Study of Blast output for domains A and B of TopoVI

Leslie REGAD

1 Data preparation

- Open file

Data were stored in the file `data/merge_060519.csv`. The two third columns contain sequence size for three subunits B and the two last columns contain sequence size for two subunits A.

```
fileIn <- read.table("data/merge_060519.csv", sep=";", header=T)
head(fileIn)
```

```
  Top6B_SACSH Top6B_METMA Top6B_HALMA Top6A_SACSH Top6A_METMA
1         530         621         796         389         369
2         530         621         796         389         369
3         530         621         796         389         369
4         530         621         795         389         369
5         530         621         796         389         369
6         471         621         796         389         369
```

The file contains 1423 rows and 5 columns.

- Concatenate size of sequence for the three subunits B and for the two subunits A

```
size.sB <- c(fileIn[,"Top6B_SACSH"], fileIn[,"Top6B_METMA"], fileIn[,"Top6B_HALMA"])
length(size.sB)
```

```
[1] 4269
```

```
size.sA <- c(fileIn[,"Top6A_SACSH"],fileIn[,"Top6A_METMA"])
length(size.sA)
```

```
[1] 2846
```

- We noted that some columns contains NA values. We removed these NA values.

```
ind.supp.sB <- which(is.na(size.sB)==TRUE)
size.sB <- size.sB[-ind.supp.sB]

ind.supp.sA <- which(is.na(size.sA)==TRUE)
size.sA <- size.sA[-ind.supp.sA]
```

After removing NA value, we have :
+ 4246 data for the subUnit B
+ 2512 data for the subUnit A

- concatenate these values in a list

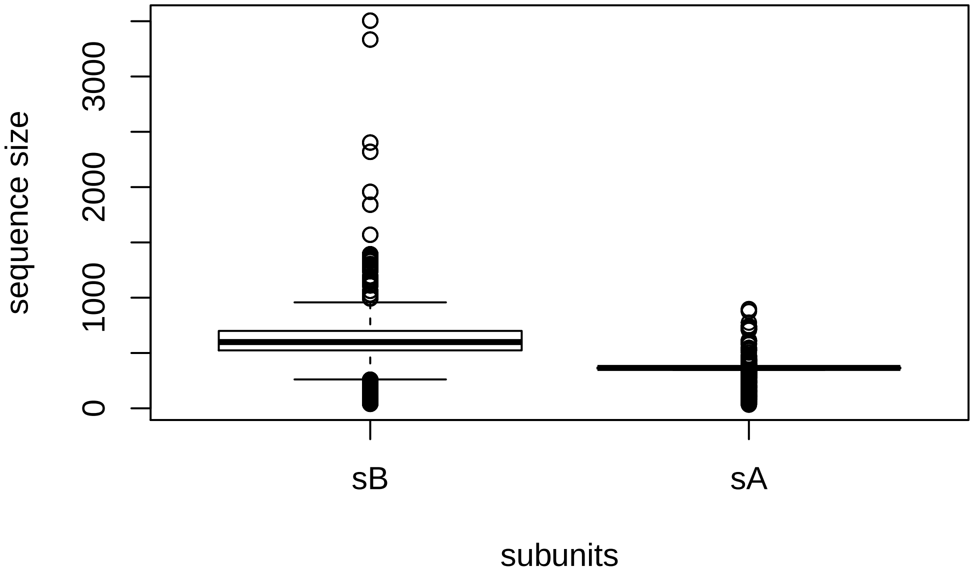
```
list.size <- vector("list", length=2)
names(list.size) <- c("sB","sA")

list.size[[1]] <-size.sB
list.size[[2]] <- size.sA
```

2 Distribution of sequence size for subUnits A and B

We plot the distribution of sequence size for the two subunits using boxplot representation.

```
boxplot(list.size, xlab="subunits", ylab="sequence size")
```



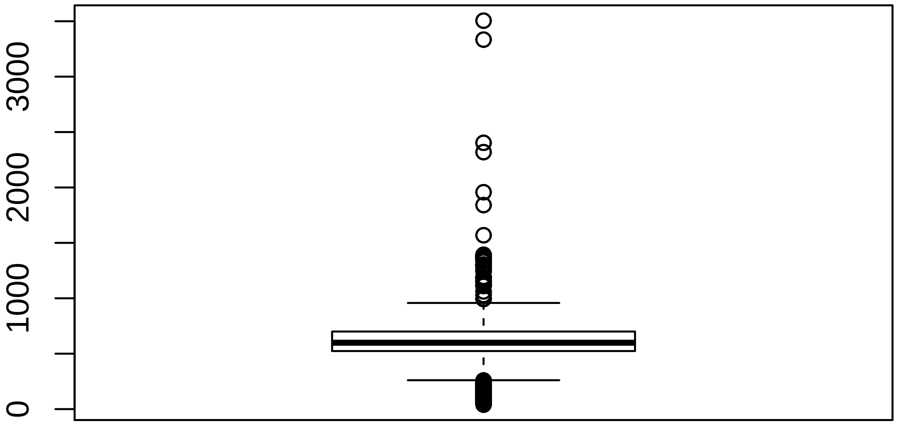
We noted that the two subunits do not have the same distribution.

