National Center for Emerging and Zoonotic Infectious Diseases

Transitioning the PulseNet USA network to whole genome sequencing: the final push

Heather Carleton, MPH, PhD

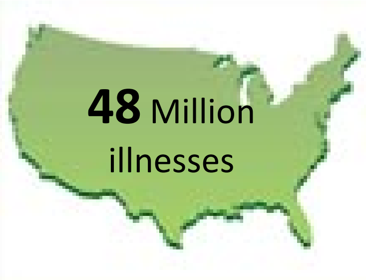
EDLB, DFWED

AMD Seminar Series

12/12/2018

Impact of Foodborne Diseases

Disease Burden in the US



48 Million illnesses

1 in **6** Americans

128,000 Hospitalizations

3,000 Deaths

Economic Impact

15+ Billion each year

Pathogens

Salmonella

Campylobacter

9 Billion $ each year

Vibrio

E coli

Listeria

Shigella

Flynn D 2014

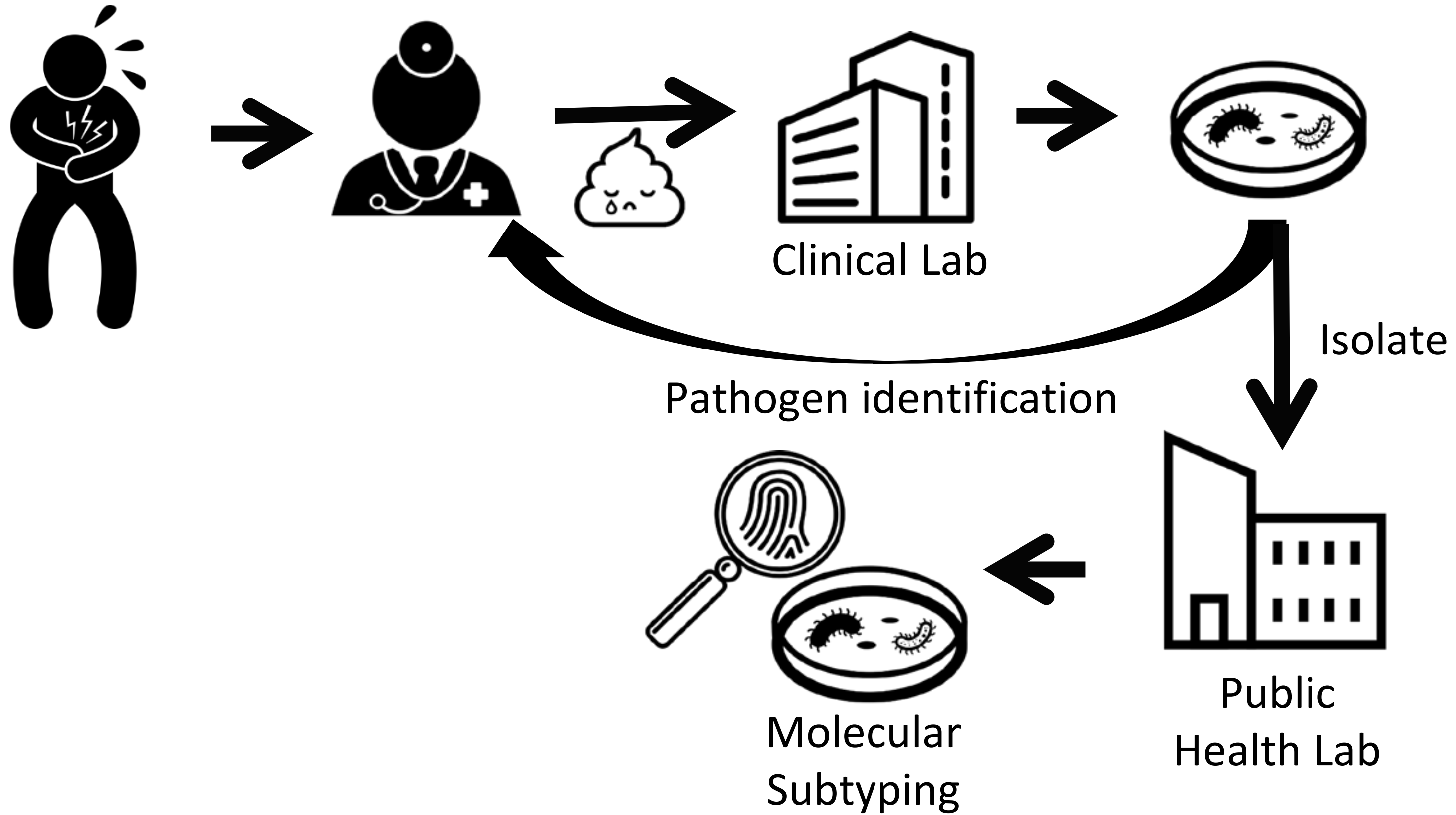
Disease Burden Worldwide

550 Million Foodborne Illnesses

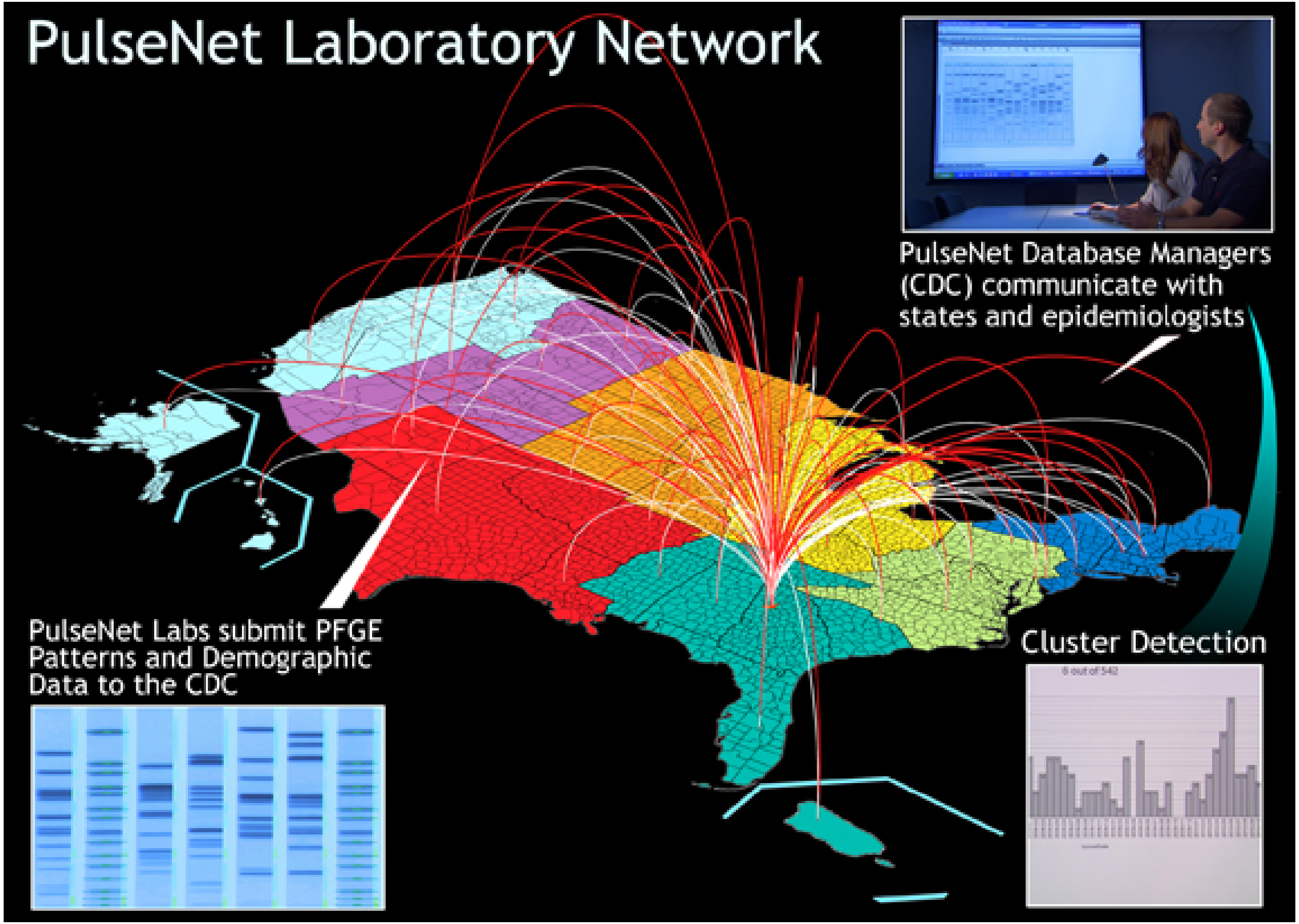
230,000 Deaths

www.who.int/foosafety

Foodborne illness: From patient to public health surveillance



PulseNet: National Molecular Subtyping Network for Foodborne Disease Surveillance

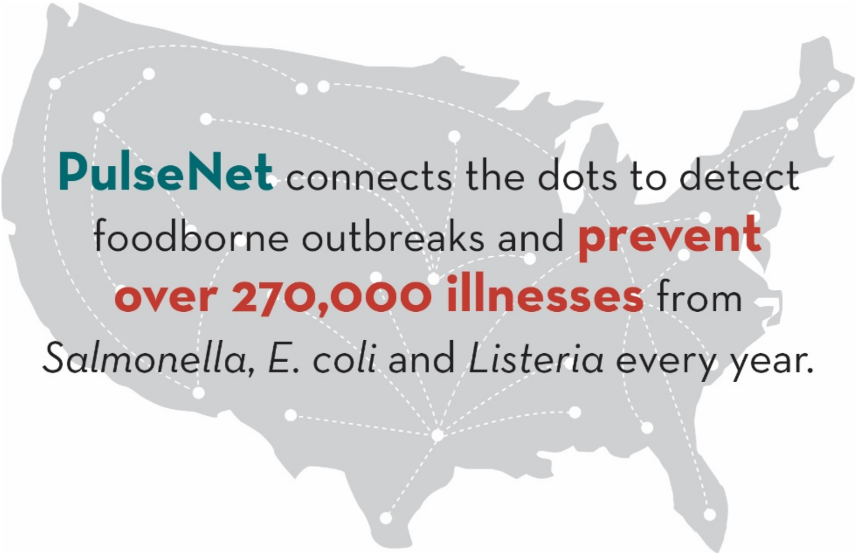


20 years of PulseNet USA
1996 CDC 2016

PulseNet: Over 20 years of detecting foodborne outbreaks with PFGE as the primary subtyping method

Every year **PulseNet** saves at least **half a billion dollars** in medical costs and lost productivity.

$1 spent = $70 saved!



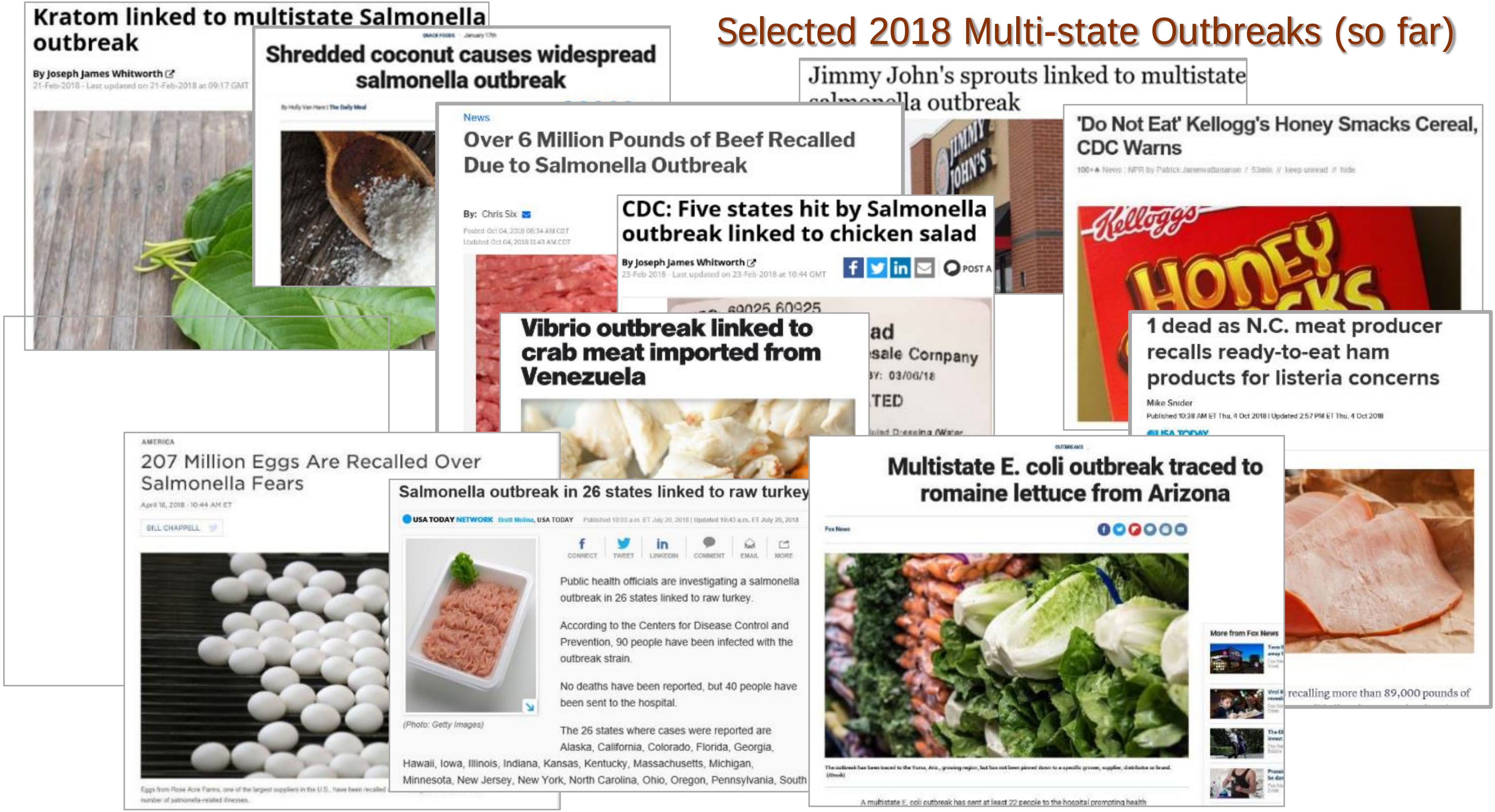
PulseNet connects the dots to detect foodborne outbreaks and **prevent over 270,000 illnesses** from *Salmonella*, *E. coli* and *Listeria* every year.

Scharff RL *et al.* 2016. An Economic Evaluation of PulseNet: A Network for Foodborne Disease Surveillance. Am J Prev Med 50:S66-73.

PulseNet has made these foods safer to eat:

Peanut butter, Sprouts, Eggs, Tree nuts, Poultry products, Leafy greens, Tomatoes, Frozen entrees, Lunch meat, Spices, Melons

Selected 2018 Multi-state Outbreaks (so far)



Kratom linked to multistate Salmonella outbreak

By Joseph James Whitworth
21-Feb-2018 - Last updated on 21-Feb-2018 at 09:17 GMT

SNACK FOODS · January 17th

Shredded coconut causes widespread salmonella outbreak

By Holly Van Hare | The Daily Meal

News

Over 6 Million Pounds of Beef Recalled Due to Salmonella Outbreak

By: Chris Six

Posted: Oct 04, 2018 08:34 AM CDT
Updated: Oct 04, 2018 11:43 AM CDT

CDC: Five states hit by Salmonella outbreak linked to chicken salad

By Joseph James Whitworth
23-Feb-2018 - Last updated on 23-Feb-2018 at 10:44 GMT

POST A

Vibrio outbreak linked to crab meat imported from Venezuela

Jimmy John's sprouts linked to multistate salmonella outbreak

---

'Do Not Eat' Kellogg's Honey Smacks Cereal, CDC Warns

100+ ♠ News : NPR by Patrick Jarenwattananon / 53min // keep unread // hide

---

1 dead as N.C. meat producer recalls ready-to-eat ham products for listeria concerns

Mike Snider
Published 10:38 AM ET Thu, 4 Oct 2018 | Updated 2:57 PM ET Thu, 4 Oct 2018
USA TODAY

recalling more than 89,000 pounds of

AMERICA

207 Million Eggs Are Recalled Over Salmonella Fears

April 16, 2018 · 10:44 AM ET

BILL CHAPPELL

Eggs from Rose Acre Farms, one of the largest suppliers in the U.S., have been recalled number of salmonella-related illnesses.

Salmonella outbreak in 26 states linked to raw turkey

USA TODAY NETWORK Brett Molina, USA TODAY Published 10:03 a.m. ET July 20, 2018 | Updated 10:43 a.m. ET July 20, 2018

CONNECT TWEET LINKEDIN COMMENT EMAIL MORE

(Photo: Getty Images)

Public health officials are investigating a salmonella outbreak in 26 states linked to raw turkey.

According to the Centers for Disease Control and Prevention, 90 people have been infected with the outbreak strain.

No deaths have been reported, but 40 people have been sent to the hospital.

