The Origins of Replication within the Human Cytomegalovirus

Brandon Tran, Kevin Wong

May 2021

1 Introduction

While the SARS-Cov-2 virus ravages the world's population, a similar virus has already been infecting millions of people. The Human Cytomegalovirus (HCMV), also known as herpes, is a virus that is similar to the SARS-Cov-2 virus, except that it contains DNA instead of RNA. To find out ways to eliminate the virus, scientists have targeted its origins of replication which are marked as palindromes containing the bases found in DNA: A, G, C, and T. We can see the origins by looking for clusters within dna that are significantly different compared to the others. In this report, we will identify these clusters by uniform randomly scattering the palindromes such that we can compare the simulated data with the actual data. Additionally, we will look at the spacings of consecutive palindromes, look at the counts of palindromes in specific areas, and identify a possible biggest cluster within the DNA sequence. By doing this, we will be able to determine whether or not the locations of origin are randomly distributed or if they follow a set pattern.

2 Random Scatter Analysis

To identify whether or not the original data contained clusters of palindromes that were randomly selected, we generated a number of randomly scattered samples in order to compare their distribution with HCMV. We found no noticeable differences between the distribution of points between HCMV and the 5 random samples.

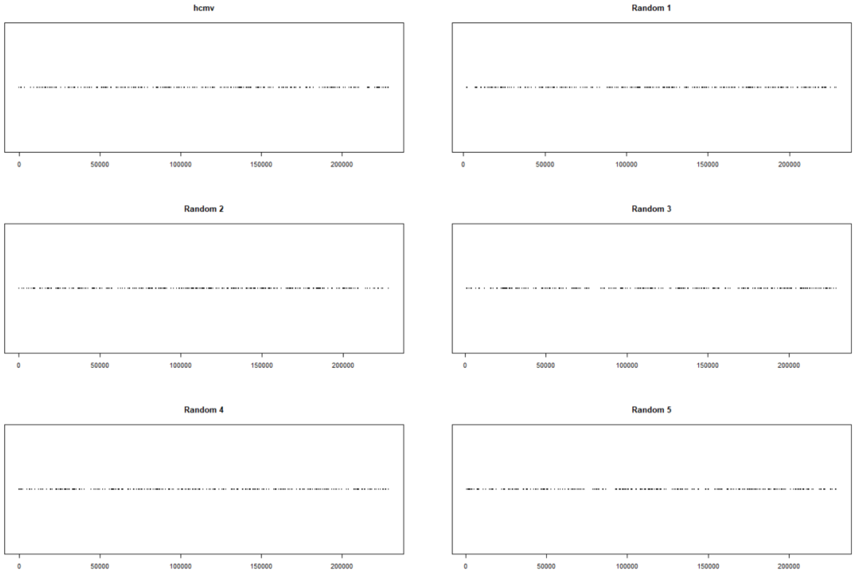


Figure 1: Comparison of Random Scatter with Original Data

3 Location and Spacing Analysis

After performing pairwise and triplet-wise transformations to HCMV, we calculated the difference between each pair and triplet and plotted the Empirical Distribution Function (ECDF). If the ECDF follows the ECDF of a sample generated by a gamma distribution, then the data follows the Homogenous Poisson Process(HPP).