- subunit B have more outliers than subunit A
- the sequence size associated to subunit B is more variable than this of subunit A

3 Determination of descriptive statistic

3.1 For subunit B :

```
boxplot.sB <- boxplot(size.sB)
```



- average size : 615.39 amino acids
- standard deviation : 245.79 amino acids
  → on average a sequence has 615.39 ± 245.79 amino acids

- minimum size : 40 amino acids
- maximum size : 3505 amino acids
- median size : 599 amino acids
  → 50% of sequences have less than 599 amino acids
- first quartile size : 524 amino acids
  → 25% of sequences have less than 524 amino acids
- third quartile size : 700 amino acids
  → 75% of sequences have less than 700 amino acids

Recompute statistic parameters after removing outliers.

- Outliers = valeurs inférieures à la valeur pivot de gauche ou supérieures à la valeur pivot de droite

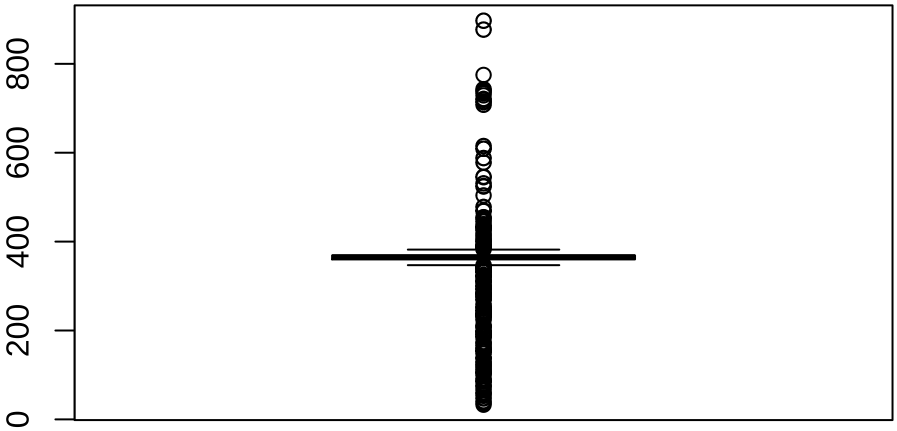
- valeur pivot de gauche = Q1-1.5*(Q3-Q1) avec Q1 et Q3 = 1er et 3ème quartiles
- valeur pivot de droite = Q3 + 1.5*(Q3-Q1) avec Q1 et Q3 = 1er et 3ème quartiles

```
size.sB.ssOutliers <- size.sB[which((size.sB> boxplot.sB$stats[1,1] ) &
                                    (size.sB< boxplot.sB$stats[5,1]))]
```

- average size after removing outliers : 614 amino acids
- standard deviation after removing outliers : 128 amino acids
- minimum size after removing outliers : 261 amino acids
- maximun size after removing outliers : 958 amino acids

3.2 For subunit A :

```
boxplot.sA <- boxplot(size.sA)
```



- average size : 354.61 amino acids
- standard deviation : 75.35 amino acids
  → on average a sequence has 354.61 ± 75.35 amino acids
- minimum size : 33 amino acids
- maximum size : 897 amino acids
- median size : 365 amino acids
  → 50% of sequences have less than 365 amino acids

- first quartile size : 360 amino acids
  → 25% of sequences have less than 360 amino acids
- third quartile size : 369 amino acids
  → 75% of sequences have less than 369 amino acids

Recompute statistic parameters after removing outliers.

```
size.sA.ssOutliers <- size.sA[which((size.sA> boxplot.sA$stats[1,1] ) &
                                       (size.sA< boxplot.sA$stats[5,1]))]
```

- average size after removing outliers : 365 amino acids
- standard deviation after removing outliers : 6 amino acids
- minimum size after removing outliers: 347 amino acids
- maximun size after removing outliers: 382 amino acids

4 Comparison of the size variance for the two subunits

- Fisher test

We performed a Fisher test to compare the two variances

```
var.test(size.sB, size.sA)
```

```

    F test to compare two variances

data:  size.sB and size.sA
F = 10.64, num df = 4245, denom df = 2511, p-value < 0.00000000000000022
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  9.919357 11.405312
sample estimates:
ratio of variances 
          10.63979 
```

Les conditions de validité du test (les échantillons doivent suivre une loi normale) n'est pas vérifiée.

- Wilcoxon test
  On va donc réaliser un test de wilcoxon sur les données suivantes :

```
data.sB <- abs(size.sB-mean(size.sB))
data.sA <- abs(size.sA-mean(size.sA))

wilcox.test(data.sB,data.sA)
```

```

    Wilcoxon rank sum test with continuity correction

data:  data.sB and data.sA
W = 8707500, p-value < 0.00000000000000022
alternative hypothesis: true location shift is not equal to 0
```

The obtained p-value is smaller than 0.05. the test is significant, thus we conclude that the variance of sequence size for subunit B is significant larger than sequence size for subunit A.