# Galaxy Genome Trakr User Guide

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

# **Version History**

Version	Implemented	Revision	Approved	Approval	Description of
Number	Ву	Date	Ву	Date	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**

V Li 1	ERSIOI IST OF INTI	N HISTORY FIGURESI RODUCTION	 V 1
	1.1 1.2	GALAXY GENOME TRAKR INFORMATION PURPOSE	1 1
2	ONE	BOARDING	1
	2.1 2.2	INITIAL ACCESS ACCESS TO GALAXYTRAKR.ORG	1 3
3	USI	NG GALAXY- QUICK START GUIDE	4
	3.1 3.2 3.2. 3.2. 3.3 3.4 3.5	CREATE AND NAME A HISTORY         UPLOAD DATA	4 5 <i>9</i> 1 3
4 5 6	SEF GEN ASS	ROTYPE PREDICTION WITH SEQSERO	.6 .8 2
	6.1	OUTPUTS OF QUAST	3
7	USI	NG THE SNP PIPELINE WORKFLOW 2	5
	7.1	Additional SNP Pipeline Information	7

# List of Figures

Figure 1. Galaxy Trakr Password Reset	2
Figure 2. GalaxyTrakr.org Login	3
Figure 3. Username and Password	3
Figure 4. Create a New History	4
Figure 5. Change the History Name	4
Figure 6. Log into SFTP client	5
Figure 7. Unknown Host Key Message	5
Figure 8. Files to Upload	6
Figure 9. Get Data	6
Figure 10. Choose FTP File	7
Figure 11. Select Uploaded Files	7
Figure 12. Upload Status	8
Figure 13. Choose local file	9
Figure 14. Select a local file	9
Figure 15. Uploading files into the Galaxy server	10
Figure 16. Data upload in progress	10
Figure 17. History View	10
Figure 18. Switch to a different history	11
Figure 19. Finalizing history selection	11
Figure 20. Data Libraries	11
Figure 21. Create New Folder	12
Figure 22. Add data set from history	12
Figure 23. Shared Data	13
Figure 24. Data Libraries	13
Figure 25. Galaxy Data Libraries	14
Figure 26. Target Data Folder	14
Figure 27. Import Selected Datasets into History	14
Figure 28. Select an existing history	15
Figure 29. Create a new history	15
Figure 30. Visible Data Set	15
Figure 31. GenomeTrakr Tools	16
Figure 32. SeqSero Batch – Paired-End Reads	16
Figure 33. Select runs and Execute	17
Figure 34. SeqSero Results	17
Figure 35. Locating SPAdes	18
Figure 36. SPAdes Input	18
Figure 37. Execute SPAdes	19
Figure 38. SPAdes Log	19
Figure 39. SPAdes Contigs	20
Figure 40. SPAdes Scaffolds	20
Figure 41. Scattold Stats	21
Figure 42. Contig Stats	21
Figure 43. QUAST in the Galaxy Toolbar	22
Figure 44. QUAST configuration	22
Figure 45. QUAST interactive HTML report	23

Figure 46. Summary statistics tooltips	23
Figure 47. Icarus Contig Size viewer	24
Figure 48. Basic SNP-Pipeline Workflow	25
Figure 49. Uploading files for a collection of paired reads	
Figure 50. Run the workflow	
Figure 51. Inputs to the pipeline	
Figure 52. Resulting SNP Distance Matrix	

### **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 <u>https://account.galaxytrakr.org</u>.
- 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

Secure   https://account.galaxytrakr.org							
	Galaxy Trakr Password	l Reset					
	Password Reset						
	Please complete all fiel	ds below. Passwords must meet the following requirements:					
	<ul> <li>At least 8 characters</li> <li>Not the same as the last 24 passwords</li> <li>Does not contain username or parts of name</li> <li>Contains at least one upper case, lower case, digit and Non-alphabetic characters</li> </ul>						
	Username	jsanders					
	Password						
	New Password						
	Confirm Password						
		Change Password					

Figure 1. Galaxy Trakr Password Reset

### 2.2 Access to GalaxyTrakr.org

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

- Open a browser to <u>https://galaxytrakr.org</u>. 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.

€ Secure   https://galaxytrakr.org							
xy / GENOME	RAKR Analyze Data Workflow Shared Data - Visualization - Help - Login or Register -						
<u>±</u>	Login						
ls 😢							
	Hello, Galaxy is running!						

Figure 2. GalaxyTrakr.org Login

Login	
Username / Email Address:	
jsanders	
Password:	
•••••	
Forgot password? Reset here	

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.

🔁 sf	tp://jsanders@upload	l.galaxytrakr.o	rg:443 - FileZilla					
File	Edit View Transfe	er Server E	Bookmarks Help	New version	available!			
111.			8 🍡 🖏 1	E 🔍 🧧	<del>80</del>			
Host:	sftp://upload.galax	Username:	jsanders	Password:	•••••	Port:	443	Quickconnect

Figure 6. Log into SFTP client

The first time connecting you will be asked to trust the host being connected to. Click OK to trust the connection.

See Figure 7.

