Bringing Whole Genome Sequencing on Board in a State Regulatory Laboratory

Brian D. Saunders, PhD
NY State Dept. of Agriculture & Markets
Food Laboratory
The Food Laboratory!
Major laboratory sections:

- **Food and Dairy Microbiology**: analyze over 15,000 food samples/yr. ~ 2500 of these for foodborne pathogens and the remainder for other indicators.

- **Food and Dairy Chemistry**: analyze over 5000 food samples/yr. ~ 1500 for chemical hazards.

- **Pesticide residues**: analyze over 2500 produce and water samples/yr. for 165 different pesticide residues and pharmaceutical products.
Program Overview

- Primarily testing for specific health hazards and/or accuracy of labeling
- Focus on high risk, ready-to-eat foods and feeds
- Collaborate and coordinate work planning with other allied public health agencies (FDA, USDA, EPA, NYSDOH, etc)
Where do the samples come from?

NY State Dept. of Agriculture and Markets
- Collected by field staff from the following
  - farms
  - food processing plants
  - food/feed/fertilizer manufacturers
  - distribution centers
  - import terminals
  - farmers markets
  - retail stores

Federal agencies
- USDA
- FDA
Primary Pathogen Testing Capability

- *Salmonella enterica*
- *Listeria monocytogenes*
- *E. coli* O157:H7 and other STECs (top 6)
- *Campylobacter jejuni* and *coli*
- *Staphylococcus aureus* & enterotoxin
Methods and technology

• Regulatory action is not initiated until a culture confirmed isolate is generated

• Regulatory food safety labs are required to use specific methods and technologies to screen food for pathogens (FDA & ISO)

• Molecular subtyping is used on confirmed culture positive samples to inform public health surveillance for the detection of possible outbreaks (PulseNet/GenomeTrakr)
“The Usual Suspects”

http://biocomicals.blogspot.com
Transition from PFGE to WGS

- PFGE patterns submitted to PulseNet either by NYSDOH on our behalf (since 1998) or by our laboratory (since 2008)

- PFGE on all foodborne pathogens isolated by our lab

- Dropping Listeria PFGE by CDC impacted our program’s ability to identify if food isolates from our lab were related to ongoing human illness clusters
  - Most of our pathogen isolates from surveillance are Listeria

- Only two certified analysts for PFGE; one of those overlaps performing WGS as well
New York State Department of Agriculture and Markets
Whole Genome Sequencing Analysis Workflow

Isolate recovered by Food Lab

DNA Extraction (Manual or QIACUBE) ~3 hours

Library Prep/Tagmentation
Nextera XT V2 ~2 hours

Library Normalization/Pooling ~3 hours

Load Miseq Cartridge ~1.5 hours

Miseq Instrument Run ~40-50 hours

Basic QC Metrics
Upload and Share Run ~1-2 hours
Prepare and share Metadata

“Black Box” of Analysis

| NCBI Pathogen Detection (wgMLST) | Internal “Pipeline” (FDA/Cornell SNP) | CDC PulseNet (wgMLST) |
WGS Work Area

- Converted BSL3 space
- Limited bench, knee space
- Bench lighting is not ideal
- Most pipetting in Biosafety Cabinets
**DNA Extraction**

- **Manual extraction**
  - QIAGEN DNEasy Blood and Tissue Kit
  - Normally run 16 per batch
    - Same Number used for each Miseq run
  - Can run up to 24 (limited by size of microfuge)

- **Automated extraction**
  - QIAGEN QIACUBE
  - Can only run 12 isolates at one time in the instrument
  - Some labs report cross contamination
  - Most labs include negative control; practically only 11/run
  - Working on validation for routine use
  - May provide higher quality DNA
  - Hands on time about the same as manual extraction kit
  - At least two runs for a full Miseq run
Library Prep

• Currently using Nextera XT V2 chemistry

• Followed as described in FDA/CDC harmonized protocol

• Use the indices until exhausted (usually goes beyond 5 freeze/thaws)

• How might adapt workflow to V3 sample volumes?
  • i.e. going from 16 to 30+ samples in a single prep
Library Normalization & Pooling

- Performed as described in CDC/FDA harmonized SOP

- Have used AMP-pure beads past expiration with success
  - Want to look at impact read-length

- Find using a smaller volume 96-well plate for NTA is much easier
Miseq Runs

• Runs take approximately 40-50 hours depending on the organism(s) loaded.

