Patterns in DNA

Yong Liu, Yunlin Tang, Jian Jiao

MATH 189, Spring 2020, HW3

I. Introduction

DNA is extremely significant for studying various types of virus in biology. A virus's DNA contains all necessary and important information for it to grow, survive, and replicate. To develop strategies for combating the virus, scientists study the way in which the virus replicates by searching the origin of replication, where is a special place on the virus' DNA that contains instructions for its reproduction. In particular, we will investigate the DNA of the human cytomegalovirus (CMV), which is a potentially life-threatening disease. To find the origin of replication, scientists suspected that clusters of palindromes in CMV may be the potential answer. The data used for investigation contains the 296 locations of the palindromes in the CMV DNA (229,354 letters long) that are at least 10 letters long. In general, we will perform data analysis in different features of data, such as locations, spacings, counts, and the biggest cluster by using statistical methods. A conclusion will be drawn with assistance from these methods in order to advise biologists who are about to start experimentally searching for the origin of replication.

II. Analysis

A. Location

First, we use the location feature to learn the distribution of palindromes on the CMV DNA. To achieve this goal, we graphically compare the sample palindrome locations to random uniform scatter. Specifically, we visualize the distribution of the sample locations, the distribution of random uniform scatter instances, and the theoretical uniform distribution by using histograms and the plot of the theoretical probability density function.

In general, we use the homogeneous Poisson process as the reference model for making comparisons since it is a natural model for uniform random scatter. In our data, there is a total of 296 palindromes on the CMV DNA along 229,354 complementary pairs of letters. Under the uniform random scatter model, the locations of palindromes can be viewed as 296 independent variables from a uniform distribution which are randomly and uniformly scattered among the DNA positions. In other words, each location on the CMV DNA has the same probability to have a palindrome. By definition of uniform distribution, this type of distribution can define equal probability over a given range for a continuous distribution. Thus we choose the uniform distribution as the theoretical model for the location feature. To visually compare the difference between sample locations and locations generated by the uniform random scatter model, a simulation study is carried out. In each run of simulations, we use runif() function in R to generate 296 uniform random deviations to represent the locations in CMV DNA. Then plot the histograms of both samples and simulated locations in one graph for better comparison. In addition, for each graph, the pdf of a theoretical distribution is plotted overlayed on the density histograms. In this case, the theoretical distribution is the uniform distribution. After running 6 times of simulation in total, we obtain the following graphs:

