



# Recherche de gènes différentiellement exprimés

David Causeur

*Laboratoire de Mathématiques Appliquées*

*Agrocampus Rennes*

*IRMAR CNRS UMR 6625*

*<http://www.agrocampus-ouest.fr/math/causeur/>*



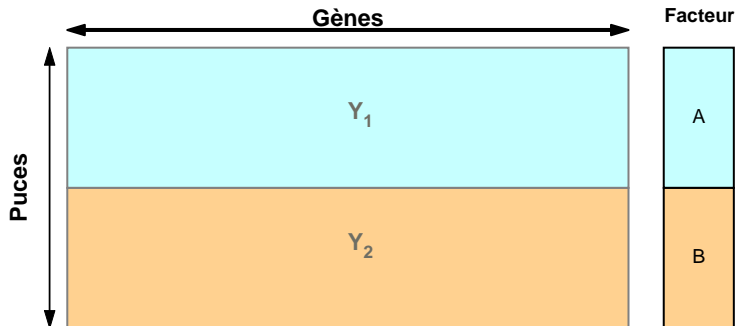
# Plan du cours

- 1 Objectifs
- 2 Tests gène par gène
  - Comparaison de deux groupes
  - Tests d'un effet
- 3 Tests multiples
  - Stratégie générale
  - Approche de Bonferroni
  - Approche de Benjamini & Hochberg
  - Étude comparative
- 4 Perspectives



## Analyse différentielle

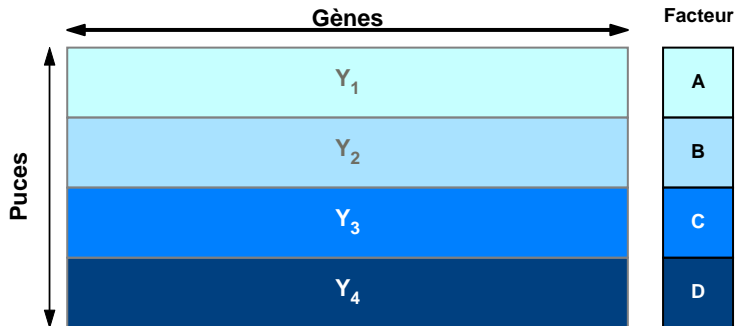
Objectif : identifier les gènes s'exprimant différemment selon les modalités d'un facteur [ou les valeurs d'une covariable]





## Analyse différentielle

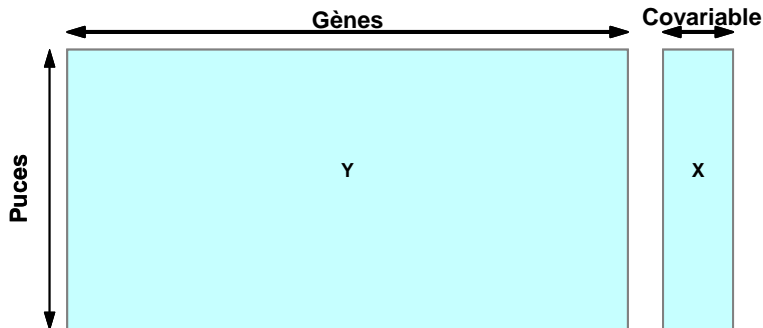
Objectif : identifier les gènes s'exprimant différemment selon les modalités d'un facteur [ou les valeurs d'une covariable]





## Analyse différentielle

Objectif : identifier les gènes s'exprimant différemment selon les modalités d'un facteur [ou les valeurs d'une covariable]





# Tests multiples

## Stratégie

- Un test par gène
  - Choix du test



# Tests multiples

## Stratégie

- Un test par gène
  - Choix du test
- Un contrôle du risque d'erreur
  - Choix d'un risque d'erreur
  - Choix de la règle de décision minimisant le risque



