Lab 2 - Palindromes in the Cytomegalovirus DNA

Stat 135

Nicholas Lai

March 18, 2018

Introduction

Palindromes in the Herpes family of viruses are of interest to biologists looking to fight disease, as their locations mark potential replication sites. These sites are a vital object of study, as the prevention of viral reproduction represents a promising avenue to curing conditions caused by the spread of these viruses.

Searching for these sites without any information about where they might be located would be time consuming and expensive. Statistical analysis of the distribution of palindrome sites may yield clues to narrow the search. Biologists conjecture that sites of replication in the CMV are characterized by an unnaturally dense cluster of palindromes, so finding this cluster is the problem that this memo will presently address.

Methodology

Definitions

Define a palindrome to be a strand of base pairs of at least length 10 such that one side's complementary strand is the strand in reverse and vice versa. Any shorter strands are disregarded, as they arise through chance too often.

The CMV DNA strand is 229354 letters long.

Data

The data we have are the locations of palindromes as defined previously in the CMV DNA. There are 296 Palindromes in total in the data, indexed by their starting locations

Theory

The pressing question of this project is this: does knowledge of the distribution of the location of the palindromes give us any information about where the replication site may be?

In nature, many rare processes can be modeled by the Poisson distribution (The Law of Rare Events). The appearance of palindromes in the CMV DNA is rare, with only 296 as defined being present in 229354 locations, so this is a natural model for their distribution.

The Poisson distribution takes one parameter, the rate of the process. This rate is unknown to us, but we can observe that the maximum likelyhood estimator and method of moments estimator of a poisson distribution are both the sample average of the process. That is: $\hat{\lambda} = 296/229354$

If the apperance of palindromes is a poisson process, then we would have few leads as to where to search along the stand for abnormal clustering, as the clustering could be reasonably be explained by natural variation.

We want this model to be insufficient, because if it is, we could look for places where more clustering is present than the model would predict and flag them as potential replication sites.

If palindromes arrive in a poisson process, several things should be true about our data. If any of them are false, it is an indication that unnatural clustering may be occuring.

The Gamma Distribution

If the palindromes arrive along the DNA by a poisson process, then the waiting times after a hit until the *nth* subsequent hit is distributed according to the $Gamma(n, \lambda)$ distribution. If they are not, it could be some evidence of unnatural clustering.

The Uniform Distribution

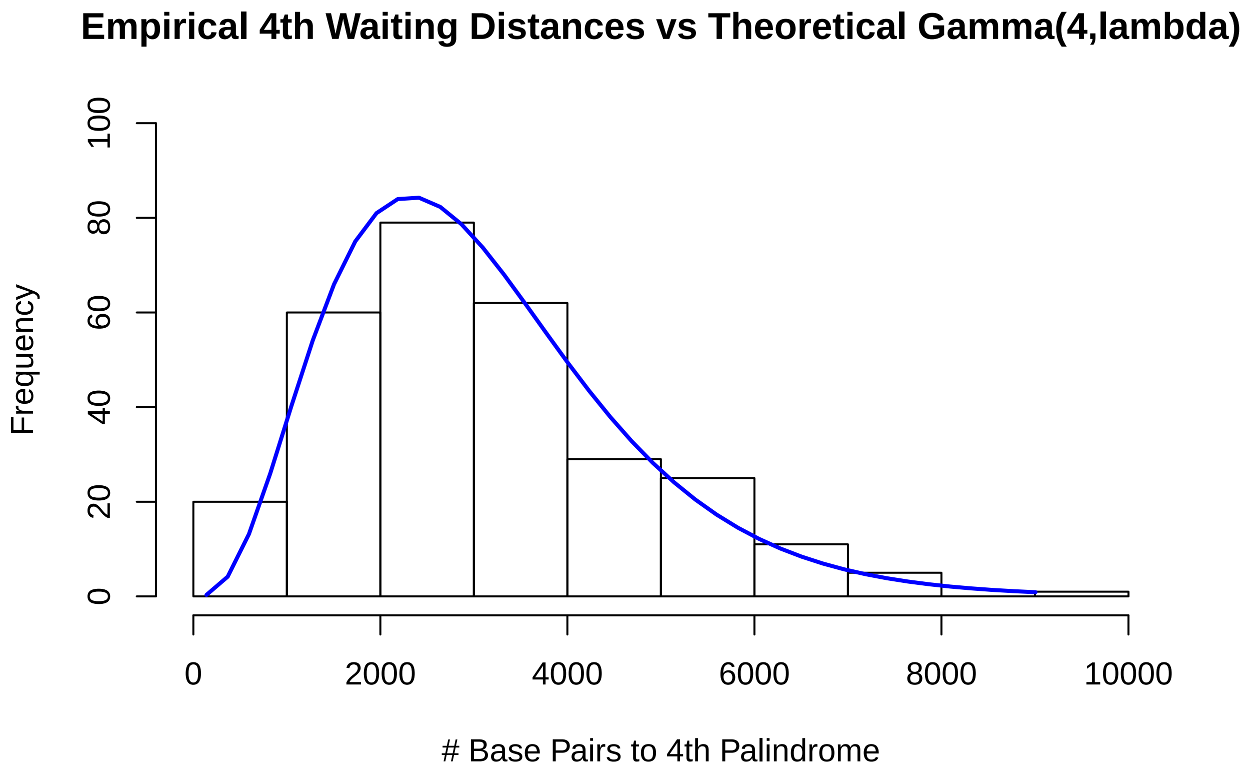
If the palindromes arrive along the DNA by a poisson process, then if we split the strand of DNA into equal lengths, the distribution of palindromes is uniform amoung those lengths. If they are not, it is evidence of unusual clustering.

