



GenomeTrakr - 2018

SCALABLE PUBLIC HEALTH BIOINFORMATICS IN THE CLOUD

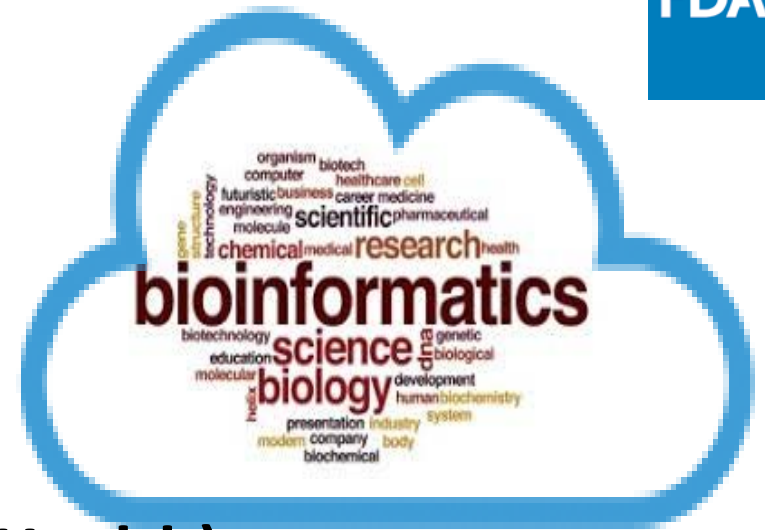
Agenda

- **Part 1 (Scalable Cloud Solution)**

- GalaxyTrakr Overview
- Challenge / Solution / Benefits
- Under the Hood

- **Part 2 (Curated Tools for Public Health)**

- JIFSAN
- Galaxy / UseGalaxy vs. GalaxyTrakr
- GalaxyTrakr Target Audience
- GalaxyTrakr



The background image shows a glowing blue globe with a grid of data points, set against a dark background with faint circuit-like patterns.

Scalable public health bioinformatics in the cloud

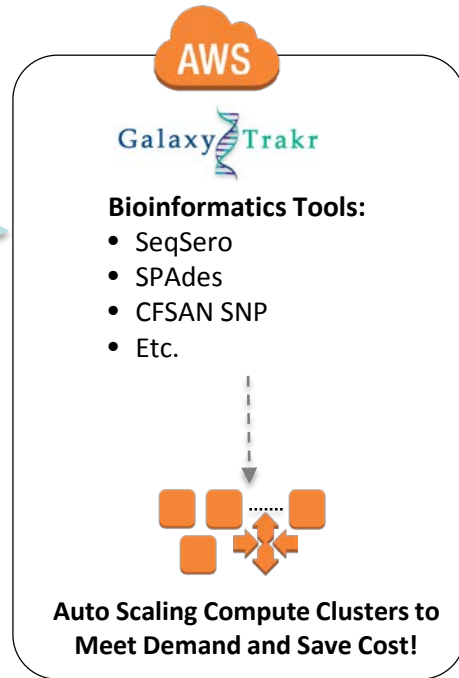
- part 1 -

Scalable Cloud Computing

Jimmy Sanders • FDA-CFSAN

GalaxyTrakr Overview

U.S. East 1 – Northern VA



External Users



<https://galaxytrakr.org>

Data Upload Options:

- SFTP
- Web Upload
- Download via SRA/ENA accession from NCBI

GalaxyTrakr Overview

- Delivers a public instance of Galaxy (galaxyproject.org) in AWS with implemented tools and workflows for analysis of foodborne bacteria
- Enhances collaboration between FDA-CFSAN and the GenomeTrakr partners
- Leverages an elastically scalable compute cluster in AWS
- Provides basic WGS analysis tools to labs within FDA and external to FDA

