



Nov 05, 2021

# Salmonella serotype prediction using the GalaxyTrakr SeqSero2 workflow

Paul Morin<sup>1</sup>, Ruth Timme<sup>1</sup>, Michelle Moore<sup>1</sup>, Shauna Madson<sup>1</sup>, Evelyn Ladines<sup>1</sup>, Julia Manetas<sup>1</sup>, Karen Jinneman<sup>1</sup>

<sup>1</sup>US Food and Drug Administration

1



[dx.doi.org/10.17504/protocols.io.bybfpsjn](https://dx.doi.org/10.17504/protocols.io.bybfpsjn)

## GenomeTrakr

Tech. support email: [genomeTrakr@fda.hhs.gov](mailto:genomeTrakr@fda.hhs.gov)

Ruth Timme

US Food and Drug Administration

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

*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 *fliB*. 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 v1.1.1 on a collection of isolates in the GalaxyTrakr environment. This is performed by utilizing a custom workflow called "SeqSero2 v1.1.1 collection workflow" and downloading the resulting table.

SCOPE: This protocol covers the following tasks:

1. set up an account in GalaxyTrakr
2. Create a new history/workspace
3. Upload data
4. Execute the SeqSero2 workflow
5. Download the results

DOI

[dx.doi.org/10.17504/protocols.io.bybfpsjn](https://dx.doi.org/10.17504/protocols.io.bybfpsjn)

Paul Morin, Ruth Timme, Michelle Moore, Shauna Madson, Evelyn Ladines, Julia Manetas, Karen Jinneman 2021. Salmonella serotype prediction using the GalaxyTrakr SeqSero2 workflow. **protocols.io**  
<https://dx.doi.org/10.17504/protocols.io.bybfpsjn>



protocol

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).

salmonella, genomic serotyping, seqsero2, Galaxy

protocol ,

Sep 16, 2021

Nov 10, 2021

53319

*Salmonella* WGS fastq files or SRA accessions

:

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

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

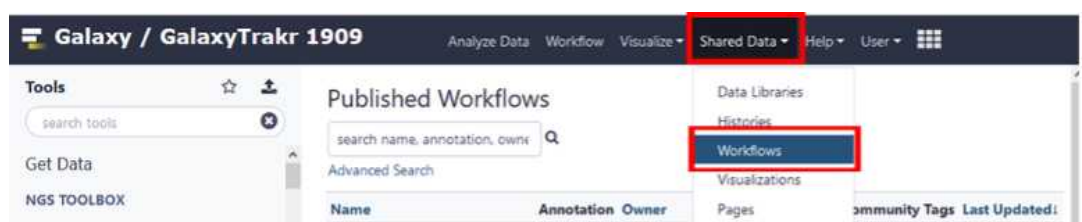
- 1 Log into GalaxyTrakr 1909 (<https://galaxytrakr.org/root/login>)

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

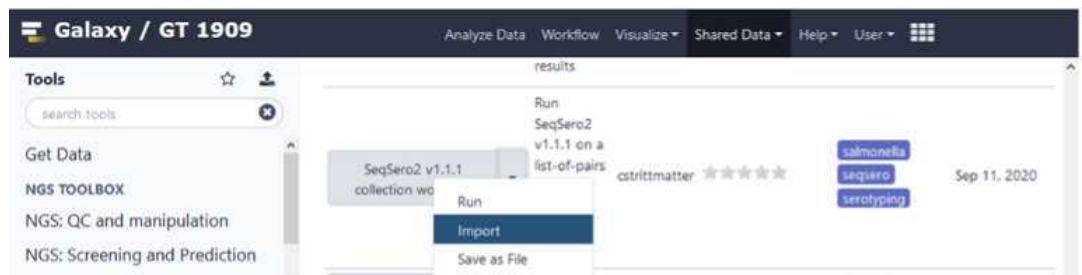
## 1 Import the SeqSero2 workflow into the Tools Panel

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 Select the shared workflow: **SeqSero2 v1.1.1 collection workflow (cstrittmatter 9/11/2020)** and select **import** from the dropdown arrow.

