1) From the ABySS output, create a table for the unitigs, contigs, and scaffolds with the number of each, N50 for each, and predicted genome length.

	Number	N50	Predicted Genome Length
Unitigs	135	267313	3885972
Contigs	106	430492	3893279
Scaffolds	87	1094055	3894497

2) https://github.com/bcgsc/abyss This is the link to the documentation for ABySS. In your own words, please summarize the function of each of the commands (e.g., abyss-pe, k, B, etc) that you included in your code.

My code was % abyss-pe name=unknown k=96 B=2G in='reads1.gz reads2.gz'

To start off, ABySS is the name of the program being run with the -pe added to the end to signify it is being used for paired-end reads. I put the name as unknown because at this point I was not yet able to perform a BLAST to identify my species. So, all of my files saved from my ABySS output start with "unknown". The k=96 controls what the k-mer size is. The B=2G controls the amount of my computer's memory that I am allowing ABySS to use. The 2G stands for 2 gigabytes. The in='reads1.gz reads2.gz' is where I imputed my sequencing files for ABySS to use for assembly. The reads1.gz are my forwards reads and the reads2.gz are the reverse reads. These are both fasta files but were renamed. The .gz ending indicates that the files are compressed.

3) Using either output, perform a BLAST search to identify your species. Write your species name here: *Bacillus velezensis*

4) Perform quality assessment using QUAST. You need find a reference genome and reference annotation to upload to QUAST for the best quality check. Which assembler gave you the higher quality output? How do you know?

The ABySS QUAST returned a higher N50 value (1094105 vs 1011820), fewer contigs (19 vs 22), and a bigger largest contig (1238574 vs 1090939) which all indicate to me that ABySS gave a higher quality input. Other values, such as L50 and number of misassemblies were either the same or very similar.

5) Describe what BUSCO is used for. What were the BUSCO values for your assembly?

The BUSCO values are used to describe the completeness of the assembly. There were no BUSCO values for either of my assemblies.

6) Perform a genome annotation using Prokka. Find 3 of the 5 genes/features in your results file and create a table of those results: **recA**, **gyrA,** 16S rRNA, **rpsB, dnaA.**

locus_tag	type	length_bp	gene	EC_number	COG	Product
IFFOGOJP _03613	CDS	2460	gyrA	5.6.2.2	COG0188	DNA gyrase subunit A
IFFOGOJP _01136	CDS	1044	recA		COG0468	Protein RecA
IFFOGOJP _01578	CDS	801	dnaA_1			Chromosom al replication initiator protein
IFFOGOJP _01181	CDS	741	rpsB		COG0052	30S ribosomal protein S2

7) https://github.com/tseemann/prokka Here is the documentation for prokka. In your own words, what is the function of each of the commands in your line of code?

My code was   % prokka --outdir prokkaannotation --prefix bacillus unknown-8.fa

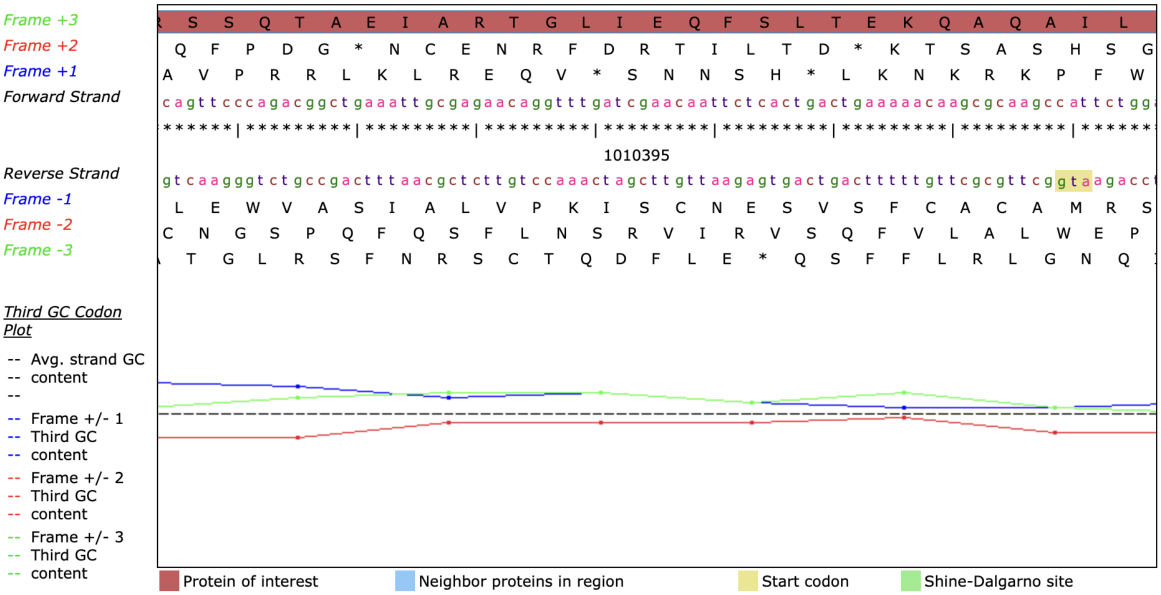
As before the first part of my code is the name of the tool being used, prokka. Next is "--outdir prokkaannotation" which means that the file on my computer where the results will be saved is named prokkaannotation. It is short for output directory will be prokkaannotation. At this point, I had already used blast to identify my species. So, rather than "unknown" I set the beginning of all of my results files to start with "bacillus" with the command "--prefix bacillus". Last, I put the name of the file with the sequencing data to be annotated, "unkown-8.fa". This file came from my ABySS output and is a fasta file.

