Case Study 3: Patterns in DNA

Dukki Hong (A98058412)
Karl Rummler (A12405136)
Nick Roberts (A11705541)
John Diez (A14751991)
Yiwen Li (A13959913)

I. Introduction

Many biological diseases and viruses are life threatening for people with suppressed or deficient immune system. In this study we particularly work on Human cytomegalovirus(HCMV), which is the type species of the virus genus Cytomegalovirus, which in turn is a member of the viral family known as *herpesvirus*. Such disease can lead to significant morbidity and even death although it remains latent within the body for a very long period. There is already one specific location interval found in previous research that has been hypothesized to be the origin of replication in the DNA sequence, and this paper will see if we can find the results. Also, CMV is typically connected with HIV, so we will be conducting further analysis to find the comorbidity rates of CMV and HIV with relation to age.

A virus' DNA has all the info to grow, survive and replicate. DNA are long coded messages from the alphabet: {A, C, G, T}. Human cytomegalovirus (CMV) is a life-threatening disease and the search for its origin of replication is vital. In herpes and epstein-barr, relatives to CMV, their origin of replication is denoted by a long palindrome of 144 length and several short palindromes in close proximity, respectively. When we look at CMV, we notice that the longest palindrome is 18 base pairs. In this study, we will take the approach of investigating clusters similar to epstein-bar to determine if there is any "unusually large cluster of palindromes". To find the origin of replication, DNA is cut into segments and each segment is tested to determine where is the critical region. The traditional approach in terms of studying this is very time consuming and expensive. A statistical investigation of DNA to identify unusually dense clusters of palindromes and critical region can help biologists reduce amount of testing needed.

II. Background

CMV is a common virus that spreads throughout the body, and once an individual is infected, they are infected for life. It's a pathogen creating immunosuppressed patients (or having a weaker immune system). Typically there are no initial symptoms, but it's risk for worst diseases and death goes significantly higher as it stays dormant and can be reactivated (Cook, 2007). Scientists have been wondering at what point in this virus's DNA sequence does it utilize to replicate. If they can find the origin of replication, then the ability to stop and possibly stop the virus from working could be key to solving this incurable disease.

Specific sequences within the base pairs of DNA are called palindromes, and some scientists have hypothesized and possibly found that unusual clusters could be the point of replication in this virus. In one study, they found that a specific palindrome sequence that they hypothesize as the origin of replication lies within repeats in the location of 92100 and 93500 in the DNA sequence (Masse, Karlin, Schachtel & Mocarski, 1992) . They also found that sequences before and after this contribute with replication, but they were not essential for the origin functions. To find these results, scientists created living organisms and infected them with the CMV virus or a control virus until the virus was influential enough where they could extract it for DNA. They cloned the CMV strain and utilized a scientific machine to find the DNA bases. To analyze their data, they utilized three statistical methods that analyzed repeated frequency of specific palindrome sequences. When they analyzed it, they found a significant point of replication to be under the previous stated area.

CMV is reactivated during critical illness. About 30% of individuals who have a dormant CMV have it reactivates and weaken once they gain a critical illness (Cook). This is an important point to those with HIV for two reasons. First CMV is regarded as a quiet disease that spreads through physical and sexuall contact, which is also the main way HIV is transmitted. Second, HIV is an autoimmune disease as well, but will make the host more susceptible to becoming sick with an intense fever or sick in general. With these two factors. In the early AIDS epidemic, epidemiologic studies indicated that through 1992, nearly half of HIV-infected patients eventually developed CMV end-organ failure or other internal diseases (Drew & Lalezari, 2006). The risk of exposure to CMV increases as age increases.

III. Data

The DNA sequence of CMV is published in 1990 by Chee et al. After one year, in 1991, Leung et al. implemented search algorithms to screen the sequence to screen out many types of patterns in DNA data. Altogether, 296 palindromes were found that were at least 10 letters long. The longest ones found were 18 letters long and occurred in locations 14719, 75812, 90763 and 173893 along the sequence.

Palindromes shorter than 10 letters were ignored because if the sequence is too short, it can occur too frequently, which may lead to biases and errors. In total, the CMV DNA is 229,354 letters long. Reading from the data, the first palindrome starts at position 177, the second is at 1321 and the last is t 228,953. The cluster of palindromes is considered as the frequent appearance of palindrome at a particular range.

One way to begin to group the data of the 296 palindromes found is to segment the DNA chain into intervals of base pairs and count the number of palindromes found in each interval. In drawing histogram with frequency of each palindrome for each interval, we can easily see where appear to be clusters. In which way, hypothesis question can be formulated. By comparing histograms of the actual palindromes to histograms based on randomly generated numbers we can see that the random sets of numbers present no pattern of clusters at any given point, no matter what size intervals we use.

IV. Methods

We model the palindromes by using a homogeneous poisson process which states that the locations of palindromes are uniformly distributed, the spacingings between palindromes are exponentially distributed, and the number of palindromes in an interval is poisson distributed. We affirm thats its a homogeneous poisson process by confirming the three necessary rules denoted in theory and use a chi-squared test to show that the observed and expected data of each come from the same distribution.

Additionally, we avoid using clustering algorithms as these are damaging to our study for two main reasons. One, if it is the case that there are no clusters, any clustering algorithm will cluster regardless of the presence of clusters. Two, if there were clusters and we were able to capture them, how could we compare one cluster to another? If we are searching for a many palindromes close together clustering cannot exactly show this.

In order to estimate the parameters for the spacing between 2 consecutive and 3 consecutive palindromes, which come from Exponential and Gamma, respectfully. We must estimate their rate parameter lambda. Note that MLE estimate for lambda here converges to the normal distribution with mean true lambda and variance: $\frac{1}{nI(\lambda)}$, where I(lambda) is the fisher information matrix. For the poisson process we see that it's the same case, and so with a large enough n we can be sure that our parameter estimates are accurate. Note that for the uniform case of palindrome locations, we do not need to estimate the parameters as the start and end points of the DNA are already known.

After finding a good estimate for lambda, we will create a contingency table that'll serve as the basis for our chi-squared test. Making sure that the expected values are greater than 5, we can form our chi-squared test and determine a probability on whether the observed data and the expected values follow from the same distribution. More details of the chi-squared processes can be found under the theory section.

The chi-square goodness-of-fit test and the test for the maximum number of palindromes in an interval, are two good examples of the hypothesis tests. Always, to investigate a claim, we introduce a hypothesis test for parameter values.

Starting with probability model, the X1, X2, ..., Xn and Y1, Y2, ..., Yn be the values for random variable X and Y. Set the standard deviation and mean of variables and compare them. The hypotheses include null hypothesis (H0) and alternative hypothesis (H1). For example, H0 : μH = μR and HR : μH ≠ μR. In our hypothesis testing, we assume H0 is true and find out how likely our data is by calculating p-value. The next step is to find out Test Statistics, such as t-test, z-test, etc.

V. Theory

Homogeneous Poisson Point Process	The homogeneous poisson process is a model for the case that when the next event will take place and how long it is expected to wait. The process comes from the notion of points discretely distributed on a line with no regularity.

A special case of the poisson process where the parameter is of the form $\Lambda = v\lambda$ where λ is the constant rate and v is the Lebesgue measure (some interval length). The parameter relates to the number of poisson points existing in some interval length.

Three characteristics of Homogeneous Poisson Point Process:

1. The underlying rate λ doesn't change with location (homogeneity).
2. The number of points falling in separate regions are independent.
3. No two points can land in exactly the same place.

