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

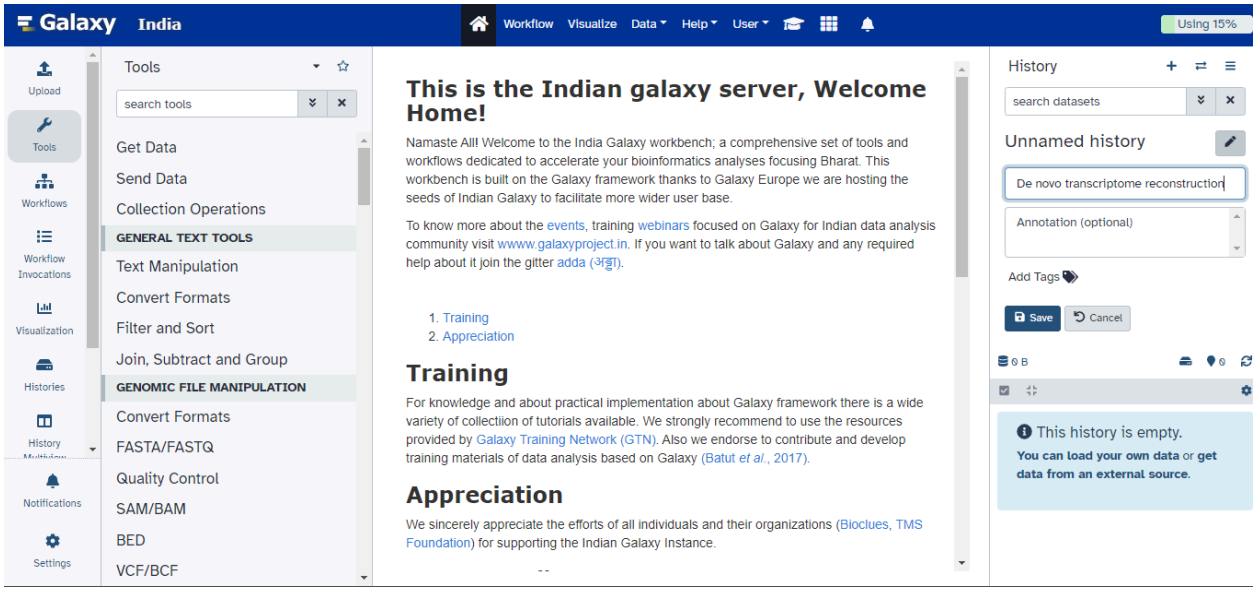
Data upload:

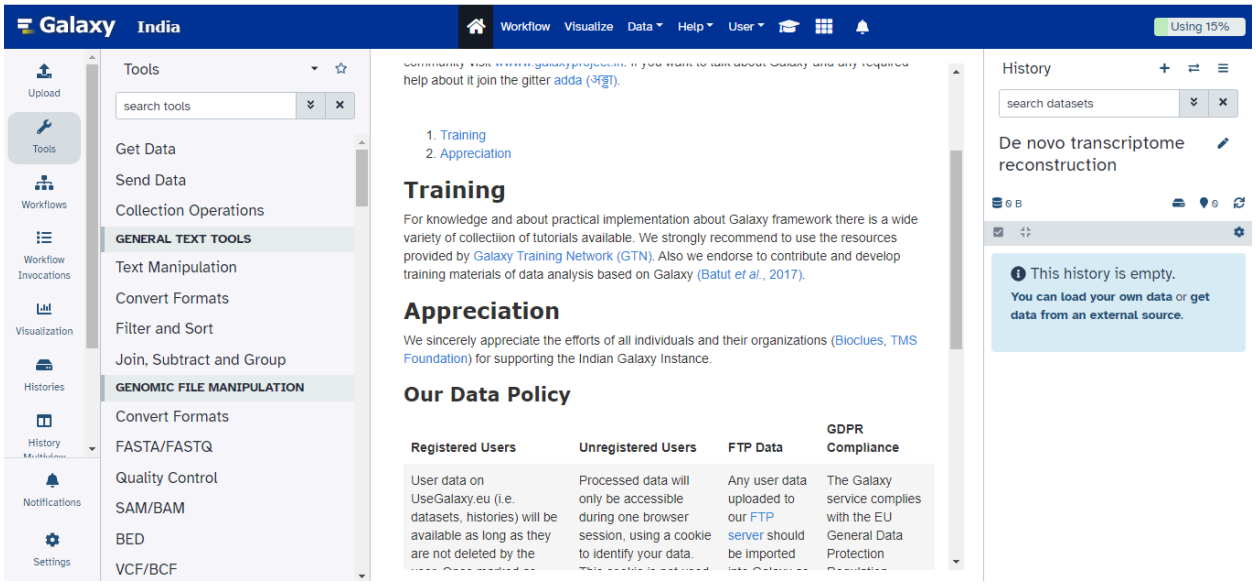
This data is available at Zenodo, where you can find the forward and reverse reads corresponding to replicate RNA-seq libraries from G1E and megakaryocyte cells and an annotation file of RefSeq transcripts we will use to generate our transcriptome database.

Data upload:

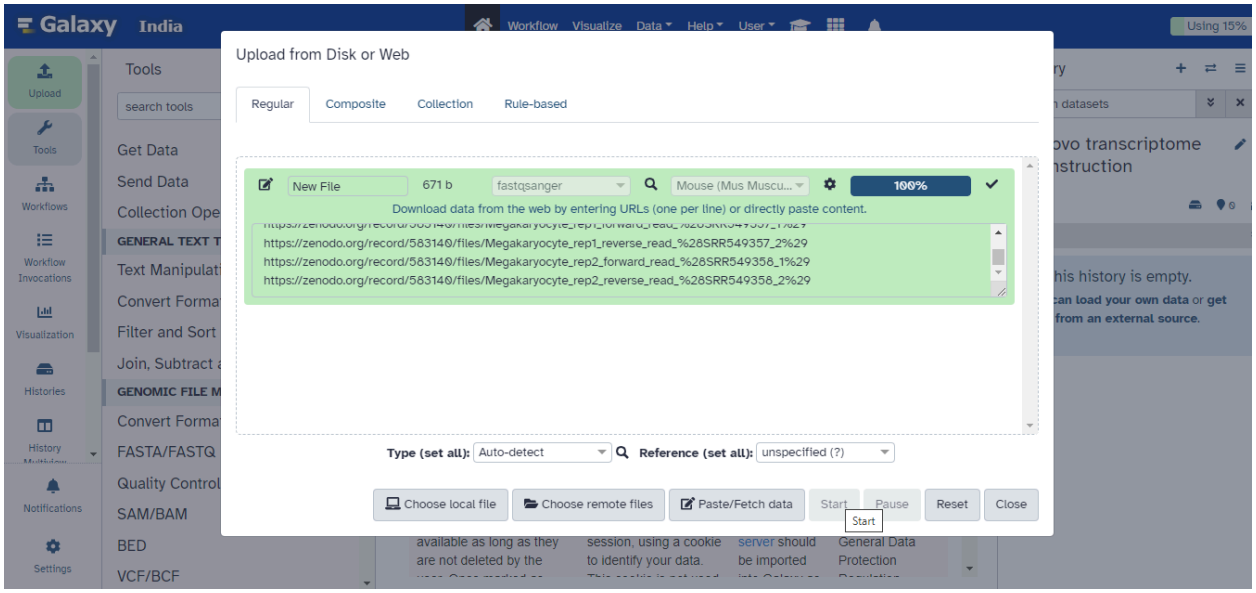
1. Create a new history for this RNA-seq exercise
- Tip: Creating a new history
- Tip: Renaming a history
2. Open the data upload manager (Get Data -> Upload file)
3. Copy and paste the links for the reads and annotation file
4. Select Paste/Fetch Data
5. Paste the link(s) into the text field
6. Change 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

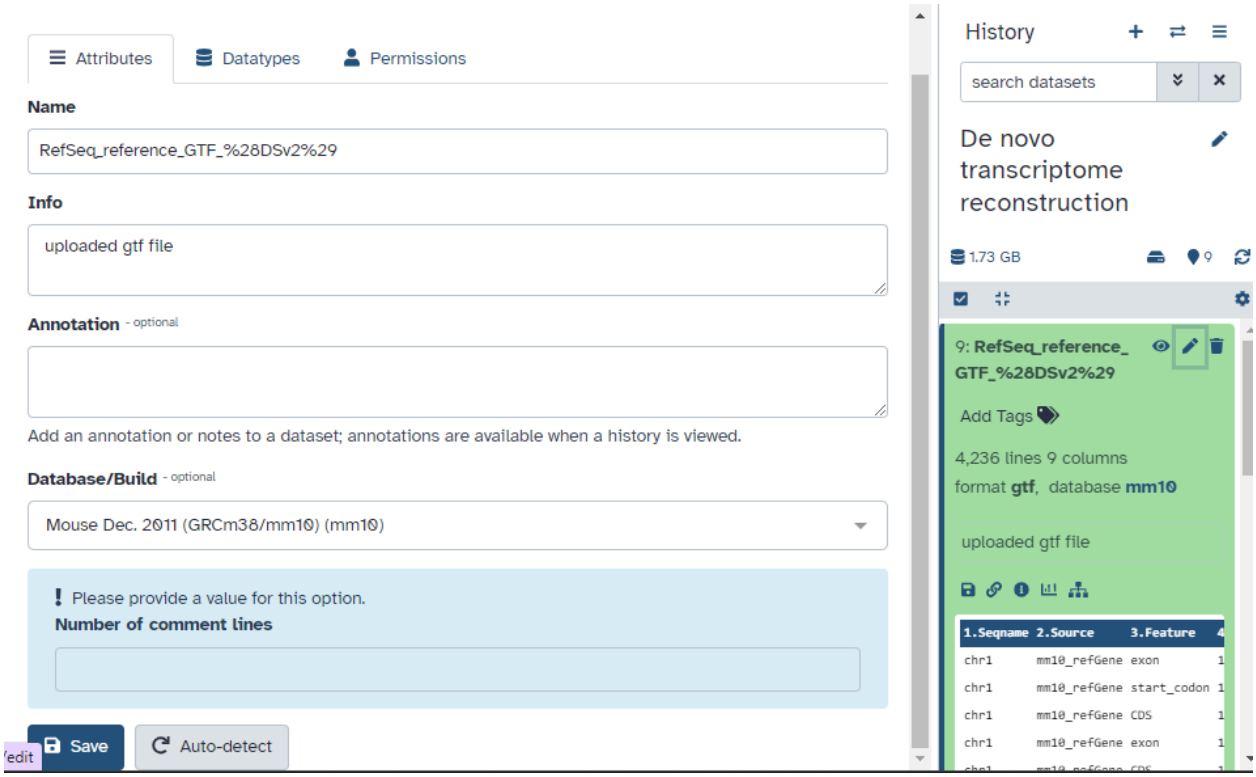
De novo transcriptome reconstruction with RNA-Seq



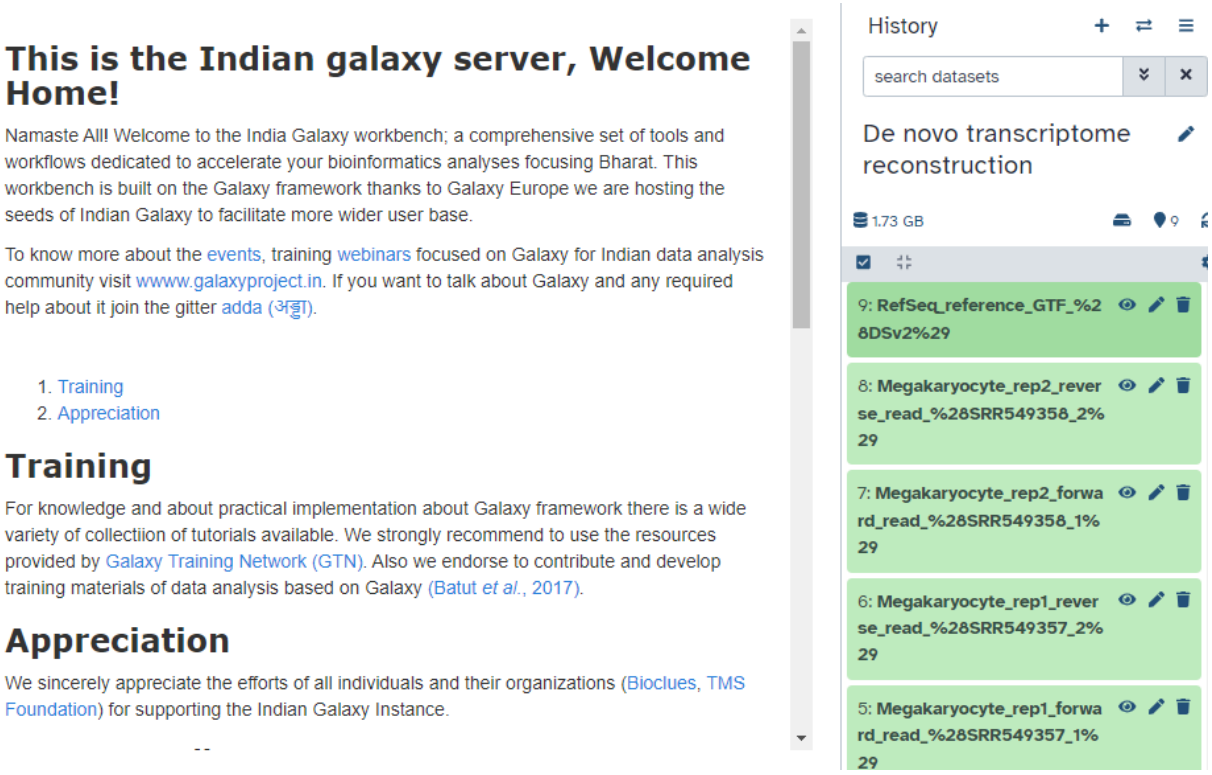


De novo transcriptome reconstruction with RNA-Seq





De novo transcriptome reconstruction with RNA-Seq



De novo transcriptome reconstruction with RNA-Seq

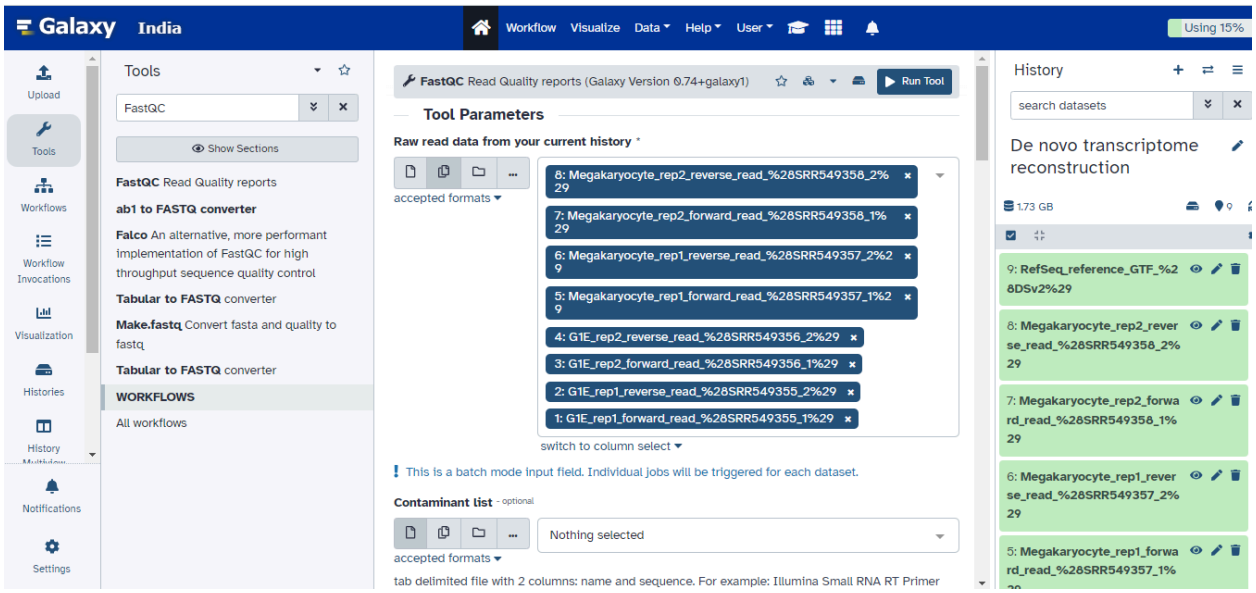
Quality control:

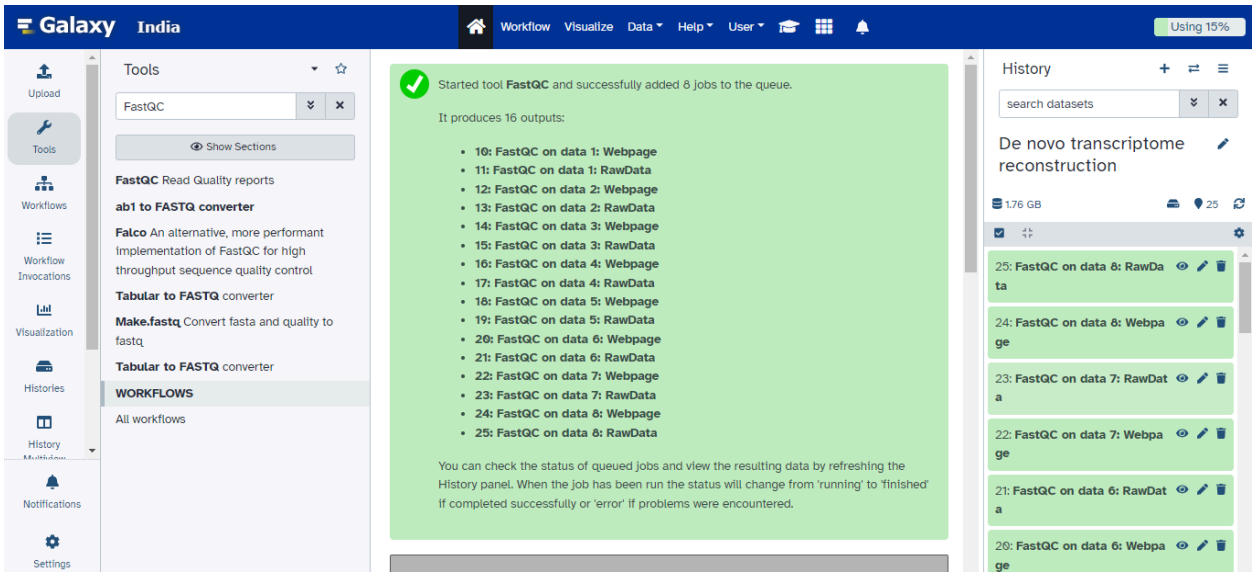
For quality control, we use similar tools as described in the NGS-QC tutorial: FastQC and Trimmomatic.

Quality control:

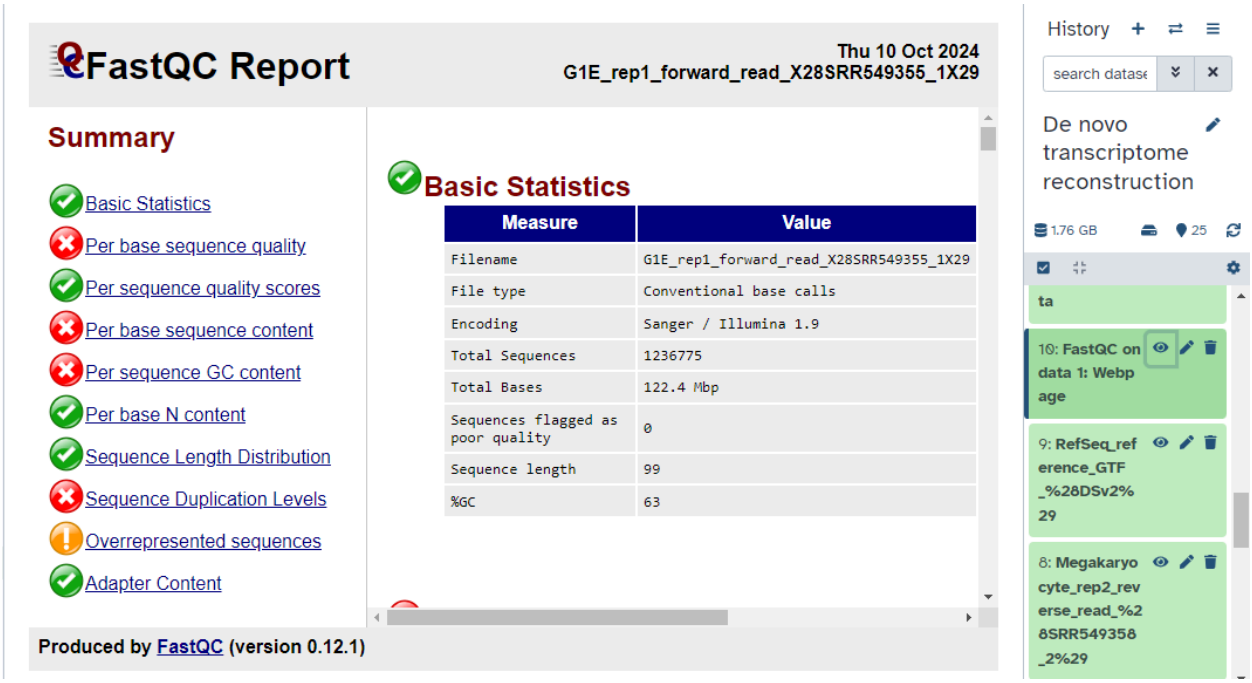
- FastQC tool Run `FastQC` on the forward and reverse read files to assess the quality of the reads.
- Trimmomatic tool Trim off the low quality bases from the ends of the reads to increase mapping efficiency
- FastQC tool Re-run `FastQC` on trimmed reads and inspect the differences.
- Trimmomatic tool: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

De novo transcriptome reconstruction with RNA-Seq

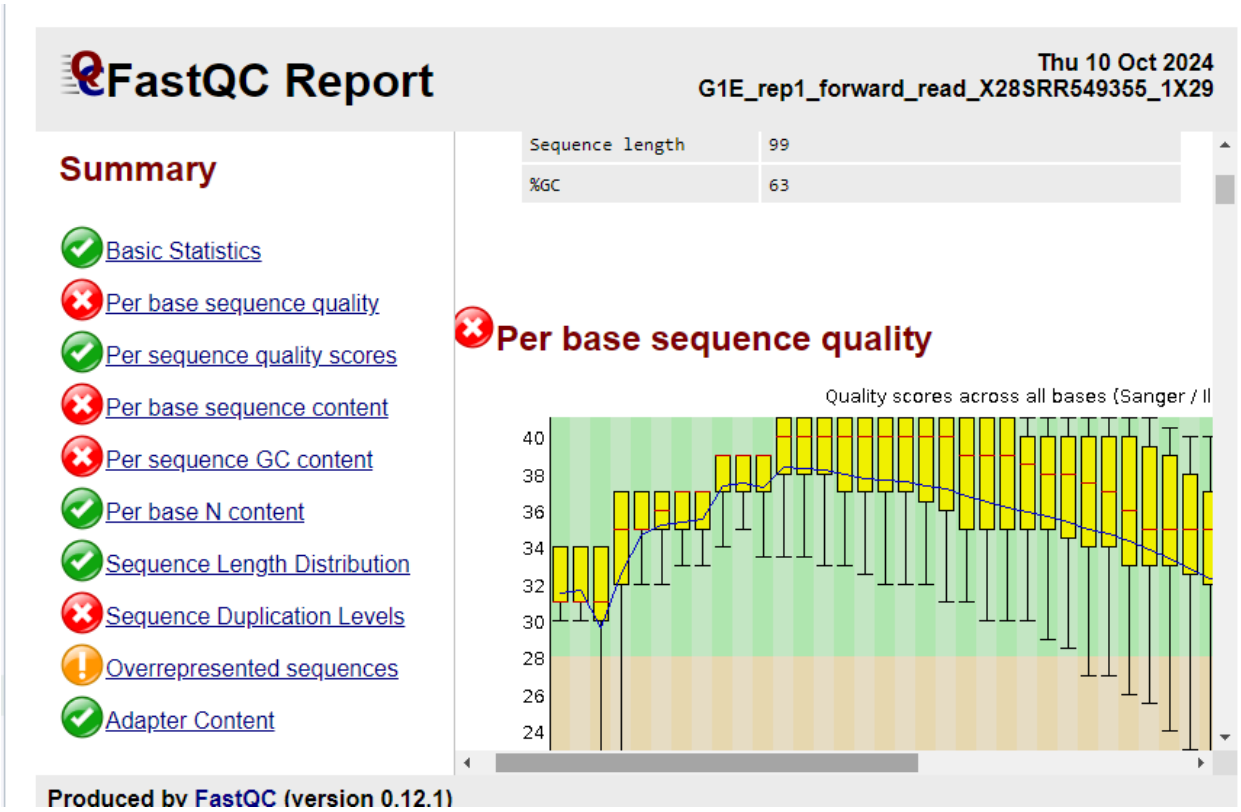




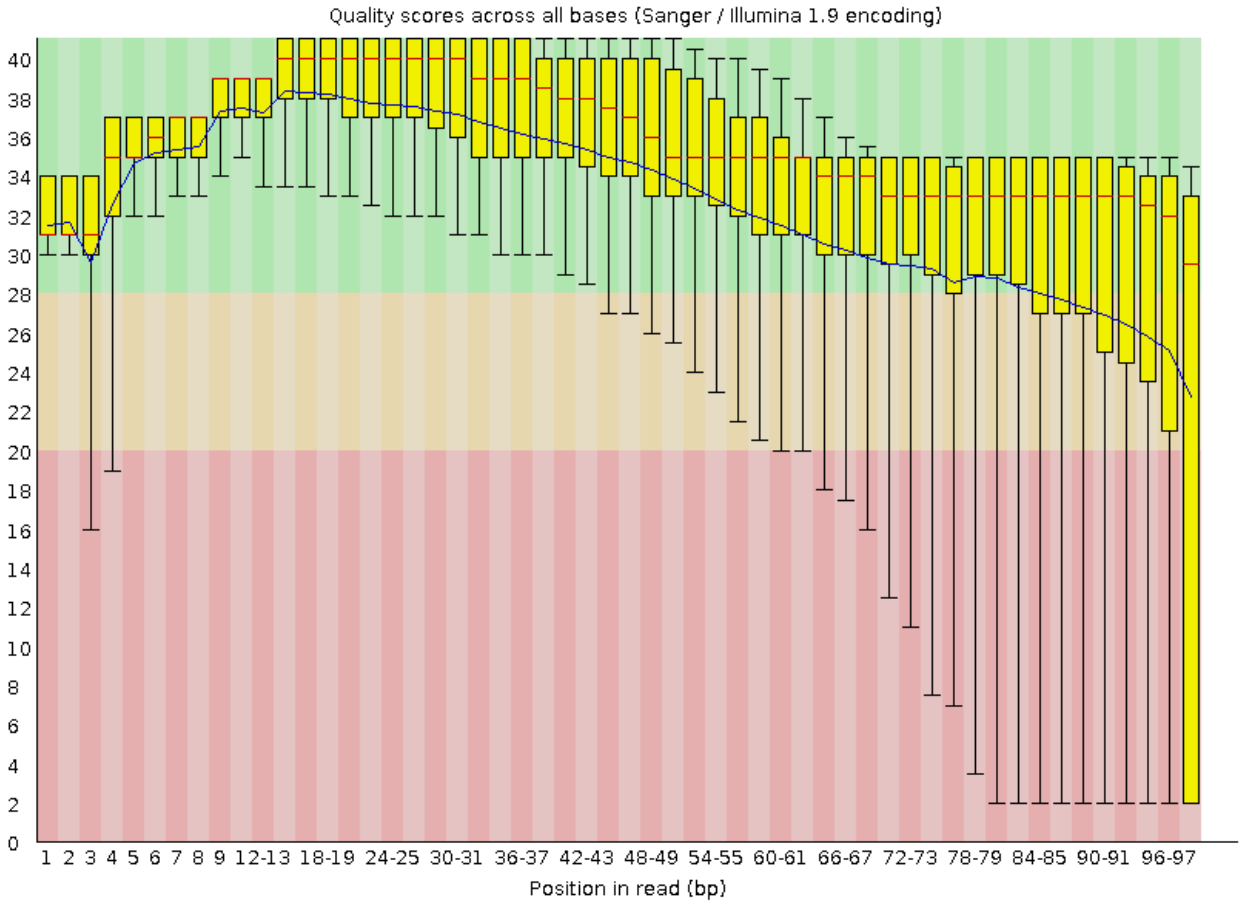
De novo transcriptome reconstruction with RNA-Seq



