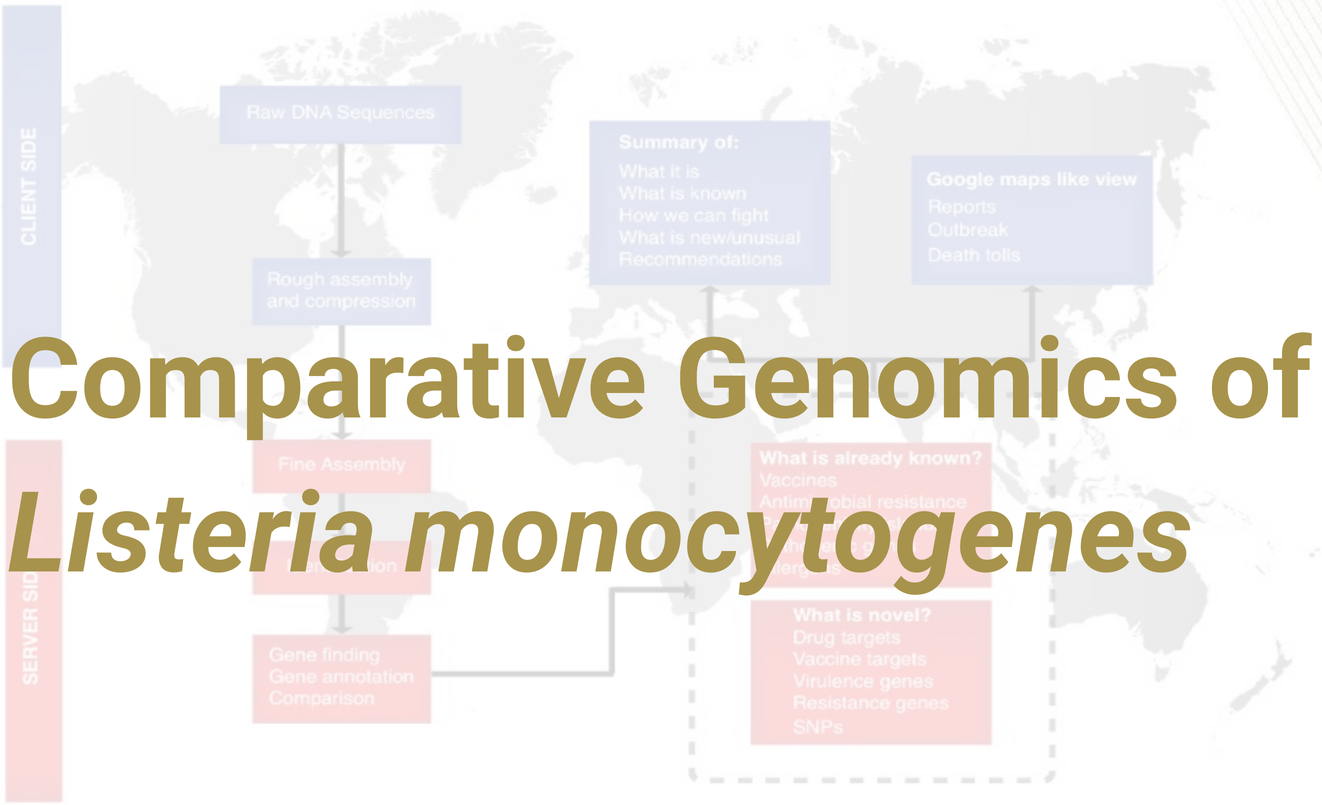


Comparative Genomics of *Listeria monocytogenes*

Swetha Singu	Gulay Bengu Ulukaya
Ruize Yang	Yuhua Zhang
Deepali Kundnani	Jie Zhou

Listeria monocytogenes - Characteristics

Listeria Sensu stricto - Clade 1 - 6 species	Organism infected
L.monocytogenes	human pathogen
L.ivanovii	animal pathogen
L.marthii	symptom free animals
L.innocua	symptom free animals
L.weshimeri	symptom free animals
L.seeligeri	symptom free animals
Listeria sensu lato - Clade 2 - 11 species	Organism infected
11 species	environmental bacteria

Listeria monocytogenes	
Type	Gram positive
Infections caused	Listeriosis
Features	hardy organism
Evolutionary lineages	4
Serotypes	14 [1/2a, 1/2b, 1/2c, 3a, 3b, 3c, 4a, 4b, 4c, 4d, 4e, and 7]
Major serotypes	**1/2a [Lineage II]** **1/2b,4b [Lineage I]**
Annual Infections	1600
Deaths	1 in 5

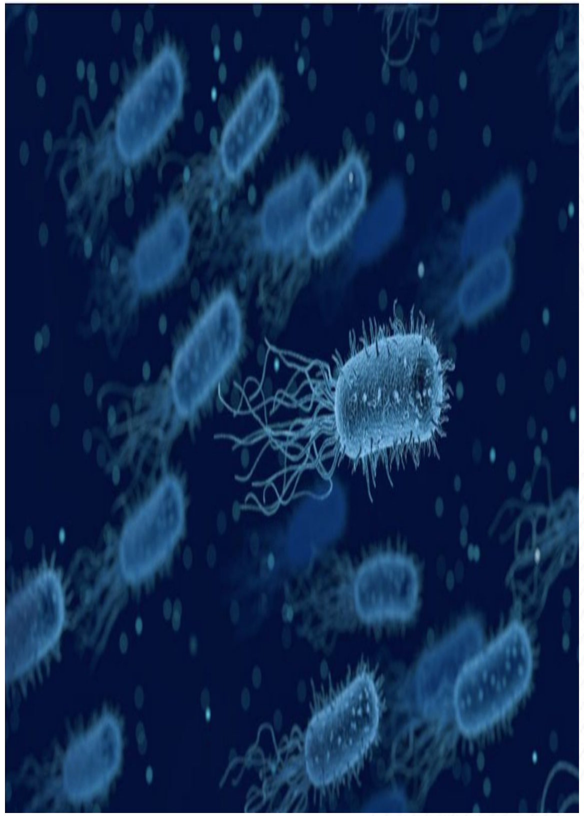
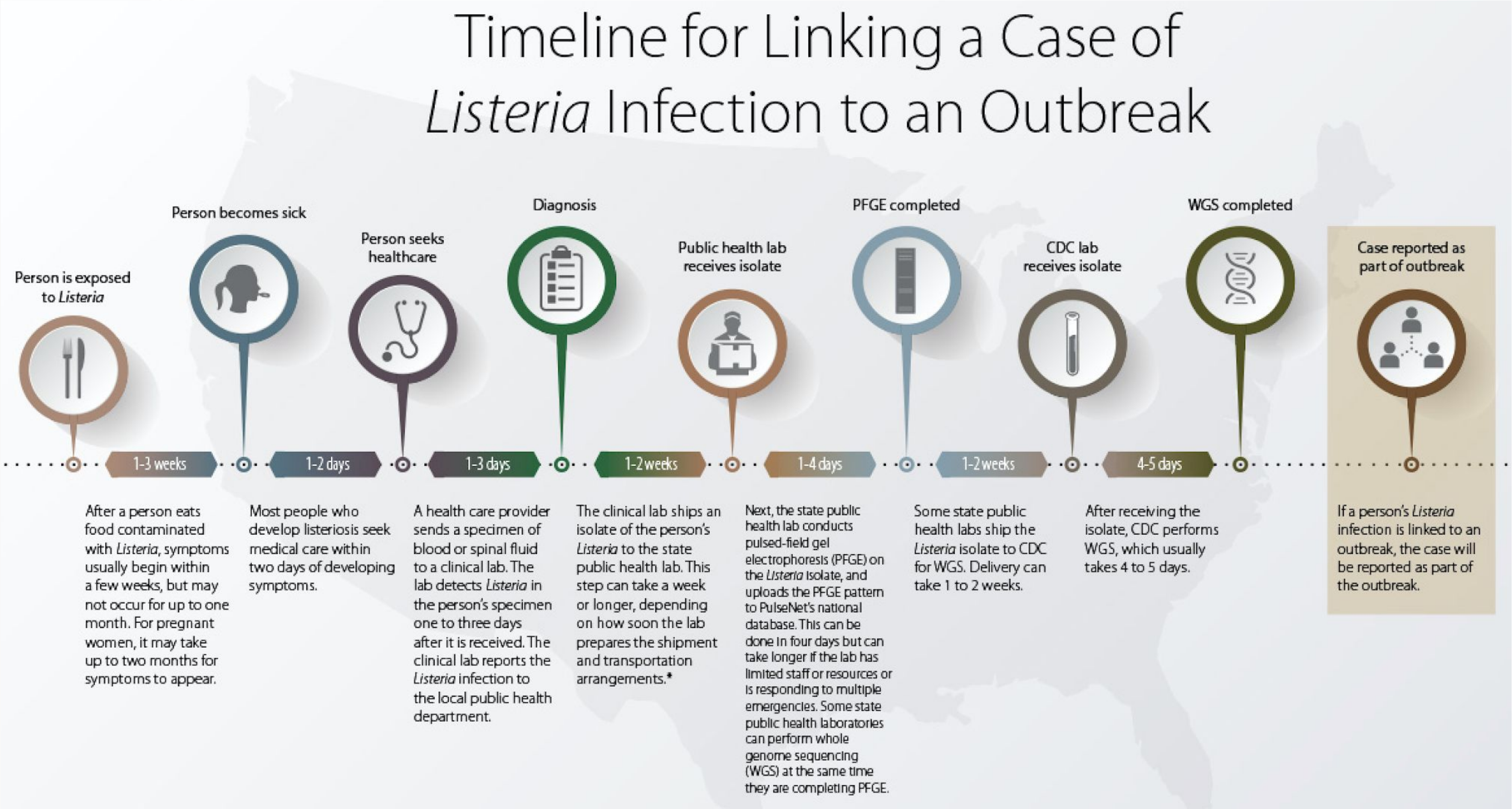


Image:https://www.forbes.com/sites/anagarciavaldivia/2019/08/23/health-alert-in-spain-listeriosis-outbreak-affects-168-people/#d21192426807

General timeline for *Listeria* infection



Timeline for Linking a Case of *Listeria* Infection to an Outbreak

Person is exposed to *Listeria*

1-3 weeks

After a person eats food contaminated with *Listeria*, symptoms usually begin within a few weeks, but may not occur for up to one month. For pregnant women, it may take up to two months for symptoms to appear.

Person becomes sick

1-2 days

Most people who develop listeriosis seek medical care within two days of developing symptoms.

Person seeks healthcare

1-3 days

A health care provider sends a specimen of blood or spinal fluid to a clinical lab. The lab detects *Listeria* in the person's specimen one to three days after it is received. The clinical lab reports the *Listeria* infection to the local public health department.