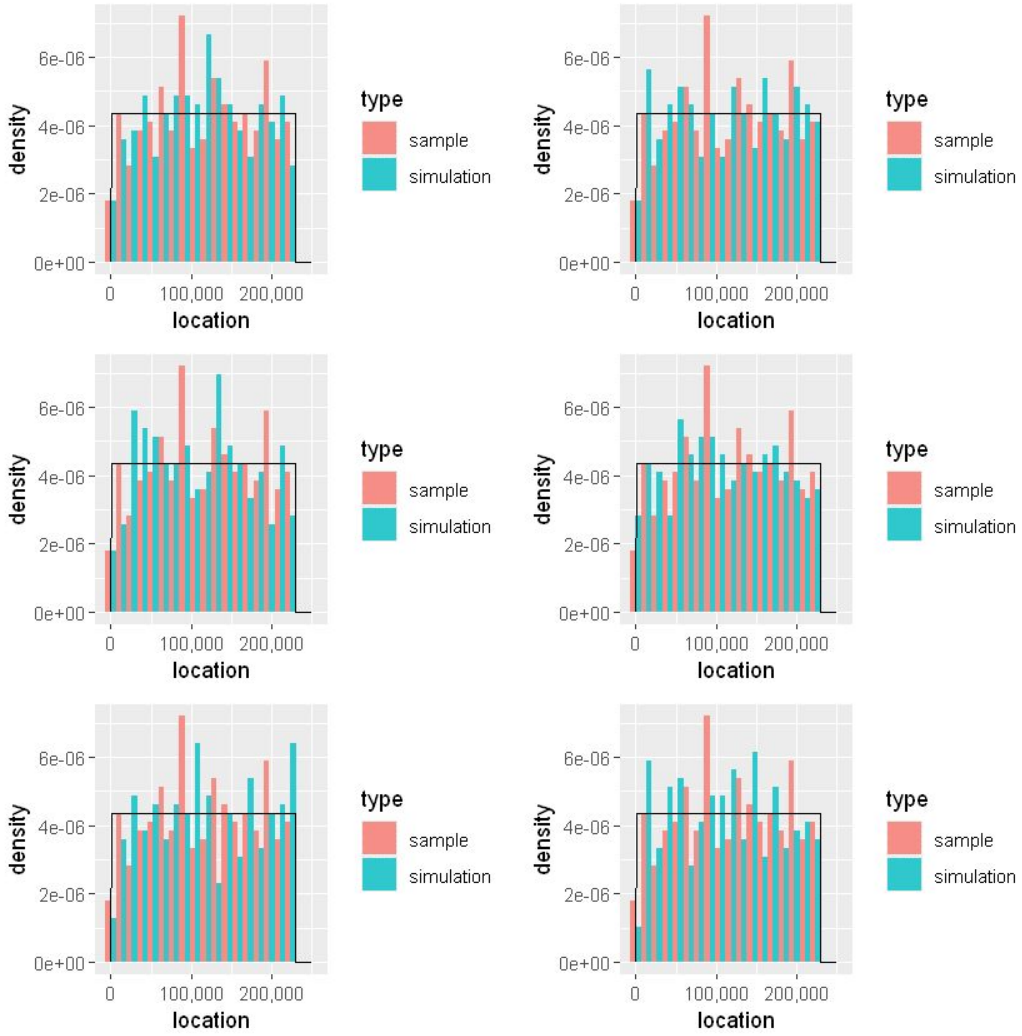


Figure 1: multiple histograms of sample and simulation locations, overlayed by pdf of uniform distribution

Each graph above is the density histogram plot for both sample and simulation locations, in which the black line indicates the pdf of uniform distribution in the interval [0,250000]. This pdf line is generated by using the dunif() function in R with the min value of 1, and the max value of 229,354. We observe that the pattern of sample and simulation are not similar, and the sample locations do not follow well on the shape of theoretical distribution. Thus we can not reasonably conclude that the location feature of the data comes from the same distribution with the random uniform scatters based on the above graphs. Furthermore, there are some obvious unusual intervals in the sample which contains more palindromes than the expected simulations. For example, in the interval [75000, 100000], the sample locations always have a higher density compared to the simulations. Therefore it is fair to suspect that there are some unusual clusters of palindrome within or near the interval [75000, 100000]. In order to further analyze the distributions in our sample data, we continue investigating other features in the following steps such as spacings and counts.

B. Spacing

In this step, we extract the spacing features from the locations in the CMV DNA data. There are three types of spacings extracted: spacings between consecutive palindromes, spacings between palindromes with one in between, and spacings between palindromes with two in between. The

goal here is to graphically examine the distribution of sample spacings by comparing these three types of spacings with random uniform scatters. In addition, identify the theoretical distributions of the spacings from random uniform scatters in order to obtain a general conclusion for the spacing features.

To achieve the goal stated above, we first compute the spacing vectors of three different types from the sample data. The first vector contains the difference between consecutive locations; the second contains the spacings between sums of pairs of consecutive; the third contains spacings of sums of triplets of consecutive. Similar to the location analysis above, we then choose to perform a simulation study which runs in a total five times. In each run, use runif() function to generate a random uniform scatter of size 296 which simulates the sample location data from a uniform distribution. After extracting the three spacing vectors from this scatter, plot the density histograms for both sample and simulation spacings in one graph, which is overlayed by the probability density function of the theoretical distribution. For each run in simulation, there are three graphs produced (for three types of spacings); in the end, there are fifteen graphs in total. There are two examples from the runs:

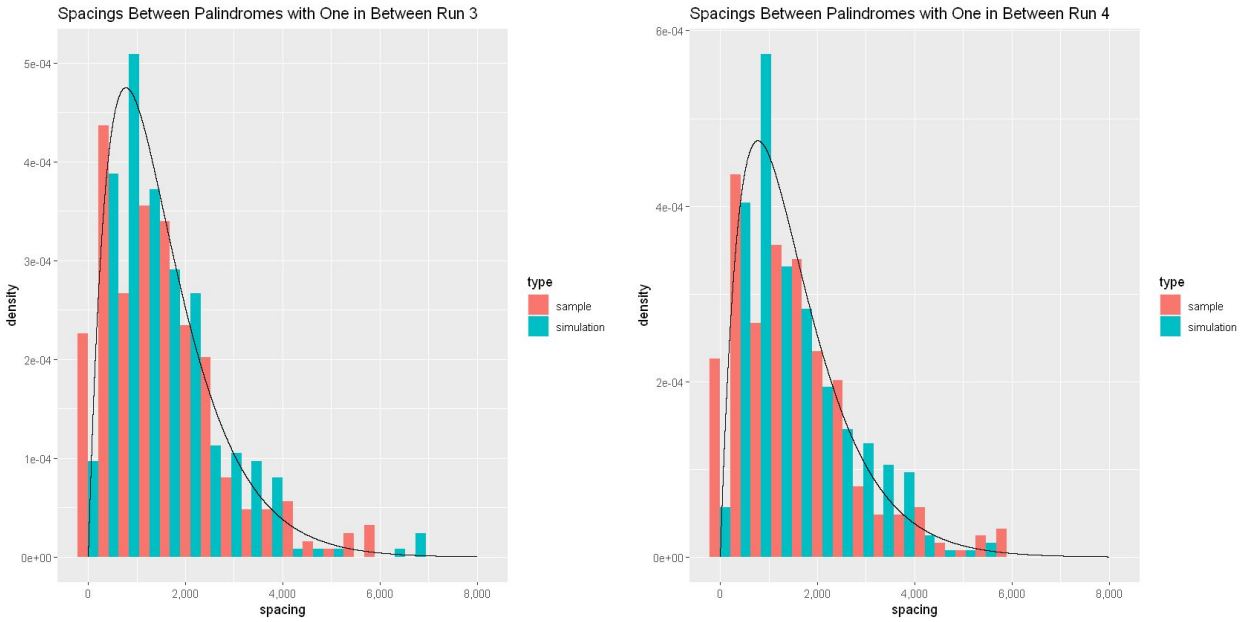


Figure 2: the histograms of sample and simulation pairs spacings in run 3 and 4

The above figures show the density histogram of sample and simulation spacings and the theoretical curve from the gamma distribution with corresponding parameters. It indicates that both distributions roughly follow the shape of gamma. In addition, we notice that there are still some odd bins always have a much higher or lower density in our sample data compared to the simulation data. It is reasonable to suspect these spacing intervals that imply the existence of unusual clusters of the palindrome.

To avoid redundancy, it is not appropriate to show all fifteen graphs here; we instead choose to produce cumulative density function for the spacing vectors (for reference, Appendix A shows other examples of the histograms with pdf). The procedure to compute the CDFs is similar to produce histogram and PDFs. However, instead of plotting histograms for both samples and simulations, employ the function stat_ecdf() in ggplot2 library to visually plot the empirical CDF for the input data. The results are shown below:

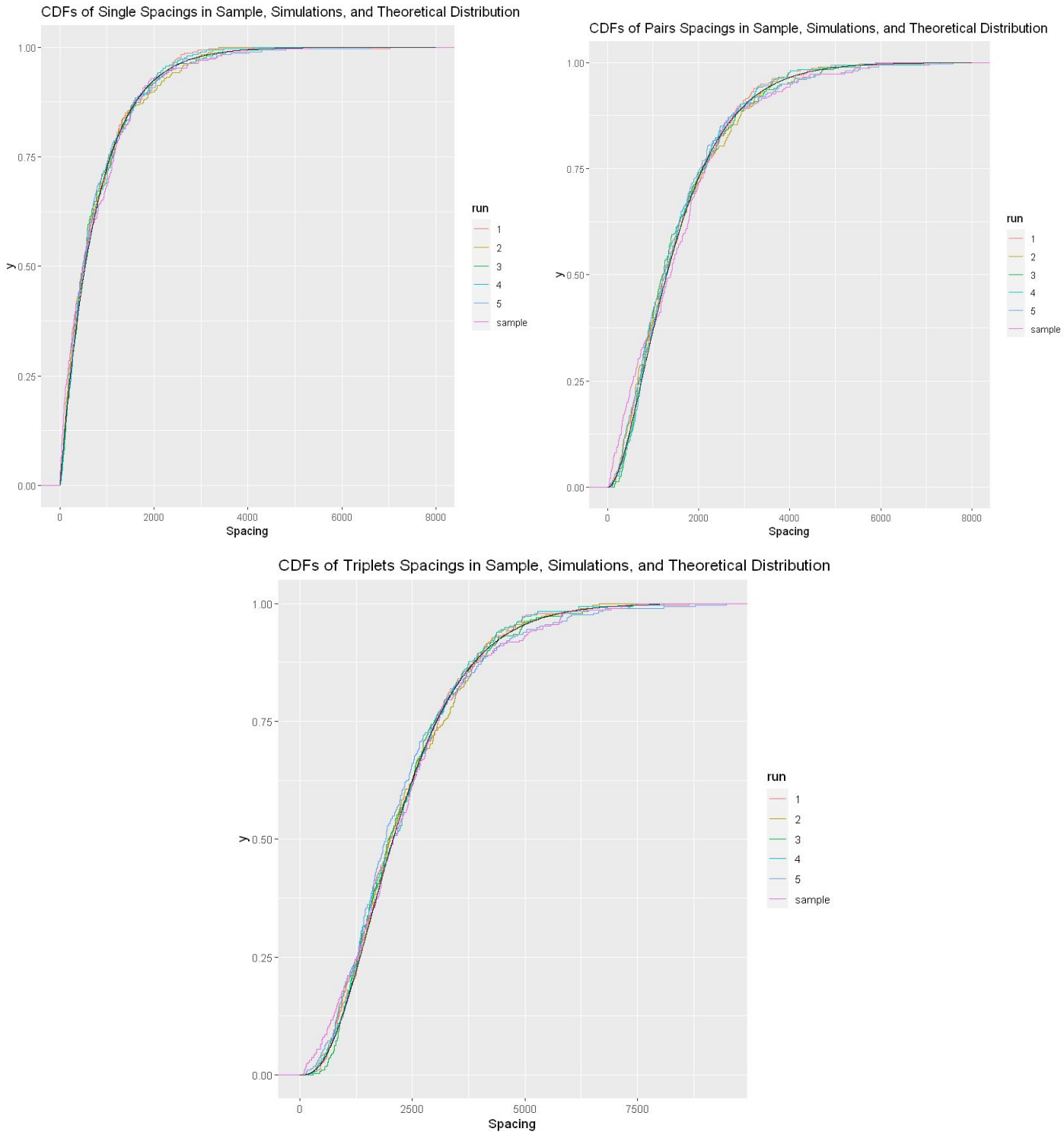


Figure 3: multiple plots of CDFs from spacings of sample, simulations, and gamma distribution

Under the Poisson process, it provides the property which states that the distance between consecutive hits would have an exponential distribution with parameter lambda. In particular, the exponential distribution is a special case of the gamma distribution with parameters 1 and lambda. For the distance between hist that is two apart away, it would also have a gamma distribution with parameters 2 and lambda. Therefore our theoretical distribution for the spacing of consecutive palindromes is the gamma distribution. The results from the simulation study also confirm this speculation. For each type of spacings, the graph contains the empirical CDFs for

the data in five simulations and the sample; in addition, the black curve which overlays on top is the CDF of gamma distribution with the corresponding parameters. In order to estimate the rate parameter lambda, we use the method of moments that substitutes the sample average for the expected value. Since the lambda is the expected number of hits per unit interval, it is reasonable to use the empirical average. In this case, the lambda would be 296/229354. By using the parameter lambda, compute the CDF of gamma distribution for three types of spacings: (1, lambda) for the single spacings; (2, lambda) for the pair spacings; for the triplet spacings, we intuitively choose the parameters (3, lambda). The graphs of CDFs indicate that both sample and simulation spacings are generally fitted well on the gamma distribution with corresponding parameters, including the distribution of triplet spacing. It is fair to conclude that all three types of spacings approximately follow the gamma distribution with some suspected bins which may reveal the unusual clusterings of the palindrome.

C. Count

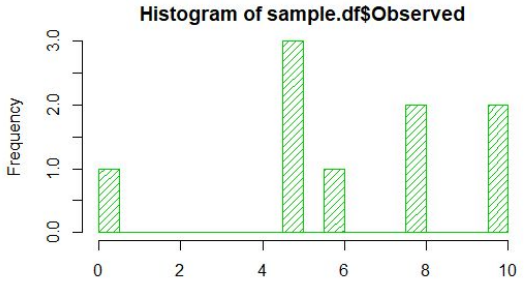
For this part, we split the data into intervals, then calculated the number of observations for each interval and did a hypothesis test to see whether the data distribution behaves in a way we expected.

We make a hypothesis that the distribution of position is random uniformly scattered. To begin, we create a genome sequence of length 229,354, with 296 palindrome sites to simulate the data. And in that case, the probability of finding k palindromes in an arbitrary interval should follow the Poisson distribution. For the unknown parameter lambda in Poisson distribution, we use the maximum likelihood estimator (mean) to estimate the lambda value, which is 296 / 57 = 5.

