Galaxy Workflow for RNAseq

David Innes

2022-01-05

Contents

1 About **5**

Workflow Step 1 5

Workflow Step 2 5

User profile 5

2 Upload and Prepare Data **7**

2.1 Import fastq files 7

2.2 Rename Files 9

2.3 Import ENSEMBL files 15

3 Workflow - Step 1 **19**

3.1 Find workflow 19

3.2 Run workflow 19

3.3 View results 22

4 Step 2 **25**

4.1 Import workflow 25

4.2 Run Workflow 25

4.3 Inspect Report 25

4.4 Save data 25

5 Mutliple Lane files **27**

Chapter 1

About

These instructions show how to use pre-made workflows on usegalaxy.org.au to analyse single stranded RNAseq files outputted from an Illumina system.

It is broken into 2 workflows. A workflow is a set of tools on galaxy organised together to do certain tasks. These workflows are shareable and are included here.

Workflow Step 1

The first workflow is designed to rename files, run quality check for each file and join any files together that are from the same sample (see 5). It also calculates some other values required for Step 2, such as the length of bp of the transcripts.

Once the output is checked (FastQC via a MultiQC report), the concatenated files can be used as input to Step 2.

Workflow Step 2

Step 2

User profile

You will need to be signed in to complete this analysis. Make sure you use university email address when setting up profile. Registered users with an Australian research institute are allocated 600GB, otherwise only 100GB is allocated.

Chapter 2

Upload and Prepare Data

2.1 Import fastq files

2.1.1 Upload

Firstly we upload the files to the Galaxy History. Follow figures below.

Figure 2.1: Click Upload Data button

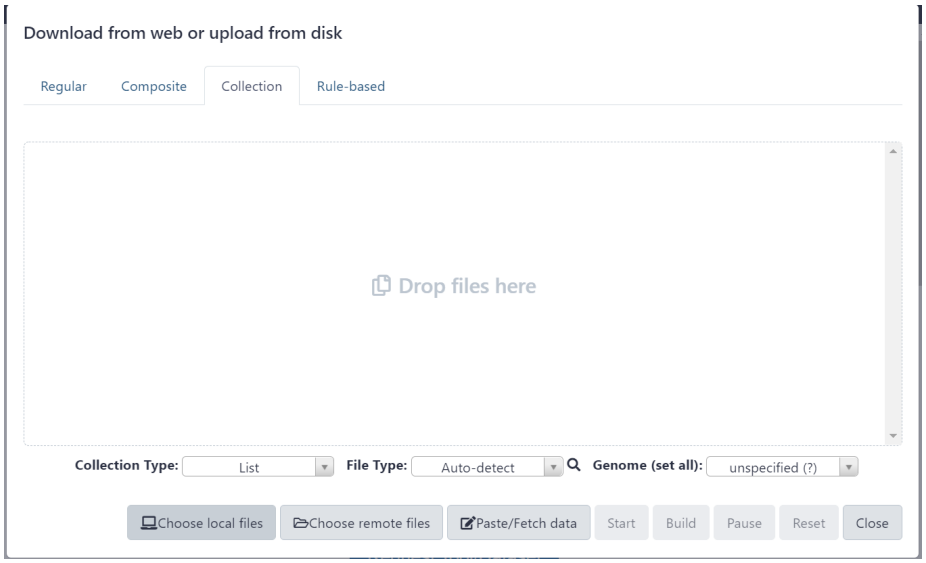


Figure 2.2: Select 'Collection' from the top ribbon and click 'Choose local files' button

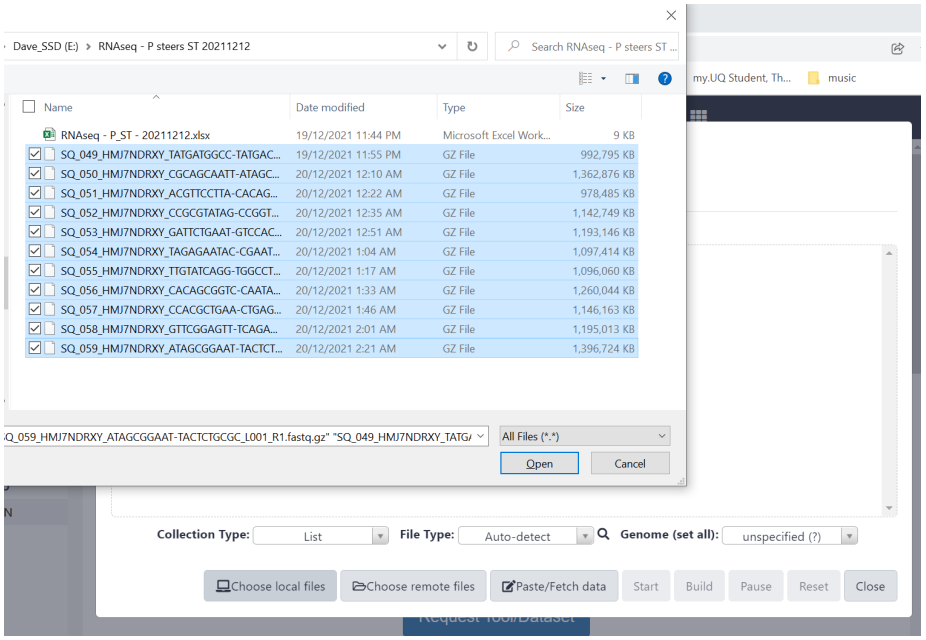


Figure 2.3: Navigate to the files to upload and highlight them all using Shift or Ctrl keys to help, then click 'Open'

