Jun 21, 2024 Version 2

Salmonella serotype prediction using the GalaxyTrakr SeqSero2 workflow V.2

DOI
https://dx.doi.org/10.17504/protocols.io.4r3l24kypg1y/v2

Paul Morin[1], Ruth Timme[1], Michelle Moore[2], Shauna Madson[1], Evelyn Ladines[1], Julia Manetas[1], Karen Jinneman[1]
[1]US Food and Drug Administration; [2]FDA

GenomeTrakr
Tech. support email: genomeTrakr@fda.hhs.gov

Ruth Timme
US Food and Drug Administration

DOI: https://dx.doi.org/10.17504/protocols.io.4r3l24kypg1y/v2

Protocol Citation: Paul Morin, Ruth Timme, Michelle Moore, Shauna Madson, Evelyn Ladines, Julia Manetas, Karen Jinneman 2024. Salmonella serotype prediction using the GalaxyTrakr SeqSero2 workflow. **protocols.io** https://dx.doi.org/10.17504/protocols.io.4r3l24kypg1y/v2Version created by **Ruth Timme**

Manuscript citation:
Gangiredla, J., Rand, H., Benisatto, D. et al. GalaxyTrakr: a distributed analysis tool for public health whole genome sequence data accessible to non-bioinformaticians. BMC Genomics 22, 114 (2021).

Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. Salmonella serotype determination utilizing high-throughput genome sequencing data. ASM Journals Vol. 53, No. 5 (2015)

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

Created: April 19, 2024

Last Modified: June 21, 2024

Protocol Integer ID: 98491

Keywords: salmonella, genomic serotyping, seqsero2, Galaxy, salmonella serotype prediction, salmonella serotype, scheme for phenotypic salmonella, phenotypic salmonella, genes responsible for serotype antigen, serotype antigen, salmonella serological somatic, molecular methods for serotype determination, serotype determination, salmonella, seqsero2, serotype prediction, traditional serotype determination, rfb gene cluster, antigenic formula, antigen, whole genome sequence, gene, isolates in the galaxytrakr environment, culture isolate, bioinformatic pipeline

Disclaimer

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.

Abstract

Salmonella serotypes are defined by two surface structures, O antigen and two H antigens. Traditional serotype determination is performed with the *Salmonella* serological somatic (O) and flagellar (H) tests and paired with biochemical confirmation. More than 2,600 *Salmonella* serotypes have been described in the White-Kauffmann-Le Minor scheme. Molecular methods for serotype determination have been developed based on genes responsible for serotype antigens. These genes are encoded in the *rfb* gene cluster, *fliC*, and *fljB*. SeqSero2 is a bioinformatic pipeline that uses whole genome sequence (WGS) data from pure-culture isolates to perform *in silico* analysis to determine the antigenic formula, including somatic (O) antigens and both flagellar (H) antigens. This provides continuity with the well-established scheme for phenotypic *Salmonella* serotypes.

PURPOSE:
This document outlines the steps required to run SeqSero2 v1.2.1 on a collection of isolates in the GalaxyTrakr environment. This is performed by utilizing a custom workflow called “SeqSero2 v1.2.1 collection workflow” and downloading the resulting table.

SCOPE: This protocol covers the following tasks:
1. Login or set up an account in GalaxyTrakr
2. Create a new history/workspace
3. Upload data
4. Execute the SeqSero2 workflow
5. Download the results

Materials

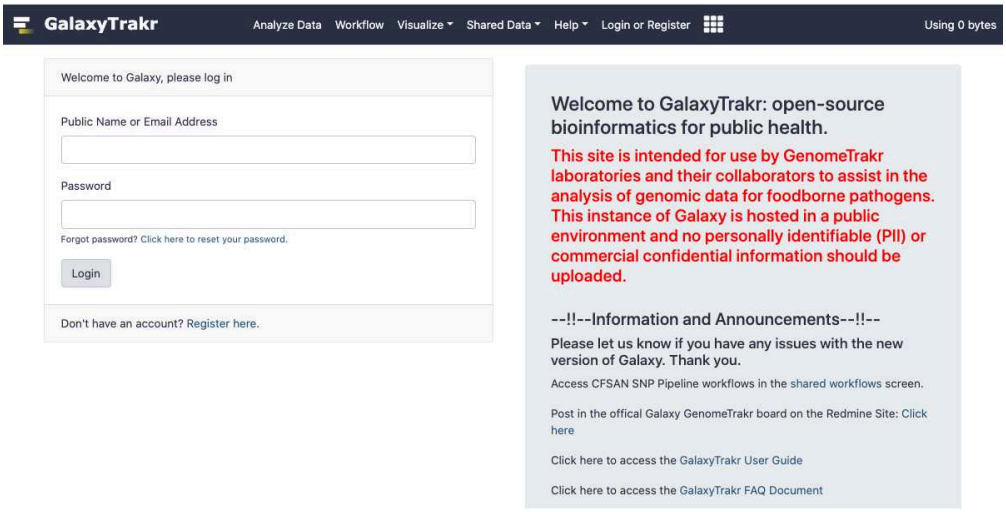
Salmonella WGS fastq files or SRA accessions

Before start

When using GalaxyTrakr, it is recommended to use Google Chrome for optimal browser experience although Microsoft Edge and Safari are also compatible browsers. Internet Explorer and Mozilla FireFox are NOT compatible with GalaxyTrakr.

