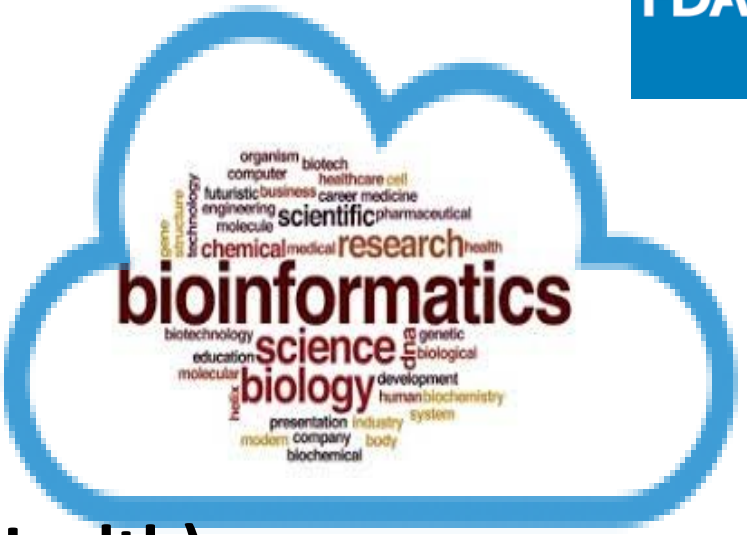
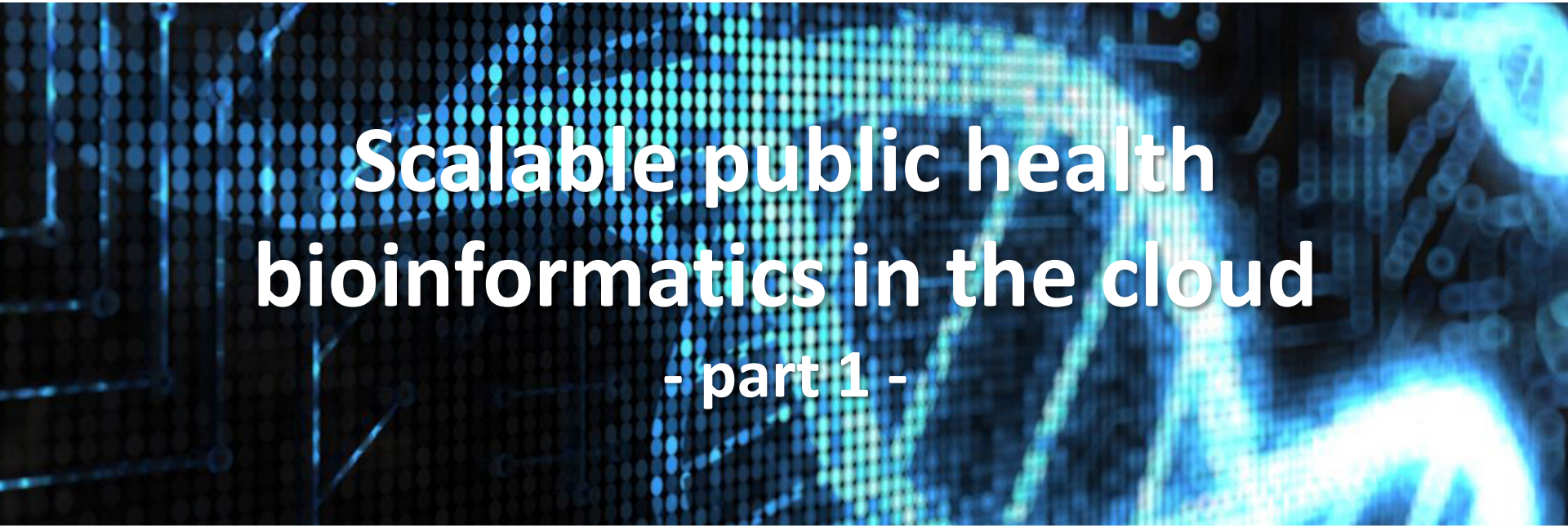
GenomeTrakr - 2018

