1) Introduction

- The human cytomegalovirus (CMV) is a life-threatening disease for people with an especially poor immune system, and it has been essential for scientists to figure out a way to combat it. One of the ways the scientists had tackled this problem was to study the way in which this virus replicates, or find the origin of replication. To aid this process, diverse statistical methods and analyses are made on the data regarding the DNA (palindrome), which is the ultimate subject in which multiple patterns are formed. The dataset used contains 296 palindromes that were at least 10 letters long, and at two distinct locations (93,000th and 195,000th base pairs), the number of palindromes were significantly higher than other points. These clusters of palindromes are important, because they may indicate the occurrence of a potential replication site. Through different statistical methods, will be investigating if these clusters are just a chance occurrence, or if they are indeed signs of a replication site.

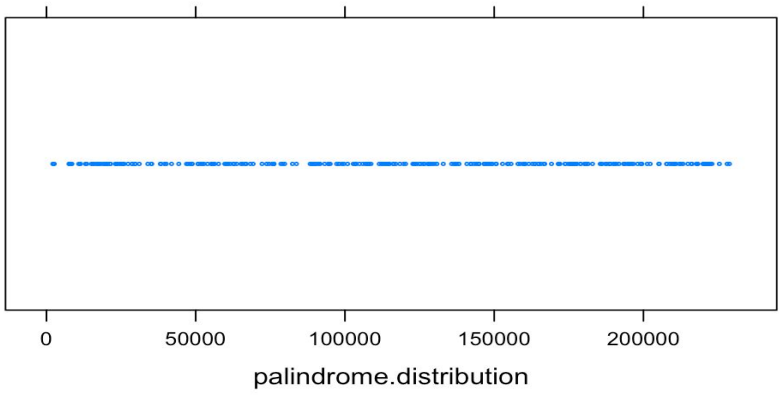
2) Analysis

1. **[Random Scatter]**

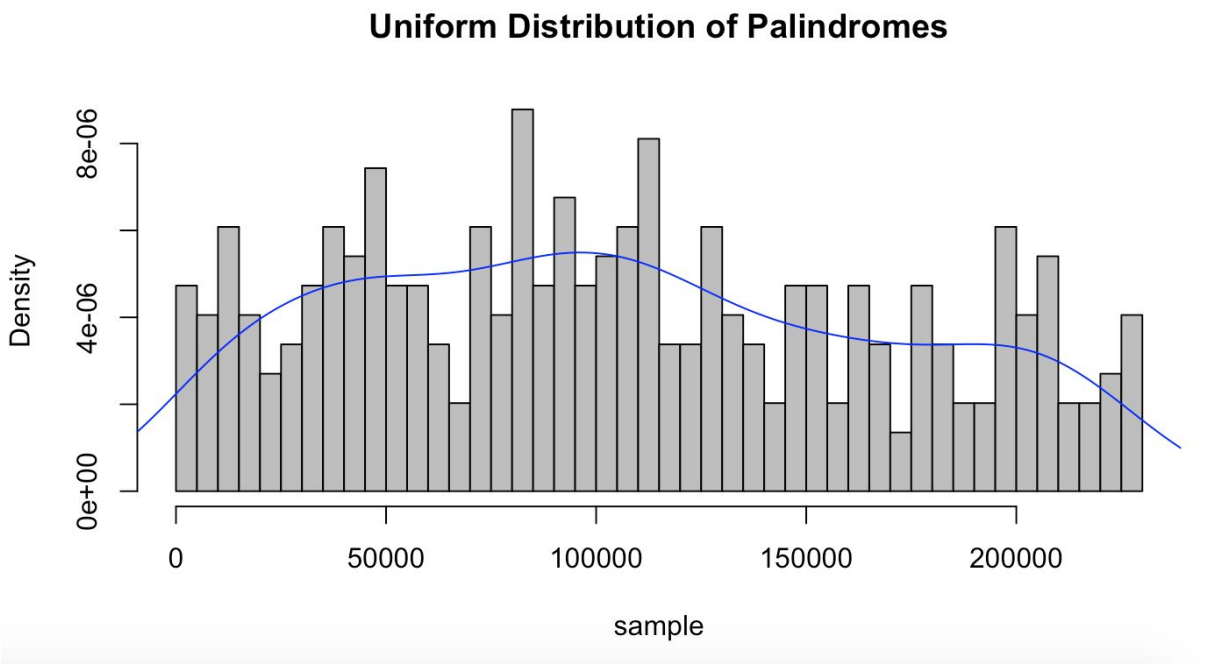
   Methods:

- The goal of identifying the random scatters in the palindrome is to see how a random scatter looks by using a computer to simulate it. To simulate a random scatter, we visualized the distribution of the uniform sample, the distributions of random uniform scatter instances, and the theoretical uniform distribution. The distribution of the uniform sample was visualized by using a strip plot. The distribution of random uniform scatter instances was visualized by using a histogram, and this histogram was simulated at least 10 times. The theoretical uniform distribution was visualized by creating a theoretical density distribution, indicated by the blue line on top of the visualized histogram.

Analysis:



- In order to visualize the distribution of the uniform random sample, we decided to create a strip plot that indicates the 296 sites chosen from the 229,354 possible bases. A strip plot was incorporated because it was useful to display the different sites chosen in a straight horizontal line, which is easy to interpret. The strip plot shows that because the sampling was run through a uniform distribution, the different sites were chosen from any base (1 ~ 229,354). The original dataset's histogram was also created in *Appendix A,* which displays two points of frequency spikes that occur between the 93,000th and 195,000th pairs of DNA bases.



- In order to visualize distributions of random uniform scatter instances, we incorporated a histogram that shows the uniform distribution of the different sites chosen from the large number of bases offered. As shown in the histogram, there are several points in the sample where the density is slightly higher than the other data points in the histogram. Also, this instance is shown throughout the histogram, suggesting the sample is indeed uniformly sampled. From the histogram, we can see that in the 90,000th pair, there seems to be a high spike, which is similar to the palindrome sequences in the CMV dataset. To visualize the theoretical uniform distribution, we incorporated a theoretical density line inside the histogram, shown in blue. This curve is important, because the 296 samples drawn are from the theoretical density curve. The density curve represents all the possible data values that the 296 bases can take.

Conclusion:

- By creating and visualizing the distribution of random uniform scatter using a histogram, we can compare the sampled histogram to the histogram of the given histogram plot. We can observe that both the sampled and original histograms have multiple spikes where the frequency (density) is higher than other points. For both histograms, we can see this

occurring in the 90,000th DNA bases, but these spikes in frequency occur more often in the sampled random distribution. However, we can see that the two histograms from above and *Appendix A* are visually similar, as the locations of high frequencies are very similar. Although we cannot conclude definitely that the original data is random, we could assume that the palindromes in the original data are random.

2. [Locations and spacings]

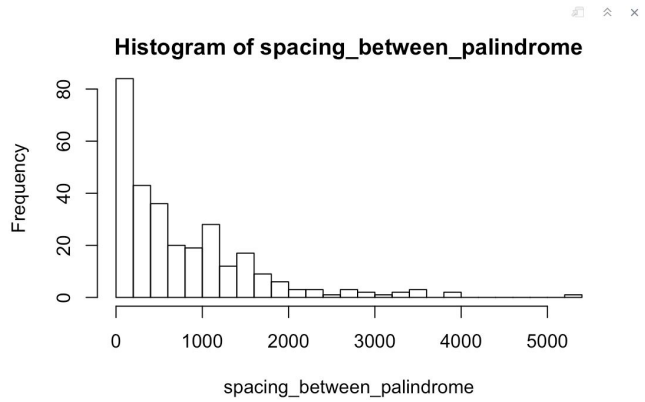
Methods:

- First, we examined the three types of spacings. The first type of spacing is spacing between two consecutive palindromes, which is simply the difference between two consecutive palindromes. The second type of spacing is the spacing between sums of two consecutive palindromes, and the third type is the spacing between sums of three consecutive palindromes. In order to get this data, we first saved the palindromes data as 'locations' and referenced the necessary indices of 'locations' by using head(), tail(), etc. Then we simply added the data together (i.e. in the case of sums of consecutive pairs, we would do 177 + 1321 = 1498) since 177 and 1321 are consecutive palindromes. Then we visualized the distribution of the three types of spacings (also known as the difference in palindromes' locations) by plotting the histograms. The spacing(breaks) between palindromes was set to 30 for all histograms. Then, we also plotted randomly (uniform) generated histograms of the three types of spacings by using runif(). Unlike the original histograms, the randomly generated histograms have a wide spectrum of x-axis which stretches from negative to positive values. What we're looking at begins from 0 and to the right of 0.

