Patterns in DNA: Combatting the Human Cytomegalovirus

Math 189 - Investigation #3

Professor Schwartzman

Spring 2020

Arthur Chang - A14410373

Raya Kavosh - A14826756

Siddharth Saha - A15572442

Contributions

Raya - Introduction and its subsections; random scatter and locations (part 1); conclusion
Arthur - Spacings including all graphs and visualizations and analysis
Siddharth - Scenarios 3 and 4 including all tests, graphs and analysis carried out there as well as the data limitation section of our analysis

Introduction

The human cytomegalovirus (CMV) can be life-threatening for people with a suppressed or deficient immune system. In order to combat the virus, scientists are in search of its origin of replication- a special place on the virus' DNA that contains instructions for its reproduction. A virus' DNA contains the information it needs to grow, survive, and replicate, in the form of sequences made of the letters A, C, G, T in various patterns. These patterns may indicate key sites in the DNA, including the origin of replication. One such pattern is the complementary palindrome, where a sequence of letters reads in reverse as the complement of the forward sequence. A and C are complementary, and G and T are complementary.

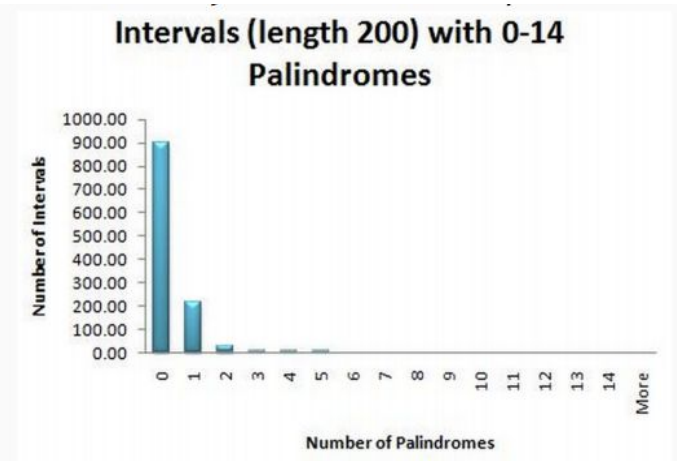
The origin of replication for two viruses from the herpes family, the same family as CMV, are marked by complimentary palindromes. Herpes simplex has a long palindrome of 144 letters, and the Epstein-Barr virus has several short palindromes and close repeats clustered at the origin of replication. The longest palindrome in CMV DNA is 18 base pairs. Altogether there are 296 palindromes between 10-18 base pairs long. Biologists conjecture that clusters of palindromes in CMV may serve the same role as the single long palindrome in Herpes simplex or the cluster of palindromes and short repeats characteristic of the Epstein-Barr virus' DNA.

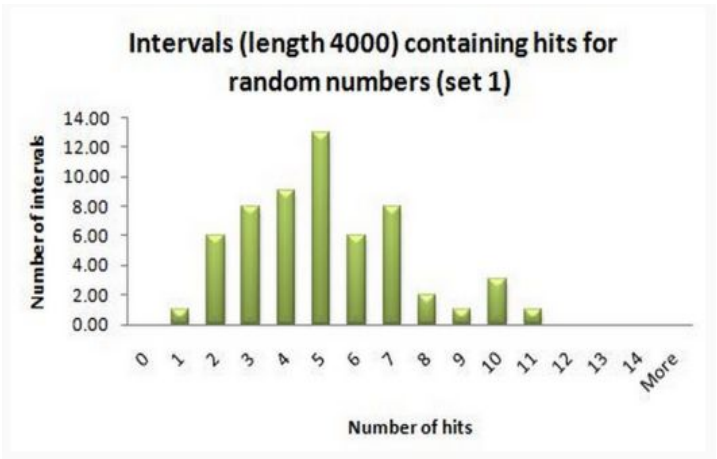
When trying to locate the origin of replication, DNA is cut into segments and each segment is tested to determine whether it can replicate. If it does not replicate, then the origin of replication is not contained in the segment. This process can be very time consuming and expensive without leads on where to begin the search. In order to narrow the search and potentially reduce the amount of testing needed to find the origin of replication, we will conduct a statistical investigation of the DNA to identify unusually dense clusters of palindromes.

Data

In 1990, the DNA sequence of CMV was published by Chee et al. In 1991, Leung et al. implemented search algorithms to screen the sequence for many types of patterns. Altogether, 296 palindromes were found that were at least 10 letters long (those shorter than 10 letters were ignored). The longest ones found were 18 letters long and occurred in locations 14719, 75812, 90763 and 173893 along the sequence. The CMV DNA is 229,354 letters long.

When comparing the number of palindromes per interval between the CMV DNA (left chart below) and the intervals of the random hits (right chart below), we can see that there are higher spikes of number of palindromes per interval and one or two outliers of intervals containing a higher number of palindromes regardless of the length of the intervals. However, no such consistent pattern of clusters of hits, or outliers, appear within the random numbers.





This evidence is enough to hypothesize that the clusters at the two locations on the DNA, as well as the outliers on the DNA are exceptions to the typical structure of the DNA chain and are not due to chance. They are worth examining for the replication code.
(Graphs in this section provided by Professor Shwartzman in Chapter 4 slides)

Background

In 1953, Franklin, Watson and Crick found that DNA has a double helical structure composed of two long chains of nucleotides. Each nucleotide has three parts: a sugar, a phosphate and a base. There are four bases which we refer to as the letters A, C, G, T. These letters stand for adenine, cytosine, guanine, and thymine, and they vary from one nucleotide to another. The two strands of nucleotides are connected at the bases, forming complementary pairs: A to T, C to G, G to C and T to A. The CMV DNA molecule contains 229,354 complementary pairs of letters or base pairs.

