Case Study 3: Search for the Unusual Cluster in the Palindromes

Question

How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence of a potential replication site?

Hypothesis

The question at hand here concerns finding clusters of palindromes within sequences of DNA. We hypothesize that, by breaking down DNA into segments and testing for large, unusual clusters in this way we will be able to determine ultimately whether or not a cluster is a one-off occurrence of a replication site. This would also serve as a more practical approach to solving a problem of this magnitude given the time and money that comes with conducting these types of experiments.

Literature Review

Research into the origin of replication for virus DNA is not limited to CMV. In the simian virus 40's (SMV) DNA, the origin of replication is marked by a series of base pair palindromes (Bergsma et al). After sequentially deleting subsets of these palindromes, it was found that replication efficiency decreased in viruses where a specific 21 length palindrome was deleted.

Likewise, specific palindrome deletion was studied in CMV. In the CMV DNA, a series of 8 base pair palindromes was responsible for virus proliferation. It was found that when these length 8 palindromes were entirely removed from the DNA, there was a 15-fold reduction in virus replication (Chang et al.). When other palindromic sites were removed, there were no significant alterations in virus replication. This research has provided scientists with the knowledge of which palindromes to delete to reduce the severity of the virus.

Introduction of Data

The main dataset we will be using in our analysis was produced by Jessica Leung et al. in 1991. A year after the DNA sequence of the cytomegalovirus was published, Leung's team

screened the sequence in an attempt to isolate any patterns within the DNA. In the 229,364 length DNA sequence, 296 palindromes of at least length 10 were found. Palindromes shorter than length 10 were deemed insignificant. The longest palindromes, located at positions: 14 719, 75 812, 90 763 and 173 893, were of length 18. In the dataset, found in file hcmv-263hxkx, the table contains a column labeled "location" that contains the positions of the 296 palindromes of length greater than 10 that were found in CMV 's DNA sequence.

In addition, two other data resources were provided. The first of these contains data relating to human CMV. For each of the 2174 Ugandan individuals that make up the sample, health metrics (e.g., CMV and HIV positivity) are given. The original study was intended to identify whether CMV was correlated with the presence of either tuberculosis or cardiovascular disease in low- and middle-income countries where the presence of human CMV is higher. The second dataset is provided by the National Center for Biotechnology Information. This dataset was produced from a study that looked for the presence of DNA palindromes in MCF-7 and IMR90 breast cancer cells line. For each of these cell types, millions of DNA samples were isolated and analyzed for palindrome occurrences.

When considering the questions we seek to answer, namely "How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?", and when taking into account that our primary dataset only contains palindrome locations for CMV DNA, it is hard to generalize our findings . It is unknown if the clustering of palindromes in CMV is unique or representative of clustering of palindromes of other DNA sequences, and thus cannot be extrapolated to other DNA sequences. If our primary goal was to determine if there was unusual clustering specifically in CMV DNA, our results would be reliable.

Background

DNA, or deoxyribonucleic acid, is a molecule composed of chains of nucleotides that form a double helical structure. Each nucleotide is composed of one of four nucleobases: cytosine, guanine, thymine and adenine. DNA molecules encode the information necessary for organism and virus growth, development, reproduction and function. Mutations in DNA can cause disruption in gene activity and lead to diseases.

Part of the herpes family, CMV is a common virus whose prevalence is very high; incidence of the disease vary from 30%-80%. When infected, you have the virus for life. While CMV very rarely causes health concerns for healthy people, it can be potentially fatal for pregnant women and those with weakened immune systems. Incidences of death from CMV are higher in low-income nations than among high-income nations.

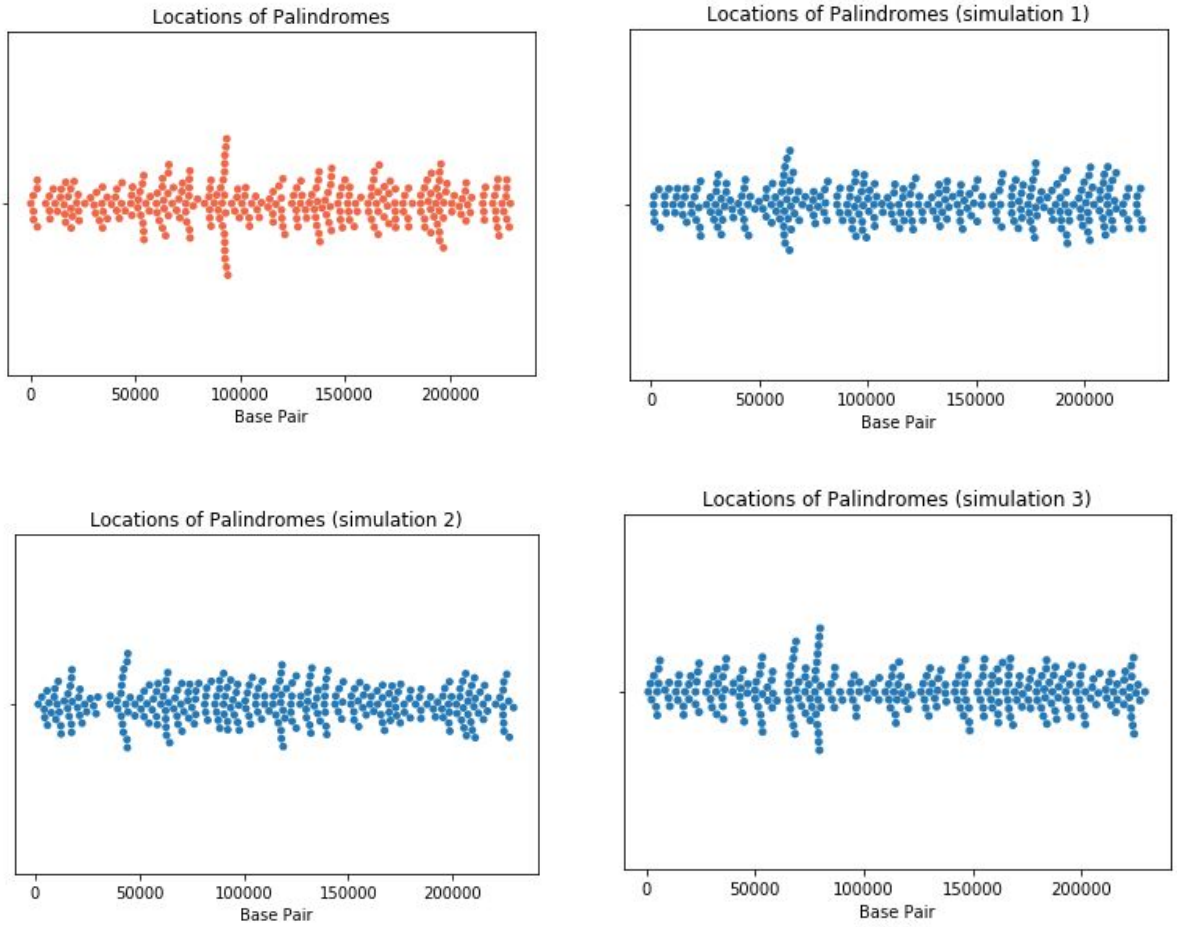
To develop treatments for viruses, scientists analyze, among other virus functions, how it replicates. As virus DNA encodes all the information required for replication and growth, scientists specifically look for the DNA's origin of replication, or where the DNA encodes for replication.

