Enhancing NCBI Pathogen Detection cluster surveillance with ncbi-cluster-tracker

Samuel Baird

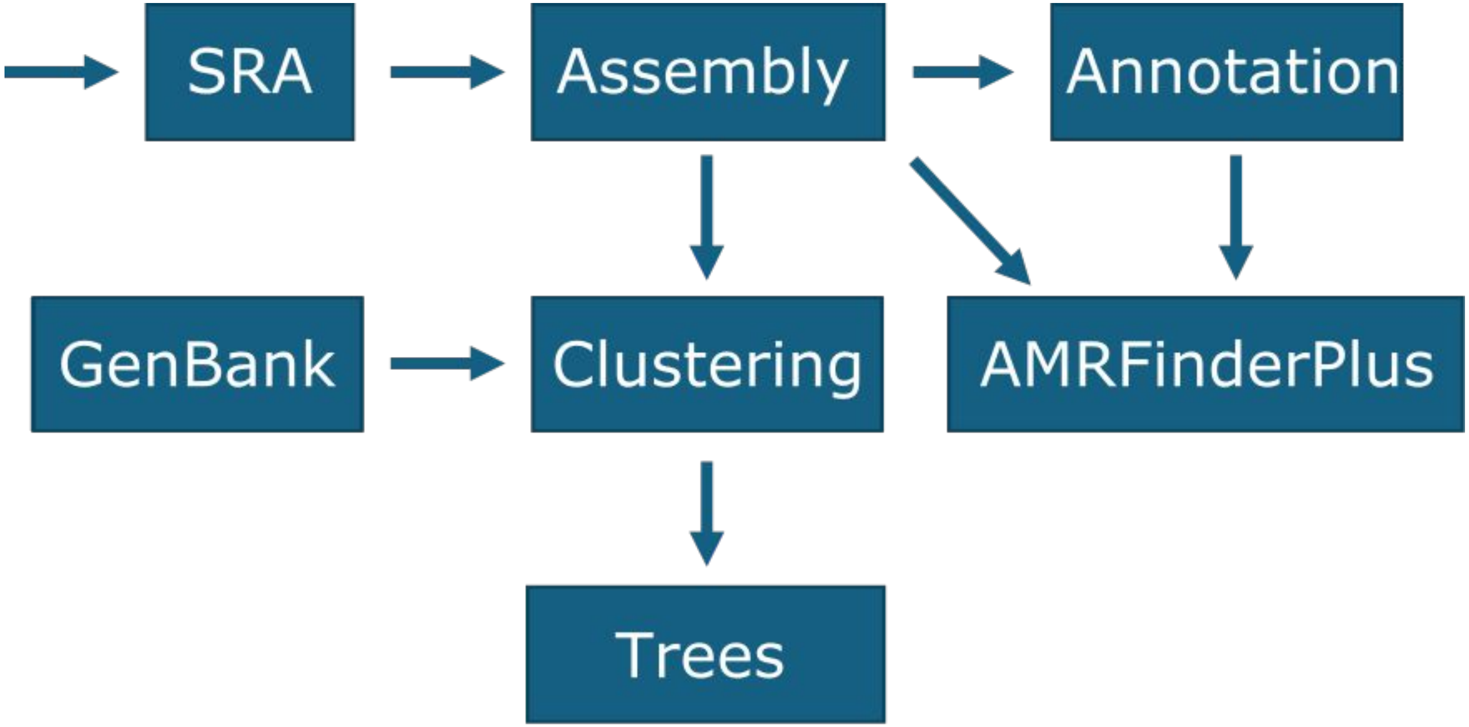
April 22, 2025

AMD Mountain Regional Bioinformatics Conference

COLORADO
Department of Public Health & Environment

NCBI Pathogen Detection

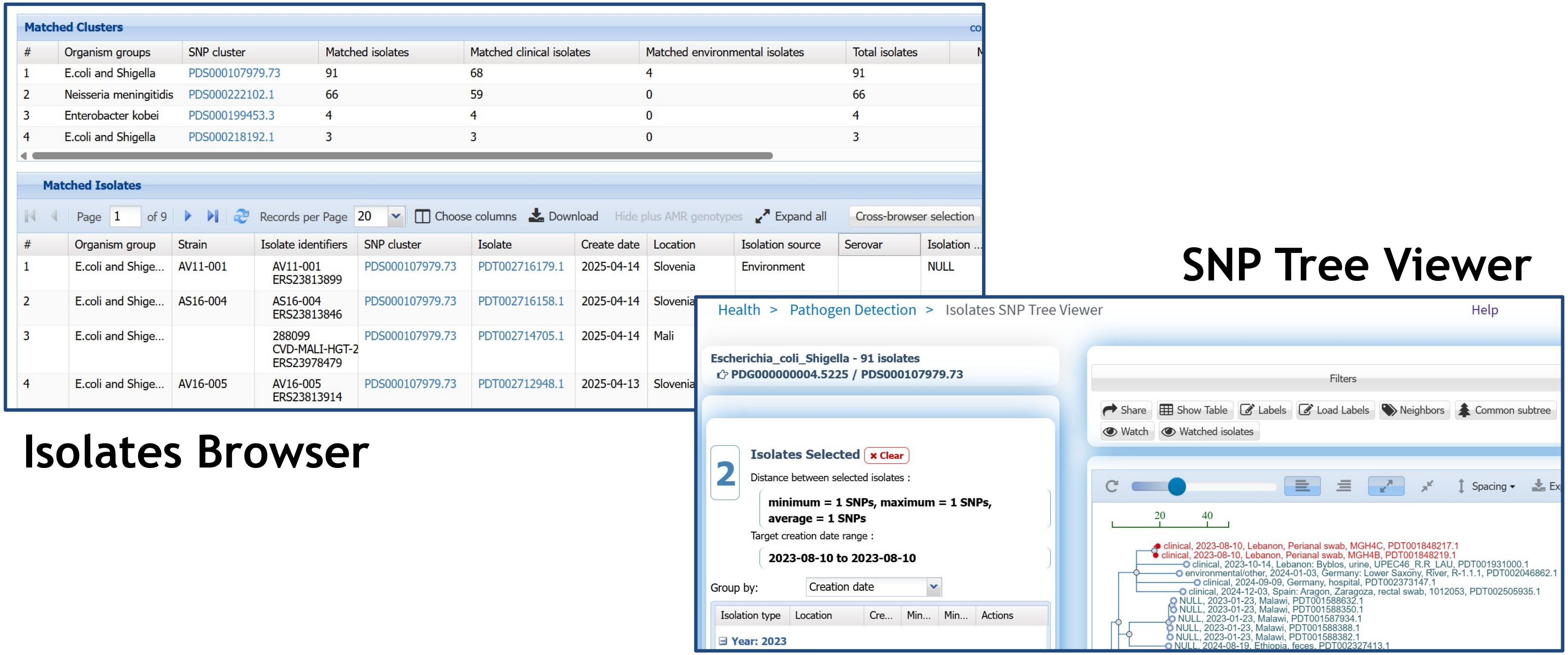
Automated system that clusters related bacterial and fungal pathogen genome sequences submitted to NCBI to help identify possible links between cases for genomic epidemiology



https://www.ncbi.nlm.nih.gov/pathogens/docs/data_processing/

NCBI Pathogen Detection

Isolates and associated clusters can be searched on the NCBI Isolate Browser website and phylogenetic trees investigated using the SNP Tree Viewer



Matched Clusters

#	Organism groups	SNP cluster	Matched isolates	Matched clinical isolates	Matched environmental isolates	Total isolates	M
1	E.coli and Shigella	PDS000107979.73	91	68	4	91	
2	Neisseria meningitidis	PDS000222102.1	66	59	0	66	
3	Enterobacter kobei	PDS000199453.3	4	4	0	4	
4	E.coli and Shigella	PDS000218192.1	3	3	0	3	

