MEMORANDUM

To: The Head of The Biology Lab

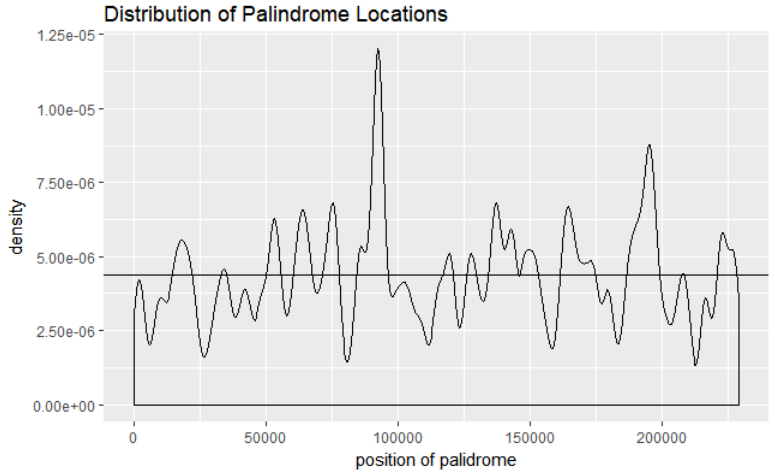
From: Graeson Gardner

Date: March 20, 2018

1. INTRODUCTION

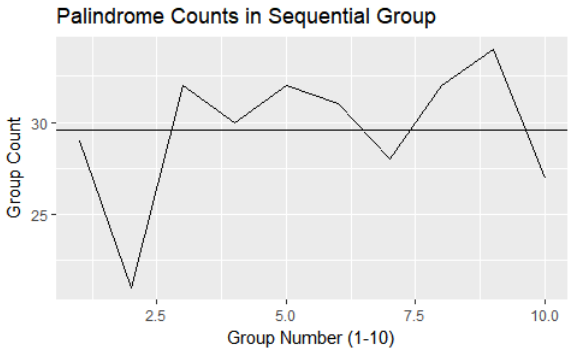
The majority of my inquiry followed the analysis done in texts closely, utilizing methods I learned mostly from the Nolan and Speed. In doing so I also tried to answer a similar question: “are there any anomalous clusters that could be the replication site for CMV.” It was pointed out in the text itself that Weston had earlier discovered a large cluster at between 195-200k base pairs. Through the process of the lab, I identified this as well as one other likely location for the replication site at around 95,000 base pairs. The tool that, more than anything, helped me identify this location was a simpler kernel plot.

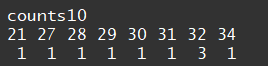
1. METHODOLOGY

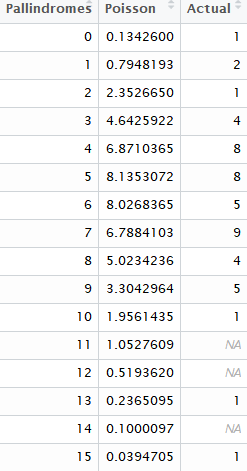
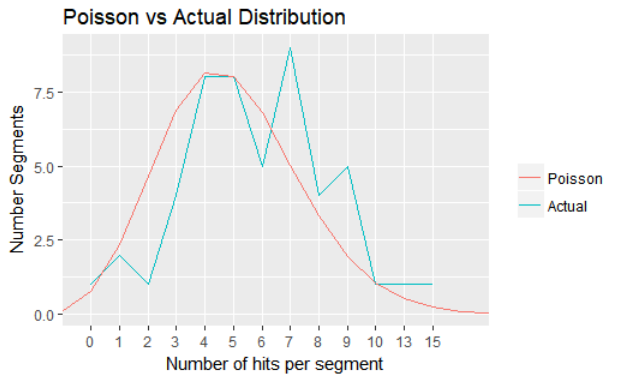
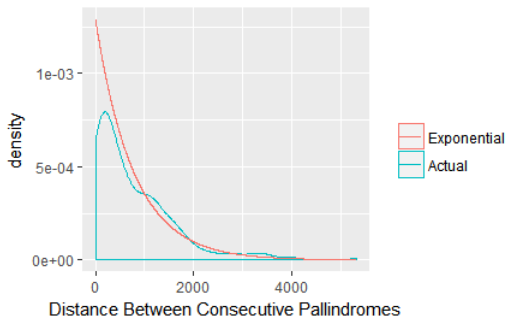
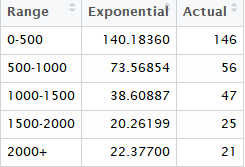
I began my investigation with some basic data graphs and analysis comparing the overall distribution to the uniform, visually using the built-in kernel density estimator for ggplot. This immediately showed a few areas where there appeared to be large concentrations of palindromes. Using thinner kernels allowed me to identify spikes more distinctly. Note the two largest peaks mentioned in the introduction. Visually, the Kernel density estimator gives us around 10 distinct peaks. One extra thing to mention: as I increased the kernel size, the distribution seemed to “become uniform”, though there was a noticeably higher density mass at the center. To test this, I wanted to see if when separated into 10 bins, the distribution was close to uniform: on a macro scale, are the peaks relevant? On a large scale, we expect the distribution of counts per bin to be more or less uniform, then become more “obviously Poisson” as bins are shrunk. This test was almost identical to the one described in the text, however I also did the recommended chi-squared test and a plot to help with my comparison to the uniform distribution.

