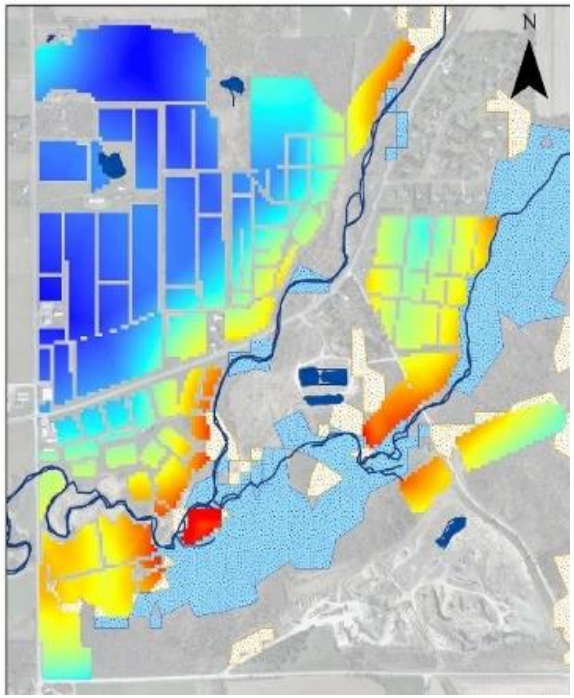
Cornell University
College of Agriculture and Life Sciences

Advanced technologies to improve food safety: WGS and beyond

Martin Wiedmann
Gellert Family Professor of Food Safety
Department of Food Science
Cornell University, Ithaca, NY
E-mail: mw16@cornell.edu
Phone: 607-254-2838

The age of precision food safety

- The same tools used in “precisions agriculture” and “personalized medicine” are leading to a paradigm shift in food safety
  - Improved outbreak detection and surveillance
  - Improved source tracking
  - Improved definition and identification of pathogens
  - Shift to a predictive and pro-active rather then reactive approaches
- “omics” tools are important for “precision food safety”, but other tools are part of this
  - Machine learning, Artificial Intelligence, Geographic information system (GIS)



Outline

- **Background: NGS, WGS, metagenomics, data analysis etc.**
- Use of WGS for Improved outbreak detection
- Use of WGS and other subtyping tools for source tracking and to understand pathogen persistence
- Use of WGS to better define pathogens and food safety hazards*: Listeria monocytogenes* and *Salmonella* as case studies
- Metagenomics
- Challenges

Sequencing technologies

- Sanger Sequencing (First generation)
  - Most widely used sequencing technology for approximately 25 years
- Next Generation Sequencing (NGS)
  - Also referred to as high-throughput sequencing
  - Includes most sequencing technologies that came after Sanger sequencing
  - Technologies
    - Second Generation (Short-read)
      - **Illumina**
      - Ion Torrent
      - 454 pyrosequencing (*Legacy Technology*)
    - Third Generation (Long-read)
      - Pac-Bio
      - Oxford Nanopore

MinION

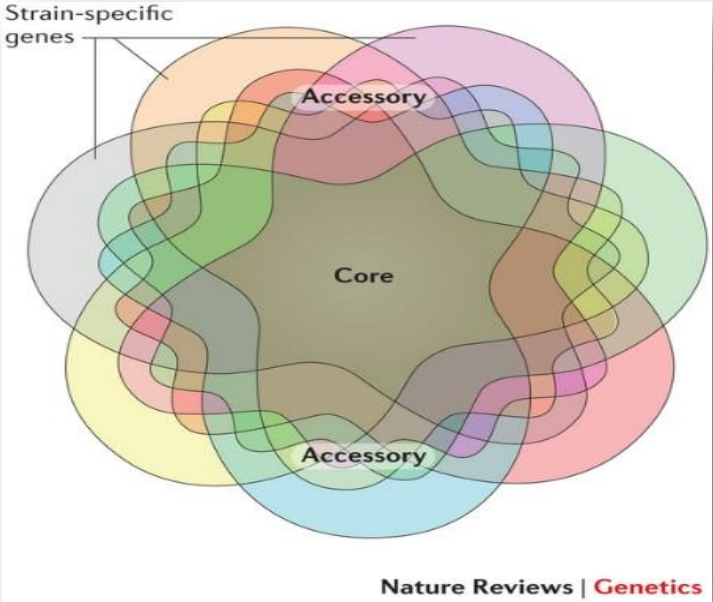
MinION

NGS and WGS

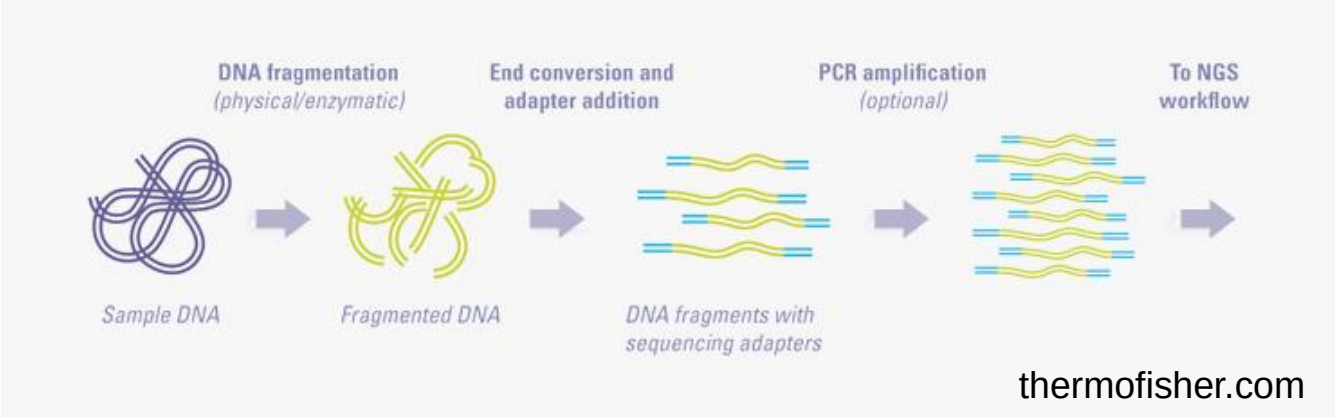
- **Whole genome sequencing (WGS)** can be performed by Sanger or NGS
- NGS can be used for many things other than WGS:
  - Metagenomics ("sequence all DNA found in a sample")
  - Highly parallel sequencing of targeted genes (identify and sequence hundreds of specific genes in a given sample)
  - RNA-seq ("sequence all genes that are active in a given sample")

Bacterial Genome

- Typically between 2 and 10 Mb (mega bases)
- Core genome: genes present in most/all strains of a given species
- Accessory genome: up to thousands of genes, not always present; can have very important functions (antimicrobial resistance, toxins)
  - Includes plasmids, prophages etc.
- "Whole genomes sequencing" rarely provides a sequence for the **whole** genome (unless genomes are finished, which today is not routinely done for standard applications)



Workflow overview



Assembly

– Reference guided mapping – When you have "map" of final product, and try to match your pieces together to look like final product
  - Example: Jigsaw puzzle – you have picture and put pieces together

– De Novo assembly– Put pieces together by what "makes sense"
  - Example: Jigsaw puzzle – you may not have picture, but can put pieces together by what fits together

Data analysis part 2: What is a SNP?