Matched Isolates

#	Organism group	Strain	Isolate identifiers	SNP cluster	Isolate	Create date	Location	Isolation source	Serovar	Isolation ...
1	E.coli and Shige...	AV11-001	AV11-001 ERS23813899	PDS000107979.73	PDT002716179.1	2025-04-14	Slovenia	Environment		NULL
2	E.coli and Shige...	AS16-004	AS16-004 ERS23813846	PDS000107979.73	PDT002716158.1	2025-04-14	Slovenia			
3	E.coli and Shige...		288099 CVD-MALI-HGT-2 ERS23978479	PDS000107979.73	PDT002714705.1	2025-04-14	Mali			
4	E.coli and Shige...	AV16-005	AV16-005 ERS23813914	PDS000107979.73	PDT002712948.1	2025-04-13	Slovenia			

Isolates Browser

SNP Tree Viewer

Escherichia_coli_Shigella - 91 isolates
PDG000000004.5225 / PDS000107979.73

2 Isolates Selected ✖ Clear

Distance between selected isolates :

minimum = 1 SNPs, maximum = 1 SNPs, average = 1 SNPs

Target creation date range :

2023-08-10 to 2023-08-10

Group by: Creation date

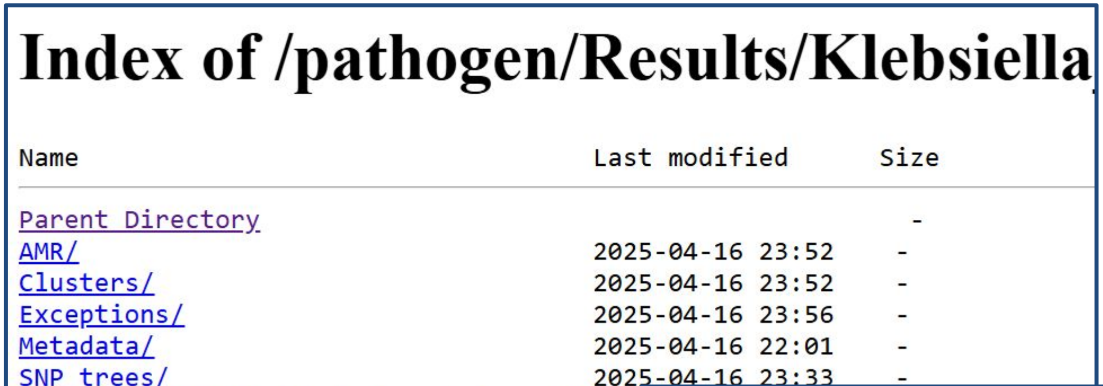
Isolation type	Location	Cre...	Min...	Min...	Actions
⊟ Year: 2023					

```
clinical, 2023-08-10, Lebanon, Perianal swab, MGH4C, PDT001848217.1
clinical, 2023-08-10, Lebanon, Perianal swab, MGH4B, PDT001848219.1
clinical, 2023-10-14, Lebanon: Byblos, urine, UPEC46_R.R_LAU, PDT001931000.1
environmental/other, 2024-01-03, Germany: Lower Saxony, River, R-1.1.1, PDT002046862.1
clinical, 2024-09-09, Germany, hospital, PDT002373147.1
clinical, 2024-12-03, Spain: Aragon, Zaragoza, rectal swab, 1012053, PDT002505935.1
NULL, 2023-01-23, Malawi, PDT001588632.1
NULL, 2023-01-23, Malawi, PDT001588350.1
NULL, 2023-01-23, Malawi, PDT001587934.1
NULL, 2023-01-23, Malawi, PDT001588388.1
NULL, 2023-01-23, Malawi, PDT001588382.1
NULL, 2024-08-19, Ethiopia, feces, PDT002327413.1
```

https://www.ncbi.nlm.nih.gov/pathogens/

NCBI Pathogen Detection

Pathogen Detection data can be also be accessed through the FTP file server and BigQuery data warehouse

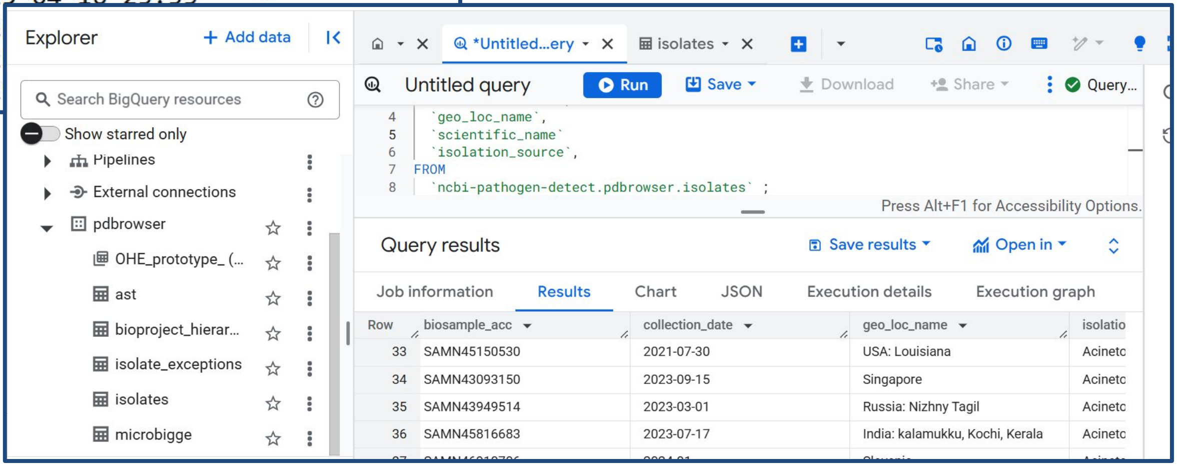


Index of /pathogen/Results/Klebsiella

```
Name                                    Last modified       Size

Parent Directory                                             -
AMR/                                    2025-04-16 23:52    -
Clusters/                               2025-04-16 23:52    -
Exceptions/                             2025-04-16 23:56    -
Metadata/                               2025-04-16 22:01    -
SNP_trees/                              2025-04-16 23:33    -
PDG000000030.960.final.descriptor.xml   202
PDG000000030.960.kmer.descriptor.xml    202
PDG000000030.960.snps.descriptor.xml    202
```