Unknow	n host key	×
1	The server's host I computer you thin	key is unknown. You have no guarantee that the server is the k it is.
~	Details	
	Host:	upload.galaxytrakr.org:443
	Hostkey algorith	m: ssh-ed25519 256
	Fingerprints:	SHA256: o+DRax+tPkY2Wi9alNF5/yd9RuE7DEBONIXTB/AEL+c= MD5: 2c:e2:a7:9e:98:b5:89:ff:19:43:fc:5d:82:3d:c0:f8
	Trust this host and	carry on connecting?
	🔽 Always trust th	nis host, add this key to the cache
		OK Cancel

Figure 7. Uknown Host Key Message

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

 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.

ename	Filesize Filetype	Last modified	Filename	- Fies	oe filetype	Last modified	Permissions	Owner/Gro
	350.592,954 FASTQ File	8/24/2017 1:35:24	SRR3933022	165,7 1.fartq 348,111.0	73 Microsoft . 16 FASTQ-File	8/29/2017 12:2. 8/29/2017 10:3.,	-74437963	20039 33264 20039 33264

Figure 8. Files to Upload

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



8. Click the Choose FTP File button.

### See Figure 10.

Regular	<u>Composite</u>	Collection				
						-
			C Dura Class			
			Drop files	shere		
						Ŧ
Ту	pe (set all):	Auto-detect	⊤ Q	Genome (set all):	Additional Species A 🔻	

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.

FTP	files		O				
This Galaxy server allows you to upload files via SFTP. To upload some files, log in to the SFTP server using a SFTP client, like FileZilla, at <b>upload.galaxytrakr.org</b> using your Galaxy credentials (username and password).							
Ava	ilable files:		🖹 3 files 🖨 666.5 MB				
⊟	Name	Size	Created				
	Records.csv	<b>161.9</b> KB	08/30/2017 10:48:45 PM				
	SRR3933021_1.fastq	332 MB	08/30/2017 10:48:46 PM				
Ø	SRR3933021_2.fastq	334.4 MB	08/30/2017 11:17:26 PM				

Figure 11. Select Uploaded Files

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

Downloa	ad from web	or upload f	rom disk				
<u>Regular</u>	<u>Composite</u>	<u>Collection</u>					
	Name	Size	Туре	Genome	Settings	Status	A
⊳ SRR39	33021_2.fastq	334.4 MB	Auto-dete 🔻 <b>Q</b>	Additional Sp 🔻	۵ (	100%	~
							•
	Type (set all):	Auto-	detect v Q	Genome (set a	all): [ A	dditional Species A	Ŧ
		😐 Choose le	ocal file 🕞 Choose	FTP file 🕼 Paste/Fetc	h data Paus	se Reset Star	Close
L							

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 ( 1) 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.

Salaxy / GENOME TRAKR	Analyze Data Winnifum Smirrid Data - Viscolization - Ump - Dent - 111	-10	Using 0 bytes
Tools 🕹	Download from web or upload from disk		History 20
search tools			Rearch datasets.
Get Data Hello,	Gale Regular Composite Collection		New_history
Send Data To custo	mize		(empty) 🗣 🕏
Text Manipulation Eilter and Soct Join, Subtract and Group	Ing C		O This followy is simply. You can been your own data or get data from an external source
Extract Fostures Take an Get Generat Sources Fetch Sequences Fetch Aligoments	inter. දා Drop files here		
Phenotype Association NGS TOOLERX BETA Galaxy NGS: OC and manimulation of many	s an t	ith the support	
Statistics The Galaxy Groph/Display Data	collect -	veraltz-	
Workflows + All workflows	Type (set all):     Auto-detect     a     Q     Genome (set all):		

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.

Transferred Genome TRA	AKR	Analyze Data Winkfile	- Shimilana	wananzatine •	miles and				Using 0 byte
Tools 👤	Download f	rom web or upload from d	isk					History	200
saarch tools	Hello, G	dory Test Data				• • • • Search New, history Test Data		C search datasets New_history	-
Send Data Text Manipulation	To custon Organize + New fol	der Hanne	Date modified	Type	See	E• 🖪 0		(empty)	nty, You can
Elline and Sort Joln, Soldrect and Group Genvert Jornatis Extract Leatures Gait Genomic Scores Fatch Sequences Fatch Alignments Phenotype Association MCS TOOL DOC MITA MSS: QC and manipulation Statistics Graph/Disoley.Data EDA Tools	Configure Constant Configure Constant Configure Constant Take an is Configure Constant Configure Cons	<ul> <li>PTC002_S1_L000_FFI_001_Antq</li> <li>PTC002_S1_L001_FF2_001_Antq</li> </ul>	6/36/20112:59 PM	FASTQ File FASTQ File	124,427.43 125,188.43		the support	tool anor over test from an external o	e or get data Sarte
All markflores	Fie	name: "PTC002,51,1001,82,001,/artq" "PTC0	02_53_L001_R1_001.festq*	-	-	Al Files (**)     Cancel			

Figure 14. Select a local file

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

#### Galaxy Genome Trakr User Guide

T Galaxy / GENOME	TRAKR	Anal	yze Data - Workflow - Silamit Da	na visuilizationa Helpa Helpa			Us	iling O byte:
Tools 2		Download from web or	upload from disk				History	200
EDA Tools	Hello, Gala	Regular Composite Colle	ection				search datacets	1
Get Data Send Data	To customize		Please wait2	out of 2 remaining.			(empty)	
Text Manipulation Filter and Sort	Configuring G	Name	Size Type	Genome Settings	Status		This history is empty.	Voli can get data
Join, Subtract and Group Convert Formats		PTC002_S3_L001_R1_ 3 001.fastq	116.8 MB Auto-deta. + Q	Additional Sp. 🔹 🔷	0		from an external sour	CC#
Extract Features Get Genomic Scores Fetch Sequences Fetch Alignments	Take an inter	PTC002_53_L001_R2_ 3 001.fastq	S17.6 MB Auto-dete + Q	Additional Sp. +	9			
Phenotype Association Inca TOOLBOX BETA NGS: OC and manipulation	<u>Galaxy</u> is an o of <u>many cont</u>					ith the support		
Statistics Graph/Display Data EDA Tools	The Galaxy Project i					iversity.		
Workflows		Type (set all):	Auto-detect Q	Genome (set all): Add	Reset Start Close			

Figure 15. Uploading files into the Galaxy server

Figure 16 below depicts uploads in progress.

