Finding the Origin of Replication in DNA

Author 1 and Author 2

2024-11-05

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?

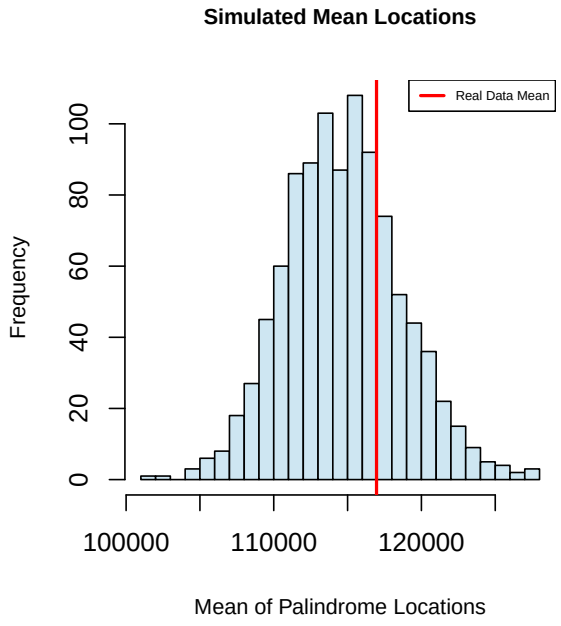
Question 1: Compare Palindrome Locations to Simulated Uniform Distributions

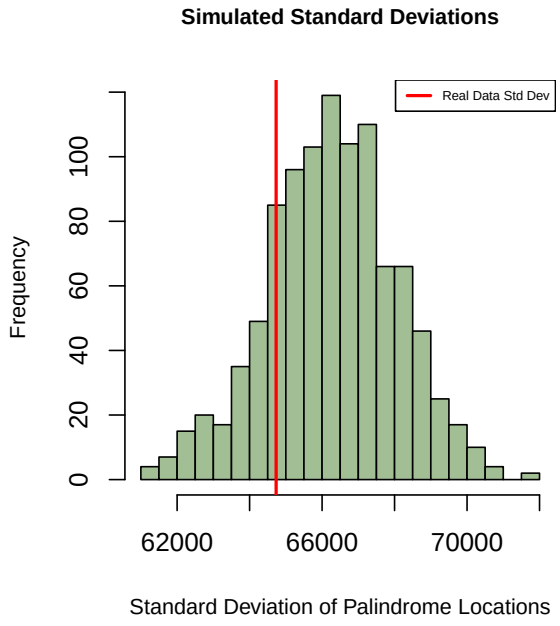
Methods

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 simulations show that there is a small deviation in terms of the means and standard deviations of the simulated data versus our real mean data. While our simulations had an overall mean location of ~117000, the real mean location was at ~114000. Similarly, our simulated standard deviation was ~64000 while the real standard deviation was ~66000.

Conclusions

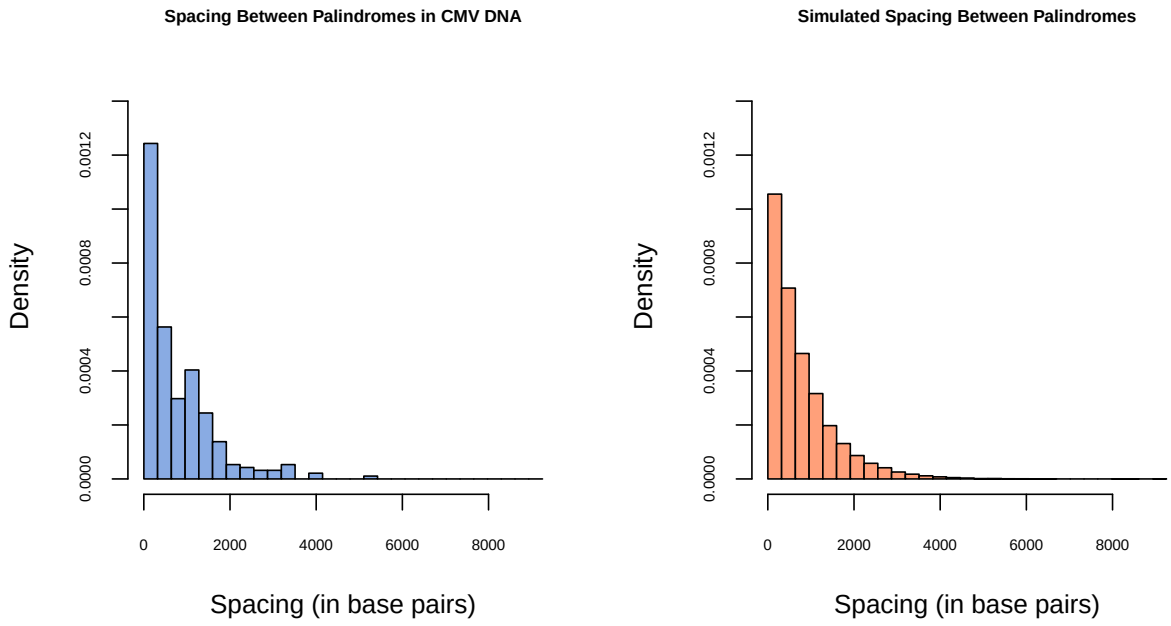
While these numbers are still very close to the simulated uniform distributions, and their differences could be attributed to chance, there is also a stronger likelihood now that the data doesn't necessarily come from a uniform distribution. Specifically, since the real mean location of the palindromes was slightly higher than our mean simulation, this could suggest that there are perhaps clusters or that the data could be left skewed, having more palindromes in later locations. Another possibility is that since the real standard deviation of the data is lower than our simulations, there are evident clusters within the data that are unlike a uniform distribution.

