



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
 - Qualité des données
 - Effet fluorochrome
 - Interaction gène \times fluorochrome
 - Création des tableaux de données
 - Plans classiques
- 4 Perspectives



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
- 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



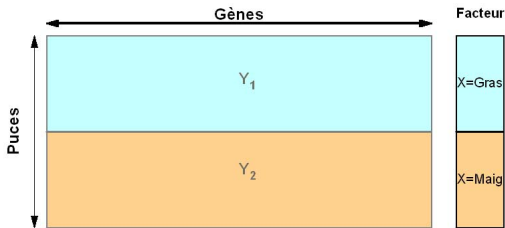
Plan du cours

- 1 Préambule
- 2 Technologie mono-couleur
 - Analyse différentielle
 - Biais techniques
 - Planification expérimentale
- 3 Technologie bi-couleur
 - Qualité des données
 - Effet fluorochrome
 - Interaction gène \times fluorochrome
 - Création des tableaux de données
 - Plans classiques
- 4 Perspectives



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»

- Technologie : dépôt (mono-couleur)
- 314 gènes, 27 puces (13 lignées grasses, 14 lignées maigres)

Exercice

- Importer les données poulets.txt
- Calculer les différences moyennes d'expressions Δ pour chaque gène
- Calculer les écart-types intra-génotypes d'expressions σ pour chaque gène
- Visualiser par un graphique la relation entre Δ et σ
- Reproduire l'exercice après passage au \log_2



Variabilité indésirable entre puces

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

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

$\varepsilon_{ir}^{(k)}$ = résidus

Exercice

- Calculer la matrice [puces \times gènes] des résidus
- Calculer les moyennes résiduelles par gène
- Calculer les moyennes résiduelles par puce



Variabilité indésirable entre puces

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

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

$\varepsilon_{ir}^{(k)}$ = résidus



Variabilité indésirable entre puces

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

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

$\varepsilon_{ir}^{(k)}$ = résidus

Exercice

- Calculer la matrice [puces \times gènes] des données corrigées de l'effet puce
- Visualiser graphiquement la variabilité de ces données par puce
- Quelles sont les valeurs estimées de σ_k ?



Variabilité indésirable entre puces

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

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

$\varepsilon_{ir}^{(k)}$ = résidus

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

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





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	Régime
	A	α
	A	β
	B	α
	B	β



Plan du cours

- 1 Préambule
- 2 Technologie mono-couleur
 - Analyse différentielle
 - Biais techniques
 - Planification expérimentale
- 3 Technologie bi-couleur
 - Qualité des données
 - Effet fluorochrome
 - Interaction gène \times fluorochrome
 - Création des tableaux de données
 - 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

Package `limma` dans `Bioconductor`

- Importation de données issues de la plupart des logiciels d'analyse d'image : Agilent Feature Extraction, ArrayVision, BlueFuse, **GenePix**, ImaGene, QuantArray, Stanford Microarray Database (SMD) et SPOT
- Contrôle qualité des données et normalisation
- Analyse différentielle

Chargement du package :

> `library(limma)`



Organisation des données

Stockage des fichiers image

<http://www.agrocampus-ouest.fr/math/causeur/Genomic/DataFC11/>

Organisation des données : `targets.txt`

```
> targets = readTargets("http://www.agrocampus-ouest.fr/.../DataFC11/targets.txt")
```

```
R Console
> targets
  ArrayName AFClass  Cy5  Cy3      FileName      AF
1         F1         F 1089 ref refCy3-1089Cy5.gpr.txt 22.3606800
2        F10         F 3789 ref refCy3-3789Cy5.gpr.txt 10.0861000
3        F11         F 3743 ref refCy3-3743Cy5.gpr.txt  9.4094550
4        F12         F  499 ref  refCy3-499Cy5.gpr.txt  8.9252110
5        F13         F 516B ref refCy3-516BCy5.gpr.txt  8.7130550
6        F14         F 3793 ref refCy3-3793Cy5.gpr.txt  8.4943240
7        F15         F 1072 ref refCy3-1072Cy5.gpr.txt  8.0738500
8        F16         F 1026 ref refCy3-1026Cy5.gpr.txt  7.4666286
9        F17         F 1042 ref refCy3-1042Cy5.gpr.txt  6.7985220
10       F18         F 641B ref refCy3-641BCy5.gpr.txt  6.1469140
11       F19         F 3754 ref refCy3-3754Cy5.gpr.txt  5.8629960
12        F2         F 1025 ref refCy3-1025Cy5.gpr.txt 18.1023800
13       F20         F  626 ref  refCy3-626Cy5.gpr.txt  5.7827070
14        F3         F 3791 ref refCy3-3791Cy5.gpr.txt 16.4491600
15        F4         F 3788 ref refCy3-3788Cy5.gpr.txt 16.3777900
16        F5         F 1069 ref refCy3-1069Cy5.gpr.txt 14.9184300
17        F6         F 1005 ref refCy3-1005Cy5.gpr.txt 14.5674700
18        F7         F 1030 ref refCy3-1030Cy5.gpr.txt 14.2151900
19        F8         F 1090 ref refCy3-1090Cy5.gpr.txt 10.3679600
```



