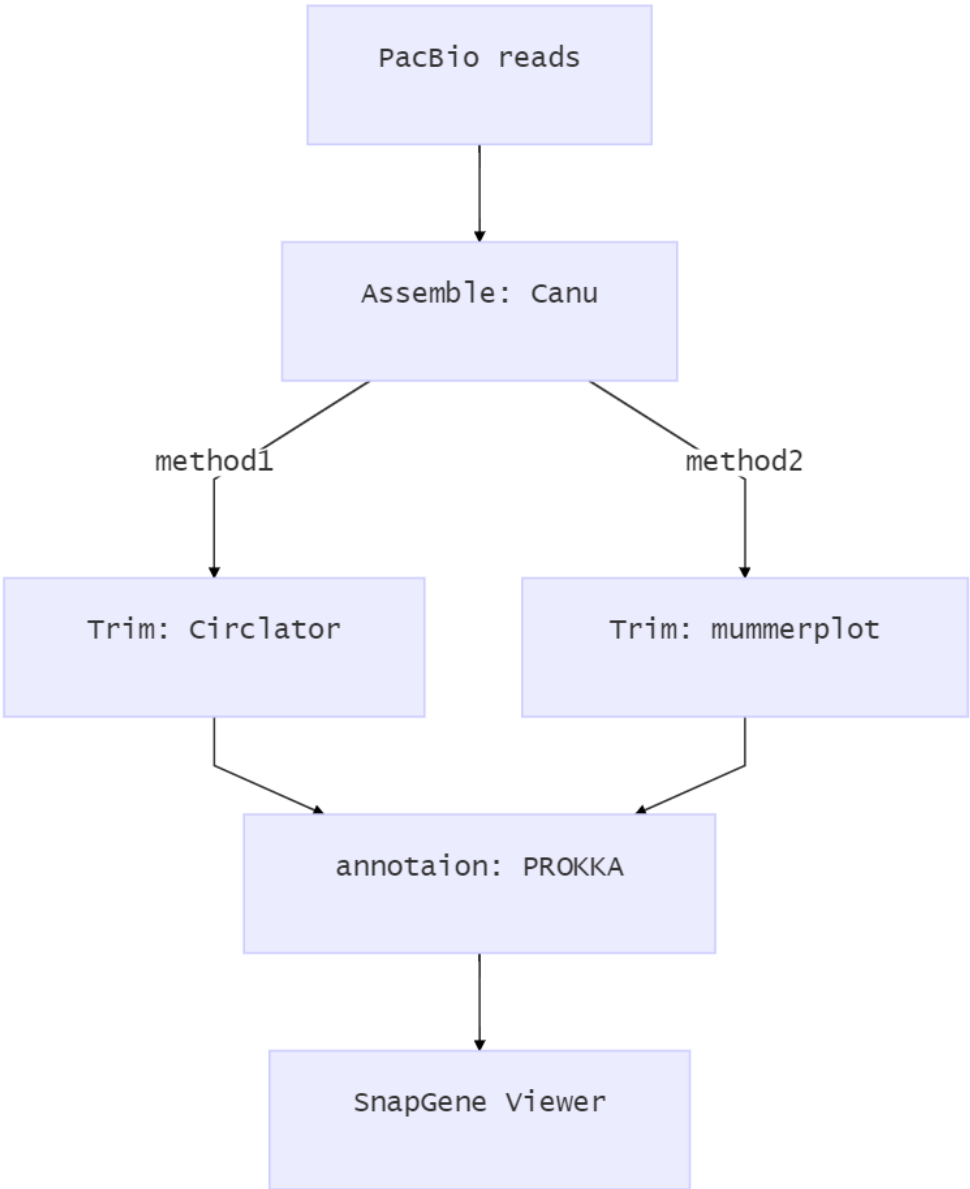
De Novo

*bacteria genome*의 *Long read sequencing* 결과에서 **MCR-1**의 유전 위치를 확인해본다.

