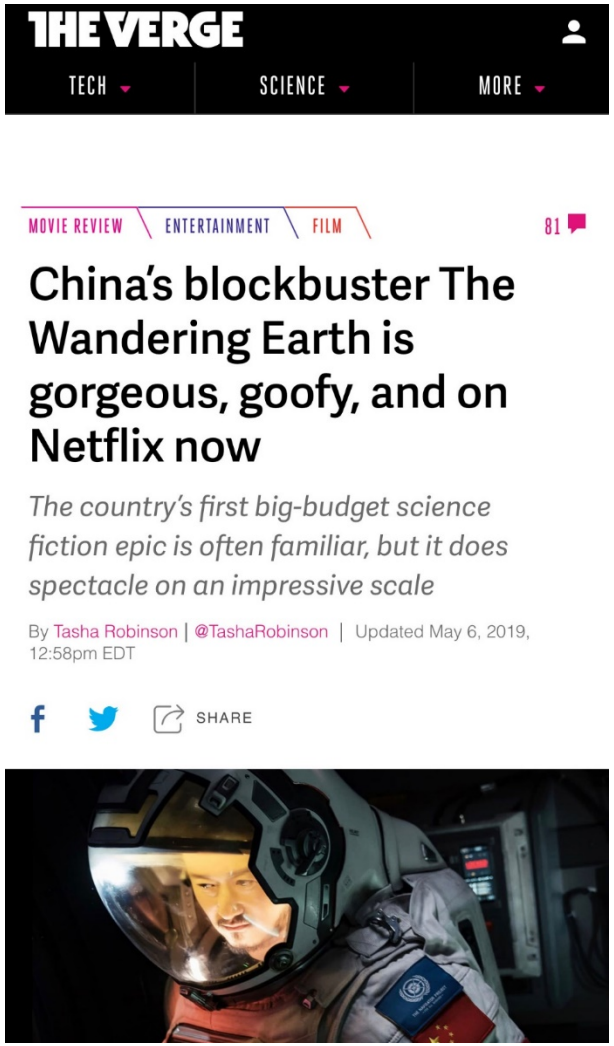
Zoonotic source attribution of *Salmonella* using WGS and machine learning

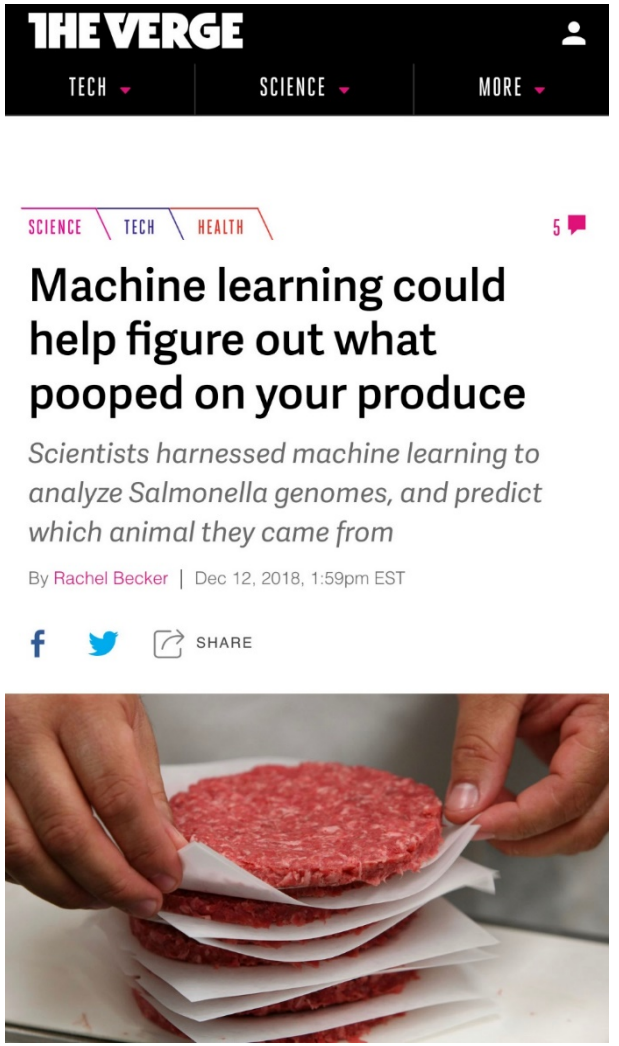
Xiangyu Deng

Associate Professor, Center for Food Safety

UNIVERSITY OF GEORGIA



Science fiction

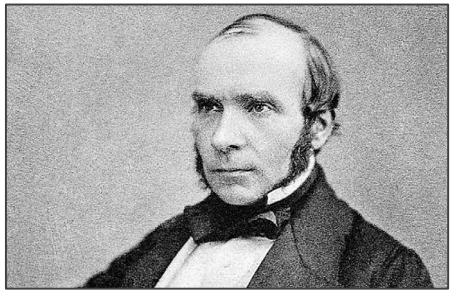


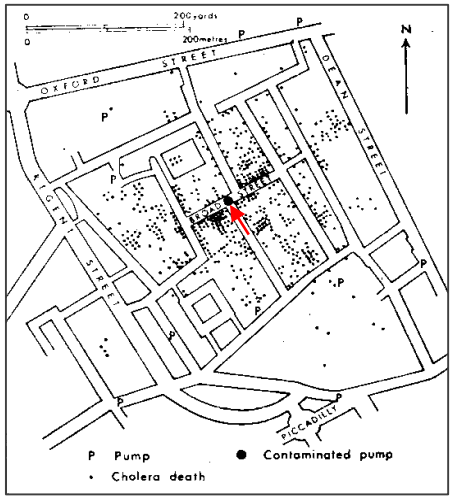
Science fiction?

Can we sequence a genome and predict where the isolate came from?

What common foodborne pathogens and their sources are amenable to genomic source attribution?

