Jumps Analysis in Target Sequences

Qian Feng

2018/7/13

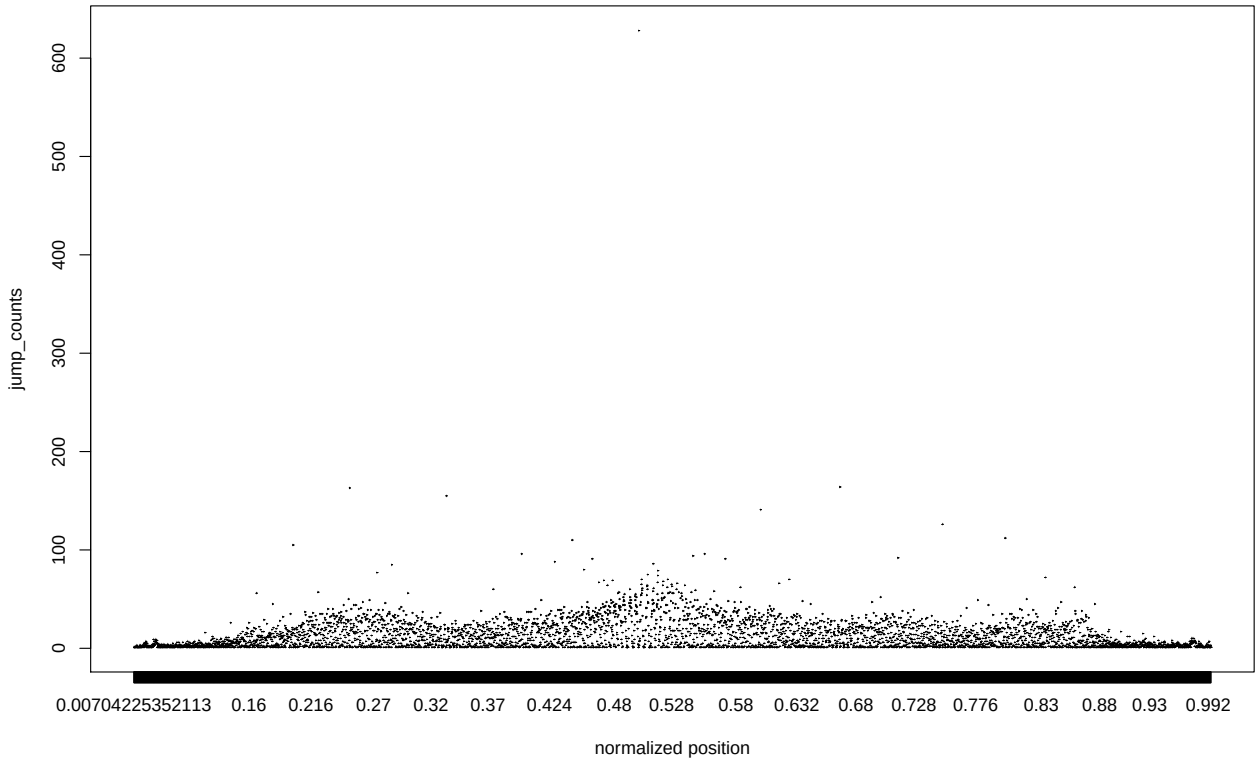
Load R libraries

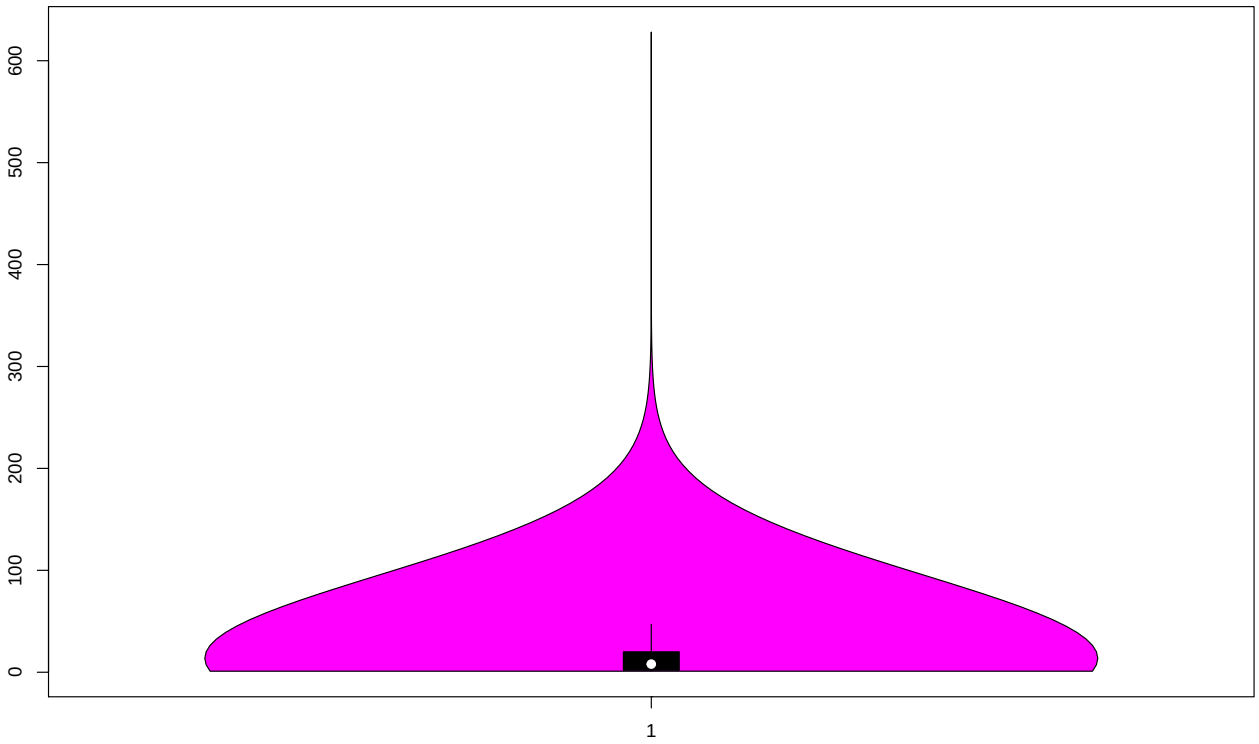
Investigate the recombination hotspots using the relative position

Let's try the global data from Gerry's paper. 31946 sequences are involved.

