

# **De novo transcriptome reconstruction with RNA-Seq**

---

Identify the role of the transcription factor Tal1 in gene regulation, a critical regulator of hematopoiesis, at multiple stages of hematopoietic differentiation.

GEO Accession: GSE51338

- 1) G1E - a GATA-null immortalized cell line derived from targeted disruption of GATA-1 in mouse embryonic stem cells
- 2) Megakaryocytes

## **AIM:**

The aim of this study is to use RNA sequencing (RNA-seq) to explore gene expression in G1E and megakaryocyte cells and to understand how the transcription factor Tal1 influences gene regulation during key stages of blood cell development.

## **OUTLINE:**

1. Analysis strategy
  1. Data upload
  2. Quality control
2. Mapping
3. De novo transcript reconstruction
4. Transcriptome assembly
5. Analysis of the differential gene expression
  1. Count the number of reads per transcript
  2. Perform differential gene expression testing
6. Visualization
7. Conclusion

# De novo transcriptome reconstruction with RNA-Seq

## Data upload:

This data is available at [zenodo](#), where you can find the forward and reverse reads corresponding to replicate RNA-seq libraries from G1E and megakaryocyte cells and an annotation file of RefSeq transcripts we will use to generate our transcriptome database.

## Data upload:

1. Create a new history for this RNA-seq exercise
  - Tip: Creating a new history
  - Tip: Renaming a history
2. Open the data upload manager (Get Data -> Upload file)
3. Copy and paste the links for the reads and annotation file
4. Select Paste/Fetch Data
5. Paste the link(s) into the text field
6. Change the datatype of the read files to fastqsanger
7. Change the datatype of the annotation file to gtf and assign the Genome as mm10
8. Press Start
9. Rename the files in your history to retain just the necessary information (e.g. "G1E R1 forward reads")
10. Import the files from [Zenodo](#)

# De novo transcriptome reconstruction with RNA-Seq

This is the Indian galaxy server, Welcome Home!

Namaste All! Welcome to the India Galaxy workbench; a comprehensive set of tools and workflows dedicated to accelerate your bioinformatics analyses focusing Bharat. This workbench is built on the Galaxy framework thanks to Galaxy Europe we are hosting the seeds of Indian Galaxy to facilitate more wider user base.

To know more about the events, training webinars focused on Galaxy for Indian data analysis community visit [www.galaxyproject.in](http://www.galaxyproject.in). If you want to talk about Galaxy and any required help about it join the gitter [adda \(गृह\)](#).

**Training**

For knowledge and about practical implementation about Galaxy framework there is a wide variety of collection of tutorials available. We strongly recommend to use the resources provided by [Galaxy Training Network \(GTN\)](#). Also we endorse to contribute and develop training materials of data analysis based on Galaxy ([Batut et al., 2017](#)).

**Appreciation**

We sincerely appreciate the efforts of all individuals and their organizations ([Bioclues](#), [TMS Foundation](#)) for supporting the Indian Galaxy Instance.

History + ⌂ ⌂ ⌂

search datasets

Unnamed history

De novo transcriptome reconstruction

Add Tags

Save Cancel

0 B

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

Training

For knowledge and about practical implementation about Galaxy framework there is a wide variety of collection of tutorials available. We strongly recommend to use the resources provided by [Galaxy Training Network \(GTN\)](#). Also we endorse to contribute and develop training materials of data analysis based on Galaxy ([Batut et al., 2017](#)).

**Appreciation**

We sincerely appreciate the efforts of all individuals and their organizations ([Bioclues](#), [TMS Foundation](#)) for supporting the Indian Galaxy Instance.

**Our Data Policy**

Registered Users	Unregistered Users	FTP Data	GDPR Compliance
User data on UseGalaxy.eu (i.e. datasets, histories) will be available as long as they are not deleted by the user. Once deleted, the data is removed from the system.	Processed data will only be accessible during one browser session, using a cookie to identify your data.	Any user data uploaded to our <a href="#">FTP server</a> should be imported.	The Galaxy service complies with the EU General Data Protection Regulation.

History + ⌂ ⌂ ⌂

search datasets

De novo transcriptome reconstruction

0 B

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

# De novo transcriptome reconstruction with RNA-Seq

Galaxy India

Upload

Tools

Get Data

Send Data

Collection Operations

GENERAL TEXT Tools

Text Manipulation

Convert Format

Filter and Sort

Join, Subtract

GENOMIC FILE MANAGEMENT

Convert Format

FASTA/FASTQ

Quality Control

SAM/BAM

BED

VCF/BCF

New File 671 b fastqsanger Mouse (Mus Muscu... 100%

Download data from the web by entering URLs (one per line) or directly paste content.

https://zenodo.org/record/583140/files/Megakaryocyte\_rep1\_reverse\_read\_%285RR549357\_2%29  
https://zenodo.org/record/583140/files/Megakaryocyte\_rep2\_forward\_read\_%285RR549358\_1%29  
https://zenodo.org/record/583140/files/Megakaryocyte\_rep2\_reverse\_read\_%285RR549358\_2%29

Type (set all): Auto-detect Reference (set all): unspecified (?)

Choose local file Choose remote files Paste/Fetch data Start Pause Reset Close

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

Attributes Datatypes Permissions

Name RefSeq\_reference\_GTF\_%28DSv2%29

Info uploaded gtf file

Annotation - optional

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

Database/Build - optional

Mouse Dec. 2011 (GRCm38/mm10) (mm10)

Please provide a value for this option.

Number of comment lines

Save Auto-detect

History

De novo transcriptome reconstruction

1.73 GB

9: RefSeq\_reference\_GTF\_%28DSv2%29

Add Tags

4,236 lines 9 columns

format gtf, database mm10

uploaded gtf file

1.Seqname	2.Source	3.Feature	4
chr1	mm10_refGene	exon	1
chr1	mm10_refGene	start_codon	1
chr1	mm10_refGene	CDS	1
chr1	mm10_refGene	exon	1
chr1	mm10_refGene	CDS	1

# De novo transcriptome reconstruction with RNA-Seq

## This is the Indian galaxy server, Welcome Home!

Namaste All! Welcome to the India Galaxy workbench; a comprehensive set of tools and workflows dedicated to accelerate your bioinformatics analyses focusing Bharat. This workbench is built on the Galaxy framework thanks to Galaxy Europe we are hosting the seeds of Indian Galaxy to facilitate more wider user base.

To know more about the [events](#), training [webinars](#) focused on Galaxy for Indian data analysis community visit [www.galaxyproject.in](http://www.galaxyproject.in). If you want to talk about Galaxy and any required help about it join the gitter adda ([അറ്റം](#)).

1. [Training](#)
2. [Appreciation](#)

## Training

For knowledge and about practical implementation about Galaxy framework there is a wide variety of collection of tutorials available. We strongly recommend to use the resources provided by [Galaxy Training Network \(GTN\)](#). Also we endorse to contribute and develop training materials of data analysis based on Galaxy ([Batut et al., 2017](#)).

## Appreciation

We sincerely appreciate the efforts of all individuals and their organizations ([Bioclues](#), [TMS Foundation](#)) for supporting the Indian Galaxy Instance.

The screenshot shows the 'History' panel of the Indian Galaxy workbench. At the top, there is a search bar labeled 'search datasets'. Below the search bar, the title 'De novo transcriptome reconstruction' is displayed. Underneath the title, the total size of the dataset is listed as '1.73 GB'. The history panel lists five datasets, each represented by a green card:

- 9: RefSeq\_reference\_GTF\_%2  
8DSv2%29
- 8: Megakaryocyte\_rep2\_rever  
se\_read\_%28SRR549358\_2%  
29
- 7: Megakaryocyte\_rep2\_forwa  
rd\_read\_%28SRR549358\_1%  
29
- 6: Megakaryocyte\_rep1\_rever  
se\_read\_%28SRR549357\_2%  
29
- 5: Megakaryocyte\_rep1\_forwa  
rd\_read\_%28SRR549357\_1%  
29

Each dataset card includes icons for viewing, editing, and deleting the dataset.

# De novo transcriptome reconstruction with RNA-Seq

## Quality control:

For quality control, we use similar tools as described in the NGS-QC tutorial: FastQC and Trimmomatic.

## Quality control:

- FastQC tool Run `FastQC` on the forward and reverse read files to assess the quality of the reads.
- Trimmomatic tool Trim off the low quality bases from the ends of the reads to increase mapping efficiency
- FastQC tool Re-run `FastQC` on trimmed reads and inspect the differences.
- Trimmomatic tool: Run `Trimmomatic` on the remaining forward/reverse read pairs with the same parameters.

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for the India instance. The left sidebar includes options for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, History (selected), Notifications, and Settings. The main panel displays the 'Tools' section with 'FastQC' selected. The 'Tool Parameters' section contains a 'Raw read data from your current history' input field containing 16 entries related to sequencing runs. Below it is a 'Contaminant list' input field with 'Nothing selected'. The right panel shows the 'History' tab with a list of 20 datasets, all of which are green and labeled as successful. The top right corner indicates 'Using 15%'.

This screenshot shows the Galaxy interface after the FastQC tool has been run. The main panel displays a success message: 'Started tool FastQC and successfully added 8 jobs to the queue. It produces 16 outputs:' followed by a bulleted list of 25 items. The bottom of this panel contains a note about checking job status. The right panel's 'History' tab now lists 25 datasets, all green and marked as successful. The top right corner shows 'Using 15%'.

# De novo transcriptome reconstruction with RNA-Seq

**FastQC Report**

Thu 10 Oct 2024  
G1E\_rep1\_forward\_read\_X28SRR549355\_1X29

**Summary**

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ 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

Produced by [FastQC](#) (version 0.12.1)

**Basic Statistics**

Measure	Value
Filename	G1E_rep1_forward_read_X28SRR549355_1X29
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1236775
Total Bases	122.4 Mbp
Sequences flagged as poor quality	0
Sequence length	99
%GC	63

History + ⌂ ⌂ ⌂  
search dataset ⌂ ⌂ ⌂

De novo transcriptome reconstruction

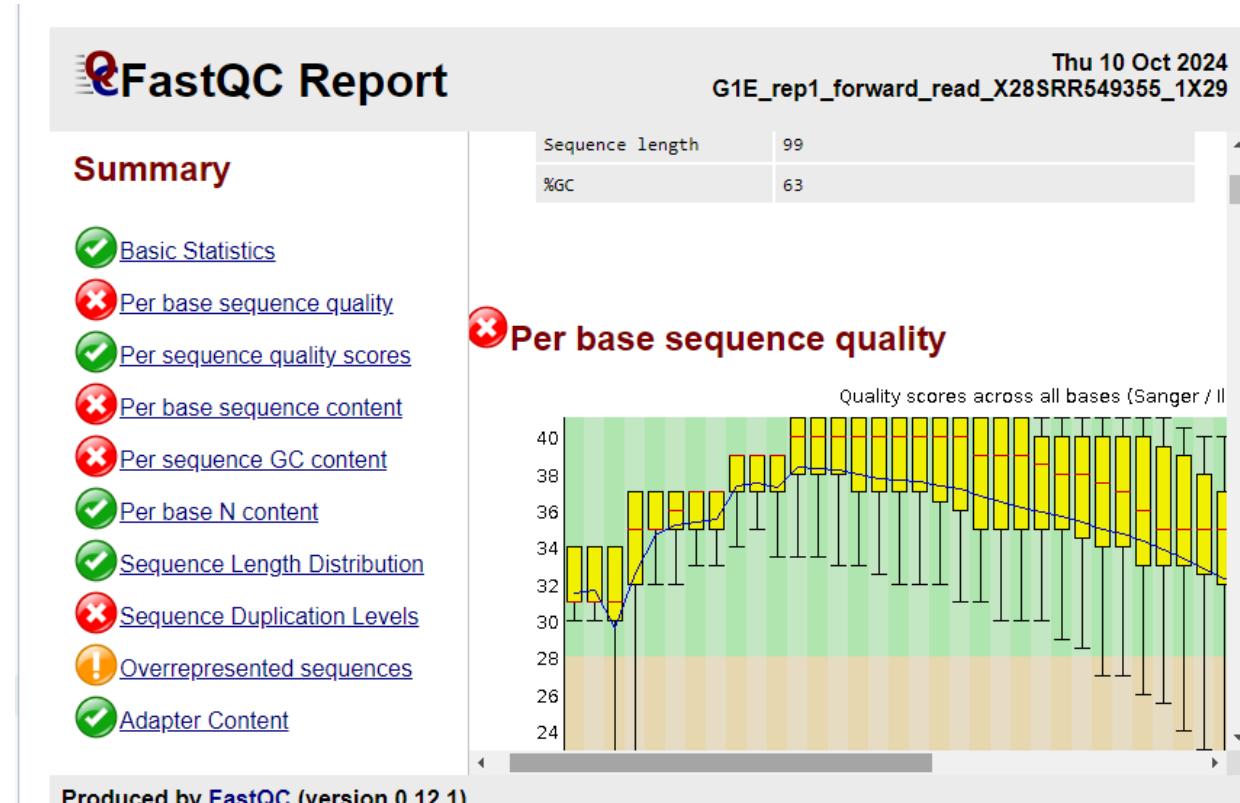
1: 1.76 GB ⌂ ⌂ ⌂ 25 ⌂  
ta

10: FastQC on data 1: Webpage

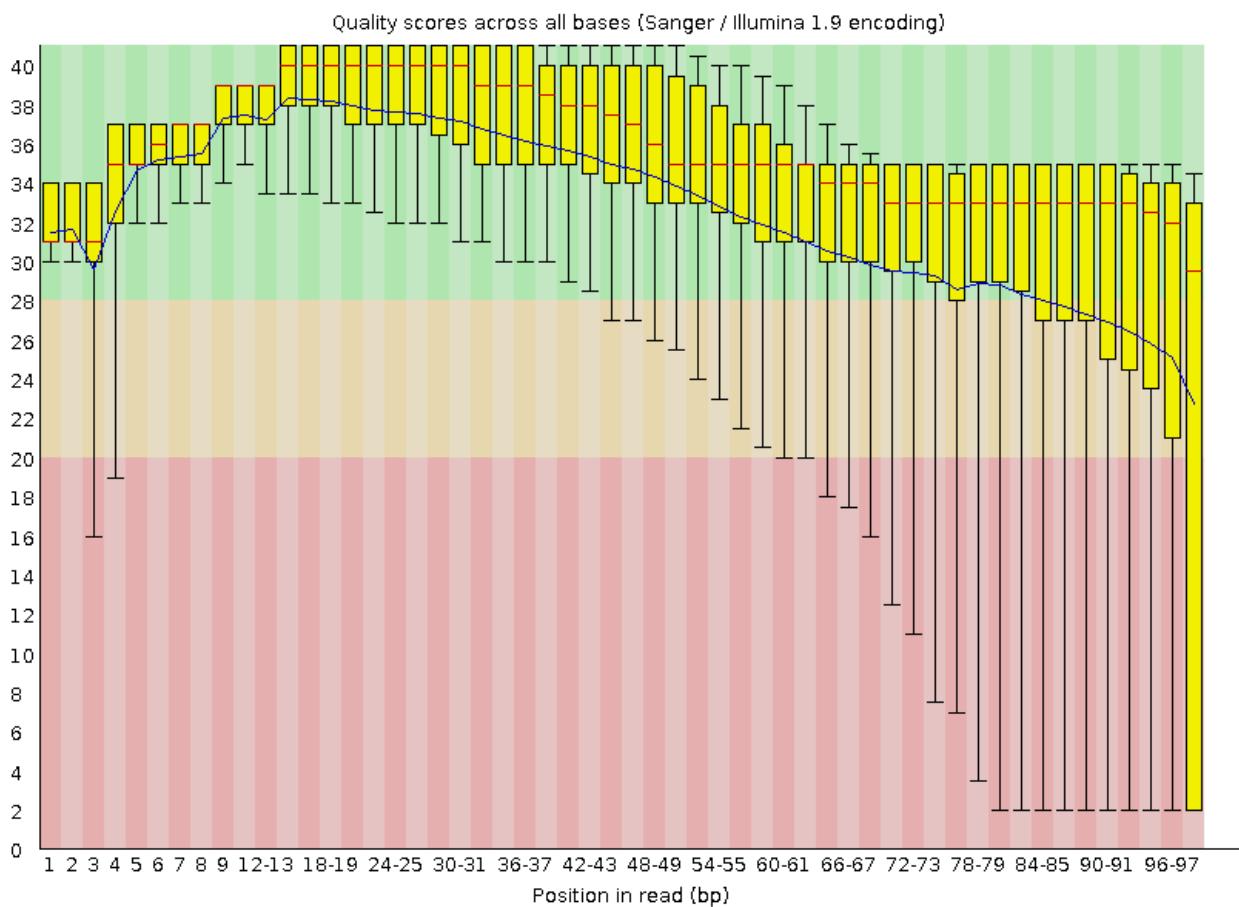
9: RefSeq\_refERENCE\_GTF\_29

8: Megakaryocyte\_rep2\_revErse\_read\_%2SRR549358\_2%29

# De novo transcriptome reconstruction with RNA-Seq



# De novo transcriptome reconstruction with RNA-Seq



# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows a bioinformatics pipeline interface with a terminal window on the left displaying FastQC results and a history panel on the right.

