Anthony Ayon
Ooduah Orafidiya

Case Study 3

Introduction

In this case study we will address the cytomegalovirus virus (CMV), which is a potentially life-threatening disease for people with weak or deficient immune systems. First off, a virus's DNA contains the necessary information for it to develop and replicate. In order to combat this disease, scientist must find the exact location within the DNA that has the instructions of the virus's replicability. The DNA of a virus is a long-coded message composed of four letters from the alphabet (A, C, G, T) which are used in a sequence with multiple patterns. In this type of DNA, the letters compliment one another such as A to T and G to C. Hence, "GGGCATGCCC" is a type of pattern named complementary palindrome, which is a sequence of those letters that can be read in reverse as the complement of the forward sequence. By flagging these types of patterns, scientist may be able to localize the replication process and develop treatments that could suppress these viruses.

Finding complimentary palindromes is important because the origin of replication was found in two viruses of the same CMV family. One of them was Herpes simplex, which had a 144-letter palindrome and the other Epstein-Barr virus, which had shorter palindromes but, constant clustered repeats. Even though, CMV has 296 palindromes between 10 and 18 base pairs, scientist believe these clusters could be indicators of the replication origin. Nevertheless, scientist must cut DNA segments and test each one in order to find the origin of replication, which can be very expensive and time consuming. One approach is to search for unusual clusters within complementary palindromes which will help narrow down the number of segments being tested.

Data

Our data is a 229,354-letter long DNA sequence of CMV published by Chee et al. in 1990. In 1991, Leung et al. used algorithms to project all types of patterns found in the data. A total of 296 palindromes where found ranging from 10 to 18 letters while palindromes shorter than 10 letters where ignored. Therefore, our data will just be a list of locations of these 296 palindromes.

Results

Scenario 1: Introduction to Location, Space & Count

In this scenarion we will explore the structure of our data with emphasis on the location, clustering and spacing of our palindromes. Two uniform randomly generated scatters will be used to compare with the original data in order to prevent potential complementary palindromes

being mistaken as chance occurrences. These random scatters will also consist of 296 palindromes within a sequence of 229,354 base pairs.

Location

In order to visualize any potential palindrome clusters, we plotted the "location" data into a histogram (plotted in red) and strip-plot. At first look we can see that the data has some high and low concentrations of palindromes. These high concentrations can be seen around the 90000th and 200000th base pairs, hence indicating potential clusters in those sections.
Additionally, we see a similar but, less conclusive clusters in sample 1 around the 30000th and 130000th base pair. In sample 2 clusters are a little less evident as base pairs seem a little more uniformly distributed. Therefore, there is still not enough evidence that would reveal weather a cluster is a chance occurrence or replication site.

