Genomic Epidemiology

Lee Katz, Ph.D.

Senior bioinformatician
Enteric Diseases Laboratory Branch

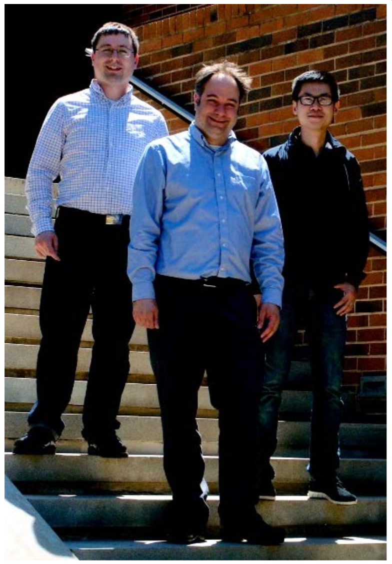
Computational Genomics course

Jan 31, 2018

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)

THIS IS THE 11TH YEAR OF THIS CLASS

April 16, 2008

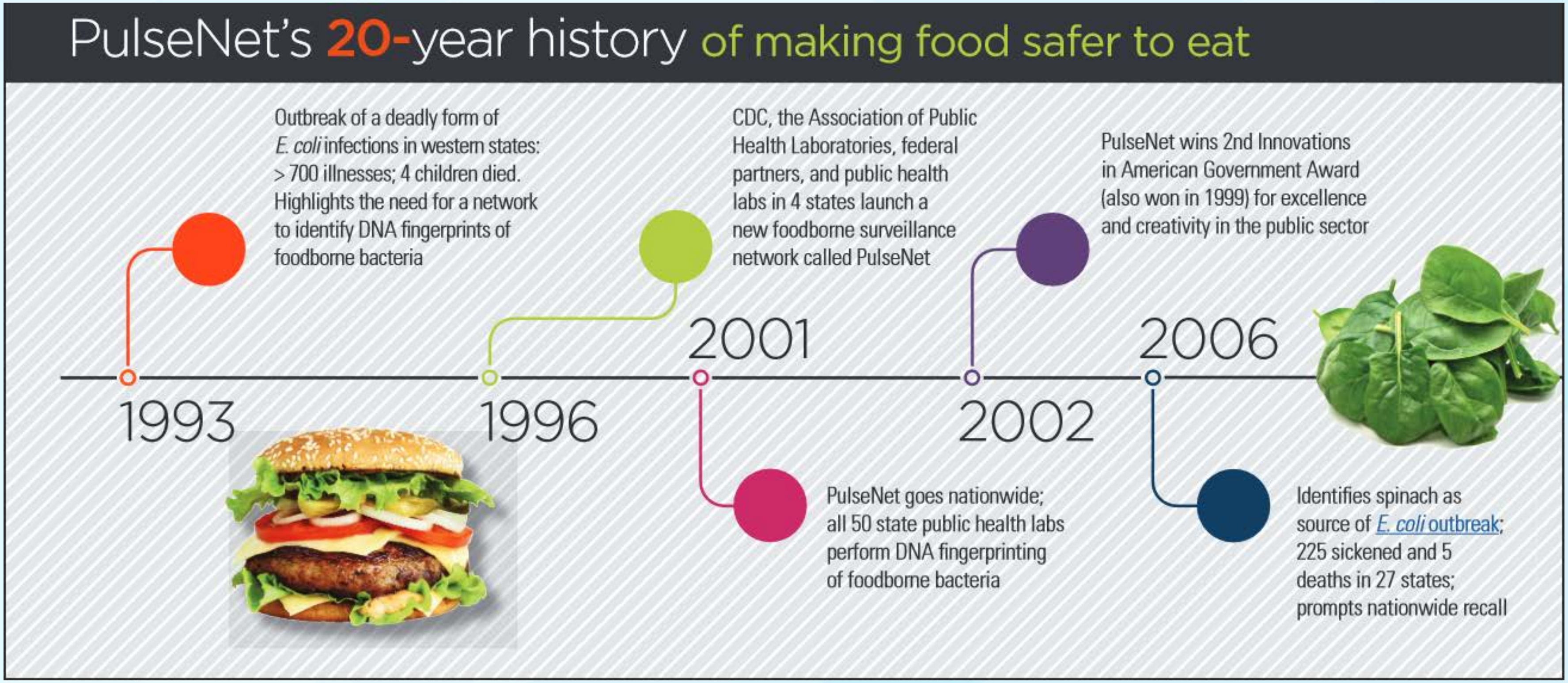
http://www.compgenomics.biology.gatech.edu/index.php/Group_photos

April 23, 2008

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

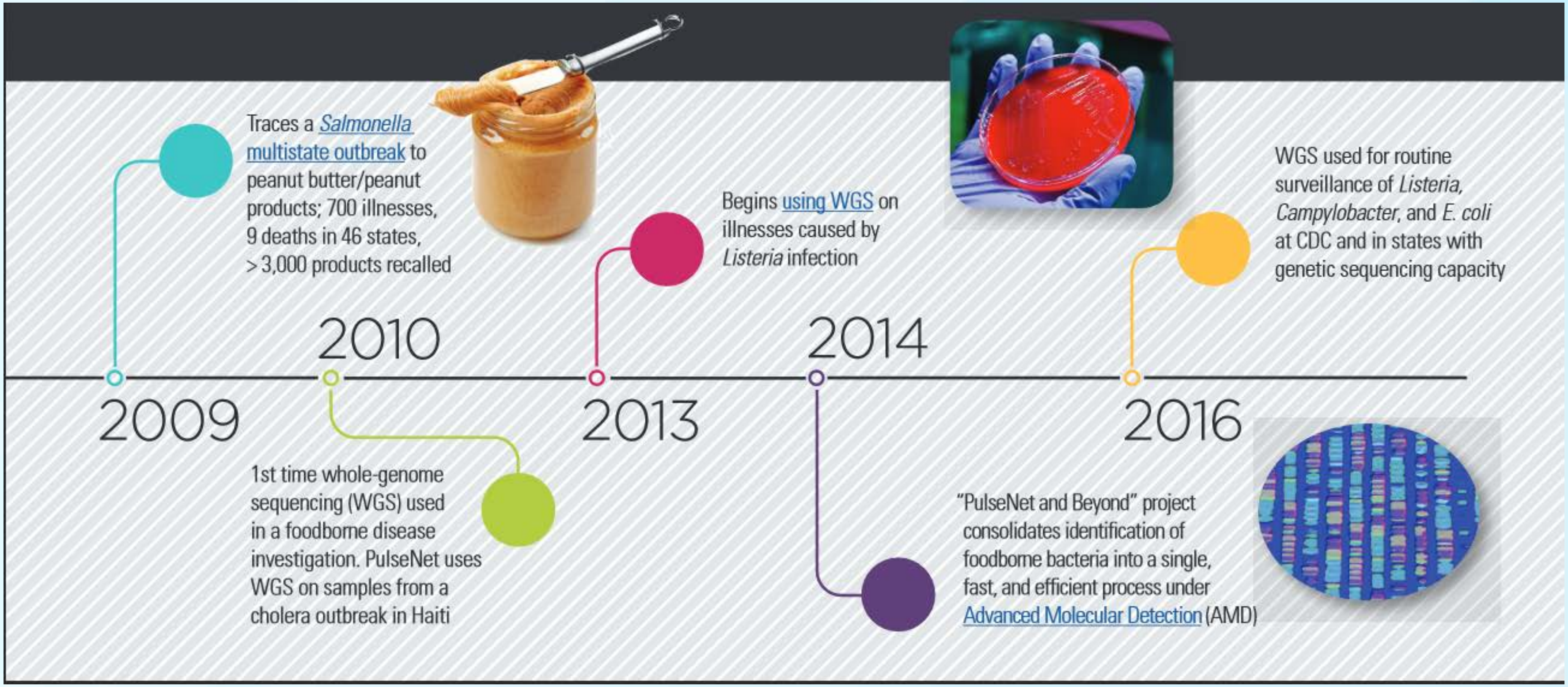
PulseNet goes nationwide; all 50 state public health labs perform DNA fingerprinting of foodborne bacteria

2002

PulseNet wins 2nd Innovations in American Government Award (also won in 1999) for excellence and creativity in the public sector

2006

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



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

“PulseNet and Beyond” project consolidates identification of foodborne bacteria into a single, fast, and efficient process under Advanced Molecular Detection (AMD)

2016

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

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

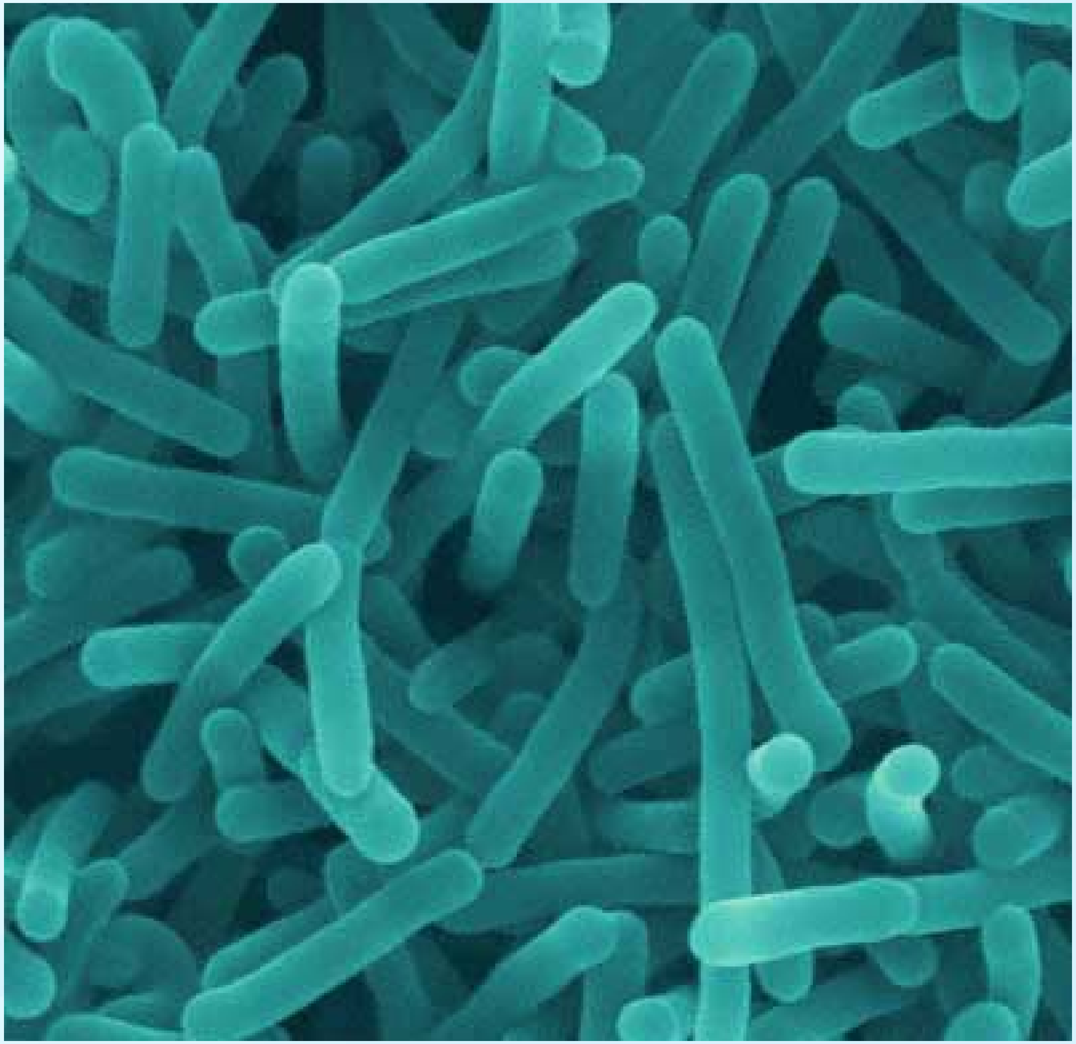
Estimated $2.8 billion in annual medical costs and lost productivity ($1.8 million/case)

Current subtyping methods are not ideal

Strong epidemiologic surveillance (Listeria Initiative)

Strong regulatory component

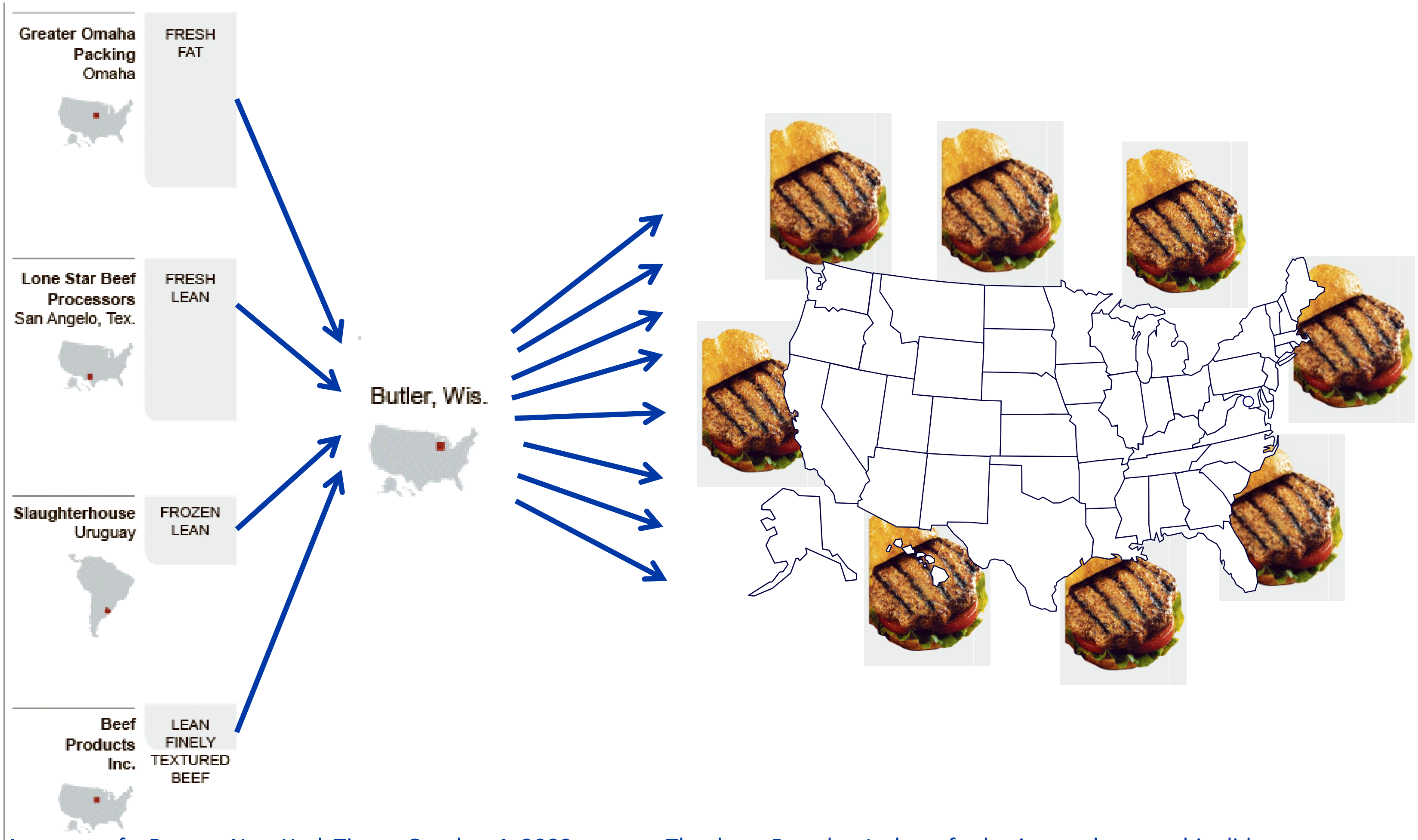
Listeria genome is fairly small, stable, and relatively easy to sequence and analyze. Most changes in the genome are due to point mutations and not phages.



