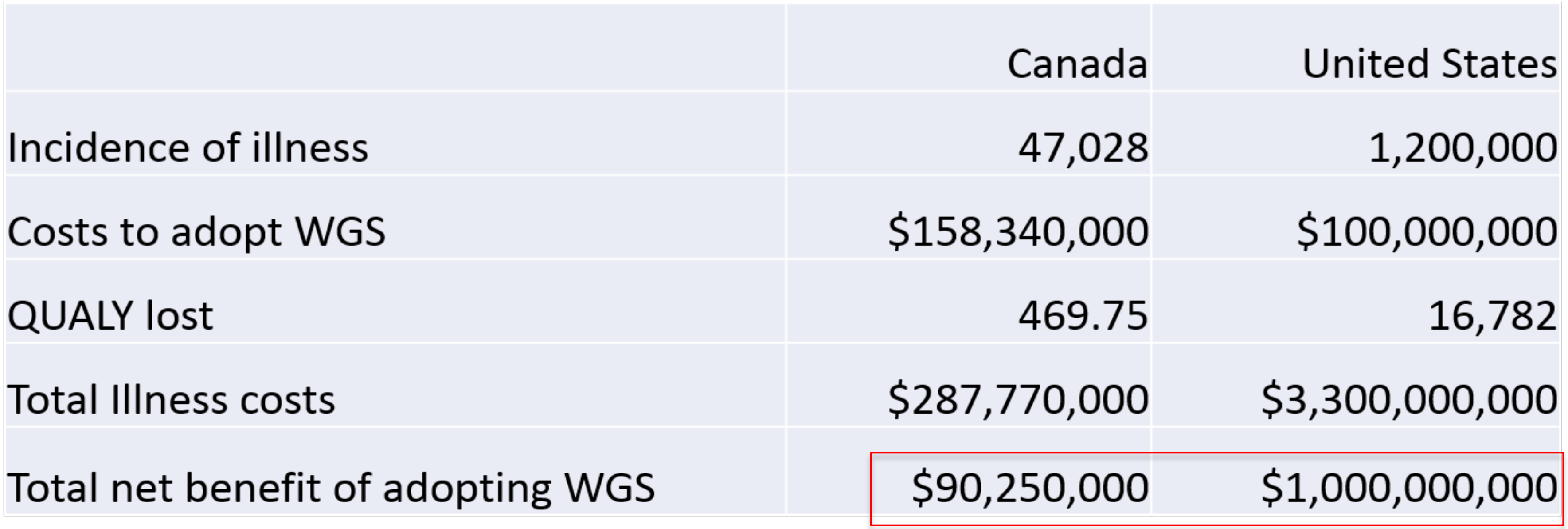
All for one and one for all: the true potential of whole-genome sequencing.

Marc Allard, Ph.D.

Center for Food Safety and Applied Nutrition,
US Food and Drug Administration

GLOBAL WATER FOOD SAFETY SUMMIT
College Park, MD, November 2019

Economic Analysis of *Salmonella* outbreaks annual reduction using WGS

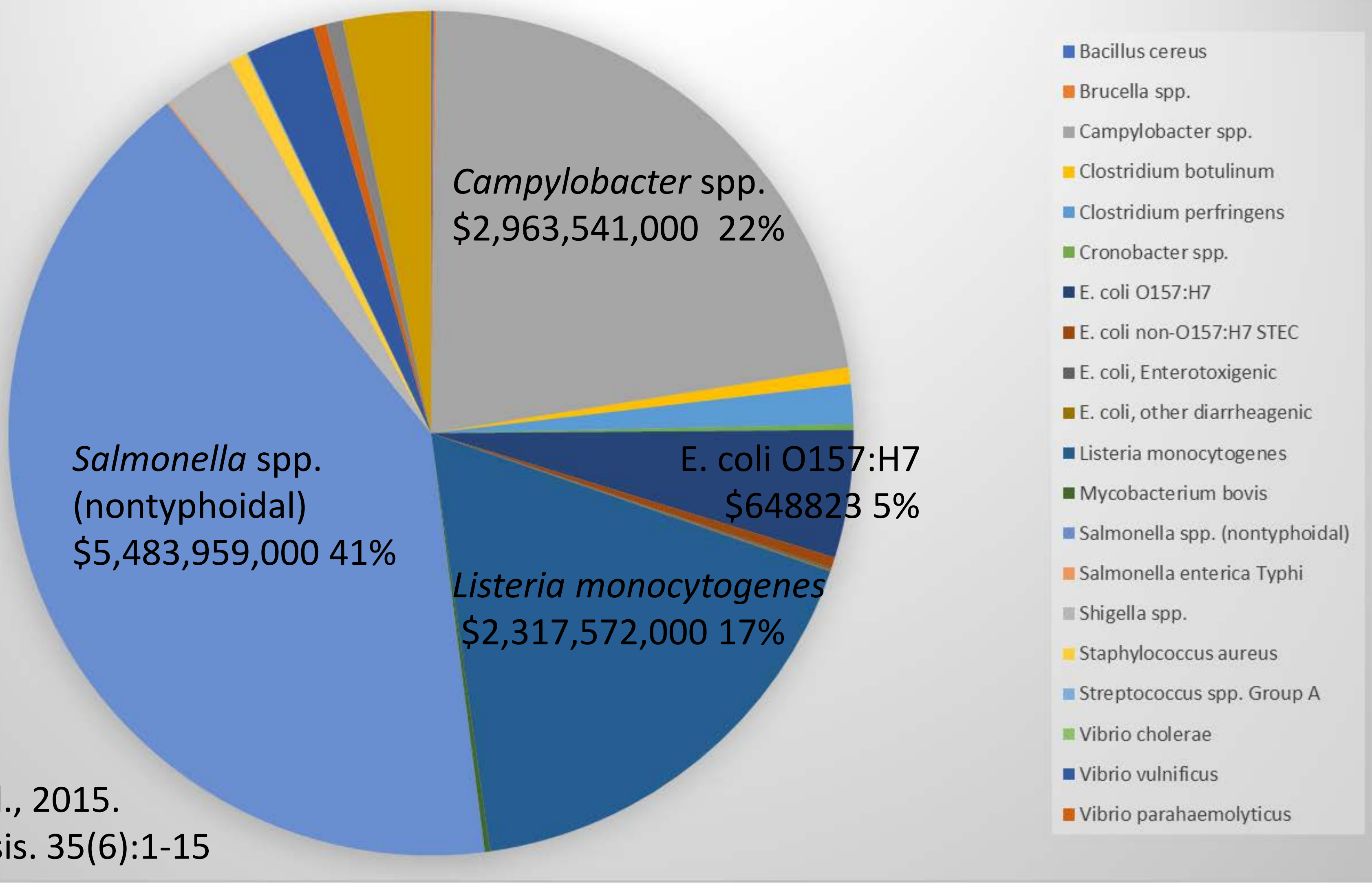


	Canada	United States
Incidence of illness	47,028	1,200,000
Costs to adopt WGS	$158,340,000	$100,000,000
QUALY lost	469.75	16,782
Total Illness costs	$287,770,000	$3,300,000,000
Total net benefit of adopting WGS	$90,250,000	$1,000,000,000

Model assumes up to 70% reduction in numbers of illnesses due to WGS implementation. Benefits gained due to earlier detection and decreased time to recall food items. United States estimates are adjusted based on increase population size. Additional analysis is needed to adjust to US illnesses and US health care costs.

Jain et al., 2019. Food Res. Int. 116: 802-809

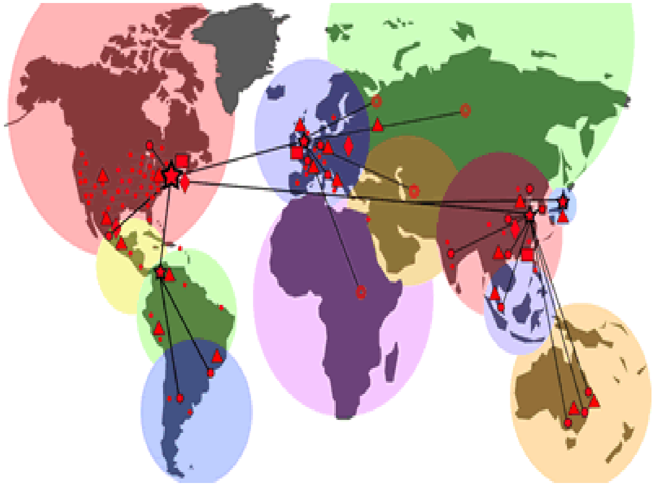
Monetary Loss from Bacteria Foodborne Illness Total $13,279,603,000



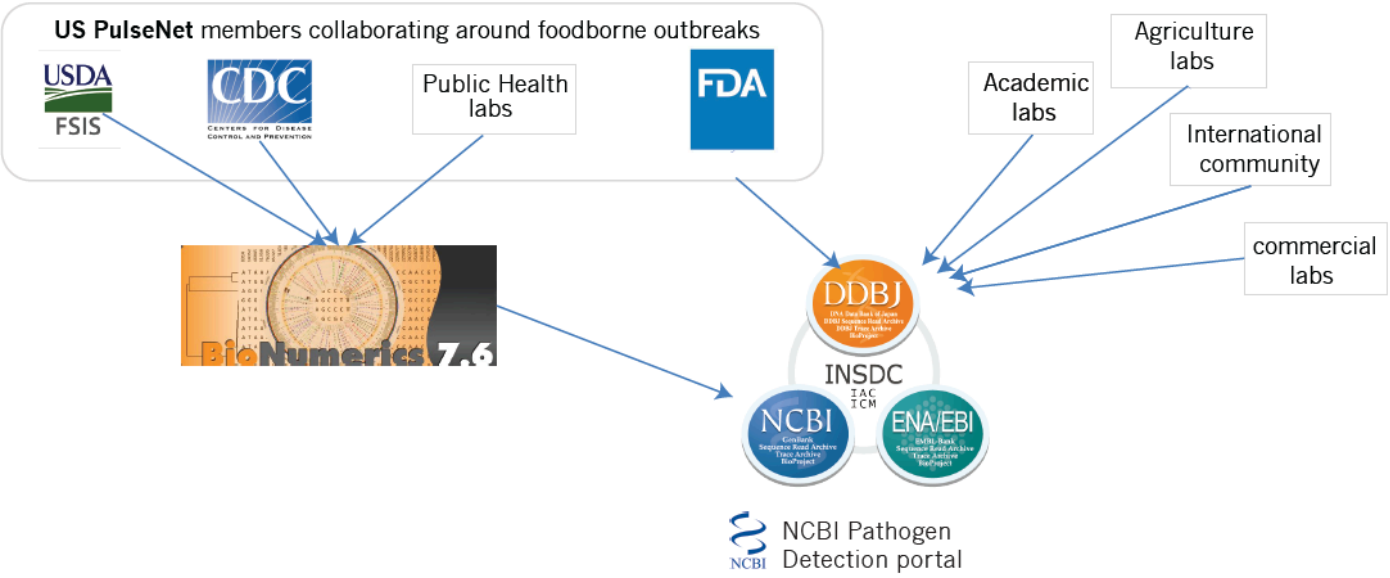
Minor et al., 2015.
Risk Analysis. 35(6):1-15

Why Develop a WGS based Network?

