017

Fostering Public Health Bioinformatics and Collaboration with GalaxyTrakr

Jayanthi Gangiredla[1], Hugh Rand[2], Daniel Benisatto[3], Justin Payne[2], Charles Strittmatter[2], Jimmy Sanders[4], William J. Wolfgang[5], Kevin Libuit[6,7], James Herrick[8], Melanie Prarat[9], Magaly Toro[10], Thomas Farrell[2], James Pettengill[2] & Errol Strain[2]

[1]Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, Laurel, MD, 20708, USA, [2]Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, MD, 20740, USA, [3]DRT Strategies, Arlington, VA, 22203, USA, [4]SDS Solutions, Inc., Springfield, VA, 22152, USA, [5]Wadsworth Center, New York State Department of Health, Albany, NY, 12201, USA, [6]Division of Consolidated Laboratory Services, Department of General Services, Richmond, VA, 23219, USA, [7]Libuit Scientific LLC, Richmond, VA 23219, USA, [8]Biology Department, James Madison University, Harrisonburg, VA, 22807, USA, [9]Animal Disease Diagnostic Laboratory, Ohio Department of Agriculture, Reynoldsburg, Ohio, 43068, USA, [10]Laboratorio de Microbiología y Probióticos, Instituto de Nutrición y Tecnología de los Alimentos, Universidad de Chile, Santiago, ChileTecnología de los Alimentos, Universidad de Chile, Santiago, Chile

Abstract

In the United States, surveillance activity of infectious diseases - foodborne, hospital-acquired, zoonotic, or otherwise – is addressed by a federated system of county, state, and national agencies managing different streams of data relatively independently. This poses a challenge to the dissemination of techniques, tools, resources, data, and analysis among these disparate groups of public health scientists, despite their aligned aims.

The objective of this project is to create a cloud based user-friendly Bioinformatics platform that enables scientists from public health and food safety research labs without any bioinformatics knowledge, to run queries and obtain reliable, comparable and consistent results. This helps in harmonized interpretations of WGS results across laboratories by providing tools optimized for food pathogen surveillance.

The US Food and Drug Administration's Center for Food Safety and Applied Nutrition (FDA-CFSAN) addressed this challenge by creating GalaxyTrakr, a cloud-hosted Galaxy environment with curated tools for pathogen biosurveillance of sequencing data generated by GenomeTrakr and from other sources. A cost-effective scaling architecture in Amazon Web Services now addresses the needs of an increasing number of users executing an increasing number of jobs, exploring an increasing number of bioinformatics tools, collaborating on an increasing number of shared data sets, and developing an increasing number of formal analysis protocols based on the GalaxyTrakr platform.

Introduction

GalaxyTrakr is a Cloud-based Galaxy bioinformatics platform for FDA scientists, public health labs, academia and GenomeTrakr partners. GalaxyTrakr is a graphical user interface-based system that enables researchers without any command line experience to perform computational analyses and to share these analyses with others. GalaxyTrakr can be accessed via GalaxyTrakr.org

