Food Safety and Inspection Service

Utilization of NCBI Pathogen Detection Tool in USDA FSIS

Glenn Tillman, Ph.D.
Branch Chief Microbiology: Characterization Branch
FSIS Office of Public Health and Science
Glenn.tillman@fsis.usda.gov

Background

- Whole Genome Sequencing (WGS) helps us identify FSIS isolates with clinical matches or matches within an establishment over time
- NCBI's pathogen detection pipeline assigns all isolates submitted to certain Bioprojects to SNP clusters if there are other isolates within 50 SNPs
- Primary FSIS purpose: compare isolate with clinical isolates or other isolates from the same facility
- Secondary FSIS purpose: identify antibiotic resistance (AMR) genes of interest in FSIS and other isolates

NCBI Pathogen Browser: Use Case 1, Sampling Periods

Primary purpose: compare isolate with clinical isolates or other isolates from the same facility

- What to Look for:
  - Inclusion in a SNP Cluster and Accession Number of the Cluster
  - Number of SNPs between like sample source (Isolation Type), between a different sample source
  - Are there additional isolates from same product type clustering (persistent strains vs. outbreak)?
- https://www.ncbi.nlm.nih.gov/pathogens/
- Input isolates to search:
  - FSIS11704798 FSIS21720508 FSIS11705677 FSIS21720844 FSIS11808152
- Choose specific nodes or isolates to show SNP Distances

NCBI Pathogen Browser: Use Case 2, persistence in a facility (harborage)

Primary purpose: compare isolate with clinical isolates or other isolates from the same facility

- What to Look for:
  - Inclusion in a SNP Cluster and Accession Number of the Cluster
  - Number of SNPs between like sample source (Isolation Type), between a different sample source
  - Are there additional isolates from same product type clustering (persistent strains vs. outbreak)?
- https://www.ncbi.nlm.nih.gov/pathogens/
- Input isolates to search:
  - FSIS21821911 FSIS21821253 FSIS21821254
- Choose specific nodes or isolates to show SNP Distances

NCBI Pathogen Browser: Use Case 3, tracking of AMR genotypes

Secondary purpose: identify antibiotic resistance (AMR) genes of interest in FSIS and other isolates

- https://www.ncbi.nlm.nih.gov/pathogens/
- Things to Look for:
  - Inclusion in a SNP Cluster and Accession Number of the Cluster
  - Does the AMR gene appear to be a distinct, homogenous cluster?
  - Do we see the gene(s) in variety of species, serotypes, etc?
- Input isolates to search:
  - FSIS1502961 FSIS11705716 FSIS31800962
- Search by a specific gene to look across sources
  - Example, AMR_genotype: blaCTX-M-65
  - Example, AMR_genotype: blaCTX-M-32
- Examine a SNP cluster more in-depth
  - PDS000003955.237

NCBI Pathogen Browser: Use Case 4, combining NCBI Pathogen Browser with Tableau visualization

- The NCBI pathogen detection pipeline is an excellent tool but there are some drawbacks to using it alone
  - There is limited metadata on NCBI especially for clinical isolates, it's ideal to have both full metadata and SNP cluster information in one place
  - To determine all of the isolates within a certain SNP range of a given set of isolates can be challenging for multiple trees and isolates
  - Geography of isolate collection can support phylogenetic trees in an investigation

NCBI Pathogen Browser: Use Case 4, combining NCBI Pathogen Browser with Tableau visualization

- Use sql query or other methods to identify isolates of interest based on collection period, serotype, or other criteria, this will serve as script input

- Two files are downloaded from NCBI's pathogen ftp site using a Bash script
  - From the Metadata folder (PDGXXX.metadata.tsv)
    - This provides link between FSIS ID and PDT accession
  - From the Cluster folder (PDGXXX.reference_target_all_isolate.tsv)
    - This provides minimum same and minimum diff columns based PDT accession

NCBI Pathogen Browser: Use Case 4, combining NCBI Pathogen Browser with Tableau visualization

- The script looks for isolates from the input file in the metadata sheet and PDGXXX.reference_target_all_isolate.tsv file

