Galaxy Genome Trakr User Guide

Document Version Number: 5.0
Document Version Date: 07/21/2021

Version History

Version Number	Implemented By	Revision Date	Approved By	Approval Date	Description of Change
1.0	BIS Support Staff	08/24/2017			Initial draft.
2.0	BIS Support Staff	09/14/2017			Added SFTP and SPAdes
3.0	Justin Payne	09/18/2017			Updates to tool layout; QUAST
4.0	Justin Payne	02/07/2018			SNP-Pipeline
5.0	Arsh Randhawa	07/21/2021			Removed FDA/CFSAN References

Table of Contents

List of Figures

1 INTRODUCTION

1.1 Galaxy Genome Trakr Information

The GenomeTrakr program currently supports whole-genome sequencing (WGS) of foodborne pathogens at more than 25 state public health and academic laboratories. The network of laboratories now routinely generates more than 1,000 isolates each month for isolates origination from food, environmental, and clinical sources.

GalaxyTrakr.org was implemented to allow laboratories to locally perform quality assessment of their sequence data and look for links between clinical isolates and positive food/environmental samples.

Galaxy, an open-source commercial license free platform, will be used as a packaging tool, GUI, and hosted runtime environment for bioinformatics software projects that will be leveraged by state and local labs.

1.2 Purpose

The purpose of this document is to outline the critical information for all end users that leverages GalaxyTrakr.org.

2 ONBOARDING

This section outlines the onboarding information required to gain access to the GalaxyTrakr.org environment.

2.1 Initial Access

The purpose of this section is to detail the password change procedure, which is required to be completed prior to first login. Please complete the following steps once initial login information has been received:

1. Open a browser to https://account.galaxytrakr.org.
2. Use the form, depicted in Figure 1, to change the temporary password that was distributed:
    a. Username: Distributed via email
    b. Password: Temporary password distributed via email
    c. New Password: A password specified by the user that to be used for future logins
    d. Confirm Password: The password specified in **New Password**

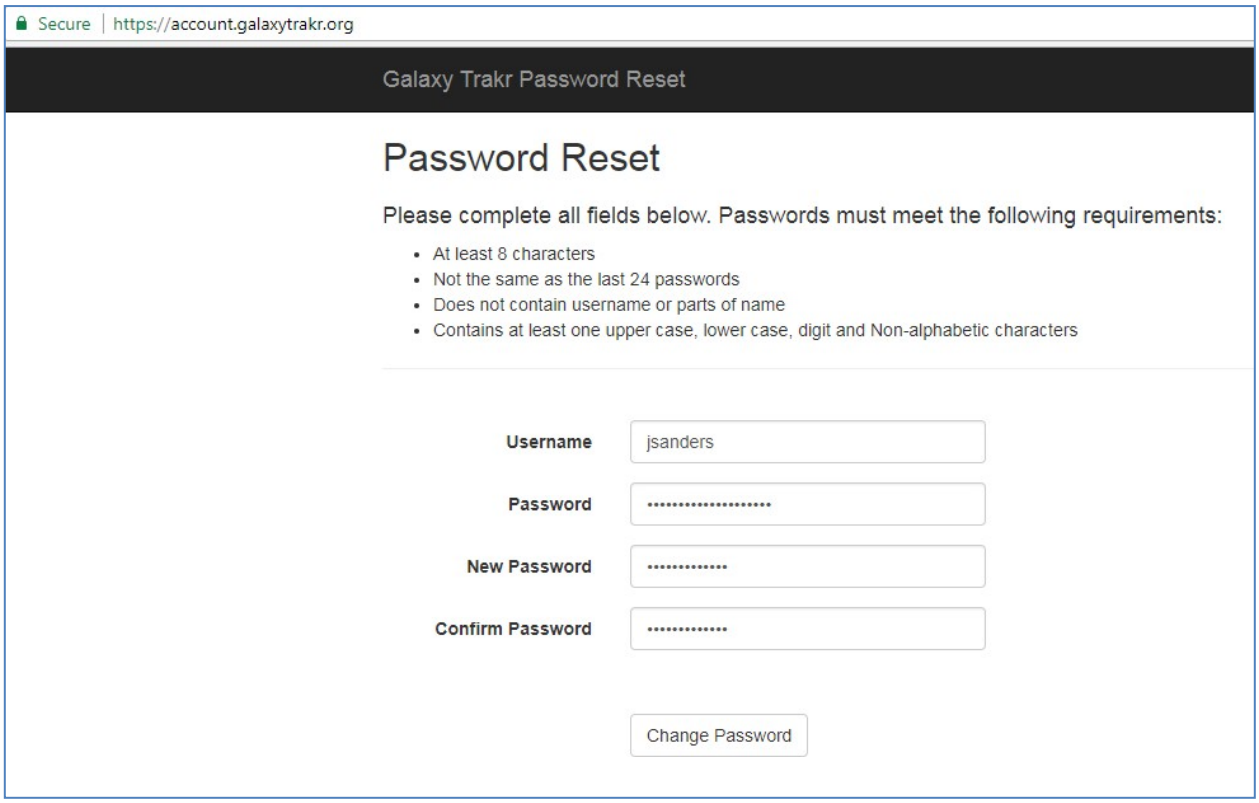


Figure 1. Galaxy Trakr Password Reset

2.2 Access to GalaxyTrakr.org

Please complete the following to gain access to Galaxy Genome Trakr.

1. Open a browser to https://galaxytrakr.org.
   Please note this URL is different than required for the initial access.
2. On the top right, click **Login**.
   See Figure 2.
3. Enter username and password and click **Login**.
   See Figure 3.

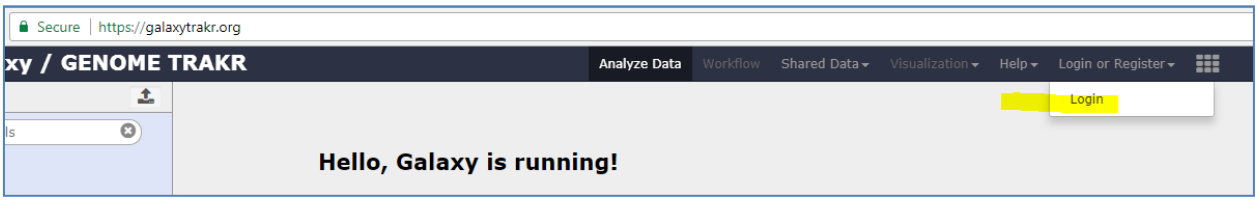


Figure 2. GalaxyTrakr.org Login

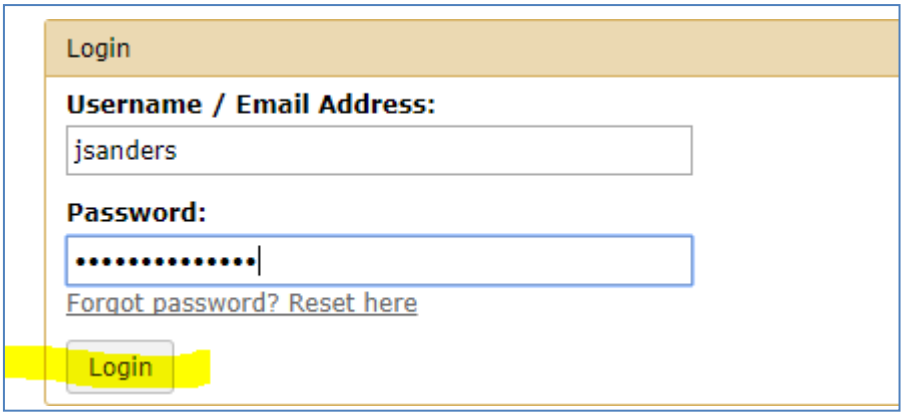


Figure 3. Username and Password

3 USING GALAXY- QUICK START GUIDE

The following section provides instructions on how to get started with Galaxy tools deployed in GalaxyTrakr.org.

3.1 Create and Name a History

Once logged into GalaxyTrakr.org, please follow these steps to create and name a history in SeqSero:

1. On the top right corner, click the cog ( ) icon.
2. Select **Create New.**
   See Figure 4.

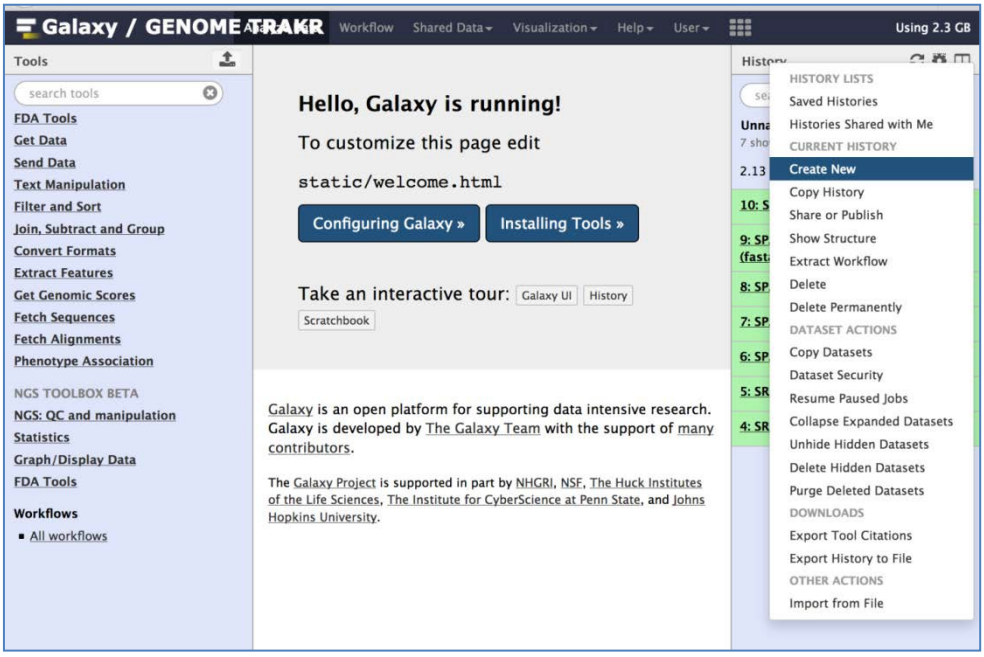


Figure 4. Create a New History

3. To name your history, click **Unnamed History** on the top right of the screen and type a new name.
4. Press **Enter** on your keyboard.
   See Figure 5.

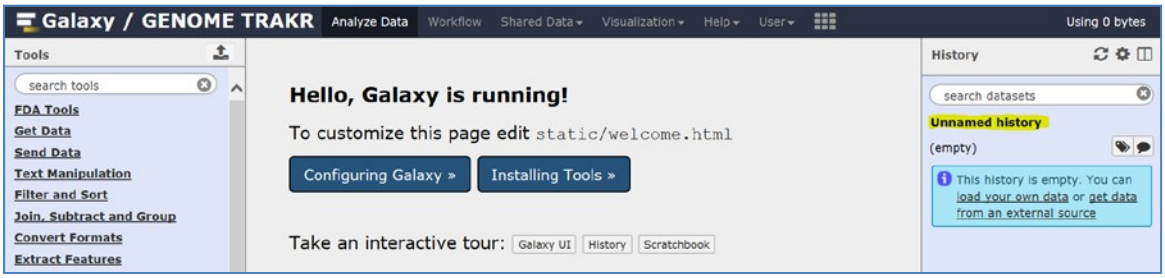


Figure 5. Change the History Name

3.2 Upload Data

3.2.1 Use the SFTP Client

To upload data using a standard SFTP client, such as Filezilla, follow the steps below:

1. Open SFTP compliant client.
2. Enter the following connection information:
    a. Host: sftp://upload.galaxytrakr.org
    b. Username (same as used to access Galaxy)
    c. Password (same as used to access Galaxy)
    d. Port: 443 or 22

    See Figure 6.

