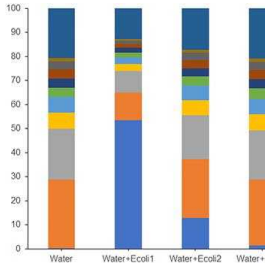
Jul 25, 2022 Version 1

Centriflaken: an automated data analysis pipeline for assembly and in silico analyses of food-borne pathogens from metagenomic samples V.1



DOI
https://dx.doi.org/10.17504/protocols.io.kxygxzdbwv8j/v1

Kranti Konganti[1], Vishal Thovarai[1], Meghan Maguire[2], Julie A. Kase[2], Narjol Gonzalez-Escalona[2]

[1]Scientific Computing Support, Office of Regulatory Science, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA;
[2]Division of Microbiology, Office of Regulatory Science, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA

CPIPES

Kranti Konganti

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.kxygxzdbwv8j/v1

External link: https://galaxytrakr.org

Protocol Citation: Kranti Konganti, Vishal Thovarai, Meghan Maguire, Julie A. Kase, Narjol Gonzalez-Escalona 2022. Centriflaken: an automated data analysis pipeline for assembly and in silico analyses of food-borne pathogens from metagenomic samples. **protocols.io** https://dx.doi.org/10.17504/protocols.io.kxygxzdbwv8j/v1

Protocol status: Working
We use this protocol and it's working

Created: July 21, 2022

Last Modified: July 25, 2022

Protocol Integer ID: 67280

Keywords: Metagenomics, Nextflow, Bioinformatics, MAG, Serotyping, Food-borne pathogen, Computational biology, Bioinformatics pipeline, entire precision metagenomics analysis, pathogens from metagenomic samples rapid, precision metagenomics approach, comprehensive analysis of metagenomic data, metagenomic data, metagenomic samples rapid, metagenomic, metagenomics approach, metagenomic sample, classifying shiga toxin, assembled genome, automated data analysis pipeline, automated analysis pipeline, escherichia coli, data analysis pipeline for assembly, bioinformatics, genome, shiga toxin, sequencing, automated data analysis workflow, silico analyses of food, borne pathogen, analysis pipeline, pipeline with nanopore data, centriflaken pipeline, silico analysis, bacterial source, nanopore data, data analysis workflow, using oxford nanopore

Funders Acknowledgements:
Chief Scientist Challenge Grants Program and and the FDA Foods Program Intramural Funds
Grant ID: 2021-200F07A

Disclaimer

Abstract

Rapid and comprehensive analysis of metagenomic data from any sample is of importance in food safety. Also important is the need for automated analysis pipelines allowing rapid and effective construction of metagenomic assembled genomes (MAGs) to enable bacterial source-tracking from metagenomic data. We developed a precision metagenomics approach for detecting and classifying shiga toxin-producing *Escherichia coli* (STEC) in enrichments of agricultural water using Oxford Nanopore long read sequencing where the bioinformatics data analysis employed many sequential manual steps (Maguire et al, 2021). Here we present centriflaken, a suite of automated data analysis workflows enabled by Nextflow, which takes metagenomic data and generates MAGs and performs *in silico*-based analysis as described in Maguire et al, 2021. The final summary plots and tables can be downloaded from the provided MultiQC HTML report generated as part of the pipeline. The centriflaken pipeline was validated with data from our previously published method and was able to replicate the detection and classification of STECs for each sample. We tested the pipeline with nanopore data obtained from 21 additional enriched samples from irrigation water and was able to perform the entire precision metagenomics analysis in less than 5 hours. We have further expanded this precision approach to include any user supplied taxa of interest (*Listeria monocytogenes* or *Salmonella*) not only STECs.

Before start

Purpose:

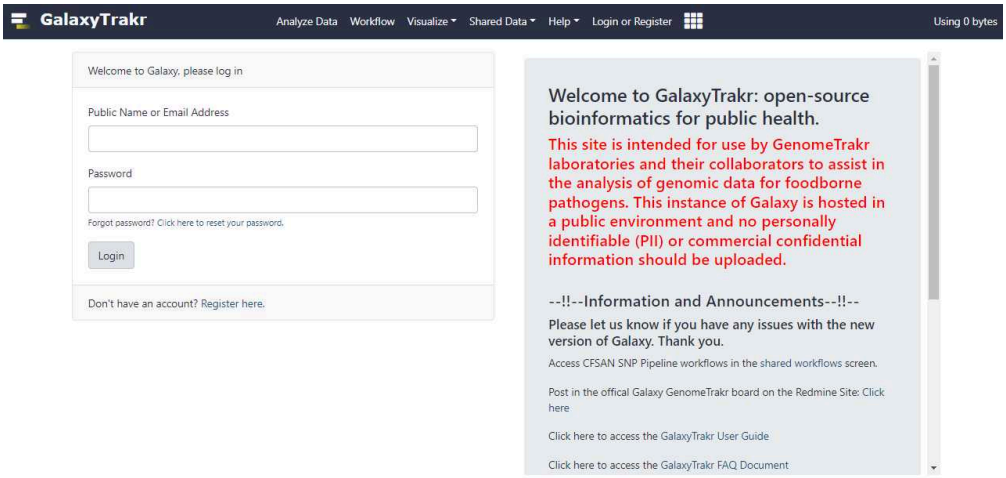
This protocol will show step-by-step instructions on how to run the Centriflaken pipeline on **GalaxyTrakr**.

Step 1

1 **Create an account and Login:**

If you do not already have an account on **GalaxyTrakr**, please create one by visiting this URL: https://account.galaxytrakr.org/Account/Register

1.1 Once your account is activated, login by visiting https://galaxytrakr.org.



Step 2

2 **Create a new history:**

Note

We recommend creating a new history for each new invocation of the **Centriflaken** pipeline.

Doing so will keep your data sets organized should you choose to carry on further analyses within the same project.

For example, you might want to further perform analyses using the Metagenomically Assembled Genomes (MAGs), which are the end products of the **Centriflaken** pipeline by predicting genes *de novo* and performing annotation.

Safety information

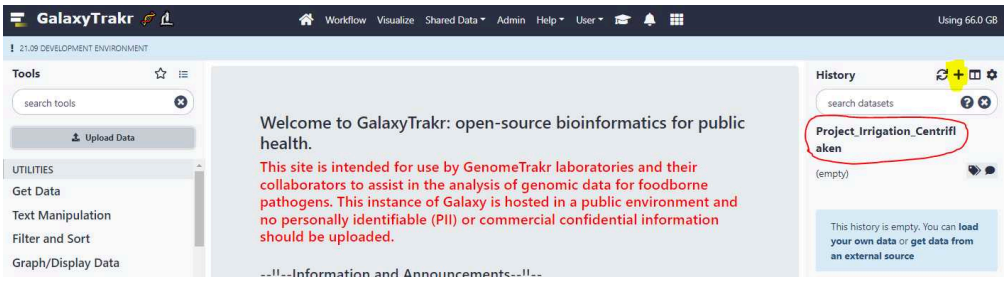
After all the analysis outputs from this run is saved to your internal data network or computer, older history's should be purged/deleted so as not to occupy the limited storage space in your account.

In some cases it may be useful to save, for a limited time, multiple histories or to run analyses concurrently in multiple histories.

In these cases you need to pay attention to your % Usage bar (shows % used of allocated storage space) in the upper right corner of the GalaxyTrakr page.

If you need additional space you can contact **galaxytrakrsupport@fda.hhs.gov** and request additional storage.

2.1 Create a new history using the "**+**" symbol in the upper right hand corner. Name your history and press "**Enter**" on your keyboard to save the name.



Step 3

3 **Upload unpaired data sets via Galaxy web interface:**

Note

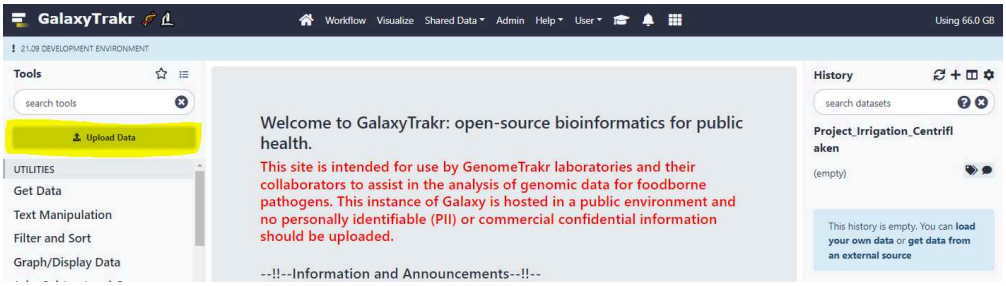
You can either upload the data sets directly from your local workstation or desktop or you can upload via **FTP** protocol.

