Case Study: Patterns in CMV DNA

May 17, 2020

Cameron Thomas
Chase Oden
Duncan Carlmark

I. Introduction

In order to cure dangerous viruses like the human cytomegalovirus (CMV), scientists first determine how a virus replicates and then use this information to develop a process drug that prevents the virus from being able to replicate. The issue with this is that the process of locating the site of replication for a virus is arduous and time consuming. Depending on the size of a virus's genome there can be thousands and thousands of different genome segments that could be the source of replication. This would mean that a scientist would have to test every single one of these segments until the correct one is found. Luckily, there are some methods by which this search can be shortened. CMV is from the Herpes family of viruses, so information from other viruses in this family and their replication sites could be used to narrow down where the replication site for CMV resides. Based on the Herpes simplex and the Epstein-Barr viruses, the replication site for CMV may be associated with the presence of complimentary palindromes as both the replication sites for these viruses are associated with either long palindromes or clusters of palindromes. Using this information we can examine the presence of palindromes in the CMV genome to locate any statistical irregularities that could clue us in on the source of replication.

To investigate statistical irregularities we analyzed a variety of metrics associated with the palindromes. This includes the distribution of palindromes and their relation to the Uniform distribution, the spacing between palindromes, the counts of palindromes over various intervals across the genome, and the largest cluster of palindromes in the genome. Through this analysis, if any irregularities or patterns are detected, these detections could provide very useful information as to where the site of replication is located.

II. Background

The Cytomegalovirus (CMV) is a potentially life threatening disease for people that have dysfunctional immune systems. It's a virus that goes into people's bodies and replicates using a strain of DNA. This strain is what we want to look at and analyze to try and identify how the virus reproduces. This information is stored at the origin of the virus' replication. Identifying the origin is a difficult task, as the DNA sequence of a virus is very long and is difficult to find any distinct patterns in. However, we do know that the origin of

the CMV is similar to other herpes based viruses[1], and as such, its origin is identifiable through a complimentary palindrome within the sequence of its DNA. A sequence of DNA is made up of four distinct neurotransmitters, which we can then look at to see where palindromes are occuring. If we operate under the belief that the origin of the disease can be identified by complementary palindromes, then we're likely to see that there are more of these complementary palindromes surrounding the location of the virus' origin. By narrowing in on this location, we can hopefully determine what the virus needs to operate with much greater efficiency than if we were to analyze the entire DNA sequence as a whole.

III. Data

The data we're working with is an analysis of a sequence of the Cytomegalovirus (CMV). To understand where the virus started, the data looks for gene Palindromes in the sequence, as the origin of replication for the virus is typically marked with complimentary palindromes. In the data, it shows where palindromes occur in the gene sequencing of the virus, however it only keeps the locations of palindromes that are longer than 10 letters. This is to minimize the amount of time searching through the sequence so that the source of the original replication can be found as quickly as possible.

IV. Theory

Goals

The goal is to understand the statistical model that describes "counts" of the number of palindromes and "uniformity" of random distribution of palindromes. We want to determine the estimation procedure in the model. Additionally, we plan to find statistical discrepancies between a model with clusters and a model without clusters. The questions we will answer include whether the model is a good model, what is a hypothesis test, and how is the uniform distribution related to our problem.

The Homogeneous Poisson Process

The Homogeneous Poisson Process is a process that arises naturally from the notion of points haphazardly distributed on a line with no obvious regularity. There are three characteristic features of the process: (1) homogeneity: the rate λ at which hits occur will never change with location, (2) independence: the number of hits falling in different intervals are independent, and (3) no two hits can occur at the exact same location.

The counts of the number of points in different intervals follow a Poisson distribution with rate λ, which represents the rate of hits per unit. We have $P(k \text{ points in a unit interval}) = \frac{\lambda^k}{k!}e^{-\lambda}$ for k=0,1,... and $P(k \text{ points in an interval of length } t) = \frac{\lambda t^k}{k!}e^{-\lambda t}$. The expected value of the number of hits per unit interval is λ. In most cases, we don't know the exact λ. A good estimate would

be the empirical average number of hits per unit interval, which we could get by either method of moments or maximum likelihood.

In this study case, we treat the strand of DNA as a line, and the location of a palindrome as a point on the line. According to the uniform random scatter model, palindromes are scattered randomly and uniformly across the DNA, which meets homogeneity. Also, the number of palindromes in any small piece of DNA is independent of the number of palindromes in another and none of any two hits occur at the same point on the DNA.

Chi-Square Goodness of Fit Test

In this study, we want to use the Homogeneous Poisson Process as a reference model to seek some unusual clusters of palindromes. However, we need to estimate how likely the Poisson distribution fits the data. A technique for accessing how well the reference model fits to the data is to apply the chi-square goodness of fit test.