Login and import workflow

1 Log into GalaxyTrakr (https://galaxytrakr.org/root/login)



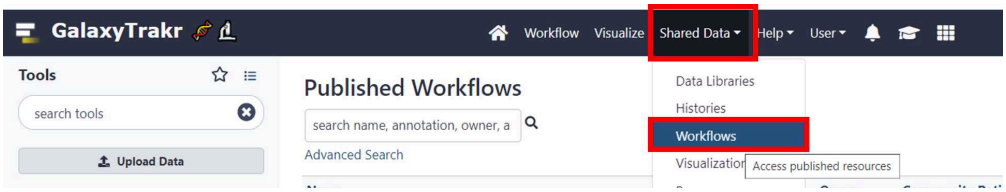
GalaxyTrakr login screen

Link to create a new GalaxyTrakr account:
https://account.galaxytrakr.org/Account/Register

1 **Import** the "SeqSero2 v1.2.1 workflow tabular and row outputs" by cstrittmatter, May30, 2023 into the Tools Panel

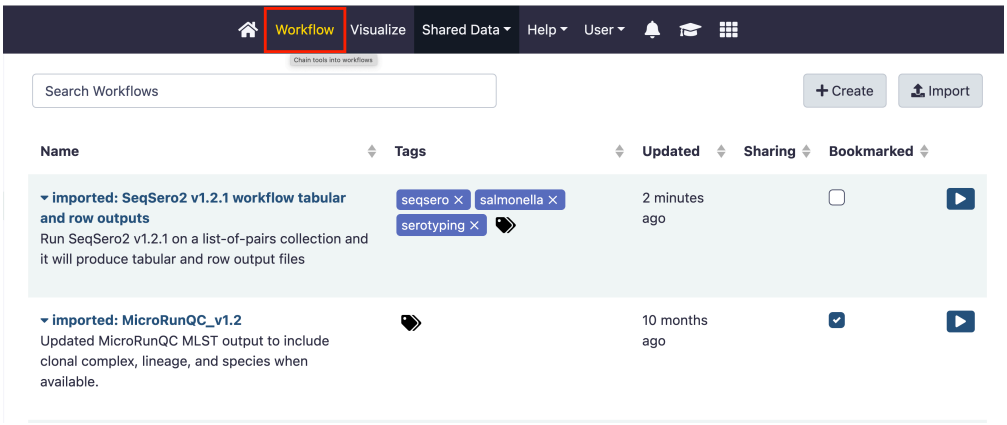
> Note
>
> Step 2 only needs to be done once. After this workflow is imported it will be available for use in your Tools Panel.

1.1 Click on **Shared Data** and then **Workflows** from the dropdown menu.

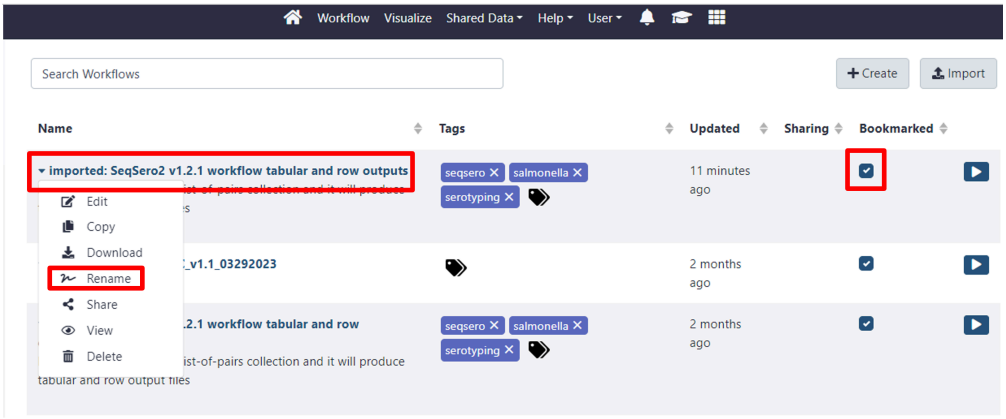


1.2 Search for "seqsero" and locate the shared workflow: "**SeqSero2 v1.2.1 workflow tabular and row outputs",** then select **Import** from the dropdown arrow.

1.3 Click on the **Workflow** tab to rename this workflow and make it visible in your tools panel.



**Adding a date to the name will help you in keeping track of newer versions of this workflow. Workflows do get updated periodically and you want to ensure you are working with the most recent version.



Check the box **“Bookmarked”**.

This will move the workflow into your tools panel permanently and you will now have this workflow easily available to you.

Step 2 only needs to be done once for each workflow that is being imported into your Tools.

Import data for analysis

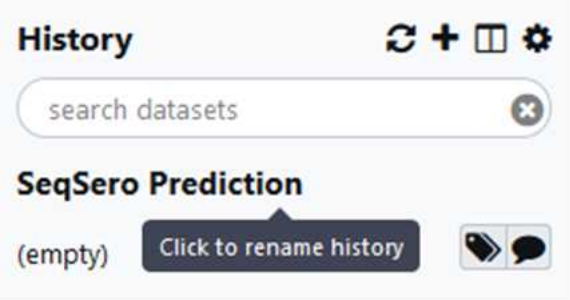
2 If your data is already in GalaxyTrakr, open the history containing that data to be analyzed or move the data to a new history for analysis and proceed to **Step # 5**. This option may be preferred if the data was already uploaded for other purposes such as MicroRunQC. It’s ok if there are non-*Salmonella* isolates in your dataset.They will not return an antigenic formula or serovar name.