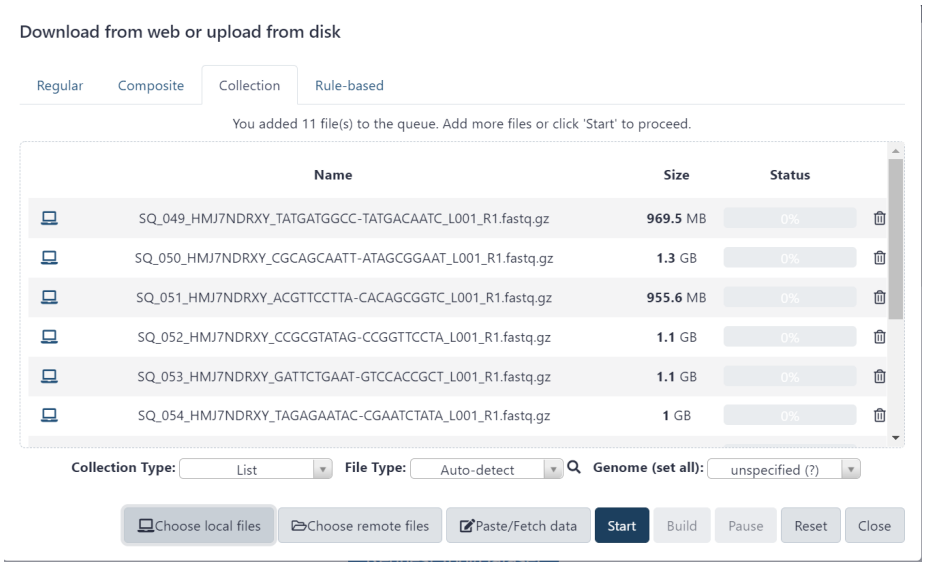


Figure 2.4: Click 'Start' button to begin uploading files. Keep this website open until it is finished.

2.1.2 Add files to a 'collection'

Once the files have uploaded they will appear in the History pane. Next, we add them to a 'collection', which is basically just a list of files. It allows all files to be parsed through a workflow one by one. For example, if a tool was used on a collection, then each item in the list/collection would invoke its own job while all output is kept within a collection in the history.

2.2 Rename Files

This step renames the files to more meaningful and user friendly names.

To do this, a new file is imported with the old names and the new names. This should be a tab delimited .txt file with 2 columns of data.

The 1st column has original name (which is likely to be filename) and the 2nd column has new names. E.g. original filename might be `SQ_049_HMJ7NDRXY_TATGATGGCC-TATGACAATC_L001_R1.fastq.gz` whereas new name could be changed to include treatment information such as `ST_811_P5_L001`. It is important that the "_L" number is included at the end, even if there is not multiple lanes per sample. This will be dealt with by the "Step 1" workflow. See section 2.2.1.

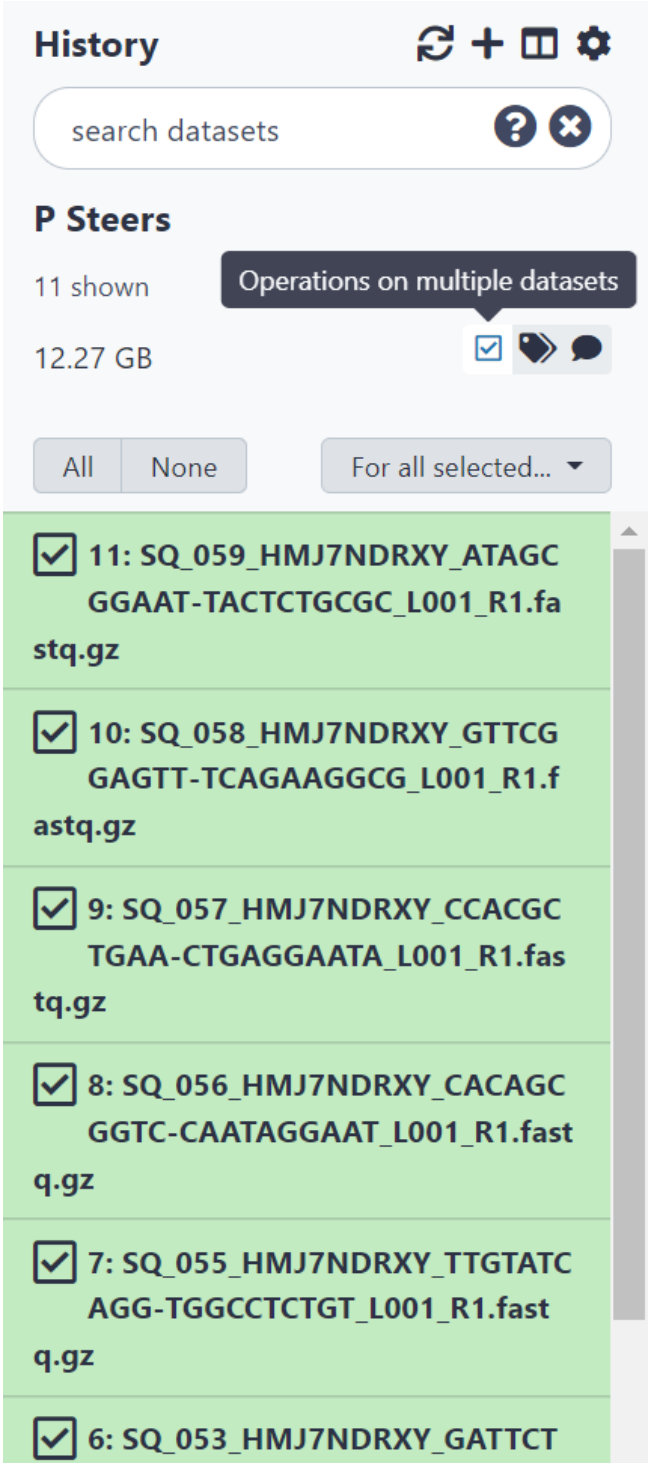


Figure 2.5: Select 'Operations on multiple datasets' then select all items to add to collection

