



Normalisation des données et planification expérimentale

David Causeur

Laboratoire de Mathématiques Appliquées

Agrocampus Ouest

IRMAR CNRS UMR 6625

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



Plan du cours

- 1 Préambule
- 2 Technologie mono-couleur
 - Analyse différentielle
 - Biais techniques
 - Planification expérimentale
- 3 Technologie bi-couleur
 - Effet fluorochrome
 - Interaction gène $\times$ fluorochrome
 - Plans classiques
- 4 Perspectives



Données de biopuce

Une structuration naturelle en sous-tableaux

Puces	Expressions géniques		
1	Y		
2			
3			
4			
5			
6			
⋮			

- Expressions géniques : quantitatives (nombreuses)



Données de biopuce

Une structuration naturelle en sous-tableaux

Puces	Expressions géniques	Facteurs de variations	
1	Y	X	
2			
3			
4			
5			
6			
⋮			

- Expressions géniques : quantitatives (nombreuses)
- Facteurs de variations : contrôlés (régime, génotype, temps, ...) ou nuisances (fluorochromes, marquage, ...)



Données de biopuce

Une structuration naturelle en sous-tableaux

Puces	Expressions géniques	Facteurs de variations	Informations externes
1	Y	X	Z
2			
3			
4			
5			
6			
⋮			

- Expressions géniques : quantitatives (nombreuses)
- Facteurs de variations : contrôlés (régime, génotype, temps, ...) ou nuisances (fluorochromes, marquage, ...)
- Informations externes : issues d'analyses parallèles



Planification expérimentale

Organiser les données de telle sorte que :

- l'interprétation ne soit pas entachée de confusion



Planification expérimentale

Organiser les données de telle sorte que :

- l'interprétation ne soit pas entachée de confusion
- les résultats soient aussi précis que possible



Planification expérimentale

Organiser les données de telle sorte que :

- l'interprétation ne soit pas entachée de confusion
- les résultats soient aussi précis que possible

Expertise préalable :

- Contraintes expérimentales (coût, technologie, ...)



Planification expérimentale

Organiser les données de telle sorte que :

- l'interprétation ne soit pas entachée de confusion
- les résultats soient aussi précis que possible

Expertise préalable :

- Contraintes expérimentales (coût, technologie, ...)
- But de l'étude ?
 - Comparer des traitements, des génotypes, ...
 - Mesurer l'évolution des niveaux d'expression dans le temps



Plan du cours

- 1 Préambule
- 2 Technologie mono-couleur
 - Analyse différentielle
 - Biais techniques
 - Planification expérimentale
- 3 Technologie bi-couleur
 - Effet fluorochrome
 - Interaction gène $\times$ fluorochrome
 - Plans classiques
- 4 Perspectives



Analyse différentielle

- Technologie mono-couleur
Sur chaque **spot**, l'expression d'**un gène** dans **une condition**.



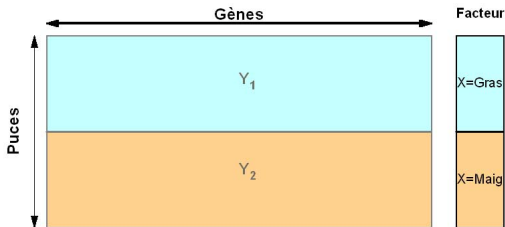
Analyse différentielle

- Technologie mono-couleur
Sur chaque **spot**, l'expression d'**un gène** dans **une condition**.
- Un seul facteur contrôlé, à 2 modalités (génotype gras/maigre)



Analyse différentielle

- Technologie mono-couleur
Sur chaque **spot**, l'expression d'**un gène** dans **une condition**.
- Un seul facteur contrôlé, à 2 modalités (génotype gras/maigre)





Comparaison de moyennes

Statistique de Student

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad S : \text{Écart-type intra-groupe}$$

Variabilité intra-groupe

- Si une puce = un patient (ou un animal) alors

S = variabilité biologique

- Si une puce = un échantillon sur un patient donné, ... alors

S = variabilité technique

Idéalement ... $\sigma = \mathbb{E}(S)$ petit



Homogénéité des variances

Exemple : données «poulets»



Homogénéité des variances

Exemple : données «poulets»

