Discovering the Origin of Replication in the CMV’s DNA

Author Contribution: Jerry wrote the Introduction and Data sections of the paper and completed the Locations part of the Analysis. Charlie completed the Counts and Biggest Cluster parts and took part in writing the conclusion. Nathan completed the Spacings part and took part in writing the conclusion.

Introduction

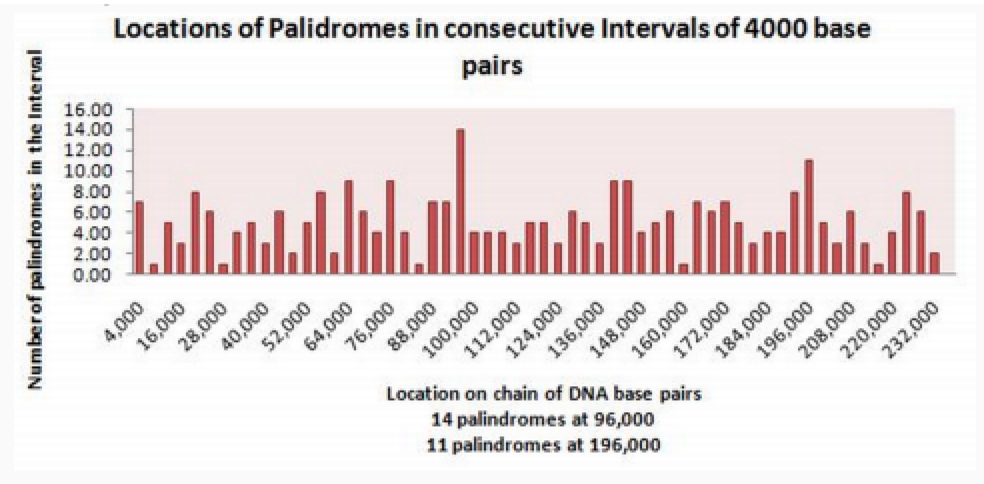
To examine the way viruses replicate, we use statistical procedures and pattern analysis to analyze the origin of replication and develop a life saving vaccine. In this study, our question asks: how do we find the significant complementary palindrome patterns in the human cytomegalovirus and determine if the cluster is a replication site? A complementary palindrome is a DNA sequence that reads in reverse as the complement of the forward sequence, which can indicate the origin of replication and in turn help virologists discover an effective vaccine. By using sampling and testing techniques, we can drastically reduce the time needed to make such a discovery, and tell the difference between chance occurrences and significant locations of palindromes.

Given data on 296 palindromes that were at least 10 letters long in the CMV DNA, which is 229354 letters long, our study first visually compares our sample palindrome distribution to the random uniform scatter, as well as the theoretical uniform distribution. Then, we examine the different types of spacings between our sample palindromes, such as between consecutive palindromes, between pairs of consecutive spacings, and between triplets of consecutive spacings. Next, we use a chi-squared test to examine the counts of palindromes for variable interval lengths. Finally, we use randomization to analyze the largest cluster of palindromes in a sub-interval for different interval sizes. At the end, we found that between the 90,000th and 108,000th pairs of DNA had the highest potential of containing the replication site.