- Tracking and Tracing of food pathogens
  - Insufficient resolution of current tools
    - -matching clinical to environmental
    - -improve the environmental database
  - Faster identification of the food involved in the outbreak
  - Limited number of investigators vs. facilities and import lines
  - Global travel
  - Global food supply



Independent submission model:



US **PulseNet** members collaborating around foodborne outbreaks

USDA FSIS

CDC

Public Health labs

FDA

BioNumerics 7.6

Academic labs

Agriculture labs

International community

commercial labs

DDBJ

INSDC

NCBI

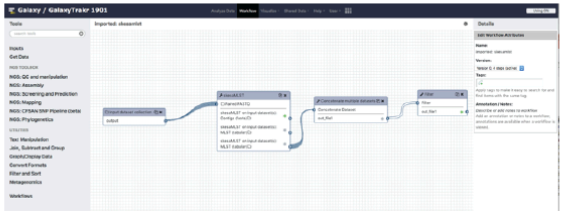
ENA/EBI

NCBI Pathogen Detection portal

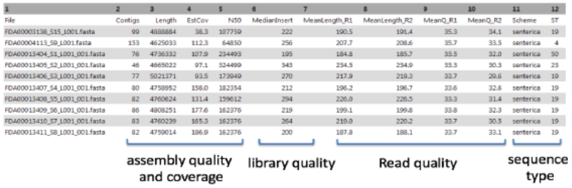
Best practices document, Sept. 2019:

1. Check the quality of your entire run:

GalaxyTrakr, MicroRunQC workflow:



QC report for entire MiSeq run:



1	2	3	4	5	6	7	8	9	10	11	12
File	Contigs	Length	EstCov	N50	MedianInsert	MeanLength_R1	MeanLength_R2	MeanQ_R1	MeanQ_R2	Scheme	ST
FDA00003138_S15_L001.fasta	99	4888884	38.3	107759	222	190.5	191.4	35.3	34.1	senterica	19
FDA00004113_S9_L001.fasta	153	4625033	112.3	64850	256	207.7	208.6	35.7	33.5	senterica	4
FDA00013404_S1_L001_001.fasta	76	4736332	107.9	234493	195	184.8	185.7	33.5	32.0	senterica	50
FDA00013405_S2_L001_001.fasta	46	4665022	97.1	324499	343	234.5	234.9	33.3	30.3	senterica	23
FDA00013406_S3_L001_001.fasta	77	5021371	93.5	173949	270	217.9	219.3	33.7	29.6	senterica	19
FDA00013407_S4_L001_001.fasta	80	4758952	158.0	182354	212	196.2	196.7	33.6	32.6	senterica	19
FDA00013408_S5_L001_001.fasta	82	4760624	131.4	159612	294	226.0	226.5	33.3	31.4	senterica	19
FDA00013409_S6_L001_001.fasta	86	4808251	177.6	162376	219	199.1	199.8	33.8	32.3	senterica	19
FDA00013410_S7_L001_001.fasta	83	4760239	165.3	162376	264	219.0	220.2	33.7	30.5	senterica	19
FDA00013411_S8_L001_001.fasta	82	4759014	186.9	162376	200	187.8	188.1	33.7	33.1	senterica	19

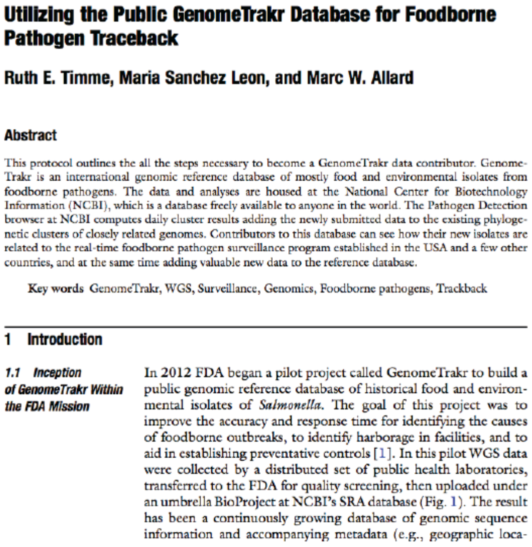
assembly quality and coverage

library quality

Read quality

sequence type

2. Detailed protocol for direct NCBI submissions:



Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback

Ruth E. Timme, Maria Sanchez Leon, and Marc W. Allard

Abstract

This protocol outlines the all the steps necessary to become a GenomeTrakr data contributor. GenomeTrakr is an international genomic reference database of mostly food and environmental isolates from foodborne pathogens. The data and analyses are housed at the National Center for Biotechnology Information (NCBI), which is a database freely available to anyone in the world. The Pathogen Detection browser at NCBI computes daily cluster results adding the newly submitted data to the existing phylogenetic clusters of closely related genomes. Contributors to this database can see how their new isolates are related to the real-time foodborne pathogen surveillance program established in the USA and a few other countries, and at the same time adding valuable new data to the reference database.

Key words GenomeTrakr, WGS, Surveillance, Genomics, Foodborne pathogens, Trackback

1 Introduction

1.1 Inception of GenomeTrakr Within the FDA Mission

In 2012 FDA began a pilot project called GenomeTrakr to build a public genomic reference database of historical food and environmental isolates of *Salmonella*. The goal of this project was to improve the accuracy and response time for identifying the causes of foodborne outbreaks, to identify harborage in facilities, and to aid in establishing preventative controls [1]. In this pilot WGS data were collected by a distributed set of public health laboratories, transferred to the FDA for quality screening, then uploaded under an umbrella BioProject at NCBI's SRA database (Fig. 1). The result has been a continuously growing database of genomic sequence information and accompanying metadata (e.g., geographic loca-

3. Data curation and data management recommendations:

- Importance of data curation.

- Where are my data? How do I update/retract records in the following NCBI databases?
  - BioProject
  - BioSample
  - SRA
  - GenBank
  - NCBI Pathogen Detection portal

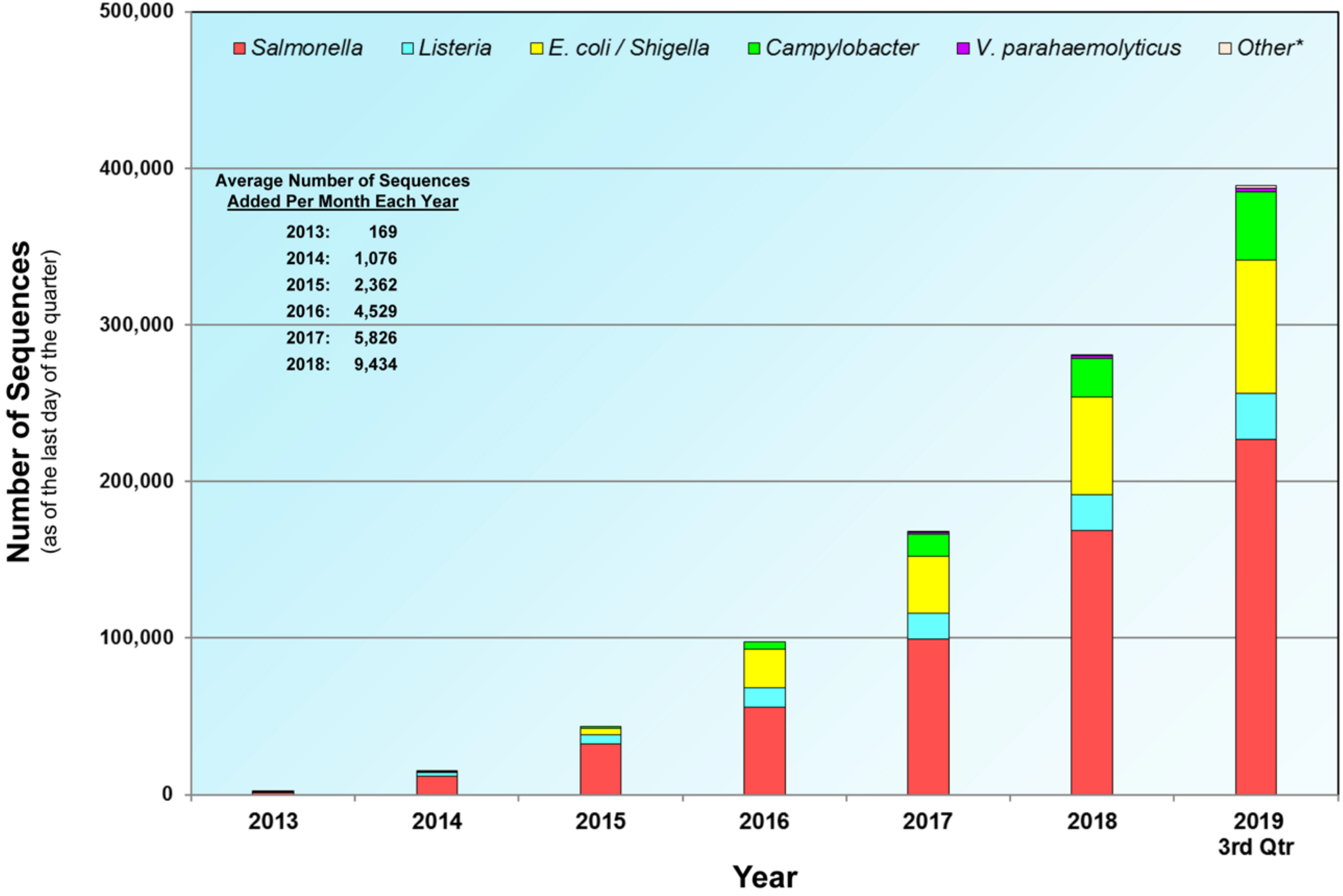
- Recommendations for managing your data, submissions, and accessions.

References:

Timme RE, Sanchez Leon M, Allard MW 2019. Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. Methods in molecular biology (Clifton, N.J.) 1918:201–212.

Daniel Benisatto et al. In review. GalaxyTrakr: A Distributed Analysis Tool for Public Health Whole Genome Sequence Data Accessible to Non-Bioinformaticans.

