Whole Genome Sequencing for National Surveillance of Enteric Pathogens

William Wolfgang Ph.D.
Wadsworth Center / NYSDOH
Division of Infectious Diseases
william.wolfgang@health.ny.gov

NEBASM Symposium 11/7/19
WGS for Enteric pathogen surveillance

Wadsworth: 2012 acquires its first bench top sequencer
  – pilot studies with *Salmonella* Enteritidis

FDA: 2013 GenomeTrackr initiative.
  – Surveillance of *Environmental* pathogens

CDC: 2014 Advanced Molecular Detection initiative.
  – Surveillance of *Clinical* pathogens

NCBI: creates public databases to hold NGS data.
  – Pathogen Detection Portal
For *Salmonella* Enteritidis (SE)
Outbreak clusters are hard to detect using PFGE

- 50% of the isolates we receive have the same PFGE DNA fingerprint.
- And 2/3 have a very common PFGE DNA fingerprint.
- These **Endemic** types are of limited use to our epidemiologists.
SNP and wgMLST analysis

**SNP**

<table>
<thead>
<tr>
<th>SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTCATAGCATTATTATTATTACGGACTA</td>
</tr>
<tr>
<td>CAGTATCGATAATAATAATAAGTCCGTAT</td>
</tr>
<tr>
<td>1bp</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>30</td>
</tr>
</tbody>
</table>

**wgMLST**

<table>
<thead>
<tr>
<th>wgMLST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain A</td>
</tr>
<tr>
<td>Strain B</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Applied Maths</td>
</tr>
</tbody>
</table>

Compare single nucleotides (SNPs)
- Highest resolution
- Typing nomenclature is not possible

Compare gene by gene (alleles)
- Lower resolution
- But typing nomenclature is possible
Salmonella phylogenetic tree

- Patient isolates are on right
- Branch points indicate putative common ancestor.
- Sum of Horizontal lines measure genetic closeness.

Tree reveals
- Genetic closeness
- Ancestral relationships
Prison outbreaks in Virginia and New York

- **5/16-Virginia** reports an SE outbreak associated with a correctional facility.

- At the same time **NYS** is investigating a SE outbreak also associated with a correctional facility.

- Both have the same PFGE pattern: JEGX01.0021

Q. Could they be from a common source?
Prison associated cluster

Pattern 21

- Clusters were distinct.
- Suggest different source for each outbreak.
Cluster Thresholds that Trigger a Report to NYS Epidemiologists

<table>
<thead>
<tr>
<th>Organism</th>
<th>Min # isolates</th>
<th>Timeframe</th>
<th>Alleles or SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listeria monocytogenes</td>
<td>2</td>
<td>Indefinite</td>
<td>20</td>
</tr>
<tr>
<td>E. coli</td>
<td>2</td>
<td>1 year</td>
<td>10</td>
</tr>
<tr>
<td>Salmonella other serovars</td>
<td>2</td>
<td>60 days</td>
<td>10</td>
</tr>
<tr>
<td>S. Enteritidis, Typhimurium &amp; Newport</td>
<td>3</td>
<td>60 days</td>
<td>5</td>
</tr>
</tbody>
</table>

Thresholds as of Oct 2019
We have made 6 revisions since May 2019
3 case studies

- *Listeria monocytogenes* – food preparation facility
- *E. coli* O103 – NYS geographical cluster
- *Salmonella* Enteritidis – The cluster that never ends
Intermittent *Listeria* outbreak

From June. 2014 to Oct. 2017

- 13 isolates with matching PFGE were detected
- NY residents were from the same counties
- Based on 120 day time frame- treated as 3 separate clusters
- But should they really be considered as one cluster?

Cluster A  
6/2014 – 10/2014

Cluster B  
6/2016 – 1/2017 *Out-of-state isolates*

Cluster C  
6/2017

NYS Environmental sampling 10/10/17

* Out-of-state isolates
wgMLST analysis shows all are 0 to 6 alleles apart

- The 2 out-of-state isolates did not match the cluster.
- The 4 environmental samples were also closely related.
Surveillance and environmental sampling continued

- In 2018 and 2019
  - 2 patient
  - 6 environmental

- Remediation of the facility was undertaken

- After March 2019 no more positive environmental or clinical specimens

- But were there out of state isolates?

<table>
<thead>
<tr>
<th>WGS ID</th>
<th>Isolation Date</th>
<th>Source Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNUSAL004689</td>
<td>2018-12-30</td>
<td>Human</td>
</tr>
<tr>
<td>NY99570229</td>
<td>2019-02-11</td>
<td>Environmental</td>
</tr>
<tr>
<td>NY99541573</td>
<td>2019-02-11</td>
<td>Environmental</td>
</tr>
<tr>
<td>NY99541318</td>
<td>2019-02-11</td>
<td>Environmental</td>
</tr>
<tr>
<td>NY99293701</td>
<td>2019-01-29</td>
<td>Environmental</td>
</tr>
<tr>
<td>PNUSAL000806</td>
<td>2014-06-02</td>
<td>Human</td>
</tr>
<tr>
<td>PNUSAL000888</td>
<td>2014-06-17</td>
<td>Human</td>
</tr>
<tr>
<td>PNUSAL001130</td>
<td>2014-10-19</td>
<td>Human</td>
</tr>
<tr>
<td>PNUSAL002709</td>
<td>2016-11-04</td>
<td>Human</td>
</tr>
<tr>
<td>PNUSAL003099</td>
<td>2017-05-30</td>
<td>Human</td>
</tr>
<tr>
<td>NY81406976</td>
<td>2017-10-10</td>
<td>Environmental</td>
</tr>
<tr>
<td>PNUSAL002802</td>
<td>2017-01-04</td>
<td>Human</td>
</tr>
<tr>
<td>NY81406936</td>
<td>2017-10-10</td>
<td>Environmental</td>
</tr>
<tr>
<td>NY100638542</td>
<td>2019-03-11</td>
<td>Environmental</td>
</tr>
<tr>
<td>PNUSAL002278</td>
<td>2016-06-23</td>
<td>Human</td>
</tr>
<tr>
<td>NY81406716</td>
<td>2017-10-10</td>
<td>Environmental</td>
</tr>
<tr>
<td>NY81406896</td>
<td>2017-10-10</td>
<td>Environmental</td>
</tr>
<tr>
<td>PNUSAL004103</td>
<td>2018-06-27</td>
<td>Human</td>
</tr>
<tr>
<td>NY102159252</td>
<td>2019-04-17</td>
<td>Environmental</td>
</tr>
</tbody>
</table>
NCBI Pathogen Detection

