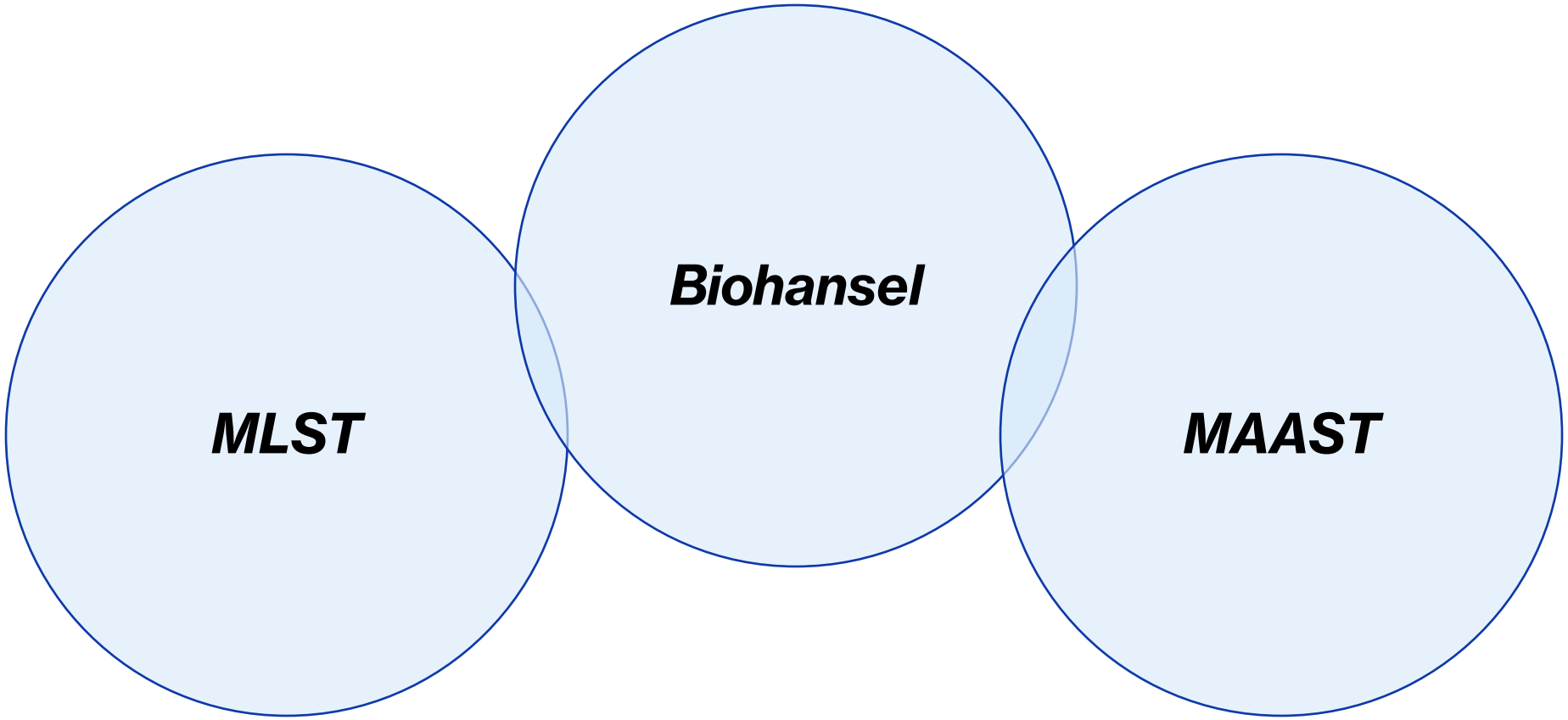
GENOTYPING & TAXONOMY RESULTS

Team F Group 3

Genotyping



MLST

- All *Listeria Monocytogenes*
  - Two different sequence types (ST's)

```
1   abcZ(3) bglA(1) cat(1)  dapE(1) dat(3)  ldh(1)  lhkA(3)
```

ST 1 (3 Samples)