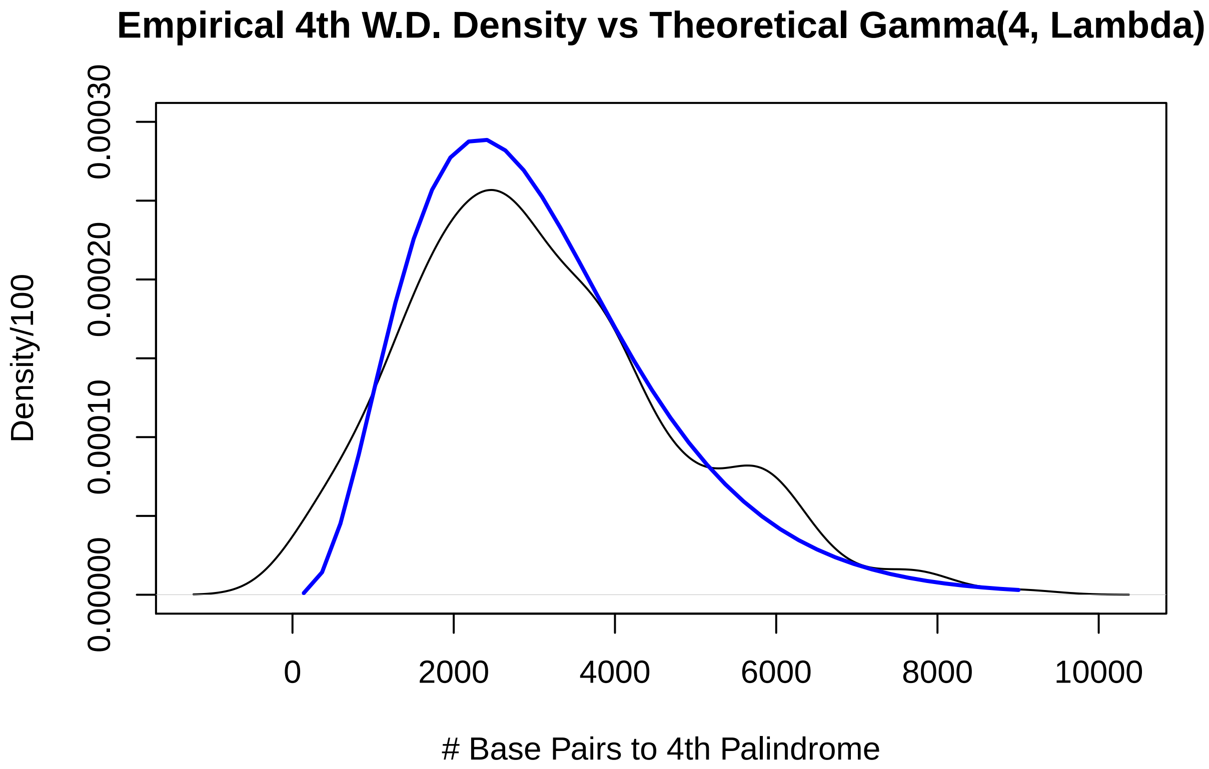
The Poisson Distribution

If the palindromes arrive along the DNA by a poisson process, then if we split the DNA into equal lengths, the distribution of the number of hits in those lengths should be $Poisson(\lambda t)$, where t is the length of one interval. If this is not true, then it could be evidence of unusual clustering.

