Finding the Origin of Replication in DNA

Author 1 and Author 2

2024-11-09

Header

Author Contributions

Author 1: Contributed to questions 2 and 4, along with doing the formatting for the graphs, tables, and the pdf.

Author 2: Contributed to questions 1, 3, and 5, along with doing the advanced analysis.

Use of GPT

ChatGPT was used as a substitute for documentation for R. Since we were unfamiliar with R, we asked ChatGPT how to use R in certain methods in order to find and filter out conditions in the dataset. We additionally used GPT to analyze reasoning and to confirm what we thought was correct about the dataset, as well as to identify extra questions that could be answered for our advanced analysis.

Introduction

The data used in this analysis is from a DNA sequence of CMV which was published in 1990. The data specifically is of one column, which consists for 296 palindromic sequences, each of which were at least 10 pairs long. Our objective in this analysis is to analyze the structure of the data, and assess how the distribution of the DNA palindromes deviates from a uniform scatter across the DNA sequence, if it even does. Essentially, we are testing if the clusters of palindromes in the DNA sequence are due to chance, or if there's a set pattern within the DNA.

Main Research Questions

1. Simulate 296 palindrome sites chosen at random along a DNA sequence of 229534 bases using a pseudo random number generator. Do this several times by making sets of simulated palindrome locations, performing a quantitative and qualitative comparison between the random scatters and real data.
2. Use graphical methods to analyze the patterns in the following. Additionally, compare observed patterns to expected uniform random distirbutions to identify significant clusters or unusual spacing in palindrome locations.

    a. Spacing between consecutive palindromes
    b. Sums of palindrome pairs
    c. Sums of palindrome triplets

3. Use graphical methods and more formal statistical tests to examine the counts of palindromes in various regions of DNA. Split the DNA into nonoverlaping regions of equal length to compare the number of palindromes in an interval to the number that you would expect from uniform random scatter.
4. Does any interval with the greatest number of palindromes indicate a potential origin of replication? Validate your results.
5. How would you advise a biologist who is about to start experimentally searching for the origin of replication?

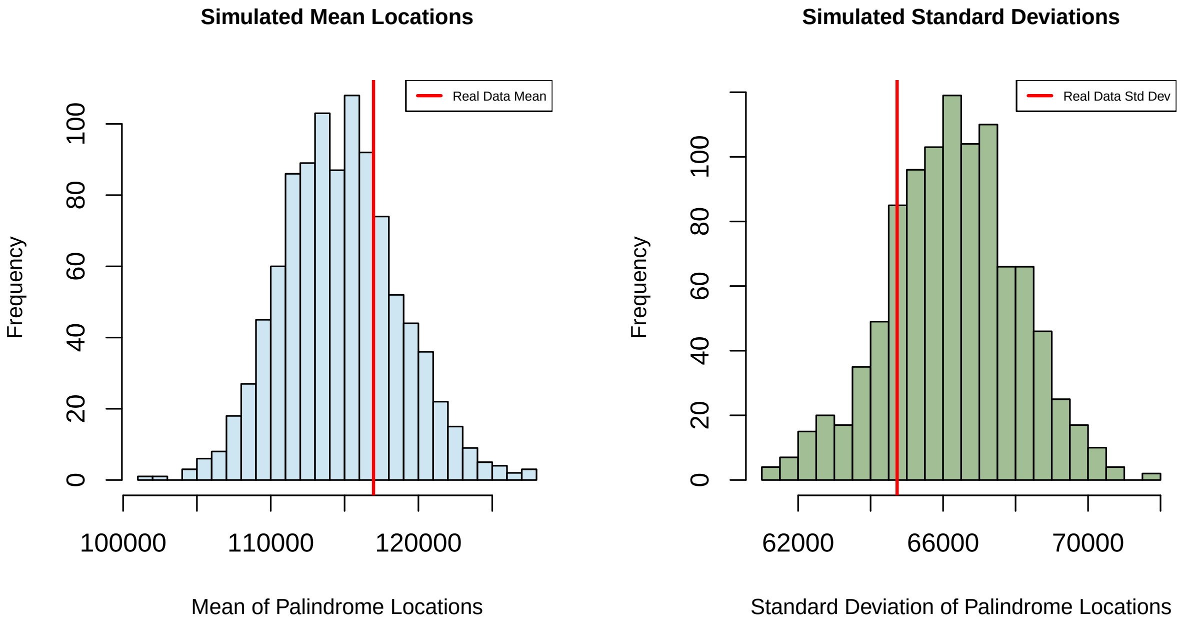
Question 1: Compare Palindrome Locations to Simulated Uniform Distributions

Methods

Our dataset contains the locations of palindromic sequences in a series of 229,354 base pairs. In order to compare our data to uniform simulations, we will simulative over 1000 uniform distributions of palindromic sites along these 229,354 base pairs and compare it to the real data, specifically looking through key stats like mean, median, variance, and standard deviation, which are shown below.