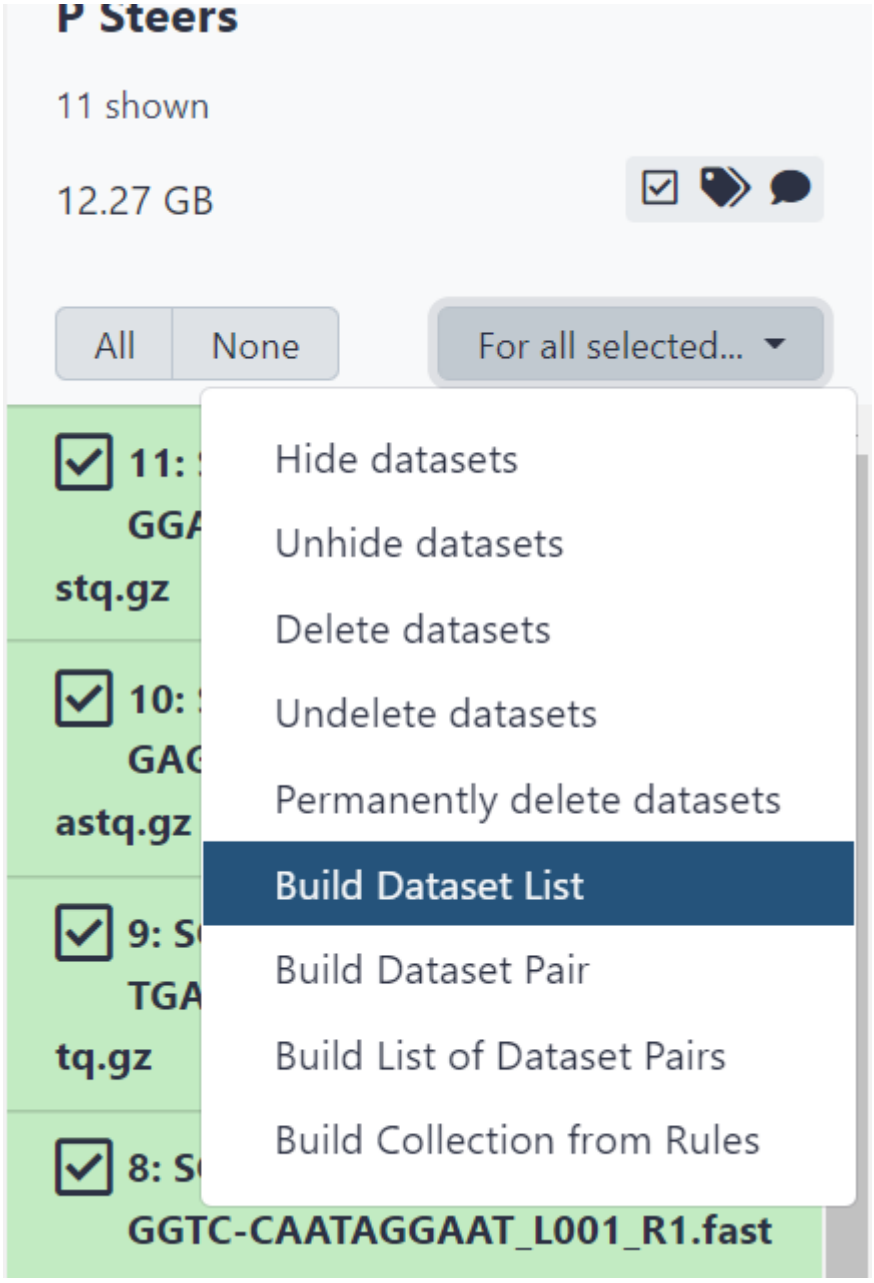


Figure 2.6: Click 'For all selected' then 'Build dataset list'

