MATH 189 Project 3

Group 29

February 2021

1 Introduction

This lab will focus on the DNA of the cytomegalovirus. Human cytomegalovirus (CMV) is a common disease that affects people of all ages. According to the Center for Disease Control and Prevention, half of people over 40 have infected CMV. Though most people show no symptoms after infection, it's life-threatening for those who have a deficiency in the immune system. To combat the CMV, scientists are trying to find the potential site of the virus replication.

To examine the replication of the virus, it's vital to start with DNA. In the herpes family, which CMV belongs to, the Complimentary palindrome has been a key pattern for the replication. Since it's time-consuming to examine the palindromes one by one, in this lab we look at the clusters of the palindrome, trying to detect unusual clusters of the complimentary palindromes.

In this lab, we will first simulate a random scatter of palindromes, and actively compare the simulated data with the observed data in terms of the locations and spacings, counts, and the biggest clusters of the palindromes.

2 Data

The dataset we use is *hcmv.txt* provided by professor Jelena Bradic. The set contains 1 columns and 297 observations. Since the last row of this dataset is NA, we omit it and use the remaining 296 observations for analysis. The "location", which is the only variable of this dataset, is a discrete numerical variable. It provides us an insight on the sites of DNA bases which the palindromes locates on. Since the bases of DNA is numbered from 1 to 229335, the observations of "location" consists of 296 numbers in this range. Based on this variable, we derive "observed", the number of counts of palindromes per interval, which is used for further investigations on the clusters of palindromes.

3 Background

A human being has more than 3 billion genes in one body. The DNA contains massive genetic information and has a unique structure of a double helix structure with two chains. Many small units-nucleotides are inside the DNA and they contained one of the four nucleobases, which are adenine, thymine cytosine and guanine. Those bases paired up another base on the other chain with certain rules. Adenine combines with thymine; and cytosine combines guanine. All sugars in the DNA are called deoxyribose. Although DNA for viruses do not have as much DNA pairs as human beings, their basic structure and functionality are similar to them.

In 1953, Watson and Crick first solved the three-dimensional structure of DNA which contributed to the conceptual framework for DNA replication and protein encoding within nucleic acids. However, the high development age of DNA did not start for the information contained in DNA molecules were much longer in nuclei acid and were similar to one another. It was not until 1972, when Walter Fiers' and his colleagues completed the first complete protein-coding gene sequence by 2-D fracnation method, more researchers from different fields began to devote themselves to sequence DNA. Computational sequence DNA is now one of the best choices for DNA study to found out the unusual segment because of an excess of information DNA is able to provide.

4 Investigation

4.1 Random Scatter

To examine the locations of the palindromes, the spacing between palindromes, and the counts of palindromes in non-overlapping regions of the DNA, we first simulate the palindromes which scatter randomly on the DNA. The simulated data has 296 palindrome sites randomly distribute on a DNA sequence of 229,335 bases.

To visualize our simulated data graphically, we generate the strip plot as shown below.

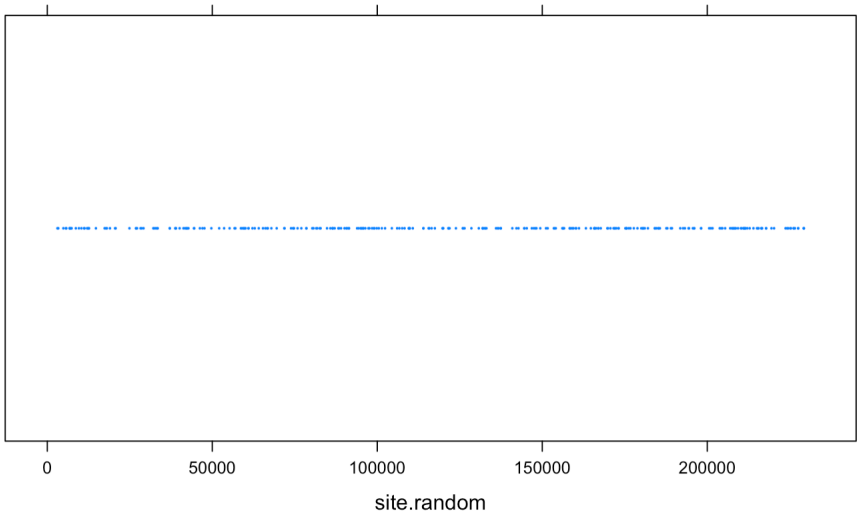

Figure 1. The strip plot of simulated data

We observe that the points are scattered uniformly across the number line, which verifies our assumption that the sites of palindromes are chosen uniformly among 229,335 sites. We also observe that there might be some clusters of palindromes around the sites 100000 and 175000 to 220000. The part of 4.4 gives further study on clusters of both observed data and simulated data.

To compare the simulated data we generate to the observed data, we put their histograms on one graph as showed below.

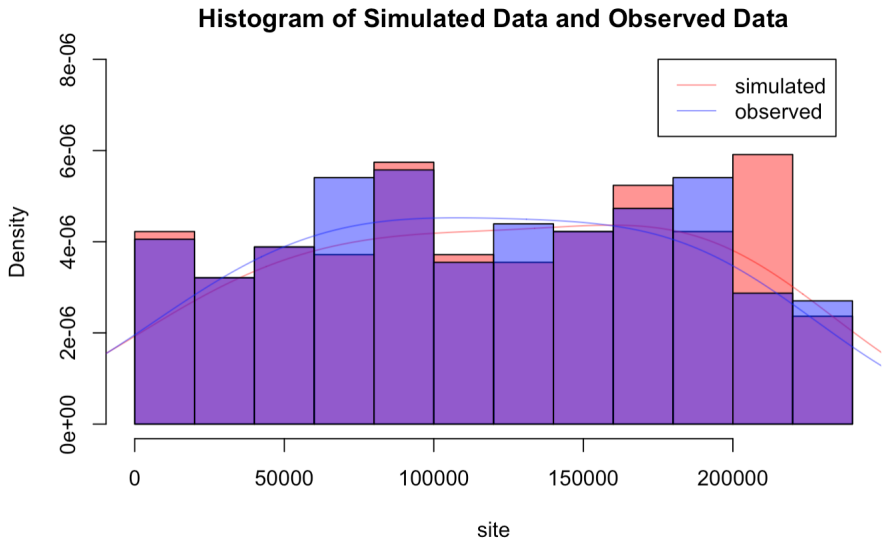

Figure 2. The histogram of both simulated data and observed data

If we just look at the simulated part of the plot, the histogram also shows that the simulated data has possible clusters around the sites 100000 and 175000 to 220000, which is the same as what we observed in the previous plot.

