U.S. FOOD & DRUG ADMINISTRATION

# Genetic Diversity in Food Facilities

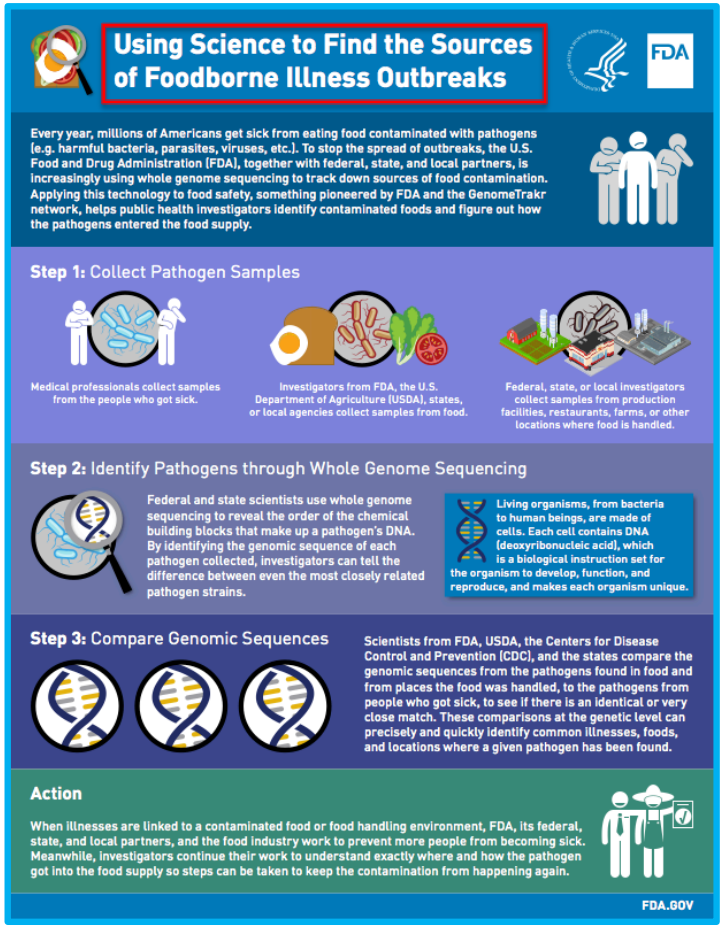
Yu (Fish) Wang, Ph.D.
Staff Fellow, Biostatistics and Bioinformatics Staff
Center for Food Safety and Applied Nutrition
U.S. Food Drug Administration

25 Sept. 2018

yu.wang1@fda.hhs.gov

# WGS in FDA


U.S. Department of Health and Human Services
U.S. FOOD & DRUG ADMINISTRATION
A to Z Index | Follow FDA | En Español
Search FDA
Home | Food | Drugs | Medical Devices | Radiation-Emitting Products | Vaccines, Blood & Biologics | Animal & Veterinary | Cosmetics | Tobacco Products
Food
Home › Food › Science & Research (Food) › Whole Genome Sequencing (WGS) Program
Whole Genome Sequencing (WGS) Program
GenomeTrakr Network
GenomeTrakr Fast Facts
GenomeTrakr Video Library
Researchers
WGS Events
International WGS Efforts
Whole Genome Sequencing (WGS) Program
SHARE TWEET LINKEDIN PIN IT EMAIL PRINT
Whole genome sequencing (WGS) is a cutting-edge technology that FDA has put to a novel and health-promoting use. FDA is laying the foundation for the use of whole genome sequencing to protect consumers from foodborne illness in countries all over the world.
On this page:
Introduction
GenomeTrakr: Using Genomics to Identify Food Contamination
How FDA Uses Whole Genome Sequencing for Regulatory Purposes
Proactive Applications of Whole Genome Sequencing Technology
Introduction
Whole genome sequencing reveals the complete DNA make-up of an organism, enabling us to better understand variations both within and between species. This in turn allows us to differentiate between organisms with a precision that other technologies do not allow. FDA is using this technology to perform basic foodborne pathogen identification during foodborne illness outbreaks and applying it in novel ways that have the potential to help reduce foodborne illnesses and deaths over the long term both in the U.S and abroad.
Using Science to Find the Sources of Outbreaks
(PDF: 544KB)