# Plan du cours

- 1 Objectifs
- 2 Tests gène par gène
  - Comparaison de deux groupes
  - Tests d'un effet
- 3 Tests multiples
  - Stratégie générale
  - Approche de Bonferroni
  - Approche de Benjamini & Hochberg
  - Étude comparative
- 4 Perspectives



## Comparaison de deux groupes

Test d'hypothèse :  $Y$  expression d'un gène

$$\left\{ \begin{array}{l} \text{Pour le 1er génotype} \quad : \quad \mathbb{E}(Y) = \mu_1 ; \text{Écart-type}(Y) = \sigma \\ \text{Pour le 2ème génotype} \quad : \quad \mathbb{E}(Y) = \mu_2 ; \text{Écart-type}(Y) = \sigma \end{array} \right.$$

$$\left\{ \begin{array}{l} H_0 \quad : \quad \mu_1 = \mu_2 \text{ [gène non différentiellement exprimé]} \\ H_1 \quad : \quad \mu_1 \neq \mu_2 \text{ [gène différentiellement exprimé]} \end{array} \right.$$



## Comparaison de deux groupes

Test d'hypothèse :  $Y$  expression d'un gène

$$\begin{cases} \text{Pour le 1er génotype} & : \mathbb{E}(Y) = \mu_1 ; \text{Écart-type}(Y) = \sigma \\ \text{Pour le 2ème génotype} & : \mathbb{E}(Y) = \mu_2 ; \text{Écart-type}(Y) = \sigma \end{cases}$$

$$\begin{cases} H_0 & : \mu_1 = \mu_2 \text{ [gène non différentiellement exprimé]} \\ H_1 & : \mu_1 \neq \mu_2 \text{ [gène différentiellement exprimé]} \end{cases}$$

Stratégie de test :



## Comparaison de deux groupes

Test d'hypothèse :  $Y$  expression d'un gène

$$\left\{ \begin{array}{l} \text{Pour le 1er génotype} \quad : \quad \mathbb{E}(Y) = \mu_1 ; \text{Écart-type}(Y) = \sigma \\ \text{Pour le 2ème génotype} \quad : \quad \mathbb{E}(Y) = \mu_2 ; \text{Écart-type}(Y) = \sigma \end{array} \right.$$

$$\left\{ \begin{array}{l} H_0 \quad : \quad \mu_1 = \mu_2 \text{ [gène non différentiellement exprimé]} \\ H_1 \quad : \quad \mu_1 \neq \mu_2 \text{ [gène différentiellement exprimé]} \end{array} \right.$$

Stratégie de test :

- Calcul de  $T$ , statistique de Student

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad S : \sigma \text{ estimé}$$



## Comparaison de deux groupes

Test d'hypothèse :  $Y$  expression d'un gène

$$\begin{cases} \text{Pour le 1er génotype} & : \mathbb{E}(Y) = \mu_1 ; \text{Écart-type}(Y) = \sigma \\ \text{Pour le 2ème génotype} & : \mathbb{E}(Y) = \mu_2 ; \text{Écart-type}(Y) = \sigma \end{cases}$$

$$\begin{cases} H_0 & : \mu_1 = \mu_2 \text{ [gène non différentiellement exprimé]} \\ H_1 & : \mu_1 \neq \mu_2 \text{ [gène différentiellement exprimé]} \end{cases}$$

Stratégie de test :

- Calcul de  $T$ , statistique de Student

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad S : \sigma \text{ estimé}$$

- Si  $|T|$  est grand, on rejette  $H_0$



## Risques d'erreurs

### Test d'hypothèse

Vérité	Décision	
	Négatif	Positif
Non DE	Vrai Négatif	Faux Positif
DE	Faux négatif	Vrai positif



## Risques d'erreurs

### Test d'hypothèse

Vérité	Décision	
	Négatif	Positif
Non DE	Vrai Négatif	Faux Positif
DE	Faux négatif	Vrai positif

- Risque que le gène soit faux positif :

