U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

Human Foods Program Regulatory Genomics and Pathogen Surveillance

Bridging Genomic Databases, Interactive Dashboards, and AI

James Pettengill
Director, Division of Surveillance and Data Integration
Office of Surveillance Strategy and Risk Prioritization
Human Foods Program, FDA

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

Outline

- What is Pathogen Surveillance?
- RIPS – Rapid Intuitive Pathogen Surveillance
- SAGE - Sequence Analysis for Genomic Epidemiology
- GalaxyTrakr – Future AI integration

Pathogen Surveillance: complex and multi-dimensional

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

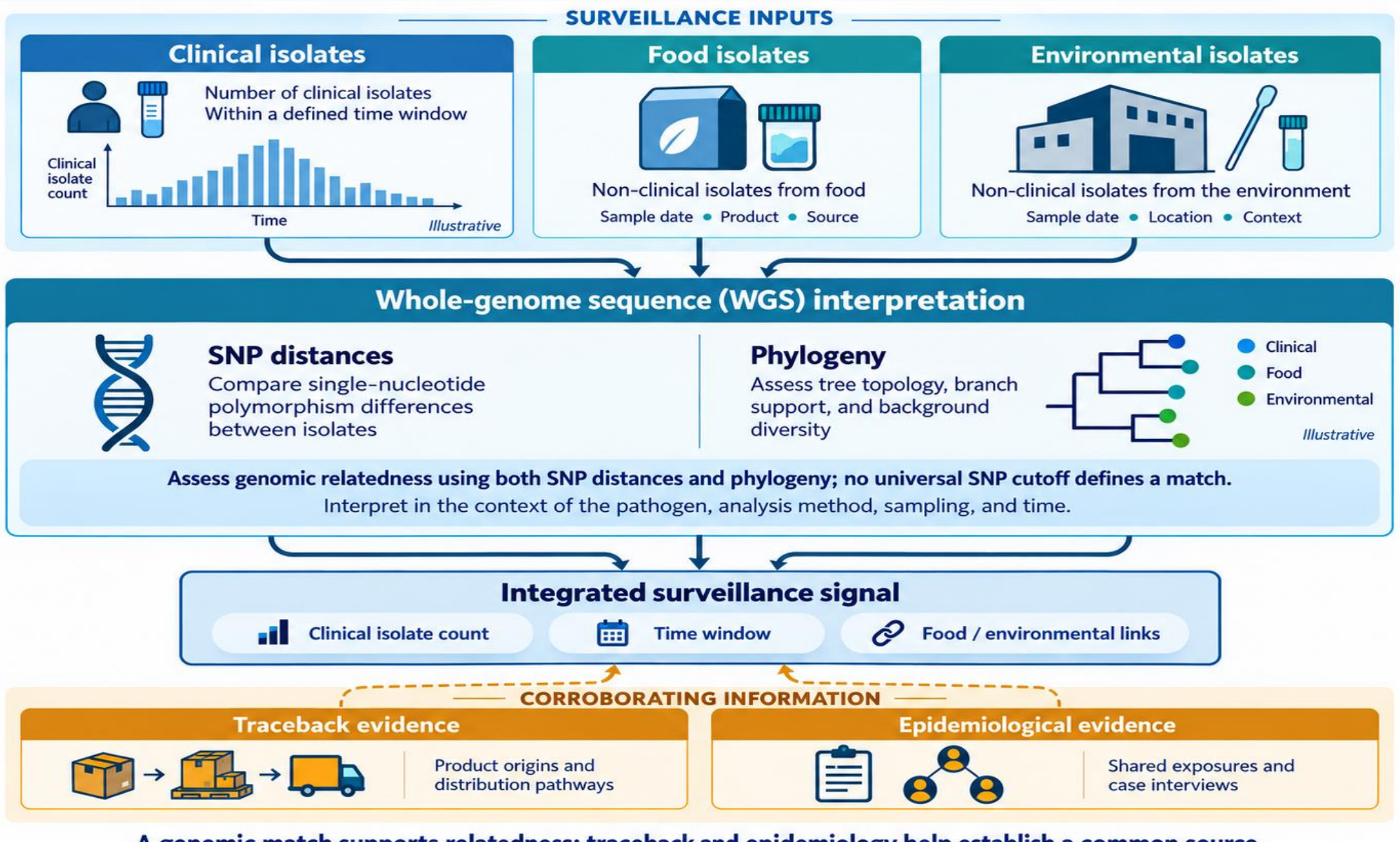


Image generated with GPT-6.1 Sol

Regulatory Genomics and Pathogen Surveillance

Regulatory Toxicology and Pharmacology 80 (2016) 342–347

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph



Workshop report

Regulatory bioinformatics for food and drug safety

Marion J. Healy [a], Weida Tong [b, *], Stephen Ostroff [c], Hans-Georg Eichler [d], Alex Patak [e], Margaret Neuspiel [f], Hubert Deluyker [g], William Slikker Jr. [b, **]

[a] Food Standards Australia New Zealand, Barton, Australian Capital Territory, 2905, Australia
[b] National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, 72079-9502, USA
[c] US Food and Drug Administration, Silver Spring, MD, 20993-0002, USA
[d] European Medicines Agency, London, E14 5EU, United Kingdom
[e] European Commission, Joint Research Centre, Ispra, Varese, 21027, Italy
[f] Canadian Food Inspection Agency, Ottawa, Ontario, Canada
[g] European Food Safety Authority, Parma, 43126, Italy

"'Regulatory Bioinformatics' strives to develop and implement a standardized and transparent bioinformatic framework to support the implementation of existing and emerging technologies in regulatory decision-making."