The 26 states where cases were reported are Alaska, California, Colorado, Florida, Georgia, Hawaii, Iowa, Illinois, Indiana, Kansas, Kentucky, Massachusetts, Michigan, Minnesota, New Jersey, New York, North Carolina, Ohio, Oregon, Pennsylvania, South

---

OUTBREAKS

Multistate E. coli outbreak traced to romaine lettuce from Arizona

Fox News

The outbreak has been traced to the Yuma, Ariz., growing region, but has not been pinned down to a specific grower, supplier, distributor or brand. (iStock)

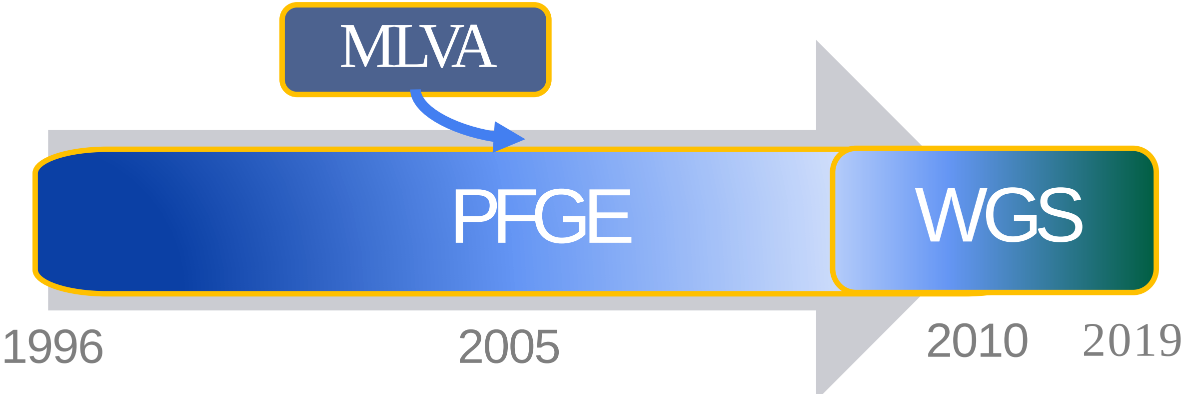
More from Fox News

A multistate E. coli outbreak has sent at least 22 people to the hospital prompting health

Transitioning to Whole Genome Sequencing

EDLB: Moving PulseNet to Whole Genome Sequencing

Standardized, automated methods to ensure comparability of data generated in different laboratories, save time and resources



MLVA

PFGE

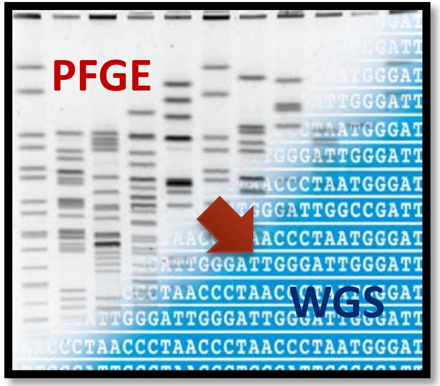
WGS

1996

2005

2010

2019



Modified from Carleton and Gerner-Smidt (ASM Microbe July 2016)

PulseNet™ International

USA, Canada, Europe, Asia Pacific, Africa, Middle East, Latin America and the Caribbean

PulseNet Surveillance by WGS – greater resolution

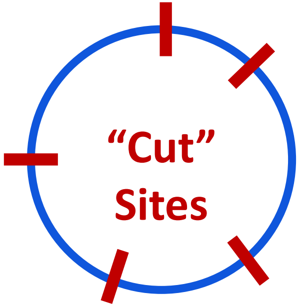
PFGE:

Comparing book chapter sizes

WGS:

Comparing the letters in each word in the book

PulseNet Surveillance by WGS – greater resolution



PFGE only gives information at a "cut" site via the banding pattern

WGS has the ability to give us information at every position in the bacterial genome

Reference Characterization by WGS:

'One Shot' Characterization Of STEC

DEPARTMENT OF HEALTH AND HUMAN SERVICES

Public Health Service

Centers for Disease Control
and Prevention (CDC)
Atlanta GA 30333

Genus/Species: *Escherichia coli*

Serotype: O104:H4

Pathotype: Shiga toxin-producing and enteroaggregative *E. coli* (STEC/EAEC)

Virulence profile: *stx2a, aggR, aggA, sigA, sepA, pic, aatA, aaiC, aap*

Sequence Type: ST678

Allele code: 102.45.26.35.3

Antimicrobial resistance genes: $bla_{TEM\text{-}1}$, $bla_{CTX\text{-}M\text{-}15}$, *strAB, sul2, tet(A)A, dfrA7*

EDLB Vision

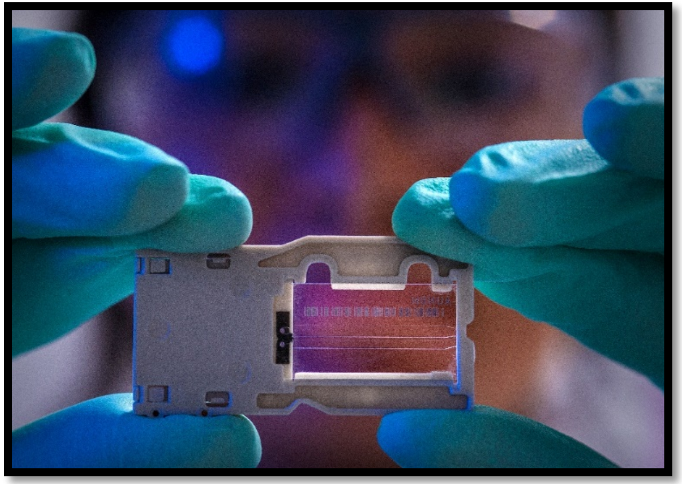
REPLACE all enteric workflows:

- Identification
- Serotyping
- Virulence profiling
- Antimicrobial susceptibility
- Subtyping for surveillance and outbreak investigations (PulseNet)

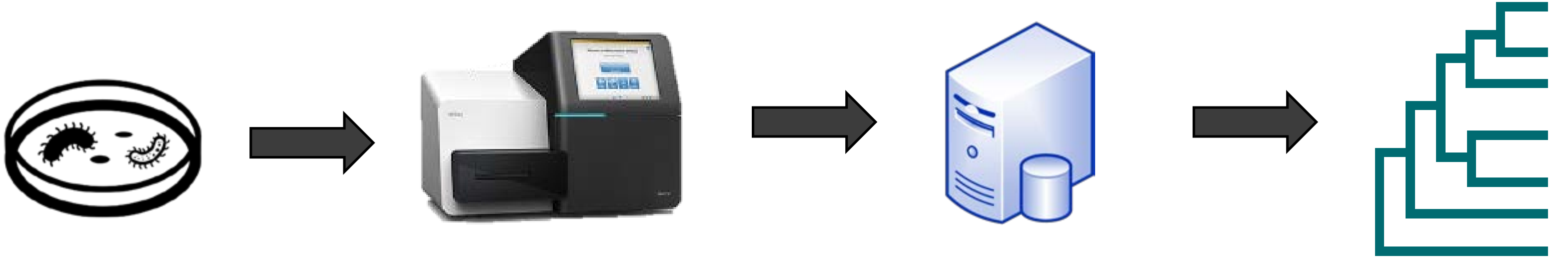
With ONE cost-efficient and precise method:

- All of this information can be derived from the genome sequence

WGS



PulseNet WGS-based Foodborne Disease Surveillance Workflow



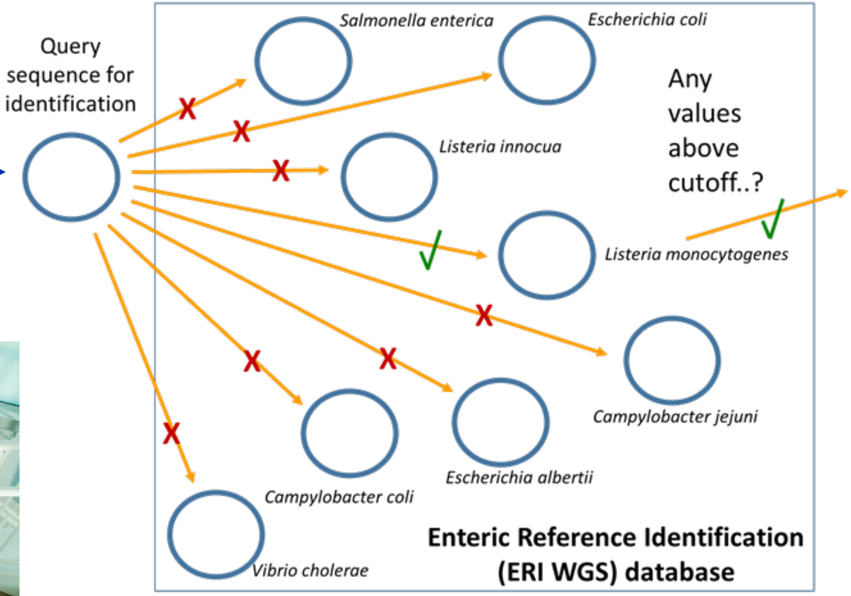
- DNA extraction and library prep – 4-8 hours (standardized protocols)
- Sequencing 24-36 hours (QC thresholds for sequence data from isolates)
- Analysis 2-4 hours (decentralized databases with centralized analysis capability)

Reference Identification Database (RefID)

Species specific databases

Raw reads, QC → *de novo* Assembly QC →

Query sequence for identification



Any values above cutoff..?

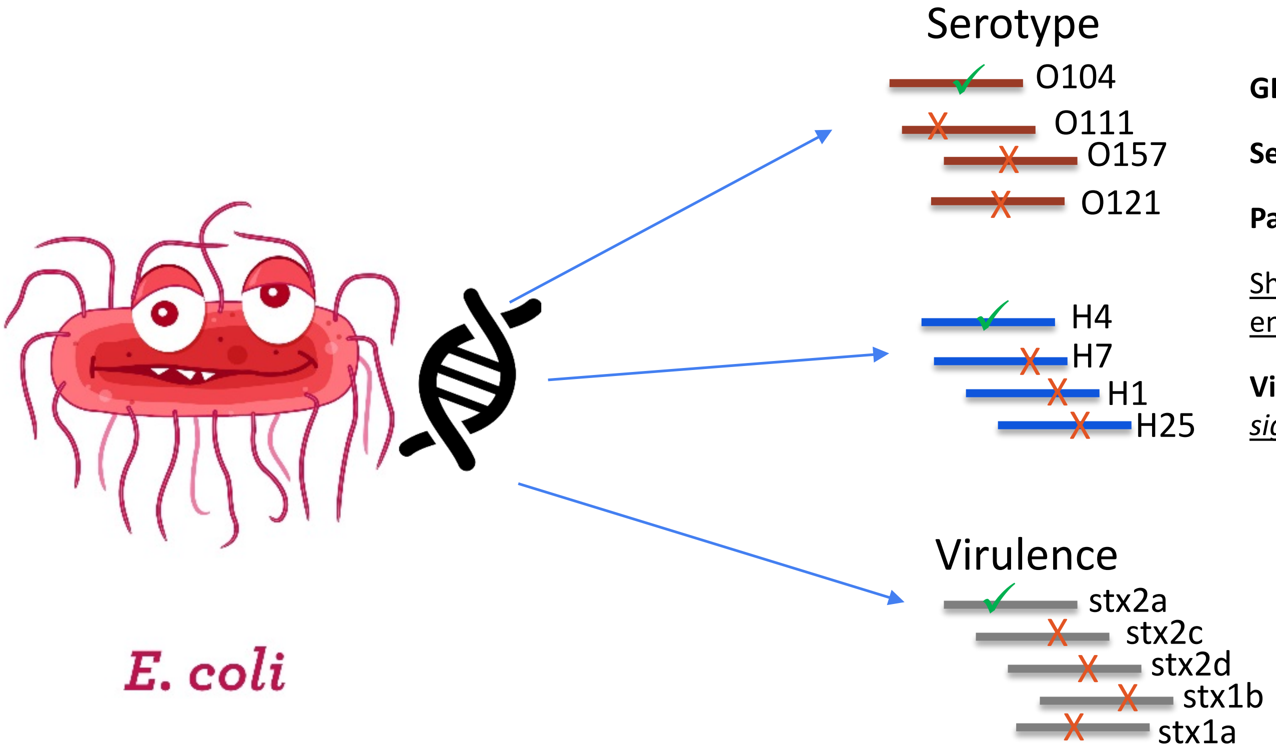
Enteric Reference Identification (ERI WGS) database

Average Nucleotide Identity (ANI)

Add to *Listeria* database for further characterization

Serotyping, MLST, virulence genes, AST, etc. (depending on organism)

Genotyper: Predicting phenotype from genotype



GENOTYPER RESULTS:

Serotype: O104:H4

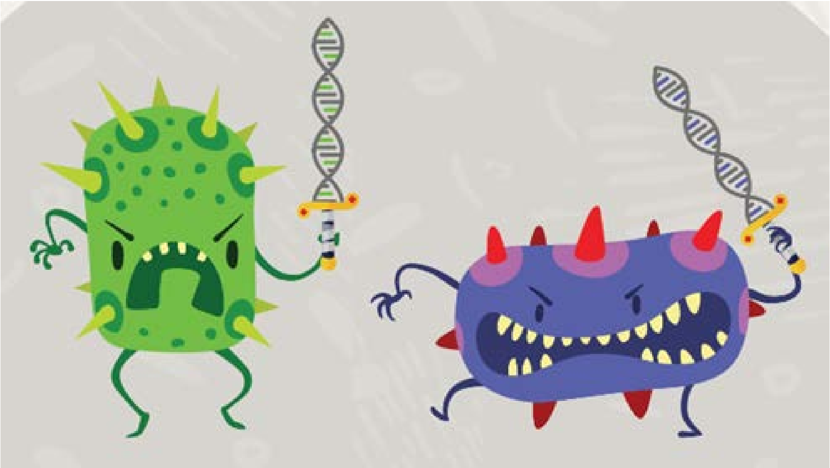
Pathotype:

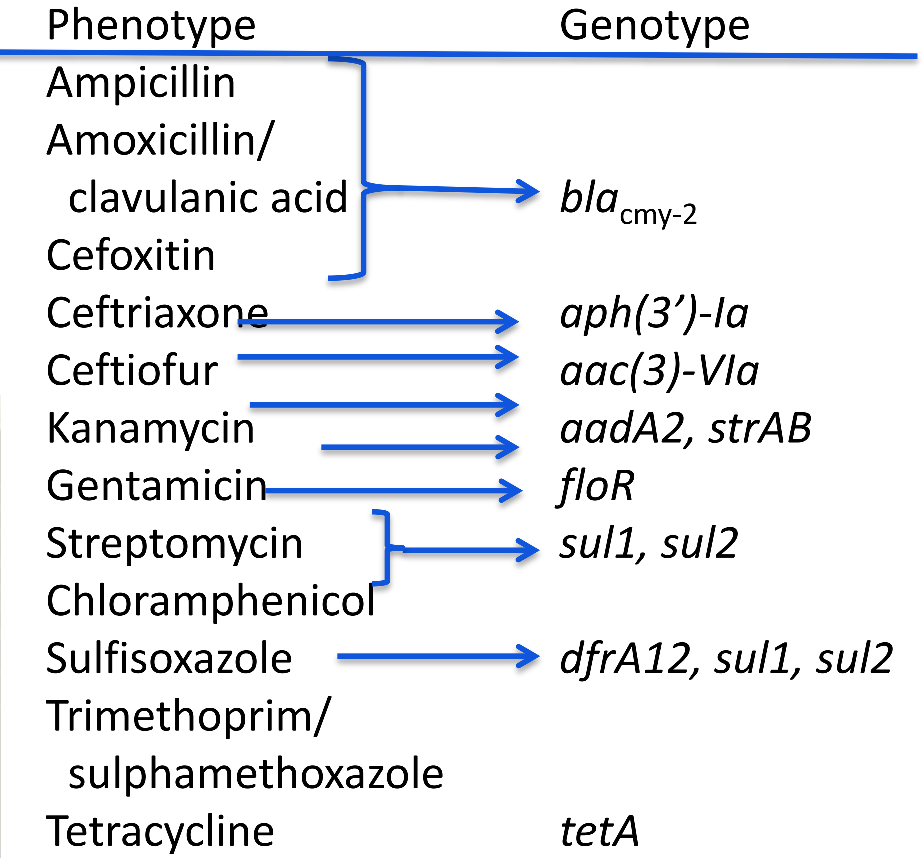
Shiga toxin-producing and enteroaggregative *E. coli* (STEC/EAEC)

Virulence profile: *stx2a, aggR, aggA, sigA, sepA, pic, aatA, aaiC, aap*

Genotyper: Acquired Resistance

- Genes associated with a particular AR phenotype
- 96.9% predictive value
  (2015 NTS NARMS data)





Phenotype

Ampicillin
Amoxicillin/
clavulanic acid
Cefoxitin
Ceftriaxone
Ceftiofur
Kanamycin
Gentamicin
Streptomycin
Chloramphenicol
Sulfisoxazole
Trimethoprim/
sulphamethoxazole
Tetracycline