Current Statistics

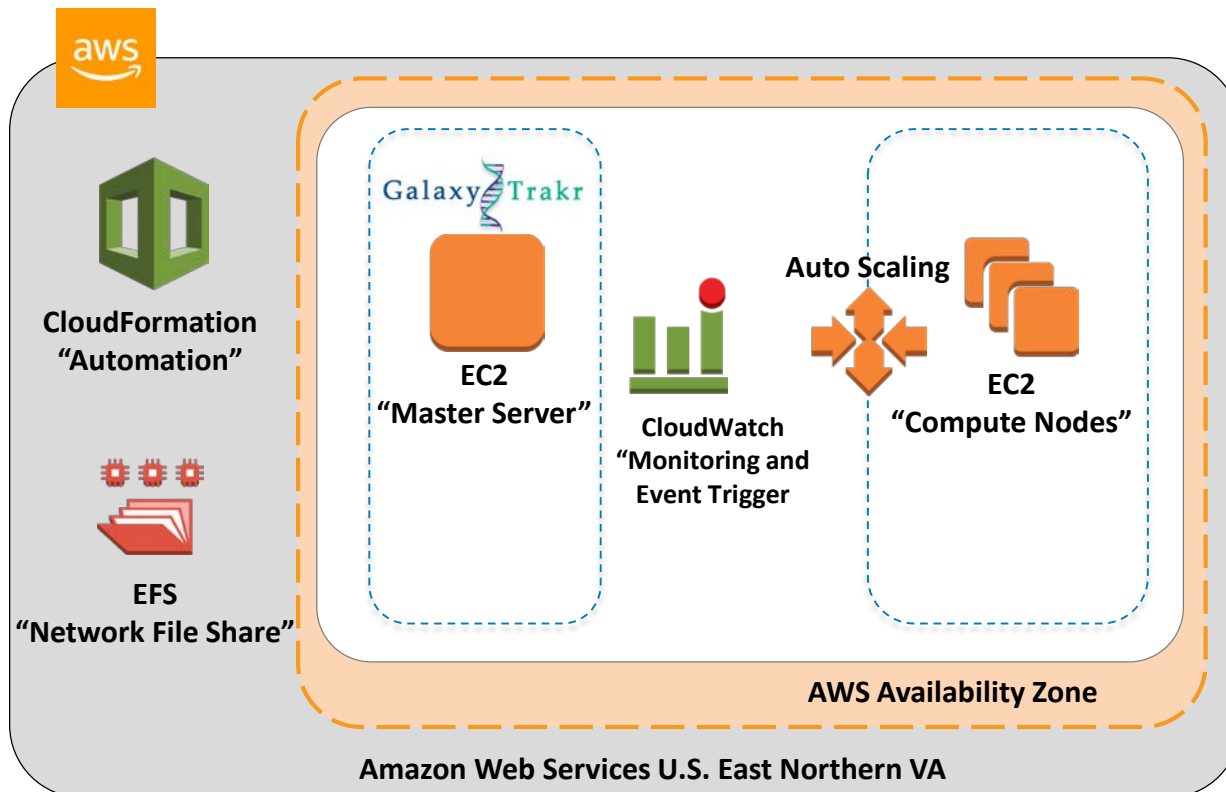
- Over 80,000 jobs processed in just over one year
- 30+ tools and workflows available
- 240 registered users
- 53 connected laboratories:
 - Public and State Health Labs
 - Academic Institutions
 - International Health Laboratories (Italy, Chile, Dublin, South Africa)
 - Other Federal Organizations (CDC, USDA)
 - Labs within FDA

Challenge

- FDA-CFSAN needed to meet the FSMA requirement for the FDA to collaborate with state and local food safety laboratories
- The network of laboratories routinely sequences more than 1000 isolates each month for isolates originating from food, environmental, and clinical sources
- FDA-CFSAN's capacity for providing bioinformatics support to these laboratories has not kept pace with the large volume of data being generated
- Non-Federal partners can't access FDA-hosted resources (tools, data, compute power)
- Required an environment that could scale on demand

