# Pattern Analysis of the CMV DNA Molecule

Ayaan Saifi and Antony Sikorski

## 0. Contribution Statement

Both Ayaan Saifi and Antony Sikorski contributed evenly to the project. The R code was evenly split, and each author contributed equally to the write-up of the report. The advanced analysis was additionally evenly split between the two. That said, specific primary contributions are as follows:
Ayaan Saifi was the primary contributor for Questions 1 & 3.
Antony Sikorski was the primary contributor for Questions 2 & 4.

## 1. Introduction

In order to better combat viruses, scientists often examine the DNA of the virus in order to study the way the virus replicates. In the case of our study, we are examining the DNA of CMV, a member of the herpes virus family. The CMV molecule's DNA consists of 229,354 base pairs. In an attempt to search for replication sites, we study the DNA in order to find locations of complimentary palindromes within it. Scientists have theorized that locations with such patterns are often replication sites for the virus.

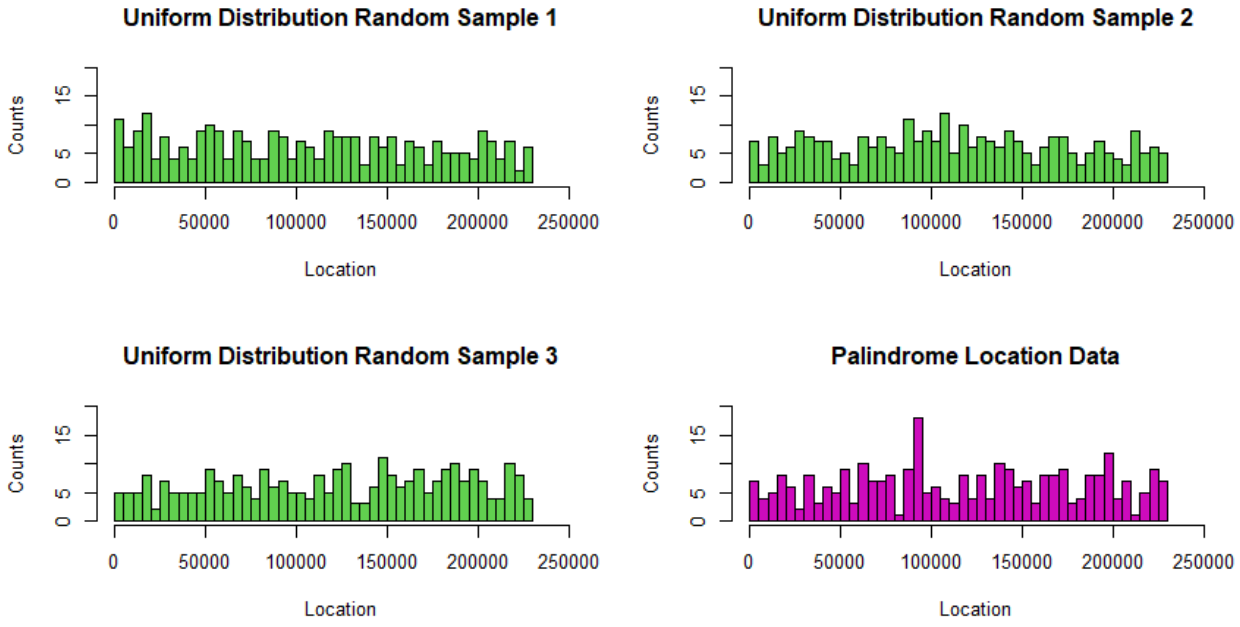
In this study, we examine a data set of 296 locations of palindromes that are between 10 and 18 base pairs long within the CMV DNA molecules. Biologists theorize that clusters of these palindromes indicate potential replication sites for the virus. Since testing each segment of the entire strand is expensive and time consuming, finding these clusters could lead to a much quicker process for examining replication sites. To begin, we generate multiple random samples from uniform distributions to compare to our distribution of locations across the sequence in an attempt to show that the location distribution is relatively uniform. Afterwards, we examine the spacings of the palindrome locations in order to better predict both the clustering and the intervals that we must study the sequence at in order to be most effective. Since departures from uniformity may indicate clustering, we used more formal statistical tests to isolate significant clusters, and found that the largest one is statistically improbable enough that it would be worthwhile to perform lab testing on it, since it is most likely a replication site. After this, we used a more advanced method in which we found the average spacing of palindromes within segments of the DNA, and then found those with the minimum average spacing, which happened to be the same zones that we found to be statistically significant enough for lab testing.

# 2. Basic Analysis

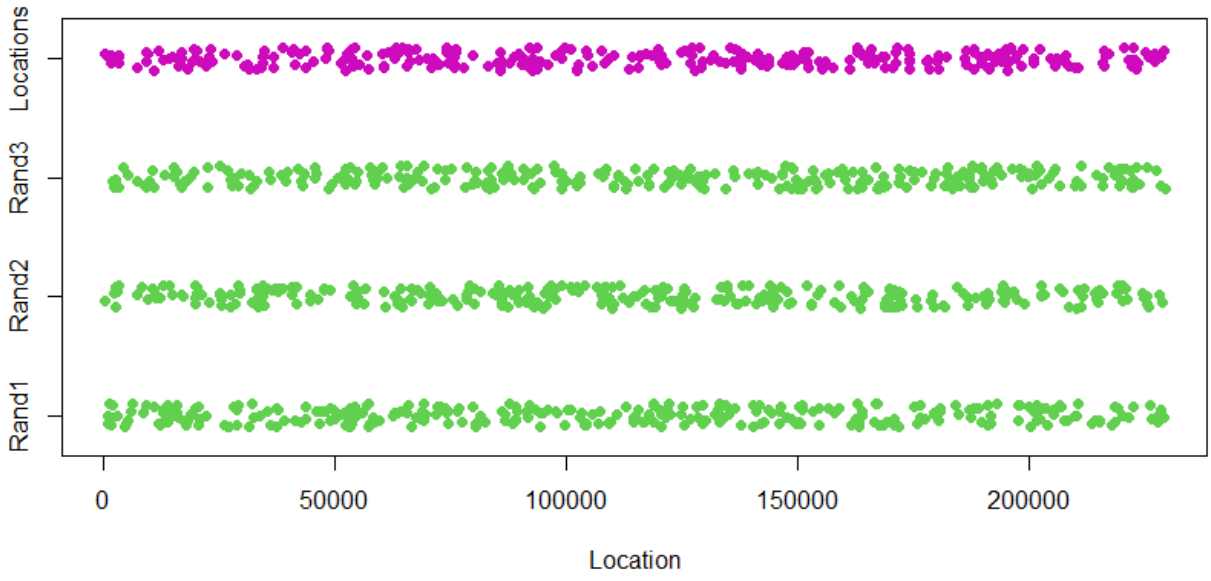
## 2.1. Comparing Random Scatter Simulations to Real Data