OverView



1. Long Read Sequencing Results

Get data

the PacBio RSII reads :

- **201600135.fasta**
- **201600138.fasta**

Sequencing Stat

Sequence Length Distribution From Fasta File.

```
#!/usr/bin/python
from Bio import SeqIO #sequence를 읽어오기 위해 biopython 사용
import sys
cmdargs = str(sys.argv)
for seq_record in SeqIO.parse(str(sys.argv[1]), "fasta"):
 output_line = '%s\t%i' % \
(seq_record.id, len(seq_record))
 print(output_line)
```

To run,

```
chmod +x lenght.py
python lenght.py input.fasta
```

Result

```
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16178/7711_13641  5930
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16178/13689_19007 5318
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16181/0_7719      7719
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16184/4399_13212  8813
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16187/0_12604     12604
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16189/31312_35736 4424
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16189/35788_45970 10182
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16190/10678_17078 6400
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16190/17123_29291 12168
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16190/29329_39515 10186
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16191/198_10212   10014
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16194/3198_14985  11787
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16195/8138_15270  7132
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16196/1924_23125  21201
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16197/5805_23394  17589
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16199/2881_25198  22317
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16200/0_15012     15012
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16201/0_429       429
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16203/1170_15309  14139
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16208/3095_5606   2511
m170523_232803_42269_c101208672550000001823286610171733_s1_p0/16209/0_9581      9581
```

Statistics FASTA

```
#!/bin/sh
sort -n | awk '
  $1 ~ /^[0-9]*(\.[0-9]*)?$/ {
```

```
    a[c++] = $1; # c = count
    sum += $1;
  }
  END {
    avg = sum / c;
    if( (c % 2) == 1 ) {
      med = a[ int(c/2) ];
    } else {
      med = ( a[c/2] + a[c/2-1] ) / 2;
    }
    OFS="\t";
    { printf ("Total:\t""%"'''d\n", sum) }
    { printf ("Count:\t""%'''''d\n", c)}
    { printf ("Mean:\t""%'''''d\n", avg)}
    { printf ("Median:\t""%'''''d\n", med)}
    { printf ("Min:\t""%'''''d\n", a[0])}
    { printf ("Max:\t""%'''''d\n", a[c-1])}
  }
'
```

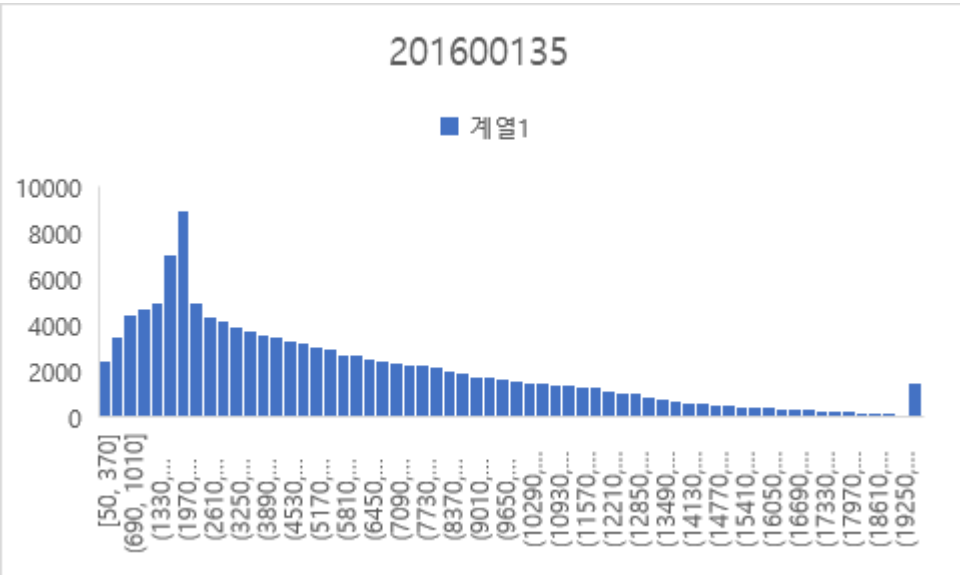
To run,

```
chomod +x stats.sh
python length.py input.fasta | cut -f 2 | sh stats.sh
```