If we consider the plot as a whole, we observe from the histogram that their density lines are very closed, and we will further their distributions and differences in the following parts.

4.2 Locations and Spacings

In this part, we used graphical method to explore the space distribution of the DNA position data. First, in order to make a direct view of the spaces between consecutive palindromes, we created a scatter plot and noticed that at around position 230000, one obvious possible outlier existed.

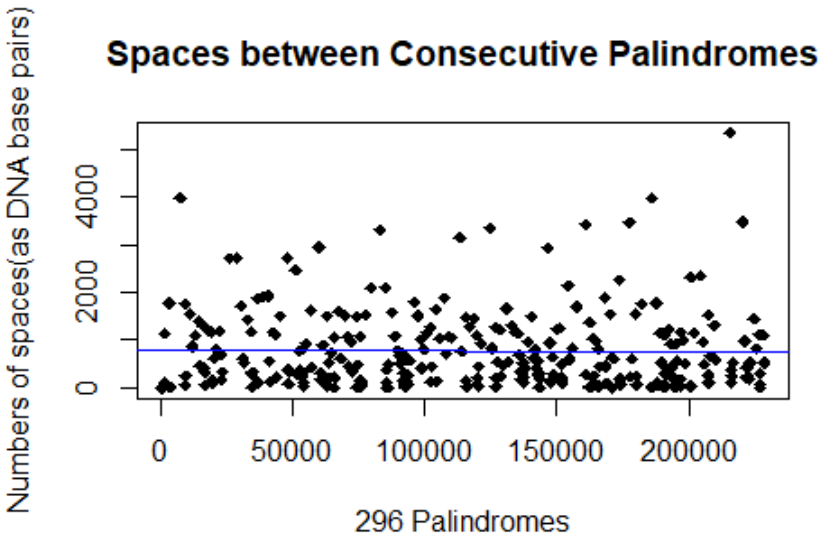

Figure 3. The scatter plot of consecutive palindromes

Then, we did further exploration on the spacings of the sum of consecutive pairs and triplets to check whether they had similar patterns as the spacing of the single consecutive palindromes.

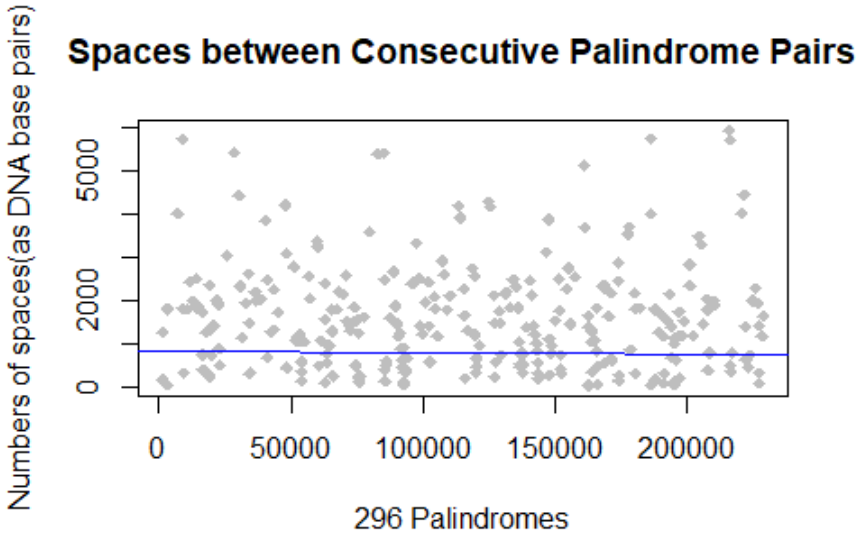

Figure 4. The scatter plot of consecutive palindrome pairs

We noticed that the dots were less condensed around the linear line in the scatter plot above. Several more possible outliers existed here, and they were at about 20000, 80000, 180000, 230000. However, the possibility of the outliers here was smaller than the possible outlier in the single consecutive palindromes graph.

Next, we did the scatter plot on the spaces between consecutive palindromes triplets.

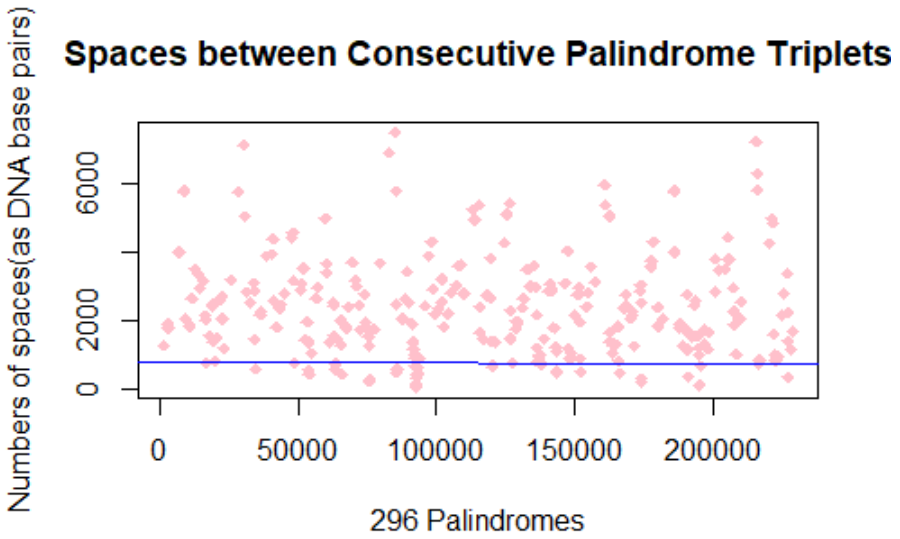

Figure 5. The scatter plot of consecutive palindrome triplets

The possible outlier could be at around 25000,80000,230000 from the triplet graph.