RIPS (RAPID INTUITIVE PATHOGEN SURVEILLANCE)

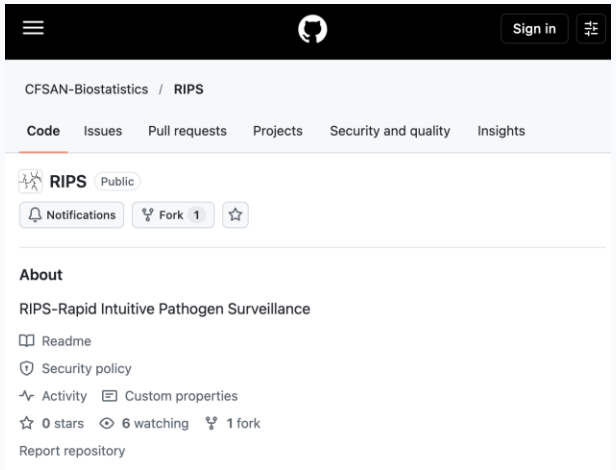
Making Raw Rapid Reports Actionable

THE CHALLENGE

NCBI Pathogen Detection files are raw text (and now parquet format) on PD's ftp site that are highly complex to parse manually for immediate signals. Public health officials need a way to quickly digest this massive influx of daily genomic data.

THE SOLUTION

RIPS is an open-source R Shiny application developed by FDA researchers (Muruvanda et al., 2024) that transforms raw data into an interactive, visual dashboard. (**NLP for metadata**)



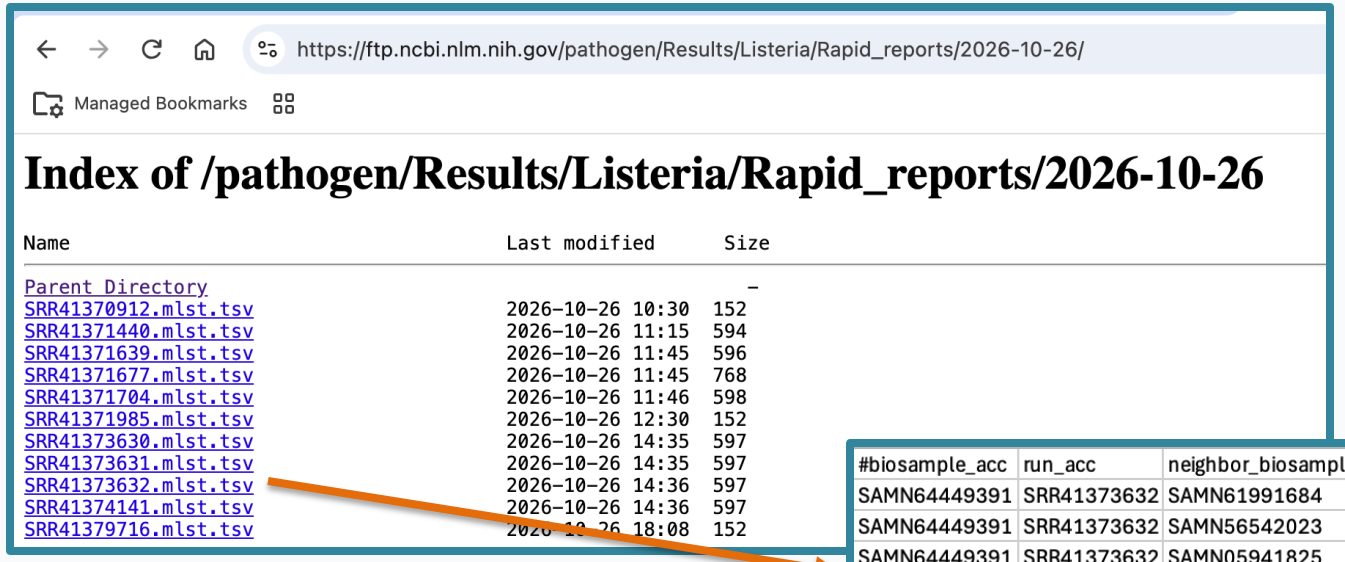
CORE FUNCTIONS

- **Automated Daily Pulls**
  Downloads NCBI Rapid Reports and matching BioSample metadata.
- **Interactive Dashboard**
  Renders clusters in easy-to-use tables and visual charts.
- **Custom Filtering**
  Allows triage of clusters based on minimum clinical counts, maximum allele distance, and matching non-clinical isolates.

RIPS (RAPID INTUITIVE PATHOGEN SURVEILLANCE)

What are NCBI's rapid reports?

- FTP Rapid Reports are based solely on wgMLST allele differences and currently reports on certain bioprojects and some organisms.
- The FTP Rapid Reports for a given organism are updated within approximately an hour of completing sequence read submissions for a new isolate.



