MAPPING RIBO-SEQ AND RNA-SEQ SAMPLES to TRANSCRIPTOME in RIBOGALAXY

This is the second part of the tutorial. Here the user can learn how to map Ribo-Seq and matching RNA-Seq reads to a transcriptome and prepare files that can be transferred to Trips-viz and analysed and visualised there using the RiboGalaxy instance available at https://ribogalaxy.genomicsdatascience.ie/. At the end of the tutorial we will show some examples of diagnostic plots available in Trips-viz and Ribo-seq subcodon profiles. For a more comprehensive tutorial on Trips-viz, please consult the corresponding publications and walkthrough example for Trips-viz https://trips.ucc.ie/help/?parent_acc=parent_walkthrough. If you're participating in the EMBO course, there will be a dedicated session where you can learn how to use Trips-viz.

As input, you will need clean FASTQ files from part 1 of the tutorial (no adapters, UMIs, untemplated additions, barcodes and ncRNA contaminants). We assume here that you have the files in the RiboGalaxy instance and we are going to continue working with it.

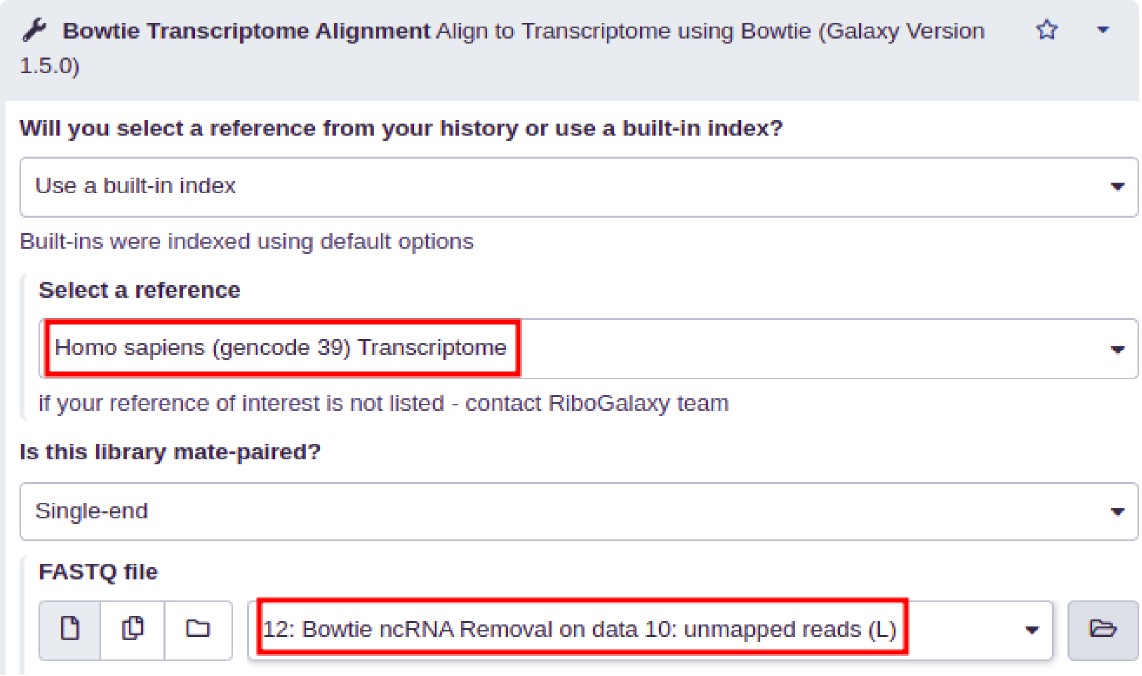
Mapping RIBO-seq to a transcriptome

Let's talk briefly about what to expect at this stage.

- First, we will map clean reads (ribosome protected fragments) from the part 1 tutorial to the transcriptome using ***Bowtie Transcriptome Alignment.***
- Next, we will show how to do PCR-deduplication with UMI-tools, although, for transcriptomic alignments it is typically not recommended (see https://umi-tools.readthedocs.io/en/latest/Single_cell_tutorial.html#mapping-to-the-transcriptome-rather-than-genome). First, we will sort the resulting alignments by coordinate. We will use ***Samtools sort***.
- Then we will deduplicate our sample based on UMIs using ***UMI-tools deduplicate***.
- Next, we will sort reads by name in the deduplicated sample using ***Samtools sort*** since this order is required for building SQLITE files. SQLITE is a format that is used to store mapped reads and associated statistics (e.g. triplet periodicity and ambiguity of mapping) for downstream analysis and visualisation in Trips-viz browser.
- Finally, we will create a SQLITE file using the ***BAM to Sqlite*** tool. We will then upload it to Trips-viz and explore subcodon profiles.

Let's begin!

1. We will map clean reads (RPFs) to transcriptome using **Bowtie Transcriptome Alignment** from **Trips-viz (transcriptome mapping) branch**. We will use the built-in index *Homo sapiens (gencode 39) Transcriptome*. Click 'Execute'.



Bowtie Transcriptome Alignment Align to Transcriptome using Bowtie (Galaxy Version 1.5.0)

Will you select a reference from your history or use a built-in index?

Use a built-in index

Built-ins were indexed using default options

Select a reference

Homo sapiens (gencode 39) Transcriptome

if your reference of interest is not listed - contact RiboGalaxy team

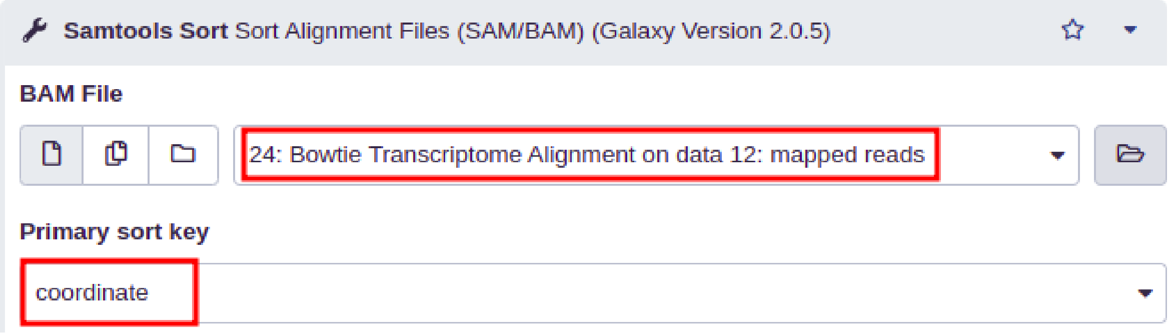
Is this library mate-paired?

Single-end

FASTQ file

12: Bowtie ncRNA Removal on data 10: unmapped reads (L)

2. Sort the BAM file by coordinate using **Samtools sort** from **GWIPS-Viz (genomic alignment) branch**. Click 'Execute'.



Samtools Sort Sort Alignment Files (SAM/BAM) (Galaxy Version 2.0.5)

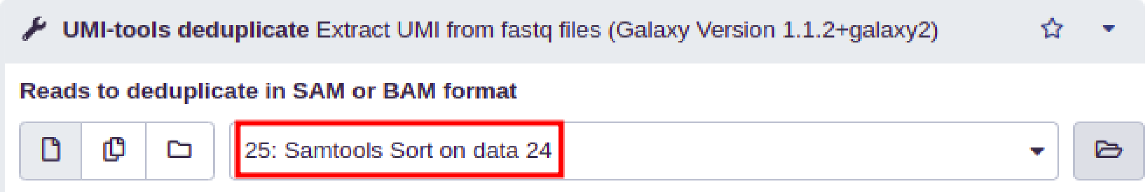
BAM File

24: Bowtie Transcriptome Alignment on data 12: mapped reads

Primary sort key

coordinate

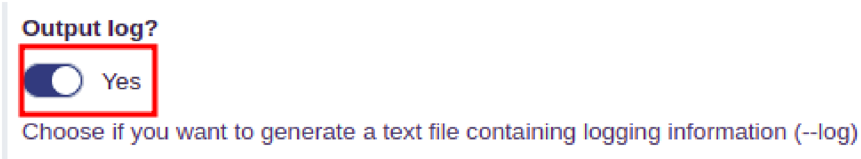
3. In this step we will remove PCR duplicates using UMIs. It can be done with **UMI-tools deduplicate** from **UMI and barcodes**. Input file is BAM sorted by coordinate.You can also choose to output log file and check how many reads were filtered out. Click 'Execute'.



UMI-tools deduplicate Extract UMI from fastq files (Galaxy Version 1.1.2+galaxy2)

Reads to deduplicate in SAM or BAM format

25: Samtools Sort on data 24



Output log?

Yes

Choose if you want to generate a text file containing logging information (--log)

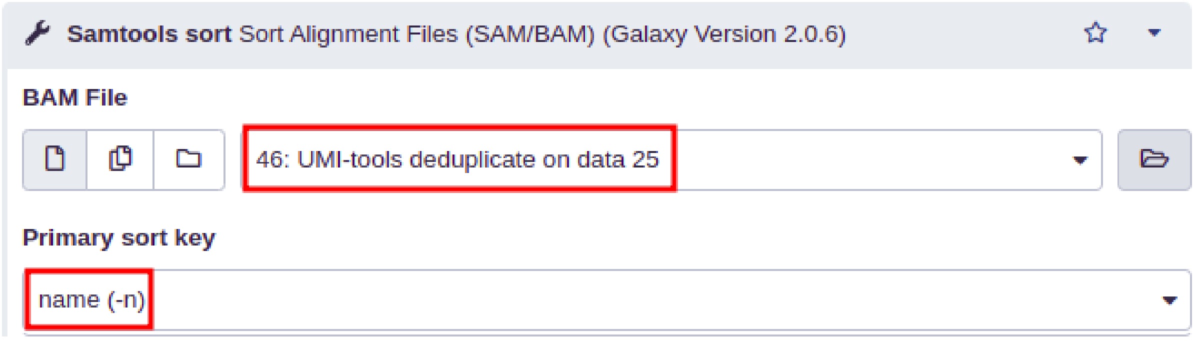
UMI-tools is recommended to be applied on genomic alignments. In the case of transcriptomic alignments, UMI-tools won't be as effective.