8) What is the function of the genes/features you chose?

The gene gyrA encodes for a subunit of DNA gyrase, an ATP-dependent Type II topoisomerase. This is an essential component of replication and transcription because it relaxes the supercoiling created by unwinding of DNA. This reduces tension on the DNA strand. The gene recA encodes a protein that is a component of DNA repair. The gene dnaA_1 is a chromosomal replication initiator protein. Initiator proteins play an important role in beginning DNA replication at the origin of replication.

9) Find those same genes/features in your RAST annotation. What information did you learn about them from RAST?

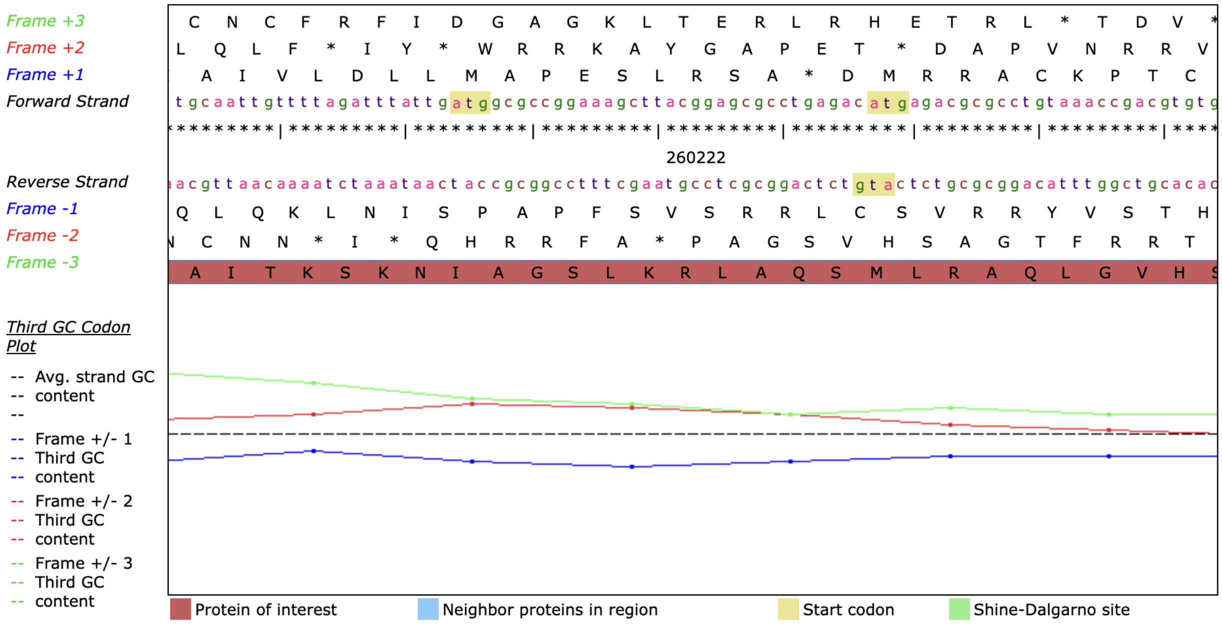
For gyrA the start codon location is 1009165 and the stop codon location is 1011624, confirming the length of 2460 bp. It is part of the following subsystems: DNA gyrase subunits, DNA replication cluster 1, DNA topoisomerases, Type II, ATP-dependent, and resistance to fluoroquinolones. It is contig 445, has a contig length of 3905978bp, a region length of 4000bp, and has a region GC content of 45.400%. It has the following DNA to protein map.



```
Frame +3        R S S Q T A E I A R T G L I E Q F S L T E K Q A Q A I L
Frame +2         Q F P D G * N C E N R F D R T I L T D * K T S A S H S G
Frame +1        A V P R R L K L R E Q V * S N N S H * L K N K R K P F W
Forward Strand  cagttcccagacggctgaaattgcgagaacaggtttgatcgaacaattctcactgactgaaaaacaagcgcaagccattctgg
                ******|*********|*********|*********|*********|*********|*********|*********|******
                                                1010395
Reverse Strand  gtcaagggtctgccgactttaacgctcttgtccaaactagcttgttaagagtgactgacttttttgttcgcgttcggtaagacct
Frame -1         L E W V A S I A L V P K I S C N E S V S F C A C A M R S
Frame -2        C N G S P Q F Q S F L N S R V I R V S Q F V L A L W E P
Frame -3          T G L R S F N R S C T Q D F L E * Q S F F L R L G N Q
```