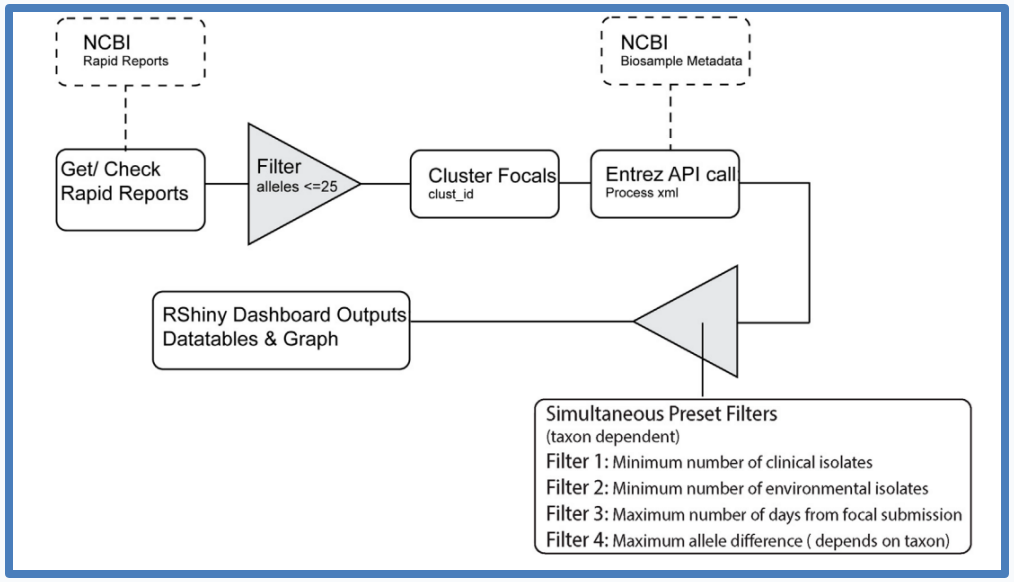
#biosample_acc	run_acc	neighbor_biosample_acc	neighbor_run_acc	neighbor_assembly_acc	neighbor_isolate	neighbor_cluster	alleles_different	loci_in_common
SAMN64449391	SRR41373632	SAMN61991684	SRR39850820	NULL	PDT003301115	PDS000259011	5	2775
SAMN64449391	SRR41373632	SAMN56542023	SRR37658358	NULL	PDT003144710	PDS000259011	5	2774
SAMN64449391	SRR41373632	SAMN05941825	SRR4450287	NULL	PDT000154107	PDS000259011	42	2798
SAMN64449391	SRR41373632	SAMN05941825	NULL	GCA_004577615.1	PDT000154107	PDS000259011	42	2747
SAMN64449391	SRR41373632	SAMEA9569205	ERR6705439	NULL	PDT001139431	PDS000174594	81	2735

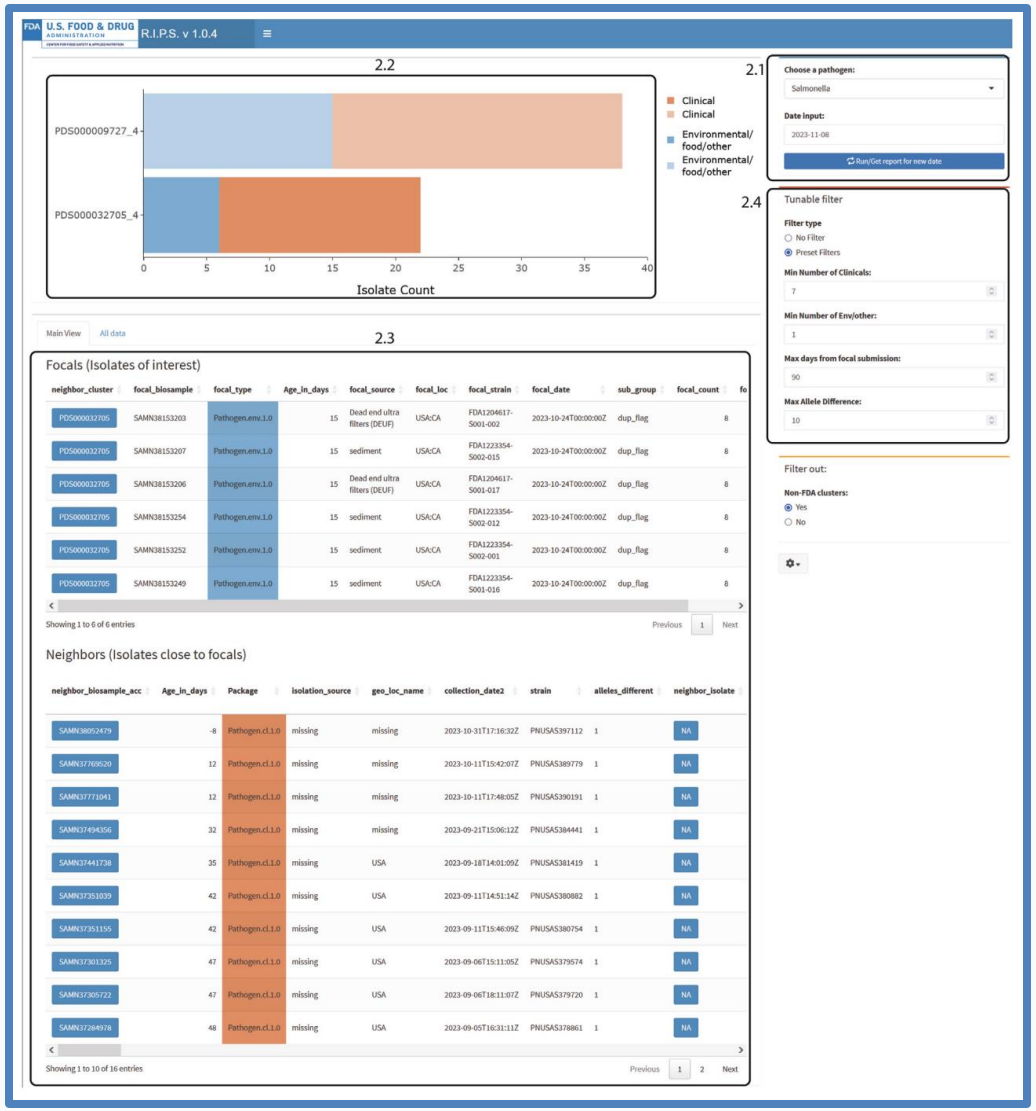
RIPS (RAPID INTUITIVE PATHOGEN SURVEILLANCE)

