Distribution of Reverse Palindromes in the DNA of Cytomegalovirus

McCord Murray

March 23, 2022

Abstract

The goal of this experiment is to determine the origin of replication in a cytomegalovirus by observing what kind of random distribution it follows and isolating outliers. To test the hypothesis that the locations of palindromes in CMV DNA follow a random distribution, the distribution of the locations will be compared to both a uniform random distribution and a Poisson random distribution. According to the palindrome locations in the CMV DNA, there are 18 palindromes between locations 89590 and 94567. Without these locations, the distribution closely resembles a uniform random distribution. It is likely that the origin of replication is located in this region.

Background and Significance

Nearly all life forms on Earth have to deal with viral infections. One method that viruses will use to infect their host is to modify the DNA of the cell so that it will replicate the virus. One such virus that does this is CMV (Cytomegalovirus). The origin of replication of CMV and other viruses in related families like Herpes simplex and the Epstien-Barr virus are marked by complimentary palindromes in the DNA of the virus. However, the challenge of this is determining which palindromes in the DNA structure are the origin of replication, as palindromes occur randomly in DNA all the time. However, there are many ways to interpret ‘random’. If the palindromes in CMV DNA occur randomly, what sort of random distribution does it follow? If we test different types of random distributions, we may be able to pinpoint the location of the origin of replication of the CMV as we know these particular palindromes do not occur randomly and therefore, should be an outlier in a particular type of random distribution. The hypothesis is that palindromes in CMV DNA follow a random distribution and we will be able to determine the most probable area for the origin of replication based on an outlier in the distribution.

Methods

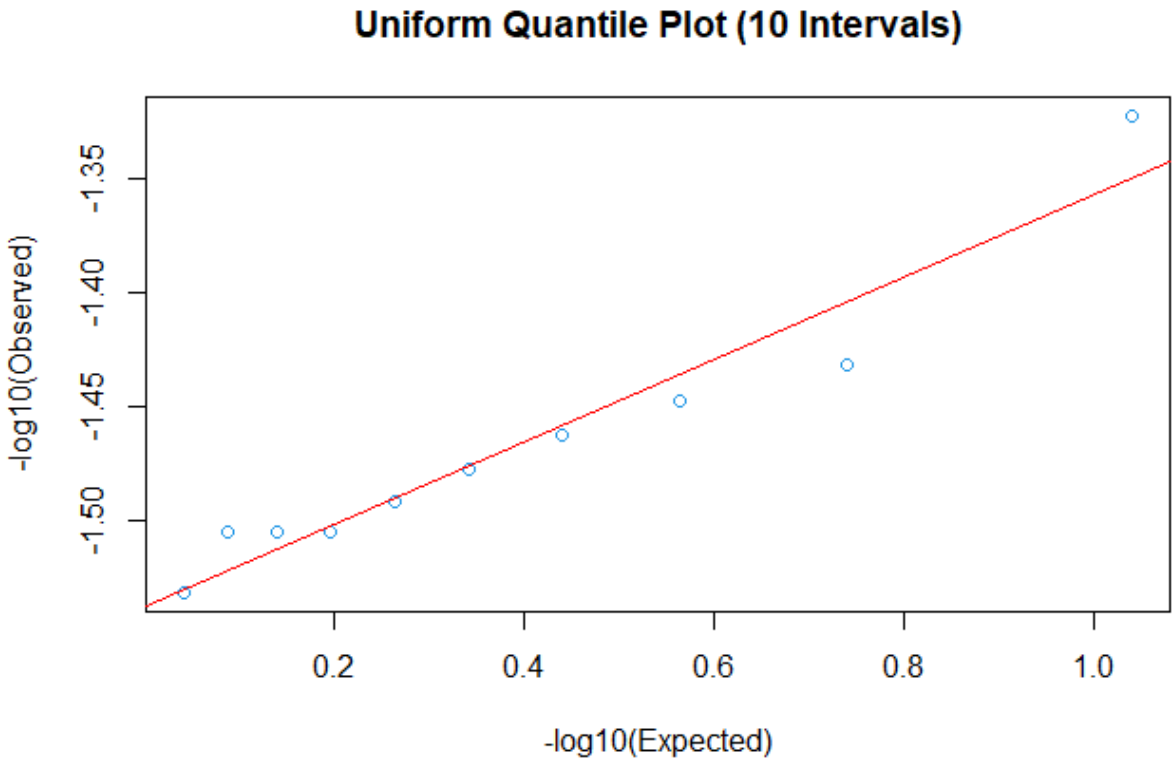
To test the hypothesis that the locations of palindromes in CMV DNA follow a random distribution, the distribution of the locations will be compared to both a uniform random distribution and a Poisson random distribution. Afterwards, a chi-squared test will be used for a “goodness of fit” test for both of these distributions. In order to test for a uniform random distribution, all locations in the CMV DNA will be split into equally sized intervals. Then the number of palindromes in each interval will be compared to the expected value. Finally, a chi-squared test will return a value which will indicate the distribution's “goodness of fit” to the uniform distribution. A similar process will be done for the Poisson distribution. In this case, λ (the average number of palindromes per interval) is unknown, so this value will have to be calculated. Once this is done, the probability of a certain number of palindromes occurring in an interval can be calculated, and a chi-squared test will be used to evaluate its “goodness of fit” for how many intervals with a certain number of palindromes were observed versus how many were expected. This same process will be repeated with different interval sizes. Repeating the process is important because the distribution of one interval size may just so happen to fit some random distribution well, but another interval size may not fit the distribution at all.