Diagnosis

1-2 weeks

The clinical lab ships an isolate of the person's *Listeria* to the state public health lab. This step can take a week or longer, depending on how soon the lab prepares the shipment and transportation arrangements.*

Public health lab receives isolate

1-4 days

Next, the state public health lab conducts pulsed-field gel electrophoresis (PFGE) on the *Listeria* isolate, and uploads the PFGE pattern to PulseNet's national database. This can be done in four days but can take longer if the lab has limited staff or resources or is responding to multiple emergencies. Some state public health laboratories can perform whole genome sequencing (WGS) at the same time they are completing PFGE.

PFGE completed

1-2 weeks

Some state public health labs ship the *Listeria* isolate to CDC for WGS. Delivery can take 1 to 2 weeks.

CDC lab receives isolate

4-5 days

After receiving the isolate, CDC performs WGS, which usually takes 4 to 5 days.

WGS completed

Case reported as part of outbreak

If a person's *Listeria* infection is linked to an outbreak, the case will be reported as part of the outbreak.

Image:https://www.cdc.gov/listeria/timeline.html

How to establish if different isolates are part of an outbreak?

Laboratory [detect clusters] → **Epidemiology** [link cases to source] → **Traceback** [stop further spread of disease]

- DNA fingerprinting: PFGE vs Whole genome Sequence analysis?

The Listeria WGS project was started by CDC, federal partners and state and local health department since 2013 to links WGS and epidemiologic data to better detect and investigate listeriosis outbreaks.

Eg: Multistate Outbreak of Listeriosis Linked to Commercially Produced, Prepackaged Caramel Apples Made from Bidart Bros. Apples [2015] - **using WGS**

- Listeria infection are rare -however higher fatality - 20-30% in high risk.
- *Listeria* initiative: provides a look at the who, where and when of Listeria infections

Listeria monocytogenes - Genetic view

Genetic view	Size
Total Genome	2.8 - 3.2 million bases
GC content	39%
core genome MLST	2014 - 2647 loci
whole genome MLST	4804 loci
Pan genome MLST	3560-6612 loci
Plasmids	14 in *Listeria*

- Quorum sensing and other signals cause the up-regulation of several **virulence genes.**

Listeriosis treatment using	Antibiotic
β-lactam antibiotic	amoxicillin, penicillin, ampicillin
aminoglycoside	gentamicin
allergy to penicillin	trimethoprim - sulfamethoxazole
alternative treatment	tetracycline and erythromycin

Listeria monocytogenes - Interested genes

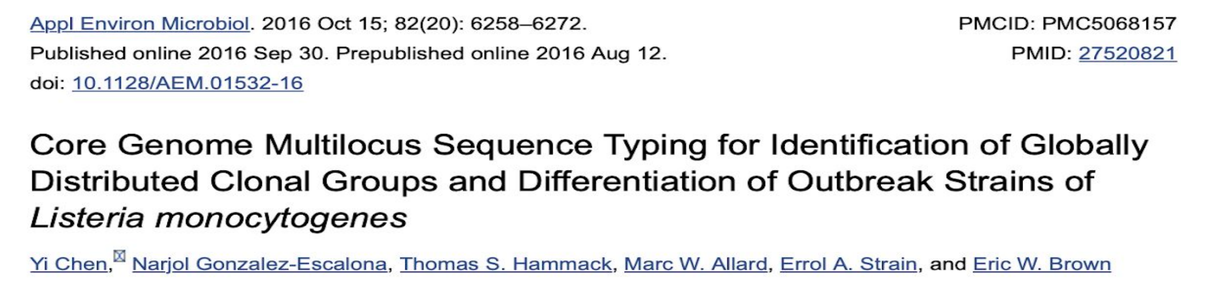
GENE	Antibiotic resistance
lmrB	lincomycin resistance protein
vanA, vanB	vancomycin resistance
dfrD and dfrG	Trimethoprim resistance
tetA, *tetK*, and *tetL*, *tetM and tetS*	Tetracycline resistance
emrA, emrB and emrC	Erythromycin resistance
lde gene	Fluoroquinolone resistance

- All L.monocytogenes species in general are inherently resistant to cephalosporins, oxacillin and fosfomycin

- Genes associated with virulence factors and pathogenicity islands - LIPI 1, LIPI 2, LIPI3, LIPI4

Genomic approaches - Subtyping

- Genomic data can be exploited with many different bioinformatics methods.
- Whole genome approach or Phylogenetic approach - ANI, MLST [core genome, whole genome] and SNP

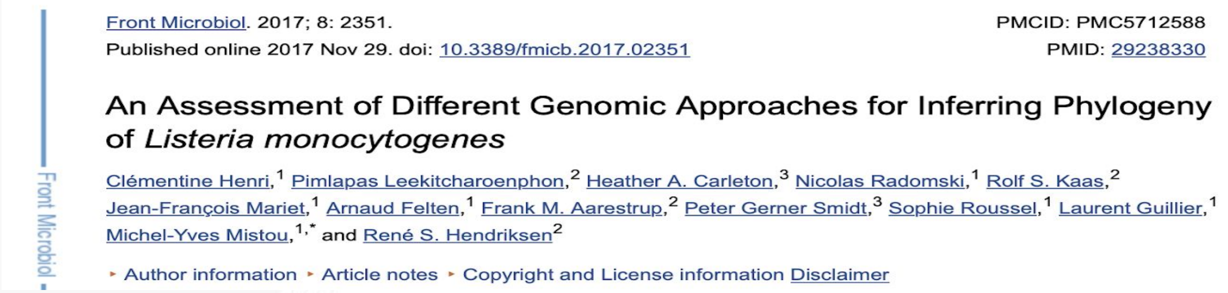


Appl Environ Microbiol. 2016 Oct 15; 82(20): 6258–6272.
Published online 2016 Sep 30. Prepublished online 2016 Aug 12.
doi: 10.1128/AEM.01532-16

PMCID: PMC5068157
PMID: 27520821

Core Genome Multilocus Sequence Typing for Identification of Globally Distributed Clonal Groups and Differentiation of Outbreak Strains of *Listeria monocytogenes*

Yi Chen,✉ Narjol Gonzalez-Escalona, Thomas S. Hammack, Marc W. Allard, Errol A. Strain, and Eric W. Brown



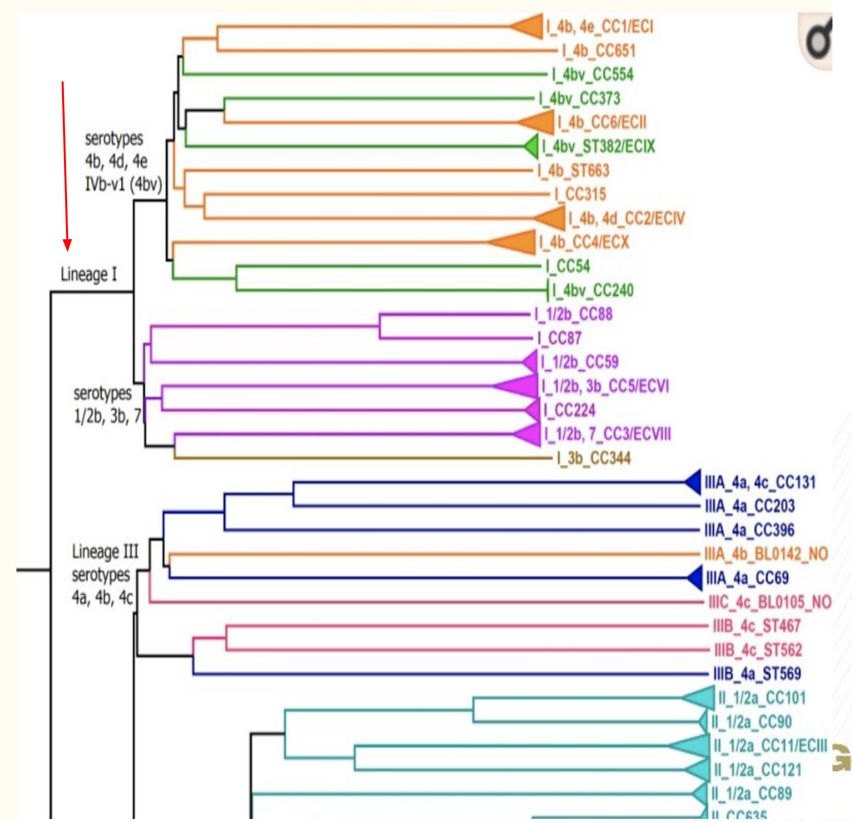
Front Microbiol. 2017; 8: 2351.
Published online 2017 Nov 29. doi: 10.3389/fmicb.2017.02351

PMCID: PMC5712588
PMID: 29238330

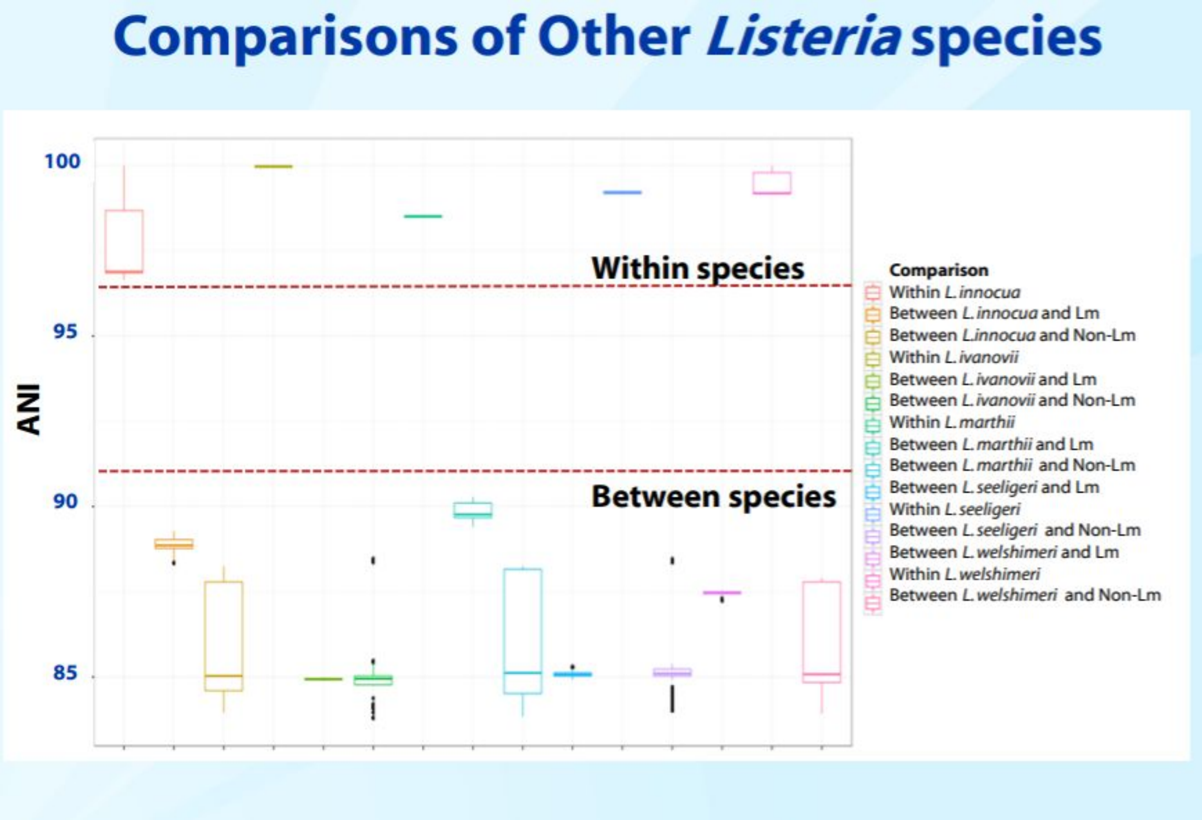
An Assessment of Different Genomic Approaches for Inferring Phylogeny of *Listeria monocytogenes*

