Whole Genome Sequencing of Food Isolates

Ai Kataoka, - GMA
Danielle Montoya, M.S. - Noblis

May 23, 2017
Project Objectives

- Promote genomic testing in the Food Industry
- To have participants gain familiarity with the Whole Genome Sequencing (WGS) technology by analyzing a set of their bacterial isolates
Participants

- 1 global food manufacturing company
- 2 food laboratories
- 1 university extension specialized in produce quality and safety
Pilot Project Flow

1. Participants submit samples (food and environmental isolates)
2. 3rd party lab sequences samples
3. Noblis analyzes the raw data using their bioinformatic software (BioVelocity)
4. Participants receive outcomes (data analysis) and provide feedback
Data Confidentiality

- Non-disclosure agreement
- Data and detailed outcomes stayed confidential
Samples (Food and Environmental isolates)

- Total of 430 samples analyzed
- Pathogenic organisms
  - *Salmonella*
  - *Listeria monocytogenes*
  - *Cronobacter sakazakii*
- Other types of organisms
  - *Lactobacillus*
  - *Citrobacter*
Pilot Project Timeline

- **Oct 2015**: Pilot started; Noblis finalized analysis requirements
- **Nov 2015**: Partners sent samples for sequencing
- **Dec 2015**: Noblis to analyze sequencing data via BioVelocity
- **Jan 2016**: Complete first round of meetings with partners
- **Mar 2016**: Complete second round of meetings with partners

- **Oct 2015**: Noblis team to complete analysis validation
- **Nov 2015**: WGS completed for 430 samples; data sent to Noblis
- **Jan 2016**: Complete first round of meetings with partners
- **Mar 2016**: Complete second round of meetings with partners
Sequence Identification Pipeline

Pathogen → Sequencing → Alignment AGAINST REFERENCE GENOME → SNPs → Analysis → Phylogenetic Tree

- Raw reads using proprietary algorithm
- Identify SNPs
- SNP matrix identifies genome and position
Pilot Output

**FASTA/FASTQ sequence files**
- Common bioinformatic formats
- Compatible with other common bioinformatic tools and pipelines

**Report**
- Table of raw results (e.g. sequence % alignment to reference)
- Phylogenetic tree
- Summary of findings

Data interpretation may be performed in collaboration with your scientists
## Raw Data