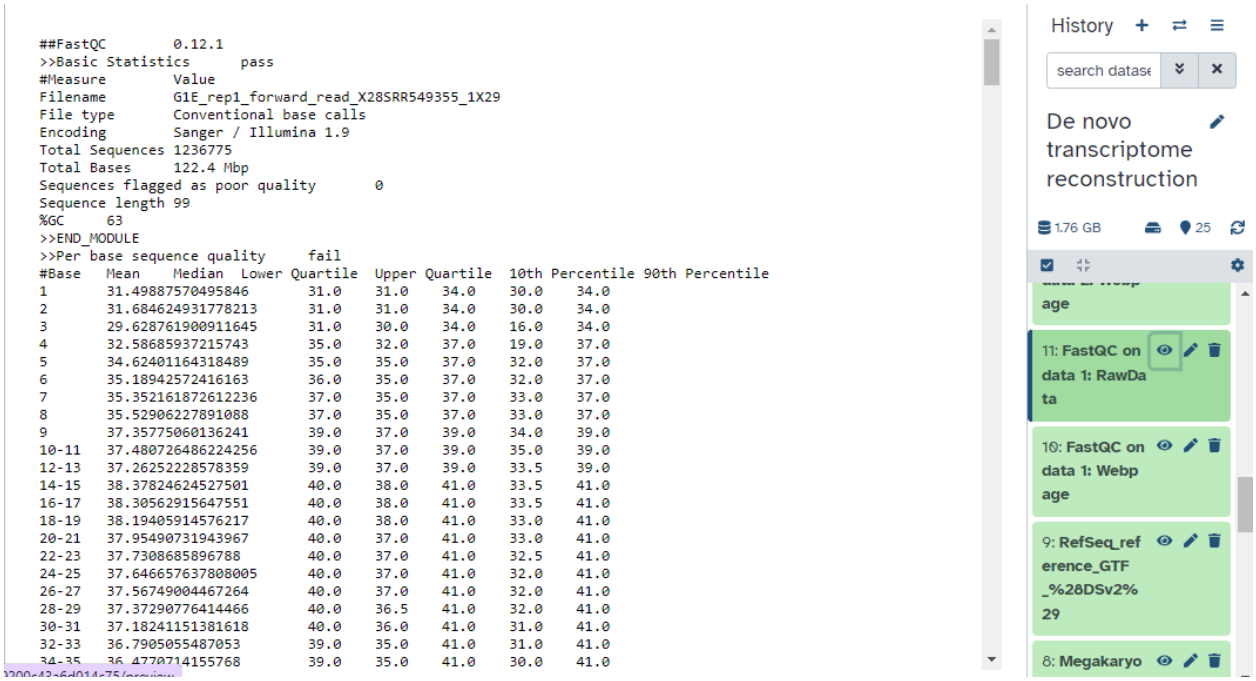
De novo transcriptome reconstruction with RNA-Seq



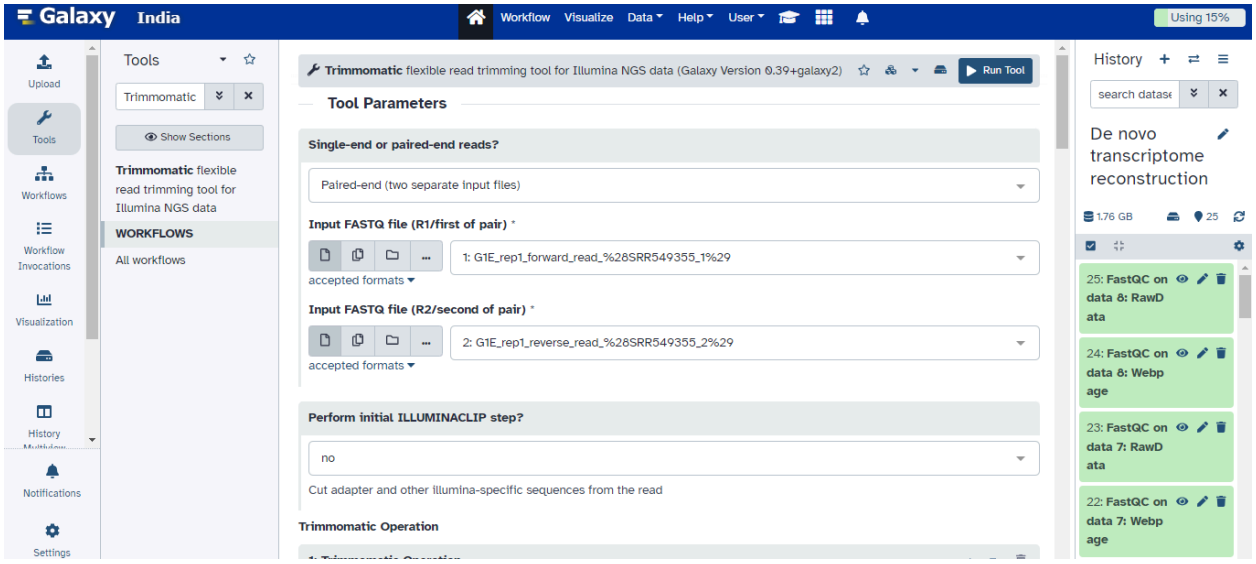
De novo transcriptome reconstruction with RNA-Seq



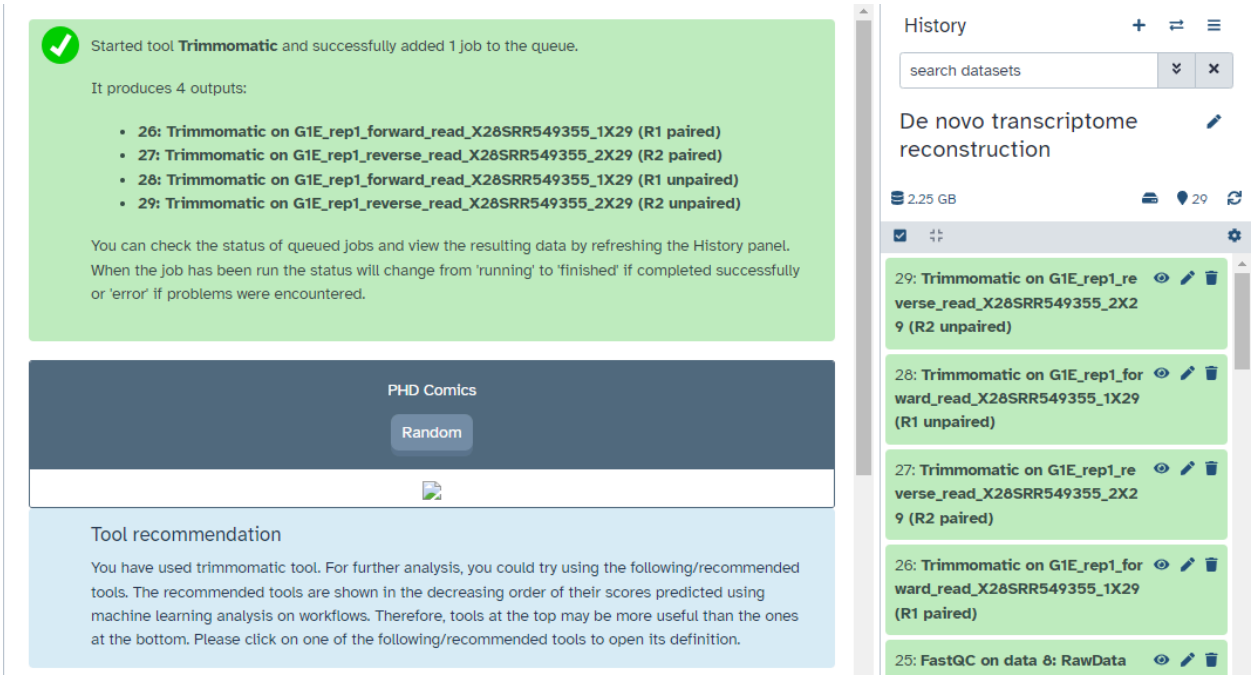
De novo transcriptome reconstruction with RNA-Seq

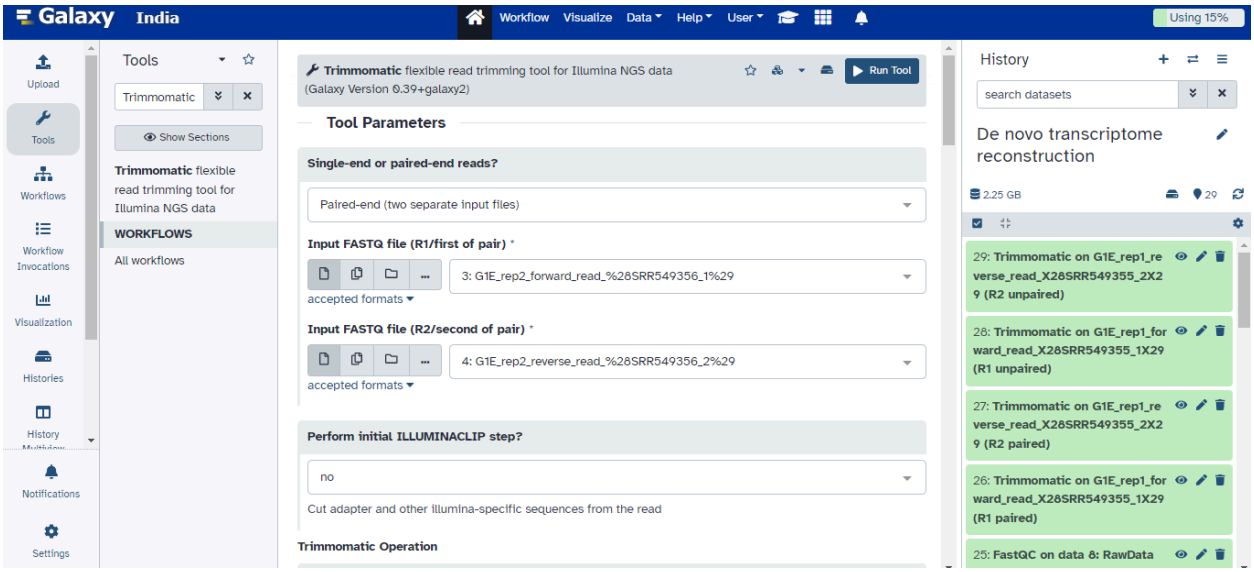


```
##FastQC        0.12.1
>>Basic Statistics      pass
#Measure        Value
Filename        G1E_rep1_forward_read_X28SRR549355_1X29
File type       Conventional base calls
Encoding        Sanger / Illumina 1.9
Total Sequences 1236775
Total Bases     122.4 Mbp
Sequences flagged as poor quality       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
```

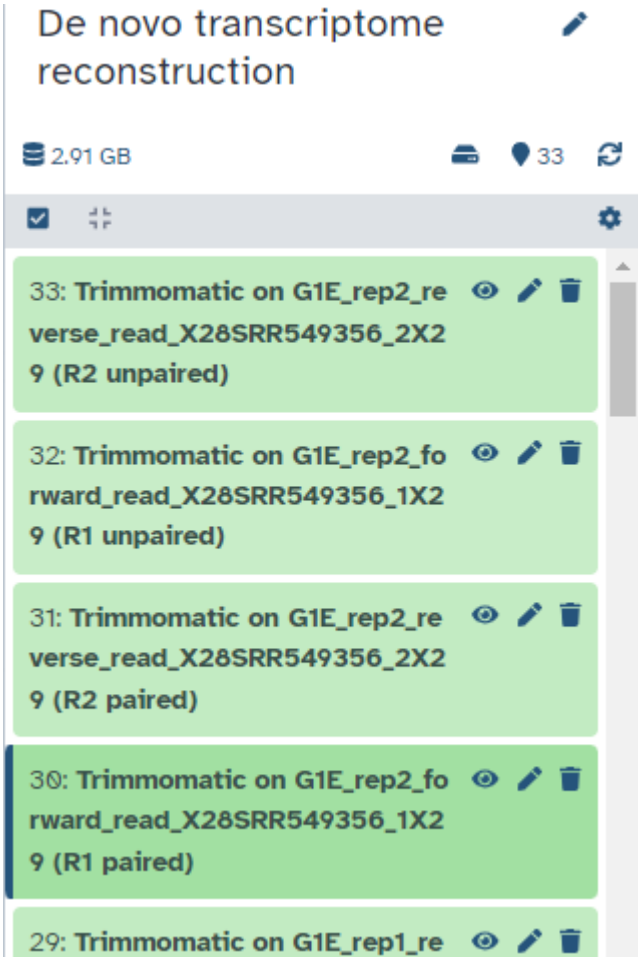



De novo transcriptome reconstruction with RNA-Seq

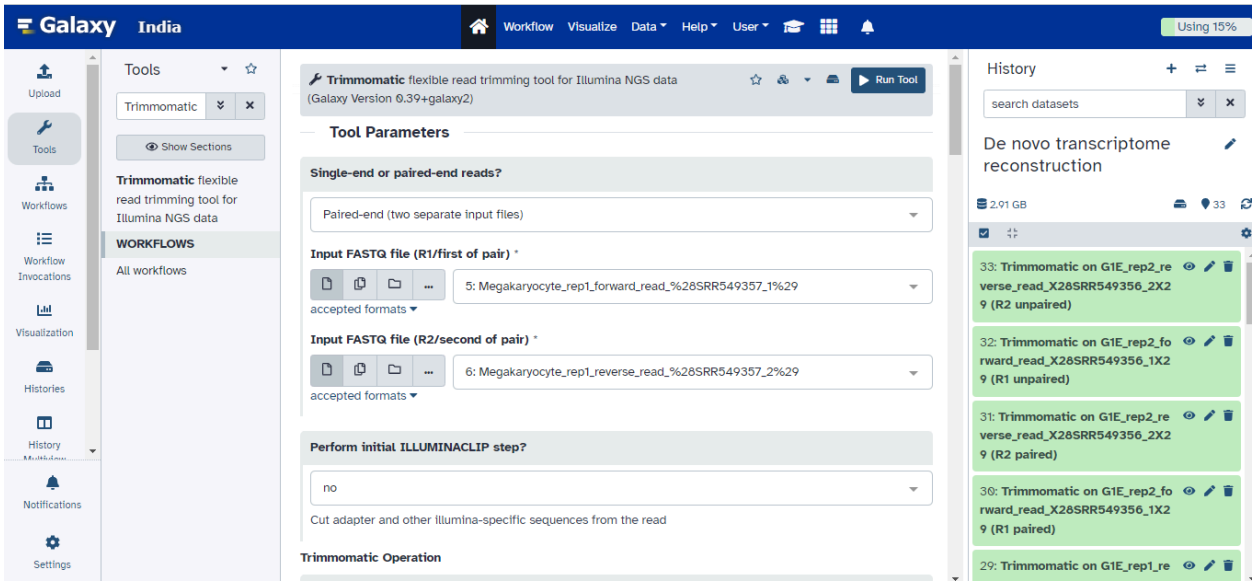


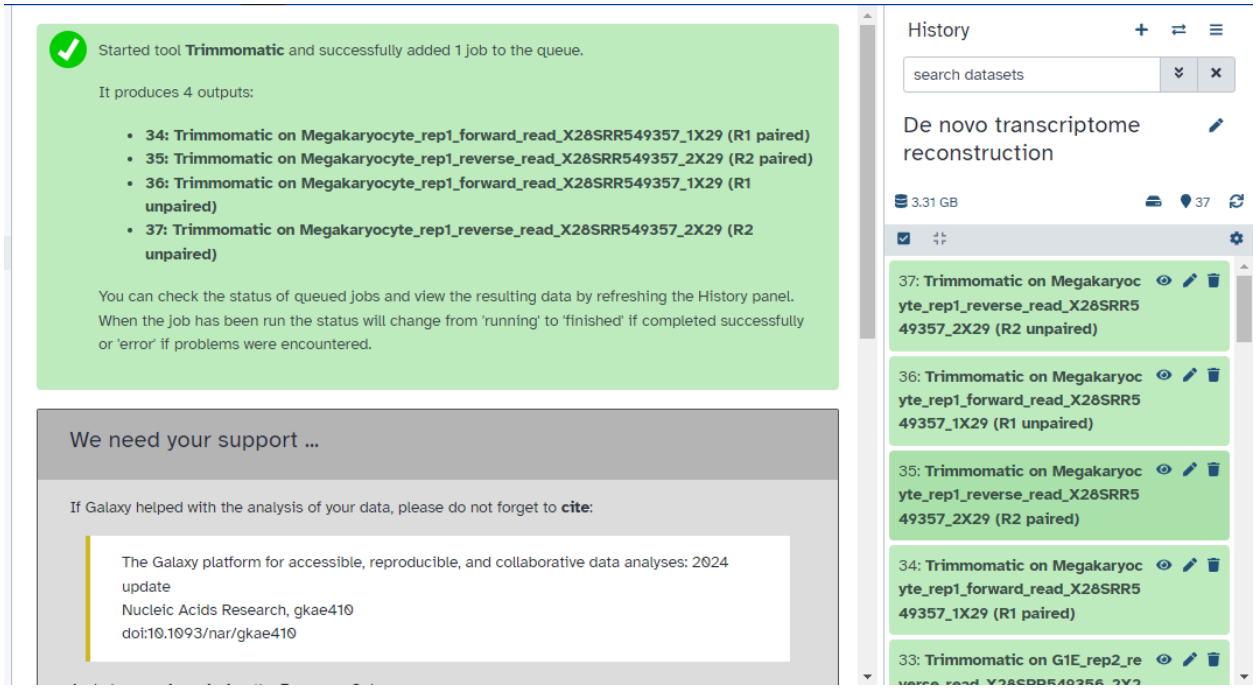


De novo transcriptome reconstruction with RNA-Seq

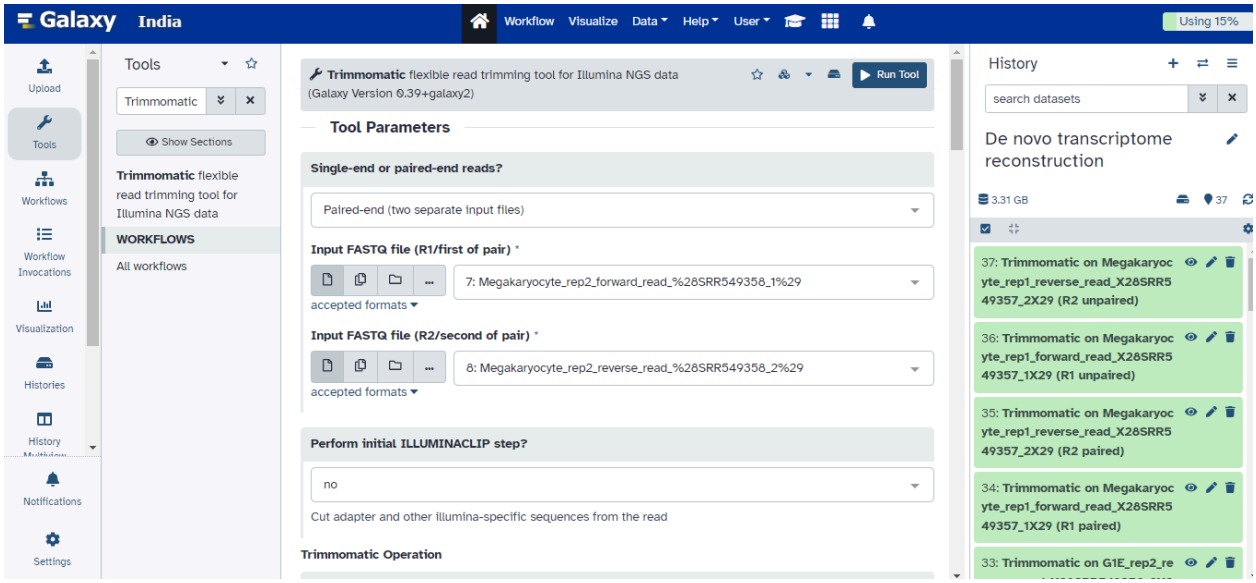


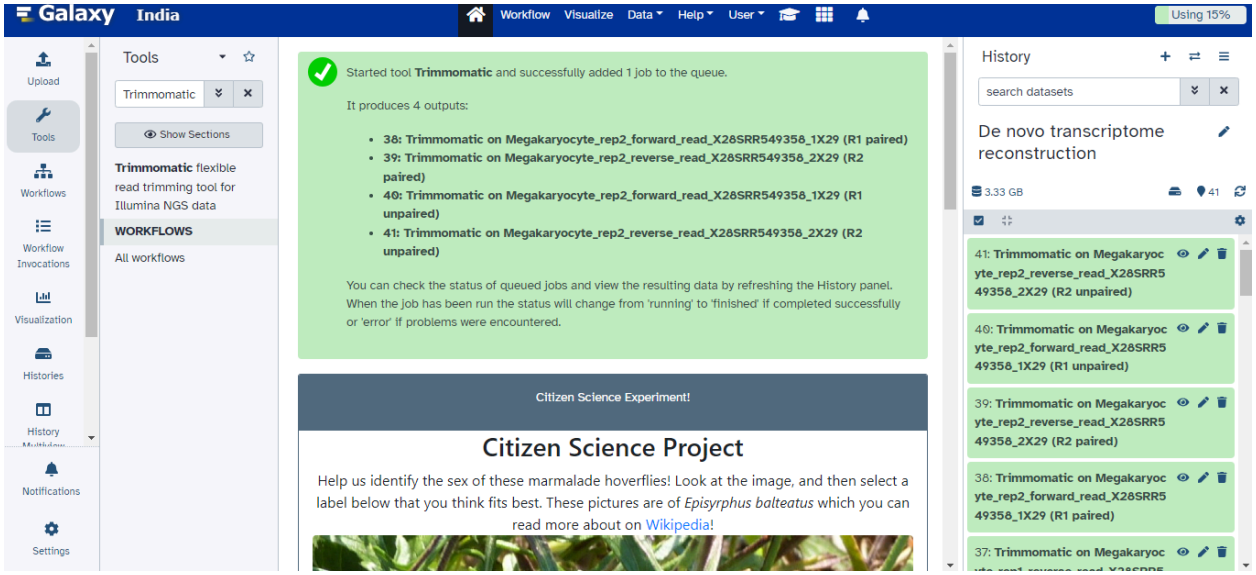
De novo transcriptome reconstruction with RNA-Seq