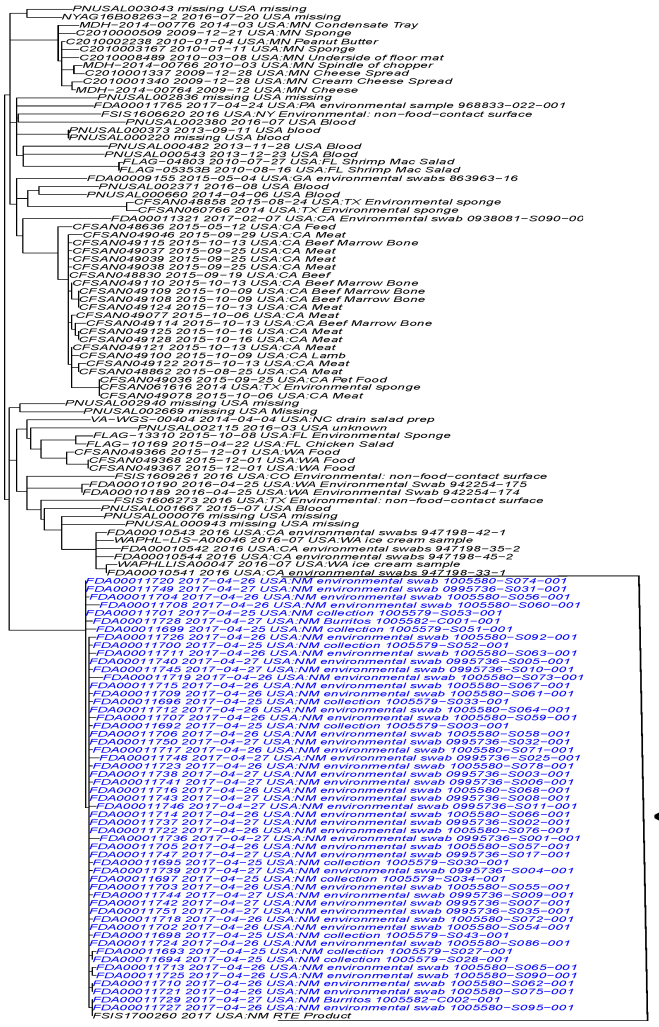

Using Science to Find the Sources of Foodborne Illness Outbreaks
Every year, millions of Americans get sick from eating food contaminated with pathogens (e.g. harmful bacteria, parasites, viruses, etc.). To stop the spread of outbreaks, the U.S. Food and Drug Administration (FDA), together with federal, state, and local partners, is increasingly using whole genome sequencing to track down sources of food contamination. Applying this technology to food safety, something pioneered by FDA and the GenomeTrakr network, helps public health investigators identify contaminated foods and figure out how the pathogens entered the food supply.
Step 1: Collect Pathogen Samples
Medical professionals collect samples from the people who got sick.
Investigators from FDA, the U.S. Department of Agriculture (USDA), states, or local agencies collect samples from food.
Federal, state, or local investigators collect samples from production facilities, restaurants, farms, or other locations where food is handled.
Step 2: Identify Pathogens through Whole Genome Sequencing
Federal and state scientists use whole genome sequencing to reveal the order of the chemical building blocks that make up a pathogen's DNA. By identifying the genomic sequence of each pathogen collected, investigators can tell the difference between even the most closely related pathogen strains.
Living organisms, from bacteria to human beings, are made of cells. Each cell contains DNA (deoxyribonucleic acid), which is a biological instruction set for the organism to develop, function, and reproduce, and makes each organism unique.
Step 3: Compare Genomic Sequences
Scientists from FDA, USDA, the Centers for Disease Control and Prevention (CDC), and the states compare the genomic sequences from the pathogens found in food and from places the food was handled, to the pathogens from people who got sick, to see if there is an identical or very close match. These comparisons at the genetic level can precisely and quickly identify common illnesses, foods, and locations where a given pathogen has been found.
Action
When illnesses are linked to a contaminated food or food handling environment, FDA, its federal, state, and local partners, and the food industry work to prevent more people from becoming sick. Meanwhile, investigators continue their work to understand exactly where and how the pathogen got into the food supply so steps can be taken to keep the contamination from happening again.
FDA.GOV

# How to Interpret WGS Analyses

**TABLE 2 |** Conditions used to determine whether whole-genome sequence analyses support a match between two or more genomes.

| | Supports | Neutral | Does not support |
|---|---|---|---|
| SNP distance | <21 | 21–100 | >100 |
| Bootstrap support | >0.89 | 0.80–0.89 | <0.80 |
| Tree topology | Monophyletic | Paraphyletic | Polyphyletic |

Arthur W. Pightling, James B. Pettengill, Yan Luo, Joseph D. Baugher, Hugh Rand, and Errol Strain. *Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. Front Microbiol*. 2018; 9: 1482.

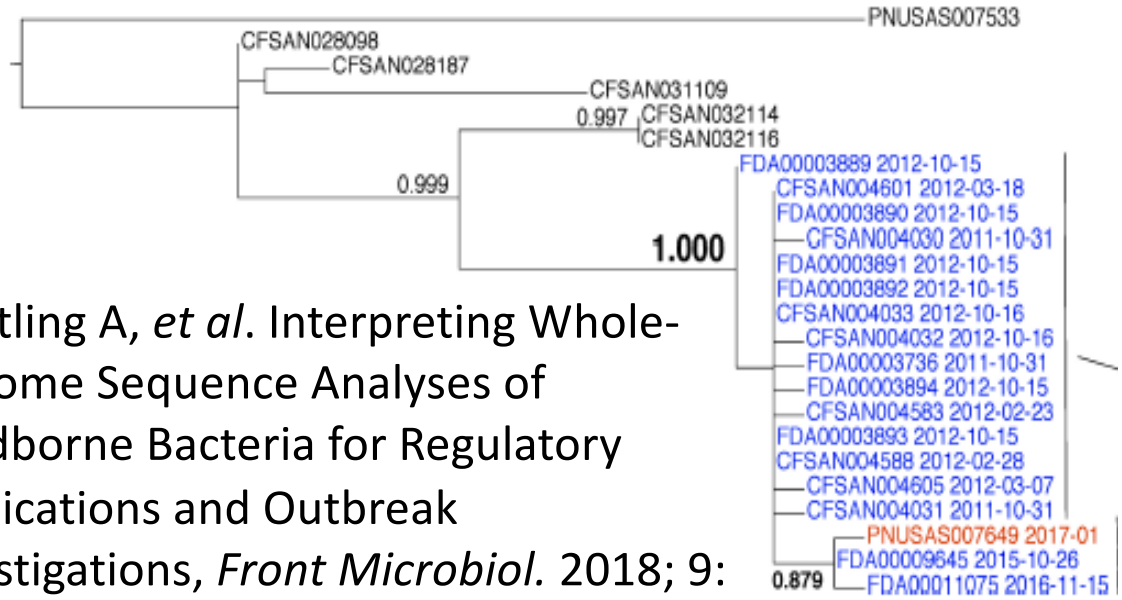



<= 9 SNPs

# *Observation I*:

Isolates with small genetic distances are generally from the same facility.