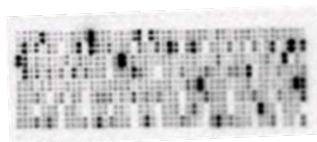
- Technologie : dépôt (mono-couleur)



Homogénéité des variances

Exemple : données «poulets»

- Technologie : dépôt (mono-couleur)





Homogénéité des variances

Exemple : données «poulets»

- Technologie : dépôt (mono-couleur)
- Nombre de gènes : 314



Homogénéité des variances

Exemple : données «poulets»

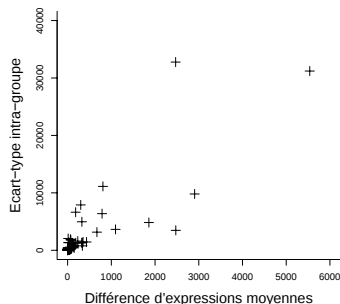
- Technologie : dépôt (mono-couleur)
- Nombre de gènes : 314
- Nombres de puces : 27
 - 13 gras
 - 14 maigres



Homogénéité des variances

Exemple : données «poulets»

- Technologie : dépôt (mono-couleur)
- Nombre de gènes : 314
- Nombres de puces : 27
 - 13 gras
 - 14 maigres



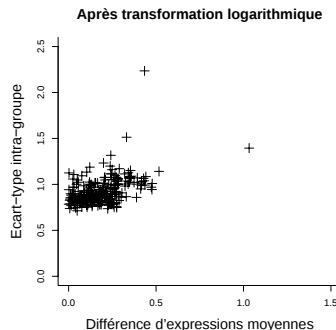


Homogénéité des variances

Exemple : données «poulets»

- Technologie : dépôt (mono-couleur)
- Nombre de gènes : 314
- Nombres de puces : 27
 - 13 gras
 - 14 maigres

Stabilisation de la variance
par passage au $\log_2$





Variabilité indésirable entre puces

$Y_{ir}^{(k)}$: expression du gène k sur la r ème puce du groupe i :

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \varepsilon_{ir}^{(k)}, \text{ Var}(\varepsilon_{ir}^{(k)}) = \sigma_k^2$$

σ_k = Variabilité intra-groupe

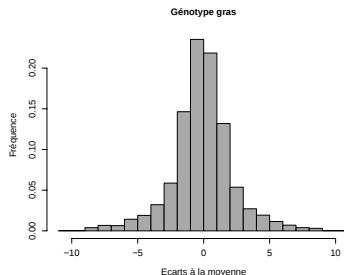


Variabilité indésirable entre puces

$Y_{ir}^{(k)}$: expression du gène k sur la r ème puce du groupe i :

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \varepsilon_{ir}^{(k)}, \text{ Var}(\varepsilon_{ir}^{(k)}) = \sigma_k^2$$

σ_k = Variabilité intra-groupe



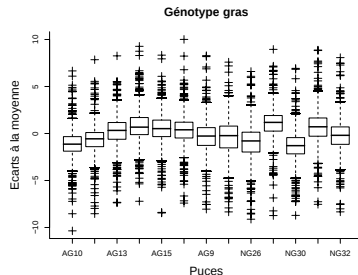
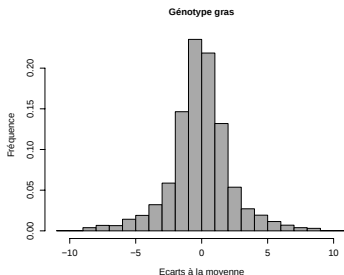


Variabilité indésirable entre puces

$Y_{ir}^{(k)}$: expression du gène k sur la r ème puce du groupe i :

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \varepsilon_{ir}^{(k)}, \text{ Var}(\varepsilon_{ir}^{(k)}) = \sigma_k^2$$

σ_k = Variabilité intra-groupe



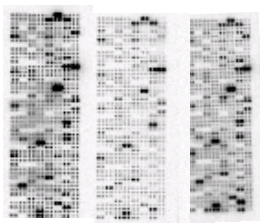


Variabilité indésirable entre puces

$Y_{ir}^{(k)}$: expression du gène k sur la r ème puce du groupe i :

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \varepsilon_{ir}^{(k)}, \text{Var}(\varepsilon_{ir}^{(k)}) = \sigma_k^2$$

