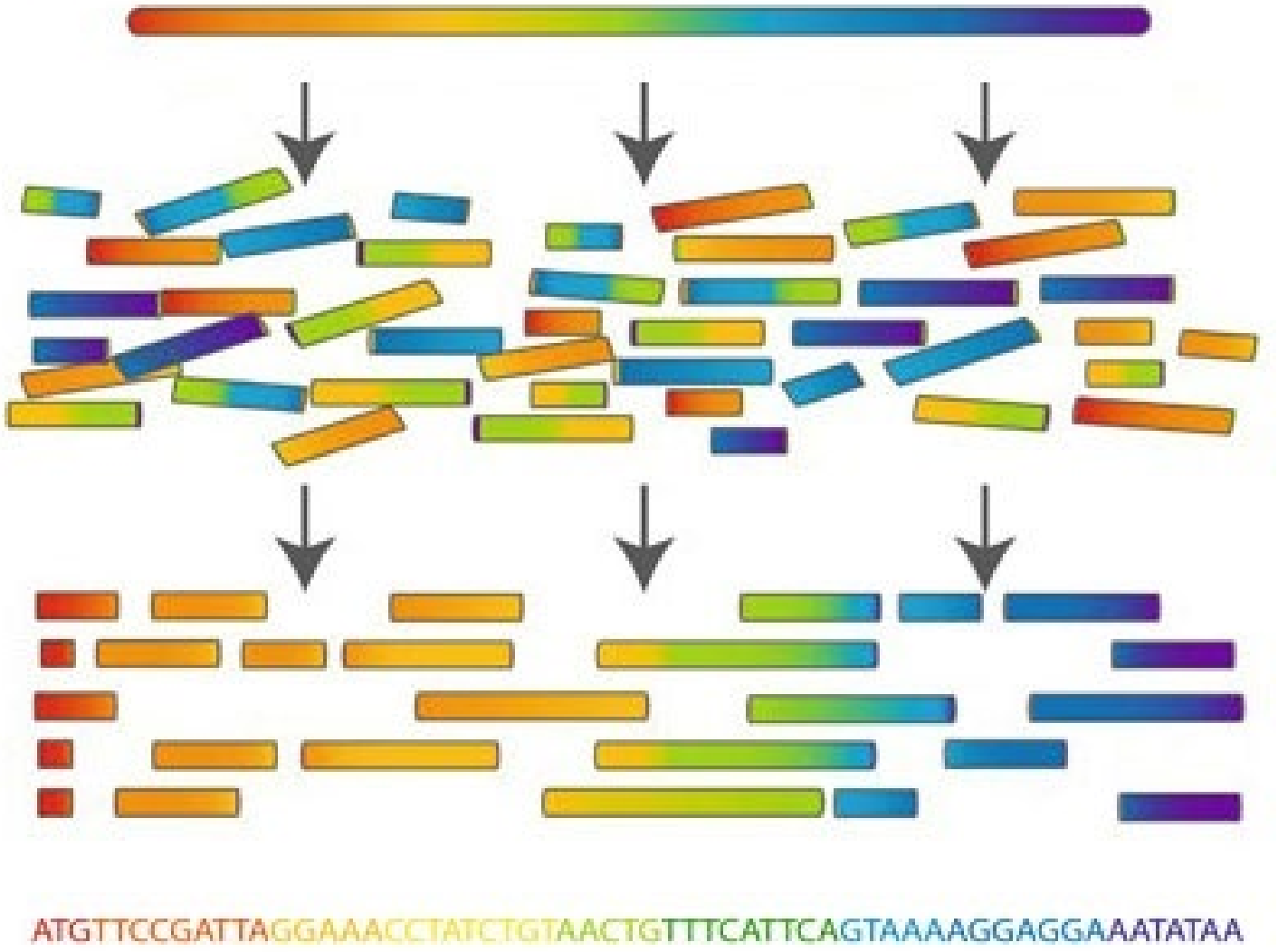
Genome Assembly and Comparative Genomics of Human microbiome from CDC sequencing data

Genome Assembly Results

Maddala Aparna
Yang Ruize
Kundnani Deepali (Slide 32-36)
Xiao Yiqiong
Singu Swetha Gowri

January 28, 2020

What is genome assembly?



original sequence

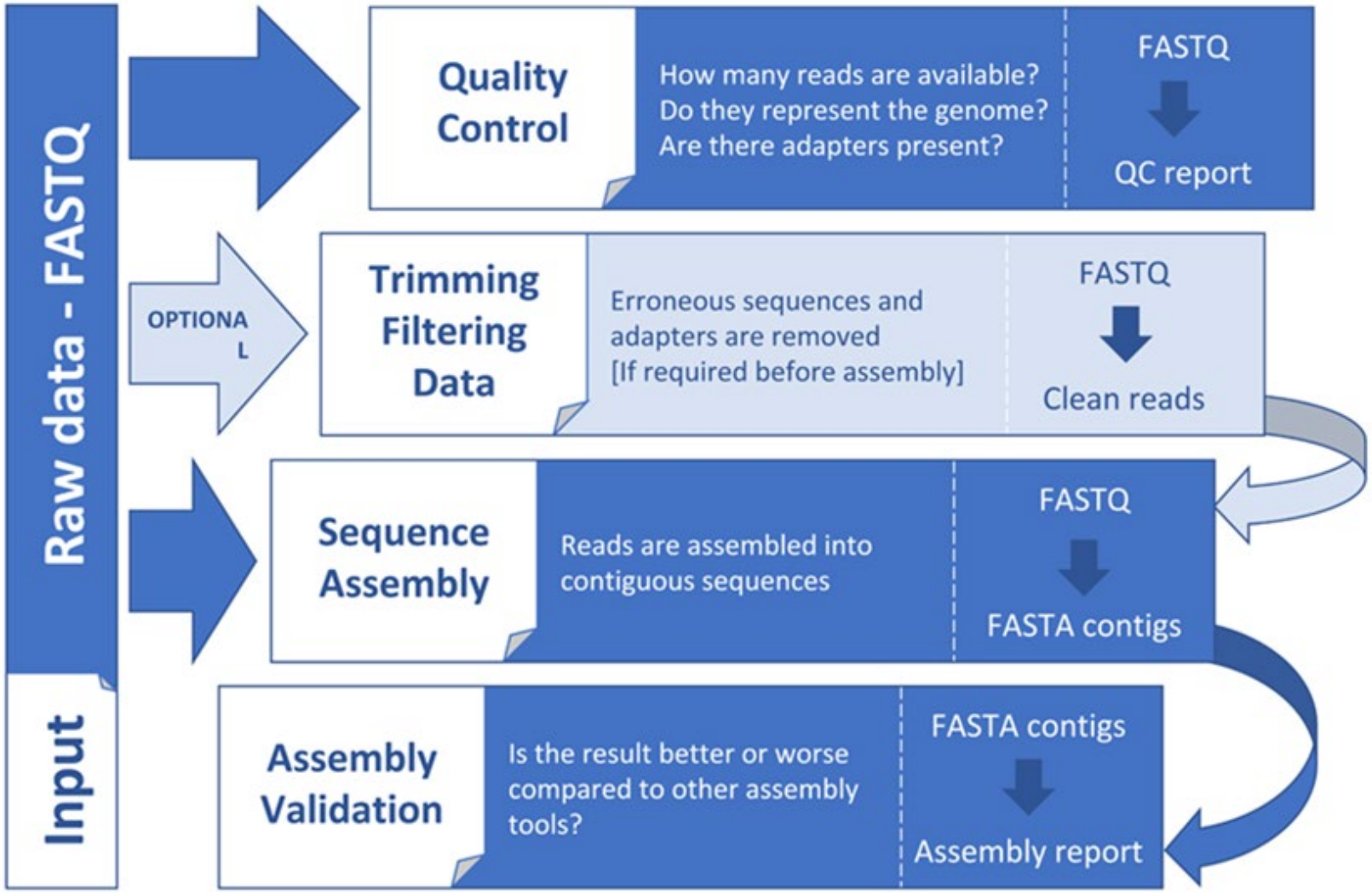
reads (output of sequencing)

alignment

contig

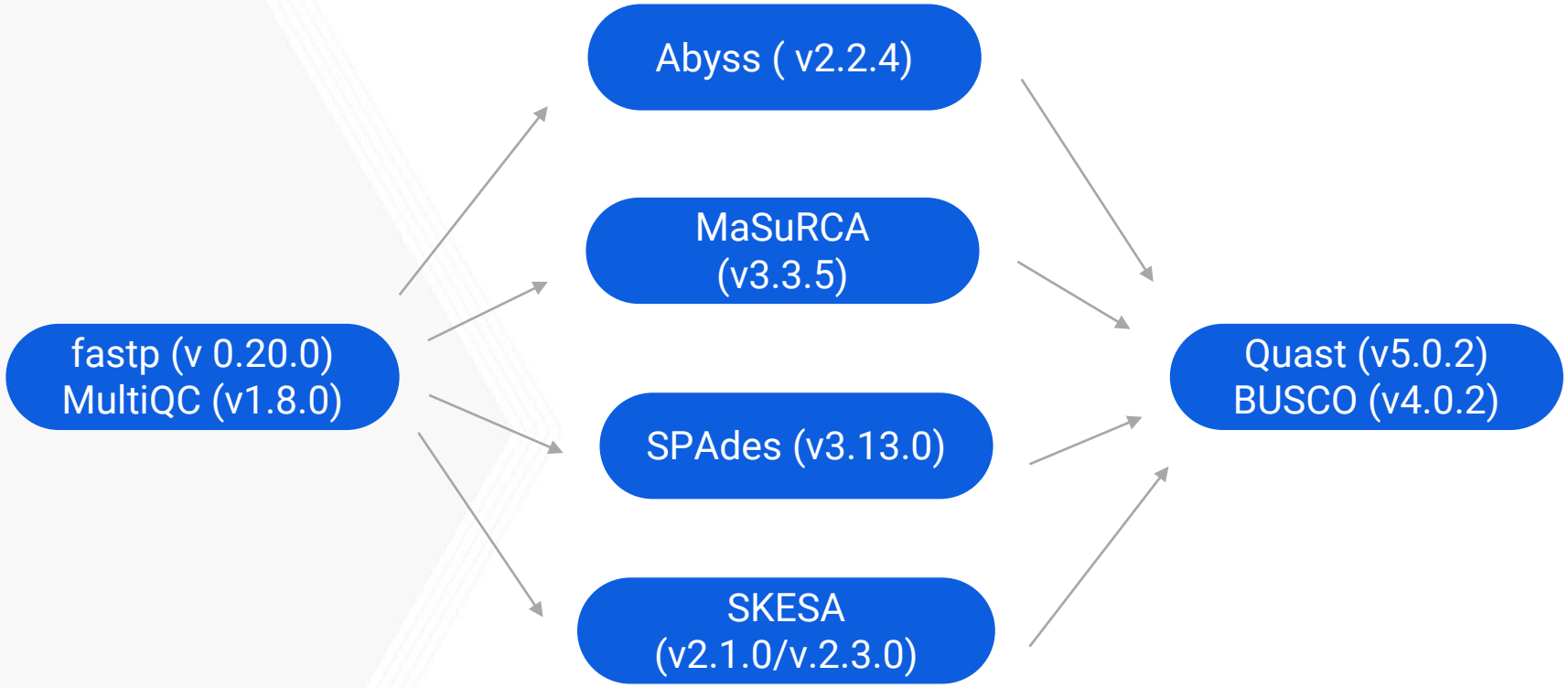
Adapted from: Commins, Jennifer et al. “Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects.” Biological Procedures vol. 11 52-78. 11 Mar. 2009.

Steps of Genome Assembly



Dominguez Del Angel V, Hjerde E, Sterck L et al. Ten steps to get started in Genome Assembly and Annotation [version 1]. F1000Research 2018, 7:148 (doi: 10.12688/f1000research.13598.1)

Genome Assembly Tools Bench Marking Plan



fastq
Quality control and trimming for reads

Feb 2nd

Feb 9th

Assembly
Test out 4 tools and complete

Quast and BUSCO
Quality control for assembled reads

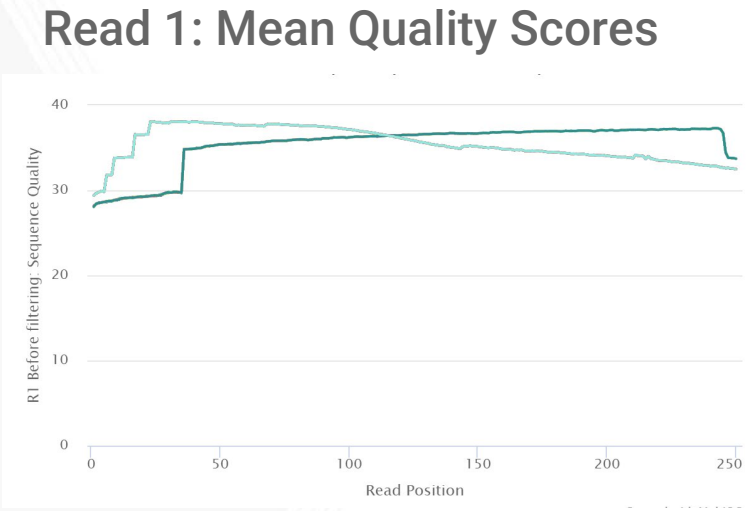
Feb 12th

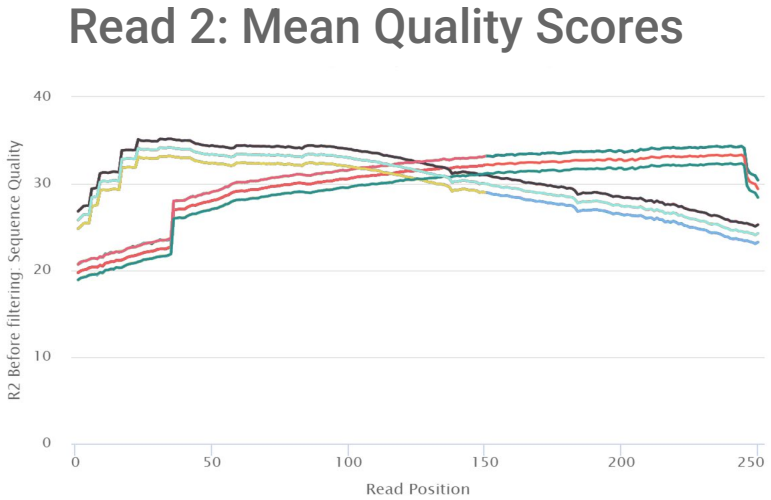
Feb 17th

2nd presentation preparation
Finish slides and update the

fastp: Pre and post trimming Read Quality

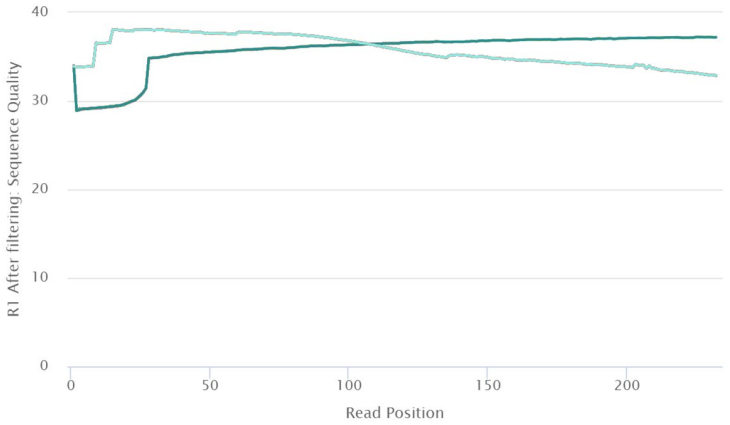
Pre-Trimming

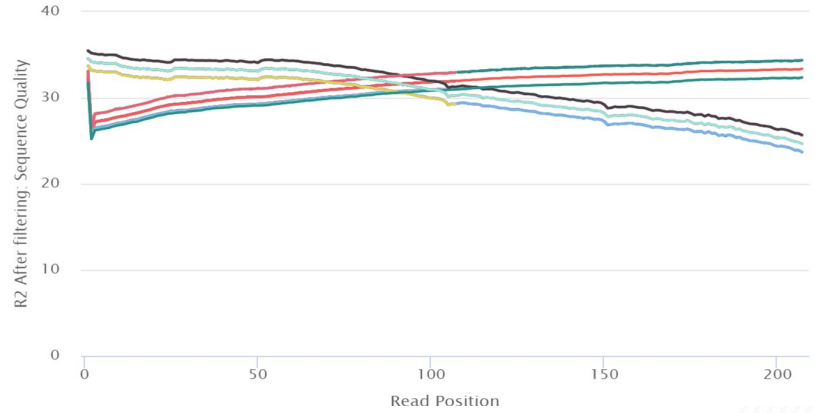




Trimming parameters: `-f 5 -F 30 -t 10 -e 28 -c -5 3 -M 27`

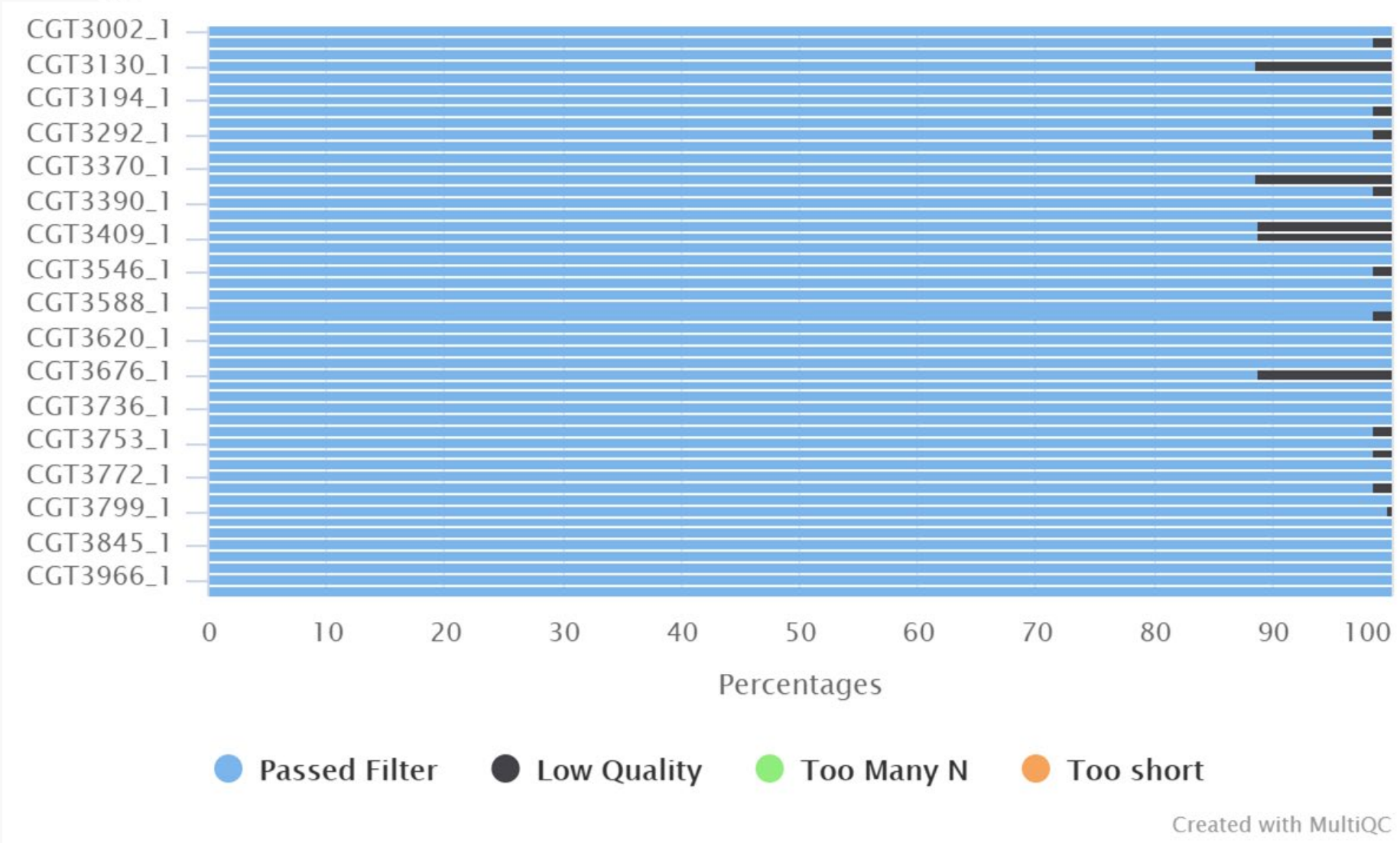
Post Trimming





Georgia Tech
CREATING THE NEXT

Percent of Reads Trimmed



Georgia Tech
CREATING THE NEXT

Assemblers Benchmarked

MaSuRCA

SKESA

SPAdes

Abyss

MaSuRCA

Version:

MaSuRCA v3.3.5

Properties:

- runs on untrimmed reads
- adapted for a mixture of long and short reads, and tolerates high sequencing error
- Pipeline
  - Jellyfish kmer counter automatically selects optimal k-mer size for each sample
  - CABOG Assembler uses “super-reads”

MaSuRCA

Parameters:

- GRAPH_KMER_SIZE = auto
- USE_LINKING_MATES = 1
- CA_PARAMETERS = cgwErrorRate=0.25 (bacteria)

Optimization:

- for each sample tested, MaSuRCA selected a k-mer size of 99

```
PARAMETERS
USE_LINKING_MATES = 1
NUM_THREADS = 16
JF_SIZE = 200000000
USE_GRID=0
GRID_ENGINE=SGE
GRID_QUEUE=all.q
GRID_BATCH_SIZE=300000000
LHE_COVERAGE=25
MEGA_READS_ONE_PASS=0
CA_PARAMETERS =  cgwErrorRate=0.25
KMER_COUNT_THRESHOLD = 1
CLOSE_GAPS=1
END
```

Assemblers Benchmarked

MaSuRCA

SKESA

SPAdes

Abyss

SPAdes

Version:

SPAdes v3.13.0

Command:

```
subprocess.call("spades.py --careful -1 " + f1 +" -2 "+f2+ " -o "+output, shell = True)
```

Parameters:

--careful: Tries to reduce the number of mismatches and short indels, recommended only for assembly of small genomes.