Shoe-leather epidemiology: field investigation outside the lab





John Snow, cholera, and the Broad Street pump

Hypothesis Testing Questionnaire Example

1. Survey Date: ___/___/___
2. Case Number: _____
3. Which of the following bakery products did you eat during April 1 – April 14? (*please check all that apply*)

- ❑ Any glazed product
- ❑ Glazed doughnut
- ❑ Glazed sweet roll
- ❑ Plain doughnut
- ❑ Bread
- ❑ Cake
- ❑ Cookies
- ❑ Pie
- ❑ Plain sweet roll
- ❑ Pastry

CDC Epidemiology Program Office, Outbreak of Jaundice in a Rural County

REVIEW

Analysis of consumer food purchase data used for outbreak investigations, a review

Frederik T Møller[1,2], Kåre Mølbak[1,3], Steen Ethelberg[1]
1. Department of Infectious Disease Epidemiology and Prevention, Statens Serum Institut, Copenhagen, Denmark
2. Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark
3. Institute of Veterinary and Animal Sciences, University of Copenhagen, Denmark

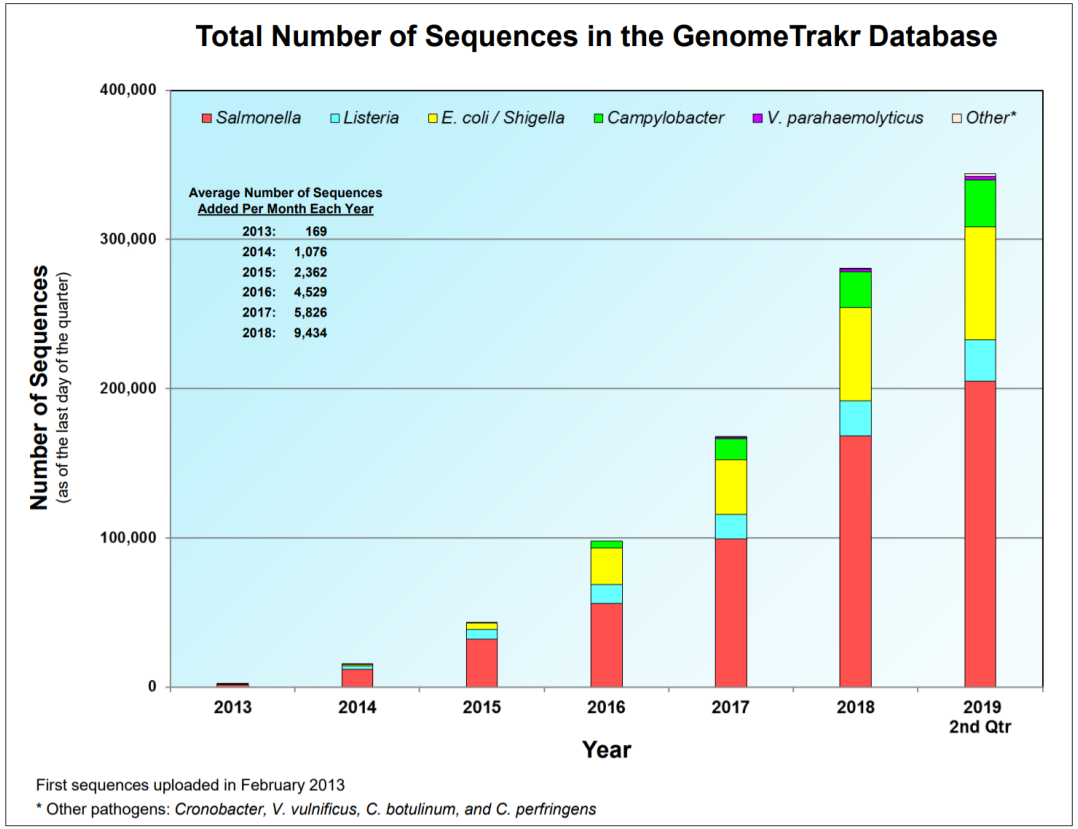
Correspondence: Steen Ethelberg (set@ssi.dk)

Credit cards, store loyalty card, shopping receipts etc.

- **9.4 million cases of foodborne illness per year in the US**
- **~95% are sporadic and non-outbreak cases that have no information about food exposure and contamination source**

- **Little knowledge in pathogen sources *v.s.* large volumes of pathogen WGS**

- **WGS: data mining outpaced by data accumulation**
  - Majority of genomes remain underutilized or untapped
  - Conventional methods for genome analysis are falling behind
  - Can machine learning help?



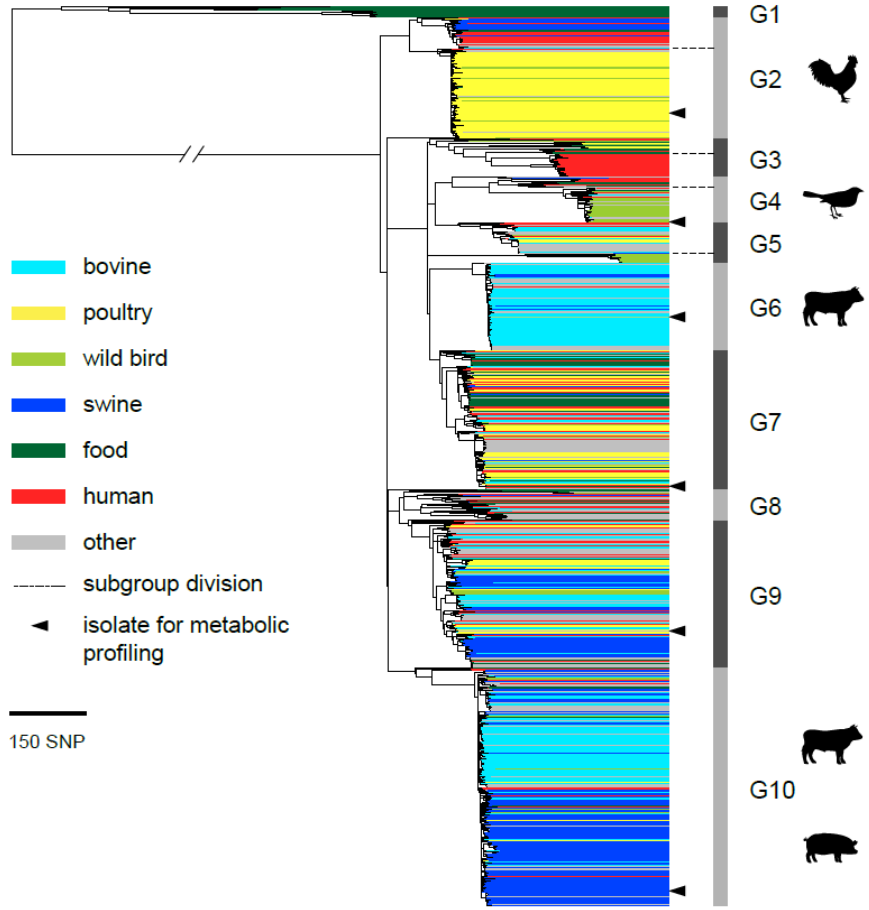
- ***S.* Typhimurium (ST) as a model organism for genomic source attribution**
  - One of the most prevalent serotypes
  - Broad livestock host range and varying degrees of host adaptation