**Methods:**
To begin our analysis of the data, we first wish to show its uniformity, so that we can later examine deviations from that general uniformity in order to isolate potential replication areas. To demonstrate general uniformity, we generate three random samples of 296 out of 229,354 locations. To do this accurately, we randomly sampled without replacement from a uniform distribution, so that we can use them as a reference for our data. We generate multiple random scatters in order to show consistency, because one comparison is not enough. In order to initially compare the data, we plot histograms of the random scatters, along with the actual location data, in order to visually compare them. We also use a stripchart to show all of the distributions alongside each other, and plot their CDF's to show similarity.

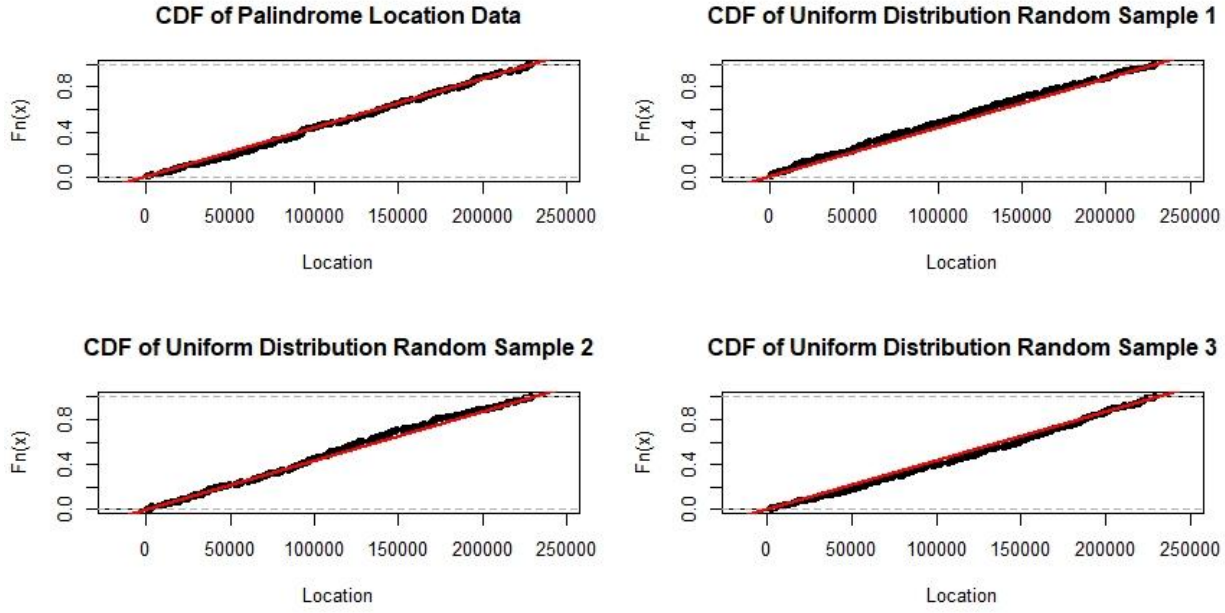
**Analysis:**
First, we generate our random samples, and plot the distribution of locations on multiple histograms to compare to our actual location data.




Next, we compare the distributions using a stripchart, which will allow us to visually analyze them side by side.




Finally, we plot the CDFs of our random scatters and our actual data to show the similarity.




**Conclusion:**

It appears that our data looks quite similar to the uniform distributions, with the exception of a few outliers, which are high numbers of palindromes in a visual cluster. The random samples from uniform distributions show no apparent maxima, minima, or outliers, and remain consistently around 6-8 counts per each 4000 location interval. The maximum number of palindromes in the uniform random distributions appear at 12, and the minimum is at 2. Our actual data shows a maximum at 18 and a minimum at 1, which is clearly a large range. The

stripchart shows the distributions of the random scatters and that of our actual data to be extremely similar on a large, side by side scale, and the cumulative distribution functions show very little difference between random samples and data as well. It appears that the principal difference between our data and the uniform random scatters is the existence of these clusters of palindromes. Aside from that, visually, on a large scale the random distributions and the actual data appear to be very similar.

## 2.2. Analyzing Locations and Spacing

**Methods:**
To further compare the random samples to our actual data, we examine the location spacings between palindromes for both. First, we calculate the distance between single pairs, doubles (skipping over a location in between), and triples (skipping over two locations), and examine the distributions of these spacings. We will find the quartiles, maxima, minima, and averages of the spacings between the random scatters for all three different categories. After this, we will plot them side by side in order to be able to perform visual analysis, using a variety of methods, such as side by side stripcharts and histograms. Since the homogeneous Poisson distribution is a good model for a uniform random scatter, we will also quickly examine the distribution of the spacings compared to the distribution of a theoretical exponential distribution with the same average parameter. This is because the distances between locations for a Poisson distribution are distributed via an exponential distribution.