The poisson process is a great reference for making comparisons between each cluster. For example, a strand of DNA can be considered as a long line and the cluster of palindrome can be thought of points on the line. Known that the number of palindromes of DNA is independent of the number of palindromes in another and the chance for one piece of DNA has palindrome in it is the same for all other pieces. Hence the data is scattered randomly and uniformly across the DNA.

Counts and the Poisson Distribution

One way to summarize a random scatter is to count the number of points in different regions. The probability model for Poisson process:

$$\frac{\lambda^k}{k!}e^{-\lambda}, \qquad \text{for } k = 0, 1, \ldots .$$

The parameter ƛ is the rate of hits per unit area. The expectation and variance of poisson distribution are also ƛ.

Chi-Square Goodness of Fit Test

Chi-square Goodness of Fit Test is used to determine how sample data is consistent with the predicted distribution. In order to compute the probabilities, the parameter of the distribution needs to be estimated. The measure of the discrepancy between sample count and the expected count is defined as follows:

$$\frac{\sum_{i=1}^{n}(N_i - \mu_i)^2}{\mu_i} = \frac{\sum_{i=1}^{n}(i^{th} observed\ count - i^{th}\ expected\ count)^2}{i^{th} expected\ count}$$

In order to calculate the p-value, χ^2 (chi-squared) distribution is used. If the probability model is right, then the test statistic has an approximate chi-squared distribution with n-k-1 degrees of freedom, n being the number of categories and k being the number of parameters estimated to obtain expected counts. If p-value is too small, it is possible that the distribution might not be fit. If the test statistic is greater than chi-squared with degrees of freedom n-k-1 and with some significance level α, then we reject the null hypothesis, H_o , otherwise we keep the null hypothesis.

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.

Uniform distribution measures the probability that a random variable is randomly distributed equally along the interval [a, b]. The following is the pdf (probability distribution) of the uniform distribution:

$f(x) = \frac{1}{b-a},\ x \in [a,b]\ and\ f(x) = \ 0,\ if\ else$

The characteristic of the uniform distribution is such that the palindromes are randomly scattered around the interval of locations [0, 229354]. It is important to note here that the scatters are not equally spaced throughout the bases. There are bases where gaps will occur too.

Exponential and Gamma

Another property that can be derived from the Poisson process is that

Distributions

the distance between successive hits follows an exponential distribution. That is,

P(the distance between the first and second hits > t)

= P(No hits in an interval of length t) = exp(−λt)

which implies that the distance between successive hits follows the exponential distribution with parameter λ. Similarly, it can be shown that the distance between hits that are two apart follows a gamma distribution with parameters 2 and λ. The exponential distribution with parameter λ is a special case of the gamma distribution for parameters 1 and λ. The χ2 k distribution is also a special case of the gamma distribution for parameters k/2 and 1/2. The χ2 k distribution can be used in place of the gamma(k/2, λ) distribution in gamma-quantile plots because λ is a scale parameter and only affects the slope of the plot.

Clusters and Maximum Number of Hits

Under the Poisson process, the sets of the number of hits in a non-overlapping and equal intervals are independent observations from a Poisson distribution. This implies that the maximum number of hits in a set of intervals behaves identical as the maximum of independent Poisson RVs (Random Variables). If n is the total number of intervals, then the probability of the maximum counts over n intervals is greater than equal to some number k. This can be rigorously defined as follows:

$$P\,(Maximum\ count\ over\ n\ intervals \geq k)$$

$$= 1 - [\lambda^0 e^{-\lambda} + \ldots + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}]^n$$

Parameter Estimation (Method of Estimation) and Maximum Likelihood Estimator (MLE)

Given $X_1, X_2, \ldots, X_n$ random variables that are iid (identically and independently distributed) with a Poisson parameter λ, the Method of Estimation technique is as follows:

1. Find E[X]= λ
2. Express λ in terms of E[X]
3. Replace E[X] with $\bar{x}$ to find an estimate of λ, $\hat{\lambda}$

MLE: Given $X_1, X_2, \ldots, X_n$ random variables that are iid (identically and independently distributed) with a Poisson parameter λ, the formula for MLE is as follows:

$L(\lambda) = e^{-\lambda} \frac{\lambda \sum_{i=0}^{n} x_i}{\prod_{i=0}^{n} x_i!}$, where $L(\lambda)$ is called the 'likelihood function'. MLE estimates the unknown parameter that maximizes the likelihood function. We use the log likelihood function. To find the maximum, we consider the first order equation:

$$l'(\lambda) = \frac{\partial[\sum_i x_i log(\lambda) - n\lambda - \sum_i log(x_i!)]}{\partial\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda - n} = 0$$

Solving the last equation yields $\hat{\lambda} = \bar{x}$.

Properties of Parameter Estimates

To compare and evaluate parameter estimates, people usually use the mean square error:

MSE(λ^) = E(λ^ – λ)^2 = Var(λ^) + [E(λ^) – λ]^2 .

Hypothesis Tests

The χ2 goodness-of-fit test and the test for the maximum number of palindromes in an interval are examples of hypothesis tests. We provide in this section another example of a hypothesis test, one for parameter values. We use it to introduce the statistical terms in testing.

VI. Investigations

Scenario I

Random Scatter

We will examine the spacings between consecutive palindromes in the DNA. Notice that a palindromes are uniformly and randomly scattered across 229,354 positions. This means that palindromes are not necessarily equally spaced across the 229,354 positions. Hence, we intend to simulate random uniform scatters to compare with the original data. 296 palindrome positions are randomly generated to observe different spacings between palindromes and different locations of the palindromes across the DNA compared to the original data.

Figures 1-4 shows the different palindrome locations along the 229,354 sites. Figure 1 is the positions of the palindromes for the original data and Figures 2-4 are simulated positions. It is difficult to visualize the differences in the palindrome locations for all four cases. However, it is apparent that the clustering of palindromes are different as all four figures show different locations with stronger color scales and lighter color scales. From this we see that the palindrome locations are not equally spaced.