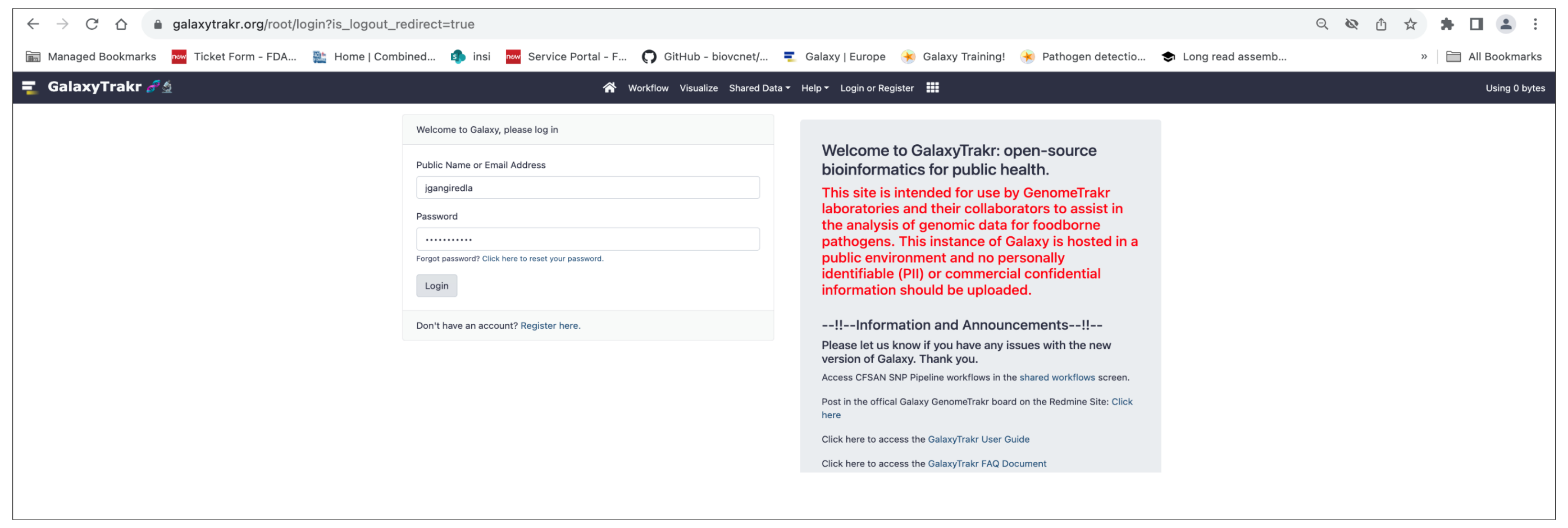
1. Enables State Public Health and agricultural laboratories, to rapidly assess quality and accuracy of WGS data prior to submission or sharing
   - Identify sample swaps, contamination, poor runs, etc.
2. Provides Public Health Laboratories/Microbiologists with access to easy-to-use bioinformatic tools
   - Sequence type, serotype, antimicrobial resistance, virulence/pathogenicity
   - Local outbreaks, integrate local WGS data with national/international data
   - Allows for early detection of emerging AMR threats through NCBI's AMRFinderPlus tool
3. Supports local outbreak analysis (e.g., restaurant cluster)
   - Identify the links between clinical isolates and positive food/environmental samples

GalaxyTrakr Statistics

- ➢ GalaxyTrakr currently has over 1700 active registered users.
- ➢ We support 200+ connected laboratories that include Public and State Health Labs, Academic Institutions, International Health Laboratories (Italy, Chile, Ireland, South Africa and India), Other Federal Organizations (CDC, USDA), Labs within FDA.
- ➢ The platform hosts over 1150 analytical tools and 26 galaxy workflows.
- ➢ To date, over 2,500,000 analytical jobs have been processed in the platform (~50,000 jobs per month)

Interface and Workflows

Figure 1. GalaxyTrakr login page where user can enter their login information and browse through documents, such as a user manual



Welcome to Galaxy, please log in

Public Name or Email Address: jgangiredla

Password: ••••••••••••

Forgot password? Click here to reset your password.

Don't have an account? Register here.

Welcome to GalaxyTrakr: open-source bioinformatics for public health.

This site is intended for use by GenomeTrakr laboratories and their collaborators to assist in the analysis of genomic data for foodborne pathogens. This instance of Galaxy is hosted in a public environment and no personally identifiable (PII) or commercial confidential information should be uploaded.

--!!--Information and Announcements--!!--

Please let us know if you have any issues with the new version of Galaxy. Thank you.

