

Jun 21, 2024 Version 2

Salmonella serotype prediction using the GalaxyTrakr SeqSero2 workflow V.2

DOI

dx.doi.org/10.17504/protocols.io.4r3l24kypg1y/v2



Paul Morin¹, Ruth Timme¹, Michelle Moore², Shauna Madson¹, Evelyn Ladines¹, Julia Manetas¹, Karen Jinneman¹

¹US Food and Drug Administration; ²FDA

GenomeTrakr

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



Ruth Timme

US Food and Drug Administration

OPEN  ACCESS



DOI: 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/v2> Version 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)

License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

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

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

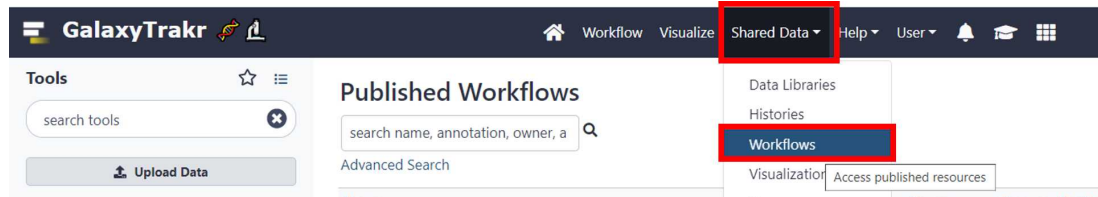
Link to create a new GalaxyTrakr account: <https://account.galaxytrakr.org/Account/Register>

- 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.

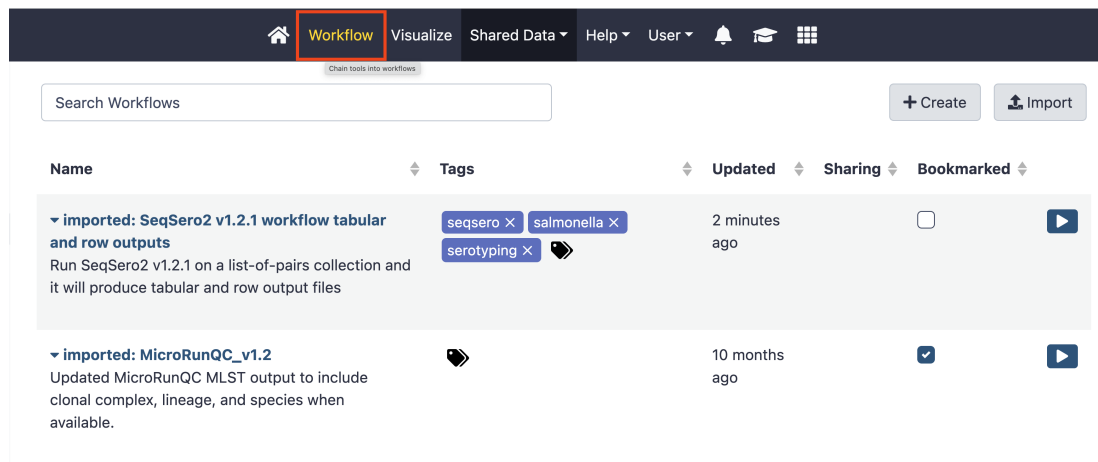
- 1.1 Click on **Shared Data** and then **Workflows** from the dropdown menu.