Figure 1: Histogram Plots of Palindrome Locations

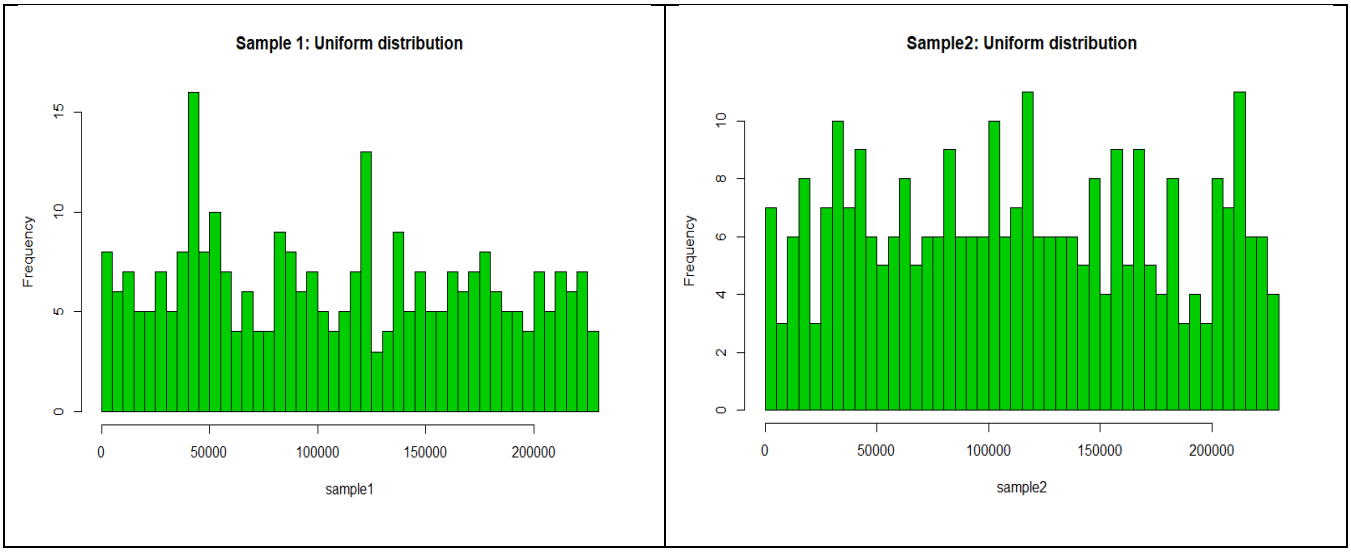
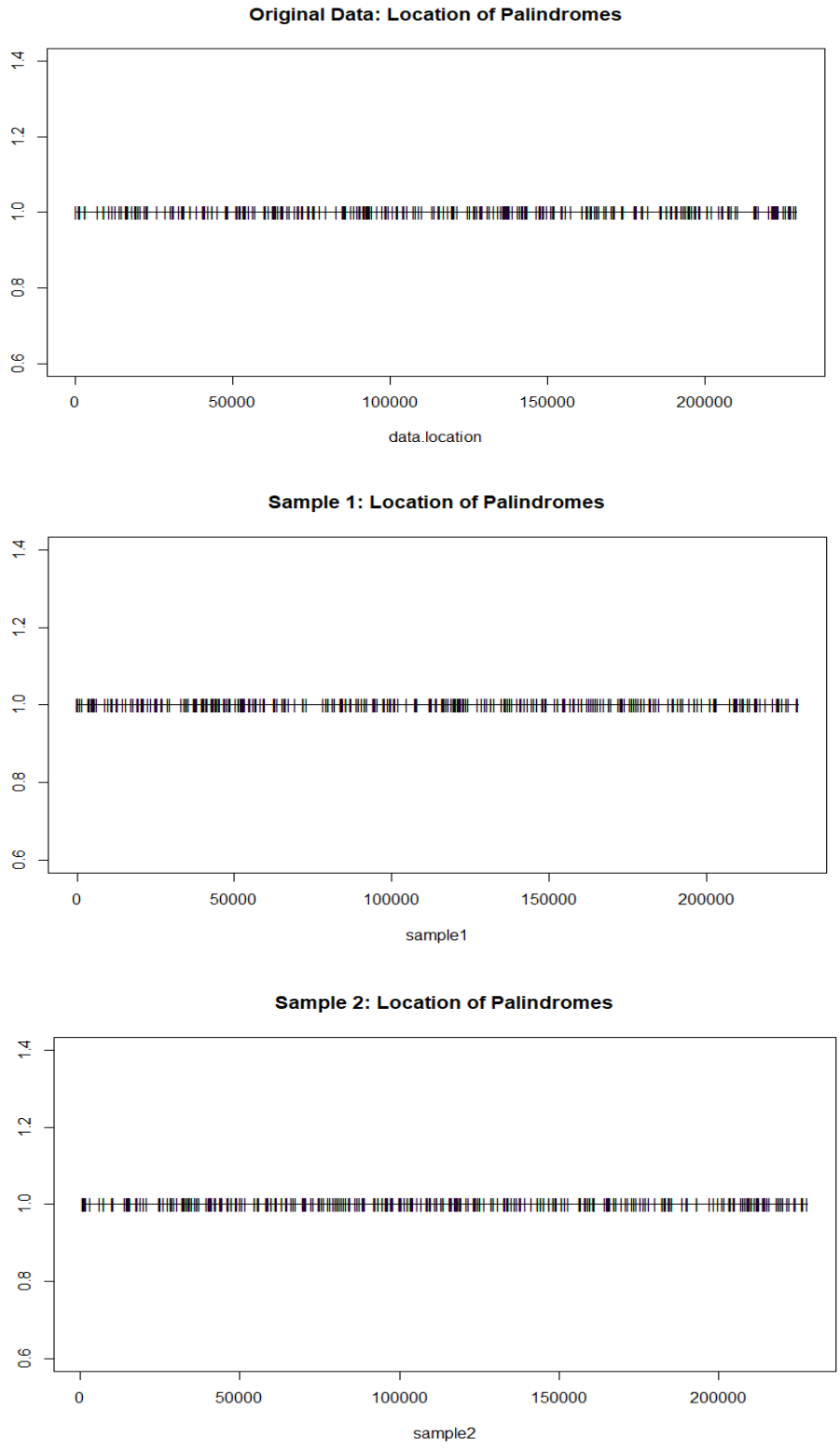