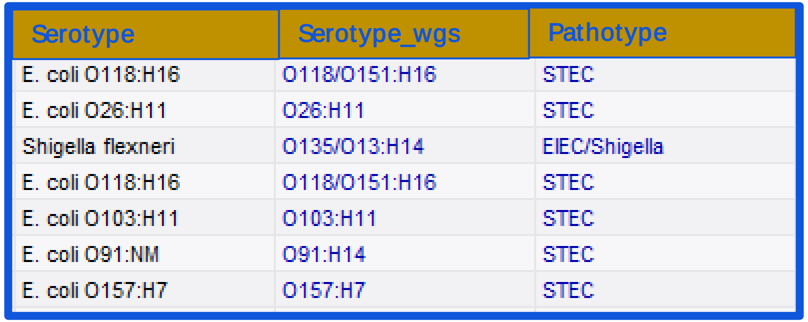
Genotype

bla$_{\text{cmy-2}}$
aph(3')-Ia
aac(3)-VIa
aadA2, strAB
floR
sul1, sul2
dfrA12, sul1, sul2
tetA

WGS Genotyping Tools Available in BioNumerics

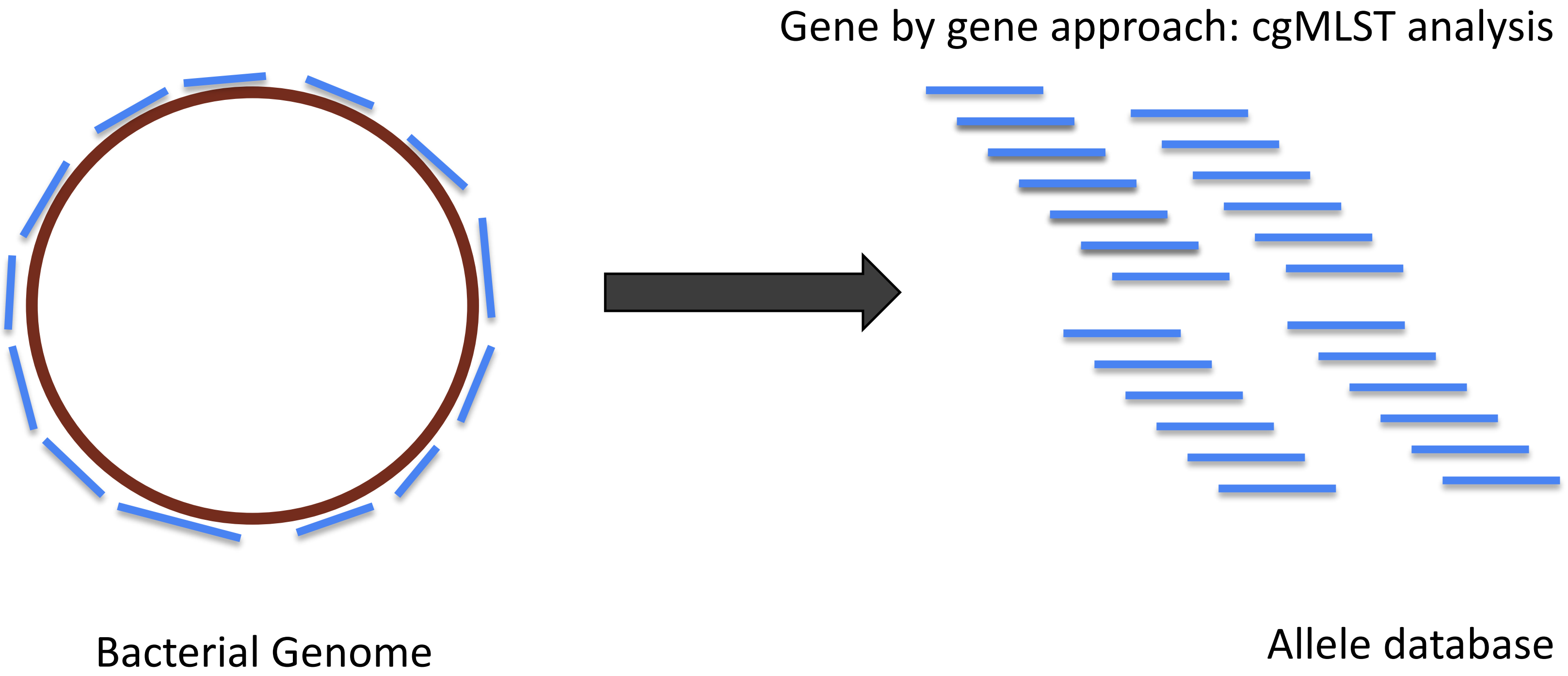
Listeria	Salmonella	Escherichia (O157/Non-O157/Shigella)	Campylobacter
Resistance	Resistance	Resistance	Resistance
Plasmid	Plasmid	Plasmid	Plasmid
	In silico PCR (genes for pathovars)	Virulence (stx/eae/etc. used to determine Pathotype)	
Lineage	Antigenic formula and serotype	Serotype	

Example of output



Serotype	Serotype_wgs	Pathotype
E. coli O118:H16	O118/O151:H16	STEC
E. coli O26:H11	O26:H11	STEC
Shigella flexneri	O135/O13:H14	EIEC/Shigella
E. coli O118:H16	O118/O151:H16	STEC
E. coli O103:H11	O103:H11	STEC
E. coli O91:NM	O91:H14	STEC
E. coli O157:H7	O157:H7	STEC

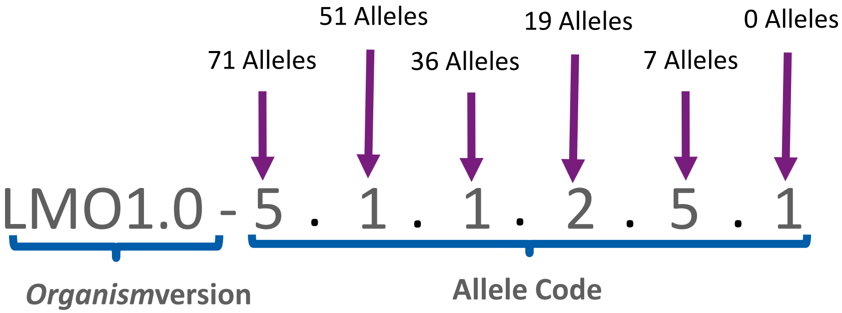
Subtyping by WGS: Core genome MLST



Subtyping by WGS: whole genome MLST

Database	Current Version	Total Loci: Whole Genome (wgMLST)	Total Loci: Core Genome (cgMLST)
Campylobacter	v5	6,651	1,343 (*C. jejuni, C. coli*)
Escherichia	v4	34,453	2,513 (*E. coli*)
Listeria	Final (v4)	4804	1,748 (*L. monocytogenes*)
Salmonella	v4	19,384	3,002 (*S. enterica*)
Vibrio	TBD	TBD	TBD

Subtyping by WGS: Allele Codes



51 Alleles 19 Alleles 0 Alleles

71 Alleles 36 Alleles 7 Alleles

LMO1.0 - 5 . 1 . 1 . 2 . 5 . 1

*Organism*version

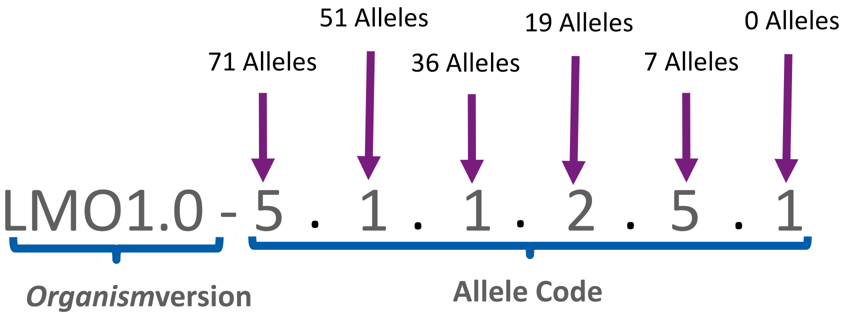
Allele Code

- When sequences have partial names, it means they are *singletons* in clusters below their last digit.
- The sequences below are approximately within 36 and 19 alleles of each other.

Sequence A **LMO1.0 - 5 . 1 . 2**

Sequence B **LMO1.0 - 5 . 1 . 2 . 2 . 5 . 1**

Listeria Nomenclature “allele code”

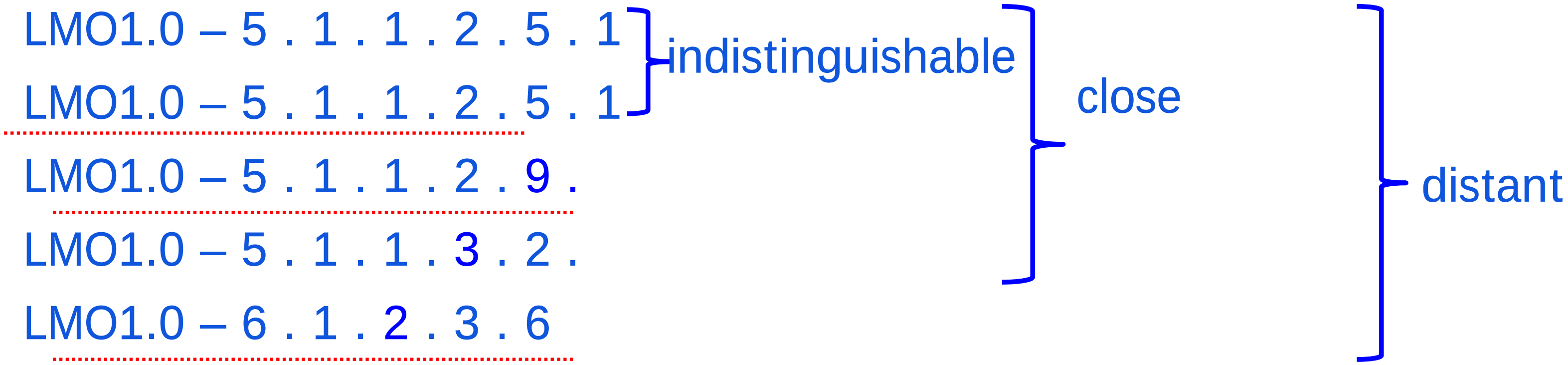


71 Alleles
51 Alleles
36 Alleles
19 Alleles
7 Alleles
0 Alleles

LMO1.0 - 5 . 1 . 1 . 2 . 5 . 1

*Organism*version

Allele Code



LMO1.0 – 5 . 1 . 1 . 2 . 5 . 1
LMO1.0 – 5 . 1 . 1 . 2 . 5 . 1
LMO1.0 – 5 . 1 . 1 . 2 . 9 .
LMO1.0 – 5 . 1 . 1 . 3 . 2 .
LMO1.0 – 6 . 1 . 2 . 3 . 6

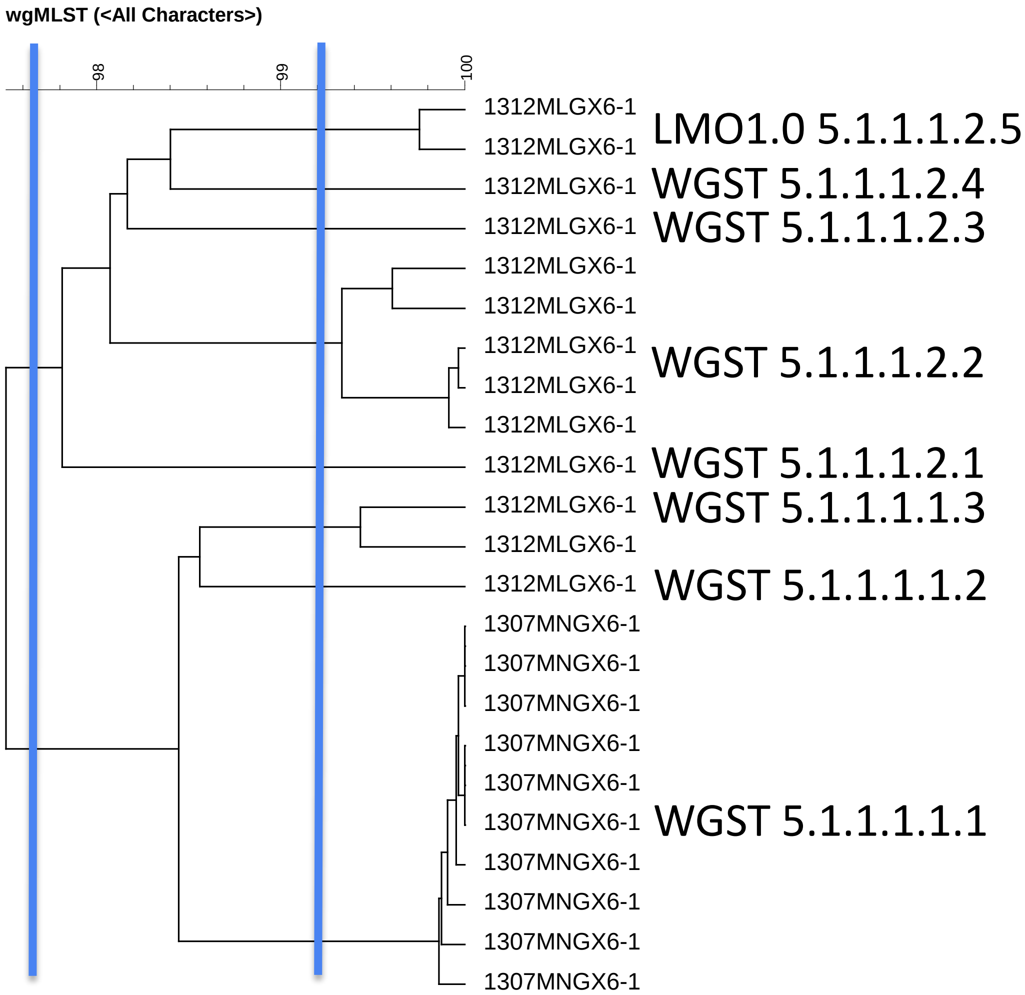
indistinguishable

close

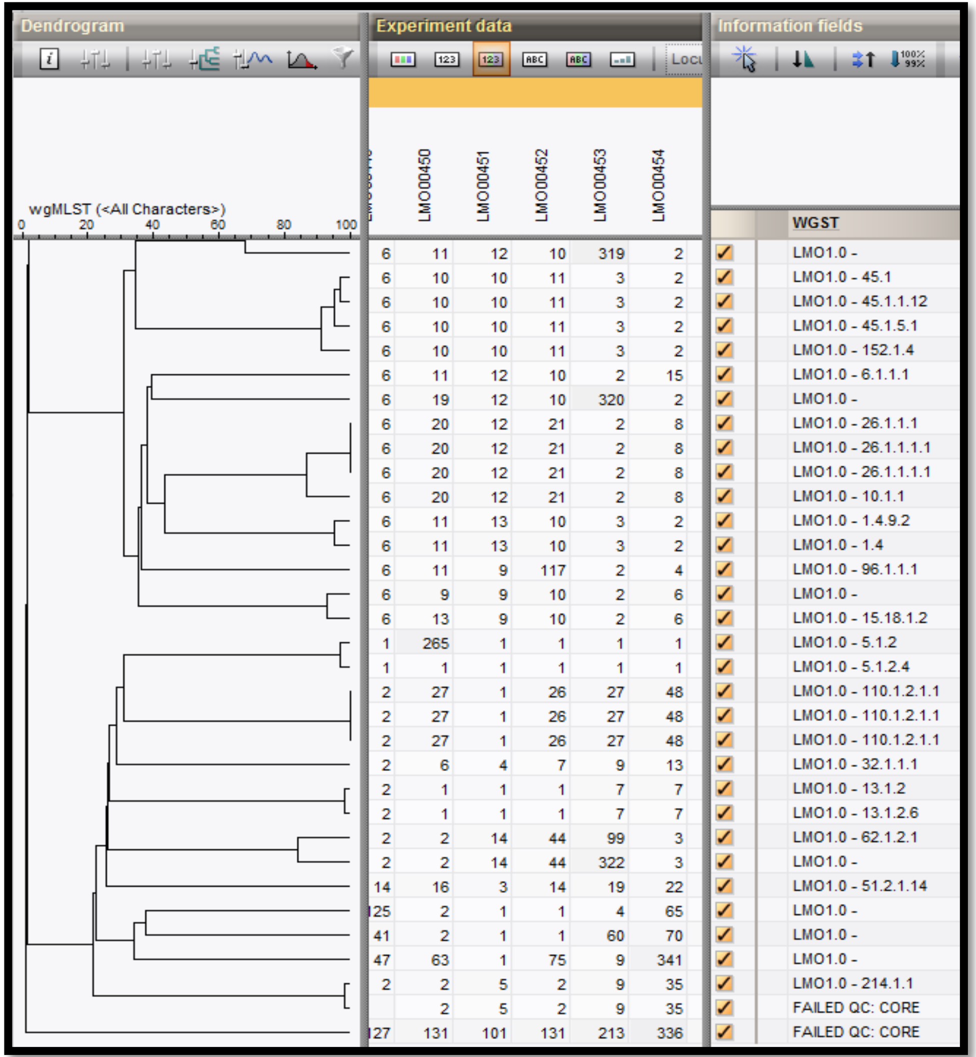
distant

Allele Code

- Allele code can be used as zip code to signify how close isolates are on a tree
- Allele codes vary by organism



Allelle Codes in BioNumerics



	LMO00450	LMO00451	LMO00452	LMO00453	LMO00454		WGST
6	11	12	10	319	2	☑	LMO1.0 -
6	10	10	11	3	2	☑	LMO1.0 - 45.1
6	10	10	11	3	2	☑	LMO1.0 - 45.1.1.12
6	10	10	11	3	2	☑	LMO1.0 - 45.1.5.1
6	10	10	11	3	2	☑	LMO1.0 - 152.1.4
6	11	12	10	2	15	☑	LMO1.0 - 6.1.1.1
6	19	12	10	320	2	☑	LMO1.0 -
6	20	12	21	2	8	☑	LMO1.0 - 26.1.1.1
6	20	12	21	2	8	☑	LMO1.0 - 26.1.1.1.1
6	20	12	21	2	8	☑	LMO1.0 - 26.1.1.1.1
6	20	12	21	2	8	☑	LMO1.0 - 10.1.1
6	11	13	10	3	2	☑	LMO1.0 - 1.4.9.2
6	11	13	10	3	2	☑	LMO1.0 - 1.4
6	11	9	117	2	4	☑	LMO1.0 - 96.1.1.1
6	9	9	10	2	6	☑	LMO1.0 -
6	13	9	10	2	6	☑	LMO1.0 - 15.18.1.2
1	265	1	1	1	1	☑	LMO1.0 - 5.1.2
1	1	1	1	1	1	☑	LMO1.0 - 5.1.2.4
2	27	1	26	27	48	☑	LMO1.0 - 110.1.2.1.1
2	27	1	26	27	48	☑	LMO1.0 - 110.1.2.1.1
2	27	1	26	27	48	☑	LMO1.0 - 110.1.2.1.1
2	6	4	7	9	13	☑	LMO1.0 - 32.1.1.1
2	1	1	1	7	7	☑	LMO1.0 - 13.1.2
2	1	1	1	7	7	☑	LMO1.0 - 13.1.2.6
2	2	14	44	99	3	☑	LMO1.0 - 62.1.2.1
2	2	14	44	322	3	☑	LMO1.0 -
14	16	3	14	19	22	☑	LMO1.0 - 51.2.1.14
125	2	1	1	4	65	☑	LMO1.0 -
41	2	1	1	60	70	☑	LMO1.0 -
47	63	1	75	9	341	☑	LMO1.0 -
2	2	5	2	9	35	☑	LMO1.0 - 214.1.1
	2	5	2	9	35	☑	FAILED QC: CORE
127	131	101	131	213	336	☑	FAILED QC: CORE

Singleton: No close matches, name not assigned

Two strains in this cluster are identical with 5 digits matching exactly. One is missing the 5th digit so it relates approximately within 19 alleles of the other two identical strains.

