Brittney Cifuentes-Tello
Justin Nguyen

# Study of Palindromes in Human Cytomegalovirus DNA

## 1. Introduction

**Background:** Human cytomegalovirus (HCMV) is a member of the herpes virus family. Scientists who want to combat the virus try to understand how it replicates by looking at the DNA sequence, believing that clusters of palindromes might mark the origin of replication. In particular, they look at patterns such as palindromes in the DNA sequence that contains 229,354 complementary pairs of base pairs A, T, C, and G. In the studied strand, they observed 296 locations in which palindromes with a length between 10 to 18 are present. This paper attempts to conduct a similar analysis in observing patterns related to base pair palindromes.

**Methods:** We first simulated 296 palindrome sites chosen at random along a DNA sequence of 229,354 and compared the real data to the simulated data. Also, we compared spacing between consecutive palindrome pairs and triplets by having theoretical, actual, and simulated data graphically. Further, we split up the data into intervals of 10, 25, 50, 75, and 100, then found palindrome frequencies for which we calculated observed total variation distance (TVD) and performed 500 simulations with a significance of 1%, yielding p-values for the TVD and the Chi-squared test compared to a theoretical distribution. When we split the data into intervals, we recorded the location in which the highest frequency of palindromes were located and compared to see if they were found in the same location. To add on, we performed a Chi-squared test between the actual observed data for spacing of pairs and spacing of triplets and on the simulated data of the spacings.

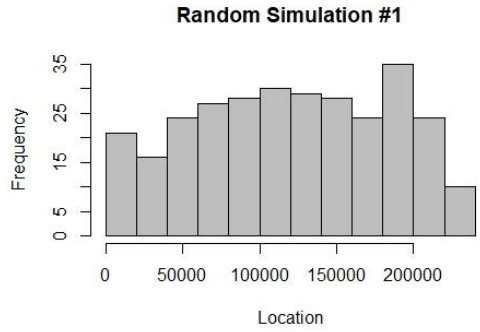
**Conclusions:** The data does look similar to the random uniform scatter simulation that we constructed, but that alone cannot fully determine if the data is uniform. The spacing between consecutive pairs and triplets is not randomly scattered when compared to the simulated spacing data. After creating multiple equally spaced intervals, the observed TVD was 0.091, 0.217, 0.365, 0.440, and 0.495, respectively. The TVD p-values for the simulations were 0.932, 0.628, 0.122, 0.102, and 0.094, and for the Chi-squared test, 0.902, 0.555, 0.049, 0.019, and 0.003. Based on the data we found the location for the biggest frequency of palindromes for each interval and noticed for intervals of 25, 50, 75, and 100, they all began at location 91,741, making it a possible starting point for replication. The mean Chi-squared tests done between spacing of pairs of palindromes and triplets gave p-values of 0.9367 and 0.9492, leading us to reject the null hypothesis.

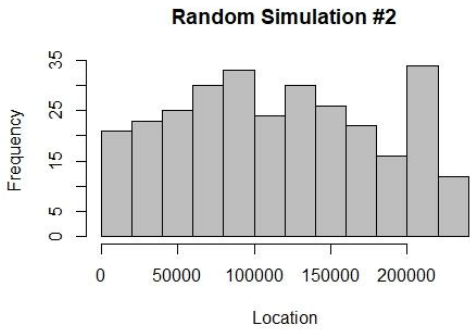
## 2. Analysis

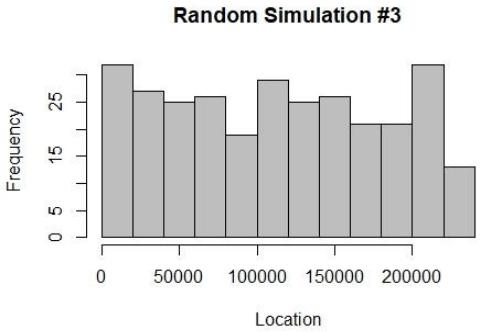
### 2.1: Random Scatter

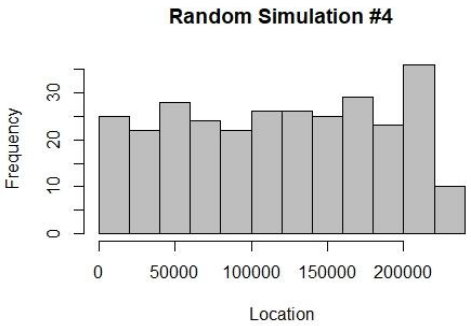
**Method:** We simulated 296 palindrome sites chosen at random in a DNA sequence of 229,354 locations using a pseudo random number generator through the sample function. Then, we simulated a uniform distribution of 296 palindrome sites 5 times using a programmed loop. We did this in order to see how the random locations would appear in the strand itself. Finally, we also graphed the actual location data to make visual comparisons between the simulations and actual data.

