Analyse des données génomiques 2020

Exercice

Sandrine Lagarrigue et David Causeur

Nom Prénom :

Etude de cas statistique

Les données d'expression `Colon` analysées par Alon *et al.* (1999) sont distribuées publiquement et, par exemple, disponibles dans le package `plsgeniomics` de `R`. Elles recensent les expressions de 2000 gènes pour 62 biopsies de colons humains, 22 sains et 40 tumoraux.

```
# load plsgenomics library
library(plsgenomics)

# load data set
data(Colon)

# how many samples and how many genes ?
dim(Colon$X)
```

```
[1]   62 2000
```

```
# how many samples of class 1 and 2 respectively ?
sum(Colon$Y==1)
```

```
[1] 22
```

```
sum(Colon$Y==2)
```

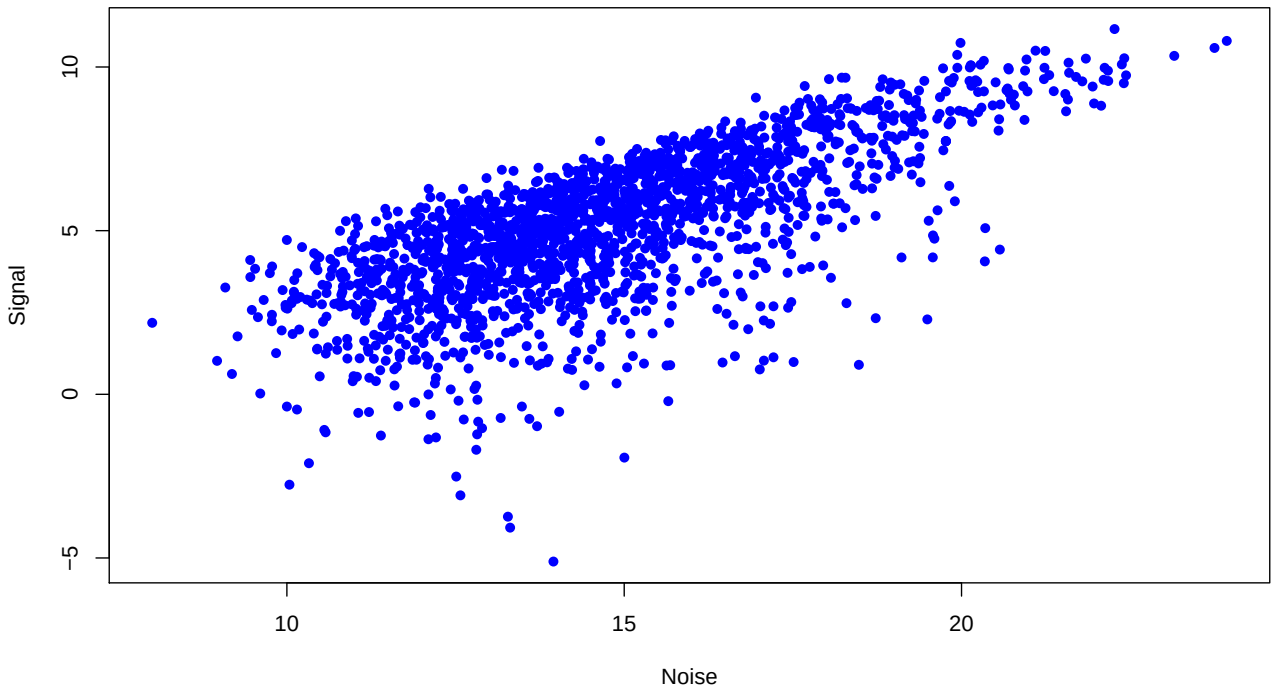
```
[1] 40
```

Afin d'analyser les différences moyennes d'expression entre les tissus sains et tumoraux, on commence par évaluer l'ampleur de ce signal biologique au regard de la variabilité intra-groupe des expressions :

```
m1 = colMeans(Colon$X[Colon$Y==1,])
m2 = colMeans(Colon$X[Colon$Y==2,])
```

```
v1 = apply(Colon$X[Colon$Y==1,],2,var)
v2 = apply(Colon$X[Colon$Y==2,],2,var)
v = (21/(21+39))*v1+((39/(21+39)))*v2

plot(log2(v),log2(abs(m2-m1)),pch=16,col="blue",
     xlab="Noise",ylab="Signal")
```



Question 1

Que contient l'objet v créé ci-dessus ?

