Lab report 2

Sequence, assemble and annotate a bacterial genome from illumina data.

In this project, DNA sample will be extracted from an unknown bacteria sample. Then it will be sequenced by using MiniSeq short read platform from Illumina and assembled by different approaches. By analyzing data, the bacterium should be identified, and their genome will be further annotated.

<u>DNA Purification, Quantitation and Quality Assessment</u>

Sample C was provided and thawed for this step. By using lysozyme and proteinase K, sample cells lysis was proceeded.

AccuGreen High Sensitivity ds DNA Quantitation Kit was used for DNA quantitation. The procedure was followed, and sample was tested by Qubit. The concentration of DNA extract sample is 51.0 ng/μL.

The Nanodrop 2000c was used for spectrophotometry assessment. A260/280 is 1.83, which generally indicates pure DNA sample. A260/230 is 2.17, which is expected for pure nucleic acid. This shows the DNA sample extracted from original sample C could consider as pure.

0.8% agarose gel was made for electrophoresis. After electrophoresis, a clear band shows above 10,000bp. (Fig.1 on the right) Lane 1: 1Kb DNA Ladder RTU(GeneDireX) Lane 2: DNA extraction sample (4μL loaded). There's a huge mess on the bottom of the lane 2, it might because of some DNA fragments generated during extraction procedure. The band above 10k bp shows DNA extracted from original sample. The band is not showing the regular shape may because of excessive sample was added. The artificial reason such as comb removal may bend

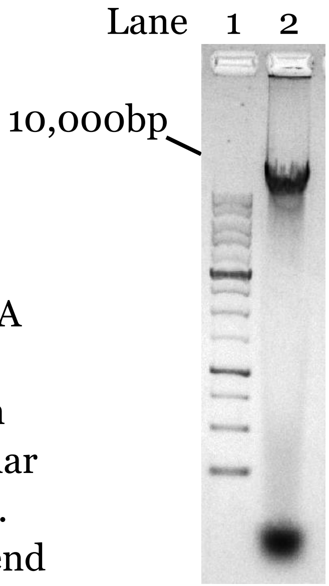


Fig.1

the well. It may lead to the band adjoin to well of lane 2 and lane 1 (clear under different exposure).

DNA library preparation

For whole genome sequencing, DNA library needs to be prepared by enzymatic reaction. Nextera DNA flex library prep kit was used to cut the DNA by transposome randomly. 50ng (1µL) DNA was used for tagmentation. After clean-up the tagmentation, 5 µL i5, H506 and 5 µL i7, H707 adapters were used in 6 cycles PCR amplification of tagmented DNA.

After PCR and libraries cleaning, the sample was tested by Qubit and it shows 14.0 ng/µL. Electrophoresis of the sample was running by 0.8% agarose gel.

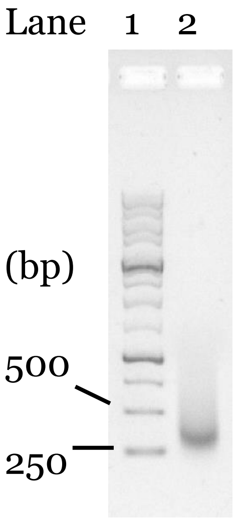


Fig.2 Lane1 1Kb DNA Ladder RTU(GeneDireX) Lane 2: DNA library sample (10µL loaded). A band between 250 and 500bp shows in lane 2. The band is little blur with a tail, this might because of degradation of DNA sample.

Library size, Molarities calculation and Pooling library

On gel, the band is between 250 and 500bp. By measuring distances between three bands, the size of DNA library fragments was estimated as 310bp.

The estimated molarity is 68.34nM. Then dilute the sample to 4nM by mixing 5µL library and 81µL resuspension buffer.

Sequencing

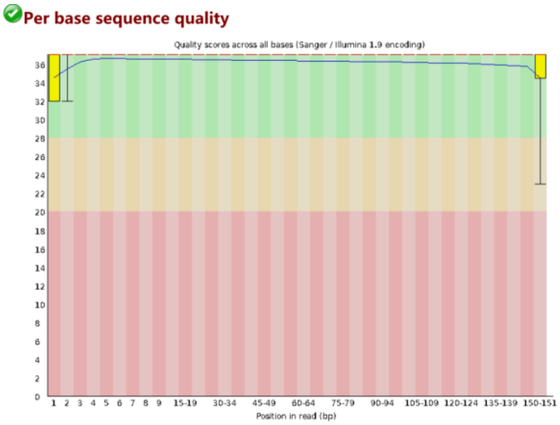
Denaturation and dilution of library was performed and PhiX was diluted and added to the library. The library was loaded on the MiniSeq and all related parameters were set up.