-k: kmer size, input a series of numbers, auto-detection

Unicycler & SPAdes

- An assembly pipeline for bacterial genomes
- Can be used as a SPAdes optimiser
- Came out in 2017, 400+ citations

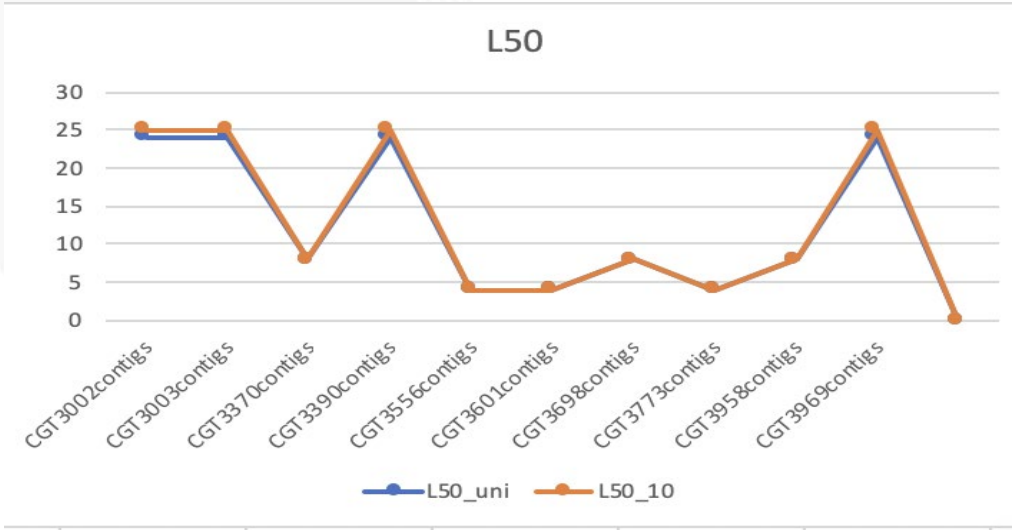
Version:

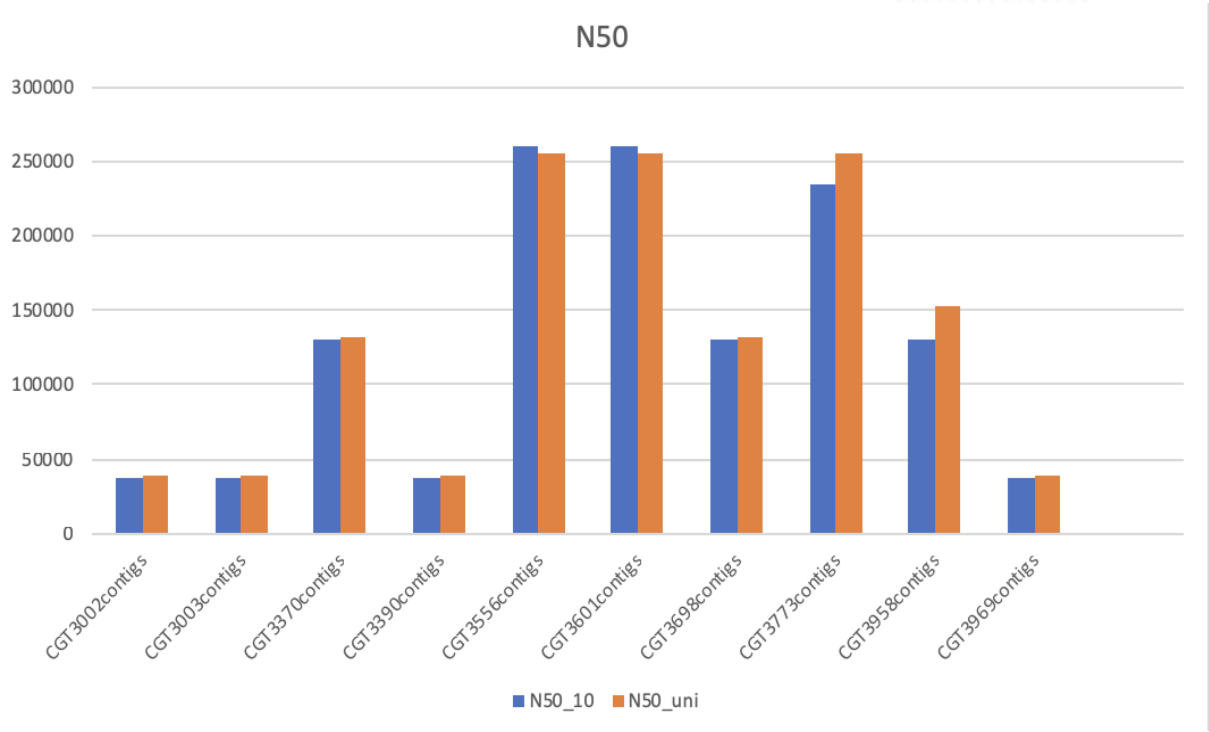
v0.4.7

Command:

```
subprocess.call("unicycler --spades_path spades.py -1 " + f1 +" -2 "+f2+ " -o "+output, shell = True)
```

Unicycler vs SPAdes

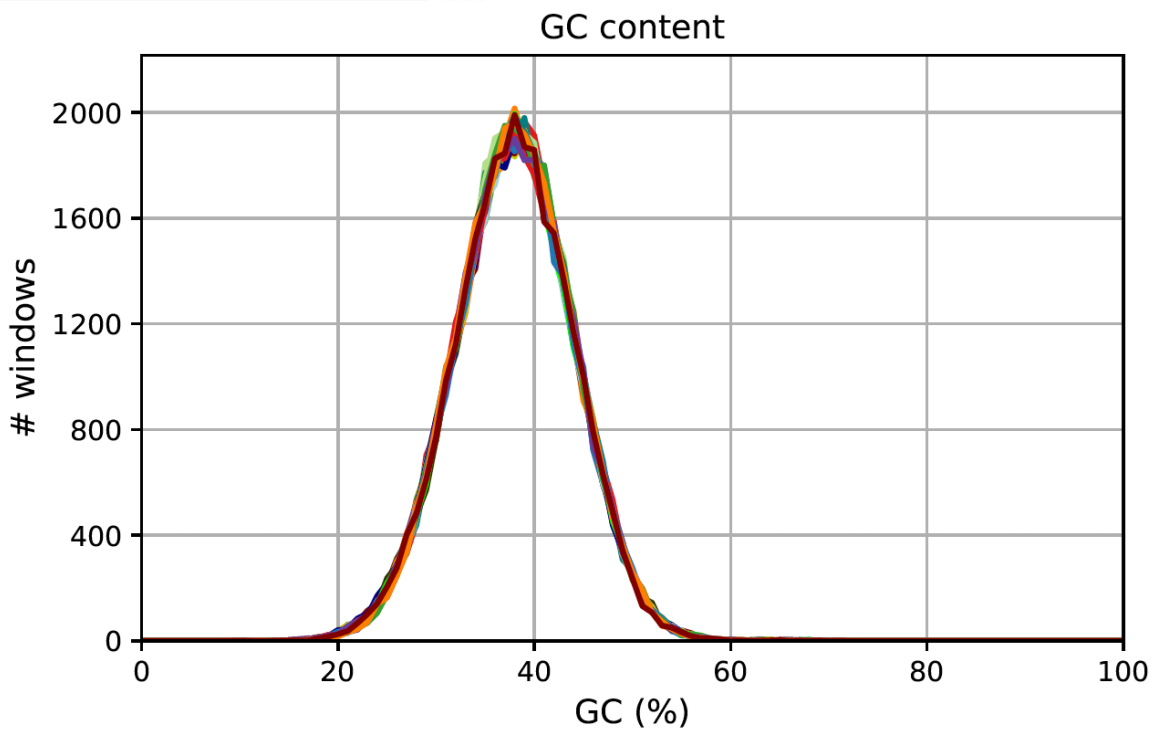


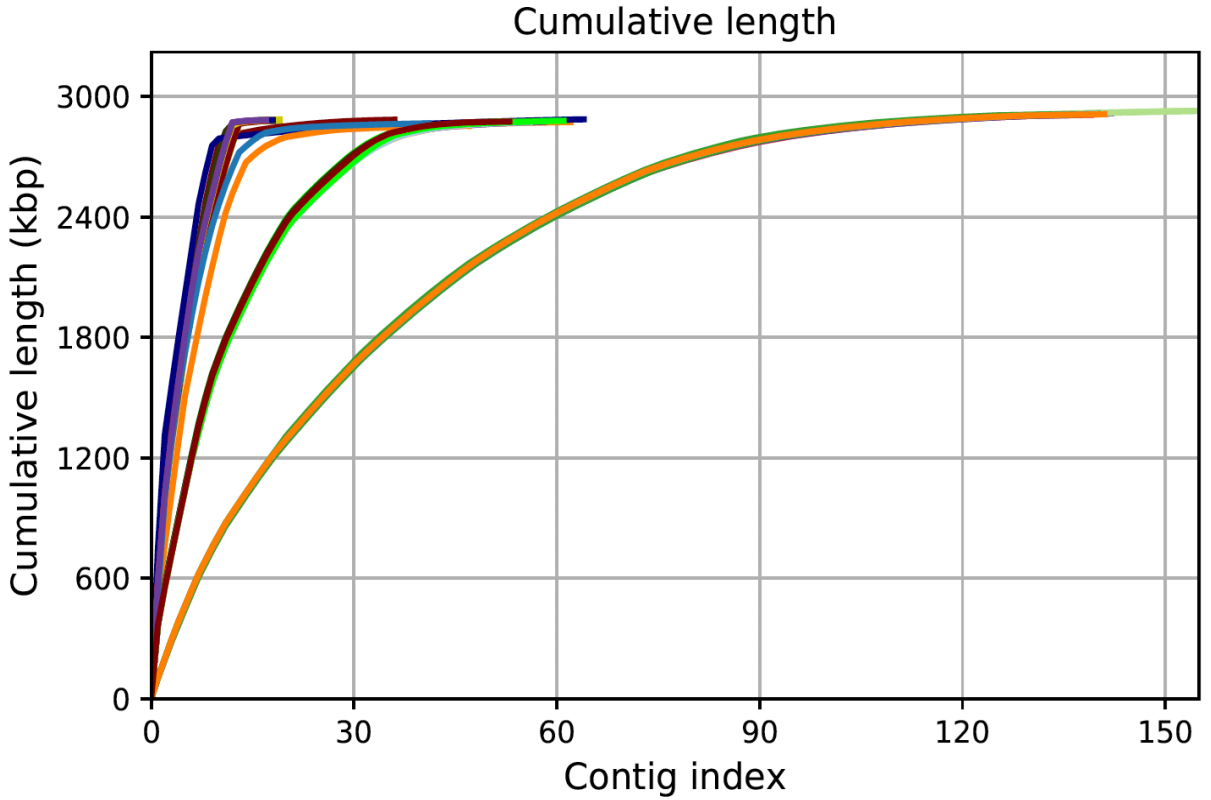



```
Thank you for using SPAdes!
        Command being timed: "spades.py --careful -1 CGT3002_r1.fq -2 CGT3002_r2.fq -o speed_spades"
        User time (seconds): 2232.40
        System time (seconds): 107.71
        Percent of CPU this job got: 450%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 8:39.37
```

```
Command being timed: "unicycler --spades_path spades.py -1 CGT3002_r1.fq -2 CGT3002_r2.fq -o speed_spades"
User time (seconds): 3365.85
System time (seconds): 165.50
Percent of CPU this job got: 327%
Elapsed (wall clock) time (h:mm:ss or m:ss): 17:57.48
```

SPAdes Output from Quast





Georgia Tech
CREATING THE NEXT

Assemblers Benchmarked

MaSuRCA

SKESA

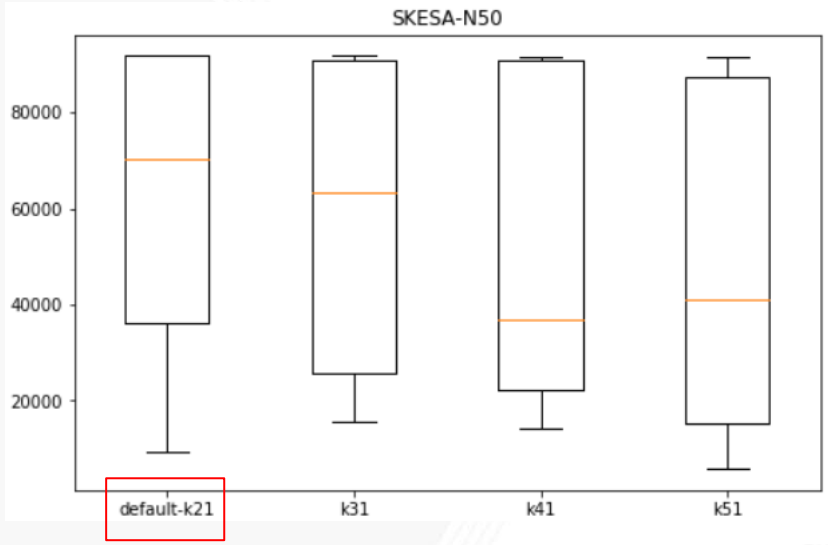
SPAdes

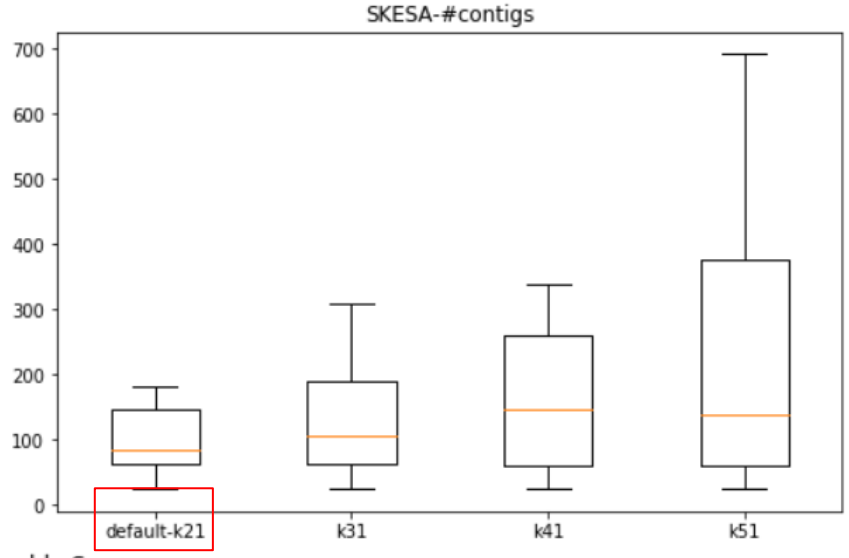
Abyss

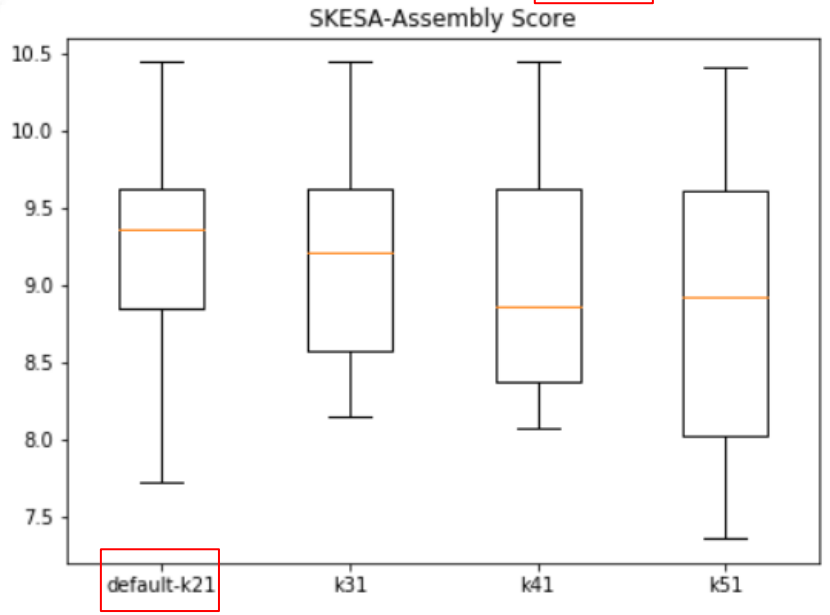
SKESA

- **Version**: 2.3.0
- **Command**: skesa --cores 4 --fasta/fastq --contigs_out
  - **--hash_count**: Use hash counter, much lower RAM (~1/3)
  - **--kmer**: Minimal kmer length for assembly, default=21
  - **--vector_percent**: Fractions of adapter, default=0.05
  - other unchanged options
- **sspace**: -l library -x 0/1 -s contigs
  - scaffolding pre-assembled contigs
  - -l: insert size: 150~450; orientation: FR
  - -x: extend input contigs using paired reads, default=0 (off)
- **Default** SKESA performs best

SKESA-kmer

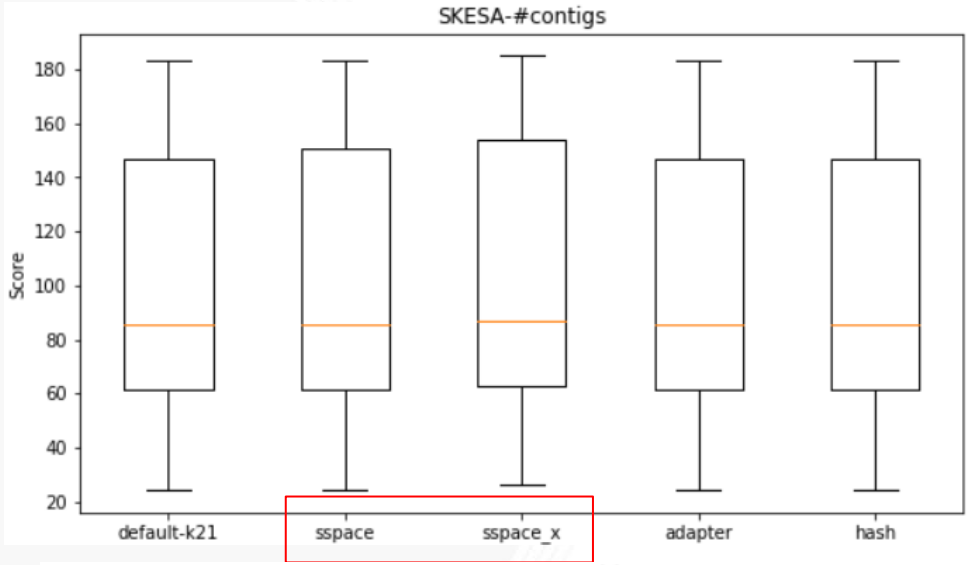


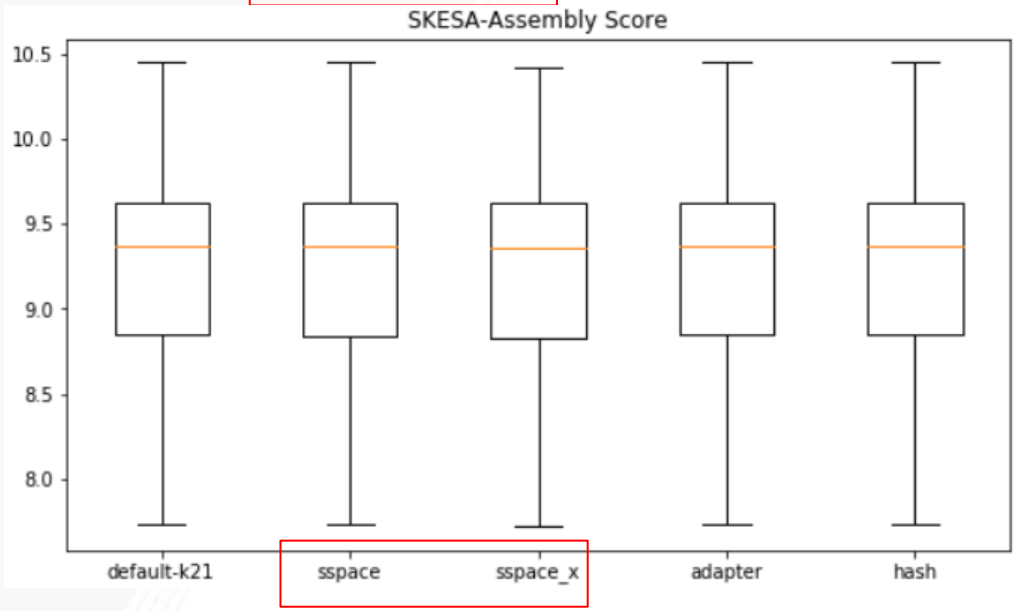


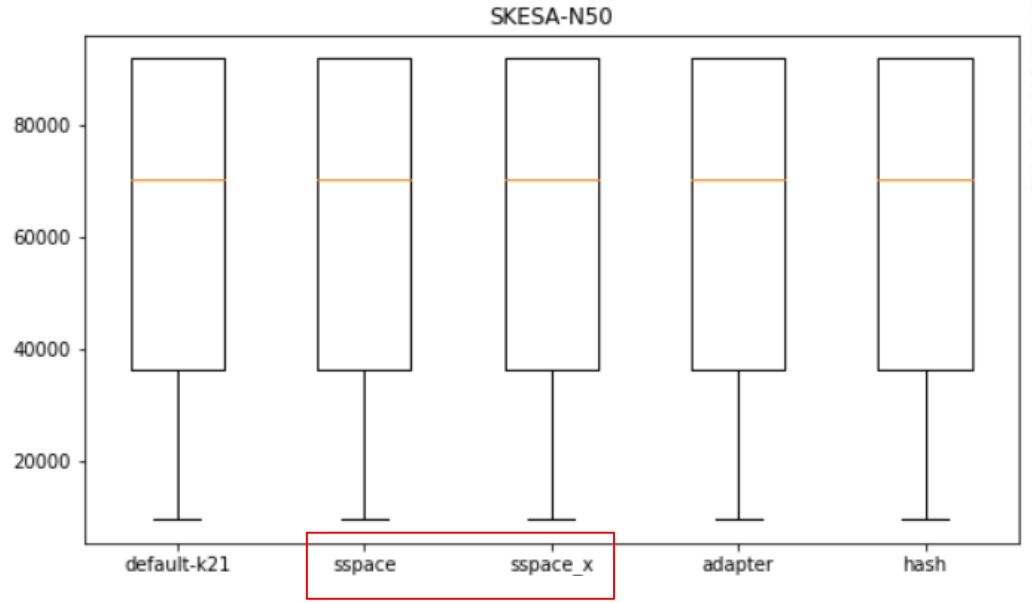


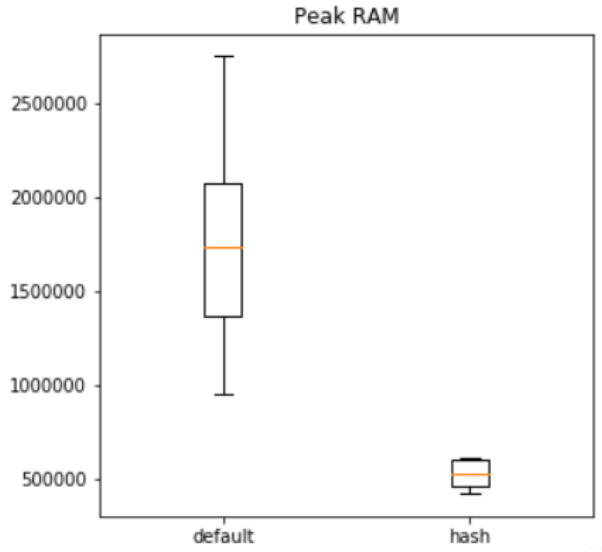
Georgia Tech
CREATING THE NEXT

SKESA-other options









Assemblers Benchmarked

MaSuRCA

SKESA

SPAdes

Abyss

Abyss

Why Abyss?

- de-novo assembler , parallel , designed for short reads
- De Bruijn graph algorithm

Which Version?

