Distribution of Genetic Palindromes on Strands of DNA affected by Human Cytomegalovirus (HCMV)

James P. Donohue

April 3, 2022

Abstract

The purpose of this study is to determine the site of replication of human cytomegalovirus (HCMV). The randomness of the distribution of palindromes will be observed in the cases of either a Poisson distribution, or a uniform random distribution. To test the hypothesis that the locations of palindromes in HCMV follow a random distribution, the provided dataset will be observed and contrasted from both these distributions. The accuracy of the data as a Poisson distribution despite a clear and prominent outlier gives credence to the dataset having a uniform random distribution, the Poisson testing yielding a false confirmation. When examining the histogram of the distribution with 45 intervals, it is very clear that the distribution closely follows a uniform random distribution. This idea is further underscored by the normalization of the outlier in the following graph. This closely reflects that of the hypothesis and brings into question whether or not this outlier could be the site of replication for this particular strand of DNA.

Background and Significance

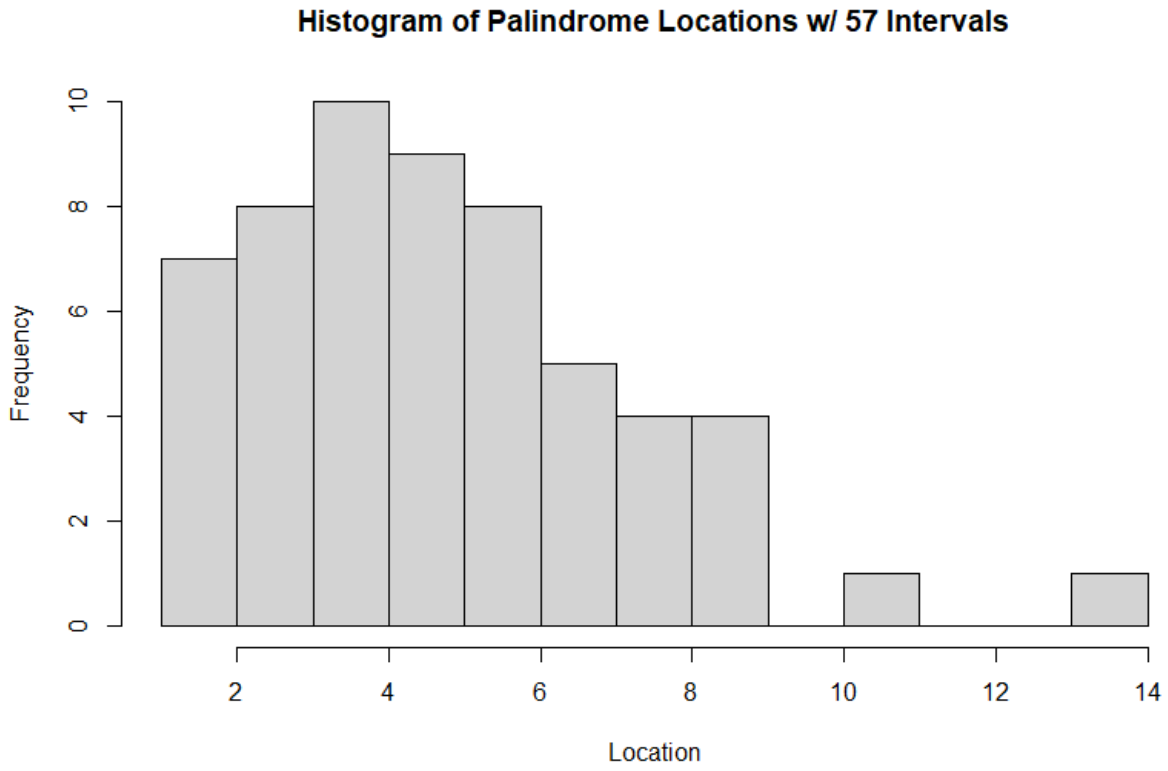
The data consists of the locations of genetic palindromes as found on strands of DNA, as a result of human cytomegalovirus (HCMV). The aforementioned virus is documented as being connected to genetic sequences that form complementary palindromes, noted as sites of replication for HCMV. Although they may be found throughout the body, HCMV infections are frequently associated with the salivary glands. HCMV infection is typically unnoticed in healthy people, but can be life-threatening for the immunocompromised, such as HIV-infected persons, organ transplant recipients, those with hyperglycemia, or newborn infants. In determining the location of the site of replication, and analyzing the distribution of palindromes, great strides could be made in the areas of viral entry and replication detection, supplementing practices such as fluorescence microscopy.