Sometimes a parameter of the distribution needs to be estimated in order to compute the probabilities. In this case, we use our data to estimate unknown parameter(s). The measure of discrepancy between the sample counts and the expected counts is $\sum_{j=1}^{m} \frac{(j^{th}\ sample\ count - j^{th}\ expected\ count)^2}{j^{th}\ expected\ count} = \sum_{j=1}^{m} \frac{(N_j - \mu_j)^2}{\mu_j}$, where m represents the number of categories (intervals), N_j stands for the number of observations that appear in category i,j=1,…,m and $\mu_j = n * P(an\ observation\ is\ in\ category\ j)$. The discrepancy we get follows an approximate chi-square distribution with m-k-1 degrees of freedom, where k is the number of parameters we estimated to obtain expected values. χ^2_{m-k-1} is a continuous distribution on the positive real line and the density has a long right tail. As the degrees of freedom increase it starts to look symmetric and a lot like normal. We use χ^2_{m-k-1} to get the p-value and if the p-value is less than significance level α, we need to doubt the fit of distribution.

Locations and the Uniform Distribution

Under the Poisson process model for random scatter, if the total number of hits in an interval is known, then the positions of the hits are uniformly scattered across the interval. In other words, the Poisson process on a region can be viewed as a process that first randomly generates the number of hits, and then generates locations for the hits according to the uniform distribution. Here, the positions of these palindromes are similar to 296 independent observations from a uniform distribution so we can apply another Chi-Square Goodness of Fit Test.

Exponential and Gamma Distributions

Distances between successive hits should follow an exponential distribution.

$$P(the\ distance\ between\ the\ first\ and\ second\ hits\ >\ t) = P(no\ hits\ in\ an\ interval\ of\ length\ t) = e^{-\lambda t}$$

Distances between the hits that are two apparatus, follows a Gamma distribution with parameters 2, λ.

Clusters and Maximum Number of Hits

Under the Poisson Process Model, the number of hits in a set of non-overlapping intervals of the same length are independent observations from a Poisson distribution. This implies that the greatest number of hits in a collection of intervals behaves as the maximum of independent Poisson random variables. If we suppose that there m such intervals, then

$$P(\textit{maximum count over m intervals } \geq k) = 1 - P(\textit{maximum count over m intervals } \leq k)$$

$$= 1 - P(\textit{all interval counts } < k) = 1 - P(\textit{first interval counts } < k)^m = 1 - [\lambda^0 e^{-\lambda} + ... + \tfrac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}]^m$$

For a given estimate of λ, from the above expression, we can find the approximate chance that the greatest number of hits is at least k. If this chance is unusually small, then it provides evidence for a cluster that is larger than expected from the Poisson process. We can use the maximum palindrome counts as a test statistic, and the computation above provides the p-value for the test statistic.

Parameter Estimation

- Method of Moments (MME)
  The method of moments is a way to estimate population parameters, like the population mean or the population standard deviation. For a Poisson distribution with unknown rate parameter λ, proceed as follows. Find E(X) where X has Poisson distribution with rate λ. Express λ in terms of E(X). Replace E(X) with $\overline{x}$ to produce an estimate of λ, called $\widehat{\lambda}$. Then, $E(X) = \lambda \rightarrow \overline{X} = \widehat{\lambda}$.
- Maximum Likelihood (MLE)
  The maximum likelihood method is the procedure of finding the value of one or more parameters for a given statistic which makes the known likelihood distribution a maximum. For a Poisson distribution, the chance of observing $X_1, ..., X_n$ is $L_n(\lambda) = \prod_{i=1}^{n} f(X_i; \lambda)$. The maximum is found by taking the derivative of the log likelihood: $l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0$. This results in the estimate $\widehat{\lambda} = \overline{X}$ which matches the MME above. For an exponential distribution, the estimator is $\widehat{\theta} = \frac{1}{\overline{X}}$.
- Mean Square Error
  The mean square error of an estimator $\widehat{\theta}$ for a parameter θ is $MSE(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2]$.
- Asymptotic Distribution
  The Fisher's Information Matrix is defined as $I(\lambda) = E(\frac{\partial}{\partial \lambda} log f_\lambda(X))^2 = -E(\frac{\partial^2}{\partial \lambda^2} log f_\lambda(X))$. Hence, as n increases $\sqrt{nI(\lambda)}(\widehat{\lambda} - \lambda) \sim N(0, 1)$. The approximate normal distribution can be used to build the 95% confidence interval for the unknown λ as $\widehat{\lambda} \pm 1.96\sqrt{nI(\lambda)}$.