You can upload either single-end short read, paired-end short read or long read data sets.

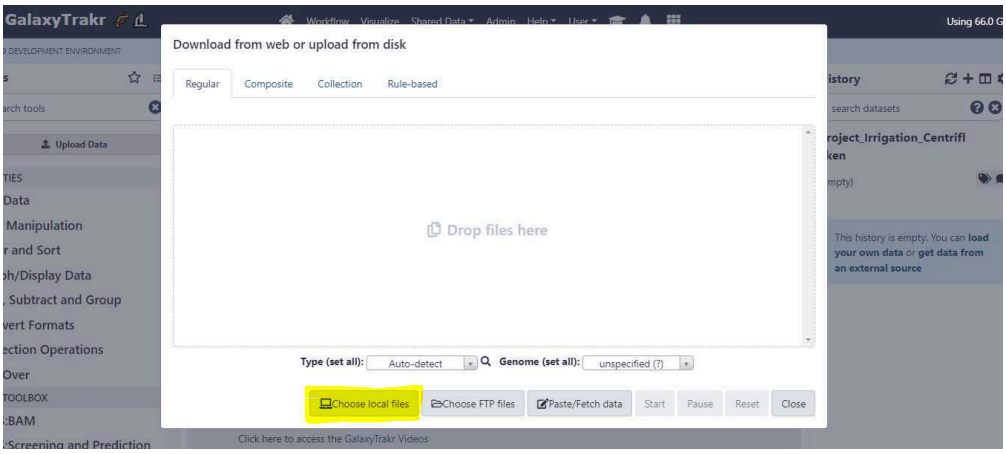
3.1 To upload files **without using FTP protocol**, click on the "**Upload Data**" button located on upper left of the page as shown below.

Safety information

Centriflaken only works on sequencing data files of **FASTQ** format.



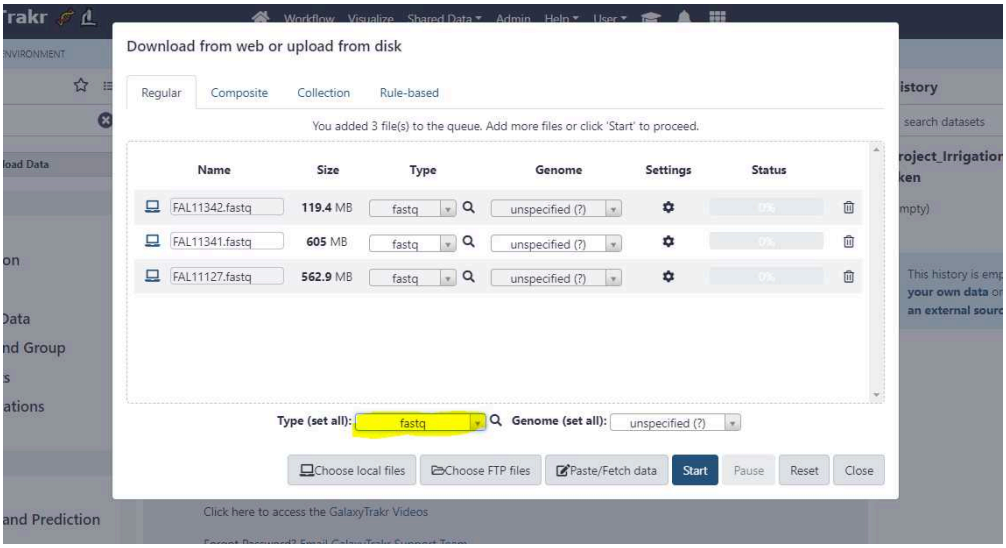
3.2 A window will appear in the middle of your screen. This is where you select your files using the "**Choose local files**" button at the bottom of the window.



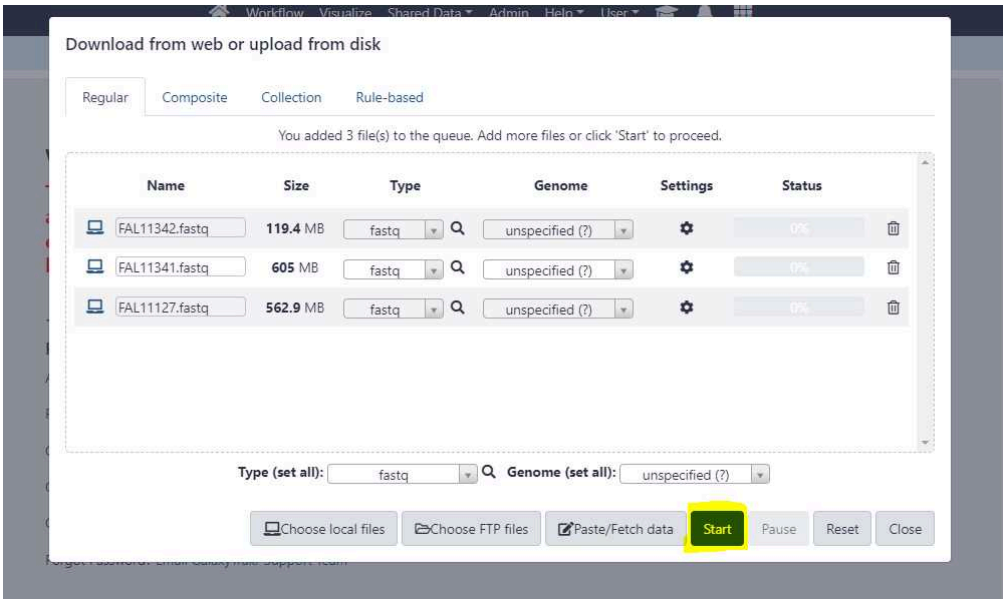
3.3 Set the appropriate "**Type (set all)**" extension.

Safety information

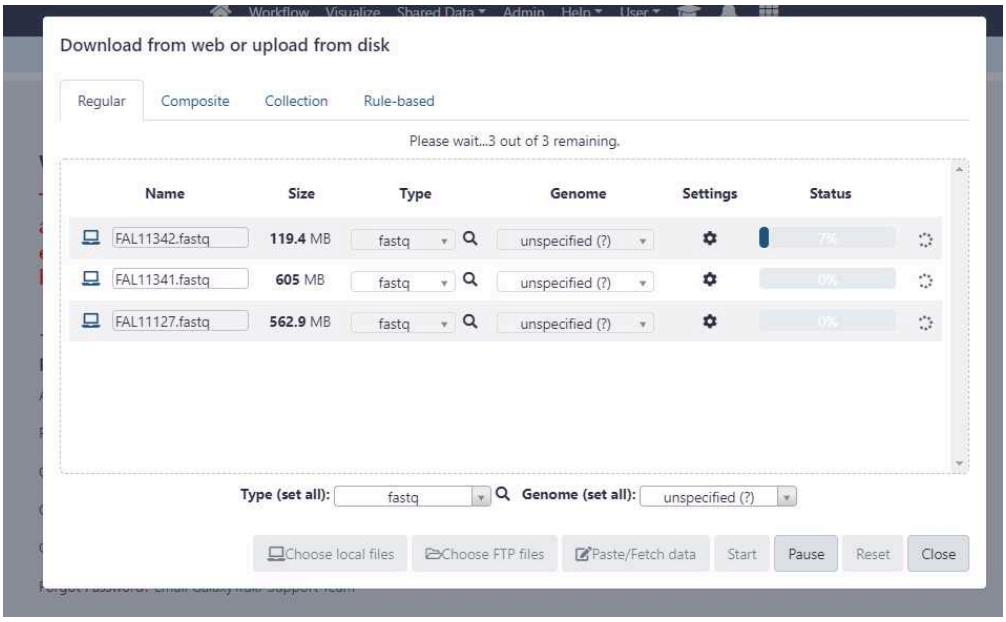
Please make sure that you select the "**Type (set all)**" option to whatever the extension of the files you are uploading is. If the files you are uploading end in extension "**.fastq**", select this as **fastq**. If it is "**.fastq.gz**", choose the option **fastq.gz**.