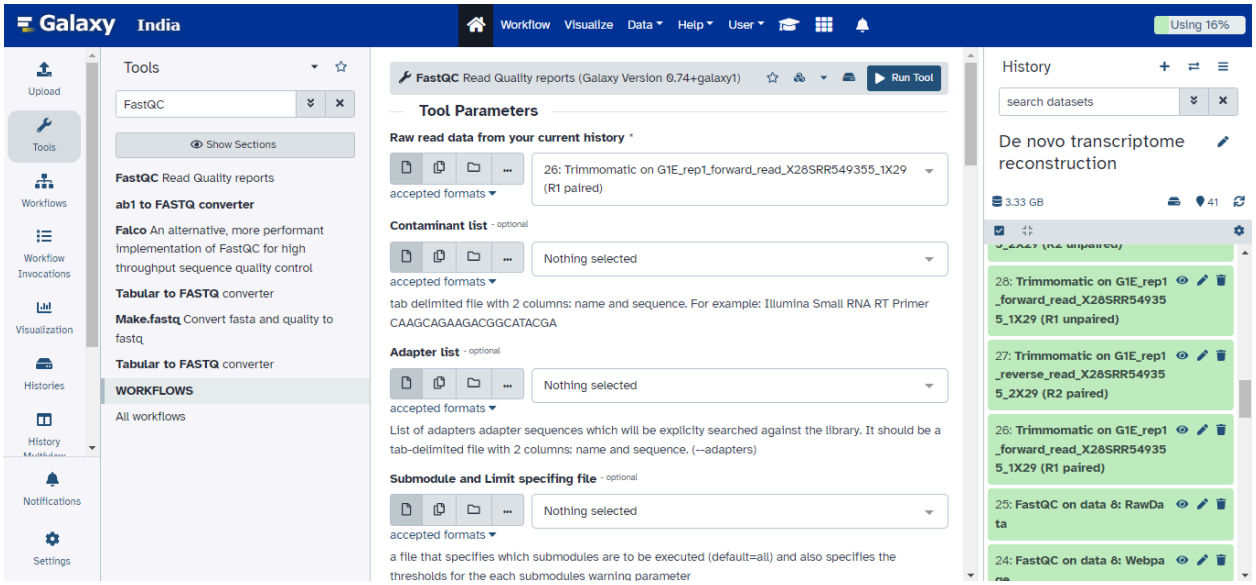


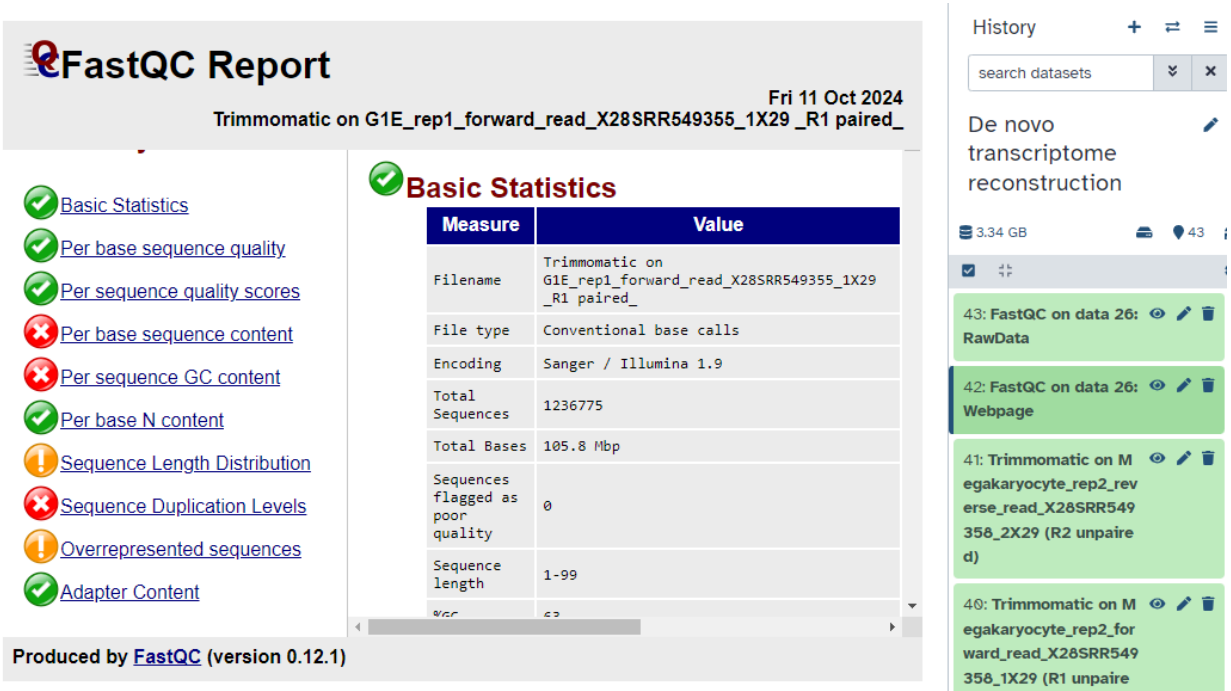
De novo transcriptome reconstruction with RNA-Seq



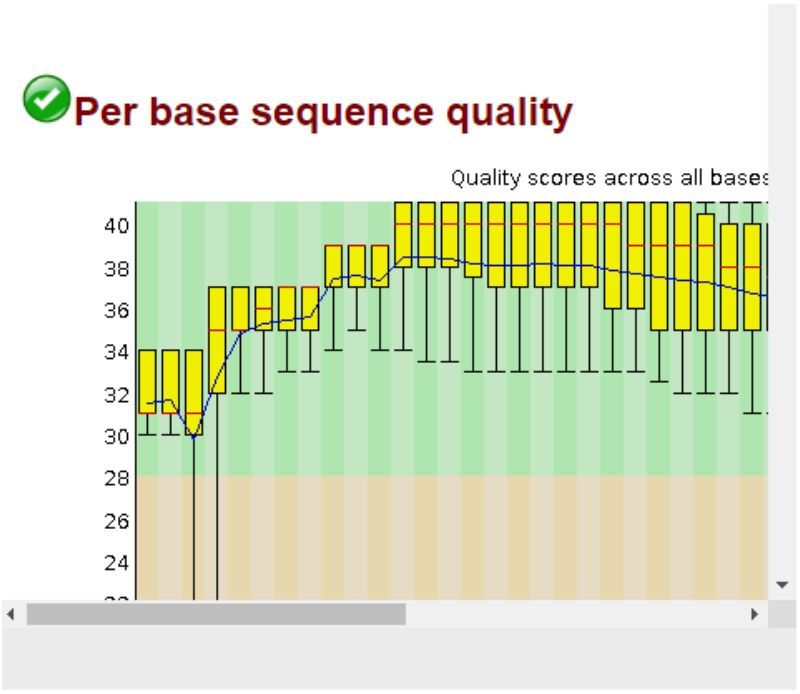


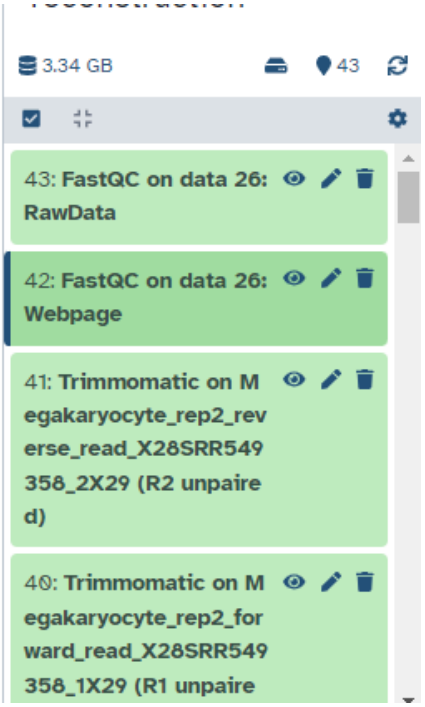
De novo transcriptome reconstruction with RNA-Seq



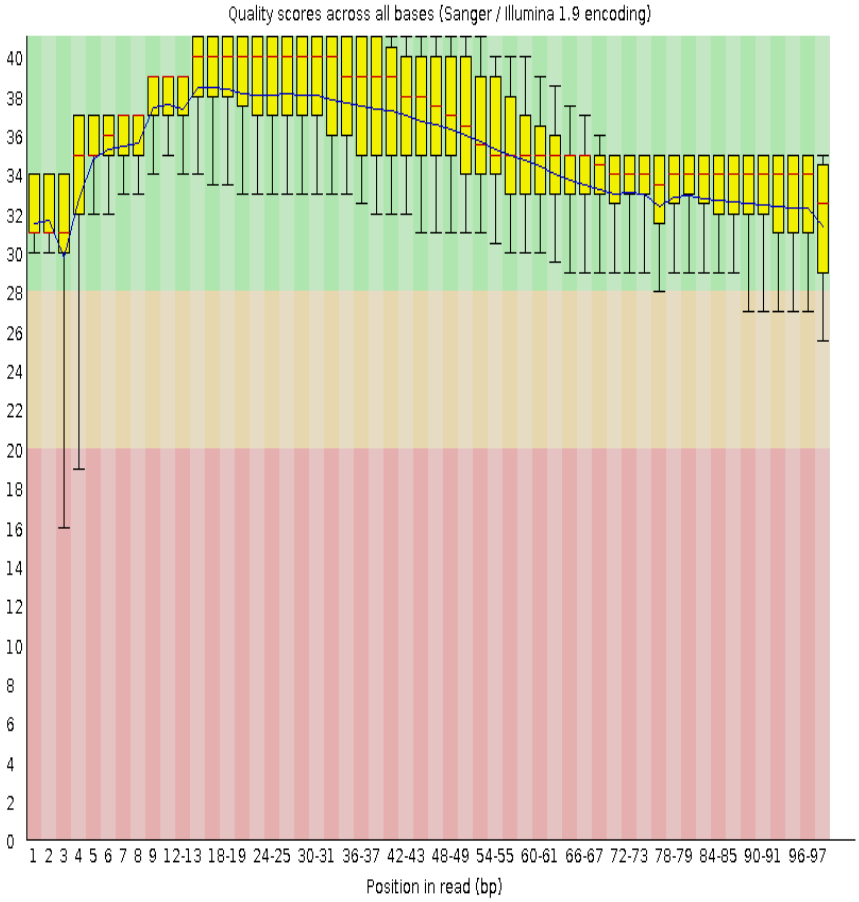


De novo transcriptome reconstruction with RNA-Seq



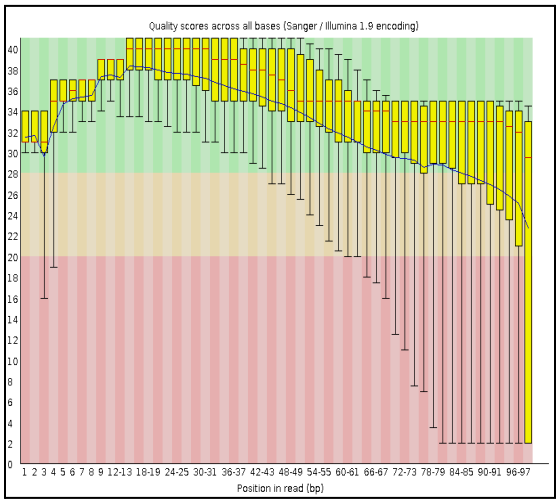


De novo transcriptome reconstruction with RNA-Seq

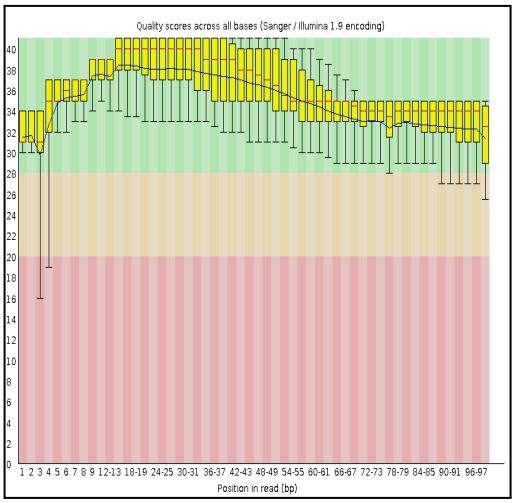


De novo transcriptome reconstruction with RNA-Seq

Before vs after filtering



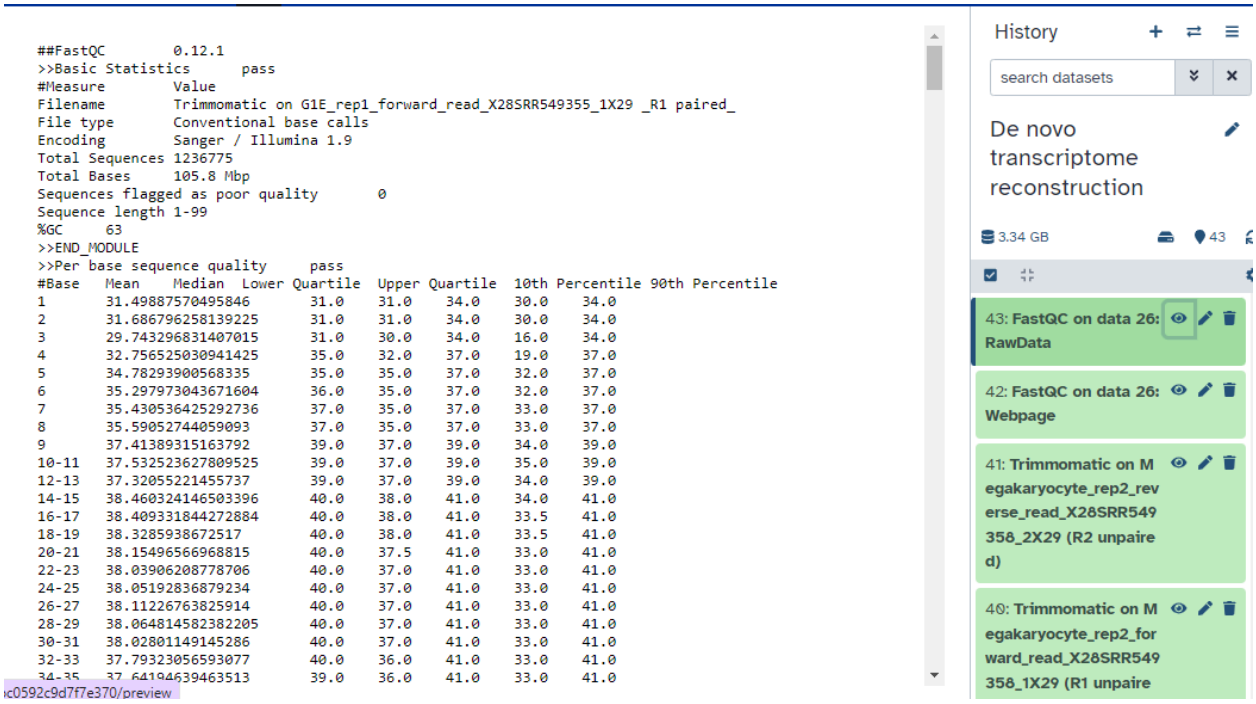
BEFORE TRIMMING



AFTER TRIMMING

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



```
##FastQC        0.12.1
>>Basic Statistics      pass
#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
%GC     63
>>END_MODULE
>>Per base sequence quality      pass
#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       31.686796258139225      31.0    31.0    34.0    30.0    34.0
3       29.743296831407015      31.0    30.0    34.0    16.0    34.0
4       32.756525030941425      35.0    32.0    37.0    19.0    37.0
5       34.78293900568335       35.0    35.0    37.0    32.0    37.0
6       35.297973043671604      36.0    35.0    37.0    32.0    37.0
7       35.430536425292736      37.0    35.0    37.0    33.0    37.0
8       35.59052744059093       37.0    35.0    37.0    33.0    37.0
9       37.41389315163792       39.0    37.0    39.0    34.0    39.0
10-11   37.532523627809525      39.0    37.0    39.0    35.0    39.0
12-13   37.32055221455737       39.0    37.0    39.0    34.0    39.0
14-15   38.460324146503396      40.0    38.0    41.0    34.0    41.0
16-17   38.409331844272884      40.0    38.0    41.0    33.5    41.0
18-19   38.3285938672517        40.0    38.0    41.0    33.5    41.0
20-21   38.15496566968815       40.0    37.5    41.0    33.0    41.0
22-23   38.03906208778706       40.0    37.0    41.0    33.0    41.0
24-25   38.05192836879234       40.0    37.0    41.0    33.0    41.0
26-27   38.11226763825914       40.0    37.0    41.0    33.0    41.0
28-29   38.064814582382205      40.0    37.0    41.0    33.0    41.0
30-31   38.02801149145286       40.0    37.0    41.0    33.0    41.0
32-33   37.79323056593077       40.0    36.0    41.0    33.0    41.0
34-35   37.64194639463513       39.0    36.0    41.0    33.0    41.0
```

ic0592c9d7f7e370/preview

History

search datasets

De novo transcriptome reconstruction

3.34 GB 43

43: FastQC on data 26: RawData

42: FastQC on data 26: Webpage

41: Trimmomatic on Megakaryocyte_rep2_reverse_read_X28SRR549358_2X29 (R2 unpaired)

40: Trimmomatic on Megakaryocyte_rep2_forward_read_X28SRR549358_1X29 (R1 unpaire

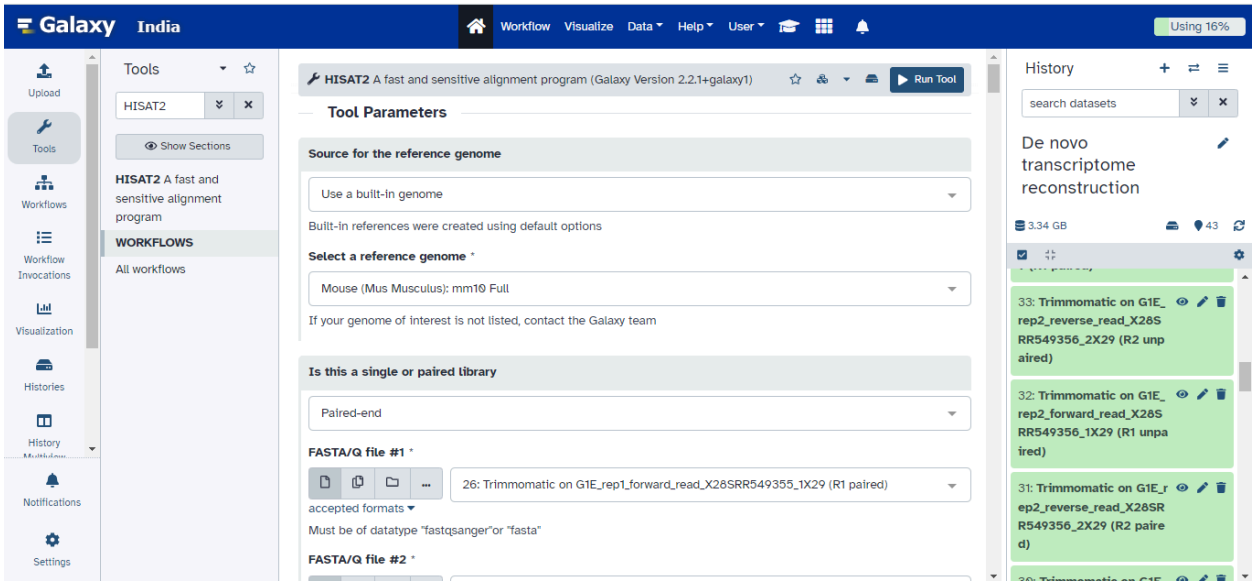
Mapping:

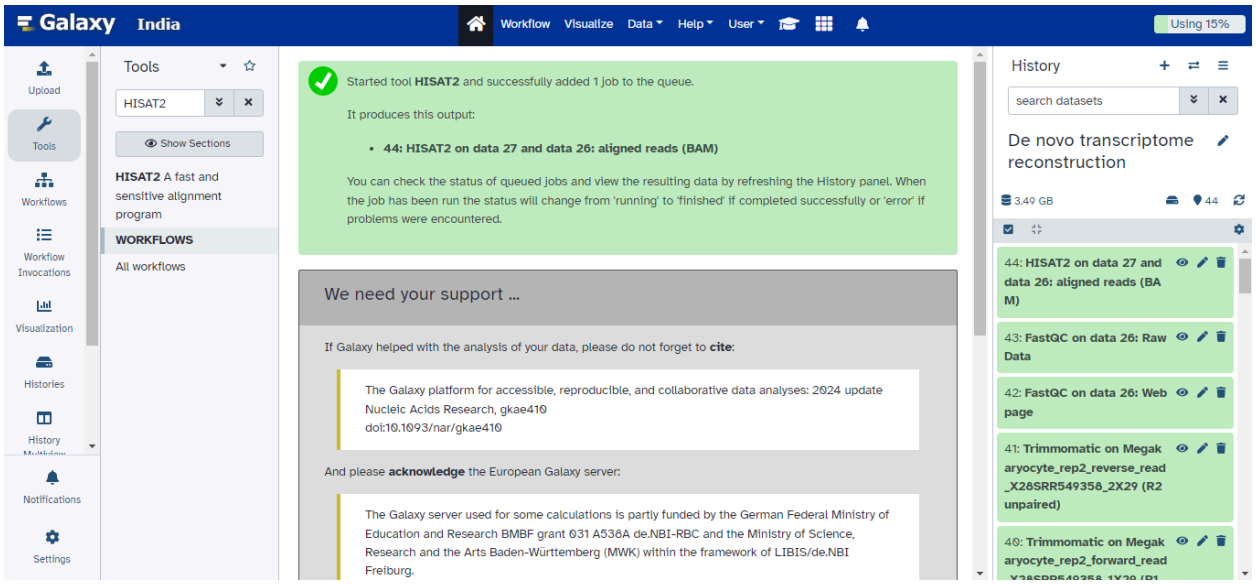
To make sense of the reads, their positions within the mouse genome must be determined. This process is known as **aligning** or '**mapping**' the reads to the reference genome.

Spliced mapping:

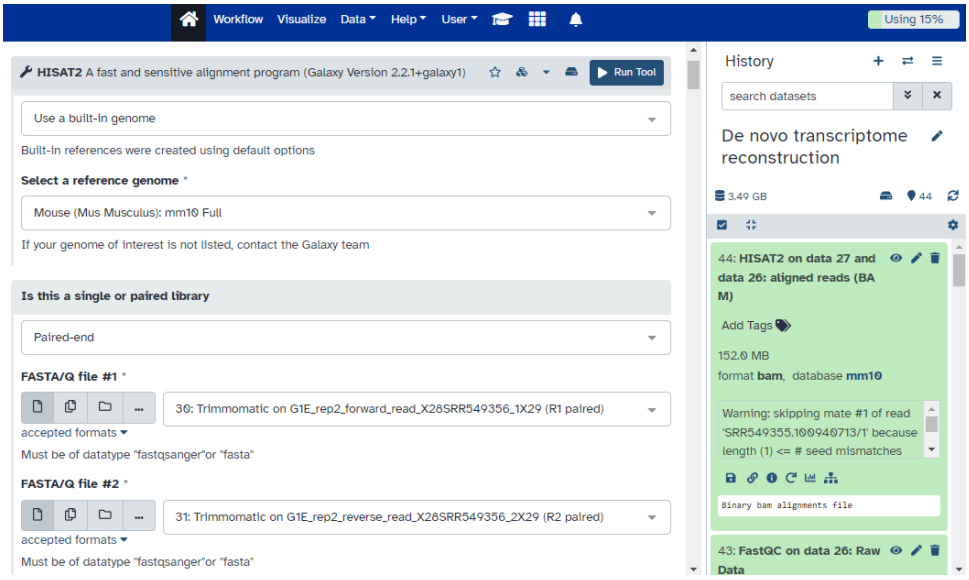
- HISAT2 tool Run `HISAT2` on one forward/reverse read pair and modify the following settings:
- HISAT2 tool: Run `HISAT2` on the remaining forward/reverse read pairs with the same parameters.

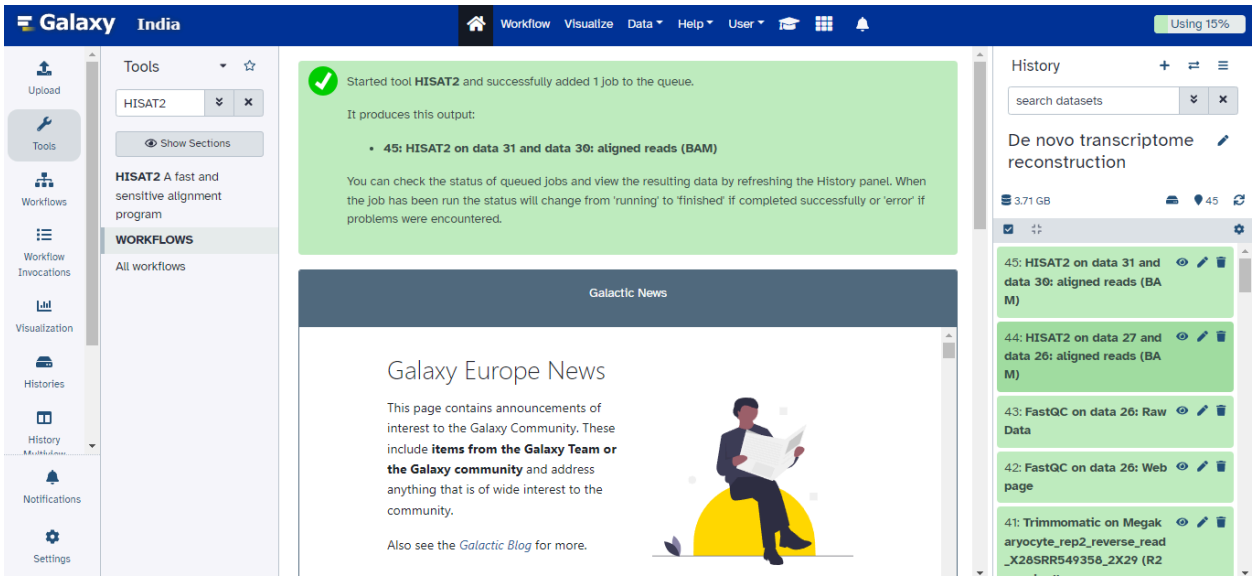
De novo transcriptome reconstruction with RNA-Seq