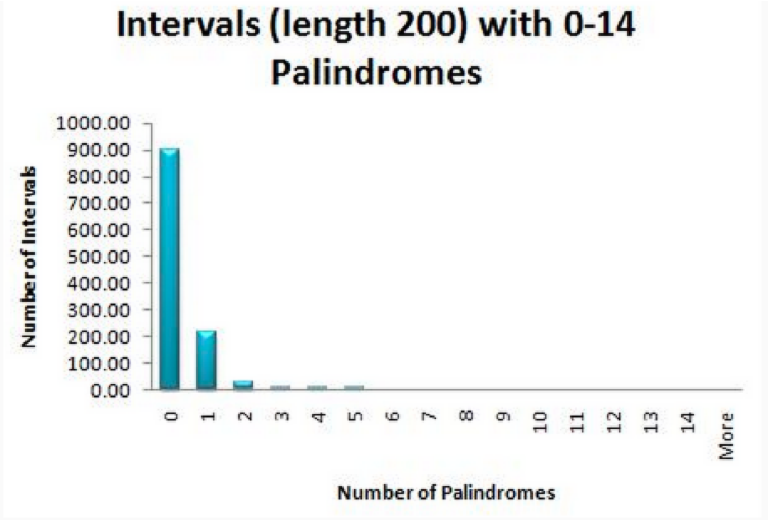
Data

Our data is created from the longest palindromes in the DNA of the human cytomegalovirus. In a sequence of 229354 characters, only palindromes longer than 10 base pairs were selected. Any palindrome length of less than 10 base pairs was not considered because they do not pose enough significance to be considered as a possible replication site.

The longest palindrome sequence is 18 base pairs long and occurred in 4 locations along the entire sequence, at locations 14719, 75812, 90763, 173893.



From this histogram, we can see that no matter the length of the interval, there always seems to be clusters of palindromes around the 93,000th and 195,000th pairs in the DNA sequence.

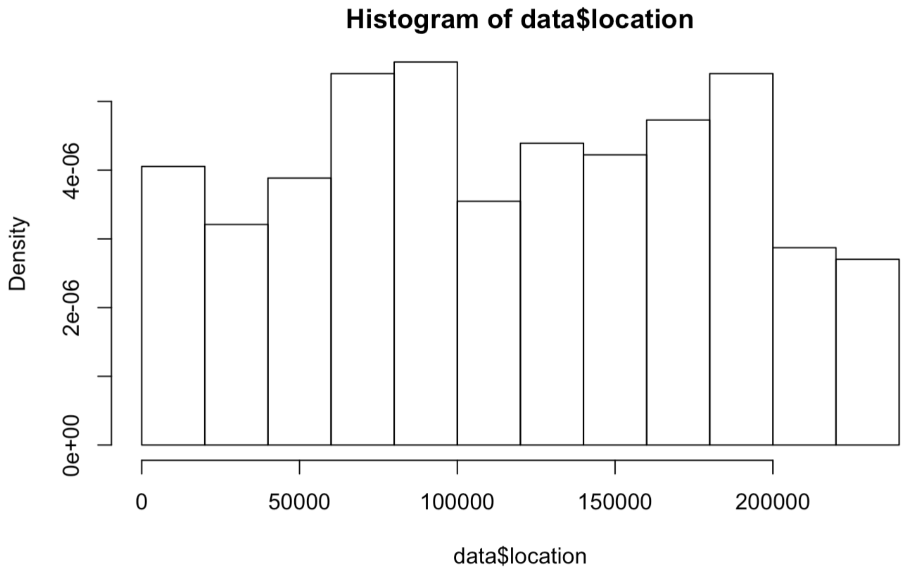


And from this histogram, we see that even when we change the length of the intervals, there are always 2 locations that have an unusually high number of palindromes.