T Galaxy / GENOME T	RAKR	An	alyze Data Wit	ritiya Shirya D	ataw Wendlesteinew	Nupe Vore	111	-			
Tools 1		Download from web o	r upload from	n disk						History	00
search tools		Terraria di secondaria dal								asarch datasets	
Get Data	Hello, Gala	Regular Composite Co	lection							New_history	
Send Data	To customize			Please wait	1 out of 1 remaining,					L SCOWEL	100 100 10
Text Manipulation Filter and Sort	Configuring G	Name	Size	Туре	Genome	Settings	Status			315.52 110	
Join, Subtract and Group Convert Formats		PTC002_53_L001_R1_ 001.fastq	316.8 MB Au	no-deta. • Q	Additional Sp *	• 1	100%	*		PTC002 S3 1001 P	1 001.fastq
Extract Features Get Genomic Scores Fetch Sequences Fetch Alignments	Take an inter	D01.fastg	317.6 MB Au	to-dets. + Q	Additional Sp. +	0	0%	o			
Phenotype Association NGB TOOLBOX BETA NGS: OC and manipoletion	<u>Galaxy</u> is an of <u>many cont</u>								ith the support		
Statistics Graph/Display Data EDA.Tools	The Galaxy Project								berraity-		
Workflows  All workflows		Type (set all):	Auto-detec	t Q	Genome (set a	H): Addi	ional Species A 🛛 👻				
				Choose I	ocal Ille 🛛 🖓 Paste/Fetch	data Pause	Reset Start	Close			
				_			_	-			

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.

🗧 Galaxy / GENOME T	RAKR Analyze Data Workfow Shared Data+ Visualization+ Heip+ Unit-	tis	ing 634.4 MB
Tools 1		History	C 0 []
(earch toble O) EPA Look Set Data Send Data Yest Kampainton Hins: and Second John, Suttract and Group Convert Formats Extract Leatures Extra Leatures Extra Leatures Extra Leatures	Hello, Galaxy is running! To customize this page edit static/welcome.html Configuring Galaxy » Installing Tools » Take an interactive tour: Gulary II (Meany Soundback)	Construct, datawata New_history 2 dilawata 634.38 MB 2: PTC002_52_1001_R2 PTC002_52_1001_R1	C C C C C C C C C C C C C C C C C C C
Phenotype Association NGS TOO BOX BETA NGS: OC and manipulation Statistics Graph/Display.Data EDA Tools Workflows - All machines	<u>Galaxy</u> is an open platform for supporting data intensive research. Galaxy is developed by <u>The Galaxy Team</u> with the support of <u>many contributors</u> . The <u>Galaxy Project</u> is reported in purt by <u>BHGEL ISE</u> . The <u>Herck Institutes of the UR Sciences</u> . The <u>Lestinde for Cales Science at Jeon State</u> , and <u>Johns Hopking University</u> .		

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 (  $\square$ ) in the upper right corner.
- Select the history you would like to use by clicking Switch to. See Figure 18.

🚍 Galaxy / GENOME TRAKR	Analyze Data Worldlow Shared Data • Visualization • Help • User • 🇱	Using 634.4 MB
Done search histories	O search all datasets O	Create new
Current History	Switch to	(v)
New_history 2 shown	Unnamed history 1 shown	
634.38 MB	130.36 MB	B 🗣 🗭
search datasets	search datasets	0
Drag detailets here to easy them to the current history	1: SRR1198853_1.fastq	
2: * * * X PTC002 S3 L001 R2 001.fasta 1: * * * * * PTC002 S3 L001 R1 001.fasta		

Figure 18. Switch to a different history

3. Click Done.

See Figure 19.

🚍 Galaxy / GENOME TRAKR	Analyze Data Workflow Shared Data • Visualization • Help • User • 🇱	Using 634.4 MB
Done search histories	O search all datasets O ==	Create new
Current History	Switch to	
New_history 2 shown 634.38 MB	Unnamed history 1 thoun 130.36 MB	2 <b>6</b> •
Search datasets O	search datasets 1: SRR1196853_1.fastq	0 • / x
2: * * * * * * * * * * * * * * * * * * *		

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.