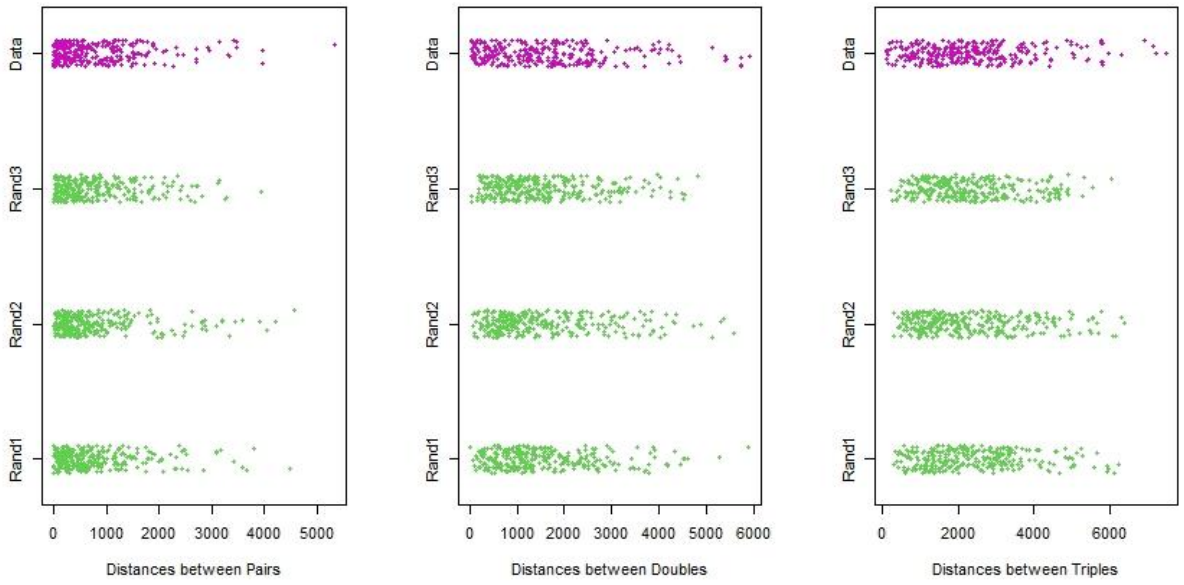
**Analysis:**
To begin, we use the diff function with lag features to find the spacings between locations for pairs, doubles, and triples. A general summary including means is shown below for each of the distributions.

| Actual Location Data for Pairs (pairs), Doubles (skipone), Triples (skiptwo) | Uniform Random Scatter Data 1 for Pairs (pairs1), Doubles (skipone1), Triples (skiptwo1) |
|---|---|
| `> summary(pairs)`<br>`Min. 1st Qu. Median Mean 3rd Qu. Max.`<br>`1.0 160.0 512.0 775.5 1144.0 5333.0`<br>`> summary(skipone)`<br>`Min. 1st Qu. Median Mean 3rd Qu. Max.`<br>`33.0 559.2 1386.0 1550.6 2147.5 5926.0`<br>`> summary(skiptwo)`<br>`Min. 1st Qu. Median Mean 3rd Qu. Max.`<br>`82 1271 2078 2327 3021 7488` | `> summary(pairs1)`<br>`Min. 1st Qu. Median Mean 3rd Qu. Max.`<br>`1.0 228.0 553.0 773.3 1024.0 4488.0`<br>`> summary(skipone1)`<br>`Min. 1st Qu. Median Mean 3rd Qu. Max.`<br>`3.0 765.8 1265.5 1550.2 2192.2 5889.0`<br>`> summary(skiptwo1)`<br>`Min. 1st Qu. Median Mean 3rd Qu. Max.`<br>`292 1357 2135 2330 3089 6212` |
| Uniform Random Scatter Data 2 for Pairs (pairs2), Doubles (skipone2), Triples (skiptwo2) | Uniform Random Scatter Data 3 for Pairs (pairs3), Doubles (skipone3), Triples (skiptwo3) |

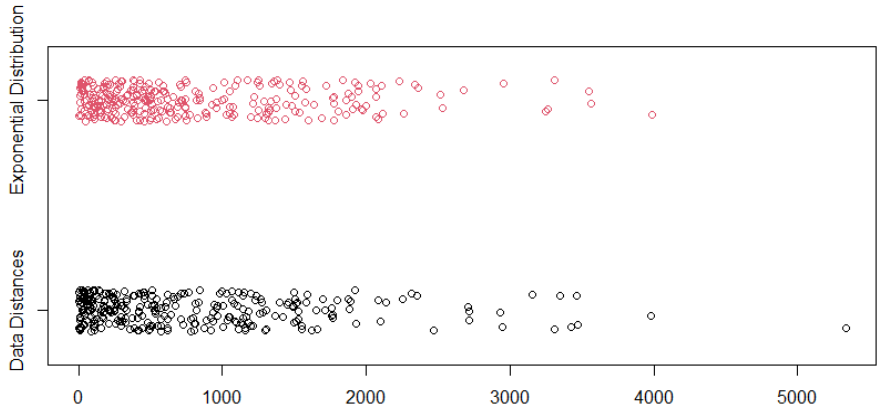
```
> summary(pairs2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    6.0   219.5   473.0   775.0   964.5  4585.0
> summary(skipone2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   77.0   657.2  1160.0  1545.9  2158.2  5575.0
> summary(skiptwo2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    298    1175    1996    2313    3182    6372
```

```
> summary(pairs3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    3.0   225.0   519.0   770.3  1054.0  3953.0
> summary(skipone3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   46.0   753.8  1353.5  1542.9  2074.5  4816.0
> summary(skiptwo3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    195    1358    2110    2315    3013    6030
```

After this, we plot stripcharts of the distributions of spacing for pairs, doubles, and triples, in order to visually analyze them. Further histograms can be found in the appendix (Appendix Figure 1).




Finally, we plot the data for our actual spacings against an exponential distribution with a mean that is identical to the mean distance between pairs for our data. The comparison is another way to determine how similar the data is to a random scatter.



**Conclusion:**
The general statistical summaries of the random and actual distributions for distances between locations appear to be extremely similar. The means for all of the pairs are roughly between 770 to 780 spaces apart, while the means for doubles are between 1540 and 1550, and the means for triples are between 2310 to 2340. The variation is obviously expected to increase a bit from pairs to doubles to triples, and the data is consistent with that expectation. The primary difference between the location data and the random samples is that the actual data appears to have a larger range and a greater variance, making the minima, maxima, and quartiles slightly more drastic in comparison to the uniform random scatters. This is most likely due to the mild skewing done by the clusters, but the effect is quite minor. Plotting the pairs, doubles, and triples against each other demonstrates high similarity between all of the distributions again, and further confirms that our data generally conforms to a uniform model. The stripcharts and histograms also explain the greater variation of the statistical summaries, because the outliers and higher variation are visually identifiable. Finally, the distribution of the pairs is very similar to the theoretical exponential distribution with the same mean, which further confirms similarity to a uniform random scatter. This is because the exponential distribution models waiting times between a homogeneous Poisson process, which is a good model for a uniform random scatter. The doubles and triples were not plotted against their respective theoretical distributions. These would simply be Gamma distributions of parameters of 2 and 3, rather than an exponential distribution, but were found to be not necessary due to all of the strong evidence pointing towards the fact that the random and actual spacings were extremely similar.