**FastQC Output:**

```

##FastQC      0.12.1
>>Basic Statistics    pass
#Measure      Value
Filename      G1E_rep1_forward_read_X28SRR549355_1X29
File type     Conventional base calls
Encoding      Sanger / Illumina 1.9
Total Sequences 1236775
Total Bases   122.4 Mbp
Sequences flagged as poor quality 0
Sequence length 99
%GC          63
>>END_MODULE
>>Per base sequence quality fail
#Base  Mean Median Lower Quartile Upper Quartile 10th Percentile 90th Percentile
1    31.49887570495846 31.0 31.0 34.0 30.0 34.0
2    31.684624931778213 31.0 31.0 34.0 30.0 34.0
3    29.628761900911645 31.0 30.0 34.0 16.0 34.0
4    32.58685937215743 35.0 32.0 37.0 19.0 37.0
5    34.62401164318489 35.0 35.0 37.0 32.0 37.0
6    35.18942572416163 36.0 35.0 37.0 32.0 37.0
7    35.352161872612236 37.0 35.0 37.0 33.0 37.0
8    35.52906227891088 37.0 35.0 37.0 33.0 37.0
9    37.35775060136241 39.0 37.0 39.0 34.0 39.0
10-11 37.480726486224256 39.0 37.0 39.0 35.0 39.0
12-13 37.26252228578359 39.0 37.0 39.0 33.5 39.0
14-15 38.37824624527501 40.0 38.0 41.0 33.5 41.0
16-17 38.30562915647551 40.0 38.0 41.0 33.5 41.0
18-19 38.19405914576217 40.0 38.0 41.0 33.0 41.0
20-21 37.95490731943967 40.0 37.0 41.0 33.0 41.0
22-23 37.7308685896788 40.0 37.0 41.0 32.5 41.0
24-25 37.646657637808005 40.0 37.0 41.0 32.0 41.0
26-27 37.56749004467264 40.0 37.0 41.0 32.0 41.0
28-29 37.37290776414466 40.0 36.5 41.0 32.0 41.0
30-31 37.18241151381618 40.0 36.0 41.0 31.0 41.0
32-33 36.7905055487053 39.0 35.0 41.0 31.0 41.0
34-35 36.4770714155768 39.0 35.0 41.0 30.0 41.0

```

**History:**

- 11: FastQC on data 1: RawData
- 10: FastQC on data 1: Webpage
- 9: RefSeq\_refERENCE\_GTF
- 29
- 8: Megakaryo

The screenshot shows the Galaxy platform interface with a tool configuration panel on the left and a history panel on the right.

**Tool Configuration:** Trimmomatic

**Tool Parameters:**

- Single-end or paired-end reads? Paired-end (two separate input files)
- Input FASTQ file (R1/first of pair): G1E\_rep1\_forward\_read\_X28SRR549355\_1X29
- Input FASTQ file (R2/second of pair): G1E\_rep1\_reverse\_read\_X28SRR549355\_2X29
- Perform initial ILLUMINACLIP step? no
- Cut adapter and other Illumina-specific sequences from the read
- Trimmomatic Operation

**History:**

- 25: FastQC on data 8: RawData
- 24: FastQC on data 8: Webpage
- 23: FastQC on data 7: RawData
- 22: FastQC on data 7: Webpage

# De novo transcriptome reconstruction with RNA-Seq

Started tool **Trimmomatic** and successfully added 1 job to the queue.

It produces 4 outputs:

- 26: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 paired)
- 27: Trimmomatic on G1E\_rep1\_reverse\_read\_X28SRR549355\_2X29 (R2 paired)
- 28: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 unpaired)
- 29: Trimmomatic on G1E\_rep1\_reverse\_read\_X28SRR549355\_2X29 (R2 unpaired)

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

PHD Comics

Random



Tool recommendation

You have used trimmomatic tool. For further analysis, you could try using the following/recommended tools. The recommended tools are shown in the decreasing order of their scores predicted using machine learning analysis on workflows. Therefore, tools at the top may be more useful than the ones at the bottom. Please click on one of the following/recommended tools to open its definition.

History

search datasets

De novo transcriptome reconstruction

2.25 GB

29: Trimmomatic on G1E\_rep1\_reverse\_read\_X28SRR549355\_2X29 (R2 unpaired)

28: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 unpaired)

27: Trimmomatic on G1E\_rep1\_reverse\_read\_X28SRR549355\_2X29 (R2 paired)

26: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 paired)

25: FastQC on data 8: RawData

Galaxy India

Workflow Visualize Data Help User Notifications Settings

Using 15%

Tools Trimmomatic Show Sections

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.39+galaxy2)

Tool Parameters

Single-end or paired-end reads? Paired-end (two separate input files)

Input FASTQ file (R1/first of pair) 3: G1E\_rep2\_forward\_read\_%28SRR549356\_1%29 accepted formats

Input FASTQ file (R2/second of pair) 4: G1E\_rep2\_reverse\_read\_%28SRR549356\_2%29 accepted formats

Perform initial ILLUMINACLIP step? no Cut adapter and other Illumina-specific sequences from the read

Trimmomatic Operation

History

search datasets

De novo transcriptome reconstruction

2.25 GB

29: Trimmomatic on G1E\_rep1\_reverse\_read\_X28SRR549355\_2X29 (R2 unpaired)

28: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 unpaired)

27: Trimmomatic on G1E\_rep1\_reverse\_read\_X28SRR549355\_2X29 (R2 paired)

26: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 paired)

25: FastQC on data 8: RawData

# De novo transcriptome reconstruction with RNA-Seq

De novo transcriptome reconstruction

2.91 GB 33

33: Trimmomatic on G1E\_rep2\_reverse\_read\_X28SRR549356\_2X2  
9 (R2 unpaired)

32: Trimmomatic on G1E\_rep2\_forward\_read\_X28SRR549356\_1X2  
9 (R1 unpaired)

31: Trimmomatic on G1E\_rep2\_reverse\_read\_X28SRR549356\_2X2  
9 (R2 paired)

30: Trimmomatic on G1E\_rep2\_forward\_read\_X28SRR549356\_1X2  
9 (R1 paired)

29: Trimmomatic on G1E\_rep1\_re

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for 'India' with the 'Trimmomatic' tool selected. The left sidebar includes links for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, History (selected), Notifications, and Settings. The main area displays the 'Trimmomatic' tool parameters:

- Tool Parameters**: Set to "Paired-end (two separate input files)".
- Input FASTQ file (R1/first of pair)**: Selected file 5: Megakaryocyte\_rep1\_forward\_read\_X28SRR549357\_1X29.
- Input FASTQ file (R2/second of pair)**: Selected file 6: Megakaryocyte\_rep1\_reverse\_read\_X28SRR549357\_2X29.
- Perform initial ILLUMINACLIP step?**: Set to "no".
- Trimmomatic Operation**: Cut adapter and other illumina-specific sequences from the read.

The right side shows the History panel with five completed jobs:

- 33: Trimmomatic on G1E\_rep2\_reverse\_read\_X28SRR549356\_2X29 (R2 unpaired)
- 32: Trimmomatic on G1E\_rep2\_forward\_read\_X28SRR549356\_1X29 (R1 unpaired)
- 31: Trimmomatic on G1E\_rep2\_reverse\_read\_X28SRR549356\_2X29 (R2 paired)
- 30: Trimmomatic on G1E\_rep2\_forward\_read\_X28SRR549356\_1X29 (R1 paired)
- 29: Trimmomatic on G1E\_rep1\_reverse\_read\_X28SRR549357\_2X29 (R2 unpaired)

**Success Message:** Started tool **Trimmomatic** and successfully added 1 job to the queue. It produces 4 outputs:

- 34: Trimmomatic on Megakaryocyte\_rep1\_forward\_read\_X28SRR549357\_1X29 (R1 paired)
- 35: Trimmomatic on Megakaryocyte\_rep1\_reverse\_read\_X28SRR549357\_2X29 (R2 paired)
- 36: Trimmomatic on Megakaryocyte\_rep1\_forward\_read\_X28SRR549357\_1X29 (R1 unpaired)
- 37: Trimmomatic on Megakaryocyte\_rep1\_reverse\_read\_X28SRR549357\_2X29 (R2 unpaired)

**Support and Citation:** We need your support ... If Galaxy helped with the analysis of your data, please do not forget to cite:  
The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update  
Nucleic Acids Research, gkae410  
doi:10.1093/nar/gkae410

The screenshot shows the Galaxy web interface for 'India' with the History panel selected. The left sidebar includes links for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, History (selected), Notifications, and Settings. The main area displays the History panel with five completed jobs:

- 37: Trimmomatic on Megakaryocyte\_rep1\_reverse\_read\_X28SRR549357\_2X29 (R2 unpaired)
- 36: Trimmomatic on Megakaryocyte\_rep1\_forward\_read\_X28SRR549357\_1X29 (R1 unpaired)
- 35: Trimmomatic on Megakaryocyte\_rep1\_reverse\_read\_X28SRR549357\_2X29 (R2 paired)
- 34: Trimmomatic on Megakaryocyte\_rep1\_forward\_read\_X28SRR549357\_1X29 (R1 paired)
- 33: Trimmomatic on G1E\_rep2\_reverse\_read\_X28SRR549356\_2X29 (R2 unpaired)

# De novo transcriptome reconstruction with RNA-Seq

Galaxy India

Workflow Visualize Data Help User

Using 15%

Upload Tools Workflows Invocations Visualization Histories History Notifications Settings

Tools Trimmomatic Show Sections

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.39+galaxy2)

Tool Parameters

Single-end or paired-end reads? Paired-end (two separate input files)

Input FASTQ file (R1/first of pair) 7: Megakaryocyte\_rep2\_forward\_read\_%28SRR549358\_1%29 accepted formats

Input FASTQ file (R2/second of pair) 8: Megakaryocyte\_rep2\_reverse\_read\_%28SRR549358\_2%29 accepted formats

Perform initial ILLUMINACLIP step? no Cut adapter and other Illumina-specific sequences from the read

Trimmomatic Operation

History search datasets De novo transcriptome reconstruction 3.31 GB

37: Trimmomatic on Megakaryocyte\_rep1\_reverse\_read\_X28SRR5 49357\_2X29 (R2 unpaired)

36: Trimmomatic on Megakaryocyte\_rep1\_forward\_read\_X28SRR5 49357\_1X29 (R1 unpaired)

35: Trimmomatic on Megakaryocyte\_rep1\_reverse\_read\_X28SRR5 49357\_2X29 (R2 paired)

34: Trimmomatic on Megakaryocyte\_rep1\_forward\_read\_X28SRR5 49357\_1X29 (R1 paired)

33: Trimmomatic on G1E\_rep2\_re

Galaxy India

Workflow Visualize Data Help User

Using 15%

Upload Tools Workflows Invocations Visualization Histories History Notifications Settings

Tools Trimmomatic Show Sections

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.39+galaxy2)

WORKFLOWS All workflows

Started tool Trimmomatic and successfully added 1 job to the queue.

It produces 4 outputs:

- 38: Trimmomatic on Megakaryocyte\_rep2\_forward\_read\_X28SRR549358\_1X29 (R1 paired)
- 39: Trimmomatic on Megakaryocyte\_rep2\_reverse\_read\_X28SRR549358\_2X29 (R2 paired)
- 40: Trimmomatic on Megakaryocyte\_rep2\_forward\_read\_X28SRR549358\_1X29 (R1 unpaired)
- 41: Trimmomatic on Megakaryocyte\_rep2\_reverse\_read\_X28SRR549358\_2X29 (R2 unpaired)

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Citizen Science Experiment!

**Citizen Science Project**

Help us identify the sex of these marmalade hoverflies! Look at the image, and then select a label below that you think fits best. These pictures are of *Episyrphus balteatus* which you can read more about on [Wikipedia](#)!



History search datasets De novo transcriptome reconstruction 3.33 GB

41: Trimmomatic on Megakaryocyte\_rep2\_reverse\_read\_X28SRR5 49358\_2X29 (R2 unpaired)

40: Trimmomatic on Megakaryocyte\_rep2\_forward\_read\_X28SRR5 49358\_1X29 (R1 unpaired)

39: Trimmomatic on Megakaryocyte\_rep2\_reverse\_read\_X28SRR5 49358\_2X29 (R2 paired)

38: Trimmomatic on Megakaryocyte\_rep2\_forward\_read\_X28SRR5 49358\_1X29 (R1 paired)

37: Trimmomatic on Megakaryoc

# De novo transcriptome reconstruction with RNA-Seq

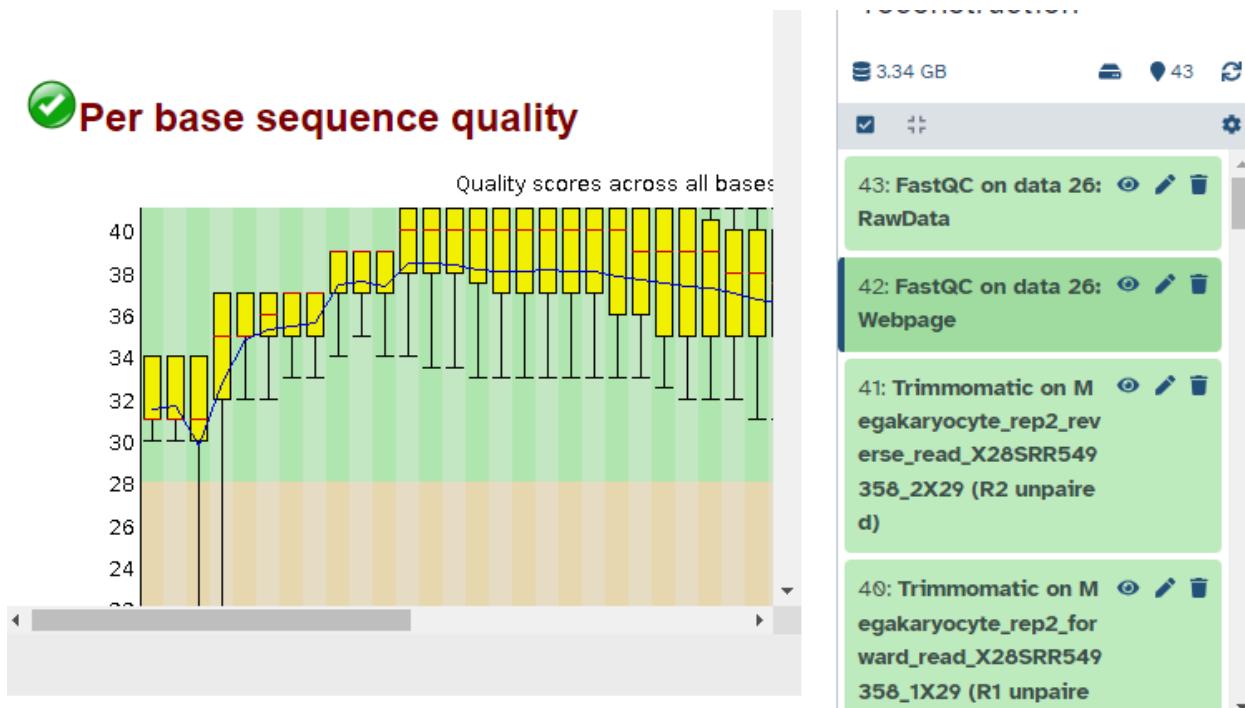
The screenshot shows the Galaxy web interface with the following details:

- Tool Selection:** FastQC Read Quality reports (Galaxy Version 0.74+galaxy1)
- Tool Parameters:**
  - Raw read data from your current history:** 26: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 paired)
  - Contaminant list:** Nothing selected
  - Adapter list:** Nothing selected
  - Submodule and Limit specifying file:** Nothing selected
- History:** Shows several recent jobs:
  - 26: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 unpaired)
  - 28: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 unpaired)
  - 27: Trimmomatic on G1E\_rep1\_reverse\_read\_X28SRR549355\_1X29 (R2 paired)
  - 26: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 paired)
  - 25: FastQC on data 8: RawData
  - 24: FastQC on data 8: Webpage