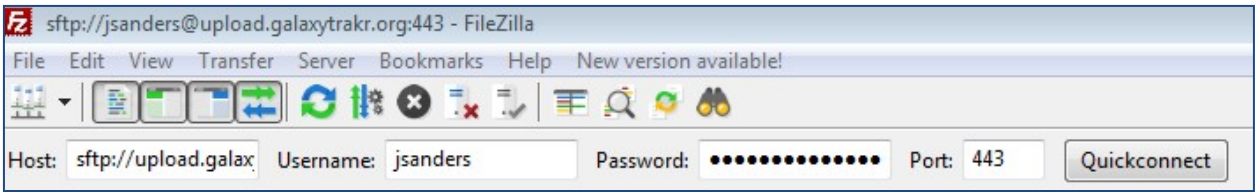


Figure 6. Log into SFTP client

3. The first time connecting you will be asked to trust the host being connected to. Click **OK** to trust the connection.
    See Figure 7.

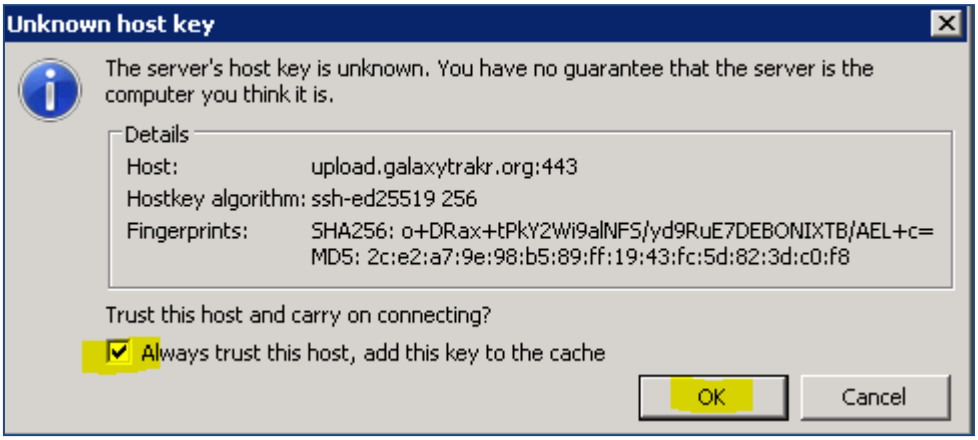


Figure 7. Uknown Host Key Message

4. Depending on the client, click **Connect** or **Quickconnect**.

5. Once connected drag the files to upload from the source to the connected server.
   The data uploads into the folder, which is the default for galaxy users.
   See Figure 8.

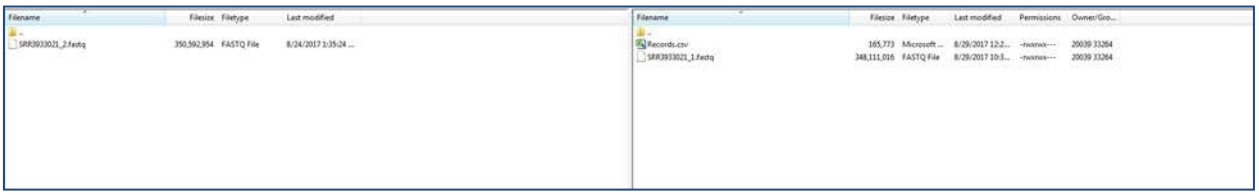


Figure 8. Files to Upload

6. Login to the galaxy web interface at https://galaxytrakr.org
7. Navigate to **Get Data** and click the **Upload File** link**.**
   See Figure 9.