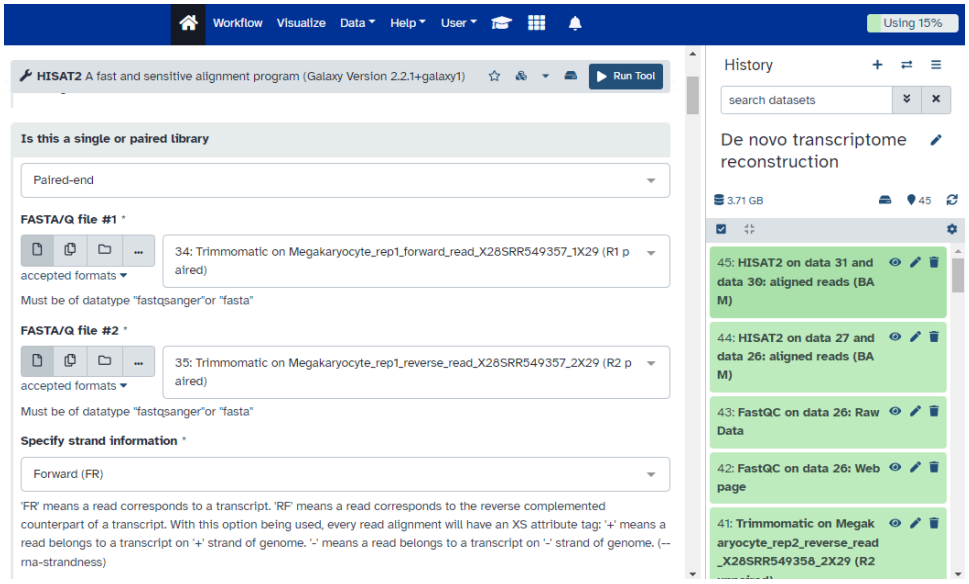


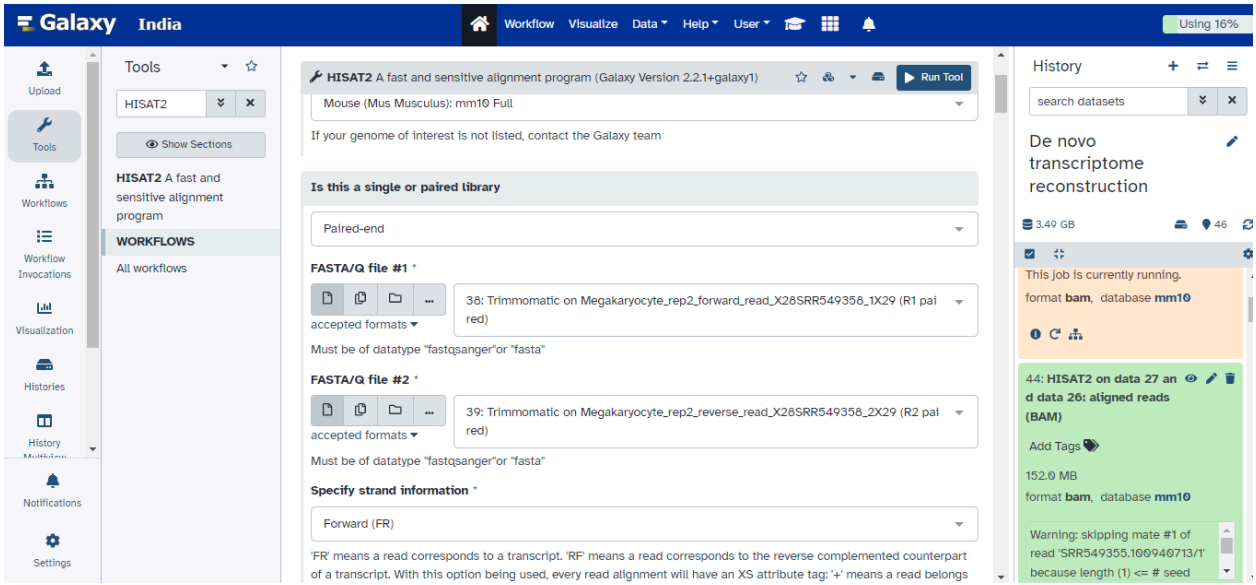
De novo transcriptome reconstruction with RNA-Seq



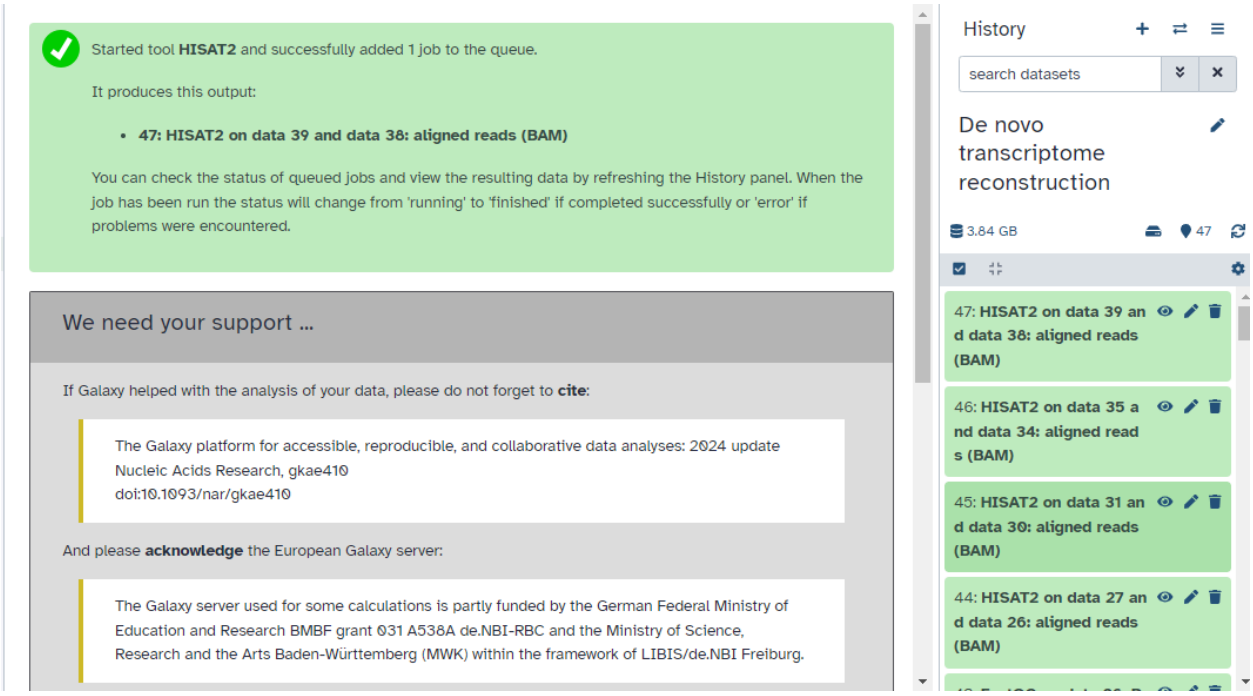


De novo transcriptome reconstruction with RNA-Seq

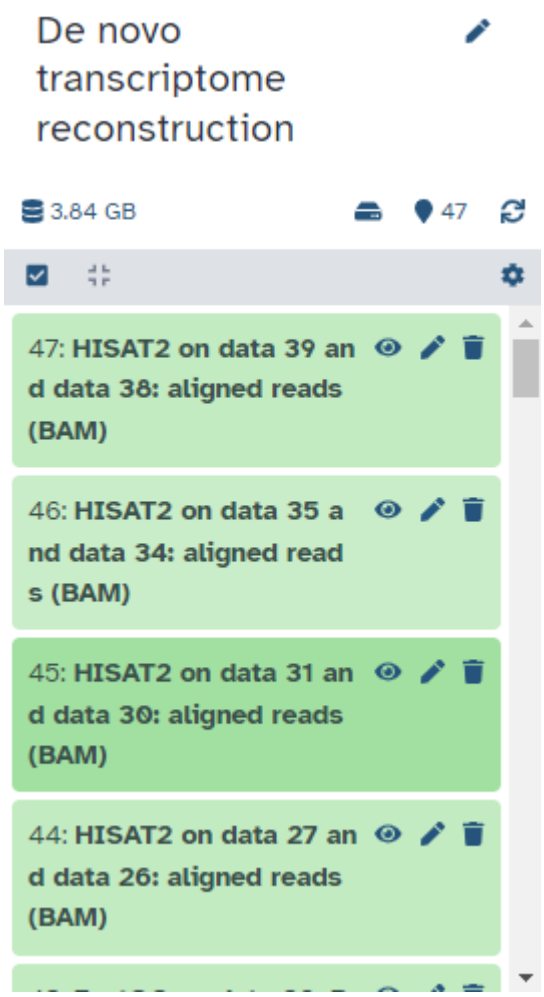




De novo transcriptome reconstruction with RNA-Seq



De novo transcriptome reconstruction with RNA-Seq



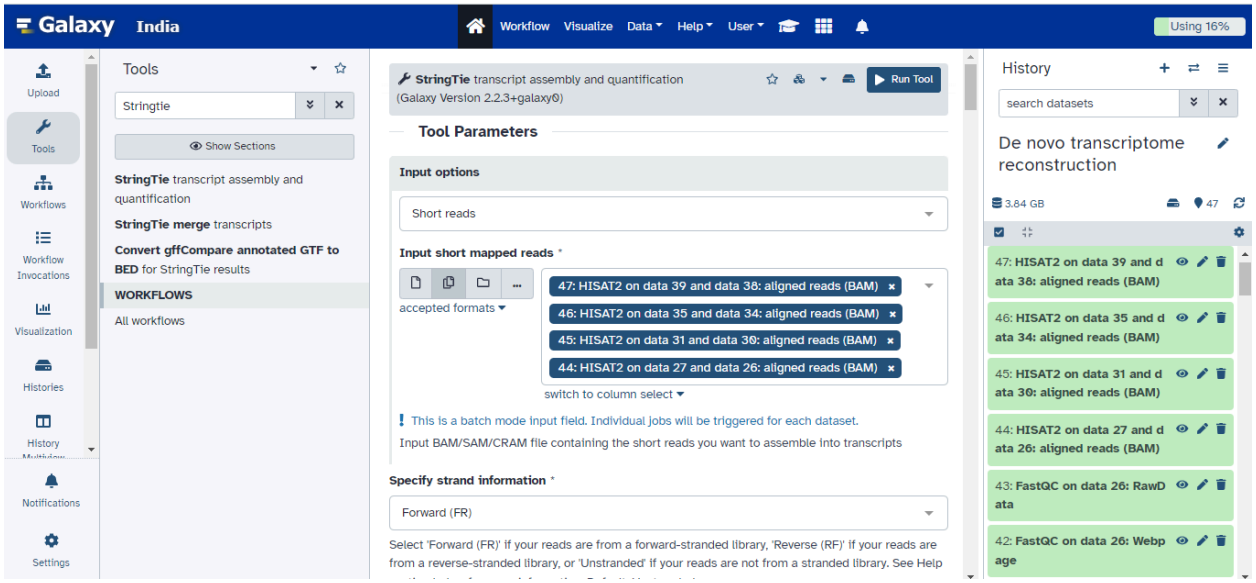
De novo transcriptome reconstruction with RNA-Seq

De novo transcript reconstruction:

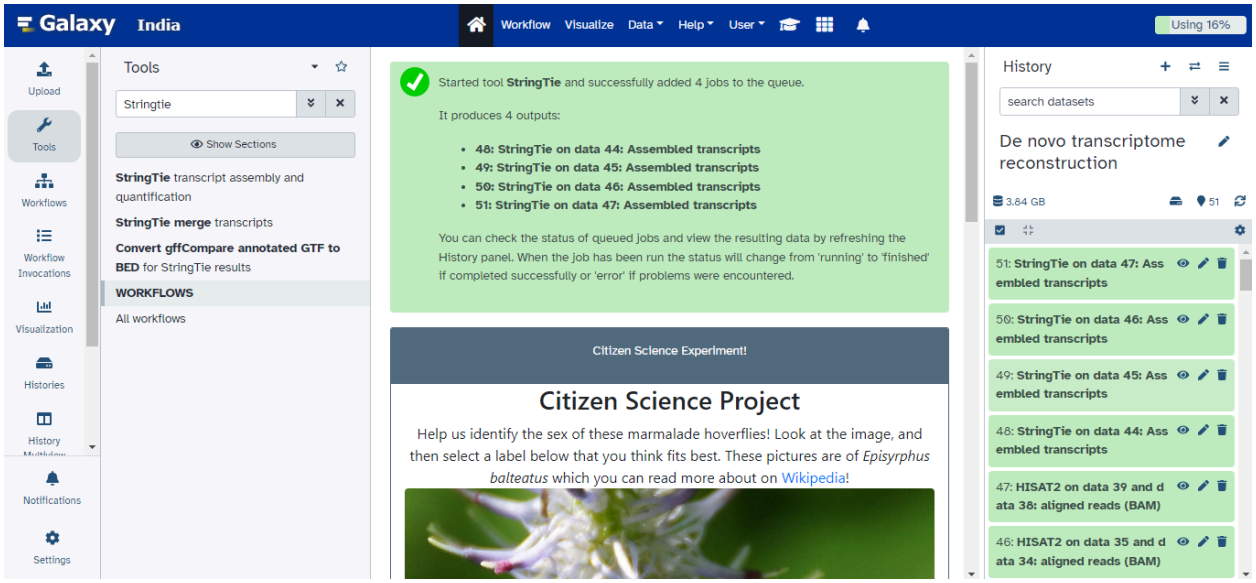
After aligning reads with HISAT2, apply **StringTie** to reconstruct the transcriptome and identify all transcripts, including novel genes and isoforms. StringTie uses the aligned BAM files to predict transcript structures and generates a GTF file. You can then run **GFFCompare** to match these predicted 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.



De novo transcriptome reconstruction with RNA-Seq



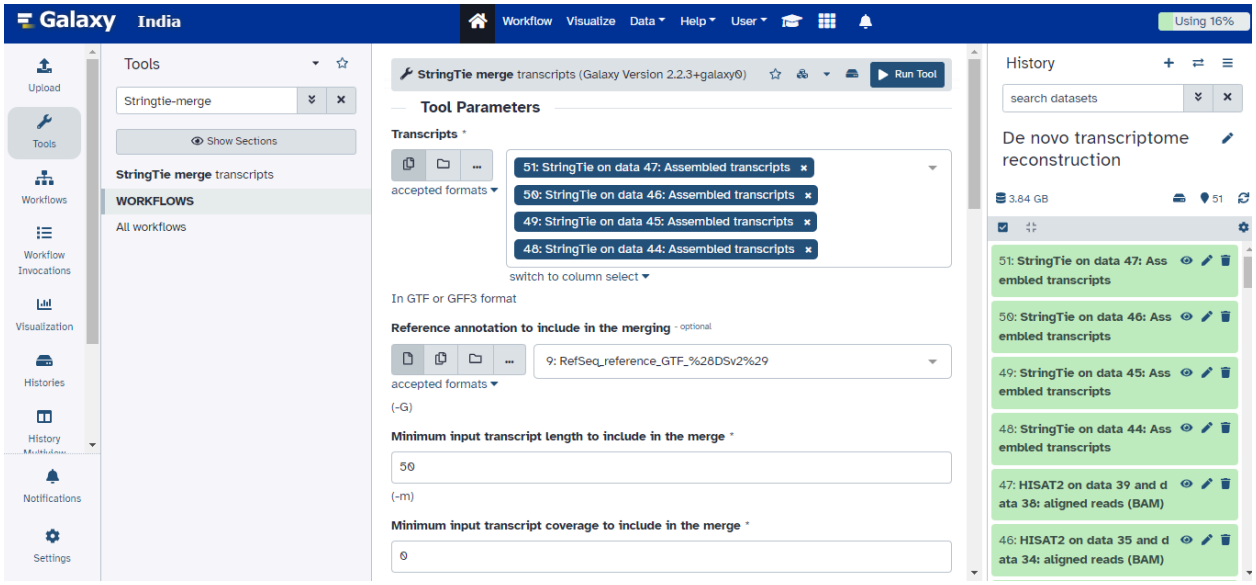
Transcriptome assembly:

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.

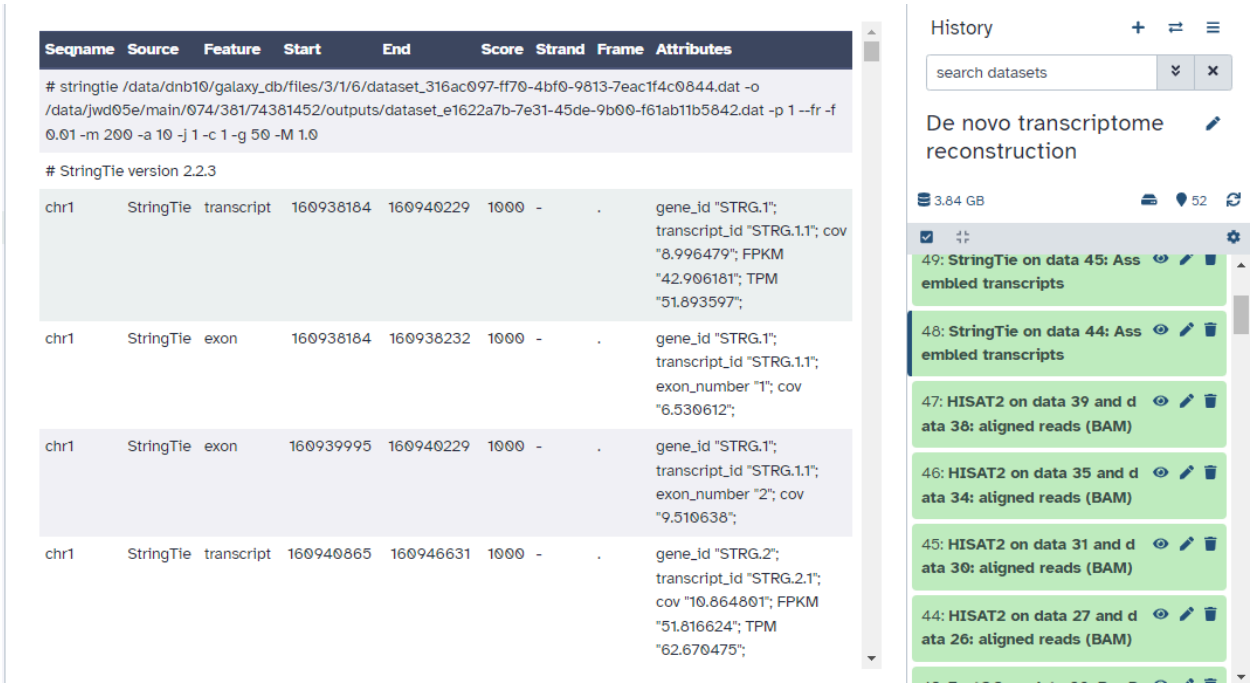
De novo transcriptome reconstruction with RNA-Seq

Transcriptome assembly:

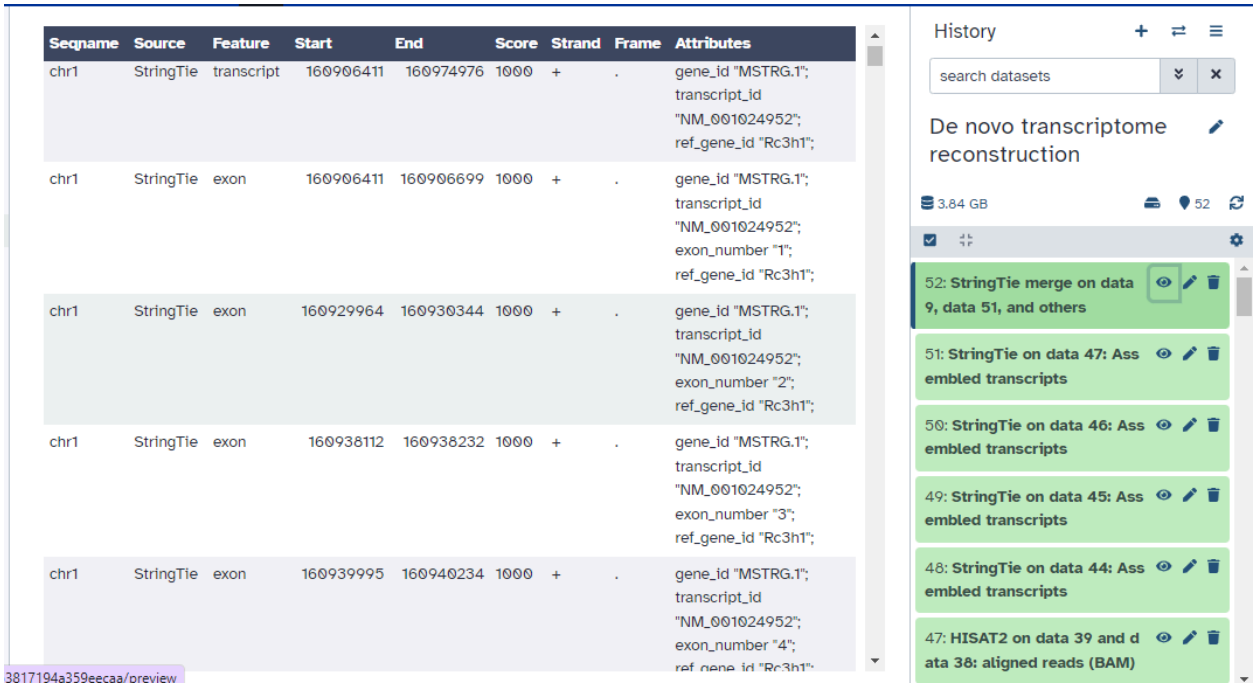
- Stringtie-merge tool Run `Stringtie-merge` on the `Stringtie` assembled transcripts along with the RefSeq annotation file we imported earlier.
- GFFCompare tool Run `GFFCompare` on the `Stringtie-merge` generated transcriptome along with the RefSeq annotation file.



De novo transcriptome reconstruction with RNA-Seq

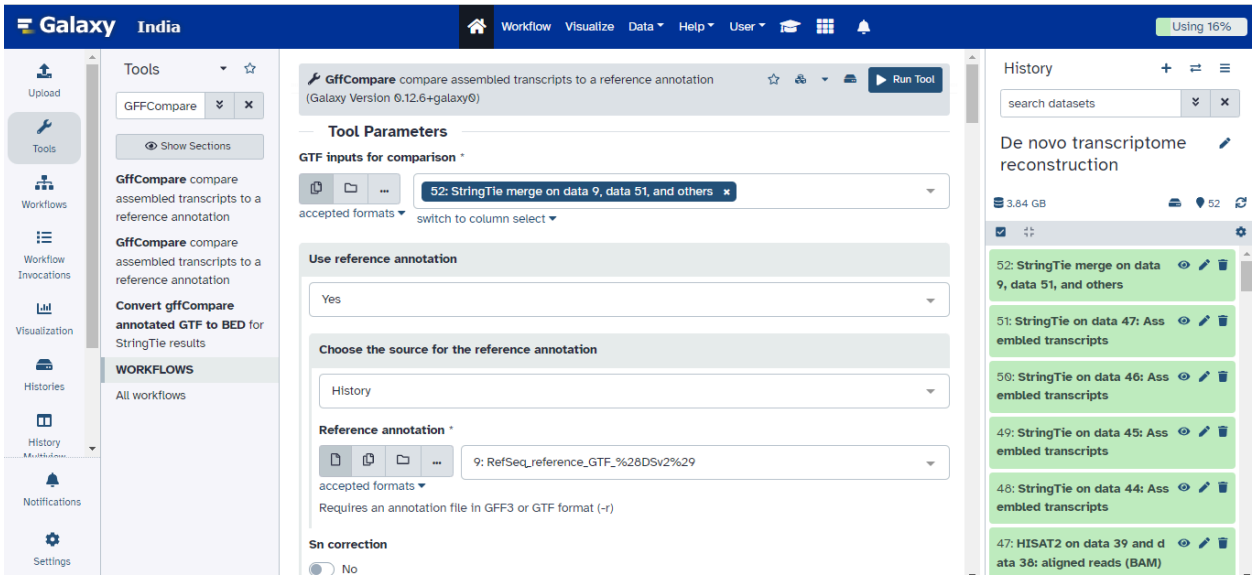


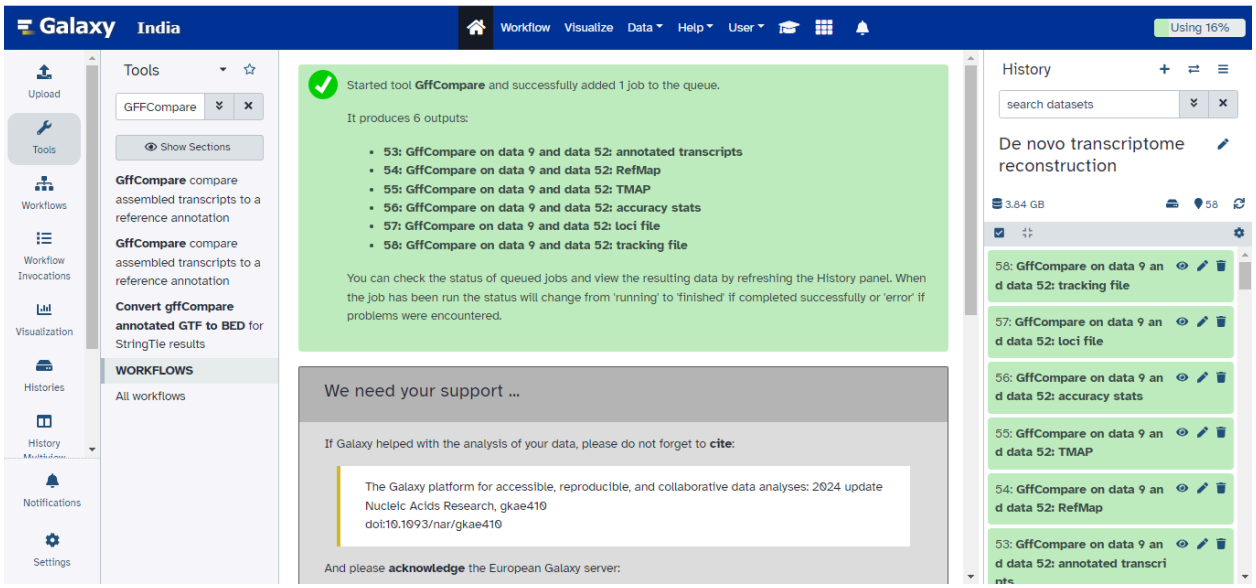
Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
# stringtie /data/dnb10/galaxy_db/files/3/1/6/dataset_316ac097-ff70-4bf0-9813-7eac1f4c0844.dat -o /data/jwd05e/main/074/381/74381452/outputs/dataset_e1622a7b-7e31-45de-9b00-f61ab11b5842.dat -p 1 --fr -f 0.01 -m 200 -a 10 -j 1 -c 1 -g 50 -M 1.0								
# StringTie version 2.2.3								
chr1	StringTie	transcript	160938184	160940229	1000	-	.	gene_id "STRG.1"; transcript_id "STRG.1.1"; cov "8.996479"; FPKM "42.906181"; 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";



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";

De novo transcriptome reconstruction with RNA-Seq





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.