From the three different type of spacings we did for the location checking, we found out that the most possible place to produce an outlier, which could be the possible unusual cluster, was at around 23000. Also, positions close to 2000 and 80000 could exist outliers.

Comparing the position data of unusual clusters we got in this section to Section 4.1 Random Scatter, we saw that this section's most possible unusual cluster position 230000 did not overlapped with the range we got from the simulated data, which was 100000 and 175000 to 220000, but it was close to the previous range. Position 25000 and 80000 we generated in this section neither fit the range we got in simulated dataset.

Therefore, we wanted to take a close look at the graphical comparison of the three spacings.

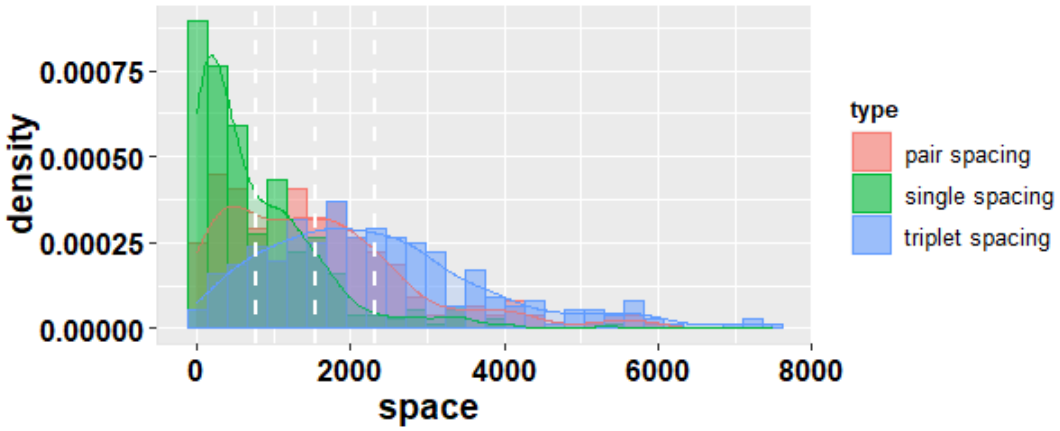

Figure 6. Comparison: Histogram with density curve

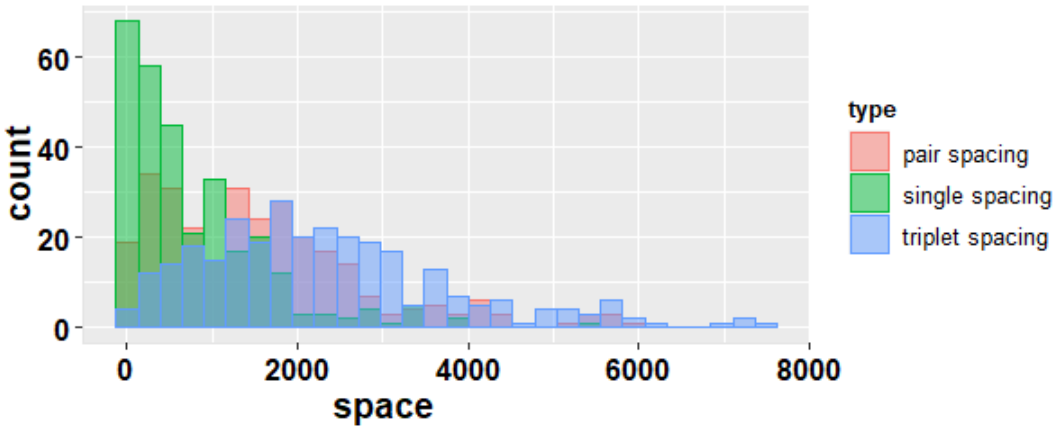

Figure 7. Comparison: Histogram with no density curve

The two graphs above were both skewed to the right. Some unusual clustering could happen at the ones of large spaces in different spacings. For single spacing, its largest spaces was around 5200. Whereas for the triplets, the largest spaces could almost reach 8000.

4.3 Counts

In this problem, we first found the applicable number of intervals to split the data using the equation

$$r = 2 * n^(2/5). \tag{1}$$

The result we found is 19.478103988461 intervals. In fact, we found that we could use 23 intervals and could ensure at least 5 counts be included in each interval in the purpose of performing chi-squared test, we finally decided to utilize 23 intervals. The size (length) of each interval is thus 10000.

Then we make a comparison between the observed data and randomly generated data. We drew two histograms to do that.