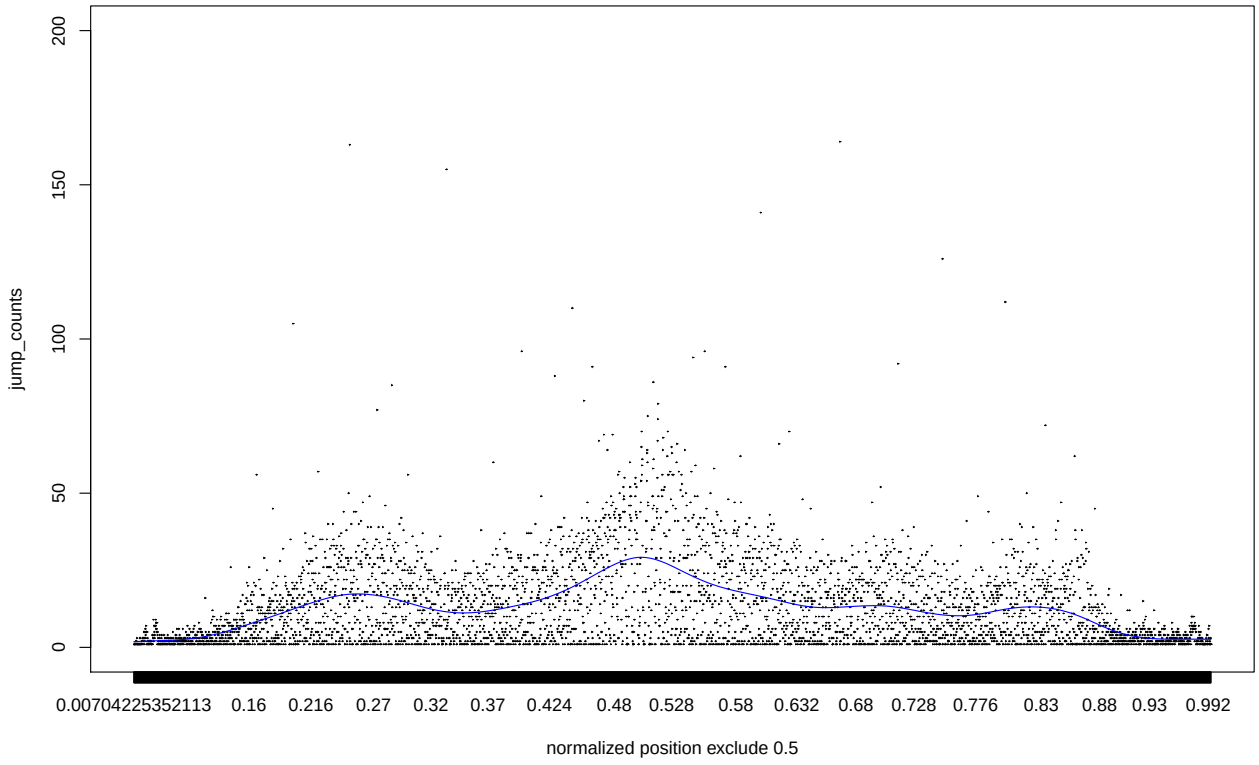
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     2.0     8.0    12.9    20.0   628.0
```

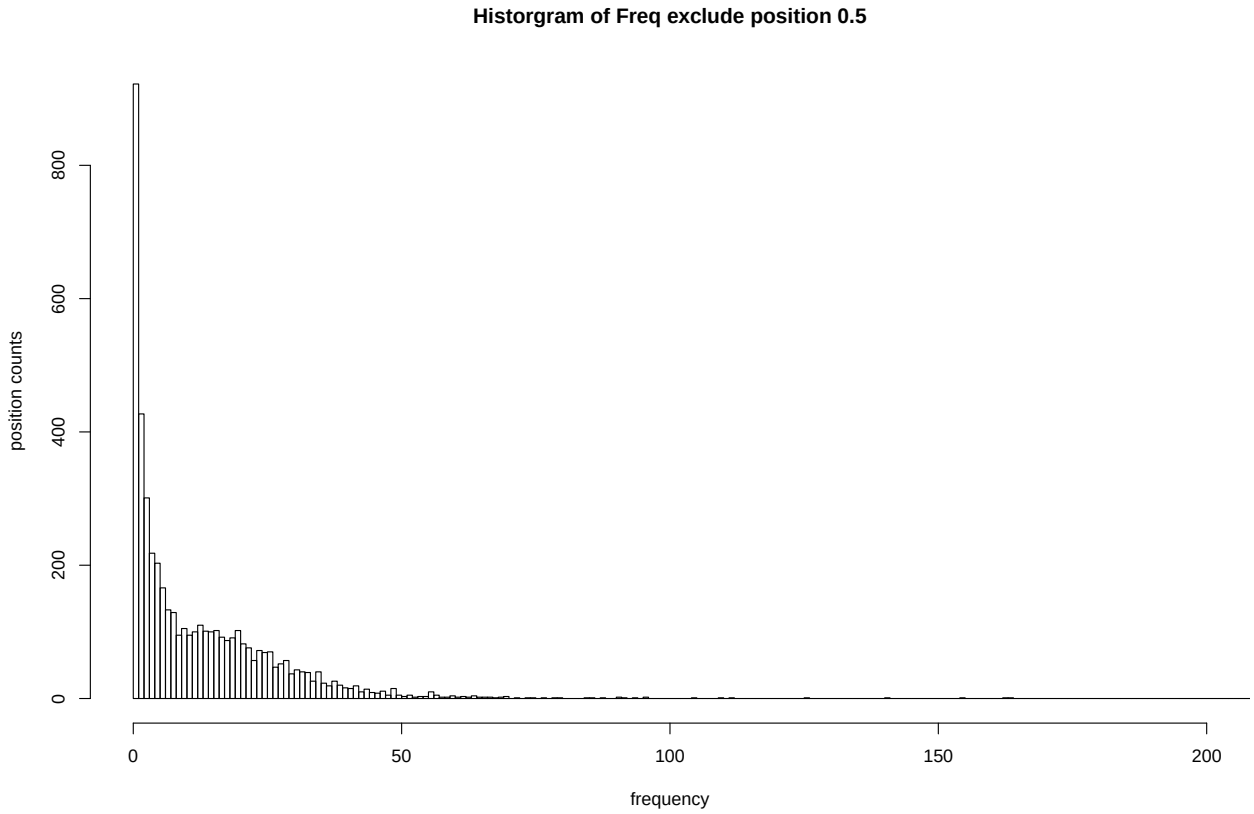
```
##      recombination Freq
## 2244           0.5  628
```





Let's take a close look at the distribution of