Analysis

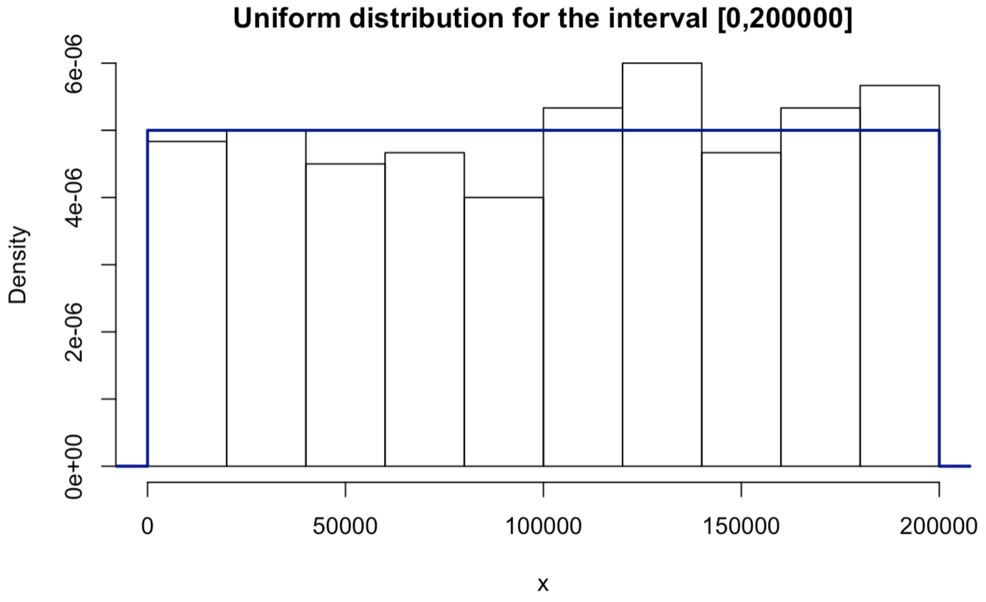
Locations

Our Sample Distribution



This is a plot of the palindrome locations given to us in our dataset.

Random Uniform Scatter Distribution (overlaid with Theoretical Uniform Distribution)



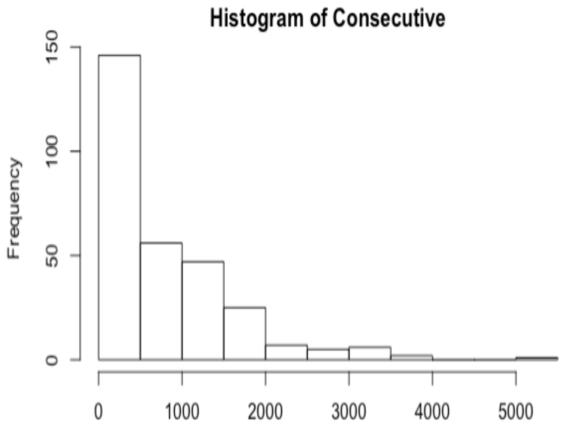
This is a simulated dataset of 300 palindrome locations on a genome sequence of 200000. From observing the sample distribution graph, we can see that there is an obvious spike in palindromes around locations 100000 and 200000. But if we look at the random uniform scatter distribution, even after conducting 5 samples, we can see that it does have an equal distribution.

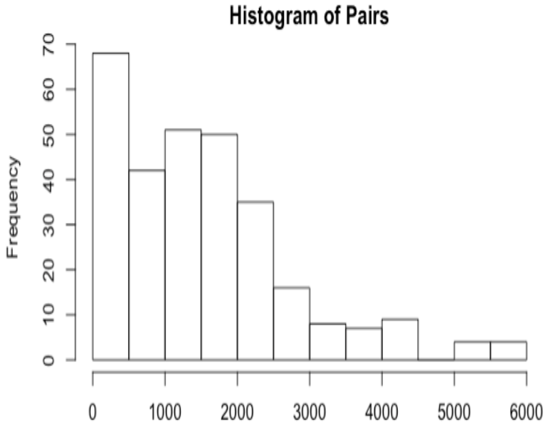
It also does not have spikes at the same locations as our sample data distribution. This distribution was simulated 5 times but only one graph was kept for conciseness.

