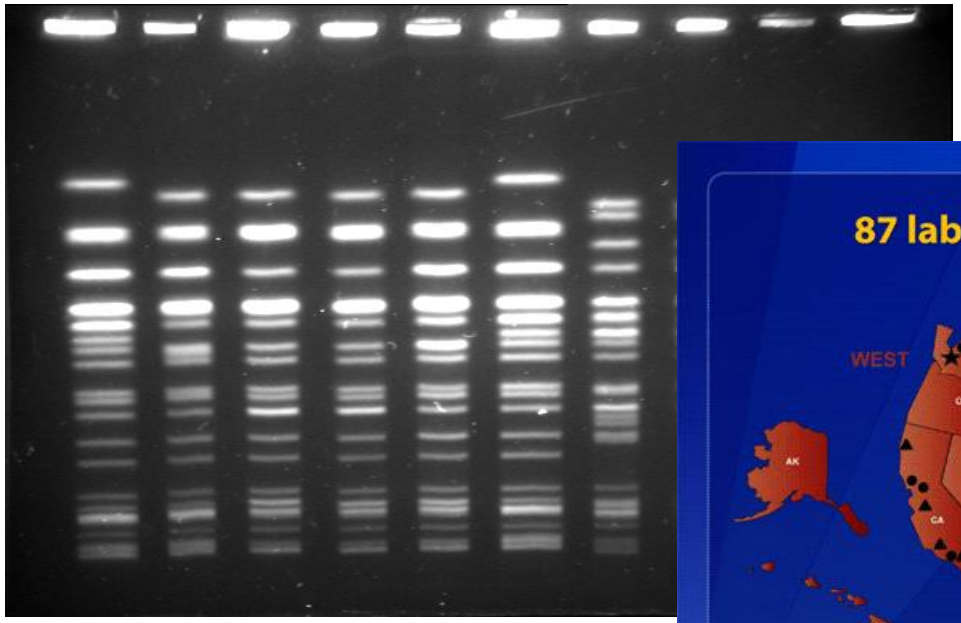
Wisconsin State Laboratory of Hygiene

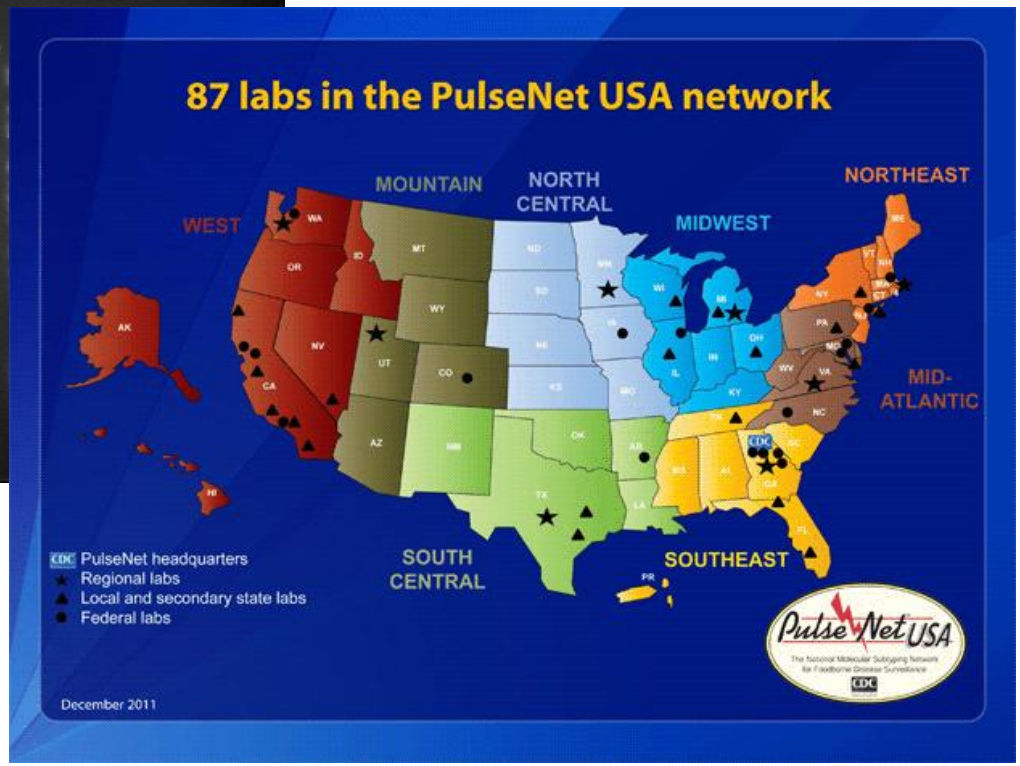
UNIVERSITY OF WISCONSIN–MADISON

Next Generation Sequencing for Outbreak Detection

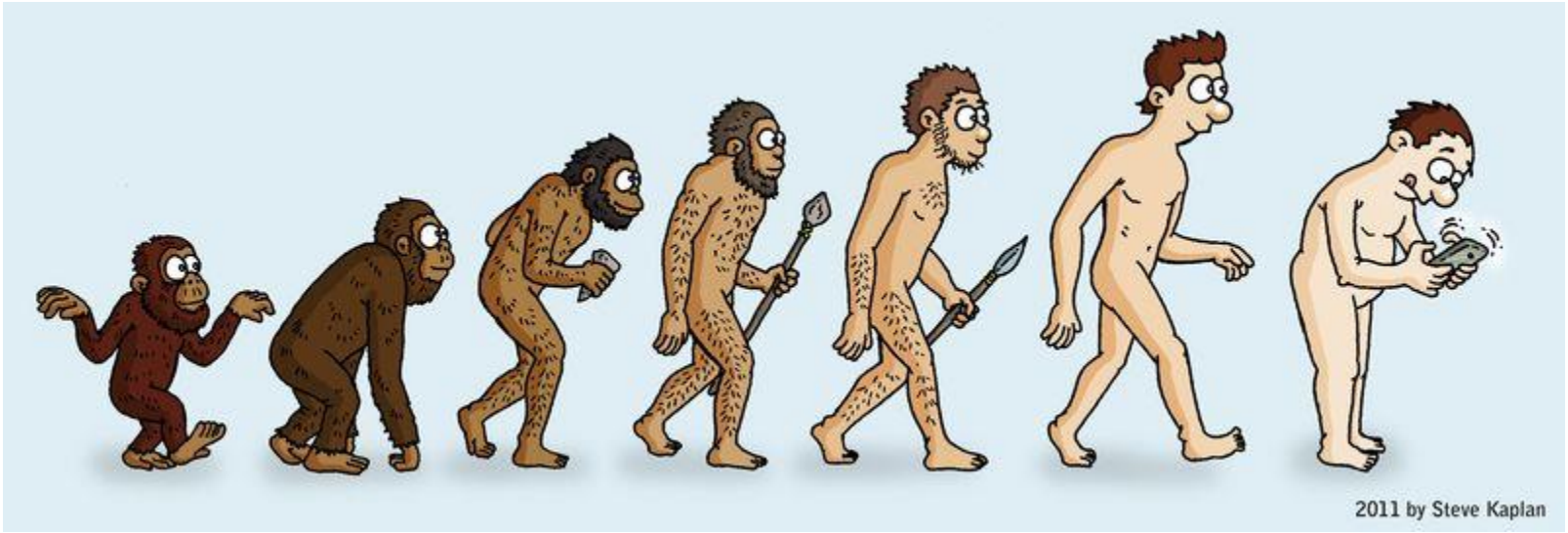
Dave Warshauer, PhD, D(ABMM)
Deputy Director, Communicable Diseases
Wisconsin State Laboratory of Hygiene
david.warshauer@slh.wisc.edu

PulseNet and PFGE



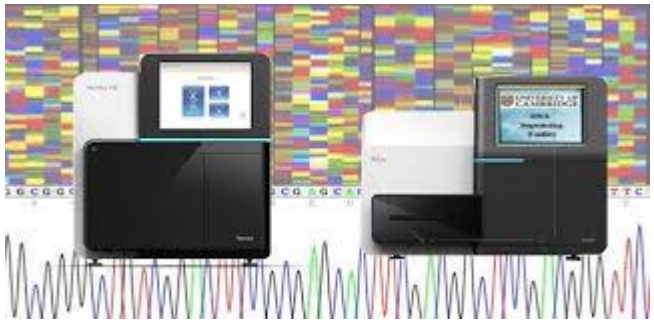


EVOLUTION



Why Evolution to WGS

- For PulseNet Labs
  - Consolidation of multiple workflows
    - ID
    - subtyping
    - Serotyping
    - antibiotic resistance
    - virulence factors



  - Fast---Decrease TAT to 2-4 days
  - Cheaper

Case: Shiga Toxin-Producing *E. coli* Cost Savings by Moving from Traditional Isolate Characterization to WGS

(Materials only)

Characterization of a Shiga toxin-producing *E. coli* isolate	Current testing costs	ID + characterization by WGS	
		MiSeq	NextSeq
Identification	$60		
Serotyping	$159		
PCR Virulence Profile – 4 targets	$10		
PFGE	$30		
MLVA	$15		
AST	$30		
WGS		$123	$60
Total	$304	$123	$60
Cost savings %		59%	80%

Annual cost savings based on # uploads to PulseNet in 2014:
$ (2239+3614) * (274-123)= $ 884,000

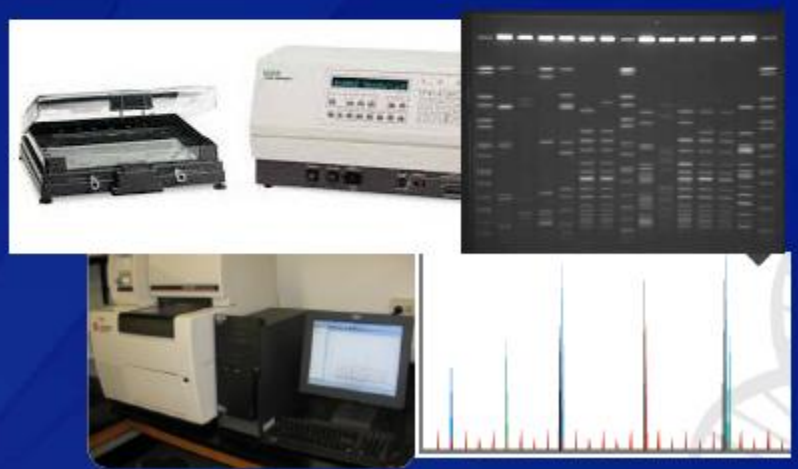