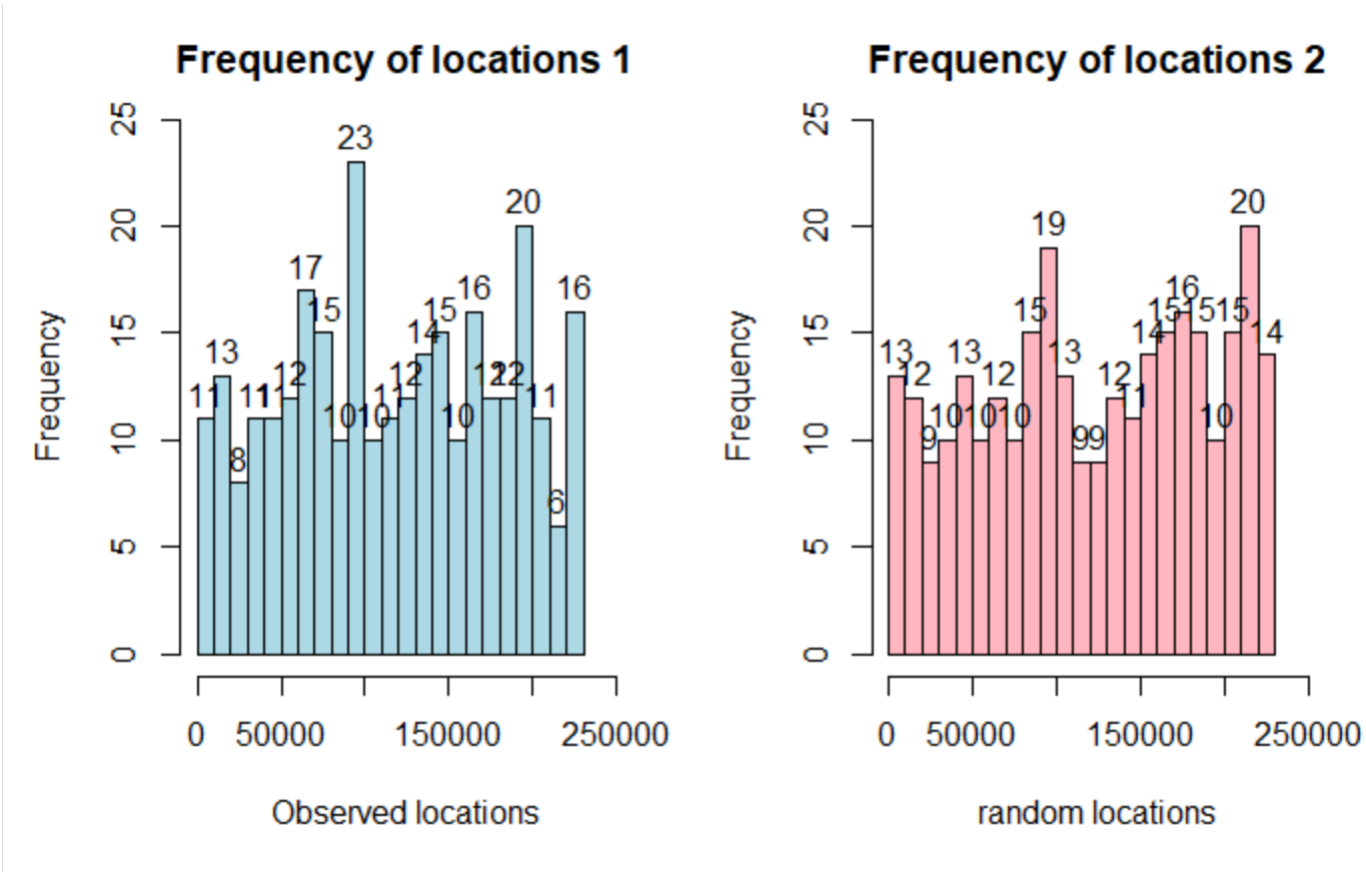

Figure 8. Frequency of locations for observed and random data

In comparison of these two graphs, we observe that the observed data are more extreme than randomly generated data, which look more like a uniform data. It's not clear to state whether there is a tail or modal in these graphs, but clearly the biggest cluster in observation is 90000 - 100000, with 23 total counts, while the corresponding randomly generated bin has 19 counts; In addition, bin 190000 - 200000 has 20 versus 10 counts, while bin 210000 - 220000 has 6 versus 20 counts. The results we discovered are that these two datasets are not the same.

However, from the two comparing distributions put into the same picture, we found these two distributions formed from the observed data and data generated from Poisson distribution are somewhat similar.

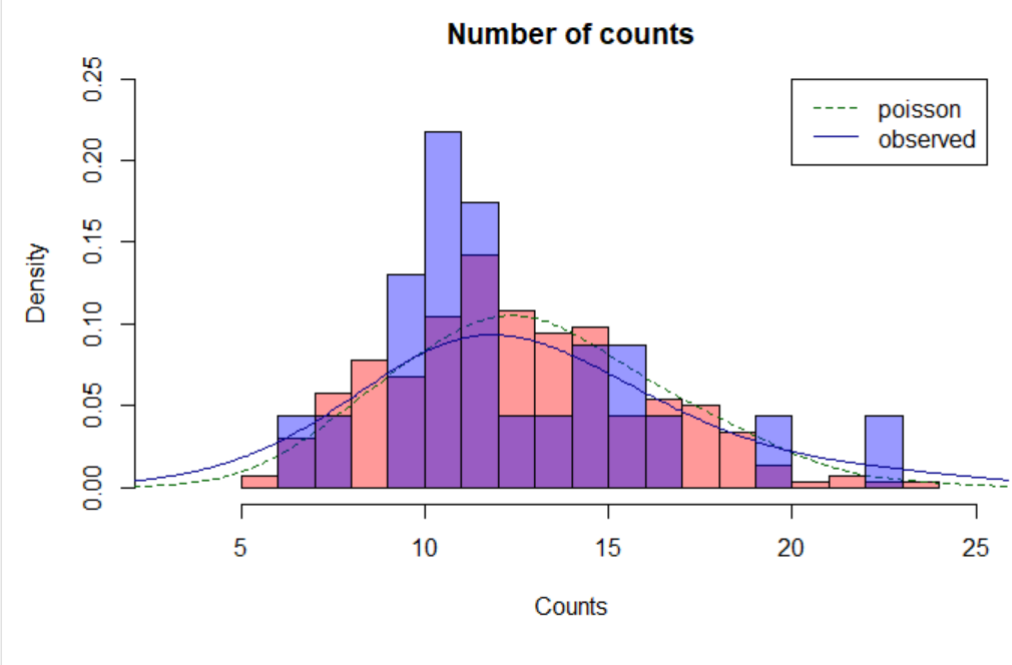

Figure 9. Observed and Poisson generated number of counts

In this graph, the lower distribution with red color is formed from the observed data and the higher distribution with purple color is formed from randomly generated data. From the curve and the overall trend shown in the graph we can observe that they look similar: they both take small number of counts at tails and large number of counts at the center around 8 - 12 counts per bin. Conclusively, we find that the overall shape of the observed data is closer to the Poisson distribution than a randomly generated data from graphical analysis.

Then we constructed the chi-squared test to test the existence of difference between observed data and each of randomly generated data and poison data as a rigorous test.

```
        Chi-squared test for given probabilities

data:  observed
X-squared = 36.288, df = 22, p-value = 0.02827

Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  observed and poiss
X-squared = 149.35, df = 143, p-value = 0.3412
```

Figure 10. Chi-squared test between observed data and randomly generated data/Poisson generated data

The p-value of the test between observed and randomly generated data is 0.02827, while the one between observed and Poisson generated data is 0.3412. At a confidence level of 95%, The conclusion is that the observed data is statistically different from the randomly scattered data while we fail to reject the null hypothesis that it is not significantly independent from the data generated from Poison distribution. The conclusion appears as what we expected.