Figure 2: Strip-plot of Palindrome Location

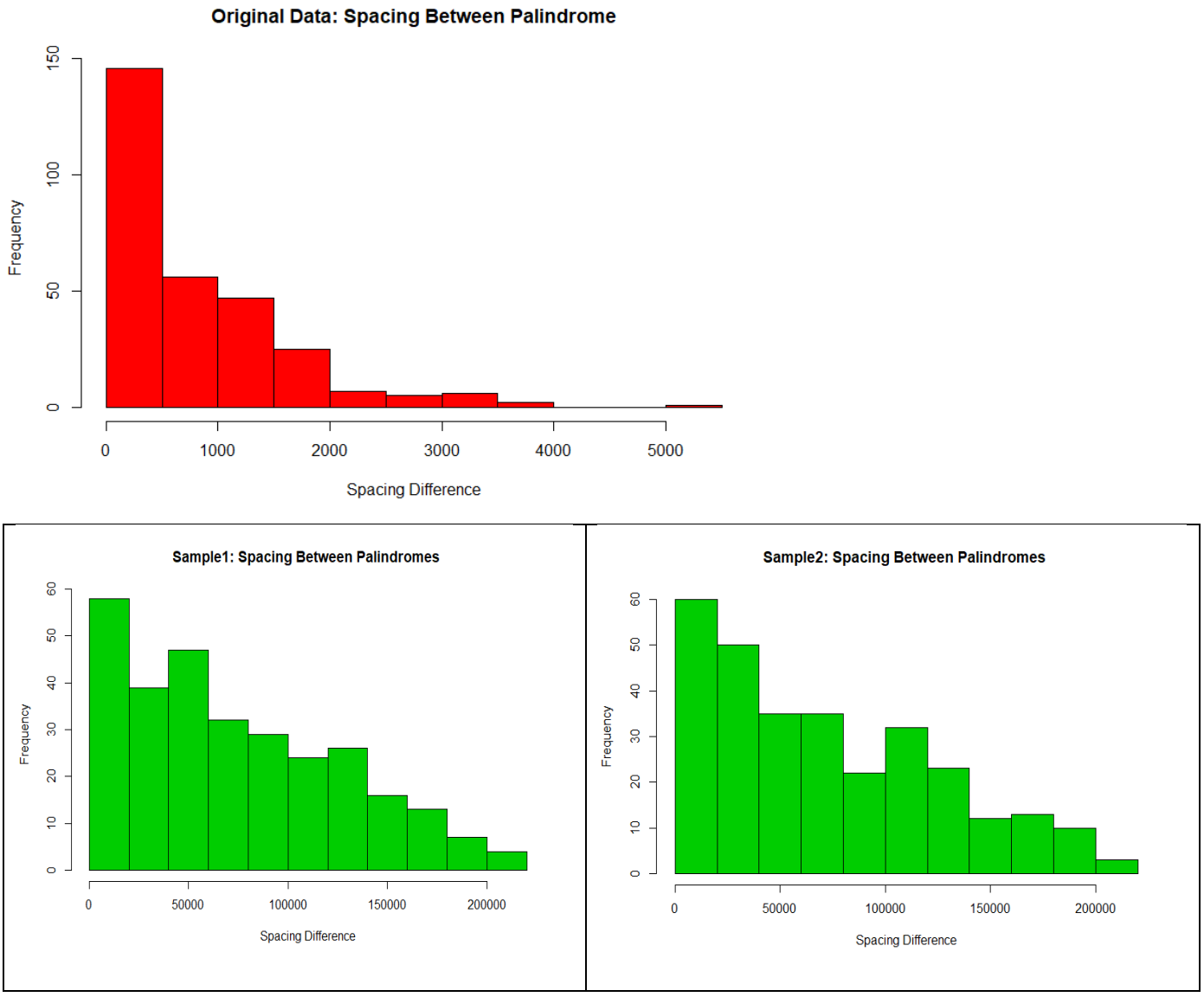


Spacing

We move on to check the spacing between consecutive palindromes in both our data and samples. This operation consisted of taking each value in our datasets and subtracting it by the value next to it, or in other words, taking the difference between palindromes. After the subtraction, the absolute value was taken thus, eliminating any negative values.

All calculation where taken and put in a histogram illustrated in *Figure 3*. After a close look it is particularly interesting the value differences between the original data and samples. In the original data differences mainly ranged below 2000, with most values having a difference of 500 or less. Whereas in both sample histograms value differences are a lot larger and evenly distributed, which makes sense considering they come from uniform distributions. This supports the evidence that our original data is in fact unique and not another random occurrence.

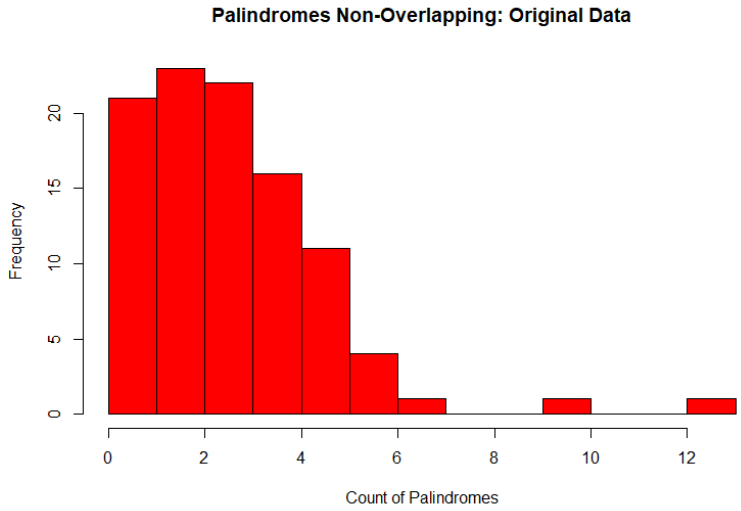
Figure 3: Histogram of Palindrome Spacing

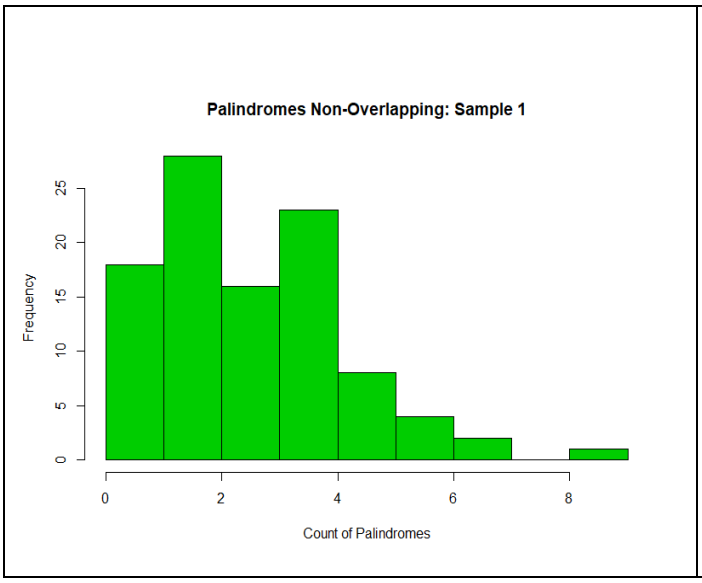


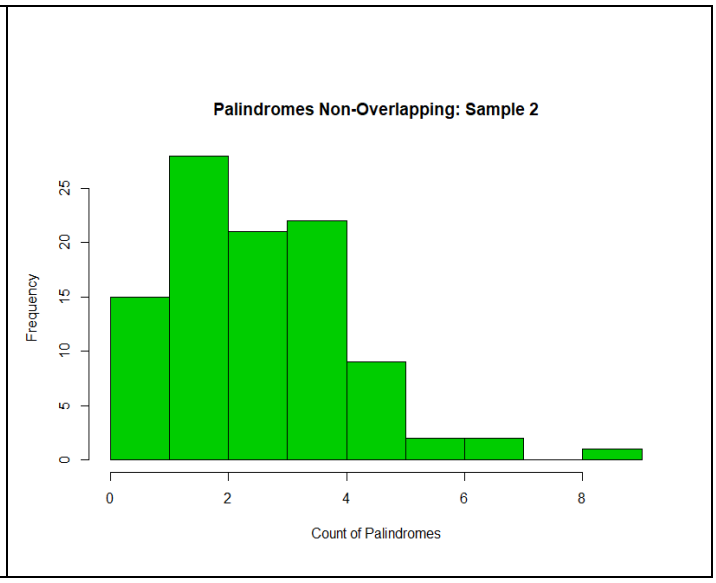
Count

In this portion we plotted the count of non-overlapping palindromes of both the location data and samples. We divided the CMV DNA into 100 non-overlapping regions with the 2293 bases each. In this case, both the location and sample histograms maintain a s general structure: skewed left. On the other hand, the original data does present steady decrease while the samples retain more of its uniform structure. Also, the original data does have a clear view of outliers which may be attributed to the large palindrome concentrations illustrated in *Figure 1*. Overall, we do see a clear distinction between the histogram structure of our samples and location data, furthering implying that these may be potential replicating origins.