<table>
<thead>
<tr>
<th>Reference</th>
<th>RefLength</th>
<th>Perfect Alignments</th>
<th>Alignments</th>
<th>Perc Perfect Reads</th>
<th>Perc Reads</th>
<th>Coverage (per bp)</th>
<th>Num SNPs</th>
<th>Cov0x</th>
<th>Cov20x</th>
<th>Cov40x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listeria monocytogenes strain L2624</td>
<td>2932501</td>
<td>621073</td>
<td>817446</td>
<td>65.47%</td>
<td>86.17%</td>
<td>68</td>
<td>234</td>
<td>0.992</td>
<td>0.972</td>
<td>0.869</td>
</tr>
<tr>
<td>Listeria monocytogenes strain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFSAN010068</td>
<td>2987563</td>
<td>617586</td>
<td>818603</td>
<td>65.10%</td>
<td>86.29%</td>
<td>67</td>
<td>567</td>
<td>0.979</td>
<td>0.956</td>
<td>0.854</td>
</tr>
<tr>
<td>Listeria monocytogenes strain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFSAN008100</td>
<td>3039897</td>
<td>618706</td>
<td>819804</td>
<td>65.22%</td>
<td>86.42%</td>
<td>66</td>
<td>603</td>
<td>0.965</td>
<td>0.943</td>
<td>0.84</td>
</tr>
<tr>
<td>Listeria monocytogenes strain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFSAN023459</td>
<td>3039887</td>
<td>626347</td>
<td>818720</td>
<td>66.03%</td>
<td>86.30%</td>
<td>66</td>
<td>181</td>
<td>0.959</td>
<td>0.939</td>
<td>0.839</td>
</tr>
<tr>
<td>Listeria monocytogenes strain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFSAN008100 plasmid pCFSAN010068_01</td>
<td>55521</td>
<td>11236</td>
<td>14937</td>
<td>1.18%</td>
<td>1.58%</td>
<td>69</td>
<td>3</td>
<td>0.976</td>
<td>0.931</td>
<td>0.836</td>
</tr>
<tr>
<td>Listeria monocytogenes strain N2306</td>
<td>2911639</td>
<td>383854</td>
<td>742472</td>
<td>40.46%</td>
<td>78.27%</td>
<td>62</td>
<td>8823</td>
<td>0.957</td>
<td>0.925</td>
<td>0.809</td>
</tr>
<tr>
<td>Listeria monocytogenes strain NTSN</td>
<td>2904500</td>
<td>379907</td>
<td>739248</td>
<td>40.05%</td>
<td>77.93%</td>
<td>62</td>
<td>8974</td>
<td>0.955</td>
<td>0.923</td>
<td>0.809</td>
</tr>
<tr>
<td>Listeria monocytogenes strain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IZSAM_Lm_hs2008</td>
<td>2904228</td>
<td>373758</td>
<td>736831</td>
<td>39.40%</td>
<td>77.67%</td>
<td>62</td>
<td>9238</td>
<td>0.954</td>
<td>0.923</td>
<td>0.807</td>
</tr>
<tr>
<td>Listeria monocytogenes strain CFSAN023463</td>
<td>2939733</td>
<td>382071</td>
<td>740260</td>
<td>40.28%</td>
<td>78.03%</td>
<td>62</td>
<td>8895</td>
<td>0.947</td>
<td>0.914</td>
<td>0.8</td>
</tr>
<tr>
<td>Listeria monocytogenes strain WSLC 1018</td>
<td>2942510</td>
<td>381094</td>
<td>739276</td>
<td>40.17%</td>
<td>77.93%</td>
<td>61</td>
<td>8953</td>
<td>0.943</td>
<td>0.912</td>
<td>0.798</td>
</tr>
<tr>
<td>Listeria monocytogenes ATCC 19117 strain</td>
<td>2950820</td>
<td>374608</td>
<td>734854</td>
<td>39.49%</td>
<td>77.46%</td>
<td>61</td>
<td>9083</td>
<td>0.936</td>
<td>0.904</td>
<td>0.792</td>
</tr>
<tr>
<td>Listeria monocytogenes strain WSLC 1047</td>
<td>2950749</td>
<td>373692</td>
<td>732807</td>
<td>39.39%</td>
<td>77.25%</td>
<td>61</td>
<td>9076</td>
<td>0.936</td>
<td>0.903</td>
<td>0.791</td>
</tr>
<tr>
<td>Listeria monocytogenes strain ATCC 19114</td>
<td>2842039</td>
<td>29462</td>
<td>335630</td>
<td>3.11%</td>
<td>35.38%</td>
<td>30</td>
<td>33759</td>
<td>0.702</td>
<td>0.52</td>
<td>0.335</td>
</tr>
<tr>
<td>Listeria monocytogenes strain WSLC 1019</td>
<td>2839870</td>
<td>29432</td>
<td>335744</td>
<td>3.10%</td>
<td>35.39%</td>
<td>30</td>
<td>33778</td>
<td>0.702</td>
<td>0.52</td>
<td>0.334</td>
</tr>
</tbody>
</table>
Pilot Results and Feedback

- **Challenges**
  - Legal implications of sample identification
  - Cost of WGS vs PCR/Current identification methodologies

- **Benefits**
  - Better understanding of technology
  - Correction of mislabeled samples
  - Overall collaboration and knowledge regarding WGS and data analytics
Future Application in Food Industry

- *In silico* PCR Primer Design (ASSET Application)
  - Serotyping
- Strain-to-strain comparison
- Quality assurance analysis (Metagenomics)
- Data obfuscation (Bloom-cipher Application)
In silico PCR Primer Design

- Select and kmerize the reference genome(s) for target organism(s)
- Select and kmerize the reference genomes for organisms from which it would be desirable to generate distinct primers
- Find kmers that are present only in the target set and are not present in the larger set
- Map signature kmers back to reference and evaluate potential for primer design.
Data Obfuscation – Why?

