Validity of Palindrome Clusters as Replication Sites in CMV DNA

Author Contributions
James Lu: Intro, R-code, Part 1, Part 2, Part 4, Conclusion
Matin Ghaffari: R-code, plots, formal statistical tests, Part 3, Advanced Analysis, Discussion, Conclusion

Introduction

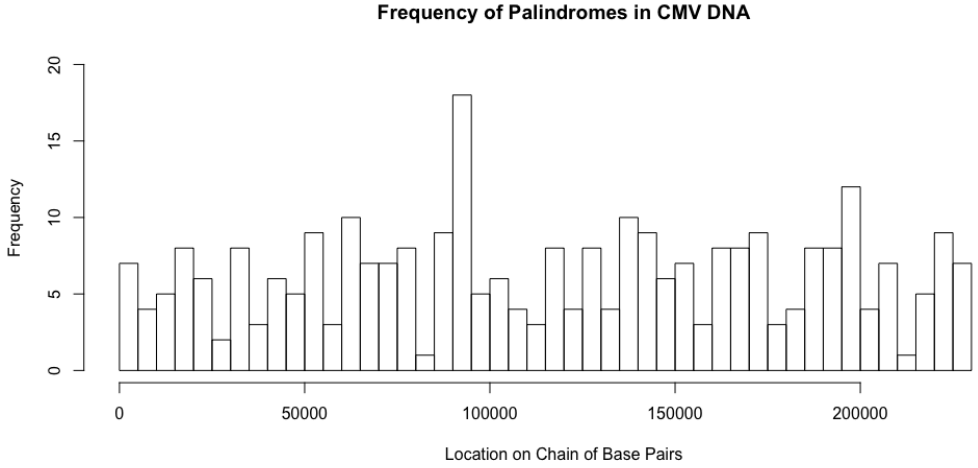
Cytomegalovirus, or CMV, is a common virus for people of all ages, however oftentimes a healthy person's immune system will keep the virus from causing illness. The incidence rate of CMV ranges geographically from as low as 30% to as high as 80%. CMV is also a part of the herpes virus family alongside herpes simplex type 1 / 2, Epstein-Barr, human herpesvirus 6 / 7. Scientists in the herpes family have already identified certain origins of replication marked by palindromes in the DNA. Traditionally scientists study the way that the virus replicates in order to find strategies to combat the virus, however oftentimes this requires extensive lab research, funding, and time to make progress. Instead we can statistically analyze if there are unusual clusters of palindromes and possibly identify if they are indeed an origin of replication or are simply an artifact of randomness.

To do this we are utilizing a dataset containing the base pair locations of 296 palindromes with length of at least 10 generated by Leung et al. (1991) from another larger dataset that contained the entire DNA sequence of CMV Chee et al. (1990). This data set has a sample size of 296 palindromes from a population of 229,354 Base pairs. Our main objective in this study is to answer the question of if we observe clusters of palindrome locations, do they represent a reliable structure that could serve as the origin of replication? To answer this question we look to compare our observed dataset to that of a uniformly random spread of palindromes to see if they might come from the same distribution. We will also compare the location, spread, and count of the observed data to try and determine how valid these clusters might be as origins of replication.

1) The Dataset Compared to Random Scattering

To begin our analysis we want to investigate the possible distribution that the observed palindromes may come from. To do this we compared the frequency of locations from the observed data to several simulations of locations sampled from a uniformly random distribution with the same sample (n = 296) and population size (N = 229,354).

Figure 1: Histogram depicting the observed frequency of locations with equal spacing of 4000



The observed frequency of palindromes and their locations are visualized in Figure 1, and from this visualization alone we can see peaks at locations ~90,000 and ~190,000 with the former having a larger and more exaggerated peak. To compare these locations against a purely random spread of palindromes we ran three simulations drawing samples of size 296 from a population of 229,354 from a uniform distribution.

Figure 2: Histogram depicting the first simulated frequency of locations from a uniformly random distribution with equal spacing of 4000