Also, depending on a sample size and the overall load on the server, the running job may result in being 'killed' (interrupted) due to the lack of computational resources. You can try to re-run it, however, if it gets killed again, it's most likely not enough resources.

We can take a look at the log file. Interestingly, the input number of reads was 100k, however we see here 68mln. This is because we allow multimappers during alignment with bowtie and also one read can come from different transcript isoforms of the same gene.

```
2023-02-04 17:12:01,146 INFO Reads: Input Reads: 67982166
2023-02-04 17:12:01,146 INFO Number of reads out: 51803806
2023-02-04 17:12:01,146 INFO Total number of positions deduplicated: 2469354
```

4. Sort BAM file by name using **Samtools sort** from **Trips-viz (transcriptome mapping) branch.** Click 'Execute'.



Samtools sort Sort Alignment Files (SAM/BAM) (Galaxy Version 2.0.6)

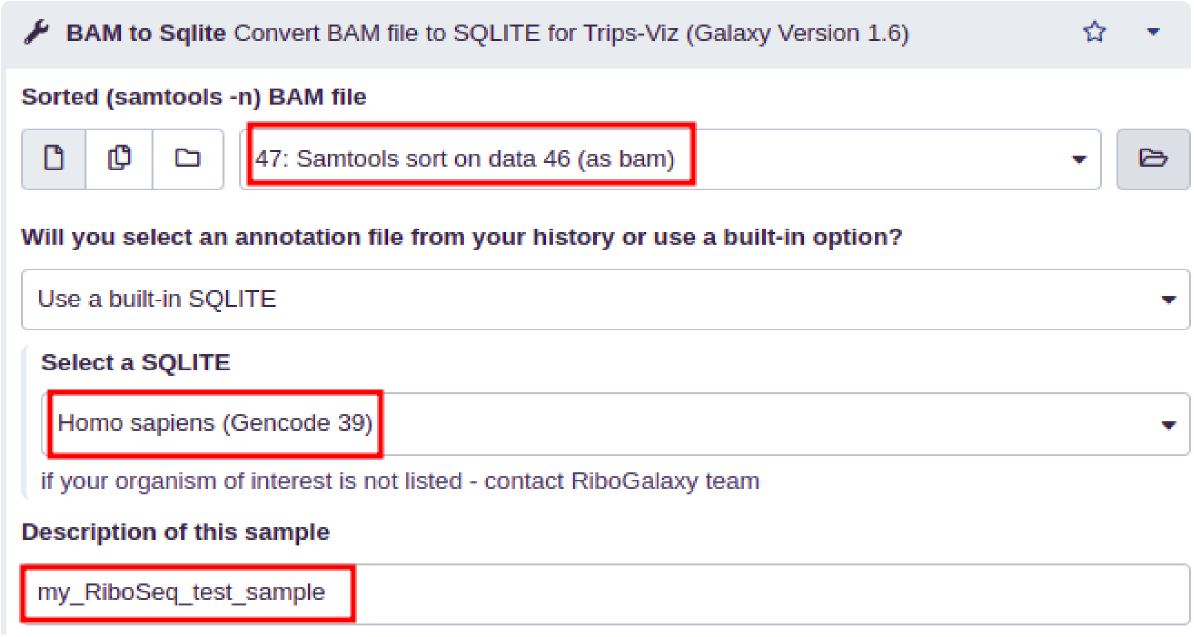
BAM File

46: UMI-tools deduplicate on data 25

Primary sort key

name (-n)

5. FInally we will get a SQLITE file using sorted by name BAM file and annotation SQLITE file (we will choose built-in option: Gencode 39). We will use the **BAM to Sqlite** tool from the **Trips-viz (transcriptome mapping) branch**. We also should add a description of the sample, e.g. '*my_RiboSeq_test_sample*'.



BAM to Sqlite Convert BAM file to SQLITE for Trips-Viz (Galaxy Version 1.6)

Sorted (samtools -n) BAM file

47: Samtools sort on data 46 (as bam)

Will you select an annotation file from your history or use a built-in option?

Use a built-in SQLITE

Select a SQLITE

Homo sapiens (Gencode 39)

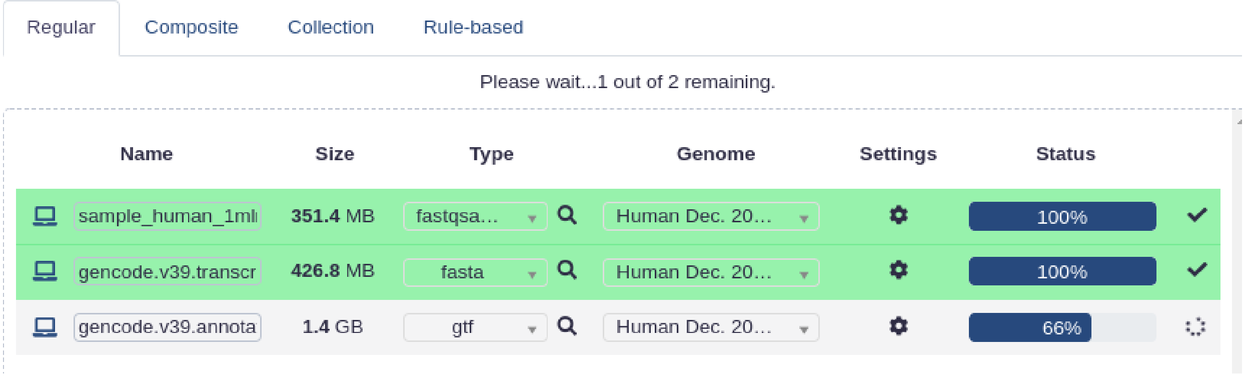
if your organism of interest is not listed - contact RiboGalaxy team

Description of this sample

my_RiboSeq_test_sample

5' (***Optional, you can skip this step since the annotation SQLITE file is provided as a built-in option***). Otherwise you can build an annotation SQLITE yourself in case there is no organism or transcriptome annotation provided. Note that it will take a while. We will use the **Create TRIPS-Viz annotation** from the **Trips-viz (transcriptome mapping) branch**. For that you first need to download files and *unzip them*, and then upload to the RiboGalaxy:

- https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_39/gencode.v39.annotation.gtf.gz - .gtf file
- https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_39/gencode.v39.transcripts.fa.gz - .fasta file

