Genomic Epidemiology

Lee Katz, Ph.D.

Senior bioinformatician
Enteric Diseases Laboratory Branch

Computational Genomics course
Jan 31, 2019
Acknowledgements up front

• Every single compgenomics class since 2008
• My branch at CDC
• Federal partners
• State partners
Enteric Diseases Laboratory Branch (EDLB)

Food Safety Informatics Group, Center for Food Safety, University of Georgia

Enteric Diseases Bioinformatics Team (EDBiT)
Enteric Diseases Laboratory Branch

2011 to present

Vibrio, Campylobacter, Escherichia, Shigella, Yersinia, Salmonella
PulseNet’s 20-year history of making food safer to eat

- **1993**: Outbreak of a deadly form of *E. coli* infections in western states; > 700 illnesses; 4 children died. Highlights the need for a network to identify DNA fingerprints of foodborne bacteria.

- **1996**: CDC, the Association of Public Health Laboratories, federal partners, and public health labs in 4 states launch a new foodborne surveillance network called PulseNet.

- **2001**: PulseNet goes nationwide; all 50 state public health labs perform DNA fingerprinting of foodborne bacteria.

- **2002**: Identifies spinach as source of *E. coli* outbreak; 225 sickened and 5 deaths in 27 states; prompts nationwide recall.

- **2006**: PulseNet wins 2nd Innovations in American Government Award (also won in 1999) for excellence and creativity in the public sector.
2009: Traces a *Salmonella multistate outbreak* to peanut butter/peanut products; 700 illnesses, 9 deaths in 46 states, >3,000 products recalled.

2010: 1st time whole-genome sequencing (WGS) used in a foodborne disease investigation. PulseNet uses WGS on samples from a cholera outbreak in Haiti.

2013: Begins using WGS on illnesses caused by *Listeria* infection.

2014: WGS used for routine surveillance of *Listeria*, *Campylobacter*, and *E. coli* at CDC and in states with genetic sequencing capacity.

2016: "PulseNet and Beyond" project consolidates identification of foodborne bacteria into a single, fast, and efficient process under Advanced Molecular Detection (AMD).
Outline

- **Background**

- **Genomic Epidemiology**
  - Algorithms
  - Software

- **Example**

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.
Listeria pilot project

As told from a bioinformatician’s perspective

(It’s an awesome perspective)
Why *Listeria monocytogenes*?

- Illness is rare but serious, costly, and commonly outbreak associated
  - Estimated $2.8 billion in annual medical costs and lost productivity ($1.8 million/case)
- Current subtyping methods are not ideal
- Strong epidemiologic surveillance (Listeria Initiative)
- Strong regulatory component
- Listeria genome is fairly small, stable, and relatively easy to sequence and analyze. Most changes in the genome are due to point mutations and not phages.
The Problem: Detecting Outbreaks in an Increasingly Globalized Food System

Thanks to Brendan Jackson for letting me borrow this slide
Limitations of Pulsed-Field Gel Electrophoresis (PFGE)
Limitation: Genetically Unrelated Isolate Might Appear Same by PFGE
Limitation: Genetically Related Isolate Might Appear Different By PFGE
Can genomics clear up this picture?
The Basics of Next Generation Sequencing (NGS)

• “Massive parallel sequencing”
• The whole genome sequenced in small random pieces (‘shotgun sequencing’, 25- >1000 bp) multiple times (‘coverage’)
• ‘Coverage’ usually 20- several 100 X
How do we compare genomes?
Three major methods we use

- Kmer-based: mile-high view (shredded paper)
- MLST-based: naked eye (book pages)
- SNP-based: microscope (book letters)

- The question in this analogy: how similar are these two books?
kmers

• **Kmer**: a length of DNA $k$ nucleotides long

1. Shred all reads in equal sizes $k$
2. How many kmers are in common?
3. Transform into a percentage **

** Known as the jaccard distance

Image credits:
“DEATH OF A SHREDDER”
https://digginginthedriftless.com/2011/01/04/death-of-a-shredder
Kmers, jaccard distance

