Are Palindrome Clusters Potential Replication Sites?

Jiawei Yang, UC San Diego

1. Introduction

Complimentary palindromes are related to virus replication, evidenced by examples of those in the DNA of members of the herpes family. To develop virus-combating strategies that would inhibit the replication process, the patterns of palindromes provide information to target locations of potential treatment imposition. In this report, we analyze the pattern of palindrome distributions in the DNA of cytomegalovirus. Cytomegalovirus, abbreviated as CMV, is a member of the herpes virus family. We try to verify the existence of statistically significant palindrome clusters. We first hypothesize that if there is no clustering of the palindromes, their location shall follow a uniform distribution. We generate simulations of uniform scatter for several groups and analyze the scattering of the palindromes in the DNA sequence and investigate the spacings between consecutive palindromes. We include the sum of consecutive pairs and triplets to analyze the spacings. We partition the sequence into segments with different intervals to investigate whether sub-intervals display non-uniform distribution. With graphical and numerical methods, we identify the counts of palindromes in each region and the intervals with the greatest palindromes number.

Data

The DNA sequence data of CMV published in 1990 contains 229,354 letters in length. In this report, we restrict the palindromes to those with a certain length, specifically the ones between 10 and 18 base pairs long. The data includes the locations of 296 qualified palindromes, each being an integer value ranging from 0 to 229,354.

2.1 Random Scatter:

Method:

We generate samples with pseudo-random numbers by drawing 296 integers from 1 to 229,354; drawing without replacement is specifically ensured since we are simulating the real case where at most one palindrome can be at one location on the DNA strand. We compare the simulations with the given data of CMV palindrome distribution. We graphically examine the distribution of the palindromes by two types of plots: empirical cumulative distribution function and location by index. The location by index approach orders the palindromes by their places (from the first to the two-hundred-and-night-sixth) on the sequence.

We exhibit simulations in two ways: individual simulation and combined information. For individual simulations, four random simulations are picked, named as group 1, 2, 3, and 4. (See Appendix for how these four groups are generated) For the combined information, we run five hundred simulations, take the median location out of the five hundred simulations for each location from the first to the two-hundred-and-night-sixth: this sample constructed by the medians is treated as one individual group with location information extracted from the five-hundred simulations. We name it the median group, and it is believed to represent an ideally typical distribution of palindromes under the normal assumption.

Analysis:

Figure I: Empirical CDF Comparison (CMV vs. group 1 2 3 4)