Hypothesis Tests

Hypothesis test is a method of inference which refers to the formal procedures used by statisticians to accept or reject statistical hypotheses. In our study, the Chi-Square goodness-of-fit test and the test for the maximum number of palindromes in an interval, are two examples of hypothesis tests. During the test, we first propose a null hypothesis, denoted by H_0, which is usually the hypothesis that sample observations result purely from chance and an alternative hypothesis, denoted by H_A, which is the hypothesis that sample observations are influenced by some non-random cause. Secondly, we select an appropriate test and state the relevant test statistic. Thirdly, we derive the distribution of the test statistic under the null hypothesis from the assumptions and compute from the observations the observed value of the test statistic. With the observed value and the distribution, we can get the p-value for our observations. Finally, we compare the p-value to the significance level (α) and decide to either reject the null hypothesis in favor of the alternative or not reject it.

When we reject the null hypothesis, we don't know if we have been unlucky with our sampling and observed a rare event or if we are making the correct decision. Thus, we have the table below to define the two types of errors we could possibly make in the hypothesis tests.

	Fail to reject H_0	Reject H_0
H_0 true	No Error	Type 1 Error = α
H_A true	Type 2 Error := β	No Error

Table (A): Types of Errors

Typically α is set in advance and β is computed for various values of the alternative hypothesis. The probability of correctly rejecting the null hypothesis is called power. High power is a sign of a good test.

V. Analysis

Investigation #1: Locations

Our investigation begins with a visual analysis of the observed distribution and two non empirical distributions: the random uniform scatter and the theoretical uniform distribution. The purpose of this is to see if there are any obvious irregularities between the observed distributions and the simulated distributions. If there are, we can conclude that the observed distribution does not follow a uniform distribution and therefore any irregularities would be investigated with greater interest. To start we compare the observed distribution of palindromes with the theoretical normal distribution. This is done by representing the observed distribution with a histogram that has bins of length 4,000 and by representing the theoretical normal distribution with a solid continuous line.

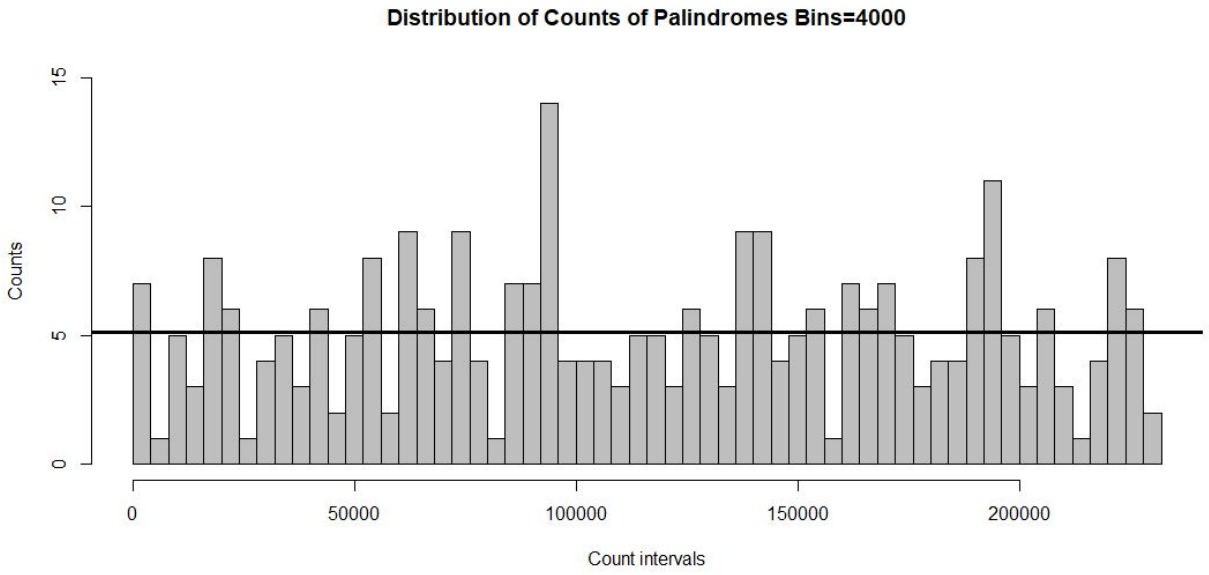


Figure (A): Histogram of palindrome occurrences. Each bin represents the count of palindromes that occur in an interval of 4,000 indexes

On its own, it seems like the observed distribution does not follow a Uniform distribution. There are constant deviations above and below the expected value for a bin and it seems that there are abnormal spikes at more than one place in the distribution. However comparing a distribution to its theoretical shape is not always the most effective way of understanding if the two are related. We also compared 5 random samples of size 293 from a Uniform distribution with the theoretical Uniform distribution and visualized the results.