Solution

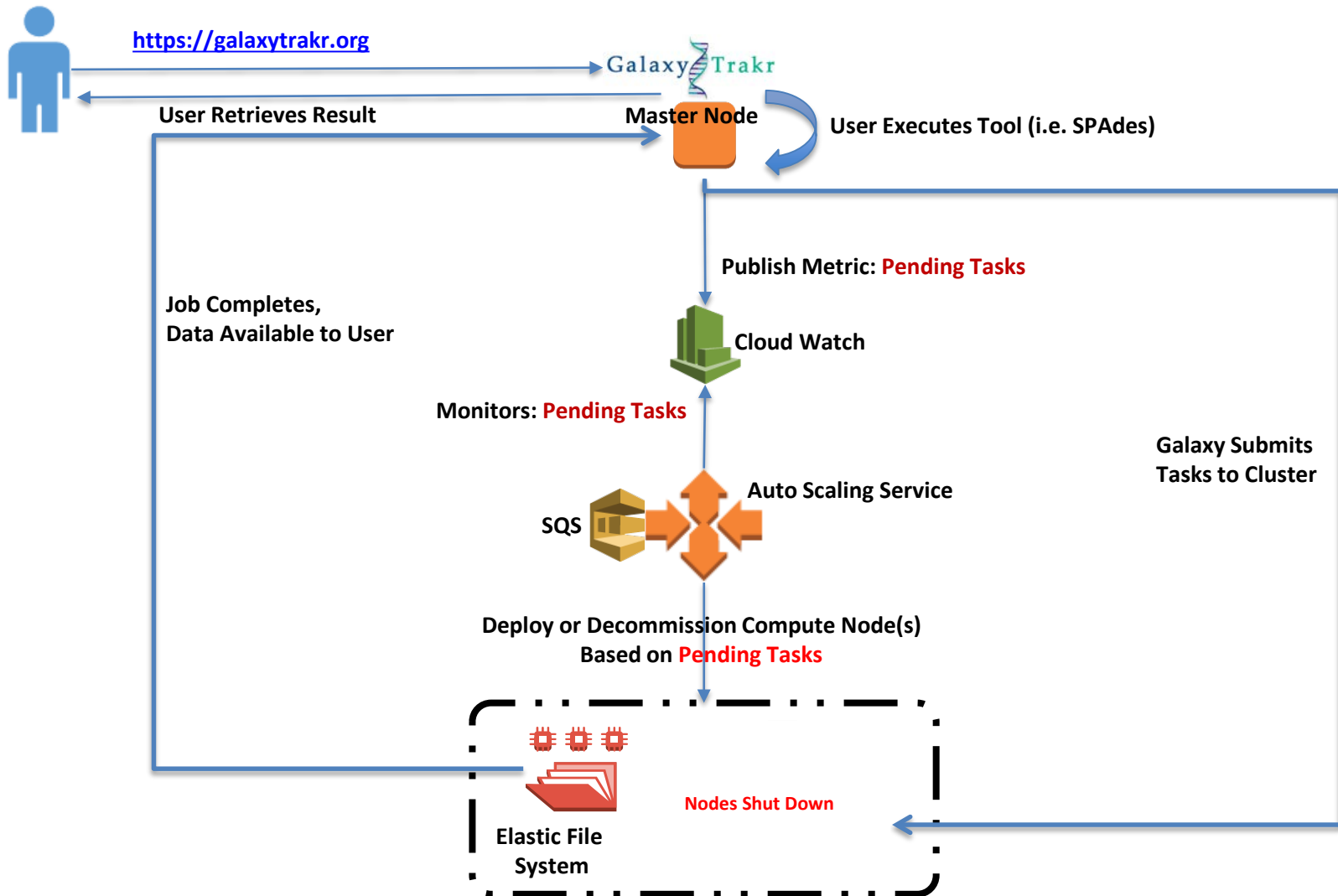
- Galaxy (<https://usegalaxy.org/>)
- Amazon Web Services (AWS) (<https://aws.amazon.com/>)
- CloudFormation Cluster (<https://cfncluster.readthedocs.io/en/latest/>)



Benefits

- Galaxy
 - Collaborative, open-source web-based platform for bioinformatics
 - Large community support
- Amazon Web Services
 - Ability to deploy a scalable solution and present to public users
 - Infrastructure services available to run a platform like Galaxy
 - Pay as you go
- CloudFormation Cluster
 - Cost savings through auto-scaling compute cluster to process jobs submitted from Galaxy (some days < 100 jobs, some days > 1000 jobs)
 - Easily modified to meet various bioinformatics computation requirements

Under the Hood





Scalable public health bioinformatics in the cloud

- part 2 -

Curated tools for public health

Justin Payne • FDA-CFSAN



JIFSAN Bioinformatics Training

The Struggle is Real

Working with the Shell

```
$ cp file_1
```

```
$ mv file_1
```

```
$ rm file_1
```

Bash shell

```
$ a_tool -
```

Most commands

```
$ man a_
```

Most commands

The dollar sign traditionally stands in for your prompt and lets you know that what follows is an example command.