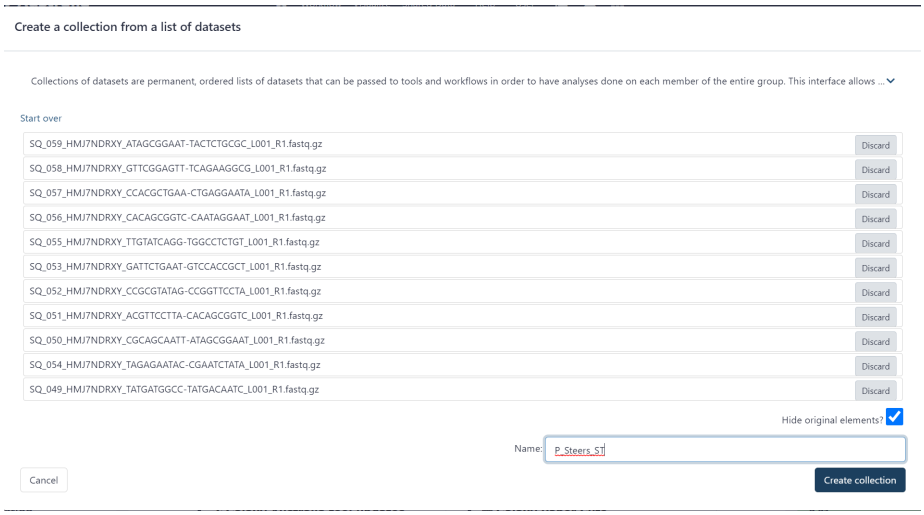


Figure 2.7: Type a name for the list and click 'Create collection' button

Use the 'upload data' feature of galaxy to import the .txt file to to the History. It should look something like 2.9.

This .txt file can be made in Excel, or it can be done much quicker using a simple R script. The following is an example script:

```
# list file names and create table with new file names for use in galaxy's "Relabel li
library(tidyverse)
library(data.table)

#list the names of the fastq files in the working directory
file_list <- list.files(path = "E:/RNAseq - P steers ST 20211212/", pattern = ".fastq"

#list .csv files to select the .csv file that contains sample information
files_csv <- list.files(path = "E:/RNAseq - P steers ST 20211212/", pattern = ".csv")

#selects the correct .csv file, in this case there is only 1 file anyway, but the corr
selected_file <- files_csv[1]
message(paste("Selected file -", selected_file))

#import the selected file, making sure no columns have the same name
sample_IDs <- fread(file = paste0("E:/RNAseq - P steers ST 20211212/",selected_file)) %
  as_tibble(.name_repair = "unique") %>%
  mutate(`Steer ID` = as.character(`Steer ID`))


#As the file names have a set structure, it can be split by the _ character
```

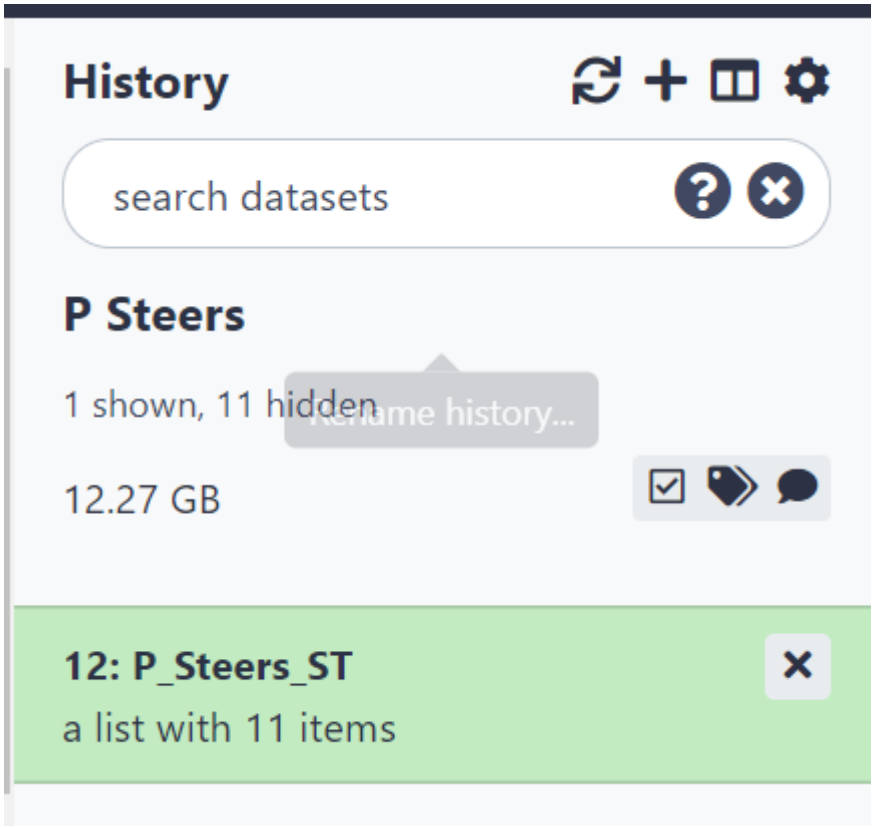
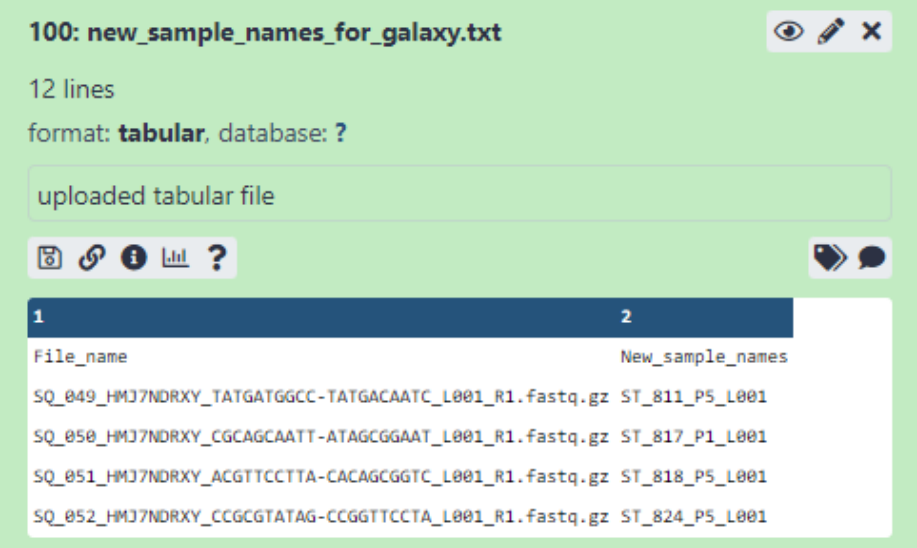


Figure 2.8: All files should now be in a collection in the History pane.