SCALABLE PUBLIC HEALTH BIOINFORMATICS IN THE CLOUD

Agenda



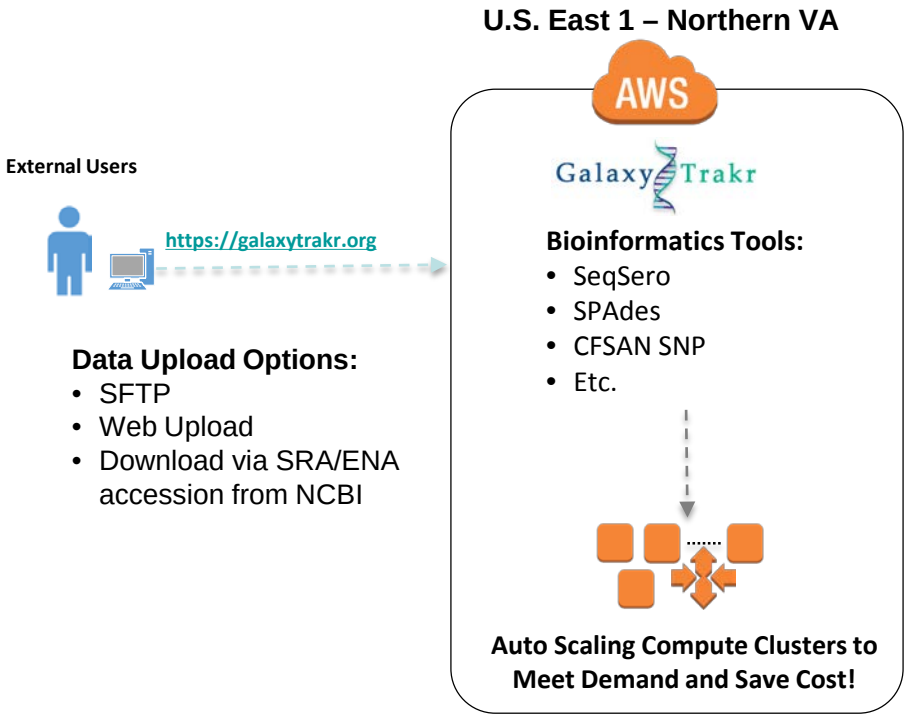
- **Part 1 (Scalable Cloud Solution)**
  - GalaxyTrakr Overview
  - Challenge / Solution / Benefits
  - Under the Hood
- **Part 2 (Curated Tools for Public Health)**
  - JIFSAN
  - Galaxy / UseGalaxy vs. GalaxyTrakr
  - GalaxyTrakr Target Audience
  - GalaxyTrakr



Scalable Cloud Computing

Jimmy Sanders • FDA-CFSAN

GalaxyTrakr Overview



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

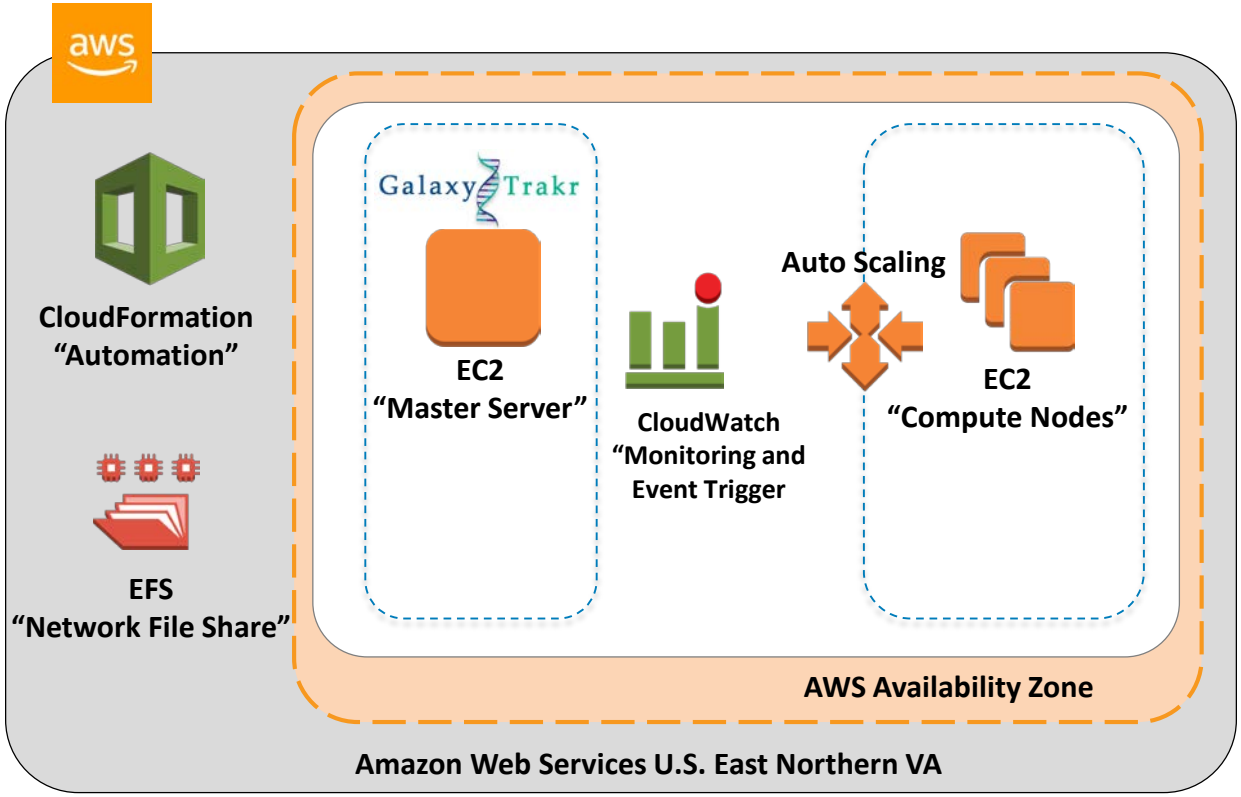
- Over 80,000 jobs processed in just over one year
- 30+ tools and workflows available
- 240 registered users
- 53 connected laboratories:
  - Public and State Health Labs
  - Academic Institutions
  - International Health Laboratories (Italy, Chile, Dublin, South Africa)
  - Other Federal Organizations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