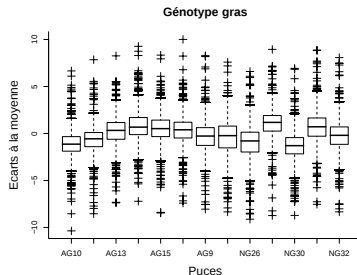
σ_k = Variabilité intra-groupe



Puce 24
(ech. 24)

Puce 25
(ech. 25)

Puce 26
(éch. 26)



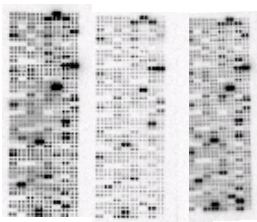


Variabilité indésirable entre puces

$Y_{ir}^{(k)}$: expression du gène k sur la r ème puce du groupe i :

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_{ir}^{(k)} + \varepsilon_{ir}^{(k)}, \text{Var}(\varepsilon_{ir}^{(k)}) = \sigma_k^2$$

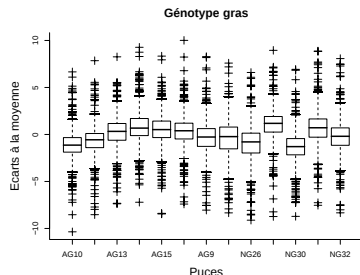
σ_k = Variabilité intra-groupe



Puce 24
(éch. 24)

Puce 25
(éch. 25)

Puce 26
(éch. 26)





Variabilité indésirable entre puces

$Y_{ir}^{(k)}$: expression du gène k sur la r ème puce du groupe i :

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r + \varepsilon_{ir}^{(k)}, \text{ Var}(\varepsilon_{ir}^{(k)}) = \sigma_k^2$$

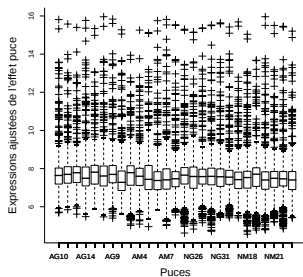
σ_k = Variabilité intra-groupe

Ajustement de l'effet puce (tous groupes confondus) :

- Invariance de gènes de contrôle
- Principe d'invariance moyenne entre puces

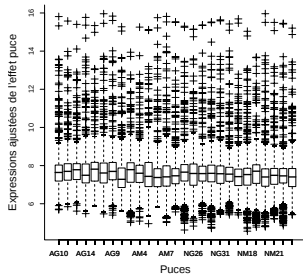


Variabilité indésirable entre puces

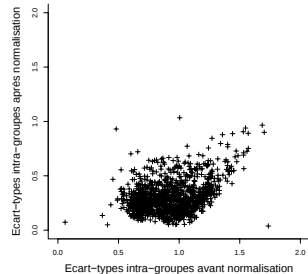




Variabilité indésirable entre puces



Effet de la normalisation sur la variabilité résiduelle





Variabilité indésirable non-homogène

Normalisation

Mêmes corrections pour tous les gènes



Variabilité indésirable non-homogène

Normalisation

Mêmes corrections pour tous les gènes

Contrôle de la qualité des données 

- Flags : spots retirés des données
- Recalcul ponctuel des valeurs d'expression  



Variabilité indésirable non-homogène

Génotype



A



A



B



B



Variabilité indésirable non-homogène

	Génotype	Régime
	A	α
	A	β
	B	α
	B	β



Variabilité indésirable non-homogène

	Génotype	Régime	Opérateur
	A	α	
	A	β	
	B	α	
	B	β	



Variabilité indésirable non-homogène



Génotype

Régime

Opérateur

A +1 α +1A +1 β -1B -1 α +1B -1 β -1



Variabilité indésirable non-homogène



Génotype

A +1A +1B -1B -1

Régime

 α +1 β -1 α +1 β -1

Opérateur

♣ +1♥ -1♥ -1♣ +1



Plan du cours

- 1 Préambule
- 2 Technologie mono-couleur
 - Analyse différentielle
 - Biais techniques
 - Planification expérimentale
- 3 Technologie bi-couleur
 - Effet fluorochrome
 - Interaction gène $\times$ fluorochrome
 - Plans classiques
- 4 Perspectives