Viruses are made up of a DNA molecule wrapped in a protein shell called a capsid. The DNA stores all necessary information for controlling life processes, including its own replication. The DNA for viruses typically ranges up to several hundred thousand base pairs in length.

CMV is a member of the herpes virus family. Its incidences vary geographically from 30% to 80%. Typically, 10-15% of children are infected with CMV before the age of 5. The virus then lays dormant until young adulthood, when it presents symptoms similar to mononucleosis. CMV only becomes harmful when it enters a reproductive cycle and quickly replicates tens of thousands of copies. During this cycle, it poses a major risk for people in immune-depressed states such as transplant patients or AIDS patients. Locating the origin of replication of CMV may help virologists develop an effective vaccine for the virus.

Research Questions

The objective of this study is to identify unusually dense clusters of palindromes within CMV DNA in order to locate its origin of replication. We will need to examine these clusters and determine whether they occur by chance, or whether they indicate a potential replication site. To do so, we will investigate the following questions:

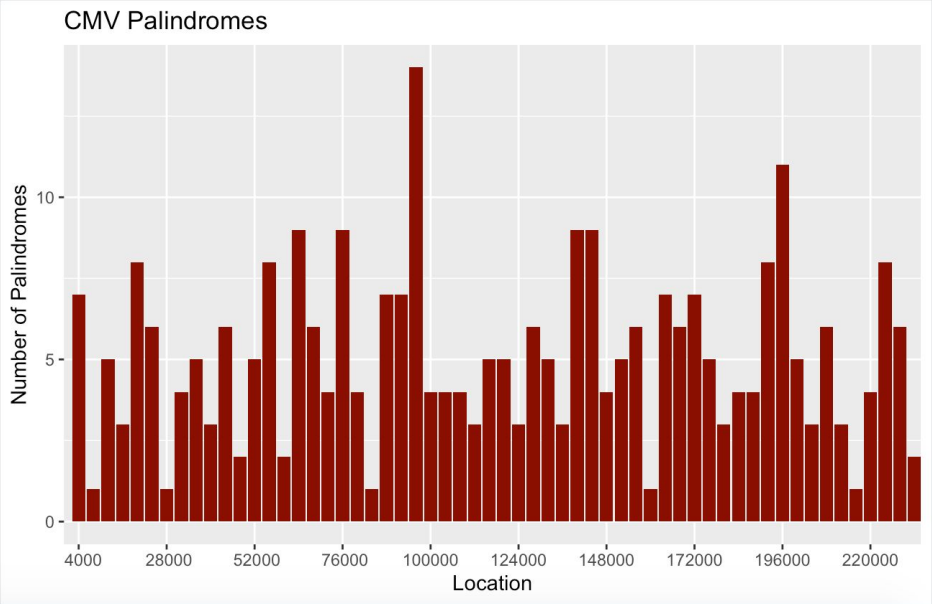
I. Are there departures from a uniform scatter of palindromes across the DNA that indicate structure?
II. Are the spacings between consecutive palindromes and their sets significantly different than what we would expect to find in a uniform random scatter?

III. Does the observed distribution of counts of palindromes match that of what we expect from a random uniform scatter model
IV. Biggest Cluster

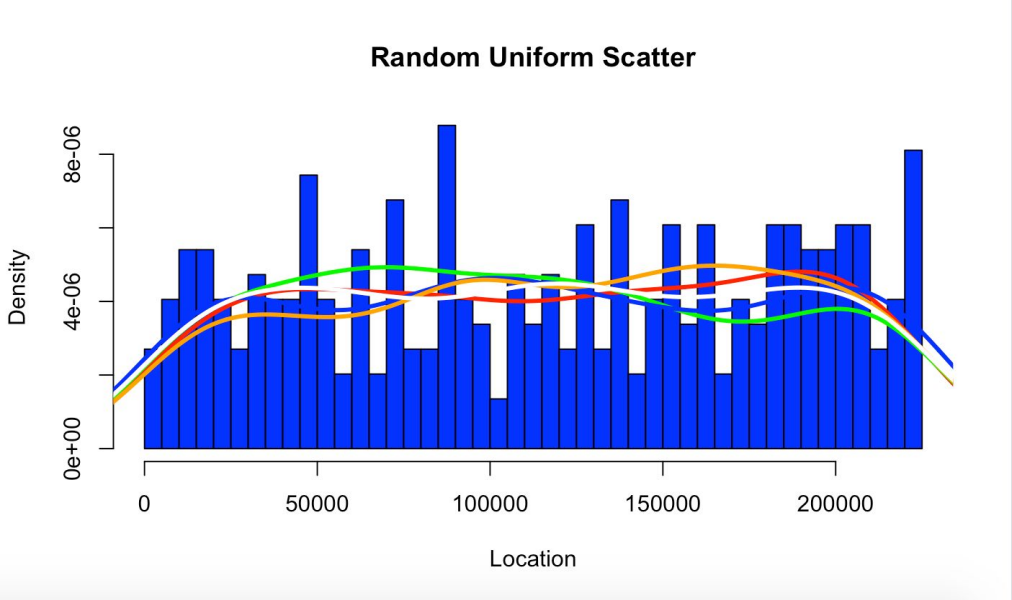
Analysis

Locations

Assuming that structure in the DNA is indicated by departures from a uniform scatter of palindromes across DNA location sites, we compared the counts of palindromes per location interval in our HCMV data to the counts of hypothetical "palindromes" per location interval generated using a uniform random scatter. To do this, we first created a histogram of the data, dividing the 296 palindromes into intervals of 4000 between the possible locations 1 to 229354. The histogram of the HCMV palindromes is shown below. We can immediately notice higher clustering in two intervals of the histogram around 95,000 and 196,000.

