Galaxy Workflow for RNAseq

David Innes

2022-01-06

Contents

1 About 5

Workflow Step 1 5

Workflow Step 2 5

User profile 6

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 Workflow - Step 2 25

4.1 Import workflow 25

4.2 Run Workflow 25

4.3 Save data 25

4.4 Inspect Report 27

5 More notes on mutliple lane files 29

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.

This assumes some familiarity with Galaxy, but also aims to help beginners complete this specific use-case. Other tutorials are more generalised and are better equipped to help users learn each stop of RNAseq pipelines.

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. Therefore many steps are automated and not discussed, but the details can be viewed when exploring these workflows within Galaxy.

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

This workflow re-runs fastQC with the input files, and uses trimmomatic, RNA STAR Aligner and featureCounts to produce the required outputs for downstream analysis. It also reports each step to MultiQC webpage.

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.

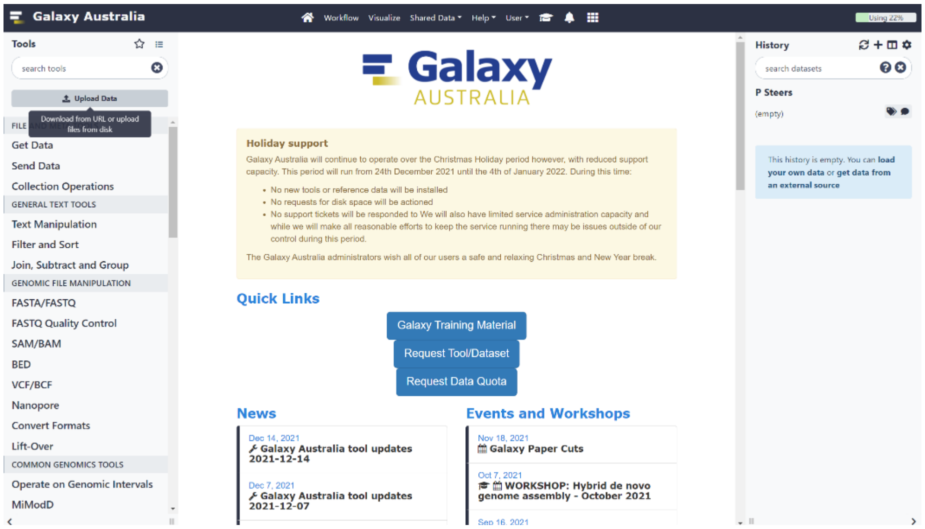
Chapter 2

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.


Galaxy Australia
Workflow
Visualize
Shared Data
Help
User
Using 22%
Tools
search tools
Upload Data
Download from URL or upload files from disk
FILE
Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
Convert Formats
Lift-Over
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
MiModD
Galaxy
AUSTRALIA
Quick Links
Galaxy Training Material
Request Tool/Dataset
Request Data Quota
History
search datasets
P Steers
(empty)

Holiday support

Galaxy Australia will continue to operate over the Christmas Holiday period however, with reduced support capacity. This period will run from 24th December 2021 until the 4th of January 2022. During this time:

- No new tools or reference data will be installed
- No requests for disk space will be actioned
- No support tickets will be responded to We will also have limited service administration capacity and while we will make all reasonable efforts to keep the service running there may be issues outside of our control during this period.

The Galaxy Australia administrators wish all of our users a safe and relaxing Christmas and New Year break.