For two other viruses in the herpes family, their origins of replication are indicated by the presence of a palindrome/palindromes. With this knowledge, it is thought that similar palindrome markers may be present in CMV. Finding origin of replication for CMV could potentially help virologists produce a vaccine for CMV in an efficient manner.

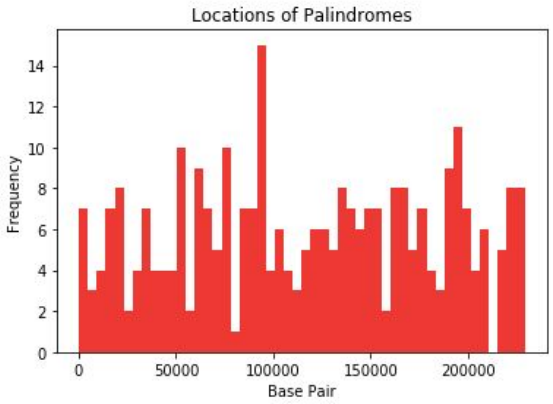
Basic Analysis

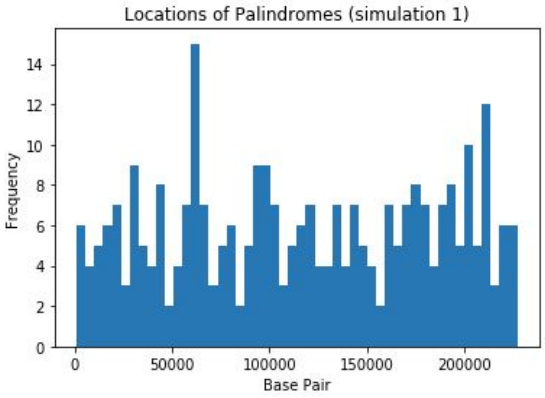
[Scenario 1]

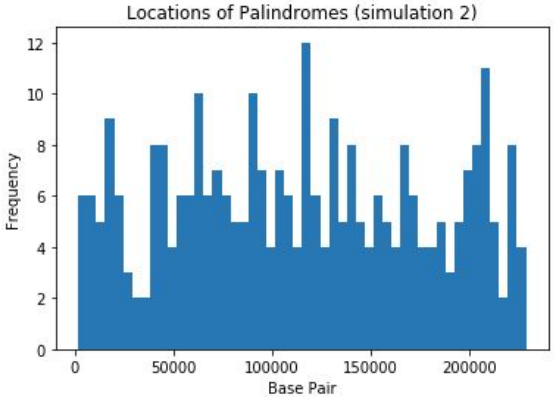
We first plotted a swarm plot of the locations of the palindromes across the DNA to help us better understand the structure of our data. The first plot was using our original data, while plots 2-4 were generated using a monte carlo simulation. The visualization of the original data is in red and the simulations in blue. We randomly generated numbers from a uniform distribution for the simulations. It is important to note that in the simulations the data will not be equally spaced, because it was generated from a uniform distribution. The way our monte carlo simulation worked was by randomly scattering 296 palindromes along a DNA sequence of 229354 base pairs. The monte carlo simulations were to replicate what DNA would look like with uniformly scattered palindromes. In the simulations we can clearly see that the palindrome locations are not equally spaced. In the original data we can see that data appears to be clustered around the 90000th base pair, which is consistent with our domain knowledge. It is a little harder to see the cluster around the 190000th base pair. The cluster around the 90000th base pair are however missing in the simulations.

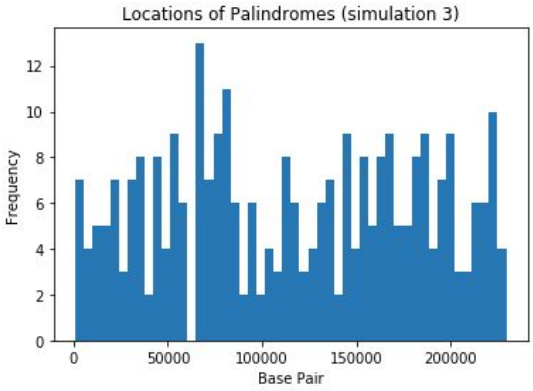


When we plotted the palindrome locations on a histogram, it became much easier to see the clusters in our original data. The clusters around the 90000th and 190000th base pairs are much more evident. Again we can see that these clusters are not present in our random scatters.