Don't type the dollar sign.

Start with the next word, which is the program you're asking the shell to run.

One Visual Conventions

These are hyphens, not dashes. The key is between '0' and '='. Microsoft products "helpfully" convert hyphens to dashes, so be careful cutting and pasting CLI commands out of Word/PowerPoint.

It's common to use example paths; if they seem 'unrealistic', you're encouraged to substitute the file location of your choice. Nonexistent example commands will be in italics, to show you things that apply to most tools.

```
$ velvet write_paper -t -f -o path/to/my/papers/ \
-i some_other_options \
-p /etc/conf/paper.conf
```

There really is a directory called 'etc' on Linux and Mac systems, this isn't a joke. When you see a realistic path, type it in as shown. Be precise and remember to use tab-completion to avoid misspellings.

Press 'return' at the end of the command; that's what makes it go.

Sometimes these commands get longer than the width of a PowerPoint slide; we'll use the backslash 'line continuation' syntax when that's the case. If you see this, you can ignore the backslash and continue typing the command all on a single line, or you can type it in as shown with line breaks. Your prompt will change to > to show you that you're continuing a line.

JIFSAN Bioinformatics Training

...But the Results Were Less So



- JIFSAN bioinformatics training feedback:
- “is this something you expect us to do?”
- “I didn’t know my Mac had a command line”
- “what should we buy?”

- Status quo might be easy to improve on.

Galaxy / GALAXY GENOME TRAKR
Using 46%

Tools

[Get Data](#)

NGS TOOLBOX

[NGS: QC and manipulation](#)

[NGS: Assembly](#)

[NGS: Screening and Prediction](#)

[NGS: Mapping](#)

[NGS: CFSAN SNP Pipeline \(beta\)](#)

[NGS: Phylogenetics](#)

UTILITIES

[Text Manipulation](#)

[Join, Subtract and Group](#)

[Graph / Display Data](#)

[Convert Formats](#)

[Filter and Sort](#)

Workflows

- [All workflows](#)
- [Batch FASTQ+SeqSero](#)
- [Assemble and QUAST](#)
- [CFSAN SNP Pipeline](#)
- [CFSAN SNP Pipeline w/ Assembly](#)
- [ABRicate Summary](#)
- [CFSAN SNP Pipeline w/ SRA Download and Assembly](#)
- [CFSAN SNP Pipeline on FASTA](#)

QUAST

Quality Assessment Tool for Genome Assemblies by [CAB](#)

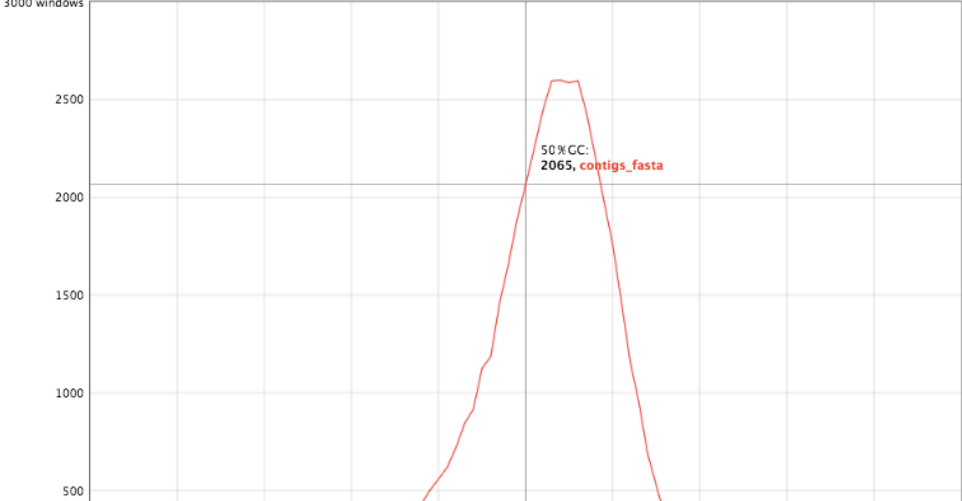
29 August 2018, Wednesday, 14:54:36

[View in Icarus contig browser](#)