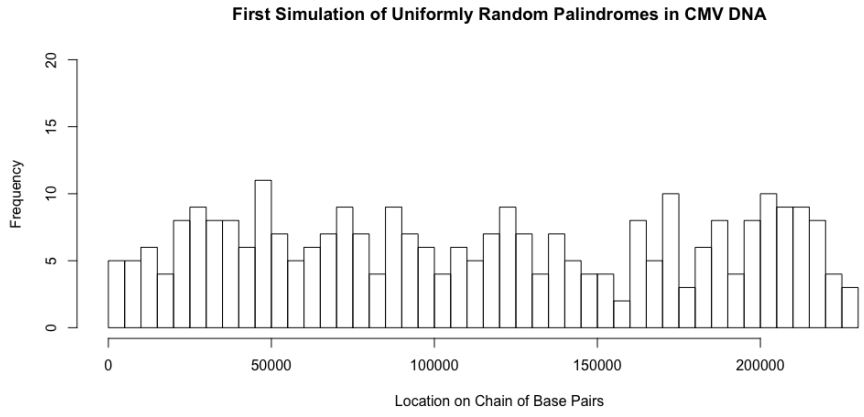


Figure 3: Histogram depicting the second simulated frequency of locations from a uniformly random distribution with equal spacing of 4000

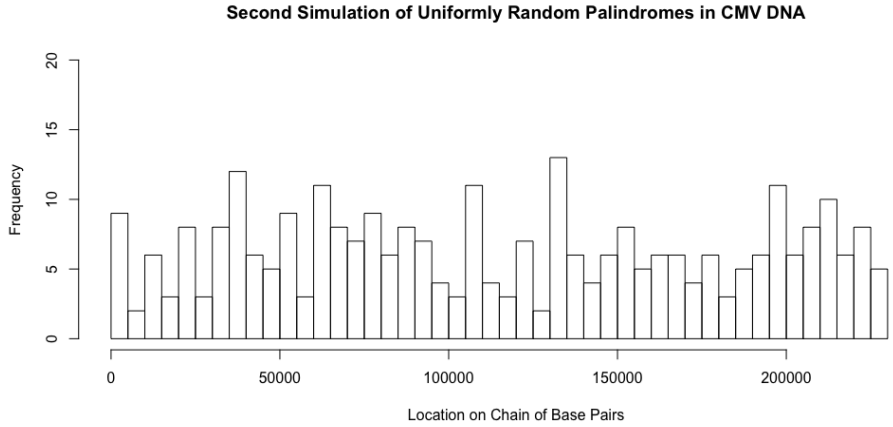
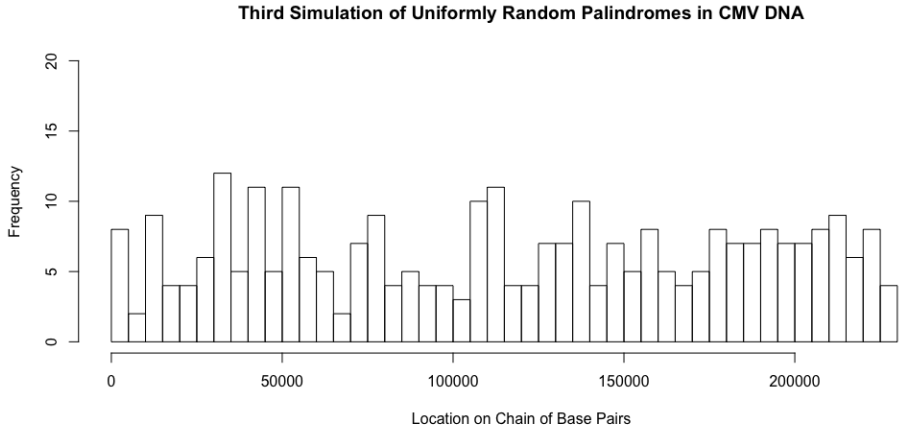


Figure 4: Histogram depicting the third simulated frequency of locations from a uniformly random distribution with equal spacing of 4000



All three simulations are depicted by Figures 2, 3, and 4 respectively. Although there are some peaks in the histograms, there are no obvious patterns that would indicate a replicability at any specific location.

Comparing these simulations to our observed location frequencies we cannot definitively say that our observed locations do not come from the uniform distribution as the two peaks that were observed may simply be outliers from the sample rather than evidence of a replication site. Further investigation is required to determine the validity of these observed peaks as replication sites.

2) Analysis of Locations and Spacings of the Palindromes

To continue our investigation we will graphically analyze the spacing between consecutive palindromes and the sum of consecutive pairs and triplets as well as the differences in locations of said palindromes. To analyze spacing we will generate various overlaid

histograms to show the density of spacing for each specific scenario using the gamma distribution as our expected distribution with our rate being 1 / mean(spacing).