- Single Nucleotide Polymorphism (SNP)

```
ATGTTCCTC sequence
ATGTTGCTC reference
```

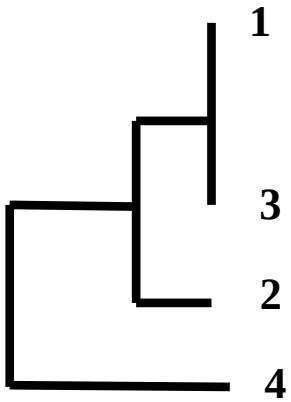
- Insertion or Deletion (Indel)

```
ATGTTCCCTC sequence
ATGTTC-CTC reference
```

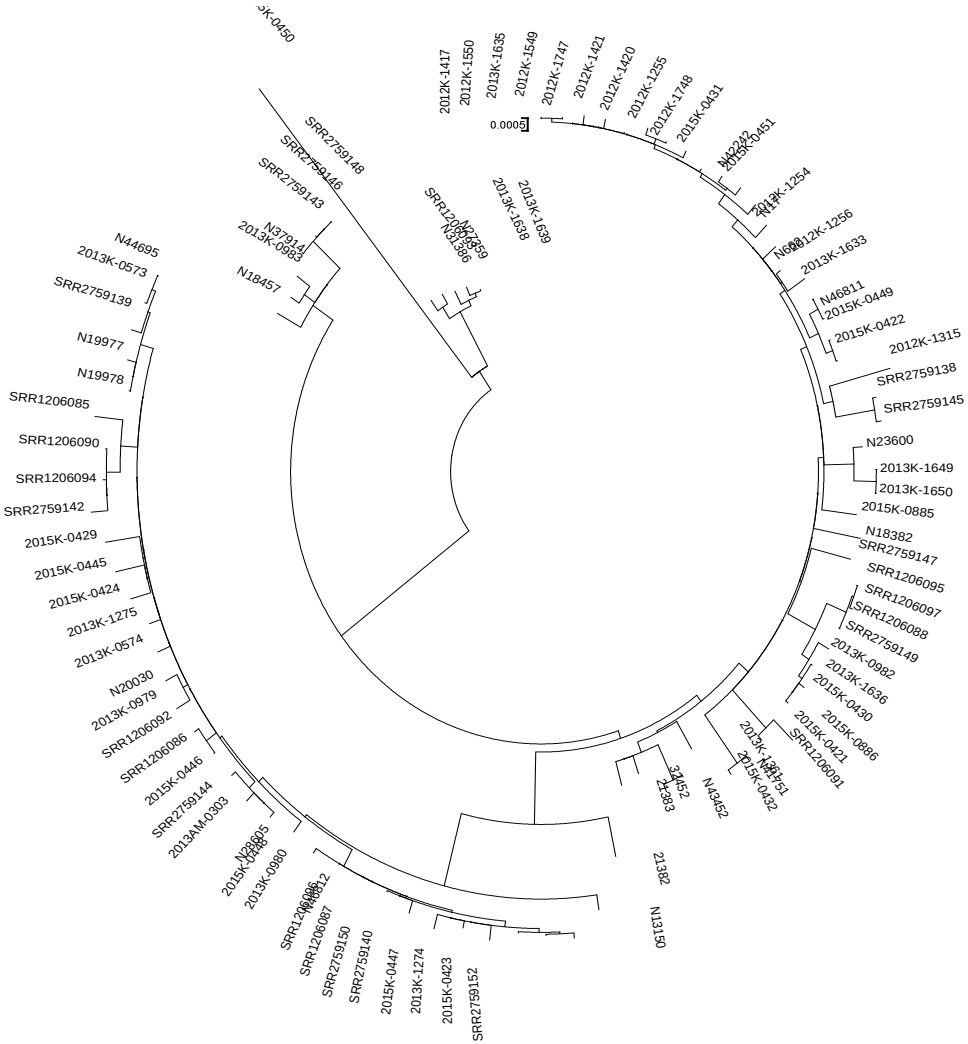
*differences not used in hqSNP analysis (SNV would includes indel)

Data analysis part 2: the hq SNP concept

```
Isolate 1 AACATGCAGACTGACGATTCGACGTAGGCTAGACGTTGACTG
Isolate 2 AACATGCAGACTGACGATTCGTCGTAGGCTAGACGTTGACTG
Isolate 3 AACATGCAGACTGACGATTCGACGTAGGCTAGACGTTGACTG
Isolate 4 AACATGCATACTGACGATTCGTCGAAGGCTAGACGTTGACTG
```



	1	2	3	4
1	0			
2	1	0		
3	0	1	0	
4	3	2	3	0


0.0005
2012K-1417
2012K-1550
2013K-1635
2012K-1549
2012K-1747
2012K-1421
2012K-1420
2012K-1255
2012K-1748
2015K-0431
2015K-0451
2012K-1254
2012K-1256
2013K-1633
N46811
2015K-0449
2015K-0422
2012K-1315
SRR2759138
SRR2759145
N23600
2013K-1649
2013K-1650
2015K-0885
N18382
SRR2759147
SRR1206095
SRR1206097
SRR1206088
SRR2759149
2013K-0982
2013K-1636
2015K-0430
2015K-0886
2015K-0421
SRR1206091
2015K-0432
N43452
21383
21382
N13150
2013K-1638
2013K-1639
SRR2759148
SRR2759146
SRR2759143
N37914
2013K-0983
N18457
N44695
2013K-0573
SRR2759139
N19977
N19978
SRR1206085
SRR1206090
SRR1206094
SRR2759142
2015K-0429
2015K-0445
2015K-0424
2013K-1275
2013K-0574
N20030
2013K-0979
SRR1206092
SRR1206086
2015K-0446
SRR2759144
2013AM-0303
2013K-0980
SRR1206087
SRR2759150
SRR2759140
2015K-0447
2013K-1274
2015K-0423
SRR2759152

Whole genome multilocus sequence typing (MLST)

- Database is built from genes found among diverse isolates of the species of interest
- Each unique gene is referred to as a "locus" – a locus may include the entire gene or a piece of the gene
- Any changes – SNP, insertions, deletions – equals a new allele call for a locus
- New alleles are named sequentially when encountered- not based on sequence

		2 SNPs	1 indel
Locus 1	ACTAGAGGGAAA	ACTAGAGGCTAA	ACT-GAGGGAAA
	allele 1	allele 2	allele 3

Whole genome multilocus sequence typing (MLST)

- Allows for simpler analysis and clear naming of subtypes
- Performs comparison on a gene by gene level

	Isolate A	Isolate B	Isolate C
Locus 1 (303 nt)	1	1	1
Locus 2 (102 nt)	1	1	2
Locus 3 (1,200nt)	1	1	1
Etc.			
Locus 2,005 (600nt)	1	1	2
MLST type	**A**	**A**	**B**

SNPs and allele differences

- *L. monocytogenes*: ~2.5×10^{-7} substitutions per site per year (approx. one substitution on the core genome every 2.5 years) (Moura et al., 2016)
  - cgMLST types diversify slowly (roughly 0.2 alleles per year)
- *Salmonella* Cerro: approx. 7×10^{-7} substitutions per site per year (Kovac et al., 2017)

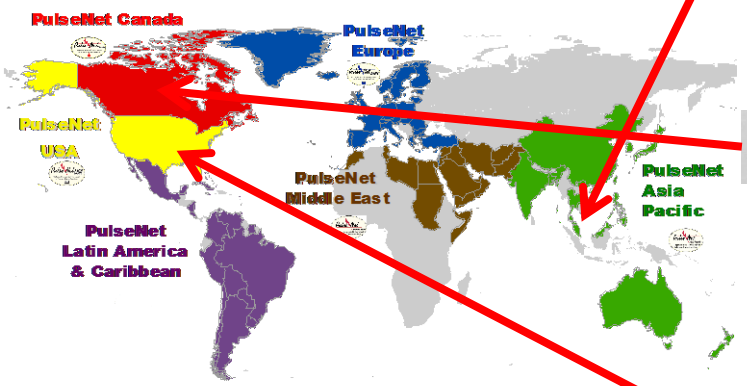
WGS and data analysis summary

- Data analysis involves many steps
- Considerable efforts and advancements in standardization and transparency
- Different analysis approaches exist, but
  - Once a genome has been sequenced data can be reanalyzed with different methods
  - If done correctly, different methods typically give comparable results (may not apply to all organisms)
- WGS analyses establish recent common ancestor, but do not establish (relevant) epidemiological link
  - Good epidemiological data and meta data are essential

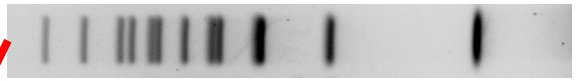
Outline

- Background: NGS, WGS, metagenomics, etc.
- **Use of WGS for Improved outbreak detection**
- Use of WGS and other subtyping tools for source tracking and to understand pathogen persistence
- Use of WGS to better define pathogens and food safety hazards*: Listeria monocytogenes* and *Salmonella* as case studies
- Metagenomics

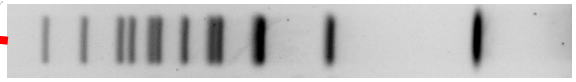
PulseNet allows international outbreak detection and traceback – a hypothetical example



