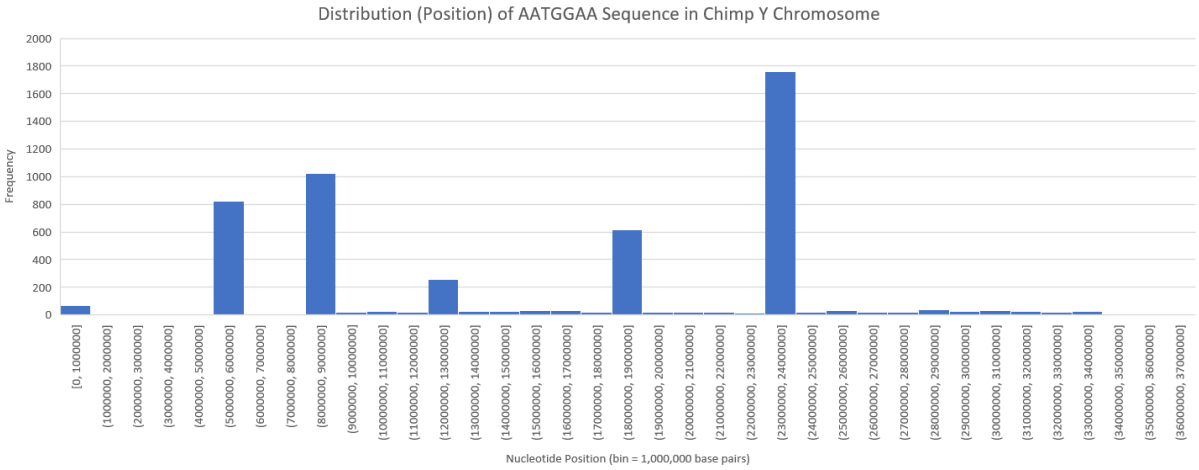
Distribution or Position (Distribution) of 7-letter Words

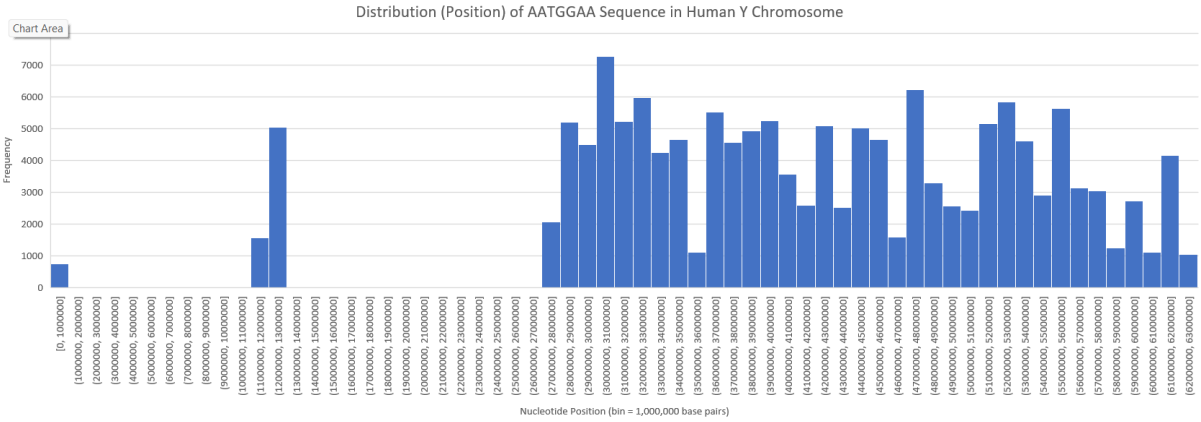
By Craig Paardekooper

Method

In this study, I looked at all of the seven letter words in human chromosome Y, and chose the one occurring with highest frequency. Then I obtained the position of each instance of this word by using the software in Appendix 1. The positions were then plotted in a histogram, to see the shape of the distribution. I carried out the same procedure for chimp chromosome Y