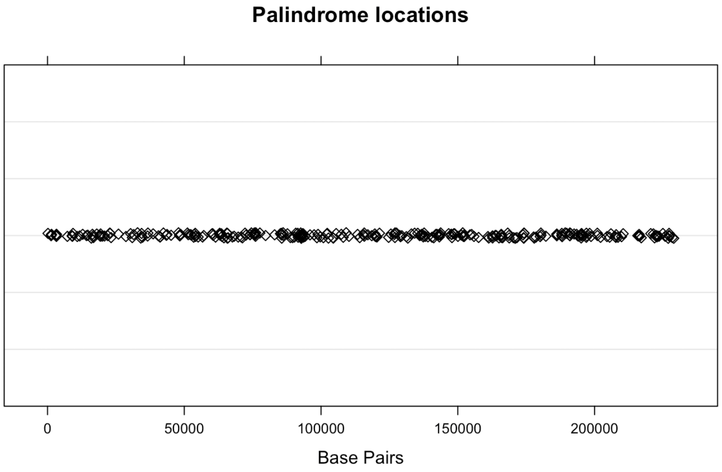


***Figure 1:** Scatter plot of Palindrome locations for original data*

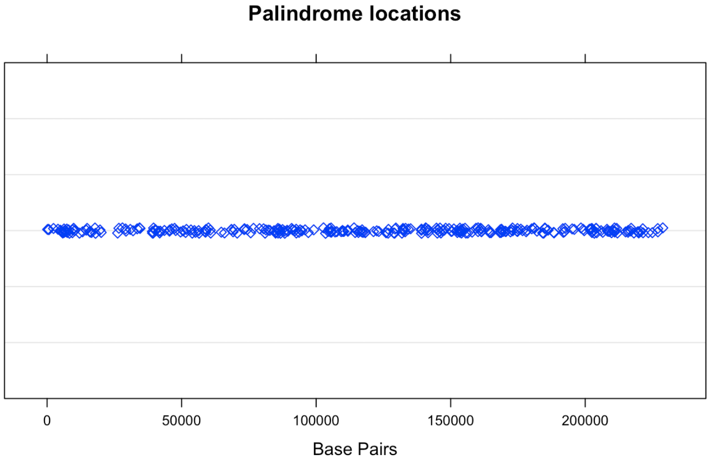


***Figure 2:** Scatter plot of Palindrome locations (simulation 1)*

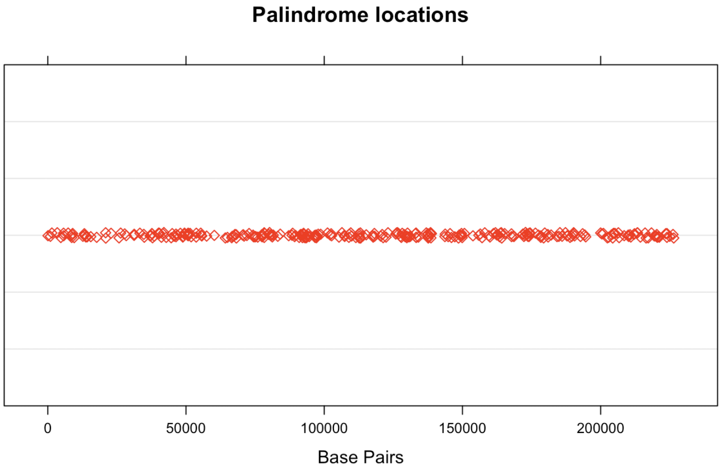


Figure 3: *Scatter plot of Palindrome locations (simulation 2)*

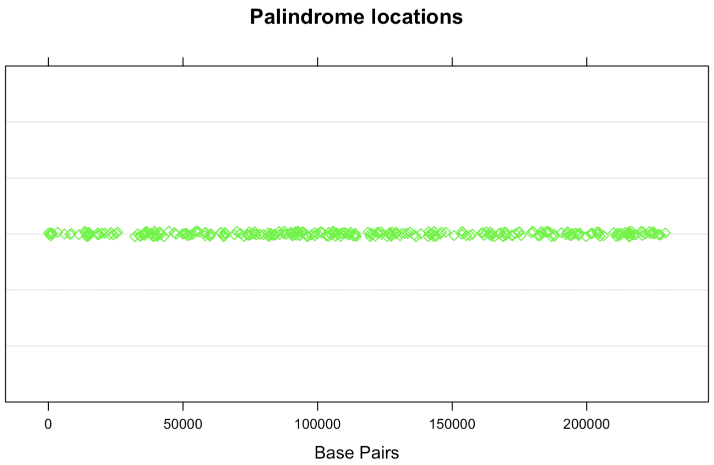


Figure 4: *Scatter plot of Palindrome locations (simulation 3)*

Figure 5 and 6 below shows overlapping histograms of palindrome locations by frequency (number of occurrences on the specific locations) for the original data and simulated data. It can be seen from Figure 5 that there are no locations that overlap with high frequency, the locations that overlap are colored in purple (i.e. the highest frequency for a given location is 5 given that the highest frequency from an individual strand was 7). It would be significant to conduct more experiments and see how palindrome location frequencies behave for a different simulation: Figure 6. Similarly, Figure 6 shows a similar trend where simulation 1 and 2 also does not have locations that overlap with high frequency that is colored in green (i.e. the highest frequency that overlapped this time was only 4 where as it was 5 in Figure 5). Hence it is important to note that despite the maximum frequency level for a given location (maximum frequency of 7 for Figure 5 and maximum frequency of 8 for Figure 6), the overlapping frequency of palindrome locations was lower. This is parallel with the idea of random scatter, indicating that the palindromes are not equally spaced along the bases but rather scattered randomly throughout.

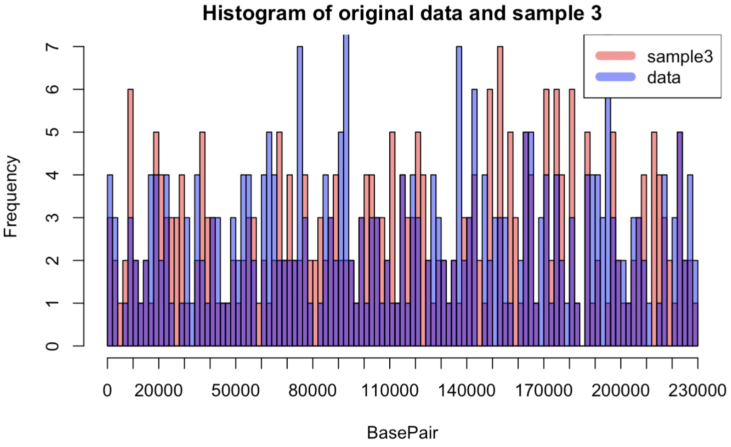


Figure 5: *Histogram of palindrome locations by occurrence (simulation 3 vs original data)*

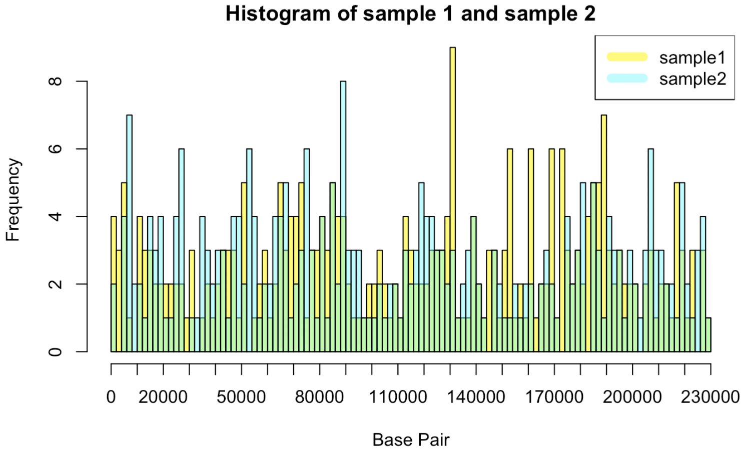


Figure 6: *Histogram of palindrome locations by occurrence (sample simulation 1 vs sample simulation 2)*

Scenario II

Spacing Between Palindromes

Figure 7 and 8 shows histograms of the spacings between pairs and triplets for both observed and generated data. It can be seen from the observed data (Figure 7) that the lowest distance between the two base pairs had the highest frequency (i.e. frequency of 80+). As the distance between the two base pairs increases, the frequency also decreases exponentially. In general, the behavior of the histogram across the x-axis is such that it is similar to the exponential distribution when graphed. Likewise, it can also be seen from the generated data (Figure 8) that the lowest distance between the two base pairs had the highest frequency (i.e. frequency of approx. 70), leading to a general trend where the frequency decreases exponentially as the distance between the two base pairs increases.

Both the observed data and generated data had some pairs that were far out from each other (i.e. pairs that were about 5200+ distances apart).

