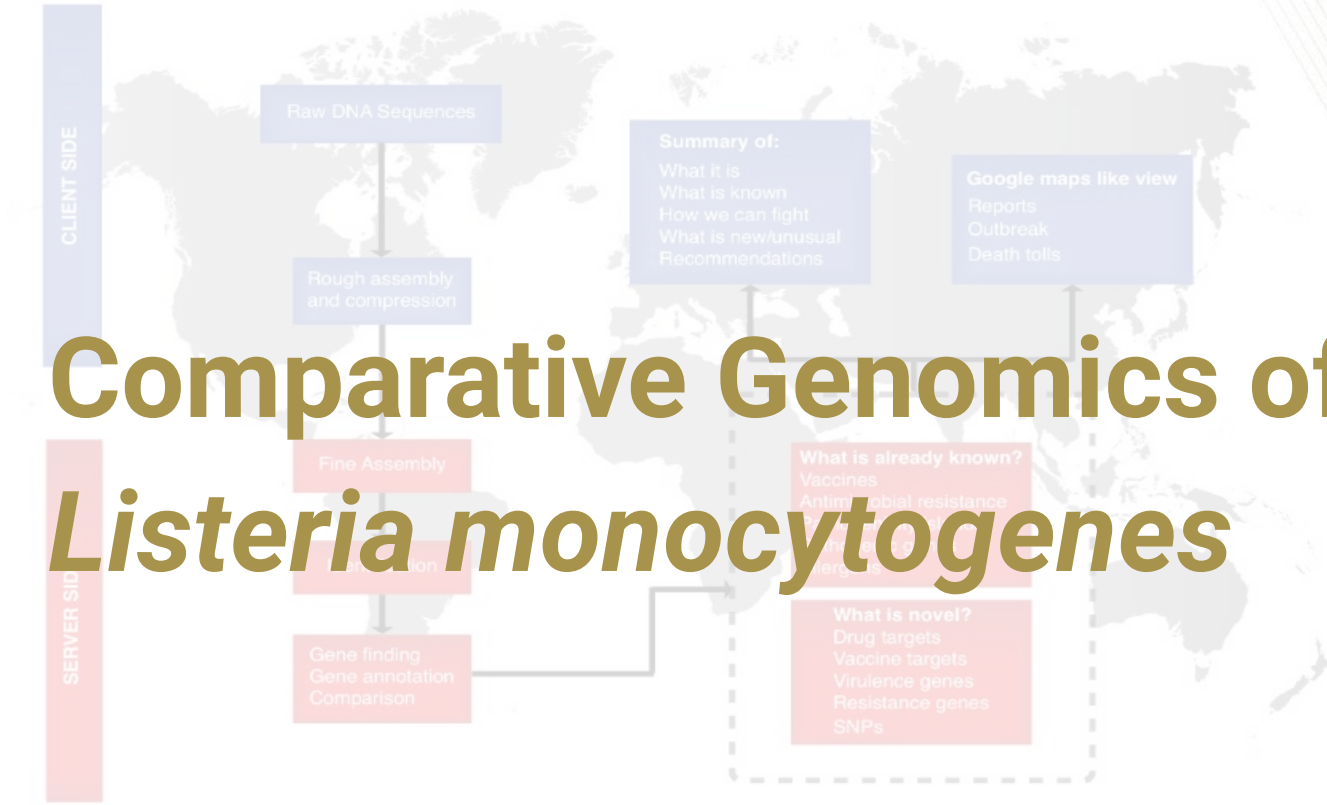


Comparative Genomics of *Listeria monocytogenes*

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

Information at hand - Analysis from previous groups

- Raw fastq, assembled files, gene prediction files

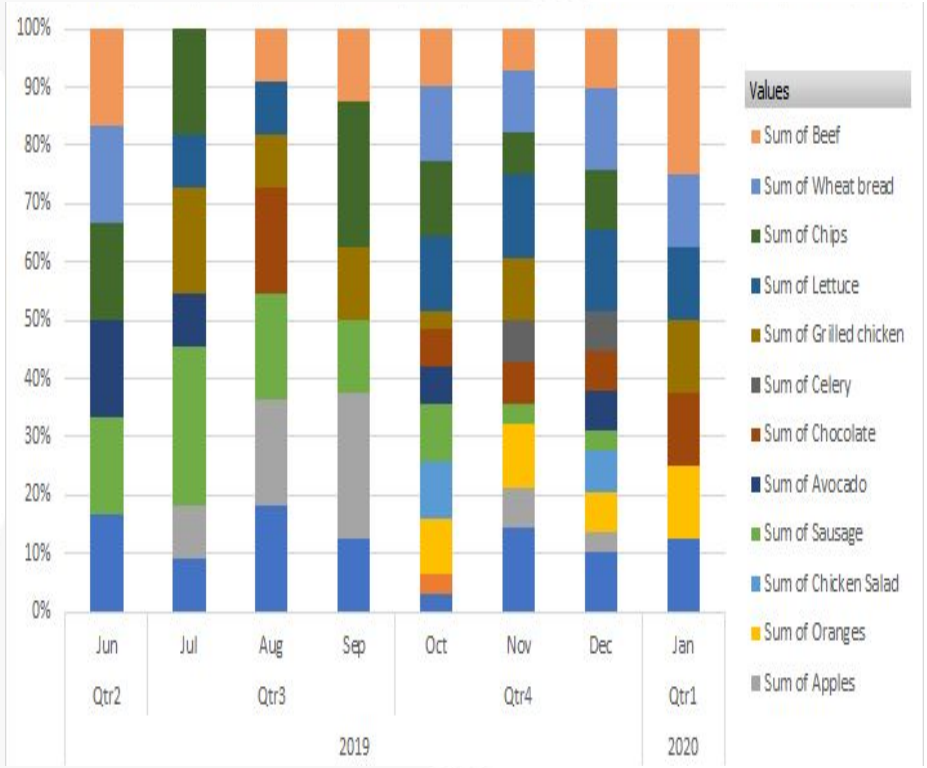
- Gff files from functionally annotation team for both genome and plasmids:
  - Merged gff files

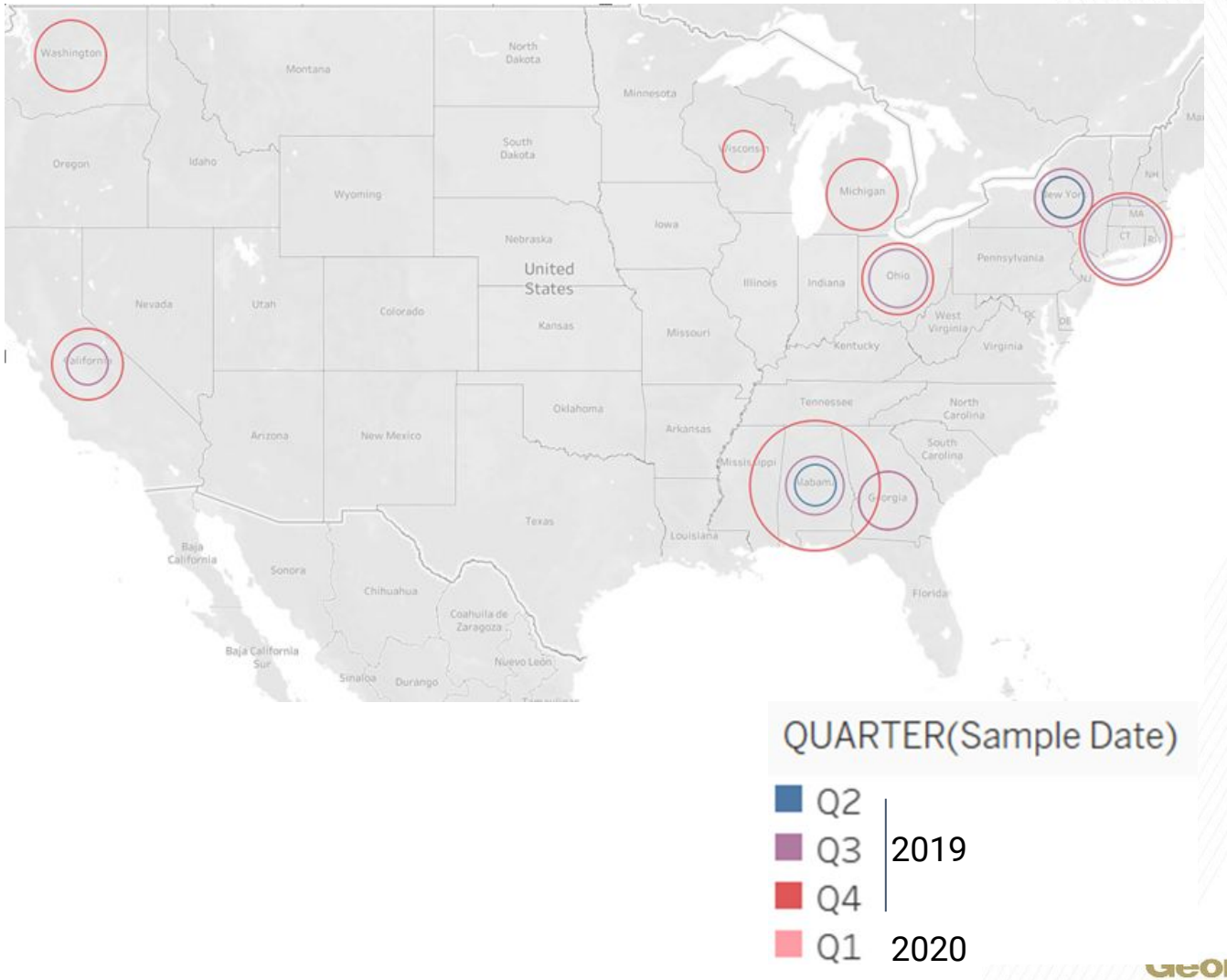
  - Gff- Virulence factors - VFDB [Virulence Factor Database]

  - Gff - Antibiotic resistance - CARD [Comprehensive Antibiotic Resistance Database]

Information at hand - Epidemiological Data

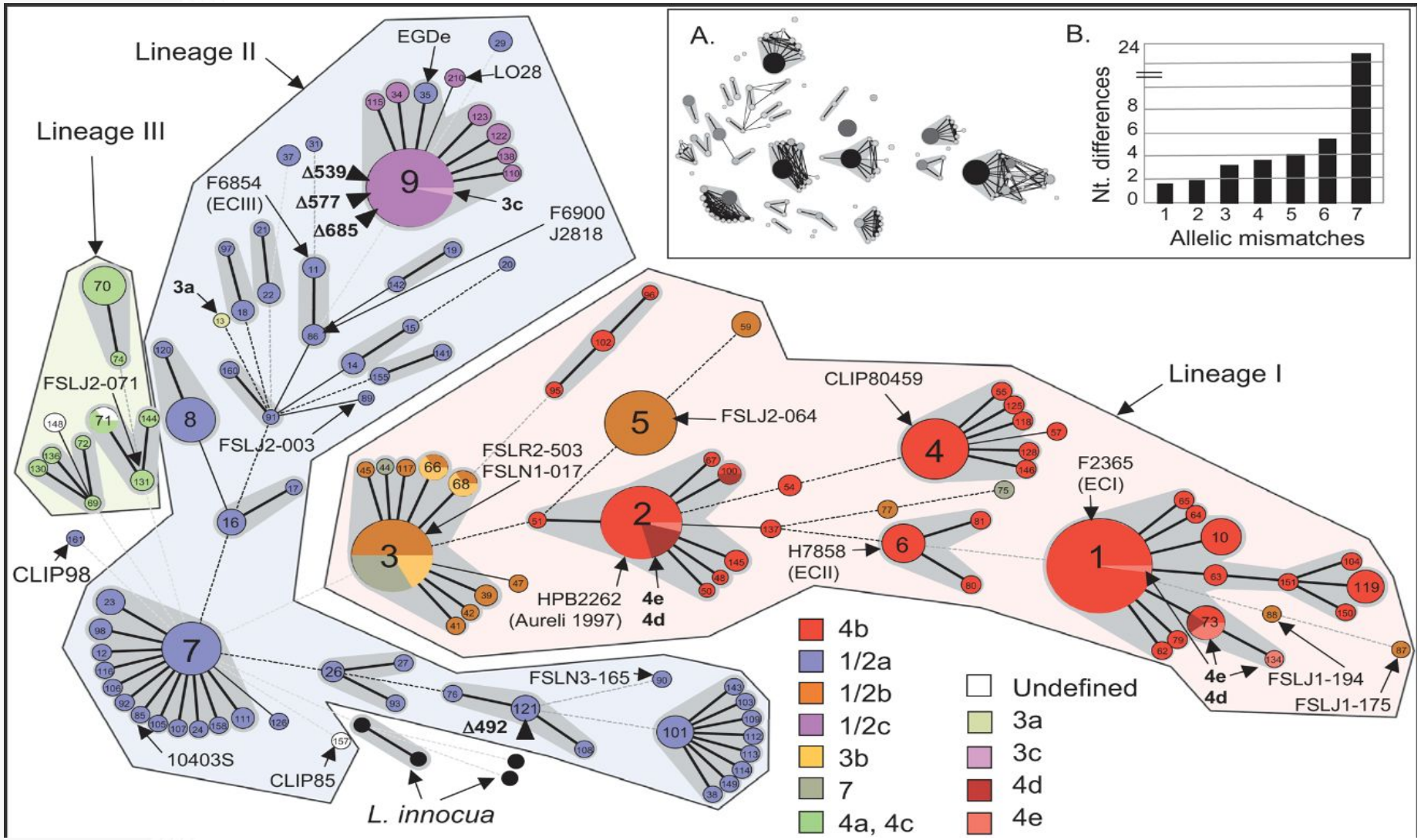
Percentage of food items consumed as per timeline