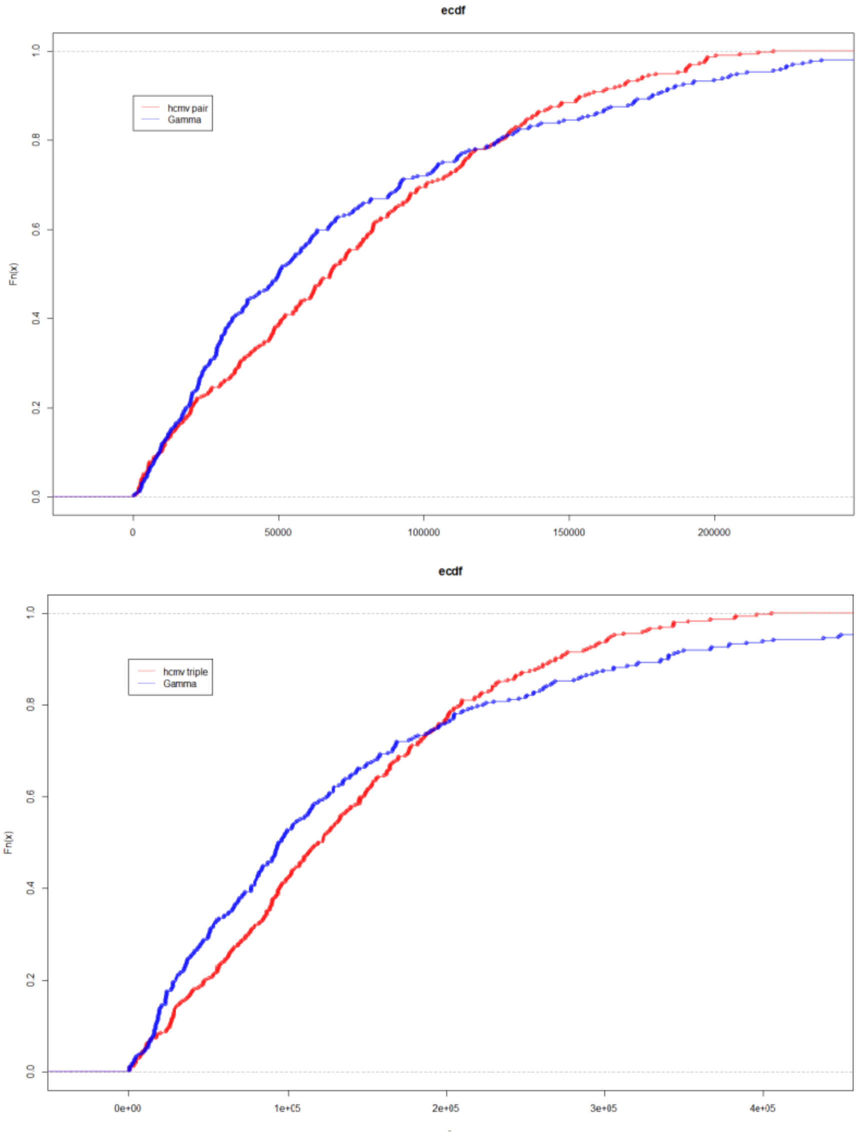


Figure 2: ECDF of the HCMV pair/triplet and The Gamma Distribution

We can observe that the ECDF of HCMV is indeed close to the ECDF of a sample generated by a gamma distribution, therefore HCMV follows the HPP.