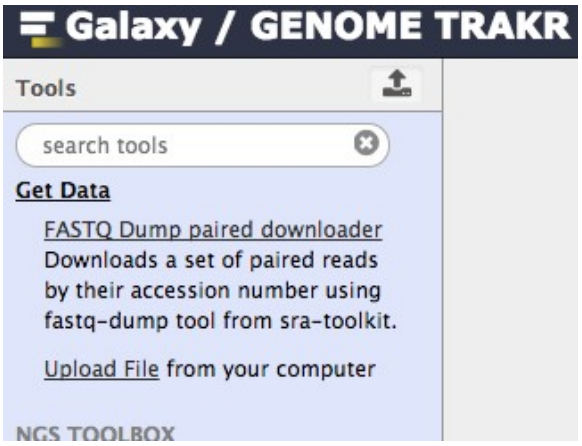


Figure 9. Get Data

8. Click the **Choose FTP File** button.
   See Figure 10.

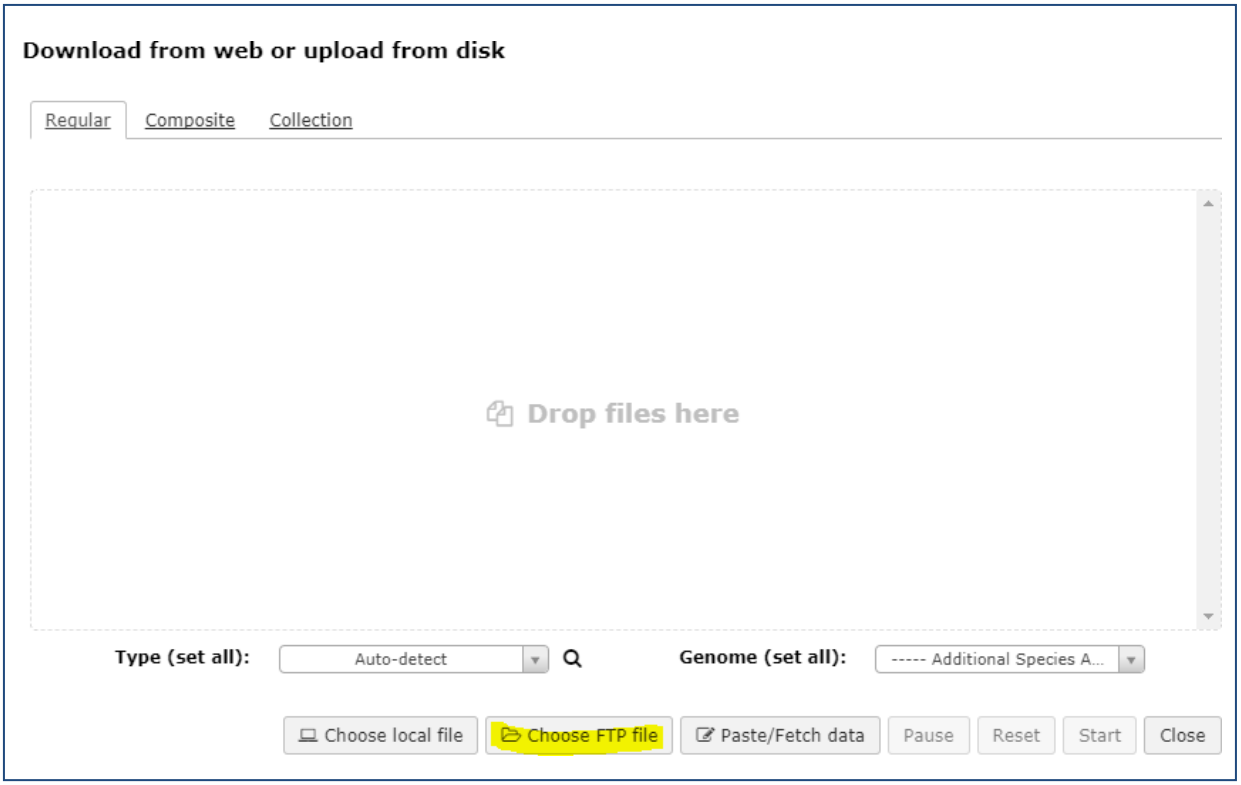


Figure 10. Choose FTP File

9. Select the file that was uploaded or files in the upload directory that you need to import.
   See Figure 11.

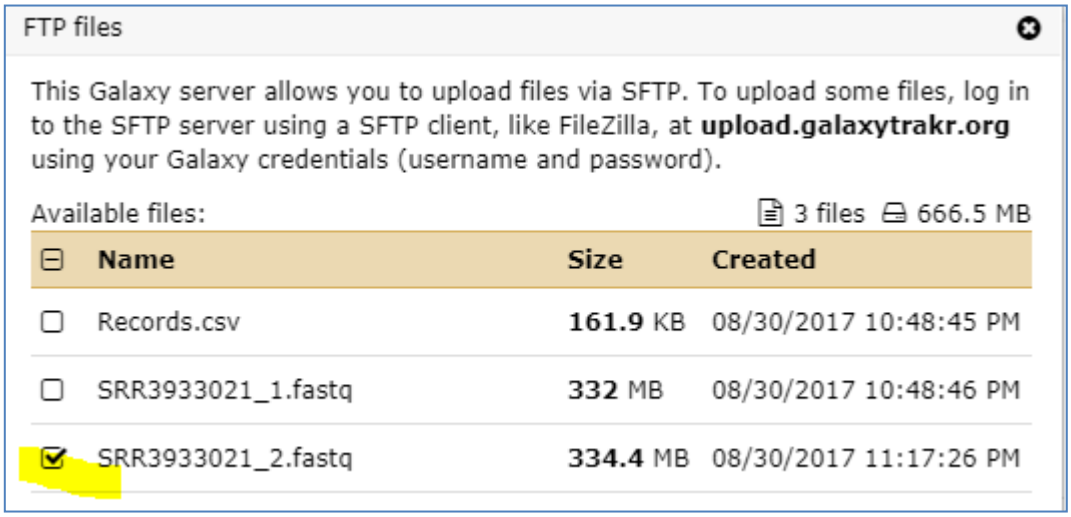


Figure 11. Select Uploaded Files

10. Click the **Start** button and observe the import status.
    See Figure 12.

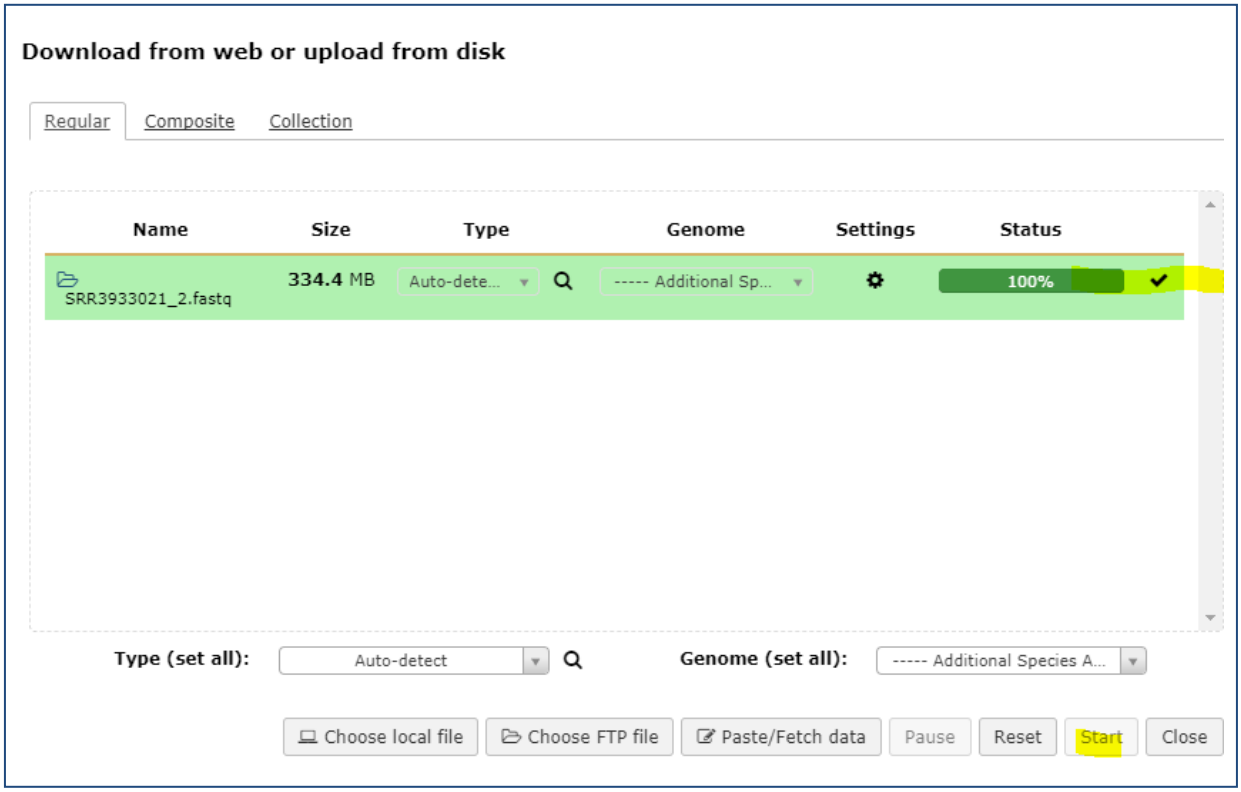


Figure 12. Upload Status

3.2.2 Use the Web Interface

To upload data to your new history, follow the steps below:

1. Click on the download icon ( ) on the top of the left menu.
2. Select **Choose local file** from the pop-up menu and navigate to your desired file. See Figure 13.

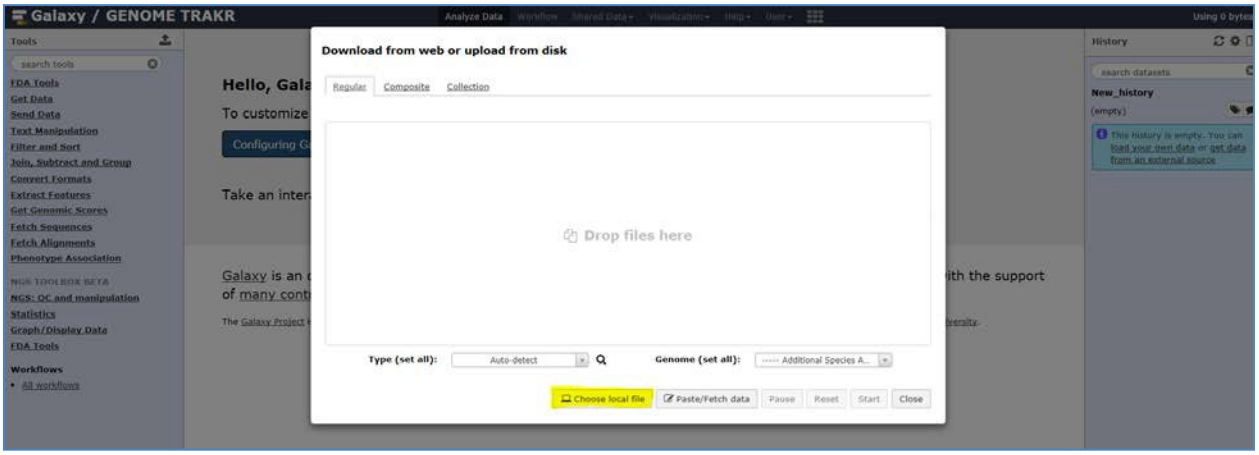


Figure 13. Choose local file

3. Select the paired end read files to be used and click **Open.** See Figure 14. Please note that files can also be dragged into Galaxy from your file explorer.

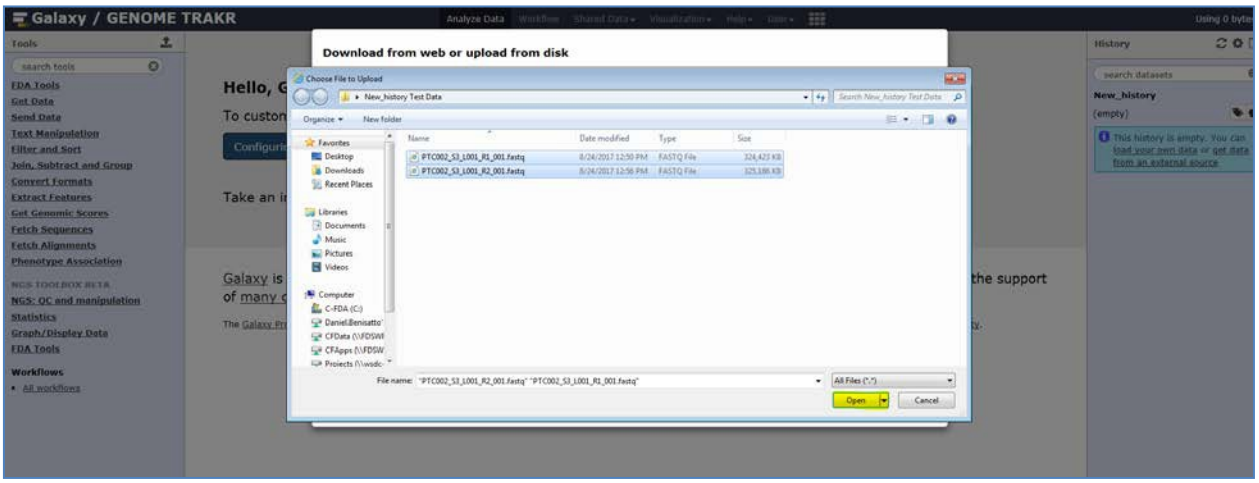


Figure 14. Select a local file

4. Click **Start** to begin uploading your files to the Galaxy server. See Figure 15.

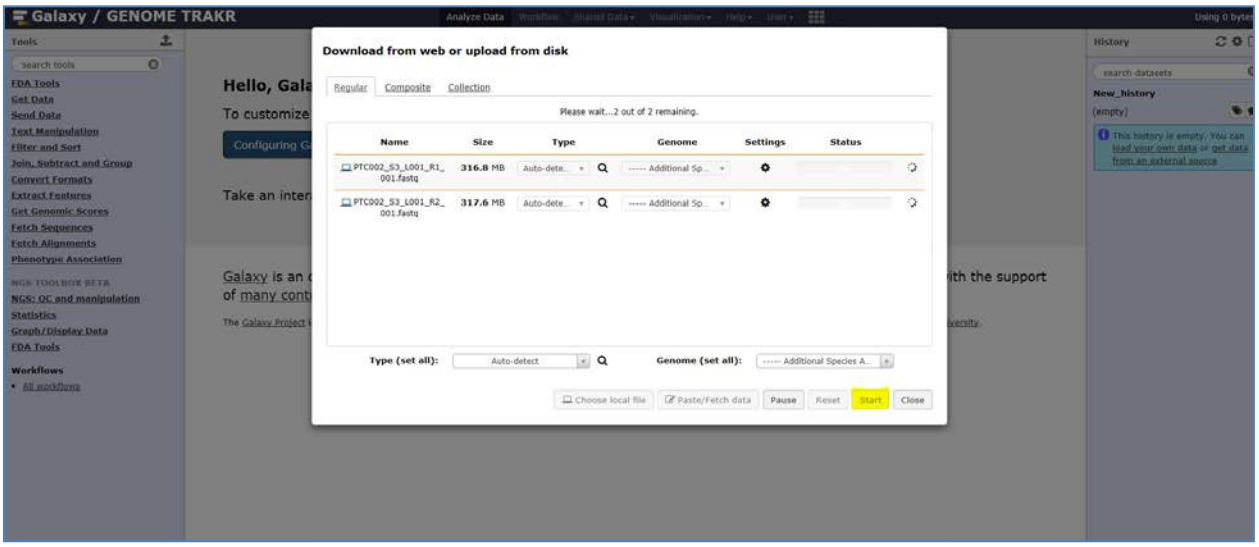


Figure 15. Uploading files into the Galaxy server

Figure 16 below depicts uploads in progress.

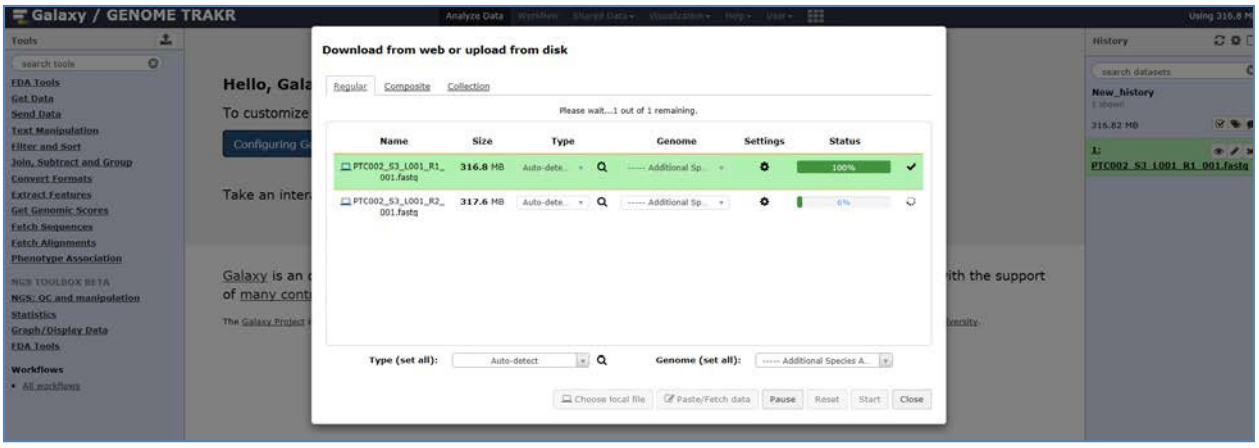


Figure 16. Data upload in progress

Once completed, the files will be visible in the history. This can be seen on the right side of the screen. See Figure 17.

Figure 17. History View

3.3 Change History

When working with multiple histories, it is easy to switch back and forth. Please use the following steps to change histories:

1. Click the book icon ( ) in the upper right corner.
2. Select the history you would like to use by clicking **Switch to**.
   See Figure 18.

