



Recherche de gènes différentiellement exprimés

David Causeur

Laboratoire de Mathématiques Appliquées

Agrocampus Ouest

IRMAR CNRS UMR 6625

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

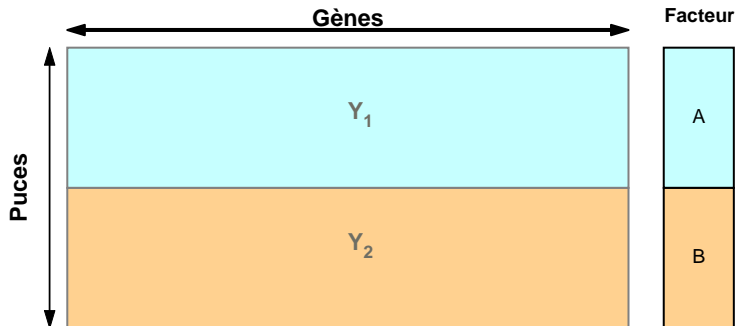
Plan du cours

- 1 Objectifs
- 2 Tests gène par gène
 - Test de la relation entre expression et covariable
 - Risques d'erreurs
- 3 Tests multiples
 - Stratégie générale
 - Contrôle du FWER
 - Contrôle du FDR
 - Optimisation des procédures
- 4 Perspectives



Analyse différentielle

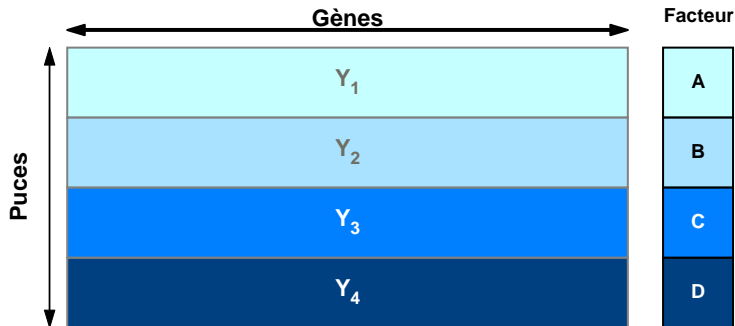
Objectif : identifier les gènes s'exprimant différemment selon les modalités d'un facteur [ou les valeurs d'une covariable]





Analyse différentielle

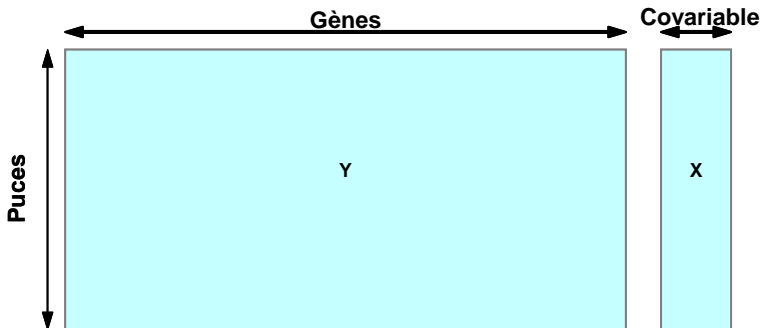
Objectif : identifier les gènes s'exprimant différemment selon les modalités d'un facteur [ou les valeurs d'une covariable]





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



Solutions dans R

De nombreux packages R pour l'analyse différentielle

- multtest [Bioconductor]
- locfdr (Stanford)
- kerfdr (INRA-Agroparistech)
- FAMT (Agrocampus) : prise en compte de la dépendance

Importations de données

```
> setwd(...)
```

```
> poulets = read.table("poulets.txt",sep="\t",header=TRUE)
```

```
> gras = read.table("gras.txt",sep="\t",header=TRUE)
```

```
> colnames(poulets) = gras$ArrayName
```



Plan du cours

- 1 Objectifs
- 2 Tests gène par gène
 - Test de la relation entre expression et covariable
 - Risques d'erreurs
- 3 Tests multiples
 - Stratégie générale
 - Contrôle du FWER
 - Contrôle du FDR
 - Optimisation des procédures
- 4 Perspectives