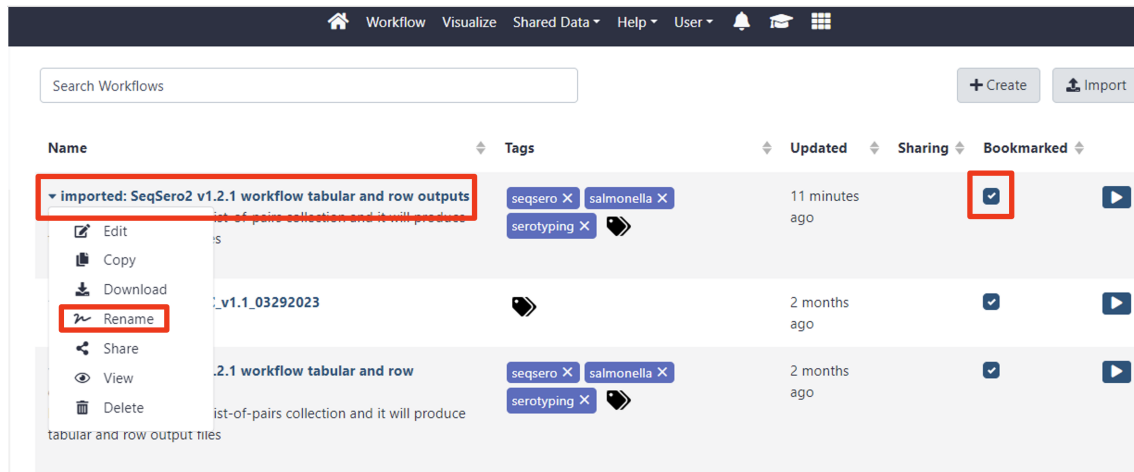
- 1.2 Search for "seqsero" and locate the shared workflow: **"SeqSero2 v1.2.1 workflow tabular and row outputs"**, then select **Import** from the dropdown arrow.



- 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

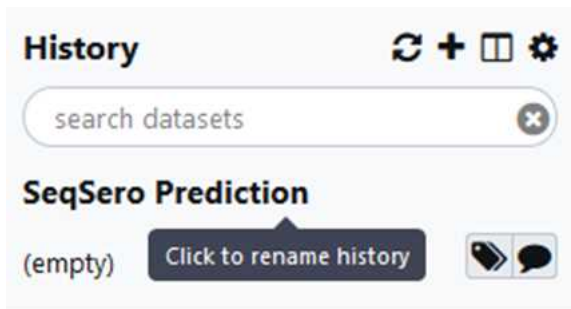
- 2 If your data is already in GalaxyTrakr, open the history containing that data to be analyzed or move the data to a new history for analysis and proceed to **Step # 5**. This option may be preferred if the data was already uploaded for other purposes such as MicroRunQC. It's ok if there are non-*Salmonella* 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:

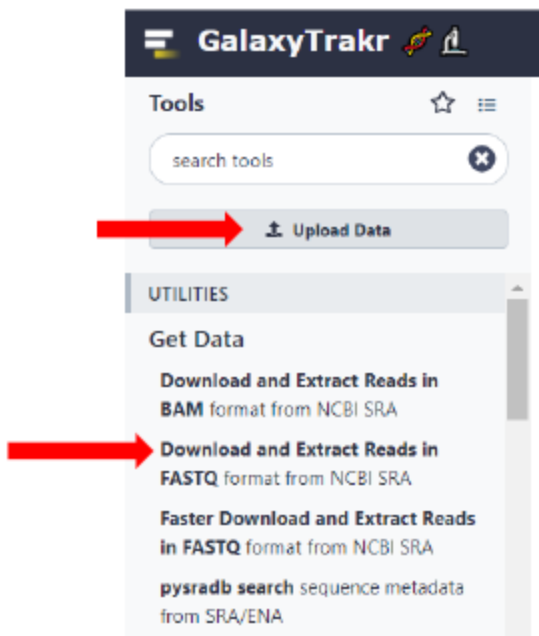
Click on the "+" button in the upper right corner.

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.








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 “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”



 **Download and Extract Reads in FASTQ** format from NCBI SRA (Galaxy Version 3.0.3+galaxy0)

select input type

SRR accession

Accession

Must start with SRR, DRR or ERR, e.g. SRR925743, ERR343809

Select output format

☒ gzip compressed fastq
☐ Uncompressed fastq
☐ bzip2 compressed fastq


Compression will greatly reduce the amount of space occupied by downloaded data. Downstream applicati input. Consider this example: an uncompressed 400 Mb fastq datasets compresses to 100 Mb or 80 Mb by

[Advanced Options](#)

Email notification

☐

Send an email notification when the job completes.