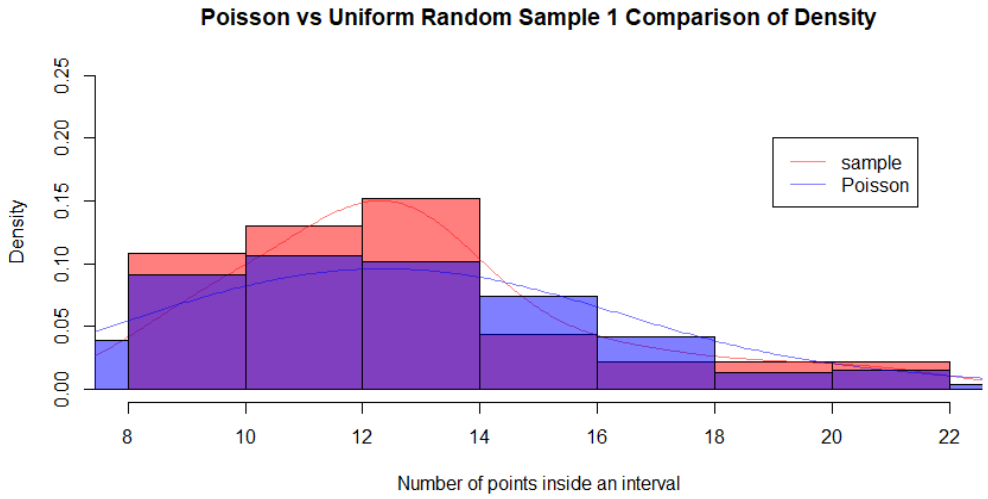
## 2.3. Formal Testing for Counts and Frequency of Palindromes

**Methods:**
Since it is already known that the Poisson is a good approximation for the model of a uniform random scatter, we decided to show that we cannot disprove this point by comparing our uniform random scatters to that of a theoretical Poisson. We plot the distribution of densities for the number of points that is expected to be found within an interval of 10,000 locations, and perform chi-square tests for goodness of fit to show that it cannot be said that the Poisson model is not a good fit for our uniform random scatters. We then continue by comparing the frequency of counts in intervals of 10,000 between our uniform random scatters and our actual data visually, and via chi square tests to demonstrate similarity. In order to be able to perform more accurate and formal analysis on deviations from uniformity in our data, we experiment visually with multiple different interval lengths (10000, 4000, 2300) in order to see which one is most viable for representing our data for further analysis. Since we chose the interval length of 4000, we isolate 6 intervals of length 4000 which appear to deviate from uniformity, and perform a chi square test in order to demonstrate that they are significantly different from the rest of our data. This rigorous process is an attempt to show the general uniformity of the palindrome location

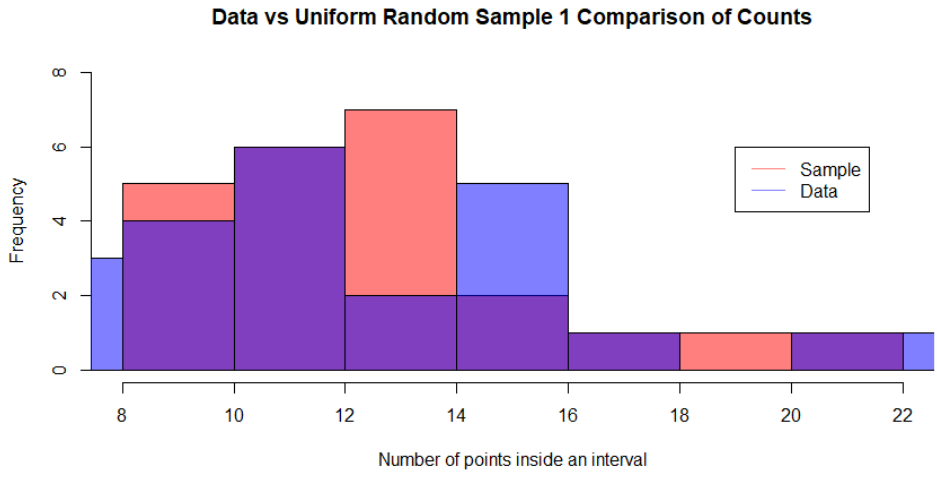
data, all the while beginning to study the significant deviations to identify potential replication sites.

**Analysis:**

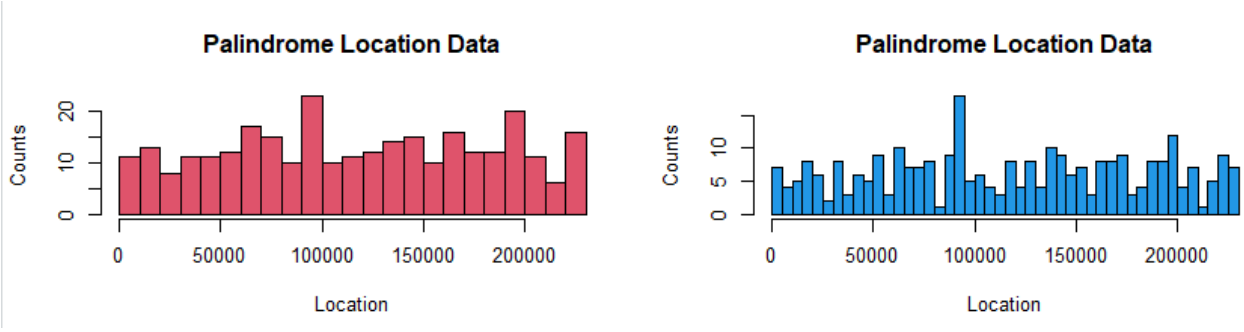
To begin, we compare the density of a theoretical Poisson to the uniform random scatters, and then perform chi-square tests to demonstrate that the Poisson is not significantly different. The chi square tests for the first, second, and third random samples return p-values of 0.8132, 0.7108, and 0.6735, respectively. An example plot is shown below for the first random scatter being tested against the theoretical Poisson. The rest can be found in the Appendix (Figures 2,3).