- **Hypothesis: WGS informs zoonotic sources of ST**
  - Certain genetic features are more informative than others and sufficient for source prediction
  - Machine learning can find them

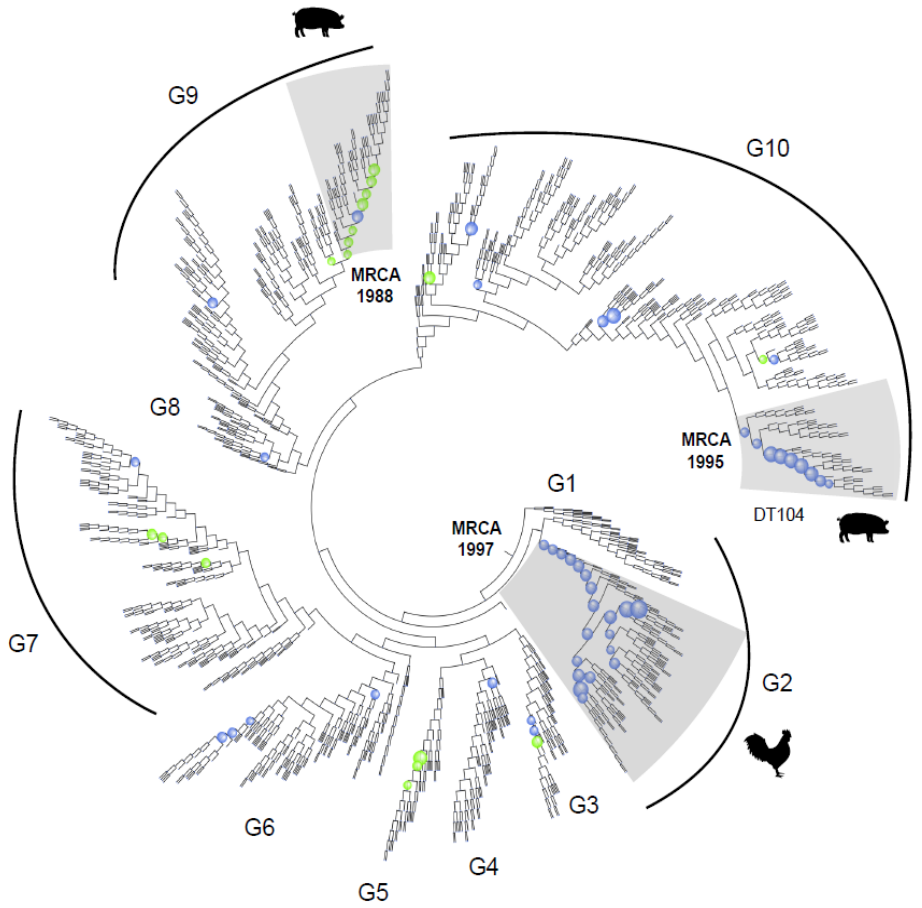
- **More than 2,000 ST genomes**
  - CDC surveillance, National Antimicrobial Resistance Monitoring System(NARMS), and FDA GenomeTrakr

- **Comprehensive interrogation of genomic features**
  - 3137 SNPs, indels, and accessory genes

- **Machine learning using Random Forrest**
  - A benchmark ML algorithm



- **Clustering of isolates from the same source**



- **Recently established livestock sources of ST**

- **Output source probabilities for each query genome:**
  80% poultry
  10% pigs
  8% cattle
  2% wild birds

- **Overall zoonotic source prediction accuracy 83%**

- **Successful retrospective source attribution of 7 of 8 US outbreaks of livestock origins (1998-2013)**