- The script extracts relevant columns (strain, PDT accession, min-diff, snp cluster accession) from files listed in previous step and joins them in a single file

Target_acc	Strain	min_same	min_diff	PDS_acc
PDT000253392.1	FSIS11704798	2	4	PDS000002757.373
PDT000260708.1	FSIS21720508	4	6	PDS000003955.220
PDT000266427.1	FSIS11705677	20	15	PDS000001946.273
PDT000277118.1	FSIS21720844	3	0	PDS000002757.373
PDT000293348.1	FSIS11808152	3	5	PDS000002757.373

- The joined file is then filtered based on min-diff or min-same column, and subsequently used for the 2nd part of the script

NCBI Pathogen Browser: Use Case 4, combining NCBI Pathogen Browser with Tableau visualization

The latest version of the SNP cluster listed is downloaded and unzipped to obtain the newick tree file

Target_acc	Strain	min_same	min_diff	PDS_acc
PDT000253392.1	FSIS11704798	2	4	PDS000002757.373
PDT000260708.1	FSIS21720508	4	6	PDS000003955.220
PDT000277118.1	FSIS21720844	3	0	PDS000002757.373
PDT000293348.1	FSIS11808152	3	5	PDS000002757.373

For each strain and tree combination, the newick file is parsed by a python script to determine the patristic distance between the isolate of interest and every other isolate on the tree to generate a set of pairwise distances

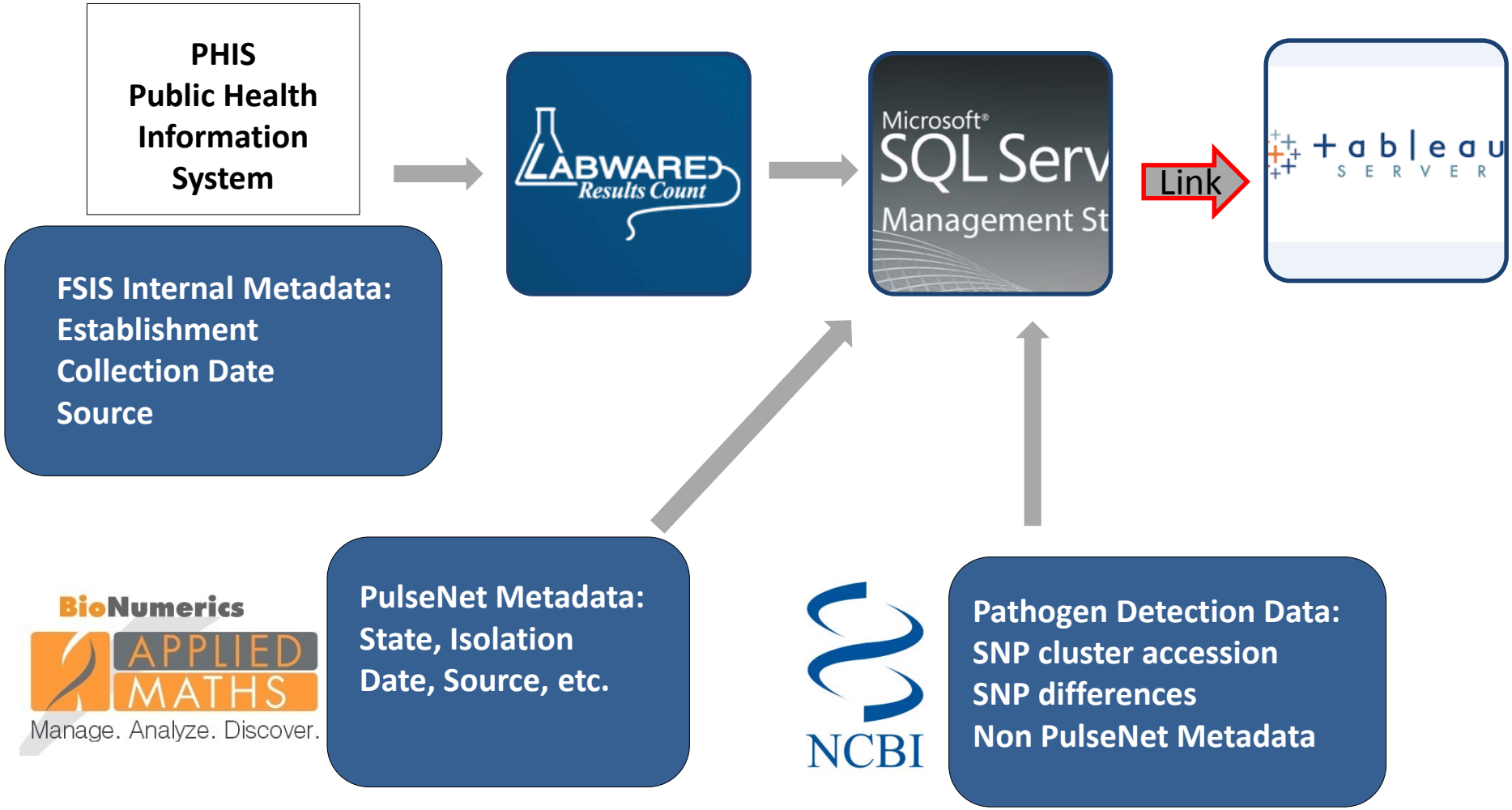
NCBI Pathogen Browser: Use Case 4, combining NCBI Pathogen Browser with Tableau visualization