Thanks to Brendan Jackson for letting me borrow this slide

Photo credit: http://www.cdc.gov/media/dpk/2013/dpk-eis-conference.html

The Problem: Detecting Outbreaks in an Increasingly Globalized Food System



Greater Omaha Packing
Omaha

FRESH FAT

Lone Star Beef Processors
San Angelo, Tex.

FRESH LEAN

Slaughterhouse
Uruguay

FROZEN LEAN

Beef Products Inc.

LEAN FINELY TEXTURED BEEF

Butler, Wis.

Anatomy of a Burger. New York Times. October 4, 2009

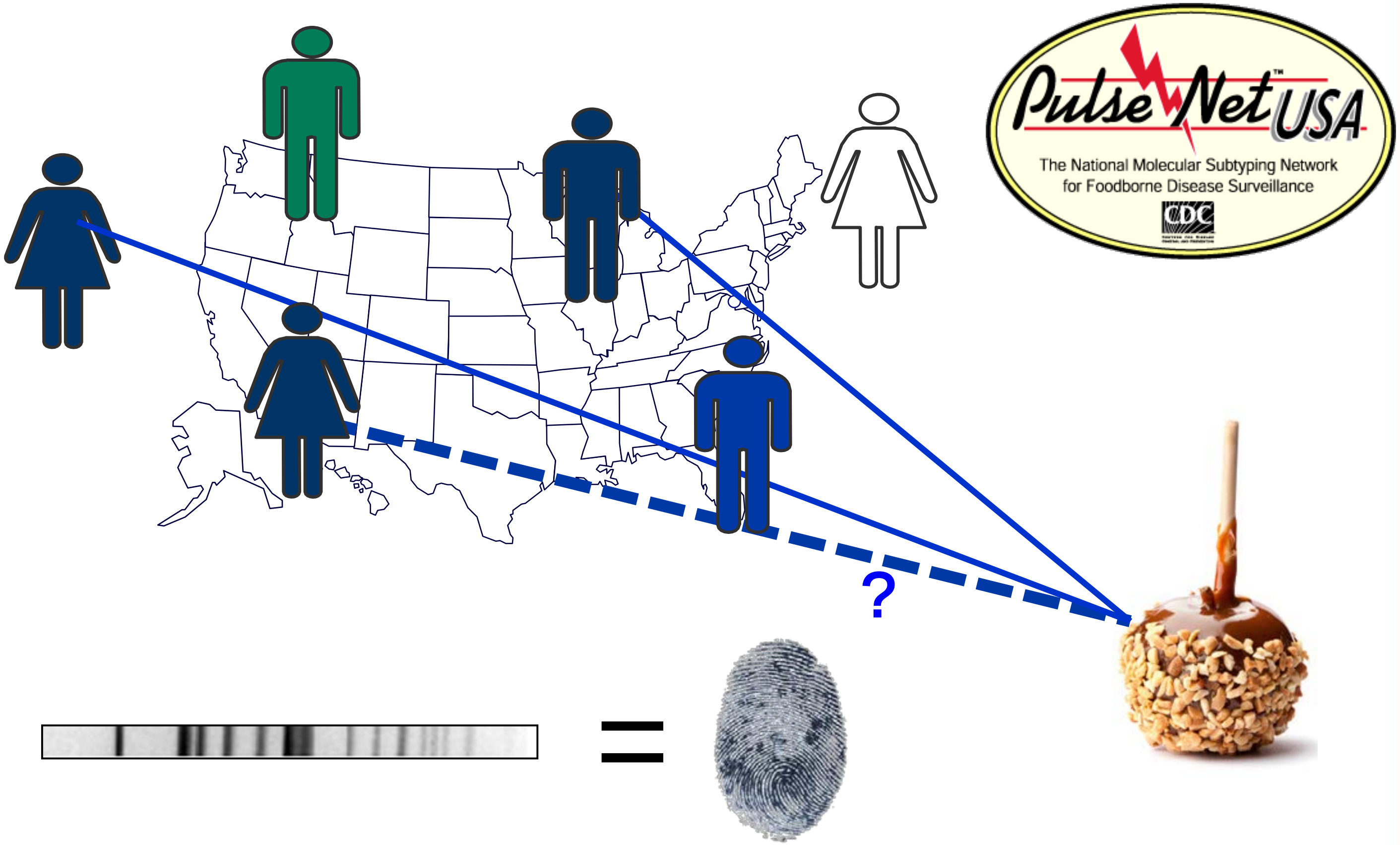
Thanks to Brendan Jackson for letting me borrow this slide

Limitations of
Pulsed-Field Gel Electrophoresis (PFGE)

PulseNet USA
The National Molecular Subtyping Network
for Foodborne Disease Surveillance
CDC
=

Thanks to Brendan Jackson for letting me borrow this slide

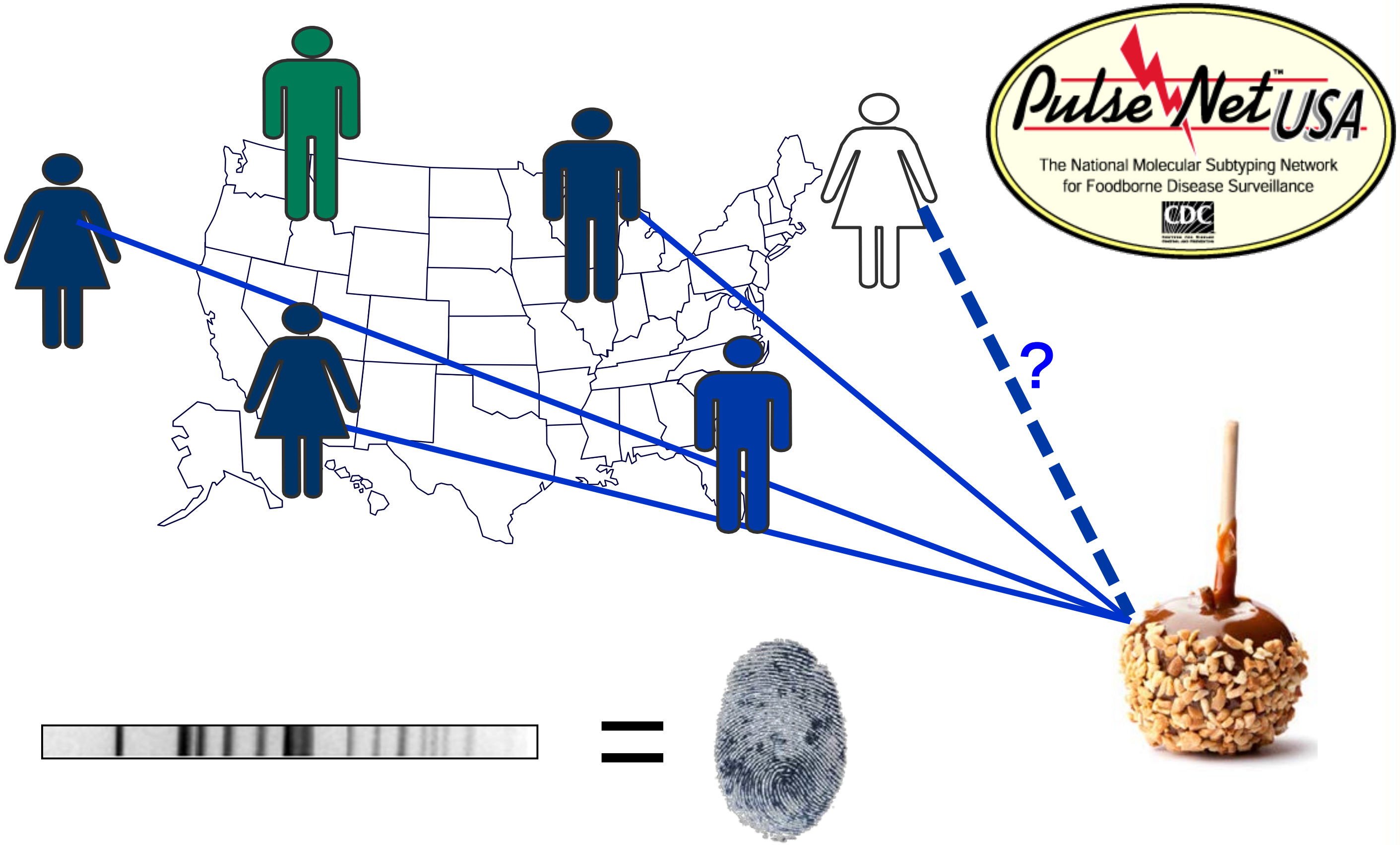
Limitation: Genetically Unrelated Isolate Might Appear Same by PFGE


PulseNet USA
The National Molecular Subtyping Network for Foodborne Disease Surveillance
CDC
?
=

PFGE is correlated with epidemiology but is not perfect

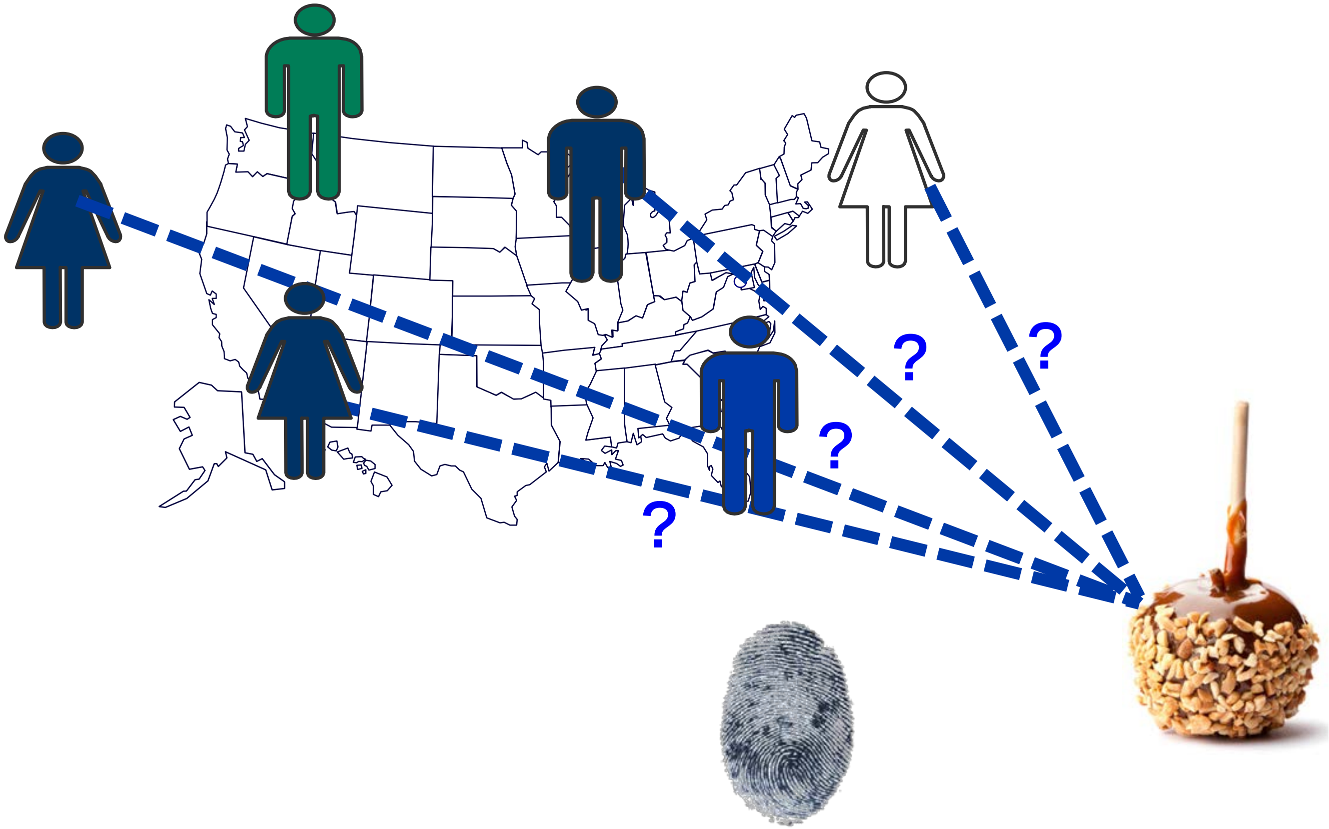
Thanks to Brendan Jackson for letting me borrow this slide

Limitation: Genetically Related Isolate Might Appear Different By PFGE


PulseNet USA
The National Molecular Subtyping Network for Foodborne Disease Surveillance
CDC
?
=

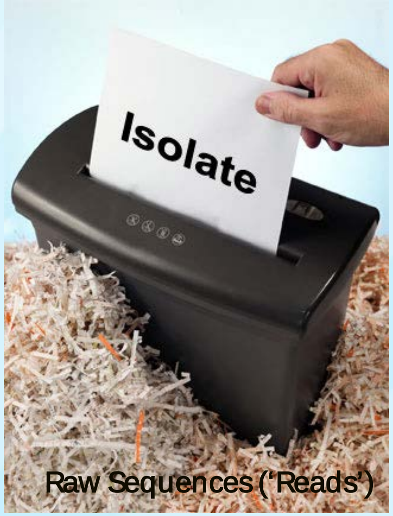
Thanks to Brendan Jackson for letting me borrow this slide

Can genomics clear up this picture?

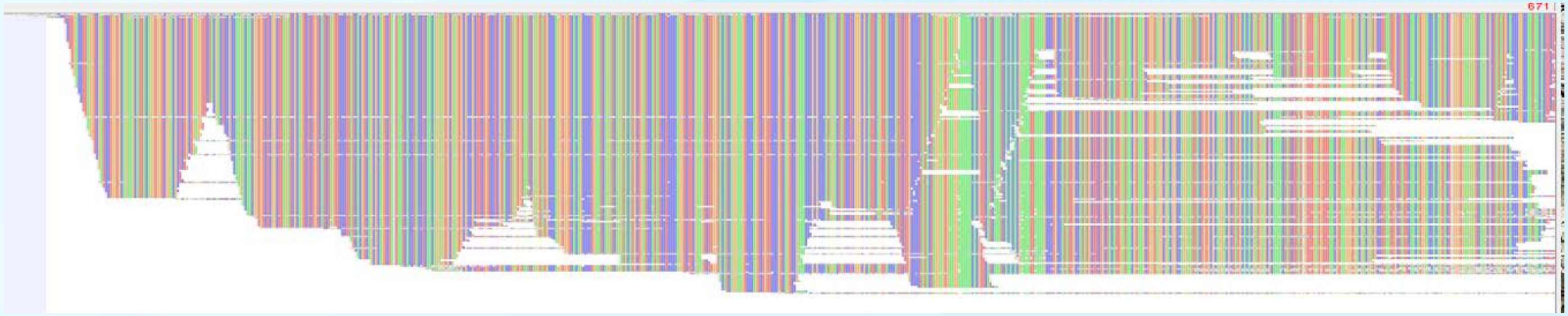


The Basics of Next Generation Sequencing (NGS)

- **“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**



The Basics of Whole Genome Sequencing (WGS)

- **Assembling and annotating the sequence**
  - Solving the puzzle using an 'assembler' software

'Reference -Based Assembly

'*de novo* Assembly'

- The puzzle usually only solved 97- 99%
  - So, even though we say 'whole genome', we don't mean that!
- Assembled in 1- 200 (- 500) fragments ('contigs')

HOW DO WE COMPARE GENOMES?