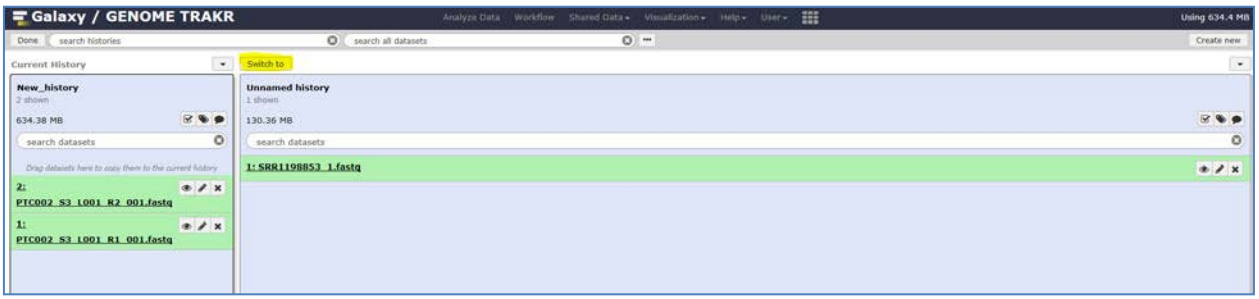


Figure 18. Switch to a different history

3. Click **Done.**
   See Figure 19.

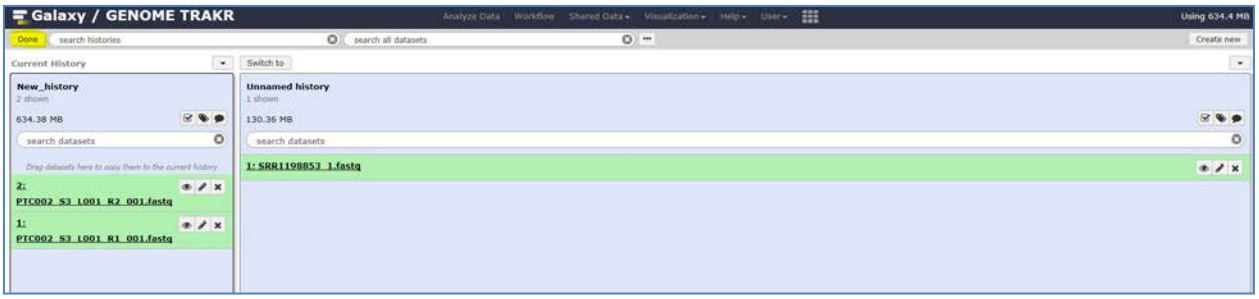


Figure 19. Finalizing history selection

3.4 Share a Data Set

Follow the below steps to share a data set:

1. Make sure the current library contains the data you want to share.
2. Click **Shared Data** and then click **Data Libraries**.
   See Figure 20.

Figure 20. Data Libraries

3. Select the library that the data set will be shared with.
4. Create a new folder.
   See Figure 21.

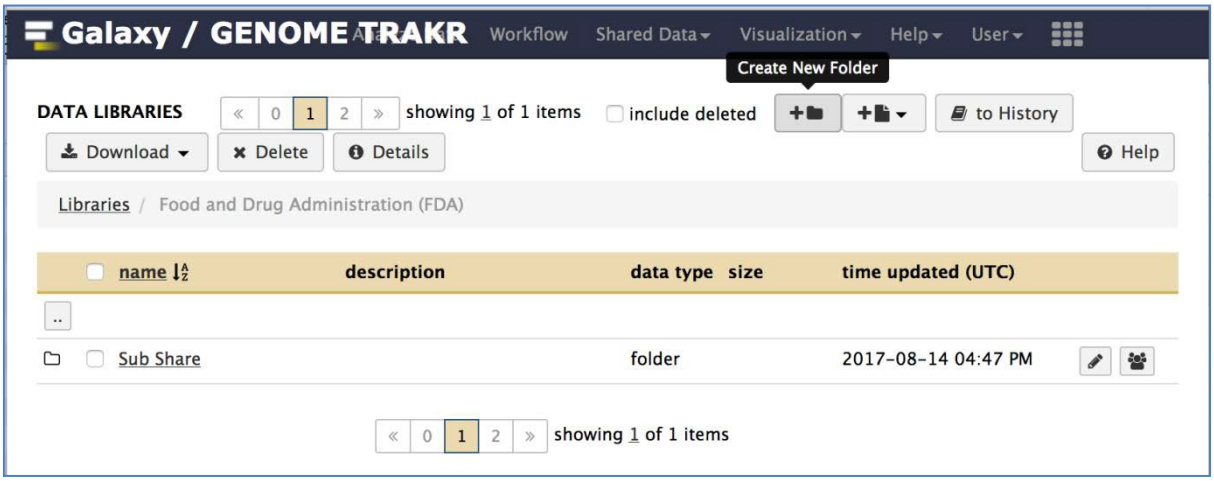


Figure 21. Create New Folder

5. Enter a name for the new folder.
6. Click the +data icon ( ) and select **from History**.
   See Figure 22.

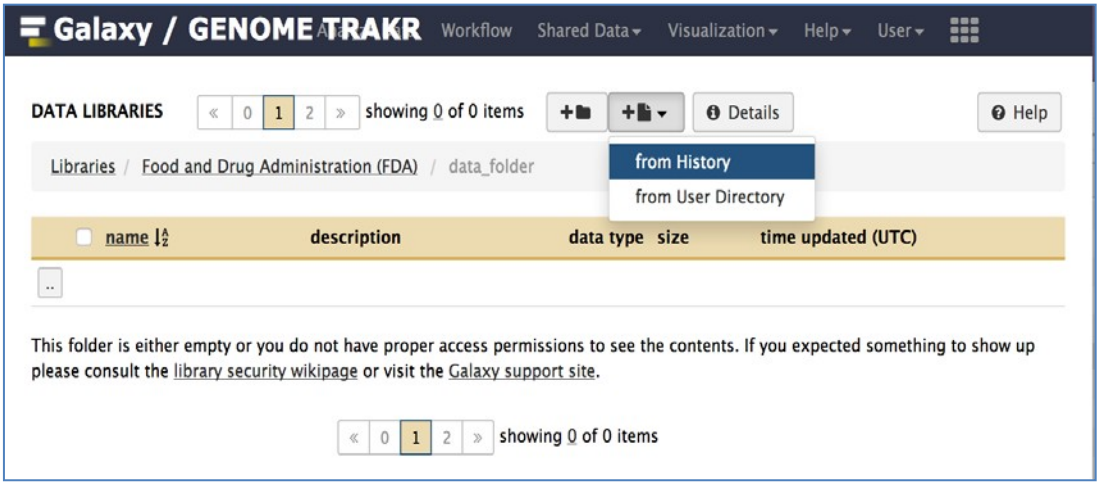


Figure 22. Add data set from history

7. Select the desired data sets and click **Add**.

3.5 Import and Queue Data Sets to History

This section details the necessary steps to retrieve and execute data sets from participating Galaxy Genome Trakr collaborators. This function enables approved users rights to access both new and archived data sets on-demand.

To import and queue data sets to history, follow these steps:

1. On the GalaxyTrakr.org home page, select **Shared Data**.
   See Figure 23.

Figure 23. Shared Data

2. Select **Data Libraries**.
   See Figure 24.

Figure 24. Data Libraries

3. Select your target **Laboratory Library** from **Galaxy Data Libraries** page.
   See Figure 25.

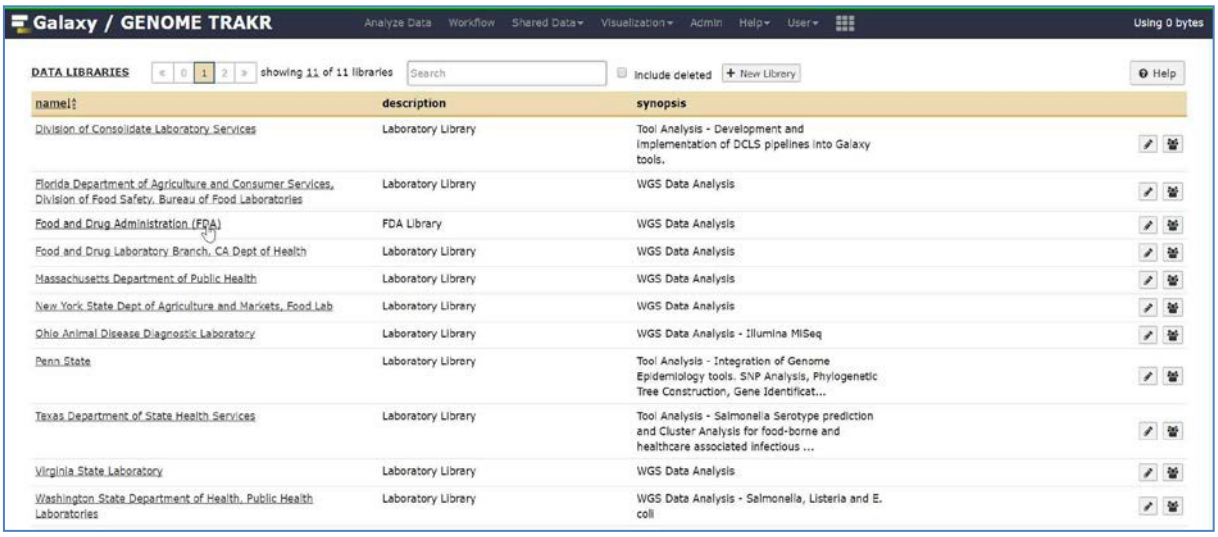


Figure 25. Galaxy Data Libraries

4. **Select the checkbox** to the left of the **Target Data Folder**.
   See Figure 26.

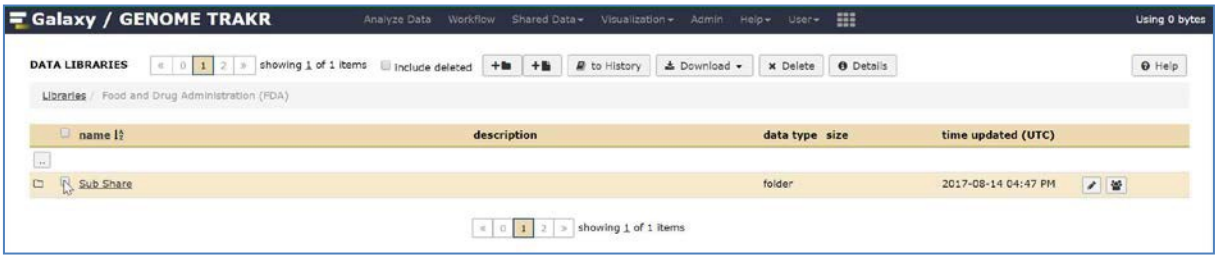


Figure 26. Target Data Folder

5. Once the data folder is selected, click the **Import Selected Datasets into History** button.
   See Figure 27. This will add the data to your Pending Queue on the Galaxy Genome Trakr home page.