Methods

To test the hypothesis that the locations of palindromes in HCMV follow a random distribution, the provided dataset will be observed and contrasted from both Poisson and uniform random distributions. Subsequently, chi-squared tests will be conducted to determine the goodness of fit of both distributions. To test for a Poisson distribution, the data is split into a set of uniform intervals. The average number of palindromes per interval is then calculated. With this average, the probability of a particular number of palindromes can be extrapolated, and the goodness of fit of the distribution then verified using a chi-squared test. This goodness of fit will be used to compare the expected number of palindromes per interval with the observed amounts. This process is then repeated with varying interval counts, as to find which amount of intervals best fits each distribution. The entire process is then repeated in the case of a uniform random distribution (bar the comparison of observed to expected counts in the distribution).

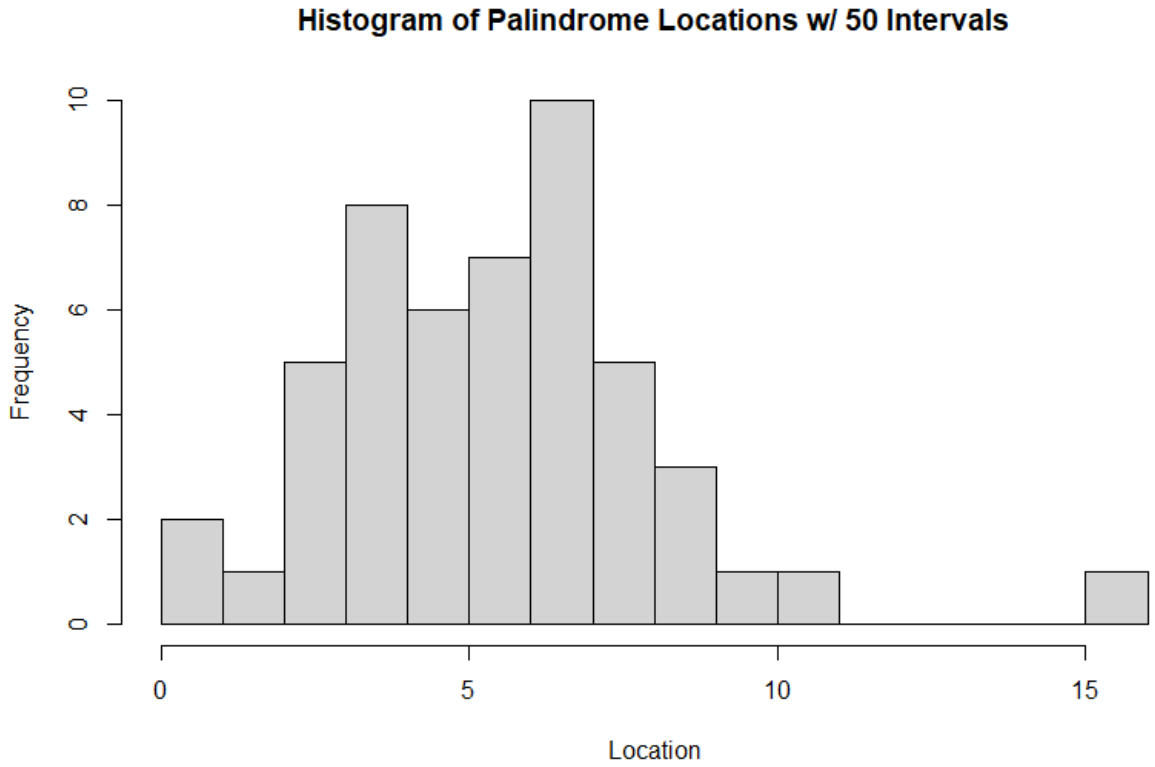
Results

For Poisson Distribution



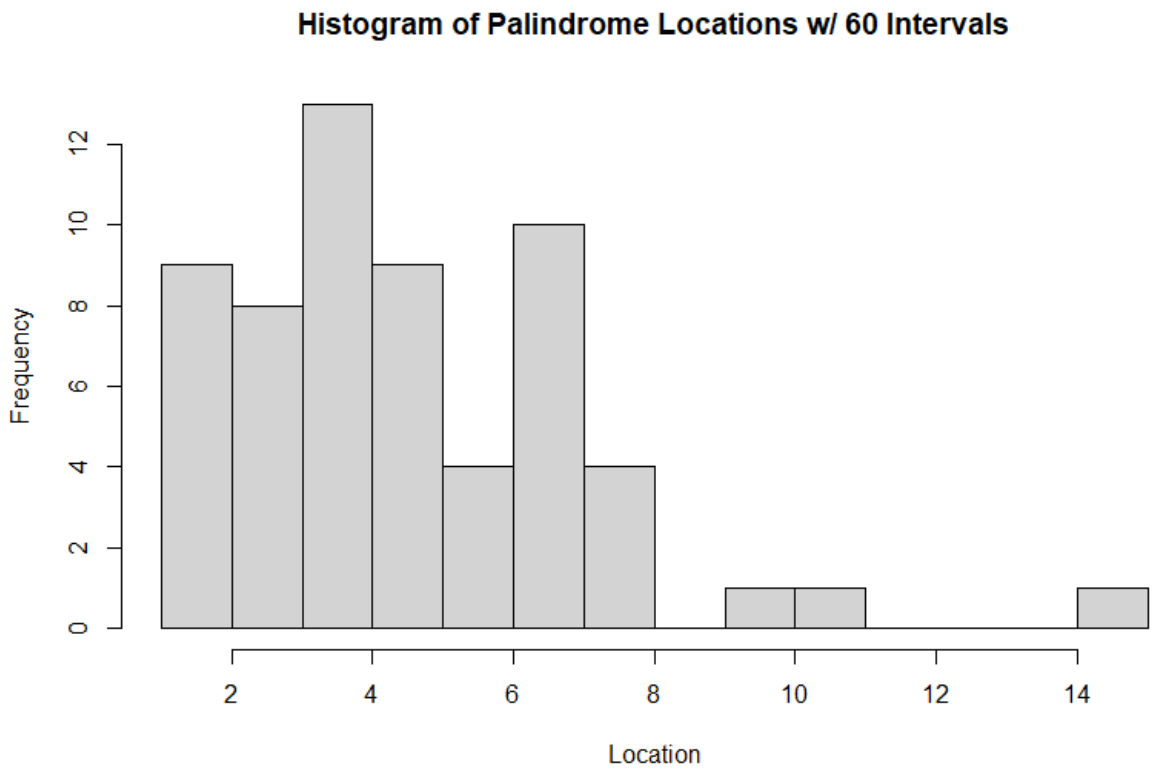
57 Intervals	0-2	3	4	5	6	7	8	9+
Expected	6.38217	7.500597	9.671822	9.977248	8.576932	6.319845	4.074637	4.496744
Observed	7	8	10	9	8	5	4	6

In this case the data was divided into 57 intervals. The chi-squared test for a Poisson distribution yielded a chi value of 1.0182. With 6 degrees of freedom, this yields for us a p-value of 0.9849. As seen above, a calculated “expected” Poisson distribution is compared to that which we observe in the given intervals.



50 Intervals	0-2	3	4	5	6	7	8	9+
Expected	3.28174	4.642592	6.871037	8.135307	8.026836	6.788410	5.023424	7.230649
Observed	3	5	8	6	7	10	5	6