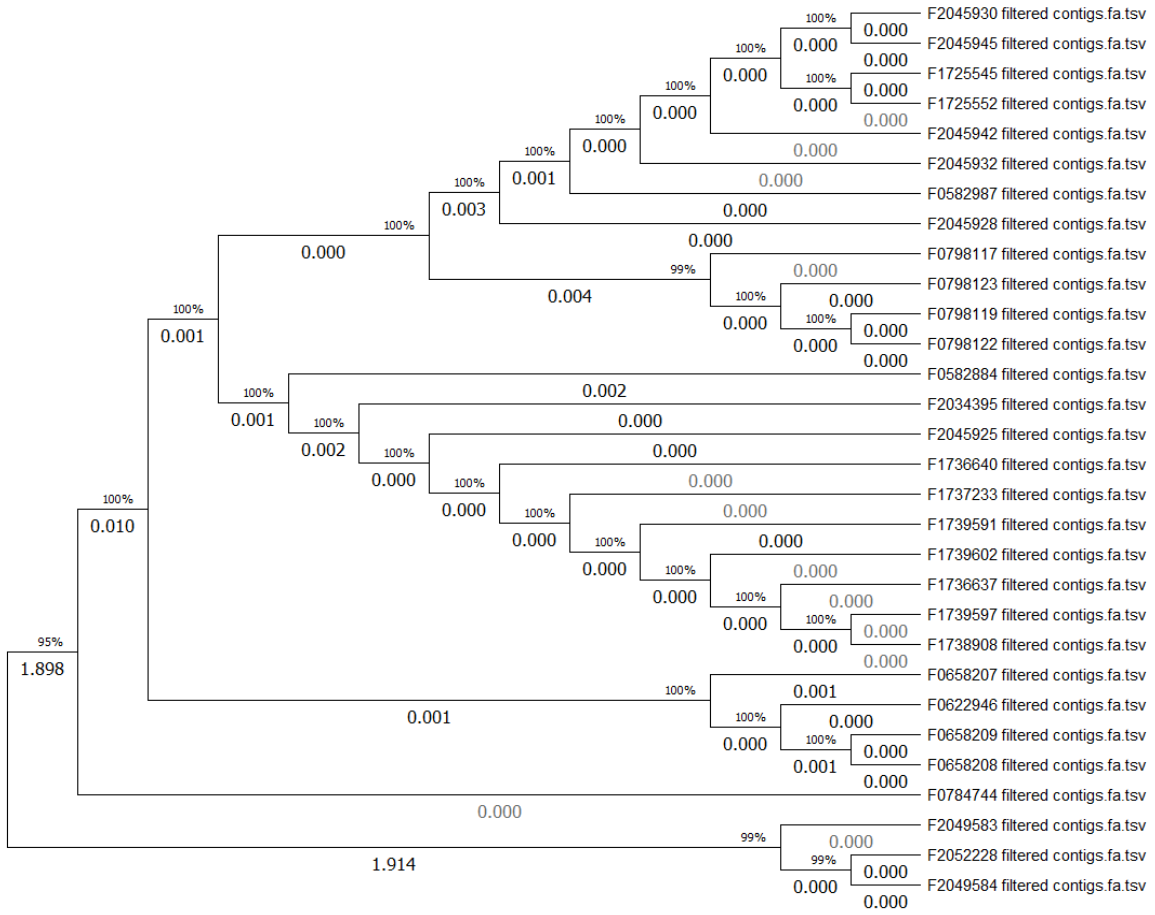
F2049583, F2049584, F2052228

```
31  abcZ(7) bglA(14)    cat(10) dapE(19)    dat(9)  ldh(8)  lhkA(1)
```