100: new_sample_names_for_galaxy.txt

12 lines

format: **tabular**, database: **?**

uploaded tabular file

1	2
File_name	New_sample_names
SQ_049_HMJ7NDRXY_TATGATGGCC-TATGACAATC_L001_R1.fastq.gz	ST_811_P5_L001
SQ_050_HMJ7NDRXY_CGCAGCAATT-ATAGCGGAAT_L001_R1.fastq.gz	ST_817_P1_L001
SQ_051_HMJ7NDRXY_ACGTTCCTTA-CACAGCGGTC_L001_R1.fastq.gz	ST_818_P5_L001
SQ_052_HMJ7NDRXY_CCGCGTATAG-CCGGTTCCTA_L001_R1.fastq.gz	ST_824_P5_L001

Figure 2.9: Example of text file imported to History with original and new filenames

```
df_file_names_split <- str_split(file_list, pattern = "_", simplify = TRUE) %>% as.data

#In this example, the 'key' column is the RNAseq_ID column. So this will be modified t
sample_IDs <-
  sample_IDs %>%
  mutate(RNAseq_ID_2 = str_remove(RNAseq_ID, "SQ_"),
         Treatment2 = str_remove(Treatment, '-'))

#join the sampleID annotations to this dataframe, concat required columns for new name
new_names <- df_file_names_split %>%
  left_join(sample_IDs, by = c("V2" = "RNAseq_ID_2")) %>%
  mutate(new_sample_names = str_c(`Tissue type`,`Steer ID`, Treatment2, V5, sep = "_"))
  magrittr::use_series(new_sample_names) #select only the required column as a list (n

#create new table with only old file names and new names
new_name_table <-
  data.frame("File_name" = file_list, "New_sample_names" = new_names)

#export to working directory, as tab delim
fwrite(new_name_table, file = "E:/RNAseq - P steers ST 20211212/new_sample_names_for_g
```

2.2.1 A note on multiple lanes

The output from Illumina sequencing is sometimes provided in multiple files, each corresponding to a 'Lane' on the sequencer. It would be easier to ask the lab to provide the output as a single file, which can be computed using the `--no-lane-splitting` option from Illumina's `bcl2fastq` program. However, it can also be handled in Galaxy. If there are multiple files, it is best practice to run FastQC on each individual file, as there is a chance that one file could be corrupt or you may identify a bias for one particular 'Lane'. If they are ok, then these files can be concatenated together before proceeding with all further steps.

This is described further in 5. Make sure the new names generated in 2.2 have a format that includes "_L". The protocol relies on there being a '_L' in the name for it to find the lane number. The rest of the name before the '_L' should be the same.

E.g. the following three files would be concatenated together by the "Step 1" workflow:

- 2139_Stage 2_Fast_L008
- 2139_Stage 2_Fast_L007
- 2139_Stage 2_Fast_L006

2.3 Import ENSEMBL files

In this step we need to import the required fastq file for sequence alignment and a gtf file for gene annotation. These can be uploaded directly to Galaxy via a URL. To find the required files, navigate to http://ftp.ensembl.org in a browser.

For sheep (ovis_aries) these might be:

- http://ftp.ensembl.org/pub/release-100/fasta/ovis_aries/dna/Ovis_aries.Oar_v3.1.dna.toplevel.fa.gz
- http://ftp.ensembl.org/pub/release-100/gtf/ovis_aries/Ovis_aries.Oar_v3.1.100.gtf.gz

For cattle (bos_taurus) these might be:

- http://ftp.ensembl.org/pub/release-100/fasta/bos_taurus/dna/Bos_taurus.ARS-UCD1.2.dna.toplevel.fa.gz
- http://ftp.ensembl.org/pub/release-100/gtf/bos_taurus/Bos_taurus.ARS-UCD1.2.100.gtf.gz

Note that unmasked files are used here (i.e. use files without `_rm` or `_sm`).

To upload these to Galaxy, use the 'Paste/Fetch data' button in the 'Upload data' dialogue box on Galaxy.