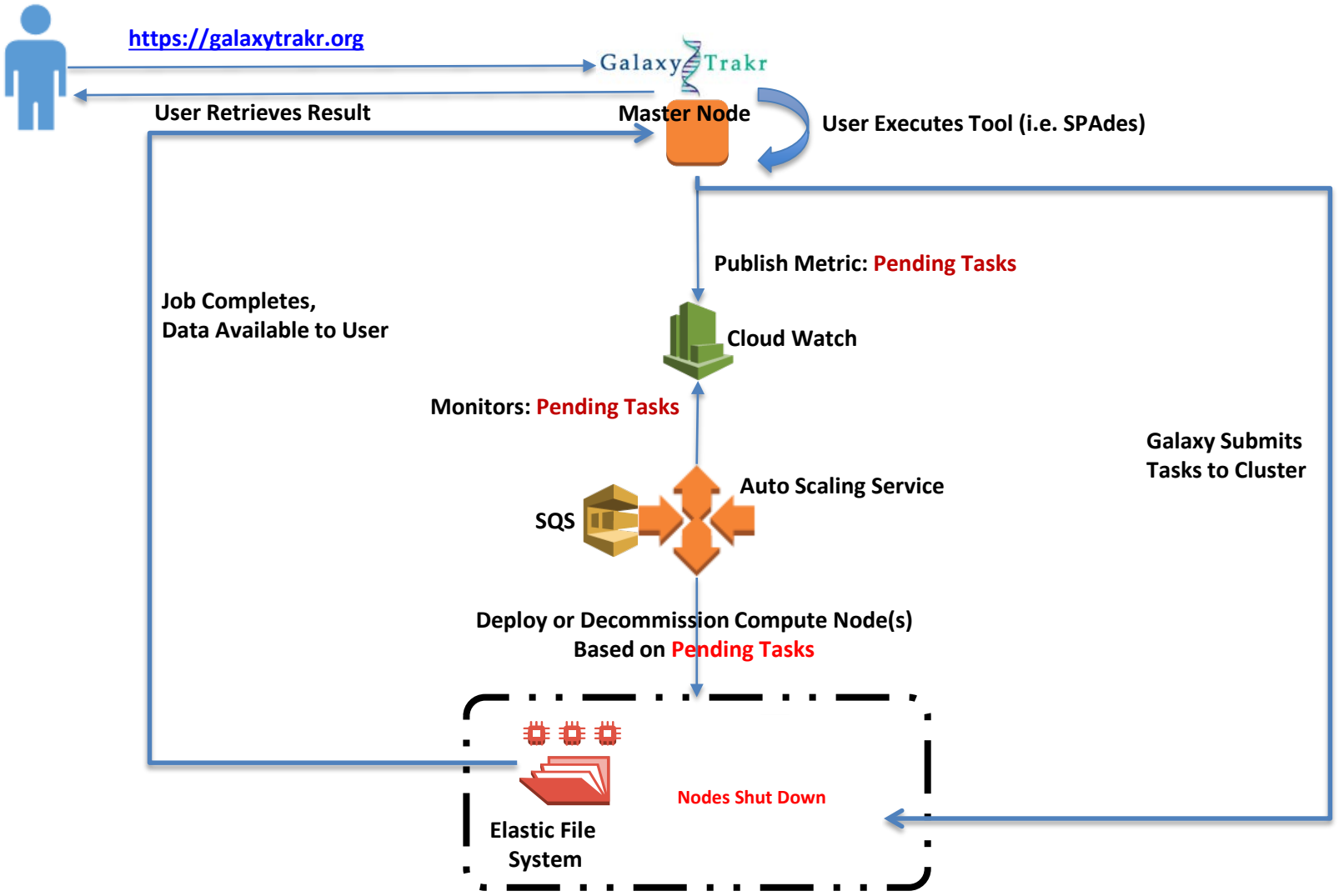
- Galaxy (https://usegalaxy.org/)
- Amazon Web Services (AWS) (https://aws.amazon.com/)
- CloudFormation Cluster (https://cfncluster.readthedocs.io/en/latest/)