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

RIPS (rapid intuitive pathogen surveillance): a tool for surveillance of genome sequence data from foodborne bacterial pathogens

Tim Muruvanda*, Hugh Rand, James Pettengill and Arthur Pightling





SAGE - Sequence Analysis for Genomic Epidemiology

Expediting NCBI SNP Cluster Review for Surveillance and Signal Detection

THE CHALLENGE

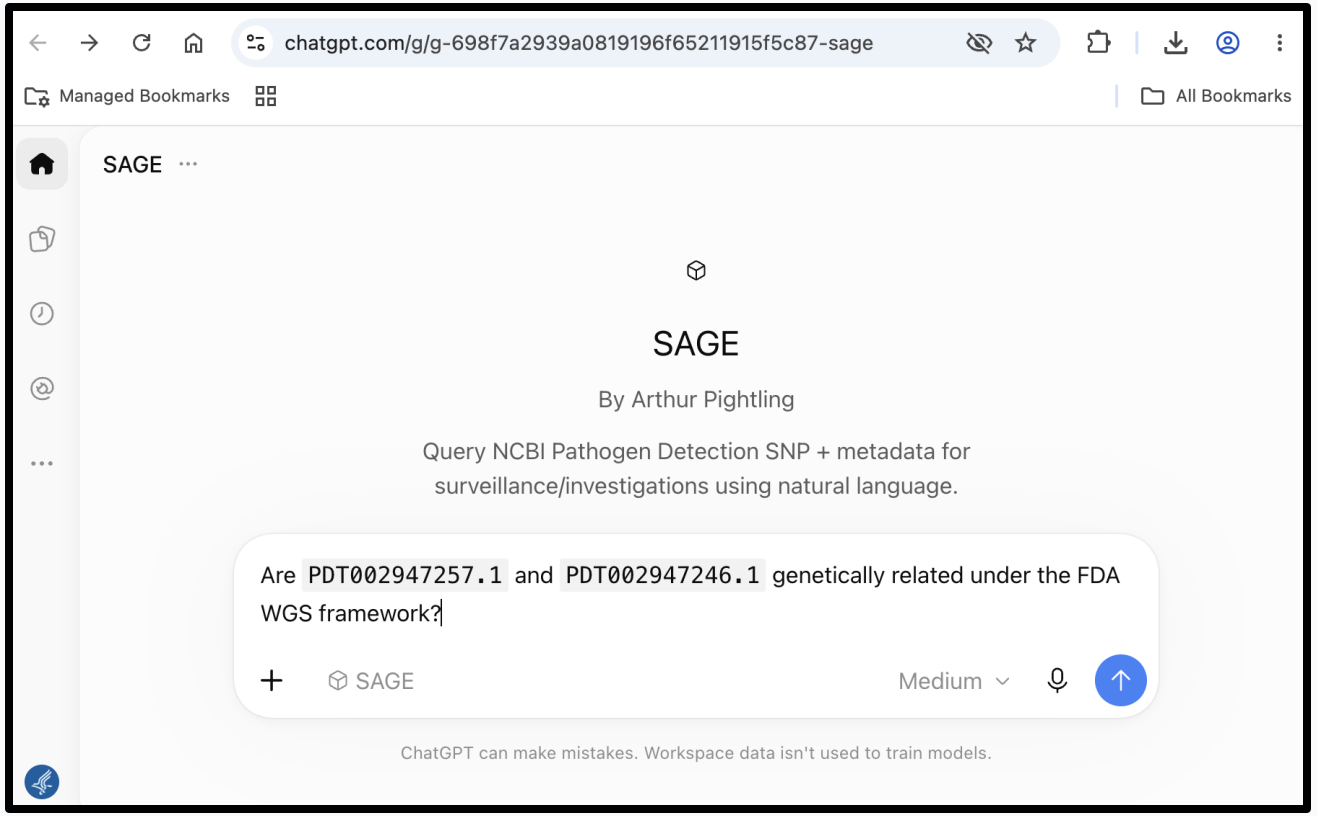
NCBI SNP clusters contain a wealth of information including the number of clinical isolates, non-clinical isolates, date of upload (and sometimes isolation), geographic information, SNP distances, and more. Synthesizing all that data for pathogen surveillance can be challenging.