*Slide courtesy of Rebecca Lindsey/*Escherichia* Reference Lab/ EDLB

Key Characteristics of the Main WGS Platforms

Platform	Read length (bp)	Isolates per run (max)	Run time	Instrument cost	Cost/ Mb	Error rate (%)
Illumina HiSeq 2500	125	600-1000	5-11 d	$740K	$0.05	0.1
Illumina MiSeq	150, 250, 300	12-16	26 h, 36 h 65 h	$99K	$1.37	0.1
Illumina NextSeq	75, 150	96	29 h	$250K	0.03-0.07	0.1
IonTorrent PGM (314, 316, 318)	200, 400	1-10	2-8h	$75K	$0.93 - $7.5	1
Ion Proton	100-200	96	2-4 h	$245K	$0.02	1
PacBio RSII	10-40kb	8 /smrt cell	0.5-2 h	$750K	$180.00	16

Transforming Public Health Microbiology – PulseNet and Beyond

- **Why WGS for surveillance?**
  - Improved outbreak detection and investigation
    - More outbreaks detected earlier with fewer cases
    - More focused investigations
    - More information available in real-time (e.g. virulence, resistance)
    - Better alignment of cases and foods/environment
  - Improved trend and attribution analyses



Whole Genome MLST (wgMLST)

- **Gene – Gene Approach**
  - Multi-locus sequence typing (MLST) and/or identification of genes for reference characterization
  - Assess variations ('alleles') within each gene:
    - SNP(s), indels, rearrangements