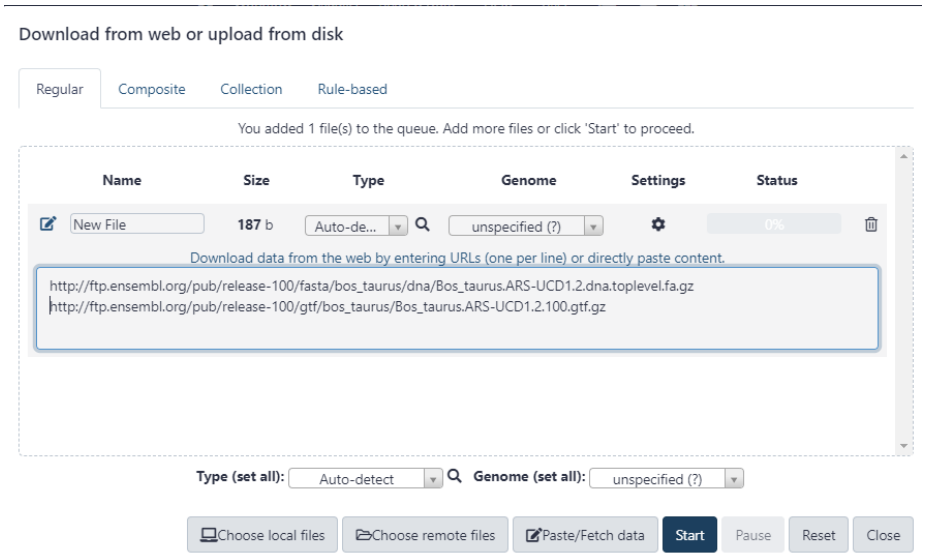


Figure 2.10: Paste each URL on a new line to upload directly from ENSEMBL to Galaxy.

Once executed, 2 new files will appear in the History, each named as the URL entered. This can take some time to finish as they are large files.

2.3.1 Uncompress .gtf file

These files are actually .gz files, which means they are compressed. Normally, this is automatically handled by Galaxy but does not currently work for the .gtf file when using it with the `STAR Aligner` in this workflow. Therefore, use the tool https://usegalaxy.org.au/root?tool_id=CONVERTER_gz_to_uncompressed to uncompress the .gtf file before proceeding.

Once it is uncompressed it may not have the correct file format attributed to it in Galaxy. Fix this by using the auto-detect feature within the Datatypes ribbon of the Edit Attributes section.

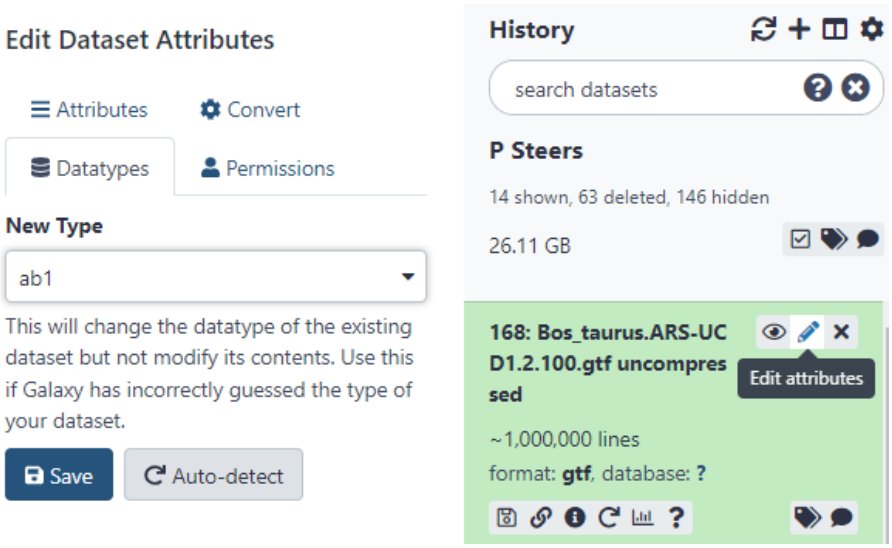


Figure 2.11: Click 'edit attributes' then navigate to 'Datatypes' to use the Auto-Detect. Otherwise manually set it to .gtf

Chapter 3

Workflow - Step 1

Once all of the data is uploaded to Galaxy it is time to run/invoke the workflows. The first workflow is relatively fast compared to Step 2, but is important for QC and to prepare data.

3.1 Find workflow

The workflow used here is published at: https://usegalaxy.org.au/u/dave-innes/w/rna-seq-step-1

This workflow will need to be imported into your personal galaxy profile. Navigate to: https://usegalaxy.org.au/workflows/import On this page you will be able to enter the workflow URL above. This will also be required for Step 2.

You can also navigate directly the workflow's URL and import it from there.

3.2 Run workflow

To view your workflows, click on the 'Workflow' link at the top of the page.

Once you've clicked 'Run Workflow' it will let you select the files it should use and show all of the steps it will complete. Select the appropriate files for 1 and 2. All other steps will not require user input.

This will invoke all steps and you will see them in the History.