jump frequencies, and on the contrary, look at the conserved jump counts.

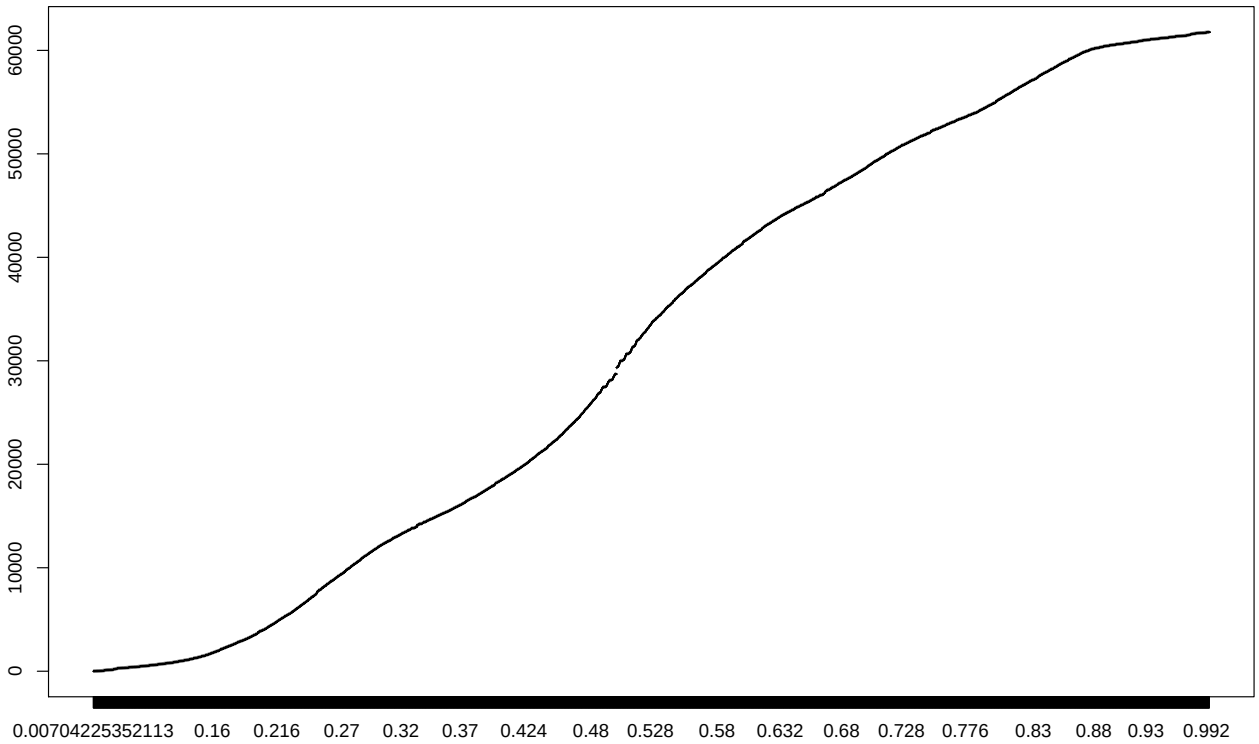




This indicates that two ends of DBLa tags are the most conserved regions. At the middle region of tags jump happens more frequently. We have 31946 sequences, when the jump counts