Results

After splitting all 229354 locations into 10 intervals of equal size, the amount of palindromes per interval was found.

Interval #	1	2	3	4	5	6	7	8	9	10
Palindrome Count	29	21	32	30	32	31	28	32	34	27
Expected #	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.6	29.6

The chi-squared test for a uniform distribution resulted in a chi value of 4.14, which with 8 degrees of freedom yields a p-value of 0.84. Additionally, a uniform quantile plot was constructed to compare this distribution with its expected values.

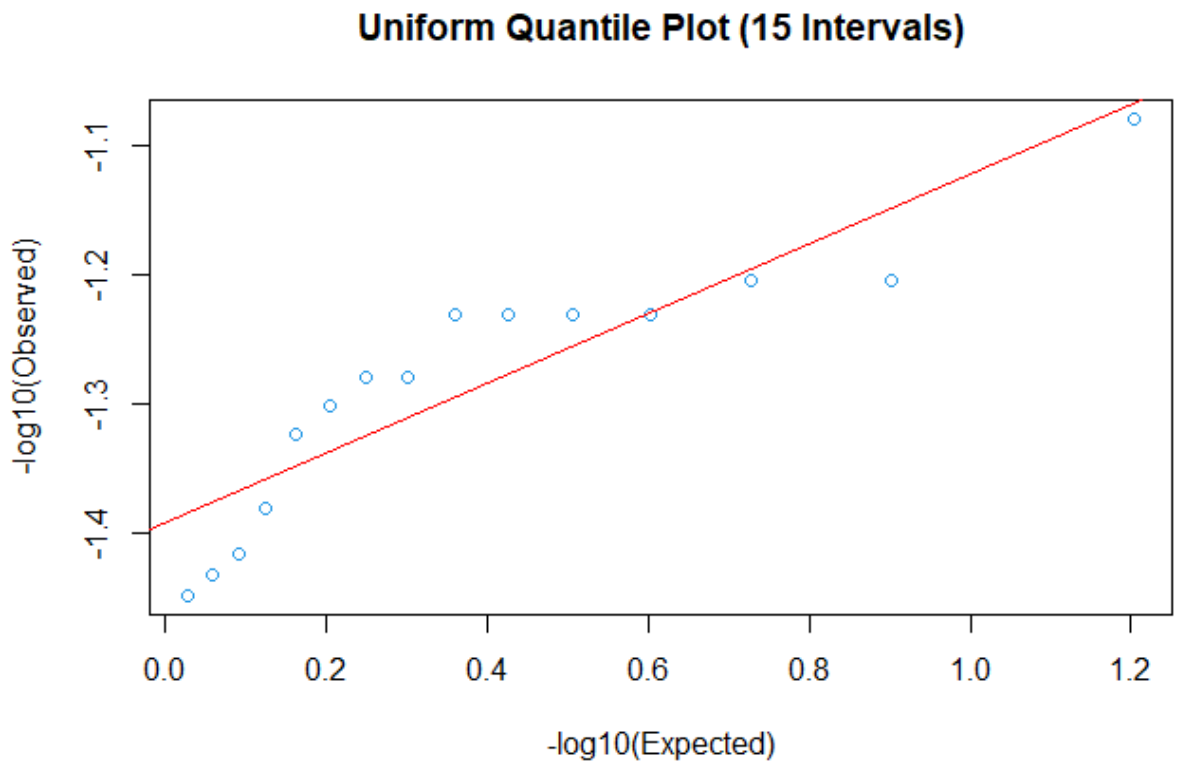


This quantile plot yielded an r-squared value of 0.937.

Next, the same process was done but with an interval size of 15. After splitting all 229354 locations into 15 intervals of equal size, the amount of palindromes per interval was found.

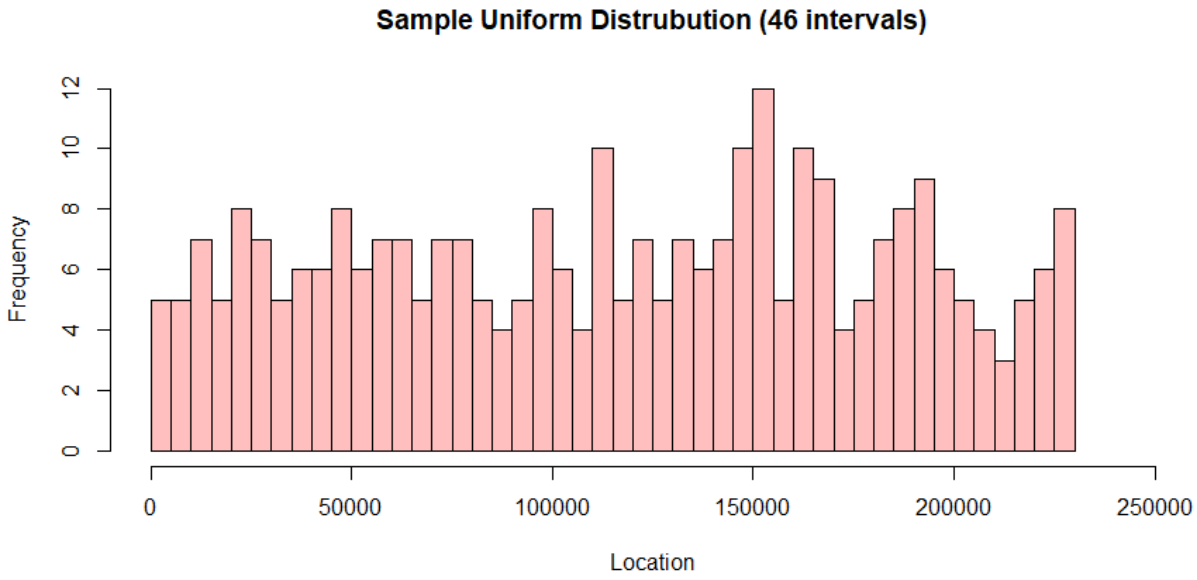
Interval #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Palindrome Count	16	17	17	19	27	16	26	17	20	24	17	19	28	12	21
Expected #	19.7	19.7	19.7	19.7	19.7	19.7	19.7	19.7	19.7	19.7	19.7	19.7	19.7	19.7	19.7