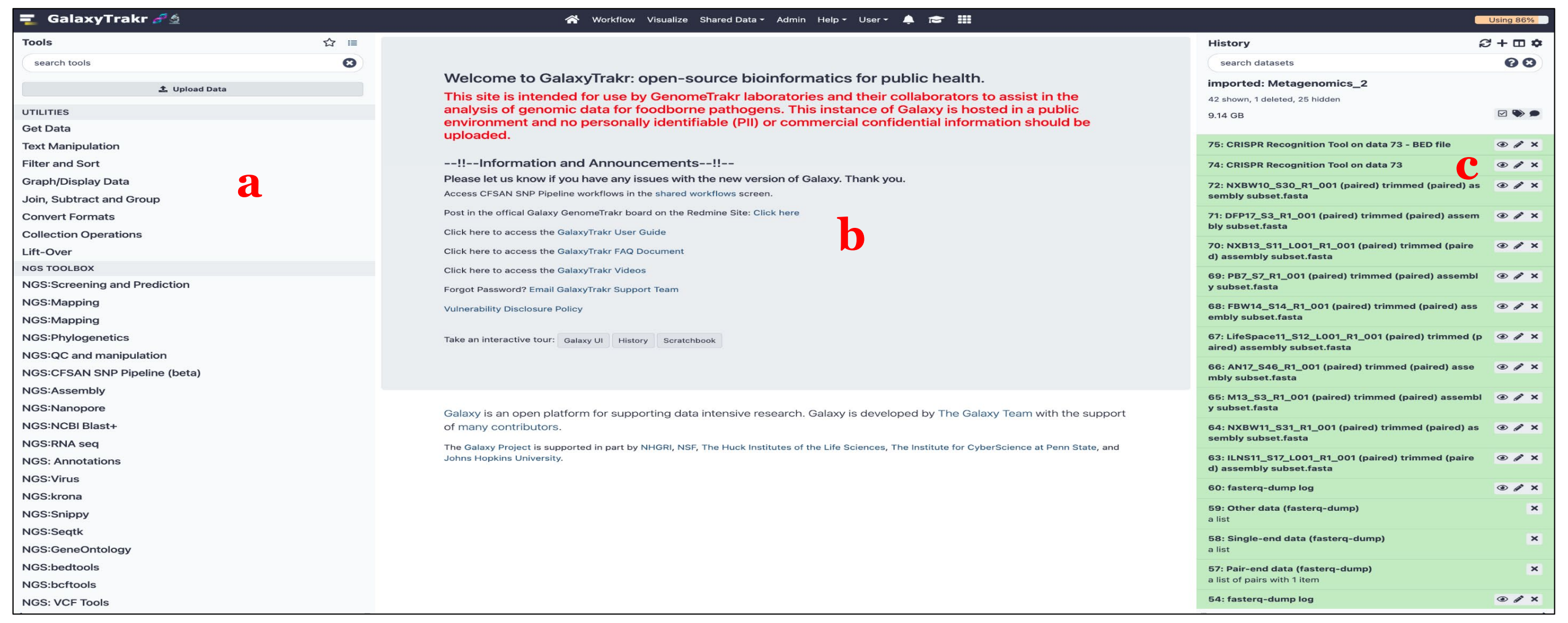
Access CFSAN SNP Pipeline workflows in the shared workflows screen.

Post in the offical Galaxy GenomeTrakr board on the Redmine Site: Click here

Click here to access the GalaxyTrakr User Guide

Click here to access the GalaxyTrakr FAQ Document

Figure 2. GalaxyTrakr user interface that contains three panels: **a** Available tools, **b** Data analysis, and **c** File histories



UTILITIES
Get Data
Text Manipulation
Filter and Sort
Graph/Display Data
Join, Subtract and Group
Convert Formats
Collection Operations
Lift-Over
NGS TOOLBOX
NGS:Screening and Prediction
NGS:Mapping
NGS:Mapping
NGS:Phylogenetics
NGS:QC and manipulation
NGS:CFSAN SNP Pipeline (beta)
NGS:Assembly
NGS:Nanopore
NGS:NCBI Blast+
NGS:RNA seq
NGS: Annotations
NGS:Virus
NGS:krona
NGS:Snippy
NGS:Seqtk
NGS:GeneOntology
NGS:bedtools
NGS:bcftools
NGS: VCF Tools

Welcome to GalaxyTrakr: open-source bioinformatics for public health.

This site is intended for use by GenomeTrakr laboratories and their collaborators to assist in the analysis of genomic data for foodborne pathogens. This instance of Galaxy is hosted in a public environment and no personally identifiable (PII) or commercial confidential information should be uploaded.

--!!--Information and Announcements--!!--

Please let us know if you have any issues with the new version of Galaxy. Thank you.

Access CFSAN SNP Pipeline workflows in the shared workflows screen.