News

Dec 14, 2021
Galaxy Australia tool updates 2021-12-14

Dec 7, 2021
Galaxy Australia tool updates 2021-12-07

Events and Workshops

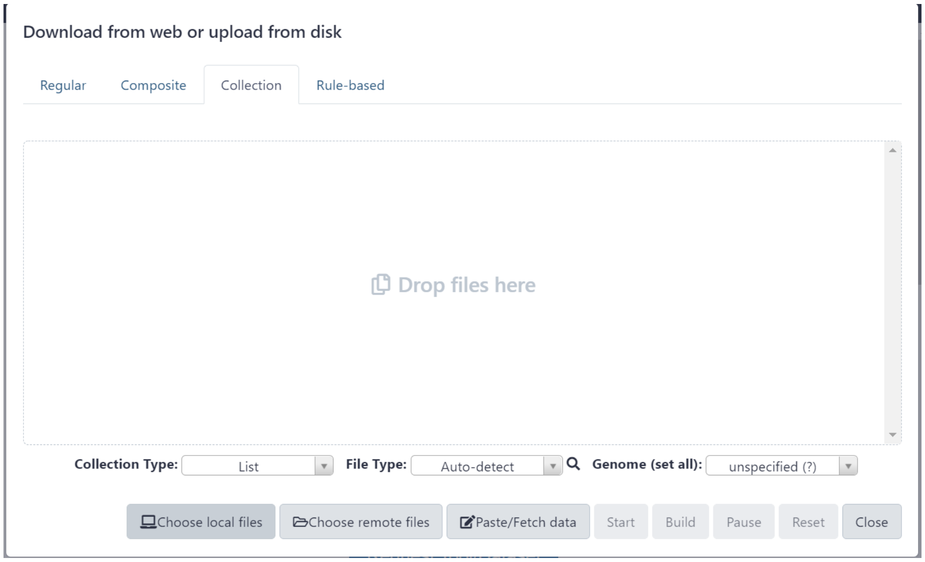
Nov 18, 2021
Galaxy Paper Cuts

Oct 7, 2021
WORKSHOP: Hybrid de novo genome assembly - October 2021

Sep 16, 2021

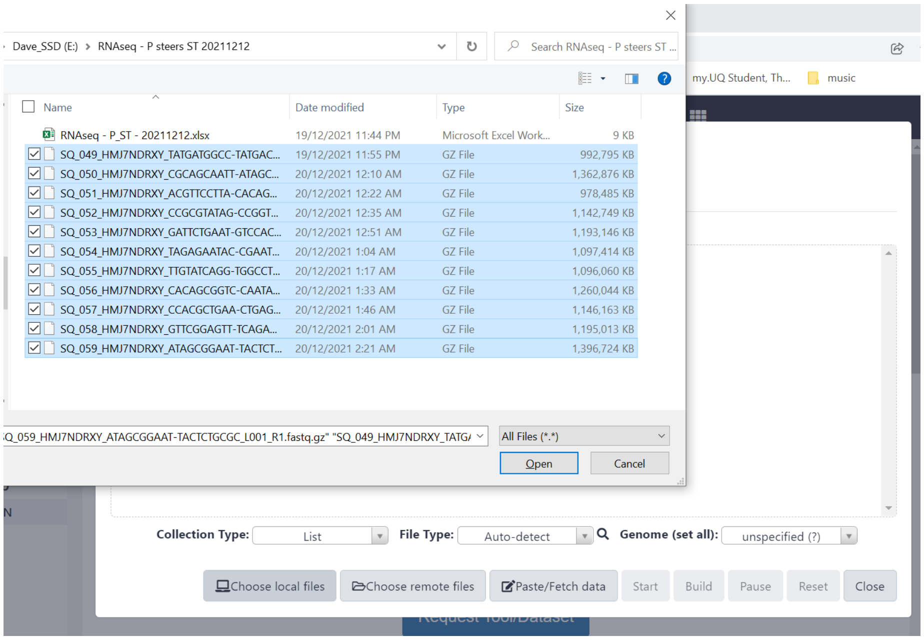
This history is empty. You can **load your own data** or **get data from an external source**

Figure 2.1: Click Upload Data button


Download from web or upload from disk
Regular
Composite
Collection
Rule-based
Drop files here
Choose local files
Choose remote files
Paste/Fetch data
Start
Build
Pause
Reset
Close

Collection Type: List **File Type:** Auto-detect **Genome (set all):** unspecified (?)

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


Dave_SSD (E:) › RNAseq - P steers ST 20211212
Search RNAseq - P steers ST ...
my.UQ Student, Th...
music
Open
Cancel
Choose local files
Choose remote files
Paste/Fetch data
Start
Build
Pause
Reset
Close

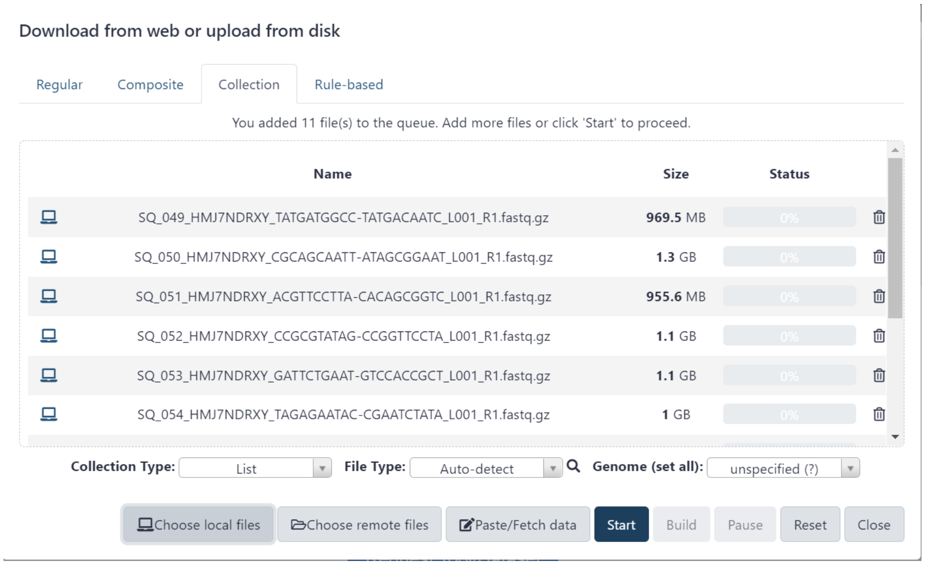
	Name	Date modified	Type	Size
	RNAseq - P_ST - 20211212.xlsx	19/12/2021 11:44 PM	Microsoft Excel Work...	9 KB
☑	SQ_049_HMJ7NDRXY_TATGATGGCC-TATGAC...	19/12/2021 11:55 PM	GZ File	992,795 KB
☑	SQ_050_HMJ7NDRXY_CGCAGCAATT-ATAGC...	20/12/2021 12:10 AM	GZ File	1,362,876 KB
☑	SQ_051_HMJ7NDRXY_ACGTTCCTTA-CACAG...	20/12/2021 12:22 AM	GZ File	978,485 KB
☑	SQ_052_HMJ7NDRXY_CCGCGTATAG-CCGGT...	20/12/2021 12:35 AM	GZ File	1,142,749 KB
☑	SQ_053_HMJ7NDRXY_GATTCTGAAT-GTCCAC...	20/12/2021 12:51 AM	GZ File	1,193,146 KB
☑	SQ_054_HMJ7NDRXY_TAGAGAATAC-CGAAT...	20/12/2021 1:04 AM	GZ File	1,097,414 KB
☑	SQ_055_HMJ7NDRXY_TTGTATCAGG-TGGCCT...	20/12/2021 1:17 AM	GZ File	1,096,060 KB
☑	SQ_056_HMJ7NDRXY_CACAGCGGTC-CAATA...	20/12/2021 1:33 AM	GZ File	1,260,044 KB
☑	SQ_057_HMJ7NDRXY_CCACGCTGAA-CTGAG...	20/12/2021 1:46 AM	GZ File	1,146,163 KB
☑	SQ_058_HMJ7NDRXY_GTTCGGAGTT-TCAGA...	20/12/2021 2:01 AM	GZ File	1,195,013 KB
☑	SQ_059_HMJ7NDRXY_ATAGCGGAAT-TACTCT...	20/12/2021 2:21 AM	GZ File	1,396,724 KB

Q_059_HMJ7NDRXY_ATAGCGGAAT-TACTCTGCGC_L001_R1.fastq.gz" "SQ_049_HMJ7NDRXY_TATG/

All Files (*.*)

Collection Type: List **File Type:** Auto-detect **Genome (set all):** unspecified (?)

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



Download from web or upload from disk

You added 11 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Status
SQ_049_HMJ7NDRXY_TATGATGGCC-TATGACAATC_L001_R1.fastq.gz	**969.5** MB	0%
SQ_050_HMJ7NDRXY_CGCAGCAATT-ATAGCGGAAT_L001_R1.fastq.gz	**1.3** GB	0%
SQ_051_HMJ7NDRXY_ACGTTCCTTA-CACAGCGGTC_L001_R1.fastq.gz	**955.6** MB	0%
SQ_052_HMJ7NDRXY_CCGCGTATAG-CCGGTTCCTA_L001_R1.fastq.gz	**1.1** GB	0%
SQ_053_HMJ7NDRXY_GATTCTGAAT-GTCCACCGCT_L001_R1.fastq.gz	**1.1** GB	0%
SQ_054_HMJ7NDRXY_TAGAGAATAC-CGAATCTATA_L001_R1.fastq.gz	**1** GB	0%

Collection Type: List **File Type:** Auto-detect **Genome (set all):** unspecified (?)