Here, K=12

```
CAAAAAAAAAAAAAAAAA
CAAAAAAAAAAAA
CAAAAAAAAAAAA
CAAAAAAAAAAA
```

```
CAAAAAAAAAAAA
CAAAAAAAAAAAA
CAAAAAAAAAAA
CAAAAAAAAAAA
```

Two out of four kmers different;
Jaccard distance = 2/4 = 0.5
Example kmer tree

- Software: pathogen detection pipeline at NCBI

Mile-high view
7,800 *Listeria monocytogenes* genomes in this tree
Kmer-based software

- NCBI Pathogen Detection Pipeline
  - Not available for individual use, but the results are comprehensive and public

- Mashtree
  - Based on min-hash, implemented in Mash

- SKA
  - Split Kmer Analysis

[https://github.com/lskatz/mashtree](https://github.com/lskatz/mashtree) (latest version: 0.37)
[https://github.com/simonrharris/SKA/releases](https://github.com/simonrharris/SKA/releases) (latest version: 1.0)
How does Mash work?

“Sketch”

@read1
GGATTAGG
+
IIIIIIII
@read2
GGATTAAA
+
IIIIIIII
...

Min-hash

Kmer counting

@read1
GGATTAGG
+
IIIIIIII
@read2
GGATTAAA
+
IIIIIIII
...

GGATT - 2
GATTA - 2
ATTAG - 1
TTAGG - 1
ATTA - 1
TTAAA - 1

hashing

66 - 2
42 - 2
82 - 1
87 - 1
64 - 1
22 - 1
...

How does Mash work?

Min-hash

How does Mash work?

“Distance” or “dist”

Min-hash

Six different hashes, two differences.
Jaccard distance = 2/6 = 0.33

The resolution gets better with more hashes.

Min-hash visualization

A = \textcolor{blue}{+} \hspace{1cm} B = \textcolor{green}{+}
S(A) = \textcolor{blue}{+} \hspace{1cm} S(B) = \textcolor{green}{+}
Mashtree

What it is and what it isn’t

<table>
<thead>
<tr>
<th>Is</th>
<th>Isn’t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Builds trees</td>
<td>Infers phylogeny</td>
</tr>
<tr>
<td>Fast</td>
<td>Slow</td>
</tr>
</tbody>
</table>

When to use it

<table>
<thead>
<tr>
<th>Use it when</th>
<th>Don’t use it when</th>
</tr>
</thead>
<tbody>
<tr>
<td>Need fast estimate</td>
<td>Need solid results</td>
</tr>
<tr>
<td>Need to know a good reference genome</td>
<td>Inferring phylogenetic relatedness</td>
</tr>
<tr>
<td>Large, diverse dataset</td>
<td>Not diverse or not large dataset</td>
</tr>
</tbody>
</table>
Mashtree is fast

- I had a tree of > 1500 genomes and ran Mashtree on the genomes of every clade with fewer than 101 taxa.
- The forward Illumina read of every genome was analyzed.
- Grey shading indicates the range of durations. (next slide)
Mashtree is fast

- I had a tree of > 1500 genomes and ran Mashtree on the genomes of every clade with fewer than 101 taxa.
- The forward Illumina read of every genome was analyzed.
- Grey shading indicates the range of durations.
The Mashtree v0.06 accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity (Sn)</th>
<th>Specificity (Sp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lyve-SET</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>kSNP3</td>
<td>100%</td>
<td>58%</td>
</tr>
<tr>
<td>RealPhy</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Snp-Pipeline</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Dataset from Katz et al., "Lyve-SET," 2017, MGEN

Mash v0.06
- Reads
- min_depth: 5x
- Sn = 100%
- Sp = 97%

Mash v0.06
- Megahit asm
- 59-3141 contigs
- Sn = 78%
- Sp = 100%

Mash v0.06
- SPAdes asm
- 23-46 contigs
- Sn = 100%
- Sp = 97%
Mashtree is command line

