Les tests basés sur les permutations

Echantillons de tailles faibles et nombre important de variables

Yohann.Foucher@univ-nantes.fr

Equipe d'Accueil 4275 "Biostatistique, recherche clinique et mesures subjectives en santé", Université de Nantes

MASTER 1 Bionformatique et Biostatistique - UE données omics

Plan

Plan

Problématique

- Les nouveaux outils d'analyse à haut débit permettent la mesure simultanée de milliers caractéristiques d'un individu

  ex : Puces ADN : elles peuvent mesurer plusieurs milliers de gènes.

- Objectif : identifier des caractéristiques différemment exprimées entre plusieurs groupes.

  ex : identifier les gènes différemment exprimés entre les malades et les non-malades.

Update on Gene Expression Analysis, Proteomics, and Network Discovery

Gene Expression Analysis, Proteomics, and Network Discovery[1]

Sacha Baginsky, Lars Hennig, Philip Zimmermann, and Wilhelm Gruissem*

Department of Biology and Zurich-Basel Plant Science Center, ETH Zurich Universitätstrasse 2, 8129 Zurich, Switzerland

402 *Plant Physiology*℠, February 2010, Vol. 152, pp. 402–410, www.plantphysiol.org

Problématique

Table II. *Overview of some of the most popular plant gene expression microarray platforms and the number of available experiments in ArrayExpress*

The Arabidopsis ATH1 array is the most frequently used microarray, followed by the CATMA 25k and 23k arrays. In all, approximately 750 Arabidopsis microarray experiments have been published so far. Rice (*Oryza sativa*) and barley (*Hordeum vulgare*) are the second and third plant species in terms of microarray experiments published. Soybean (*Glycine max*) also has a high number of arrays, but this is due to a single very large experiment containing 2,521 arrays. IPK, Leibniz Institute of Plant Genetics and Crop Plant Research; TIGR, The Institute for Genomic Research.

Species	Provider	Array Format	Array Name	Experiments	Arrays
Arabidopsis	Affymetrix	8K	AG	41	352
	Affymetrix	22K	ATH1	554	8,895
	Agilent	22K	Arabidopsis 2	34	253
	Agilent	44K	Arabidopsis 3	7	60
	CATMA	25K	CATMA2_URGV to CATMA2.3_URGV	83	851
	CATMA	23K	CATMA Arabidopsis 23K array	50	1,290
	TIGR	26K	TIGR Arabidopsis whole genome	6	264
Rice	Affymetrix	57K	GeneChip Rice Genome Array	29	418
	Agilent	21K	Agilent Rice Oligo Microarray	22	164
Barley	Affymetrix	22K	GeneChip Barley Genome Array	35	1,165
	IPK	6K + 4K	IPK barley PGRC1_A and B	7	324
Medicago	Affymetrix	61K	GeneChip Medicago Genome Array	19	218
Maize	Affymetrix	17K	GeneChip Maize Genome Array	22	370
Soybean	Affymetrix	61K	GeneChip Soybean Genome Array	22	3,236
Tomato (*Solanum lycopersicum*)	Affymetrix	10K	GeneChip Tomato Genome Array	6	127
Grape (*Vitis vinifera*)	Affymetrix	16K	GeneChip *Vitis vinifera* Genome Array	6	239
Wheat (*Triticum aestivum*)	Affymetrix	61K	GeneChip Wheat Genome Array	25	811
Total				968	19,037

Limites des méthodes traditionnelles

- Méthode : calcul des probabilités critiques (pvalue *).

→ Soit p_j la probabilité critique associée au gène j.

- Supposons $k = 1000$ gènes étudiés.
- Deux groupes A et B de tailles N_A et N_B
- On observe les expressions du gène j :
  - $x_{jA1}, x_{jA2}, \ldots, x_{jAN_A}$ dans le groupe A

  → moyenne : $\bar{x}_{jA}$ et écart-type s_{jA}
  - $x_{jB1}, x_{jB2}, \ldots, x_{jBN_B}$ dans le groupe B

  → moyenne : $\bar{x}_{jB}$ et écart-type s_{jB}

$*$. probabilité que la différence observée soit due au hasard (probabilité critique en français)

Limites des méthodes traditionnelles

- Hypothèses :

  H_0 : $\bar{X}_{jA} = \bar{X}_{jB}$, la moyenne d'expression du gène j est identique entre les deux populations.

  H_1 : $\bar{X}_{jA} \neq \bar{X}_{jB}$, la moyenne d'expression du gène j est différente entre les deux populations.

- Soit T_j la statistique de test :

$$T_j = \frac{\bar{X}_{jA} - \bar{X}_{jB}}{\sqrt{\frac{S_{jA}^2}{N_A} + \frac{S_{jB}^2}{N_B}}}$$

- En considérant H_0 comme vraie et si N_A et $N_B > 30$:

$$T_j \sim \mathcal{N}(0,1) \text{ théorème centrale limite}$$

- Si $p_j < 0.05 \Rightarrow$, il semble que les moyennes d'expression du gènes soient différentes entre les deux populations.