Post in the offical Galaxy GenomeTrakr board on the Redmine Site: Click here

Click here to access the GalaxyTrakr User Guide

Click here to access the GalaxyTrakr FAQ Document

Click here to access the GalaxyTrakr Videos

Forgot Password? Email GalaxyTrakr Support Team

Vulnerability Disclosure Policy

Take an interactive tour: Galaxy UI | History | Scratchbook

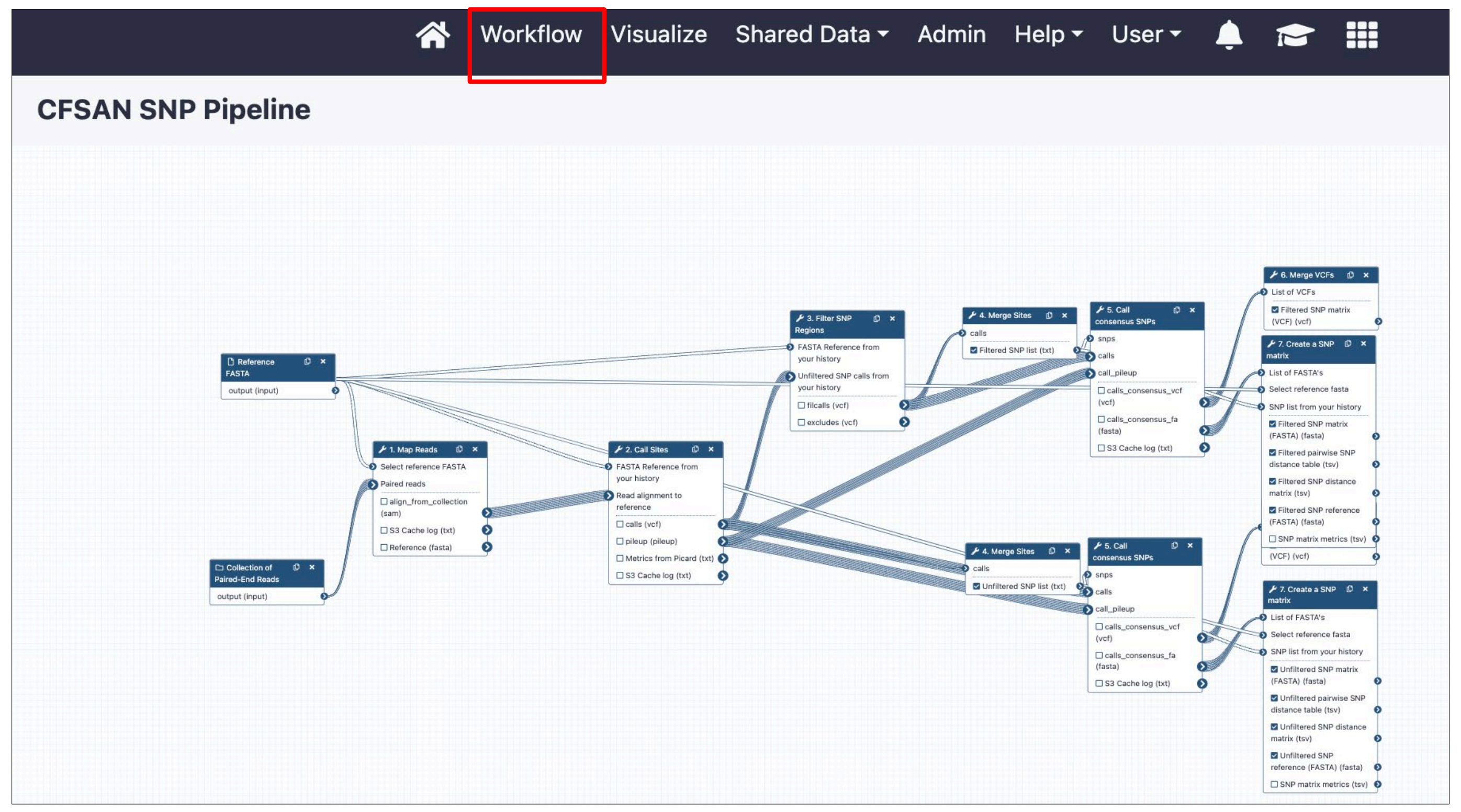
Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