FTP Server

BigQuery data warehouse



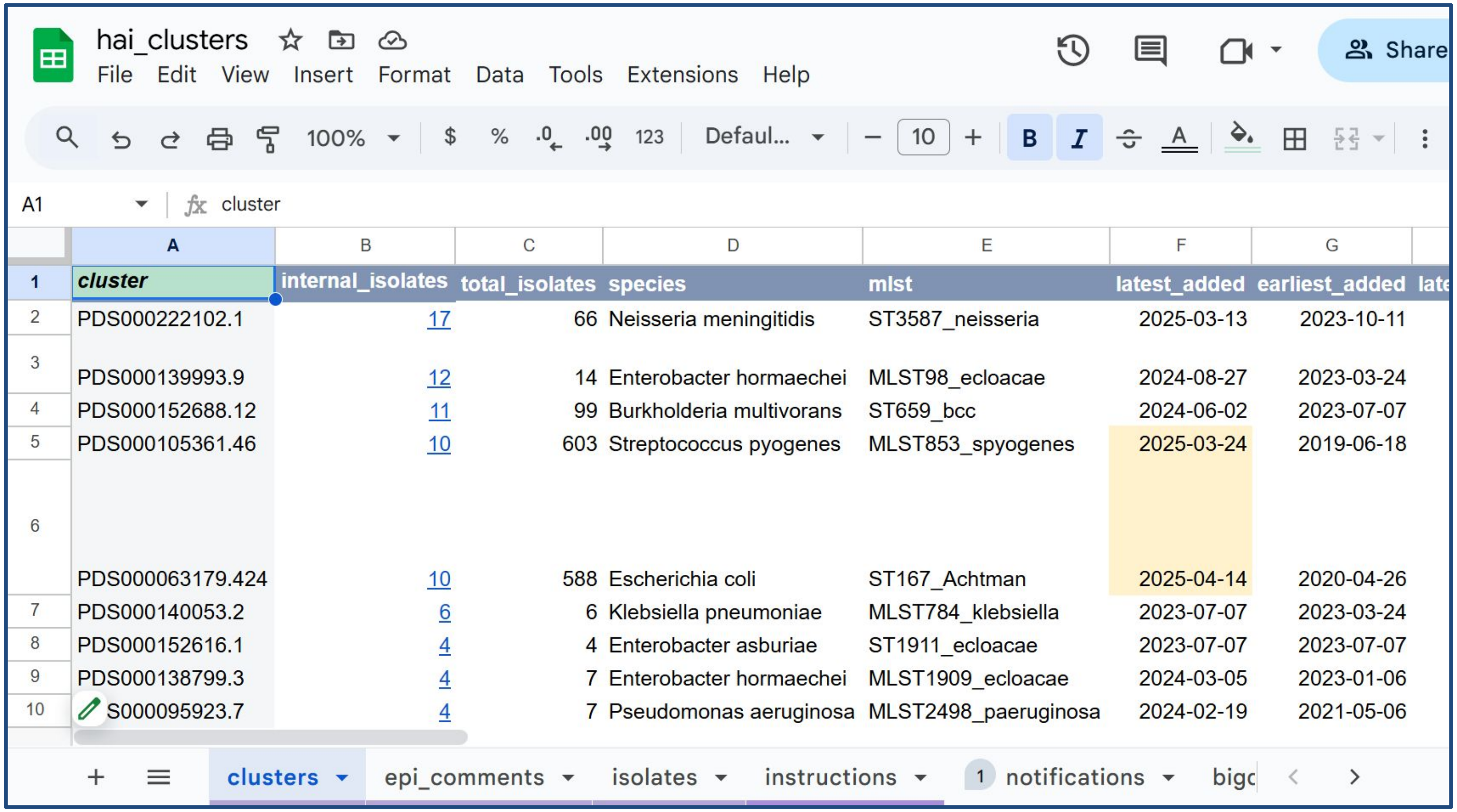
```
  `geo_loc_name`,
  `scientific_name`
  `isolation_source`,
FROM
  `ncbi-pathogen-detect.pdbrowser.isolates` ;
```

Row	biosample_acc	collection_date	geo_loc_name	isolatio
33	SAMN45150530	2021-07-30	USA: Louisiana	Acinetc
34	SAMN43093150	2023-09-15	Singapore	Acinetc
35	SAMN43949514	2023-03-01	Russia: Nizhny Tagil	Acinetc
36	SAMN45816683	2023-07-17	India: kalamukku, Kochi, Kerala	Acinetc

https://ftp.ncbi.nlm.nih.gov/pathogen/Results/
https://www.ncbi.nlm.nih.gov/pathogens/docs/getting_started_bigquery/

Original system for tracking HAI clusters

Every week: submit new sequencing data to NCBI, run BigQuery SQL query, copy results to Google sheet, review new clusters with epi at weekly meeting



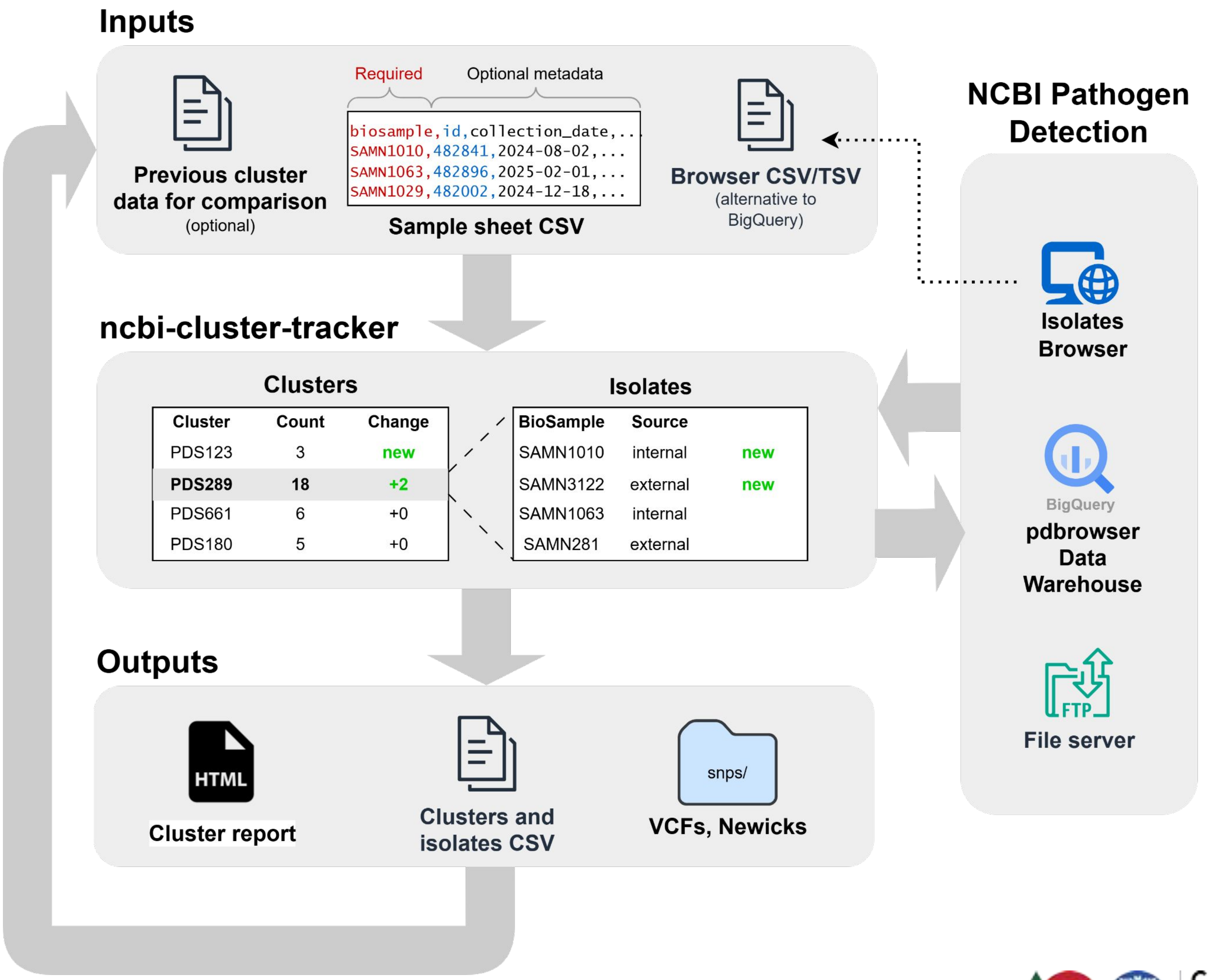
cluster	internal_isolates	total_isolates	species	mlst	latest_added	earliest_added	late
PDS000222102.1	17	66	Neisseria meningitidis	ST3587_neisseria	2025-03-13	2023-10-11	
PDS000139993.9	12	14	Enterobacter hormaechei	MLST98_ecloacae	2024-08-27	2023-03-24	
PDS000152688.12	11	99	Burkholderia multivorans	ST659_bcc	2024-06-02	2023-07-07	
PDS000105361.46	10	603	Streptococcus pyogenes	MLST853_spyogenes	2025-03-24	2019-06-18	
PDS000063179.424	10	588	Escherichia coli	ST167_Achtman	2025-04-14	2020-04-26	
PDS000140053.2	6	6	Klebsiella pneumoniae	MLST784_klebsiella	2023-07-07	2023-03-24	
PDS000152616.1	4	4	Enterobacter asburiae	ST1911_ecloacae	2023-07-07	2023-07-07	
PDS000138799.3	4	7	Enterobacter hormaechei	MLST1909_ecloacae	2024-03-05	2023-01-06	
S000095923.7	4	7	Pseudomonas aeruginosa	MLST2498_paeruginosa	2024-02-19	2021-05-06	