On the other hand, looking at the distance between triplets for the observed data, we see a different pattern. The lowest distance between the triplets had the highest frequency, however, as the distance hit near the range of 400 and 1000, the frequency decreases abruptly. Furthermore, the frequency increases again when the distance started to increase from 1000 and then decreases exponentially after the distance increases from around 2000. Hence the behavior of the histogram is similar to the graph of the gamma distribution. Comparing the observed data with the generated data, the histogram also behaves similar to the graph of the gamma distribution. As the distance between the triplets increases, there is a similar trend where the frequency is low at first and then increases and decreases as the distance increases.

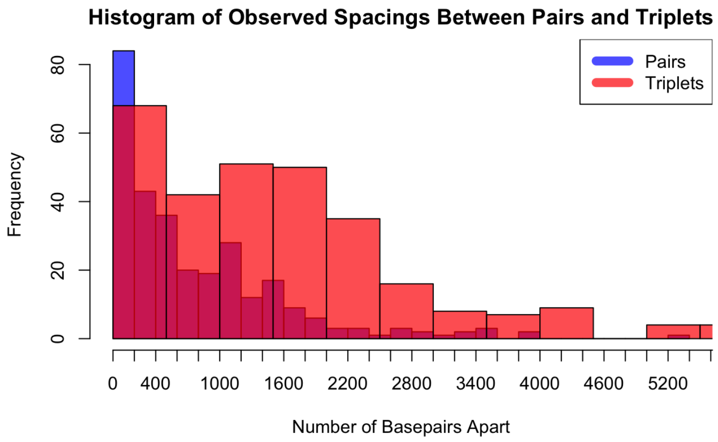


Figure 7: *Histogram of Observed Spacings (Observed data)*

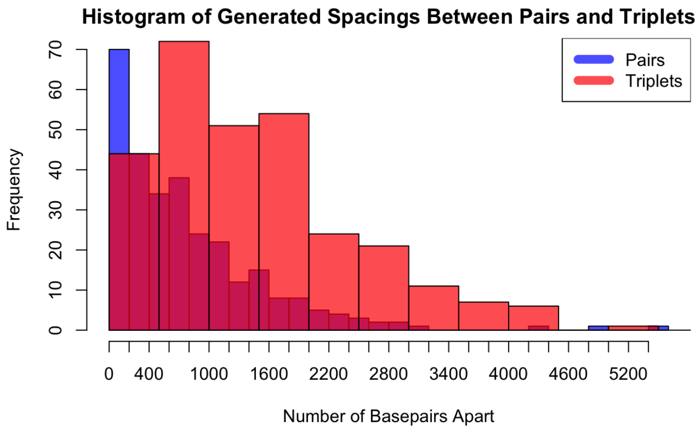


Figure 8: *Histogram of Observed Spacings (Generated data)*

Figure 9 and 10 shows overlapping histograms that compare the distance between the pairs and the triplets. Comparing the pair spacings for the observed data and generated data, both had similar trends wherein the histograms' behavior looked similar to the graph of the exponential distribution.

Similarly, Figure 10 shows how the generated data and the observed data has similar histogram behaviors wherein they look similar to the graph of the gamma distribution.

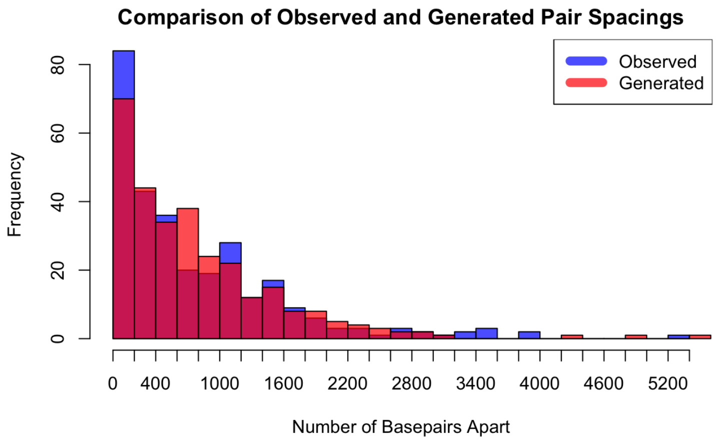


Figure 9: *Overlapping Histogram of the distance between pairs*

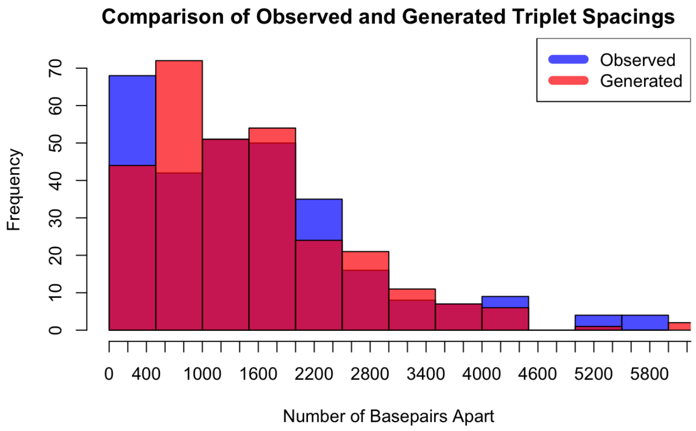


Figure 10: *Overlapping Histogram of the distance between triplets*

Estimating Lambda & Generating the Data

For generating the pair spacings in Figures 9 and 8, we had to estimate the lambda using the MLE of the parameter. In our case we found $\lambda = \frac{1}{\# \ of \ Palindromes} = \frac{1}{296} = 0.001289$. Thus, when generating the space between consecutive palindromes, we generated 295 spacings from Exponential($\lambda = 0.001289$). For the spacing between two consecutive pairs, we generated 294 spacings from Gamma(2, $\lambda = 0.001289$).

However, in order to statistically show that the spacings between palindromes follows an Exponential distribution (and show that the consecutive spacings follow Gamma), we most form contingency tables so that we may accurately do the chi-squared test.

Interval	Observed Spacing	Expected Spacing
0 - 75	*46*	27.19341
76 - 150	*26*	24.68669
151 - 225	*20*	22.41105
226 - 300	*20*	20.34517
301 - 375	*12*	18.46973
376 - 450	*14*	16.76717
451 - 525	*12*	15.22156
526 - 600	*13*	13.81842
601 - 675	*11*	12.54462
676 - 750	*5*	11.38825
751+	*116*	112.15391

Table 1: *Contingency Table of Observed Spacings of Palindrome Pairs*

Interval	Observed Spacing	Expected Spacing
0 - 200	28	8.247908
201 - 400	27	19.690700
401 - 600	22	25.504856
601 - 800	22	27.658097
801 - 1000	11	27.514393
1001 - 1200	20	26.006770
1201 - 1400	19	23.762756
1401 - 1600	21	21.195054
1601+	124	114.419466

Table 2: *Contingency Table for Spacings between three consecutive Palindromes*

For calculating the expected spacing, we multiplied the probability of observing that interval and multiplying it by the number of intervals across the entire DNA strand. For calculating the probabilities we used the distributions described above. Additionally, we check that the expected spacings are greater than 5 so that we can properly perform the chi-squared test to see if the two distributions are the same.

Goodness of Fit Test: Chi-Squared

To see if the observed data follows the suspected distributions we computed to the chi-squared test. We performed this on spacing between consecutive palindromes and between three consecutive palindromes.