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

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



Analysis

As seen above, our uniform spacing simulations show that there is a relatively small deviation between our observed mean and simulated means. While our simulations had an overall mean location of ~115000, the observed mean location was at ~117000. However, the observed standard deviation was slightly more unusual in the context of the simulated uniform spacings. The simulations tended to have a standard deviation around ~66,300 while our observed standard deviation was ~64,700. A smaller standard deviation signifies that the palindromes are less spread out at points and might be more clustered than most uniform distributions of palindromes.

A hypothesis test was conducted to test whether the observed mean was significantly different from the simulated means of uniform palindrome spreads. The p-value of **0.561** shows that the mean value is not different enough from the simulated means. A similar hypothesis test was conducted to see if the observed standard deviation was significantly different from the simulated standard deviations. The p-value for standard deviations was **0.375**, which also is not significant enough to reject the null hypothesis that the observed standard deviation in the real DNA is different from uniformly simulated DNA.

Conclusions

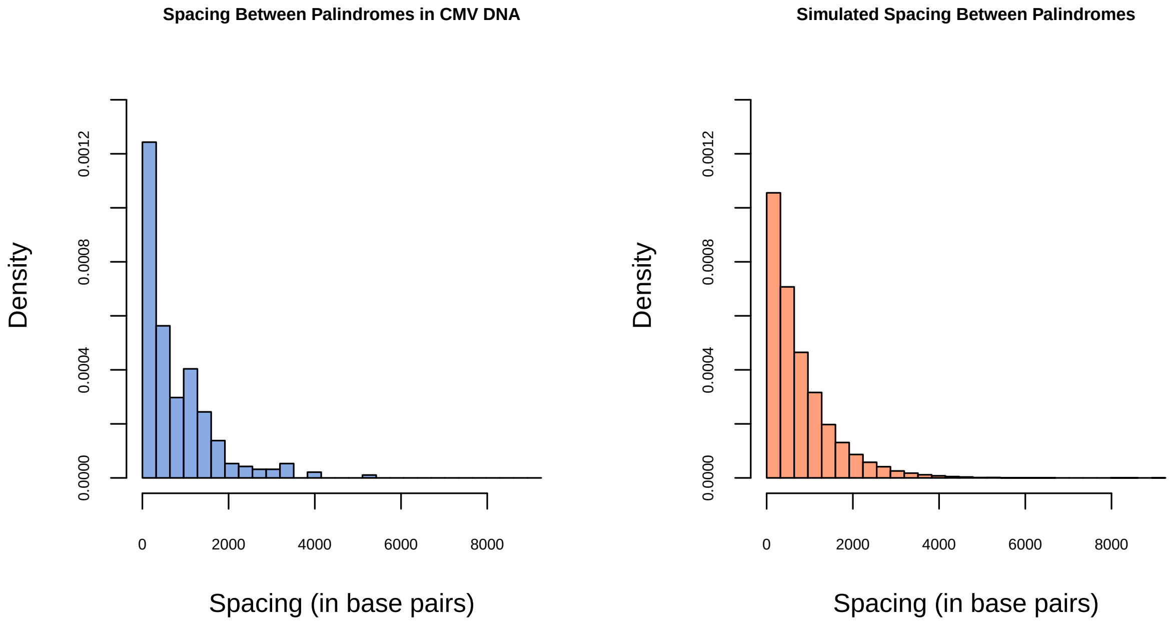
After visually inspecting the simulate data as well as conducting hypothesis tests, we are not confident that statistics in our DNA data like mean and standard deviation are significantly different from the simulated uniform data. However, the smaller standard deviation of our palindrome locations motivates further analysis because it implies that there may be clusters that have significance in the context of virus replication.