# Installation
$ cpanm -L ~ Mashtree
$ export PERL5LIB=$PERL5LIB:$HOME/lib/perl5

# Usage
$ mashtree.pl --help

# Execution
$ mashtree.pl --numcpus 12 --genomesize 4700000 \
  *.fastq.gz \
  [*fast*] [*gbk*] [*fasta.gz*] [*gbk.gz*] \
> mashtree.dnd

https://github.com/lskatz/mashtree
MLST

• **MLST**: multilocus sequence typing
• **Locus**: a place in a genome. Plural: **loci**

• Identify a set of loci (genes) in the genome
• Compare each locus in a genome against the set of loci
• Count differences and the number of loci compared

Different kinds
• 7-gene MLST
• wgMLST (whole genome MLST)
• cgMLST (core genome MLST)
• … and more

Image credit: Wikipedia.org
Software: BioNumerics
7-gene MLST

- Choose about seven loci in the genome
- Compare all genomes based on these seven loci
- This profile of alleles is called a sequence type (ST)

Maiden et al 1998 PNAS
Animation of MLST

0. Assemble the genome
1. Identify the loci
2. Call alleles
3. Compare with other genomes and their alleles
4. Create a phylogeny

• Note: many methods do not require an assembly and these are called assembly-free methods.
Whole-genome MLST

~one locus per 1,000 nucleotides (nt) in the genome.
Different species have different sizes
  e.g., *L. monocytogenes* has ~3,000,000 nt and ~3,000 loci

Strain A

Strain B

Strain C
Flavors of multilocus sequence type analysis

- Subsets of genes can be used to identify genus/species and lineage (rMLST/MLST)
- Core genome MLST are the genes that are in common in vast majority of genomes belonging to a genus species (for Listeria – 1748 genes belong to core and are present in ~98% of isolates tested)

Example wgMLST tree

- Larger circles represent more with the same sequence type (ST)
- 4800 loci represented
- Distances shown on the connecting lines

- The style of tree shown is called a minimum spanning tree
- wgMLST can also be displayed in a conventional tree
MLST software

- **StringMLST**
  - Compare kmers of raw reads against a database

- **BioNumerics**
  - Graphical user interface

- **SRST2, Ariba**
  - Map raw reads onto database

- **mlst**
  - BLAST genome assembly against database

- **Mentalist**
  - Command line, meant for wgMLST schemes

Image taken from http://www.applied-maths.com/applications/wgmlst
For more information: Page et al 2017, “Comparison of Multi-locus Sequence Typing software for next generation sequencing data.”
MLST Resources

- **Main MLST site:** https://pubmlst.org/
- **BigsDB manual:** http://bigsdb.readthedocs.io/en/latest/concepts.html
- **API:** https://pubmlst.org/rest/
- **Also see:**
  - https://enterobase.warwick.ac.uk/
  - http://bigsdb.web.pasteur.fr/listeria/

*Jolley & Maiden 2010, BMC Bioinformatics 11:595
Jolley et al. (2017) Database 2017: bax060*
SNPs

• Compare individual letters in a query genome against the reference genome
• hqSNP: high-quality SNP (ie, high confidence)
• hqSNP indicates some high threshold, e.g.,
  • 10x coverage
  • 75% consensus
SNP analysis

0. Pre-processing
   a) Identification of troublesome regions
   b) Read cleaning

1. Mapping
2. SNP calling
   a) % consensus
   b) x depth
   c) Other filters
3. Phylogeny inference

https://github.com/lskatz/lyve-SET

https://github.com/lskatz/lyve-SET
SNP software

- **Lyve-SET**
  - Optimized for outbreak surveillance.

- **SNP-Pipeline**
  - FDA SNP pipeline. Optimized for regulatory workflow. Optimized for speed and accuracy of SNPs.

- **SNVPhyl**
  - Public Health Agency of Canada. Graphical User Interface in Galaxy.

