PR2 asymmetry analysis

Leslie REGAD

date()

Objectifs

- distribution de taille de chaque proteine avec sa variance
- distribution de la variance
- normaliser par le nombre de sequences (a priori, peut-etre essayer les 2)

du coup, j'ai fait un dessin qui ressemble a des montagnes et sur la distribution des variances, on a positionne la sous-unité A et la sous-unité B

- mais ce que tu proposes, au niveau des ACP me parait aussi interessant

Remarque : Je vais plutôt utiliser l'écart-type que la variance car l'écart-type aura pour unité (acides aminés), alors que la variances son unité sera (acides aminés)^2.

Importation des données

On va travailler avec le fichier `Pyrab_sizes.csv` qui contient en lignes la taille des séquences extraites.

```
fileIn <- read.table("../data/Pyrab_sizes.csv", sep=";", row.names = 1)
dim(fileIn)
```

```
[1] 641 176
```

On va donc travailler avec 641 protéines

Nombre de séquence par protéine

Tout d'abord, on va déterminer le nombre de séquences qui ont été extraites par Blast pour chaque protéine

1. On va créer une liste qui contient la taille des séquences pour chaque protéine

```
list.seq <- vector('list', length = nrow(fileIn))
names(list.seq) <- rownames(fileIn)

for(i in 1:nrow(fileIn)){
  ind.ssNA <- which(is.na(fileIn[i,])==FALSE)
  list.seq[[i]] <- as.numeric(fileIn[i,ind.ssNA])
}
```

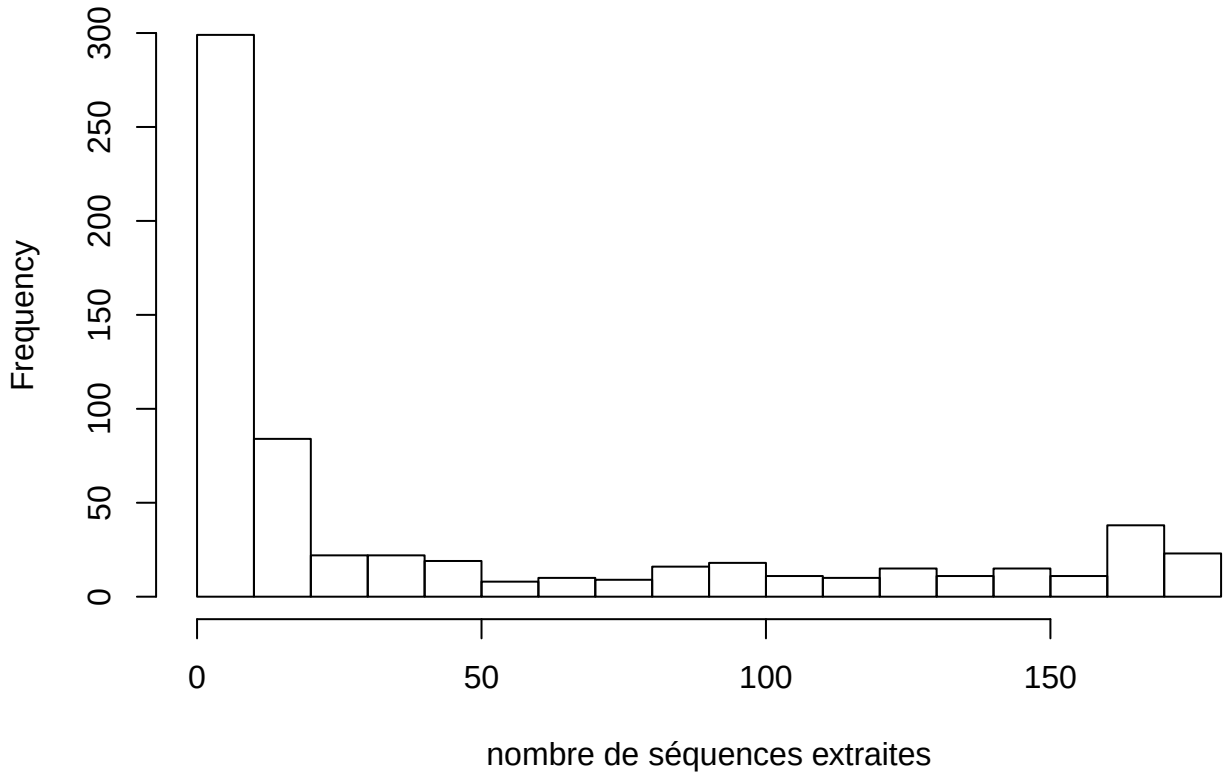
2. On détermine le nombre de séquence extraites pour chaque protéine

```
nbr.seq <- lapply(list.seq, length)
```

3. représentations graphiques :

- histogramme représentant la taille des séquences

```
hist(unlist(nbr.seq), xlab="nombre de séquences extraites",
     main="", br=20)
```



