WGS-based pathogen surveillance: international collaboration

Codex Committee on Food Hygiene (CCFH47)
November 2015

Ruth E. Timme, PhD
Research Microbiologist
GenomeTrakr data coordinator
International outreach:

• **Global Microbial Identifier (GMI)**
  - Data and analytical cooperation

• **WHO/PAHO**
  - Framework for supporting new labs
Global Microbial Identifier

What is to be gained

GMI envisions a global system of DNA genome databases for microbial and infectious disease identification and diagnostics. Such a system will benefit those tackling individual problems at the frontline, clinicians, veterinarians, etc., as well as policy-makers, regulators, and industry. By enabling access to this global resource, a professional response on health threats will be within reach of all countries with basic laboratory infrastructure.

NEWSLETTERS FROM THE GMI STEERING COMMITTEE

- 06 October 2015
  ASM Conference on Rapid Next-Generation Sequencing and Bioinformatic Pipelines for Enhanced Molecular...

- 24 September 2015
  Announcing the 9th Meeting on Global Microbial Identifier (GMI) in Rome, Italy

- 07 August 2015
  GMI Proficiency Test 2015 - sign-up is now closed

- 01 June 2015
  GMI Proficiency Test 2015 – sign-up is now open

BECOME A MEMBER OF GMI

To ensure involvement and progress in the GMI initiative, we ask everybody who wishes to participate as an active member in one or more working groups to accept the charter and sign up for membership.

GMI Proficiency Test 2015

The GMI Working Group 4 has launched the 2015 proficiency test on next-generation sequencing, focusing on analysis of Salmonella, Escherichia coli and Staphylococcus aureus.
Global Microbial Identifier

- Make novel genomic technologies and informatics tools available for improved global patient diagnostics, surveillance, research and public health response. Develop a global system to aggregate, share, mine and use microbiological genomic data to address global public health and clinical challenges, a high impact area in need of focused effort. 500 members in 30 countries

http://www.g-m-i.org/

**Work groups**

1. Political challenges, outreach and building a global network
2. Repository and storage of sequence and meta-data
3. Analytical approaches
4. Ring trials and quality assurance
5. Pilot project
Background

WGS Standards and Analysis workgroup

• Participants – FDA, CDC, NCBI and FSIS
• Developing a uniform approach for
  – Analysis comparison and standardization of results
  – Sample inclusion standards

Related Efforts

• Genome Comparison and Analytic Testing website
  – http://www.bioplanet.com/gcat
• Zook et. al. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology, 32, 246–251
• NIST. Reference sequencing material
GMI Workgroup 3 goals

• Produce a set of benchmark datasets
• Store them in an easily accessible public location.
• set up framework for the expansion of benchmark datasets from other sources.
Benchmark Datasets are Critical

Benchmarks allow:

• Evaluation of performance of methods
• Guidance on improvement of methods
• Standards for assessing reproducibility of results
• Bolster use of results for regulatory action
Benchmark Outbreak Datasets