Result

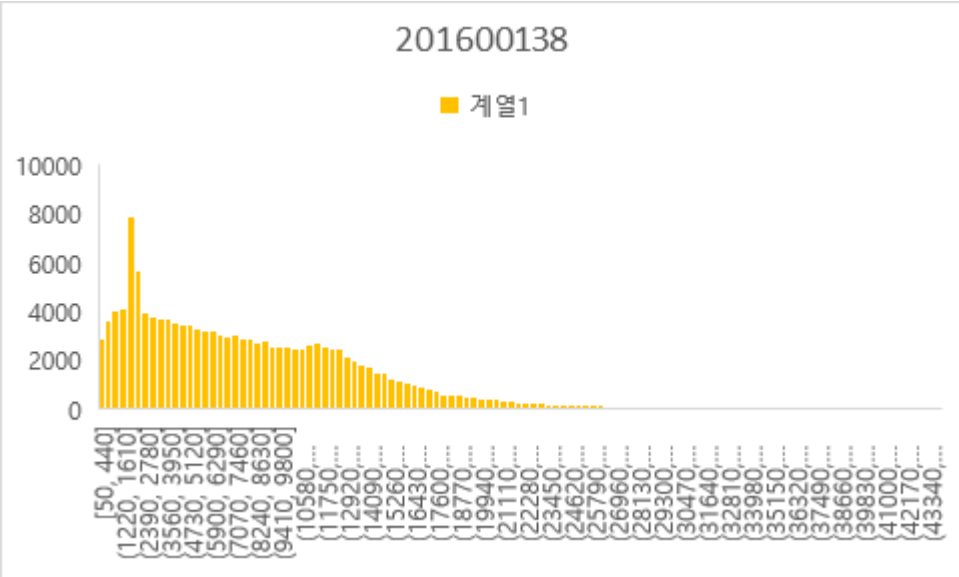
- **201600135.fasta**

NUMBER OF BASES	NUMBER OF READS	N50 READ LENGTH	MEAN READ LENGTH	MAX READ LENGTH
729,796,771	126,743	8,798	5,758	38,967



- **201600138.fasta**

NUMBER OF BASES	NUMBER OF READS	N50 READ LENGTH	MEAN READ LENGTH	MAX READ LENGTH
1,019,907,640	131,711	11,576	7,743	43,929



2. Genome De Novo Assembly

CANU

Canu specializes in assembling PacBio or Oxford Nanopore sequences.

1. install

https://github.com/marbl/canu

```
wget https://github.com/marbl/canu/releases/download/v1.8/canu-1.8.Linux-amd64.tar.xz
tar xvfcanu-1.8.Linux-amd64.tar.xz
```

2. Run

> Usually, the size of the genome of Bacteria is about 5.5 meters.

```
canu -p canu -d outdir genomeSize= -pacbio-raw input.fasta
```

- `-p canu` names prefix for output files (“canu”)
- `-d outdir` names output directory

3. Output

```
201600138.contigs.fasta                201600138.unitigs.fasta
201600138.contigs.gfa                  201600138.unitigs.gfa
201600138.contigs.layout               201600138.unitigs.layout
201600138.contigs.layout.readToTig     201600138.unitigs.layout.readToTig
201600138.contigs.layout.tigInfo       201600138.unitigs.layout.tigInfo
201600138.correctedReads.fasta.gz      canu-logs
201600138.report                       canu-scripts
201600138.seqStore                     circlator_outdir
201600138.seqStore.err                 correction
201600138.seqStore.ssi                 haplotype
201600138.trimmedReads.fasta.gz        tig0000005.fasta
201600138.unassembled.fasta            trimming
201600138.unitigs.bed                  unitigging
```

- **contigs.fasta** is assembled sequences.
- Display basic information about sequences: `infoseq` is a tool from EMBOSS

```
infoseq ~.contigs.fasta
```

201600135.contigs.fasta

```
(base) dnalink02@tmp-dnalinkserver:~/data/20160035-pacbio$ infoseq 201600135.contigs.fasta
Display basic information about sequences
USA                     Database  Name           Accession      Type Length %GC    Organism            Description
fasta::201600135.contigs.fasta:tig00000001 -              tig00000001    -              N    5072308 50.80                       len=5072308 reads=15584 covStat=10248.34 ga
ppedBases=no class=contig suggestRepeat=no suggestCircular=no
fasta::201600135.contigs.fasta:tig00000007 -              tig00000007    -              N    154500 49.45                       len=154500 reads=292 covStat=398.94 gappedBa
ses=no class=contig suggestRepeat=no suggestCircular=yes
fasta::201600135.contigs.fasta:tig00000008 -              tig00000008    -              N    58588  42.16                       len=58588 reads=58 covStat=168.60 gappedBase
s=no class=contig suggestRepeat=no suggestCircular=no
fasta::201600135.contigs.fasta:tig00000011 -              tig00000011    -              N    5967   48.20                       len=5967 reads=2157 covStat=-1476.11 gappedB
ases=no class=contig suggestRepeat=no suggestCircular=yes
```

