Statistical Analysis of CMV DNA Data

2024-11-10

Contribution

Student 1 was responsible for Analysis Questions 3, 4, and 5. Student 2 was responsible for Analysis Questions 1, 2 and the Advanced Analysis. Both Student 1 and 2 were responsible for the Introduction and Discussion.

1. Introduction

Cytomegalovirus (CMV) is a complex DNA virus with a genome comprising 229,354 base pairs. DNA sequences contain many patterns, and some of these patterns may flag important sites on the DNA, such as the origin of replication. The origin of replication is special patterns in the virus' DNA that contains instructions for its reproduction. To develop strategies for combating the virus, scientists want study the way in which the virus replicates. Our analysis uses data from the DNA sequence of CMV that was published in 1990 (Chee etal.) that is 229,354 letters long. Leung et al. (1991) implemented search algorithms to screen the sequence for many types of patterns. Altogether, 296 palindromes were found that were at least 10 letters long, with the longest being 18 letters. Palindromes are DNA sequences that read the same forwards and backwards, often associated with key genetic elements such as replication origins, regulatory regions, and DNA binding sites. By identifying and studying these palindromic patterns, researchers can gain a better understanding of how the virus replicates and interacts with the host cell.

The primary goal of this statistical analysis is to assess how the distribution of palindromes deviates from a uniform scatter across the DNA sequence and if any clusters are due to random chance. We applied statistical methods to compare the distribution of actual palindromic sequences in the CMV genome against what would be expected from random distribution models. We used a combination of Kolmogorov-Smirnov tests, Chi-square goodness-of-fit tests, and Poisson distribution models to analyze spacing, interval counts, and locations of palindromic sequences, testing whether observed clustering could be attributed to random chance or if it might indicate an underlying biological structure.

We found that palindromic sequences are uniformly distributed across the genome, with no significant variability in the number of palindromes observed in different regions. Analysis of palindrome counts revealed that shorter intervals had greater variability than larger intervals and that certain region lengths (500, 20000, 50000) showed clustering, suggesting that these areas are less likely to occur by random chance. Further exploration of these clusters showed statistically significant clustering at region length 500.

The findings suggest that while some small-scale clustering is evident, particularly in the 500-base pair regions, the overall distribution of palindromes does not significantly deviate from random scatter on a larger scale. This could imply that palindromes are dispersed across the genome in a manner consistent with random chance, but local variations might still hold biological significance.

The report is structured as follows: Section 1 introduces this study's objectives, background, and key questions, Section 2 analyzes the randomness of palindrome distribution, spatial patterns of palindromic sequences, palindrome counts in different genome regions, and the presence of significant clusters. Section 3 explores the robustness of statistical assumptions in studying palindrome distribution across the CMV DNA sequence. Section 4 summarizes the findings and discusses possible methods for searching for the origin of replication.

2. Analysis

2.1 Random Scatter

2.1.1 Methods

To determine if the observed palindromic sites in CMV DNA are randomly scattered or exhibit clustering, which could indicate biological significance, we simulate what a random scatter of palindromic sites along a DNA sequence of 229,354 bases would look like and use statistical comparisons to determine if the observed distribution is significantly different from what we would expect under random conditions.

Using a pseudo-random number generator, we simulate 296 palindrome locations randomly distributed along a DNA sequence of length 229,354. This process is repeated to obtain a robust sample of random distributions. To simulate random scattering, we use a uniform distribution to place 296 palindrome sites along the DNA sequence. Since we're assessing whether the real data deviates from a random scatter, we assume each location is equally likely (to have an equal probability of containing a palindrome under the null hypothesis of randomness), meaning each position on the DNA sequence has a uniform probability of selection, justifying the use of a uniform distribution.

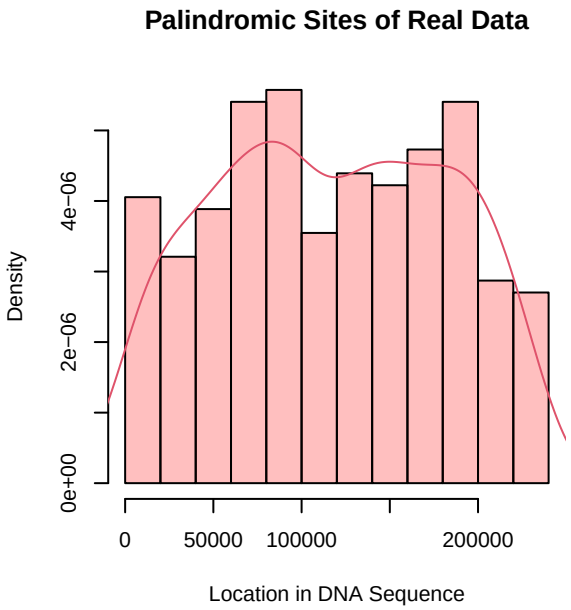
A uniform distribution for our purposes is defined by the probability $P(X = x) = \frac{1}{b-a}$ for all x in the interval $[a, b]$, where $a = 1$ and $b = 229354$, the sequence length. Each randomly generated site should be uniformly distributed over the DNA sequence, producing an expected random scatter.

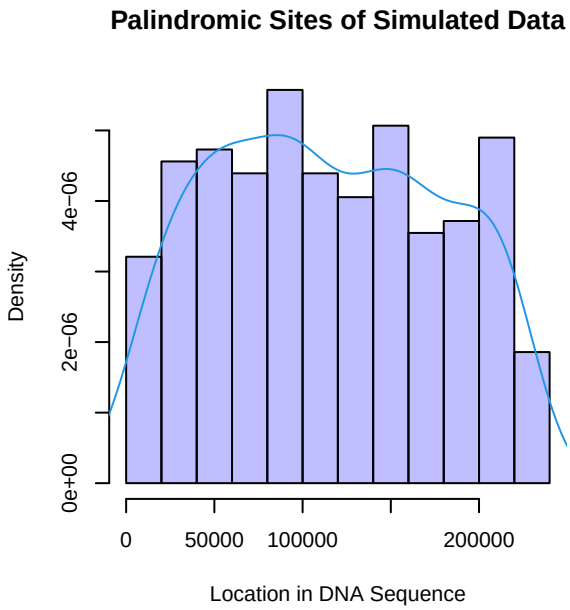
We will visually compare the distribution of actual palindrome locations, palindrome counts per intervals of 1000, and location spacing, against the simulated distributions using histograms to see if the observed results match the expectation under a random scattering model and identify potential clustering in the actual data. We use summary statistics, including mean and variance, to quantitatively assess whether the observed distribution deviates significantly from the simulated random distributions.