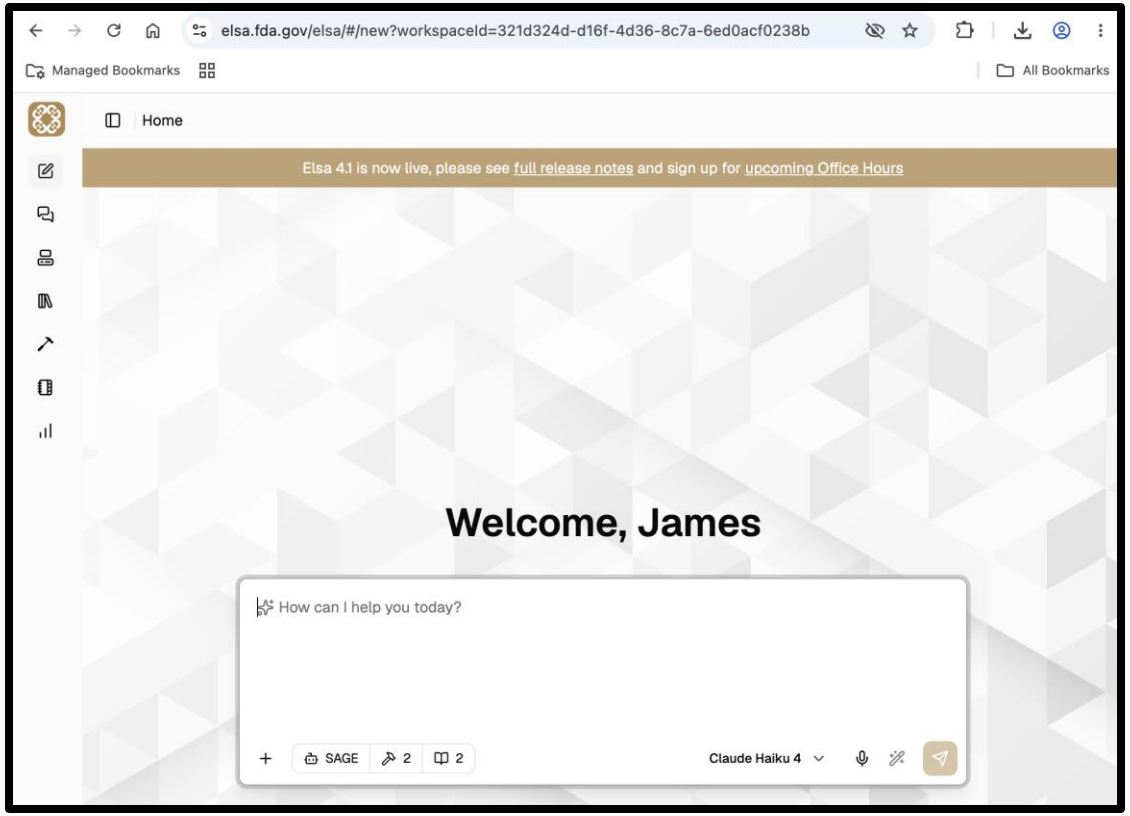
THE SOLUTION

SAGE is an AI-agent (prompt + libraries/documents) that takes the files associated with an NCBI SNP cluster (metadata, tree files, SNP distances) and performs a series of both technical bioinformatic tasks in addition to LLM synthesis and generative AI to provide information useful for pathogen surveillance.
Instructions: 887 words – 97 lines (~2-page Word doc)

SAGE - Sequence Analysis for Genomic Epidemiology

Two environments





SAGE - Sequence Analysis for Genomic Epidemiology

Expediting NCBI SNP Cluster Review for Surveillance and Signal Detection

CORE FUNCTIONS

- **Phylogenetic analysis** - parses the full Newick tree, identifies MRCA, and determines monophyly.
- **SNP analysis -** extracts all required compatible_distance values and calculates cluster-level mean pairwise SNP distance.
- **Temporal and metadata analysis -** uses target_creation_date to reassess non-monophyletic queries within the relevant time window.
- **FDA framework interpretation -** integrates tree topology, SNP distances, and metadata to identify supported genetic clusters and report uncertainties.

SAGE - Sequence Analysis for Genomic Epidemiology

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

Example SAGE output

1) **Question -** Are PDT002947257.1 and PDT002947246.1 genetically related under the FDA WGS framework?

2) Data used - Full Newick phylogeny, pairwise SNP-distance table using compatible_distance, metadata, and the FDA WGS interpretation framework. [python]

3) PHASE 0 - Validation

- Tree: usable
- SNP table: readable
- Query IDs linked across sources: yes
- Required metadata present: yes

4) PHASE 1 - Tree topology

- Queries: PDT002947257.1, PDT002947246.1
- MRCA descendants: 2 isolates
- Queries are **2 of 2 descendants**
- Monophyly: **YES**
- Interpretation: the two query isolates form an exclusive clade in the full phylogeny.

SAGE - Sequence Analysis for Genomic Epidemiology

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

5) PHASE 2 - SNP distances

- Isolates in candidate clade: **2**
- Expected unique pairs: **1**
- Missing pairs: **0**
- Mean pairwise SNP distance: **2 SNPs**
- FDA SNP criterion: **Supportive**

6) PHASE 3 - Temporal analysis- Not required because the query isolates are monophyletic in the full tree.

7) FDA framework interpretation

The phylogenetic topology and low pairwise SNP distance provide **genetic support that these isolates are closely related**. The selected genetic cluster contains:

- PDT002947257.1
- PDT002947246.1

Classification: Supported genetic cluster / closely related isolates.

WGS evidence alone does not establish that the isolates came from the same food, facility, or exposure event; epidemiologic and traceback evidence should be considered separately.