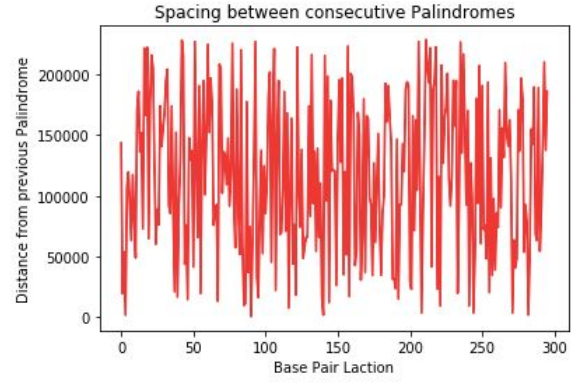


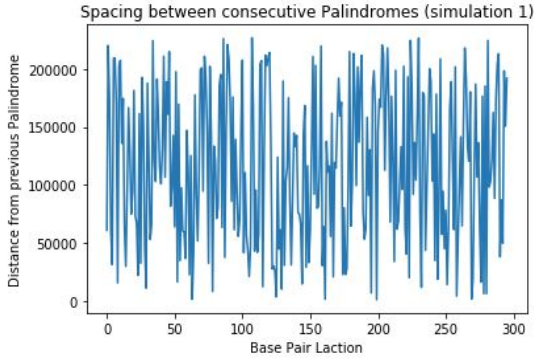


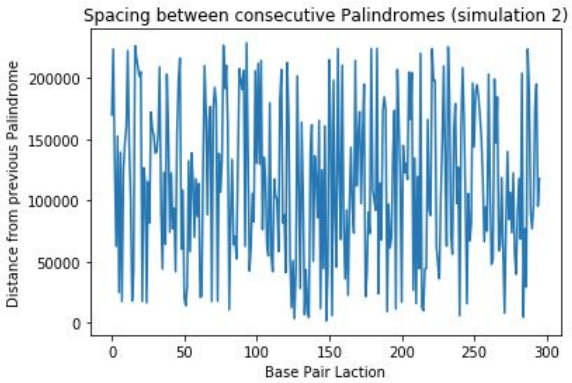


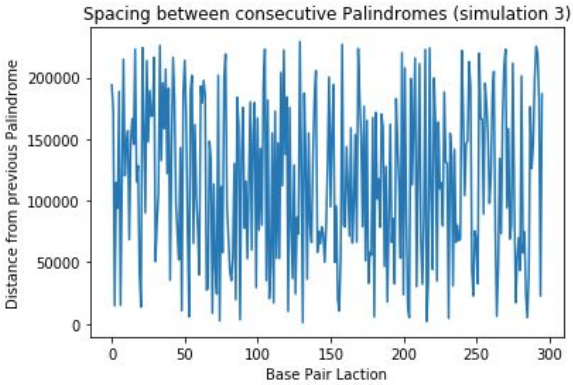


We then decided to explore the spacing between consecutive palindromes using a line graph. It is very hard to recognise any trends in the line graphs for both the original data as well as the simulations. We will need to further investigate this in scenario 2.

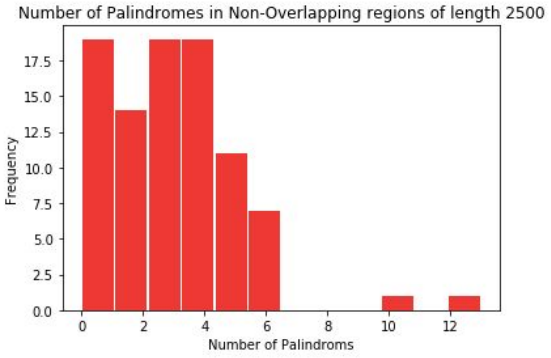


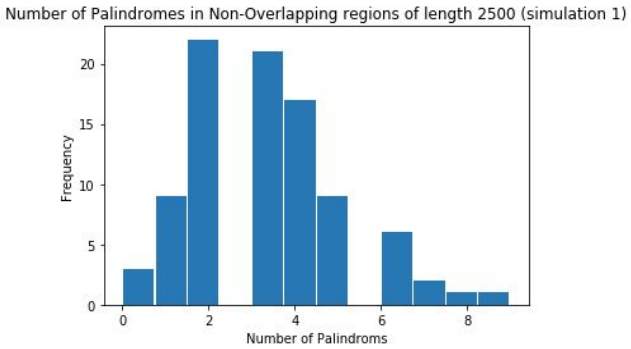


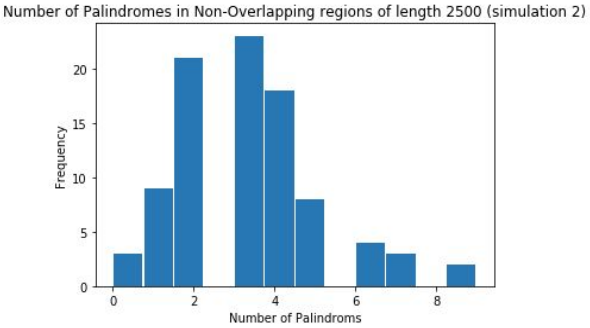


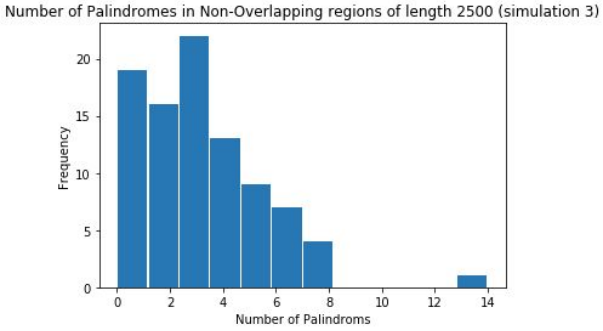


We then investigated the counts of palindromes in non-overlapping regions of the DNA. In the histogram below we plot the number of palindromes in intervals of 2500 base pairs. The original seems to have a high frequency in the left half of the graph and outliers in the right tail. These outliers are caused by the large clusters around the 90000th and 190000th base pairs in the original data.