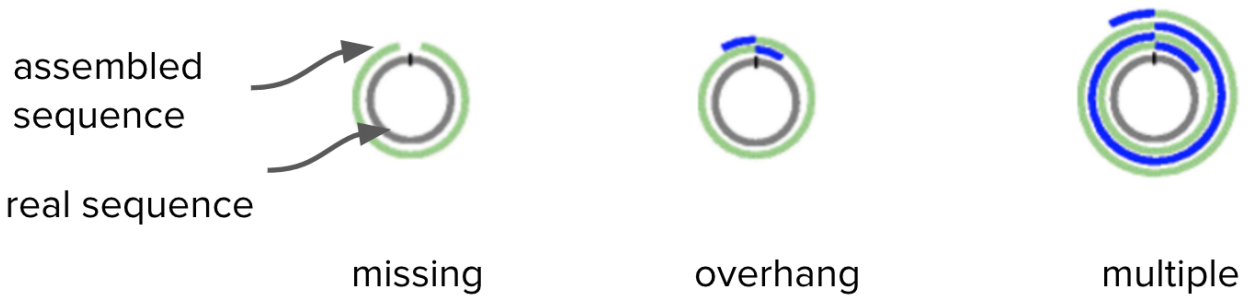
201600138.contigs.fasta

```
(base) dnalink02@tmp-dnalinkserver:~/data/201600138-pacbio$ infoseq 201600138.contigs.fasta
Display basic information about sequences
USA                     Database  Name           Accession      Type Length %GC    Organism            Description
fasta::201600138.contigs.fasta:tig00000001 -              tig00000001    -              N    4881734 50.70                       len=4881734 reads=11087 covStat=7544.78 gap
pedBases=no class=contig suggestRepeat=no suggestCircular=yes
fasta::201600138.contigs.fasta:tig00000002 -              tig00000002    -              N    26919  53.68                       len=26919 reads=10 covStat=24.87 gappedBases
=no class=contig suggestRepeat=no suggestCircular=no
fasta::201600138.contigs.fasta:tig00000005 -              tig00000005    -              N    174892 49.81                       len=174892 reads=482 covStat=169.11 gappedBa
ses=no class=contig suggestRepeat=no suggestCircular=yes
fasta::201600138.contigs.fasta:tig00000006 -              tig00000006    -              N    74647  42.49                       len=74647 reads=152 covStat=83.12 gappedBase
s=no class=contig suggestRepeat=no suggestCircular=yes
fasta::201600138.contigs.fasta:tig00002299 -              tig00002299    -              N    3969   48.12                       len=3969 reads=1 covStat=0.00 gappedBases=no
 class=contig suggestRepeat=no suggestCircular=no
fasta::201600138.contigs.fasta:tig00002300 -              tig00002300    -              N    3968   48.29                       len=3968 reads=1006 covStat=-687.56 gappedBa
ses=no class=contig suggestRepeat=no suggestCircular=yes
```

3. Circularize Check

Circlator

A tool to circularize genome assemblies. Circlator identifies and trims overhangs (on chromosomes and plasmids) and orients the start position at an appropriate gene (e.g. dnaA). It takes in the assembled contigs from Canu, as well as the corrected reads prepared by Canu.



1. install

```
pip3 install circlator
```

2. Run

Given an assembly `assembly.fasta` in FASTA format and corrected PacBio reads in a file called `reads`, run

```
circlator all canu.contigs.fasta canu.correctedReads.fasta.gz outdir
```