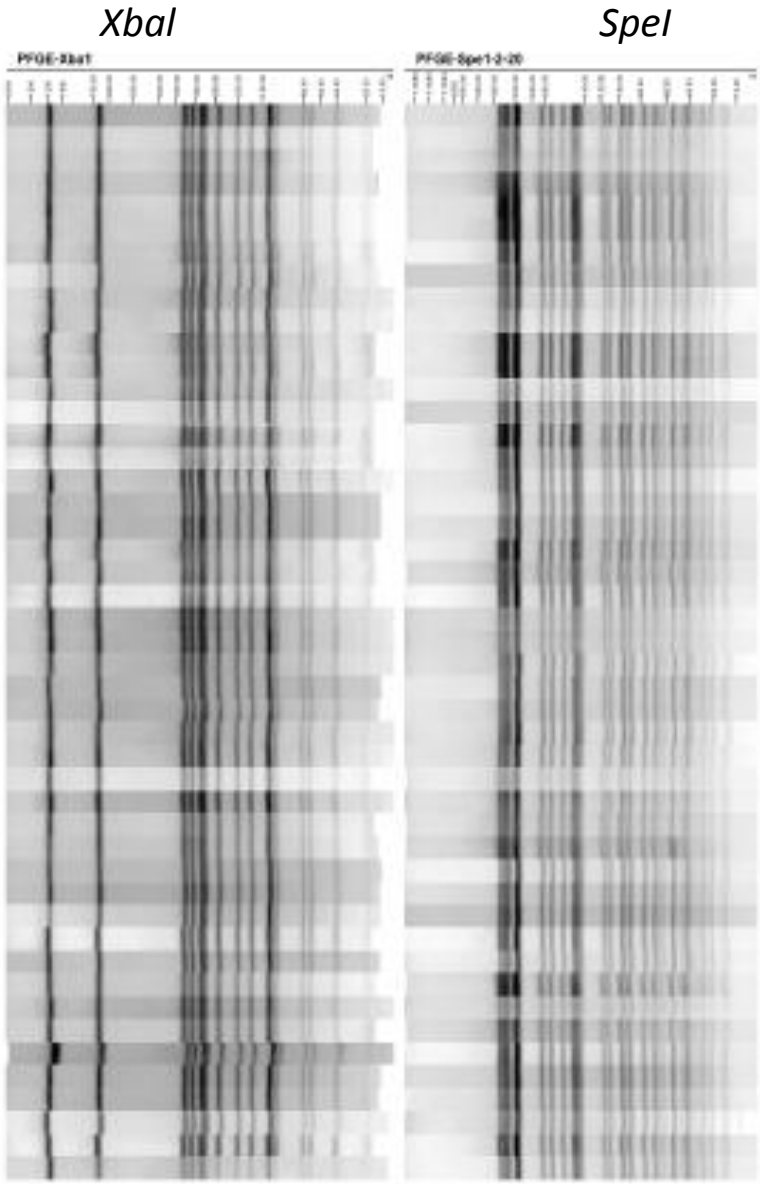
Food isolate, deposited into PulseNet



Human case



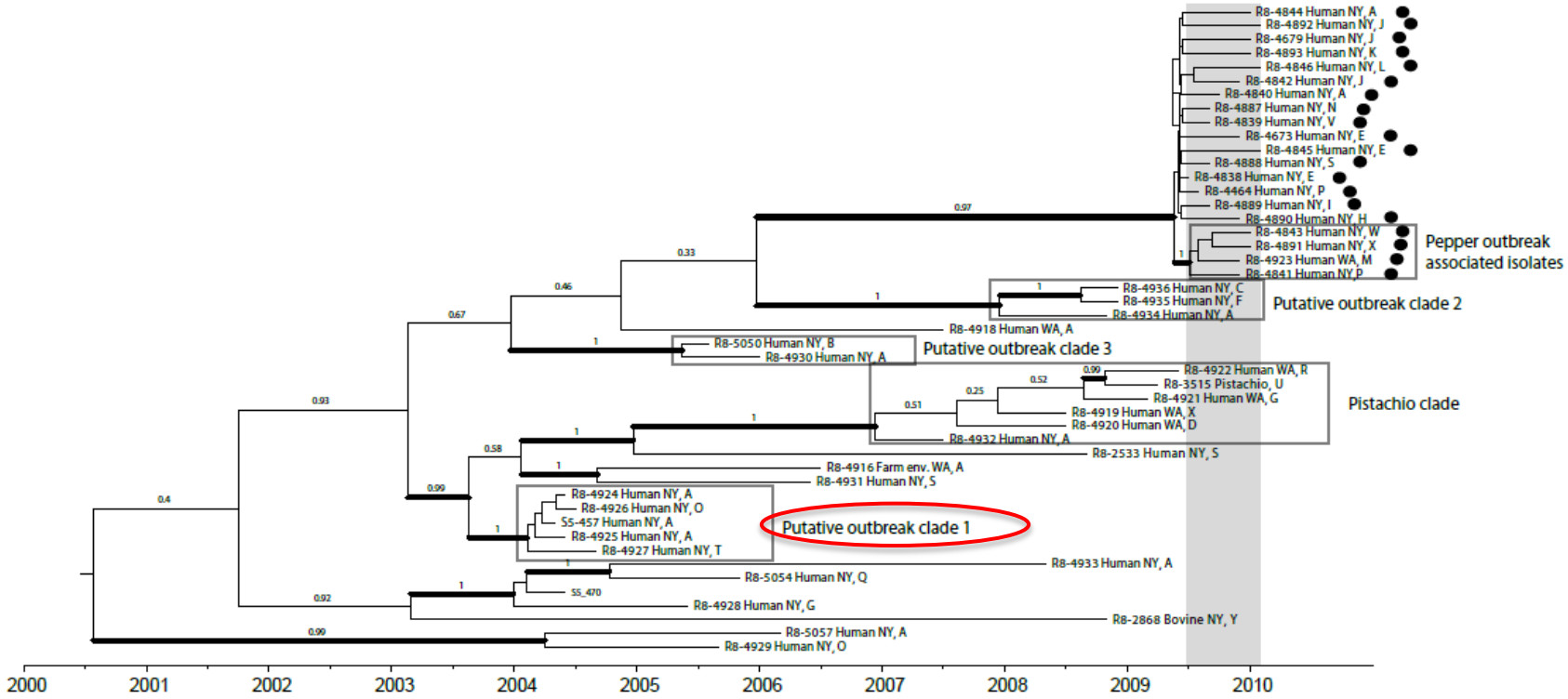
Human case



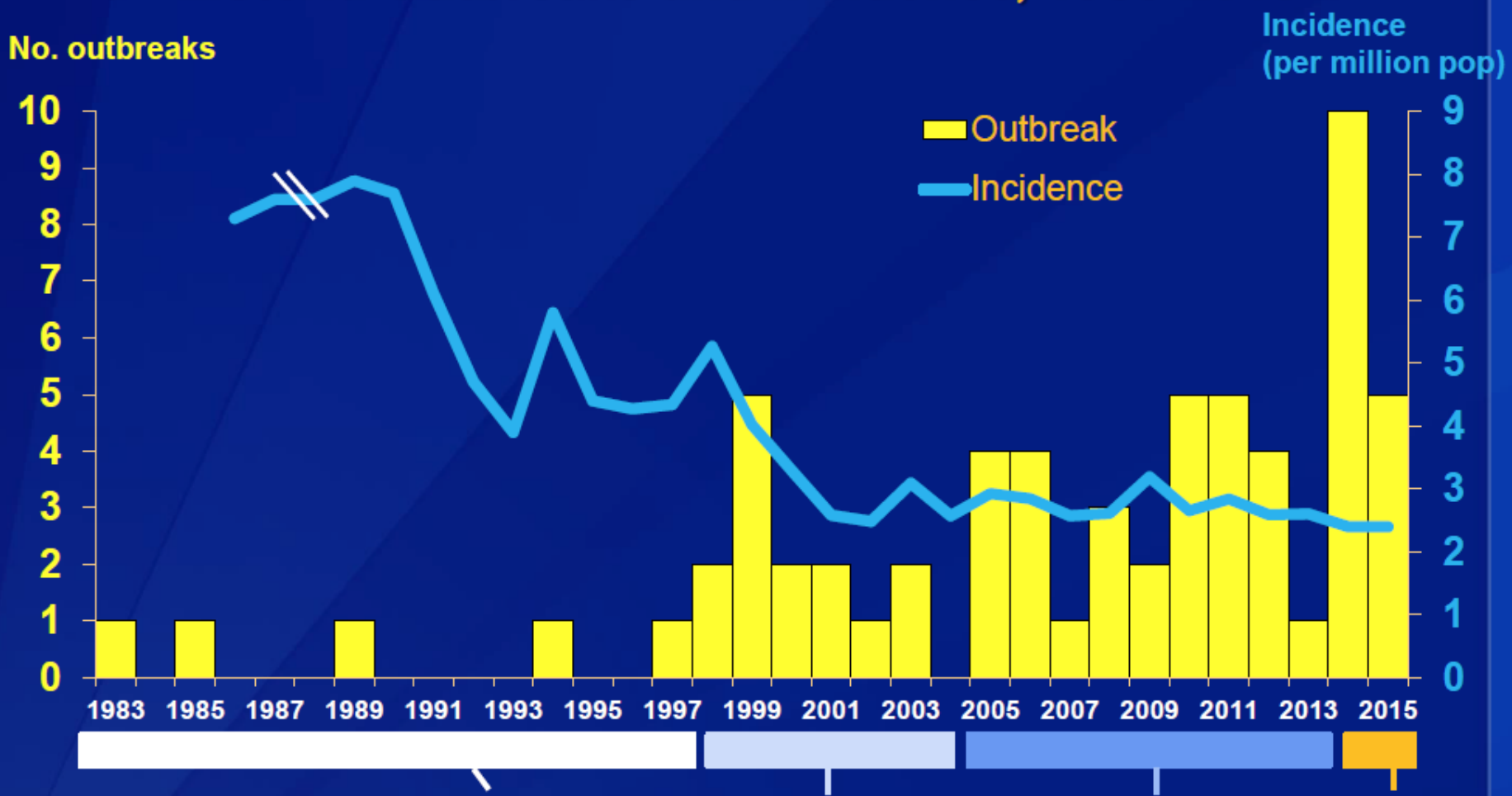
Includes isolates form Salmonella outbreak linked to sausages (Rhode Island) and isolates from pistachios