Limites des méthodes traditionnelles

```
> N.A <- 100
> N.B <- 100
> k <- 1000
> E.A <- rnorm(N.A, mean=0, sd=1)
> E.B <- rnorm(N.B, mean=0, sd=1)
> E.A
  [1]  0.259767012  0.104756575  0.309616563  0.624475170 -0.627431940
  [6]  1.262903438  0.835544333 -0.806822577  0.636655175  0.001920257
 [11]  1.308219156 -0.593271319 -1.781291756  0.928474836  0.047191736
 [16]  0.097193687  0.188600260 -0.896055910 -0.780684894 -0.903431156
 [21] -1.404657666 -1.112908505 -0.482515098 -0.305200701 -2.959189767
 [26] -0.619398436  0.964401705 -0.677996115  0.570158887  0.500985102
 [31]  1.310491882 -0.026689214 -0.834014767  1.487892811 -0.806169717
 [36]  0.875037438 -0.615388338 -0.767584042 -1.006066605 -1.790111173
 [41] -0.711082650  2.402996565  1.579152671 -1.360627507  0.557139166
 [46] -0.233817692  0.906085507 -0.816444847 -0.745482092 -1.768722704
 [51]  0.653054726 -1.139583695 -0.799815577 -1.194266380 -0.224372307
 [56] -0.892277328 -0.986516414  1.168128361 -1.765169418 -0.127545667
 [61]  2.073436445 -0.679708254 -1.397863208 -1.922954639 -0.310621692
 [66]  0.691923479  0.378232604  0.304669060 -1.386852628 -0.352114839
 [71] -1.356227305  0.265875723 -1.064897180  0.506978879  0.241024861
 [76] -0.959748501  0.678611734 -0.068313695  0.593750095 -0.370472628
 [81]  0.454820690  0.342254341 -0.746135230 -0.955464046  1.456725980
 [86]  0.388536819 -0.162063823  0.708618518  0.358394361 -0.336222683
 [91]  0.580267446  0.008241462  0.808807889  0.680924584 -0.872851582
 [96]  0.024976298 -2.223802297 -0.706544494 -1.916743943  2.188871939
> E.B
  [1]  2.274006261 -0.177904256 -0.517614196  1.867644610 -0.889814720
```

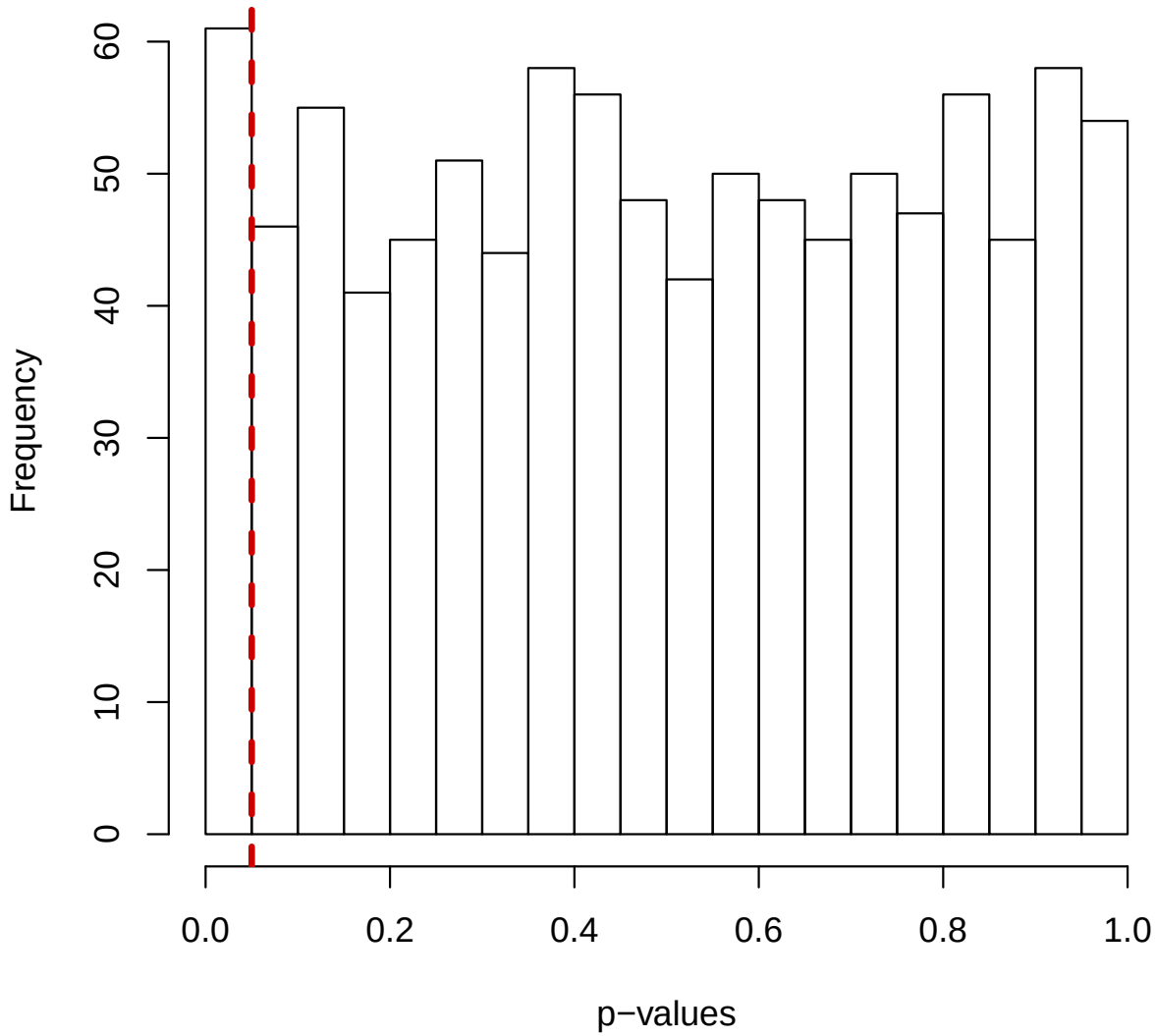
Limites des méthodes traditionnelles

```
> t.test(E.A, E.B)

Welch Two Sample t-test

data:  E.A and E.B
t = -1.868, df = 197.301, p-value = 0.06324
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.56670225  0.01535635
sample estimates:
 mean of x  mean of y
-0.1703542  0.1053187

> t.test(E.A, E.B)$p.value

[1] 0.06324423
```