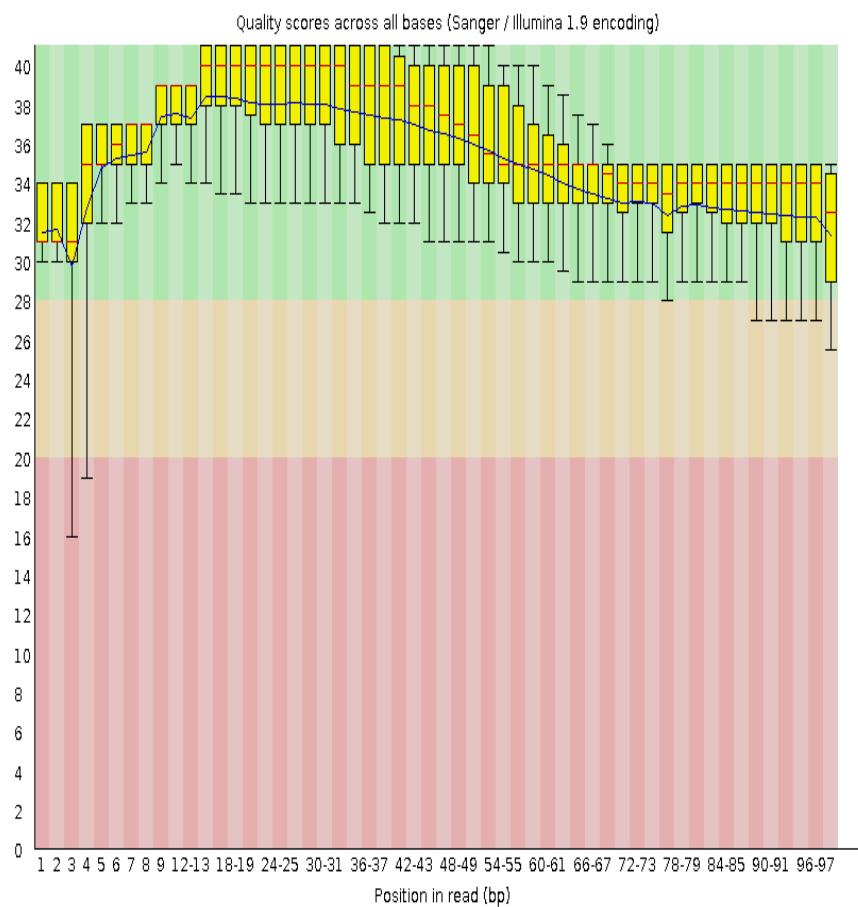
The screenshot shows the FastQC Report with the following details:

- Report Title:** FastQC Report
- Date:** Fri 11 Oct 2024
- Sample ID:** Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29\_R1 paired
- Basic Statistics:**
  - Measure: Value
  - Filename: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29\_R1 paired
  - File type: Conventional base calls
  - Encoding: Sanger / Illumina 1.9
  - Total Sequences: 1236775
  - Total Bases: 105.8 Mbp
  - Sequences flagged as poor quality: 0
  - Sequence length: 1-99
  - %N: 4%
- Produced by:** FastQC (version 0.12.1)
- History:** Shows several recent jobs:
  - 43: FastQC on data 26: RawData
  - 42: FastQC on data 26: Webpage
  - 41: Trimmomatic on Megakaryocyte\_rep2\_revise\_read\_X28SRR549358\_2X29 (R2 unpaired)
  - 40: Trimmomatic on Megakaryocyte\_rep2\_forward\_read\_X28SRR549358\_1X29 (R1 unpaired)

# De novo transcriptome reconstruction with RNA-Seq

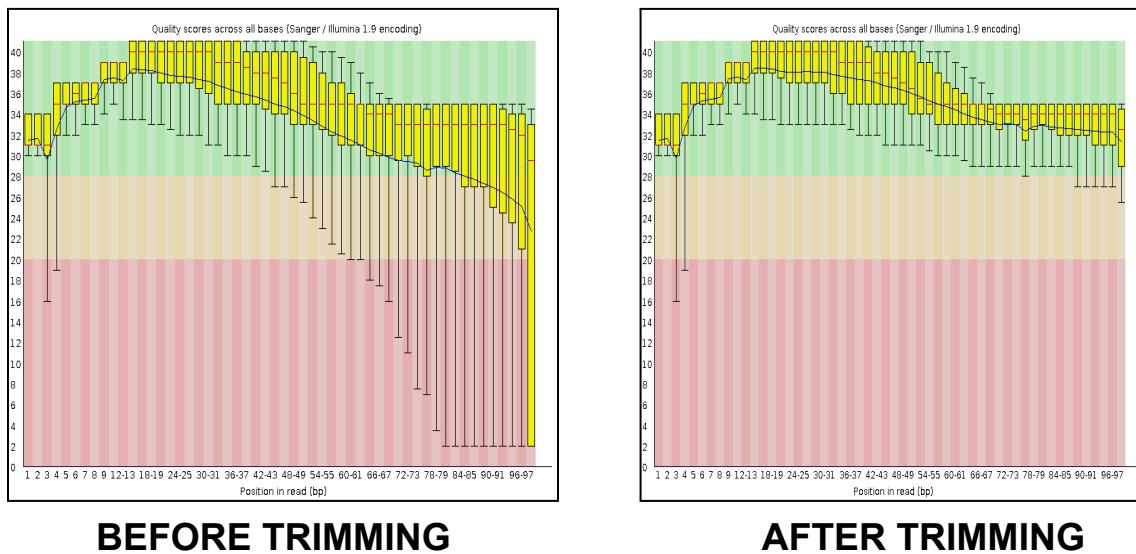


# De novo transcriptome reconstruction with RNA-Seq



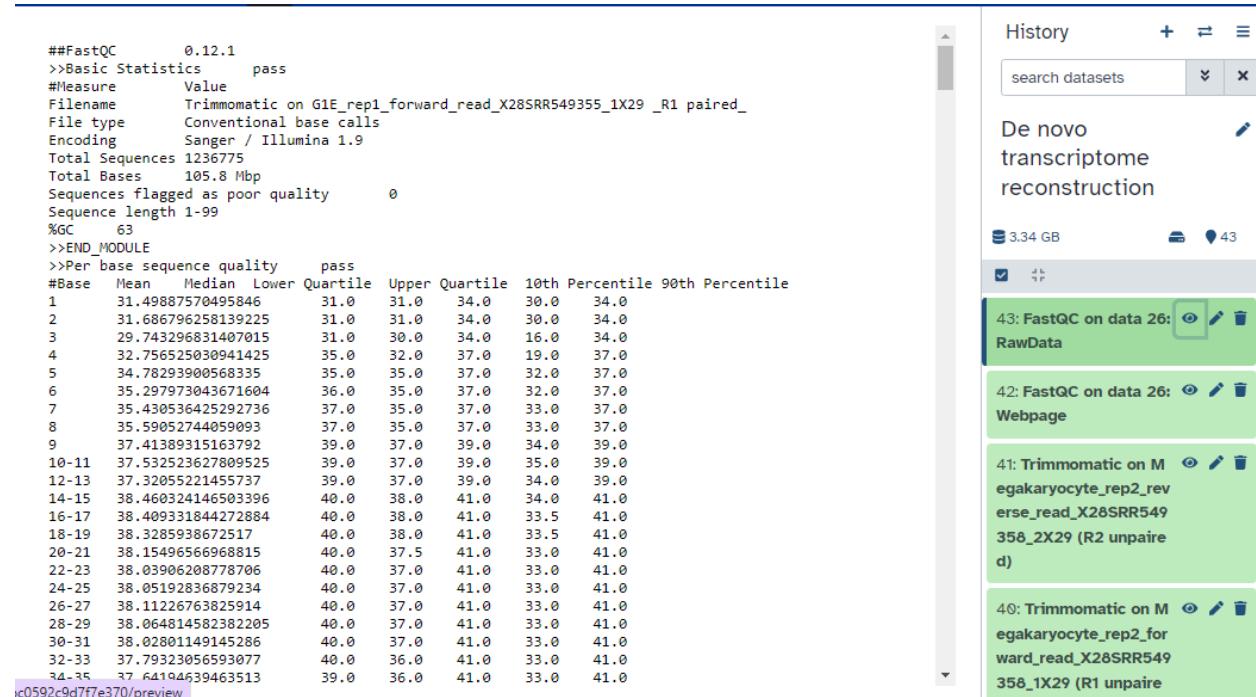
# De novo transcriptome reconstruction with RNA-Seq

## Before vs after filtering



The images illustrate FASTQC quality score plots, comparing sequencing data before and after trimming. Initially, quality scores decline toward the end of the reads, with many bases in the poor-quality range (red zone). After trimming, scores improve significantly, with most bases in the high-quality range (green zone). This enhancement in data quality will positively impact downstream analyses, resulting in more reliable outcomes in studies like variant calling and gene expression.

# De novo transcriptome reconstruction with RNA-Seq



## Mapping:

To make sense of the reads, their positions within the mouse genome must be determined. This process is known as **aligning** or '**mapping**' the reads to the reference genome.

### Spliced mapping:

- HISAT2 tool Run [HISAT2](#) on one forward/reverse read pair and modify the following settings:
- HISAT2 tool: Run [HISAT2](#) on the remaining forward/reverse read pairs with the same parameters.

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for the India instance. On the left, the navigation bar includes 'Upload', 'Tools' (selected), 'Workflows', 'Workflow Invocations', 'Visualization', 'Histories', 'History' (selected), 'Notifications', and 'Settings'. The main panel displays the 'HISAT2 A fast and sensitive alignment program' tool. Under 'Tool Parameters', the 'Source for the reference genome' section shows 'Use a built-in genome' selected. In the 'Select a reference genome' dropdown, 'Mouse (Mus Musculus); mm10 Full' is chosen. The 'Is this a single or paired library?' section has 'Paired-end' selected. The 'FASTA/Q file #1' field contains '26: Trimmomatic on G1E\_rep1\_forward\_read\_X28SRR549355\_1X29 (R1 paired)' with accepted formats 'fastq,sanger,fasta'. The 'FASTA/Q file #2' field is empty. The right panel shows the 'History' tab with several completed workflows: '33: Trimmomatic on G1E\_rep2\_reverse\_read\_X28SRR549356\_2X29 (R2 unpaired)', '32: Trimmomatic on G1E\_rep2\_forward\_read\_X28SRR549356\_1X29 (R1 unpaired)', '31: Trimmomatic on G1E\_rrep2\_reverse\_read\_X28SRR549356\_2X29 (R2 paired)', and '26: Trimmomatic on G1E...'. The status bar at the top right indicates 'Using 16%'. The title bar says 'Galaxy India'.

The screenshot shows the Galaxy web interface after the HISAT2 tool has been run. The 'History' panel on the right shows a green entry: '44: HISAT2 on data 27 and data 26: aligned reads (BAM)'. The main panel displays a success message: 'Started tool HISAT2 and successfully added 1 job to the queue. It produces this output: 44: HISAT2 on data 27 and data 26: aligned reads (BAM)'. Below this, a note says: 'You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from "running" to "finished" if completed successfully or "error" if problems were encountered.' A 'We need your support ...' section encourages citation and acknowledgment. The citation information is: 'The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update Nucleic Acids Research, gkae410 doi:10.1093/nar/gkae410'. The acknowledgment information is: 'The Galaxy server used for some calculations is partly funded by the German Federal Ministry of Education and Research BMBF grant 031A536A de.NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI Freiburg.' The right panel shows the 'History' tab with several completed workflows: '44: HISAT2 on data 27 and data 26: aligned reads (BAM)', '43: FastQC on data 26: Raw Data', '42: FastQC on data 26: Web page', '41: Trimmomatic on Megakaryocyte\_rep2\_reverse\_read\_X28SRR549356\_2X29 (R2 unpaired)', and '40: Trimmomatic on Megakaryocyte\_rep2\_forward\_read\_X28SRR549356\_1X29 (R1 unpaired)'. The status bar at the top right indicates 'Using 15%'. The title bar says 'Galaxy India'.

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for running the HISAT2 tool. In the main panel, the user has selected "HISAT2 A fast and sensitive alignment program (Galaxy Version 2.2.1+galaxy1)". The configuration includes:

- Use a built-in genome:** Built-in references were created using default options.
- Select a reference genome:** Mouse (Mus Musculus); mm10 Full.
- Is this a single or paired library:** Paired-end.
- FASTA/Q file #1:** 30: Trimmomatic on G1E\_rep2\_forward\_read\_X28SRR549356\_1X29 (R1 paired).
- FASTA/Q file #2:** 31: Trimmomatic on G1E\_rep2\_reverse\_read\_X28SRR549356\_2X29 (R2 paired).

The History panel on the right shows the following workflow steps:

- 44: HISAT2 on data 27 and data 26: aligned reads (BA M)
- 43: FastQC on data 26: Raw Data

The screenshot shows the Galaxy India web interface. The left sidebar navigation includes:

- Upload
- Tools
- Workflows
- Workflow Invocations
- Visualization
- History
- Notifications
- Settings

The main panel shows the results of a completed HISAT2 run:

- Started tool HISAT2 and successfully added 1 job to the queue.**
- It produces this output:**
  - 45: HISAT2 on data 31 and data 30: aligned reads (BAM)
- Galaxy News:** This page contains announcements of interest to the Galaxy Community. These include items from the Galaxy Team or the Galaxy community and address anything that is of wide interest to the community. Also see the [Galactic Blog](#) for more.

The History panel on the right shows the following workflow steps:

- 45: HISAT2 on data 31 and data 30: aligned reads (BA M)
- 44: HISAT2 on data 27 and data 26: aligned reads (BA M)
- 43: FastQC on data 26: Raw Data
- 42: FastQC on data 26: Web page
- 41: Trimmomatic on Megakaryocyte\_rep2\_reverse\_read\_X28SRR549356\_2X29 (R2)

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for running the HISAT2 tool. The top navigation bar includes Workflow, Visualize, Data, Help, User, and a Run Tool button. The main panel displays the HISAT2 tool configuration form. The first section asks if it's a single or paired library, with 'Paired-end' selected. The next two sections, 'FASTA/Q file #1' and 'FASTA/Q file #2', show dropdown menus for selecting files. Both dropdowns list '34: Trimmomatic on Megakaryocyte\_rep1\_forward\_read\_X28SRR549357\_1X29 (R1 paired)' and '35: Trimmomatic on Megakaryocyte\_rep1\_reverse\_read\_X28SRR549357\_2X29 (R2 paired)'. Below these dropdowns, notes state that the files must be in fastq-sanger or fasta format. The 'Specify strand information' section has 'Forward (FR)' selected. A note explains that 'FR' means a read corresponds to a transcript and 'RF' means it corresponds to the reverse complemented counterpart. The right side of the interface shows the History panel with several completed jobs listed, including HISAT2 runs and FastQC reports. The total memory usage is shown as 3.71 GB.

This screenshot shows the Galaxy web interface with a different history. The left sidebar is expanded, showing the 'WORKFLOWS' section with 'All workflows'. The main panel shows the HISAT2 tool configuration for a mouse genome (mm10). The 'Is this a single or paired library' section has 'Paired-end' selected. The 'FASTA/Q file #1' and 'FASTA/Q file #2' sections show dropdown menus for selecting files. Both dropdowns list '38: Trimmomatic on Megakaryocyte\_rep2\_forward\_read\_X28SRR549356\_1X29 (R1 paired)' and '39: Trimmomatic on Megakaryocyte\_rep2\_reverse\_read\_X28SRR549356\_2X29 (R2 paired)'. Below these dropdowns, notes state that the files must be in fastq-sanger or fasta format. The 'Specify strand information' section has 'Forward (FR)' selected. A note explains that 'FR' means a read corresponds to a transcript and 'RF' means it corresponds to the reverse complemented counterpart. The right side of the interface shows the History panel with several completed jobs listed, including HISAT2 runs and FastQC reports. The total memory usage is shown as 3.49 GB. A warning message in the history panel states: 'Warning: skipping mate #1 of read SRR549355.109946713/1 because length (1) <= # seed'.

# De novo transcriptome reconstruction with RNA-Seq

Started tool **HISAT2** and successfully added 1 job to the queue.

It produces this output:

- 47: HISAT2 on data 39 and data 38: aligned reads (BAM)

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

We need your support ...

If Galaxy helped with the analysis of your data, please do not forget to [cite](#):

The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update  
Nucleic Acids Research, gkae410  
doi:10.1093/nar/gkae410

And please [acknowledge](#) the European Galaxy server:

The Galaxy server used for some calculations is partly funded by the German Federal Ministry of Education and Research BMBF grant 031A536A de.NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI Freiburg.