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

Galaxy /		Shared Data <del>-</del> Visualiz	zation <del>-</del> Help - User -	
DATA LIBRARIES	<ul> <li>&lt; 0 1 2 » showing 1 of 1 items</li> <li>X Delete O Details</li> <li>and Drug Administration (FDA)</li> </ul>	Create N	to History	🕑 Help
□ name ↓ <sup>A</sup>	description	data type size	time updated (UTC)	
Sub Share		folder	2017-08-14 04:47 PM	/
		wing $\underline{1}$ of $1$ items		

Figure 21. Create New Folder

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

Libraries / Food and Drug Administration (FDA) / data_folder from History from User Directory	
name J <sup>A</sup> description data type size time updated (UTC)	

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:

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



Figure 23. Shared Data

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

= Galaxy /	GENOME TRAI	CR Analyze Data Workflow Shared Data Visualization + Admin Help + User + III	Using 0 bytes
Tools	2	Data Ubraries da	History 201
( search tools	0	Histories	C search datasets
EDA Tools Get Data Send Data		Hello, Galaxy is runi Vypulizations To customize this page ed	Unnamed history (empty)
end Data ext Manipulation lifer and Sort in. Subtract and Group onvert. Formats stract.Features et.Genomic.Scores et.Genomic.Scores	Configuring Galaxy > Installing Tools > Take an interactive tour: Galaxy UI (Installing Scretchbook)	This hotory is empty. You can land your own date or gat, date from an external source	
Fetch Alignments Phenotype Associat NGS TOOLBOX BETA NGS: QC and manip Statistics Graph/Display Data FDA Tools Workflows • Aliworkflows	ion Vulation	<u>Galaxy</u> is an open platform for supporting data intensive research. Galaxy is developed by <u>The Galaxy Team</u> with the support of <u>many contributors</u> . The Galaxy Project is supported in part by NHGRL NSF. The Huck Institutes of the Life Sciences. The Institute for CyberScience at Penn State, and Johns Hopkins University.	

Figure 24. Data Libraries

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

DATA LIBRARIES	braries Search	Include deleted + New Library	Help
namelt	description	synopsis	
Division of Consolidate Laboratory Services	Laboratory Library	Tool Analysis - Development and Implementation of DCLS pipelines into Galaxy tools.	
Florida Department of Agriculture and Consumer Services, Division of Food Safety. Bureau of Food Laboratories	Laboratory Library	WGS Data Analysis	
Food and Drug Administration (FDA)	FDA Library	WGS Data Analysis	1 2
Food and Drug Laboratory Branch, CA Dept of Health	Laboratory Library	WGS Data Analysis	/ 2
Massachusetts Department of Public Health	Laboratory Library	WGS Data Analysis	/ 4
New York State Dept of Agriculture and Markets, Food Lab	Laboratory Library	WGS Data Analysis	/ 1
Ohio Animal Disease Diagnostic Laboratory	Laboratory Library	WGS Data Analysis - Illumina MiSeq	1 3
Penn State	Laboratory Library	Tool Analysis - Integration of Genome Epidemiology tools. SNP Analysis, Phylogenetic Tree Construction, Gene Identificat	/
Texas Department of State Health Services	Laboratory Library	Tool Analysis - Salmonelia Serotype prediction and Cluster Analysis for food-borne and healthcare associated infectious	
Virginia State Laboratory	Laboratory Library	WGS Data Analysis	/ 3
Washington State Department of Health, Public Health Laboratories	Laboratory Library	WGS Data Analysis - Salmonella, Listeria and E. coli	23

Figure 25. Galaxy Data Libraries

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

Galaxy / Gl	NOME TRAKR Analyze Data Workf	ow Shared Data - Visualization - Admin	Help+ User+			Using
DATA LIBRARIES	C 0 1 2 > showing 1 of 1 items Include deleted	+ + + + to History 🕹 Download	• X Delete	Details		0 H
Libraries / Food an	d Drug Administration (FDA)					
🔲 name l‡	d	escription	data type	size	time updated (UTC)	
Sub Share			folder		2017-08-14 04:47 PM	/ 4
	ſ	< 0 1 2 > showing 1 of 1 items				
	.[	0   1   2   > showing 1 of 1 items				



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.

Galaxy / GE	NOME TRAKR Analyze Data Work	flow Sh Import selected dataset history	s into Admin He	lp∀ User≁			ι	Jsing 0 byt
DATA LIBRARIES	C C I 2 S showing 1 of 1 items. I include deleted id Drug Administration (FDA)	+ History	≛ Download -	× Delete	0 Details			O Help
🕘 name l‡		description		data type	size	time updated (UTC)		
🖿 🏾 Sub Share				folder		2017-08-14 04:47 PM	2 8	
		< 0 1 2 > showing 1 of 1	items					

Figure 27. Import Selected Datasets into History

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

na Galaxy / GENOME TRAKR	Analyze Data Wonkflow Shared Data+ Visualization+ Admin Help+ User+	Using 0 byte
DATA LIBRARIES = 0 1 2 > showing 1 c	Import into History	e Help
Libraries - Food and Drag Administration (FDA)	Select history: Unnamed history N or create new: Unnamed history	time updated (UTC)
an 🖉 Sub-Share	Import Close	2017-00-14 04:47 791

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.

🖥 Galaxy / GENOME TRAKR	Amelyze Deta Workflow Shared Data + Visualization + Admin Help + User + 🇱	Using 0 bybe
DATA LIBRARIES = 0 3 2 + showing 1 c	Import into History	@ Help
Ubracies / Food and Drug Administration (FDA)	Select history: Unnamed history	
Cita in anne 12	or create new: hame of the new history	time updated (UTC)
ter.) In 🥙 Sub Shere		2017-08-14 04:47 PM . 🛛 🗑
	Import Close	

Figure 29. Create a new history

- 8. Click Import.
- 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:

 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.
- Click the eye icon ( ) at the upper right box to view a table of your results. See Figure 34.