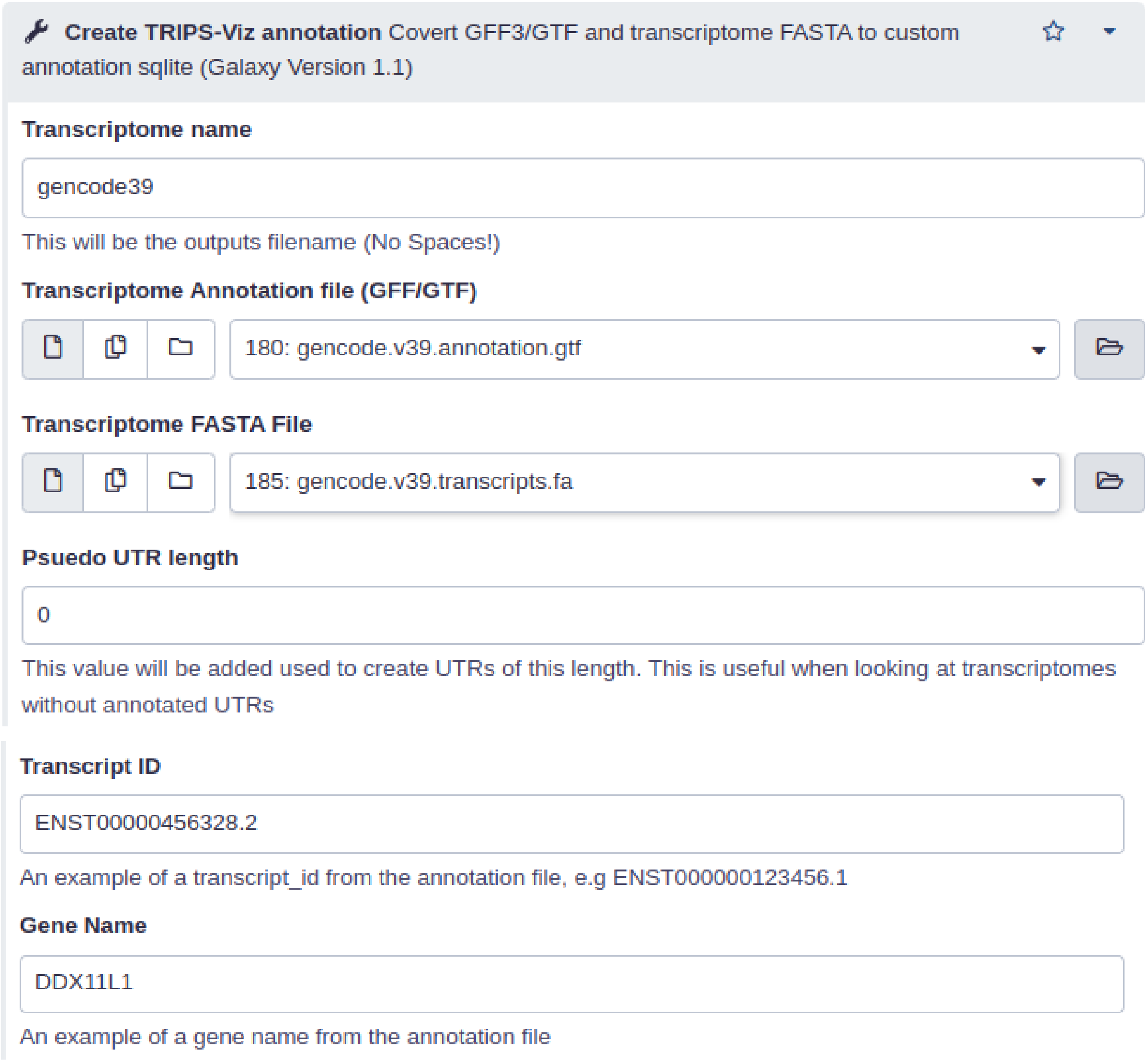


Please wait...1 out of 2 remaining.

Name	Size	Type	Genome	Settings	Status
sample_human_1mli	351.4 MB	fastqsa...	Human Dec. 20...		100% ✓
gencode.v39.transcr	426.8 MB	fasta	Human Dec. 20...		100% ✓
gencode.v39.annota	1.4 GB	gtf	Human Dec. 20...		66%

Then you need to use following input files and parameters:

- Transcriptome name: '*gencode39*'
- Transcriptome annotation file: uploaded GTF
- Transcriptome fasta file: uploaded FASTQ
- Pseudo UTR length: 0
- Transcript id: *ENST00000456328.2*
- Gene name: *DDX11L1*



Create TRIPS-Viz annotation Covert GFF3/GTF and transcriptome FASTA to custom annotation sqlite (Galaxy Version 1.1)

Transcriptome name

gencode39

This will be the outputs filename (No Spaces!)

Transcriptome Annotation file (GFF/GTF)

180: gencode.v39.annotation.gtf

Transcriptome FASTA File

185: gencode.v39.transcripts.fa

Psuedo UTR length

0

This value will be added used to create UTRs of this length. This is useful when looking at transcriptomes without annotated UTRs

Transcript ID

ENST00000456328.2

An example of a transcript_id from the annotation file, e.g ENST000000123456.1

Gene Name

DDX11L1

An example of a gene name from the annotation file

Gene and transcript id are taken from the beginning of the GTF file, it helps to better parse the GTF file since they typically have different formats depending on a gene annotation provider and organisms. Pseudo UTR length is not set on 0 when the transcriptome does not have UTRs (like yeast transcriptomes, in this case typically you would choose 300). Click ‘Execute’.

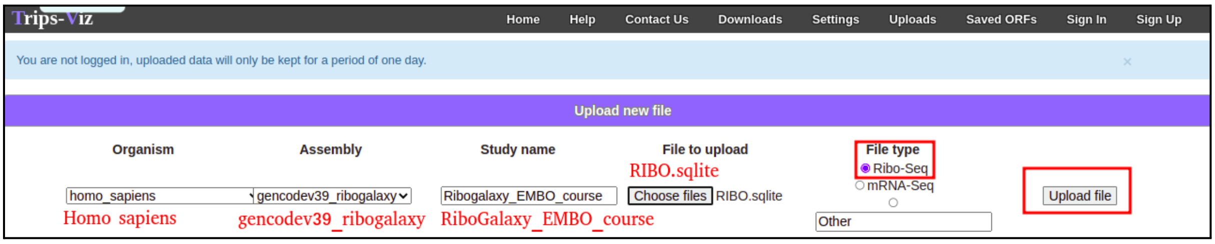
6. We can download the SQLITE file from step 4 and upload it to the Trips-viz (trips.ucc.ie). Ideally you will need to sign up to Trips-viz first and then sign in so that you can access your uploaded data for a longer time.

GENCODE v39 is available in Trips-viz and it’s called *gencodev39_ribogalaxy*. We re-named file to RIBO.sqlite.

Choose Upload new file to upload the sample bam sqlite.

- Organism: homo_sapiens
- Assembly: gencode39_ribogalaxy
- Study name: e.g. Ribogalaxy_EMBO_course
- choose sqlite file for uploading.

- File type by default is Ribo-Seq.



You are not logged in, uploaded data will only be kept for a period of one day.

Organism	Assembly	Study name	File to upload	File type
homo_sapiens	gencodev39_ribogalaxy	Ribogalaxy_EMBO_course	RIBO.sqlite	● Ribo-Seq ○ mRNA-Seq ○ Other

RIBO.sqlite

Homo sapiens gencodev39_ribogalaxy RiboGalaxy_EMBO_course

You'll need to wait till you see that 'File uploaded successfully'.

Optional step, can be performed together with 5':
In case if you're working with a new transcriptome that is not available in RiboGalaxy and Trips-viz and you created an annotation SQLITE file on step 4', you also need to upload it to Trips-viz. In our case it's GENCODE v39 annotation as an example.
Go to https://trips.ucc.ie/uploads/ and choose Upload new transcriptome.

- Organism name: '*homo_sapiens*'
- Assembly name: '*gencode39_ribogalaxy'* (though you can put any name of assembly and organism)
- Default transcript: *ENST00000559916*
- Choose annotation sqlite file for uploading (it's ~3Gb)

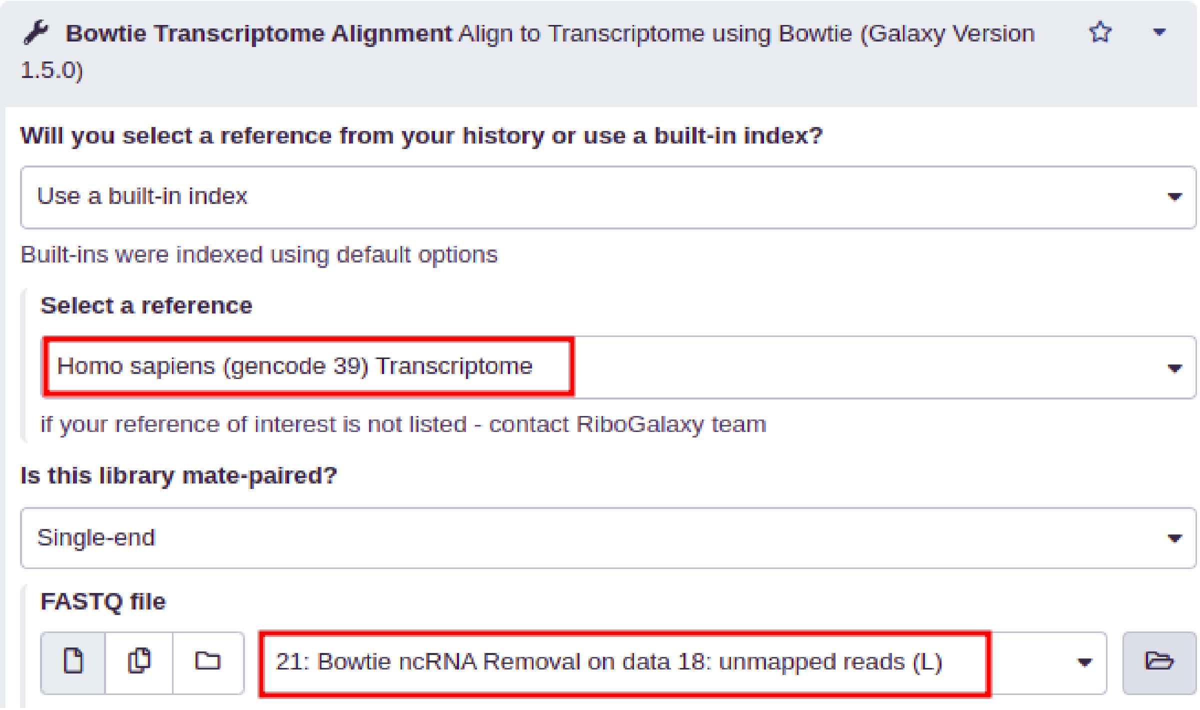
Then when you need to upload your alignments in .sqlite format, you can select your already uploaded custom transcriptome.

Mapping RNA-seq to a transcriptome

Let's talk briefly about what to expect at this stage.

- First, we will map clean RNA-seq reads (without adapters and ncRNAs) to the transcriptome using ***Bowtie Transcriptome Alignment.***
- Next, we will sort reads by name in the deduplicated sample using ***Samtools sort*** since this order is required for building SQLITE files.
- Finally, we will create a SQLITE file using the ***BAM to Sqlite*** tool. We will then upload it to Trips-viz.

1. We will map clean RNA-seq reads to the transcriptome using **Bowtie Transcriptome Alignment** from **Trips-viz (transcriptome mapping) branch**. We will use the built-in index *Homo sapiens (gencode 39) Transcriptome*. Click 'Execute'.



Bowtie Transcriptome Alignment Align to Transcriptome using Bowtie (Galaxy Version 1.5.0)

Will you select a reference from your history or use a built-in index?