'Locus' (gene)	Strain 1	Strain 2	Strain3	Strain 4
A	ACTAGAGGGAA allele 1	ACTAGAGGCAA allele 2	ACT-GAGGGTAA allele 3	ACGGGAGATAA allele 4
B	TAGCCAGGGTC allele 1	TAGCAAGGGTC allele 2	TAGC---GGTC allele 3	TAGGCAGGGTC allele 1
C, D, E, etc....	alleles 5,2,8…	alleles 1,4,7…	alleles 1,3,9…	alleles 6,2,9…

  - The gene- gene approach should give you all the information you need from a reference laboratory

- Plain language reporting of WGS reference data from wgMLST database

Centers for Disease Control & Prevention
National *Escherichia* and *Shigella* Reference Laboratory

Patient Name: [redacted]
Sex: **Female** Birthdate: Age: [redacted]

Public Health / International Submitter IDs

Patient ID: Alt. Patient ID:
Specimen ID: [redacted] Alt. Specimen ID:

CDC Specimen ID: **2014003970** CDC Unique ID: **N8K7DBC1** CDC Local ID: **2014C-3008**

GENUS/SPECIES: *Escherichia coli*

SEROTYPE: O104:H4*

PATHOTYPE: Shiga toxin producing and Enteroaggregative *E. coli* (STEC & EaggEC)

VIRULENCE PROFILE: *stx2a, aagR, aagA, sigA, sepA, pic, aatA, aaiC, aap*

SEQUENCE TYPE: ST34

ANTIMICROBIAL RESISTANCE GENES: bla_{TEM-1} , $bla_{CTX-M-15}$

All characteristics have been determined by whole genome sequencing (WGS)

Comments:

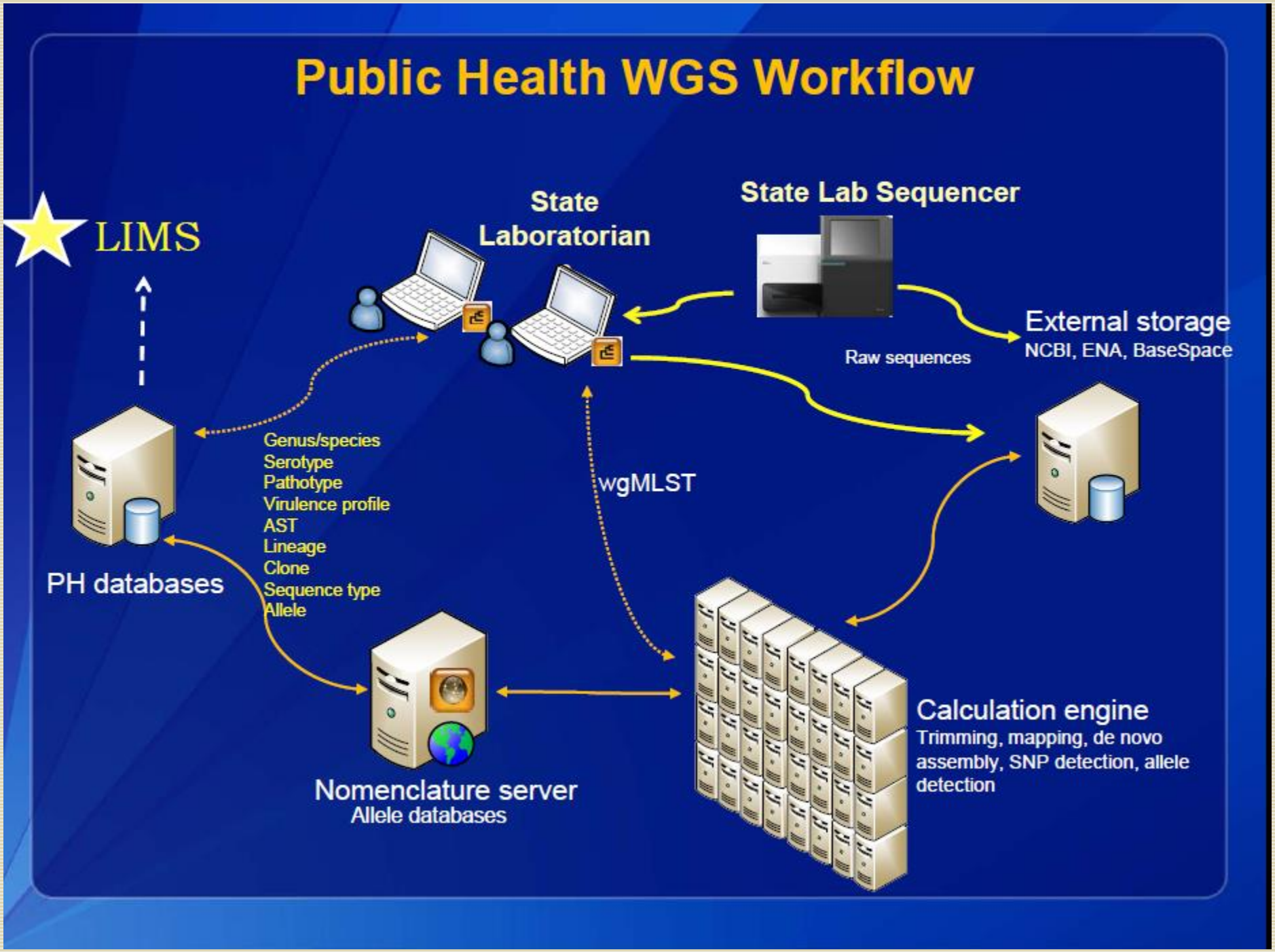
***Disclaimer** - This test has not been cleared or approved by the FDA. The performance characteristics have not been fully established. The results of this test should NOT be used for the diagnosis, treatment, or assessment of patient health or management.