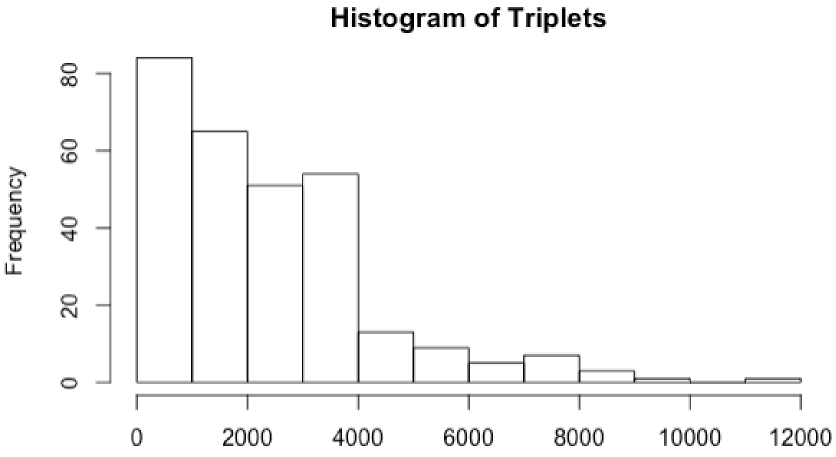
Spacings

Use graphical methods to examine the spacings between consecutive palindromes and sum of consecutive pairs, triplets, etc, spacings. Compare what you find to what would you expect to find in a random scatter. Also, use graphical methods to compare locations of the palindromes.

Now we will examine the three different types of spacings: spacings between consecutive palindromes, spacings between palindromes with one in between (sums of pairs of consecutive spacings), and spacings between palindromes with two in between (i.e. sums of triplets of consecutive spacings). We can generate histograms from the original dataset to analyze each type of palindrome spacing.