Contenu des fichiers image

Structure d'un fichier genepix

```

"jpegOrigin=2500, 4400"
"Creator=GenePix Pro 5.1.0.16"
"Scanner=GenePix 4000A [55444]"
"FocusPosition=0"
"Temperature=44.04"
"LinesAveraged=1"
"Comment="
"PMTGain=500 600"
"ScanPower=100 100"
"LaserPower=2.02 2.26"
"Filters="
"ScanRegion=200,412,2080,5876"
"Supplier=BioRobotics"
"ArrayerSoftwareName=TAS Application Suite (MicroGrid II)"
"ArrayerSoftwareVersion=2.4.0.2"
"Block" "Column" "Row" "Name" "ID" "X" "Y" "Dia." "F635 Median" "F635 Mean" "F635 SD" "F
1 1 1 "GAPDH cntrl" "RIGG14714" 2790 4720 120 235 210 80 38 24 24 25 4 16 100 99
1 2 1 "GAPDH cntrl" "RIGG14714" 2970 4730 130 229 210 80 38 24 24 25 4 16 100 10
1 3 1 "Gallus gallus mRNA for hypothetical protein, clone 6a18" "RIGG00105" 3170 4720 12
1 4 1 "ENSGALT00000017516.1" "RIGG15897" 3350 4720 130 38 39 10 25 24 24 25 4 16
1 5 1 "Weakly similar to Q9NX88 (Q9NX88) Hypothetical protein FLJ20374" "RIGG04616" 3550 47
1 6 1 "Genome Hit Contig103.14" "RIGG01680" 3740 4730 130 48 49 11 22 24 24 25 4
1 7 1 "ENSGALT00000008978.1" "RIGG12912" 3930 4720 120 72 75 25 33 24 24 25 5 20
1 8 1 "Genome Hit Contig1.750" "RIGG04987" 4120 4720 130 287 291 101 34 24 24 25 5
1 9 1 "Contig Hit 037550.1" "RIGG03113" 4310 4720 120 40 42 10 23 24 24 25 4 16
1 10 1 "Genome Hit Contig228.14" "RIGG06115" 4500 4710 130 327 389 139 35 24 24 25 4

```



Contenu des fichiers image

Structure d'un fichier genepix

"B635"	"B635 Median"	"B635 Mean"	"B635 SD"	"B635 CV"	"% > B635+1SD"	"% > B635+2SD"
450 170 37 37 37 38	9 23 100 100 0	0.462 0.450	0.454 0.436	1.653 0.433		
442 173 39 36 36 38	9 23 100 100 0	0.443 0.458	0.459 0.456	1.390 0.446		
6 18 25 25 25 4	16 66 38 0 66 71 36 50 37	37 39 9 23 91 79 0 0.2				
0 74 74 17 22 37	37 39 11 28 95 84 0	0.378 0.405	0.401 0.385	2.345		
37 38 8 21 24 24	25 5 20 84 65 0 69 70 18 25 36 36 38	11 28 90 75				
95 0 93 95 22 23	36 36 38 9 23 99 96 0	0.421 0.424	0.421 0.398	1.9		
0 119 124 37 29 36	36 38 10 26 98 96 0	0.578 0.580	0.573 0.559	1.684		
98 0 606 642 176 27	36 36 38 10 26 100 100 0	0.461 0.441	0.447 0.392	2.0		
0 81 83 21 25 36	36 38 9 23 99 91 0	0.356 0.383	0.345 0.354	2.018		
100 0 627 731 249 34	37 37 38 9 23 100 100 0	0.514 0.526	0.520 0.521	1.2		