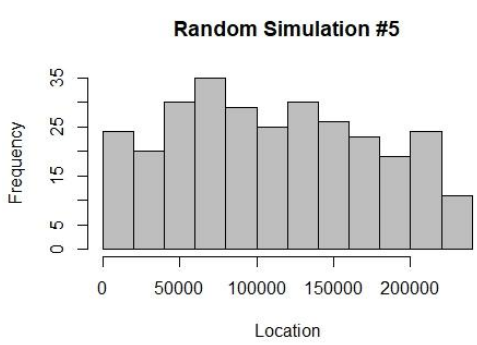
**Analysis:** We have the 5 histograms for the random uniform samples as well as the histogram for the actual palindrome data (with plot titled "Location Data"). As one can see, the real data does resemble the data for the random simulations. Since we kept the scales the same, it almost looks as if the actual data is indistinguishable from the simulated data.

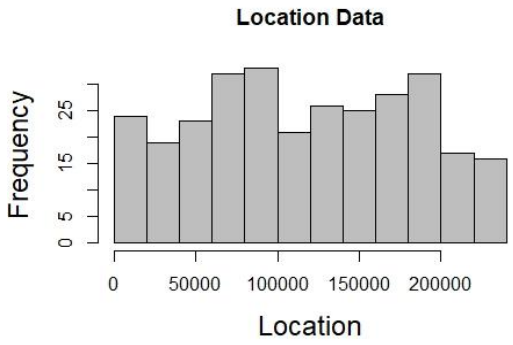




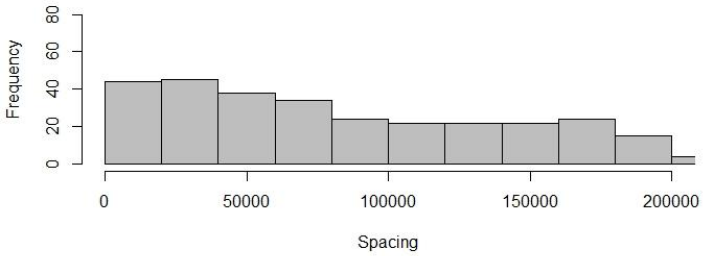

**Conclusion:** Although the goal of comparing the histograms is to see if the location data distribution is random, we cannot conclude that it follows a uniform distribution just by looking at the graph. Further we also cannot conclude that the palindromes are not part of the replicating process. It could be possible that some palindromes carry a higher significance in replication although it might still be a random uniform distribution.

## 2.2: Location and Spacing

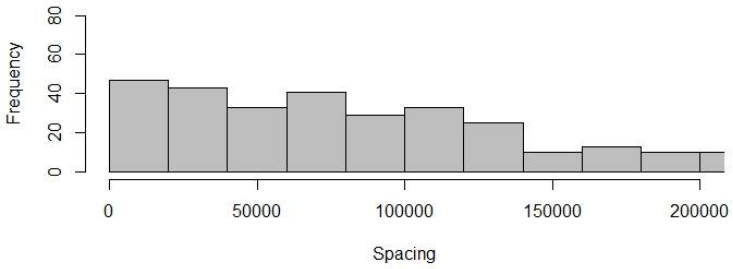
**Method:** We used graphical methods to examine the spacings between consecutive palindromes and sum of consecutive pairs then triplets. Then, we programmed a function to calculate the consecutive sums of pairs and also used a very similar method to calculate the consecutive sums of triplets as well. The function takes in the "location" data as an array and returns a new array of the sums; the pairs' array had 295 sums while the triplets had 294. After, we created a consecutive difference function that takes these arrays as a parameter and returns the spacing of the pairs and spacing of the triplets. In order to compare our results to what we expect to find in a random uniform scatter, we ran a simulation using a "sample" function that utilized a random number generator. Later, we ran 5 simulations of random spacing for pairs and 5 for random spacing of triplets. Lastly, we made theoretical data that we titled "null spacing" and have graphs for null spacing of pairs and null spacing of triplets.