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'

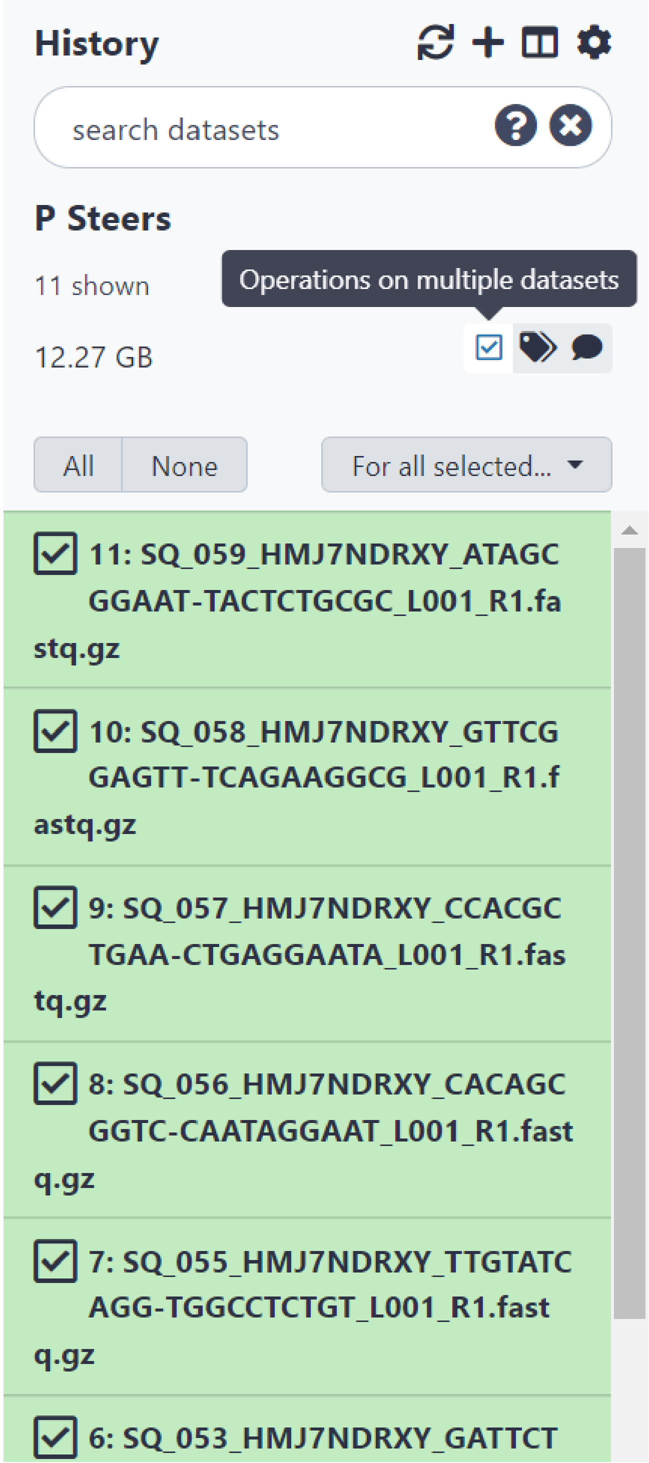
Once the files have uploaded they will appear in the History pane. Next, we add them to a 'collection', which is basically just a list of files. It allows all files to be parsed through a workflow one by one. For example, if a tool was used on a collection, then each item in the list/collection would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


History
search datasets
P Steers
11 shown
Operations on multiple datasets
12.27 GB
All
None
For all selected...

