Assignment #6

The goal of this project is to analyze next generation sequencing data taken from two *C. elegans* strains. Unlike previous projects/assignments, we will not only design scripts, but we will also use these scripts to analyze and understand common issues that arise from next-generation sequencing.

On canvas is a template file that you should rename as **Assignment_6_yourlastname.py**.

I have also uploaded a reference fasta file, as well as 4 sets of bam/index files:

- **LSJ2_II.bam** and **N2_II.bam** contain sorted indexed bam files that map to CHROMOSOME_II.
- **LSJ2_II_reduced.bam** and **N2_II_reduced.bam** are smaller files that you can use for debugging purposes (i.e., you don't want to debug code on huge files that might take a long time to analyze).

The example images and output are run on the reduced/smaller bam files, but you will be graded on your output from the large files! You will turn in (1) your script as a .py file, (2) a pdf of your write-up to the analysis questions, including figures supporting your arguments.

1. Script requirements:

A. Follow the template code to take in the following positional arguments: <reference file> <bam file LSJ2> <bam file N2>

B. Open these files using the appropriate methods from *pysam* [to install, use install -c bioconda pysam]

C. For each bam file, identify:

1. number of reads
2. number of reads with mapping quality zero
3. number of reads that contain a mismatch to the reference

Store and print this information back to the user.

2. Data analysis: [include all answers in a **separate pdf file**]

A. Calculate the coverage of LSJ2 and N2 on CHROMOSOME_II in 50,000 bp bins. Use matplotlib to plot a scatter plot of these coverages comparing LSJ2 and N2. The x-axis should show coverage in LSJ2 and the y-axis should show cover from N2. Also plot a normalizing line that accounts for the different sequencing depth these strains were sequenced to. Save this as **figure1_yourlastname.png.**

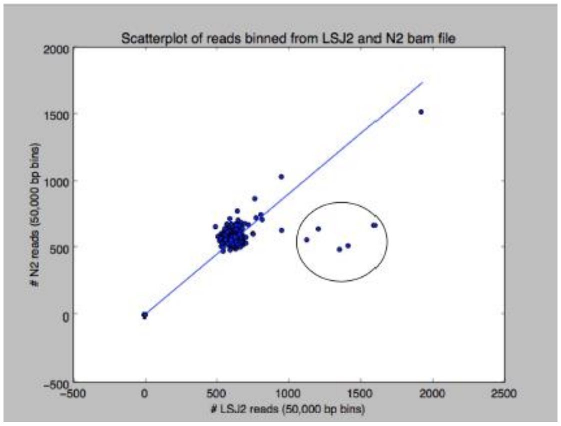
B. You will notice a set of unusual bins (circled in first graphical output example). What might cause these reads fall off the axis so? Confirm this hypothesis by identifying the location of these unusual bins and plot the average coverage of LSJ2 and N2 in this region at 5000 bp resolution. Save this as **figure2_yourlastname.png**

C. Identify all the positions in LSJ2 with 2 or more reads that mismatch the reference. For these positions calculate the percent of reads that are mutant vs. total in LSJ2. Also calculate the percent of reads that are mutant vs. total in N2 at these same positions. Create a scatter plot that show the LSJ2 mutant frequency on the x-axis and the N2 mutant frequency on the y-axis. Save this as **figure3_yourlastname.png**

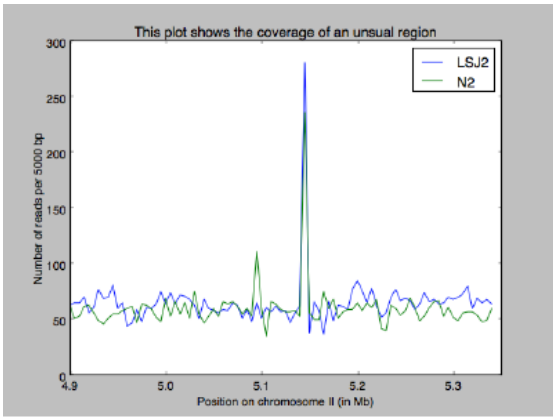
D. Using your outputted figure 3, identify the regions on the graph that represent actual differences in DNA sequences between LSJ2 and N2, errors in the *C. elegans* reference genome, errors due to misaligned reads, and errors due to mistakes in the sequencing run. What percentage of all of these are *bona fide* mutations?