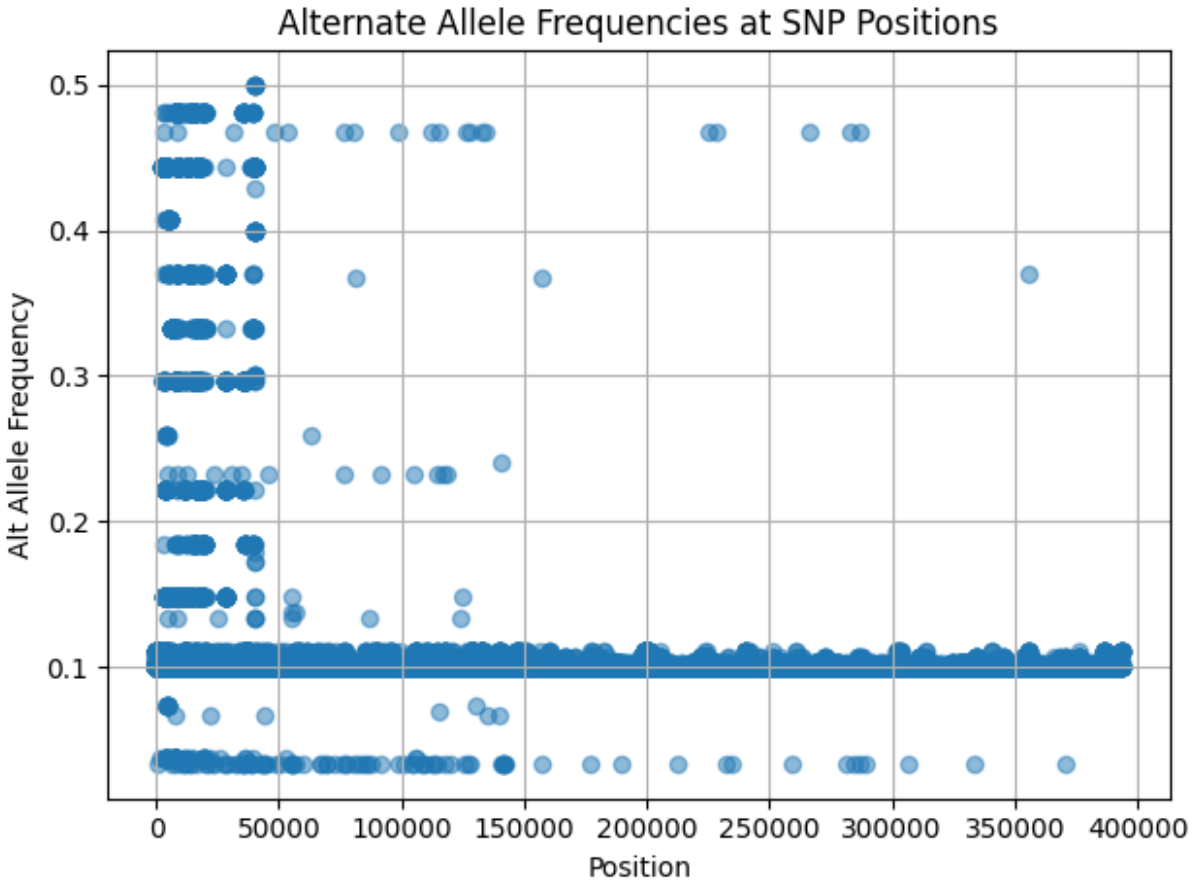
ST 31

(27 samples)

MAAST



Phylogenetic tree depicting genomic relationships of 30 bacterial strains based on SNPs



Scatter Plot depicting the distribution of alternate allele frequencies at SNP positions.

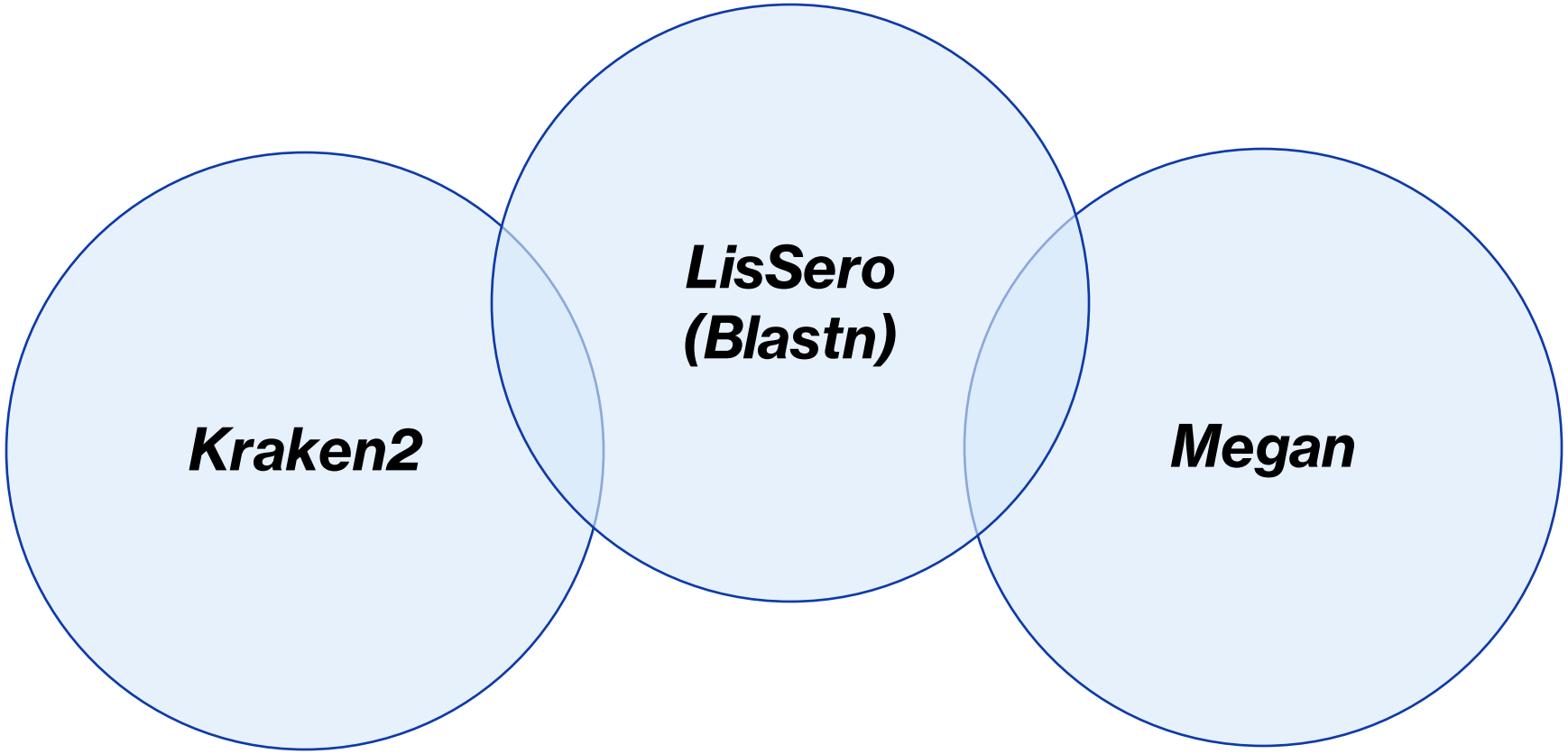
BioHansel

```
BioHansel version 2.5.1: Subtype microbial genomes using SNV targeting k-mer subtyping schemes.

Built-in schemes:

* heidelberg:  Salmonella enterica spp. enterica serovar Heidelberg
* enteritidis: Salmonella enterica spp. enterica serovar Enteritidis
* typhimurium: Salmonella enterica spp. enterica serovar Typhimurium
* typhi:       Salmonella enterica spp. enterica serovar Typhi
* tb_lineage:  Mycobacterium tuberculosis
```

- Had genomes for only these 5 strains so did not yield any results when compared with our strain.