 **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:

History

search datasets

×

SeqSero Prediction

4 shown

843.76 MB

✓

🔒

💬

4: FNW19a70_S2_L001_R2_00

1.fastq.gz

👁

✎

✕

3: FNW19a70_S2_L001_R1_00

1.fastq.gz

👁

✎

✕

2: FNW19a69_S3_L001_R2_00

1.fastq.gz

👁

✎

✕

1: FNW19a69_S3_L001_R1_00

1.fastq.gz

👁

✎

✕

Example of SRR data downloaded from NCBI:

History

search datasets

?

×

SeqSero Prediction

4 shown, 4 hidden

819.06 MB

✓

🔒

💬

8: Single-end data (fastq-dump)

a list

✕

7: Paired-end data (fastq-dump)

a list of pairs with 1 item

✕

4: Single-end data (fastq-dump)

a list

✕

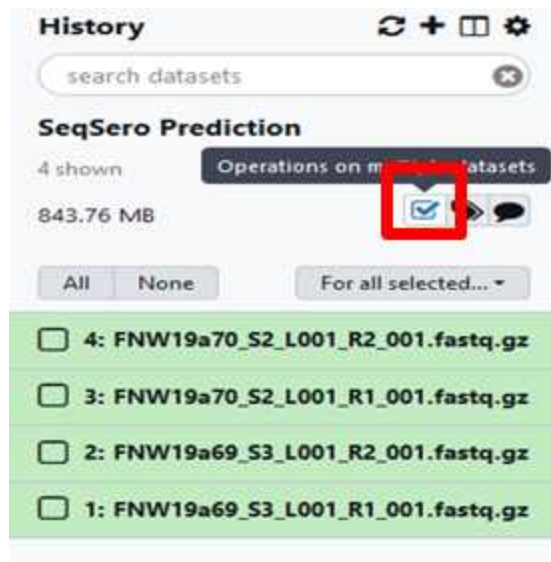
3: Paired-end data (fastq-dump)

a list of pairs with 1 item

✕

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.

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. [cancel](#) and reselect new elements. ×

Collections of paired datasets are ordered lists of dataset pairs (often forward and reverse reads). These collections can be pass... ▼

0 unpaired forward - 4 filtered out

▼

Clear Filters
Auto-pair

0 unpaired reverse - 4 filtered out

▼

No datasets were found matching the current filters.



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.

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. **cancel** and reselect new elements. ×

Collections of paired datasets are ordered lists of dataset pairs (often forward and reverse reads). These collections can be pass... ▼