History    +  ≡  x

search datasets

De novo transcriptome reconstruction

3.84 GB    47

- 47: HISAT2 on data 39 and data 38: aligned reads (BAM)
- 46: HISAT2 on data 35 and data 34: aligned reads (BAM)
- 45: HISAT2 on data 31 and data 30: aligned reads (BAM)
- 44: HISAT2 on data 27 and data 26: aligned reads (BAM)

# De novo transcriptome reconstruction with RNA-Seq

De novo transcriptome reconstruction

3.84 GB 47

The screenshot shows a list of completed tasks for transcriptome reconstruction. Each task is represented by a green card with white text. The tasks are:

- 47: HISAT2 on data 39 and data 38: aligned reads (BAM)
- 46: HISAT2 on data 35 and data 34: aligned reads (BAM)
- 45: HISAT2 on data 31 and data 30: aligned reads (BAM)
- 44: HISAT2 on data 27 and data 26: aligned reads (BAM)

Each card includes icons for viewing, editing, and deleting the task.

# De novo transcriptome reconstruction with RNA-Seq

## De novo transcript reconstruction:

After aligning reads with HISAT2, apply **StringTie** to reconstruct the transcriptome and identify all transcripts, including novel genes and isoforms. StringTie uses the aligned BAM files to predict transcript structures and generates a GTF file. You can then run **GFFCompare** to match these predicted transcripts against known annotations. Finally, use tools like **Trackster** or **IGV** to visually inspect and analyze both novel and annotated transcripts. complete transcriptome(s) identification from the experimental samples. The leading tool for transcript reconstruction is **Stringtie**. Here, we will use **Stringtie** to predict transcript structures based on the reads aligned by **HISAT**.

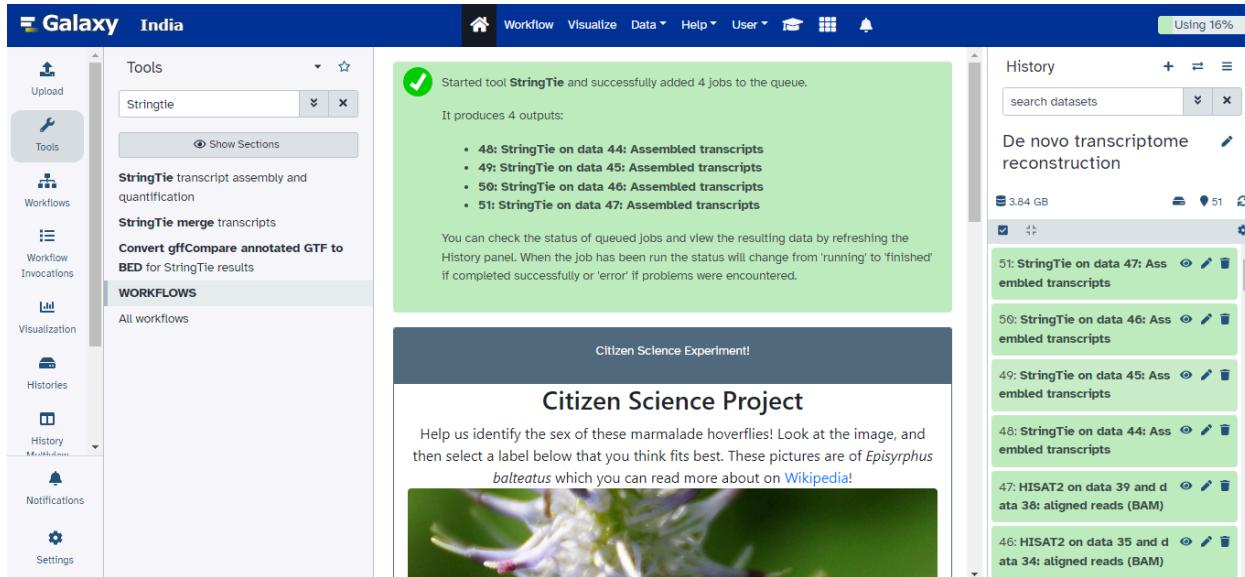
### Transcriptome reconstruction:

- Stringtie tool Run **Stringtie** on the **HISAT2** alignments using the default parameters.

The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy India
- Left Sidebar:** Tools, Workflows, Visualization, Histories, Notifications, Settings.
- Tool Selection:** Tools dropdown selected, Stringtie selected.
- Tool Details:** StringTie transcript assembly and quantification (Galaxy Version 2.2.3+galaxy9)
- Tool Parameters:**
  - Input options:** Short reads (Input short mapped reads: accepted formats: 47: HISAT2 on data 39 and data 38: aligned reads (BAM), 46: HISAT2 on data 35 and data 34: aligned reads (BAM), 45: HISAT2 on data 31 and data 30: aligned reads (BAM), 44: HISAT2 on data 27 and data 26: aligned reads (BAM)).
  - Specify strand information:** Forward (FR)
- History:** De novo transcriptome reconstruction (3.64 GB). History items include:
  - 47: HISAT2 on data 39 and data 38: aligned reads (BAM)
  - 46: HISAT2 on data 35 and data 34: aligned reads (BAM)
  - 45: HISAT2 on data 31 and data 30: aligned reads (BAM)
  - 44: HISAT2 on data 27 and data 26: aligned reads (BAM)
  - 43: FastQC on data 26: RawData
  - 42: FastQC on data 26: Webpage

# De novo transcriptome reconstruction with RNA-Seq



## Transcriptome assembly:

**StringTie - Merge** is used to combine the GTF files from four RNA-seq libraries along with the RefSeq reference, reducing redundancy and creating a unified transcriptome. This process generates a merged GTF file that includes transcripts from all libraries and the reference. Next, GFFCompare is run with the merged GTF and the RefSeq GTF to annotate the transcriptome. The tool identifies the relationships between the merged transcripts and the RefSeq reference. As a result, transcripts are classified as known, novel, or overlapping based on their alignment with the reference.

# De novo transcriptome reconstruction with RNA-Seq

## Transcriptome assembly:

- Stringtie-merge tool Run `Stringtie-merge` on the `Stringtie` assembled transcripts along with the RefSeq annotation file we imported earlier.
- GFFCompare tool Run `GFFCompare` on the `Stringtie-merge` generated transcriptome along with the RefSeq annotation file.

The screenshot shows the Galaxy web interface with the following details:

- Left Sidebar:** Includes links for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, Notifications, and Settings.
- Top Bar:** Shows "Galaxy India" and various navigation links like Workflow, Visualize, Data, Help, User, and a search bar.
- Tool Panel:** Shows the "Stringtie-merge" tool selected under "Tools".
- Tool Parameters:**
  - Transcripts:** A list of datasets selected for merging:
    - 51: StringTie on data 47: Assembled transcripts
    - 50: StringTie on data 46: Assembled transcripts
    - 49: StringTie on data 45: Assembled transcripts
    - 48: StringTie on data 44: Assembled transcripts
  - In GTF or GFF3 format:** A dropdown menu.
  - Reference annotation to include in the merging:** A dropdown menu set to "9: RefSeq\_reference\_GTF\_%26DSv%29".
  - Minimum input transcript length to include in the merge:** Set to "50".
  - Minimum input transcript coverage to include in the merge:** Set to "0".
- History Panel:** Shows a list of recent workflows and datasets, including:
  - 51: StringTie on data 47: Assembled transcripts
  - 50: StringTie on data 46: Assembled transcripts
  - 49: StringTie on data 45: Assembled transcripts
  - 48: StringTie on data 44: Assembled transcripts
  - 47: HISAT2 on data 39 and data 38: aligned reads (BAM)
  - 46: HISAT2 on data 35 and data 34: aligned reads (BAM)

# De novo transcriptome reconstruction with RNA-Seq

History + ⌂ ⌂

search datasets

De novo transcriptome reconstruction

3.84 GB 52

- 49: StringTie on data 45: Assembled transcripts
- 48: StringTie on data 44: Assembled transcripts
- 47: HISAT2 on data 39 and d ata 38: aligned reads (BAM)
- 46: HISAT2 on data 35 and d ata 34: aligned reads (BAM)
- 45: HISAT2 on data 31 and d ata 30: aligned reads (BAM)
- 44: HISAT2 on data 27 and d ata 26: aligned reads (BAM)

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr1	StringTie	transcript	160938184	160940229	1000	-	.	gene_id "STRG.1"; transcript_id "STRG.1.1"; cov "8.996479"; FPKM "42.906161"; TPM "51.893597";
chr1	StringTie	exon	160938184	160938232	1000	-	.	gene_id "STRG.1"; transcript_id "STRG.1.1"; exon_number "1"; cov "6.530612";
chr1	StringTie	exon	160939995	160940229	1000	-	.	gene_id "STRG.1"; transcript_id "STRG.1.1"; exon_number "2"; cov "9.510638";
chr1	StringTie	transcript	160940865	160946631	1000	-	.	gene_id "STRG.2"; transcript_id "STRG.2.1"; cov "10.864801"; FPKM "51.816624"; TPM "62.670475";

History + ⌂ ⌂

search datasets

De novo transcriptome reconstruction

3.84 GB 52

- 52: StringTie merge on data 9, data 51, and others
- 51: StringTie on data 47: Assembled transcripts
- 50: StringTie on data 46: Assembled transcripts
- 49: StringTie on data 45: Assembled transcripts
- 48: StringTie on data 44: Assembled transcripts
- 47: HISAT2 on data 39 and d ata 38: aligned reads (BAM)

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr1	StringTie	transcript	160906411	160974976	1000	+	.	gene_id "MSTRG.1"; transcript_id "NM_001024952"; ref_gene_id "Rc3h1";
chr1	StringTie	exon	160906411	160906699	1000	+	.	gene_id "MSTRG.1"; transcript_id "NM_001024952"; exon_number "1"; ref_gene_id "Rc3h1";
chr1	StringTie	exon	160929964	160930344	1000	+	.	gene_id "MSTRG.1"; transcript_id "NM_001024952"; exon_number "2"; ref_gene_id "Rc3h1";
chr1	StringTie	exon	160938112	160938232	1000	+	.	gene_id "MSTRG.1"; transcript_id "NM_001024952"; exon_number "3"; ref_gene_id "Rc3h1";
chr1	StringTie	exon	160939995	160940234	1000	+	.	gene_id "MSTRG.1"; transcript_id "NM_001024952"; exon_number "4"; ref_gene_id "Rc3h1";

3817194a359eeca/oreview

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface with the 'India' instance selected. The left sidebar includes 'Upload', 'Tools' (selected), 'Workflows', 'Workflow Invocations', 'Visualization', 'Histories', 'History' (selected), 'Notifications', and 'Settings'. The main panel displays the 'GffCompare' tool configuration. Under 'Tool Parameters', the 'GTF inputs for comparison' section lists '52: StringTie merge on data 9, data 51, and others'. The 'Use reference annotation' section has 'Yes' selected. In the 'Choose the source for the reference annotation' dropdown, 'History' is chosen, and the 'Reference annotation' dropdown shows '9: RefSeq\_reference\_GTF\_%28DSv2%29'. The 'Snp correction' section has 'No' selected. The right panel shows the 'History' tab with several completed jobs related to transcriptome reconstruction, such as '52: StringTie merge on data 9, data 51, and others' and '49: StringTie on data 45: Assembled transcripts'.

The screenshot shows the Galaxy web interface with the 'India' instance selected. The left sidebar includes 'Upload', 'Tools' (selected), 'Workflows', 'Workflow Invocations', 'Visualization', 'Histories', 'History' (selected), 'Notifications', and 'Settings'. The main panel displays a success message: 'Started tool GffCompare and successfully added 1 job to the queue.' It also lists 6 outputs: '53: GffCompare on data 9 and data 52: annotated transcripts', '54: GffCompare on data 9 and data 52: RefMap', '55: GffCompare on data 9 and data 52: TMAP', '56: GffCompare on data 9 and data 52: accuracy stats', '57: GffCompare on data 9 and data 52: loci file', and '58: GffCompare on data 9 and data 52: tracking file'. Below this, a note says 'You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.' The right panel shows the 'History' tab with several completed jobs related to transcriptome reconstruction, such as '56: GffCompare on data 9 and data 52: tracking file' and '57: GffCompare on data 9 and data 52: loci file'.

# De novo transcriptome reconstruction with RNA-Seq

## Analysis of the differential gene expression:

FeatureCounts are used to count reads per transcript for the G1E and megakaryocyte samples. The resulting counts are input into **DESeq2** for **normalization** and **differential expression analysis**. DESeq2 calculates transcript abundance and performs significance testing. This identifies differentially expressed transcripts between the two cellular states.

## Count the number of reads per transcript:

Use FeatureCounts to count reads aligning to exons in the transcriptome database generated by GFFCompare for accurate transcript abundance quantification. which counts reads that partially overlap with genomic features, while excluding reads that overlap multiple features. This method ensures precise read counts for each transcript. The resulting data will be used for subsequent differential expression analysis

### Counting the number of reads per transcript:

- FeatureCounts tool Run `FeatureCounts` on the aligned reads (`HISAT2` output) using the `GFFCompare` transcriptome database as the annotation file.

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for the 'India' instance. On the left, the navigation bar includes 'Upload', 'Tools' (selected), 'Workflows', 'Invocations', 'Visualization', 'Histories', 'History' (selected), 'Notifications', and 'Settings'. The main panel displays the 'featureCounts' tool, which measures gene expression in RNA-Seq experiments from SAM or BAM files. The 'Tool Parameters' section shows the 'Alignment file' input field containing four BAM files: '47: HISAT2 on data 39 and data 38: aligned reads (BAM)', '46: HISAT2 on data 35 and data 34: aligned reads (BAM)', '45: HISAT2 on data 31 and data 30: aligned reads (BAM)', and '44: HISAT2 on data 27 and data 26: aligned reads (BAM)'. Below this, a note states: 'This is a batch mode input field. Individual jobs will be triggered for each dataset. The input alignment file(s) where the gene expression has to be counted. The file can have a SAM or BAM format; but ALL files must be in the same format. Unless you are using a Gene annotation file from the History, these files must have the database/genome attribute already specified e.g. hg38, not the default: ?'. The 'Specify strand information' dropdown is set to 'Stranded (Forward)'. The 'Gene annotation file' input field is empty. The right panel shows the 'History' tab with a list of completed jobs, including '58: GffCompare on data 9 and data 52: tracking file', '57: GffCompare on data 9 and data 52: loci file', '56: GffCompare on data 9 and data 52: accuracy stats', '55: GffCompare on data 9 and data 52: TMAP', '54: GffCompare on data 9 and data 52: RefMap', and '53: GffCompare on data 9 and data 52: annotated transcript'. The status bar at the top right indicates 'Using 16%'. A green progress bar at the bottom of the main panel shows 'Using 16%'.

The screenshot shows the Galaxy web interface after the featureCounts tool has been run. The main panel now displays a green success message: 'Started tool featureCounts and successfully added 4 jobs to the queue. It produces 8 outputs:'. Below this, a list of 8 output files is shown: '67: featureCounts on data 53 and data 44: Counts', '68: featureCounts on data 53 and data 44: Summary', '69: featureCounts on data 53 and data 45: Counts', '70: featureCounts on data 53 and data 45: Summary', '71: featureCounts on data 53 and data 46: Counts', '72: featureCounts on data 53 and data 46: Summary', '73: featureCounts on data 53 and data 47: Counts', and '74: featureCounts on data 53 and data 47: Summary'. A note below the outputs says: 'You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.' At the bottom, a 'Citizen Science Experiment!' section features a 'Citizen Science Project' image with the text: 'Help us identify the sex of these marmalade hoverflies! Look at the image, and then select a label below that you think fits best. These pictures are of *Episyphus balteatus* which you can read more about on [Wikipedia](#)!'. The right panel shows the 'History' tab with a list of completed jobs, including '74: featureCounts on data 53 and data 47: Summary', '73: featureCounts on data 53 and data 47: Counts', '72: featureCounts on data 53 and data 46: Summary', '71: featureCounts on data 53 and data 46: Counts', '70: featureCounts on data 53 and data 45: Summary', and '69: featureCounts on data 53 and data 45: Counts'. The status bar at the top right indicates 'Using 16%'. A green progress bar at the bottom of the main panel shows 'Using 16%'.

# De novo transcriptome reconstruction with RNA-Seq

## Perform differential gene expression testing:

Transcript expression is determined from read counts, and at least two biological replicates are needed for reliable results. DESeq2 processes the read counts obtained from FeatureCounts by calculating the geometric mean for each gene across all samples. It then divides each count by this mean and uses the median of these ratios as the normalization size factor. This method reduces variability and ensures accurate transcript expression estimates.

