Gaëlle Lelandais

gaelle.lelandais@univ-paris-diderot.fr

Méthodes d'analyses bioinformatiques d'expériences de puces à ADN

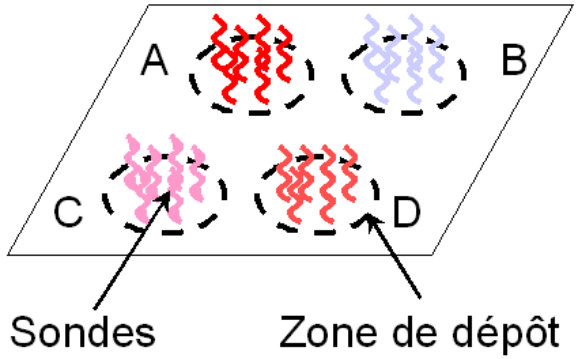
12 novembre 2008

Introduction et principe général

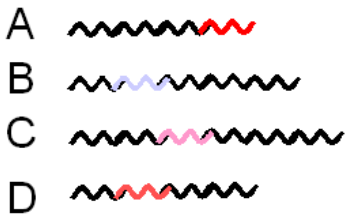
Zoom sur une puce à ADN

• Une puce à ADN est un <u>support rigide</u> sur lequel de courtes séquences d'ADN ont été déposées.

A Surface d'une puce à ADN



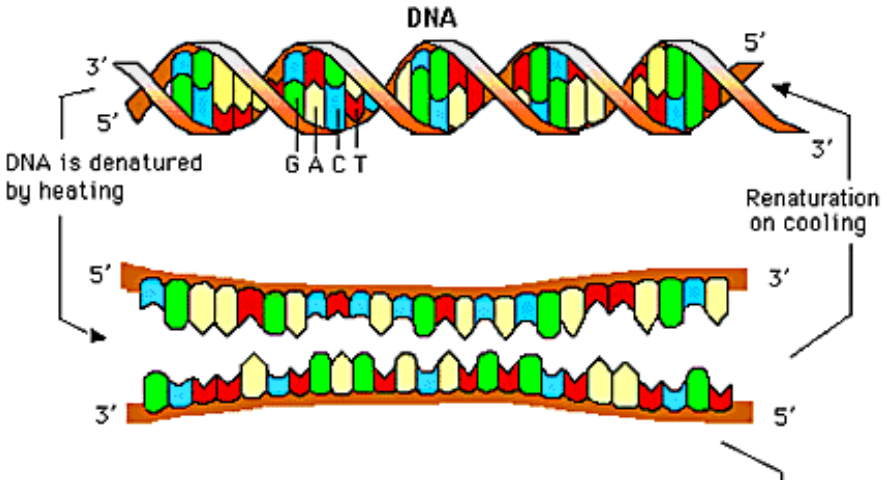
B Séquences complètes des gènes



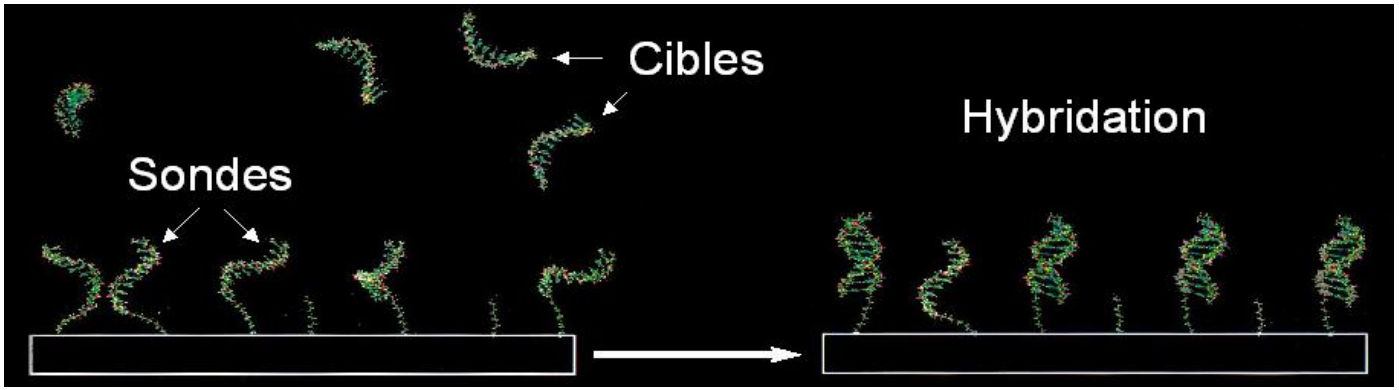
• Les sondes ont la particularité d'avoir été choisies de manière à être <u>spécifiques</u> d'un seul et unique gène.

Principe d'hybridation à la base du fonctionnement des puces à ADN

• Le fonctionnement des puces à ADN repose sur le principe d'hybridation qui stipule que deux fragments d'acides nucléiques complémentaires peuvent s'associer et se dissocier de façon réversible.



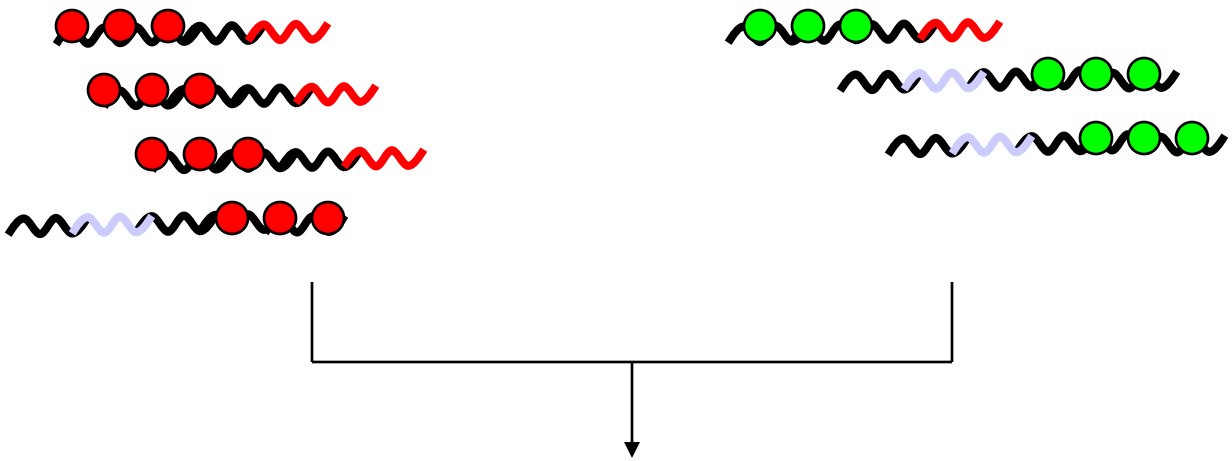
• Les acides nucléiques d'un mélange à tester s'hybrident au niveau des sondes qui leurs sont spécifiques.

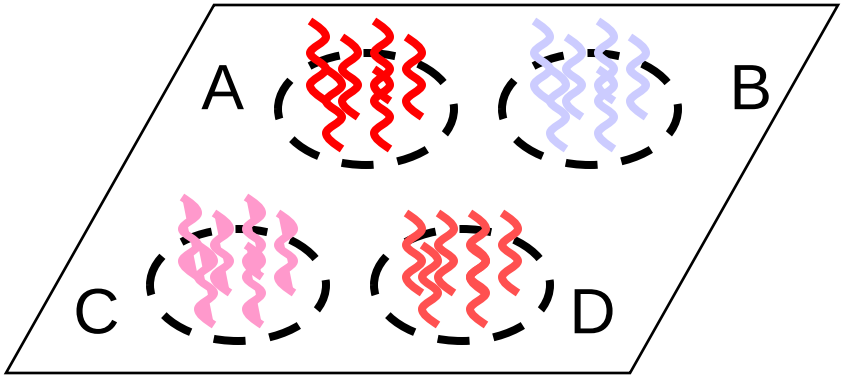


Comment ça marche ?

Condition 1 :

Condition 2 :



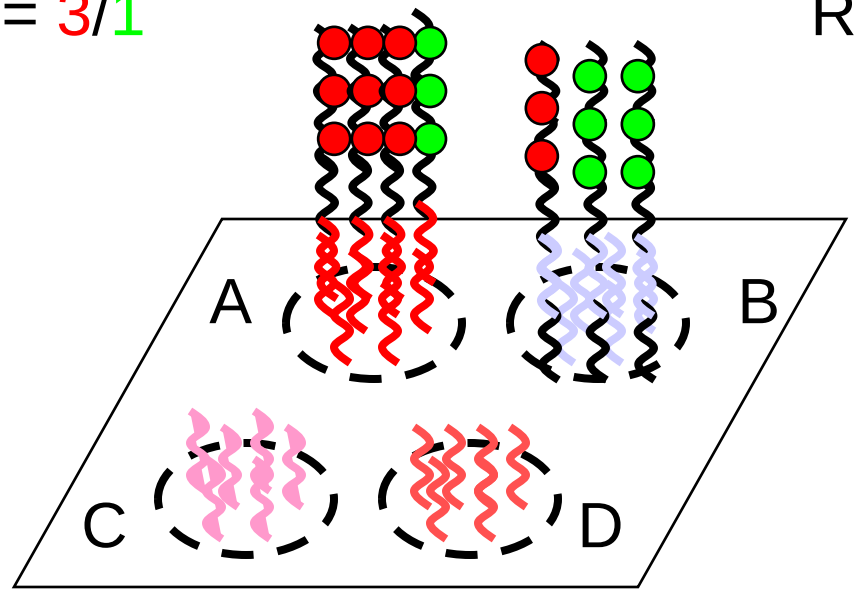


Estimation d'un rapport d'expression (le ratio)

✓ A chaque gène est associé une <u>valeur de ratio</u> :

Ratio = 3/1

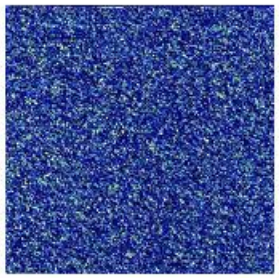
Ratio = 1/2



⇨ Les puces à ADN permettent de <u>comparer</u> la composition de deux populations d'ARNm

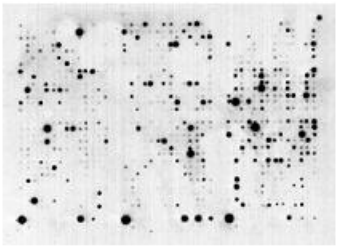
Les principaux types de puces à ADN

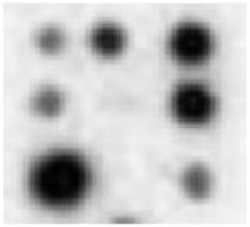
Puces Affymetrix



Taille : 1,28cm x 1,28cm

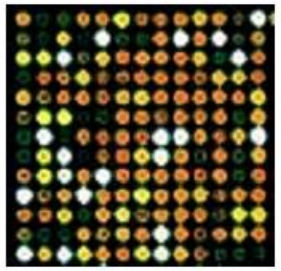
Filtres à haute densité (macroarrays)





Taille : 12cm x 8cm

Lames de verre (microarrays)

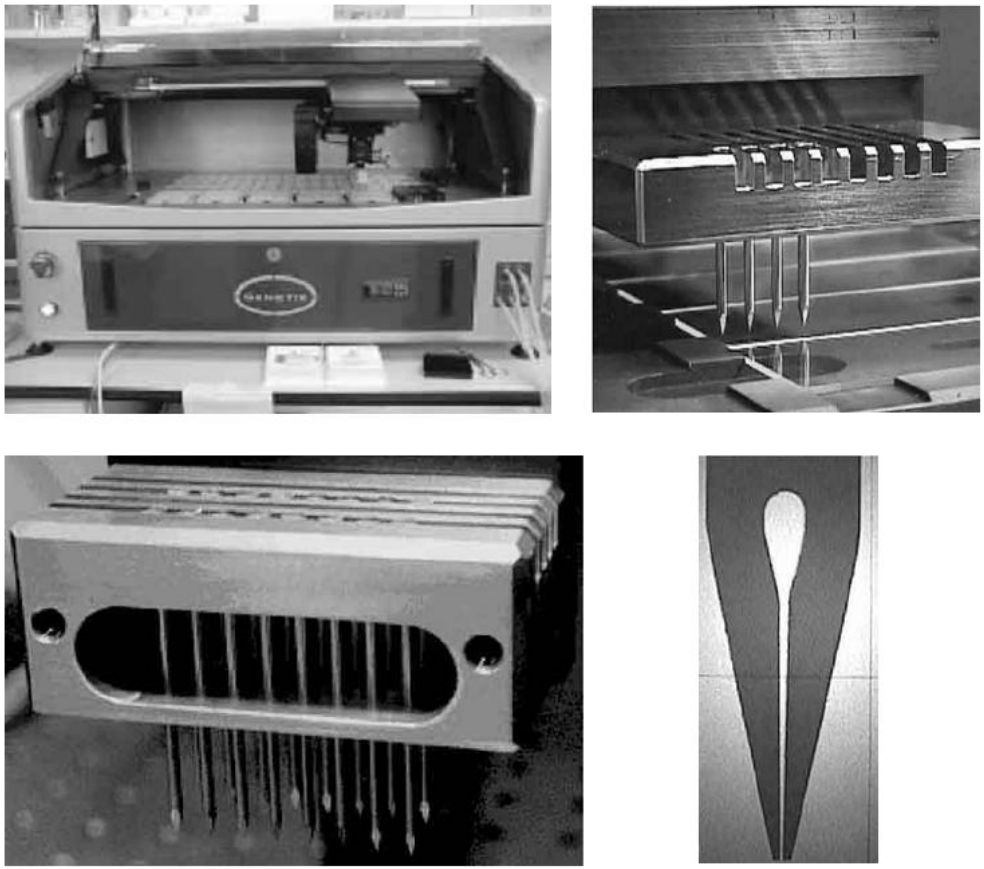


Taille : 5,4cm x 0,9cm

Et bien d'autres …

Fabrication d'une puce à ADN

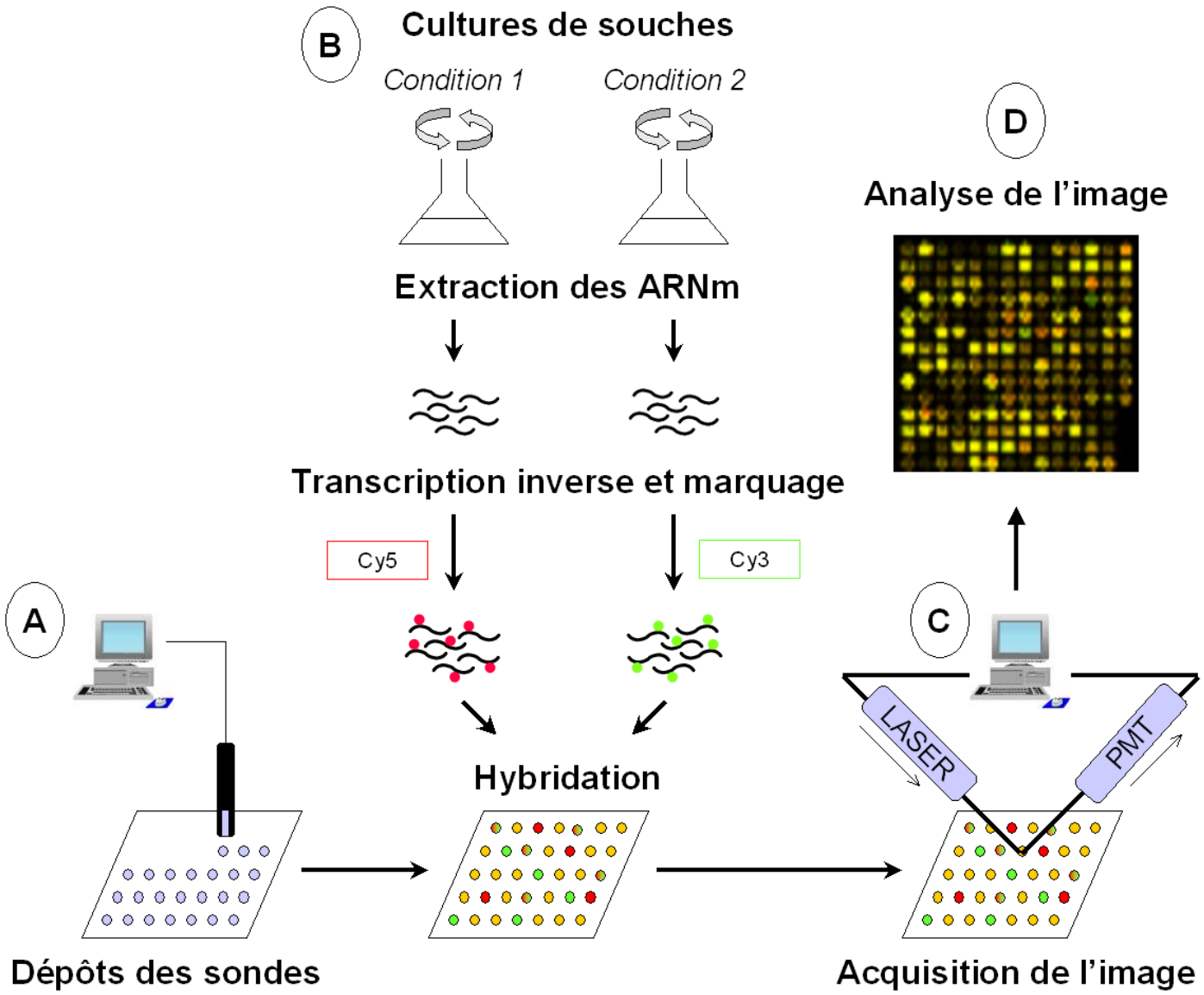
• La fabrication d'une puce consiste à fixer des sondes sur un support rigide en des endroits bien déterminés.



D'après « Microarray Bioinformatics », Dov Stekel

• Des micro-gouttes de solutions d'ADN sont déposées par un robot spécialisé.

Vue générale d'une expérience de puce à ADN



Partie 1 :

Analyse de l'image

Partie 2 :

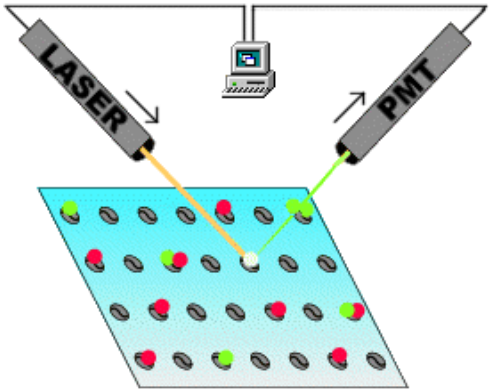
Correction des biais expérimentaux

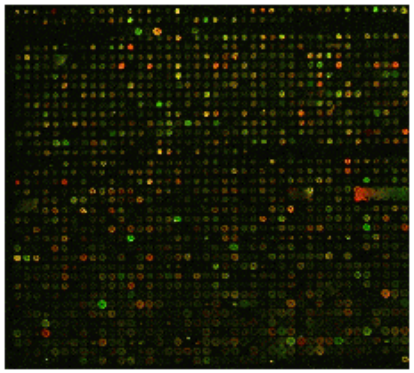
Analyse de l'image

Acquisition des images

• Consiste à détecter la fluorescence émise à la surface de la lame.

• Utilisation d'un lecteur de fluorescence capable d'exciter les fluorochromes et d'en recueillir l'émission fluorescente.

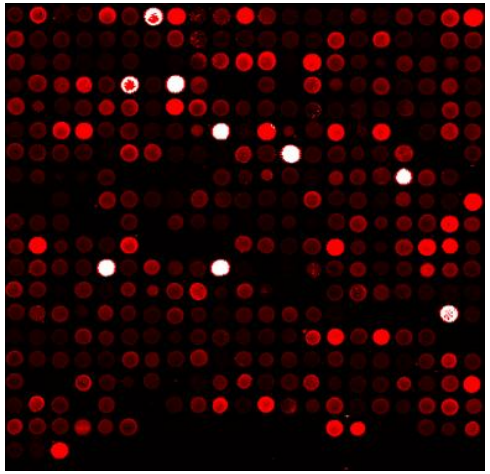




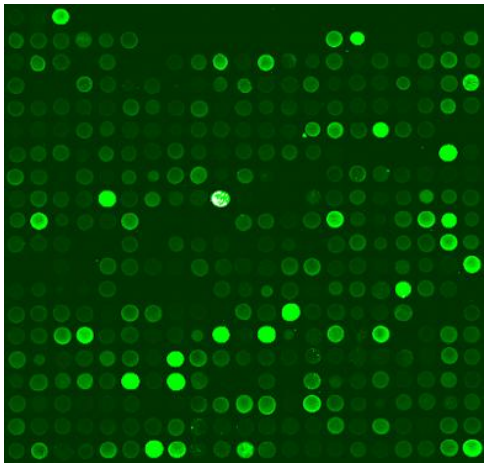
GenePix 4000 - Axon Instrument

(http://www.axon.com/GN_Genomics.html)

Obtention de l'image



Longueur d'onde du Cy5
(635 nm)



Longueur d'onde du Cy3
(532 nm)

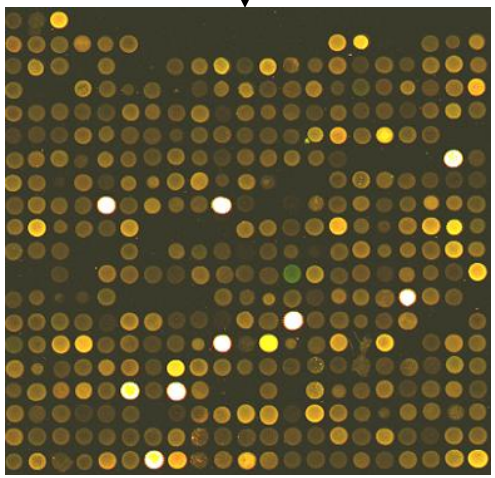
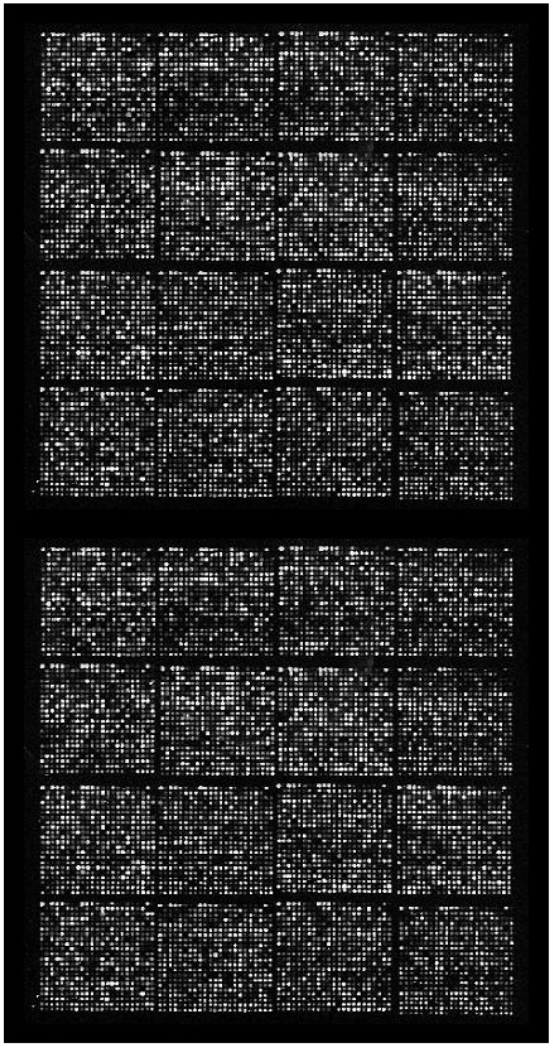


Image finale

En réalité …

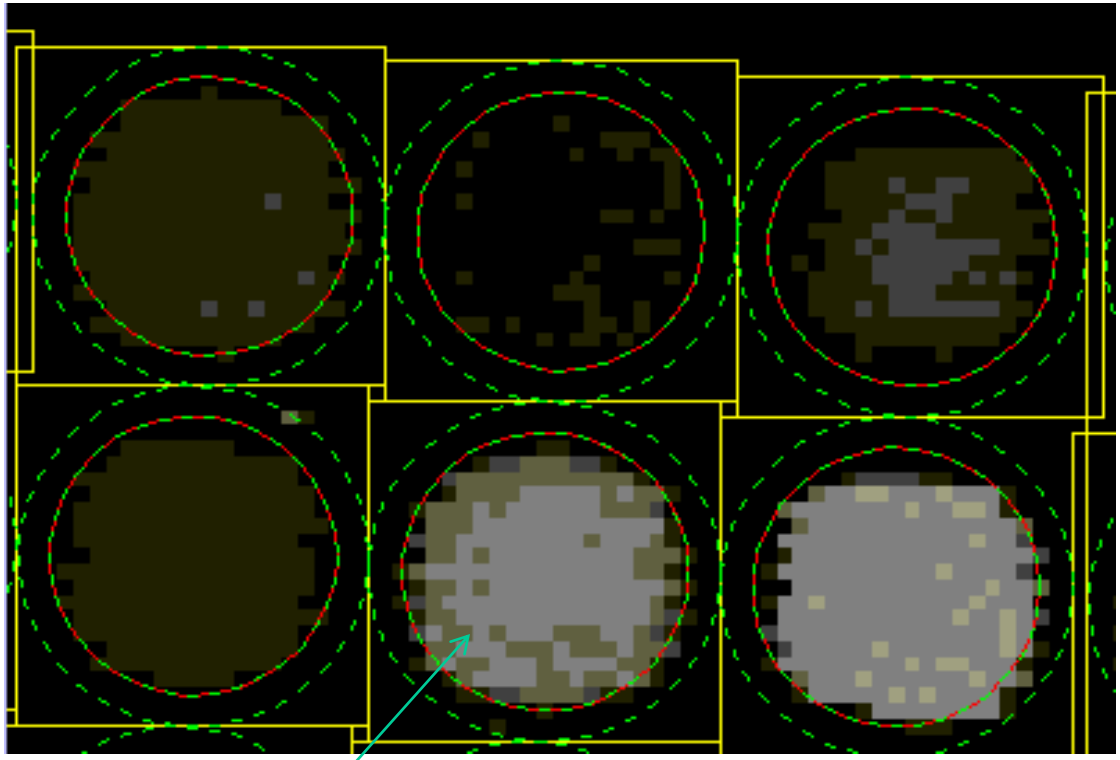


On a une image en niveaux de gris.