Den Bakker et al. 2011. AEM.

Tip-dated maximum clade credibility tree based on SNP data for 47 Montevideo isolates



Listeria Outbreaks and Incidence, 1983-2015



Era	Pre-PulseNet	Early PulseNet	*Listeria* Initiative	WGS
Outbreaks per year	0.3	2.3	2.6	7.5
Median cases per outbreak	69	11	5.5	3.5

Courtesy: Amanda Conrad, Outbreak Preparedness and Response Branch

Food Safety News
Breaking news for everyone's consumption

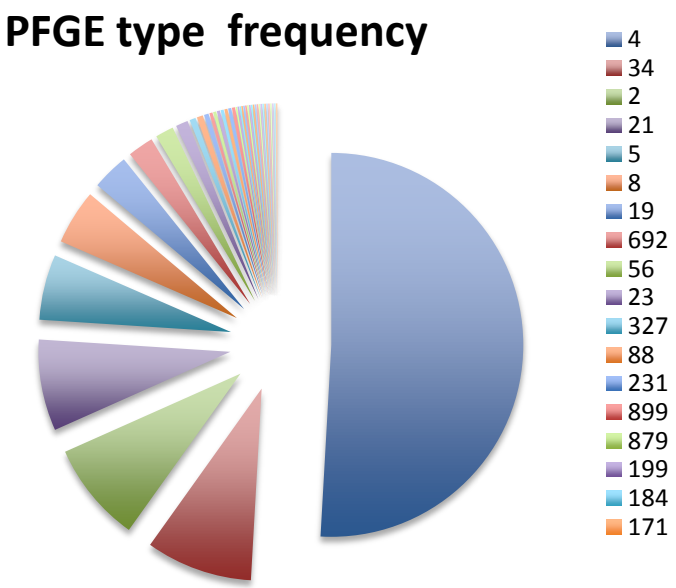
CDC/FDA Partnership Targets Whole Genome Sequencing of Listeria Monocytogenes

By Brian Sauders | November 27, 2013

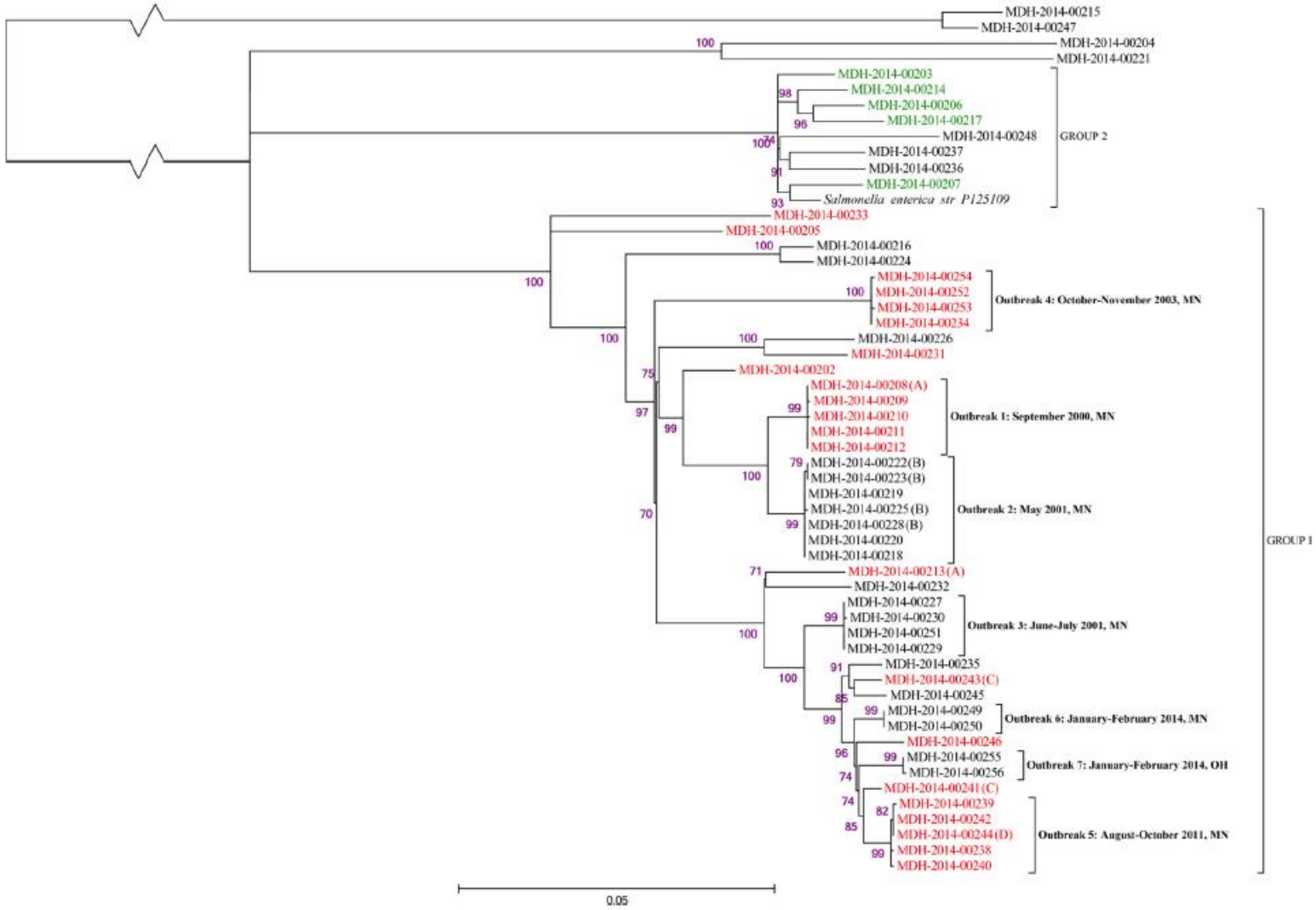
In a prior APHLTech blog post (NGS in Action: FDA's Genome TRAKR Network), Victor Waddell of the Arizona State Public Health Laboratory described the newly formed network of laboratories formed by the U.S. Food and Drug Administration (FDA). Known collectively as Genome TRAKR, the member laboratories perform whole genome sequencing (WGS) on bacterial foodborne pathogens isolated primarily from food and environmental sources.

On Sept. 1, 2013, the Centers for Disease Control and Prevention (CDC) began a partnership with the FDA Genome TRAKR network to utilize the network to conduct WGS of all Listeria monocytogenes collected from reported human illness cases in the United States. This effort leverages public health resources to evaluate and

Salmonella Enteritidis is most common cause of human salmonellosis in the US – poorly resolved by current subtyping technologies.