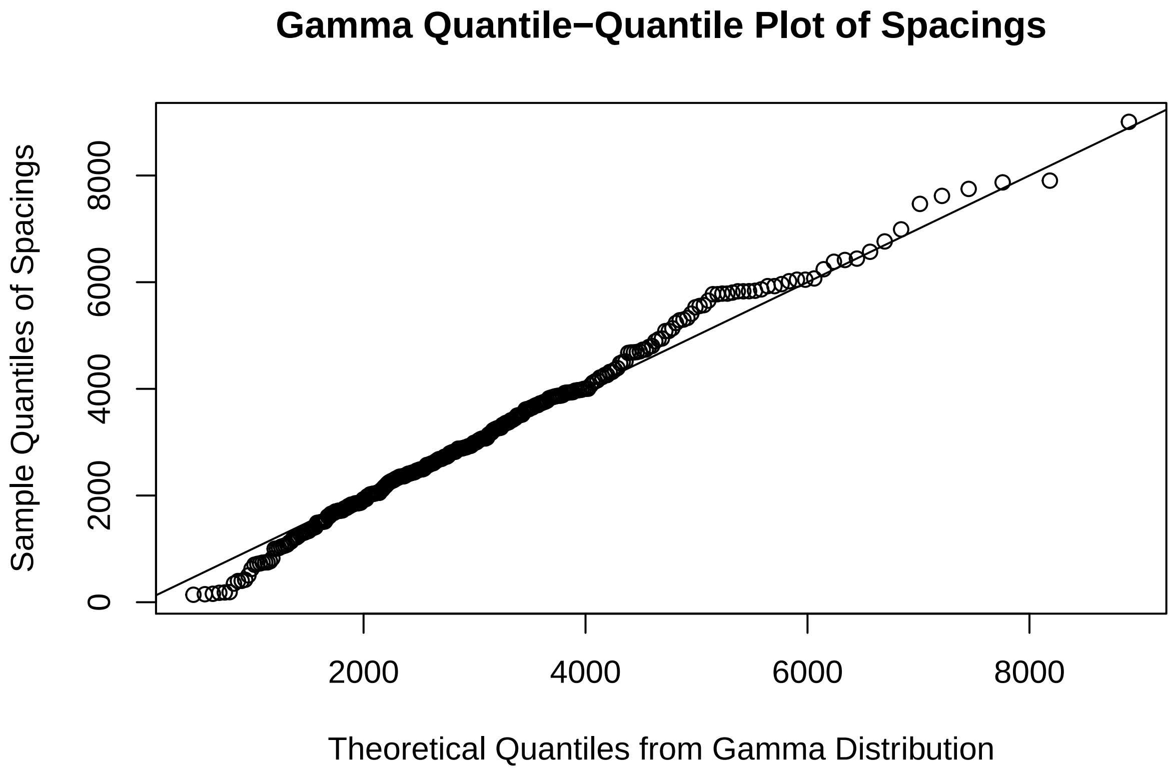
Results

The Gamma Distribution





As we can observe by imposing the $Gamma(4, \lambda)$ distribution on the observed data of 4th occurrence waiting times, the fit is far from perfect. This is some evidence that there is an unusual cluster of palindromes in the data. We can further confirm this with a quantile plot:



We observe from the quantile plot that the distribution has longer tails than normal. This means that there exist more short spacings between palindromes than we would normally expect, which is good evidence that an unnatural cluster exists.

The Uniform Distribution

Here, we split the DNA strand into 12 segments of length 19112.