Probabilité critique  $p$



## Risques d'erreurs

### Test d'hypothèse

Vérité	Décision	
	Négatif	Positif
Non DE	Vrai Négatif	Faux Positif
DE	Faux négatif	Vrai positif

- Risque que le gène soit faux positif :

Probabilité critique  $p$

- Règle de décision :

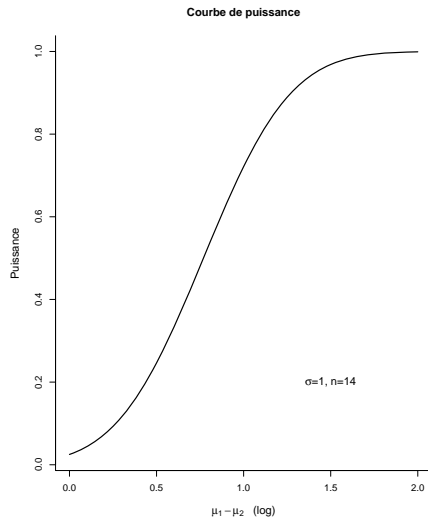
Si  $p \leq 0.05$ , alors le gène est positif

- Puissance de la règle de décision :

Probabilité qu'un gène DE soit positif

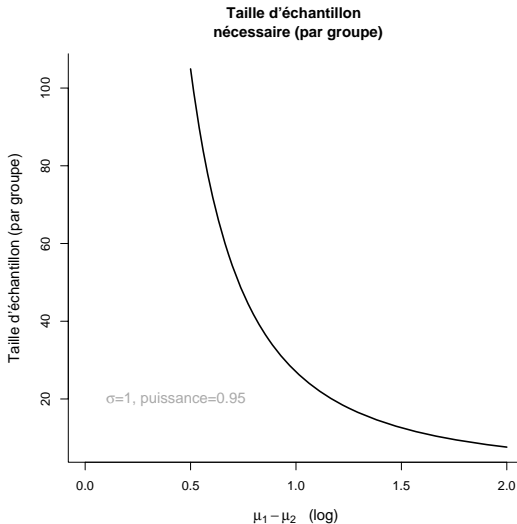


# Puissance et taille d'échantillon





# Puissance et taille d'échantillon





## Extensions

Intégration de connaissance : tests unilatéraux

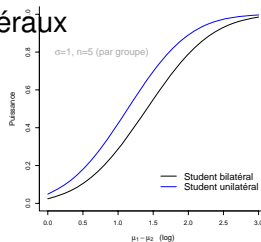
$$\begin{cases} H_0 & : \mu_1 = \mu_2 \text{ [non DE]} \\ H_1 & : \mu_1 > \mu_2 \text{ [DE]} \end{cases}$$



## Extensions

Intégration de connaissance : tests unilatéraux

$$\begin{cases} H_0 : \mu_1 = \mu_2 \text{ [non DE]} \\ H_1 : \mu_1 > \mu_2 \text{ [DE]} \end{cases}$$

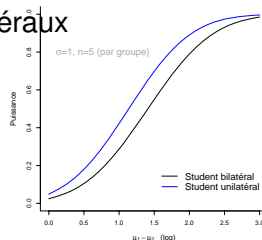




## Extensions

Intégration de connaissance : tests unilatéraux

$$\begin{cases} H_0 : \mu_1 = \mu_2 \text{ [non DE]} \\ H_1 : \mu_1 > \mu_2 \text{ [DE]} \end{cases}$$



Affaiblissement des hypothèses

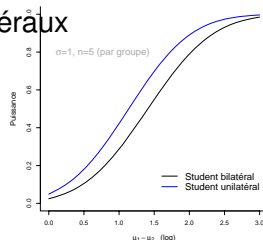
- Écart-types intra-groupes différents : test de Welch



## Extensions

Intégration de connaissance : tests unilatéraux

$$\begin{cases} H_0 : \mu_1 = \mu_2 \text{ [non DE]} \\ H_1 : \mu_1 > \mu_2 \text{ [DE]} \end{cases}$$



Affaiblissement des hypothèses

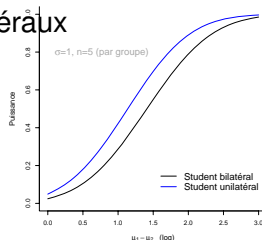
- Écart-types intra-groupes différents : test de Welch
- Distributions intra-groupes non-normales : test de Wilcoxon, ...