( galaxy.ravencloudhpc.com:808	0			c	Q fret serot	ype	→ ☆ 自 ♥	• ♠ ≡
🚍 Galaxy / FDA-CFS/	AN Analyze Data	Workflow Sha	red Data <del>-</del> Visualiza	ition <del>+</del> Help <del>+</del> Use	er~ III			Using 4.6 GB
Tools 🕹	SeqSero Results						History	2 O 🗆
Search tools Search tools Search tools Map with BWA - map short	Input Files	O antigen prediction	H1 antigen prediction(fliC)	H2 antigen prediction(fljB)	Predicted antigenic profile	Predicted serotype(s)	John1 40 shown, 4 <u>deleted</u>	8
reads (< 100 bp) against reference genome	dataset_204_SRR3933082.fastq dataset_205_SRR3933082.fastq	O-4	i	1,2	4:i:1,2	Typhimurium	5.34 GB	g 🔌 🏓
<u>Map with BWA-MEM</u> - map medium and long reads (> 100 bp) against reference genome	dataset_10_SRR1202985.fastq dataset_9_SRR1202985.fastq	O-8	l,v	1,2	8:1,v:1,2	Pakistan or Litchfield*	44: SeqSero Batch - Paired-End Reads on data 42 data 41 and	⊕ / ×
Kraken-report view a sample report of your classification	dataset_157_SRR1198854.fastq dataset_158_SRR1198854.fastq	O-7	k	1,5	7:k:1,5	Thompson	7 lines format: txt. database: 7	viners.
Kraken-mpa-report view report of classification for multiple	dataset_161_SRR3933079.fastq dataset_162_SRR3933079.fastq	0-4	i	1,2	4:i:1,2	Typhimurium	n CLASSPATH=/nfs/sw/apps /varscan/2.3.9 /VarScan.v2.3.9 jar ;export CLASSPATH;LOADEDMODULES= ;export LOADEDMODULES;PATH=/nfs /sw/lang/java/re1.8.0_112	
samples SeoSero Batch - Paired-End	dataset_159_SRR3933080.fastq dataset_160_SRR3933080.fastq	O-8	i	z6	8:i:z6	Kentucky		
Reads Salmonella Serotype Prediction from Paired-End Reads	dataset_202_SRR3933081.fastq dataset_203_SRR3933081.fastq	O-4	i	1,2	4:i:1,2	Typhimurium		
Get Data Send Data Text Manipulation							/sw/apps/spades/spa 3.9.0/bin:/nfs/sw/app	des- s/sratoolkit
Filter and Sort Join, Subtract and Group Convert Formats							Download iles 0_antig dataset_204_SRR3933082	gen_prediction
Extract Features Fetch Sequences							dataset_10_SRR1202985. dataset_157_SRR1198854	fastą dataset. fastą dataset

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.

NGS TOOLBOX
NGS: QC and manipulation
NGS: Assembly
SPAdes genome assembler for
regular and single-cell projects
Quast Genome assembly Quality
NCS: Screening and Prediction

Figure 35. Locating SPAdes

 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.

	E TRAKE Analyze Data Workflow Shared Data - Visualization - Help - User-	
ools	1: Libraries	A History
search tools	Library type	
A Tools	Paired-end / Single reads	search datasets
t Data	Orientation	New_history
nd Data	-> <- (fr)	2. shown, & general
. Manipulation	Files	1.24 GB
r and Sort	1: Files	8:
Subtract and Group	Select file format	PTC002_53_1001
act Features	Separate input files	Zi
Genomic Scores	Forward reads	PTC002 53 1001
h Sequences	8: PTC002_53_L001_R2_001_fastq	
h Alignments	FASTQ format	
notype Association	Reverse reads	
TOOLBOX BETA	8: PTC002_53_L001_92_001.flaitq	
: OC and manipulation	FASTQ format	
tistics	Insert Files	
eph/Display Data	Insert Libraries	
Dates genome assembler for	It is not possible to specify only mate-pair libraries. Scatfulds are not produced if neither a paired-end nor a mate-pair library is provided.	
CORES VILLES CONTRACTOR CONTRACTOR	Darfile CI R made	
gular and single-cell projects	PacBio CLR reads	
pular and single-cell projects [	Pacific CLR root5	
gular and single-cell projects I kflows	Pacific CIX reals	
gular and single-cell projects I ktlows Lworkflows	Pacific CLR reads Protoco 53, L001, Jr2, 001 Jania 7: PEC002_53_L001_R1_001 Jania	
ngular and single-cell projects BL rkflows ILworkflows	Pacific CLR reads  R: PFC002_S3_L001_R1_0015ielia  F: PFC002_S3_L001_R1_0015ielia  Nanopore reads	
gular and single-cell projects I ktiows workflows	Pacific CLR reads Process 51,1001_82_0015### Process 51,1001_81_0015### Process 51,1001_81_0015### Process 51,1001_82_0015### Process 51,1001_82_0015###	
gular and single-cell projects I kthows Lworkfloma	Pacific CLR reads Protoco 53, L001_P12_001_Seting 7: PFC002_53_L001_P12_001_Seting Ranopore reads RefC002_53_L001_P12_001_Seting 7: PFC002_53_L001_P12_001_Seting 7: PFC002_53_L001_P12_001_Seting	
ngular and single-cell projects BI rktiows Al mortflows	Pacific CIX reads F : PrCoc2 : 53 .1001_R3_001_Asting 7 : PrCoc2 : 53 .1001_R3_001_Asting Hanopore reads E : PrCoc2 : 53 .1001_R3_001_Asting 7 : PrCoc2 : 53 .1001_R3_001_Asting	
ngular and single-cell projects BI ricklows II. mothflows	Pacific CLR reads  Percess 51,1001,142,001/anita P; PFC002_53_1001_R1_001/anita Nanopore reads  Ranger reads Sanger reads	
pular and single-cell projects [ [ [diows .wackflowa	Pacific CLR reads Protoco 51,1001_P12_0015#888 7: PFC002_53_1001_P12_0015#888 Ranopore reads 8: PFC002_53_1001_P12_0015#888 9: PFC002_53_1001_P12_0015#888 8: PFC002_53_1001_P12_0015#888 8: PFC002_53_1001_P12_0015#888	
oviar and single-cell projects E Hollows workflows	Pactile CLR reads           F: PICO20_51_001_81_001_8488           F: PICO2_51_001_81_001_8488           Innepore reads           F: PICO2_51_001_81_001_8488           Sampler reads           E: PICO2_51_001_81_001_8488           E: PICO2_51_001_81_001_8488	
gular and single-cell projects I Kitows Liventifions	Bantion CLR reads           P. PICK02_53_1001_R1_001_Asting           P. PICK02_53_1001_R1_001_Asting           Banopore reads           Ex. PICK02_53_1001_R1_001_Asting           Sangee reads           Ex. PICK02_53_1001_R1_001_Asting           P. PICK02_53_1001_R1_001_Asting	
nguara no ungle-cell projects BI Intéléous Il markfilona	Fractic CLR mode           F: PrOcce_33_1001_R1_001_Rstm           F: Procce_33_1001_R2_001_Rstm           F: Procce_33_1001_R2_001_Rstm           Sanger reads           F: Procce_33_1001_R2_001_Rstm           F: Procce_33_1001_R2_001_Rstm           F: Procce_33_1001_R2_001_Rstm           Trusted contigs	
regular and single-cell projects BI whitewes NL workflows	Sampler reads           E: PTCOD2_51_001_81_001_8488           Sampler reads           E: PTCOD2_51_001_81_001_8488           B: PTCOD2_51_001_81_001_8488           F: PTCOD2_51_001_81_001_8488	