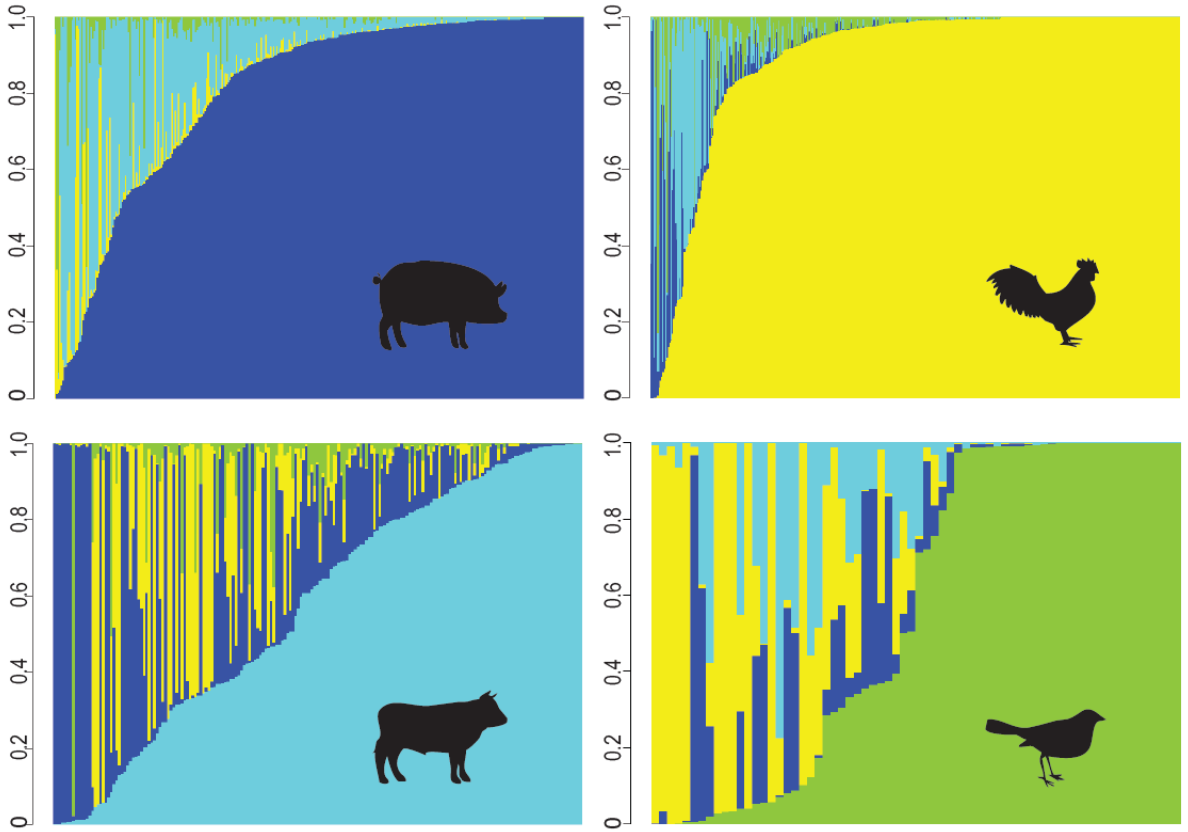
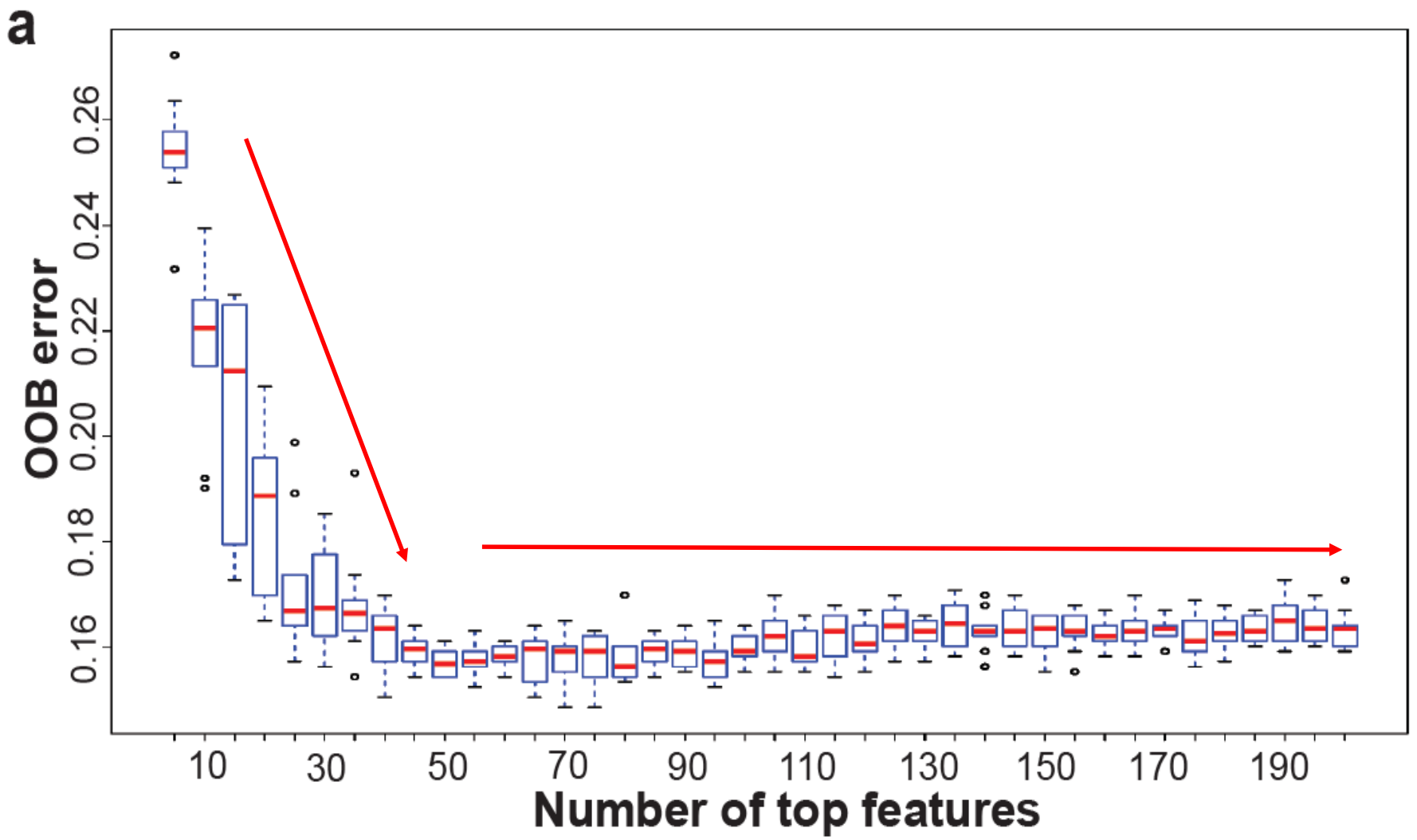


Table 1. Retrospective source attribution of zoonotic outbreaks of *Salmonella enterica* serotype Typhimurium, United States*