Limites des méthodes traditionnelles

```
> pvalues <- rep(NA, k)
> for(j in 1:k)
+ {
+ E.A <- rnorm(N.A, mean=0, sd=1)
+ E.B <- rnorm(N.B, mean=0, sd=1)
+ pvalues[j] <- t.test(E.A, E.B)$p.value
+ }
> pvalues[1:10]

 [1] 0.8066186 0.8363131 0.8869664 0.4589966 0.3805023 0.2423012
 [7] 0.2317830 0.5548061 0.9269442 0.1275307

> table(pvalues<0.05)

FALSE  TRUE
  939    61

> table(pvalues<0.05)/k

FALSE  TRUE
0.939 0.061

> hist(pvalues, nclass=30, xlab="p-values", main="")
> abline(v=0.05, col="red3", lwd=3, lty=2)
```

Limites des méthodes traditionnelles



Problématique

Nombre variables >>> Nombre de sujets

1. Impossibilité de connaitre la distribution des statistiques de test sous l'hypothèse nulle.
2. Comparaisons multiples et augmentation du nombre d'erreurs de 1ère espèce.

Plan

Multiple Test Procedure (MTP) : Références

Article M.J. van der Laan, S. Dudoit, K.S. Pollard (2004), Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives, Statistical Applications in Genetics and Molecular Biology, 3(1).

Livre S. Dudoit and M.J. van der Laan. Multiple Testing Procedures and Applications to Genomics. Springer Series in Statistics. Springer, New York, 2008.

R Package "multtest".

MTP : Deux principes fondamentaux

1. Estimation non-paramétrique des distributions des statistiques de test sous H_0.

2. Pénalisation des probabilités critiques pour la prise en compte des comparaisons multiples.

Plan

Principe de la permutation

- Permutation des libellés des colonnes (groupes A/B).
- Distributions des gènes deviennent indépendantes des groupes.

Données observées : H_0 ou H_1 ?

Groupes	A	A	A	A	A	B	B	B	B	B
Valeurs	1.2	2.4	0.5	0.7	1.0	2.2	3.4	1.5	0.7	0.9

$\Downarrow$

Données permutées : H_0 est vraie

Groupes	B	A	B	B	A	B	A	B	A	A
Valeurs	1.2	2.4	0.5	0.7	1.0	2.2	3.4	1.5	0.7	0.9

Construction des distributions sous H_0 et calcul des p-values

- On calcul classiquement la statistique de test pour les k gènes :

$$t_1, t_2, ..., t_k$$

- On réalise B itérations. Pour chaque itération ($b = 1, 2, ..., B$) :
  - Permutation des colonnes.
  - Calcul des statistiques de test pour chaque gène :

$$t_1^{(b)}, t_2^{(b)}, ..., t_k^{(b)}$$

- Calcul des probabilités critiques (non corrigées) :

$$p_j^* = \frac{\sum_{b=1}^{B} I(|t_j^{(b)}| \geq |t_j|)}{B}$$

où $I(a)$ est égale à 1 si la condition a est respectée et 0 sinon.

Exemple

```
> Groupes <- c("A", "A", "A", "A", "A", "B", "B", "B", "B", "B")
> Valeurs <- c(1.2, 2.4, 0.5, 0.7, 1.0, 2.2, 3.4, 1.5, 0.7, 0.9)
> Valeurs[Groupes=="A"]

[1] 1.2 2.4 0.5 0.7 1.0

> Valeurs[Groupes=="B"]

[1] 2.2 3.4 1.5 0.7 0.9

> t.test(Valeurs[Groupes=="A"], Valeurs[Groupes=="B"])

Welch Two Sample t-test

data:  Valeurs[Groupes == "A"] and Valeurs[Groupes == "B"]
t = -0.9787, df = 7.036, p-value = 0.3602
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.9798766  0.8198766
sample estimates:
mean of x mean of y
     1.16      1.74
```