8) Compliance check

✓ Full tree examined
✓ Monophyly determined from tree topology
✓ All required SNP pairs extracted exactly
✓ Complete cluster membership reported
✓ Temporal analysis correctly not triggered
✓ No missing SNP values estimated

SAGE - Sequence Analysis for Genomic Epidemiology

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

Caveats and Current Status of SAGE

- SAGE is **not** being used in routine surveillance at this time
- Work still needs to be done to make it operational and integrate it more seamlessly with PD's ftp site
- Tokens, tokens, tokens

GALAXYTRAKR PLATFORM

Accessible Bioinformatics for Public Health

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

01

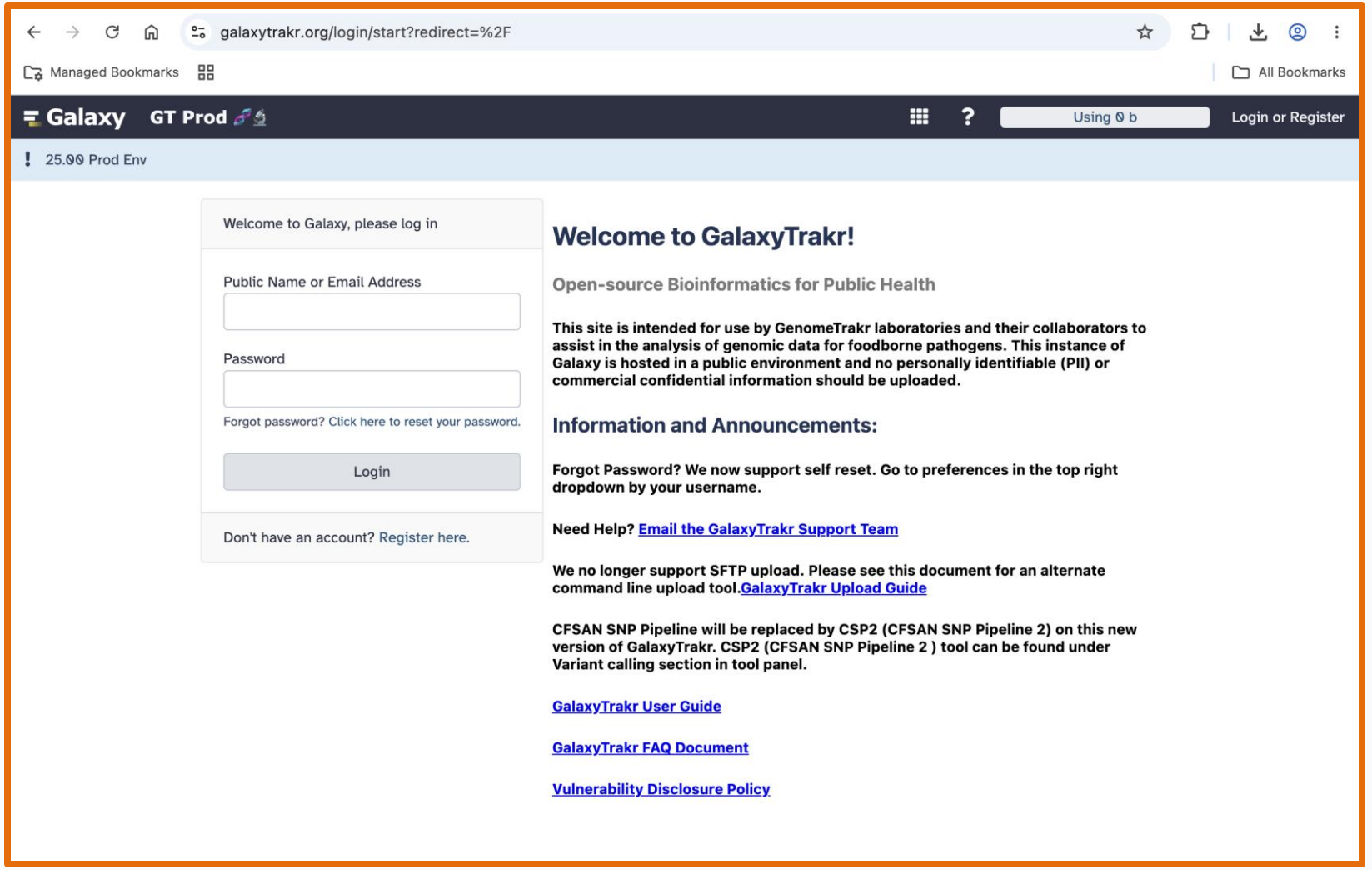
Free & Web-Based

A completely free, web-accessible GUI platform allowing researchers to execute complex command-line pipelines without coding.

02

Standardized Reproducibility

Unifies workflows so that globally distributed labs can replicate pathogen analyses using identical parameters.