2 unpaired forward - 2 filtered out

_R1

FDA1256931-C001-002_S1_L001_R1_001.fastq

FDA1257236-C001-001_S3_L001_R1_001.fastq

Clear Filters

Auto-pair

Pair these datasets

Pair these datasets

2 unpaired reverse - 2 filtered out

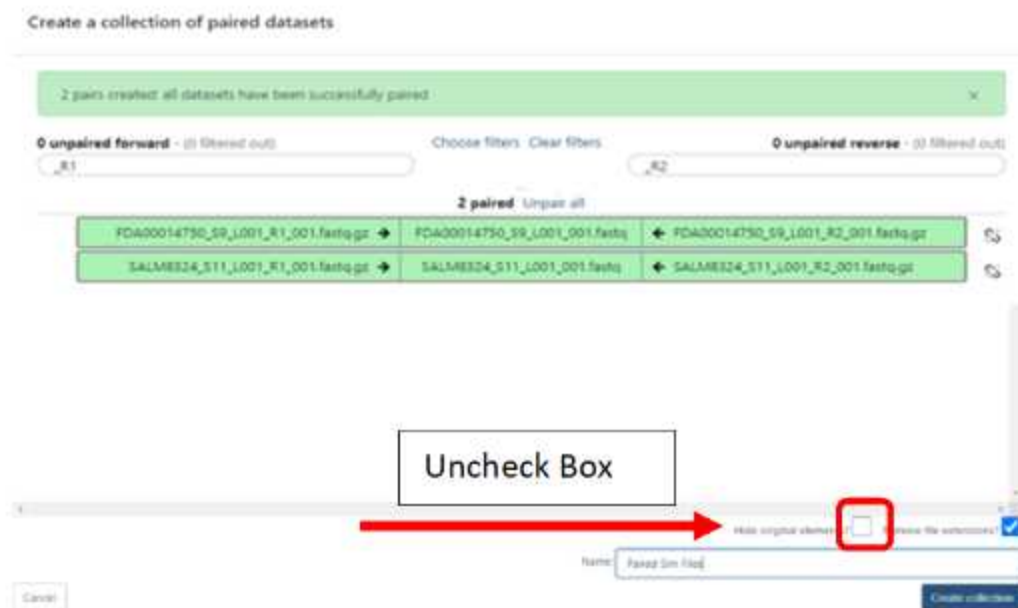
_R2

FDA1256931-C001-002_S1_L001_R2_001.fastq

FDA1257236-C001-001_S3_L001_R2_001.fastq

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

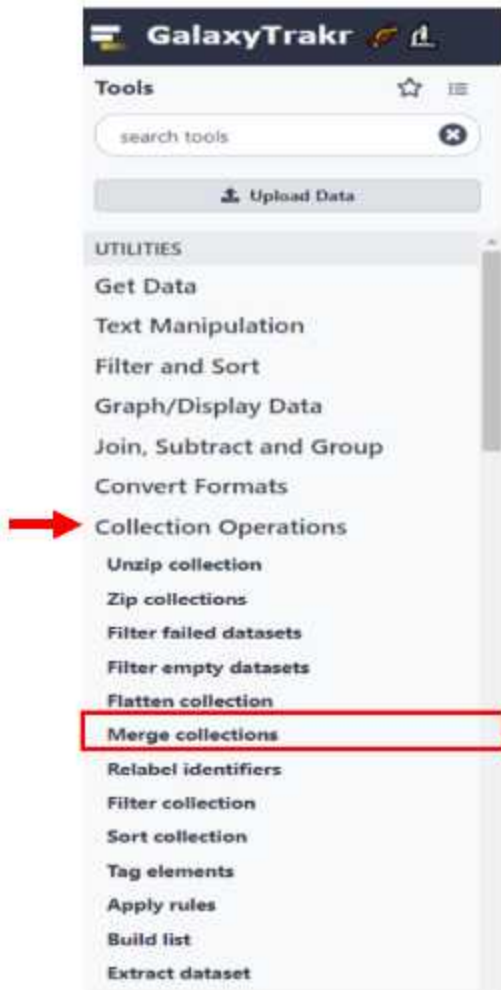
Click “**Create list**”



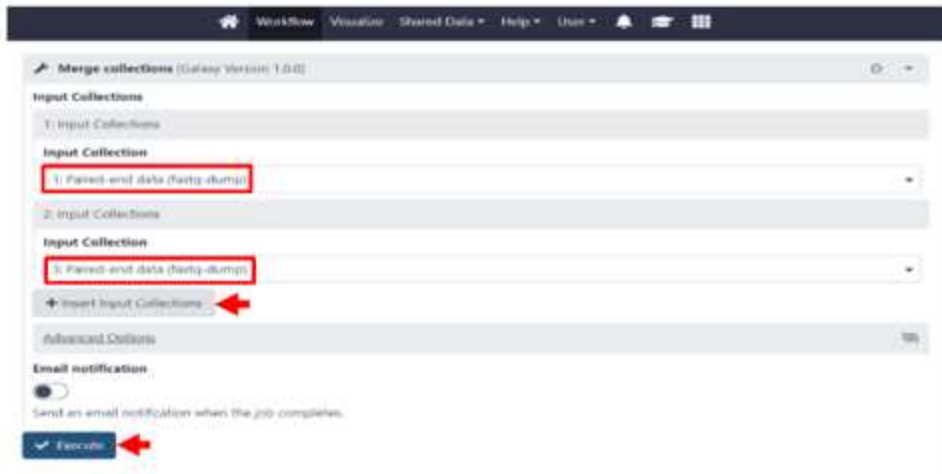
You should see your named list in the history panel. Continue with step #5.