Total Number of Sequences in the GenomeTrakr Database



Average Number of Sequences Added Per Month Each Year

2013:	169
2014:	1,076
2015:	2,362
2016:	4,529
2017:	5,826
2018:	9,434

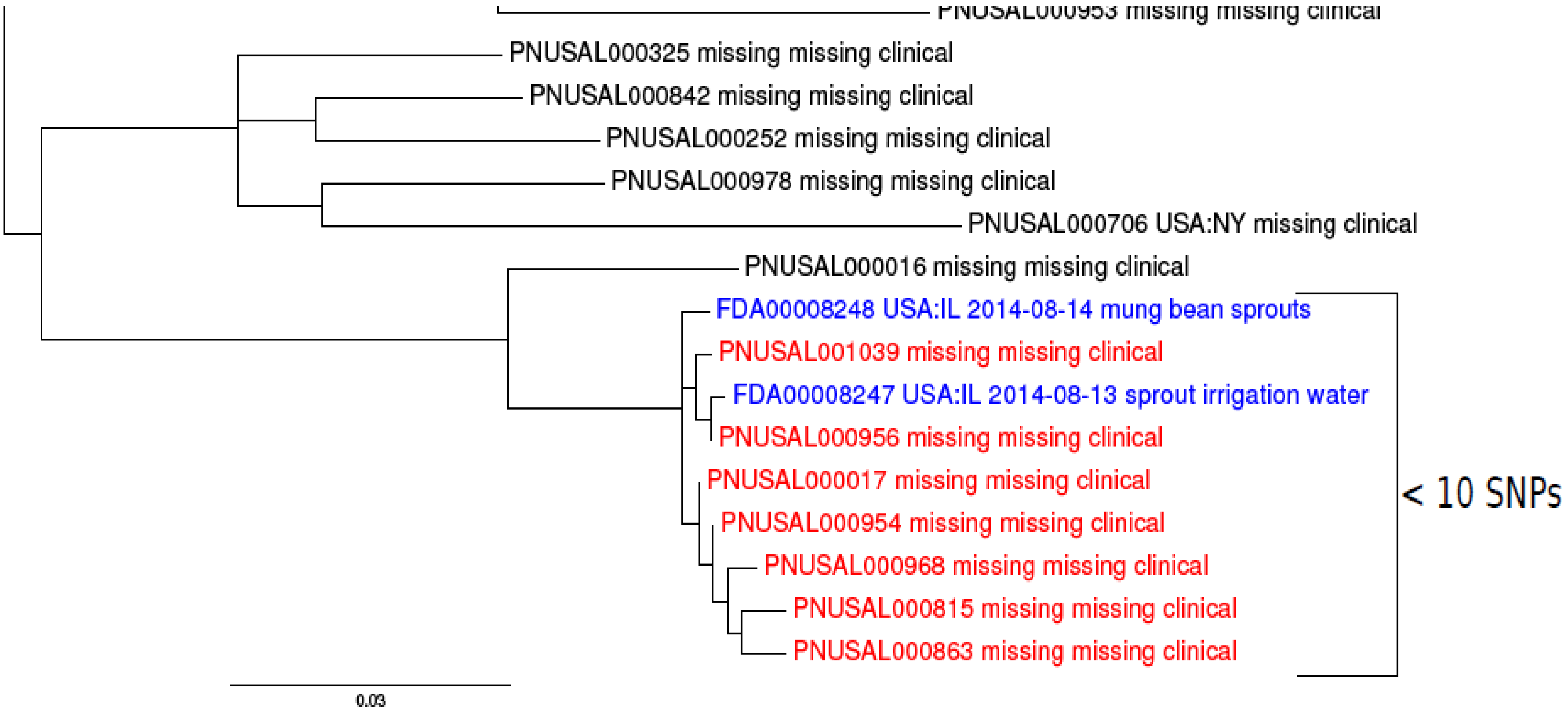
First sequences uploaded in February 2013

* Other pathogens: *Cronobacter, V. vulnificus, C. botulinum, and C. perfringens*

September 30, 2019 GenomeTrakr Numbers

Species	**Total Isolates**
Salmonella enterica	277,051
E.coli and Shigella	85,157
Campylobacter jejuni	43,705
Listeria monocytogenes	28,978
Vibrio parahaemolyticus	2,519
Cronobacter	636
Vibrio vulnificus	359
Clostridium botulinum	298
Clostridium perfringens	242
Total	**388,945**

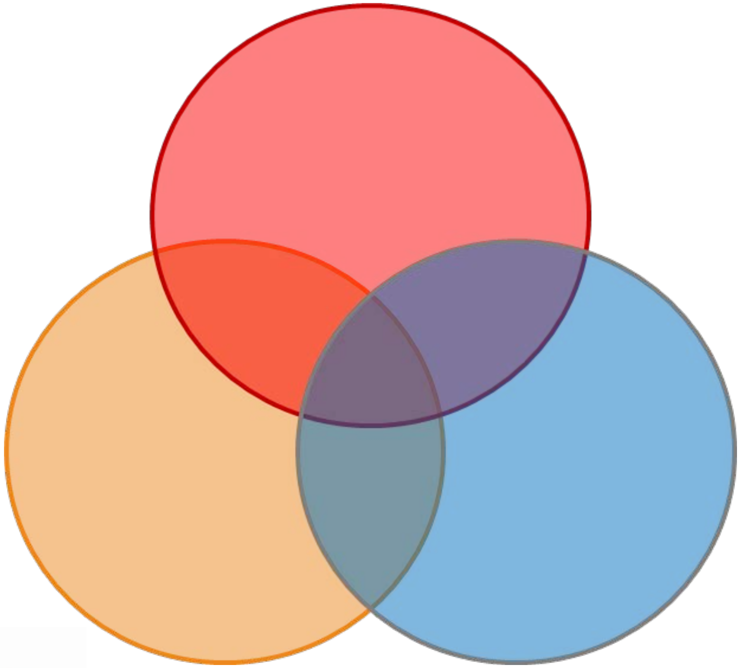
How do we use the GenomeTrakr information? Example of Listeria in sprouts using a phylogenetic perspective.

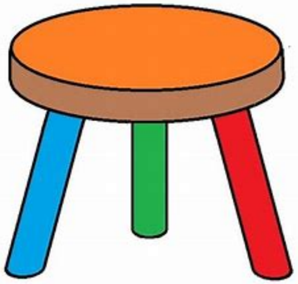


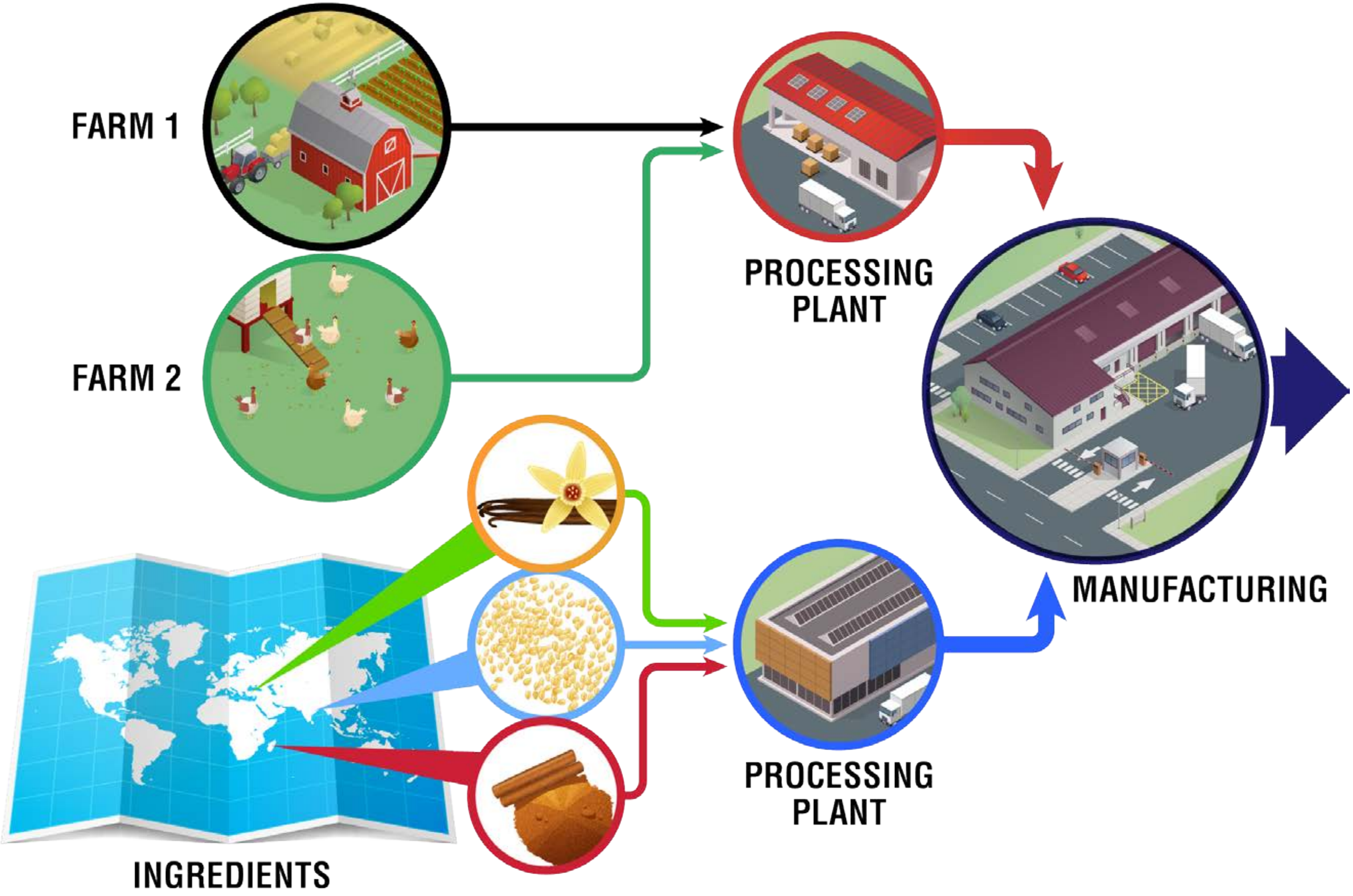
Identifying an Outbreak Vehicle: Lines of Evidence

Three types of evidence used:

- **Epidemiologic:** association between illness and food exposure
- **Traceback:** suspected food item links back to a common source of contamination
- **Microbiologic/laboratory:** pathogen found in the food, farm or facility