Principe

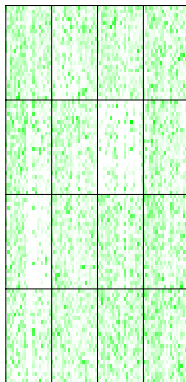
Technologie bi-couleur : une puce = un couple de mesures d'expression par gène

Couple = 1 couleur (**vert** ou **rouge**) par condition

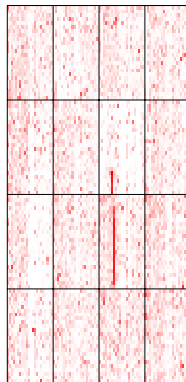


Principe

Données « zebrafish » : 8448 gènes



x-range 363.7 to 55699 (saturation 363.7, 55699)



x-range 164.9 to 60030.6 (saturation 164.9, 60030.6)



Principe

Technologie bi-couleur : une puce = un couple de mesures d'expression par gène

Couple = 1 couleur (**vert** ou **rouge**) par condition

Si conditions i et j sur la r ème puce

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \varepsilon_{ir}^{(k)},$$

$$Y_{jr}^{(k)} = \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \varepsilon_{jr}^{(k)},$$



Principe

Technologie bi-couleur : une puce = un couple de mesures d'expression par gène

Couple = 1 couleur (**vert** ou **rouge**) par condition

Si conditions i et j sur la r ème puce

$$\begin{aligned} Y_{ir}^{(k)} &= \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \varepsilon_{ir}^{(k)}, \\ Y_{jr}^{(k)} &= \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \varepsilon_{jr}^{(k)}, \end{aligned}$$

Pas d'effet puce sur le log-ratio des expressions

$$M_{ijr}^{(k)} = Y_{ir}^{(k)} - Y_{jr}^{(k)}.$$



Effet fluorochrome

Expressions sur la *rème* puce (conditions *i* et *j*)

$$M_{ijr}^{(k)} = \alpha_i^{(k)} - \alpha_j^{(k)} + \text{Bruit}$$



Effet fluorochrome

Expressions sur la r ème puce (conditions i et j)

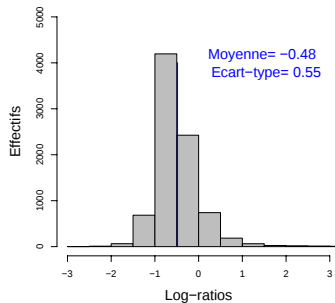
$$M_{ijr}^{(k)} = \alpha_i^{(k)} - \alpha_j^{(k)} + \text{Bruit}$$

Hypothèse d'invariance moyenne : pour presque tous les k ,
 $\alpha_i^{(k)} - \alpha_j^{(k)} = 0$



Effet fluorochrome

Histogramme des valeurs M
Puce 1





Effet fluorochrome

Expressions sur la r ème puce (conditions i et j)

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \gamma_{\text{rouge}}^{(k)} + \varepsilon_{ir}^{(k)},$$

$$Y_{jr}^{(k)} = \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \gamma_{\text{vert}}^{(k)} + \varepsilon_{jr}^{(k)},$$



Effet fluorochrome

Expressions sur la r ème puce (conditions i et j)

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \gamma_{\text{rouge}}^{(k)} + \varepsilon_{ir}^{(k)},$$

$$Y_{jr}^{(k)} = \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \gamma_{\text{vert}}^{(k)} + \varepsilon_{jr}^{(k)},$$

Exemple : 2 puces, 2 génotypes

Puce	Génotype	Couleur
1	G	Rouge
1	M	Vert
2	G	Vert
2	M	Rouge



Effet fluorochrome

Expressions sur la r ème puce (conditions i et j)

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \gamma_{\text{rouge}}^{(k)} + \varepsilon_{ir}^{(k)},$$

$$Y_{jr}^{(k)} = \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \gamma_{\text{vert}}^{(k)} + \varepsilon_{jr}^{(k)},$$

Exemple : 2 puces, 2 génotypes

Puce	Génotype	Couleur
1	G	Rouge
1	M	Vert
2	G	Vert
2	M	Rouge

Plan « dye swap »



Effet fluorochrome

Expressions sur la r ème puce (conditions i et j)