Question 2: Graphically Analyze Patterns in the Palindrome Data

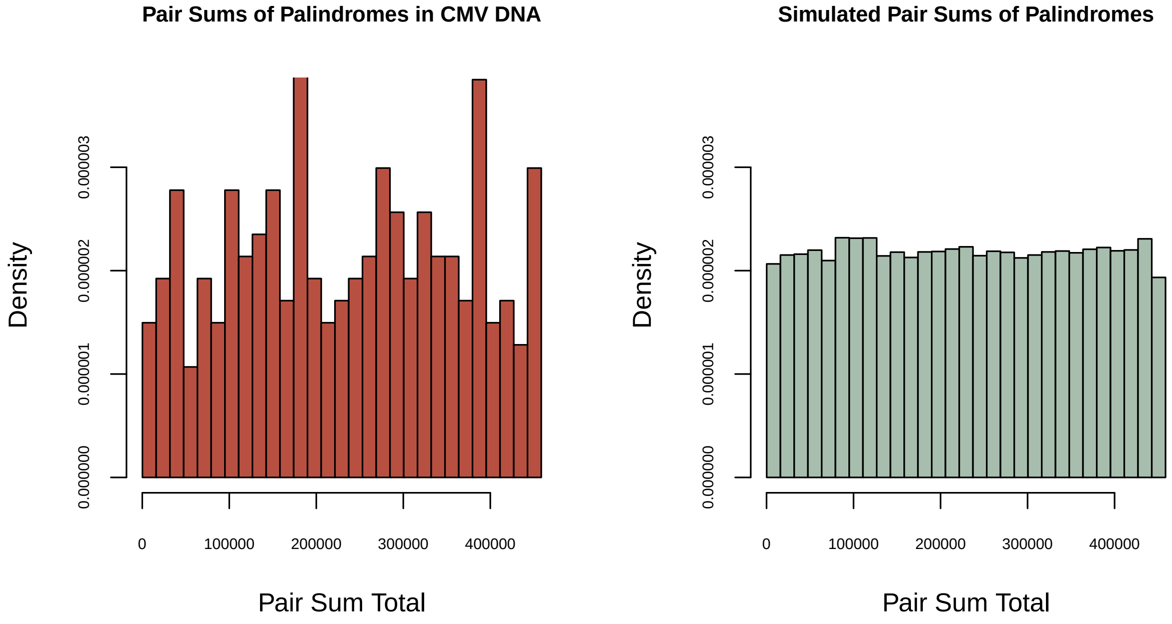
Methods

In order to further analyze the distribution of the palindromes, let's see three different variations of the data and how they deviate from a uniform scatter across the DNA sequence. To do this, we will compare the observed patterns below to expected uniform random distributions to identify any significant clsuters of unusual spacing in these palindrome locations. One important thing to notice is that we used 100 simulations of the uniform distribution to make our graphs on the right.

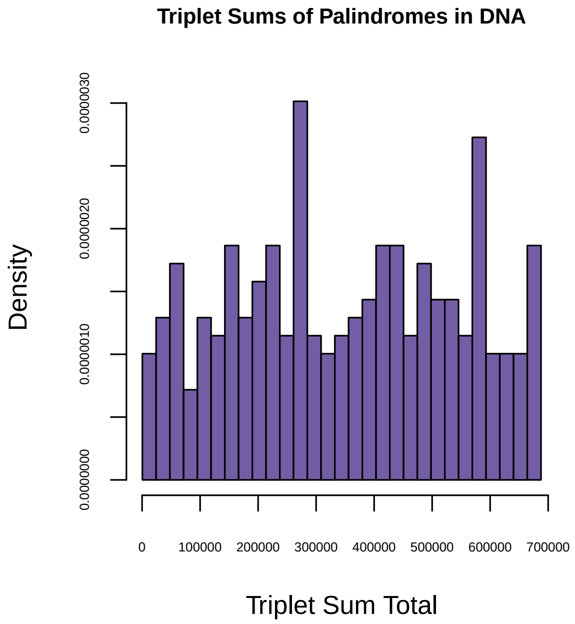
a) Spacing between consecutive palindromes