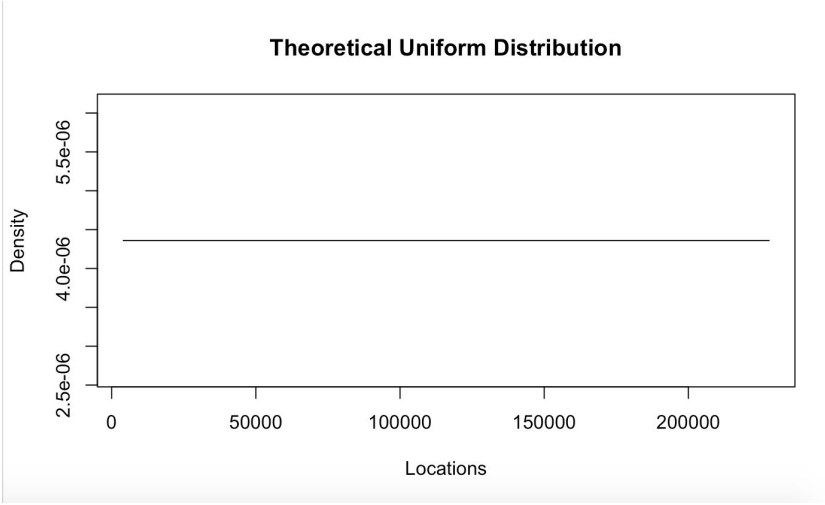


Next we simulated uniform random scatter. We generated five samples of 296 numbers between 1 and 229354 and plotted the cumulative distribution functions of their respective densities over the histogram of the first sample. We used the same intervals of 4000 to represent locations for each sample. The distributions of uniform random scatter are shown below.



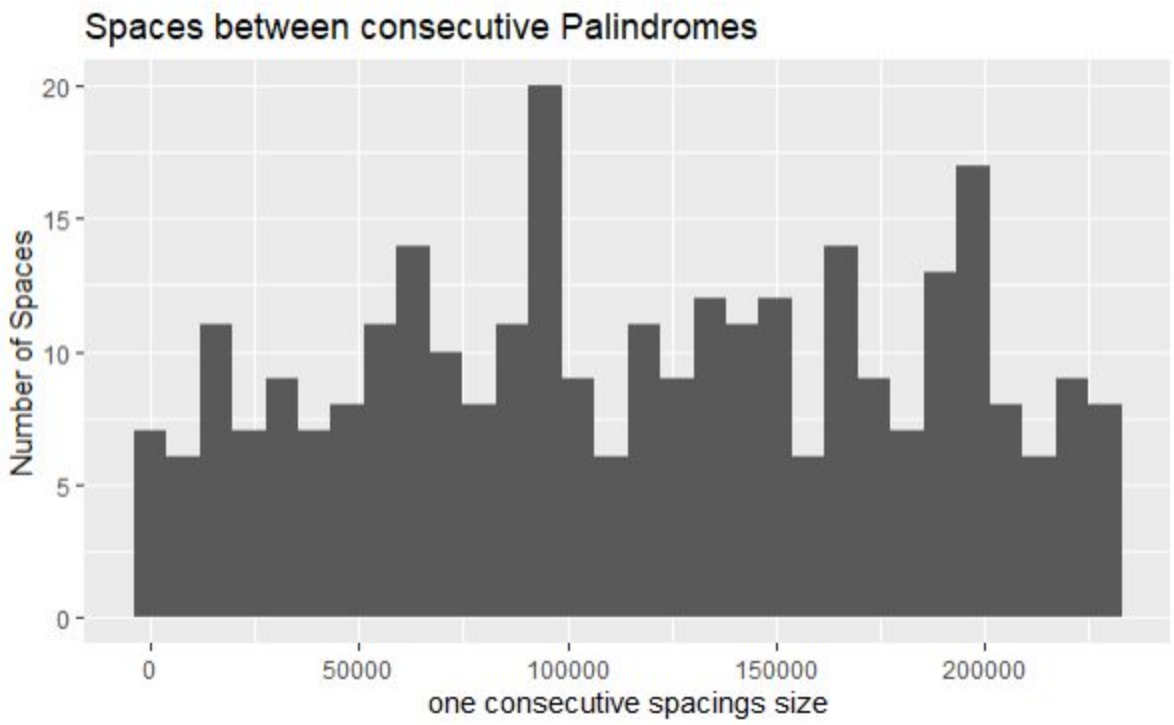
As we can see, in all of our samples there were no consistently concentrated clusters around any specific locations when simulating random scatter. It appears that the chance any given interval will contain a certain number of hits is the same for each interval over time, when many samples are taken. While every interval's respective number of hits is not necessarily equal, the distribution does not favor specific intervals over others.