	Degrees of Freedom	Chi-Squared Value	p-value
Space Between 2 Consecutive Palindromes	9	20.700	0.014
Space Between 3 Consecutive Palindromes	7	64.712	1.717567e-11

Table 3: *Chi-squared Test for Spacing between 2 and 3 consecutive palindromes*

Looking at the p-values of the chi-squared test, it seems easy to conclude that the spacings do not follow an Exponential and Gamma distribution, however, looking at the contingency table we see that the smallest spacing intervals had the largest residual in comparison to the other rows. This follows with our original hypothesis that there may be many palindromes close to one another indicating the source of replication. Furthermore, we can clearly see from figure 11 and 12 that the main source of residuals come from the smallest spacing interval for both spacing of consecutive pairs and three consecutive pairs. This supports our reasoning that the failure of the chi-squared test comes from the existence of the source of replication in the CMV.

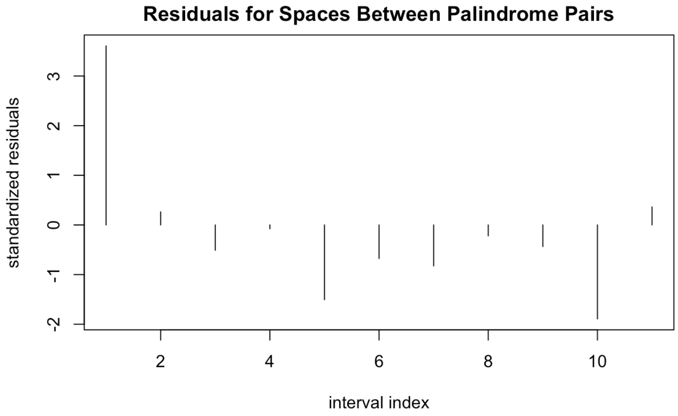


Figure 11: *Residuals for Spacing between Consecutive Palindromes*

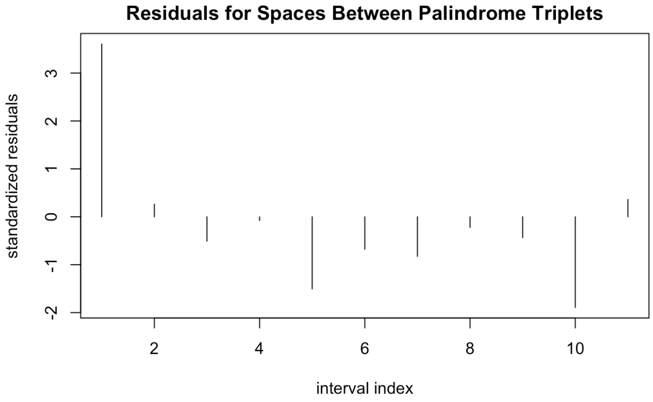


Figure 12: *Residuals for Spacing between 3 Consecutive Palindromes*

Location of Palindromes

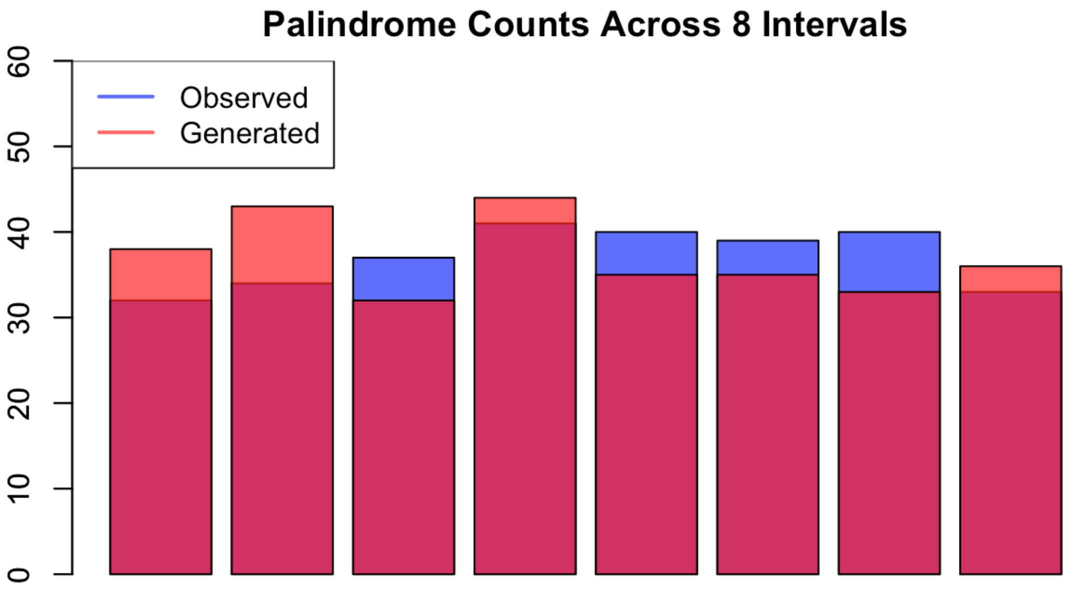


Figure 13: *Count of Palindrome locations over 8 intervals which partition the DNA*

Looking a the the palindrome counts across 8 intervals it looks that the observed and generated datasets follow the uniform distribution in terms of palindrome locations. We created the generated graph by using a Uniform distribution with parameters 0 and 229354. We then computed 296 random points from this uniform distribution and plotted them and compared them with the observed data.

To check that the observed palindrome locations likely follows a Uniform distribution we build a contingency table so that we may properly create a chi-squared test. Forming the contingency table, we partitioned the DNA sequence into 8 equally sized portions. The expected counts were simply computed by getting the total number of palindromes observed and dividing by 8. Since the observed counts are all obviously above 5 we continued to the chi-squared test.

Interval	*Observed Counts*	*Expected Counts*
1	*32*	*37*
2	*34*	*37*
3	*37*	*37*
4	*41*	*37*
5	*40*	*37*
6	*39*	*37*
7	*40*	*37*
8	*33*	*37*

Table 4: *Contingency Table for Palindromes Locations*

Goodness of Fit Test: Chi-squared

We created a chi-squared test on the observed counts over 8 equally sized intervals with the uniform distribution. However, when computing the degrees of freedom, we calculated it differently by using the number of rows - 1. This is because with the uniform distribution we do not need to estimate any of the parameters since the begin and end points are known. Looking at table 5 we see that it's very likely the observed palindrome locations follow a uniform distribution. Looking at the residuals in figure 14, we very little error across the 8 intervals.

Degrees of Freedom	Chi-Squared Value	p-value
7	2.3783784	0.936

Table 5: *Chi-squared Test for Locations of Palindromes*

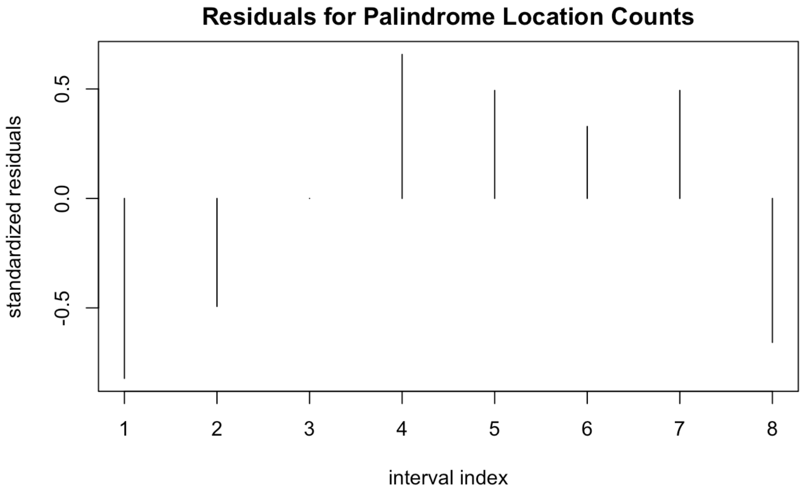


Figure 14: *Count of Palindrome locations over 8 intervals which partition the DNA*