Tool	Time per file (sec)	Max RAM (kb)
MLST	1.16	186788
MAAST	15.08	3365516

Taxonomic Classification

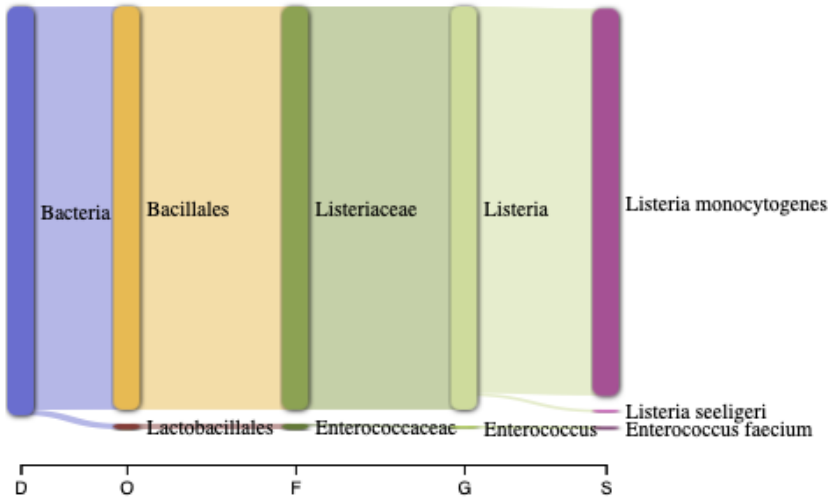


Kraken2

- All samples are classified as Listeria monocytogenes
- Use Standard Kraken 2 database (Refeq archaea, bacteria, viral, plasmid, human1, UniVec_Core)
- For install database 100 GB of disk space required and ~ 4 hours 30 min with 128 threads
- Use KrakenTools and Pavian for summary and Sankey plot the results

#perc	tot_all	type	taxid	name
0.028	1	U	0	unclassified
99.972	3567	R	1	root
99.9439	3566	R1	131567	cellular organisms
99.9439	3566	D	2	Bacteria
99.9159	3565	D1	1783272	Terrabacteria group
99.9159	3565	P	1239	Bacillota
99.9159	3565	C	91061	Bacilli
97.8139	3490	O	1385	Bacillales
97.4215	3476	F	186820	Listeriaceae
97.4215	3476	G	1637	Listeria
89.1536	3181	S	1639	Listeria monocytogenes

Summary for all samples



Sankey plot of F0582987

LisSero & Blastn

LisSero: *in silico* serogrouping (presence of surface antigen) for *Listeria monocytogenes* using PCR

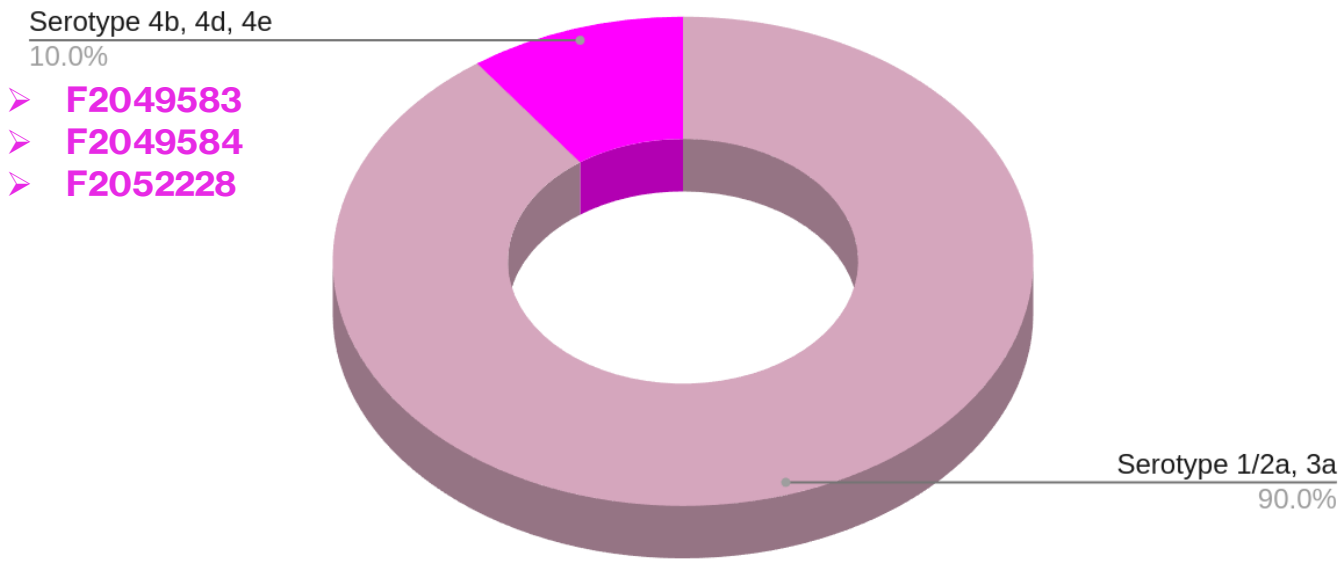
Detects: presence/absence of 5 genes (lmo1118, lmo0737, ORF2110, ORF2819, Prs)