• Try to run over weekends, so that instrument is free for other analysis, maintenance or requests.

• 2\textsuperscript{nd} instrument? Cost? Additional maintenance? When does workload, need for redundancy justify?
QC Metrics

• Record initial metrics from end of the run screen as indicated in the CDC/FDA harmonized workbook

• Cut and paste run data from BaseSpace for Coverage Calculations

• As long as metrics meet minimum requirements, data are shared from BaseSpace and metadata sent to FDA to upload to NCBI
## QC Metrics - Workbook

### Run Name
- NYAG: M1082911 190109

### MiSeq Run Start Date
- 1/9/19

### MiSeq Reagent Kit
- 500 cycle V2 kit

### Post-Run Metrics:

<table>
<thead>
<tr>
<th></th>
<th>Q30</th>
<th>Cluster Density (K/mm²)</th>
<th>Clusters Passing Filter</th>
<th>Estimated Yield (Gb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC</td>
<td>&gt; 75%</td>
<td>600 - 1300</td>
<td>&gt; 80%</td>
<td>7.5 - 8.5</td>
</tr>
<tr>
<td>Actual Values</td>
<td>9.0G 82.5%</td>
<td>1312 K/mm²</td>
<td>88%</td>
<td>11249.4 MB</td>
</tr>
</tbody>
</table>

### Minimum Coverage:

<table>
<thead>
<tr>
<th>Escherichia/Shigella</th>
<th>Salmonella</th>
<th>Esteria</th>
<th>Campyloabacter</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC</td>
<td>48X</td>
<td>30X</td>
<td>20X</td>
</tr>
</tbody>
</table>