Third GC Codon Plot
-- Avg. strand GC content
--
-- Frame +/- 1 Third GC content
-- Frame +/- 2 Third GC content
-- Frame +/- 3 Third GC content
🟥 Protein of interest
🟦 Neighbor proteins in region
🟨 Start codon
🟩 Shine-Dalgarno site

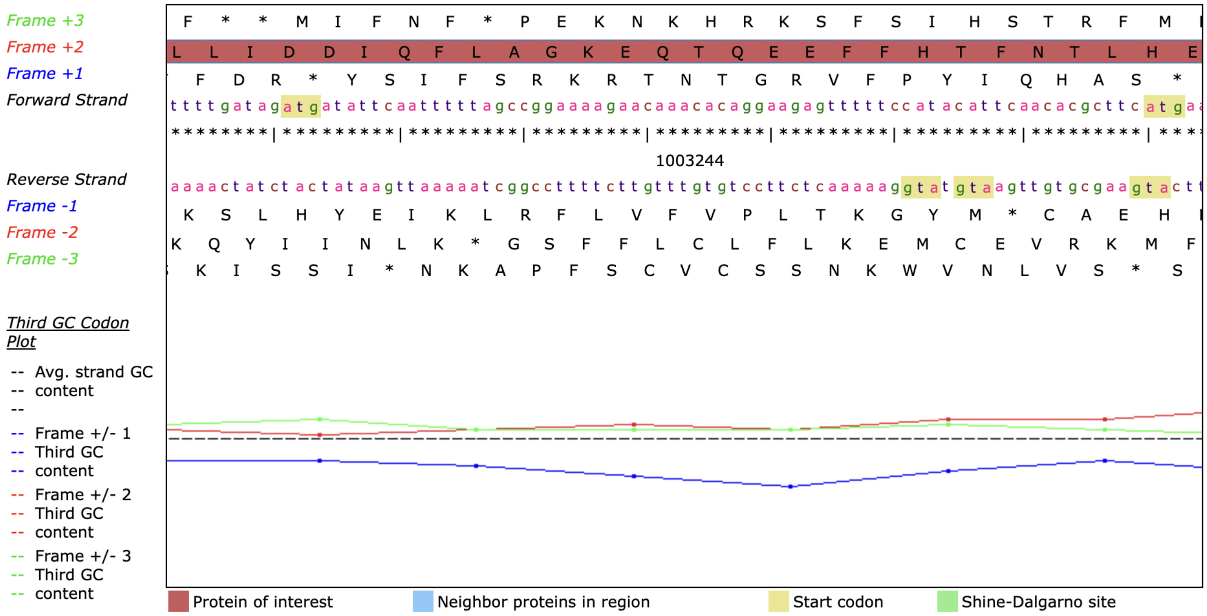
For recA the start codon location is 260743 and the stop codon location is 259700, confirming the length of 1044 bp. It is part of the following subsystems: DNA repair, bacterial, DNA repair, bacterial RecFOR pathway, DNA repair system including RecA, MutS and a hypothetical protein, RecA and RecX. It is contig 443 which has a length of 1,094,105 bp, a region length of 4000 bp, and has a region GC content of 46.025%. It has the following DNA to protein map.



```
Frame +3          C  N  C  F  R  F  I  D  G  A  G  K  L  T  E  R  L  R  H  E  T  R  L  *  T  D  V  *
Frame +2          L  Q  L  F  *  I  Y  *  W  R  R  K  A  Y  G  A  P  E  T  *  D  A  P  V  N  R  R  V
Frame +1          A  I  V  L  D  L  L  M  A  P  E  S  L  R  S  A  *  D  M  R  R  A  C  K  P  T  C
Forward Strand    tgcaattgttttagatttattgatggcgccggaaagcttacggagcgcctgagacatgagacgcgcctgtaaaccgacgtgtg
                  *********|*********|*********|*********|*********|*********|*********|*********|****
                                                     260222
Reverse Strand    aacgttaacaaaatctaaataactaccgcggcctttcgaatgcctcgcggactctgtactctgcgcggacatttggctgcacac
Frame -1          Q  L  Q  K  L  N  I  S  P  A  P  F  S  V  S  R  R  L  C  S  V  R  R  Y  V  S  T  H
Frame -2          N  C  N  N  *  I  *  Q  H  R  R  F  A  *  P  A  G  S  V  H  S  A  G  T  F  R  R  T
Frame -3          A  I  T  K  S  K  N  I  A  G  S  L  K  R  L  A  Q  S  M  L  R  A  Q  L  G  V  H  S
```

