Homework 8

Instructions. This homework is due at the beginning of lab, Wednesday, November 1st. When executing the commands of this homework on the `hpc-class` cluster, please use the `salloc` or `sbatch` commands. We may get in trouble if we just type the commands at the command prompt of the head node (the node you log onto). In addition, every command in this lab/homework is "real" bioinformatics, meaning they take real time and things can go wrong. (I myself have had to do every step about three times to get it to this point. Do not assume it should just work. Only thourougly pre-chewed, roughage-removed, watered-down homeworks just work, and this one is not one of those.)

1. Your first step is to download some data from the Short Read Archive.
   (a) Navigate to the website and enter the experiment accession `SRX026594`. What machine was used to generate these sequence data? When were the data generated? Note the run accession(s) associated with this experiment.

   (b) It is no longer possible to download the data using the web interface. Instead, NCBI provides the sratoolkit command-line tools. These tools are installed on `hpc-class`. Make them available by issuing the command `module load sratoolkit`. Read this helpful blog about paired-end reads, the help for the `fastq-dump` command, and explain what the `--split-files` command-line option does.

   (c) Using the run accession you noted in Part a, issue a `fastq-dump` command that uses option `--split-files` and compresses the file using gzip technology. Since our `hpc-class` home directories are limited to 5Gb of storage and this is big data, you may encounter your disk quota while attempting this lab/homework. You can create a personal directory in `/ptmp` and store the files there. You will need to check the `fastq-dump` help to get it to dump the data in this directory. Please copy and paste the command you used below.

(d) Since these data are quite old, we should investigate the quality score encoding, which has changed over the years. The Wikipedia entry claims that SRA has converted all quality scores to the Sanger standard, but there is no citation provided for this claim, and I have only found confirmatory rumors.

i. On the same Wikipedia page, there is a helpful diagram of the different ASCII quality score encodings. Can you use this diagram to assess whether the quality scores are in standard Sanger format? Explain how you came to your conclusion.

ii. In the same **Encoding** section of the page, it states that Illumina, for some platforms, used Phred scores 0 through 2 to communicate something other than probability of error. Can you determine if these data are subject to this special encoding? Explain how you came to your conclusion.

2. The data you downloaded are reads of the roundworm *C. elegans*. WormBase is a website repository of everything *C. elegans*, including the reference genome. Navigate to the public ftp site and drill down to *C. elegans* genome for the reference strain Bristol N2. (PRJNA# accession numbers identify BioProjects, which you can look up and learn more about in NCBI's BioProject Database.) Once you find the file you need, download it to `hpc-class` using the Linux command `wget`. How many chromosomes does *C. elegans* have? How long are each of the chromosomes in the current version of the reference genome? (**Hint**: remember `BioPython`)

3. We will align the reads to the reference genome using the software `bwa`, which becomes available if you type `module load bwa`.

   (a) The first step is to build an index for the reference genome. You do this with the `bwa index gzipped_fasta_file`, where `gzipped_fasta_file` is the reference genome you downloaded in Part **??**. How much disk space does the index occupy compared to the gzipped fasta file?

   (b) The next step is to align the reads to the reference genome. This step can take a long time, but the `hpc-class` cluster has 16 processor cores per node that can be utilized to speed up `bwa`. We will use the `bwa mem` command to align the reads. Read the help (just type `bwa mem`) to find out how to request 16 threads and provide the index and fastq files to `bwa`. Also, you should use the `-M` flag in order for the results to be compatible with other steps in this Lab/Homework. Note, the output is a `sam` file that is delivered to `stdout`: you should redirect it to a file ending in `sam`. Read about the SAM format and hard vs. soft-clipping. Record the command you used below, randomly select one aligned entry in the SAM file with at least one mismatch and show its local alignment with the reference genome below.

(c) Now we need to convert the SAM file to a binary BAM file. Issue these commands (please note some can be threaded):

```
samtools view -b samfile > bamfile # convert SAM to BAM
samtools sort -o sorted_bamfile bamfile # sort BAM file (along genome)
samtools faidx reference_fasta_file # create an index of the reference genome
samtools index sorted_bamfile # create an index of the BAM file
```

To view the result, you can use `samtools tview sorted_bam_file`. (Note: You may prefer using the Integrative Genomics Viewer (IGV), but this graphical interface is not available on the `hpc-class` cluster. You may be able to install it on your local computer. To use it, you need all the same files that `samtools tview` needs.) Wander around until you find evidence of a sequencing error. Take a screenshot of the evidence and explain why you believe it to be a sequencing error.

4. This question is about read depth.

(a) The coverage of a sequencing experiment is the expected number of reads covering each genome position under the assumption of random fragmentation. Ignoring the effects of chromosome ends, explain why coverage can be estimated as

$$\frac{LN}{G},$$

where L is the read length, N is the number of reads, and G is the genome length. What is this estimated value for the *C. elegans* data? What is the predicted probability that a particular base is not covered by any read? (Show your work.)

(b) The actual coverage will be lower because of read errors, clipping, and alignment problems. We can estimate the average coverage per base using `samtools` with the command `samtools depth sorted_bam_file` combined with `awk`. Report the result.

5. This question is about variant discovery.
   (a) In this section, we will look for Single Nucleotide Variants (SNVs).
      i. Typically during library preparation, there are amplification steps (before the bridge amplification that occurs on the Illumina flow cell) where single molecules are copied many times. As a result, the same original molecule can be represented by multiple reads. Many pipelines remove the duplicated reads. Why are datasets deduplicated (dedupped)? Why would it *not* be a good idea to remove duplicate reads in amplicon sequencing of a metagenomic sample?

      ii. You can deduplicate with the `samtools rmdup`. How many reads were removed during deduplication?

      iii. Use `samtools mpileup` to create a `.bcf` file, which is a binary version of the Variant Call Format (vcf). You can convert to `.vcf` format using `bcftools view` (available with `module load bcftools`) to see the data in text format. Select one interesting variant and explain why it is listed as a variant.

      iv. Finally, you can call SNVs that are likely to be true variation in the sequenced individual relative to the reference genome using `bcftools call`. Output the results in `.vcf` format, so you can easily work with the results. How many SNVs do you call? Which calling method did you use? Why is it appropriate?

6. In this final part, you will compare the SNV calls between two methods.

   (a) Repeat Part iv with the other calling method. **Bonus version I.** Repeat the analysis using the GATK pipeline, through the Variant Discovery step only. You can access the GATK tools with the command `module load gatk`. (**Hint:** You probably need to `locate` the necessary `.jar` files in the filesystem and explicitly name them in the GATK pipeline commands.) **Bonus version II.** Repeat the analysis after filtering the reads for lower quality (see `module fastx-toolkit` and consider that quality scores less than 30 are often considered of too low quality). **Bonus version III.** Repeat the analysis after using the `bowtie2` aligner (`module load bowtie2`).

   (b) The `bedtools` suite of software, available after you issue `module load bedtool2`, provides methods to compare `.vcf` files. Use `bedtools intersect` and `bedtools subtract` to prepare a Venn diagram showing how many SNVs the two methods both found and how many each uniquely found.