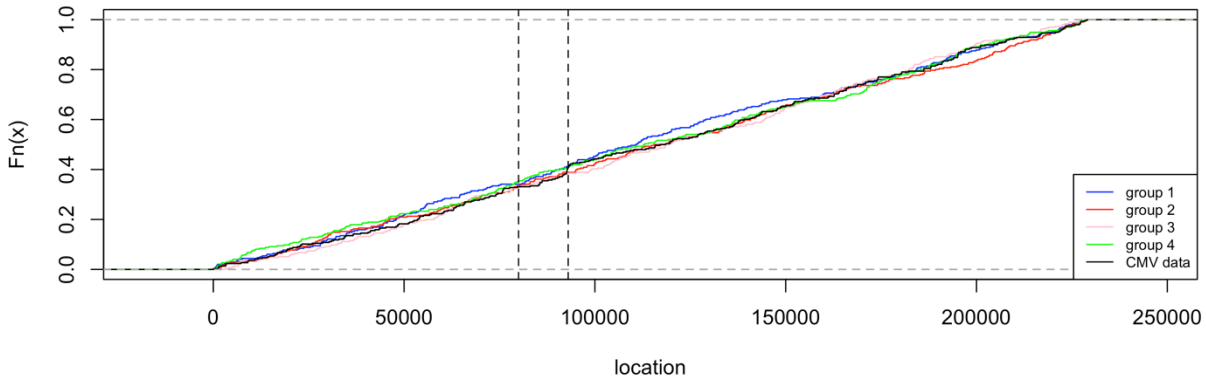
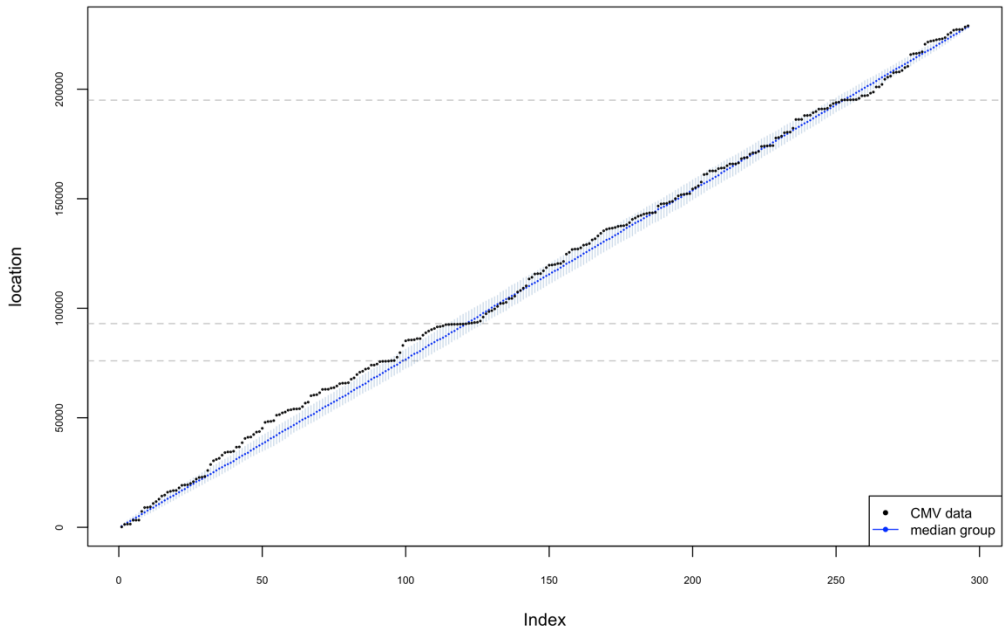


Figure I is the combined CDF comparison of the CMV data against the 4 individual simulation groups. We observed that, for most of the locations, the empirical CDF of the CMV data is sandwiched by those of the simulation groups: there is at least one group featuring a higher and lower cumulative distribution at most of the locations. It is noticeable that for one specific interval approximately from 80,000 to 93,000, the CDF function of the CMV data shows a lower cumulative distribution than all the four groups (graphically captured by the dashed line interval). The rapid rebound after 93,000 is also noticeable which implies a dense allocation of palindromes.

We further verify the claim made in the previous paragraph with location by index plot.

Figure II: Location by Index (CMV vs. Median group with width of middle 50%)



Note. For each index, the location of the middle fifty percentiles is plotted, centered at the median

Figure II and Appendix I are location-by-index scatterplots. From the distribution of simulated palindrome distribution from the median group (the dark blue points), we observe the baseline case: for a typically uniform distribution, we expect the location to increase linearly with the position ranking of the palindrome. From Appendix I, no palindrome in the CMV DNA features a location considered as an outlier defined by the 500 simulations we generate. We then narrow the margin to the interquartile range

as shown in Figure II. Here we observe that, for the first one-third palindromes, the increase in location for an additional index is greater on average for the CMV palindromes than those from the median group: for this part of the DNA, the CMV palindromes are further apart from each other. We also observe piece-wise stagnation of location in Figure II, most obviously near the 76000^{th}, 93000^{th}, and 195000^{th} locations. (marked with dashed line) At these places, the consecutive palindromes are close to each other. This pattern confirms our earlier findings from the CDF graph, and the pattern also provides insights to our later discussion on locating clusters of palindromes.