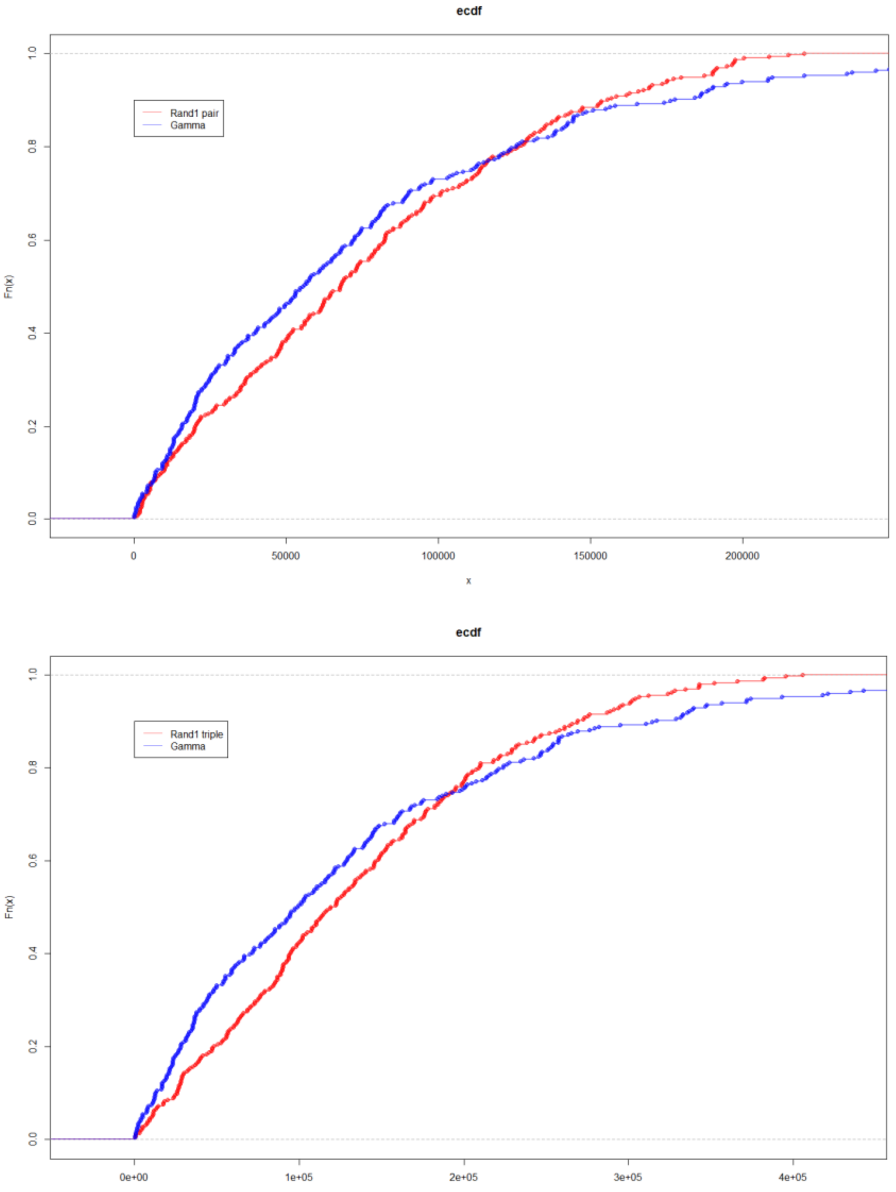
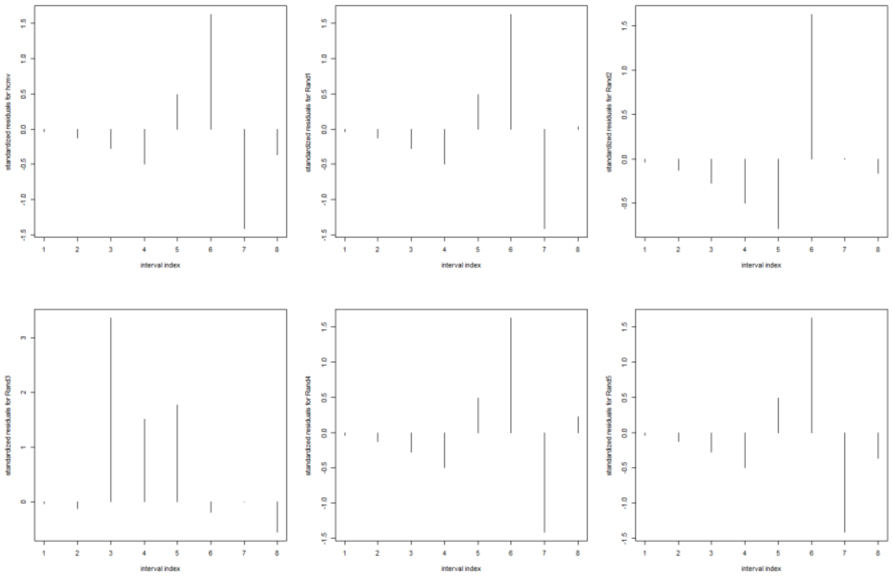