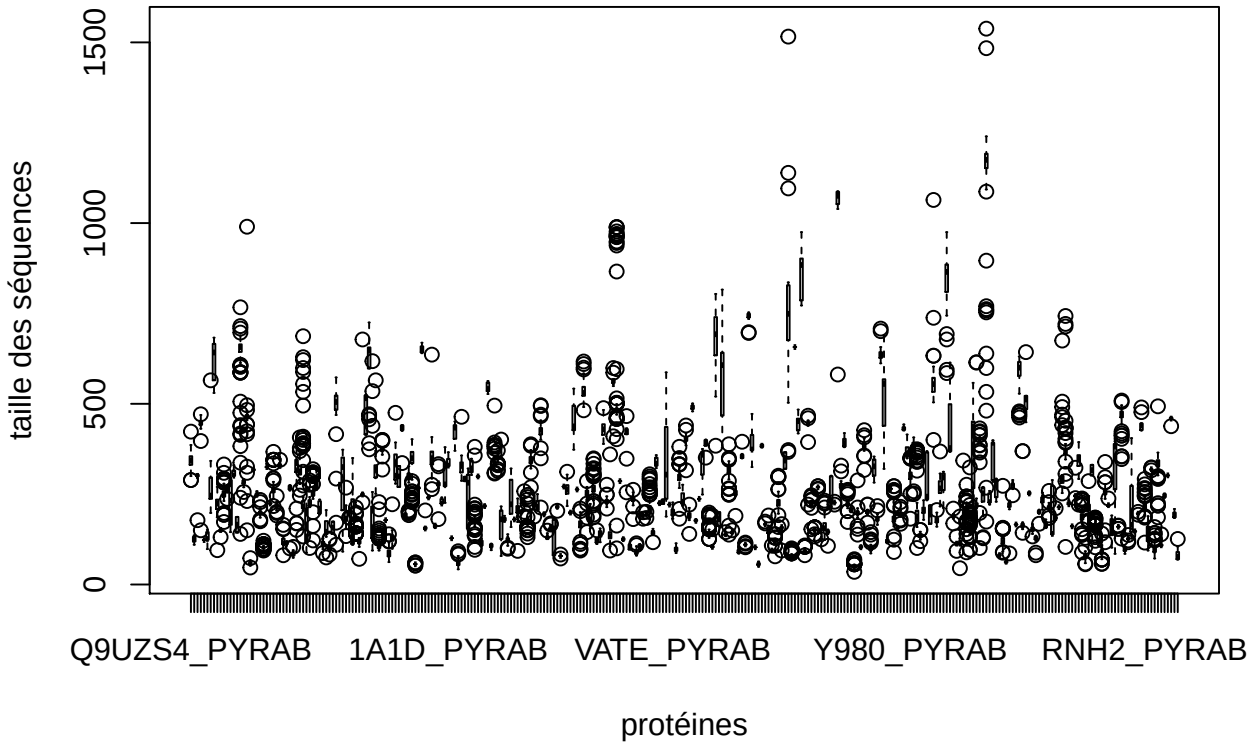
On voit que pour 40 protéines aucune séquence n'ont été extraites. On va supprimer ces protéines :

```
ind.noSeq <- which(unlist(nbr.seq)==0)
ind.avecSeq <- (1:length(nbr.seq))[-ind.noSeq]
list.seq2 <- list.seq[ind.avecSeq]
```

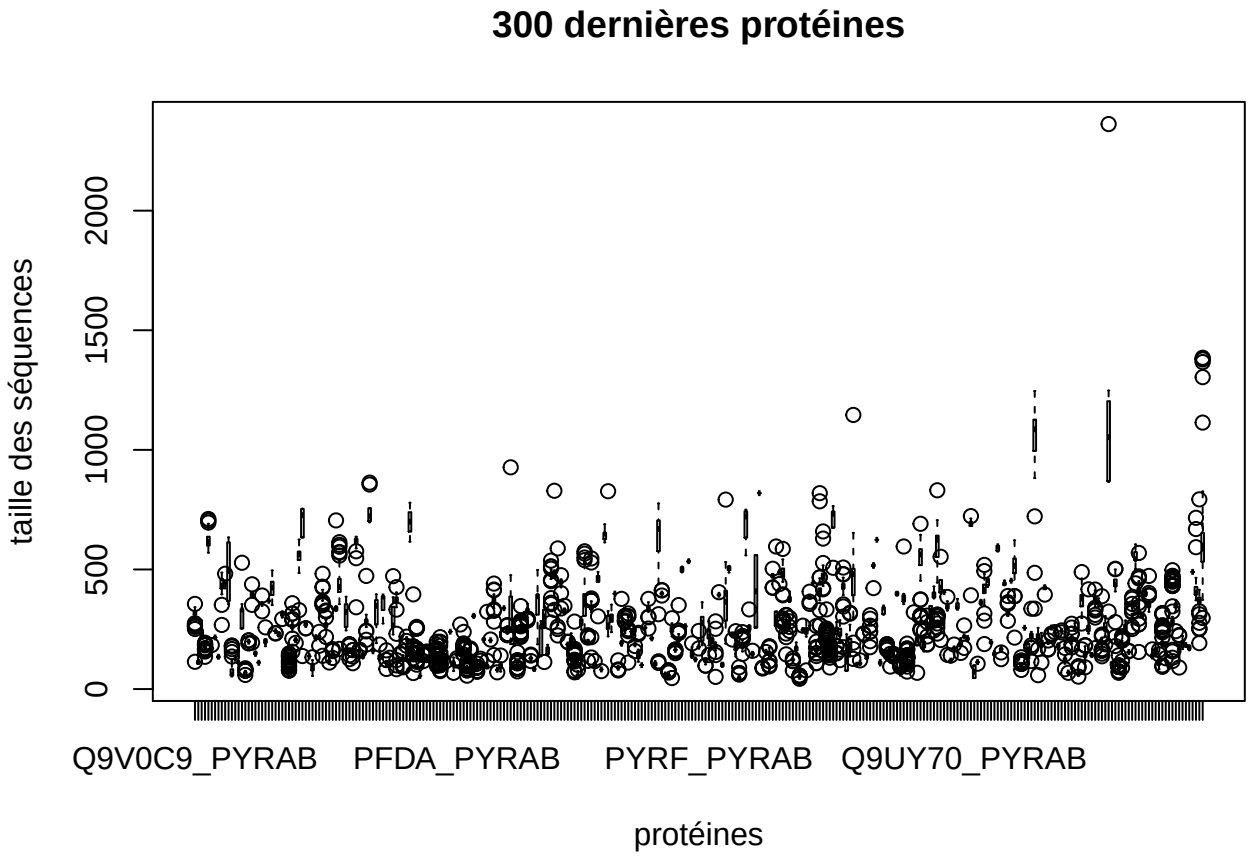
- réprésentattion de la taille de l'ensemble des séquences extraites pour chaque protéine.

