Application of bettercallsal pipeline on environmental samples

Padmini Ramachandran
FDA/CFSAN
Outbreaks- *Salmonella*

*Salmonella* outbreak- one among the leading cause.

NORS – CDC – Mostly foodborne

~3% of *Salmonella* outbreaks are multi-serovar

https://www.cdc.gov/norsdashboard/
Quasi metagenomics- Microbiological steps to recover individual isolates of *Salmonella*

1. UPB, MBPW, 24 hrs
2. RV, TT 24 hrs
3. XLD, HE 24 hrs
4. 24 hrs

130 hrs- total

180 hrs- total
Challenges on metagenomics dataset on *Salmonella* serotyping

Papaya Outbreak- 2017

**Aerobic**
- Pre-enrichment
- mTT
- TT

**Anaerobic**
- Pre-enrichment
- mTT
- TT

Relative Abundance
- 100
- 75
- 50
- 25
- 0

Sample

- Other genera < 5%
- Bordetella
- Lysinibacillus
- Propionibacterium
- Raletobia
- Klebsiella
- Enterococcus
- Alcaligenes
- Morganella
- Pantea
- Proteus
- Enterobacteriaceae
- E. coli
- Providencia
- Citrobacter
- Enterobacter
- Pseudomonas
- S. enterica
- S. enterica infantis
- S. enterica Newport

Aerobic (MBPW), 24 H

Anaerobic, 24 H

RV, TT 24 hrs

XLD, HE 24 hrs

24 hrs

Thompson, Kiambu
Senftenberg, Agona

Newport
Infantis,

Anatum

24 hrs
Kallisto abundance

SERO-BLAST

- Pulling *Salmonella* reads -
  - Assembly based approaches
  - SeqSero2 on the *Salmonella* only reads
  - SeqSero2 on the assembly
  - SISTR/SRST2
- Many other tools
**Nextflow** – Domain Specific Language (BCS – HTML) Platform, portability layer between logic and execution, checkpoint tracking, execution resumed at any step.

**MASH** – Genome/metagenome distance estimation using MinHash, generates sketch to represent large datasets.

**KMA** – kmer alignment, ConClave score to address ties, reference assembly and statistics. Good for highly redundant databases.

**SALMON** – quantification of reads, performance superior to kallisto, G+C content and positional bias adjustment.
Obtain precise *Salmonella* serotype/serovar information for each sample *insilico*.  

Characterize presence of potential multiple *Salmonella* serotypes in a single sample.  

Work well with both metagenomics and quasi-metagenomics (enriched) datasets. Very fast.  

Leverage *Salmonella* WGS isolate data made available (FDA GenomeTrakr) at NCBI.  

**Automated workflows. v0.4.1:**  
[https://github.com/CFSAN-Biostatistics/bettercallsal](https://github.com/CFSAN-Biostatistics/bettercallsal)  

Latest version is on Galaxytrakr-bcs 0.6.0.  
Comprehensive multiQC report
Papaya Outbreak 2017

[Link to the full manuscript]