imported: Metagenomics_2
42 shown, 1 deleted, 25 hidden
9.14 GB

75: CRISPR Recognition Tool on data 73 - BED file
74: CRISPR Recognition Tool on data 73
72: NXBW10_S30_R1_001 (paired) trimmed (paired) assembly subset.fasta
71: DFP17_S3_R1_001 (paired) trimmed (paired) assembly subset.fasta
70: NXB13_S11_L001_R1_001 (paired) trimmed (paired) assembly subset.fasta
69: PB7_S7_R1_001 (paired) trimmed (paired) assembly subset.fasta
68: FBW14_S14_R1_001 (paired) trimmed (paired) assembly subset.fasta
67: LifeSpace11_S12_L001_R1_001 (paired) trimmed (paired) assembly subset.fasta
66: AN17_S46_R1_001 (paired) trimmed (paired) assembly subset.fasta
65: M13_S3_R1_001 (paired) trimmed (paired) assembly subset.fasta
64: NXBW11_S31_R1_001 (paired) trimmed (paired) assembly subset.fasta
63: ILNS11_S17_L001_R1_001 (paired) trimmed (paired) assembly subset.fasta
60: fasterq-dump log
59: Other data (fasterq-dump)
a list
58: Single-end data (fasterq-dump)
a list
57: Pair-end data (fasterq-dump)
a list of pairs with 1 item
54: fasterq-dump log

Figure 3. GalaxyTrakr workflows are analysis pipelines orchestrating multiple tools


Workflow
Visualize
Shared Data
Admin
Help
User
CFSAN SNP Pipeline
Reference FASTA
output (input)
Collection of Paired-End Reads
output (input)
1. Map Reads
Select reference FASTA
Paired reads
align_from_collection (sam)
S3 Cache log (txt)
Reference (fasta)
2. Call Sites
FASTA Reference from your history
Read alignment to reference
calls (vcf)
pileup (pileup)
Metrics from Picard (txt)
S3 Cache log (txt)
3. Filter SNP Regions
FASTA Reference from your history
Unfiltered SNP calls from your history
filcalls (vcf)
excludes (vcf)
4. Merge Sites
calls
Filtered SNP list (txt)
4. Merge Sites
calls
Unfiltered SNP list (txt)
5. Call consensus SNPs
snps
calls
call_pileup
calls_consensus_vcf (vcf)
calls_consensus_fa (fasta)
S3 Cache log (txt)
5. Call consensus SNPs
snps
calls
call_pileup
calls_consensus_vcf (vcf)
calls_consensus_fa (fasta)
S3 Cache log (txt)
6. Merge VCFs
List of VCFs
Filtered SNP matrix (VCF) (vcf)
7. Create a SNP matrix
List of FASTA's
Select reference fasta
SNP list from your history
Filtered SNP matrix (FASTA) (fasta)
Filtered pairwise SNP distance table (tsv)
Filtered SNP distance matrix (tsv)
Filtered SNP reference (FASTA) (fasta)
SNP matrix metrics (tsv)
(VCF) (vcf)
7. Create a SNP matrix
List of FASTA's
Select reference fasta
SNP list from your history
Unfiltered SNP matrix (FASTA) (fasta)
Unfiltered pairwise SNP distance table (tsv)
Unfiltered SNP distance matrix (tsv)
Unfiltered SNP reference (FASTA) (fasta)
SNP matrix metrics (tsv)

GalaxyTrakr workflows for QC and characterization for batch analysis of bacterial samples

- **MicroRunQC :** De-Novo Assembly using NCBI's SKESA assembler and followed by MLST profiling
- **QC_Reads :** Calculate basic summary statistics (Q30, length, etc.) using FASTQC
- **ConFindr :** Contamination detection using rMLST (rps genes)
- **MicroRunKraken2 :** Metagenomic analysis using Kraken2
- **CFSAN SNP Pipeline :** Phylogenetic relatedness
- **Metagenomics_Taxonomy_Metaphlan:** Taxonomic profiling of Metagenomics data
- **NARMS AMR workflows:** Species level AMR characterization