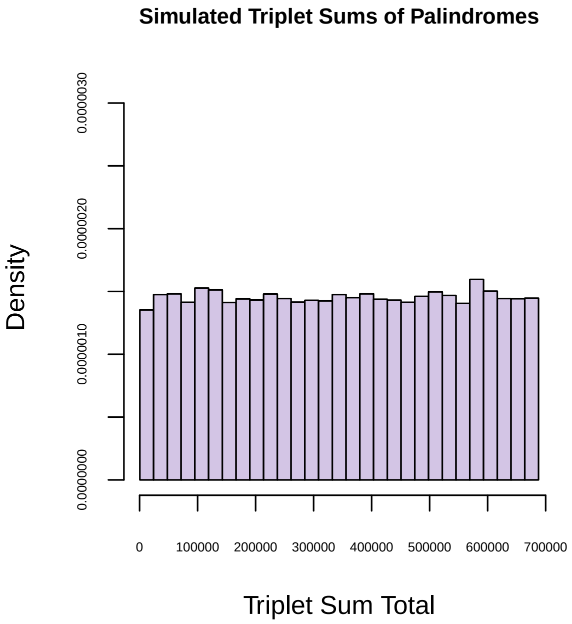


b: Sums of Palindrome Pairs



c) Sums of Palindrome Triplets





Analysis

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

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

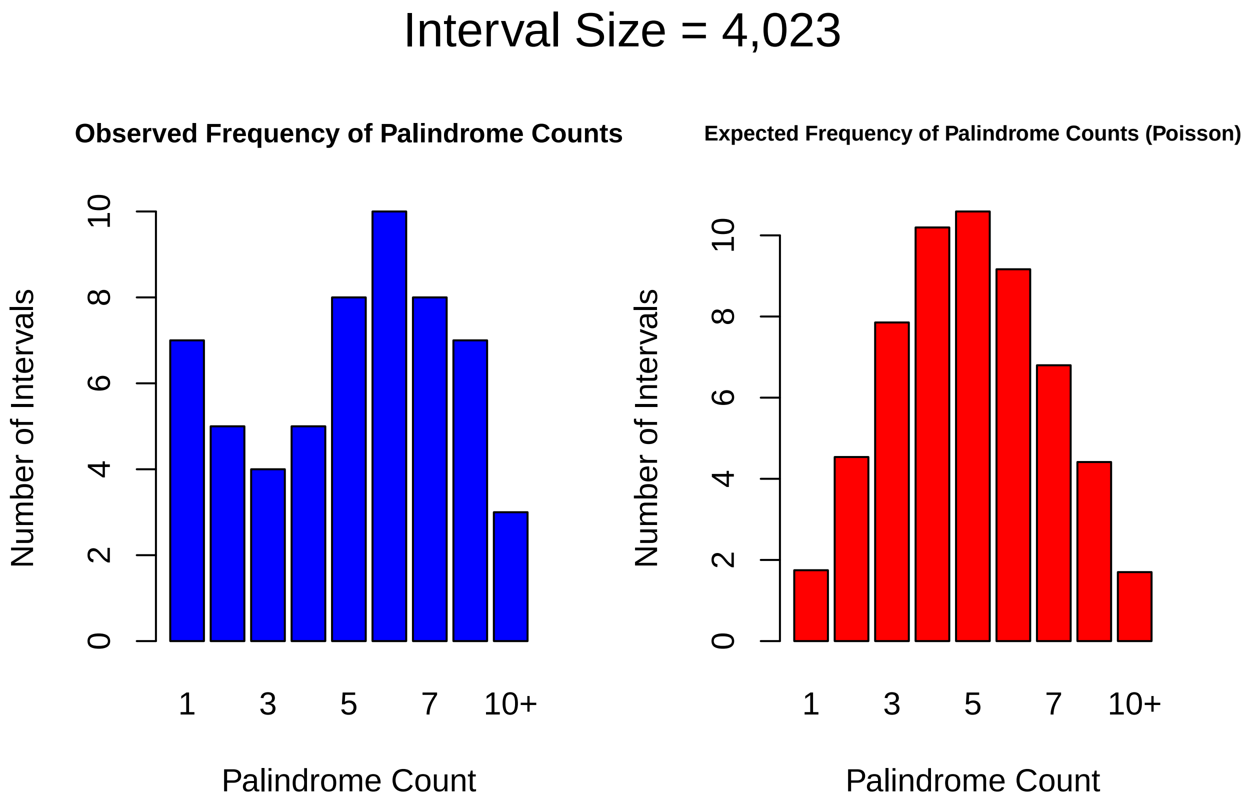
Conclusions

While spacing between consecutive palindromes follows a similar distribution to our simulated consecutive palindromes, the pair sums and triple sums do not, with unusual spacing between their distributions compared to their respective simulated uniform distributions. Due to these, along with evident clusters and unusual spacings between many parts of the data, it seems likely that the CMV DNA does not follow a uniform distribution, even though it might seem like it.

Question 3: Examine the Counts of Palindromes in Various Regions of the DNA

Methods

First, intervals of size 4023 are used to partition the DNA sequence. We then count the number of intervals that contain each palindrome count and compare it to the expected distribution of intervals. The expected distribution is obtained by using the Poisson distribution to calculate the probability of getting each palindrome count and then scaling it so that the sum of counts matches that of the observed distribution.



After completing a graphical comparison, a chi-squared goodness-of-fit test is conducted to see if the observed distribution is different from the expected.

```
## Warning in chisq.test(x = count_frequency_df$Intervals_Observed, p =
## count_frequency_df$Intervals_Expected/sum(count_frequency_df$Intervals_Expected),
## : Chi-squared approximation may be incorrect
```