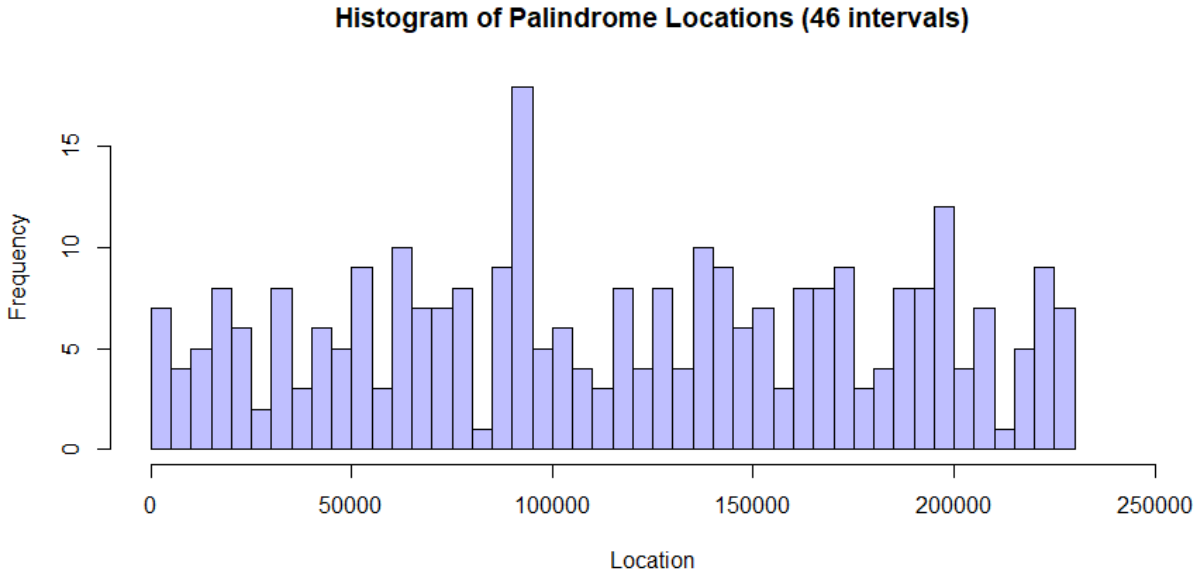
The chi-squared test for a uniform distribution resulted in a chi value of 15.15, which with 13 degrees of freedom yields a p-value of 0.30. Additionally, a uniform quantile plot was constructed to compare this distribution with its expected values.



This quantile plot yielded an r-squared value of 0.824.

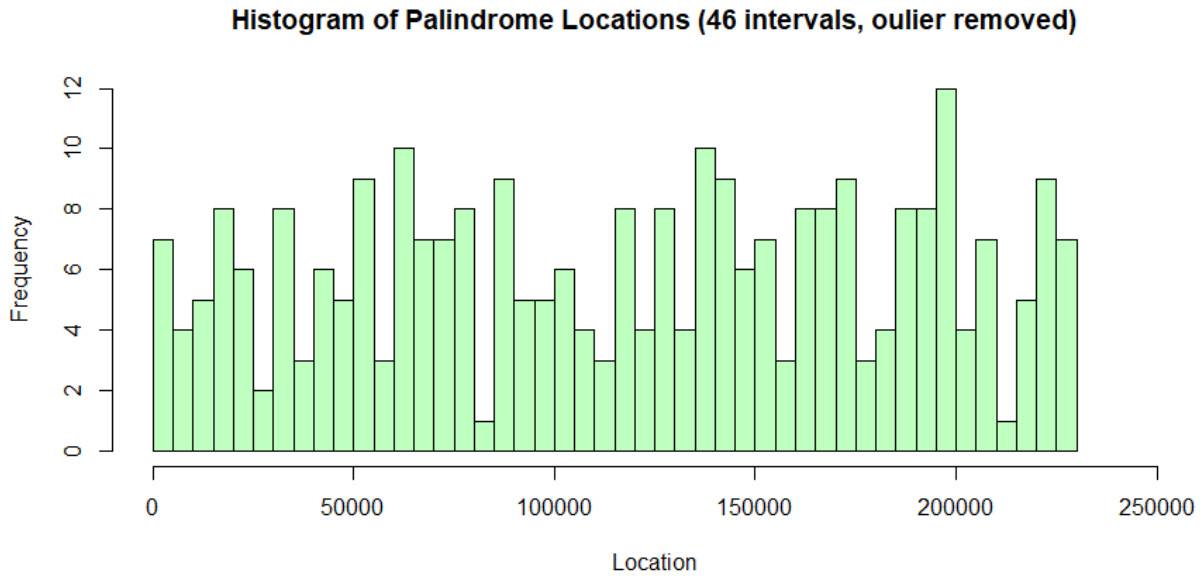
A sample uniform distribution was constructed and compared to the observed number of palindromes for all 229,000 locations separated into 46 intervals.