All statistics are based on contigs of size ≥ 500 bp, unless otherwise noted (e.g., "# contigs ($>= 0$ bp)" and "Total length ($>= 0$ bp)" include all contigs).

| Statistics without reference | | contigs_fasta |
|------------------------------|-----------|---------------|
| # contigs | 99 | |
| # contigs ($>= 0$ bp) | 113 | |
| # contigs ($>= 1000$ bp) | 88 | |
| Largest contig | 217 425 | |
| Total length | 4 556 982 | |
| Total length ($>= 0$ bp) | 4 562 313 | |
| Total length ($>= 1000$ bp) | 4 548 428 | |
| N50 | 88 543 | |
| N75 | 50 223 | |
| L50 | 18 | |
| L75 | 36 | |
| GC (%) | 52.28 | |
| Mismatches | | |
| # N's | 0 | |
| # N's per 100 kbp | 0 | |

Plots: [Cumulative length](#) [Nx](#) [GC content](#) Normal / logarithmic scale



50% GC:
2065, contigs_fasta

contigs_fasta by contigs

History

Unnamed history

21 shown

235.5 MB

- 21: Quast report.html
361.9 KB
format: html, database: 2
- 20: Icarus Contig size viewer
- 19: Quast report.tex
- 18: Quast report.tsv
- 17: Quast report.txt
- 16: contigs.fasta
- 15: SRR7521740 (sra-pileup)
- 14: SRR7521740 (fastq-dump)
- 13: Shovill on collection 1: Contig Graph
a list with 5 Items
- 12: Shovill on collection 1: Contigs
a list with 5 Items
- 11: Shovill on data 683 and data 682: Contig Graph
- 10: Shovill on data 683 and data 682: Contigs
- 9: Shovill on data 683 and data 682: Contig Graph

Galaxy (workflows)



Galaxy / GALAXY GENOME TRAKR Analyze Data Workflow Visualize Shared Data Admin Help User Using 46%

Tools Workflow Canvas | CFSAN SNP Pipeline (Lite)

search tools

Inputs
Get Data

NGS TOOLBOX
NGS: QC and manipulation
NGS: Assembly
NGS: Screening and Prediction
NGS: Mapping
NGS: CFSAN SNP Pipeline (beta)
NGS: Phylogenetics

UTILITIES
Text Manipulation
Join, Subtract and Group
Graph/Display Data
Convert Formats
Filter and Sort

Data Manager Tools

Workflows

Reference FASTA
output

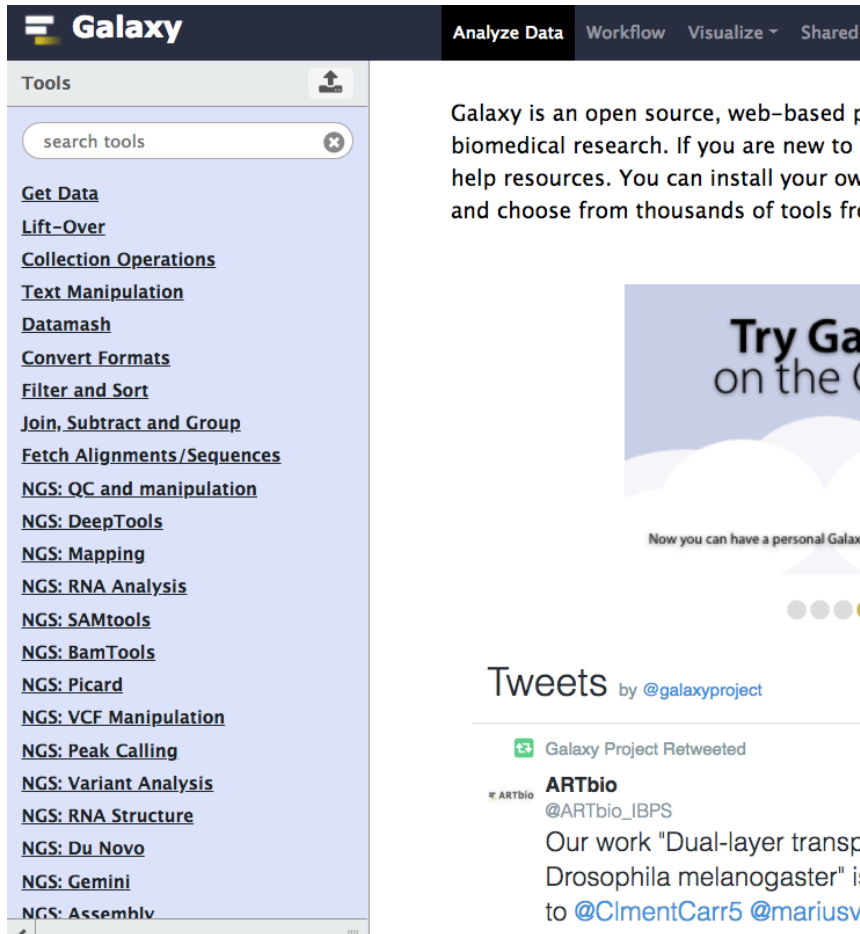
1. Map Reads
Select reference FASTA
Paired reads
align_from_collection (sam)
align_from_history (sam)
cache_log (txt)
ref_out (fasta)