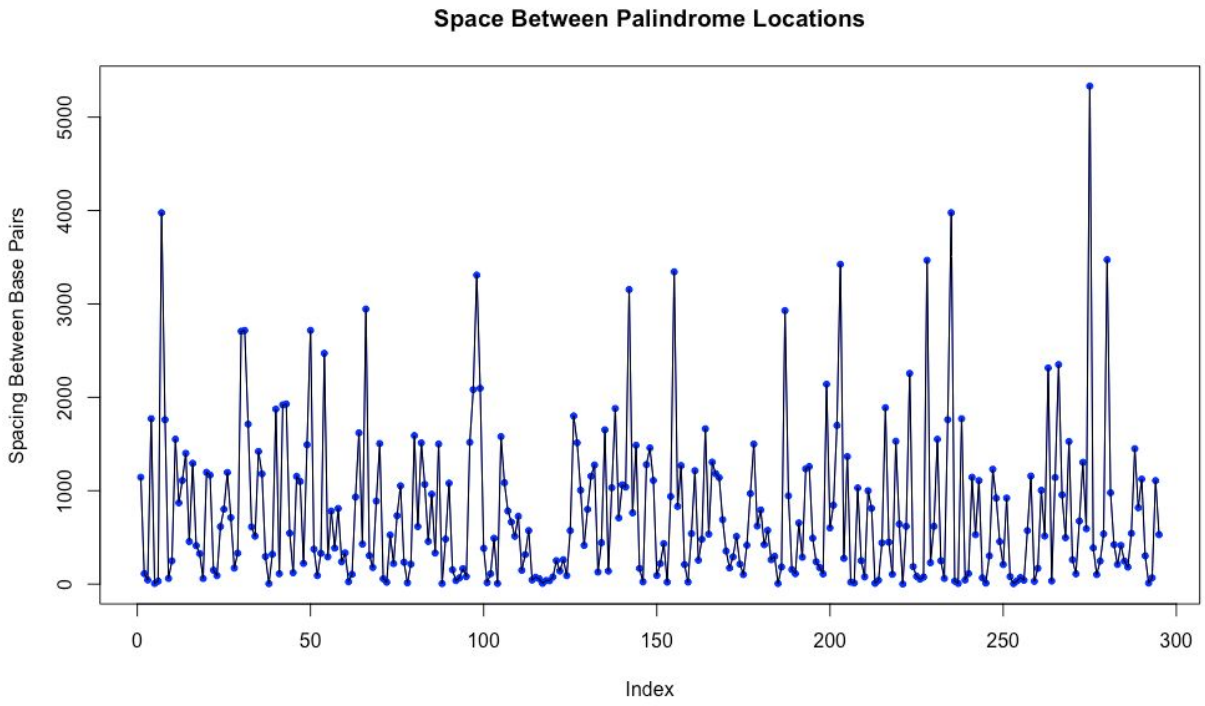




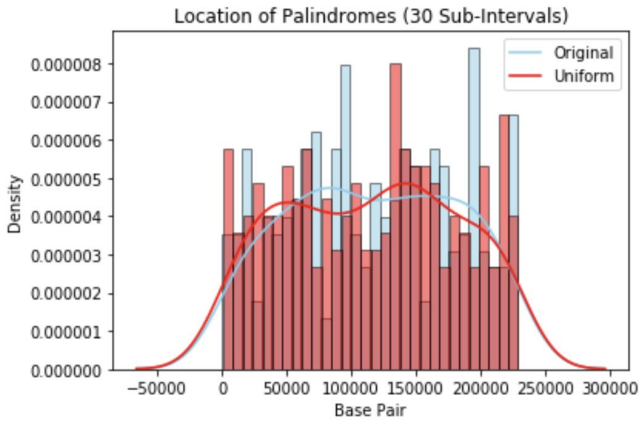
After all the above investigation, we can conclude the presence of clusters in our original data. We can see that the site of replication of the DNA is clearly around the 90000th base pair. We will investigate these claims further using graphical analysis and statistical tests.

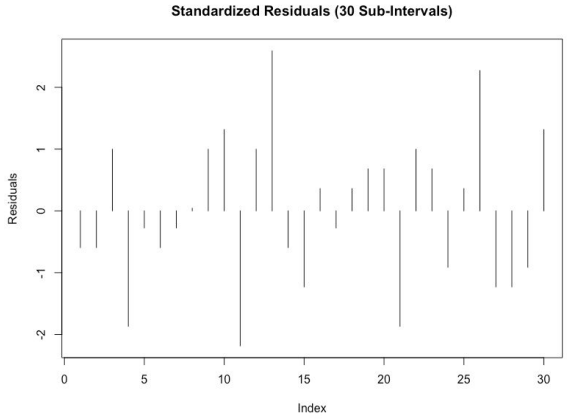
[Scenario 2]

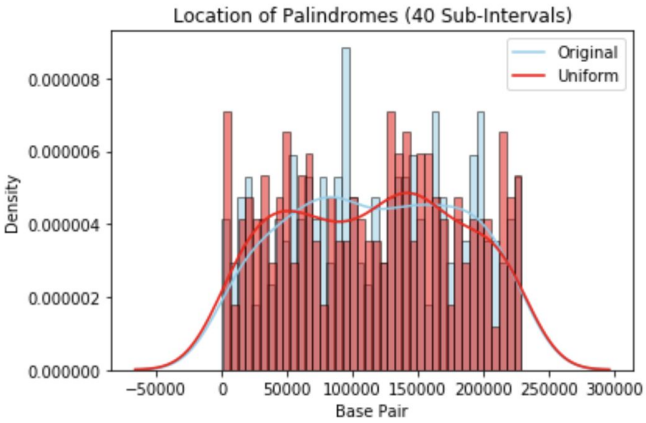
As was mentioned in the introduction to the data, putting length aside, the palindromes appear to cluster in two locations: the 93,000th and 195,000th pairs of DNA. By generating several uniformly distributed data sets, our goal is to compare the average amount of consecutive palindromes within the simulation to that of the given dataset. Then, we will be able to see whether more palindromes are concentrated in one over the other, thus giving us insight into the spacing of the palindromes. Plotted below for reference is the spacing between consecutive palindromes in the data set:

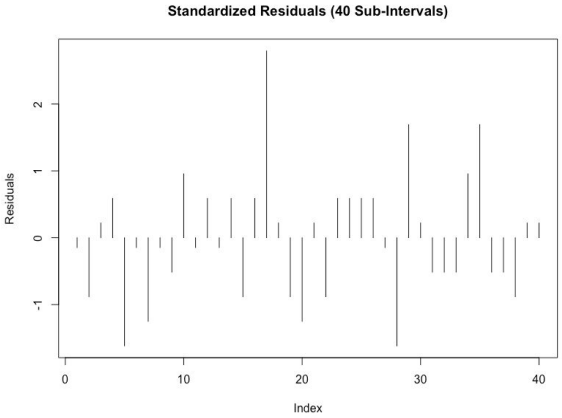


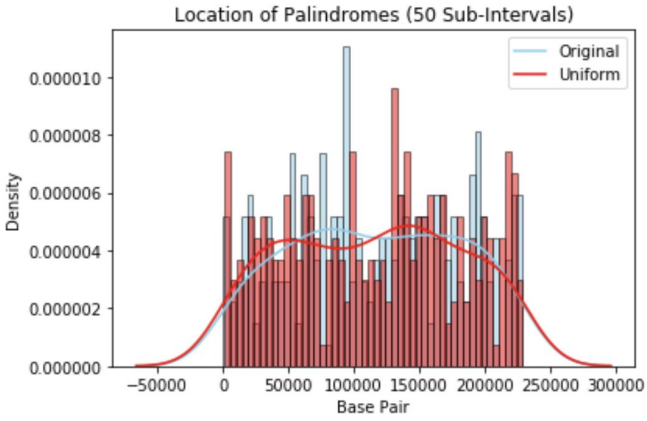
Next, we will divide the data set into three different sub-intervals of size 30, 40, and 50 in order to get an idea of how the distribution of the locations of palindromes (within each interval) fits the uniform distribution. These intervals will ensure that the expected value for each bin is at least five. This will help us identify any abnormalities in the location of the base pairs from the original data set. Also, using these graphical methods in conjunction with standardized residuals will help us analyze the difference between the two distributions.