- 2.2.4 version
- Old version - more memory consumption
- From 2.0 version - Bloom filter

Stages:

- multistage assembly pipeline - unitigs, contigs and scaffold stages

Abyss

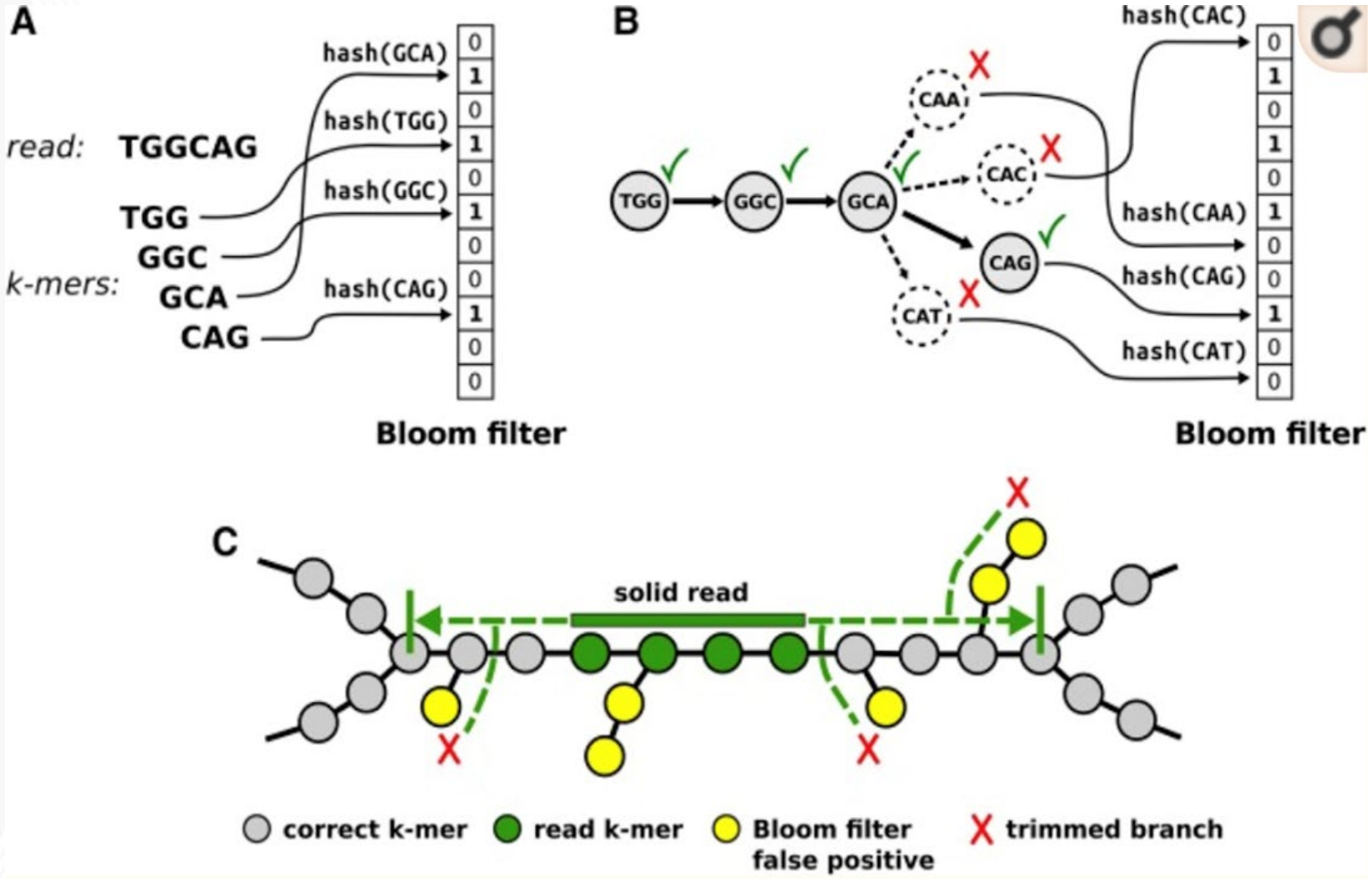
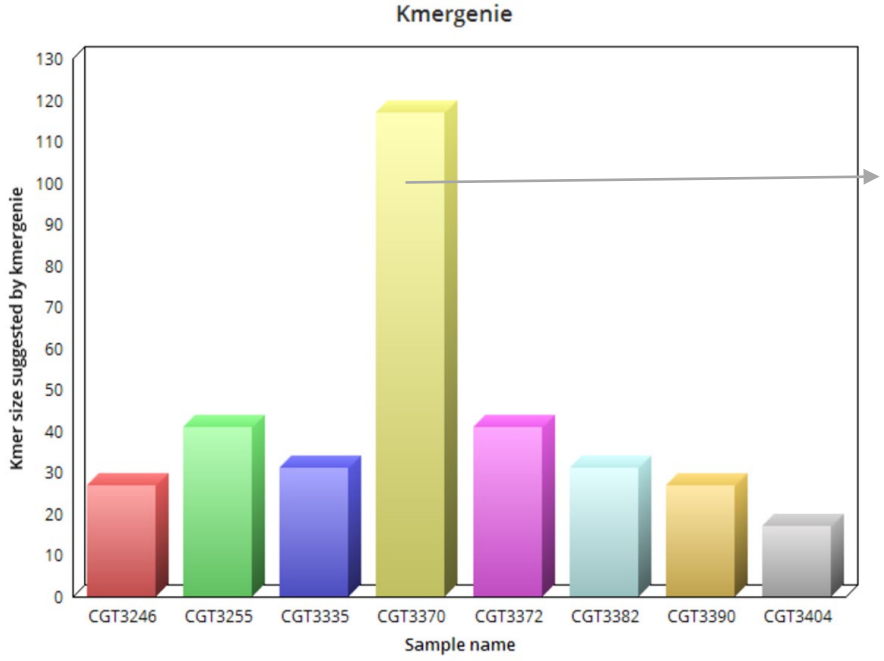
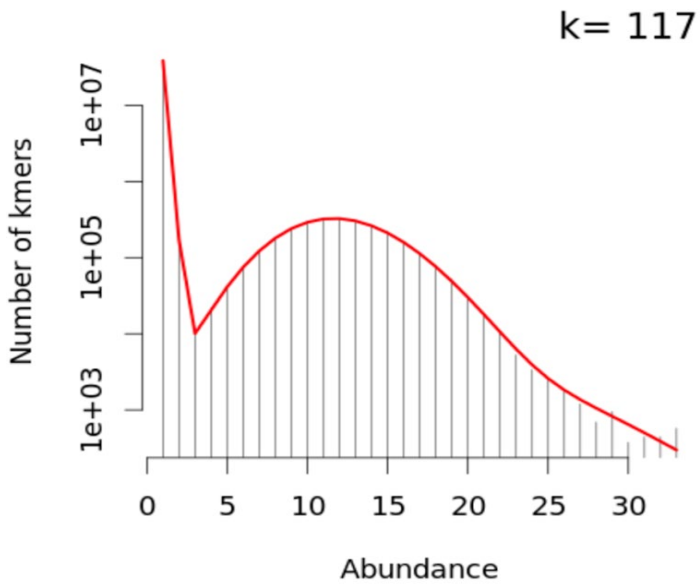


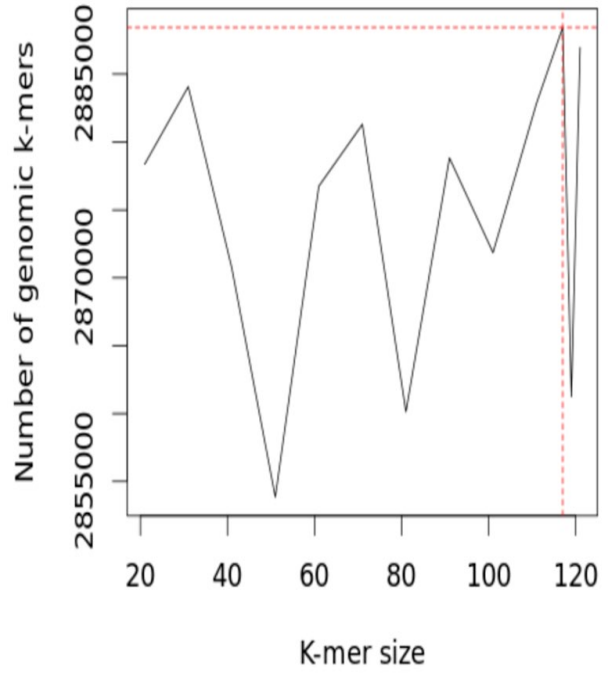
Figure - Shaun D. Jackman et.al [2017] “ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter”

kmer counter tools

- Some popular tools: Jellyfish, **kmergenie,** DSK, ntCard







Kmergenie suggested 117 sample - CGT3370

kmergenie suggestions

- kmers suggested in 17 - 117 range, mostly less than 41
- why is it suggesting very low kmers?

SAMPLE NAME	KMERGENIE SUGGESTION
CGT3002	27
CGT3058	41
CGT3130	27
CGT3136	25
CGT3158	41
CGT3246	27
CGT3292	41
CGT3323	41
CGT3335	31
CGT3370	117

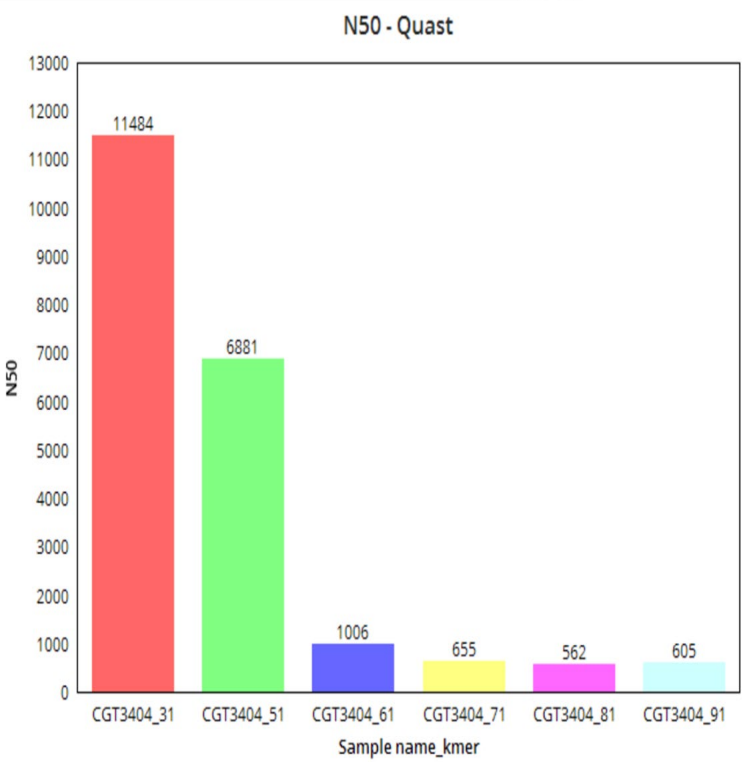
Abyss customized test run

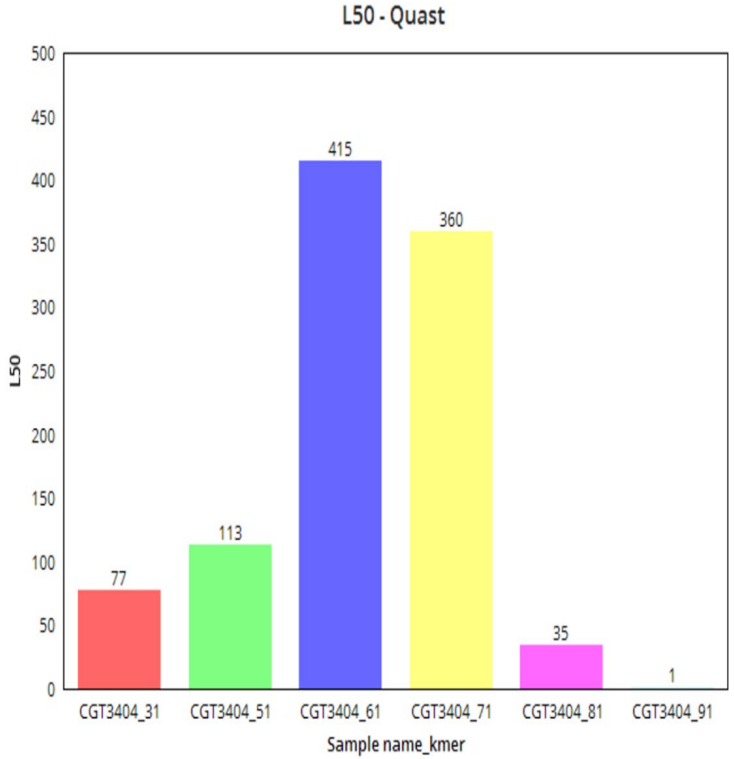
- kmers tested - 31,51,61,71,81,91,101

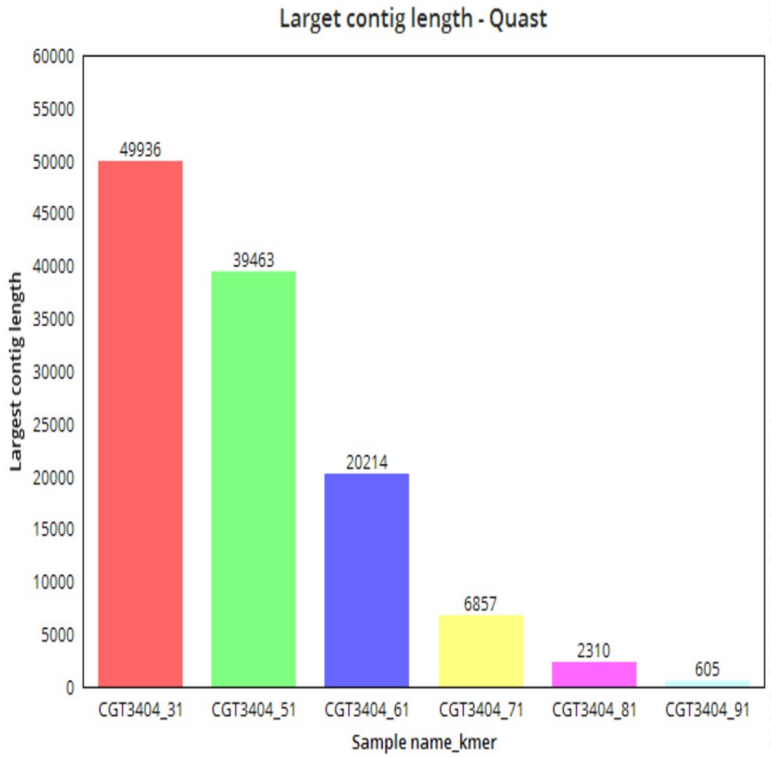
Group	Samples
Group 1	CGT3409[third], CGT3335[seventh], CGT3404[ninth]
Group 2	CGT3002[first], CGT3588[second], CGT3757 fourth], CGT3768[fifth] , CGT3827[sixth], CGT3390[eighth]
Worst	CGT3757[fourth], CGT3390[eighth]