Clémentine Henri,[1] Pimlapas Leekitcharoenphon,[2] Heather A. Carleton,[3] Nicolas Radomski,[1] Rolf S. Kaas,[2] Jean-François Mariet,[1] Arnaud Felten,[1] Frank M. Aarestrup,[2] Peter Gerner Smidt,[3] Sophie Roussel,[1] Laurent Guillier,[1] Michel-Yves Mistou,[1,*] and René S. Hendriksen[2]

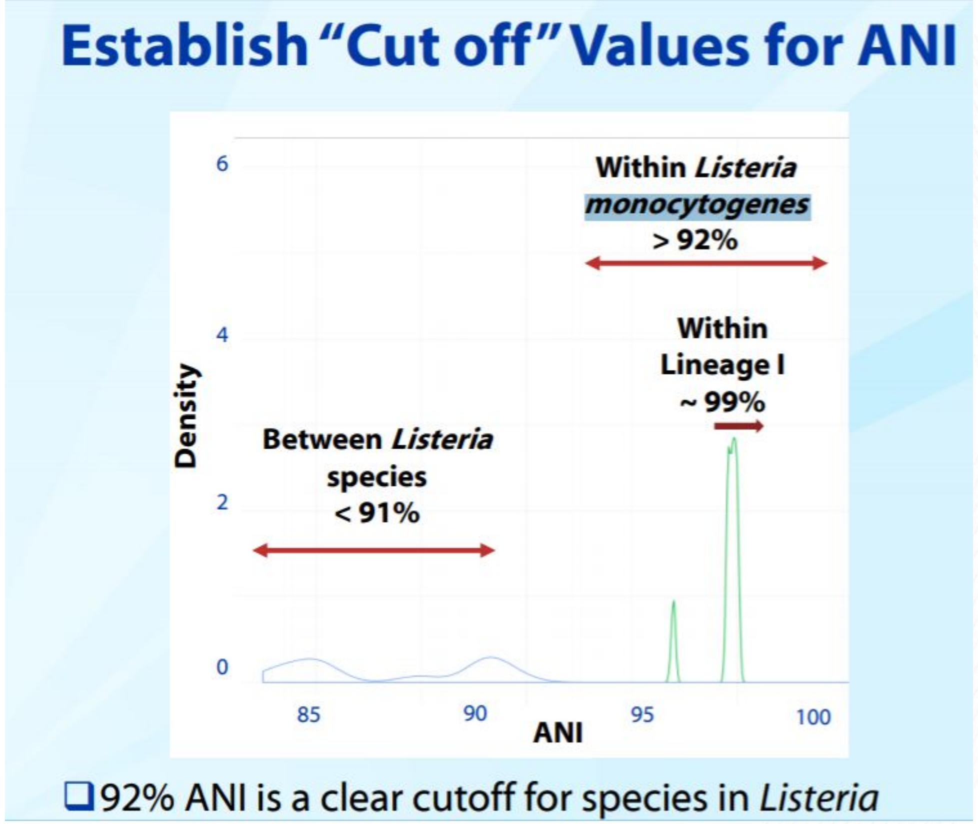
▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer



ANI for Listeria



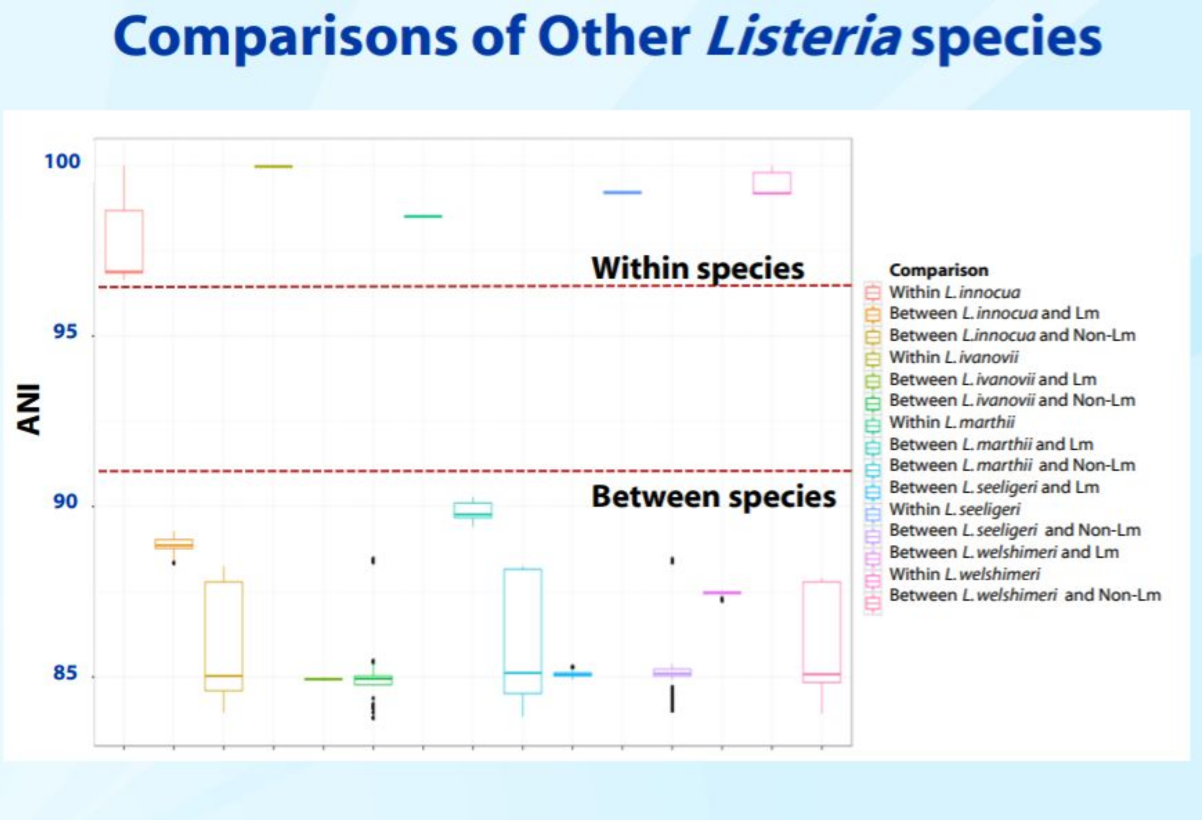
Comparisons of Other *Listeria* species



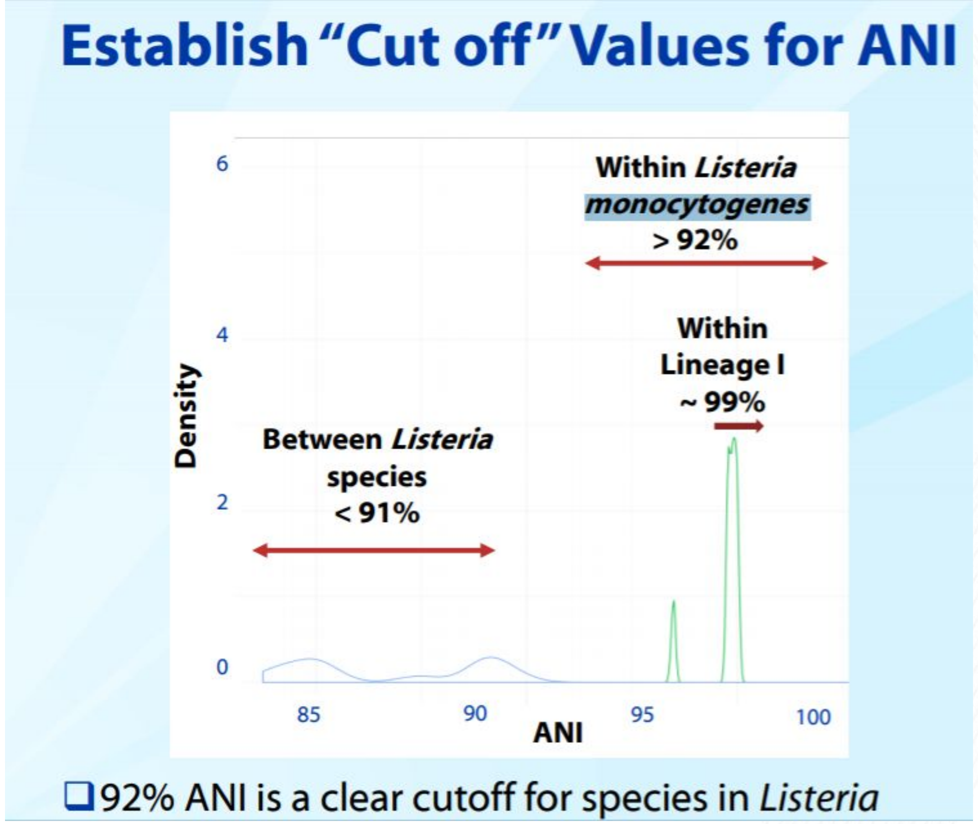
Establish "Cut off" Values for ANI

❑92% ANI is a clear cutoff for species in *Listeria*

ANI for Listeria



Comparisons of Other *Listeria* species



Establish "Cut off" Values for ANI

☐ 92% ANI is a clear cutoff for species in *Listeria*

Alignment-based ANI

- ANI values are based on pairwise alignment of the genome stretches.
- Reliability depends on the quantity and quality of the aligned fragment.
- We can calculate the ANI based on BLAST (ANIb) and MUMmer (ANIm)
- JSpecies is able to run BLAST-based and MUMmmer-based ANI
- Limitation: Needs a lot of time.