Strain	Compared_Isolate	SNP_difference
FSIS11704798	FSIS11704798	0
FSIS11704798	FSIS21720397	2
FSIS11704798	FSIS31800240	2
FSIS11704798	FSIS11704795	3
FSIS11704798	FSIS11704913	3
FSIS11704798	PNUSAS028465	4
FSIS11704798	PNUSAS029421	4
FSIS11704798	PNUSAS038562	8
FSIS11704798	PNUSAS039081	8
FSIS11704798	PNUSAS038342	8
FSIS11704798	PNUSAS039668	8
FSIS11704798	PNUSAS041708	8
FSIS11704798	FSIS1608489	9

Compared isolates are then used as a query to obtain extended metadata such as establishment number, exact collection dates, and locations for clinical isolates

Extended metadata and SNP differences are used as a data source to view the information in Tableau

NCBI Pathogen Browser: Use Case 4, combining NCBI Pathogen Browser with Tableau visualization

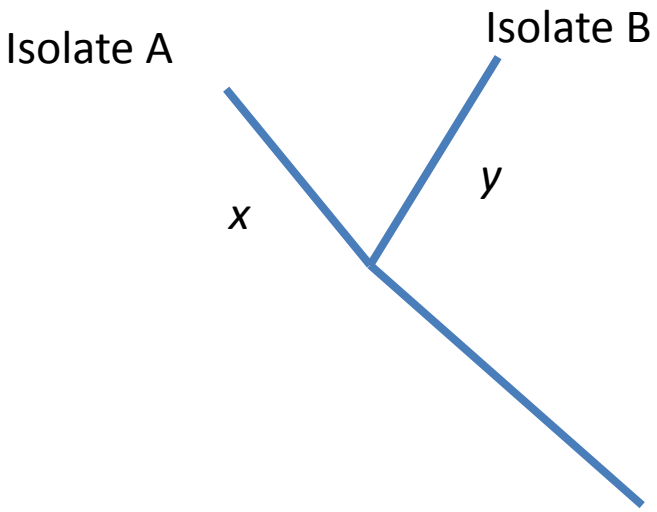


Live Demo of Tableau Visualization for *Listeria monocytogenes* isolates from USDA-FSIS collected Samples

https://tableau.fsis.usda.gov/#/site/OPHS/workbooks/210/views

Supplemental Slide showing parsing using patristic distance

For each strain and tree combination, the newick file is parsed by a python script to determine the patristic distance between the isolate of interest and every other isolate on the tree to generate a set of pairwise distances



Patristic distance AB = $x + y$

A patristic distance is the sum of the lengths of the branches that link two nodes in a tree, where those nodes are typically represent extant gene sequences or species