Abyss test run - Quality check - Quast

- Sample CGT3404 [Group 1] - best kmer - 31

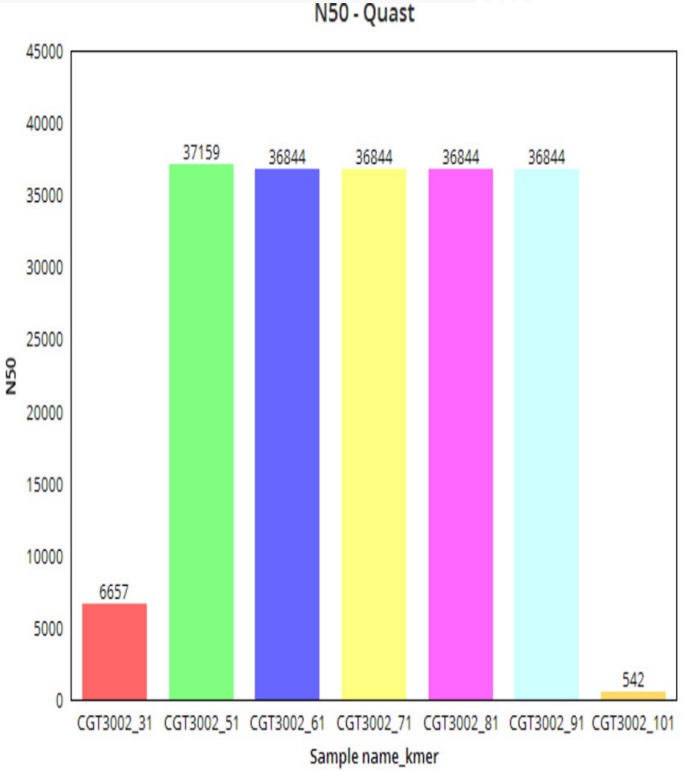


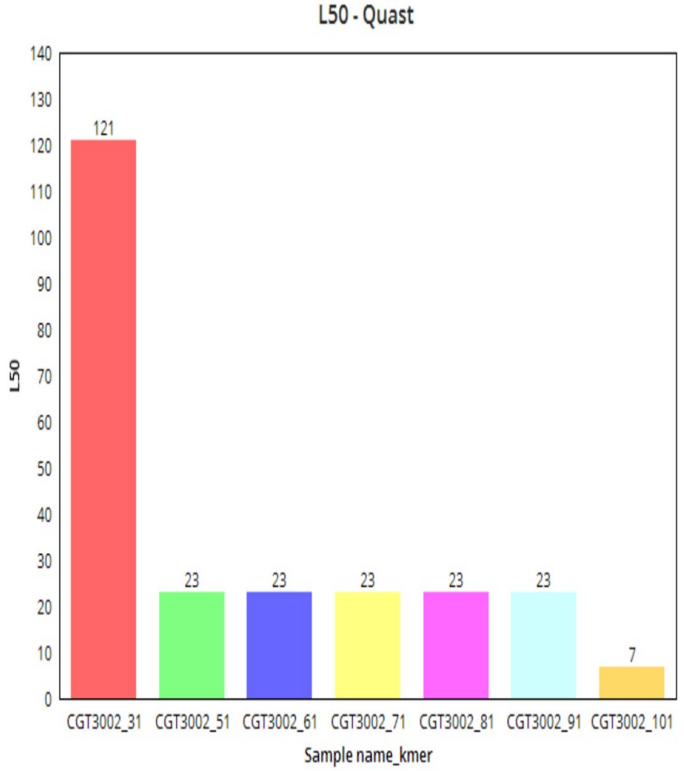


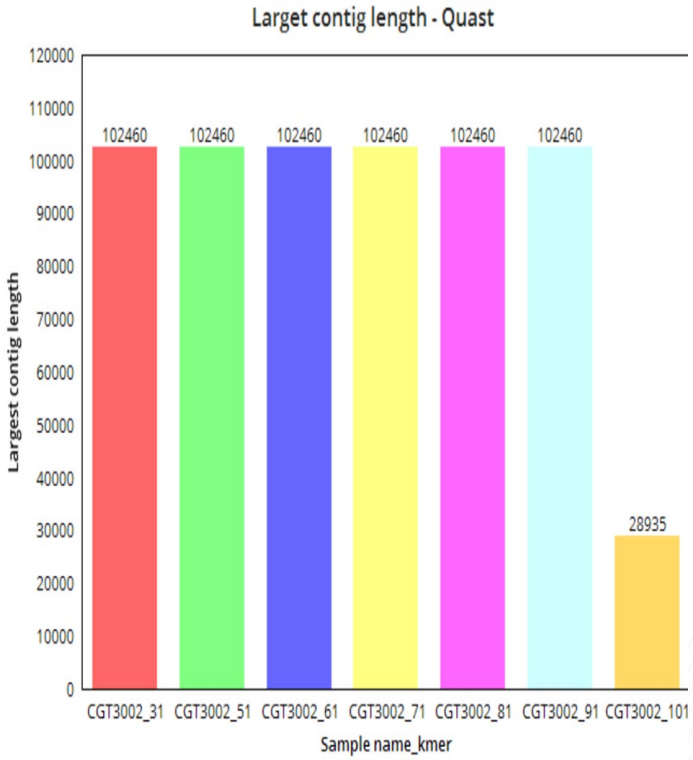


Abyss test run - Quality check - Quast

- Sample CGT3002 [Group 2] - best kmer - 51

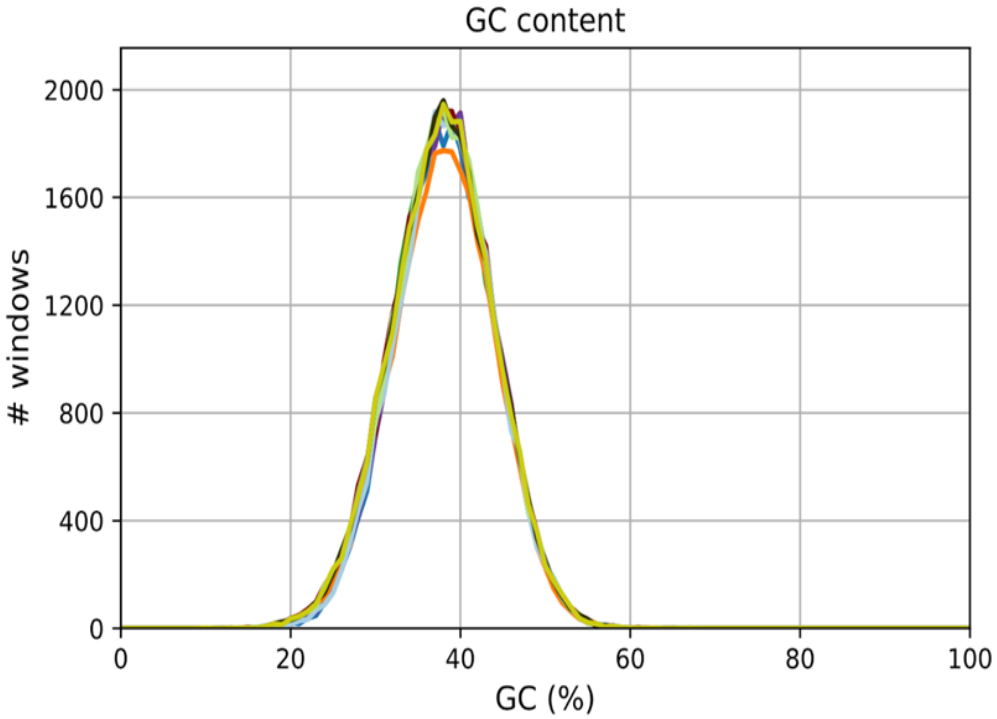


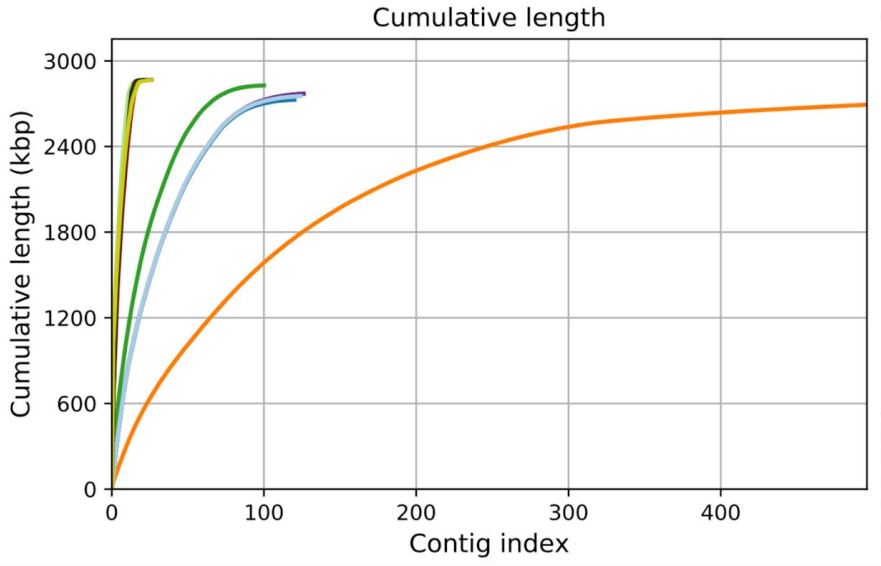




Abyss test run - Quality check - Quast

- kmer 31 for 9 samples

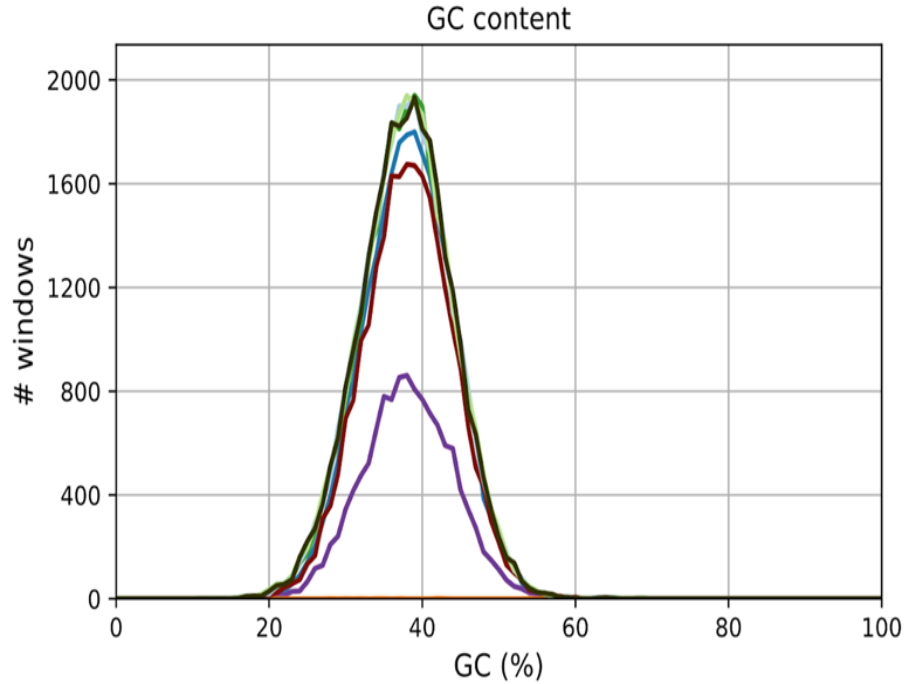


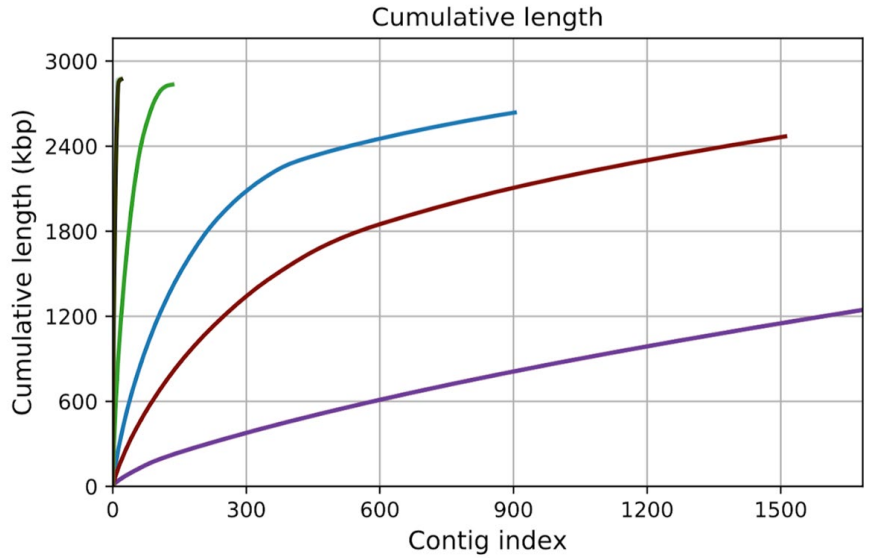


CGT3390_31CGT3390_31_contigs
kmerCGT3002_31kmerCGT3002_31_contigs
kmerCGT3335_31kmerCGT3335_31_contigs
kmerCGT3390_31kmerCGT3390_31_contigs
kmerCGT3404_31kmerCGT3404_31_contigs
kmerCGT3409_31kmerCGT3409_31_contigs
kmerCGT3588_31kmerCGT3588_31_contigs
kmerCGT3757_31kmerCGT3757_31_contigs
kmerCGT3768_31kmerCGT3768_31_contigs
kmerCGT3827_31kmerCGT3827_31_contigs

Abyss test run - Quality check - Quast

kmer 101 for 9 samples





CGT3390_101CGT3390_101_contigs
kmerCGT3002_101kmerCGT3002_101_contigs
kmerCGT3335_101kmerCGT3335_101_contigs
kmerCGT3390_101kmerCGT3390_101_contigs
kmerCGT3409_101kmerCGT3409_101_contigs
kmerCGT3588_101kmerCGT3588_101_contigs
kmerCGT3757_101kmerCGT3757_101_contigs
kmerCGT3768_101kmerCGT3768_101_contigs
kmerCGT3827_101kmerCGT3827_101_contigs

Kmergenie vs Custom test run

Sample	kmer by Kmergenie	kmer by customized test run
CGT3002	27	51
CGT3390	27	71
CGT3335	31	31
CGT3404	17	31
CGT3409	41	61
CGT3588	27	61
CGT3757	41	71
CGT3768	41	31
CGT3827	41	71

Abyss Final run

- Kmergenie suggestion for each sample
- parameters used :
- standard de Bruijn graph

```
abyss-pe k=kmervalue name=outputfilename in='file1 file2' v=-v
```

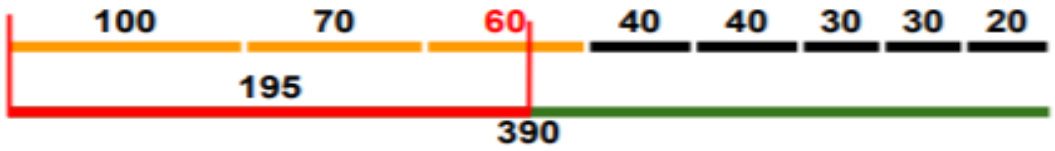
- Bloom filter de Bruijn graph

```
abyss-pe k=kmervalue name=filename  in='file_r1.fq file_r2.fq' B=100M H=3 kc=3 v=-v
```

Genome Assembly Quality metrics – QUAST and BUSCO

Quast-N50 family metrics

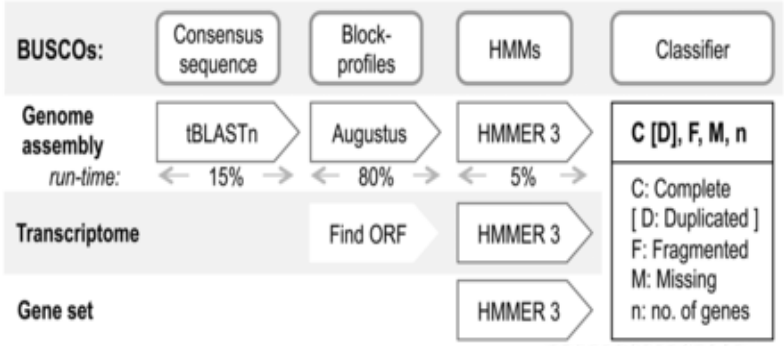
- Length of largest contig
- number of contigs
- N50 and L50
- N75 and L75



BUSCOs: Benchmarking Universal Single Copy Orthologs

- Single-Copy and Duplicated BUSCOS - BUSCO Completeness
- Fragmented BUSCOs - partially present
- Missing BUSCOs

- Lineage database for bacteria was used which contains 124 BUSCOs

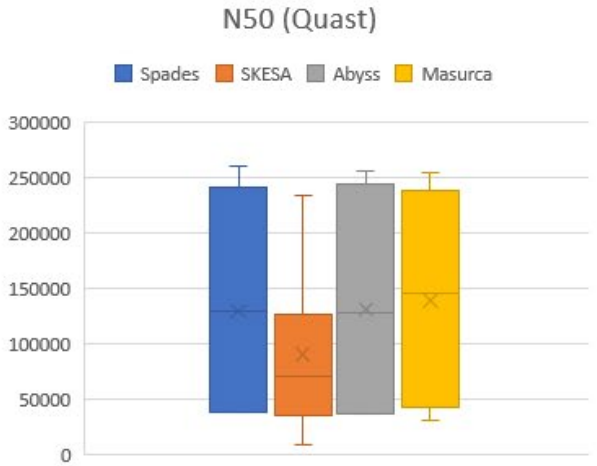


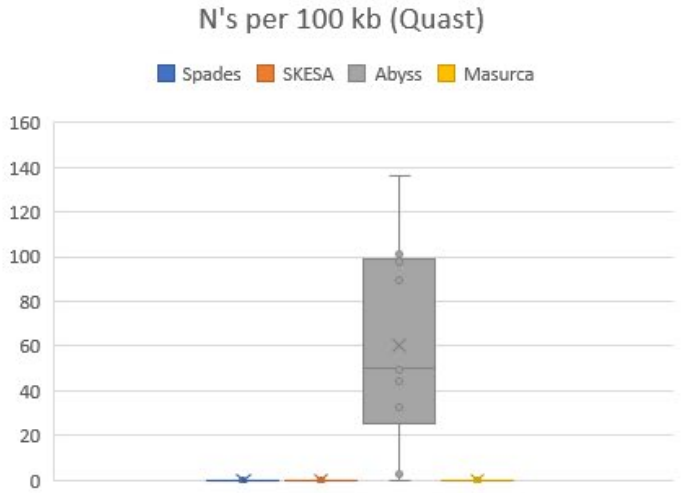

```
busco -m Genome -i config.fa -l bacteria_odb10 -o output
```

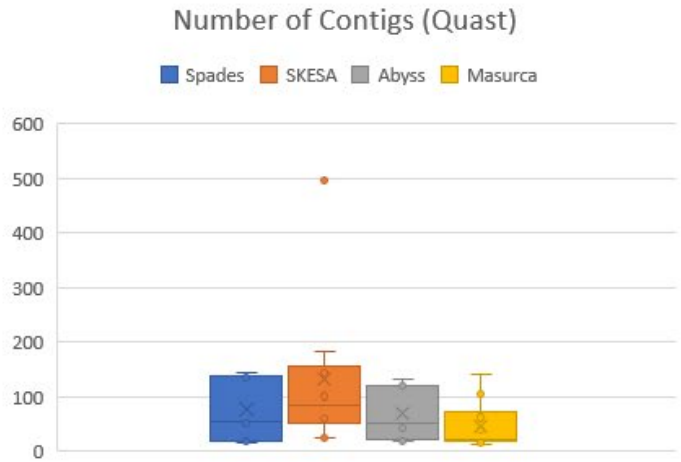
Quast N50 metrics with BUSCO scores on subset of samples

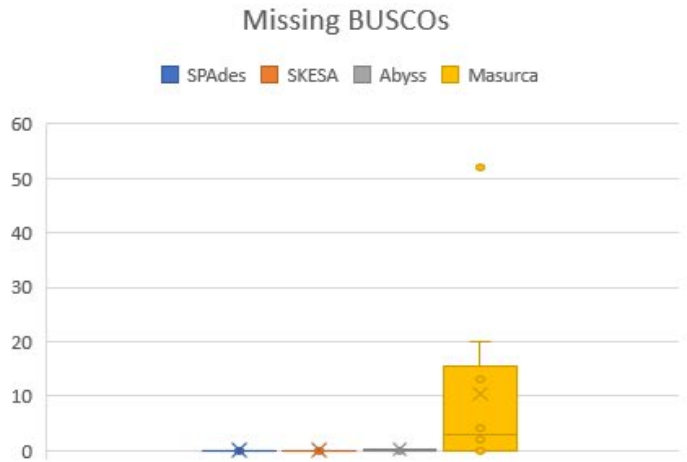
Abyss ruled out at this stage

Masurca ruled out at this stage



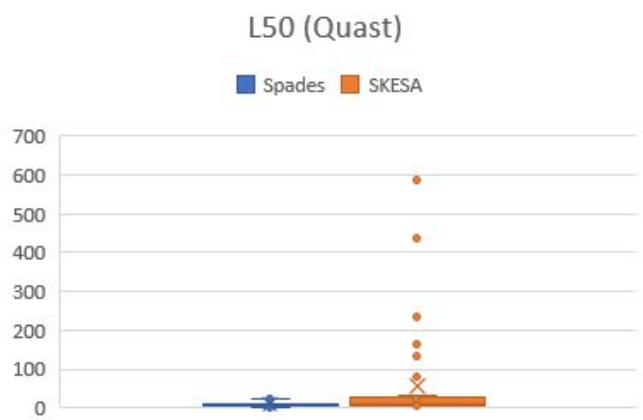


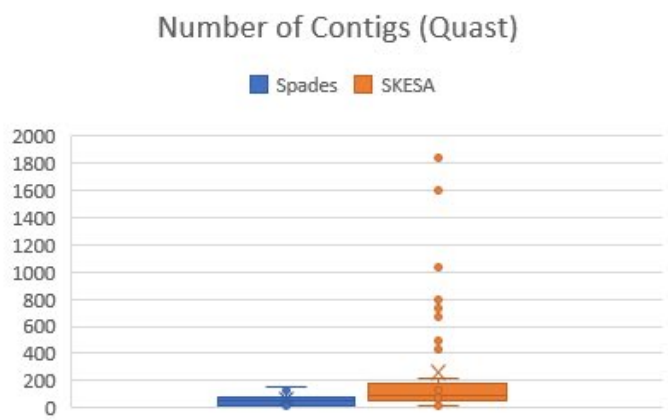


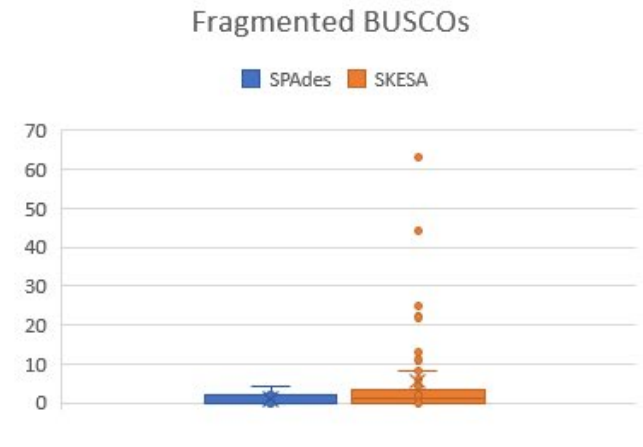


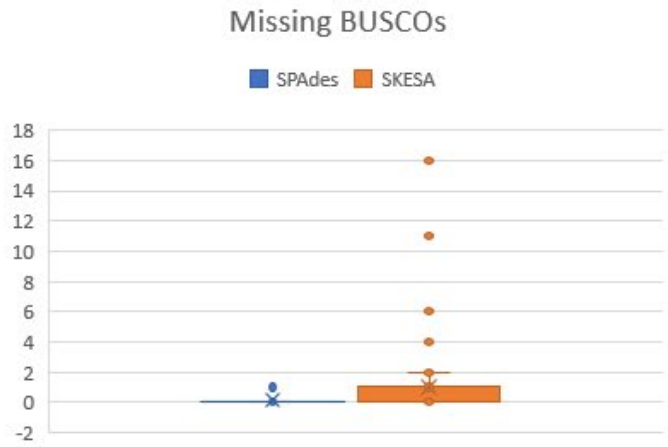
Georgia Tech
CREATING THE NEXT

SPAdes vs SKESA



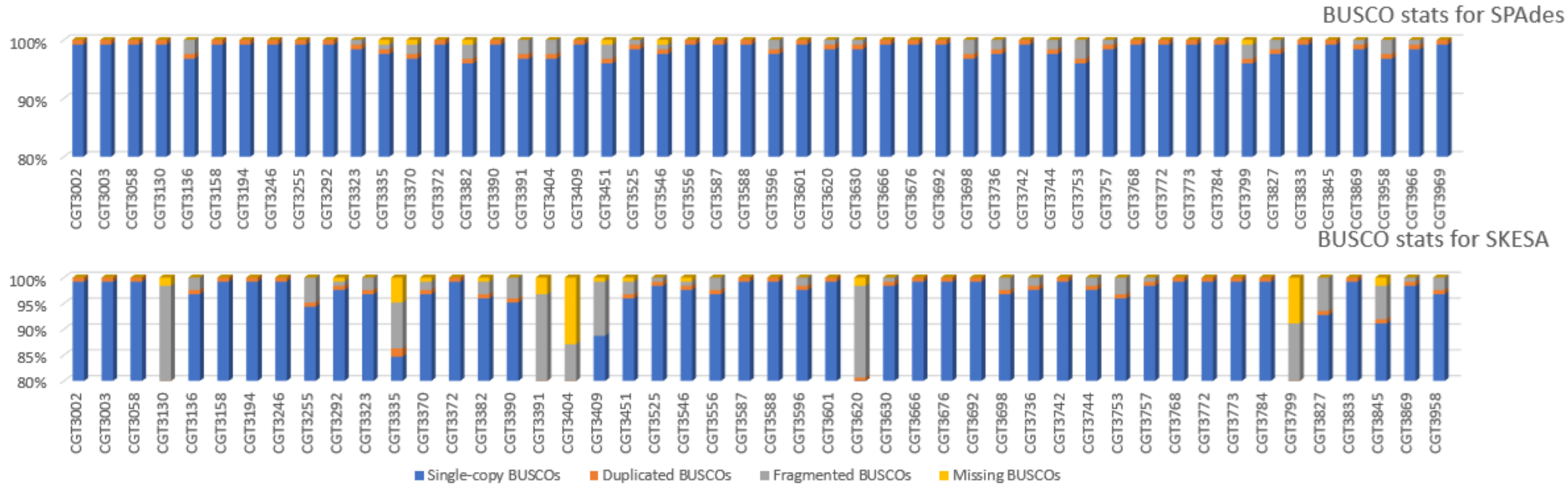


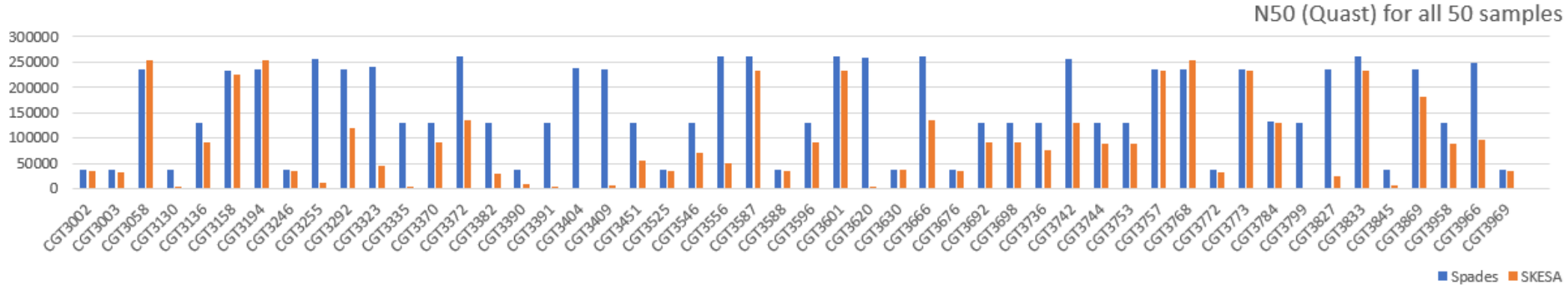




SPADES seems to be better with completeness of BUSCOs

Correlation of N50 with BUSCO for all the samples





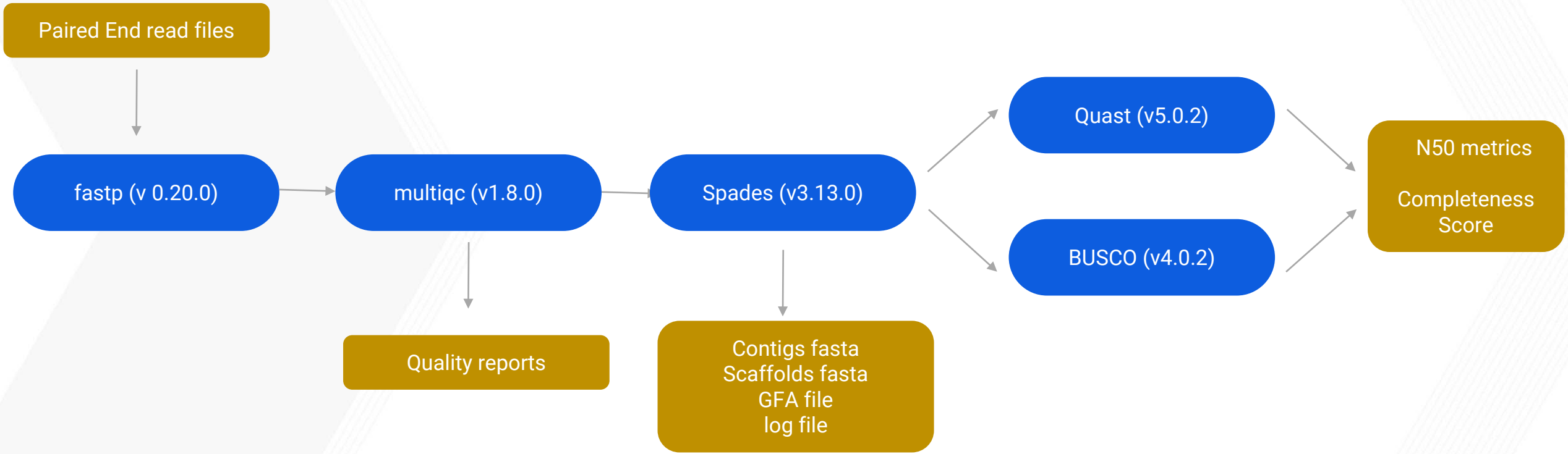
SKESA was not able to handle all the samples provided.

