Project C Report:
Evaluate the genetic diversity between bacterial species, strains and/or isolate.

Dawei Wang

In this project, the dataset of 15 genomes were used to assess the sequence similarities. Their gene complements and architectures were evaluated.

First, the genomes data were downloaded with queryNCBI.pl and part of them were downloaded manually. The genomes are listed below:

Clostridium botulinum A str. Hall	NC_009698.1	(1)
Clostridium botulinum A str. ATCC 3502	NC_009495.1	
Clostridium botulinum B str. Eklund 17B (NRP)	NC_010674.1	
Clostridium botulinum BKT015925	NC_015425.1	(2)
Clostridium botulinum E3 str. Alaska E43	NC_010723.1	
Clostridium botulinum F str. Langeland	NC_009699.1	
Clostridium sporogenes NCIMB 10696	NZ_CP009225.1	(3)
Clostridium sporogenes DSM 795	NZ_CP011663.1	
Clostridium sporogenes ATCC 15579	GCA_000155085.1	(4)
Clostridium sporogenes PA 3679	GCA_000240115.1	
Clostridium sporogenes PA 3679 1990	GCA_001444575.1	
Clostridium tetani E88	NC_004557.1	(5)
Clostridium tetani 12124569	NC_022777.1	(6)
Clostridium perfringens ATCC 13124	NC_008261.1	(7)
Clostridium perfringens SM101	NC_008262.1	(8)

The number (1) listed above on the right are used as sample numbers for ANI and dDDH and further process.

To investigate the Average Nucleotide Identity (ANI) , the EZBioCloud was used (https://www.ezbiocloud.net/tools/ani)

For digital DNA-DNA hybridization (dDDH) evaluation, Genome-to-Genome Distance Calculator 2.1 was used. (https://ggdc.dsmz.de/ggdc.php#) The data used here is Formula: 2(identities / HSP length)

8 samples were used to calculate identity or distance and the results listed below:

	OrthoANIu value (%)	**Distance by GGDC**	**DDH estimate (GLM-based):**
Sample 2 – Sample 1 (similarly hereinafter)	72.19	0.2007	21.90% [19.6 - 24.3%]
2-3	72.38	0.2012	21.80% [19.5 - 24.2%]
2-6	71.91	0.2099	20.90% [18.7 - 23.3%]
2-7	71.38	0.2017	21.70% [19.5 - 24.2%]
3-1	92.15	0.0782	47.30% [44.7 - 49.9%]
3-4	99.28	0.0078	94.00% [92.2 - 95.4%]
3-5	73.36	0.1950	22.50% [20.2 - 24.9%]
3-8	72.52	0.1763	24.70% [22.4 - 27.2%]
5-1	73.56	0.2052	21.40% [19.1 - 23.8%]
5-4	73.34	0.1987	22.10% [19.8 - 24.5%]
5-6	96.50	0.0347	71.00% [68 - 73.8%]
5-7	70.98	0.2120	20.70% [18.5 - 23.1%]
8-1	71.73	0.1842	23.70% [21.4 - 26.2%]
8-4	70.95	0.1918	22.80% [20.5 - 25.3%]
8-5	71.00	0.1999	21.90% [19.7 - 24.4%]
8-7	96.93	0.0305	74.20% [71.2 - 77%]

Table.1 ANI and DDH results. These shows a general idea identify and distance between selected genomes. The distance and identity obtained by these two methods are mutually confirmed.

Then mash was used for all the genomes by using run_Mash.pl. Then use MashToDistanceCSV.pl and MashR_plotter.pl to generate heatmap and phylogenic tree.

