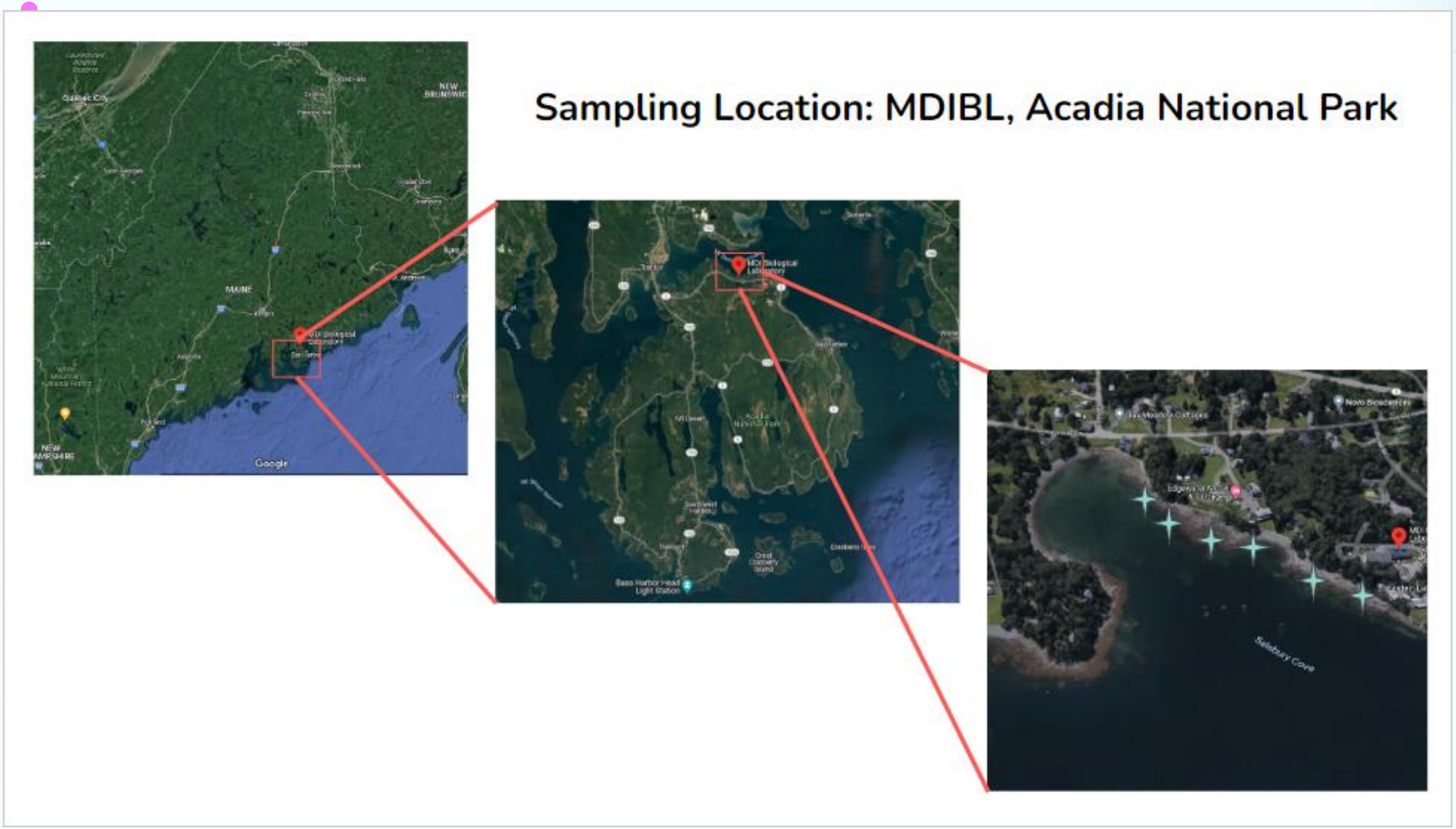
Gen 711 Final Project – Bacterial Genome Assembly

By, Caylin Grove, Ethan Morgan, and Graham Collinsworth

Background Information

- Data was collected by Anthony Hay, Steven Weicksel, Dana-Lynn Koomoa-Lange, Leah Elliot, Melissa Chisholm, Princess Rodriguez
- Pathway was created by Joseph Sevigny on Github
- Seaweed eating microbes were collected in MDIBL, Acadia National Park
- Their DNA was extracted and Illumina libraries were generated
- Illumina sequencing of the samples occurred



Sampling Location: MDIBL, Acadia National Park

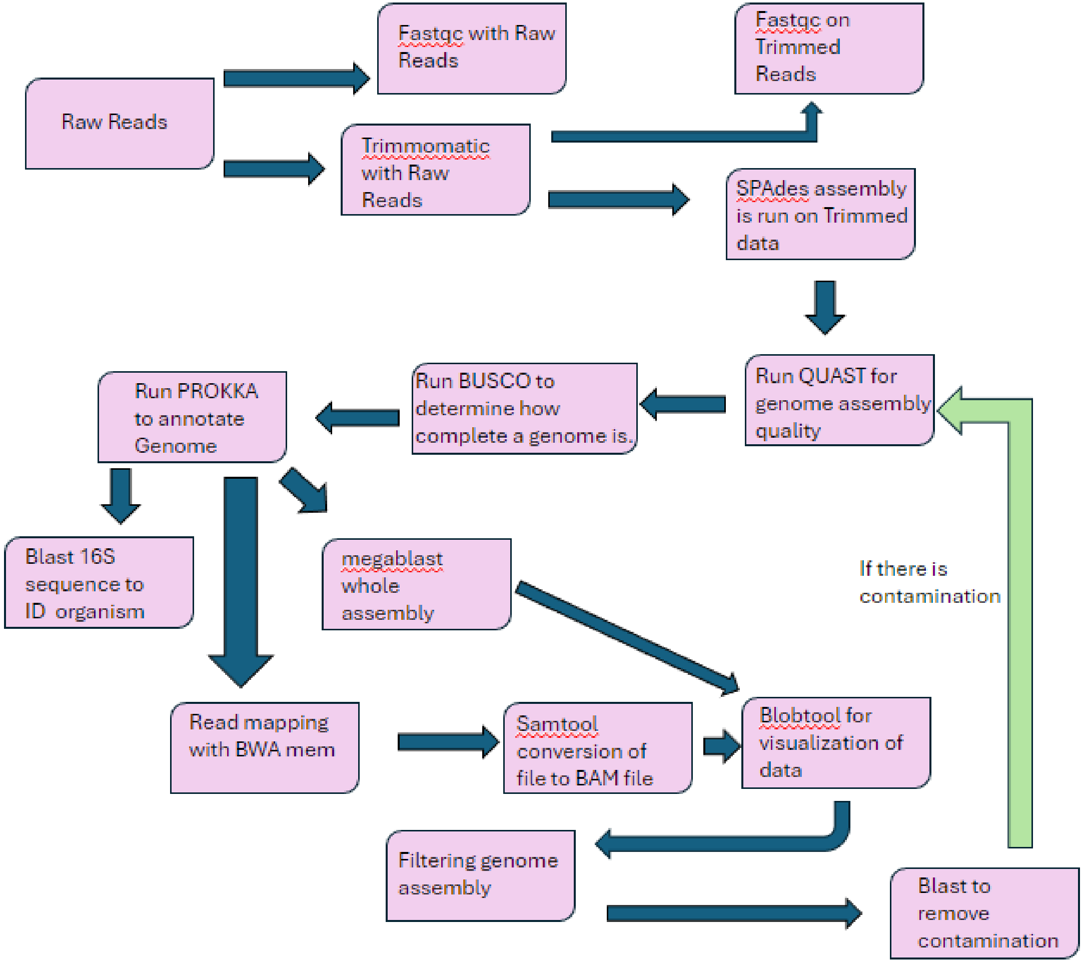
Joseph7e/MDIBL-T3-WGS-Tutorial: Bacterial Genome Assembly and Assessment (github.com)

Some General Information

- We each ran one set of samples

  Ethan Morgan: 69_S8_L001_R*_001.fastq.gz
  Graham Collinsworth: DC1_S41_L001_R*_001.fastq.gz
  Caylin Grove: 2_S26_L001_R*_001.fastq.gz

- We worked through each step together so our method was combined into one .sh file



Raw Reads

Fastqc with Raw Reads

Trimmomatic with Raw Reads

Fastqc on Trimmed Reads

SPAdes assembly is run on Trimmed data

Run QUAST for genome assembly quality

Run BUSCO to determine how complete a genome is.