Alignment-free ANI

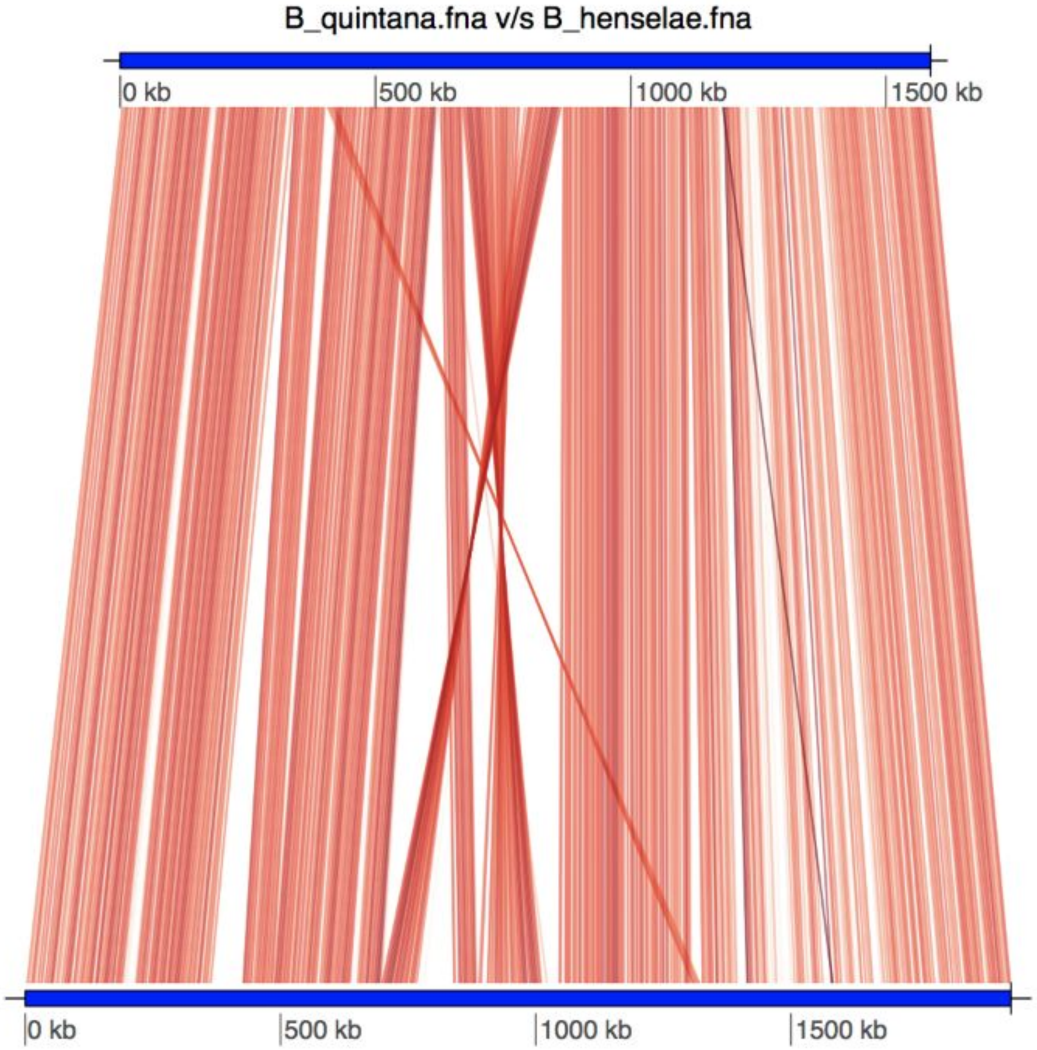
$$I(A,B)/100 = 1 + \frac{1}{k} \times \log\left(\frac{2 \cdot J(A,B)}{1 + J(A,B)}\right)$$

- Fast-ANI
- Avoids expensive sequence alignments
- Uses Mashmap as its MinHash based sequence mapping engine to compute the orthologous mappings and alignment identity estimates
- Estimates a k-mer based Jaccard similarity

```
$ ./fastANI -q [QUERY_GENOME] -r [REFERENCE_GENOME] -o [OUTPUT_FILE]
```

```
./contig/CGT3002contigs.fasta   ./contig/CGT3003contigs.fasta   99.9739 893     899
```

Fast ANI


B_quintana.fna v/s B_henselae.fna
0 kb
500 kb
1000 kb
1500 kb
0 kb
500 kb
1000 kb
1500 kb

Multilocus Sequence Typing (MLST)

- Aim
  - Characterizing DNA sequence variations in bacterial isolates by focusing on allelic diversity across housekeeping genes (highly-conserved genes)
  - Evaluating relationships between strains based on their unique allelic profiles or sequences (Maiden, 2006)
  - Important in pathogen outbreak surveillance

- Deliverables: Allelic profiles of analyzed genes, sequence type for each isolate, phylogenetic tree generated with MLST output

- Types: 7-gene MLST, wgMLST, cgMLST, rMLST

Possible Tools to Use

- MentaLIST
  - Performs allele calling directly from reads, relies on existing schemas and allele definitions (from PubMLST and cgMLST.org) (Silva, 2018)
  - Faster than other tools for larger schemas like cgMLST and wgMLST (“MentaLIST”)
- ChewBBACA
  - Complete stand-alone pipeline including constructing and validating novel cg/wgMLST schemas and performing allele calling
  - De novo assemblers on complete or draft genomes
  - Suitable for large scale studies
- StringMLST
  - Easy and fast to run
  - Self-reported 100% accuracy

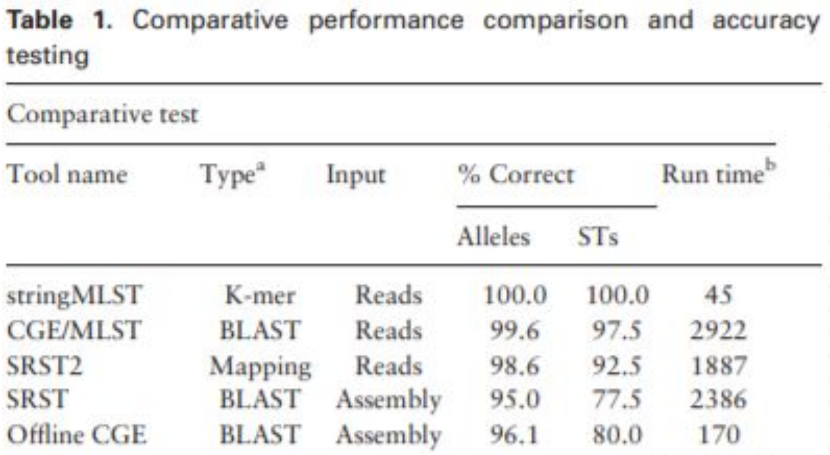


Table 1. Comparative performance comparison and accuracy testing

Comparative test

Tool name	Type[a]	Input	% Correct		Run time[b]
			Alleles	STs	
stringMLST	K-mer	Reads	100.0	100.0	45
CGE/MLST	BLAST	Reads	99.6	97.5	2922
SRST2	Mapping	Reads	98.6	92.5	1887
SRST	BLAST	Assembly	95.0	77.5	2386
Offline CGE	BLAST	Assembly	96.1	80.0	170

Gupta, Anuj, et al. “StringMLST: a Fast k-Mer Based Tool for Multilocus Sequence Typing.” *Bioinformatics (Oxford, England)*, U.S. National Library of Medicine, 1 Jan. 2017, www.ncbi.nlm.nih.gov/pubmed/27605103.

Our First MLST Tool of Choice: StringMLST

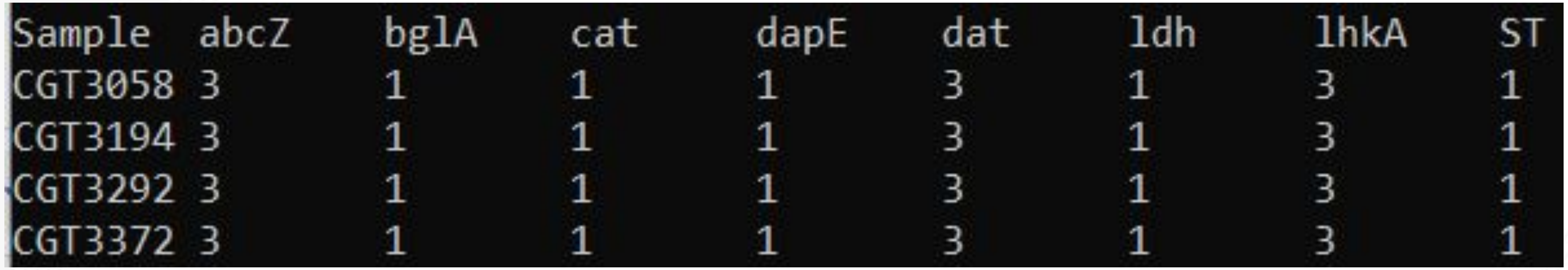
Exploratory tool for MLST, 7 housekeeping genes

Utilized the existing MLST scheme from PubMLST

Very fast and efficient

Plan: Construct phylogenetic tree from this initial output to visualize the sequence types of isolates at hand, research heteroresistance and susceptibility of sequence types

First five lines of output:



```
Sample  abcZ    bglA    cat     dapE    dat     ldh     lhkA    ST
CGT3058 3       1       1       1       3       1       3       1
CGT3194 3       1       1       1       3       1       3       1
CGT3292 3       1       1       1       3       1       3       1
CGT3372 3       1       1       1       3       1       3       1
```

Next Tool of Choice: MentaLIST

- Existing or constructed schema
  - A traditional MLST schema exists for our species on PubMLST and cgMLST schema exists on cgmlst.org
  - Verified and comprehensive MLST schemes take time & funding
  - We choose to use this existing scheme along with known phenotypic profiles of our samples to easily and accurately get variances in significant genes for 26,395 strains of Listeria
- Options
  - 7-gene MLST, cgMLST ---> Plan: Construct phylogenetic tree, compare with the tree from StringMLST output
  - Detects novel alleles and their mutation(s)
- Efficiency
  - Faster than ChewbaCCa with same or better accuracy, less computational resources needed when running larger schemas like wgMLST (a few thousand loci) and cgMLST (a few hundred loci) (Feijao, 2017)

(SNP)-based Phylogenetic Analysis

What

- Identifies and compares SNPs between isolate genomes
- Measures variations of SNPs between isolates
- Construct a tree based on comparisons to differentiate isolates

Read mapping
- The NGS reads are mapped on a reference genome

Variant calling
- High quality variants are identified for each isolate using a predefined filtering parameters

SNP matrix construction
- The variant calls of individual isolates are selected based on specifies rules and combined into a population-wide SNP matrix