Three major methods we use

- Kmer-based: mile-high view
- MLST-based: naked eye
- SNP-based: microscope

- 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

CAAAAAAAAAAAAT

CAAAAAAAAAAAAG

Here, K=12

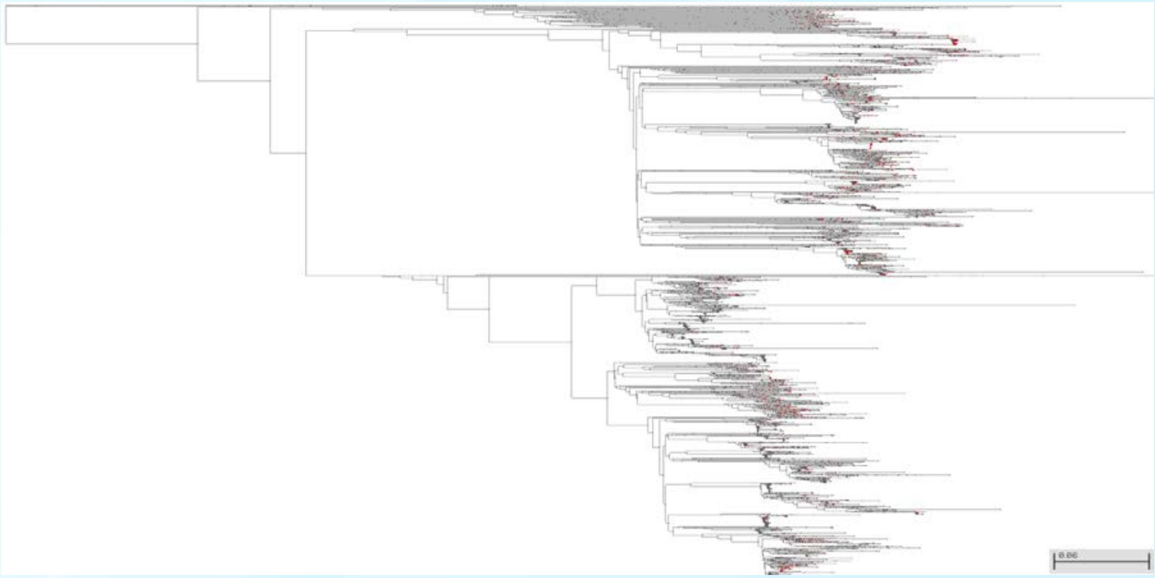
CAAAAAAAAAAA	1
AAAAAAAAAAAA	2
AAAAAAAAAAAT	3

1	CAAAAAAAAAAA
2	AAAAAAAAAAAA
4	AAAAAAAAAAAG

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

Example kmer tree

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



http://www.ncbi.nlm.nih.gov/pathogens/

Software: pathogen detection pipeline at NCBI

Kmer-based software

NCBI Pathogen Detection Pipeline

Not available for individual use, but the results are comprehensive and public

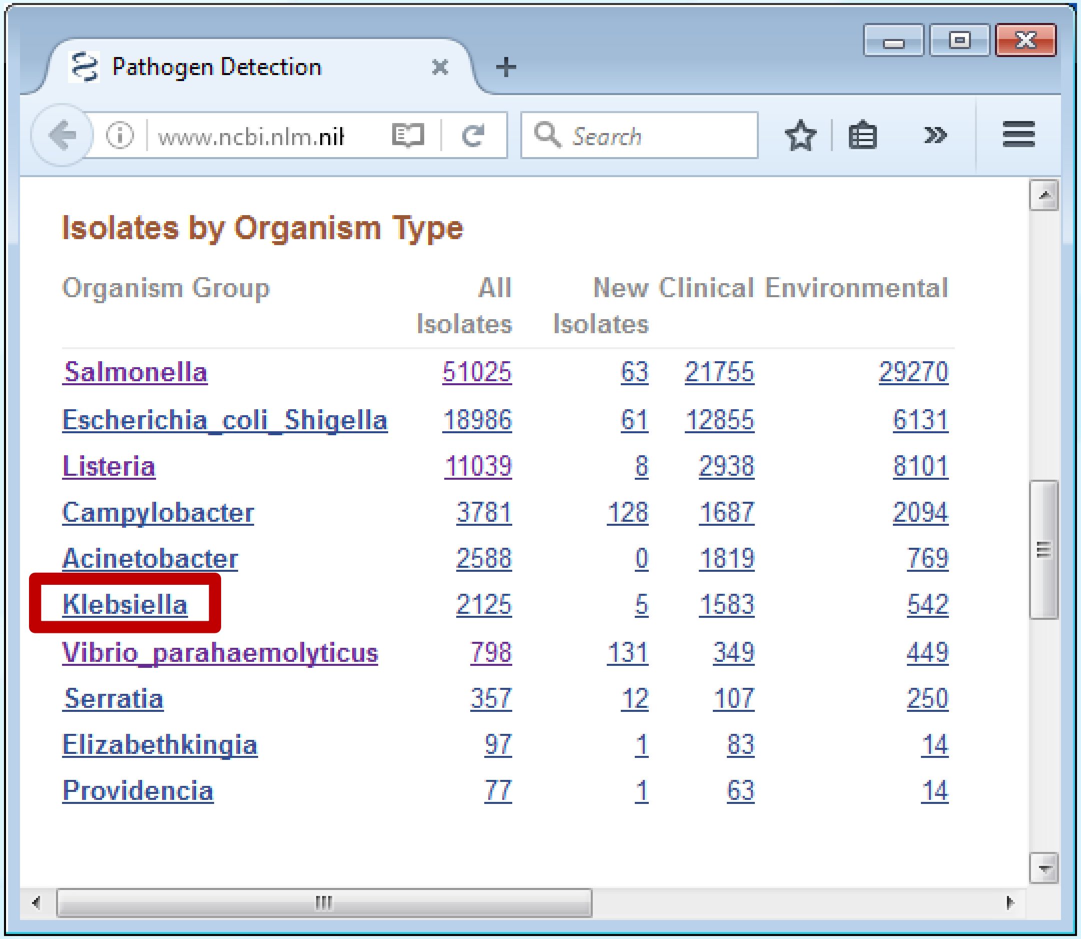
Mashtree

Documentation and installation instructions are at https://github.com/lskatz/mashtree

Built on top of **Mash**

KSNP

Alignments of the middle base in kmers. Arguably, KSNP is actually a SNP pipeline instead.



Isolates by Organism Type

Organism Group	All Isolates	New Isolates	Clinical	Environmental
Salmonella	51025	63	21755	29270
Escherichia_coli_Shigella	18986	61	12855	6131
Listeria	11039	8	2938	8101
Campylobacter	3781	128	1687	2094
Acinetobacter	2588	0	1819	769
Klebsiella	2125	5	1583	542
Vibrio_parahaemolyticus	798	131	349	449
Serratia	357	12	107	250
Elizabethkingia	97	1	83	14
Providencia	77	1	63	14

NCBI kmer trees screen shot taken Sept 23, 2016

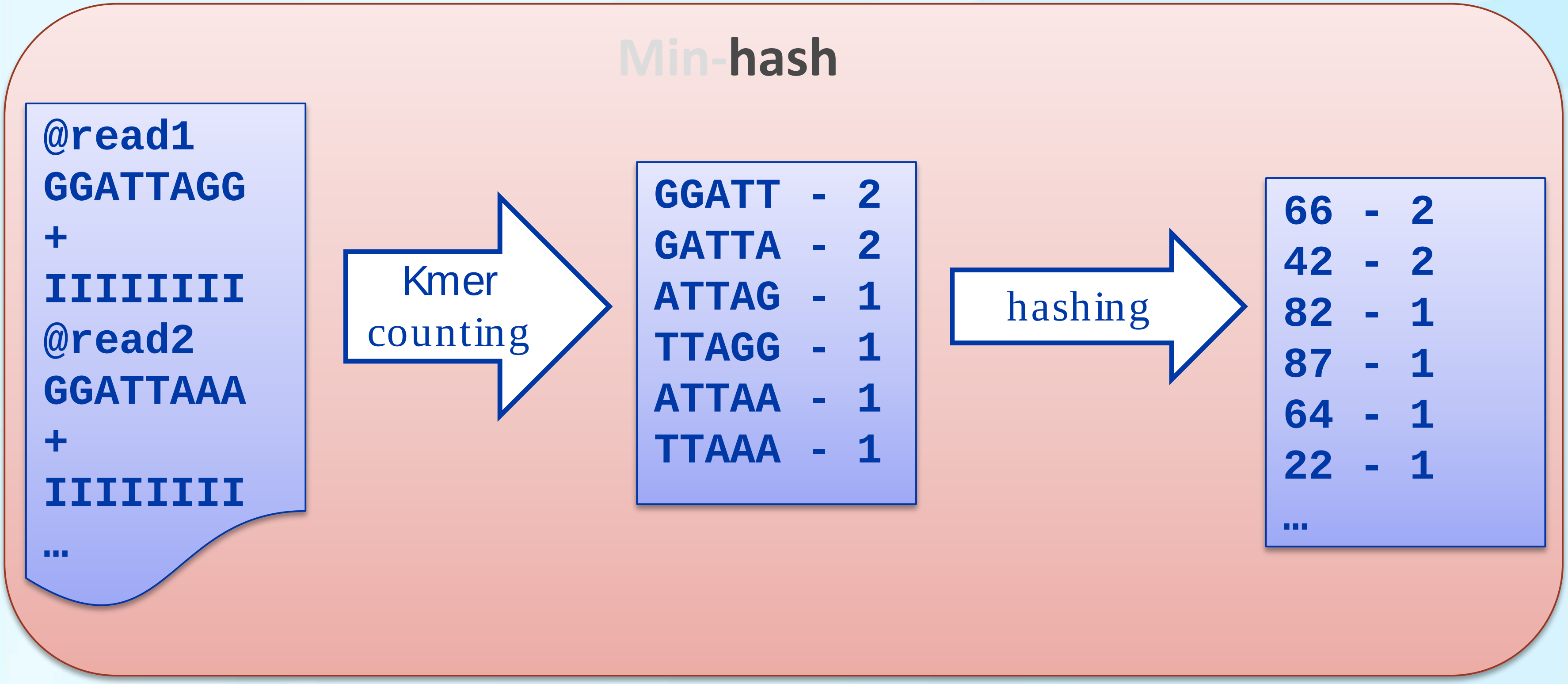
https://www.ncbi.nlm.nih.gov/pathogens

How does Mash work?

Based on the software Mash

Mash implements the **min-hash** algorithm for sequence data

“Sketch”



Min-hash

```
@read1
GGATTAGG
+
IIIIIIII
@read2
GGATTAAA
+
IIIIIIII
…
```

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

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

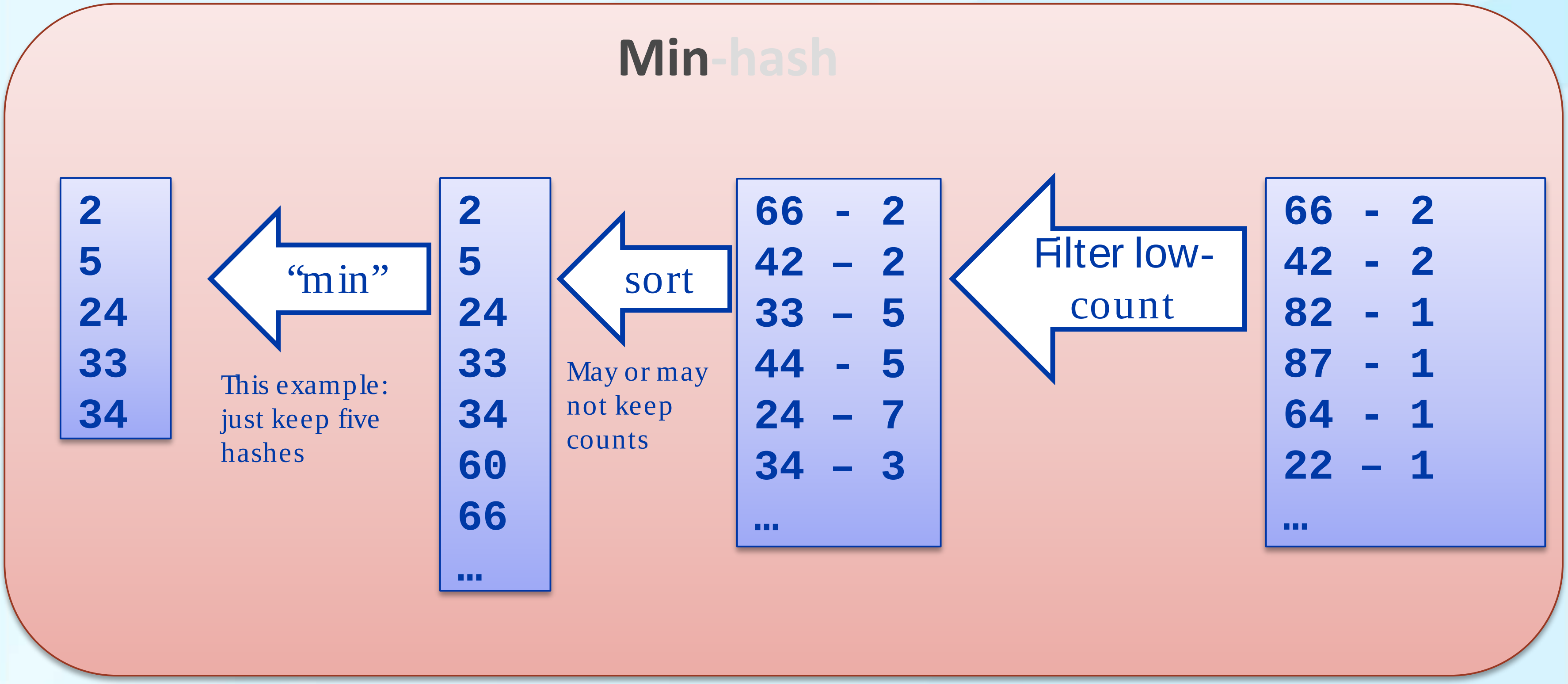
Ondov et al, “Mash”, Genome Biology. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0997-x

How does Mash work?

