**TRACK 1 - GENOME SKIMMING.**

**ASSIGNMENT 2 – DATA ANALYSIS**

1. You will have to transfer five (5) files from the Professor’s cloud server (azureuser@monicaoct14.cloudapp.net / Password: Cluster2014 / Directory: Assignment) into your own cloud server instance (i.e. Virtual Machine in Azure) using sftp.
2. Study your dataset with clean reads in the cluster using the Terminal application
   1. How big is this dataset? (in GB)
   2. How many reads do you have in this file? (Hint: Use grep command)
   3. What is the index that you used in your library preparation?
3. Study your dataset with clean reads using FastQC GUI. Download your cleaned reads dataset to your local computer using a sftp connection, and then open your file in FastQC.
   1. Are your reads of good quality? (FastQC)
   2. What is the range of quality scores per base?
   3. What is the most common quality score per read
   4. How long are your reads?
4. Try to find the sequence of the adaptor (GATCGGAAGAGCACACGTCTGAA) in your cleaned reads dataset using the “grep” command.
   1. How many reads did you find with the adaptor still included? (Hint: When running a command in the Terminal, you can redirect any output files by adding the command “>” and the new file name at the end)
   2. By comparison to your entire cleaned dataset, are you OK with the quality of your reads?
5. Now look at your reference-based results (from YASRA) in Sequencher. Upload the file named FinalAssembly.
   1. How many contigs did YASRA find?
   2. How long is the longest contig?
6. Assemble together all contigs from YASRA in Sequencher (using a high percent similarity).
   1. Did you improve on your data?
   2. How long is your biggest scaffold (contig of contigs) now?
   3. Would you consider keeping this YASRA contigs or just deleting them? Não deleta.
7. In Sequencher, delete all YASRA contigs and now upload your Velvet contigs only.
   1. How many contigs did Velvet obtain? 52
   2. How long is your longest contig? 18412
   3. What is the average coverage of the chloroplast contigs? 28 -30
   4. How many mitochondrial contigs do you have in these results? 27
   5. What is the average coverage of the mitochondrial contigs? 3 - 4
8. After deleting all possible mitochondrial contamination. Assemble all your putative chloroplast contigs together using Sequencher.
   1. How many scaffolds (contigs of contigs) did you get? 9
   2. How long is your longest scaffold? 31007 bps
   3. What genes are present in your longest scaffold? No archive wev
   4. What part of the chloroplast genome does the longest scaffold represent? (i.e. Large Single Copy, Small Single Copy or Inverted Repeat) Inverted repeat
   5. How many single contigs (not assembled) do you still have? 3
   6. Is this an improvement over your YASRA analysis? sim
9. Using the “grep” command try to put together as many scaffolds as you can, until you get approximately 80000 bps of your chloroplast genome together
   1. How many of the original Velvet contigs were you able to piece together?
   2. What part of the chloroplast genome does this piece represent? (i.e. Large Single Copy, Small Single Copy or Inverted Repeat)
   3. How many genes do you have in this section of the genome?
   4. List three intergenic regions you found in this piece
   5. List two introns you found in this piece
10. Now, let’s look at the full chloroplast genome sequence that another researcher put together using the same raw reads dataset that you have (Assignment\_cpGenomeFINAL.fsa)
    1. Upload this dataset in Sequencher, and try to align your longest Velvet scaffolds and contigs to it.
    2. How many scaffolds or contigs were you able to assemble to the genome?
    3. What are the regions in the genome that do not have any supporting Velvet contigs attached to them? (Hint: you can give the positions, in bps, that are not covered by Velvet)
11. Open the results of the Jellyfish coverage analysis (problem\_areas\_file). As you can see, there are lots of mistakes (zero coverage regions in this genome).
    1. Find two small (less than 40 bps) regions with mistakes (i.e. zero coverage). [For example: 78543-78552 bps]
    2. By using “grep” around the flanks of these regions, try to reconstruct the correct sequence in the genome, according to your cleaned reads dataset.
    3. Were you able to recover the correct sequence?
    4. What was the error found?
12. Look now at position 5457 in the Jellyfish problem areas output.
    1. Why is this area problematic?
    2. Is it a real error in assembling the chloroplast sequence? Or is it a biological reality?
13. Look now at position 114361 in both Jellyfish output files (the problem areas file and the total coverage file)
    1. Can you find this position in both files? Why? or Why not?
    2. How many times is that particular 20bps sequence repeated in the chloroplast genome?
    3. Is this an error in the assembly process or is it biological reality?
    4. What could be the utility of this type of sequences for future studies?
14. Now open your Jellyfish total coverage output file in Excel [Hint: Use the import data function]
    1. What is the sequence of the most commonly found repeat?
    2. How many times was it found?
    3. In what positions was this repeat found?
    4. How many repeat regions (i.e. with overly high coverage, found more than 300 times) can you find in this genome?

Respostas

2. a) 4988081832 aprox. 5 GB

b)reads = 20843884

c)Index = TGACCA

4. a)6554 reads

b) não é muito boa a qualidade, de 20.843.384 tem 6554 reads contaminadas.

5. a) How many contigs did YASRA find? 133 contig.

b) How long is the longest contig? 6285 bps

6. Did you improve on your data? Acho sim, porque tem agora só 6 contig.

1. b) How long is your biggest scaffold (contig of contigs) now? 6497 Bps