Explanantion of Virulence Markers

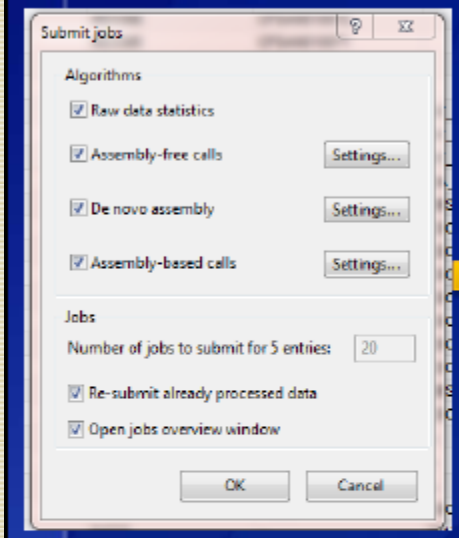
The strain contains Shiga toxin subtype 2a typically associated with virulent STEC
It does <u>not</u> contain adherence and virulence factors (*eae, ehxA*) typically associated with virulent STEC
It contains adherence and virulence factors typically associated with virulent EaggEc (*aagR, aagA, sigA, sepA, pic, aatA, aaiC, aap*)
This genotype is associated with extremely high (>10%) rates of hemolytic uremic syndrome (HUS)

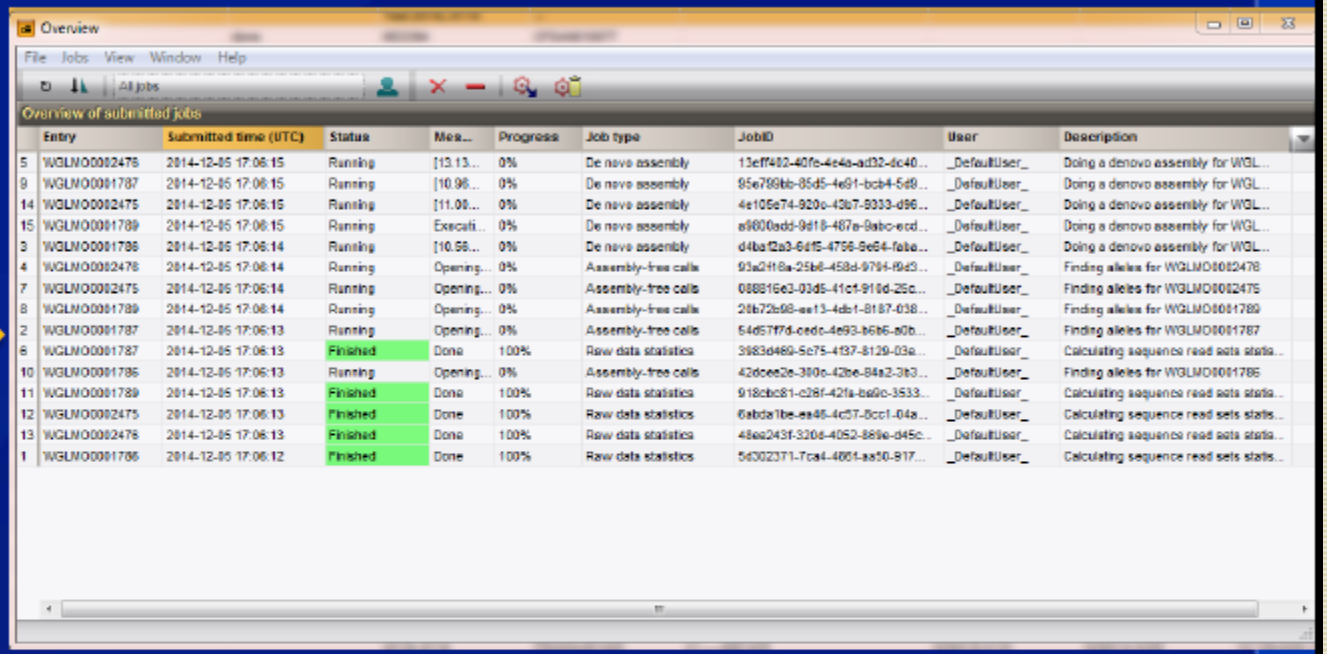
Approved by: Nancy Strockbine, Ph.D.
Ph: 404-639-4186
Fax: 404-639-3333
E-mail: nas6@cdc.gov