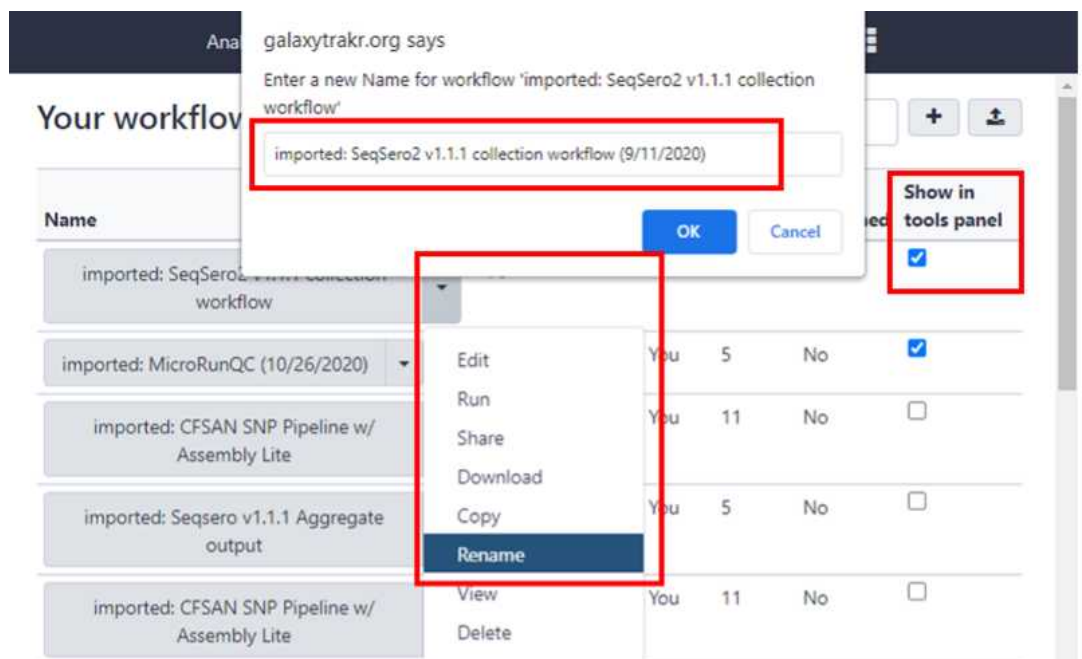


### 1.3 In **Your Workflows** you can use the dropdown arrow to rename of this workflow.

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 **"Show in tools panel"**.

This will move the Seqsero2 v1.1.1 collection workflow into your tools panel permanently and you will now have this workflow available to you every time you log into GalaxyTrakr.



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 #**. This option may be preferred if the data was already uploaded for other purposes such as MicroRunQC. It's ok if there are

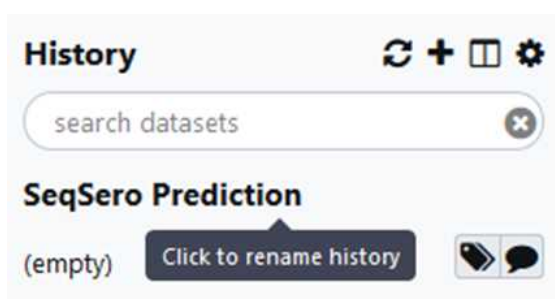
non-*Salmonella* isolates in your dataset. They will not return an antigenic formula or serovar name.

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.

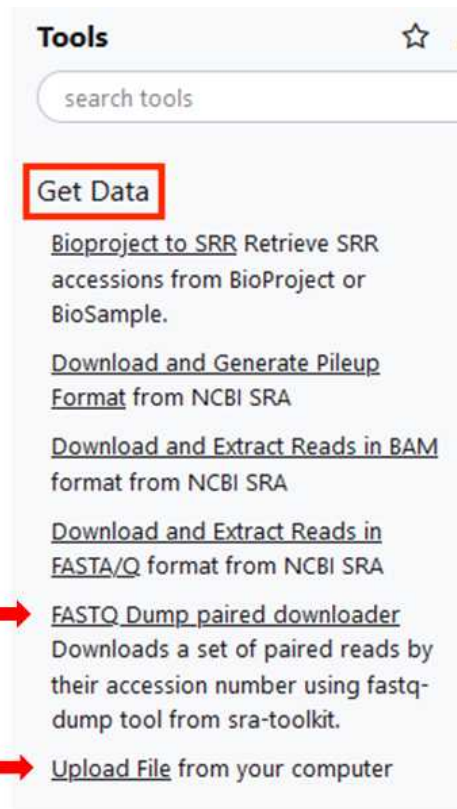
Type in a custom name (i.e., “SeqSero Prediction”)



## 2.2 Import data:

Click on the Tool “**Get Data**”, top of left panel.