Figure 4: Histogram of Non-overlapping Palindrome Spacing







Scenario 2: Location and Spacing Analyzes

In this section we will go more in depth for analyzing both location and spacing in our given dataset. First, we will analyze the location of the data while making comparisons with the random uniform samples provided above. Even though in *figure 1 and 2* we got a glimpse of the location of each palindrome and its clusters, now we will be analyzing the main differences between the given and random datasets. To do this we used QQ plots, Chi-squared goodness fit test and residual plots.

Location

To examine location, we divided our data into 25, 30, 45 and 57 sub-intervals where each would get 9174, 7645, 5097 and 4024 base pairs respectively. A QQ-plot was used to compare the location data with both samples. As we can see in *Figure 5 & 6*, the scatter points in both plots seem to somewhat follow a linear path but, there are some discrepancies along the way that may imply they do not come from the same distribution. In the case of sample 1, we see the points derail from the linear line between palindrome 50000 to 150000, which reflects what we saw in *Figure 1* with low and high concentrations of our original data. For sample 2, it's a little less clear as the points follow the linear path, however there is various bumps along the way that seem to indicate a difference in palindrome values. Overall, we do see a general difference between our uniform distributions and location data especially among potential cluster sites.

Figure 5: QQ-Plot: Sample 1

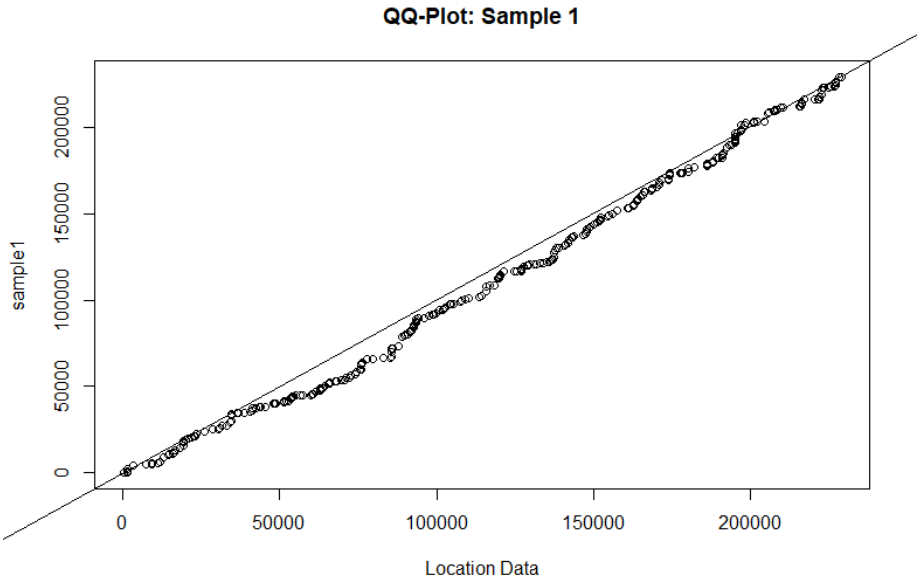
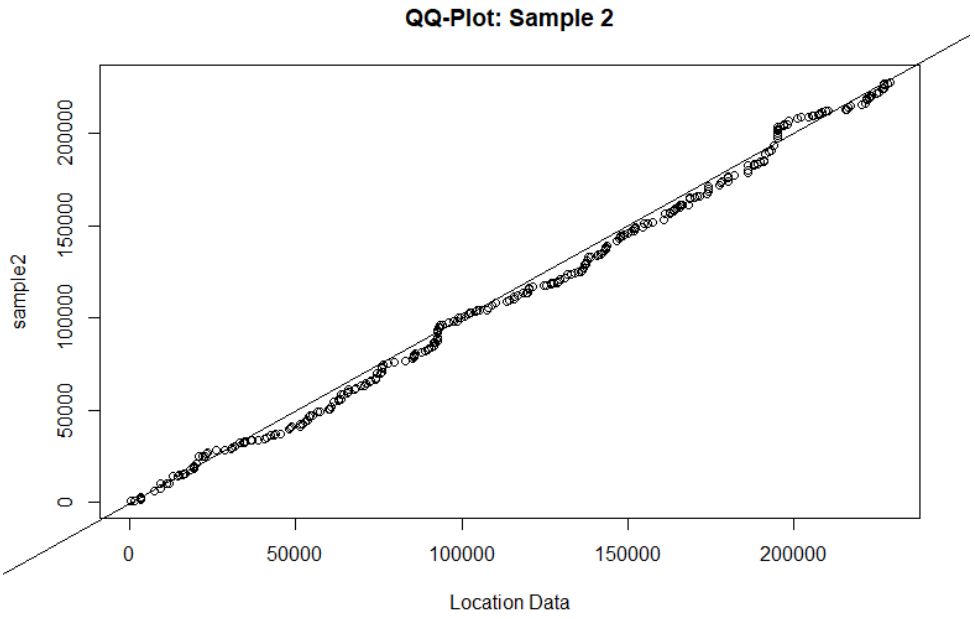


Figure 6: QQ-Plot: Sample 2



Additionally, we integrated another uniform random sample (sample 3) along with residual plots in order to further analyze location. As illustrated in *Figure 7*, the QQ-plot between our new sample and the location data seems fairly similar to our previous plots. Again, the points derail from a straight linear path and the bumps appear within the same range as *sample 1 and 2*.