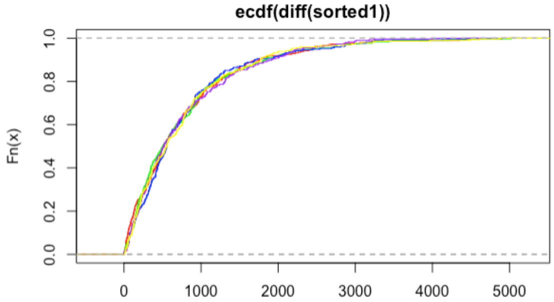


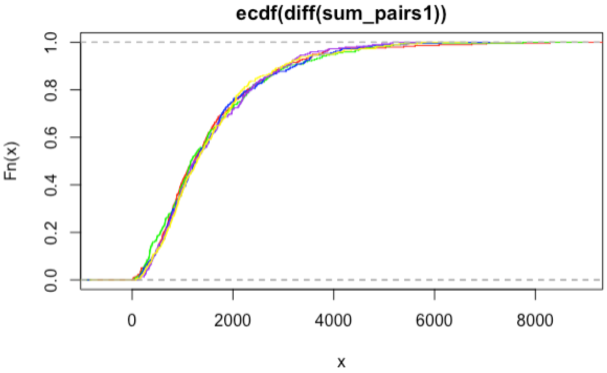


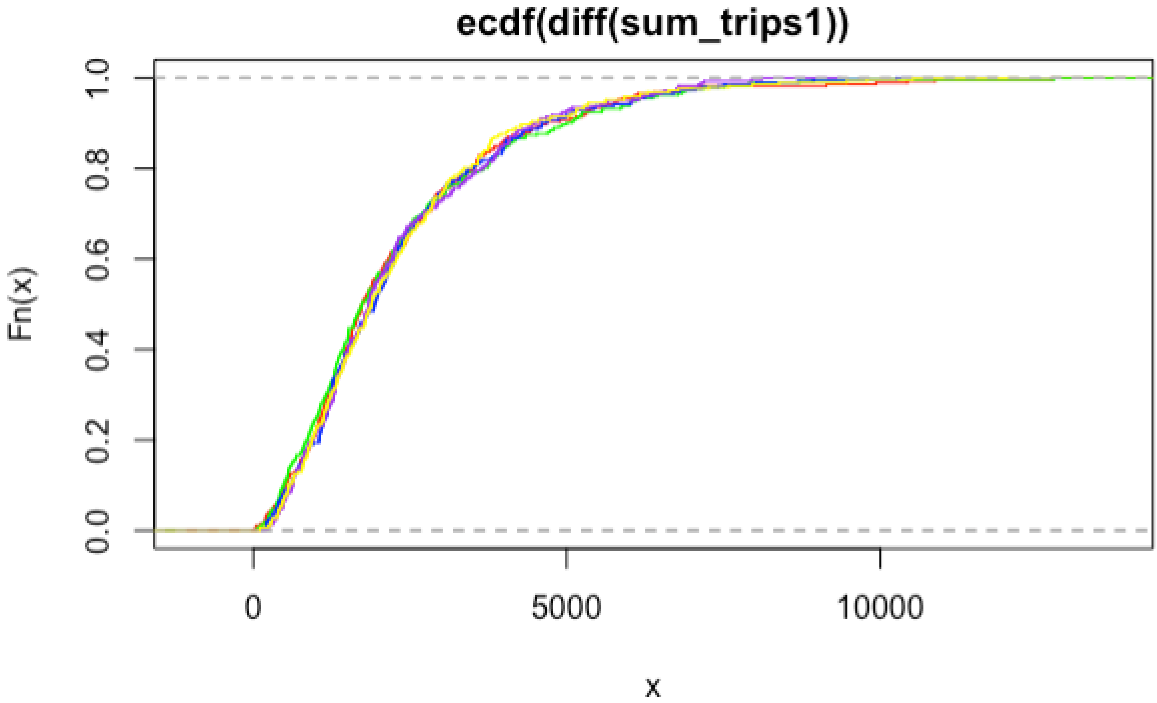


According to these visualizations, there is a pattern that all three types of spacings follow - they are all right skewed which implies a greater mean than median. Having a greater mean and median means that the palindrome spacing data has a few large outlier values that drive the data’s average higher but do not greatly change the median of the data.

To further analyze the locations of the data, we can create empirical CDFs to compare the 3 types.



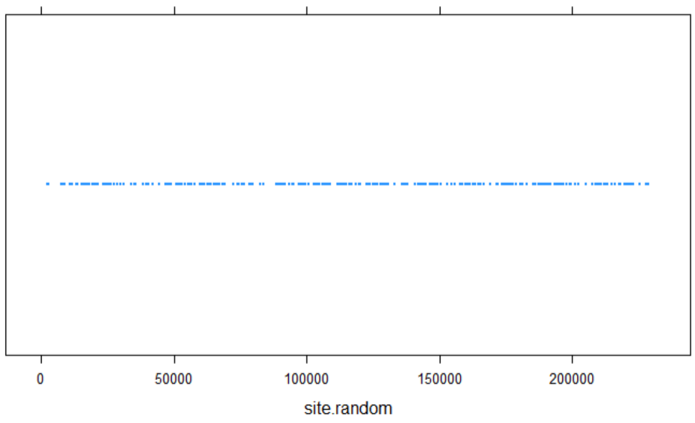


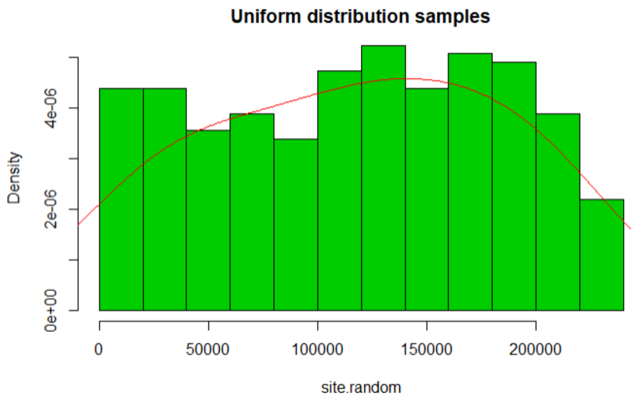


To construct these empirical CDFs, we took five different random uniform samples from the data and after identifying the three types of spacings of data within each random sample, we overlay the five different ecdfs that were generated. These graphs confirm what we found out earlier - that our data is right skewed and most of the data is clustered to smaller numbers with a small amount of values being extremely high.