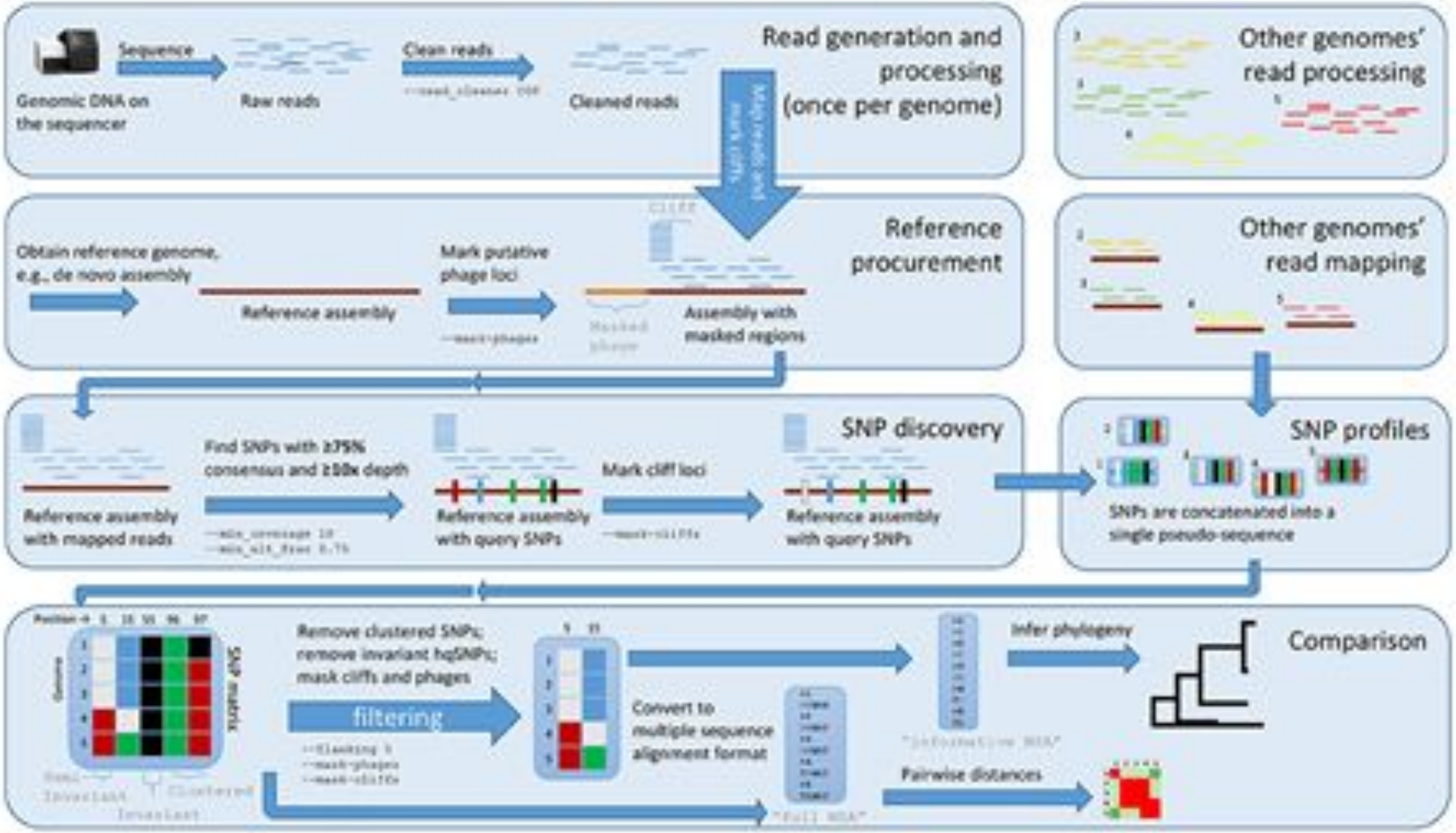
Phylogenetic analysis
- Concatenated SNPs from the SNP matrix are used to construct a phylogenetic tree

Comparing with wgMLST

- More flexible as they do not require a predefined scheme
- Provide an exceptionally high subtyping resolution
- Computationally demanding
- Distinguish between isolates that have been identified as closely related

Tool	Citation	Input	Reference Genome	Features	Preference
ParSNP	580/2014	Draft assemblies or finished genomes	Yes	• Multi- maximal unique matches • Only aligns the core genomes • Requires finished or assembled genomes	• Developed as a solution to the problem of aligning large numbers of microbial genomes • RealPhy depends on accurate mapping of raw reads (or contigs) to the reference genomes • kSNP3.0
RealPhy	226/2014	FASTQ (short reads), FASTA or Genbank format	Yes (multiple)	• Either FASTA or Genbank format (contigs or fully sequenced genomes) as reference genome • Option to combine individual reference alignments	
kSNP3.0	221/2015	A list of sequence file path containing a genome and a name for that genome (txt file)	No	• K-mer • Without genome alignment or reference genome	
SNPhylo	186/2014	SNP/genotype format (vcf/hapmap file), SNP data format file, GDS file	No	• Reduce SNP redundancy by linkage disequilibrium (LD) • Decreases running time without losing informative sites	• SNPhylo
CFSAN SNP Pipeline	93/2015	FASTQ (short reads)	Yes	• Focus on closely related sequences, not suited for the analysis of relatively distantly related organisms	• Developed with the objective of creating high quality SNP matrices for sequences from closely-related pathogens • Lyve-SET
Lyve-SET	55/2017	FASTQ (short reads)	Yes	• Customized pipeline for different species • Phage masking	
SNVPhyl	49/2017	FASTQ (short reads), Invalid positions file (bed file)	Yes	• Mask out regions on the reference genome with variants. Masked regions will not be included in the phylogeny	

LYVE-Listeria, Yersinia, Vibrio and Enterobacteriaceae reference lab SET

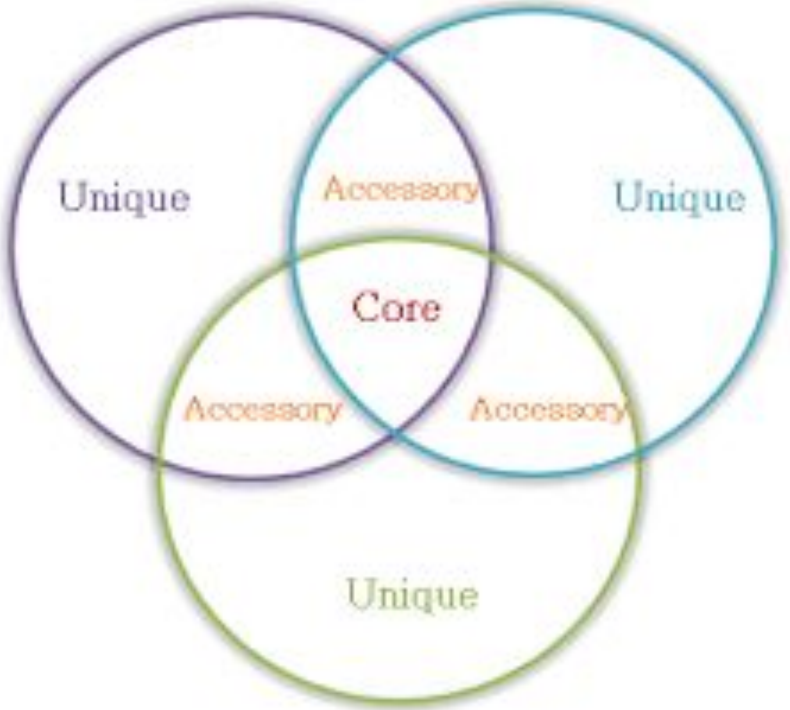


- Reads cleaning: CG-Pipeline
- Reads mapping: SMALT
- Variant calling: VarScan
- SNP matrix: bcftools
- SNP matrix to MSA
- Phylogeny: RAxML v8

The Lyve-SET workflow
Lee, et al. 2017, Front. Microbiol.

Pan-genome analysis

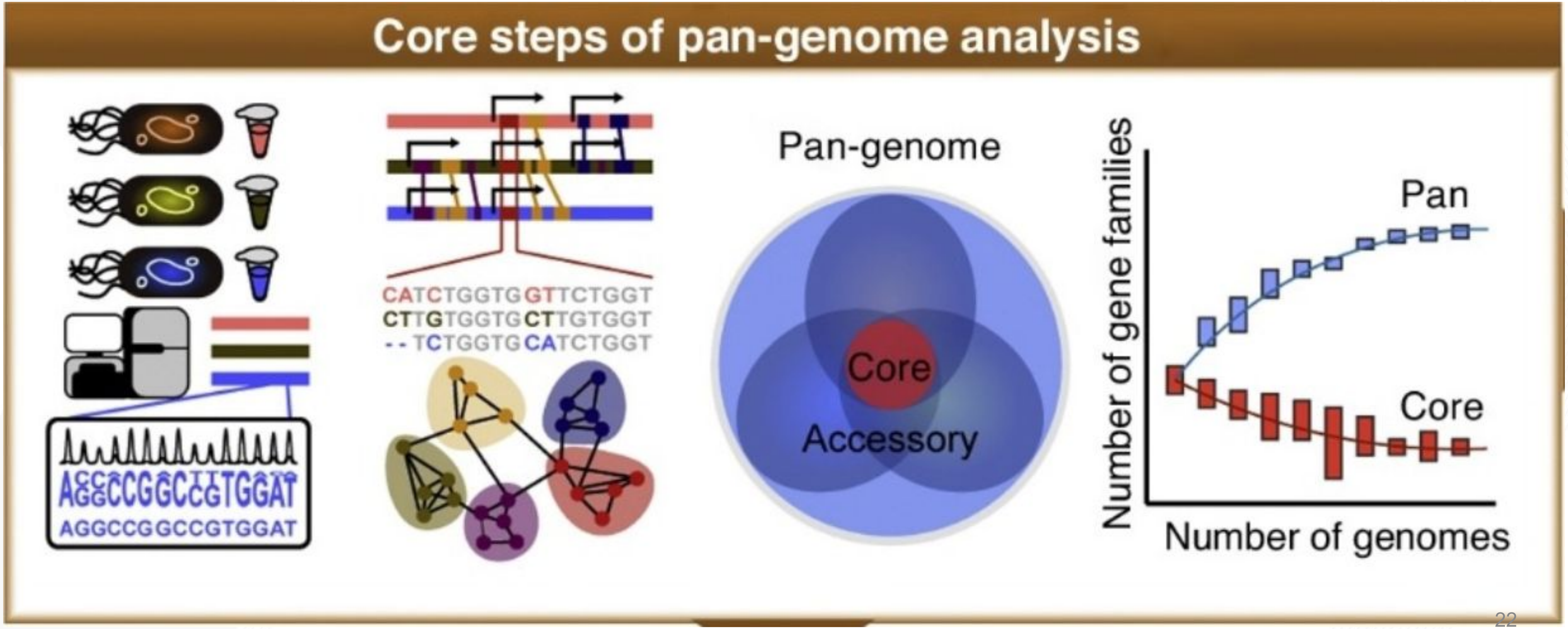
- Pan-genome
  - Pan-genome: all the genes found in the given sample set
  - Core-genome: genes shared among all samples
  - Accessory genome: pan-genome minus core-genome
- Core steps of p
- Biological infor e analysis
- Applications