Le rouge et le vert sont des « fausses couleurs » !

Les dépôts sont arrangés à la surface de la puce en deux ensembles de 4 x 4 blocs.

Une image = des pixels de différents niveaux de gris

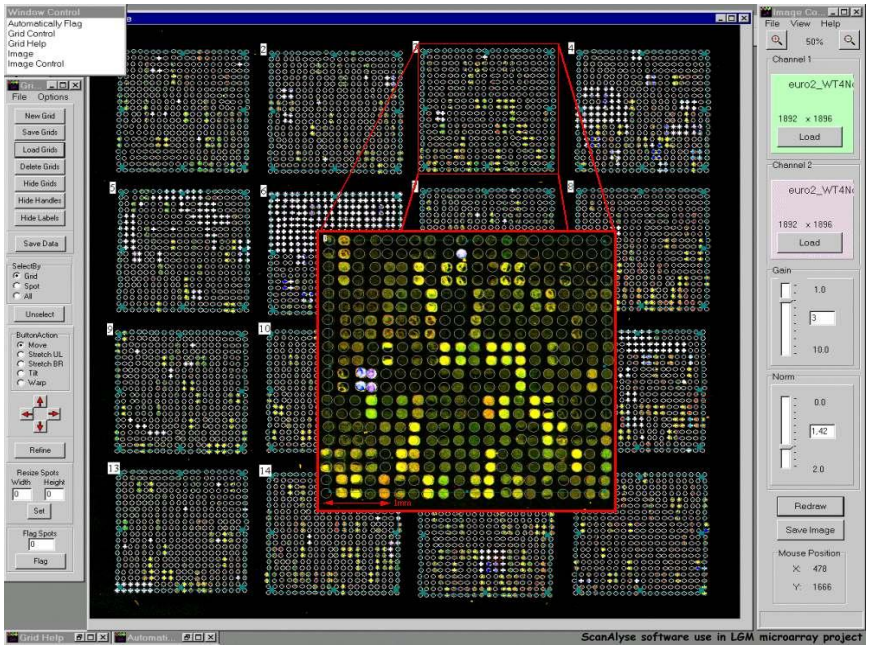


A chaque pixel correspond une intensité de niveaux de gris (entre 0 et 256)

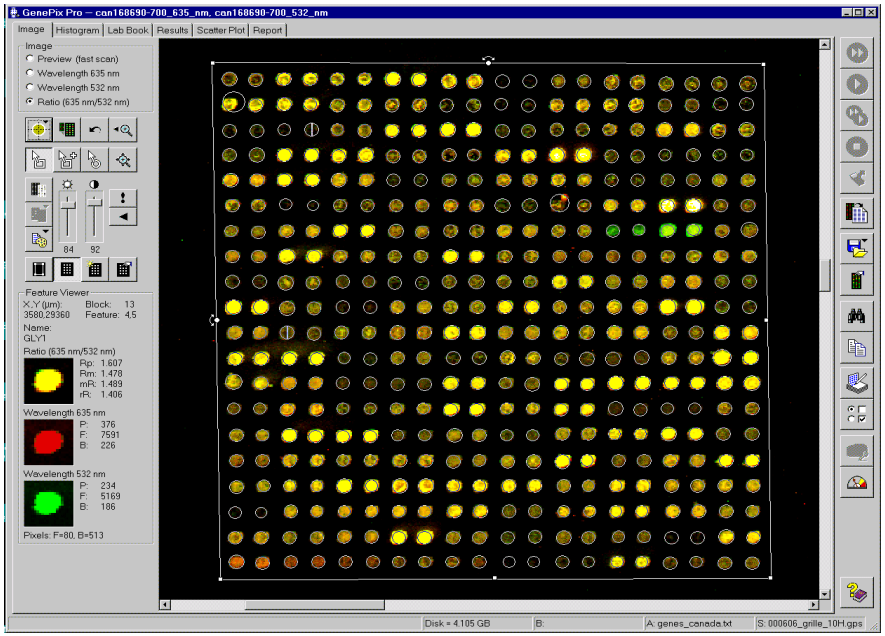
Principes généraux de l'analyse d'image

• <u>But</u> : Convertir l'image en valeurs numériques quantifiant l'expression des gènes

Il existe des logiciels d'analyse d'image …



ScanAlyse
(M. Eisen Stanford University)



Genepix 3.0
(Axon software)

Les différentes étapes de l'analyse d'image

1 – Localisation des dépôts sur la lame

Pour chaque spot :

2 – Délimitation des pixels correspondant à la zone de dépôt

3 – Délimitation des pixels pour l'estimation du bruit de fond

4 – Calcul de l'intensité globale de fluorescence

Sur l'ensemble de la lame :

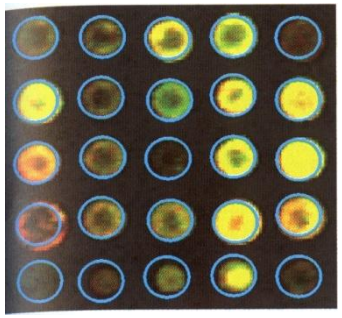
5 – Identification des dépôts abîmés par des artéfacts

1– Localisation des dépôts sur la lame

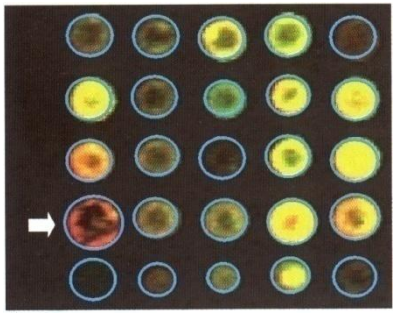
- Il est important d'assigner à chaque dépôt un identifiant de gène correct !

2 — Délimitation des pixels correspondant à la zone de dépôt

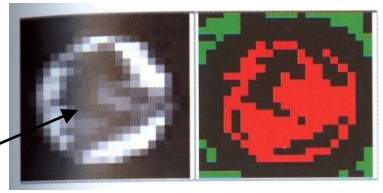
• Différentes méthodes sont possibles :



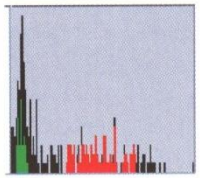
Cercles à diamètre fixe



Cercles à diamètre variable



Intensité variable



Méthode par histogramme

D'après « Microarray Bioinformatics », Dov Stekel

3 – Délimitation des pixels pour l'estimation du bruit de fond

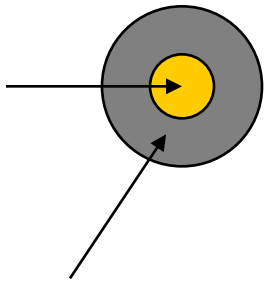
Les signaux observés ont deux composantes :

Fluorescence issue d'une hybridation spécifique

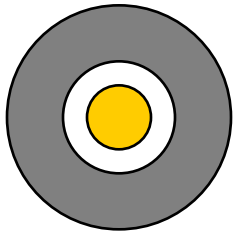
Fluorescence non spécifique: **Bruit de fond**

• Analyse des pixels localisés à proximité de la zone de dépôt :

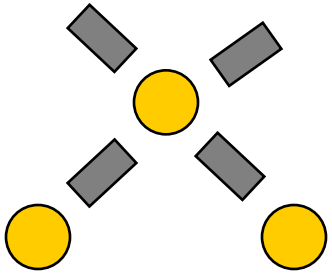
Méthode « ScanAlyze »



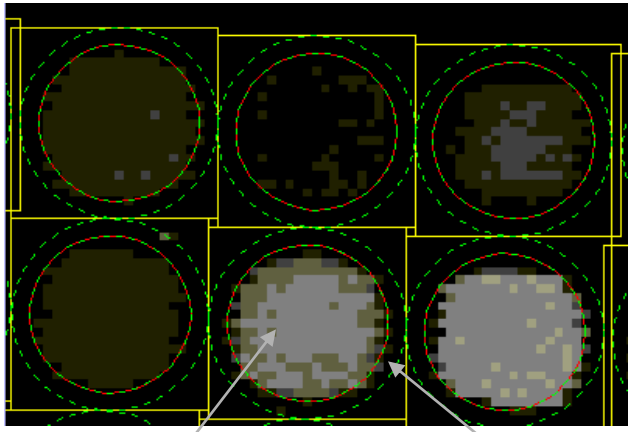
Méthode « ImaGene »



Méthode « GenePix »



4 – Calcul de l'intensité globale de fluorescence



Intensité à l'intérieur du dépôt

Mesure du bruit de fond

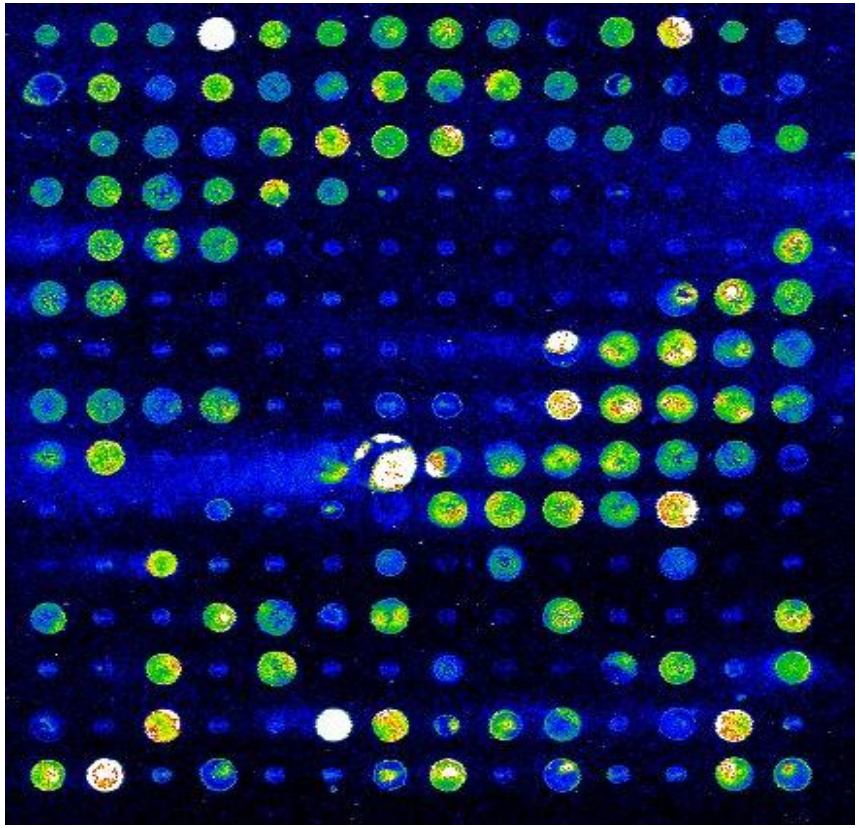
Intensité par pixel → **Moyenne ou médiane ?** → Intensité globale du spot

Critère qualité : taille du dépôt, déviation standard…

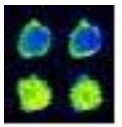
5 – Annotation des dépôts non conformes

• Quelques exemples :

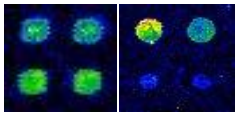
500 µ



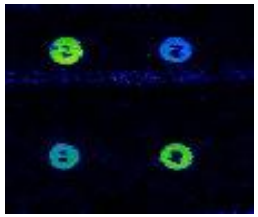
Spots diffus



Spots hétérogènes



Spots en « beignets »

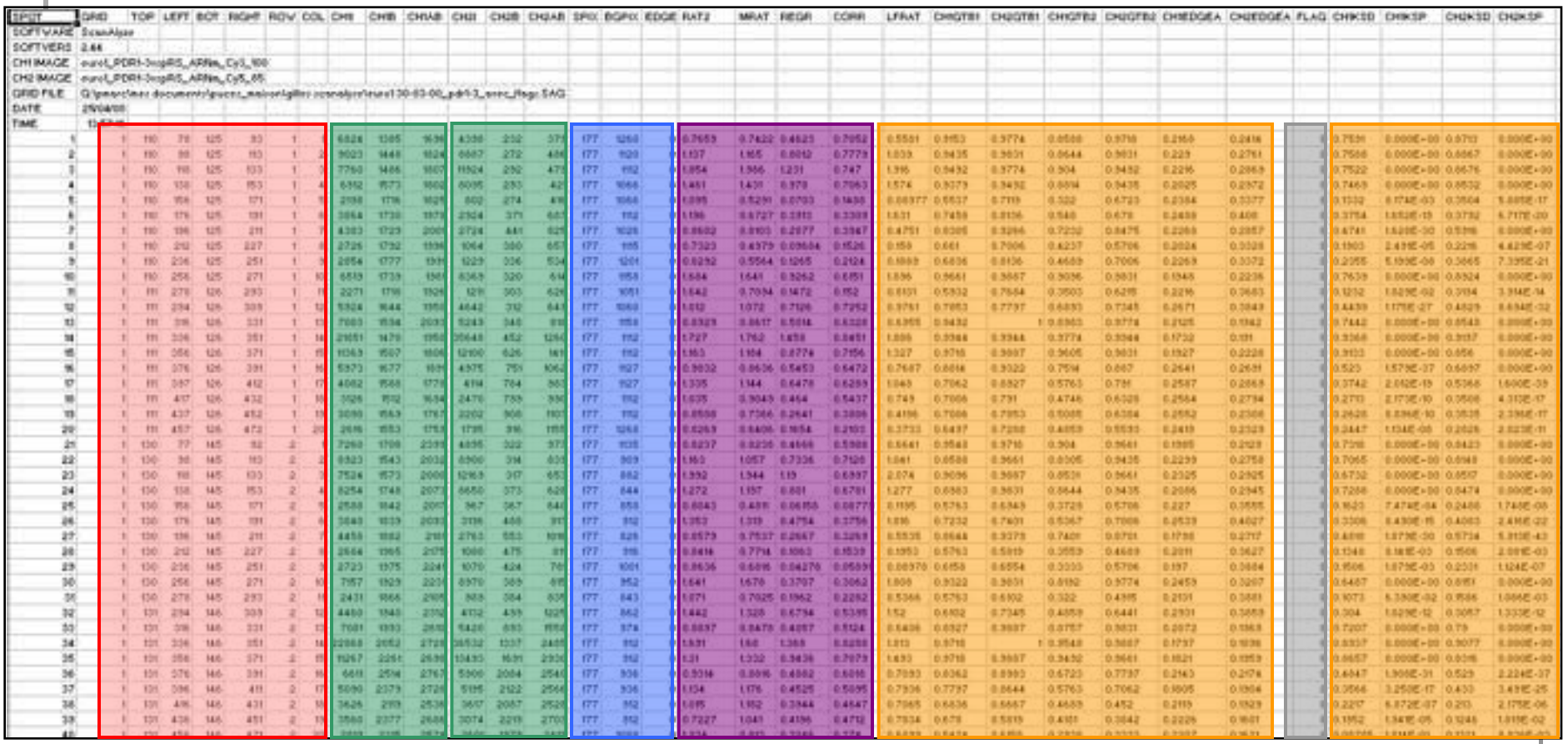


La récupération des résultats de l'analyse d'image

- Exemple d'un fichier de sortie du logiciel GENEPIX (.gpr) :

Annotations

Position | Intensités (Brutes, + BF) | Taille | Information (homogénéité) | Statistiques



Le « nettoyage des données »

- Élimination des gènes dont l'annotation est différente de 0

  (éventuellement, retournez voir l'image originale)

- Filtrage sur les intensités :

  Saturation du scanner

  Écart avec le bruit de fond trop faible

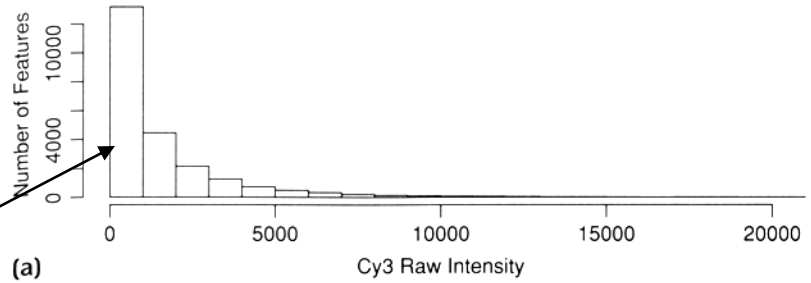
- Transformation logarithmique

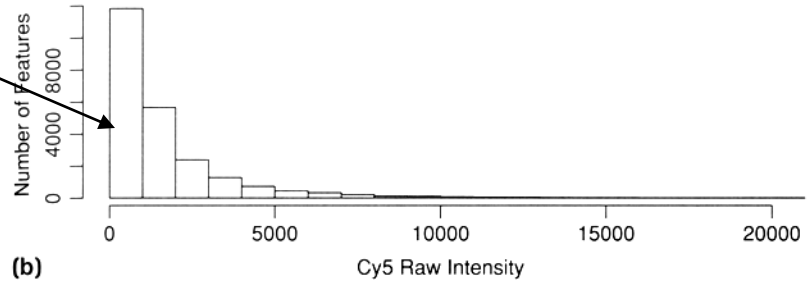
La transformation logarithmique

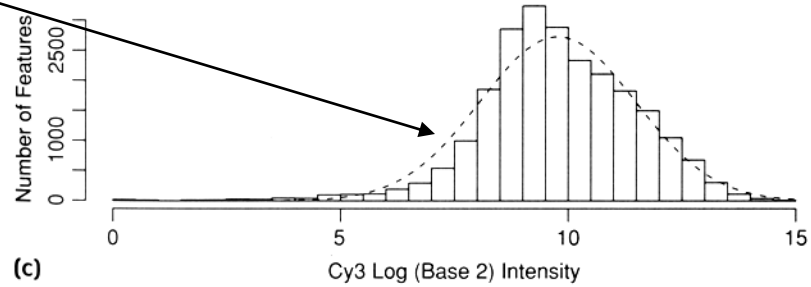
• Effet sur la distribution des intensités :

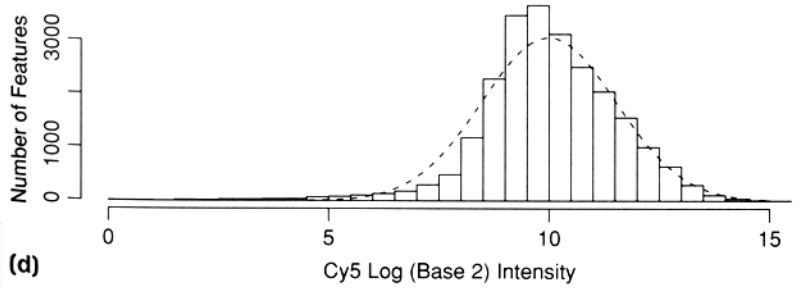
La plupart des intensités mesurées sont faibles

Distribution normale







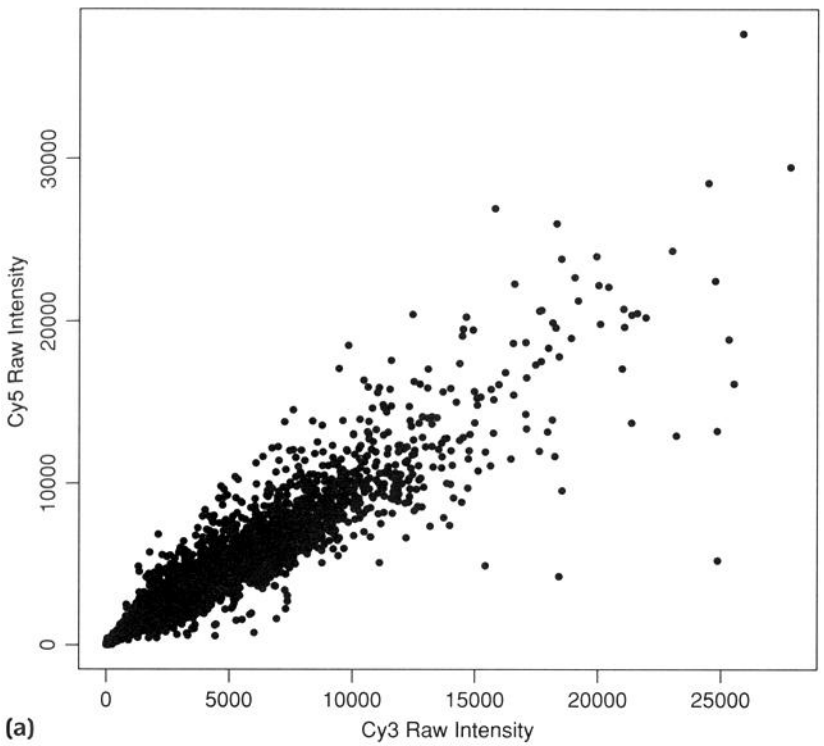


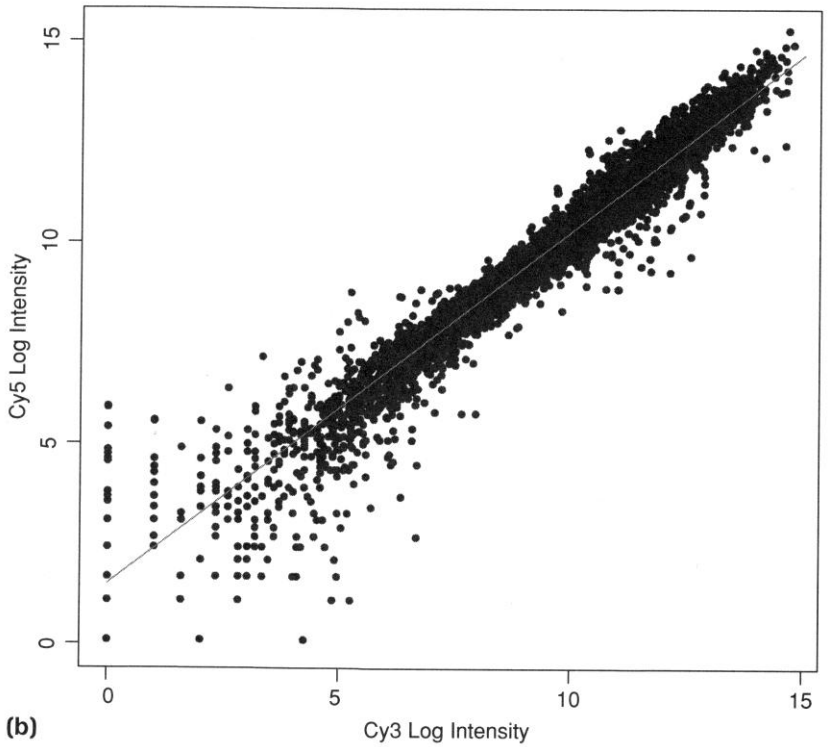
• Recentrage de la distribution :

