I-519 Report HW4 Part 2

For part 2/3 of HW4, genome *Staphylococcus aureus subsp* (accession number **NC_007795.1**) was split up into reads of lengths 70 and 250. Both sets of reads were indexed, mapped, and sorted utilizing BWA and Samtools. BWA assisted with indexing and mapping, whereas Samtools helped with reviewing the stats of the mapped reads. The genome used had a length of **1,000,000**. Below outlines results for reads of length 70 and of length 250.

Reads of Length 70

Number of Reads: **214,286**

Number of Mapped Reads: **214,286**

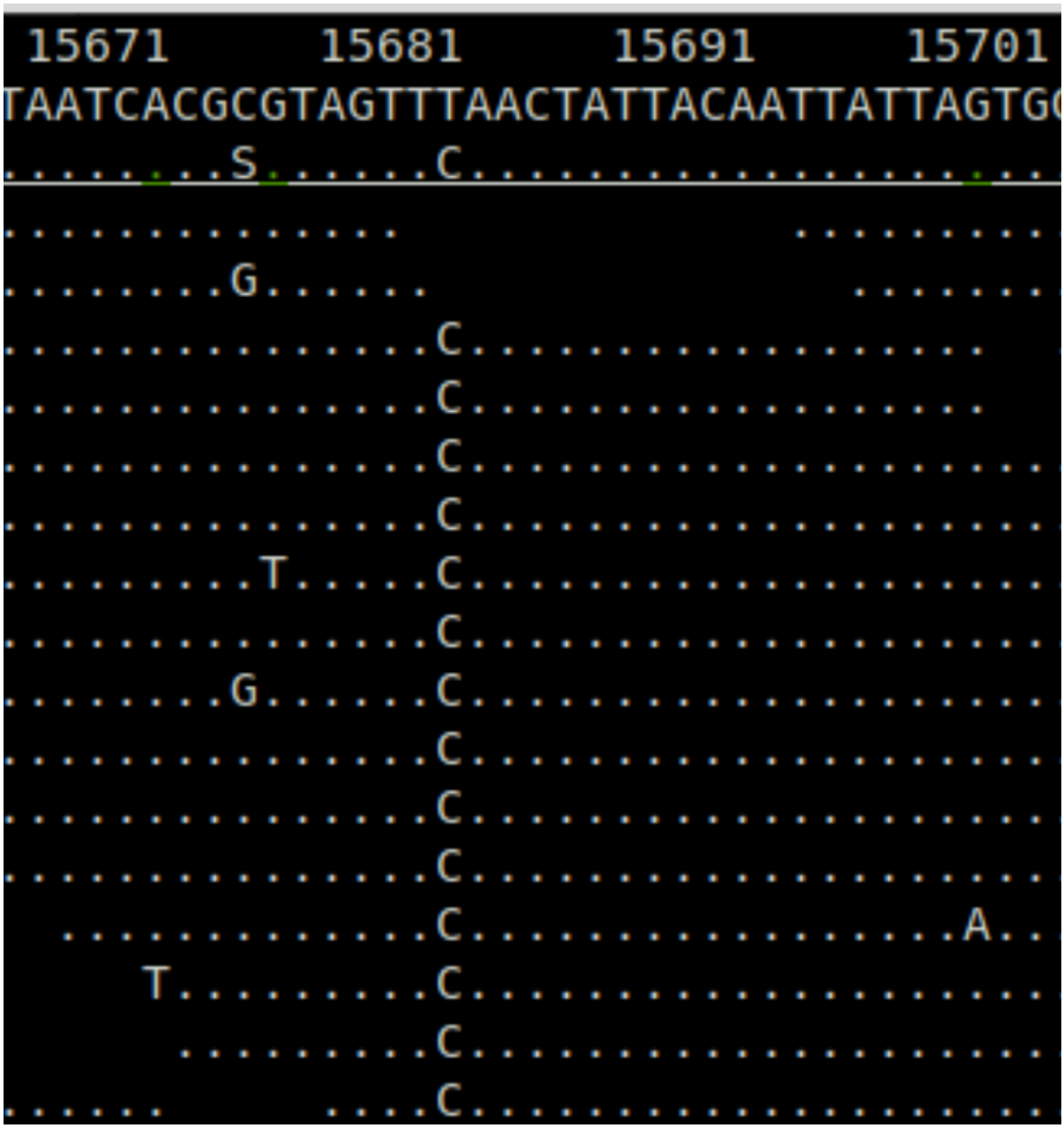
Number of Unmapped Reads: **0**

Tview Mutations: *(See code and picture below)*

Here, you can see mutations and technical errors. The mutation is represented at around position 15,681 where you see the vertical line of "C" nucleotides, where it was supposed to be a "T". The technical errors are the scattered nucleotides that do not follow a vertical line.