- All Genometrakr and PulseNet samples end up here.

- NCBI builds trees daily for Enteric pathogens.
  - 29,000 $Lm$ samples
  - 2,548 $Lm$ trees

- Easily accessed through a web portal.

- Data is publicly available.
NCBI Pathogen Detection shows no closely related out of state isolates

- All isolates fell into a single NCBI tree.
- No other isolates on the tree.
- 0 to 13 SNPs

And so were not closely related to 29,000 other *Lm*
So how is this working with *Listeria monocytogenes*?

**Pretty Good**

Improved resolution of WGS allowed:

- Identification of a long term source of *Lm* contamination
- Exclusion of out of state samples
- Able to accurately track remediation
- No additional isolates were detected nationally

- Supports allele range of 0-20
- Supports time frame of forever
NYS *E. Coli* O103 cluster

From July 11 to Oct 7, 2019

- 20 isolates within a single genomic cluster (0-20 alleles) were detected.
- No PFGE was done

- Epidemiology supported two origins- Western NY and the Metro area.
- Some Metro isolates were associated with kosher food consumption.
cgMLST allele analysis

- 0 – 20 alleles

WNY
- One sub-cluster with 1-8 alleles; associated with kosher food consumption

1-8 alleles
- cgMLST did not support geographic clustering

Metro
SNP Analysis is similar but different

- 2-61 SNP diversity
- One sub-cluster with 3-15 SNP diversity
- SNP analysis supported geographic clustering
NCBI Pathogen detection

- 2 - 51 SNPs overall
- One sub-cluster 2-15 SNPs
- SNP analysis supported geographic clustering
- WNY interspersed among many other isolates
### Comparing Allele vs SNP trees

<table>
<thead>
<tr>
<th>Type of tree</th>
<th>Full diversity</th>
<th>Sub-cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGMLST</td>
<td>0-20 alleles</td>
<td>1-8 alleles</td>
</tr>
<tr>
<td>FDA SNP</td>
<td>2-61 SNPs</td>
<td>3-15 SNPs</td>
</tr>
<tr>
<td>NCBI SNP</td>
<td>2-51 SNPs</td>
<td>2-15 SNPs</td>
</tr>
</tbody>
</table>

- Alleles underestimate full diversity.
- Structure of the allele tree less concordant with epi. data.
- When resolution is needed SNP trees should be built.
So how is this working with STEC?

Pretty good but
In this case cgMLST was discordant with geography
  • SNP tree was concordant and had higher resolution

Demonstrated need to rethink cluster reporting thresholds for STECs
  • reduce allele or SNP diversity
  • shorten timeframe
The *Salmonella* Enteritidis cluster that won’t stop

From 4/30/19 to 10/08/19 we sent 8 separate reports tracking a single cluster to our epidemiologists.

- **The cluster** eventually contained 84 patient samples.
- With no strong epidemiological links.
- 0-11 allele diversity.  
  - Why is the cluster allele diversity greater than 5?
The problem of chaining

- New samples are received within 60 days that are within 5 alleles to at least one other sample.
- Clustering criteria are inadequate.
- Overlaying epi. data was not helpful.
And worse at NCBI

- 353 samples within 0 SNPs of this cluster.
- 1645 within 5 SNPs.
- 2375 within 15 SNPs.
- And these numbers are constantly increasing
So how is this working with SE?

Not so well for common S. Enteritidis types

Improved resolution was not helpful
- Lab results can not inform epi.
- Yet a huge amount of work.

Solutions for *Salmonella*
- Consider other parts of the genome.
- Try to identify these common types prior to analysis.
- then do not analyze unless requested
Where do we stand now?

WGS does improve surveillance.
- More clusters will be identified than with PFGE.
- More sources will be identified.
- And cases of foodborne illness should decline.

But challenges remain
- Too many clusters.
  - Need to prioritize, but how?
  - Refine cluster definitions.
- Identification of endemic clusters.
Acknowledgements

Wadsworth Center
• Kim Musser
• Samantha Wirth
• Lisa Thompson
• Lisa Mingle
• Nellie Dumas
• Michelle Dickinson
• Deborah Baker
• Sai Gubbala

• Sequencing core
• Bioinformatics core

BCDC
• Amy Robbins
• Kate Purcell
• Alexandra Newman
• Paula Huth
• Madhu Anand
• Kadri Ajileye

BCEHFP
• David Nicholas

FDA
• Eric Brown
• Marc Allard
• Errol Strain
• Ruth Timme

CDC
• Eija Trees
• Kelley Hise
• Steven Stroika