FARM 1

FARM 2

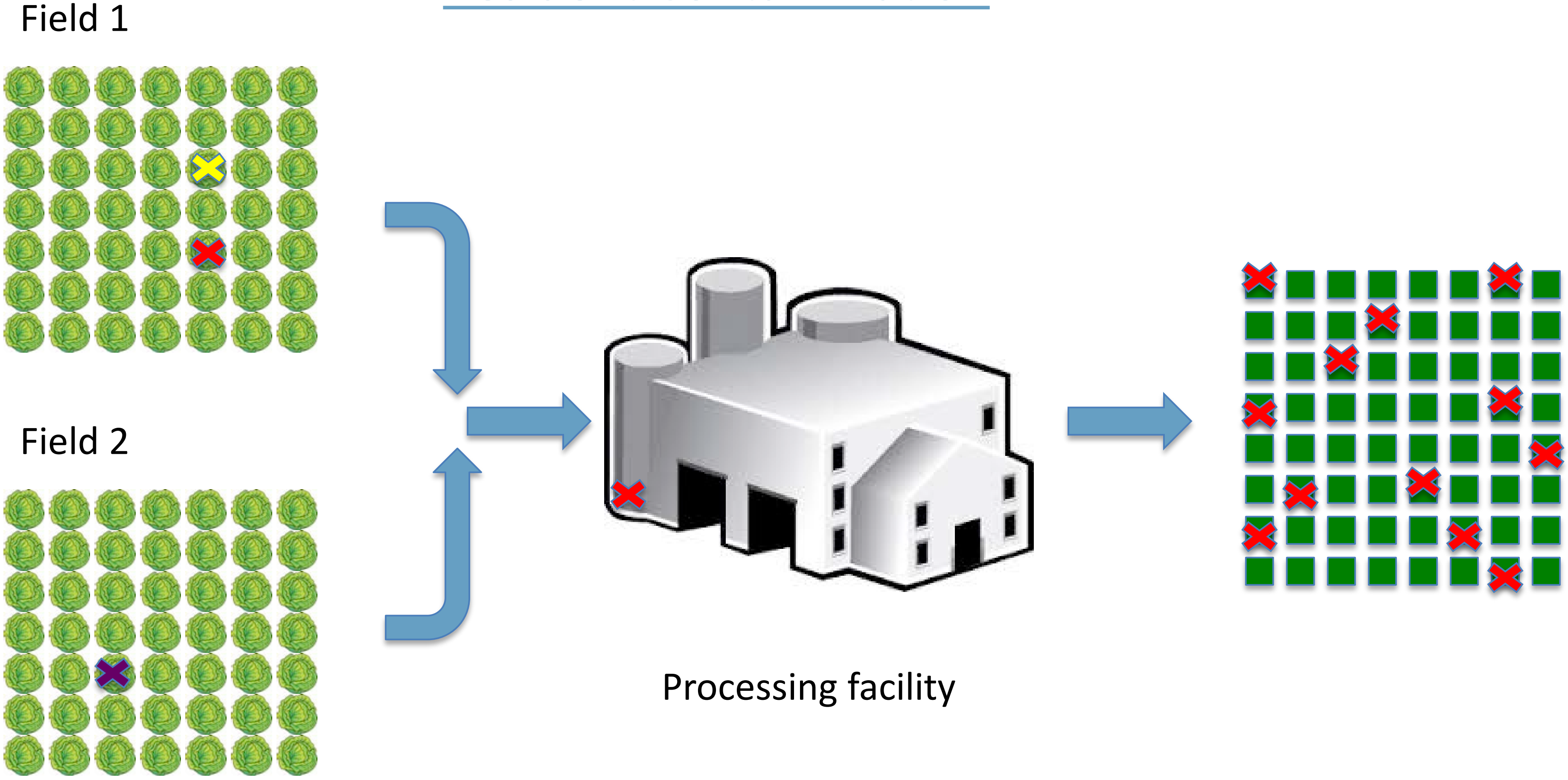
PROCESSING
PLANT

MANUFACTURING

PROCESSING
PLANT

INGREDIENTS

Environmental or Resident Contamination



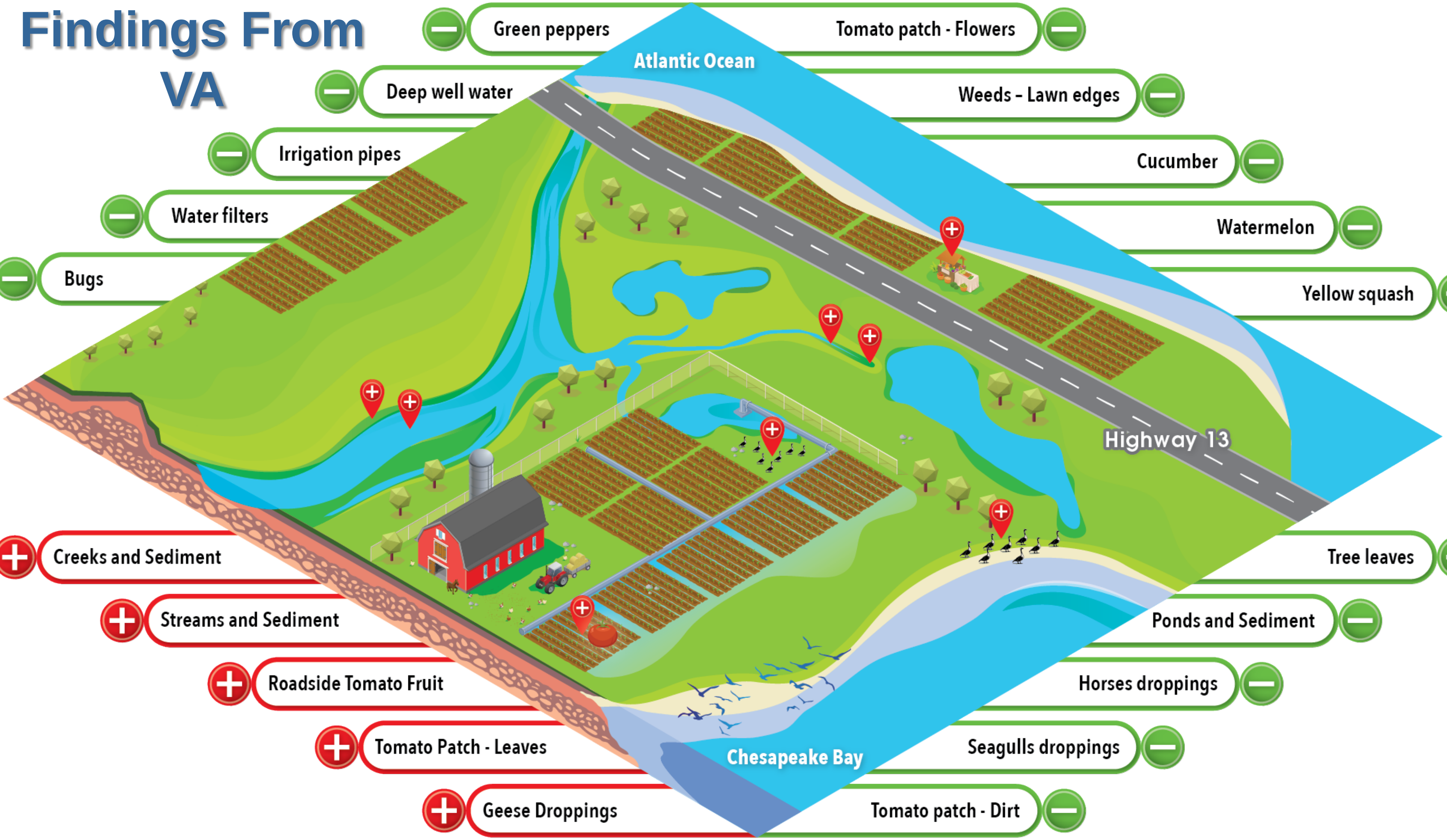
Field 1

Field 2

Processing facility

U.S. FOOD & DRUG
ADMINISTRATION

Findings From VA



- ⊖ Green peppers
- ⊖ Deep well water
- ⊖ Irrigation pipes
- ⊖ Water filters
- ⊖ Bugs
- ⊖ Tomato patch - Flowers
- ⊖ Weeds - Lawn edges
- ⊖ Cucumber
- ⊖ Watermelon
- ⊖ Yellow squash
- ⊕ Creeks and Sediment
- ⊕ Streams and Sediment
- ⊕ Roadside Tomato Fruit
- ⊕ Tomato Patch - Leaves
- ⊕ Geese Droppings
- ⊖ Tree leaves
- ⊖ Ponds and Sediment
- ⊖ Horses droppings
- ⊖ Seagulls droppings
- ⊖ Tomato patch - Dirt

Collected over 500 isolates:

2010-2011: **237** 2012: **137** 2014: **133+**

Key Conclusions from VA surveys Identify Water, Water, and Water

- Sampled Areas:
  - Virginia Tech Agricultural Research and Education Center (AREC): 2009-2011
  - 6 environmental waters and sediments: 2011, 2012, 2014-2015
- *Salmonella* Pattern 61 and other clinically relevant isolates found consistently in waters and sediments at AREC and the 6 other locations
- Environmental waters and sediments are potential reservoirs for *Salmonella*
- Enhanced agricultural practices related to the use of surface waters are important to mitigating *Salmonella* transfer on to crop plants

Bell RL et al. Ecological prevalence, genetic diversity, and epidemiological aspects of Salmonella isolated from tomato agricultural regions of the Virginia Eastern Shore. Front Microbiol. 2015 May 7;6:415.