# *Observation 1*:



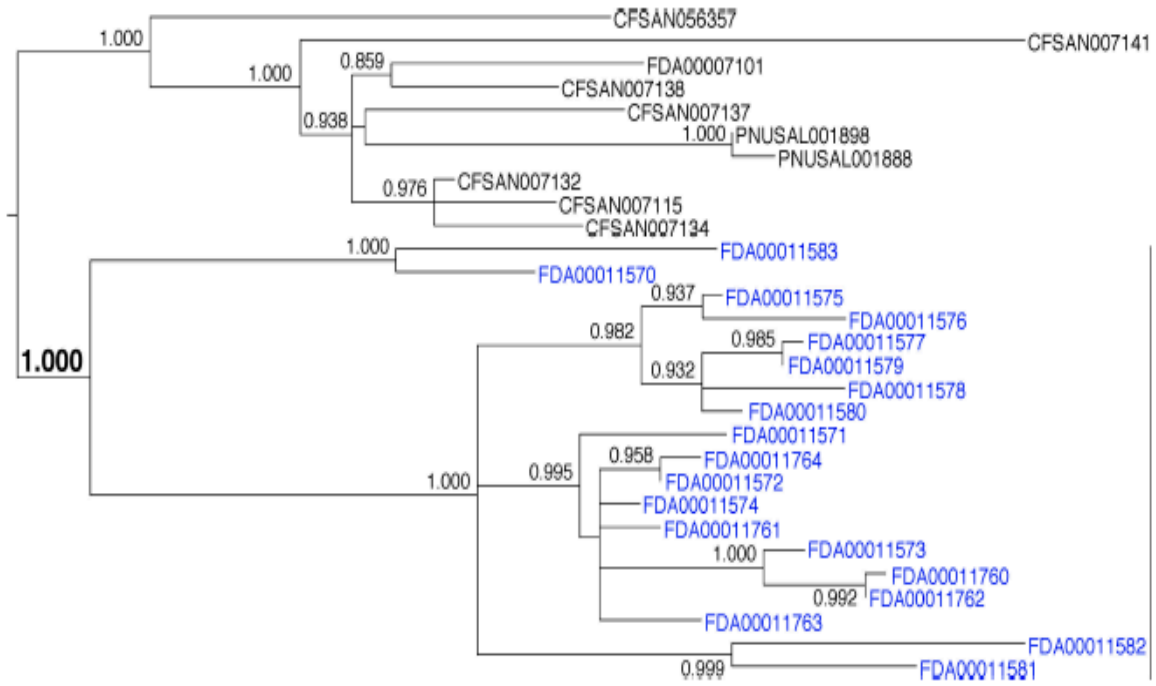

Pightling A, *et al*. Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations, *Front Microbiol.* 2018; 9: 1482.

Range: 0-4 SNPs

Median: 1 SNP

Isolates collected from a single facility could be genetically very close.

# *Observation II*:




But not always.

Range: 1-73 SNPs

Median: 29 SNPs

Pightling A, *et al*. Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations, *Front Microbiol.* 2018; 9: 1482.

# Initial Focus

- P(F|D ≤ ***d***): What is the probability that 2 isolates are collected from the same food facility if their genetic distance is no more than ***d*** SNPs?
- P(D ≤ ***d***|F): What is the probability that the genetic distance of 2 isolates is no more than ***d*** SNPs if both of them are collected from the same food facility?

# Available Data

| Database | Owner | Data |
|---|---|---|
| Pathogen Detection | NCBI | • Clusters (cluster ID, biosample accessions, SRA accessions)<br>• Phylogenetic trees |
| SRA | NCBI | • Sequence data |
| FACTS | FDA/ORA | • FACTS ID (FDA Sample ID)<br>• Sample information (responsible firm, other facilities' information, collection information, product description, etc.) |
| GIMS | FDA/CFSAN | • Biosample accessions<br>• FACTS ID<br>• Some collection information |

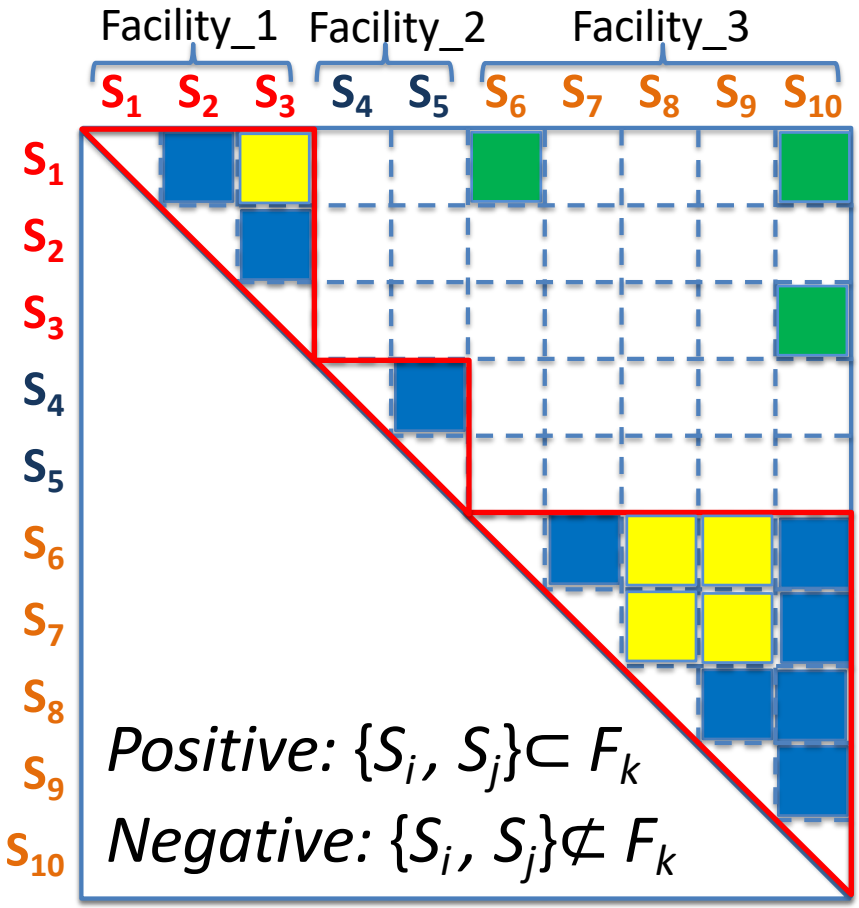
# Data Summary

| Organism | *Salmonella* | *Listeria* |
|---|---|---|
| Collection date | 04/04/1999 – 03/24/2017 | 07/07/1999 – 07/24/2017 |
| # of Isolates | 6351 | 5321 |
| # of Facilities | 2196 | 846 |
| # of Isolates in NCBI Clusters | 4029 | 3044 |
| # of NCBI Clusters | 779 | 248 |