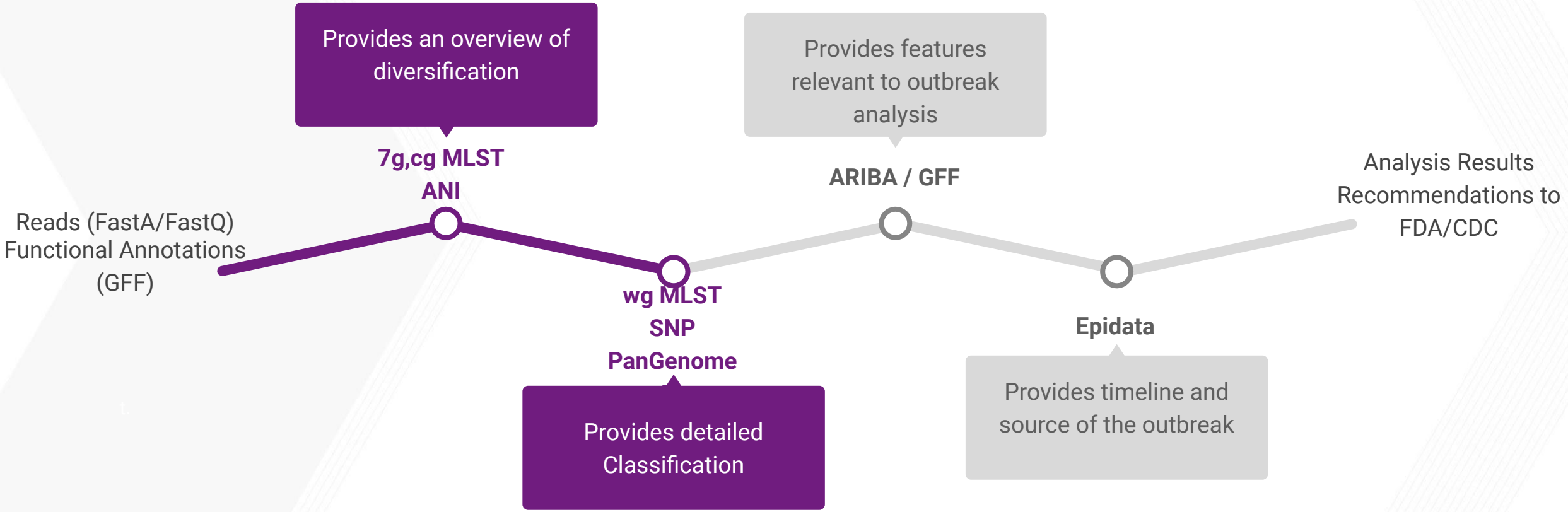
Georgia Tech
CREATING THE NEXT

What we tried to analyze?



Picture from: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1000146

Comparative Genomics Pipeline



Georgia Tech
CREATING THE NEXT

Average Nucleotide Identity (ANI)

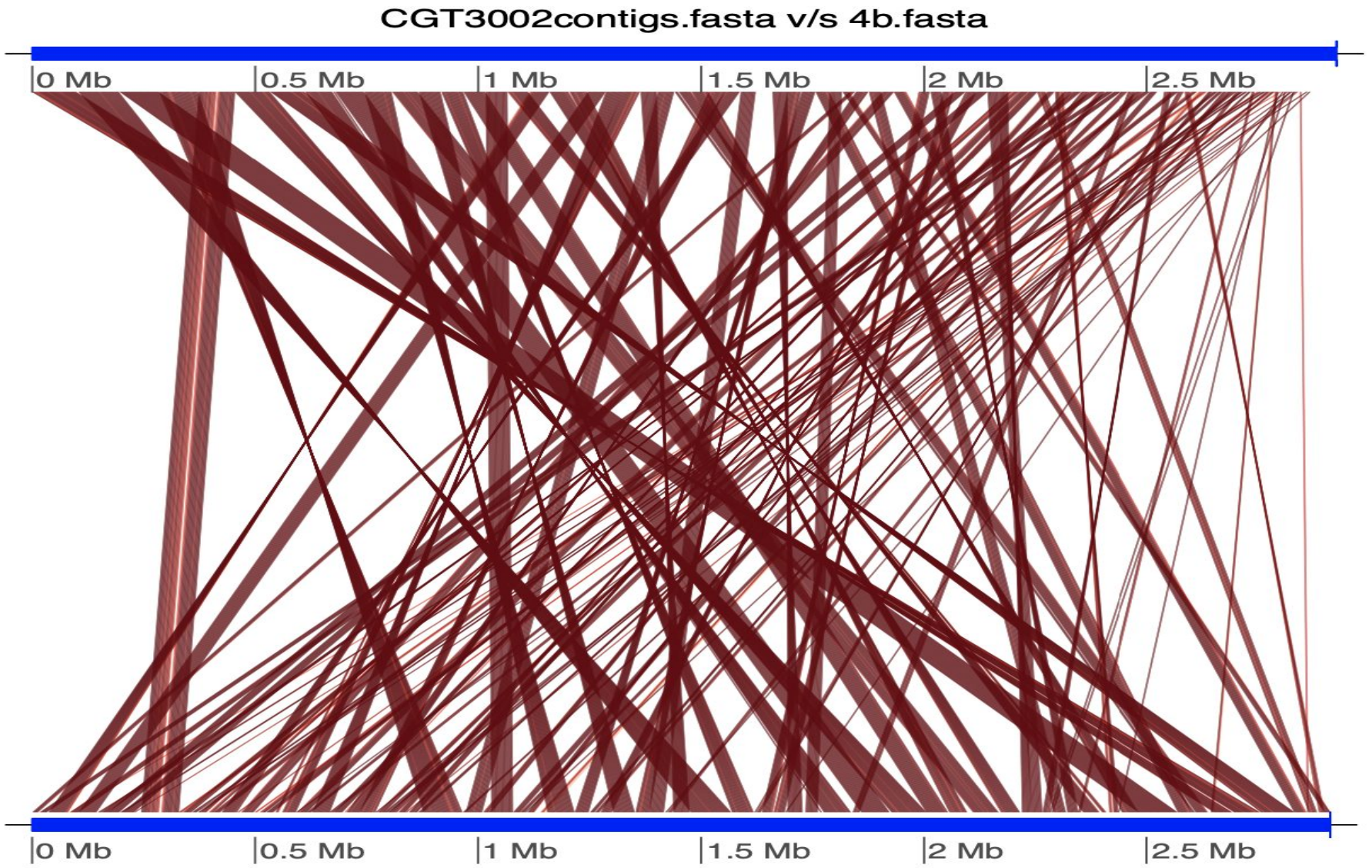
- We used FastANI
- Command line:

  fastANI --ql query.txt --rl ref.txt -o output.csv
- Using Listeria (serotype: 1/2a, 1/2b, 4b), Campylobacter and COVID-19 as reference genome.
- The result shows that Listeria (serotype: 4b) has the highest average ANI value.

ANI results

Species	Average ANI
Listeria 1/2a	99.443%
Listeria 1/2b	94.736%
Listeria 4b	99.641%
Campylobacter	Below 80%
COVID-19	Below 80%

ANI result



Tool 1: StringMLST

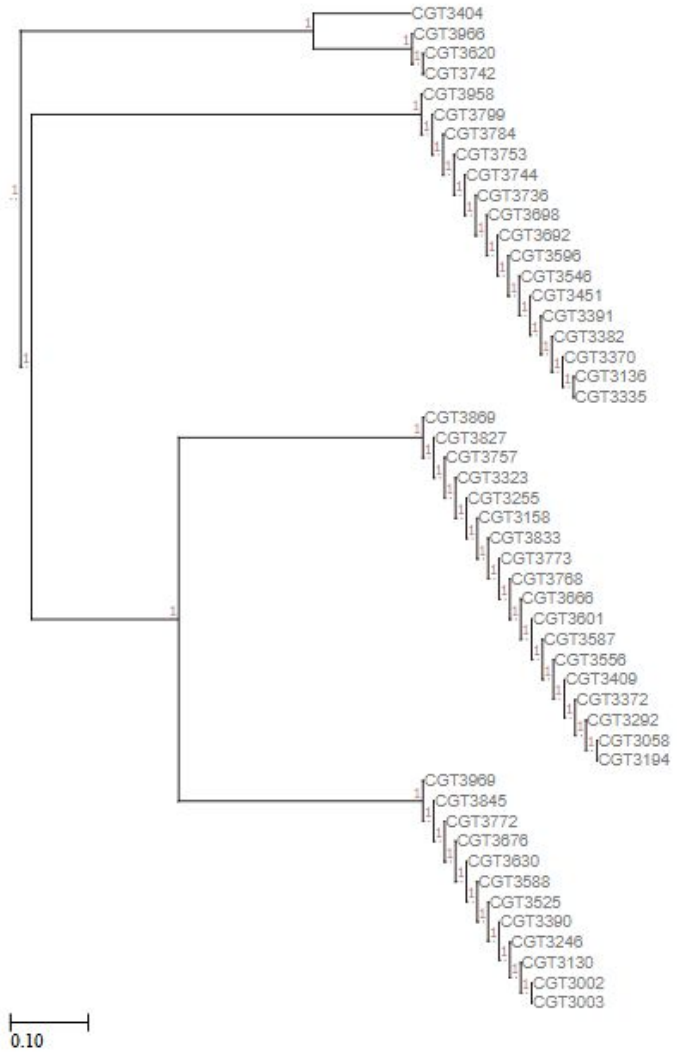
- Input: raw FASTQ files
- 7 housekeeping genes
- Used existing PubMLST schema of Listeria monocytogenes

  stringMLST.py --buildDB
- Output format:

  stringMLST.py --predict

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

Phylogenetic Tree from 7-gene StringMLST



Based on the traditional MLST analysis, there are 5 distinct sequence types among our 50 samples.