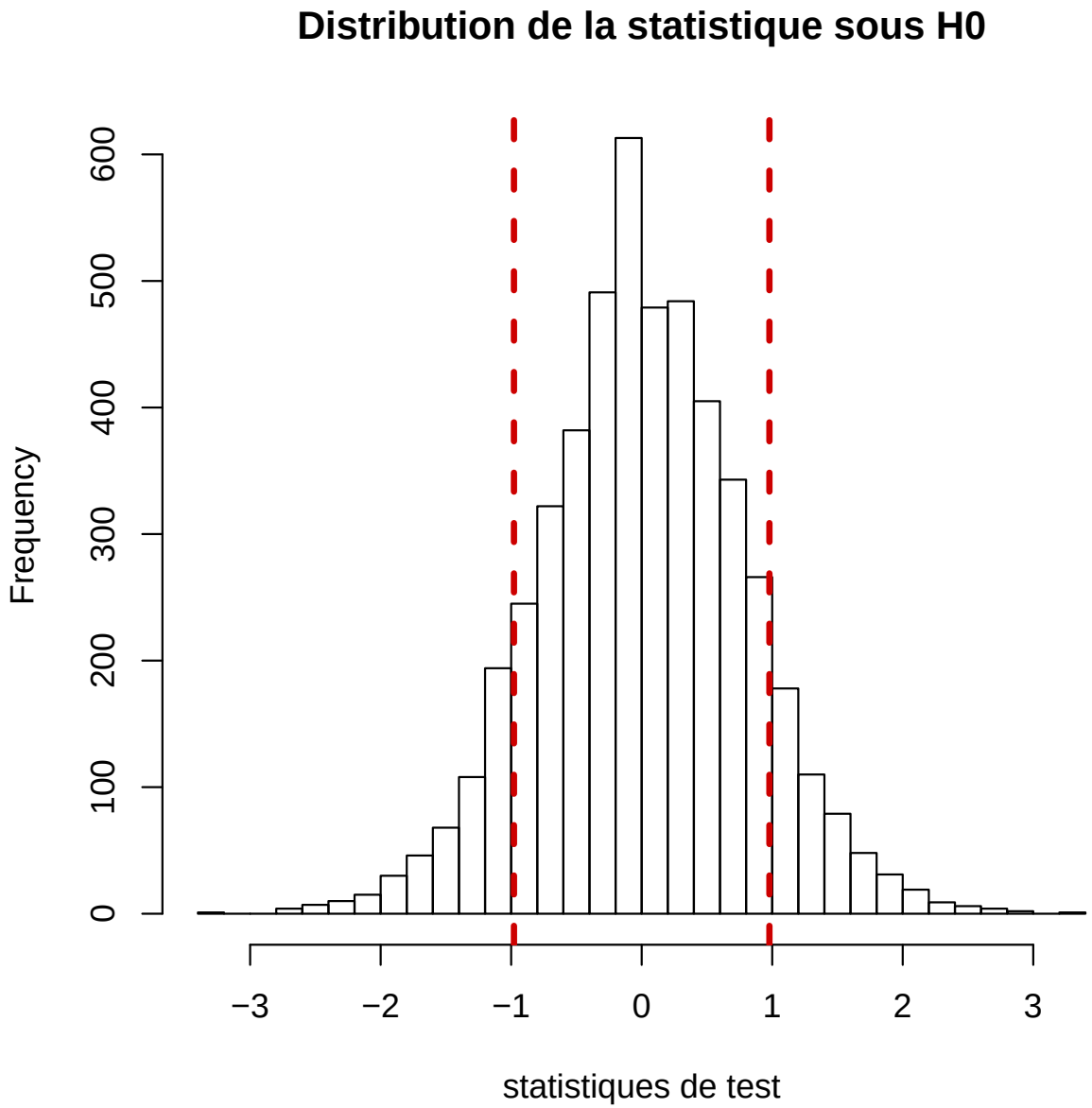
Exemple

```
> sample(Groupes)
 [1] "A" "A" "A" "A" "B" "B" "B" "A" "B" "B"
> sample(Groupes)
 [1] "A" "B" "A" "B" "A" "A" "B" "A" "B" "B"
> t.test(Valeurs[sample(Groupes)=="A"], Valeurs[sample(Groupes)=="B"])$statistic
        t
0.5791038
> t.test(Valeurs[sample(Groupes)=="A"], Valeurs[sample(Groupes)=="B"])$statistic
        t
0.6356417
```

Exemple

```
> B<-5000
> statistics <- rep(NA, B)
> for(b in 1:B)
+ {
+ statistics[b] <- t.test(Valeurs[sample(Groupes)=="A"],
+  Valeurs[sample(Groupes)=="B"])$statistic
+ }
> hist(statistics, nclass=30, xlab="statistiques de test", main="Distribution de la
> stat.ini <- t.test(Valeurs[Groupes=="A"], Valeurs[Groupes=="B"])$statistic
> abline(v=stat.ini, col="red3", lwd=3, lty=2)
> abline(v=-1*stat.ini, col="red3", lwd=3, lty=2)
```