Contenu des fichiers image

Structure d'un fichier genepix

	"F532 Median"	"F532 Mean"	"F532 SD"	"F532 CV"	"B532"	"B532 Median"	"B532 Mean"	"B532 SD"	"B532 CV"
100	668 599	-1.115	211 457	186 413	25182	53957	46.250	45.778	0 0 0
100	668 592	-1.175	205 463	186 406	25217	53081	46.250	44.889	0 0 0
57	0.250	2.566	0.103	0.185	120 594	100 36	41	-2.051	7 29 7 34 3878 8478 1.750
120	648 100	51 52	-1.402	14 37	15 37	4708	8878	3.500	3.182 0 0 0
12	0.379	0.376	2.089	0.286	0.352	120 595	100 46	48	-1.344 13 33 14 34 4500 8424
54	120 647	100 81	84	-1.248	24 57	25 59	5831	11380	6.000 6.333 0 0 0
120	594 100	131 139	-0.790	48 83	51 88	8942	14938	10.000	8.600 0 0 0
30	120 662	100 833	873	-1.116	263 570	267 606	34917	76992	53.200 60.400 0 0 0
120	591 100	61 65	-1.492	16 45	18 47	4984	9924	4.250	5.000 0 0 0
97	120 648	100 893	1059	-0.961	303 590	365 694	46692	87731	91.000 77.000 0 0 0



Contenu des fichiers image

Structure d'un fichier genepix

```
"F532 Mean - B532" "F635 Total Intensity" "F532 Total Intensity" "SNR 635" "SNR 532" "Flags"
```



Contenu des fichiers image

Importation des fichiers image

```
> RG = read.maimages(files=target$FileName,
  source="genepix",wt.fun=MonFiltrage)
```

`wt.fun` (weighting function) = `MonFiltrage` : attribue 0 aux spots de mauvaise qualité, 1 aux autres

Flags = -100 (Bad), -75 (Absent), -50 (Not found), +100 (Good)

```
MonFiltrage = fonction(X) { # X, tableau Genepix
  okFLAG = X$Flags > -49 # okFLAG=TRUE si Flags>-49
  return(as.numeric(okFLAG)) } # la fonction as.numeric convertit
# selon la règle suivante : TRUE ↔ 1, FALSE ↔ 0
```



Contrôle qualité des données

Qualité = intensité et homogénéité du signal

```

MonFiltrage = fonction(X, seuilSNR=2, seuilH=0.2) {
# Par défaut, seuilSNR=2 et seuilH=0.2
okFLAG = X$Flags > -49 # Début test sur le rapport signal/bruit
okSNRred = X[,"SNR 635"] > seuilSNR
okSNRgreen = X[,"SNR 532"] > seuilSNR
okSNR = okSNRred & okSNRgreen # Début test d'homogénéité
NumRed = abs(X[,"F635 Median"]-X[,"F635 Mean"])
DenomRed = 0.5*(X[,"F635 Median"]+X[,"F635 Mean"])
okHRed = (NumRed/DenomRed) < seuilH
NumGreen = abs(X[,"F532 Median"]-X[,"F532 Mean"])
DenomGreen = 0.5*(X[,"F532 Median"]+X[,"F532 Mean"])
okHGreen = (NumGreen/DenomGreen) < seuilH
okH = okHRed & okHGreen
ok = okFLAG & okSNR & okH
# ok=TRUE si okFLAG=TRUE et okSNR=TRUE et okH=TRUE
return(as.numeric(ok)) }