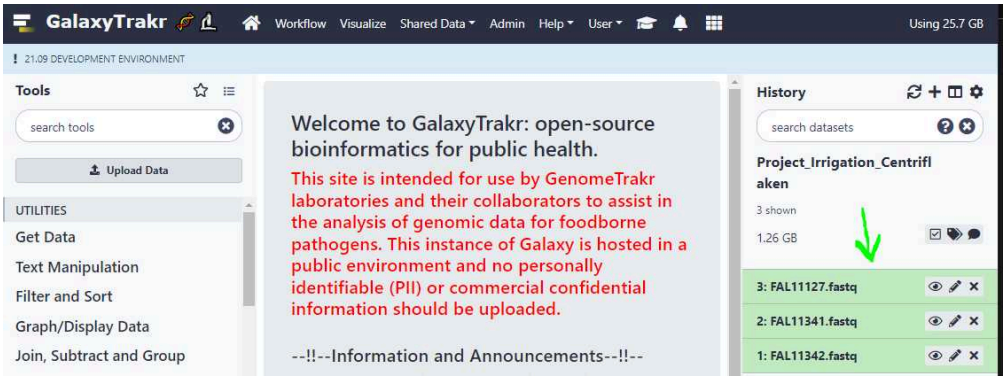
3.4 Once you have correctly set the "**Type (set all)**" option, click on "**Start**" button to start the upload of **FASTQ** files from your local workstation or desktop to **GalaxyTrakr**.



3.5 The "**Status**" bar changes to show the progress of the upload.



3.6 Once, the uploads are complete, they should appear in your history as shown below.



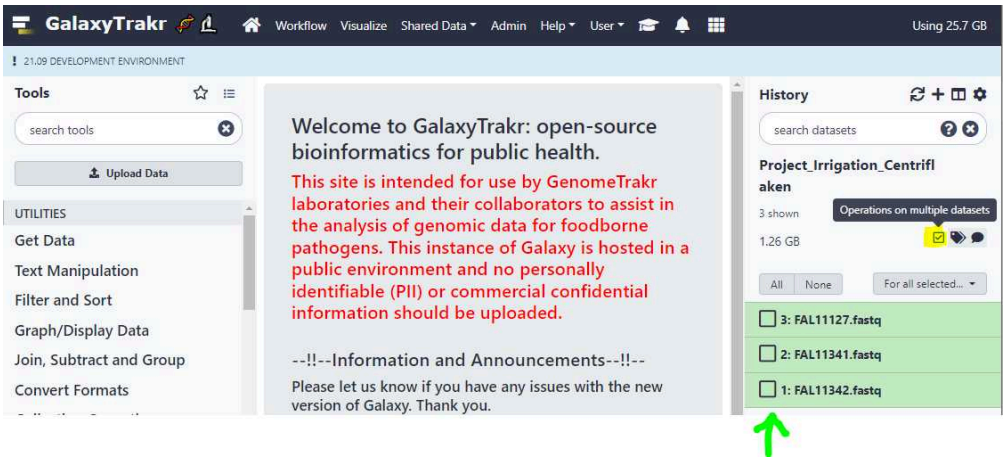
3.7 We need to create a data set list to run **Centriflaken** pipeline.

For single-end or long reads, we create an unpaired data set list and for paired-end reads we create a list of data set pairs.

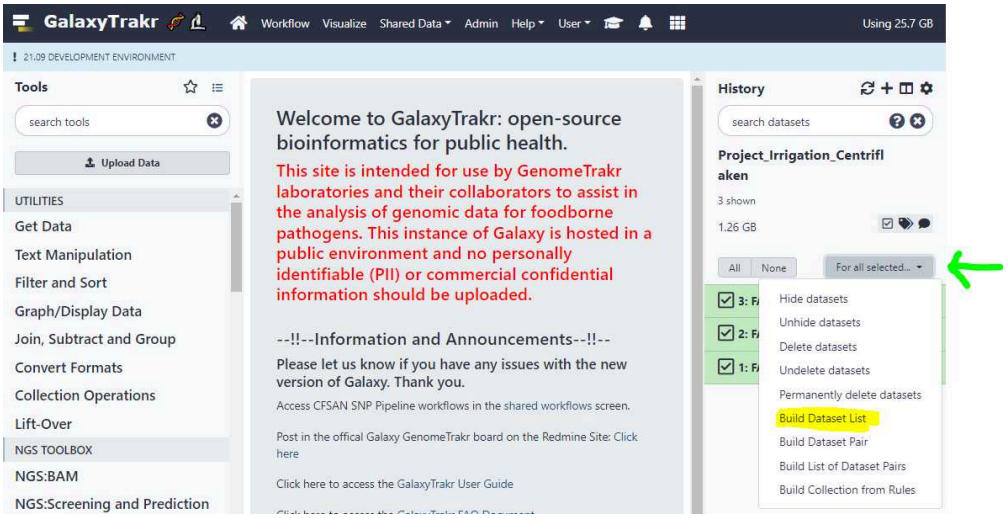
3.8 **Create an unpaired data set collection:**

Using the above uploads as an example, we need to create a list unpaired data sets as these are Oxford Nanopore long reads.

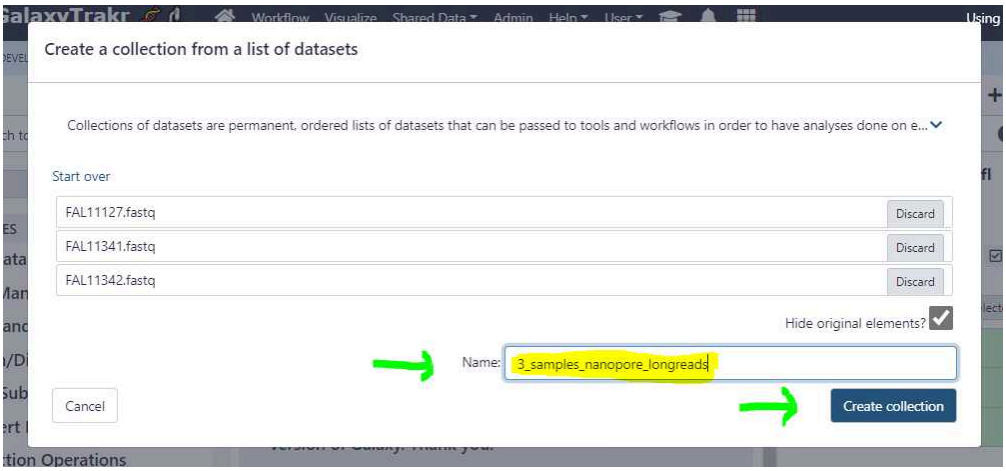
First, go to your history and click on the "**Checkbox**" icon as highlighted in yellow below. This will put check boxes in front of the data sets to select.



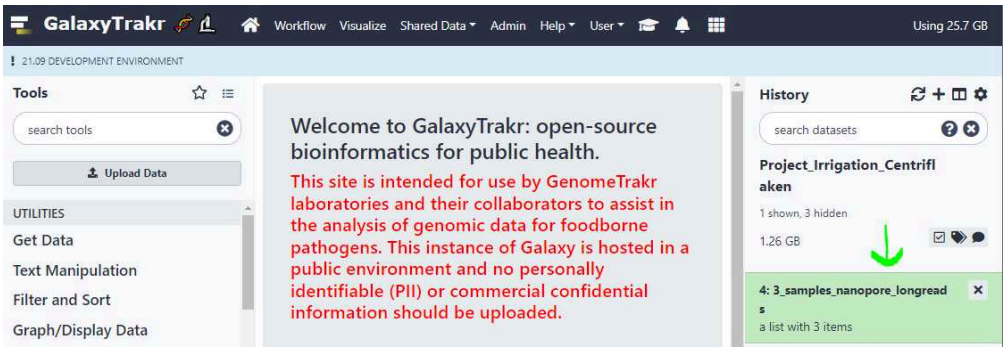
3.9 Now, select the data sets for which you want to create an unpaired data set list and click on the dropdown menu as show below and select the option "**Build dataset list**".



3.10 This will bring up a window where you should name the this collection of data sets and click on "**Create Collection**" to save it to your history.



The newly named collection of long read data sets should now appear in your history.