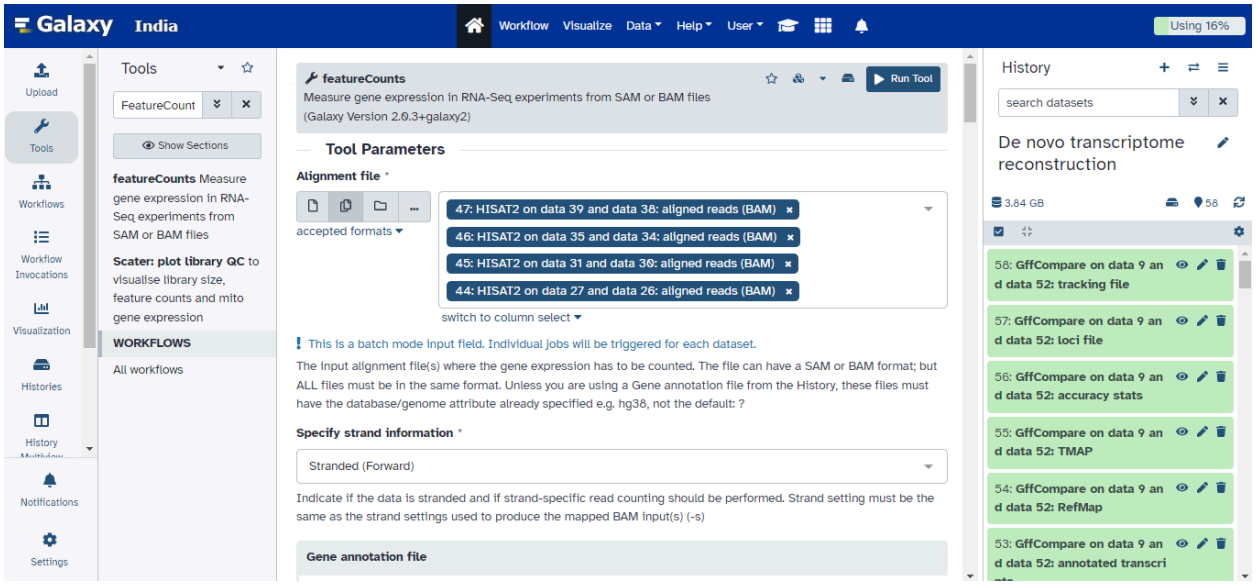
Count the number of reads per transcript:

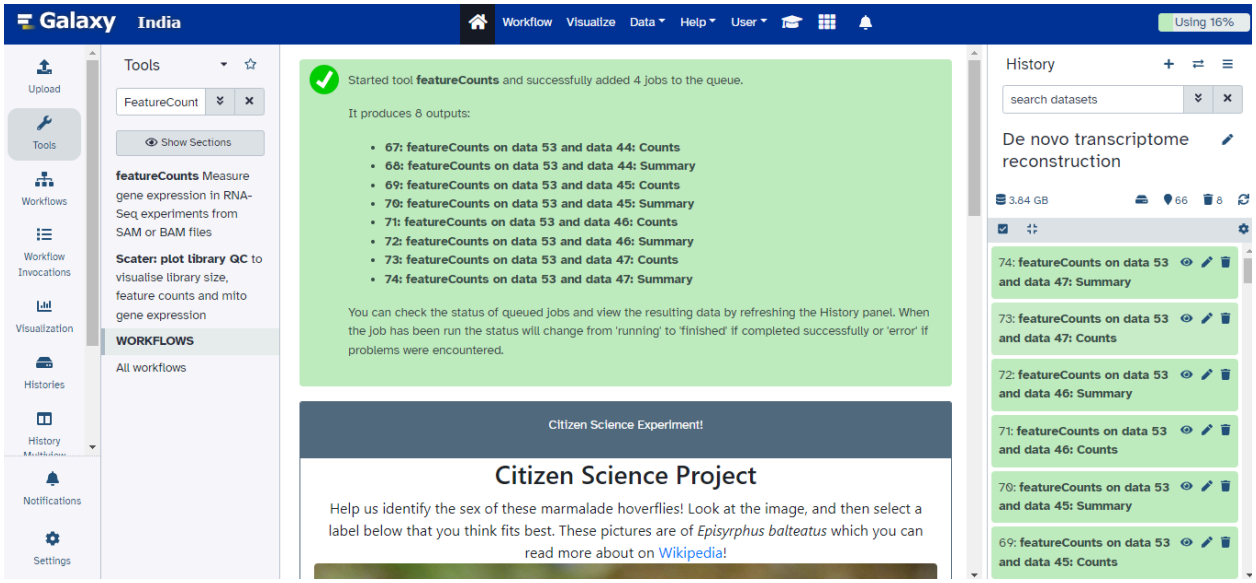
Use FeatureCounts to count reads aligning to exons in the transcriptome database generated by GFFCompare for accurate transcript abundance quantification. which counts reads that partially overlap with genomic features, while excluding reads that overlap multiple features. This method ensures precise 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





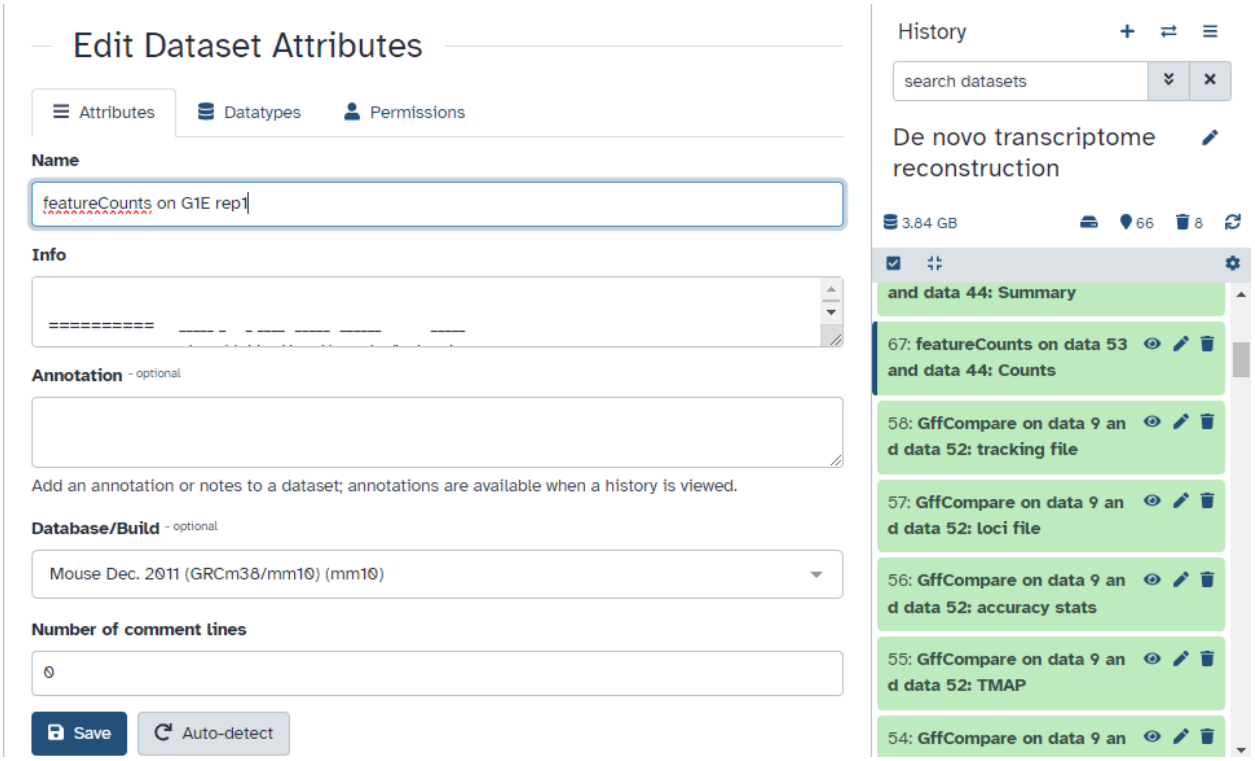
De novo transcriptome reconstruction with RNA-Seq

Perform differential gene expression testing:

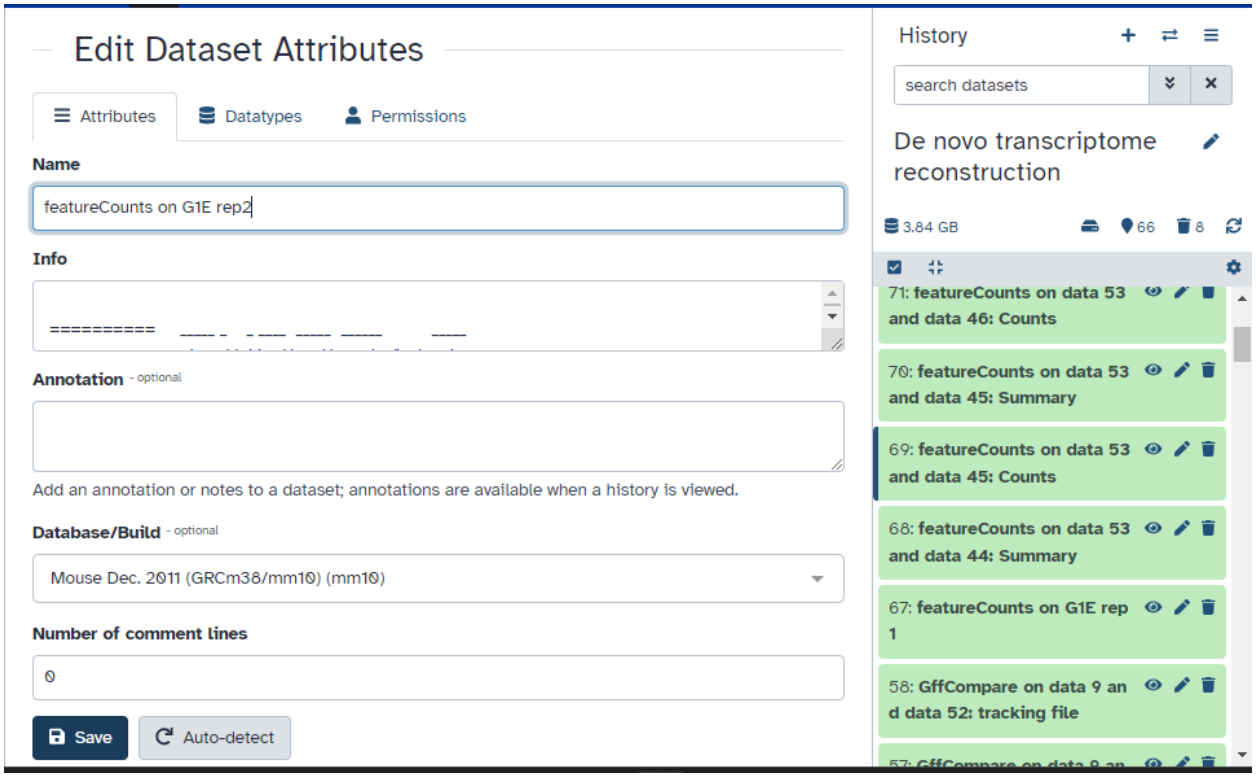
Transcript expression is determined from read counts, and at least two biological replicates are needed for reliable results. DESeq2 processes the read counts obtained from FeatureCounts by calculating the geometric mean for each gene across all samples. It then divides each count by this mean and uses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

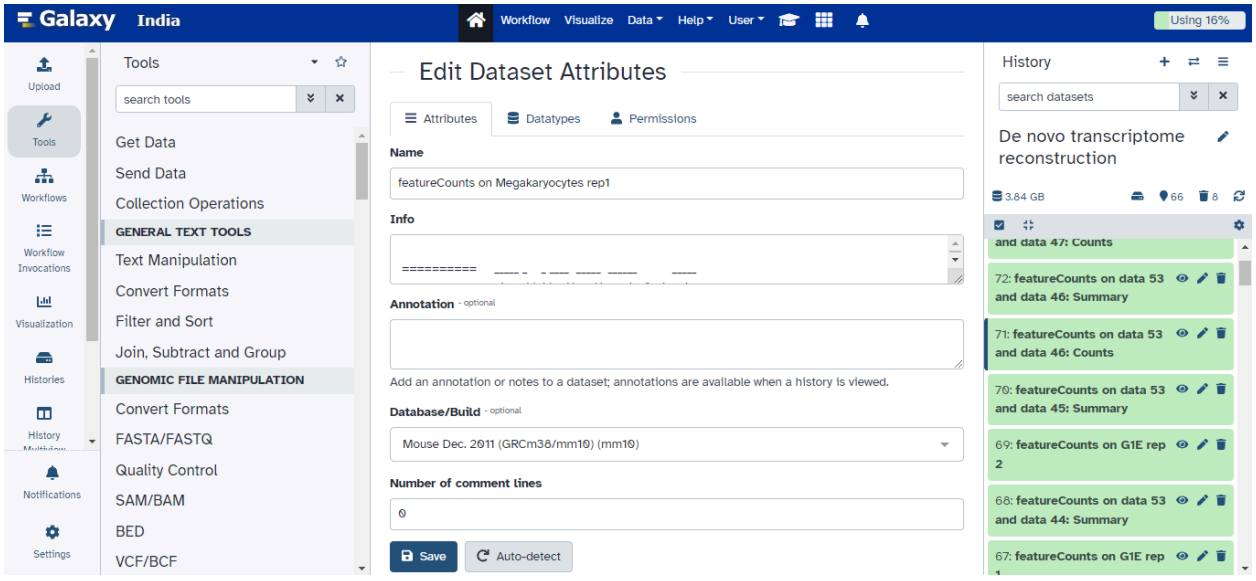
- Computation for each gene of the geometric mean of read counts across all samples
- Division of every gene count by the geometric mean
- Use of the median of these ratios as sample's size factor for normalization
- DESeq2 tool Run `DESeq2` with the following parameters:
- *"1: Factor"*
    - *"1: Factor 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

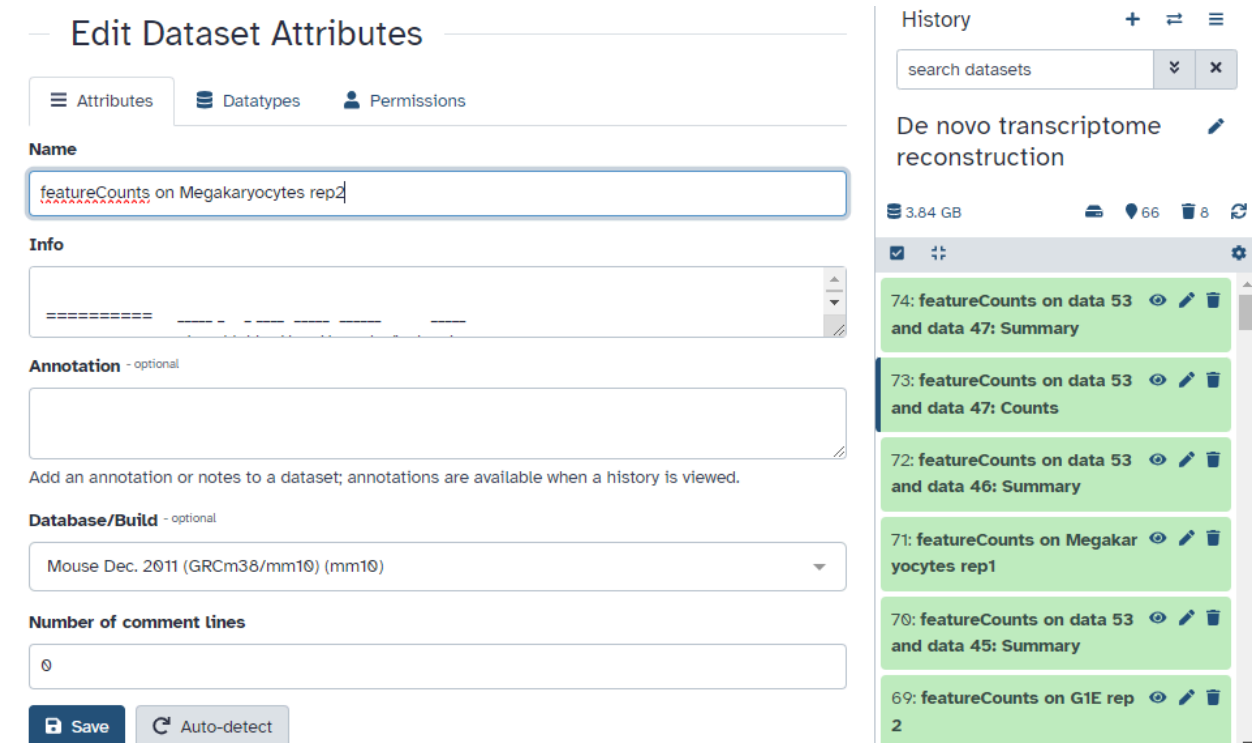


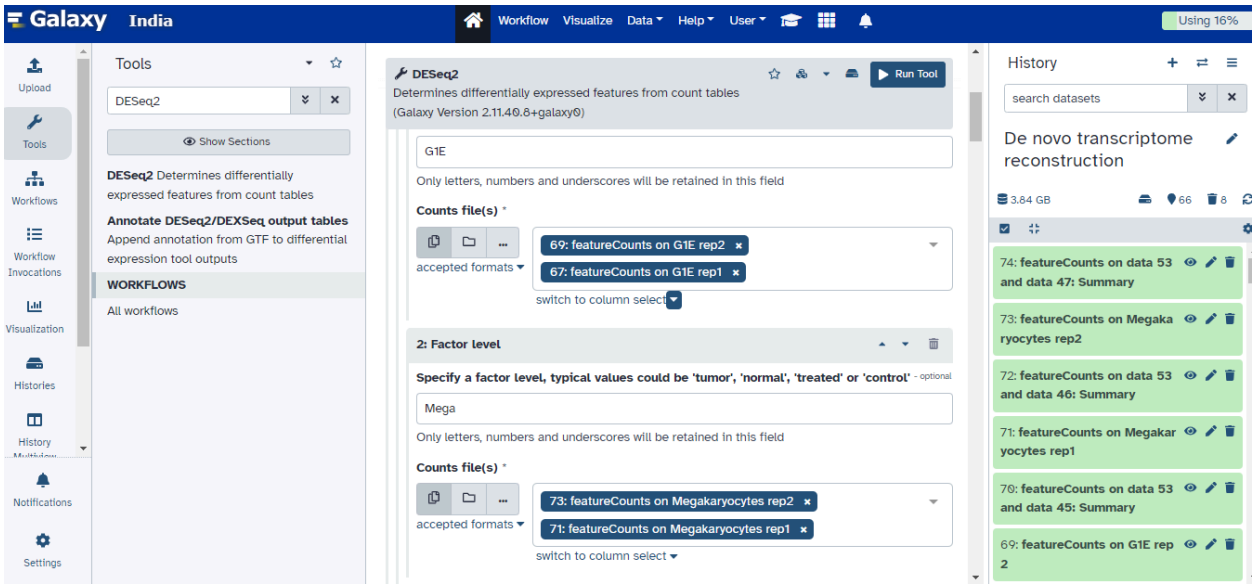
De novo transcriptome reconstruction with RNA-Seq



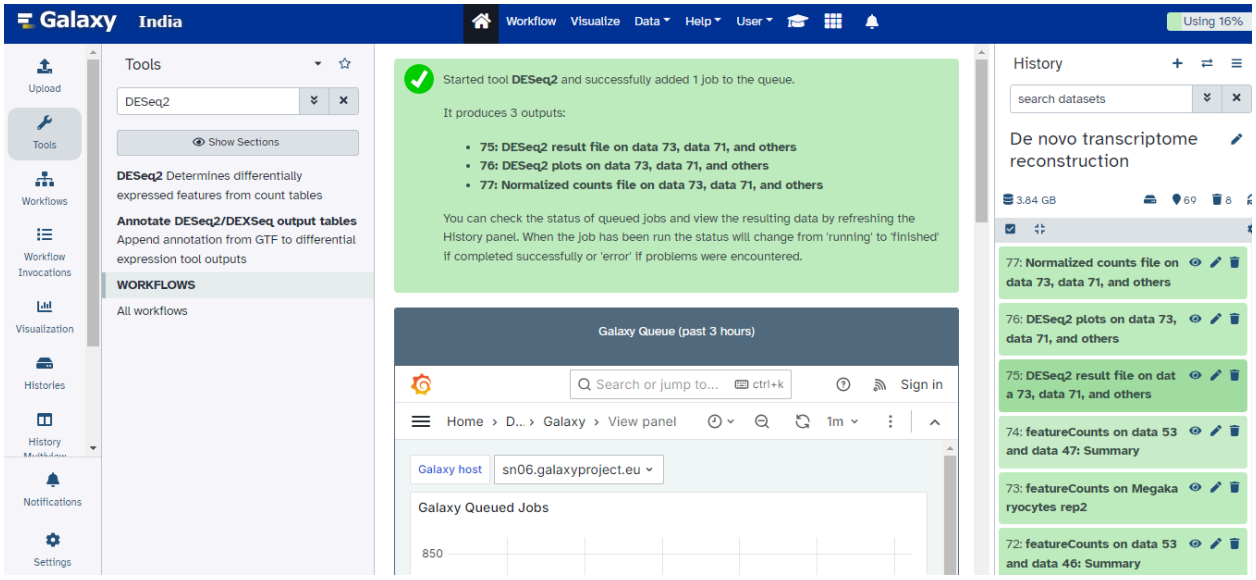


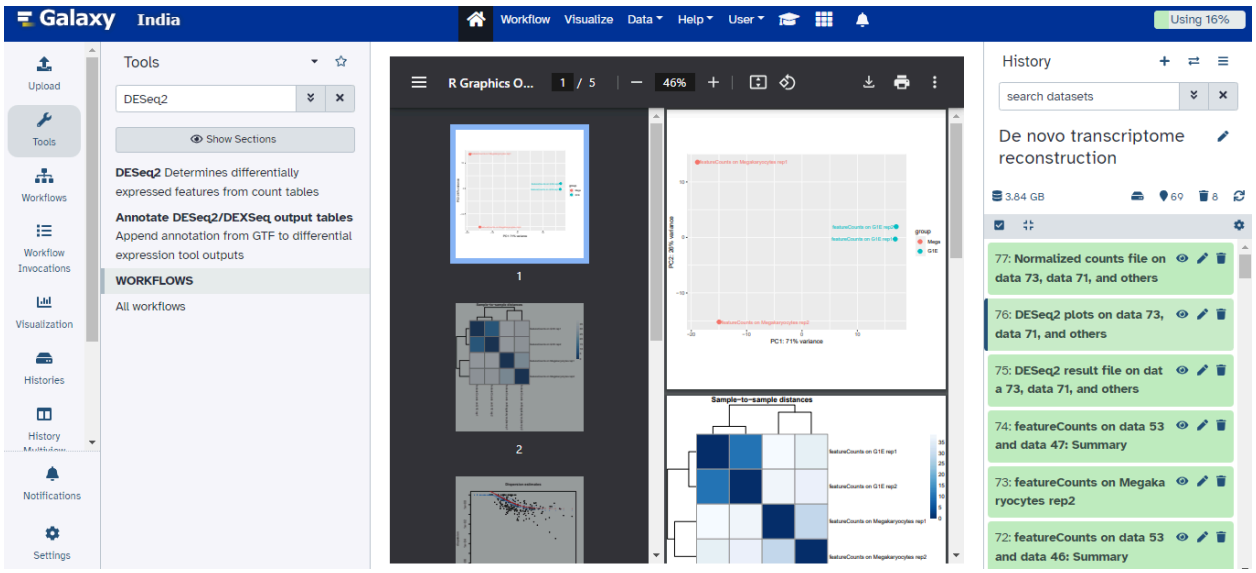
De novo transcriptome reconstruction with RNA-Seq



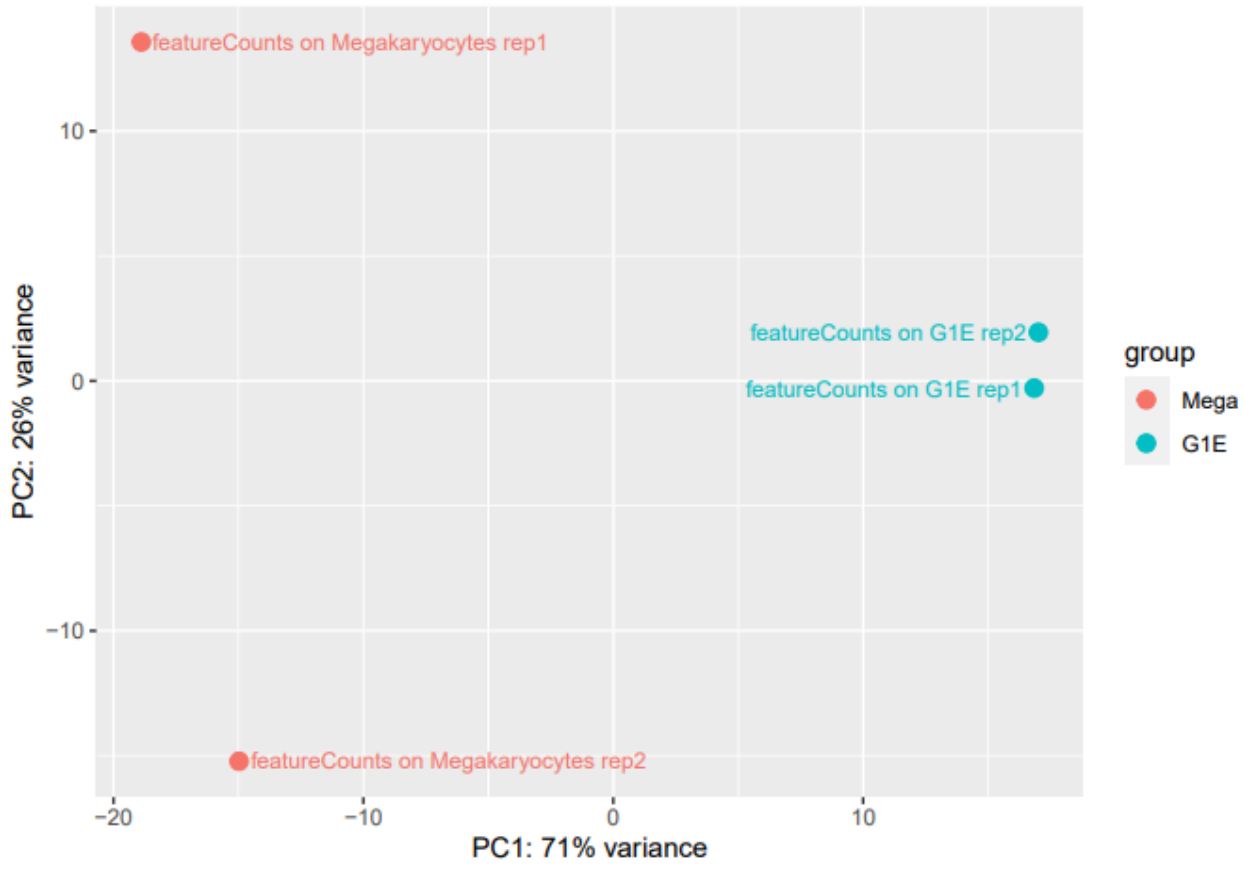


De novo transcriptome reconstruction with RNA-Seq

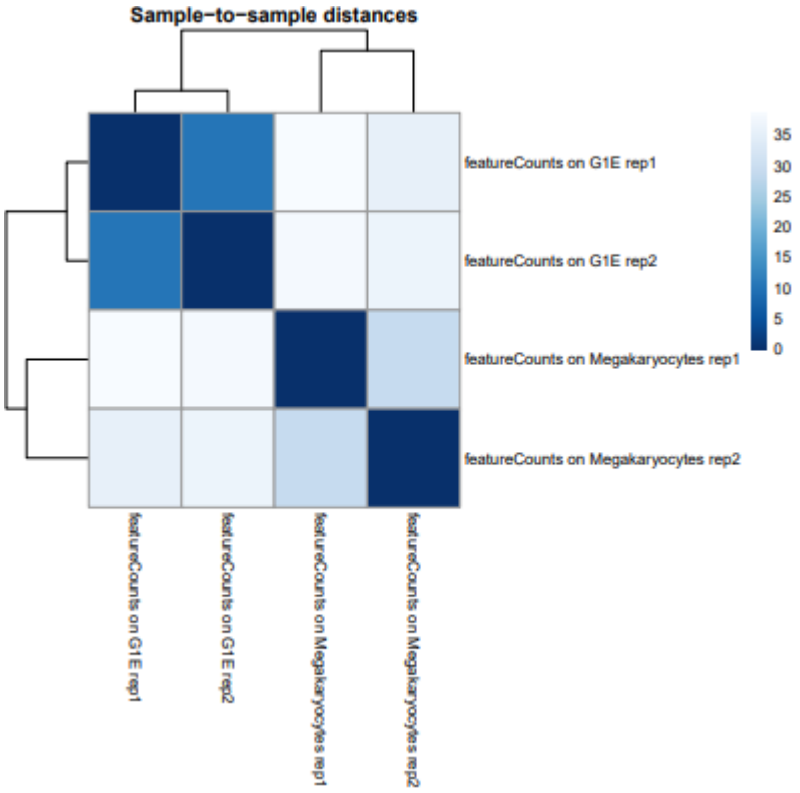




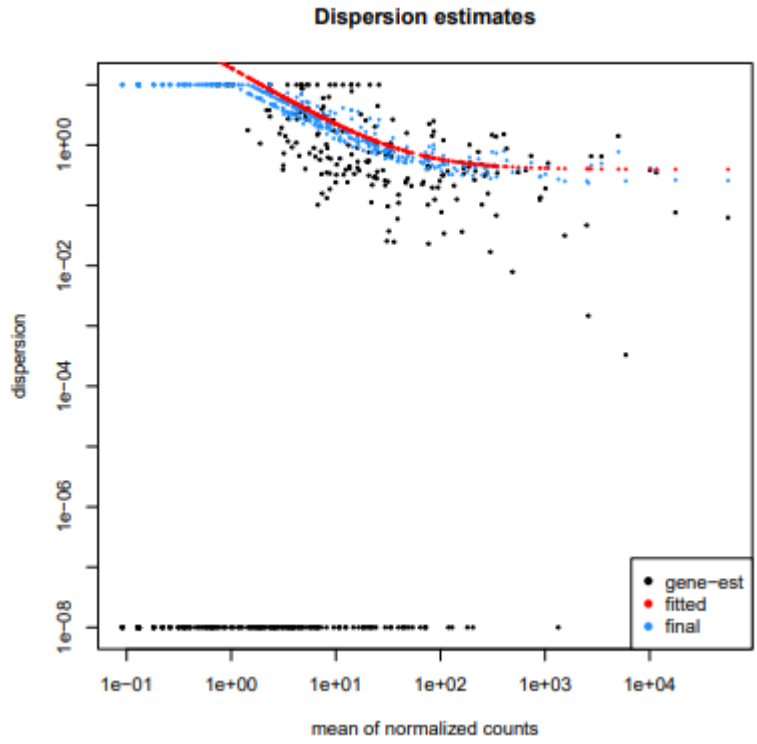
De novo transcriptome reconstruction with RNA-Seq