Georgia Tech
CREATING THE NEXT

Summary of Assemblers

Criteria	SPAdes	SKESA	MaSuRCA	Abyss
N50	Relatively large	Small	Relatively large	Relatively large
Optimization	Auto-detects k-mers	Auto-detects k-mers	Auto-detects k-mers	Requires external k-mer counter or optimization tool
Time	~8min	<5 min	30-40 mins	<5mins [Bloom filter]
Post-assembly QC	Good completeness scores	A few samples have very low completeness scores	missing BUSCOs, potential loss of data	as good as the kmer counter, possibility of N's in the contigs

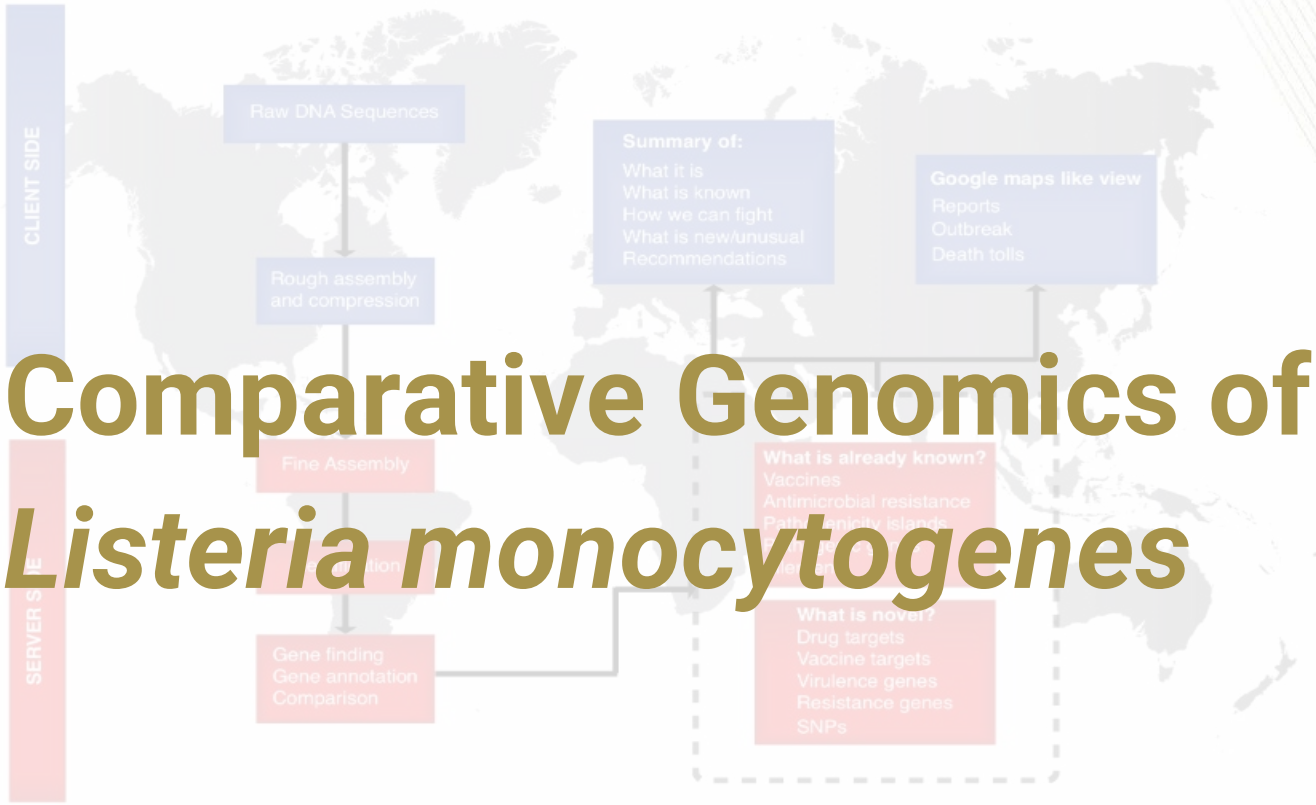
Final Pipeline Overview



Georgia Tech
CREATING THE NEXT

References

1. Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. Biological Procedures Online, 11, 52–78. doi:10.1007/s12575-009-9004-1
2. Dominguez Del Angel V, Hjerde E, Sterck L et al. Ten steps to get started in Genome Assembly and Annotation [version 1; peer review: 2 approved]. F1000Research 2018, 7(ELIXIR):148
3. Abdul Rafay Khan et.al [2018] - “A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective” - PMID: 29511353, doi: 10.1177/1176934318758650
4. Tanja Magoc et.al [2013] - “GAGE-B: an evaluation of genome assemblers for bacterial organisms” - PMID: 23665771, doi: 10.1093/bioinformatics/btt273
5. Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, Alexey Gurevich, Versatile genome assembly evaluation with QUAST-LG, Bioinformatics (2018) 34 (13): i142-i150. doi: 10.1093/bioinformatics/bty266
6. Bankevich, A.; Nurk, S. et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 19(5), 455–477. doi: 10.1089/cmb.2012.0021
7. Huang, Y.-T., & Liao, C.-F. (2016). Integration of string and de Bruijn graphs for genome assembly. Bioinformatics, 32(9), 1301–1307. doi: 10.1093/bioinformatics/btw011
8. Souvorov, A., Agarwala, R., & Lipman, D. J. (2018). SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biology, 19(1). doi: 10.1186/s13059-018-1540-z



Comparative Genomics of *Listeria monocytogenes*

Swetha Singu

Ruize Yang

Deepali Kundnani (Slide 40-41,57-62)

Gulay Bengu Ulukaya

Yuhua Zhang

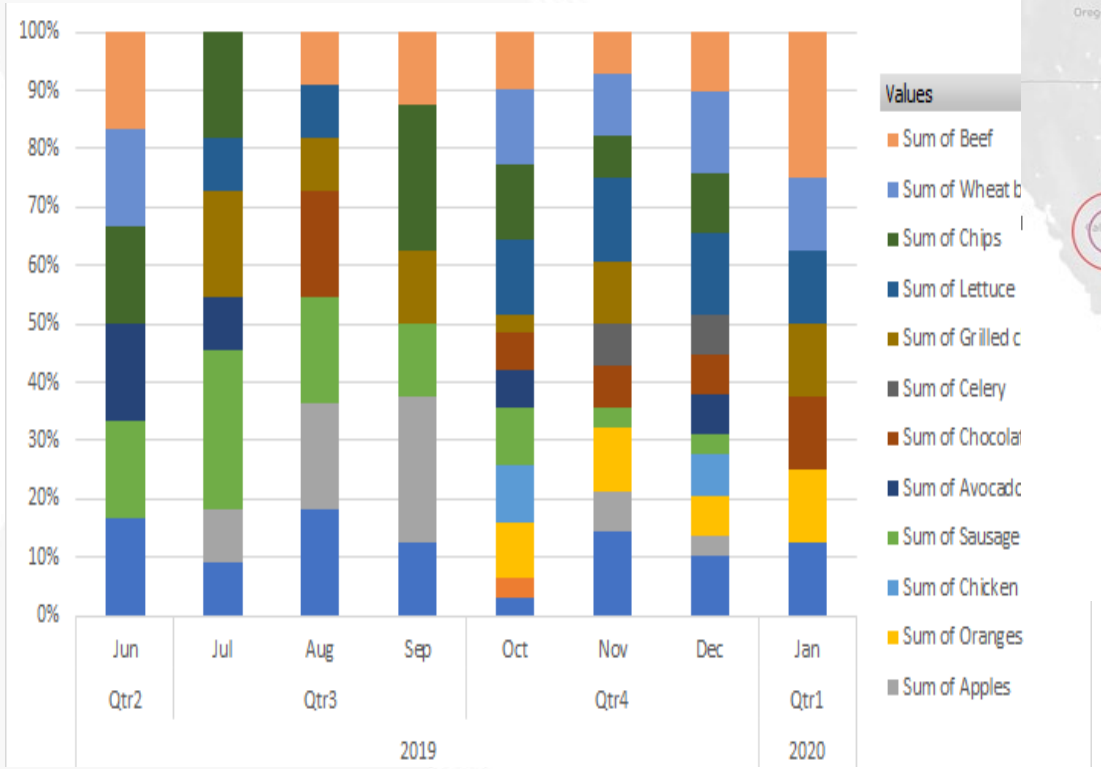
Jie Zhou

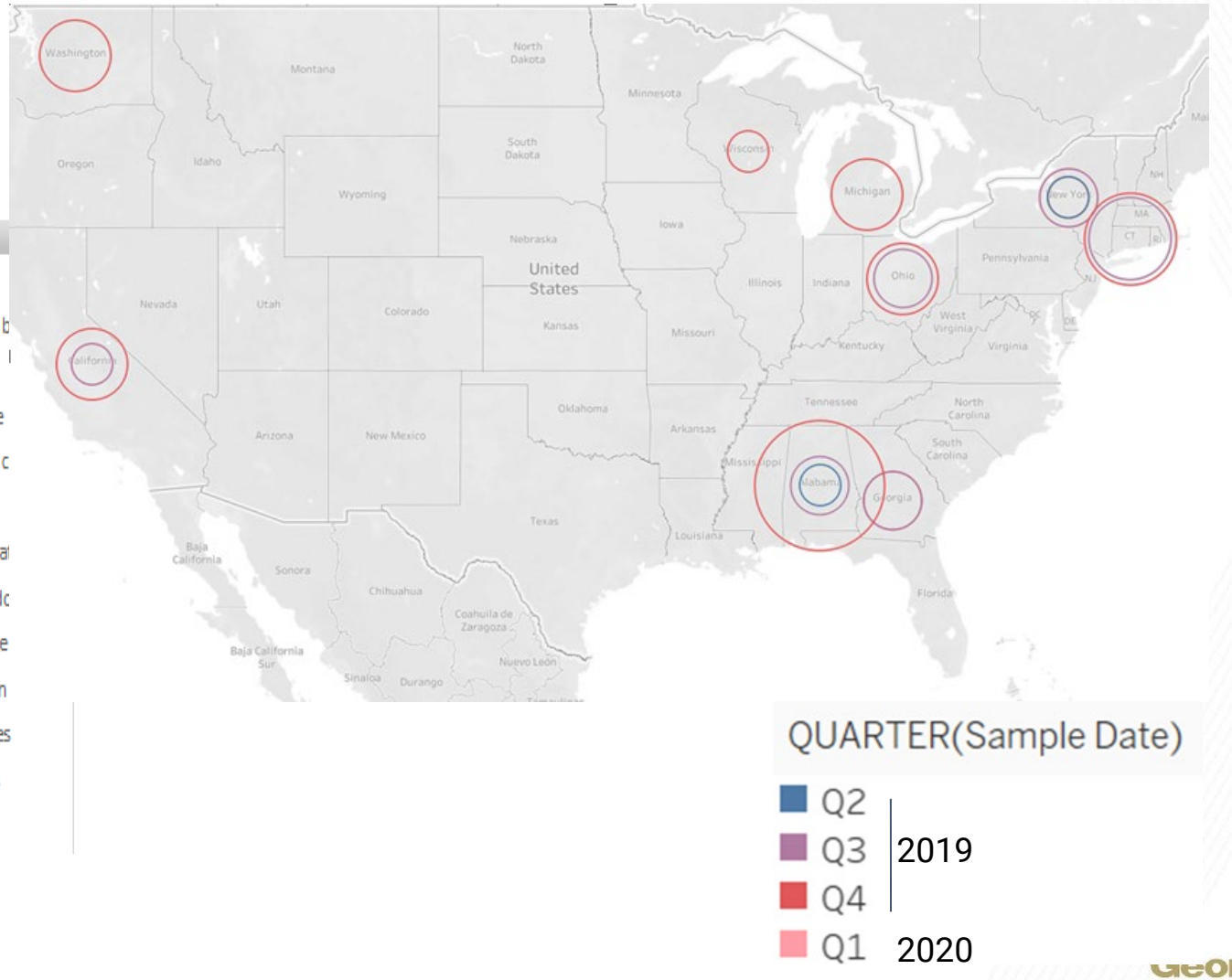
Information at hand - Analysis from previous groups

- Raw fasta, trimmed data, genes predicted, other functionally annotated genes.
- Genes - Virulence factors - VFDB [Virulence Factor Database]
- Genes - Antibiotic resistance - CARD [Comprehensive Antibiotic Resistance Database]
- Plasmid genes for Virulence and antibiotic resistance

Information at hand - Epidemiological Data

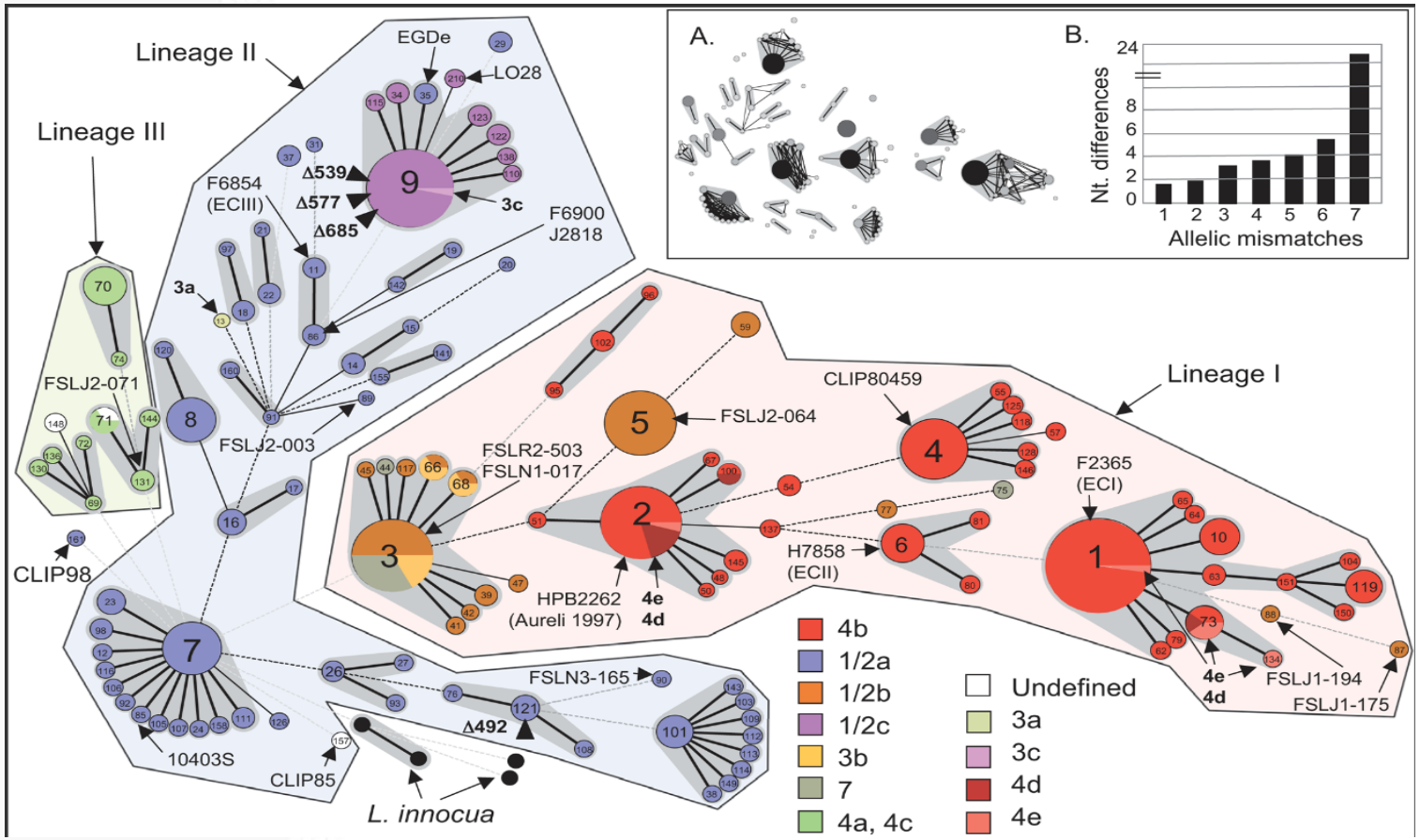
Percentage of food items consumed as per timeline





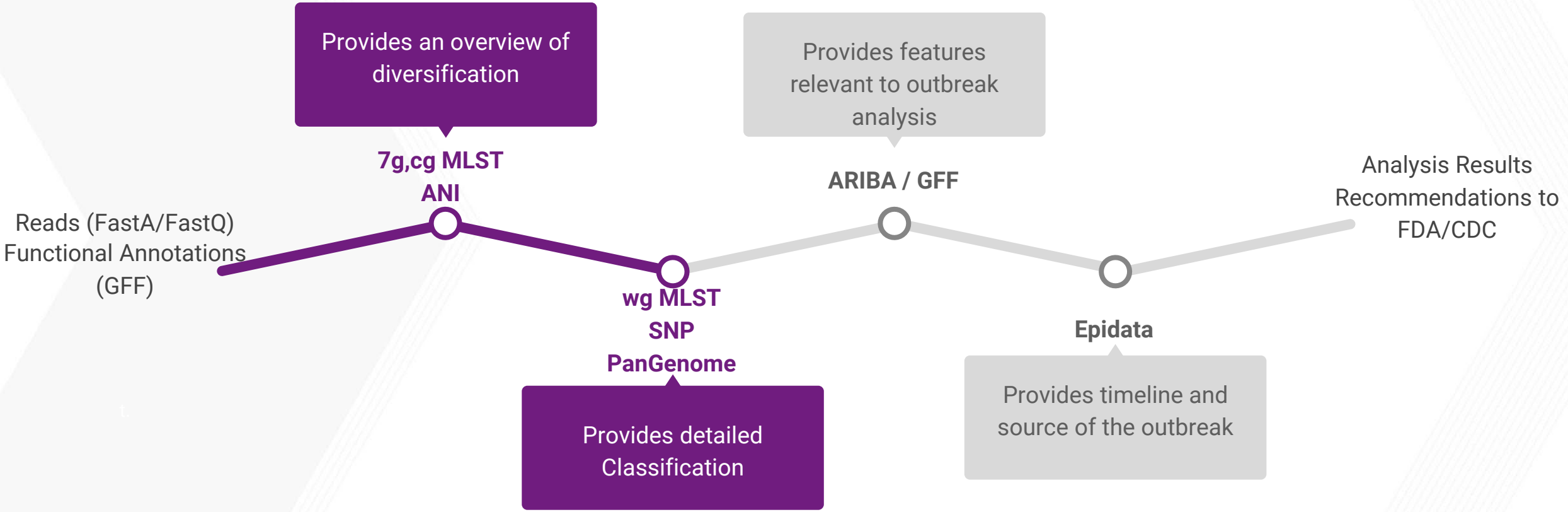
Georgia Tech
CREATING THE NEXT

What we tried to analyze?



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

Comparative Genomics Pipeline



Georgia Tech
CREATING THE NEXT

Average Nucleotide Identity (ANI)

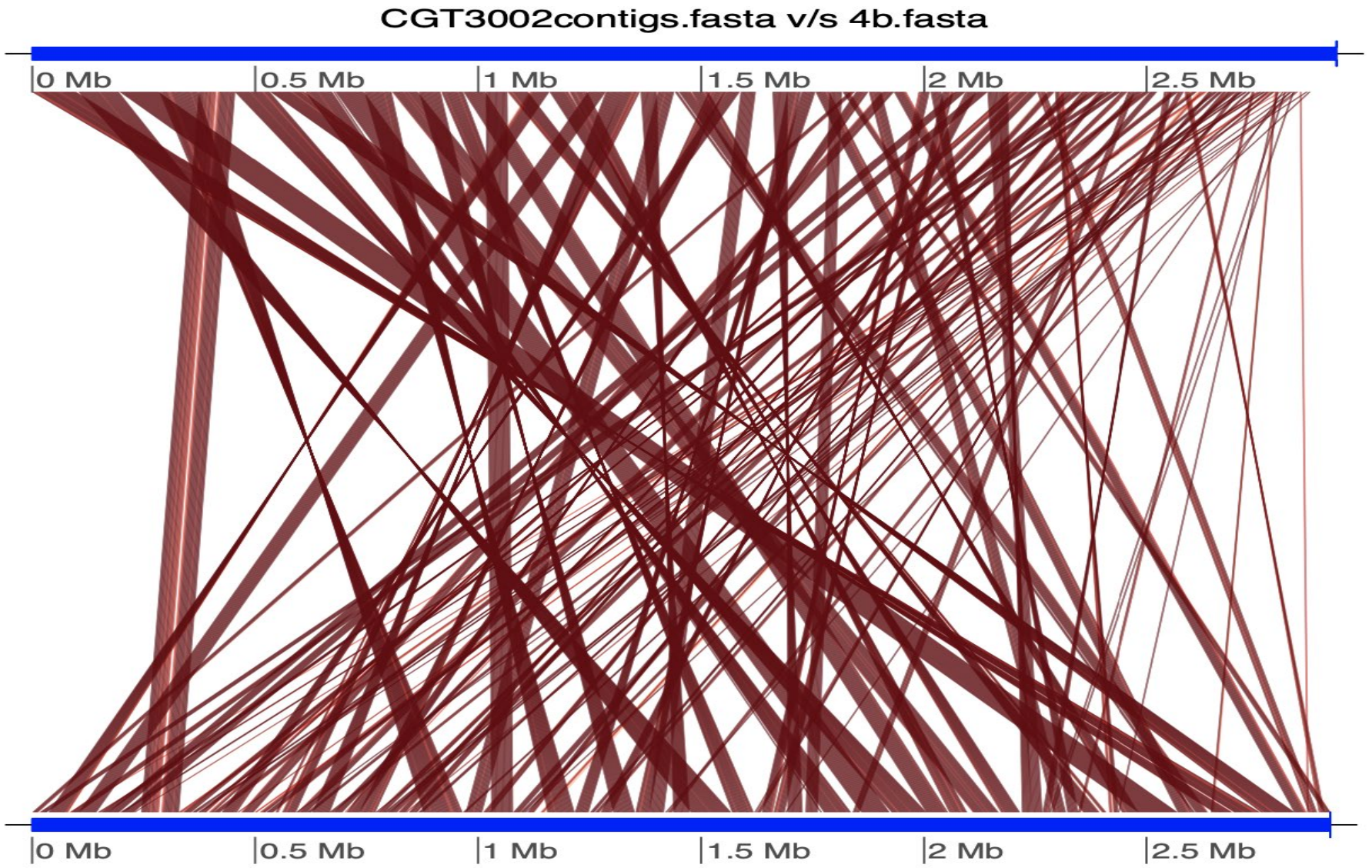
- We used FastANI
- Command line:

  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

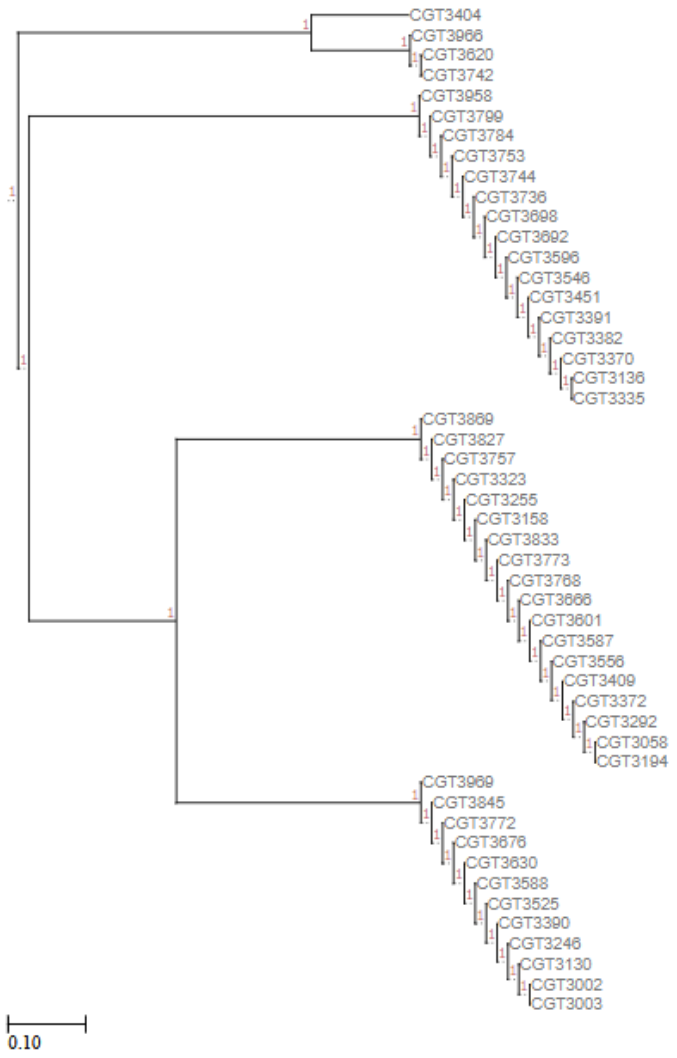
- Input: raw FASTQ files
- 7 housekeeping genes
- Used existing PubMLST schema of Listeria monocytogenes

  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)