Listeria monocytogenes Sequence Types:

219 (1 sample)

397 (3 samples)

1 (18 samples)

37 (16 samples)

6 (12 samples)

Tool 2: ChewBBACA

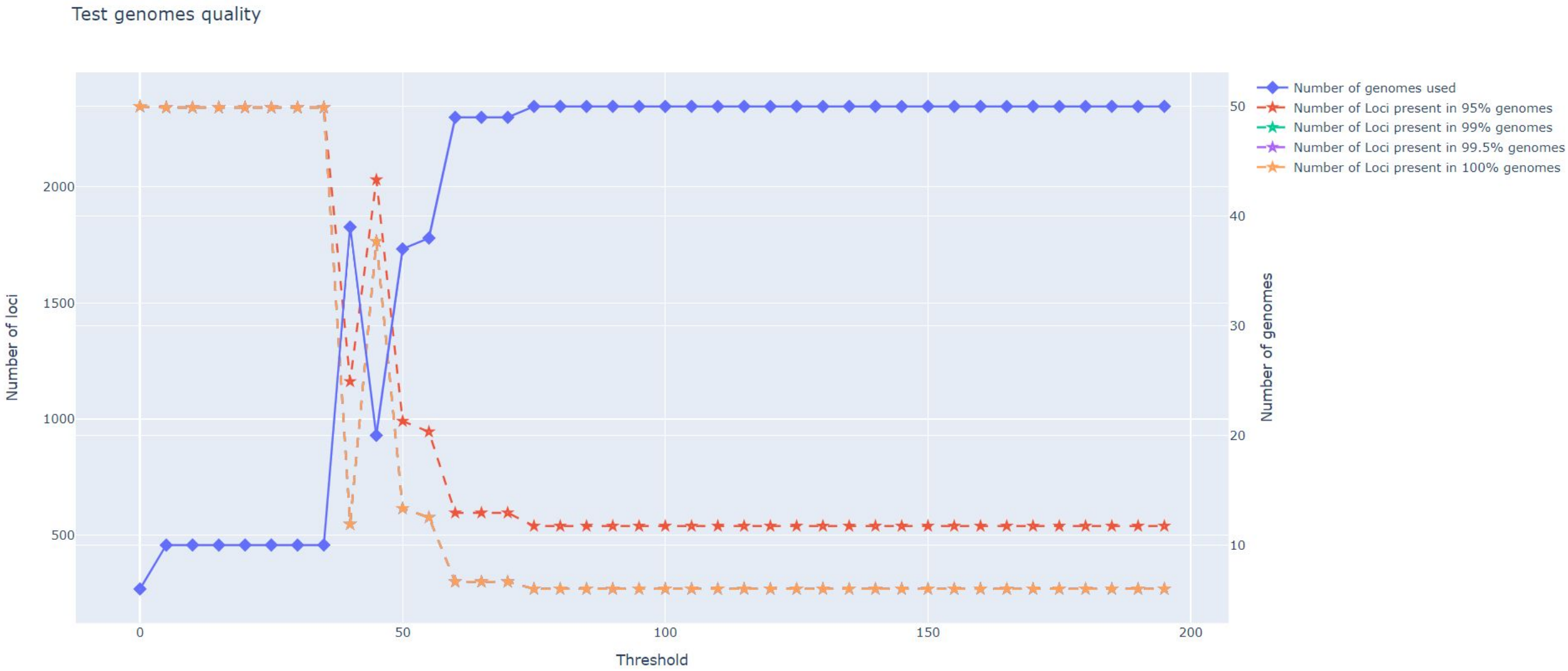
- 2997 loci in total, 540 loci used for cgMLST
- Input: Merged FASTA files from Gene Prediction group
- Construct allele schema based genes from all isolates

  chewBBACA.py CreateSchema
- Calling alleles from the schema

  chewBBACA.py AlleleCall
- Run MLST analysis only with the loci present in 95% of the matrix

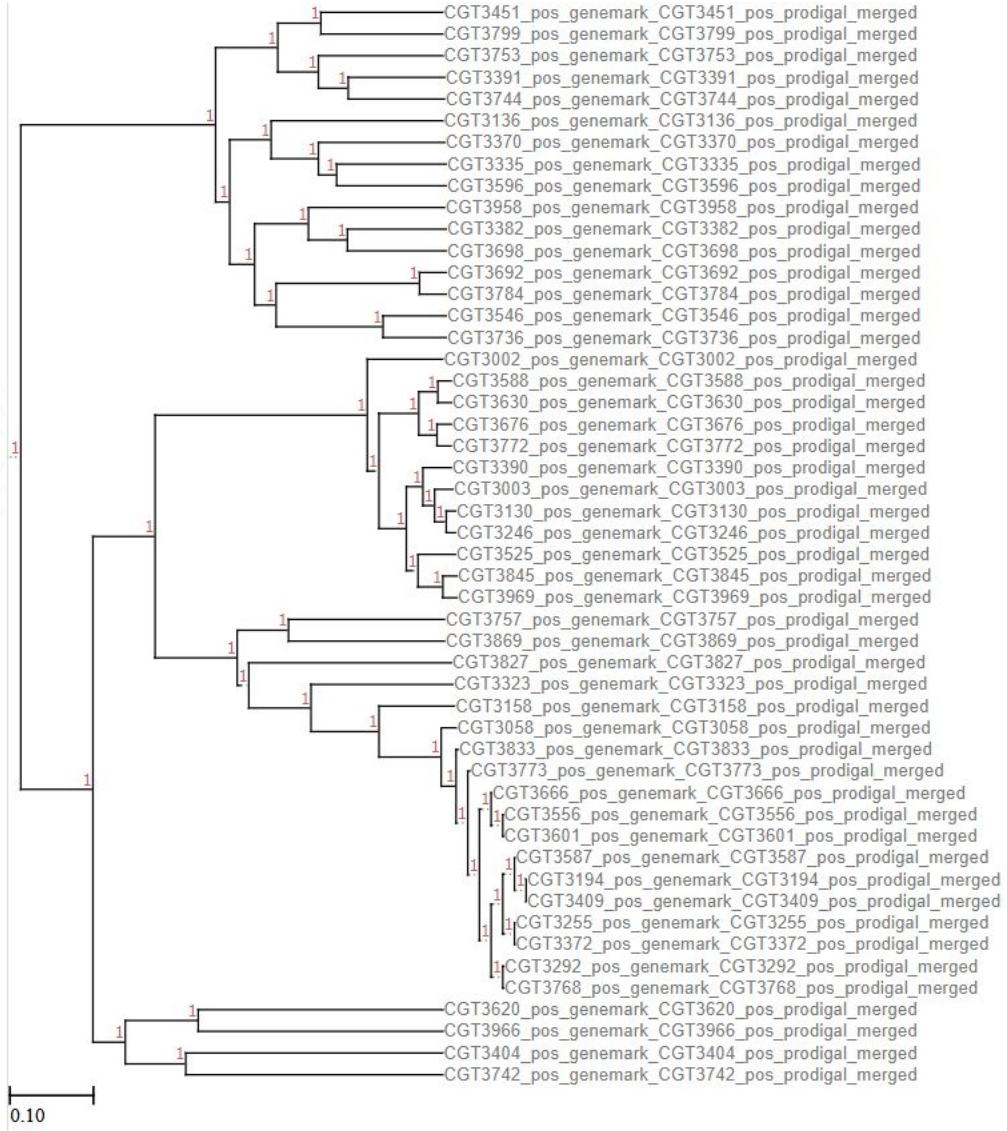
  chewBBACA.py ExtractCgMLST

ChewBBACA



CREATING THE NEXT

Phylogenetic Tree from ChewBBACA cgMLST



Georgia Tech
CREATING THE NEXT

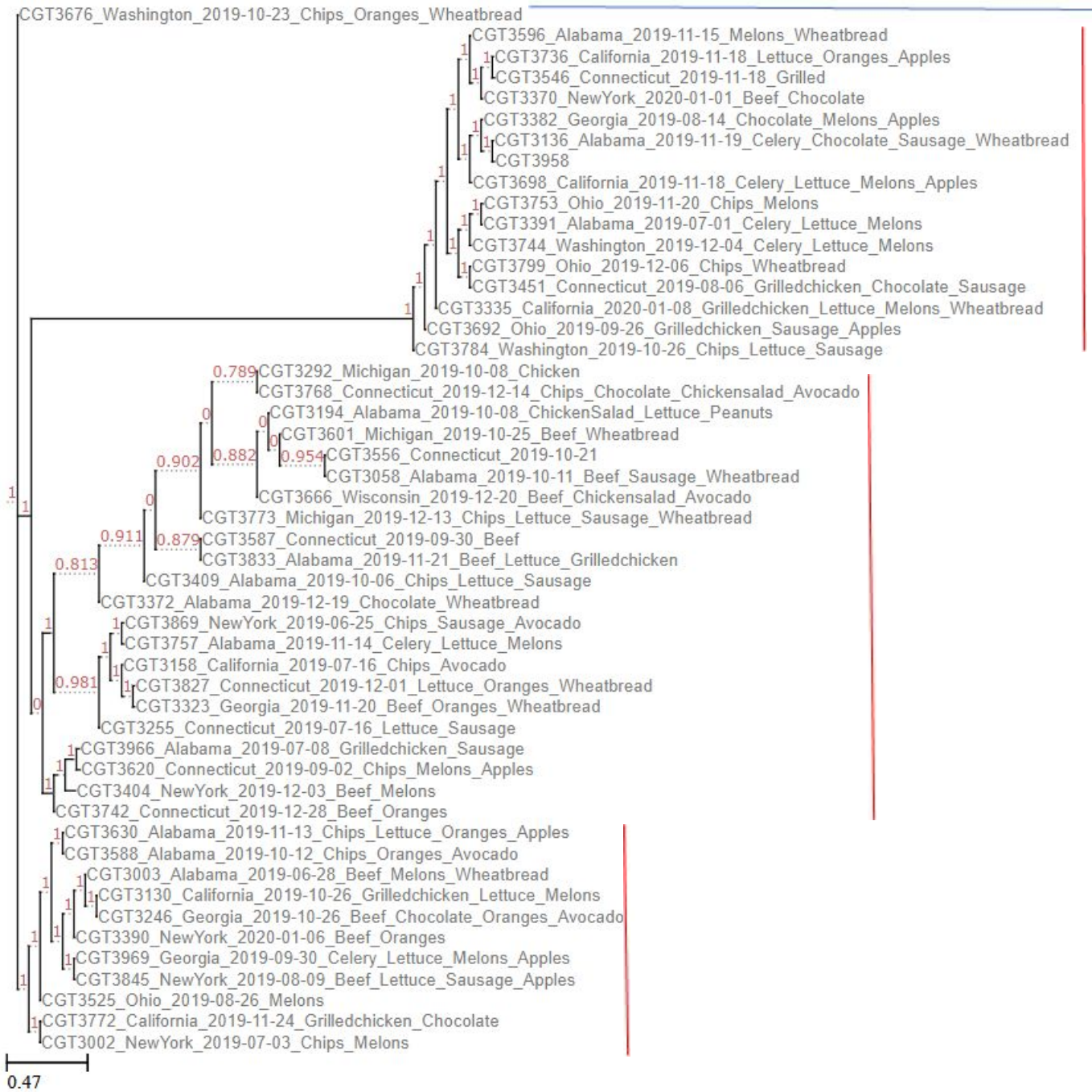
SNP-based Typing

kSNP	Output	Best k
• input • k-mer • less memory	• lower resolution • clustering	• 19 • 99.74%

```
(base) [yzhang3466@biogenome2020 SNP]$ cat Kchooser.report
Initial value of k is 13.
When k is 13 0.872395562926884 of the kmers from the median length sequence are unique.
When k is 15 0.981747630863476 of the kmers from the median length sequence are unique.
When k is 17 0.995887747660249 of the kmers from the median length sequence are unique.
The optimum value of K is 19.
When k is 19 0.997407662620663 of the kmers from the median length sequence are unique.

There were 50 genomes.
The median length genome was 2886883 bases.
The time used was 641 seconds

From a sample of 997 unique kmers  594 are core kmers.
0.595787362086259 of the kmers are present in all genomes.
```