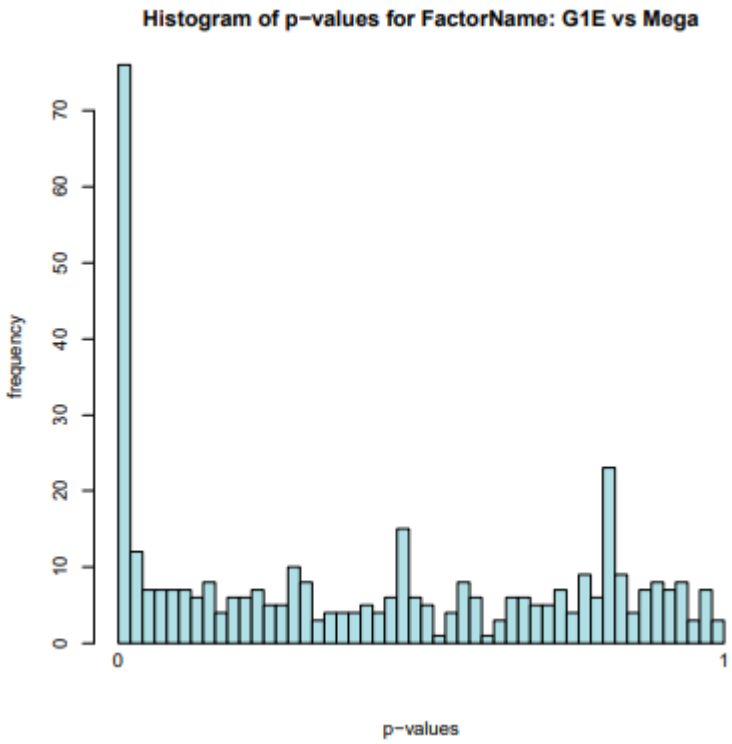
De novo transcriptome reconstruction with RNA-Seq



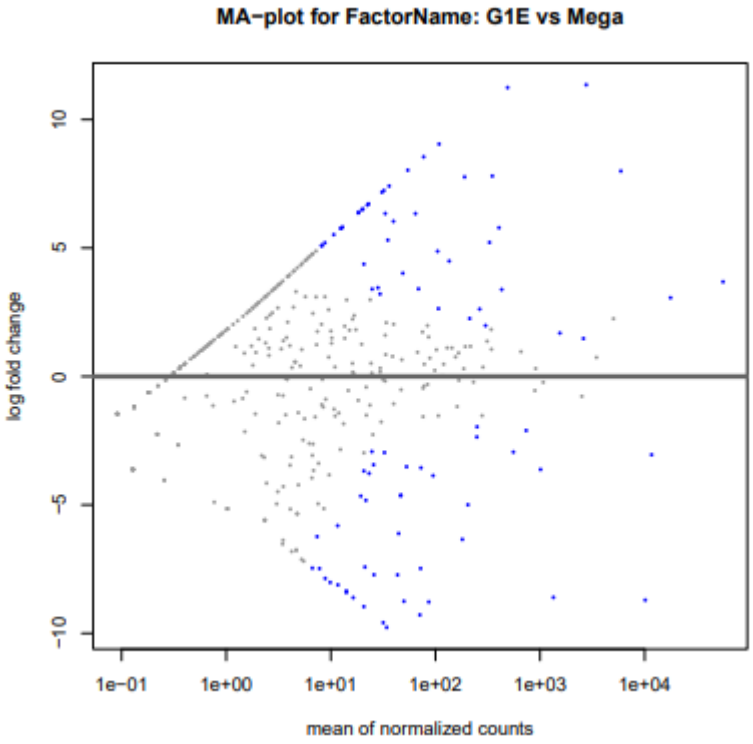
De novo transcriptome reconstruction with RNA-Seq



De novo transcriptome reconstruction with RNA-Seq

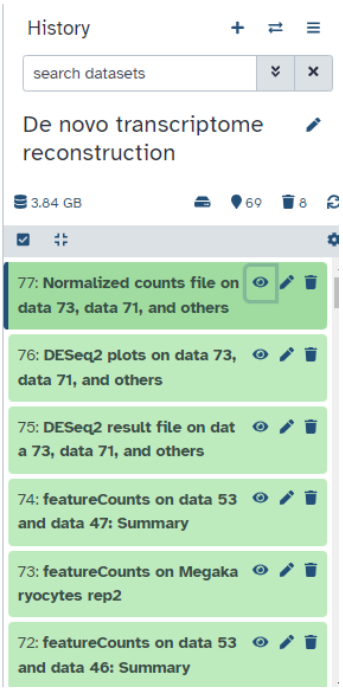


De novo transcriptome reconstruction with RNA-Seq



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	18.5668523097742
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

9d359a2b16edb75/preview



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	2.830263247
MSTRG.32.1	5919.00826748956	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	-6.34383274393348	0.932041908456529	-6.80638143668768	1.000841396
MSTRG.80.1	349.659601824575	7.80173083071458	1.19203314001101	6.54489423896596	5.953750433
MSTRG.162.1	204.511784877632	-4.99657797405334	0.834792516998134	-5.98541298863189	2.158414854
MSTRG.213.1	108.33466944873	9.046857817317	1.61521111199896	5.60103738149792	2.130727769
MSTRG.94.1	189.887161635607	7.76495620004759	1.440554897 10335	5.39025358607387	7.035832972
MSTRG.112.1	71.0091298952804	-9.28319044590951	1.76247397630784	-5.26713617942696	1.385684168
MSTRG.212.1	77.1320305287957	8.54648289719812	1.65018504927631	5.1791057620757	2.229520487
MSTRG.345.1	86.3330215935554	-8.77928643686929	1.721279231318	-5.10044290149652	3.388595974
NM_019932	72.0859143731627	-7.47428361215963	1.47081881048825	-5.08171698825053	3.740401851
MSTRG.216.1	56129.4800922312	3.68464107873795	0.733072786920486	5.02629635757794	5.000429443
MSTRG.308.1	44.4670196780545	-6.11496849972545	1.22335223720917	-4.99853461148322	5.776763872

History

search datase

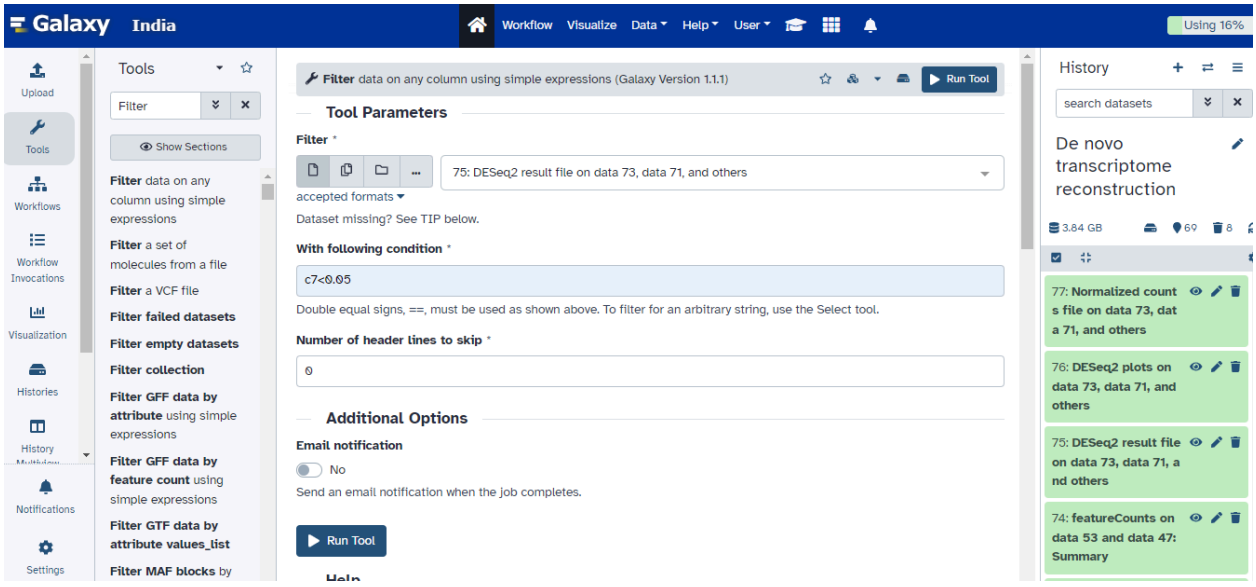
De novo transcriptome reconstruction

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

Using 16%

Upload Tools Workflows Workflow Invocations Visualization Histories History Notifications Settings

Tools

Filter

Show Sections

Filter data on any column using simple expressions

Filter a set of molecules from a file

Filter a VCF file

Filter failed datasets

Filter empty datasets

Filter collection

Filter GFF data by attribute using simple expressions

Filter GFF data by feature count using simple expressions

Filter GTF data by attribute values_list

Filter MAF blocks by

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

Run Tool

Tool Parameters

Filter *

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

accepted formats

Dataset missing? See TIP below.

With following condition *

c7<0.05

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip *

0

Additional Options

Email notification

No

Send an email notification when the job completes.

Run Tool

Help

History

search datasets

De novo transcriptome reconstruction

3.84 GB 69 8

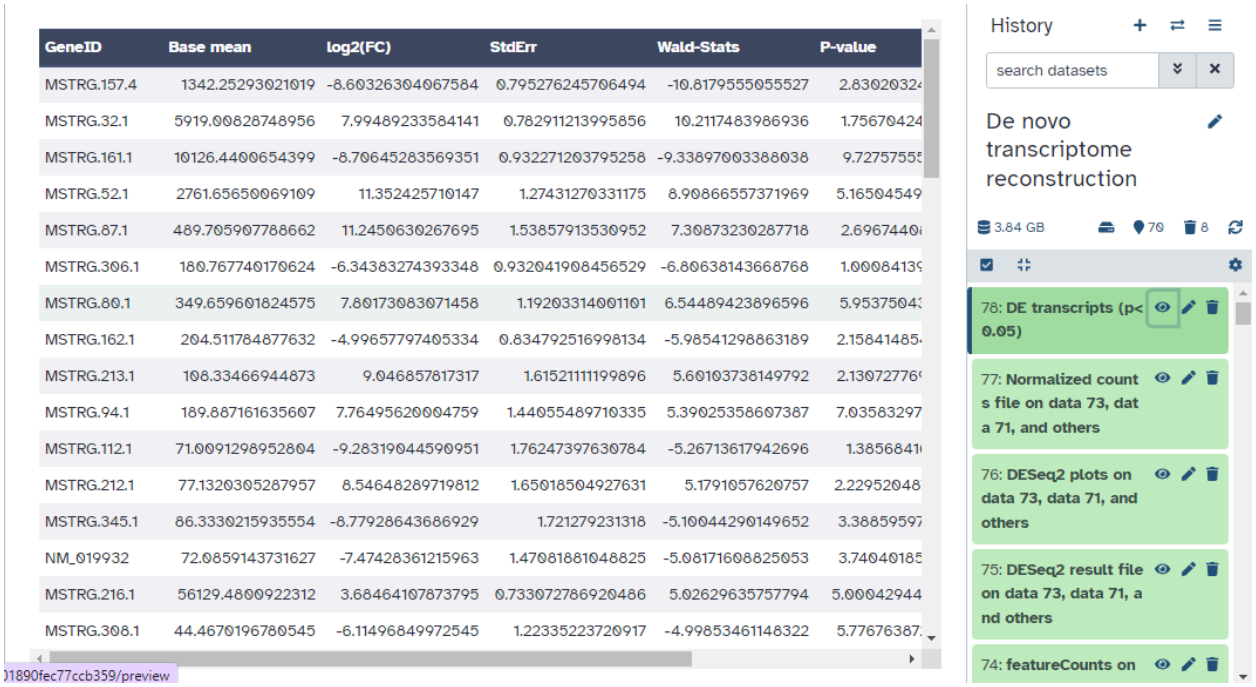
77: Normalized counts file on data 73, data 71, and others

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



GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value
MSTRG.157.4	1342.25293021019	-8.60326304067584	0.795276245706494	-10.8179555055527	2.8302032
MSTRG.32.1	5919.00828748956	7.99489233584141	0.782911213995856	10.2117483986936	1.75670424
MSTRG.161.1	10126.4400654399	-8.70645263569351	0.932271203795258	-9.33897003388038	9.7275755
MSTRG.52.1	2761.65650069109	11.352425710147	1.27431270331175	8.90866557371969	5.16504549
MSTRG.87.1	489.705907788662	11.2450630267695	1.53857913530952	7.30873230287718	2.6967446
MSTRG.396.1	180.767740170624	-6.34383274393348	0.932041908456529	-6.80638143668768	1.0008413
MSTRG.80.1	349.659601824575	7.80173083071458	1.19203314001101	6.54489423896596	5.9537504
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	189.887161635607	7.76495620004759	1.44055469710335	5.39025358607387	7.03583297
MSTRG.112.1	71.0091298952804	-9.28319044590951	1.76247397630784	-5.26713617942696	1.3856841
MSTRG.212.1	77.1320305287957	8.54648269719812	1.65018504927631	5.1791057620757	2.22952048
MSTRG.345.1	86.3339215935554	-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

)1890fec77ccb359/preview

History

search datasets

De novo transcriptome reconstruction

3.84 GB 70 8

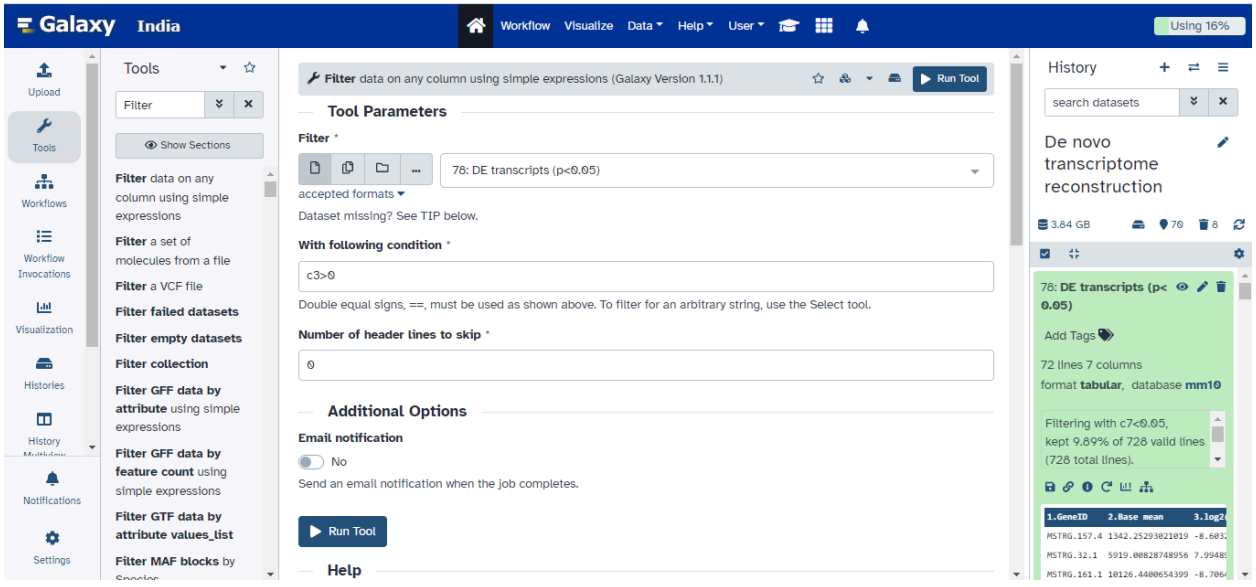
78: DE transcripts (p< 0.05)

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



Galaxy India

Workflow Visualize Data Help User

Using 16%

Upload Tools Workflows Workflow Invocations Visualization Histories History Notifications Settings

Tools

Filter

Show Sections

Filter data on any column using simple expressions

Filter a set of molecules from a file

Filter a VCF file

Filter failed datasets

Filter empty datasets

Filter collection

Filter GFF data by attribute using simple expressions

Filter GFF data by feature count using simple expressions

Filter GTF data by attribute values_list

Filter MAF blocks by

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

Run Tool

Tool Parameters

Filter *

78: DE transcripts (p<0.05)

accepted formats

Dataset missing? See TIP below.

With following condition *

c3>0

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip *

0

Additional Options

Email notification

No

Send an email notification when the job completes.

Run Tool

Help

History

search datasets

De novo transcriptome reconstruction

3.84 GB 70 8

78: DE transcripts (p< 0.05)

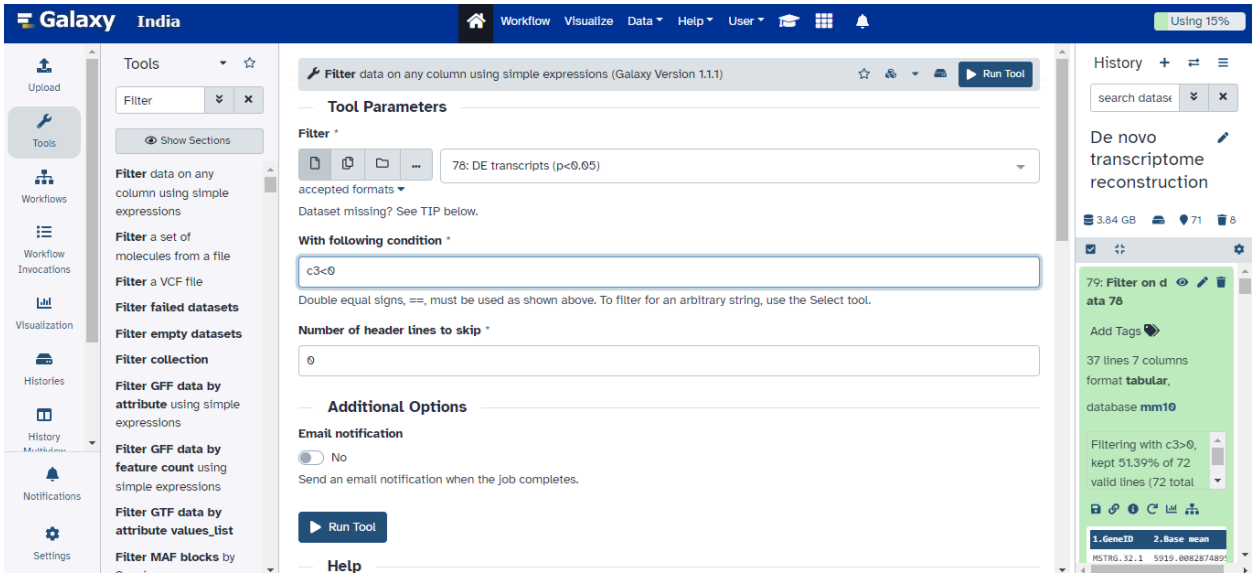
Add Tags

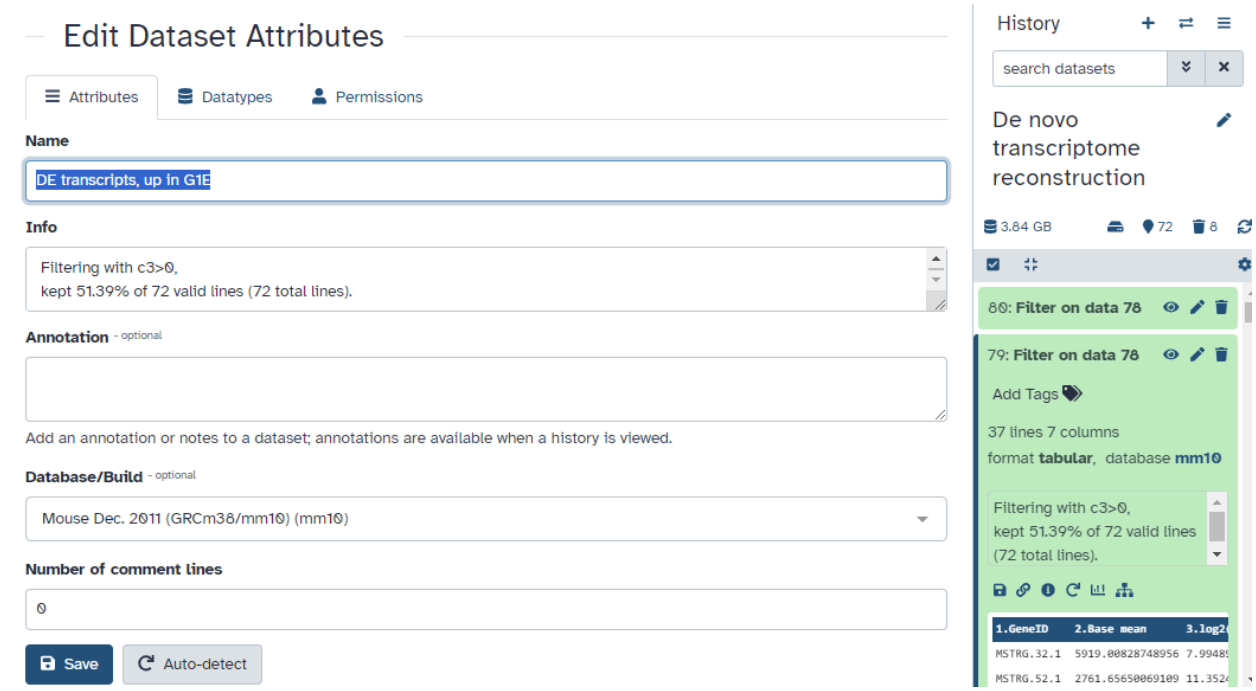
72 lines 7 columns format **tabular**, database **mm10**

Filtering with c7<0.05, kept 9.89% of 728 valid lines (728 total lines).