52 PFGE types



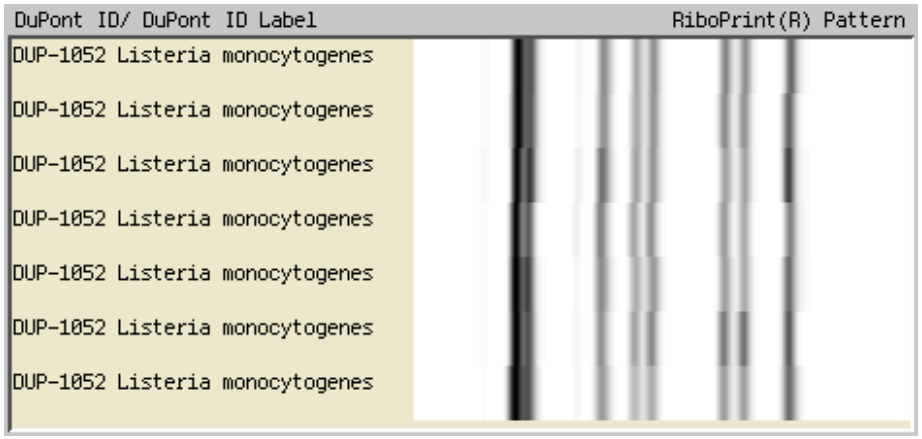
Taylor et al. JCM. 2015

Outline

- Background: NGS, WGS, metagenomics, etc.
- Use of WGS for Improved outbreak detection
- **Use of WGS and other subtyping tools for source tracking and to understand pathogen persistence**
- Use of WGS to better define pathogens and food safety hazards*: Listeria monocytogenes* and *Salmonella* as case studies
- Metagenomics

2000 US outbreak - Environmental persistence of *L. monocytogenes*

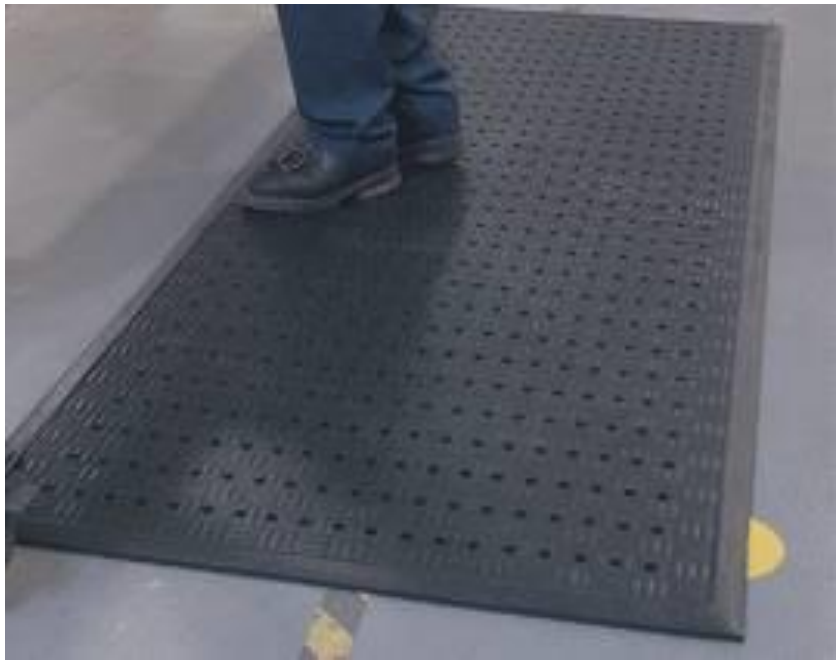
- 1988: one human listeriosis case linked to hot dogs produced by plant X
- 2000: 29 human listeriosis cases linked to sliced turkey meats from plant X



1988 and 2000 isolates differed by as little as 1 SNP (Orsi et al., 2008)

Plant A2	2/28/01	3/26/01	4/24/01	5/22/01	6/19/01	7/17/01	8/14/01	9/18/01	10/9/01	11/6/01	12/12/01	1/29/02	2/25/02	3/5/02	4/2/02	4/16/02	5/14/02	6/10/02	7/1/02	7/23/02	8/20/02	9/17/02	10/15/02	11/12/02	12/10/02
														1-L.spp			2-L.spp			1-L.spp					1-L.spp
Raw Product Samples	1062D	1060A	-	-	-	-	-	-	-	L.spp	-	L.spp		1025A	-		1053C	-	-	1039C	-	-	-	-	1039C
	1 of 6	1 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	3 of 6	6 of 6	1 of 6		1 of 6	6 of 6		1 of 6	6 of 6	6 of 6	1 of 6	6 of 6	6 of 6	6 of 6	6 of 6	1 of 6
Raw/In-Process Areas																									
E3: Floor drain, raw salmon ro	1053A	-	-	-	-	-	-	-	-	L.spp	L.spp	L.spp		1053A	-		-	-	-	-	-	-	-	-	-
Salmon receiving floor drain				-	-	L.spp	1053A	-	-	L.spp	-	-		L. spp	-		-	-	-	L.spp	L.spp	L.spp	-	-	-
Raw salmon room, Drain (SB-FD1)														1053A	1053A		-	-	1053a	1053A	-	1053A	1062A	L.spp	1053A
Raw salmon room, Drain (SB-FD2)														1053A	1053A		1053A	-	L.spp	1053A	1053A	L.spp			-
Raw salmon room, 3 floor mats														1053A	1053A			1053A	1053A	1053A	1053A	1053A	1053A	1053A	1053A
RawSalmon room, mats- post cleaning																1053A	1053A								
Raw salmon room, platic pallet														1053A	-			-	L.spp	-	-	-	-	-	-
Raw Salmon room, pallet, post cleaning																-	-								
Raw salmon room, pallet jacket handle														-	-		-	-	-	-	-	-	-	-	-
E8: Apron, employee in raw are	1062D	-	-	1053A	-	1053A	1025A	-	1053A	-	1053A	-		-	-		-	-	-	-	-	-	-	-	-
Incoming raw material packaging												-		-	-		-	-	-	-	-	L.spp	-	L.spp	-
Finished Product Areas																									
E1: Trench Drain, processing r	L.spp	-	-	116-693	L.spp	L.spp	L.spp	-	L.spp	L.spp	-	L.spp		1042C	L.spp		1042C	L.spp	-	-	L.spp	L.spp	L.spp	L.spp	L. spp
E2: Trench Drain, smoke room	-	-	-	-	-	-	-	-	-	-	-	-		-	-		-	-	L.spp	-	-	L.spp	-	-	L. spp
Smoke room trench drain, in use																1053A	-	-	-	-	-	-	-	-	-
E4: Cart wheels, for box transp	L.spp	-	-	-	-	-	-	L.spp	-	L.spp	-	-		-	-		-	-	-	-	-	L.spp	L.spp	-	L. spp
E5: Floor, under conveyor belt	L.spp	-	-	-	-	L.spp	L.spp	L.spp	L.spp	L.spp	L.spp	-		-	-		-	-	-	-	-	-	1053	1053	-
Finish Room, floor mats #1													L.spp												
Finish room, floor mats #2													L.spp												
Finish room, floor mats, reg. Clean														L.spp	L.spp	L.spp	L.spp	L.spp	L.spp	L.spp	L.spp	1042B	L.spp	L.spp	L. spp
Finish room, floor mats, reg. Clean													L.spp												
Finish room, 1200 ppm Quat, weekend													L.spp												
Finish room, 1200 ppm Quat, weekend													-												
Bootdip valve cover, processin	-	-	L.spp																						
E6: Platform under Geba #1 sli	-	-	-	-	-	-	L.spp	L.spp	L.spp	-	-	-		-	-		-	-	-	-	-	-	-	-	-
E9: Sliding door handle, skinni	L.spp	-	-	-	-	-	-	1053A	L.spp	-	-	-		-	-		-	-	-	L.spp	-	-	-	-	-
Food Contact Surfaces																									
E7: Gloved hands, finish prod.	-	-	-	-	-	-	-	-	-	-	-	-		-	-		-	-	-	-	-	-	1053	-	L. spp
E10: Skinning machine	L.spp	-	L.spp	-	-	-	L.spp	L.spp	L.spp	-	L.spp	-		L.spp	-		-	-	-	-	-	L.spp	-	L. spp	-
E11: Geba #5 slicer	L.spp	-	-	-	-	-	-	-	-	-	L.spp	-		-	-		-	-	-	-	-	-	1053	-	-
E12: 20/20 vac belt	-	-	-	L.spp	-	-	-	-	-	-	-	-		-	-		-	-	-	-	-	-	-	-	-
	6 of 6	6 of 6	6of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6		6 of 6	6 of 6		6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6	6 of 6
Finished Product Sample	-	-	-	-	-	-	-	-	-	-	-	-		-	-		-	-	-	-	-	-	-	1039A	-