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \gamma_{\text{rouge}}^{(k)} + \varepsilon_{ir}^{(k)},$$

$$Y_{jr}^{(k)} = \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \gamma_{\text{vert}}^{(k)} + \varepsilon_{jr}^{(k)},$$

Exemple : 2 puces, 2 génotypes

Puce	Génotype	Couleur
1	G	Rouge
1	M	Vert
2	G	Vert
2	M	Rouge

$M_{ij1} = \text{Condition} + \text{Fluorochrome}$

$M_{ij2} = \text{Condition} - \text{Fluorochrome}$

$M_{ij1} + M_{ij2} = \text{Effet condition}$



Interaction gène $\times$ fluorochrome

Expressions d'un gène sur une puce

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \gamma_{\text{rouge}}^{(k)} + \varepsilon_{ir}^{(k)},$$

$$Y_{jr}^{(k)} = \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \gamma_{\text{vert}}^{(k)} + \varepsilon_{jr}^{(k)},$$



Interaction gène $\times$ fluorochrome

Expressions d'un gène sur une puce

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \gamma_{\text{rouge}}^{(k)} + \varepsilon_{ir}^{(k)},$$

$$Y_{jr}^{(k)} = \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \gamma_{\text{vert}}^{(k)} + \varepsilon_{jr}^{(k)},$$

$$M_{ijr}^{(k)} = Y_{ir}^{(k)} - Y_{jr}^{(k)} = \left[\alpha_i^{(k)} - \alpha_j^{(k)} \right] + \left[\gamma_{\text{rouge}}^{(k)} - \gamma_{\text{vert}}^{(k)} \right] + \text{Bruit}$$

$$A_{ijr}^{(k)} = Y_{ir}^{(k)} + Y_{jr}^{(k)} = 2\mu^{(k)} + 2\beta_r^{(k)} + \text{Bruit}$$



Interaction gène $\times$ fluorochrome

Expressions d'un gène sur une puce

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \gamma_{\text{rouge}}^{(k)} + \varepsilon_{ir}^{(k)},$$

$$Y_{jr}^{(k)} = \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \gamma_{\text{vert}}^{(k)} + \varepsilon_{jr}^{(k)},$$

$$M_{ijr}^{(k)} = Y_{ir}^{(k)} - Y_{jr}^{(k)} = \left[\alpha_i^{(k)} - \alpha_j^{(k)} \right] + \left[\gamma_{\text{rouge}}^{(k)} - \gamma_{\text{vert}}^{(k)} \right] + \text{Bruit}$$

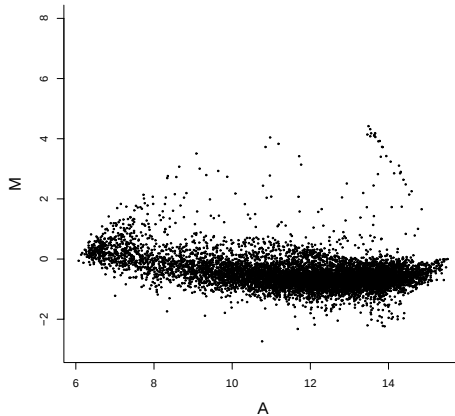
$$A_{ijr}^{(k)} = Y_{ir}^{(k)} + Y_{jr}^{(k)} = 2\mu^{(k)} + 2\beta_r^{(k)} + \text{Bruit}$$