On représente les 300 premières protéines puis les 300 dernières ($\rightarrow$ 2 graphs)

```
boxplot(list.seq2[1:300], ylab="taille des séquences", xlab="protéines",
        main="300 premières proteines")
```




```
boxplot(list.seq2[301:length(list.seq2)], ylab="taille des séquences", xlab="protéines",
        main="300 dernières protéines")
```



Distribution de l'écart-type de la taille des séquences

1. Distribution

```
summary(unlist(lapply(list.seq2,sd)))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   6.924  15.280  28.680  31.060 578.500      33
```

En moyenne, l'écart-type de la taille des séquences est de 28.68 acides aminés

- Les `NA` sont associées aux protéines qui n'ont qu'une séquence
- Les `0` sont associées aux protéines qui ont des séquences qui ont toutes la même taille.

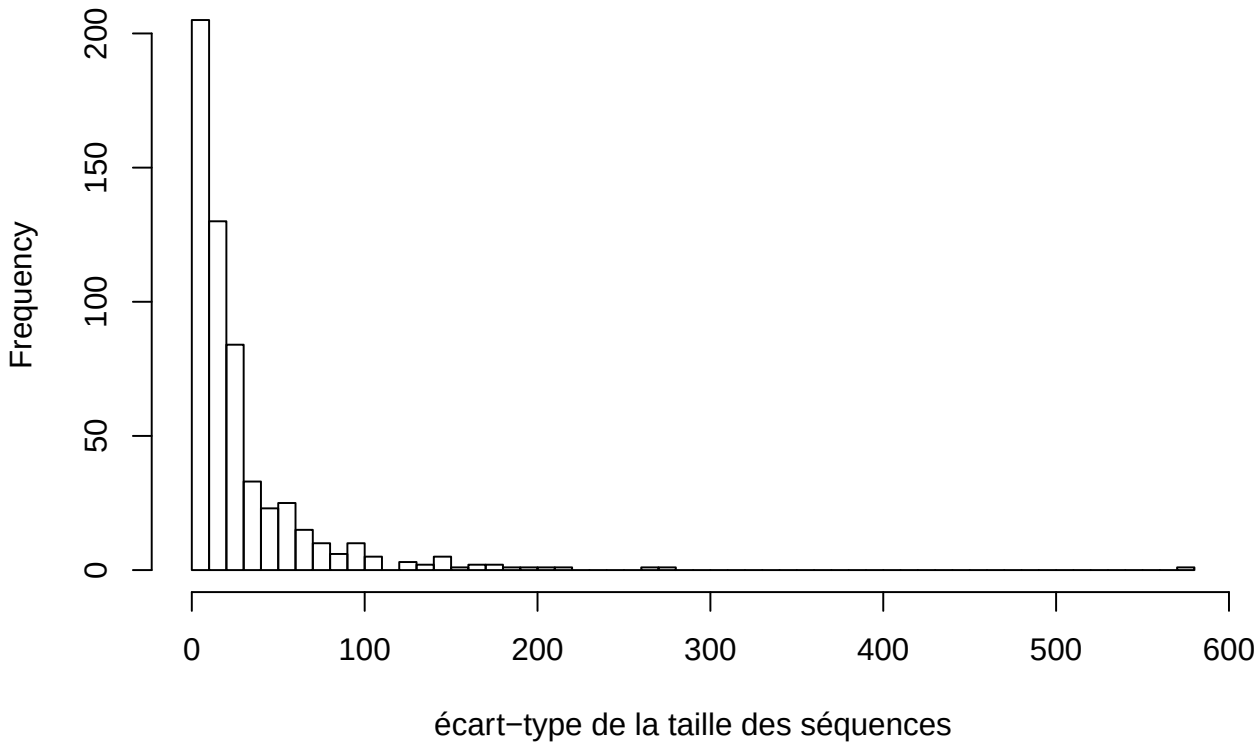
Dans ce que tu m'avais écrit par mail, tu proposais de normaliser l'écart-type de la taille des séquence par le nombre de séquence. Cette valeur est déjà prise en compte dans le calcul. L'écart-type (en acides aminés) de la taille des séquences extraites pour la protéine j, noté sd_j se calcule :

$$sd_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2}$$

avec m = la taille moyenne des séquences extraites pour la protéine j,
x_i = la taille d'une séquence i
et n le nombre de séquences extraites pour la protéine j

2. Détermination de l'écart-type de la taille des séquences extraites pour chaque protéine

```
hist(unlist(lapply(list.seq2,sd)), br=50, main="", xlab="écart-type de la taille des séquences")
```