The chi-squared test for a uniform distribution resulted in a chi value of 62.05, which with 44 degrees of freedom yields a p-value of 0.04. According to this histogram there are 18 palindromes between locations 89590 and 94567.

This is what the previous distribution looks like with the outlier normalized:

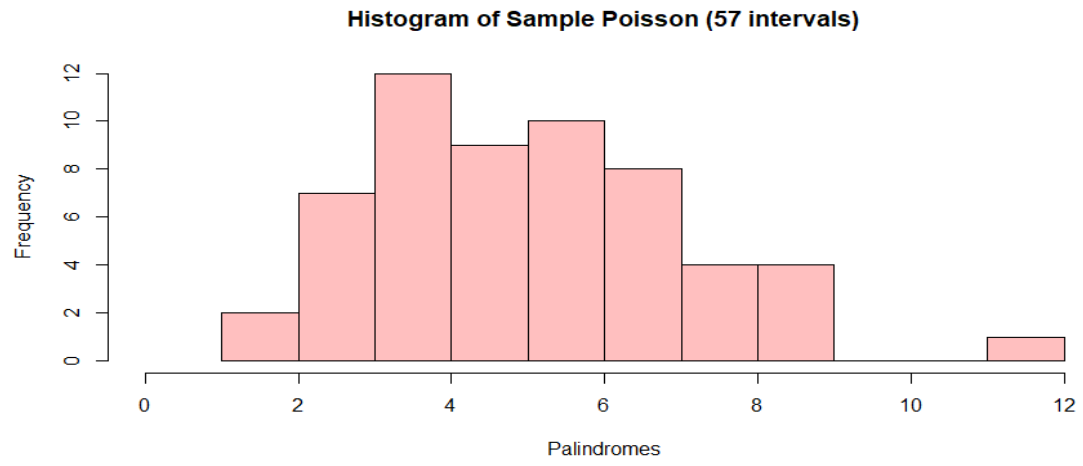


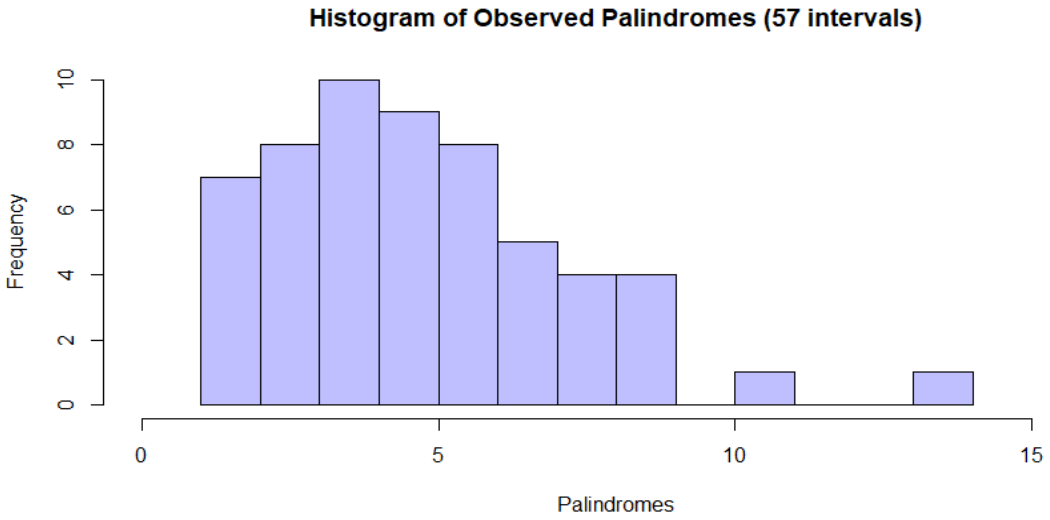
Moving on to a Poisson distribution. This time, data was divided into 57 intervals of size 4000. These intervals did not include all 296 palindromes (two were excluded). The value λ was calculated to be 294/57 = 5.16. These were the number of palindromes for each of the 57 intervals:

7 1 5 3 8 6 1 4 5 3 6 2 5 8 2 9 6 4 9 4 1 7 7 14 4 4 4 3 5 5 3 6 5 3 9 9 4 5 6 1 7 6 7 5 3 4 4 8 11 5 3 6 3 1 4 8 6

Palindrome Count	0-2	3	4	5	6	7	8	9+
Observed	7	8	10	9	8	5	4	6
Expected	6.4	7.5	9.7	10.0	8.6	6.3	4.1	4.5

The chi-squared test for a Poisson distribution resulted in a chi value of 1.02, which with 6 degrees of freedom yields a p-value of 0.98. A sample Poisson distribution was compared to the observed distribution of palindrome locations per interval.





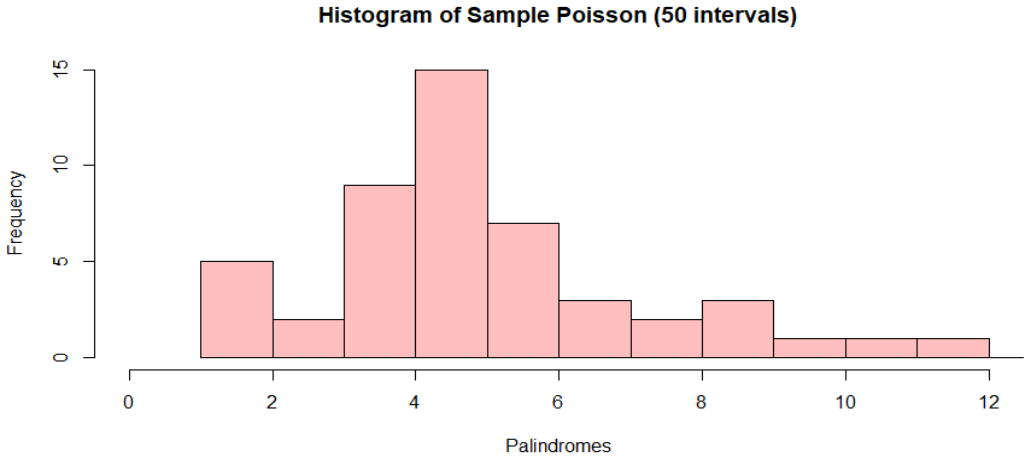
The probability of an interval with 14 palindromes in a Poisson distribution where λ = 5.16 is 0.0006.