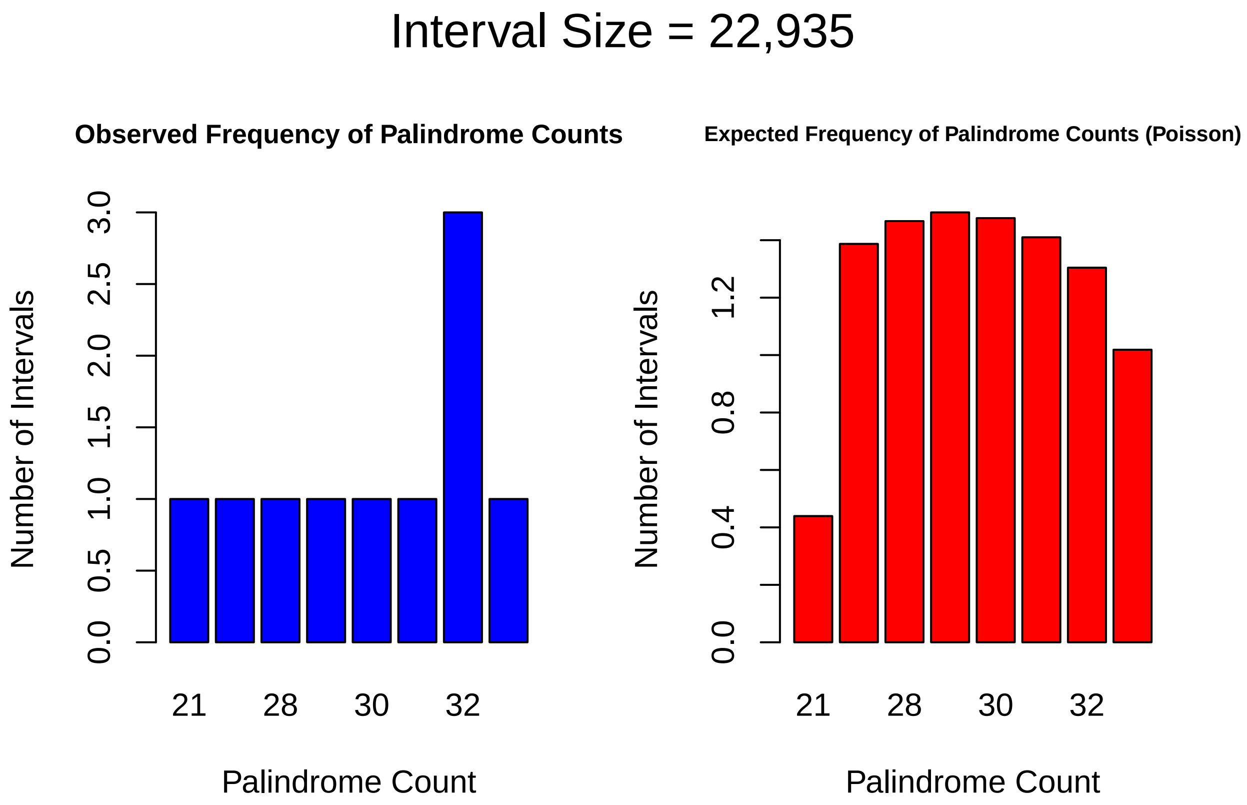
```
##
```

```
##  Chi-squared test for given probabilities
##
## data:  count_frequency_df$Intervals_Observed
## X-squared = 23.806, df = 8, p-value = 0.00247
```

The above warning appears because there are some categories that only contain a small number of intervals, which could affect the quality of the chi-squared test's results. Further tests are conducted with different interval sizes to see if the results change.

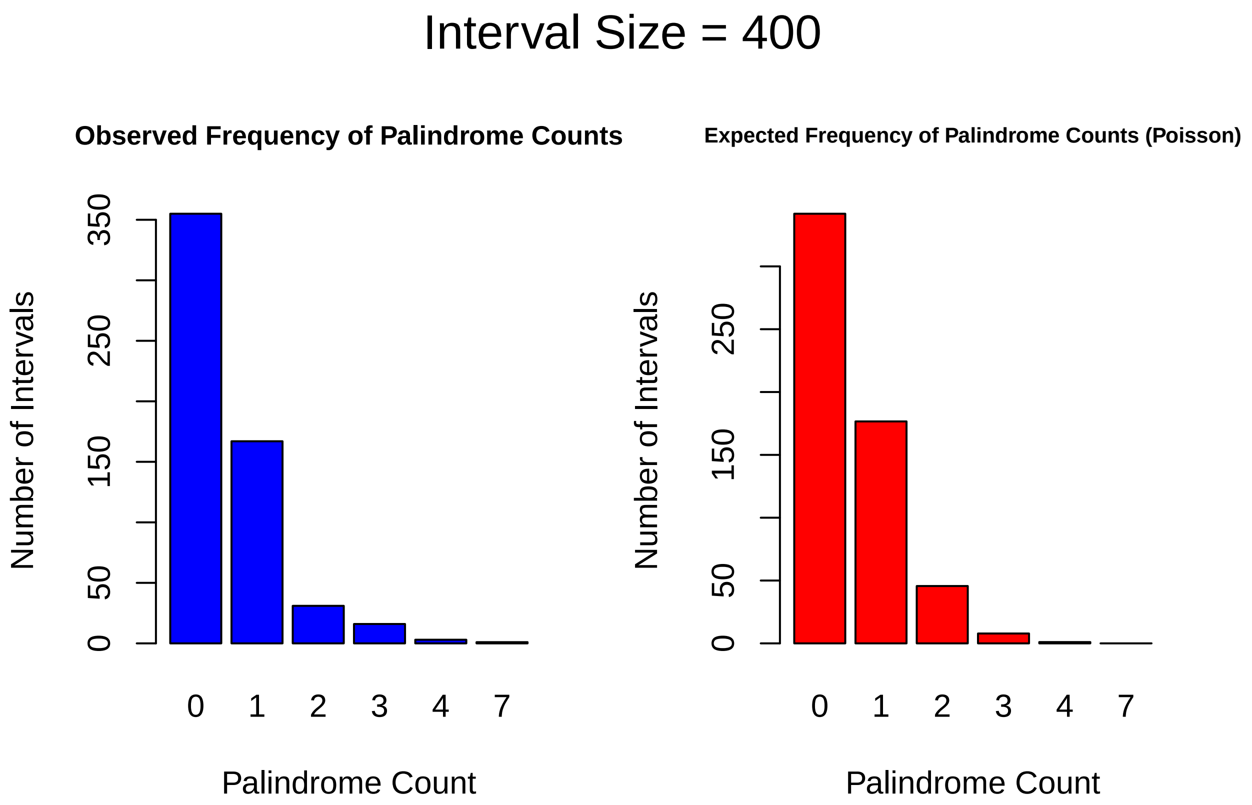
Since the p-value is less than our significance level of 0.05, there is significant evidence to reject the null hypothesis, meaning the distribution of palindrome counts among intervals is significantly different from one generated from the Poisson process.