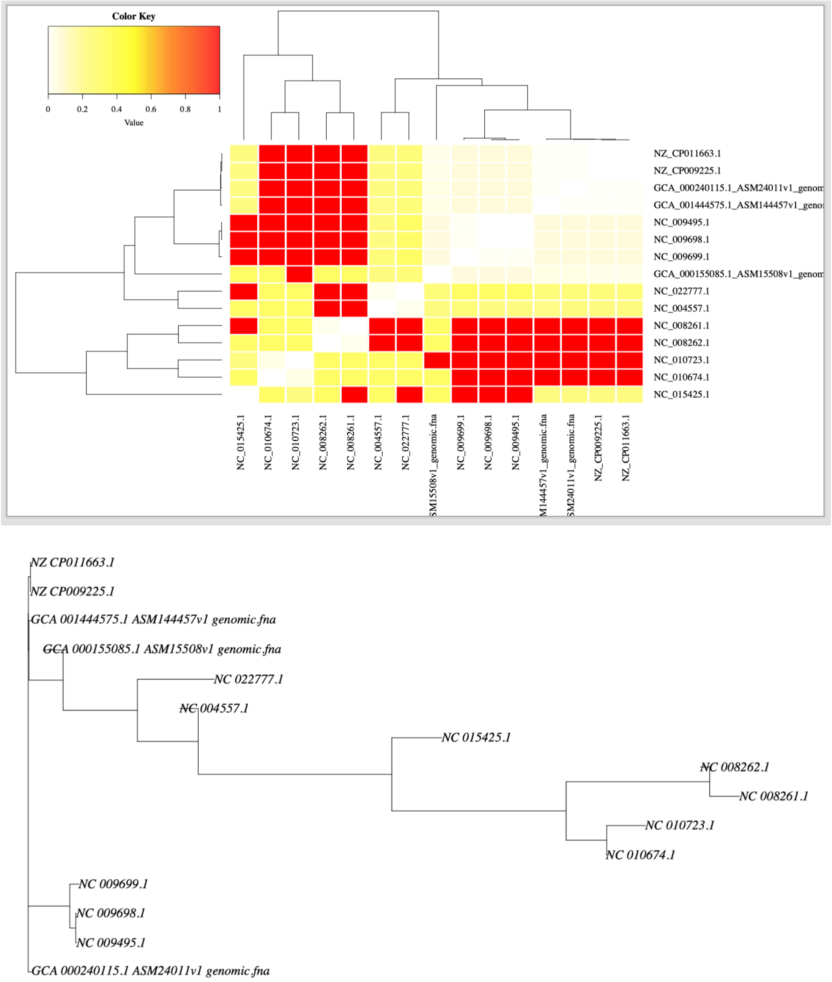
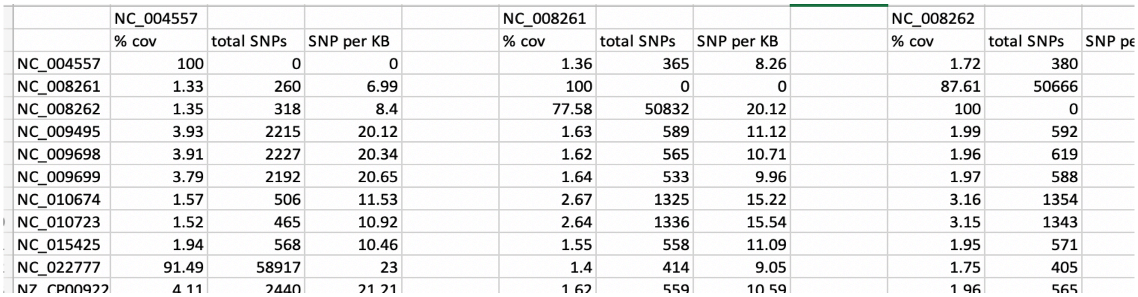


Fig.1 Heatmap and phylogenic tree. From these figures, the distances between different genomes.

To call the variants, bowtie2 was used. But first, SSRG.pl was used to break the complete genomes from FASTA file and generate sets of FASTQ files. In this case, the read size was selected as 100bps, and Paired ends fastq files for each genome in the dataset was created.

After synthetic reads created, get_SNPs.pl was helping to call the variants by using bowtie2. To get a better view of all the variants, sort_stats.pl was used to create a tab-delimited (TSV) file.



	NC_004557				NC_008261				NC_008262		
	% cov	total SNPs	SNP per KB		% cov	total SNPs	SNP per KB		% cov	total SNPs	SNP pe
NC_004557	100	0	0		1.36	365	8.26		1.72	380	
NC_008261	1.33	260	6.99		100	0	0		87.61	50666	
NC_008262	1.35	318	8.4		77.58	50832	20.12		100	0	
NC_009495	3.93	2215	20.12		1.63	589	11.12		1.99	592	
NC_009698	3.91	2227	20.34		1.62	565	10.71		1.96	619	
NC_009699	3.79	2192	20.65		1.64	533	9.96		1.97	588	
NC_010674	1.57	506	11.53		2.67	1325	15.22		3.16	1354	
NC_010723	1.52	465	10.92		2.64	1336	15.54		3.15	1343	
NC_015425	1.94	568	10.46		1.55	558	11.09		1.95	571	
NC_022777	91.49	58917	23		1.4	414	9.05		1.75	405	
NZ_CP00922	4.11	2440	21.21		1.62	559	10.59		1.96	565	

Fig.2 Part of the tsv file.

