Problem Set 4 - Next-generation sequencing

This homework should be sent to mgymrek@ucsd.edu with subject line [CSE291 PS4:LASTNAME] by the beginning of class on **Tuesday, February 28**. The assignment is worth 10 points total.

As in previous problem sets, template code is provided for some problems. Using the template code is optional, it is simply there to guide you.

For all plots, please include the plot in the writeup. For code, please paste the relevant snippets that you wrote into the writeup.

Objectives

- Gain experience using standard short read alignment, genotyping, and visualization techniques.
- Learn how to detect artifacts in alignment or variant calling that can arise from next generation sequencing analysis.
- Explore new long read sequencing technologies.

A description of data files for this problem set and several setup steps can be found at PS4 resources.

Part 1: Sequence alignment and visualization (4 points)

Overview

In this problem, we'll get comfortable dealing with standard next-generation sequencing file formats and command line tools. We'll be analyzing the genome of sample NA12878, perhaps the most sequenced genome in the world and a commonly used standard for evaluating variant callers. We'll be primarily working with samtools. Some example samtools commands, and other helpful command line tips, are given in the slides for lecture 11.

In the ps4 data directory, you'll find sequence alignments for NA12878 in CRAM format, aligned to the GRCh38 (hg38) reference genome, along with accompanying index files:

```
NA12878.alt_bwamem_GRCh38DH.20150706.CEU.illumina_platinum_ped.cram
NA12878.alt_bwamem_GRCh38DH.20150706.CEU.illumina_platinum_ped.cram.crai
```

You will also find the hg38 reference genome:

```
GRCh38_full_analysis_set_plus_decoy_hla.fa
```

The SAM Specification will provide a helpful reference for how to interpret SAM, BAM, and CRAM files. We'll also talk extensively about the different file formats in class.

Exercises

1. **(0.5 points)** Describe and draw the difference between "single-end" and "paired-end" sequencing reads.

2. **(1 point)** Read alignments for sample NA12878 are available in CRAM format at the path specified above. You can use samtools to view all reads, or only reads in certain locations. e.g.:

   ```
   samtools view $CRAMFILE chr1:949140-949150
   ```

   Note, this command will only work on CRAM (or BAM) files that have been sorted and indexed (this one has been already). Let's take a look at some reads:

   - For read "ERR194147.1761673", what does a CIGAR score of "23M4D78M" mean?
   - For read "ERR194147.1761658", what does a CIGAR score of "101M" mean?
   - At what position do the two mates with ID "ERR194147.1761677" map?

3. **(1 point)** Now we will do some quality controls on the resulting alignment:

   - Plot a histogram of the coverage per base for the chromosome 1, chrX, and chrY. Is this a male or female sample? How do you know? How can you explain that some reads were mapped to chromosome Y?

     Hint: samtools mpileup command can be used to pull out the coverage for each base, e.g.:

     ```
     samtools mpileup -r chr1 $CRAMFILE | cut -f 4
     ```

     will give you the coverage at each base in chromosome 1 only. See the lecture 11 slides for helpful UNIX tips in dealing with this data.

   - Plot a histogram of template length for each mate pair mapped to chromosome 1. Explain what this number means. What is the average? Standard deviation? Do you notice any artifacts in your histogram? Hint: what's going on with pairs with a template length "0"?

     Hint: the 9th column of the CRAM file gives the template length:

     ```
     samtools view $CRAMFILE chr1 | cut -f 9
     ```

4. **(1 point)** Use samtools tview to visualize the resulting alignment:

   ```
   samtools tview $CRAMFILE $REFFASTA
   ```

   Navigate to positions chr1:926250, chr1:1035844, and chr1:1056421. Do these look like heterozygous or homozygous SNPs or indels? Now navigate to chr1:1129961. You'll find this region is pretty messy. Hypothesize what's going on there.

5. **(0.5 points)** Use samtools view to look at reads that could not be successfully aligned to the reference genome:

   ```
   samtools view -f 4 $CRAMFILE | less -S
   ```

   Hypothesize possible reasons these reads couldn't be aligned.

   Now look at a read that has many alternate mapping locations listed (see the XA tag). For instance:

   ```
   samtools view $CRAMFILE | grep "ERR194147.755144932"
   ```

   Hypothesize where this read is coming from, and why it might map to multiple locations in the genome.