☑ 11: SQ_059_HMJ7NDRXY_ATAGCGGAAT-TACTCTGCGC_L001_R1.fastq.gz

☑ 10: SQ_058_HMJ7NDRXY_GTTCGGAGTT-TCAGAAGGCG_L001_R1.fastq.gz

☑ 9: SQ_057_HMJ7NDRXY_CCACGCTGAA-CTGAGGAATA_L001_R1.fastq.gz

☑ 8: SQ_056_HMJ7NDRXY_CACAGCGGTC-CAATAGGAAT_L001_R1.fastq.gz

☑ 7: SQ_055_HMJ7NDRXY_TTGTATCAGG-TGGCCTCTGT_L001_R1.fastq.gz

☑ 6: SQ_053_HMJ7NDRXY_GATTCT

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

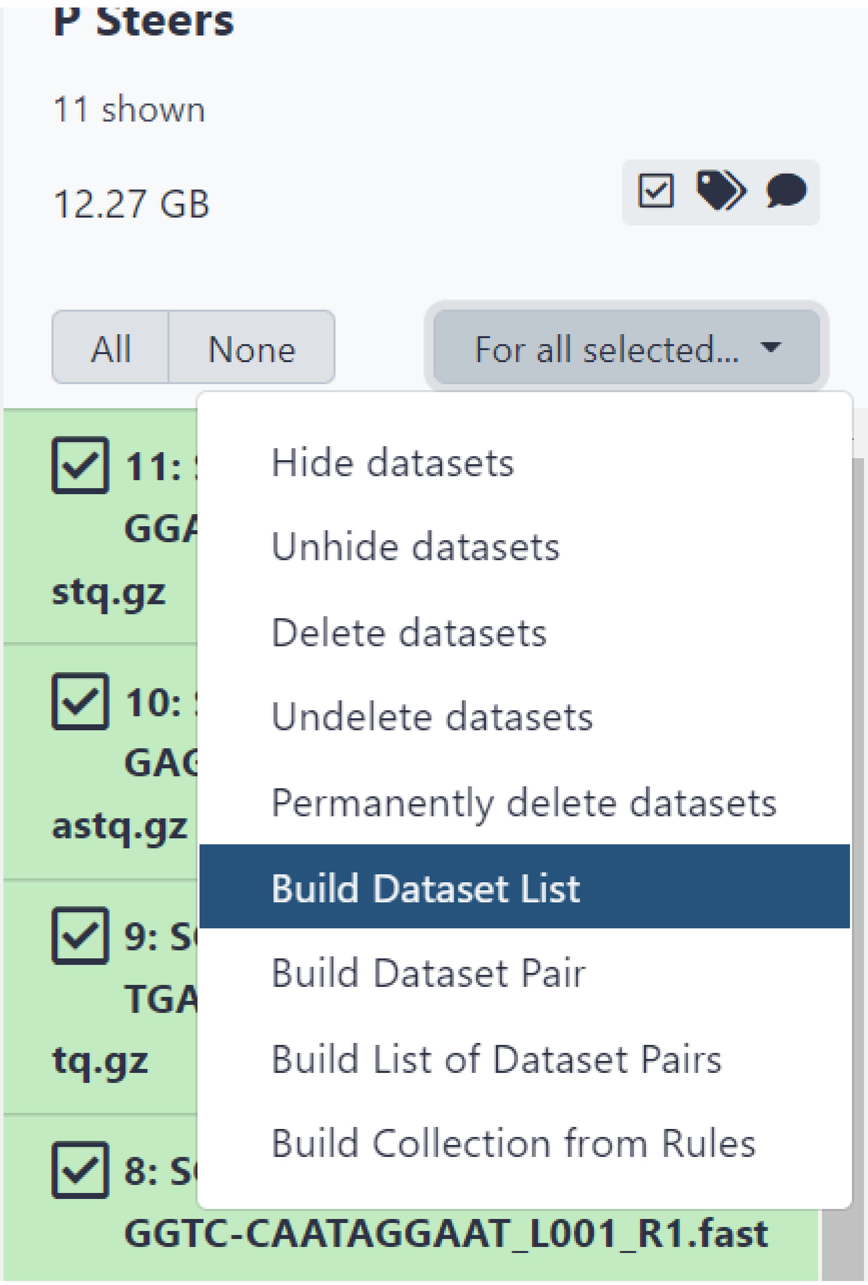
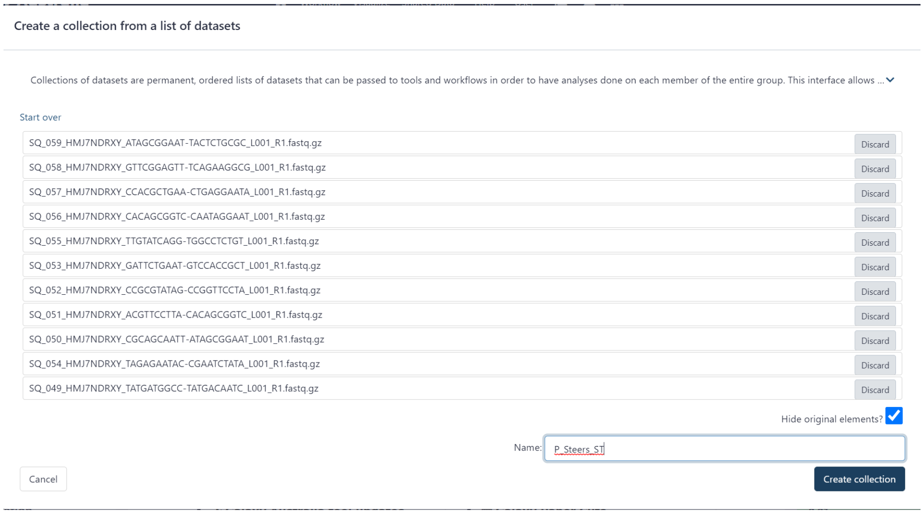

P Steers
11 shown
12.27 GB
☑
All
None
For all selected...
Hide datasets
Unhide datasets
Delete datasets
Undelete datasets
Permanently delete datasets
Build Dataset List
Build Dataset Pair
Build List of Dataset Pairs
Build Collection from Rules
☑ 11:
GGA
stq.gz
☑ 10:
GAG
astq.gz
☑ 9: S
TGA
tq.gz
☑ 8: S
GGTC-CAATAGGAAT_L001_R1.fast

Figure 2.6: Click ‘For all selected’ then ’Build dataset list’


Create a collection from a list of datasets

Collections of datasets are permanent, ordered lists of datasets that can be passed to tools and workflows in order to have analyses done on each member of the entire group. This interface allows ...

SQ_059_HMJ7NDRXY_ATAGCGGAAT-TACTCTGCGC_L001_R1.fastq.gz	Discard
SQ_058_HMJ7NDRXY_GTTCGGAGTT-TCAGAAGGCG_L001_R1.fastq.gz	Discard
SQ_057_HMJ7NDRXY_CCACGCTGAA-CTGAGGAATA_L001_R1.fastq.gz	Discard
SQ_056_HMJ7NDRXY_CACAGCGGTC-CAATAGGAAT_L001_R1.fastq.gz	Discard
SQ_055_HMJ7NDRXY_TTGTATCAGG-TGGCCTCTGT_L001_R1.fastq.gz	Discard
SQ_053_HMJ7NDRXY_GATTCTGAAT-GTCCACCGCT_L001_R1.fastq.gz	Discard
SQ_052_HMJ7NDRXY_CCGCGTATAG-CCGGTTCCTA_L001_R1.fastq.gz	Discard
SQ_051_HMJ7NDRXY_ACGTTCCTTA-CACAGCGGTC_L001_R1.fastq.gz	Discard
SQ_050_HMJ7NDRXY_CGCAGCAATT-ATAGCGGAAT_L001_R1.fastq.gz	Discard
SQ_054_HMJ7NDRXY_TAGAGAATAC-CGAATCTATA_L001_R1.fastq.gz	Discard
SQ_049_HMJ7NDRXY_TATGATGGCC-TATGACAATC_L001_R1.fastq.gz	Discard