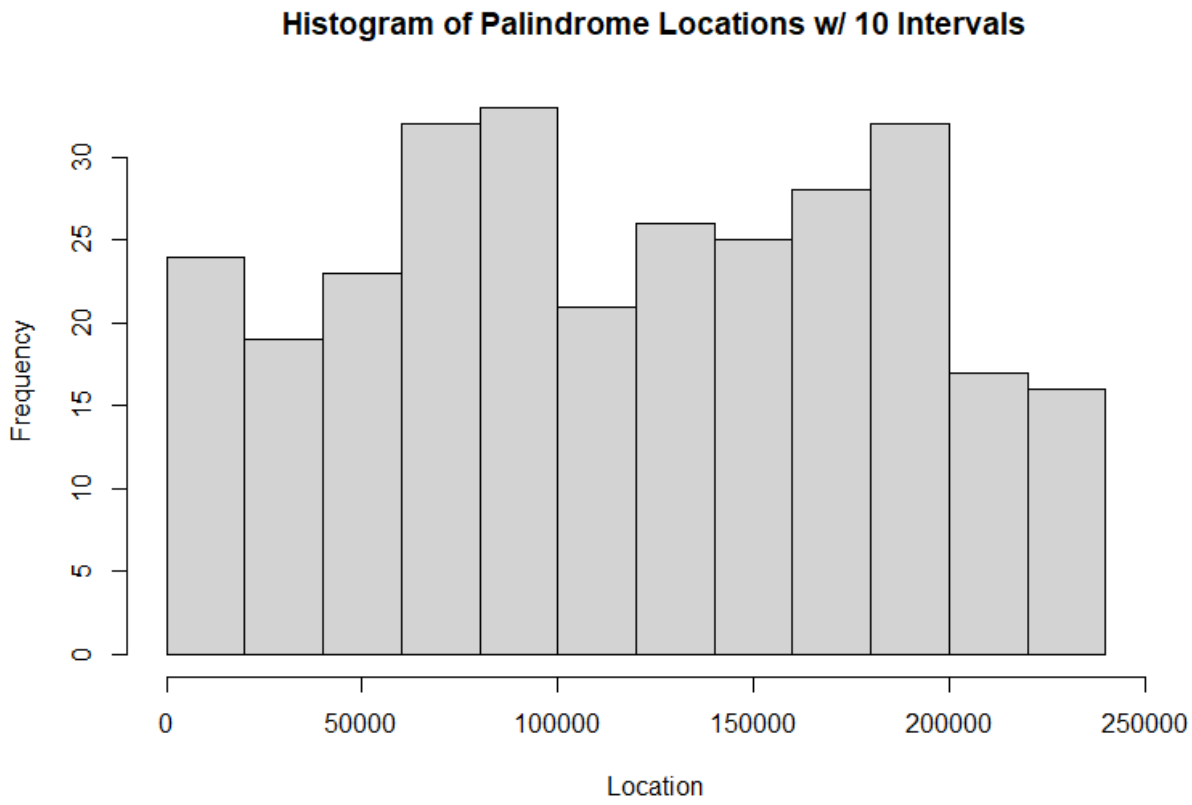
In this case the data was divided into 50 intervals. The chi-squared test for a Poisson distribution yielded a chi value of 2.6579. With 6 degrees of freedom, this yields for us a p-value of 0.8503. As seen above, a calculated “expected” Poisson distribution is compared to that which we observe in the given intervals.



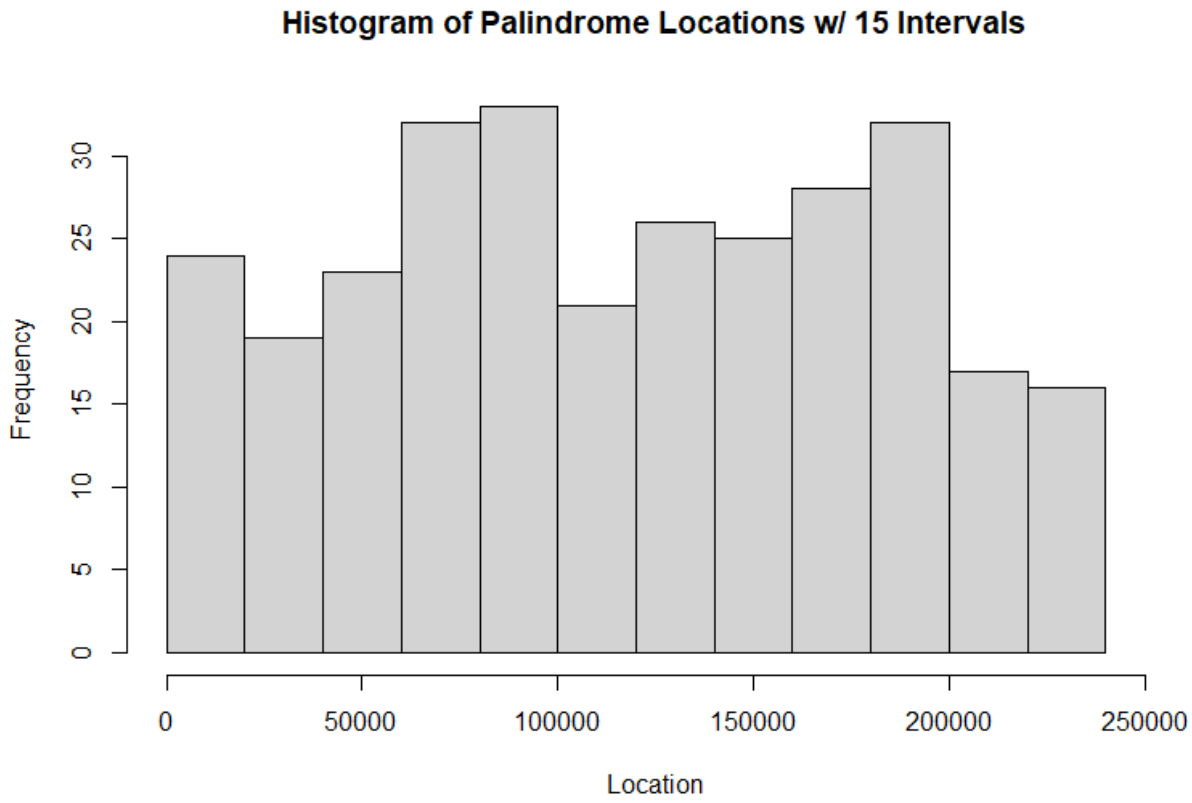
60 Intervals	0-2	3	4	5	6	7	8	9+
Expected	7.99986	8.760831	10.73201	10.51737	8.589191	6.012434	3.682616	3.705670
Observed	9	8	13	9	4	10	4	3