We split the data into intervals with lengths of 30, 57, and 100. Then we create histograms and tables to compare observed values and expect values.

levels <fctr>	Observed <int>	Expected <dbl>
0	0	0.3166586
1	6	1.6444023
2	5	4.2696762
3	5	7.3907846
4	5	9.5950537
5	10	9.9653891
6	8	8.6250151
7	8	6.3985075
>=8	10	8.7945128

(#interval = 57, interval length = 4000)



Finally, we do chi-square tests. Using the alpha value of 0.01, since the p-values of tests on different interval lengths are all greater than 0.01, we have a result which is that we failed to reject H0 and conclude that the observed data is modeled by Poisson distribution.

```
        Chi-squared test for given probabilities with simulated
        p-value (based on 2000 replicates)

data:  sample.trunc
X-squared = 6.0053, df = NA, p-value = 0.4123
```

(interval length = 30, 100. See Appendix C)

D. Biggest Cluster:

In this part, we will try different intervals to examine the probability of obtaining, in any subinterval, a count as large or larger than the count observed in the sample. In other words, we will do hypothesis tests using the chi-squared test through different interval sizes.

The null hypothesis would be that: The poisson model fits the data. This implies that the largest number of hits in a collection of intervals behave as the maximum of independent poisson random variables. The alternative hypothesis would be that : The poisson model does not fit the data.

While making the interval, we should avoid making really small intervals or large intervals. Because if the interval is too small , a cluster of palindromes may be split between adjacent intervals and not appear as a high-count interval. If the interval is too large, it is easy to go undetected if the regions examined are too large. Therefore, based on the 4000 interval length we tested in the previous steps , I decided to examine the interval with length [200, 2000, 4000, 5000, 10000].

result

interval_length	intervals	lambda	maxCount	p_value
200	1146	0.2582897	5	2.690642e-29
2000	114	2.5964912	12	2.596567e-125
4000	57	5.1929825	13	1.286927e-03
5000	45	6.5777778	15	1.320740e-01
10000	22	13.4545455	26	2.216363e-06

As you can see the table above, by setting our significance level a = 0.05, we can notice that there are 4 out of 5 p-values that are less than the significance level , which means the cluster we observed is larger than that expected from the poisson process. We reject the null hypothesis. The number of hits within a collection does not follow the poisson distribution.

E. Advanced Analysis:

Based on the analysis we performed above, we can not determine whether the clustering is unusual or by random because the count part told us the data does from the poisson distribution, while the biggest cluster part told us the poisson distribution fits the data. Thus, we can not really give the advice to the biologist whether the cluster is worth investigating or not.

For the Advanced Analysis part, we would like to construct a test to compare the variance of the observed dataset and the variance of simulated distribution of random uniform.

We start by claiming the null hypothesis:

$H_0 : \sigma_{obs} = \sigma_{sim}$

$H_\alpha : \sigma_{obs} \neq \sigma_{sim}$

Using $\alpha = 0.05$, we construct a F-test for variance. N = 296, M = 296.
We would expect $\frac{S_y^2}{S_x^2}$ to behave like an F random variable.
The test statistics is calculated as:
$F = \frac{S_y^2}{S_x^2} = 1.037$.
$F_{\alpha/2,m,n} = F_{0.025, 296,296} = 0.8$. $F_{1-\alpha/2,n,m} = F_{0.975, 296,296} = 1.26$.
Since 1.037 falls between the interval (0.8, 1.26), we failed to reject H0 and conclude that Variances are similar.

III. Importance and Conclusion

Our group successfully achieved our research goals by doing numerical analysis, graph visualizations and hypothesis tests.

For the Location part, we found that it is not enough to visually conclude that the location feature of the data comes from the same distribution with the random uniform scatters, which is the uniform distribution; and we suspect that there are some unusual clusters of palindrome within or near the interval [75000, 100000].

For the Spacing part, we have found that it is fair to conclude that all three types of spacings approximately follow the gamma distribution with some suspected bins which may reveal the unusual clusterings of the palindrome.

For the Count part, we have found that although there are outliers, the observed data is proved to be modeled by Poisson distribution.

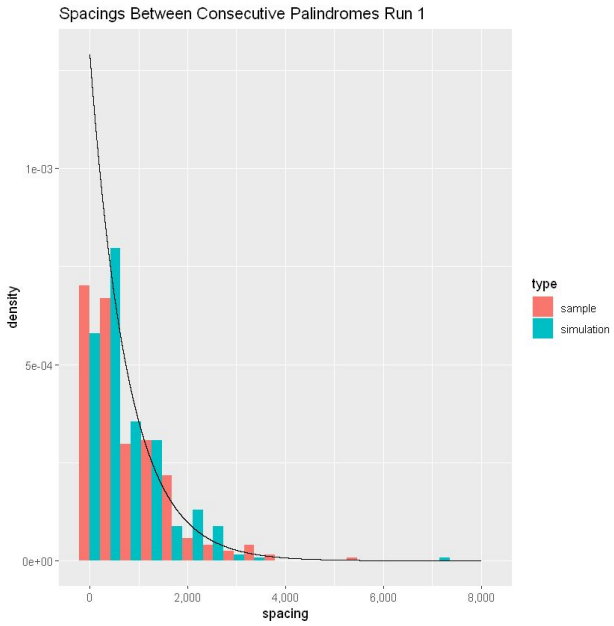
For the Biggest Cluster part,we tried different interval sizes and came to the conclusion that the poisson model does not fit the data we have.