Hypothèse : pour la plupart des gènes, effet condition = 0



Interaction gène $\times$ fluorochrome

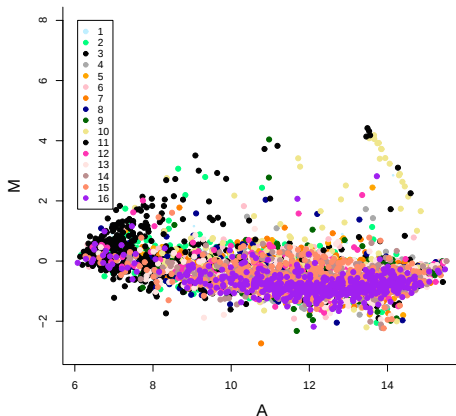
Graphe MA – Données zebrafish – Puce 1





Interaction gène $\times$ fluorochrome

Graphe MA – Données zebrafish – Puce 1





Interaction gène $\times$ fluorochrome

Expressions d'un gène sur une puce

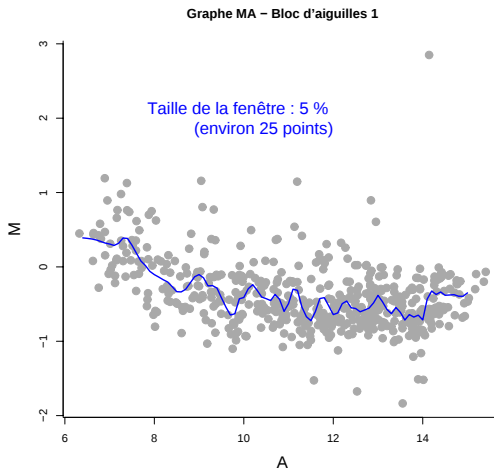
$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \gamma_{\text{rouge}}^{(k)} + \varepsilon_{ir}^{(k)},$$

$$Y_{jr}^{(k)} = \mu^{(k)} + \alpha_j^{(k)} + \beta_r^{(k)} + \gamma_{\text{vert}}^{(k)} + \varepsilon_{jr}^{(k)},$$

$$\text{Effet fluorochrome} = f(\text{potentiel gène})$$



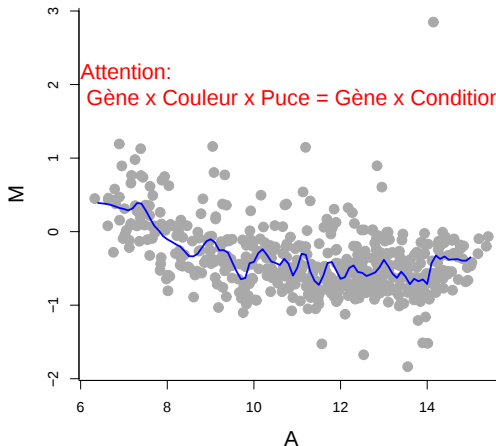
Correction de l'interaction gène $\times$ fluorochrome





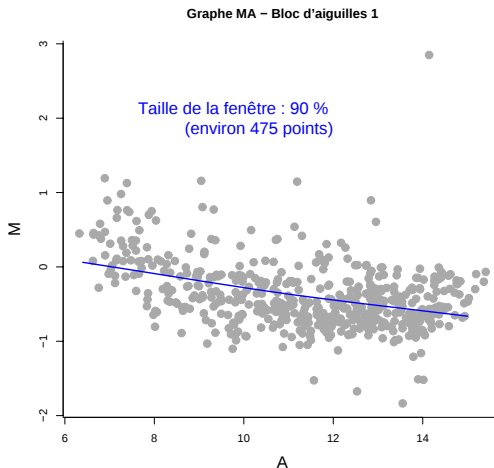
Correction de l'interaction gène $\times$ fluorochrome

Graphe MA – Bloc d'aiguilles 1



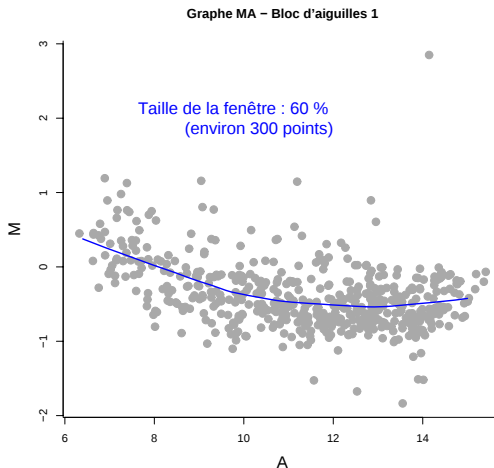


Correction de l'interaction gène $\times$ fluorochrome



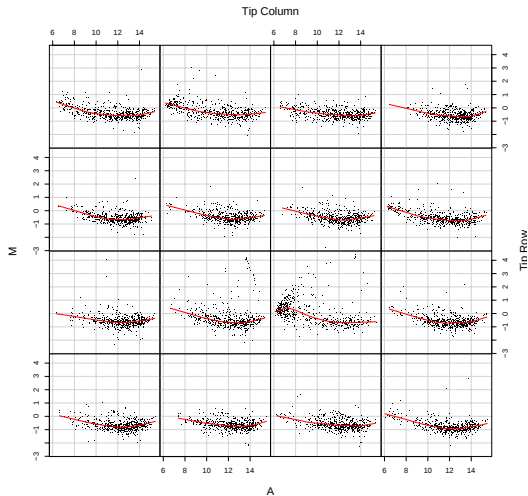


Correction de l'interaction gène $\times$ fluorochrome





Correction de l'interaction gène $\times$ fluorochrome





Correction de l'interaction gène $\times$ fluorochrome

Normalisation des données:

$$Y_{ir}^{(k)} \Leftarrow Y_{ir}^{(k)} - \hat{\gamma}_{\text{rouge},r}^{(k)}$$

Modèle:

$$Y_{ir}^{(k)} = \mu^{(k)} + \alpha_i^{(k)} + \beta_r^{(k)} + \varepsilon_{ir}^{(k)}$$

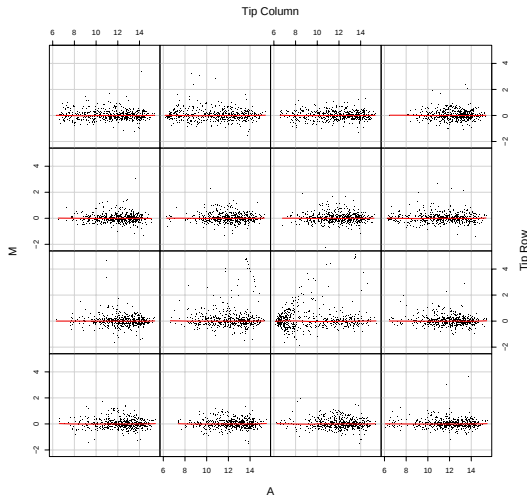
Données:

Gènes					Puce	Facteur
1	2	3	...	m		
☐	☐	☐	...	☐	1	A
☐	☐	☐	...	☐	1	B
☐	☐	☐	...	☐	2	A
☐	☐	☐	...	☐	2	C
		$\vdots$			$\vdots$	$\vdots$
☐	☐	☐	...	☐	R	C
☐	☐	☐	...	☐	R	B

☐ = Y



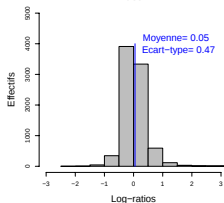
Correction de l'interaction gène $\times$ fluorochrome



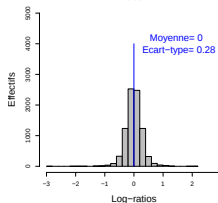


Correction de l'interaction gène $\times$ fluorochrome

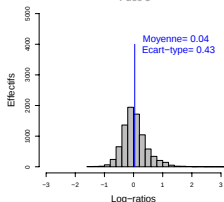
Histogramme des valeurs M
Puce 1



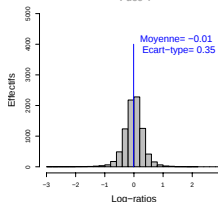
Histogramme des valeurs M
Puce 2



Histogramme des valeurs M
Puce 3



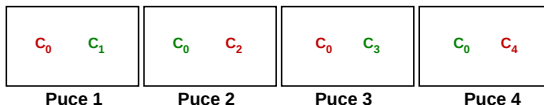
Histogramme des valeurs M
Puce 4





Plan « en étoile »

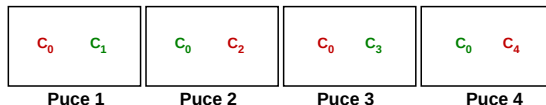
Lorsqu'une condition de référence s'impose ...





Plan « en étoile »

Lorsqu'une condition de référence s'impose ...



Données:

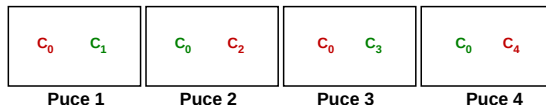
Gènes					Puce	Facteur
1	2	3	...	m		
☐	☐	☐	...	☐	1	C ₁
☐	☐	☐	...	☐	1	C ₀
☐	☐	☐	...	☐	2	C ₀
☐	☐	☐	...	☐	2	C ₂
		⋮			⋮	⋮
☐	☐	☐	...	☐	4	C ₀
☐	☐	☐	...	☐	4	C ₄

☐ = Y



Plan « en étoile »

Lorsqu'une condition de référence s'impose ...