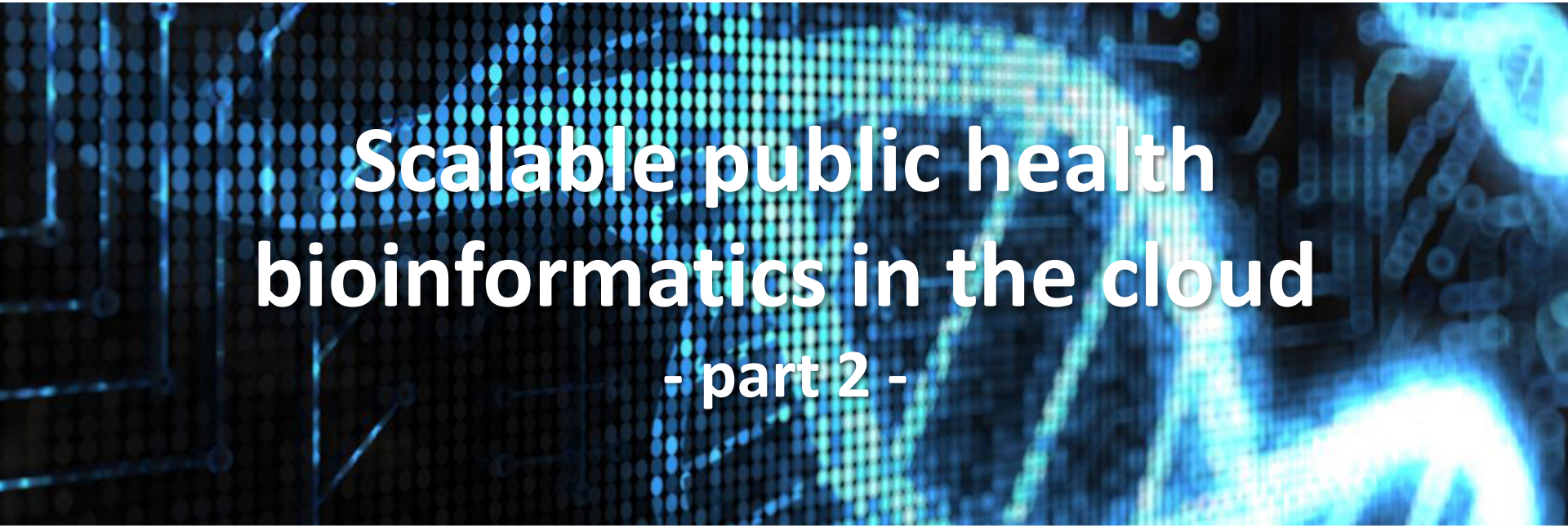


Benefits

- 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





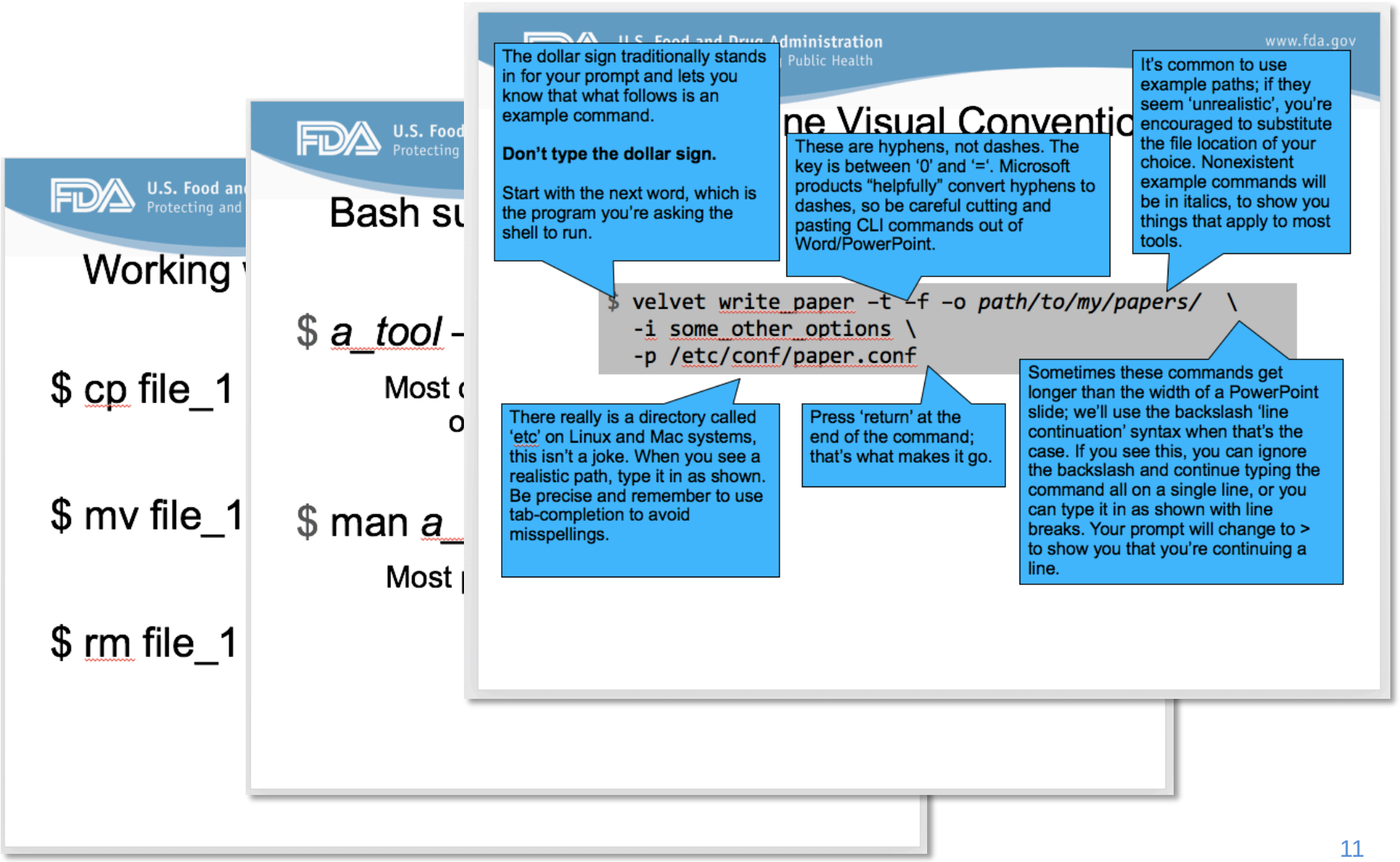
Curated tools for public health

Justin Payne • FDA-CFSAN


JOINT INSTITUTE for FOOD SAFETY