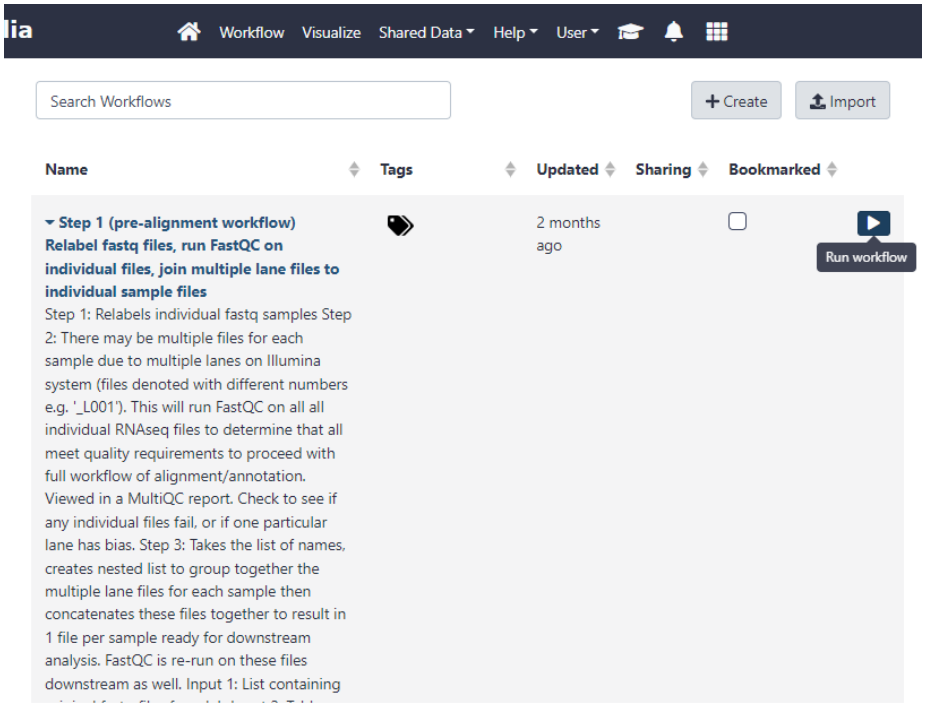


Figure 3.1: Click the Run Workflow button

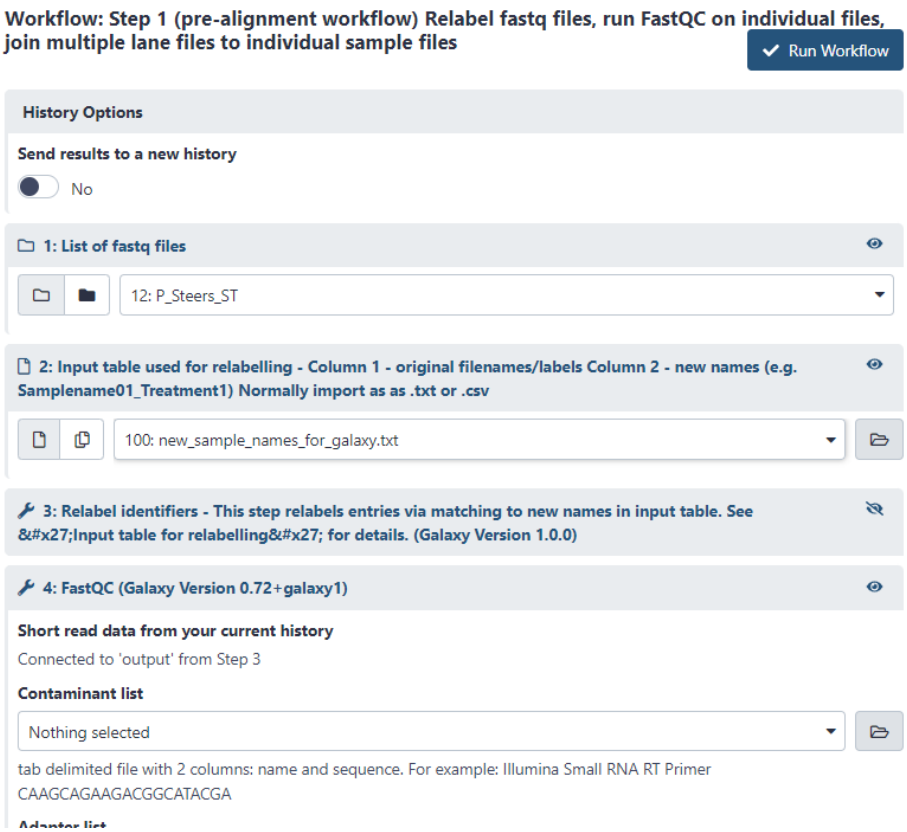


Figure 3.2: Select list of fastq files for number 1 and select the table of new names for number 2, then click 'Run Workflow'

3.3 View results

Once all steps are completed you will see the each part of the History has turned green.

The most interesting data will be found in the MultiQC webpage. It can be Viewed within galaxy but it may view better if downloaded and opened up in web browser.

There is plenty of help included in the report and it is worth noting that some sections will normally fail for RNAseq data. Although unlikely, it is also important to check if there are any major quality differences between lanes before proceeding (if multiple lanes exist).

Figure 3.3: Example output in history after running first workflow

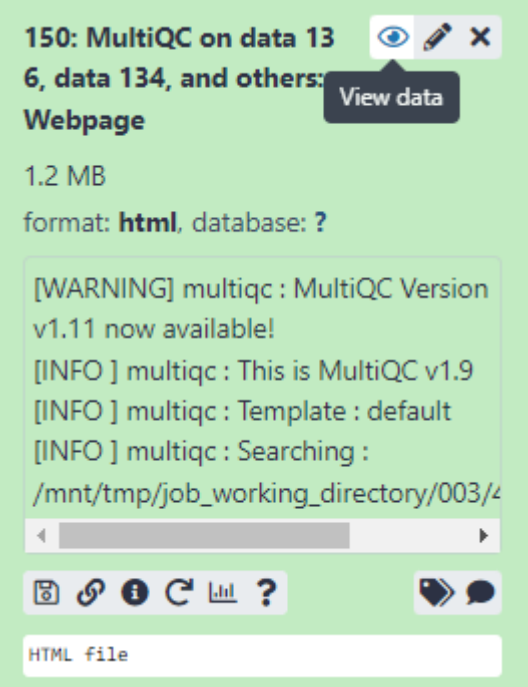


Figure 3.4: View the MultiQC webpage report

Chapter 4

Step 2

Step 2 completes multiple steps and reports output to one MultiQC report at the end, with various important outputs to the History. The most useful output from this workflow for DE analysis is the table of counts with rows as gene names and columns as sample names.