Facilite l'utilisation des statistiques…

D'après « Microarray Bioinformatics », Dov Stekel

La transformation logarithmique



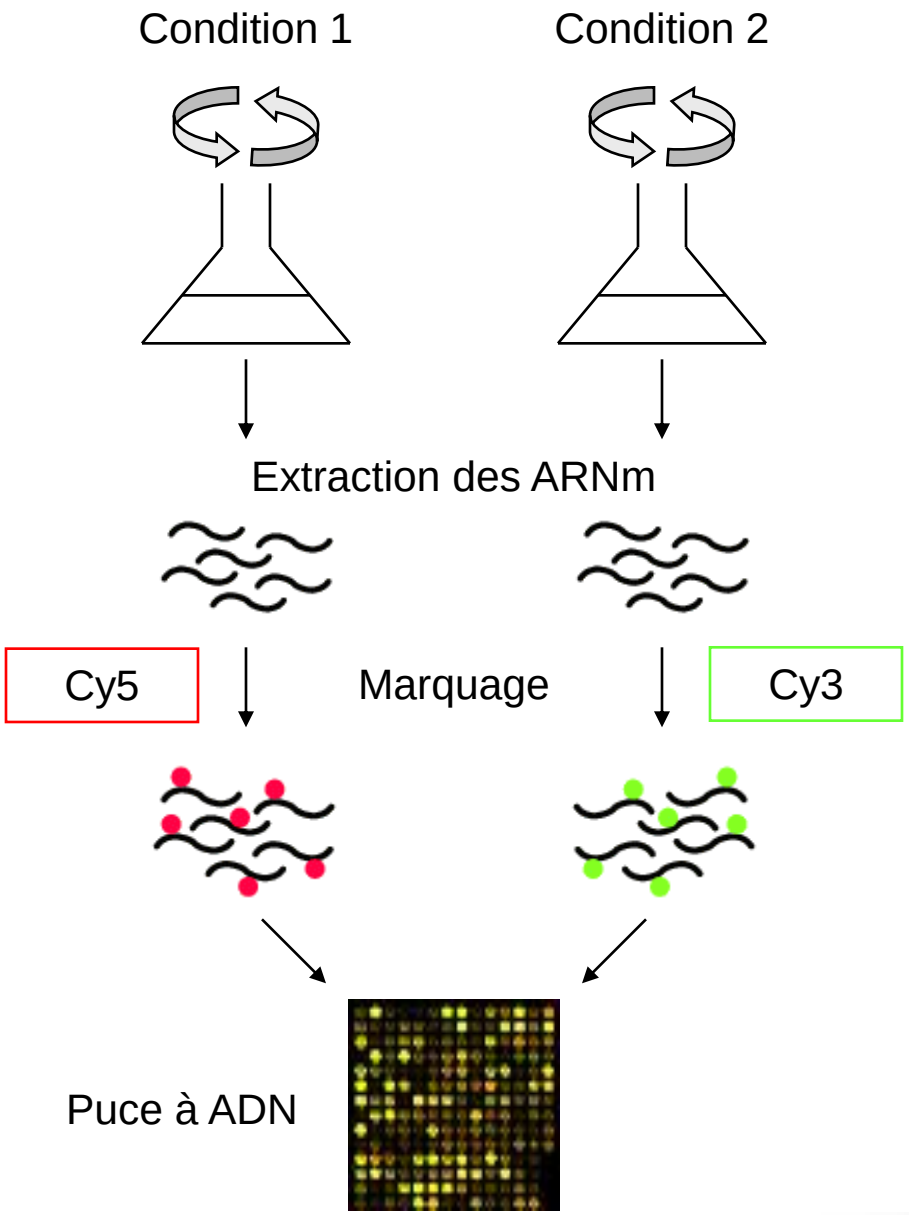


• Augmentation de la variabilité avec l'intensité

• Variabilité constante

Remarque : On choisira souvent le logarithme en base 2 [log2(2) = 1]

Le logarithme du ratio, quelques repères …



Analyse du rapport R/G

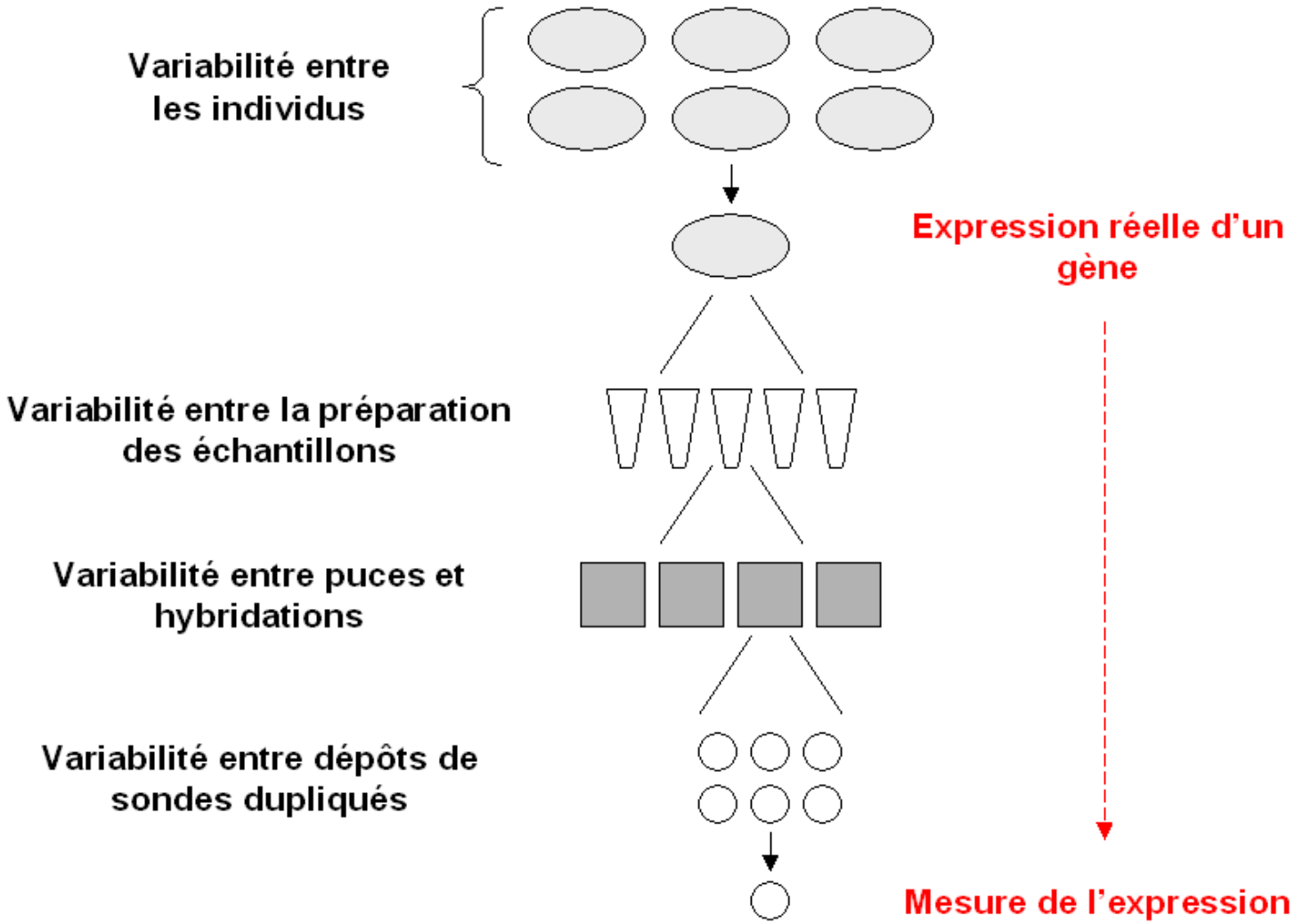
Log2(R/G)

Si R > G alors R/G > 1
et log2(R/G) > 0
(= Induction du gène)

Si R < G alors R/G < 1
et log2(R/G) < 0
(= Répression du gène)

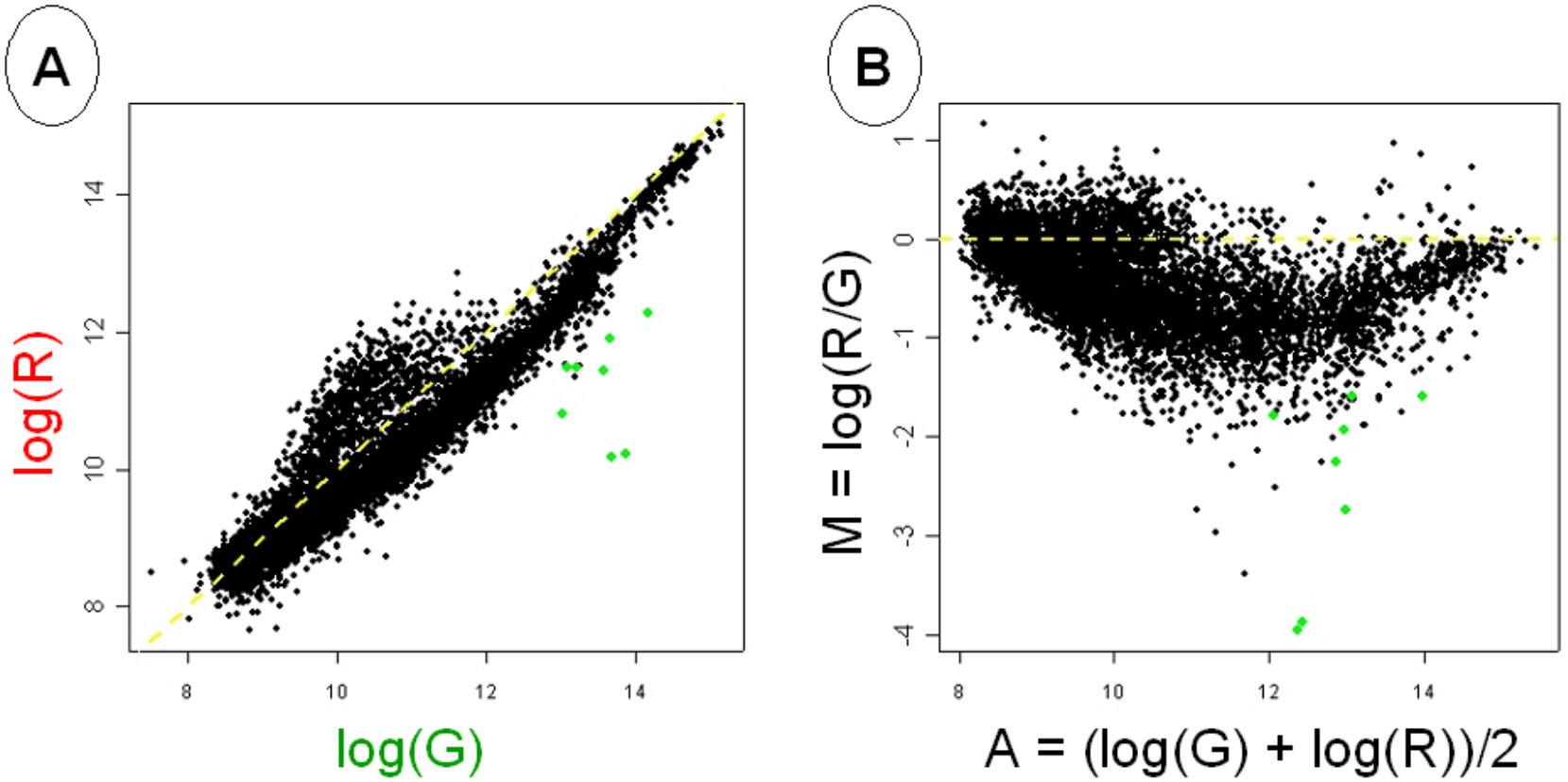
Correction des biais expérimentaux

Les sources de variabilité dans une expérience de puce à ADN



Mise en évidence des biais expérimentaux (1/3)

• Représentation des gènes sous la forme d'un nuage de points :

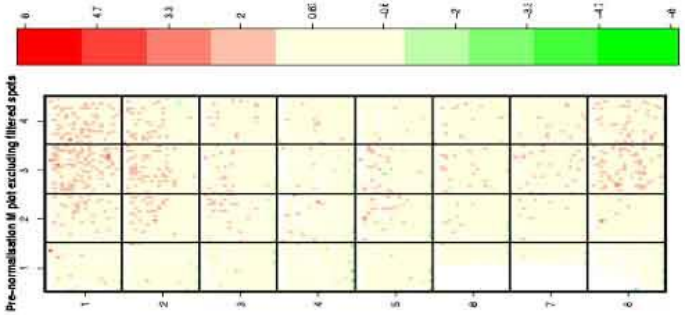


→ Visualisation de l'existence d'éventuels biais systématiques entre les mesures R et G.

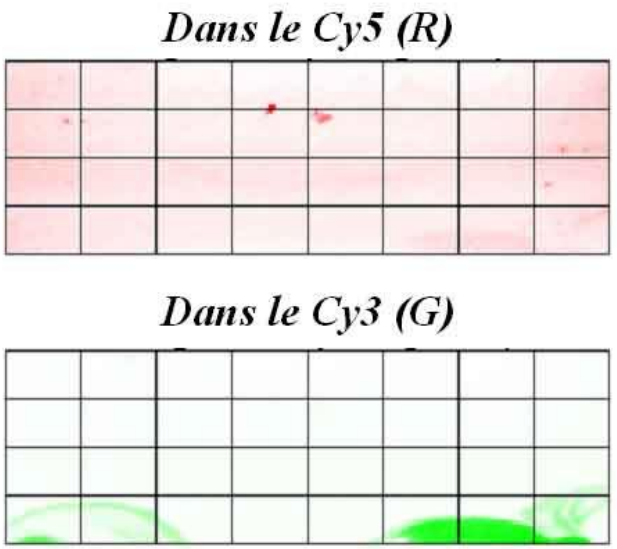
Mise en évidence des biais expérimentaux (2/3)

• Représentation de la répartition spatiale des données sur la lame :

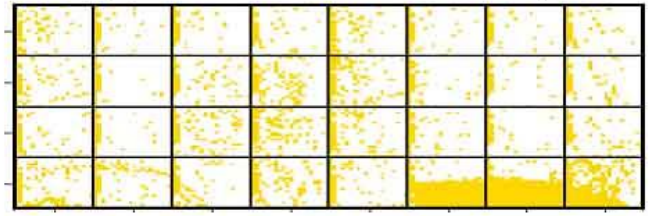
A **Répartition des valeurs de log(*R*/*G*)**



B **Répartition du bruit de fond**



C **Dépôts exclus de l'analyse**

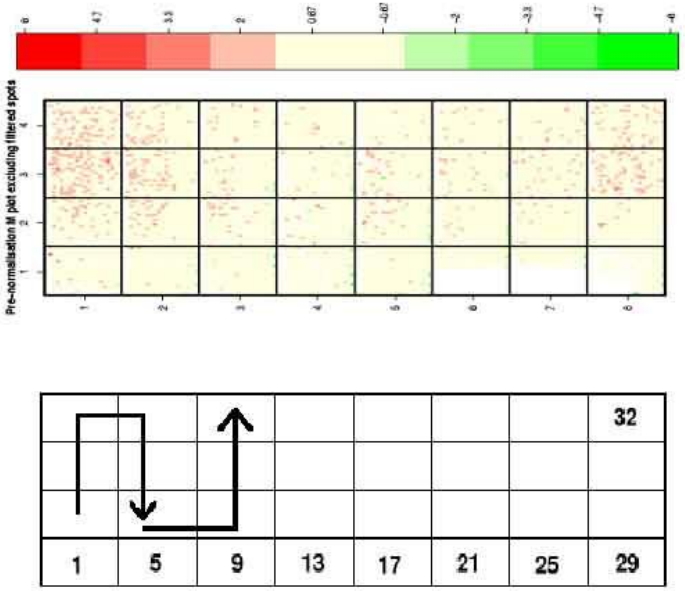


→ Ce type de représentation permet de détecter des <u>biais systématiques</u> parfois inattendus !

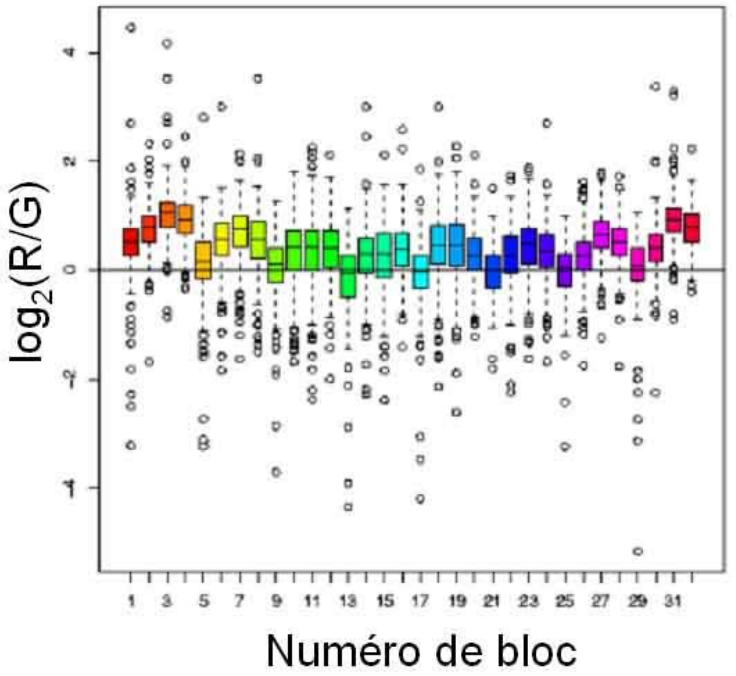
Mise en évidence des biais expérimentaux (3/3)

• Représentation de la répartition spatiale des données sur la lame :

A **Répartition des valeurs de log(R/G)**

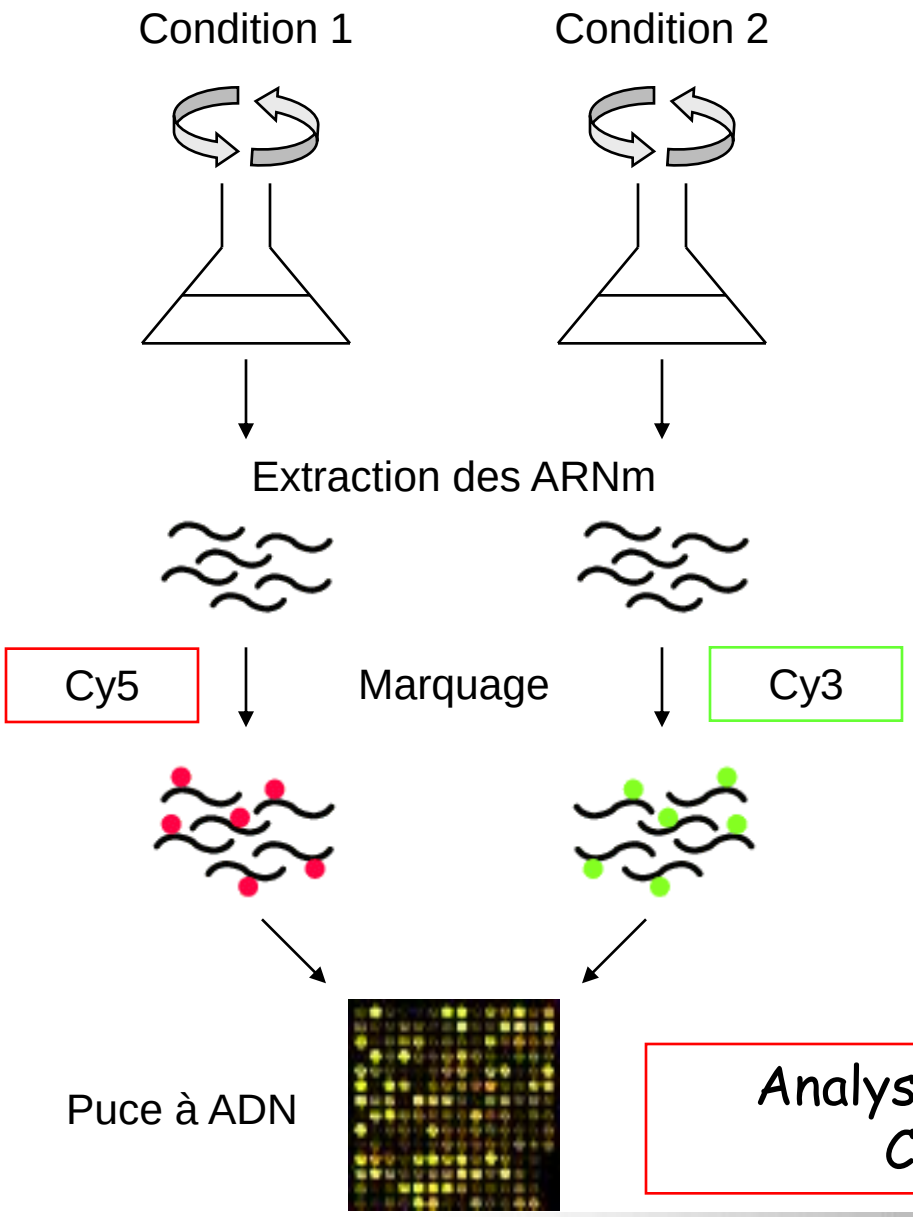


B **Représentation Boxplots**



→ Un biais important est détecté pour les blocs 3,4 et 31, 32.