In this case the data was divided into 60 intervals. The chi-squared test for a Poisson distribution yielded a chi value of 6.1476. With 6 degrees of freedom, this yields for us a p-value of 0.4068. As seen above, a calculated “expected” Poisson distribution is compared to that which we observe in the given intervals.

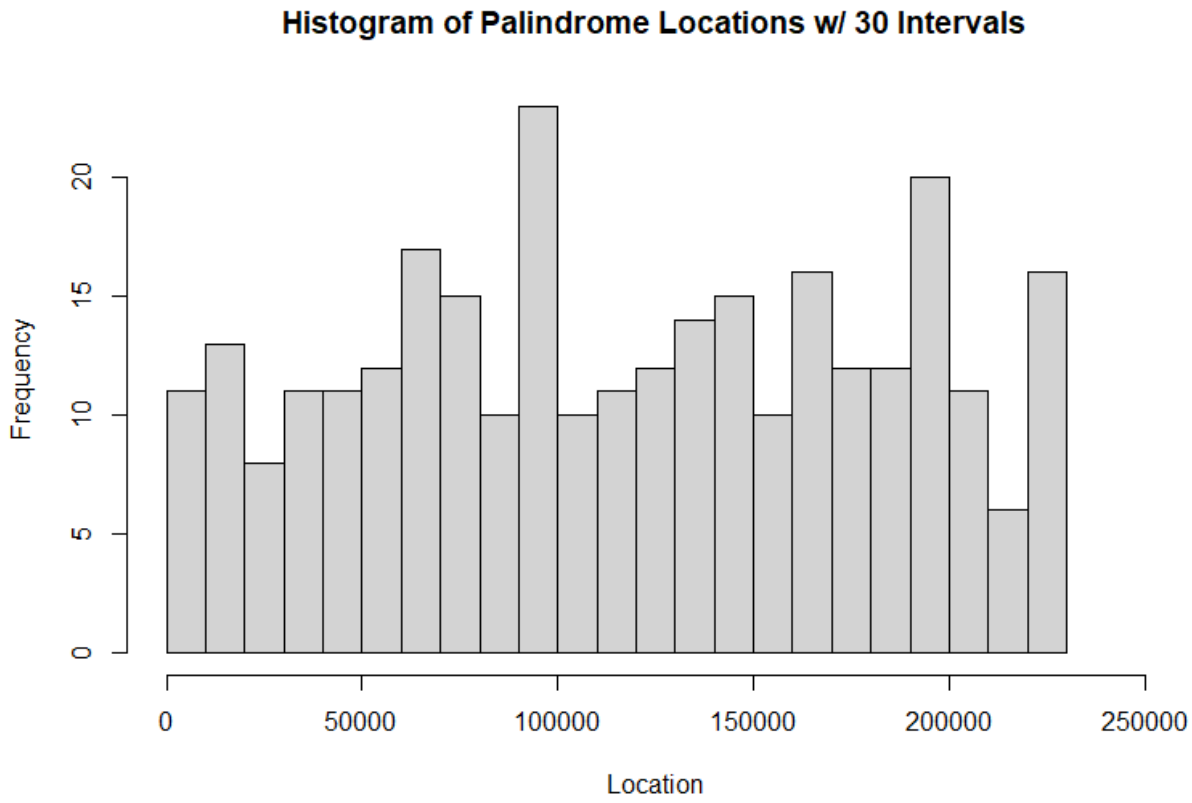
For Uniform Random Distribution



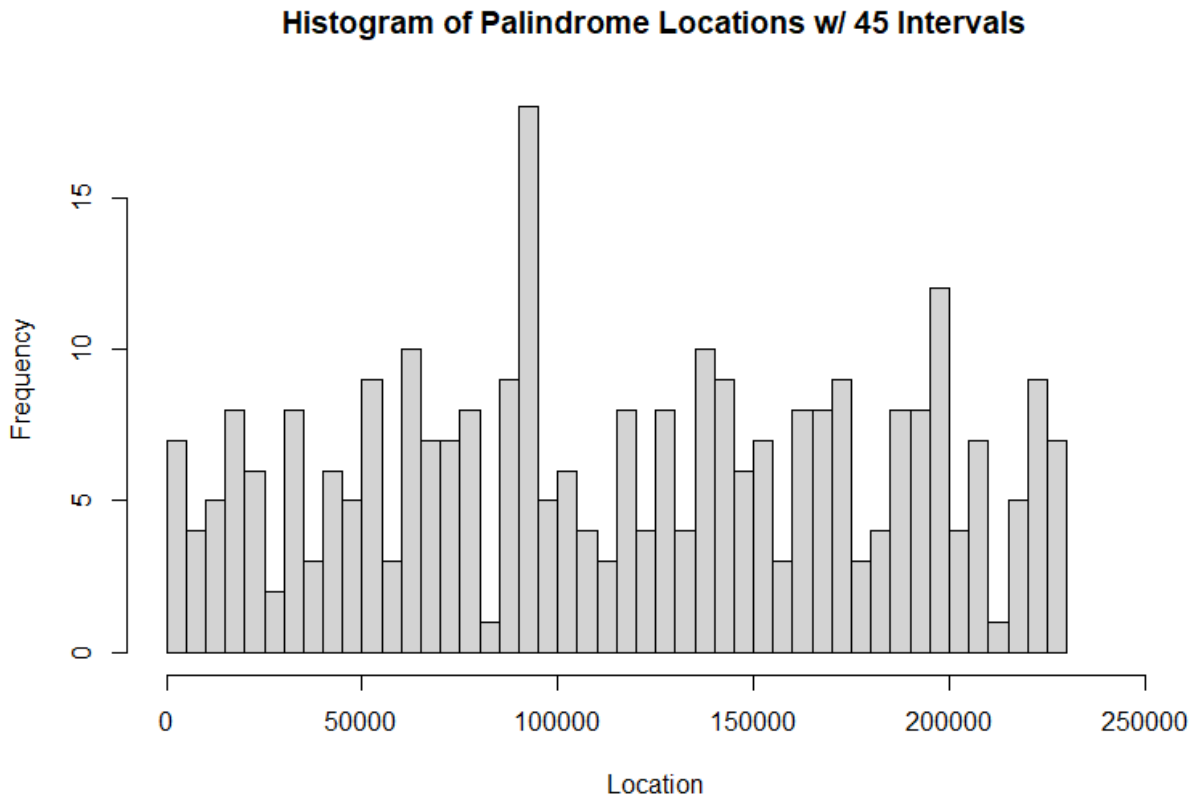
The chi-squared test for a uniform random distribution on 10 intervals resulted in a value of 4.135. With 8 degrees of freedom the resultant p-value is that of 0.8447.