Maximum Parsimony Tree

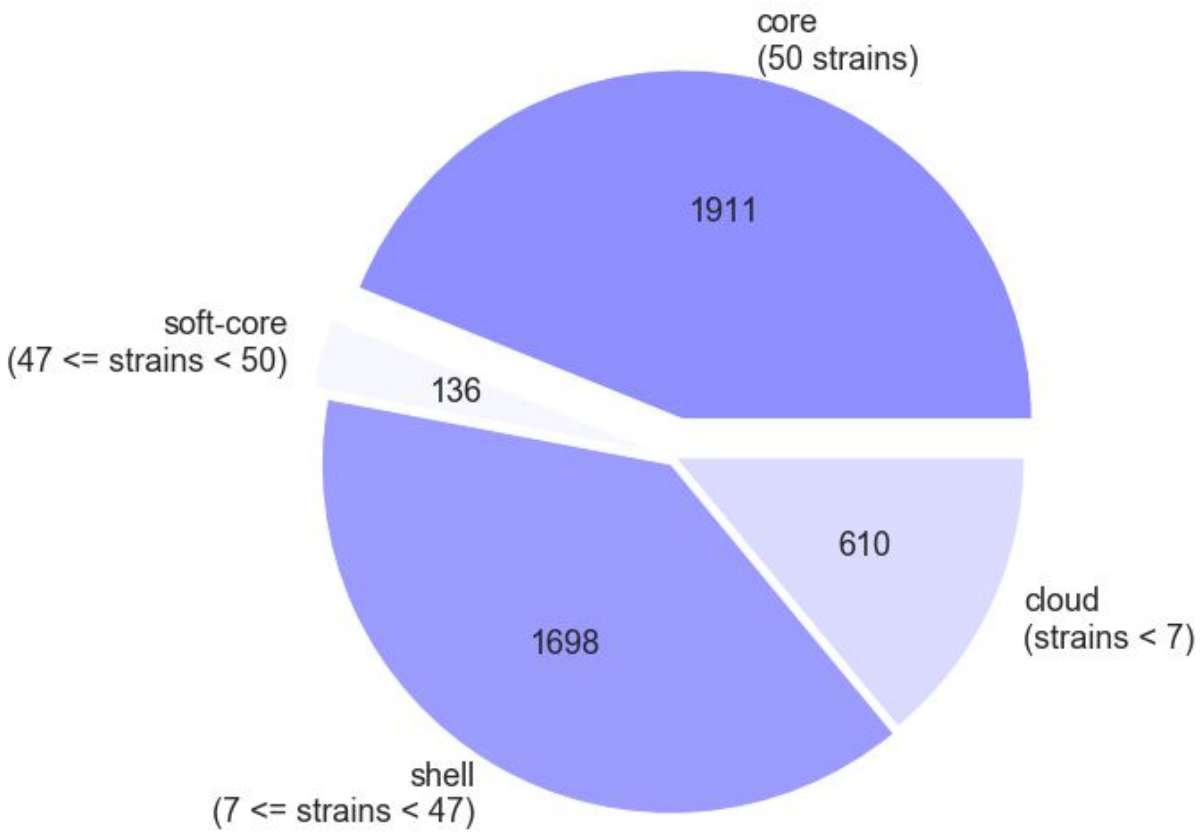
- Highest accuracy
- Fewest evolutionary change
- Fail to take into account many factors of sequence evolution
- 3 clusters
- Exclude 1 isolate

Georgia Tech
CREATING THE NEXT

Pan-genome analysis

- Roary:
  - Input: GFF file from PROKKA
  - Command line: roary -f output -i 95 -cd 96 -r *.gff
- BPGA:
  - Input: FASTA file from PROKKA
  - Command line interface