Kmer-based calling variants method (kestrel) was used to call part of genomes tentatively.

	Variants called by Kestrel	Variants called by Bowtie2
Sample 2 - Sample 1	3905	1073
Sample 2 - Sample 3	3350	1008
Sample 2 - Sample 6	2952	619
Sample 5 - Sample 1	3115	2228
Sample 5 - Sample 6	67736	58918
Sample 5 - Sample 7	2121	261

Table.2 Variants called by kestrel and bowtie2. Each amount of variants called by kestrel are significantly higher than traditional way. As it mentioned in Audano and colleagues' publication, Kestrel works fine with only specific regions of genome. The limitation of kestrel is read context lost during process. It appears that results from bowtie2 is reliable in this case.

Here the minimap2 alignment was tried to call variants. Two individual genomes were used for the alignment manually, However, the last step of bcftools call wasn't success and the vcf file generated was empty. This step was just for satisfying curiosity.

panX was used for gene complements assessment. The procedure of installing and testing was followed strictly. It worked for the Test data.
panX visualization package was also installed successfully. And by testing by firefox on Mozart, it works appropriately.

Unfortunately, error reminder appears when the result was tried to send to visualization parts. The result dataset was checked and partly empty. It shows that panX didn't work as expected. By trying to use different folder and even file format, checking installation/working environment and re-run, it still not working.

To evaluate gene collinearity, progressiveMauve was applied for the whole database. The Mauve graph was created as well as phylogenic tree.

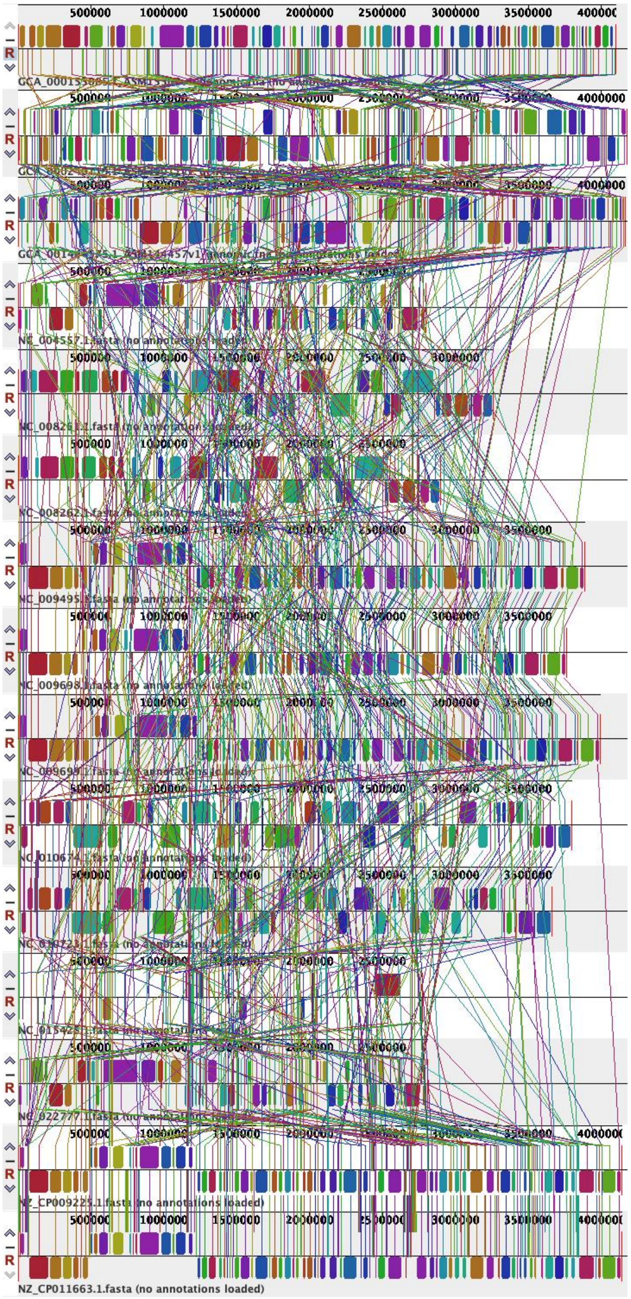

500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
R
NC_004557.1.fasta (no annotations loaded)
NC_009495.1.fasta (no annotations loaded)
NC_009698.1.fasta (no annotations loaded)
NC_010674.1.fasta (no annotations loaded)
NC_022777.1.fasta (no annotations loaded)
NZ_CP009225.1.fasta (no annotations loaded)
NZ_CP011663.1.fasta (no annotations loaded)

Fig.3 Mauve graph. This graph shows gene collinearity between all 15 genomes. Each same color represents same conserved gene and the line linked them together.