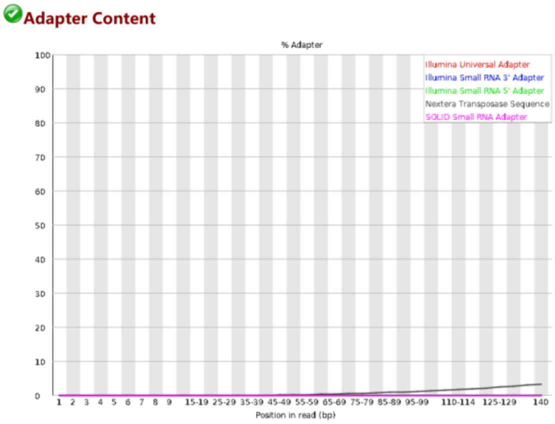
Computational analysis

A pair of paired ends fastq files were generated by MiniSeq.

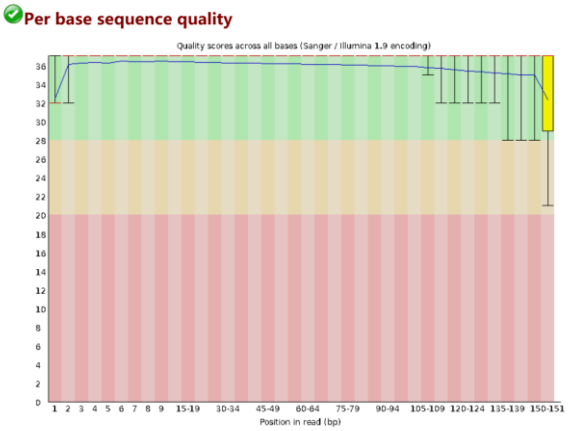
Check the quality by using fastqc

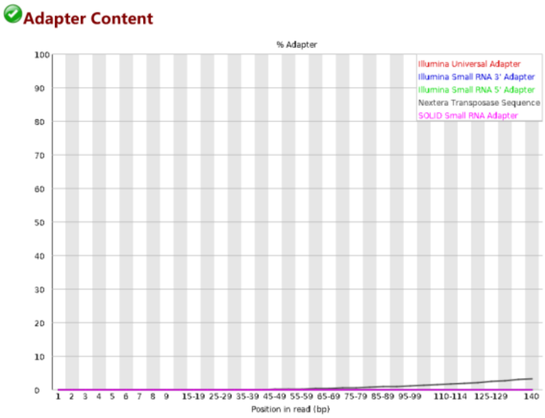
For Read 1:





For Read 2:



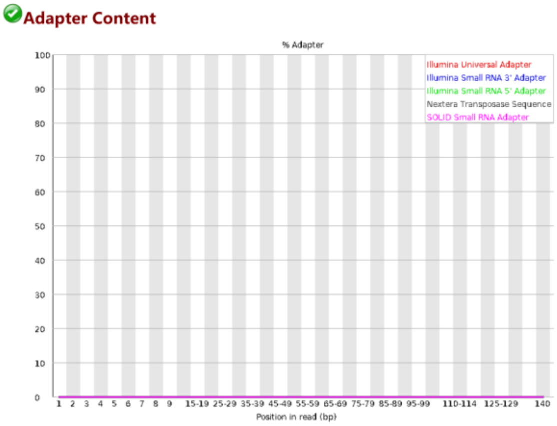


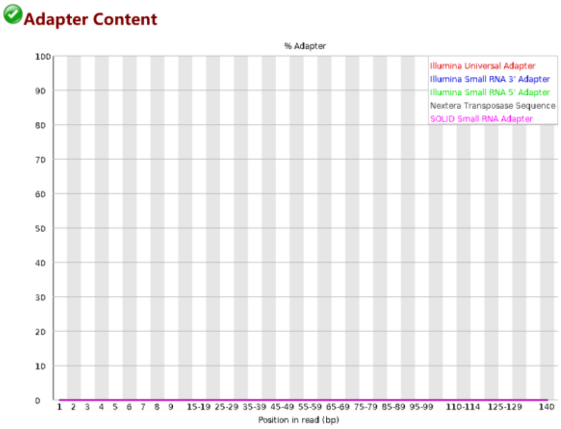
This shows for both reads, the base sequence quality is good, and they all have adapter inside.