Use a built-in index

Built-ins were indexed using default options

Select a reference

Homo sapiens (gencode 39) Transcriptome

if your reference of interest is not listed - contact RiboGalaxy team

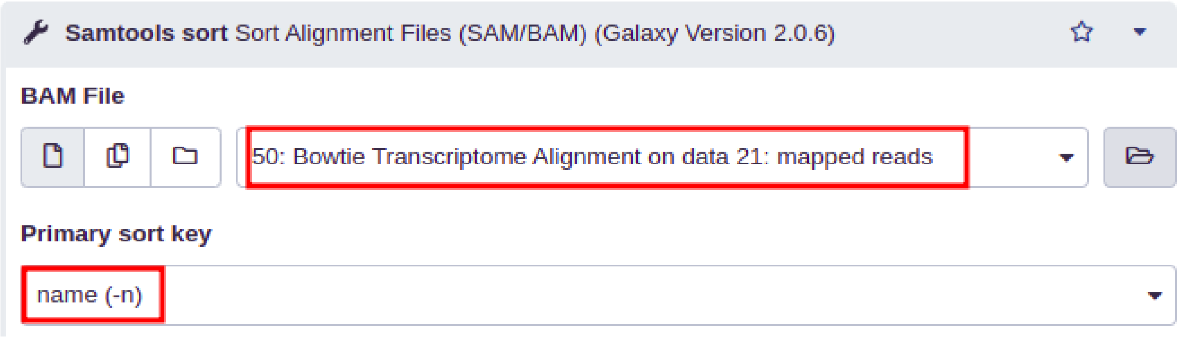
Is this library mate-paired?

Single-end

FASTQ file

21: Bowtie ncRNA Removal on data 18: unmapped reads (L)

2. Then we will sort the **BAM** file by name using **Samtools sort** from the **Trips-viz (transcriptome mapping) branch.** Click ‘Execute’.



Samtools sort Sort Alignment Files (SAM/BAM) (Galaxy Version 2.0.6)

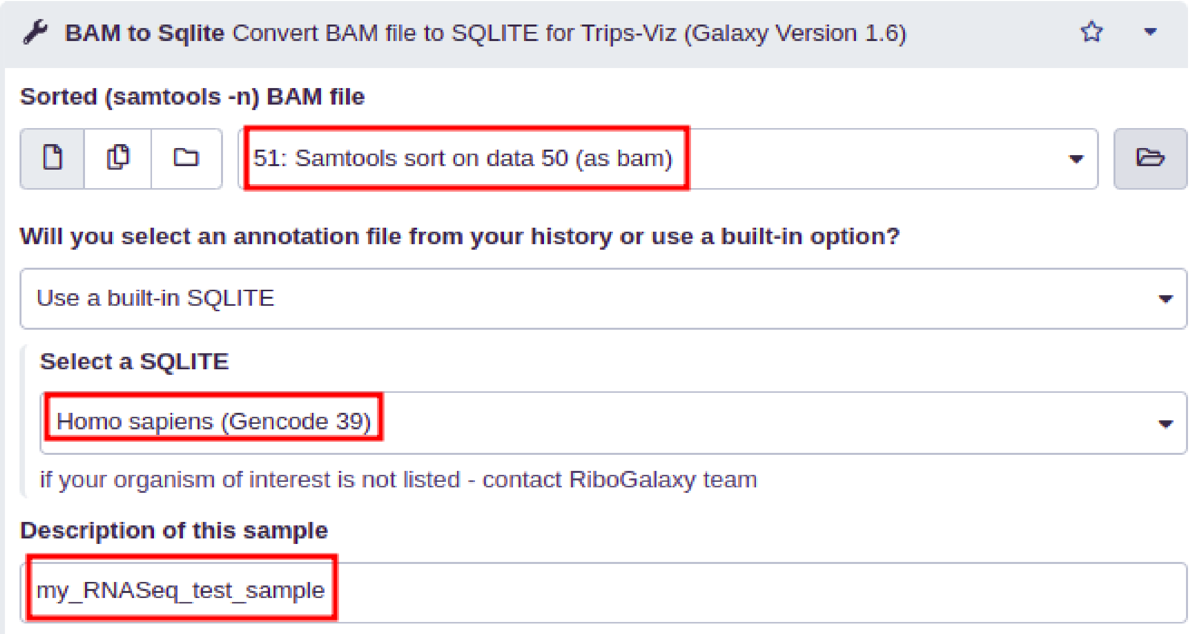
BAM File

50: Bowtie Transcriptome Alignment on data 21: mapped reads

Primary sort key

name (-n)

3. FInally we will get a SQLITE file using sorted by name BAM file and annotation SQLITE file (we will choose built-in option: Gencode 39). We will use the **BAM to Sqlite** tool from the **Trips-viz (transcriptome mapping) branch**. We also should add a description of the sample, e.g. ‘*my_RNASeq_test_sample*’.



BAM to Sqlite Convert BAM file to SQLITE for Trips-Viz (Galaxy Version 1.6)

Sorted (samtools -n) BAM file

51: Samtools sort on data 50 (as bam)

Will you select an annotation file from your history or use a built-in option?

Use a built-in SQLITE

Select a SQLITE

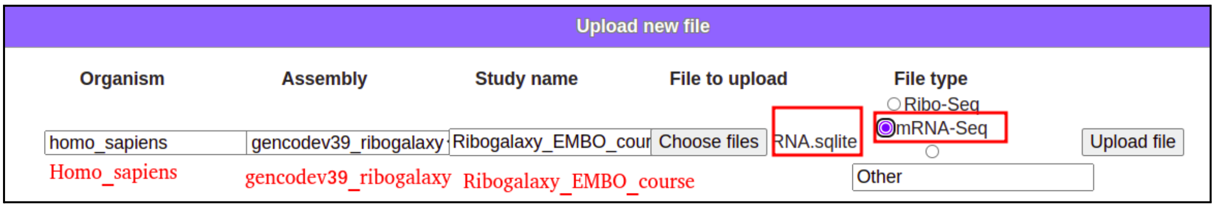
Homo sapiens (Gencode 39)

if your organism of interest is not listed - contact RiboGalaxy team

Description of this sample

my_RNASeq_test_sample

4. Now we can upload the resulting SQLITE file to the Trips-viz. The only difference from uploading a Ribo-seq SQLITE file would be ‘File Type’ - ‘mRNA-seq’.



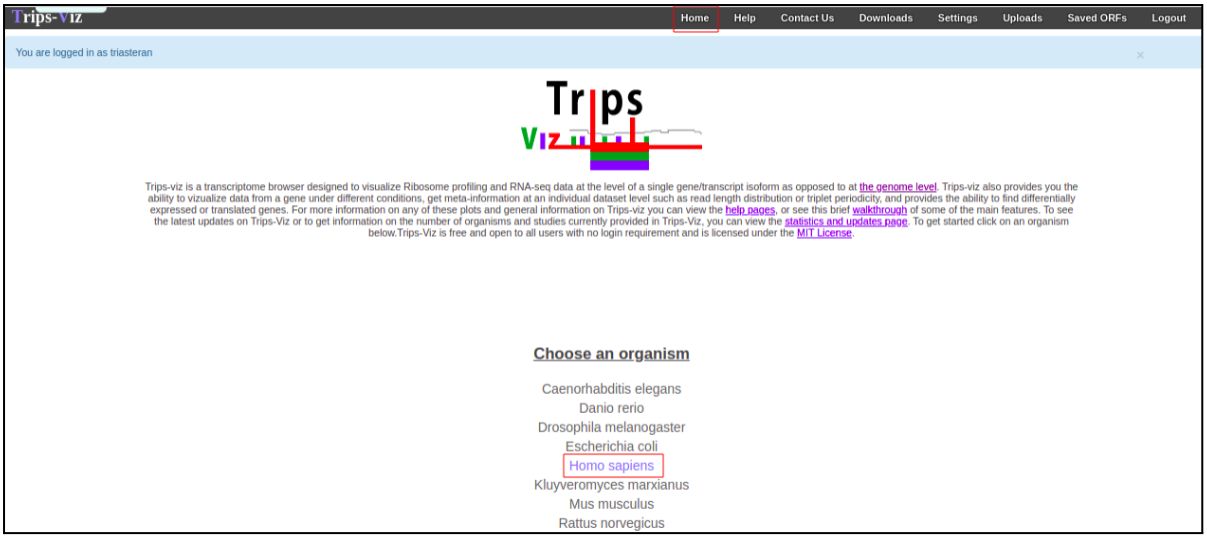
Upload new file

Organism	Assembly	Study name	File to upload	File type	
homo_sapiens	gencodev39_ribogalaxy	Ribogalaxy_EMBO_cour	Choose files RNA.sqlite	○ Ribo-Seq ● mRNA-Seq ○ Other	Upload file
Homo_sapiens	gencodev39_ribogalaxy	Ribogalaxy_EMBO_course			

Post-alignment Ribo-seq quality control and visualisation in Trips-viz

Now we are going to assess the post-alignment quality of the Ribo-seq data using the Trips-viz transcriptome browser (trips.ucc.ie). Of note, this is just a brief introduction of Trips-viz functionality since the full description is out of the scope of this tutorial.
First we will look at the distribution of read length of the Ribo-seq sample.

