Genomes in BSstring packages - Drosophila melanogaster

BSgenome is Biostring based package with whole genom of an organism. To check all available BSgenomes use:

available.genomes()

Drosophila melanogaster reference genome structure

BSgenome.Dmelanogaster.UCSC.dm6 contains 1870 Biostrings with sequences.

```
library(BSgenome.Dmelanogaster.UCSC.dm6)
droso <- BSgenome.Dmelanogaster.UCSC.dm6
```

Sequences have different lengh, with median 1576

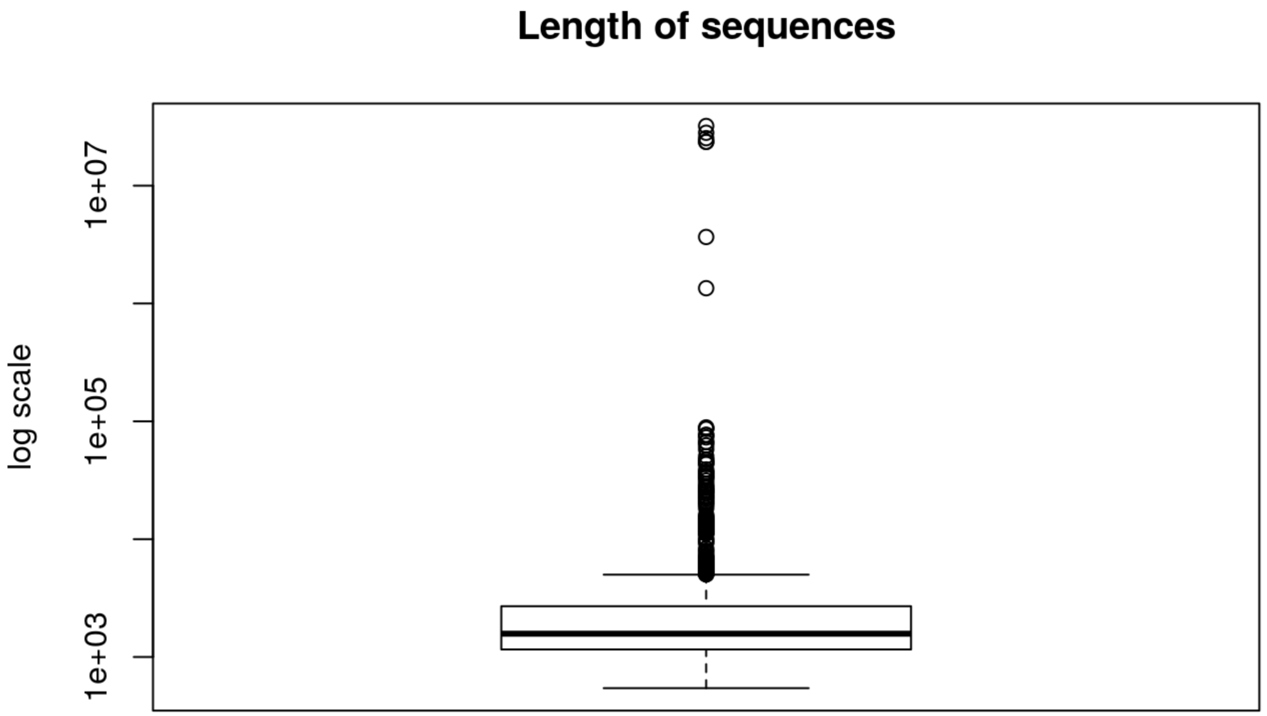
```
head(seqlengths(droso), n=8)
```

```
##    chr2L    chr2R    chr3L    chr3R     chr4     chrX     chrY     chrM
## 23513712 25286936 28110227 32079331  1348131 23542271  3667352    19524
```

```
median(seqlengths(droso))
```

```
## [1] 1576
```

After visualization is visible that first 7 of them has significant longer seqence.Logarythmic scale makes plot more clear.



Let’s look into first chromosome:

```
droso$chr2R
```

```
##   25286936-letter "DNAString" instance
## seq: CTCAAGATACCTTCTACAGATTATTTAAAGCTAGTG...ACTTTGCTGGTGGAGGTACGGAAACAGAATGAATTC
```

BSgenomes contain only forward strands, not reverse (without any exceptions).