We also used a residual plot to compare both distributions and see if some of the intervals varied. In *Figure 8* we can observe a vast difference between mean values as you increment the number of intervals for both distributions. Most values surpassing two residuals indicate that there is no general fit between the interval of the original and sample data, thus correlating with what we found in the QQ-plot. Therefore, we can state that the location data does not follow the uniform distribution of sample 3.

Figure 7: QQ-Plot: Sample 3

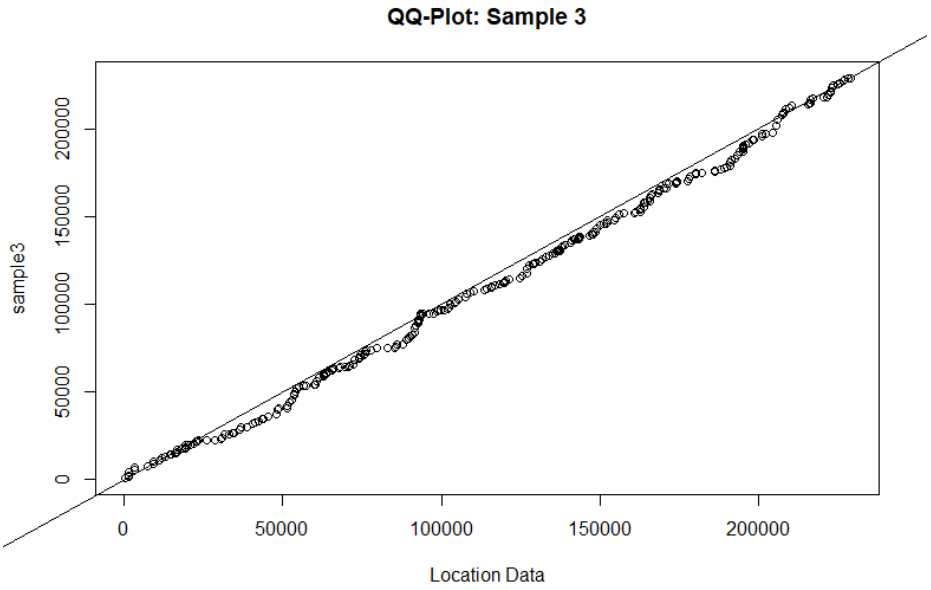
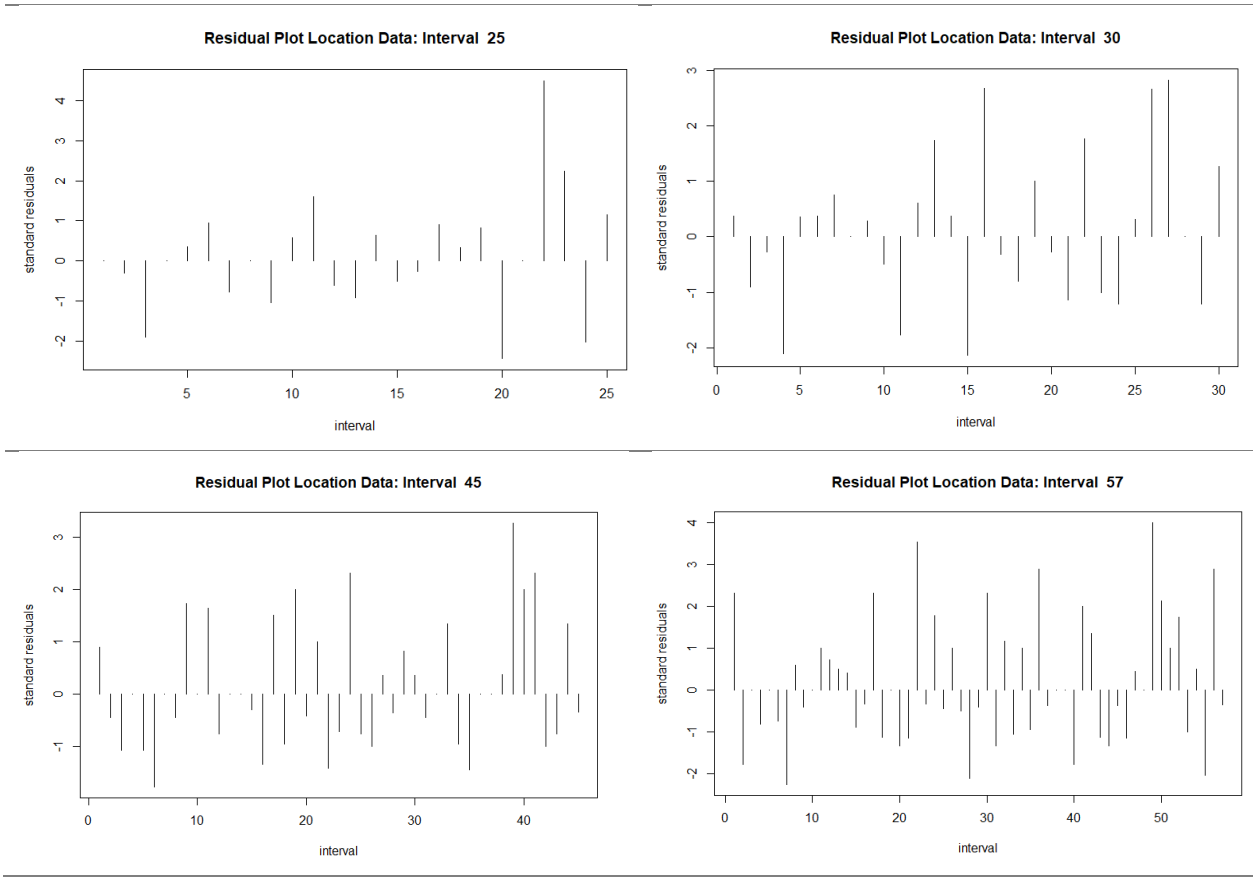