To compare the distribution to its “expected Poisson”, I started by estimating the lambda parameter to be the total palindromes divided by bin number to be 5.92 (\*mean arrivals are the ml estimate for Poisson). I then separated the data into 50 intervals of equal length. 50 intervals were chosen because the 229354-long genome only loses 4 base pairs this division is made, it was an easy number to work with. Another Chi-squared test was performed to compare the distributions, using the bins: 2 or less, 3, 4, 5, 6, 7, 8, 9, 10 or more. I also looked at the probability distribution of the expected maximum value in this set of 50 intervals (again, this procedure being from the stats lab).

I did a similar comparison between the set of first differences in location and the exponential distribution, estimating lambda using the multiplicative inverse of the mean. I did not have high hopes for this to return anything too interesting, however.

1. RESULTS

The chi-squared test initially used to compare the distribution of the 10 segments to the uniform resulted in a test statistic value 4.13. Given the test-stat’s distribution had 8 degrees of freedom (10 bins – 1 parameter -1) we cannot reject at 50% certainty (critical value: 7.3), meaning this is a “good” fit for the data. Similarly, the graph looks “somewhat” uniform, but distinctly not so. The peak at around 200k is visible here, but not the other peak is not. From the distributions of counts, it is fairly clear that this could not be Poisson; there is a huge right peak.

The chi-square test described for the Poisson distribution resulted in a test-stat of 3.59 which again is low (critical value for rejecting the null at 50% that the data does not follow the distribution is 6.3 given 7 degrees of freedom). This by itself did not yield a lot of interesting information, but the testing the distribution for the max value did. Notice that there is one group with 15 palindromes in it. I checked the sequential location of this data and as I had expected, it corresponded to the area around 95k base pairs. Using the formula 1-P(X<max)^m from our text to estimate the probability of getting 1 group of this size (where m = 50, p = ppoiss(14, lambda = 5.92). This value was only .059. This means, we expect 94% percent of max values to be less (than 15) given 50 trials and the same lambda.

The chi-squared test done to compare the exponential distribution to the actual distribution of the times between subsequent hits resulted in a statistic of 7.4 with three degrees of freedom (see table). The critical value to reject at 95% percent confidence is 7.8, so this is a surprisingly bad fit (contrary to the appearance of the graph). The biggest discrepancy seems to be from 500-1500, where at 500-1000 our values are above our estimate, they were noticeably smaller from 1000-1500, leading to large t value.

1. CONCLUSION

Through the process of this investigation, I was able to pinpoint a “new” possible location for the CMV virus’s replication site (besides the location Weston had discovered) occurring at around 95000 base pairs. This was supported graphically by the distinctly greater densities found at these two areas and by estimating the probability of having an outlier as larger as we did in our sample, which was quite small (6%). Though having less apparent value to immediate inquiry, comparisons between the palindromes distribution and the Poisson (as well as its “corresponding” exponential) gave us (or me at least) better insight into how palindromes were distributed and confirmed the fact that they were more or less Poisson. The result that they were Poisson itself was quite important as it allowed us to make the assertion that there was indeed an outlier through comparison.