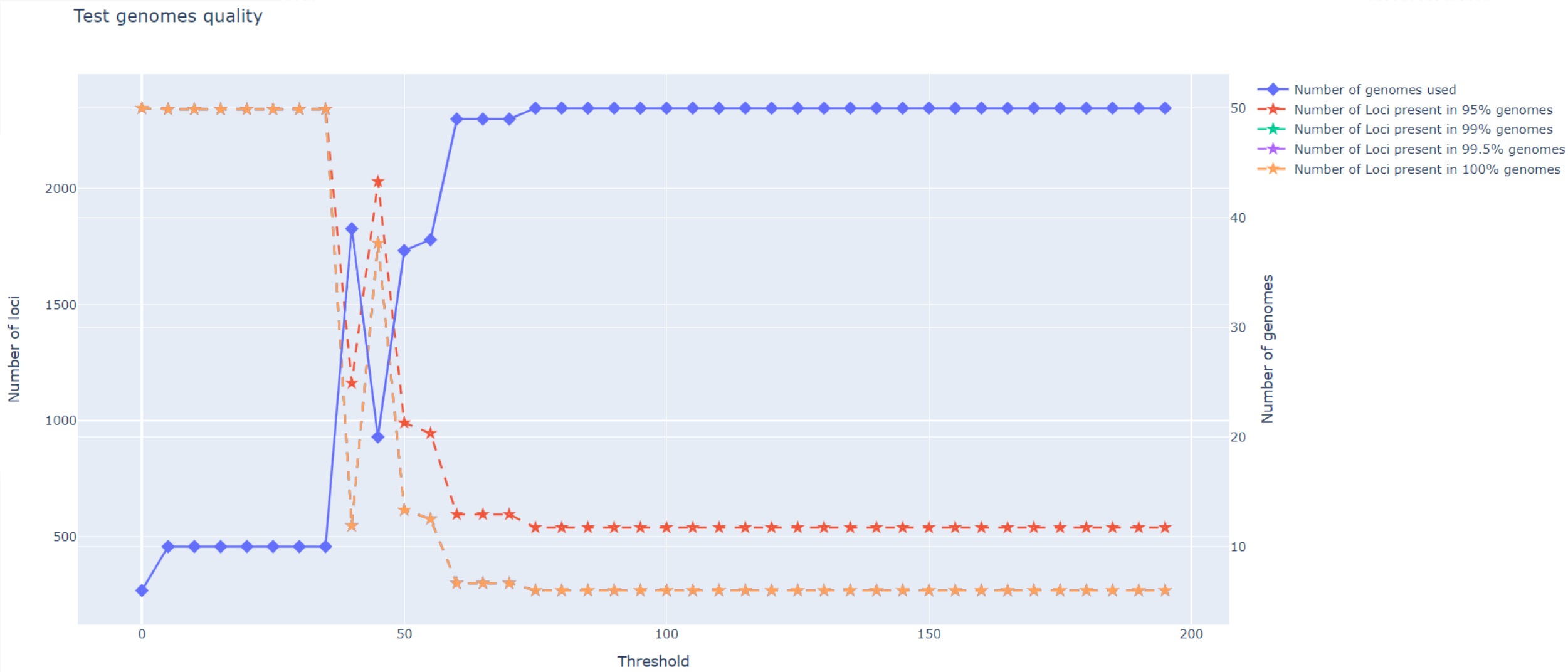
Tool 2: ChewBBACA

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

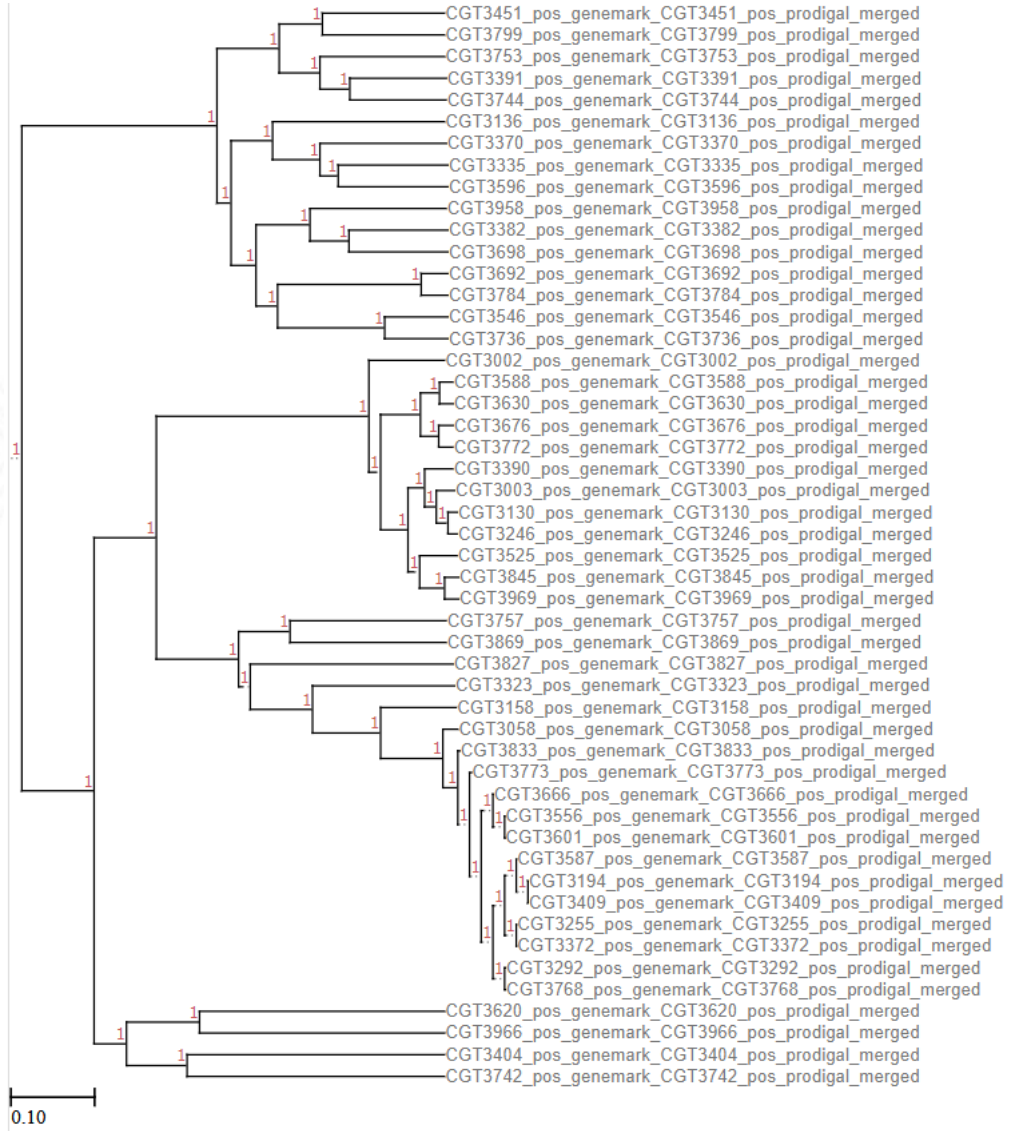
  chewBBACA.py CreateSchema
- Calling alleles from the schema

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

  chewBBACA.py ExtractCgMLST


Test genomes quality
Number of genomes used
Number of Loci present in 95% genomes
Number of Loci present in 99% genomes
Number of Loci present in 99.5% genomes
Number of Loci present in 100% genomes
Number of loci
Number of genomes
Threshold
2000
1500
1000
500
50
40
30
20
10
0
50
100
150
200