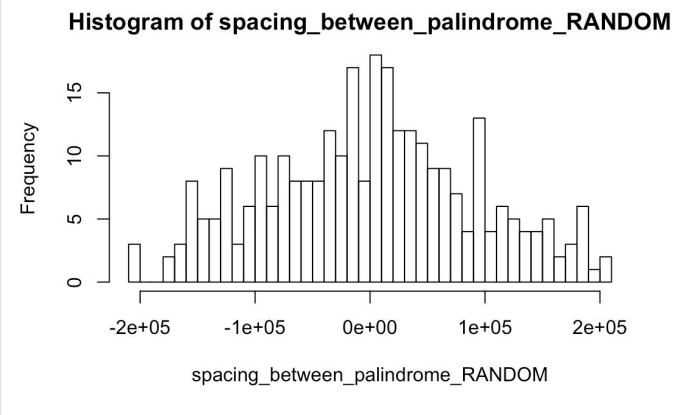
Analysis:

- The three histograms had a commonality in that they were right-skewed (much of the data falls to the right), but the level in which they were right-skewed were different: the first type of spacing was the most right-skewed in general. The maximum values of

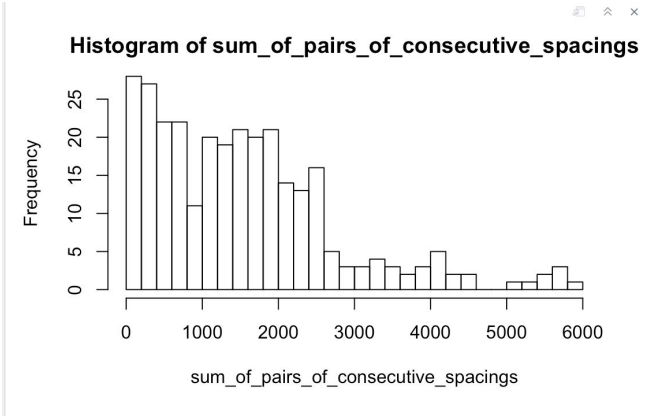
spacing were different as well, with the first spacing around 5000, second with 6000, and third approximately 8000. To appropriately compare between the original and randomly generated histograms, we specifically looked at the bars to the right of 0 (only positive values of x-axis) We found that the randomly generated histograms were generally similar to the original histograms, having right skewness for all three histograms. The slight difference was that randomly generated histograms had more spikes (bars with high frequency) than the original histograms. We analyzed that the theoretical distribution would come from the Poisson distribution, which counts the number of points in different regions. It is the best standardized reference for our model because it is a natural model for uniform random scatter.



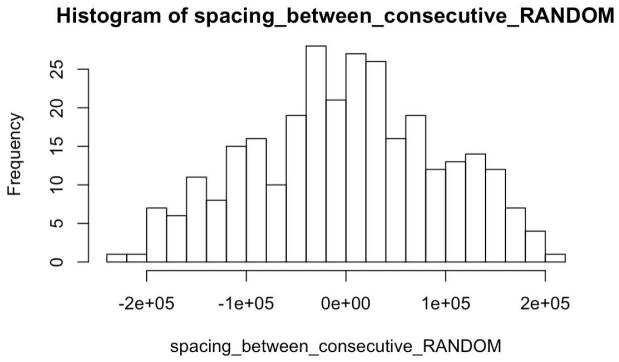
- Original histogram of the first spacing type



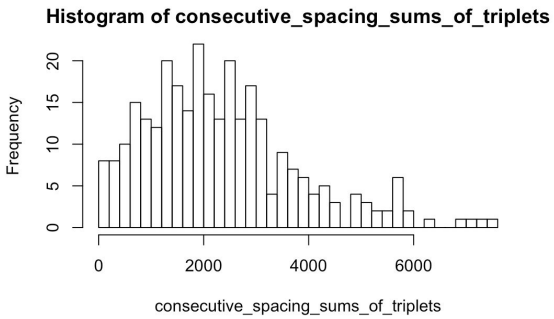
- Randomly (uniform) generated histogram of the first type of spacing



