The Role of Whole Genome Sequencing in Outbreak Detection and Investigation

Peter Evans
FDA-CFSAN
College Park, MD

Pathogen Control and Regulatory Compliance in Beef Processing Conference
September 9-10, Rosemont, IL
The Rise of the Genomes II: Practical Integration of Whole Genome Sequencing into Food Safety

Back to Basics: What is Whole Genome Sequencing? Why the Hype?  
David Engelthaler, Translational Genomics Research Institute

Whole Genome Sequencing Really Replacing Traditional Microbiology?  
Peter Gerner-Smidt, Centers for Disease Control and Prevention

Whole Genome Sequencing Has Transformed Detection and Investigation of Outbreaks!  
Kathie Grant, Public Health England

Whole Genome Sequencing for Surveillance of the Food Supply – Stopping Outbreaks before They Appear!  
Peter Evans, U.S. Food and Drug Administration

Practical Experience with Whole Genome Sequencing in the Food Industry  
Leen Baert, Nestlé Research Center
Foodborne Illness is common and costly

- Each year, 1 in 6 Americans gets sick from foodborne disease and 3,000 die as a result
- Reducing foodborne illness by just 10% would keep 5 million people a year from getting sick
- *Salmonella* infections alone are responsible for $365 million in direct medical costs annually
- Foodborne illness are preventable
Tracking (And Controlling) Pathogens In Food Is A Joint Responsibility

- Food producers, manufacturers and retailers
Different Roles—Shared Mission

Non-regulatory

**CDC**
- Disease surveillance
- Outbreak detection and investigation
- Analyzing burden, trends, and effectiveness of prevention efforts
- Attribution to sources
- Education and training
- Information for policy

Regulatory

**FDA and USDA**
- Inspection
- Enforcement
- Investigating farm and production facilities
- Product recall
- Product traceback
- Risk assessment and management
Steps in an OUTBREAK INVESTIGATION

1. DETECT A possible outbreak
2. FIND Cases in an outbreak
3. GENERATE Hypotheses through interviews
4. TEST Hypotheses through analytic studies and laboratory testing
5. SOLVE Point of contamination and original source of outbreak vehicle
6. CONTROL Outbreak through recalls, facility improvements, and industry collaboration
7. DECIDE An outbreak is over

If cases continue
- Not finding associations

If cases stop

[http://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/investigations/]
Measure of Association

Investigation process and interview content

<table>
<thead>
<tr>
<th></th>
<th>case</th>
<th>control</th>
</tr>
</thead>
<tbody>
<tr>
<td>exposed</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>not exposed</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

OR = \frac{AD}{CB}
### Specificity and Association

**Exposure:** tomatoes  
**Case:** *S. Montevideo*  
**Control:** not ill

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Not Exposed</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

**odds ratio = 1.6**
### Specificity and Association

**Exposure:** tomatoes  
**Case:** S. Montevideo *subt. A*  
**Control:** not ill

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Not Exposed</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

**odds ratio = 6.7**
Genomic Diversity

Ancestor

Evolutionary Time

= Mutational Event
Typing methods for outbreak detection and epidemiological surveillance

<table>
<thead>
<tr>
<th>Traditional</th>
<th>Molecular</th>
<th>Sequence-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serology</td>
<td>PFGE</td>
<td>Microarray</td>
</tr>
<tr>
<td>Phage</td>
<td>AFLP</td>
<td>Sanger</td>
</tr>
<tr>
<td>Biochemistry</td>
<td>RAPD</td>
<td>Next generation</td>
</tr>
<tr>
<td>Antimicrobial resistance</td>
<td>Rep-PCR</td>
<td>Third generation</td>
</tr>
<tr>
<td>DNA-DNA hybridization</td>
<td>MLVA (VNTR)</td>
<td></td>
</tr>
<tr>
<td>Enzyme electrophoresis</td>
<td>Optical map</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Microarray</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mass-spec</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(MaldiTOF)</td>
<td></td>
</tr>
</tbody>
</table>
Detecting Genomic Differences

<table>
<thead>
<tr>
<th>Mutational event</th>
<th>PFGE</th>
<th>MLVA, CRISPR</th>
<th>MLST</th>
<th>WGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertion / Deletion</td>
<td>X</td>
<td>O</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SNPs</td>
<td>O</td>
<td>O</td>
<td>X?</td>
<td>X</td>
</tr>
<tr>
<td>Recombination</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>
Next Gen Sequencing

Capillary electrophoresis (Sanger)
Between 96 and 384 samples
(76 - 308 Kb/run)

454
400,000 samples
(120Mb / run)

Solexa
32M samples
(2Gb / run)

Solid
40M samples
(2-6 Gb / run)

4GB – 20 hrs

1GB – 4.5hrs
Moore’s Law vs. Carlson’s Curve

Cost per Raw Megabase of DNA Sequence

- Moore’s Law

National Human Genome Research Institute

genome.gov/sequencingcosts
Bioinformatic Analyses

1. Whole Genome Sequencing
   - Whole Genome Sequence
   - Reference Comparison
   - De novo assembly
   - Reference Comparison
   - Initial Assembly
   - Ref. guided assembly
   - Refinement: Alignment/Comparison
   - Scaffolding
   - Finished Sequence Analysis

2. Sequence Alignment and Assembly
   - Alignment
   - High Quality Ref Genomes

3. Comparative Genomics
   - SNP and Indel Identification
   - Syntenic Analysis
   - Molecular Clock
   - Core (shared) Genome, gene homolog/family
   - Gene mutations, divergence, recombination
   - Metadata Analysis
   - Accessory Genome New gene/families/islands
   - Phenotype factors: Virulence, Clinical, Environmental
   - Phylogenetic Analysis
The promise of WGS for outbreak and traceback investigation
Traditional Foodborne Illness Investigation

- Contaminated food enters commerce
- Identify illnesses and get PFGE pattern from clinical samples
- Identify contaminated food and confirm that product or environmental sample PFGE pattern matches the clinical sample pattern
- Source of contamination identified too late to prevent most illnesses

Number of Cases

Days
Foodborne Illness Investigation Using Whole Genome Sequencing

FDA, CDC, FSIS, and States use WGS in real-time and in parallel on clinical, food, and environmental samples.