Figure 5: Frequency of consecutive palindrome spacing overlaid with the density of A randomly sampled gamma distribution with shape 1 (exponential)

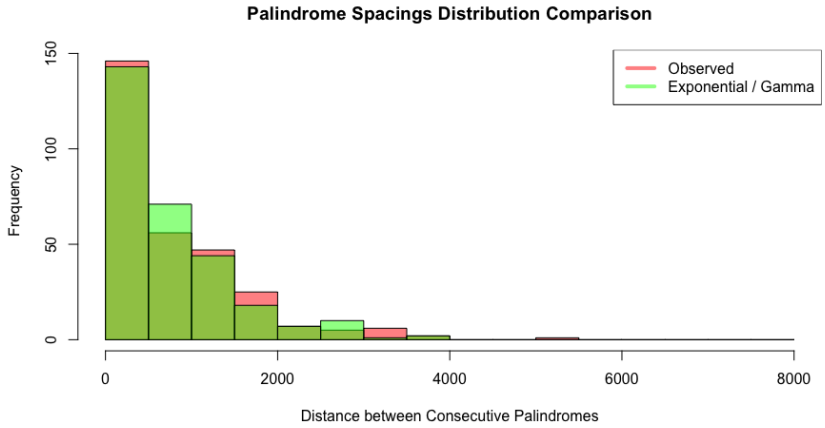


Figure 6: Frequency of consecutive pairs