3. Output

```
(base) dnalink02@tmp-dnalinkserver:~/data/20160035-pacbio/circlator_outdir$ ls
00.info.txt                          04.merge.merge.iterations.log
00.input_assembly.fasta              05.clean.contigs_to_keep
00.input_assembly.fasta.fai          05.clean.coords
01.mapreads.bam                      05.clean.fasta
01.mapreads.bam.bai                  05.clean.log
02.bam2reads.fasta                   05.clean.remove_small.fa
02.bam2reads.log                     06.fixstart.ALL_FINISHED
03.assemble                          06.fixstart.contigs_to_not_change
04.merge.circularise.coords          06.fixstart.detailed.log
04.merge.circularise_details.log     06.fixstart.fasta
04.merge.circularise.log             06.fixstart.log
04.merge.circularise.start_act.sh    06.fixstart.prodigal.for_prodigal.fa
04.merge.fasta                       06.fixstart.prodigal.prodigal.gff
04.merge.merge.iter.1.coords         06.fixstart.promer.contigs_with_ends.fa
04.merge.merge.iter.1.crunch         06.fixstart.promer.promer
04.merge.merge.iter.1.start_act.sh   PROKKA_01252019
```

trimmed contig sizes

```
infoseq 06.fixstart.fasta
```

201600135

```
(base) dnalink02@tmp-dnalinkserver:~/data/20160035-pacbio/circlator_outdir$ infoseq 06.fixstart.fasta
Display basic information about sequences
USA                      Database  Name           Accession      Type Length %GC    Organism            Descr
iption
fasta::06.fixstart.fasta:tig00000001 -              tig00000001    -              N    5072308 50.80

fasta::06.fixstart.fasta:tig00000007 -              tig00000007    -              N    154500 49.45

fasta::06.fixstart.fasta:tig00000008 -              tig00000008    -              N    60953  42.32

fasta::06.fixstart.fasta:tig00000011 -              tig00000011    -              N    5967   48.20
```

201600138

```
fasta::00.input_assembly.fasta:tig00002300 -              tig00002300    -              N    3968   48.29
(base) dnalink02@tmp-dnalinkserver:~/data/201600138-pacbio/circlator_outdir$ infoseq 06.fixstart.fasta
Display basic information about sequences
USA                      Database  Name           Accession      Type Length %GC    Organism            Description
fasta::06.fixstart.fasta:tig00000001 -              tig00000001    -              N    4866035 50.69
fasta::06.fixstart.fasta:tig00000006.tig00000005 -              tig00000006.tig00000005 -              N    284848 47.48
fasta::06.fixstart.fasta:tig00002299 -              tig00002299    -              N    3969   48.12
```

Mummerplot

4. Gene Prediction

PROKKA

1. install

https://github.com/tseemann/prokka

```
conda create -n prokka_env -c conda-forge -c bioconda prokka
```

2. Run

```
prokka --outdir mydir --prefix genome contigs.fa
```

3. Output

```
(base) dnalink02@tmp-dnalinkserver:~/data/201600138-pacbio/circlator_outdir/PROKKA_0118
2019$ ls
PROKKA_01182019.err  PROKKA_01182019.fna  PROKKA_01182019.gff  PROKKA_01182019.tbl
PROKKA_01182019.faa  PROKKA_01182019.fsa  PROKKA_01182019.log  PROKKA_01182019.tsv
PROKKA_01182019.ffn  PROKKA_01182019.gbk  PROKKA_01182019.sqn  PROKKA_01182019.txt
```