Results





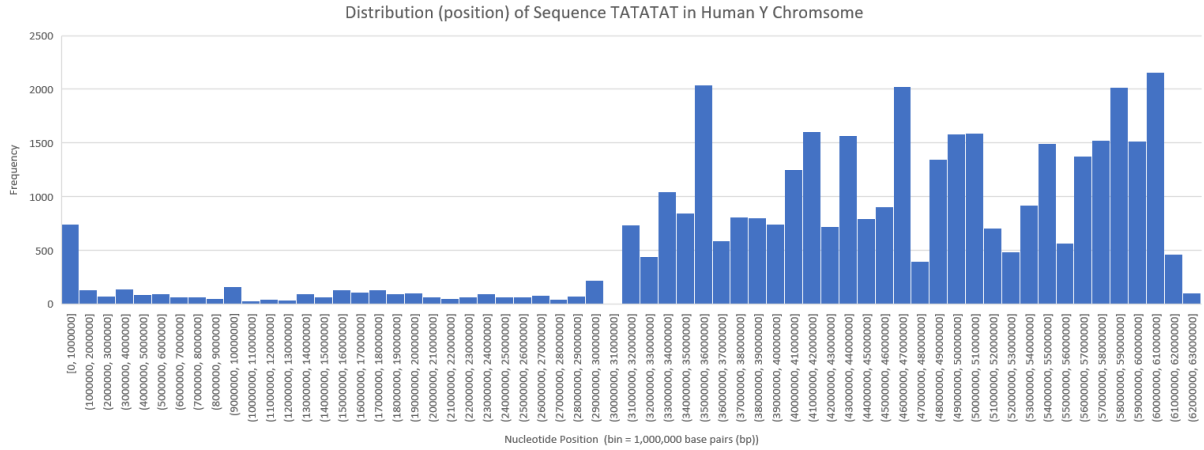
Observations

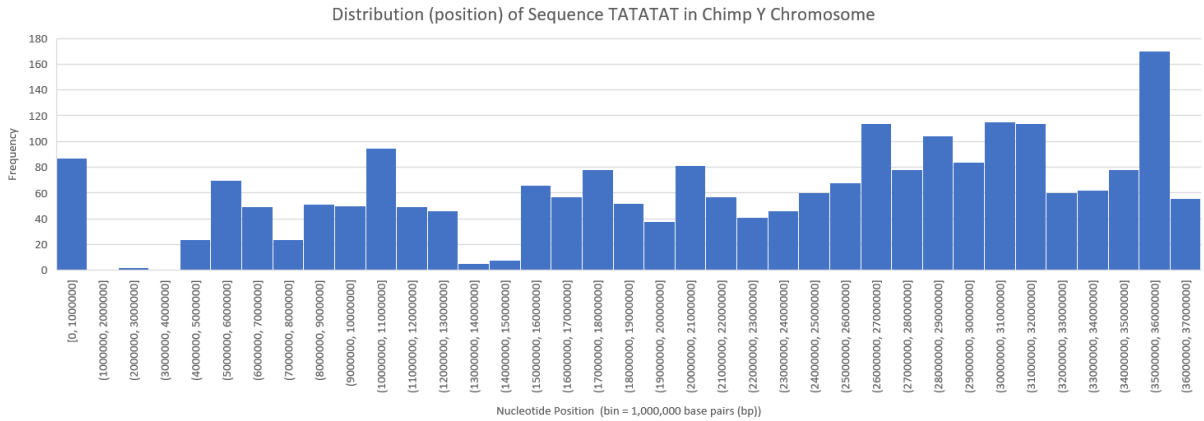
The most frequent 7 letter word in human Y chromosome is AATGGAA. It occurs 147,540 times, and is highly concentrated in all of the latter half of the chromosome, in all the bins from 28 million bp to 62 million bp, as the chart above shows.

The same 7 letter word in the chimp Y chromosome occurs 5096 times, and is distributed mainly in 5-6, 8-9, 12-13, 18-19, and 24-24 bins.

So, not only is the absolute frequency of this word different by a factor of 30 times, the distribution of the word is completely different.

Next, I looked at another 7-letter word – TATATAT. It occurs 38,564 times in the human Y chromosome, and 2239 times in the chimp Y chromosome. Here are the distributions –



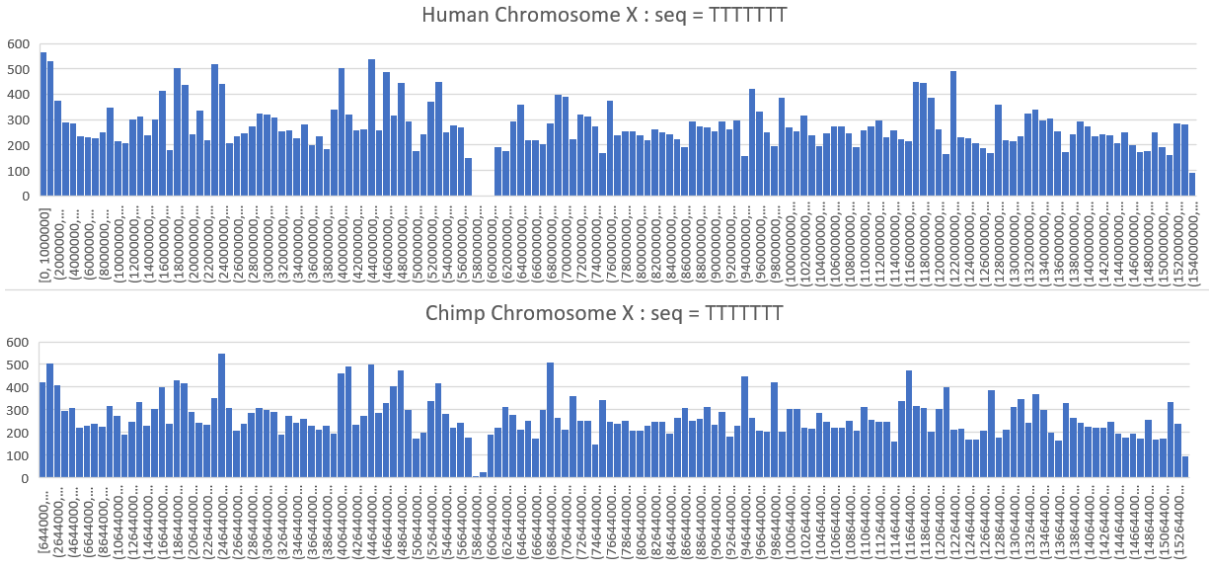


Once again there is no correlation between the two distributions.

The TATA sequence is a DNA sequence that **indicates where a genetic sequence can be read and decoded**. It is a type of promoter sequence, which specifies to other molecules where transcription begins. Transcription is a process that produces an RNA molecule from a DNA sequence.

This means that in the human Y chromosome, transcription promoter sequences are concentrated in the later half of the chromosome. In comparison, in chimp Y chromosomes transcription sequences are more evenly distributed through out the chromosome.

Next, I compared the human X chromosome with the chimp X chromosome, for the positions of the sequence TTTTTTT, which is the most frequent sequence in both chromosomes. Here are the results –



There is a very strong correlation between the positions of TTTTTTT sequences in human compared to chimp chromosome X. The correlation is 0.9999.

Visually, you can see a close fit.