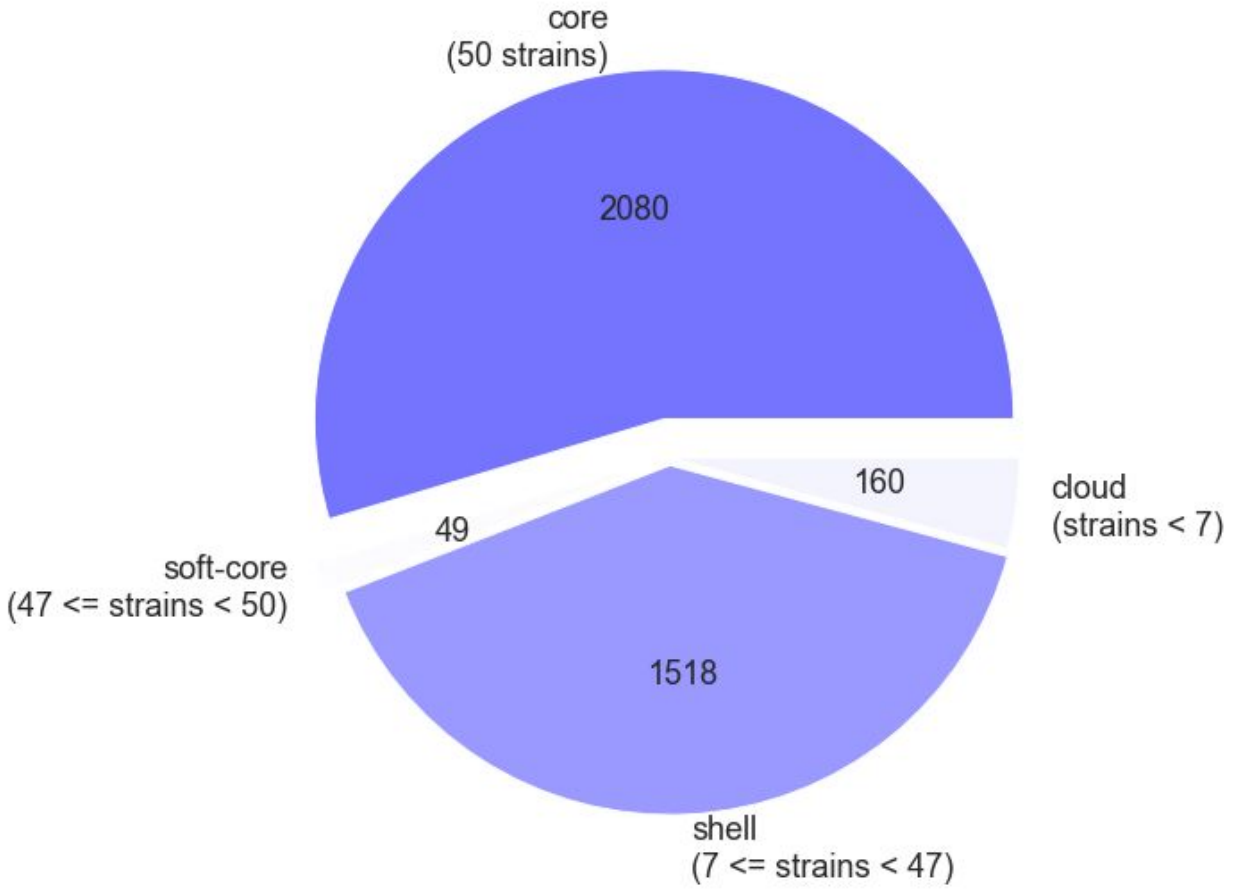
Pan-genome analysis



by Roary

# of genes in pan-genome: 4356

# of genes in core-genome: 1911

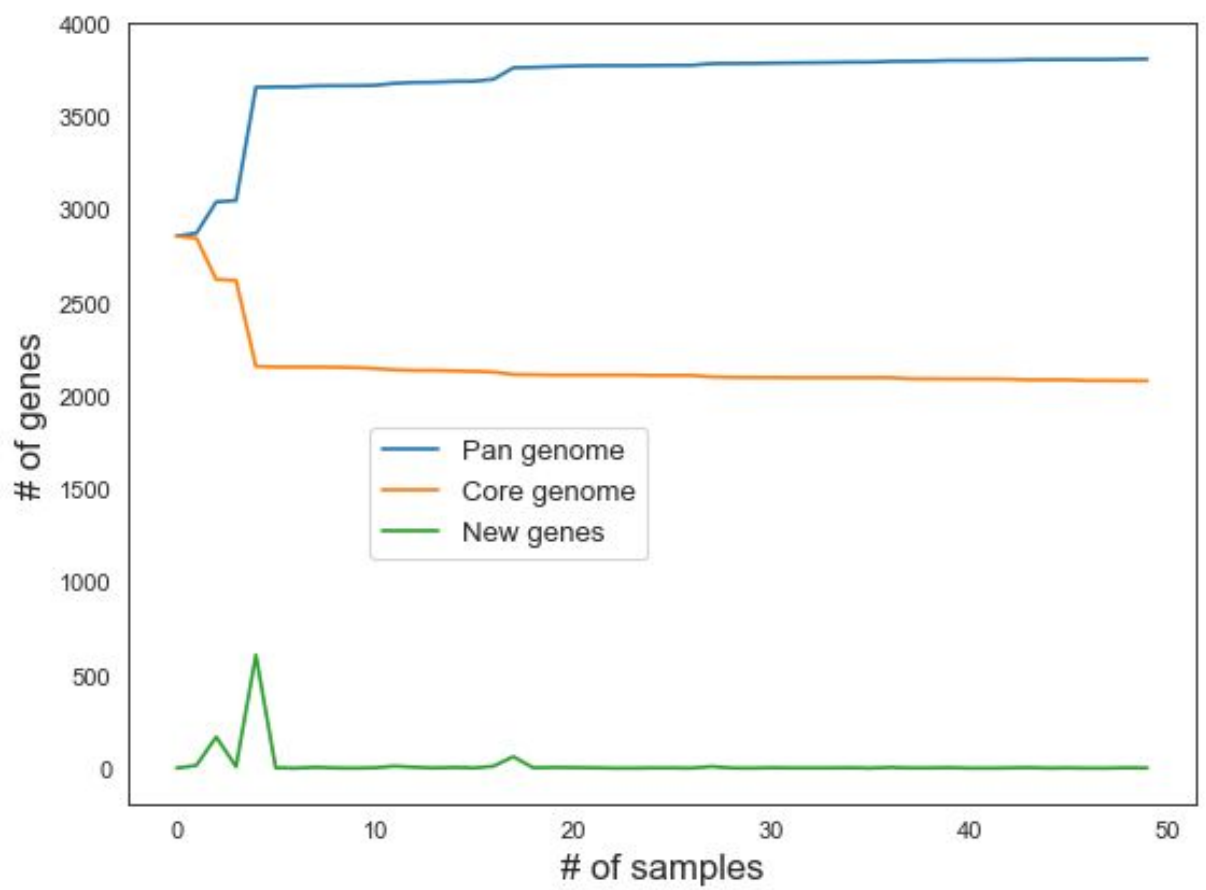


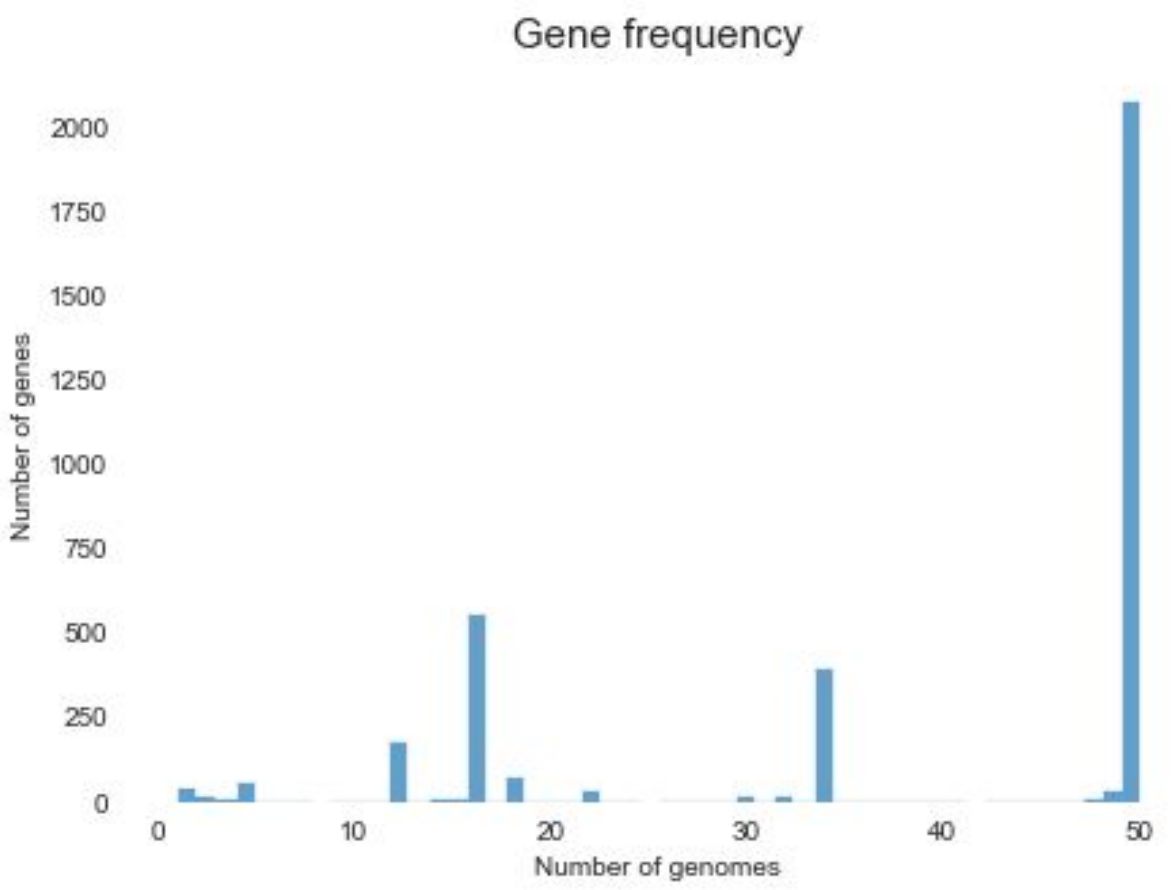
by BPGA

# of genes in pan-genome: 3808

# of genes in core-genome: 2080

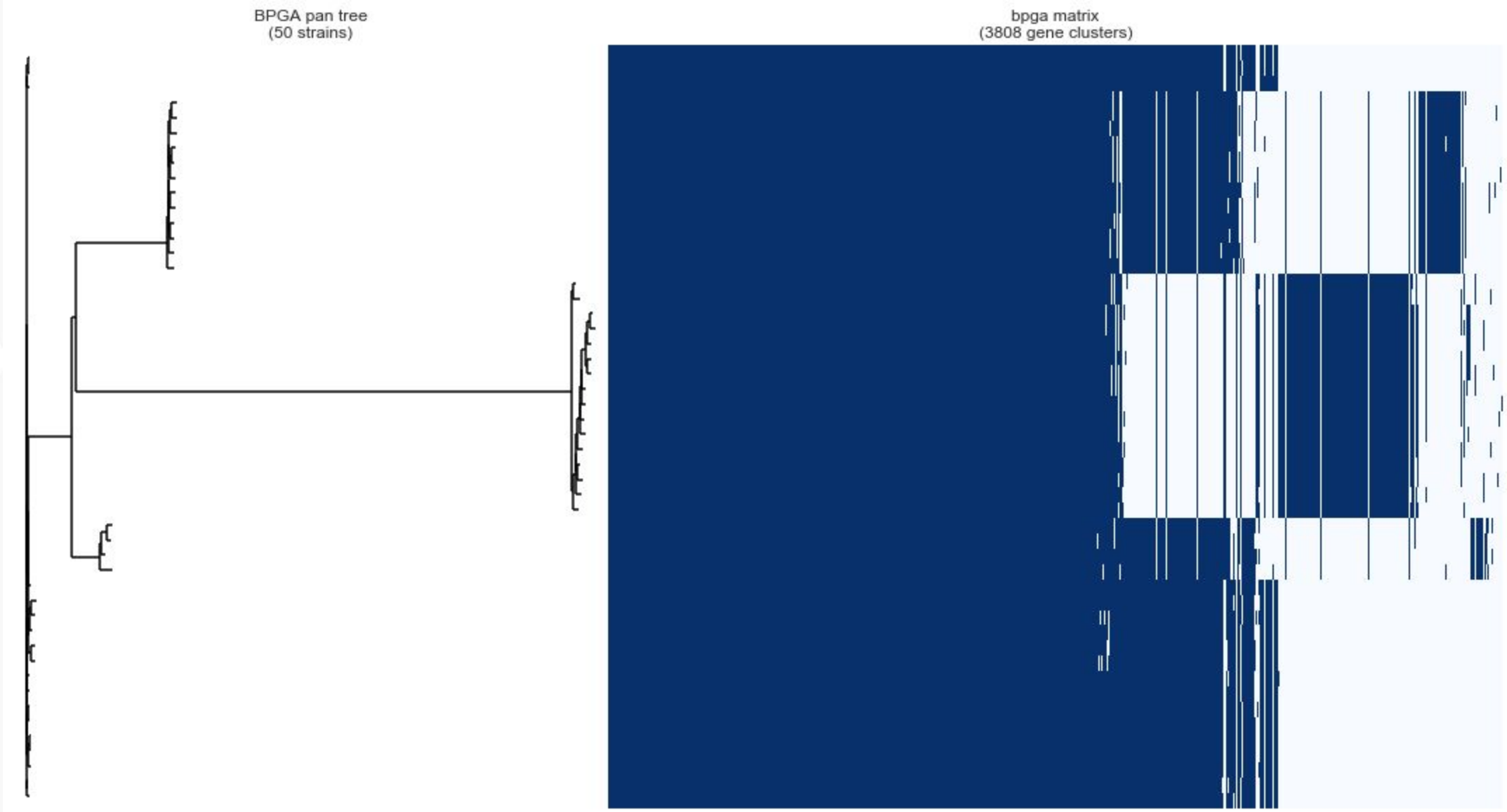
Pan-genome analysis



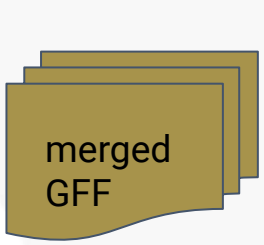


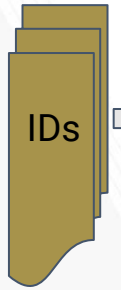
# of genes in pan-genome: 3808
# of genes in core-genome: 2080

Pan-genome analysis



Information extraction from merge annotated data





0	lmo1421	lmo2557	LMOSA_24590	LMOSA_1270	lmo0319	LMOSA_28520	LMC
CGT3002_merged_	0	1	1	0	0	1	1	1	1
CGT3003_merged_	0	1	1	0	1	1	1	1	1
CGT3058_merged_	0	1	1	1	1	1	1	1	1
CGT3130_merged_	0	1	1	1	0	1	1	1	1
CGT3136_merged_	1	1	1	0	0	0	0	0	1
CGT3158_merged_	0	1	1	1	1	1	1	1	1
CGT3194_merged_	0	0	1	1	0	1	1	1	1
CGT3246_merged_	0	0	1	0	0	1	1	1	1
CGT3255_merged_	0	0	1	1	0	1	1	1	1

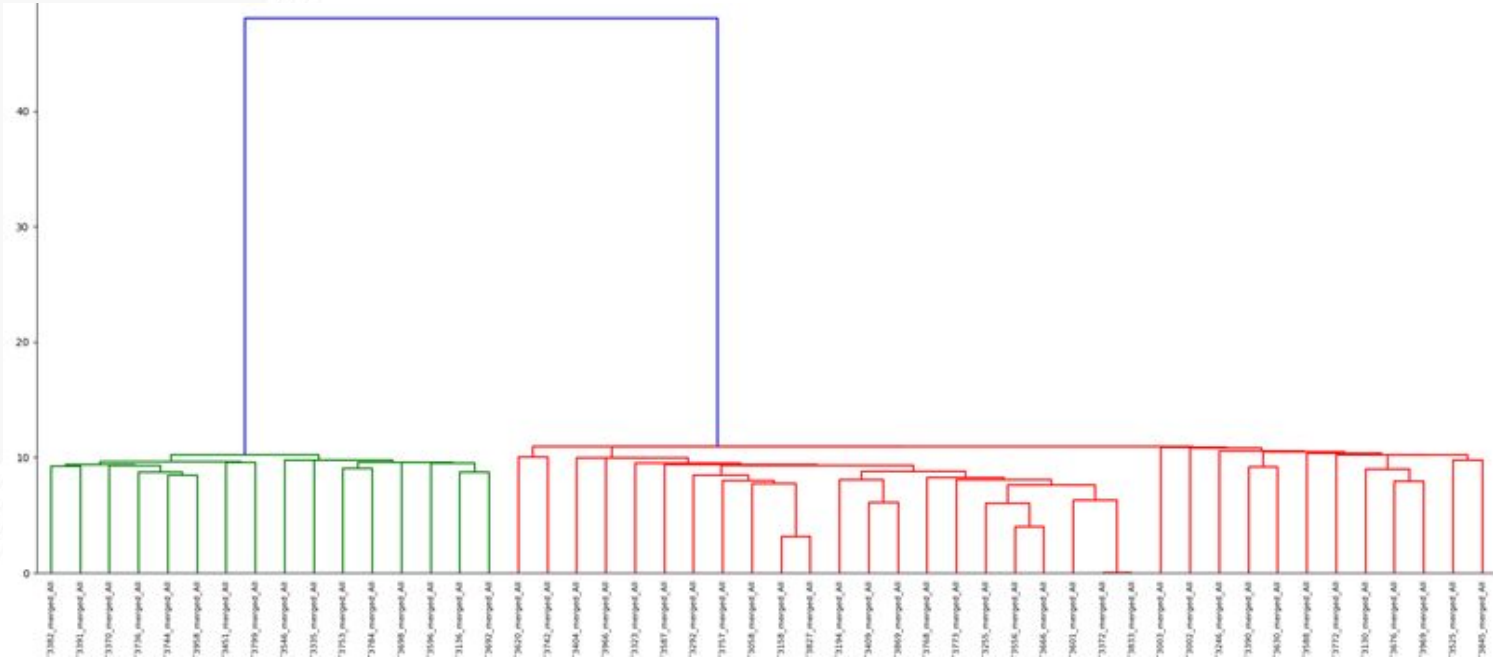
Presence/Absence Matrix

Online tools like Heatmapper can be used to generate plots to look at group of genes present/absent

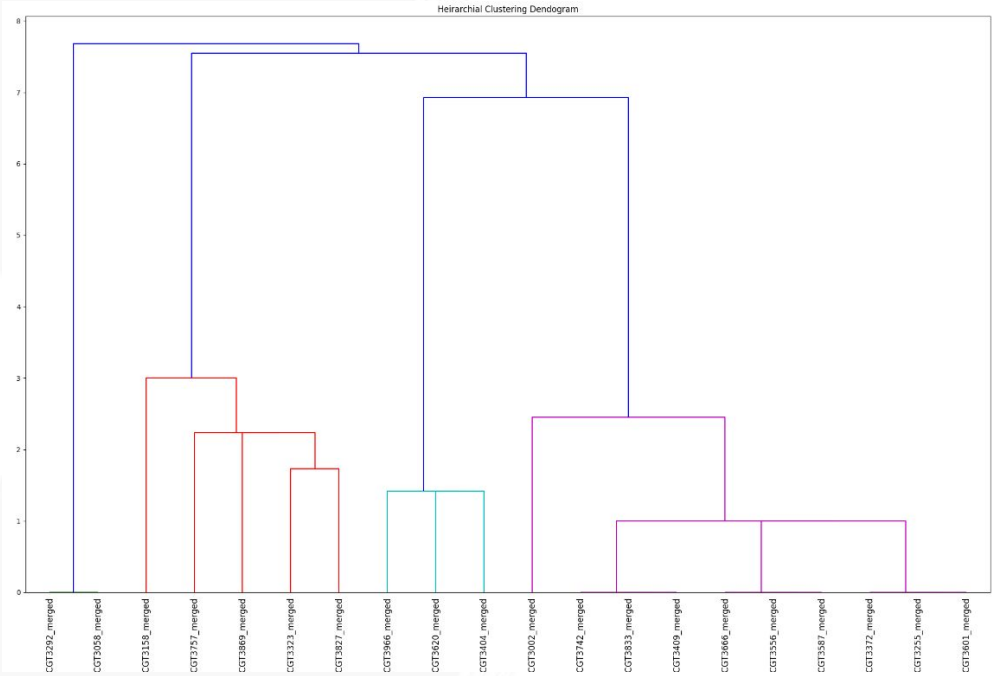
Hierarchical Clustering (Scipy)