### Index Table

<table>
<thead>
<tr>
<th>Index Number</th>
<th>Sample ID</th>
<th>Project</th>
<th>Index 1 (I7)</th>
<th>Index 2 (I5)</th>
<th>% Reads Identified (PF)</th>
<th>PF Reads (Paste Top Row ONLY)</th>
<th>Total # of Reads (Calculated-Not from BaseSpace)</th>
<th>Coverage</th>
<th>Genome Size</th>
<th>Read Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>06805246C-1</td>
<td>default</td>
<td>TAAGGCGA</td>
<td>CCTAGAGT</td>
<td>3.5722</td>
<td>43,517,954</td>
<td>1554548</td>
<td>129.55</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>2</td>
<td>06805246F-6</td>
<td>default</td>
<td>CGTACTAG</td>
<td>CCTAGAGT</td>
<td>8.0213</td>
<td>43,517,954</td>
<td>3490706</td>
<td>290.89</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>3</td>
<td>06805246G-1</td>
<td>default</td>
<td>AGGCGAGA</td>
<td>CCTAGAGT</td>
<td>6.0703</td>
<td>43517954</td>
<td>2641670</td>
<td>220.14</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>4</td>
<td>068056111-1</td>
<td>default</td>
<td>TCCTGAGC</td>
<td>CCTAGAGT</td>
<td>4.0804</td>
<td>43517954</td>
<td>1775707</td>
<td>117.98</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>5</td>
<td>06805669C-1</td>
<td>default</td>
<td>GGACTCTCT</td>
<td>CCTAGAGT</td>
<td>6.6576</td>
<td>43517954</td>
<td>2897251</td>
<td>241.44</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>6</td>
<td>06805669D-1</td>
<td>default</td>
<td>TAGGCGAG</td>
<td>CCTAGAGT</td>
<td>5.2189</td>
<td>43517954</td>
<td>2271159</td>
<td>189.26</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>7</td>
<td>06805669H-1</td>
<td>default</td>
<td>CTCTCTAC</td>
<td>CCTAGAGT</td>
<td>6.8711</td>
<td>43517954</td>
<td>2990162</td>
<td>249.18</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>8</td>
<td>06805814A-1</td>
<td>default</td>
<td>CGAGCGCTG</td>
<td>CCTAGAGT</td>
<td>5.2689</td>
<td>43517954</td>
<td>2292917</td>
<td>191.08</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>9</td>
<td>06805815C-1</td>
<td>default</td>
<td>TAAGGCCGA</td>
<td>GCGTAAAGA</td>
<td>8.329</td>
<td>43517954</td>
<td>3624610</td>
<td>302.05</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>10</td>
<td>06805815D-1</td>
<td>default</td>
<td>CGTACTAG</td>
<td>GCGTAAAGA</td>
<td>5.2627</td>
<td>43517954</td>
<td>2292019</td>
<td>190.85</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>11</td>
<td>06805815E-1</td>
<td>default</td>
<td>AGGCAGAA</td>
<td>GCGTAAAGA</td>
<td>7.5894</td>
<td>43517954</td>
<td>3302752</td>
<td>275.23</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>12</td>
<td>06805815F-1</td>
<td>default</td>
<td>TCTCTGAC</td>
<td>GCGTAAAGA</td>
<td>7.1304</td>
<td>43517954</td>
<td>3103004</td>
<td>258.58</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>13</td>
<td>06805815G-1</td>
<td>default</td>
<td>GGACTCTCT</td>
<td>GCGTAAAGA</td>
<td>7.2393</td>
<td>43517954</td>
<td>3150395</td>
<td>262.53</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>14</td>
<td>06805880H-1</td>
<td>default</td>
<td>TAGGCGAG</td>
<td>GCGTAAAGA</td>
<td>5.2795</td>
<td>43517954</td>
<td>2297530</td>
<td>191.46</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>15</td>
<td>06805948F-1</td>
<td>default</td>
<td>CTCTCTAC</td>
<td>GCGTAAAGA</td>
<td>4.3744</td>
<td>43517954</td>
<td>1903649</td>
<td>158.64</td>
<td>3000000</td>
<td>250</td>
</tr>
<tr>
<td>16</td>
<td>06805948I-1</td>
<td>default</td>
<td>CGAGCGCTG</td>
<td>GCGTAAAGA</td>
<td>6.3157</td>
<td>43517954</td>
<td>2748463</td>
<td>229.04</td>
<td>3000000</td>
<td>250</td>
</tr>
</tbody>
</table>
New York State Department of Agriculture and Markets
Whole Genome Sequencing Analysis Workflow

Isolate recovered by Food Lab

DNA Extraction (Manual or QIACUBE) ~3 hours

Library Prep/Tagmentation Nextera XT V2 ~2 hours

Library Normalization/Pooling ~3 hours

Load Miseq Cartridge ~1.5 hours

Miseq Instrument Run ~40-50 hours

Basic QC Metrics

Upload and Share Run ~1-2 hours

Prepare and share Metadata

“Black Box” of Analysis

<table>
<thead>
<tr>
<th>NCBI Pathogen Detection (wgMLST)</th>
<th>Internal “Pipeline” (FDA/Cornell SNP)</th>
<th>CDC PulseNet (wgMLST)</th>
</tr>
</thead>
</table>
Data “Pipeline”
Black Box

- NCBI - Originally SNPs, now wgMLST
- FDA - SNP
- Internal Cornell - SNP
- CDC PulseNET - wgMLST

"We are bioinformaticians, that's what we do"
Data Pipelines

• Will this ever be an ecosystem of standardized analysis? Can it be?

• Will all labs submitting into GenomeTrakr also submit to PulseNet?

• Will NCBI be the only place where all public health data can be compared easily?
Other sources of Isolates

- USDA
- Ecuador
- Academia?
Summary

• Bringing WGS online in a state regulatory agency brings next generation technology to an important input of public health surveillance for foodborne pathogens

• One challenge is to find the right mix of sequencing isolates in real-time and limiting the cost impacts of sequencing partial runs by including sequencing historical isolates

• Another regulatory challenge is that there are now multiple streams of analysis that yield similar “yet not identical results”
As I suspected, you’re full of bacteria. We’re going to have to throw you away.