Step 4

4 **Upload paired-end data sets via FTP protocol**:

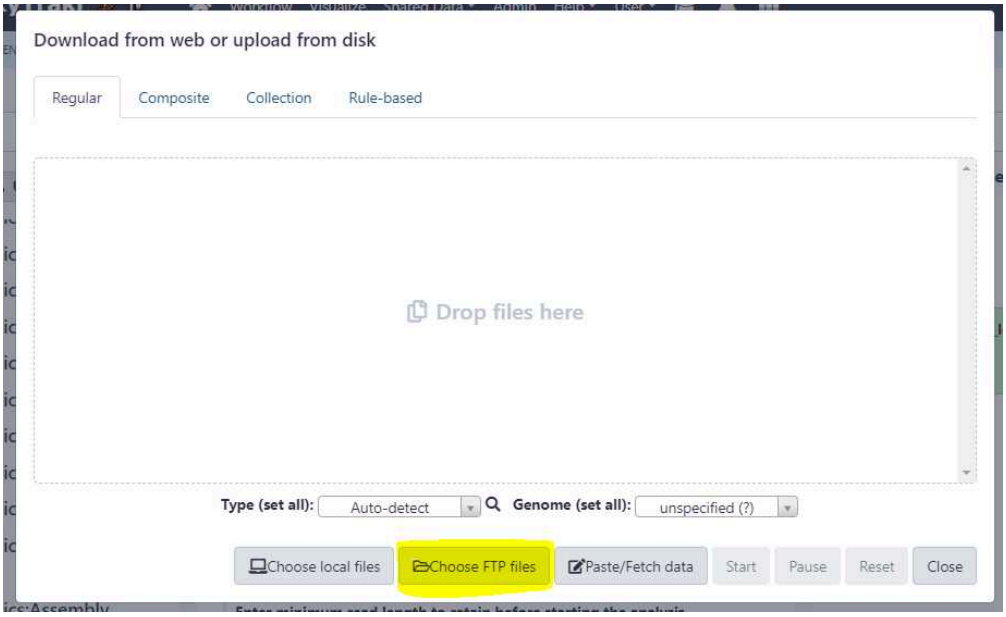
Note

It is generally recommended to use the **FTP** protocol to upload large number of data sets to **GalaxyTrakr**.

Please refer to the section **3.2.1** in the User Guide below and follow the directions to upload the data sets using the **FTP** protocol to **GalaxyTrakr**.

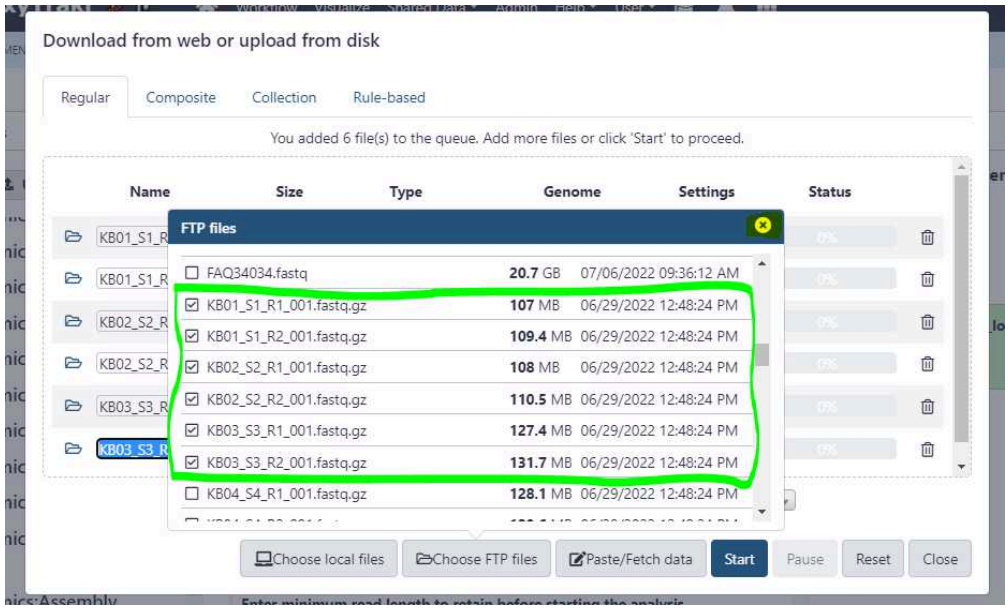
Galaxy+Genome+Trakr+User+Guid...

4.1 Once your paired-end data sets are uploaded, go to **GalaxyTrakr** and click on the "**Upload Data**" button. This should bring up the upload window overlay, and this time, click on the the "**Choose FTP files**" button as shown below.



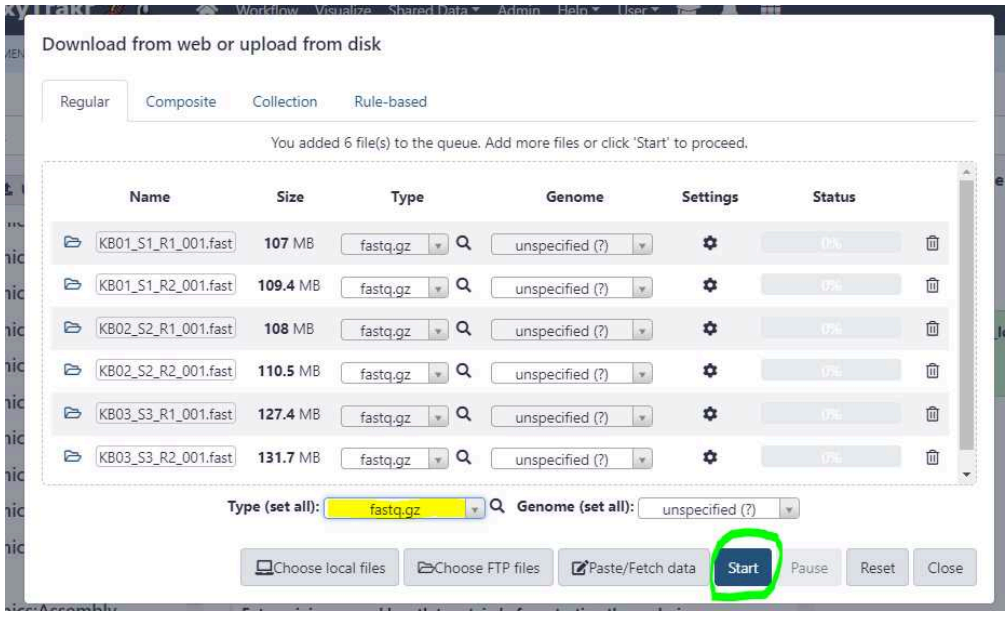
4.2 Now, select the relevant files you uploaded via **FTP** protocol. In the example below, we are selecting 3 samples (circled in green) which equals to 6 files to be uploaded.

Then, first close the "**FTP files**" window (highlighted in yellow).

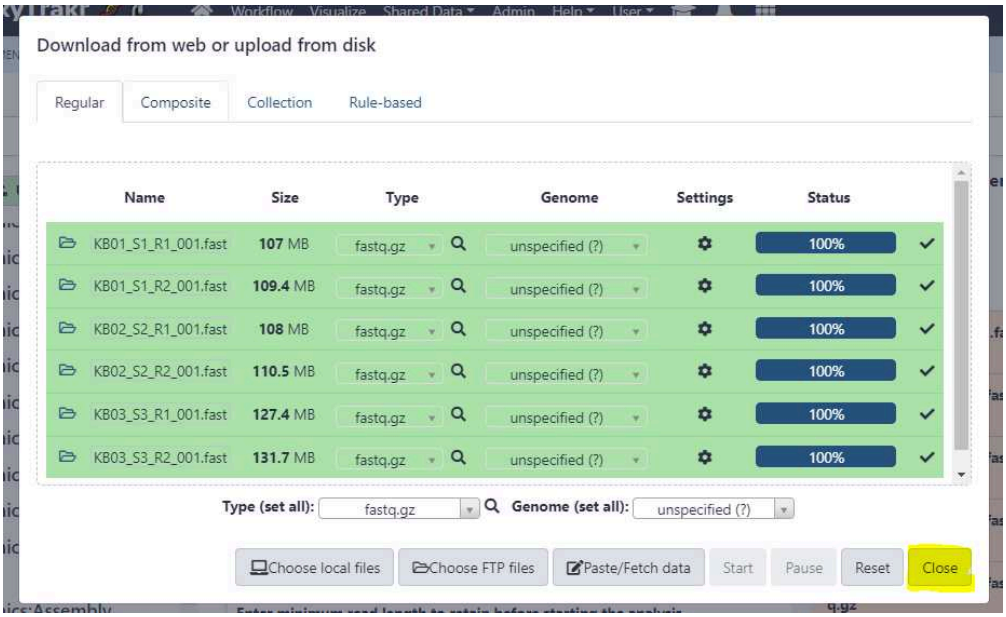


Safety information

Now, make sure to change the "**Type (set all)**" option to the corresponding extension of the files that were selected (see below) and then click on "**Start**" button to finish uploading the **FTP** files.