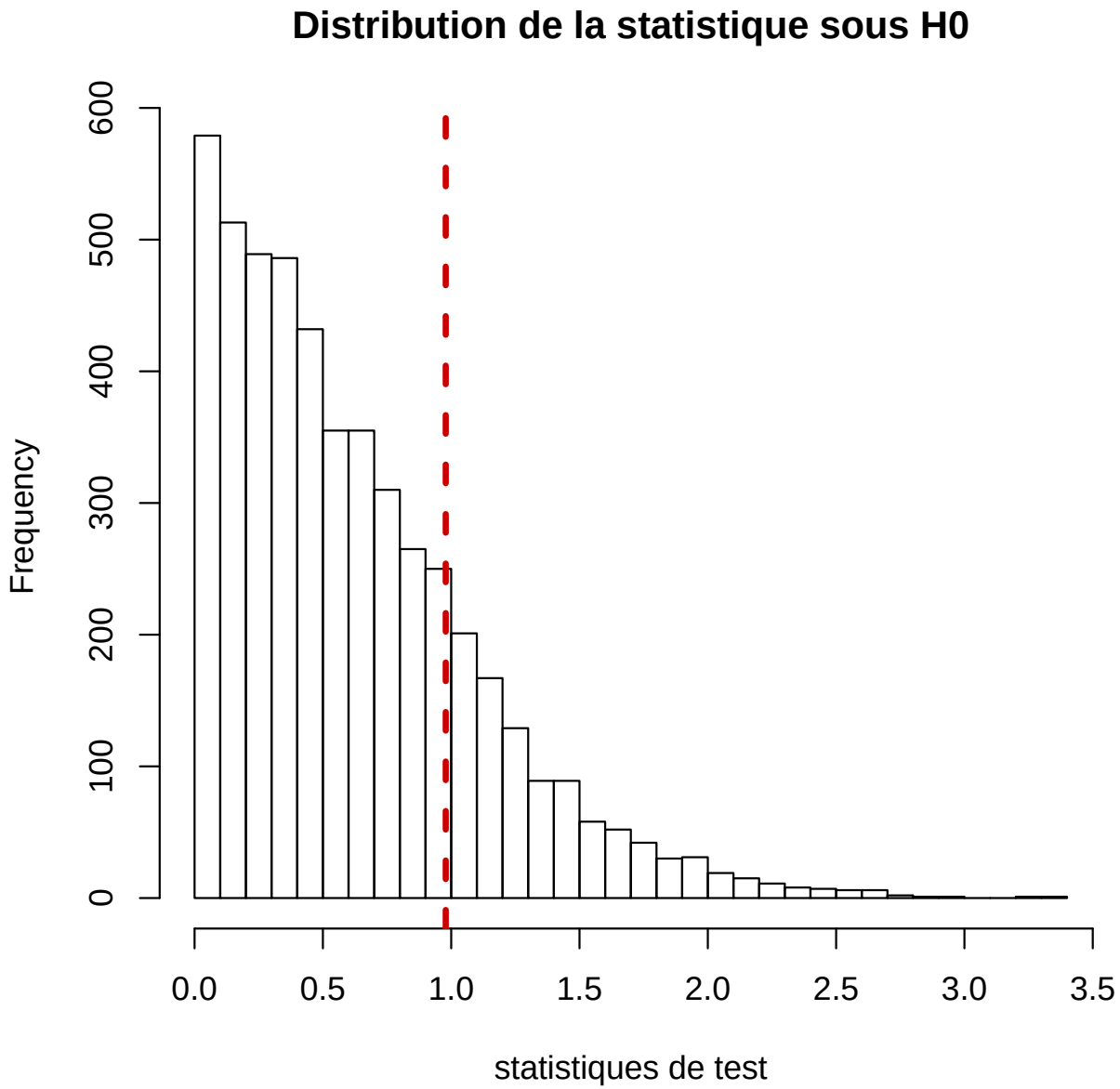
Exemple


Distribution de la statistique sous H0
Frequency
0
100
200
300
400
500
600
−3
−2
−1
0
1
2
3
statistiques de test

Un exemple pour mieux comprendre...

```
> hist(abs(statistics), nclass=30, xlab="statistiques de test", main="Distribution
> abline(v=abs(stat.ini), col="red3", lwd=3, lty=2)
```

Exemple


Distribution de la statistique sous H0
Frequency
0
100
200
300
400
500
600
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
statistiques de test

Un exemple pour mieux comprendre...

```
> sum(abs(statistics)>=abs(stat.ini))
[1] 1008
> sum(abs(statistics)>=abs(stat.ini))/B
[1] 0.2016
> t.test(Valeurs[Groupes=="A"], Valeurs[Groupes=="B"])$p.value
[1] 0.3601715
```

$$p_j^* = \frac{\sum_{b=1}^{B} I(|t_j^{(b)}| \geq |t_j|)}{B} = 0.2016$$

$$p_j = 0.3602 \text{ (sans permutation)}$$

Plan

Correction de Bonferroni

- La méthode de Bonferroni est la plus connue.
- Pour le gène j, soit $\tilde{p}_j$ la probabilité corrigée associée à p_j^*.

$$\tilde{p}_j = min(kp_j^*, 1)$$

Exemple pour le $j$ème gène :

- $p_j^* = 0.0023$ et $k = 1000$ gènes.

$\rightarrow$ $\tilde{p}_j = min(1000 * 0.0023, 1) = min(2.3, 1) = 1.$

Problème :

1. Méthode très conservative.
2. Faible puissance due au non rejet quasi-systématique de H_0.

Procédure de Holm

- Procédure de Holm moins conservative
- Idée :
  - Ordonner les gènes selon les probabilités critiques.
  - A chaque fois qu'un test est significatif, le gène suivant est inclus.
  - La probabilité critique du gène inclus est corrigée selon le nombre de gènes restant à inclure.
- Posons $p_{r1} \leq p_{r2} \leq ... \leq p_{rk}$, les probabilités critiques ordonnées en utilisant les valeurs obtenues par permutation $(p_j^*, j = 1, 2, ..., k)$:

$$\tilde{p}_{r1} = kp_{r1}$$

$$\tilde{p}_{rj} = max(\tilde{p}_{r(j-1)}, (k - j + 1)p_{rj}) \text{ pour } 2 \leq j \leq k$$

- Les probabilités critiques supérieures à 1 sont corrigées à 1.