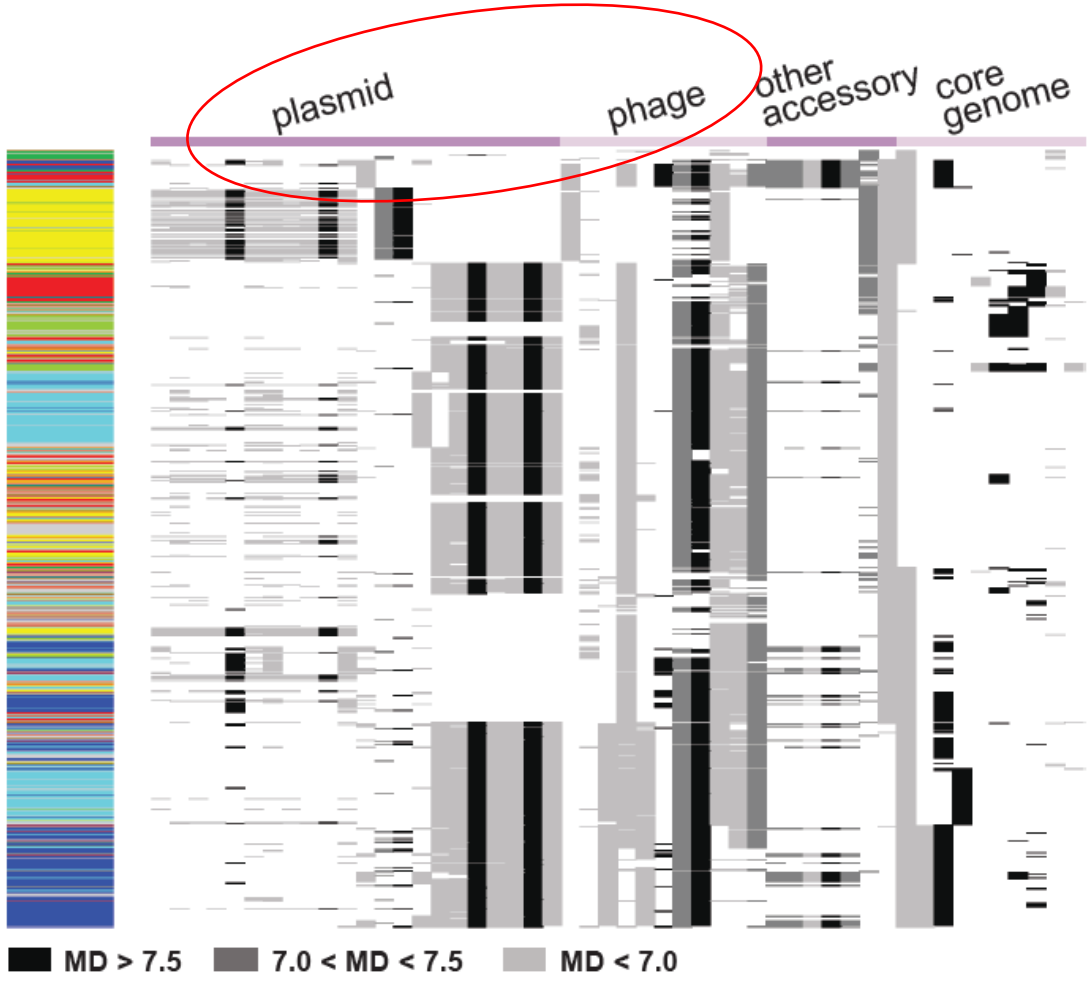
	Outbreak		Phylogenetic reference†			Population	Phylogeny	RF
Isolate	Year	Confirmed vehicle	Isolate	Year	Source	group	prediction	prediction
STM2207	2013	Ground beef	STM296	2006	Bovine	G9	+	+
STM2208	2013	Ground beef	STM296	2006	Bovine	G9	+	+
STM2209	2007	Pot pie turkey	STM093	2005	Poultry	G7	+	+
STM2210	2007	Pot pie turkey	STM093	2005	Poultry	G7	+	+
STM2211	2007	Pot pie turkey	STM093	2005	Poultry	G7	+	+
STM2212	2007	Pot pie turkey	STM093	2005	Poultry	G7	+	+
STM2213	2013	Live poultry	STM2114	2016	Bovine	G7	–	+
STM2214	2013	Live poultry	STM2114	2016	Bovine	G7	–	+
STM2215	2011	Ground beef	STM1563	2011	Bovine	G6	+	+
STM2216	2011	Ground beef	STM1563	2011	Bovine	G6	+	+
STM2217	2015	Pork	STM2116	2016	Swine	G2a	+	+
STM1016	2010	Cattle contact	STM328	2008	Bovine	G9	+	+
STM1075	2010	Cattle contact	STM978	2010	Bovine	G2a	+	+
STM995	2010	Cattle contact	STM978	2010	Bovine	G2a	+	–
STM996	2010	Cattle contact	STM978	2010	Bovine	G2a	+	–
STM1065	1998	Raw milk	STM034	2004	Bovine	G9	+	+
STM988	2009	Chicken	STM1975	2015	Bovine	G9	–	+

*RF, Random Forest; +, correct prediction; –, incorrect prediction.
†The livestock genome that had the most recent common ancestor with an outbreak query genome.

- **Source attribution by phylogenetic placement:**
  1. Scalability 2. Qualitative vs. quantitative 3. HGT 4. What genetic features are source markers?