- 27/30 samples – Serotype 1/2a, 3a
- 3/30 samples – Serotype 4b, 4d, 4e

Default min coverage: 95%

BLAST Reference genomes: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4733179/

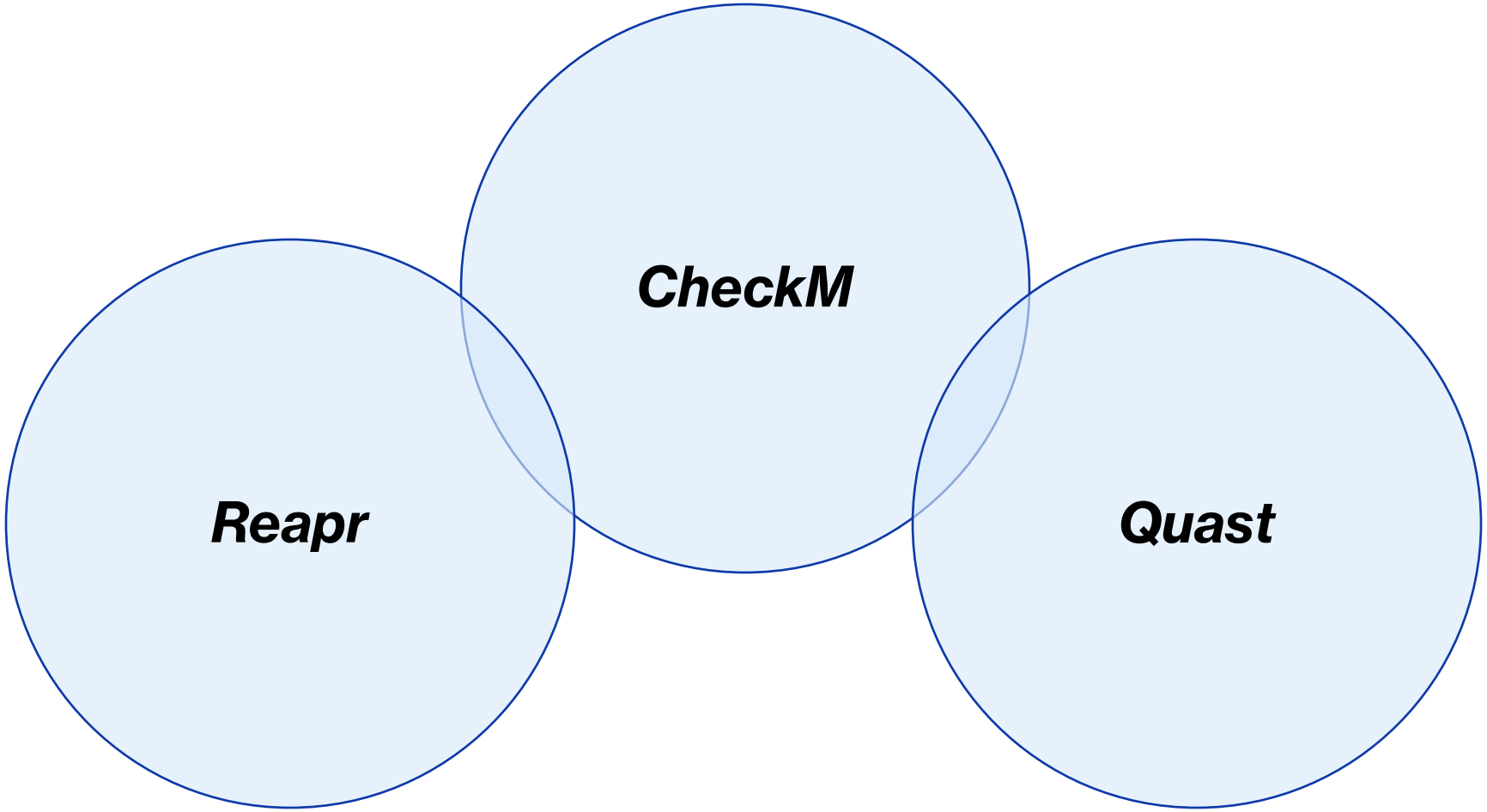
*Most outbreaks of human disease are serotypes 1/2a, 1/2b, 4b