Run PROKKA to annotate Genome

Blast 16S sequence to ID organism

megablast whole assembly

Read mapping with BWA mem

Samtool conversion of file to BAM file

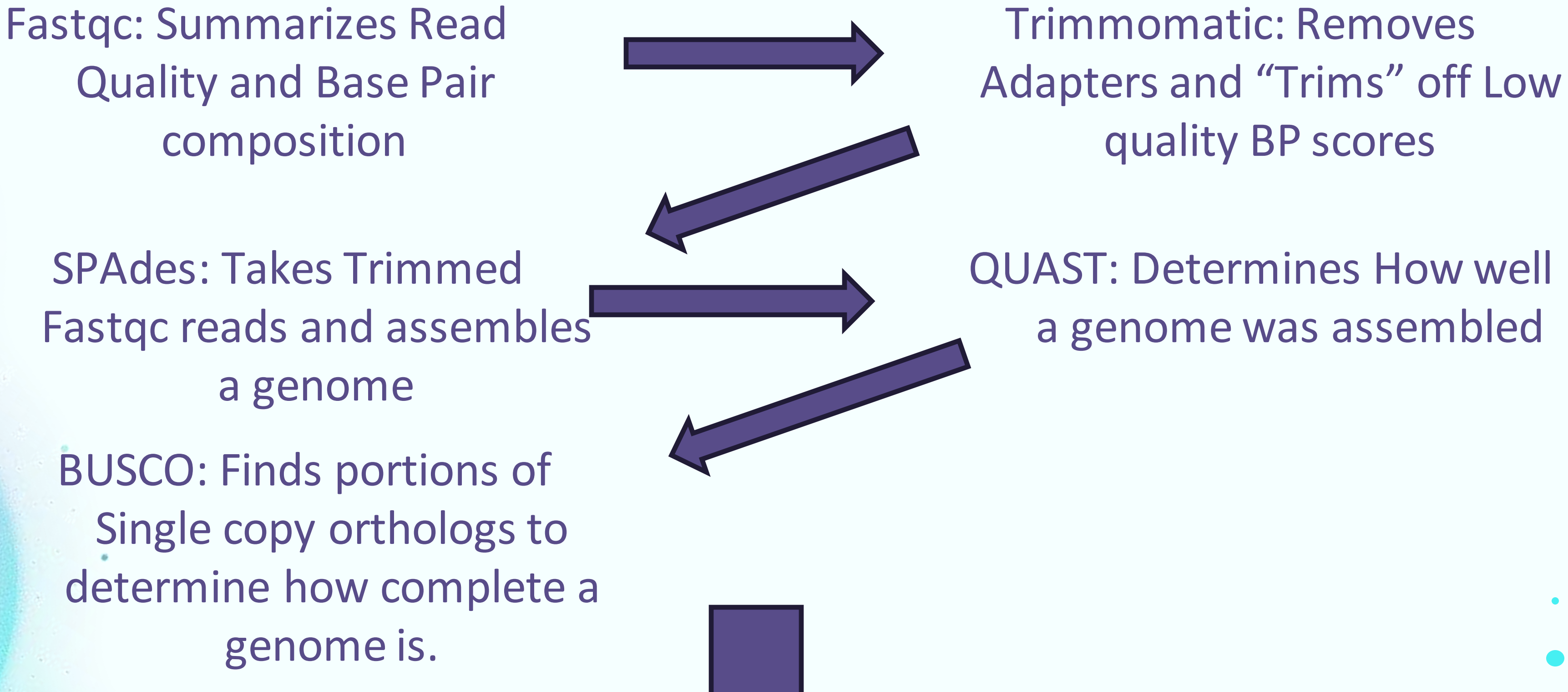
Blobtool for visualization of data

Filtering genome assembly

Blast to remove contamination

If there is contamination

Methods Overview



Fastqc: Summarizes Read Quality and Base Pair composition

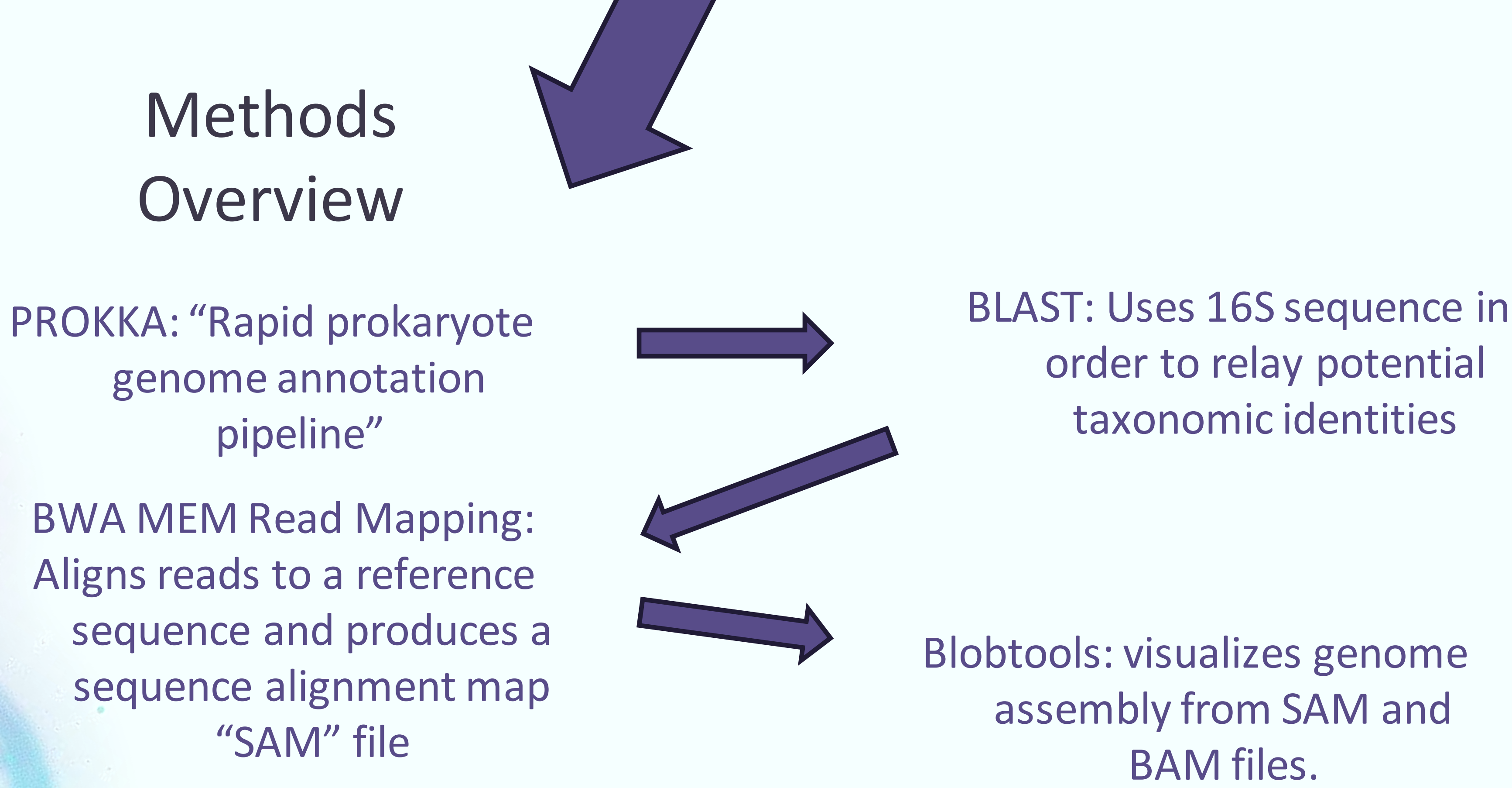
Trimmomatic: Removes Adapters and “Trims” off Low quality BP scores

SPAdes: Takes Trimmed Fastqc reads and assembles a genome

QUAST: Determines How well a genome was assembled

BUSCO: Finds portions of Single copy orthologs to determine how complete a genome is.

Methods Overview



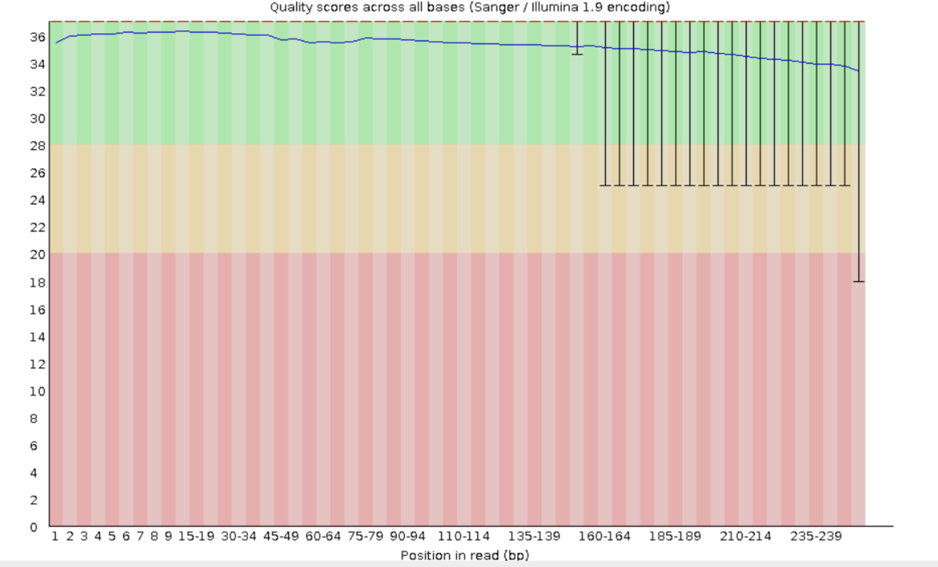
PROKKA: “Rapid prokaryote genome annotation pipeline”

BLAST: Uses 16S sequence in order to relay potential taxonomic identities

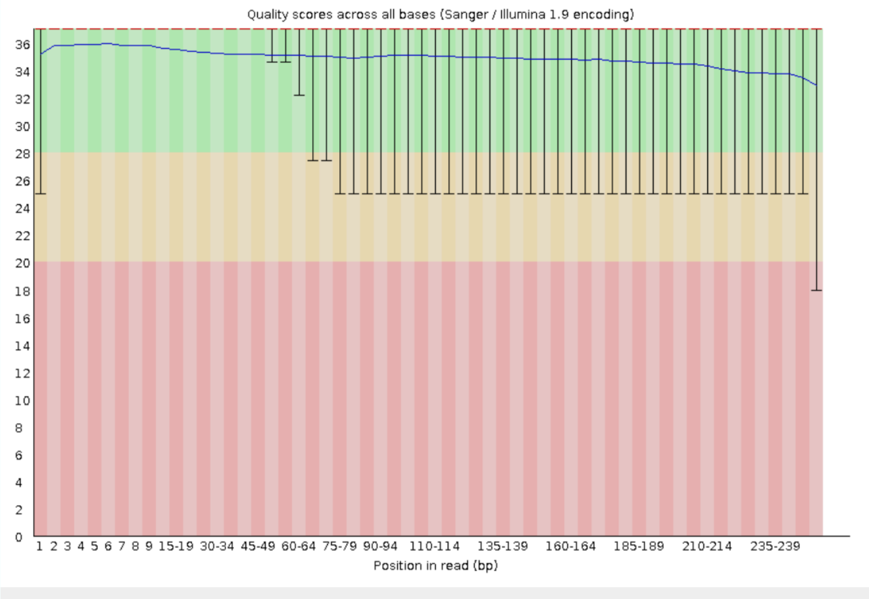
BWA MEM Read Mapping: Aligns reads to a reference sequence and produces a sequence alignment map “SAM” file

Blobtools: visualizes genome assembly from SAM and BAM files.