Locate the MCR-1

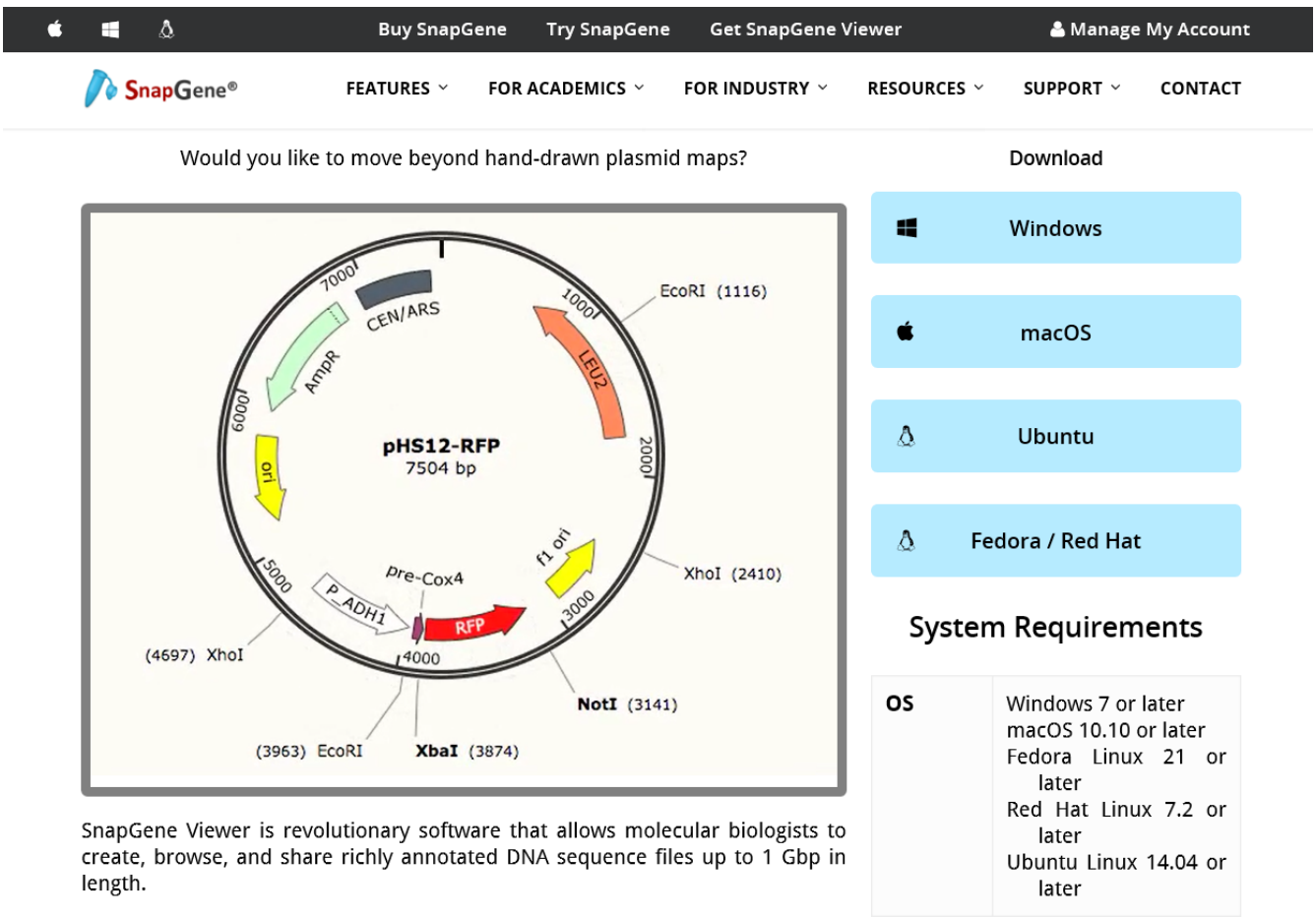
```
grep 'mcr' PROKKA.gff
```

```
(base) dnalink02@tmp-dnalinkserver:~/data/201600138-pacbio/circlator_outdir/PROKKA_0118
2019$ grep 'mcr' PROKKA_01182019.gff
tig00000006.tig00000005 Prodigal:2.6    CDS     14664   16289   .       -       0     I
D=BLCGAEJP_04919;Name=mcr-1.1_1;gene=mcr-1.1_1;inference=ab initio prediction:Prodigal:
2.6,similar to AA sequence:BARRGD:A7J11_03461;locus_tag=BLCGAEJP_04919;product=phosphoe
thanolamine--lipid A transferase MCR-1.1
tig00000006.tig00000005 Prodigal:2.6    CDS     76769   78394   .       -       0     I
D=BLCGAEJP_05000;Name=mcr-1.1_2;gene=mcr-1.1_2;inference=ab initio prediction:Prodigal:
2.6,similar to AA sequence:BARRGD:A7J11_03461;locus_tag=BLCGAEJP_05000;product=phosphoe
thanolamine--lipid A transferase MCR-1.1
```