Part 2: Writing a simple SNP caller (4 points)

Overview

We'll be working with the same data files as in the above problem, namely the sequence alignment in CRAM format and the human reference genome for build hg38. We'll also use a previously published gold standard set of SNP calls:

```
HG001_GRCh38_GIAB_highconf_CG-IllFB-IllGATKHC
-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf_PGandRTGphasetransfer.vcf.gz
```

This has been converted into a more convenient tab-delimited format here:

```
NIST_NA12878_hg38_chr22.tab
```

In the ps4 templates directory, you'll find template scripts for implementing and running your SNP caller, as well as a script to compare your SNP calls to previously published calls.

```
ps4_snpcaller_template.py
run_ps4_snpcaller.sh
ps4_comparesnps_template.py
run_ps4_comparesnps.sh
```

Exercises

1. **(0.5 point)** Convert sequences from chromosome 22 of the BAM file of your sequence alignment from the last problem to samtools "pileup" format. Describe the resulting format. How big is this file?

   This example command can get you started:

   ```
   samtools mpileup -r chr22 -f $REFFASTA $CRAMFILE > $RESULTS/NA12878.chr22.pileup
   ```

2. **(1.5 points)** Write a SNP caller that takes in the pileup format and outputs a list of putative SNPs. To work with templates for the next part, your script should output columns: chromosome, position, reference allele, alternate allele, coverage, score, alternate allele frequency, genotype (allele 1), genotype (allele 2).

   Here is an example output (numbers are fake though):

   ```
   chr22 200 G A 100 5.2 0.45 A G
   ```

   To simplify things, you can restrict to sites with no insertions or deletions. You may also ignore sites with more than 2 alleles present. The template file ps4_snpcaller_template.py is given to get you started. Look in that file for the place you should fill in with your SNP caller code. The file run_ps4_snpcaller.sh shows how to run this script, and already deals with filtering indels.

   Briefly describe your method. Note, you can define a score however you want, as long as you describe what you did and it's reasonable. How many homozygous reference, heterozygous, and homozygous non-reference sites did you find?

   NOTE: if your solution beats mine (shouldn't be that hard to do) for overall accuracy, you'll get 1 point extra credit!

3. **(1 point)** Compare your results to those obtained by the NIST gold standard. Based on this comparison, what is your accuracy rate overall? At heterozygous SNPs? At homozygous SNPs? How does your accuracy change if you restrict to loci covered by at least 1 read? 5 reads? 10 reads? 20? Is your score correlated with accuracy? You may find the template run_ps4_comparesnps.sh and ps4_comparesnps_template.py helpful. These scripts are mostly complete, but you may want to tweak them e.g. based on your score.

4. **(1 point)** Examine, for instance using samtools tview, 3 cases of SNPs you got "wrong". Hypothesize where your SNP caller went wrong, or whether you think you are actually correct! If you use tview, include screenshots in your writeup.

Part 3: Long read technologies (2 points)

Overview

Exercises

1. **(0.5 points)** 10X Genomics has a freely available browser: http://loupe.10xgenomics.com/. Navigate to the whole genome of NA12878. In the "structural variants" view, navigate to position "chr7:54,202,001-54,402,001;chr7:54,289,009-54,489,009". Describe what is being displayed by the heatmap plots. What seems to be going on at this locus? What about chr3:63,850,233-63,989,138?

   Now, for the chr3 locus, go to the "linked reads" view. Hypothesize why, or investigate using loupe or another means, why in several places there are large piles of gray (unphased) reads that could not be confidently phased onto either haplotype?

2. **(0.5 points**) You are designing a 10X experiment. You have available 1.2 million barcodes. You would like to use enough input DNA such that each barcode will correspond to at most 10 molecules. How many molecules should you load onto the sequencer?

3. **(0.5 points**) The sequencer will generate 250 million reads that are 100bp each. What will be the average coverage of each molecule? The average overall coverage (assuming an average molecule size of 50kb and a genome size of 3 billion base pairs)?

4. **(0.5 points)** Nanopore now offers 3 different sequencers: https://nanoporetech.com/how-it-works in increasing order of capacity (the SmidgION, MinION, PromethION). Name one potential application (you can get creative) of each.