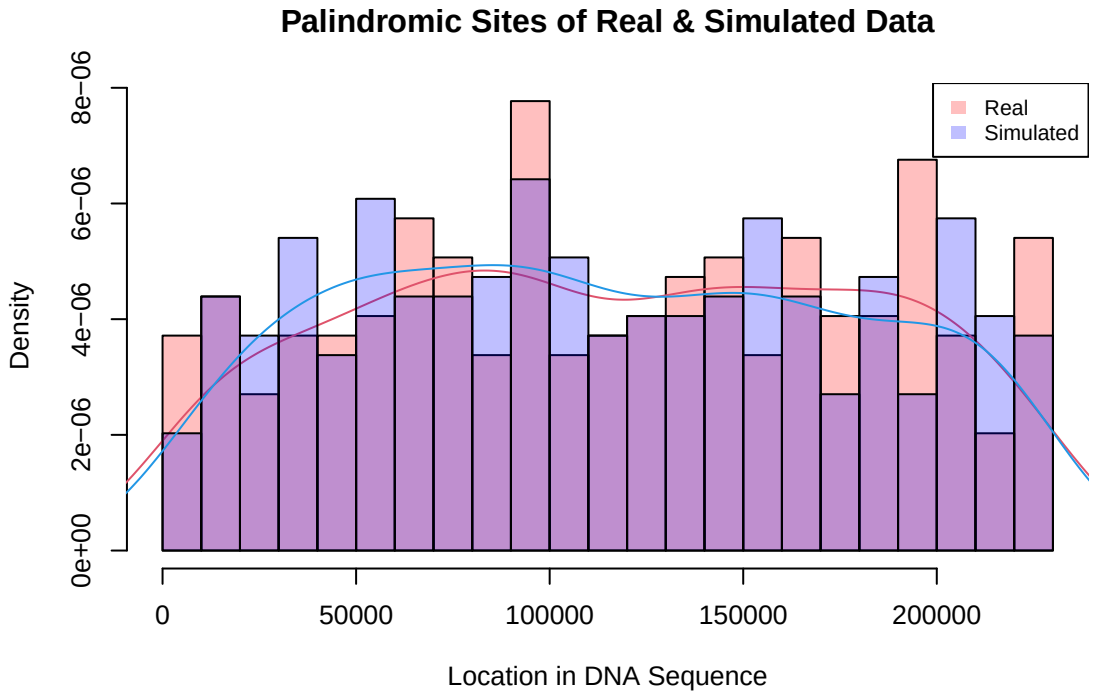
To test whether the observed distribution matches a random scatter, we use the Kolmogorov-Smirnov test, comparing the distribution of actual palindrome locations to the randomly simulated distributions. The K-S test quantifies the maximum distance between the cumulative distributions of two samples, returning a p-value that can help us determine whether to reject the null hypothesis (i.e., that the distribution of palindrome locations is random). A small p-value (e.g., < 0.05) would suggest a significant deviation from a random distribution.

2.1.2 Analysis

In the simulations, random palindromic sites are scattered across the DNA sequence, with uniform distribution expected under the null hypothesis. The histograms below visualize the distribution of randomly simulated palindromic sites, which we use as a reference to compare the actual distribution of observed palindrome locations.

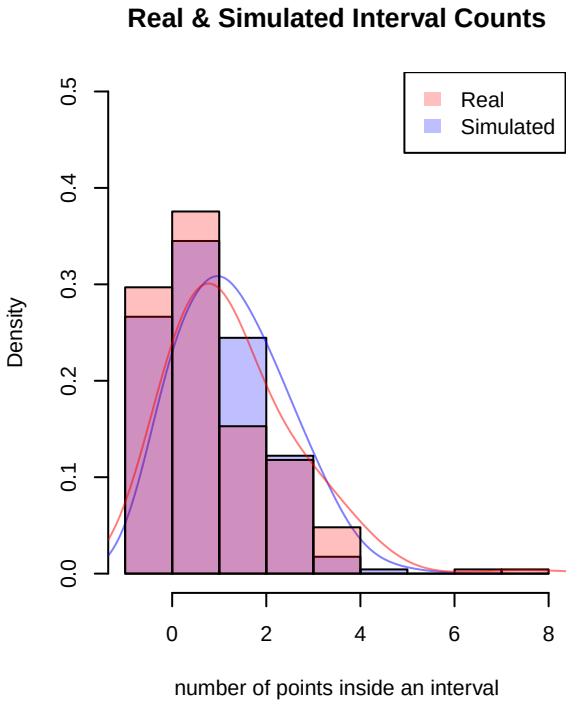


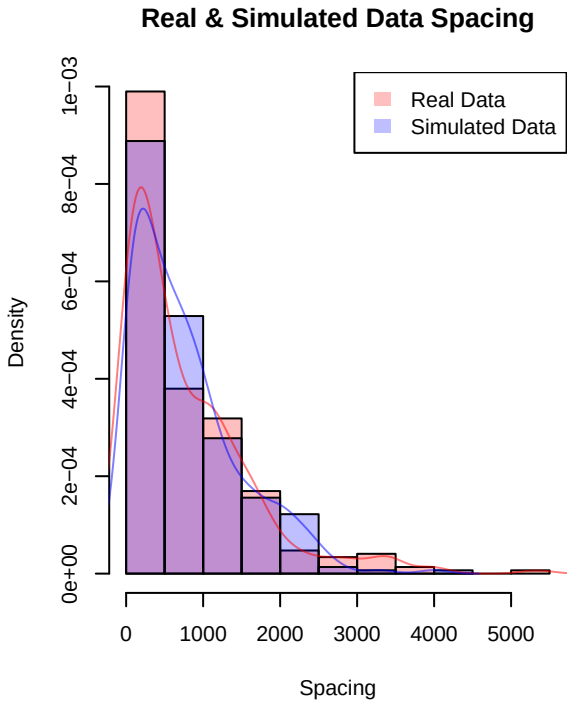




Additional histograms of repeated random scatters can be found in the Appendix.

The histograms for spacing and counts should look similar if the real data follows a random scatter. A notable deviation in real data (e.g., many more intervals with high or low counts than expected in the simulations) would suggest clustering. The spacing and counts histograms can be fund below.





We calculate the mean and variance of spacings between palindromes in both the real and simulated data. If the real data has significantly lower spacing variance, this could indicate clustering (i.e., the palindromic sites are closer together than we would expect from a random scatter).