Source of contamination identified early through WGS combined database queries.

Contaminated food enters commerce

Averted Illnesses
### Example: *Listeria monocytogenes*, sprouts, 2014

<table>
<thead>
<tr>
<th>Date</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jun - Aug</td>
<td>CDC reports 5 listeriosis cases in 2 states</td>
</tr>
<tr>
<td>Aug – Sep</td>
<td>FDA conducts routine inspection in facility manufacturing sprouts and other soy products</td>
</tr>
<tr>
<td></td>
<td>28 samples are positive for <em>Listeria monocytogenes</em> (Lm)</td>
</tr>
<tr>
<td>Aug 28</td>
<td>Manufacturer agrees to product recall</td>
</tr>
<tr>
<td>Sep 25</td>
<td>WGS found to be highly related to 5 recent clinical samples and a 2013 clinical sample not in CDC definition</td>
</tr>
<tr>
<td>Oct 16</td>
<td>9 additional samples from facility test positive for Lm</td>
</tr>
<tr>
<td>Nov 3</td>
<td>WGS found to be highly related to clinical samples</td>
</tr>
<tr>
<td>Nov 7</td>
<td>Firm agrees to close facility and cease production of mung beak sprouts</td>
</tr>
<tr>
<td>WGS data characteristic</td>
<td>Advantage</td>
</tr>
<tr>
<td>-------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Highly robust (~1e3 SNPs or alleles over ~1e6 datapoints)</td>
<td>• Earlier and smaller cluster detection</td>
</tr>
<tr>
<td></td>
<td>• Enhanced opportunity to investigate “sporadic” case</td>
</tr>
<tr>
<td>More specific (&quot;endless&quot; number of types)</td>
<td>• Improved binning of outbreak related and unrelated</td>
</tr>
<tr>
<td></td>
<td>• Improved outbreak hypotheses</td>
</tr>
<tr>
<td>Evolutionary ordering of mutations (apply rigorous phylogenic analyses to data)</td>
<td>• Detect clusters by time and space</td>
</tr>
<tr>
<td>Digital</td>
<td>• Robust / automatic determination of type</td>
</tr>
<tr>
<td></td>
<td>• Data easily sharable</td>
</tr>
</tbody>
</table>
Listeria Outbreaks and Incidence, 1983-2014

Outbreaks per year:
- Pre-PulseNet: 0.3
- Early PulseNet: 2.3
- Listeria Initiative: 2.9
- WGS: 9

Median cases per outbreak:
- Pre-PulseNet: 69
- Early PulseNet: 11
- Listeria Initiative: 5.5
- WGS: 4.0

Data are preliminary and subject to change.
S. Bareilly, scrape tuna, 2012

Same PFGE but not part of the outbreak

Processed within 8 km of Moon Fishery

Outbreak Isolates 2-5 SNPs

Moon Fishery (India)
S. Bareilly, scrape tuna, 2012

West coast of India

S. Bareilly 189 (shrimp)

8 km

S. Bareilly isolated from tuna samples collected at Moon Fishery, India

Maryland & New York

Indonesia
use of WGS by public sector
AMD Initiative
‘Advanced Molecular Detection’

- 5 year budget initiative that started in fiscal year 2014
  - initial investment of $30 million; level funding requested for each of the remaining years

Transforming disease detection and response

www.cdc.gov/amd/index.html
CDC WGS Pilot Project -- *Listeria monocytogenes*

- Sequencing as close to real-time as possible in parallel with current surveillance
  - all clinical isolates referred from PulseNet laboratories
  - food and environmental isolates from FDA, USDA-FSIS, GenomeTrakr States

- Upload sequences to a public database NCBI (Genbank), a public database, as the sequences are generated
  - Sensitive patient metadata stored in non-public PulseNet database

- Evaluate data on a weekly basis
  - Sequence analysis and epi data (Listeria Initiative)

- Further investigation of PFGE and WGS clusters
Basic Data Flow for Global WGS Public Access Databases

**DATA ACQUISITION**
Sequence and upload genomic and geographic data

**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

**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
GenomeTrakr Strategy

- Sequencing responsibility is distributed
- Build shared database containing draft sequence and minimal metadata to enhance opportunity to find similars or matches
- Open source data and methods provide transparency
- Leverage NCBI assets and expertise to store and serve genomic data
- Publicly-accessible data spurs discovery and innovation
- Encourage partnerships, not firewalls
<table>
<thead>
<tr>
<th>WGS Use Case</th>
<th>Regulatory</th>
<th>Research</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illness Cluster Detection</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product traceback</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environmental assessment</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Source attribution</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Replace legacy workflows</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Define food adulterant</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infer origin and evolution of pathogenicity</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Determine basis for microbial growth and survival</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Characterize process microbes</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Discover new enzymes, mechanisms, pathways</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Outreach to industry
Appreciation and credit to:

• Drs. Peter Gerner-Smidt, Brendan Jackson and John Besser, CDC, Atlanta
• Dr. David Engelthaler, T-Gen North, Flagstaff
• Dr. Mohammed Koohmaraie, IEH, Lake Forest Park
• Drs. Ruth Timme, Marc Allard, and Eric Brown, FDA-CFSAN, College Park