Hide original elements? ☑

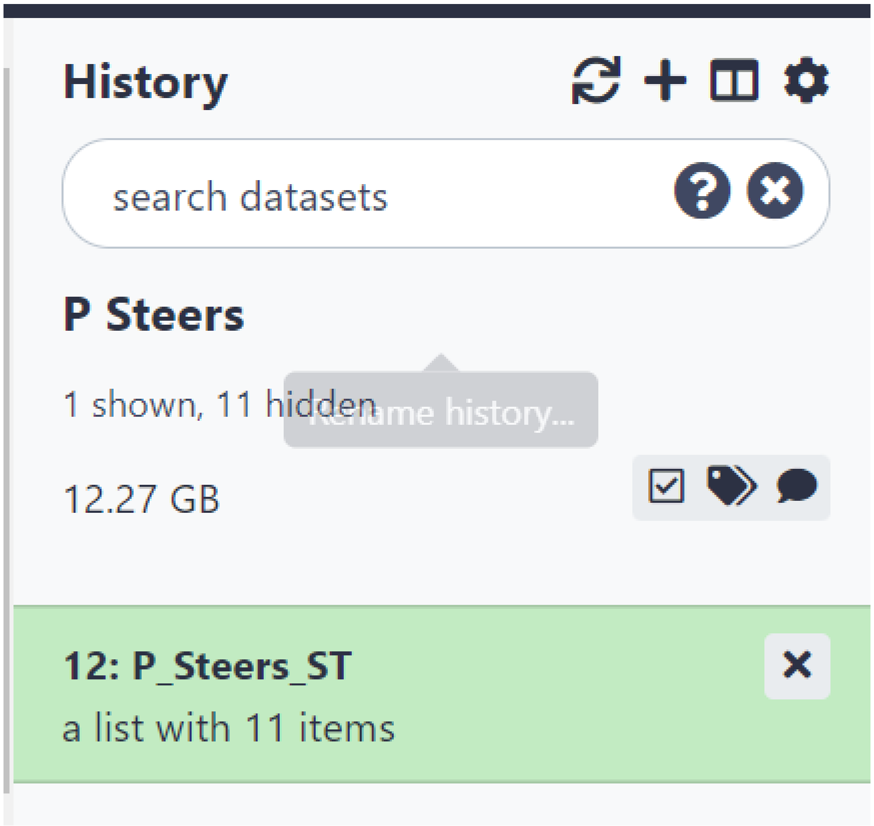
Name: P_Steers_ST

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
```


History

search datasets

P Steers

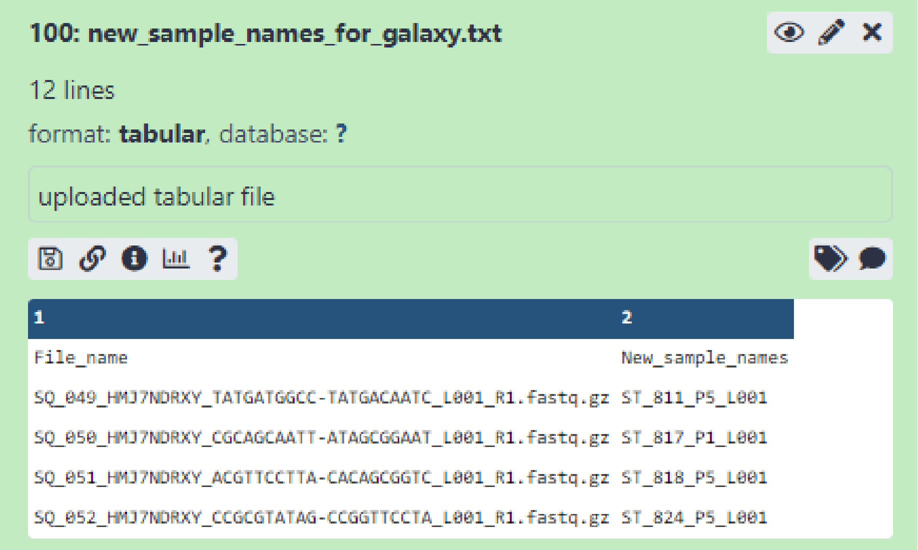
1 shown, 11 hidden

12.27 GB

12: P_Steers_ST

a list with 11 items

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.dat

#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 Chapter 5. Make sure the new names generated in Section 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:

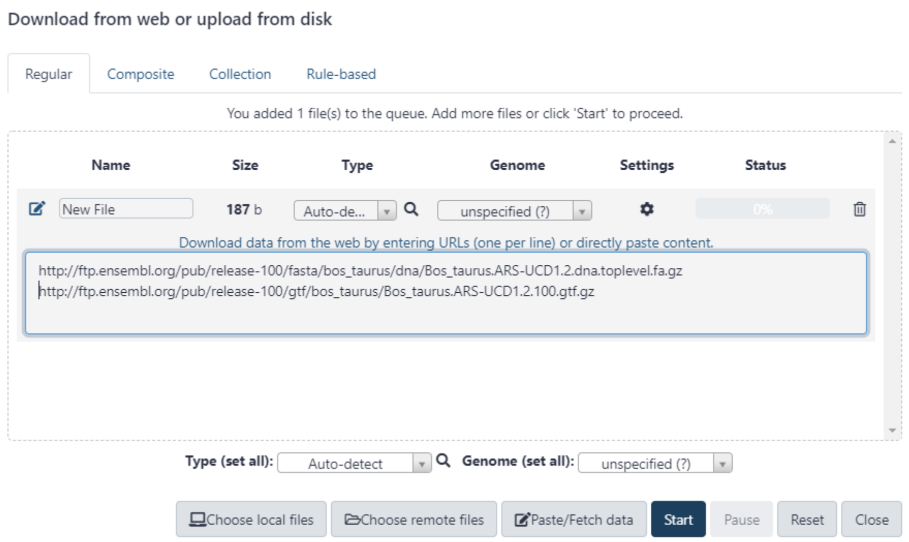
- http://ftp.ensembl.org/pub/release-100/fasta/ovis_aries/dna/Ovis_aries.Oar_v3.1.dna.toplevel.fa.gz
- http://ftp.ensembl.org/pub/release-100/gtf/ovis_aries/Ovis_aries.Oar_v3.1.100.gtf.gz

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.



Download from web or upload from disk

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
New File	187 b	Auto-de...	unspecified (?)		0%

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