AND APPLIED NUTRITION
UM
FDA

JIFSAN Bioinformatics Training

The Struggle is Real



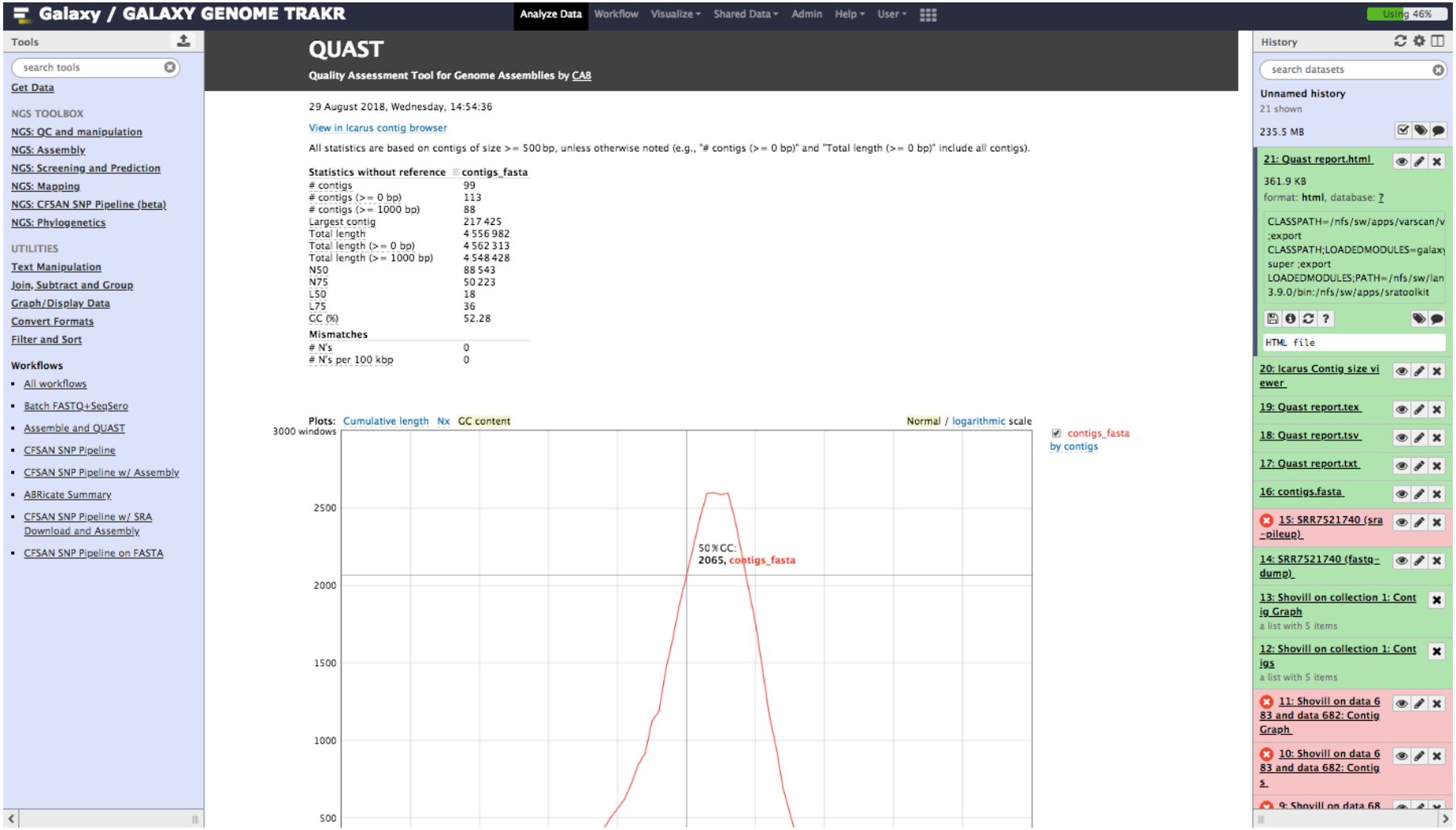
JIFSAN Bioinformatics Training
…But the Results Were Less So

- 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?”