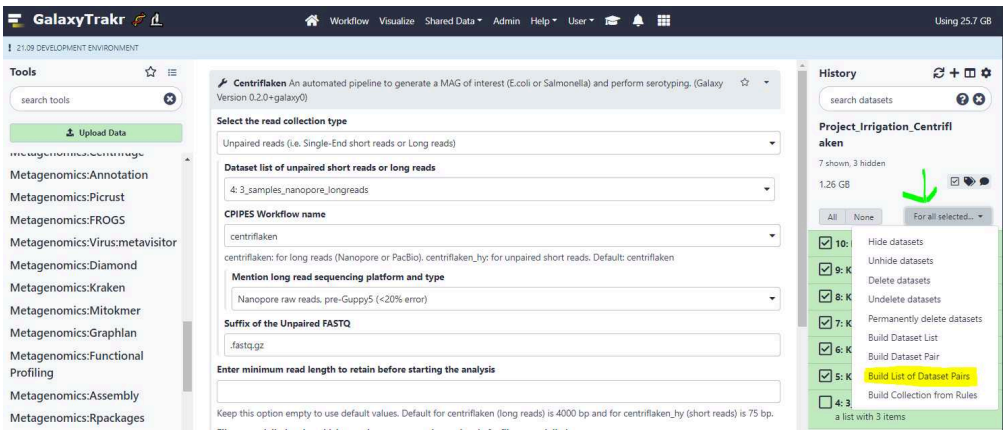


Since the files were pre-uploaded using an **FTP** client, the upload from the **GalaxyTrakr** interface should finish relatively quickly. Next, close the upload window by clicking on "**Close**" button.

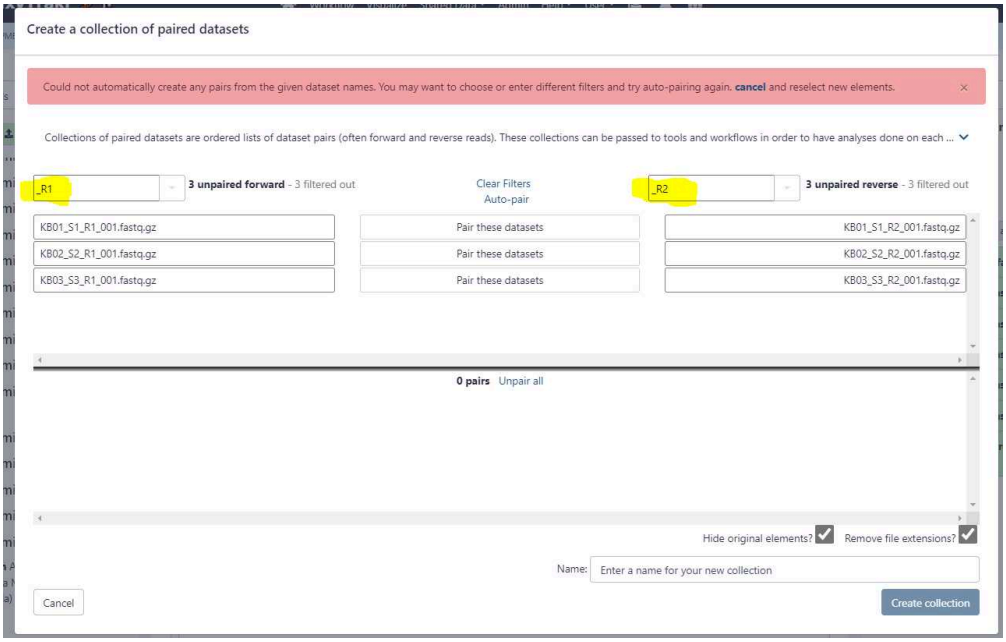


4.3 Now the **FTP** uploaded files should appear in your history. We will create a new list of data pairs collection using these 6 files (3 samples).

Similar to step **3.8**, we click on the "**Checkbox**" to select our 6 data set files (3 samples) and this time, we click on "**Build List of Dataset Pairs**" (highlighted in yellow) from the dropdown menu (green arrow).



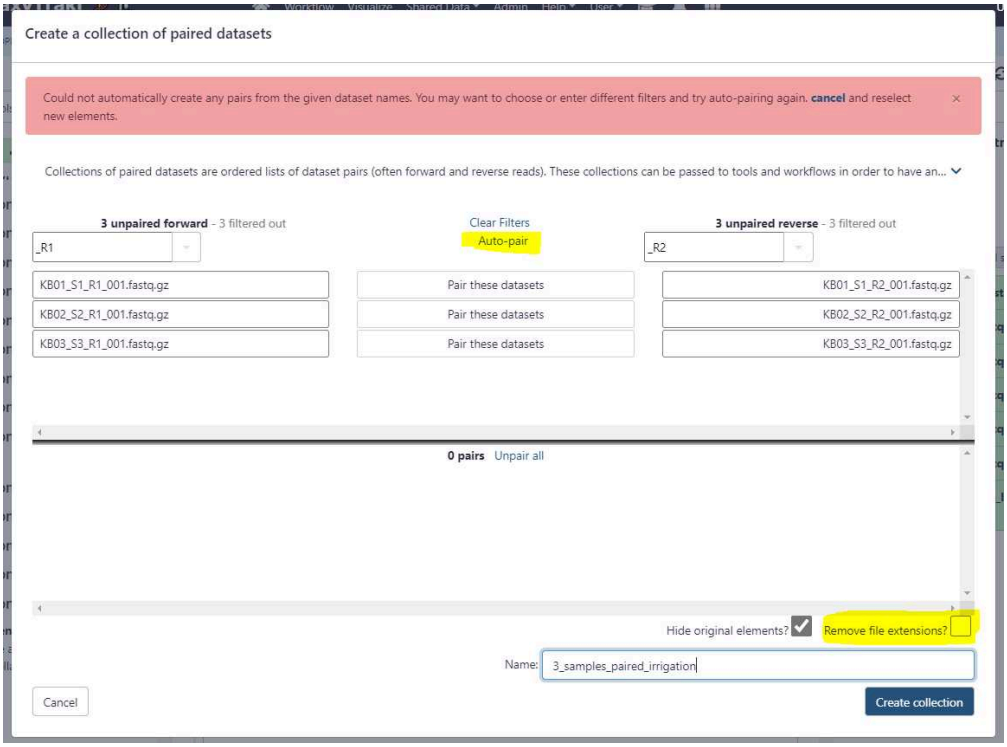
4.4 Since the file names of our uploaded data sets ended in the following suffixes: **_R1_001.fastq.gz** and **_R2_001.fastq.gz**, we select **_R1** to indicate our forward read and **_R2** to indicate our reverse read from the drop down menus (highlighted in yellow).



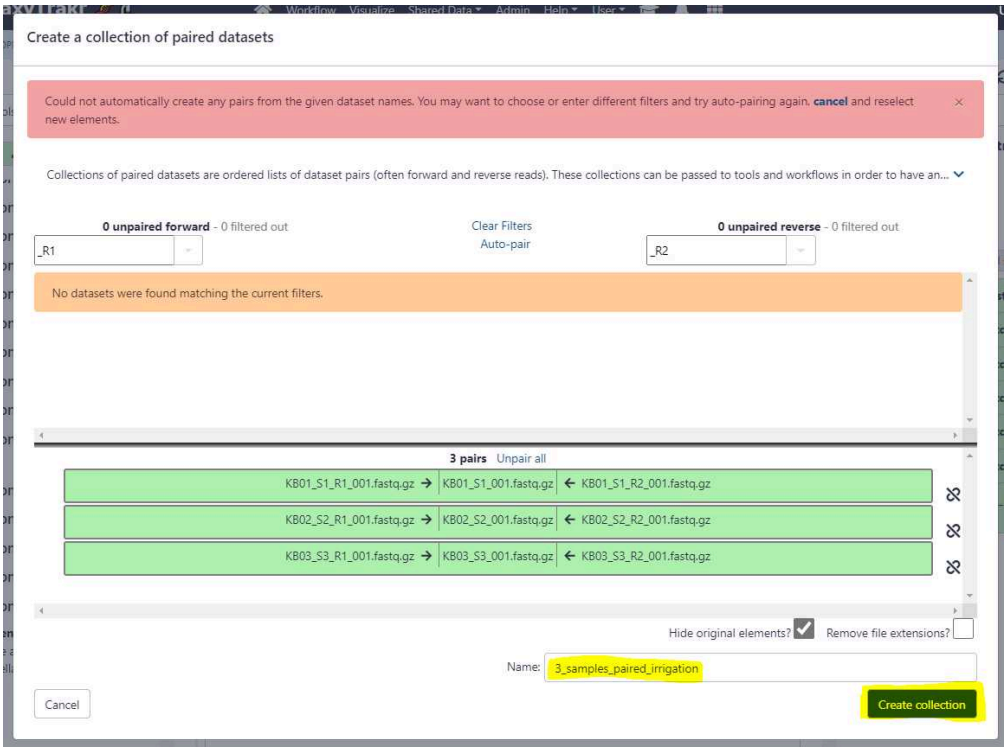
4.5 Next, click on "**Auto-pair**" to pair these data set files.