Pan-genome analysis

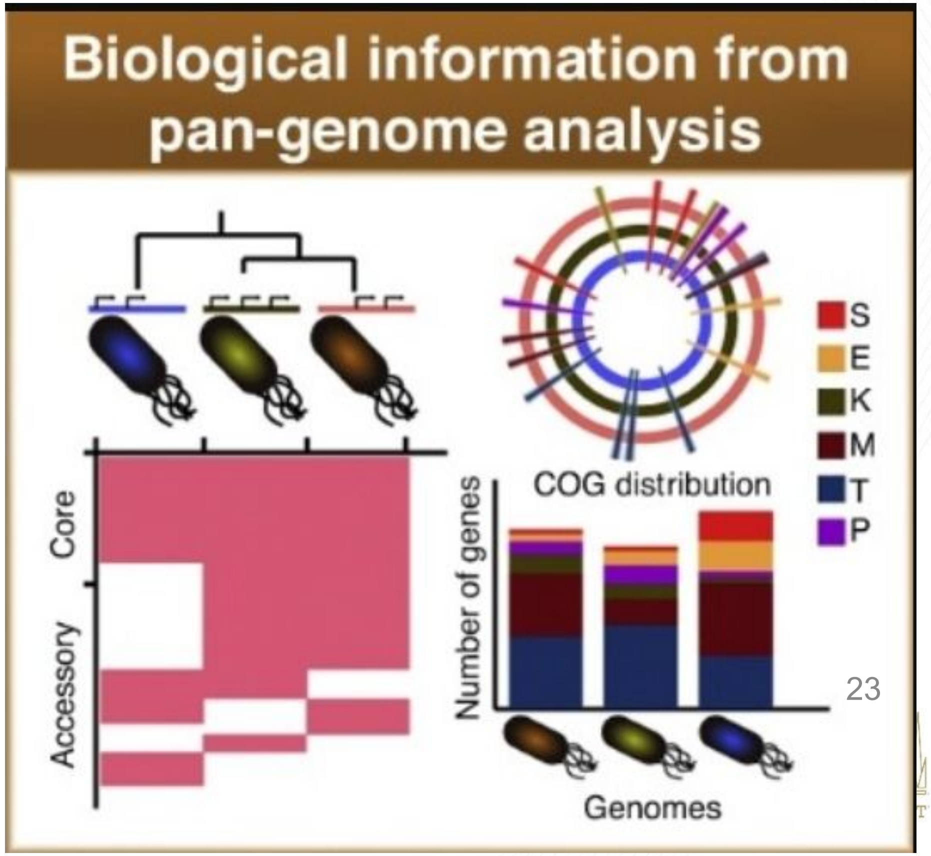
- Pan-genome
- Core steps of pan-genome analysis
  - Collection of genome data
  - Homology clustering
  - Profiling of pan- and core-genomes
- Biological information from pan-genome analysis
- Applications

Pan-genome analysis



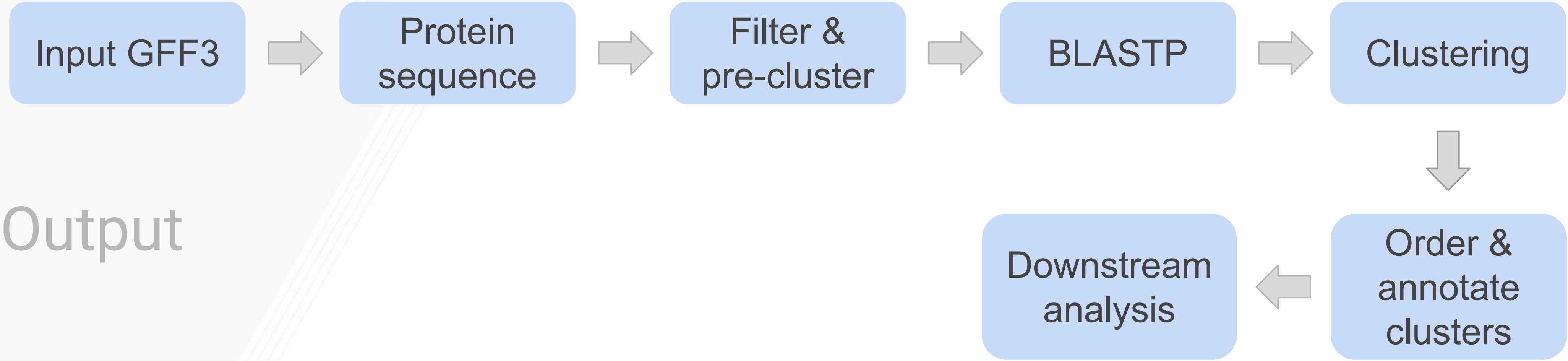
Pan-genome analysis

- Pan-genome
- Core steps of pan-genome analysis
- Biological information from pan-genome analysis
  - Phylogenetic tree
  - Presence and absence of genes
  - Functional distribution of proteins
- Applications



Tools for pan-genome analysis: Roary

- 2015; 1045 citations.
- Input: one annotated GFF3 file per sample
- Workflow:

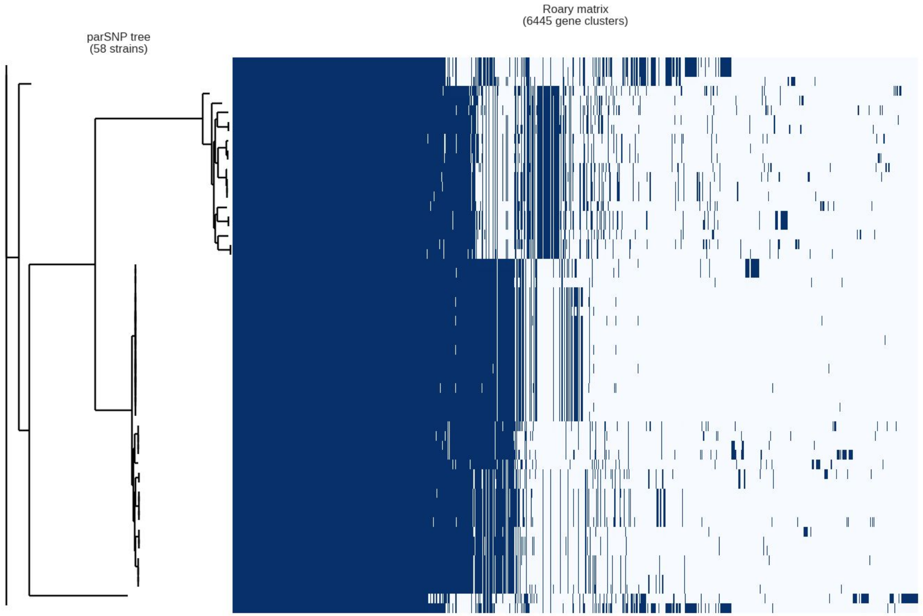


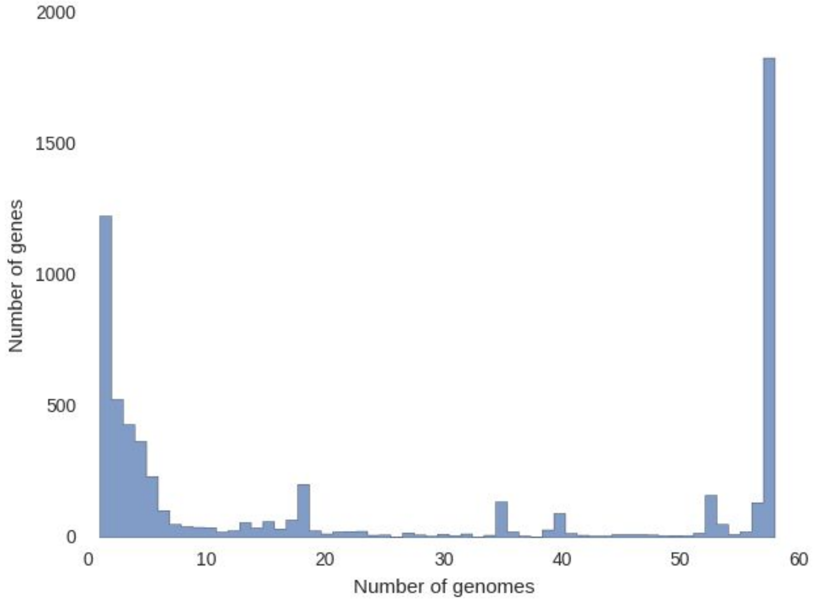
Input GFF3 → Protein sequence → Filter & pre-cluster → BLASTP → Clustering → Order & annotate clusters → Downstream analysis

- Output

Tools for pan-genome analysis: Roary

- Output:
    - Profiling of pan- and core-genomes
    - Gene presence/absence matrix
    - Representative sequence for each cluster
    - Core/accessory genome phylogenetic tree





Tools for pan-genome analysis: BPGA

- 2016; 187 citations
- Input: one protein sequence file per sample
- Workflow:

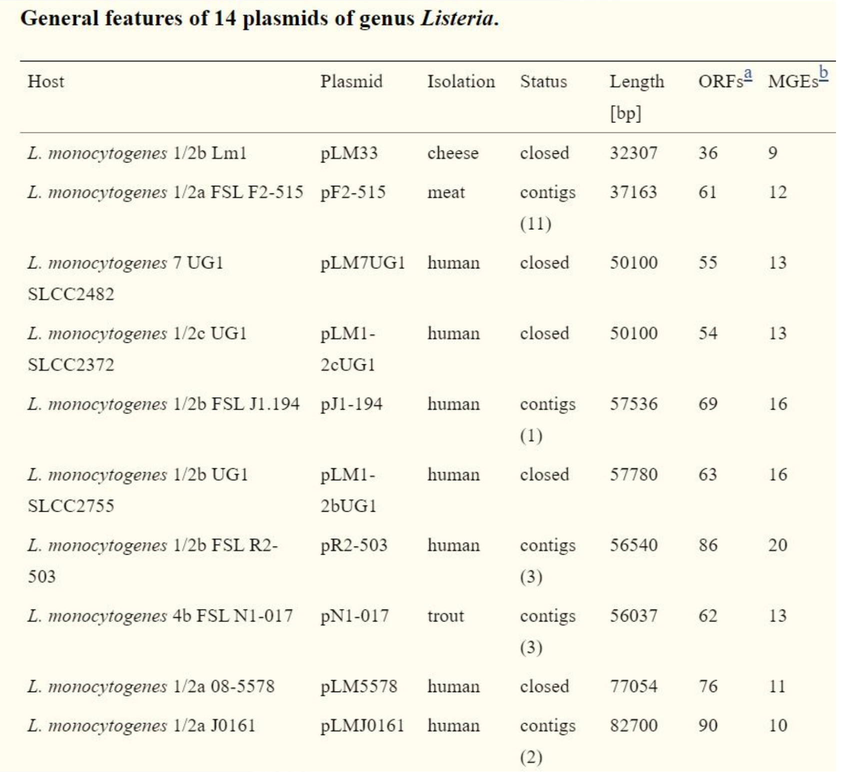
- Output

Tools for pan-genome analysis: BPGA

- Output:
  - Profiling of pan- and core-genomes
  - Representative protein sequence for each cluster
  - Gene presence/absence matrix
  - Atypical GC content
  - Gene function distribution
  - Core/accessory genome phylogenetic tree
  - ...

