PulseNet Updates: Transitioning to WGS for Reference Testing and Surveillance

GenomeTrakr Meeting, 2018

Steven Stroika, PulseNet CDC

National Center for Emerging and Zoonotic Infectious Diseases
Office of Advanced Molecular Detection
Enteric Diseases Laboratory Branch
Peter Gerner-Smidt, MD, ScD, Branch Chief
John Besser, PhD, Deputy Branch Chief
Sherricka Stephens, MPA, Branch Manager
Alex Mercante, PhD, Branch Quality Manager
Devon Stoneburg, BSAB, Branch Quality Manager

PulseNet Reference Outbreak Surveillance Team (PROS)
Efrain Ribot, PhD
Nancy Strockbine, PhD (Deputy)

National Botulism and Enteric Toxins Team (NBET)
Carolina Lüquez, PhD

National Antimicrobial Resistance Surveillance Team (NARST)
Jean Whichard, DVM, PhD

Bioinformatics and Metagenomics Team (BioMe)
Heather Carleton, PhD, MPH

Pseudomonas and Other Enterobacteriaceae Unit
Michael Plocharski, PhD

Campylobacter and Helicobacter Unit
Rachael Aubert, PhD

Escherichia and Shigella, Unit
Nancy Strockbine, PhD

Salmonella Unit
*Patricia Fields, PhD

Listeria, Yersinia, Vibrio, and other Enterobacteriaceae Unit
Cheryl Tarr, PhD

PulseNet Outbreak Detection and Surveillance Unit
Kelley Hise, MPH

PulseNet Next Generation Subtyping Methods Unit
Eija Trees, PhD

PulseNet PFGE Reference Unit
Molly Freeman, PhD

*Patricia Fields Ph.D. is also the Associate Director for Laboratory Science in DFWED

NARST Surveillance Unit
Hayat Caidi, PhD

NARST Applied Research Unit
Jason Folster, PhD

May 2018
PulseNet Reference Outbreak Surveillance Team (PROS)

- PulseNet Outbreak Detection and Surveillance Unit
- PulseNet Next Generation Subtyping Methods Unit
- PulseNet PFGE Reference Unit
- Campylobacter and Helicobacter Unit
- Escherichia and Shigella, Unit
- Salmonella Unit
- Listeria, Yersinia, Vibrio, and other Enterobacteriaceae Unit
EDLB Vision

REPLACE all of these workflows:
• Identification - develop agnostic approach
• Serotyping
• Virulence profiling
• Antimicrobial susceptibility
• Subtyping for outbreak investigations

With ONE cost-efficient and precise method: All of this information can be derived from the genome sequence
PulseNet Transition Timeline to WGS Surveillance

- **NOW:** prepare the future workflow in your lab without using traditional reference and subtyping methods (e.g. serotyping, ID confirmation, virulence and resistance testing)
- **October 1, 2018:** stop pulsing *Campylobacter*
- **November 2, 2018:** labs begin transitioning to BioNumerics v7
  - Instructions and further guidance will be forthcoming
- **January 15, 2019:** stop pulsing all organisms
  - Conduct WGS on as many isolates as funds permit using the following priority schedule: (1) *Listeria monocytogenes*, (2) STEC, (3) *Salmonella*, (4) Other species
- **WGS and PFGE Prioritization until routine PFGE stops**
  - *Salmonella* and STEC: pulse all isolates (only 1st enzyme for *Salmonella*) and sequence all isolates if possible; if not prioritize isolates with cluster codes
    - **If the isolate has an outbreak code assigned in the PFGE national database, please submit a request for a WGS ID and sequence those isolates**
  - *Listeria*: sequence all isolates
  - *Campylobacter* and *Shigella*: sequence all isolates, but prioritize other organisms first
- CaliciNet is not transitioning to BioNumerics 7 at this time. Because BioNumerics licenses are backwards compatible, you will be able to access your CaliciNet databases in version 6 as well as your PulseNet databases in version 7. You will need to change your home directory depending on which version you are using. Please email pulsenet@cdc.gov if you have questions about this.
- **Starting January 1st TAT will be calculated from the date the isolate was received (or recovered) in the PHL to the date of upload to the national database**
  - Should be 7 working days or less
What will the analysis workflow look like?
1. Move sequence data to local storage
2. Link sequence data to RefID Database
3. Submit raw reads, and retrieve assembly with basic QC metrics
4. Submit sequence data for taxonomic identification
5. Export de novo assemblies, QC metrics, taxa ID to correct organism-specific database
6. Submit sequence data for annotation, and retrieve allele calls
7. Upload allele calls and metadata
8. Download allele code, outbreak code, etc.
9. Upload raw sequence data with minimal metadata

Data Analysis Workflow with National Database:

- Raw Sequence Data
- Private Raw Sequence Storage
- Reference ID Database
- Organism-specific Database
- Calculation Engine
- Allele Databases
- PulseNet National Databases
- National Center for Biotechnology Information (NCBI)
Reference Identification Database (RefID)

Raw reads, QC, read quality, predicted coverage, contamination

de novo Assembly, QC, N50, genome size and coverage

Average Nucleotide Identity (ANI) – For 18 species

Species specific databases

Add to Listeria database for further characterization

Serotyping, MLST, virulence genes, AST, etc. (depending on organism)
## Reference ID Database: Identification

<table>
<thead>
<tr>
<th>Genera</th>
<th>Species</th>
<th>ANI value (%)</th>
<th>Genome size (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Campylobacter</strong></td>
<td>coli fetus jejuni lari upsaliensis</td>
<td>≥92</td>
<td>1.4-2.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Escherichia</strong></td>
<td>albertii coli and Shigella fergusonii*</td>
<td>≥95</td>
<td>4.5-5.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Listeria</strong></td>
<td>innocua* ivanovii* marthii* monocytogenes seeligeri*</td>
<td>≥92</td>
<td>2.7-3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Salmonella</strong></td>
<td>bongori enterica</td>
<td>≥93</td>
<td>4.5-5.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Vibrio</strong></td>
<td>cholerae parahaemolyticus vulnificus</td>
<td>≥95</td>
<td>4.0-5.0</td>
</tr>
</tbody>
</table>

*will not set to map for organism specific database export*
How does Reference ID look in BioNumerics?
Using the BioNumerics RefID Database: Linking Sequence Data

- After export and mapping settings are complete, link sequence data.
- To link sequence data to your database entries, click the Import icon.
- Expand the options for Sequence read sets data.
- Select Import sequence read sets data as links.
- Select Import.
Submission to the Calculation Engine: Linking Sequence Data

- You can link to SRA via SRR id’s on NCBI/EMBL-EBI for analysis of publically available strains and submit directly to the calculation engine.
- If your sequence data is linked on BaseSpace, you can submit to the calculation engine directly after establishing link to your BaseSpace account.
- If sequence data is linked locally (desktop, local or networked file share) you can link and submit those files through the CEStore Uploader.
Your database entries should be selected and your wgs link present in the database (green dots visible for wgs experiment indicate linked data).

You are now ready to submit entries to the calculation engine (CE) for denovo assembly.

Authenticate to the PulseNetWGS Firewall.

Select the Submit jobs icon and only submit the denovo job, click OK.
When the denovo job finishes in the calculation engine, you can retrieve results by clicking the icon.

Now that you have an assembly for the genome, you can resubmit to the calculation engine for ANI results (Genus and Species determination) and retrieve those results when finished.
When the ANI job finishes in the calculation engine, you can retrieve results by clicking the icon. You will see Genus and Species populated in the RefID database. You can now export this data by going to File → Transfer Current Selection. You can transfer at one time any Genus/Species results you have; this does not have to be done by organism.
Importing from Ref ID Database to Organism database

- Click “File” tab
- Click “Check for data transfer updates…”
- A progress bar indicating files being checked will appear
Importing from Ref ID Database (continued)

- An “Updates and additions” window will appear.
  - If new entries are being imported into database, that number will be “Created.”
  - If entries are already in the database, then those entries will be “Updated.”
  - MAKE SURE THESE NUMBERS ARE CORRECT

- Another progress bar appears after you hit “OK”

- A “Finished import” window will appear indicating success. Click “OK”
Organism-Specific ANI Results and Genotyping
Organism-Specific ANI Results

Additional information provided beyond genus/species when run in organism-specific databases:

- Lineage (*Listeria*)
- Subtype (*Salmonella*)
- Subspecies (*Campylobacter*)
Click the “Submit jobs to calculation engine” icon to open the Submit Jobs window.

Click “Settings” next to Genotyping to see the organism-specific tools and settings.

Note: Default settings are provided by CDC and should not be changed.
Genotyping Tools

- **Center for Genomic Epidemiology Tools**
  - Online tools hosted by CGE ([http://genomicepidemiology.org/](http://genomicepidemiology.org/)) and built into BioNumerics. Use BLAST to compare sequences to reference genomes:
    - Serotype (Serotype Finder (*Escherichia*) and SeqSero (*Salmonella*))
      - SeqSero is a curated version incorporating versions 1 and 2 with identifying 100 serotypes and antigenic formulas when serotype is not assigned
    - Resistance genes (ResFinder, PointFinder in all organism databases)
    - Virulence genes (parts of VirulenceFinder (*Escherichia*))
    - Plasmids (PlasmidFinder)
  - **Pathotype (*Escherichia*)**
    - Results (STEC, ETEC, etc.) based on presence/absence of specific genetic markers
  - **In silico PCR (all organism databases)**
    - Sequences are searched for particular markers using PCR primer pairs
Genotyping Tools: View Results in the Main Window