---

### Table of Isolates

<table>
<thead>
<tr>
<th>Organism group</th>
<th>Sample</th>
<th>Isolate identifiers</th>
<th>Server</th>
<th>Create date</th>
<th>Location</th>
<th>Isolation source</th>
<th>Isolation...</th>
<th>Min-same</th>
<th>Min-diff</th>
<th>RetSample</th>
<th>Assembly</th>
<th>Completed types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salmonella spp.</td>
<td>MD-001...</td>
<td>CF3SNX67219</td>
<td>USDA-ABD</td>
<td>2017-08-18</td>
<td>USA, MD</td>
<td>papaya</td>
<td>environm...</td>
<td>0</td>
<td>0</td>
<td>SARM674954</td>
<td>GCA_00575815.1</td>
<td></td>
</tr>
<tr>
<td>Salmonella spp.</td>
<td>FDA01985...</td>
<td>CF3SNX662565</td>
<td>FDA019899-5004</td>
<td>2017-08-16</td>
<td>Mexico</td>
<td>papaya</td>
<td>environm...</td>
<td>0</td>
<td>1</td>
<td>SARM674954</td>
<td>GCA_00563505.5</td>
<td></td>
</tr>
<tr>
<td>Salmonella spp.</td>
<td>FDA01985...</td>
<td>CF3SNX66255</td>
<td>FDA019899-5004</td>
<td>2017-08-16</td>
<td>Mexico</td>
<td>papaya</td>
<td>environm...</td>
<td>0</td>
<td>1</td>
<td>SARM674954</td>
<td>GCA_00563505.5</td>
<td></td>
</tr>
<tr>
<td>Salmonella spp.</td>
<td>FDA01985...</td>
<td>CF3SNX66252</td>
<td>FDA019899-5004</td>
<td>2017-08-16</td>
<td>Mexico</td>
<td>papaya</td>
<td>environm...</td>
<td>0</td>
<td>1</td>
<td>SARM674954</td>
<td>GCA_00563505.5</td>
<td></td>
</tr>
<tr>
<td>Salmonella spp.</td>
<td>FDA01985...</td>
<td>CF3SNX66251</td>
<td>FDA019899-5004</td>
<td>2017-08-16</td>
<td>Mexico</td>
<td>papaya</td>
<td>environm...</td>
<td>0</td>
<td>0</td>
<td>SARM674954</td>
<td>GCA_00563505.5</td>
<td></td>
</tr>
</tbody>
</table>

---

### Diagram

[Image of a digital interface displaying data related to the outbreak]
Environmental sampling- application of bettercallsal

Chief Scientist Grant
CVM

Brandon Kocurek, Andrea Ottesen, and Errol Strain
Florida

- A total of 9 sites were samples across the State of Florida

- Sites were selected from Developed, Agriculture & Natural hotspots
  - 3 Developed
  - 3 Agriculture
  - 3 Natural

- 3 treatment groups were collected, prepared, sequenced & analyzed:
  - Culture Independent (CI)
  - General Enrichment (2x BPW)
  - *Salmonella* selective enrichment (RV/TT)
<table>
<thead>
<tr>
<th>Organism group</th>
<th>Strain</th>
<th>Isolate identifiers</th>
<th>Serovar</th>
<th>Isolate</th>
<th>Create date</th>
<th>Locat...</th>
<th>Isolation source</th>
<th>Isolation ...</th>
<th>Min-same</th>
<th>Min-diff</th>
<th>BioSample</th>
<th>Assembly</th>
<th>Computed types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salmonella enterica</td>
<td>FL_FLDACS...</td>
<td>FL_FLDACS-0926 SR528365</td>
<td></td>
<td>PDT000000428.2</td>
<td>2013-08-28</td>
<td>USA-FL</td>
<td>alligator meat</td>
<td>environnm...</td>
<td>10</td>
<td>5</td>
<td>SAMN02253011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>NY_IDR110...</td>
<td>MDP_10_60066 NY_IDR11000061 SR528562</td>
<td></td>
<td>PDT000002519.5</td>
<td>2018-06-26</td>
<td>USA:...</td>
<td>food</td>
<td>environnm...</td>
<td>13</td>
<td>12</td>
<td>SAMN01902381 GCA_007773065.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>FL-MDD-75</td>
<td>FL-MDD-75 SR580059</td>
<td>Baildon</td>
<td>PDT0000024357.2</td>
<td>2014-03-26</td>
<td>USA-FL</td>
<td>surface water</td>
<td>environnm...</td>
<td>33</td>
<td>22</td>
<td>SAMN02564723 GCA_010655465.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>FLDOH-38</td>
<td>CFSA180828 FLDOH-38 SR549134</td>
<td></td>
<td>PDT0000054901.2</td>
<td>2014-06-27</td>
<td>USA-FL</td>
<td>clinical</td>
<td>GCA_011491175.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Diagram](https://www.ncbi.nlm.nih.gov/pathogens/isolates/#PDT000274118.2)