Each bioinformatician to have their own personal short-read aligner by 2016

Each bioinformatician to have their own personal short-read aligner by 2016

Each bioinformatician to have their own personal short-read aligner by 2016

Each bioinformatician to have their own personal short-read aligner by 2016

https://thescienceweb.wordpress.com/2015/03/23/each-bioinformatician-to-have-their-own-personal-short-read-aligner-by-2016/
Installation and sample run

$ cd ~/bin/
$ git clone https://github.com/lskatz/lyve-SET
$ cd Lyve-SET
$ git checkout v1.1.4f
$ make install
$ export PATH=$PATH:~/bin/lyve-SET/scripts
# You may also add this to your bash profile
$ echo >> ~/.bash_profile "export PATH=$PATH:~/bin/lyve-SET/scripts"
$ which launch_set.pl
$ set_test.pl lambda lambda --numcpus 4
# Takes about two minutes
$ ls lambda/msa/tree.dnd
Comparison of Lyve-SET with other SNP pipelines

### L. monocytogenes

<table>
<thead>
<tr>
<th>Pipeline</th>
<th>y = mx + b</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>kSNP</td>
<td>y = 0.26x + 24</td>
<td>0.69</td>
</tr>
<tr>
<td>RealPhy</td>
<td>y = 1.14x + 31</td>
<td>0.96</td>
</tr>
<tr>
<td>SNP-Pipeline</td>
<td>y = 1.8x - 13</td>
<td>0.97</td>
</tr>
<tr>
<td>SNVPhyl</td>
<td>y = 0.27x + 19</td>
<td>0.58</td>
</tr>
</tbody>
</table>

### S. enterica

<table>
<thead>
<tr>
<th>Pipeline</th>
<th>y = mx + b</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>kSNP</td>
<td>y = 0.11x + 47</td>
<td>0.23</td>
</tr>
<tr>
<td>RealPhy</td>
<td>y = 0.92x - 5</td>
<td>0.95</td>
</tr>
<tr>
<td>SNP-Pipeline</td>
<td>y = 1.0x + 5.4</td>
<td>0.96</td>
</tr>
<tr>
<td>SNVPhyl</td>
<td>y = 0.91x - 5.1</td>
<td>0.94</td>
</tr>
</tbody>
</table>

### E. coli

<table>
<thead>
<tr>
<th>Pipeline</th>
<th>y = mx + b</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>kSNP</td>
<td>y = 1.1x + 2.9</td>
<td>0.43</td>
</tr>
<tr>
<td>RealPhy</td>
<td>y = 0.78x + 39</td>
<td>0.27</td>
</tr>
<tr>
<td>SNP-Pipeline</td>
<td>y = 1.2x + 58</td>
<td>0.3</td>
</tr>
<tr>
<td>SNVPhyl</td>
<td>y = 0.69x + 2.1</td>
<td>0.92</td>
</tr>
</tbody>
</table>

### C. jejuni

<table>
<thead>
<tr>
<th>Pipeline</th>
<th>y = mx + b</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>kSNP</td>
<td>y = 0.23x + 4</td>
<td>0.89</td>
</tr>
<tr>
<td>RealPhy</td>
<td>y = 0.4x - 15</td>
<td>0.88</td>
</tr>
<tr>
<td>SNP-Pipeline</td>
<td>y = 1.6x - 17</td>
<td>0.97</td>
</tr>
<tr>
<td>SNVPhyl</td>
<td>y = 0.18x + 49</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Each data point is a SNP distance as determined by Lyve-SET (x-axis) and the distance of an alternative SNP pipeline (y-axis). The slope indicates the number of SNPs per Lyve-SET SNP.
SNPs overlayed on MLST loci

Strain A

Strain B

Strain C
Comparison with whole-genome MLST (Listeria monocytogenes only)

Which algorithm should you use?