Next, we create different interval sizes to investigate the effects on detection of uniformity.


Interval Size = 22,935

Observed Frequency of Palindrome Counts

Expected Frequency of Palindrome Counts (Poisson)



Another chi-squared goodness-of-fit test is conducted, but this time, much smaller interval sizes are used to see if results change when we manipulate interval size.

```
## Warning in chisq.test(x = smaller_count_frequency_df$Intervals_Observed, :
## Chi-squared approximation may be incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  smaller_count_frequency_df$Intervals_Observed
## X-squared = 1517.7, df = 5, p-value < 0.00000000000000022
```

Analysis

Using interval sizes of 4023 bases reveals that the true distribution of palindrome counts differs from the expected (Poisson) distribution in many ways. Firstly, the true distribution has a much greater count of intervals with only 1 palindrome than expected. Secondly, the true distribution has far greater frequencies of intervals with 7 palindromes or more. The chi-squared goodness-of-fit test corroborates these observations because the p-value of **0.0025** means that there is a significant difference between our distribution and what would be expected from a Poisson data generating process. It's worth acknowledging that, because our expected distribution has some very small counts, the chi-squared test may not be the most accurate.

Increasing the interval size to 22,935 bases made the distribution less insightful because we have far fewer intervals, meaning less data. It doesn't make sense to conduct a chi-squared goodness-of-fit test with this data because the counts of intervals are so small.

Decreasing the interval size to 400 means we have far larger interval frequencies, but these frequencies have less meaning in context because the vast majority of intervals now contain 0 palindromes. Thus, given the context of palindromes in a DNA sequence, the chi-squared goodness-of-fit test's results are not as useful with this interval size.