Analytical Tools

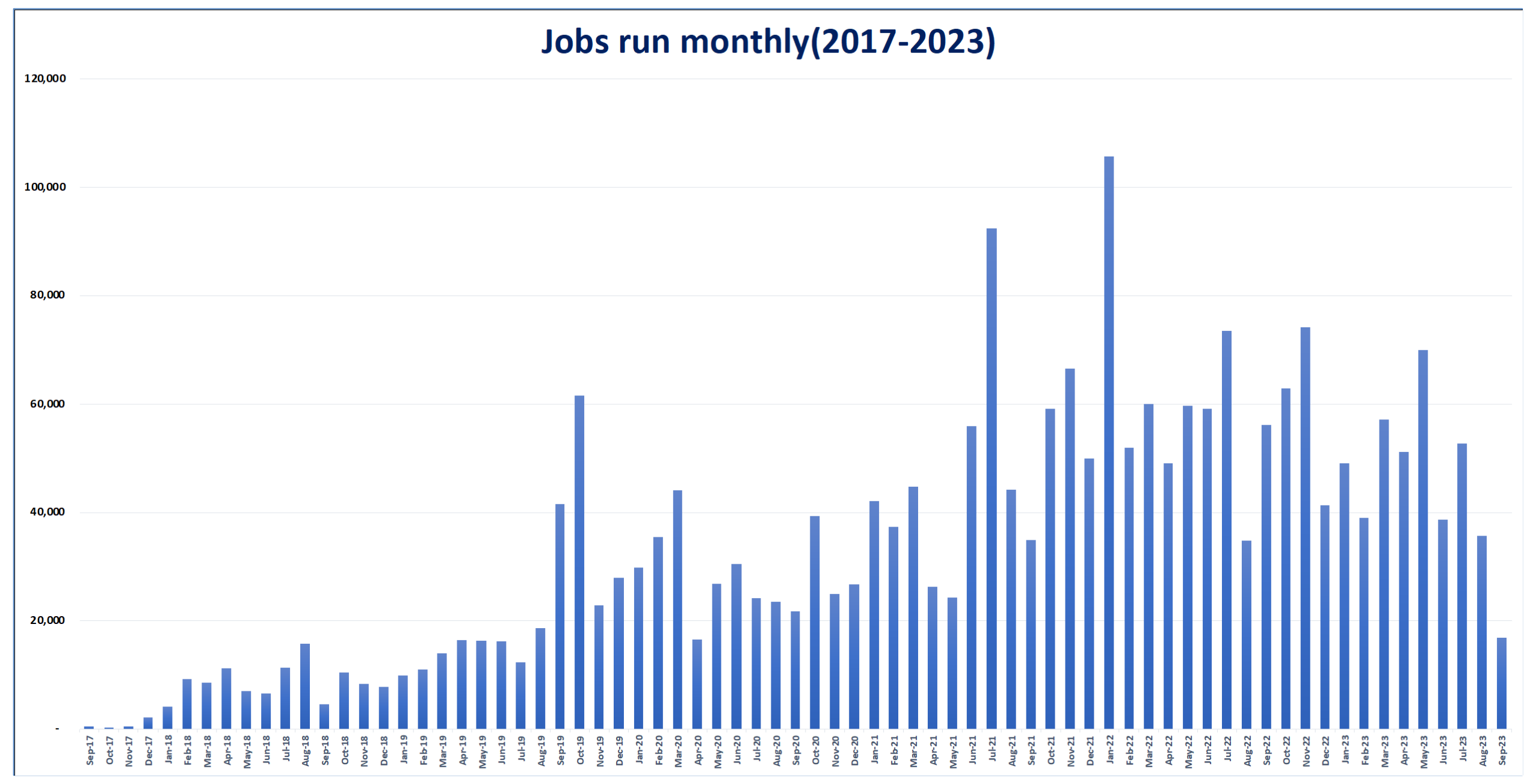
GalaxyTrakr tool panel consists of curated set of analytical tools for NGS data processing