Comparaison de deux groupes

Test d'hypothèse : Y expression d'un gène

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

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

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



Test de l'effet d'une covariable continue

Test d'hypothèse : Y expression d'un gène

{ Si la covariable vaut x : $\mathbb{E}(Y) = \beta_0 + \beta_1 x$; Écart-type(Y) = σ

{ H_0 : $\beta_1 = 0$ [gène non différentiellement exprimé]
 { H_1 : $\beta_1 \neq 0$ [gène différentiellement exprimé]

Stratégie de test :

- Calcul de F , statistique de Fisher

$$F = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{S^2},$$

- Si F 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

- 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



Probabilité critique

Exercice

- Importer et normaliser les données `poulets` mono-couleurs.
- A l'aide de la fonction `rmvnorm` [package `mvtnorm`], simuler des données *images* des données `poulets` (27 puces, 314 gènes) mais de même moyenne pour les lignées grasses et maigres.
- Calculer les probabilités critiques des tests de comparaisons des moyennes d'expression entre lignées grasses et maigres.
- Calculer la proportion de faux positifs.



Tests multiples

Importations des données `poulets` bi-couleurs

```
> setwd(...)
```

```
> poulets = read.table("poulets.txt",sep="\t",header=TRUE)
```

```
> gras = read.table("gras.txt",sep="\t",header=TRUE)
```

```
> colnames(poulets) = gras$ArrayName
```

Exercice :

- Pour chaque gène, calculer la probabilité critique du test du lien entre son expression et le *poids de gras abdominal*.
- Donner les annotations des 10 gènes les plus positifs.
- Calculer [voir fonction `power.t.test`] la différence entre les expressions moyennes des poulets gras et des poulets maigres détectable par un test T de comparaison de moyennes avec probabilité 0.95.



Tests multiples avec FAMT

Création d'un ensemble expressions-covariables-annotations

```
> library(FAMT)
> annotations = read.table("annotations.txt",header=TRUE,sep="^")
> annotations$ID = as.character(annotations$ID)
> annotations$Name = as.character(annotations$Name)
> poulets = as.FAMTdata(poulets,gras,annotations,idcovar=1,idannot=1)
> summaryFAMT(poulets)
```

Tests multiples

```
> tests = raw.pvalues(poulets,x=6,test=6)
> ord = order(tests$pval)
> annotations$Name[ord][1:10]
> tests$pval[ord][1:10]
```



Puissance d'un test individuel

Exercice: calculer [voir fonction `power.t.test`] la différence entre les expressions moyennes des poulets gras et des poulets maigres détectable par un test T de comparaison de moyennes avec probabilité 0.95.



Puissance et Proportion de Vrais Positifs

Exercice

- A l'aide de la fonction `rmvnorm` [package `mvtnorm`], simuler des données *images* des données `poulets` (27 puces, 314 gènes) mais, pour les 100 premiers gènes, de moyenne $\delta = 0.1$ pour les lignées grasses et 0 maigres.
- Calculer les probabilités critiques des tests de comparaisons des moyennes d'expression entre lignées grasses et maigres.
- Représenter la répartition des probabilités critiques par un histogramme.
- Calculer la proportion de faux positifs et la proportion de vrais positifs.
- Même exercice avec $\delta = 0.5$.

Plan du cours

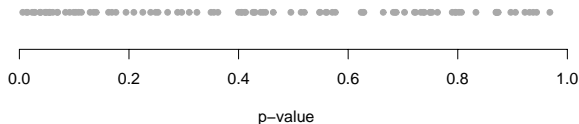
- 1 Objectifs
- 2 Tests gène par gène
 - Test de la relation entre expression et covariable
 - Risques d'erreurs
- 3 Tests multiples
 - Stratégie générale
 - Contrôle du FWER
 - Contrôle du FDR
 - Optimisation des procédures
- 4 Perspectives