Click on the sub-tool of choice to bring in data for analysis.



Next steps will show how to import data from NCBI OR upload from your computer.

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

1. Click on “Choose 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.

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 4 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
FNW19a69_S3_L001_R	202.9 MB	Auto-det...	unspecified (?)		
FNW19a69_S3_L001_R	226.6 MB	Auto-det...	unspecified (?)		
FNW19a70_S2_L001_R	194.7 MB	Auto-det...	unspecified (?)		
FNW19a70_S2_L001_R	219.5 MB	Auto-det...	unspecified (?)		

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

Choose local file Choose FTP file Paste/Fetch data Pause Reset **Start** Close

## 2.4 "FASTQ Dump Paired downloader" to import data from NCBI.

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

**FASTQ Dump paired downloader** Downloads a set of paired reads by their accession number using fastq-dump tool from sra-toolkit. (Galaxy Version 1.1.4) ☆ Favorite ▼ Options

**Accession Number**

✓ Execute

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

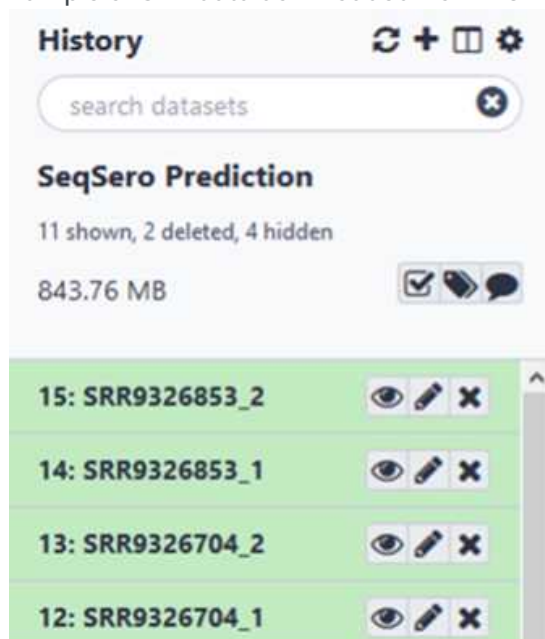
Files will be highlighted in RED if they were NOT successfully uploaded.

Example of .gz files uploaded:





Example of SRR data downloaded from NCBI:

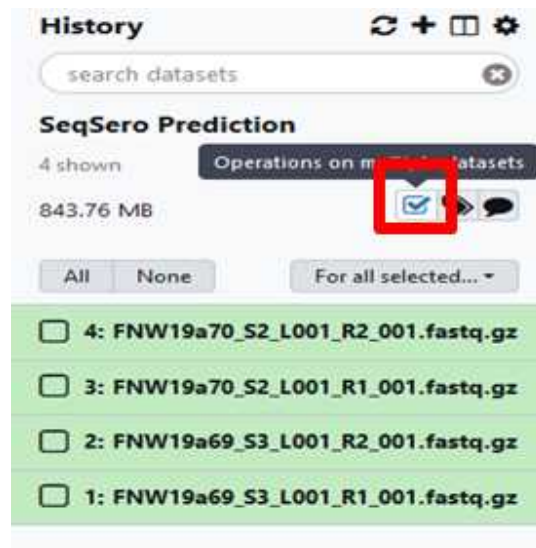


Build your dataset of paired-reads

### 3 Build “list of data set pairs”



- 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 “Choose Filters”

Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names. You may want to choose or enter different filters and try auto-pairing again. Close this message using the X on the right to view more help.