Public Health WGS Workflow



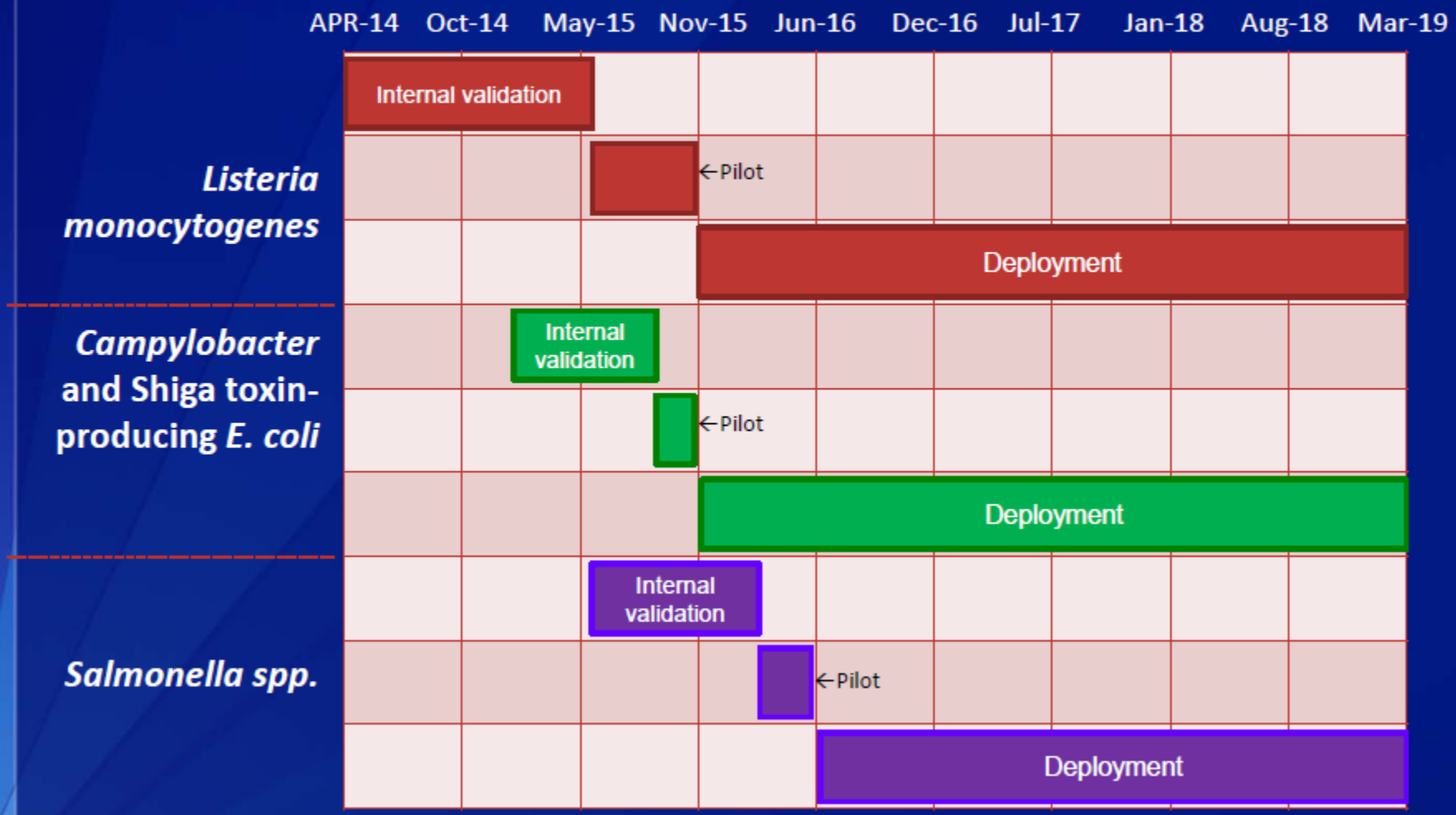
Analysis: Easy and Rapid





- No need for on-site bioinformatician or large computer resources; labs in network can submit sequence data for analysis to CDC computing resources

Projected wgMLST Database Validation and Deployment Timeline



PulseNet Sequencing Priorities with AMD Funds – Clinical Isolates

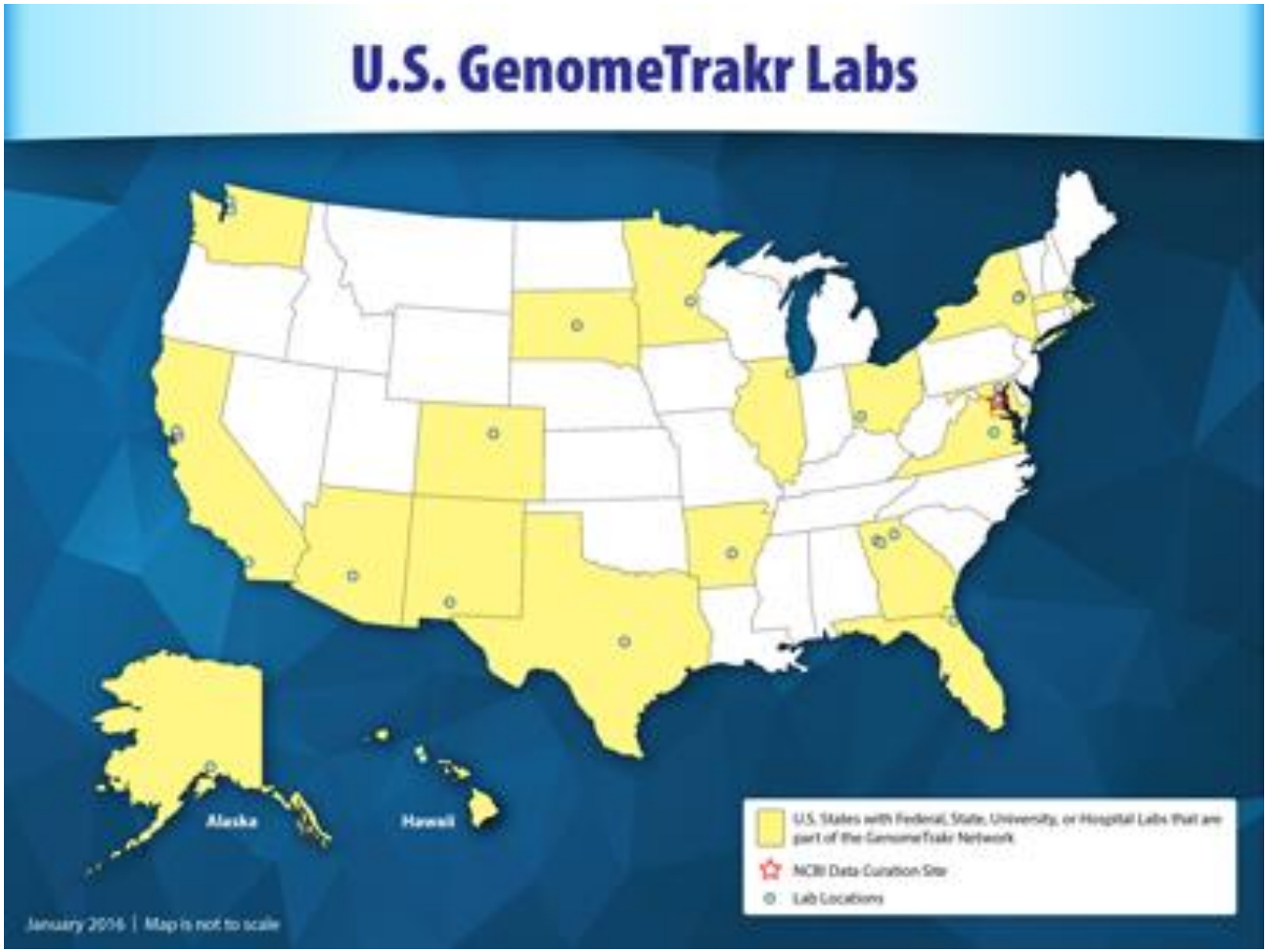
- ***Listeria monocytogenes:***
  - All isolates received by supported laboratories
- **Non-O157 STEC**
  - All sporadic isolates
  - No more than 3 from the same outbreak. If multiple PFGE patterns within an outbreak, include one representative of each pattern
- **O157 STEC**
  - Only if requested by CDC
    - One representative from each outbreak or if MLVA data is inconclusive

PulseNet Sequencing Priorities with AMD Funds – Clinical Isolates

- ***Campylobacter***
  - All isolates that are PFGE-tested and uploaded by your lab
- ***Salmonella***
  - Isolates requested by CDC
  - Prioritization based on epi data
- **Each laboratory expected to sequence about 150-300 isolates during the first year**