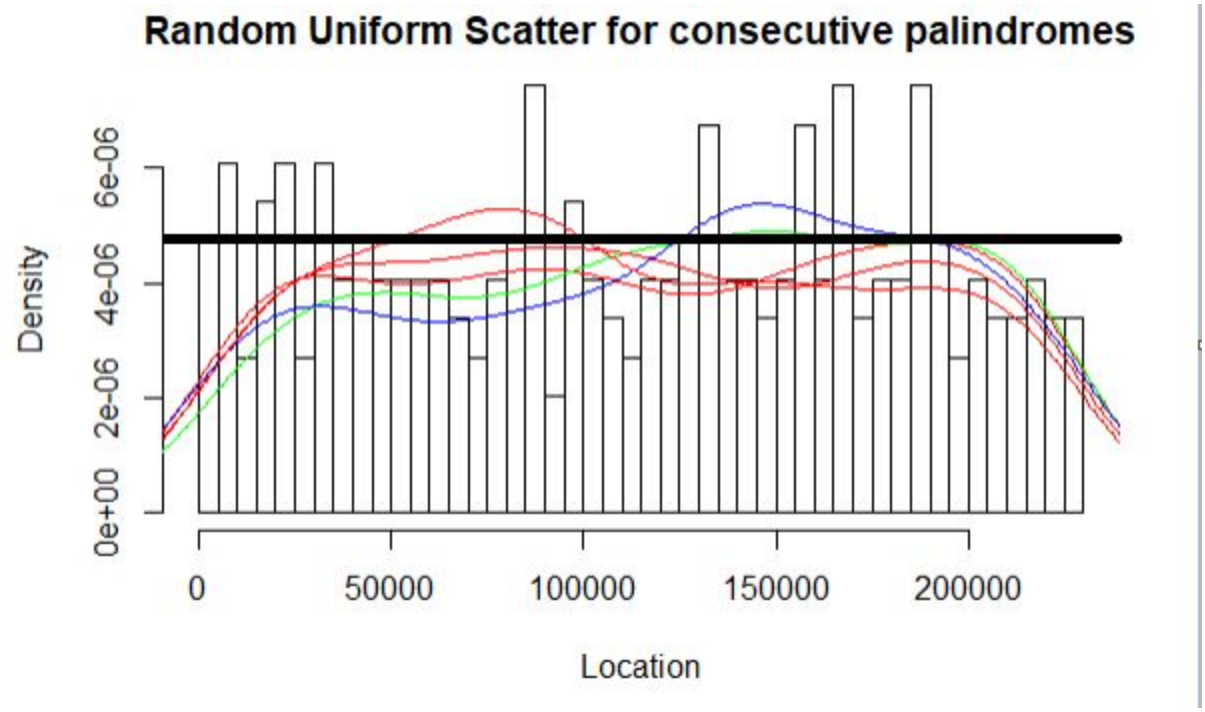
Lastly, we plotted the theoretical uniform distribution for intervals of 4000 from 1 to 229354. As shown below in the probability density function line, we would expect roughly an even number of palindromes per interval on average.



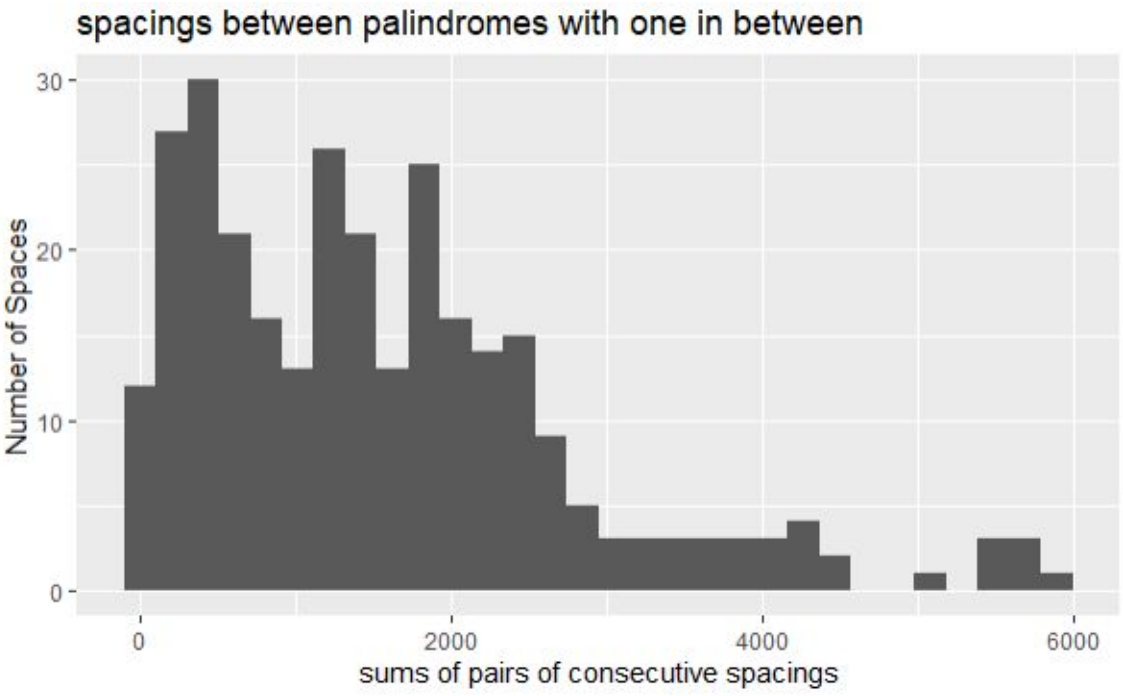
Comparing the histograms and CDFs for uniform random scatter and the function of the theoretical uniform distribution to the histogram for our data, we can see some clear differences. There appears to be a departure from a uniform random scatter of palindromes across the DNA in our CMV data that may indicate structure.