To check what is the % of AT or GC in chr2L there is a fucntion:

```
## percentage of AT:   0.582175710921355
```

```
## percentage of GC:   0.417815783403318
```

Lets check how many “TTAGG” pattern is in Drosophila genome? This is typical for telomers in some invertebrates.

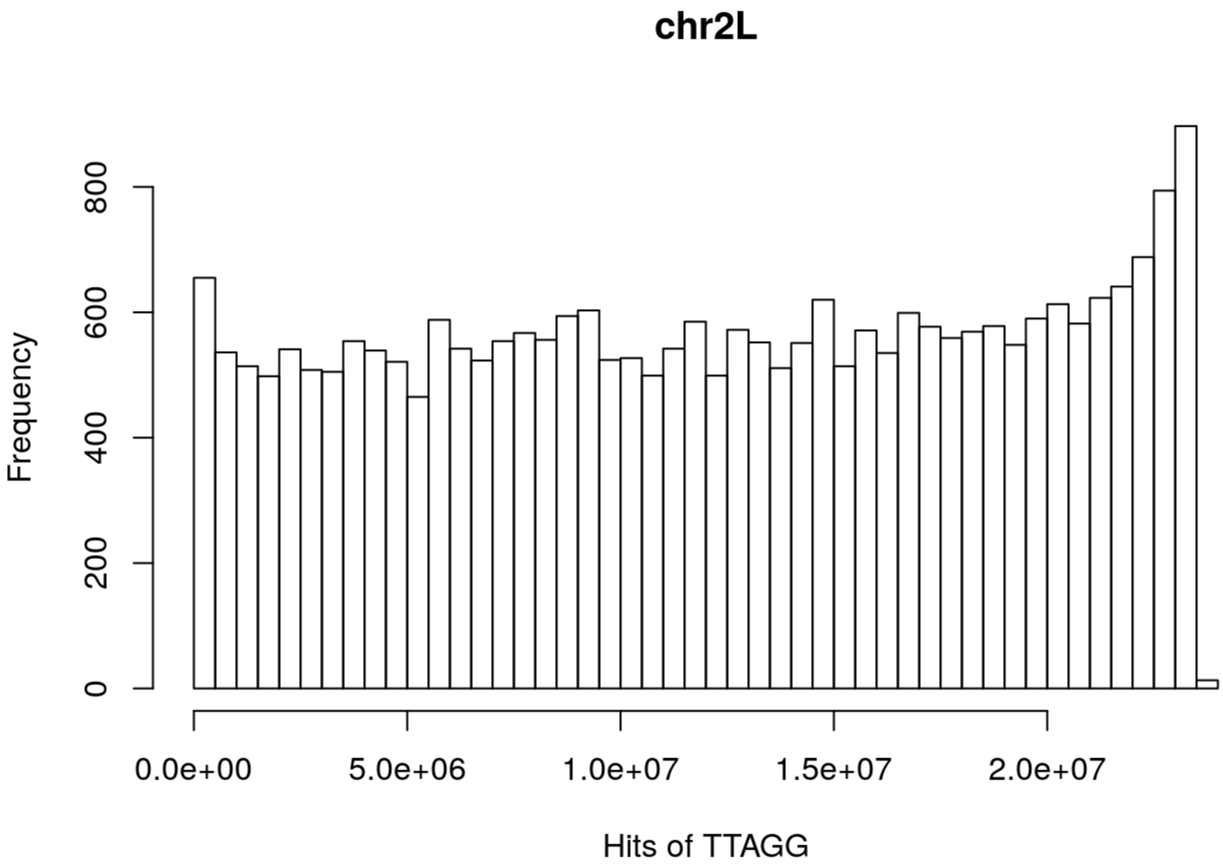
```
dna_seq = DNAString("TTAGG")
tel <- vmatchPattern(dna_seq, droso)
length(tel)
```

```
## [1] 166550
```

```
dna_seq = DNAString("TTAGG")
tel <- vmatchPattern(dna_seq, droso)
length(tel)
```

```
## [1] 166550
```

What is the distribution of this pattern depending on the chromosome? According to literature there is no repeated sequence in Drosophila melanogaster telomeres (http://telomerase.asu.edu/sequences_telomere.html (http://telomerase.asu.edu/sequences_telomere.html)) Check on chr2L as an example



##What is the difference between masekd/unmasked on chr2L as an example?