Results output into database fields can be viewed in the main window:

- Predicted serotype (Serotype_wgs)
- Predicted pathotype (Pathotype)

<table>
<thead>
<tr>
<th>Serotype</th>
<th>Predicted serotype</th>
<th>Predicted pathotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli 0118:H16</td>
<td>O118/O151:H16</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 026:H11</td>
<td>O26:H11</td>
<td>STEC</td>
</tr>
<tr>
<td>Shigella flexneri</td>
<td>O135/O13:H14</td>
<td>EEC/Shigella</td>
</tr>
<tr>
<td>E. coli 0118:H16</td>
<td>O118/O151:H16</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 0103:H11</td>
<td>O103:H11</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 091:NM</td>
<td>O91:H14</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 0157:H7</td>
<td>O157:H7</td>
<td>STEC</td>
</tr>
<tr>
<td>Shigella sonnei (Subg.)</td>
<td>O121:HNT</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 0121:HNT</td>
<td>O121:H19</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 091:NM</td>
<td>O91:H14</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 0157:H7</td>
<td>O157:H7</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 026:H11</td>
<td>O26:H11</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 0157:H7</td>
<td>O157:H7</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 026:H11</td>
<td>O26:H11</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 0157:H7</td>
<td>O157:H7</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 018:H16</td>
<td>O118/H16</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 0157:H7</td>
<td>O157:H7</td>
<td>STEC</td>
</tr>
<tr>
<td>Shigella sonnei (Subg.)</td>
<td>O121:HNT</td>
<td>STEC</td>
</tr>
<tr>
<td>E. coli 0121:HNT</td>
<td>O121:H19</td>
<td>STEC</td>
</tr>
<tr>
<td>Shigella flexneri</td>
<td>O135/O13:H14</td>
<td>EEC/Shigella</td>
</tr>
<tr>
<td>E. coli 045:Undetermined</td>
<td>045:H2</td>
<td>STEC</td>
</tr>
</tbody>
</table>
Genotyping Tools: View Results in a Comparison

- Results associated with experiments can be viewed in a comparison:
  - Resistance
  - Virulence
  - Plasmids
- Select entries and create a comparison
- Click the eyeball next to a genotyping experiment to view results in the Experiment Data window
Genotyping Tools: View Results in an Experiment Card

- Results associated with experiments can also be viewed in an experiment card:
  - Resistance
  - Virulence
  - Plasmids
- Click the green dot in the main window for a genotyping experiment
- An experiment card will appear with output data
BioNumerics Analyses and National Databases
Allele databases and Analysis

- Developed allele databases:
  - *Escherichia*: wgMLST, cgMLST, Chromosome-Associated, Plasmid-Associated, MLST (Achtman, Pasteur, Whittam)
  - *Campylobacter*: wgMLST, cgMLST (jejuni/coli), MLST (species specific)
  - *Listeria*: wgMLST, cgMLST, MLST
  - *Salmonella*: wgMLST, cgMLST (Enterobase), MLST (Achtman)

- For outbreak detection: cgMLST (*Salmonella* and *Escherichia*); wgMLST (*Listeria* and *Campylobacter*)

- Allele database in development for *Vibrio*

- wgSNP analysis available in BioNumerics v7.6 (pipeline most similar to CFSAN)

- Nomenclature development upon finalized version of cgMLST schemes
Partial Names

When sequences have partial names, it means they are *singleton* in clusters below their last digit.

The sequences above are approximately within 36 and 19 alleles of each other.
Future of the National Databases

- **Tenative November 5, 2018:** PHLs begin converting to BioNumerics v7.6
  - Database clean-up completed (PFGE databases)
  - Procedures will be piloted with a few labs prior to instructions being sent out
  - Possible webinar between November-December
  - BioNumerics trainings offered at PulseNet Regional Meetings and CDC, May 2019
  - Get analysis certified for WGS (likely available next month)
  - Sequence all PulseNet organisms in as real-time as possible and upload to the combined national databases
  - Testing combined national databases
  - Validating allele databases
  - Validating Reference ID database and Genotyping tools

- **Tenative Transition to use of wg/cgMLST by January 15, 2019**
  - All PHL’s migrated to BioNumerics v7.6
  - Testing and validating WGS nomenclature (Allele Codes)
  - Automatic WGS id assignments
Acknowledgements

Center for Genomic Epidemiology, DTU
University of Oxford, M. Maiden, K. Jolly
Public Health England, T. Dallman

Disclaimers:
“The findings and conclusions in this presentation are those of the author and do not necessarily represent the official position of the Centers for Disease Control and Prevention”

“Use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention or by the U.S. Department of Health and Human Services.”
The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.