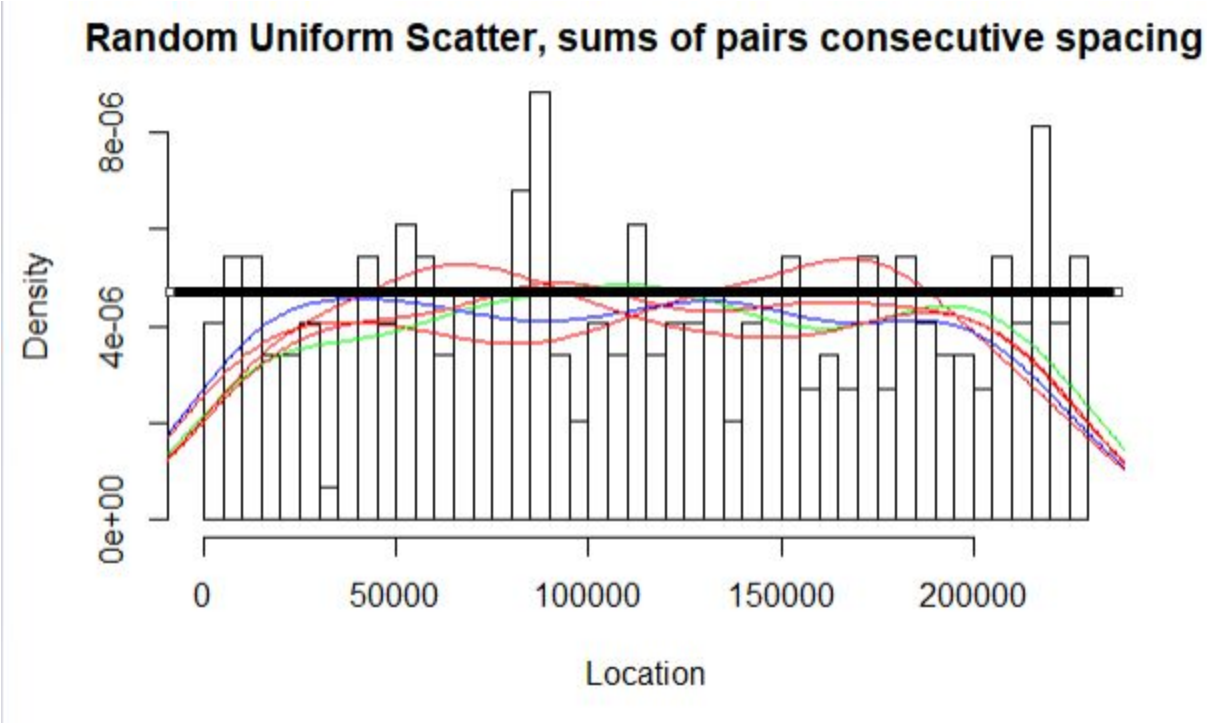
Spacings



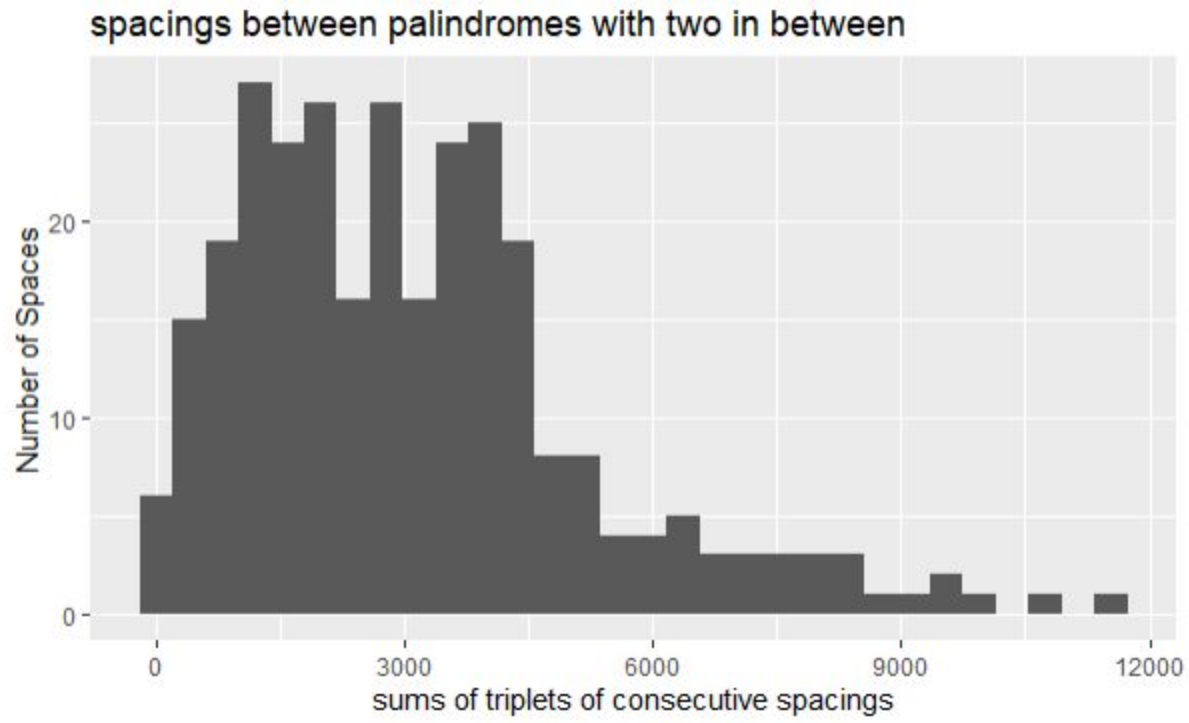


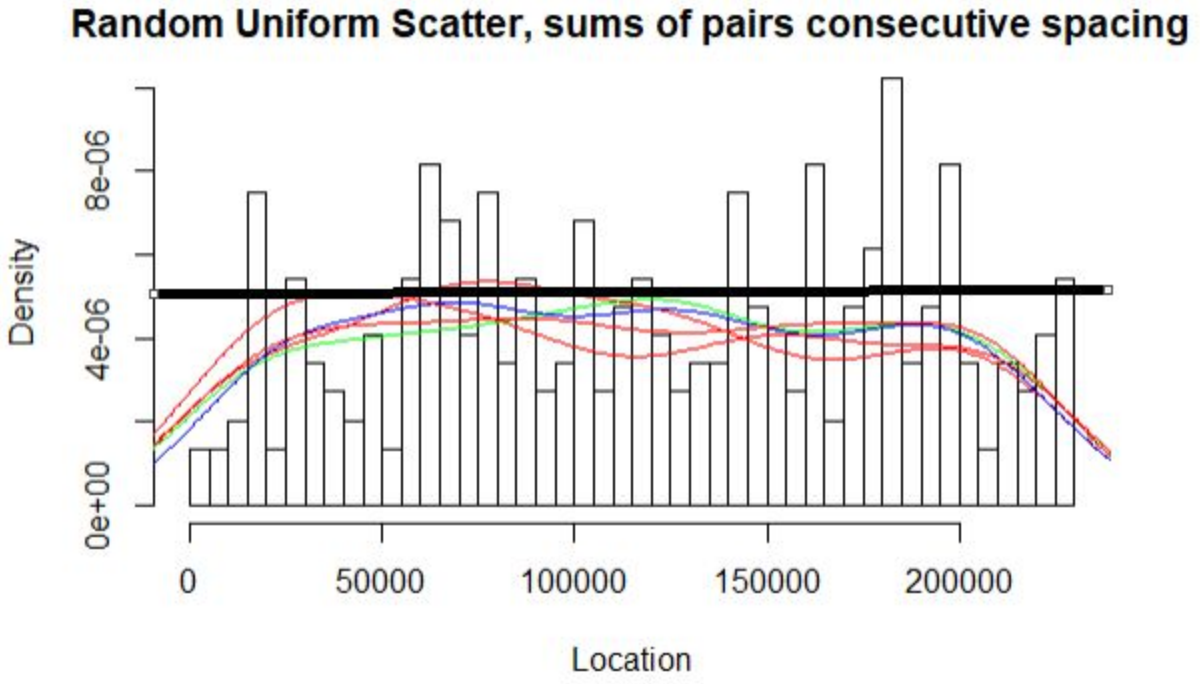
The two charts above show the spacings between consecutive palindromes visualization. The first chart graphs the consecutive spacing size differences using histograms. The second chart shows the random uniform scatter sampling from the consecutive spacing size differences.





The two charts above show the spacings between palindromes with one in between visualizations. The first chart graphs the consecutive spacing size differences using histograms. The second chart shows the random uniform scatter sampling from the sum of pairs consecutive spacing.





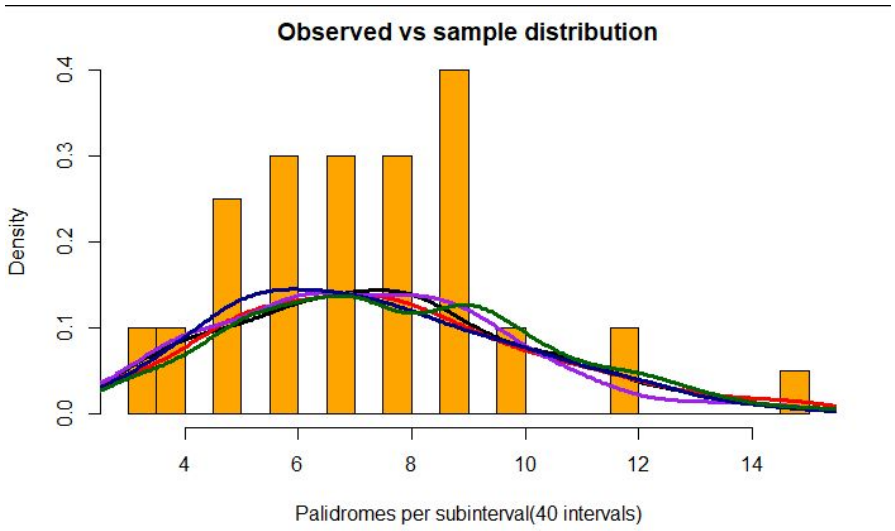