Fastqc run on samples -Caylin

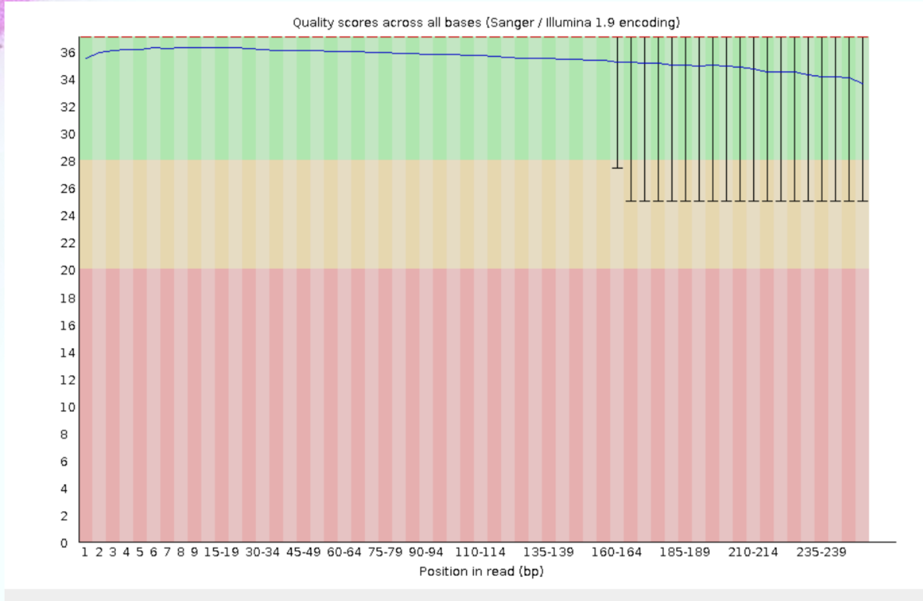


R1

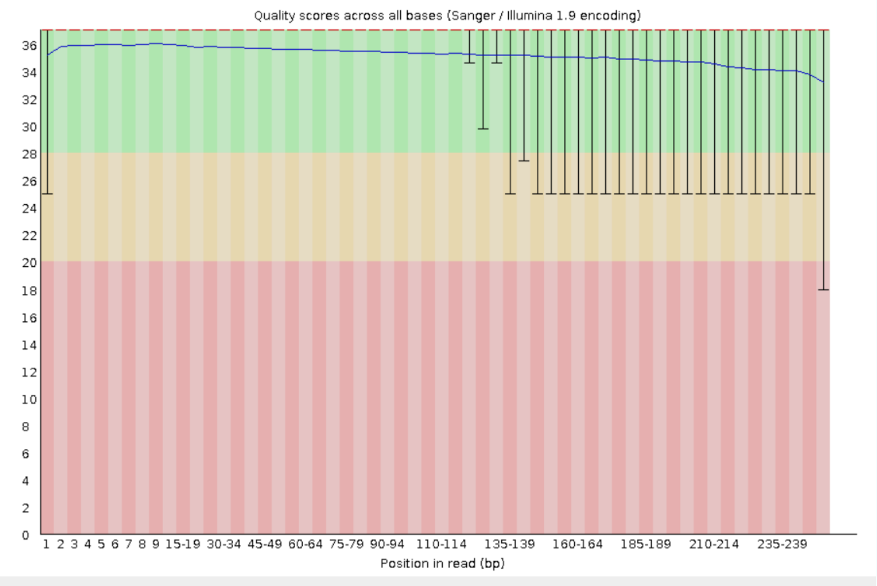


R2

Trimmomatic - Caylin



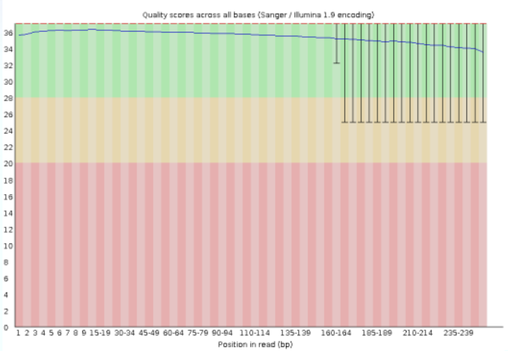
R1



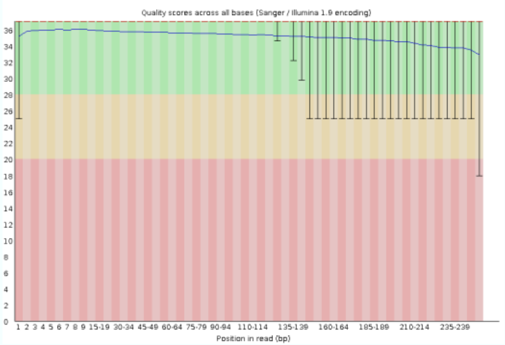
R2

Fastqc run on samples -Ethan

Raw Data

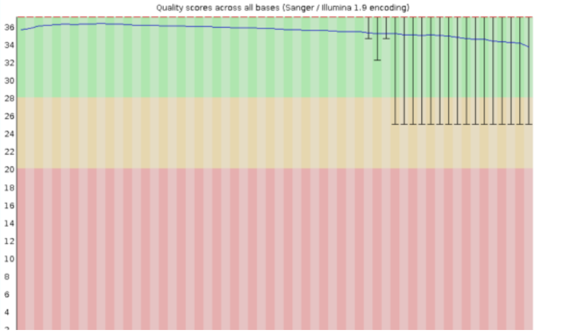


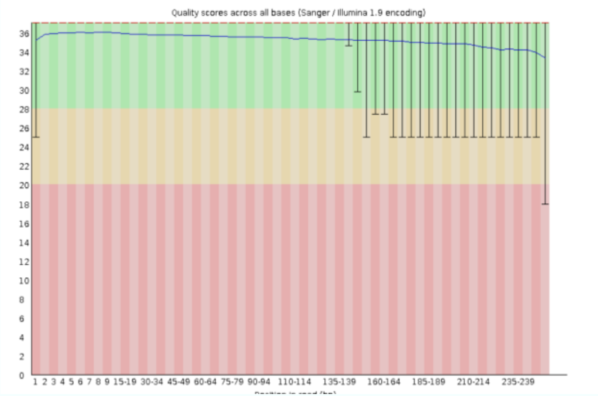
R1



R2

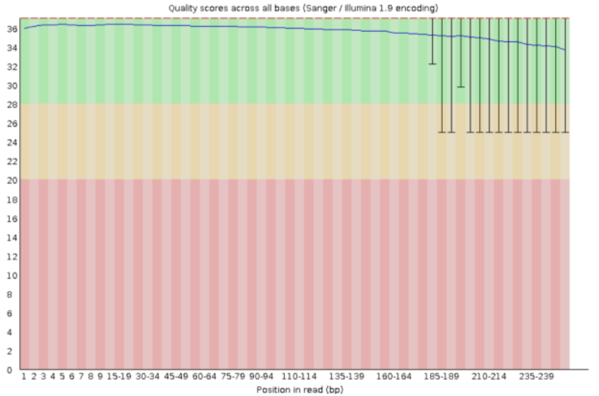
Trimmomatic Data

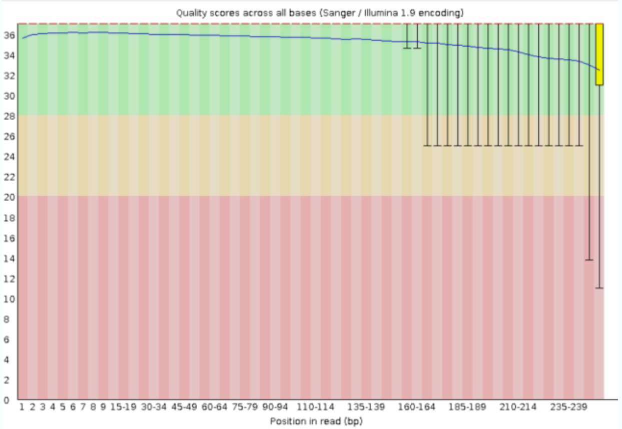




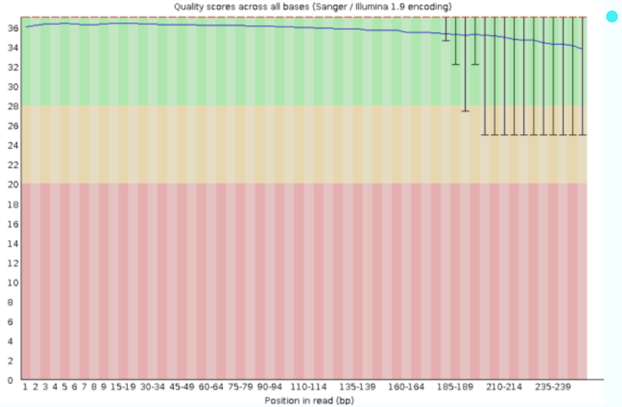
Fastqc - Graham

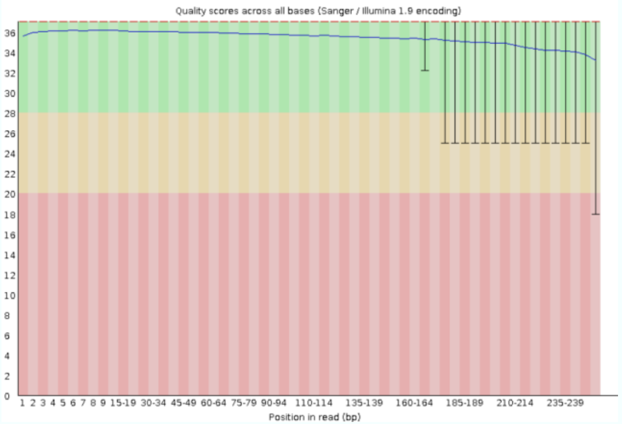
Raw Data



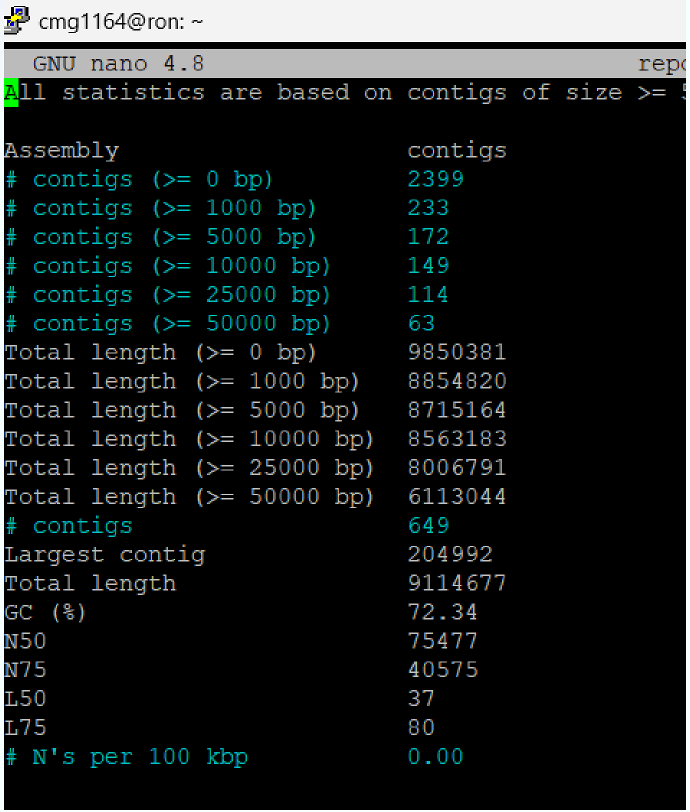


Trimmomatic Data





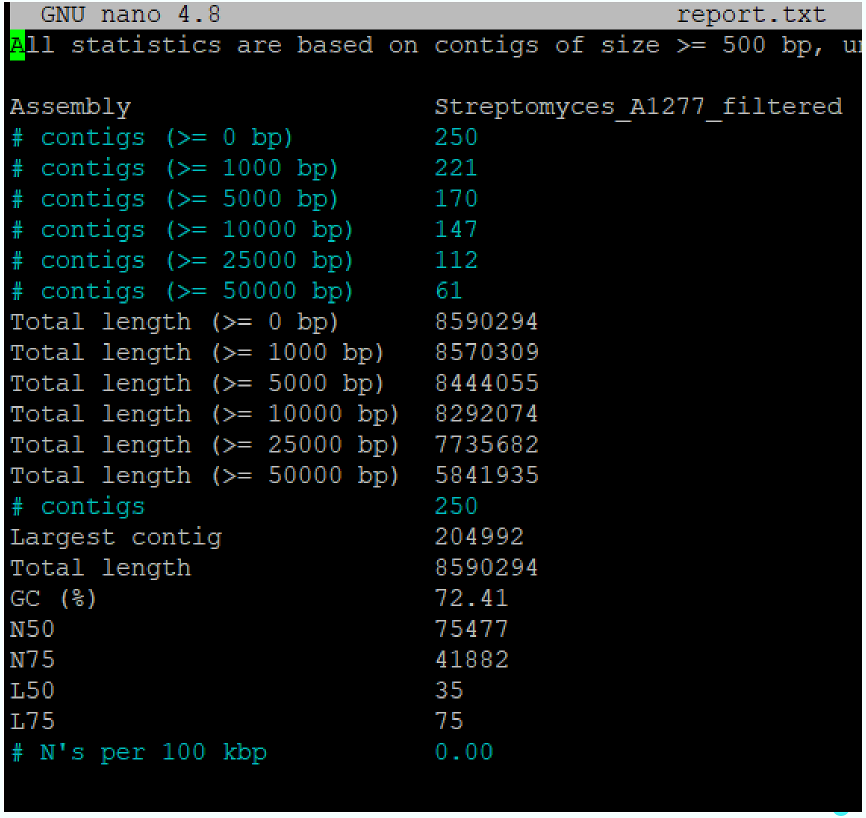
Quast – Caylin



```
  GNU nano 4.8                                          repo
All statistics are based on contigs of size >= 5

Assembly                    contigs
# contigs (>= 0 bp)         2399
# contigs (>= 1000 bp)      233
# contigs (>= 5000 bp)      172
# contigs (>= 10000 bp)     149
# contigs (>= 25000 bp)     114
# contigs (>= 50000 bp)     63
Total length (>= 0 bp)      9850381
Total length (>= 1000 bp)   8854820
Total length (>= 5000 bp)   8715164
Total length (>= 10000 bp)  8563183
Total length (>= 25000 bp)  8006791
Total length (>= 50000 bp)  6113044
# contigs                   649
Largest contig              204992
Total length                9114677
GC (%)                      72.34
N50                         75477
N75                         40575
L50                         37
L75                         80
# N's per 100 kbp           0.00
```

Before filtering



```
  GNU nano 4.8                                          report.txt
All statistics are based on contigs of size >= 500 bp, un