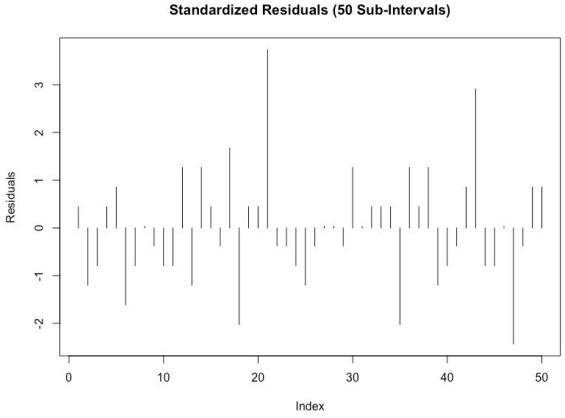






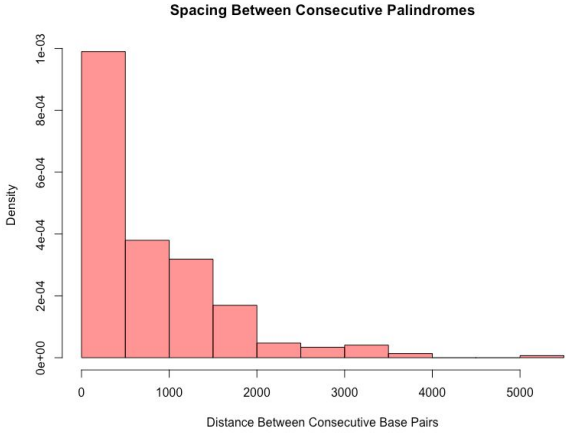


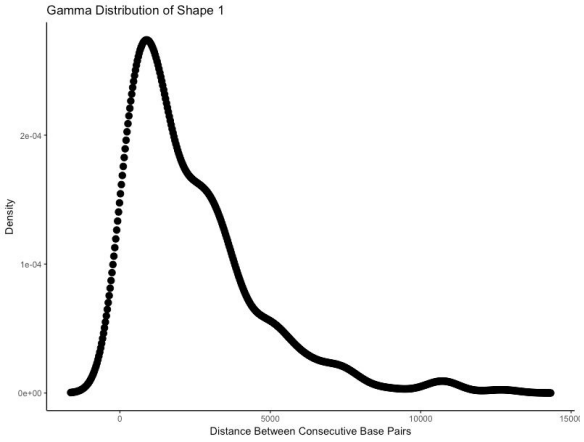


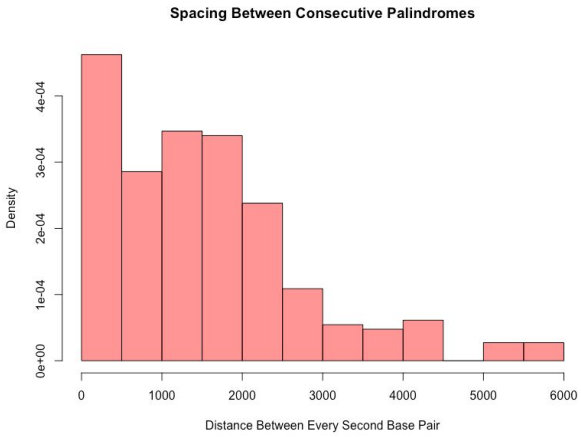


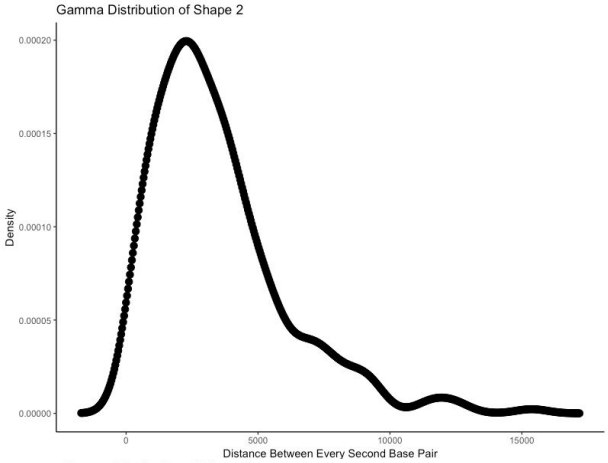
Looking at the residual plots, it seems as though the uniform distribution may not be the best model from which to model the locations of our palindrome clusters. Below, we have calculated the distances between palindromes with different lag values (*1, 2* and *3*) with the goal of identifying any unique patterns amongst the consecutive, double, and triple pairings of palindromes. It appears as though the first plot somewhat follows an exponential distribution while the second two follow a gamma distribution to a degree ($k = 2$ and $k = 3$ respectively). We can see a variance between the two distributions for each shape. This leads us to the conclusion

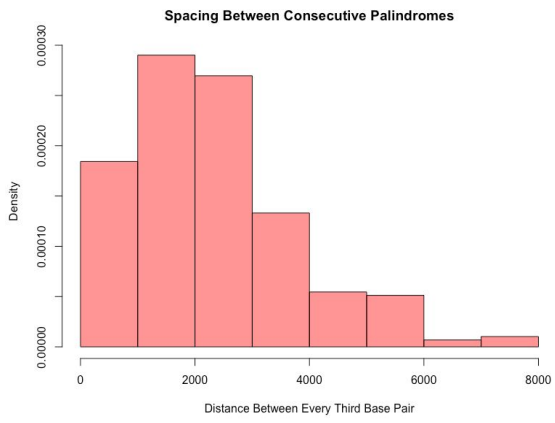
that there is in fact a cluster of palindromes present in our data set based on the difference in spacing between palindromes.