Exemple

Gènes	1	2	3	4	5	6	7
p_j^*	0.122	0.001	0.523	0.013	0.029	0.987	0.342

$\Downarrow$

Gènes	2	4	5	1	7	3	6
p_j^*	0.001	0.013	0.029	0.122	0.342	0.523	0.987
Ordre	$r1$	$r2$	$r3$	$r4$	$r5$	$r6$	$r7$
$\tilde{p}_{rj}$							

Exemple

Gènes	1	2	3	4	5	6	7
p_j^*	0.122	0.001	0.523	0.013	0.029	0.987	0.342

$\Downarrow$

Gènes	2	4	5	1	7	3	6
p_j^*	0.001	0.013	0.029	0.122	0.342	0.523	0.987
Ordre	$r1$	$r2$	$r3$	$r4$	$r5$	$r6$	$r7$
$\tilde{p}_{rj}$	0.007						

Calcul de la probabilité critique corrigée

$\tilde{p}_{r1} = kp_{r1} = 7 \times 0.001 = 0.007$

Exemple

Gènes	1	2	3	4	5	6	7
p_j^*	0.122	0.001	0.523	0.013	0.029	0.987	0.342

$\Downarrow$

Gènes	2	4	5	1	7	3	6
p_j^*	0.001	0.013	0.029	0.122	0.342	0.523	0.987
Ordre	$r1$	$r2$	$r3$	$r4$	$r5$	$r6$	$r7$
$\tilde{p}_{rj}$	0.007	0.078					

Calcul de la probabilité critique corrigée

$\tilde{p}_{r1} = kp_{r1} = 7 \times 0.001 = 0.007$

$\tilde{p}_{r2} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.007, (7-2+1) \times 0.013) = 0.078$

Exemple

Gènes	1	2	3	4	5	6	7
p_j^*	0.122	0.001	0.523	0.013	0.029	0.987	0.342

$\Downarrow$

Gènes	2	4	5	1	7	3	6
p_j^*	0.001	0.013	0.029	0.122	0.342	0.523	0.987
Ordre	$r1$	$r2$	$r3$	$r4$	$r5$	$r6$	$r7$
$\tilde{p}_{rj}$	0.007	0.078	0.145				

Calcul de la probabilité critique corrigée

$\tilde{p}_{r1} = kp_{r1} = 7 \times 0.001 = 0.007$

$\tilde{p}_{r2} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.007, (7-2+1) \times 0.013) = 0.078$

$\tilde{p}_{r3} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.078, (7-3+1) \times 0.029) = 0.145$

Exemple

Gènes	1	2	3	4	5	6	7
p_j^*	0.122	0.001	0.523	0.013	0.029	0.987	0.342

$\Downarrow$

Gènes	2	4	5	1	7	3	6
p_j^*	0.001	0.013	0.029	0.122	0.342	0.523	0.987
Ordre	$r1$	$r2$	$r3$	$r4$	$r5$	$r6$	$r7$
$\tilde{p}_{rj}$	0.007	0.078	0.145	0.488			

Calcul de la probabilité critique corrigée

$$\tilde{p}_{r1} = kp_{r1} = 7 \times 0.001 = 0.007$$
$$\tilde{p}_{r2} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.007, (7-2+1) \times 0.013) = 0.078$$
$$\tilde{p}_{r3} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.078, (7-3+1) \times 0.029) = 0.145$$
$$\tilde{p}_{r4} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.145, (7-4+1) \times 0.122) = 0.488$$

Exemple

Gènes	1	2	3	4	5	6	7
p_j^*	0.122	0.001	0.523	0.013	0.029	0.987	0.342

$\Downarrow$

Gènes	2	4	5	1	7	3	6
p_j^*	0.001	0.013	0.029	0.122	0.342	0.523	0.987
Ordre	$r1$	$r2$	$r3$	$r4$	$r5$	$r6$	$r7$
$\tilde{p}_{rj}$	0.007	0.078	0.145	0.488	1.000	1.000	1.000

Calcul de la probabilité critique corrigée

$$\tilde{p}_{r1} = kp_{r1} = 7 \times 0.001 = 0.007$$
$$\tilde{p}_{r2} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.007, (7-2+1) \times 0.013) = 0.078$$
$$\tilde{p}_{r3} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.078, (7-3+1) \times 0.029) = 0.145$$
$$\tilde{p}_{r4} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.145, (7-4+1) \times 0.122) = 0.488$$
$$\tilde{p}_{r5} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) = max(0.488, (7-5+1) \times 0.342) = 1.026$$

Remarque : l'intérêt d'une filtration préalable

- Objectif : limiter le nombre de gènes candidats.

$$\tilde{p}_{r1} = kp_{r1}$$

$$\tilde{p}_{rj} = max(\tilde{p}_{r(j-1)}, (k-j+1)p_{rj}) \text{ pour } 2 \leq j \leq k$$

- La correction diminue quand k diminue.
- Méthodes couramment rencontrées :
  - Elimination des gènes ayant de trop faible ou de trop fort coefficient de variation.
  - Elimination des gènes qui ont trop d'individus au-dessus d'un certain seuil.

Plan

Les données †

- 128 patients avec un diagnostic de leucémie aiguë lymphoblastique (ALL).
- Des puces contenant l'expression de 12625 gènes ont été collectés.
- Objectif de l'exemple : Identifier les gènes qui sont associés à la réponse thérapeutique des patients.

