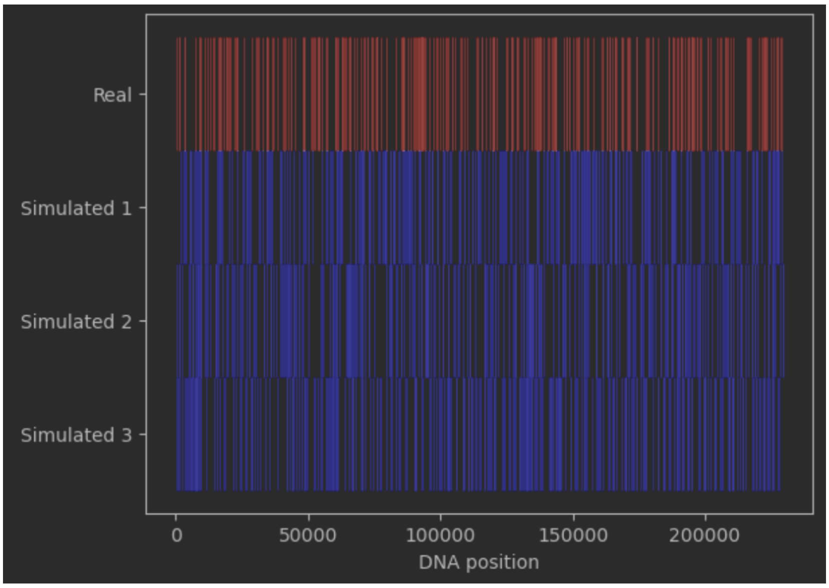
Abdulayev Damir BSDS21-02 @speedfiref
Optional assignment on Statistics

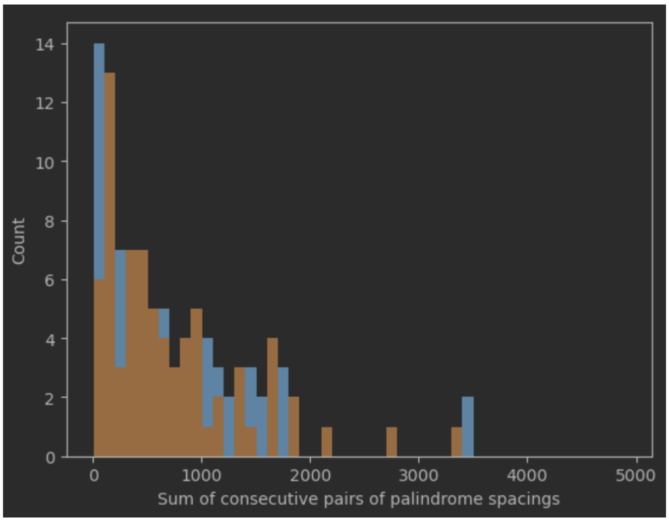
Random scatter results:

Based on the scatter plot we can see that real palindromes are not randomly scattered, because they are not evenly distributed across the DNA sequence. Although we can see some gaps, they are not even. For example, we can see that there are more palindromes in the first 100000 positions than in the last 100000 positions.



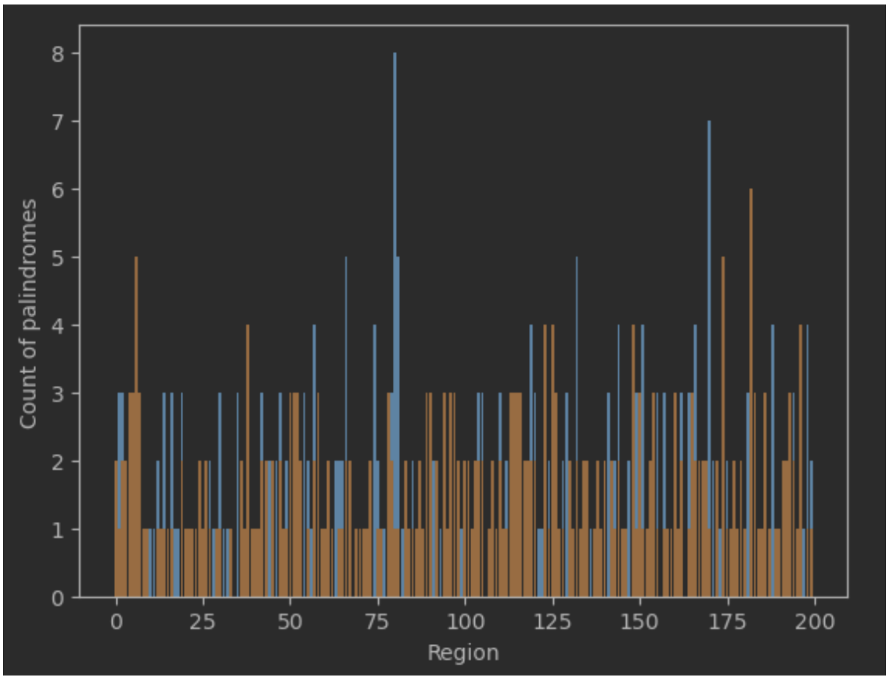
Location and spacings results:

Based on the histogram we can see that spacing in CMV DNA is different with the random scatter. While random scatter seems to be more uniform, the real scatter is more concentrated around 1000. This means that the real scatter is more likely to have palindromes that are close to each other.



Results of counts:

I divided regions into 50-550 regions with step 50 and 100-1100 regions with step 100. Based on graphs and results I can say that the best results are with 200-300 regions, so I take a graph with 200 regions and perform a chi-square test. Null hypothesis is that the distribution of palindromes is uniform. I used the scipy.stats.chisquare function to perform a chi-square test. The result is 0.0312075, so we can reject the null hypothesis and say that the distribution of palindromes is not uniform (with 0.05 significance level).



Conclusion:

That means that the Anomalous cluster of palindrome locations is not random, but it is caused by some biological process. The most probable explanation is that the palindromes are caused by the replication of DNA. So my advice to the head biologist would be to check if the palindromes are caused by replication of DNA in the region 91000-95000 and then check other anomalous clusters.