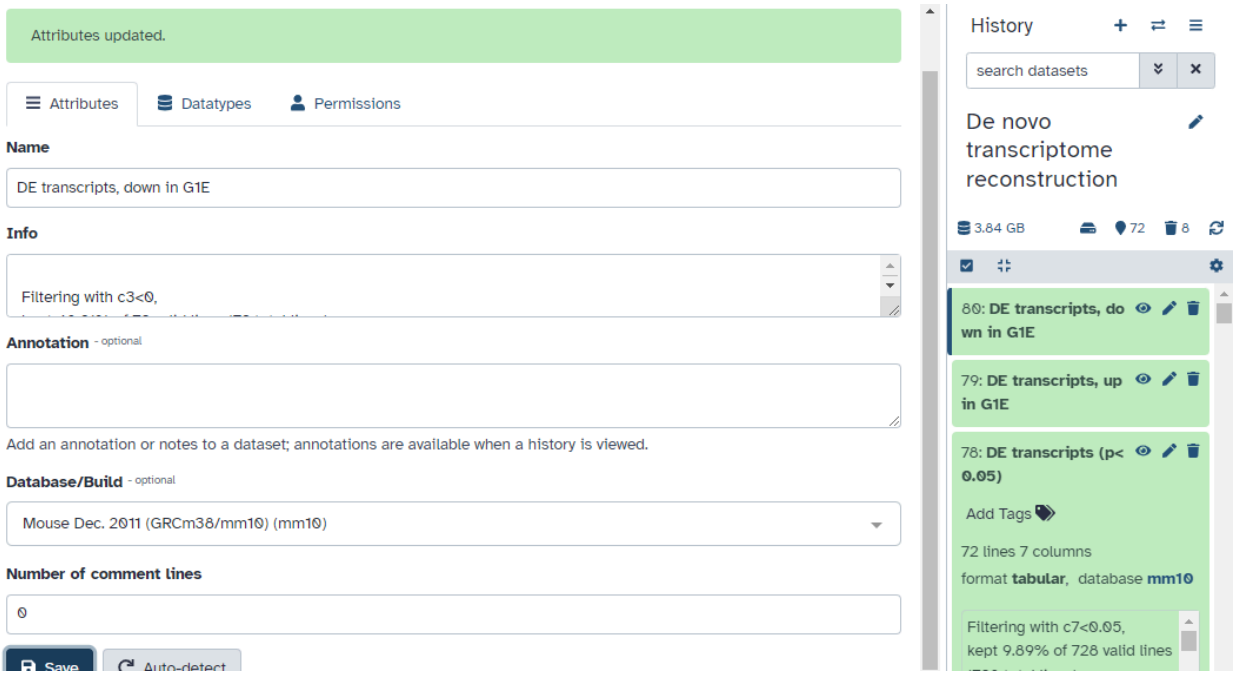
1.GeneID	2.Base mean	3.log2
MSTRG.157.4	1342.25293021019	-8.6032
MSTRG.32.1	5919.00828748956	7.9948
MSTRG.161.1	10126.4400654399	-8.706

De novo transcriptome reconstruction with RNA-Seq





De novo transcriptome reconstruction with RNA-Seq



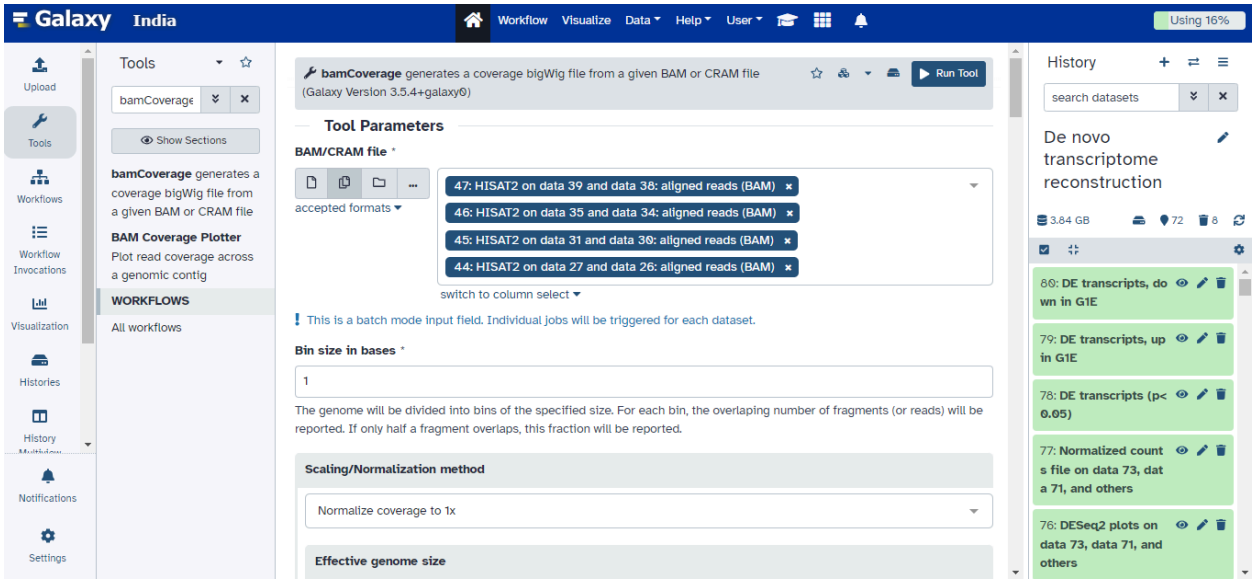
Visualization:

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.

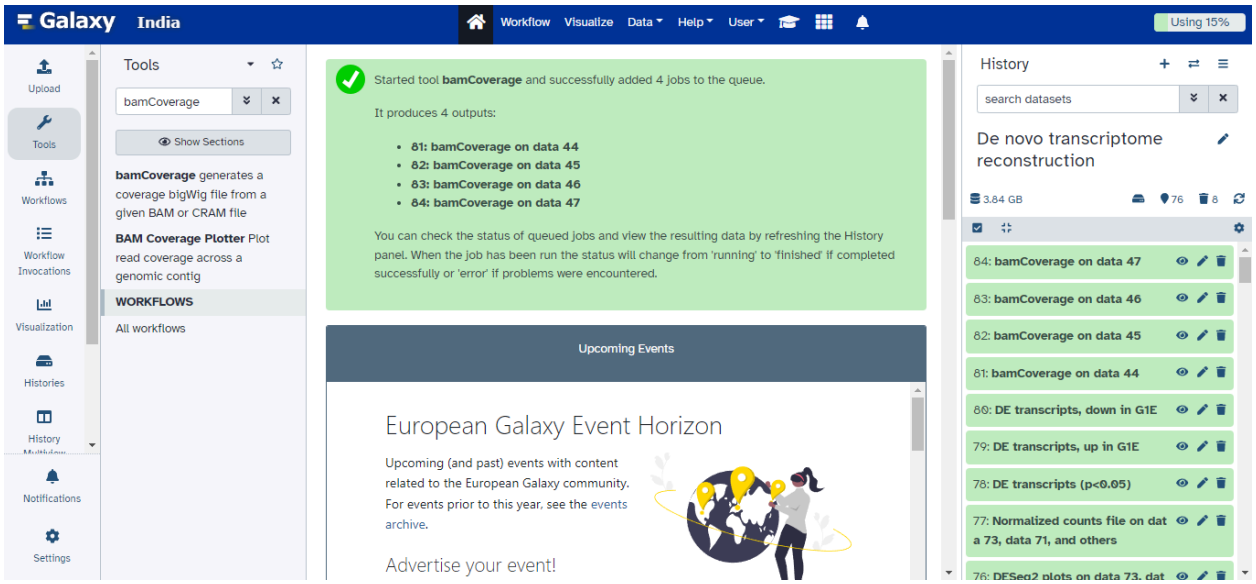
De novo transcriptome reconstruction with RNA-Seq

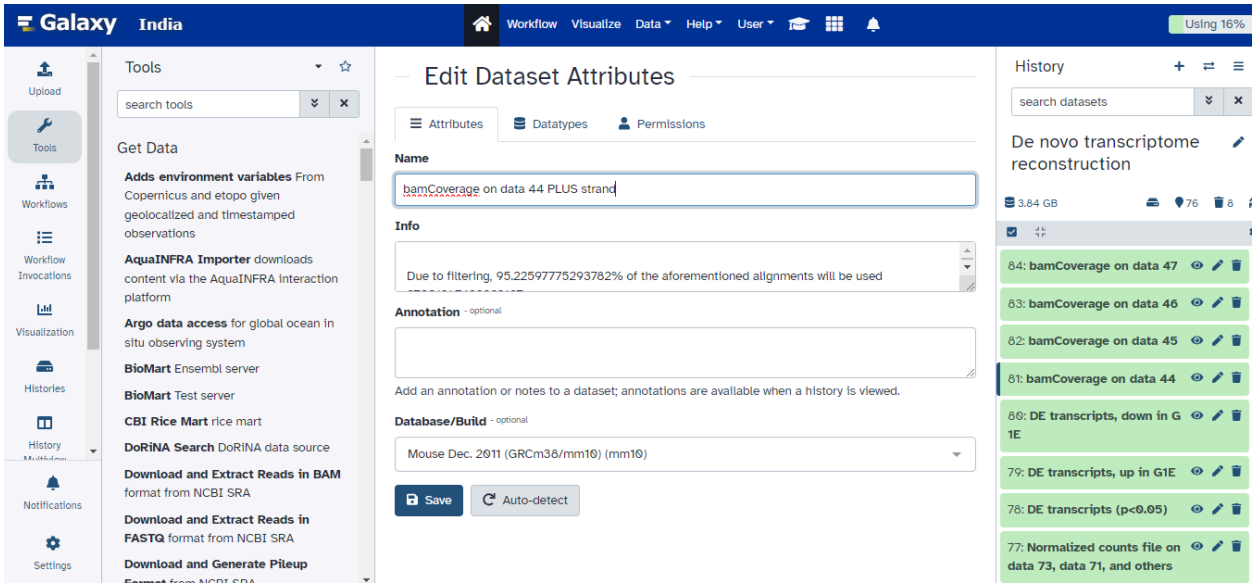
Converting aligned read files to bigWig format:

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.

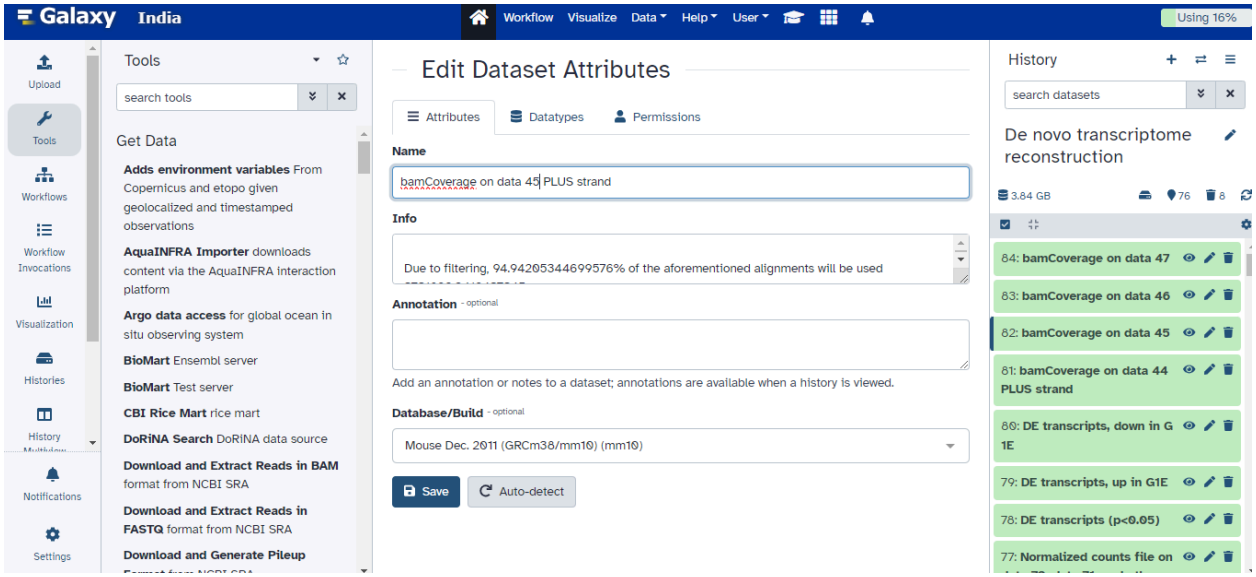


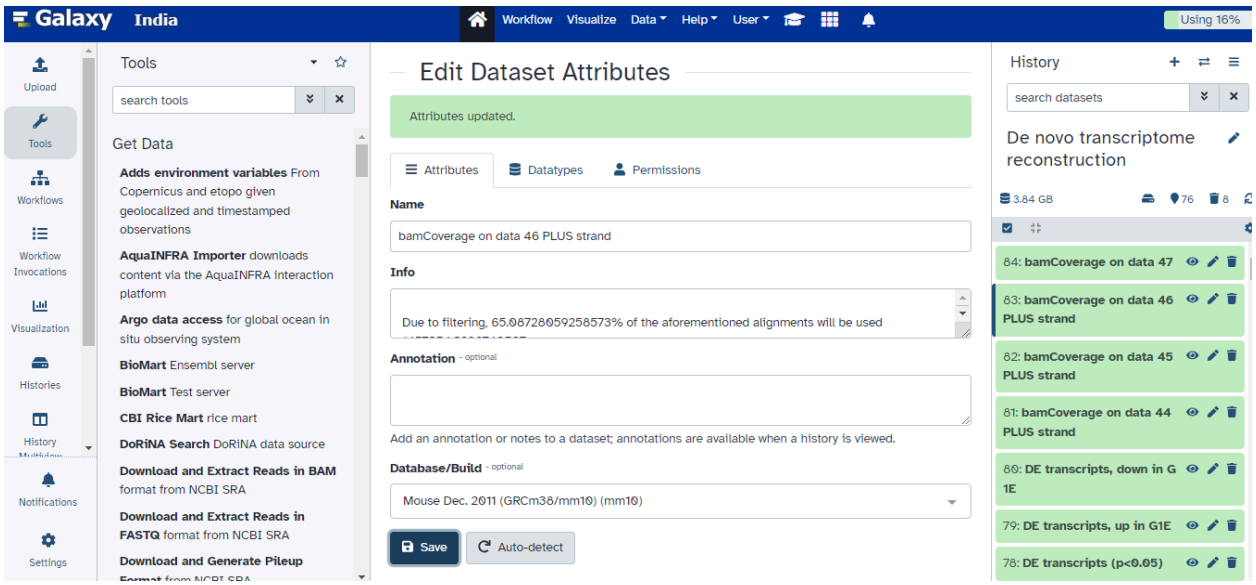
De novo transcriptome reconstruction with RNA-Seq



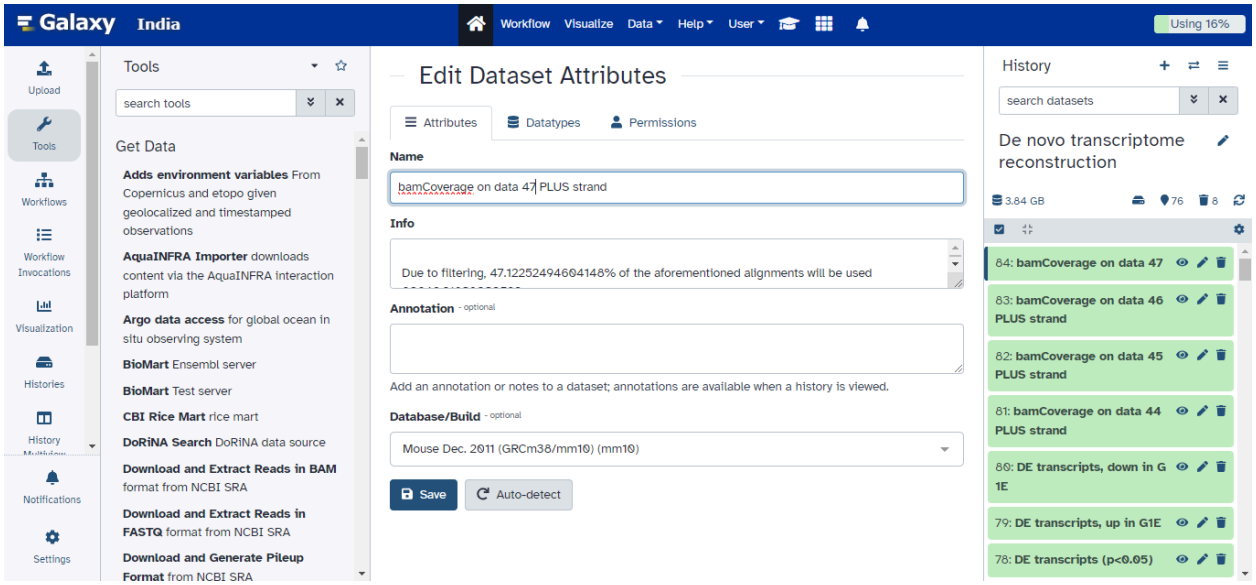


De novo transcriptome reconstruction with RNA-Seq

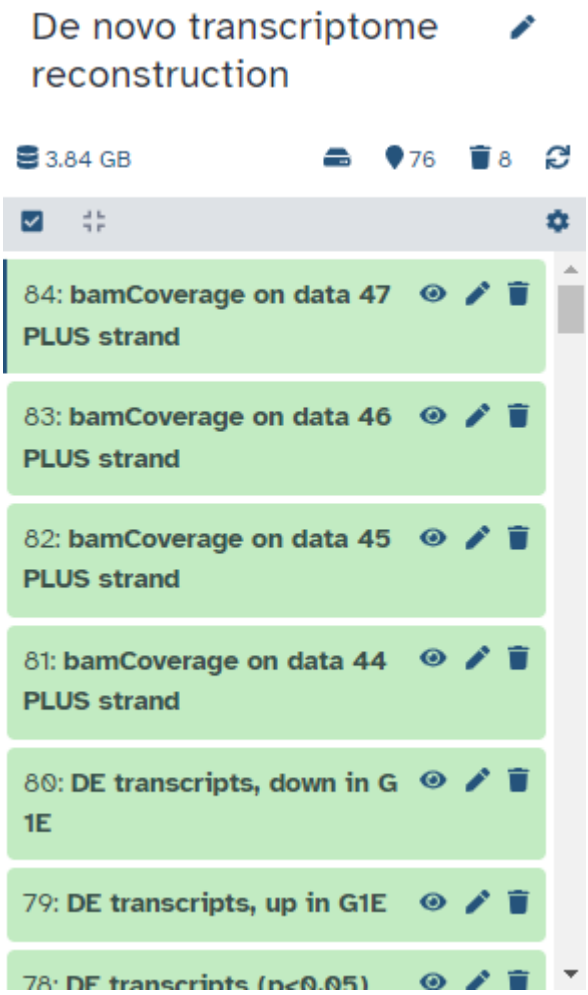




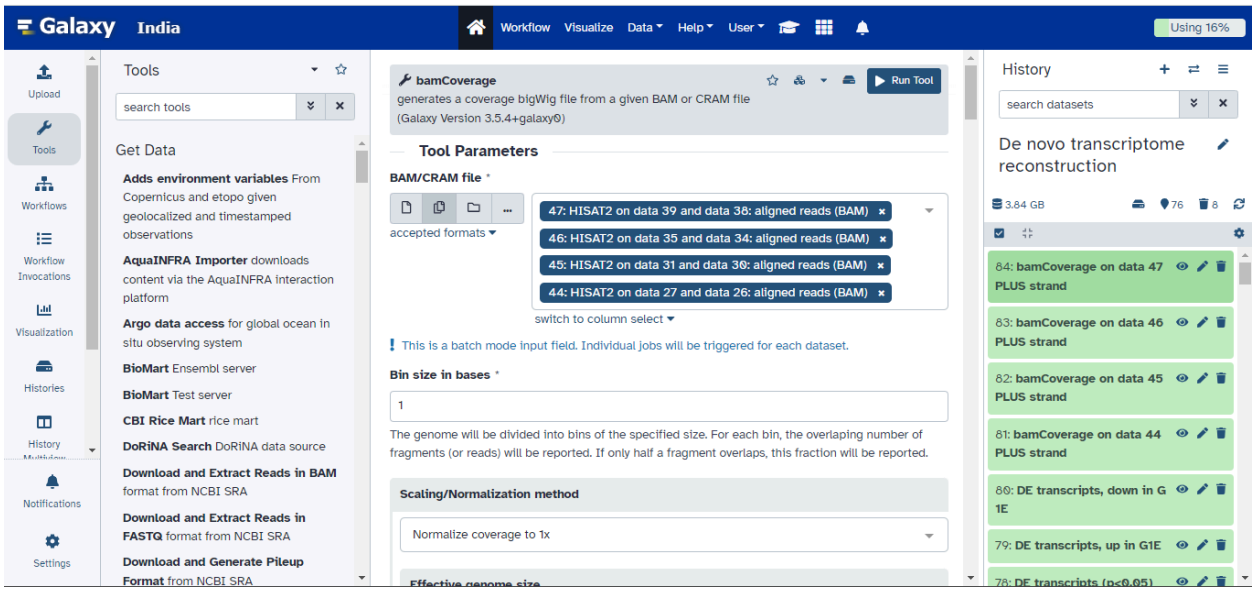
De novo transcriptome reconstruction with RNA-Seq

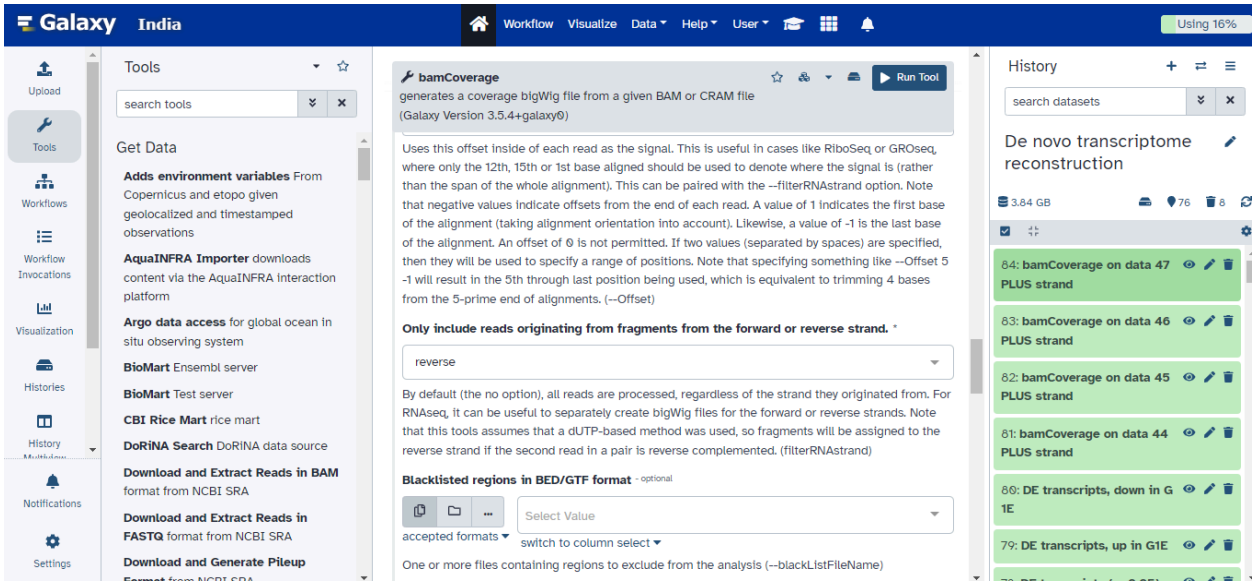


De novo transcriptome reconstruction with RNA-Seq

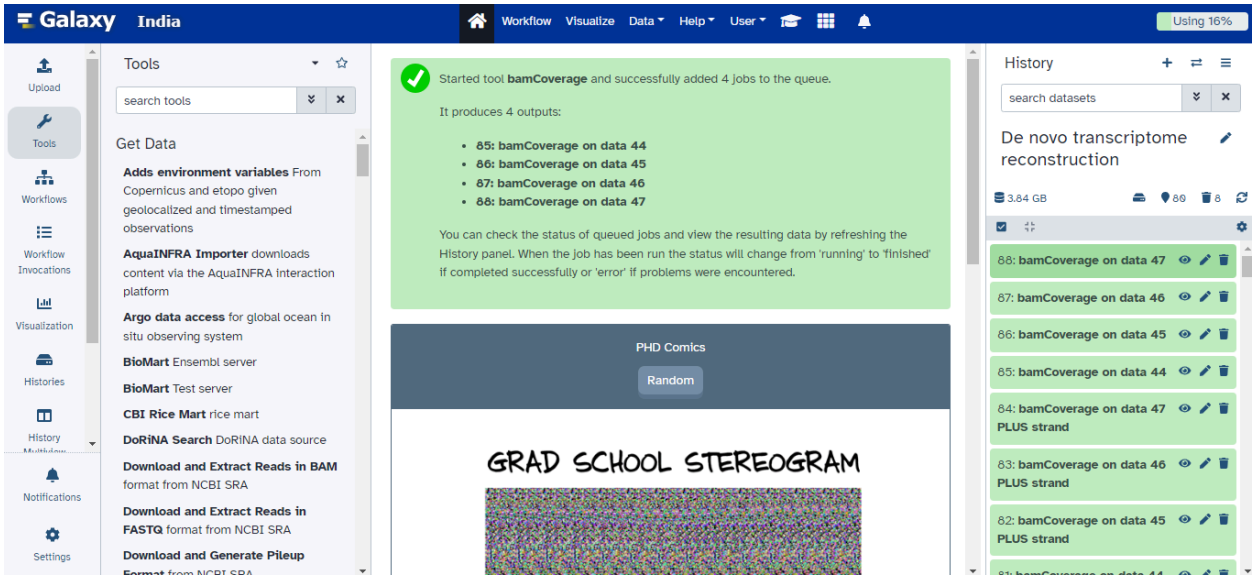


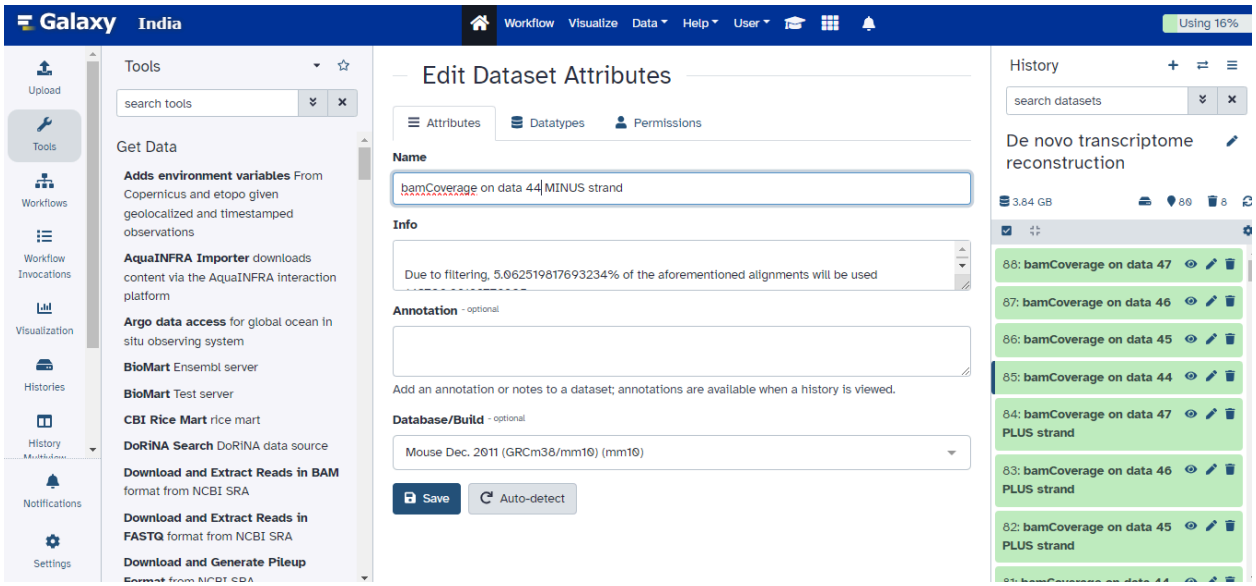
De novo transcriptome reconstruction with RNA-Seq