**Analysis:** As one can see, our graphs for the spacing between consecutive pairs and spacing between consecutive triplets differ greatly from what we expect to see in a random scatter when set to scale. Similarly, the random simulations for pairs and triplets follow the same pattern. However, when looking at the actual data for pairs and spacing we see that with an x-limit of 200,000, they are hardly visible. There is much less variation and spread; thus, the palindrome locations are much closer together, displaying a deviation from the simulation. In comparison, the theoretical data looks much like the actual data for pairs and triplets spacing. However, the theoretical data has an even smaller spread than the actual data. As one can see, the null / theoretical graphs appear as more of a single bar rather than the actual data that has some dimension.

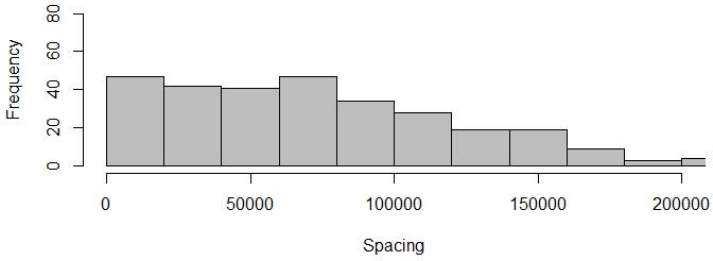
**Spacing Between Consecutive Pairs Random Simulation #1**