For uploading new data proceed to next step to create a new history and upload your data to be analyzed.

2.1 **Create new History:**

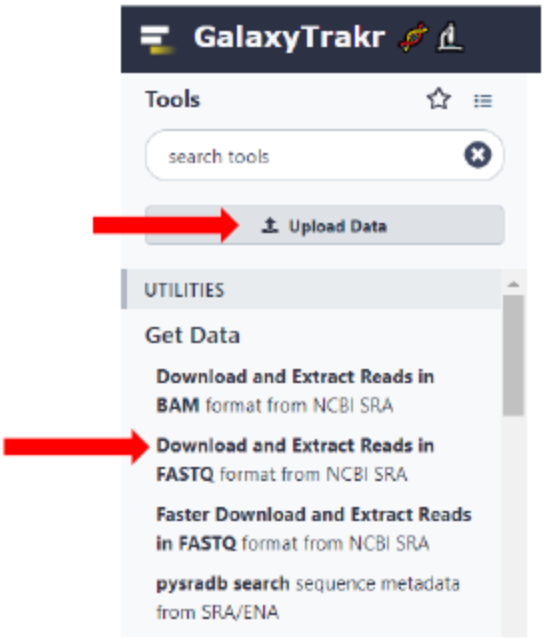
Click on the “+” button in the upper right corner.

Type in a custom name (i.e., “SeqSero Prediction”)



2.2 **Import data:**

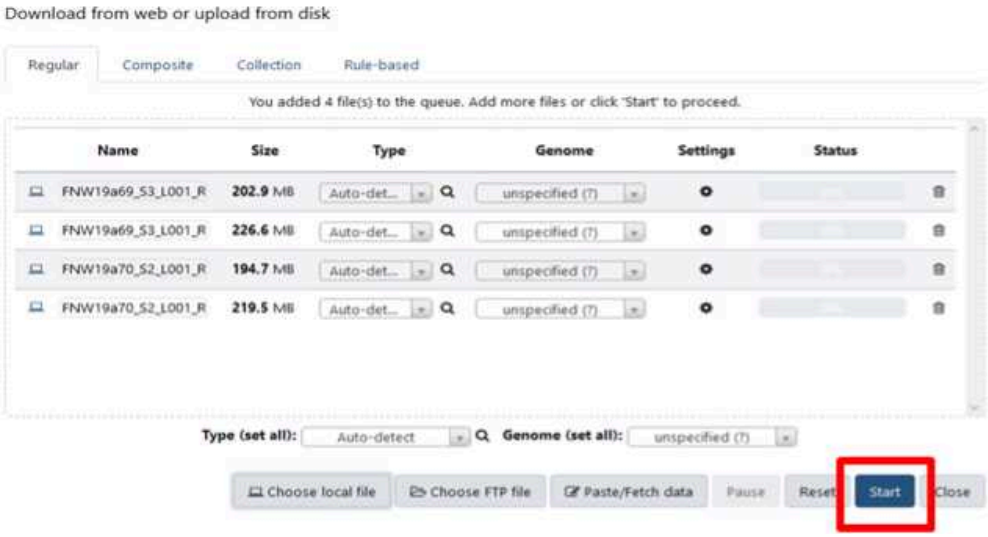
"Upload Data", step 3.3 or "Get Data > "**Download and Extract Reads in FASTQ** format from NCBI SRA" step 3.4.



Next steps will show how to upload data or import data from NCBI.

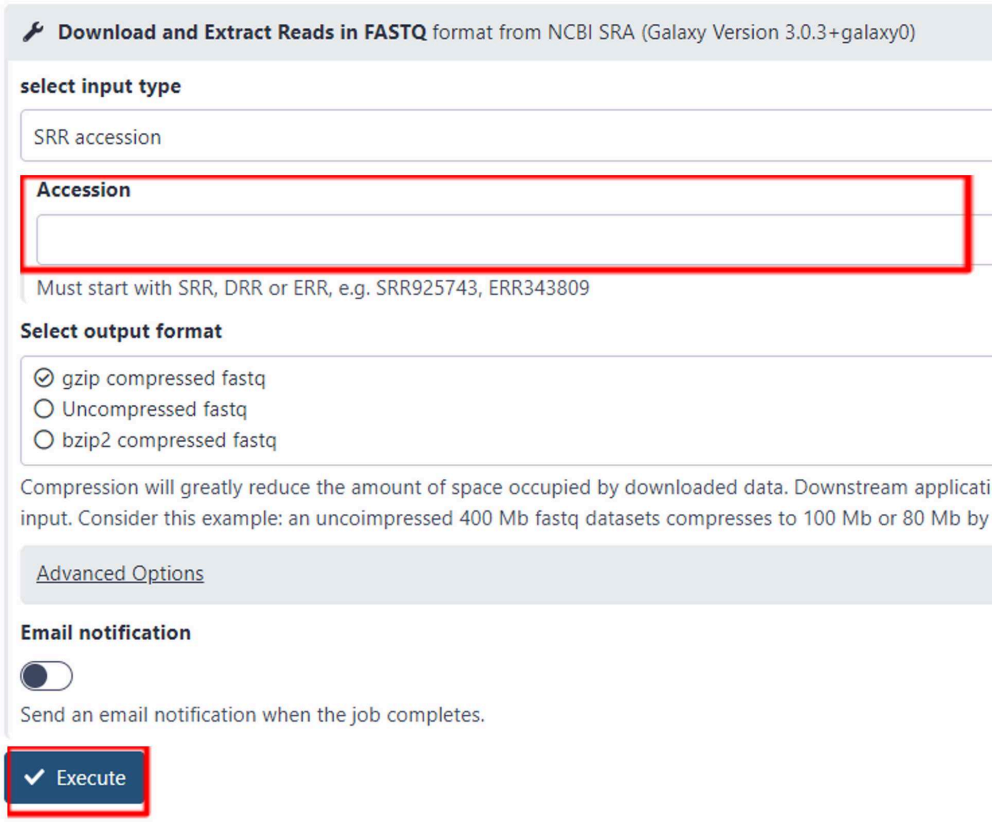
2.3 “Upload File” for .gz files stored locally.