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.

🗮 Galaxy / GENOME	TRAKR Analyze Data Workflow Shared Data - Visualization - Help - Uber -	Using 2%
Tools	6: P1C002_51_0001_82_001.radq FASTO format	A History
( search tools )	Reverse reads	Canada datasala
EDA Tools Get.Data Send.Data	#1 PTC002_53_L001_R2_001.Fatq     KaS10 format     Tonget File	New history 2 hours of states 126 GB
Text Manipulation Filter and Sort Join, Subtract and Group Convert Formats	It is not possible to specify only make-pair libraries. Scattidies are not produced if nother a paired-end nor a mate-pair library is provided. Pacific CLR reads.	8: PTC002_S3_L001_R2_001.fastq
Extract Features Get Genomic Scores Fetch Sequences	8: PTC002_S3_L001_R2_001_Astq 7: PTC002_S3_L001_R1_001_Astq	Z; PTC002_53_L001_R1_001.festa
Fetch Alignments	Nasopore reads	
Phenotype Association NGS TOOLBOX BETA NGS: OC and manipulation	8: PTC002_53_L001_R2_001_hatq 7: 9TC002_53_L001_R1_001_hatq	
Statistics Graph / Display Data	Sanger reads	
EDA Tools <u>SPAdes</u> genome assembler for regular and single-cell projects	8: PTC002_53_L001_RE_001_Retq 7: PTC002_53_L001_RL_001_Retq	
NCBI	Trusted configs	
Workflows <ul> <li>All workflows</li> </ul>	8: FTC002_53_L003_R2_001_Auto 7: FTC002_53_L001_R1_001_fate	
	Untrusted contigs	
	6: PTC002_53_L001_R4_001_R4tq 7: PTC002_53_L001_R1_001_R4tq	
	Exercise	

Figure 37. Execute SPAdes

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

Command line:		History	2 <b>0</b> 0
/nfs/sw/apps/galaxy/database/jobs_directory/00/119/workingdisable-gzip-out careful -t l -k 21.33.55pel-frpel-fr	tput	search da	tasets C
fastg:/nfs/sw/apps/galaxy/database/files/000/dataset_159.datpel-2 fastg:/nfs/sw/apps/galaxy/database/files/000/dataset_160.datpacbio fastg:/nfs/sw/apps/galaxy/database/files/000/dataset_162.dat		Charles S 13 shown	
System information:		3.13 GB	<b>S A</b>
SPAdes version: 3.9.0		13: SPAdes	log 👁 🖋 🗙
Python Version: 3.5.3 OS: Linux-3.10.0-327.36.3.e17.x86_64-x86_64-with-centos-7.2.1511-Core		957 lines	
Output dir: /nfs/sw/apps/galaxy/database/jobs_directory/000/119/working Mode: read error correction and assembling Debug mode is turned OFF		format: txt, CLASSPATH	database: <u>?</u> H=/nfs/sw/apps/varscar
Dataset parameters: Multi-cell mode (you should set 'sc' flag if input data was obtained with MDA cell) technology ormeta flag if processing metagenomic dataset) Reads: Library number: 1, library type: paired-end control for for	(single-	CLASSPATH super ;exp LOADEDMO 3.9.0/bin:,	H;LOADEDMODULES=gal; ort DDULES;PATH=/nfs/sw/I /nfs/sw/apps/sratoolkit
orientation: ir left reads: ['/nfs/sw/apps/galaxy/database/files/000/dataset_159.dat'] right reads: ['/nfs/sw/apps/galaxy/database/files/000/dataset_160.dat'] interlaced reads: not specified single reads: not specified Library number: 2, library type: pacbio left reads: not specified right reads: not specified interlaced reads: not specified		Command lin base/files/ System info SPAdes vers	Lul ? Solution: he: /nfs/sw/apps/galaxy/c /000/dataset_159.datp prmation: sion: 3.9.0
single reads: ['/nfs/sw/apps/galaxy/database/files/000/dataset 162.dat']		-	

