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
```

Une transformation logarithmique des données d'expression est d'abord réalisée, donnant lieu à la création d'un nouvel objet R nommé *Colon2*.

```
Colon2 = Colon
Colon2$X = log2(Colon$X)
```

Une normalisation des données par centrage sur la médiane est également appliquée aux données, donnant lieu à la création d'un nouvel objet R nommé *Colon3* :

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

Le package `limma` de `BioConductor` est maintenant utilisé pour identifier les gènes dont l'expression moyenne n'est pas la même selon que le tissu est sain ou tumoral :

```
require(limma)
design = model.matrix(~Type,data=data.frame(Type=factor(Colon3$Y)))
fit = lmFit(t(Colon3$X),design,weights=NULL)
fit = eBayes(fit)
head(fit$coefficients)
```

```
  (Intercept)   Type2
1        5.03  0.1104
2        4.58  0.0110
3        4.23  0.1201
4        4.27 -0.0375
5        3.81  0.0461
6        4.36  0.1353
```

```
logFC = fit$coefficients[,2]
```

Question 1

Quelle signification pratique a le contenu de l'objet `logFC` créé ci-dessus ?

Réponse

La méthode de Benjamini-Hochberg peut être implémentée pour obtenir une sélection de gènes :

```
pval = fit$p.value[,"Type2"]
BHpval = p.adjust(pval,method="BH")
```

Question 2

Expliquez comme vous le feriez à un collègue ne connaissant rien à la statistique génomique les différences en matière de sélection de gènes entre la méthode consistant à considérer qu'un gène est positif si la

p-value correspondante est inférieure à 0.05 et celle consistant à considérer qu'un gène est positif si la p-value ajustée par la méthode de Benjamini-Hochberg correspondante est inférieure à 0.05.

Réponse

A titre illustratif, on s'intéresse ci-après au gène dont les mesures d'expression sont dans la 520ème colonne du tableau des données d'expression (la p-value pour ce gène est au rang 100 si on classe toutes les p-values par ordre croissant) :

```
# Numéro du gène dont la p-value est au rang 100 (ordre croissant)
ord_100 = order(pval)[100]
ord_100
```

```
[1] 520
```

Les commandes ci-dessus donnent la p-value pour ce gène et la p-value ajustée par la méthode de Benjamini-Hochberg :

```
pval[ord_100]
```

```
    520
0.00017
```

```
BHpval[ord_100]
```

```
   520
0.0034
```

Question 3

En prenant l'exemple ci-dessus du gène de la colonne 520, donnez l'opération permettant d'obtenir la p-value ajustée par la methode de Benjamini-Hochberg à partir de la p-value.

Réponse

On choisit de considérer comme positifs les gènes dont la p-value stockée dans l'objet `pval` est inférieure ou égale à 0.00017.

Question 4

Combien obtient-on de gènes positifs ?

Réponse

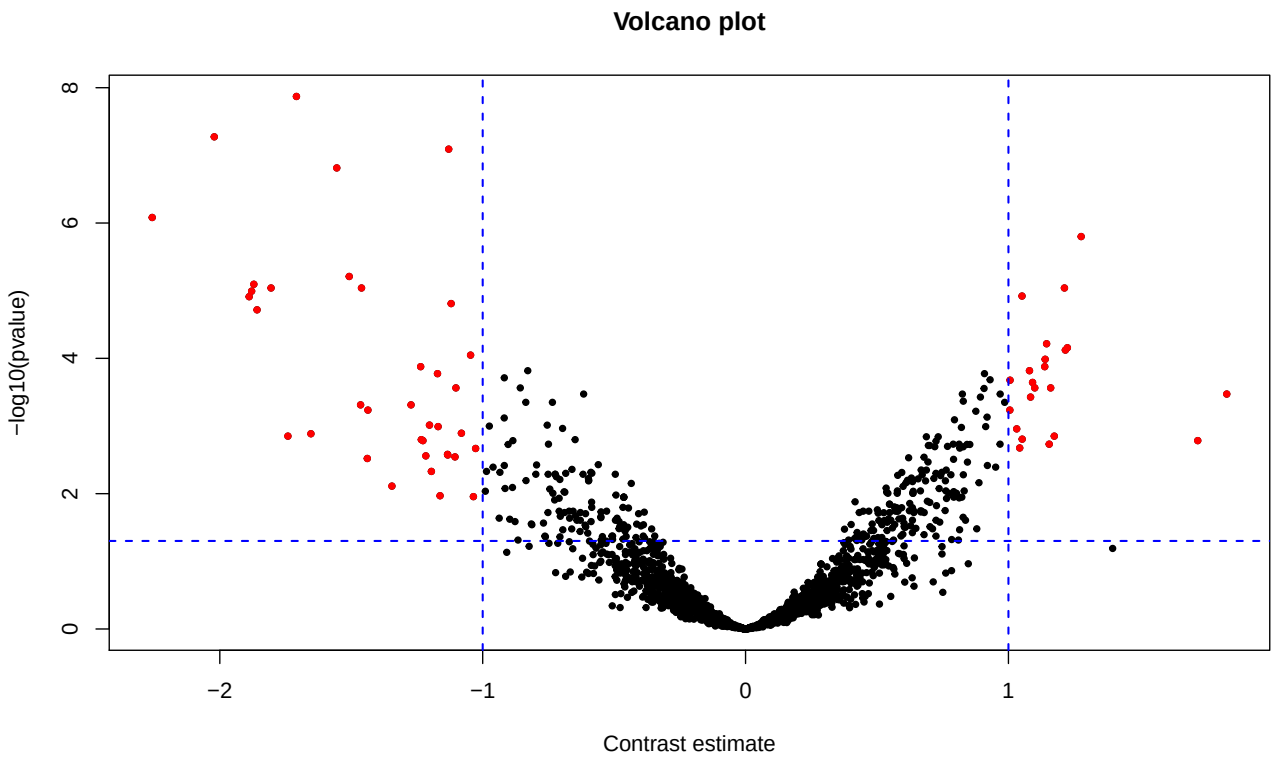
Question 5

Quelle est la valeur estimée du taux de gènes faux positifs dans cette liste ?

Réponse

On construit maintenant une règle de sélection des gènes basée à la fois sur `logFC` et sur `BHpval`. Les points en rouge sur le graphique ci-après correspondent aux gènes sélectionnés :

```
plot(logFC,-log10(BHpval),xlab = "Contrast estimate",
     ylab = "-log10(pvalue)",main = "Volcano plot",cex = 0.6, pch = 19)
points(logFC[(BHpval < 0.05)&(abs(logFC)>1)],
       -log10(BHpval)[(BHpval < 0.05)&(abs(logFC)>1)],
       cex = 0.6, pch = 19, col = "red")
abline(h = -log10(0.05),col = "blue",lty =2,lwd = 1.5)
abline(v = c(-1,1),col="blue",lty=2,lwd=1.5)
```



Question 6

Expliquez la règle de sélection de gènes ci-dessus.

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.