Considering that we are logged in and we have already uploaded all necessary files, let’s go to the homepage of Trips-viz, choose *Homo sapiens:*



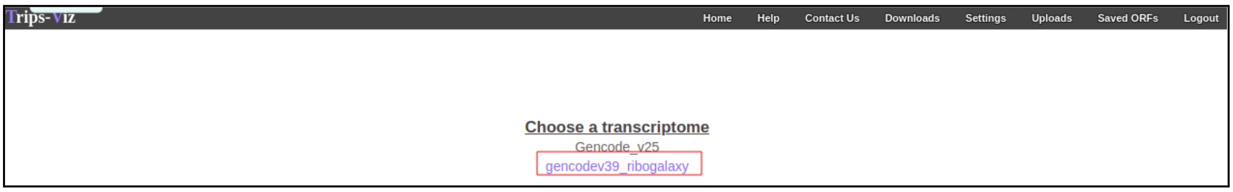
You are logged in as triasteran

Trips-viz is a transcriptome browser designed to visualize Ribosome profiling and RNA-seq data at the level of a single gene/transcript isoform as opposed to at the genome level. Trips-viz also provides you the ability to vizualize data from a gene under different conditions, get meta-information at an individual dataset level such as read length distribution or triplet periodicity, and provides the ability to find differentially expressed or translated genes. For more information on any of these plots and general information on Trips-viz you can view the help pages, or see this brief walkthrough of some of the main features. To see the latest updates on Trips-Viz or to get information on the number of organisms and studies currently provided in Trips-Viz, you can view the statistics and updates page. To get started click on an organism below.Trips-Viz is free and open to all users with no login requirement and is licensed under the MIT License.

Choose an organism

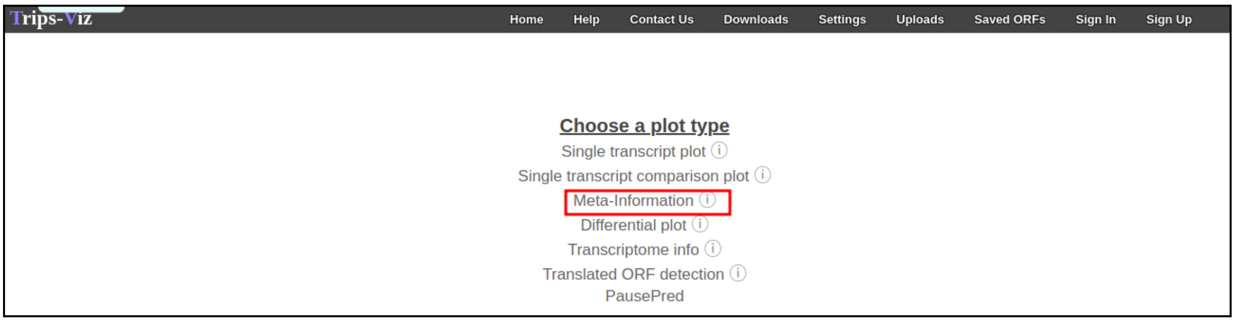
Caenorhabditis elegans
Danio rerio
Drosophila melanogaster
Escherichia coli
Homo sapiens
Kluyveromyces marxianus
Mus musculus
Rattus norvegicus

Then choose *gencode39_ribogalaxy* as transcriptome and choose Meta-information:



Choose a transcriptome

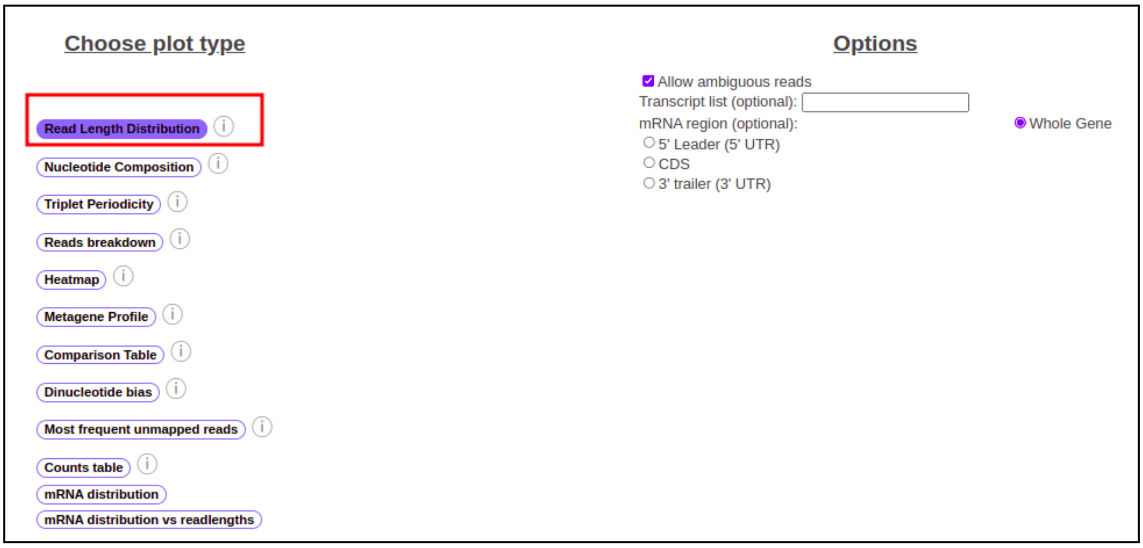
Gencode_v25
gencodev39_ribogalaxy



Choose a plot type

Single transcript plot ⓘ
Single transcript comparison plot ⓘ
Meta-Information ⓘ
Differential plot ⓘ
Transcriptome info ⓘ
Translated ORF detection ⓘ
PausePred

Next, in ‘Choose plot type’, select ‘Read Length Distribution’ (all options can be set on default values). Then select the uploaded file (RIBO.sqlite) and run ‘View plot’.

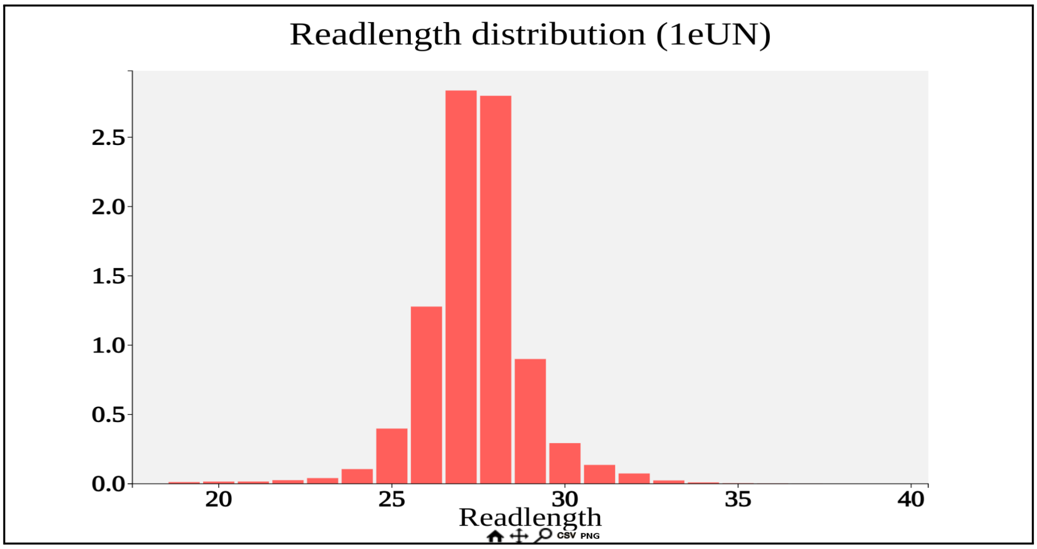


Choose plot type

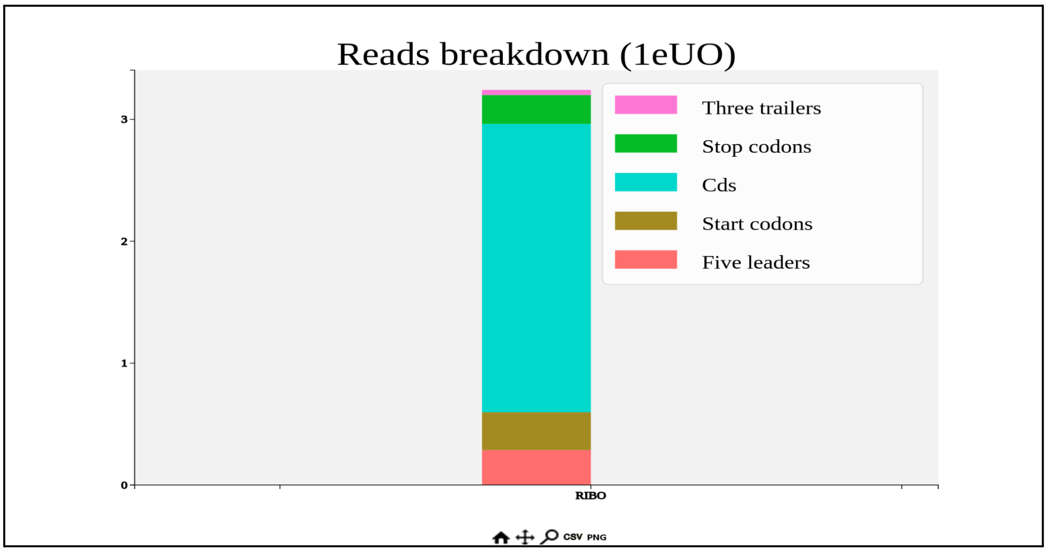
Options

☑ Allow ambiguous reads
Transcript list (optional):
mRNA region (optional): ● Whole Gene
○ 5' Leader (5' UTR)
○ CDS
○ 3' trailer (3' UTR)

You can see that most of the reads are 25-30nt long as expected.