These three strains are identical, 0 alleles different based on the core genome

These strains failed QC and should be re-sequenced. Core genome less than 95%.

Reference Characterization by WGS:

’One Shot’ Characterization Of STEC

DEPARTMENT OF HEALTH AND HUMAN SERVICES

Public Health Service

Centers for Disease Control
and Prevention (CDC)
Atlanta GA 30333

Genus/Species: *Escherichia coli*

Serotype: O104:H4

Pathotype: Shiga toxin-producing and enteroaggregative *E. coli* (STEC/EAEC)

Virulence profile: *stx2a, aggR, aggA, sigA, sepA, pic, aatA, aaiC, aap*

Sequence Type: ST678

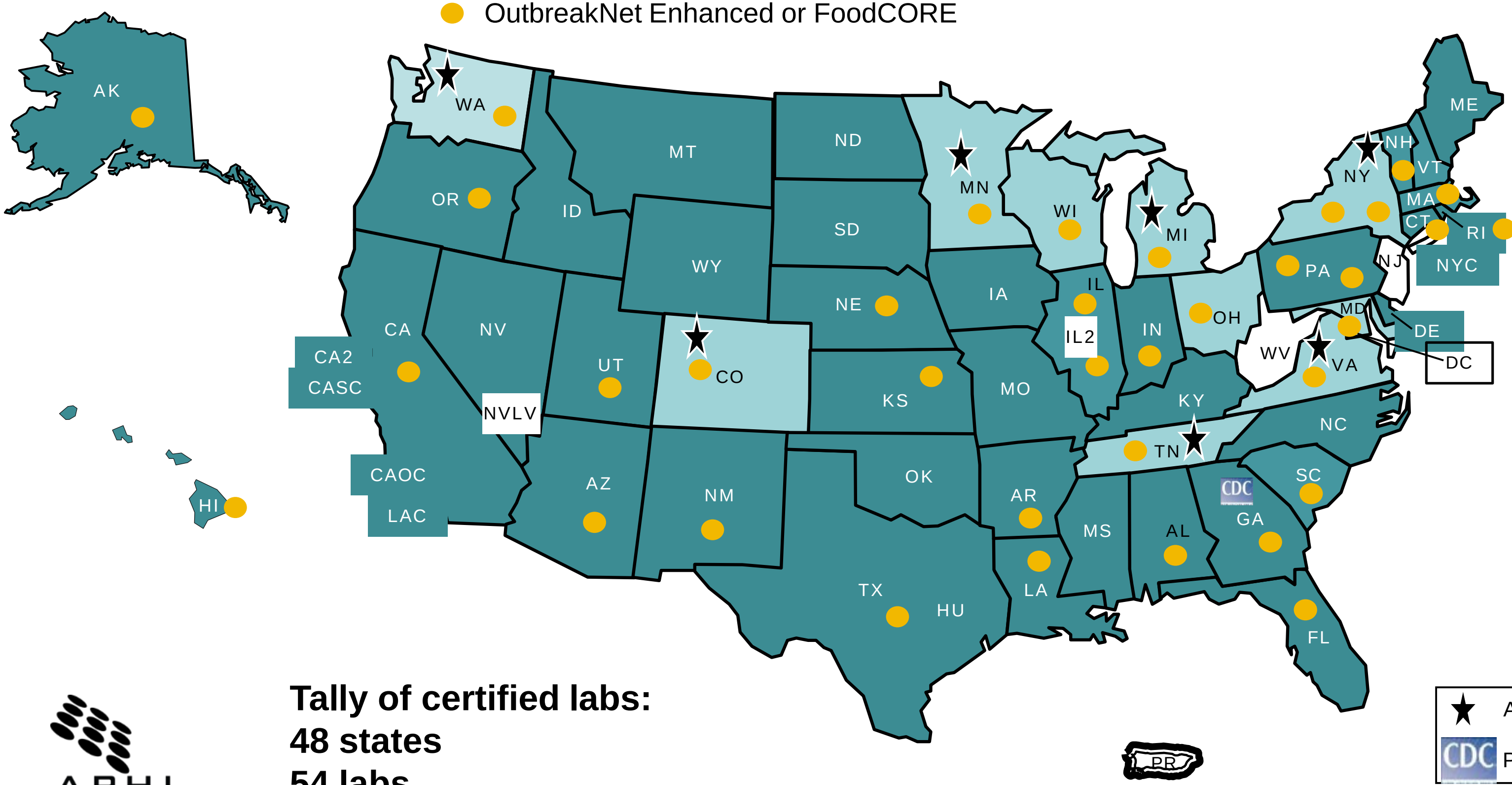
Allele code: 102.45.26.35.3

Antimicrobial resistance genes: $\textit{bla}_{\text{TEM-1}}$, $\textit{bla}_{\text{CTX-M-15}}$, *strAB, sul2, tet(A)A, dfrA7*

Implementing Whole Genome Sequencing Analysis Workflow

PulseNet, WGS and Enhanced Epidemiological Capacity

- 🟦 WGS certified, BioNumerics 7.6 pilot site
- ■ WGS certified
- ● OutbreakNet Enhanced or FoodCORE



Tally of certified labs:
48 states
54 labs

★ Area Laboratories
CDC PulseNet Central

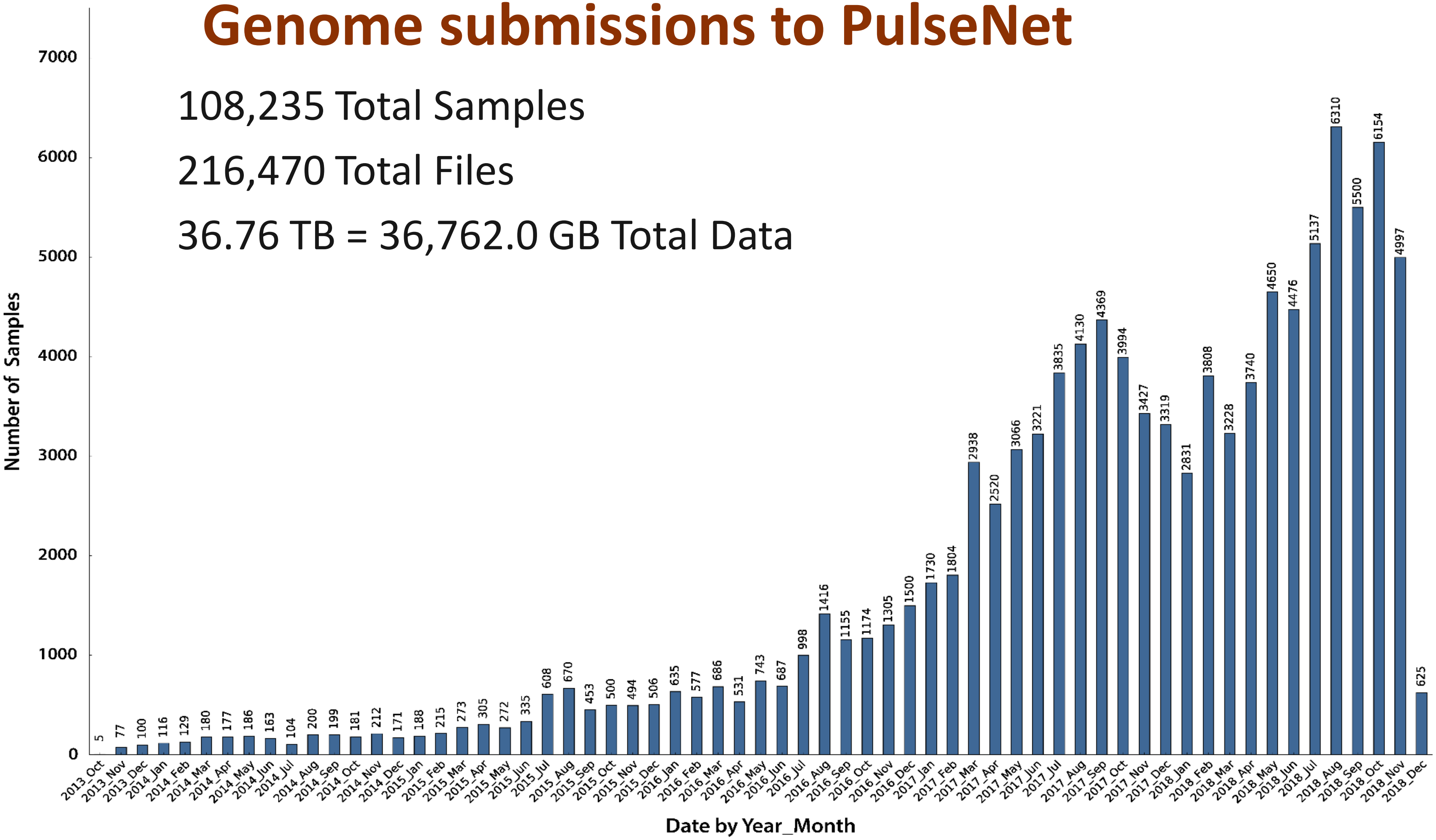
Modified: December 6, 2018

Genome submissions to PulseNet

108,235 Total Samples

216,470 Total Files

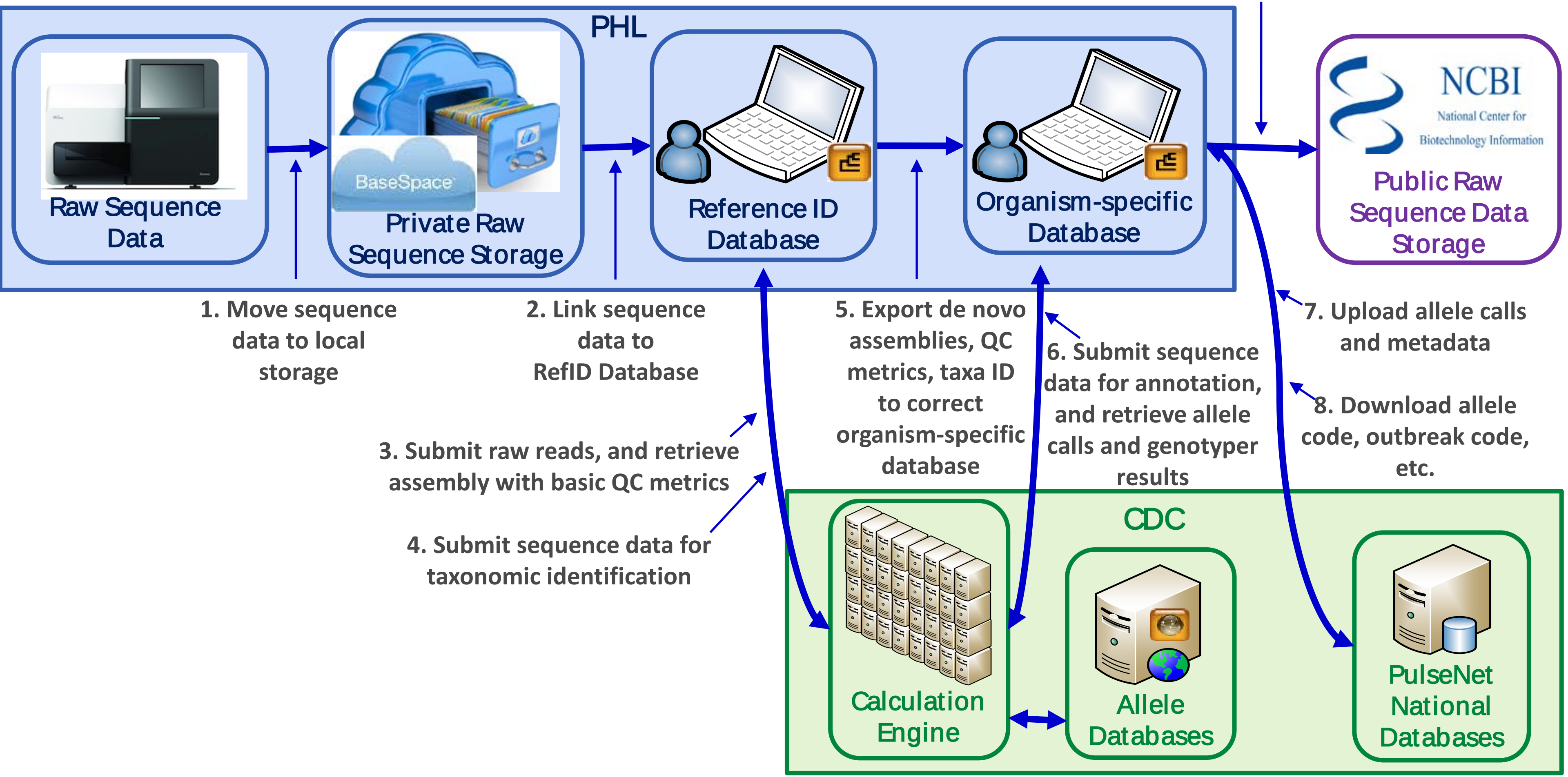
36.76 TB = 36,762.0 GB Total Data



Enteric Bacterial Genomes Sequenced and Uploaded to NCBI

PulseNet Organisms **(December 11, 2018)**	**Genomes on NCBI**
Salmonella spp	62,579
E. coli/Shigella	21,643
Campylobacter spp	8,086
Listeria monocytogenes	4,999
Vibrio spp	541
Yersinia spp	16
PulseNet Total	97,864