Figure 3: ECDF of the Randomly Sampled pair/triplet and The Gamma Distribution

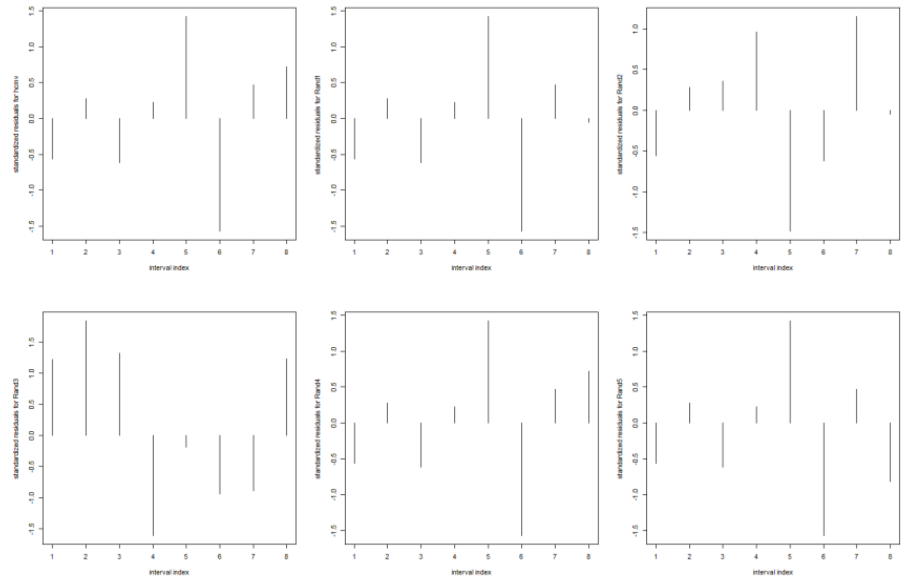
Likewise, we performed the same test on one of our randomly generated samples and this too shows that the sample Rand1 follows the HPP.

4 Palindrome Counts

With 30 intervals of 7000, there were 4 randomized samples including HCMV with a p-value greater than .05 after performing the chi-square test. For rand3 however resulted in a p-value less than .05 which is most likely due to chance because this sample was generated by a pseudo random number generator. For chi-square tests with p-values greater than the significance level shows that the distribution of points within these samples are uniform random scatters, including rand3. On top of the chi-square test, we graphically plotted the residuals of the 6 samples and 5 of the 6 indicated a good fit because the residuals are within 3 standard deviations from the expected number of counts. Rand3 however did not fit this criteria but we already recognize that it is due to chance.

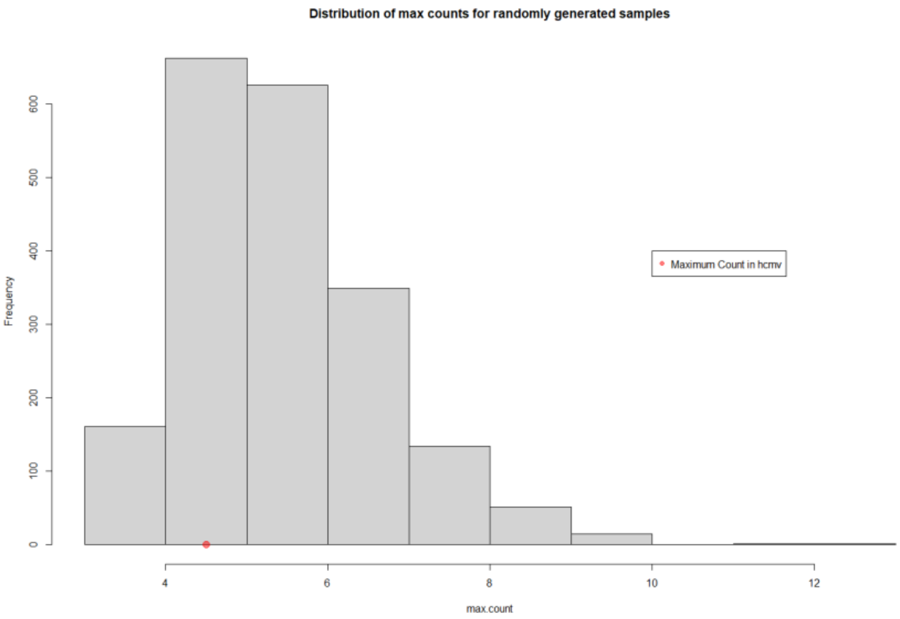


Increasing the number of intervals to 57 (intervals of 4000), the chi-square test again indicated a p-value greater than .05 which includes Rand3 this time around. Therefore, after decreasing the length of each interval yields stronger support that these samples have points that are uniformly scattered.