1. Click on “Choose local files”
2. Find your WGS fastq.gz files and select those (2 data files: Read 1 and Read 2 per organism).
3. Click “Start” The amount of time to upload depends on how many files have been selected and the size of those files. The status bar will start to fill as upload progress is made and turn green when completed.



2.4 “**Download and Extract Reads in FASTQ** format from NCBI SRA” to import data from NCBI.

1. Enter the NCBI SRR for each sequence to be retrieved.
2. Click “Execute”

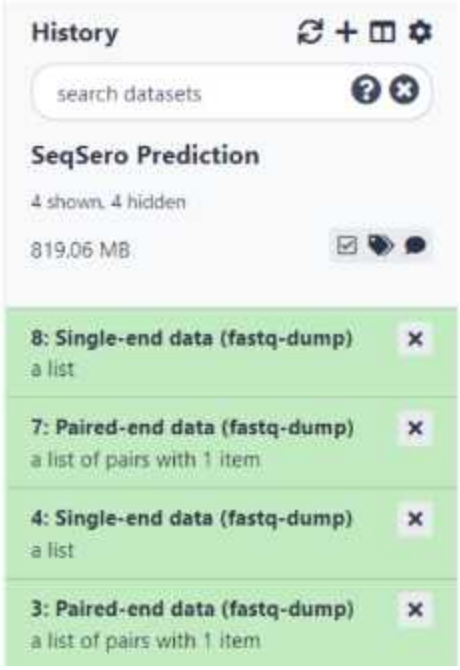


2.5 When the data has finished importing, you should see the successfully uploaded files listed in green in the right panel.

Files will be highlighted in RED if they were NOT successfully uploaded.

Example of .gz files uploaded:

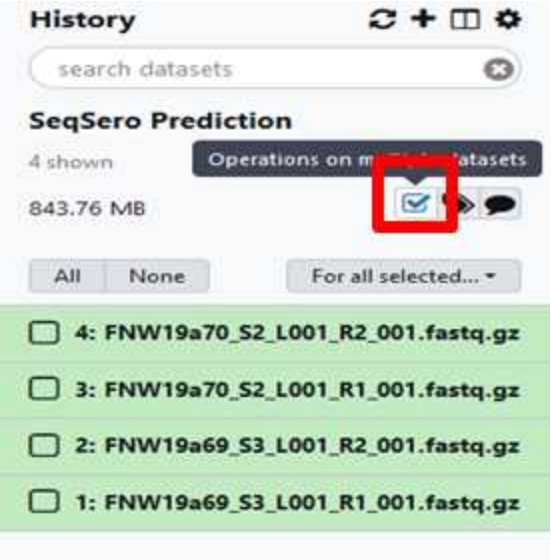
Example of SRR data downloaded from NCBI:



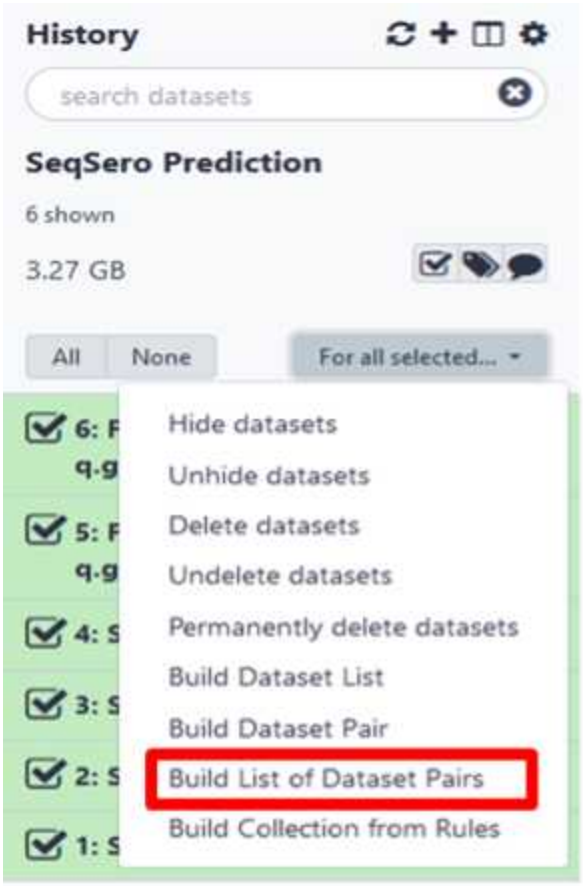
Build your dataset of paired-reads

3 For uploaded local .gz files **Build "list of data set pairs" .** following steps 4.1 through 4.6. For SRR data downloaded from NCBI merge data set collections, following step 4.7.