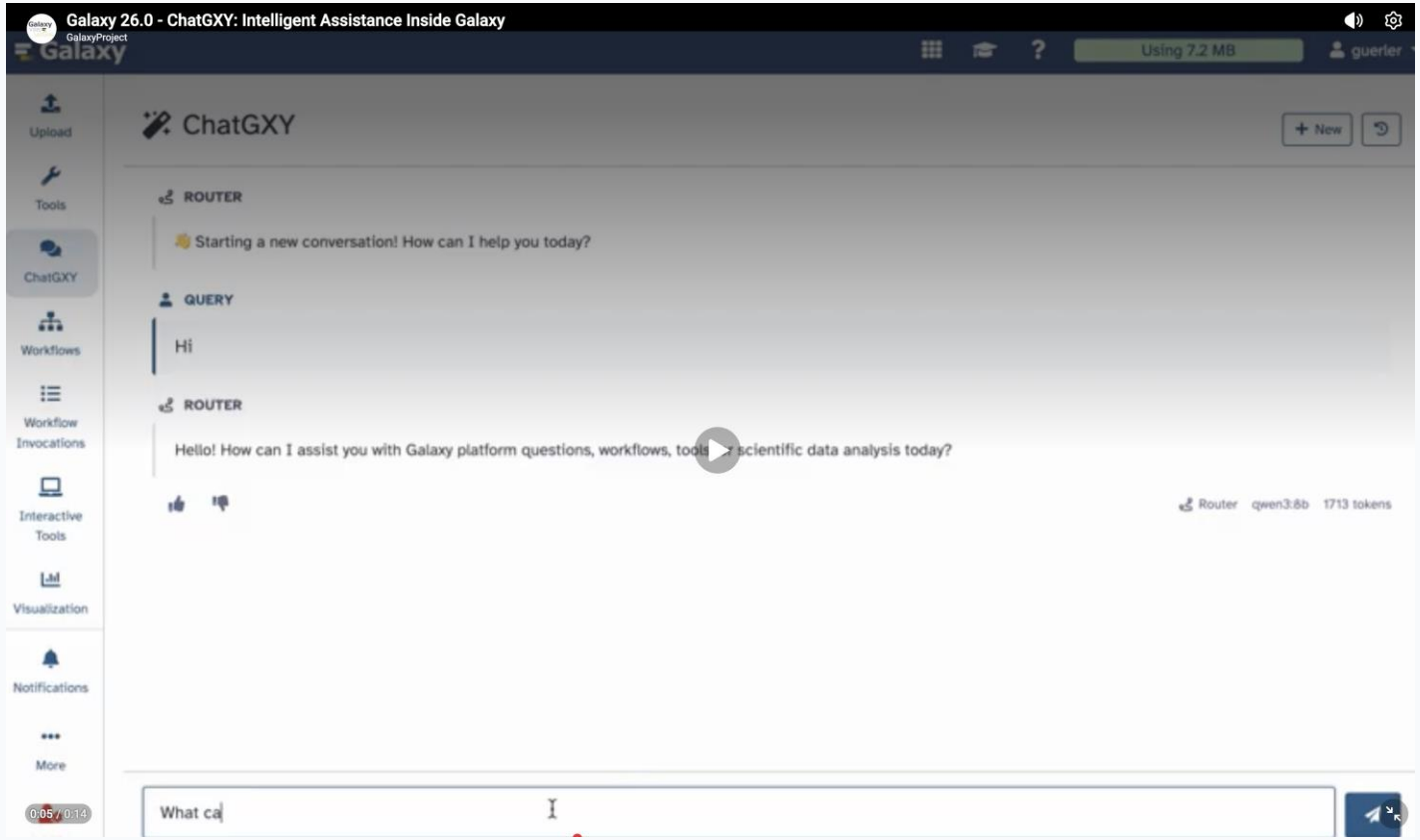


GalaxyTrakr Platform

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

AI (and other upgrades) in version 26 (GalaxyTrakr is currently run on v25)

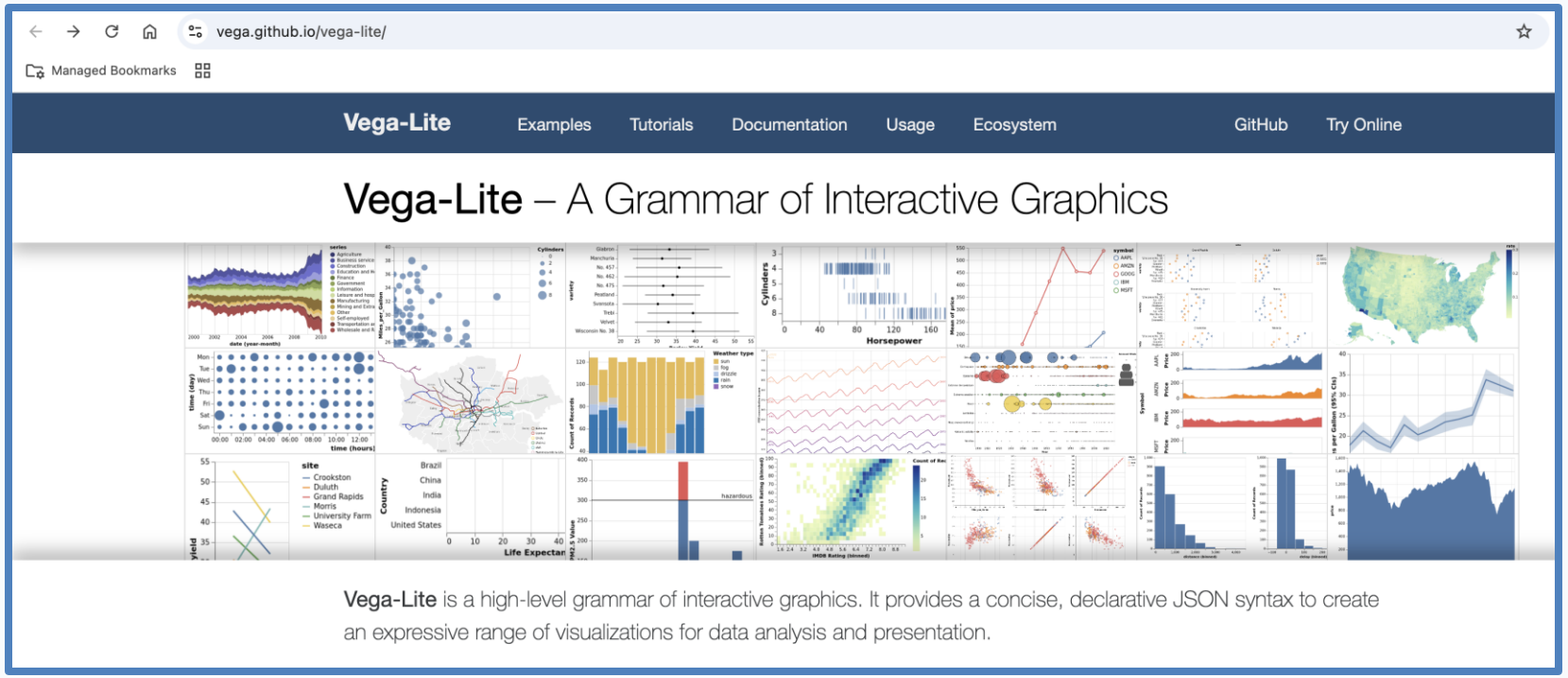
- **ChatGXY (claude-sonnet-5 on GalaxTrakr)**: A built-in conversational AI assistant integrated into the Galaxy interface.
  - **Conversational help**: Ask questions about tools, workflows, outputs, or error logs without leaving Galaxy.
  - **Error analysis guidance**: Receive plain-language explanations and troubleshooting steps when jobs fail.
  - **Tool recommendations**: Describe your analysis goals to get suggestions tailored to the tools installed on your specific Galaxy instance.