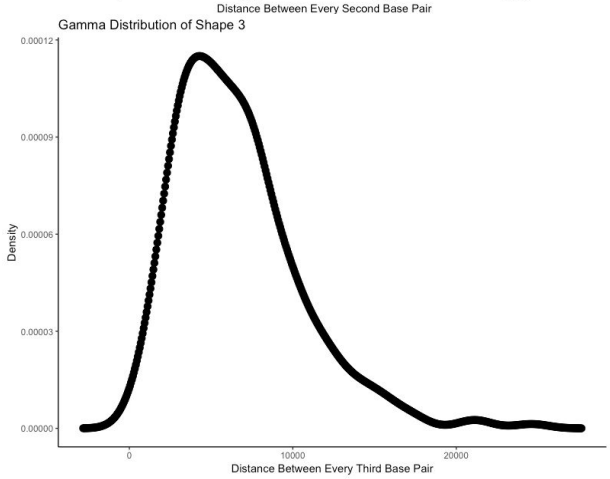












[Scenario 3]

By analyzing the distribution of counts of palindromes with varying interval sizes, we can assess how unlikely it is to see various clusters. We can determine how fitting a homogenous Poisson model would be for our data using random scatter simulations and Chi-Squared Tests to assess the fit of our models.

For the tables featured below, observed counts were generated by first setting a number of "bins" - for example, 82. Then, the DNA data was divided into that number of non-overlapping equal parts. Then, the "observed counts" columns were generated by counting the number of palindromes within those intervals and summing total observed counts of each palindrome sum. Expected counts were generated using the Homogenous Poisson model, in order to gauge the validity of this model as a general fit of the data. The following tables are for bin sizes of 10, 41, 82, 400, 1000, and 5000 and are indexed by the counts of palindromes. We used a wide range of sizes in order to see patterns in how the fit changes as the interval size changes.

Observed and Expected Counts for Size = 10

	21	27	28	29	30	31	32	34
Observed Count	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	3.000000	1.000000
Expected Count	0.215608	0.680419	0.719301	0.734183	0.724394	0.691679	0.639803	0.499617

Observed and Expected Counts for Size = 41

	2	3	4	5	6	7	8	9	10	11	14	19
Observed Count	1.000000	1.000000	4.000000	5.000000	5.000000	9.000000	8.000000	1.00000	4.000000	1.000000	1.000000	1.000000
Expected Count	0.782303	1.882617	3.397893	4.906226	5.903427	6.088552	5.494547	4.40755	3.182036	2.088431	0.359825	0.005058

Observed and Expected Counts for Size = 82

	0	1	2	3	4	5	6	7	9	13
Observed Count	1.000000	9.0000	12.00000	21.000000	16.000000	14.000000	5.000000	2.000000	1.000000	1.000000
Expected Count	2.218793	8.0093	14.45581	17.393982	15.697008	11.332474	6.817911	3.515857	0.636289	0.006296

Observed and Expected Counts for Size = 200

	0	1	2	3	4	5	7	8
Observed Count	43.000000	83.000000	33.000000	27.000000	9.000000	3.000000	1.0000	1.000000
Expected Count	45.527538	67.380756	49.861759	24.598468	9.101433	2.694024	0.1405	0.025992

Observed and Expected Counts for Size = 1000

	0	1	2	3	4	5
Observed Count	759.000000	200.000000	32.00000	6.000000	1.000000	2.000000
Expected Count	743.787428	220.161079	32.58384	3.214939	0.237905	0.014084

Observed and Expected Counts for Size = 5000

	0	1	2	3
Observed Count	4729.000000	254.000000	16.000000	1.00000
Expected Count	4719.193482	272.769383	7.883035	0.15188

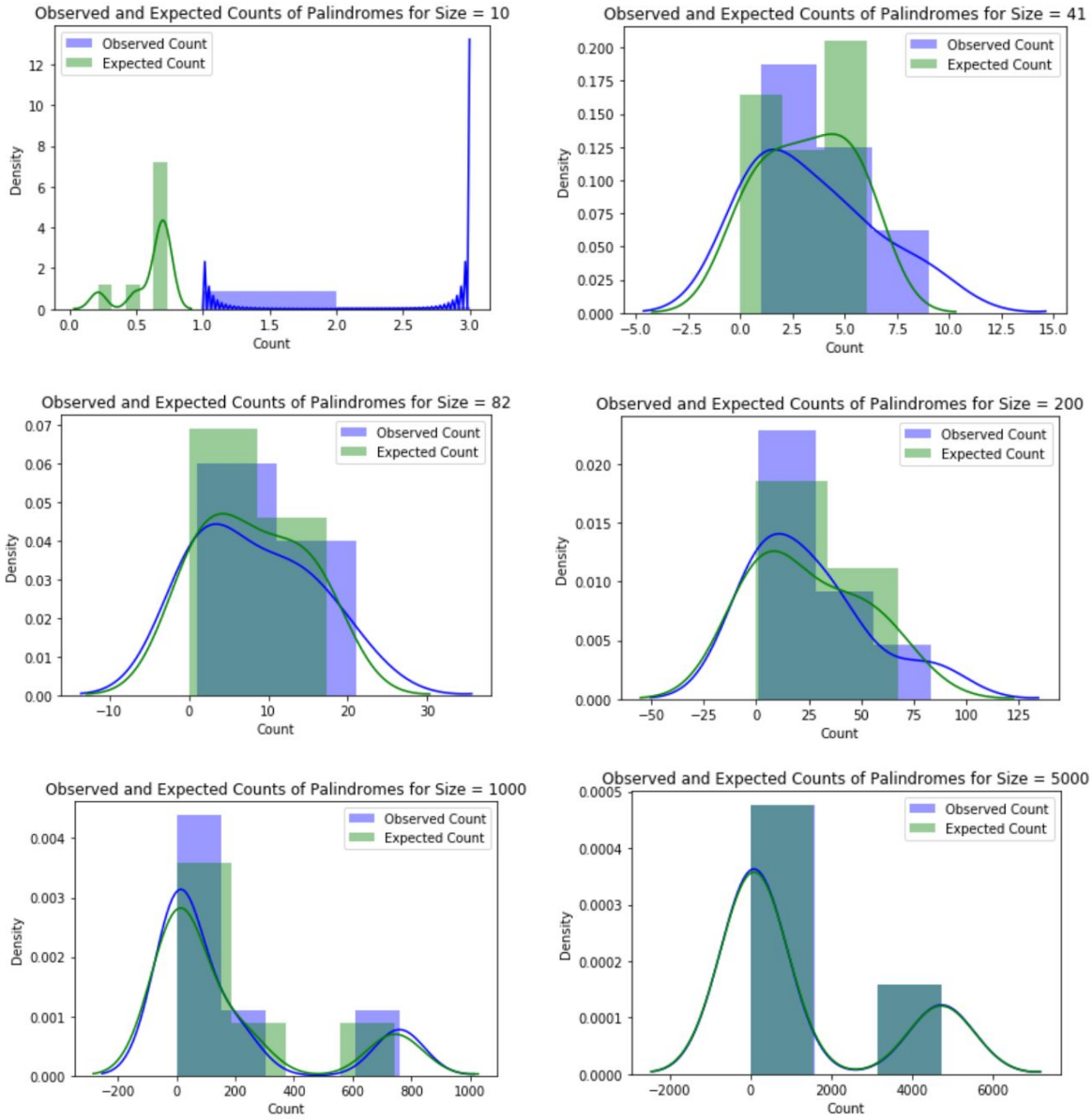
The resulting Chi-Squared Statistics are displayed below to clarify our findings:

	Chi-Squared Statistic
10	12.659609
41	203.534134
82	160.779964
200	51.490075
1000	287.045982
5000	41.400234

In theory, we would prefer not to have expected counts of < 5. Those counts in particular seem to be causing strange behavior in the Chi-Squared data, as dividing by very low expected counts will result in dubious high Chi-Squared statistics. Disregarding expected counts less than 0.5 is not a perfect solution but does demonstrate the results of this phenomenon and result in much more reasonable Chi-Squared statistics:

	Chi-Squared Statistic
10	9.304808
41	6.667737
82	3.936749
200	9.733436
1000	4.580500
5000	10.561756

Finally, we can look at overlaid histograms of expected counts versus observed counts to assess how unusual certain clusters appear:



Clearly, when performing analysis we can see that as interval sizes become larger, the data more closely reflects a random scatter model.

[Scenario 4]

The Poisson process is a good reference model for making comparisons because it is a natural model for uniform random scatter (slide 29). This allows us to use maximum likelihood estimation to calculate the value of lambda. Our interval length was calculated by dividing the range of the locations by the amount of intervals being used. We calculated the probability that the maximum count of the current interval was larger than the maximum count of other intervals. If we set a threshold for the probabilities to be 0.05, we can see that only for indices 20,30 and

40 have probabilities below the threshold. That corresponds to the fact that the clusters are not unusual and the replication site for the DNA is unknown, since the interval length was too large. That's why as we decrease the interval length, our probability decreases. After index 60 our probability becomes too low and again it becomes unlikely for a cluster to exist in that region as the interval length is too small. The maximum counts was taken for all the intervals of the interval length and we also got the corresponding interval that coincided with the max count. We can see that all the intervals are around the 90000th base pair location, which confirms what we suspected initially. This confirms that this is the area of clustering and is the replication site.

	Lambda	Interval Length	Probability	Interval	Max Count
20	14.800000	11467.700000	0.443264	(91733.6, 103200.3]	23
30	9.866667	7645.133333	0.317525	(91729.6, 99373.733]	18
40	7.400000	5733.850000	0.308448	(91725.6, 97458.45]	15
50	5.920000	4587.080000	0.059746	(91721.6, 96307.68]	15
60	4.933333	3822.566667	0.036209	(91717.6, 95539.167]	14
70	4.228571	3276.485714	0.009397	(91713.6, 94989.086]	14
80	3.700000	2866.925000	0.002694	(91709.6, 94575.525]	14
90	3.288889	2548.377778	0.000850	(91705.6, 94252.978]	14
100	2.960000	2293.540000	0.001406	(91701.6, 93994.14]	13

[Theory]

Basic Principle:

For the table on the Basic Principle (cont) slide in the lecture slides, the last 2 rows in the table are incorrect since the values are less than 5. From our discussion during lecture, 5 is an arbitrary value/threshold. In addition, values are expected to be greater than 5. Here is the process behind this:

$$\begin{aligned} Expected &= 57 * P(Y = 0 \text{ or } Y = 1 \text{ or } Y = 2) \\ &= 57 * (P(Y=0) + P(Y=1) + P(Y=2)) \\ (Null\ Hypothesis) \rightarrow &= 57 * \left(\left(\frac{e^{-\lambda} * \lambda^0}{0^1} \right) + \left(\frac{e^{-\lambda} * \lambda^1}{1^1} \right) + \left(\frac{e^{-\lambda} * \lambda^2}{2^1} \right) \right) \\ &= 57 * e^{-\lambda}(1 + \lambda + \lambda^2) \end{aligned}$$

Degrees of Freedom:

The reason that the degrees of freedom was 6 in the slides is because of the following:

- 8 = number of rows
- -1 = theory behind degrees of freedom
- -1 = because we're approximating lambda-hat
- Therefore, degrees of freedom = 8-1-1 = 6

We can use this reasoning of degrees of freedom to determine the degrees of freedom when performing chi-square analysis.

Test Statistic/Maximum Number of Hits:

$$Y_i = \text{number of palindromes in interval } i$$

$$P(\max\{Y_1, Y_2, Y_3, \dots, Y_{57}\} \geq k), \text{ where } \max\{Y_1, Y_2, Y_3, \dots, Y_{57}\} \text{ is the test statistic}$$

$$= 1 - P(\max\{Y_1, Y_2, Y_3, \dots, Y_{57}\} < k)$$

$$= 1 - P(\text{all } Y_i\text{'s are } < k)$$

$$= 1 - (P(Y_1 < k) * P(Y_2 < k) * P(Y_3 < k) * \dots * P(Y_{57} < k))$$

$$= 1 - [P(Y_1 < k)]^{57}$$

$$\text{Since we know } Y_1 \sim P(\lambda):$$

$$= 1 - [P(Y_1 = 0) + P(Y_1 = 1) + \dots + P(Y_1 = k-1)]^{57}$$

The test statistic is IID (independent and identically distributed) since:

- Across intervals, we have independence in data
- Data is identically distributed across intervals

Poisson Process:

The Homogeneous Poisson Process is a model for random phenomena, which in this case is the locations of palindromes in DNA strands. This process comes about naturally from the idea that points are distributed on a plot with irregularity.