Voor Hees Algorithm

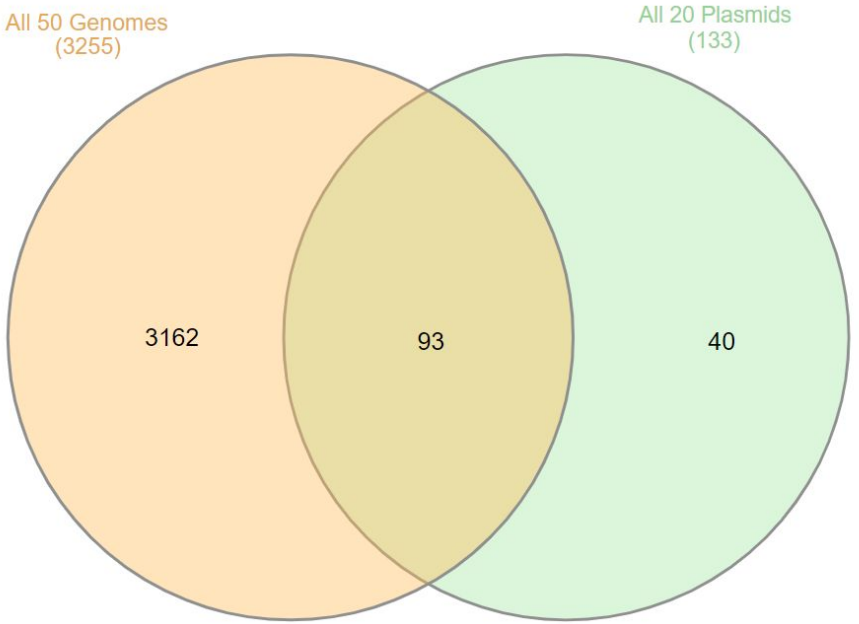
$$d(u, v) = \max(dist(u[i], v[j]))$$



GFF analysis of Plasmids



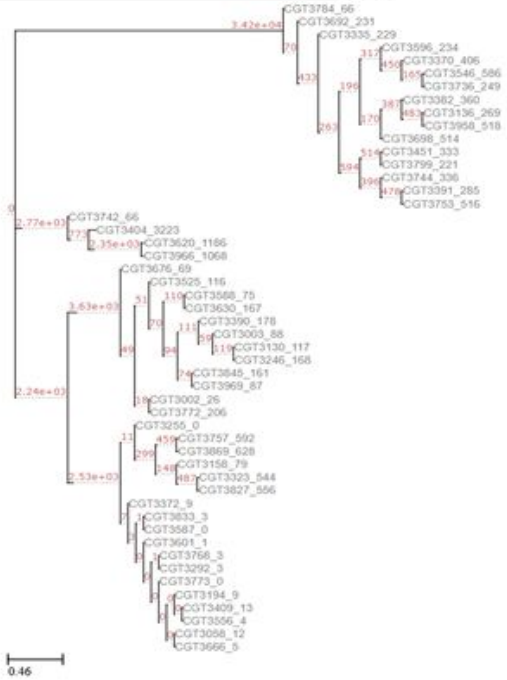
Hierarchical clustering of merged GFF files annoted on assembly files generated using plasmidSPades



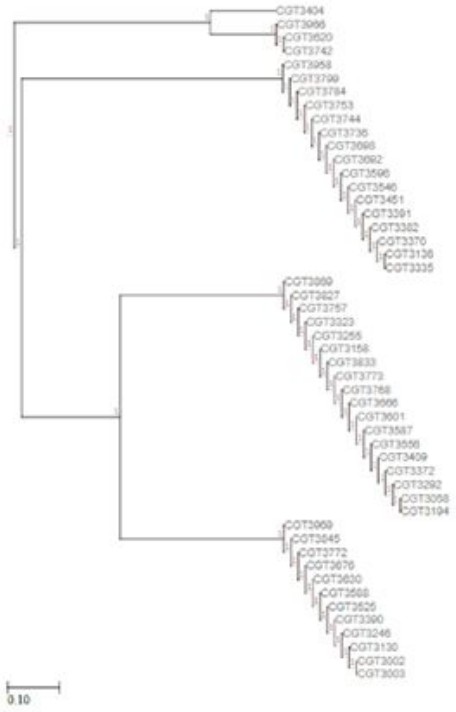
Unique annotations uncovered in plasmid data

Georgia Tech
CREATING THE NEXT

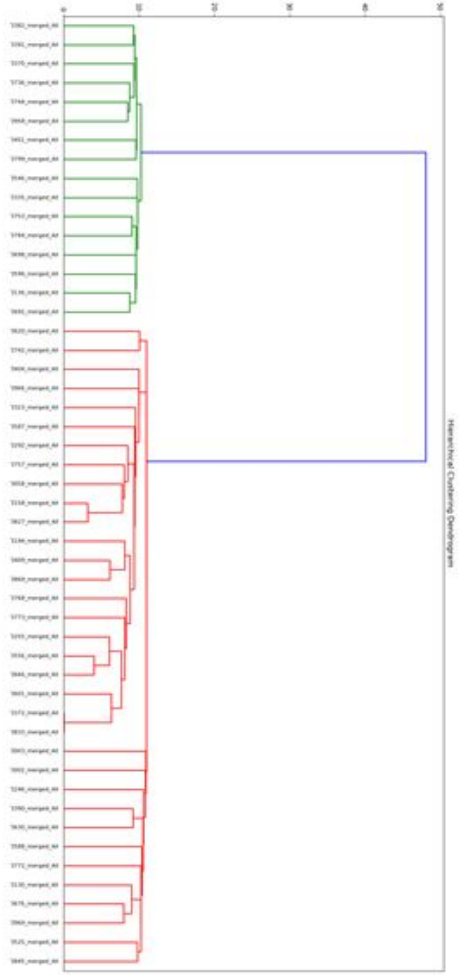
Correlation of clusters with different typing analysis



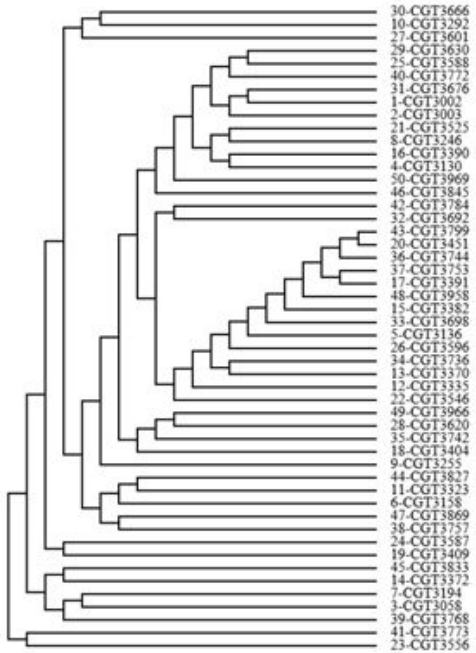
Tip Allele SNP ML Tree



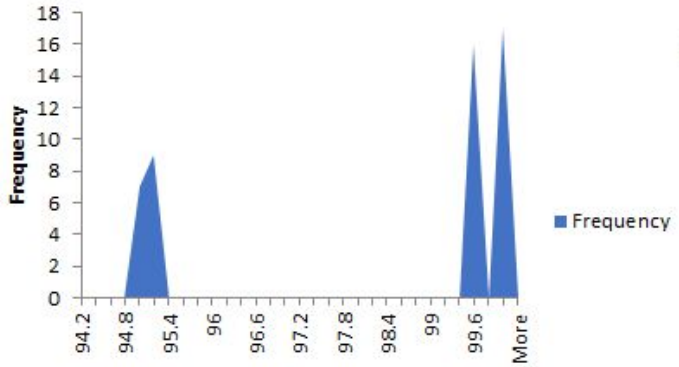
7gMLST



Hierarchical Clustering from merged annotations



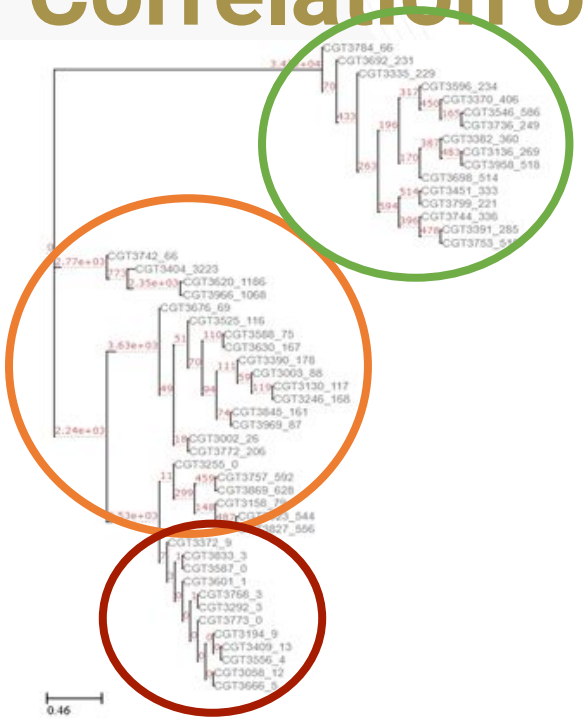
BPGA Pan Genome Analysis



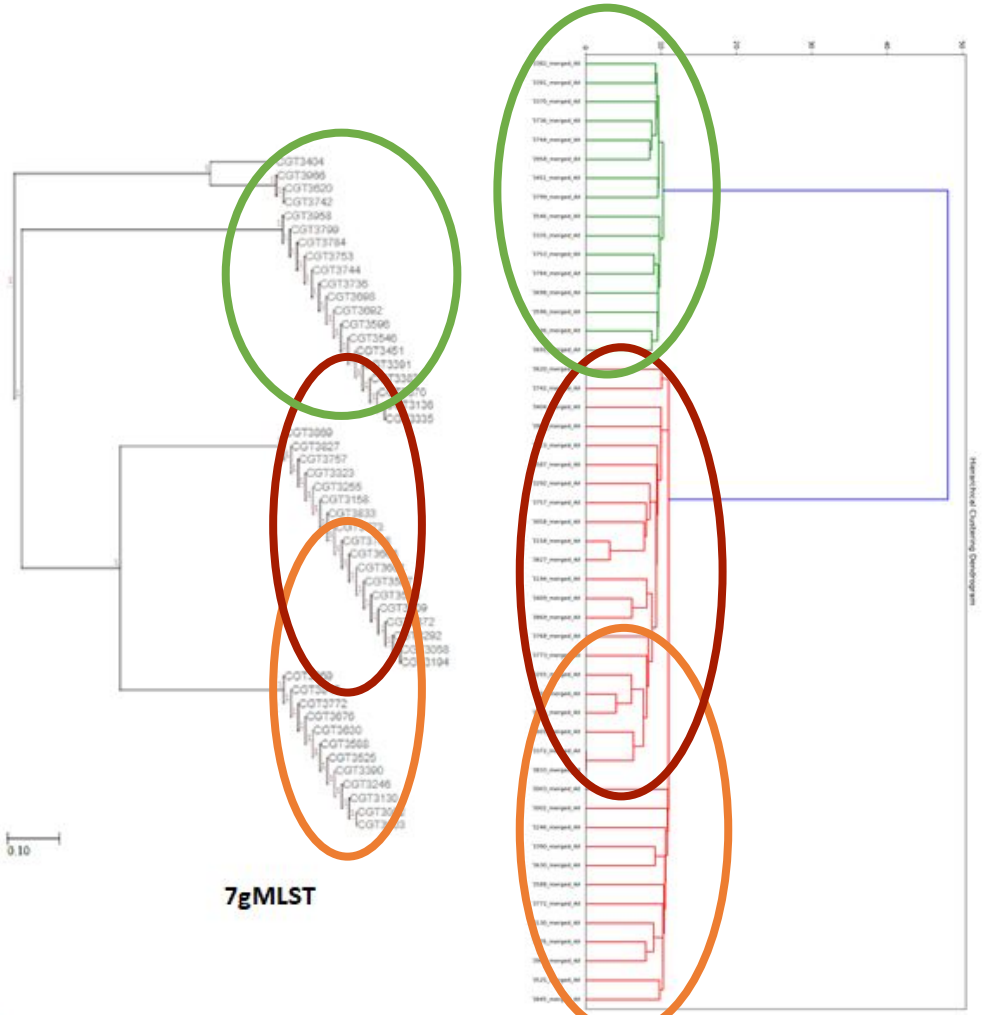
ANI scores against *L.monocytogenees* Seratype 4b Genome