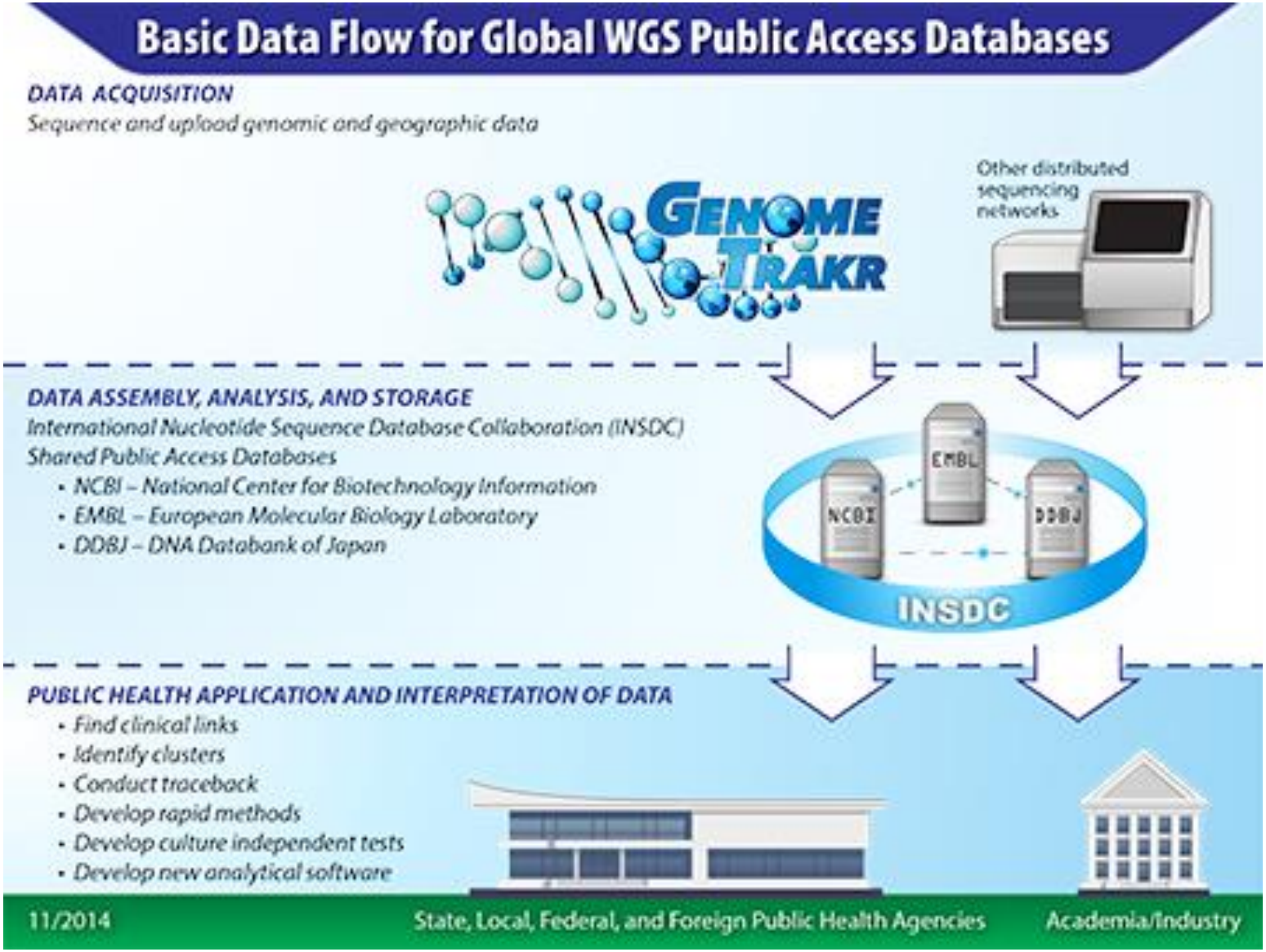
Genome Trakr Network

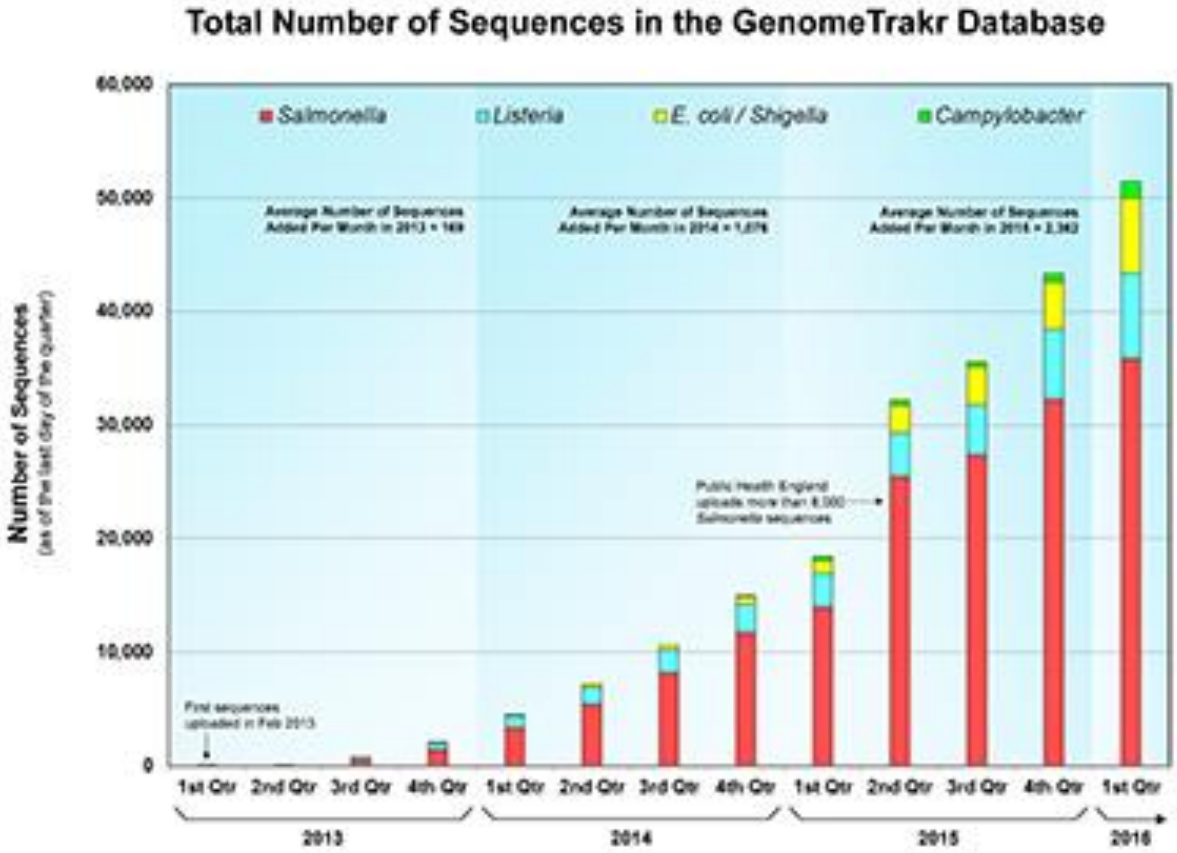
- FDA led network
  - Public health labs
  - University and hospital labs
  - Federal labs
- Fooodborne pathogens
  - Foodborne outbreaks
  - Contaminated food products
  - Environmental sources
- Data in an open-access genomic reference database called Genome Trakr at the NCBI