```



Structure des données

Importation des fichiers `image` et contrôle qualité

```
> RG = read.maimages(files=target$FileName,  
source="genepix",wt.fun=MonFiltrage)
```

```
R Console  
> RG<-read.maimages(files=target$FileName,source="genepix",wt.fun=MonFiltrage)  
Read refCy3-1089Cy5.gpr.txt  
Read refCy3-3789Cy5.gpr.txt  
Read refCy3-3743Cy5.gpr.txt  
Read refCy3-499Cy5.gpr.txt  
Read refCy3-516BCy5.gpr.txt  
Read refCy3-3793Cy5.gpr.txt  
Read refCy3-1072Cy5.gpr.txt  
Read refCy3-1026Cy5.gpr.txt  
Read refCy3-1042Cy5.gpr.txt  
Read refCy3-641BCy5.gpr.txt  
Read refCy3-3754Cy5.gpr.txt  
Read refCy3-1025Cy5.gpr.txt  
Read refCy3-626Cy5.gpr.txt  
Read refCy3-3791Cy5.gpr.txt  
Read refCy3-3788Cy5.gpr.txt  
Read refCy3-1069Cy5.gpr.txt  
Read refCy3-1005Cy5.gpr.txt  
Read refCy3-1030Cy5.gpr.txt  
Read refCy3-1090Cy5.gpr.txt  
Read refCy3-643BCy5.gpr.txt  
Read refCy3-3761Cy5.gpr.txt  
Read refCy3-649BCy5.gpr.txt
```



Structure des données

Importation des fichiers `image` et contrôle qualité

```
> RG = read.maimages(files=target$FileName,
  source="genepix",wt.fun=MonFiltrage)
```

Structure de RG

```
> names(RG)
```

```
Read refCy3-644ACy5.gpr.txt
Read refCy3-527Cy5.gpr.txt
Read refCy3-627Cy5.gpr.txt
Read refCy3-1078Cy5.gpr.txt
Read refCy3-1169Cy5.gpr.txt
Read refCy3-1099Cy5.gpr.txt
Read refCy3-611Cy5.gpr.txt
Read refCy3-3776Cy5.gpr.txt
Read refCy3-522ACy5.gpr.txt
> names(RG)
[1] "R"      "G"      "Rb"     "Gb"     "weights" "targets" "genes"  "source" "printer"
```



Structure des données

Importation des fichiers `image` et contrôle qualité

```
> RG = read.maimages(files=target$FileName,
  source="genepix",wt.fun=MonFiltrage)
```

Structure de `RG`

```
> names(RG)
```

Structure de `RG$genes`

```
> names(RG$genes)
```

```
Read refCy3-611Cy5.gpr.txt
Read refCy3-3776Cy5.gpr.txt
Read refCy3-522ACy5.gpr.txt
>
> names(RG)
[1] "R"      "G"      "Rb"     "Gb"     "weights" "targets" "genes"  "source" "printer"
> names(RG$genes)
[1] "Block" "Row"   "Column" "ID"    "Name"
```



Statut des spots

Certain spots sont blancs :

```
> spottypes = readSpotTypes("SpotTypes.txt")
```

```
> spottypes <- readSpotTypes("SpotTypes.txt")
> spottypes
  SpotType      ID      Name Color
1      gene      *      *  black
2 blancBlank blanc* blank* orange
3 blancEmpty  Empty*  Empty*  pink
4 blancOperon opHsV04NC* operon QC control* red
5 blancBuffer Spotting buffer Spotting buffer brown
6      blancVide
> |
```



Statut des spots

Certain spots sont blancs :

```
> spottypes = readSpotTypes("SpotTypes.txt")
```

Affectation de son statut à chaque spot

```
> RG$genes$Status = controlStatus(spottypes, RG)
```



Identification des puces défectueuses

Exercice

- Pour chaque puce, calculer le taux de spots de bonne qualité.
- Quelles puces ont $< 60\%$ de spots de bonne qualité ?



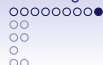
Analyse exploratoire du bruit de fond

Analyse comparative des puces :

```
> boxplot(data.frame(log2(RG$Rb)),names=targets$ArrayName)
```

Analyse détaillée d'une puce

```
> imageplot(log2(RG$Rb[,3]),layout=RG$printer)
```



Analyse exploratoire du signal

Analyse comparative des puces :

- > `boxplot(data.frame(log2(RG$R)),names=targets$ArrayName)`
- > `boxplot(data.frame(log2(RG$G)),names=targets$ArrayName)`

Analyse détaillée d'une puce

- > `par(mfrow=c(1,2))`
- > `imageplot(log2(RG$R[,22]),layout=RG$printer,low="red")`
- > `imageplot(log2(RG$G[,22]),layout=RG$printer,low="green")`
- > `par(mfrow=c(1,1))`



Modèle d'analyse différentielle pour données bi-couleurs

Si conditions i et j sur la même puce

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

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

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*)

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

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



Effet fluorochrome

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

Exercice

- Visualiser par un histogramme la répartition des valeurs M de la 1ère puce
- Quelle conclusion tirer de cet histogramme ?