- Large amounts of read set data that people do not want to share
  - Legal liability concerns
  - Unique and/or proprietary sample
  - Desire to keep sequence repository in house

- Large data sets that are not easily transferrable
  - The volume of data is too large to handle appropriately
  - Can the data be compressed for transmission

https://commons.wikimedia.org/w/index.php?curid=47910919
Data Obfuscation - Preliminary Results

- Ice Cream metagenomic sample
- Almost 99% of the bacterial genomes are excluded due to low kmer hit rate
Data Obfuscation – Preliminary Results

Top Hit: 99.237%
- All Top Kmer Hit Rates are Listeria monocytogenes
- Verified using Noblis’ BioVelocity alignment algorithm
Contact Us

Ai Kataoka
Scientist
akataoka@gmaonline.org
GMA
1350 I Street, NW
Suite 300
Washington, DC 20005
(202) 639-5000
info@gmaonline.org
www.gmaonline.org

Danielle Montoya
Project Manager, Bioinformaticist
Danielle.Montoya@noblis.org

Noblis for the best of reasons
2002 Edmund Halley Drive
Reston, VA 20191
(703) 610-2000
answers@noblis.org
www.noblis.org
Appendix
Metagenomics Case Study

- Worked with NIH National Institute of Neurological Disorders and Stroke (NINDS) to demonstrate identifying multiple organisms in the presence of human DNA.
- We tested our metagenomics capability on blinded data that was intentionally spiked with a variety of pathogens (bacteria, viruses, parasites).
- We were able to correctly identify 32 out of 33 spiked in pathogens (missed one bacteria that was not in our reference database).

<table>
<thead>
<tr>
<th>Virus Found</th>
<th>Bacteria Found</th>
<th>Parasite Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chikungunya virus</td>
<td>Bacillus anthracis</td>
<td>Cryptosporidium parvum Iowa II</td>
</tr>
<tr>
<td>Dengue virus</td>
<td>Chlamydia psittaci</td>
<td>Plasmodium vivax Sal-1</td>
</tr>
<tr>
<td>Hepatitis C virus</td>
<td>Chlamyphila psittaci</td>
<td></td>
</tr>
<tr>
<td>Human enterovirus A</td>
<td>Clostridium botulinum</td>
<td></td>
</tr>
<tr>
<td>Human enterovirus B</td>
<td>Escherichia coli</td>
<td></td>
</tr>
<tr>
<td>Human enterovirus D</td>
<td>Haemophilus influenzae</td>
<td></td>
</tr>
<tr>
<td>Human herpesvirus</td>
<td>Klebsiella variicola</td>
<td></td>
</tr>
<tr>
<td>Human metapneumovirus</td>
<td>Legionella pneumophila</td>
<td></td>
</tr>
<tr>
<td>Human papillomavirus</td>
<td>Staphylococcus aureus</td>
<td></td>
</tr>
<tr>
<td>Human respiratory syncytial virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human rhinovirus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lymphocytic choriomeningitis virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marburg virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measles virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mumps virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norwalk virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rabies virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rubella virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SARS coronavirus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tick-borne encephalitis virus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zaire ebolavirus</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Virus</th>
<th>Bacteria</th>
<th>Parasite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>21/21</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>Sample 2</td>
<td>1/1</td>
<td>9/10</td>
<td>2/2</td>
</tr>
</tbody>
</table>