Table 1: Quantitative Comparison of Real vs Simulated Data

Statistic	Real	Simulated
Mean Spacing	775.512	772.527
Variance of Spacing	693560.500	477052.800
Std Dev of Spacing	832.803	690.690
Mean Counts	12.727	12.955
Variance of Counts	14.589	10.807
Std Dev of Counts	3.820	3.287

The Kolmogorov-Smirnov (KS) tests were conducted to compare the distributions of real data with simulated data in three different cases: spacing, counts, and palindrome locations. Here are the results for each:

Spacing (Real vs. Simulated)

Test Statistic (D): 0.081356

p-value: 0.283

The p-value of 0.283 is greater than the common significance level of 0.05, suggesting that there is no significant difference between the distributions of real and simulated spacing. Thus, we do not have sufficient evidence to reject the null hypothesis, and the real and simulated spacing distributions appear similar.

Counts (Real vs. Simulated)

Test Statistic (D): 0.22727

p-value: 0.4767

The p-value of 0.4767 is also greater than 0.05, indicating no significant difference between the distributions of real and simulated counts. The real and simulated counts distributions are similar.

Palindrome Locations (Real vs. Simulated)

Test Statistic (D): 0.040541

p-value: 0.9681

The very high p-value of 0.9681 suggests that there is no significant difference between the distributions of palindrome locations in the real and simulated data. The locations of palindromes in the real data closely resemble those in the simulated data.

2.1.3 Conclusion

The comparison between the actual distribution of palindromic locations and the randomly simulated distributions does not show any significant deviations from a Uniform Distribution. The histogram of actual palindrome locations does not show significant areas with higher concentration of palindromic sequences than what is typically seen in the random simulations, suggesting the lack of potential clustering. The overlaid histograms visually demonstrate that the distribution of palindromic sites in both the real and simulated data are quite similar. This similarity extends to the interval counts and the spacing between sites, where the shapes of the distributions in both datasets appear comparable. These results suggest that the real data does not diverge significantly from what would be expected under a random scattering model, as the real and simulated data align closely in their overall distribution patterns.

The analysis of palindrome distributions in real and simulated DNA sequence data reveals minimal differences; palindromic sites in the real data do not significantly deviate from a random scatter. The quantitative summary shows that the mean spacing between palindromes is similar for both real (775.5) and simulated (772.5) data, though the variance in spacing for the real data (693,560.5) is somewhat higher than in the simulated data (477,052.8). While this variance difference could suggest mild clustering tendencies, the consistency in mean values implies that this clustering is not substantial. Similar values in mean and variance of palindrome counts per interval further support a uniform distribution assumption.

Kolmogorov-Smirnov tests comparing spacing, interval counts, and palindrome locations between real and simulated data all resulted in high p-values ($p > 0.05$); we have no significant evidence to reject the null hypothesis in any of these cases, confirming that the real and simulated distributions align closely. These findings suggest that the distribution of palindromic sites in the real data closely resembles a random scatter, with little evidence of significant clustering or structured patterns.

2.2 Locations and Spacings

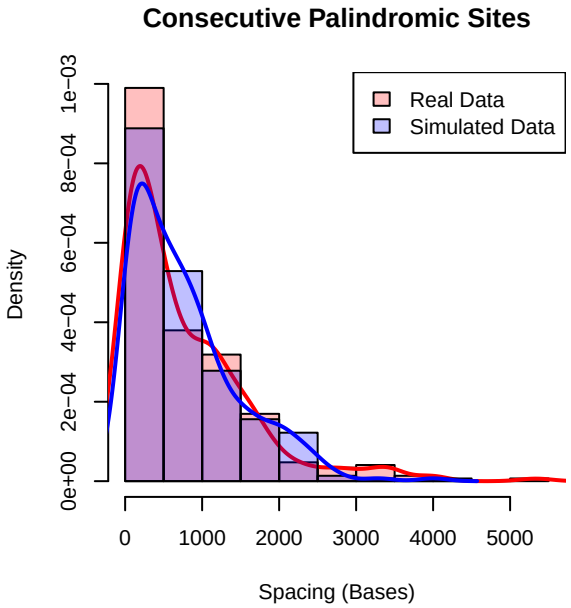
2.2.1 Methods

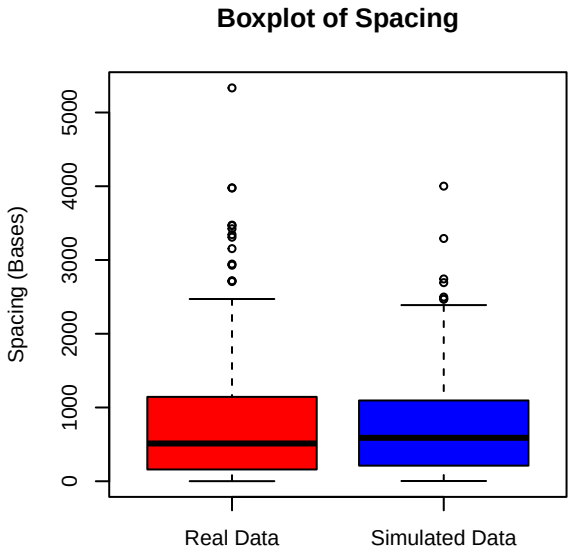
To examine the spatial distribution of palindromic sequences we will analyze the spacings between consecutive palindromic sites, as well as sums of consecutive pairs and triplets of palindromic sites to assess if clustering is more prevalent than expected under random distribution.

The simulated random distribution in Question 1 will be used to generate spacings, pair sums, and triplet sums for comparison. Histograms and box plots will be used to display the distributions of spacings, pair sums, and triplet sums for comparison to their theoretical distributions under random scatter. A Kolmogorov-Smirnov (KS) test is used to test if the observed distributions significantly deviate from the expected random distributions.

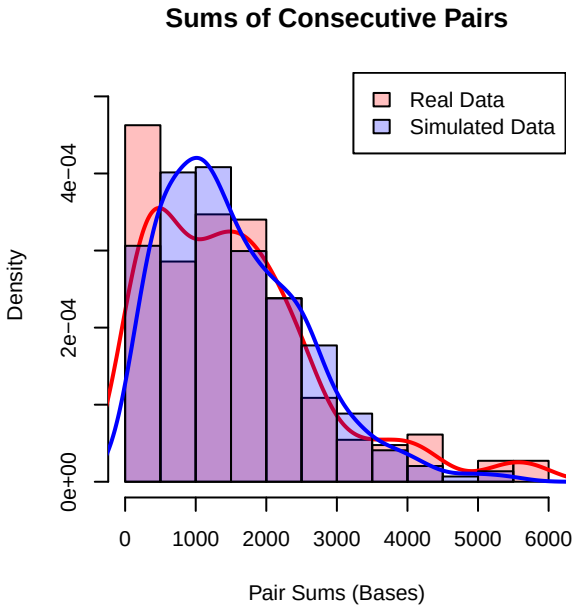
2.2.2 Analysis

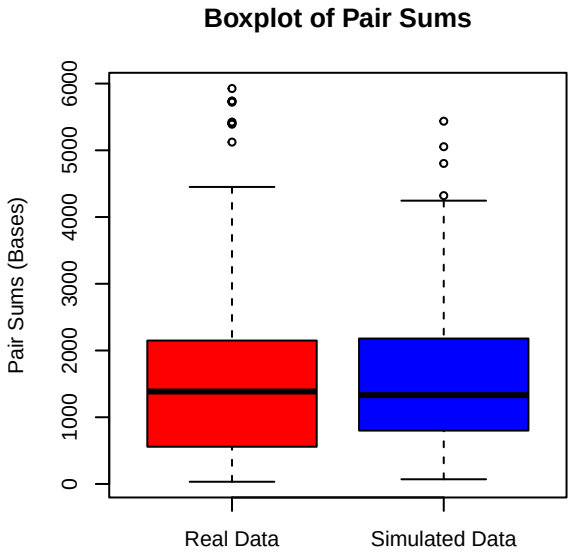
The graphs for the spacing between consecutive palindromes are as follows:



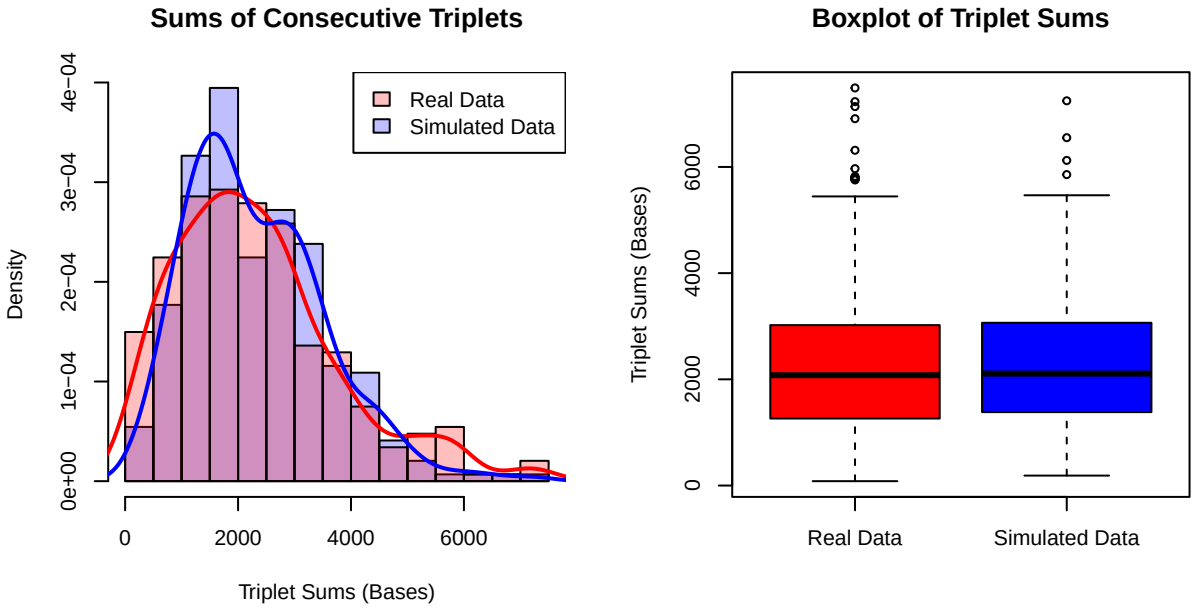


The distributions of the sums of palindrome pairs are as follows:



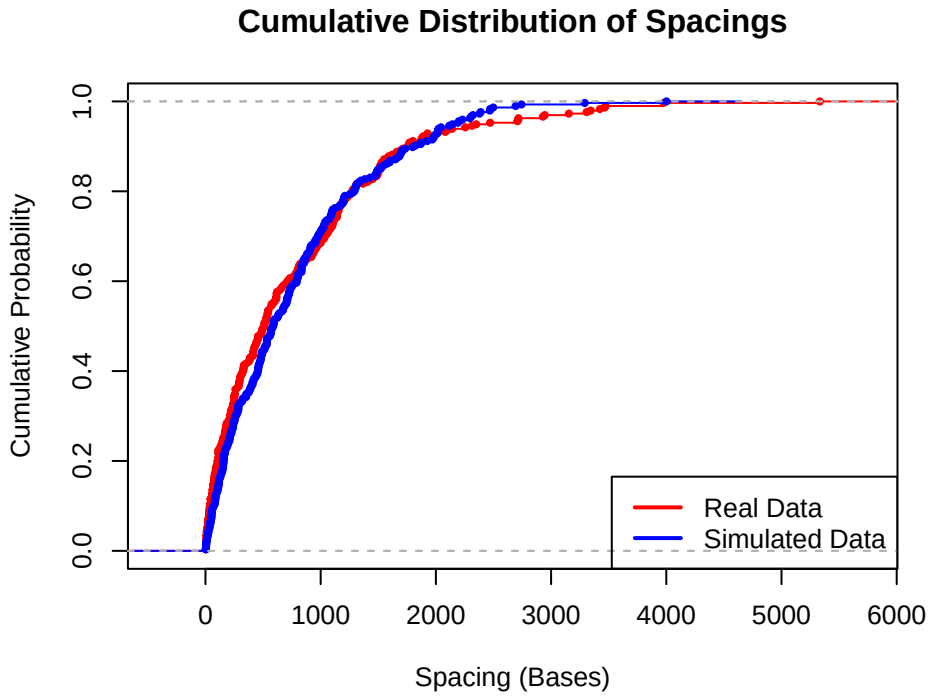


The distributions of the sums of palindrome triplets are as follows:



The Kolmogorov-Smirnov test for spacing yielded a p-value of 0.283, indicating that the observed distribution is consistent with the expected exponential distribution for random scatter. The KS test for pair sums yielded a p-value of 0.09365, the test for triplet sums returned a p-value of 0.1675. These p-values indicate that the observed distributions of pair and triplet sums are consistent with the uniform distribution expected under random scattering.

The Cumulative distribution plot for spacings can be seen below:



2.2.3 Conclusion

The distribution of spacings between palindromic sequences in the actual data shows similar variance to the simulated random data, suggesting that palindromes may regularly spaced, with no significant regions of clustering. The histograms of sums of consecutive pairs and triplets of palindromic locations do not indicate specific regions of higher concentration. The box plots similarly have close medians and quartiles.

The K-S test on spacings indicates whether the distribution of observed spacings significantly deviates from a uniform scatter. A low p-value would imply non-random clustering, while a high p-value would suggest the distribution could be random. All the tests resulted in a p-value exceeding 0.05, suggesting that the observed distributions align with random scatter expectations; we can conclude that the palindromic sequences are randomly scattered along the sequence.