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} = Condition + Fluorochrome

M_{ij2} = Condition – Fluorochrome

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



Interaction gène \times fluorochrome

Expressions d'un gène sur une puce

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

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

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

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

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



Interaction gène \times fluorochrome

Exercice :

- Pour les puces 1 et 2, représenter graphiquement les nuages de points (A,M)
- Que conclure de ces graphiques ?



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

Exercice : à l'aide de la fonction `plotPrintTipLoess`,
représenter f pour la puce 2



Correction de l'interaction gène \times fluorochrome

Normalisation des données:

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



Correction de l'interaction gène × fluorochrome

Normalisation des données:

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

Exercice : à l'aide de la fonction `normalizeWithinArrays`, normaliser les données



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		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	1	A
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	1	B
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	2	A
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	2	C
		⋮			⋮	⋮
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	R	C
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	R	B

= Y



Tableaux de données

Données `M`, `Weights`, `Annotations` et `Gras` :

- > `Poulets = M[MA$genes$Status=="gene",NonFiltres>0.60]`
- > `Weights = MA$weights[MA$genes$Status=="gene",NonFiltres>0.60]`
- > `ID = MA$genes$ID[MA$genes$Status=="gene"]`
- > `Name = MA$genes$Name[MA$genes$Status=="gene"]`
- > `Gras = targets[NonFiltres>0.60,]`



Tableaux de données

Données M, Weights, Annotations et Gras :

- > Poulets = MA\$M[MA\$genes\$Status=="gene",NonFiltres>0.60]
- > Weights = MA\$weights[MA\$genes\$Status=="gene",NonFiltres>0.60]
- > ID = MA\$genes\$ID[MA\$genes\$Status=="gene"]
- > Name = MA\$genes\$Name[MA\$genes\$Status=="gene"]
- > Gras = targets[NonFiltres>0.60,]

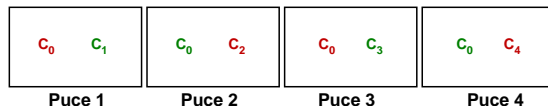
Suppression des gènes dont l'expression est mal mesurée :

- > PropBon = apply(Weights,1,mean)
- > Poulets = Poulets[PropBon>0.90,]
- > Weights = Weights[PropBon>0.90,]
- > ID = ID[PropBon>0.90]
- > Name = Name[PropBon>0.90]
- > rownames(Poulets) = ID
- > Annotations = data.frame(ID=ID,Name=Name)



Plan « en étoile »

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



Données:

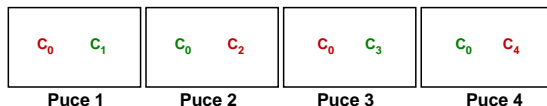
Gènes					Puce	Facteur
1	2	3	...	m		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	1	C_1
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	1	C_0
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	2	C_0
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	2	C_2
		⋮			⋮	⋮
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	4	C_0
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	4	C_4

= Y



Plan « en étoile »

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



Données:

Gènes					Puce	Facteur
1	2	3	...	m		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	1	C_1
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	2	C_2
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	3	C_3
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	4	C_4

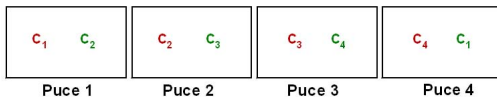
$$\square = M = Y - Y_0$$

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 ...



Données:

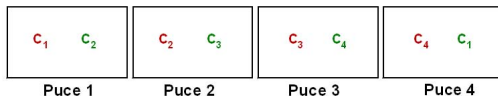
Gènes					Puce	Facteur
1	2	3	...	m		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	1	C_1
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	1	C_2
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	2	C_2
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	2	C_3
		⋮			⋮	⋮
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	4	C_4
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	...	<input type="checkbox"/>	4	C_1

= Y



Plan « en boucle »

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



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

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
 - Qualité des données
 - Effet fluorochrome
 - Interaction gène \times fluorochrome
 - Création des tableaux de données
 - 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

Tout est en place pour

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