# Computation of Genetic Distances

- CFSAN SNP Pipeline v0.7.0
  - A Python-based pipeline for the production of SNP matrices from sequence data
  - Using reference-based alignments to create a matrix of SNPs for a given set of samples
  - *Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, Strain E. (2015). PeerJ Computer Science 1:e20 https://doi.org/10.7717/peerj-cs.20*
  - https://github.com/CFSAN-Biostatistics/snp-pipeline

# Probabilities Estimation




■ $\{S_i, S_j\} \subset F_k, D(S_i, S_j) \le \boldsymbol{d}$

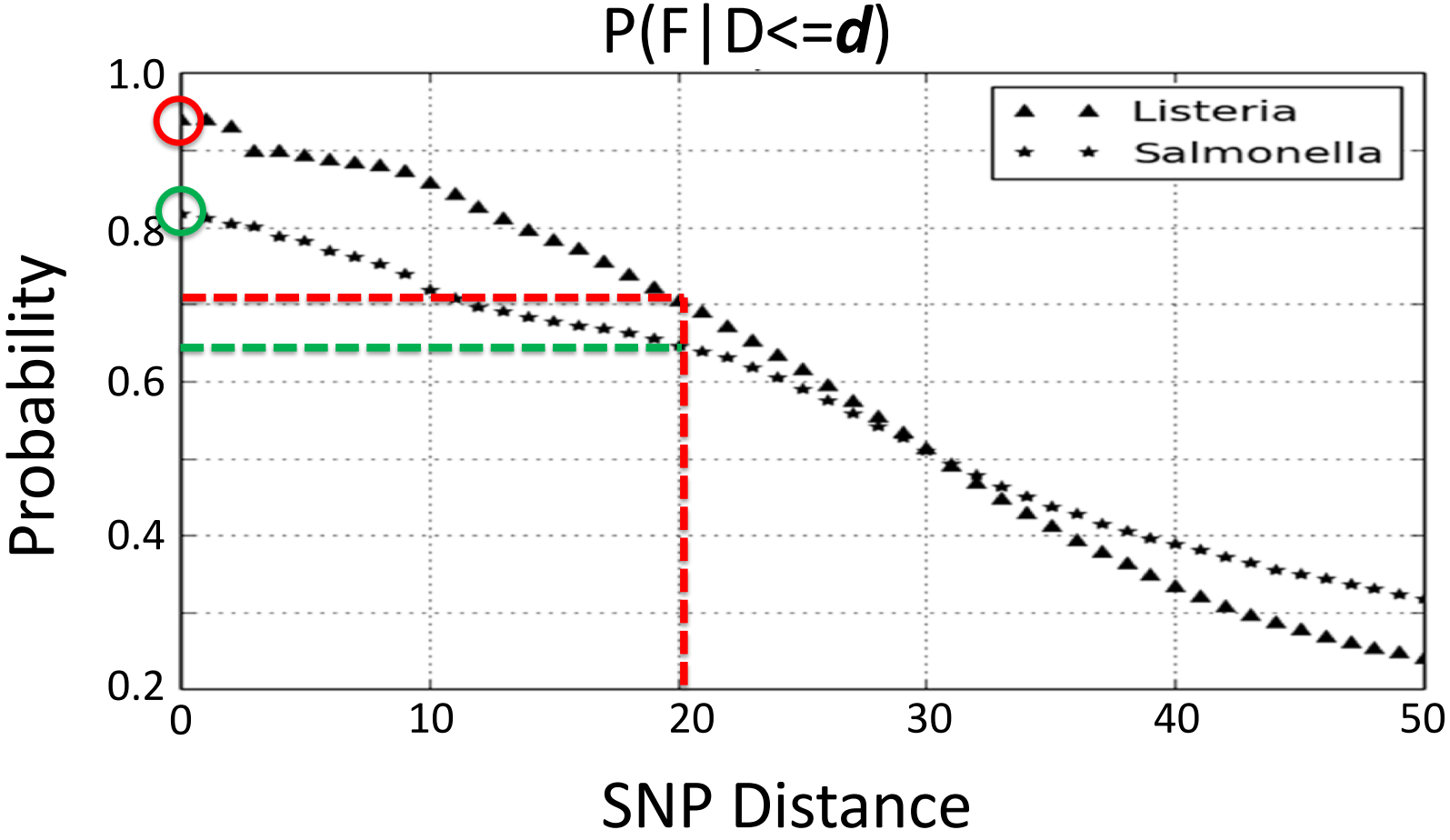
■ $\{S_i, S_j\} \subset F_k, D(S_i, S_j) > \boldsymbol{d}$