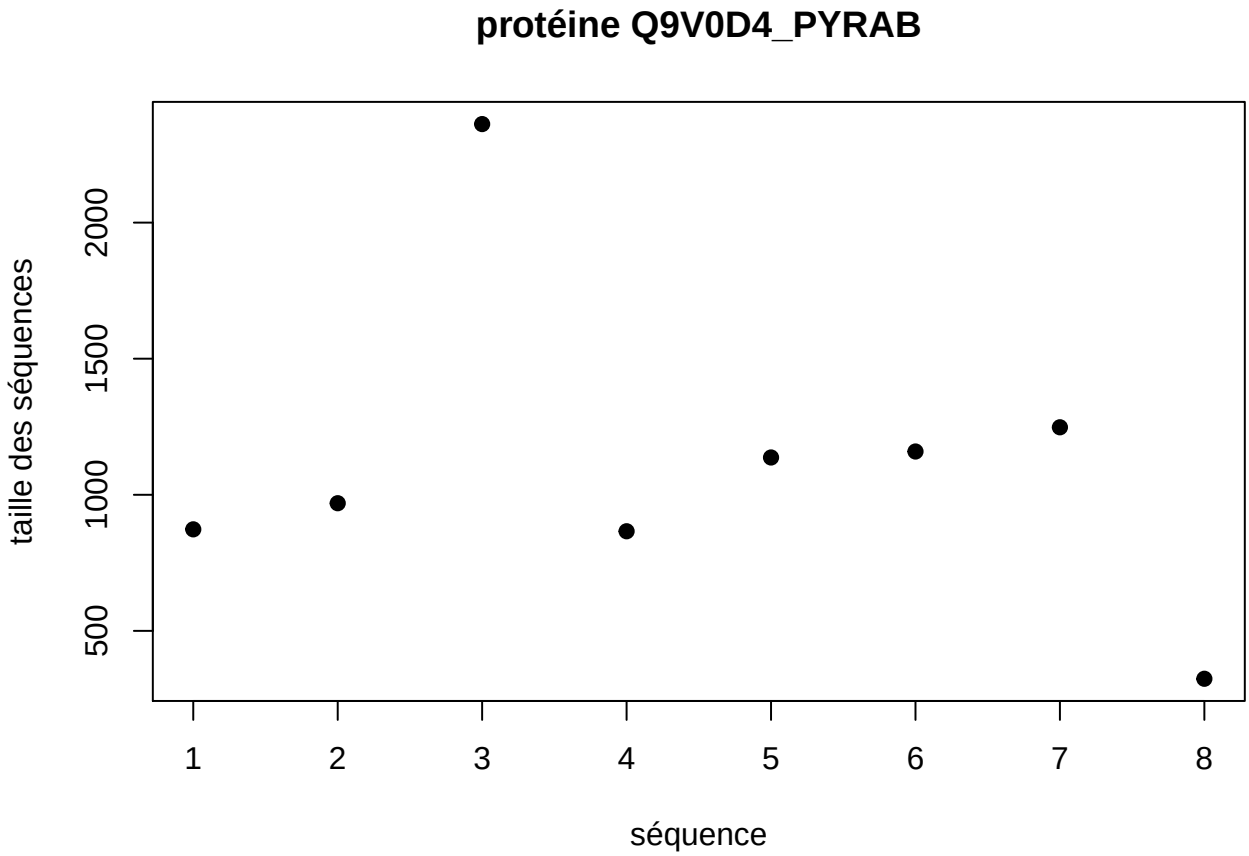


- On voit qu'il y a une protéine qui a une très grande diversité dans la taille de ses séquences (> 500). Cette protéine correspond à Q9V0D4_PYRAB. 8 séquences ont été extraites pour cette protéine. On étudie la distribution de la taille des séquences extraites pour cette protéine :

```
summary(list.seq2$Q9V0D4_PYRAB)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  324.0   871.2  1053.0  1117.0  1181.0  2362.0
```

```
plot(list.seq2$Q9V0D4_PYRAB, pch=19, xlab="séquence", ylab="taille des séquences",
     main = "protéine Q9V0D4_PYRAB")
```



On voit que pour cette protéine les séquences ont une taille allant de 324 à 2362 acides aminés.

Etudie du lien entre écart-type en termes de taille de séquences et le nombre de séquences extraites.

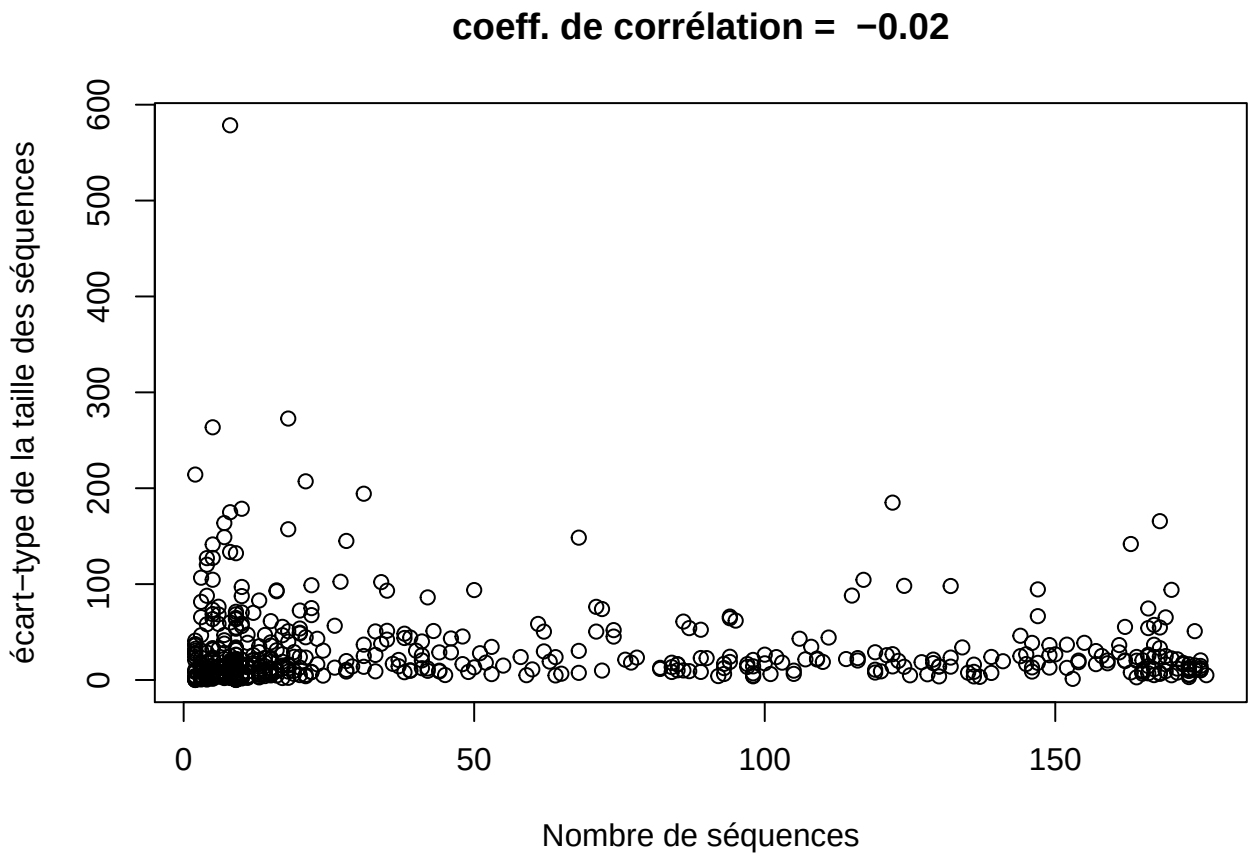
On plotte l'écart-type en fonction du nombre de séquences extraites pour chaque protéine Pour cela, on va conserver que les protéines qui ont au moins deux séquences

```
x.val <- unlist(lapply(list.seq2,length))
y.val <- unlist(lapply(list.seq2,sd))

seq.plus1 <- which(x.val>1)

coef.cor <- round(cor(x.val[seq.plus1], y.val[seq.plus1]),2)

plot(x.val[seq.plus1], y.val[seq.plus1],
     xlab="Nombre de séquences", ylab="écart-type de la taille des séquences",
     main = paste("coeff. de corrélation = ", coef.cor)
     )
```