L. monocytogenes persisted in rubber floor mats despite sanitation



Listeria can be protected from sanitizer in “micro-cracks”, but can be squeezed out by pressure if people stand on mats

Implications: a hypothetical case study

- January 2016: sliced turkey from plant A found positive for *L. monocytogenes* when a sample collected at a supermarket was tested

Implications: a hypothetical case study

- January 2016: sliced turkey from plant A found positive for *L. monocytogenes* when a sample collected at a supermarket was tested
- December 2016: ham from plant A found positive for *L. monocytogenes* when a sample was tested

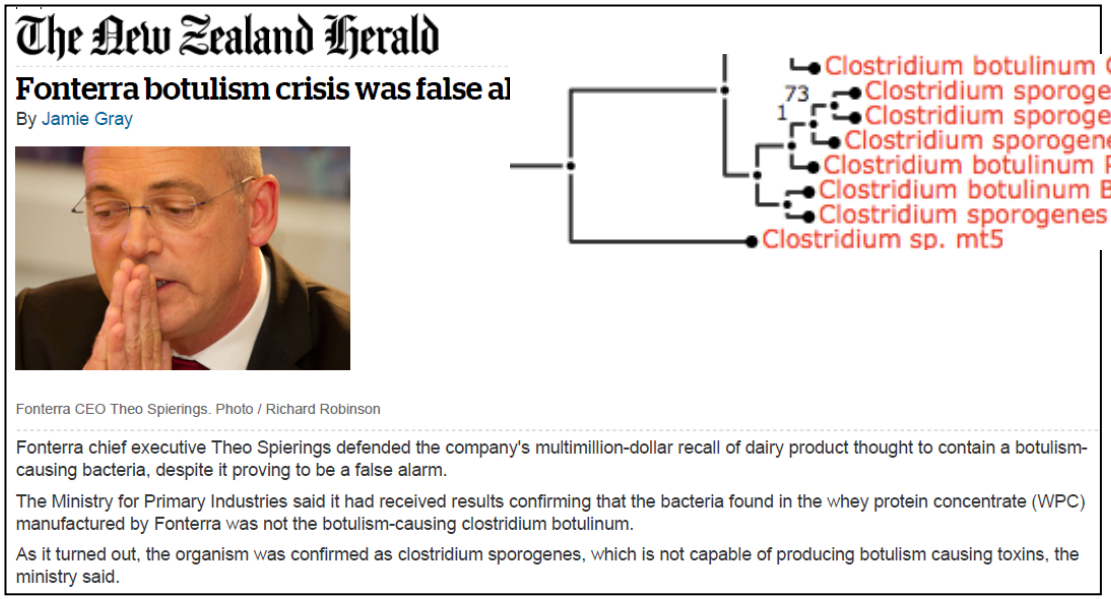
Implications: a hypothetical case study

- January 2016: sliced turkey from plant A found positive for *L. monocytogenes* when a sample collected at a supermarket was tested
- December 2016: ham from plant A found positive for *L. monocytogenes* when a sample was tested
- Isolates from January and December match by WGS

Outline

- Background: NGS, WGS, metagenomics, etc.
- Use of WGS for Improved outbreak detection
- Use of WGS and other subtyping tools for source tracking and to understand pathogen persistence
- **Use of WGS to better define pathogens and food safety hazards*: Listeria monocytogenes* and *Salmonella* as case studies**
- Metagenomics

Consequences of unreliable differentiation between pathogens and non-pathogens can be costly



The New Zealand Herald

Fonterra botulism crisis was false al

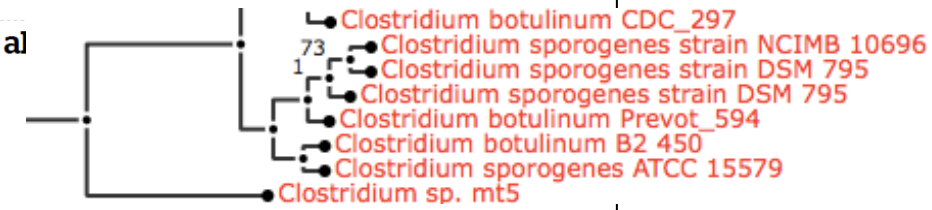
By Jamie Gray

Fonterra CEO Theo Spierings. Photo / Richard Robinson

Fonterra chief executive Theo Spierings defended the company's multimillion-dollar recall of dairy product thought to contain a botulism-causing bacteria, despite it proving to be a false alarm.

The Ministry for Primary Industries said it had received results confirming that the bacteria found in the whey protein concentrate (WPC) manufactured by Fonterra was not the botulism-causing clostridium botulinum.

As it turned out, the organism was confirmed as clostridium sporogenes, which is not capable of producing botulism causing toxins, the ministry said.



- Suspected botulism-causing bacteria identified in whey products
- ~1,000 tones of products recalled in 7 countries
- No disease cases
- Detailed strain characterization confirmed **species misclassification**

Cornell University
College of Agriculture and Life Sciences

INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

TAXONOMIC DESCRIPTION

Dobritsa *et al.*, *Int J Syst Evol Microbiol* 2017;67:2317–2322
DOI 10.1099/ijsem.0.001948

MICROBIOLOGY SOCIETY

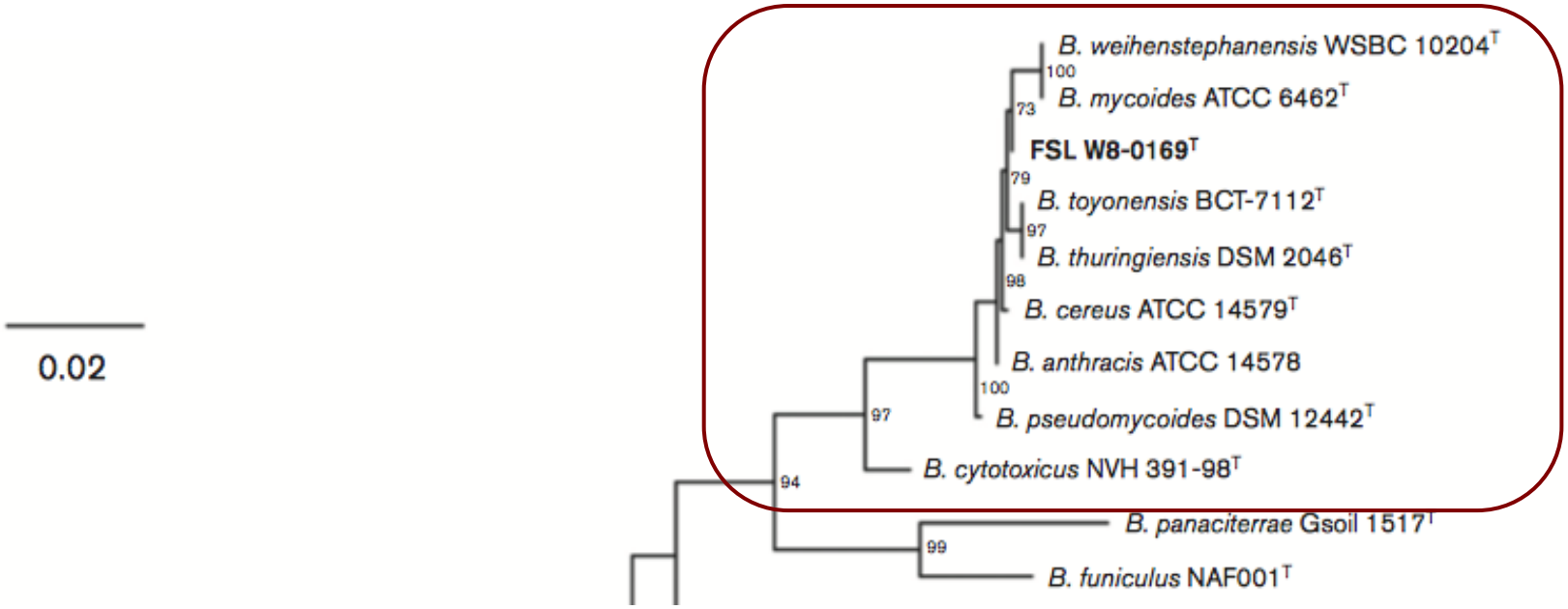
Clostridium tepidum sp. nov., a close relative of *Clostridium sporogenes* and *Clostridium botulinum* Group I