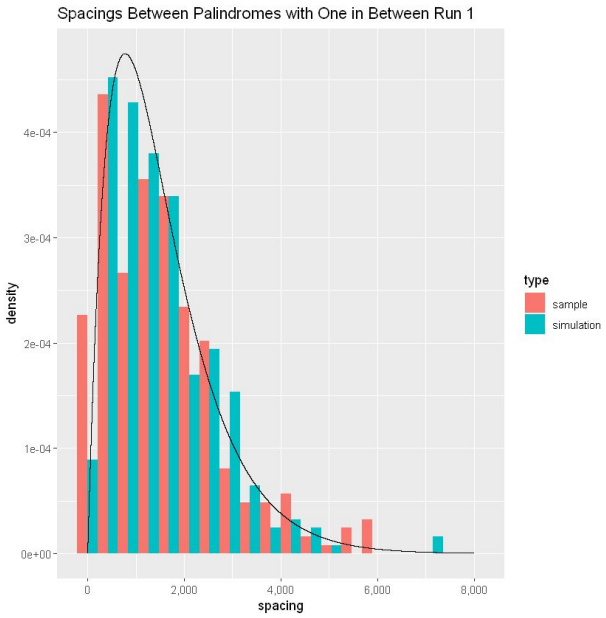
We have also discovered the variances are similar for the original dataset and the simulated data of random uniform distribution.

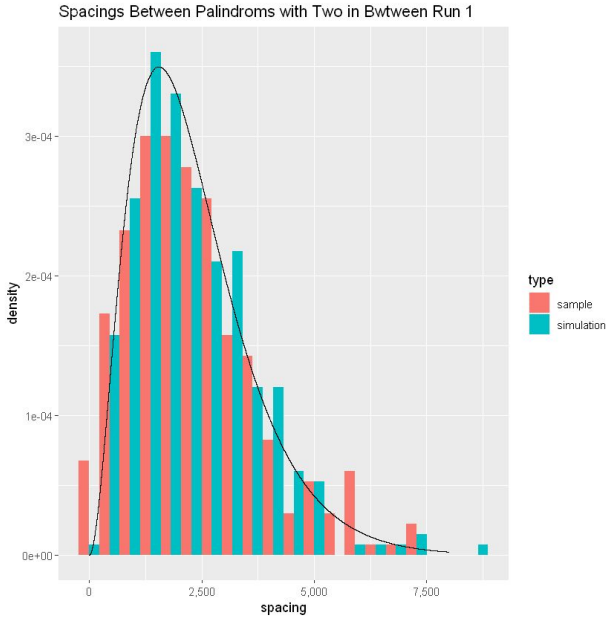
The limitation of our study is that our findings could not be generalized since we are restrained to the only given DNA location dataset and the dataset size is relatively small (296 out of 229354).