Lets explore BSgenome.Dmelanogaster.UCSC.dm3.masked.

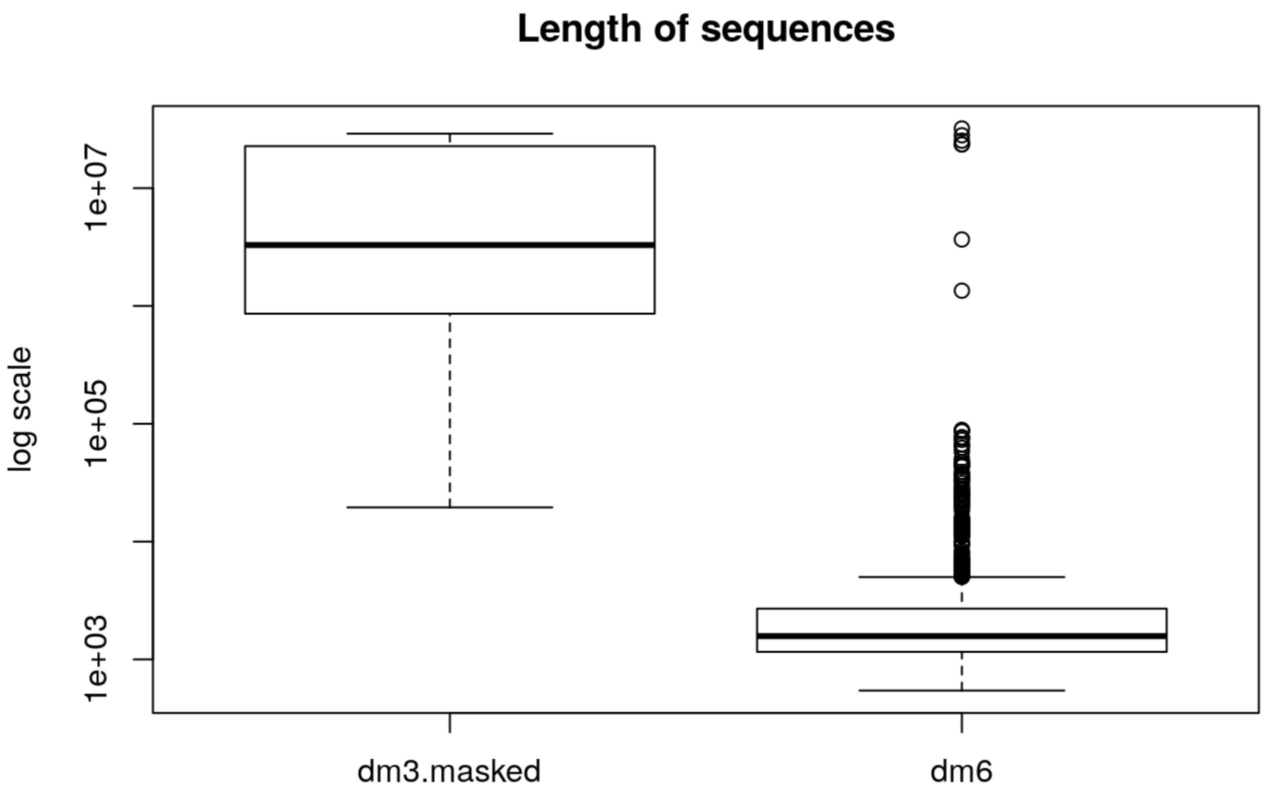
```
library(BSgenome.Dmelanogaster.UCSC.dm3.masked)
droso_masked <- BSgenome.Dmelanogaster.UCSC.dm3.masked
```

Intresting thing is that this genome contains only 15 sequnces (in comparison dm.6: 1870)

```
seqlengths(droso_masked)
```

```
##     chr2L     chr2R     chr3L     chr3R      chr4      chrX      chrU      chrM 
##  23011544  21146708  24543557  27905053   1351857  22422827  10049037     19517 
##  chr2LHet  chr2RHet  chr3LHet  chr3RHet   chrXHet   chrYHet chrUextra 
##    368872   3288761   2555491   2517507    204112    347038  29004656
```

After visualization is visible that there are 15 long reads.



To check what is the % of AT or GC in chr2L there is a fucntion:

```
## percentage of AT:   0.581646382758
```

```
## percentage of GC:   0.418353617242
```