```
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
```

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

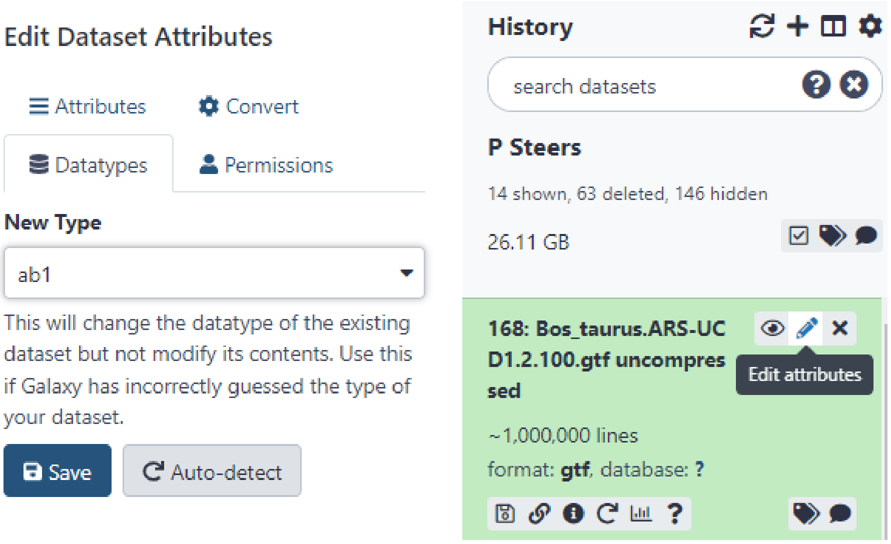
Figure 2.10: Paste each URL on a new line to upload directly from ENSEMBL to Galaxy.

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.



Edit Dataset Attributes

Attributes
Convert
Datatypes
Permissions

New Type: ab1

This will change the datatype of the existing dataset but not modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

Save
Auto-detect
History
search datasets

P Steers

14 shown, 63 deleted, 146 hidden

26.11 GB

168: Bos_taurus.ARS-UCD1.2.100.gtf uncompressed

Edit attributes

~1,000,000 lines

format: gtf, database: ?

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.

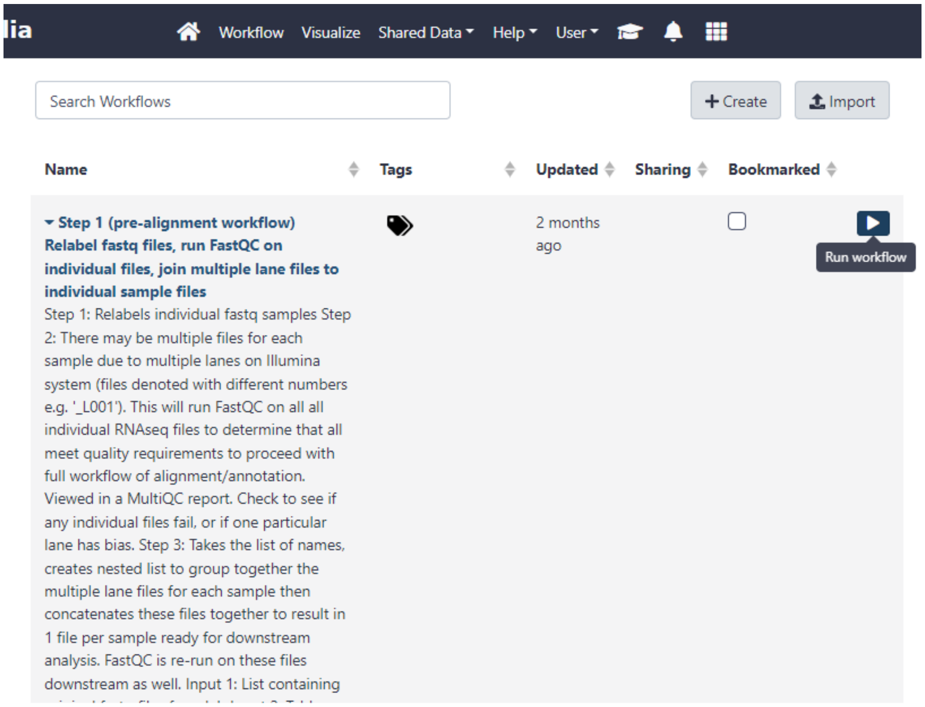
You can also navigate directly the workflow's URL and import it from there.

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.


lia
Workflow
Visualize
Shared Data
Help
User
Search Workflows
Create
Import
Name
Tags
Updated
Sharing
Bookmarked
Run workflow

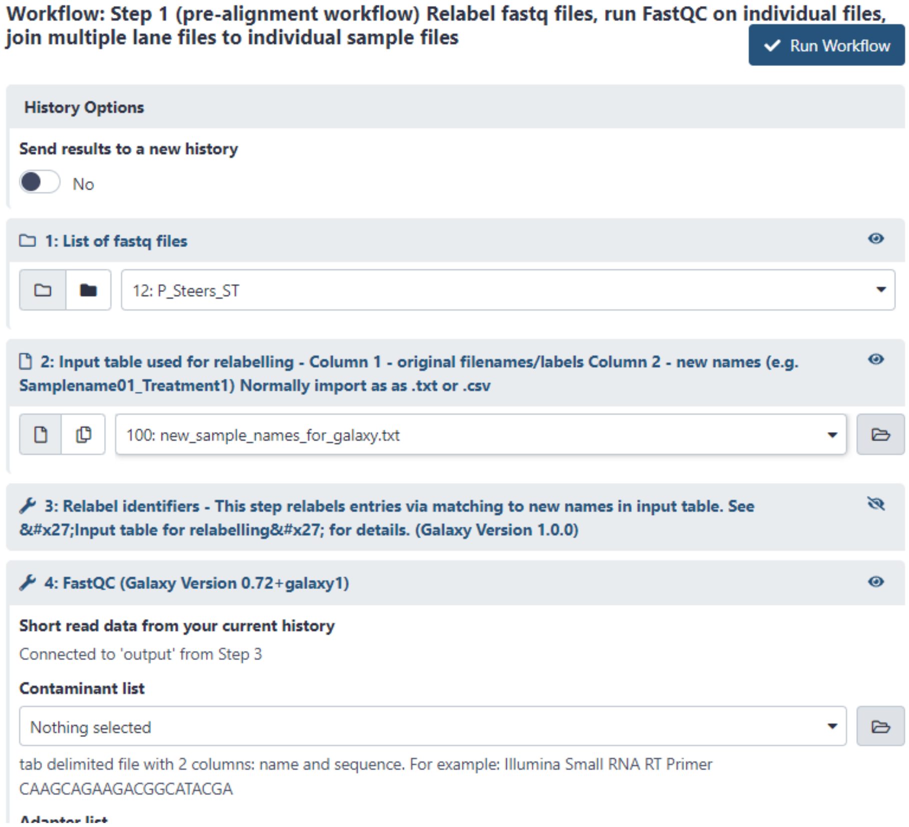
Step 1 (pre-alignment workflow) Relabel fastq files, run FastQC on individual files, join multiple lane files to individual sample files

Step 1: Relabels individual fastq samples Step 2: There may be multiple files for each sample due to multiple lanes on Illumina system (files denoted with different numbers e.g. '_L001'). This will run FastQC on all all individual RNAseq files to determine that all meet quality requirements to proceed with full workflow of alignment/annotation. Viewed in a MultiQC report. Check to see if any individual files fail, or if one particular lane has bias. Step 3: Takes the list of names, creates nested list to group together the multiple lane files for each sample then concatenates these files together to result in 1 file per sample ready for downstream analysis. FastQC is re-run on these files downstream as well. Input 1: List containing