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.

This dataset is large and only the first megabyte is shown below. <u>Show all   Save</u>	search datasets
>NODE_1_length_485118_cov_44.1279 AGTGGATGGCATGGATGGCCGTTGATGGTCTCCACGGCCTGGTCGGCG GCGTTATCGTTACGGCACGTGGCTGGCGCCAAAGACGATACCATGCGCCGGCAG CGCATCGTGCGCAGCATGTTTTTTCTGTGGCCGGGATGGGGACTGGGACTGCGCCGGCG	Charles S 13 shown 3.13 GB
ATTATECATCTEGGTTGGCCGATCGCGCGCGTTTAACCGCGTAGCGCGTCC GCACTGAGTAATGGCACCAGGATCGGCGTGTTTAACCGCGTAGCGGCATCGGGGG CTGGTGCGGTACGGCTAAAATGCCCCCGGCGCTGGGTAAAGTGTGGCTGCTGGTCAGT ATGGCGCGCGGCGTGCATCAACTGGCATGACGGCGTCGTTTATCAGATAGAT	12: SPAdes scaffolds (fasta)
CCGGTGTTTGCGGCGCTACTGCTACGCATCGGCGCGCCCCCATTTGCAGCGTGACGTTT GCCAGTATTAGCGGCCTGGCATGGTGGGTGGCGTGACGGTCATCGTGACGTACGGTGCTG CTCTCTCTACTACAGTCCATGCAGCGCAGC	11: SPAdes contigs (ra       sta)       175 sequences       format: fasta, database: 2
CTANTCCGCGCGCGAACCCCATACCGTTGGTGTGCGCGCTGCTGCGCGCTG GCAGGCGAAATTATTGGCCGCGGGCGACTGTTTTTATGGCCTGCACATAGACCGTAGGTAG	CLASSPATH=/nfs/sw/apps/varscan, ;export CLASSPATH;LOADEDMODULES=gala super ;export LOADEDMODULES:PATH=/nfs/sw/Ja
CAGGOCGAFGGCCGCTTGATGCCGAGGCGCTTGGGGCCTGTGGGGCTATGTTTAAGACC CACAGCGAAGAGTGGTGCCACAGGCCTGGCAGCGTCTGTTTATGGCCCTTAGGCCTTG CCGTCTCCGCCGTGGGGTCGCTGGGATGGGA	3.9.0/bin:/nfs/sw/apps/sratoolkit

Figure 39. SPAdes Contigs

 Spades Scaffolds (fasta): Scaffolded contigs. SPAdes will attempt to use the pairedend 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: <u>http://spades.bioinf.spbau.ru/release3.9.0/manual.html</u>

## 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.

NGS: Assembly SPAdes genome assembler for regular and single-cell projects Quast Genome assembly Quality NGS: Screening and Prediction

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.

T Galaxy / GENOME	TRAKR Analyze Data Workflow Shared Data - Visualization - Admin Help - User -		Using 2.8 GB	
Tools 🚨	Quast Genome assembly Quality (Galaxy Version 4.5) = Op	History	C O 🗆	
(search tools (	Input assembly files	search datasets	0	
FDA Tools Get Data	1: Input assembly files Contigs/scaffolds output file	Unnamed history 30 shown, 9 deleted, 7 his	dden	
Send Data	37: SPAdes contigs (fasta)	• 2.83 G8		
Text Manipulation	Type of data	9 46: FLAG-10047 512	10	
Join, Subtract and Group	Contig 17: SPAdes contins (Esta)	01 R1 001.fastq		
Convert Formats Extract Features	+ Insert Input asse 24: SPAdes scaffolds (fasta)	45: FLAG-10047 512 01 R2 001.fastg	10 * * ×	
Get Genomic Scores Fetch Sequences	Size of reference ge 13: SPAdes scaffolds (fasta) 13: Orden en (fasta)	44: Quast report.html	* / ×	
Fetch Alignments	Estimated reference	43: karus Contig size	YI @ / X	
Phenotype Association	Reference File	41: Quast report tay	-	
NGS TOOLBOX BETA	0 0 D Nothing selected	-		
NGS: QC and manipulation Statistics	Many metrics can't be evaluated without a reference. If this is omitted, QUAST will only report the metrics that can be evaluat without a reference. (-R)	a)	× • • 21	
Graph/Display Data	Gene Annotations	33: SeqSero Batch - Pr	ii e/x	
FDA Tools	C 2 C Nothing selected	<ul> <li>ed-Lnd Reads on data</li> <li>0 and data 19</li> </ul>	12	
NGS: Assembly	File with gene positions in the reference genome. (-G)	32: Quast report.html		
Quast Genome assembly Quality	Operon Annotations	emoty		
NGS: Screening and Prediction	C C Nothing selected	format: html, database	1	
Workflows      All workflows	Type of organism	CLASSPATH=/nfs/sw/	apps/varscan/va	
	Prokarvotes	CLASSPATHI CADEDM	ODULES-malane	
	Lower Threshold	super ;export	super ;export LOADEDMODULES:PATH=/nfs/sw/lang 3.9.0/bin:/nfs/sw/apps/sratoolkit	
	500	LOADEDMODULES:PAT 3.9.0/bin:/nfs/sw/app		
	Set the lower threshold for a contig length. Shorter contigs won't be taken into account (default is 500) (min-contig) Thresholds	007		
	0,1000 Set the thresholds for contig length. Comma-separated list of contig length thresholds (default is 0,1000) (==contig=thresholds)	31: Icarus Contig size	vi 👁 🖌 🗙	
	✓ Execute	30: Quast report.tex	* / ×	
<	Description		>	