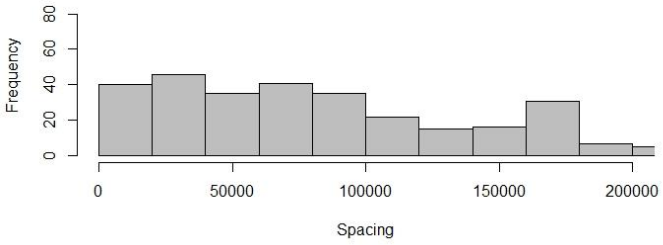

**Spacing Between Consecutive Pairs Random Simulation #2**

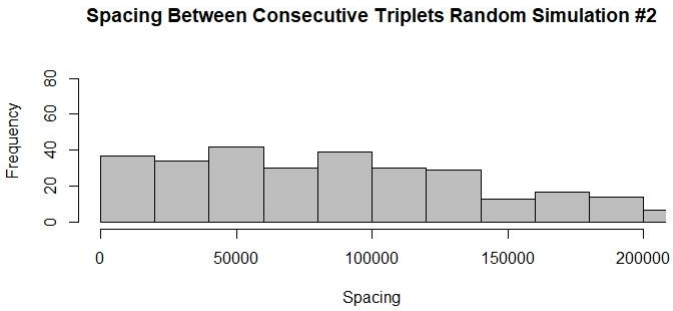


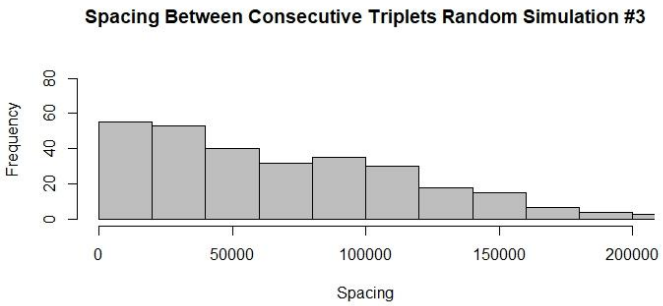

**Spacing Between Consecutive Pairs Random Simulation #3**



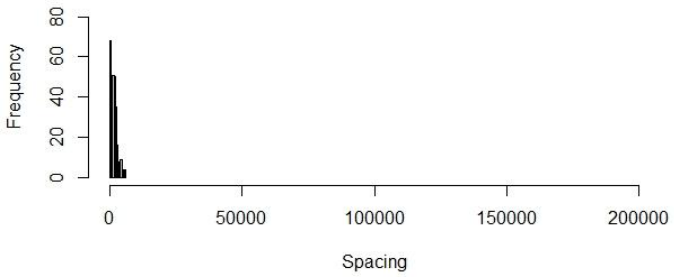

**Spacing Between Consecutive Triplets Random Simulation #1**



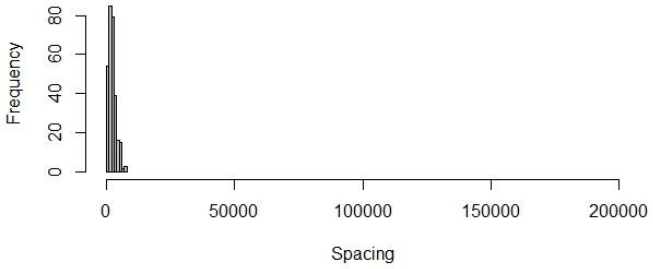

Spacing Between Consecutive Triplets Random Simulation #2
Frequency
0
20
40
60
80
0
50000
100000
150000
200000
Spacing


Spacing Between Consecutive Triplets Random Simulation #3
Frequency
0
20
40
60
80
0
50000
100000
150000
200000
Spacing

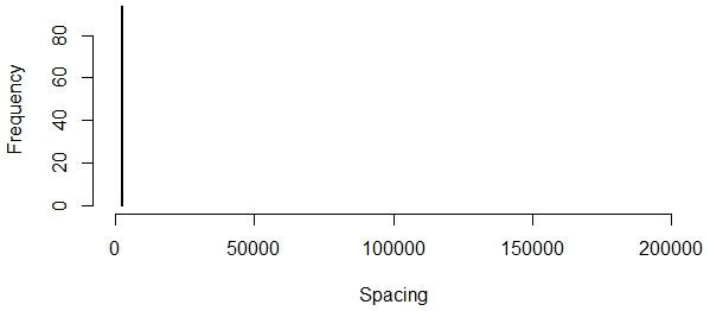
**Spacing Between Consecutive Pairs**



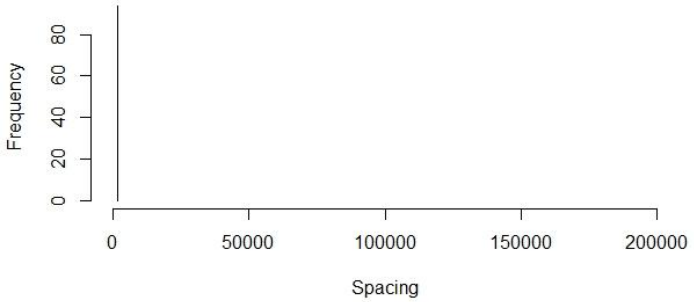

**Spacing Between Consecutive Triplets**




**Spacing Between Consecutive Triplets Under the Null Distributio**




**Spacing Between Consecutive Pairs Under the Null Distribution**

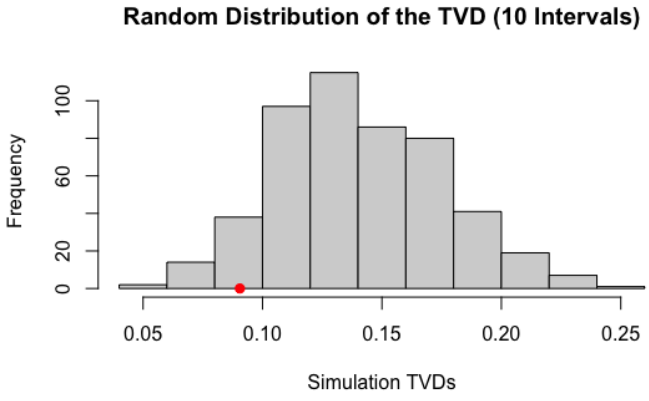


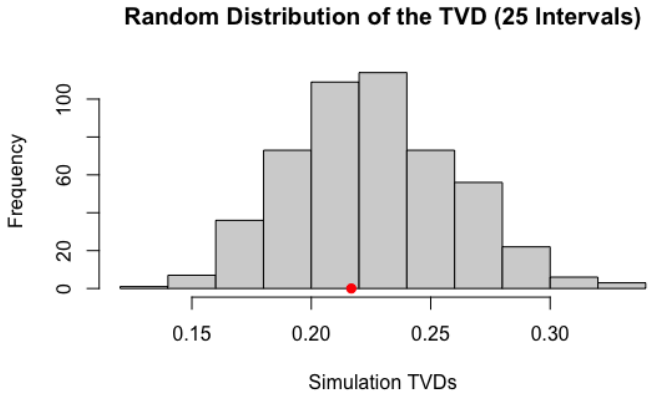
**Conclusion:** We conclude that the spacing between the sum of consecutive pairs and triplets is not randomly scattered. When compared to the simulated data, the variation is much different, showing that the spacing does not match. However, it does appear to have a closer variation to the theoretical data than the simulated data.