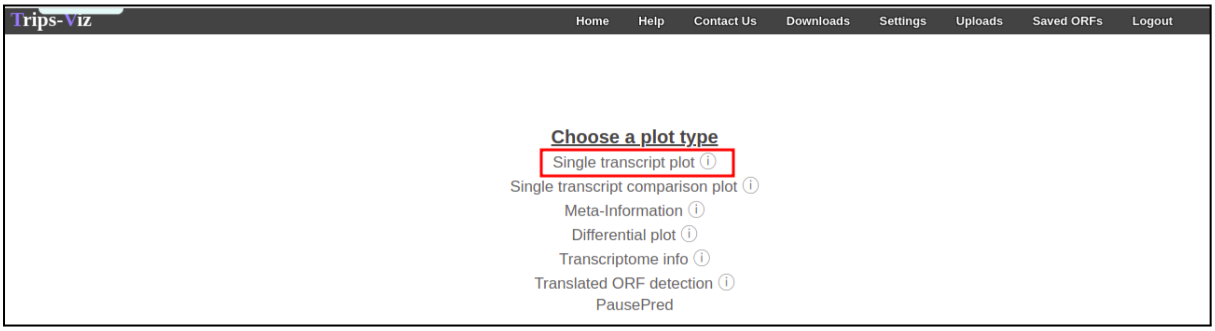
Next, we can select ‘mRNA distribution’. We can see that most reads fall within CDS and more reads map on 5’ leaders than on 3’ trailers as expected of Ribo-seq.



Since it is quite a small sample of reads that map to only 6 transcripts, some important plots such as metagene profile and triplet periodicity are empty.

Finally, we can study the sub-codon Ribo-seq profile. Sample contains reads that map on a set of 6 transcripts: ENST00000559916 (B2M), ENST00000371222 (JUN), ENST00000621592 (MYC), ENST00000373316 (PGK1), ENST00000674681 (ACTB), ENST00000396861 (GAPDH).

Instead of choosing Meta-information, now we choose Single transcript plot:



Choose a plot type

Single transcript plot ⓘ

Single transcript comparison plot ⓘ

Meta-Information ⓘ

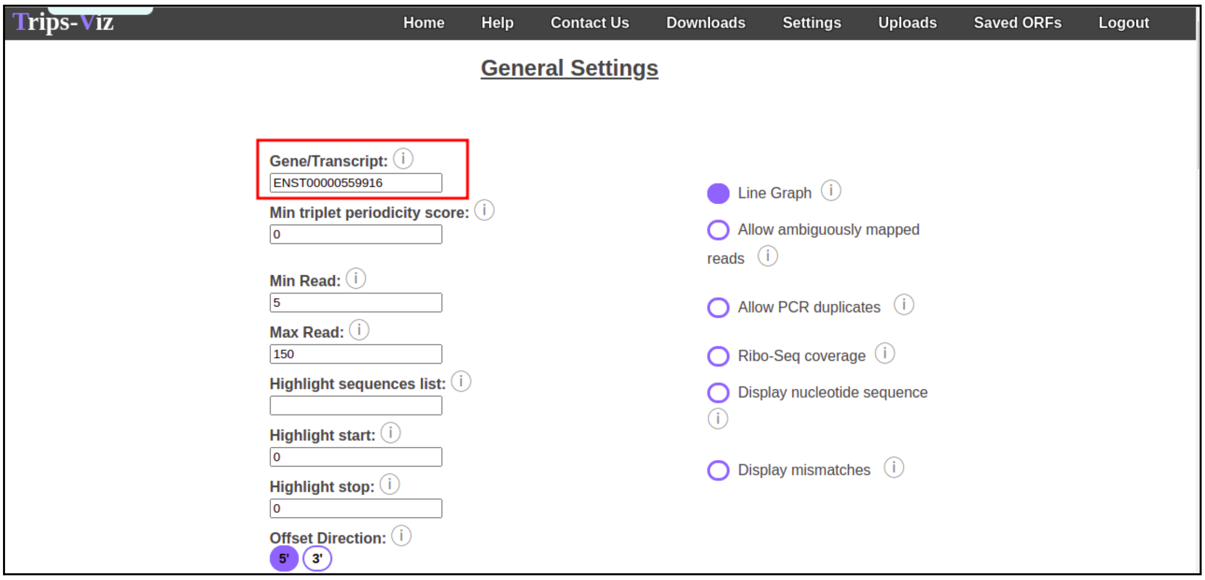
Differential plot ⓘ

Transcriptome info ⓘ

Translated ORF detection ⓘ

PausePred

And put ‘ENST00000559916’ in Gene/Transcript, select Ribo-seq sample that you uploaded and then run ‘View plot’.



General Settings

Gene/Transcript: ⓘ ENST00000559916

Min triplet periodicity score: ⓘ 0

Min Read: ⓘ 5

Max Read: ⓘ 150

Highlight sequences list: ⓘ

Highlight start: ⓘ 0

Highlight stop: ⓘ 0

Offset Direction: ⓘ ● 5' ○ 3'

● Line Graph ⓘ

○ Allow ambiguously mapped reads ⓘ

○ Allow PCR duplicates ⓘ

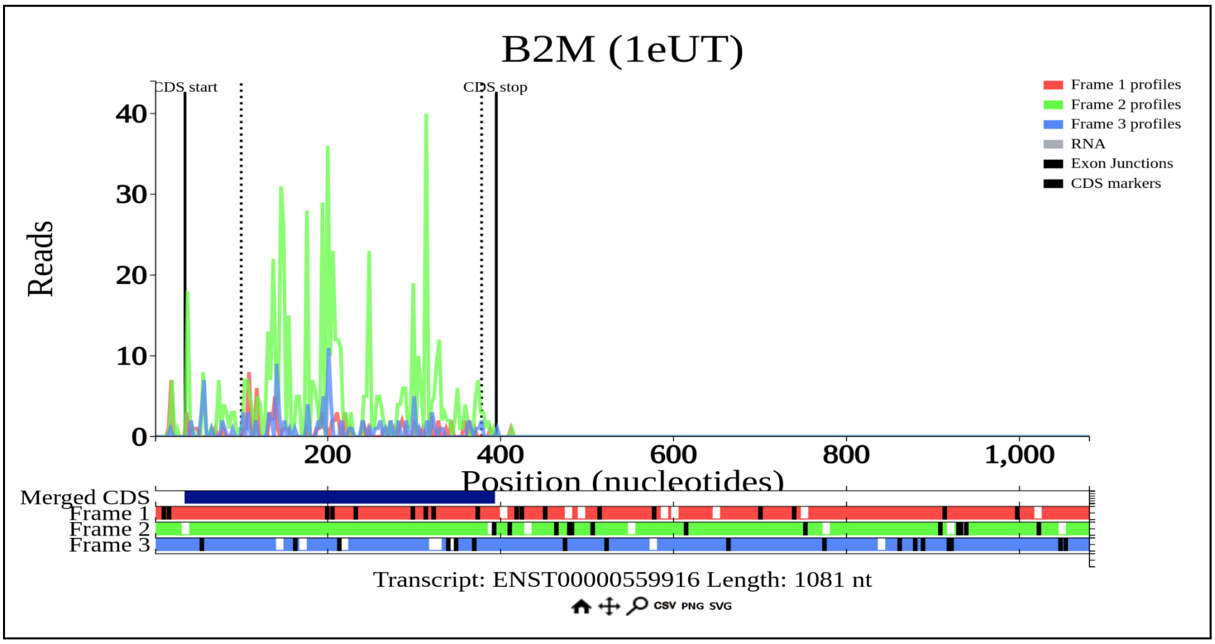
○ Ribo-Seq coverage ⓘ

○ Display nucleotide sequence ⓘ

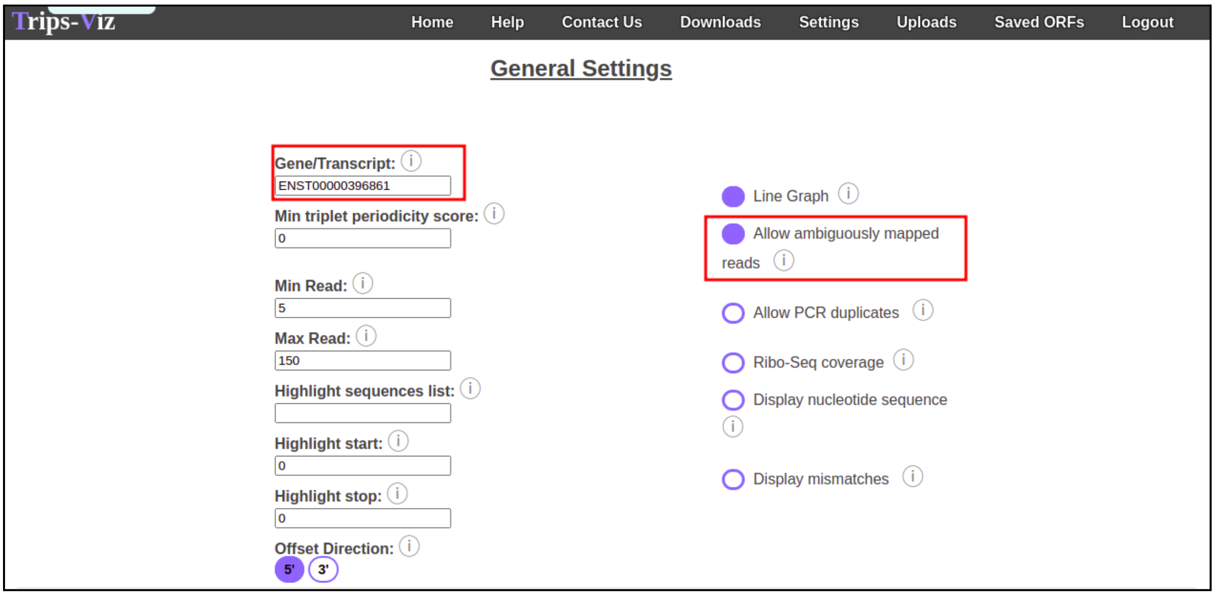
○ Display mismatches ⓘ

You will see a subcodon Ribo-seq profile for *B2M* gene. There are three reading frames that are shown in **red**, **blue** and **green**. Y-axis shows read counts, while x-axis depicts the position in a transcript. Bottom ORF plot consists of 3 horizontal bars that stand for 3 reading frames and have AUGs as white lines and stop codons as black lines. On top of the ORF plot there is another horizontal bar (dark blue) that shows merged coding sequences. Annotated CDS start and CDS stop are shown as vertical solid black lines. Dotted lines represent exon-exon junctions.