US Federal Agency collaboration:
FDA, CDC, NCBI, USDA

<table>
<thead>
<tr>
<th>Species</th>
<th>Isolates</th>
<th>Outgroups</th>
<th>Reference</th>
<th>Outbreak</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td>10</td>
<td>6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><em>Listeria monocytogenes</em></td>
<td>31</td>
<td>3</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td><em>Salmonella enterica</em></td>
<td>24</td>
<td>4</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>biosample_acc</td>
<td>strain</td>
<td>genBankAssembly</td>
<td>SRArun_acc</td>
<td>outbreak</td>
</tr>
<tr>
<td>---------------</td>
<td>---------</td>
<td>----------------</td>
<td>------------</td>
<td>---------------</td>
</tr>
<tr>
<td>SAMP01823701</td>
<td>CFSAN000189</td>
<td>CP006053, CP006054</td>
<td>SRR498276</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00860590</td>
<td>CFSAN000191</td>
<td>JMMH00000000</td>
<td>SRR498369</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00989085</td>
<td>CFSAN000211</td>
<td>JMMH00000000</td>
<td>SRR498373</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00862341</td>
<td>CFSAN000212</td>
<td>JRD00000000</td>
<td>SRR500494</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00862340</td>
<td>CFSAN000228</td>
<td>JRCV00000000</td>
<td>SRR500493</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00991042</td>
<td>CFSAN000661</td>
<td>JMMGG00000000</td>
<td>SRR498397</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00991044</td>
<td>CFSAN000669</td>
<td>JRCG00000000</td>
<td>SRR498399</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00991046</td>
<td>CFSAN000700</td>
<td>JRCO00000000</td>
<td>SRR498402</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00991047</td>
<td>CFSAN000752</td>
<td>JRCN00000000</td>
<td>SRR498403</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00991048</td>
<td>CFSAN000753</td>
<td>JRCM00000000</td>
<td>SRR498404</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00991050</td>
<td>CFSAN000951</td>
<td>JRCI00000000</td>
<td>SRR498422</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00991051</td>
<td>CFSAN000952</td>
<td>JRCC00000000</td>
<td>SRR498423</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00991053</td>
<td>CFSAN000954</td>
<td>JRCG00000000</td>
<td>SRR498425</td>
<td>outbreak</td>
</tr>
<tr>
<td>SAMP00991081</td>
<td>CFSAN000958</td>
<td>JRCO00000000</td>
<td>SRR498431</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP00991087</td>
<td>CFSAN000960</td>
<td>JRC00000000</td>
<td>SRR498433</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP00991100</td>
<td>CFSAN000961</td>
<td>JRC00000000</td>
<td>SRR498434</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP00941246</td>
<td>CFSAN000963</td>
<td>JRC00000000</td>
<td>SRR498436</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP00942468</td>
<td>CFSAN000968</td>
<td>JMMF00000000</td>
<td>SRR498442</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP00996629</td>
<td>CFSAN000970</td>
<td>JR00000000</td>
<td>SRR498444</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP01816358</td>
<td>CFSAN001112</td>
<td>JRBL00000000</td>
<td>SRR1258439</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP01816359</td>
<td>CFSAN001115</td>
<td>JRBL00000000</td>
<td>SRR1258442</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP01816351</td>
<td>CFSAN001118</td>
<td>JRBL00000000</td>
<td>SRR1258443</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP01816352</td>
<td>CFSAN001140</td>
<td>JRBL00000000</td>
<td>SRR1258440</td>
<td>1203NYJAP-1</td>
</tr>
<tr>
<td>SAMP01816353</td>
<td>CFSAN001140</td>
<td>JRBL00000000</td>
<td>SRR1258440</td>
<td>1203NYJAP-1</td>
</tr>
</tbody>
</table>
Metadata and Code at GitHub

datasets
  Escherichia_coli_1405WAEXK-1.xlsx
  Listeria_monocytogenes_1408MLGX6-3WGS.xlsx
  Salmonella_enterica_1203NYJAP-1.xlsx

Scripts
  downloadDataset.pl
  readChecksum.pl
  readChecksum.sh

https://github.com/WGS-standards-and-analysis/datasets
Data Archiving

Permanent repository for data:

- **NCBI** – sequence data, assemblies

- **openTreeOfLife** – tree files
  - [tree.opentreeoflife.org/curator/study/edit/ot](http://tree.opentreeoflife.org/curator/study/edit/ot)
Working with the Datasets

1. Visit WGS standards GitHub site
2. "Git clone" to download tables and script (or zip file)
3. Run `downloadDataset.pl` on table file
4. Downloaded files, verify integrity with checksums:
   - `fastq` files
   - `fasta` assemblies
   - `tree` file (newick format)
5. Perform your own clustering pipeline(s)
   (FDA: SNP pipeline, CDC: LyveSet, NCBI: Pathogen Detection, ...)
6. Compare results!
Preliminary Analysis Results

SNP comparison

- 1st round of comparison saw minor differences in SNP calling, which is leading to improvements in all three pipelines (FDA/CDC/NCBI).
- working towards a validated, gold standard SNP set in a stable file format.