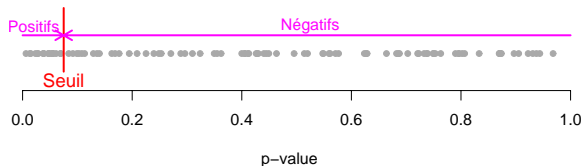
Stratégie générale

De manière générale

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

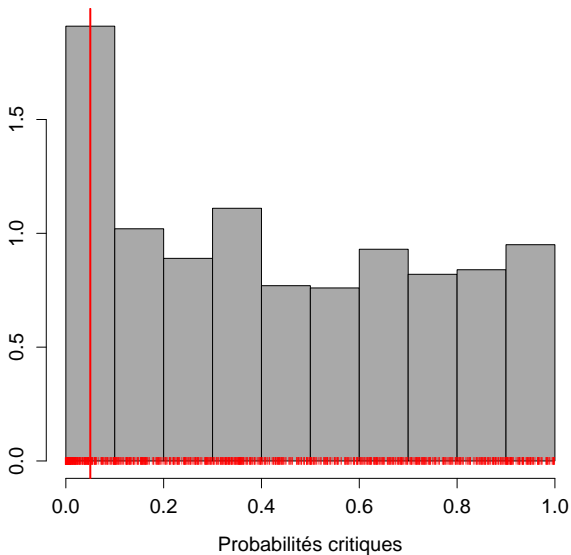


- Choix d'un seuil de décision





Stratégie générale





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 = α

Exercice

- A l'aide de la fonction `rmvnorm` [package `mvtnorm`], simuler des données *images* des données `poulets` (27 puces, 314 gènes) mais, pour les 100 premiers gènes, de moyenne $\delta = 0.25$ pour les lignées grasses et 0 maigres.
- Calculer les probabilités critiques des tests de comparaisons des moyennes d'expression entre lignées grasses et maigres.
- Donner tous les termes du tableau des décomptes d'erreurs tel que présenté dans le transparent précédent.
- Calculer le FDR et la proportion de positifs parmi les variables vraiment sous H_1 .



Contrôle du FWER

Approche de Bonferroni - \mathcal{M}_0 : indices des gènes non DE

$$\begin{aligned}
 \text{FWER}_t = \mathbb{P}(V_t > 0) &= \mathbb{P}\left\{ \bigcup_{i \in \mathcal{M}_0} [\text{gène } i \text{ positif}] \right\}, \\
 &\leq \sum_{i \in \mathcal{M}_0} \mathbb{P}[\text{gène } i \text{ positif}], \\
 &\leq \sum_{i \in \mathcal{M}_0} \mathbb{P}[P_i \leq t], \\
 &\leq m_0 t
 \end{aligned}$$

Contrôle au niveau α : Si $t = \alpha/m$, $\text{FWER}_t \leq \frac{m_0}{m} \alpha \leq \alpha$

Probabilités critiques ajustées : $p_i \leq \alpha/m \Leftrightarrow \underbrace{mp_i}_{\tilde{p}_i} \leq \alpha$



Contrôle du FWER

Approche de Bonferroni. $t = \alpha/m$

Approche de Šidák - \mathcal{M}_0 : indices des gènes non DE

$$\begin{aligned}
 1 - \text{FWER}_t = \mathbb{P}(V_t = 0) &= \mathbb{P}\left\{ \bigcap_{i \in \mathcal{M}_0} [\text{gène } i \text{ négatif}] \right\}, \\
 &= \prod_{i \in \mathcal{M}_0} \mathbb{P}[\text{gène } i \text{ négatif}], \text{ [si indépendants]} \\
 &= \prod_{i \in \mathcal{M}_0} \mathbb{P}[P_i \geq t], \\
 &= (1 - t)^{m_0}
 \end{aligned}$$

Contrôle au niveau α : Si $t = 1 - (1 - \alpha)^{1/m_0}$, $\text{FWER}_t \leq \alpha$



Contrôle du FWER

Approche de Bonferroni. $t = \alpha/m$

Approche de Šidák. $t = 1 - (1 - \alpha)^{1/m}$

Contrôle du FWER à 5 %

m	t Bonferroni	t Šidák
10	5e-03	2.14e-01
100	5e-04	2.38e-02
1000	5e-05	2.41e-03
5000	1e-05	4.81e-04
10000	5e-06	2.41e-04
50000	1e-06	4.82e-05



Propriétés de la méthode de Bonferroni

Exercice

- A l'aide de la fonction `rmvnorm` [package `mvtnorm`], simuler des données *images* des données `poulets` (27 puces, 314 gènes) mais, pour les 100 premiers gènes, de moyenne $\delta = 0.25$ pour les lignées grasses et 0 maigres.
- Calculer les probabilités critiques des tests de comparaisons des moyennes d'expression entre lignées grasses et maigres.
- A l'aide de la fonction `p.adjust`, corriger les probabilités critiques par la méthode de Bonferroni.
- Donner tous les termes du tableau des décomptes d'erreurs tel que présenté dans le transparent précédent.
- Calculer le FDR et la proportion de positifs parmi les variables vraiment sous H_1 .