- Computation for each gene of the geometric mean of read counts across all samples
- Division of every gene count by the geometric mean
- Use of the median of these ratios as sample's size factor for normalization
- DESeq2 tool Run `DESeq2` with the following parameters:
  - “1: Factor”
    - “1: Factor level”: G1E
      - param-file
      - “Counts file(s)": featureCount files corresponding to the two G1E replicates
    - “2: Factor level”: Mega
      - param-file
      - “Counts file(s)": featureCount files corresponding to the two Mega replicates
  - Filter tool Run `Filter` to extract genes with a significant change in gene expression (adjusted  $p$ -value less than 0.05) between treated and untreated samples
  - Filter tool Determine how many transcripts are up or down regulated in the G1E state.

# De novo transcriptome reconstruction with RNA-Seq

### Edit Dataset Attributes

Attributes    Datatypes    Permissions

**Name**  
featureCounts on G1E rep1

**Info**  
=====

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

**Database/Build** - optional  
Mouse Dec. 2011 (GRCm38/mm10) (mm10)

**Number of comment lines**  
0

**Save**    Auto-detect

### History

search datasets

De novo transcriptome reconstruction

3.84 GB    66    8

- and data 44: Summary
- 67: featureCounts on data 53
- and data 44: Counts
- 58: GffCompare on data 9 and data 52: tracking file
- 57: GffCompare on data 9 and data 52: loci file
- 56: GffCompare on data 9 and data 52: accuracy stats
- 55: GffCompare on data 9 and data 52: TMAP
- 54: GffCompare on data 9 and data 52: TMAP

# De novo transcriptome reconstruction with RNA-Seq

Edit Dataset Attributes

Name: featureCounts on G1E rep2

Info:

Annotation (optional):

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

Database/Build (optional): Mouse Dec. 2011 (GRCm38/mm10) (mm10)

Number of comment lines: 0

Save Auto-detect

History:

- 3.84 GB 66 8
- 71: featureCounts on data 53 and data 46: Counts
- 70: featureCounts on data 53 and data 45: Summary
- 69: featureCounts on data 53 and data 45: Counts
- 68: featureCounts on data 53 and data 44: Summary
- 67: featureCounts on G1E rep 1
- 58: GffCompare on data 9 and data 52: tracking file
- 57: GffCompare on data 9 and data 52: tracking file

Galaxy India

Tools

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Convert Formats

Filter and Sort

Join, Subtract and Group

GENOMIC FILE MANIPULATION

Convert Formats

FASTA/FASTQ

Quality Control

SAM/BAM

BED

VCF/BCF

Edit Dataset Attributes

Name: featureCounts on Megakaryocytes rep1

Info:

Annotation (optional):

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

Database/Build (optional): Mouse Dec. 2011 (GRCm38/mm10) (mm10)

Number of comment lines: 0

Save Auto-detect

History:

- 3.84 GB 66 8
- and data 47: Counts
- 72: featureCounts on data 53 and data 46: Summary
- 71: featureCounts on data 53 and data 46: Counts
- 70: featureCounts on data 53 and data 45: Summary
- 69: featureCounts on G1E rep 2
- 68: featureCounts on data 53 and data 44: Summary
- 67: featureCounts on G1E rep 1

# De novo transcriptome reconstruction with RNA-Seq

— Edit Dataset Attributes —

Attributes   Datatypes   Permissions

**Name**  
featureCounts on Megakaryocytes rep2

**Info**  
=====

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

**Database/Build** - optional  
Mouse Dec. 2011 (GRCm38/mm10) (mm10)

**Number of comment lines**  
0

**Save**   **Auto-detect**

History

search datasets

De novo transcriptome reconstruction

3.84 GB   66   8

74: featureCounts on data 53 and data 47: Summary  
73: featureCounts on data 53 and data 47: Counts  
72: featureCounts on data 53 and data 46: Summary  
71: featureCounts on Megakaryocytes rep1  
70: featureCounts on data 53 and data 45: Summary  
69: featureCounts on G1E rep 2

Galaxy India

Workflow Visualize Data Help User Home Notifications

Using 16%

Tools

DESeq2

Show Sections

DESeq2 Determines differentially expressed features from count tables

Annotate DESeq2/DEXSeq output tables Append annotation from GTF to differential expression tool outputs

WORKFLOWS

All workflows

DESeq2

Determines differentially expressed features from count tables (Galaxy Version 2.11.40.6+galaxy0)

Run Tool

G1E

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

Counts file(s) \*

accepted formats

69: featureCounts on G1E rep2  
67: featureCounts on G1E rep1

switch to column select

2: Factor level

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

Mega

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

Counts file(s) \*

accepted formats

73: featureCounts on Megakaryocytes rep2  
71: featureCounts on Megakaryocytes rep1

switch to column select

History

search datasets

De novo transcriptome reconstruction

3.84 GB   66   8

74: featureCounts on data 53 and data 47: Summary  
73: featureCounts on Megakaryocytes rep2  
72: featureCounts on data 53 and data 46: Summary  
71: featureCounts on Megakaryocytes rep1  
70: featureCounts on data 53 and data 45: Summary  
69: featureCounts on G1E rep 2

# De novo transcriptome reconstruction with RNA-Seq

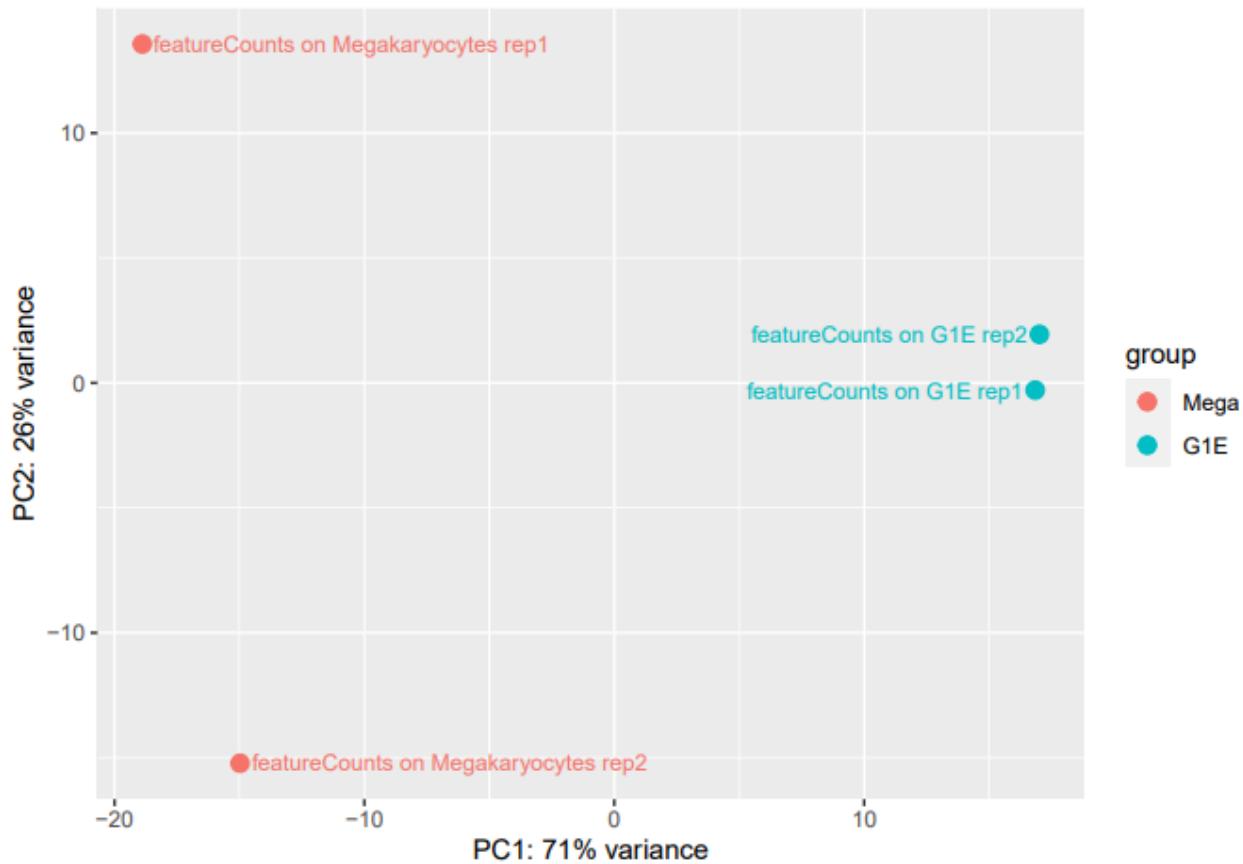
This screenshot shows the Galaxy web interface for a user named 'India'. The left sidebar includes links for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, History (selected), Notifications, and Settings. The main area displays the results of a DESeq2 tool run. A green success message states: 'Started tool DESeq2 and successfully added 1 job to the queue.' It lists three outputs: '75: DESeq2 result file on data 73, data 71, and others', '76: DESeq2 plots on data 73, data 71, and others', and '77: Normalized counts file on data 73, data 71, and others'. Below this, a note says: 'You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from "running" to "finished" if completed successfully or "error" if problems were encountered.' To the right, the 'History' panel shows the following items:

- 77: Normalized counts file on data 73, data 71, and others
- 76: DESeq2 plots on data 73, data 71, and others
- 75: DESeq2 result file on data 73, data 71, and others
- 74: featureCounts on data 53 and data 47: Summary
- 73: featureCounts on Megakaryocytes rep2
- 72: featureCounts on data 53 and data 46: Summary

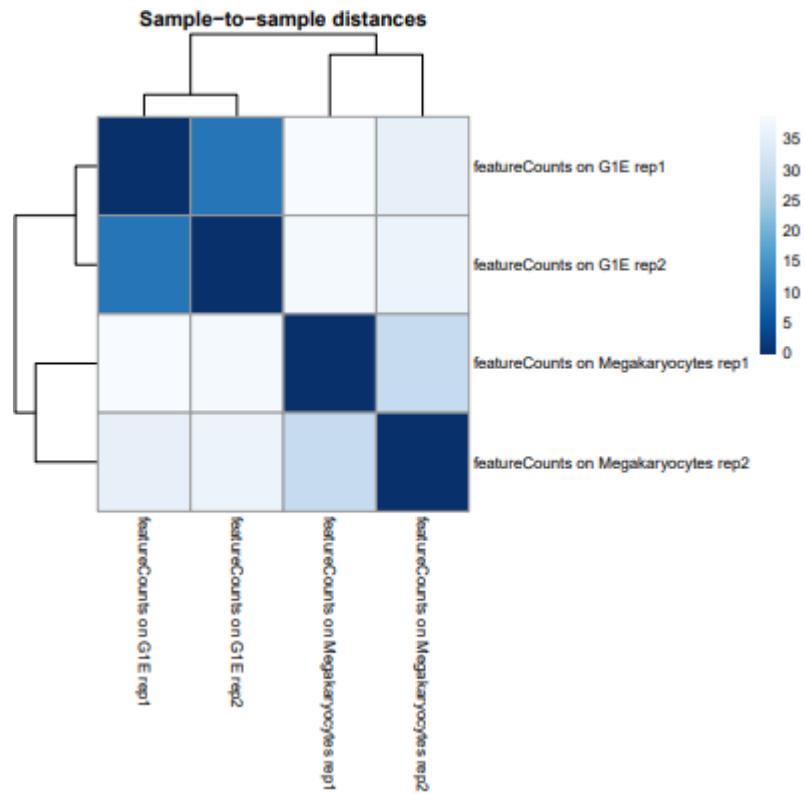
This screenshot shows the Galaxy web interface for a user named 'India'. The left sidebar includes links for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, History (selected), Notifications, and Settings. The main area displays R Graphics output. The top part shows a PCA plot with PC1: 71% variance and PC2: 26% variance. The bottom part shows a heatmap titled 'Sample-to-sample distances' with samples grouped into four categories: 'NatureCounts on G1E rep1', 'NatureCounts on G1E rep2', 'NatureCounts on Megakaryocytes rep1', and 'NatureCounts on Megakaryocytes rep2'. To the right, the 'History' panel shows the following items:

- 77: Normalized counts file on data 73, data 71, and others
- 76: DESeq2 plots on data 73, data 71, and others
- 75: DESeq2 result file on data 73, data 71, and others
- 74: featureCounts on data 53 and data 47: Summary
- 73: featureCounts on Megakaryocytes rep2
- 72: featureCounts on data 53 and data 46: Summary

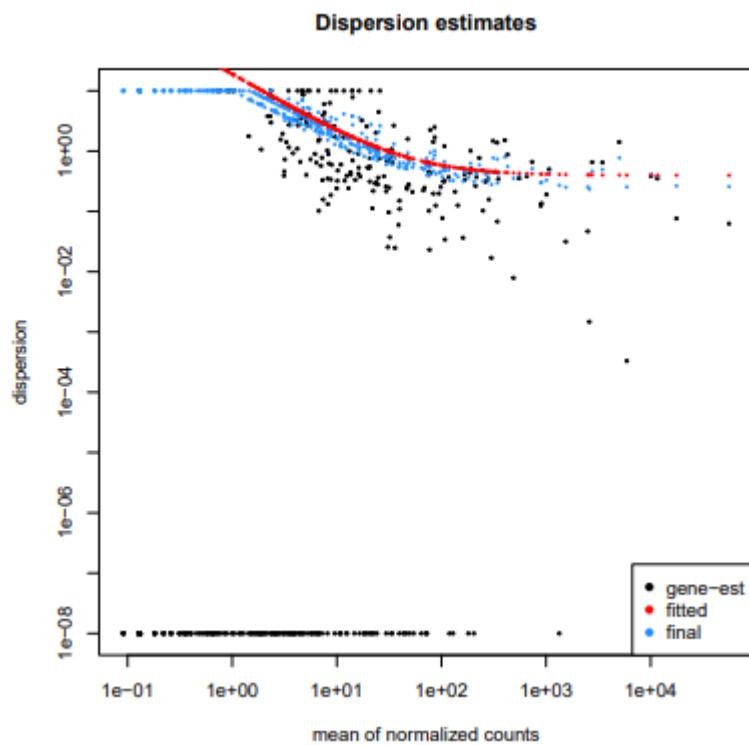
# De novo transcriptome reconstruction with RNA-Seq



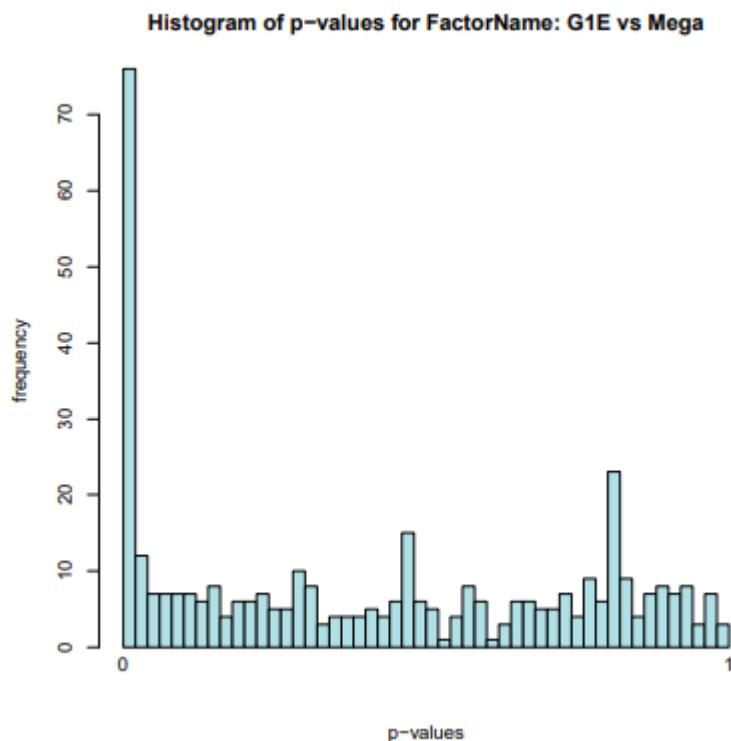
# De novo transcriptome reconstruction with RNA-Seq