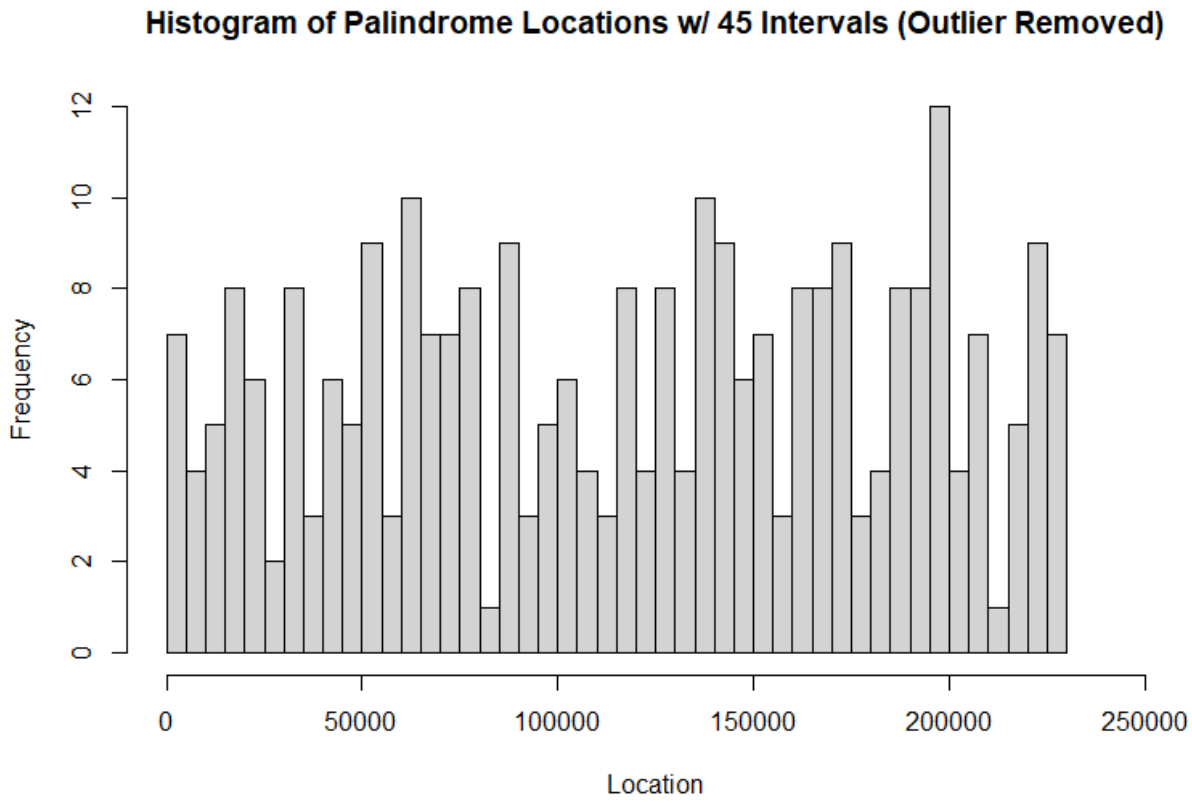
The chi-squared test for a uniform random distribution on 15 intervals resulted in a value of 15.1486. With 13 degrees of freedom the resultant p-value is that of 0.2981.



The chi-squared test for a uniform random distribution on 30 intervals resulted in a value of 40.6891. With 28 degrees of freedom the resultant p-value is that of 0.0573.



The chi-squared test for a uniform random distribution on 45 intervals resulted in a value of 50.3175. With 43 degrees of freedom the resultant p-value is that of 0.2063. Something of note here is that the p-value has slightly increased.



With the outlier normalized, we can see that the distribution more closely follows a uniform random distribution.

Conclusion

The purpose of this study is to determine the site of replication of human cytomegalovirus (HCMV). The randomness of the distribution of palindromes will be observed in the cases of either a Poisson distribution, or a uniform random distribution. The hypothesis of this study being that the palindrome locations would follow one of these distributions, and the site of replication being the outlier of the distribution. It is quite difficult to extrapolate random distributions given large outliers as seen in the above Poisson and uniform random distributions. The accuracy of the data as a Poisson distribution despite a clear and prominent outlier gives credence to the dataset having a uniform random distribution, the Poisson testing yielding a false confirmation. When examining the histogram of the distribution with 45 intervals, it is very clear that the distribution closely follows a uniform random distribution. This idea is further underscored by the normalization of the outlier in the following graph. This closely reflects that

of the hypothesis and brings into question whether or not this outlier could be the site of replication for this particular strand of DNA.