South American Salmonella Environmentals

Global Water & Food Safety Summit

21 November 2019

Maria Sanchez
James Pettengill
Hugh Rand

•Serotypes Found

45	Infantis
26	Typhimurium
15	IIIb 48:i:z
14	Anatum
12	Enteritidis
10	Paratyphi B
10	Brandenburg
7	I 4[5]12:i:-
4	Rissen
4	Muenchen
4	Mbandaka
4	Manhattan
3	IIIb 58:k:z
2	Johannesburg
2	IIIb 61:i:z
1	-:z10:enz15
1	Senftenberg
1	Predicted Serotype
1	IV 43:z4z23:-
1	Dublin
1	Cerro
1	18:k:z

Cluster Characteristics (N= 28)

	SNPcluster	Water Isolates	Cluster Size
1	PDS000053260.2	2	3
2	PDS000053263.2	2	3
3	PDS000053276.1	4	5
4	PDS000053277.1	2	2
5	PDS000053278.1	2	2
6	PDS000053279.1	3	3
7	PDS000053280.1	2	2
8	PDS000053281.1	4	5
9	PDS000053282.1	4	4
10	PDS000053283.1	3	3
11	PDS000003955.531	10	4330
12	PDS000053285.1	6	4
13	PDS000053286.1	5	5
14	PDS000053288.1	2	2

Cluster Characteristics (N= 28)

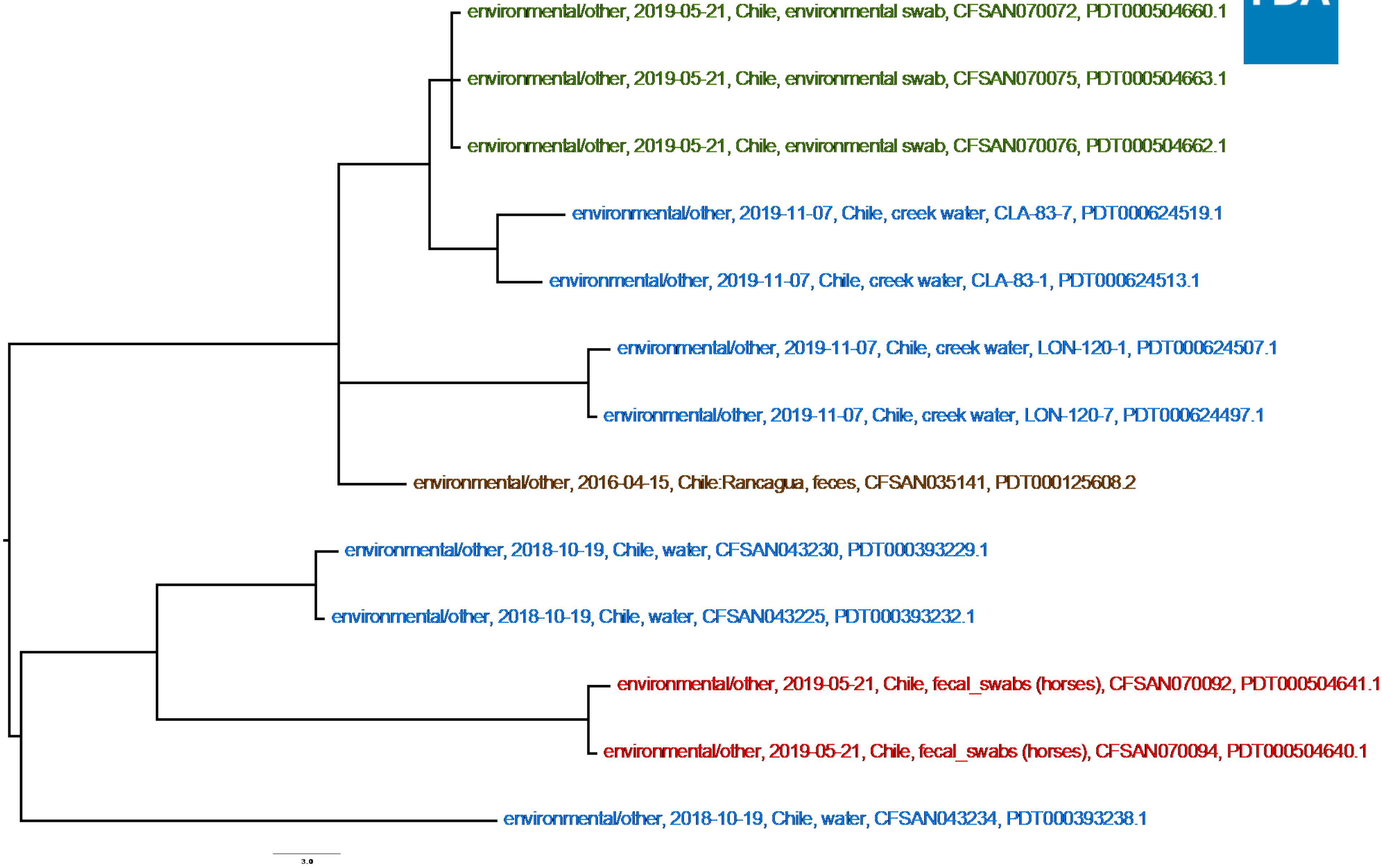
	SNPcluster	Water Isolates	Cluster Size
15	PDS000053289.1	2	2
16	PDS000053290.1	2	2
17	PDS000053291.1	6	7
18	PDS000053292.1	3	3
19	PDS000053293.1	2	2
20	PDS000053294.1	2	2
21	PDS000026867.7	3	55
22	PDS000026991.4	2	7
23	PDS000029248.7	2	24
24	PDS000029912.2	6	8
25	PDS000030058.41	1	89
26	PDS000032393.134	23	861
27	PDS000037691.3	4	13
28	PDS000051343.30	4	2009

Geographic Overlap

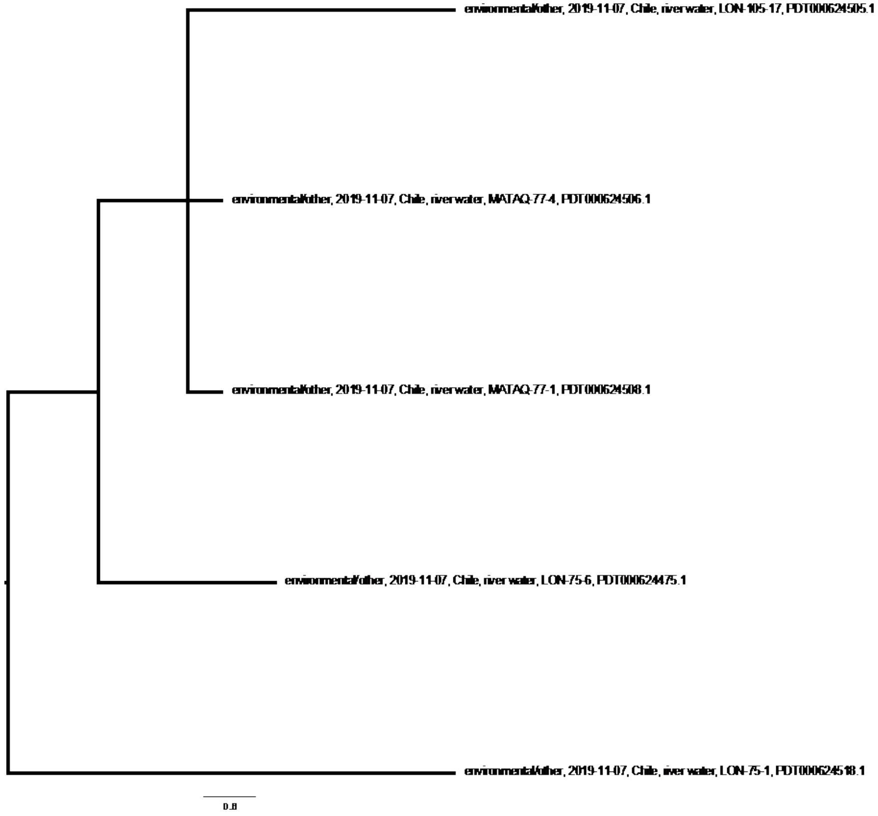
SNP Cluster	Brazil	Canada	Chile	Colombia	Ecuador	Germany	Guyana	Indonesia	Italy	Mexico	not collected	Paraguay	Peru	Suriname	Switzerland	United Kingdom	Missing	USA	Venezuela	Viet Nam
PDS000003955		1	38	56	208				1				193		3	69	15	3754	2	1
PDS000026867			51															4		
PDS000026991			2															5		
PDS000029248	4		3													17	1			
PDS000029912	2		6																	
PDS000030058	1		1							1		3				5		78		
PDS000032393		5	28					1		3	2			1		63	8	763		
PDS000037691			13																	
PDS000051343		45	4		1	9	3				11				1	58	58	1832		
PDS000053260			3																	
PDS000053263			2															1		
PDS000053276			3																	
PDS000053277			2																	
PDS000053278			2																	
PDS000053279			3																	
PDS000053280			2																	
PDS000053281			4															1		
PDS000053282			3																	
PDS000053283			3																	
PDS000053285			2																	
PDS000053286			5																	
PDS000053288			2																	
PDS000053289			2																	
PDS000053290			2																	
PDS000053291			6															1		
PDS000053292			3																	
PDS000053293			2																	
PDS000053294			2																	

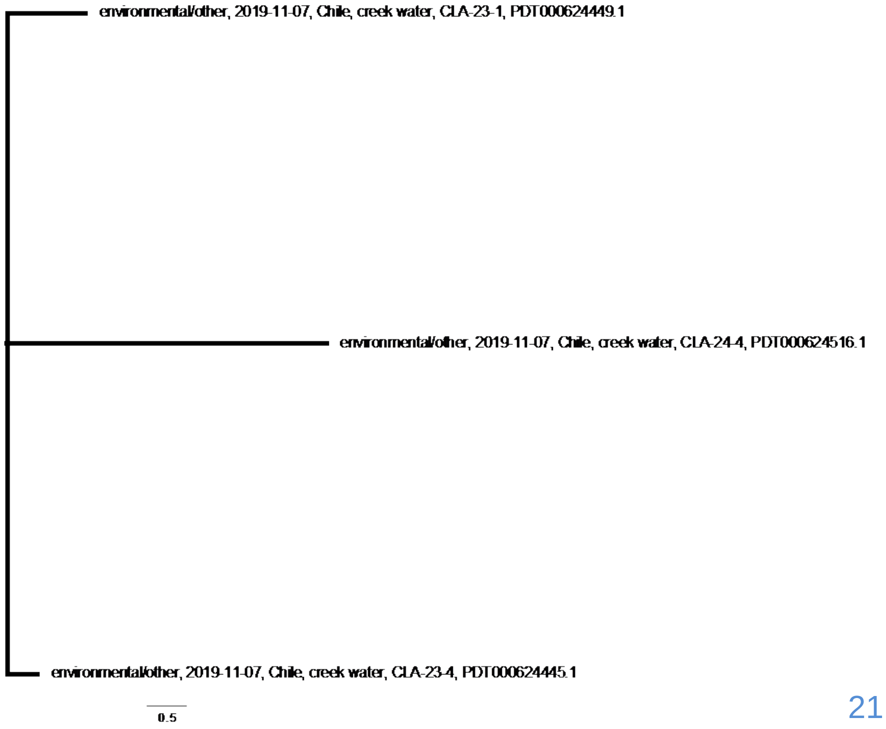
SNP Cluster	AMR	BIOCIDE	METAL	VIRULENCE
PDS000003955.533	63	7	56	14
PDS000026867.7	6	0	9	0
PDS000029248.7	4	0	8	0
PDS000029912.2	10	0	15	0
PDS000030058.41	3	0	3	0
PDS000032393.135	32	0	64	0
PDS000037691.3	8	0	12	0
PDS000051343.31	8	0	12	0
PDS000053260.2	3	0	3	0
PDS000053263.2	4	0	6	0
PDS000053277.1	2	0	3	0
PDS000053278.1	0	0	2	4
PDS000053279.1	6	0	9	0
PDS000053280.1	6	0	6	0
PDS000053281.1	19	0	9	0
PDS000053282.1	4	0	6	0
PDS000053283.1	6	0	9	0
PDS000053286.1	10	0	15	0
PDS000053288.1	4	0	6	0
PDS000053289.1	0	0	2	4
PDS000053290.1	4	0	6	0
PDS000053291.1	12	0	18	0
PDS000053292.1	4	0	6	0
PDS000053293.1	4	0	6	0
PDS000053294.1	4	0	6	0