Assembly                    Streptomyces_A1277_filtered
# contigs (>= 0 bp)         250
# contigs (>= 1000 bp)      221
# contigs (>= 5000 bp)      170
# contigs (>= 10000 bp)     147
# contigs (>= 25000 bp)     112
# contigs (>= 50000 bp)     61
Total length (>= 0 bp)      8590294
Total length (>= 1000 bp)   8570309
Total length (>= 5000 bp)   8444055
Total length (>= 10000 bp)  8292074
Total length (>= 25000 bp)  7735682
Total length (>= 50000 bp)  5841935
# contigs                   250
Largest contig              204992
Total length                8590294
GC (%)                      72.41
N50                         75477
N75                         41882
L50                         35
L75                         75
# N's per 100 kbp           0.00
```

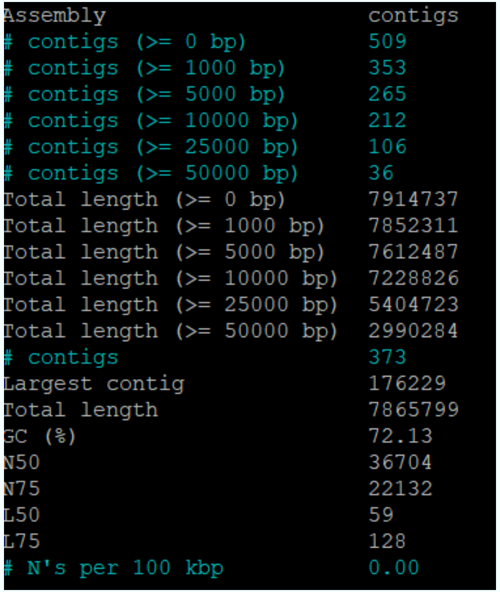
After filtering

QUAST Results: Ethan

N50: 36704

Total Length: 7,914,737

contigs: 373



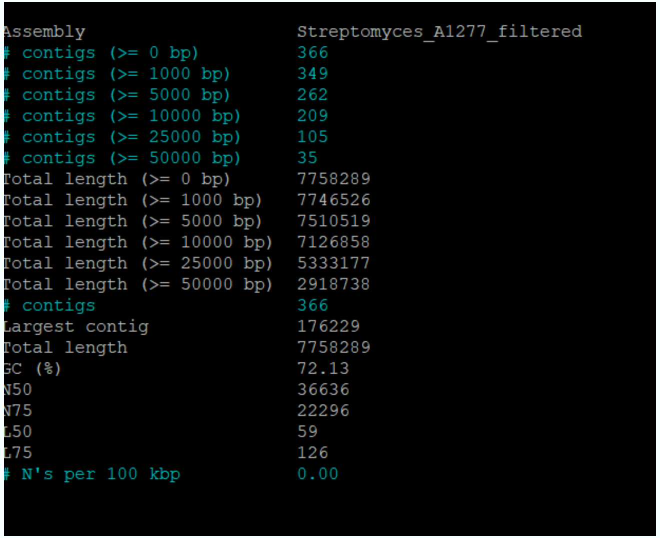
```
Assembly                     contigs
# contigs (>= 0 bp)          509
# contigs (>= 1000 bp)       353
# contigs (>= 5000 bp)       265
# contigs (>= 10000 bp)      212
# contigs (>= 25000 bp)      106
# contigs (>= 50000 bp)      36
Total length (>= 0 bp)       7914737
Total length (>= 1000 bp)    7852311
Total length (>= 5000 bp)    7612487
Total length (>= 10000 bp)   7228826
Total length (>= 25000 bp)   5404723
Total length (>= 50000 bp)   2990284
# contigs                    373
Largest contig               176229
Total length                 7865799
GC (%)                       72.13
N50                          36704
N75                          22132
L50                          59
L75                          128
# N's per 100 kbp            0.00
```