Remove adaptor by using cutadapter then check with both reads again with fastqc.

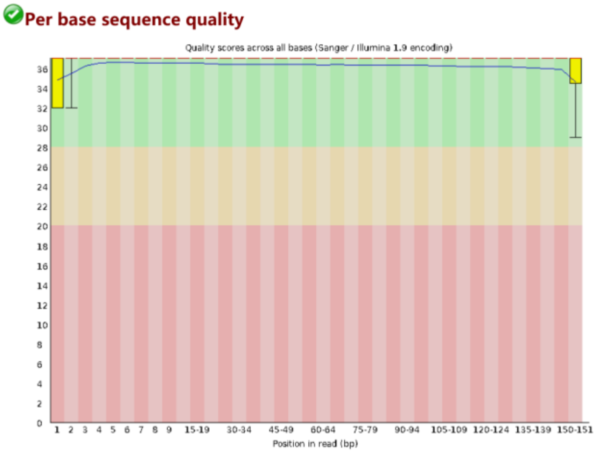
For both reads, adapter was removed after cutadapter

(The following fig showing there’s no adapter. Left: read1 Right: read2)

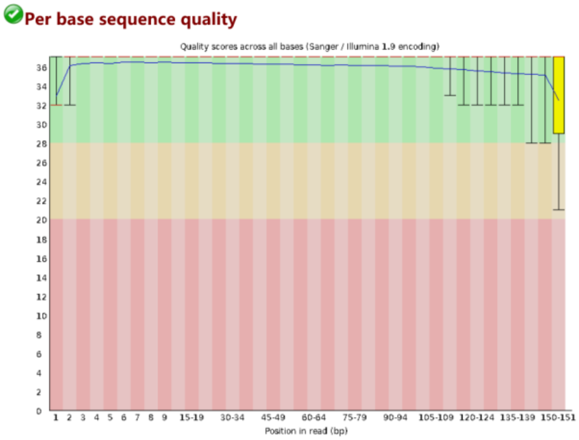




However, both reads shows: the Sequence length: 0-151.
Sickle the length under 50 bp and check with fastqc again.
Read 1:



Read 2:



Both reads are ready for further assemble after these steps.

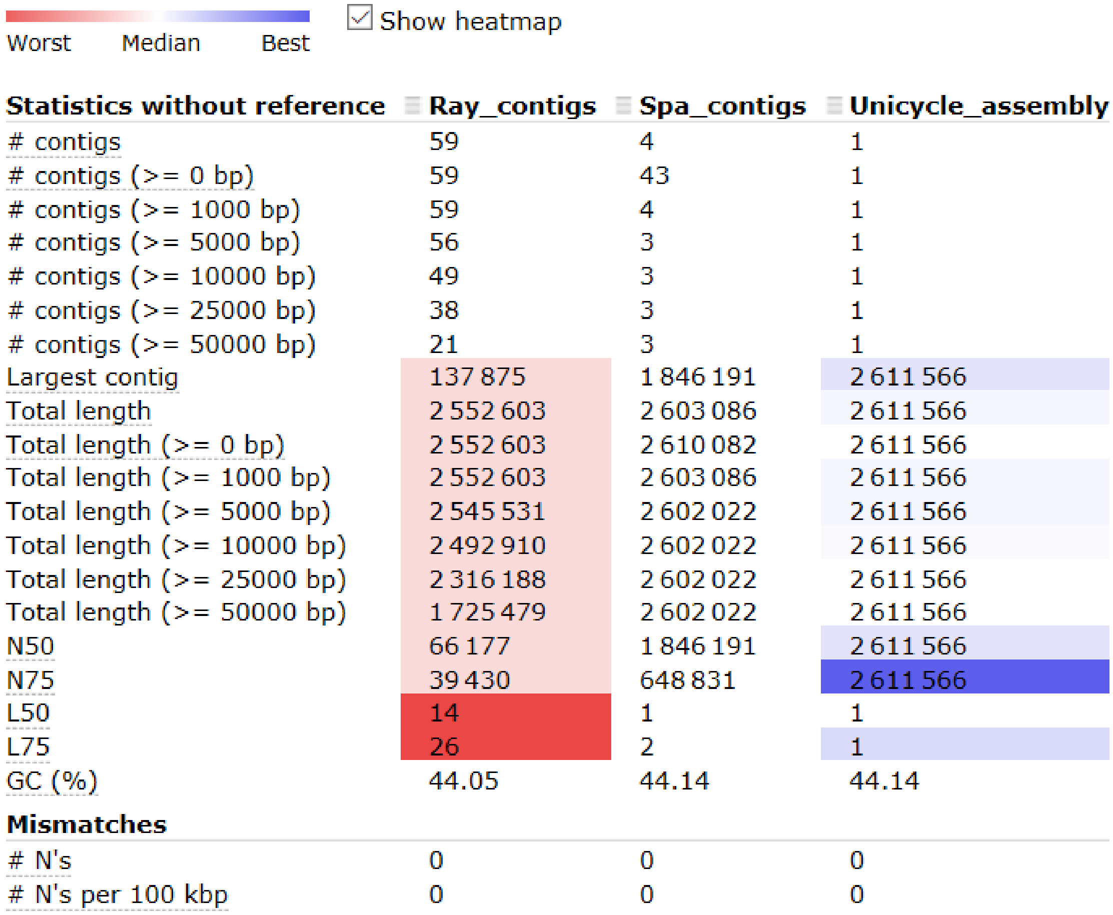
Find the best kmer by kmergenie and the best kmer is 61
De novo assembles by using Ray

```
mpirun -np 8 Ray -k 61 -p Trim_Notp_R1.fastq Trim_Notp_R2.fastq
-o Ray-61
```