Safety information

Make sure to uncheck the "**Remove file extensions?**" checkbox.



4.6 Finally, create a name for the collection and click on "**Create collection**" button.

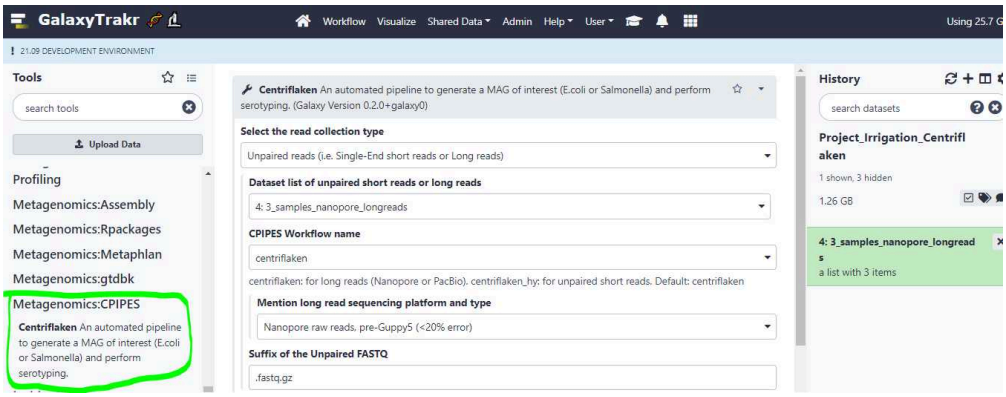


Step 5

5 **Run Centriflaken pipeline on long reads:**

Now, choose "**Centriflaken**" pipeline from the tool navigation menu on the left under "**CPIPES: Metagenomics**".

> Note
>
> Alternatively, you can search for "**Centriflaken**" by typing in the "**search tools**" box.



5.1 Selecting the "**Centriflaken**" under "**Metagenomics:CPIPES**" will bring up the job submission form in GalaxyTrakr. Here we set a few required parameters to run the pipeline on our created collection of long read data sets.

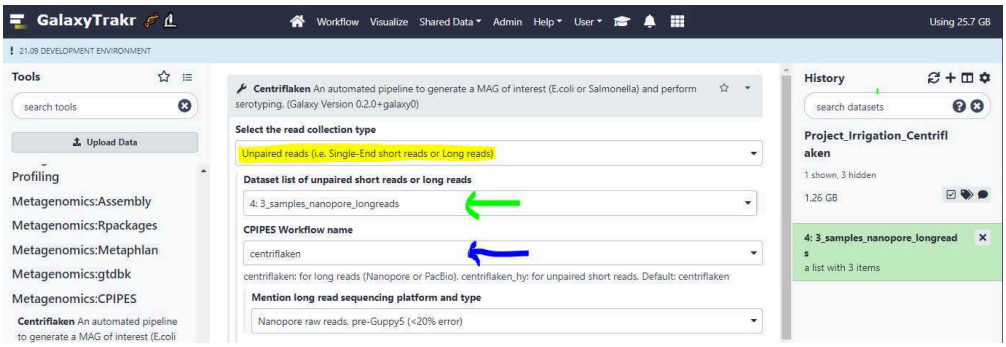
Select the read collection type:

Unpaired reads (i.e. Single-End short reads or Long reads): Select this option if your newly created data set collection contains Single-End short reads or Long reads. Based on the data set collection we created in the example above, we will be selecting this option (highlighted in yellow; see below).

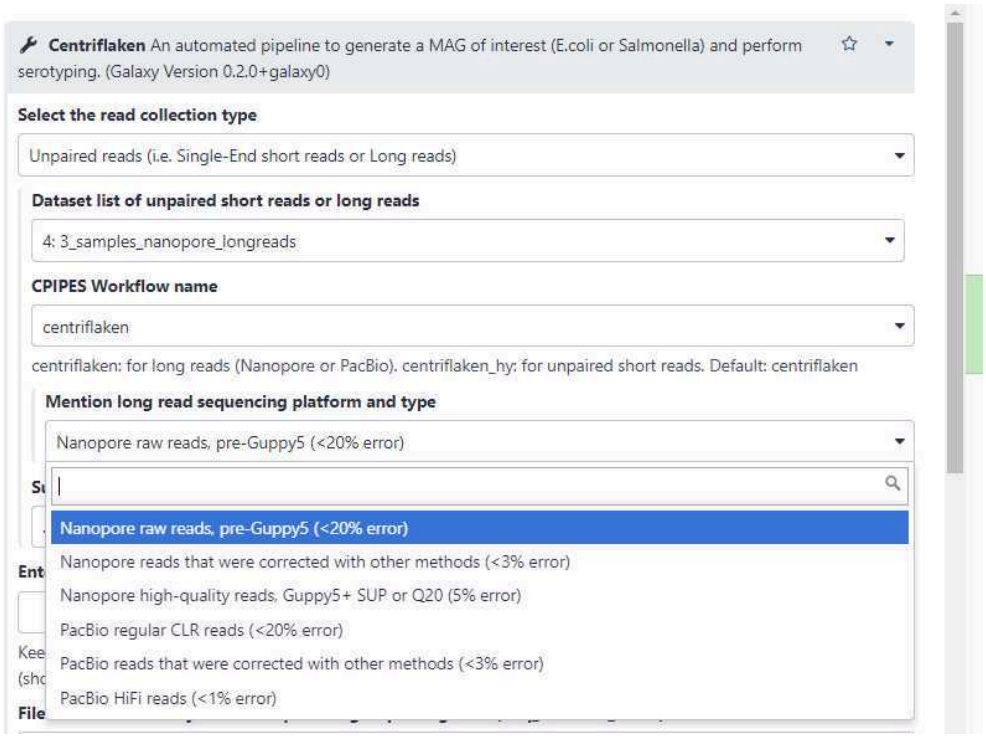
Paired-End reads: Select this option if your newly created data set collection contains Paired-End short reads.

5.2 **Dataset list of unpaired short reads or long reads**: This is where we point to our created data set collection. Based on the example above, the collection name "**3_samples_nanopore_longreads**" is selected (green arrow; see below).

CPIPES Workflow name: Select "**centriflaken**" if your data set collection contains long reads and select "**centriflaken_hy**" if your data set collection contains short reads, either single-end or paired-end (blue arrow; see below)



5.3 **Mention long read sequencing platform and type:** Use this option to select the type of long reads that were uploaded or present in your read collection. For the example above, we uploaded raw Oxford Nanopore reads and that option is selected below.



5.4 **Suffix of the Unpaired FASTQ:**

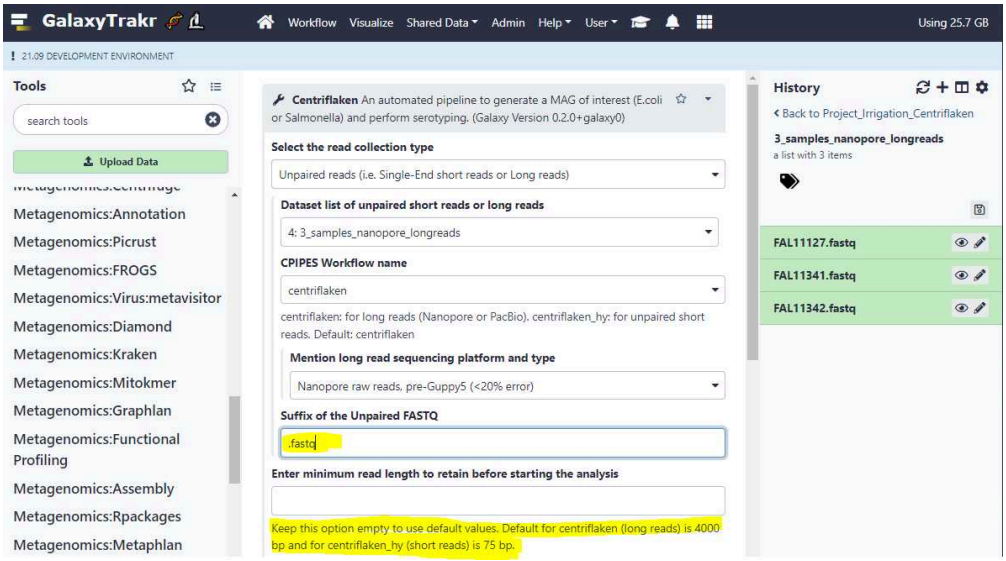
Since the extensions of the **FASTQ** files we originally uploaded was **.fastq** (Ex: FAL11127.fastq, FALL11341.fastq etc.), we change this option from the default value of "**.fastq.gz**" to "**.fastq**".