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

N50 is the contig length such that	ontigs of size >= 500 bp, unless otherwise note
using longer or equal length contigs	e SPAdes_contigsfasta_
the assembly. Usually there is no	51
value that produces exactly 50%, so	144
the technical definition is the op)	41
maximum length x such that using	629 354
contigs of length at least x accounts	4 775 953
for at least 50% of the total p)	4 796 724
assembly length. 00 bp)	4 768 821
N50	298 806
N75	149 520
L50	6
L75	12
GC (%)	52.11
Mismatches	
# N's	0
# N's per 100 kbp	0

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: <u>http://quast.bioinf.spbau.ru/manual.html</u>.

# 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.

Canvas   CPSAN SNP Pipeline (imported	from uphoaded file)				
Cheference FASTA X angue			# 3 filter SMP Reports     #       # 4 Morge SMP     #       Colls     Colls       Colls     Colls       Doubleted SMP colls from your history     *       Master onD     *       excludes bort/D     *	A. Call consents (MP) X You Craits Craits Jointy calls, consents and, ref mcCD calls, consents and france calls, consents and france consents and france calls, consents and france cal	S. Merge VO's X Use of VO's X upmar (with a      X. Create a SNP manne X Use of ASTAs Extent reference Fues Del tas then upwer history regime (flasse)     4
	L. Noo Reals     R Select reference FASTA Difference reals	At Call Sites. 8 FASTA Reference from your hispary Dises Journet to reference			paines (tsv) distma (tsv) referenceSNP (tasta) metrics (tsv)
	align_from_collection (sam)	calls (set)D private (private(D)		_	▲ 6. Herge VCFs ★ Use of VCFs
Cl Collection of Panel-End M Radis autout	Classifying Classifying	<ul> <li>Classifier</li> <li>Classifier</li> <li>Classifier</li> </ul>	✓ 4. Merge Sant X calls unputs that	AS Call concess SWA X 1995 Calls Calls Call, particle alli, concessor, fr (fits) call, concessor, fr (fits) call, ipp fits)	Angenav ked) *7. Create a SAP matrix * Usr of FASTA's Select inference fasta SAP last from your history snorms (fasta) = aains (tai)
					distria (tsv) = referenceSNP (fasta) = metrics (tsv) 0

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.
- Set File Type to fastqsanger.
   This ensures maximum compatibility with the SNP-Pipeline tools and the rest of the GalaxyTrakr ecosystem.
- 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..

Regular	Composite Collection				
	Yo	u added 10 file(s) to the queue. Add	more files or click 'Start' to proce	ed.	
		Name	Size	Status	
	MOD1-EC4447_S3_L0	01_R2_001.fastq.gz	831.8 MB	0%	⑪
묘	MOD1-EC4447_S3_L0	01_R1_001.fastq.gz	781.6 MB		创
묘	AZ-TG98595_S1_L00	1_R2_001.fastq.gz	142.5 MB		甸
묘	AZ-TG98595_S1_L00	1_R1_001.fastq.gz	133.5 MB		甸
	CDPHFDLB-F11M01157-R1-2	_S1_L001_R2_001.fastq.gz	137.9 MB		圃
	CDPHFDLB-F11M01157-R1-2	_S1_L001_R1_001.fastq.gz	113.2 MB		圃
	MD-DOH-18-000049-1_51	4_L001_R1_001.fastq.gz	355.3 MB		圃
	MD-DOH-18-000049-1_51	4_L001_R2_001.fastq.gz	411.4 MB		⑪
Collect	ion Type: List of Pairs 🔻	File Type: fastqsanger	Q Genome (set all):	Additional Speci	ies 🔻

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.

Your workflows			
Name			
CFSAN SN	IP Pineline 👻	_	
	Edit		
CFSAN S	Run		
SRA ass	Share or Download		
	Сору		

Figure 50. Run the workflow

Users will have a chance to configure the workflow stages, but the default options are preconfigured 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: <a href="http://snp-pipeline.readthedocs.io/en/latest/usage.html#snp-filtering">http://snp-pipeline.readthedocs.io/en/latest/usage.html#snp-filtering</a>

• The SNP Pipeline generates most, if not all, of the analytic outputs described at the following link:

<u>http://snp-pipeline.readthedocs.io/en/latest/usage.html#outputs</u> 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