After Filtering:

N50: 36636

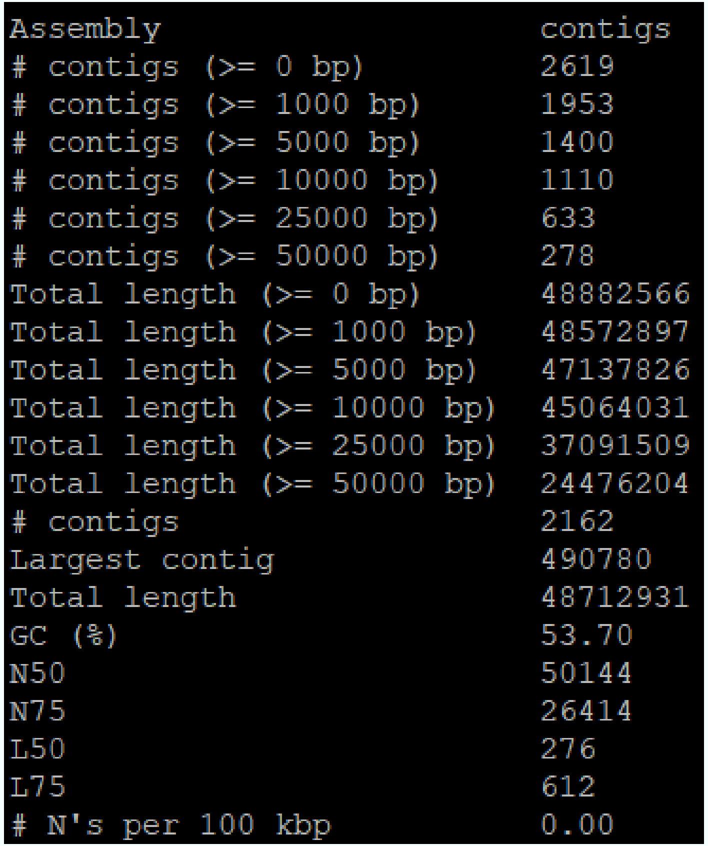
Total Length: 7,758,289

contigs: 366



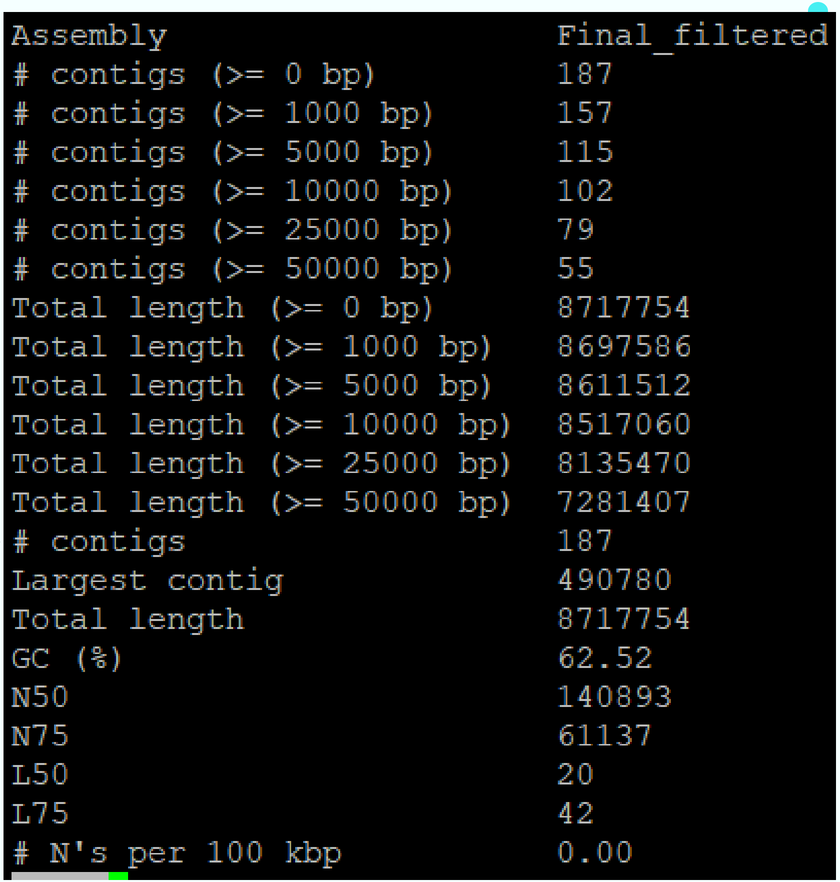
```
Assembly                     Streptomyces_A1277_filtered
# contigs (>= 0 bp)          366
# contigs (>= 1000 bp)       349
# contigs (>= 5000 bp)       262
# contigs (>= 10000 bp)      209
# contigs (>= 25000 bp)      105
# contigs (>= 50000 bp)      35
Total length (>= 0 bp)       7758289
Total length (>= 1000 bp)    7746526
Total length (>= 5000 bp)    7510519
Total length (>= 10000 bp)   7126858
Total length (>= 25000 bp)   5333177
Total length (>= 50000 bp)   2918738
# contigs                    366
Largest contig               176229
Total length                 7758289
GC (%)                       72.13
N50                          36636
N75                          22296
L50                          59
L75                          126
# N's per 100 kbp            0.00
```

Quast – Graham



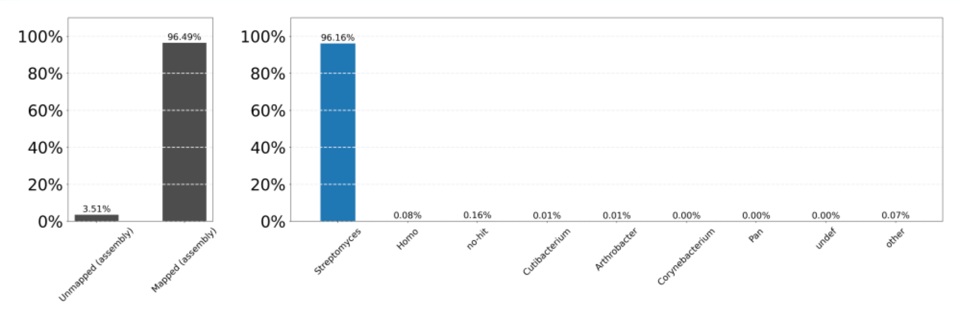
```
Assembly                    contigs
# contigs (>= 0 bp)         2619
# contigs (>= 1000 bp)      1953
# contigs (>= 5000 bp)      1400
# contigs (>= 10000 bp)     1110
# contigs (>= 25000 bp)     633
# contigs (>= 50000 bp)     278
Total length (>= 0 bp)      48882566
Total length (>= 1000 bp)   48572897
Total length (>= 5000 bp)   47137826
Total length (>= 10000 bp)  45064031
Total length (>= 25000 bp)  37091509
Total length (>= 50000 bp)  24476204
# contigs                   2162
Largest contig              490780
Total length                48712931
GC (%)                      53.70
N50                         50144
N75                         26414
L50                         276
L75                         612
# N's per 100 kbp           0.00
```

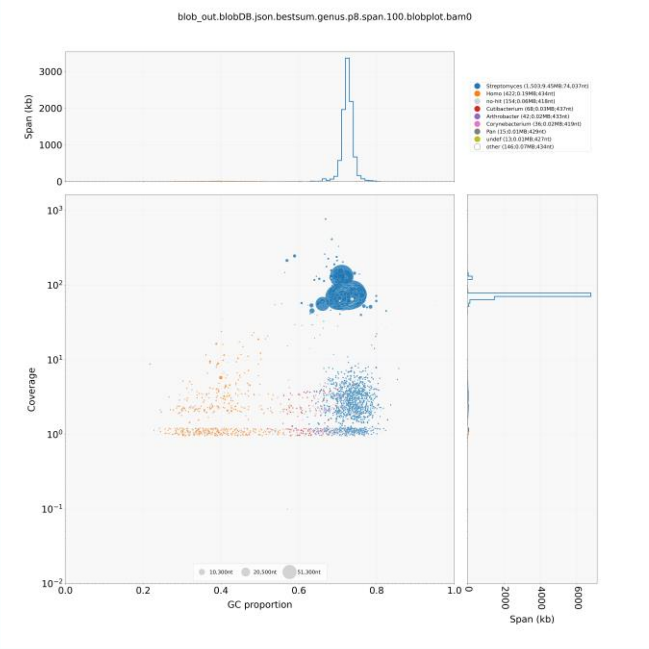
After Filtering



```
Assembly                    Final_filtered
# contigs (>= 0 bp)         187
# contigs (>= 1000 bp)      157
# contigs (>= 5000 bp)      115
# contigs (>= 10000 bp)     102
# contigs (>= 25000 bp)     79
# contigs (>= 50000 bp)     55
Total length (>= 0 bp)      8717754
Total length (>= 1000 bp)   8697586
Total length (>= 5000 bp)   8611512
Total length (>= 10000 bp)  8517060
Total length (>= 25000 bp)  8135470
Total length (>= 50000 bp)  7281407
# contigs                   187
Largest contig              490780
Total length                8717754
GC (%)                      62.52
N50                         140893
N75                         61137
L50                         20
L75                         42
# N's per 100 kbp           0.00
```