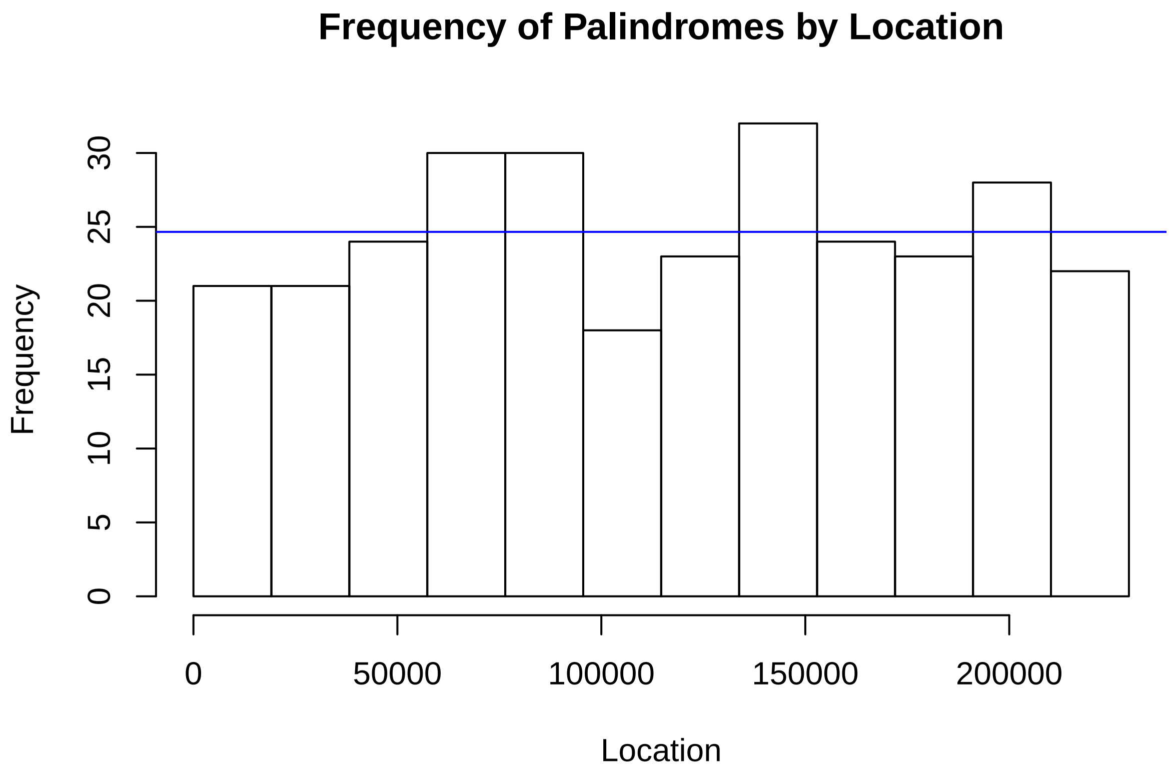
segment	count	expected	ch_sq_term
1	21	24.66667	0.5450450
2	21	24.66667	0.5450450
3	24	24.66667	0.0180180
4	30	24.66667	1.1531532
5	30	24.66667	1.1531532
6	18	24.66667	1.8018018
7	23	24.66667	0.1126126
8	32	24.66667	2.1801802
9	24	24.66667	0.0180180
10	23	24.66667	0.1126126
11	28	24.66667	0.4504505
12	22	24.66667	0.2882883

Observe that segment 8 is the largest contributor to the Chi-Squared Test Statistic, so there is some evidence that base pairs 133784-152896 may contain the replication site.