## Extensions

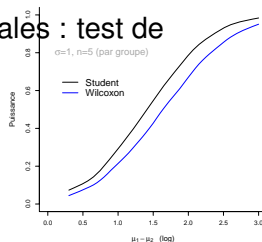
Intégration de connaissance : tests unilatéraux

$$\begin{cases} H_0 : \mu_1 = \mu_2 \text{ [non DE]} \\ H_1 : \mu_1 > \mu_2 \text{ [DE]} \end{cases}$$



Affaiblissement des hypothèses

- Écart-types intra-groupes différents : test de Welch
- Distributions intra-groupes non-normales : test de Wilcoxon, ...





## Tests d'un effet

Comparaison de plusieurs groupes

- analyse de la variance

$$\begin{cases} H_0 & : \mu_1 = \mu_2 = \mu_3 \text{ [non DE]} \\ H_1 & : \mu_1 \neq \mu_2 \text{ ou } \mu_1 \neq \mu_3 \text{ ou } \mu_2 \neq \mu_3 \text{ [DE]} \end{cases}$$



## Tests d'un effet

### Comparaison de plusieurs groupes

- analyse de la variance

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 \text{ [non DE]} \\ H_1 : \mu_1 \neq \mu_2 \text{ ou } \mu_1 \neq \mu_3 \text{ ou } \mu_2 \neq \mu_3 \text{ [DE]} \end{cases}$$

- Test d'un contraste

$$\begin{cases} H_0 : \mu_3 = \frac{\mu_1 + \mu_2}{2} \text{ [non DE]} \\ H_1 : \mu_3 \neq \frac{\mu_1 + \mu_2}{2} \text{ [DE]} \end{cases}$$



## Tests d'un effet

### Comparaison de plusieurs groupes

- analyse de la variance

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 \text{ [non DE]} \\ H_1 : \mu_1 \neq \mu_2 \text{ ou } \mu_1 \neq \mu_3 \text{ ou } \mu_2 \neq \mu_3 \text{ [DE]} \end{cases}$$

- Test d'un contraste

$$\begin{cases} H_0 : \mu_3 = \frac{\mu_1 + \mu_2}{2} \text{ [non DE]} \\ H_1 : \mu_3 \neq \frac{\mu_1 + \mu_2}{2} \text{ [DE]} \end{cases}$$

### Effet de covariables $x$ : régression

$$\begin{cases} H_0 : x \text{ ne modifie pas l'expression génique [non DE]} \\ H_1 : x \text{ modifie l'expression génique [DE]} \end{cases}$$



# Plan du cours

- 1 Objectifs
- 2 Tests gène par gène
  - Comparaison de deux groupes
  - Tests d'un effet
- 3 Tests multiples
  - Stratégie générale
  - Approche de Bonferroni
  - Approche de Benjamini & Hochberg
  - Étude comparative
- 4 Perspectives



## Stratégie générale

De manière générale

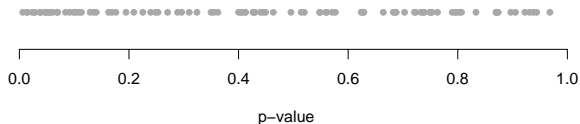
- Pour le gène  $i$ , risque d'être faux positif  $p_i$



## Stratégie générale

De manière générale

- Pour le gène  $i$ , risque d'être faux positif  $p_i$
- Classement des gènes par risque croissant