†. S. Chiaretti et al. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood, 2004, Vol. 103, No. 7

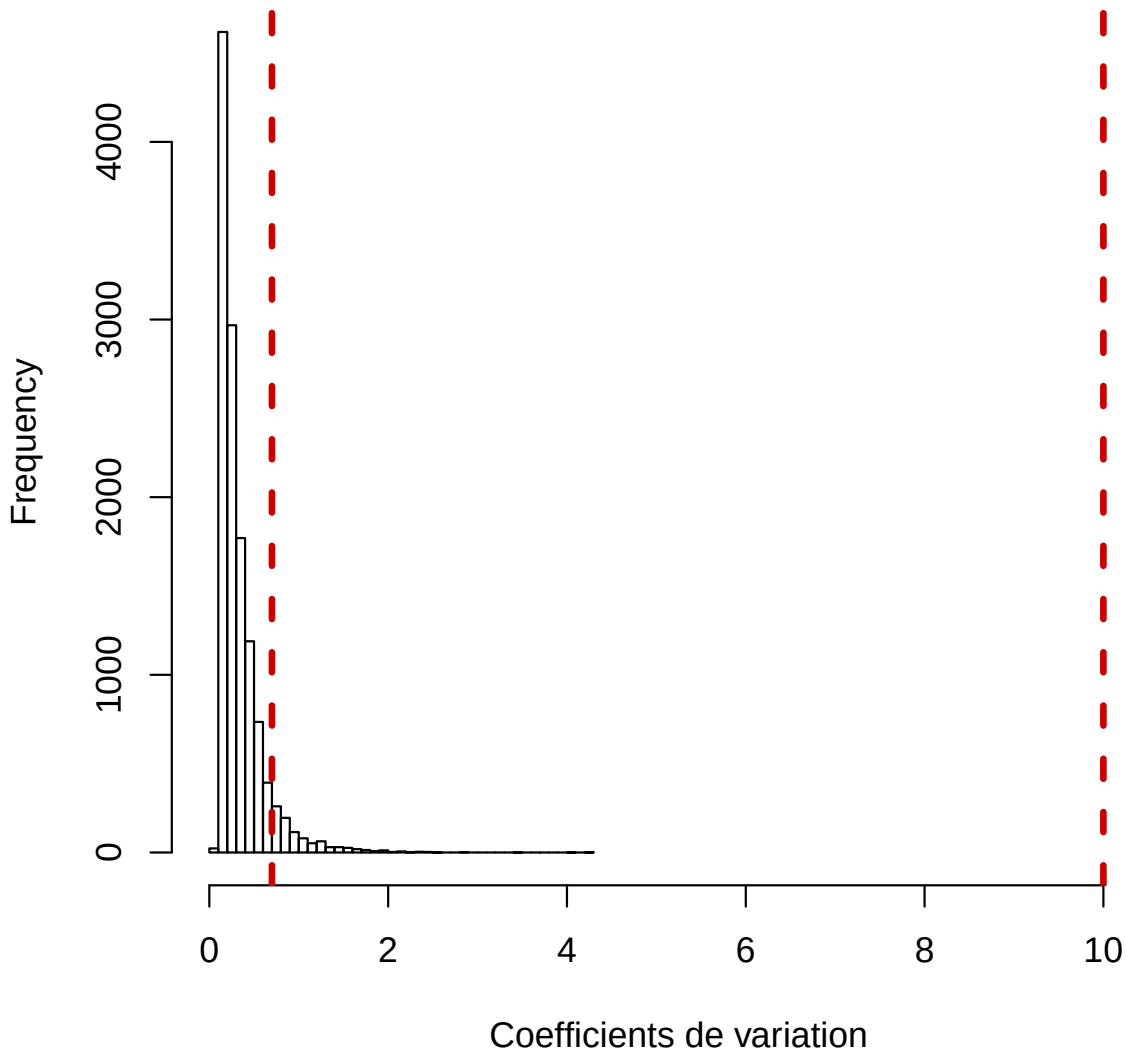
Les analyses sous R

```
> library(Biobase)
> library(multtest)
> library(ALL)
> library(genefilter)
> data(ALL)
> dim(exprs(ALL))
```

```
[1] 12625   128
```

```
> X<-exprs(ALL)
> pheno<-pData(ALL)
> coef.var<-function(x) {sd(x)/abs(mean(x))}
> coef.var.vector<-apply(2^X, 1, coef.var)
> hist(coef.var.vector, nclass=30, xlab="Coefficients de variation",
+  main="", xlim=c(0, 10.5))
> abline(v=0.7, col="red3", lwd=3, lty=2)
> abline(v=10, col="red3", lwd=3, lty=2)
```