Conclusion:

The empirical CDF plots show a lack of significant evidence that the overall distribution of palindromes differs from specific simulated random scatters. However, when we compare the location-by-index plot of the CMV sample to that of the median group where we intentionally control the sampling variability, we discover fewer uniform characteristics of the CMV palindromes. However, the median group may not be representative in the following analysis due to pre-manipulated randomness. The first one-third of the palindromes are more loosely located with piece-wise clustering the most predominantly around locations of 76000, 93000, and 195000.

2.2 Locations and Spacings

Method:

To assess whether the palindrome clusters are located in specific spots, we compare the differences in spacings of consecutive palindromes, the sum of consecutive pairs, and the sum of consecutive triplets between the CMV dataset and randomly generated samples. We calculate the sum of consecutive pairs and triplets; and we calculate the spacings between consecutive palindromes, spacings between the sum of pairs and triplets; and then we compare them in numerical and graphical methods. We give statistical summaries for spacings and visualize the comparison by computing and plotting the empirical cumulative distribution functions.

Analysis:

Below is a table of descriptive statistics of the spacings between consecutive palindromes:

Table I : Summary Statistics of Spacings between Consecutive Palindromes

Group	1st quartile	median	3rd quartile	mean $\bar{x}$	Standard deviation s
CMV Data	160	512	1144	775	833
Random(Group1)	207	535	1026	763	804
Random(Group2)	200	579	1035	770	756
Random(Group3)	198	529	1095	773	762
Random(Group4)	215	516	1081	775	810

For the numerical comparison, we tell from the table that the palindromes have similar spacings with all groups in the simulated data, evidenced by similar means, medians, and standard deviations.

Below is the plot of the empirical cumulative distribution function of spacings between the consecutive palindromes. The distribution comparison also shows similarities in spacings given the overlaps in lines. And the two scatterplots of spacings below show consistency with our speculation.

Figure III: Ecdf of Spacings between Consecutive Palindromes

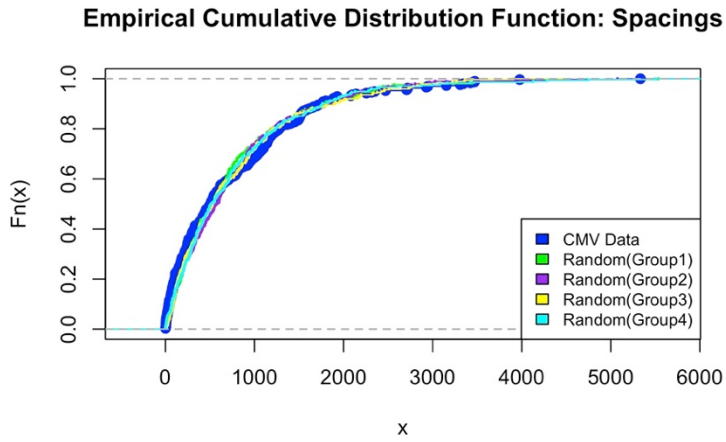
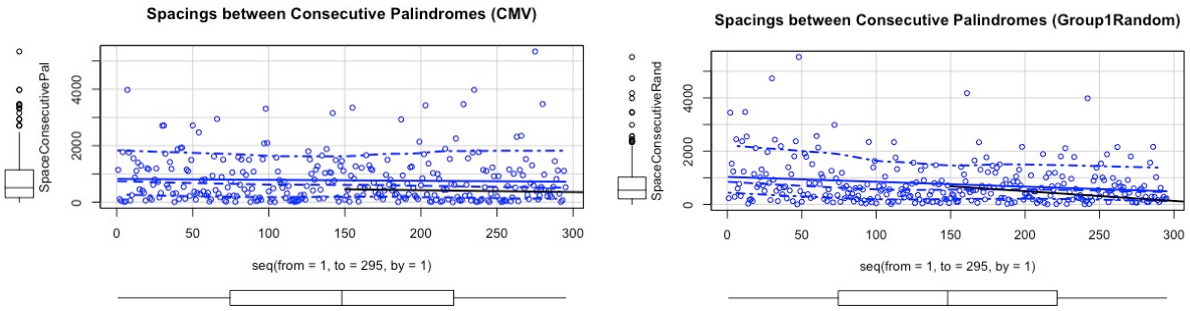


Figure III: Scatterplots of Spacings between Consecutive Palindromes



The spacings between sum of pairs illustrate the maximum distance that covers only one palindrome after a palindrome until meeting the next. And the spacings between sum of triplets mean the maximum distance that covers two adjacent palindromes until meeting the third one.