Diversity of isolates from Chili

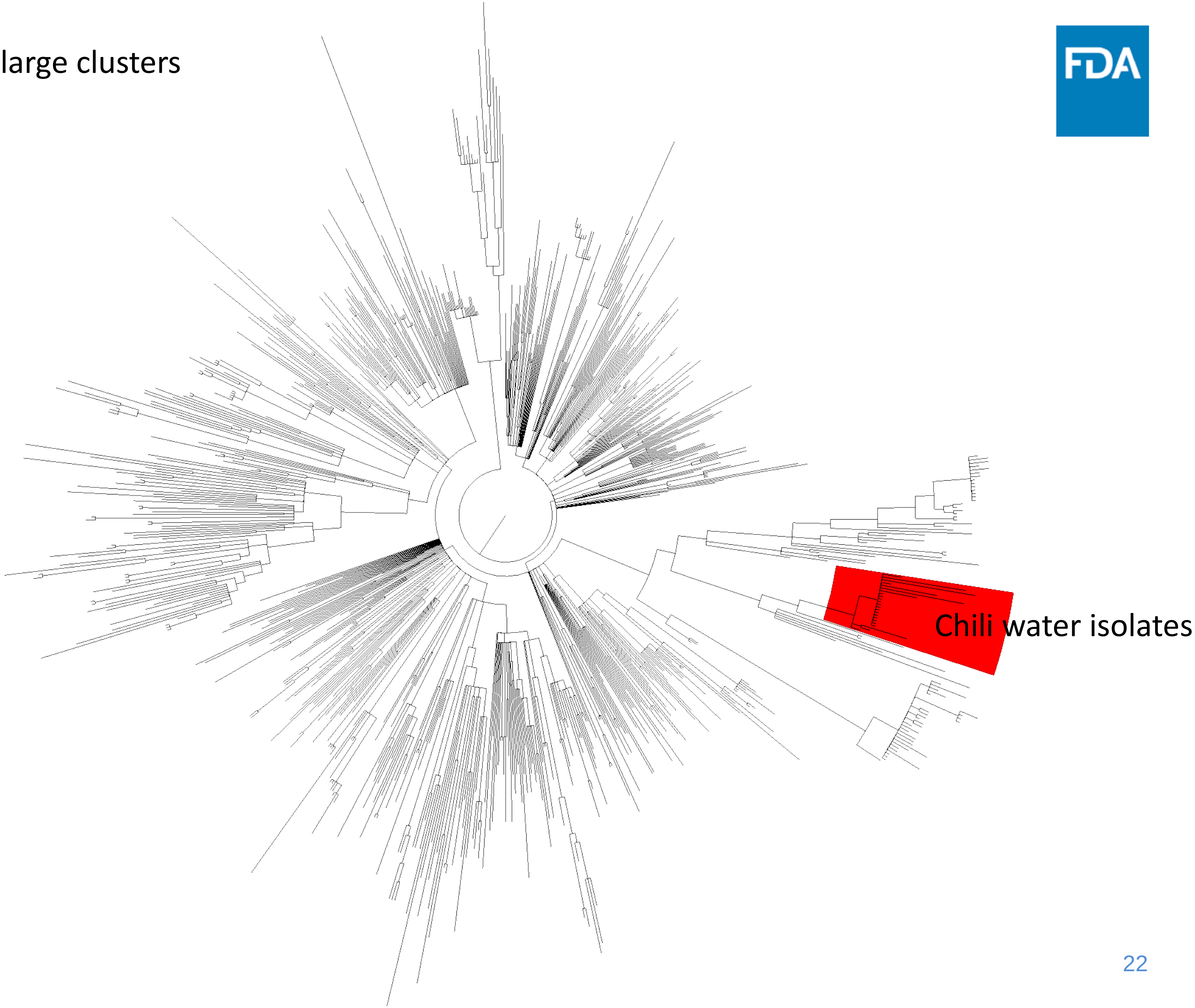


Many small clusters

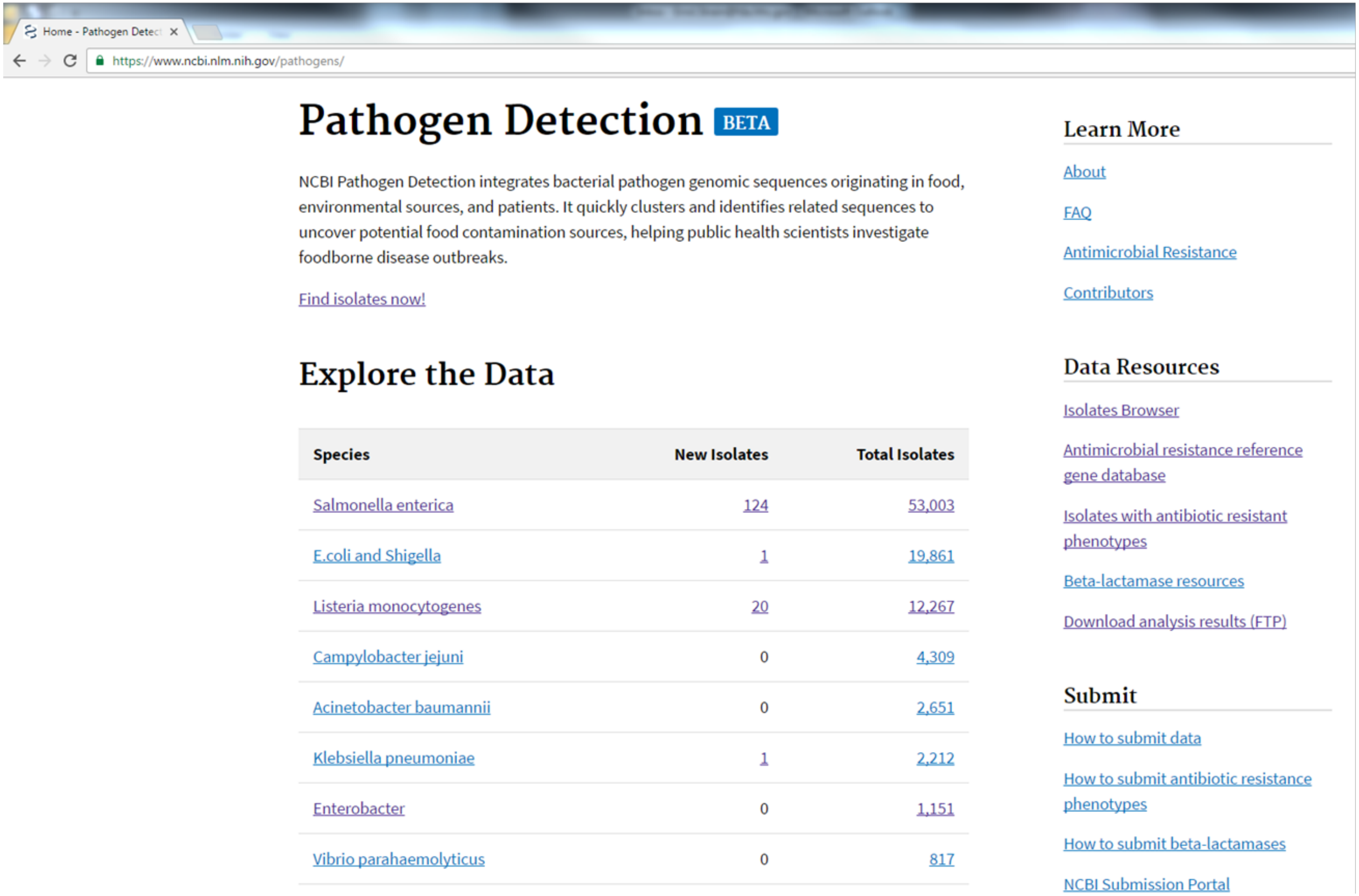




But also large clusters



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



Pathogen Detection BETA

NCBI Pathogen Detection integrates bacterial pathogen genomic sequences originating in food, environmental sources, and patients. It quickly clusters and identifies related sequences to uncover potential food contamination sources, helping public health scientists investigate foodborne disease outbreaks.

Find isolates now!

Explore the Data

Species	New Isolates	Total Isolates
Salmonella enterica	124	53,003
E.coli and Shigella	1	19,861
Listeria monocytogenes	20	12,267
Campylobacter jejuni	0	4,309
Acinetobacter baumannii	0	2,651
Klebsiella pneumoniae	1	2,212
Enterobacter	0	1,151
Vibrio parahaemolyticus	0	817

Learn More

- About
- FAQ
- Antimicrobial Resistance
- Contributors

Data Resources

- Isolates Browser
- Antimicrobial resistance reference gene database
- Isolates with antibiotic resistant phenotypes
- Beta-lactamase resources
- Download analysis results (FTP)

Submit

- How to submit data
- How to submit antibiotic resistance phenotypes
- How to submit beta-lactamases
- NCBI Submission Portal

Publicly released metadata

Food/environmental Submission

Pathogen: environmental/food/other sample from Listeria monocytogenes
BioSample: SAMN02709234;
Sample name: FDA00007620
Organism Listeria monocytogenes
cellular organisms; Bacteria; Firmicutes; Bacilli; Bacillales; Listeriaceae; Listeria
Pathogen: environmental/food/other; v.1.0

strain	FDA00007620
host disease	missing
latitude and longitude	missing
collection date	3/26/2012
isolate	CFSAN003790
geographic location	Italy
isolation source	moliterno al tartufo cheese
collected by	FDA
PFGE_SecondaryEnzyme_pattern	GX6A12.0280
PFGE_PrimaryEnzyme_pattern	GX6A16.0085

Submission
FDA, Justin Payne; 2014-03-27
ID: 2709234

Clinical Sample Submission

Pathogen: clinical or host-associated sample from Salmonella enterica
BioSample: SAMN02927343;
Sample name:/2013K-0563
Organism Salmonella enterica
cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Salmonella
Pathogen: clinical or host-associated; version 1.0