Based on the above hypothesis tests, we would like to perform a sensitivity analysis of our methods used on both distributions. We chose to separately and slightly change the size of the intervals by adding or subtract a small amount of intervals and then perform the hypothesis tests again.

```
        Chi-squared test for given probabilities

data:  observed.dec
X-squared = 68.253, df = 23, p-value = 2.257e-06


        Chi-squared test for given probabilities

data:  observed.inc
X-squared = 52.494, df = 21, p-value = 0.0001625
```

Figure 11. Chi-squared test between observed data and randomly generated data for sensitivity analysis

The p-values are both smaller than 0.05. By both increasing and decreasing the intervals, we would reject the null hypothesis that the observed data and randomly generated data are statistically the same at a 95% confidence level. Then by sensitivity analysis, we found that the result is consistent and solid to be applied to build next step of our research.

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  observed.dec and poiss.dec
X-squared = 175.47, df = 176, p-value = 0.4972

Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  observed.inc and poiss.inc
X-squared = 216.33, df = 192, p-value = 0.11
```

Figure 12. Chi-squared test between observed data and Poisson generated data for sensitivity analysis

By both increasing and decreasing the intervals, we would fail to the null hypothesis that the observed data and Poisson generated data are statistically the same at a 95% confidence level. Then by sensitivity analysis, we found that the result is consistent and solid to be applied to build next step of our research.

Based on these tests, we found that the results would not change and are solidly prepared for our next step.

Additionally, we formulate a way to classify the data based on the difference between observed and randomly generated data.

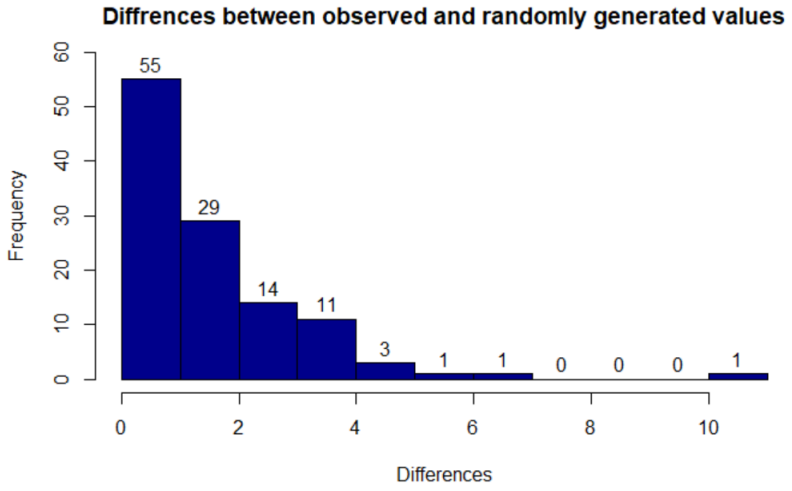

Figure 13. Differences between observed and randomly generated values

We generate a new model that uses our observed data to pair and subtract randomly generated data. Then we calculate the standard deviation of the data. We then classify the data: If there is no difference, it is classified as region type 1. Then if the value is between 0 - 3.575663 (1 standard deviation above mean difference), it is classified as region type 2 (98 of them). Region type 3 is defined as value between 4 - 5.238282 (2 standard deviation above mean difference) while region type 4 is those differences equal or larger than 6. It is also important to notice that one of the difference is an obvious outlier. We may find interesting to investigate it.

4.4 The Biggest Cluster

Our goal is to test if the interval with the greatest number of palindromes indicate a potential origin of replication.

We use smaller intervals of size 2000 since we intend to examine the specific location of cluster. We check if replication occurs in this cluster. In other words, we test if this interval with maximum palindromes could happen in randomly scattered data. cannot perform chi-squared test since some bins have less than 5 elements. In this case proportional test to examine exceptional bins.

We observed that potential clusters occurred at location around 93000 and 195000. We need to test if repetitions could happen in these two locations. We apply proportion test and obtain the p-value of 0.00252 for the first location. Under 95% confidence level, we could say that we reject the null hypothesis and conclude that it is statistically not likely that randomly scattered samples would contain our maximum. For the second location, the p-value we obtain is 0.1908, which is greater than 0.05. As a result, we fail to reject the null. It means that randomly scattered sample could generate cluster like that at position 195000. We also check the sensitivity for exceptional bins. We first adjust the interval size to 2100 instead of 2000. By performing the same proportion test above, we obtain p-values of 0.004457 and 0.07239. The first is smaller than 0.05 while the second is greater. This accords our findings above. We also tried the interval size of 1900 and obtain p-values of 0.01243 and 0.1117, so the conclusion is consistent. Thus, we contend that the interval size of 2000 is not unique. We could conclude that the only cluster that is the site of repetition occurs at position 195000.