COLORADO
Department of Public Health & Environment

Goals for a new system and report format

1. Indicate new isolates and new clusters from the previous week
2. Annotate with additional metadata not published on NCBI (exact collection dates, internal lab IDs…)
3. Additional data visualizations (isolate counts over time, pairwise SNP distance matrix)
4. Report can be securely shared and explored
5. Agnostic to CDPHE data systems, could be deployed and adopted by other agencies

ncbi-cluster-tracker



Inputs

Previous cluster data for comparison
(optional)

Required · Optional metadata

```
biosample,id,collection_date,...
SAMN1010,482841,2024-08-02,...
SAMN1063,482896,2025-02-01,...
SAMN1029,482002,2024-12-18,...
```

Sample sheet CSV

Browser CSV/TSV
(alternative to BigQuery)

NCBI Pathogen Detection

ncbi-cluster-tracker

Clusters

Cluster	Count	Change
PDS123	3	new
PDS289	**18**	**+2**
PDS661	6	+0
PDS180	5	+0

Isolates

BioSample	Source	
SAMN1010	internal	new
SAMN3122	external	new
SAMN1063	internal	
SAMN281	external	

Outputs

Cluster report

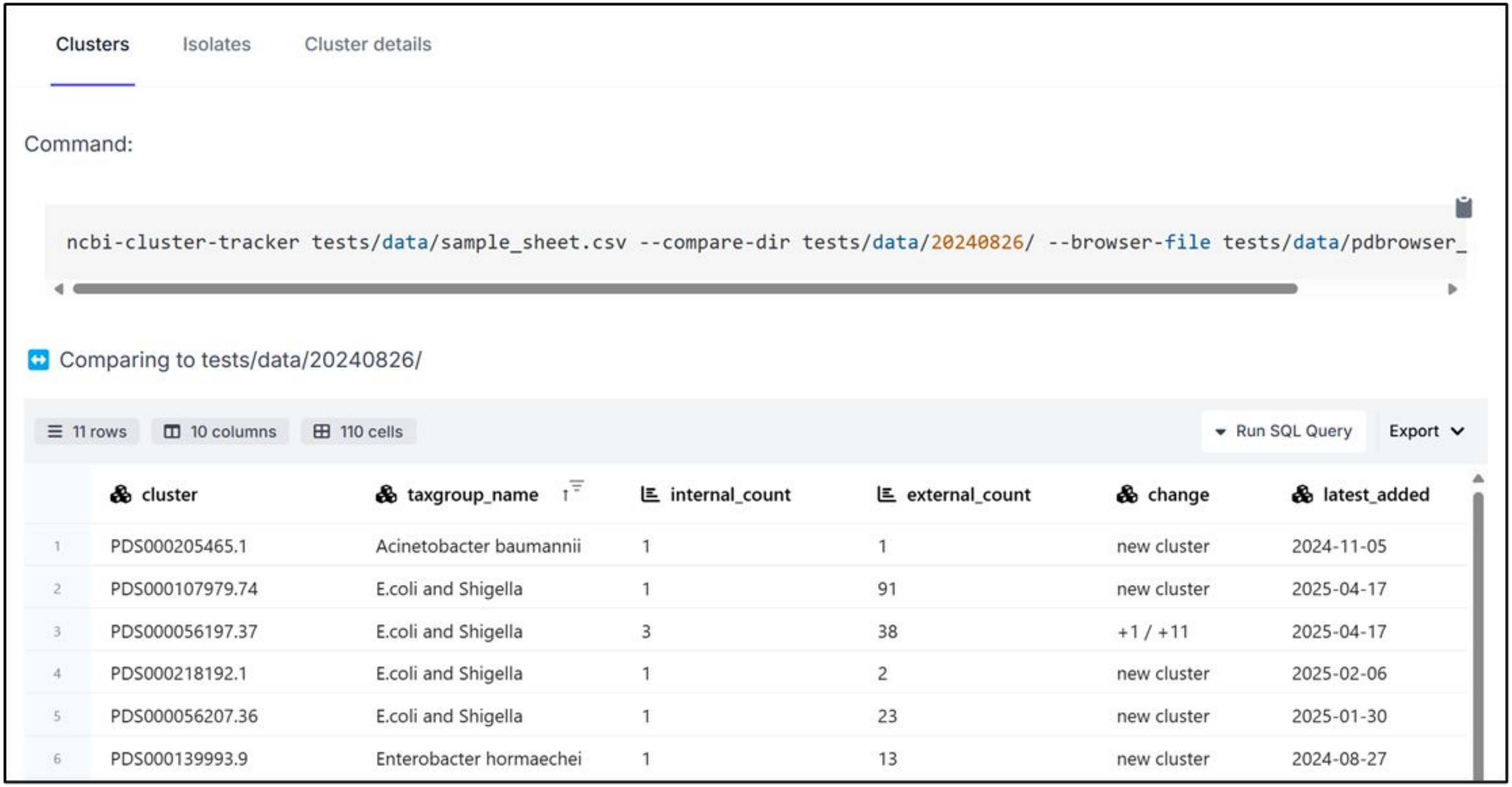
Clusters and isolates CSV

VCFs, Newicks

COLORADO
Department of Public Health & Environment

Demo

Clusters table: View clusters associated with internal isolates and changes to isolate counts since the previous report was created.



Command:

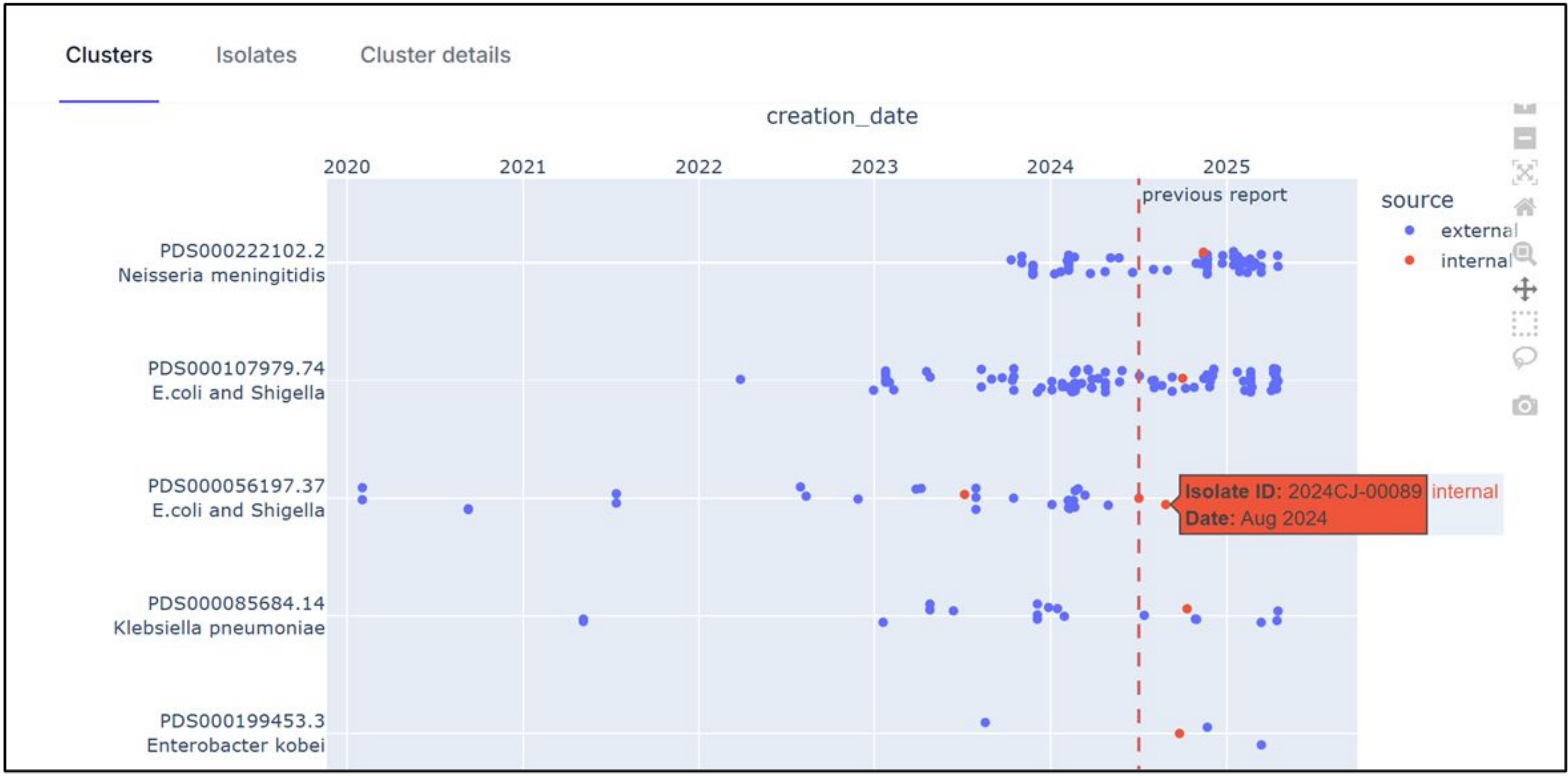
```
ncbi-cluster-tracker tests/data/sample_sheet.csv --compare-dir tests/data/20240826/ --browser-file tests/data/pdbrowser_
```