Contrôle du FDR

Approche de Benjamini-Hochberg - \mathcal{M}_0 : indices des gènes non DE

$$\text{FDR}_t = \mathbb{E}(V_t/R_t).$$

Estimation du FDR

$$\widehat{\text{FDR}}_t = \frac{m_0 t}{R_t}.$$

Contrôle au niveau α :

Si $t = \max \left\{ t : \widehat{\text{FDR}}_t \leq \alpha \right\}$, $\text{FDR}_t \leq \alpha$



Contrôle du FDR

Approche de Benjamini-Hochberg - \mathcal{M}_0 : indices des gènes non DE

$$\text{FDR}_t = \mathbb{E}(V_t/R_t).$$

Estimation du FDR

$$\widehat{\text{FDR}}_t = \frac{m t}{R_t}.$$

Contrôle au niveau α :

$$\text{Si } t = \max \left\{ t : \widehat{\text{FDR}}_t \leq \alpha \right\}, \text{FDR}_t \leq \underbrace{\frac{m_0}{m}}_{\pi_0} \alpha \leq \alpha$$



Propriétés de la méthode de Benjamini & Hochberg

Exercice

- A l'aide de la fonction `rmvnorm` [package `mvtnorm`], simuler des données *images* des données `poulets` (27 puces, 314 gènes) mais, pour les 100 premiers gènes, de moyenne $\delta = 0.25$ pour les lignées grasses et 0 maigres.
- Calculer les probabilités critiques des tests de comparaisons des moyennes d'expression entre lignées grasses et maigres.
- A l'aide de la fonction `p.adjust`, corriger les probabilités critiques par la méthode de Benjamini-Hochberg.
- Donner tous les termes du tableau des décomptes d'erreurs tel que présenté dans le transparent précédent.
- Calculer le FDR et la proportion de positifs parmi les variables vraiment sous H_1 .



Analyse différentielle des données *poulets* bi-couleurs

Package FAMT

```
> poulet.FAMT = modelFAMT(poulet,x=6,nbf=0)
```

```
> summaryFAMT(poulet.FAMT,alpha=seq(0,0.3,0.02),pi0=1)
```



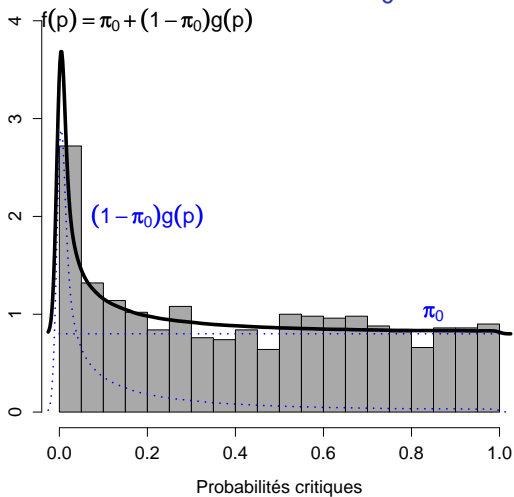
Estimation de π_0

Estimation à partir des probabilités critiques - Pour $0 \leq p \leq 1$,

$$\begin{aligned}
 F(p) = \mathbb{P}(P \leq p) &= \pi_0 G_0(p) + (1 - \pi_0) G(p), \\
 \text{[densité]} \quad f(p) &= \pi_0 g_0(p) + (1 - \pi_0) g(p),
 \end{aligned}$$



Estimation de π_0





Estimation de π_0

Estimation à partir des probabilités critiques - Pour $0 \leq p \leq 1$,

$$F(p) = \mathbb{P}(P \leq p) = \pi_0 G_0(p) + (1 - \pi_0)G(p),$$

[densité] $f(p) = \pi_0 g_0(p) + (1 - \pi_0)g(p),$

Un estimateur possible : $\hat{m}_0 = \hat{f}(1)$

> pi0 = pi0FAMT(poulet.raw,diagnostic.plot=TRUE)