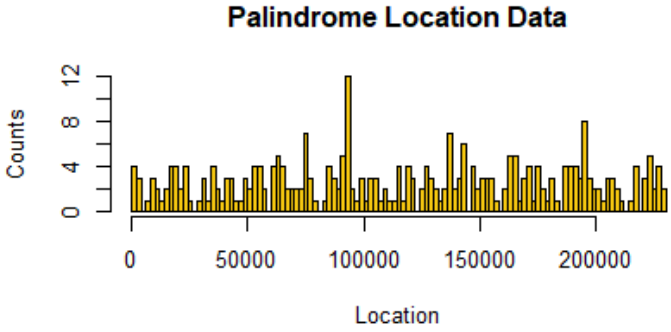



Next we split the uniform random scatter data and the actual location data into intervals of 10,000 locations, and then plot the number of counts and the frequency for the different distributions against each other. Chi square tests are performed comparing the data to each of the random distributions in order, and return p-values of 0.4413, 0.4138, and 0.182 for the data against the first, second, and third random scatter, respectively. An example plot with our data being compared to the first random scatter is shown below. The rest can be found in the Appendix (Figures 4,5).



We experiment with different interval lengths of 10000, 4000, and 2300 locations below (in order right to left), to see which one is most viable for visually isolating clusters.




Finally, we isolate the following intervals of size 4000: (76000, 80000), (80001, 84000), (87500, 91500), (91501, 95500), (196000, 200000), (200001, 204000). These intervals appear to deviate from the uniformity of our DNA location data, and perform a chi-square test against the uniform, which returns a p-value of $2.2 \times 10^{-16}$.

```
        Chi-squared test for given probabilities

data:  setup
X-squared = 88.015, df = 5, p-value < 2.2e-16
```

**Conclusion:**

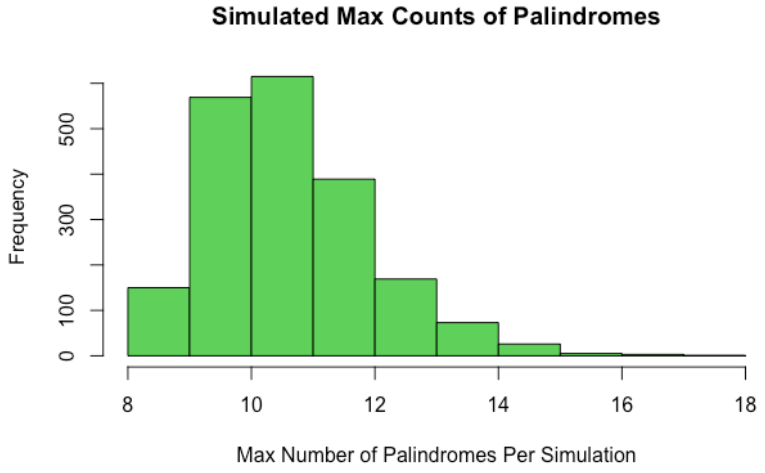
More formal statistical testing allows us to show that the Poisson distribution is not significantly different from our uniform random scatters due to the high p-values that are much greater than alpha = 0.5, which further reiterates the already known hypothesis that the Poisson model is a good approximation for a uniform random scatter. We also show that none of the uniform random scatters are significantly different from our actual data, with slightly lower p-values that nevertheless exceed the alpha = 0.5 significance level by a large margin. This yet again shows that the uniform model may be a good general fit for our DNA palindrome locations. To allow for further and more precise testing, we experiment with different interval sizes, and find that the most efficient one for visual analysis appears to be around 4000 locations long. This interval

length is not too large that it hides individual clusters, but not too small that it splits them up so that they are harder to find. After this interval length was decided, the 6 locations were tested against uniformity, and the extremely low p-value shows that these are clearly significant deviations from the general uniform model. These locations are most likely the ones with the highest chance to be replication sites, and should be tested further so that it can be determined which ones need to be analyzed in a laboratory.

## 2.4. Further Examining the Largest Clusters of Palindromes

**Methods:**
In order to identify possible origins of replication within the CMV DNA, we need to find the intervals with the greatest number of palindromes and test their statistical significance against theoretical uniform distributions. We did this by first simulating 2,000 uniformly distributed random samples of 269 palindrome locations (without replacement) from our population of 229,354 base pairs. Each of these 2,000 samples were then plotted on histograms with an interval size of 4,000 (which was chosen as the optimal interval size based on our findings from 2.3). We then took the maximum number of palindromes in an interval from each histogram, thus generating a set of 2,000 maximum values, which were subsequently plotted on a separate histogram. Lastly, we tested the significance of the maximum number of palindromes in an interval of 4,000 from our given data against the distribution of maxima from our simulation of theoretical data.

**Analysis:**
To begin, after simulating theoretical palindrome locations, we plot a histogram of the maximum number of palindromes within an interval in order to provide a visual representation of the theoretical distribution.



We then test the statistical significance of the maximum number of palindromes in our given data with a null hypothesis that the theoretical distribution of maxima is a good parent fit for our original data's maximum value, and an alternative hypothesis that it is not a good fit for a parent distribution. We did this by taking our original data's maximum value (18 palindromes) and finding the number of times our theoretical distribution generated a maximum of greater than or equal to 18 palindromes (1 instance). Lastly, we divided the number of such instances by the total number of simulations conducted (2,000) to find a p-value of 0.0005.

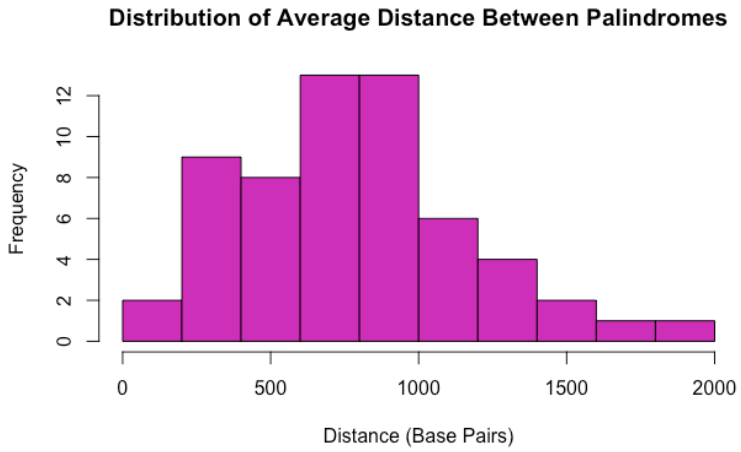
**Conclusion:**
From our test of statistical significance, we found the p-value to be 0.0005, which is significant at both the 5% and 1% significance levels. As a result, we reject the null hypothesis, thus showing us that the maximum number of palindromes in our given data is not a good fit for the theoretical distribution of maxima. From this, we conclude that the maximum number of palindromes present in an interval of our given data is out of the norm of our normal theoretical distributions. Therefore, clusters of palindromes such as the one tested above prove to be irregularities, and would thus likely be good candidates for further biological testing.