Question 2: Graphically Analyze Patterns in the Palindrome Data

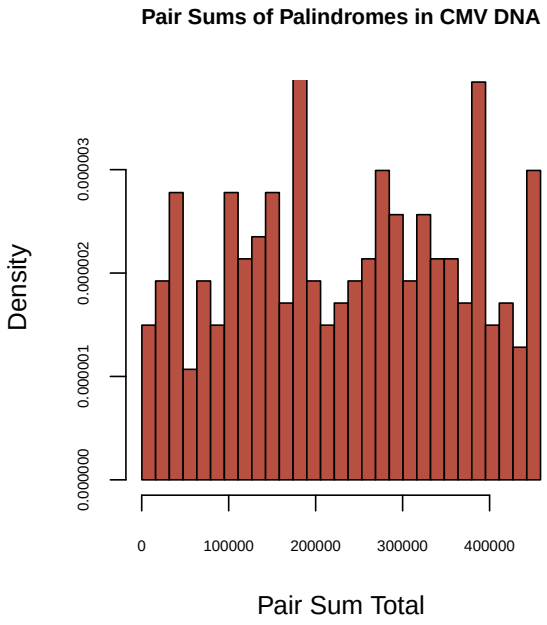
Methods

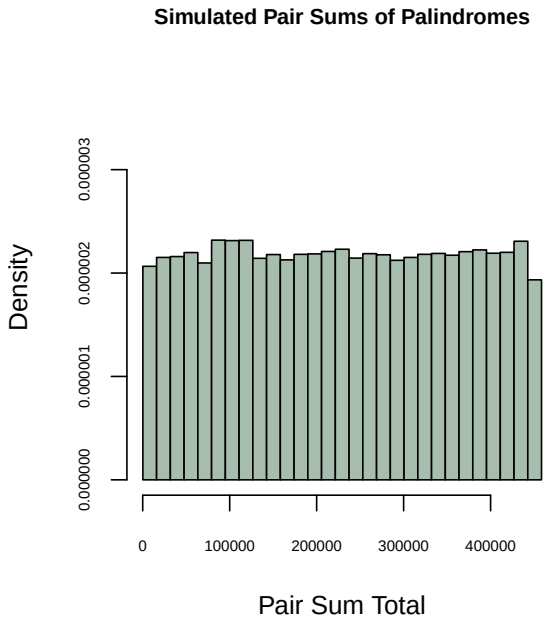
In order to further analyze the distribution of the palindromes, let's see three different variations of the data and how they deviate from a uniform scatter across the DNA sequence. To do this, we will compare the observed patterns below to expected uniform random distributions to identify any significant 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

