Integration of Regulatory and Clinical Data: an FDA Perspective

Eric L. Stevens, PhD
U.S. Food & Drug Administration
Center for Food Safety and Applied Nutrition
Safe, Wholesome, Sanitary Foods
The Fresh-Cut Tomato Supply Chain
Understanding a SNP-based Approach

Bacterial Genome
3-5 million nucleotides

Cell Division

0.003 mutations per genome per cell generation

www.fda.gov
Understanding a SNP-based Approach

- 1 cell
- 2 cells
- 4 cells
- 8 cells
- 16 cells
- 32 cells
- 64 cells
- 128 cells
- 256 cells
- 512 cells
Contamination occurs at farm

Represents 1 SNP
Goal of Phylogenetic Trees using WGS Data:
Infer evolutionary relationships based on SNPs & match clinical to food/environmental isolates
FDA’s GenomeTrakr

- First distributed network of labs to use WGS:
  - 25 US State health and university labs
  - 15 federal labs
  - 1 US hospital lab
  - 20 international labs

- Focuses on environmental and food isolates to be a large reference database for clinical samples (150,000)

- Each genomic sequence has associated metadata
  - Species
  - Geographic location (within US has states, outside US has countries)
  - Date of isolation and Collector
  - Isolation source (food product, environmental swab)
# Minimal Pathogen Metadata

<table>
<thead>
<tr>
<th>What</th>
<th>When</th>
<th>Where</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample_name</td>
<td>collection_date</td>
<td>Geographic location</td>
</tr>
<tr>
<td>organism</td>
<td></td>
<td>6a) geo_loc_name</td>
</tr>
<tr>
<td>strain/isolate</td>
<td></td>
<td>OR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6b) lat_lon</td>
</tr>
</tbody>
</table>

**Category (attribute package)**

1a) **Clinical/Host-associated**
   - 1a1) specific host
   - 1a2) isolation source
   - 1a3) host-disease
   OR

1b) **Environmental/Food/Other**
   - 1b1) isolation source
Using NCBI to Access Data

Pathogen Detection BETA

NCBI Pathogen Detection integrates bacterial pathogen genomic sequences originating in food, environmental sources, and patients. It quickly clusters and identifies related sequences to uncover potential food contamination sources, helping public health scientists investigate foodborne disease outbreaks.

Find isolates now!

Explore the Data

<table>
<thead>
<tr>
<th>Species</th>
<th>New Isolates</th>
<th>Total Isolates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salmonella enterica</td>
<td>1,060</td>
<td>91,365</td>
</tr>
<tr>
<td>E.coli and Shigella</td>
<td>117</td>
<td>35,293</td>
</tr>
<tr>
<td>Listeria monocytogenes</td>
<td>11</td>
<td>15,717</td>
</tr>
<tr>
<td>Campylobacter jejuni</td>
<td>26</td>
<td>12,600</td>
</tr>
</tbody>
</table>

Data Resources

- [Isolates Browser](#)
- [Antimicrobial resistance reference gene database](#)
- [Isolates with antibiotic resistant phenotypes](#)
- [Beta-lactamase resources](#)
- [Download analysis results (FTP)](#)
## Using NCBI to Access Data

**Search Query:**
```plaintext
taxgroup_name:"Salmonella enterica"
```

**Results:**
- *Salmonella enterica PDS000002757.316* (5339 out of 5339)
- *Salmonella enterica PDS000015319.31* (3936 out of 3936)

**Clusters:**
- Show more clusters (up to 100)

### Table

<table>
<thead>
<tr>
<th>#</th>
<th>Organism G</th>
<th>Strain</th>
<th>Serovar</th>
<th>Isolate</th>
<th>Create</th>
<th>Locatiw</th>
<th>Isolatiw</th>
<th>Isolation</th>
<th>Host</th>
<th>SNP clust</th>
<th>Min</th>
<th>Min</th>
<th>BioSam</th>
<th>Assembly</th>
<th>K-mer grot</th>
<th>AST phenoty</th>
<th>AMR genotypy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>Salmonella enterica</em></td>
<td>WAP/A0003</td>
<td>enteric</td>
<td>PDT0000000</td>
<td>2015-02-11</td>
<td>USA:W</td>
<td>Cow</td>
<td>environme</td>
<td>PDS000000</td>
<td>0</td>
<td>3</td>
<td>SAMN02</td>
<td>PDG000000</td>
<td>ant(2&quot;)-la</td>
<td>aph(3&quot;)-lb</td>
<td>Show all 11 genes</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td><em>Salmonella enterica</em></td>
<td>WAP/A0003</td>
<td>enteric</td>
<td>PDT0000000</td>
<td>2015-02-11</td>
<td>USA:W</td>
<td>Cow</td>
<td>environme</td>
<td>PDS000000</td>
<td>0</td>
<td>3</td>
<td>SAMN02</td>
<td>PDG000000</td>
<td>ant(2&quot;)-la</td>
<td>aph(3&quot;)-lb</td>
<td>aph(3&quot;)-la</td>
<td>Show all 11 genes</td>
</tr>
<tr>
<td>3</td>
<td><em>Salmonella enterica</em></td>
<td>CFSA</td>
<td></td>
<td>PDT0000000</td>
<td>2013-09-04</td>
<td>USA:W</td>
<td>Ground Beef</td>
<td>environme</td>
<td>PDS000000</td>
<td>2</td>
<td>3</td>
<td>SAMN02</td>
<td>PDG000000</td>
<td>aadA</td>
<td>aph(3&quot;)-lb</td>
<td>Show all 8 genes</td>
<td></td>
</tr>
</tbody>
</table>
Why Open Data?

Why submit?

- To contribute information on pathogen sequences that may help discover sources of contamination or help in solving outbreaks more quickly
- To provide valuable real-time information about the relationship of an isolate to other isolates and outbreaks
- To enhance the set of pathogen genomes that can be used by the scientific community
- To supply information on the set of resistance genes present in a pathogen

The NCBI Pathogen Detection system is built on the foundation of open data. Data are intended to be submitted and released to the public immediately. Currently four major foodborne pathogens (Campylobacter, Escherichia coli and Shigella, Listeria, and Salmonella) are being analyzed in real time as participating public health agencies submit the sequences. Pathogens are also being analyzed for antimicrobial resistance.
Importance of a Balanced Approach

Clinical Samples

Maximum WGS Benefit

Food and Environmental Samples

www.fda.gov
Global Source Traceback

• In 2012, 410 individuals in the U.S. became sick from ingesting food that contained S. Bareilly

• CDC investigated a multistate (29 states) outbreak

• The illnesses were linked to frozen raw yellowfin tuna (Nakaochi Scrape) which was imported from India

• FDA retroactively examined the WGS of these isolates and compared them to previous S. Bareilly events
S. Bareilly Phylogeny

- 20-25 SNPs
- <=5 SNPs
- 110-130 SNPs

**PFGE Match**
Putting it all together

- Focus on collecting environmental and food isolates to be a large reference database for clinical samples

- Each genomic sequence has associated metadata

- Data made public so that:
  - Everyone can benefit from using this technology
  - Transparency of methods/data for regulatory decisions