**FL-Agr-RV-Site4-A**
## Comparing bettercallsal to WGS

<table>
<thead>
<tr>
<th>Group</th>
<th>Site</th>
<th>RV/TT</th>
<th>Bettercallsal (RV/TT QMGS)</th>
<th>Isolates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developed</td>
<td></td>
<td>RV</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>RV</td>
<td>Barranquilla*</td>
<td>Gaminara</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>Gaminara</td>
<td>Gaminara</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>RV</td>
<td>Muenchen</td>
<td>Muenchen</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>Muenchen</td>
<td>Muenchen</td>
</tr>
<tr>
<td>Natural</td>
<td>5</td>
<td>RV</td>
<td>Weltevreden</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>Weltevreden</td>
<td>Na</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>RV</td>
<td>IIIb 35;i,v;235</td>
<td>IIIb 60:r;z</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>RV</td>
<td>IIIb 60:r;z</td>
<td>IIIb 60:r;z</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>IIIb 60:r;z</td>
<td>IIIb 60:r;z</td>
</tr>
<tr>
<td>Agriculture</td>
<td>4</td>
<td>RV</td>
<td>Baildon</td>
<td>IIIb 61;r;z</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>Baildon</td>
<td>Rubislaw</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>RV</td>
<td>IIIb 47:k;253</td>
<td>IV 50;z4,z23;1,7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>Gaminara</td>
<td>IIIb 47:k;253</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>RV</td>
<td>Glostrup or Chomedey</td>
<td>Poona</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>Glostrup or Chomedey</td>
<td>Hartford</td>
</tr>
</tbody>
</table>
Arkansas

• A total of 15 sites were samples across the State of Arkansas

• Sites were selected from Developed, Agriculture & Natural hotspots
  • 5 Developed
  • 5 Agriculture
  • 5 Natural

• 3 treatment groups were collected, prepared, sequenced & analyzed:
  • Culture Independent (CI)
  • General Enrichment (2x BPW)
  • Salmonella selective enrichment (RV/TT)
Salmonella serovars (Bettercallsal)

**Culture Positive**
- Culture positive Salmonella found in the following sites:

<table>
<thead>
<tr>
<th>Site</th>
<th>Dev/Agr/Nat</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Natural</td>
</tr>
<tr>
<td>3</td>
<td>Natural</td>
</tr>
<tr>
<td>4</td>
<td>Natural</td>
</tr>
<tr>
<td>6</td>
<td>Developed</td>
</tr>
<tr>
<td>7</td>
<td>Developed</td>
</tr>
<tr>
<td>8</td>
<td>Developed</td>
</tr>
<tr>
<td>9</td>
<td>Agriculture</td>
</tr>
<tr>
<td>10</td>
<td>Agriculture</td>
</tr>
<tr>
<td>12</td>
<td>Agriculture</td>
</tr>
</tbody>
</table>
## Agr RV Site10 - Arkansas