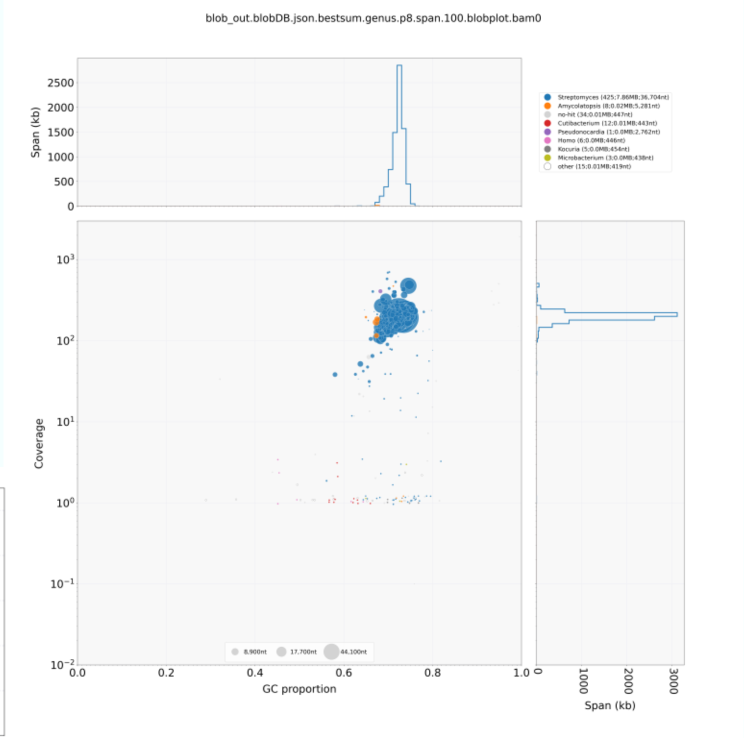
Blobtool (visualize genome assembly) - Caylin