Comparing to tests/data/20240826/

	cluster	taxgroup_name	internal_count	external_count	change	latest_added
1	PDS000205465.1	Acinetobacter baumannii	1	1	new cluster	2024-11-05
2	PDS000107979.74	E.coli and Shigella	1	91	new cluster	2025-04-17
3	PDS000056197.37	E.coli and Shigella	3	38	+1 / +11	2025-04-17
4	PDS000218192.1	E.coli and Shigella	1	2	new cluster	2025-02-06
5	PDS000056207.36	E.coli and Shigella	1	23	new cluster	2025-01-30
6	PDS000139993.9	Enterobacter hormaechei	1	13	new cluster	2024-08-27

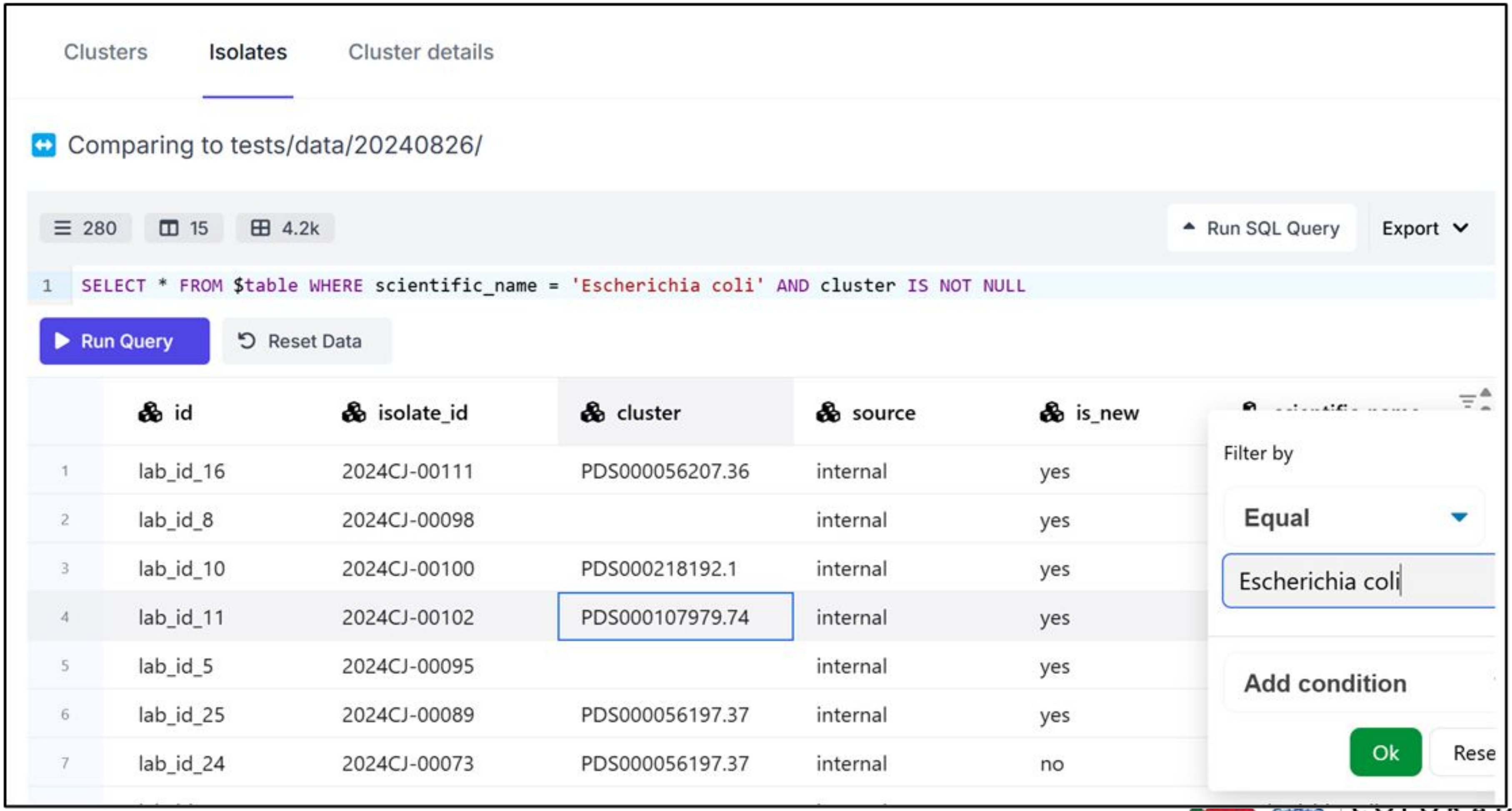
COLORADO
Department of Public
Health & Environment

Cluster timelines: View when isolates were added to each cluster.


Clusters
Isolates
Cluster details
creation_date
2020
2021
2022
2023
2024
2025
previous report
source
● external
● internal
PDS000222102.2 Neisseria meningitidis
PDS000107979.74 E.coli and Shigella
PDS000056197.37 E.coli and Shigella
PDS000085684.14 Klebsiella pneumoniae
PDS000199453.3 Enterobacter kobei
Isolate ID: 2024CJ-00089
Date: Aug 2024
internal

COLORADO
Department of Public Health & Environment

Isolates table: View metadata and associated clusters for specific isolates. Table can be sorted, filtered, and exported for further analysis.



Clusters | Isolates | Cluster details

Comparing to tests/data/20240826/

☰ 280 | ◫ 15 | ⊞ 4.2k | ▲ Run SQL Query | Export ⌄

```
SELECT * FROM $table WHERE scientific_name = 'Escherichia coli' AND cluster IS NOT NULL
```

▶ Run Query | ↺ Reset Data

	id	isolate_id	cluster	source	is_new
1	lab_id_16	2024CJ-00111	PDS000056207.36	internal	yes
2	lab_id_8	2024CJ-00098		internal	yes
3	lab_id_10	2024CJ-00100	PDS000218192.1	internal	yes
4	lab_id_11	2024CJ-00102	PDS000107979.74	internal	yes
5	lab_id_5	2024CJ-00095		internal	yes
6	lab_id_25	2024CJ-00089	PDS000056197.37	internal	yes
7	lab_id_24	2024CJ-00073	PDS000056197.37	internal	no

