annotation** - UCSC, Ensembl, NCBI
 - Get gene annotations for reference genome
- 7. Quantification** - RSEM, HTSeq, Salmon, Kallisto
 - Get gene read counts
- 8. Differential gene expression** - DEXSeq (exons), baySeq (genes)
 - Differences between conditions
- 9. Gene ontology and pathways** - g:Profiler, KEGG
 - Check ontologies and pathways for selected genes