<table>
<thead>
<tr>
<th>#</th>
<th>Organism group</th>
<th>Strain</th>
<th>Isolate identifiers</th>
<th>Serovar</th>
<th>Isolate</th>
<th>Create date</th>
<th>Locat...</th>
<th>Isolation source</th>
<th>Isolation ...</th>
<th>Min-same</th>
<th>Min-diff</th>
<th>BioSample</th>
<th>Assembly</th>
<th>Computed types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Salmonella ente...</td>
<td>PNUSA0215...</td>
<td>PNUSA0215035 SRS9693875</td>
<td>POT00103610.1</td>
<td>2021-08-07</td>
<td>USA</td>
<td>clinical</td>
<td>18</td>
<td>n/a</td>
<td>S002600798</td>
<td>GCA_019485005.1</td>
<td>Serotype: Newport Antigen formula: B1.ch1,2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Salmonella ente...</td>
<td>PNUSA007...</td>
<td>PNUSA0076941 SRS1952141</td>
<td>Newport</td>
<td>POT00103620.2</td>
<td>2017-02-01</td>
<td>USA</td>
<td>urine</td>
<td>clinical</td>
<td>1</td>
<td>n/a</td>
<td>S006231875</td>
<td>GCA_001483905.1</td>
<td>Serotype: Newport Antigen formula: B1.ch1,2</td>
</tr>
<tr>
<td>3</td>
<td>Salmonella ente...</td>
<td>PNUSA006...</td>
<td>PNUSA0060941 SRS1980020</td>
<td>Newport</td>
<td>POT00103630.2</td>
<td>2017-02-27</td>
<td>USA</td>
<td>stool</td>
<td>clinical</td>
<td>1</td>
<td>n/a</td>
<td>S006320658</td>
<td>GCA_00123385.1</td>
<td>Serotype: Newport Antigen formula: B1.ch1,2</td>
</tr>
<tr>
<td>4</td>
<td>Salmonella ente...</td>
<td>PNUSA021...</td>
<td>PNUSA0213557 SRS1453526</td>
<td>POT002036819.2</td>
<td>2017-08-23</td>
<td>USA</td>
<td>clinical</td>
<td>10</td>
<td>n/a</td>
<td>S007524581</td>
<td>GCA_006140275.1</td>
<td>Serotype: Newport Antigen formula: B1.ch1,2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Salmonella ente...</td>
<td>PNUSA037...</td>
<td>PNUSA037673 SRS3145496</td>
<td>POT003036512</td>
<td>2018-04-09</td>
<td>USA</td>
<td>Clinical</td>
<td>0</td>
<td>n/a</td>
<td>S00839948</td>
<td>GCA_008319465.1</td>
<td>Serotype: Newport Antigen formula: B1.ch1,2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Salmonella ente...</td>
<td>PNUSA037...</td>
<td>PNUSA037673 SRS3145500</td>
<td>POT003036552</td>
<td>2018-04-09</td>
<td>USA</td>
<td>Clinical</td>
<td>0</td>
<td>n/a</td>
<td>S00839243</td>
<td>GCA_008319465.1</td>
<td>Serotype: Newport Antigen formula: B1.ch1,2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Diagram:
- clinical, 2017-02-17, USA, stool, PNUSA006941, PDT000188299.2
- clinical, 2017-02-01, USA, urine, PNUSA0076941, PDT000184250.2
- clinical, 2017-08-23, USA, PNUSA021557, PDT000235819.2
- clinical, 2019-06-26, USA, PNUSA080605, PDT000532687.1
- clinical, 2018-12-11, USA, PNUSA062419, PDT000416313.1
- clinical, 2021-09-22, USA, PNUSA007649, PDT000141696.1
- clinical, 2019-07-31, USA, PNUSA058874, PDT000551305.1
- clinical, 2018-04-09, USA, PNUSA077682, PDT000509686.1
- clinical, 2018-04-09, USA, PNUSA037673, PDT00034855.1
- clinical, 2018-04-14, USA, PNUSA037677, PDT00036609.2
- clinical, 2021-08-07, USA, PNUSA021535, PDT001103610.1
CFSAN- Florida sampling- Culture Independent

6 Sites- South Florida

Dr. Rebecca Bell and Dr Eric Brown

Samples collected- Aug 2015 to July 2016

<table>
<thead>
<tr>
<th>Site</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Canal Near Agriculture</td>
</tr>
<tr>
<td>B</td>
<td>Canal Near Agriculture</td>
</tr>
<tr>
<td>C</td>
<td>Natural Swamp</td>
</tr>
<tr>
<td>D</td>
<td>Natural Swamp</td>
</tr>
<tr>
<td>E</td>
<td>Man Made Canal</td>
</tr>
<tr>
<td>F</td>
<td>Man Made Canal</td>
</tr>
</tbody>
</table>
Correlation of relative abundance of Salmonella to water temperature
Culture independent samples - 1% genome coverage

Salmon: Read counts

- Bredeney
- Manhattan
- Montevideo
- Poona
- Reading
- Rubislaw
- Schwarzengrund
- No genome hit

Created with MultiQC
Conclusions and next steps

- Bettercallsal helps in detection of multi serovars of Salmonella in metagenomics dataset. Could be used in WGS when serotype information is not available.

- Many check points to avoid false positives. CFSAN longitudinal study on environmental samples, we observed close to ~20% samples are dual serovars, ~10% are 3 serovars

- Easy to access/navigate report with Sequence typing and computed serotype information

Next steps

- mashtree + iTOL for the main workflow and then gunc + checkm2 check for the db creation

- Expand the workflow for other organisms. We are already working on Cronology for Cronobacter.

- Use bettercallsal for WGS data and make some custom changes.
Acknowledgements

CFSAN
Kranti Konganti
Chris Grim
Amanda Windsor
Mark Mammel
Elizabeth Reed
Rebecca Bell
Sandra Tallent
Eric Brown

CVM
Brandon Kocurek
Andrea Ottesen

GenomeTrakr labs and NCBI PD