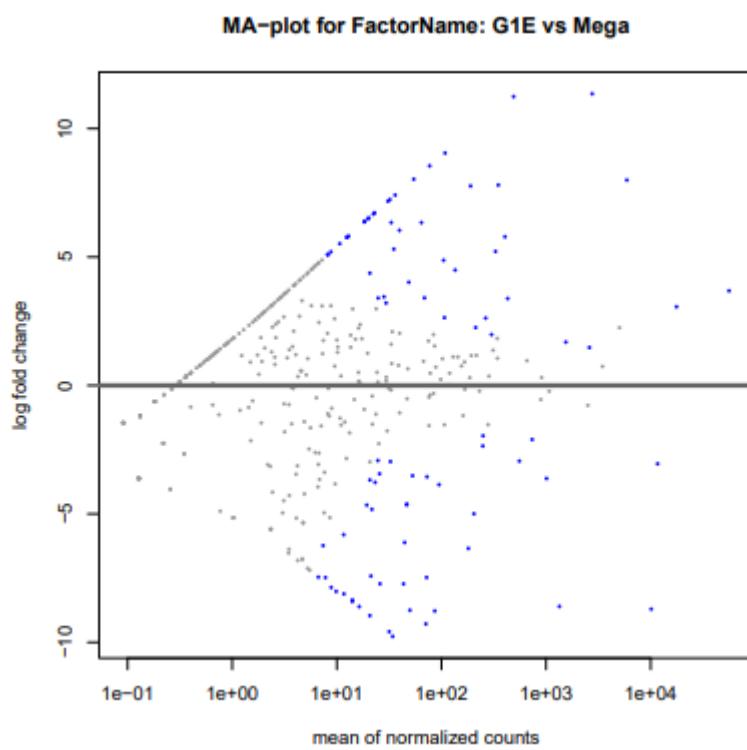
# De novo transcriptome reconstruction with RNA-Seq



# De novo transcriptome reconstruction with RNA-Seq



# De novo transcriptome reconstruction with RNA-Seq



Column 1	Column 2	Column 3	Column 4	Column 5
	featureCounts on Megakaryocytes rep1	featureCounts on Megakaryocytes rep2	featureCounts on G1E rep1	featureCounts on G1E rep2
NM_001024952	0	9.35406710564498	4.73861300580976	2.91244742114105
NM_080844	0	0	0	0
MSTRG.6.1	13.8294949457185	0	3.68558789340759	8.00923040813788
MSTRG.7.1	20.4881406603237	0	4.21210044960868	3.64055927642631
MSTRG.8.1	2.04881406603237	0	5.26512556201085	2.91244742114105
MSTRG.9.1	0	0	9.47722601161953	2.18433556585578
MSTRG.11.1	0	0	2.10605022480434	18.5666523097742
MSTRG.12.1	2.56101758254046	0	9.47722601161953	14.9262930333479
NR_002840	7.68305274762139	0	4.73861300580976	8.00923040813788
NR_028543	0	0	0	0
MSTRG.16.2	0	0	0	0
MSTRG.16.1	0	0	3.15907533720651	4.00461520406894
MSTRG.16.3	0	0	0	0.364055927642631
NR_131029	0	0	0	0
NR_131030	0	0	0	0

History + = ⌂

search datasets

De novo transcriptome reconstruction

3.84 GB

77: Normalized counts file on data 73, data 71, and others

76: DESeq2 plots on data 73, data 71, and others

75: DESeq2 result file on data 73, data 71, and others

74: featureCounts on data 53 and data 47: Summary

73: featureCounts on Megakaryocytes rep2

72: featureCounts on data 53 and data 46: Summary

# De novo transcriptome reconstruction with RNA-Seq

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value
MSTRG.157.4	1342.25293021019	-8.60326304067584	0.795276245706494	-10.8179555055527	2.830203247
MSTRG.32.1	5919.00828748956	7.99489233584141	0.782911213995856	10.2117483986936	1.756704243
MSTRG.161.1	10126.4400654399	-8.70645283569351	0.932271203795258	-9.33897003388038	9.727575551
MSTRG.52.1	2761.65650069109	11.352425710147	1.27431270331175	8.90866557371969	5.165045496
MSTRG.87.1	489.705907788662	11.2450630267695	1.53857913530952	7.30873230287718	2.696744081
MSTRG.306.1	180.767740170624	-6.34383274393346	0.932041908456529	-6.80638143668768	1.000841390
MSTRG.80.1	349.659601824575	7.80173083071458	1.19203314001101	6.54489423896596	5.953750433
MSTRG.162.1	204.511784877632	-4.99657797405334	0.834792516998134	-5.98541298863189	2.1584148542
MSTRG.213.1	108.33466944873	9.046857817317	1.61521111199896	5.60103738149792	2.1307277692
MSTRG.94.1	189.887161635607	7.76495620004759	1.44055489710335	5.39025358607387	7.035832972
MSTRG.112.1	71.0091298952804	-9.28319044590951	1.76247397630784	-5.26713617942696	1.385684166
MSTRG.212.1	77.1320305287957	8.54646289719812	1.65018504927631	5.1791057620757	2.2295204876
MSTRG.345.1	86.3330215935554	-8.77928643686929	1.721279231318	-5.10044290149652	3.388595974
NM_019932	72.0859143731627	-7.47426361215963	1.47081881048825	-5.08171608825053	3.740401851
MSTRG.216.1	56129.4800922312	3.68464107873795	0.733072786920486	5.02629635757794	5.0004294431
MSTRG.308.1	44.4670196780545	-6.11496849972545	1.22335223720917	-4.99853461148322	5.7767638725

History + ⌂ ⌂

search dataset

De novo transcriptome reconstruction

3.84 GB 69 8

77: Normalized counts file on data 73, data 71, and others

76: DESeq2 plots on data 73, data 71, and others

75: DESeq2 result file on data 73, data 71, and others

Galaxy India

Workflow Visualize Data Help User Home Notifications

Using 16%

Upload Tools Workflows Workflow Invocations Visualization Histories History Notifications Settings

Tools

Filter

Filter data on any column using simple expressions (Galaxy Version 1.1.1)

Tool Parameters

Filter \*  accepted formats

With following condition \*

Number of header lines to skip \*

Additional Options

Email notification  Send an email notification when the job completes.

Help

History + ⌂ ⌂

search datasets

De novo transcriptome reconstruction

3.84 GB 69 8

77: Normalized counts file on data 73, data 71, and others

76: DESeq2 plots on data 73, data 71, and others

75: DESeq2 result file on data 73, data 71, and others

74: featureCounts on data 53 and data 47: Summary

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows a bioinformatics pipeline interface with a table of gene expression data and a history panel.

**Table Data:**

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value
MSTRG.157.4	1342.25293021019	-8.60326304067584	0.795276245706494	-10.8179555055527	2.83020324
MSTRG.32.1	5919.00828748956	7.99489233564141	0.782911213995856	10.2117483986936	1.75670424
MSTRG.161.1	10126.4400654399	-8.70645283569351	0.932271203795258	-9.33897003388038	9.72757555
MSTRG.52.1	2761.65650069109	11.352425710147	1.27431270331175	8.90866557371969	5.16504549
MSTRG.87.1	489.705907788662	11.2450630267695	1.53857913530952	7.30873230287718	2.69674404
MSTRG.306.1	180.767740170624	-6.34383274393348	0.932041908456529	-6.80638143668766	1.00084139
MSTRG.80.1	349.659601824575	7.80173083071458	1.19203314001101	6.54489423896596	5.95375045
MSTRG.162.1	204.511784877632	-4.99657797405334	0.834792516998134	-5.98541298863189	2.15841485
MSTRG.213.1	108.33466944873	9.046857817317	1.61521111199896	5.60103738149792	2.13072776
MSTRG.94.1	189.687161635607	7.76495620004759	1.44055469710335	5.390253586607387	7.03583297
MSTRG.112.1	71.0091298952804	-9.28319044590951	1.76247397630784	-5.26713617942696	1.38568410
MSTRG.212.1	77.1320305287957	8.54648289719812	1.65018504927631	5.1791057620757	2.22952048
MSTRG.345.1	86.3330215935554	-8.77928643686929	1.721279231318	-5.10044290149652	3.38859597
NM_019932	72.0859143731627	-7.47428361215963	1.47081881048825	-5.08171608825053	3.74040185
MSTRG.216.1	56129.4800922312	3.68464107873795	0.733072786920486	5.02629635757794	5.00042944
MSTRG.308.1	44.4670196780545	-6.11496849972545	1.22335223720917	-4.99853461148322	5.77676387

**History Panel:**

- 78: DE transcripts (p< 0.05)
- 77: Normalized count file on data 73, data 71, and others
- 76: DESeq2 plots on data 73, data 71, and others
- 75: DESeq2 result file on data 73, data 71, and others
- 74: featureCounts on

The screenshot shows the Galaxy platform interface with a filtering tool and a history panel.

**Tool Parameters:**

**Filter data on any column using simple expressions (Galaxy Version 1.1.1)**

**Tool Parameters:**

**Filter** (selected) **Show Sections**

**Filter data on any column using simple expressions**

**accepted formats**: Dataset missing? See TIP below.

**With following condition \***: c3>0

**Number of header lines to skip \***: 0

**Additional Options**

**Email notification**: No

**Run Tool**

**Help**

**History Panel:**

- 78: DE transcripts (p< 0.05)
- Add Tags
- 72 lines 7 columns  
format tabular, database mm10
- Filtering with c7<0.05, kept 9.89% of 728 valid lines (728 total lines).

1. GeneID	2. Base mean	3. log2
MSTRG.157.4	1342.25293021019	-8.603
MSTRG.32.1	5919.00828748956	7.9948
MSTRG.161.1	10126.4400654399	-8.706

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy India, Workflow, Visualize, Data, Help, User, Notifications.
- Left Sidebar:** Upload, Tools (selected), Workflows, Invocations, Visualization, History, Notifications, Settings.
- Tool Panel:** Filter data on any column using simple expressions (Galaxy Version 1.1.1).
  - Tool Parameters:** Filter condition: 78: DE transcripts ( $p < 0.05$ ).
  - With following condition:** c3<0.
  - Number of header lines to skip:** 0.
  - Additional Options:** Email notification (No).
- Run Tool:** Run Tool button.
- History Panel:** De novo transcriptome reconstruction (3.84 GB, 71 datasets, 8 filters).
- Output:** Dataset 78: Filter on data 78 (37 lines, 7 columns, tabular format, mm10 database). The table includes columns: GeneID, Base mean, log2.

## Edit Dataset Attributes

Attributes Datatypes Permissions

**Name**  
DE transcripts, up in G1E

**Info**  
Filtering with c3>0, kept 51.39% of 72 valid lines (72 total lines).

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

**Database/Build** - optional  
Mouse Dec. 2011 (GRCm38/mm10) (mm10)

**Number of comment lines**  
0

**Save** Auto-detect

The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy India, Workflow, Visualize, Data, Help, User, Notifications.
- Left Sidebar:** Upload, Tools (selected), Workflows, Invocations, Visualization, History, Notifications, Settings.
- Tool Panel:** Filter data on any column using simple expressions (Galaxy Version 1.1.1).
  - Tool Parameters:** Filter condition: 78: DE transcripts ( $p < 0.05$ ).
  - With following condition:** c3<0.
  - Number of header lines to skip:** 0.
  - Additional Options:** Email notification (No).
- Run Tool:** Run Tool button.
- History Panel:** De novo transcriptome reconstruction (3.84 GB, 72 datasets, 8 filters).
- Output:** Dataset 79: Filter on data 78 (37 lines, 7 columns, tabular format, mm10 database). The table includes columns: GeneID, Base mean, log2.

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for creating a new dataset. At the top, a green banner indicates "Attributes updated." Below this, the "Attributes" tab is selected in the navigation bar. The main form fields include:

- Name:** DE transcripts, down in G1E
- Info:** Filtering with  $c3 < 0$ .
- Annotation:** (optional) A large text area for notes.
- Database/Build:** Mouse Dec. 2011 (GRCm38/mm10) (mm10)
- Number of comment lines:** 0

At the bottom right of the form are two buttons: "Save" and "Auto-detect". To the right of the form is a vertical "History" panel. The history panel shows the following entries:

- De novo transcriptome reconstruction (dataset ID 80)
- 79: DE transcripts, up in G1E
- 78: DE transcripts ( $p < 0.05$ )

Details for entry 80 are expanded, showing:

- Size: 3.84 GB
- Location: 72 lines
- Format: tabular, database mm10
- Filtering: Filtering with  $c7 < 0.05$ , kept 9.89% of 728 valid lines

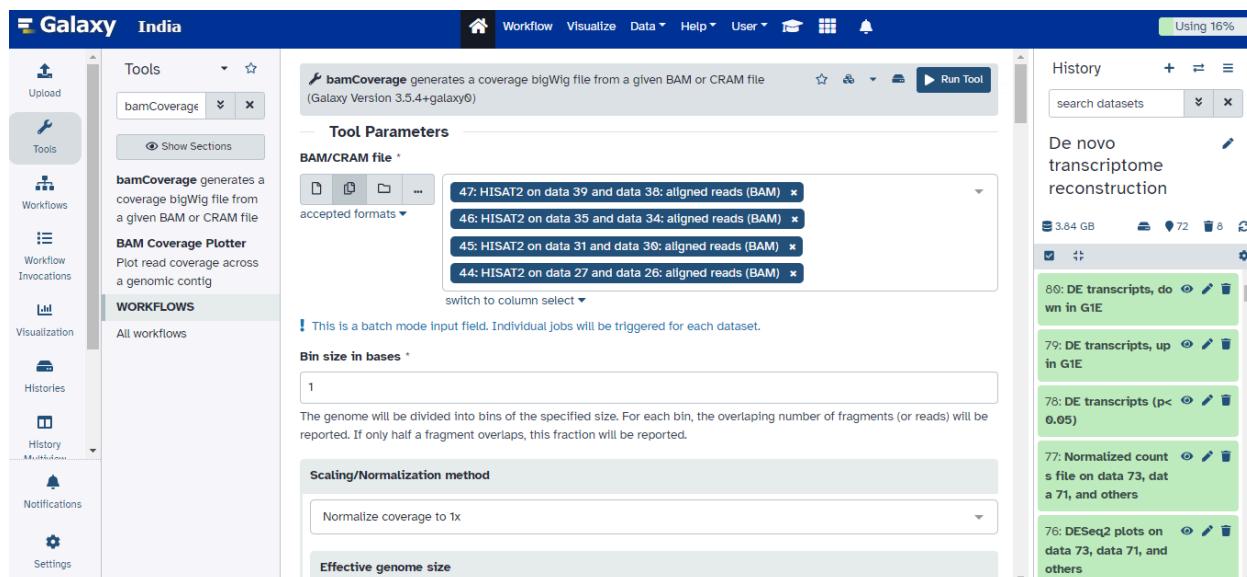
## Visualization:

Transform your aligned read data from **BAM format** to **bigWig format** for simplified visualization. Launch Trackster in Galaxy to initiate a new session, and upload your bigWig files along with the appropriate transcriptome annotations. You can navigate to specific regions of interest to visually inspect the read alignments and transcript structures. This visual approach helps effectively evaluate expression levels and the characteristics of transcripts.

# De novo transcriptome reconstruction with RNA-Seq

## Converting aligned read files to bigWig format:

1. bamCoverage tool Run `bamCoverage` on all four aligned read files (`HISAT2` output) with the following parameters:
  - “*Bin size in bases*”: ‘1’
  - “*Effective genome size*”: ‘mm9 (2150570000)’
  - “*Advanced options*”
    - “*Only include reads originating from fragments from the forward or reverse strand*”: ‘forward’
2. Rename Tool Rename the outputs to reflect the origin of the reads and that they represent the reads mapping to the PLUS strand.
3. bamCoverage tool: Repeat Step 1 except changing the following parameter:
  - “*Only include reads originating from fragments from the forward or reverse strand*”: ‘reverse’
4. Rename tool: Rename the outputs to reflect the origin of the reads and that they represent the reads mapping to the MINUS strand.



# De novo transcriptome reconstruction with RNA-Seq

Galaxy India

Workflow Visualize Data Help User

Using 15%

Tools

bamCoverage

Show Sections

bamCoverage generates a coverage bigWig file from a given BAM or CRAM file

BAM Coverage Plotter Plot read coverage across a genomic contig

WORKFLOWS

All workflows

Upcoming Events

European Galaxy Event Horizon

Upcoming (and past) events with content related to the European Galaxy community.

For events prior to this year, see the events archive.

Advertise your event!