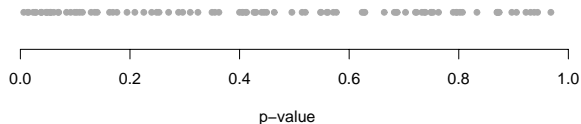




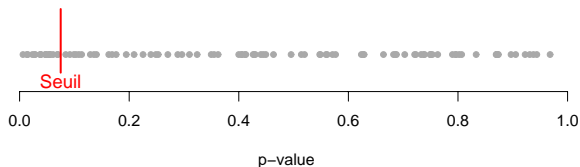
## Stratégie générale

De manière générale

- Pour le gène  $i$ , risque d'être faux positif  $p_i$
- Classement des gènes par risque croissant



- Choix d'un seuil de décision

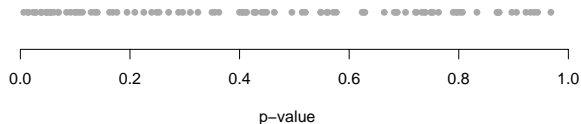




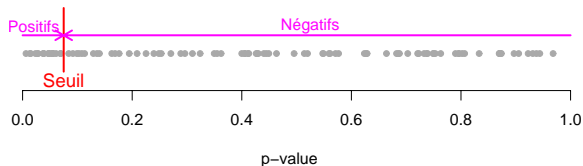
## Stratégie générale

De manière générale

- Pour le gène  $i$ , risque d'être faux positif  $p_i$
- Classement des gènes par risque croissant



- Choix d'un seuil de décision





## Risques d'erreurs

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$



## Risques d'erreurs

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

Objectif : contrôle du risque d'erreur (de 1ère espèce)

- Risque qu'il y ait au moins un faux positif (Bonferroni)

$$\text{FWER}_t = \mathbb{P}(V_t > 0) \quad [t \text{ tel que } \text{FWER}_t \leq \alpha]$$



## Risques d'erreurs

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

Objectif : contrôle du risque d'erreur (de 1ère espèce)

- Risque qu'il y ait au moins un faux positif (**Bonferroni**)

$$\text{FWER}_t = \mathbb{P}(V_t > 0) \quad [t \text{ tel que } \text{FWER}_t \leq \alpha]$$

- Taux de faux positifs (**Benjamini & Hochberg**)

$$\text{FDR}_t = \mathbb{E} \left[ \frac{V_t}{R_t} \right] \quad [t \text{ tel que } \text{FDR}_t \leq \alpha]$$



## Approche naïve : seuil = $\alpha$

Pour chaque choix d'un seuil  $t$

Vérité	Décision		Total
	Négatif	Positif	
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

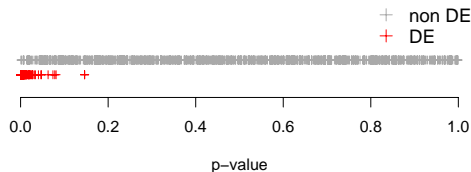


## Approche naïve : seuil = $\alpha$

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

Exemple artificiel (1000 gènes dont 100 DE)



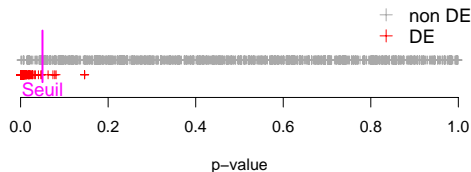


## Approche naïve : seuil = $\alpha$

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

Exemple artificiel (1000 gènes dont 100 DE)



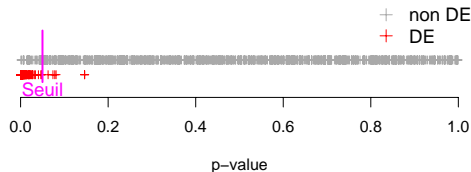


## Approche naïve : seuil = $\alpha$

Pour  $t = 0.05$

Vérité	Décision			
	Négatif	Positif	Total	
Non DE	839	61	900	FDR = $\frac{61}{161} = 0.38$ puissance = $\frac{96}{100} = 0.96$
DE	4	96	100	
Total	843	161	1000	