- NGS: NCBI Downloads from SRA,WGS databases.
- NGS:QC and manipulation
- NGS: Screening and Prediction
- NGS: Assembly and Annotations.
- NGS: Reference based Mapping and Variant calling
- NGS: Phylogenetics
- NGS: Nanopore long read sequencing
- Metagenomics: Taxonomic profiling
- Metagenomics: Functional profiling
- Metagenomics: AMR pipeline
- Metagenomics: Statistical and Visualizations

Table 1. The most commonly used analytical tools (top 30) from GalaxyTrakr tool panel, based on the number of jobs run over the years 2017–2023

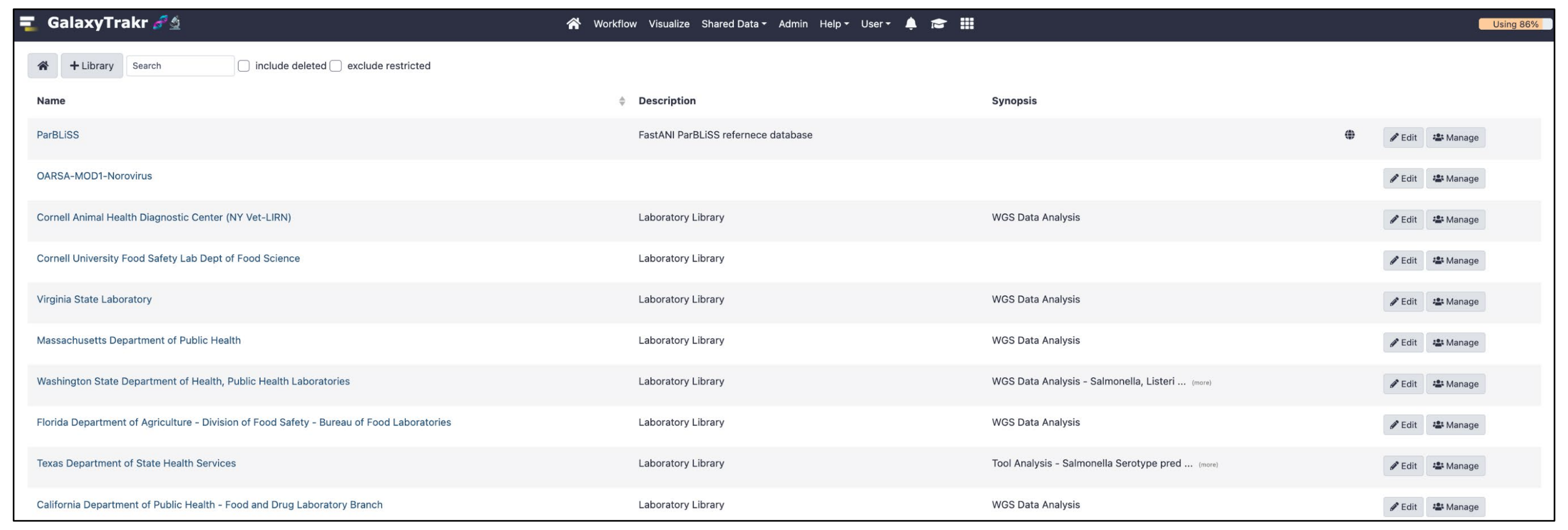
Tool	Description	Number of jobs
CFSAN snp_pipeline	Galaxy implementation of the CFSAN SNP Pipeline	468770
upload1	Data uploads	452663
Seqsero	Salmonella serotype prediction	262079
Trimmomatic	A flexible read trimming tool for Illumina NGS data	153741
Skesamlst	Skesa assembly and MLST	112148
FastQC	Read QC reports using FastQC	97660
Abricate	Mass screening of contigs for antiobiotic resistance genes	59680
Ectyper	in silico serotyping of Escherichia coli species	51632
Kraken2	Taxonomic classification system	49137
Spades	St. Petersburg genome assembler	46001
Skesa	de-novo sequence read assembler for microbial genomes	37171
Amrfinder	NCBI Antimicrobial Resistance Gene Finder	36413
Datamash	Grouping and summarizing tool on tabular data files	26991
Sum_fastqc	summarizes raw FASTQC output	25678
Fastqdump_paired	Downloads a set of paired reads by their accession number	24789
NCBI_blast_plus	Find regions of similarity between biological sequences	22283
Shovill	Faster de novo assembly pipeline based around Spades	21063
Srst2	Short Read Sequencing Typing	18758
Snippy	Rapid haploid variant calling and core genome alignment	17314
Samtools_stats	Generate statistics for BAM dataset	15825
Ecoliserotyper	in silico serotyping of Escherichia coli species	15560
Bbmap_sendsketch	Identifying species using sketch	14262
Microrunqc	QC Metrics for Illumina Bacterial Whole-Genome Sequencing	13831
Quast	Quast (Quality Assessment Tool) evaluates genome assemblies	12493
Mitoprokka	Rapid annotation of bacteria, archaeal, viral and mitochondria genomes	11660
FastANI_db	Fast Whole-Genome Similarity (ANI) Estimation	11278
bowtie2	Map reads against reference genome	10119
Mob_suite	Clustering, reconstruction and typing of plasmids from draft assemblies	9026
Prokka	Prokaryotic genome annotation	8400