of palindrome spacing overlaid with the density of A randomly sampled gamma distribution with shape 2

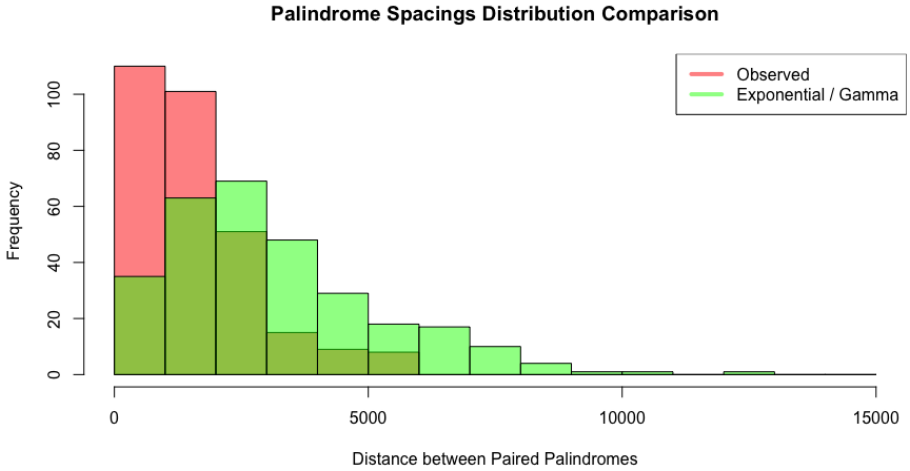
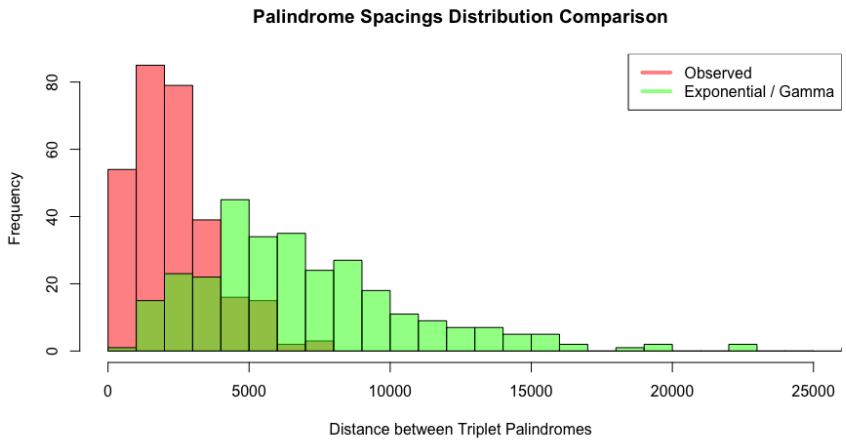
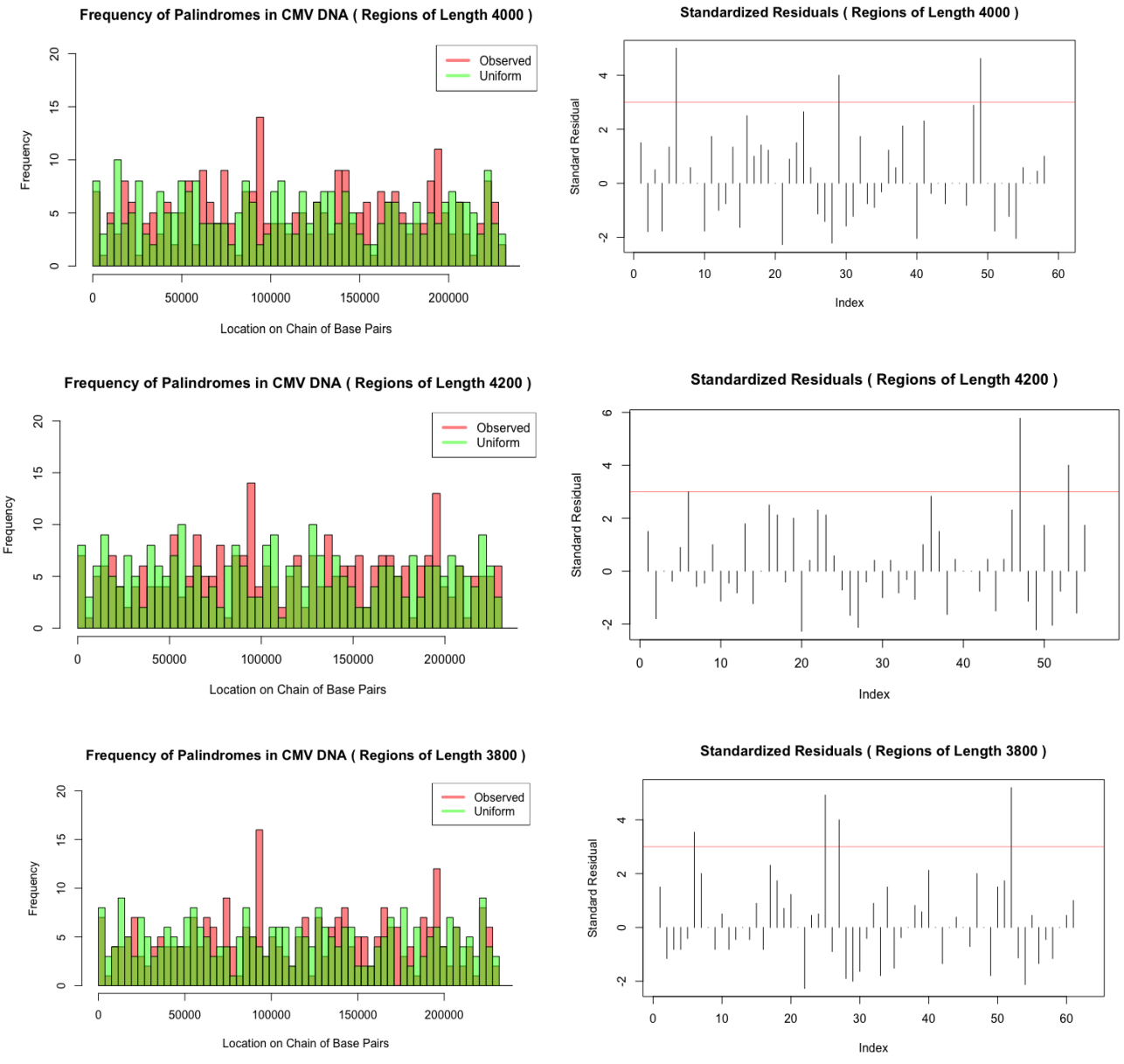