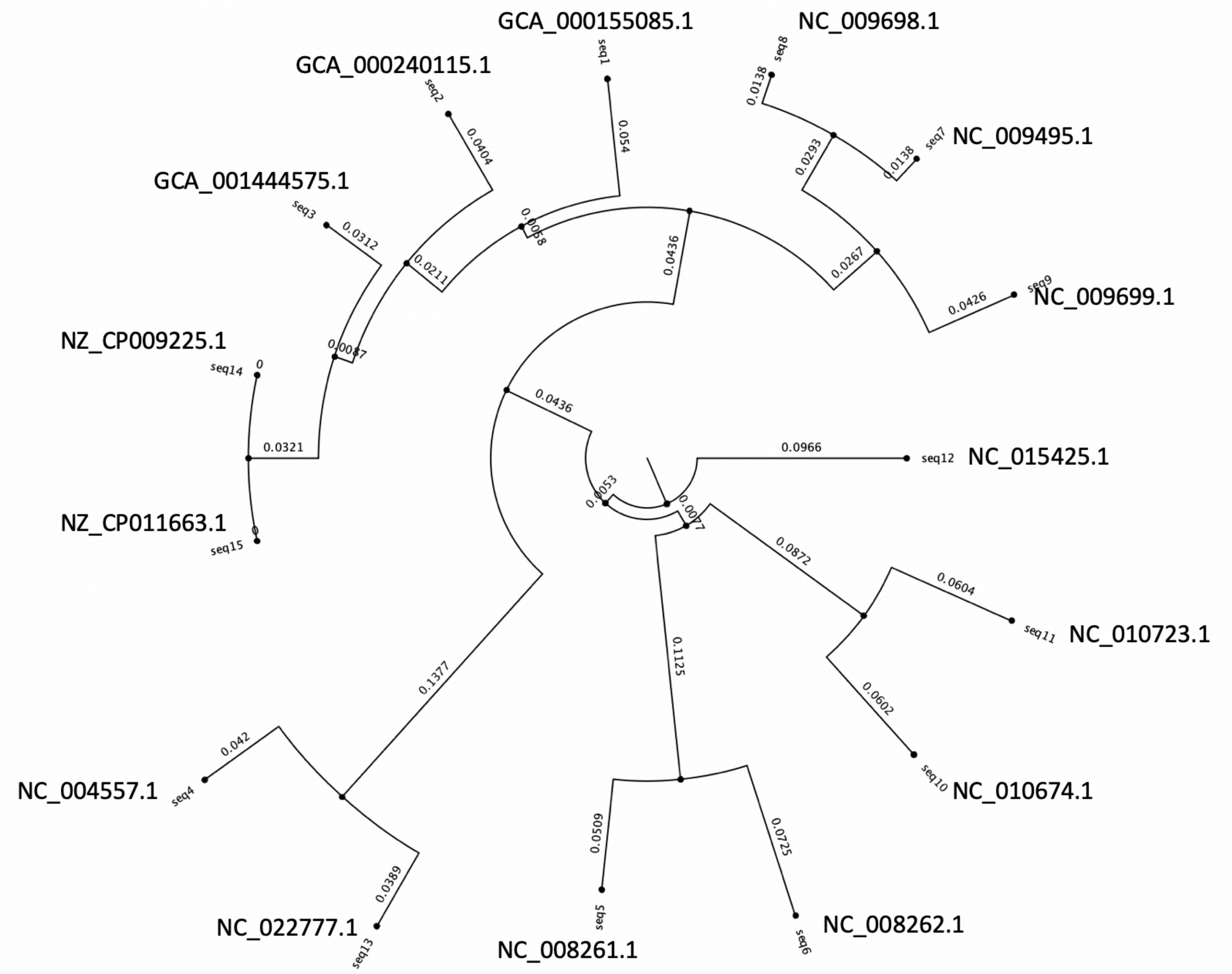


Fig.4 Phylogenic tree by mauve (Polar tree layout)

Mega 7 was also used to help analyze the dataset. Part of functions such as phylogenic tree was successfully realized. However, it requires more background knowledge to run the other functions.