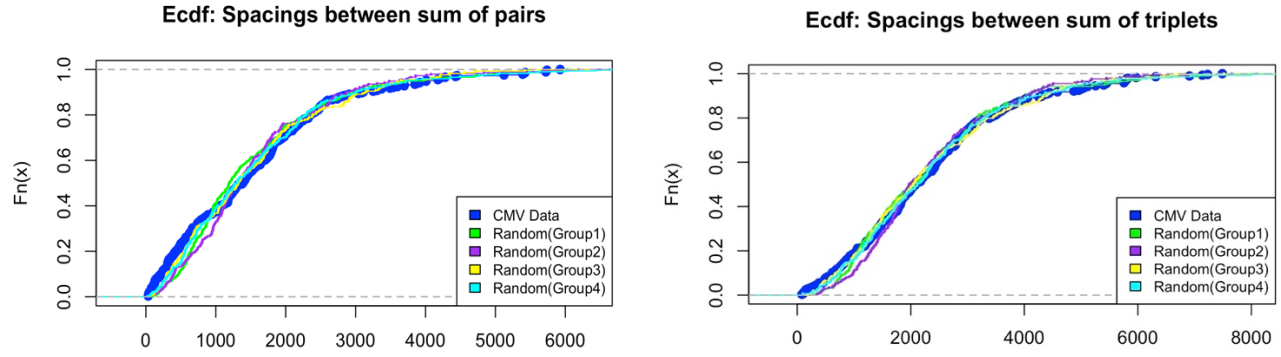
Below is a table of descriptive statistics of the calculated spacings of sum of pairs and triplets from the CMV dataset and simulation datasets, where we choose the results for group 1 for reference.

Table II : Summary Statistics of Spacings of Sum of Pairs and Triplets

	Group	1st quartile	median	3rd quartile	mean	standard deviation
Pairs	CMV Data	559	1386	2148	1150	1230
	Random (Group1)	721	1193	1984	1530	1125
Triplets	CMV Data	1271	2078	3021	2327	1472
	Random (Group1)	1207	2048	3047	2290	1404

Below are the plots of empirical cumulative distribution functions presenting spacings between sum of pairs and triplets drawn from the CMV and simulated data.

Figure IV: Ecdf Plots of Spacings between Sum of Pairs and Triplets



The spacings between the sum of pairs and triplets give more detail about the locations of palindromes and act as indicators of how adjacent palindromes are spread out. By the numerical and graphical comparison, we tell the similarities in statistical summaries and the empirical cumulative distributions of spacings between pairs and triplets. We infer from the comparison results that spacings do not make a difference between the real CMV data and our simulated data.

Conclusion:

Through comparing the spacings of individual consecutive palindromes, the sum of pairs, and the sum of triplets, we conclude that the spacings are similar between real data and simulated ones, indicating similar locations of palindromes in two categories.

2.3 Count:

Method:

We investigate the number of palindromes occurring at different regions of the DNA: we derive the regions by portioning the DNA strand into non-overlapping consecutive unit intervals of equal length. Under the null hypothesis that the distribution of palindromes is uniform, we assume the expected instances of palindromes to be independent and, at most, one at a location. We experiment with different interval lengths and compare the sample count distribution with that from the four random groups. Supported by our previous assumption, we expect the counts to follow a uniform Poisson process:

$$X \sim Pois(\lambda)$$

$$P(k\ points\ in\ a\ unit\ interval) = \frac{\lambda^k}{k!}e^{-\lambda},\ k \in \mathbb{N}$$

λ, the rate parameter, is the expected number of palindromes per unit interval calculated as:

$$\lambda = No.\ of\ palindromes \cdot \frac{unit\ interval\ length}{DNA\ length}$$