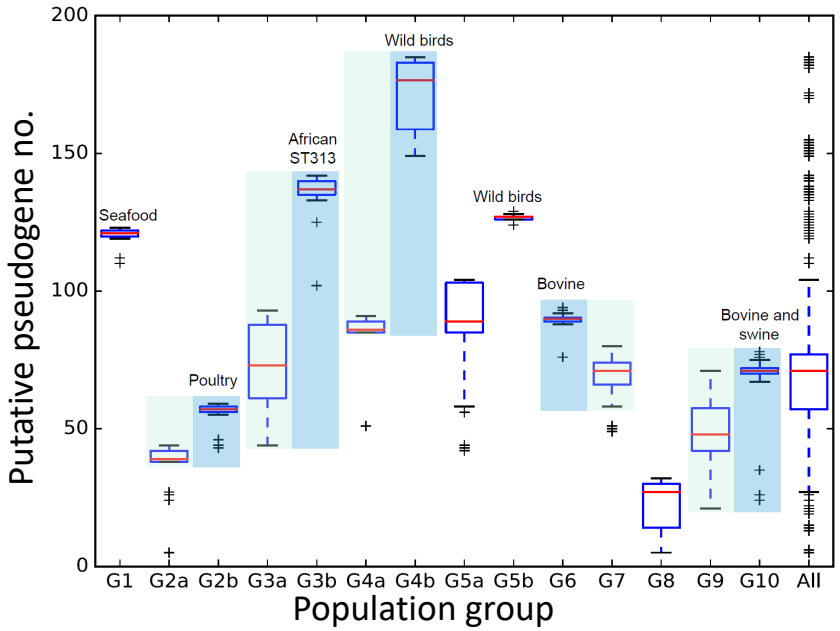


- **Top 50 features are sufficient for robust zoonotic source prediction**

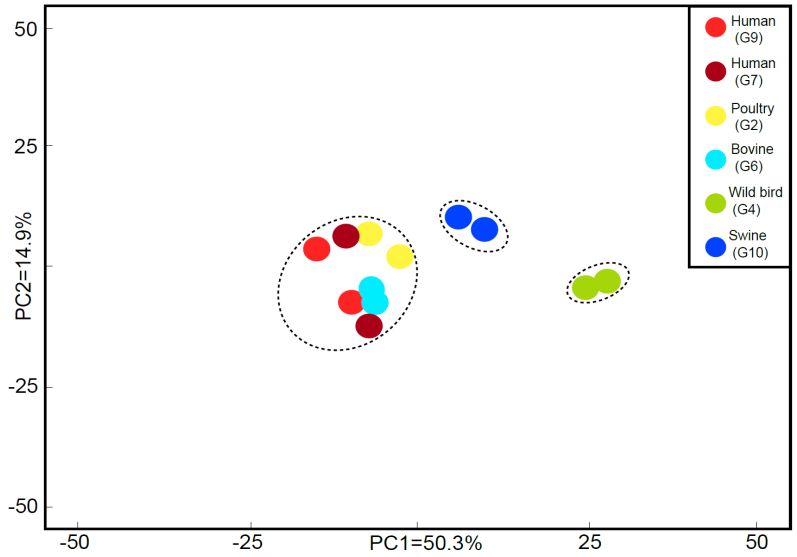


- **Mobile genetic elements as source markers**
- **No.1 feature is a point mutation in *fliC* gene, (flagellar subunit, surface protein, host interaction?)**
- **Causation of or correlation with host preference?**
  - 24 of the top 50 features had been functionally studied
  - 14 had a role in animal host colonization (Chaudhuri et al. 2013)

- **Bacterial host adaptation often coincides with pseudogene accumulation**
  - Human restricted serotypes Typhi and Paratyphi A (Holt et al. 2009)
- **Livestock clades diverged from clades of diverse sources (i.e., generalists)**
- **Elevated pseudogene accumulation in recently diverged livestock clades**

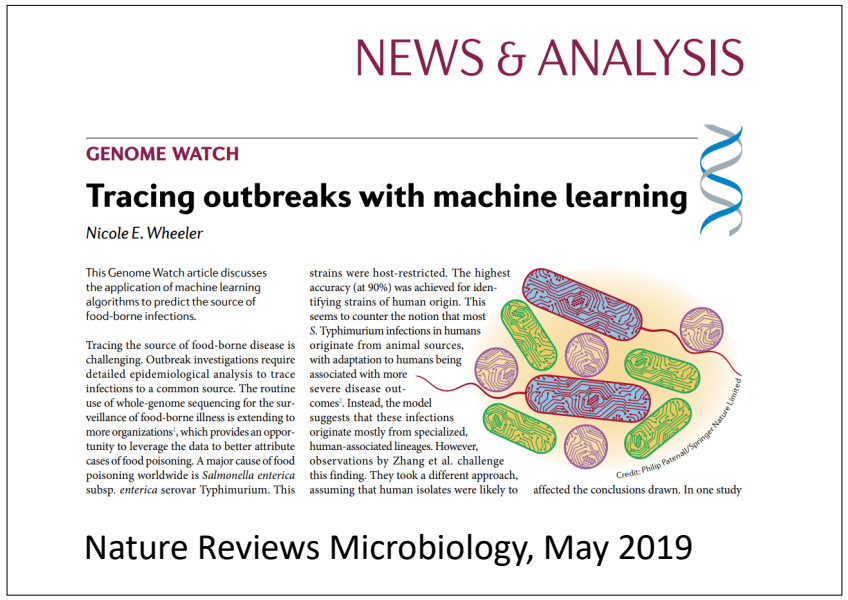


- **Comparison of metabolic profiles**
  - Utilization of 384 substrates of carbon, nitrogen, sulfur and phosphorus sources
  - Distinct metabolic profiles of isolates from wild bird and pig, indicating metabolic acclimation to host



- **Machine learning: black box?**
  - Back up ML findings with logic and theory
  - When ML findings contradict established notions: Specialized ST lineages associated with humans?
  - Most human infections of ST originate from animals. Human is not a reservoir and attributable source of ST.

- **Machine learning study design: training data matter!**



NEWS & ANALYSIS

GENOME WATCH

Tracing outbreaks with machine learning

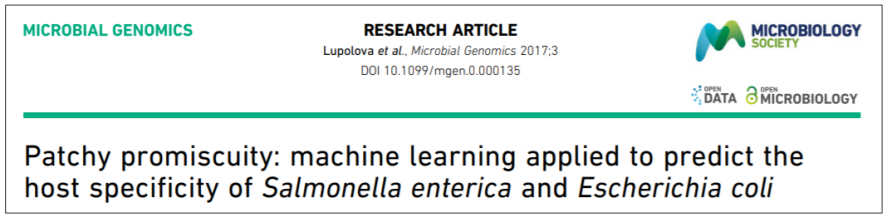
Nicole E. Wheeler

This Genome Watch article discusses the application of machine learning algorithms to predict the source of food-borne infections.