SnapGene Viewer

1. install

https://www.snapgene.com/snapgene-viewer/



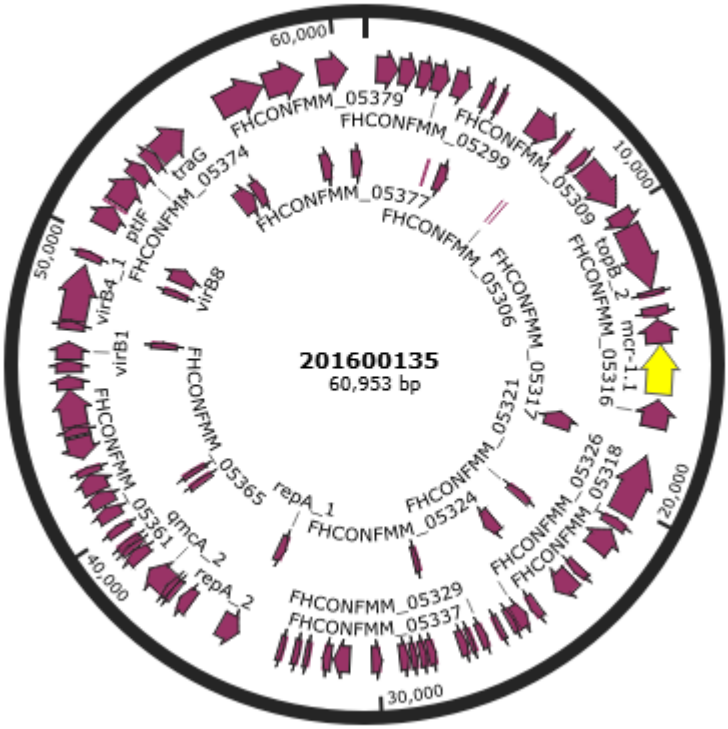
2. Run

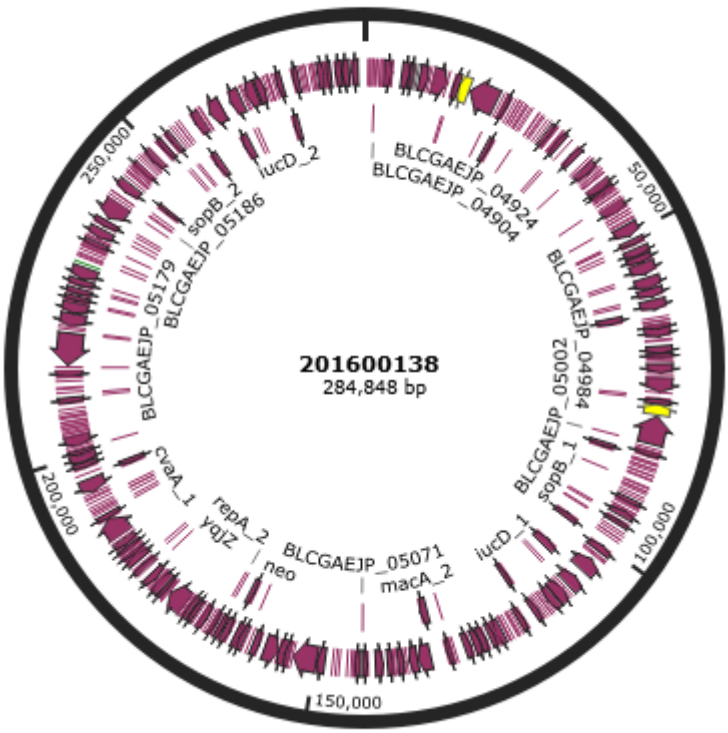
Change "linear" to "circular".