Assemble with long reads by using SPAdes and Unicycler

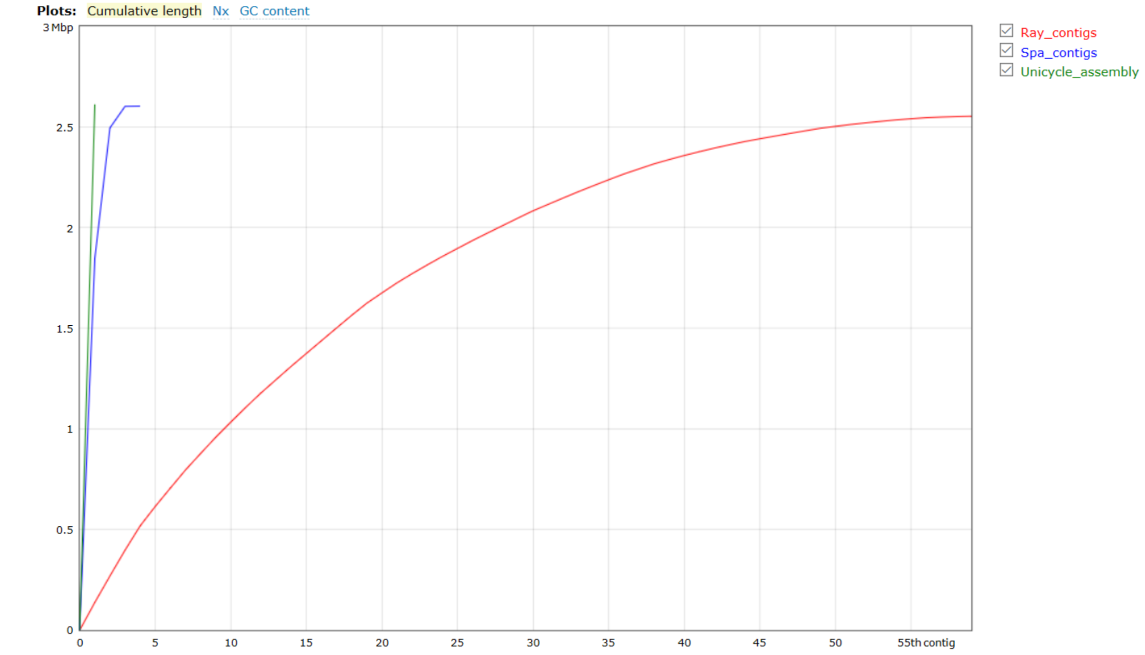
```
spades.py -k 57,61,75 --careful --pe1-1 Trim_Notp_R1.fastq --pe1-2
Trim_Notp_R2.fastq -o Spades --s1 Bacteria_C.pacbio.fastq

unicycler -t 16 -1 Trim_Notp_R1.fastq -2 Trim_Notp_R2.fastq --
pilon_path /opt/pilon/pilon-1.22.jar -l Bacteria_C.pacbio.fastq -o
Hydrid
```