WGS Analysis Workflow



PHL

Raw Sequence Data

Private Raw Sequence Storage

Reference ID Database

Organism-specific Database

Public Raw Sequence Data Storage

CDC

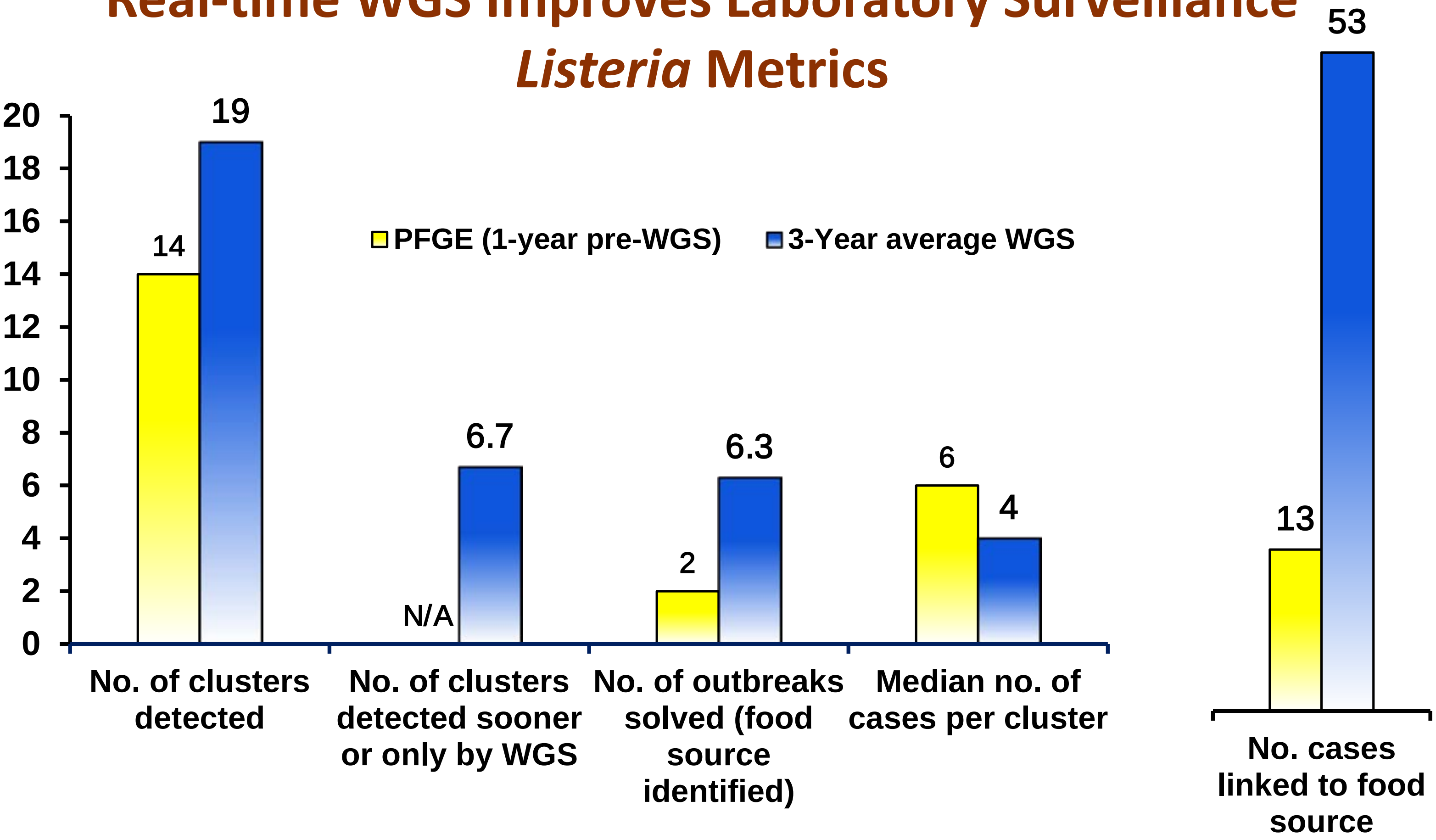
Calculation Engine

Allele Databases

PulseNet National Databases

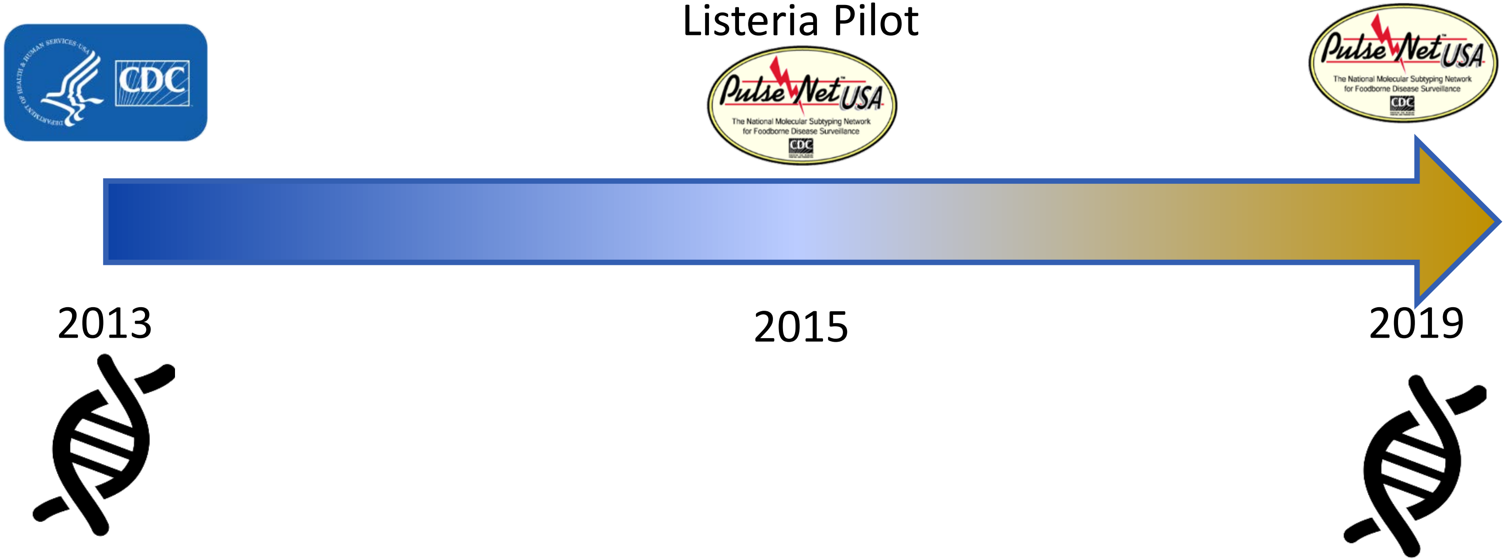
1. Move sequence data to local storage
2. Link sequence data to RefID Database
3. Submit raw reads, and retrieve assembly with basic QC metrics
4. Submit sequence data for taxonomic identification
5. Export de novo assemblies, QC metrics, taxa ID to correct organism-specific database
6. Submit sequence data for annotation, and retrieve allele calls and genotyper results
7. Upload allele calls and metadata
8. Download allele code, outbreak code, etc.
9. Upload raw sequence data with minimal metadata

Real-time WGS Improves Laboratory Surveillance *Listeria* Metrics



Courtesy Amanda Conrad, CDC Outbreak Response & Preparedness Branch

WGS Analysis: decentralizing from CDC to PulseNet PHL members



2013

Listeria Pilot

2015

2019

WGS in Action

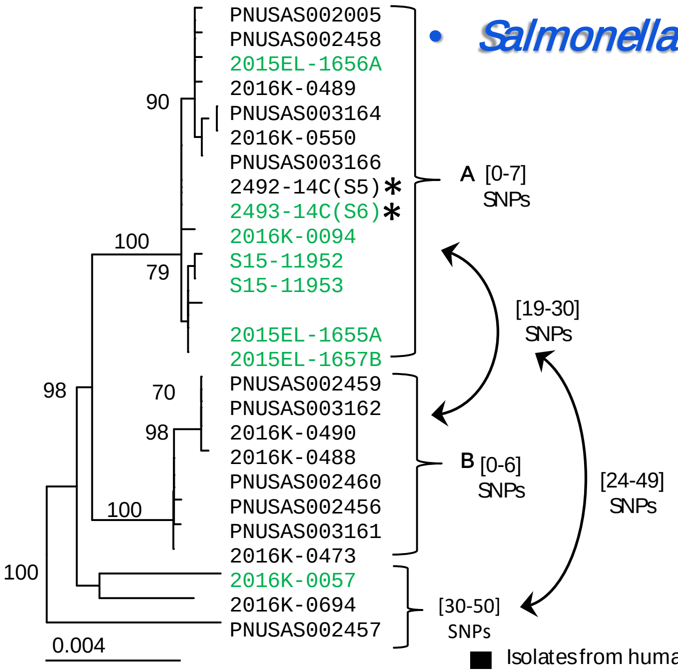
WGS in Action: How WGS is used for outbreak clusters

- Linking cases with potential sources
- Improving case definitions
- Linking historical clinical cases with ongoing outbreaks

How close is close?

- **Zoonotic outbreaks often quite diverse**

- *Salmonella* ser. Pomona outbreak



■ Isolates from human
🟩 Isolates from animal or Environmental
* Isolates from Chile

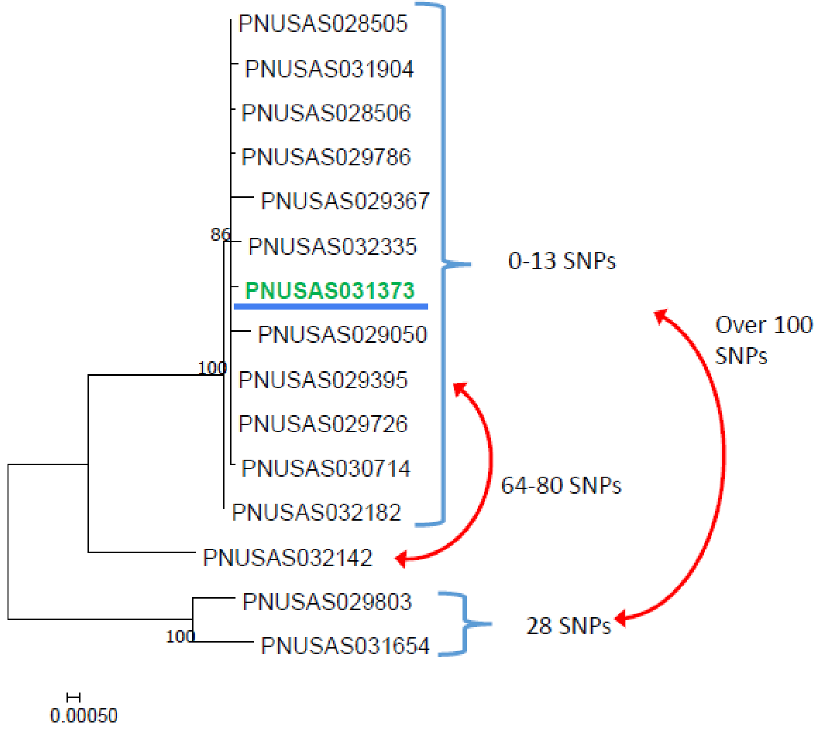
- **Turtles** were shipped from 8 turtle farms in **Louisiana** to **47 countries** from January 1, 2008 to March 1, 2016
- WGS sequencing confirmed the relationship between US and **Chile** *Salmonella* Pomona isolates

Don’t let the WGS data fool you!
Any supporting information should always be considered

Salmonella ser. Typhimurium strain from Egg Nog clustering with isolates from outbreak associated with laboratory exposure, 2017

WGS_id	Key	SourceSite	UploadDate
PNUSAS032142	[redacted]	Stool	1/11/2018
PNUSAS031904	[redacted]	Stool	1/3/2018
PNUSAS028505	[redacted]	Stool	11/21/2017
PNUSAS032182	[redacted]	Stool	1/5/2018
PNUSAS029395	[redacted]	STOOL	11/22/2017
PNUSAS029726	[redacted]	STOOL	11/30/2017
PNUSAS029050	[redacted]	Stool	11/17/2017
PNUSAS031373		Egg Nog	12/22/2017
PNUSAS032335	[redacted]	stool	12/22/2017
PNUSAS028506	[redacted]	Stool	11/21/2017
PNUSAS029786	[redacted]	Stool	11/21/2017
PNUSAS029803	[redacted]	Stool	12/4/2017
PNUSAS031654	[redacted]	Stool	12/8/2017
PNUSAS030714	[redacted]	Stool	12/14/2017
PNUSAS029367	[redacted]	Stool	11/28/2017

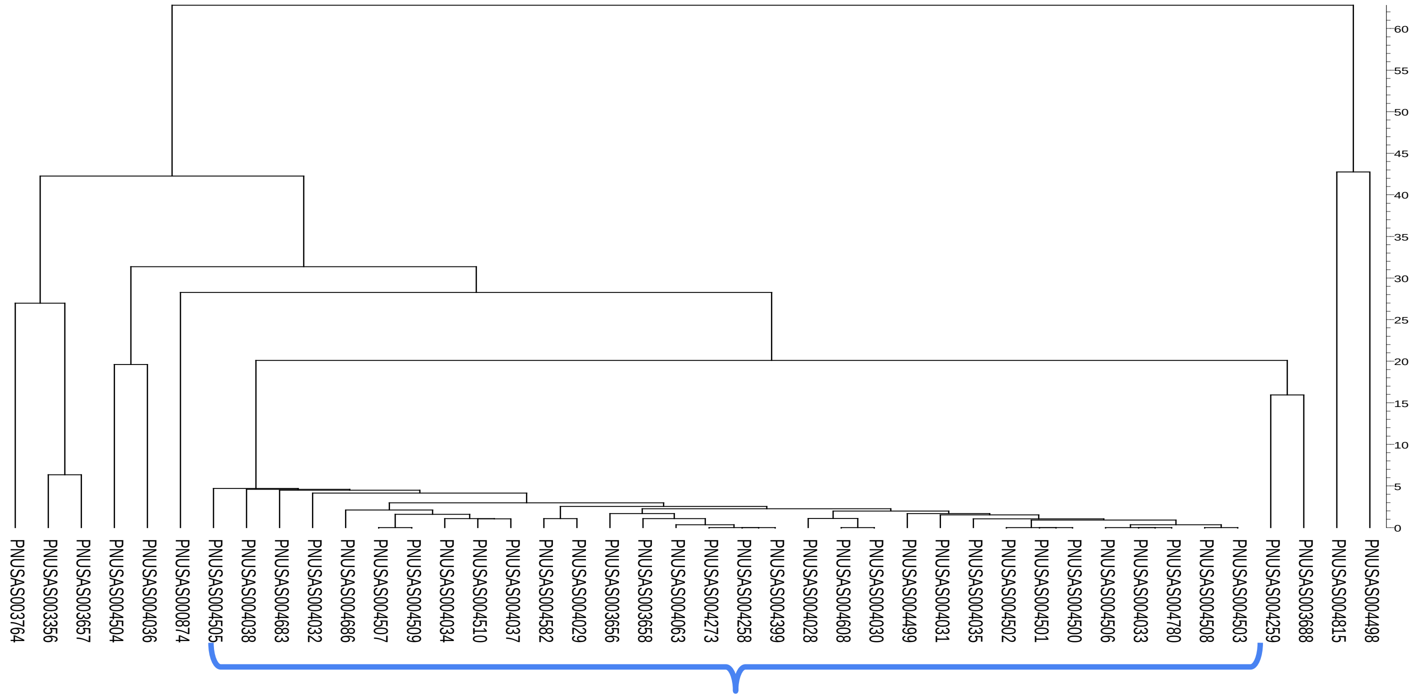
- wgMLST coincides with hqSNP
- **Resfinder:** No resistance genes



Source NOT Identified in All Outbreaks

Salmonella ser. Newport 1610MLJJP-1 (JJPX01.0011)