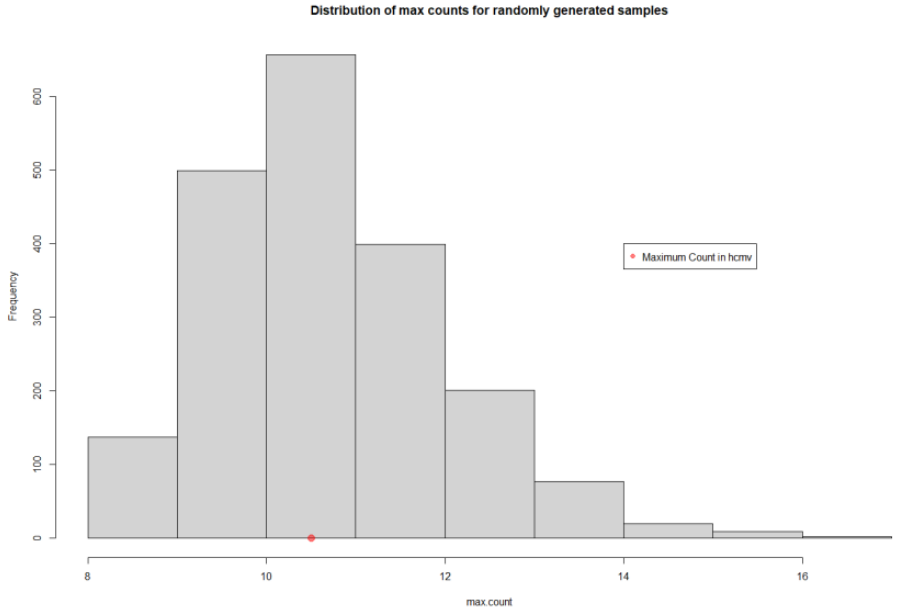


Residuals also support this as the samples counts are within 3 standard deviations from the expected counts.

5 The Biggest Cluster



For 30 intervals across the length of the gene, we simulated 2000 random samples and plotted a histogram of the maximum number of counts out of these intervals for each sample. We then plotted a point of reference for the maximum number of counts from hcmv. We can see that the biggest cluster follows the HPP because the max count is close to the most frequent occurrence of max count from the 2000 samples of uniform random scatters.



For 57 intervals, the result remains the same, therefore the biggest cluster follows the HPP. This tells us that this large cluster is most likely due to chance and is not a region for replication.

6 Chi-Square Goodness of Fit

We used the chi square goodness of fit test in order to determine if the distribution of the spacing in triplets and pairs were a good match for the poisson distribution. To do so, we used the values that we found for the spacing between the pairs and triplets when we plotted the graph using them. We also used the rgamma function to generate data that would be representative of what the data would look like if it followed a hypothetical poisson distribution. Additionally, we only used one of the samples to compare against the original dataset since we found that there were no notable differences between the distribution of points for the random sample. Using the chisq.test function, we found that the p value from the test was 0.2495 for the spacing for both the triplet and pairs. For the random scatter sample, we found that the p-value was 0.2484 for the pairs, and 0.2421 for the triplets. Based on a significance level of 0.001, we can conclude that the random scatter sample and the original data both follow the poisson distribution.

7 Conclusion

If we are able to identify the origin of replication for this virus, we would be able to prevent the virus from growing rapidly. However, based on our findings, we conclude that the origins of replication cannot be found.

After we created several samples of random scatters, we found no noticeable difference between the distribution of points between HCMV and the five random samples. Furthering our analysis, we plotted the ECDF of the differences between the pairs and the differences between triplets for HCMV and rand1. We plotted this against the ECDF of a sample generated using a gamma distribution and found similarities, therefore both samples follow the HPP. We then compared the expected counts and the observed counts using the chi-square test and residuals test and found reasonable evidence that the distribution of counts follows the HPP. Finally, when we identified the biggest count for the original dataset with the largest count of the uniform random scatter, we found that the maximum count for a cluster was similar in frequency.

We can conclude that the DNA sequence of palindrome follows a uniform scatter, and that the clusters of palindromes in the HCMV dataset does not represent a possible origin of replication. Some of the limitations of this dataset were that the HCMV dataset didn't have a large enough sample size to accurately represent the population, and that the dataset only included palindromes that were within a certain range. We could have also tested more intervals to identify discrepancies for the maximum counts between the HCMV dataset and the random sample. If a scientist were to investigate the origins of replication for the human cytomegalovirus, they must understand that an unusual cluster of palindromes will not represent any origin of replication.