The two charts above show the spacings between palindromes with two in between visualizations. The first chart graphs the consecutive spacing size differences using histograms. The second chart shows the random uniform scatter sampling from the consecutive spacing size differences.

Counts

To carry out an analysis of the counts of palindromes in various regions of the DNA we split our palindromes into intervals of size: 40, 50 and 60. This is to ensure we still have a decent amount of data in each interval(if we were to believe that our distribution is truly randomly scattered) i.e we expect

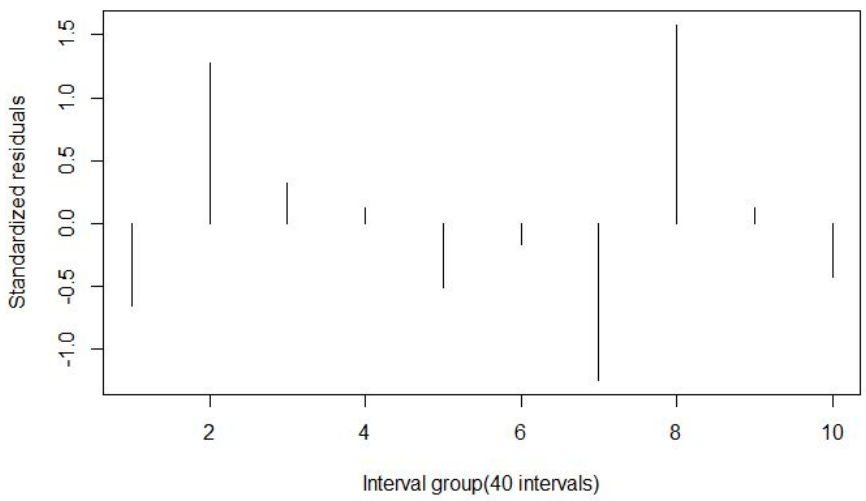
- 297/40 = 7.43 hits in our 40 interval splits
- 297/50 = 5.94 hits in our 50 interval splits
- 297/60 = 4.95 hits in our 60 interval splits