Based on the software Mash

Mash implements the **min-hash** algorithm for sequence data

“Sketch”



Min-hash

2
5
24
33
34

“min”

This example: just keep five hashes

2
5
24
33
34
60
66
…

sort

May or may not keep counts

66 - 2
42 – 2
33 – 5
44 - 5
24 – 7
34 – 3
…

Filter low-count

66 - 2
42 - 2
82 - 1
87 - 1
64 - 1
22 – 1
…

Ondov et al, “Mash”, Genome Biology. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0997-x

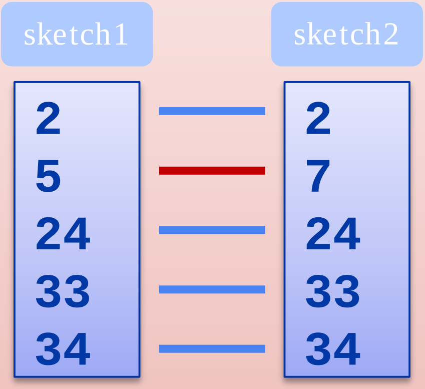
How does Mash work?

Based on the software Mash

Mash implements the **min-hash** algorithm for sequence data

“Distance” or “dist”

Min-hash



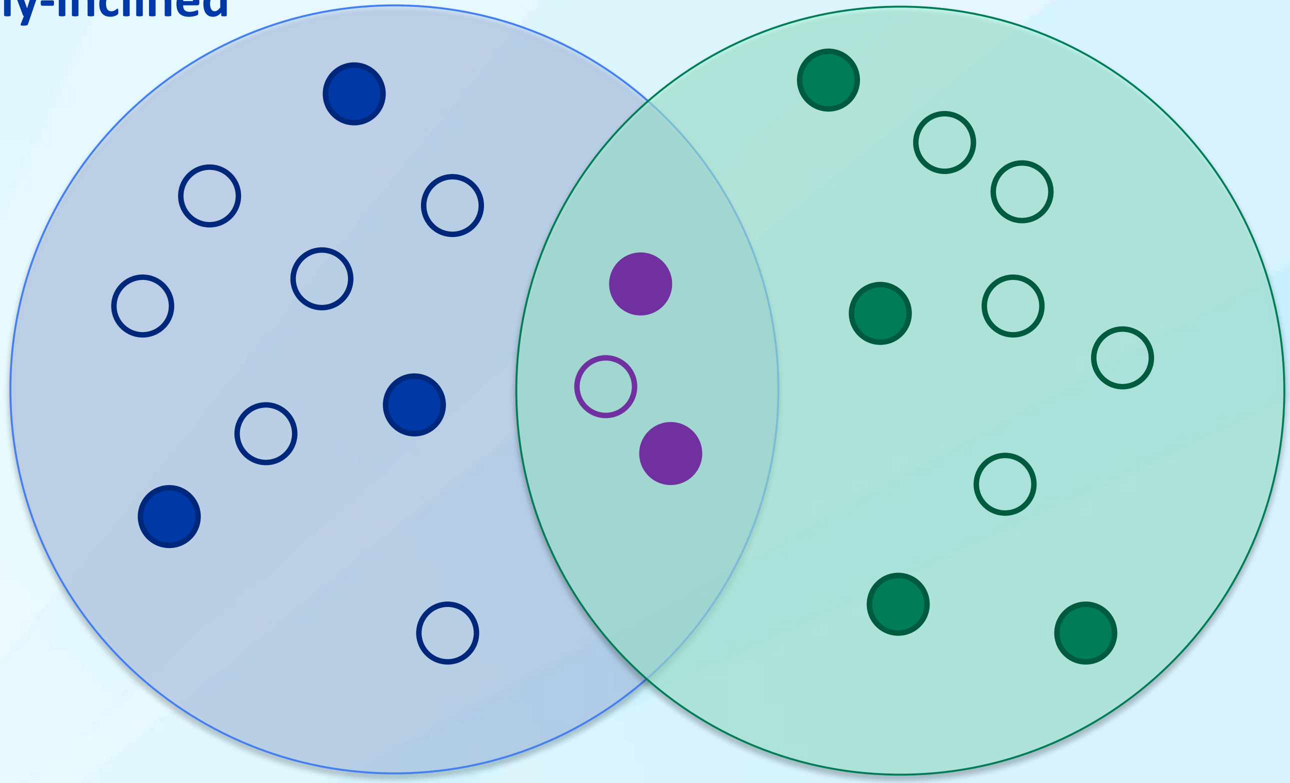
sketch1	sketch2
2	2
5	7
24	24
33	33
34	34

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

The resolution gets better with more hashes.

Ondov et al, “Mash”, Genome Biology. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0997-x

Min-hash for the more mathematically-inclined



A = ● + ○

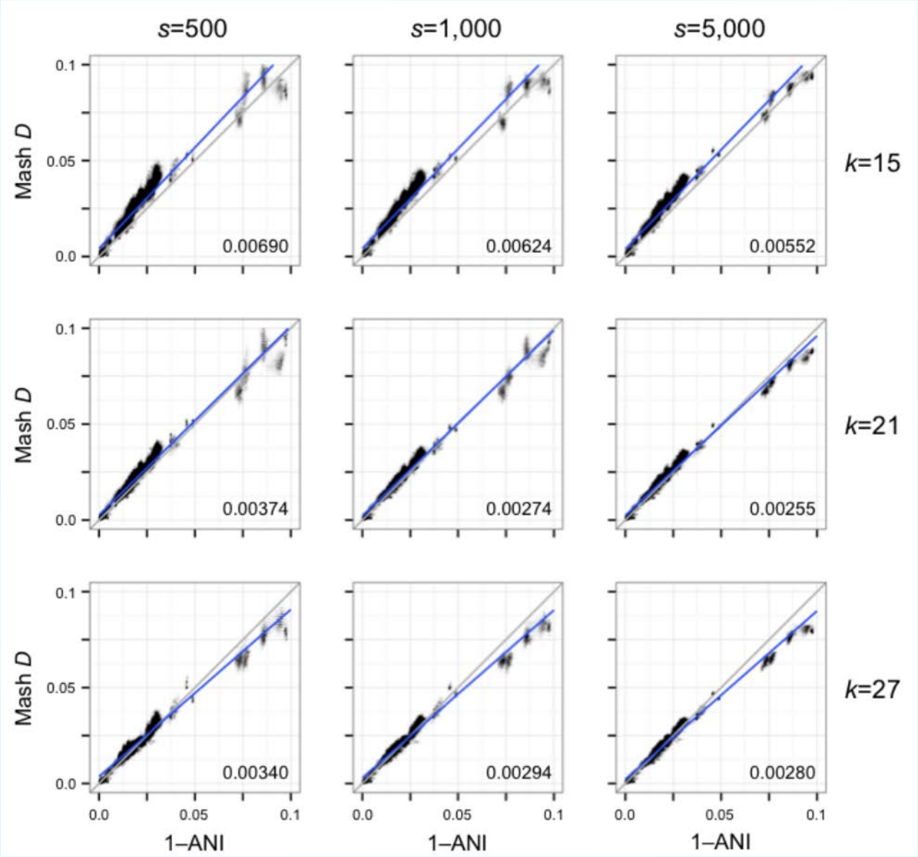
S(A) = ●

B = ● + ○

S(B) = ●

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

Comparison to ANI



k=kmer length
s=sketch size, ie, number of hashes

Values on the bottom right of graphs indicate the root-mean-square error when comparing Mash vs ANI.

Mash time trials

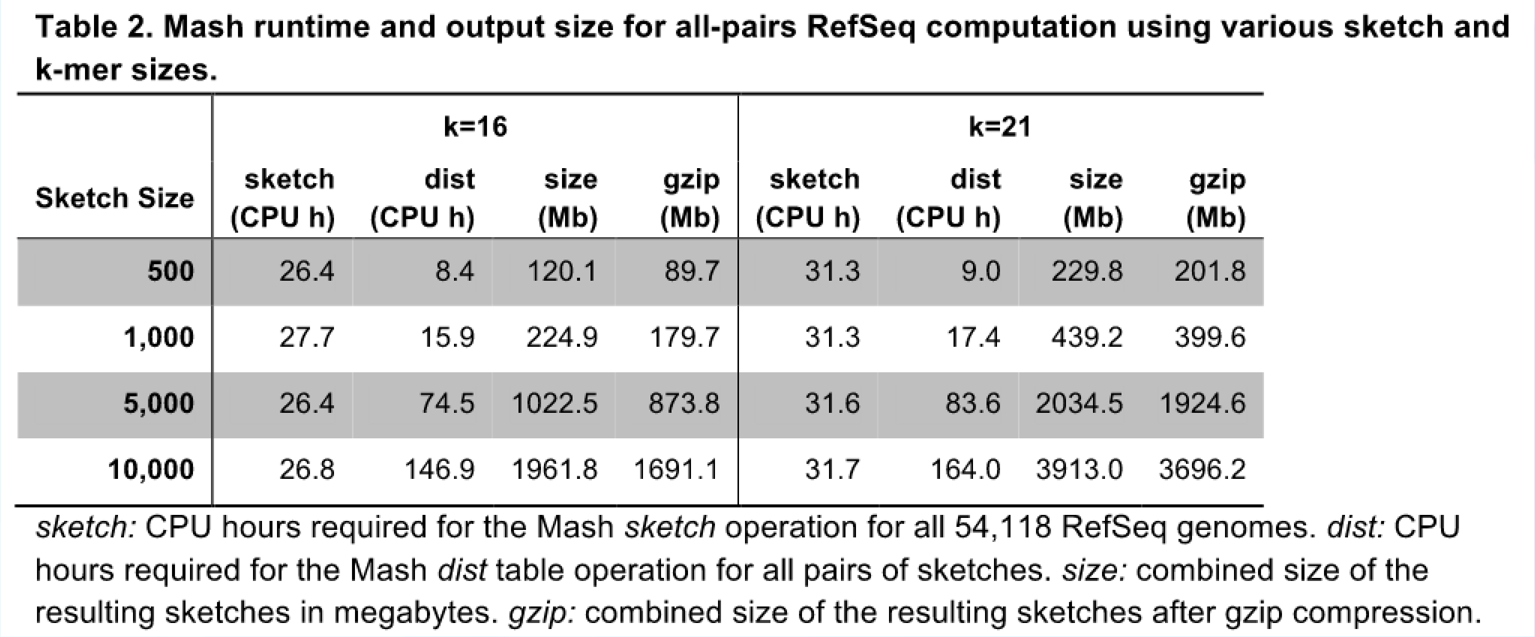


Table 2. Mash runtime and output size for all-pairs RefSeq computation using various sketch and k-mer sizes.

	k=16				k=21			
Sketch Size	**sketch (CPU h)**	**dist (CPU h)**	**size (Mb)**	**gzip (Mb)**	**sketch (CPU h)**	**dist (CPU h)**	**size (Mb)**	**gzip (Mb)**
500	26.4	8.4	120.1	89.7	31.3	9.0	229.8	201.8
1,000	27.7	15.9	224.9	179.7	31.3	17.4	439.2	399.6
5,000	26.4	74.5	1022.5	873.8	31.6	83.6	2034.5	1924.6
10,000	26.8	146.9	1961.8	1691.1	31.7	164.0	3913.0	3696.2

sketch: CPU hours required for the Mash *sketch* operation for all 54,118 RefSeq genomes. *dist:* CPU hours required for the Mash *dist* table operation for all pairs of sketches. *size:* combined size of the resulting sketches in megabytes. *gzip:* combined size of the resulting sketches after gzip compression.

Ondov et al, “Mash”, Genome Biology. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0997-x

MASHTREE

Useful for creating a dendrogram quickly

Does not create a phylogeny

https://github.com/lskatz/mashtree

2+2

What it is and what it isn’t

Is	Isn’t
Builds trees	Infers phylogeny
Fast	Slow

When to use it

Use it when	Don’t use it when
Need fast estimate	Need solid results
Need to know a good reference genome	Inferring phylogenetic relatedness
Large, diverse dataset	Not diverse or not large dataset

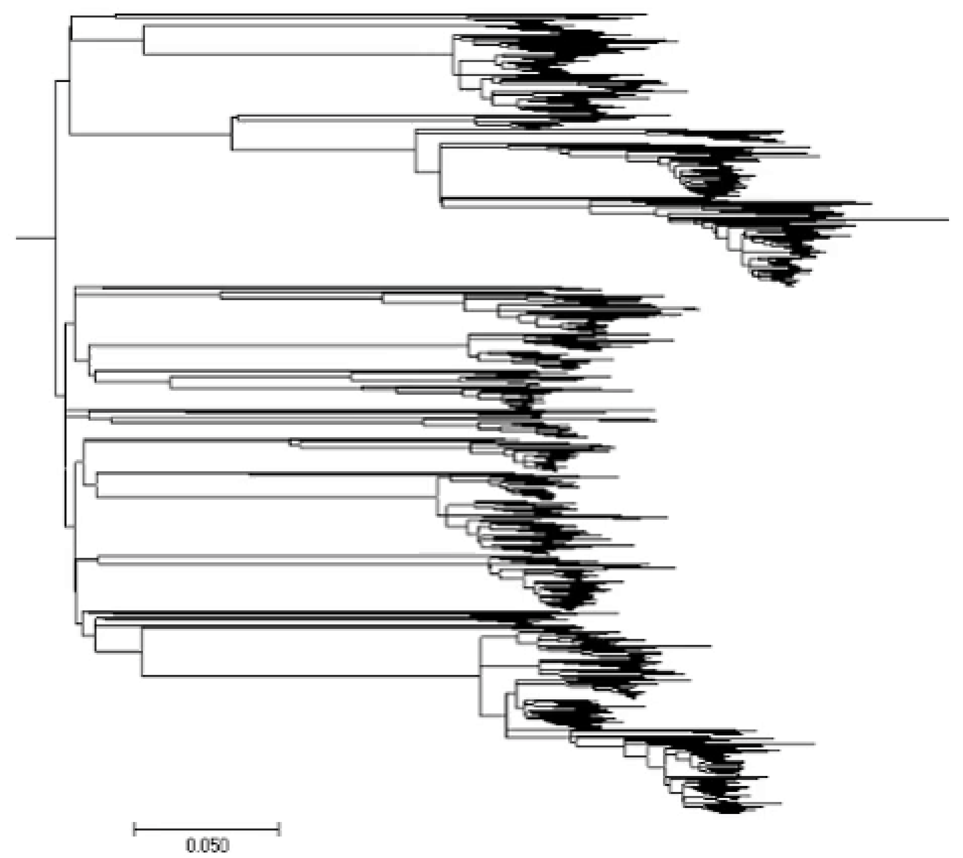
Mashtree is fast

Methods

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.