Figure 8: Residual Spacing: Location Data & Sample 3



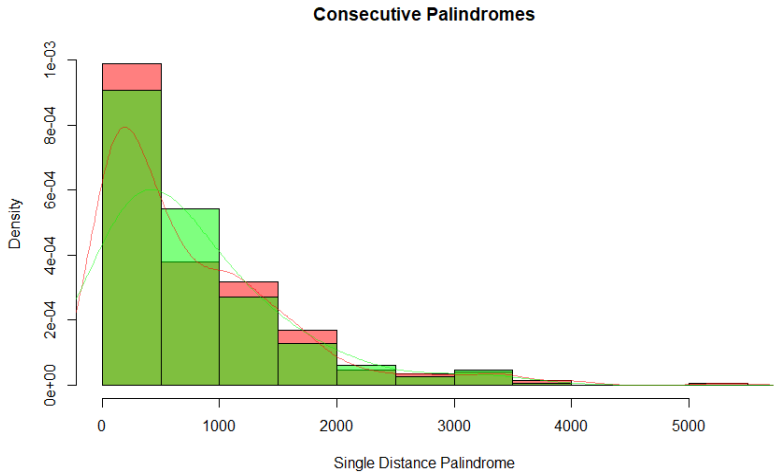
Spacing

H0: Location data follows an exponential/gamma distribution
H1: Does not follow an exponential/gamma distribution

In this section we examined single, pair and triple spacing among palindromes. Single, meaning for each consecutive palindrome the difference was acquired. Pair means we took two consecutive palindromes to acquire the difference, hence taking two values instead of one. Same case for triplets where three consecutive values where taken. For all cases we overlapped palindromes, which means that if we have a list of values [1, 2, 3, 4] for a pair instance we took [1, 2] and [2, 3] to calculate the difference.

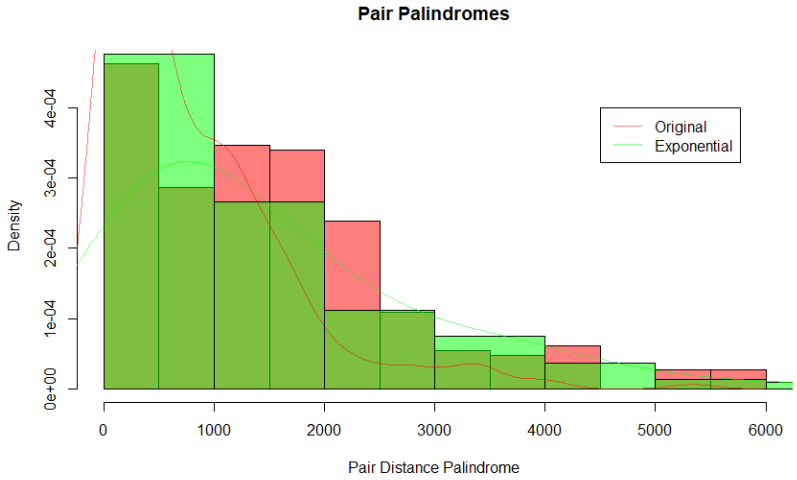
Consecutive:
For consecutive palindromes we plotted a histogram displaying the single-spaced difference(red) and an exponential distribution(green) based on the same interval. As we can see there is reduced space among single palindromes. The original's data curve seems slightly more narrow and higher, but both appear to follow a similar distribution. Further statistical test will be done below to compare these distributions.

Figure 9: Consecutive Palindrome Distribution



Pair:
For the pair distributions, in *figure 10* we see a similar structure to single pair distances. Contrarily, there is a greater visual difference among distributions which may be an indication that the original data does not follow an exponential distribution.

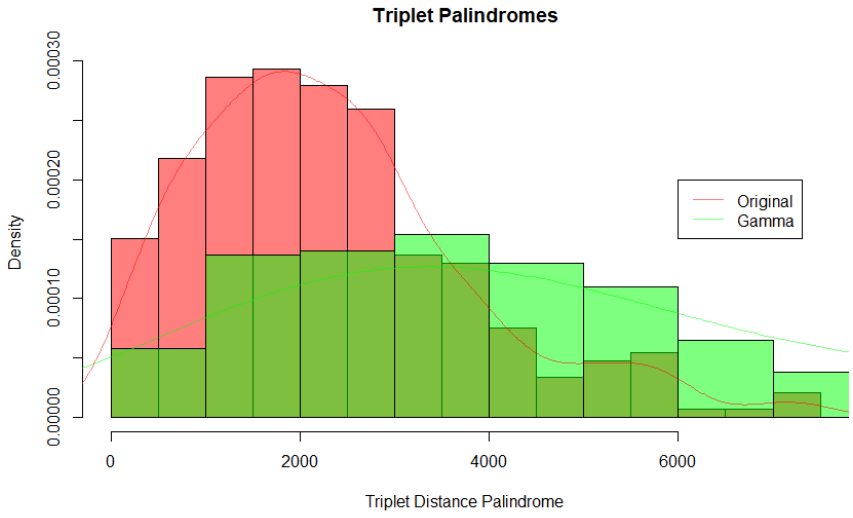
Figure 10: Pair Distance Palindrome Distribution



Triplets:

In this case, there is a clear distinction between the triplet gamma distribution and the original data distribution. Particularly, the triple paired original data is heavily skewed left while the gamma distribution follows a clean shaped curve. In short, we can state that these consecutive triple values do not come from the same distribution.