Anatoly P. Dobritsa,[1,*] Kirthi K. Kutumbaka,[1] Kirsten Werner,[1] Martin Wiedmann,[2] Aaron Asmus[3] and Mansour Samadpour[1]

Abstract

Obligately anaerobic, Gram-stain-positive, spore-forming bacteria indistinguishable by pulsed-field gel electrophoresis were isolated from non-dairy protein shakes in bloated bottles. One of the isolates, strain IEH 97212^T, was selected for further study. The strain was closely related to *Clostridium sporogenes* and *Clostridium botulinum* Group 1 based on 16S rRNA gene sequence similarities. Phylogenetic analysis also showed that strain IEH 97212^T and strain PE (=DSM 18688),

16S rRNA-based differentiation among *B. cereus* group species



Miller et al., 2016

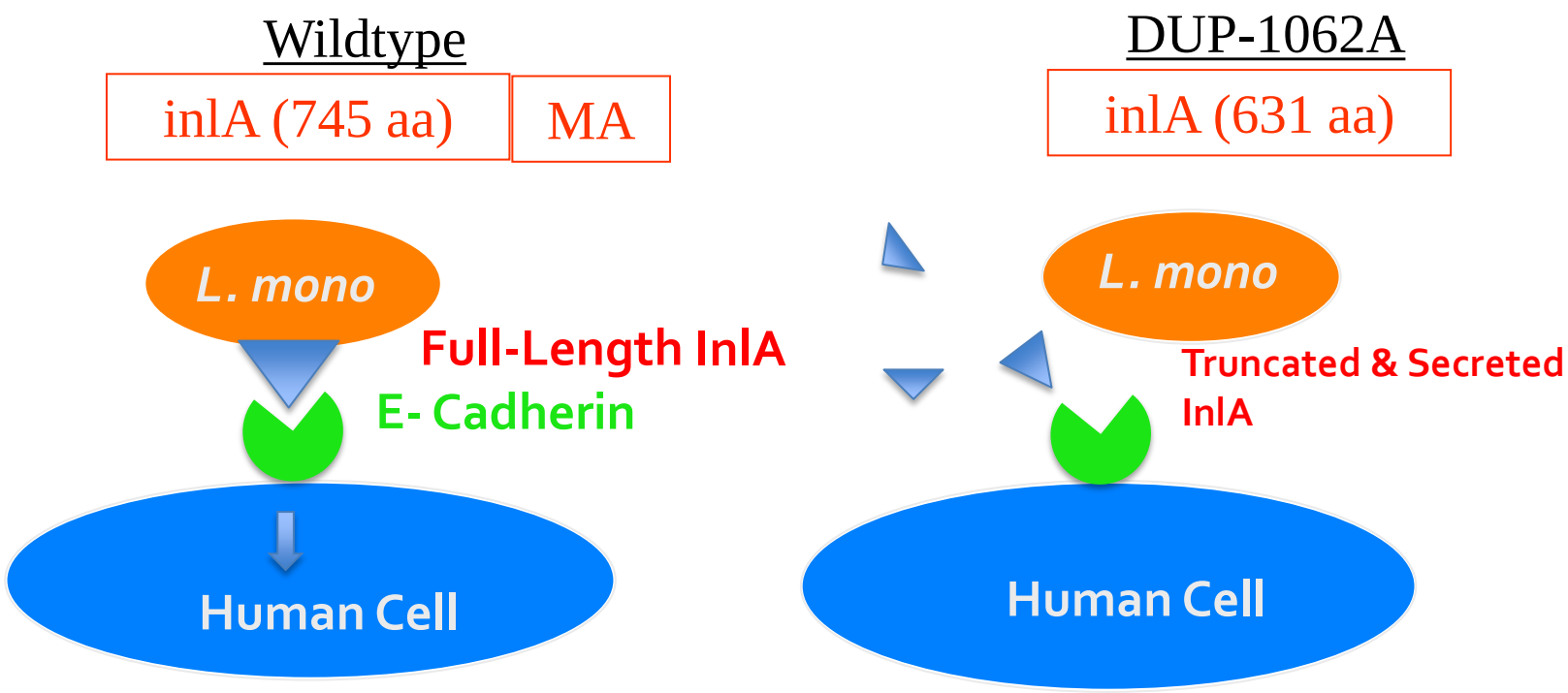
Molecular characterization of *Listeria monocytogenes* human, and food isolates

Ribotype	Number of isolates		P-value [1)]	Comments
	Food	Human		
DUP-1030A	8	8	NS	
DUP-1030B	0	10	**	not found in food
DUP-1038B	15	63	****	
DUP-1039A	12	31	**	
DUP-1039B	18	43	**	
DUP-1039C	35	25	NS	
DUP-1042A	11	16	NS	
DUP-1042B	18	72	****	
DUP-1042C	14	0	***	multiple food types, not in humans
DUP-1043A	30	16	*	
DUP-1044A	11	28	**	
DUP-1044B	1	19	***	rarely found in food
DUP-1044E	10	0	**	blue cheese only
DUP-1045B	14	11	NS	
DUP-1052A	58	39	*	
DUP-1053A	24	41	*	
DUP-1062A	151	9	****	rarely found in humans
DUP-1062D	28	1	****	rarely found in humans
rare	22	42	*	Ribotypes with 1-4 isolates
uncommon	22	33	NS	Ribotypes with 5-8 isolates
Total	502	507	****	Overall analysis of ribotype vs. origin

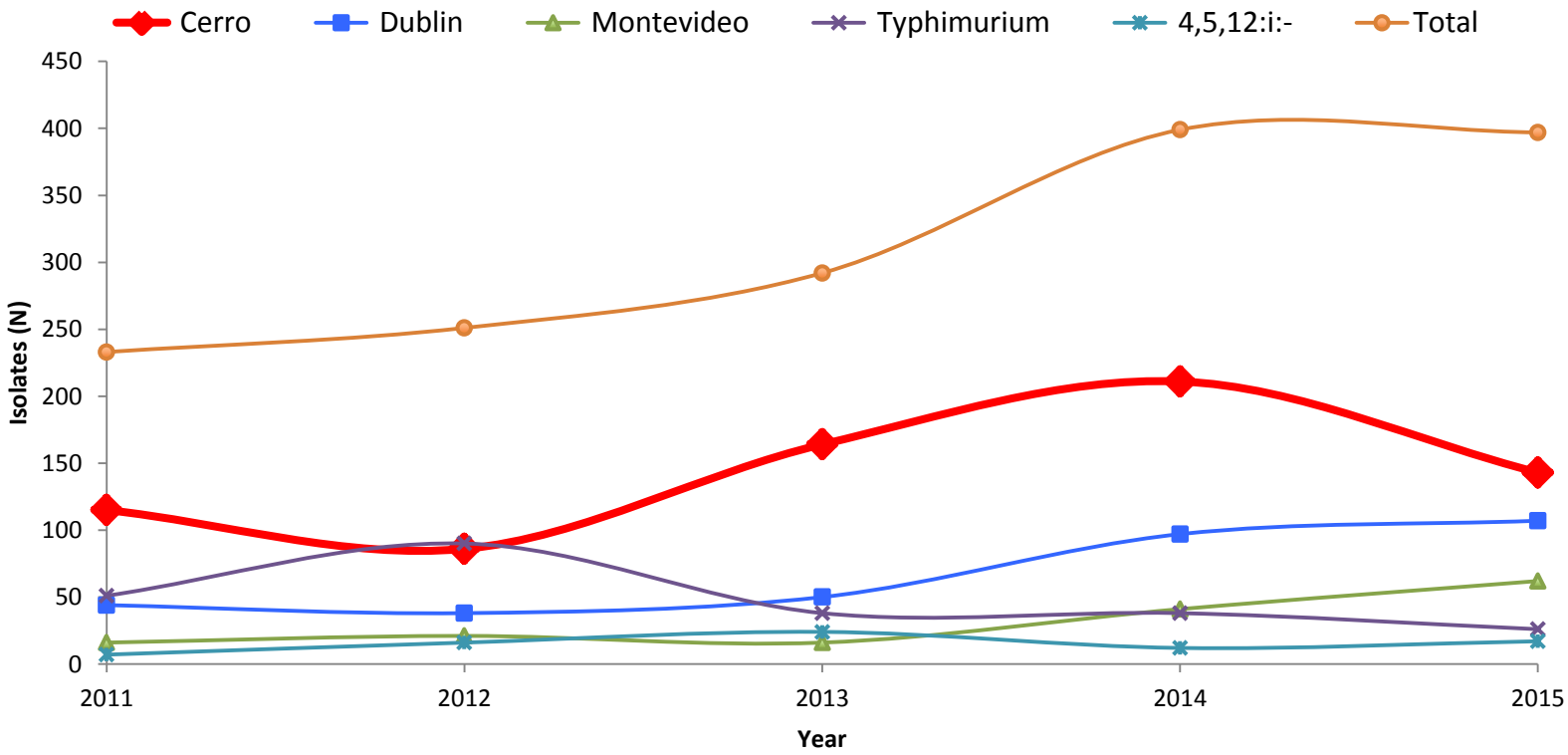
[1)] *P*-values refer to comparison of origin between ribotype specified in that row vs. all other ribotypes where NS = not significant, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$