Phylogenetic Tree from ChewBBACA cgMLST



Georgia Tech
CREATING THE NEXT

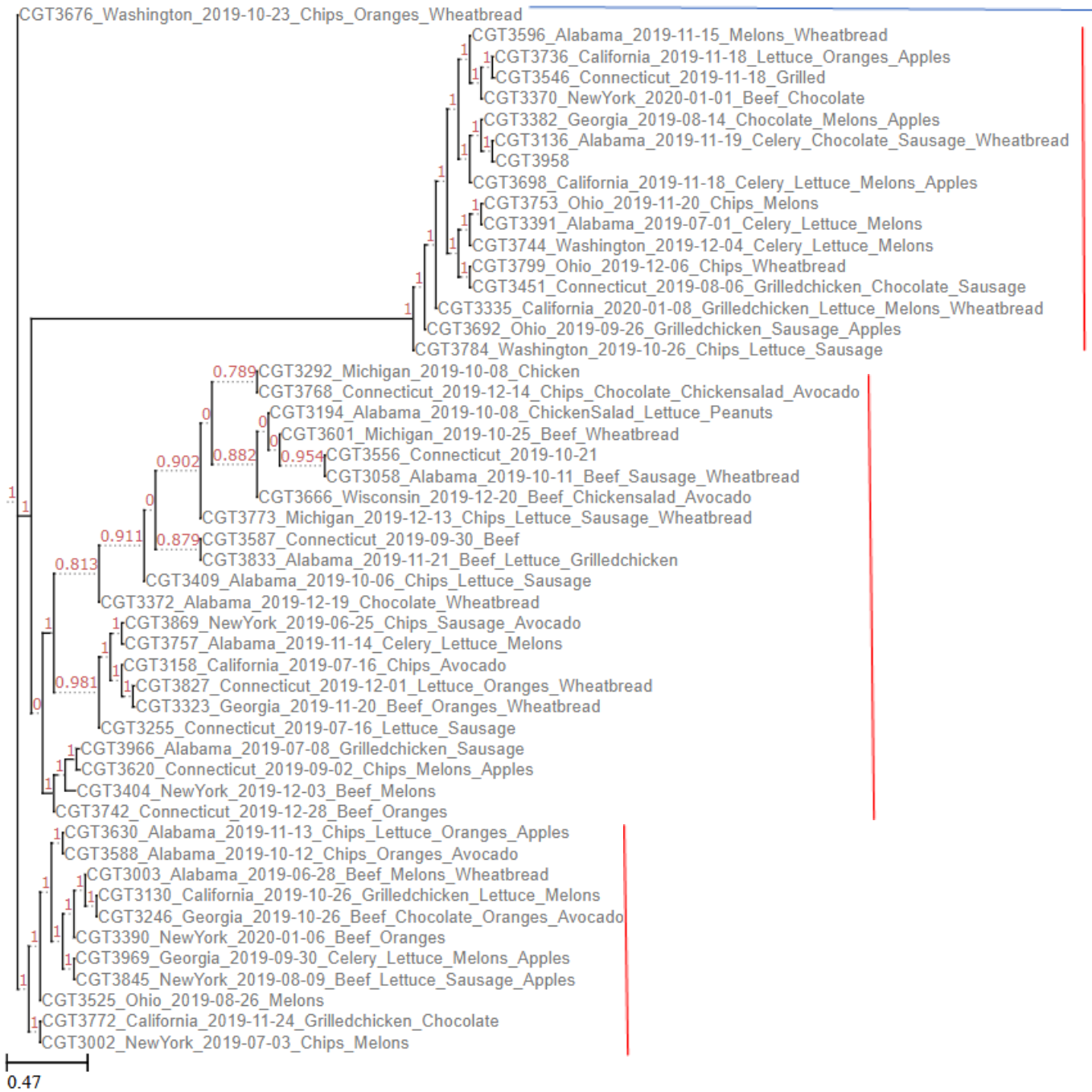
SNP-based Typing

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.
```

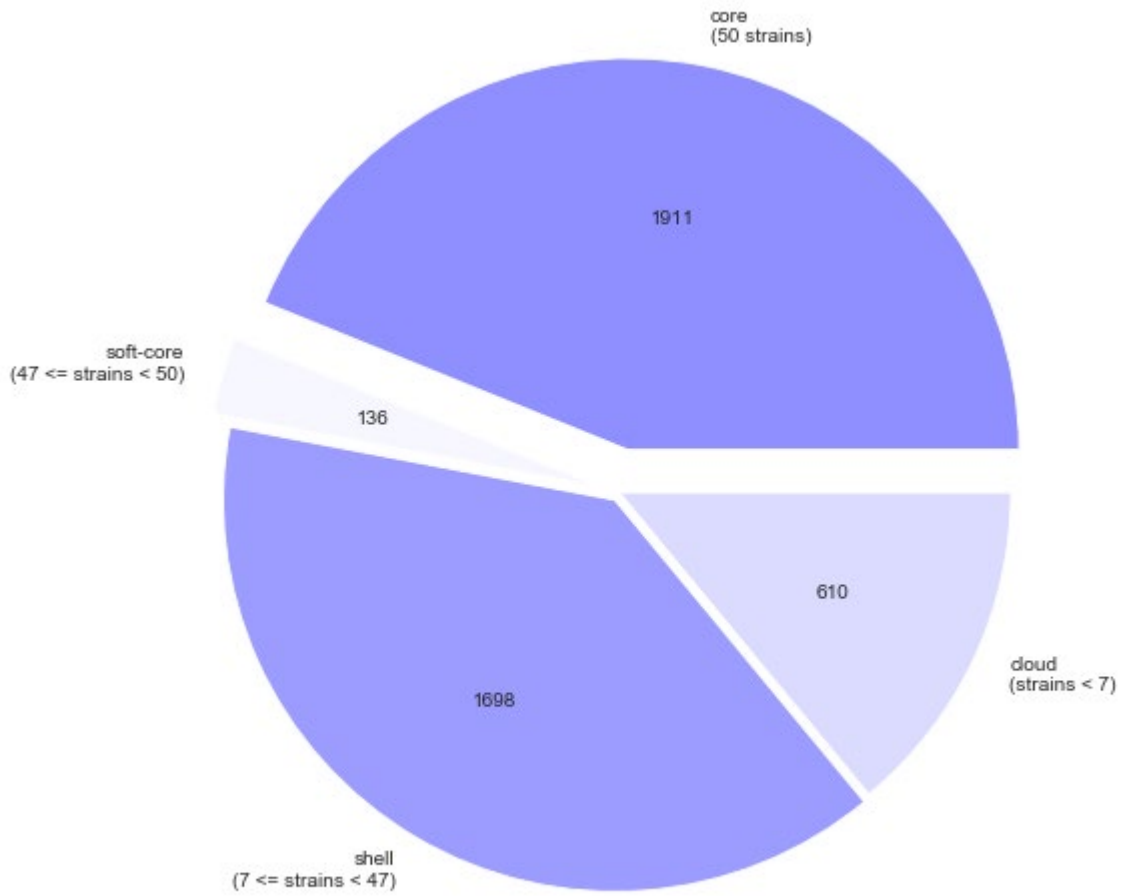


Maximum Parsimony Tree

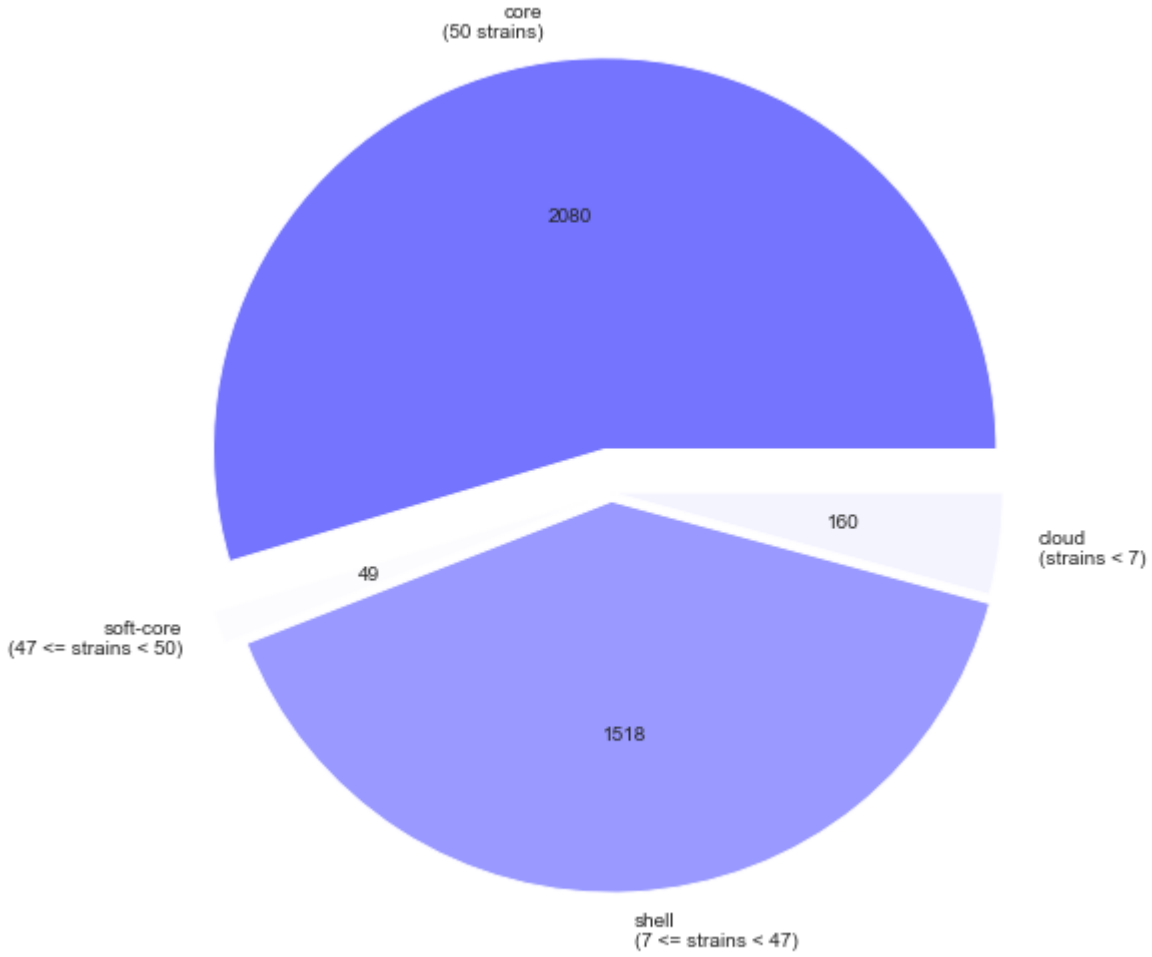
- Highest accuracy
- Fewest evolutionary change
- Fail to take into account many factors of sequence evolution
- 3 clusters
- Exclude 1 isolate

Georgia Tech
CREATING THE NEXT

Pan-genome analysis

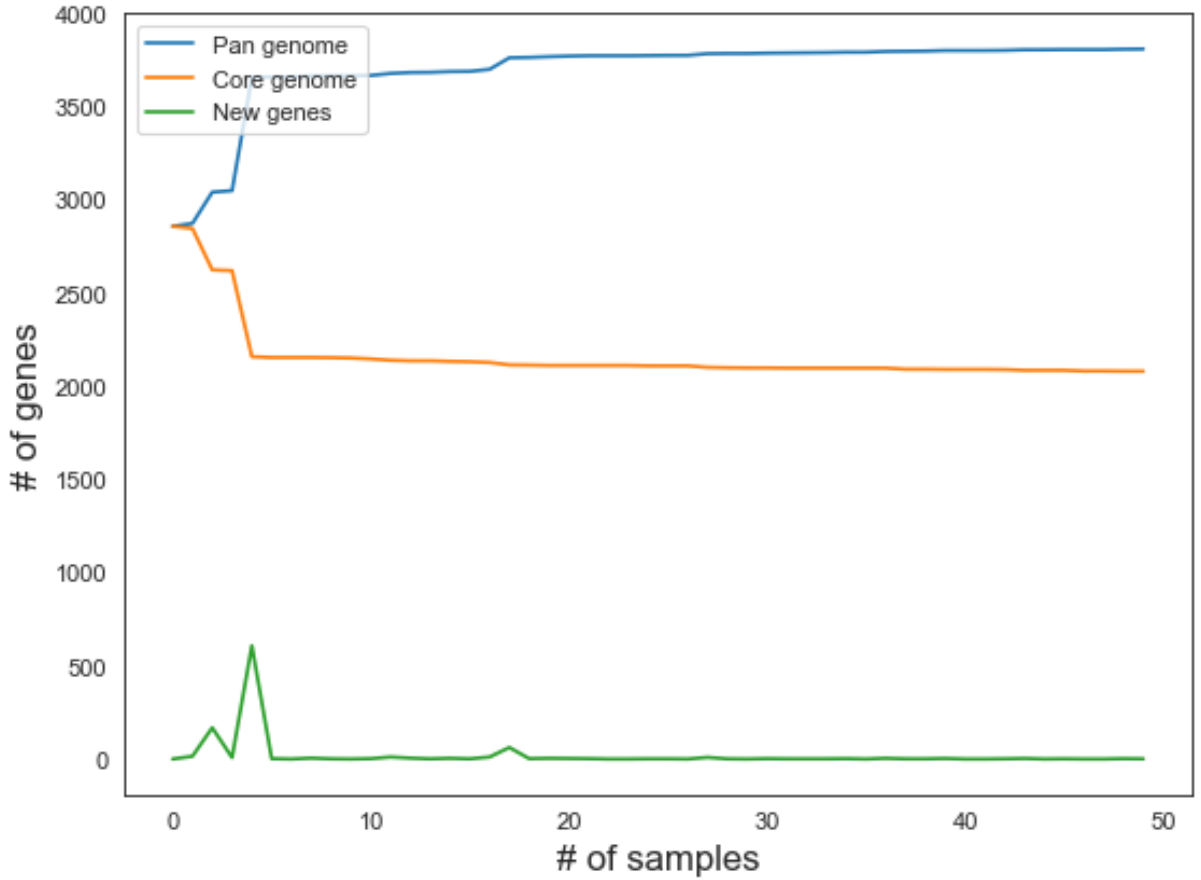


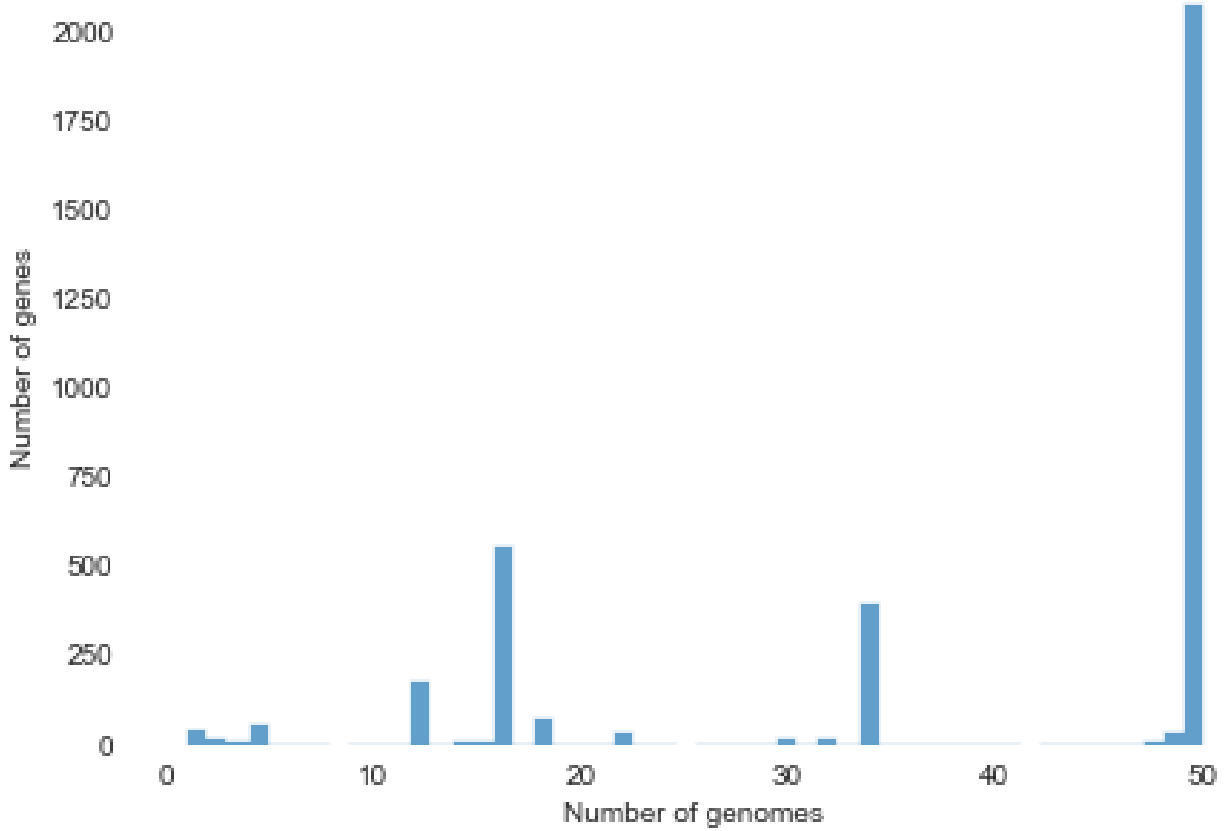
All genes



Coding genes

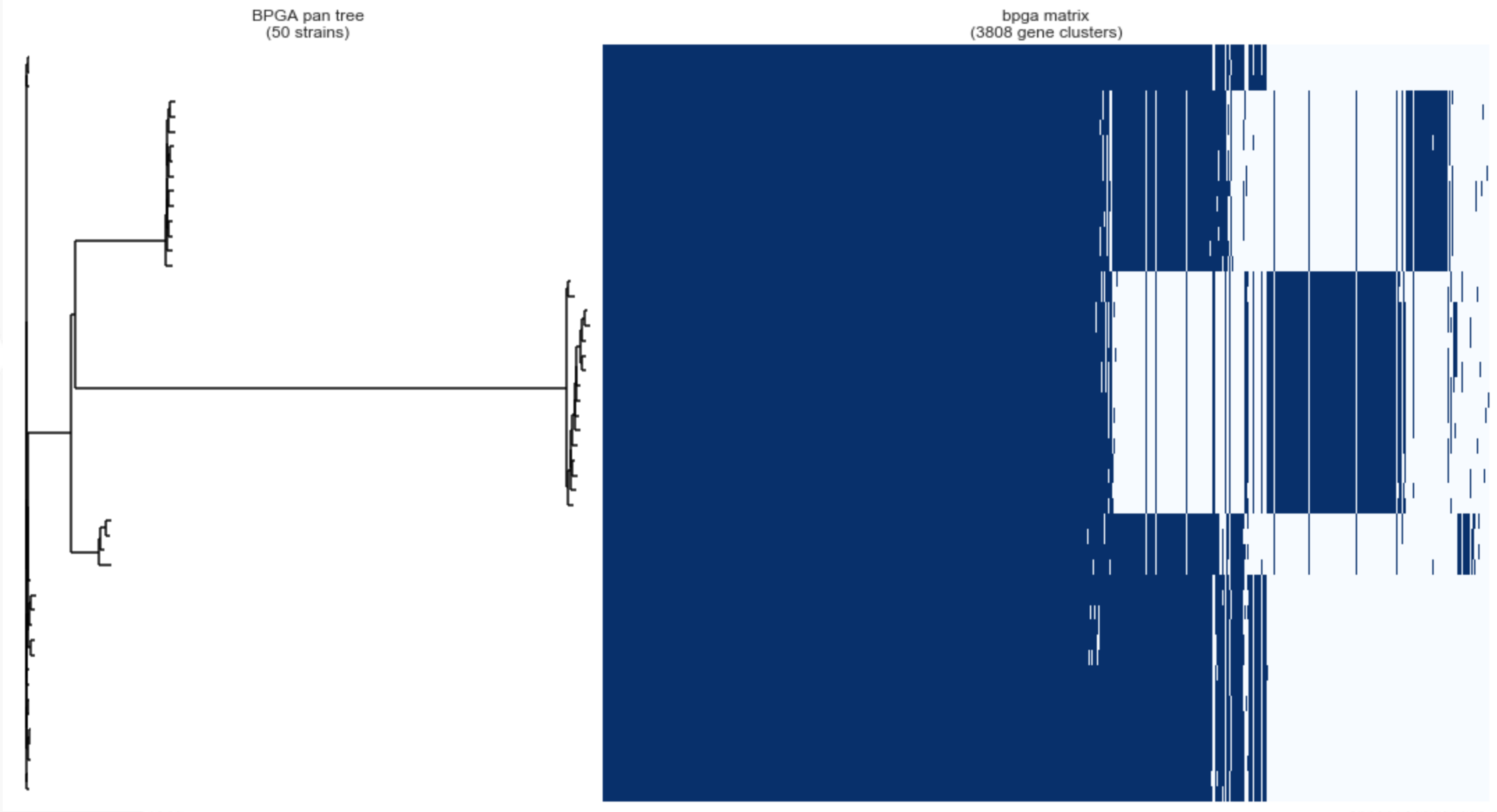
Pan-genome analysis



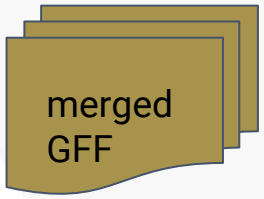


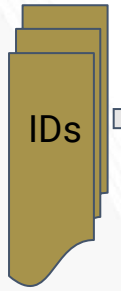
Georgia Tech
CREATING THE NEXT

Pan-genome analysis



Information extraction from merge annotated data





0	lmo1421	lmo2557	LMOSA_24590	LMOSA_1270	lmo0319	LMOSA_28520	LMO
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

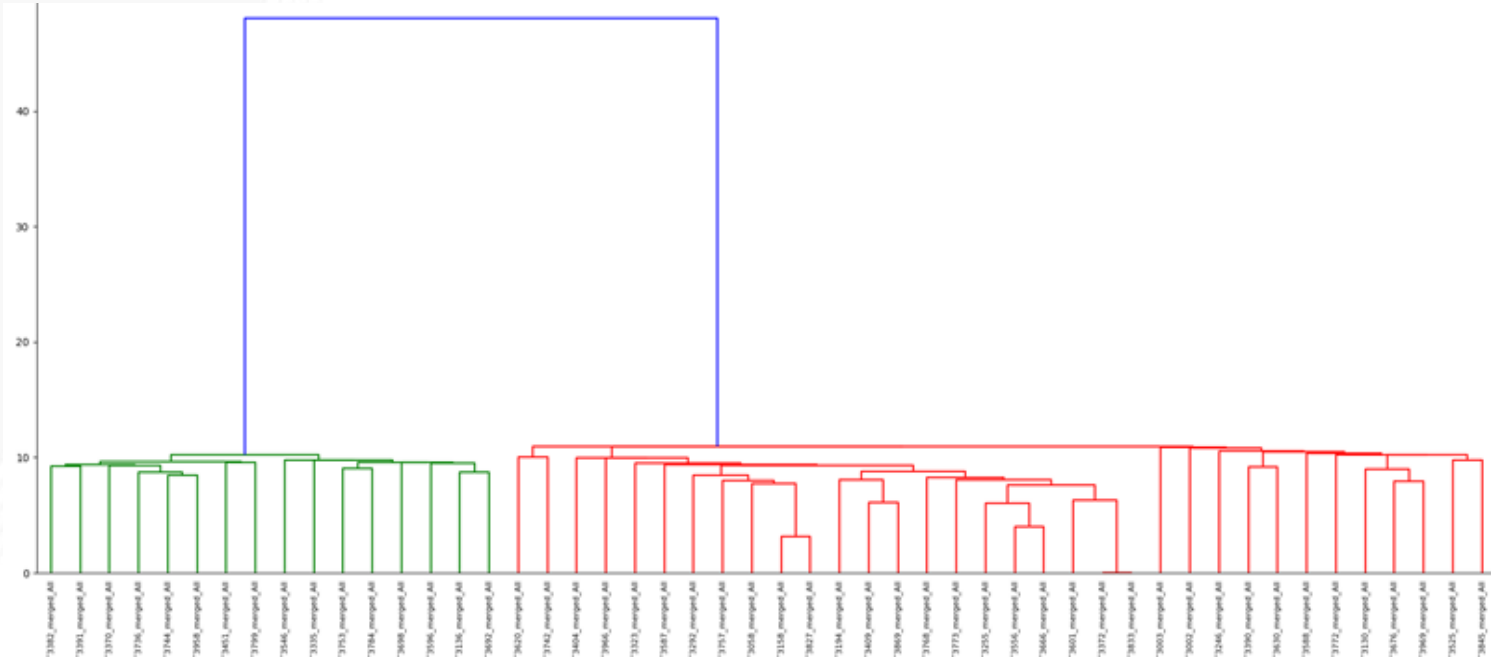
Presence/Absence Matrix

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

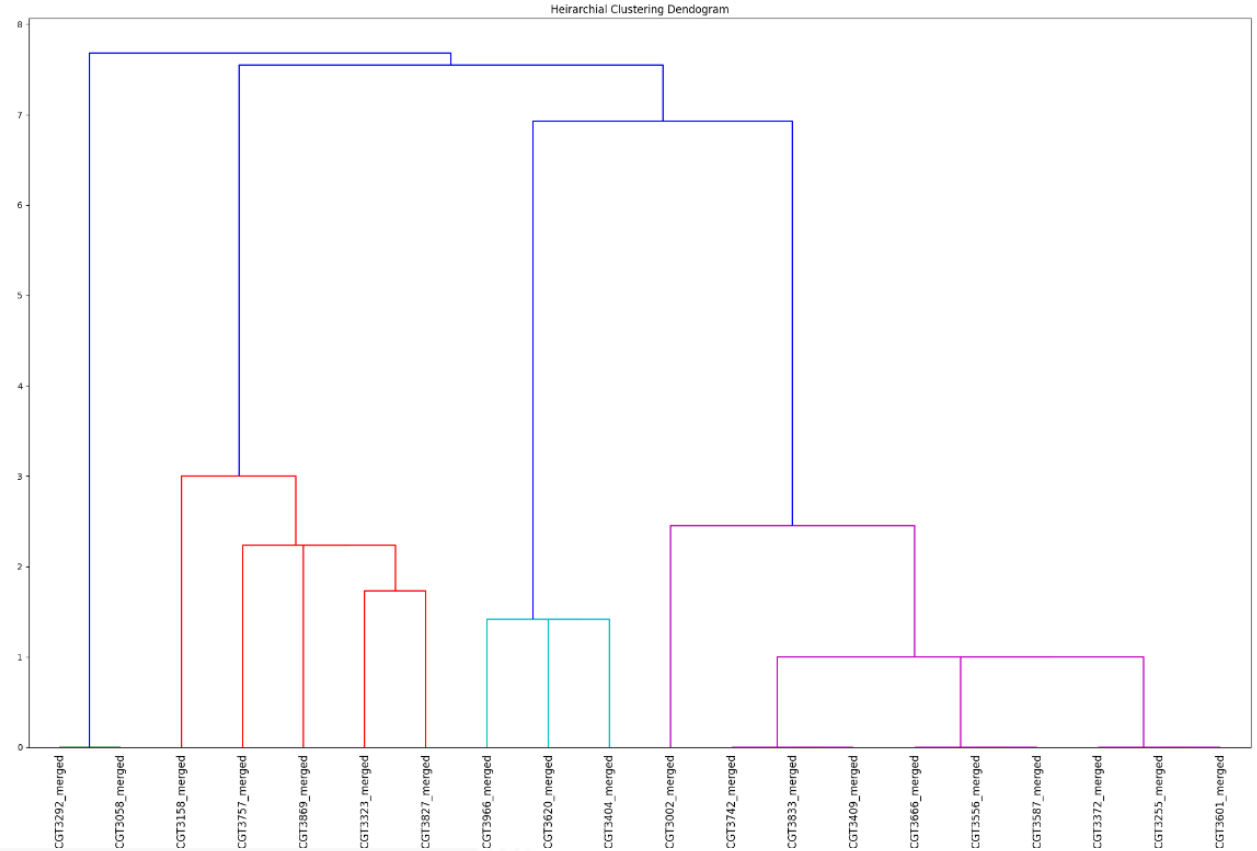
Hierarchical Clustering (Scipy)

Voor Hees Algorithm

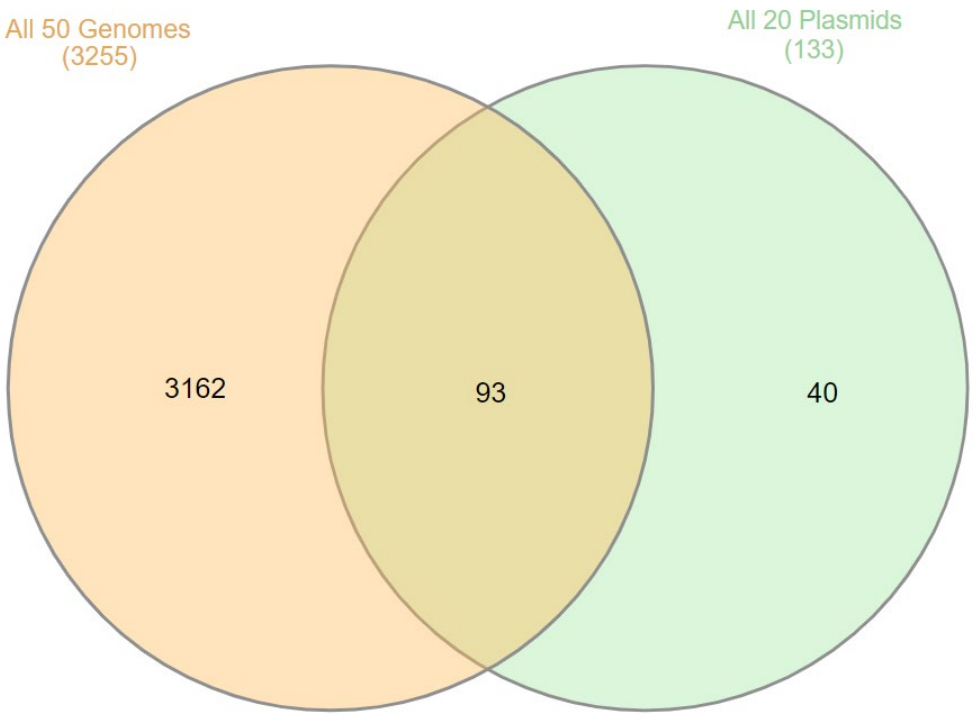
$$d(u, v) = \max(dist(u[i], v[j]))$$



GFF analysis of Plasmids

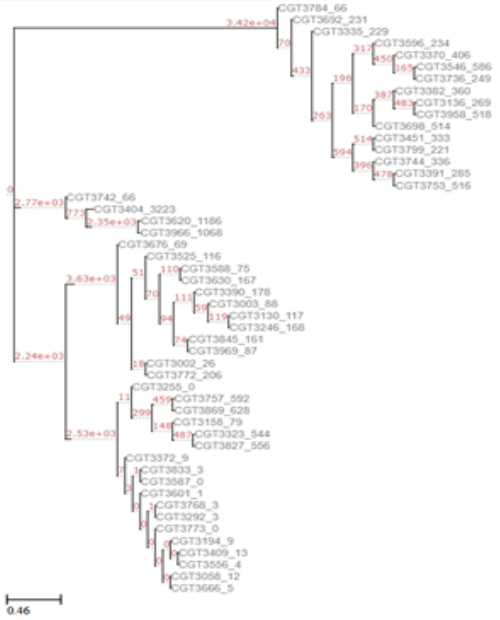


Hierarchical clustering of merged GFF files annoted on assembly files generated using plasmidSPades

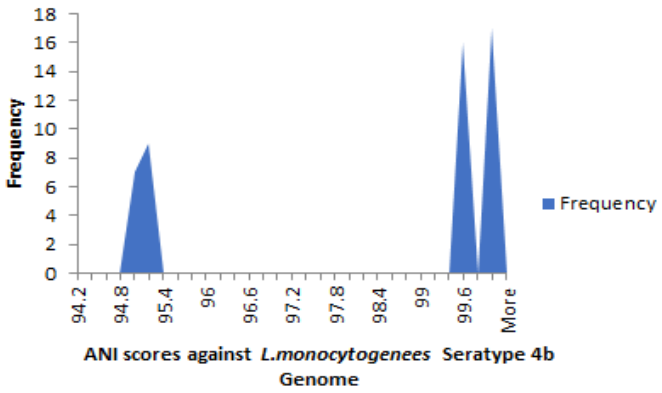


Unique annotations uncovered in plasmid data

Correlation of clusters with different typing analysis

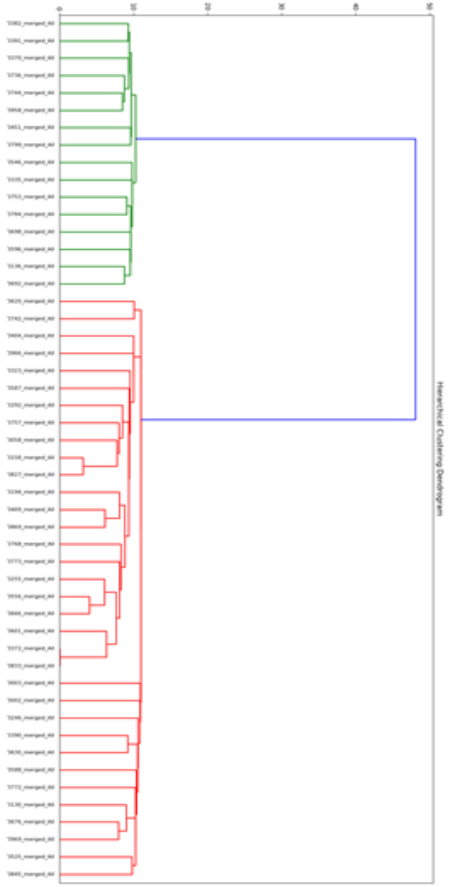


Tip Allele SNP ML Tree

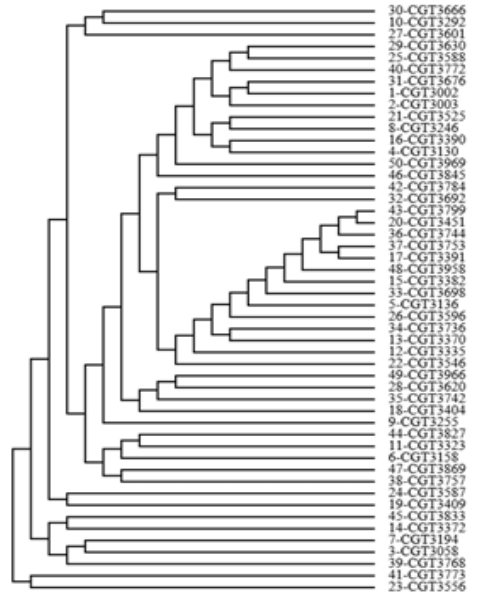


ANI scores against *L.monocytogenees* Seratype 4b Genome

7gMLST



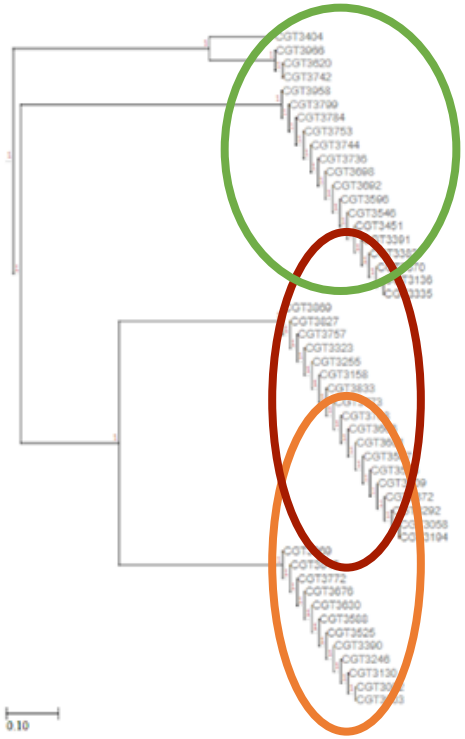
Hierarchical Clustering from merged annotations



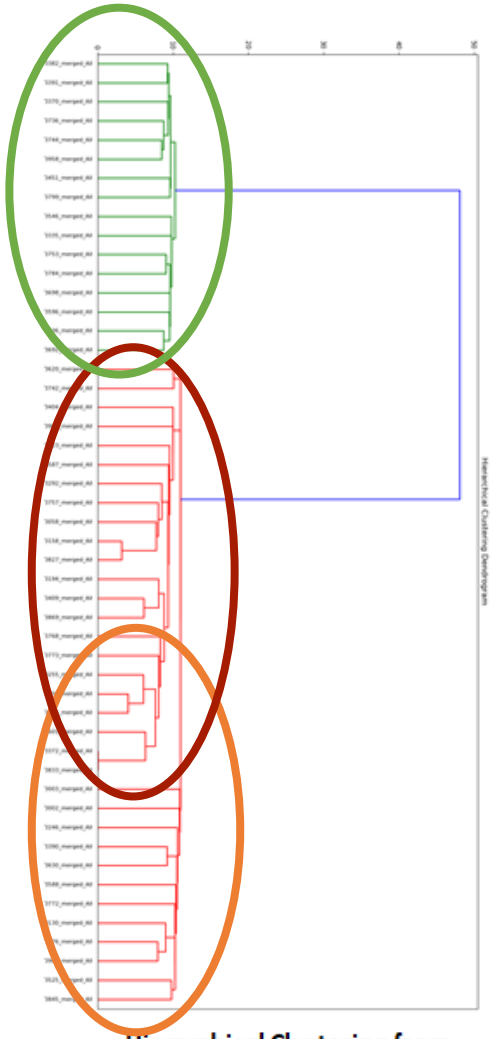
BPGA Pan Genome Analysis

Correlation of clusters with different type of analysis

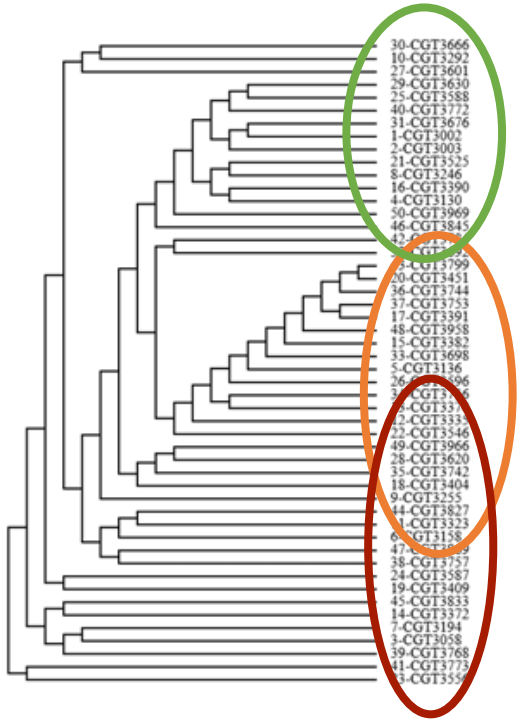
Tip Allele SNP ML Tree



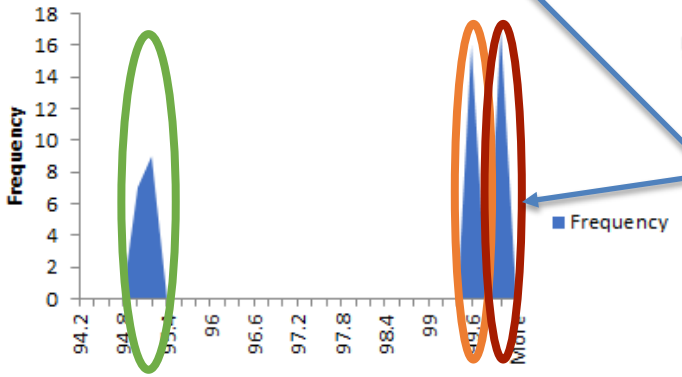
7gMLST



Hierarchical Clustering from merged annotations