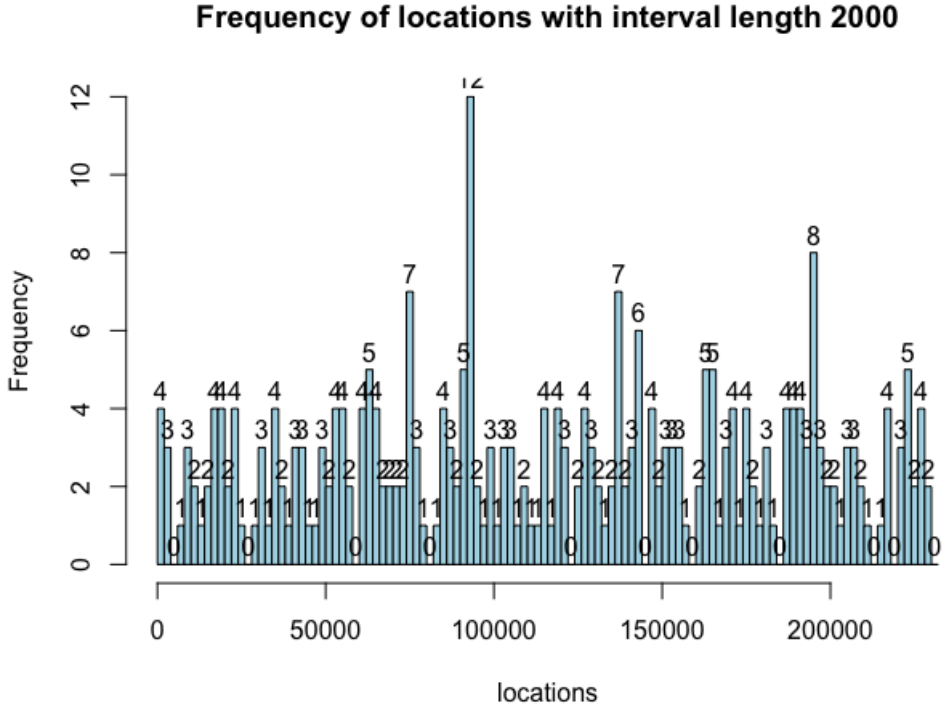

Figure 14. Frequency of locations with interval length 2000

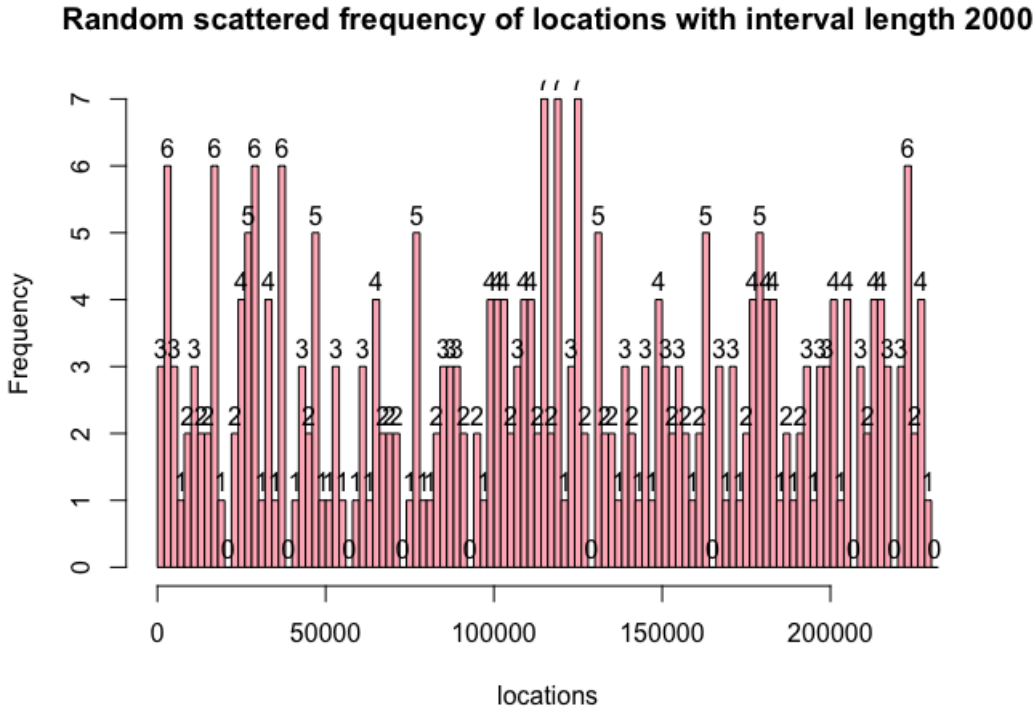

Figure 15. Random scattered frequency of locations with interval length 2000

To further assure our findings, we would test through simulations to make sure that the cluster at around 93000 is the site of repetition. We find the maximum occurrence of palindromes in our data set divided into intervals of size 2000. Then, we compare the maximum we find with data generated through simulation of random scatters 1000 times with interval size of 2000. Our null hypothesis would be that under random scatter In this case, we find that the chance of occurrence, which is the estimation

of p-value, is 0.003. Under a 95% confidence level, we could statistically deduce that this doesn't occur in random scatters. Figure 10. shows the distribution of random sample maximum we generated with interval size of 2000.

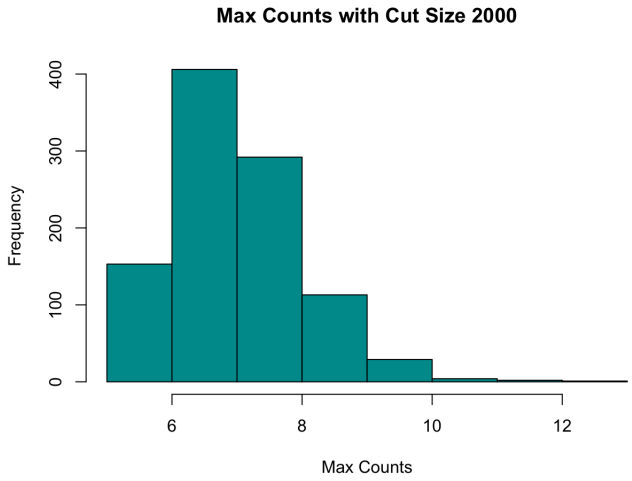

Figure 16. The Max Counts of Simulated DNA with Cut Size of 2000

We also conduct the same test but divide our DNA into intervals of size 1000 and 500. We find that the chance of occurrence, which is the estimation of p-value, is 0.011 and 0.011. This result is slightly greater than that of the 2000-sized intervals, and it still overrides the null hypothesis under 95% confidence level. Figure 11. and Figure 12. show the distribution of sample maximums with interval size of 1000 and 500. As a result, we could conclude that even when we generate randomly scattered data many times, the chance of obtaining a sample maximum as we observed in our data is statistically small. We can say that repetition does occur in the position of maximum occurrence, which is the cluster around 93000.

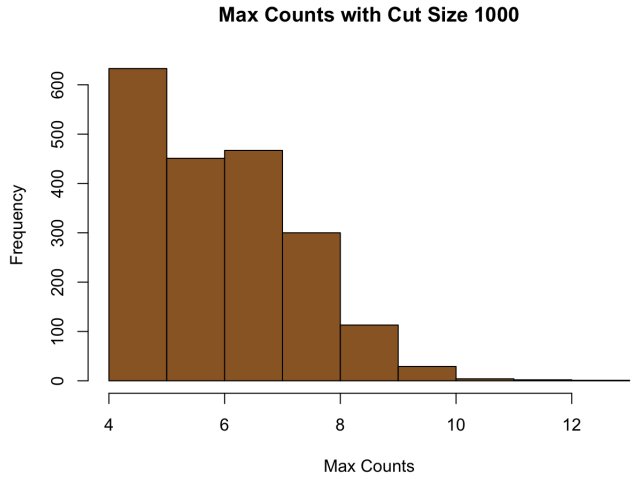

Figure 17. The Max Counts of Simulated DNA with Cut Size of 1000

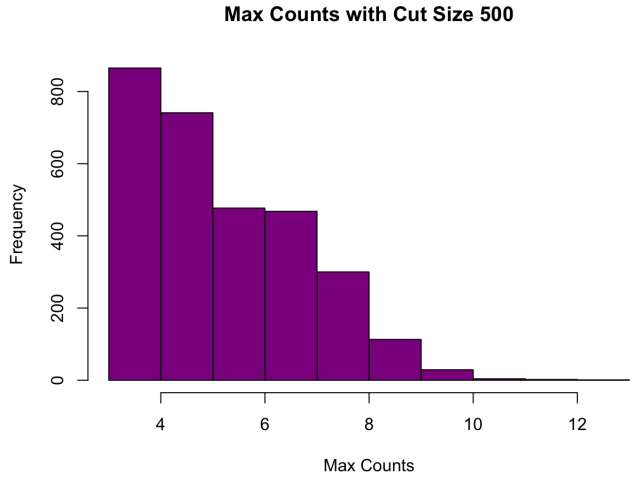

Figure 18. The Max Counts of Simulated DNA with Cut Size of 500

5 Conclusion

In order to discover the answer for the main question, we mainly separate the big question into 4 topics. In the first one, we use the dataset provided by professor Jelena Bradic to stimulate the palindromes based on the random scatter corresponding to the location of spacing and construct the histograms to compare the stimulated data and observed data. The simulation will be further applied in the following three topics.

In the second one (Location and Spacings), we use graphical method to explore the space distribution of DNA location data by first creating a scatter plot of consecutive palindrome, a scatter plot of consecutive palindrome pairs, and a scatter plot of consecutive palindrome triples. As we noticing the location of number lines where the dots condensed, we want to make the comparison between different clusters of these three spacings more direct. Hence, we construct two histograms showing the spacing of three variables with density curve and with no density curve and find skewed right tendency.

In the third topic (Counts), we first create the number intervals and then construct three tests: the first one is test the difference between observed data and randomly generated data and difference between observed data and poison data with the constructed histograms; the second is the chi-squared test which is further used to test the difference among the three datasets; the last test aims at testing the sensitivity of our results by slightly changing the size of intervals.

In the last topic (The Biggest Cluster), the goal is to test whether the interval with the greatest number of palindromes indicate a potential origin of replication, so we try to construct the test which can examine the exceptional bins. In order to ensure our result is trustworthy, we further test through the simulations finding the maximum occurrence of palindrome in the dataset and the chance of occurrence. Based on the research part showed above, we finally find the replication site is in the interval of 92000-94000 as the corresponding answer to the main question.