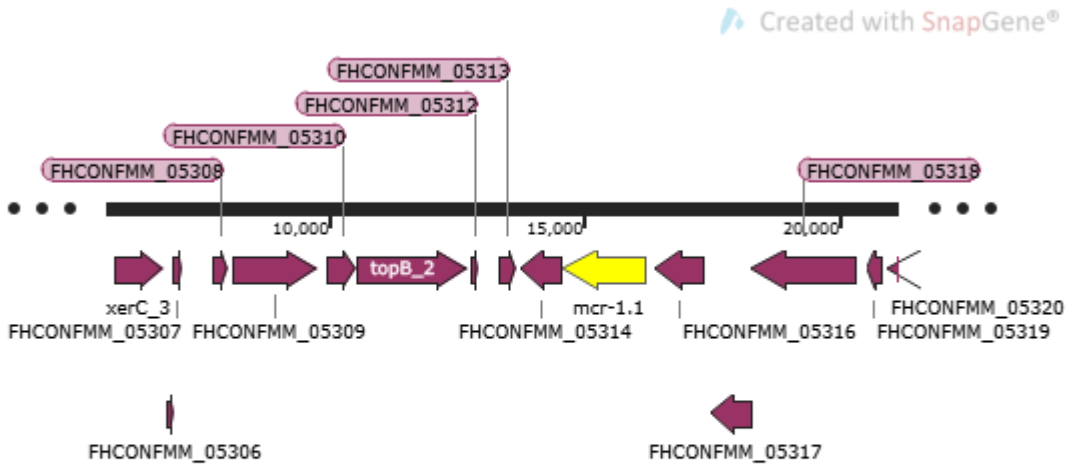
```
vi PROKKA_.gbk
```

```
LOCUS       tig00000001          4866035 bp    DNA     linear       18-JAN-2019
DEFINITION  Genus species strain strain.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Genus species
  ORGANISM  Genus species
            Unclassified.
COMMENT     Annotated using prokka 1.13.3 from
            https://github.com/tseemann/prokka.
FEATURES             Location/Qualifiers
     source          1..4866035
                     /organism="Genus species"
                     /mol_type="genomic DNA"
                     /strain="strain"
     CDS             70..1404
                     /gene="dnaA"
                     /locus_tag="BLCGAEJP_00001"
                     /inference="ab initio prediction:Prodigal:2.6"
                     /inference="similar to AA sequence:UniProtKB:P03004"
                     /codon_start=1
                     /transl_table=11
                     /product="Chromosomal replication initiator protein DnaA"
                     /db_xref="COG:COG0593"
                     /translation="MWIRPLQAELSDNTLALYAPNRFVLDWVRDKYLNNINGLLTSFC
                     GADAPQLRFEVGTKPVTQTPQAAVTSNVAAPAQVAQTQPQRAAPSTRSGWDNVPAPAE
                     PTYRSNVNVKHTFDNFVEGKSNQLARAAARQVADNPGGAYNPLFLYGGTGLGKTHLLH
                     AVGNGIMARKPNAKVVYMHSERFVQDMVKALQNNAIEEFKRYYRSVDALLIDDIQFFA
                     NKERSQEEFFHTFNALLEGNQQIILTSDRYPKEINGVEDRLKSRFGWGLTVAIEPPEL
                     ETRVAILMKKADENDIRLPGEVAFFIAKRLRSNVRELEGALNRVIANANFTGRAITID
                     FVREALRDLLALQEKLVTIDNIQKTVAEYYKIKVADLLSKRRSRSVARPRQMAMALAK
                     ELTNHSLPEIGDAFGGRDHTTVLHACRKIEQLREESHDIKEDFSNLIRTLSS"
     CDS             1409..2509
```

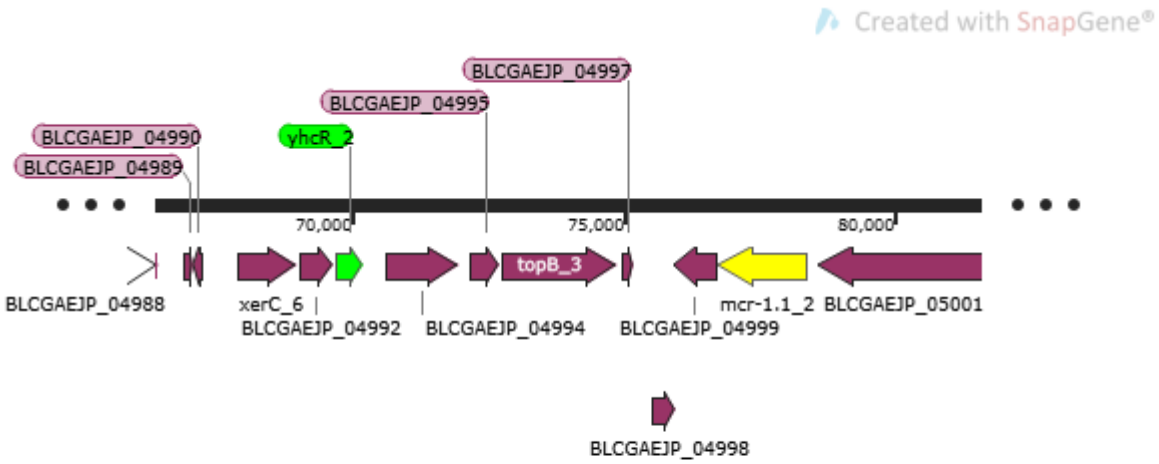
3. Output







201600135
60,953 bp



201600138
284,848 bp