Scenario III

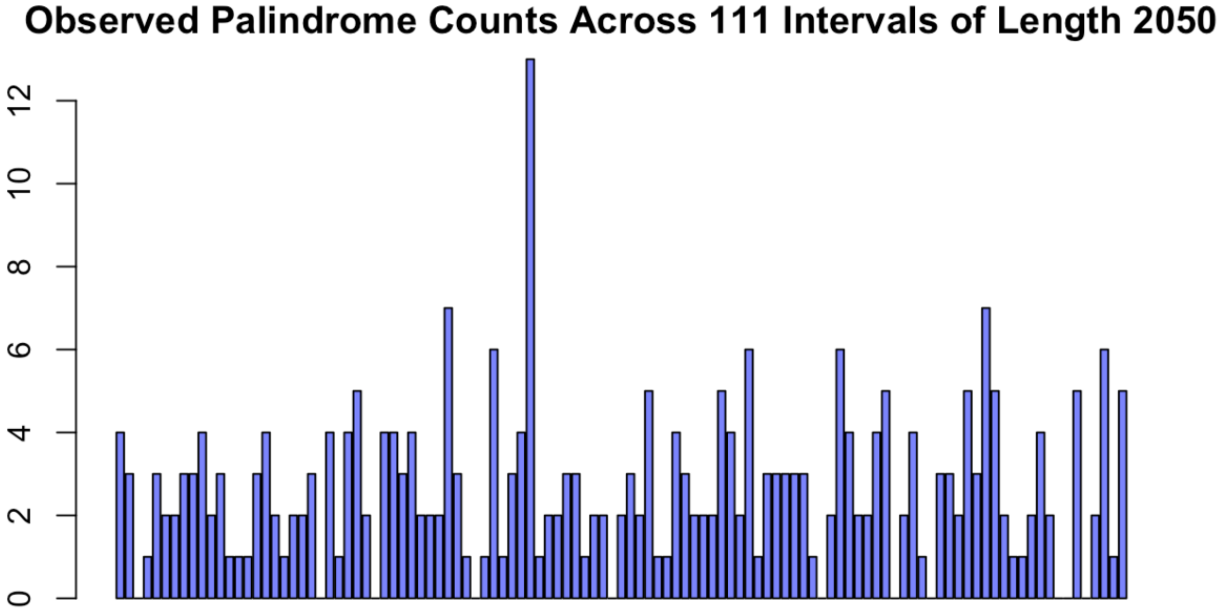


Figure 13*: Observed Palindrome Counts across 111 Intervals of Length 2050*

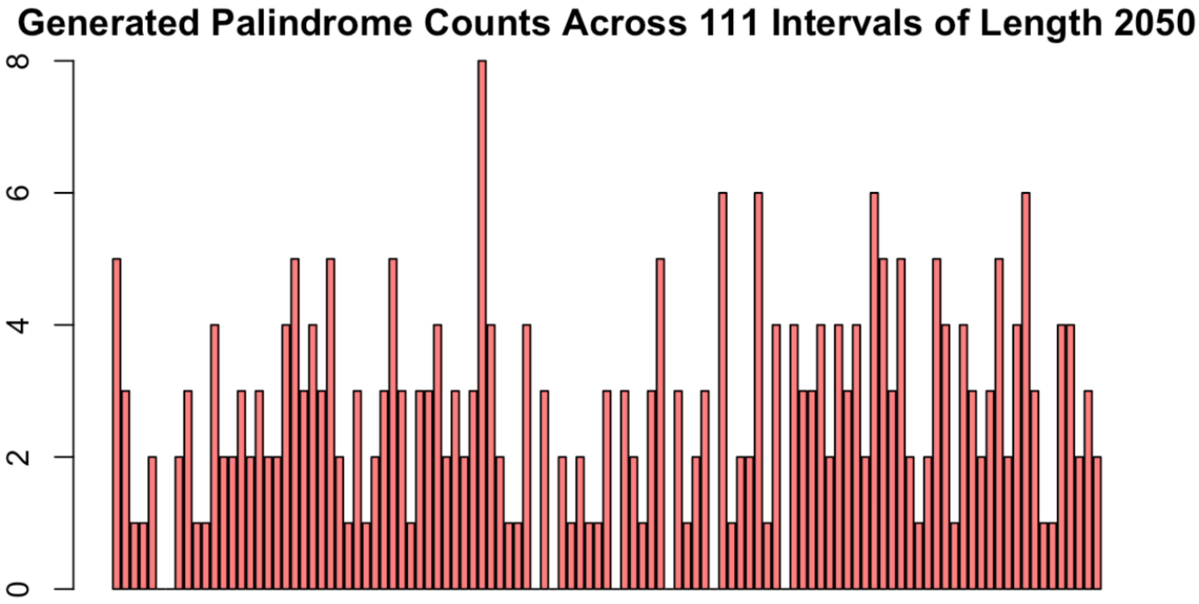


Figure 14*: Generated Palindrome Counts across 57 Intervals of Length 2050*

We generated a graph that simulates a poisson distribution of palindrome counts across 57 intervals of length 2050 each. We did generated the poisson distribution by estimating lambda form the observed data. Utilizing the MLE estimate for lambda with the number of palindromes divided by the number of intervals. Comparing the two graphs we see that everything below 7 in terms of count follows closely with both the observed the generated graph. It is only at around interval 25 that we see an outlier in the observed data with a count of 14.

Here we calculated: $\lambda = \frac{\#\ of\ Palindromes}{\#\ Of\ Intervals} = \frac{294}{111} = 2.649$

In order to perform a chi-squared test on this data to determine if the observed data follows a poisson distribution, we must build a contingency table that allows us to properly make this test.

Palindrome Counts	**Observed Counts**	**Expected Counts**
0-2	11	7.852
3	19	20.800

4	29	27.545
5	22	24.319
6	15	16.103
7	8	8.530
8+	7	5.849

Table 6: *Chi-Squared Contingency Table Across 57 Intervals of Length 2050*

In order to determine if the number of palindromes per interval follows a poisson distribution we first build a chi-squared contingency table that counts the number of palindromes across all intervals, tallying them up. Additionally, we check that the expected counts is greater than 5 such that performing a chi-squared test is valid. We combined certain rows together such that we get values above 5. The expected values were calculated using the probability of exponential distributions from the palindrome count range using the estimated parameter explained above and multiplying this probability by the number of intervals. This gives us the expected number of palindromes across all the intervals.

Goodness of Fit Test: Chi-Squared

We perform the chi-squared test with 5 degrees of freedom. This is calculated from number of rows - number of parameters being estimators - 1. We get a chi-squared value of 2.0499 and find that our p-value becomes 0.842. From the data we see an 84% likelihood that the observed data follows a poisson distribution. (Note that chi-squared null hypothesis states that the two distributions are the same, and the alternative hypothesis is that they're different). So we may conclude that its likely the two datasets come from the same distribution.