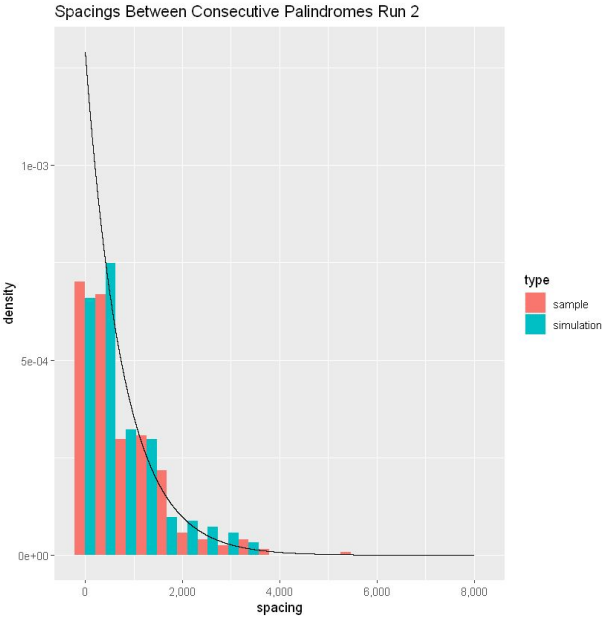
IV. Appendix

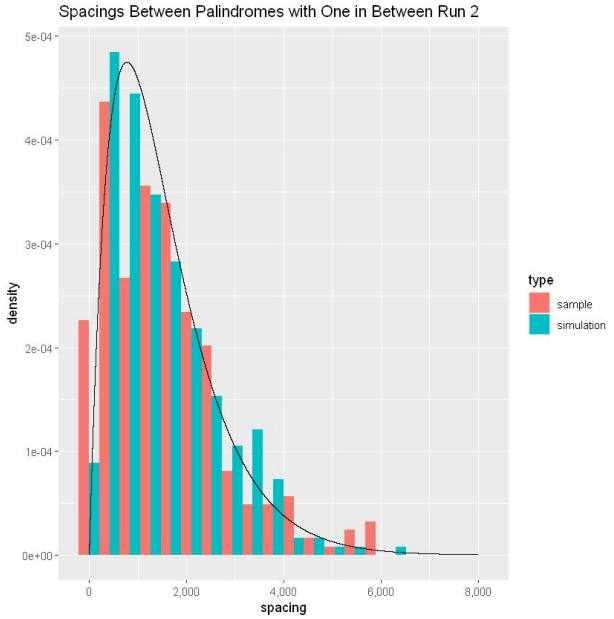
A. Histograms of sample and simulation spacings overlayed by PDF of the gamma distribution (from step 2 in analysis, showing the partial results (the first two runs) from the five runs)

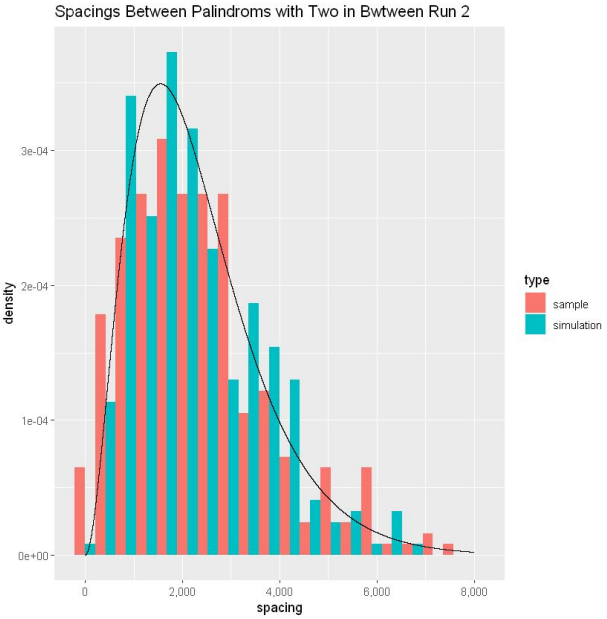

Spacings Between Consecutive Palindromes Run 1
density
1e-03
5e-04
0e+00
0
2,000
4,000
6,000
8,000
spacing
type
sample
simulation


Spacings Between Palindromes with One in Between Run 1
density
4e-04
3e-04
2e-04
1e-04
0e+00
0
2,000
4,000
6,000
8,000
spacing
type
sample
simulation


Spacings Between Palindroms with Two in Bwtween Run 1
density
3e-04
2e-04
1e-04
0e+00
0
2,500
5,000
7,500
spacing
type
sample
simulation

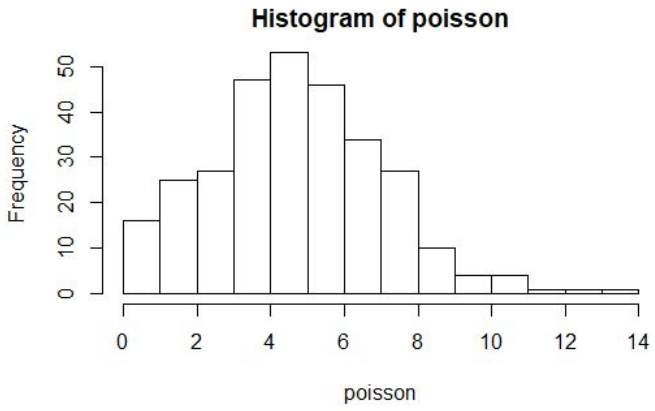

Spacings Between Consecutive Palindromes Run 2
density
1e-03
5e-04
0e+00
0
2,000
4,000
6,000
8,000
spacing
type
sample
simulation


Spacings Between Palindromes with One in Between Run 2
density
5e-04
4e-04
3e-04
2e-04
1e-04
0e+00
0
2,000
4,000
6,000
8,000
spacing
type
sample
simulation