Figure 27. Import Selected Datasets into History

6. Select an existing history for the data set to be imported into.
   See Figure 28.

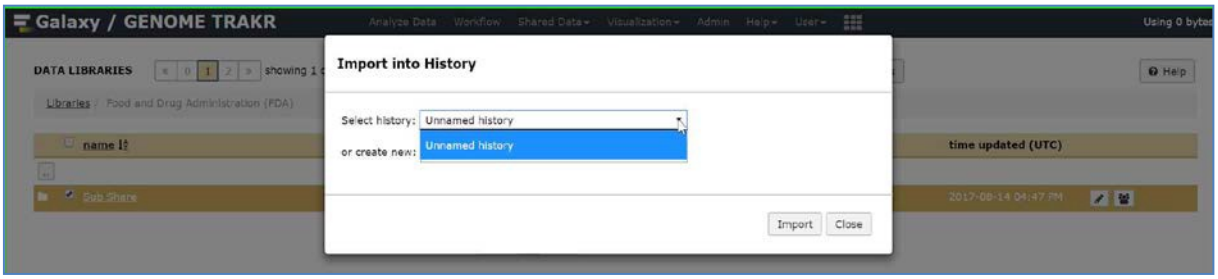


Figure 28. Select an existing history

7. If one does not already exist, create a new history by entering in a unique and identifiable history name.
   See Figure 29.

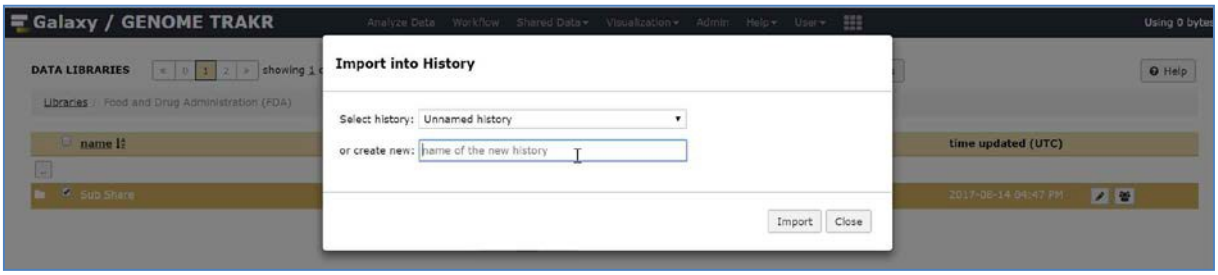


Figure 29. Create a new history

8. Click **Import**.
9. Once you select import, the data set is visible on the Galaxy Genome Trakr home page on the right side of the screen.
   See Figure 30.

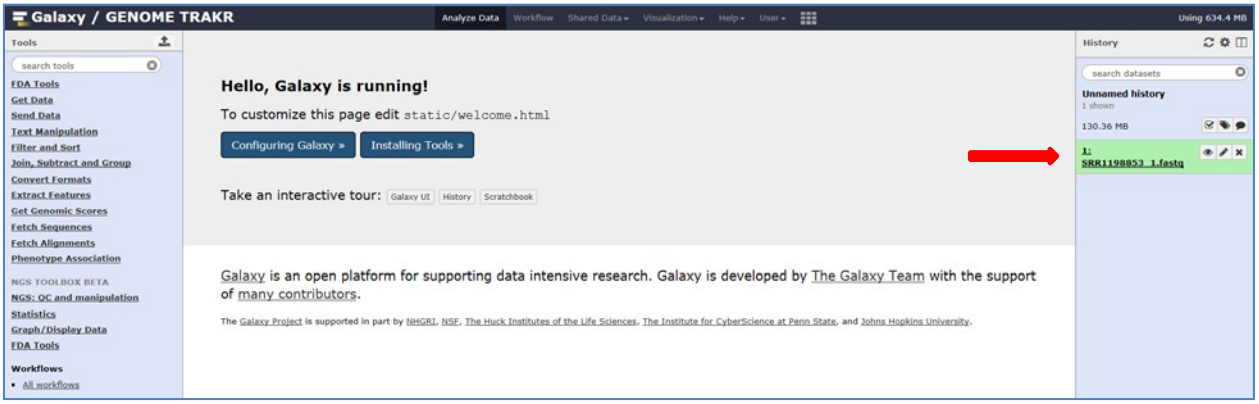


Figure 30. Visible Data Set

4 SEROTYPE PREDICTION WITH SEQSERO

SeqSero uses whole genome sequence (WGS) data to predict Salmonella enterica serotypes. SeqSero achieves such through the following:

- Maps read to database of antigen alleles using Burrows-Wheeler Aligner (BWA) in multiple steps.
- Chooses alleles best mapped-to by the most reads.
- Uses Basic Local Alignment Search Tool (BLAST) to clear up ambiguities.
- Allelic antigen profile is matched to Kaufmann-White serotypes, where known.

Follow the steps to execute a SeqSero run:

1. In the left navigation pane, click **NGS: Screening and Prediction.**
   See Figure 31.

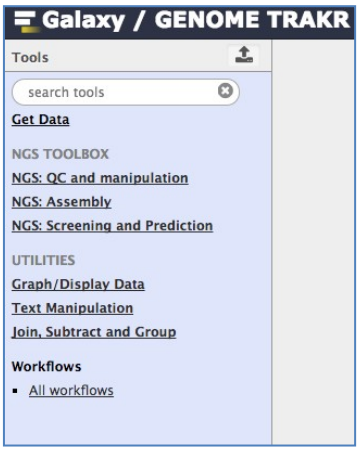


Figure 31. GenomeTrakr Tools

2. Click **SeqSero Batch – Paired-End Reads**.
   See Figure 32.

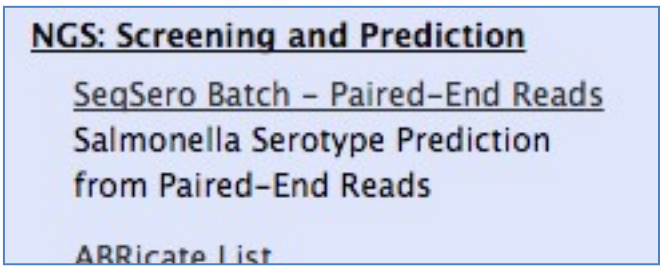
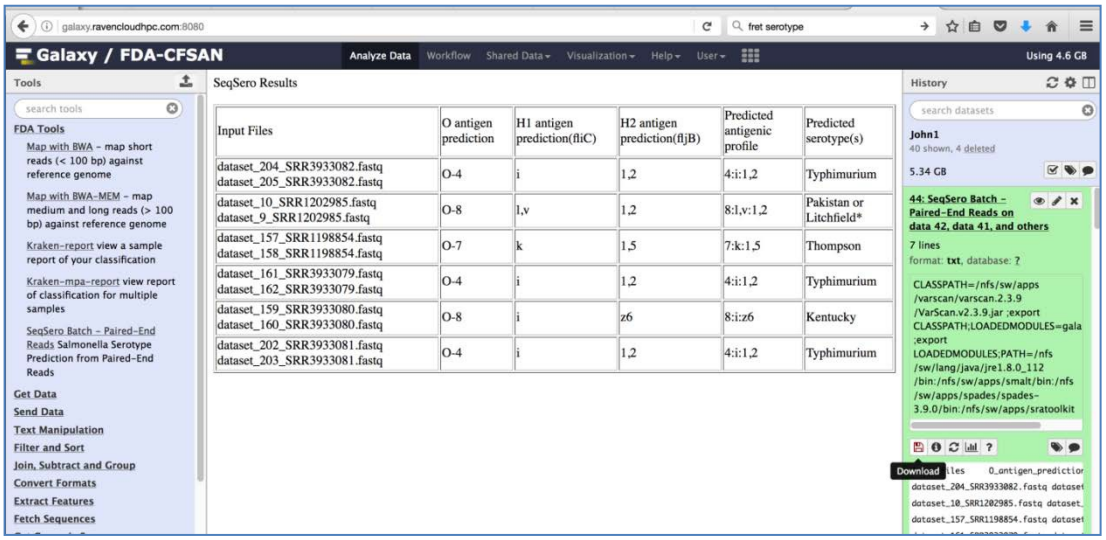


Figure 32. SeqSero Batch – Paired-End Reads

3. Select the pairs of desired sequencing runs and click **Execute.**
See Figure 33.

Figure 33. Select runs and Execute

4. If the run was successful, you a notification displays.
5. Click the eye icon ( ) at the upper right box to view a table of your results.
See Figure 34.



Input Files	O antigen prediction	H1 antigen prediction(fliC)	H2 antigen prediction(fljB)	Predicted antigenic profile	Predicted serotype(s)
dataset_204_SRR3933082.fastq dataset_205_SRR3933082.fastq	O-4	i	1,2	4:i:1,2	Typhimurium
dataset_10_SRR1202985.fastq dataset_9_SRR1202985.fastq	O-8	l,v	1,2	8:l,v:1,2	Pakistan or Litchfield*
dataset_157_SRR1198854.fastq dataset_158_SRR1198854.fastq	O-7	k	1,5	7:k:1,5	Thompson
dataset_161_SRR3933079.fastq dataset_162_SRR3933079.fastq	O-4	i	1,2	4:i:1,2	Typhimurium
dataset_159_SRR3933080.fastq dataset_160_SRR3933080.fastq	O-8	i	z6	8:i:z6	Kentucky
dataset_202_SRR3933081.fastq dataset_203_SRR3933081.fastq	O-4	i	1,2	4:i:1,2	Typhimurium

Figure 34. SeqSero Results

5 GENOMIC ASSEMBLY WITH SPADES

SPAdes (St. Petersburg genome assembler) is a high-performance de Bruijin-graph assembler for single or multi-cell libraries with single-end, paired-end, or mate-pair layouts. SPAdes produces draft assemblies useful for genomic annotation, antibiotic resistance prediction, and other gene-finding tasks.

Follow the steps below to use the SPAdes Genome Assembler:

1. Once data is uploaded into Galaxy Genome Trakr, access SPAdes, which is located under the **NGS Toolbox**.
2. Select **NGS: Assembly** and then click **SPAdes**.
   See Figure 35.

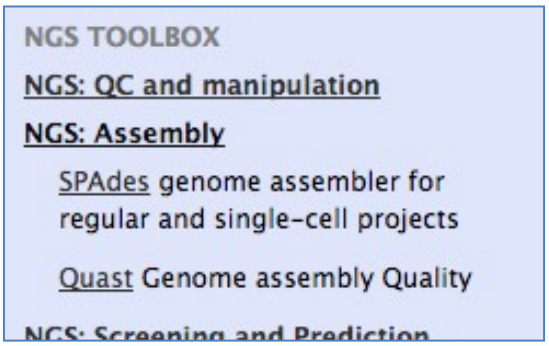


Figure 35. Locating SPAdes

3. Select **Library Type, Orientation,** and **Reads** for your genome assembly.
   Note: It is possible to select more than one library and file pairs.
   See Figure 36.