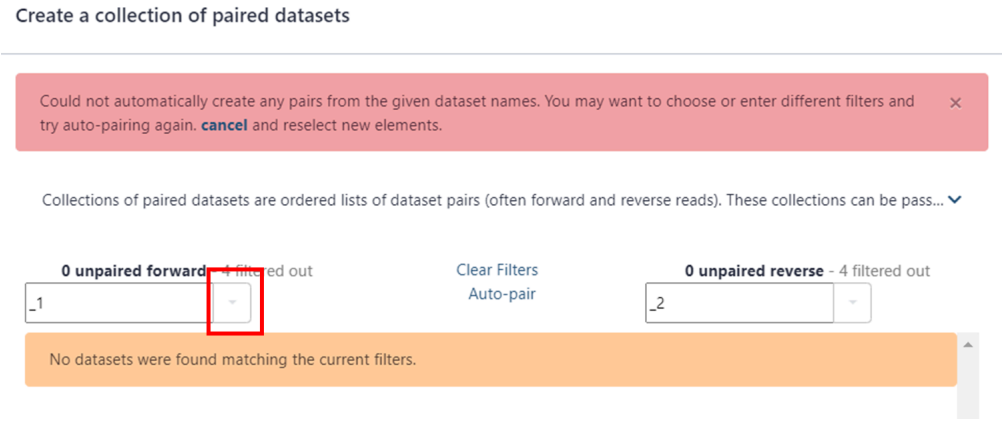
3.1 Click on the check mark in the history panel then select all files you want to include in the data set for SeqSero analysis.



3.2 Open options under "For all selected" and then choose "Build List of Dataset Pairs"



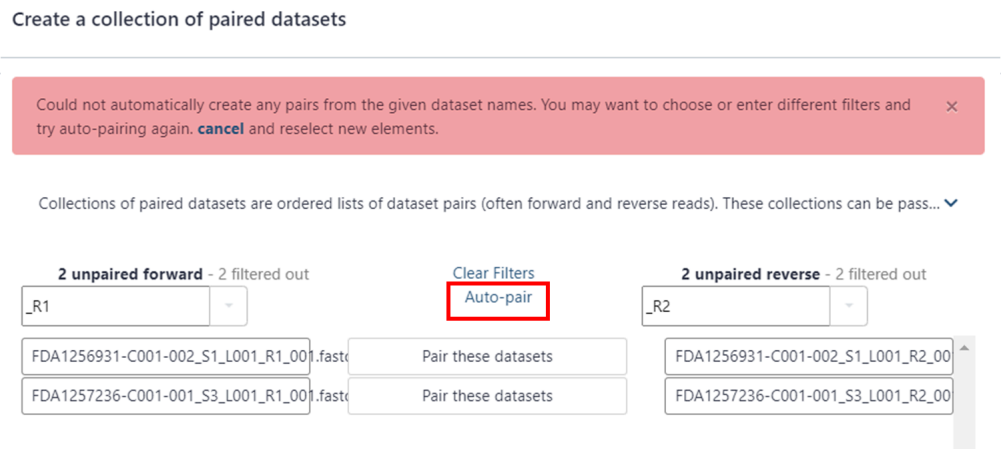
3.3 Click dropdown arrow.



3.4 Click the correct file extension, e.g. "**_R1"**

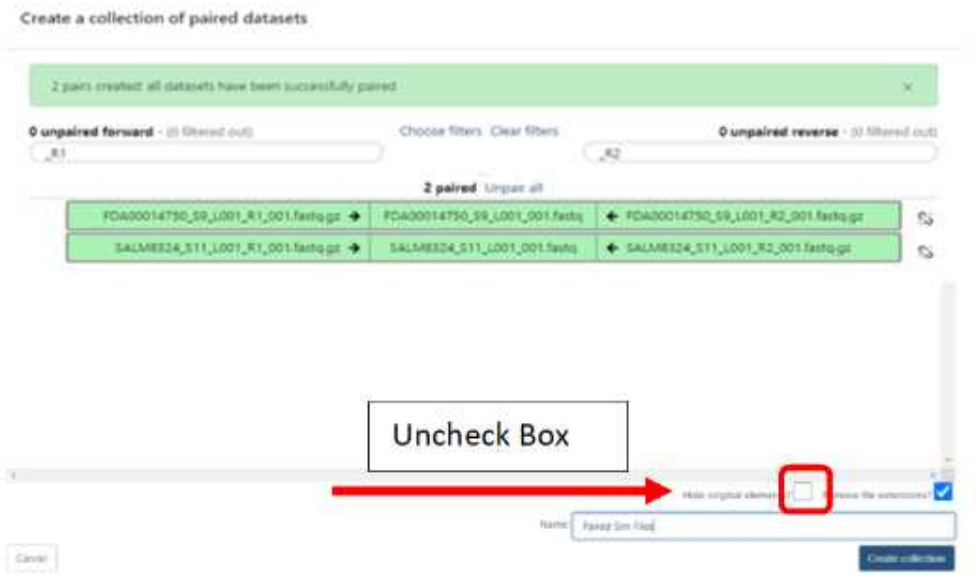
3.5 Click “**Auto-pair**”

The Read 1 and Read 2 fastq.gz files should automatically pair together.