Graphical Outputs:

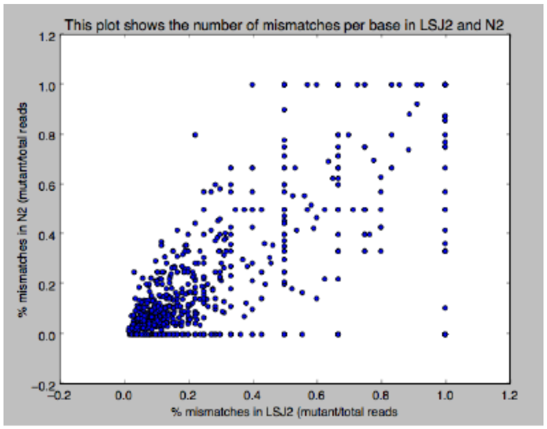
The following examples below show the expected output figures your script should also produce. **Your script should output these as separate .png files**. In total, there should be at least three figures.



<u>Ex. Fig1</u>: N2 and LSJ2 reads. *Circle is option here I'm simply showing you which bins are 'unusual'):*



<u>Ex. Fig2:</u> Coverage of an unusual region. *This example doesn't show the actual correct region (it's up to you to figure that out!) but this is the general format of your figure.*

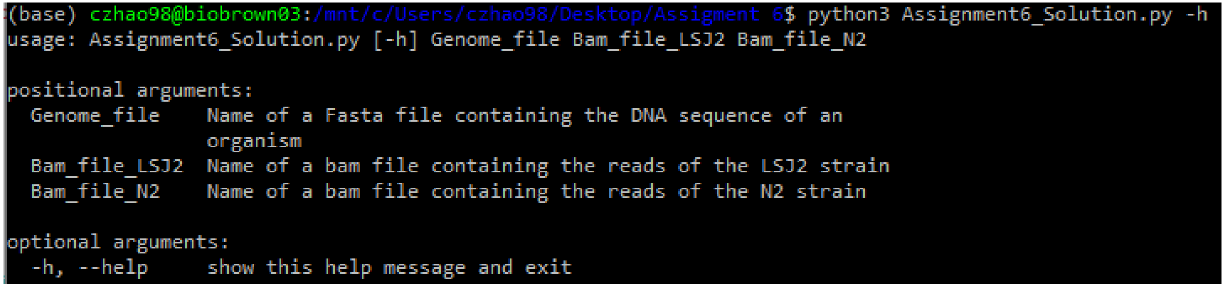


<u>Ex. Fig3:</u> Mismatch frequencies.

Note that all of these examples are run on the *reduced bam files*. I have done this to help you debug. However, your write-up (including figures) should be based on the *full bam files*.

Example Usage:

$ python3 Assignment6_Solution.py -h

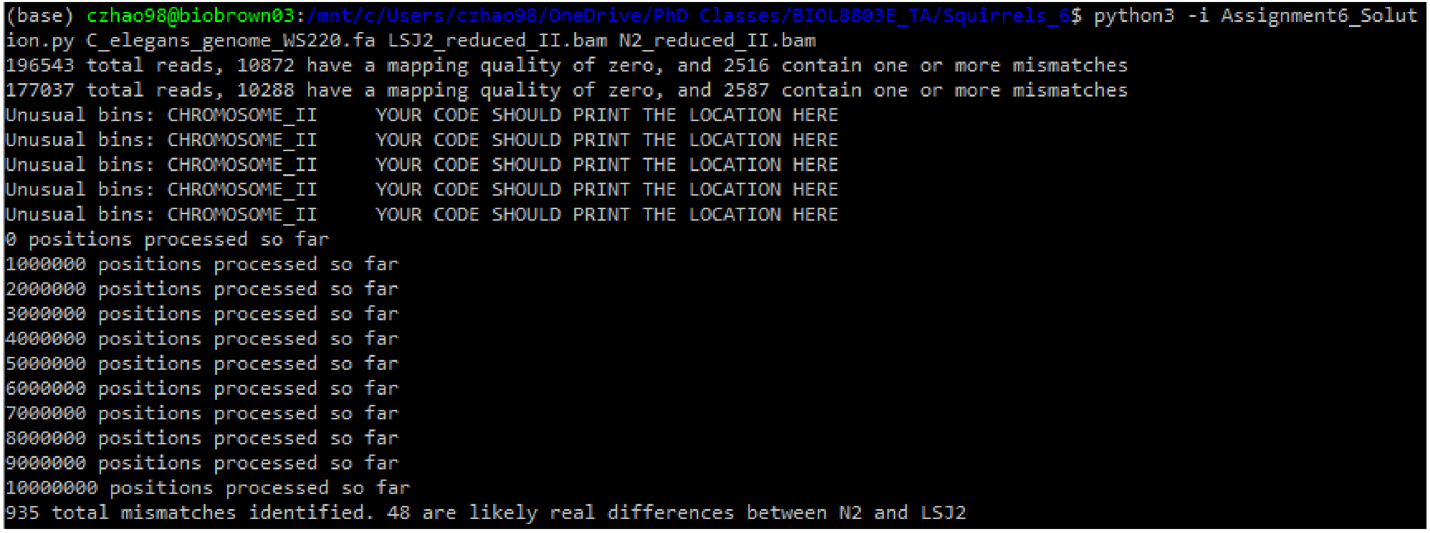


```
(base) czhao98@biobrown03:/mnt/c/Users/czhao98/Desktop/Assigment 6$ python3 Assignment6_Solution.py -h
usage: Assignment6_Solution.py [-h] Genome_file Bam_file_LSJ2 Bam_file_N2

positional arguments:
  Genome_file    Name of a Fasta file containing the DNA sequence of an
                 organism
  Bam_file_LSJ2  Name of a bam file containing the reads of the LSJ2 strain
  Bam_file_N2    Name of a bam file containing the reads of the N2 strain

optional arguments:
  -h, --help     show this help message and exit
```

$ python3 Assignment6_Solution.py C_elegans_genome_WS220.fa LSJ2_reduced_II.bam N2_reduced_II.bam



```
(base) czhao98@biobrown03:/mnt/c/Users/czhao98/OneDrive/PhD Classes/BIOL8803E_TA/Squirrels_6$ python3 -i Assignment6_Solut
ion.py C_elegans_genome_WS220.fa LSJ2_reduced_II.bam N2_reduced_II.bam
196543 total reads, 10872 have a mapping quality of zero, and 2516 contain one or more mismatches
177037 total reads, 10288 have a mapping quality of zero, and 2587 contain one or more mismatches
Unusual bins: CHROMOSOME_II     YOUR CODE SHOULD PRINT THE LOCATION HERE
Unusual bins: CHROMOSOME_II     YOUR CODE SHOULD PRINT THE LOCATION HERE
Unusual bins: CHROMOSOME_II     YOUR CODE SHOULD PRINT THE LOCATION HERE
Unusual bins: CHROMOSOME_II     YOUR CODE SHOULD PRINT THE LOCATION HERE
Unusual bins: CHROMOSOME_II     YOUR CODE SHOULD PRINT THE LOCATION HERE
0 positions processed so far
1000000 positions processed so far
2000000 positions processed so far
3000000 positions processed so far
4000000 positions processed so far
5000000 positions processed so far
6000000 positions processed so far
7000000 positions processed so far
8000000 positions processed so far
9000000 positions processed so far
10000000 positions processed so far
935 total mismatches identified. 48 are likely real differences between N2 and LSJ2
```