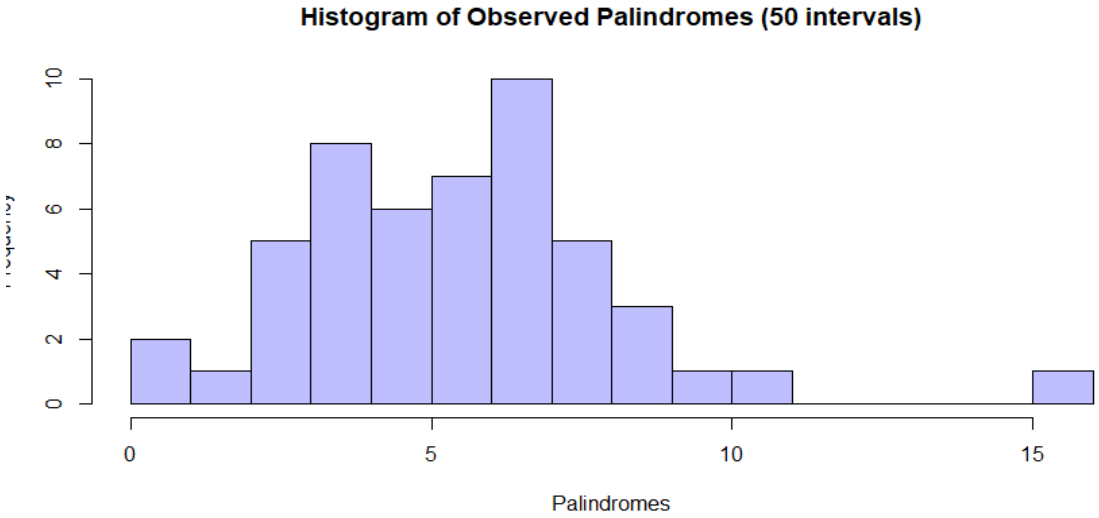
This process was repeated for 50 intervals with a size of 5800. This time all 296 palindrome locations were included in the distribution. The value λ was calculated to be 296/50 = 5.92. These were the number of palindromes for each of the 50 intervals:

7 3 4 7 7 3 4 6 4 5 4 9 3 9 7 5 10 1 7 6 16 4 6 4 3 5 6 6 5 8 6 7 7 7 2 8 8 5 7 4 3 9 11 7 4 6 0 5 8 8

Palindrome Count	0-2	3	4	5	6	7	8	9+
Observed	3	5	8	6	7	10	5	6
Expected	3.3	4.6	6.9	8.1	8.0	6.8	5.0	7.2

The chi-squared test for a Poisson distribution resulted in a chi value of 2.66, which with 6 degrees of freedom yields a p-value of 0.85. A sample Poisson distribution was compared to the observed distribution of palindrome locations per interval.





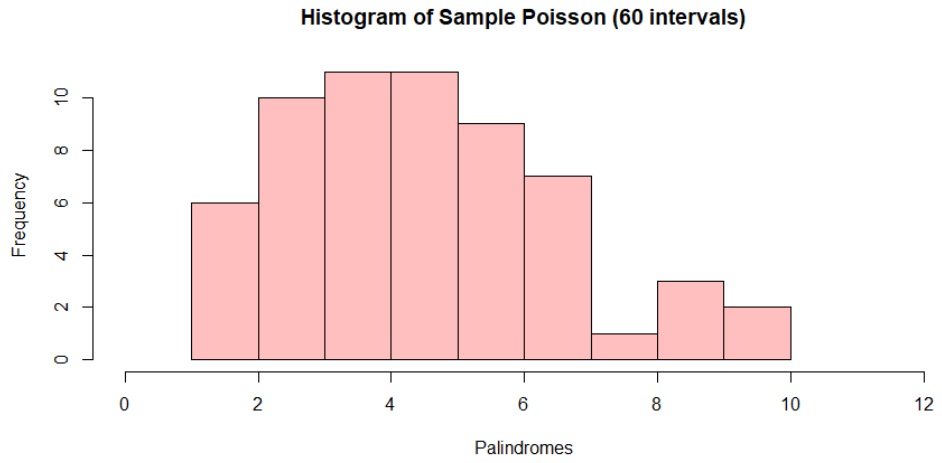
The probability of an interval with 16 palindromes in a Poisson distribution where $\lambda = 5.92$ is 0.0003.