2 months ago

☐

Figure 3.1: Click the Run Workflow button



Workflow: Step 1 (pre-alignment workflow) Relabel fastq files, run FastQC on individual files, join multiple lane files to individual sample files

History Options

Send results to a new history

No

1: List of fastq files

12: P_Steers_ST

2: Input table used for relabelling - Column 1 - original filenames/labels Column 2 - new names (e.g. Samplename01_Treatment1) Normally import as as .txt or .csv

100: new_sample_names_for_galaxy.txt

3: Relabel identifiers - This step relabels entries via matching to new names in input table. See &#x27;Input table for relabelling&#x27; for details. (Galaxy Version 1.0.0)

4: FastQC (Galaxy Version 0.72+galaxy1)

Short read data from your current history

Connected to 'output' from Step 3

Contaminant list

Nothing selected

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

Adapter list

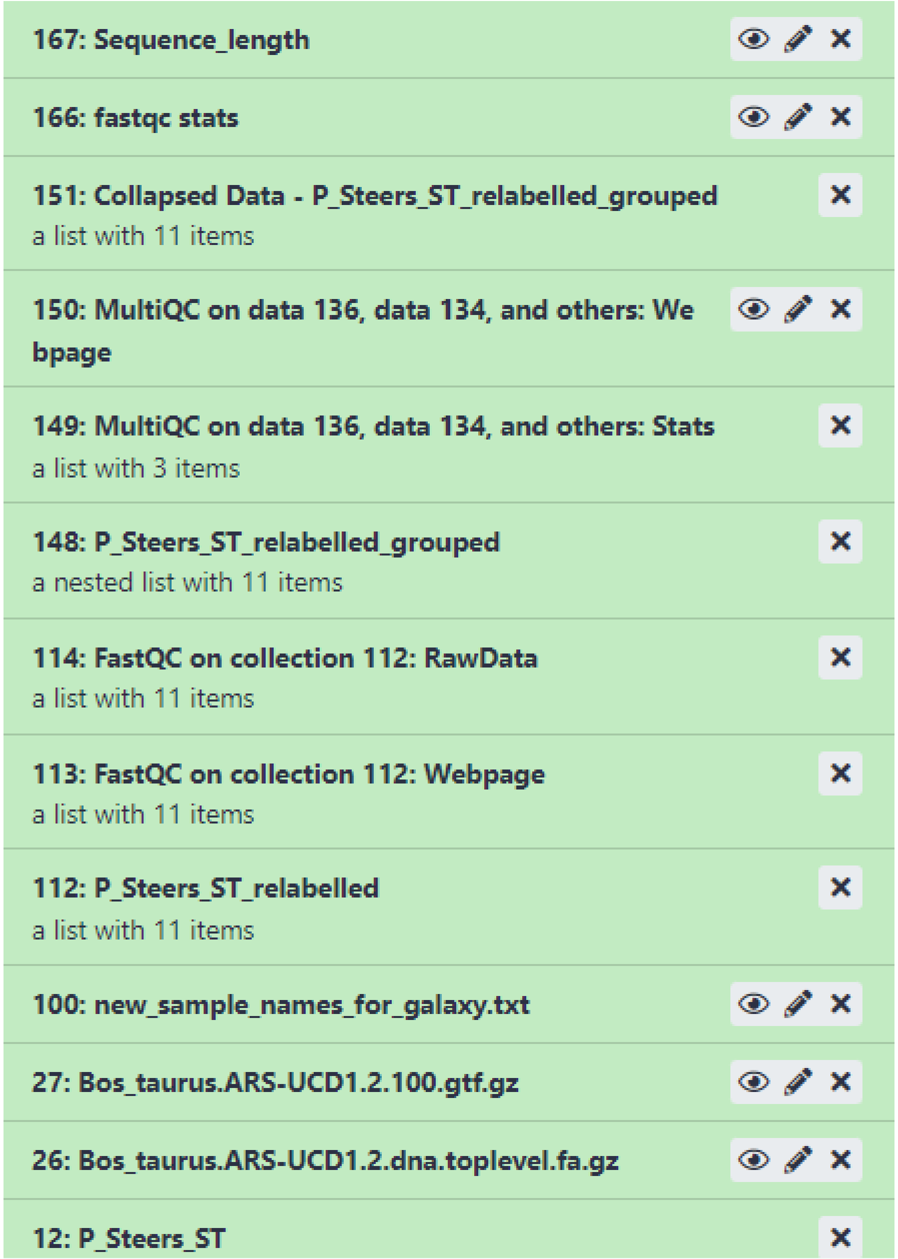
Figure 3.2: Select list of fastq files for number 1 and select the table of new names for number 2, then click 'Run Workflow'

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).



167: Sequence_length

166: fastqc stats

151: Collapsed Data - P_Steers_ST_relabelled_grouped
a list with 11 items

150: MultiQC on data 136, data 134, and others: Webpage

149: MultiQC on data 136, data 134, and others: Stats
a list with 3 items

148: P_Steers_ST_relabelled_grouped
a nested list with 11 items

114: FastQC on collection 112: RawData
a list with 11 items

113: FastQC on collection 112: Webpage
a list with 11 items

112: P_Steers_ST_relabelled
a list with 11 items

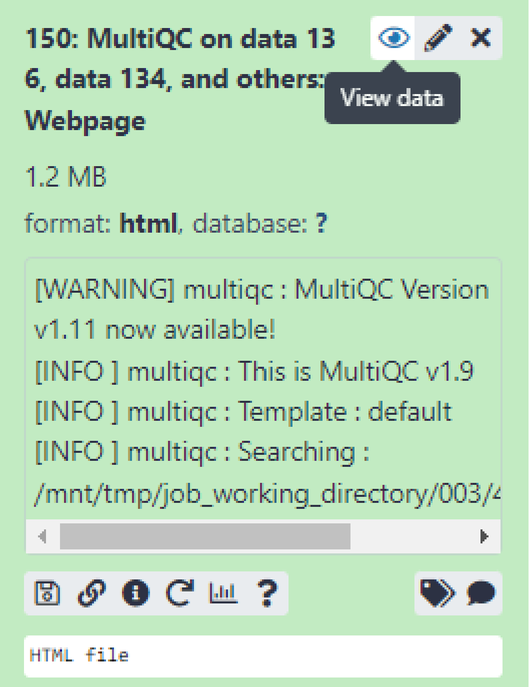
100: new_sample_names_for_galaxy.txt

27: Bos_taurus.ARS-UCD1.2.100.gtf.gz

26: Bos_taurus.ARS-UCD1.2.dna.toplevel.fa.gz

12: P_Steers_ST

Figure 3.3: Example output in history after running first workflow



150: MultiQC on data 136, data 134, and others: Webpage

1.2 MB

format: **html**, database: **?**