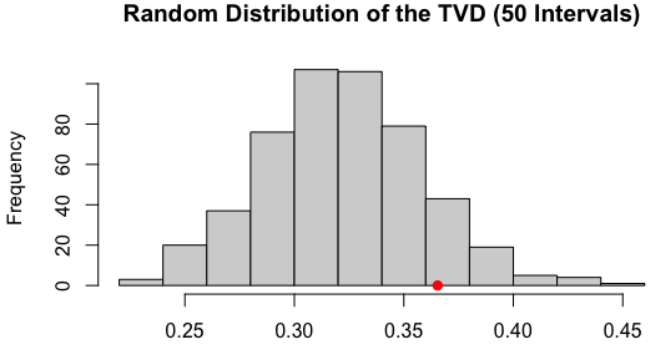
### 2.3 Counts

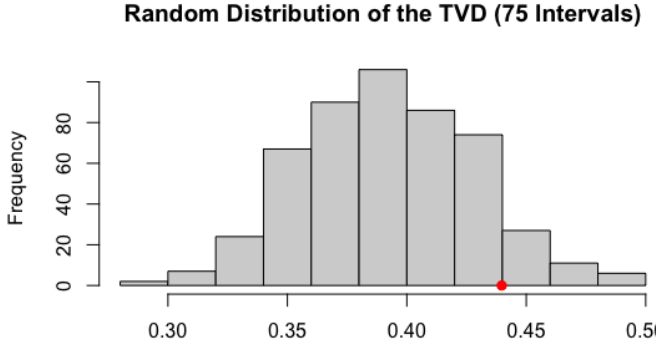
**Method:** We categorized the observed data into equally spaced intervals. Then, we aggregated by count to find the palindrome frequency for each interval and from there, converted the counts into proportions. Using those proportions, we found the observed TVD compared to the theoretical null distribution. Afterwards, we fixed a significance level of 1%, ran 500 simulations, randomly sampling 296 observations from all possible palindrome locations and computing the TVD with the theoretical null distribution. The results were stored in an array, and a p-value was calculated from the results. The random distribution of TVDs was graphed with a red point to signify the observed TVD value. After the simulations, we ran a Chi-squared test on the observed data to obtain p-values for the test. We ran through this entire process for splitting the data into equal intervals of 10, 25, 50, 75, and 100.

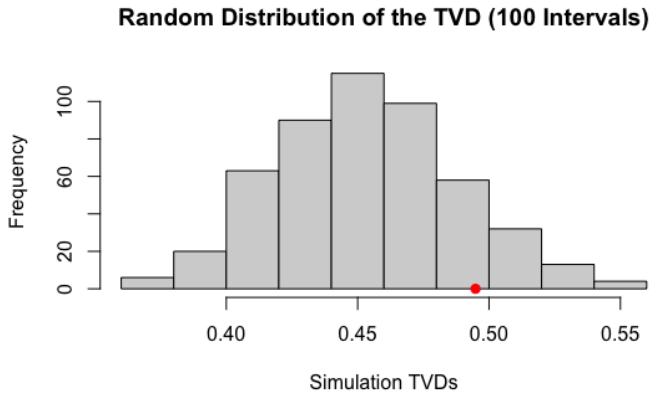
**Analysis:** Respectively, for interval values of 10, 25, 50, 75, and 100, the observed TVD statistics came out to 0.091, 0.217, 0.365, 0.440, and 0.495. The p-values for the TVD permutation test were 0.932, 0.628, 0.122, 0.102, and 0.094. The p-values for the Chi-squared test were 0.902, 0.555, 0.049, 0.019, and 0.003. As the histograms below indicate, the observed value nears closer and closer to the extreme values, as the intervals by which the data is split increase.






**Conclusion:** For all tests except the Chi-squared test in splitting the data into 100 equal intervals, each p-value was found to be greater than 0.01, meaning that we fail to reject the null hypothesis in that it is possible that the observed palindrome frequencies come from the same distribution as a random uniform distribution.