Figure 4. Number of monthly jobs run on GalaxyTrakr through September 2023



Sharing and Collaboration

Figure 5. Sharing and Collaboration in GalaxyTrakr via Shared Data libraries



☐ include deleted ☐ exclude restricted

Name	Description	Synopsis
ParBLiSS	FastANI ParBLiSS refernece database	
OARSA-MOD1-Norovirus		
Cornell Animal Health Diagnostic Center (NY Vet-LIRN)	Laboratory Library	WGS Data Analysis
Cornell University Food Safety Lab Dept of Food Science	Laboratory Library	
Virginia State Laboratory	Laboratory Library	WGS Data Analysis
Massachusetts Department of Public Health	Laboratory Library	WGS Data Analysis
Washington State Department of Health, Public Health Laboratories	Laboratory Library	WGS Data Analysis - Salmonella, Listeri ... (more)
Florida Department of Agriculture - Division of Food Safety - Bureau of Food Laboratories	Laboratory Library	WGS Data Analysis
Texas Department of State Health Services	Laboratory Library	Tool Analysis - Salmonella Serotype pred ... (more)
California Department of Public Health - Food and Drug Laboratory Branch	Laboratory Library	WGS Data Analysis

To illustrate how laboratories are making use of this resource, we describe how six institutions use GalaxyTrakr to quickly analyze and review their data.

- *Wadsworth Center, New York State Department of Health:* QC and serotyping of *Salmonella* isolates.
- *Ohio Department of Agriculture/Animal Disease Diagnostic Laboratory (ADDL).* Antibiotic susceptibility and *Salmonella* serotyping.
- *Virginia Division of Consolidated Laboratory Services (DCLS).* Bioinformatics training to state public health labs.
- *State Public Health Bioinformatics (StaPH-B).* Bioinformatics training to state public health labs.
- *James Madison University*. an undergraduate advanced microbiology course for students to study the genomics of *Salmonella enterica* isolated from environmental sources.
- *Laboratorio de Microbiología y Probióticos, Instituto de Nutrición y Tecnología de los Alimentos (INTA), Universidad de Chile, Santiago, Chile.* To characterize and understanding of Shiga toxin-producing *Escherichia coli* (STEC) and *Salmonella* isolates.

Conclusion

- GalaxyTrakr advances food safety by providing reliable and harmonized WGS analyses for public health laboratories and promoting collaboration across laboratories with differing resources.
- User friendly bioinformatics resource with curated set of analytical tools and workflows, helps scientists to analyze their WGS data.
- Anticipated enhancements to this resource will include workflows for additional foodborne pathogens, viruses, and parasites, as well as new tools and services.

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). https://doi.org/10.1186/s12864-021-07405-8

FDA Mission Relevance

Creating and maintaining an effective pathogen surveillance system is essential to global public health. GalaxyTrakr facilitates surveillance by increasing the number of laboratories able to participate in surveillance and outbreak investigations and helping those with limited budgets analyze their data by removing expense and expertise barriers.