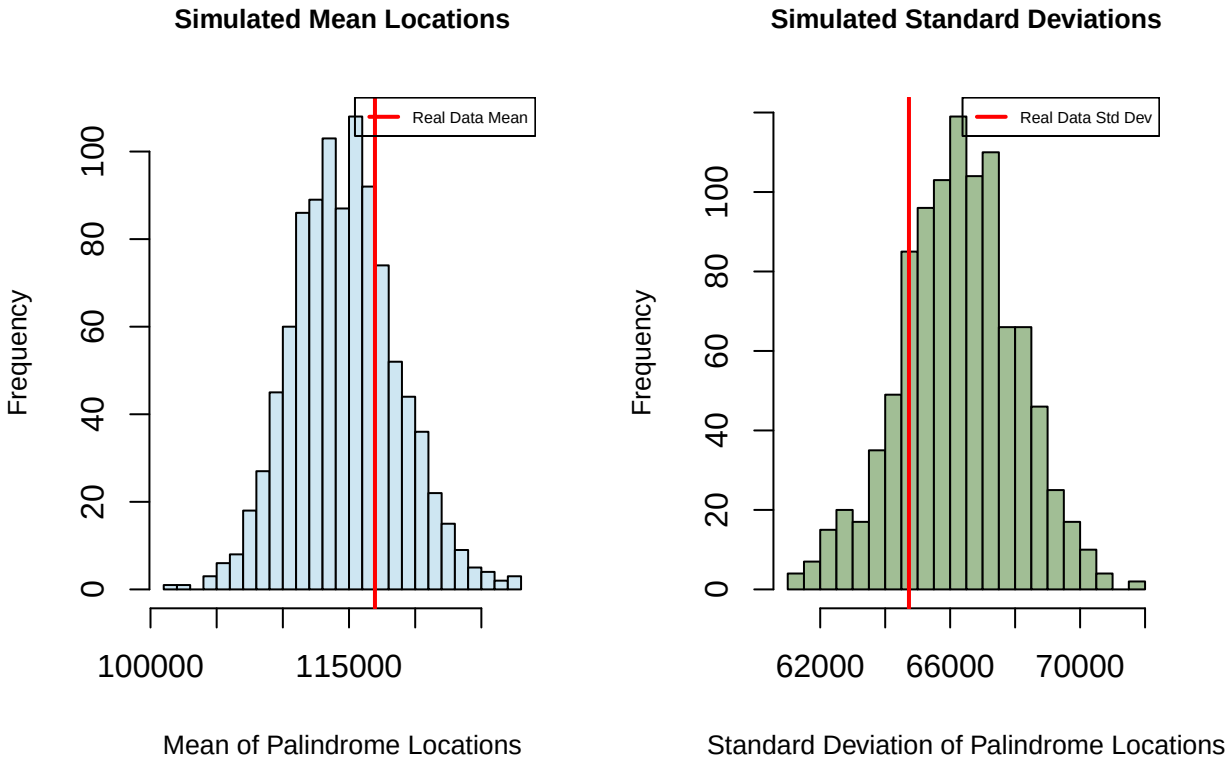
Question 1: Compare Data with Random Scattering of Palindromes

The dataset contains the locations of palindromic sequences in a series of 229,354 base pairs. We will do multiple simulations of random locations of palindromic sites along these 229,354 base pairs and compare it to the real data.

```
##            Statistic Real_Locations Simulated_Locations
## 1               Mean   1.169601e+05        1.147081e+05
## 2             Median   1.178260e+05        1.146895e+05
## 3           Variance   4.190236e+09        4.394407e+09
## 4 Standard Deviation   6.473203e+04        6.626665e+04
```