Exemple artificiel (1000 gènes dont 100 DE)





## Approche de Bonferroni

Sur 3 gènes  $G_1$ ,  $G_2$ ,  $G_3$

- Pour chaque gène  $i$ , risque d'être faux positif  $p_i$



## Approche de Bonferroni

Sur 3 gènes  $G_1$ ,  $G_2$ ,  $G_3$

- Pour chaque gène  $i$ , risque d'être faux positif  $p_i$
- Risque qu'il y ait au moins un faux positif :

$$\begin{aligned}\mathbb{P}(V_t > 0) &= \mathbb{P}([G_1 \text{ est FP}] \text{ ou } [G_2 \text{ est FP}] \text{ ou } [G_3 \text{ est FP}]) \\ &= \mathbb{P}(G_1 \text{ est FP}) + \mathbb{P}(G_2 \text{ est FP}) + \mathbb{P}(G_3 \text{ est FP}) \\ &= p_1 + p_2 + p_3\end{aligned}$$



## Approche de Bonferroni

Sur 3 gènes  $G_1$ ,  $G_2$ ,  $G_3$

- Pour chaque gène  $i$ , risque d'être faux positif  $p_i$
- Risque qu'il y ait au moins un faux positif :

$$\begin{aligned}\mathbb{P}(V_t > 0) &= \mathbb{P}([G_1 \text{ est FP}] \text{ ou } [G_2 \text{ est FP}] \text{ ou } [G_3 \text{ est FP}]) \\ &= \mathbb{P}(G_1 \text{ est FP}) + \mathbb{P}(G_2 \text{ est FP}) + \mathbb{P}(G_3 \text{ est FP}) \\ &= p_1 + p_2 + p_3\end{aligned}$$

- Seuil :  $\alpha/3 \Rightarrow \mathbb{P}(V_t > 0) \leq \alpha$



## Propriétés de la méthode de Bonferroni

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

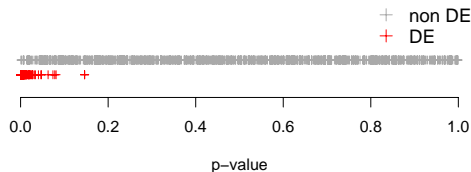


## Propriétés de la méthode de Bonferroni

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

Exemple artificiel (1000 gènes dont 100 DE)



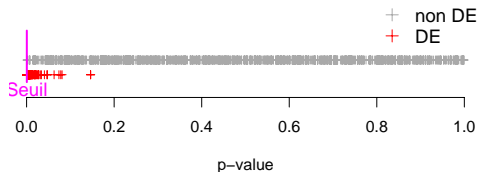


## Propriétés de la méthode de Bonferroni

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

Exemple artificiel (1000 gènes dont 100 DE)



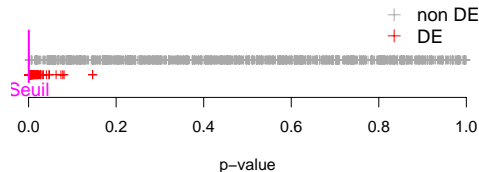


## Propriétés de la méthode de Bonferroni

Pour  $t = 0.0005$

Vérité	Décision			
	Négatif	Positif	Total	
Non DE	900	0	900	$\text{FDR} = \frac{0}{23} = 0 !$ $\text{puissance} = \frac{23}{100} = 0.23$
DE	77	23	100	
Total	977	23	1000	

Exemple artificiel (1000 gènes dont 100 DE)





## Approche de Benjamini & Hochberg

Contrôle du taux de faux positifs

- Pour chaque gène  $i$ , risque d'être faux positif  $p_i$

$$p_1 < p_2 < p_3 < \dots$$



## Approche de Benjamini & Hochberg

### Contrôle du taux de faux positifs

- Pour chaque gène  $i$ , risque d'être faux positif  $p_i$

$$p_1 < p_2 < p_3 < \dots$$

- Si  $t = p_k$  [ les  $k$  premiers gènes sont positifs ] :
  - Estimation du nombre de faux positifs :  $m_0 p_k$  [  $\approx m p_k$  ]
  - Estimation du FDR :  $\widehat{FDR} = \frac{m_0 p_k}{k}$  [  $\approx \frac{m p_k}{k}$  ]