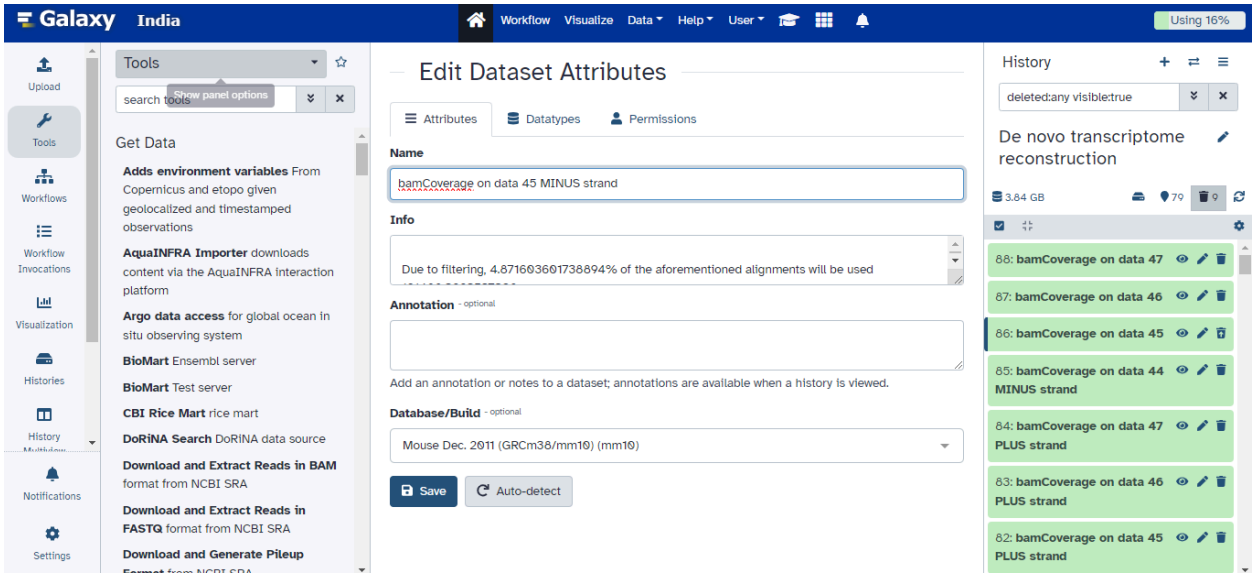


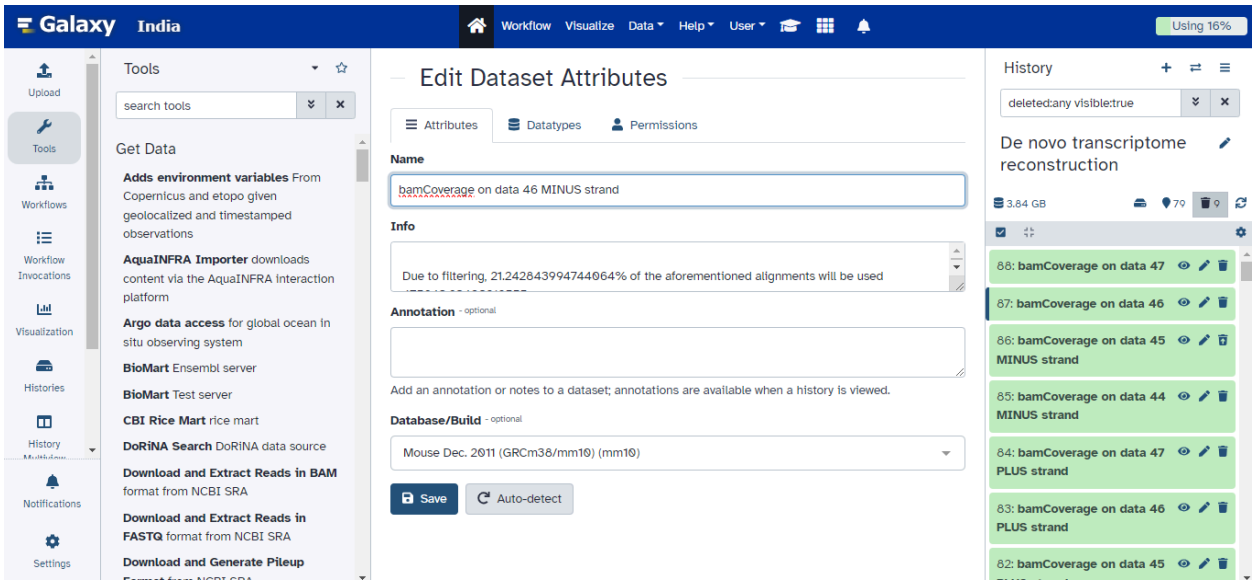
De novo transcriptome reconstruction with RNA-Seq



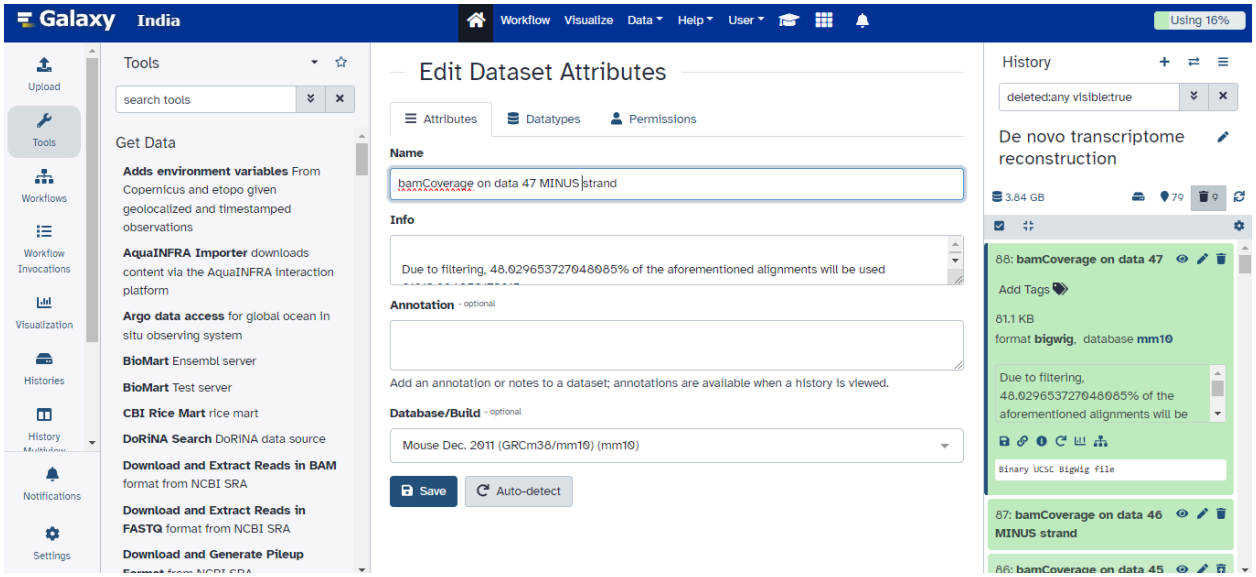


De novo transcriptome reconstruction with RNA-Seq

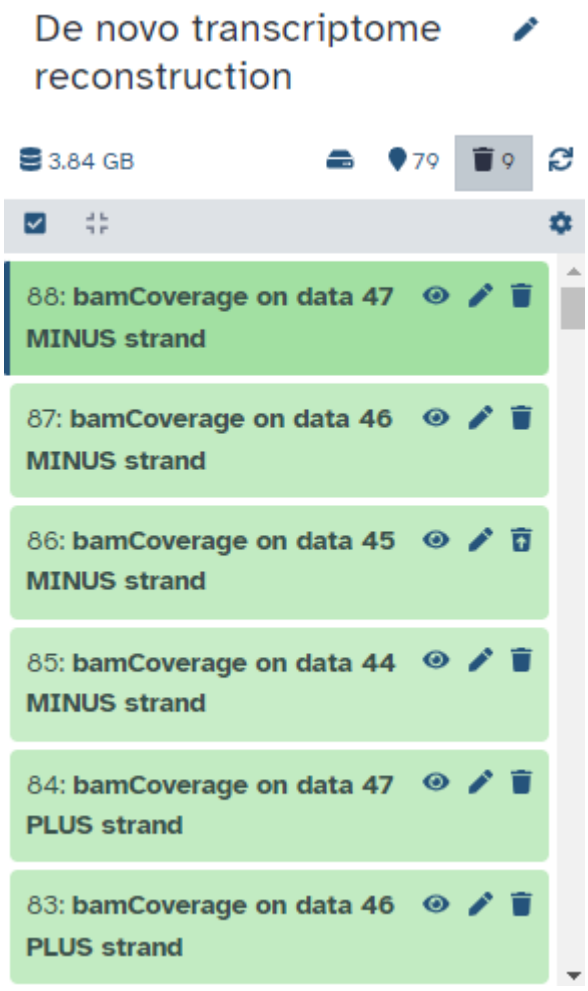




De novo transcriptome reconstruction with RNA-Seq



De novo transcriptome reconstruction with RNA-Seq



De novo transcriptome reconstruction with RNA-Seq

Trackster based visualization:

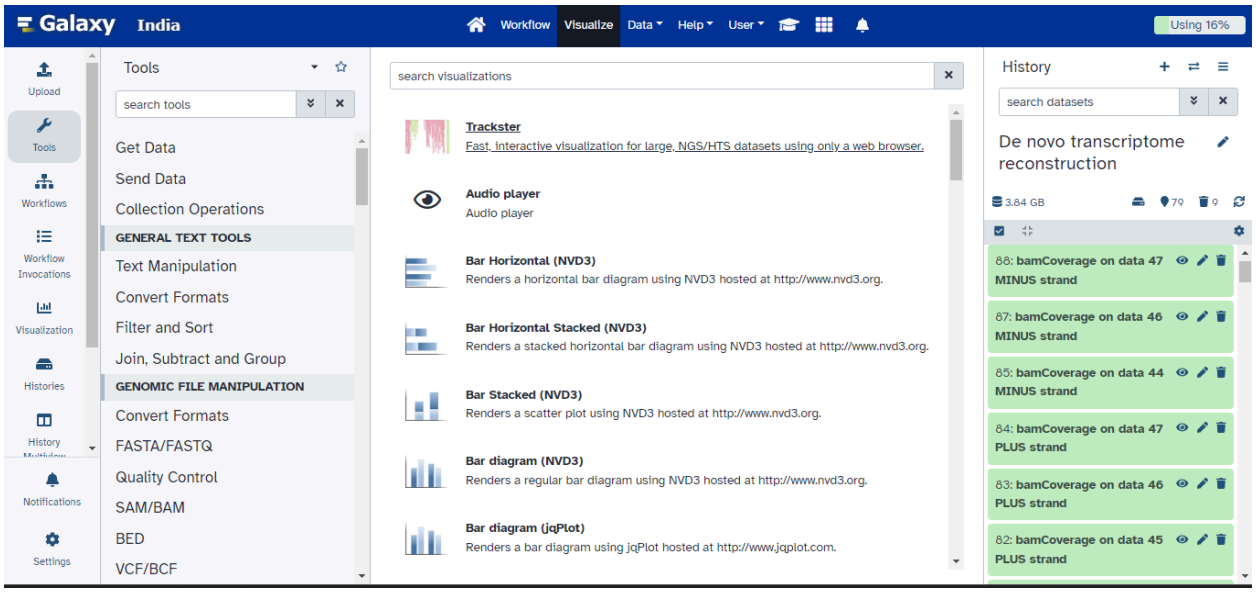
At **chr11:96191452-96206029**, transcript structures from RefSeq, StringTie, and GFFCompare enable comparison of known and predicted transcripts. The bigWig tracks display RNA-seq read density, with blue for the positive strand and red for the negative strand, reflecting transcript abundance. This visualization 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"

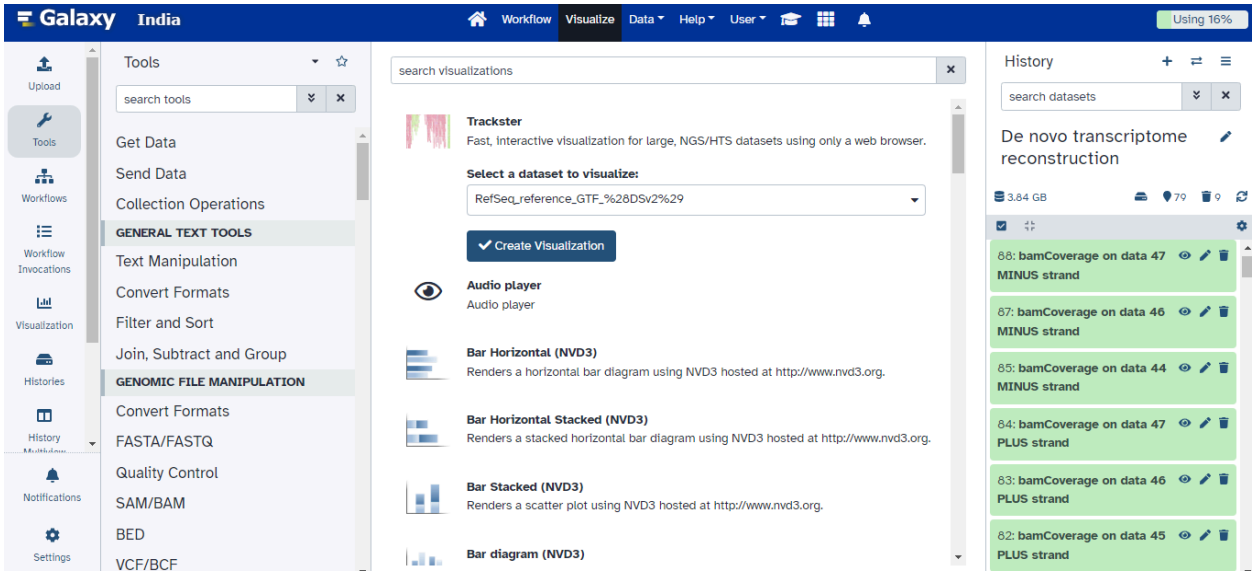
- Name your visualization someting descriptive under "Browser name:"
- Choose "Mouse Dec. 2011 (GRCm38/mm10) (mm10)" as the "Reference genome build (dbkey)
- Click "Create" to initiate your Trackster session
- Viz tool : Click "Add datasets to visualization"
    - Select the "RefSeq GTF mm10" file
    - 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.
- Too l: 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)

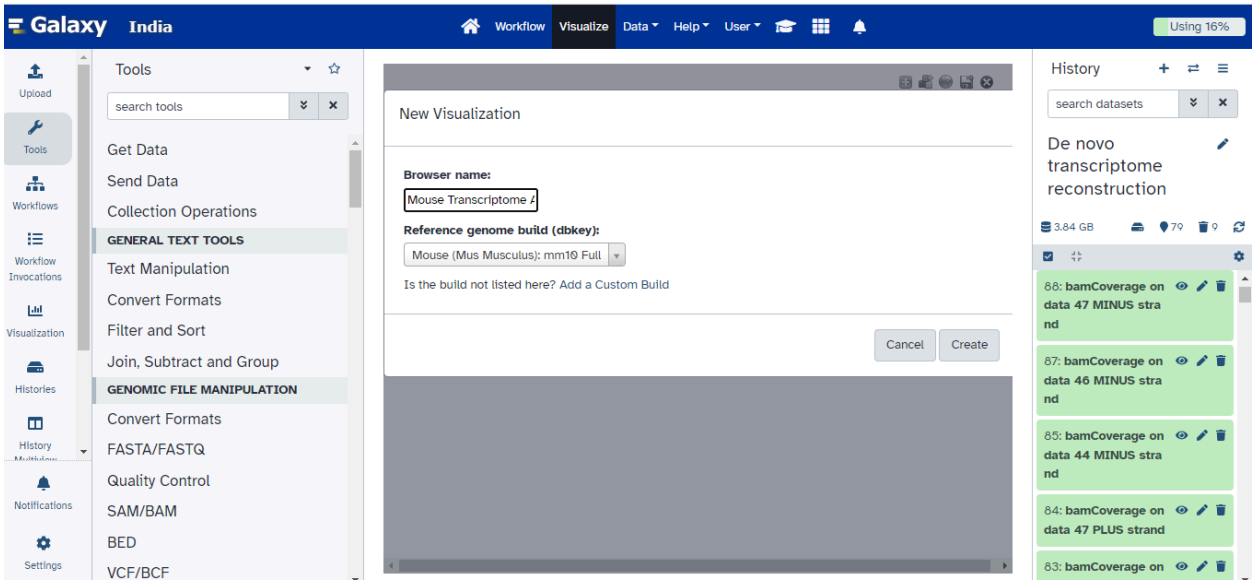
De novo transcriptome reconstruction with RNA-Seq

- 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.

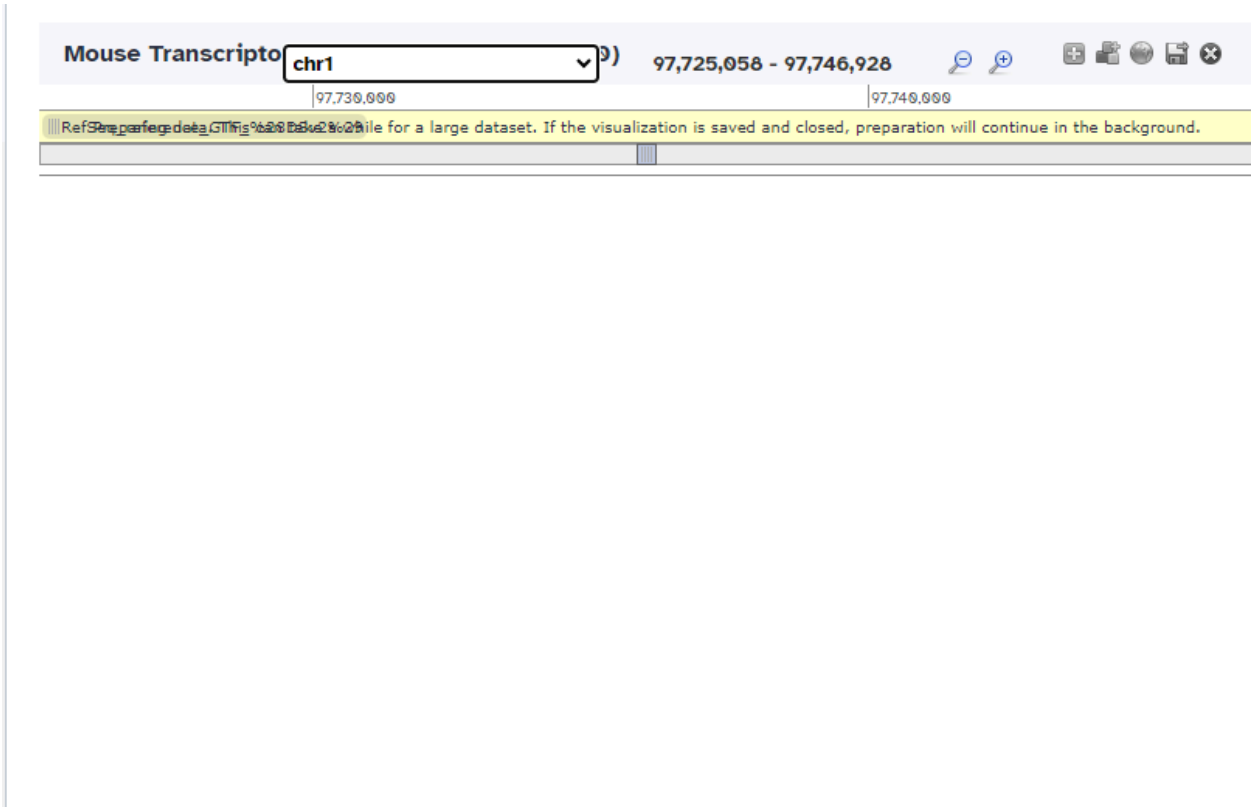


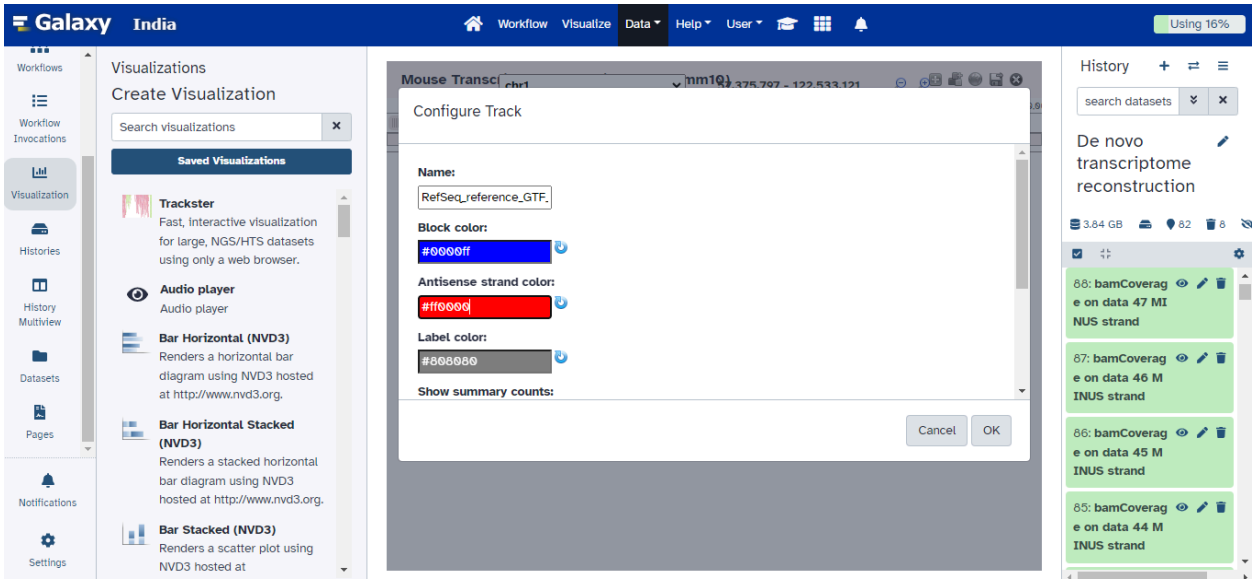
De novo transcriptome reconstruction with RNA-Seq



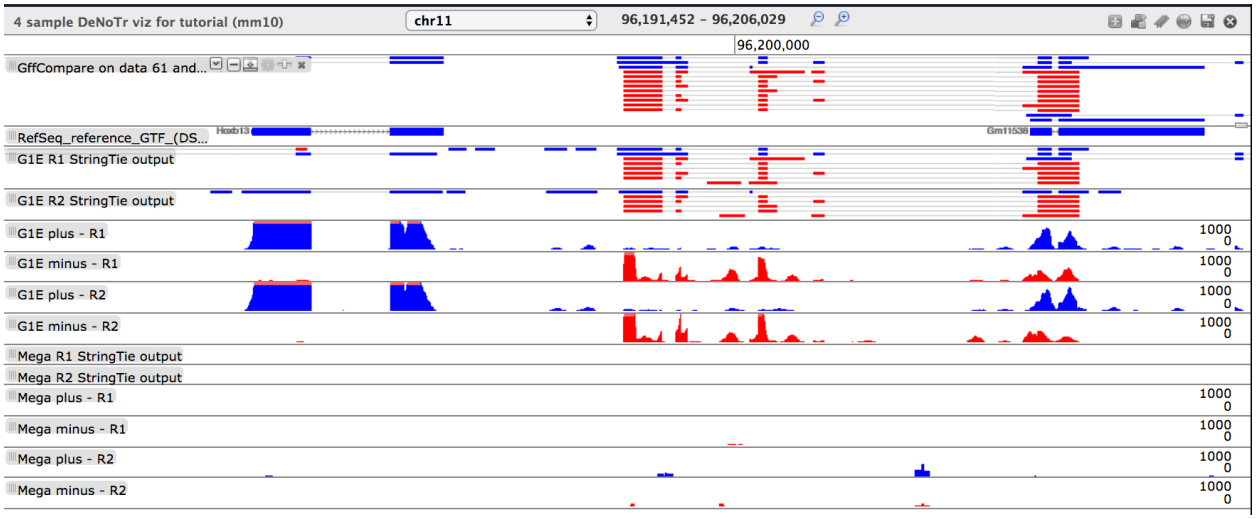


De novo transcriptome reconstruction with RNA-Seq





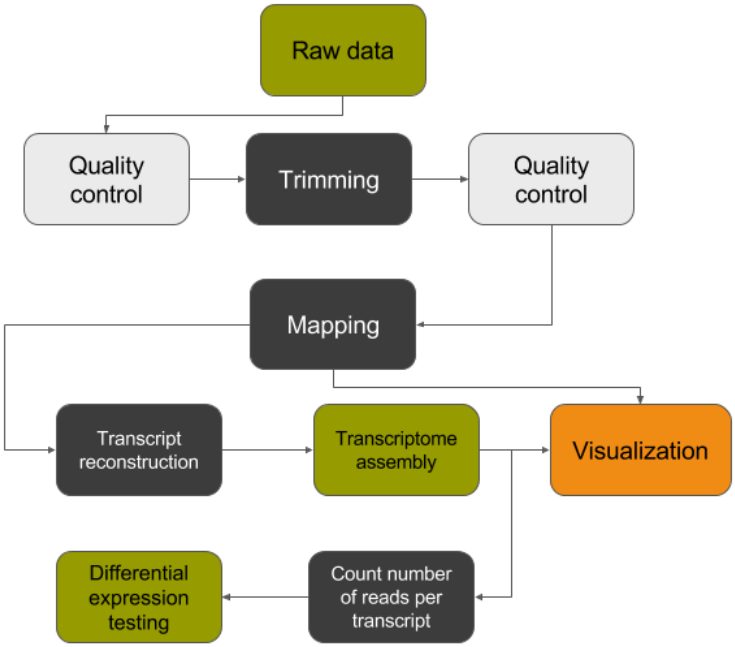
De novo transcriptome reconstruction with RNA-Seq



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.