Plasmids in Listeria

Number of plasmids curated: >57
Length of Listeria plasmids: 30K-100K
Number of ORFs: 35 - 100
MGEs: 9-20

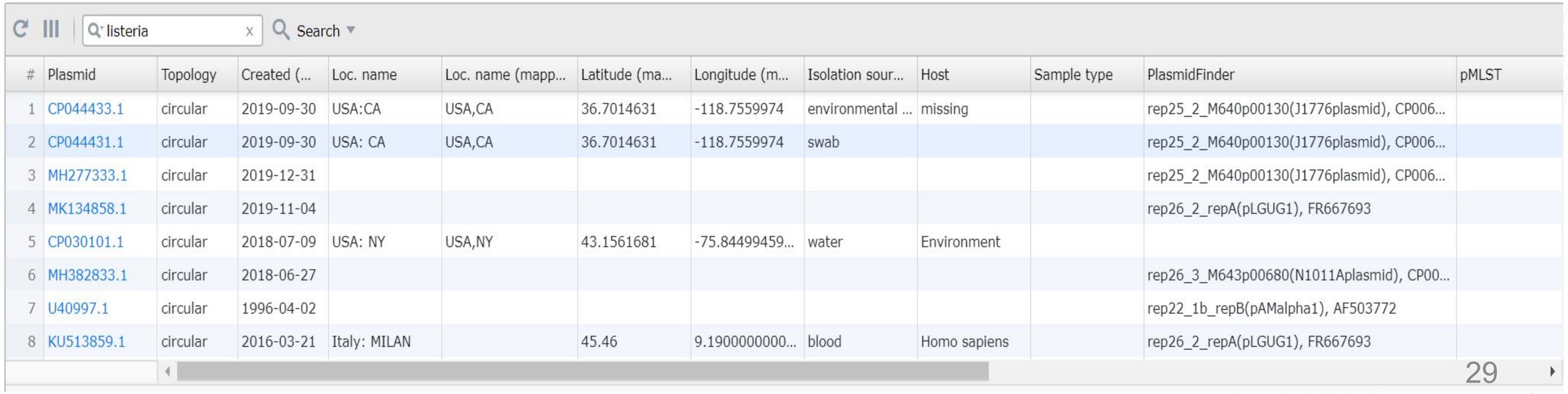


General features of 14 plasmids of genus *Listeria*.

Host	Plasmid	Isolation	Status	Length [bp]	ORFs[a]	MGEs[b]
L. monocytogenes 1/2b Lm1	pLM33	cheese	closed	32307	36	9
L. monocytogenes 1/2a FSL F2-515	pF2-515	meat	contigs (11)	37163	61	12
L. monocytogenes 7 UG1 SLCC2482	pLM7UG1	human	closed	50100	55	13
L. monocytogenes 1/2c UG1 SLCC2372	pLM1-2cUG1	human	closed	50100	54	13
L. monocytogenes 1/2b FSL J1.194	pJ1-194	human	contigs (1)	57536	69	16
L. monocytogenes 1/2b UG1 SLCC2755	pLM1-2bUG1	human	closed	57780	63	16
L. monocytogenes 1/2b FSL R2-503	pR2-503	human	contigs (3)	56540	86	20
L. monocytogenes 4b FSL N1-017	pN1-017	trout	contigs (3)	56037	62	13
L. monocytogenes 1/2a 08-5578	pLM5578	human	closed	77054	76	11
L. monocytogenes 1/2a J0161	pLMJ0161	human	contigs (2)	82700	90	10

Plasmid Databases and Tools

Databases/Tools	Year	Citations	Number of plasmids/sequences	Description	Limitations
PLSDB	2019	29	13789	Sources from **RefSeq, INSDC(DDBJ, EMBL-EBI, GenBank)**	Novel, not been used for typing before
pATLAS	2019	6	12746	web server containing comprehensive information about bacterial plasmids	online, Limited functionality
pMLST	2014	1061	769	sourced from pubMLST, updated weekly	no command line alternate



#	Plasmid	Topology	Created (...	Loc. name	Loc. name (mapp...	Latitude (ma...	Longitude (m...	Isolation sour...	Host	Sample type	PlasmidFinder	pMLST
1	CP044433.1	circular	2019-09-30	USA:CA	USA,CA	36.7014631	-118.7559974	environmental ...	missing		rep25_2_M640p00130(J1776plasmid), CP006...	
2	CP044431.1	circular	2019-09-30	USA: CA	USA,CA	36.7014631	-118.7559974	swab			rep25_2_M640p00130(J1776plasmid), CP006...	
3	MH277333.1	circular	2019-12-31								rep25_2_M640p00130(J1776plasmid), CP006...	
4	MK134858.1	circular	2019-11-04								rep26_2_repA(pLGUG1), FR667693	
5	CP030101.1	circular	2018-07-09	USA: NY	USA,NY	43.1561681	-75.84499459...	water	Environment			
6	MH382833.1	circular	2018-06-27								rep26_3_M643p00680(N1011Aplasmid), CP00...	
7	U40997.1	circular	1996-04-02								rep22_1b_repB(pAMalpha1), AF503772	
8	KU513859.1	circular	2016-03-21	Italy: MILAN		45.46	9.1900000000...	blood	Homo sapiens		rep26_2_repA(pLGUG1), FR667693	

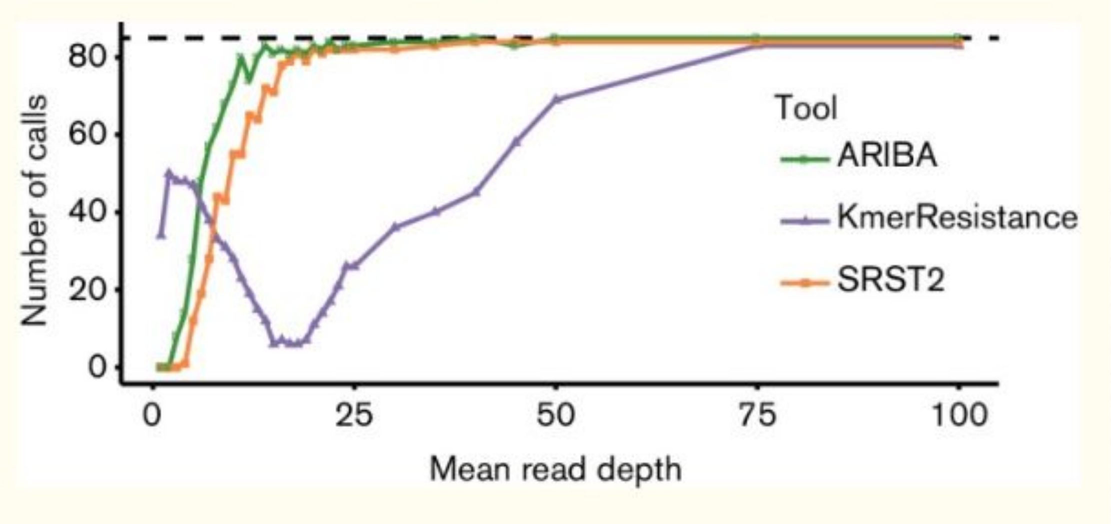
ARIBA

- Rapid antimicrobial resistance genotyping
- Uses fastq reads and can extract information relevant to both Genome and Plasmids that we might have missed out in assemblies

- ARG-ANNOT. PMID: 24145532
- CARD. PMID: 23650175
- MEGARes PMID: 27899569
- NCBI BioProject: PRJNA313047
- plasmidfinder PMID: 24777092
- resfinder. PMID: 22782487
- VFDB. PMID: 26578559
- SRST2's version of ARG-ANNOT. PMID: 25422674.
- VirulenceFinder PMID: 24574290.

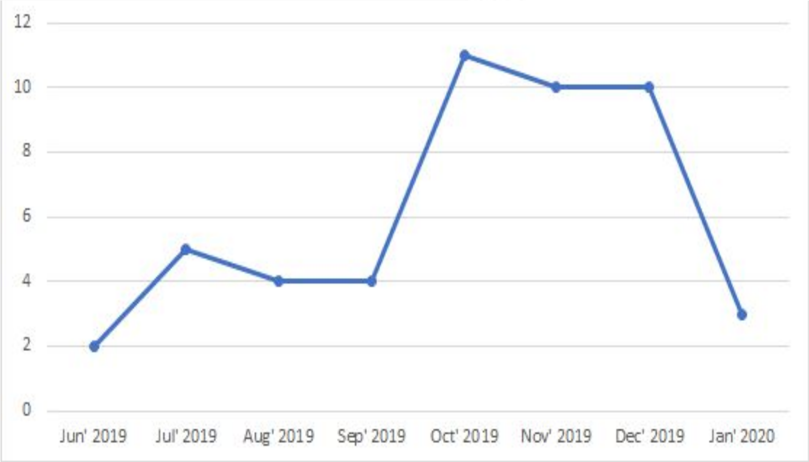
Output will be compared and combined with results obtained from annotation group

Tools	**Year**	**Citations**
ARIBA(rapid antimicrobial resistance genotyping)	2017	156
KmerResistance	2016	60
SRST2(Short read sequence typing)	2014	481