As you may notice, the reading frame of CDS is green (CDS start corresponds to AUG in green frame and CDS stop corresponds to STOP codon in green frame, also there are no stops interrupting the CDS). Also you can see that the dominant Ribo-seq signal in CDS is coloured in green - the colour of the CDS reading frame (as expected). We do observe some translation in 5’leader (looks like it could be non-AUG N-terminal extension) while we do not see any noticeable signal of translation in 3’trailer (UTR).



Let’s now have a look at ENST00000396861 (GAPDH). In this case we will need to tick the ‘Allow ambiguously mapped reads’. Typically, if you see that a profile looks ‘patchy’ and has dips in it or no signal in CDS where it is expected, the first thing you could try is to allow ambiguously mapped reads (reads that map to multiple locations).



General Settings

Gene/Transcript: ENST00000396861

Min triplet periodicity score: 0

Min Read: 5

Max Read: 150

Highlight sequences list:

Highlight start: 0

Highlight stop: 0

Offset Direction: [✓ 5'] 3'

● Line Graph

● Allow ambiguously mapped reads

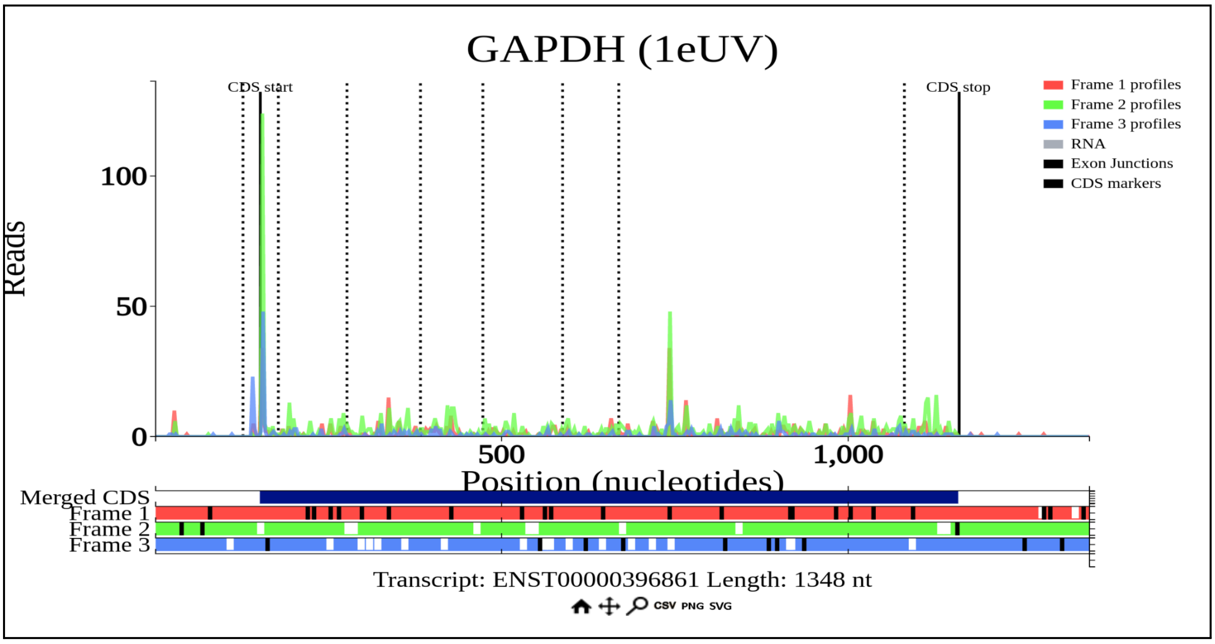
○ Allow PCR duplicates

○ Ribo-Seq coverage

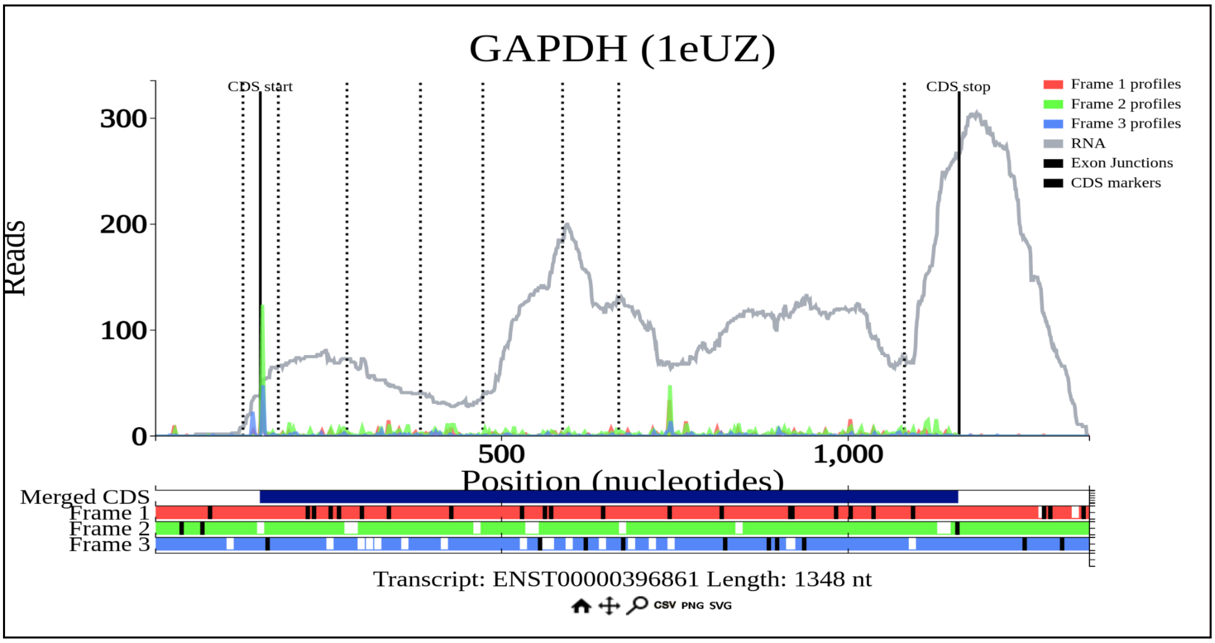
○ Display nucleotide sequence

○ Display mismatches

Now you can see that the main reading frame (CDS reading frame) is also green as in the previous example.

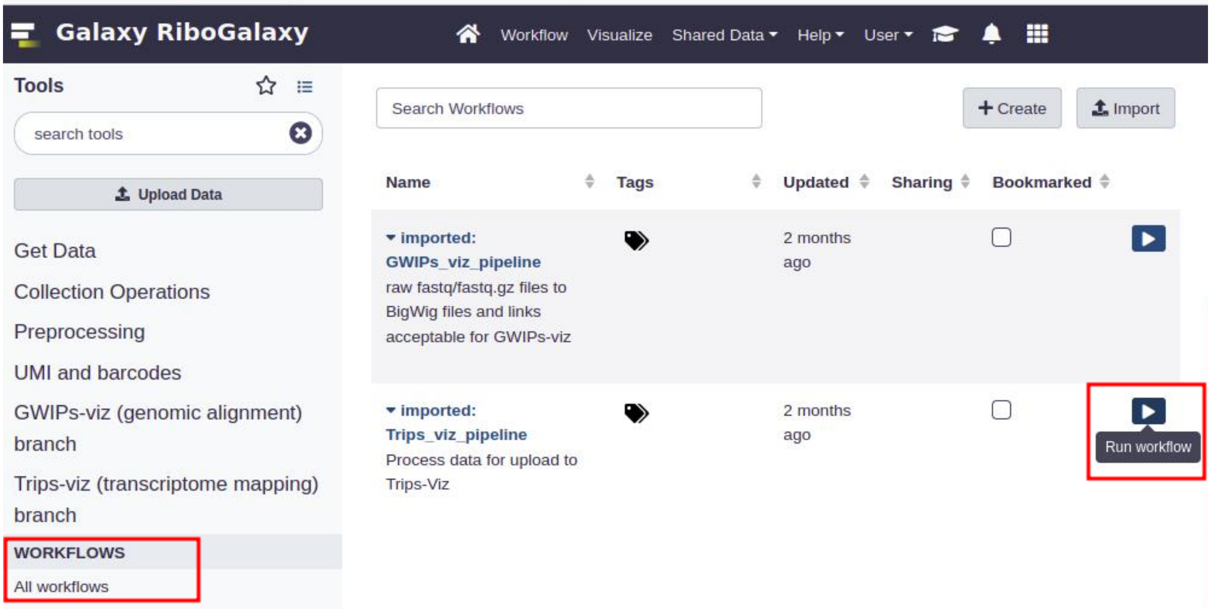


Similarly, we can add RNA-seq data to the plot:



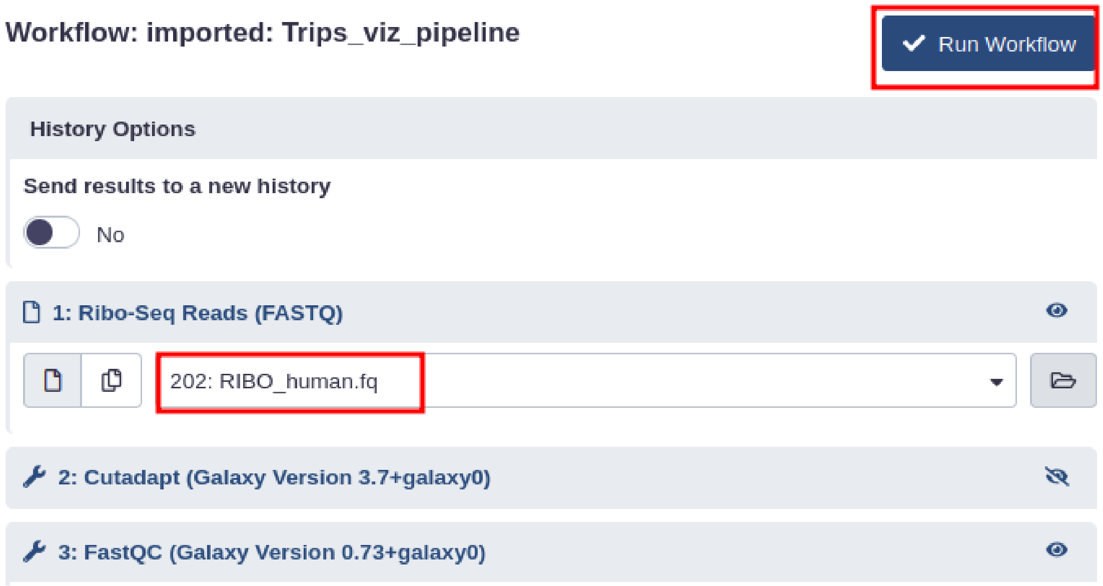
Workflow for Ribo-seq sample

For the Ribo-seq sample, we also created a Workflow which was tailored (all the parameters are set) to process the particular read structure from raw data to SQLITE file that can be studied in Trips-viz. First, you need to choose ‘All workflows’ on the left panel and then select ‘Run workflow’ for Trips_viz_pipleine.



Name	Tags	Updated	Sharing	Bookmarked
imported: GWIPs_viz_pipeline raw fastq/fastq.gz files to BigWig files and links acceptable for GWIPs-viz		2 months ago		□
imported: Trips_viz_pipeline Process data for upload to Trips-Viz		2 months ago		□

Next you need to make sure that the input file is correct and hit ‘Run Workflow’.



Workflow: imported: Trips_viz_pipeline

History Options

Send results to a new history

No

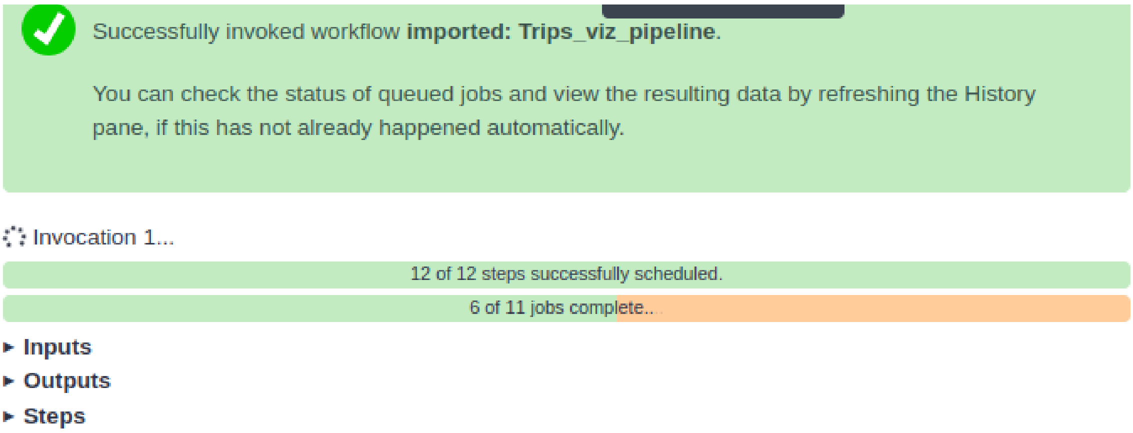
1: Ribo-Seq Reads (FASTQ)

202: RIBO_human.fq

2: Cutadapt (Galaxy Version 3.7+galaxy0)

3: FastQC (Galaxy Version 0.73+galaxy0)

Now can can just monitor the progress of the Workflow:



Successfully invoked workflow **imported: Trips_viz_pipeline.**

You can check the status of queued jobs and view the resulting data by refreshing the History pane, if this has not already happened automatically.

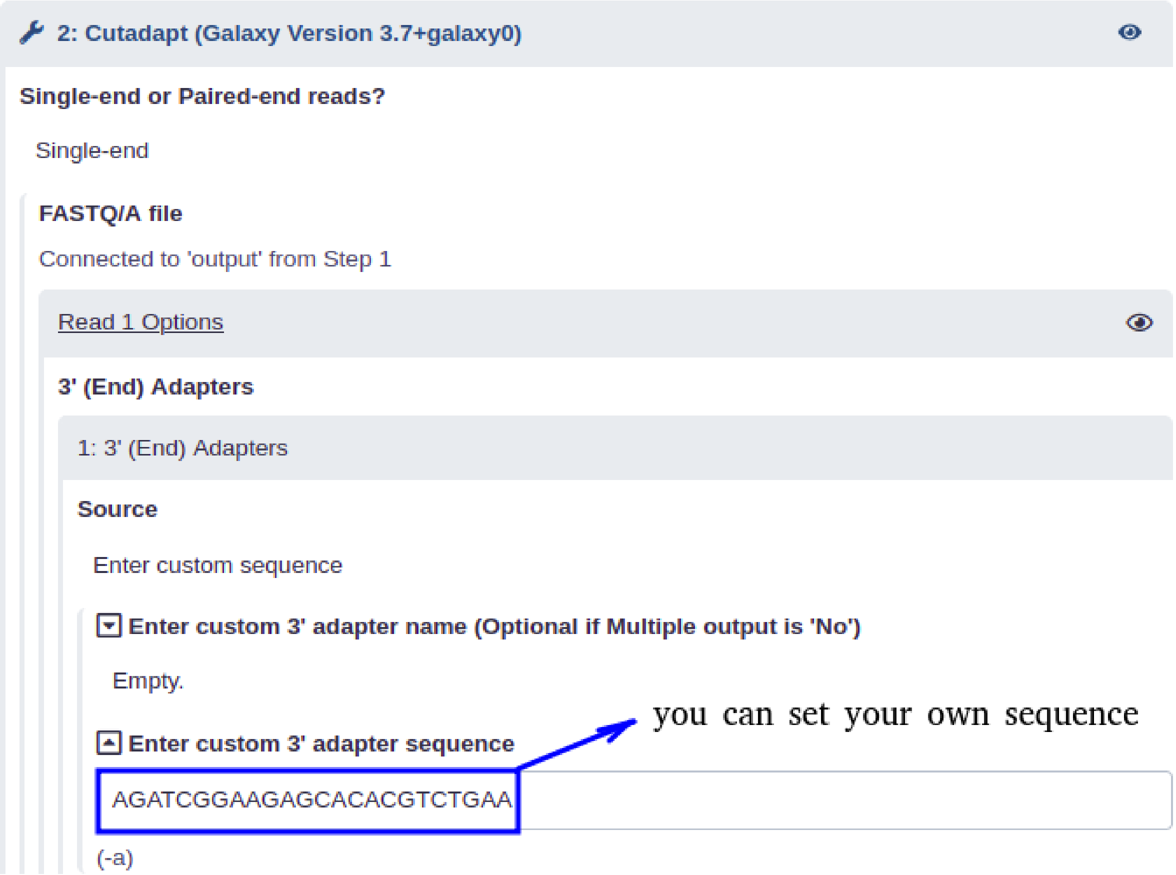
Invocation 1...

12 of 12 steps successfully scheduled.

6 of 11 jobs complete..

Once it’s finished, you will have SQLITE file ready for import to Trips-viz.

The workflow parameters are customisable:



2: Cutadapt (Galaxy Version 3.7+galaxy0)

Single-end or Paired-end reads?

Single-end

FASTQ/A file

Connected to 'output' from Step 1

Read 1 Options

3' (End) Adapters

1: 3' (End) Adapters

Source

Enter custom sequence

Enter custom 3' adapter name (Optional if Multiple output is 'No')

Empty.

Enter custom 3' adapter sequence

AGATCGGAAGAGCACACGTCTGAA

(-a)

you can set your own sequence

You can also create your own workflows.