Figure 7: Frequency of consecutive triplets of palindrome spacing overlaid with the density of A randomly sampled gamma distribution with shape 3



After visualizing the different spacings in comparison to their respective expected distributions, we can see that the consecutive palindrome spacing *appears* to follow the exponential distribution while the sum of pairs and triplets stray away from theirs (gamma shape 2 / 3). It appears that for pairs and triplets, both distributions do not follow their expected distributions which may imply unusual clustering with spacings greater than 2.

Figure 8: Overlaid histograms and standardized residuals for locations of palindromes for region lengths 4200, 4000, and 3800.
**Red line indicates standard threshold of 3 for standardized residuals*

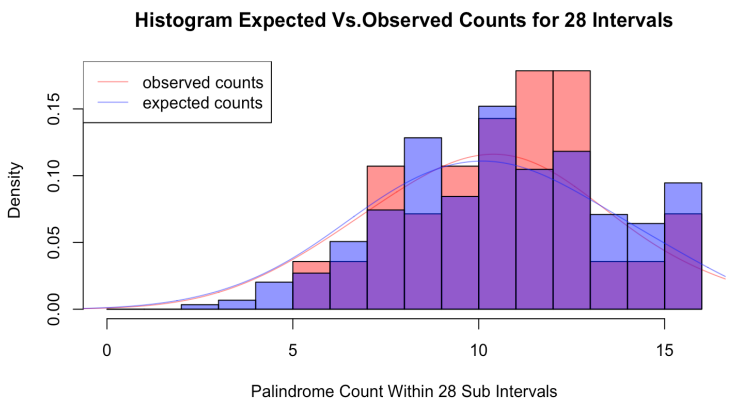


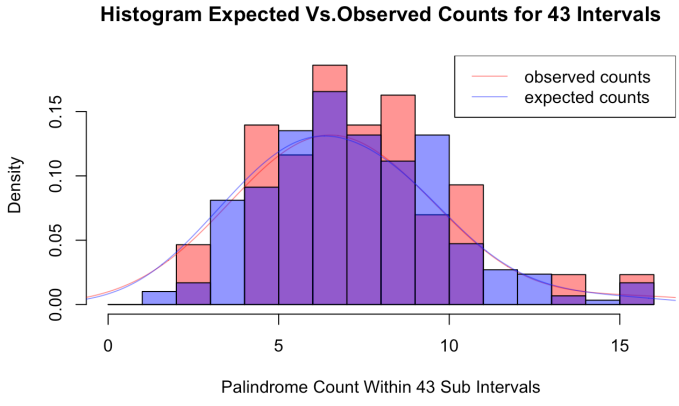
We can see from these visualizations that despite modifying the region length ± 200, there are still clusters that stray from the expected distribution as depicted by the standardized residuals with clusters over the threshold 3. This implies that there are certain locations of bases where clustering occurs however more formal investigation is required to confirm.

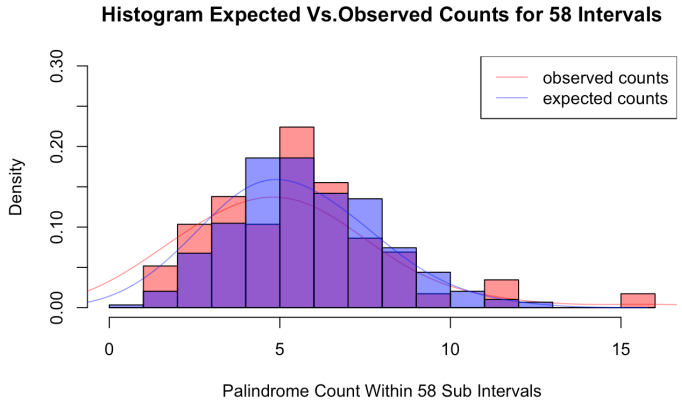
3) Analysis of Counts of Palindromes in Various Regions of Equal Length of DNA

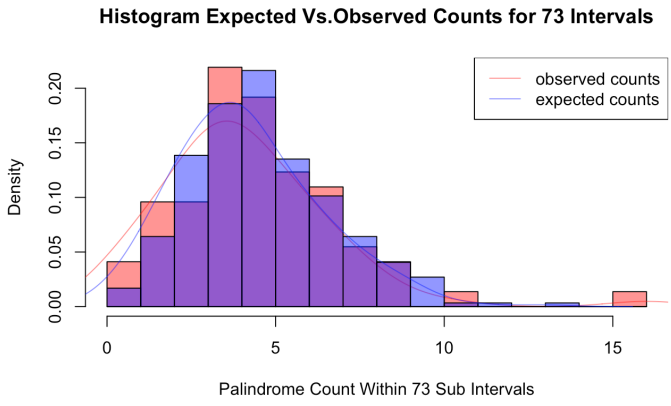
Furthermore, we analyze the counts in various regions of the DNA using both formal and graphical statistical methods in order to help us determine if the number of palindromes within an interval groupings corresponds to a uniform random scatter. The initial step of our analysis here was to determine appropriate subintervals that aren't too large or too small in order to decide if the frequencies of the counts within these sub-intervals follow the distribution of an uniform random scatter. Below in figure 3.1 we try different smaller, larger, and medium sizes of sub-intervals to see what interval sizes provide the best fit for the observed counts with the expected counts for a uniform random scatter that we can find using the Poisson process model with the rate (lambda) for a random scatter.