Données:

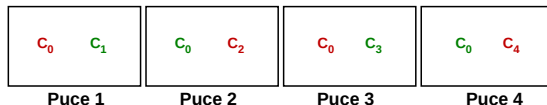
Gènes					Puce	Facteur
1	2	3	...	m		
☐	☐	☐	...	☐	1	C_1
☐	☐	☐	...	☐	2	C_2
☐	☐	☐	...	☐	3	C_3
☐	☐	☐	...	☐	4	C_4

$$\square = M = Y - Y_0$$



Plan « en étoile »

Lorsqu'une condition de référence s'impose ...



Données:

Gènes					Puce	Facteur
1	2	3	...	m		
☐	☐	☐	...	☐	1	C_1
☐	☐	☐	...	☐	2	C_2
☐	☐	☐	...	☐	3	C_3
☐	☐	☐	...	☐	4	C_4

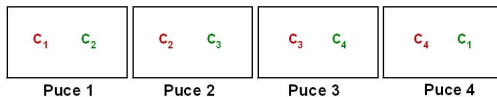
$$\square = M = Y - Y_0$$

Estimateur	Variance
M_{i0}	$2 \frac{\sigma^2}{R}$
$M_{i0} + M_{0j}$	$4 \frac{\sigma^2}{R}$



Plan « en boucle »

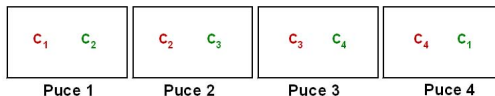
Lorsqu'une notion d'ordre existe entre les conditions ...





Plan « en boucle »

Lorsqu'une notion d'ordre existe entre les conditions ...



Données:

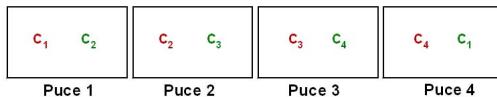
Gènes					Puce	Facteur
1	2	3	...	m		
☐	☐	☐	...	☐	1	C_1
☐	☐	☐	...	☐	1	C_2
☐	☐	☐	...	☐	2	C_2
☐	☐	☐	...	☐	2	C_3
		⋮			⋮	⋮
☐	☐	☐	...	☐	4	C_4
☐	☐	☐	...	☐	4	C_1

☐ = Y



Plan « en boucle »

Lorsqu'une notion d'ordre existe entre les conditions ...

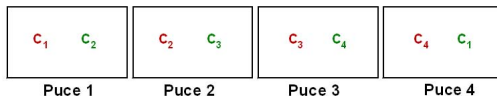


Estimateur	Variance
$M_{1,2}$	$2 \frac{\sigma^2}{R}$
$M_{2,3} + M_{1,2}$	$4 \frac{\sigma^2}{R}$



Plan « en boucle »

Lorsqu'une notion d'ordre existe entre les conditions ...



Estimateur	Variance
$M_{1,2}$	$2 \frac{\sigma^2}{R}$
$M_{2,3} + M_{1,2}$	$4 \frac{\sigma^2}{R}$

Estimateur	Variance
$\frac{3}{4} M_{1,2} + \frac{1}{4} [M_{1,4} + M_{4,3} + M_{3,2}]$	$\frac{3}{4} 2 \frac{\sigma^2}{R}$
$\frac{1}{2} [M_{1,2} + M_{2,3}] + \frac{1}{2} [M_{1,4} + M_{4,3}]$	$\frac{1}{2} 4 \frac{\sigma^2}{R}$



Plan du cours

- 1 Préambule
- 2 Technologie mono-couleur
 - Analyse différentielle
 - Biais techniques
 - Planification expérimentale
- 3 Technologie bi-couleur
 - Effet fluorochrome
 - Interaction gène $\times$ fluorochrome
 - Plans classiques
- 4 Perspectives



Perspectives

Une première réflexion a permis

- la réduction de σ par l'identification d'une variabilité « technologique »
- l'organisation optimale des données



Perspectives

Une première réflexion a permis

- la réduction de σ par l'identification d'une variabilité « technologique »
- l'organisation optimale des données

Tout est en place pour

- l'analyse des effets
- l'identification des gènes s'exprimant différemment d'une condition à une autre