# 3. Advanced Analysis

## 3.1. Analyzing Average Group Distance Between Palindromes for Clusters

**Methods:**
In order to further gather information on the presence of palindrome clusters within the given data, we conducted an in depth analysis of the average space between palindromes. This was done by first calculating the distance between each palindrome location, and subsequently breaking these distances into intervals of 5 palindromes and finding the average of each interval. This gave us a set of 59 average distances (with each entry being the average distance between 5 consecutive palindromes). This data was then sorted in order to find the lowest average distances and their corresponding palindrome locations.

**Analysis:**
After calculating the average distance between each palindrome in a given interval, we plotted a histogram of the averages in order to visualize the distribution of spacing between palindromes in our data, thus providing us with the necessary tools to search for potential clusters of palindromes.




**Conclusion:**
After sorting our averages from least to greatest distance, we picked the three smallest distances as representations of possible clusters of palindromes within the given data. The smallest average distance (43.2 base pairs) was found to be between palindromes 116 - 120 (between base pairs 92,570 - 92,783). The next smallest average distance (100.4 base pairs) was found to be between palindrome 91 - 95 (between base pairs 75,622 - 76,043). Lastly, the third smallest average distance (222.0 base pairs) was found to be between palindrome 251 - 255 (between base pairs 194,111 - 195,151). Due to these three intervals having the smallest average distances between palindromes, we concluded that they were the most likely to contain clusters of palindromes, and therefore had a greater likelihood of containing origins of replication. Our findings through this analytical method reaffirmed our previous discovery of base pair intervals that significantly deviated from uniformity in terms of palindrome occurrences.

# 4. Discussion and Conclusion

The objective of this study was to determine where potential replication sites could be in the DNA molecule of CMV, a member of the herpes virus family. The potential replication sites of the virus are marked by high numbers of complimentary palindromes of length 10-18 base pairs in one area. Our data, which consisted of 296 locations of these palindromes, was compared to random scatters that were generated by sampling from a random distribution. Visually comparing the data to the random scatters showed that the DNA could be generally described as uniform, but had deviations from uniformity which were most likely replication zones. Further visual analysis was done by analyzing the spacing between palindromes for the data and the random scatters, along with comparing our data spacing to the theoretical exponential distribution. Since the data and the random scatters matched up with the theoretical exponential distribution, which is the distribution for waiting times for a Poisson model, we obtained further evidence towards our hypothesis of general uniformity. This was proven via a rigorous set of tests comparing the

uniform random scatters, Poisson distributions, and our own data, and it was shown that we did not have statistically significant evidence to disprove the hypothesis.

With our assumption of uniformity, we moved on to find the optimum interval length for examination, and showed statistically significant intervals in which the number of palindromes was notably higher. These intervals were (76000, 80000), (80001, 84000), (87500, 91500), (91501, 95500), (196000, 200000), (200001, 204000). The interval with the highest number of palindromes occurred within (91501, 95500), and contained 18 palindromes. This maximum was tested for significance via a simulation of maxima of uniform random scatters, and it was found that the p-value of a maximum this great occurring was $5 \times 10^{-4}$, which is incredibly low, meaning this largest cluster was most likely very significant enough for testing. In order to confirm our results, we chose to group palindromes into groups of 5, and find the minimum average distances for all groups, in hopes that the groups with the smallest average distances would lie inside our maximum clusters. The interval with the smallest average distance was at interval locations (92750, 92783), which was right in the middle of our largest cluster, with the other two being within 100 locations of other clusters that we found to be statistically significant. Comparing our results to another biological study (1) done by Yiyang Xu and the rest of her colleagues at the University of Nevada, it appears that the replication site is located between the location of 90500 and 93930 nucleotides. Our estimates of the most significant cluster proved to be quite accurate considering they both encompass and fall within this interval.

Of course, while there are many improvements that could be made to this method, such as a more refined algorithm for clustering, and repeated testing on other virus DNA, the general method is one of incredible use to those who work in the field of biology. Testing each segment of DNA is incredibly expensive and time consuming, and one would be well advised to first perform some data analysis. Many patterns in DNA can be modeled as sourcing from a uniform random scatter, although if not, one must find a probabilistic model that most closely fits to what is being studied about their strand. For the sake of simplicity, we will continue on the example of uniformity. Once the model has been determined, one must simply look for irregularities in the data (deviations in uniformity), and perform a couple statistical tests to see whether or not those irregularities are significant enough to warrant further analysis. The irregularities with the highest significance are the ones that are most likely to be replication zones (or some other characteristic being searched for), and those should be taken to the laboratory so that they can be tested and confirmed. Testing each strand is far too difficult when we can simply look at the data, find a similar model, and search for deviations from that model.

# 5. Work Cited


1. Human Cytomegalovirus DNA Replication Requires Transcriptional Activation via an IE2- and UL84-Responsive Bidirectional Promoter Element within *ori*Lyt

   Yiyang Xu, Sylvia A. Cei, Alicia Rodriguez Huete, Kelly S. Colletti, Gregory S. Pari

   Journal of Virology Oct 2004, 78 (21) 11664-11677; DOI: 10.1128/JVI.78.21.11664-11677.2004