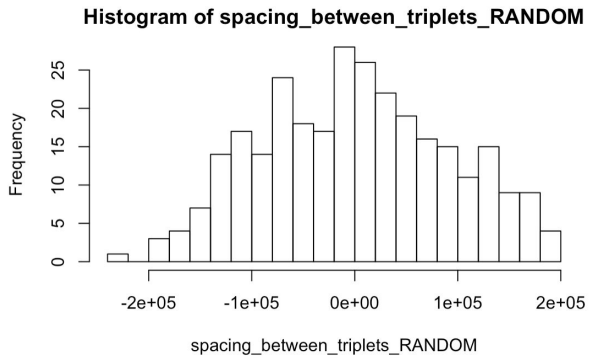
- Original histogram of the second spacing type



- Randomly (uniform) generated histogram of the second type of spacing



- Original histogram of the third spacing type



- Randomly (uniform) generated histogram of the third type of spacing

Conclusion:

- In conclusion, we were able to see that the similarities between the randomly generated histograms and original palindrome data histogram are similar, although the randomly generated histograms had more spikes. This supports the hypothesis that the palindrome data's clusters have occurred due to randomness, rather than a replication site.

3. [Counts]

Method:

- Even though we used graphical methods, it is hard to conclude whether our data is uniform random scatter or not. Thus, we split the non overlapping regions, which are 20, 50, 100, and use chisq test to check our data follows poisson distribution, which is uniform random distribution in this case. Also, we compared counts of each region based on the results.
- H0: This is follow the poisson distribution
- H1: This is not follow the poisson distribution
- Confidence level : 0.05
- Trunc : 9

Analysis:

- When we have too small regions such as 20, all the counts ,which is observed, is >=9. This has high p-value in our chi sq test. However, this is not a good observation since we can not observe the various values. On the other hand, if we choose 50 as our regions. Our p-value in chi sq test is 0.3128, so we fail to reject a null hypothesis, so we can conclude that our distribution could follow the poisson distribution. Our observation variance of the counts is also fairly distributed. We have more observations on each category. Lastly, if we increase more regions to the 200, the p-value is 0.0005 which is under 0.05. Thus, we reject our null hypothesis. Therefore, we can not use these observations. However, our reasonable selection of regions can be changed by variance of counts. If we have less counts of variance as reducing the trunc to the 7, 100 regions have higher p-value. Also distribution of observation is more fair than 50 regions. Thus, the regions can be considered according to the counts.

levels	Observed	Expected
0	0	7.472599e-06
1	0	1.105945e-04
2	0	8.183990e-04
3	0	4.037435e-03
4	0	1.493851e-02
5	0	4.421799e-02
6	0	1.090710e-01
7	0	2.306073e-01
8	0	4.266236e-01
>=9	20	1.916957e+01

levels	Observed	Expected
0	1	0.1342600
1	2	0.7948193
2	1	2.3526650
3	4	4.6425922
4	8	6.8710365
5	8	8.1353072
6	5	8.0268365
7	9	6.7884103
8	4	5.0234236
>=9	8	7.2306494

levels	Observed	Expected
0	118	1.118452e+02
1	117	1.103539e+02
2	32	5.444127e+01
3	27	1.790513e+01
4	2	4.416599e+00
5	2	8.715422e-01
6	1	1.433203e-01
7	0	2.020133e-02
8	0	2.491498e-03
>=9	1	3.027243e-04

(left = 20, mid= 50, right =200)

Conclusion:

- When we have appropriate regions to split the data, we will get the evidence about the uniform random scatter. Also, if we increase the regions inappropriately, the p-value of chi square will decrease. As a result, the observation will not follow the poisson

distribution. Shorter region or too longer region will affect the variance of the counts. On the other hand, we can consider counts first. When we set up counts first, we can re-address our size of regions

4. **[The Biggest Cluster]**

Method:

- We tested whether the large sample is better or not under the t-test. Before testing it, we find the difference of variance between samples.
- H0: The mean of two sample is same
- H1: The mean of two sample is different
- Confidence level: 0.05
- If the mean is the same, the greatest number of palindromes samples are not a potential origin of replication. Otherwise, it is potential origin replication

Analysis:

- By using the var.test function, we compare the variance between table 1 (20), table 2(50), and table 3 (70). Table 1 and table 2 are the same under since p-value are under our confidence level. Table 2 and Table 3 are different. Through the result, we set up the t-test. As a result, both p-values(0.19,0.5) are greater than our confidence level (0.05). Thus, we can reject the null hypothesis, and we can conclude that the mean of the samples is different. Chisq also table3(0.8)>table(0.3). Thus, table3 is a good fitting table. we can say that we can get the potential of origin data, when the values are large enough. We exclude comparing Chisq of the table1, table1 skewness is right skew. However, table2 and table3 are normally distributed. Thus, we can assume that table 2 and table 3 is
  better table.