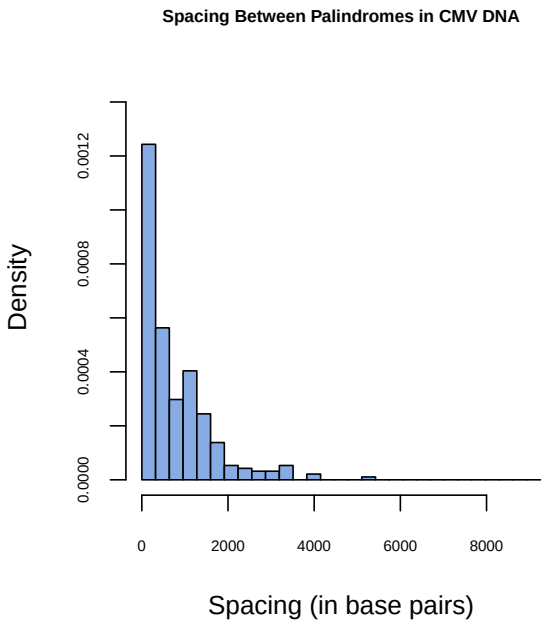
To qualitatively compare the real palindrome locations to the simulated locations, I plot a histogram of the means of each simulated palindrome sequence and analyze where the real mean is on the distribution. The same process is also done for the standard deviations.

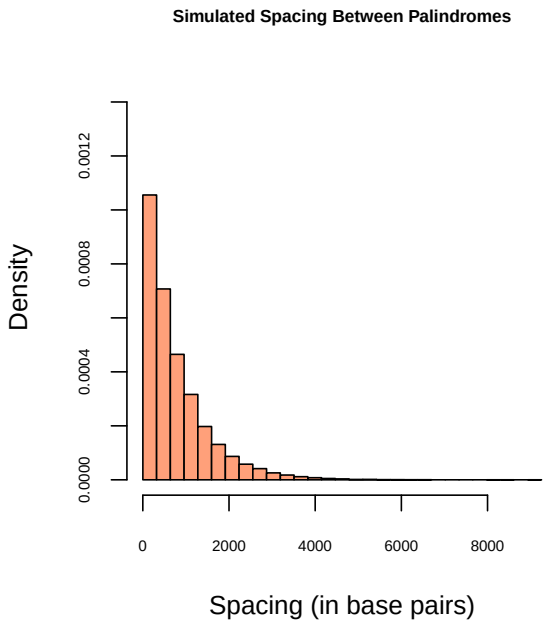


Question 2) [Locations and Spacings] Use graphical methods to analyze the patterns in:

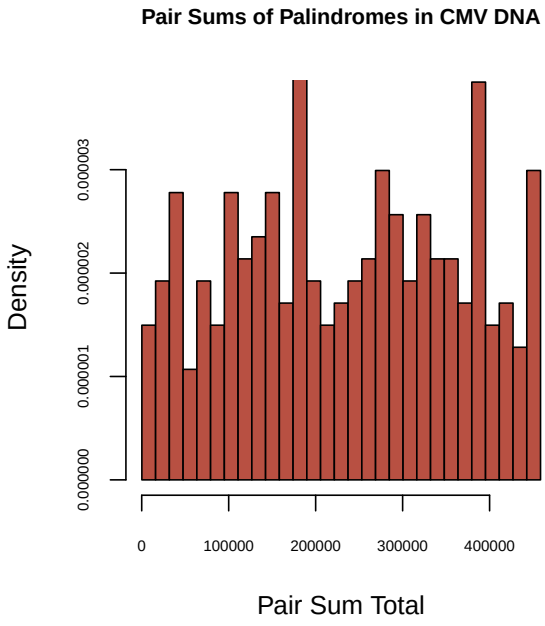
In order to further analyze the distribution of the palindromes, let's see three different variations of the data and how they deviate from a uniform scatter across the DNA sequence. To do this, we will compare the observed patterns below to expected uniform random distributions to identify any significant clsuters of unusual spacing in these palindrome locations. One important thing to notice is that we used only 10 simulations of the palindromes

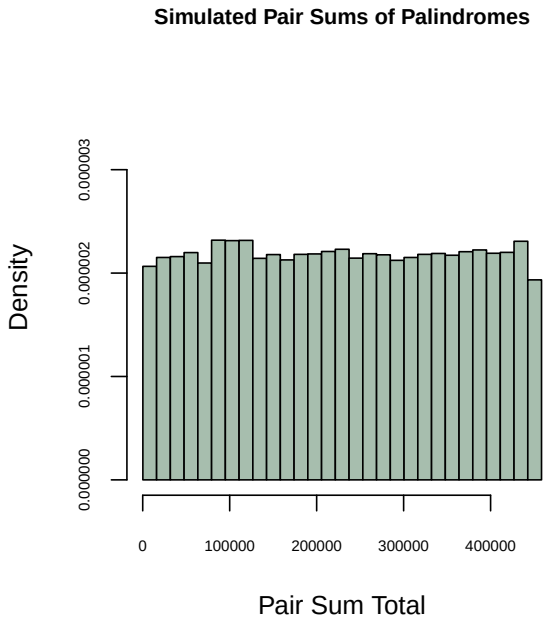
a) Spacing between consecutive palindromes