Quast the three assembly and the results are listed as following:



Statistics without reference	Ray_contigs	Spa_contigs	Unicycle_assembly
# contigs	59	4	1
# contigs (>= 0 bp)	59	43	1
# contigs (>= 1000 bp)	59	4	1
# contigs (>= 5000 bp)	56	3	1
# contigs (>= 10000 bp)	49	3	1
# contigs (>= 25000 bp)	38	3	1
# contigs (>= 50000 bp)	21	3	1
Largest contig	137 875	1 846 191	2 611 566
Total length	2 552 603	2 603 086	2 611 566
Total length (>= 0 bp)	2 552 603	2 610 082	2 611 566
Total length (>= 1000 bp)	2 552 603	2 603 086	2 611 566
Total length (>= 5000 bp)	2 545 531	2 602 022	2 611 566
Total length (>= 10000 bp)	2 492 910	2 602 022	2 611 566
Total length (>= 25000 bp)	2 316 188	2 602 022	2 611 566
Total length (>= 50000 bp)	1 725 479	2 602 022	2 611 566
N50	66 177	1 846 191	2 611 566
N75	39 430	648 831	2 611 566
L50	14	1	1
L75	26	2	1
GC (%)	44.05	44.14	44.14
Mismatches			
# N's	0	0	0
# N's per 100 kbp	0	0	0

The Quast shows the Unicycle has the best result. Unicycle assemble all the data into one contig. And the contig length is 2,611,566 bp. It shows that there is no contaminant in the database and Miniseq preforms well. Ray is *de novo* assemble and the total length is very close to hybrid assembly. This shows a decent character of assembly for Ray. SPAdes use short reads and long reads to assemble. The results show that SPAdes assembled most fragments into 2 contigs and the amount of contigs is 4. It’s also shows good performance. However, compare to Unicycle, both of these assemblers are not good enough in this experiment.



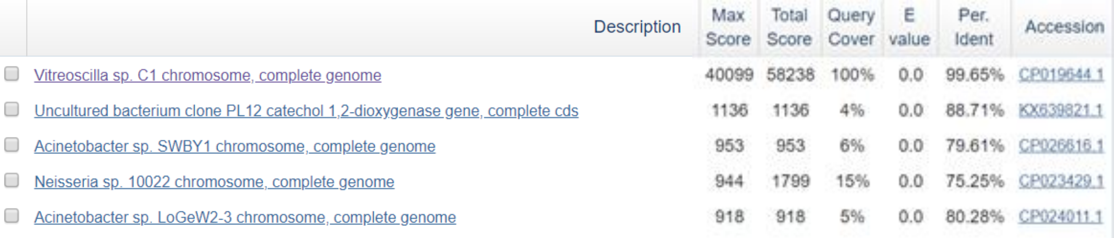
Contigs are ordered from largest (contig #1) to smallest.

Run the blastn and compare with the local mini_nt database and use it on the blast website to compare with the latest database. Set evalue as 1e-10.

For the local database, there are lots of matches, but the query coverage is low. The most frequent matches are from *Neisseria* and they might be 16 S amplicon.

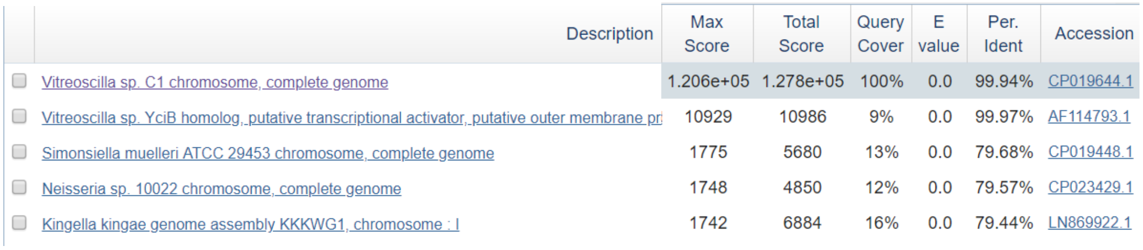
For NCBI online blast, select 3 different pieces of assembly and the results is here:

For the first fragment (21980 bp)