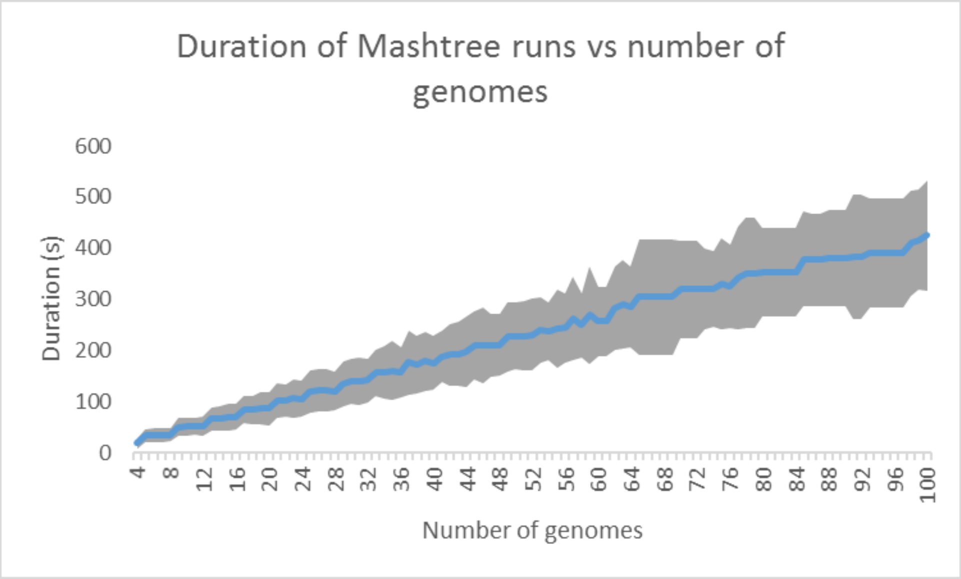
Mashtree is fast

Methods

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.

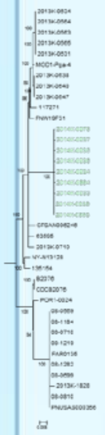
Grey shading indicates the range of durations.



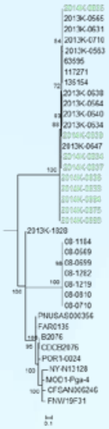
The Mashtree v0.06 tree is usually at least as good as the kSNP3 tree.

*based on only a few comparisons

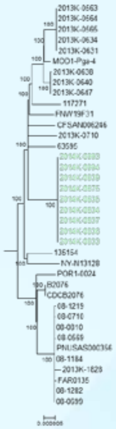
** currently on v0.29



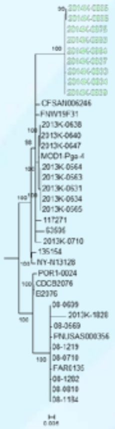
Lyve-SET 100%
Sp =
Sn = 100%



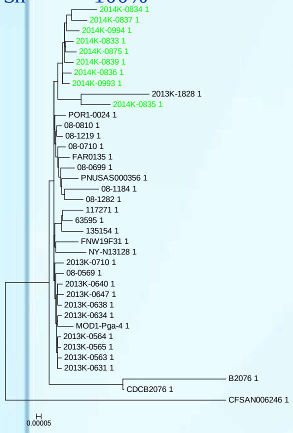
kSNP3 100%
Sp = 58%
Sn =



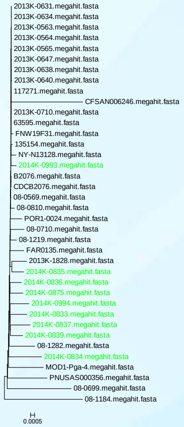
RealPhy 100%
Sp =
Sn = 100%



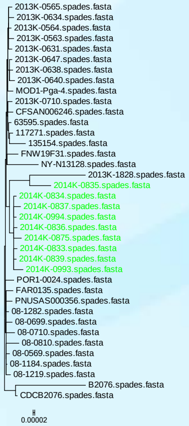
Snp-Pipeline Sn = 100%
Sp = 100%



Mash v0.06 Sn = 100%
Raw reads Sp = 97%
min_depth: 5x



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



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

1409MLJN6-1

$n_{pos} = 9$

$n_{neg} = 29$

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

🟩 part of outbreak

Mashtree is command line

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

# Usage
$ mashtree.pl --help

# Execution
$ mashtree.pl --numcpus 12 --genomesize 5300000 \
  *.fastq.gz \
  [*.fasta] [*.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

Image credit: Wikipedia.org

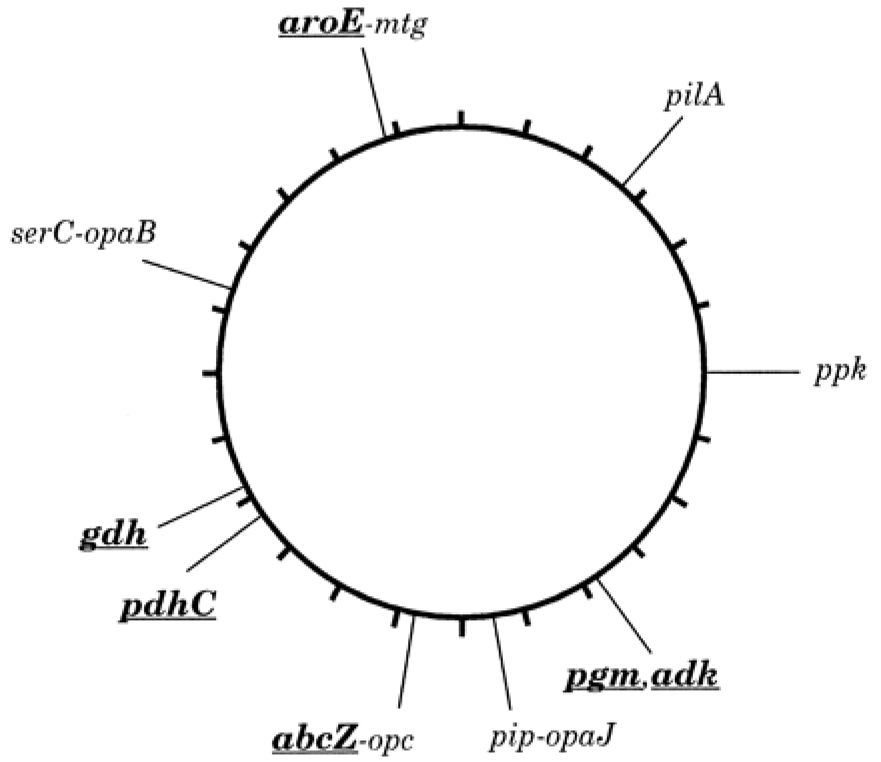
Software: BioNumerics

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)**

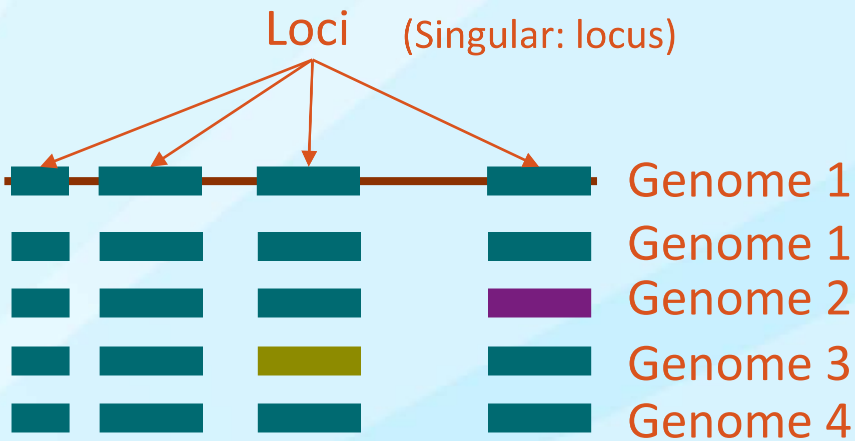


Maiden et al 1998 *PNAS*

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**.



Loci (Singular: locus)

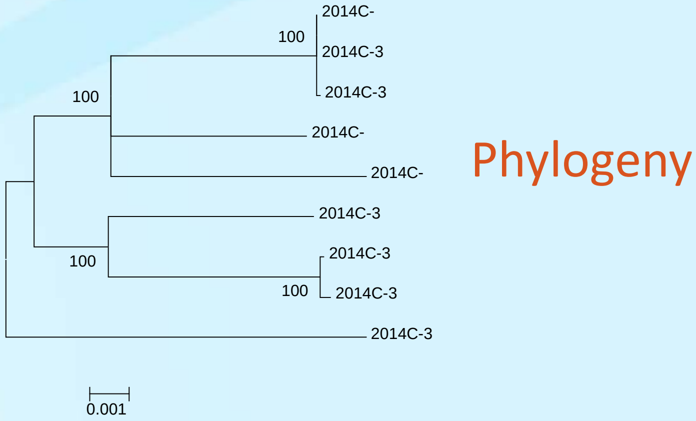
Genome 1

Genome 1

Genome 2

Genome 3

Genome 4



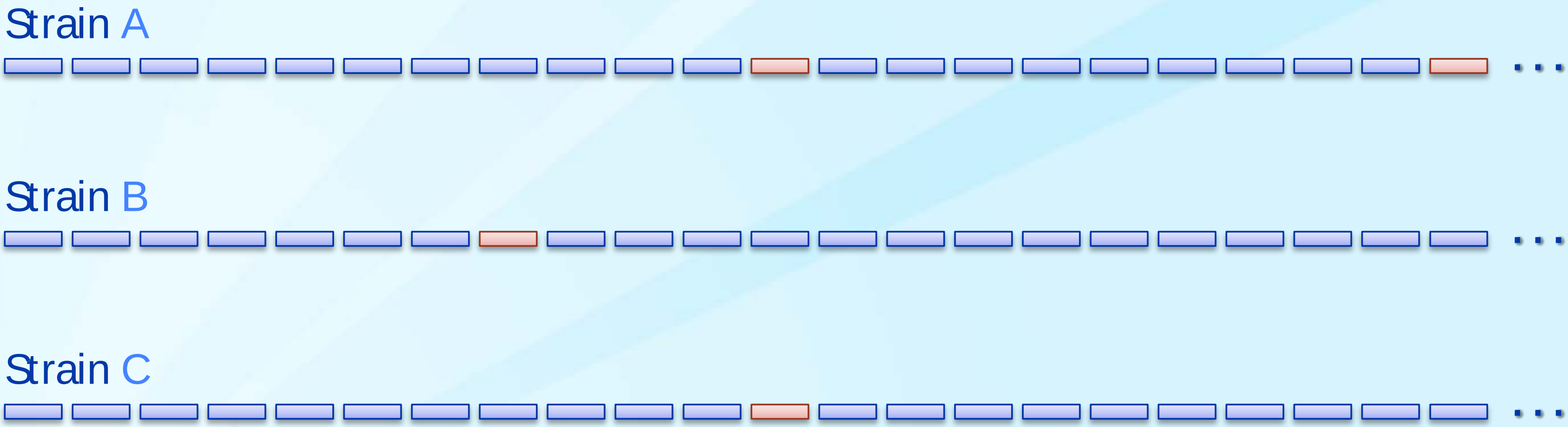
Phylogeny

Whole-genome MLST

Rule of thumb: there is about one locus per 1,000 nucleotides in the genome.

Different species have different sizes, e.g., *L. monocytogenes* has about 3,000,000 nucleotides (~3,000 loci)

In an outbreak, again rule of thumb, we expect 0-10, or perhaps as many as 50 allele differences between genomes.



Strain A

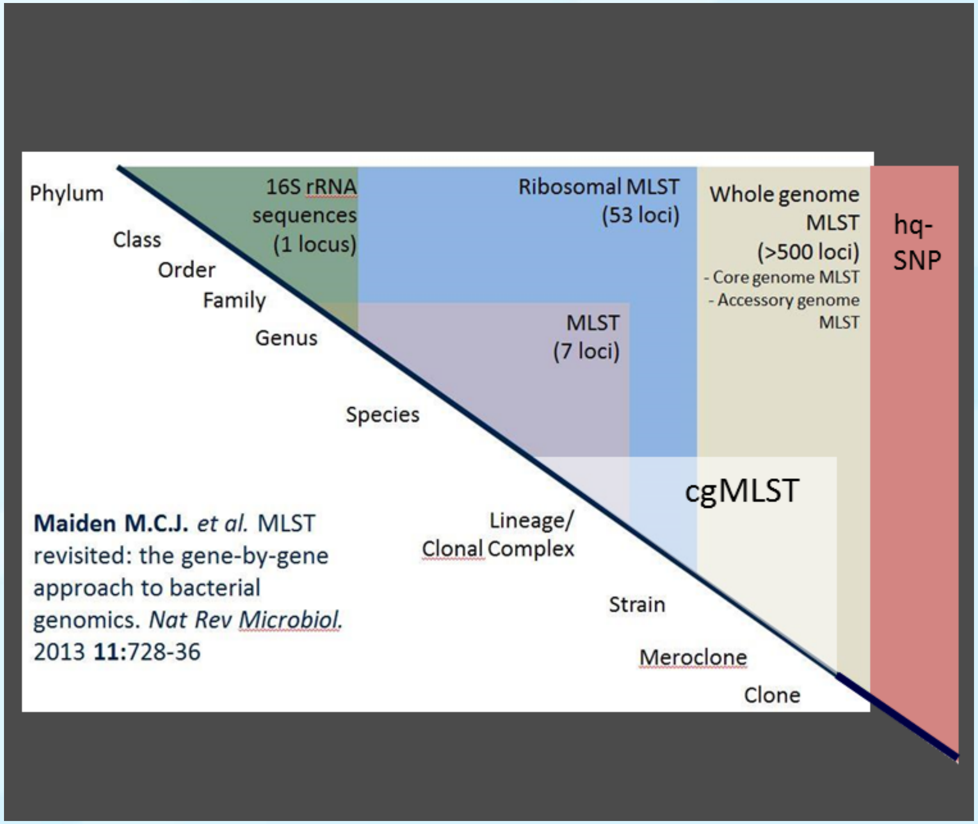
Strain B

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)


Phylum
Class
Order
Family
Genus
Species
Lineage/ Clonal Complex
Strain
Meroclone
Clone
16S rRNA sequences (1 locus)
Ribosomal MLST (53 loci)
Whole genome MLST (>500 loci)
- Core genome MLST
- Accessory genome MLST
hq-SNP
MLST (7 loci)
cgMLST
Maiden M.C.J. et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013 11:728-36

Maiden M.C.J. et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol.* 2013 **11**:728-36

MLST software

BioNumerics

Graphical user interface.

StringMLST

Compare kmers of raw reads against a database.