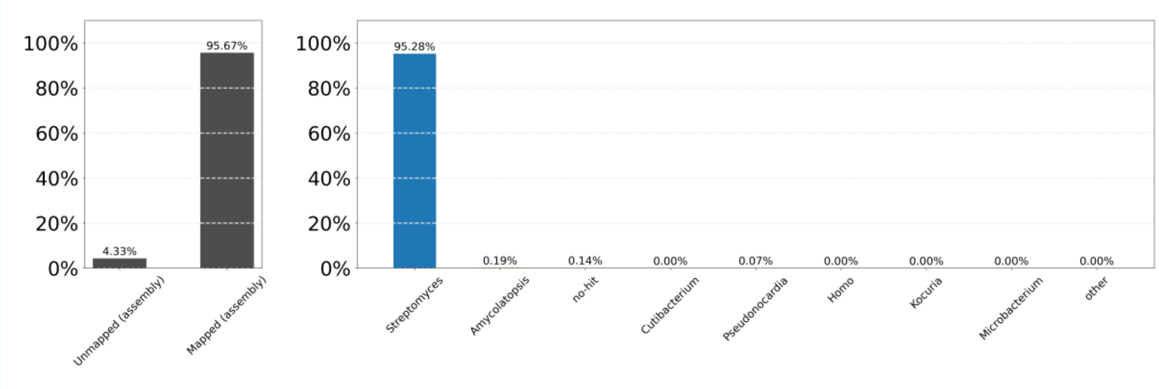




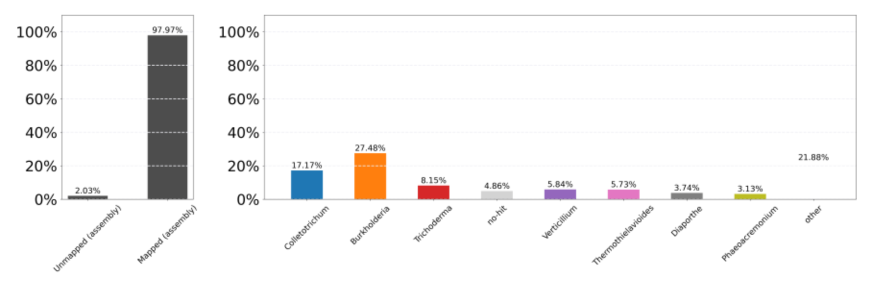
Before filtering

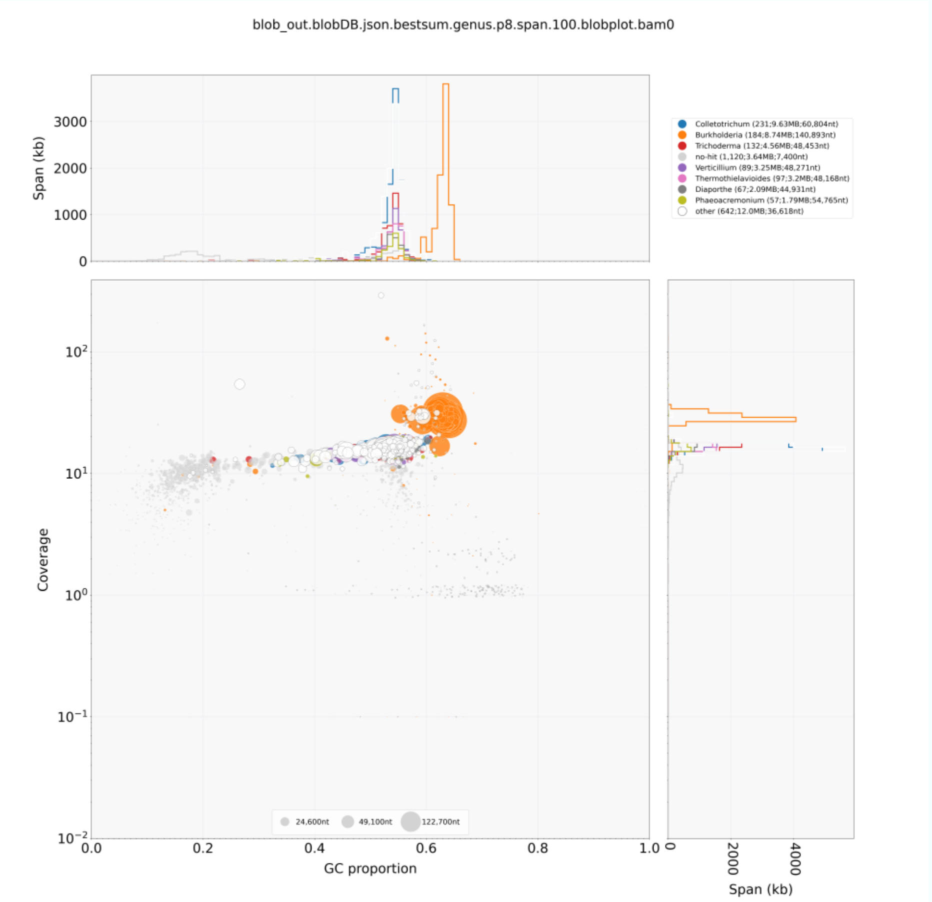
Blobtools - Ethan



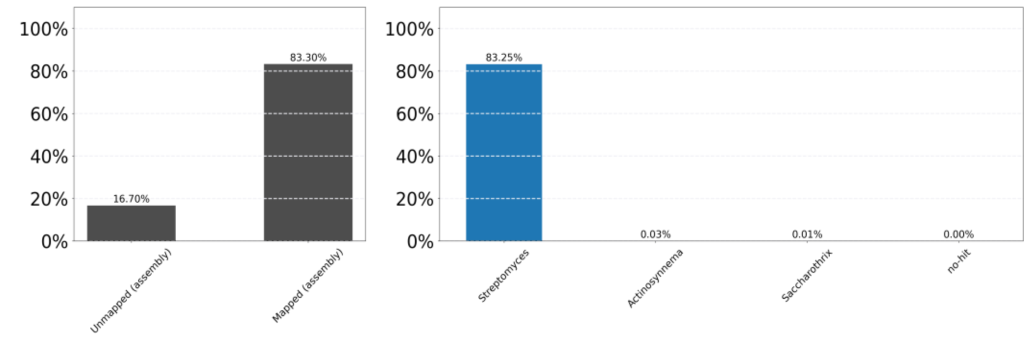


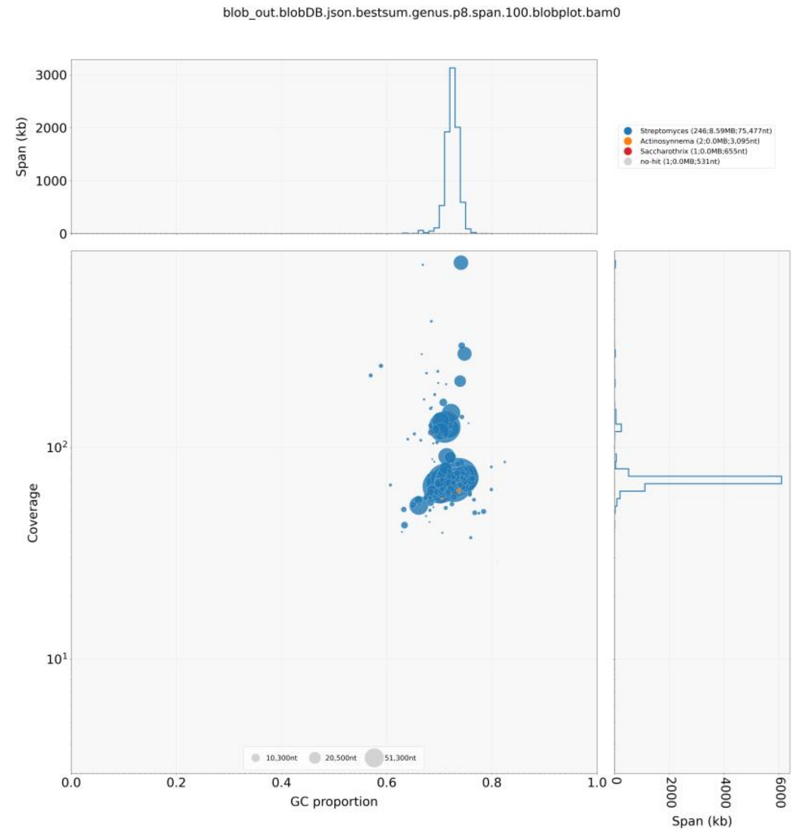
Blobtools – Graham





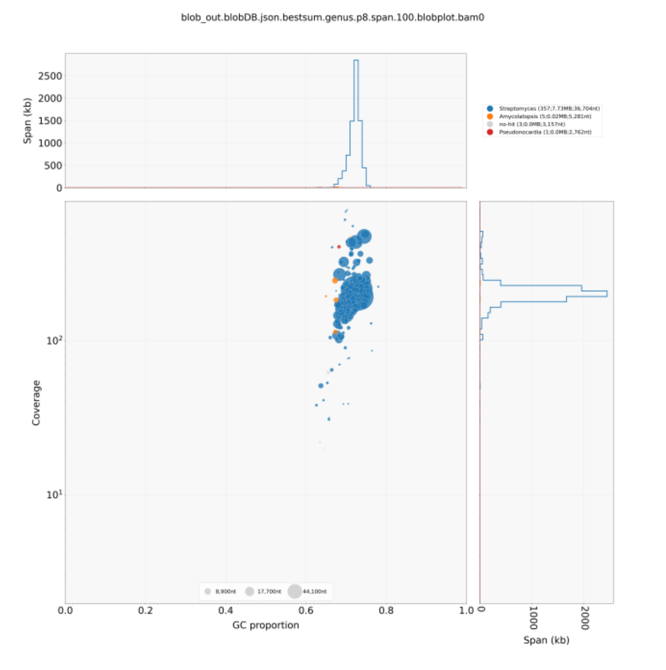
Blobtools ReRun with Filtered Data -Caylin



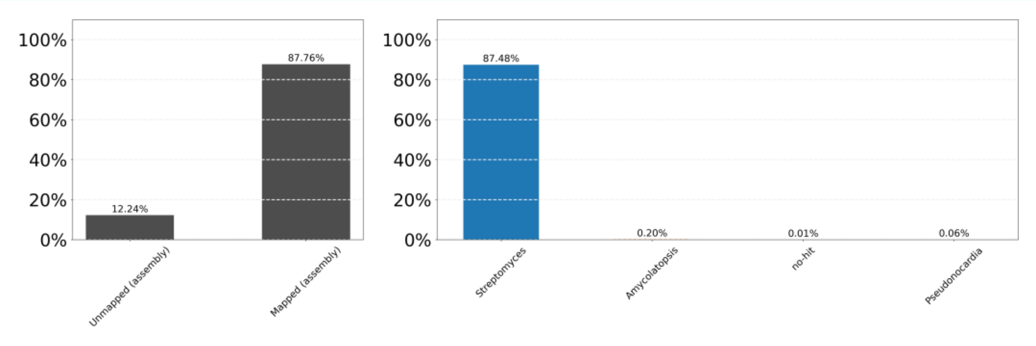


After filtering

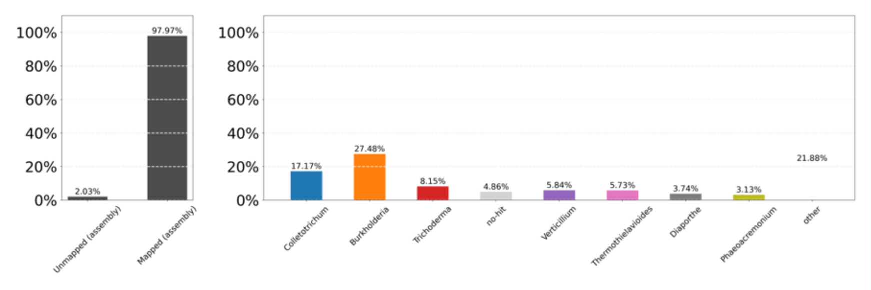
Blobtools ReRun with Filtered Data -Ethan