In [5]:
```
# Mutation Screenshot Below
from IPython.display import Image
Image(filename='C://Users//13177//Desktop//mutation_error70.PNG')
```

Out[5]:



Reads of Length 250

Number of Reads: **60,000**

Number of Mapped Reads: **60,000**

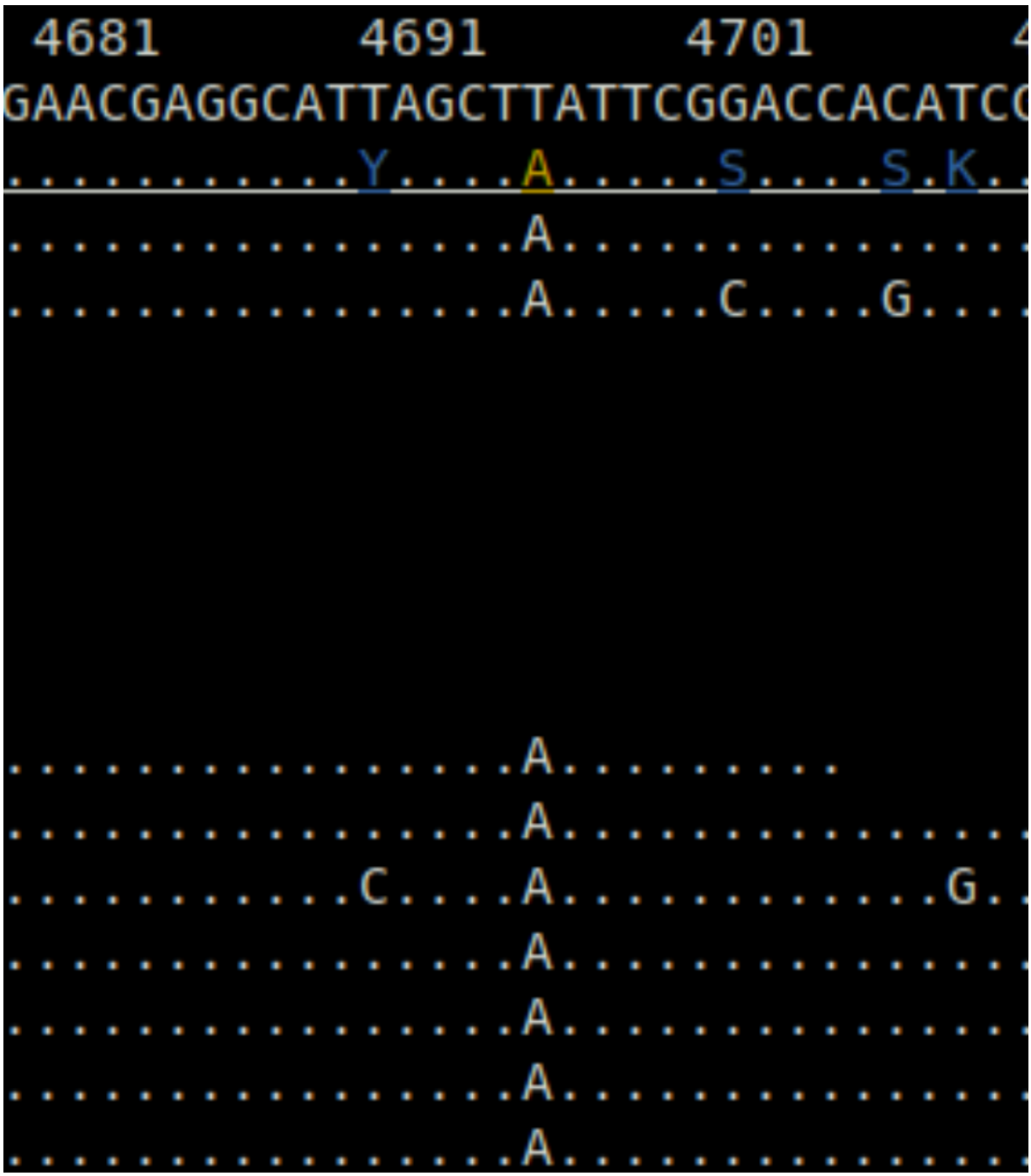
Number of Unmapped Reads: **0**

Tview Mutations: *(See code and picture below)*

Again, the picture below shows a mutation and technical errors. The mutation is shown with the vertical line of "A" nucleotides between positions 4,691 and 4,701. The technial errors are the scattered nucleotides that do not not match the sequence above.

In [6]:
```
# Mutation Screenshot Below
from IPython.display import Image
Image(filename='C://Users//13177//Desktop//mutation_error250.PNG')
```

Out[6]:



Coverage Plot

Below shows the code to generate a coverage plot for reads of lengths 70 and 250. To extract the alignment data for reads of length 70, I ran the command **samtools stats sorted_alignment70.bam | grep ^COV | cut -f 2- > stats70**. For reads of length 250, I ran **samtools stats sorted_alignment250.bam | grep ^COV | cut -f 2- > stats250**.

To create the plot, I used matplotlib.pyplot and pandas libraries to construct the histogram. The data was stored in two different text files and read in/stored as dataframes. Appropriate columns were extracted from dataframes to plot the coverage/number of positions.

The histogram shows the distrubtion of mapped reads at different positions. The two data sets show the overlap.

In [64]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

reads70 = pd.read_csv('stats70',header=None,sep="\t")
reads250 = pd.read_csv('stats250',header=None,sep="\t")
plt.bar(reads250[1], reads250[2],.9,color='cyan',label='250BP')
plt.bar(reads70[1], reads70[2],.2,color='black',label='70BP')
plt.xlabel('Position')
plt.ylabel("Position Frequency")
plt.title('Coverage Plot')
legend=['250-BP','70-BP']
plt.legend(legend)
plt.show()
```