Rappel du principe des puces à ADN



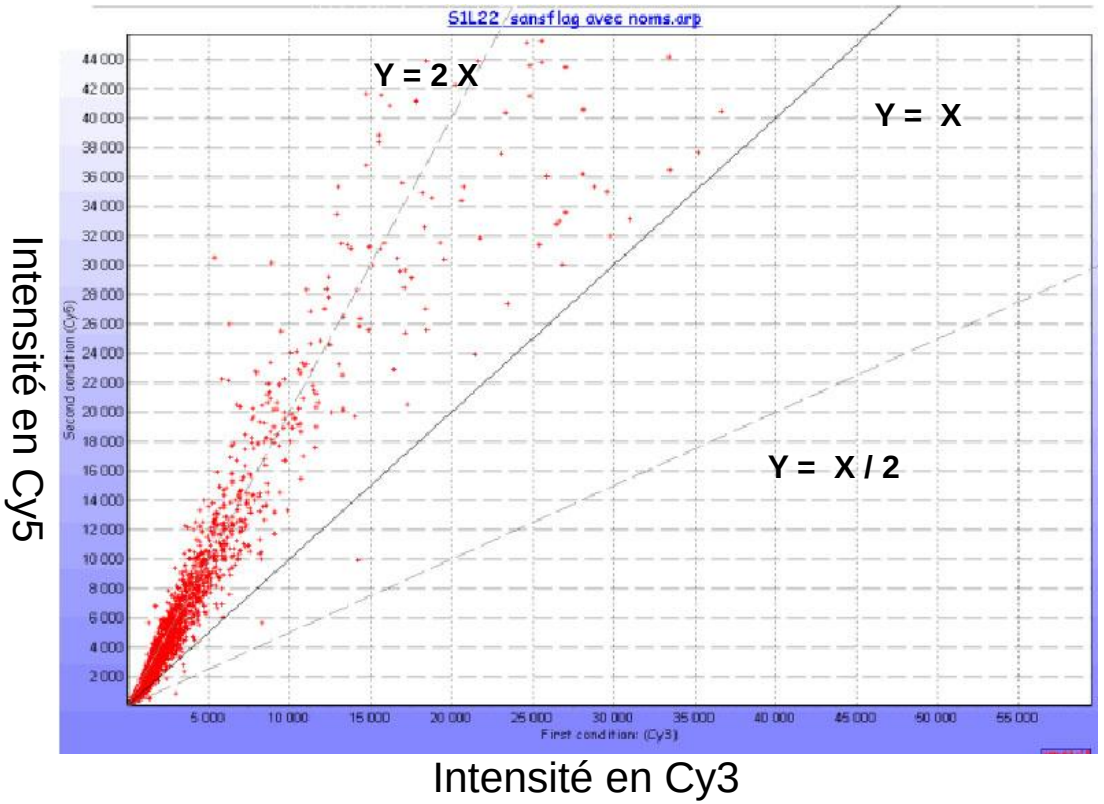
La normalisation inter-canaux

- **Pourquoi normaliser ?**

Marquage par des fluorochromes différents lors de réactions chimiques indépendantes

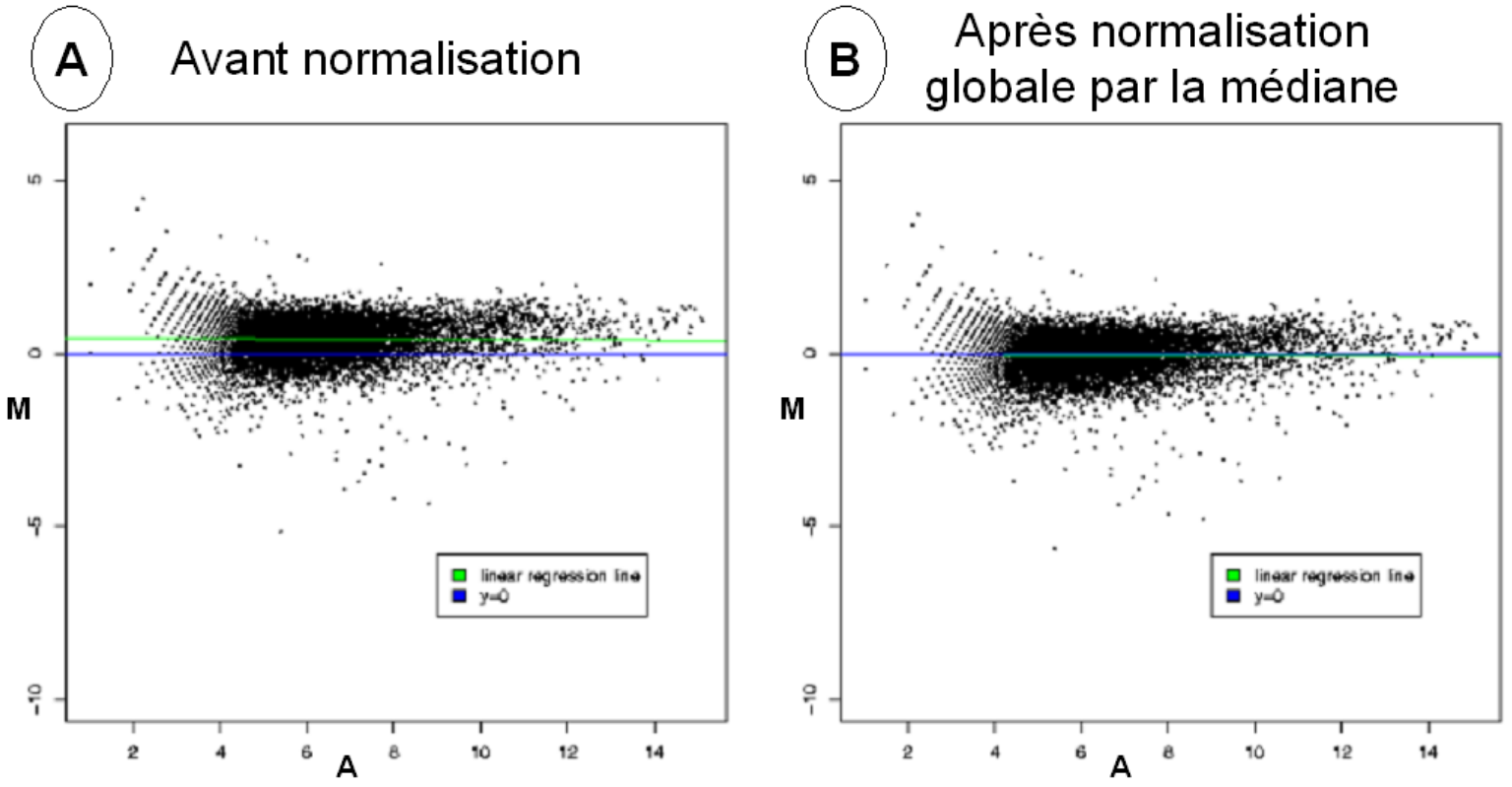
Les intensités de fluorescence sont mesurées par des lasers différents dans des longueurs d'onde différentes

- **Même lot d'ARN marqué en Cy5 et Cy3 :**

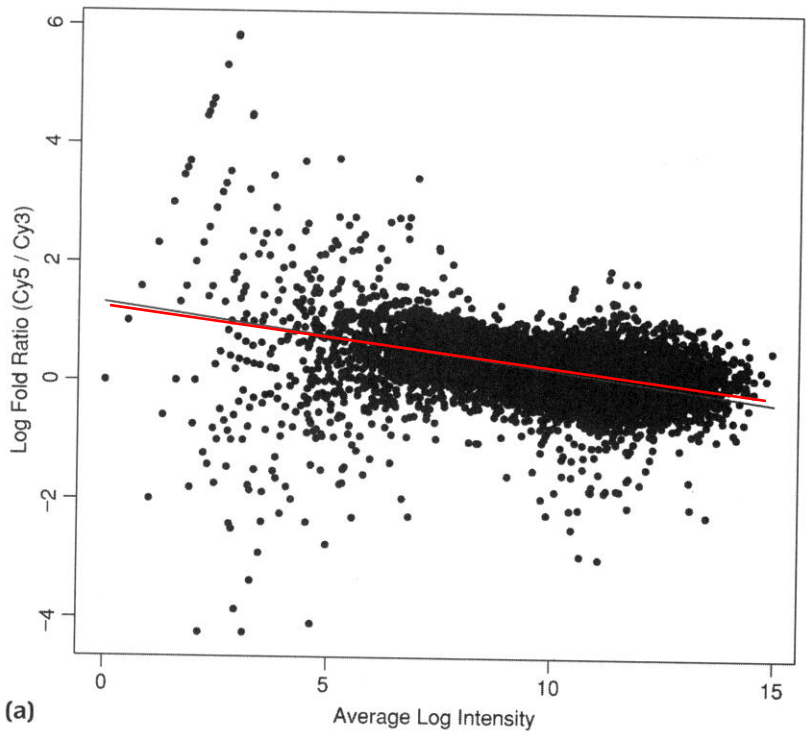


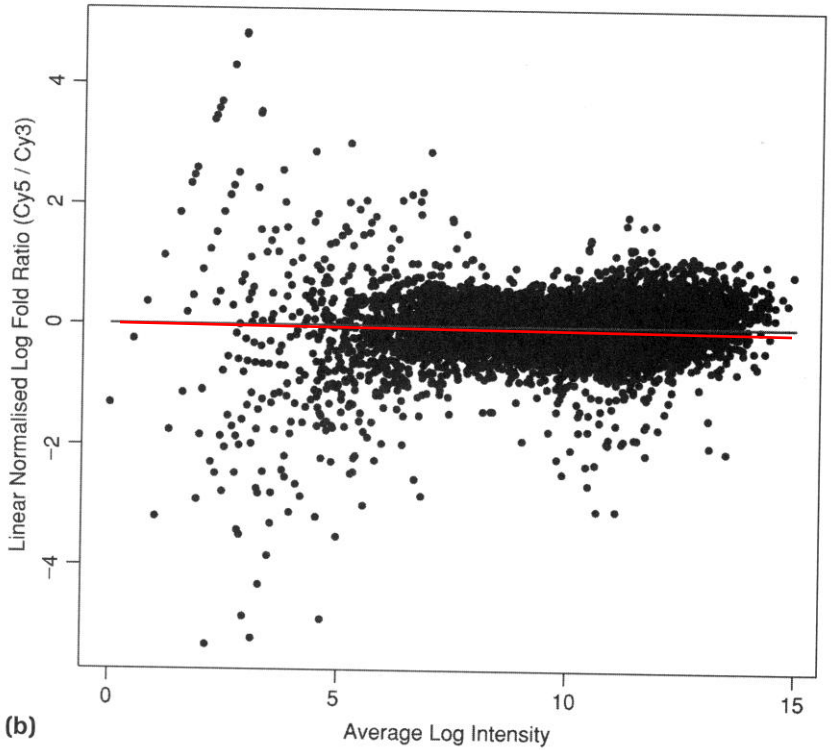
Normalisation par la médiane ou la moyenne

• Il s'agit d'une méthode de normalisation globale qui repose sur l'hypothèse que les intensités R et G sont reliées par un facteur k constant.



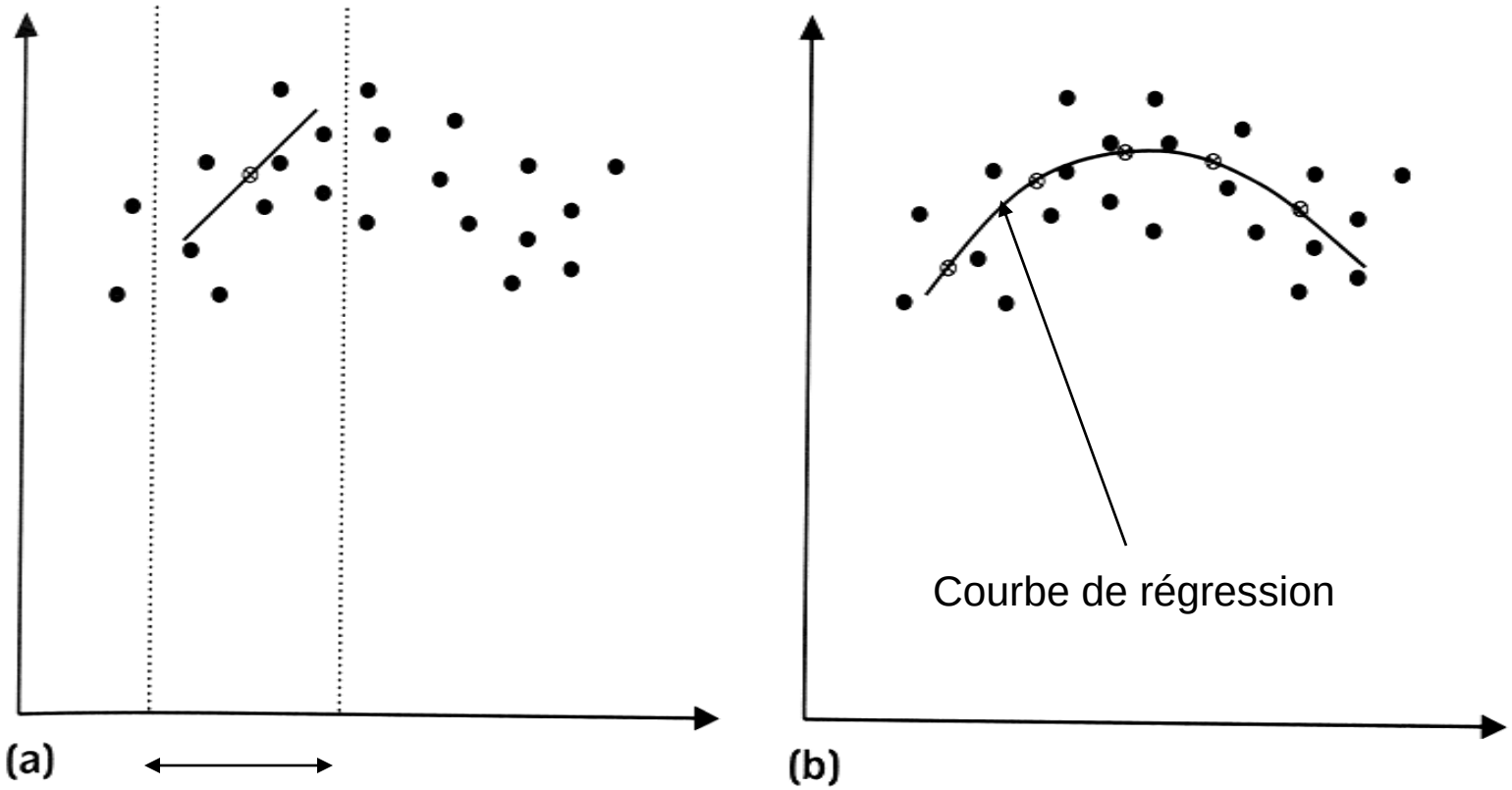
Normalisation par régression linéaire



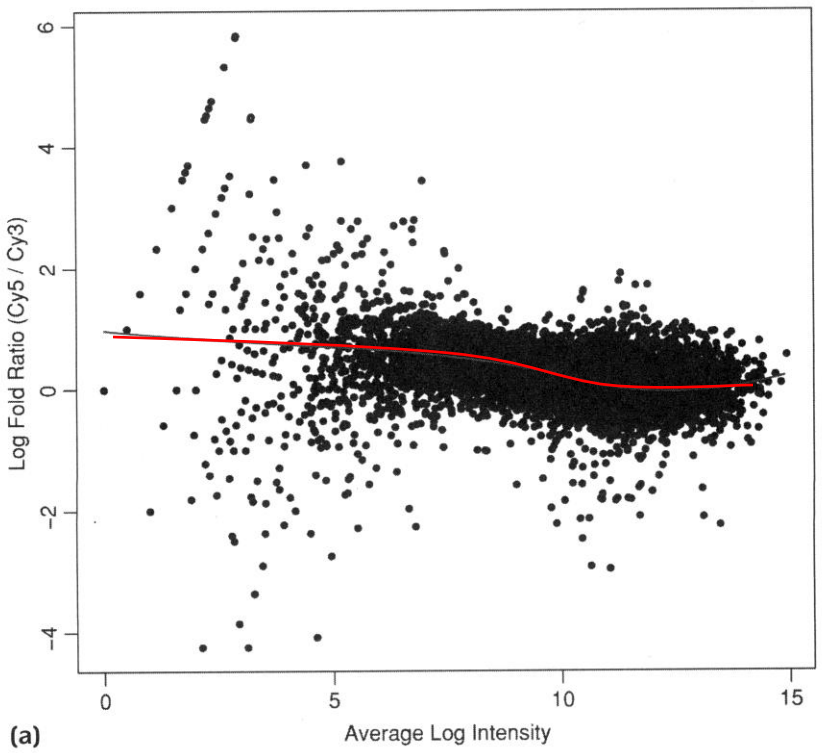


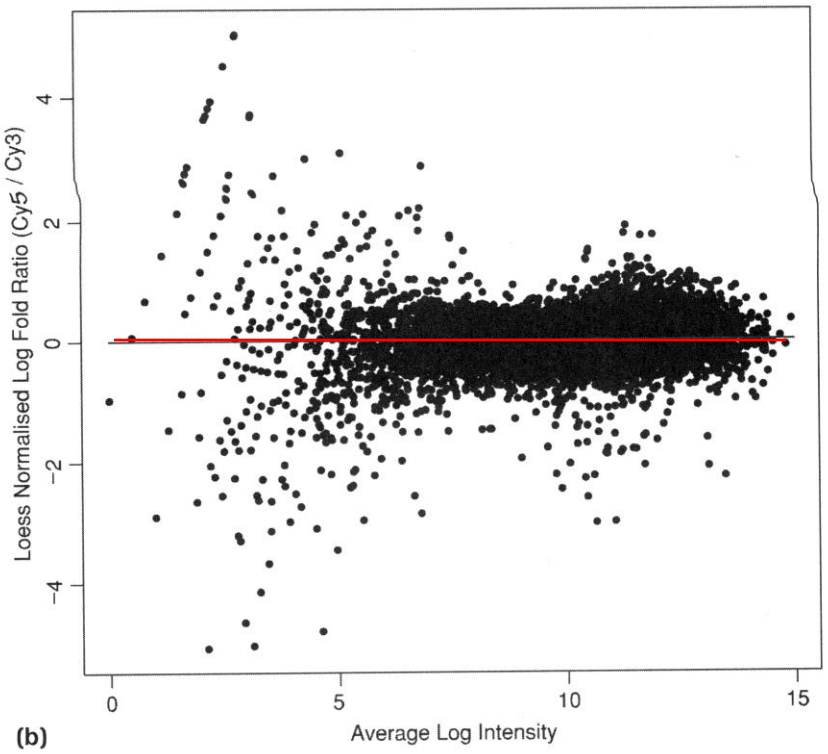
Normalisation intensité dépendante : méthode loess (1/2)

$\log_2 R/G \rightarrow \log_2 R/G - c(A) = \log_2 R/(k(A)G)$.

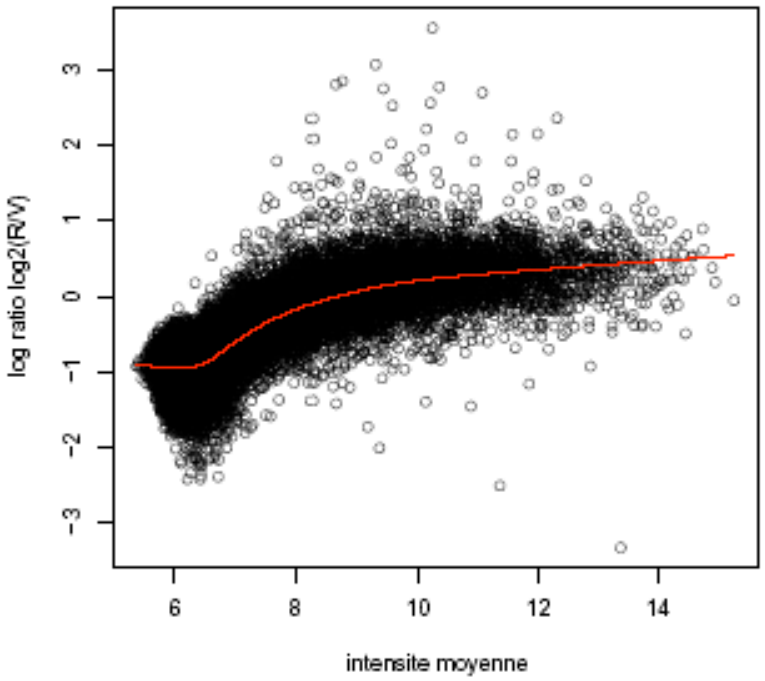


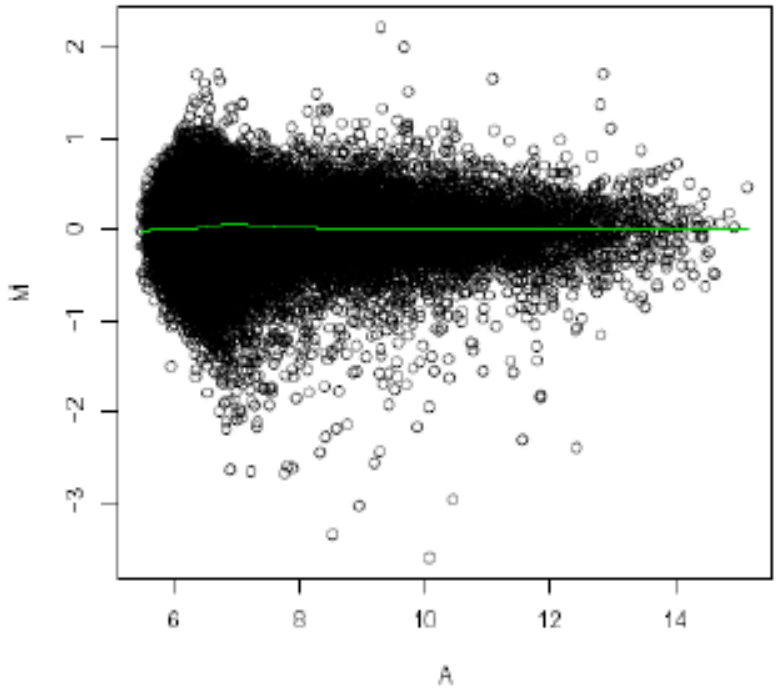
Normalisation intensité dépendante : méthode loess (2/2)



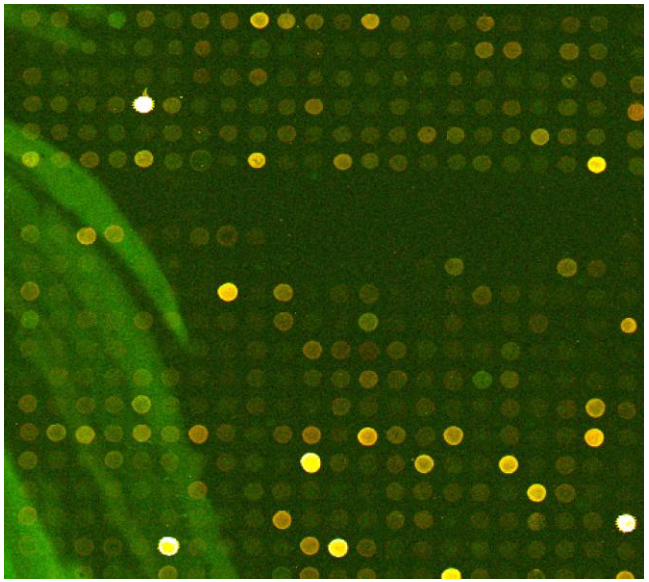


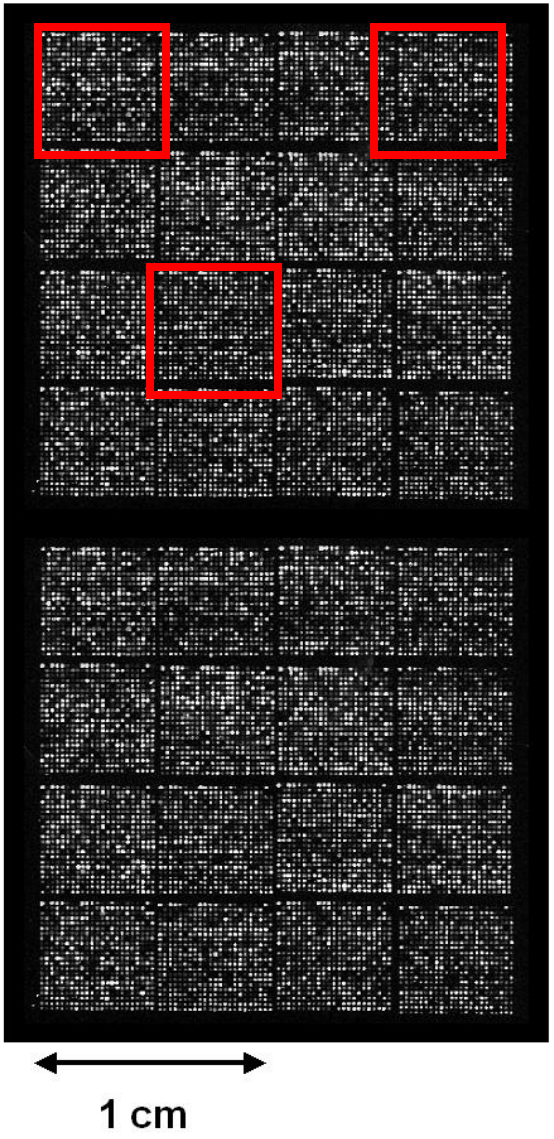
Normalisation intensité dépendante : méthode loess (2/2)