Analysis:

We experiment by segmenting the DNA strand into 100 and 60 intervals to expect approximately 3 and 5 palindromes on average in each unit interval if uniformly distributed. Results as shown:

Figure IV: Density Plot of Intervals Containing X count of palindromes

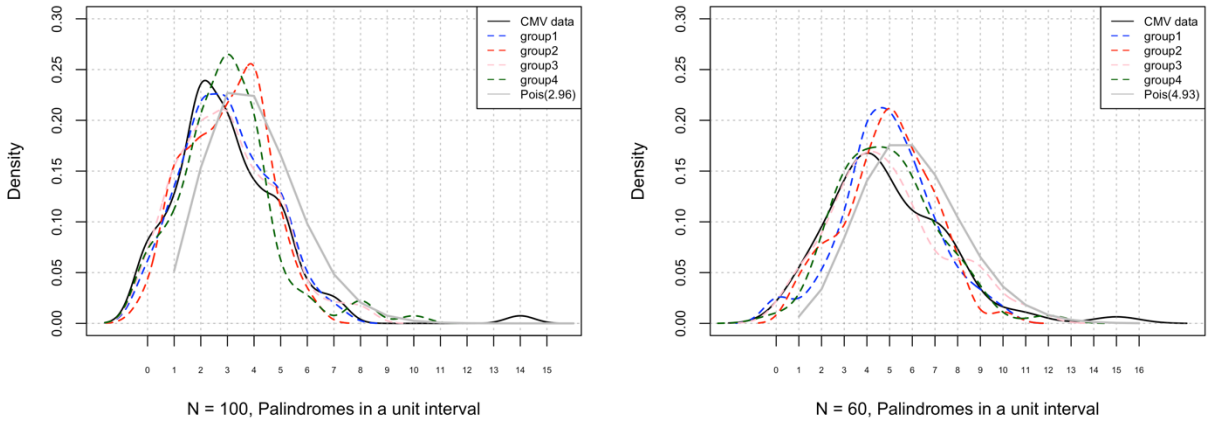


Figure IV shows the density plot of counts of palindromes in a unit interval for cases of 100- and 60-unit intervals. The theoretical Poisson approximation is plotted in grey. When comparing the CMV data against the random samples, we do not observe significant differences in the density plot: especially the density plot of the CMV counts is close to group 1 and 3 for N = 100, group 3 and 4 for N = 60. (see figure IV). This provides evidence that the randomly generated group 3, in particular, behave similarly to the CMV data under both partition criteria. However, if we compare the CMV count density to the theoretical Poisson density, we find the CMV count density is approximately a parallel left shift of the theoretical density curve: precisely, the CMV data shows a higher density for intervals with fewer counts of palindrome and a lower density for intervals with higher counts of palindrome.

Conclusion:

We observe the number of counts within unit interval in the CMV data is overall compatible with those from our simulated groups. However, we observe a systematic left-shift of the density curve when comparing against the theocratical Poisson, and we therefore cannot conclude compatibility with theoretical Poisson. We propose that this could be likely due to dense clustering of the data where more than 10 counts; Such clustering dilutes the density for the intervals with around 5 to 10 counts, and simultaneously increases the density for low-count intervals.