Spacings Between Palindroms with Two in Bwtween Run 2
density
3e-04
2e-04
1e-04
0e+00
0
2,000
4,000
6,000
8,000
spacing
type
sample
simulation

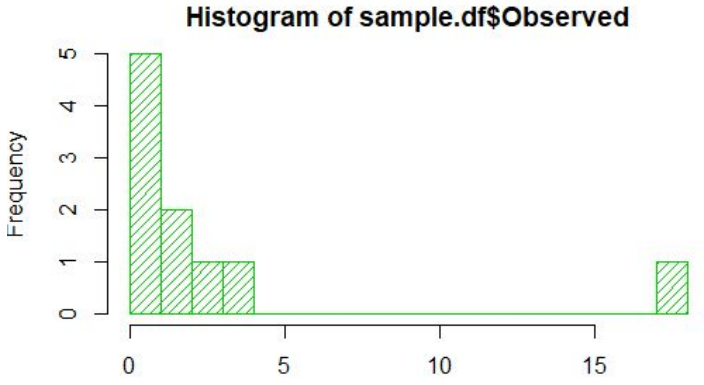
B.

C. Poisson Distribution with lambda = 5



30 intervals

levels <fctr>	Observed <int>	Expected <dbl>
0	0	0.001556261
1	0	0.015355106
2	0	0.075751857
3	1	0.249139441
4	2	0.614543954
5	0	1.212700069
6	3	1.994217891
7	2	2.810897599
8	4	3.466773705
>=9	18	19.559064117

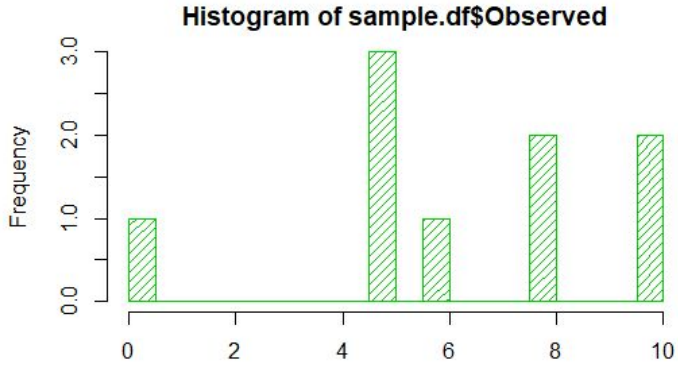



```
        Chi-squared test for given probabilities with simulated
        p-value (based on 2000 replicates)

data:  sample.trunc
X-squared = 7.6392, df = NA, p-value = 0.2934
```

57 intervals

levels <fctr>	Observed <int>	Expected <dbl>
0	0	0.3166586
1	6	1.6444023
2	5	4.2696762
3	5	7.3907846
4	5	9.5950537
5	10	9.9653891
6	8	8.6250151
7	8	6.3985075
>=8	10	8.7945128

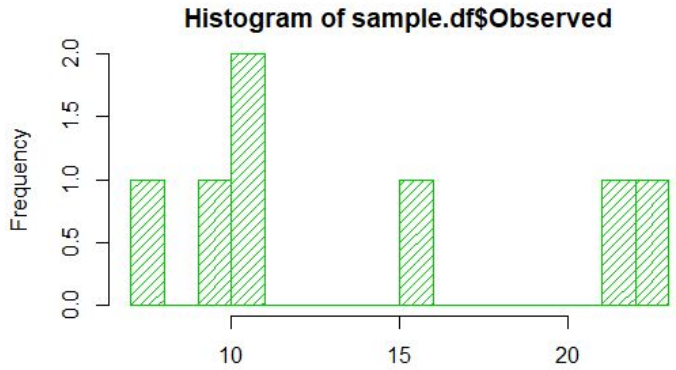



```
        Chi-squared test for given probabilities with simulated
        p-value (based on 2000 replicates)

data:  sample.trunc
X-squared = 15.564, df = NA, p-value = 0.05047
```

100 intervals

levels <fctr>	Observed <int>	Expected <dbl>
0	10	5.181892
1	11	15.338399
2	23	22.700831
3	22	22.398153
4	16	16.574634
5	11	9.812183
>=6	7	7.993907



```
        Chi-squared test for given probabilities with simulated
        p-value (based on 2000 replicates)

data:  sample.trunc
X-squared = 6.0053, df = NA, p-value = 0.4123
```

V. Contribution Statement

Yunlin worked on the introduction, locations, and spacings.
Jian worked on the counts, partially advancing analysis and conclusion.
Yong Liu worked one the biggest cluster part and organize some of the conclusion.