wgMLST analysis



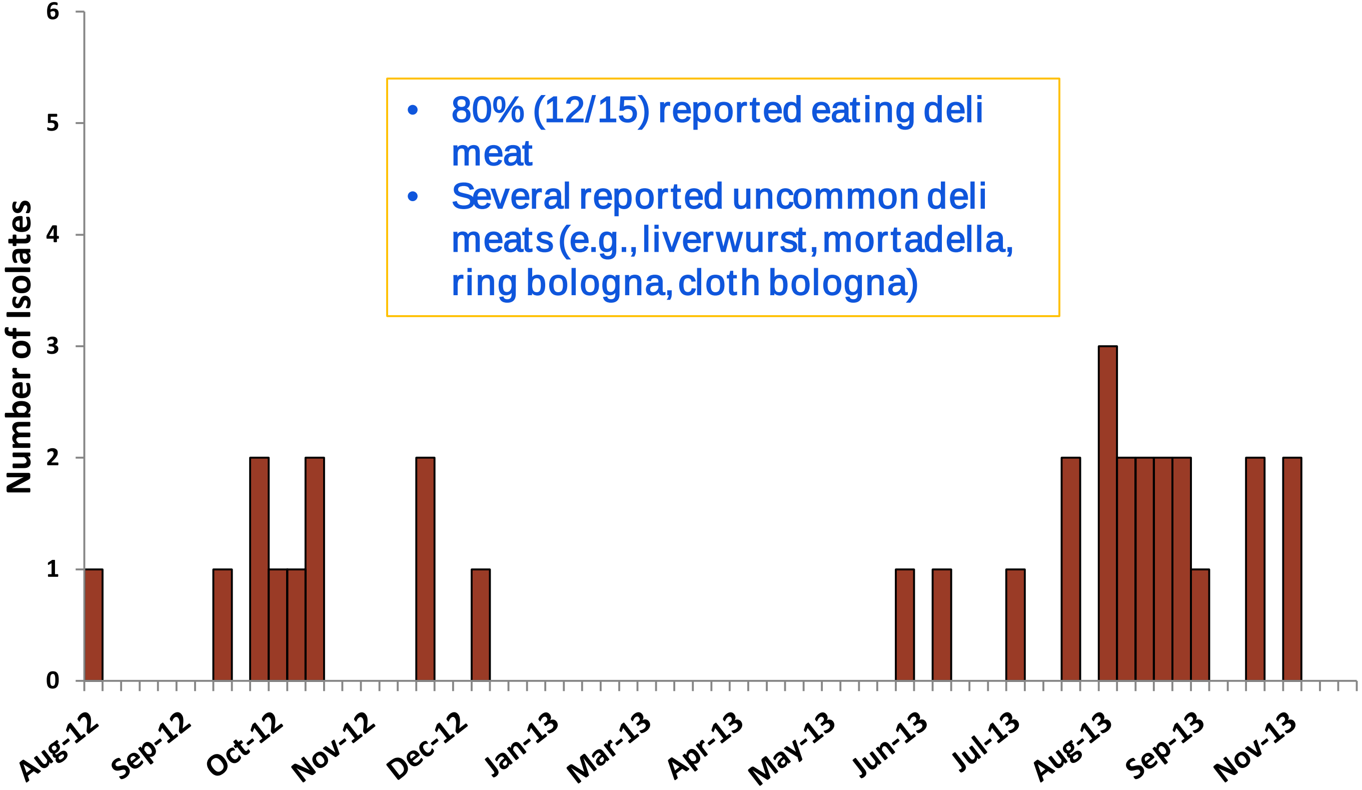
0 – 7 alleles;
median 2

Resfinder: all panS

BREAKING UP PFGE CLUSTERS IMPROVE CASE DEFINITION

PFGE Cluster Possibly Associated with Uncommon Deli Meats, 2012–2013

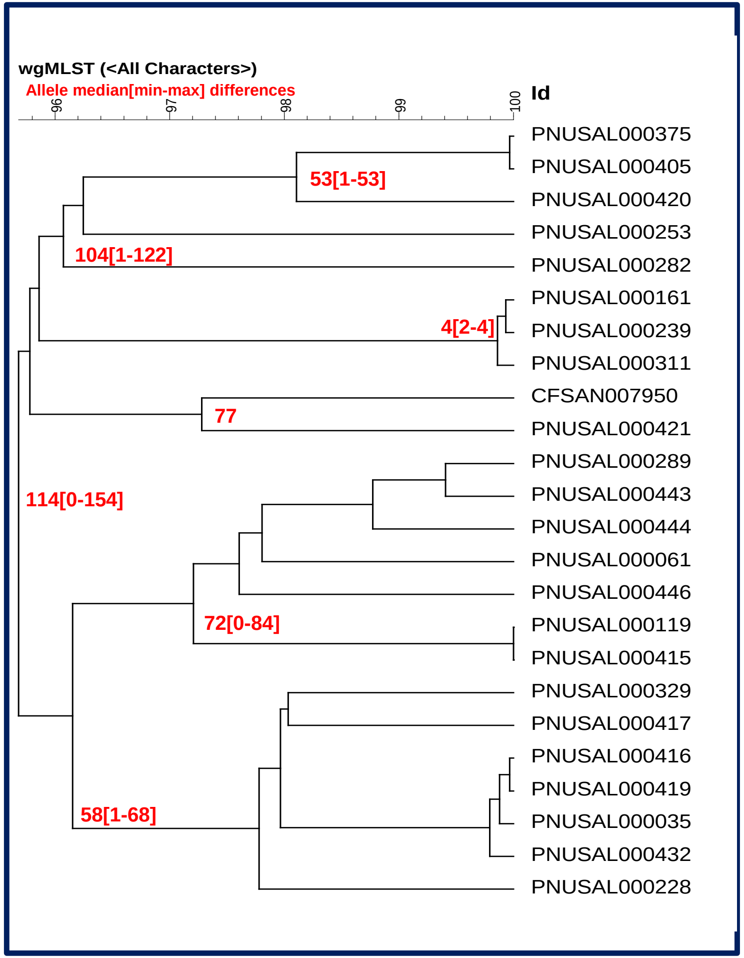
An Outbreak That Wasn't

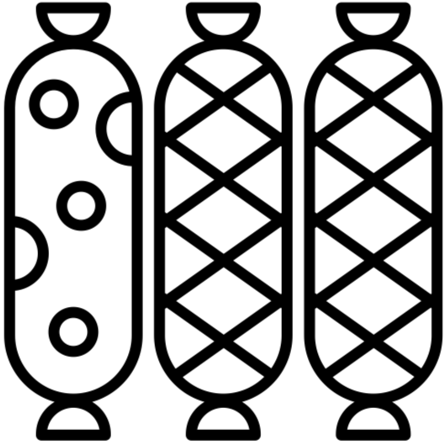


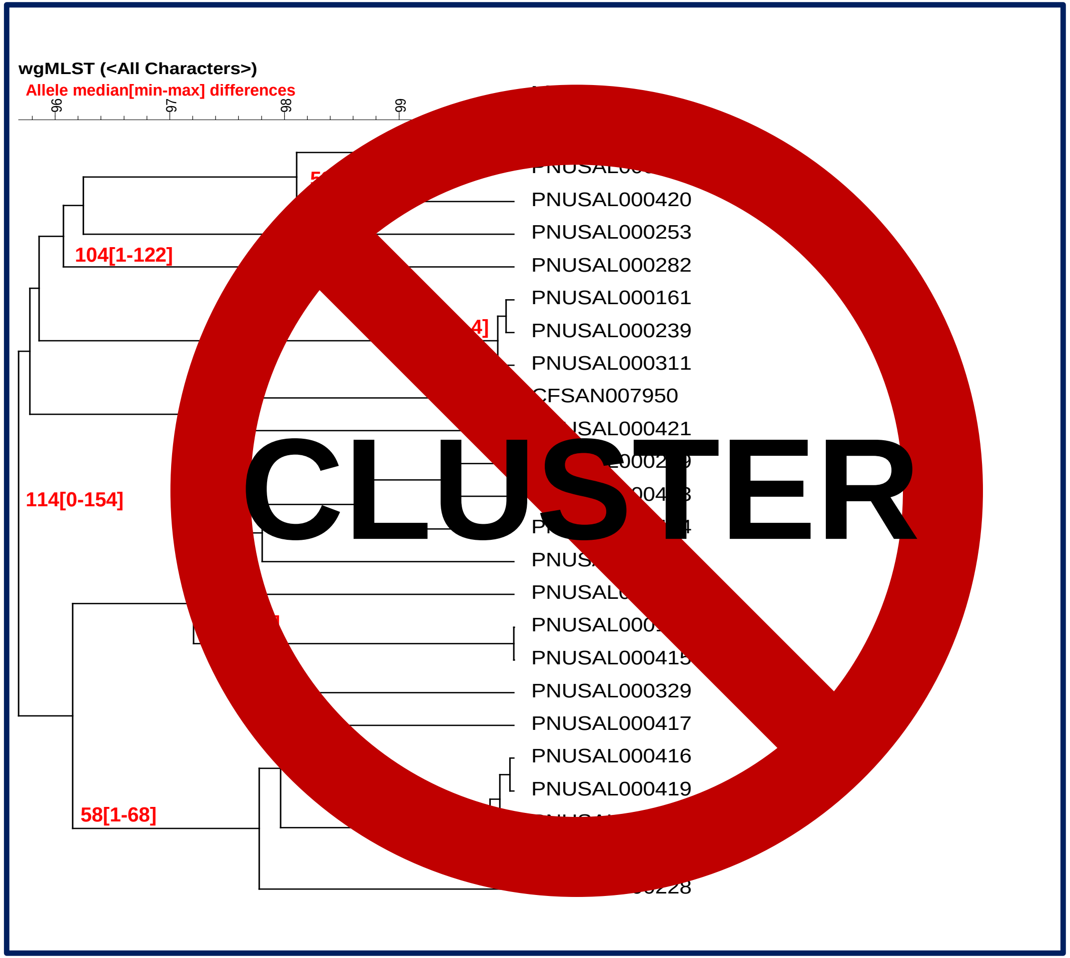
- 80% (12/15) reported eating deli meat
- Several reported uncommon deli meats (e.g., liverwurst, mortadella, ring bologna, cloth bologna)

PFGE Cluster Possibly Associated with Uncommon Deli Meats, 2012–2013

An Outbreak That Wasn't






wgMLST (<All Characters>)
Allele median[min-max] differences
96
97
98
99
104[1-122]
114[0-154]
58[1-68]
PNUSAL000420
PNUSAL000253
PNUSAL000282
PNUSAL000161
PNUSAL000239
PNUSAL000311
CFSAN007950
PNUSAL000415
PNUSAL000329
PNUSAL000417
PNUSAL000416
PNUSAL000419
CLUSTER

Improving Case Definition

Salmonella Enteritidis Outbreaks Related to Travel to Mexico (2015)

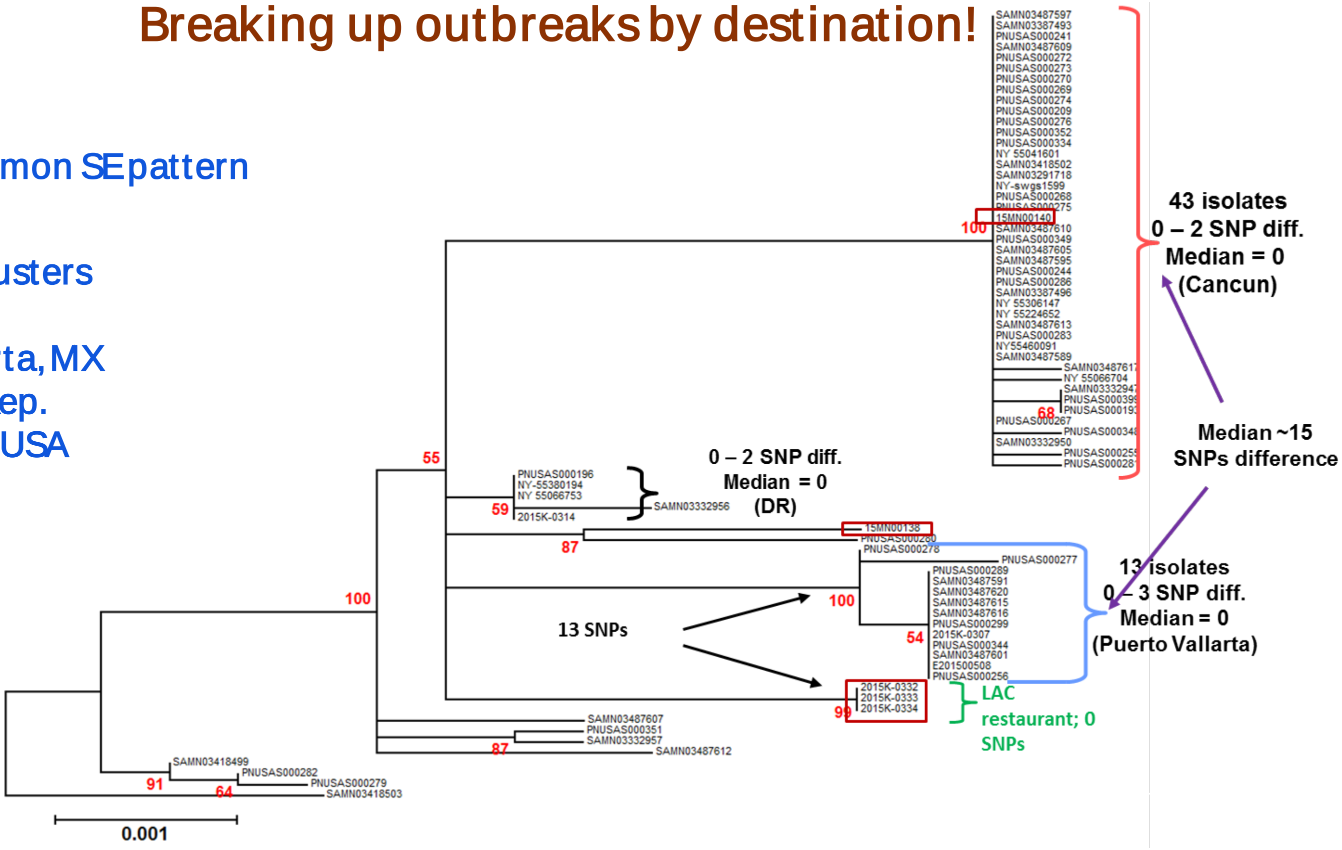
Breaking up outbreaks by destination!

Third most common SE pattern

~500 cases

-Distinct sub-clusters

- Cancun, MX
- Puerto Vallarta, MX
- Dominican Rep.
- Los Angeles, USA



43 isolates
0 – 2 SNP diff.
Median = 0
(Cancun)

Median ~15
SNPs difference

0 – 2 SNP diff.
Median = 0
(DR)

13 SNPs

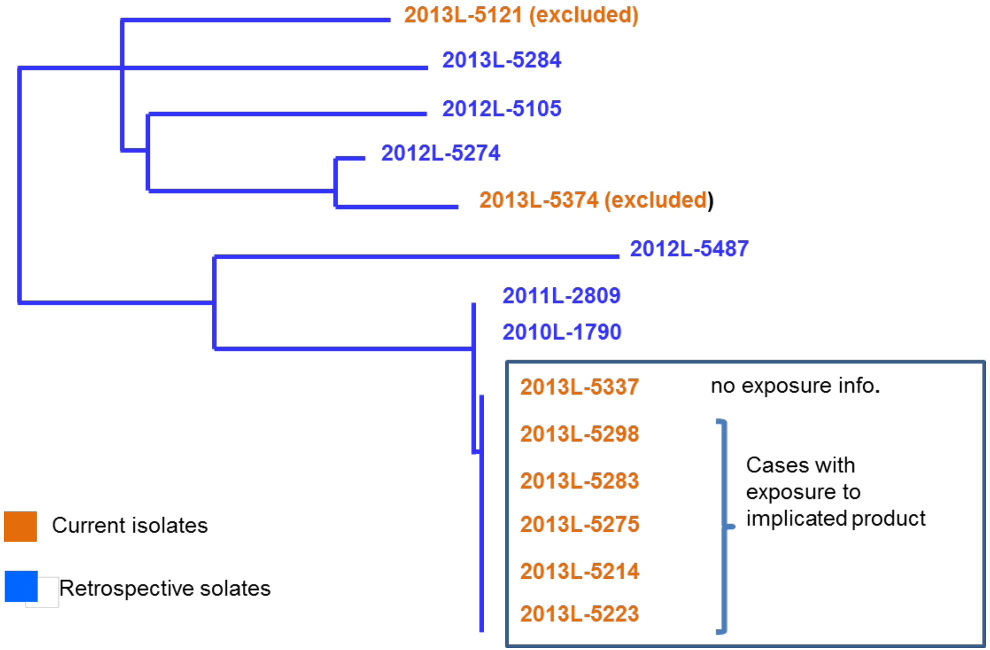
13 isolates
0 – 3 SNP diff.
Median = 0
(Puerto Vallarta)

LAC
restaurant; 0
SNPs

Improving Case Definition

Listeria Outbreak Linked to Artisan Cheese (2013)

All the same PFGE pattern



2013L-5121 (excluded)

2013L-5284

2012L-5105

2012L-5274

2013L-5374 (excluded)

2012L-5487

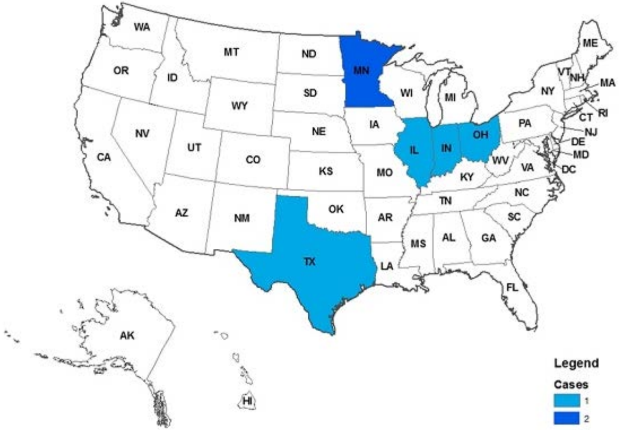
2011L-2809

2010L-1790

2013L-5337	no exposure info.
2013L-5298 2013L-5283 2013L-5275 2013L-5214 2013L-5223	Cases with exposure to implicated product

🟧 Current isolates

🟦 Retrospective solates



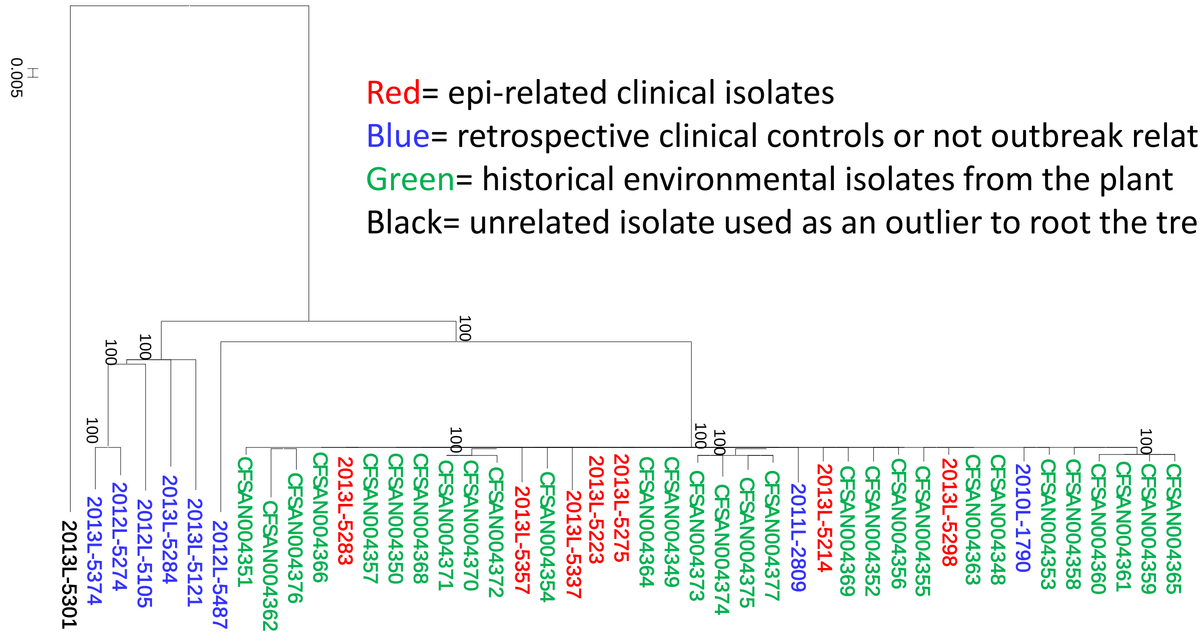
At a Glance:
- **Case Count: 6**
- **States: 5**
- **Deaths: 1**
- **Hospitalizations: 6**
- **Recall: Yes**