## 2.4 The Biggest Cluster

**Method:** While we ran simulations for each interval value, we also obtained the interval with the highest frequency, recorded the corresponding count, and converted the count into a proportion. From here, for each interval value to split the data by, we found the range of each biggest cluster in the data and compared them to see if any biggest clusters for one interval value contained other biggest clusters of another interval value.
**Analysis:** Respectively, for interval values of 10, 25, 50, 75, and 100, we found the biggest clusters to be the ranges (183483, 206418), (91741, 100915), (91741, 96328), (91741, 94799), and (91741, 94035). As one might note, there is an overlap in the clusters for the intervals 25, 50, 75, 100.
**Conclusion:** From the obtained biggest clusters, one can see that the biggest cluster for splitting the data into 100 intervals is contained within the biggest cluster for 75 intervals, both of which are also contained within the biggest cluster for 50 intervals. Those biggest clusters are also included in the biggest cluster for 25 intervals. From the fact that we tested partitioning the data into equally spaced intervals for different interval values and obtained a biggest cluster that coincides with the biggest cluster for other intervals, it seems likely that the interval (91741, 94799) could contain a potential origin for replication and is not just by chance.

## 3. Advanced Analysis

**Method:** We decided to quantify the difference between spacings distribution. We began by creating a simulation of 296 random sample locations from 229,354 possible locations. We then got the consecutive differences for pairs and triplets in spacing and performed Chi-squared tests on both. Lastly, we found the average of all the Chi-squared tests for pair spacing and triplet spacing against the randomized data.

**Analysis:** When the simulation was performed, and the Chi-squared test results were obtained, we had 500 values for pairs spacing and triplets spacing. Hence, we decided to find the mean Chi-squared p-values of all the simulations. For pair spacing it was 0.9367 and for triplet spacing it was 0.9492.

| | Mean Chi squared p-value |
|---|---|
| Pair Spacing | 0.9367 |
| Triplet Spacing | 0.9492 |

**Conclusion:** As one can see the mean Chi-square p-value for the spacing of pairs and triplets is large. The null hypothesis would be that the observed spacings are distributed like the null distribution of spacings, resembling the tightly concentrated vertical bars. On the other hand, the alternative hypothesis would be that the observed spacings have a different distribution compared to the null distribution. The higher the value, the more difference between the distributions. We reject the null hypothesis, suggesting that observed and null distributions are different.

## 4. Conclusion

Using the HCMV DNA, scientists observed 296 palindromes located on a sequence of DNA that has a length of 229,354. The scientists only accounted for the palindromes that were of length 10 to 18 base letters long. They did this on a theory that palindromes could be a possible origin of replication for the virus. After simulating 296 palindrome sites along the DNA length of 229,354, we plotted the data and noticed that the actual data did look like 5 graphs of uniform scatters. Thus, the actual data could possibly be a uniform distribution but we cannot be certain based solely off of graphs. Possibly one location shows more information on palindromes or carries more significance. Ergo, we advise biologists who are searching for the origin of replication for HCMV not to look at the locations individually. After looking at graphical spacings between the sum of consecutive palindrome pairs and triplets and comparing them to simulated spacing data we believe that the spacing is not uniformly distributed. However, the spacing looks more similar to the theoretical distribution than the random distribution. We advise biologists to look at the spacing since it seems to be much more spread out in the simulation compared to the actual data. When we took the p-values for Chi-squared and TVD tests for intervals

of 10, 25, 50, 75, and 100, for all except in the case of the Chi-squared test for the intervals of 100, we see the values are greater than 0.01. Thus, only in this case do we reject the null hypothesis. Therefore, we advise biologists that it seems the palindrome frequencies are random uniform distribution and to pay closer attention to intervals of 100. Using the same intervals we looked at the intervals with the highest frequencies. We then found the locations in which highest frequencies occur and saw there was an overlapping region with intervals of 25, 50, 75, and 100, all starting at location 91,741 and ending in different locations due to interval size. We would advise biologists to look into the highest frequencies in intervals of 100 as a possible replication site. On the side, we ran one more test on the data and performed 500 simulations in the spacing while calculating the resulting Chi-squared test values for all. The Chi-squared p-values for spacing pairs was 0.9367 and for triplets was 0.9492. This further solidified our idea that the spacings do not follow a uniform distribution; thus our advice remains the same: look into spacings of pairs and triplets for a possible origin replication.