Filter by

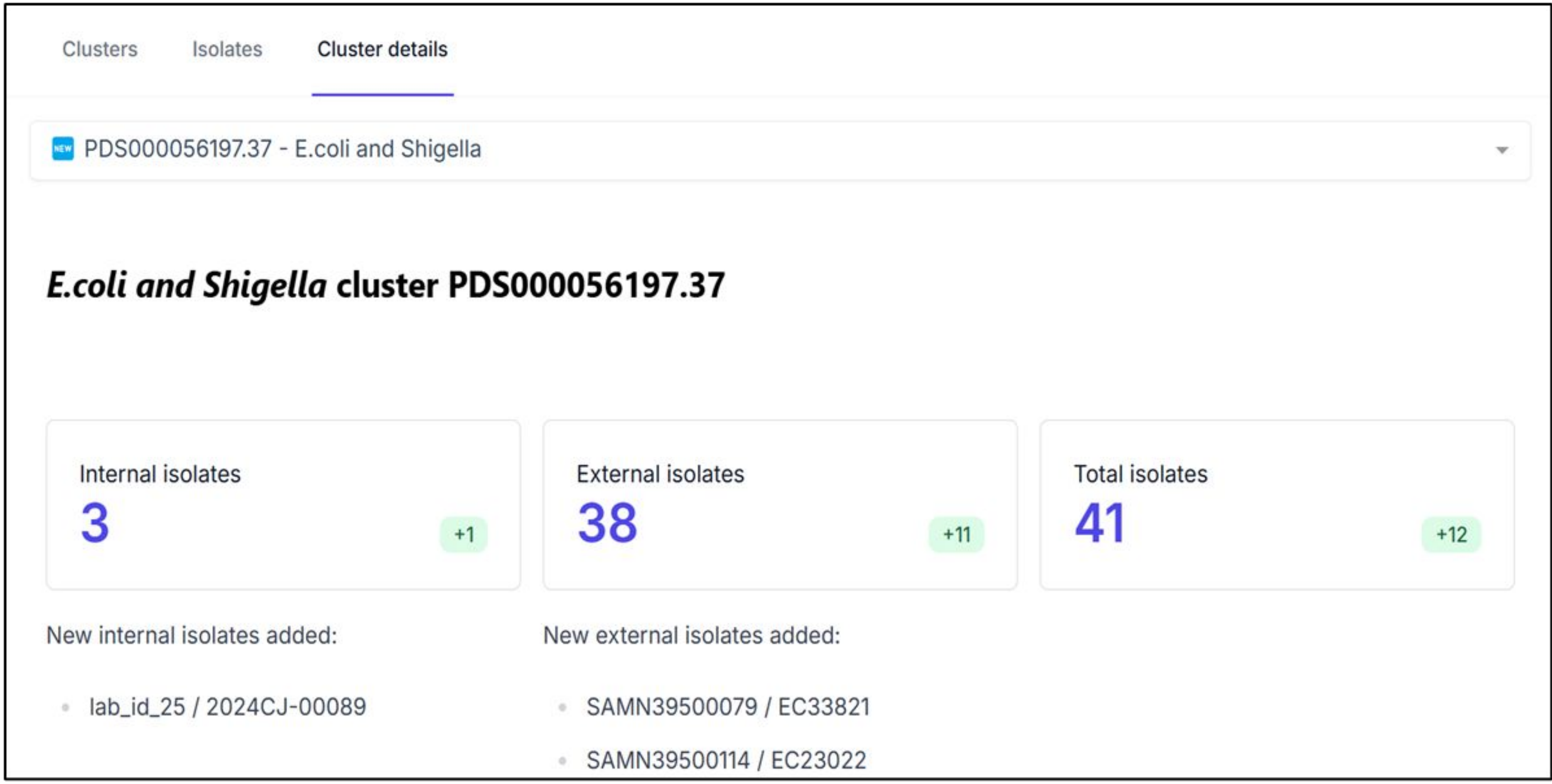
Equal ▾

Escherichia coli

Add condition

Ok | Rese

Cluster details: Look up details about specific clusters using the dropdown menu or search bar.



PDS000056197.37 - E.coli and Shigella

***E.coli and Shigella* cluster PDS000056197.37**

Internal isolates
3
+1

External isolates
38
+11

Total isolates
41
+12

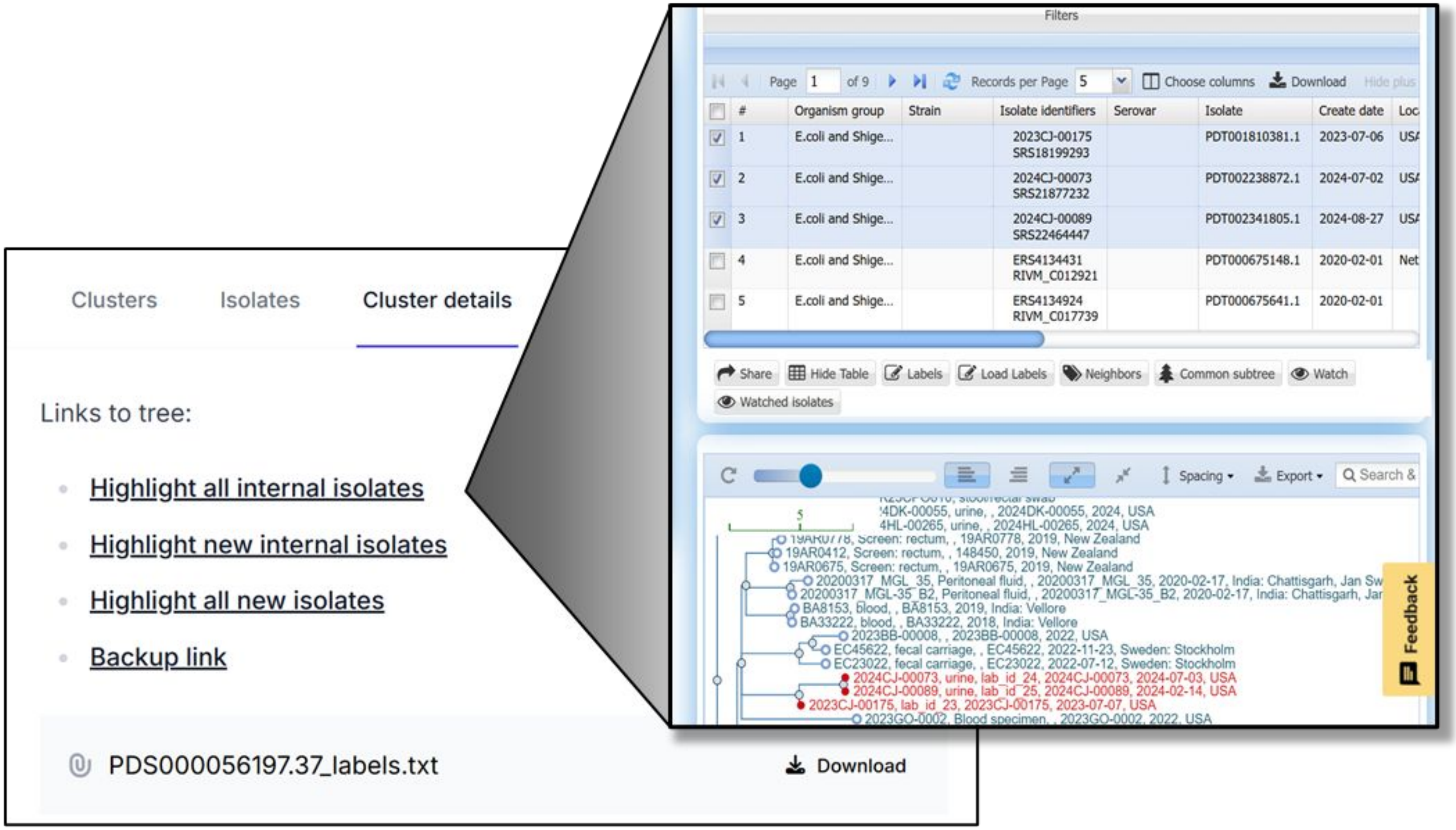
New internal isolates added:

- lab_id_25 / 2024CJ-00089

New external isolates added:

- SAMN39500079 / EC33821
- SAMN39500114 / EC23022

Tree links and labels: Load the Pathogen Detection tree with defined isolates selected and download label file to display custom metadata on the tree.