6 Theory

6.1 Hypothesis Testing for Proportion

Goal of Hypothesis Testing is to determine whether the data provide enough evidence to conclude that some "null hypothesis" about a parameter is false and some "alternative hypothesis" is true.

H0 : there will be no observed effect for an experiment

H1: there will be an observed effect for the experiment

It's test statistics for population proportion are defined as the following:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{2}$$

Then the p-value is calculated by the test statistics above, which is the probability, if H0 is true, of getting a value for the test statistic that is at least as extreme (in the direction of the alternative hypothesis, i.e., favorable to H1) as what was actually observed.

6.2 Histogram

Histogram is a straightforward visual estimator of probability density function. As we have a dataset, we can categorize the data points into a number of bins and use the number of data collected in each bin to reconstruct the PDF. In fact, histogram follows the philosophy of using sample data to estimate population distribution. The choice of the number bin depends on the number of samples in the data set. Theoretically, the optimal choice of the bins that balances the bias and variance trade off is $^3, where n is the number of the dataset.$

6.3 Chi-squared Goodness-Of-Fit Test

A chi-squared test is a hypothesis test used to determine whether a sample follows the chi-squared distribution. Researchers can use the chi-squared test to determine whether two groups of categorical data are statistically independent or not, and whether the data fall in bins are following the desired distribution. In this case study, the chi-square goodness of fit test is used to determine if the observed data of palindromes follows the certain distribution. Then

Continuous Distributions:

$$H_0 : f_Y(y) = f_0(y)$$

$$H_1 : f_Y(y) \neq f_0(y)$$

Discrete Distributions:

$$H_0 : p_1 = p_{1_0}, p_2 = p_{2_0}, ..., p_t = p_{t_0}$$

$$H_1 : p_i \neq p_{i_0} \text{ for at least one i}$$

Test statistics:

$$X^2 = \sum \frac{(obs - exp)^2}{exp} \tag{3}$$

Using the t-test statistics we can obtain the corresponding p-value. If H0 is true, then the p-value will be large, and vice versa for H1.

6.3.1 Testing for Poisson Distribution

We can use Chi-Squared Goodness-Of-Fit Test to test if the observed data follows a Poisson Distribution. Poisson distribution is a discrete distribution. Then

H0 : The observed data follows a Poisson Distribution

H1: The observed data does not follow a Poisson Distribution

In other words

$$H_0 : p_1 = p_{1_0}, p_2 = p_{2_0}, ..., p_t = p_{t_0}$$

$$H_1 : p_i \neq p_{i_0} \text{ for at least one i}$$

and

$$p_j = Prob(X = j)$$

Test statistics:

$$X^2 = \sum \frac{(\hat{V}_j - n * \hat{p}_j)^2}{n * \hat{p}_j} \tag{4}$$

6.3.2 Testing for Uniform Distribution

We can use Chi-Squared Goodness-Of-Fit Test to test if the observed data follows a Uniform Distribution. Uniform distribution could be either discrete or continuous. Then

H0 : The observed data follows a Uniform Distribution

H1: The observed data does not follow a Uniform Distribution

For discrete Uniform Distributions:

$$H_0 : p_1 = p_{1_0}, p_2 = p_{2_0}, ..., p_t = p_{t_0}$$

$$H_1 : p_i \neq p_{i_0} \text{ for at least one i}$$

Test statistics:

$$X^2 = \sum \frac{(\hat{V}_j - E(\hat{V}_j))^2}{E(\hat{V}_j)} \tag{5}$$

6.4 Decision Errors

Decision from the Test

	Fail to reject H0	Reject H0
H0 is true	This is correct	Type I Error
H1 is true	Type II Error	This is true

6.4.1 Type I Error

Type I Error is the error that when the null hypothesis is in fact true, the decision we make based on the statistical test yields to reject the null hypothesis. This is due to the fact that confidence interval does not guarantee the true value will 100% fall into the interval, but tolerate a failing level alpha. This alpha is the possibility of the occurrence of Type I Error.

6.4.2 Type II Error

Type II Error is the error that when the alternative hypothesis is true, the decision we make based on the statistical test is "fail to reject the null hypothesis". This is not due to a mistaken testing process, but a naturally existing drawback of statistical hypothesis testing. We commonly denote Beta as the probability of the occurrence of Type II Error.