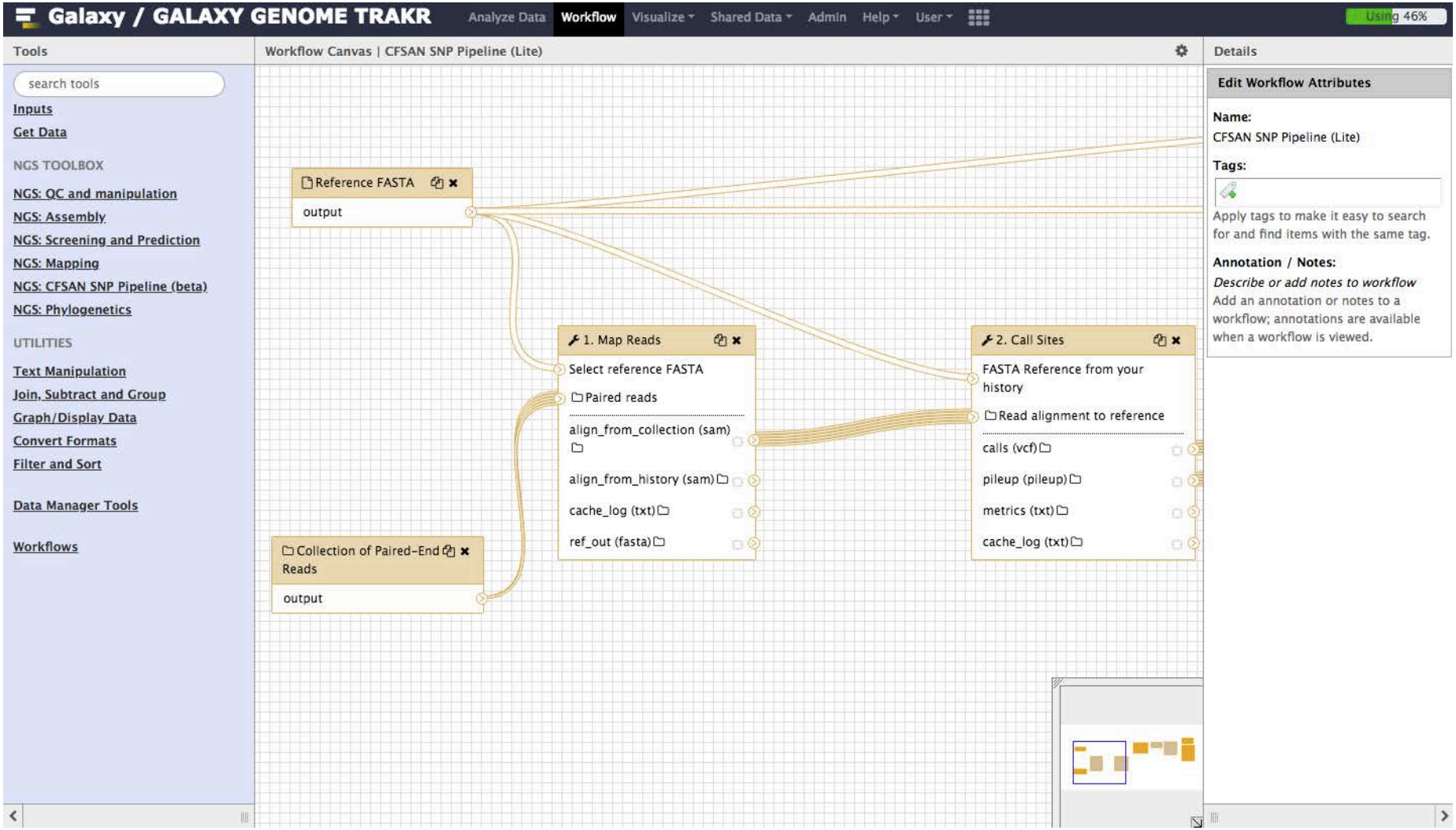
- Status quo might be easy to improve on.

Galaxy

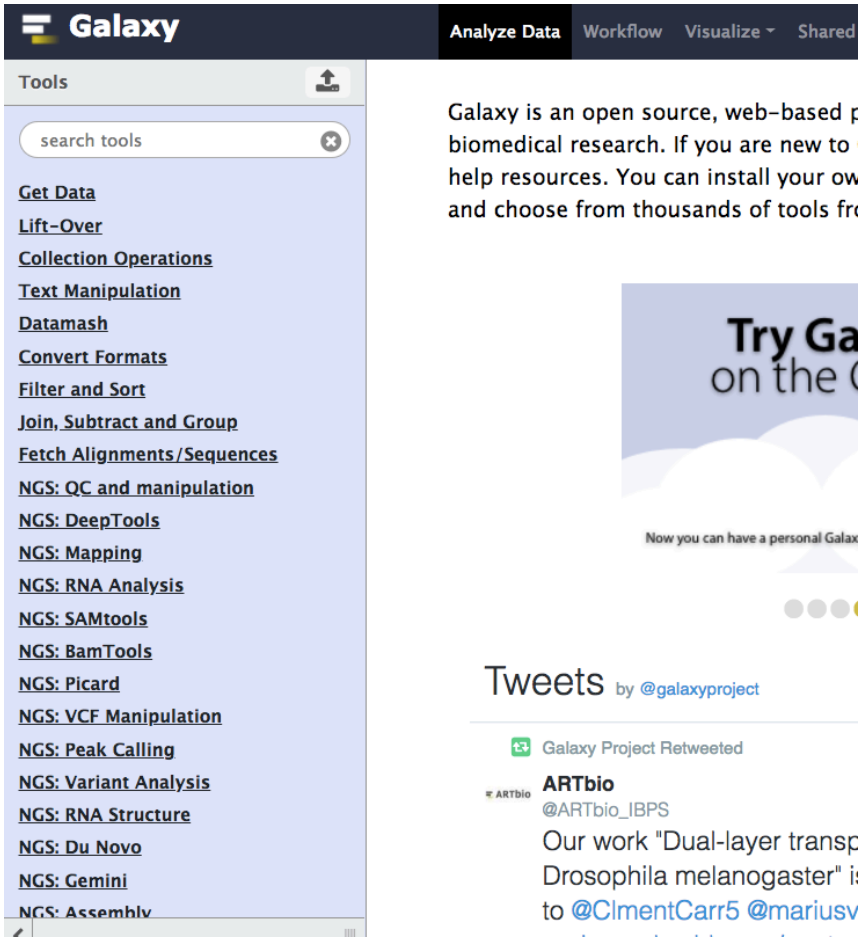
Blankenberg et al. Curr. Protoc. Mol. Biol. **89:19.10.1-19.10.21. (2010)**

