Bi 621 – Problem Set 6

Our goal with this assignment is to generate some measures of accuracy for whole genome assemblies and then use these measures to assay our success in several Velvet assemblies.

Part 1 – Contig length distributions

1. Parse the `contigs.fa` file that is output by `velvetg`. Extract the FASTA ID lines as you parse the file (remember: these strings will begin with the ">" character).
2. You can use the sample data `contigs.fa` from HPC to test your code.
3. Using Python regular expressions, extract k-mer length of each contig (in **red** below). In addition, extract the k-mer coverage for the contig (in **blue**).
   `>NODE_11_length_3717_cov_19.315845`
4. Adjust the k-mer length to represent the physical length. Calculate the number of contigs, the maximum contig length, the mean contig length, and the total length of the genome across the contigs. Calculate the mean depth of coverage for the contigs.
5. Calculate the N50 value of your assembly.
6. Calculate the distribution of contig lengths, and bucket the contig lengths into groups of 100bp. So, all contigs with lengths between 0 and 99 would be in the 0 bucket, those with lengths between 100 and 199 would be in the 100 bucket, etc.
7. Print out the distribution.

```
# Contig length   Number of contigs in this category
0     0
100   5324
200   3345
300   1130
...
```

Part 2 – Velvet

You will need to install Velvet in your home directory on HPC to use it.

- On HPC, in your home directory:
  `wget https://www.ebi.ac.uk/~zerbino/velvet/velvet_1.2.10.tgz`
- In your home directory, make a directory called `bin` (if you don't already have one)
- Move `velvet_1.2.10.tgz` into `~/bin`
- Untar the file
- Move into the `velvet_1.2.10` directory
- Issue the command `make 'MAXKMERLENGTH=50'`
- Move `velveth` and `velvetg` up one directory (into `~/bin`)

1. All your work in this section should be completed using the queuing system on HPC. (See Nick's lecture notes on HPC and https://hpcrcf.atlassian.net/wiki/display/TCP/How-to+Submit+a+Job to remind yourself how the queuing system works)
2. You can find the data here:
   `/projects/bgmp/Bi621/rs_female_1983.13.1.fq.gz`
   `/projects/bgmp/Bi621/rs_female_1983.13.2.fq.gz`

Please do not copy the data, but rather refer to its location in your script. Remember to NOT write to the group project directory.

3. Run Velvet on the dataset. You can find the Velvet manual here: https://www.ebi.ac.uk/~zerbino/velvet/Manual.pdf
    a. Calculate the k-mer coverage for the dataset assuming a stickleback genome size is 460 Mb
    b. Run velveth/velvetg with k-mer sizes of 31, 41, and 49
    c. Use your code to collect assembly statistics on each result
4. With a k-mer size of 49, adjust the coverage cutoff to 4x, 8x, and 'auto'. Again, assay your results with your code.
5. Finally, adjust the minimum contig length to 500bp and again, assay your results.

Part 3 – Questions

1. Describe how the assembly changes with different k-mer values using the assembly statistics you have collected. How does the contig length distribution change?
2. How does an increased coverage cutoff affect the assembly? What is happening to the de Bruijin graph when you change the value of this parameter? How does velvet calculate its value for 'auto'?
3. How does increasing minimum contig length affect your contig length distribution?

To turn in your work for this assignment, do the following:

Be sure to turn in your code, your output (mean, max, N50, etc) and plots, as well as the answers to the questions above to GitHub.