Figure 11: Triplet Palindrome



Overall, we have compared three different distributions with the location data. A t-test, which is a simple statistical calculation, was used to compare both distributions and recorded in *table 1*. The p-value must be smaller than 0.05 in order to reject the null hypothesis, which is

satisfied in both pair and triplet palindromes but, not for single consecutive ones. This may suggest that either the random exponential distribution correlated by chance or the spacing in our location data follows an exponential distribution. Upon further analysis we decided to compare the t-test result for single space with a Chi-Squared test. The values are displayed in *Table 2*, where we can see a very low p-value of 2.2e-16, which heavily contrast with the t-test. In reflection, we believe the t-test's p-value may have been misled mainly because this test is used primarily for normal distributions. Hence, the Chi-Squared test provides a better statistical approach for the type of distributions present in single spacing. Shifting our focus to the pair and triplet spacing distributions, we see a very small p-value, hence suggesting that both exponential and gamma distributions are not related to the location spacing distributions.

Table 1: t-test for Spacing Distributions

	SINGLE	PAIR	TRIPLET
T – STATISTIC	0.48234	9.6612	-10.239
DF	578	477	424
P-VALUE	0.6297	2.2e-16	2.2e-16

Table 2: Chi-Squared Test for Single Spacing

	X-Squared	***DF***	***p-Value***
SINGLE	1169700	294	2.2e-16

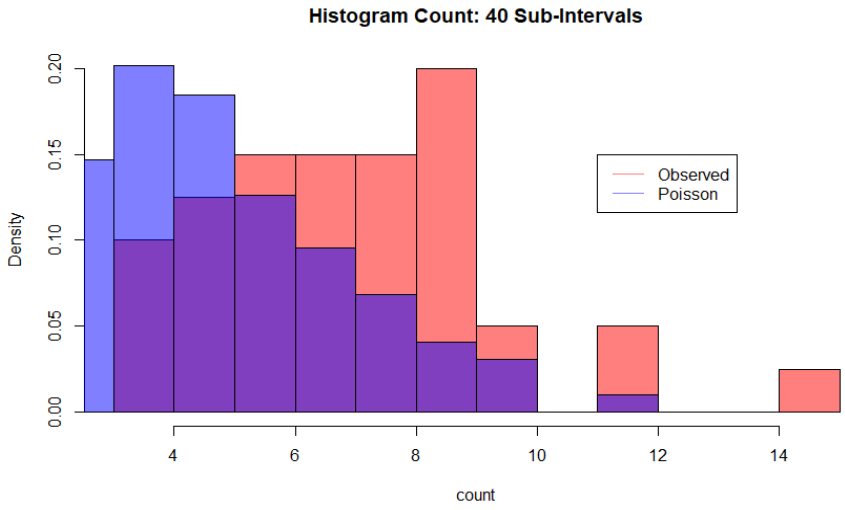
Scenario 3: Count

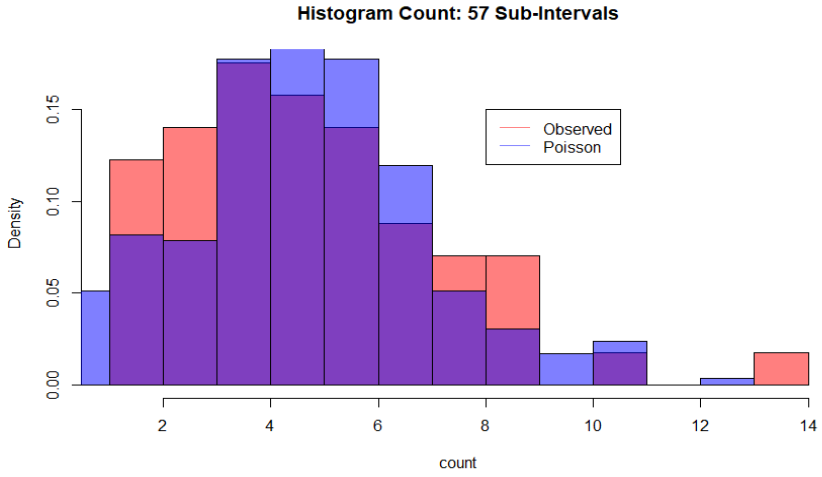
H0: Count distribution follow a Poisson Distribution
H1: Count distribution does not follow a Poisson Distribution

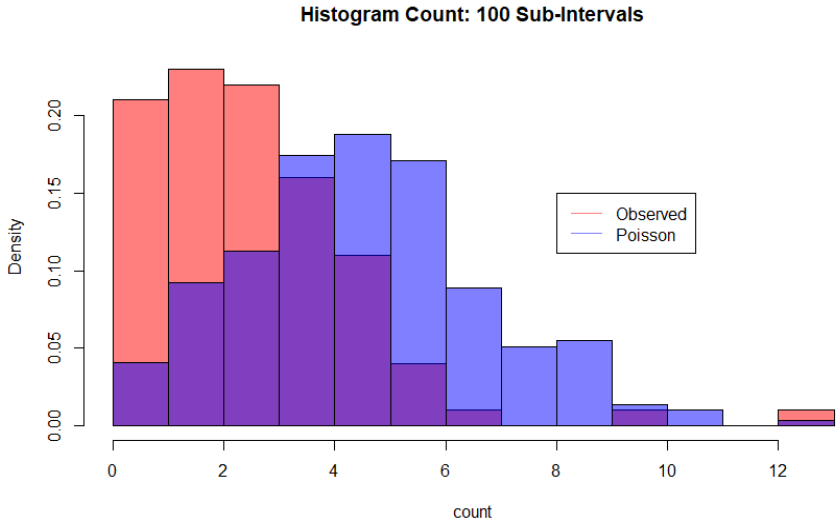
In this case we divided our dataset into three different sub-regions: 40, 57, 100. Each is expected to have a base pair of: 5734, 4000, 2293 respectively. We split up each case into non-overlapping sub-regions. Histograms were used to display the frequency of each location count and a Poisson distribution. Also, we added a Chi-Squared test in order to see if the statistical findings correlated with the visuals.