■ $\{S_i, S_j\} \not\subset F_k, D(S_i, S_j) \le \boldsymbol{d}$

$$P(F \mid D \le \boldsymbol{d}) = \frac{Num(■)}{Num(■)+Num(■)}$$

$$P(D \le \boldsymbol{d} \mid F) = \frac{Num(■)}{Num(■)+Num(■)}$$

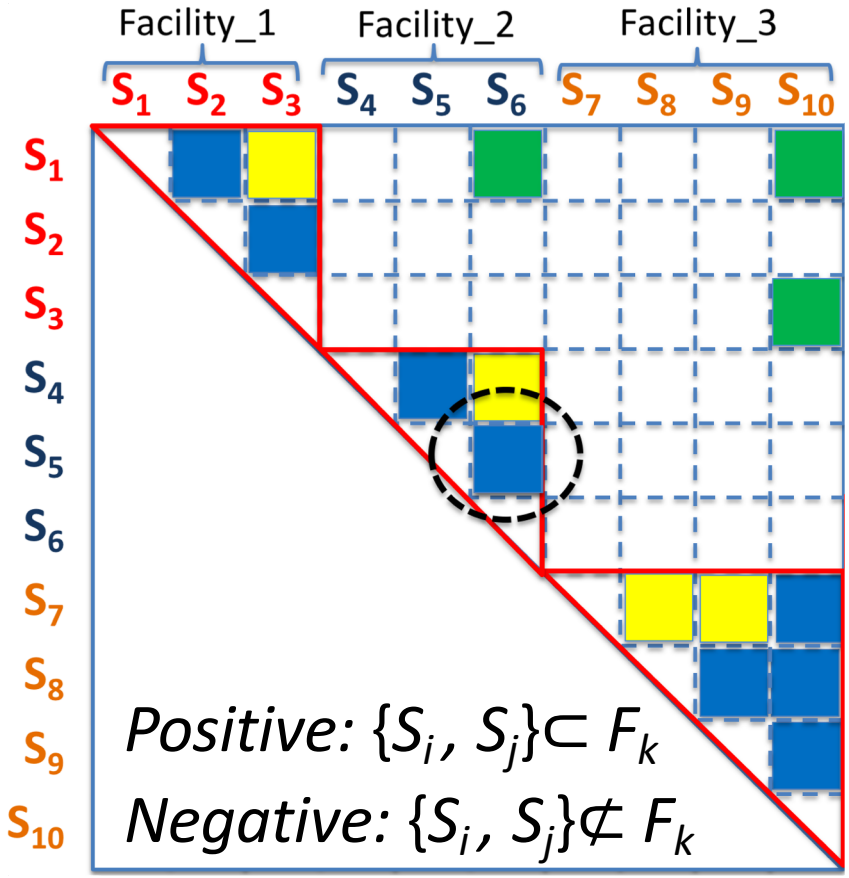
# Facility Match Probability



# Critiquing the Results

- How much should we believe our results
  - Bias
    - Sampling program is targeted, not random
    - Dirty facilities are overrepresented in data set
    - An isolate-pair is classified as positive or negative only by responsible firms
    - Analysis methodology
    - Consistent, unknown patterns in supply chains
  - Variability
    - Experiment-to-experiment
    - Sampling noise

# Bias: More than Responsible Firm




$D(S_i, S_j) \le \boldsymbol{d}$

$D(S_i, S_j) > \boldsymbol{d}$

$D(S_i, S_j) \le \boldsymbol{d}$

$$P(F \mid D \le \boldsymbol{d}) = \frac{Num(\blacksquare)}{Num(\blacksquare)+Num(\blacksquare)}$$

# Bias Evaluation for P(F|D<=*d*)

- ***False Negative***: isolate-pair ($S_i$, $S_j$)
  - $S_i$, $S_j$ have different Responsible Firms
  - Upon review, determined to come from a single facility
- Isolates with different Responsible Firms: possible "overlap" in their supply chains
  - Same facility in supply chains
  - Facilities are adjacent to each other.

# False Negative: Example 1

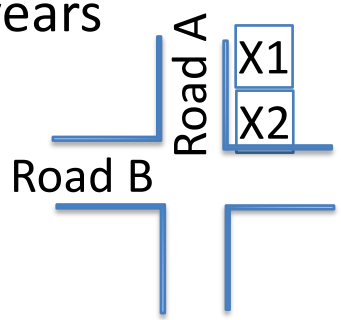
- *Salmonella* contamination
  - Sample 1
    - Responsible Firm: Company X1 in Georgia
    - Shipper: Company Y1 in California
  - Sample 2
    - Responsible Firm: Company X2 in Massachusetts
  - Sample 3
    - Responsible Firm: Company X3 in California
  - Genetic distances among samples: < 5 SNPs
  - Connection:
    - Y1 and X3 are the same company
    - All samples: frozen tuna imported from India
    - Difference in collection dates: < 1 week

# False Negative: Example 2

- *Listeria* contamination
  - Sample 1
    - Responsible Firm: Company X1 in California
    - Grower: Company Y1 in North Dakota
  - Sample 2
    - Responsible Firm: Company X2 in Texas
    - Ingredient Supplier: Company Y2 in Minnesota
  - Genetic distance: 0 SNP
  - Connection:
    - Y1 is doing business as Y2
    - Both samples: soybean products
    - Difference in collection dates: < 1 week

# False Negative: Example 3

- *Listeria* contamination
  - Sample 1
    - Responsible Firm: Company X1 in New York
  - Sample 2
    - Responsible Firm: Company X2 in New York
  - Genetic distance: 1 SNP
  - Difference in collection dates: > 2 years
  - Connection:
    - X1 and X2 are adjacent
      - X1: Road A address
      - X2: Road B address



# Bias Evaluation

- False Omission Rate (***FOR***)

$$\textbf{\textit{FOR}} = \frac{\textbf{\textit{Number of False Negative}}}{\textbf{\textit{Total Number of Negative Calls}}}$$

- Method:
  - $d \in \{0, 5, 10, 15, 20\}$, find all ***negative*** isolate-pairs
  - For each distance, randomly pick 30 isolate-pairs → 5 sets of negative isolate-pairs ($Q_1$, $Q_2$, $Q_3$, $Q_4$, $Q_5$)
  - For each isolate-pair, verify whether false-negative
  - $\boldsymbol{N_i}$ : the number of ***False Negatives*** isolate-pairs identified in $Q_i$ ($i$=1, 2, 3, 4, 5)

# Bias Evaluation: Method (cont.)

- Construct 5 groups $G_k$ ($k$=1, 2, 3, 4, 5) of isolate-pairs.

| $G_1$ | $G_2$ | $G_3$ | $G_4$ | $G_5$ |
|---|---|---|---|---|
| $Q_1$ | $Q_1$ $Q_2$ | $Q_1$ $Q_2$ $Q_3$ | $Q_1$ $Q_2$ $Q_3$ $Q_4$ | $Q_1$ $Q_2$ $Q_3$ $Q_4$ $Q_5$ |