Réponse

Question 2

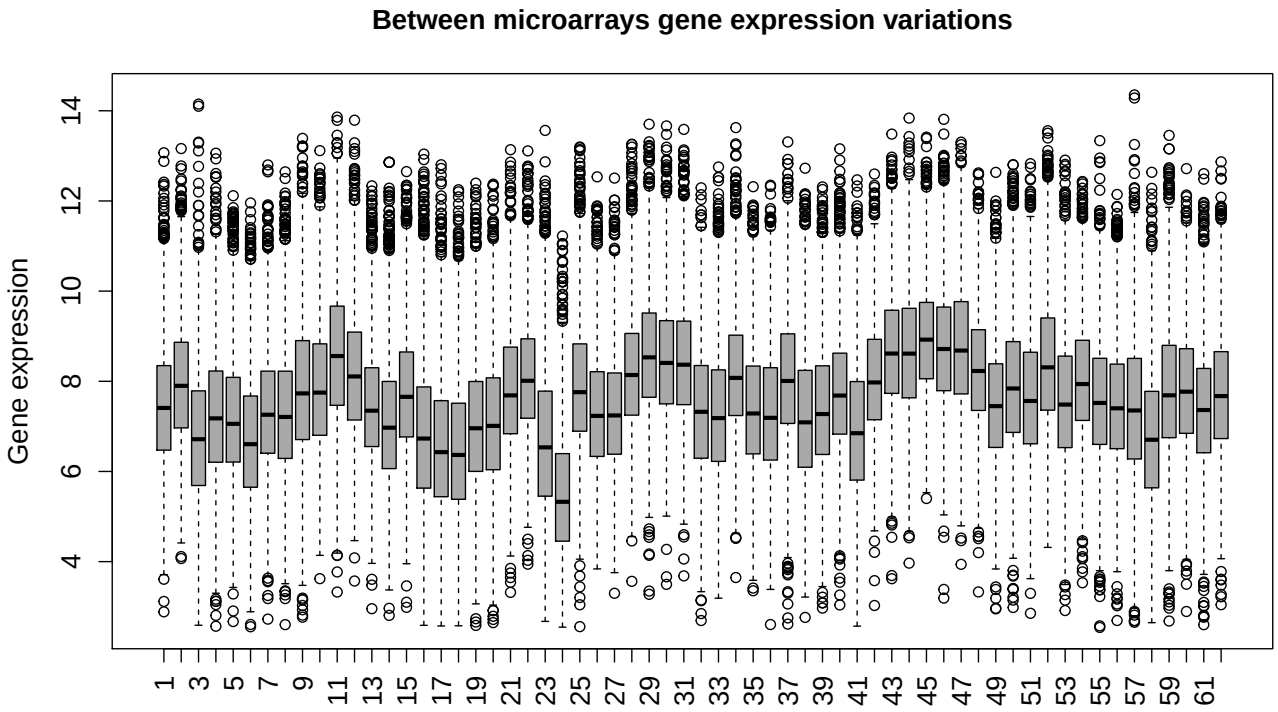
D'après le graphique créé ci-dessus, quelle opération sur les données d'expression serait souhaitable selon vous ? Expliquez en quelques mots en quoi cette opération est utile.

Réponse

Dans la suite, cette opération est réalisée, donnant lieu à la création d'un nouvel objet R nommé *Colon2*.

Le graphique suivant est obtenu à partir de ces données transformées :

```
boxplot(t(Colon2$X),col="darkgray",bty="l",xlab="",ylab="Gene expression",
        main="Between microarrays gene expression variations",
        cex.axis=1.25,cex.lab=1.25,cex.main=1.25,las=3)
```



Question 3

D'après le graphique créé ci-dessus, quelle nouvelle opération sur les données Colon2 serait souhaitable ? Expliquez en quelques mots en quoi cette opération est utile.

Réponse

Cette nouvelle opération est implémentée ci-après, donnant lieu à la création d'un nouvel objet R nommé *Colon3* :

```
Colon3 = Colon2

m = rowMeans(Colon2$X)
Colon3$X = sweep(Colon2$X,1,FUN="-",STATS=m)
```

Question 4

Que contient l'objet m créé ci-dessus ?

Réponse

On évalue à nouveau, sur les données *Colon3*, l'ampleur du signal biologique d'intérêt au regard de la variabilité intra-groupe des expressions :

```
m1 = colMeans(Colon3$X[Colon3$Y==1,])
m2 = colMeans(Colon3$X[Colon3$Y==2,])

v1 = apply(Colon3$X[Colon3$Y==1,],2,var)
v2 = apply(Colon3$X[Colon3$Y==2,],2,var)
v = (21/(21+39))*v1+((39/(21+39)))*v2
```

Question 5

Les valeurs de v calculées ci-dessus sont-elles généralement plus élevées ou plus faibles que celles qui seraient calculées de la même manière sur les données Colon2 ? Expliquez en quelques mots d'où provient cette différence.

Réponse

La ligne de code R ci-après permet d'évaluer la puissance du dispositif expérimental de l'étude `Colon` :

```
power.t.test(power=0.9,sig.level=0.05,sd=sqrt(mean(v)),delta=1)$n
```

```
[1] 12.9
```

Question 6

Que peut-on déduire du calcul ci-dessus quant à la puissance du dispositif expérimental de l'étude Colon ?

Réponse

Références

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. & Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, 96(12), 6745-6750.