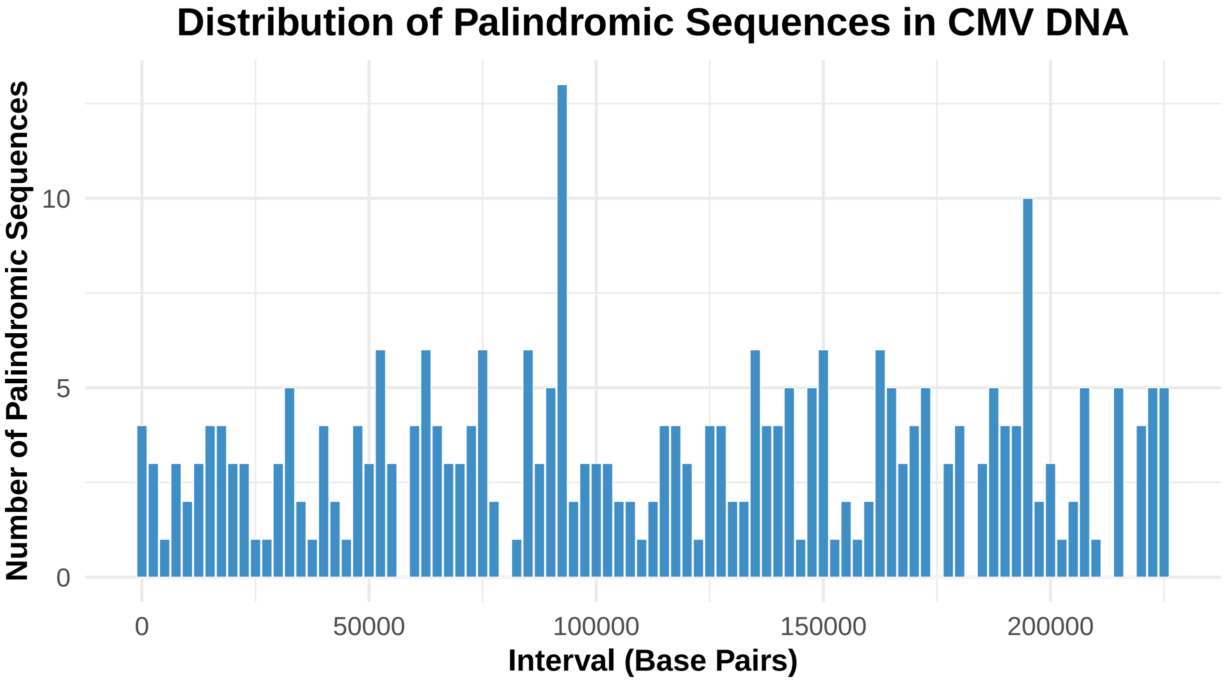
Conclusions

The distribution of palindrome counts in intervals of size 4023 is significantly different from what would be expected of a uniform random scatter. This means the scatter of palindromes in the DNA is significant and not just random. There are more intervals with high counts of palindromes (7+) than expected. This insight, combined with the earlier insight that there are possibly significant clusters of palindromes, motivate further analysis of specific high-density clusters of palindromes.

Question 4: Find a Potential Origin of Replication

Methods

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



Additionally, let's conduct a Poisson test to test the level of significance, validating our results that the intervals of 92500 - 95000 or 195000 - 197500 are potential origins of replication.

```
## P-value for interval 92500-95000: 0.00003071359
```

```
## P-value for interval 195000-197500: 0.001744772
```

Analysis

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

Additionally, we used a hypothesis test to validate these two intevals at a value of p = 0.05. Our null hypothesis is that all of the intervals follow a Poisson distribution, while our alternative hypothesis is that it doesn't follow a Poisson distribution and these intervals are statistically significant. We can see that these two intervals are statistically significant, with p-values of 3.0713e-05 and 1.744e-03, both lower than 0.05.

Conclusions

It seems likely that the interval between 92500 - 95000 is the area where the origin of replication occurs. This is due to the extreme number of palindromes in this interval when compared to all other 2500 location-length intervals, potentially giving us an insight where the origin of replication is. Due to having so many more palindromes (~4x more than a normal interval), this is likely to be the origin of replication. Additionally, by testing with a Poisson distribution, we figured that this interval was significantly significant.

Question 5: Advice for a Biologist

The above analyses revealed that the distribution of palindrome sites throughout the CMV DNA has significant clusters and is different from a uniform scatter. This means I would recommend the biologist to first conduct experiments focused on the largest clusters, which are from locations 92500 - 95000 and 195000 - 197500. Additionally, I would recommend the biologist to study consecutive pairs of palindrome sites whose locations add up to around 385,000 because there is a surprisingly large frequency of pairs that match this description compared to what would be expected based on the number of palindrome sites with locations near 192,500. This can be seen by viewing peaks in the histogram of palindrome pair sums and comparing it to the histogram of individual counts within intervals.

There is numerous evidence presented in our analyses to suggest that the DNA sequence has significant clusters of palindromes for finding the origin of replication. Firstly, the standard deviation of palindrome locations in the CMV DNA is lower than those of data simulated from uniform scatters. This means that palindromes are closer together on average (more clustered) than most uniformly scattered DNA sequences. Additionally, the chi-squared goodness-of-fit test strongly suggested that the distribution of palindrome counts among intervals did not match that of a Poisson distribution.

