CASE STUDY 3: SEARCH FOR THE UNUSUAL CLUSTER IN THE PALINDROMES

Question

How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?

Setup

```
locations <- read.table('hcmv-25kgjn1-1rfrtkc.txt', header=TRUE)$location  # Original
health <- read.csv('RAW_DATA-2iwcznn-2kr2xw0.csv', header = TRUE)  # Additional

N <- 229354  # Base pairs
n <- 296  # Palindromes
```

Scenario 1: Random Scatter

To begin, pursue the point of view that structure in the data is indicated by departures from a uniform scatter of palindromes across the DNA.

Of course, a random uniform scatter does that mean that palindromes will be equally spaced as milestones on a freeway. There will be some gaps on the DNA where no palindromes occur, and there will be some clumping together of palindromes.

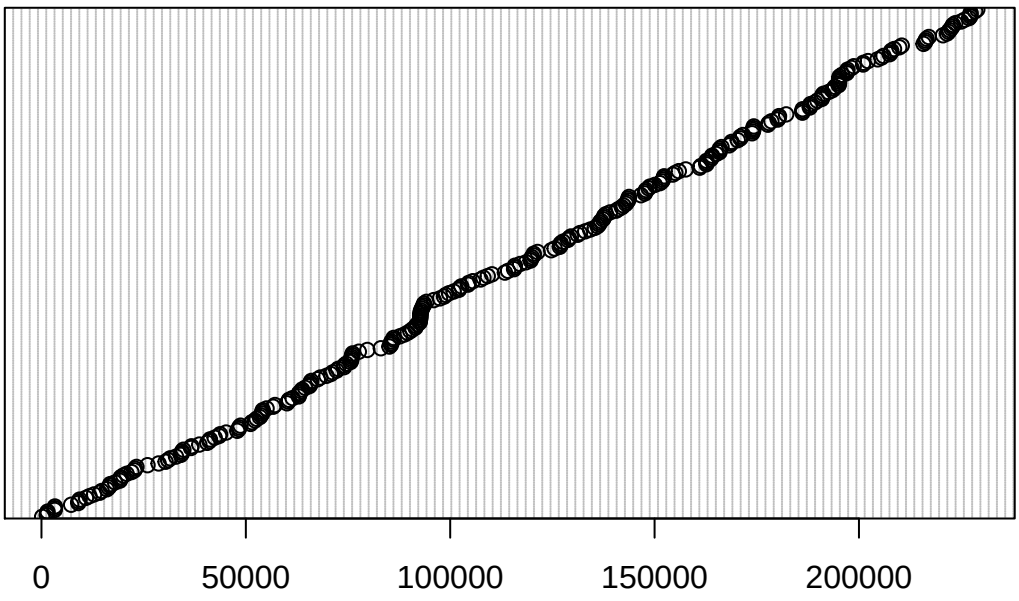
To look for structure examine the locations of the palindromes, the spacing between palindromes, and the counts of palindromes in non overlapping regions of the DNA. One starting placing might be to see first how random scatter looks by using a computer to simulate it.

A computer can simulate 296 palindrome sites chosen at random along a DNA sequence of 229,354 bases using a pseudo random number generator. When this is done several times, by making seller sets of simualted palindrome locations, then the real data can be compared to the simulated data.

```
set.seed(0)
uniform <- sample.int(N, size=n)
?uniform
```

```
## No documentation for 'uniform' in specified packages and libraries:
## you could try '??uniform'
```

```
dotchart(locations)
```




```
# hist(locations, breaks=20, main='Locations of Palindromes', xlab='Base Pair')
```