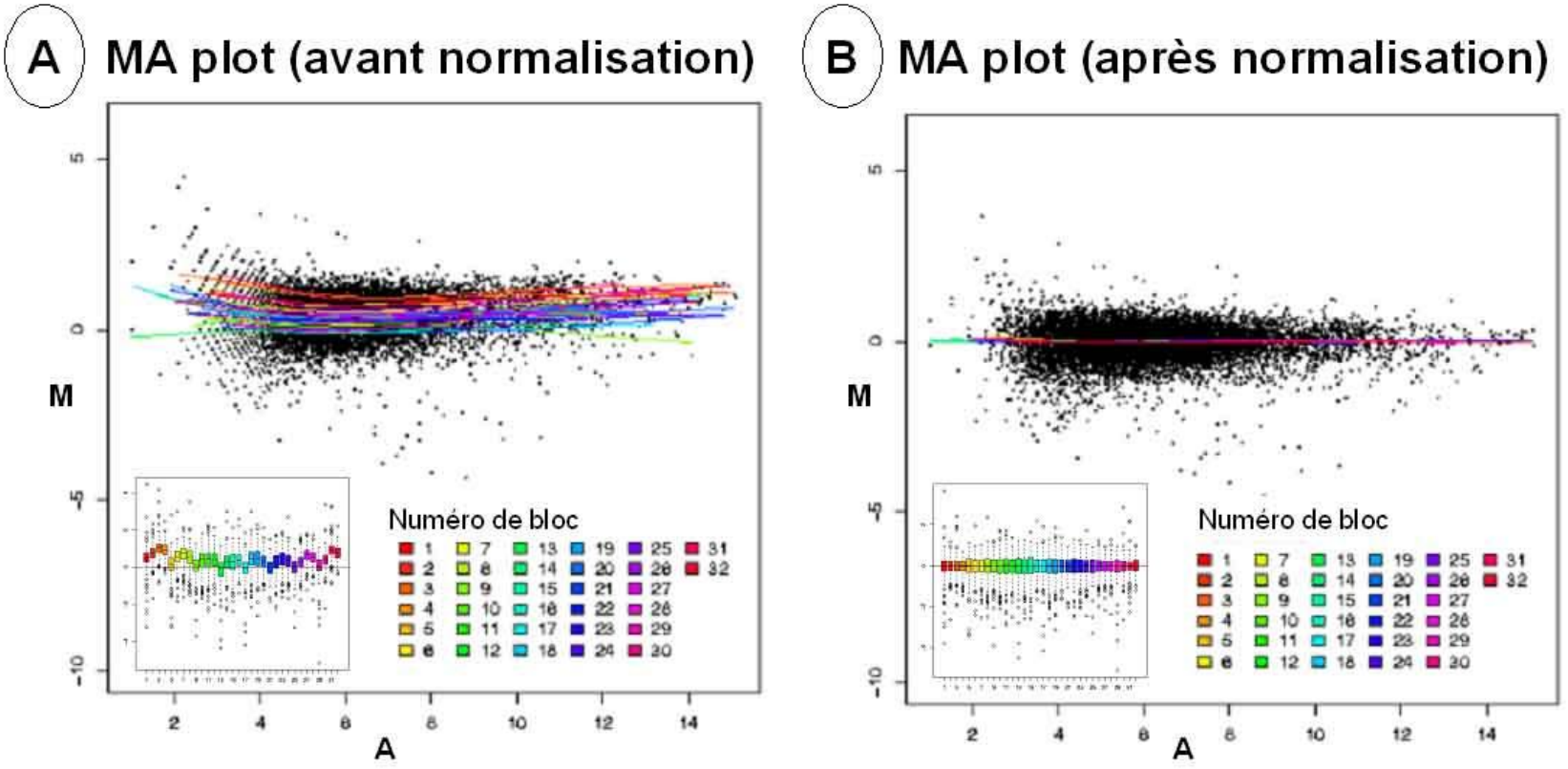
Correction de l'effet de localisation sur la puce



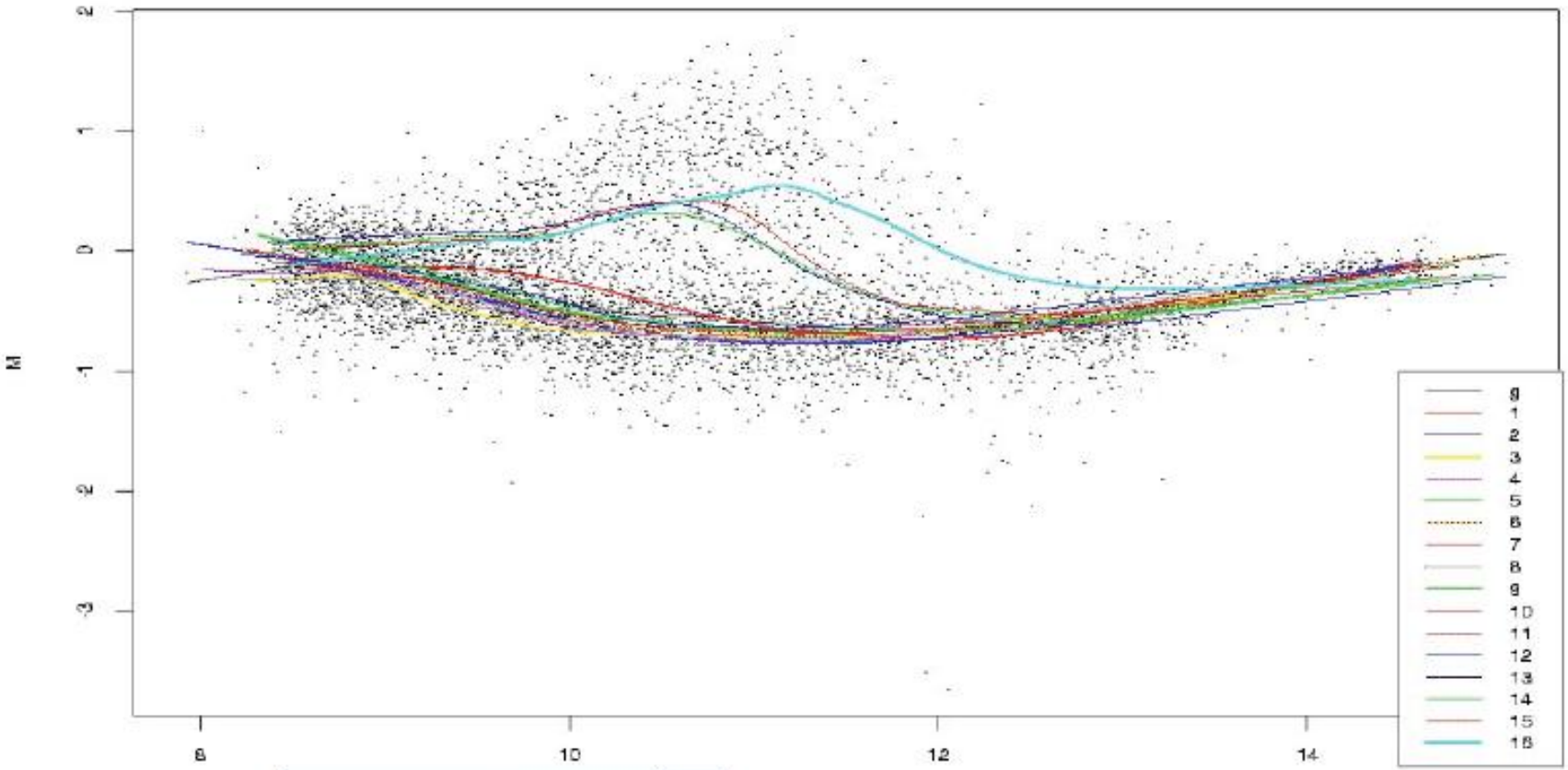


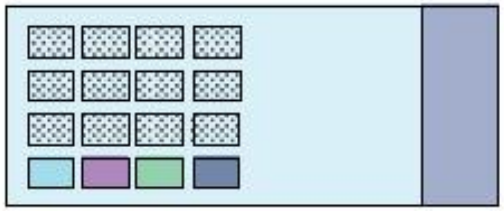
Le plus simple est de faire une
<u>normalisation par bloc</u>

Normalisation loess par bloc



Normalisation loess par bloc





Lame de verre

Lames de verre avec ADNc

4x4 blocs = 16 groupes de pointes

Les conditions pour effectuer une normalisation

• Les méthodes de normalisation reposent sur les hypothèses suivantes :

Hypothèse 1 : La grande majorité des gènes utilisés pour estimer les biais entre les deux fluorochromes est supposé avoir une expression constante entre les deux conditions comparées sur la puce.

Hypothèse 2 : Les biais corrigés ne sont pas confondus avec les effets biologiques.

- ✓ Une fois que l'analyse de l'image et la normalisation des données ont été réalisées, les fichiers de données sont utilisables pour répondre à des questions biologiques :

Question 1 :

Quels sont les gènes différentiellement exprimés dans une condition par rapport à l'autre ?

Question 2 :

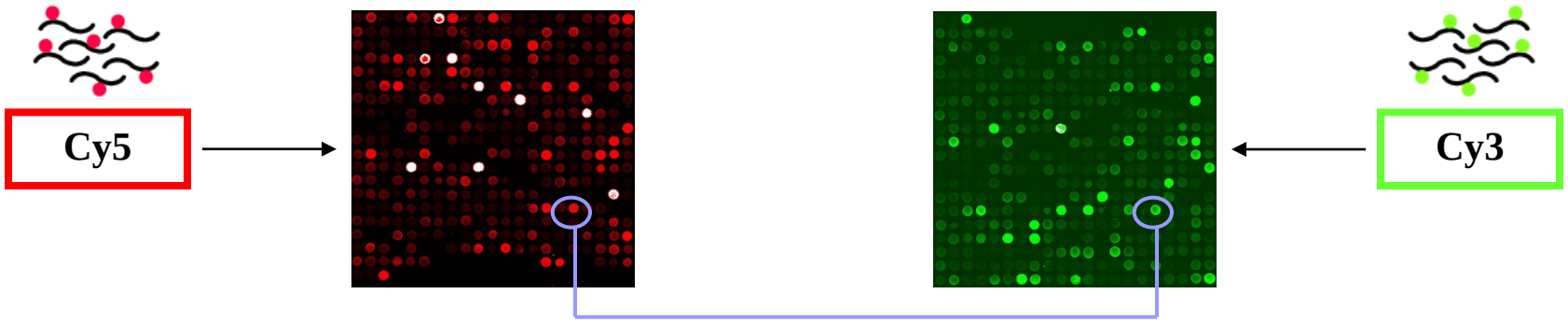
Comment classer les gènes relativement à leurs mesures d'expression dans plusieurs expériences ?

Question 1 :

Quels sont les gènes différentiellement exprimés dans une condition par rapport à l'autre ?

Recherche des gènes différentiellement exprimés

- ✓ **Objectif** : On souhaite identifier les gènes qui sont différentiellement exprimés entre les deux conditions hybridées sur la puce (ou entre deux puces différentes).



Expression différentielle ?

- ✓ Les gènes sont considérés un à un (et indépendamment) afin de déterminer s'ils sont différentiellement exprimés ou non. La méthode choisie sera alors appliquée à chaque gène présent sur la puce.

→ La puce est utilisée comme un outil pour étudier des milliers de gènes en parallèle

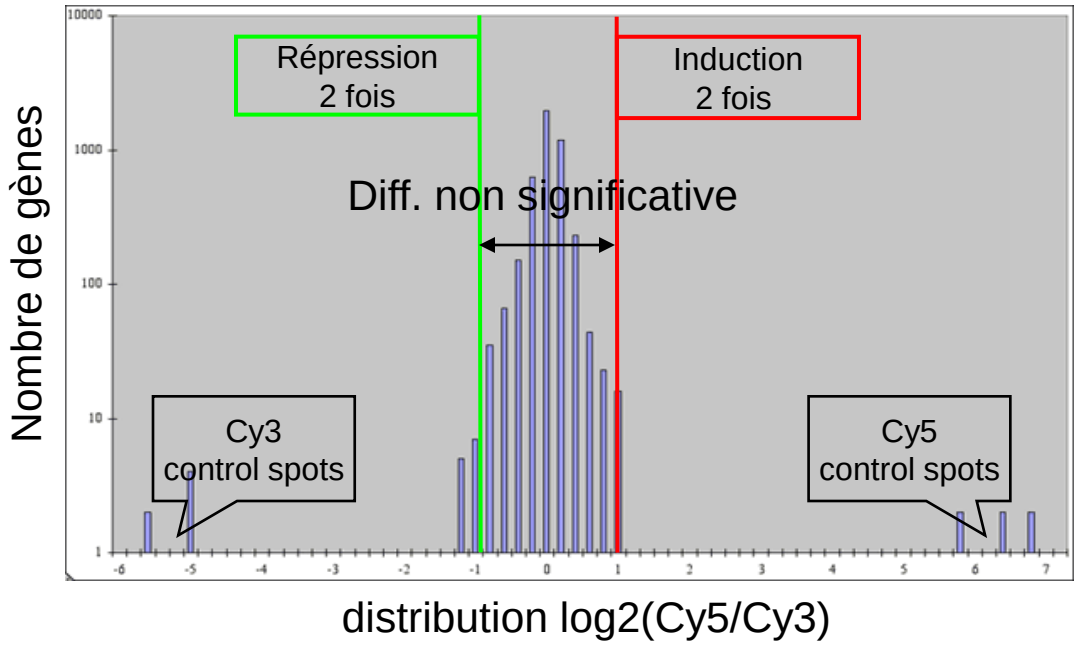
Une approche simple : choix d'un seuil arbitraire

- ✓ **Principe :** Un seuil **C** est choisi (par exemple différence d'expression = 2), et les gènes dont la différence d'expression est supérieure à ce seuil sont sélectionnés.

Si : Intensité-Cy5 > **C** * Intensité-Cy3 : **expression différentielle**

Si : Intensité-Cy5 < **C** * Intensité-Cy3 : ***pas* d'expression différentielle**

- ✓ Si un même lot d'ARNm est marqué en Cy3 et Cy5 puis co-hybridé sur la même puce, après normalisation pour corriger les biais expérimentaux, on observe:



Principales limites de cette approche

- ✓ Considérablement utilisée dans les premières études réalisées avec des puces à ADN, cette approche simple n'est pas la meilleure pour différentes raisons :

  - Comment choisir un seuil approprié ?

  - La variabilité globale des différences d'expression sur l'ensemble de la puce n'est pas prise en compte.

  Si cette variabilité est faible, un gène peut être seulement 1.5 fois différentiellement exprimé pour être sélectionné…

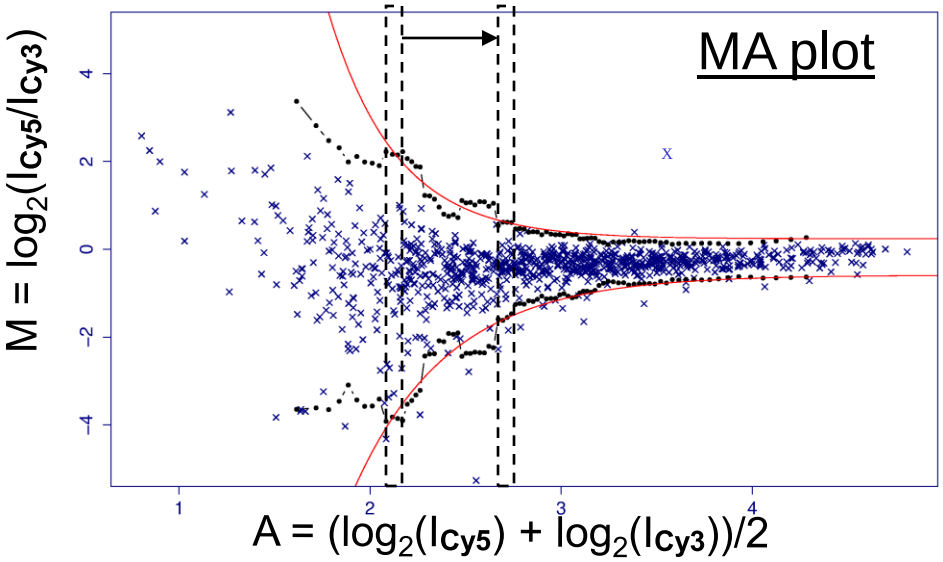
  - Le niveau des intensités (Cy5 et Cy3) n'est pas pris en compte.

  Une différence d'expression de 2 peut résulter d'un ratio d'intensité de 10/5 ou 10000/5000 …

→ Il faut utiliser des méthodes plus élaborées qui reposent sur des approches statistiques !

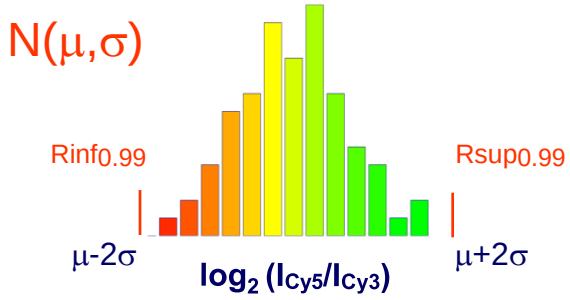
Analyse de la variabilité d'une expérience de puces à ADN : VARAN

- ✓ VARAN est un outil WEB qui réalise une analyse des intensités en tenant compte de la variabilité des valeurs de log2(Ratio) issues d'une expérience de puce à ADN.



Golfier et al., Bioinformatics (2004)

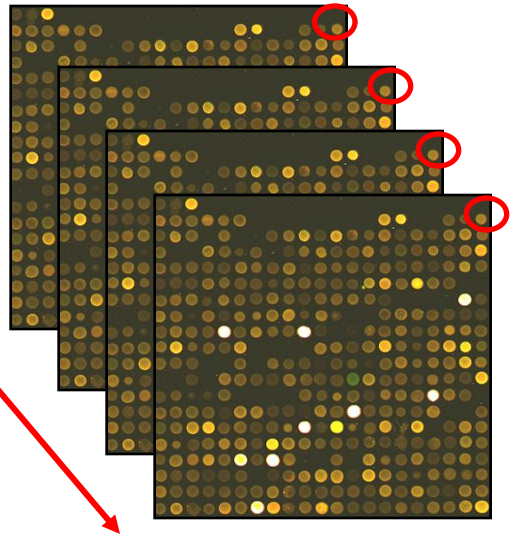
Un gène avec de fortes intensités est sélectionné sur une valeur de log2(Ratio) plus faible qu'un gène avec de faibles intensités.



Le principe de VARAN repose sur l'analyse des distributions des log2(Ratio) dans une fenêtre coulissante sur le MA plot.

http://www.bionet.espci.fr/varan/varan_info.htm

Pourquoi répéter les expériences ?



Répétition de la même expérience

→ Création d'un échantillon

✓ Pour chaque gène, une moyenne d'expression différentielle entre les deux conditions peut être calculée.

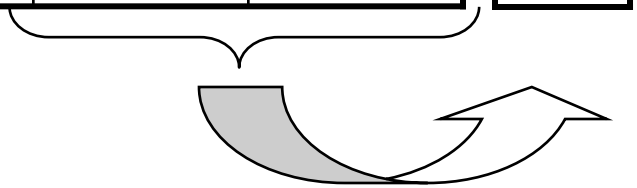
→ Les moyennes sont moins variables que les valeurs individuelles.

→ En utilisant les répétitions, il est possible de déterminer si un gène est ou n'est pas différentiellement exprimé, en utilisant les tests d'hypothèses.

Données appariées ou non appariées ?

✓ Données appariées

Exp	Cond 1	Cond 2	Ratio
Puce 1	$I_{Cy5}(1,1)$	$I_{Cy3}(1,2)$	Ratio1
Puce 2	$I_{Cy5}(2,1)$	$I_{Cy3}(2,2)$	Ratio2
Puce 3	$I_{Cy5}(3,1)$	$I_{Cy3}(3,2)$	Ratio3
Puce 4	$I_{Cy5}(4,1)$	$I_{Cy3}(4,2)$	Ratio4
Puce 5	$I_{Cy5}(5,1)$	$I_{Cy3}(5,2)$	Ratio5



- Deux mesures pour un seul groupe d'expériences.
- Les intensités sont combinées afin d'obtenir une valeur unique : le ratio.

✓ Données non-appariées

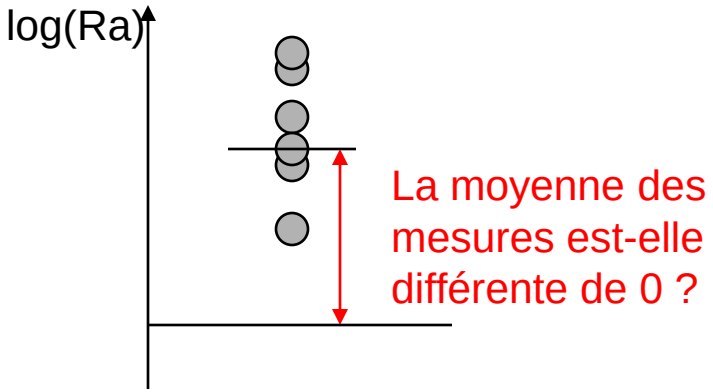
Répétitions cond 1	Ratio du gène X
1	Ratio(1,1)
2	Ratio(2,1)
3	Ratio(3,1)
4	Ratio(4,1)
5	Ratio(5,1)

Répétitions cond 2	Ratio du gène X
1	Ratio(1,2)
2	Ratio(2,2)
3	Ratio(3,2)

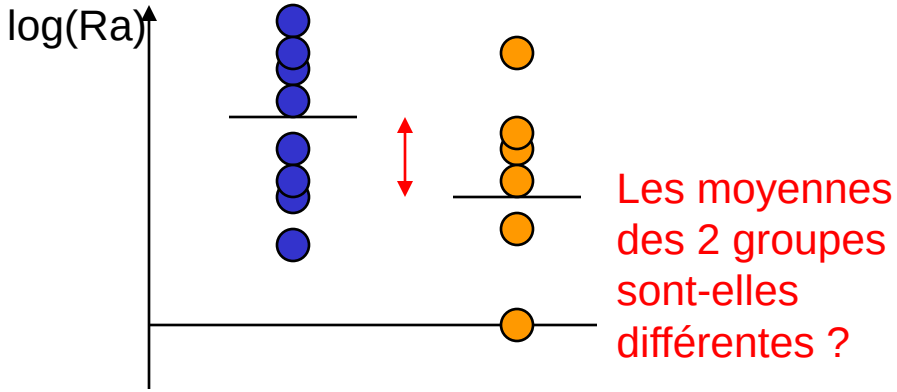
- Deux groupes d'expériences (conditions 1 et 2) avec seulement une mesure.
- Le nombre de répétitions pour chaque conditions peut être différent.