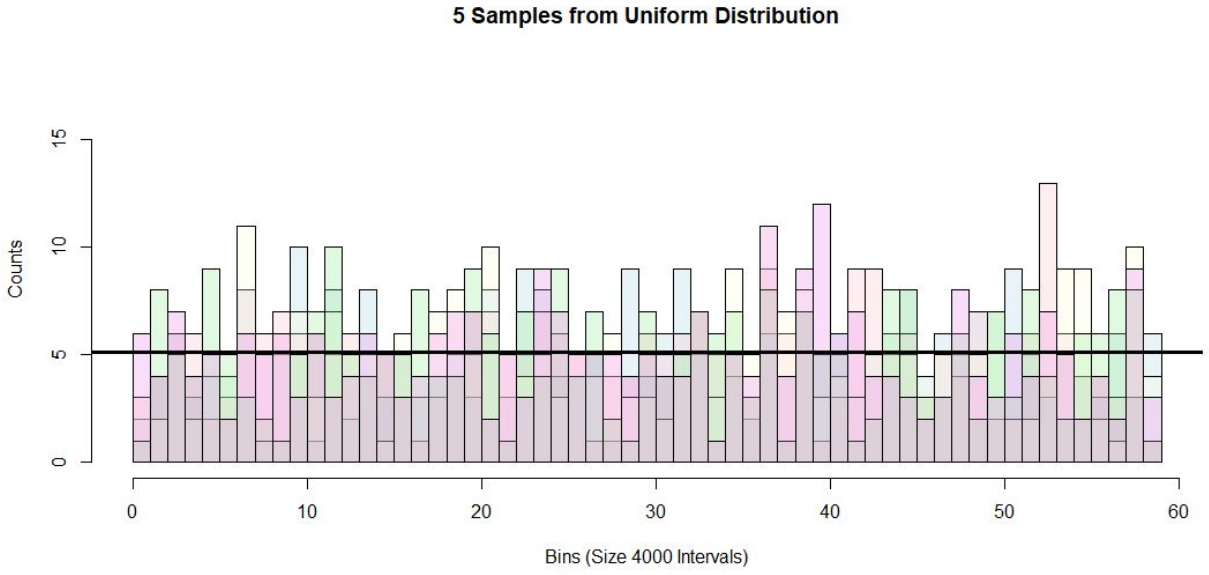


Figure (B): 5 different samples from a Uniform distribution denoted by different transparent colors. Each sample is meant to represent the conditions of the observed distribution.

From this we can see that samples from a Uniform distribution have similar deviations to our observed distribution. Each distribution has its own distinct peaks and valleys so their presence is not necessarily abnormal, and there is not necessarily a pattern in any of these samples either. The peaks are also close to or the same as the peaks in the observed distribution: counts of 14 or less. After analyzing the distributions in this manner we cannot confidently conclude on a purely visual basis that the distribution of palindromes is not Uniform or that the outliers in the distribution have any statistical significance.

Investigation #2: Spacings

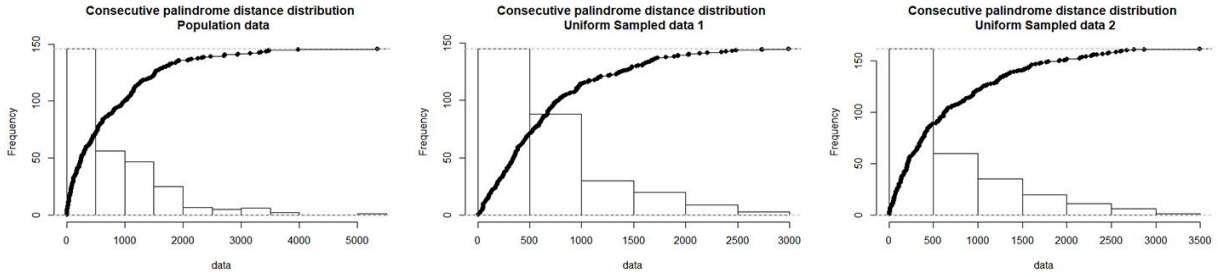


Figure (C): Consecutive palindrome spacing distributions with uniform sampled data comparisons

In this first figure, we can see that there is a generally similar curve for the CDF of both the population and uniformly sampled distributions, even though the histograms tell different stories. Looking at the spaces between consecutive palindromes it seems that there's a higher overall density in the population data than in the other two uniform sample distributions.