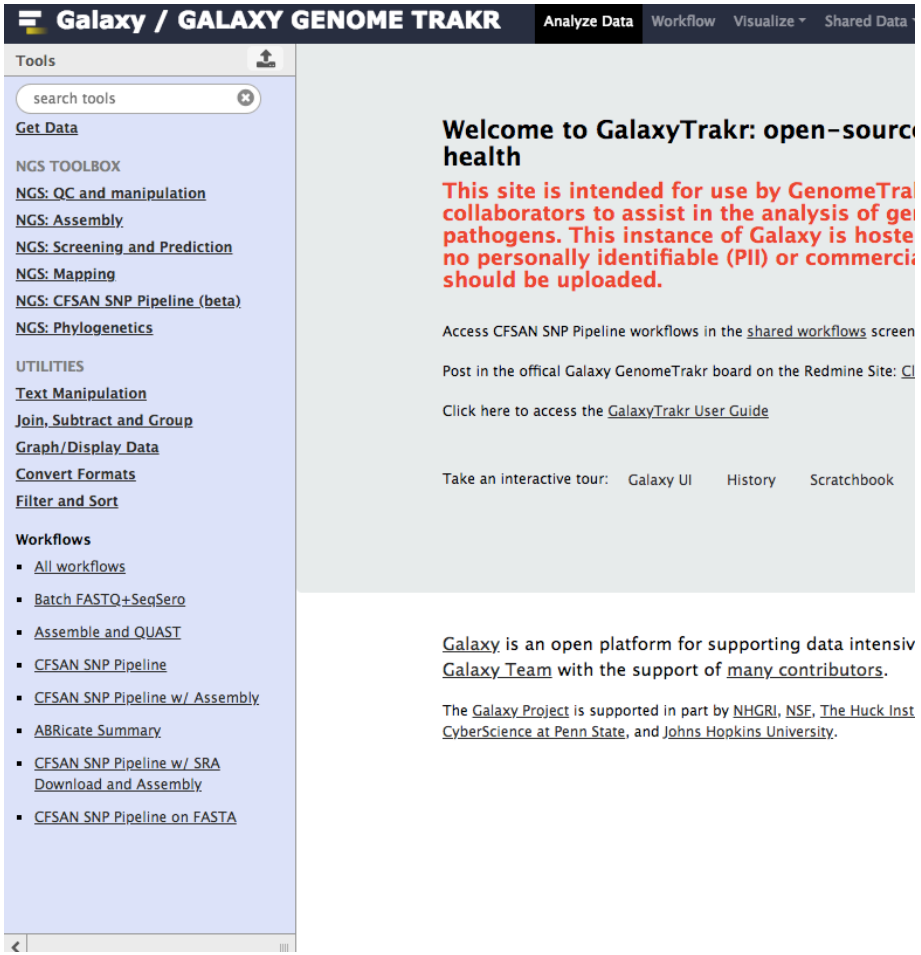


Galaxy (workflows)



UseGalaxy vs GalaxyTrakr





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."

GalaxyTrakr

Who do we target

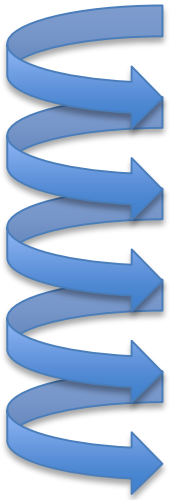
1. “My web browser? Google, I think”
2. “Is the internet down? I can’t get to the Facebook”
3. “Oh, there’s an Excel formula for that.”
4. “You mean you *didn’t* change your router’s default password?”
5. “Where’s Terminal on this thing?”
6. “There’s a funny easter egg in echo’s man page.” ***A scurrilous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