In this case, the locations are in a uniform discrete distribution and can be found using a Q-Q plot or chi-square calculation. In addition, the spacings are in an exponential distribution and can be found using a Q-Q plot or chi-square calculation. Furthermore, the number of palindromes per interval is in a Poisson distribution and can be found using a chi-square calculation, which is optimal. We are testing for the maximal cluster. All of these items are in the data. The control group is no clusters, which is "unusual".

The specific features that are unique to the Poisson Process are the following: the underlying rate λ at which points, called hits, occur and is such that id does not change with location, which is known as homogeneity, the count of points in separate regions are independent, and each point has to land in a unique place. This process is a natural model for the uniform random scatter, so

it is a good model for making comparisons. Using the Poisson distribution to determine the number of hits, this method of estimation is known as method of moments. Another method is known as the maximum likelihood method.

In addition, the null hypothesis is that $Y_1, Y_2, \ldots, Y_{57}$ follows P(λ*4000), where Y_1 is the number of times that 0 occurred in $Y_1, \ldots, Y_{57}$, Y_2 is the number of times that 1 occurred in $Y_1, \ldots, Y_{57}$, and so on. For example, P(Y_1 =7) = $\frac{e^{-\lambda} * \lambda^7}{7!}$

So, when we determine the number of palindromes in a certain range, if we make the range of locations in groups of 4,000, we see that, for example:
Y_1 is the number of locations from positions (0, 4,000) ~ P(λ*4,000), where ~ describes the null hypothesis and λ*4,000 is λ_1.
Y_2 is the number of locations from positions (4,000, 8,000) ~ P(λ*4,000), where ~ describes the null hypothesis.
Extrapolating this pattern out, we see that Y_N is the number of locations from the last 4,000 positions ~ P(λ*4,000), where ~ describes the null hypothesis.

Building off of this, when talking about large clusters, the null hypothesis is that there is not an unusually large cluster. We recognize that the data follows the Poisson Process.
We define P(λ) to be the probability of a match within a certain range of positions. From there, we define the max i.i.d. P(λ) to be the most number of matches that are contained within a certain specified range. Finally, we define an unusually large cluster to be a cluster that has more matches than the max i.i.d. P(λ), which can be determined by comparing the number of matches from the unusually large cluster with that of the max i.i.d. P(λ).

Chi-Square:
The formula for chi-square measures discrepancy between the observed sample counts and expected counts. The formula for chi-square is the following:

$$\chi^2 = \sum \frac{(O-E)^2}{E}, \text{ where } O \text{ is an observed value and } E \text{ is the expected value}$$

If the random scatter model is true, then the test statistic has an approximate chi-square distribution with some number of degrees of freedom. For the example in the lecture slides, we used 6 degrees of freedom, and the reasoning behind this is shown under the Degrees of Freedom portion of the theory.

In general, for a chi-square analysis, to create a hypothesis test for a discrete distribution, we create a distribution table from the data, where b represents the number of categories or values for the response and Nt represents the number of observations that appear in category t, where t = 1, . . . , b. These counts are then compared to the expectation under the null hypothesis, or in other words, under the assumption that the data does follow the poisson distribution.

Assuming that the data are generated from the hypothesized distribution, we can calculate the chance that the test statistics would be as large as, or larger than, the observed values. We call this change the p-value, or observed significance level.

To compute p-value, we use the chi-square distribution. If the probability model is correct, then the test statistic has an approximate chi-squared distribution with m − k − 1 degrees of freedom, where m is the number of categories and k is the number of parameters estimated to obtain the expected counts. The way to interpret p-value is if the p-value is too small, there is something wrong with the fit of the distribution.

χ^2_{m-k-1} is a continuous distribution on the positive real line, and the density has a long right tail. As the degrees of freedom increase, the chi-square distribution starts to look symmetric and a lot like a normal distribution.

Mean Squared Error (MSE):

As we learned from lecture, the formula to calculate mean squared error is the following:

$$\text{MSE}(\hat{\lambda}) = \mathbb{E}(\hat{\lambda} - \lambda)^2 = \text{Var}(\hat{\lambda})\text{variance} + \left[\mathbb{E}(\hat{\lambda}) - \lambda\right]^2 \text{squared BIAS}$$

We use MSE to compare and evaluate parameter estimates. The theorem behind MSE is that under regularity conditions, as the sample size increases, the Maximum-likelihood estimator (lambda-hat) satisfies the following:

$$\hat{\lambda} \to \lambda$$

$$\hat{\lambda} \sim \mathcal{N}\left(\lambda, \frac{1}{nI(\lambda)}\right)$$

From this condition, I(lambda) is called the Fisher's Information Matrix.

Type I and Type II Errors:

Essentially, a Type I error is alpha, or the test significance. In addition, a Type II error is beta, where the power of a test is 1 - beta.

Conclusion

From **[Scenario 1]** we saw, using the Monte Carlo simulation, that there appears to be outliers in the middle portion of the original data, possibly corresponding to the cluster discussed earlier around the 90000th base pair. This data is also far from uniform which leads us to believe that there is a need for further testing. This will help us determine whether or not there is a potential replication site in the DNA.

For [**Scenario 2]** and [**Scenario 3]**, we used both graphical methods and goodness of fit statistical testing for the purpose of comparing the actual data against the simulated uniform data. Using standardized residuals (on both the locations and spacing of the palindromes), we saw that the uniform distribution may have not been the ideal model to compare our data with. This implied that there may have been large clusters of palindromes in the data set that messed with the uniform distribution. We also saw, when taking the difference in locations of consecutive, every second, and every third palindrome there was a deviation from both the corresponding exponential and gamma distributions.

From **[Scenario 4]**, we divided the total length into different interval lengths and calculated the probability of the chance that the maximum count over *m* intervals is greater than or equal to the maximum counts generated from a different length of intervals. The maximum counts was taken for all the intervals of the interval length and we also got the corresponding interval that coincided with the max count. The intervals of the clusters calculated from the nine different amounts of intervals all include the interval between 91740 and 93900, so this is interval with the greatest amount of palindromes. This also indicates a potential origin of replication. Therefore, biologists should focus on the interval between 91740 and 93900 to find the origin of replication efficiently.

Each of the aforementioned scenarios have shown that our data does not closely follow the uniform distribution. We would expect the data set to follow this distribution if there was no suspicion of replication sites, however our visualizations lead us to believe there are clusters of palindromes that exist, and thus potential replication sites.

Works Cited

N/A