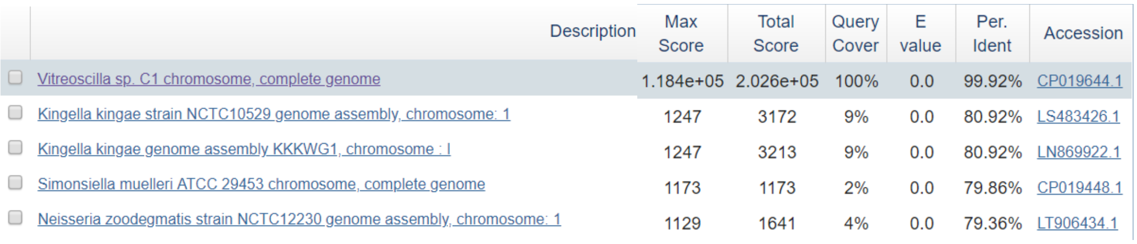
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
□	Vitreoscilla sp. C1 chromosome, complete genome	40099	58238	100%	0.0	99.65%	CP019644.1
□	Uncultured bacterium clone PL12 catechol 1,2-dioxygenase gene, complete cds	1136	1136	4%	0.0	88.71%	KX639821.1
□	Acinetobacter sp. SWBY1 chromosome, complete genome	953	953	6%	0.0	79.61%	CP026616.1
□	Neisseria sp. 10022 chromosome, complete genome	944	1799	15%	0.0	75.25%	CP023429.1
□	Acinetobacter sp. LoGeW2-3 chromosome, complete genome	918	918	5%	0.0	80.28%	CP024011.1

The second fragment (65450bp)



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
□	Vitreoscilla sp. C1 chromosome, complete genome	1.206e+05	1.278e+05	100%	0.0	99.94%	CP019644.1
□	Vitreoscilla sp. YciB homolog, putative transcriptional activator, putative outer membrane pr	10929	10986	9%	0.0	99.97%	AF114793.1
□	Simonsiella muelleri ATCC 29453 chromosome, complete genome	1775	5680	13%	0.0	79.68%	CP019448.1
□	Neisseria sp. 10022 chromosome, complete genome	1748	4850	12%	0.0	79.57%	CP023429.1
□	Kingella kingae genome assembly KKKWG1, chromosome : I	1742	6884	16%	0.0	79.44%	LN869922.1

The third fragment (64330bp)



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
□	Vitreoscilla sp. C1 chromosome, complete genome	1.184e+05	2.026e+05	100%	0.0	99.92%	CP019644.1
□	Kingella kingae strain NCTC10529 genome assembly, chromosome: 1	1247	3172	9%	0.0	80.92%	LS483426.1
□	Kingella kingae genome assembly KKKWG1, chromosome : I	1247	3213	9%	0.0	80.92%	LN869922.1
□	Simonsiella muelleri ATCC 29453 chromosome, complete genome	1173	1173	2%	0.0	79.86%	CP019448.1
□	Neisseria zoodegmatis strain NCTC12230 genome assembly, chromosome: 1	1129	1641	4%	0.0	79.36%	LT906434.1

Vitreoscilla sp. Strain C1 chromosome shows 100% query coverage and significantly high identity percentage. The length of Vitreoscilla sp. C1 chromosome in NCBI is 2610419 bp which also shares no significant differences with unicycler assembly. Thus, the sample C provided is *Vitreoscilla*. Part matches with *Neisseria* is expected since *Neisseria* and *Vitreoscilla* are from the same family – *Neisseriaceae*.

The best assembly made by Unicycler was used for the further annotation.

Use prokka and dfast to annotate the assembly. The results attached as zip files.

```
prokka --addgenes --compliant -genus vitreoscilla -gram neg --outdir prokka Unicycle_assembly.fasta

dfast -g Unicycle_assembly.fasta -o DFAST --cpu 8
```

By using grep, the amount of predicted CDS for each method are:
prokka/PROKKA_04192019.gbk:5424
DFAST/genome.gbk:2602

```
grep -c '/locus_tag=' /prokka/PROKKA_04192019.gbk /DFAST/genome.gbk
```

For prokka, the file structure has two "/locus_tag=" for each gene, so the number of genes from prokka should divide by 2. The actual amount of genes from Prokka is 2712. Despite the DFAST reference database is 20 times larger than Prokka, their prediction number are quite similar.

For protein predication, the amount from each method are:

```
prokka/PROKKA_04192019.faa:2515
DFAST/protein.faa:2501
```

```
grep -c ‘>’ /prokka/PROKKA_04192019.faa /DFAST/protein.faa
```

The amount of protein number is very close to each other, this is a good result for protein annotation.