The details of the pipeline are visible by following the workflow URL. In short, it uses trimmo

4.1 Import workflow

https://usegalaxy.org.au/u/dave-innes/w/rna-seq-step-2

4.2 Run Workflow

There are 4 required inputs to this workflow. Set it up as follows:

This workflow will take a very long time to run. It runs on the Galaxy server and you can log in from any computer to see it's status. You're computer does not need to stay on.

4.3 Inspect Report

4.4 Save data

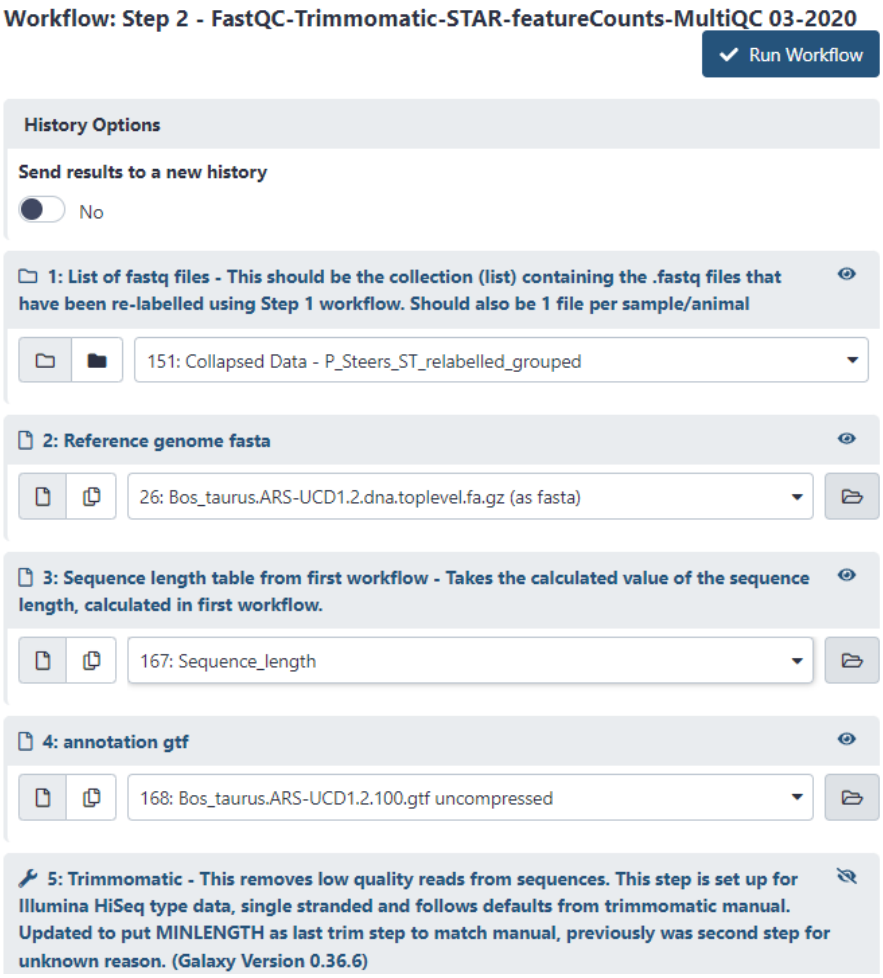


Figure 4.1: Setup Step 2 then press Run Workflow

Chapter 5

Mutliple Lane files

The output from Illumina sequencing is sometimes provided in multiple files, each corresponding to a 'Lane' on the sequencer. It would be easier to ask the lab to provide the output as a single file, which can be computed using the `--no-lane-splitting` option from Illumina's `bcl2fastq` program. However, it can also be handled in Galaxy.

If there are multiple files, it is best practice to run FastQC on each individual file, as there is a chance that one file could be corrupt or you may identify a bias for one particular 'Lane'. If they are ok, then these files can be concatenated together before proceeding with all further steps. There are 2 workflows that will be used in Galaxy, with the first designed to work with each individual lane file and the second requiring 1 file per sample. This can be done by following these steps:

1. 'Apply Rule to Collection' tool
2. Input collection is the list with the re-labelled data
3. Press Edit button
4. There should be 1 column titled 'A' with the names of the files (the re-labelled names set previously)

Make new columns with grouping information:

5. Press +Column -> Using a regular expression
6. Select 'Create column matching expression groups'
7. Paste the following code into the Regular Expression box: `(.*?)_L(.*)`
8. Set the number of groups to 2
9. Press Apply

Set columns as identifiers for grouping:

10. Press +Rules -> Add / Modify Column Definitions
11. Press +Add Definition -> List Identifiers
12. Select column B, then click on 'Assign another column' and select C
13. Press Apply
14. Press Save, then Execute the job

This outputs a nested list to the history. The number of items in the list should match the number of samples/animals, and then each sample in the list should contain the number of individual files (e.g. 2 or 3 files).

To then join this datasets together, the tool 'Collapse Collection' will work in the background to be the same as using the 'concatenate datasets tail-to-head' tool to concatenate the files individually. Although not clear in the tool's description, if it is provided a nested list, it will collapse the lowest level of groups together, in this case it is the Column C from above, which was the individual lanes. It should output a new list with the same number of files as the number of samples/animals. The names should be the names defined in Column B above: e.g. 2139_Stage 2_Fast:

1. Open Collapse Collection tool
2. Choose to use a dataset collection, not an individual file
3. Select the nested list from above
4. Execute with all other default settings