Les analyses sous R



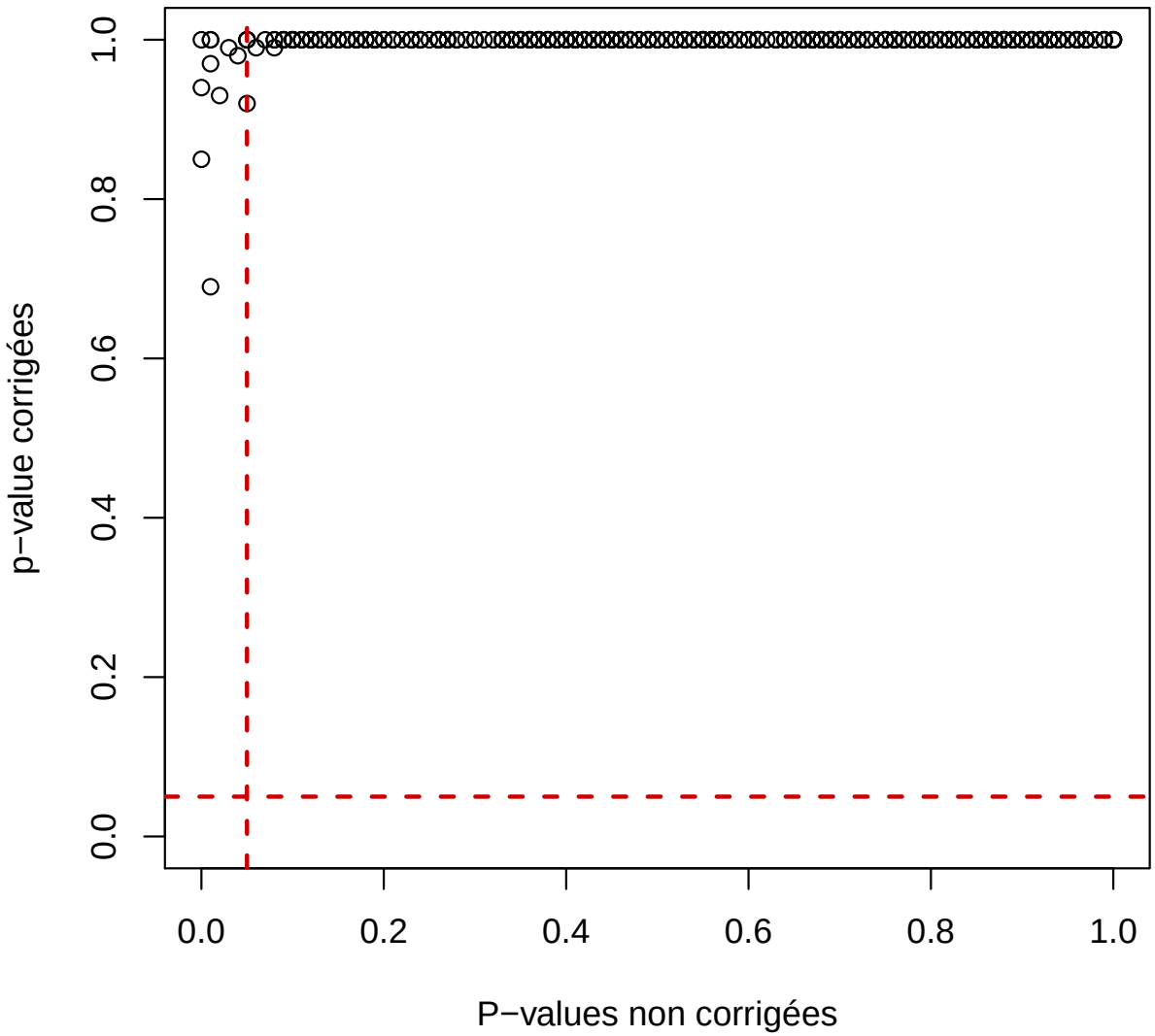
Les analyses sous R

```
> sum(coef.var.vector<=0.7)/length(coef.var.vector)

[1] 0.926495

> sum(coef.var.vector>10)/length(coef.var.vector)

[1] 0

> ffun<-filterfun(pOverA(p=0.2, A=100), cv(a=0.7, b=10))
> filt<-genefilter(2^X, ffun)
> sum(filt)

[1] 431

> mb<-as.character(pheno$mdr)
> table(mb)

mb
NEG POS
101  24

> filtX<-X[filt,!is.na(mb)]
> dim(filtX)

[1] 431 125

> mb <- mb[!is.na(mb)]
> length(mb)

[1] 125
```

Les analyses sous R

```
> mb.boot<-MTP(X = filtX, Y=mb, test = "t.twosamp.equalvar", robust=TRUE, alpha=c(0

running bootstrap...
iteration = 100

> sum(mb.boot@rawp<=0.05)/sum(filt)

[1] 0.0324826

> sum(mb.boot@adjp<=0.05)/sum(filt)

[1] 0

> plot(mb.boot@rawp, mb.boot@adjp, ylim=c(0,1), xlim=c(0,1), xlab="P-values non cor
> abline(v=0.05, col="red3", lty=2, lwd=2)
> abline(h=0.05, col="red3", lty=2, lwd=2)
```

Les analyses sous R



Plan

Limites

- Ne prend pas en compte la structure de dépendance des gènes
  - → Algorithme de Westfall et Young. ‡
- Même après permutations et corrections, les probabilités critiques sont peu fiables
  - → Cette méthode permet de retrouver des résultats plus réalistes et de classer les gènes en fonction de leur potentiel intérêt.
- Survol de la méthode MTP : nombreuses autres possibilités.
- Après classement/sélection des gènes par MTP :
  1. Calcul d'une taille d'échantillon minimale nécessaire pour une validation externe
  2. Méthodes de mesure spécifique des expressions des caractéristiques (PCR, etc.)
  3. On retrouve une situation acceptable pour l'utilisation des méthodes statistiques :

$$\text{Nombre variables} <<< \text{Nombre de sujets}$$

‡. Resampling-based multiple testing. Peter H. Westfall, S. Stanley Young, Wiley, New York, 1993