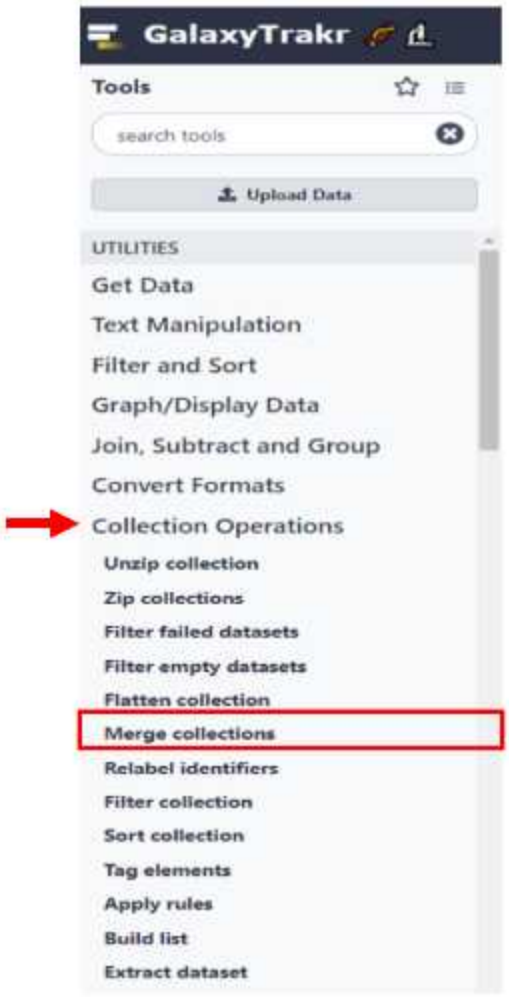
3.6 Uncheck "Hide original elements?" and type in a custom name for the dataset (i.e., “Paired Slm Files”)

Click “**Create list”**

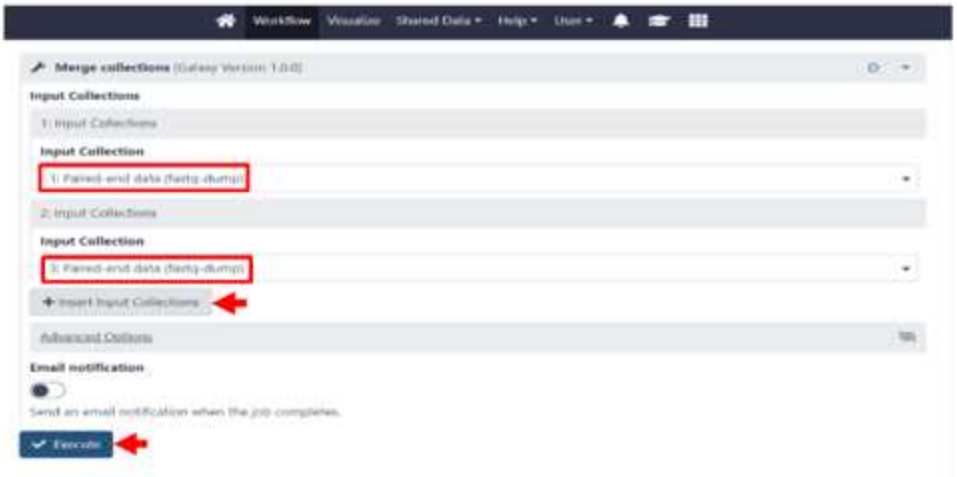


You should see your named list in the history panel. Continue with step #5.
Note: You may also use the same dataset list from other workflows such as MicroRunQC. It's ok if there are non-Salmonella isolates in your dataset. They will not return an antigenic formula or serovar name.

3.7 For SRR data that was downloaded from NCBI and uploaded to GalaxyTrakr as paired end data merge the datasets into a list of dataset pairs.
1. Navigate on tools panel to Collection Operations and open options
2. Select Merge collections



3. Select input collections (paired-end data (fastq-dump) files to be merged. Additional collections can be specified with the “+ Insert Input Collections” button. Then click “Execute”



4. The resulting data file ending with “(merged) list of pairs” in your history panel can be used in the SeqSero v1.2.1 workflow. Coantinue with step 5.