Georgia Tech
CREATING THE NEXT

Correlation of clusters with different typing analysis

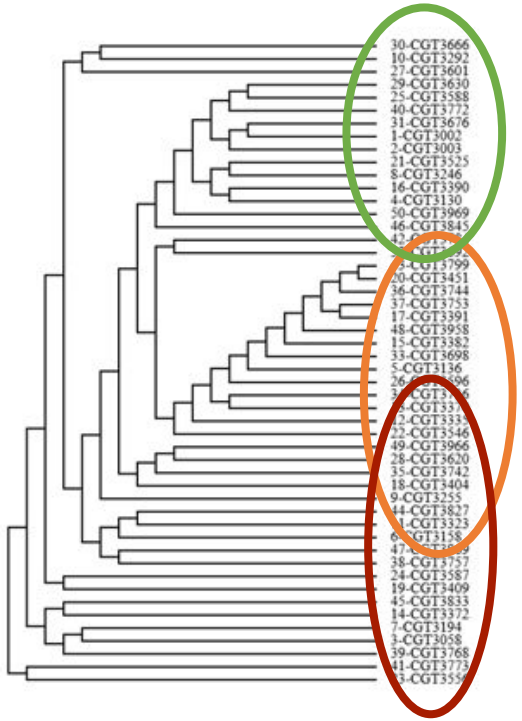


Tip Allele SNP ML Tree

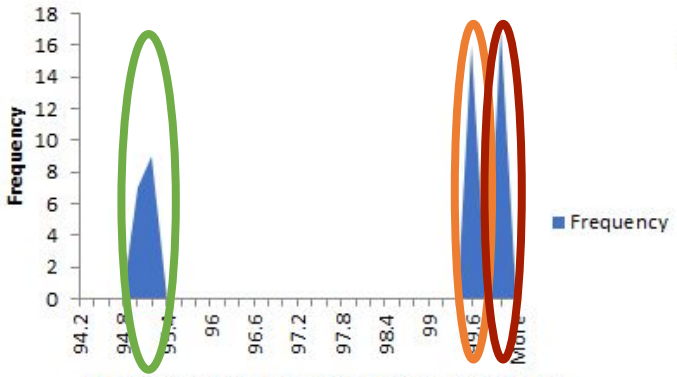


7gMLST

Hierarchical Clustering from merged annotations



BPGA Pan Genome Analysis

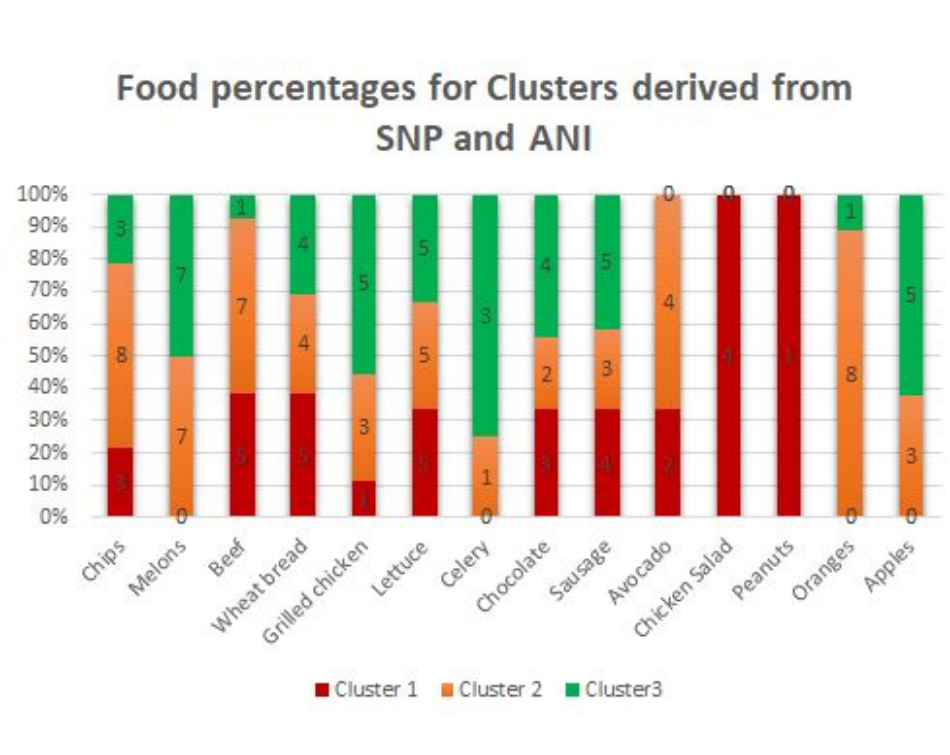


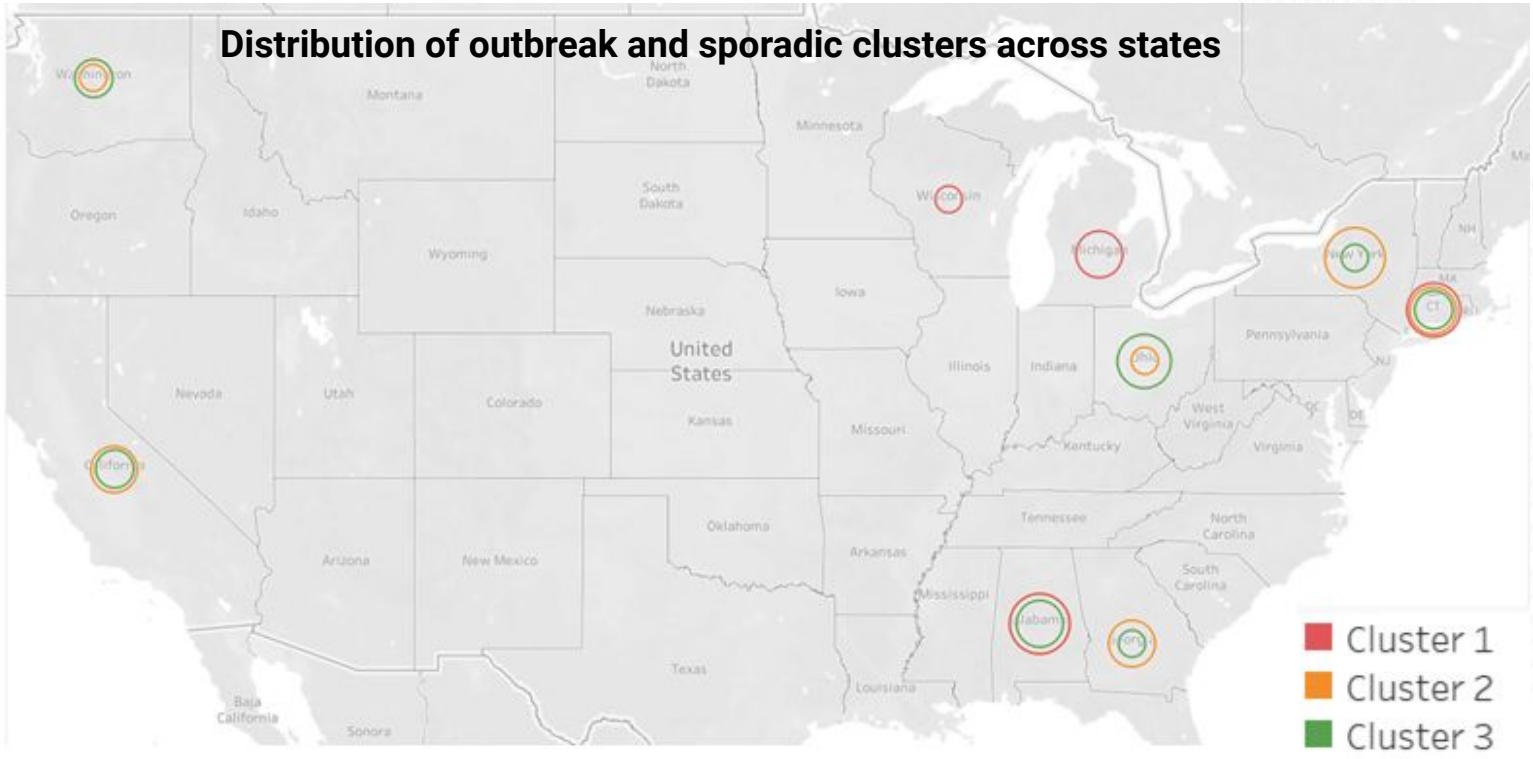
ANI scores against *L.monocytogenees* Seratype 4b Genome

Cluster 1 - Outbreak Cluster 2 - Similar to outbreak Cluster 3 - Different than outbreak

Georgia Tech
CREATING THE NEXT

Food source and Outbreak locations





Chicken Salad fits the requirement for being the outbreak source for Listeria

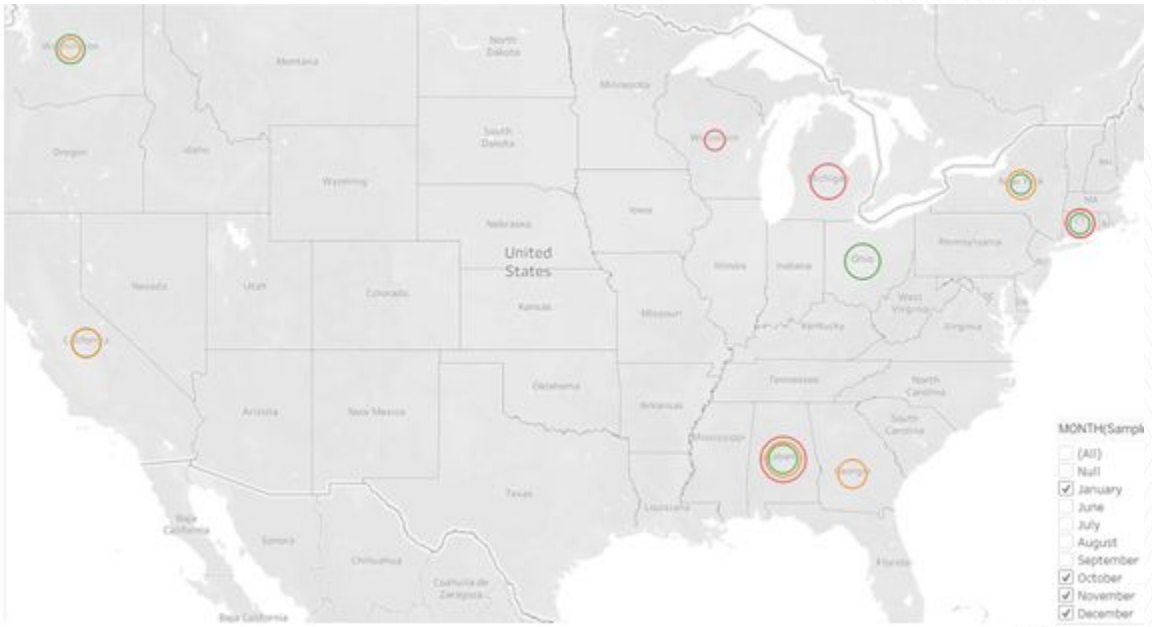
Interesting observation: You see Outbreak cluster(Red) and Cluster(Orange) similar to the outbreak cluster only existing in **Connecticut**

Cluster 1 - Outbreak Cluster 2 - Similar to outbreak Cluster 3 - Different than outbreak

Georgia Tech
CREATING THE NEXT

Timeline and location of clusters

Distribution of outbreak and sporadic clusters at the beginning of the outbreak



Distribution of outbreak and sporadic clusters at the peak of the outbreak

The outbreak source is from Connecticut!