The Chi-Squared Goodness of Fit Test statistic:

```
## [1] 8.378378
```

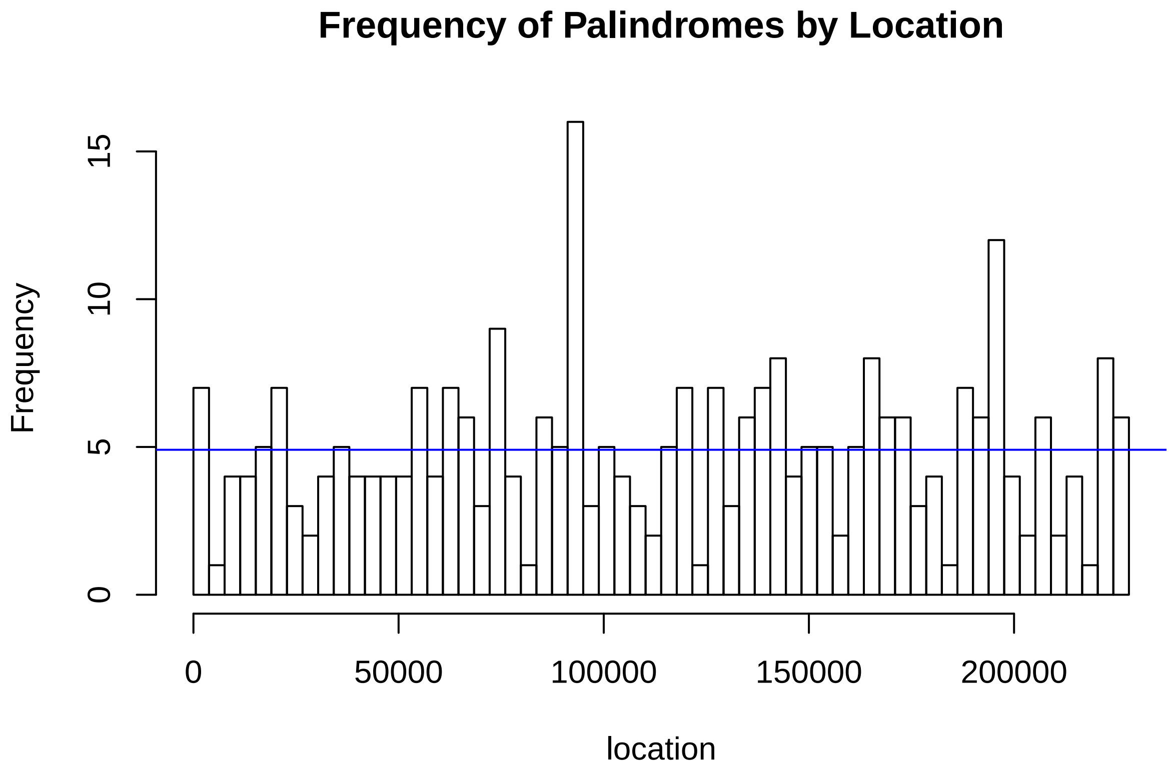
The Degree of Freedom for this test is $12 - 1 - 1 = 10$, and for a significance level of 5%, the Chi-Square statistic corresponding to α is 18.31. Therefore, we fail to reject the null hypothesis that the distribution is uniform.



Observe that segments 4,5,8, and 11 have an above average number of palindromes and can be considered potential replication sites as well.

The Poisson Distribution

We split our data now into 60 segments of length 3800.



The clustering of palindromes is much more apparent with shorter chosen segments. Segments 25, 52 have many more palindromes than average, so they are possible replication sites.

ct_Palindromes	Freq	expected
1	5	2.1819723
2	5	5.3504244
3	6	8.7465336
4	14	10.7237076
5	8	10.5182611
6	8	8.5972922
7	8	6.0232742
8	3	3.6924246
9	1	2.0120470
12	1	0.9867499
16	1	0.4399298

The probability of the maximum palindrome frequency being 16 or higher given that the data are distributed $Poisson(3800\lambda)$:

```
## [1] 0.003313349
```

This is very strong evidence that segment 25, base pairs 91200-95000, is a possible replication site.