GALAXYTRAKR PLATFORM

AI in version 26 (GalaxyTrakr is currently run on v25)

- **Vintent (AI Data Visualization)**: An AI-driven visualization engine that turns natural language prompts into reproducible, interactive Vega-Lite charts running in a versioned Pyodide/Python browser environment.

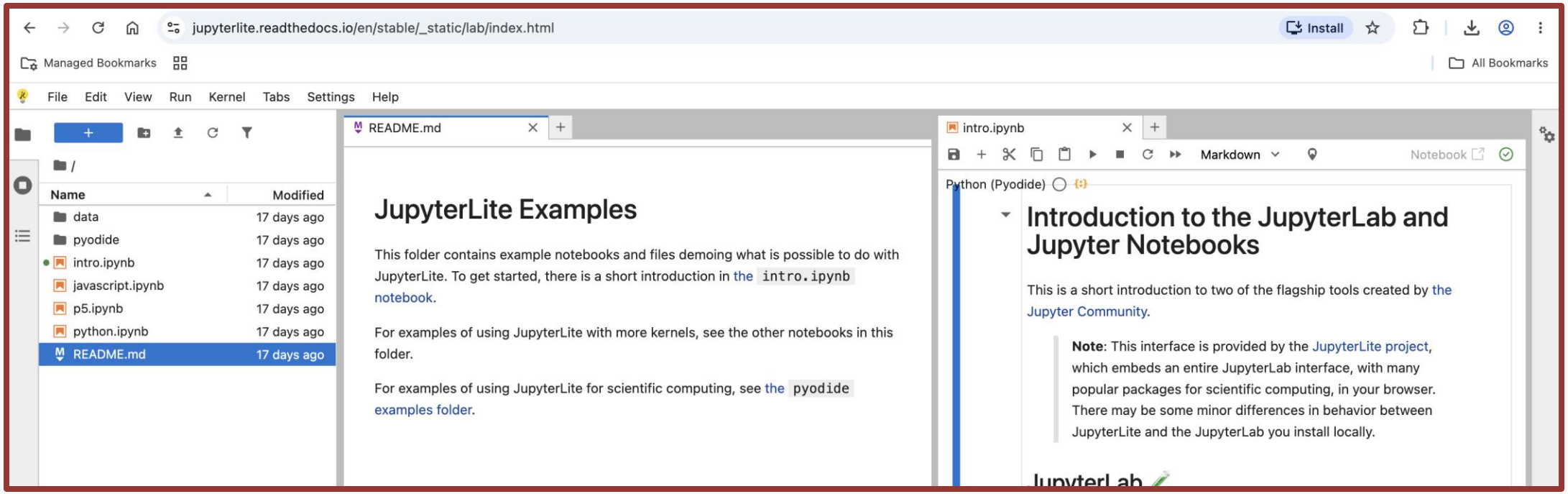


GALAXYTRAKR PLATFORM

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

AI in version 26 (GalaxyTrakr is currently run on v25)

- **AI-Assisted Notebooks with JupyterLite**: Launch browser-based, interactive JupyterLite notebooks connected directly to your Galaxy datasets without manual file downloads (we are yet to explore this capability)



Conclusions

- Surveillance is multi-dimensional and requires the integration across those dimensions to detect signals of interest
- RIPS is a publicly available platform that can help with more timely surveillance via integration with rapid reports
- SAGE provides a framework for using AI to also perform surveillance and cluster summary functions
  - Requires additional tooling and orchestration
- GalaxyTrakr - future AI integration will increase accessibility of bioinformatics by providing chat functions to guide the user

FDA U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

U.S. FOOD & DRUG ADMINISTRATION
Human Foods Program

Acknowledgements

- Tim Muruvanda (RIPS)
- Arthur Pightling (SAGE)
- OMITS
- OLOAS
- NCBI PD Team