Tree comparison

- Trees are generally the same with respect to correctly reconstructing the outbreak clade. More precise comparison in progress.
Comparison of the *E. coli* outbreak trees

- Both outbreak genomes cluster with the correct clade with 100% in all trees
- Most trees have almost the exact same topology with high confidence values for outbreak clades
Simulated data pipeline
Collaboration with UK PostDoc, Emily Jane McTavish

Download here:
https://github.com/snacktavish/TreeToReads
Parameters of interest

Data:
- Sequence coverage (5x, 10x, 20x, etc)
- NGS read length (150bp, 250bp, etc)
- Amount of sequence divergence (total number of SNPs or variable sites)
- Tree shape (length of node leading to OB)
- Clumpiness of SNP distribution

Analysis:
- SNP-based methods
  - reference vs. ref-free SNP collection
  - Distance of reference genome
- Effects of phylogenetic inference method? NJ, MP, ML.
- Maximum likelihood model misspecification?
- Alternate analysis methods (wgMLST, Kmer)

Validation
Test the effect of different parameters and analysis methods.
Future Developments

• Additional benchmark datasets (new outbreaks, taxa, multi-clonal outbreaks, etc)
• Appeal to the community to add additional validated datasets
• Host at SeqAnswers for easier collaboration
• Complete and easily reproducible SNP comparison
• Experimental validation of results
Carlos G. Malbrán National Institute of Infectious Diseases (INEI)
Salmonella enterica

GenomeTrakr Project: Argentina

Whole genome sequencing of cultured Salmonella enterica as part of the FDA GenomeTrakr surveillance project for rapid source tracking of foodborne illness outbreaks.

Related Resources:
- Malbran Institute, Argentina

Project Data Type: Genome sequencing
Attributes: Scope: Multiisolate; Material: Genome; Capture: Whole; Method type: Sequencing
Relevance: Medical

Project Data:

<table>
<thead>
<tr>
<th>Resource Name</th>
<th>Number of Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRA Experiments</td>
<td>7</td>
</tr>
<tr>
<td>BioSample</td>
<td>139</td>
</tr>
</tbody>
</table>

SRA Data Details

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data volume, Gbases</td>
<td>4</td>
</tr>
<tr>
<td>Data volume, Mbytes</td>
<td>2751</td>
</tr>
</tbody>
</table>

Lineage: Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Salmonella; Salmonella enterica [Taxonomy ID: 28901]

Submission:
Registration date: 20-Dec-2014
FDA Center for Food Safety and Applied Nutrition
Acknowledgements

- GMI Working group 3
- WGS Standards and Analysis workgroup (CDC/FDA/NCBI/FSIS)
  - FDA: Ruth Timme, Hugh Rand, Steven Davis
  - CDC: Lee S. Katz, Eija Trees, Heather A. Carleton
  - NCBI: Richa Agarwala, Martin Shumway, Bill Klimke
  - FSIS: Mustafa Simmons, Glenn Tillman, Philip Bronstein, Stephanie Defibaugh-Chávez
- Emily Jane McTavish (TreeToReads)
- WHO, PAHO, Peter Evans
# Results

<table>
<thead>
<tr>
<th># SNPS</th>
<th>outgroup branch</th>
<th>Correct tree?</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>LONG</td>
<td>YES</td>
</tr>
<tr>
<td>200</td>
<td>SHORT</td>
<td>YES</td>
</tr>
<tr>
<td>50</td>
<td>LONG</td>
<td>YES</td>
</tr>
<tr>
<td>50</td>
<td>SHORT</td>
<td>NO</td>
</tr>
</tbody>
</table>