On voit qu'il n'existe pas de lien entre le nombre de séquences extraites et la variabilité de la taille des séquences.

Localisation des protéines TopoVI sur ces données

On va localiser les données de la sous-unité A et la sous-unité B des Topo VI

Ouverture des fichiers

Ouverture du fichier contenant les tailles des séquences pour les six sous-unités des TopoVI.

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

```
  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
```

calcul les écart-types pour chaque protéine

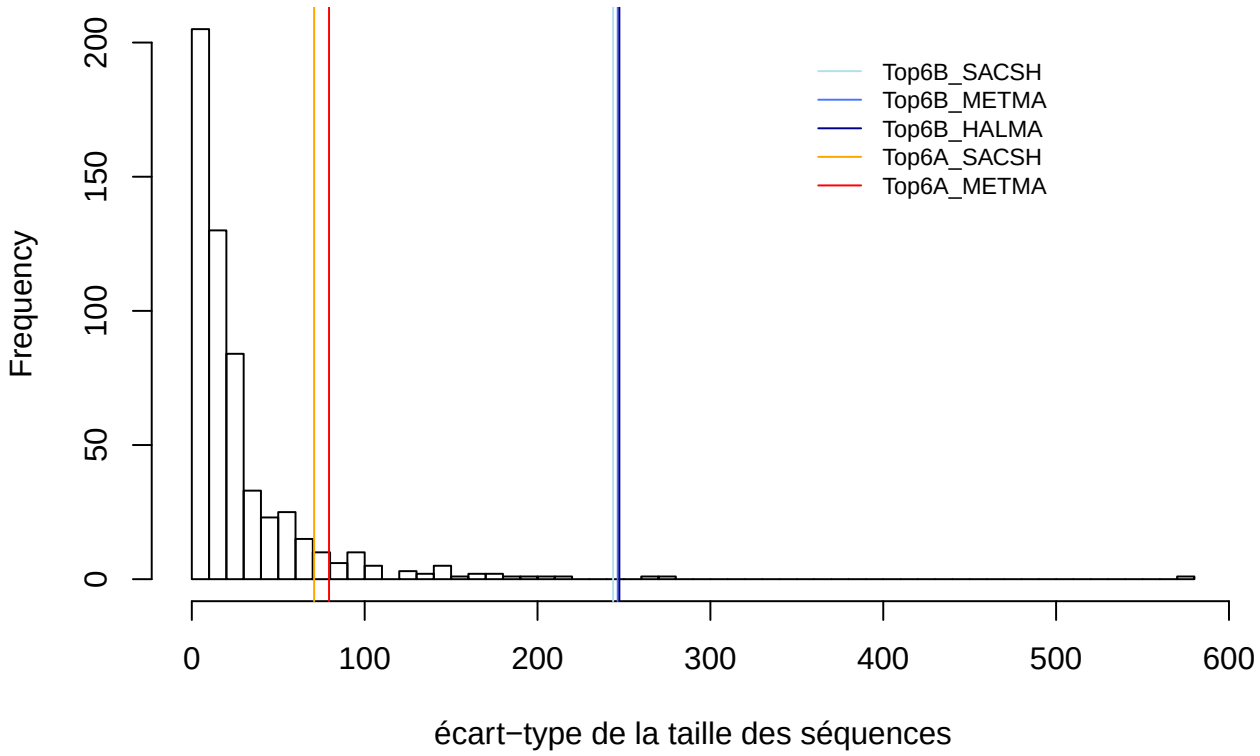
```
seq.Topo6.sd <- apply(fileIn.AB,2,sd, na.rm=T)
```

localise ces valeurs d'écart-type sur l'histogramme des écart-types

- definit les couleurs :
  - Top6B_SACSH : bleu clair
  - Top6B_METMA : bleu
  - Top6B_HALMA : bleu foncé
  - Top6A_SACSH : orange
  - Top6A_METMA : red

```
vect.col.Tp6 <- c("lightblue2", "royalblue1", "blue4", "orange", "red")
names(vect.col.Tp6) <- names(seq.Topo6.sd)
```

```
hist(unlist(lapply(list.seq2,sd)), br=50, main="", xlab="écart-type de la taille des séquences")
abline(v=seq.Topo6.sd, col = vect.col.Tp6)
legend(x=350, y=200, legend=names(seq.Topo6.sd), col = vect.col.Tp6,
       bty="n",lty=1, cex=0.76)
```



On voit que les séquences extraites pour les sous-unités B des TopoVI ont une très grande variabilité par rapport à l'ensemble des protéines.