Utilisation d'un test paramétrique classique : t test

✓ Données appariées



✓ Données non-appariées



Une « p-value » est calculée en comparant la valeur t calculée, à une distribution de Student (avec un degré de liberté approprié).

t-test apparié

$$t = \frac{\bar{x}}{s/\sqrt{n}}$$

t-test non-apparié

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

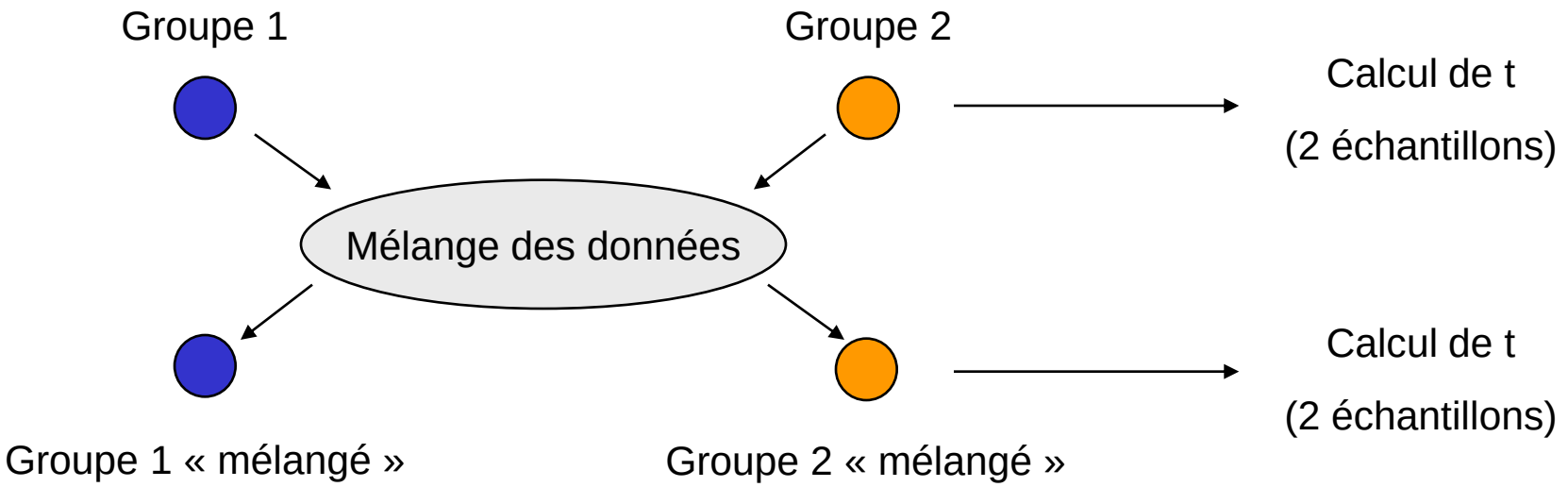
Le t-test nécessite que la distribution des données testées soit normale.
Contestable → TESTS NON-PARAMETRIQUES !

Analyse par ré-échantillonnage : ex des données non appariées (1/2)

- ✓ **Objectif :** On veut déterminer si les moyennes de deux échantillons sont différentes.

- Les analyses par « bootstrap » (ou ré-échantillonnage) ne supposent pas que les données sont distribuées selon un loi normale.

On fait l'hypothèse qu'il n'y a pas de différences entre les moyennes des deux groupes (hypothèse nulle Ho). Dans ce cas, chaque mesure du groupe 1 aurait pu être observée pour le groupe 2, et inversement.

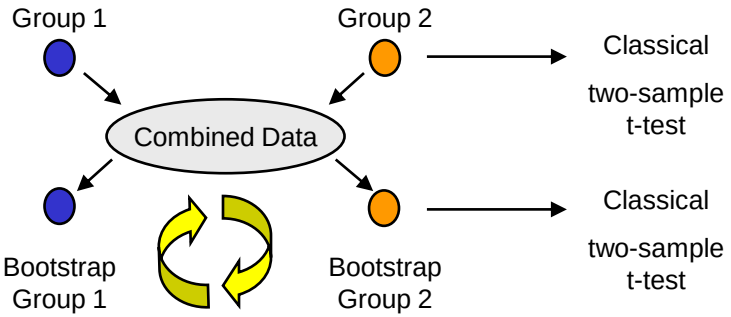


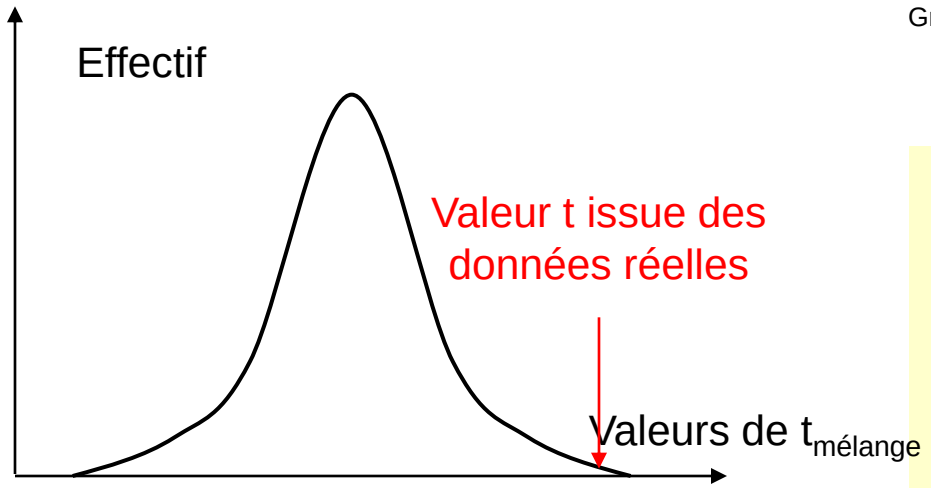
Le jeu de données mélangées ressemble aux données réelles puisque ce sont les mêmes valeurs, mais n'a pas de sens biologique.

Analyse par ré-échantillonnage : ex des données non appariées (2/2)

- ✓ **Principe :** Un grand nombre de jeux de données aléatoires sont construits en mélangeant les données initiales. On obtient alors une distribution des valeurs t obtenues pour chaque jeux de données mélangées.

- ✓ La valeur de t, calculée sur les données réelles est comparée à la distribution des valeurs de « t aléatoire ».





Qu'est ce qu'une p-value ?

Une p-value de 0.01 signifie qu'il y a 1 chance sur 100 d'observer au moins ce niveau d'expression différentielle uniquement par le hasard.

→ Une « p-value » empirique est calculée en regardant la proportion de valeurs t aléatoires supérieures à celle obtenue sur les données réelles.

Tests Multiples

- ✓ **Principe :** En considérant la définition d'une p-value, si un gène à 1 % de chance d'avoir une p-value inférieure à 0.01 par le hasard, en analysant 10 000 gènes sur une puce, on peut attendre une centaine de gènes « différentiellement exprimés » (CAD avec une p-value de 0.01) uniquement par le hasard !

Il faut faire attention lorsque l'on effectue le même test statistique sur de nombreux gènes en parallèle…

→ ESTIMATION DU TAUX DE FAUX POSITIFS !

Un outil disponible : SAM (Tusher et al., 2000)

SAM (Significance Analysis of Microarrays) effectue une analyse par ré-échantillonnage des données et estime le taux de faux positifs.

La méthode SAM

- ✓ **Principe :** Pour chaque gène, calcul d'un score qui quantifie la différence d'expression du gène par rapport à 0.

Expression moyenne sur les répétitions

$$d(i) = \frac{\bar{x}(i) - 0}{s(i) + so}$$

Variabilité observée sur les répétitions

Coefficient correctif

→ **Tri des gènes selon le d(i) calculé. Plus la valeur du score est grande, plus l'expression du gène peut être considérée comme différente de 0.**

Gène 1	Gène 2	Gène 3	….	Gène (n-1)	Gène n
d(1)	d(2)	d(3)	….	d(n-1)	d(n)

La méthode SAM

Tableau de données initial

	R1	R2	
Gène 1	+Val(1,1)	+Val(1,2)	**d(1)**
Gène 2	-Val(2,1)	-Val(2,2)	**d(2)**
…	…	…	**…**
Gène (n-1)	-Val(n-1,1)	-Val(n-1,2)	**d(n-1)**
Gène n	+Val(n,2)	+Val(n,2)	**d(n)**

→ Etape 1 : calcul des valeurs de « d » réelles

→ Etape 2 : Changement des signes et calcul des valeurs d^{perm}

	R1	R2	
Gène 1	+Val(1,1)	-Val(1,2)	$d^1(1)$
Gène 2	+Val(2,1)	-Val(2,2)	$d^1(2)$
…	…	…	…
Gène (n-1)	+Val(n-1,1)	-Val(n-1,2)	$d^1(n-1)$
Gène n	+Val(n,2)	-Val(n,2)	$d^1(n)$

Le nombre de permutations possible est 2 x 2 = 4

→ Etape 3 : Tri décroissant des valeurs d^{perm} et calcul d'une moyenne pour chaque rang

$d^1(n-1)$	$d^2(2)$	$d^3(n-1)$	$d^4(2)$	$\longrightarrow d^{E1}$
$d^1(1)$	$d^2(1)$	$d^3(n)$	$d^4(1)$	$\longrightarrow d^{E2}$
…	…	…	…	…
$d^1(n)$	$d^2(n-1)$	$d^3(2)$	$d^4(n-1)$	$\longrightarrow d^{En-1}$
$d^1(2)$	$d^2(n)$	$d^3(1)$	$d^4(n)$	$\longrightarrow d^{En}$

→ Etape 4 : Comparaison des valeurs $d^{réelles}$ et d^{perm}

d(2)	⟷	d^{E1}
d(n)	⟷	d^{E2}
…		…
d(n-1)	⟷	d^{En-1}
d(1)	⟷	d^{En}

La méthode SAM

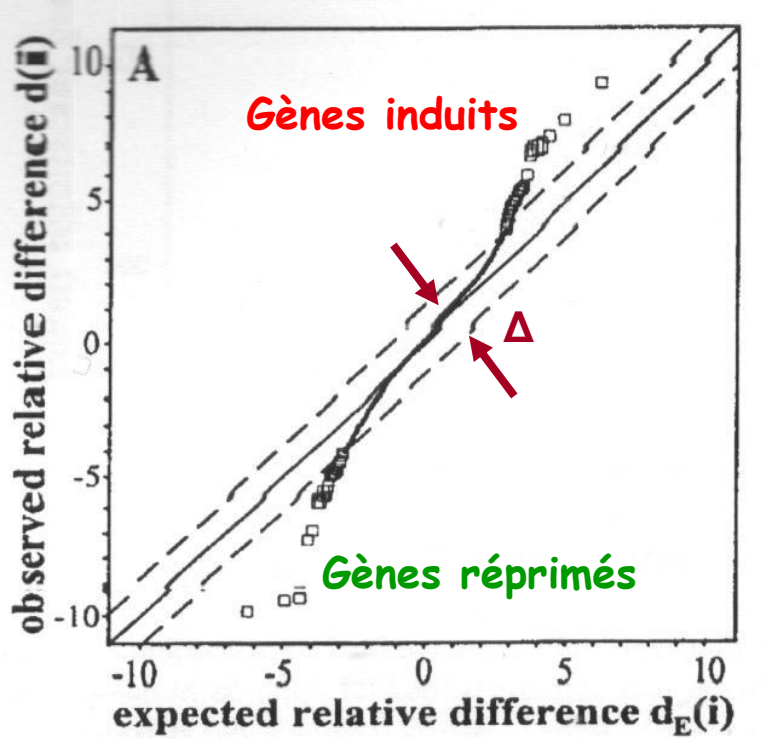
- ✓ Des permutations de signe sont réalisées dans les données et des valeurs d(i)(hasard) sont calculées (en utilisant la formule précédente) puis ordonnées :

d(3) (hasard)	d(n-1) (hasard)	d(1) (hasard)	….	d(2) (hasard)	d(n) (hasard)
d(1)	d(2)	d(3)	….	d(n-1)	d(n)

- ✓ Les valeurs d(i) obtenues sur les données réelles sont comparées à celles d(i)(hasard) obtenues suite aux permutations :

- ✓ Choix de la valeur du paramètre Δ, limite à partir de laquelle les gènes induits et réprimés sont gardés.

- ✓ Estimation du taux de faux positifs (FDR = False Discovery Rate). Plus Δ est grand plus le FDR est faible.



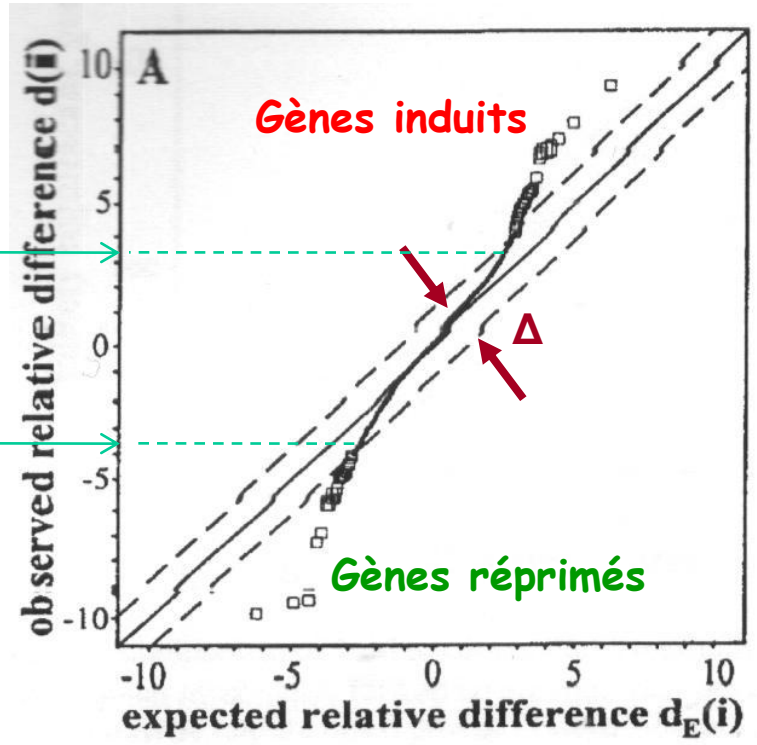
Calcul du taux de faux positifs (FDR)

Une fois le Δ fixé, on a une valeur d_{lim} à partir de laquelle un gène est considéré comme différentiellement exprimé

En utilisant les valeurs d^{perm}, on peut estimer le taux de faux positifs :

P1	P2	P3	P4	
$d^1(n-1)$	$d^2(2)$	$d^3(n-1)$	$d^4(2)$	→ d^{E1}
$d^1(1)$	$d^2(1)$	$d^3(n)$	$d^4(1)$	→ d^{E2}
…	…	…	…	…
$d^1(n)$	$d^2(n-1)$	$d^3(2)$	$d^4(n-1)$	→ d^{En-1}
$d^1(2)$	$d^2(n)$	$d^3(1)$	$d^4(n)$	→ d^{En}

→ Décompte du nombre de $d^{perm} > d_{lim}$ dans chacune des permutations



FDR = Nombre moyen de $d^{perm} > d_{lim}$

Ainsi, plus Delta est grand, plus le FDR est petit…

Exemple de tableau de résultat SAM

Parameter	Number falsely significant	Number called significant	FDR
SAM			
$\Delta = 0.4$	134.9	288	47%
$\Delta = 0.5$	78.1	192	41%
$\Delta = 0.6$	56.1	162	35%
$\Delta = 0.9$	19.1	80	24%
$\Delta = 1.2$	8.4	46	18%
$\Delta = 1.2$; R = 1.5	4.5	34	12%

(Tusher et al., 2000)

→ Le choix de Δ dépend du nombre de gène sélectionnés comme différentiellement exprimés et du taux de faux positifs associé.

→ Il faut **trouver un juste milieu** entre un taux de faux positifs acceptable et le risque d'avoir des faux négatifs (gènes différentiellement exprimés qui ne sont pas gardés).

Question 2 :

Comment classer les gènes relativement à leurs mesures d'expression dans plusieurs expériences ?

Constitution d'une série d'expérience

Plusieurs expériences de puces à ADN

Constitution d'une série d'expérience

	Exp 1	**Exp 2**	**Exp 3**	…..	**Exp m**
Gène 1	Val(1,1)	Val(1,2)	Val(1,3)	…..	Val(1,m)
Gène 2	Val(2,1)	Val(2,2)	Val(2,3)	…..	Val(2,m)
…..	…..	…..	…..	…..	…..
Gene n	Val(n,1)	Val(n,2)	Val(n,3)	…..	Val(n,m)

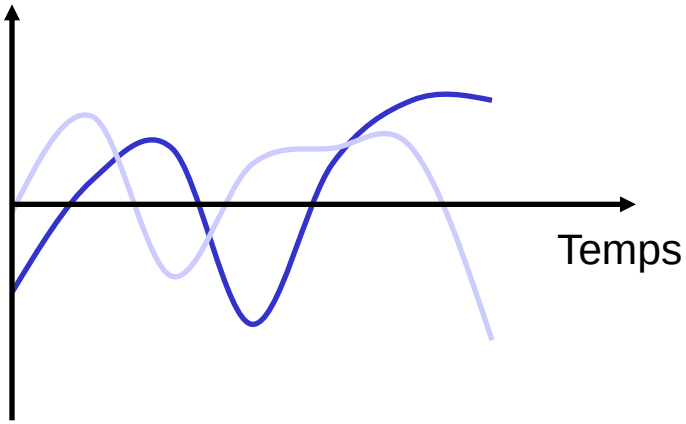
Profil d'expression

→ L'évolution de la quantité des transcrits est suivie en fonction de diverses conditions expérimentales (ex: temps)

Pourquoi analyser les profils d'expression des gènes ?

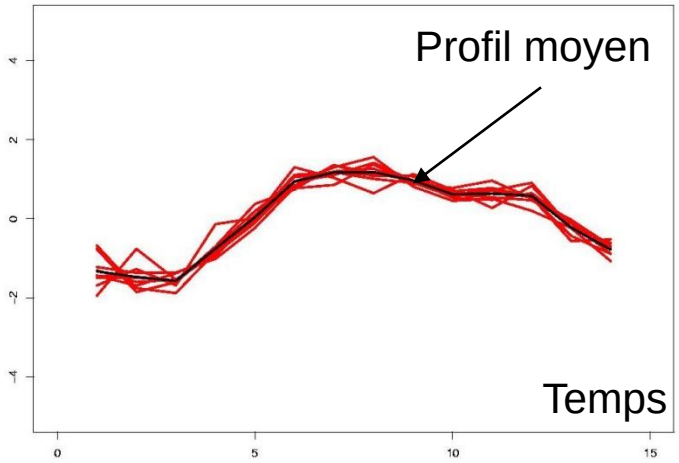
✓ Superposition des profils d'expression

log2(Ratio)



✓ Exemple : Profils d'expression des gènes codants les protéines Histones

log2(Ratio)



(*Cycle cellulaire, Spellman et al., 1998*)

→ Les gènes dont les <u>profils d'expression sont similaires</u> sont de bons candidats pour être régulés par les mêmes facteurs ou intervenir dans le même processus biologique.

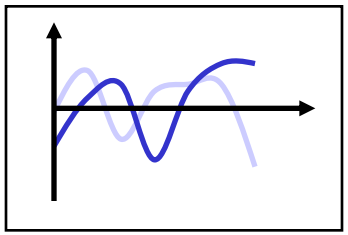
Identification des gènes dont les profils sont similaires

A chaque gène est associé un profil d'expression

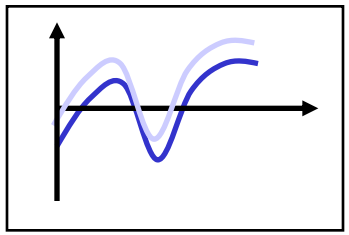
↓

Il s'agit alors de trier ces profils en fonction de leur ressemblance

Profils différents



↓



Profils similaires

De nombreuses approches peuvent être appliquées aux données puces à ADN

→ Deux étapes principales :

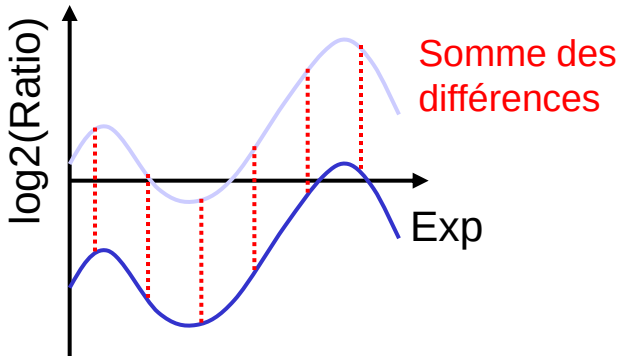
(1) Quantification du degré de ressemblance entre les profils pris deux à deux (calcul d'une distance)

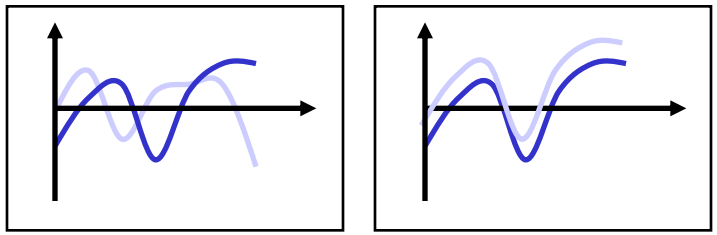
(2) Tri des profils en fonction de la distance qui les sépare

Calcul d'une distance entre des profils d'expression

✓ Pour identifier les gènes dont les profils d'expression sont similaires, un « critère de ressemblance » doit être défini. Classiquement, une distance est calculée.

- Distance euclidienne

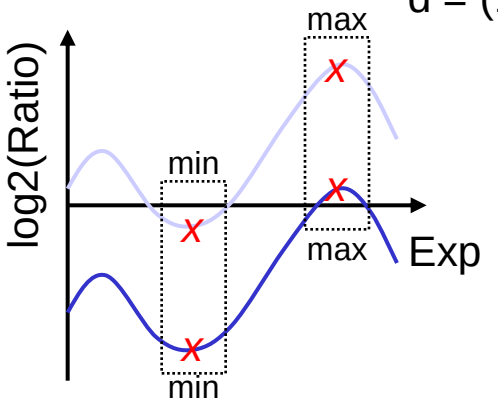


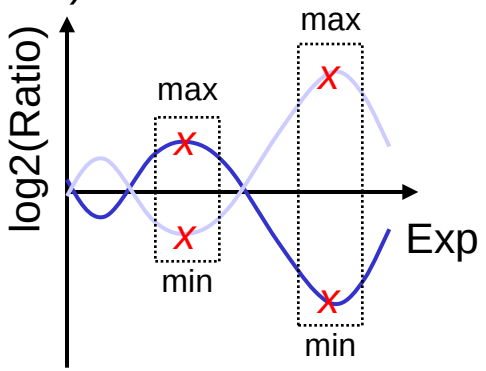


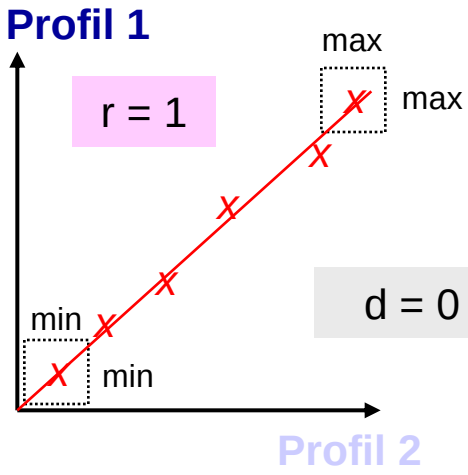
→ Deux profils peuvent être plus ou moins ressemblants selon la mesure de distance utilisée !