2.4 The Biggest Cluster

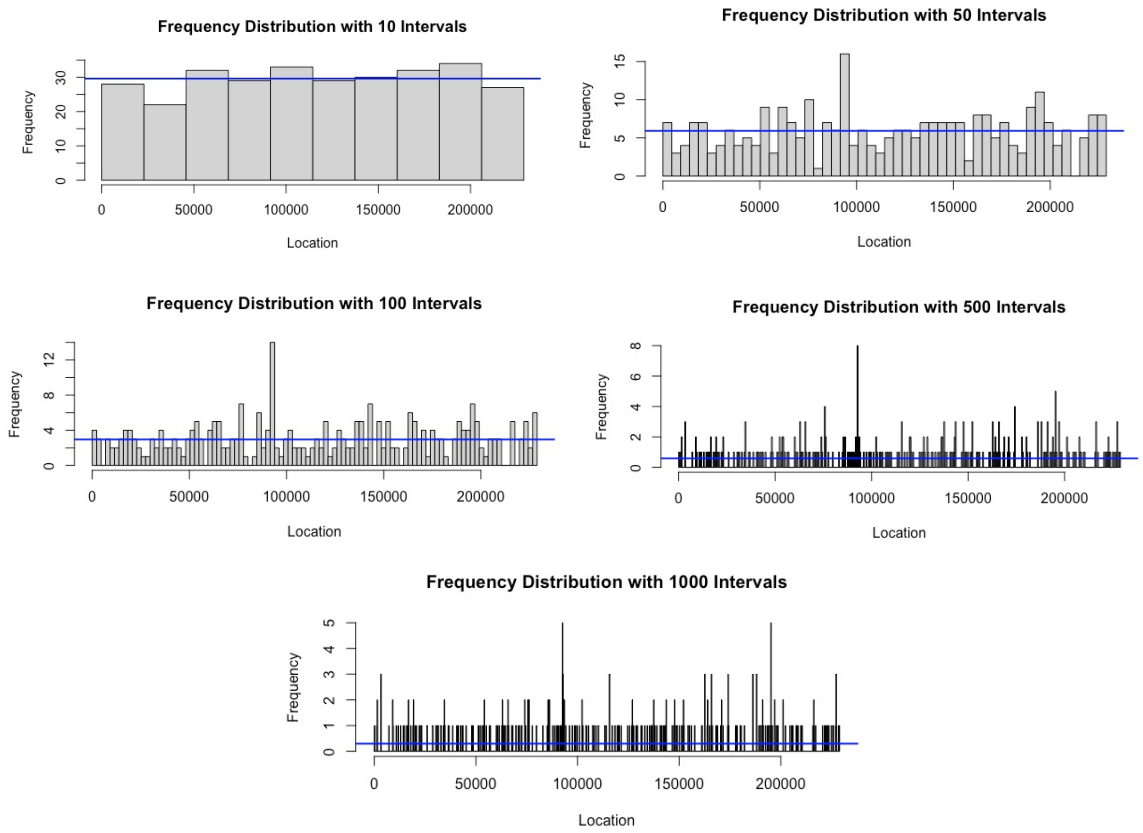
Method:

To identify whether intervals with the greatest number of palindromes would indicate a potential origin of replication, we visualize the frequency of palindromes under different numbers of intervals, including interval numbers of 10, 50, 100, 500, 1000. We plot histograms for frequency distributions. As the interval width decreasing, we narrow down the scope of each interval group and identify the cluster with the greatest number. Within the specific clusters, we compare it with a uniform distribution to see if it indicates an abnormal distribution.

Analysis:

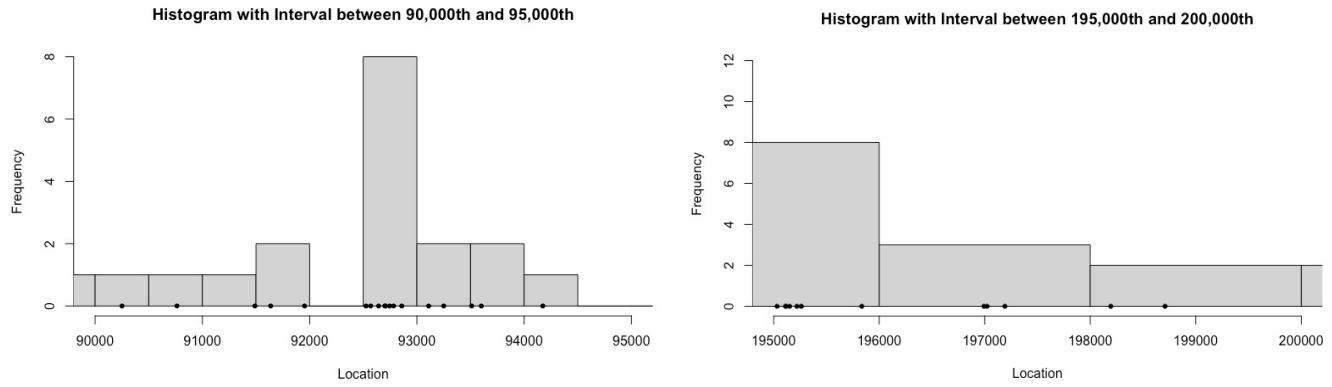
Below are the histograms of the frequency distributions with interval numbers of 10, 50, 100, 500, 1000. A baseline is supplemented for each histogram for reference.