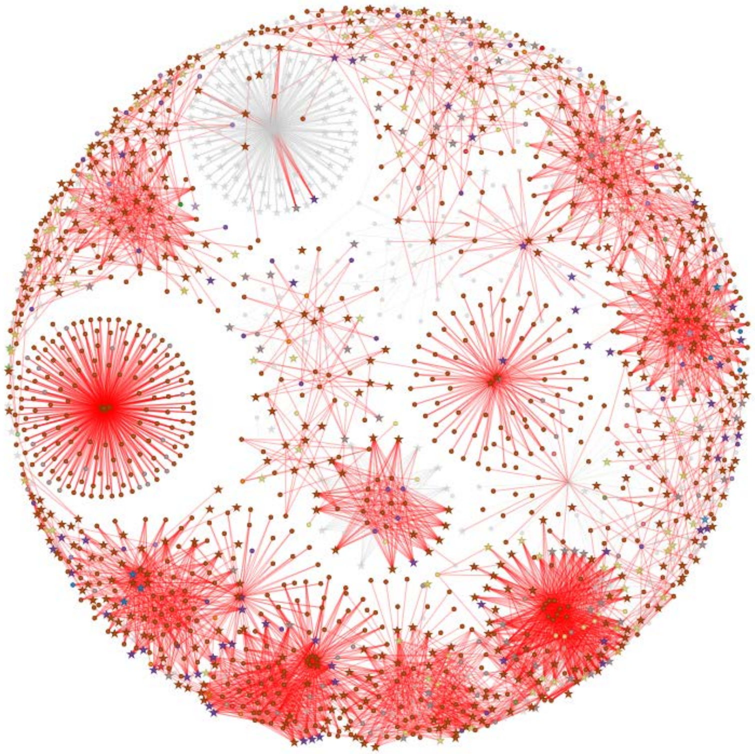
strain	2013K-0563
collected by	CDC
collection date	Missing
geographic location	USA
host	Missing
host disease	Missing
isolation source	Missing
latitude and longitude	Missing

Submission
Pulsenet, Eija Trees; 2014-07-18
ID: 2927343

U.S. FOOD & DRUG ADMINISTRATION

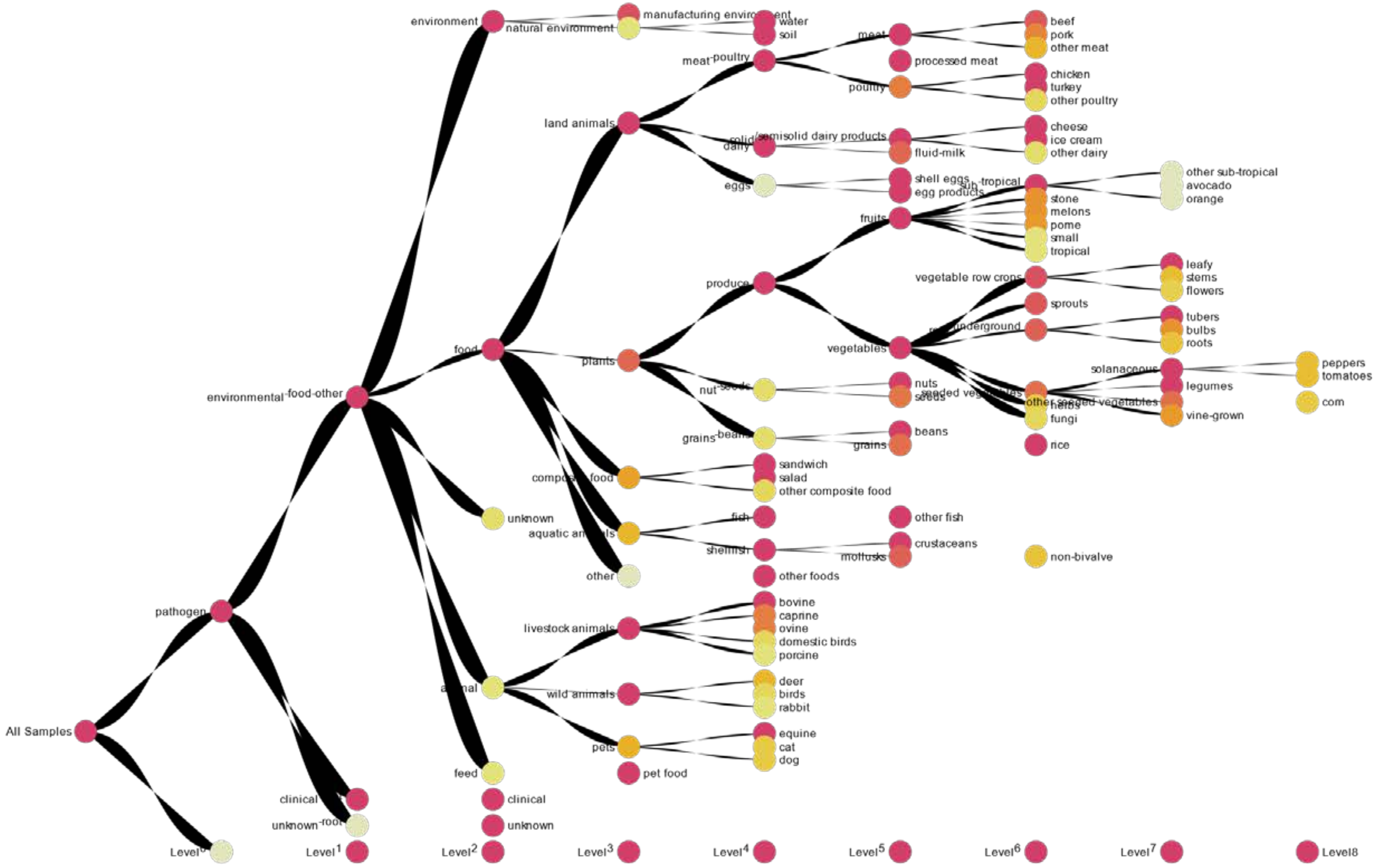
U.S. FOOD & DRUG ADMINISTRATION

GENOME GRAPHR



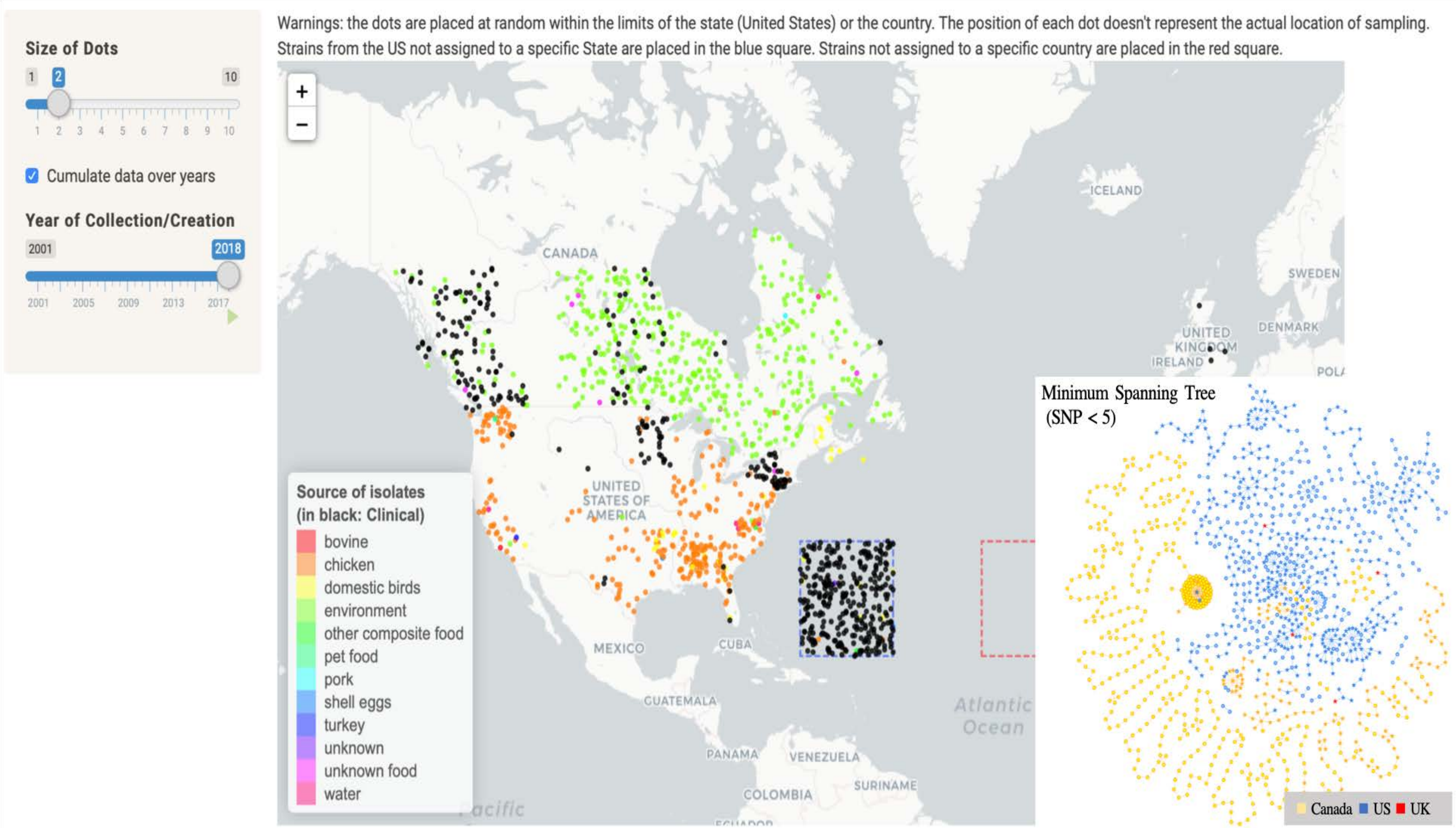
Leveraging GenomeTrakr & NCBI Pathogen Detection WGS Data to Enhance

- Risk Assessment
- Attribution
- Large-scale Epidemiology studies



GenomeGraphR: WGS data integration, analysis, and visualization for risk assessment and management: https://fda-riskmodels.foodrisk.org/genomegraphr/

Moez Sanaa, Régis Pouillot, Francisco J Garces-Vega, Errol Strain, Jane M Van Doren doi: https://doi.org/10.1101/495309 2018.



Warnings: the dots are placed at random within the limits of the state (United States) or the country. The position of each dot doesn't represent the actual location of sampling. Strains from the US not assigned to a specific State are placed in the blue square. Strains not assigned to a specific country are placed in the red square.

Size of Dots: 2

☑ Cumulate data over years

Year of Collection/Creation: 2001 – 2018

FDA U.S. FOOD & DRUG ADMINISTRATION

Center for Veterinary Medicine
Protecting Human and Animal Health

NARMS
National Antimicrobial Resistance Monitoring System
Food and Drug Administration • Centers for Disease Control and Prevention • United States Department of Agriculture
FDA CDC USDA

From WGS to Antibiotic Resistance Genotype

DNA from Single colony →

Sequencing
With Illumina Miseq →

CLCbio A QIAGEN Company

Assembly
CLC Genomics Workbench

↗ BLAST ®

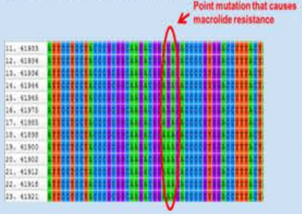
Local Blast AR Gene Database →

aac(3)-IIa, aadA1, aph(3')-Ia
catA1, tetO …

Acquired AR genes

↘ Sequences alignment
gyrA gene
23S rRNA gene →

Campylobacter 23S rDNA alignment



Point mutations related to AR

AR Genotype

RESISTOME TRACKER

Salmonella

Select an icon or alert below to get started.