The spacings of palindromes in the data seem to have no significant deviations from what is expected of a uniform random scatter, suggesting a lack of significant clusters or unusual spacings in palindrome locations.

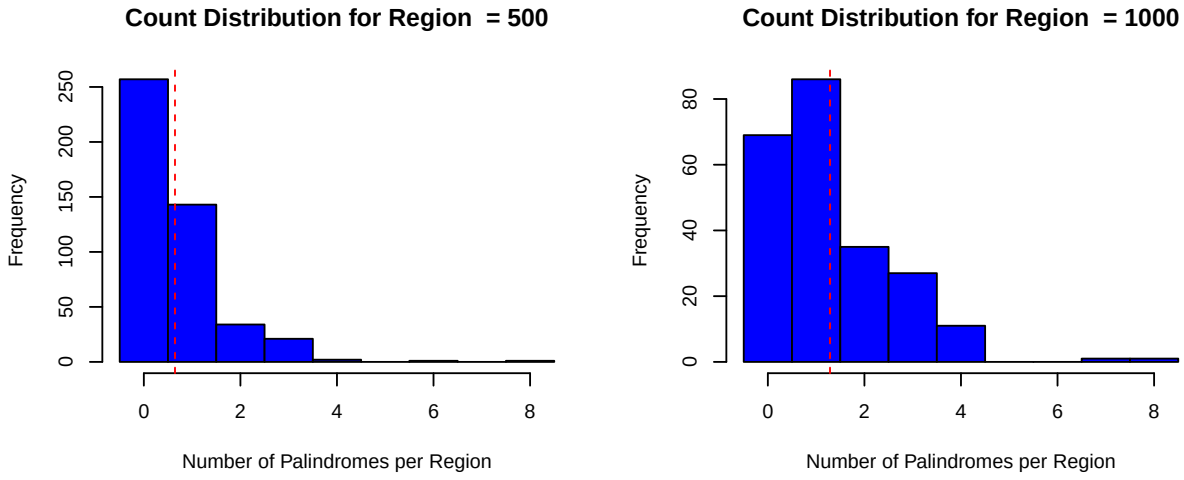
2.3 Palindrome Counts

2.3.1 Methods

To examine palindrome counts in different regions of DNA, we split the DNA sequence into equal-length, non-overlapping regions, counted the palindromic sequences in each, and compared these counts to those expected from a uniform random scatter. To visualize the results, we graphed histograms to visualize the distribution of palindrome counts across regions. We then used Chi-Square Goodness of Fit Tests to compare observed palindrome counts to expected counts, testing the hypothesis of a uniform random distribution of palindromic sequences across regions.

2.3.2 Analysis

The histograms and Chi-Square Test results are shown below.



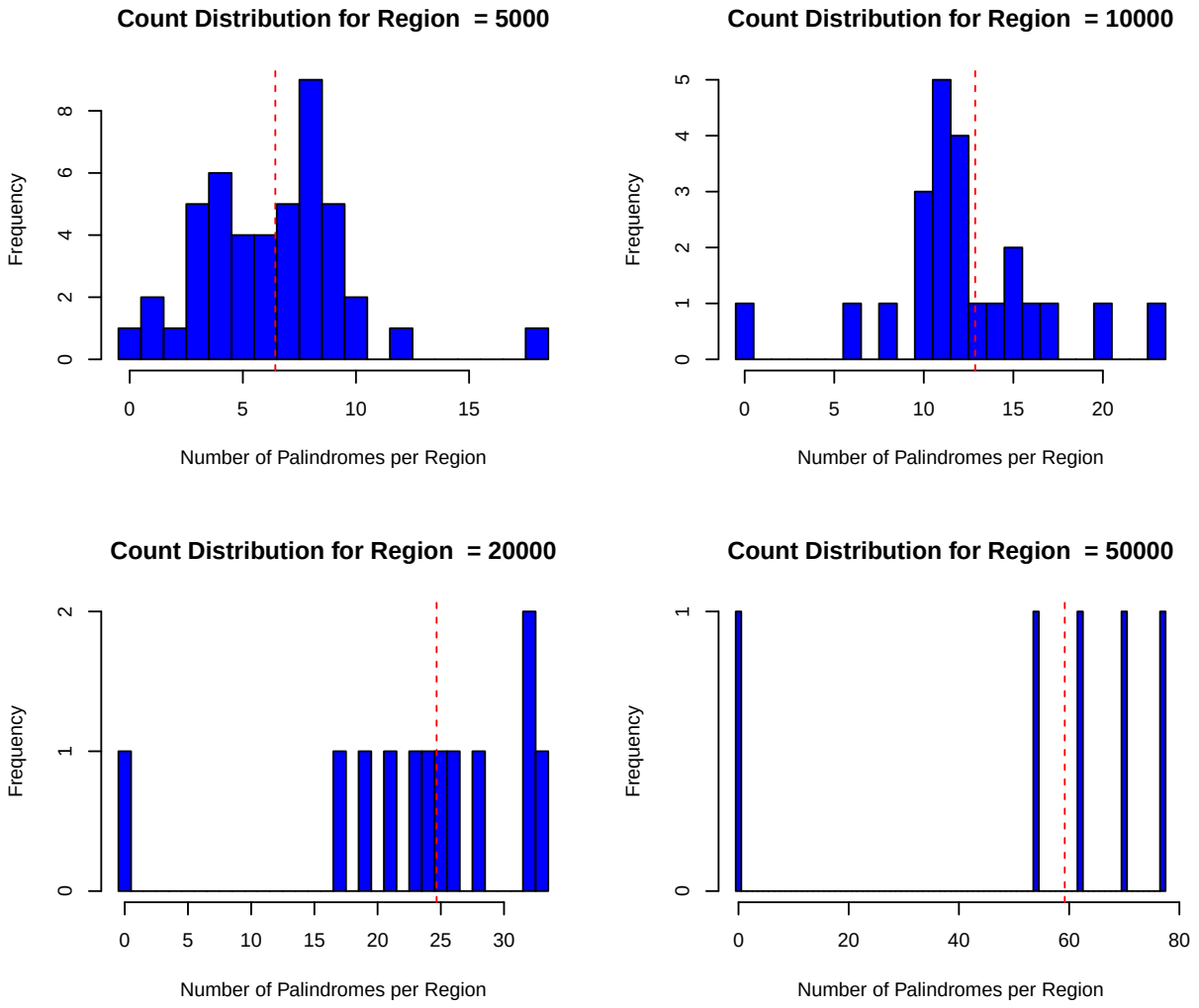


Table 2: Chi-Square Test Results

	Region.Length	Chi.Square.Statistic	p.value	Interpretation
X-squared	500	634.40541	0.0004998	Significant
X-squared1	1000	292.98649	0.0049975	Not Significant
X-squared2	5000	74.70242	0.0044978	Not Significant
X-squared3	10000	37.89286	0.0204898	Not Significant
X-squared4	20000	37.91429	0.0004998	Significant
X-squared5	50000	71.39163	0.0004998	Significant