U.S. GenomeTrakr Labs
Alaska
Hawaii
U.S. States with Federal, State, University, or Hospital Labs that are part of the GenomeTrakr Network
NCBI Data Curation Site
Lab Locations
January 2016 | Map is not to scale

Purpose of Genome Trakr

- Find contamination sources of outbreaks
- Better understand the environmental conditions associated with the contamination of agricultural products
- Help develop new rapid methods and culture independent tests
- Monitor emerging pathogens
- Determine persistence of pathogens in the environment


Basic Data Flow for Global WGS Public Access Databases
DATA ACQUISITION
Sequence and upload genomic and geographic data
GENOME TRAKR
Other distributed sequencing networks
DATA ASSEMBLY, ANALYSIS, AND STORAGE
International Nucleotide Sequence Database Collaboration (INSDC)
Shared Public Access Databases
• NCBI – National Center for Biotechnology Information
• EMBL – European Molecular Biology Laboratory
• DDBJ – DNA Databank of Japan
NCBI
EMBL
DDBJ
INSDC
PUBLIC HEALTH APPLICATION AND INTERPRETATION OF DATA
• Find clinical links
• Identify clusters
• Conduct traceback
• Develop rapid methods
• Develop culture independent tests
• Develop new analytical software
11/2014
State, Local, Federal, and Foreign Public Health Agencies
Academia/Industry


Total Number of Sequences in the GenomeTrakr Database
Number of Sequences
(as of the last day of the quarter)
60,000
50,000
40,000
30,000
20,000
10,000
0
Salmonella
Listeria
E. coli / Shigella
Campylobacter
Average Number of Sequences Added Per Month in 2013 = 169
Average Number of Sequences Added Per Month in 2014 = 1,076
Average Number of Sequences Added Per Month in 2015 = 2,362
Public Health England uploads more than 8,000 Salmonella sequences
First sequences uploaded in Feb 2013
1st Qtr
2nd Qtr
3rd Qtr
4th Qtr
2013
2014
2015
2016

WGS in a Foodborne Outbreak

- 2010 nationwide salmonellosis outbreak
  - Over 1,900 illnesses with S. Enteritidis associated with eggs
  - Shared a common PFGE pattern
    - Pattern 4 (JEGX01.004)
    - 40% of SE share this pattern
    - Could not determine if the increase in illnesses was due to a single outbreak or multiple outbreaks

Salmonella Enteritidis Outbreak

- Traceback investigation pointed to two egg producers in the Midwest
  - Environmental specimens collected
  - Positive for SE with same PFGE pattern from both producers
    - Indistinguishable from clinical isolates
    - Because of common pattern could not know for sure if they were the same strain

WGS Comes to Save the Day

- Genomic sequences for the SE found at the 2 egg producers very closely related, but distinguishable
- Egg producer sequences also closely related to the sequences from the clinical samples
  - So closely that they were both deemed to match
  - Because there was a slight difference in the genomes, investigators could further delineate the specific egg processor to which an individual illness was linked

Outbreak 2
Cereal Killer

Cereal Killer

- 2008 outbreak of Salmonella Agona in 33 people linked to Midwest dry cereal manufacturer
- 3 different PFGE patterns
  - 3 different sources of contamination?
- WGS revealed isolates had a recently common lineage and were in fact the same strain

Cereal Killer

- CDC archived isolate of SA collected in 1998 from an outbreak linked to the same cereal manufacturer
  - WGS showed it virtually identical to the strain causing the 2008 outbreak
- How?
  - Would expect greater genetic diversity over 10 years

The Theory

- 1998 outbreak linked to contaminated water
- Water also used in 1998 renovation in the mortar
- SA lay dormant in the mortar
- 2008 another renovation
  - Mortar was disrupted and SA released into the environment
  - Multiplied and contaminated the plant and the cereal product