- In group $G_k$,
  - $D$ <= ($k$-1)*5
  - 30*$k$ negative isolate-pairs (i.e., ***Total Negatives***)
  - Number of False Negative: $\sum_{i=1}^{k} N_i$
  - Estimated ***FOR***: $\widetilde{p_k} = \frac{\sum_{i=1}^{k} N_i}{30 \times k}$

# Bias Evaluation: Method (cont.)

- ***FOR*** and its margin of error for binomial proportion 95% confidence interval ($k$=1,2,3,4,5)

$$\widetilde{\boldsymbol{FOR}}_k = \widetilde{p_k} \pm 1.96 \sqrt{\frac{\widetilde{p_k}}{30 \times k}\left(1 - \widetilde{p_k}\right)}$$

- For $P = P(F|D \leq \boldsymbol{d})$ , $\boldsymbol{d} \in \{0, 5, 10, 15, 20\}$

$$\boldsymbol{Bias} = (\ 1 - P\ ) \times \boldsymbol{FOR}$$

After-bias-adjustment probabilities $\hat{P}$ = $P$ + ***Bias***.

# Bias Adjustment Result: $P(F|D \leq d)$

## *Listeria*

| *d* | *P* | *FOR* | *Bias* | $\hat{P}$ |
|---|---|---|---|---|
| 0 | 0.94 | 1.00–0.00 | 0.06-0.00 | 1.00-0.00 |
| 5 | 0.89 | 0.60±0.12 | 0.07±0.01 | 0.96±0.01 |
| 10 | 0.86 | 0.40±0.10 | 0.06±0.01 | 0.92±0.01 |
| 15 | 0.79 | 0.30±0.08 | 0.06±0.02 | 0.85±0.02 |
| 20 | 0.70 | 0.24±0.07 | 0.07±0.02 | 0.77±0.02 |

*P* : *Result before bias-adjustment.*

$\hat{P}$ : *After-bias-adjustment probabilities* $\hat{P}$= *P* + ***Bias***

# Bias Adjustment Result: $P(F|D \leq d)$

## *Salmonella*

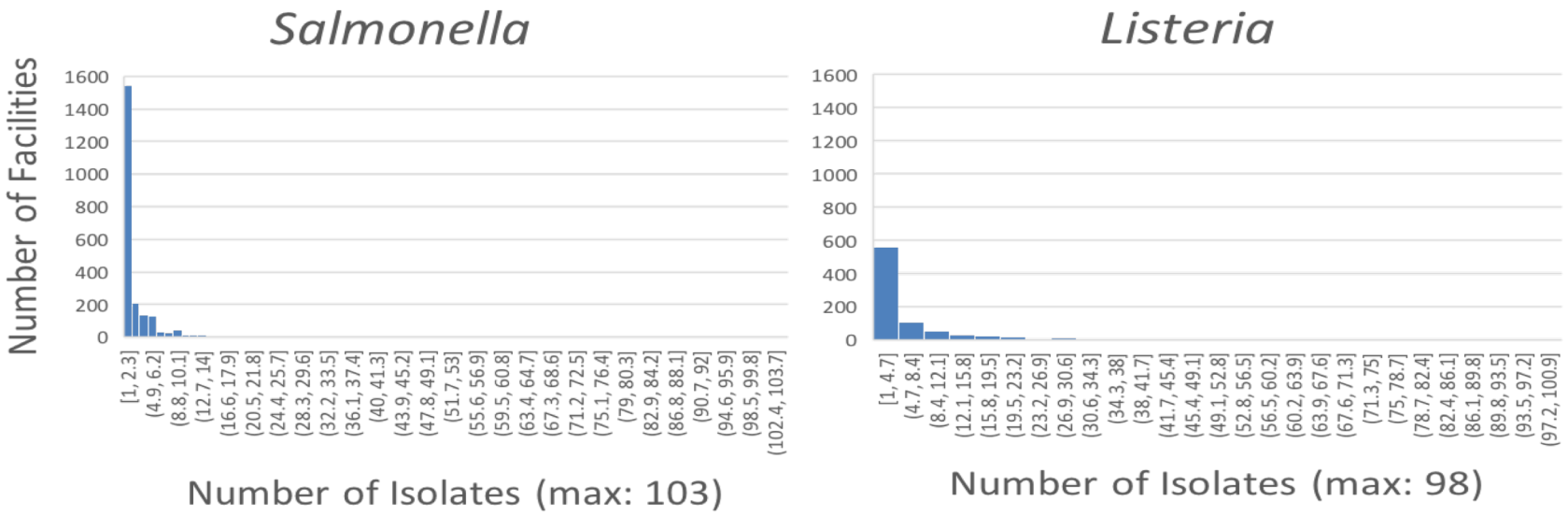
| $d$ | $P$ | $FOR$ | Bias | $\hat{P}$ |
|---|---|---|---|---|
| 0 | 0.82 | 0.47±0.18 | 0.08±0.03 | 0.90±0.03 |
| 5 | 0.78 | 0.35±0.12 | 0.08±0.03 | 0.85±0.03 |
| 10 | 0.72 | 0.39±0.10 | 0.11±0.03 | 0.83±0.03 |
| 15 | 0.68 | 0.30±0.08 | 0.10±0.03 | 0.78±0.03 |
| 20 | 0.65 | 0.25±0.07 | 0.09±0.02 | 0.74±0.02 |