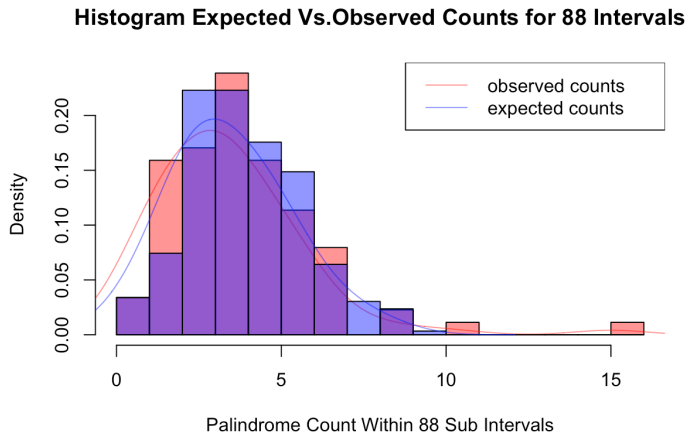
Figure 3.1: These overlaid density histograms show the effect on the distribution of the counts within each various equal non-overlapping regions of the DNA with the bases divided into the sub-intervals of size 28, 43, 58, 73, and 88. (Refer to Appendix Table 3.1 for observed and expected count values)











We see in Figure 3.1 that when choosing too long of interval length (small amount of intervals) that this results in a left skew since too many of the data points are aggregated together in the same few bins resulting in high frequencies for just a few bins and misrepresentation smaller and larger values that now fall under the same category as largely differing values, and ultimately not providing clear enough detail of the values to compare with our expectation. On the other hand if we choose too short of intervals length (large number of intervals), we see that the distribution then becomes right skewed since this would widely disperse and reduce the data frequencies and consequently giving us too much granular detail where it becomes insufficient to compare with our expected values since too many values would now contain zero palindromes.

Table 3.1: Results From Pearson's Chi-Squared Goodness of fit test on the counts in each interval

Equal Non-Overlapping Interval sizes	P-Value	X-Squared	Degree of Freedom
28 Intervals	0.33	85	80
43 Intervals	0.3184	102	96
58 Intervals	0.2882	153	144
73 Intervals	0.2976	136	128
88 Intervals	0.2976	136	128

Furthermore, we use formal statistical testing with the chi-squared goodness of fit framework since we meet the required assumptions of having enough counts and the data making up the counts are also independent. Therefore, we can use this framework to test the null hypothesis of the observed data being a good fit to the expected values modeled from the Poisson process model for a random scatter, versus the null hypothesis that it is instead not a good fit of the true proportions (expectation). After conducting the test and attaining the values in table 3.1, we see that these values support the differences in the density plots in Figure 3.1 and how the medium size of 58 intervals provides the p-value that is most in our favor when assuming the null hypothesis. However since this smallest p-value from 58 intervals is still quite greater than the conventional significance level of 5%, and the data suggests that we may

support the alternative hypothesis that these intervals do not provide a good fit to the frequencies expected with an uniform random scatter.

4) Analysis of Palindrome Clusters Within Intervals in Various Regions of Equal DNA Length

To test whether or not the interval with the greatest number of palindromes indicates a potential origin of replication we test the probability in which the interval with the maximum number of palindromes will appear. Formally our null hypothesis is that the interval with the greatest number of palindromes is a typical occurrence in our data, using the MLE estimator for the poisson process as the expected outcome. Our alternative hypothesis is that the interval with the greatest number of palindromes is *not* a typical occurrence in our data. For our testing we used the standard significance level of 0.05 to determine whether or not a cluster of that size is unusual. We used intervals starting at 38 ranging up to 78 to try and see the behavior of our p-value as the number of intervals change.

Table 1: Results from our chi squared goodness of fit tests on intervals ranging from 38 to 78

# Intervals	$\hat{\lambda}$ - Estimator	P-value	Max Count	Interval of Max Count
38	7.789	0.044	18	(90534, 96570]
48	6.166	0.032	16	(90785, 95564]
58	5.103	0.0050	16	(90950, 94905]
68	4.35	0.0009	16	(91067, 94439]
78	3.79	0.0008	15	(91153, 94093]

After conducting our tests we found that every interval that we tested was under our threshold of 0.05, had max counts ranging from 15 - 18, and were roughly found in the interval range of 90,500 to 95,500. With our p-values < 0.05, we reject the null hypothesis that the interval with the greatest number of palindromes is a usual occurrence in favor of the alternative hypothesis. This implies that these clusters of palindromes may potentially be a site of replication for CMV.

Advanced Analysis

In our advanced analysis we further analyze the distribution of the number of palindromes across the regions of the dataset using more conclusive methods through simulations. These more definitive approaches better analyze the distribution of our data by comparing it to the simulated theoretical uniform distribution of our data and by using further statistical tests that allow us to better understand the distribution of our data and to more confidently determine if our data follows a random uniform distribution. When conducting this analysis, we considered the properties of uniform distributions and used the property that

uniform distributions have equal intervals of data and respective equal probabilities to construct this analysis approach. In turn, this property tells us that if we were to divide the data into 2 equal intervals that the number of observations in each halved interval should be the same according to this definition of uniform distributions.