Third GC Codon Plot
-- Avg. strand GC -- content --
-- Frame +/- 1 -- Third GC -- content
-- Frame +/- 2 -- Third GC -- content
-- Frame +/- 3 -- Third GC -- content
Protein of interest
Neighbor proteins in region
Start codon
Shine-Dalgarno site

For dnaA_01 the start codon location is 1002574 and the stop codon location is 1003914, which gives a length of 1341 bp. It is part of the following subsystems: DNA replication cluster 1. The function is listed as chromosomal replication initiator protein DnaA. It is located on contig 445 which has a length of 3905978 bp, a region length of 4001 bp, and has a region GC content of 42.089%. It has the following DNA to protein map.



```
Frame +3          F  *  *  M  I  F  N  F  *  P  E  K  N  K  H  R  K  S  F  S  I  H  S  T  R  F  M
Frame +2          L  L  I  D  D  I  Q  F  L  A  G  K  E  Q  T  Q  E  E  F  F  H  T  F  N  T  L  H  E
Frame +1          F  D  R  *  Y  S  I  F  S  R  K  R  T  N  T  G  R  V  F  P  Y  I  Q  H  A  S  *
Forward Strand    ttttgatagatgatattcaatttttagccggaaaagaacaaacacaggaagagtttttccatacattcaacacgcttcatgaa
                  ********|*********|*********|*********|*********|*********|*********|*********|****
                                                     1003244
Reverse Strand    aaaactatctactataagttaaaaatcggccttttcttgtttgtgtccttctcaaaaaggtatgtaagttgtgcgaagtactt
Frame -1          K  S  L  H  Y  E  I  K  L  R  F  L  V  F  V  P  L  T  K  G  Y  M  *  C  A  E  H
Frame -2          K  Q  Y  I  I  N  L  K  *  G  S  F  F  L  C  L  F  L  K  E  M  C  E  V  R  K  M  F
Frame -3          K  I  S  S  I  *  N  K  A  P  F  S  C  V  C  S  S  N  K  W  V  N  L  V  S  *  S
```

Third GC Codon Plot
-- Avg. strand GC -- content --
-- Frame +/- 1 -- Third GC -- content
-- Frame +/- 2 -- Third GC -- content
-- Frame +/- 3 -- Third GC -- content
Protein of interest
Neighbor proteins in region
Start codon
Shine-Dalgarno site

10) Upload the folder of this information to your GitHub in your Bioinfomatics Repository. Please share the link to your repository.

https://github.com/Aswatkins3/Bioinformatics/tree/main/Assembly/GenomeAssemblyandAnnotationSectionReport%20