2. Call Sites
FASTA Reference from your history
Read alignment to reference
calls (vcf)
pileup (pileup)
metrics (txt)
cache_log (txt)

Collection of Paired-End Reads
Reads
output

Edit Workflow Attributes
Name: CFSAN SNP Pipeline (Lite)
Tags:
Annotation / Notes: Describe or add notes to workflow. Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

UseGalaxy vs GalaxyTrakr



Galaxy Analyze Data Workflow Visualize Shared Data

Tools

- [Get Data](#)
- [Lift-Over](#)
- [Collection Operations](#)
- [Text Manipulation](#)
- [Datamash](#)
- [Convert Formats](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Fetch Alignments/Sequences](#)
- [NGS: QC and manipulation](#)
- [NGS: DeepTools](#)
- [NGS: Mapping](#)
- [NGS: RNA Analysis](#)
- [NGS: SAMtools](#)
- [NGS: BamTools](#)
- [NGS: Picard](#)
- [NGS: VCF Manipulation](#)
- [NGS: Peak Calling](#)
- [NGS: Variant Analysis](#)
- [NGS: RNA Structure](#)
- [NGS: Du Novo](#)
- [NGS: Gemini](#)
- [NGS: Assembly](#)

Galaxy is an open source, web-based platform for sharing and analyzing data from biomedical research. If you are new to Galaxy, we have a number of help resources. You can install your own Galaxy instance and choose from thousands of tools from the Galaxy Toolshed.