Cluster 1 - Outbreak Cluster 2 - Similar to outbreak Cluster 3 - Different than outbreak

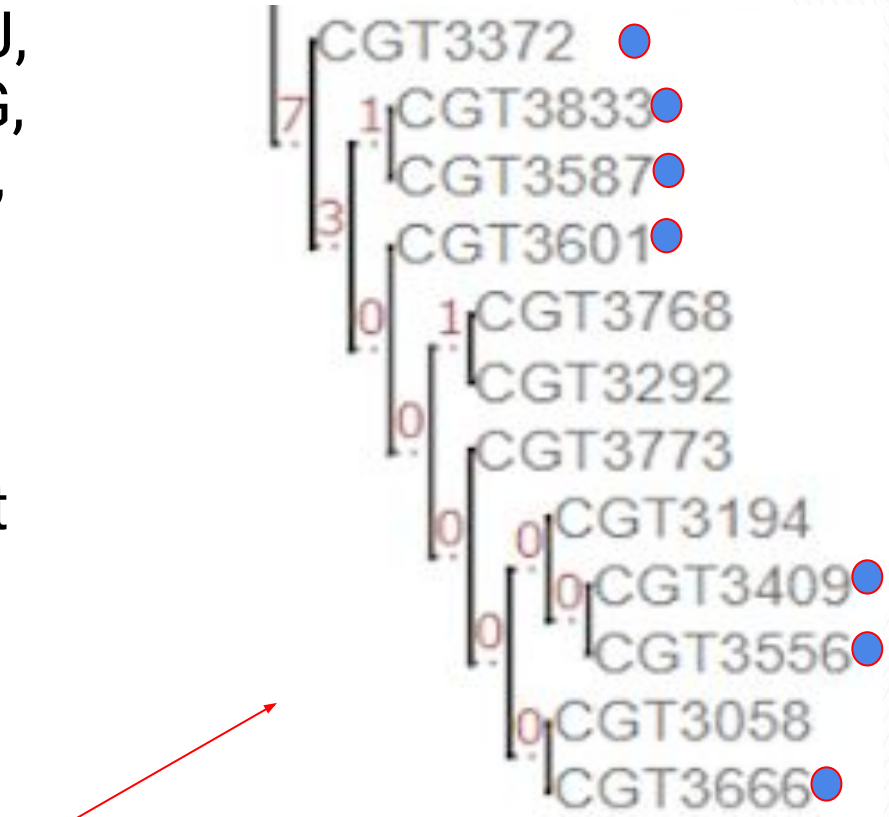
Georgia Tech
CREATING THE NEXT

Outbreak Analysis - VFDB

- 36 common virulence factor genes - lapB, inlJ, oatA, hpt, prsA2, lspA, prfA, llsY, llsB, llsH, llsG, llsD, llsX, lpeA, plcA, plcB , actA, pdgA, vip, hly, inlF, inlA, inlB, inlC, clpE, inlP, mpl, clpP, inlK, iap/cwhA, fbpA, clpC, lntA, ami, lap, bsh

- 3 genes absent in outbreak group but present in other isolates- llsP, gtcA, aut

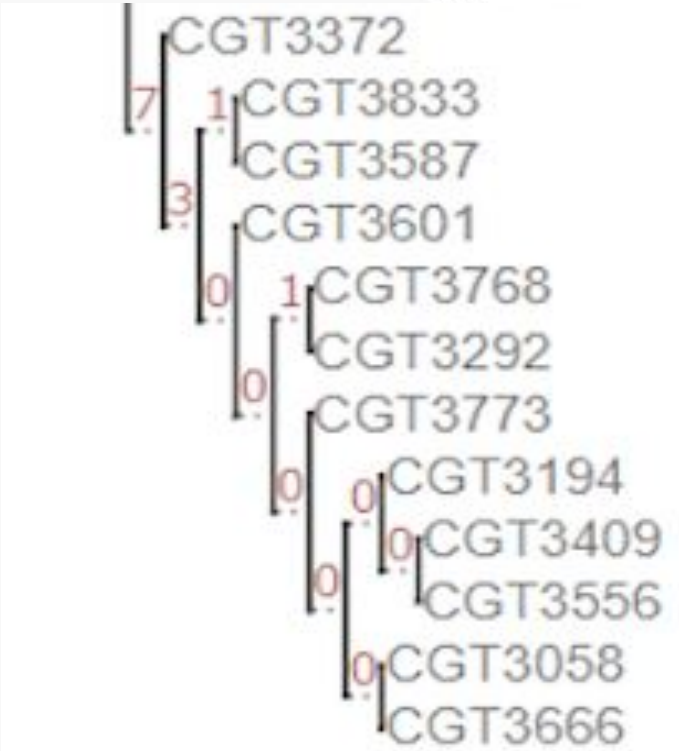
- plasmid analysis of VFDB gave lplA1 gene associated with plasmid.

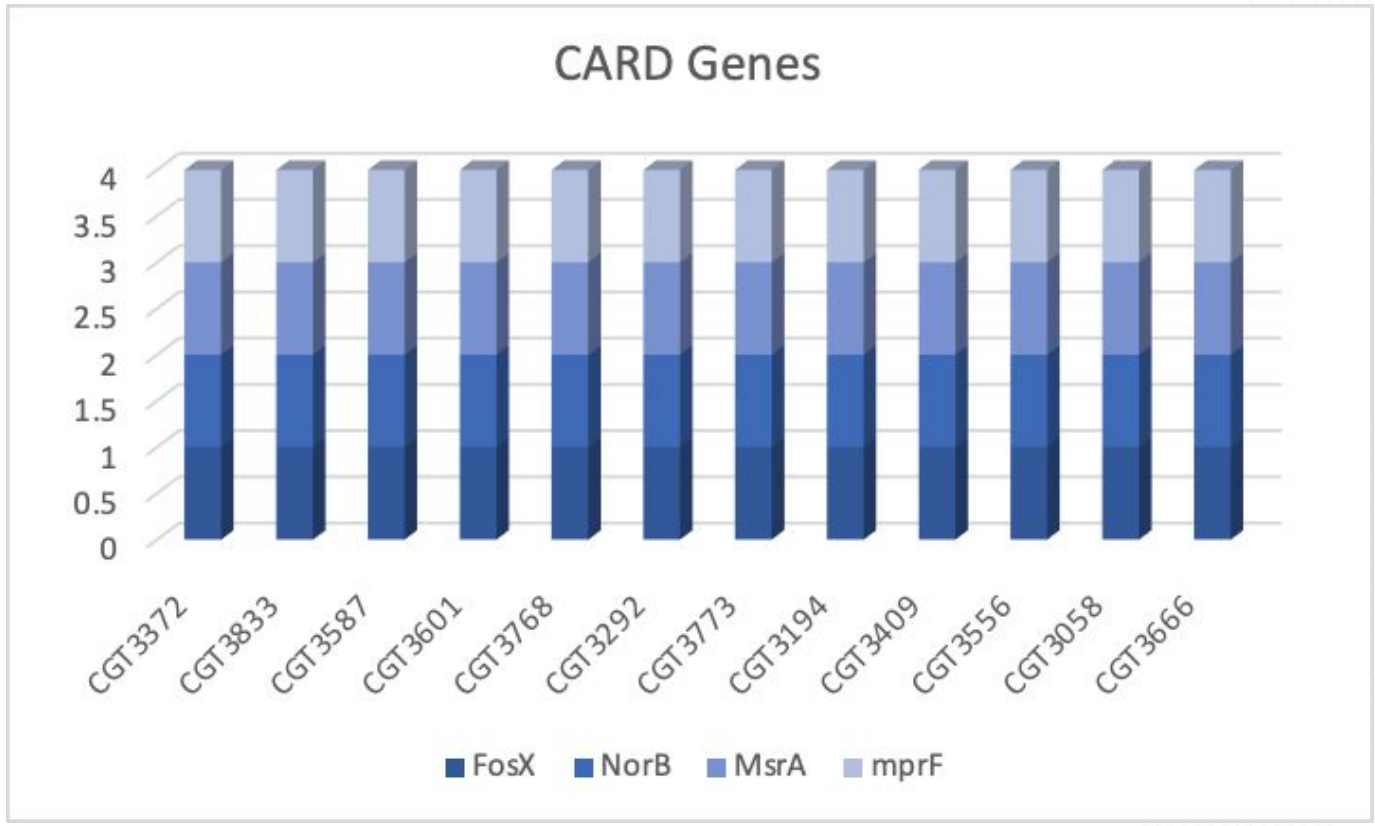


Georgia Tech
CREATING THE NEXT

Outbreak Analysis - CARD gff

Isolates with OUTBREAK strains --> Antibiotic resistance genes based on GFF from functional annotation team





Antibiotic resistance

Database	Gene	Present on	Drug resistance	Resistance mechanism	AMR gene family	Drug class
CARD	FosX	Chromosome	Fosfomycin	antibiotic inactivation	fosfomycin thiol transferase	fosfomycin
CARD	msrA	plasmid or chromosome	Erythromycin and streptogramin B	antibiotic target protection	ABC-F ATP-binding cassette ribosomal protection protein	streptogramin, tetracycline, pleuromutilin, macrolide, oxazolidinone, lincosamide, phenicol antibiotic
CARD	norB	chromosome	fluoroquinolones and other structurally unrelated antibiotics like tetracycline.	antibiotic efflux	major facilitator superfamily (MFS) antibiotic efflux pump	fluoroquinolone antibiotic
CARD	Listeria monocytogenes mprF	chromosome	defensin resistance	antibiotic target alteration	defensin resistant mprF	peptide antibiotic

Recommendation for Antibiotic

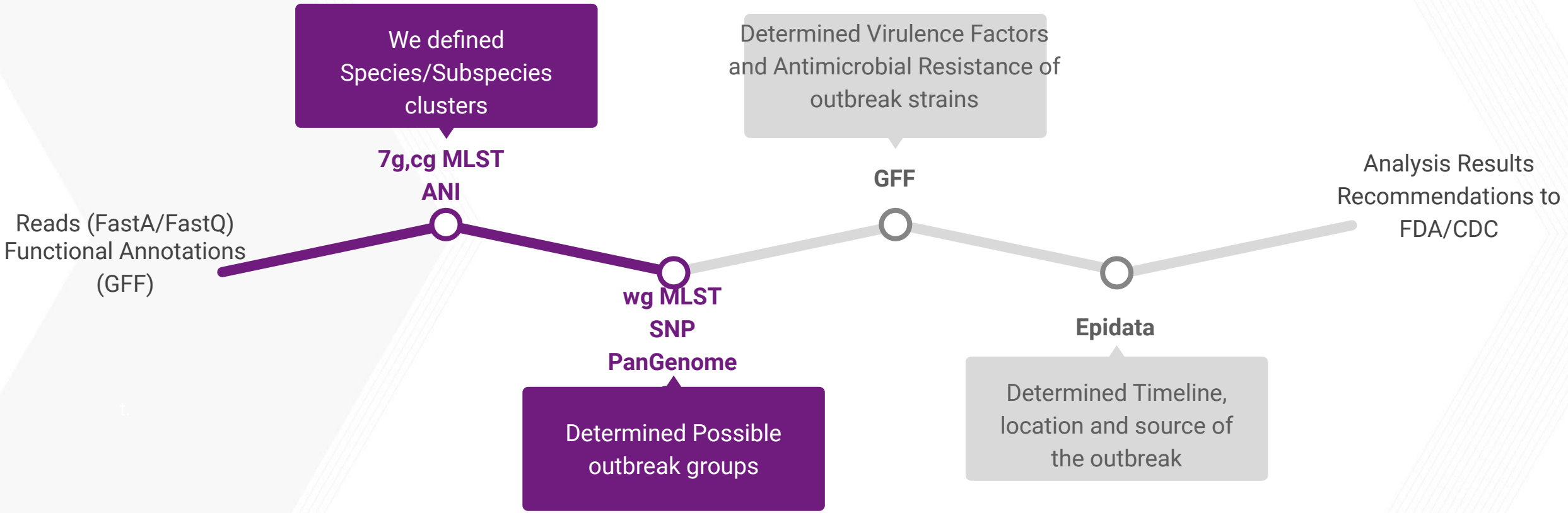
Listeriosis treatment using	Antibiotic	Recommendation
β-lactam antibiotic	**ampicillin**	YES
aminoglycoside	gentamicin [+ampicillin]	YES
β-lactam antibiotic	penicillin	YES
β-lactam antibiotic	amoxicillin [not used mostly]	NO
allergy to penicillin	trimethoprim - sulfamethoxazole	YES
allergy to penicillin	vancomycin, meropenem, or a macrolide [not widely used]	YES
alternative treatment	tetracycline	NO
alternative treatment	erythromycin	NO
alternative treatment	Fosfomycin	NO
alternative treatment	Fluoroquinolone	NO

*Cephalosporins, Chloramphenicol are not effective against *Listeria monocytogenes*.

Output to user

- ANI analysis: ANI score
- MLST analysis: StringMLST - 7-gene mlst allele matrix, Chewbacca - cgMLST allele matrix, Genome Quality plot, results_statistics.tsv, results_contigsinfo.tsv
- SNP analysis: Phylogenetic tree
- Gff file clustering : Presence absence matrix and dendrograms and tab delimited Gff files.

Comparative Genomics Pipeline



Georgia Tech
CREATING THE NEXT

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
- 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
- "Identification of acquired antimicrobial resistance genes", Zankari et al 2012, PMID: 22782487

Thankyou!