Blastn	Min percent Identity (%)	Average E-value
Serotype 1/2a	71.154%	1.183 e -06
Serotype 3a	69.973%	9.46 e –07
Serotype 4b	69.998%	1.345 e –08
Serotype 4d	69.057%	2.807 e –07
Serotype 4e	68.998%	1.335 e -08

Tool	Time per file (sec)	Max RAM (kb)
Kraken	88	7,852,476
LisSero	2.56	191,766
BlastN	2.13	188,365

Genome Quality Assessment



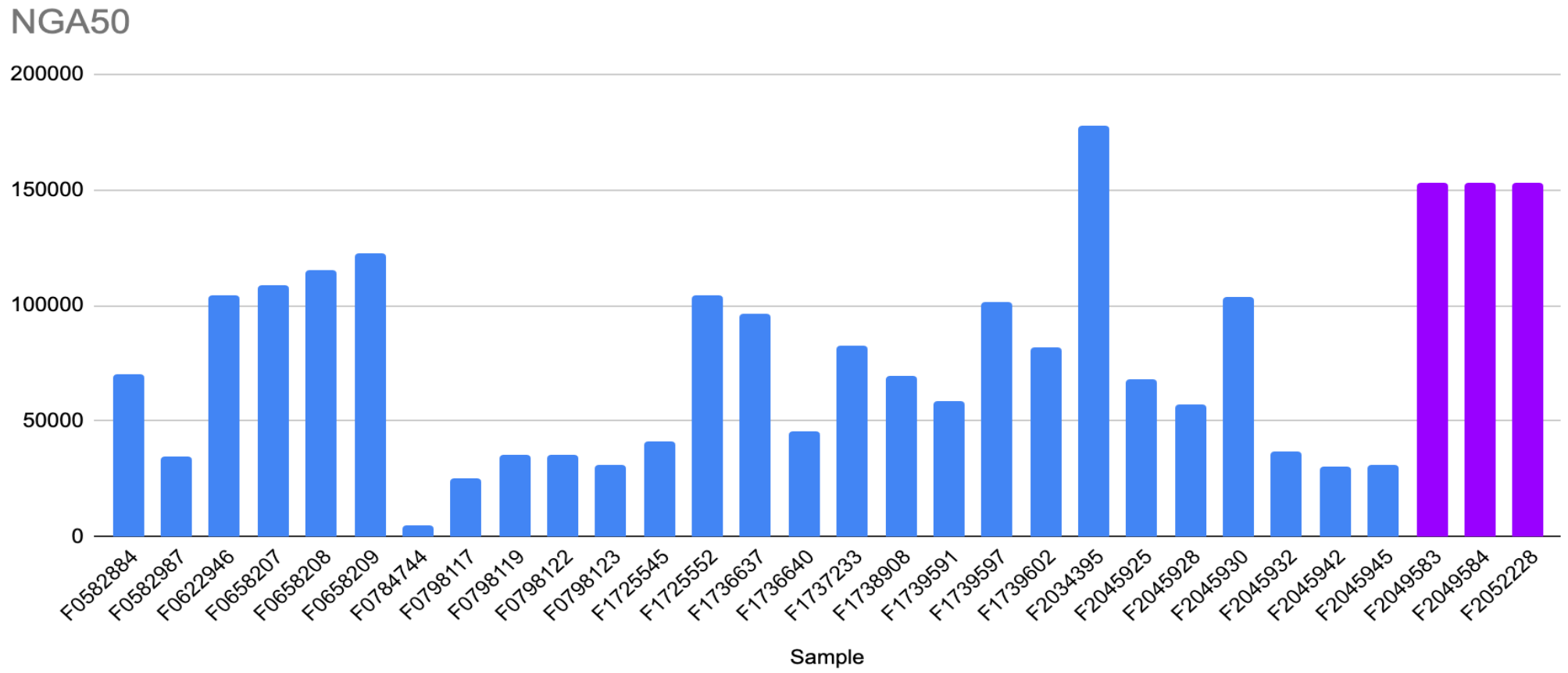
CheckM

Lowest completeness: 98.5, highest contamination: 1.27

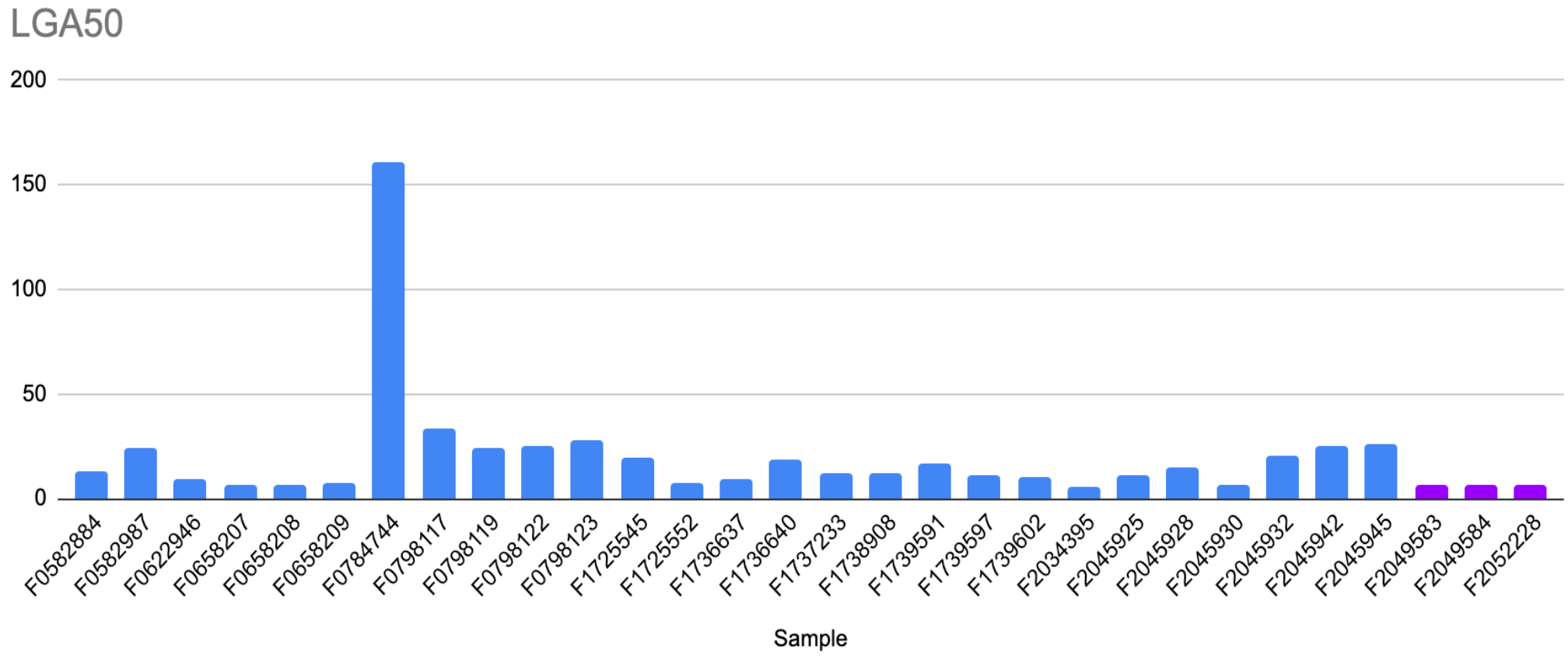
Average Completeness	Average Contamination
99.155	0.821

Bin Id	Marker lineage	# genomes	# markers	# marker sets	0	1	2	3	4	5+	Completeness	Contamination	Strain heterogeneity
F1725552	Listeria monocytogenes (6)	20	1262	179	5	1247	9	1	0	0	99.34	0.93	0
F2052228	Listeria monocytogenes (6)	20	1262	179	5	1254	3	0	0	0	99.27	0.32	0
F2049584	Listeria monocytogenes (6)	20	1262	179	5	1254	3	0	0	0	99.27	0.32	0
F2049583	Listeria monocytogenes (6)	20	1262	179	5	1254	3	0	0	0	99.27	0.32	0