# 6. Appendix

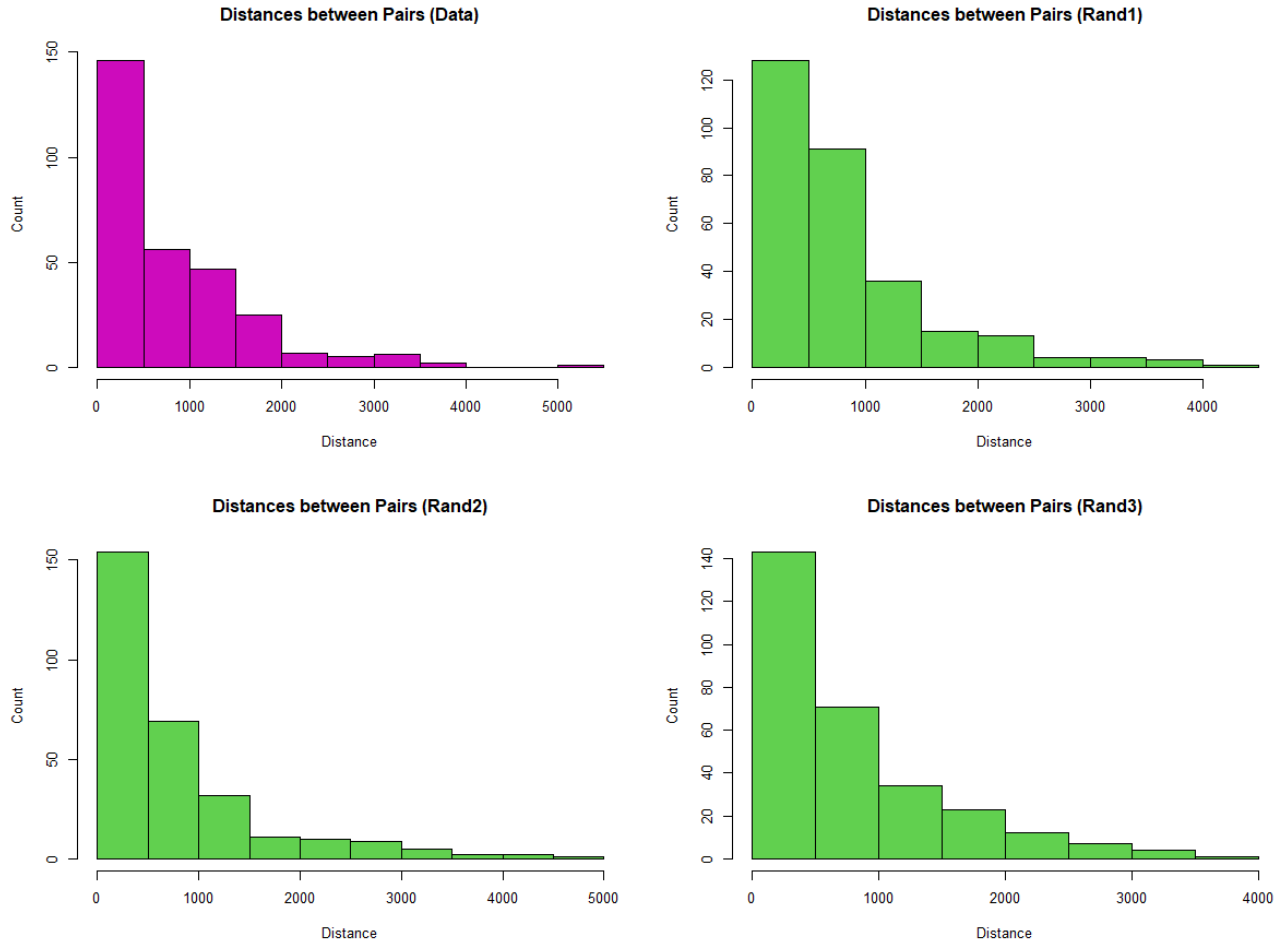



Figure 1: Histograms of the distribution of distances between pairs for both the actual location data and the random scatter.

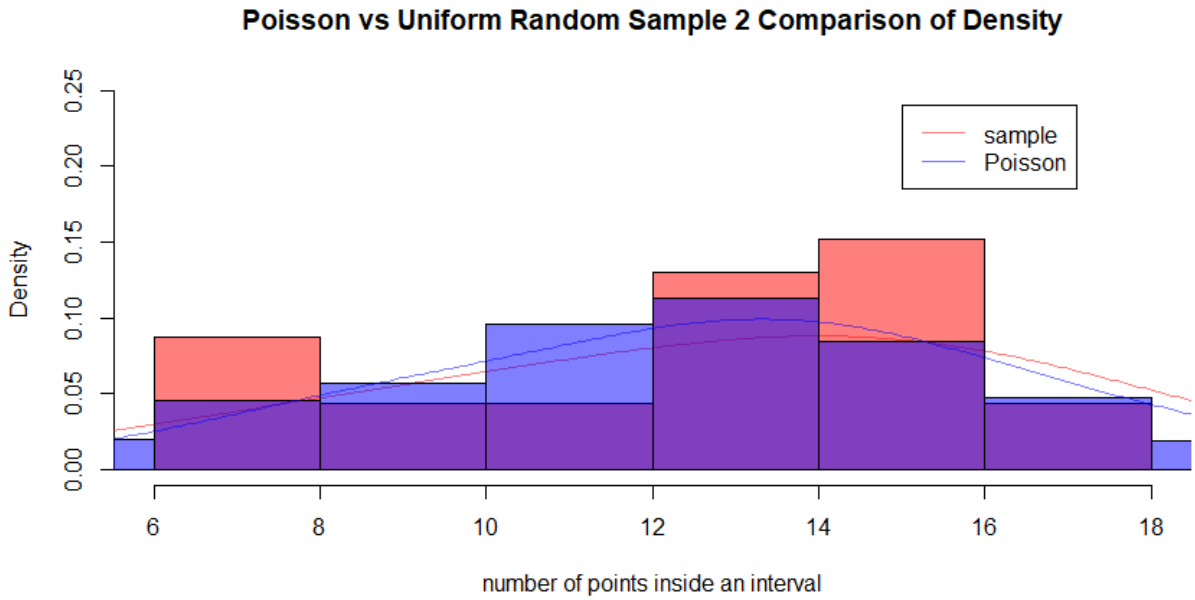



Figure 2: Random scatter 2 against the theoretical Poisson.