Virulence attenuation of ribotype DUP-1062A

Isolates with ribotype DUP-1062A carry a premature stop codon in *inlA*, which leads to reduced invasion of human epithelial cells



Salmonella Cerro – an emerging cattle pathogen (NY data)

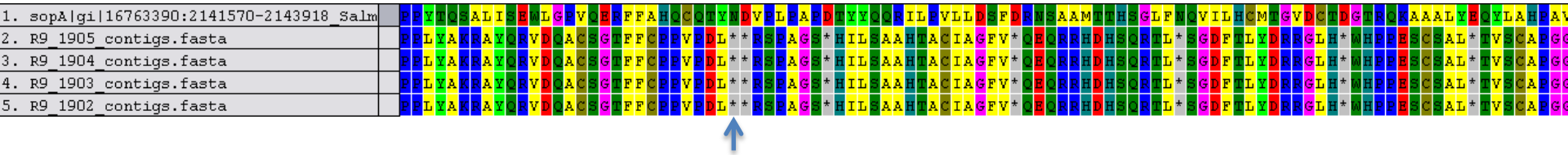


Kovac et al. unpublished.

S. Cerro displays mutations that may reduce human virulence

Non-synonymous mutation in *sopA*, which results in premature stop codon on position 434/783

- All analyzed *S.* Cerro isolates carried this mutation



sopA encodes an effector protein

- A E3 ubiquitin ligase that interferes with host signaling pathway
- Alters host cell physiology and promotes bacterial survival in host tissues

- Additional mutations in D-alanine transporter, which is Involved in intracellular survival in murine macrophages

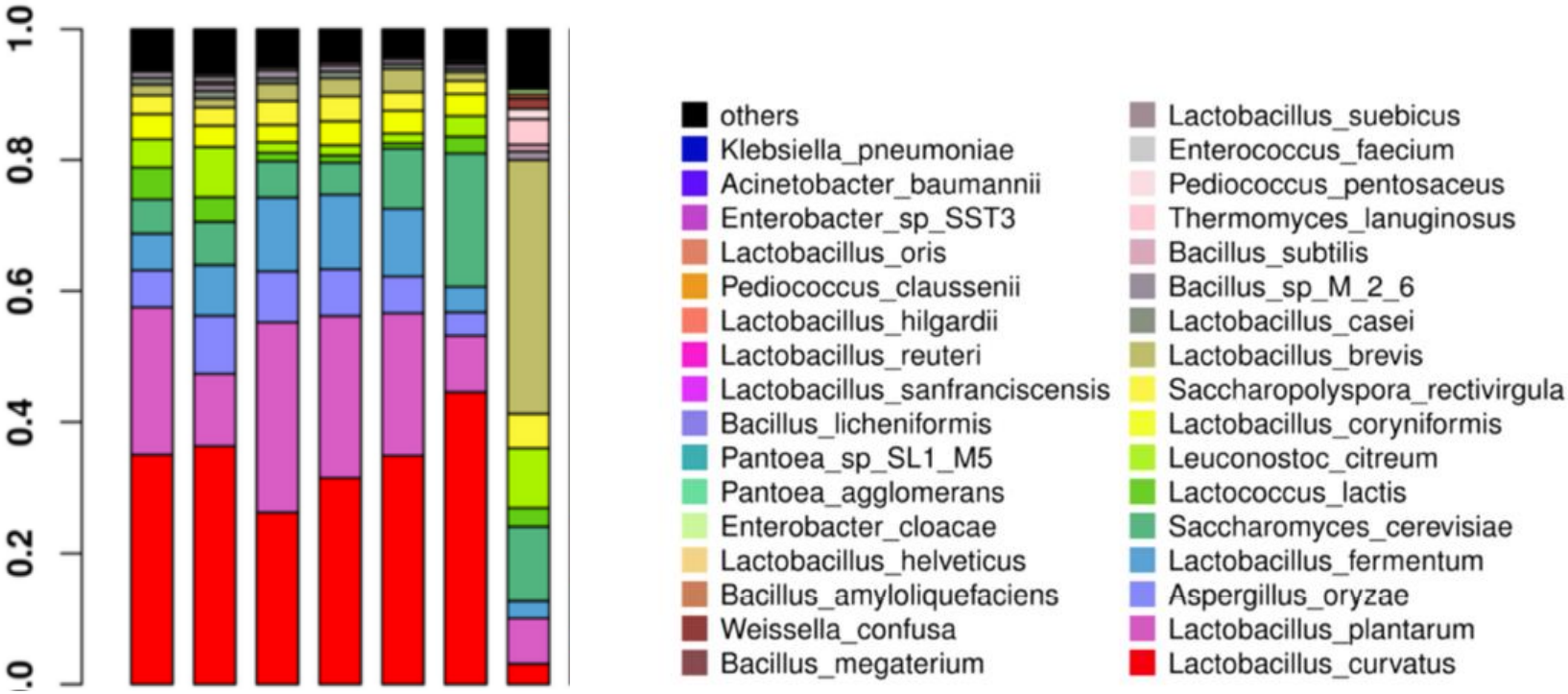
Outline

- Background: NGS, WGS, metagenomics, etc.
- Use of WGS for Improved outbreak detection
- Use of WGS and other subtyping tools for source tracking and to understand pathogen persistence
- Use of WGS to better define pathogens and food safety hazards*: Listeria monocytogenes* and *Salmonella* as case studies
- **Metagenomics**

Could this be the future?

- Auditor conducts audit of supplier facility X; collects 5 samples of ingredient (e.g., pepper), which are subsequently characterized by metagenomic analysis
- Subsequently, metagenomic analysis is conducted, at a risk-based frequency, of lots received at customer
- Lots that show "substantial" deviation will trigger further analyses (and possibly other actions?)
  - Statistical approaches to reliably identify meaningful deviations as main challenge

A hypothetical example



4 samples taken during plant audit

Figure taken from Hong et al., 2016. Nature Scientific reports

Challenges

- Bacteria with very few/no SNP differences can be found in different locations, foods, and food associated environments
  - WGS will rarely give a final answer
  - Epidemiology is more important than ever
- Metagenomics detects both live and dead cells
  - Is presence of certain genes a public health hazard (business risk) even if we don't know if they are found in a living cell
- Uncertainty around data interpretation, different data analyses approaches, and data use may affect industry willingness to use technologies
  - Can we have a "safe harbor" for industry use of WGS and metagenomics tools

Conclusions

- Precision food safety is here
- Improved outbreak detection, surveillance, source tracking and bacterial identification due to WGS use is the new reality
- Roadmap for other uses of WGS and NGS is less clear
  - Will metagenomics and WGS replace “hygiene indicators” and will these tools be used for process control
  - Will we change the approach to defining hazards and move away form bacterial species and serotypes to genomic –based hazard characterization
- Training is essential to facilitate efficient and appropriate use of WGs, NGS, and other “precision food safety tools”