Samantha Prince

Genome assembly

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.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Number | N50 | Predicted Genome Length |
| Unitigs | 6863 | 434 | 3062631 |
| Contigs | 6732 | 461 | 3093161 |
| Scaffolds | 6630 | 471 | 3087839 |

2) [https://github.com/bcgsc/abyssLinks to an external site.](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.

N/A. Command line not working no commands used in code. All information obtained through galaxy.

3) [https://ablab.github.io/spades/index.htmlLinks to an external site.](https://ablab.github.io/spades/index.html) This is the documentation for SPAdes. Based on this manual, can you identify how you could modify the code you used to do a hybrid assembly with nanopore reads? Please explain what a hybrid assembly is and why someone might want to do that.

To modify the code to do a hybrid assembly with nanopore reads you would add

--nanopore in the command line after your file names and before spades output

You would also add hybrid to the spades output command…

Spades\_hybrid\_output

A hybrid assembly combines short and long reads in the genome, this can be used to get a more complete genome with less contigs and scaffolds.

4) Include a screenshot of the QUAST assembly statistics for the ABySS and SPAdes assembly.

5) Based on the statistics from your genome, which assembly do you think is best? Why? This is the assembly you can use going forward.

Based on the assemblies I believe spades to have the best results. While Abyss has less contigs and a better continuity of information the total genome length is much less than the results of Spades. Indicating spades gives a much larger and more complete assembly of the genome.

6) How can we use barrnap to figure out what species we have? Why is using the 16S rRNA sequence a good, but imperfect, tool for identifying species identity?

Barrnap is used to scan genome assemblies for rRNA genes including the 16S rRNA gene. This gene is highly conserved in bacteria and there is a large database available to compare 16S rRNA with other species making it a good sequence to use for identification purposes but not perfect because this may be present in many bacteria of the same genus making it hard to narrow down the specific species.

7) What species do you have? Include a screenshot of your top NCBI results.

Rhodococcus gingshengii

8) What is genome annotation? Why is it important to do that?

Genome annotation is taking a genome sequence determining its scaffolds, contigs, and other useful coding information, then using this to identify what proteins are coded in the sequence and where they can be found.

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

What is the location of the genes you chose? What does each program tell you about the gene? How are the outputs different between the two programs.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Prokka Results | | | | | | | |
| gene | locus\_tag | ftype | length\_bp | EC# | COG | Gene Description | |
| gyrA\_1 | MNHNOIJP\_00035 | CDS | 1881 | 5.6.2.2 |  | DNA gyrase subunit A | |
| rpsB | MNHNOIJP\_00121 | CDS | 783 |  | COG0052 | 30S ribosomal protein S2 | |
| recA\_1 | MNHNOIJP\_00129 | CDS | 1098 |  | COG0468 | Protein RecA | |
| RAST Results | | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

I was unable to get results from RAST as I never received an email from them allowing me to set up my account, I reached out to them several times starting on Monday but no luck. When annotating the genome Prokka provides the following information about each gene. It provides the genes common name and its product. A locus\_tag which is a unique identifier for each gene, ensuring consistent tracking across the genome. The ftype (feature type) describes what kind of genetic element it is, such as a coding sequence (CDS). EC\_number is the Enzyme Commission number, which is the specific biochemical reaction the gene’s enzyme product catalyzes. Finally COG number tells the evolutionary relationship and Length\_bp is the length of each gene in base pairs.

10) Create a table for your ANI results. How do you interpret these results? What do each of the columns represent? Hint: You can refer to the original paper to find that information.