are compared to this value, it's still pretty small. Most positions in all sequences only jump once, most positions have less than 50 jumps,its maximum number should be the number of all sequences,namely 31946.

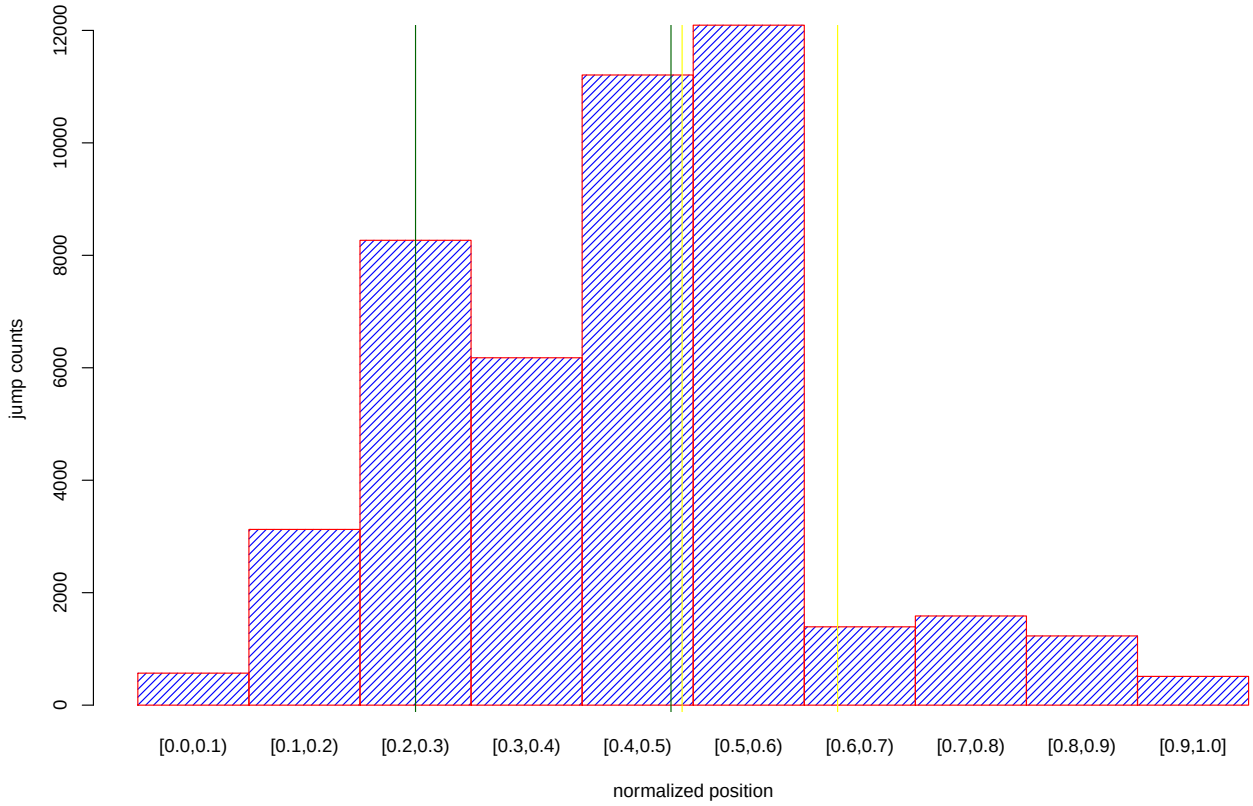
This is coded by Yao-ban, a cumulative jump frequency plot is obtained.