0 unpaired forward - (2 filtered out) **Choose filters** Clear filters 0 unpaired reverse - (2 filtered out)

\_1 Auto-pair \_2

(no datasets were found matching the current filters)

### 3.4 Click the correct file extension “Forward: \_R1, Reverse: \_R2”

Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names. You may want to choose or enter different filters and try auto-pairing again. Close this message using the X on the right to view more help.

0 unpaired forward - (4 filtered out) Choose filters Clear filters 0 unpaired reverse - (4 filtered out)

\_1 Choose from the following filters to change which unpaired reads are shown in the display:

Forward: \_1, Reverse: \_2

**Forward: \_R1, Reverse: \_R2**

### 3.5 Click “Auto-pair”

The Read 1 and Read 2 fastq.gz files should automatically pair together. Note: For data downloaded from NCBI the two reads will already be paired and you will not need to select filter and auto-pair.

Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names. You may want to choose or enter different filters and try auto-pairing again. Close this message using the X on the right to view more help.

1 unpaired forward - (1 filtered out) Choose filters Clear filters 1 unpaired reverse - (1 filtered out)

\_R1 **Auto-pair** \_R2

FNW19879\_S2\_L001\_R1\_001.fastq.gz Pair these datasets FNW19879\_S2\_L001\_R2\_001.fastq.gz

### 3.6 Type in a custom name for the dataset (i.e., “Paired SIm Files”)

Click **“Create list”**

Create a collection of paired datasets

2 pairs created: all datasets have been successfully paired

0 unpaired forward - (0 filtered out) Choose filters Clear filters 0 unpaired reverse - (0 filtered out)

\_R1 \_R2

2 paired Unpair all

FDA00014750_S9_L001_R1_001.fastq.gz →	FDA00014750_S9_L001_001.fastq	← FDA00014750_S9_L001_R2_001.fastq.gz	🗑️
SALM8324_S11_L001_R1_001.fastq.gz →	SALM8324_S11_L001_001.fastq	← SALM8324_S11_L001_R2_001.fastq.gz	🗑️

Remove file extensions from pair names? ☒ Hide original elements? ☐

Name: Paired SRA files

Cancel Create list

You should see your named list in the history panel

#### Analyze your data using the SeqSero2 workflow

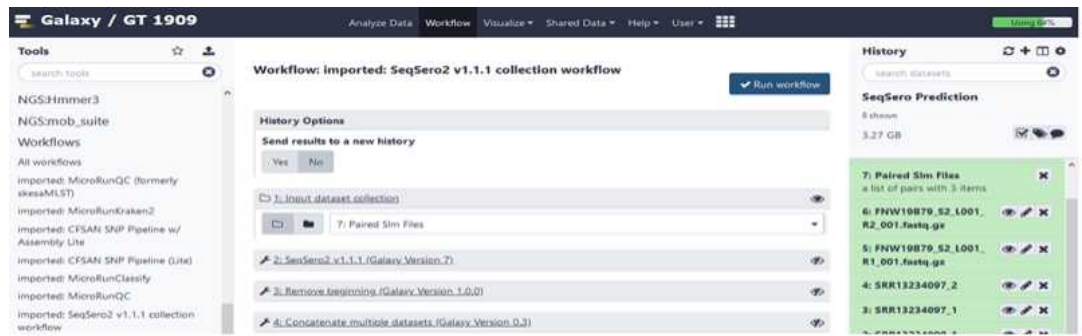
4 In NGS TOOLBOX, left panel:

Click on the imported and saved version of the **“SeqSero2 v1.1.1 collection workflow”**.

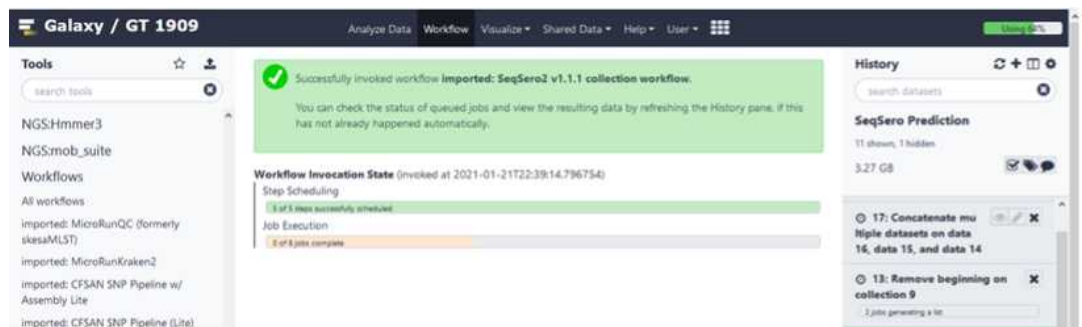
4.1 In the Main window, the newly created list of paired files should automatically show up in the “Input dataset collection” window.