Amélioration de la procédure de Benjamini-Hochberg :

$$\widehat{\text{FDR}}_t = \frac{\hat{m}_0 t}{R_t}.$$

Contrôle au niveau α : si $t = \max \left\{ t : \widehat{\text{FDR}}_t \leq \alpha \right\}, \text{FDR}_t \leq \alpha$



Hétérogénéité des données

Exercice

- A l'aide de la fonction `rmvnorm` [package `mvtnorm`], simuler des données `images` des données `poulets` (27 puces, 314 gènes) mais, pour les 100 premiers gènes, de moyenne $\delta = 0.5$ pour les lignées grasses et 0 maigres.
- Ajouter à chaque colonne du tableau de données un facteur d'hétérogénéité z ne modifiant pas les moyennes par génotype :
 - $z = \text{rnorm}(27)$
 - $z = \text{residuals}(\text{lm}(z \sim \text{poulets3}\$genotype))$
- Calculer les probabilités critiques des tests de comparaisons des lignées grasses et maigres et représenter leur répartition par un histogramme.
- Calculer le FDR et la proportion de positifs parmi les variables vraiment sous H_1 pour la méthode de Benjamini-Hochberg.



Modèle pour données hétérogènes

Modèle de la dépendance entre gènes. Pour le k ème gène,

$$Y^{(k)} = \beta_0 + \beta_1 X + \varepsilon^{(k)}, \quad \text{Var}(\varepsilon^{(k)}) = \sigma_k^2$$

$$Y^{(k)} = \beta_0 + \beta_1 X + b_1^{(k)} Z_1 + \dots + b_q^{(k)} Z_q + \epsilon^{(k)}, \quad \text{Var}(\epsilon^{(k)}) = \Psi_k^2 < \sigma_k^2$$

Analyse en facteurs : Z_1, Z_2, \dots, Z_q indépendants $\sim \mathcal{N}(0; 1)$.

- Les facteurs concentrent la dépendance [interprétation]
- Indépendance inter-gènes des résidus $\epsilon^{(k)}$
- Décomposition de la variance en composantes **commune** et **spécifique**

$$\sigma_k^2 = \sum_{i=1}^q [b_i^{(k)}]^2 + \Psi_k^2$$



Tests ajustés des facteurs

Statistiques de test. Pour le k ème gène,

$$T^{(k)} = T(Y^{(k)})$$

Statistiques de test ajustées des facteurs. Pour le k ème gène,

$$T^{(k)} = T(Y^{(k)})$$

$$T_z^{(k)} = T(Y^{(k)} - [b_1^{(k)} Z_1 + \dots + b_q^{(k)} Z_q])$$

Calcul des $T_z^{(k)}$

- Nombre q de facteurs
- Paramètres $b^{(k)}$ et Ψ_k du modèle d'analyse en facteurs



Procédure FAMT

Données hétérogènes obtenues par simulation

```
> expr = t(data)
```

```
> colnames(expr) = 1:27
```

```
> covar = data.frame(ld=1:27,x=groupes)
```

```
> simul = as.FAMTdata(expr,covar,idcovar=1)
```

Détermination du nombre de facteurs.

```
> nbf = nbfactors(simul,x=2,diagnostic.plot=TRUE)
```

Ajustement du modèle d'analyse en facteurs.

```
> simul.FAMT = modelFAMT(simul,x=2,nbf=1)
```

Calcul des probabilités critiques ajustées des facteurs

```
> pval = simul.FAMT$adjpval
```



Analyse différentielle de données hétérogènes

Exercice : Pour les données hétérogènes simulées, calculer le FDR et la proportion de positifs parmi les variables vraiment sous H_1 pour la méthode FAMT + Benjamini-Hochberg.

Exercice : Pour les données poulets bi-couleurs, donner la liste des gènes positifs par la méthode FAMT + Benjamini-Hochberg.

Plan du cours

- 1 Objectifs
- 2 Tests gène par gène
 - Test de la relation entre expression et covariable
 - Risques d'erreurs
- 3 Tests multiples
 - Stratégie générale
 - Contrôle du FWER
 - Contrôle du FDR
 - Optimisation des procédures
- 4 Perspectives



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