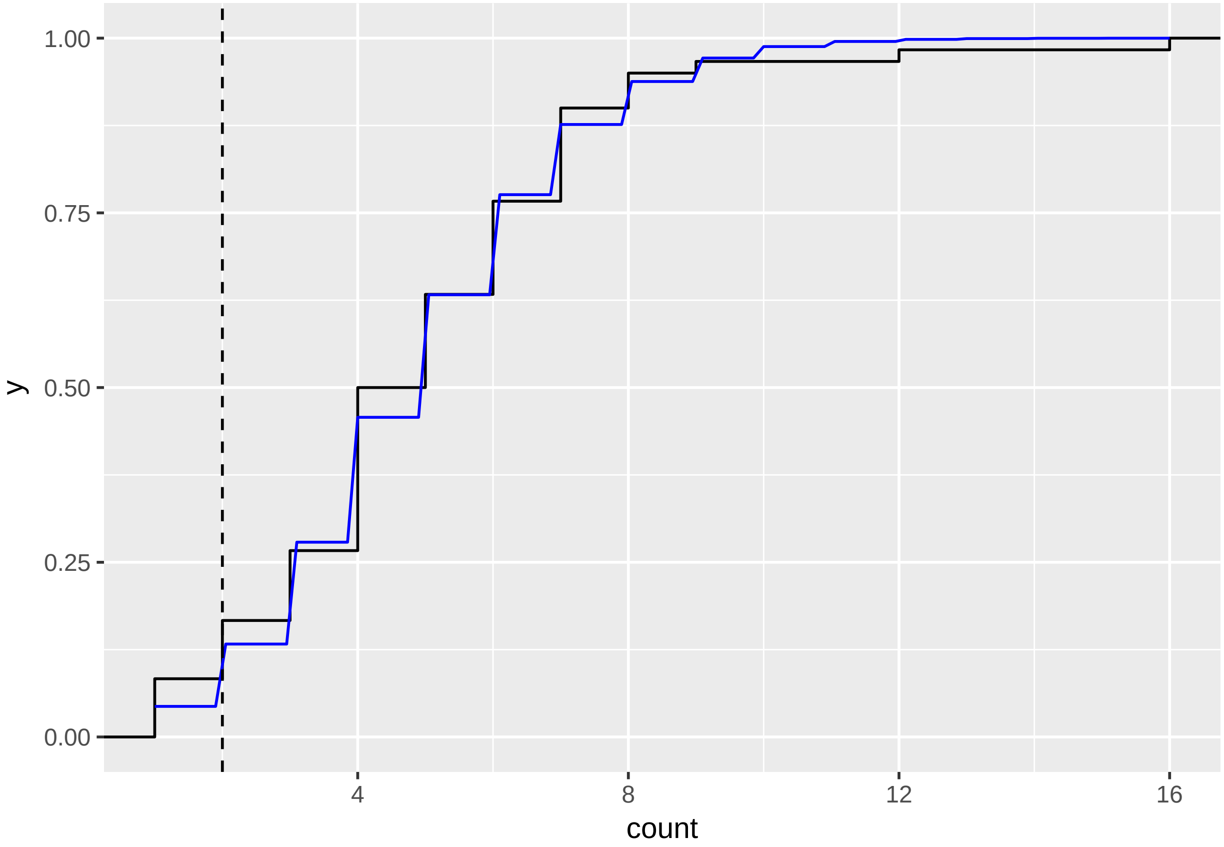
The probability of the maximum palindrome frequency being 12 or higher given that the data are distributed $Poisson(3800\lambda)$:

```
## [1] 0.2465795
```

This is fairly weak evidence that segment 52 is a possible replication site.

To see how well the Poisson distribution in general fits this data, we perform a KS Test:

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  pois_data$count
## D = 0.19079, p-value = 0.02535
## alternative hypothesis: two-sided
```



At the 5% significance level, we reject the null hypothesis that the data is distributed $Poisson(3800\lambda)$. This is good news, as it means that it is likely that there are unnatural clusters of palindromes in the DNA.

Discussion

Based on our analysis of the distribution of the palindromes, we can conclude that the Poisson Process does not fully explain the data. This means that an analysis of the distribution of palindromes is likely to yield candidates for the replication site.

The shape of the distribution of spacings before the 4th palindrome, which we expected to be $Gamma(4, \lambda)$ was not perfectly explained by the distribution. This, along with the tiny likelyhood of getting an interval of 3800 base pairs with 16 palindromes and the low p-value on our KS test suggests that there is unnatural clustering in palindromes.

However, when our intervals were larger, the distribution of palindromes were adequately explained by the Uniform distribution (We failed to reject that hypothesis). Therefore, there is some evidence that the data could have arisen from a poisson distribution, but the other evidence is compelling towards the conclusion that this is not the case.

Conclusion

Based on our 60 interval Poisson Distribution test, our best candidates for conducting a search for replication sites are base pairs 91200-95000 and base pairs 193800-197600, with very good evidence for the former.

If these two sites do not yield the replication site, there is weak evidence that base pairs 133784-152896 contain the replication site.