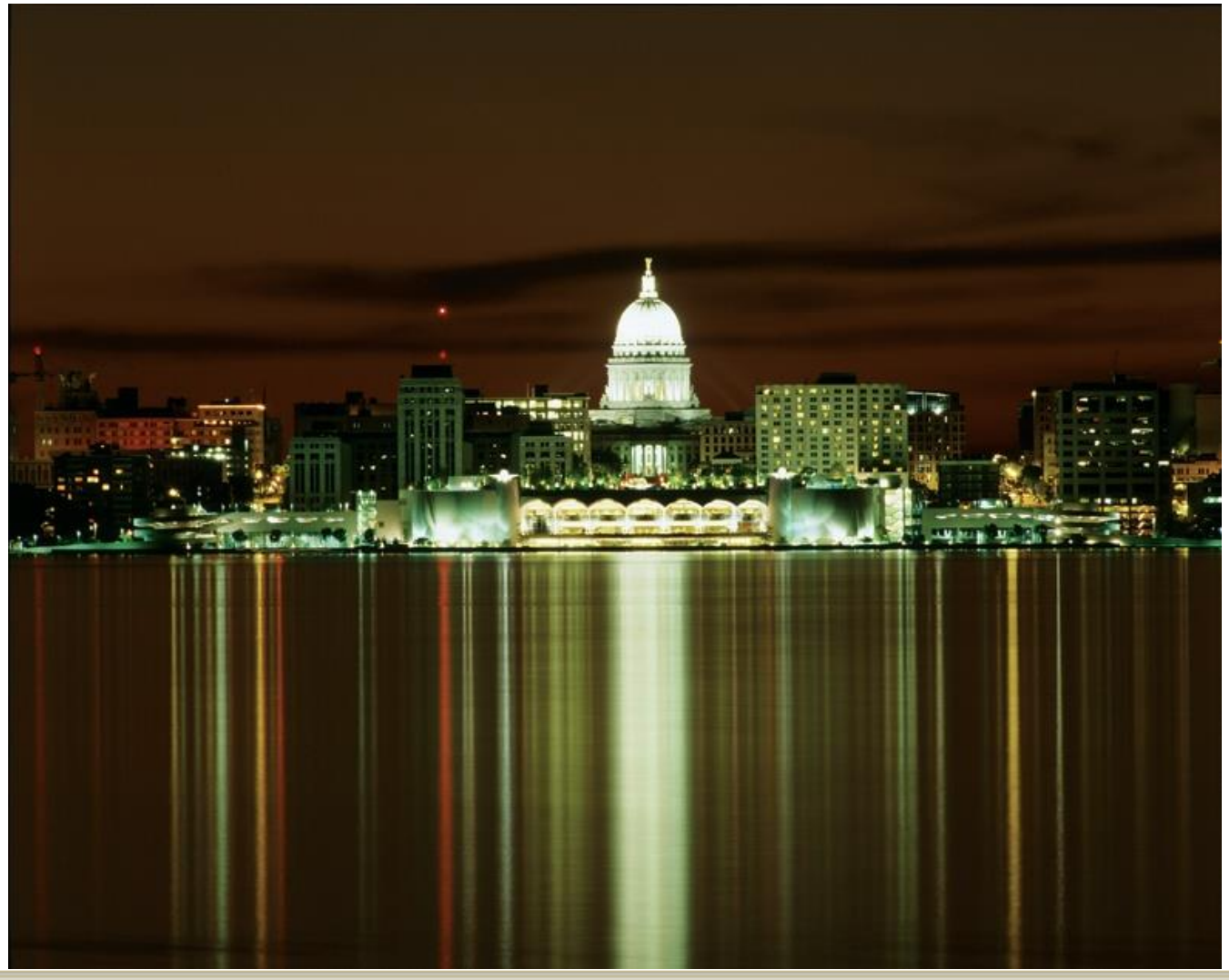
Other WGS at WSLH

- Influenza A and B directly from specimens
  - Partnership with CDC for surveillance of genetic changes and selection of vaccine strains
- Norovirus
- Vaccine Preventable Diseases
  - Measles
  - Mumps
  - Rubella

Acknowledgements

- Heather Carleton Ph.D., M.P.H., CDC
- The Staff of WSLH

Thank You



Case: *Campylobacter* Cost Savings by Moving from Traditional Isolate Characterization to WGS

(Materials only)

Characterization of a *Campylobacter jejuni* isolate	Current testing costs	ID + characterization by WGS	
		MiSeq	NextSeq
Identification	$74.20		
PFGE	$30.00		
MLST	$71.80		
AST	$20.00		
WGS		$73	$54
Total	$196.00	$73	$54
Cost savings %		63%	72%

Annual cost savings based on # uploads to PulseNet in 2014:

$ 3346 * (104.20-73)= $ 105,000

*Slide courtesy of *Campylobacter* Reference Lab/ EDLB

Comparison of Different Benchtop Platforms

Factors	Illumina MiSeq	Ion Torrent PGM
Time from DNA to Sequence*	30 -32 hours	30 hours
Total hands on time*	3-4 hours	8 hours
Number of isolates per run	12-16	1-12
Cost per isolate**	~ $85-125 (12-16 isolates/ 300 cycle cartridge)	~$306-325 (4-5 isolates/316 chip)
Data Quality	Q30	Q20

*Using in-house protocols

**Our costs for our most common usages of this technology; cost will go up for MiSeq if you have less isolates to multiplex; costs will change for PGM based on chip used and chemistry

PulseNet AMD Funds to States in 2014

- **PulseNet FY 2014 support through AMD funding for the states**
  - ~ $750,000 through ELC for 6 states (CT, MI, MN, OH, TN, WI)
    - Instrumentation, personnel, reagents
    - Mainly FoodCore states
  - $100,000 as direct Illumina reagent support for 4 states
    - Mainly GenomeTrakr labs (MD, NY, VA, WA)
  - 2 bench training laboratory workshops: Aug 4-7, Sept 22-25
    - 10 laboratorians trained in MiSeq sequencing and limited data analysis

Library Preparation: Multiplexing

- **Illumina MiSeq**
  - 300 cycle cartridge (60 MB)
    - 16 *Listeria*
    - 12 *Salmonella/Ecoli*
    - 30 *Campylobacter*
  - 500 cycle cartridge (80 MB)
    - 20-22 *Listeria*
    - 16 *Salmonella/Ecoli*
    - 45-50 *Campylobacter*
- **Ion Torrent PGM**
  - 314 chip
    - 2 *Listeria*
    - 1 *Salmonella/Ecoli*
  - 316 chip
    - 5 *Listeria*
    - 3 *Salmonella/Ecoli*
  - 318 chip
    - 10 *Listeria*
    - 6 *Salmonella/Ecoli*

Results of Benchtop NGS Comparison

Factor	MiSeq	PGM
Coverage	128 (58x-266x)	47x (21x-73x)
Contigs per assembly	22 (assembled using CLC)	28 (assembled using MIRA)
N50	391,927	306,604
hqSNP calls	0-2 differences	
wgMLST loci detected	Average of 16 more loci identified by MiSeq	
wgMLST allele call differences	0-2 discrepancies	

* Preliminary analysis suggests data is compatible to use in surveillance and outbreak detection