GalaxyTrakr

Curation and Layout Philosophy

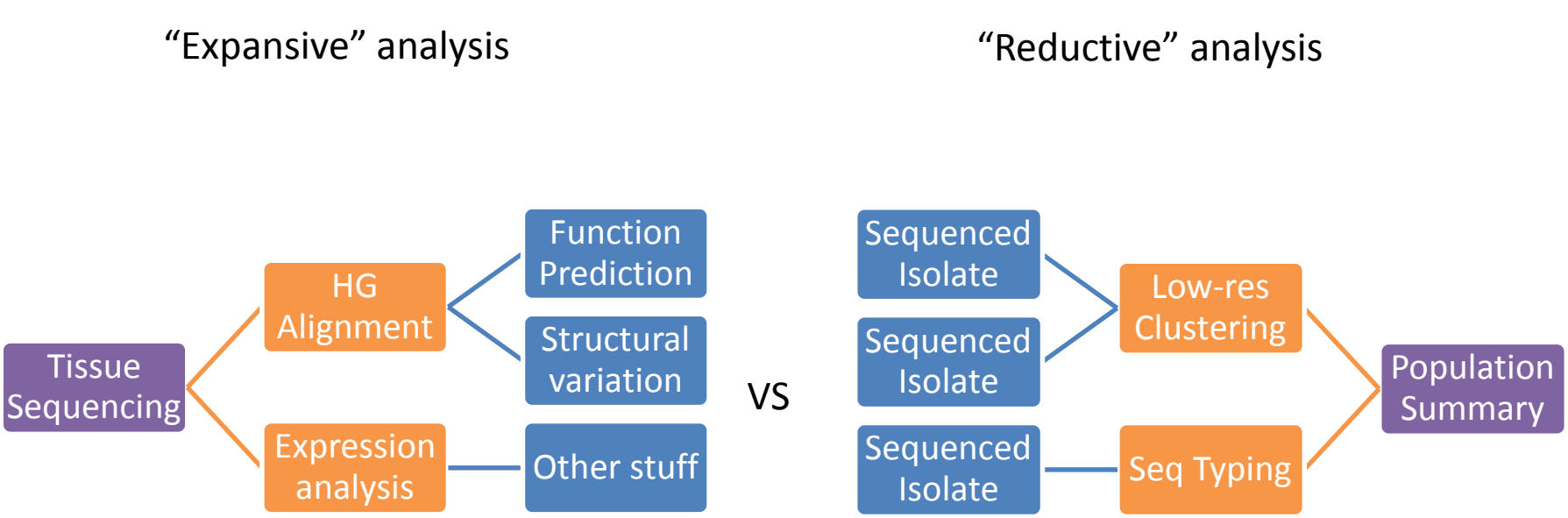
- ”One (or two) ways to do things”
- “Obvious path forward”

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



GalaxyTrakr

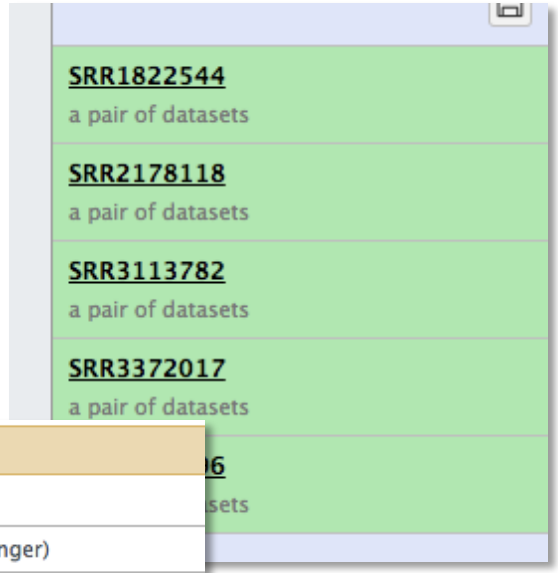
Expansive vs. Reductive Analyses

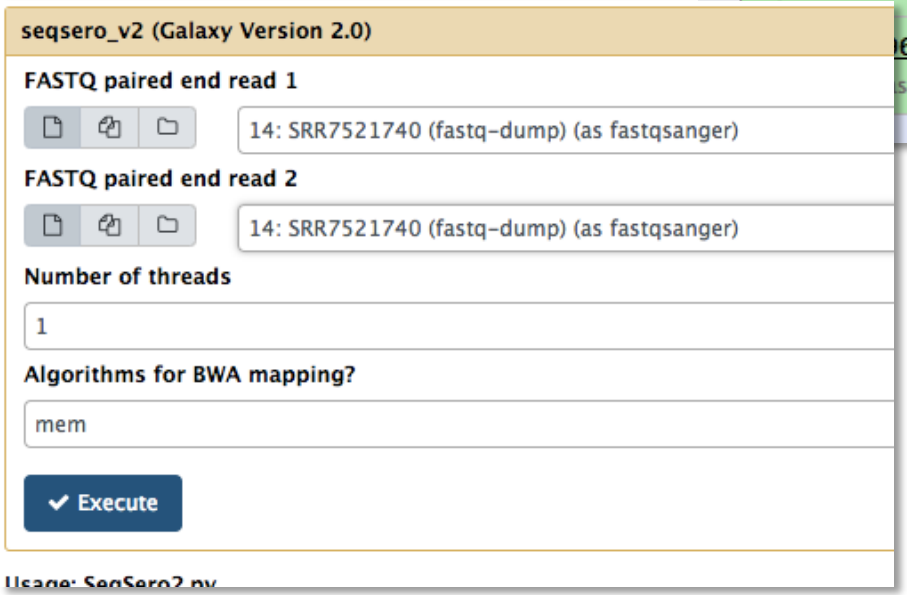


GalaxyTrakr

Challenges to Population Analysis

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





FDA U.S. FOOD & DRUG ADMINISTRATION