Figure V: Histograms of Frequency Distributions



From the five histograms with different intervals, the frequency differences become significant: when interval is small as 10, there's little discrepany and deviation from the baseline; when it increases to 50 and 100, the location left to 100,000th place indicates a highest frequency; when interval is 500 or 1000, another cluster near 200,000th become prominent with high frequencies, too. The latter two distibutions suggest that between around 90,000th to 95,000th and between 195,000th to 200,000th place, it indicates the highest and second highest frequency of palindromes. We display the histiograms within these intervals and we also plot the locations of palindromes on the x-axis.

Figure VI: Histograms with Specific Intervals



Both closed-interval distributions show non-uniform patterns of frequencies, and we also observe that the palindromes are densely distributed near 93000th and 195000th , which show a significance of palindrome numbers within the intervals. As it does not follow a uniform distribution, we infer that it is a sign for a potential site of replication.

Conclusion:

We conclude that the palindromes are spread out with a potentially non-uniform pattern as the interval widths are narrowed. By taking a closer look, we observe that the highest frequencies would occurr between around 90,000th to 95,000th and between 195,000th to 200,000th intervals; and the distribution within these intervals tend to have non-uniform patterns as well. With the information, we conclude that the interval with the greatest number of palindromes indicates a potential origin of replication with different intervals.

3. Advanced analysis.

Method:

To assess whether the palindromes are uniformly scattered, we divide the locations of palindromes into 10 equal subintervals, with a unit interval length of around 23,000. Under our assumption that palindromes are uniformly scattered, the expected frequency within each subinterval would be 29.6. We summarize the actual frequency and carry out a Chi-square test to verify whether the observed frequencies match the expected ones. Before performing the Chi-square goodness of fit test, we state our hypothesis as followed, using a significance level of 0.05:

$$H_0: There\ is\ no\ significant\ difference\ between\ the\ observed\ and\ expected\ frequency.$$

$$H_a: There\ is\ significant\ difference\ between\ the\ observed\ and\ expected\ frequency.$$

Analysis:

Below is a table of the observed and expected values for frequency:

Table III : Observed and Expected Frequency

Interval	1	2	3	4	5	
Observed	29	21	32	30	32	
Expected	29.6	29.6	29.6	29.6	29.6	
Interval	6	7	8	9	10	Total
Observed	31	28	32	34	27	296
Expected	29.6	29.6	29.6	29.6	29.6	296

In the Chi-square goodness of fit test, the X-squared value is 4.1351 with a df of 9. The p-value is 0.9023, much higher than the significance level. Therefore, we fail to reject the null hypothesis and suggest that the frequency of palindromes within each interval is not different from the expected ones.

Conclusion:

Through the Chi-square goodness of fit test, at a 5% significance level, we fail to reject the null hypothesis and conclude that there is no significant difference in the observed and expected frequency in each unit interval.

4. Conclusions and Discussion

Our previous analysis indicates a lack of evidence that the overall distribution of CMV palindromes is significantly different from one generated from random samples. However, noticeable patterns, especially one considering dense clustering of the palindromes, are observed when we look at smaller sub-intervals of the DNA strand. Supported by results from palindrome spacing and interval analysis, we recommend prioritizing the testing on the neighborhood of these two locations: 93110 and 195032, both featuring an abnormally high density of palindromes.