<table>
<thead>
<tr>
<th></th>
<th>Kmer-based</th>
<th>wgMLST</th>
<th>hqSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity</td>
<td>✔️ ✔️</td>
<td>✔️</td>
<td>✗ ✗</td>
</tr>
<tr>
<td>Outbreak-level resolution</td>
<td>✗</td>
<td>✔️ ✔️</td>
<td>✔️ ✔️</td>
</tr>
<tr>
<td>Further genomic information</td>
<td>✗</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Minimal upfront effort</td>
<td>✔️</td>
<td>✗ ✗</td>
<td>✔️</td>
</tr>
<tr>
<td>Fast</td>
<td>✔️ ✔️</td>
<td>✔️ ✔️</td>
<td>✗</td>
</tr>
<tr>
<td>Easy to use for anyone</td>
<td>✗</td>
<td>✔️</td>
<td>✗</td>
</tr>
</tbody>
</table>
Fun examples

More information can be found in the virtual lab talk given on Jan 7, 2019: https://youtu.be/YPnU63Le53Y?t=1234
Multistate outbreak of farmstead cheeses

Listeria (Listeriosis)

Listeria (Listeriosis) » Outbreaks

Multistate Outbreak of Listeriosis Linked to Crave Brothers Farmstead Cheeses (Final Update)

Posted September 24, 2013 1:00 PM ET

This outbreak appears to be over. Listeria monocytogenes infection (listeriosis) is an important cause of illness in the United States. More information about listeriosis, and steps people can take to reduce their risk of infection, can be found on the CDC Listeria Web Page.

Highlights

- Read the Advice to Consumers & Cheese Retailers:
- A total of six persons infected with the outbreak strain of Listeria monocytogenes were reported from five states:
  - The number of ill persons identified in each state was as follows: Illinois (1), Indiana (1), Minnesota (2), Ohio (1), and Texas (1).
- All six ill persons were hospitalized. One death was reported in Minnesota. In addition, one illness in a pregnant woman resulted in a miscarriage.
- No new ill persons were reported since the last update on August 22, 2013.
- A collaborative investigation by local and state public health and regulatory agencies, CDC, and the U.S. Food and Drug Administration (FDA) indicated that Les Frères, Petit Frère, and Petit Frère with Truffles

At a Glance:
- Case Count: 6
- States: 5
- Deaths: 1
- Hospitalizations: 6
- Recall: Yes

More Information:
- Recall & Advice to Consumers
- Signs & Symptoms
- Key Resources

Contact Us:
- Centers for Disease Control and Prevention
  1600 Clifton Rd
  Atlanta, GA 30333
- 800-CDC-INFO (800-232-4636)
  TTY: (888) 222-6348
  New Hours of Operation
  8am-8pm ET, Monday-Friday
  Closed Holidays
- cdcinfo@cdc.gov
How to read a phylogeny

Scale bar
0.01

Direction of evolution

Outgroup1

Outgroup2

Hypothetical last common ancestor (LCA) of taxa 3-7

Vertical distance is irrelevant

Root (LCA of all taxa)

Percent confidence in hypothetical ancestor

Hypothetical last common ancestor (LCA) of taxa 3-7

taxon

taxon

taxa
2013 outbreak linked to farmstead cheese

Red = epi-related clinical isolates
Blue = retrospective clinical cases or not outbreak related
Green = historical environmental isolates from the plant

★ Exposure
★ No Exposure
Phylogenetically related outbreak of unknown etiology, December 2013

1a

1ai

2013L-5
2013L-5
2013L-5

1aii
cold smoked white fish recall 1hqSNP [0-2]

kale

NYC

2013L-5
2013L-5

2013L-5
2013L-5

Between 1ai and 1aii: 14 hqSNPs [13-16]

10.5 hqSNPs [0-22]

48 hqSNPs [0-64]

Farmstead cheese outbreak July 2013
In conclusion

- WGS provides high resolution
- We have many tools for differing levels of resolution
- We can and have used it on outbreak investigations
Questions?

Lee Katz

gzu2@cdc.gov

lskatz

github.com/lskatz