```
[WARNING] multiqc : MultiQC Version v1.11 now available!
[INFO ] multiqc : This is MultiQC v1.9
[INFO ] multiqc : Template : default
[INFO ] multiqc : Searching :
/mnt/tmp/job_working_directory/003/4
```

HTML file

Figure 3.4: View the MultiQC webpage report

Chapter 4

Workflow - 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 Differential Expression (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 trimmomatic, RNA STAR Aligner and featureCounts.

4.1 Import workflow

Follow instructions in Section 3.1 to add the following workflow:

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

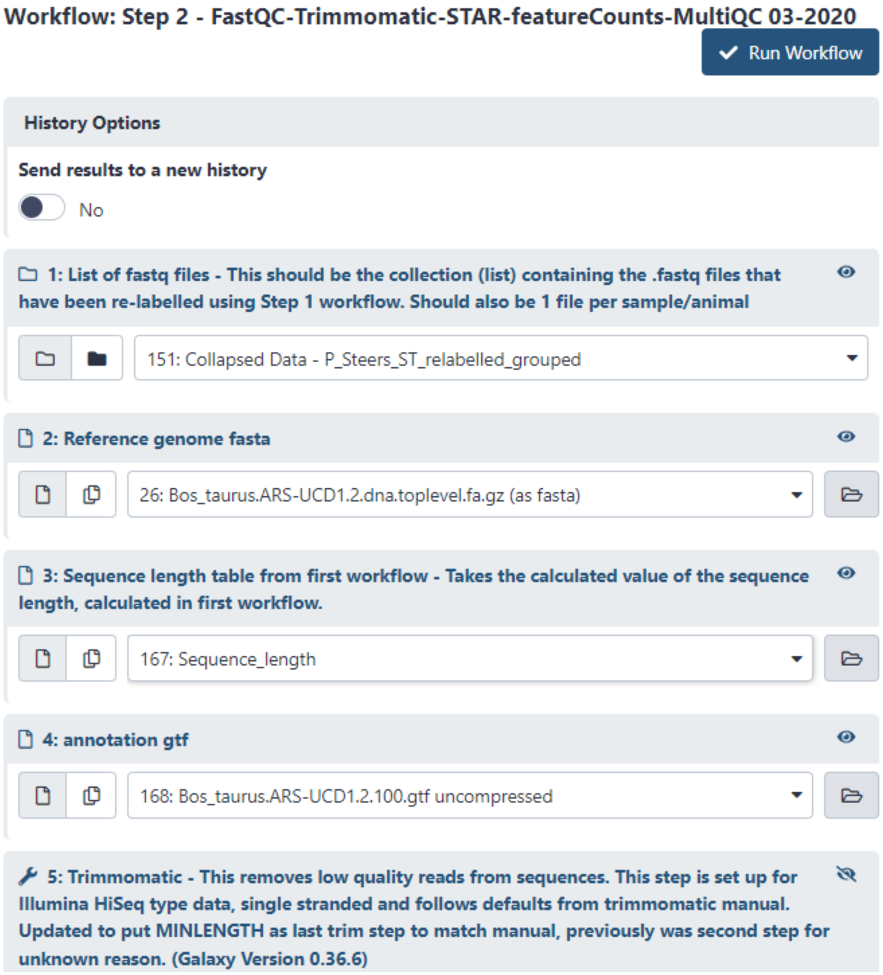
4.2 Run Workflow

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. An email will also be sent to you once the last step is completed.

4.3 Save data

Each analysis will require different outputs, but a simple DE analysis will normally only require the featureCounts output that is normally labelled as



Workflow: Step 2 - FastQC-Trimmomatic-STAR-featureCounts-MultiQC 03-2020

History Options

Send results to a new history

No

1: List of fastq files - This should be the collection (list) containing the .fastq files that have been re-labelled using Step 1 workflow. Should also be 1 file per sample/animal

151: Collapsed Data - P_Steers_ST_relabelled_grouped

2: Reference genome fasta

26: Bos_taurus.ARS-UCD1.2.dna.toplevel.fa.gz (as fasta)

3: Sequence length table from first workflow - Takes the calculated value of the sequence length, calculated in first workflow.

167: Sequence_length

4: annotation gtf

168: Bos_taurus.ARS-UCD1.2.100.gtf uncompressed

5: Trimmomatic - This removes low quality reads from sequences. This step is set up for Illumina HiSeq type data, single stranded and follows defaults from trimmomatic manual. Updated to put MINLENGTH as last trim step to match manual, previously was second step for unknown reason. (Galaxy Version 0.36.6)

Figure 4.1: Setup Step 2 then press Run Workflow

`featureCounts_matrix.tabular`. Download this file. It can be previewed in Microsoft Excel by dragging and dropping into an active Excel window, however it is sometimes a large file and best handled in R or other similar software. Also be careful not to save changes if viewing in Excel as some gene names will be converted to date formats in excel and may introduce errors e.g. `MARCH1` gene.

It is also recommended to save the MultiQC webpage and stats output. This is particularly useful for describing methods when writing up.

You may also want to view the specific alignments and therefore you will need the .bam files. See https://software.broadinstitute.org/software/igv/BAM for more information.

4.4 Inspect Report

It is important to view the MultiQC webpage and check that all output meets your quality requirements.

Chapter 5

More notes on 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.

The following is a description of how to handle this manually. Please note that this is automated within Step 1 Workflow.

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