Conclusion:

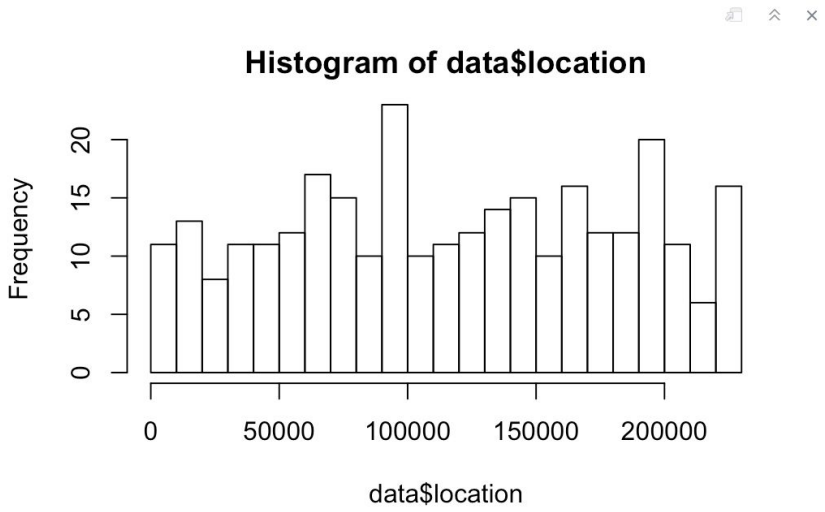
- When we get the greater number of palindromes, we can be closer to the origin of replication. If we get a worse fitting model when we increase our interval, we can't say the result. However, we tested two t-tests, and found the difference in means. Moreover, the greater intervals got the better results.

3) Report Conclusion

- By investigating the different scenarios, we found useful information and statistics to show that the increased frequencies in the two distinct spots of (93,000th and 195,000th base pairs) are due to chance, and are not actually replication sites. By investigating the random scatter by creating a uniform random sample, and graphically comparing the distributions, we found that there is a possibility that the clusters found in the original dataset is due to random chance. Also, through sample spacing and finding the six different histograms and simulating at least 5 random uniform scatters, we found that the histograms of randomly sampled and original are similar, showing that again, the clusters may be formed due to chance, and are not replication sites. Through examination of counts, when we have appropriate regions to split the data, we again have evidence of a uniform random scatter. If we increase the regions inappropriately, the p-value of chi square will decrease, showing that the poisson distribution will not be achieved. Through the biggest cluster, we can conclude that when we get a large number of palindromes, we can be closer to the origin of replication. Through different t-tests and difference in means and the different statistical analysis above, we can conclude that the clusters are indeed found from a chance occurrence, and are evidently not a potential replication site.

4) Appendix

- Appendix A:



- Histogram of the original HCMV data file.

5) Author Contribution Statement

- Seokmin Hong developed the code and analytical reasoning (method, analysis, conclusion) for the random scatter analysis. Yeongjae Kim and Youngseo Do encouraged Seokmin Hong to study the concept of theoretical density curve by referring to the slides and discussion labs provided in class and supervised the findings of his work.
- Youngseo Do developed the code and analytical reasoning (method, analysis, conclusion) for the locations and spacings analysis. Seokmin Hong and Yeongjae Kim helped create the different histograms for the different locations and spacings in the palindromes.

- Yeongjae Kim developed the code and analytical reasoning (method, analysis, conclusion) for counts and the biggest cluster analysis. Seokmin Hong and Youngseo Do helped visualize the tables, images, and assisted incorporating the previous sections' ideas into the analytical analysis of the biggest cluster.
- Youngseo Do edited grammar and code (if it can be made more efficient) throughout the code part of the homework as well as the written report.
- Seokmin Hong and Yeongjae Kim added to the appendix where statistical information was useful, and also wrote the introduction and final conclusions of the paper.