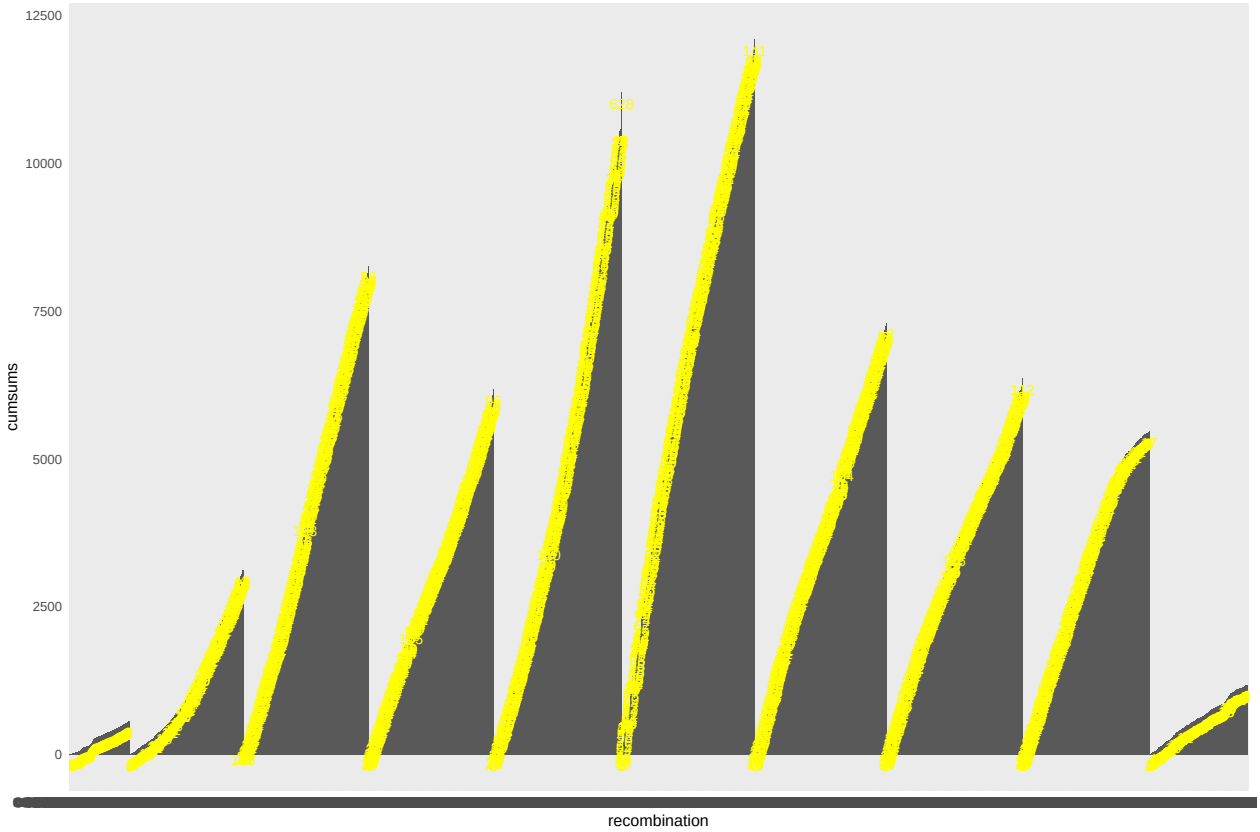
```
rec2 <- recombination
rec2$sumFreq <- rec2$Freq
for (i in 2:4787) {
    rec2$sumFreq[i] <- rec2$sumFreq[i] + rec2$sumFreq[i - 1]
}
plot(rec2$recombination, rec2$sumFreq, type = "l")
```



Then let's look at its barplot.