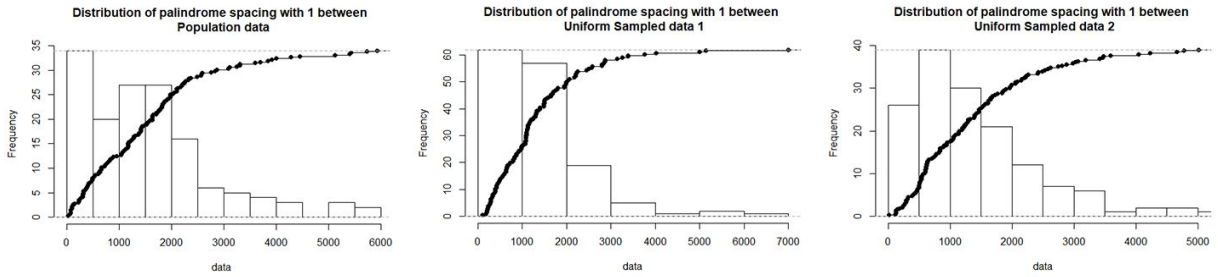


Figure (D): Distribution of single spaced palindrome spacings with uniform sampled data comparisons

In this second figure, the slope of the population's CDF line has straightened out to be more linear rather than quadratic. It also appears that there's now a more distributed amount of spaces between every other palindrome, seeing as the histogram now looks nearly uniform for the distribution of distances up to 2000. The population results look nearly in line with what's shown in the uniform sampled data plots, as the random samples also seem to show a new bias in distribution towards slightly higher distances.

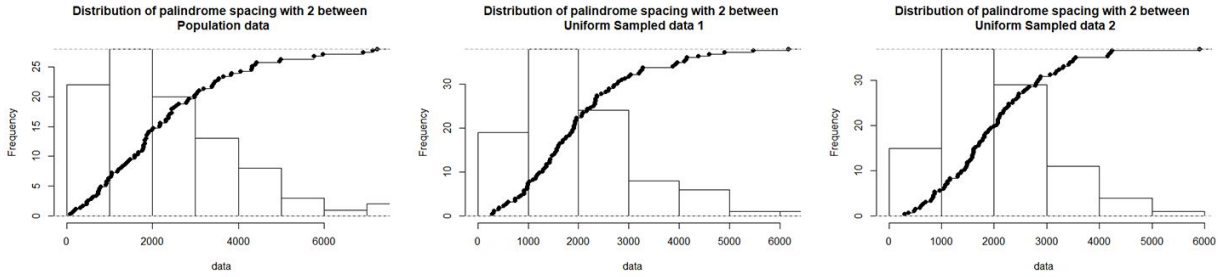
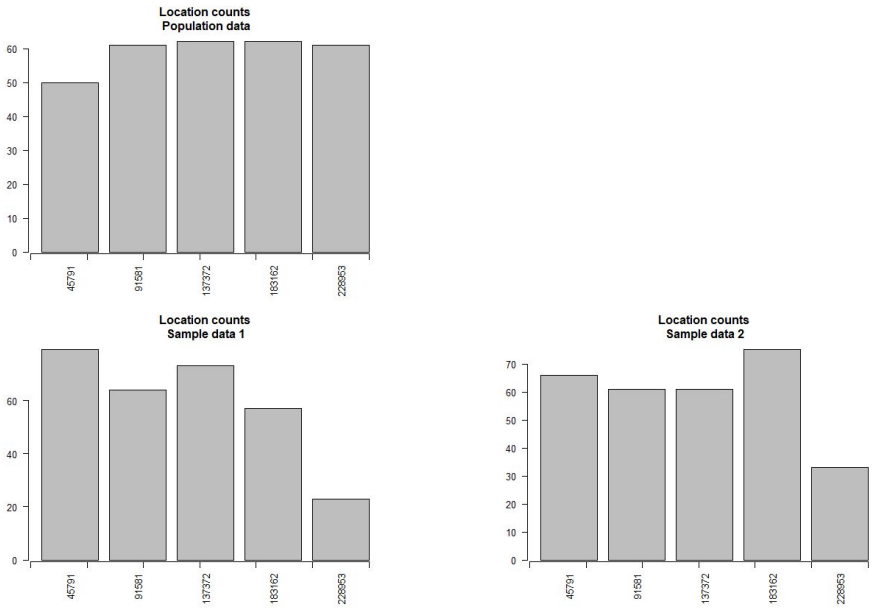


Figure (E): Distribution of double spaced palindrome spacings with uniform sampled data comparisons

This last figure shows another further linearization of the population CDF line, as the slope gets less steep as the distribution of spaces begins to open up even further. A distinction to still note between the population and sample data is that the population histogram still has a sizable amount of data with a lower locational distance measurement.This contrasts with the sampled distribution because the sample distributions are skewing more and more to the right, noticeably more than the population distribution.