Degrees of Freedom	Chi-Squared Value	p-value
5	2.0499380	0.842

Table 7: *Chi-squared Test for Counts of Palindromes*

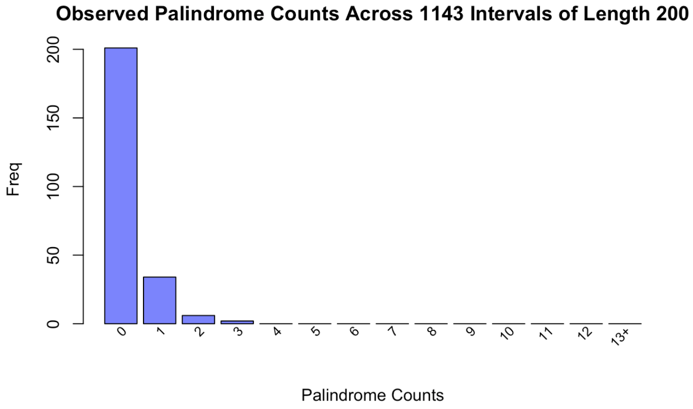


Figure 15: *Palindrome Counts Across 1143 Intervals of Size 200*

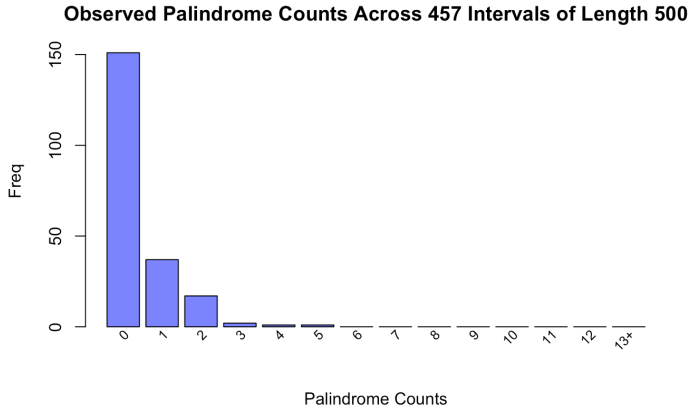


Figure 16*: Palindrome Counts Across 457 Intervals of Size 500*

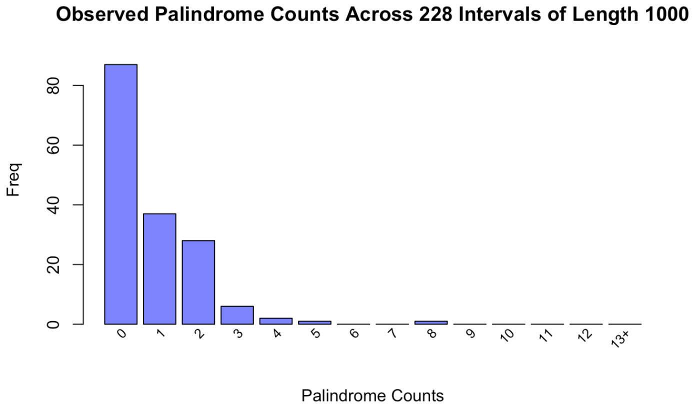


Figure 17*: Palindrome Counts Across 228 Intervals of Size 1000*

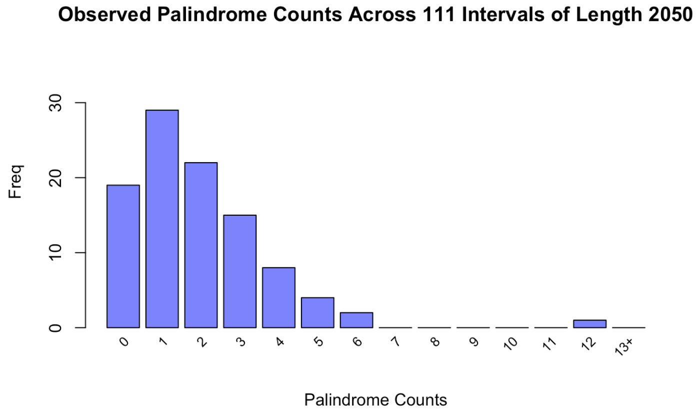


Figure 18*: Palindrome Counts Across 111 Intervals of Size 2050*

Looking at observed palindrome counts across multiple interval sizes. We see a consistent pattern that most of the palindrome counts do follow a poisson distribution. This is noted by the fact that the smaller interval sizes have more intervals and thus smaller lambdas. This is well represented in poisson distributions with a stepper curve, similar to an exponential, which is indicated by figures 15, 16, and 17. Looking at interval 18 we note the lambda estimate is larger, which also follows a poisson distribution with the same lambda.

Scenario IV

In hypothesis testing we assume that the H0 is true and find out how likely our data are under this model. If the p-value was < 5% we would have concluded that the observations do not support the null, and we would reject the null in favour of the alternative. The cutoff of 5% is called significance level of a test.

In generating the maximum counts, m, of various intervals length, we want to show that the probability that the interval with the greatest number, n, of palindromes is larger or equal to m. According to previous analysis, we picked different interval length as different interval counts, including 40, 50 and 60. With each interval counts, we calculate the corresponding interval length, rate parameter, and p-value related to specific n. The interval length is calculated by total interval length divided by interval counts; the rate parameter ƛ is estimated with maximum likelihood Estimator (MLE).

	40 Intervals	**50 Intervals**	**60 Intervals**
Interval Length	5733.850	4587.080	3822.567
Parameter ƛ	7.400000	5.920000	4.933333
Max Count	15	14	14
P-value	0.308448064	0.05974615	0.03620938

Table 8 *Interval Counts and p-values*

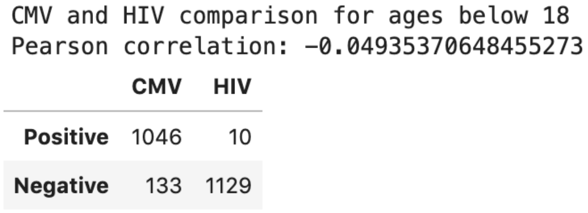
Based on calculation, we found out that the interval with the greatest number of palindromes indicate a potential origin of replication (91800, 95600), which is the 24th interval in n = 60 interval counts. The p-value calculated is the table demonstrates how possibly the interval indicates a potential origin of replication. Locating to the last column, as interval count equal to 60, the corresponding p-value is approximately 0.0362, which is smaller than our threshold 0.05. This manifests that the cluster is unusual and it is highly possible that the interval (91800, 95600) is the replication site. Consider the other two cases, as n = 40 and n = 50, they obtain higher p-value, which shows the interval is not unusual and the replication goes undetected because the regions examined are too large. Moreover, consider the interval (91800, 95600), it is always contained in the interval for n = 40 and n = 50 because they have larger interval length. Hence, we would advise biologist to investigate on the interval range of (91800, 95600).

Additional Hypothesis

Throughout this point, we've been analyzing the possible origin of replication point in the CMV data strand. Now we can take a step back and review some of the literature of CMV in a broader scope, specifically how it is related to HIV. Based on previous research, we know a few things. First that CMV can get reactivated once someone receives a critical illness (like a high fever), and HIV (an autoimmune disease) makes people susceptible to that. Due to this factor, we are interested in the relationship between HIV and CMV. Based on previous research we found that people with HIV have a high comorbid (jointly occuring) rate with CMV, due to both possibly being transferred sexually. Another factor was how higher age groups have a higher chance of developing or having both HIV and CMV.