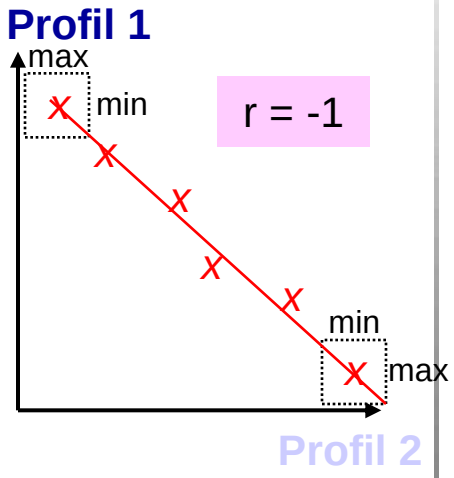
- Distance de corrélation

$d = (1 - r^2)$





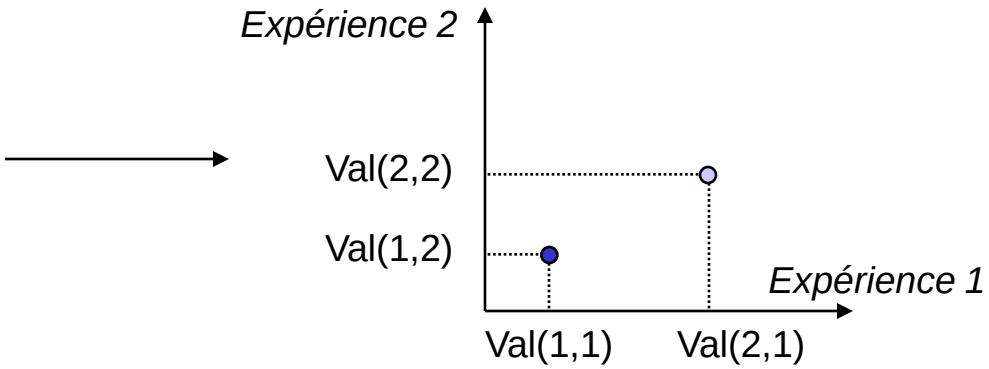




Représentation géométrique d'une série d'expérience

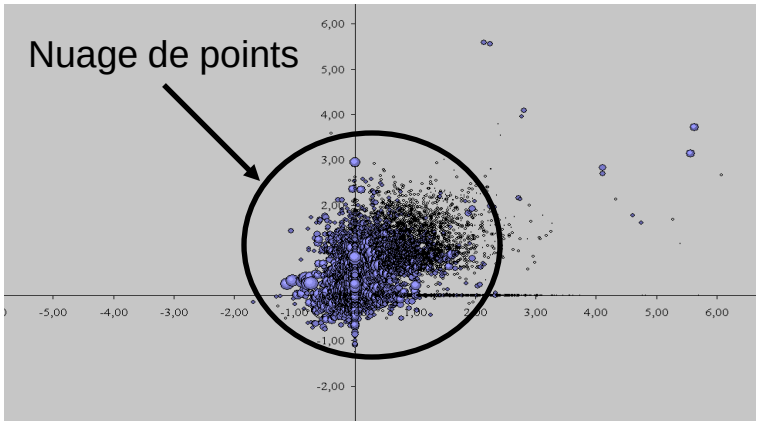
✓ Exemple :

	Exp 1	Exp 2
Gène 1	Val(1,1)	Val(1,2)
Gène 2	Val(2,1)	Val(2,2)

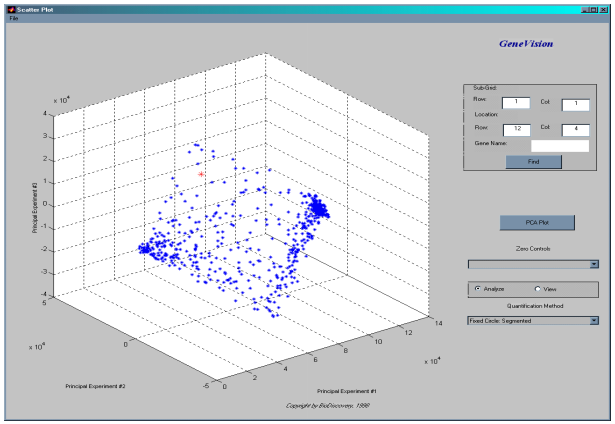


✓ Généralisation :

n gènes, 2 expériences



n gènes, 3 expériences



?

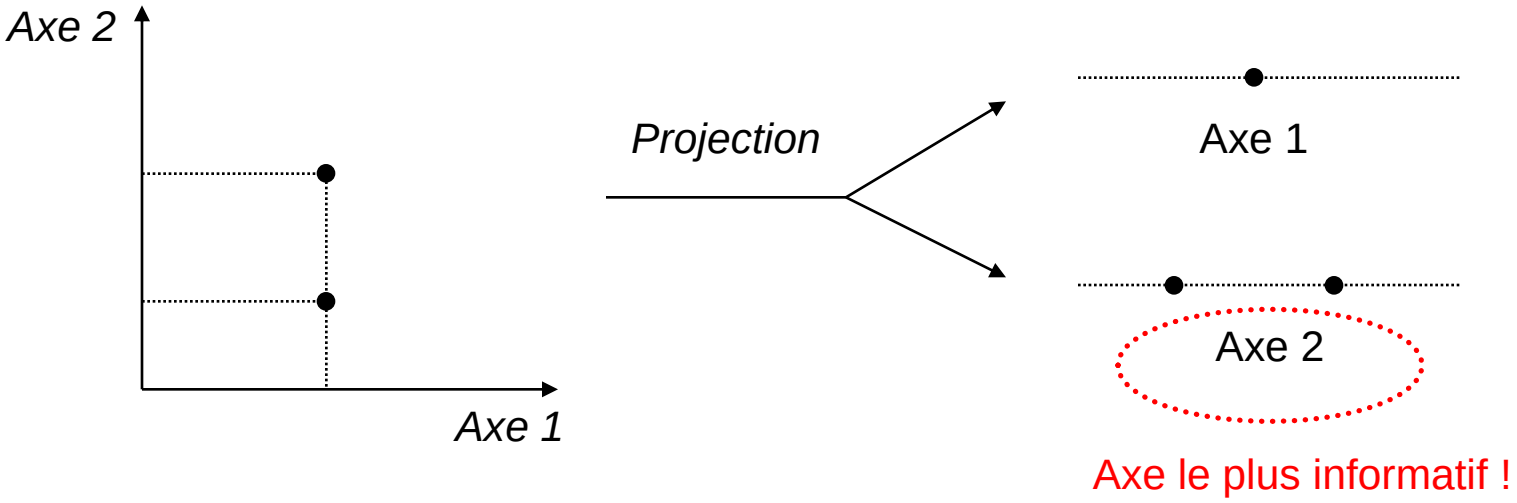
Visualisation du nuage dans le plan (2D) le plus informatif :
→ Méthode de réduction de dimensions

Une méthode factorielle : l'Analyse en Composante Principale (ACP)

- ✓ **Problème :** Difficulté de visualiser les distances entre les gènes lorsque la représentation géométrique d'une série requiert un espace à plus de 2 ou 3 dimensions.
- ✓ **Objectif :** Recherche d'un sous-espace qui ajuste au mieux le nuage de point de façon à ce que les proximités qui y sont mesurées reflètent les proximités réelles.

▪ Exemple :

Espace 2-dimensions → Espace 1-dimension



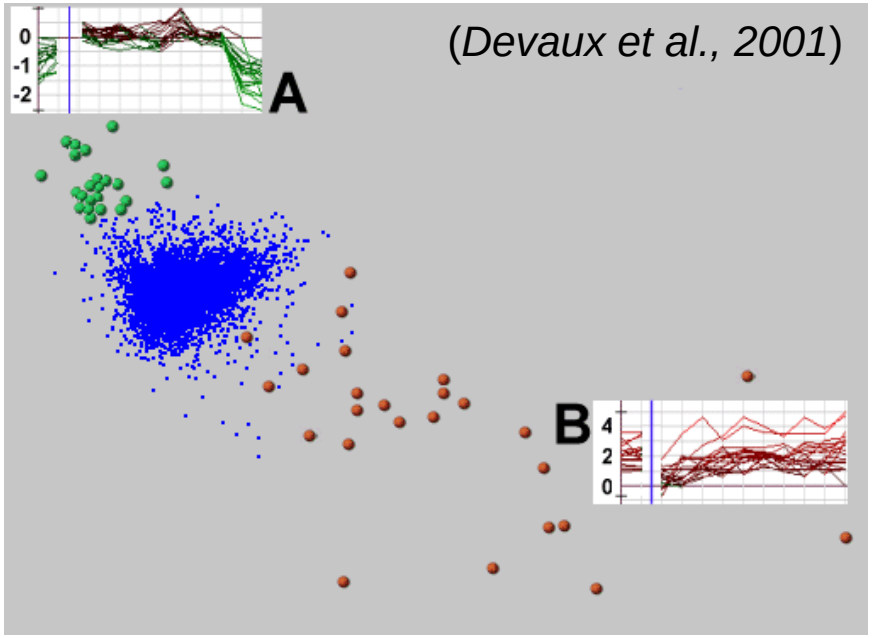
Analyse en Composante Principale : Exemple

Espace N-dimensions ⟶ Espace 2-dimensions

ACP

- Exemple :

Résultat d'une ACP sur la cinétique d'expression d'une protéine chimère :



Cette analyse permet de distinguer clairement le groupe des gènes réprimés (à gauche, vert) et le groupe des gènes activés (droite, rouge) de l'ensemble des gènes invariants (centre, bleu).

Le pour et le contre des méthodes

✓ Avantages

- Aucun paramètre à choisir

----- ACP -----

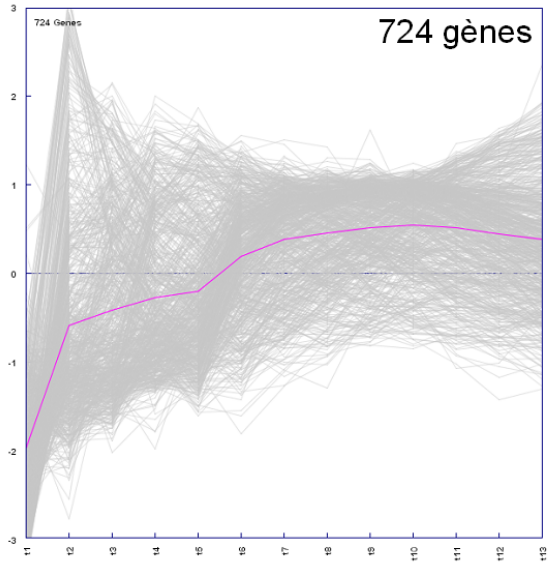
✓ Inconvénients

- Parfois difficile à interpréter

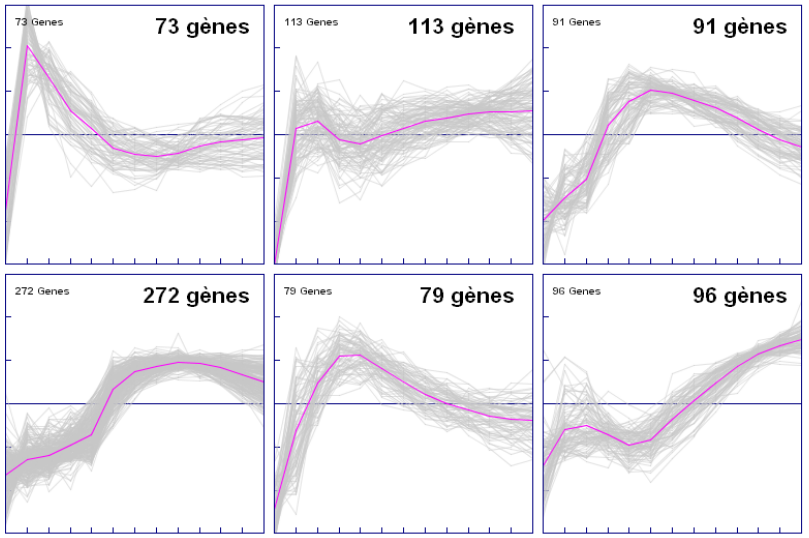
Objectif des méthodes de classification

✓ Le méthodes de classification ont pour objectif de regrouper les gènes sur la base d'une ressemblance entre leurs profils d'expression.

Exemple :



Avant classification

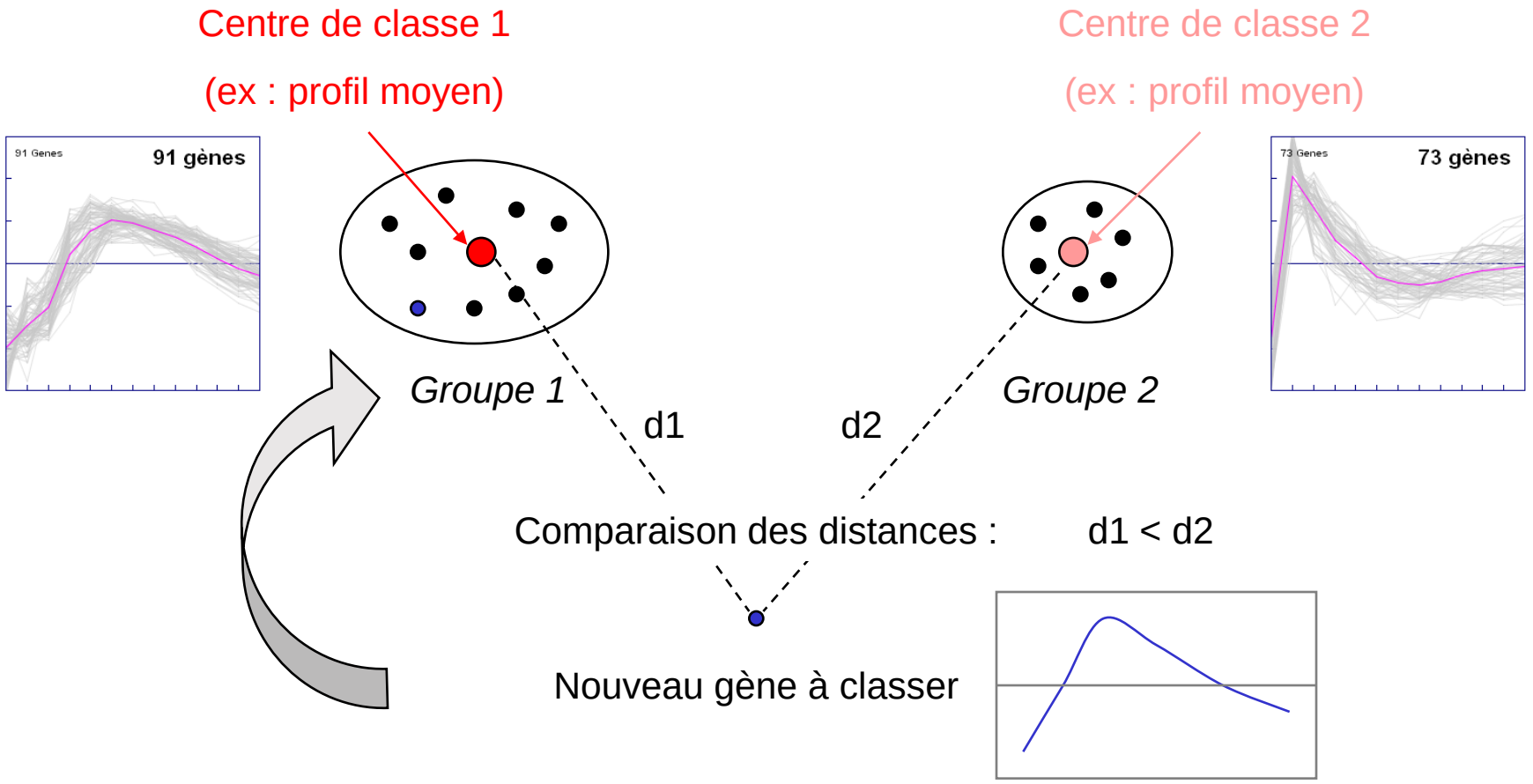


Après classification

→ La classification fait ressortir des groupes de gènes de taille plus restreinte pour lesquels des caractéristiques fortes sur les profils peuvent être observées

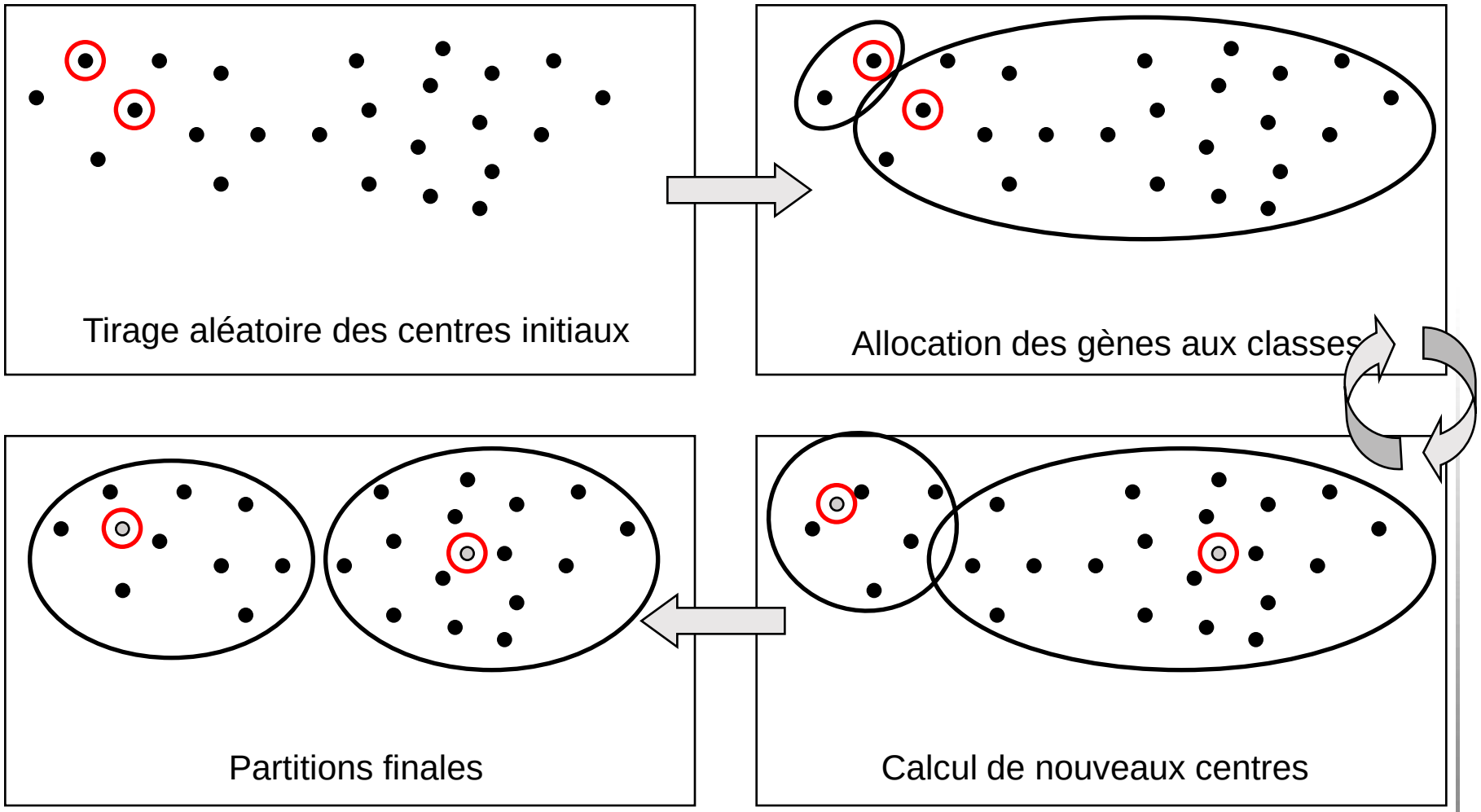
Une méthode de classification : les k-means

✓ Principe général de l'algorithme :



Une méthode de classification : les k-means

✓ Les différentes étapes de l'algorithme :



Le pour et le contre des méthodes

✓ Avantages

✓ Inconvénients

----- ACP -----

- Aucun paramètre à choisir
- Sometimes difficult to read

----- k-means -----

- Algorithme très rapide
- Bien adapté au jeu de données important
- Un travail préalable est nécessaire pour choisir le nombre de groupes approprié

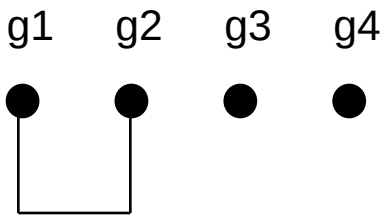
Une méthode de classification : le regroupement hiérarchique

✓ C'est une méthode qui fournit une <u>hiérarchie de partition</u> des gènes

- Matrice de distance initiale

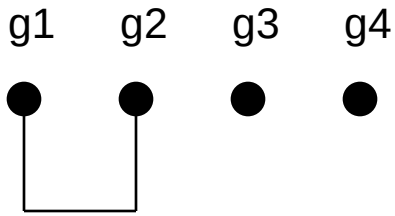
d(min)

	g1	g2	g3	g4
g1		d1	d2	d3
g2	d1		d4	d5
g3	d2	d4		d6
g4	d3	d5	d6	



- Matrice de distance recalculée

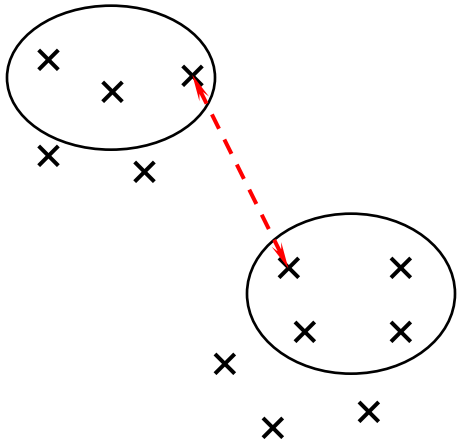
	1,2	g3	g4
1,2		d'	d''
g3	d'		d6
g4	d''	d6	



Le regroupement hiérarchique

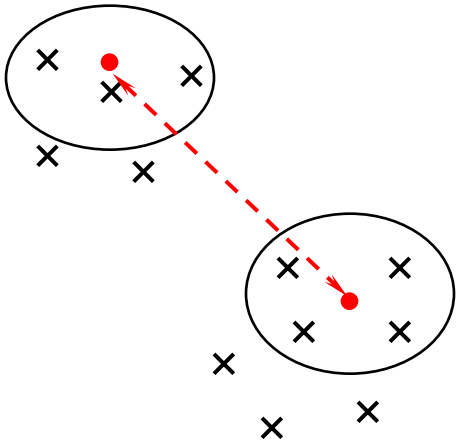
✓ Calcul de la distance entre plusieurs gènes :

- « Simple linkage »



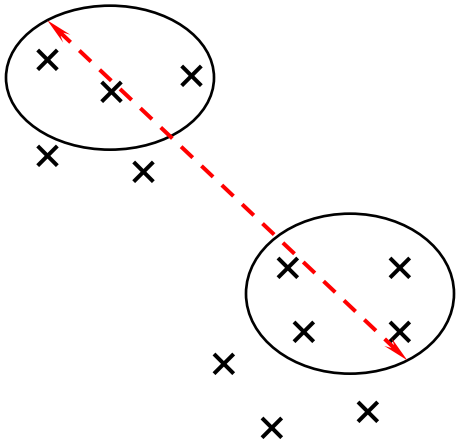
La distance entre les deux groupes est la distance minimale