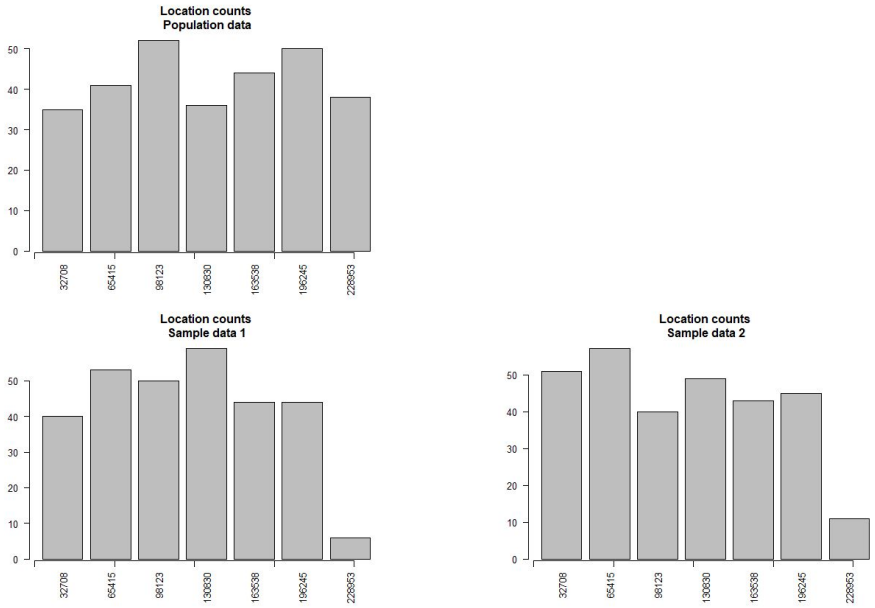
Investigation #3: Counts



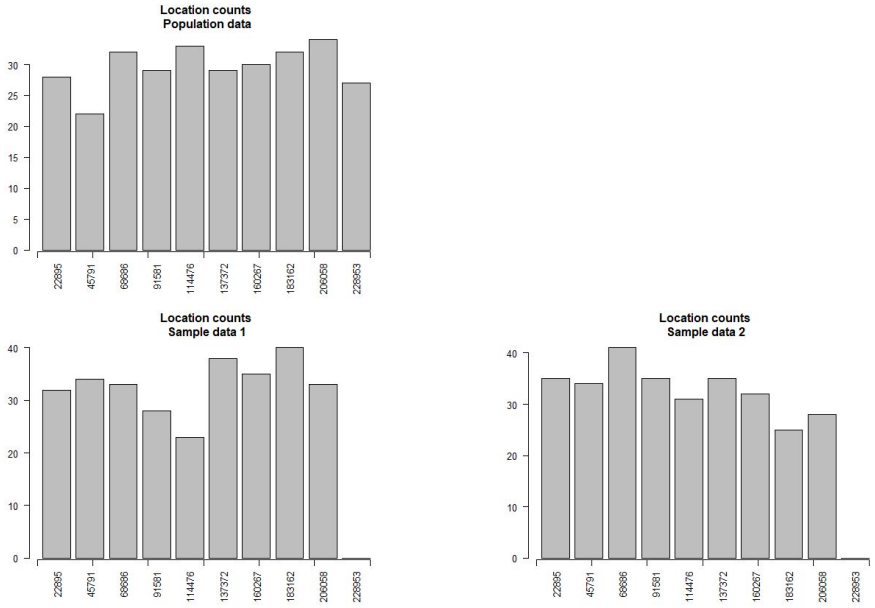
Figure(F): Interval counts of population and sample distributions with 5 bins

Looking at the distribution of palindrome counts between the population and a uniform sample distribution it seems like there's definitely something going on with this plot using 5 bins. In the population data, it seems like a very steady uniform distribution, but the sample distributions are dropping off in count near the end of the range, likely due to the population counts having a higher max range than the uniform distribution.

Accounting for that, the other bins look relatively reasonably distributed in comparison, which bodes well for the likelihood of a uniform distribution in the population.