ALERTS

There are 11 records with 2 flagged genes uploaded in the last 30 days

Manuscripts using structured metadata ontologies, machine learning and artificial intelligence.

Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, Tyson GH, Zhao S, Davis JJ. 2019. **Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella.** J Clin Microbiol 57:e01260-18. https://doi.org/10.1128/JCM.01260-18.

Zhang et al. 2019. **Zoonotic Source Attribution of Salmonella enterica Serotype Typhimurium Using Genomic Surveillance Data, United State**. Emerg Infect Dis. 25(1): 82–91. doi: 10.3201/eid2501.180835

Adaptations of particular interest to food safety specialists:

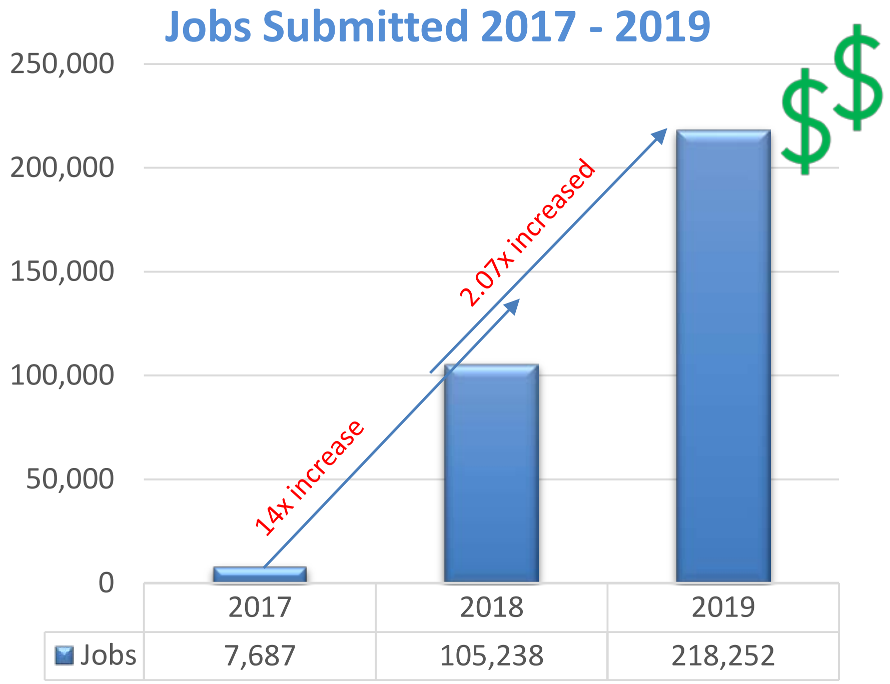
(1) Thermal tolerance
(2) Dessication resistance
(3) Osmotic/Ionic tolerance
(4) Quat resistance
(5) Chlorine resistance
(6) Biofilm persistence
(7) Surface adherence
(8) Antibiotic resistance
(9) Antimicrobial resistance
(10) Ecological fitness
(11) Heavy metal resistance
(12) Metabolic persistence
(13) Enhanced hydrophobic fitness
(14) Produce invasiveness
(15) Flower invasiveness
(16) Root system invasiveness
(17) Acid resistance
(18) Surface water fitness
(19) In vivo plant migratory fitness
(20) Soil fitness
(21) Capsaicin resistance
(22) Swarming
(23) Trans-ovarian poultry colonization
(24) Fecal persistence (poultry)
(25) Yolk content invasion
(26) Multidrug resistance
(27) External amoeba harborage
(28) Internal amoeba harborage
(29) Acyl-homoserine lactone (AHL)
(30) KatE stationary-phase catalase
(31) In vivo migratory fitness
(32) RDAR phenotype
(33) The ‘Weltevreden’ type
(34) Peristence within the tomato**

A partial gene list is now incorporated into AMRFinderPlus v3.2 (https://github.com/ncbi/amr/wiki)and the Reference Gene Catalog as of 2019-10-31 (https://www.ncbi.nlm.nih.gov/pathogens/isolates#/refgene/)

GalaxyTrakr

- **2019 Objective: Increase Jobs Submitted to Galaxy by 1.5 times to 160,000**

- **End of Year Status: Achieved!**
  - ✓ **From 2017 to 2018, number of jobs submitted increased by 14x**
  - ✓ **From 2018 to 2019, number of jobs submitted increased by 2.07x**

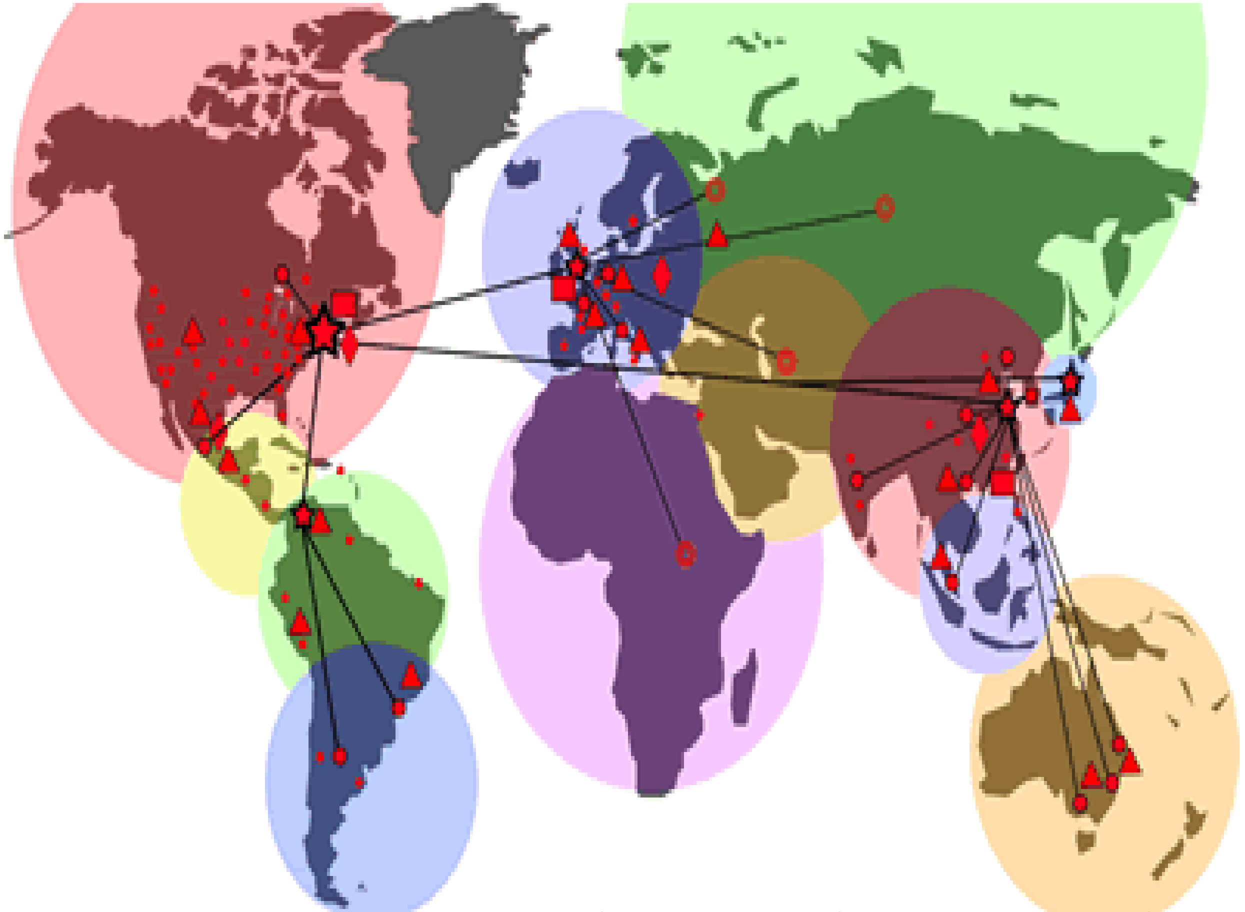


Jobs Submitted 2017 - 2019

- **2019 Objective: Increase Jobs Submitted to Galaxy by 1.5 times to 160,000**

- **Benefit / Impact:**
  - ✓ **GalaxyTrakr Training Environment**
    - ▪ Deployed a separate training environment to support various GalaxyTrakr training events without impacting production
  - ✓ **Updated to AWS S3**
    - ▪ Updated object storage services (from AWS EFS to AWS S3) allowing an increase in user storage quotas AND additional 90% reduction in storage costs ($.30 to $.003 per GB)
  - ✓ **Increased Number of Users:**
    - ▪ Current Number of Registered Users: 522 (10% increase since last quarterly meeting:475 registered users)
  - ✓ **Over 70 tools and 11 workflows available on GalaxyTrakr**
    - ▪ Increase in tools supports overall mission objectives and expands the overall GalaxyTrakr community

- Future Projects
- GenomeTrakr Latin America



For more discussions please seek out Gonzalez-Escalona, Narjol <Narjol.Gonzalez-Escalona@fda.hhs.gov>; Stevens, Eric <Eric.Stevens@fda.hhs.gov>; Cornell, Jason <Jason.Cornell@fda.hhs.gov>

Acknowledgements

- **FDA**
  - Center for Food Safety and Applied Nutrition
  - Center for Veterinary Medicine
  - Office of Regulatory Affairs
- **National Institutes of Health**
  - National Center for Biotechnology Information
- **State Health and University Labs**
  - Alaska
  - Arizona
  - California
  - Florida
  - Hawaii
  - Maryland
  - Minnesota
  - New Mexico
  - New York
  - South Dakota
  - Texas
  - Virginia
  - Washington
- **USDA/FSIS and ARS**
- **CDC**
  - Enteric Diseases Laboratory
- **INEI-ANLIS "Carolos Malbran Institute," Argentina**
- **Centre for Food Safety, University College Dublin, Ireland and Irish FSA**
- **Melbourne (FSA). Australia**
- **Public Health England, UK**
- **Institute for Food Safety and Health (IFSH)**
- **WHO and FAO**
- **Illumina**
- **Pac Bio**
- **Other independent collaborators**

FDA U.S. FOOD & DRUG
ADMINISTRATION