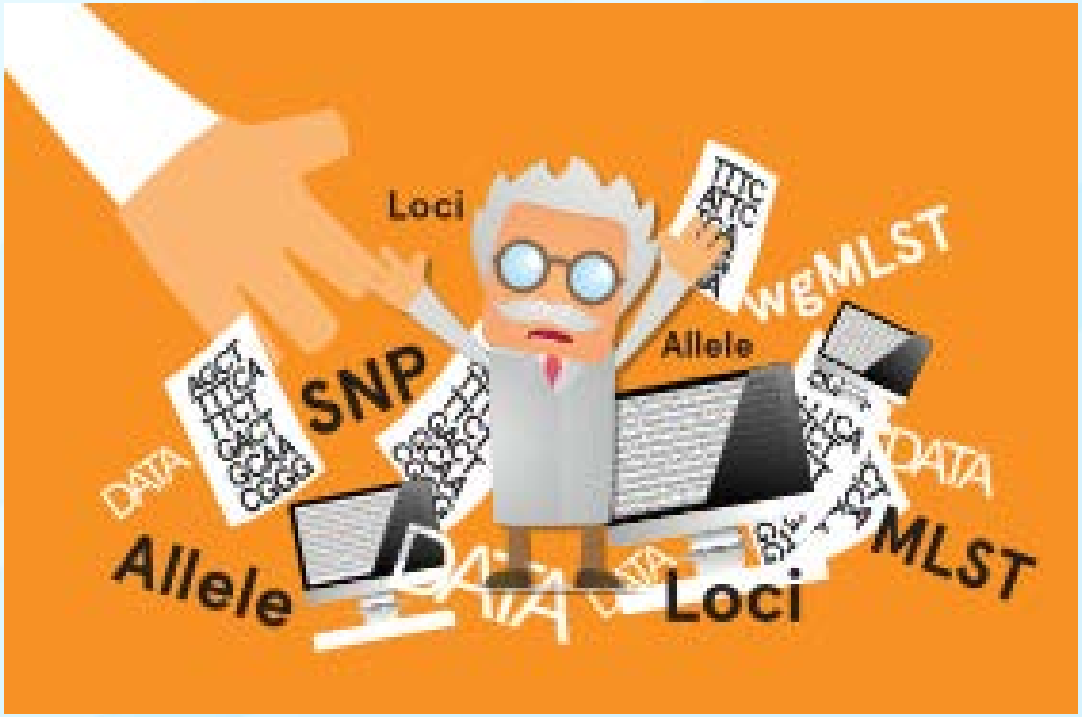
(also: Sting)

Ridom SeqSphere+

Graphical user interface. Mostly used in Europe.

mlst

BLAST genome assembly against database. Not rated (yet) for wgMLST.



MentaLiST

Another command line MLST caller, focused on large schemes.

Image taken from http://www.applied-maths.com/applications/wgmlst

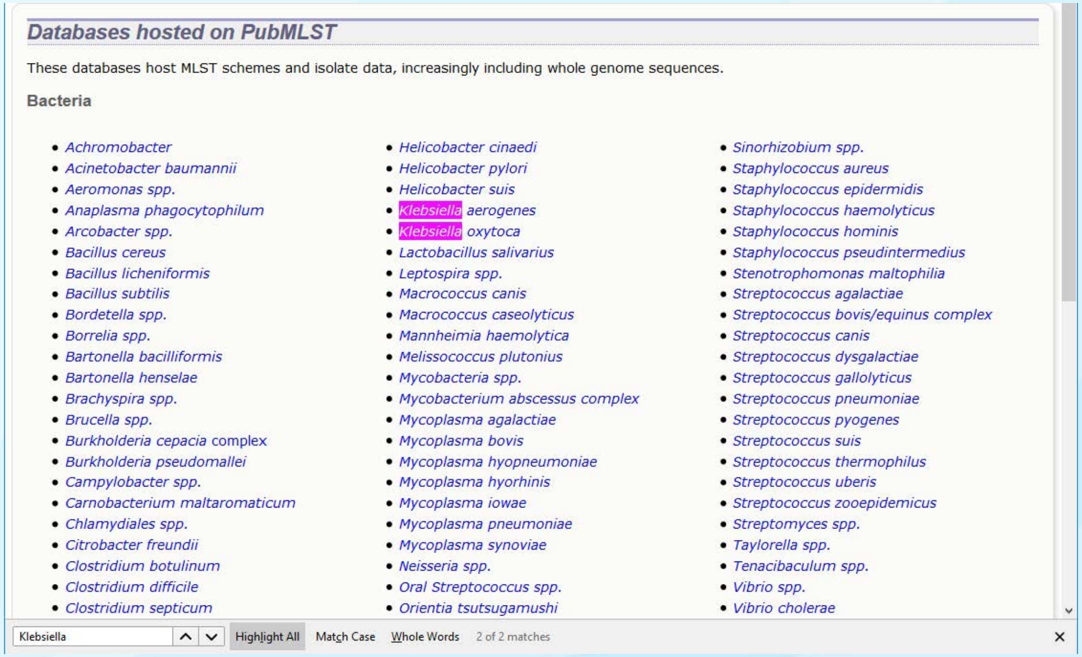
Page et al 2017, Comparison of Multi-locus Sequence Typing software for next generation sequencing data. *MGEN*.

MLST Resources

Main MLST site: https://pubmlst.org/

Good resource on MLST terms on the BigsDB manual: http://bigsdb.readthedocs.io/en/latest/concepts.html

API: https://pubmlst.org/rest/

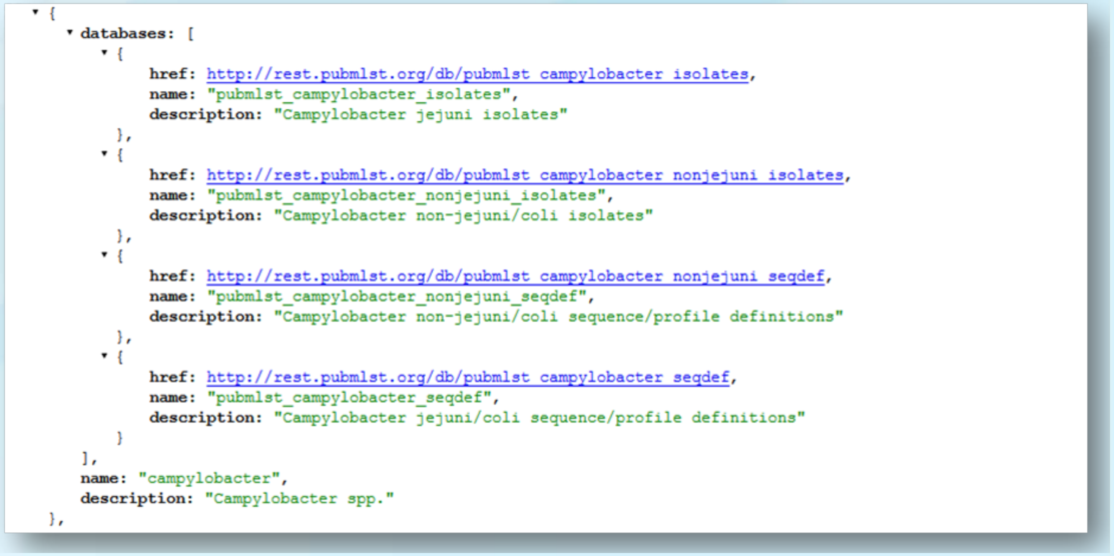


Databases hosted on PubMLST

These databases host MLST schemes and isolate data, increasingly including whole genome sequences.

Bacteria

- Achromobacter
- Acinetobacter baumannii
- Aeromonas spp.
- Anaplasma phagocytophilum
- Arcobacter spp.
- Bacillus cereus
- Bacillus licheniformis
- Bacillus subtilis
- Bordetella spp.
- Borrelia spp.
- Bartonella bacilliformis
- Bartonella henselae
- Brachyspira spp.
- Brucella spp.
- Burkholderia cepacia complex
- Burkholderia pseudomallei
- Campylobacter spp.
- Carnobacterium maltaromaticum
- Chlamydiales spp.
- Citrobacter freundii
- Clostridium botulinum
- Clostridium difficile
- Clostridium septicum
- Helicobacter cinaedi
- Helicobacter pylori
- Helicobacter suis
- Klebsiella aerogenes
- Klebsiella oxytoca
- Lactobacillus salivarius
- Leptospira spp.
- Macrococcus canis
- Macrococcus caseolyticus
- Mannheimia haemolytica
- Melissococcus plutonius
- Mycobacteria spp.
- Mycobacterium abscessus complex
- Mycoplasma agalactiae
- Mycoplasma bovis
- Mycoplasma hyopneumoniae
- Mycoplasma hyorhinis
- Mycoplasma iowae
- Mycoplasma pneumoniae
- Mycoplasma synoviae
- Neisseria spp.
- Oral Streptococcus spp.
- Orientia tsutsugamushi
- Sinorhizobium spp.
- Staphylococcus aureus
- Staphylococcus epidermidis
- Staphylococcus haemolyticus
- Staphylococcus hominis
- Staphylococcus pseudintermedius
- Stenotrophomonas maltophilia
- Streptococcus agalactiae
- Streptococcus bovis/equinus complex
- Streptococcus canis
- Streptococcus dysgalactiae
- Streptococcus gallolyticus
- Streptococcus pneumoniae
- Streptococcus pyogenes
- Streptococcus suis
- Streptococcus thermophilus
- Streptococcus uberis
- Streptococcus zooepidemicus
- Streptomyces spp.
- Taylorella spp.
- Tenacibaculum spp.
- Vibrio spp.
- Vibrio cholerae



```
▾ {
    ▾ databases: [
        ▾ {
              href: http://rest.pubmlst.org/db/pubmlst_campylobacter_isolates,
              name: "pubmlst_campylobacter_isolates",
              description: "Campylobacter jejuni isolates"
          },
        ▾ {
              href: http://rest.pubmlst.org/db/pubmlst_campylobacter_nonjejuni_isolates,
              name: "pubmlst_campylobacter_nonjejuni_isolates",
              description: "Campylobacter non-jejuni/coli isolates"
          },
        ▾ {
              href: http://rest.pubmlst.org/db/pubmlst_campylobacter_nonjejuni_seqdef,
              name: "pubmlst_campylobacter_nonjejuni_seqdef",
              description: "Campylobacter non-jejuni/coli sequence/profile definitions"
          },
        ▾ {
              href: http://rest.pubmlst.org/db/pubmlst_campylobacter_seqdef,
              name: "pubmlst_campylobacter_seqdef",
              description: "Campylobacter jejuni/coli sequence/profile definitions"
          }
      ],
      name: "campylobacter",
      description: "Campylobacter spp."
  },
```

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

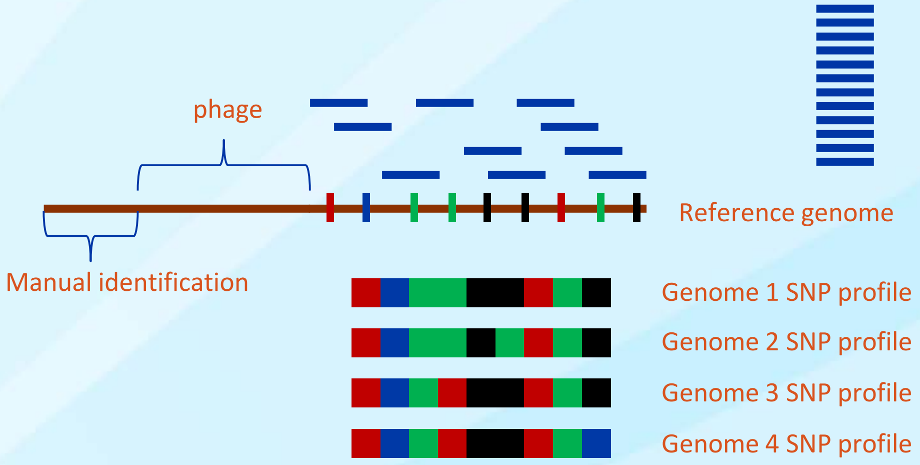
hqSNP: high-quality SNP (ie, high confidence)

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



phage

Reference genome

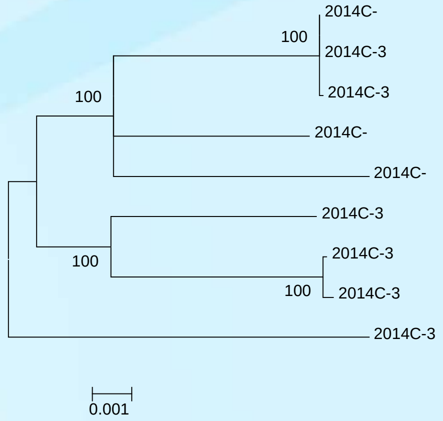
Manual identification

Genome 1 SNP profile

Genome 2 SNP profile

Genome 3 SNP profile

Genome 4 SNP profile

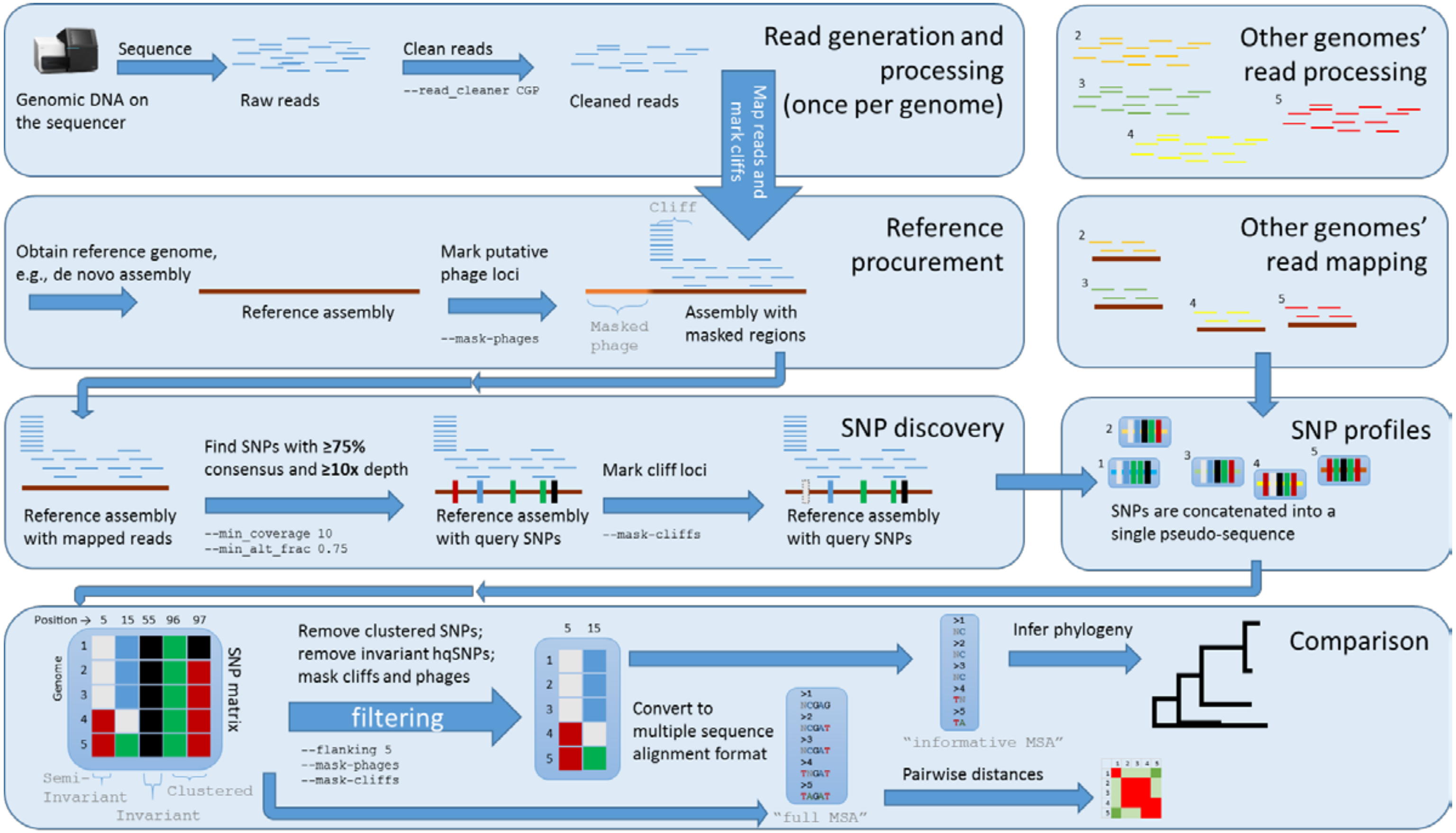


Phylogeny

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

Katz et al. (2017) A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology for foodborne pathogens. Frontiers in Microbiology 8: 375.