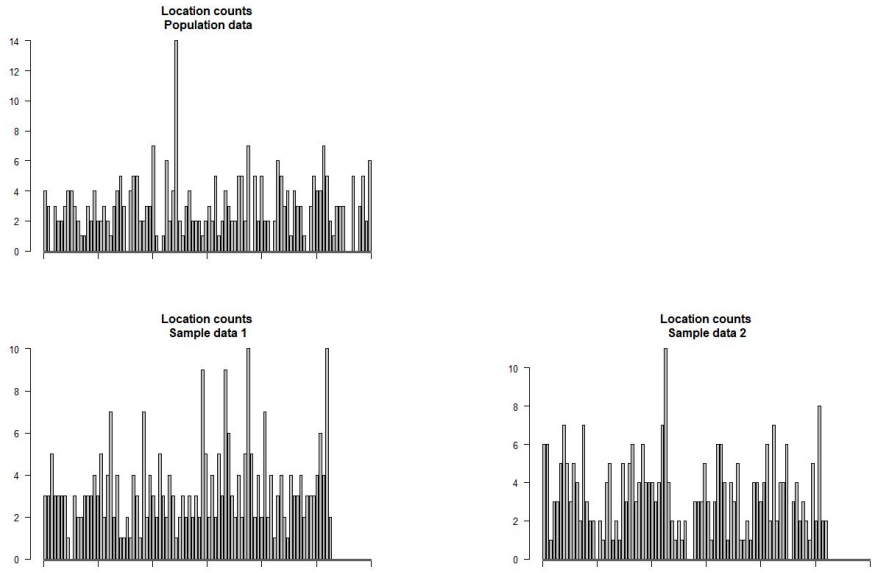
Figure(G): Interval counts of population and sample distributions with 7 bins



Figure(H): Interval counts of population and sample distributions with 10 bins

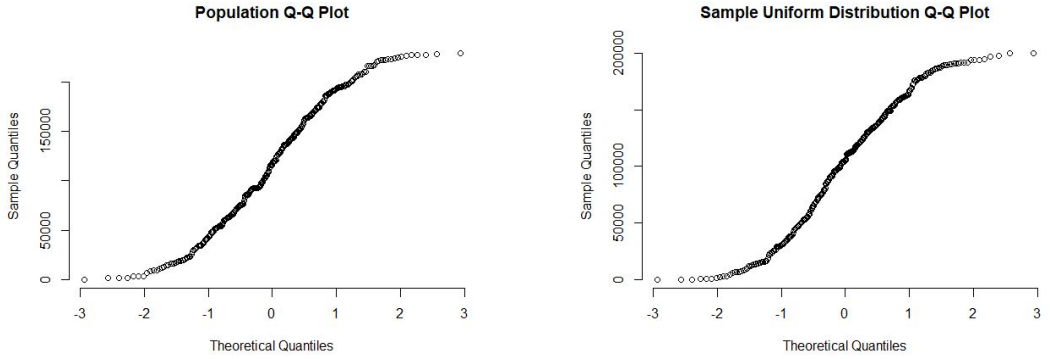
Moving on to the latter figures showing the plots with 7 and 10 bins respectively, it appears that the issue of the population having samples go further than the sample uniform distribution is still present, but besides that there still seems to be a relatively similar count distribution throughout. For the 7 bin plot, the population distribution seems to have more variation with spikes around 90,000 and 200,000, while the uniform sample distributions have less obvious spikes throughout.

The 10 bin plot looks to have a mostly uniform distribution with minimal drops, looking very similar to the population distribution. However, the population distribution does have a noticeable drop around the 50,000 mark. This drop isn't too dissimilar from the sample distributions though, so it can still probably be a uniform distribution.



Figure(I): Interval counts of population and sample distributions with 10 bins

Moving into a larger plot of 100 bins, we start to see more distinct spikes occurring in both the population and sample plots. Even though the uniform sample distributions have spikes in their distributions, the population distribution has one spike that has over double the amount of the next highest spike. Although there seems to be a uniform distribution for the lower binned plots, this high bin plot points out that there is an abnormally high amount of complementary palindromes around the location of 90,000 in the DNA sequence.



Figure(J): Q-Q plots of population and sample distributions

```
> chisq.test(data$location)

        Chi-squared test for given probabilities

data:  data$location
X-squared = 10569000, df = 295, p-value < 2.2e-16

> chisq.test(site.random)

        Chi-squared test for given probabilities

data:  site.random
X-squared = 9831000, df = 295, p-value < 2.2e-16
```

Figure(J): Chi-squared test for population and sample uniform distributions

Moving on to the Q-Q plots of the population and a uniform sample distribution, the plots show that there's relatively the same variability between the distributions, with both plots looking remarkably similar to one another. Performing a Chi-squared test on the distributions yields p-values for both distributions of next to 0, meaning we don't have enough evidence to reject the null hypothesis that the population is uniformly distributed.