*$P$ : Result before bias-adjustment.*

*$\hat{P}$ : After-bias-adjustment probabilities $\hat{P}$= P +* ***Bias***

# Effect of Sampling

- Different number of isolates collected from facilities



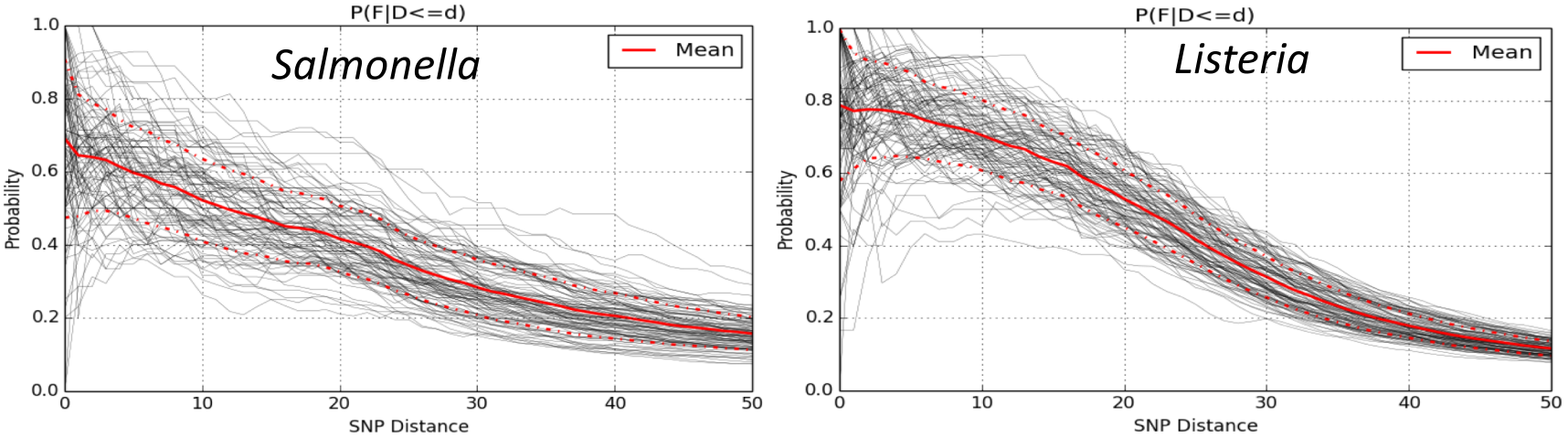

*Salmonella*: 6351 isolates from 2196 facilities

*Listeria*: 5321 isolates from 846 facilities

# Sub-sampling Analysis

- Sub-sampling process:
  - Step 1: for each facility, if it has
    - 1 isolate: remove from data set
    - 2 isolates: keep 2 isolates
    - >2 isolates: randomly select 2 isolates.
  - Step 2: apply the pipeline on the subset of the data
- Repeat the sub-sampling process 100 times
- Calculate mean and SD for $P(F|D \leq d)$