2.3.3 Conclusion

The histograms of palindrome counts revealed that the smaller the region length, the more right skewed the count histogram was, which indicates that shorter regions have a higher frequency of intervals with low or zero palindrome counts, while a few regions have unusually high counts. This is consistent with the idea that smaller regions are more susceptible to variability in palindrome counts. Using a cut-off value of 0.001, the chi-square tests yield significant p-values for very small and very high region lengths, but insignificant

p-values for middle region lengths. The significant result at region length 500 suggests that palindromic sequences are clustered in certain small regions rather than being uniformly scattered across the DNA. The non-significant results at the intermediate scales (region lengths 1,000, 5,000, and 10,000) imply that, when viewed at these sizes, the distribution of palindromic sequences appears more uniform. At these lengths, local clusters of palindromic sequences are likely averaged out, resulting in counts that are more consistent with a uniform random distribution. The significant results at region lengths 20,000 and 50,000 indicate that there are large-scale patterns in the distribution of palindromic sequences across broad DNA segments.

2.4 The Biggest Cluster

2.4.1 Methods

We now want to investigate whether the interval with the highest number of palindromic sequences in a DNA sequence could suggest a potential origin of replication. To do this, we determined if a high-count palindrome cluster deviates significantly from a random scatter and whether such clusters align with features characteristic of origins of replication. Based on prior analysis, region lengths were selected to avoid both overly large and overly small intervals. Small intervals risk splitting clusters across adjacent regions, while large intervals may obscure tight clusters. For each interval, the number of palindromic sequences was counted, yielding a distribution of palindrome counts across intervals and the interval with the highest count of palindromic sequences was identified. Finally, Chi-square goodness of fit tests were applied to determine if the palindrome count in the highest-count interval deviates significantly from expected counts under a random distribution. Histograms were generated to visualize palindrome counts across intervals.

2.4.2 Analysis

The results of the Chi-Square Tests across different region lengths is shown below.

Table 3: Chi-Square Test Results

Region.Length	Max.Count	Max.Interval	Expected.Count	Chi.Square.p.value	Interpretation
500	8	186	0.6462882	0.0004998	Significant
1000	8	93	1.2925764	0.0024988	Not Significant
5000	18	19	6.5777778	0.0184908	Not Significant
10000	23	10	13.4545455	0.2858571	Not Significant
20000	33	5	26.9090909	0.3283358	Not Significant
50000	77	2	74.0000000	0.2258871	Not Significant

2.4.3 Conclusion

As shown in the table above, only the smallest region length (500) showed a significant p-value, while the larger region lengths (1000, 5000, 10000, 20000, and 50000) did not. This suggests that clustering of palindromic sequences exists on a small scale but not over broader intervals. The presence of a high number of palindromic sequences within an interval suggests non-random clustering, which may signal biological importance. Since the statistical analysis shows that the count in this interval is significantly higher than expected by chance, this could imply functional relevance, such as a replication origin.

2.5 Recommendations for Biologists

Our statistical analysis can provide insight on how to start experimentally searching for the origin of replication. The results showed a significant clustering of palindromic sequences in small regions (length 500), while

larger intervals showed a more uniform distribution. These findings suggest that the origin of replication may lie in one of these dense clusters of palindromic sequences within shorter segments of DNA. Given the statistically significant clustering at region length 500, biologists could begin the experimental search within these smaller regions.

For a more detailed explanation and further discussion on these findings, refer to the Discussion.

3. Advanced Analysis

3.1 Methods

The goal of this analysis is to evaluate the robustness of statistical assumptions used to analyze palindrome distribution across the CMV DNA sequence. Specifically, we examine how results from statistical tests—such as the Chi-square and Poisson tests—vary across different interval lengths. By assessing if observed palindrome counts align with expected values for each interval size, we aim to identify the most appropriate interval length for studying palindrome clustering and to determine if a uniform distribution model adequately describes palindrome distribution.

We will calculate palindrome counts within a range of interval lengths (100, 500, 1000, 5000, 10000). For each interval length, we apply Chi-square and Poisson tests to assess if observed palindrome counts match the expectations under a uniform random scatter assumption. A significant variation in test outcomes may suggest that the uniform scatter assumption is sensitive to interval length.

We begin with dividing the 229,354 bases into non-overlapping intervals of different lengths. For each chosen interval length, we calculate the observed number of palindromes within each interval across the DNA sequence. Based on a uniform distribution, the expected count per interval is calculated as:

$$E_i = \frac{\text{Total Palindromes}}{\text{Total Intervals}}$$

- **Chi-square Goodness-of-Fit Test**: For each interval size, we use the Chi-square test to compare observed vs. expected counts:

  $$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

  where O_i and E_i are the observed and expected counts for each interval i.

- **Poisson Distribution Test**: If the palindromic sites are randomly scattered across the DNA sequence, the counts in intervals should follow a Poisson distribution because a Poisson process models random events occurring at a constant rate over space. Assuming that palindrome locations follow a Poisson process, we calculate the probability of observed counts within each interval and test if they align with a Poisson distribution using a goodness-of-fit approach. A Poisson distribution is defined for count data where events occur independently, with a constant average rate λ over an interval. The probability of observing k palindromic sites in an interval is:

  $$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

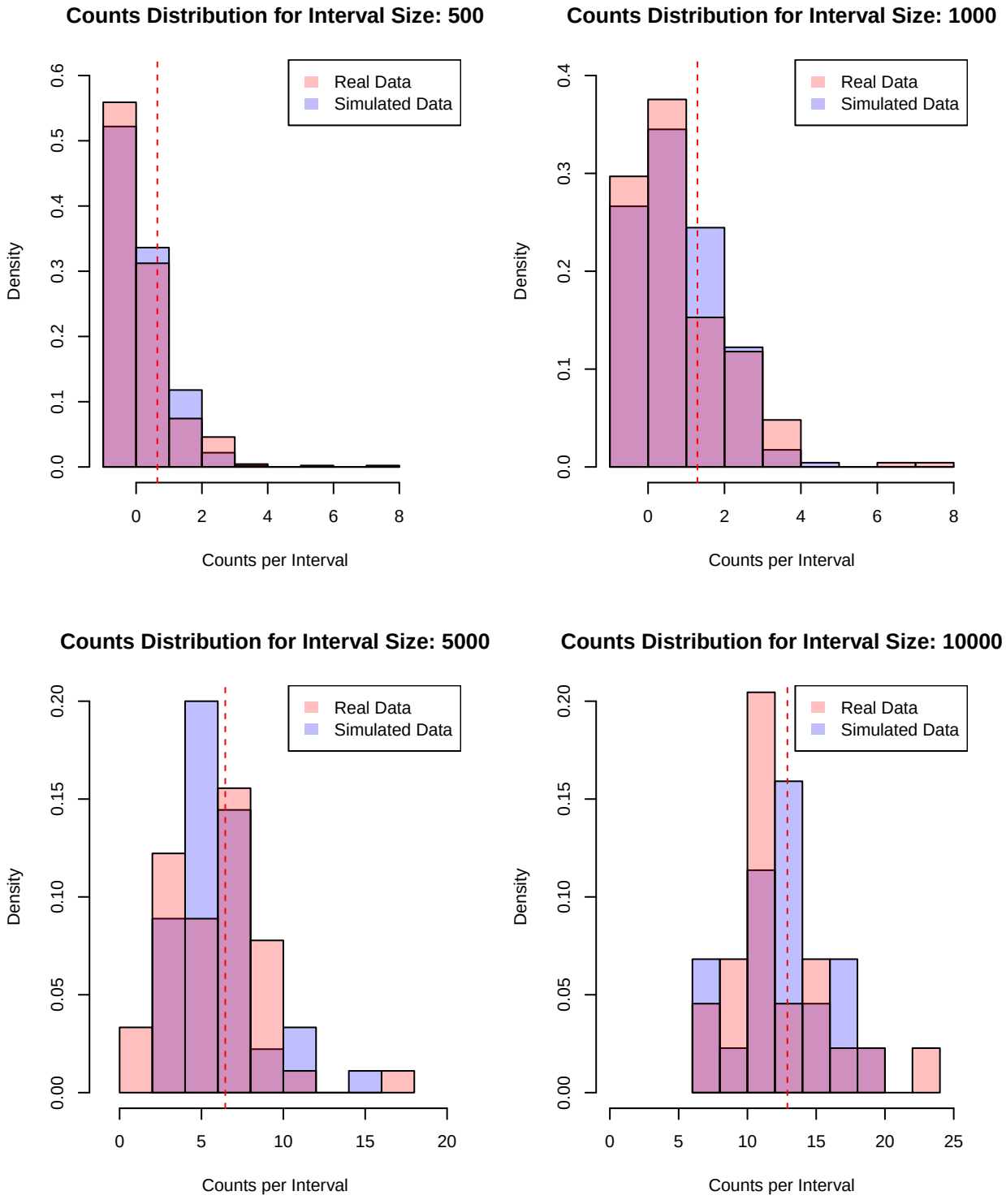
  For our analysis, λ is calculated as the **mean number of palindromes per interval**: $\lambda = \frac{\text{total palindromes}}{\text{number of intervals}}$. This assumption is justified because each interval is of equal length, so we expect a constant rate if the sites are randomly scattered.

If **p-value** $> \alpha$ (e.g., $\alpha = 0.05$), we *fail to reject* the null hypothesis, suggesting that the observed palindrome counts align with a Poisson distribution. This alignment supports the assumption of random scattering in the data.

If **p-value** $\leq \alpha$, we *reject* the null hypothesis, indicating that palindrome counts likely deviate from a Poisson distribution, suggesting clustering or non-random scatter.

3.2 Analysis

Histograms of palindrome counts for each interval length are as follows:



The summary provided displays results from simulations at four different interval sizes (500, 1000, 5000, and 10,000 base pairs) to test the assumption of a Poisson distribution for palindrome counts across these

intervals:

Table 4: Count Summary for Various Interval Sizes

Interval	Real.Mean	Sim.Mean	Real.Var	Sim.Var	Chi.Square	Expected.Count	Poisson
500	0.646	0.646	0.894	0.623	0.0005	0.6453	1
1000	1.293	1.293	1.646	1.146	0.0055	1.2906	1
5000	6.423	6.489	9.749	6.028	0.0175	6.4529	1
10000	12.727	12.955	14.589	10.807	0.2939	12.9058	1

`mean_real_counts` and `mean_sim_counts` show the average number of palindromic sites per interval for real data and randomly scattered data, respectively. These values should be similar if the observed data matches the randomness of the simulation.

`var_real_counts` and `var_sim_counts` reflect the variance in the count of palindromic sites per interval for both real and simulated data. In a true Poisson distribution, we expect the mean and variance to be roughly equal, especially in larger intervals where random scatter is more evident.

`chi_p_value` assesses how well the observed distribution of palindromic counts aligns with the simulated distribution under the Chi-square test. A low p-value (e.g., <0.05) suggests a statistically significant difference, indicating that the real data distribution may deviate from the expected Poisson scatter.

`poisson_p_value` shows results from a goodness-of-fit test specifically for the Poisson distribution.

3.3 Conclusion

The results indicated that, as interval size increased, the fit between real and simulated data generally improved, as evidenced by increasing Chi-square p-values. For smaller intervals (500 and 1000 base pairs), the Chi-square p-values were below the 0.05 significance threshold, suggesting a significant deviation from randomness; this implies clustering at finer scales or patterns not accounted for by a uniform model, indicating that palindromic sites are more densely distributed in certain areas than would be expected under a random scatter model. As interval sizes increased to 5000 and 10,000 base pairs, the Chi-square p-values become larger (e.g., 0.29 at 10,000 base pairs), indicating that the real data more closely aligns with the simulated data, as would be expected under random scatter over larger scales.

The Poisson p-values remained high across all interval sizes, suggesting that, on average, palindrome counts align with a Poisson distribution. This supports the random scattering assumption when considering large-scale patterns across the DNA sequence, even if there are minor local deviations at smaller scales.

This implies that palindrome distribution in the DNA sequence may show clustering at smaller scales, yet appears more random when observed over larger intervals. Smaller intervals provide finer granularity, while larger intervals aggregate more data, which might smooth out fluctuations. Although the Poisson distribution offers a useful general model for palindromic distribution in the CMV DNA sequence, clustering patterns become apparent at finer scales, questioning the assumption of a purely uniform scatter. This may inform researchers on the appropriate interval size for studying palindrome distribution, as smaller intervals could reveal structural patterns, while larger intervals support the assumption of uniform random scatter.