To further look into this, we looked into a more detailed dataset by Nash & Colleagues at 2018. Their data is taken from a rural Ugandan area which is vastly different from the other data set of just the dna sequence of CMV. They had a total of 2175 participants with an age range of 0 to 80+. We're analyzing in the data set 3 columns: Age, CMV Status, and HIV status. What we've done is create a correlation coefficient for 5 different age groups to see if there is a correlation between CMV and HIV to confirm previous research.

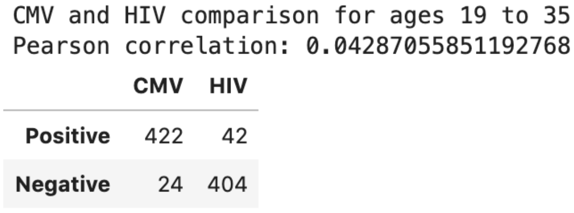
In the five tables below, the values or numbers located within each box is the number counts in that data that satisfy the two different levels of conditions. The column titles represent the two different diseases of CMV and HIV, while the rows represent the status of their disease of Positive (they have it) and negative (they don't have it). We've separated our data into five different sub groups of age to see if there was any significant correlation between CMV and HIV. From our analysis, for each of the five groups, we found no significant difference between any or all of the groups.



```
CMV and HIV comparison for ages below 18
Pearson correlation: -0.04935370648455273
```

	CMV	HIV
Positive	1046	10
Negative	133	1129

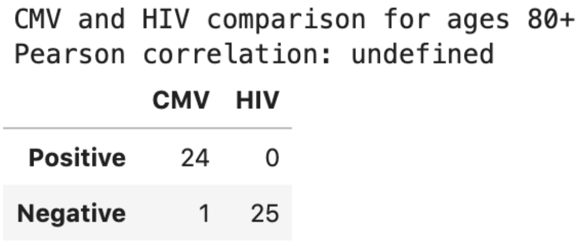
***Table 9** CMV and HIV for <18*



```
CMV and HIV comparison for ages 19 to 35
Pearson correlation: 0.04287055851192768
```

	CMV	HIV
Positive	422	42
Negative	24	404

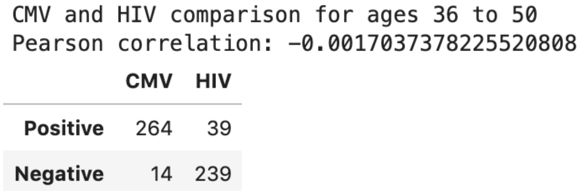
***Table 10** CMV and HIV for 19-35*



```
CMV and HIV comparison for ages 80+
Pearson correlation: undefined
```

	CMV	HIV
Positive	24	0
Negative	1	25

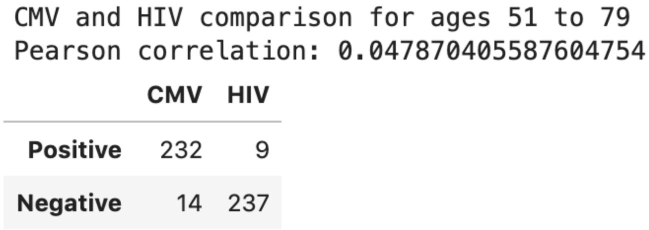
Table 11: *CMV and HIV for 80+*



```
CMV and HIV comparison for ages 36 to 50
Pearson correlation: -0.0017037378225520808
```

	CMV	HIV
Positive	264	39
Negative	14	239

Table 12: *CMV and HIV for 80+*



```
CMV and HIV comparison for ages 51 to 79
Pearson correlation: 0.047870405587604754
```

	CMV	HIV
Positive	232	9
Negative	14	237

Table 13 *Interval Counts and p-values*

These findings could be attributed to a specific limitation. One limitation from our findings is that the CMV and HIV status markers are only markers, and they don't provide a time frame of when they occur. Typically, a dormant CMV would occur after HIV, so we don't know whether this is an accurate depiction. So the lack of comorbidity in the results could be due to a measurement artifact.

VII. Conclusion

In our study, we wanted to help biologist find a possible origin of replication for the CMV virus. We were provided with the data sequence of base pairs of a CMV virus, and we conducted multiple analysis to see if it followed a specific theoretical distribution that would allow us to determine if there was a specific sequences or palindromes that were unusual. In scenario 1, we tested whether the palindrome locations followed a uniform distribution, and we found a specific range which we regarded as an outlier that we will explain in the next paragraph. In the second and third scenario, we tested whether the spacings between each palindrome as well as counts within a specific interval we tested followed and exponential and poisson distribution. We found that it mostly followed those distributions, but a specific section skewed it slightly. Now we will explain our recommendations and conclusions

According to our previous analysis, we recommend biologist to experimentally searching for the origin of replication in the range of interval (91800, 95600). Based on previous analysis on data distribution, we consider 40, 50, 60 are reasonable interval counts to evaluate the most optimal replication region. In dividing the total length of DNA sequence by interval counts, the interval length is calculated, which contributes to decide how many counts in each interval. In taking the maximum counts over all intervals, we narrow down our replication region to (91800, 97534) for n = 40, (91800, 96387) for n = 60, and (91800, 95600) for n = 80. In calculating the p-value for different interval counts, we use hypothesis testing to determine if the region is unusual enough to be replication of origin. As it turns out, the p-value for n = 80 is smaller than the critical value 0.05. Hence we recommend biologist the region (91800, 95600) by n = 60.

In addition to our next hypothesis on the relationship between CMV and HIV, we found that they are not significantly correlated, however, we have a possible confound for our analysis due to a measurement artifact. Due to having only the status of CMV and HIV for the participants, we don’t know at what time it was specifically identified that the individuals had the disease. Also the data was collected in a rural community in Uganda. We recommend that future researchers to figure out the longitudinal rate or process of CMV and HIV. Specifically, for them to record the severity levels and the estimated length of how long they have the disease. Also we recommend future biologists to sample a more diverse population of individuals who have CMV and or HIV as previous research have been focused on homosexual community.

VIII. Works Cited

Comprehensive, up-to-date information on HIV/AIDS treatment and prevention from the University of California San Francisco. (2006, April 28). Retrieved from http://hivinsite.ucsf.edu/InSite?page=kb-05-03-03

Cook, C. (2007). Cytomegalovirus Reactivation in "Immunocompetent" Patients: A Call for Scientific Prophylaxis. *The Journal of Infectious Diseases,196*(9), 1273-1275. doi:10.1086/522433

Erice, A., Tierney, C., Hirsch, M., Caliendo, A. M., Weinberg, A., Kendall, M. A., & Polsky, B. (2003). Cytomegalovirus (CMV) and Human Immunodeficiency Virus (HIV) Burden, CMV End-Organ Disease, and Survival in Subjects with Advanced HIV Infection (AIDS Clinical Trials Group Protocol 360). *Clinical Infectious Diseases,37*(4), 567-578. doi:10.1086/375843

Fickenscher, H., Stamminger, T., Ruger, R., & Fleckenstein, B. (1989). The Role of a Repetitive Palindromic Sequence Element in the Human Cytomegalovirus Major Immediate Early Enhancer. *Journal of General Virology,70*(1), 107-123. doi:10.1099/0022-1317-70-1-107

Lerner, C. W., & Tapper, M. L. (1984). Opportunistic Infection Complicating Acquired Immune Deficiency Syndrome. *Medicine,63*(3), 155-164. doi:10.1097/00005792-198405000-00002

Masse, M. J., Karlin, S., Schachtel, G. A., & Mocarski, E. S. (1992). Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. *Proceedings of the National Academy of Sciences,89*(12), 5246-5250. doi:10.1073/pnas.89.12.5246