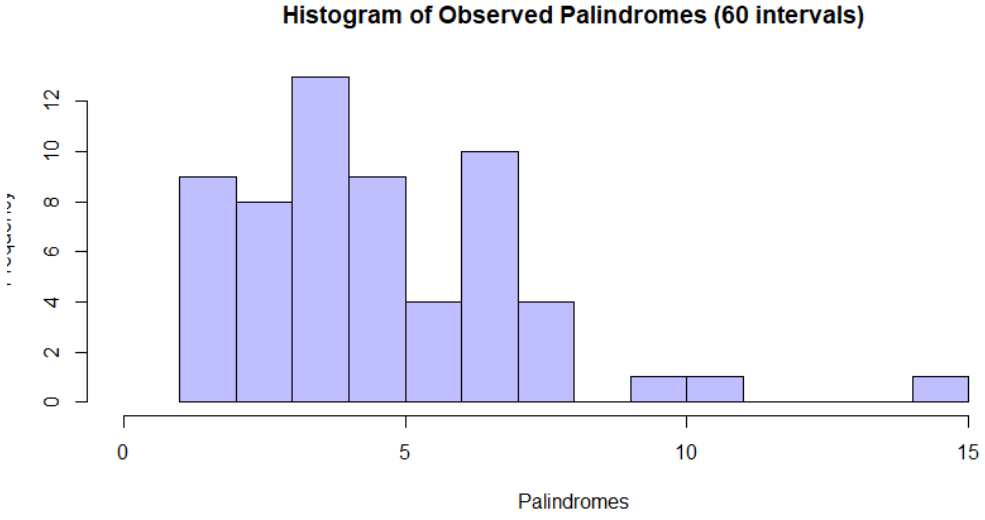
This process was repeated one final time for 60 intervals with a size of 4833. The value λ was calculated to be 296/60 = 4.93. These were the number of palindromes for each of the 60 intervals:

7 1 4 4 5 7 3 2 4 5 4 4 4 4 8 3 7 6 4 10 2 1 7 5 15 4 5 3 4 2 4 7 2 7 3 7 6 7 6 6 3 1 8 5 7 5 3 4 3 5 8 11 4 3 5 1 5 1 8 7

Palindrome Count	0-2	3	4	5	6	7	8	9+
Observed	9	8	13	9	4	10	4	3
Expected	7.8	8.6	10.7	10.5	8.7	6.1	3.8	3.8

The chi-squared test for a Poisson distribution resulted in a chi value of 6.15, which with 6 degrees of freedom yields a p-value of 0.41. A sample Poisson distribution was compared to the observed distribution of palindrome locations per interval.





The probability of an interval with 14 palindromes in a Poisson distribution where $\lambda = 4.93$ is 0.0004.

Conclusion

The goal of this experiment was to analyze the distribution of the locations of reverse palindromes in a cytomegalovirus to determine if they followed a uniform or poisson random distribution. The hypothesis was that the palindrome locations would follow one of these distributions and the origin of replication could be detected as an outlier in the distribution. In hindsight, it did not make much sense to predict the locations would fit a random distribution in addition to having a large outlier. This is why it was unusual that the locations split into 57 intervals of 4000 fit a Poisson distribution so well, especially since one of those intervals has 14 palindromes which has a 0.0006 probability of occurring with a λ of 5.16. Considering this particular distribution fits the Poisson distribution so well when there is non-random data means that the palindrome locations in the CMV DNA most likely do not follow a Poisson random distribution. However, when examining the histogram of the palindrome locations with 46 intervals appears close to the sample uniform random distribution with the exception of the one large interval. This sort of distribution is exactly what was expected from the hypothesis. This outlier with 18 palindromes occurs between locations 89590 and 94567. If the hypothesis is accurate, then the origin of replication is in this region of the CMV's DNA.