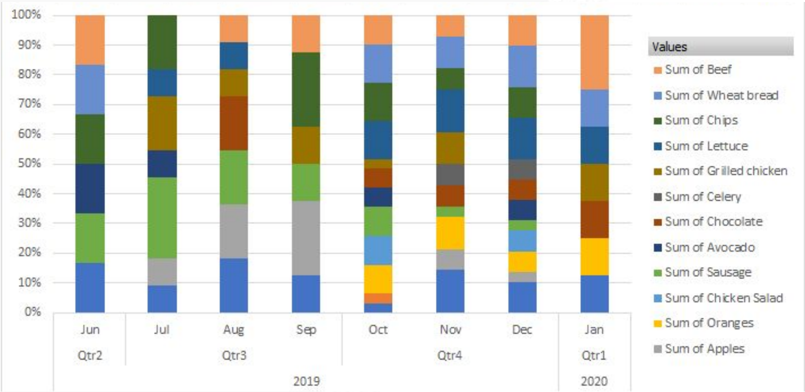


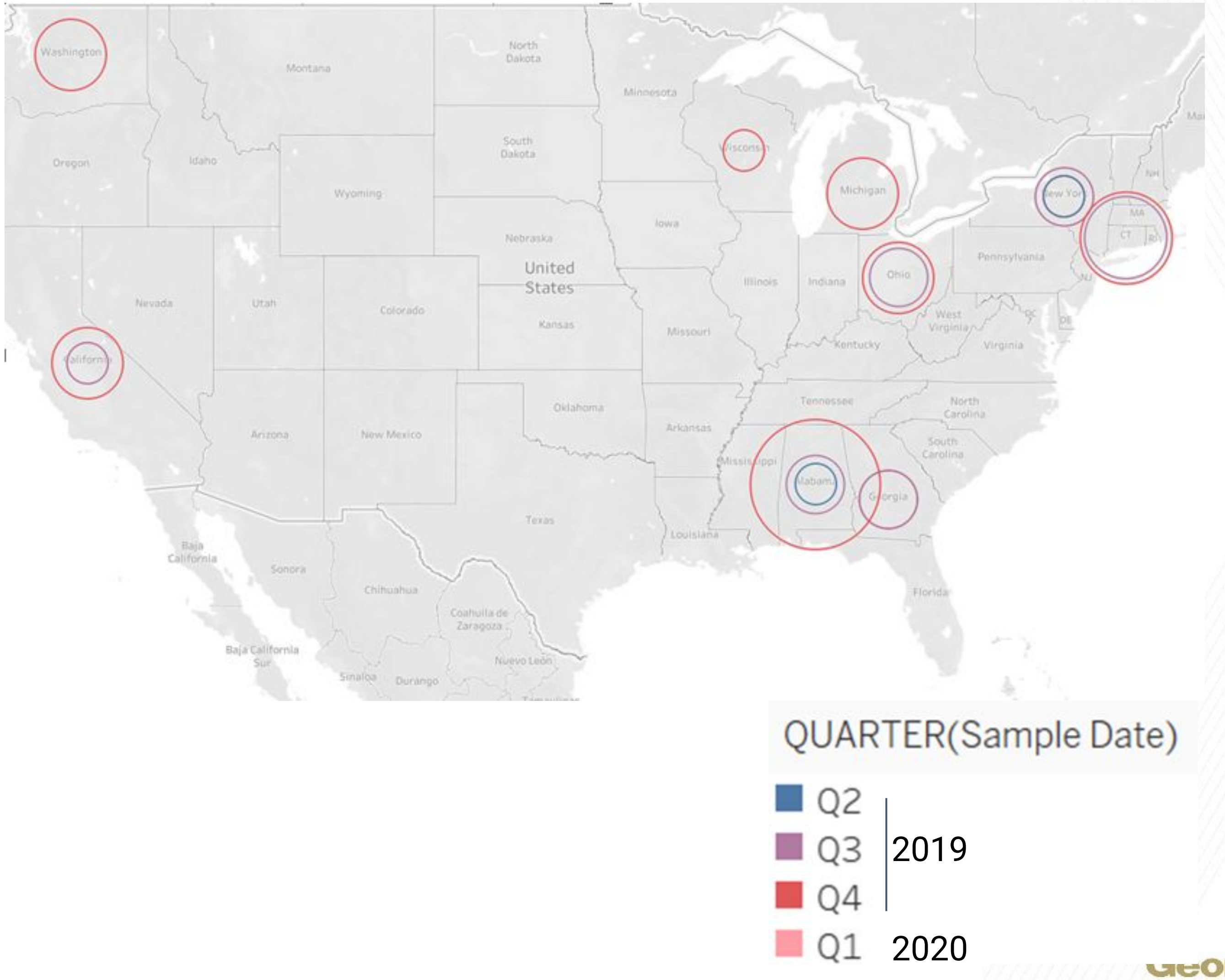
Epidemiological Data Exploration

Number of samples/cases as per timeline



Percentage of food items consumed as per timeline





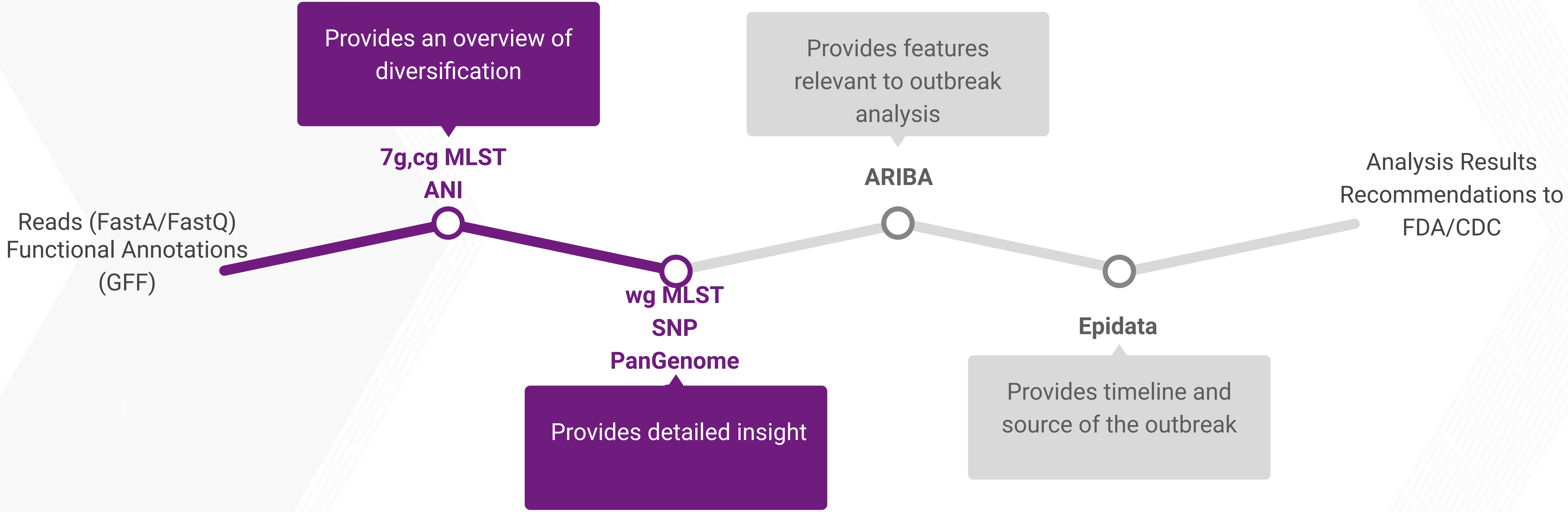
Comprehensive Analysis

- Outbreak vs Sporadic Strains
  - Combined analysis if clusters from MLST, SNP and ANI tools
  - Compare with information received from ARIBA/SRST2 and additional results from Gene Prediction/Annotation Groups

- Narrow down on the location and food source using the Epidata

- Recommendations to FDA/CDC
  - List of recommended Antibiotics based on resistance profiles
  - Further WGS analysis on the food source and imposing limitations on distribution

Comprehensive Analysis



Reads (FastA/FastQ)
Functional Annotations
(GFF)

Provides an overview of diversification

7g,cg MLST
ANI

wg MLST
SNP
PanGenome

Provides detailed insight

Provides features relevant to outbreak analysis

ARIBA

Epidata

Provides timeline and source of the outbreak

Analysis Results
Recommendations to FDA/CDC

References

- Filliol I, et al. Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J. Bacteriol. 2006;188:759–772. doi: 10.1128/JB.188.2.759-772.2006.
- Adam D. Leaché1 and Jamie R. Oaks2, The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. Annual Review of Ecology, Evolution, and Systematics. 2017; Vol. 48:69-84. https://doi.org/10.1146/annurev-ecolsys-110316-022645
- Shea N Gardner, Tom Slezak, Barry G. Hall, kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome, Bioinformatics, Volume 31, Issue 17, 1 September 2015, Pages 2877–2878, https://doi.org/10.1093/bioinformatics/btv271
- Maiden, Martin C J. “Multilocus Sequence Typing of Bacteria.” Annual Review of Microbiology, U.S. National Library of Medicine, 2006, www.ncbi.nlm.nih.gov/pubmed/16774461.
- Silva, Mickael, et al. “ChewBBACA: A Complete Suite for Gene-by-Gene Schema Creation and Strain Identification.” Microbial Genomics, Microbiology Society, Mar. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC5885018/.
- “MentaLiST.” OmicX, omictools.com/mentalist-tool.
- Feijao, Pedro, et al. “MentaLiST – A Fast MLST Caller for Large MLST Schemes.” BioRxiv, Cold Spring Harbor Laboratory, 1 Jan. 2017, www.biorxiv.org/content/10.1101/172858v2.
- Kim, Yeji, et al. “Current Status of Pan-Genome Analysis for Pathogenic Bacteria.” Current Opinion in Biotechnology, vol. 63, 2020, pp. 54–62., doi:10.1016/j.copbio.2019.12.001.
- Page, Andrew J., et al. “Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis.” Bioinformatics, vol. 31, no. 22, 2015, pp. 3691–3693., doi:10.1093/bioinformatics/btv421.
- Chaudhari, Narendrakumar M., et al. “BPGA- an Ultra-Fast Pan-Genome Analysis Pipeline.” Scientific Reports, vol. 6, no. 1, 2016, doi:10.1038/srep24373.
- Valentina Galata, Tobias Fehlmann, Christina Backes, Andreas Keller, PLSDB: a resource of complete bacterial plasmids, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D195–D202

References

- Hunt, Martin et al. “ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads.” *Microbial genomics* vol. 3,10 e000131. 4 Sep. 2017, doi:10.1099/mgen.0.000131
- Annaleise Wilson et al. “Phenotypic and Genotypic Analysis of Antimicrobial Resistance among *Listeria monocytogenes* Isolated from Australian Food Production Chains”. Feb 9, 2018. Genes doi: 10.3390/genes9020080
- Clementine Henri et al “An Assessment of Different Genomic Approaches for Inferring Phylogeny of *Listeria monocytogenes”*Front. Microbiol., 29 November 2017 | https://doi.org/10.3389/fmicb.2017.02351
- Yi Chen et al “Core Genome Multilocus Sequence Typing for Identification of Globally Distributed Clonal Groups and Differentiation of Outbreak Strains of *Listeria monocytogenes”* Appl Environ Microbiology, 2016 Oct 15 doi: 10.1128/AEM.01532-16

Thankyou!