In *figure 12* we can see that distribution for location varied depending on the number of sub-intervals. When compared with a random Poisson distribution, the frequency for each count differs, making them seem like opposite distributions. On one hand, the distributions with 57 sub-intervals can be said to closely resemble one another however, the outliers are potential factors that invalidate this statement. Therefore, we can conclude that our count data does not resemble a Poisson distribution for this instance.

Figure 12: Count Palindrome Histograms







In order to reject the null hypothesis, the p-value must be smaller than 0.05. We performed a Chi-Squared test on all three sub-intervals and obtained p-values all greater than 0.05. Therefore, independently of what we observed in the above count distributions, *figure 12*, we can presume that the location data does fit a Poisson distribution.

Table 3: Chi-Squared Test for Count Distributions

	40 Sub-intervals	**57 sub-intervals**	**100 sub-intervals**
p-value	0.214553	0.1665476	0.9621318

Scenario 4: Cluster Separation

In this case, we wanted to get a better understanding of the count distribution for the location data. Assuming the count distribution is Poisson, we wanted to determine if interval size played a role in obtaining a higher or lower p-value. Hence, we started at 20 sub-intervals and used a Maximum Likelihood Estimator to calculate lambda and p-value for each subsequent interval until reaching 80. The same Chi-Squared test used in scenario 3 was used to calculate each p-value and all values lower than 0.05 where recorded in *Table 4.* It is important to point out that only 13 sub-intervals rejected the null hypothesis, where most ranged among the 20s. Furthermore, this can be an indicator that sub-intervals lower than 30 may potentially be too broad for accurate comparisons among distributions. Also, it is worth to note that after 35 sub-intervals, p-values start to drastically increase, therefore making 47, 64, 71 and 77 important intervals to consider. These four values may hold the best fit for separating palindromes that otherwise could be lost in large regions or split in small ones.

Table 4: All Intervals Between 20-80 with < 0.05 p-value

Sub-Interval	lambda	p-value
20	14.800	0.00
23	12.869	0.00
24	12.333	0.00
27	10.962	0.00
28	10.571	0.00
29	10.206	0.041
30	9.866	0.00
34	8.705	0.00
35	8.457	0.010
47	6.297	0.026
64	4.625	0.009
71	4.169	0.043
77	3.844	0.016

Theory

QQ-Plot

Graphical tool used to identify if data provided comes from some theoretical distribution such as Normal or exponential. Two sets of quantiles are plotted in a two dimensional space against one another and if the points form a roughly straight line, it's an indication that they come from the same distribution.

Residuals

A powerful diagnostic tool used to calculate the difference between the actual observed values and the expected values from a specific model. In other words, it's a detailed look at what is left over from the data and predictor. Ideal residuals approximate to zero, meaning the analysis was successful in acquiring the desired result.

T-Test

Inferential statistical tool used to determine if there is a significant difference between the means of two groups. Three main factors are taken into account to determine the probability difference among datasets which are t-statistic, t-distribution and degrees of freedom. Null hypothesis can either be rejected or not rejected based on the p-value.

Chi-Squared Test

Statistical tool used for comparing observed data to a model that distributes the data according to the expectation that the variable is independent. In the case the observed data does not fit the model, likelihood that the variables are dependent become stronger, hence rejecting the null hypothesis. This test was initially designed for categorical data and will not work for parametric or continuous data.

The Poisson Process

This statistical process enables us to find the probability that a certain event occurs within a specified interval. The symbol lambda is used to signal the average number of events in given time interval.

Exponential Distribution

A continuous probability distribution used to model the time we need to wait before a given event occurs. Its strictly related to the Poisson distribution in the case an event can occur more than once and the time between two successive occurrences is independent and exponentially distributed.

Gamma Distribution
Statistical distribution that arises naturally in Poisson processes for which the waiting times between events are relevant. In other words, is generally used to predict the waiting times between each event. Main parameters are alpha and rate(beta).

Conclusion

Overall, we analyzed specific components that historically have proven to be effective in finding potential replication sites. From Scenario 1 we were able to visualize the scatter of palindromes across the entire string of base pairs as well as compare it to the uniform distributions. From this analyzes we found that graphically location, spacing and count distributions did not have the traits of uniform distributions. For further analysis, in scenario 2 added a third uniform random distribution and used QQ-plots to run more visual comparisons, which all discarded a possible correlation among distributions. Also, we divided the location data into different sub-intervals and measured the distance between consecutive palindromes, pairs and triplets, where after visual and statistical analysis concluded that the data distribution was unique. A similar procedure occurred in scenario 3, where we compared counted palindromes for each sub-region with a Poisson distribution. In this specific case our visual and statistical analysis contradicted so we decided to opt for our Chi-Squared test result and not reject the null hypothesis. Afterwards, assuming the counted palindrome distribution followed a Poisson distribution, in scenario 4 we calculated lambda and p-value for each interval value from 20 to 80. This was done in effort to prevent any intervals with a large size of palindromes go undetected due under or over splitting the data. This procedure resulted in the finding of four potential sub-interval fits that increases the chances of finding potential cluster sites in our data.

In general, my recommendation to biologist would be to run further testing on palindrome count distributions as there may be significant cluster findings if the interval split is done right. Particularly because based on most test, the location data seemed unique and not a circumstance of chance. Likewise, take into consideration the potential concentration zones mentioned above as these may be potential starting points to for reproduction origin testing.