ACP à partir des distributions des tailles des séquences

1. On neconserve que les protéines qui ont plus qu'une séquence

```
prot.plus1seq <- which(unlist(lapply(list.seq2,length))> 1)
list.seqPlus1 <- list.seq2[prot.plus1seq]
length(list.seqPlus1)
```

```
[1] 568
```

On travaille avec 568 protéines

2. On ajoute à ces données de taille de séquences, les données des topoVI

- transofrme les données des TopoVI en liste

```
list.seqT6 <- vector("list", length = ncol(fileIn.AB))
names(list.seqT6) <- colnames(fileIn.AB)


for(i in 1: ncol(fileIn.AB)){
    taille.seq <- fileIn.AB[,i]
    ind.ssNA2 <- which(is.na(taille.seq)==FALSE)
    list.seqT6[[i]] = taille.seq[ind.ssNA2]
}
```

- concatène les deux listes

```
list.seqPlus1.T6 <- vector("list", length = length(list.seqPlus1)+ncol(fileIn.AB))
names(list.seqPlus1.T6) <- c(names(list.seqPlus1), colnames(fileIn.AB))

list.seqPlus1.T6[1:length(list.seqPlus1)] = list.seqPlus1
list.seqPlus1.T6[(length(list.seqPlus1)+1):(length(list.seqPlus1.T6))] = list.seqT6
```

3. Calcul les différents paramètres caractérisant la distribution des tailles des séquences de l'ensemble des protéines

On créé une matrice qui contiendra :

- en colonne 1 : le nombre de séquences extraites pour chaque protéine
- en colonne 2 : la taille minimale des séquences extraites pour chaque protéine
- en colonne 3 : la taille moyenne des séquences extraites pour chaque protéine
- en colonne 4 : la taille médiane des séquences extraites pour chaque protéine
- en colonne 5 : la valeur quantile à 25% de la séquences extraites pour chaque protéine
- en colonne 6 : la valeur quantile à 75% de la séquences extraites pour chaque protéine
- en colonne 2 : la taille max des séquences extraites pour chaque protéine
- en colonne 3 : la valeur de l'écart-type de la taille des séquences extraites pour chaque protéine

```
mat.desc <- cbind(unlist(lapply(list.seqPlus1.T6,length)),
                  unlist(lapply(list.seqPlus1.T6,min)),
                  unlist(lapply(list.seqPlus1.T6,mean)),
                  unlist(lapply(list.seqPlus1.T6,median)),
                  unlist(lapply(list.seqPlus1.T6,quantile,0.25)),
                  unlist(lapply(list.seqPlus1.T6,quantile,0.75)),
                  unlist(lapply(list.seqPlus1.T6,max)),
                  unlist(lapply(list.seqPlus1.T6,sd))
)
colnames(mat.desc) <-c("size", "min","mean", "median","Quantile0.25",
                       "Quantile0.75","max","sd")
rownames(mat.desc) <- names(list.seqPlus1.T6)
```

4. Représentation de l'ACP (analyse en composante principale)

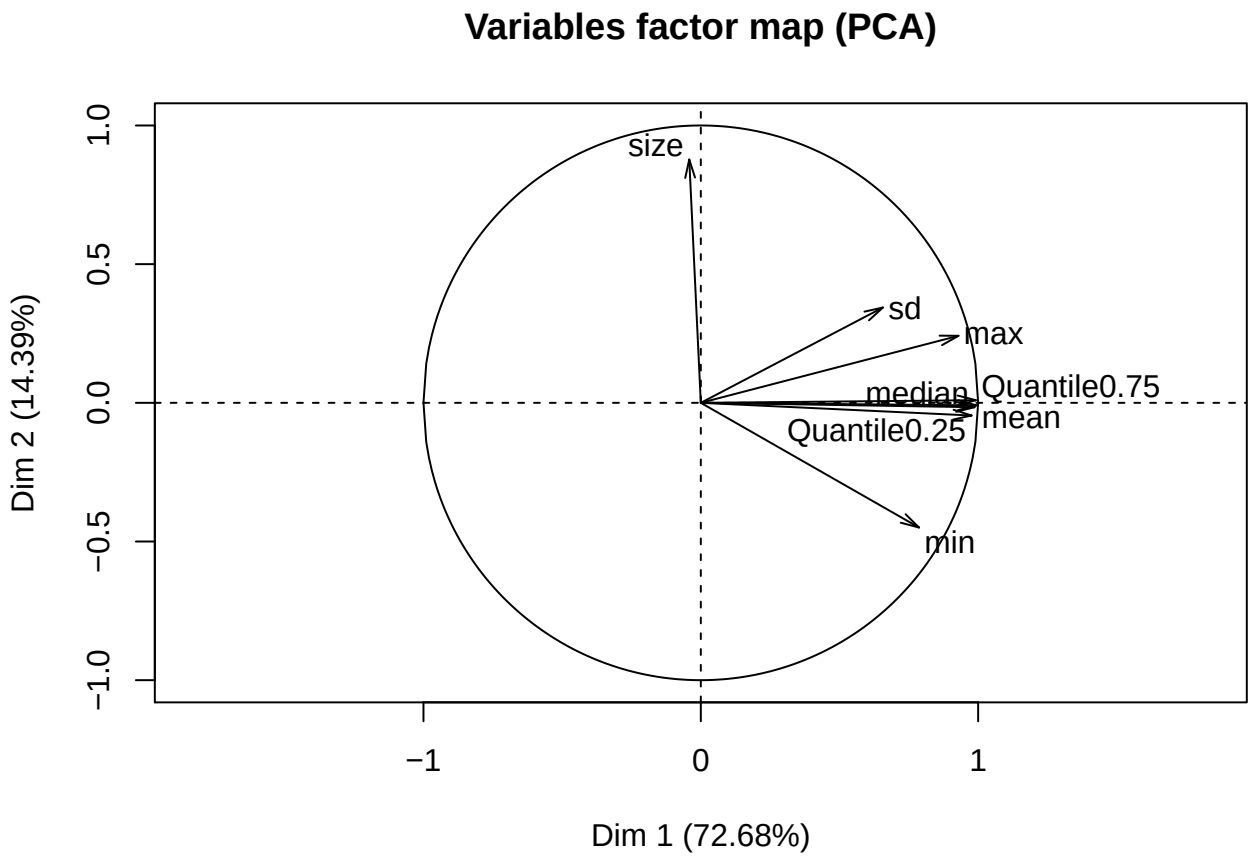
- calcul de l'ACP

Pour calculer l'ACP, on n'utilise pas les topo VI

```
ind.Topo6 <- (length(list.seqPlus1)+1):length(list.seqPlus1.T6)
pca.res <- PCA(mat.desc, graph=F, ind.sup = ind.Topo6)
```