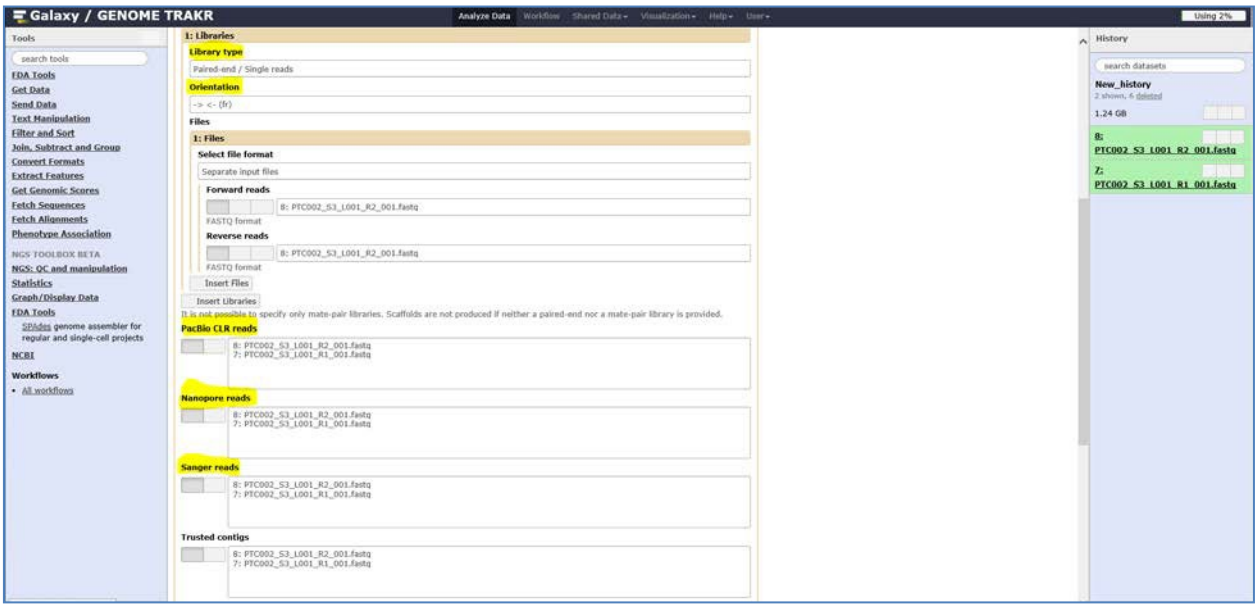


Figure 36. SPAdes Input

4. Galaxy suggests default k-mer values of 21,33,55; you may either run with these values, supply your own, or enable “Automatically choose k-mer values” to allow SPAdes to determine the optimum length based on your reads data. This is typically the best option.
5. At the bottom of your screen, click **Execute.**
   See Figure 37.

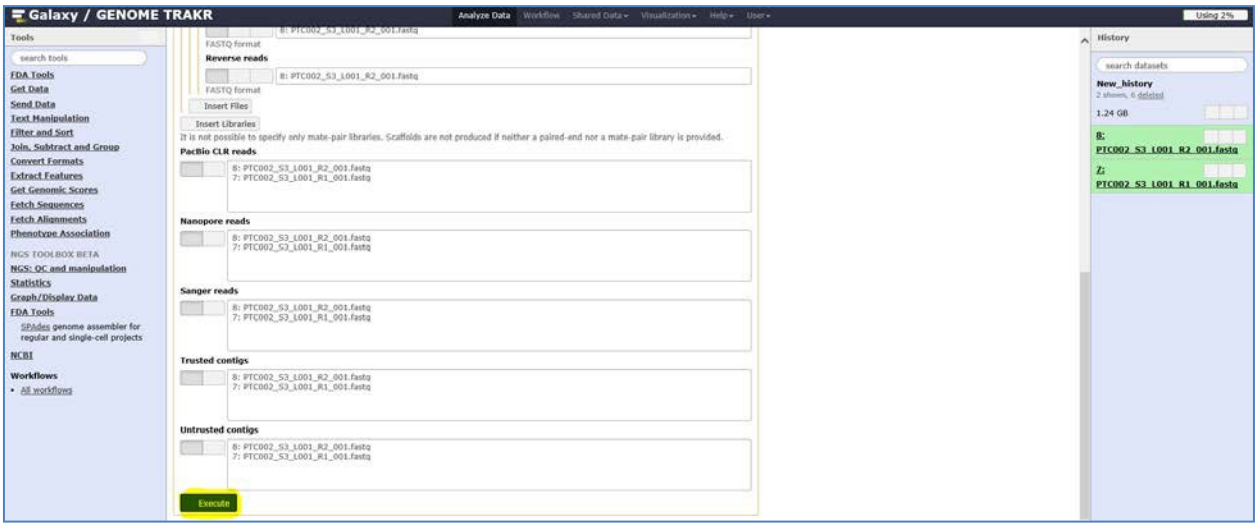


Figure 37. Execute SPAdes

6. SPAdes produces the following files:
   - **SPAdes Log:** A log of activity that shows everything SPAdes does.
     See Figure 38.

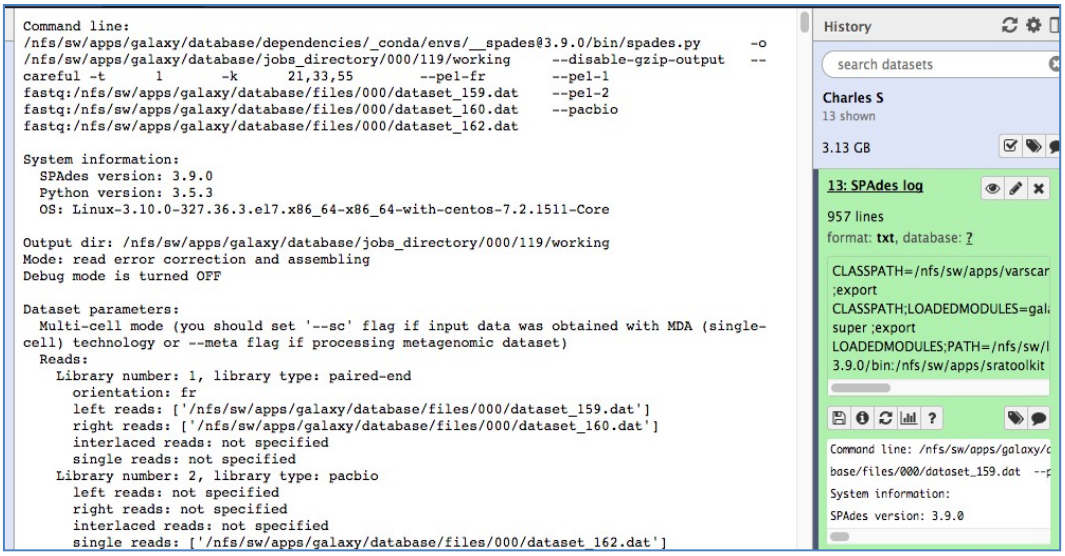


Figure 38. SPAdes Log

- **Spades Contigs (fasta):** Unscaffolded contigs in order by descending length. Each contig begins with a *defline* that reports length in bases and the estimated coverage depth. See Figure 39.

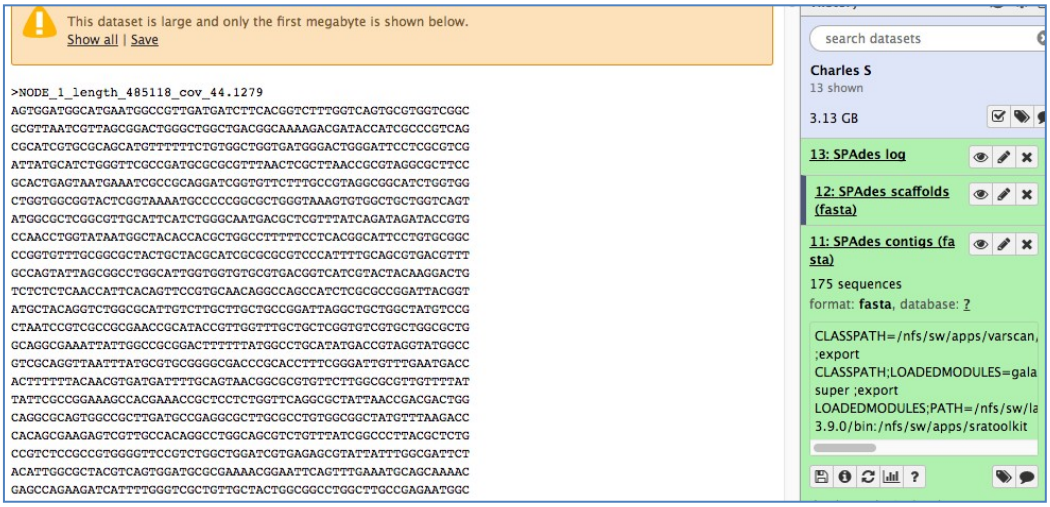


Figure 39. SPAdes Contigs

- **Spades Scaffolds (fasta):** Scaffolded contigs. SPAdes will attempt to use the paired-end relationship between reads to orient contigs relative to each other along the genome. Regions of unknown sequence between contigs (gaps) are bridged by poly-*N* sequences.
See Figure 40.

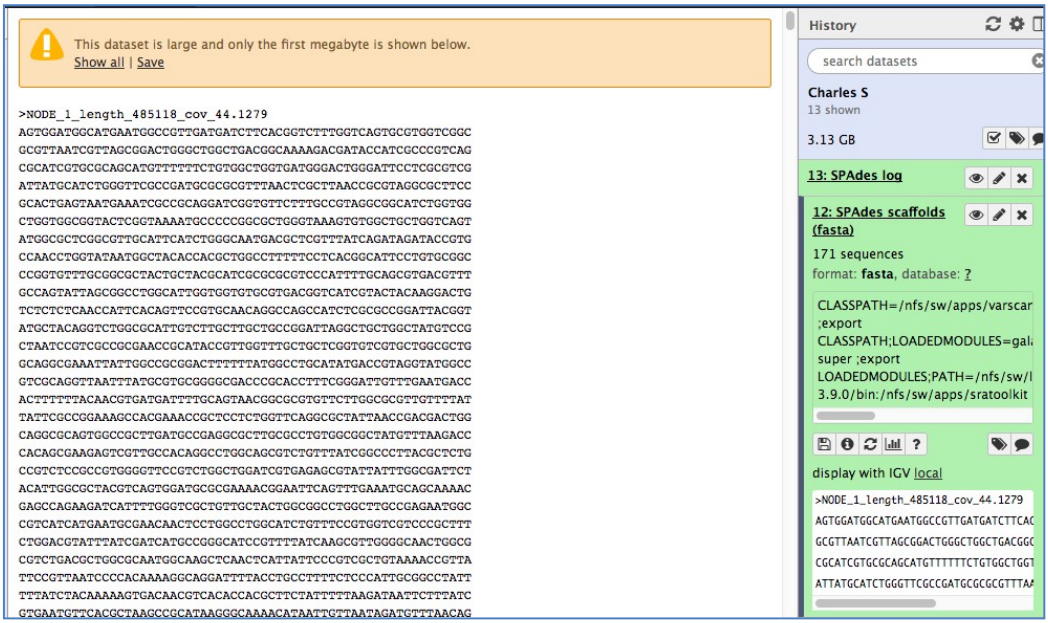


Figure 40. SPAdes Scaffolds

- **SPAdes Scaffold Stats:** Provides the length of the Scaffold files. See Figure 41.

name	length	coverage
#name	length	coverage
NODE_1	59188	77.3447
NODE_2	45576	86.9273
NODE_3	36214	77.1701
NODE_4	34221	54.8056

Figure 41. Scaffold Stats

- **SPAdes Contig Stats:** Provides the length of the Contig files. See Figure 42.

name	length	coverage
#name	length	coverage
NODE_1	48007	80.6177
NODE_2	45576	86.9273
NODE_3	36214	77.1701

Figure 42. Contig Stats

Further information on SPAdes and its output can be found in the SPAdes manual: http://spades.bioinf.spbau.ru/release3.9.0/manual.html

6 ASSEMBLY CHARACTERIZATION WITH QUAST

Use the QUAST (Quality Assessment Tool for genome assemblies) to find the *N50* of a genome assembly and gather other metrics on quality and contiguity. QUAST can assess either contig or scaffold data in FASTA format.