# Sub-sampling Results: $P(F|D \leq d)$

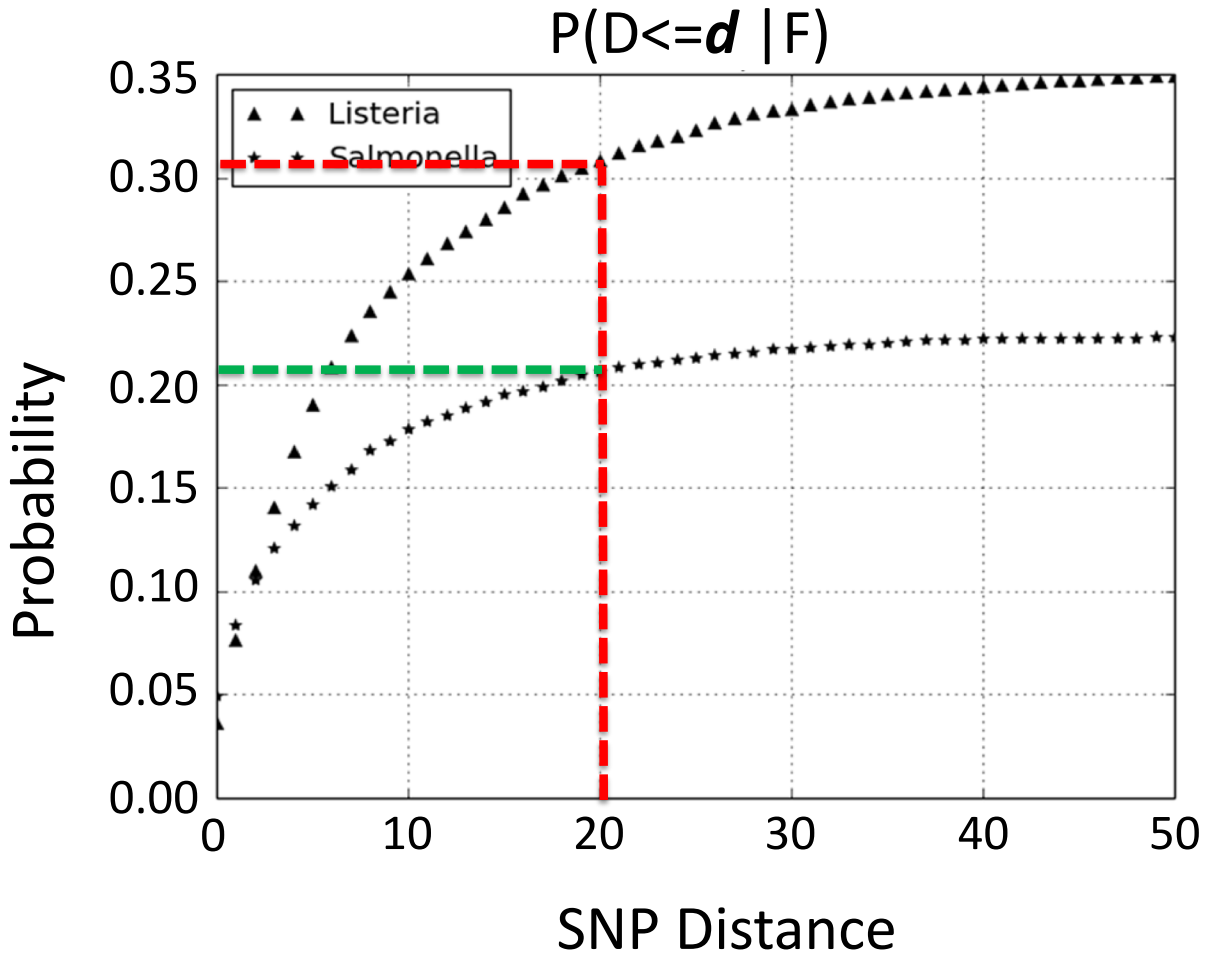



| | *Salmonella* | | | | *Listeria* | | |
|---|---|---|---|---|---|---|---|
| | *N* | *M* | *N/M* | | *N* | *M* | *N/M* |
| Original | 31,439 | 20,164,425 | 0.156% | | 55,916 | 14,153,860 | 0.395% |
| Sub-sampling | 1,135 | 2,575,315 | 0.044% | | 591 | 697,971 | 0.085% |

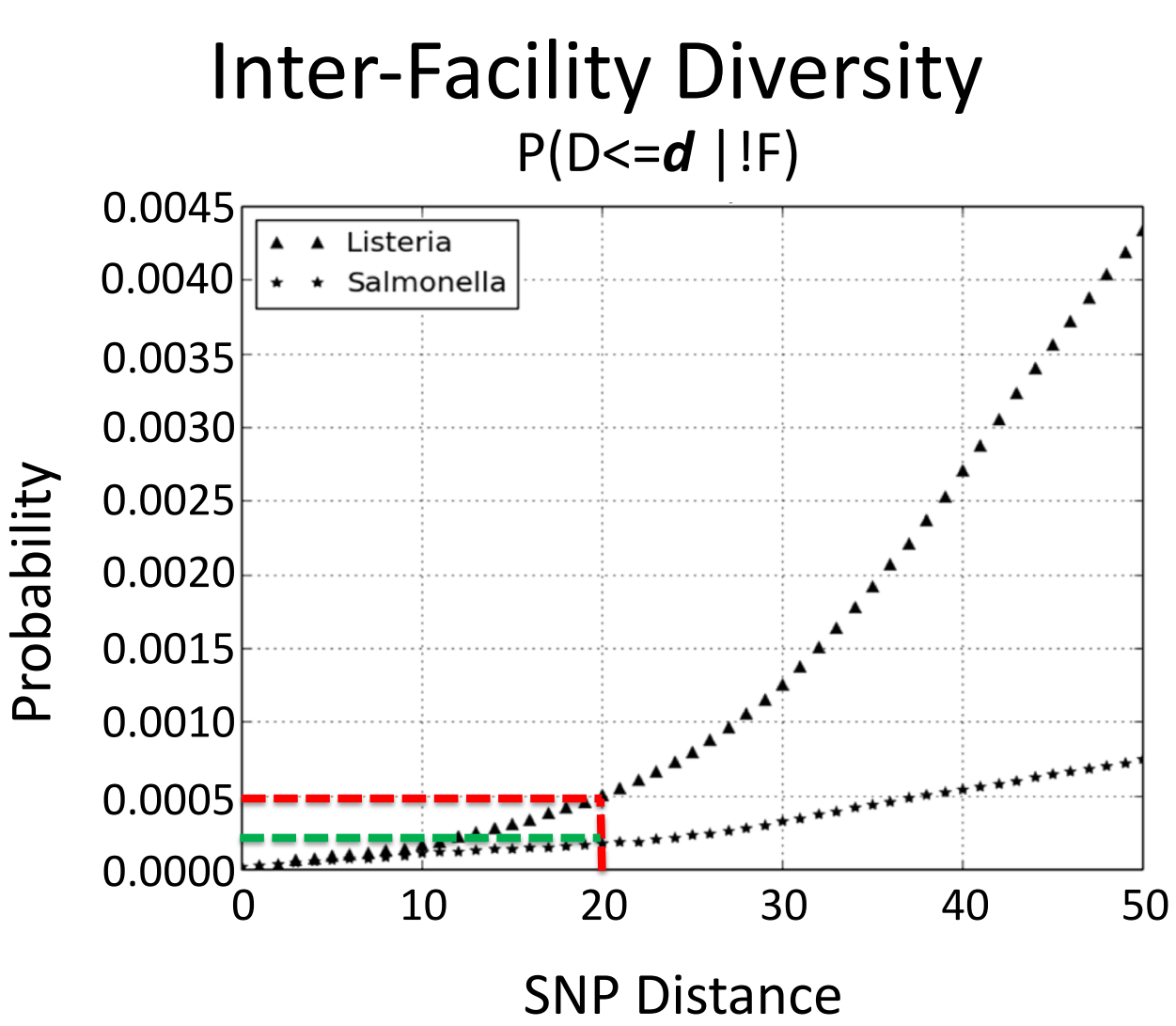
*M*: total number of isolate-pairs $\{S_i, S_j\} \not\subset F_k$

*N*: number of positive isolate-pairs $\{S_i, S_j\} \subset F_k$

# Within-Facility Diversity



# Inter-Facility Diversity



# Summary

- For *Salmonella* and *Listeria*, if the genetic distance of 2 isolates is <= 20 SNPs, then the 2 isolates are very likely collected from the same facility ($\hat{P}$(F|D<=20) $\approx$ 0.75).
- Noteworthy:
  - For *Listeria*: $\hat{P}$(F|D=0)=1.00, from the same facility
  - For *Salmonella*: $\hat{P}$(F|D=0)=0.90<1, not necessarily from the same facility

# Future Work

- Examine the influence of commodity type on our results
- Examine the influence of facility type on our results
- Examine time-between-collections as an explanatory factor in genetic distances

# Acknowledgement


- FDA/CFSAN/OAO/BBS
  - Hugh Rand
  - Arthur Pightling
  - James Pettengill
  - Yan Luo
  - Steven Davis
  - Errol Strain
- FDA/CFSAN/ORS
  - Ruth Timme
  - Marc Allard