Tracing the source of food-borne disease is challenging. Outbreak investigations require detailed epidemiological analysis to trace infections to a common source. The routine use of whole-genome sequencing for the surveillance of food-borne illness is extending to more organizations[1], which provides an opportunity to leverage the data to better attribute cases of food poisoning. A major cause of food poisoning worldwide is *Salmonella enterica* subsp. *enterica* serovar Typhimurium. This strains were host-restricted. The highest accuracy (at 90%) was achieved for identifying strains of human origin. This seems to counter the notion that most *S.* Typhimurium infections in humans originate from animal sources, with adaptation to humans being associated with more severe disease outcomes[1]. Instead, the model suggests that these infections originate mostly from specialized, human-associated lineages. However, observations by Zhang et al. challenge this finding. They took a different approach, assuming that human isolates were likely to affected the conclusions drawn. In one study

Credit: Philip Pattenall/Springer Nature Limited

Nature Reviews Microbiology, May 2019



MICROBIAL GENOMICS

RESEARCH ARTICLE

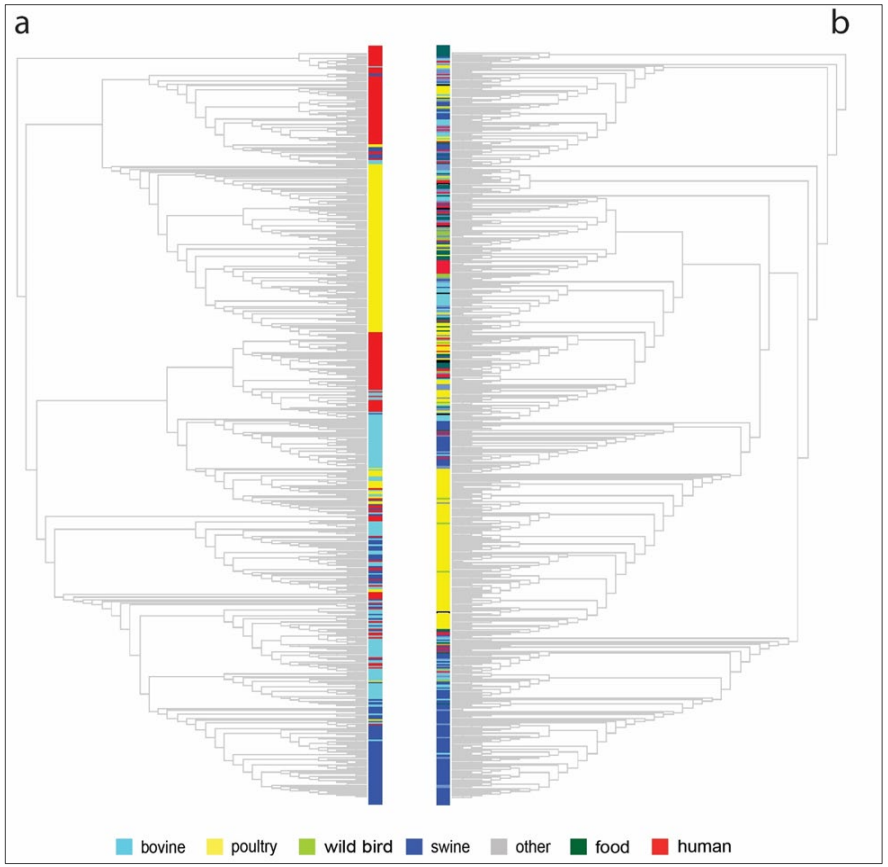
Lupolova *et al.*, *Microbial Genomics* 2017;3

DOI 10.1099/mgen.0.000135

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*



Limitations and future work

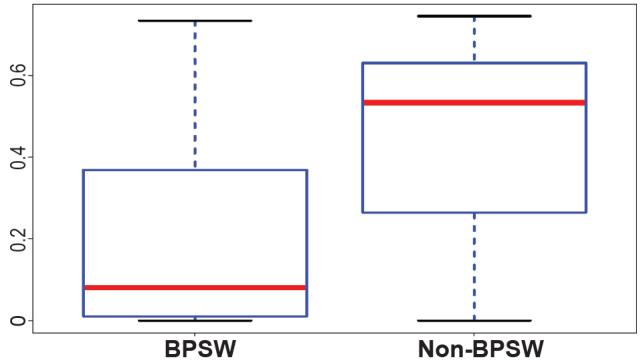
- Isolates from sources unknown to the machine learning model?

- Generalist strains?

- More sources and serotypes for source attribution?

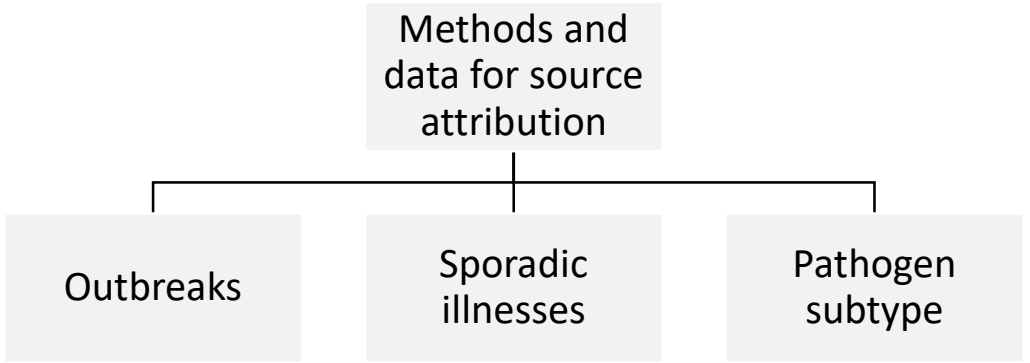
Confused:
25% Bovine
23% Poultry
27% Swine
25% Wild birds