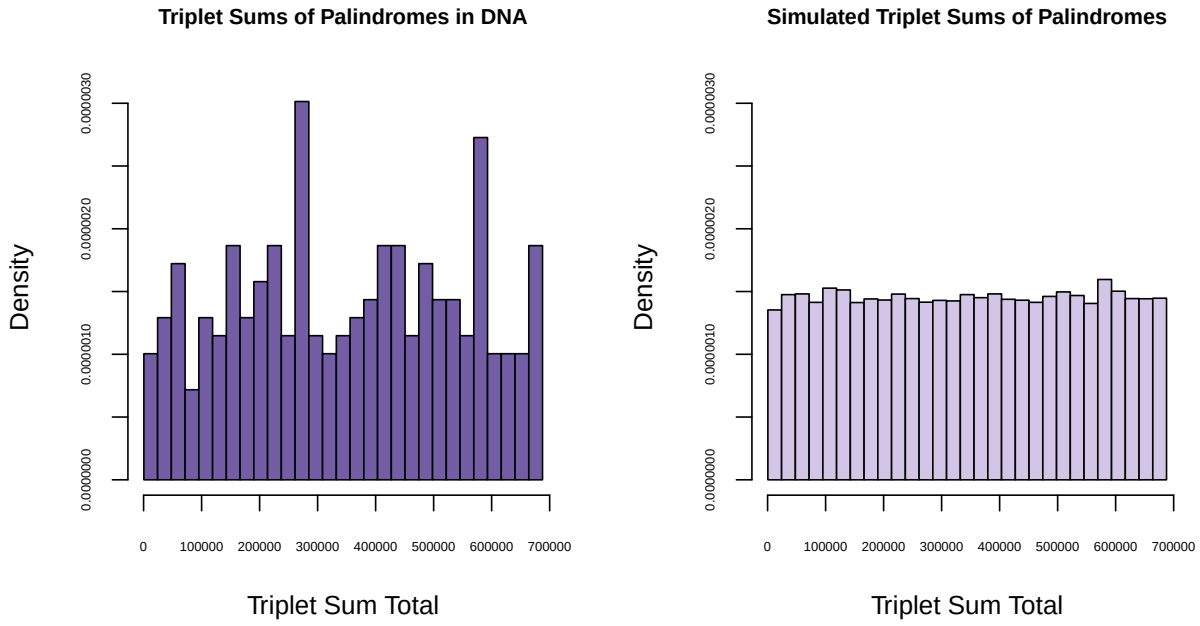


b: Sums of Palindrome Pairs





c) Sums of Palindrome Triplets



Looking through these three comparisons of palindromes and their simulated statistics, a few things can be noticed. We'll go through these one section at a time:

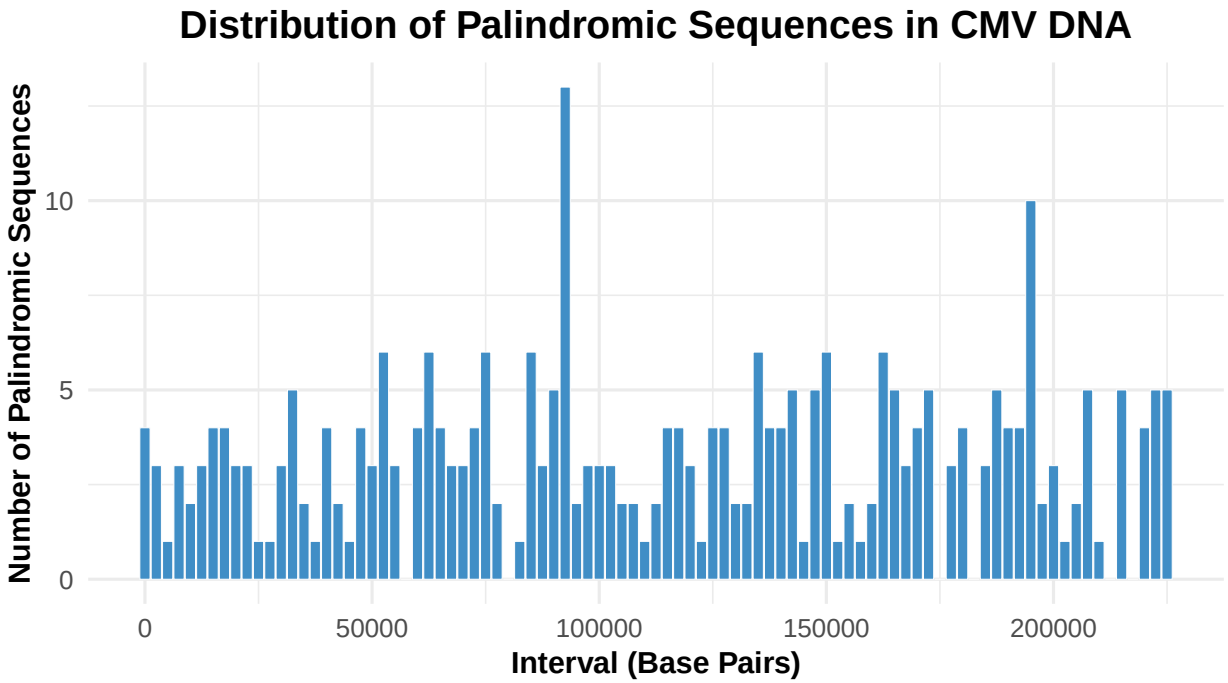
1) For spacing between consecutive palindromes, it is shown that the palindromes in the CMV DNA emulate a right-skewd distribution in terms of the spacing differences. There are peaks at the start of the spacing differences of the CMV data, which perhaps show that there's a slightly higher chance that palindromes could be closer together, but the difference is not very high (<0.004 difference). There is also a small dip in spacings between 800 - 1000 palindrome locations long. However, most of the data is clustered in between the 0 - 2000 spacing mark, meaning that most palindrome locations are somewhat near each other. There is also a potential outlier in between 5000 - 6000 palindrome locations long, which shows there's potential for longer spacing between palindrome locations. These are rarely seen in our simulated spacings between palindromes, with only a few markings past the 4000 spacing mark.
2) The pair sums for palindronmes in the CMV DNA do not follow a uniform distribution. There are peaks of pair sum totals, with one being near the ~200000 pair sum mark and the other being near the ~400000 pair sum mark. There are also dips at ~50000 and ~450000 On the other hand, the simulated pair sum of palindromes follows a uniform distribution, with most of the pair sums having around a ~0.00002 density in comparison to the CMV DNA, which has ranges from 0.00001 to 0.000045. For the CMV DNA Pair Sums, there are clusters from 100000 - 150000 as well as the 250000 - 325000 pair sum ranges.
3) Finally, for the palindrome triples, there is a similar pattern to the pair sums. They don't follow a uniform distribution, with many peaks as well as a dip. Two peaks include a peak near the ~300000 mark as well as near the ~600000 marks, while a dip is near ~100000. There are no significantly large clusters, though one could call in between 400000 - 500000 a small cluster. Compared to the simulated triplet sums, which is flat and follows a uniform distribution, the CMV DNA looks like it follows a different pattern, as there are peaks at seemingly random places.

While spacing between consecutive palindromes follows a similar distribution to our simulated consecutive palindromes, the pair sums and triple sums do not, with unusual spacing between their distributions com-

pared to their respective simulated uniform distributions. These suggest that the CMV DNA does not follow a uniform distribution, even though it might seem like it.

Question 4) Potential for Origin of Replication

In order to identify a palindrome sequence that has potential for origin of replication, we decided to create clusters in our dataset to see whether or not a significant number of palindrome sequences showed up in any interval. To identify clusters of palindromic sequences, we created intervals along the DNA sequence and counted the number of palindromes in each interval. For our distribution, we used intervals of 2500 because we wanted to be extremely specific, but also have a large enough interval size that there would be significant differences that could be seen immediately from our data.



The two main intervals that have potential for an origin of replication are from 92500 - 95000 or 195000 - 197500. In particular, the interval between 92500 and 95000 has the highest chance for an origin of replication, as it contains ~4x more sequences than an average sequence, at 13 sequences compared to the average 3.23.