Therefore, using this information we conducted a simulation where we generate 50,000 theoretical random data values using the same sample size and sample space as our actual data, in order to see where our actual data falls under the simulated theoretical uniform distribution of our data. This will allow us to better determine how well our data fits or doesn't fit the theoretical random uniform distribution through this simulation and the use of statistical tests. The test statistic that we are performing is the absolute difference of counts between the 2 halved intervals which we perform 50,000 times to get the theoretical distribution shown in figure 9 along with the red line for our observed data difference that we also calculate. With these we were about to both analyze graphically and more formally with a p-value that was computed by finding the percent of simulated observations that had greater absolute differences than that of our observed data. Thus, using this test statistic of absolute differences we are able to conduct a hypothesis test where our null hypothesis is that there is no difference between the observed and theoretical difference in counts suggesting that the data follows a random uniform distribution. Versus the alternative hypothesis that there is a difference in these counts suggesting that the data does not follow a random uniform distribution.

Figure 3.1: Histogram showing the distribution of 50,000 differences in the 2 halves of simulated intervals along with the red line for the observed difference in counts of the actual dataset.

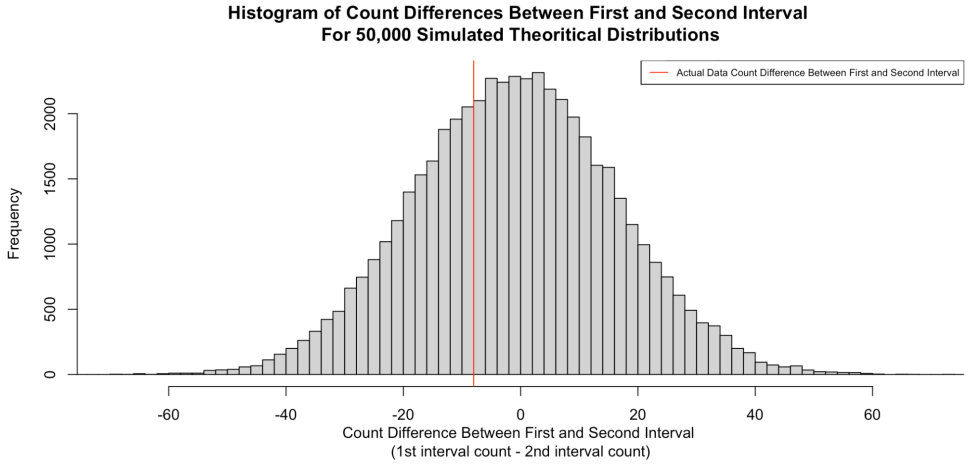


Table 9: P-value for simulation using the test statistic of absolute differences, and to the right is the value for the observed difference of counts (interval half 1 - interval half 2).

P-Value for Simulation Hypothesis Test	Observed Difference in Counts (1st half - 2nd half interval) of Observed Data (red reference line on histogram)
0.6034	-8

We can see above that the observed difference of -8 in the 2 halved intervals of our data almost aligns with the most frequent values towards the center of 0 which provide higher p-values for these values that are suggested to be acceptably small differences under the theoretical random uniform distribution of our data. Furthermore, because our observed value is near the center of the acceptable range of values, we ultimately get a large p-value of 0.6034 which is much greater than the conventional significance level of 0.05, which suggests that our

observation does fit under the theoretical simulated random uniform distribution of this data. Therefore, since we have a p-value of 0.6034 > 0.05 we fail to reject the null hypothesis of there being no difference between the observed and theoretical difference in counts, which suggests that our observed difference in counts is small enough to be considered consistent with the random uniform distribution of this data, however despite having a large p-value which does help our confidence in our hypothesis however it does not indicate causation since it is possible still that this outcome occurred due to chance and random variability.

Conclusions and Discussion

For the first section of our analysis we compared the locations of our observed values to the locations of several other randomly sampled locations from a uniform distribution of the same sample size and population size. We found that the random samples seemed similar to the observed samples, however there were outliers in the observed samples at locations ~90,000 and ~190,000. From this analysis alone we could not determine whether or not these were outliers or potential replication sites, however it did mark a point of interest for further analysis.

Next, we compared the spacings and locations of our observed values to that of a random scatter. To do this we compared the spacing of consecutive palindromes and the sum of pairs and triplets of palindromes to see if we could distinguish any particular sites with more frequent palindromes. We found that at our previously found outlier sites there were clusters of palindromes that could imply a potential replication site.