Confident:
8% Bovine
80% Poultry
8% Swine
2% Wild birds

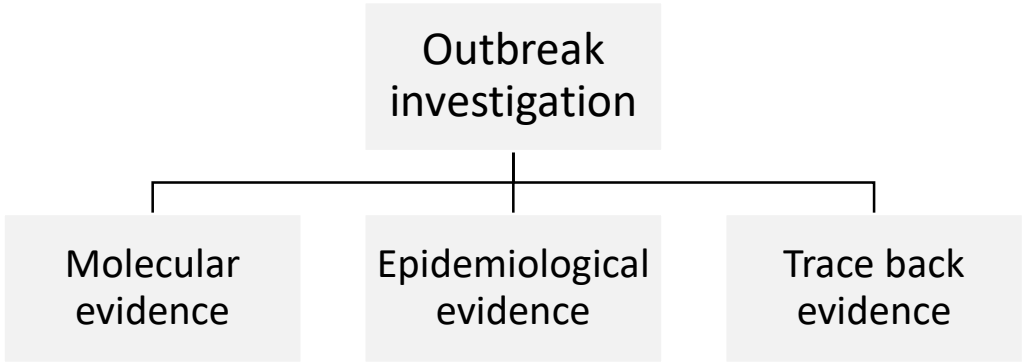


Simpson index of source probabilities
(BPSW: bovine, poultry, swine, and wild birds)

- **Attribution of foodborne illness**



- **Outbreak investigation**
  - Educated guess, not evidence
  - Guidance and clue for seeking evidence



- **Root cause of contamination?**

Trace back analysis

- **Isolates from the 2010 alfalfa sprouts outbreak showed strong poultry signals**

THE VERGE ☰

Machine learning could help figure out what pooped on your produce

Scientists harnessed machine learning to analyze Salmonella genomes, and predict which animal they came from

By Rachel Becker on December 12, 2018 1:59 pm

DEPARTMENT OF HEALTH AND HUMAN SERVICES
FOOD AND DRUG ADMINISTRATION

DISTRICT OFFICE ADDRESS AND PHONE NUMBER	DATE(S) OF INSPECTION
USFDA Chicago District Office 550 W Jackson Blvd Ste 1500 Chicago, IL 60661 phone 312-353-5863 Industry Information: www.fda.gov/oc/industry	12/20/10, 12/21/10, 12/28/10, 01/06/11, 01/07/11, 01/19/11, 01/28/11
	FEI NUMBER 1000515256

NAME AND TITLE OF INDIVIDUAL TO WHOM REPORT IS ISSUED

TO: [redacted]

FIRM NAME	STREET ADDRESS
[redacted]	[redacted]
CITY, STATE AND ZIP CODE	TYPE OF ESTABLISHMENT INSPECTED
[redacted]	sprout grower/manufacturer, dealer, distributor

THIS DOCUMENT LISTS OBSERVATIONS MADE BY THE FDA REPRESENTATIVE(S) DURING THE INSPECTION OF YOUR FACILITY. THEY ARE INSPECTIONAL OBSERVATIONS; AND DO NOT REPRESENT A FINAL AGENCY DETERMINATION REGARDING YOUR COMPLIANCE. IF YOU HAVE AN OBJECTION REGARDING AN OBSERVATION, OR HAVE IMPLEMENTED, OR PLAN TO IMPLEMENT CORRECTIVE ACTION IN RESPONSE TO AN OBSERVATION, YOU MAY DISCUSS THE OBJECTION OR ACTION WITH THE FDA REPRESENTATIVE(S) DURING THE INSPECTION OR SUBMIT THIS INFORMATION TO FDA AT THE ADDRESS ABOVE. IF YOU HAVE ANY QUESTIONS, PLEASE CONTACT FDA AT THE PHONE NUMBER AND ADDRESS ABOVE.

DURING AN INSPECTION OF YOUR FIRM (I) (WE) OBSERVED:

1. The production facility has one entrance by the loading docks, one entrance into the reception area, two entrances into the greenhouse, one entrance into the kitchen and one entrance into an office. All of the entrances are pathways for people and equipment into the production area. Employees were observed using all of the entrances.

a. On 12/21/10 Investigator Speer observed an employee in a red hooded sweat shirt dumping production waste into a compost pile then return to the production area via the south greenhouse entrance. The grade of the land is sloped down towards the greenhouse entrance. Run off water was observed pooling from the compost pile into drain along the walkway 11 ft from the entrance to the greenhouse. After walking through the compost pile and water, this employee returned to production wearing the same clothing and boots. This is the site of positive identification of Salmonella 4, 5, 12, i:, identified in the outbreak.

EMERGING INFECTIOUS DISEASES®

EID Journal > Volume 25 > Number 1—January 2019 > Main Article

RESEARCH

Zoonotic Source Attribution of *Salmonella enterica* Serotype Typhimurium Using Genomic Surveillance Data, United States

Shaokang Zhang, Shaoting Li, Weidong Gu, Henk den Bakker, Dave Boxrud, Angie Taylor, Chandler Roe, Elizabeth Driebe, David M. Engelthaler, Marc Allard, Eric Brown, Patrick McDermott, Shaohua Zhao, Beau B. Bruce, Eija Trees, Patricia I. Fields, Xiangyu Deng

Using Genomic Data to Find the Source of Salmonella enterica Typhimurium

Dr. Xiangyu Deng, an assistant professor at the University of Georgia, discusses the use of genomic sequencing to identify the source of salmonella infection.

More info on this topic

Topics: EID Journal (Emerging Infectious Diseases), Emerging Infectious Diseases, Genetics and Genomics Health, Public Health, Salmonella

Published: 12/19/2018

Last Updated: 12/19/2018

This podcast belongs to the Emerging Infectious Diseases series.

Subscribe to this Series

Press Play to listen to this CDC Podcast

Running time = 13:50

0:00 / 13:49

To save the Podcast, right click the "Save this file" link below and select the "Save Target As..." option.

Save This File (19.0 MB) [right click]

Other Formats: Read Transcript (Pdf)

Frank Yiannas ✓
@FrankYiannasFDA

Follow

A new era of smarter food safety & epidemiology. Study suggests machine learning of whole genome sequences might help identify root, zoonotic sources of some foodborne outbreaks.
wwwnc.cdc.gov/eid/article/25 …

8:11 PM - 13 Dec 2018

Thanks

Food Safety Informatics Group@UGA CFS

Center for Food Safety

UNIVERSITY OF GEORGIA

www.denglab.site