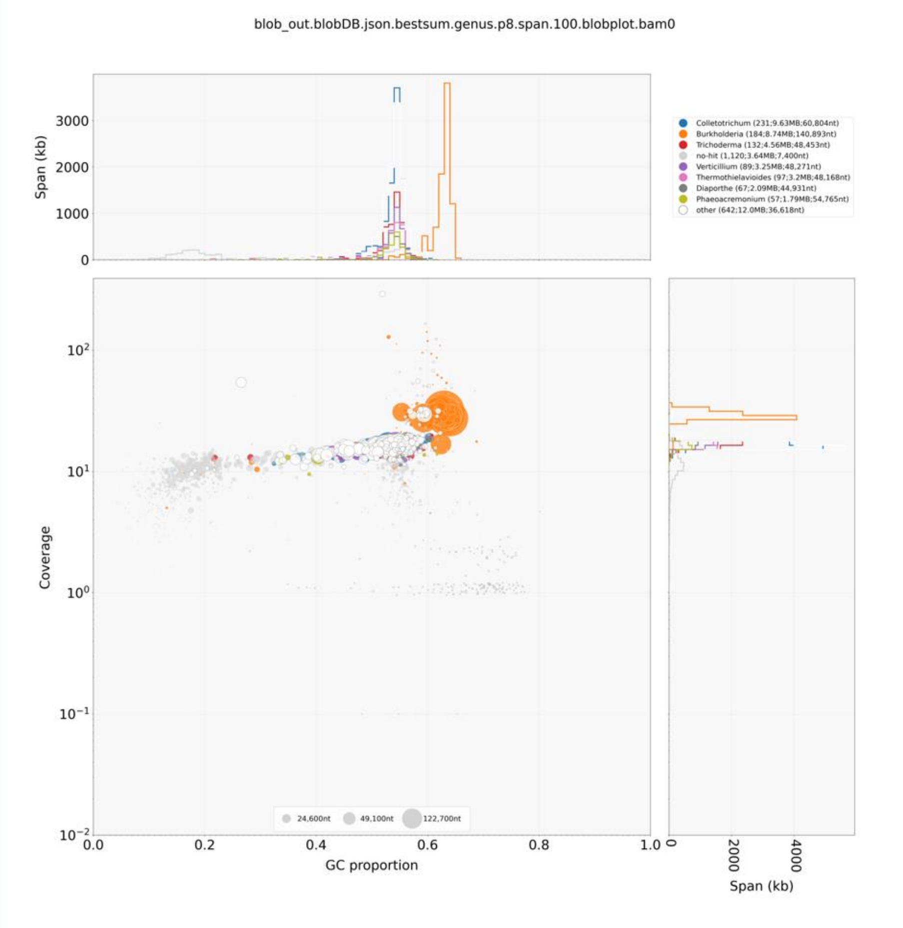
Identification: Streptomyces



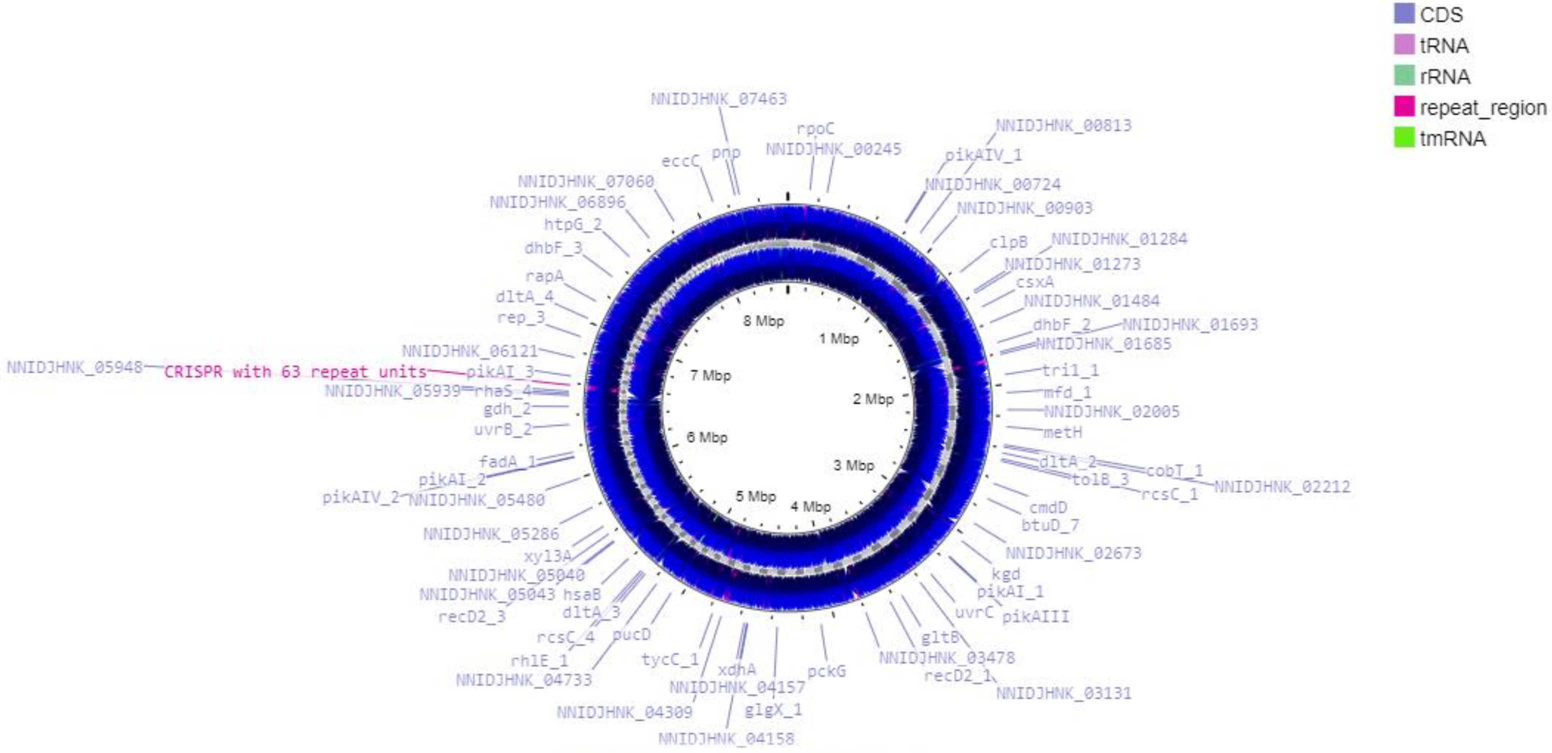
Blobtools ReRun with Filtered Data -Graham



Burkholderia Species

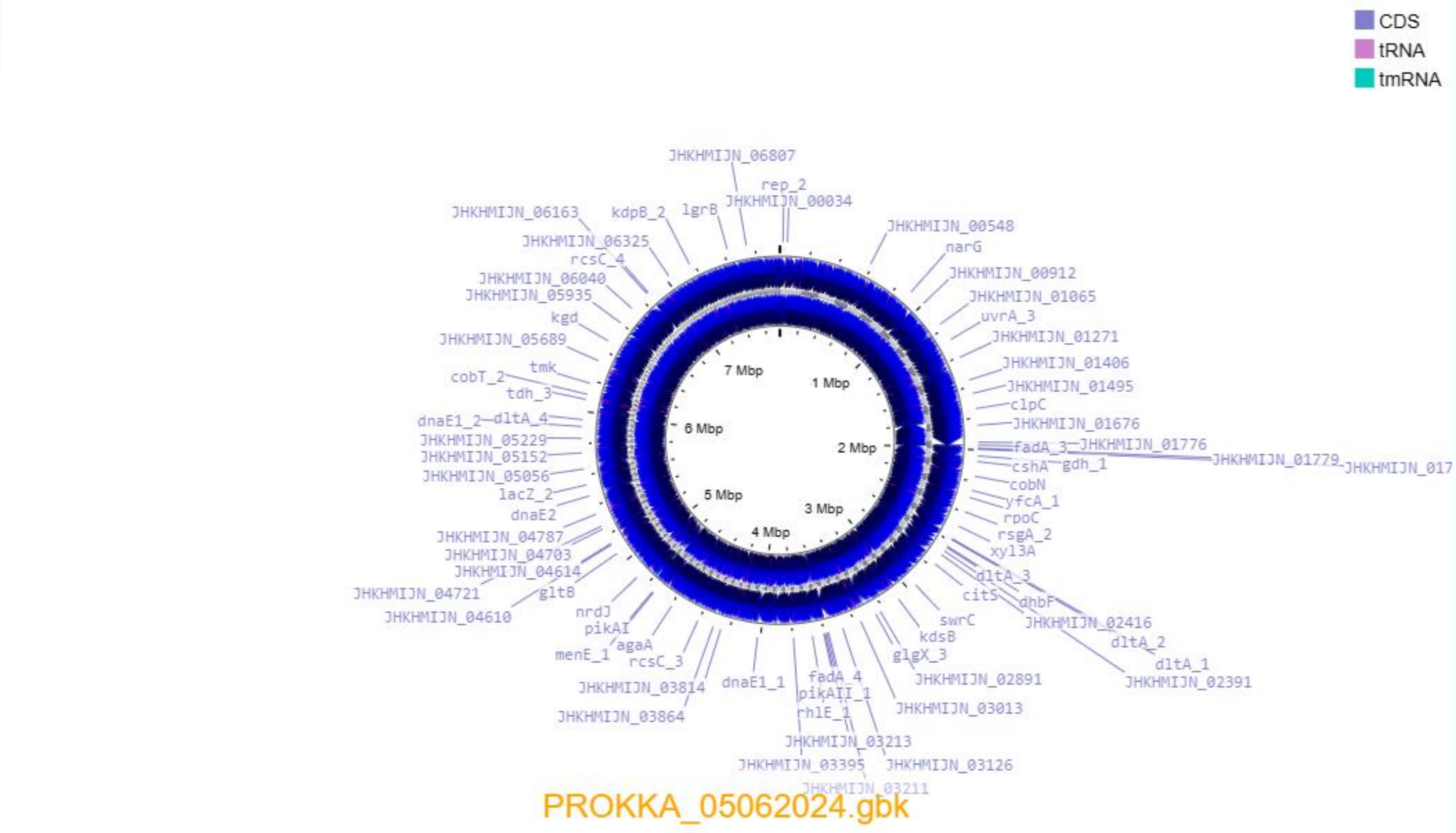


Genome Visualization - Caylin

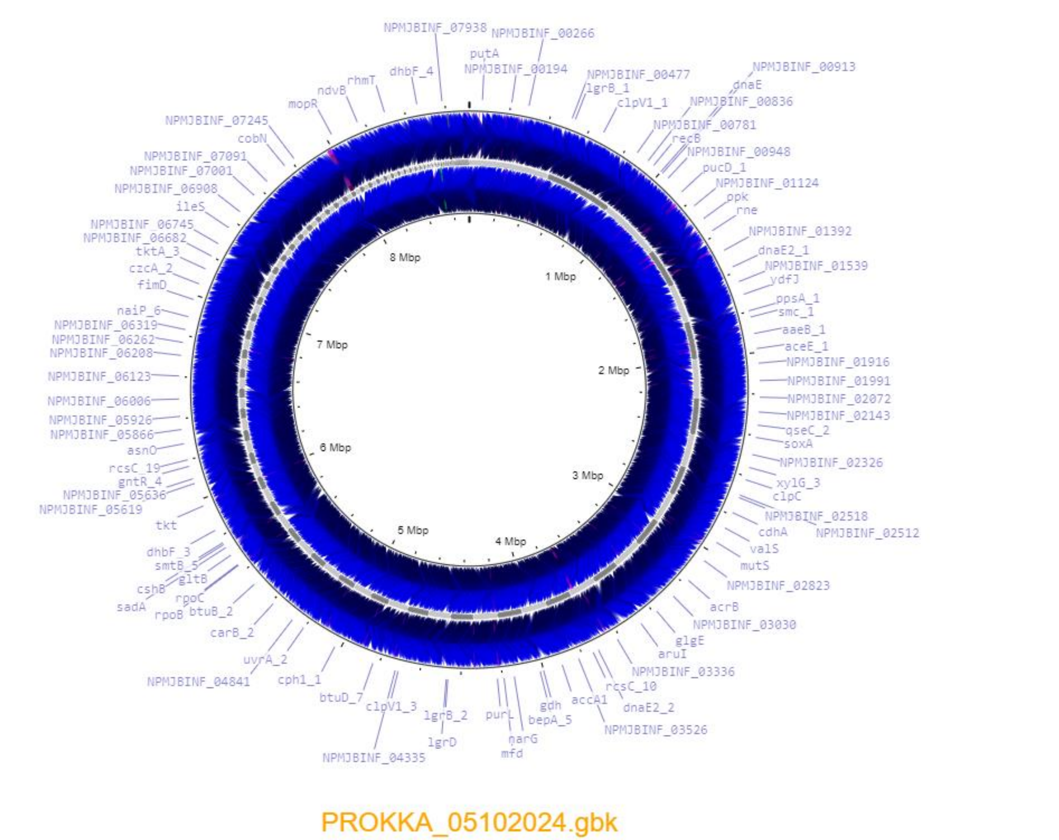


PROKKA_05032024.gbk

Genome Visualization – Ethan



Genome Visualization - Graham



Concluding Remarks

- Our experience showed us that coding takes time and patience

- Some Conda environments did not contain all necessary commands and we needed to download new condas

- BWA mem failed multiple times due to files not being in the proper format we had to find the correct format to run BWA mem on

Bibliography

Grant JR, Enns E, Marinier E, Mandal A, Herman EK, Chen C, Graham M, Van Domselaar G, and Stothard P Proksee: in-depth characterization and visualization of bacterial genomes Nucleic Acids Research, 2023, gkad326, https://doi.org/10.1093/nar/gkad326

Sevigny, Joseph. (2024). MDIBL-T3-WGS-Tutorial. https://github.com/Joseph7e/MDIBL-T3-WGS-Tutorial

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**

THANKS

DO YOU HAVE ANY QUESTIONS?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**

Please keep this slide for attribution