Advanced Analysis

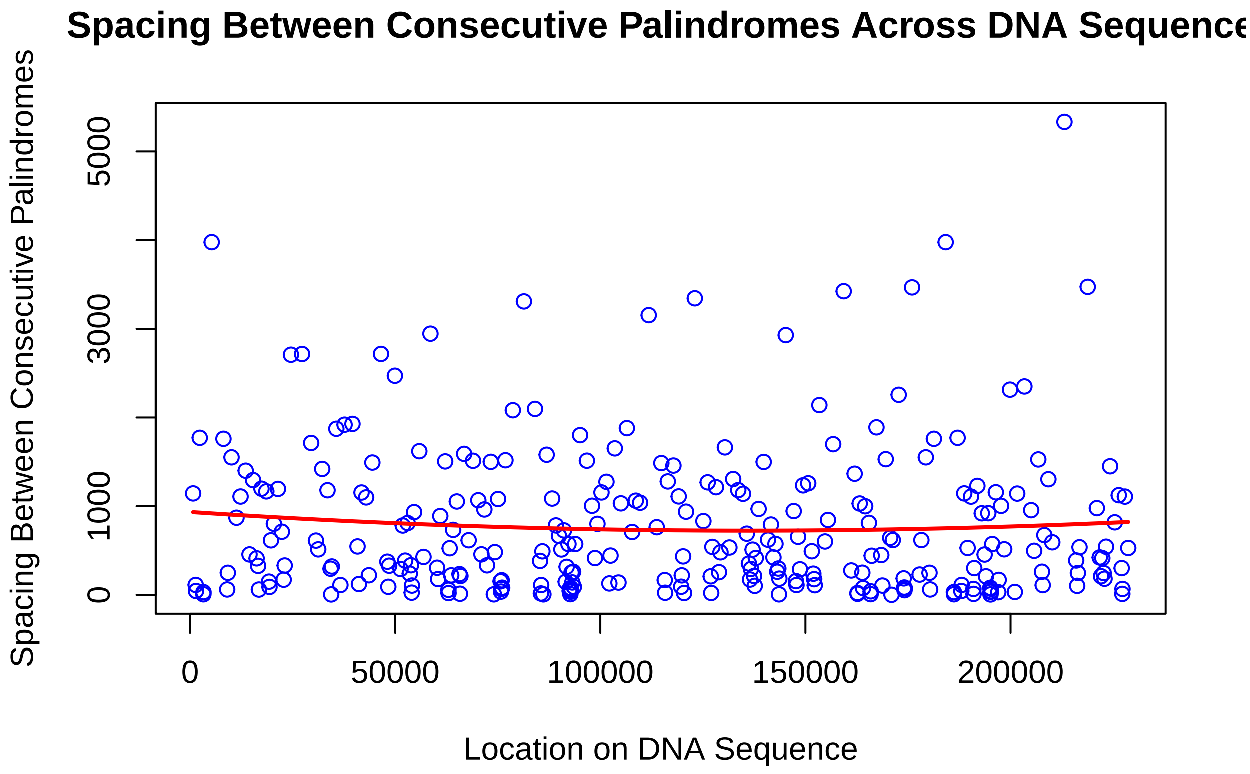
Find Correlation Between Location and Spacing Between Palindromes

Methods

In order to qualitatively search for a correlation between location in DNA sequence and spacing between consecutive palindromes, we first create a scatter plot with location on the x-axis and palindrome distance on the y-axis. Visual inspection led us to believe that a quadratic shape may be necessary for an effective regression, so we engineered a feature by squaring the location value. Finally, a quadratic function was fit and plotted.

##

```
## Call:
## lm(formula = Spacing ~ Location + Location_Squared, data = spacing_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -917.9 -606.6 -263.2  336.0 4541.0
##
## Coefficients:
##                         Estimate      Std. Error t value      Pr(>|t|)
## (Intercept)      934.82411709783 155.18782626870   6.024 0.00000000512 ***
## Location          -0.00309808071   0.00302590607  -1.024         0.307
## Location_Squared   0.00000001139   0.00000001260   0.904         0.367
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 834 on 292 degrees of freedom
## Multiple R-squared:  0.003998,   Adjusted R-squared:  -0.002823
## F-statistic: 0.5861 on 2 and 292 DF,  p-value: 0.5571
```



Analysis

The regression gave us the following coefficients: **1.14e-8** for Location_Squared and **-3.10e-3** for Location. The coefficient on Location_Squared is so negligible that the regression does not take on much of a quadratic form. The coefficient on Location is also very small. Additionally, both coefficients have p-values greater than 0.3, meaning they both don’t have very strong significance in terms of predicting palindrome spacing from location. Finally, the

$$R^2$$

value of **0.004** is very close to 0, showing that there is very low correlation between our features and palindrome spacing.

Conclusions

There is almost no correlation between location in the DNA sequence and palindrome spacing that could be detected by either linear or quadratic regression. This means that, throughout the DNA sequence, space between consecutive palindromes does not follow an increasing or decreasing trend. Neither does it follow a cyclic trend like the shape of a quadratic. However, this result should not be interpreted as meaning palindromes are randomly spaced throughout the sequence. The earlier analyses showed that there are indeed palindrome clusters of significance.

Conclusions and Discussion

Summary of Findings

Throughout our analysis on the dataset, it looks like the dataset does not confirm to a uniform random distribution of palindrome locations across ~229000 DNA locations. This can be proven by our tests of simulating the palindrome sites, along with also analyzing patterns of spacing and sums in the palindrome data. Our tests of significance, using a Chi-Square Test and a Poisson distribution along with hypothesis tests also demonstrated that it is unlikely that this data is from a uniform distribution, meaning it is unique. Additionally, our analysis of the palindrome locations and their counts in various regions of the data exposed that some intervals of locations were significantly more likely to contain palindromes, potentially showcasing that an origin of replication was somewhere in the data, likely in the location between the 92500 - 95000 mark.