LONG-TERM PATHOGEN RESERVOIRS IN THE FOOD CHAIN

Listeria Outbreak Linked to Artisan Cheese (2013)

hqSNP

Historical isolates from the plant environment added to the comparison (courtesy FDA/CFSAN)



Red= epi-related clinical isolates
Blue= retrospective clinical controls or not outbreak related
Green= historical environmental isolates from the plant
Black= unrelated isolate used as an outlier to root the tree

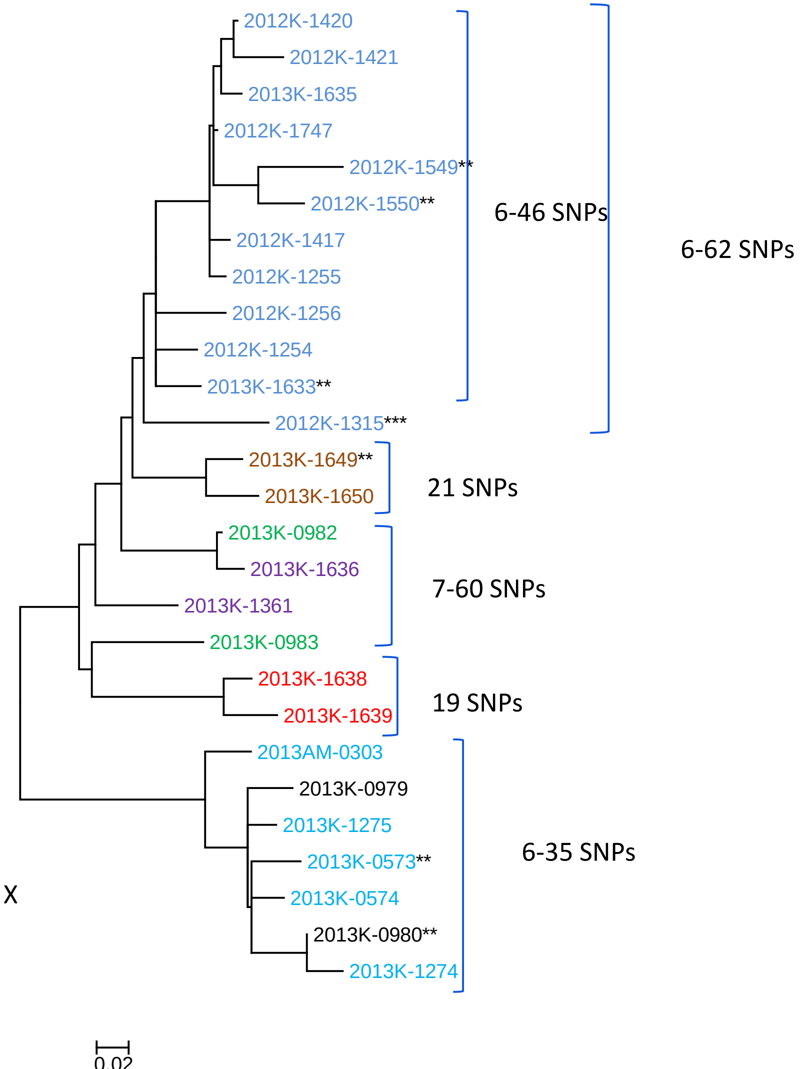
Salmonella Heidelberg multistate outbreak associated with chicken from manufacturer X (2012- 13)

Multiple PFGE patterns- By WGS Some Related, some not related

PFGE patterns

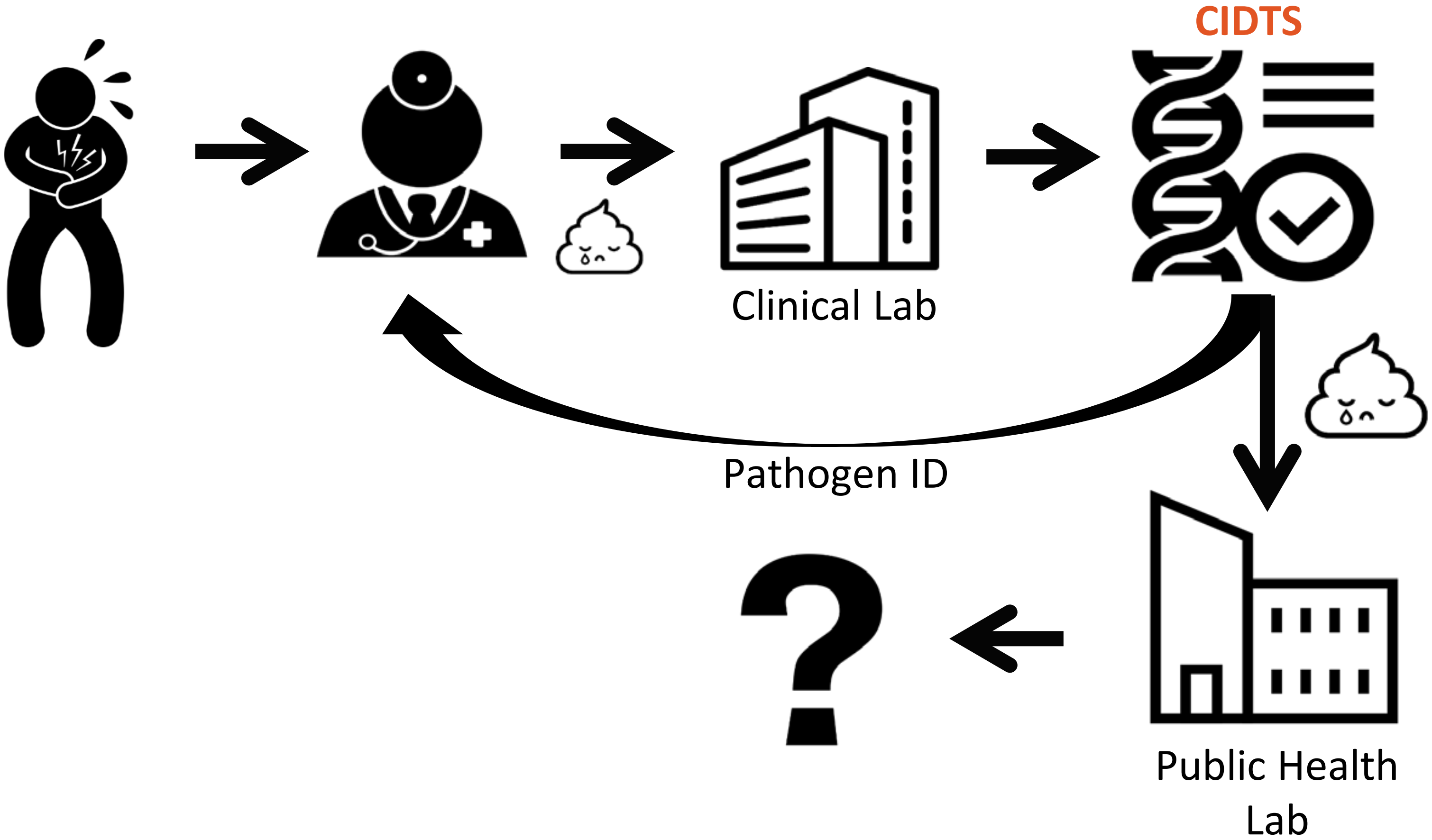
- 🟦 Pattern 122
- ⬛ Pattern 672
- 🟩 Pattern 45
- 🟪 Pattern 22
- 🟥 Pattern 326
- 🟦 Pattern 41
- 🟦 Pattern 258

** Chicken from manufacturer X
*** Clinical isolate from 2011

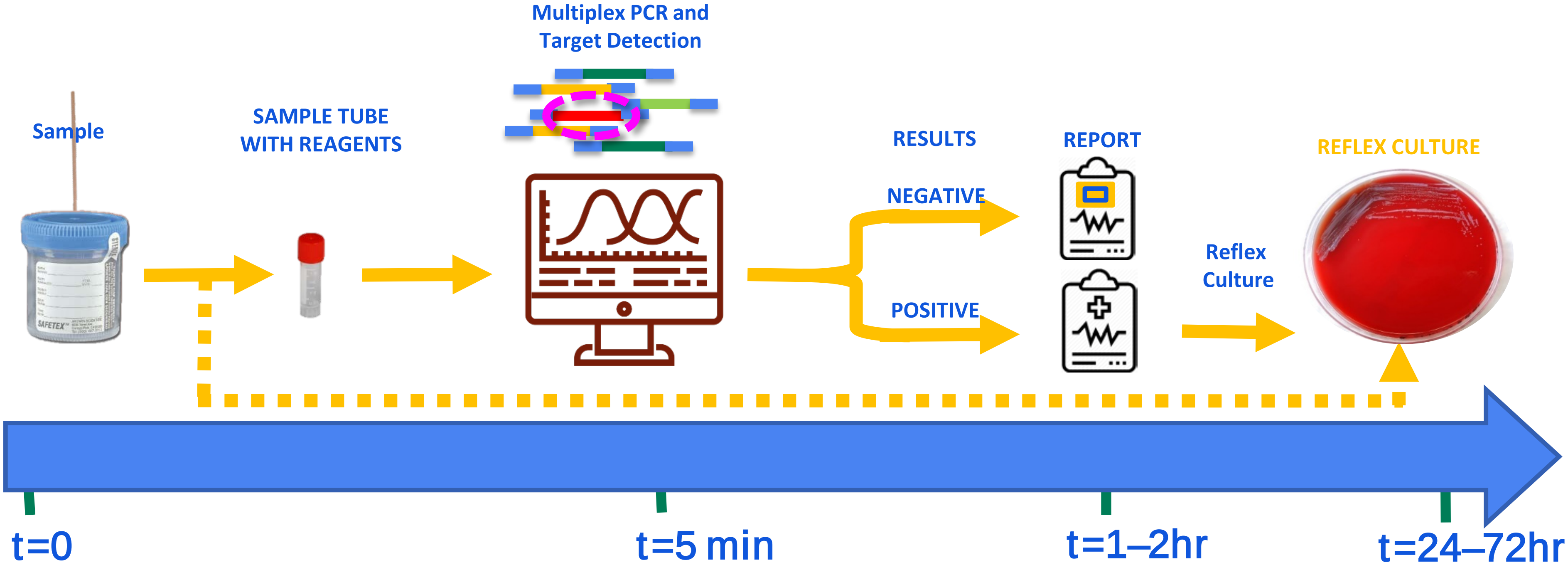


Beyond WGS

Public health surveillance workflow in transition



Why use CIDT – go from sample to answer faster



Sample

SAMPLE TUBE
WITH REAGENTS

Multiplex PCR and
Target Detection

RESULTS

NEGATIVE

POSITIVE

REPORT

Reflex
Culture

REFLEX CULTURE

t=0

t=5 min

t=1–2hr

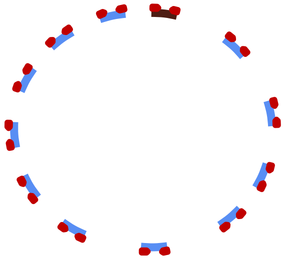
t=24–72hr

What is the new subtyping workflow?

- How to go from sample to subtype rather than isolate to subtype

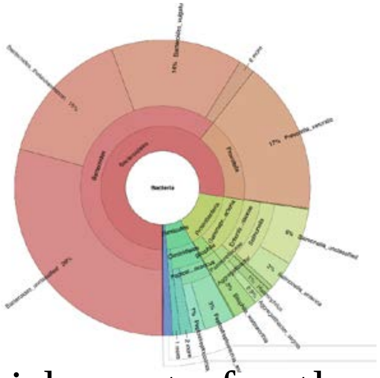
Approaches to sample based subtyping

Amplicon sequencing



1,000's of MLST targets

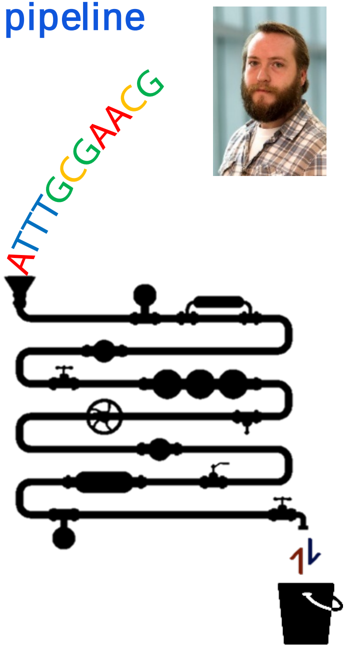
Shotgun metagenomics



- Enrichment of pathogen targets
- Unbiased sequencing

Amplicon sequencing approach for subtyping

Primer design pipeline

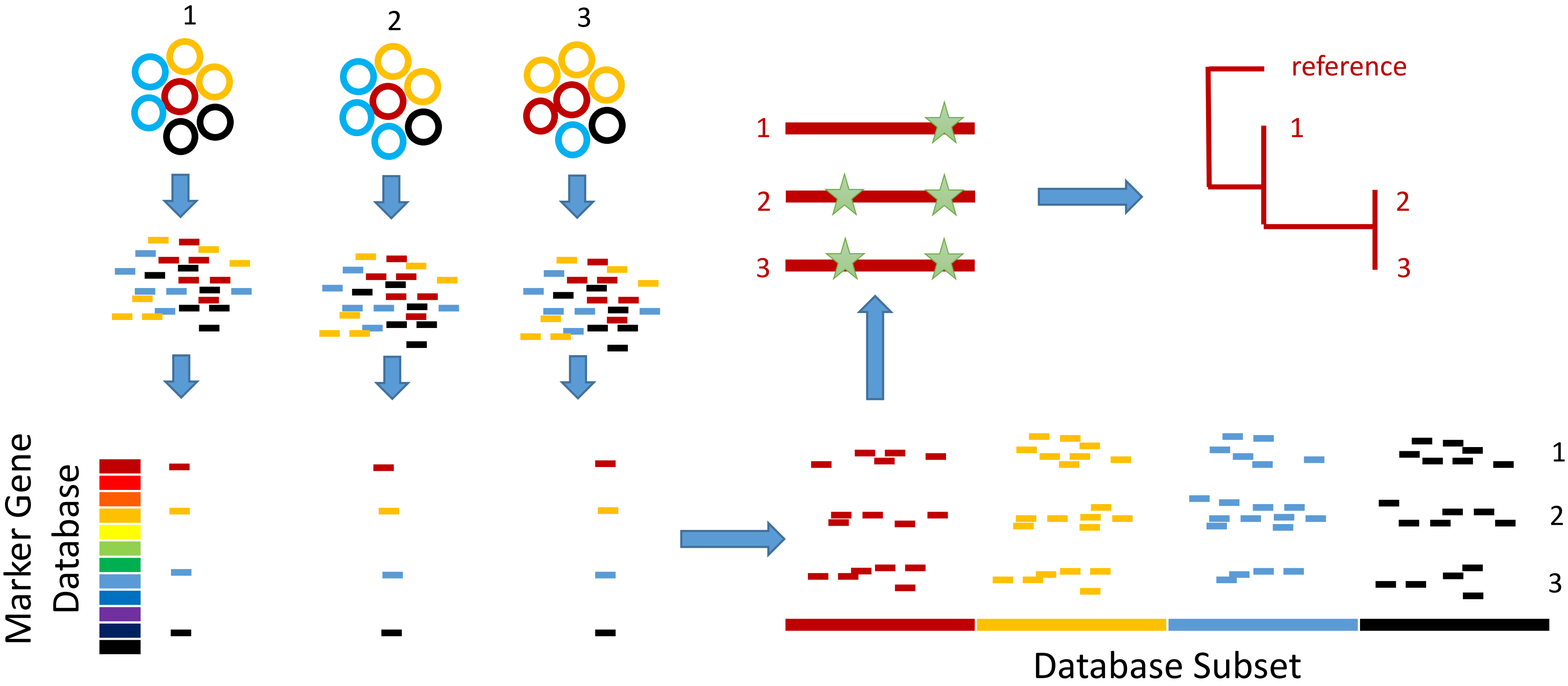


Resolution similar between amplicon and isolate WGS approach

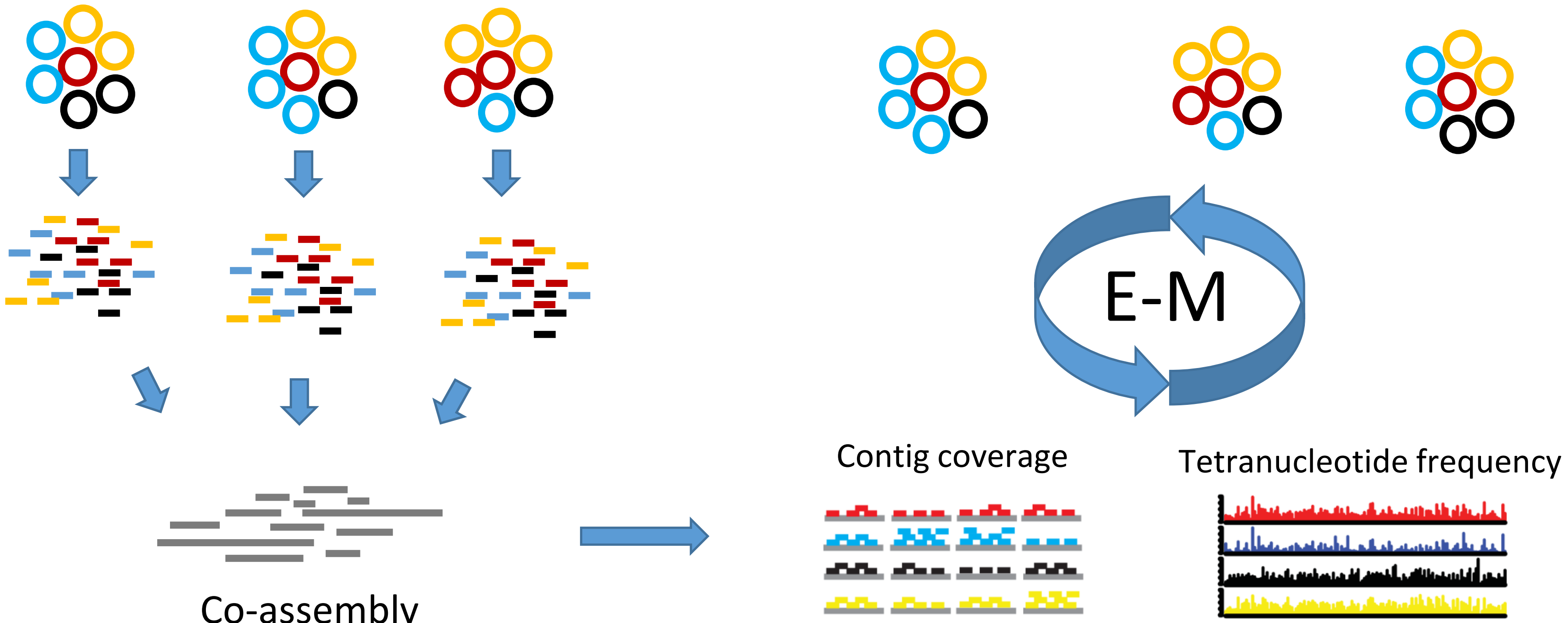
Strain-level characterization of metagenomics samples