History

search datasets

De novo transcriptome reconstruction

3.84 GB

64: bamCoverage on data 47  
63: bamCoverage on data 46  
62: bamCoverage on data 45  
61: bamCoverage on data 44  
60: DE transcripts, down in G1E  
79: DE transcripts, up in G1E  
78: DE transcripts (p<0.05)  
77: Normalized counts file on data 73, data 71, and others  
76: DESeq2 plots on data 73.dat

This screenshot shows the Galaxy web interface for the 'India' instance. On the left, the navigation bar includes 'Workflow', 'Visualize', 'Data', 'Help', 'User', and various system icons. The main content area displays a success message for running 'bamCoverage' on four datasets (44, 45, 46, 47). It also features an 'Upcoming Events' section titled 'European Galaxy Event Horizon' with a placeholder for advertising. The right side shows a history panel with a list of completed and pending jobs, including 'bamCoverage' runs and differential expression analysis (DE transcripts) and DESeq2 plots. The total disk usage is listed as 3.84 GB.

Galaxy India

Workflow Visualize Data Help User

Using 16%

Tools

search tools

Get Data

Adds environment variables From Copernicus and etopo given geolocalized and timestamped observations

AquaINRA Importer downloads content via the AquaINRA interaction platform

Argo data access for global ocean in situ observing system

BioMart Ensembl server

BioMart Test server

CBI Rice Mart rice mart

DoRiNA Search DoRiNA data source

Download and Extract Reads in BAM format from NCBI SRA

Download and Extract Reads in FASTQ format from NCBI SRA

Download and Generate Pileup Format from NCBI SRA

Edit Dataset Attributes

Name

bamCoverage on data 44 PLUS strand

Info

Due to filtering, 95.22597775293782% of the aforementioned alignments will be used

Annotation - optional

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

Database/Build - optional

Mouse Dec. 2011 (GRCm38/mm10) (mm10)

Save Auto-detect

History

search datasets

De novo transcriptome reconstruction

3.84 GB

64: bamCoverage on data 47  
63: bamCoverage on data 46  
62: bamCoverage on data 45  
61: bamCoverage on data 44  
60: DE transcripts, down in G1E  
79: DE transcripts, up in G1E  
78: DE transcripts (p<0.05)  
77: Normalized counts file on data 73, data 71, and others

This screenshot shows the 'Edit Dataset Attributes' page for a 'bamCoverage' dataset named 'bamCoverage on data 44 PLUS strand'. The page includes fields for 'Name', 'Info' (describing filtering), 'Annotation' (empty), and 'Database/Build' (set to 'Mouse Dec. 2011 (GRCm38/mm10) (mm10)'). At the bottom are 'Save' and 'Auto-detect' buttons. The right side of the interface is identical to the one in the first screenshot, showing the history of completed and pending jobs.

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for 'India'. The left sidebar includes 'Upload', 'Tools' (selected), 'Workflows', 'Workflow Invocations', 'Visualization', 'Histories', 'History' (selected), 'Notifications', and 'Settings'. The main area displays the 'Edit Dataset Attributes' form for dataset 'bamCoverage on data 45 PLUS strand'. The 'Name' field contains 'bamCoverage on data 45 PLUS strand'. The 'Info' section notes 'Due to filtering, 94.94205344699576% of the aforementioned alignments will be used'. The 'Annotation' and 'Database/Build' fields are empty. A green status bar at the bottom indicates 'Attributes updated.' The right panel shows a history of datasets, including 'bamCoverage on data 47', 'bamCoverage on data 46', 'bamCoverage on data 45', 'bamCoverage on data 44', 'DE transcripts, down in G1E', 'DE transcripts, up in G1E', 'DE transcripts (p<0.05)', and 'Normalized counts file on'.

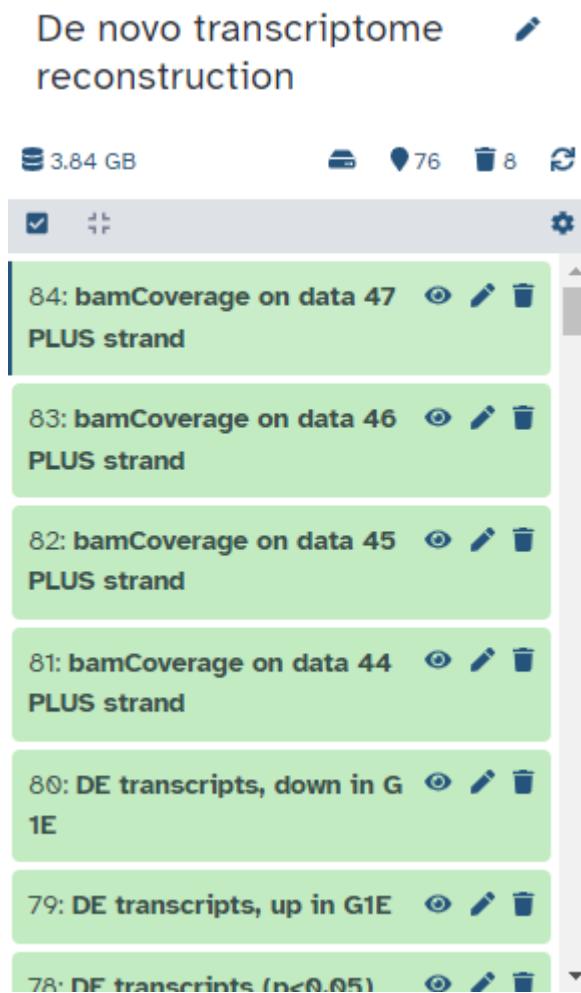
This screenshot shows the same Galaxy interface after saving changes. The 'Edit Dataset Attributes' form now displays a green message 'Attributes updated.' in the top bar. The dataset name remains 'bamCoverage on data 46 PLUS strand'. The 'Info' section now states 'Due to filtering, 65.08728059258573% of the aforementioned alignments will be used'. The right panel's history is identical to the previous screenshot.

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy India, Workflow, Visualize, Data, Help, User, Notifications.
- Left Sidebar:** Upload, Tools (selected), Workflows, Invocations, Visualization, Histories, Notifications, Settings.
- Middle Panel - Edit Dataset Attributes:**
  - Name:** bamCoverage on data 47 PLUS strand
  - Info:** Due to filtering, 47.12252494604148% of the aforementioned alignments will be used.
  - Annotation:** Add an annotation or notes to a dataset; annotations are available when a history is viewed.
  - Database/Build:** optional, set to Mouse Dec. 2011 (GRCm38/mm10) (mm10).
- Bottom Buttons:** Save, Auto-detect.
- Right Panel - History:** De novo transcriptome reconstruction, 3.64 GB, 76 datasets. A list of datasets is shown:
  - 84: bamCoverage on data 47 PLUS strand
  - 83: bamCoverage on data 46 PLUS strand
  - 82: bamCoverage on data 45 PLUS strand
  - 81: bamCoverage on data 44 PLUS strand
  - 80: DE transcripts, down in G1E
  - 79: DE transcripts, up in G1E
  - 78: DE transcripts (p<0.05)

# De novo transcriptome reconstruction with RNA-Seq



# De novo transcriptome reconstruction with RNA-Seq

**Galaxy India**

Workflow Visualize Data Help User Workflow Invocations Visualization Histories Notifications Settings

Using 16%

**Tools**

search tools

**Get Data**

- Adds environment variables From Copernicus and etopo given geolocalized and timestamped observations
- AquaINFRA Importer downloads content via the AquaINFRA interaction platform
- Argo data access for global ocean in situ observing system
- BioMart Ensembl server
- BioMart Test server
- CBI Rice Mart rice mart
- DoRINA Search DoRINA data source
- Download and Extract Reads in BAM format from NCBI SRA
- Download and Extract Reads in FASTQ format from NCBI SRA
- Download and Generate Pileup Format from NCBI SRA

**bamCoverage**  
generates a coverage bigWig file from a given BAM or CRAM file (Galaxy Version 3.5.4+galaxy0)

**Tool Parameters**

**BAM/CRAM file \***

accepted formats

47: HISAT2 on data 39 and data 38; aligned reads (BAM) \*  
 46: HISAT2 on data 35 and data 34; aligned reads (BAM) \*  
 45: HISAT2 on data 31 and data 30; aligned reads (BAM) \*  
 44: HISAT2 on data 27 and data 26; aligned reads (BAM) \*

switch to column select ▾

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

**Bin size in bases \***

1

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**

History

search datasets

De novo transcriptome reconstruction

3.84 GB 76 8

84: bamCoverage on data 47 PLUS strand  
 83: bamCoverage on data 46 PLUS strand  
 82: bamCoverage on data 45 PLUS strand  
 81: bamCoverage on data 44 PLUS strand  
 80: DE transcripts, down in G 1E  
 79: DE transcripts, up in G 1E  
 78: DE transcripts (p<0.05)

**Galaxy India**

Workflow Visualize Data Help User Workflow Invocations Visualization Histories Notifications Settings

Using 16%

**Tools**

search tools

**Get Data**

- Adds environment variables From Copernicus and etopo given geolocalized and timestamped observations
- AquaINFRA Importer downloads content via the AquaINFRA interaction platform
- Argo data access for global ocean in situ observing system
- BioMart Ensembl server
- BioMart Test server
- CBI Rice Mart rice mart
- DoRINA Search DoRINA data source
- Download and Extract Reads in BAM format from NCBI SRA
- Download and Extract Reads in FASTQ format from NCBI SRA
- Download and Generate Pileup Format from NCBI SRA

**bamCoverage**  
generates a coverage bigWig file from a given BAM or CRAM file (Galaxy Version 3.5.4+galaxy0)

Uses this offset inside of each read as the signal. This is useful in cases like RiboSeq or GROseq, where only the 12th, 15th or 1st base aligned should be used to denote where the signal is (rather than the span of the whole alignment). This can be paired with the --filterRNAstrand option. Note that negative values indicate offsets from the end of each read. A value of 1 indicates the first base of the alignment (taking alignment orientation into account). Likewise, a value of -1 is the last base of the alignment. An offset of 0 is not permitted. If two values (separated by spaces) are specified, then they will be used to specify a range of positions. Note that specifying something like --Offset 5 -1 will result in the 5th through last position being used, which is equivalent to trimming 4 bases from the 5-prime end of alignments. (-Offset)

Only include reads originating from fragments from the forward or reverse strand.\*

reverse

By default (the no option), all reads are processed, regardless of the strand they originated from. For RNASEq, it can be useful to separately create bigWig files for the forward or reverse strands. Note that this tool assumes that a dUTP-based method was used, so fragments will be assigned to the reverse strand if the second read in a pair is reverse complemented. (filterRNAstrand)

**Blacklisted regions in BED/GTF format - optional**

accepted formats

Select Value

switch to column select ▾

One or more files containing regions to exclude from the analysis (--blackListFileName)

History

search datasets

De novo transcriptome reconstruction

3.84 GB 76 8

84: bamCoverage on data 47 PLUS strand  
 83: bamCoverage on data 46 PLUS strand  
 82: bamCoverage on data 45 PLUS strand  
 81: bamCoverage on data 44 PLUS strand  
 80: DE transcripts, down in G 1E  
 79: DE transcripts, up in G 1E  
 78: DE transcripts (p<0.05)

# De novo transcriptome reconstruction with RNA-Seq

Galaxy India

Workflow Visualize Data Help User Home History Notifications Settings

Using 16%

Tools

Get Data

- Adds environment variables From Copernicus and etopo given geolocalized and timestamped observations
- AquaINFRA Importer downloads content via the AquaINFRA interaction platform
- Argo data access for global ocean in situ observing system
- BioMart Ensembl server
- BioMart Test server
- CBI Rice Mart rice mart
- DoRINA Search DoRINA data source
- Download and Extract Reads in BAM format from NCBI SRA
- Download and Extract Reads in FASTQ format from NCBI SRA
- Download and Generate Pileup Format from NCBI SRA

Started tool **bamCoverage** and successfully added 4 jobs to the queue.

It produces 4 outputs:

- 85: bamCoverage on data 44
- 86: bamCoverage on data 45
- 87: bamCoverage on data 46
- 88: bamCoverage on data 47

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

PHD Comics Random

GRAD SCHOOL STEREOGRAM

History

De novo transcriptome reconstruction

3.84 GB

88: bamCoverage on data 47  
87: bamCoverage on data 46  
86: bamCoverage on data 45  
85: bamCoverage on data 44  
84: bamCoverage on data 47 PLUS strand  
83: bamCoverage on data 46 PLUS strand  
82: bamCoverage on data 45 PLUS strand  
81: bamCoverage on data 44

Galaxy India

Workflow Visualize Data Help User Home History Notifications Settings

Using 16%

Tools

Get Data

- Adds environment variables From Copernicus and etopo given geolocalized and timestamped observations
- AquaINFRA Importer downloads content via the AquaINFRA interaction platform
- Argo data access for global ocean in situ observing system
- BioMart Ensembl server
- BioMart Test server
- CBI Rice Mart rice mart
- DoRINA Search DoRINA data source
- Download and Extract Reads in BAM format from NCBI SRA
- Download and Extract Reads in FASTQ format from NCBI SRA
- Download and Generate Pileup Format from NCBI SRA

Edit Dataset Attributes

Name

bamCoverage on data 44|MINUS strand

Info

Due to filtering, 5.062519817693234% of the aforementioned alignments will be used

Annotation - optional

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

Database/Build - optional

Mouse Dec. 2011 (GRCm38/mm10) (mm10)

Save Auto-detect

History

De novo transcriptome reconstruction

3.84 GB

88: bamCoverage on data 47  
87: bamCoverage on data 46  
86: bamCoverage on data 45  
85: bamCoverage on data 44  
84: bamCoverage on data 47 PLUS strand  
83: bamCoverage on data 46 PLUS strand  
82: bamCoverage on data 45 PLUS strand  
81: bamCoverage on data 44

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for 'India'. The left sidebar includes 'Upload', 'Tools', 'Workflows', 'Workflow Invocations', 'Visualization', 'Histories', 'History', 'Notifications', and 'Settings'. The main area displays the 'Edit Dataset Attributes' page for a dataset named 'bamCoverage on data 45 MINUS strand'. The 'Name' field contains the dataset name. The 'Info' section notes that 4.871603601738894% of alignments will be used after filtering. The 'Annotation' and 'Database/Build' fields are empty. At the bottom are 'Save' and 'Auto-detect' buttons. To the right is a 'History' panel showing a list of datasets, all of which are 'bamCoverage on data' followed by a number and either 'MINUS strand' or 'PLUS strand'.

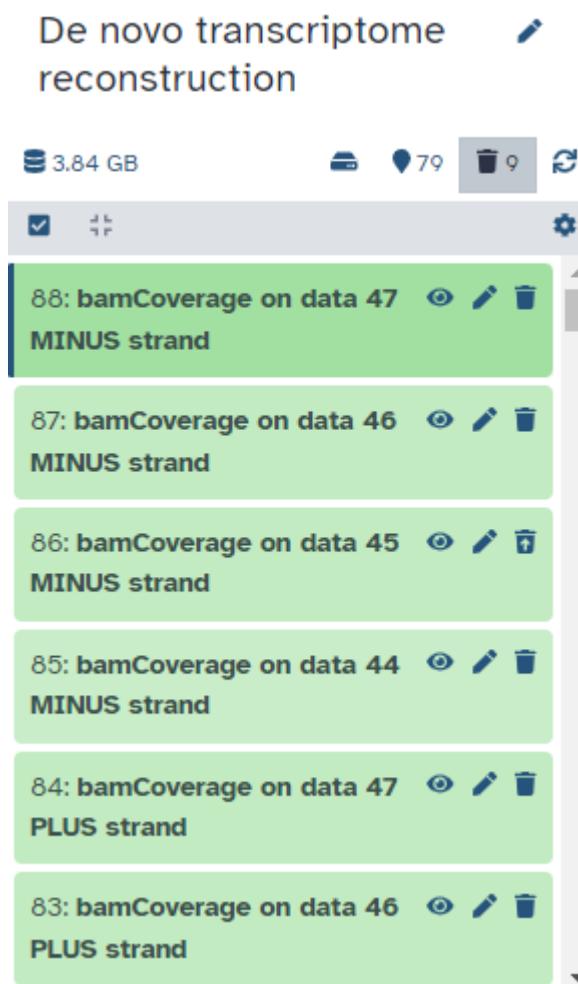
This screenshot shows the same Galaxy interface as the first one, but for a different dataset. The 'Name' field now contains 'bamCoverage on data 46 MINUS strand'. The 'Info' section notes that 21.242843994744064% of alignments will be used after filtering. The 'Annotation' and 'Database/Build' fields are empty. The 'Save' and 'Auto-detect' buttons are at the bottom. The 'History' panel on the right shows a list of datasets, all of which are 'bamCoverage on data' followed by a number and either 'MINUS strand' or 'PLUS strand', matching the structure in the first screenshot.

# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy India, Workflow, Visualize, Data, Help, User, Notifications.
- Left Sidebar:** Upload, Tools (selected), Workflows, Invocations, Visualization, Histories, History (selected), Notifications, Settings.
- Main Content:**
  - Tools:** Search bar for tools.
  - Get Data:**
    - Adds environment variables From Copernicus and etopo given geolocalized and timestamped observations
    - AquaINTRA Importer downloads content via the AquaINTRA Interaction platform
    - Argo data access for global ocean in situ observing system
    - BioMart Ensembl server
    - BioMart Test server
    - CBI Rice Mart rice mart
    - DoRINA Search DoRINA data source
    - Download and Extract Reads in BAM format from NCBI SRA
    - Download and Extract Reads in FASTQ format from NCBI SRA
    - Download and Generate Pileup Format from NCBI SRA
  - Edit Dataset Attributes:** Name: bamCoverage on data 47 MINUS strand. Info: Due to filtering, 48.029653727048085% of the aforementioned alignments will be used. Annotation: optional. Database/Build: optional. Buttons: Save, Auto-detect.- Right Sidebar:** History (deleted: any visible: true). De novo transcriptome reconstruction (3.84 GB, 79 items). A list of datasets:
  - 88: bamCoverage on data 47 (Binary UCSC BigWig file)
  - 87: bamCoverage on data 46 MINUS strand
  - 86: bamCoverage on data 45

# De novo transcriptome reconstruction with RNA-Seq



# De novo transcriptome reconstruction with RNA-Seq

## Trackster based visualization:

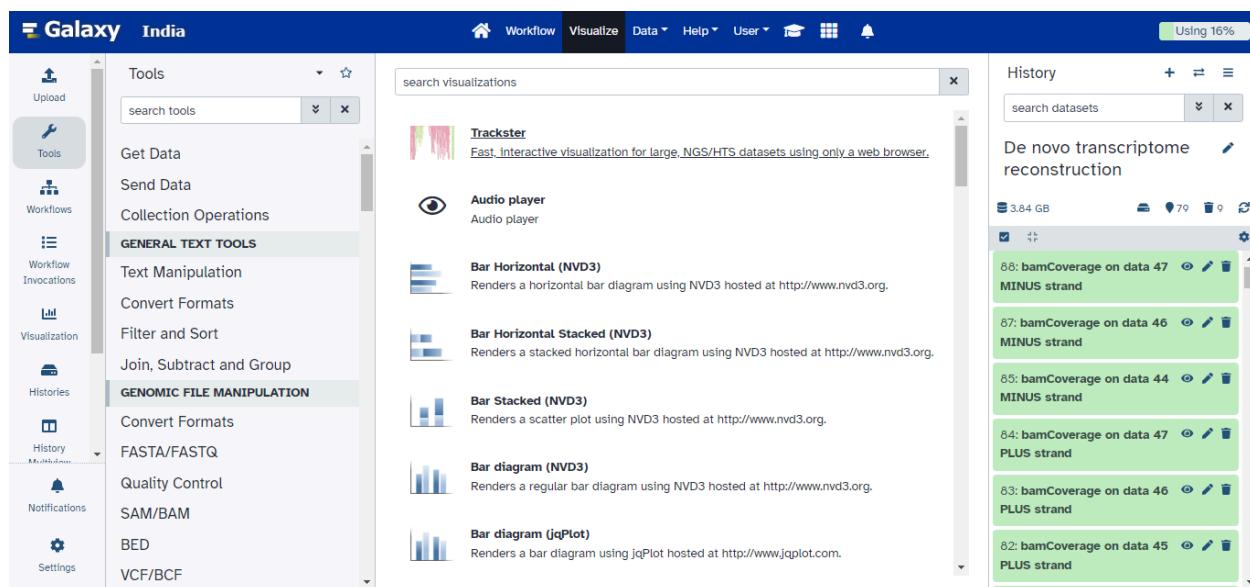
At **chr11:96191452-96206029**, transcript structures from RefSeq, StringTie, and GFFCompare enable comparison of known and predicted transcripts. The bigWig tracks display RNA-seq read density, with blue for the positive strand and red for the negative strand, reflecting transcript abundance. This visualization aids in identifying expressed transcripts and validating predictions. It also provides insights into gene expression variations at this locus, highlighting important aspects relevant to biological research.

Viz tool On the center console at the top of the Galaxy interface, choose “Visualization” -> “New track browser”

- Name your visualization something descriptive under “Browser name:”
- Choose “Mouse Dec. 2011 (GRCm38/mm10) (mm10)” as the “Reference genome build (dbkey)”
- Click “Create” to initiate your Trackster session
- Viz tool : Click “Add datasets to visualization”
  - Select the “RefSeq GTF mm10” file
  - Select the output files from Stringtie
  - Select the output file from GFFCompare
  - Select the output files from bamCoverage
- Tool : Using the grey labels on the left side of each track, drag and arrange the track order to your preference.
- Tool : Hover over the grey label on the left side of the “RefSeq GTF mm10” track and click the “Edit settings” icon.
  - Adjust the block color to blue (#0000ff) and antisense strand color to red (#ff0000)

# De novo transcriptome reconstruction with RNA-Seq

- Tool : Repeat the previous step on the output files from StringTie and GFFCompare.
- Tool : Hover over the grey label on the left side of the “G1E R1 plus” track and click the “Edit settings” icon.
  - Adjust the color to blue (#0000ff)
- Tool : Repeat the previous step on the other three bigWig files representing the plus strand.
- Tool : Hover over the grey label on the left side of the “G1E R1 minus” track and click the “Edit settings” icon.
  - Adjust the color to red (#ff0000)
- Tool : Repeat the previous step on the other three bigWig files representing the minus strand.
- Direct Trackster to the coordinates: chr11:96191452-96206029
- Tool : Adjust the track height of the bigWig files to be consistent for each set of plus strand and minus strand tracks.

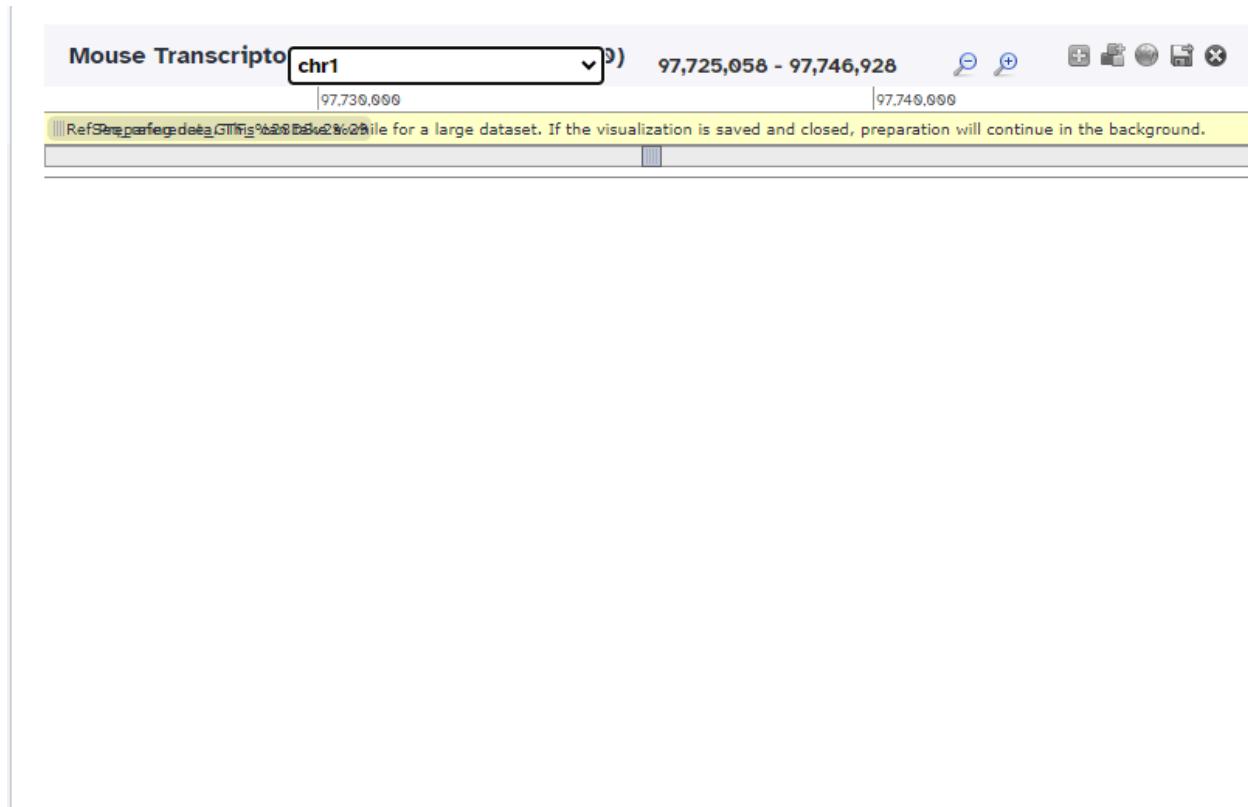


# De novo transcriptome reconstruction with RNA-Seq

The screenshot shows the Galaxy web interface for the 'India' instance. The left sidebar includes sections for Upload, Tools, Workflows, Workflow Invocations, Visualization, Histories, Notifications, and Settings. The main area displays a search bar for visualizations and a list of available tools. A dropdown menu under 'Select a dataset to visualize' shows 'RefSeq\_reference\_GTF\_%26DSv2%29'. Below this is a button labeled 'Create Visualization'. To the right, a 'History' panel shows a list of datasets, all of which are 'bamCoverage' type on various data sets (47, 46, 44, 47, 46, 45) for both MINUS and PLUS strands.

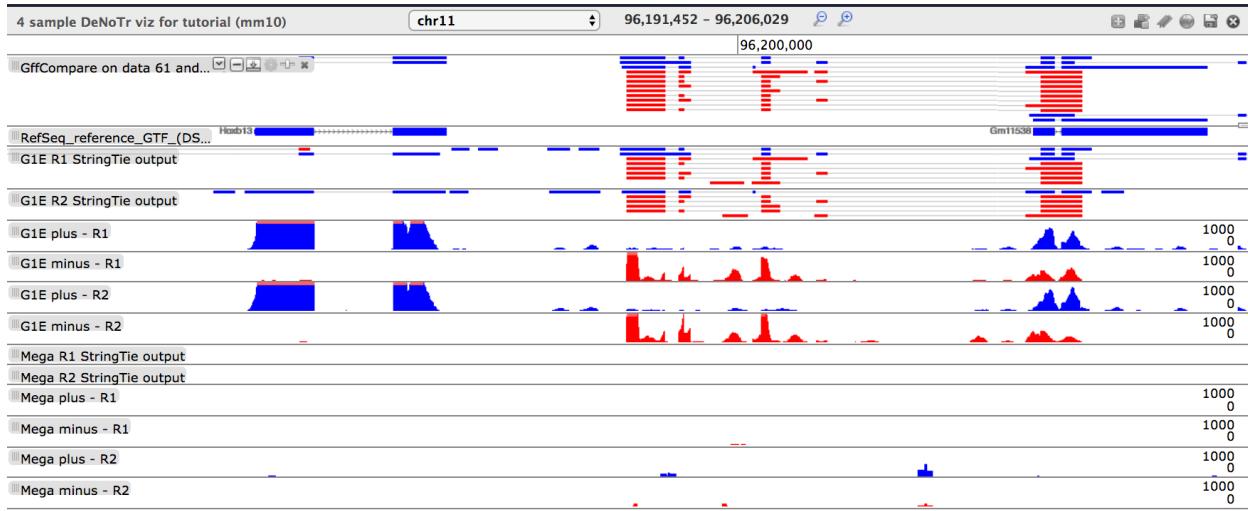
This screenshot shows the Galaxy interface with a 'New Visualization' dialog open. The dialog fields include 'Browser name' set to 'Mouse Transcriptome A' and 'Reference genome build (dbkey)' set to 'Mouse (Mus Musculus); mm10 Full'. A note below states 'Is the build not listed here? Add a Custom Build'. At the bottom of the dialog are 'Cancel' and 'Create' buttons. The rest of the interface is identical to the first screenshot, showing the same sidebar and history panel.

# De novo transcriptome reconstruction with RNA-Seq



A screenshot of the Galaxy web interface. On the left, there is a sidebar with various navigation options: Workflow Invocations, Visualization (which is selected), Histories, Multiview, Datasets, Pages, Notifications, and Settings. The main content area shows a "Configure Track" dialog for a "Mouse Trans" visualization. The dialog has fields for "Name" (set to "RefSeq\_reference\_GTF"), "Block color" (set to "#0000ff"), "Antisense strand color" (set to "#ff0000"), and "Label color" (set to "#808080"). There is also a checkbox for "Show summary counts". At the bottom right of the dialog are "Cancel" and "OK" buttons. To the right of the dialog, there is a "History" panel showing a workflow named "De novo transcriptome reconstruction" with four steps: "bamCoverage" (step 66), "bamCoverage" (step 67), "bamCoverage" (step 68), and "bamCoverage" (step 65). The "bamCoverage" steps are associated with "NUS strand" and "INUS strand" data.

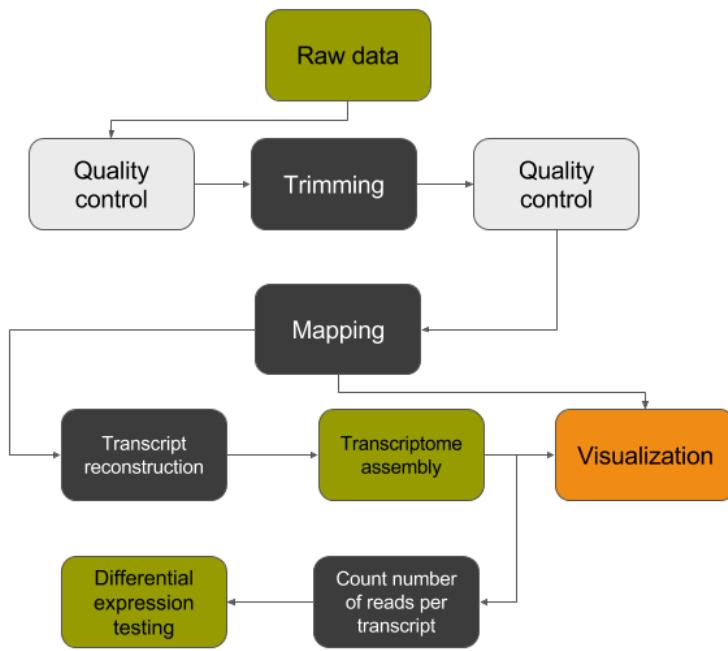
# De novo transcriptome reconstruction with RNA-Seq



## Conclusion:

In this study, we analyzed RNA sequencing (RNA-seq) data from two different cell states, **G1E** and **megakaryocytes**, to understand which genes are active and how they differ between these states. We used a de novo transcriptome reconstruction approach to identify both known and new genes involved in cell differentiation. A key focus was on how the transcription factor Tal1 regulates gene expression during hematopoiesis (the formation of blood cells), providing insights into its role at various stages of blood cell development.

# De novo transcriptome reconstruction with RNA-Seq



The workflow depicted in the image represents the general process of RNA-seq data analysis, particularly focusing on de novo transcriptome reconstruction, starting from raw RNA sequencing data. This workflow involves several key steps:

1. **Raw data acquisition:** Starting with the collection of RNA-seq data.
2. **Quality control:** Ensuring the raw data is free from errors, such as sequencing artifacts, before and after trimming.
3. **Trimming:** Removing low-quality bases or sequences, such as adapters.
4. **Mapping:** Aligning reads to a reference genome or transcriptome, which allows for the identification of where transcripts are expressed.
5. **Transcriptome assembly:** Building a representation of all the transcripts based on the RNA-seq data.
6. **Visualization:** Representing the assembled transcriptome and gene expression levels in a visual form for interpretation.

# De novo transcriptome reconstruction with RNA-Seq

7. **Count number of reads per transcript:** Quantifying how many reads map to each transcript.
8. **Differential expression testing:** Identifying which genes are expressed differently between the two cellular states (G1E and megakaryocytes).
9. **Transcript reconstruction:** Rebuilding the transcripts from mapped reads to explore novel or unannotated transcripts.

The provided GEO accession (GSE51338) corresponds to a dataset that includes RNA-seq data from G1E (a GATA-null cell line derived from mouse embryonic stem cells) and megakaryocytes. The goal is to examine gene expression in these cell states and determine how the transcription factor Tal1, which is crucial for hematopoiesis, regulates gene expression during different stages of hematopoietic differentiation.

In this case, the workflow enables an investigation into how Tal1 influences gene regulation in G1E cells and megakaryocytes, which may provide insights into hematopoietic differentiation.