Analyze your data using the SeqSero2 workflow

4 In NGS TOOLBOX, left panel:

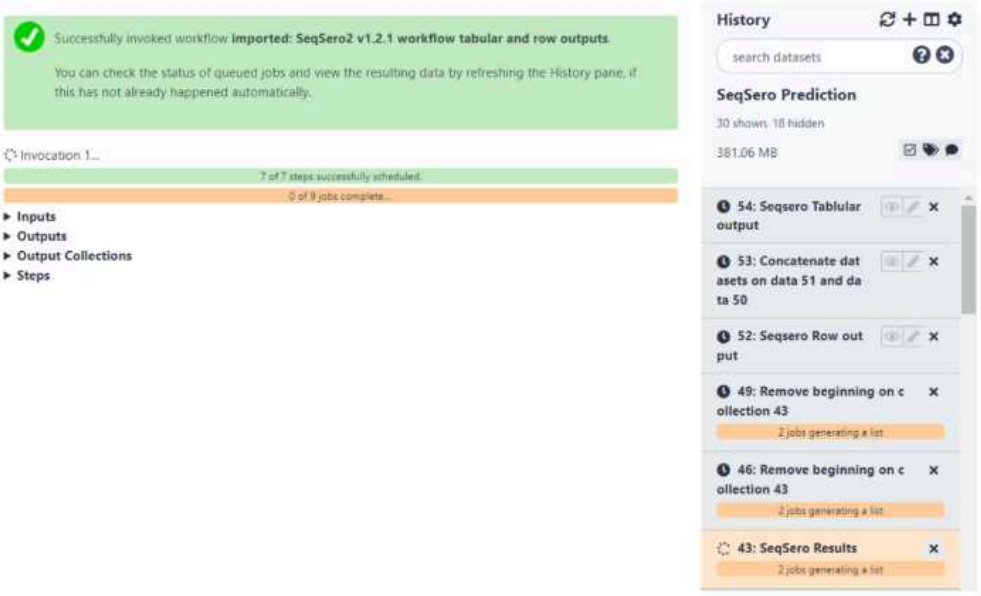
Click on the imported and saved version of the “**SeqSero2 v1.2.1 workflow tabular and row outputs**”.

4.1 In the Main window, the newly created list of paired files should automatically show up in the “Input dataset collection” window.

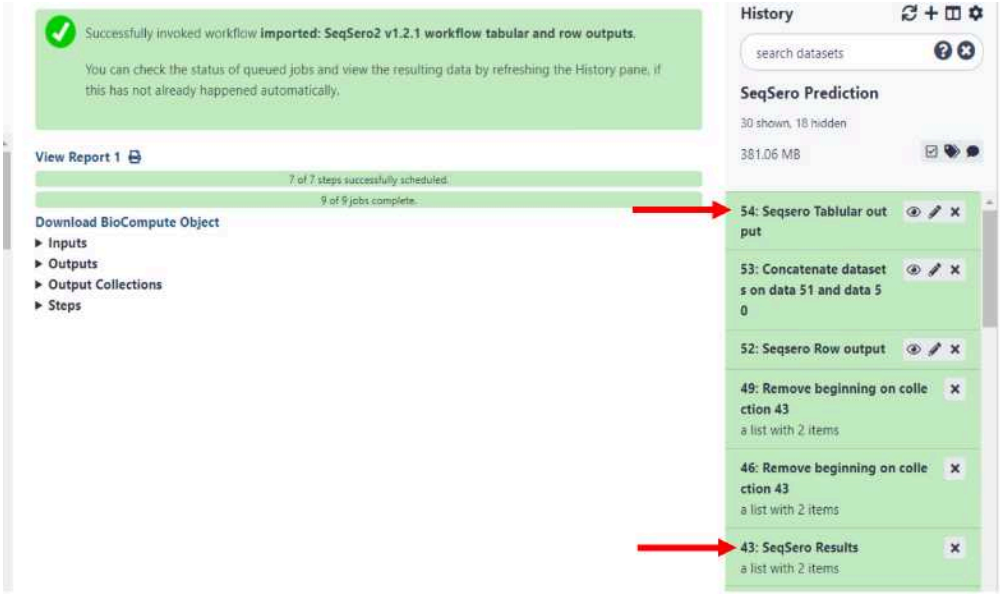
If it doesn’t, click and drag the file from your history panel into the “Input dataset collection” window.

Click “**Run Workflow”**

4.2 Your working panel should appear green with a white check mark on the upper left-hand corner.

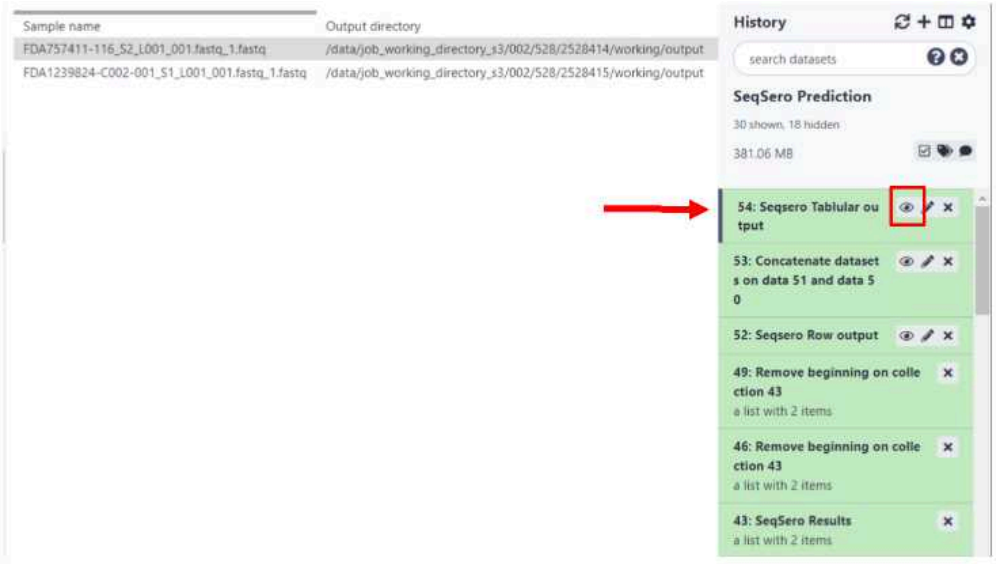


4.3 After the SeqSero analysis is complete, the “Seqsero Results” will appear green as well as the “Seqsero Tabular output”. “Seqsero Results” provide serotyping results for each individual strain while “Seqsero Tabular output” provides a table of all the paired datasets.