QUAST
with reference genome

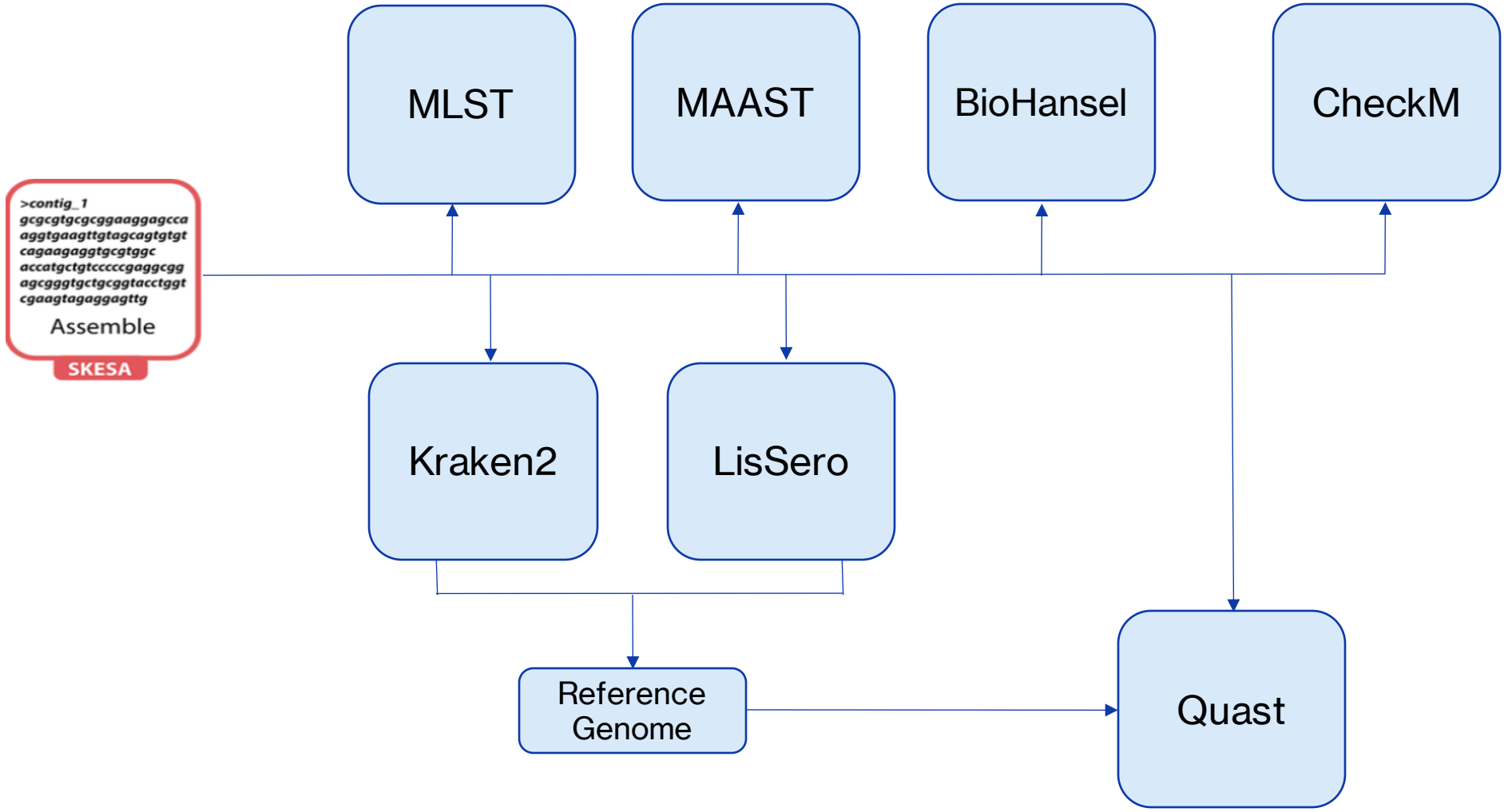


QUAST
with reference genome



Tool	Time per file (sec)	Max RAM (kb)
CheckM	7	320840
Quast	~12	149,168

Final Pipeline



Sample of Pipeline Script

```
conda install -c bioconda mummer4
conda install -c bioconda pigz lbzip2 lz4
conda install -c bioconda fasttree
conda deactivate maast

#creation of checkm environment
conda create -n checkm -c conda-forge -c bioconda checkm-genome -y


### RUN COMMANDS

#MLST COMMAND

conda activate mlst
mkdir ./mlst
#copies raw data into form suitable for MLST
cp -r "$input_directory"/* ./mlst/
cd mlst
#Run mlst
mlst *.fna > MLST_Summary.tsv
cd ..
conda deactivate


#MAAST COMMANDS
conda activate maast
#Copied a copy of Maast to my local computer
git clone https://github.com/zjshi/Maast.git
cd Maast # Navigating to directory where maast is present
make #This compiles the source code of maast
chmod 755 maast #to make GT-Pro ready to execute
#Maast command for genotyping

mkdir ./maast_output
maast end_to_end --in-dir $input_directory --out-dir ./maast_output --min-pre
#the step generates a list of input pairs.
paste <(find ./maast_output/gt_results/ -name '*.tsv' | sort) <(find ./maast_
./Maast tree --input-list ./genotypes.input.tsv --out-dir ./tree_results/ #Pl
```

Citations

- Wood, D.E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 257 (2019). https://doi.org/10.1186/s13059-019-1891-0
- Parks, Donovan H., et al. "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes." Genome research 25.7 (2015): 1043-1055.
- Shi, Z.J., Nayfach, S. & Pollard, K.S. Maast: genotyping thousands of microbial strains efficiently. Genome Biol 24, 186 (2023). https://doi.org/10.1186/s13059-023-03030-8
- Labbé, G., Kruczkiewicz, P., Robertson, J., Mabon, P., Schonfeld, J., Kein, D., Rankin, M. A., Gopez, M., Hole, D., Son, D., Knox, N., Laing, C. R., Bessonov, K., Taboada, E. N., Yoshida, C., Ziebell, K., Nichani, A., Johnson, R. P., Van Domselaar, G., & Nash, J. H. E. (2021). Rapid and accurate SNP genotyping of clonal bacterial pathogens with BioHansel. Microbial genomics, 7(9), 000651. https://doi.org/10.1099/mgen.0.000651
- "This publication made use of the PubMLST website (https://pubmlst.org/) developed by Keith Jolley (Jolley & Maiden 2010, BMC Bioinformatics, 11:595) and sited at the University of Oxford. The development of that website was funded by the Wellcome Trust".
- Seemann T, mlst Github https://github.com/tseemann/mlst
- Josh Zhang, LisSero Github https://github.com/MDU-PHL/LisSero
- Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 May 1;29(8):1072-5. doi: 10.1093/bioinformatics/btt086. Epub 2013 Feb 21. PMID: 23426934.