Safety information

It is of **utmost importance** to set the correct suffix as the pipeline will fail in the backend if it cannot find files that end with the mentioned suffix.

Enter minimum read length to retain before starting the analysis:

Enter the minimum length of the **FASTQ** reads to keep before starting the analysis. Leave this box empty to use the default values. It is 4000 bp for long reads and 75 bp for short reads.



5.5 File name delimitor:

Safety information

Most of the pipeline failures result from incorrectly setting the "**Suffix**" field, "**File name delimitor**" and the "**File name delimitor index**" fields.

Use this option to perform sample grouping. Sample grouping is entirely based on file names of the uploaded **FASTQ** files. The suffixes of the file names of the uploaded files

are first removed based on the entry in the "**Suffix**" field and then the sample grouping is performed.

For example, if the uploaded FASTQ files are as below:

- KB01_biological_replicate1.fastq
- KB01_biological_replicate2.fastq
- KB02_biological_replicate1.fastq
- KB02_biological_replicate2.fastq
- KB03_biological_replicate1.fastq
- KB03_biological_replicate2.fastq
- KB04_biological_replicate1.fastq
- KB04_biological_replicate2.fastq

Here, we have 2 biological replicates per sample. Now, to create 4 sample groups, KB01, KB02, KB03 and KB04, we set the "**File name delimitor**" to **_** (an underscore character), because, if you split the file names above (after removing the suffix **.fastq**) by an underscore character, you end up with the following words for each file name:

- KB01 biological replicate1
- KB01 biological replicate2
- KB02 biological replicate1
- KB02 biological replicate2
- KB03 biological replicate1
- KB03 biological replicate2
- KB04 biological replicate1
- KB04 biological replicate2

Note

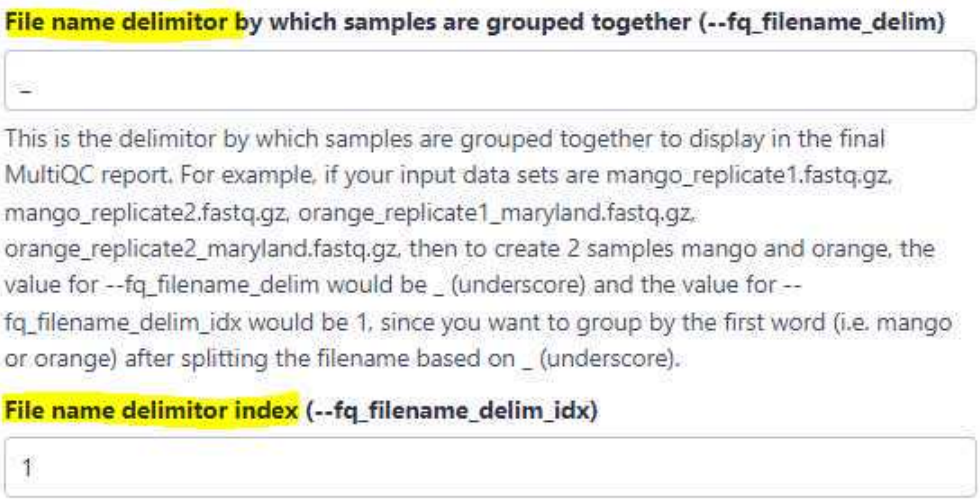
Since the suffixes (in the example above, **.fastq**) are removed before sample grouping, they are not considered as words after splitting by the "**File name delimitor**"

Safety information

If you are working with data sets i.e. **FASTQ** files downloaded from SRA through Galaxy itself using any of the tools like **fasterq-dump**, it is highly likely that downloaded files will have an extension of **.fastq**, so please set it to that value for the "**Suffix**" field and leave the "**File name delimitor**" and "**File name delimitor index**" values as the default values of **_** and **1**.

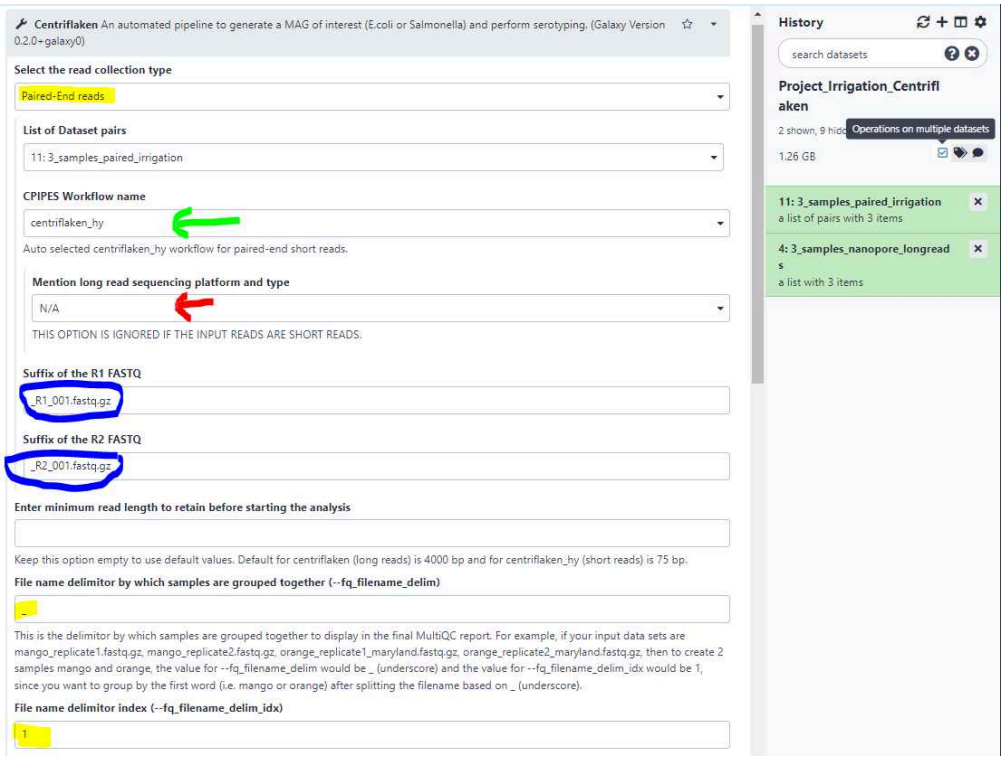
5.6 **File name delimitor index:**

Following the above example, now to create 4 sample groups (KB01, KB02, KB03 and KB04), we need to select the first word and therefore we set the value to 1.



5.7 **Suffix of the paired-end FASTQ:**

Changing the "**Select the read collection type**" (highlighted in yellow; see below) to "**Paired-End reads**" dynamically changes the form fields and auto-selects the "**centriflaken_hy**" (green arrow; see below) pipeline.



It also automatically disables the "**Mention long read sequencing platform and type**" field (red arrow; see above).

Safety information

Finally, make sure to correctly set the complete suffix of the upload **paired-end** FASTQ files. When you upload the sequencing **FASTQ** files that comes off of the Illumina DNA Sequencing Instrument, it is highly likely that the suffixes will be **_R1_001.fastq.gz** and **_R2_001.fastq.gz**. If there are any non-standard suffixes, make sure to correctly set it here.

Safety information

If you are working with data sets i.e. **FASTQ** files downloaded from SRA through Galaxy itself using any of the tools like **fasterq-dump**, it is highly likely that downloaded files will have an extension of **.fastq**, so please set it to that value for the "**Suffix**" field and leave the "**File name delimitor**" and "**File name delimitor index**" values as the default values of **_** and **1**.

