Ryan Johnson

Instructions on how to run my Nextflow pipeline.

There are three files that need to be placed in a directory

1. Nextflow.config
2. Final\_project.nf
3. Run\_my\_nf\_process.sh

Since we are on the hpc you do not need to change any of the paths to the location of the week2\_fastqs.txt, reference genome location, or directory that contains all of the pair end read files. I used the paths that were provided to the class, so anyone that has access to these files for this class can run this pipeline on the HPC without worry.

If you have your own pair end reads, and reference genome you would like to run, please look through the three files using nano text editor and update any path locations to the paths of your corresponding files and directories.

The week2\_fastqs.txt file contains the names of the pair end reads and also the ID’s. So if you have your own pair end files please make a .txt file that has the ID and the two pair end reads one the same line, all three separated by a tab to create three columns. Each line will be a new sample with the ID in the first column, forward read name in the second column, reverse read name in the third column. Then look in the run\_my\_nf\_pipeline.sh file and change the name from ‘week2\_fastqs.txt’ to whatever you named the .txt file you just created.

In the run\_my\_nf\_pipeline.sh file, we are making an array job based on how many samples you want to analyze. This file has an array set to 1-3, for the purposes of the project to show that it can run 3 samples successfully, but you can change it to run only 1 or run how many samples (rows) you have in your .txt file. If you have ten just set the array to equal 1-10. It’s that simple.

Now we just run the command sbatch run\_my\_nf\_process.sh.

This will create 3 instances of the file, each uses its own number from 1-3. So this will choose that number row from the .txt file and will pass the values generated into the ‘final\_project.nf’ script. Give the full pipeline a total of 3 to 4 hours to run. However one of the jobs ends up finishing after about 1 hour 10 minutes, so you can check back in to look at that first one.

To find the generated files look in the slurm.out file for that job and you will see process, the name of my process, and its completion at 100 percent. For the third process, at the beginning of the line will be the directory the output will be located in. Just copy that. First you need to cd to the work directory where everything is located, then paste the two directories you just copied but also press tab on the second directory since it is just a small part of that directory’s name. Now hit enter and you’re in the directory with the .bam and .bai files, which shows the pipeline worked. To see the outputs for the previous two processes of that job just get the directory’s name the same way from the slurm.out file and repeat the following steps.