Afterwards, we then compare various interval lengths and the effect it may have on the distribution of the interval count frequencies. We see that using shorter intervals captures more granular data, however this causes the distribution to be right skewed since too many of the values will now be zero or have small frequencies due to the granularity which prevents us from accurately determining if the values match the expected pattern. On the other hand we see that when choosing too long of intervals that the distribution of frequencies become left skew since too many of the data points are grouped together in the same few bins resulting in high frequencies for just a few bins and misrepresentation smaller and larger values, which also prevents us from accurately fitting our observations to the expected uniform random scatter. Additionally, we performed chi-square goodness of fit testing as well to further support these plots in how intervals that aren't too large or small provide the best fit, which ultimately we didn’t support the hypothesis that the observed counts fit the expected uniform random scatter since all the intervals tested resulted in p-values much higher than the conventional significance level of 5%.

Lastly, we analyzed the largest cluster in the sample (~90,000) to see if it may or may not be a potential replication site. To accomplish this we performed a chi squared test using the poisson process to estimate lambda. We found that for varying interval sizes ranging from 38 to 78 this cluster of palindromes remained the largest and most significant by calculating the p-value. All of our p-values were under our threshold of 0.05 and suggested that this cluster is unlikely to have occurred purely by chance.

Ultimately, this dataset provided us with enough information in order for us to conduct a basic analysis statistically by analyzing if there are unusual clusters of palindromes and possibly identify if they are indeed an origin of replication or are simply an artifact of randomness. However, although this data is independent and doesn't seem to have much limitations since it was obtained through an algorithm that sequenced many types of patterns in DNA, it is limited in the sense that we were only provided with one sequence, which has the potential to have variability and not generalizable since we are not certain that the algorithm is truely generalizable since it is not specified that the algorithm accounts for possible mutations. With this being said we would like to consider this in our future work, because by having more sets of sequences we can more accurately interpret many of our results since we can cross-reference our results with multiple other sequences' results to help ensure that our observation wasn't just out of chance and random variability.

Works Cited

"About Cytomegalovirus and Congenital CMV Infection." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 18 Aug. 2020, www.cdc.gov/cmv/overview.html.

Appendix

Table 3.A: Observed and Expected values under Poisson distribution for the counts for the subinterval sizes of 28, 58, 88.

Palindrome Count Values	**Observed Counts For 28 Intervals**	Expected Counts for 28 Intervals	**Observed Counts For 58 Intervals**	Expected Counts for 58 Intervals	**Observed Counts For 88 Intervals**	Expected Counts for 88 Intervals
0	**0**	0.0007178686	**0**	0.352394051	**3**	3.045608
1	**0**	0.007588896	**5**	1.79842481	**12**	10.244317
2	**0**	0.040112737	**7**	4.58908401	**14**	17.229078
3	**0**	0.141349646	**4**	7.80671762	**23**	19.317451
4	**0**	0.373566922	**9**	9.96029490	**18**	16.244220
5	**1**	0.789827207	**8**	10.16636997	**8**	10.927930
6	**2**	1.391600316	**6**	8.64725721	**3**	6.126264
7	**1**	2.101600478	**13**	6.30440427	**6**	2.943789
8	**3**	2.777114917	**2**	4.02177514	**0**	1.237730
9	**4**	3.262007998	**2**	2.28054682	**0**	0.462586
10	**5**	3.448408455	**0**	1.16386527	**0**	0.155597
11	**4**	3.314054879	**0**	0.53997511	**0**	0.047579
12	**2**	2.919524536	**1**	0.22964459	**0**	0.013337
13	**2**	2.374118853	**0**	0.09015225	**0**	0.003451
14	**0**	1.792701991	**0**	0.03286338	**0**	0.000829
15	**2**	1.263428070	**0**	0.01118110	**1**	0.000186
16+	**2**	2.002994098	**1**	0.35744353	**0**	3.045656

Table A: Formal testing upon the count data within our simulation to support and verify our results from our test statistic simulation in advanced analysis (Extra, used to check work in advanced analysis)

P-Value of Observed Data	X-Squared of Observed Data	X-Squared of Expected Data under theoretical distribution and using significance level of 5%
0.6419	0.21622	3.84

Figure A: The top histogram is the x-squared distribution of our simulated theoretical data with the actual observed x-squared value as the overlaid red line, then when comparing this to the plot on the chart below it, we can visualize the p-value of this statistic in red which is much more than the conventional blue shaded area when using a 5% significance level. (Extra, used to check work in advanced analysis)