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. Continue 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”**




Workflow: imported: SeqSero2 v1.2.1 workflow tabular and row outputs



✓ Run Workflow

1

30: Paired Sim Files



Expand to full workflow form.

History   

search datasets  

SeqSero Prediction


24 shown, 18 hidden

381.06 MB   

30: Paired Sim Files 

a list of pairs with 2 items

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

 Successfully invoked workflow **imported: SeqSero2 v1.2.1 workflow tabular and row outputs**




You can check the status of queued jobs and view the resulting data by refreshing the History pane, if this has not already happened automatically.



Invocation 1...

7 of 7 steps successfully scheduled.

0 of 8 jobs complete...




- Inputs
- Outputs
- Output Collections
- Steps




History   




search datasets  




SeqSero Prediction




30 shown, 18 hidden

381.06 MB   




54: Seqsero Tabular output   

53: Concatenate datasets on data 51 and data 50   




52: Seqsero Row output   

49: Remove beginning on collection 43   

2 jobs generating a list

46: Remove beginning on collection 43   

2 jobs generating a list

43: SeqSero Results   

2 jobs generating a list

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.

✓ Successfully invoked workflow **Imported: SeqSero2 v1.2.1 workflow tabular and row outputs.**

You can check the status of queued jobs and view the resulting data by refreshing the History pane, if this has not already happened automatically.

View Report 1

7 of 7 steps successfully scheduled.

9 of 9 jobs complete.

Download BioCompute Object

- ▶ Inputs
- ▶ Outputs
- ▶ Output Collections
- ▶ Steps

History

search datasets

SeqSero Prediction

30 shown, 18 hidden

381.06 MB

- 54: Seqsero Tabular output
- 53: Concatenate datasets on data 51 and data 50
- 52: Seqsero Row output
- 49: Remove beginning on collection 43
- 46: Remove beginning on collection 43
- 43: SeqSero Results

View and export results

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

Sample name	Output directory
FDA757411-116_S2_L001_001.fastq_1.fastq	/data/job_working_directory_s3/002/528/2528414/working/output
FDA1239824-C002-001_S1_L001_001.fastq_1.fastq	/data/job_working_directory_s3/002/528/2528415/working/output

History

search datasets

SeqSero Prediction

30 shown, 18 hidden

381.06 MB

- 54: Seqsero Tabular output
- 53: Concatenate datasets on data 51 and data 50
- 52: Seqsero Row output
- 49: Remove beginning on collection 43
- 46: Remove beginning on collection 43
- 43: SeqSero 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	/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	
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.