4. Discussion

The primary goal of this analysis was to investigate whether the distribution of palindromes in the CMV DNA sequence deviates from a uniform scatter or if any clusters are due to random chance, comparing the actual palindrome distribution to simulated random scatterings. We examined the spacing and counts of palindromic sequences across DNA intervals, looking for potential clusters that might indicate functional

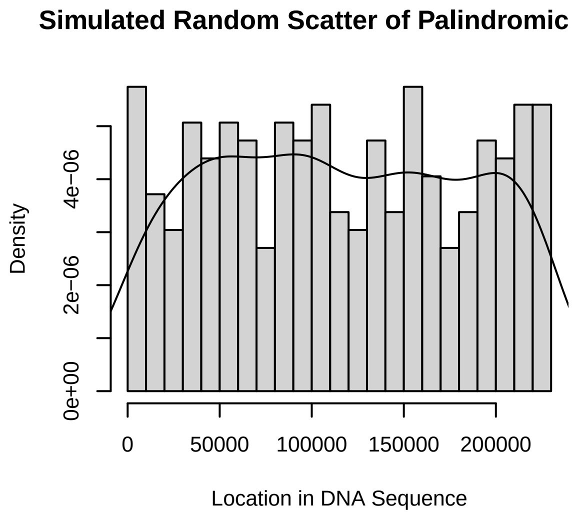
regions like replication origins. To do this, we simulated random palindrome locations, compared them to the real data, and used Chi-square and Poisson tests to assess how well the observed distribution fit with the expected random pattern across different interval sizes.

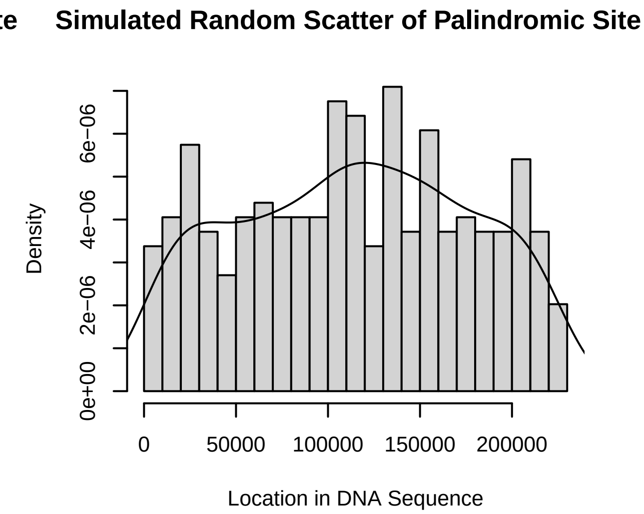
Our findings showed that, overall, the distribution of palindromic sequences in the CMV genome closely resembled a random scatter, with no large-scale clusters that might suggest specific functional regions. The Kolmogorov-Smirnov tests, comparing real data to random simulations, produced high p-values, indicating no significant deviations from randomness. However, when we looked at smaller intervals (like 500 base pairs), we found some significant clustering of palindromes. This was confirmed by Chi-square tests, which showed a notable difference in palindrome counts compared to the random model. These results suggest that, while the overall genome-wide distribution of palindromes appears random, there are localized clusters that could have functional significance, possibly related to replication origins or other regulatory elements.

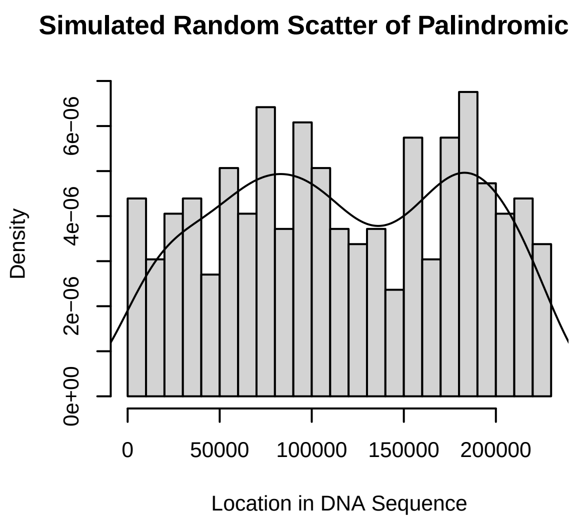
Additionally, our analysis highlighted the importance of choosing the right interval size for statistical tests. Smaller intervals (like 500 or 1000 bases) revealed significant deviations from uniformity, suggesting potential clusters of palindromes, whereas larger intervals (e.g., 5000 or 10,000 bases) seemed to follow a more random distribution. This suggests that smaller intervals can uncover finer details of biological patterns, while larger ones smooth out small-scale fluctuations and make the data look more random. The significant clustering of palindromes in the 500-base pair intervals could provide clues for locating the origin of replication in CMV, as replication origins often show specific sequence patterns. These findings could guide future research on identifying replication origins or other important biological sites.

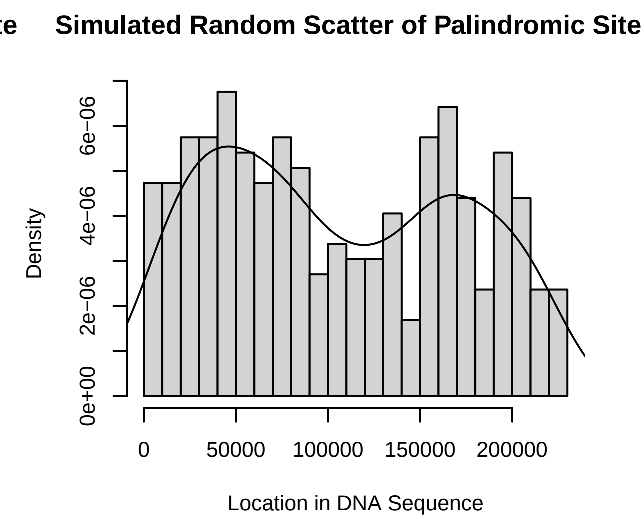
There are a few limitations to this analysis. Although we used a range of interval sizes, the choice of breaks was still somewhat arbitrary and might not reflect biologically meaningful regions within the sequence. Our analysis also assumes that palindromic sites are randomly distributed if they don't serve a functional purpose. However, biological systems are often more organized than random distributions suggest, and factors we didn't measure—like specific sequence motifs or binding sites—could impact where palindromes occur. Including more variables in future analyses could clarify these relationships and lead to a more accurate model for understanding palindromic site distributions.

5. Appendix


Simulated Random Scatter of Palindromic Site
Density
0e+00
2e−06
4e−06
0
50000
100000
150000
200000
Location in DNA Sequence


Simulated Random Scatter of Palindromic Site
Density
0e+00
2e−06
4e−06
6e−06
0
50000
100000
150000
200000
Location in DNA Sequence


Simulated Random Scatter of Palindromic Site
Density
0e+00
2e−06
4e−06
6e−06
0
50000
100000
150000
200000
Location in DNA Sequence


Simulated Random Scatter of Palindromic Site
Density
0e+00
2e−06
4e−06
6e−06
0
50000
100000
150000
200000
Location in DNA Sequence