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

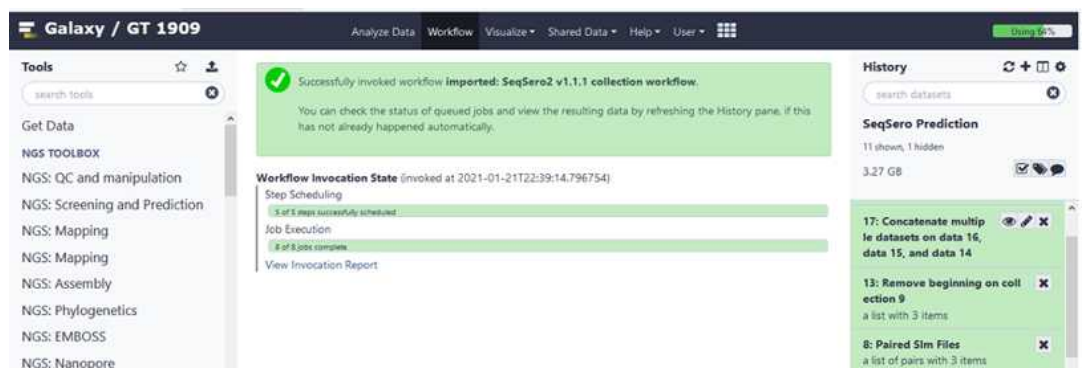
Click **“Run Workflow”**



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

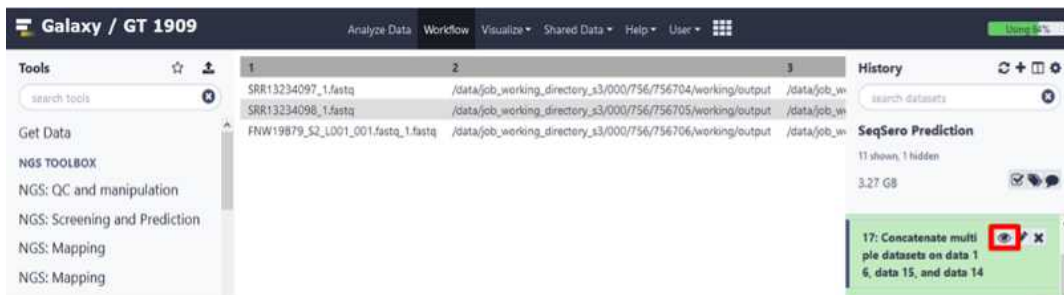


4.3 After the SeqSero analysis is complete, the “Concatenate multiple datasets on data” will appear green.



## View and export results

5 Click on the “eyeball” in the “Concatenate multiple datasets on data” to view table of predicted serotypes for the collection.



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	/data/job	/data/job	60	r	e,n,x,z15	IIlb	60:r:e,n,x, IIlb	60:r:e,n,x,z15
SRR13234098_1	/data/job	/data/job	9,46	z29	-	I	9,46:z29:-	Ouakam
FNW19879_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