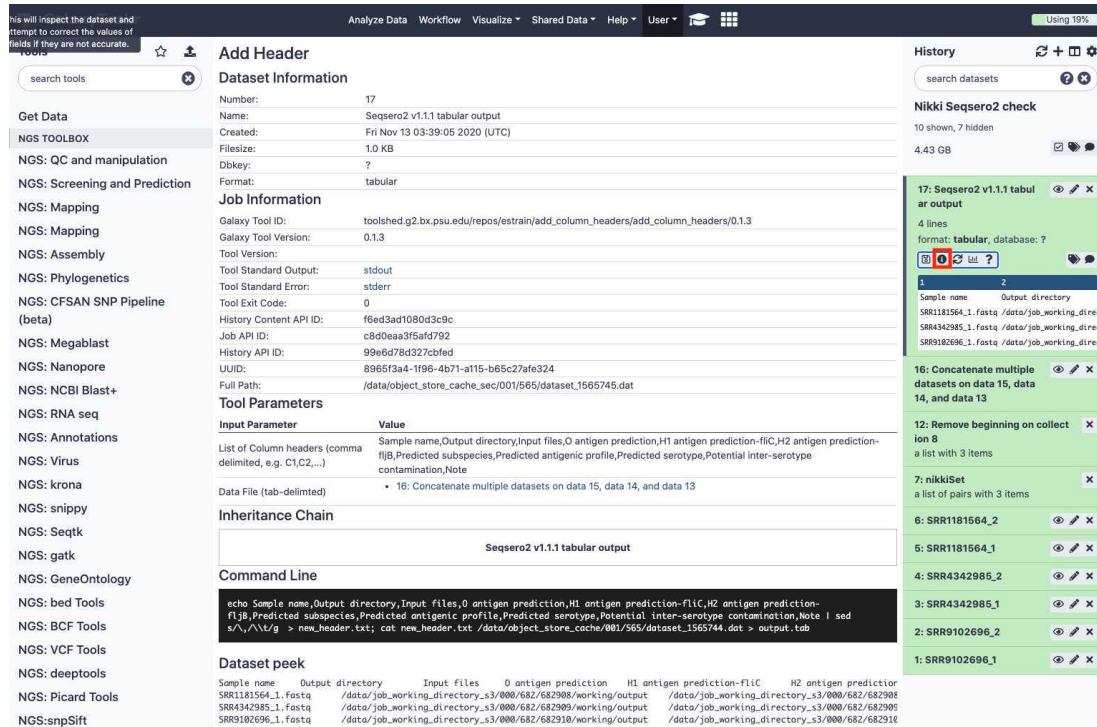
The screenshot shows the SeqSero2 web interface. On the left, there is a sidebar with a search bar and a list of tools under 'Get Data'. The main area displays a table with columns for Sample name, Output directory, and Input files. The table contains three rows of data. On the right, there is a 'History' panel showing a list of datasets. The dataset 'Nikki Seqsero2 check' is selected, and its details are shown, including a 'Save' button (a floppy disk icon) and a 'Download' button (a download icon). The 'Save' button is highlighted with a red box.

 SeqSeroExampleResults.tabular

Example results file:

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.



The screenshot displays the GalaxyTrakr web interface. On the left is a sidebar with a 'Get Data' section and an 'NGS TOOLBOX' containing various tools like 'NGS: QC and manipulation', 'NGS: Screening and Prediction', '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', and 'NGS: snpSift'. The main panel shows the 'Add Header' tool configuration for 'SeqSero2 v1.1.1 tabular output'. It includes sections for 'Dataset Information' (Number: 17, Name: Seqsero2 v1.1.1 tabular output, Created: Fri Nov 13 03:39:05 2020 (UTC), Filesize: 1.0 KB, Dbkey: ?, Format: tabular), 'Job Information' (Galaxy Tool ID, Galaxy Tool Version: 0.1.3, Tool Version, Tool Standard Output: stdout, Tool Standard Error: stderr, Tool Exit Code: 0, History Content API ID, Job API ID, History API ID, UUID, Full Path), 'Tool Parameters' (Input Parameter, Value, List of Column headers, Data File), 'Inheritance Chain' (Seqsero2 v1.1.1 tabular output), 'Command Line' (echo Sample name, Output directory, Input files, O antigen prediction, H1 antigen prediction-fl1C, H2 antigen prediction-fl1B, Predicted subspecies, Predicted antigenic profile, Predicted serotype, Potential inter-serotype contamination, Note | sed s/\t/g > new_header.txt; cat new_header.txt /data/object_store_cache/001/565/dataset_1565744.dat > output.tab), and 'Dataset peek' (Sample name, Output directory, Input files, O antigen prediction, H1 antigen prediction-fl1C, H2 antigen predictor). On the right, a 'History' panel shows a list of datasets, including '17: Seqsero2 v1.1.1 tabular output' and '16: Concatenate multiple datasets on data 15, data 14, and data 13'.

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)