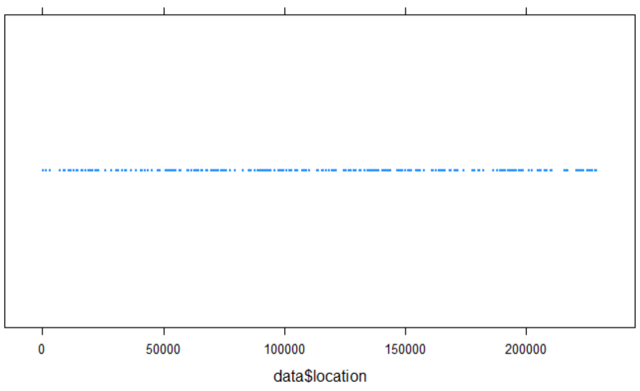
Counts

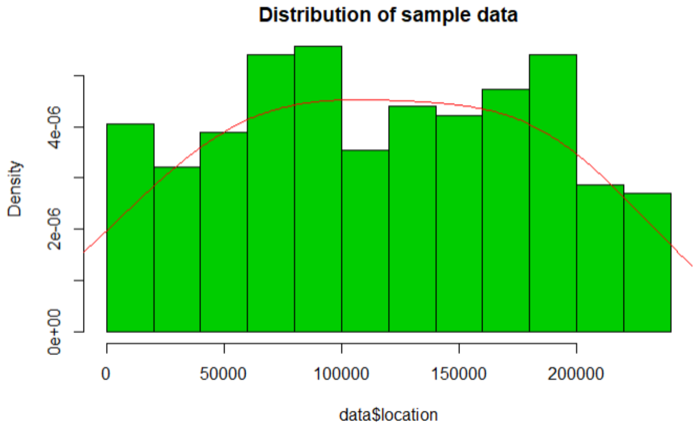
In this section, we want to use graphical and formal statistical methods to examine the counts of palindromes in regions of the DNA. To begin, we create a genome sequence of length 229,354, with 296 palindrome sites. Instead of simulating the runif function multiple times, we select our sites uniformly among 229,354 genome locations in the simulated sample. Graphically, our distributions like:





This verifies our simulated data are scattered uniformly across the number line. Now we compare this to our actual data:



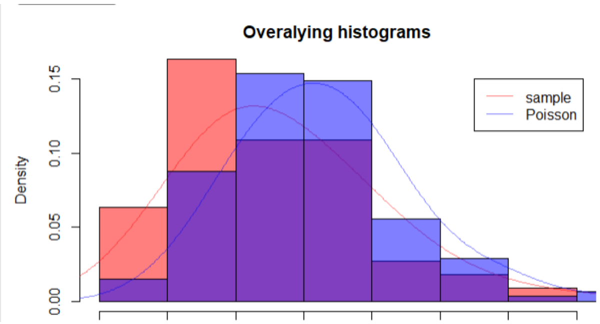


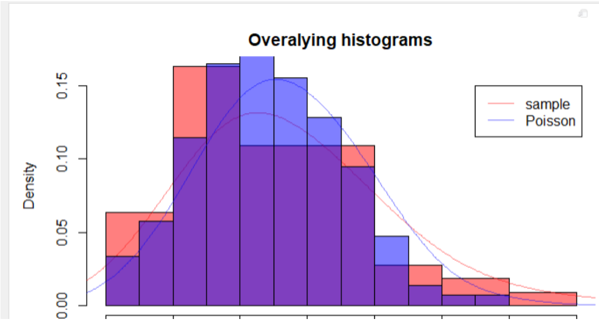
As one can see, just by comparing the two graphs, our simulated uniform data hardly has much difference from our sample data. Now we would like to compare the distribution of counts of our palindromes to that of a random uniform scatter. We will divide our data into non-overlapping regions of equal length and examine the counts. We know that the counts of the number of points in different regions follow Poisson distribution. Since there are too many possible positions (229,354) with only 296 observed sites, it's not really doable to perform a chi-squared Test directly to the dataset. Rather, we performed the following procedure:

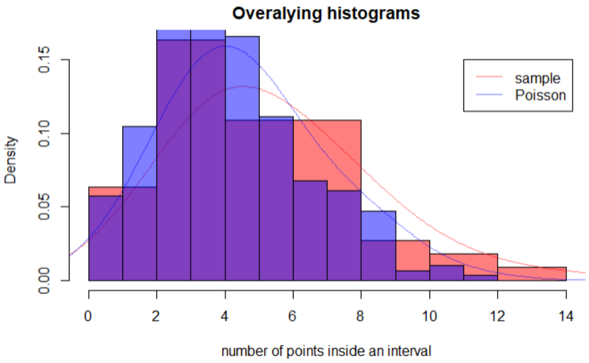
1. Split our data into non-overlapping regions of equal length
2. Calculate the counts in each region
3. Truncate (group) the counts that are greater or equal to 9 into one cell. (We are able to do this because there are counts that are able to be grouped together.)
4. Compare our observed counts to expected counts calculated using the probability mass function of the Poisson Distribution
5. Perform a chi-squared test to test whether the counts follow a Poisson distribution and obtain a p-value
6. Repeat for 3 different interval lengths and 5 different random uniform scatter simulations

Interval Length: 45 Interval Length: 55 Interval Length: 65

Levels	**Expected Counts**	Observed (random)	Observed (data)		**Expected Counts**	Observed (random)	Observed (data)		**Expected Counts**	Observed (random)	Observed (data)
0	0.06	0	0		0.25	0	0		.68	2	2
1	0.41	0	0		1.36	1	1		3.11	1	3
2	1.35	2	2		3.66	4	6		7.09	8	5
3	2.97	4	2		6.57	9	5		10.78	15	10
4	4.88	6	5		8.84	9	13		12.26	4	17
5	6.43	4	9		9.52	6	8		11.17	14	8
6	7.04	6	5		8.54	9	4		8.48	8	8
7	6.62	8	9		6.56	6	8		5.51	8	8
8	5.44	7	5		4.42	6	4		3.12	2	2
9	9.80	8	8		5.27	5	6		2.78	3	2







We select these intervals because if we based our test on much shorter interval lengths, we will get many more intervals but a larger proportion of them would contain 0 palindromes. We don’t use too large of intervals either because we have hardly enough data to compare observed and expected numbers of intervals for a particular palindrome count.

We can see that the count for our sample is very close to that of a Poisson distribution. We perform a Pearson’s chi-squared test which establishes whether an observed frequency distribution differs from an observed theoretical distribution. We want to check whether our sample does follow the Uniform distribution and thus, it’s count will follow the Poisson distribution. We find:

	Random Uniform Distribution p-value	Sample Distribution p-value
Interval Length: 45	0.91	0.88
Interval Length: 55	0.95	0.59
Interval Length: 65	0.12	0.57

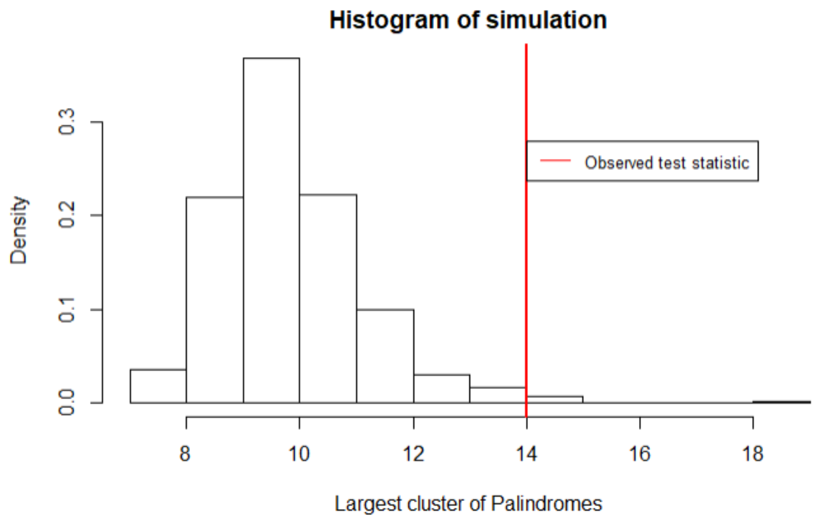
Keep in mind that the null hypothesis of the testing procedures above assumes the site locations are formed by a random scatter. Thus, the conclusions show that with a significance level of 0.05, we fail to reject the null hypothesis for our sample.