View and export results

5 Click on the “eyeball” in the “**Seqsero Tabular output”** to view the tabular results.



Note: Scroll across the table to see additional information.

5.1 **Export SeqSero results:** cut/paste method

1	2	3	4	5	6	7	8	9
SRR13234097_1.f	/data/job_	/data/job_	60	r	e,n,x,z15	IIIb	60:r:e,n,x,	IIIb 60:r:e,n,x,z15
SRR13234098_1.f	/data/job_	/data/job_	9,46	z29	-	I	9,46:z29:-	Ouakam
FNW19B79_S2_L	/data/job_	/data/job_	4	k	e,n,z15	I	4:k:e,n,z1!	Texas

1. Click and drag to highlight text
2. Copy
3. Paste Special as “Text” or “Unicode Text” into Excel

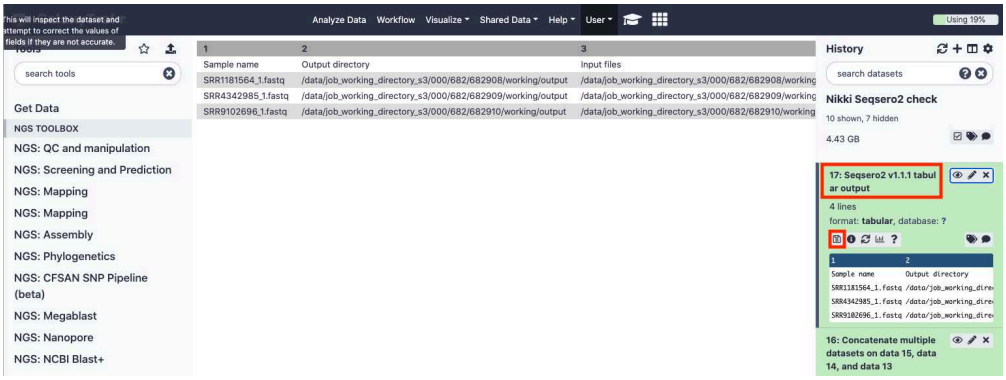
Alternatively, click on the table and “Ctrl-A” to select the entire Table, “Ctrl-C” to copy data and paste the copied data into Excel by “Ctrl-V”

5.2 **Export SeqSero2 results**: download tab-delimited text file

Click the dataset name.

The panel will expand, enabling more options.

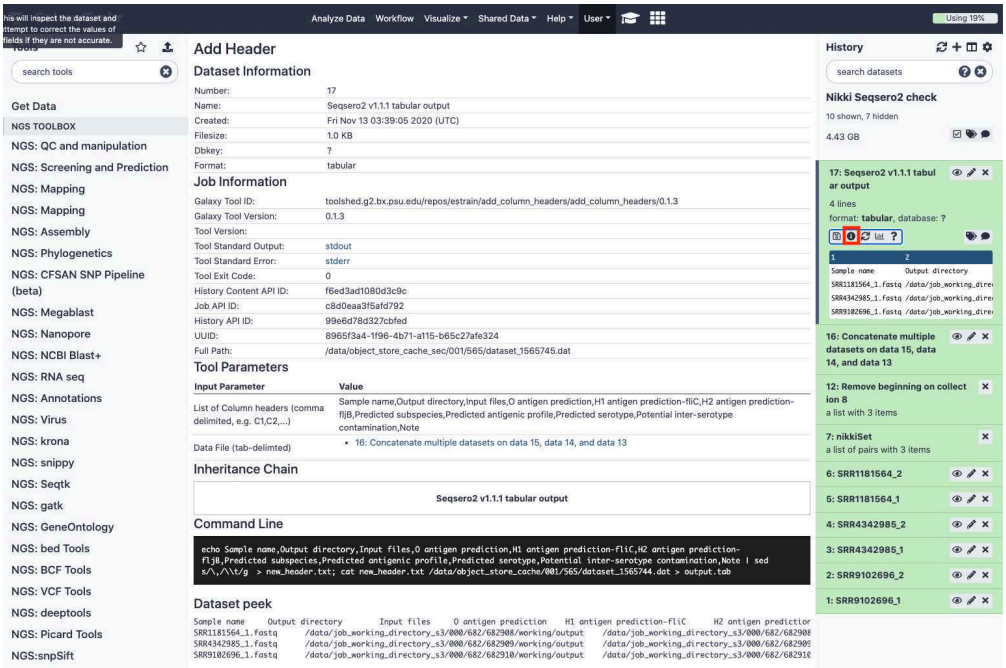
Click the "**Save**" icon to download a tab-delimited file of results.



Example results file: SeqSeroExampleResults.tabular

5.3 **Optional:**

The small "Info" icon results in a detailed view of the dataset, analysis, parameters used, etc., which can be helpful for troubleshooting.



Protocol references

Gangiredla, J., Rand, H., Benisatto, D. et al. GalaxyTrakr: a distributed analysis tool for public health whole genome sequence data accessible to non-bioinformaticians. BMC Genomics 22, 114 (2021).

Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. Salmonella serotype determination utilizing high-throughput genome sequencing data. ASM Journals Vol. 53, No. 5 (2015)