Investigation #4: Biggest Cluster

Bins	$\widehat{\lambda}$	Interval Width	P-value	Max Counts
40	7.400000	5733.850	0.308448064	15
60	4.933333	3822.567	0.036209375	14
80	3.700000	2866.925	0.002693866	14

Table (B): P-values and Cluster Sizes Based on Different Interval Lengths

We use $\alpha = 0.05$ as the threshold. Table (B) shows the probability of the chance that the maximum count over m intervals is larger than or equal to k, where k is the maximum count generated from different lengths of intervals. Table (B) has three different amounts of intervals, which are 40, 60 and 80, and each of the three separations has a different lambda and probability. "Lambda" is estimated by using the method of MLE since we assume the model follows the Poisson distribution. "Interval Width" is calculated by dividing the corresponding amount of intervals from the total length. In the "P-value" column, the first p-value is above α and the following two are below α. Maximum counts are generated by taking the maximum count of palindromes among all the intervals. The p-value above α indicates the cluster is not unusual and the replication site is undetected since the regions examined are too large. This explains why, from Table (B), the probability of getting the maximum cluster larger than k decreases as the length of interval decreases. However, once we narrow the intervals, the cluster is well detected under $\alpha = 0.05$. The p-value (0.002693866) becomes very small when we seperate the total length into 80 intervals, which can be inferred that it is

very unlikely for the unusual cluster between the interval of [91700, 94600] to occur by chance. In addition, the interval where the maximum cluster is located for the three different interval widths are all in the range of [91700, 94600], which implies the maximum amount of palindrome gathered within that particular interval is unusual and thus may be a potential replication site.

VI. Conclusion

From our initial analysis we were not able to determine that the observed distribution of palindromes differed from a standard Uniform distribution. The samples we generated from a theoretical Uniform distribution behave similarly to our observed distribution and because of this we cannot state that the abnormalities in the observed have any explicit significance. When analyzing the spacing between each palindrome we discovered that the distribution of the distances in the observed data was relatively similar to the distribution of the distances in the data sampled from a uniform distribution for the distances between individual palindromes. This similarity faded away once we started analyzing the distances between every other and between every two other palindromes. The distribution of distances for the observed became more skewed to the left while the distribution of distances for the Uniform samples became more skewed to the right. Building off our original analysis, we analyzed the distribution of palindrome counts again, however we analyzed them with varying interval lengths for the bins. Though there was some slight variation in our observed data with smaller bin sizes we found no difference between the simulated normal distribution and the observed distribution after performing a Chi-Square test. In investigation #4, we looked for the maximum cluster size for different interval sizes. We found that the different subintervals result in different p-values and based on the length of the subinterval, the p-value will change. Hence, we conclude that there is statistically significant evidence that the largest clusters of certain intervals are larger than expected by the Poisson process. This suggests the interval with the most palindromes is a potential site of replication within the DNA. For our analysis specifically this means that the interval [91700, 94600] which contained a larger than normal number of palindromes when grouping the data in 80 bins would be a cite of interest for replication.

Given the results of our analysis there are no obvious outstanding outliers that a biologist could immediately identify and use to specify their research. A majority of our analysis found that the distributions of data were very similar to the Uniform distribution. If we could provide any recommendations to a biologist studying CMV we would suggest that they focus their analysis on the results from Figure (I) in Analysis Section 3. When separating the data in bins of 100 there was a very extreme spike at approximately the 90,000th index. This area of the genome was also found to be relatively significant from our Analysis Section 4 as when separating the data via a histogram with 80 bins the interval [91700, 94600] proved to be significant with a small p-value. The extremity of this spike was not well replicated at all by our simulations of a normal distribution so we would consider this one of the few significant takeaways from our analysis.

Limitations of the Data

It is worth noting that we are limited by the data as the length of the palindromes remain unknown. Without this information, we are limited to just the placement of the palindromes. Also, palindromes of length less than 10 were also excluded from our data. It is possible that if we had either then lengths of each palindrome or the complete distribution of palindromes we might have come to different conclusions with our analysis. Another limitation of this dataset is the absence of other mutations.This study would have benefited from information regarding the characteristics of the palindromes in known mutations of CMV or viruses from the Betaherpesvirinae or Alphaherpesvirinae families of which CMV is related with.[2]

VII. Works Cited

[1]: "Cytomegalovirus (CMV) Infection." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 14 Mar. 2020, www.mayoclinic.org/diseases-conditions/cmv/symptoms-causes/syc-20355358.

[2]: Ryan KJ, Ray CG, eds. (2004). *Sherris Medical Microbiology* (4th ed.). McGraw Hill. pp. 556, 566–9. ISBN 978-0-8385-8529-0.

VIII. Appendix

Variable	Description	Type of Data
location	The location of a palindrome on the genome sequence of length 229,354	integer

Table (C): Data Dictionary for hcmv.txt.