Follow these steps to assess an assembly with QUAST:

1. Select **NGS: Assembly** and then select **QUAST**. See Figure 43.

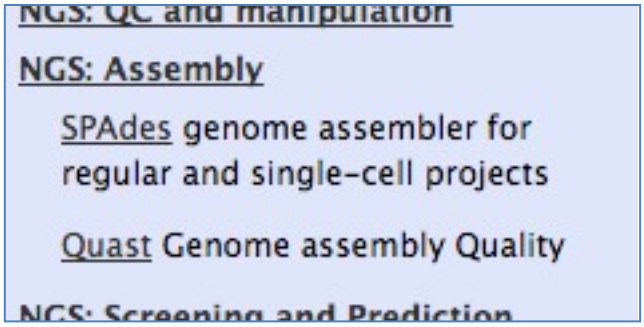


Figure 43. QUAST in the Galaxy Toolbar

2. Select one or more datasets.
   Users can also chose to provide a reference assembly and/or gene annotation file (in GFF2/3 BED, or ASN.1 format). See Figure 44.
3. Select **Execute**.

   QUAST can build quality statistics on both contig and scaffold FASTA assemblies. If multiple assemblies are provided, QUAST will compare and rank them; this is a useful way to compare the performance of assemblers.

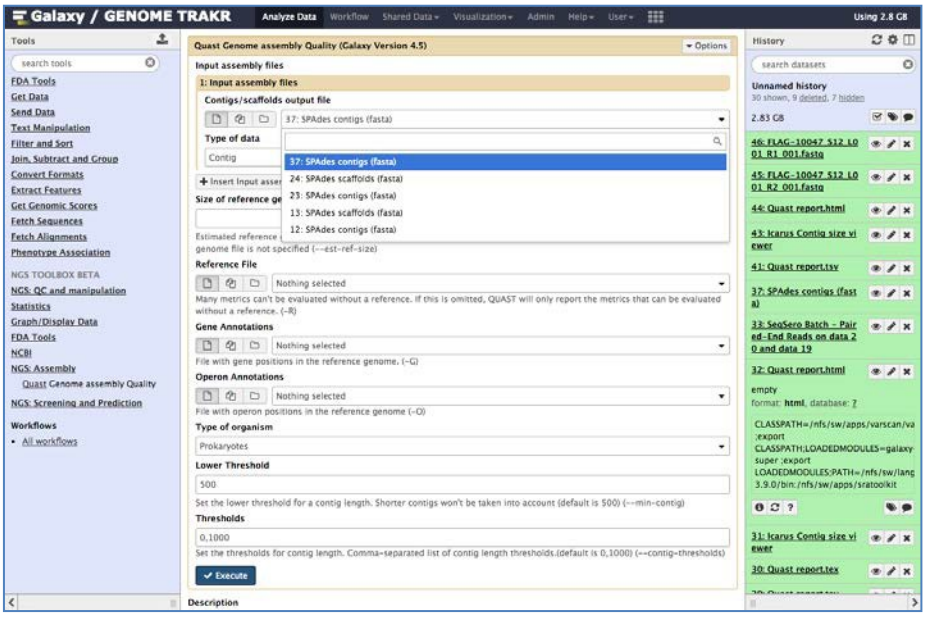


Figure 44. QUAST configuration

6.1 Outputs of QUAST

QUAST produces output datasets as follows:

- **QUAST Report.html:**
  An interactive HTML5 report is produced with summary statistics and plots, including a contig count, N50 (the contig length such that the set of contigs this long or longer contain at least half of the bases in the assembly; a measure of assembly contiguity), G/C content, and other metrics. Detailed explanations of the summary metrics are given as mouseover tooltips. See Figure 45 and Figure 46 below.

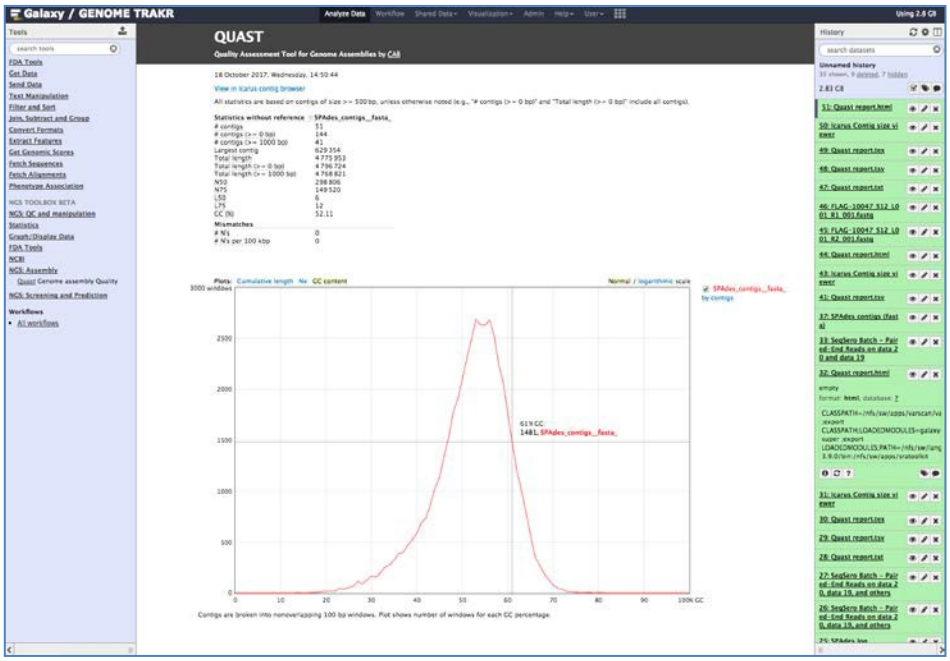


Figure 45. QUAST interactive HTML report

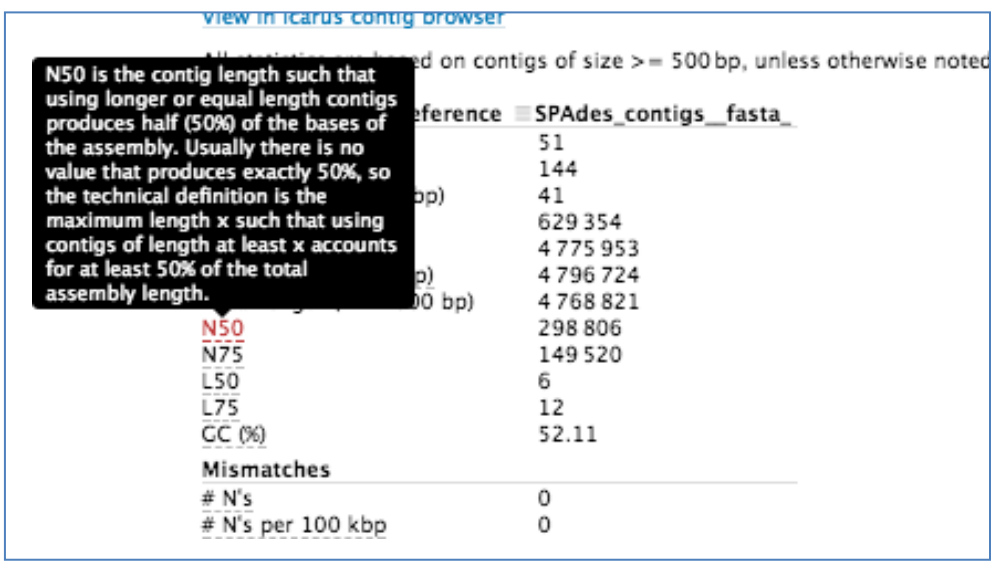


Figure 46. Summary statistics tooltips

- **Icarus Contig Size Viewer:**

An interactive contig length distribution viewer shows the distribution of contigs lengths in the assembly, as well as “landmarks” such as the N50 and N75 of the assembly. If you ran QUAST on multiple assembly files, they’ll be compared in “tracks”, one above the other. In the lower track view, you can drag the yellow viewbox left and right to move the upper viewing window. See Figure 47.

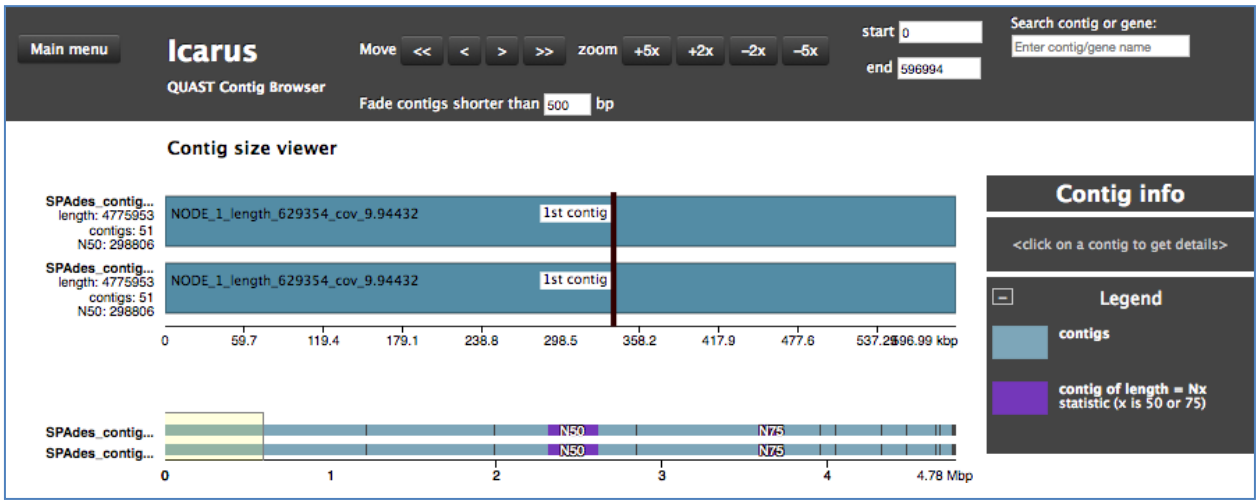


Figure 47. Icarus Contig Size viewer

- **QUAST Report:**
  Quast report.tex, Quast report.tsv, Quast report.txt give the same summary stats as the HTML report, but in additional structured formats (LaTeX, TSV, and ASCII text).

You can find additional information on the use of QUAST to assess assembly quality in its online manual: http://quast.bioinf.spbau.ru/manual.html.