5.8 **Reads belonging to this taxa are extracted and a MAG is generated to allow for serotyping:**

Use this option to set the name of a taxa for which the reads belonging to the set taxa are extracted, *de novo* assembled and further serotyping and AMR analyses are performed.

Note

If the input is long reads, then the **FLYE** assembler is used and if the input reads are short reads, **MEGAHIT** assembler is used.

Note

Similarly, if the taxa of interest is "**Escherichia coli**" SerotypeFinder tool is used for serotyping where as SeqSero2 is used for "**Salmonella**".

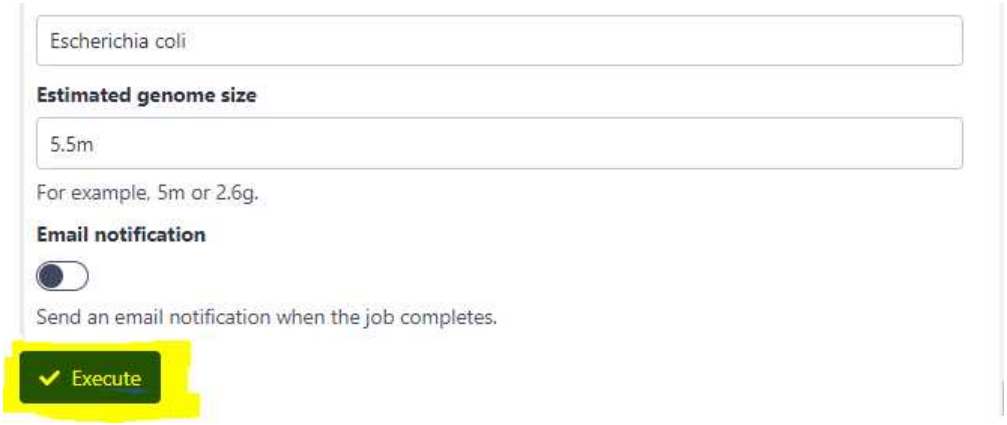
5.9 **Estimated genome size:**

This option is only required if the input is long reads as **FLYE** assembler requires this information to perform a de novo assembly. For short reads, the value in this field is ignored.

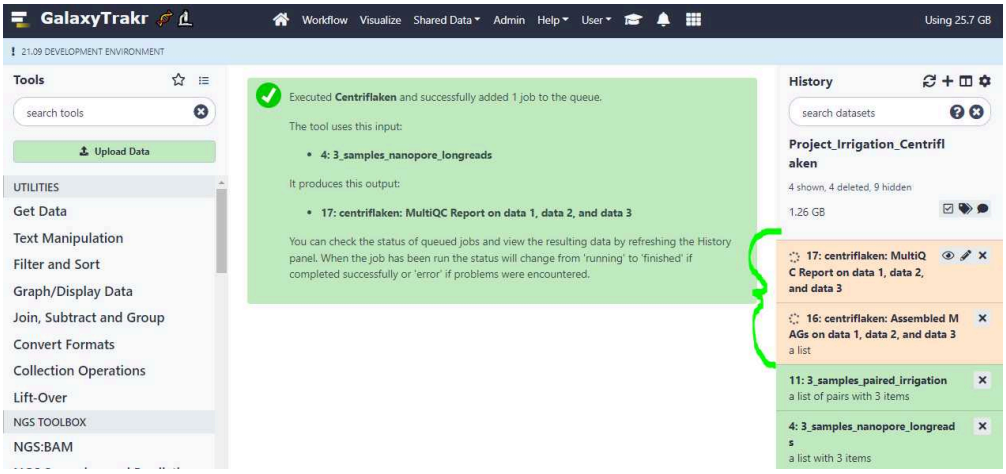
Step 6

6 **Submit the Centriflaken pipeline for execution:**

Finally, click on the "Execute" button to submit the pipeline for analysis.



After the job is submitted, your history should indicate that the job is running with 2 expected outputs: one, a **MultiQC HTML report** and another: a collection of assembled **MAG**s (**Metagenomically assembled genomes**) in **FASTA** format as shown below (highlighted in curly green bracket).

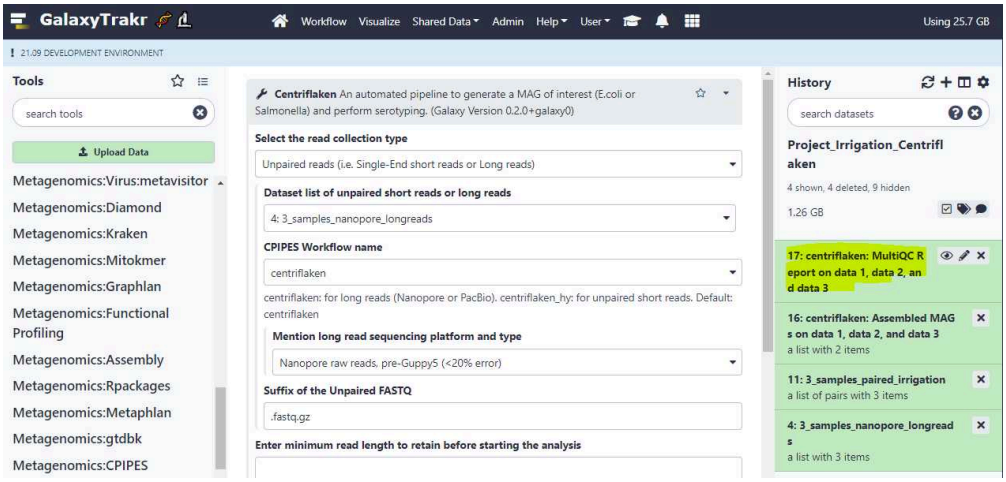


Step 7

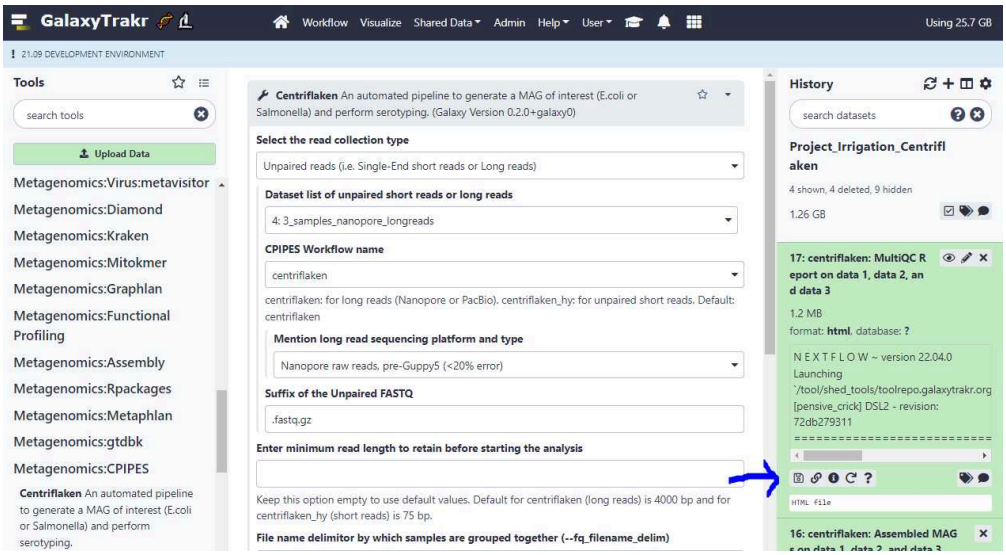
7 **Download results:**

If the pipeline finishes successfully, your history will have 2 outputs: one, a **MultiQC HTML report**, which contains brief summary about the quality of your raw reads and any

results from the analysis, which should be downloaded by first clicking on the results tab to expand the section (highlighted in yellow) as shown below.



Next click on the "**floppy**" icon to download the **MultiQC HTML** report (blue arrow) as shown below.



The downloaded HTML report can be opened in your web browser directly by double-clicking on it.

See an example report generated by the Centriflaken pipeline:

Expected result

centriflaken__MultiQC_Report.html

Note

Note that in the example report, **FAL11342** sample does not have an entry in serotyping and subsequent AMR result tables and the reason is that the **FLYE** assembler failed to generate any **Escherichia coli** related contigs for this sample. The **Centriflaken** pipeline gracefully ignores such failures and proceeds to work on any samples that have a valid contig assembly.

Step 8

8 **Example data sets:**

You can use the following example data sets to try out the **Centriflaken** pipeline.

Oxford Nanopore long reads (Escherichia coli):

FAL11341.fastq

FAL11127.fastq