On examining the 40 interval split we see:



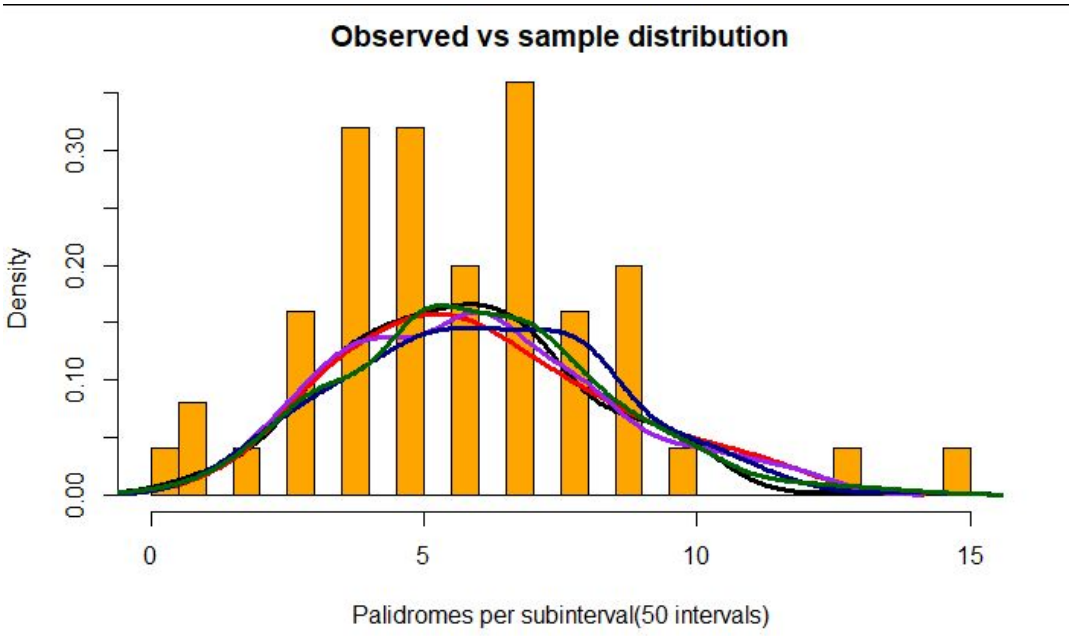
As indicated by the graph, the distributions are slightly different with the observed distributions of counts being more right skewed than the expected distribution as seen from 5 different random uniform scatters. These random uniform scatters are generated from a Poisson distribution as if the clusters are truly randomly uniformly separated we can expect an average of a certain number of hits from each one of them. This is displayed above as we expect 7.43 hits per interval and our Poisson distribution is centered around the 7.5 range.

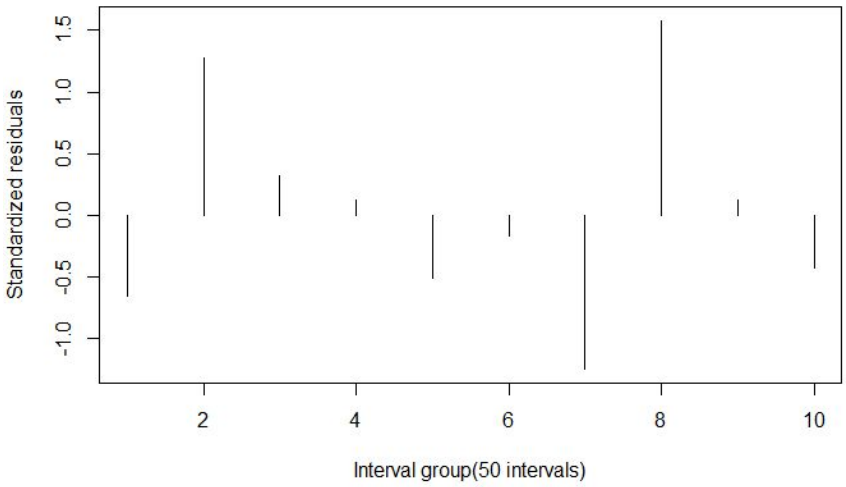
Furthermore we run a chi square goodness of fit test to see how well our observed distribution fits the random scatter distribution. On running this test we get a p values ranging from 0.997 to 0.999 indicating that we can't conclude the 2 distributions are different and that it is very likely that our palindromes per subinterval actually matches what we expect from a random uniform scatter. Finally we also plot the residuals and see how much the 2 distributions differ. Since all 5 random scatters give nearly the same residual plot:

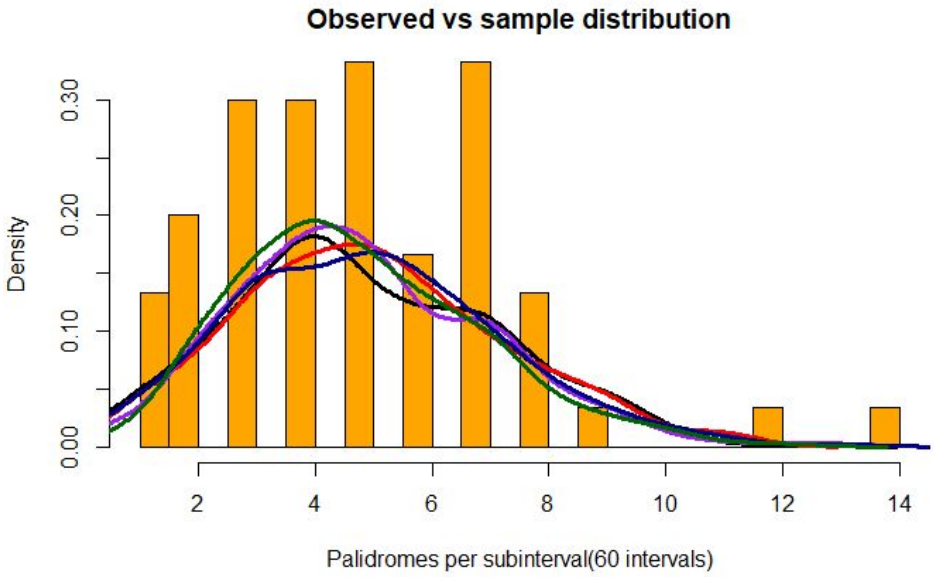


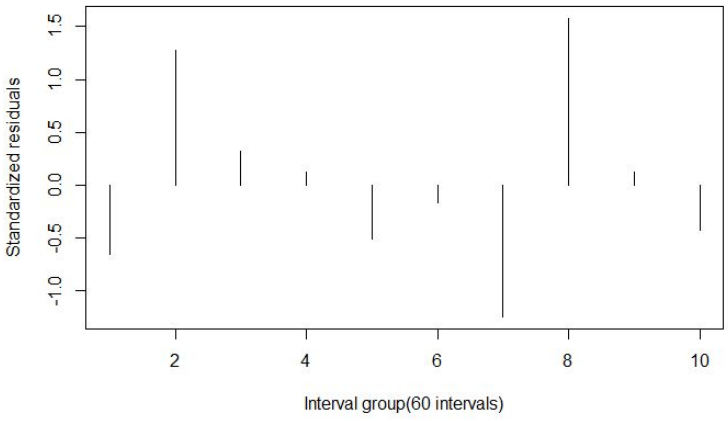
As we can see none of the residuals ever differ by more than 3 and thus we can say with more confidence that our counts of palindromes in each subinterval follows a random scatter.

Repeating the above analysis for number of intervals 50 and 60 gives us the following plots:




Standardized residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
2
4
6
8
10
Interval group(50 intervals)


Observed vs sample distribution
Density
0.30
0.20
0.10
0.00
2
4
6
8
10
12
14
Palidromes per subinterval(60 intervals)


Standardized residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
2
4
6
8
10
Interval group(60 intervals)

The results match what we get when we use a number of sub intervals equal to 40. Moreover as the number of sub intervals increases the graphs of the observed and expected distributions begin to match up more. Running chisq tests on the above intervals result in:

- 50: P values ranging from 0.2934-0.3063 across 5 random uniform scatters
- 60: P values ranging from 0.6322-0.6712 across 5 random uniform scatters

Overall while there is no pattern to the p values we can see that they are usually very high and as such do indicate that the observed distributions of count match that of what we expect from random uniform scatters

Biggest Cluster

In this question we look to answer whether or not intervals with a large number of palindromes are a potential site of replication. We answer this by analyzing the probability of such an event happening. If such an event is common enough (i.e it's common to get a high number of palindromes) then we may not worry too much about such sites. However if they are rare then they can very well be potential sites of replication. Using a theoretical idea(expanded on in the appendix) we calculate the probabilities of the maximum count occurring across 3 different interval sizes(40, 50 and 60) as: 0.3, 0.03 and 0.002. As the interval size increases we see that the chances of getting the maximum count lower to indicating that it can be a potential site of replication. Moreover this always happens in the site b/w $9.17 * 10^4$ and $9.56 * 10^4$.

Conclusion

From the initial visualizations of the locations of 296 palindromes found in CMV DNA, it appeared that the palindromes cluster around two particular locations which could potentially be the origins of replication. This was further verified by visualizations of various palindrome spacings, which did not distribute as equally as we would expect if they were scattered randomly. While we found that the observed distribution of counts of palindromes matches that of a random uniform scatter, we found that the site with the biggest cluster has a very high chance of being a potential site of replication. Moreover, the highest site of replication always occurs at $[9.17 * 10^4, 9.56 * 10^4]$ indicating that this is the best site to start at when looking for potential issues of replication.

Data Limitations

The biggest limitation we observe in our analysis is that we are only given a vector of locations. While that is a very important feature (as indicated by the amount of analysis done above) it is insufficient on its own. Paired with more data on other effects of each site it's value can drastically increase and provide much more information

Appendix

Question 4:

$$
\begin{aligned}
&P(\text{maximum count over m intervals} \geq k) \\
&= 1 - P(\text{maximum count over m intervals} < k) \\
&= 1 - P(\text{all interval counts} < k) \\
&= 1 - P(\text{first interval counts} < k)^m \\
&= 1 - \left[\lambda^0 e^{-\lambda} + \cdots + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}\right]^m
\end{aligned}
$$

In this formula k is the maximum count we observe in each interval, m is the number of intervals and lambda is the average hit we expect in each group. We usually get k values of 14-17 and m values are 40, 50 and 60 in each case.