- plot de la projection des variables

```
plot(pca.res, choix="var")
```



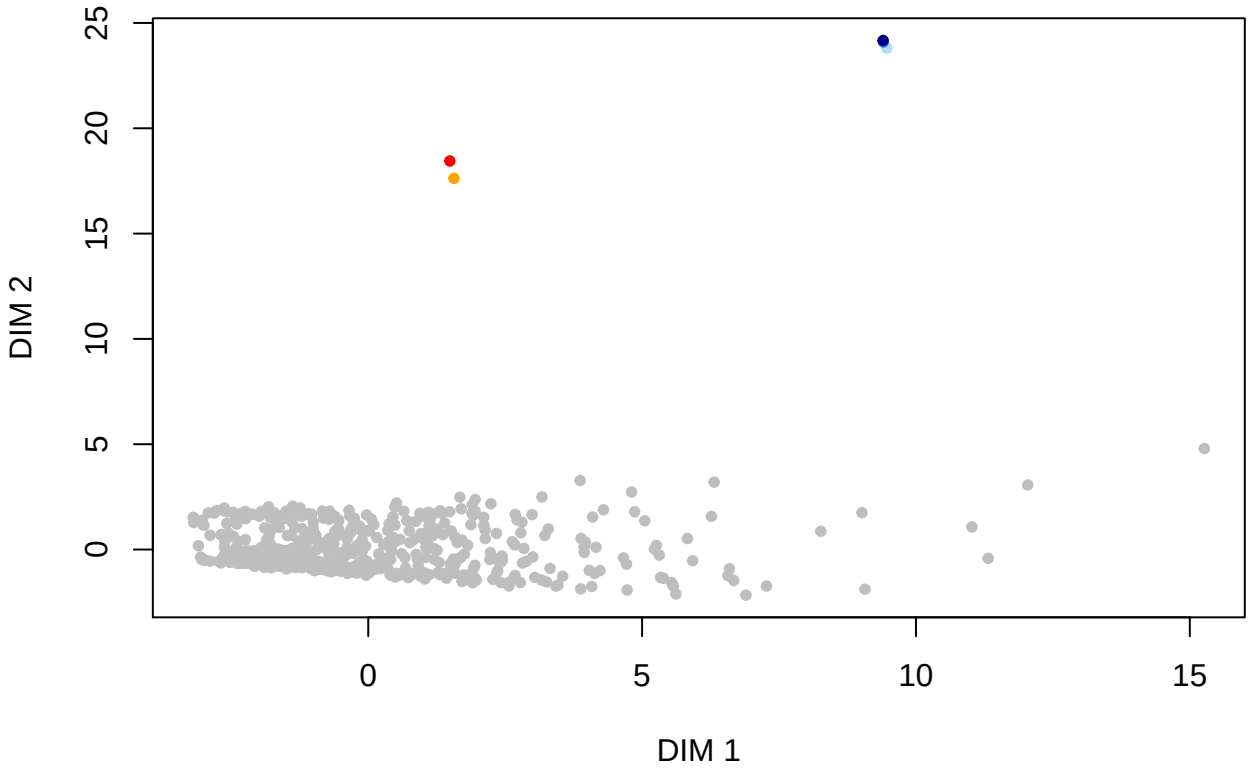
Comme attendu, il y a une forte corrélation entre les descripteurs : médiane, moyenne, valeur du quantile à 25 et 75%.

- plot la projection des individus

```
mat.coord <- rbind(pca.res$ind$coord,pca.res$ind.sup$coord)

vcol.all = rep("gray", length = nrow(mat.coord))
names(vcol.all) <- rownames(mat.coord)
vcol.all[names(vect.col.Tp6)] <- vect.col.Tp6

plot(mat.coord[,1], mat.coord[,2], pch=20, col = vcol.all,
     xlab="DIM 1", ylab="DIM 2")
```



On remarque que les protéines Topo VI présente des particularités par rapports aux autres protéines. Cela est du au fait quelle ont beaucoup plus de séquences extraites que les autres protéines. On va supprimer ce descripteur dans l'ACP.