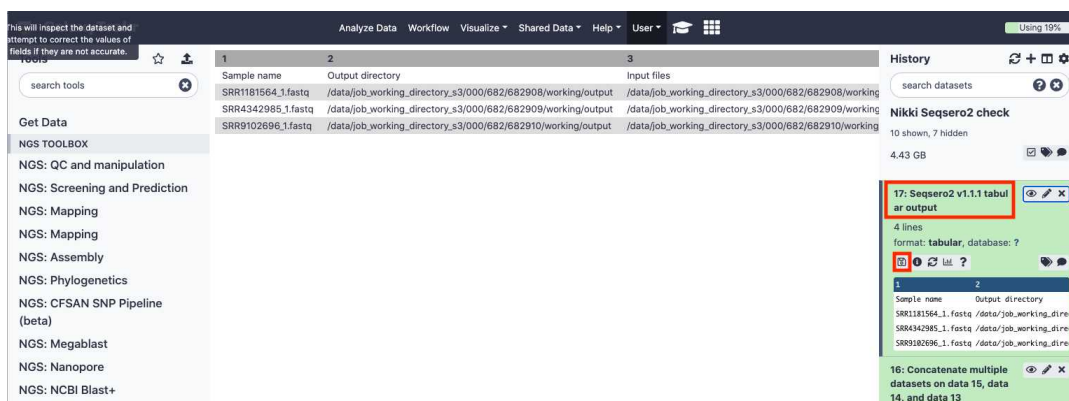
Alternatively, click on the table and "Ctrl-A" to select the entire Table, "Ctrl-C" to copy data and paste the copied data into Excel by "Ctrl-V"

## 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.



## 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.

You will inspect the dataset and attempt to correct the values of fields if they are not accurate.
Analyze Data   Workflow   Visualize »   Shared Data   Help »
Using 19%

**NGS TOOLBOX**

- Get Data
- NGS: QC and manipulation
- NGS: Screening and Prediction
- NGS: Mapping
- NGS: Mapping
- NGS: Assembly
- NGS: Phylogenetics
- NGS: CFSAN SNP Pipeline (beta)
- NGS: Megablast
- NGS: Nanopore
- NGS: NCBI Blast+
- NGS: RNA seq
- NGS: Annotations
- NGS: Virus
- NGS: krona
- NGS: snippy
- NGS: Seqtk
- NGS: gatk
- NGS: GeneOntology
- NGS: bed Tools
- NGS: BCF Tools
- NGS: VCF Tools
- NGS: deepTools
- NGS: Picard Tools
- NGS: snpSift

## Add Header

### Dataset Information

Name:	17
Name:	Seqzero2 v1.1.1 tabular output
Created:	Fri Nov 13 03:39:05 2020 (UTC)
Filesize:	1.0 KB
Dkeyy:	?
Format:	tabular

### Job Information

Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/estran/add_column_headers/add_column_headers/0.1.3
Galaxy Tool Version:	0.1.3
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
History Content API ID:	f6ed3ad108d03dc9c
Job API ID:	c8d0eaa3f5afd792
History API ID:	99e6d78d327cbfd
UUID:	8b65f3a4-1f96-4b71-a115-b65c27afe324
Full Path:	/data/object_store_cache_sec/001/656/dataset_1565744.dat

Input Parameter	Value
List of Column headers (comma delimited, e.g. C1,C2,...)	Sample name,Output directory,input files,O antigen prediction,H1 antigen prediction-fltC,H2 antigen prediction-fltB,Predicted subspecies,Predicted antigenic profile,Predicted serotype,Potential inter-serotype contamination>Note • 16: Concatenate multiple datasets on data 15, data 14, and data 13

Data File (tab-delimited)

### Inheritance Chain

```
Seqzero2 v1.1.1 tabular output
```

### Command Line

```
echo Sample name,Output directory,input files,O antigen prediction,H1 antigen prediction-fltC,H2 antigen prediction-fltB,Predicted subspecies,Predicted antigenic profile,Predicted serotype,Potential inter-serotype contamination>Note | sed \n/,H1/L/g > new_header.txt; cat new_header.txt /data/object_store_cache/001/656/dataset_1565744.dat > output.tab
```

### Dataset peek

Sample name	Output directory	Input files	O antigen prediction	H1 antigen prediction-fltC	H2 antigen prediction-fltB
SRR1181564_1.fastq	/data/job_working_directory.s3/000/682/682980/working/output	/data/job_working_directory.s3/000/682/682980/working/output			
SRR4342985_1.fastq	/data/job_working_directory.s3/000/682/682980/working/output	/data/job_working_directory.s3/000/682/682980/working/output			
SRR9102696_1.fastq	/data/job_working_directory.s3/000/682/682918/working/output	/data/job_working_directory.s3/000/682/682918/working/output			