Assemble your genome using ABySS and SPAdes.

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.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Unitigs** | **Number of Contigs** | **Number Scaffolds** | **N50** | | | **Genome Length** |
| 4124459 | 4127911 | 4128811 | **Unitigs** | **Contigs** | **Scaffolds** | 4128811 |
| 477182 | 477807 | 1024464 |

1. **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.**

* For the command prompt that was used to assemble my genome through Abyss was
  + abyss-pe name=*givenameofjob* k=96 B=2G in= *‘filenames’*
    - Abyss-pe
      * in Abyss run the program with pair ends (pe)
    - Name
      * job name that is given
    - k=96
      * means setting the kmers to a set indicated constraint. Mine was set at 96
    - B=2G
      * The computer will use 2 gigabytes of ram while running the assembly.
    - in=
      * Searches for the file names in that specific folder.

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

* After running a Nucleotide Blast I discovered my species is Bacillus *subtilis.*

**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?**

A screenshot of a report

Description automatically generatedA table with numbers and letters

Description automatically generated

* When comparing the two genome assemblers using QUAST abyss is a much more accurate and higher quality assembler based on the N50 and L50 number highlighted. The bigger the N50 number the better overall quality the assembled genome is because a N50 number indicates that more than 50% of the assembled genome is that size or bigger. The L50 should be low. In both assembles the number is 2 which means with all other data needed for assembled genome the better assembler is Abyss.

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

* BUSCO has the ability to show how complete a genome is based on the data that is entered. My BUSCO values for my assembly was 100%.

**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.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| Prokka Annotation | | | | | |
| Gene Symbol | **recA** | **Gyra** | **16s** **rRNA** | **rpsB** | **dnaA** |
| Length bp | 1047 | 2466 | NA | 741 | 636 |
| Locus Tag | EKMHDMFA  01518 | EKMHDMFA  04272 | NA | EKMHDMFA  01475 | EKMHDMFA  01066 |
| Product | Protein RecA | DNA gyrase subunit A | NA | 30S ribosomal protein S2 | Chromosomal replication initiator protein DnaA |

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

The line of code that I entered into my command prompt was

prokka --outdir prokkaannotation --prefix Bacillus\_subtilis abyssoutput\_April9-8.fa

* Prokka
  + Command Prokka calls the program in the genome assembly.
* --outdir
  + Creates a annotation using prokka program
* --prefix
  + When crating the annotation it gives it a name that is chosen by programmer and looks for file that is indicated in line of code to annotate.

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

**Genes featured in my species are: recA, gyra, rpsB, dnaA**

* recA
  + recA is responsible for protein RecA.
* gyra
  + gyra uses DNA gyrase subunit A.
* rpsB
  + 30s ribosomal protein S2 is required to grow in temperatures below 8 degress Celsius.
* dnaA
  + chromsosal replication initiator protein DnaA.

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

* recA
  + A DNA repair system that includes RecA, MutS and a hypothetical protein. RecX, which is known to regulate RecA activity is often included in the cluster.
* gyra
  + Gene gyra has two subunits. Gyra subunit A and B—both of which have a function of DNA replication cluster located on the positive strand contig 381.
* rpsB
  + Also knonw as ribosomal protein S2 is a highly conserved in all forms of life and its counterparts are referred to S0 in yeast and SA in higher eukaryotes. It is essential for all translational machinery in all prokaryote, eukaryotes, mitochondria and chloroplasts. The overall function is still unclear.
* dnaA
  + DnaA is a chromosomal replication intiatiator protein located on the positive strand on contig 381.

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

<https://github.com/walkercano/Genomeassembly_Bacillus_subtilis>