More details



Read generation and processing (once per genome)

Genomic DNA on the sequencer

Sequence

Raw reads

Clean reads

`--read_cleaner CGP`

Cleaned reads

Map reads and mark cliffs

Other genomes' read processing

Reference procurement

Obtain reference genome, e.g., de novo assembly

Reference assembly

Mark putative phage loci

`--mask-phages`

Masked phage

Cliff

Assembly with masked regions

Other genomes' read mapping

SNP discovery

Reference assembly with mapped reads

Find SNPs with ≥75% consensus and ≥10x depth

`--min_coverage 10`
`--min_alt_frac 0.75`

Reference assembly with query SNPs

Mark cliff loci

`--mask-cliffs`

Reference assembly with query SNPs

SNP profiles

SNPs are concatenated into a single pseudo-sequence

Comparison

Remove clustered SNPs; remove invariant hqSNPs; mask cliffs and phages

filtering

`--flanking 5`
`--mask-phages`
`--mask-cliffs`

SNP matrix

Semi-Invariant

Invariant

Clustered

Convert to multiple sequence alignment format

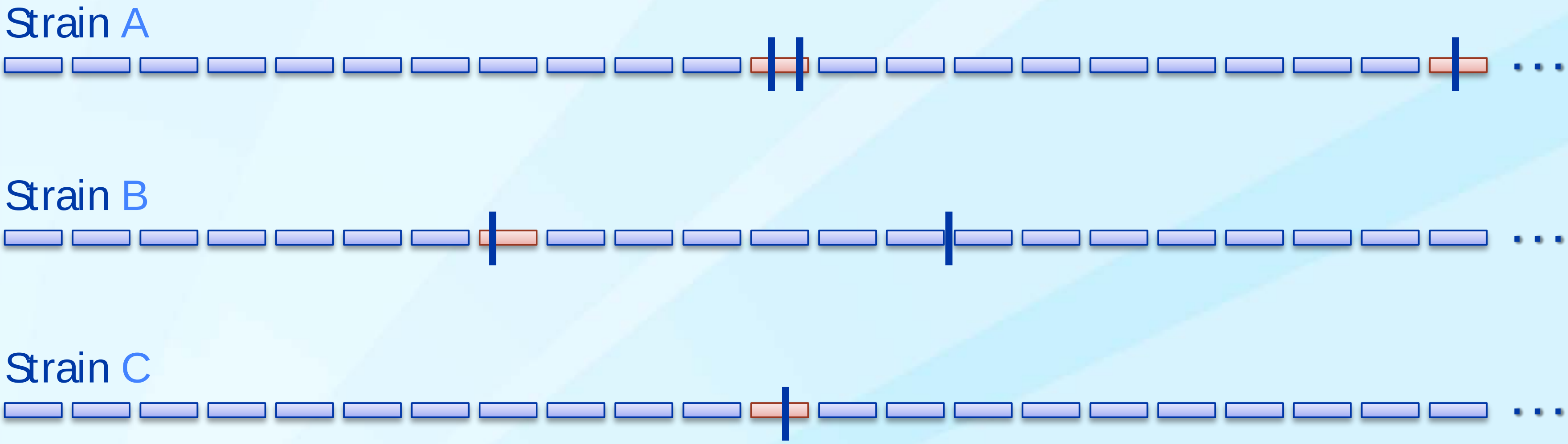
"full MSA"

"informative MSA"

Infer phylogeny

Pairwise distances

SNPs overlayed on MLST loci



Strain A

Strain B

Strain C

SNP software

Lyve-SET

Optimized for outbreak surveillance.

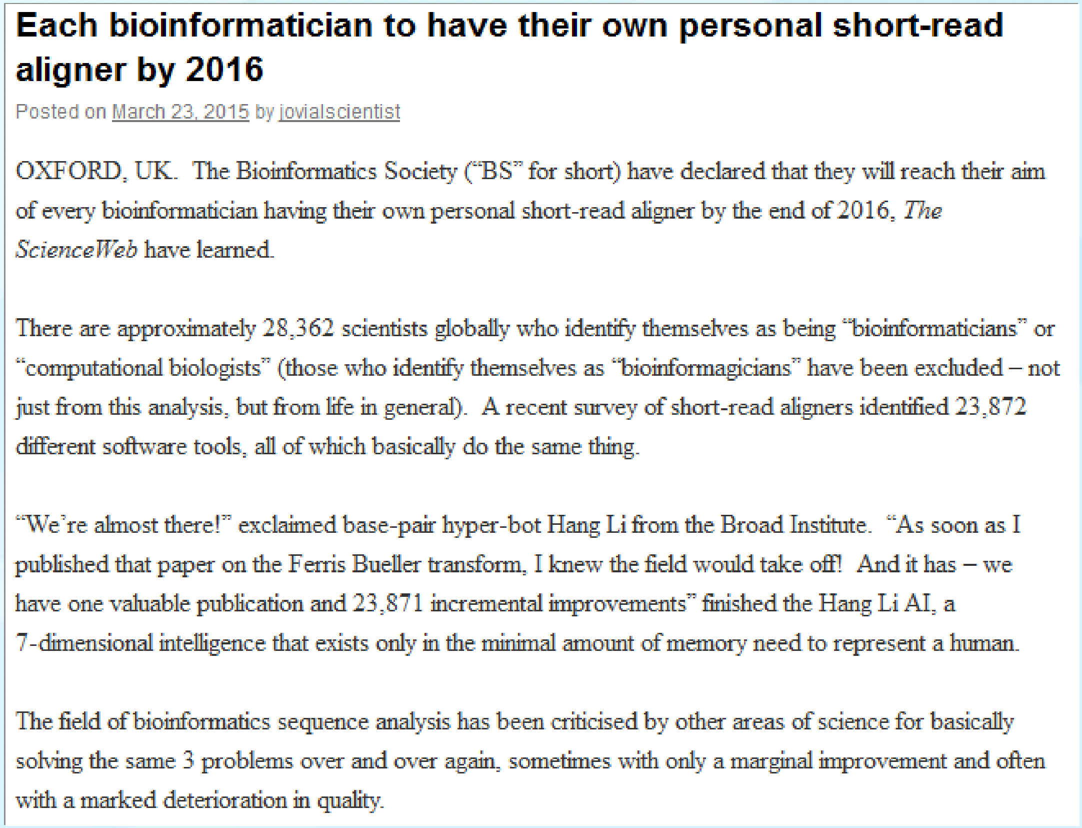
SNP-Pipeline

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.

KSNP (?)



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

Posted on March 23, 2015 by jovialscientist

OXFORD, UK. The Bioinformatics Society ("BS" for short) have declared that they will reach their aim of every bioinformatician having their own personal short-read aligner by the end of 2016, *The ScienceWeb* have learned.

There are approximately 28,362 scientists globally who identify themselves as being "bioinformaticians" or "computational biologists" (those who identify themselves as "bioinformagicians" have been excluded – not just from this analysis, but from life in general). A recent survey of short-read aligners identified 23,872 different software tools, all of which basically do the same thing.

"We're almost there!" exclaimed base-pair hyper-bot Hang Li from the Broad Institute. "As soon as I published that paper on the Ferris Bueller transform, I knew the field would take off! And it has – we have one valuable publication and 23,871 incremental improvements" finished the Hang Li AI, a 7-dimensional intelligence that exists only in the minimal amount of memory need to represent a human.

The field of bioinformatics sequence analysis has been criticised by other areas of science for basically solving the same 3 problems over and over again, sometimes with only a marginal improvement and often with a marked deterioration in quality.