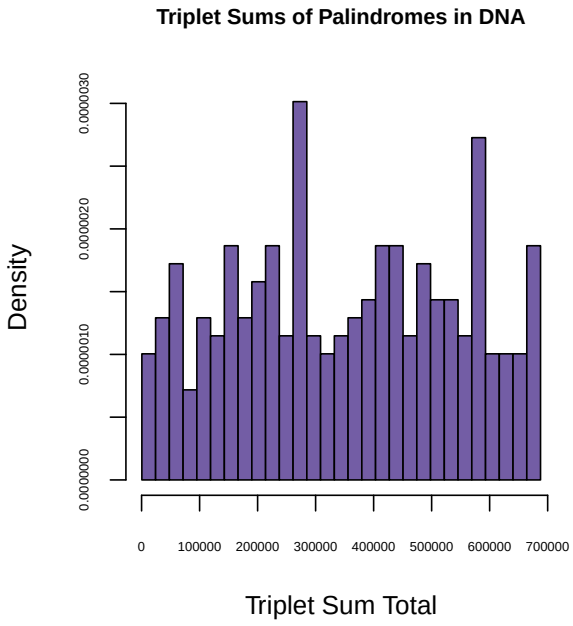


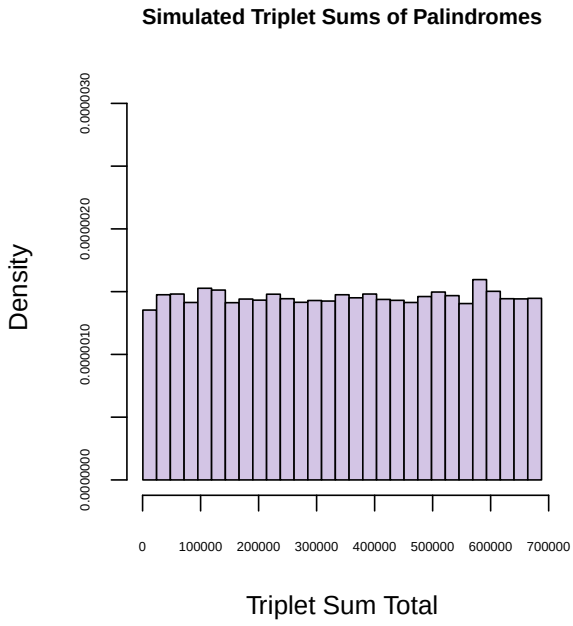
b: Sums of Palindrome Pairs





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

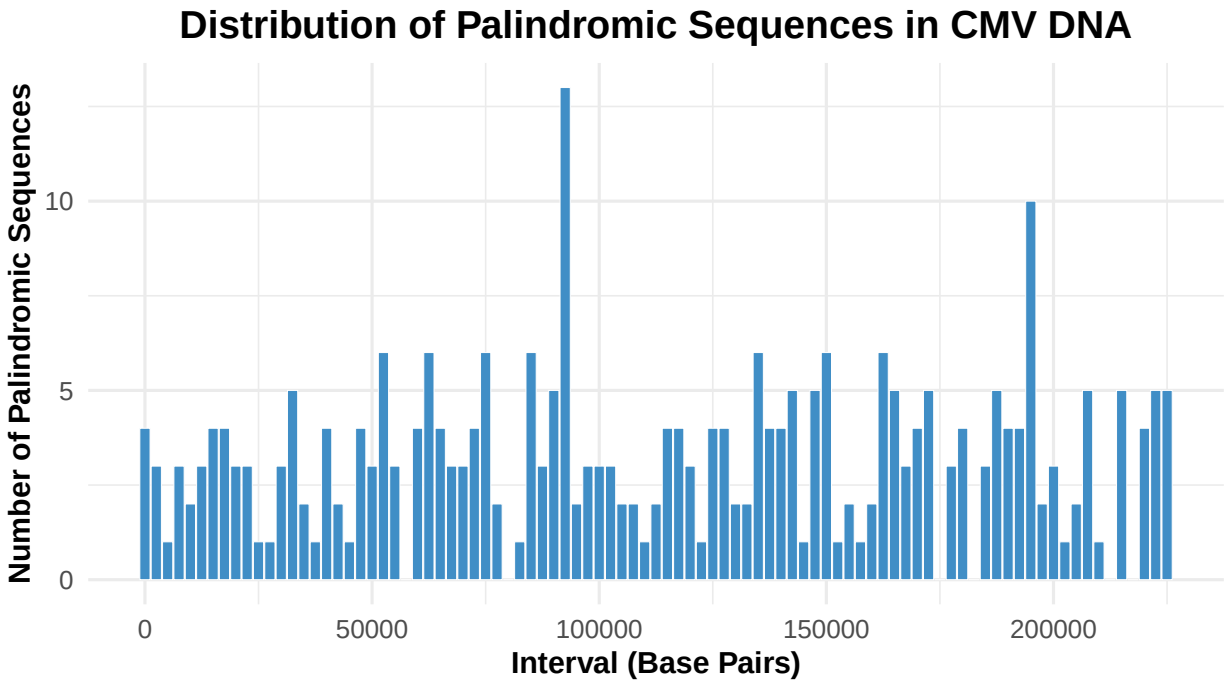
Analysis

Conclusions

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: Advise for a Biologist

just answer the question lmfao

Advanced Analysis

Methods

Analysis

Conclusions

Conclusions and Discussion

Summary of Findings

Discussion