## Approche de Benjamini & Hochberg

### Contrôle du taux de faux positifs

- Pour chaque gène  $i$ , risque d'être faux positif  $p_i$

$$p_1 < p_2 < p_3 < \dots$$

- Si  $t = p_k$  [ les  $k$  premiers gènes sont positifs ] :
  - Estimation du nombre de faux positifs :  $m_0 p_k$  [  $\approx m p_k$  ]
  - Estimation du FDR :  $\widehat{FDR} = \frac{m_0 p_k}{k}$  [  $\approx \frac{m p_k}{k}$  ]
- Seuil :  $p_k$  tel que  $\frac{m p_k}{k} \leq \alpha$



# Propriétés de la méthode de Benjamini & Hochberg

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

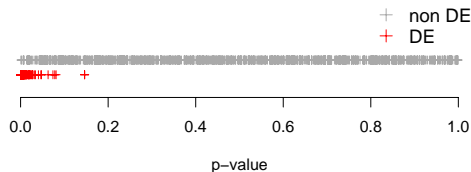


# Propriétés de la méthode de Benjamini & Hochberg

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

Exemple artificiel (1000 gènes dont 100 DE)



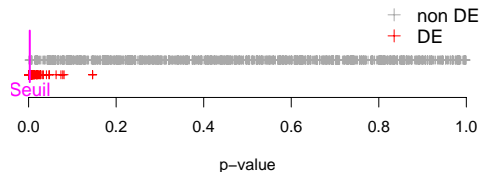


# Propriétés de la méthode de Benjamini & Hochberg

Pour chaque choix d'un seuil  $t$

Vérité	Décision		
	Négatif	Positif	Total
Non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

Exemple artificiel (1000 gènes dont 100 DE)



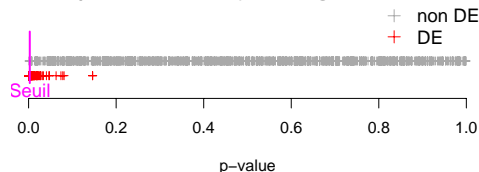


# Propriétés de la méthode de Benjamini & Hochberg

Pour  $t = 0.0033$

Vérité	Décision			
	Négatif	Positif	Total	
Non DE	898	2	900	$\text{FDR} = \frac{2}{61} = 0.03$ $\text{puissance} = \frac{59}{100} = 0.59$
DE	41	59	100	
Total	939	61	1000	

Exemple artificiel (1000 gènes dont 100 DE)





## Propriétés comparées

Contrôle du risque	Tests par gènes			
	Student		Wilcoxon	
	FDR	Puissance	FDR	Puissance
Naïf	0.38	0.96	0.33	0.87
Bonferroni	0.00	0.23	0.00	0.00
Benjamini & Hochberg	0.03	0.59	0.00	0.00



# Plan du cours

- 1 Objectifs
- 2 Tests gène par gène
  - Comparaison de deux groupes
  - Tests d'un effet
- 3 Tests multiples
  - Stratégie générale
  - Approche de Bonferroni
  - Approche de Benjamini & Hochberg
  - Étude comparative
- 4 Perspectives



# Perspectives

## Stratégie de tests

- choix du test spécifique aux gènes
- choix de la stratégie de contrôle du risque



## Perspectives

### Stratégie de tests

- choix du test spécifique aux gènes
- choix de la stratégie de contrôle du risque

### Tout est en place pour

- une analyse biologique confirmatoire [ mise en relation avec ontologie ]
- construction d'une typologie des gènes ou des individus
- construction d'une règle de diagnostic