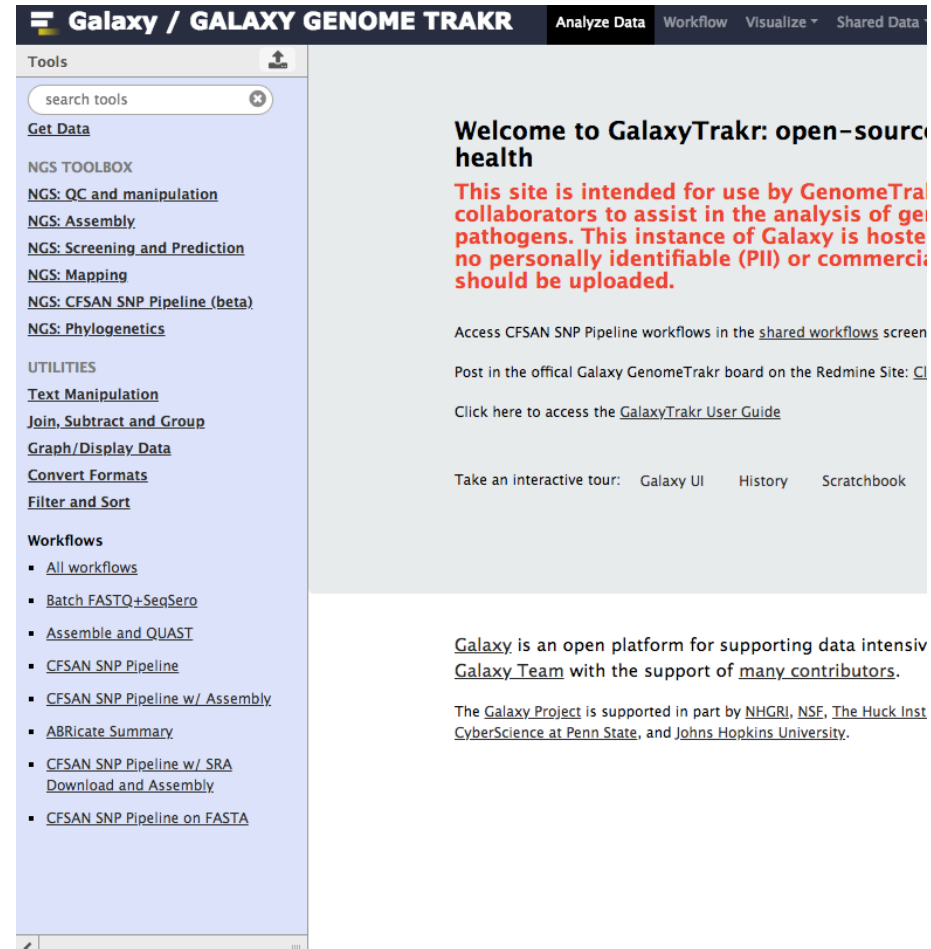
Try Galaxy on the Cloud

Now you can have a personal Galaxy instance.

Tweets by @galaxyproject

Galaxy Project Retweeted

ARTbio @ARTbio_IBPS
Our work "Dual-layer transcriptome analysis in *Drosophila melanogaster*" is available on Galaxy. Contact us to @ClmentCarr5 @mariusv



Galaxy / GALAXY GENOME TRAKR Analyze Data Workflow Visualize Shared Data

Tools

- [Get Data](#)
- NGS TOOLBOX**
 - [NGS: QC and manipulation](#)
 - [NGS: Assembly](#)
 - [NGS: Screening and Prediction](#)
 - [NGS: Mapping](#)
 - [NGS: CFSAN SNP Pipeline \(beta\)](#)
 - [NGS: Phylogenetics](#)
- UTILITIES**
 - [Text Manipulation](#)
 - [Join, Subtract and Group](#)
 - [Graph/Display Data](#)
 - [Convert Formats](#)
 - [Filter and Sort](#)
- Workflows**
 - [All workflows](#)
 - [Batch FASTQ+SeqSero](#)
 - [Assemble and QCAST](#)
 - [CFSAN SNP Pipeline](#)
 - [CFSAN SNP Pipeline w/ Assembly](#)
 - [ABRicate Summary](#)
 - [CFSAN SNP Pipeline w/ SRA Download and Assembly](#)
 - [CFSAN SNP Pipeline on FASTA](#)

Welcome to GalaxyTrakr: open-source health

This site is intended for use by GenomeTrakr collaborators to assist in the analysis of genomic pathogens. This instance of Galaxy is hosted on a cloud platform and no personally identifiable (PII) or commercial data should be uploaded.

Access CFSAN SNP Pipeline workflows in the [shared workflows](#) screen.

Post in the official Galaxy GenomeTrakr board on the Redmine Site: [Click here](#)

Click here to access the [GalaxyTrakr User Guide](#)

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

[Galaxy](#) is an open platform for supporting data intensive research. We are part of the [Galaxy Team](#) with the support of [many contributors](#).

The [Galaxy Project](#) is supported in part by [NHGRI](#), [NSE](#), [The Huck Institute of Life Sciences](#) at Penn State, and [Johns Hopkins University](#).

GalaxyTrakr

Who's the Intended User?



1. "My web browser? Google, I think"
2. "Is the internet down? I can't get to the Facebook"
3. "Oh, there's an Excel formula for that."
4. "You mean you *didn't* change your router's default password?"
5. "Where's Terminal on this thing?"
6. "There's a funny easter egg in echo's man page."
7. "Arch really screams on my dev box since I re-compiled the kernel."