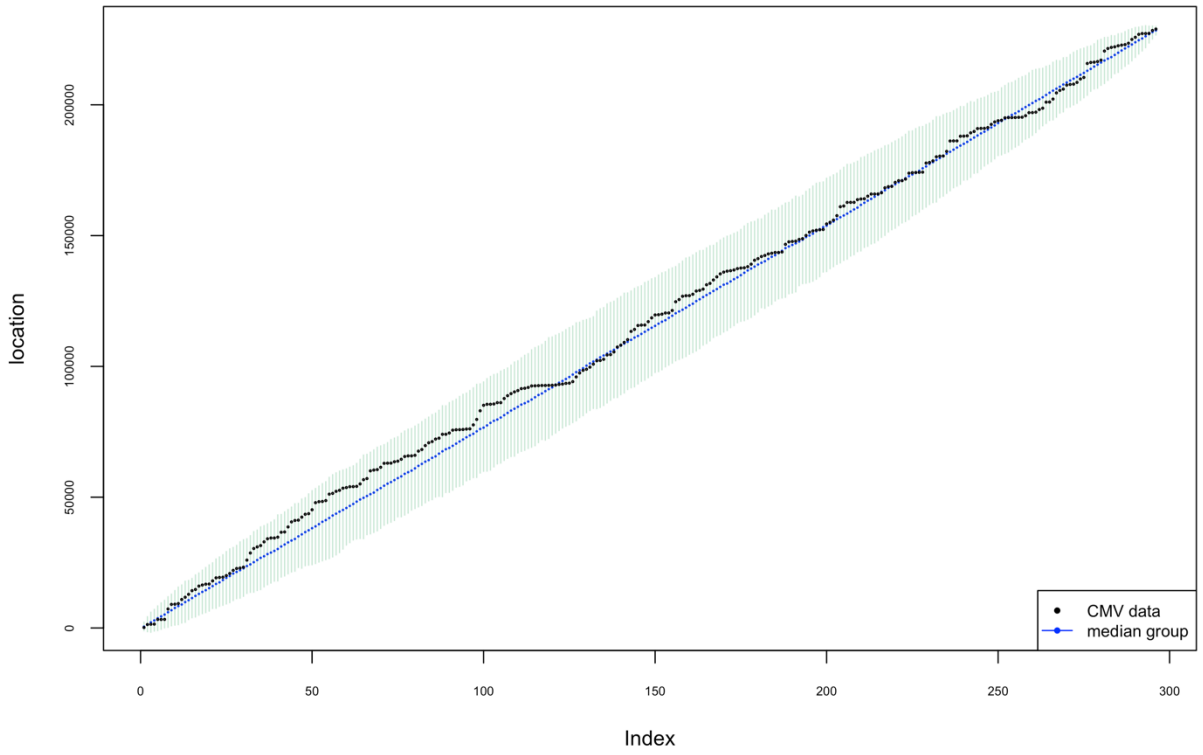
However, our analysis could be furthered with a more comprehensive background in biology. In this report, we base our analysis on the assumption that the CMV data is tested against the uniform distribution to find a place where dense clustering of palindromes is the sign of replication sites. The direction of testing would be more credible if we have a better model if any than the uniform distribution to approximate the natural scattering of the palindromes. Also, if there are specific permutations, other than simple clustering, of palindromes that could also signal origins of replication, they could be neglected and require further exploration.

We also want to stress the limitations of our process that our assumption concerns uniform distribution and the uniform Poisson process that are both applicable under assumptions of independence between palindromes. Here is a possible drawback: if the occurrences of one palindrome increase the probability of other palindromes around it, then our model may lose credit and covariance-related analysis could be a more reliable tool for assessing the distribution of the palindromes.

5. Appendices

Appendix I:

Location by Index (CMV vs. Median group with width of IQR)



Note. For each index, the location of the maximum and minimum whisker, centered at the median

Maximum whisker = Q3+1.5IQR, Minimum whisker = Q1-1.5IQR