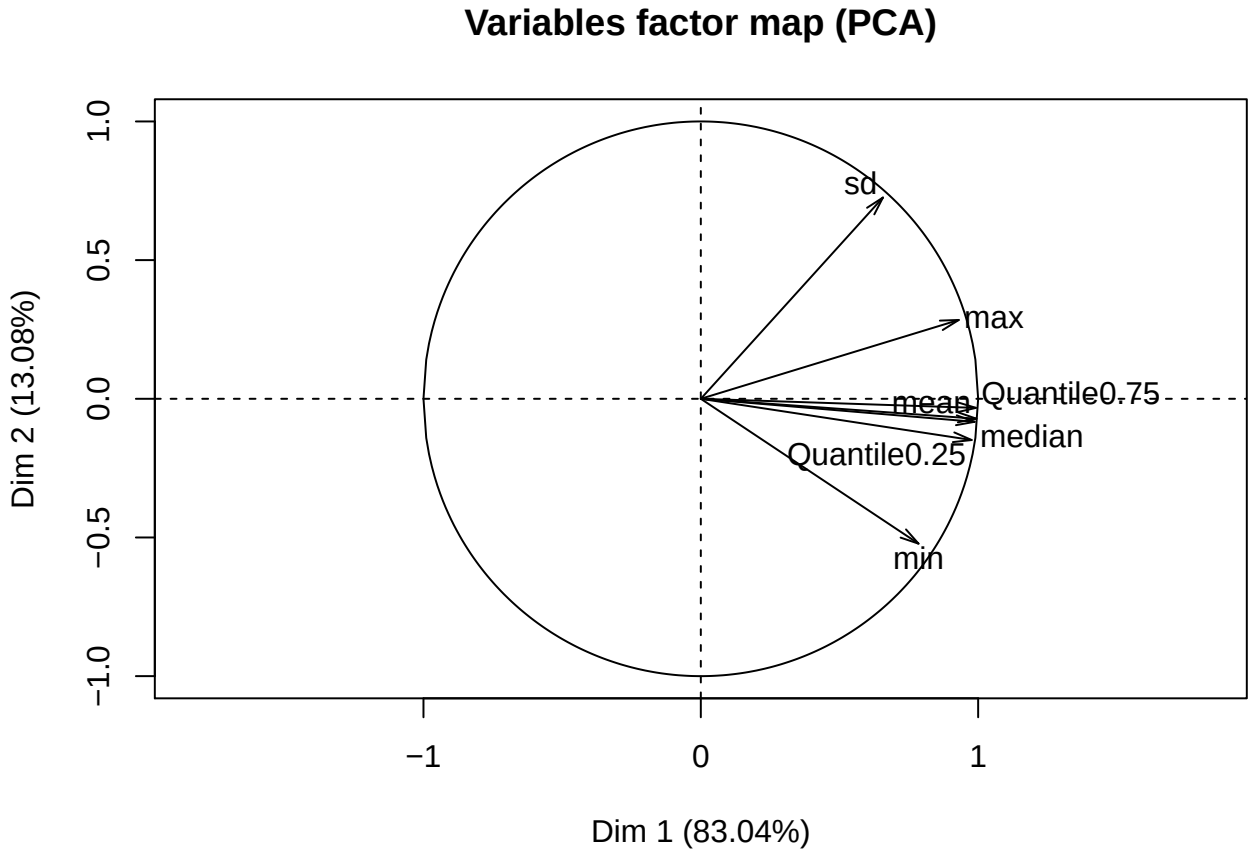
Refait l'ACP sans le descripteur "nombre de séquences""

- calcul de l'ACP

```
pca.res2 <- PCA(mat.desc[,-1], graph=F, ind.sup = ind.Topo6)
```

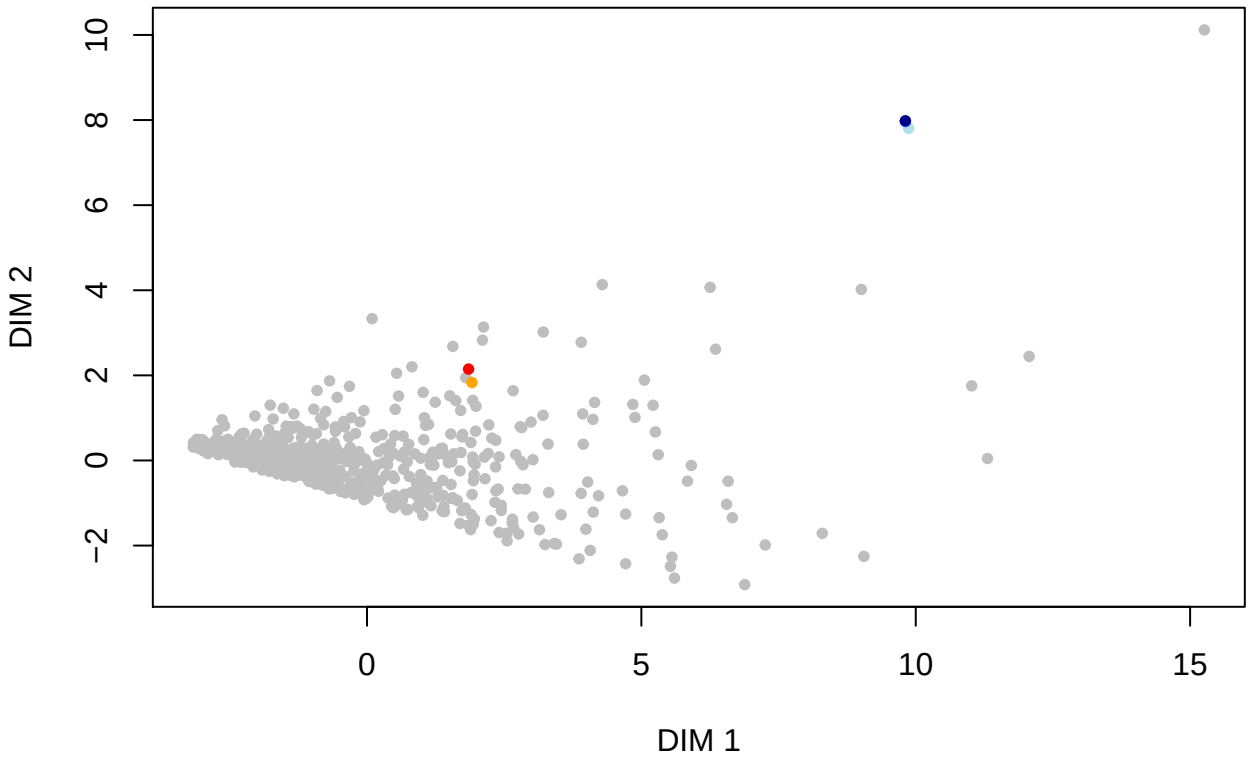
- plot la projection des variables

```
plot(pca.res2, choix="var")
```



- plot la projection des individus

```
mat.coord2 <- rbind(pca.res2$ind$coord,pca.res2$ind.sup$coord)
plot(mat.coord2[,1], mat.coord2[,2], pch=20, col = vcol.all,
     xlab="DIM 1", ylab="DIM 2")
```



On remarque que les sous-unités B des Topo VI sont particulières par rapport à l'ensemble des autres protéines.