The Biggest Cluster

Here, we utilize randomization through simulation to examine the largest cluster of palindromes in a sub-interval. We have shown that our samples counts appear to follow a Poisson distribution and thus we can begin to search for unusual clusters. Once again, we choose interval lengths of 45, 55, and 65. In our simulation, we generate a genome sequence of length 229,354, with 296 palindrome sites in each iteration. We have verified in the section ‘Counts’ that this simulated a random uniform scatter. We then split the new sample into non-overlapping of equal length and calculate the counts in each region. Then, we append the max, or largest cluster to our array and repeat the process 1000 times. Finally, we calculate the probability of finding in any subinterval, a count as large or larger than the observed count in our sample and set our significance level to be 0.05.

	P-Value
Interval of length 45	0.016
Interval of length 55	0.001
Interval of length 65	0.0

In each simulation, we reject the null hypothesis that our observed largest cluster is likely to be seen in a uniform random scatter. This is the histogram for choosing 55 subintervals:



Therefore, no matter the length of the interval, our observed largest cluster in our sample is not very likely to occur under a uniform random scatter distribution. Hence, we would advise the team and biologists to begin their search for the origin of replication at the site of the largest cluster we found, which happens to be in the range, depending on interval lengths, between the 90,000th and 108,000th pairs of DNA.

Results

Through the locations analysis, we visualized the distribution of palindrome locations in our given sample and compared it to the random uniform scatter. After conducting 5 samples of the random uniform scatter, we found that it did not produce spikes in palindrome occurrences similar to our sample data distribution. Even though this provides evidence that those locations could be significant, we continue to analyze the spacing between the palindromes. From our graphical analysis of the three different types of spacings, we found that the data is right skewed, meaning that our data was mostly clustered at smaller values with some large values occurring spiking the mean of the dataset. After examining our sample spacings, we used graphical and statistical methods to examine the counts of palindromes in regions of the DNA. First, we simulated a uniform random scatter distribution to compare our sample against. We find that by the eye test, they do not look that much different but proceed to look further into the counts. We divided our data into non-overlapping regions of equal length (45, 55, 65) and examined the counts. Since we know that the counts of the number of points in different regions follow Poisson distribution. We found that after 5 simulations with different uniform random scatter distributions, we find through graphical means that our samples counts appear to follow a Poisson distribution. We perform a Pearson's chi-squared test and find that each time, we fail to reject the null hypothesis with a significance level of 0.05. After examining our sample counts, we proceed to look at the largest cluster of palindromes in a sub-interval. We use interval lengths of 45, 55, adn 65 once again and use randomization to obtain the probability of finding in any sub-interval, a count as large or larger than the one observed in our sample. After multiple simulations and different interval lengths, we found that our observed cluster is unlikely to be found under a uniform scatter distribution and rejected the null hypothesis of our simulation study each time with a significance level of 0.05.

Conclusion

(discuss data limitations)

Answer this: How would you advise biologist who is about to start experimentally searching for the origin of replication?

In our analysis of locations, we found that there is not an equal distribution so some locations may be better to look into for the origin of replication. Furthermore, looking at the spacings, we found out data to be right skewed with a large amount of the data being small values and a small amount being large values. Lastly, when we analyzed the random clusters we concluded that the largest cluster was actually between the 90,000th and 108,000th pairs of DNA which is where we would advise biologists to start searching for the origin of replication.