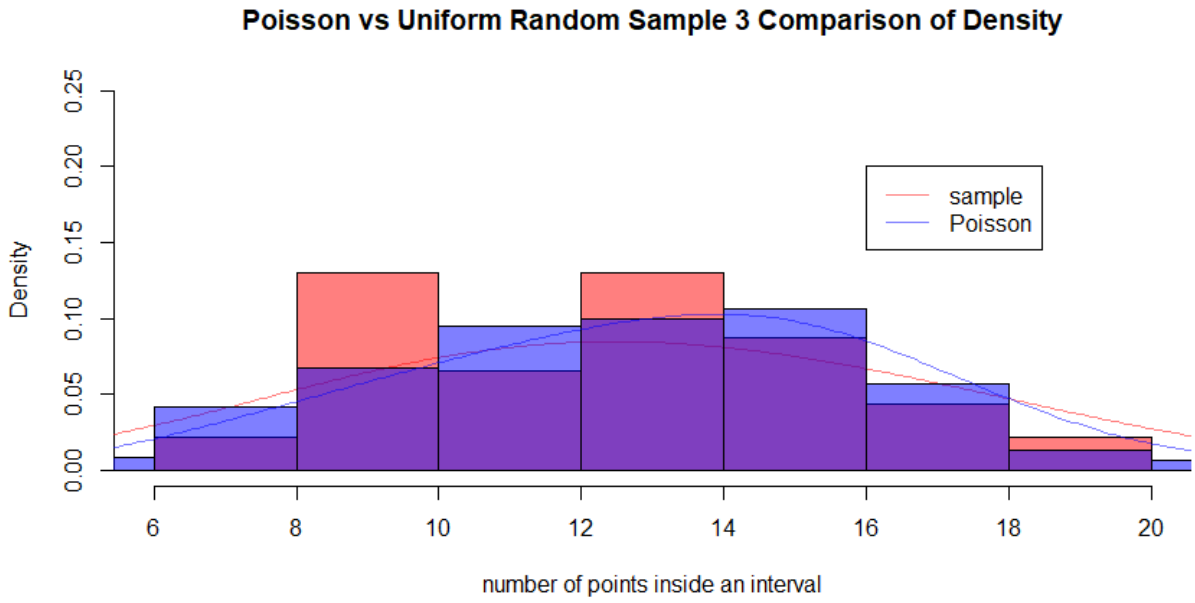



Figure 3: Random scatter 3 against the theoretical Poisson.

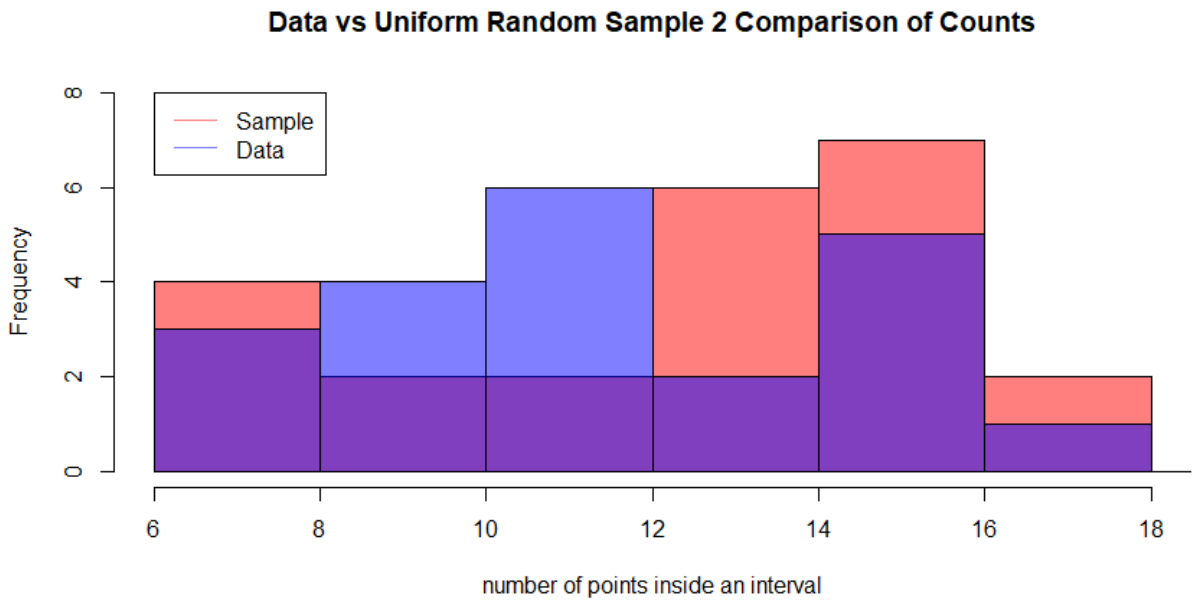



Figure 4: Number of counts histogram for random scatter 2 against actual data.

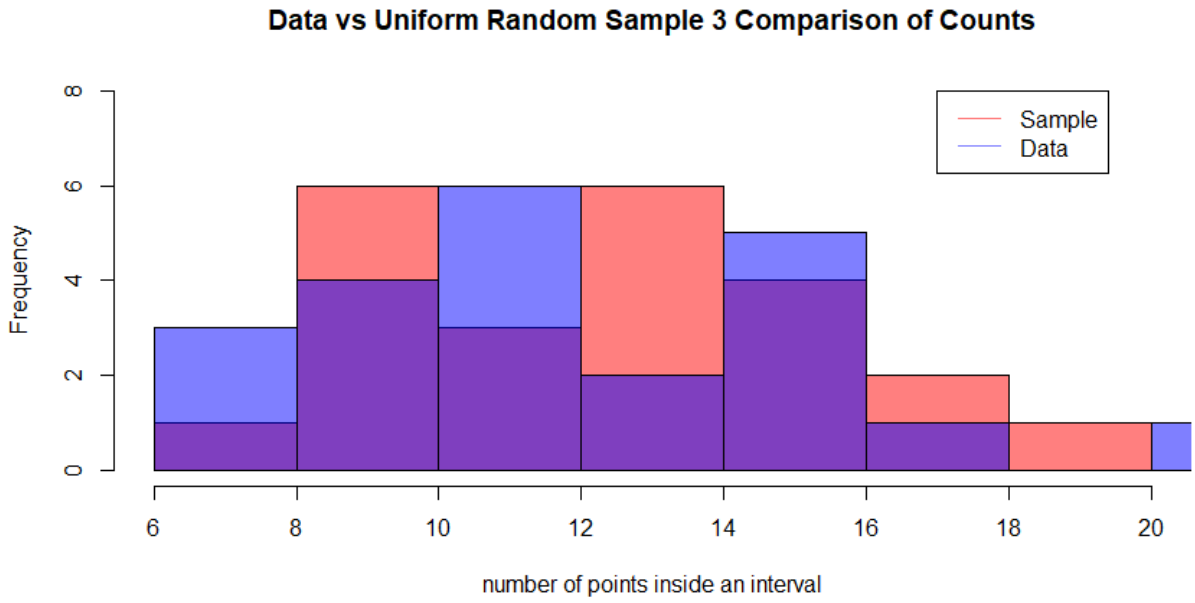



Figure 5: Number of counts histogram for random scatter 3 against actual data.