- Challenge of strain level characterization of bacteria from stool – identifying databases and analysis tools to correctly identify and classify pathogen-associated sequence and AR genes

Reference-based Binning with MIDAS



Reference-free Binning with MaxBin



Co-assembly
(all of the samples together)

E-M

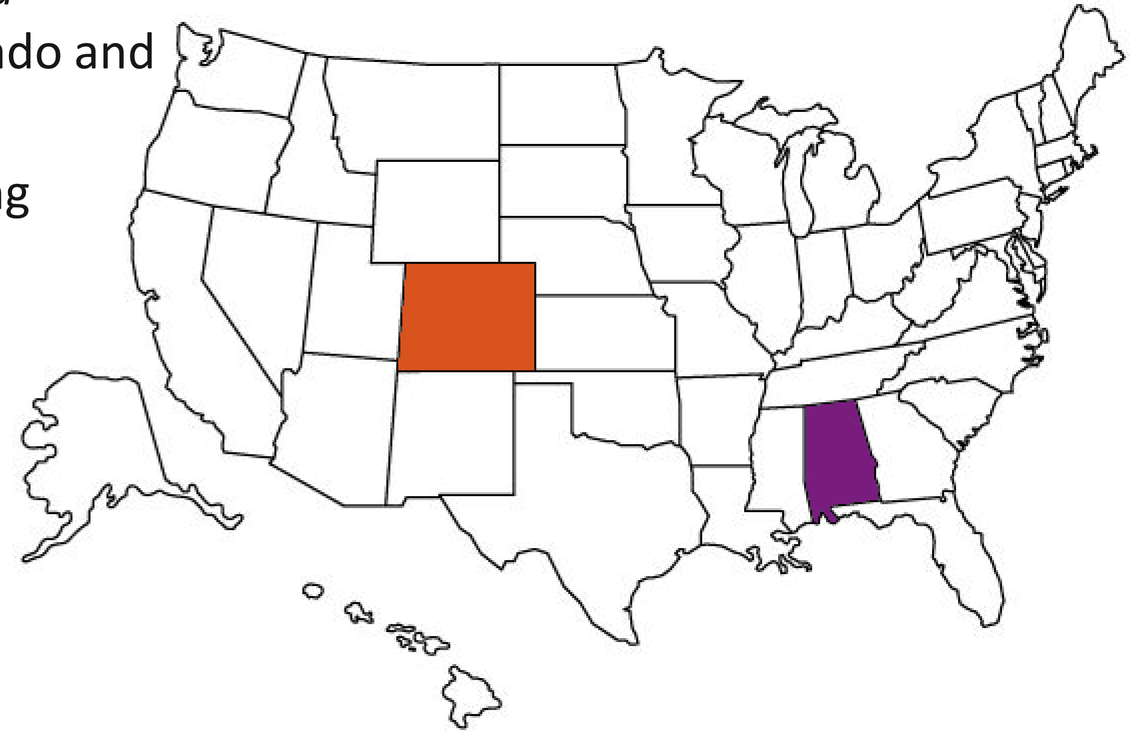
Contig coverage

Tetranucleotide frequency

Adapted from Kang *et al* (2015) *PeerJ*, 3, e1165, Figure 1

Metagenomics Case Study: Colorado and Alabama Outbreaks

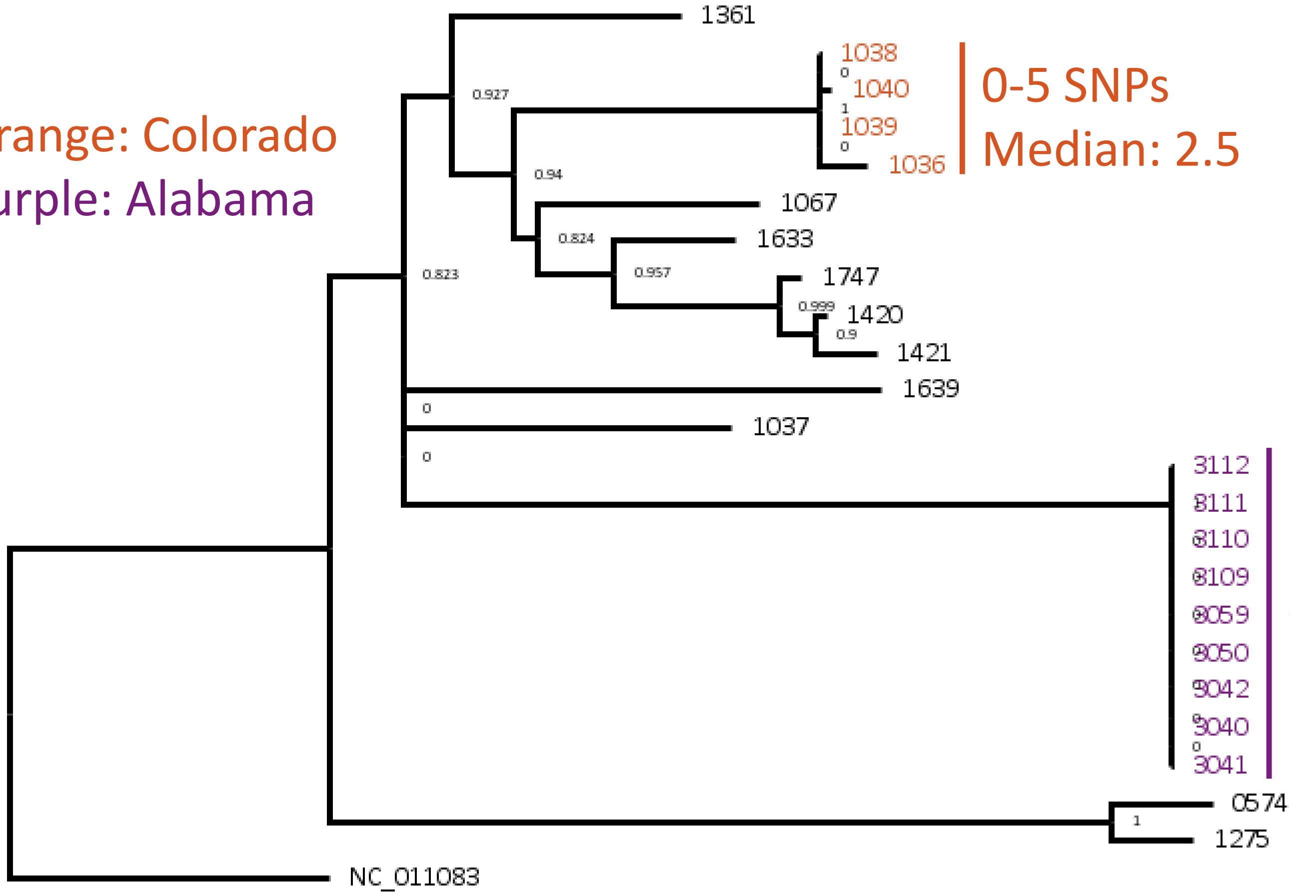
- Two similar 2013 *Salmonella enterica* outbreaks in Colorado and Alabama
- Originally distinguished using epidemiological and isolate genomic data



Isolate Tree (Gold Standard): Outbreaks are <u>Separated</u>

Orange: Colorado

Purple: Alabama



0-5 SNPs

Median: 2.5

0 SNPs

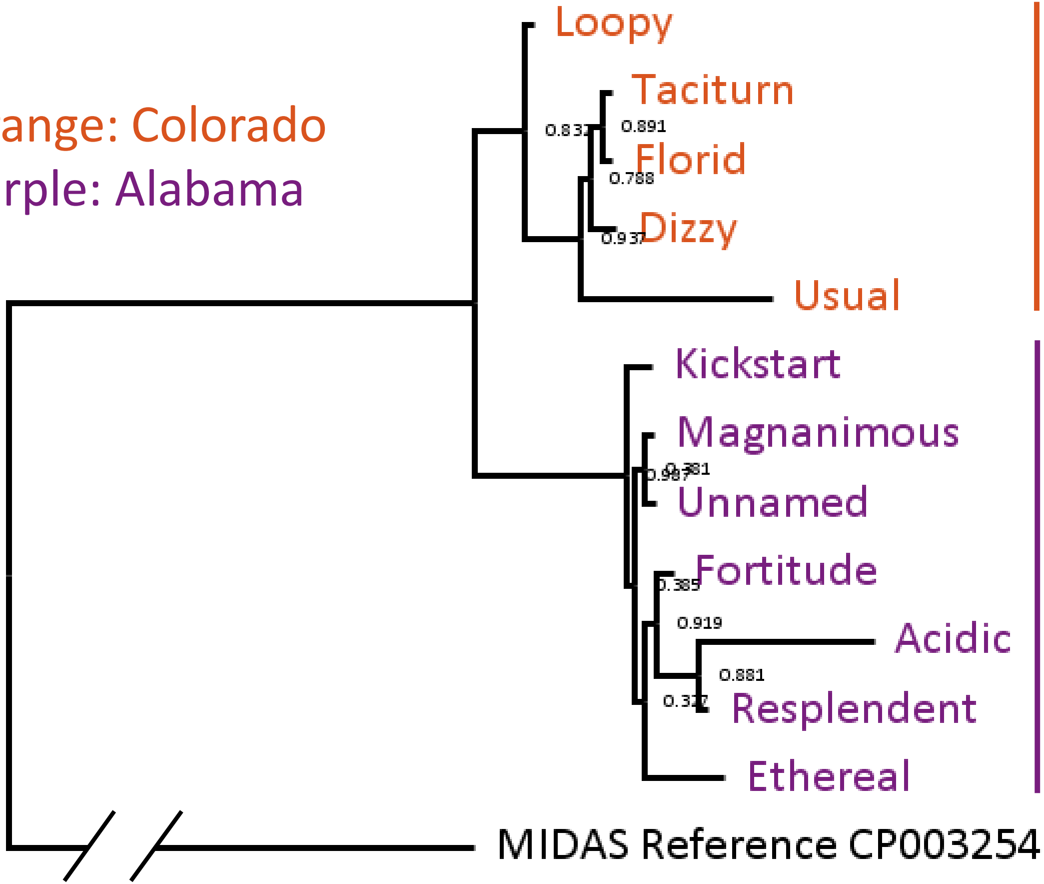
1-95 SNPs

Median: 79

MIDAS Tree (Reference-based): Outbreaks are Separated

Orange: Colorado

Purple: Alabama



1-15 SNPs

Median: 13

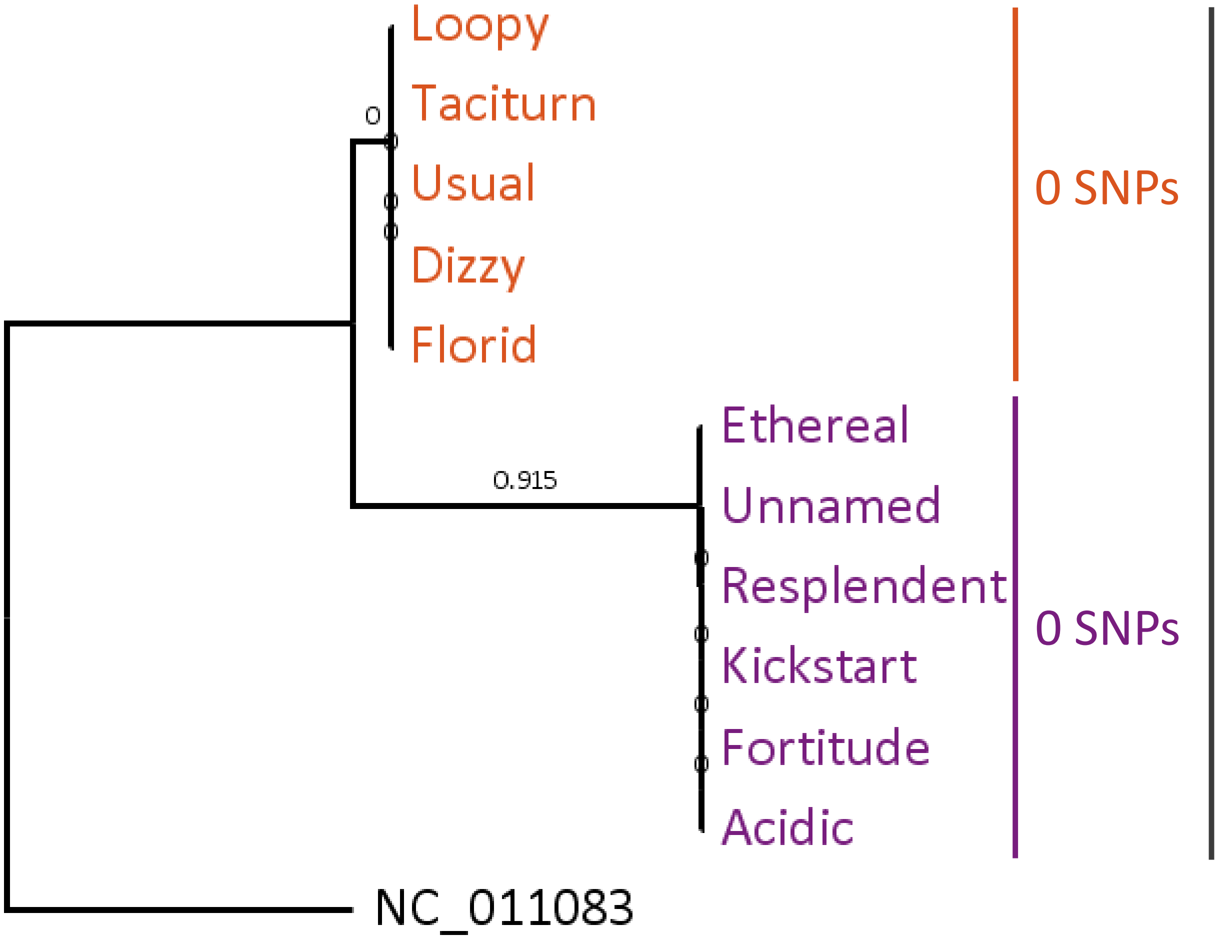
7-19 SNPs

Median: 17

1-33 SNPs

Median: 22.5

MaxBin Tree (Reference-free): Outbreaks are Separated



0-34 SNPs
Median: 34

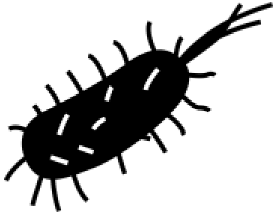
Orange: Colorado
Purple: Alabama

Conclusions

- WGS provides more resolution to identify outbreak clusters
- WGS data gives outbreak subtyping and genotyping information through the same workflow
- WGS is just the first step for PulseNet, next stop is culture independent subtyping methods

Noun project credits

- Gregor Cresnor
- Fahmionline
- Nikita Kozin
- Oli Mohr
- Made by Made
- Gan Khoon Lay



Acknowledgements

CDC's AMD Program

Advanced Molecular Detection

Innovate • Transform • Protect

APHL

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

CS 296673-A


WGS
Whole Genome Sequencing
Revolutionizing food safety with a clearer picture of bacterial DNA