Links to tree:

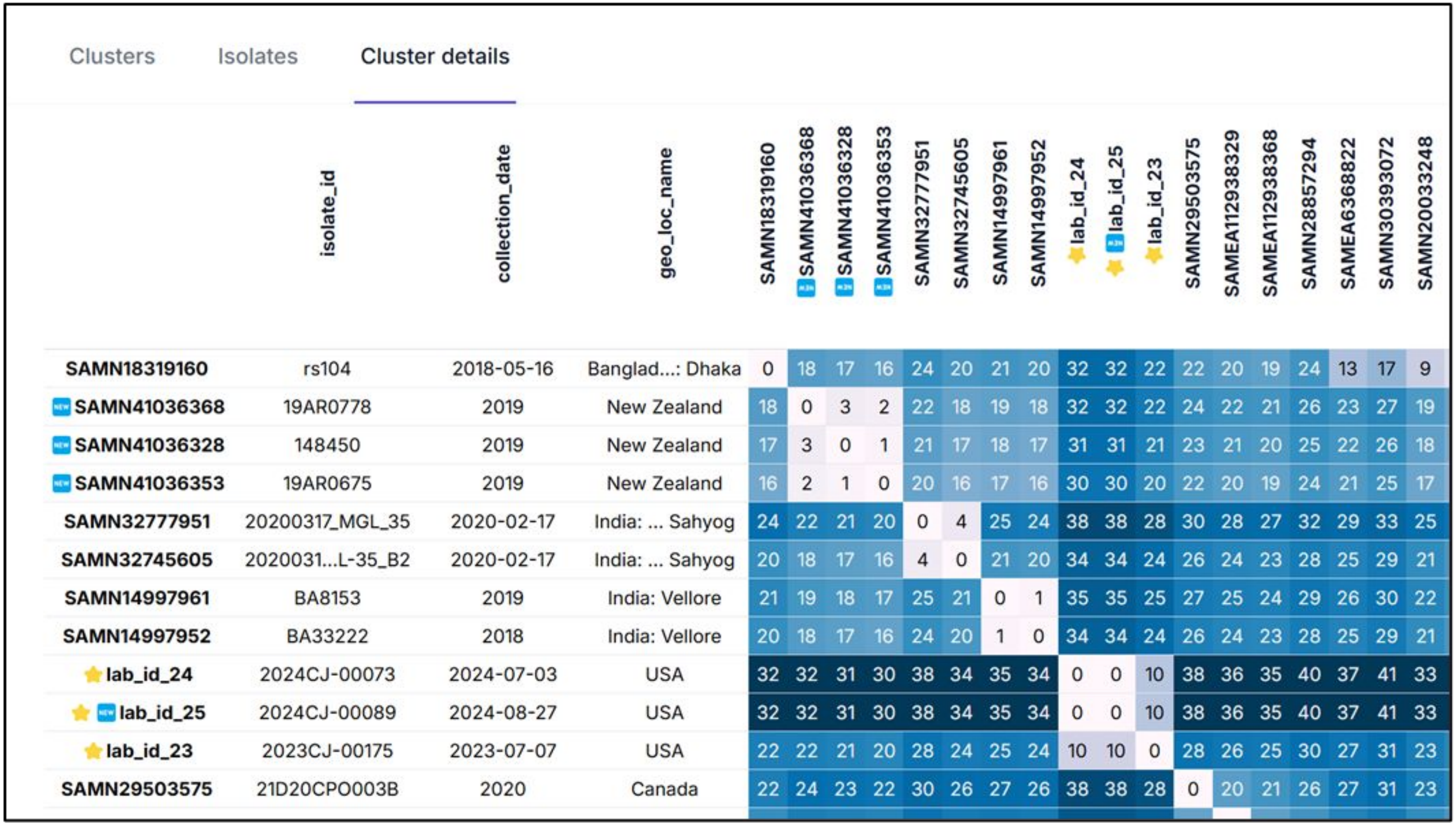
- Highlight all internal isolates
- Highlight new internal isolates
- Highlight all new isolates
- Backup link

PDS000056197.37_labels.txt

	#	Organism group	Strain	Isolate identifiers	Serovar	Isolate	Create date	Loc
☑	1	E.coli and Shige...		2023CJ-00175 SRS18199293		PDT001810381.1	2023-07-06	USA
☑	2	E.coli and Shige...		2024CJ-00073 SRS21877232		PDT002238872.1	2024-07-02	USA
☑	3	E.coli and Shige...		2024CJ-00089 SRS22464447		PDT002341805.1	2024-08-27	USA
□	4	E.coli and Shige...		ERS4134431 RIVM_C012921		PDT000675148.1	2020-02-01	Net
□	5	E.coli and Shige...		ERS4134924 RIVM_C017739		PDT000675641.1	2020-02-01	

```
R23CPO010, stool/rectal swab
24DK-00055, urine, , 2024DK-00055, 2024, USA
4HL-00265, urine, , 2024HL-00265, 2024, USA
19AR0778, Screen: rectum, , 19AR0778, 2019, New Zealand
19AR0412, Screen: rectum, , 148450, 2019, New Zealand
19AR0675, Screen: rectum, , 19AR0675, 2019, New Zealand
20200317_MGL_35, Peritoneal fluid, , 20200317_MGL_35, 2020-02-17, India: Chattisgarh, Jan Sw
20200317_MGL-35_B2, Peritoneal fluid, , 20200317_MGL-35_B2, 2020-02-17, India: Chattisgarh, Jar
BA8153, blood, , BA8153, 2019, India: Vellore
BA33222, blood, , BA33222, 2018, India: Vellore
2023BB-00008, , 2023BB-00008, 2022, USA
EC45622, fecal carriage, , EC45622, 2022-11-23, Sweden: Stockholm
EC23022, fecal carriage, , EC23022, 2022-07-12, Sweden: Stockholm
2024CJ-00073, urine, lab_id_24, 2024CJ-00073, 2024-07-03, USA
2024CJ-00089, urine, lab_id_25, 2024CJ-00089, 2024-02-14, USA
2023CJ-00175, lab_id_23, 2023CJ-00175, 2023-07-07, USA
2023GO-0002, Blood specimen, , 2023GO-0002, 2022, USA
```

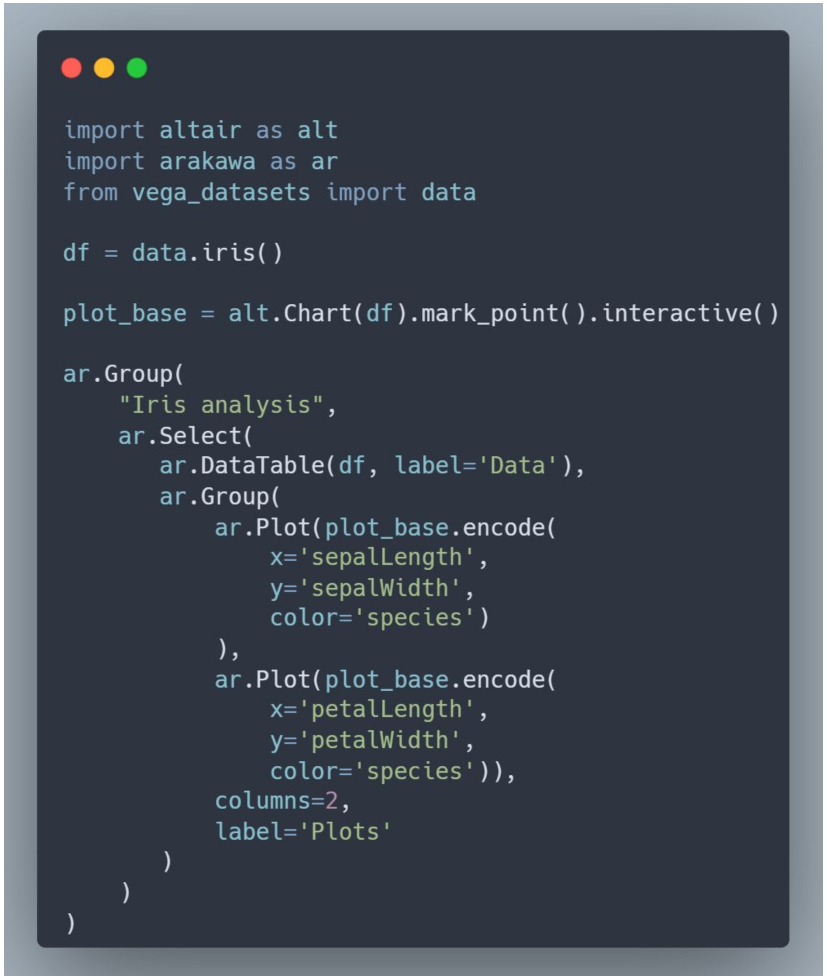
SNP matrix: View pairwise SNP distances between internal and external isolates.