6.4.3 Power

The power of a statistical hypothesis test is the probability of rejecting the null hypothesis when the alternative hypothesis is true. This power is equal to 1 – Beta, the Type II Error. As the power of a test increases, possibility of Type II Error decreases. In addition, we can say power of a test is a curve depends on the alternative hypothesis. Power ranges from 0 to 1. As power increases, we can simply claim that the possibility of rejecting a wrong null hypothesis and follow the alternative hypothesis is high.

6.5 Testing for Exponential Distribution

When we draw a exponential distribution, the distance between two consecutive palindromes (locations of palindromes) has exponential distribution (λ)

$$p[distanceb/wfirst"hit"andsecond"hit" > t] = exp^{-*t} \tag{6}$$

There are no palindromes in any interval of length t.

6.6 Gamma Distribution

6.6.1 Gamma Distribution (2,λ)

Distance "b/w" consecutive pairs of palindromes and there is one palindrome in each interval of length t.

$$H_0 : Distance"b/w"pairsofpalindromes$$

6.6.2 Gamma Distribution (3,λ)

Distance "b/w" consecutive teiplets of palindromes and there are two palindromes in each interval of length t.

$$H_0 : Distance"b/w"teipletsofpalindromes$$

6.7 Poisson distribution

The Poisson distribution is a discrete probability distribution that distribute number of events occurred during a time period divided into fixed intervals based on a constant observed average number per interval, λ.

We can estimate λ with

$$\hat{\lambda} = \frac{1}{n}\sum X \tag{7}$$

The probability mass function is

$$P[X - k] = \frac{\lambda^k}{k!}exp^{-\lambda} \tag{8}$$

We also have the following formula for expected value and variance of a Poisson distributed sample.

$$E[X] = \lambda \tag{9}$$

$$Var[X] = \lambda \tag{10}$$

6.8 The Homogeneous Poisson Process

The homogeneous Poisson process id the model for random phenomena; to be more specifically, we call the process as. the homogeneous Poisson process if:

1. The underlying rate λ at which points, called hits, occur and is such that id doesn't change with location.
2. The number of points falling in separate regions are independent.
3. No two points can land in exactly the same place.

Let $N(t); t \geq 0$ denotes a homogeneous Poisson process, then we have

1. For any time points $t_0 = 0 < t_1 < t_2 < ... < t_n$, the process increments $N(t_1) - N(t_0), N(t_2) - N(t_1), ..., N(t_n) - N(t_{n-1})$ are independent random variables,
2. for $s \geq 0$ and $t > 0$, the random variable $N(s + t) - N(s)$ has the Poisson distribution

$$P(N(s + t) - N(s) = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{11}$$

for k = 0, 1, ...;

3. N(0) = 0

6.9 Scan Statistics

Scan statistics have been widely used to detect the potential clustering. The term Tm indicates the maximum number of events (e.g. palindromes) over m intervals.

Ho: Tm has a distribution of a maximum of independent Poisson random(χ) variables

$$P[Tm < a] = 1 - P[max\, independent\, Poisson(\chi) <= a] \tag{12}$$

$$P[Tm < a] = e^{-\hat{\chi}m}[(1 + \hat{\chi} + \frac{\hat{\chi}^2}{2} + ... + \frac{\hat{\chi}^n}{n})]^m$$

(13)

p - values = P [Tm is greater than the observed test statistic]

$$p - values = 1 - [e^{-\hat{\chi}}(1 + \hat{\chi} + \frac{\hat{\chi}^2}{2} + ... + \frac{\hat{\chi}^n}{n})]^m$$

(14)

$\hat{\chi}$ = Maximum likelihood estimator of parameter χ of Poisson Distribution

n = number computed at the data level

7 Appendix

7.1 Contribution

1. Shuyang Zhang: Wrote background and did 4.2 Locations and Spacings
2. Zetong Lai: Wrote conclusion and 4.3 analysis part.
3. Yibei Cai: Wrote introduction and did part of 4.1 Random Scatter.
4. Yuan Lin: Review the report; wrote part 4.4 The Biggest Cluster
5. Junqian Liu: Wrote Data and did part of 4.1 Random Scatter.
6. Yiteng Lu: Review the code; Finish part of 4.3 and theory.

7.2 Citation

1. Heather, James M., and Benjamin Chain. "The sequence of sequencers: The history of sequencing DNA." Genomics 107.1 (2016): 1-8.
2. Pinsky, Mark A., and Karilin Samuel. "An Introduction to Stochastic Modeling". 4th edition, ISBN: 978-0-12-381416-6.
3. Professor Bradic's ppt, Week 7 4-6.