**MCB 525, Fall 2016**

**2.4: Multi-sample processing**

Now that you are well versed in the art of read quality control, read mapping, and genotype calling, your next assignment is to work through these steps with 48 additional samples! For this exercise, we will work in pairs, and **each pair will be responsible for creating a genotypes.tab file (through Step 2.3b from yesterday) for 4 samples.** To help complete this task, we will be learning about a very useful functionality of SGE\_Batch.

**Objectives:**

* Scale up bioinformatic analyses to multiple samples

**Hint:**

In lab exercises from the past few days, we have submitted jobs through SGE\_Batch by specifying each individual file name. It is also possible to submit multiple jobs at once, using the clever matching variable $SGE\_TASK\_ID. For example, if I had a group of 10 fasta files for samples named Arthur1 through Arthur10, I could submit the following job to run the program for all 10 samples, where the first argument was the input file and the second argument the output file:

$ SGE\_Batch –c ‘./myscript.pl Arthur$SGE\_TASK\_ID/Arthur$SGE\_TASK\_ID.fa Dent$SGE\_TASK\_ID.txt’ –r hg –t 1-10

#This script would submit 10 separate jobs, substituting ‘$SGE\_TASK\_ID’ with each number 1-10. For the first job, the script would read in the file Arthur1.fa contained in the Arthur1 directory and create an output file called Dent1.txt, which would be stored in the current working directory (the directory you were in when you submitted the job).

#Don’t be a queue hog! If you submit 1000 jobs (e.g. t 1-1000), the scheduler will submit all of them before your poor lab mate, or anyone else using the CGRB infrastructure for that matter, can get a job onto a machine.

**Protocol:**

1. Create separate directories for each of the 4 samples you have been assigned.
2. Copy gzipped read files into each folder.
3. Unzip each read file.
4. Trim the last 2 bp from each read to create 36 bp reads from 38 bp reads using fastx\_trimmer.
5. Filter the read files to remove all reads where <83% of the read has quality <26 with fastq\_quality\_filter.
6. Map reads to the BcgI reference with gmapper (you should already have the BcgI reference prepared).
7. Filter the resulting alignment to exclude weak or ambiguous matches using SamFilter.pl.
8. Count the nucleotides at each position with SAMBaseCaller.pl.
9. Call genotypes at each position using NFGenotyper.pl.

At this point, you will have genotype files for 4 different samples, plus one unknown sample (the first sample you processed). Flag down an instructor or TA to help you check these files and copy them to the appropriate directory in the class folder.

Once all groups have copied their data to their respective folder, all the genotypes will be combined into one file, along with genotypes from an additional 48 samples processed previously, using CombineGenotypes.pl. The file will then be filtered for polymorphic sites with PolyFilter.pl, another script you used yesterday. Finally, we will exclude rare variants that are polymorphic in <5 individuals using a script called MinPolyFilter.pl and remove loci with an excessive number of missing genotypes with the LowcovSampleFilter.pl. These latter two scripts you have not used, but they should be available in your scripts folder. Run the script without any arguments if you are interested in learning more about them.

In our next class meeting, we will take the file of combined genotypes and convert it to a format suitable for input to the program STRUCTURE.