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

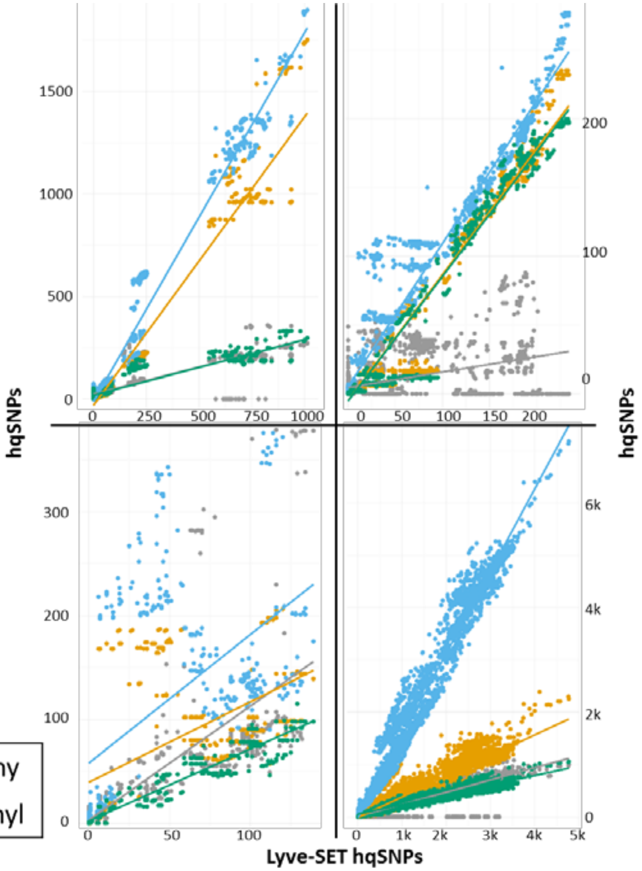
Installation and sample run

```
$   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		
Pipeline	**y=mx+b**	**R^2**
kSNP	y=0.26+24	0.69
RealPhy	y=1.14+31	0.96
SNP-Pipeline	y=1.8x-13	0.97
SNVPhyl	y=0.27x+19	0.58

S. enterica		
Pipeline	**y=mx+b**	**R^2**
kSNP	y=0.11x+4.7	0.23
RealPhy	y=0.92x-5	0.95
SNP-Pipeline	y=1.0x+5.4	0.96
SNVPhyl	y=0.91-5.1	0.94

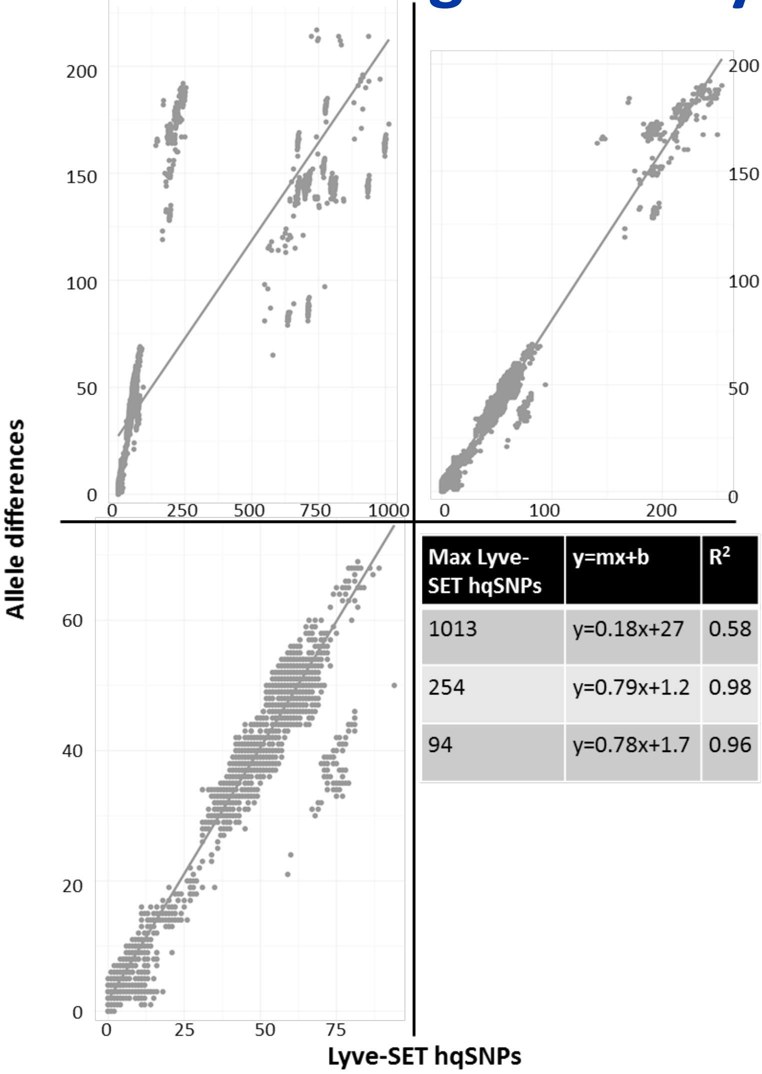


E. coli		
Pipeline	**y=mx+b**	**R^2**
kSNP	y=1.1x+2.9	0.43
RealPhy	y=0.78+39	0.27
SNP-Pipeline	y=1.2x+58	0.3
SNVPhyl	y=0.69x+2.1	0.92

C. jejuni		
Pipeline	**y=mx+b**	**R^2**
kSNP	y=0.23x+4	0.89
RealPhy	y=0.4x-15	0.88
SNP-Pipeline	y=1.6x-17	0.97
SNVPhyl	y=0.18+49	0.92

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.

Comparison with whole-genome MLST (Listeria monocytogenes only)



Max Lyve-SET hqSNPs	y=mx+b	R^2
1013	y=0.18x+27	0.58
254	y=0.79x+1.2	0.98
94	y=0.78x+1.7	0.96

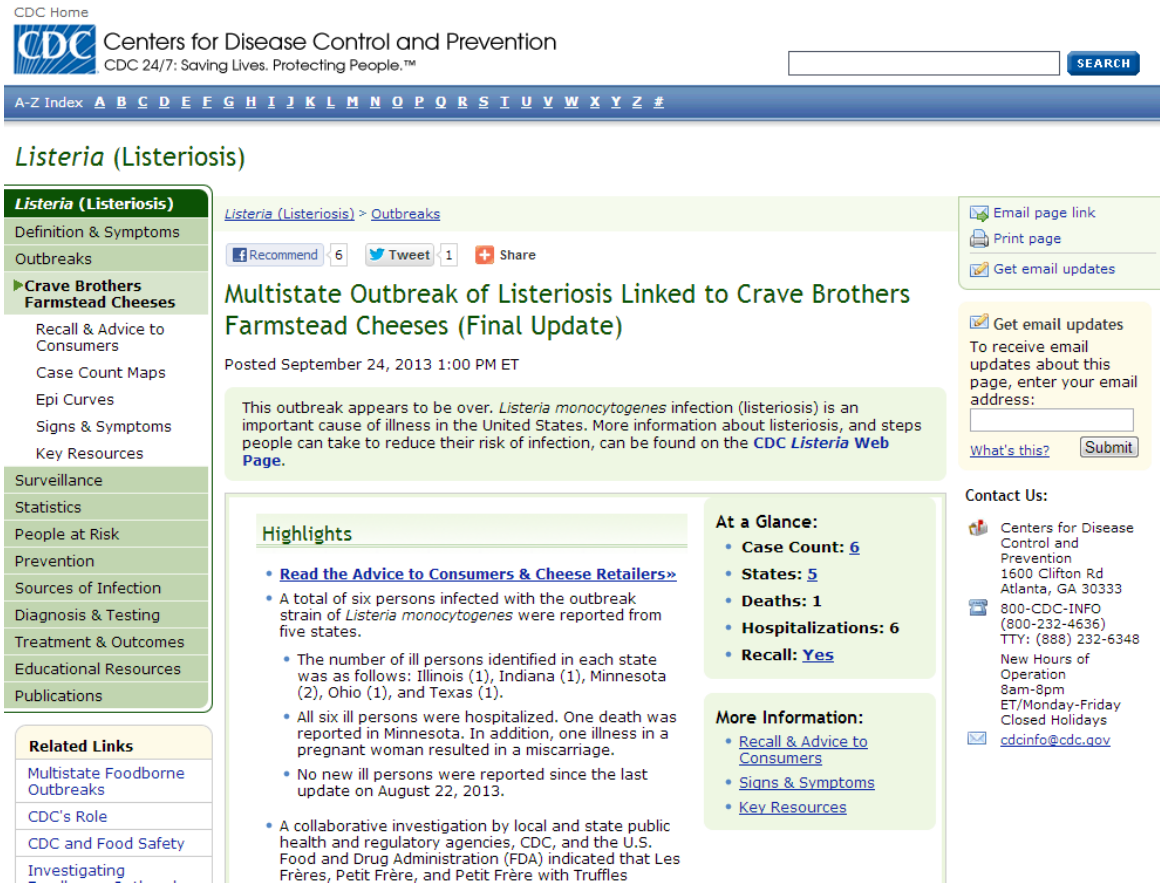
Katz et al 2017, *Lyve-SET*,
Frontiers in Microbiology.

Which algorithm should you use?

	Kmer-based	wgMLST	hqSNP
Diversity	✓✓	✓	✗✗
Outbreak-level resolution	✗	✓	✓
Further genomic information	✗	✓	✓
Minimal upfront effort	✓	✗✗	✓
Fast	✓✓	✓✓	✗
Easy to use for anyone	✗	✓	✗

The best level resolution for outbreaks is theoretically hqSNP but empirically wgMLST has performed approximately as well

Multistate outbreak of farmstead cheeses



CDC Home

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.™

A-Z Index A B C D E F G H I J K L M N O P Q R S T U V W X Y Z #

Listeria (Listeriosis)

Listeria (Listeriosis) > Outbreaks
Recommend 6 Tweet 1 Share

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

Get email updates

To receive email updates about this page, enter your email address:

Contact Us:

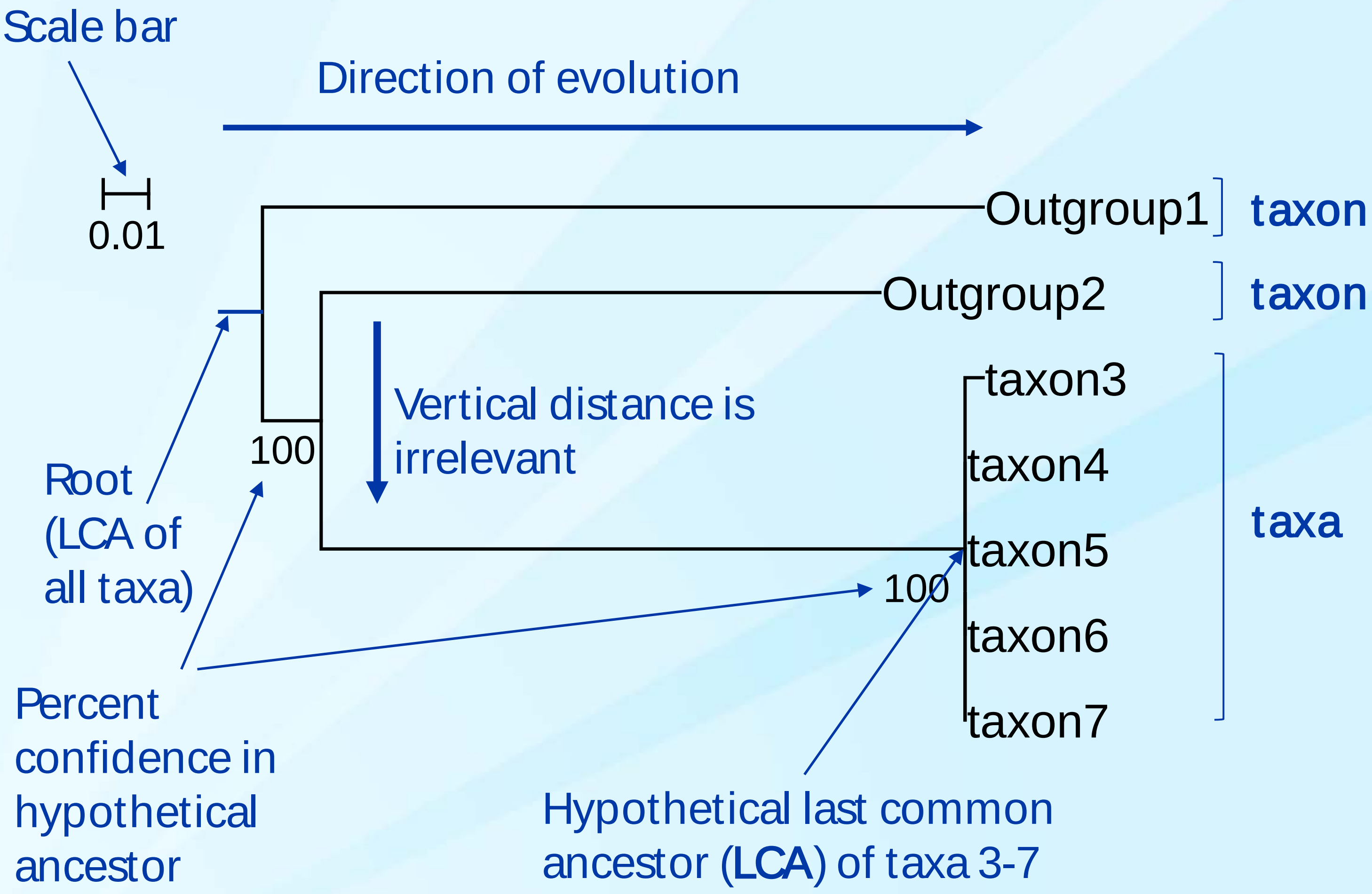
Centers for Disease Control and Prevention
1600 Clifton Rd
Atlanta, GA 30333

800-CDC-INFO
(800-232-4636)
TTY: (888) 232-6348

New Hours of Operation
8am-8pm
ET/Monday-Friday
Closed Holidays

cdcinfo@cdc.gov

How to read a phylogeny


Scale bar
Direction of evolution
0.01
Outgroup1
taxon
Outgroup2
taxon
taxon3
taxon4
taxon5
taxon6
taxon7
taxa
Vertical distance is irrelevant
100
100
Root (LCA of all taxa)
Percent confidence in hypothetical ancestor
Hypothetical last common ancestor (LCA) of taxa 3-7

2013 outbreak linked to farmstead cheese

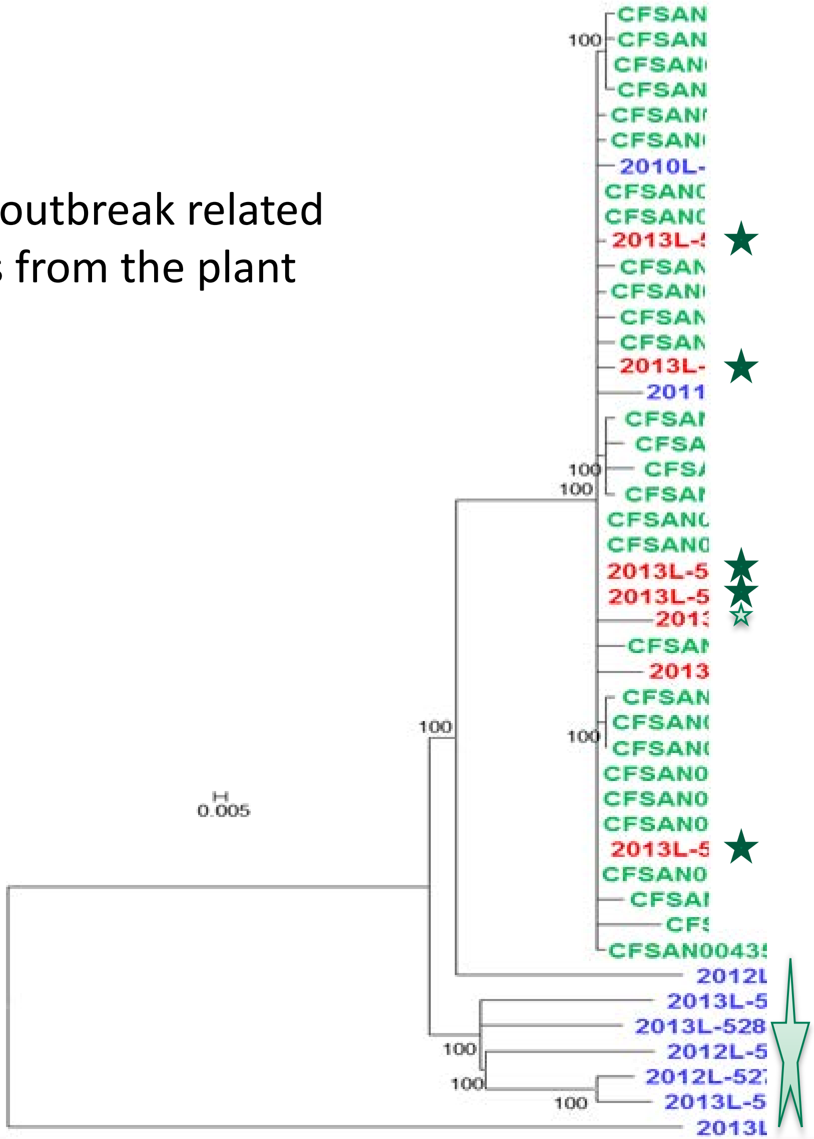
Red= epi-related clinical isolates

Blue= retrospective clinical cases or not outbreak related

Green= historical environmental isolates from the plant

★ Exposure

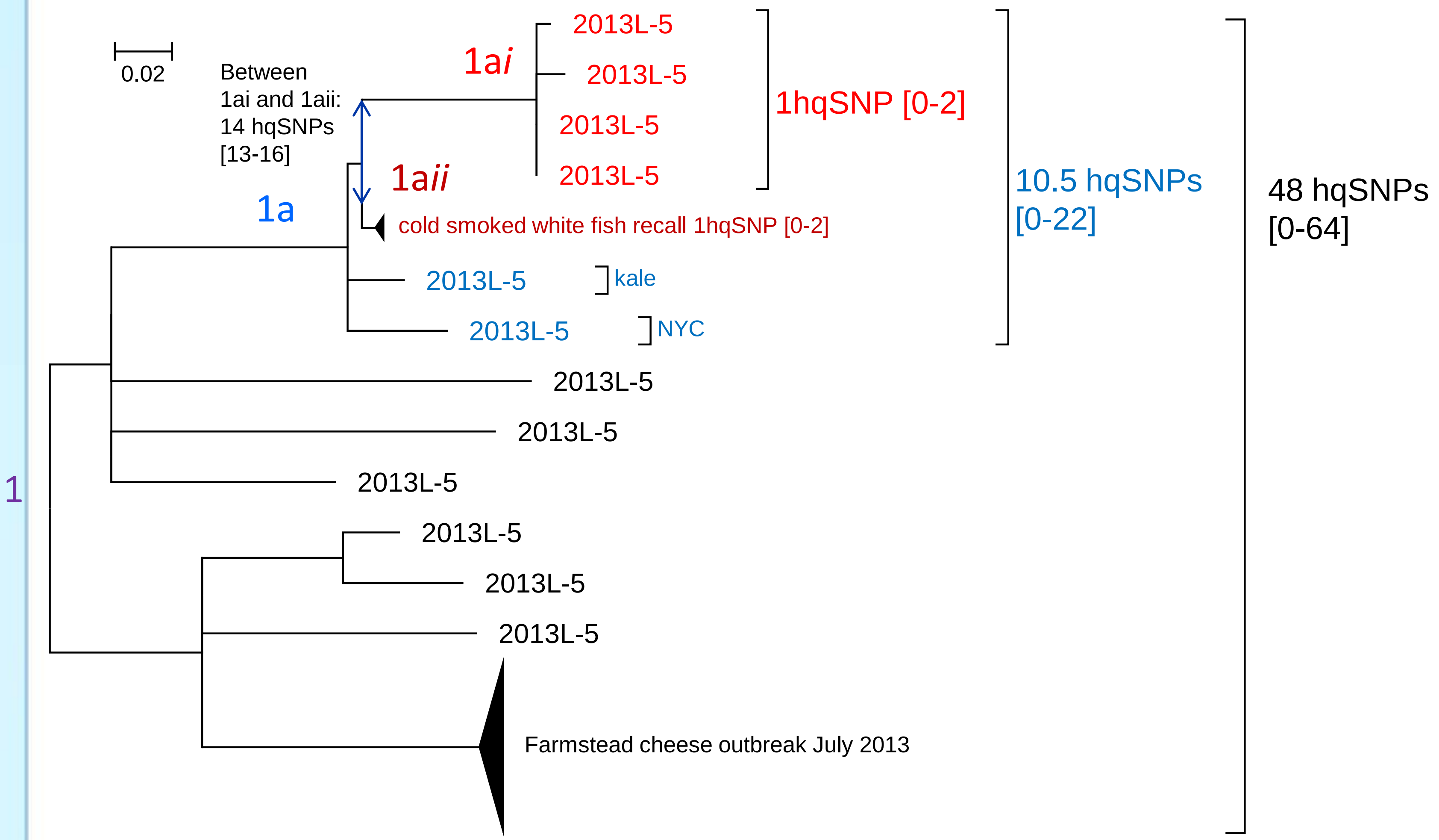
☆ No Exposure



For another outbreak of Lmono of the same year but in-depth analysis:

Chen, Yi, et al. "Whole genome and core genome multilocus sequence typing and single nucleotide polymorphism analyses of Listeria monocytogenes isolates associated with an outbreak linked to cheese, United States, 2013." *Applied and environmental microbiology* 83.15 (2017): e00633-17.

Phylogenetically related outbreak of unknown etiology, December 2013



In conclusion

- **Genomic epidemiology is awesome.**
- **There are several methods to compare genomes.**
  - Kmer
  - MLST
  - SNP
- **We are using genomic epi in real time to solve real world problems.**
- **… but genomic epidemiology does not work in a vacuum. Other data are needed for real world conclusions.**

Questions?

Lee Katz

gzu2@cdc.gov

lskatz

github.com/lskatz