BPGA Pan Genome Analysis



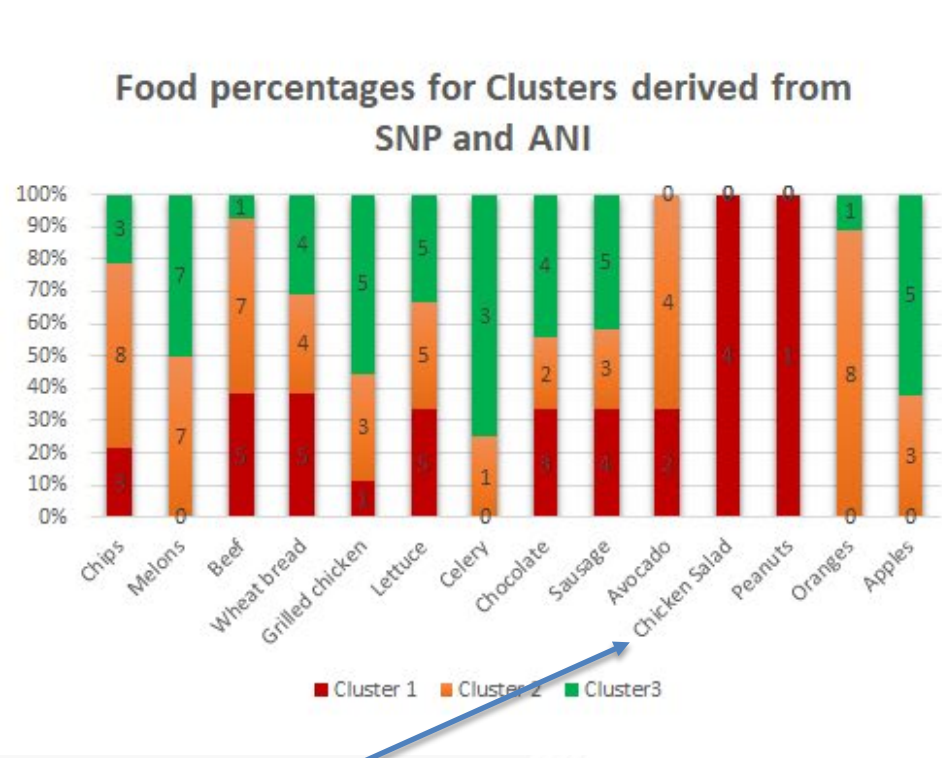
ANI scores against *L.monocytogenees* Seratype 4b Genome

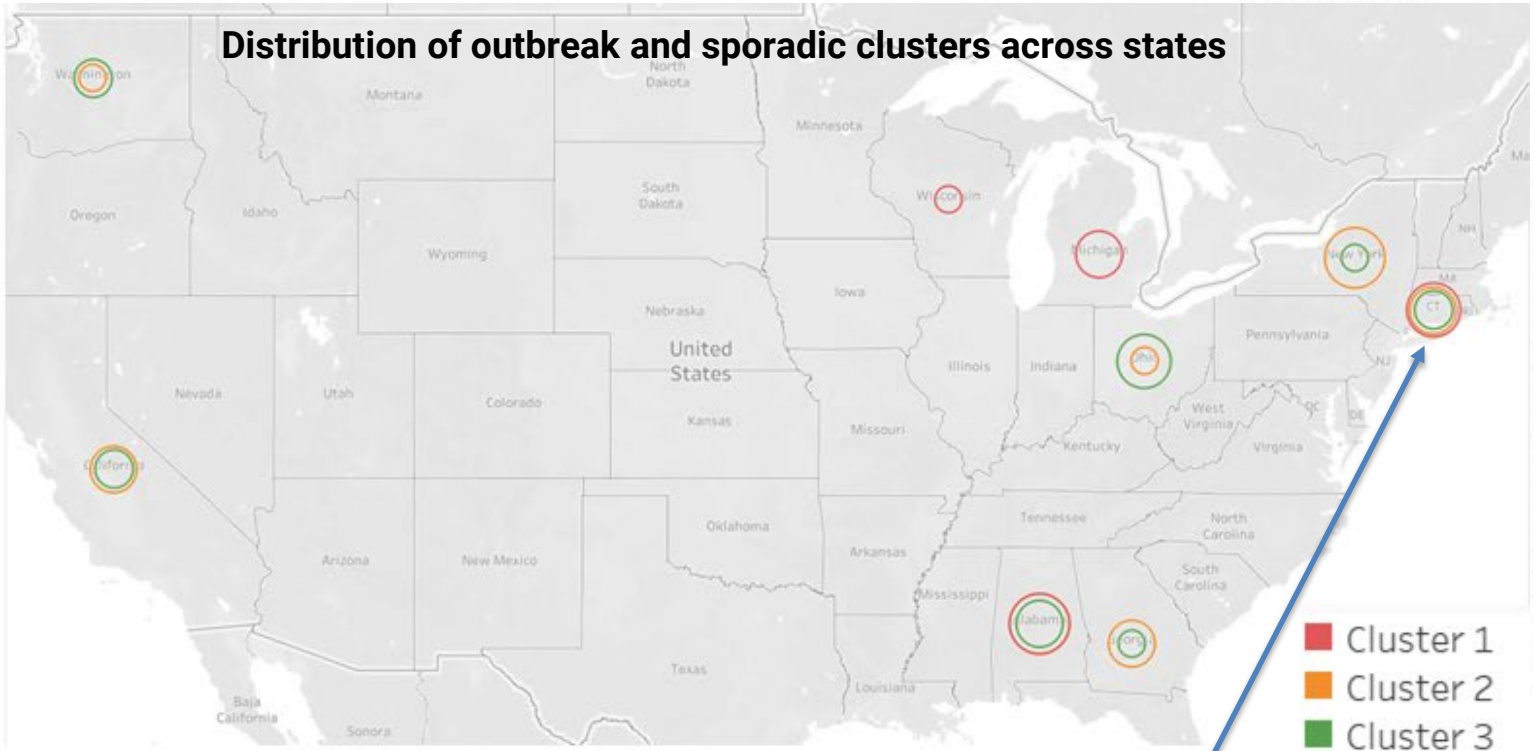
Least diversity suggests close association of source. It's the outbreak cluster

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

Georgia Tech
CREATING THE NEXT

Food source and Outbreak locations





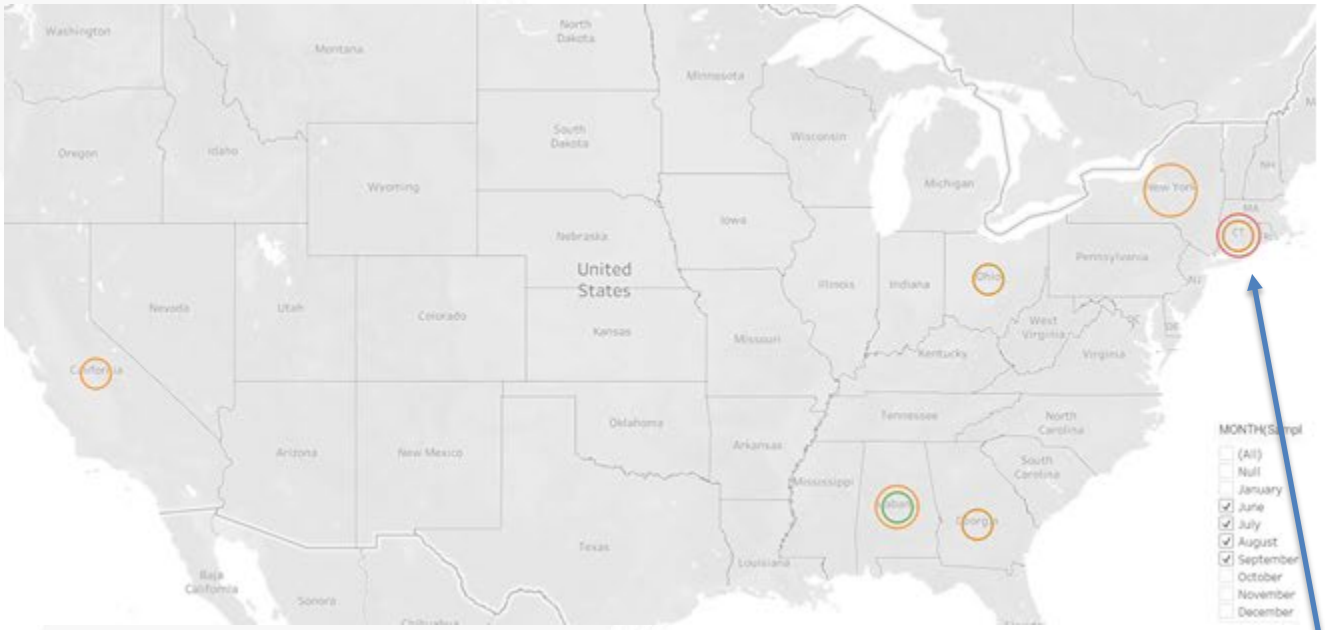
Chicken Salad fits the requirement for being the outbreak source for Listeria. (Listeria is seen mostly in ready to eat meats and people who consumed chicken salad were exclusively from the outbreak cluster)

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

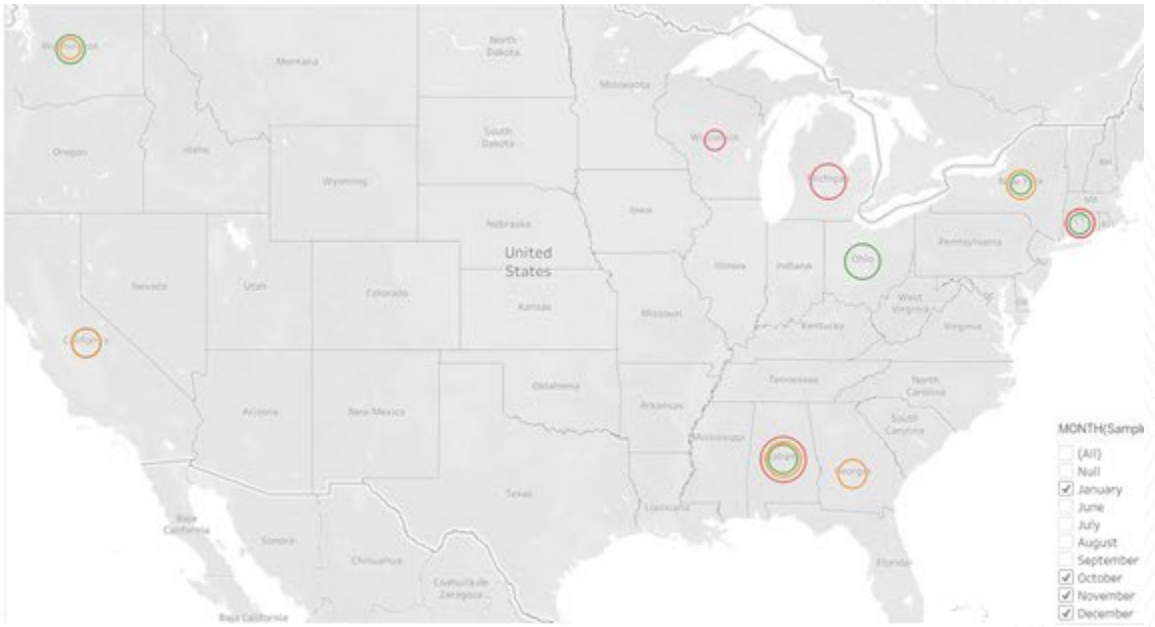
Cluster 1 - Outbreak Cluster 2 - Similar to outbreak Cluster 3 - Different than outbreak

Georgia Tech
CREATING THE NEXT

Timeline and source of Outbreak



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

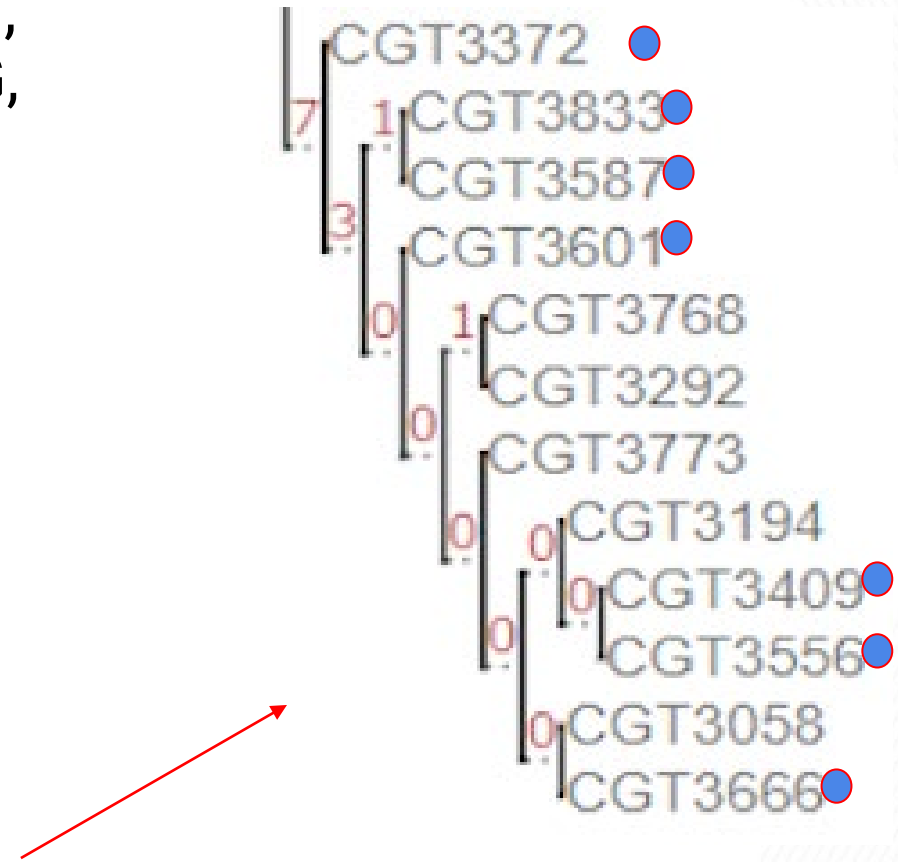
Georgia Tech
CREATING THE NEXT

Outbreak Analysis - VFDB

- 35 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

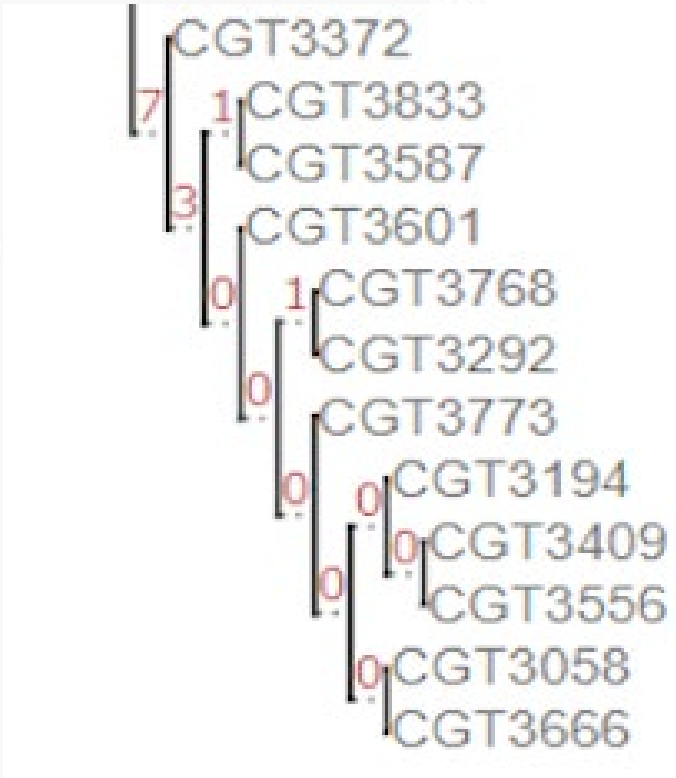
- 3 genes absent in outbreak group but present in other isolates- llsP, gtcA, aut

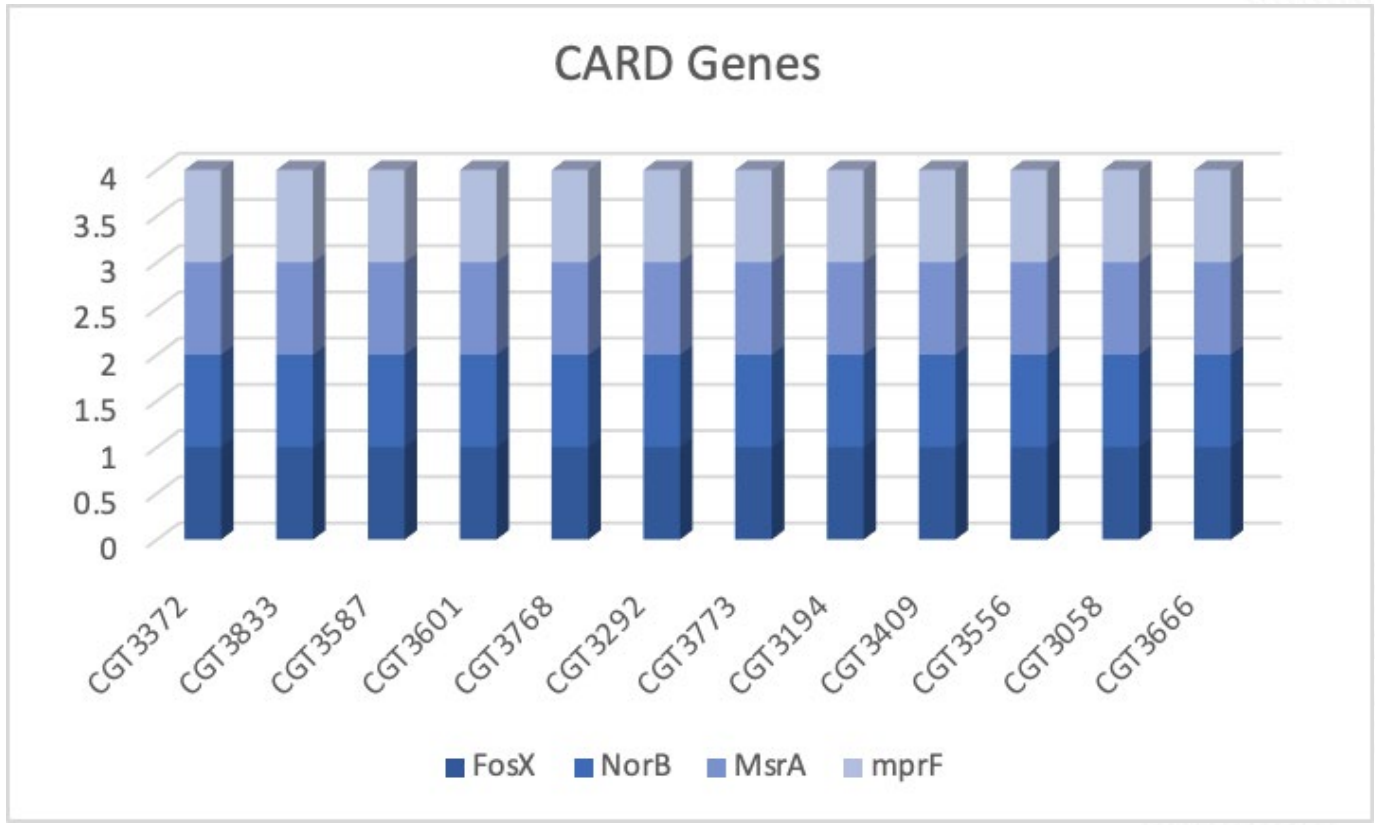
- plasmid analysis of VFDB gave lplA1 gene associated with plasmid.



Outbreak Analysis - CARD gff

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





Antibiotic resistance

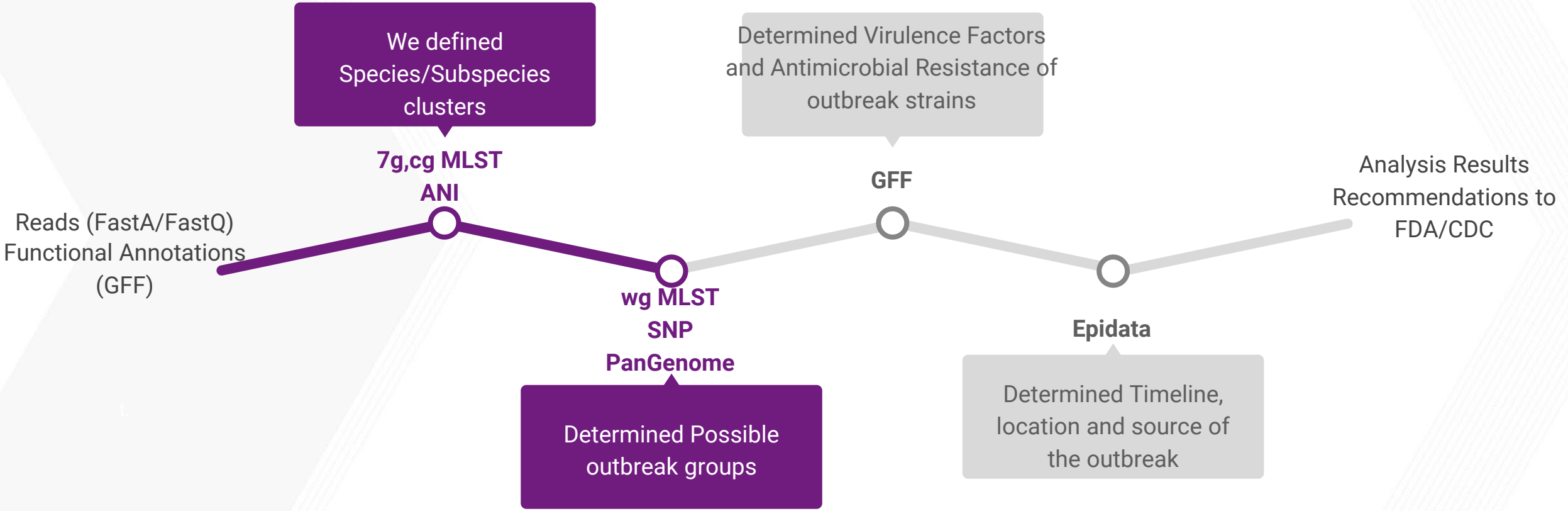
Database	Gene	Present on	Drug resistance	Resistance mechanism	AMR gene family	Drug class
CARD	FosX	Chromosome	Fosfomycin	antibiotic inactivation	fosfomycin thiol transferase	fosfomycin
CARD	msrA	plasmid or chromosome	Erythromycin and streptogramin B	antibiotic target protection	ABC-F ATP-binding cassette ribosomal protection protein	streptogramin antibiotic, tetracycline antibiotic, pleuromutilin antibiotic, macrolide antibiotic, oxazolidinone antibiotic, lincosamide antibiotic, 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*.

Comparative Genomics Final Pipeline



Georgia Tech
CREATING THE NEXT

Refernces

- 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!