```
barplot(rec_cumsum[temp$sumFreq, 4], names.arg = temp[, 1], density = 20, border = "red",
    col = "blue", xlab = "normalized position", ylab = "jump counts", space = 0)
abline(v = c(4.9, 6.3), col = "yellow")
abline(v = c(2.5, 4.8), col = "darkgreen")
```





From the first bar plot, the position between [0.5,0.6) has the highest jump counts,followed by position range from 0.4 to 0.5. There is one decrease about the jump count in position at [0.3,0.4). Still, at the two endpoints

of target sequences, they are silent regions which seledom occur jumps. Moreover, based on second plot, each block is cumulated, and the increasing trend is very interesting in each interval.[0.1,0.2) and [0.4,0.5) increase much more dramatically than other intervals. Two ends of targets sequences increase slowly.

Dr **Qixin He** told me more conserved region has a higher probability to occur recombination and share higher similarity sequence block. HB5 should be in range at position from [0.4,0.6),its related amino acid sequence starts from "redww". The position [0.2,0.3) is related with HB14.

Then let's look at the aligned protein sequences and unaligned sequence by AliView (seaview is also Ok). It seems "redww" is in the middle of sequences, HB2 is at the rightmost of sequences.The leftmost sequences are pretty conserved, but am not sure it's HB3. so I should look at the HB14. Moreover, I calculate the location of intervals for HB5 and HB14 in sequences.

The references for extracting HB5 and HB14 are below:

---

1. Rask, Thomas S., et al. "Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes, divide and conquer." PLoS computational biology 6.9 (2010): e1000933.
2. Larremore, Daniel B., et al. "Ape parasite origins of human malaria virulence genes." Nature communications 6 (2015): 8368.

```
## [1] "HB5 left endpoint location:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3039  0.4737  0.4919  0.4913  0.5083  0.8111
```

```
## [1] "HB5 right endpoint location:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4706  0.6050  0.6279  0.6270  0.6471  1.0000
```

Therefore, HB5 is in the location between 49.1% to 62.7%, combined with previous barplot, this region has the highest number of jump counts.

```
## [1] "HB14 left endpoint location:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04425 0.20800 0.22308 0.24497 0.24460 0.64444
```

```
## [1] "HB14 right endpoint location:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3415  0.4667  0.4851  0.4851  0.5036  0.8000
```

In general, from 24.5% to 48.5% of sequence length, HB14 exists there, they are also in the region of high jump probabilities. Then region from 48.5% to 49.1% should be the highly variable region.