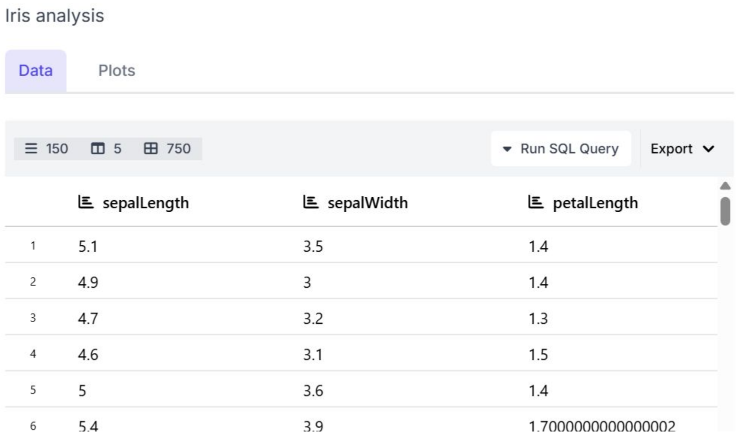
	isolate_id	collection_date	geo_loc_name	SAMN18319160	NEW SAMN41036368	NEW SAMN41036328	NEW SAMN41036353	SAMN32777951	SAMN32745605	SAMN14997961	SAMN14997952	★ lab_id_24	★ NEW lab_id_25	★ lab_id_23	SAMN29503575	SAMEA112938329	SAMEA112938368	SAMN28857294	SAMEA6368822	SAMN30393072	SAMN20033248
SAMN18319160	rs104	2018-05-16	Banglad...: Dhaka	0	18	17	16	24	20	21	20	32	32	22	22	20	19	24	13	17	9
NEW SAMN41036368	19AR0778	2019	New Zealand	18	0	3	2	22	18	19	18	32	32	22	24	22	21	26	23	27	19
NEW SAMN41036328	148450	2019	New Zealand	17	3	0	1	21	17	18	17	31	31	21	23	21	20	25	22	26	18
NEW SAMN41036353	19AR0675	2019	New Zealand	16	2	1	0	20	16	17	16	30	30	20	22	20	19	24	21	25	17
SAMN32777951	20200317_MGL_35	2020-02-17	India: ... Sahyog	24	22	21	20	0	4	25	24	38	38	28	30	28	27	32	29	33	25
SAMN32745605	2020031...L-35_B2	2020-02-17	India: ... Sahyog	20	18	17	16	4	0	21	20	34	34	24	26	24	23	28	25	29	21
SAMN14997961	BA8153	2019	India: Vellore	21	19	18	17	25	21	0	1	35	35	25	27	25	24	29	26	30	22
SAMN14997952	BA33222	2018	India: Vellore	20	18	17	16	24	20	1	0	34	34	24	26	24	23	28	25	29	21
★ lab_id_24	2024CJ-00073	2024-07-03	USA	32	32	31	30	38	34	35	34	0	0	10	38	36	35	40	37	41	33
★ NEW lab_id_25	2024CJ-00089	2024-08-27	USA	32	32	31	30	38	34	35	34	0	0	10	38	36	35	40	37	41	33
★ lab_id_23	2023CJ-00175	2023-07-07	USA	22	22	21	20	28	24	25	24	10	10	0	28	26	25	30	27	31	23
SAMN29503575	21D20CPO003B	2020	Canada	22	24	23	22	30	26	27	26	38	38	28	0	20	21	26	27	31	23

Technical note

Report-generating code written in Python (no HTML/CSS/JavaScript) using Arakawa to create interactive tables and plots within the report

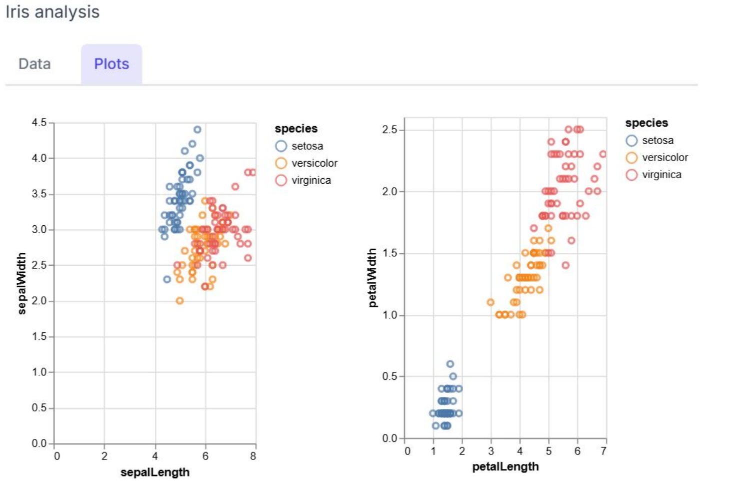


```
import altair as alt
import arakawa as ar
from vega_datasets import data

df = data.iris()

plot_base = alt.Chart(df).mark_point().interactive()

ar.Group(
    "Iris analysis",
    ar.Select(
       ar.DataTable(df, label='Data'),
       ar.Group(
           ar.Plot(plot_base.encode(
               x='sepalLength',
               y='sepalWidth',
               color='species')
           ),
           ar.Plot(plot_base.encode(
               x='petalLength',
               y='petalWidth',
               color='species')),
           columns=2,
           label='Plots'
       )
    )
)
```



Iris analysis

	sepalLength	sepalWidth	petalLength
1	5.1	3.5	1.4
2	4.9	3	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5	3.6	1.4
6	5.4	3.9	1.7000000000000002



Iris analysis

https://ninoseki.github.io/arakawa/latest/blocks/layout-blocks/#group

COLORADO
Department of Public Health & Environment

Limitations of tracking clusters with Pathogen Detection

- Limitations of ncbi-cluster-tracker:
  - Currently does not incorporate AMR results
  - Cannot add notes directly to report (this was easy to do in the Google Sheets system!)
- General limitations with using Pathogen Detection:
  - Clustering distance threshold and timeline are fixed, resulting in sensitivity and specificity issues with detecting potential outbreaks
  - Can be difficult to find useful or actionable information about closely related external isolates
  - Particularly relevant to HAI: inability to detect plasmid clusters to help track the spread of AMR genes

Limitations of tracking clusters with Pathogen Detection

- Limitations of ncbi-cluster-tracker:
  - Currently does not incorporate AMR results
  - Cannot add notes directly to report (this was easy to do in the Google Sheets system!)
- General limitations with using Pathogen Detection:
  - Clustering distance threshold and timeline are fixed, resulting in sensitivity and specificity issues with detecting potential outbreaks
  - Can be difficult to find useful or actionable information about closely related external isolates
  - Particularly relevant to HAI: inability to detect plasmid clusters to help track the spread of AMR genes

Overall: Clusters identified through ncbi-cluster-tracker and Pathogen Detection can be a good *starting point* (hypothesis generation), and can help confirm existing links identified through epidemiology, but require follow-up analysis and investigation

Future directions

- Automated querying of Pathogen Detection without needing BigQuery access
- Include AMR genes
- Ability to add notes to clusters / isolates
- Include annotated tree directly within report
- Incorporate references to any literature that exists about external isolates and clusters
- Provide more options to configure analysis
- Easier installation and usage (Pip/Docker options and cross-platform support)

THANKS!

Documentation and source code can be found here:

github.com/CDPHE-bioinformatics/ncbi-cluster-tracker

Feature requests, bug reports, and pull requests are welcomed!

More questions?

sam.baird@state.co.us
cdphe_bioinformatics_lab@state.co.us

COLORADO
Department of Public
Health & Environment