7 USING THE SNP PIPELINE WORKFLOW

The CFSAN SNP-Pipeline is implemented in GalaxyTrakr as 7 connected stages and is available for use as a shared workflow. This workflow encapsulates the basic SNP-Pipeline functionality, but users should feel free to use the pipeline stages in their own workflows or clone and extend the provided workflow, etc. See Figure 48 for a view of the workflow.

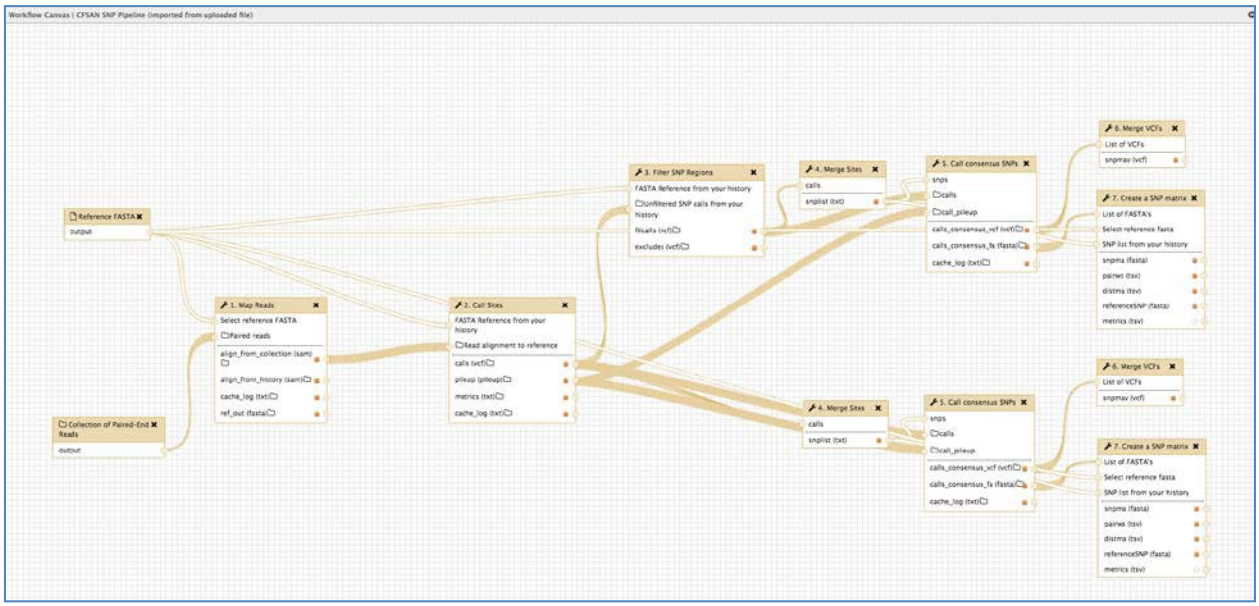


Figure 48. Basic SNP-Pipeline Workflow

Need an intro sentence to these steps—need to know where I am…in the system?

1. Upload paired reads and build a list (paired collection).
2. Set **Collection Type** to **List of Pairs.**
3. Set **File Type** to **fastqsanger**.
   This ensures maximum compatibility with the SNP-Pipeline tools and the rest of the GalaxyTrakr ecosystem.
4. Click **Start** to begin the upload.
   Once upload has been completed, the **Build** button will become available next to **Start**.
5. Select **Build**. See **Error! Reference source not found.**.

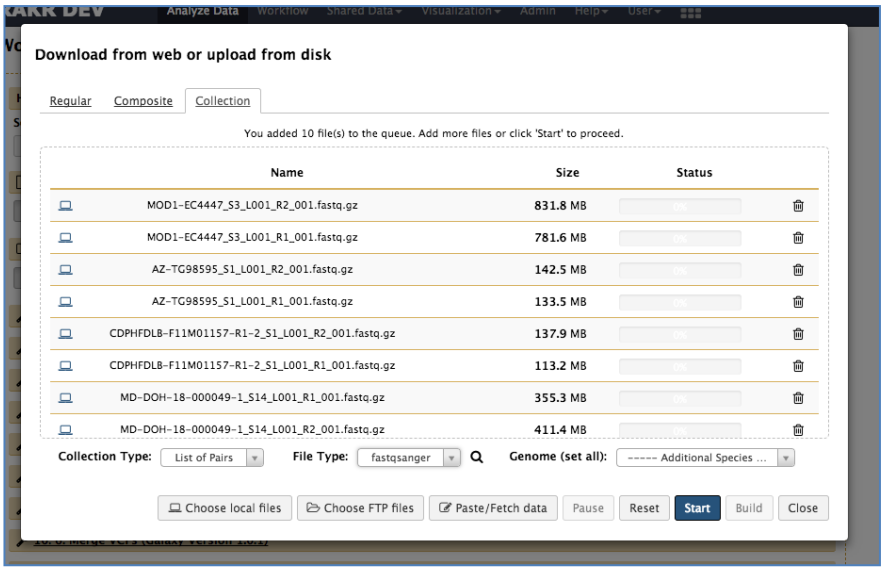


Figure 49. Uploading files for a collection of paired reads

6. From the **Workflow** screen, click **CFSAN SNP Pipeline** to show the contextual workflow menu.
7. Select **Run** to initiate the workflow.

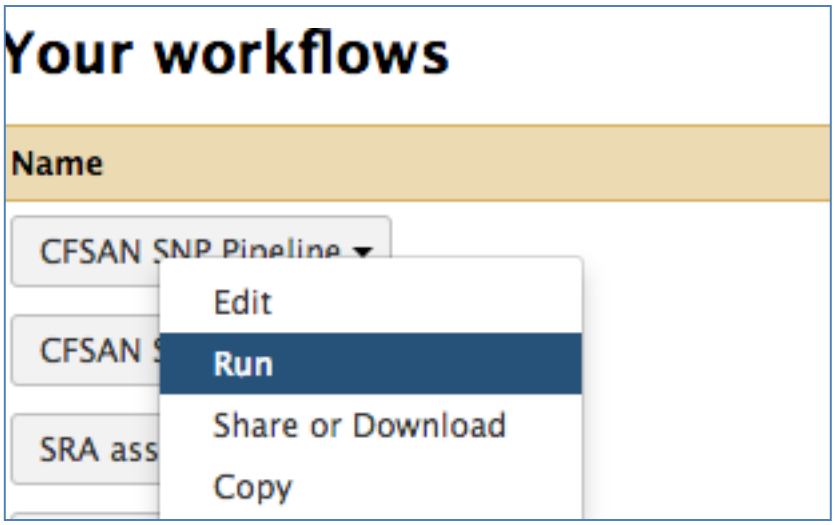


Figure 50. Run the workflow

Users will have a chance to configure the workflow stages, but the default options are pre-configured for most use cases. Configure your inputs by selecting a FASTA from your history as a reference, and a collection of paired-end reads from your history as input. See Figure 51.

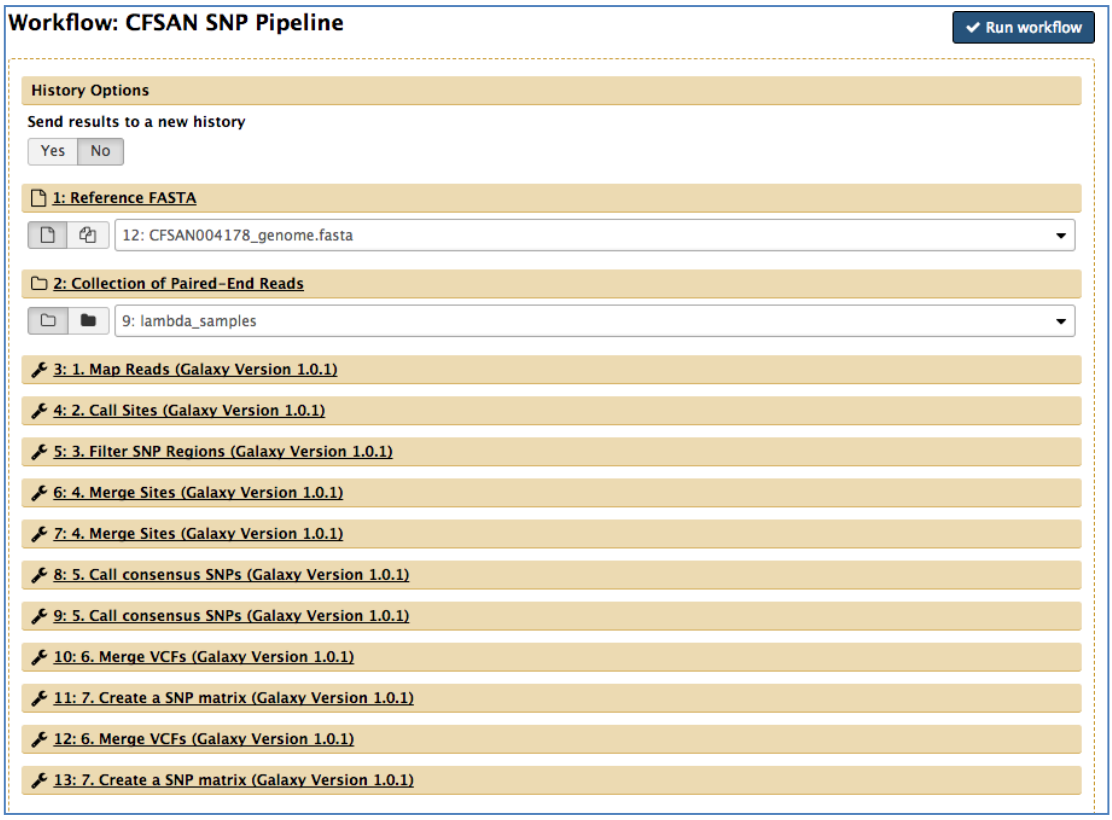


Figure 51. Inputs to the pipeline

7.1 Additional SNP Pipeline Information

The following identifies helpful information regarding the use of a reference based SNP analysis pipeline.

- Running the workflow will produce about 30 datasets in your history per sample, but after execution these will collapse into collections or be hidden. It can be helpful to execute the pipeline on a new history, just to keep things organized.

  The workflow branches into a filtered and unfiltered flow approximately halfway through, and subsequent results are tagged with **filtered** or **unfiltered** depending on which of those branches they are produced by. The filtering is the result of ignoring SNPs promimal to the ends of reads and in regions in which many SNP's are found in proximity. For an in-depth description of the region-based filtering step in the SNP Pipeline, please follow the below link: http://snp-pipeline.readthedocs.io/en/latest/usage.html#snp-filtering

- The SNP Pipeline generates most, if not all, of the analytic outputs described at the following link:
  http://snp-pipeline.readthedocs.io/en/latest/usage.html#outputs
  However, many job execution metrics are not produced because of differences in the way Galaxy is used as a job scheduler. Individual job metrics can be viewed in the Galaxy interface by expanding the dataset and clicking the **View Details** button. See Figure 52.

	SRR1822544	SRR2178118	SRR3113782	SRR3372017	SRR3545396
	SRR1822544	SRR2178118	SRR3113782	SRR3372017	SRR3545396
SRR1822544	0	15	29	47	24
SRR2178118	15	0	25	42	20
SRR3113782	29	25	0	42	19
SRR3372017	47	42	42	0	37
SRR3545396	24	20	19	37	0

Figure 52. Resulting SNP Distance Matrix