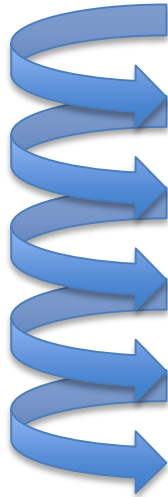
1. "My web browser? Google, I think"
2. "Is the internet down? I can't get to the Facebook"
3. "Oh, there's an Excel formula for that."
4. "You mean you *didn't* change your router's default password?"
5. "Where's Terminal on this thing?"
6. "There's a funny easter egg in echo's man page." ***A scurrilous falsehood***
7. "Arch really screams on my dev box since I re-compiled the kernel."

GalaxyTrakr

Curation and Layout Philosophy

- “One (or two) ways to do things”
- “Obvious path forward”

- “Dogfooding”
 - SNP-Pipeline
 - SeqSero
 - ECTyper

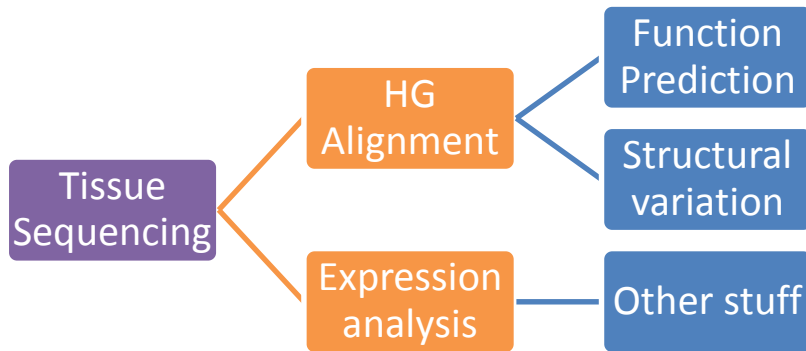


GalaxyTrakr

Expansive vs. Reductive Analyses

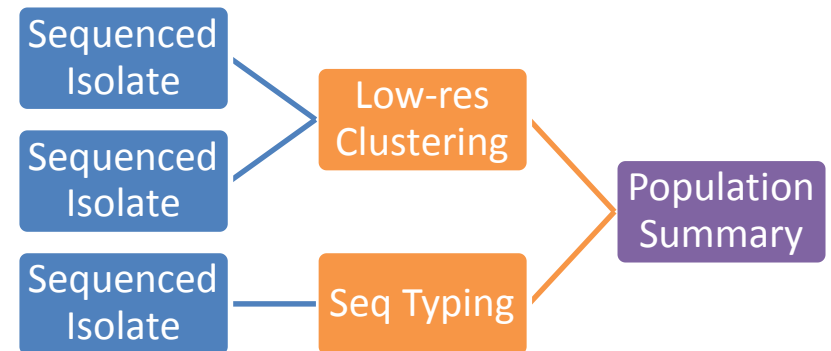


“Expansive” analysis

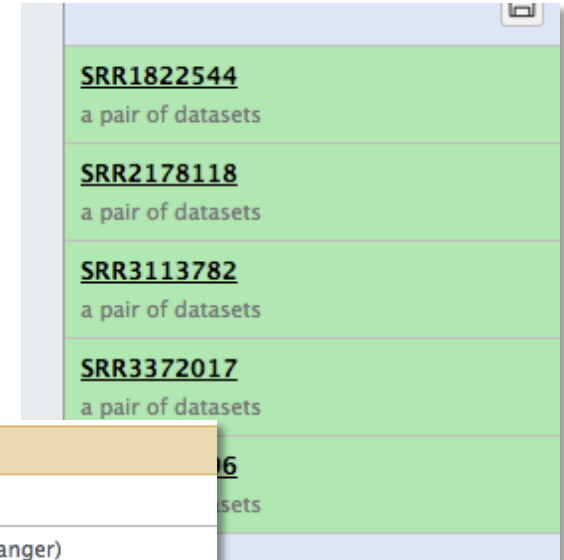


VS

“Reductive” analysis



- Large-scale sample handling
- Tool support for complex collections



seqsero_v2 (Galaxy Version 2.0)

FASTQ paired end read 1

FASTQ paired end read 2

Number of threads

Algorithms for BWA mapping?

Usage: SeqSero2.py

Q&A