- « Average linkage »



La distance entre les deux groupes est la distance entre les profils moyens

- « Complete linkage »



La distance entre les deux groupes est la distance maximale

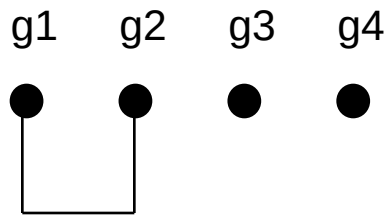
Une méthode de classification : le regroupement hiérarchique

✓ C'est une méthode qui fournit une hiérarchie de partition des gènes

- Matrice de distance initiale

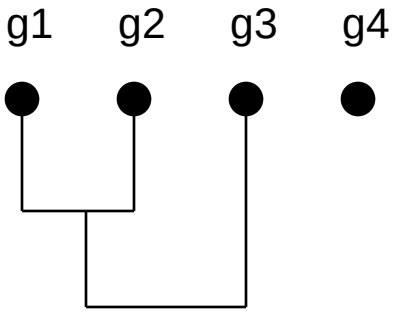
d(min)

	g1	g2	g3	g4
g1		d1	d2	d3
g2	d1		d4	d5
g3	d2	d4		d6
g4	d3	d5	d6	



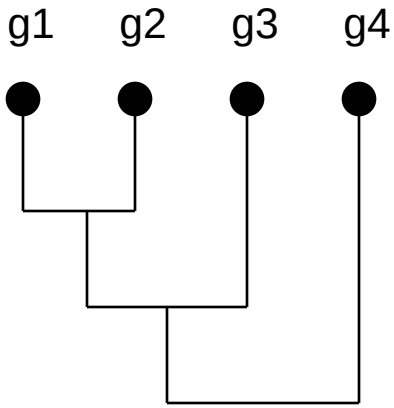
- Matrice de distance recalculée

	1,2	g3	g4
1,2		d'	d''
g3	d'		d6
g4	d''	d6	



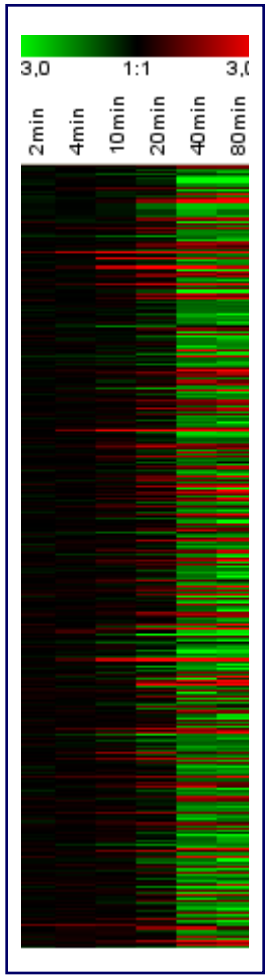
- Matrice de distance recalculée

	1,2,3	g4
1,2,3		d'''
g4	d'''	

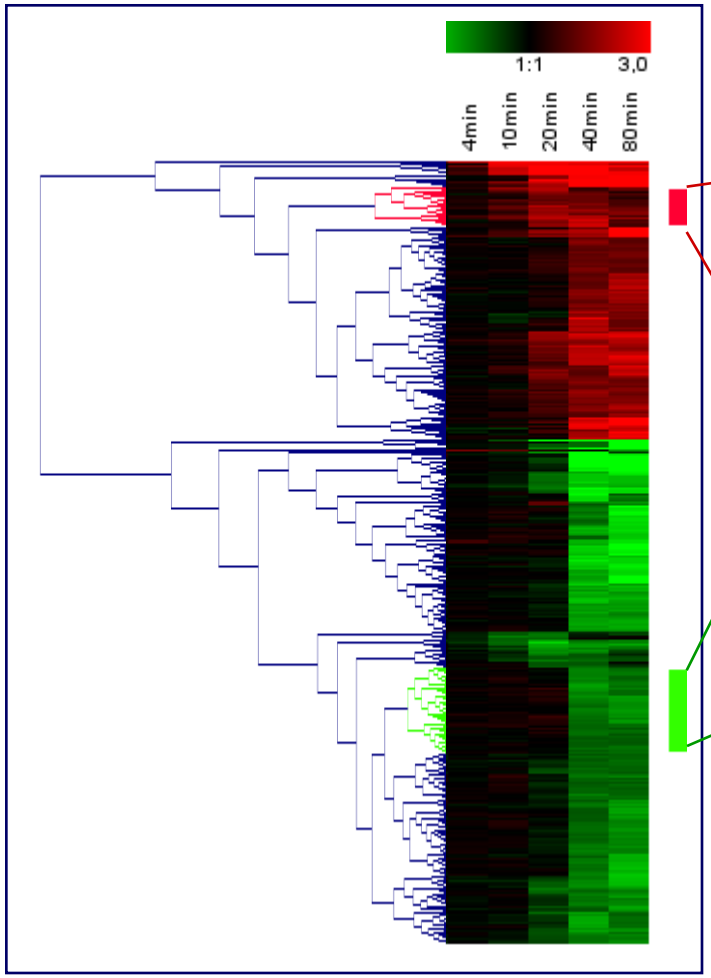


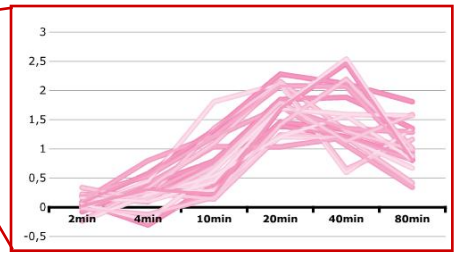
Résultat d'une classification hiérarchique en « Eisengramme »

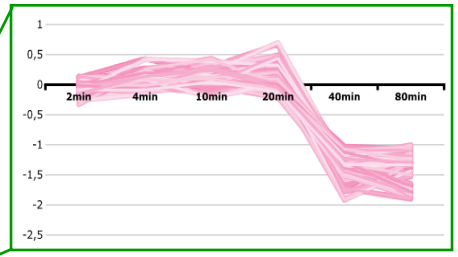
Avant classification



Après classification







Le pour et le contre des méthodes

✓ Avantages	✓ Inconvénients
----- ACP -----	
▪ Aucun paramètre à choisir	▪ Sometimes difficult to read
----- k-means -----	
▪ Algorithme très rapide ▪ Bien adapté au jeu de données important	▪ Un travail préalable est nécessaire pour choisir le nombre de groupes approprié
----- Classification hiérarchique -----	
▪ En regardant l'arbre, on peut retrouver l'ordre dans lequel les gènes ont été regroupés	▪ Difficile de traiter des jeux de données importants (coûteux en ressources-mémoires)

Une série d'expériences, deux classifications possibles

✓ Série d'expérience

	Exp 1	Exp 2	Exp 3
Gene 1	log2(1,1)	log2(1,2)	log2(1,3)
Gene 2	log2(2,1)	log2(2,2)	log2(2,3)
Gene 3	log2(3,1)	log2(3,2)	log2(3,3)
Gene 4	log2(4,1)	log2(4,2)	log2(4,3)

Classement des gènes

Exemple :
Identification des gènes co-régulés

Classement des expériences

Exemple :
Classification des tumeurs

→ D'un point de vu biologique, ce sont des questions très différentes. Mais les méthodes d'analyse sont semblables

Avant la classification : le prétraitement des données (1/3)

✓ **La gestion des valeurs manquantes :**

Il n'est pas rare que les ratios obtenus pour certains gènes ne soient pas utilisables.

	Exp 1	**Exp 2**	**Exp 3**	…..	**Exp m**
Gene 1	Val(1,1)	Val(1,2)	Val(1,3)	…..	Val(1,m)
Gene 2	X	Val(2,2)	Val(2,3)	…..	Val(2,m)
Gene 3	Val(3,1)	Val(3,2)	X	…..	Val(3,m)
…..	…..	…..	…..	…..	…..
Gene n	Val(n,1)	Val(n,2)	Val(n,3)	…..	Val(n,m)

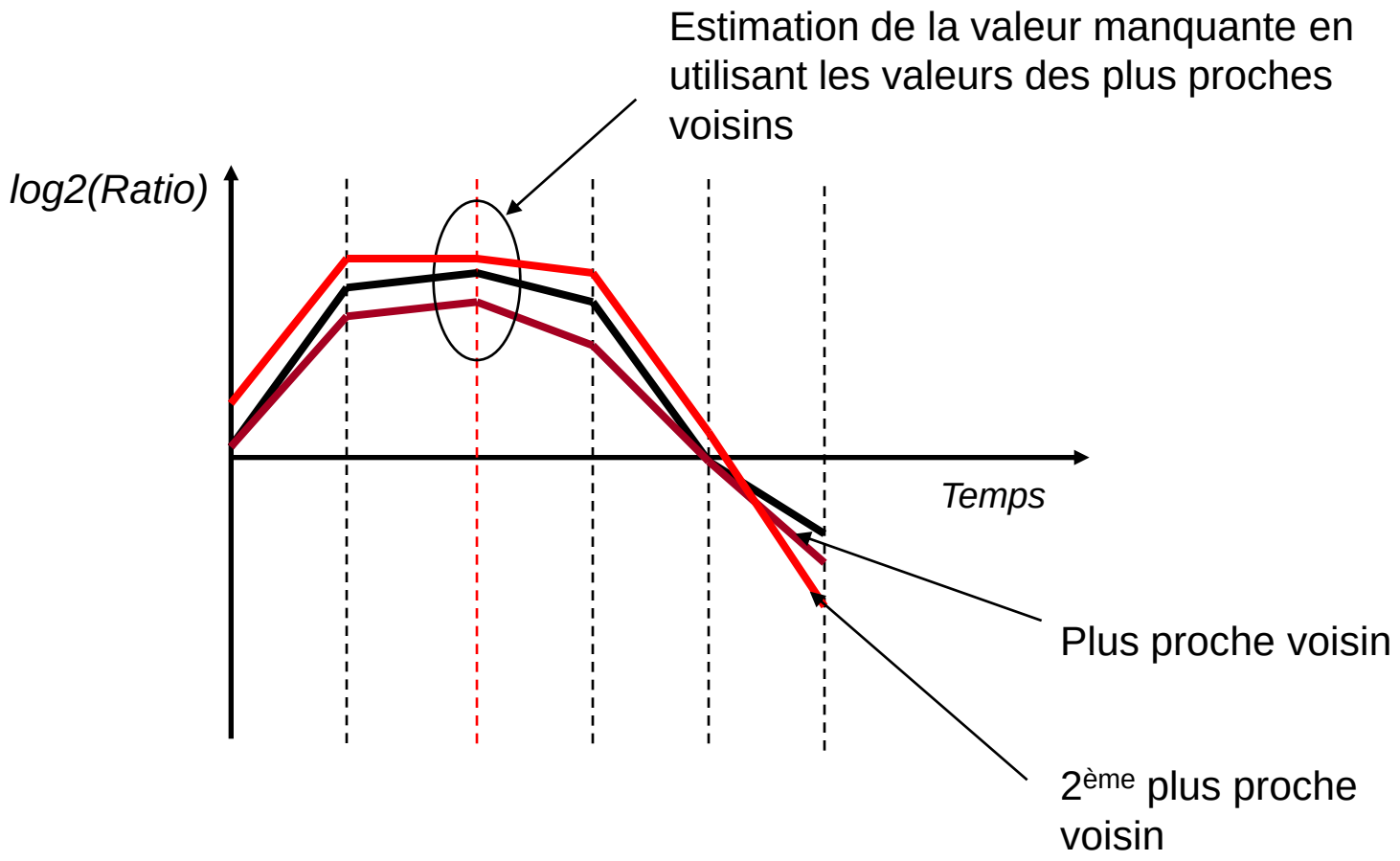
Valeurs manquantes

→ Les profils d'expression avec trop de valeurs manquantes sont éliminés.

→ Les valeurs manquantes sont remplacées en utilisant différentes méthodes :

- Valeur moyenne
- Méthode des plus proches voisins

La gestion des valeurs manquantes : méthode des plus proches voisins



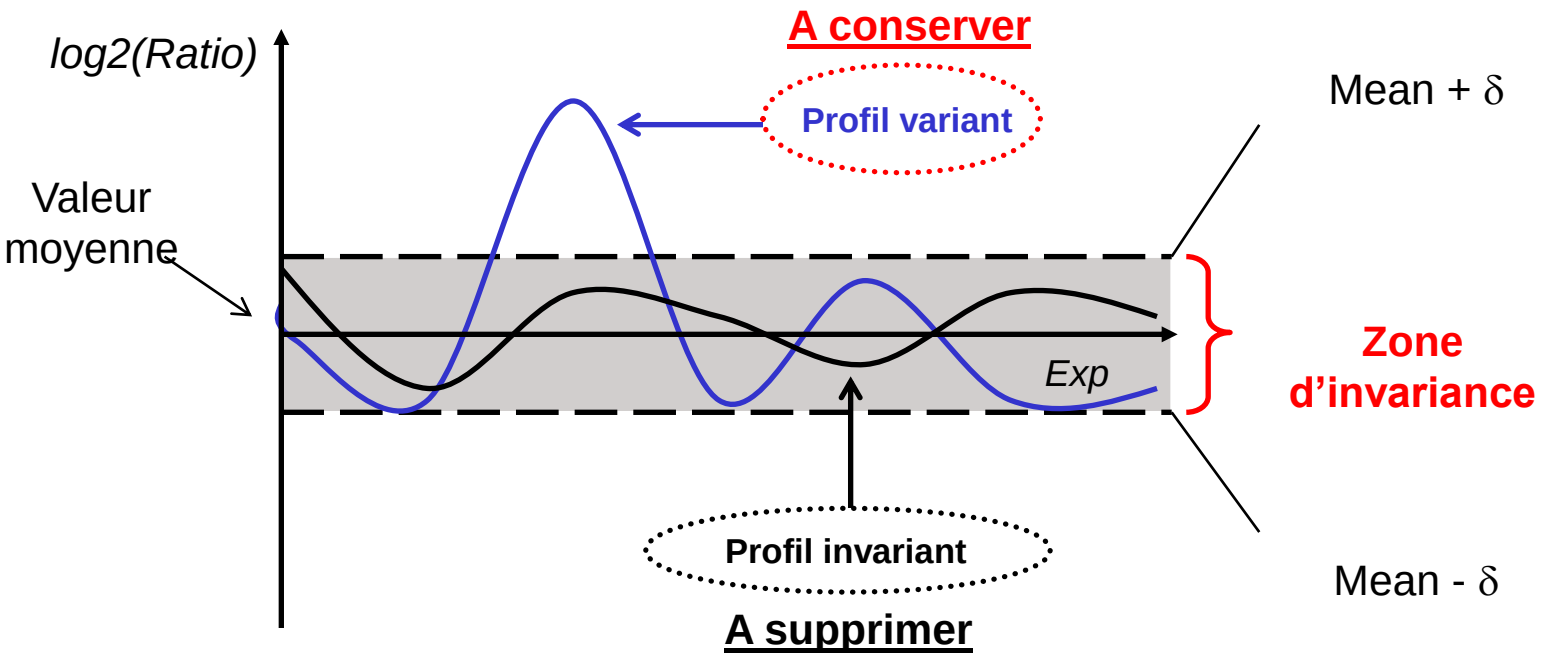
✓ Pour rechercher les plus proches voisins, un calcul de distance est réalisé en utilisant uniquement les valeurs existantes du profil initial.

✓ Dans la pratique, plusieurs dizaines de voisins seront utilisés pour estimer la valeur manquante.

Avant la classification : le prétraitement des données (1/3)

✓ **Filtrage des profils :**

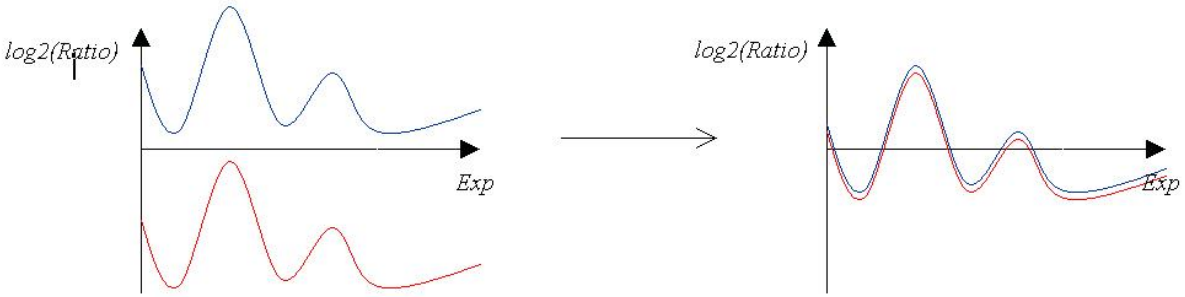
Il a pour but d'éliminer les profils des gènes dont l'expression ne varie pas dans une série d'expérience.



→ Une approche simple consiste à élimer les gènes dont les profils d'expression restent dans une "zone d'invariance".

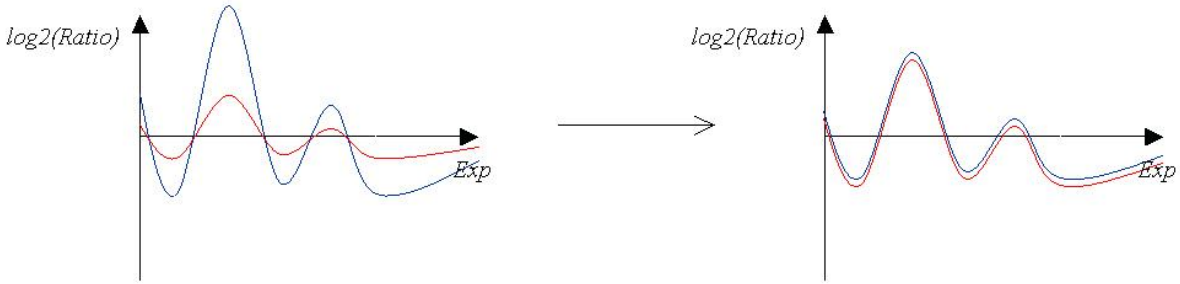
Avant la classification : le prétraitement des données (3/3)

✓ **Centrer** un profil d'expression consiste à soustraire la moyenne du profil à chacune des valeurs d'expression. La moyenne d'un profil centré est ainsi de 0.



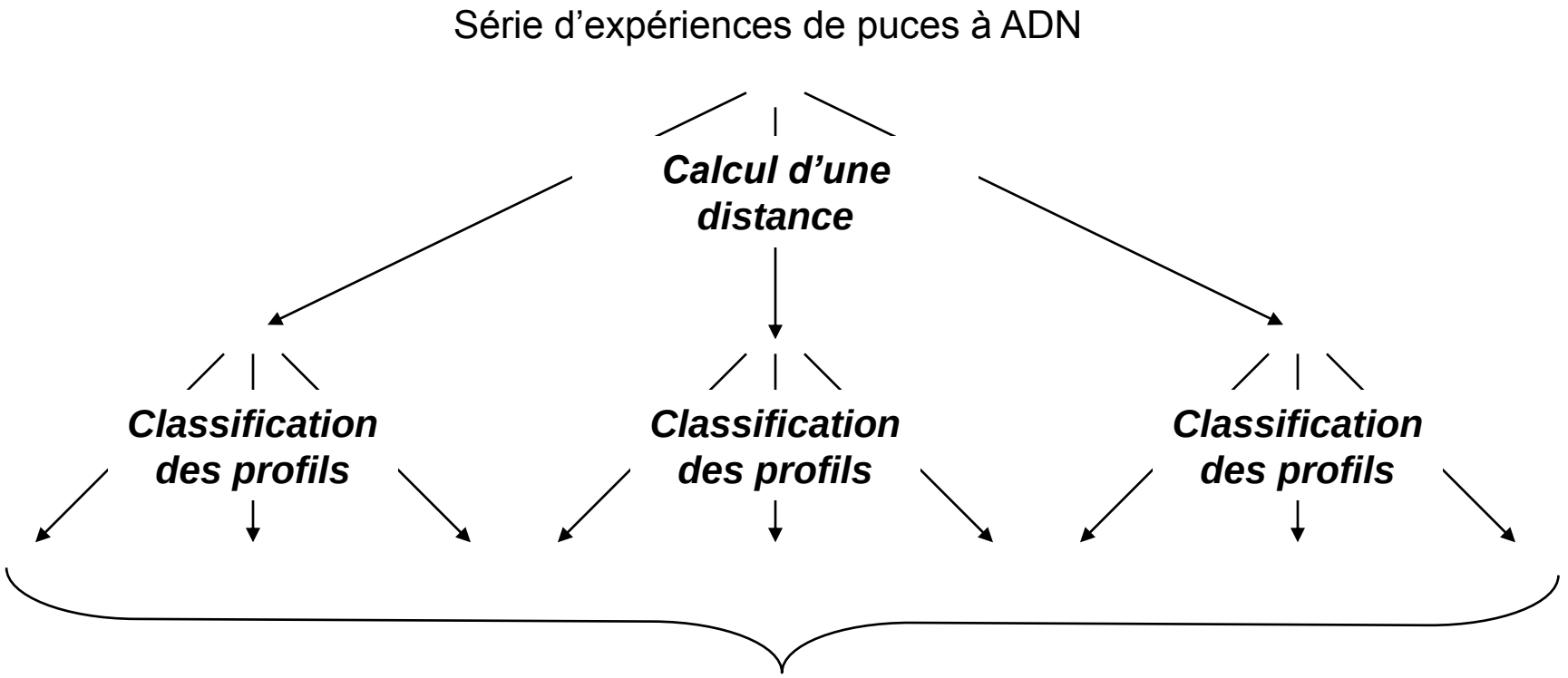
→ Le centrage élimine l'influence de la valeur de référence.

✓ **Réduire** un profil d'expression consiste à diviser chaque valeurs d'expression par l'écart-type du profil. La variance d'un profil réduit est alors de 1.



→ Réduire des profils d'expression permet de comparer les variations d'expression sans tenir compte des amplitudes.

La classification est un problème difficile !

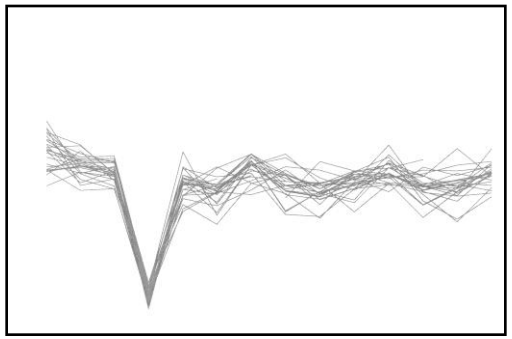


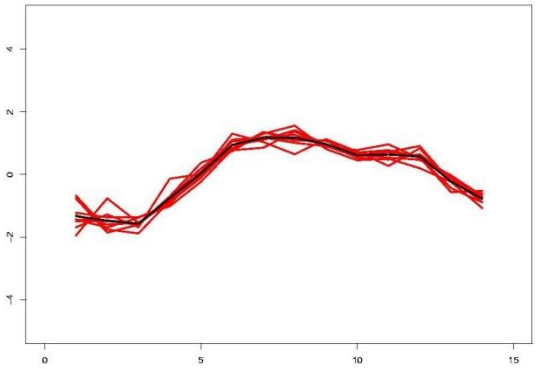
Une partie des résultats peut être différente selon la méthode utilisée !

→ Les groupes de gènes doivent être validés…

Après la classification : La validation des groupes

✓ Il faut tracer les profils d'expression :





Artéfact expérimental ?

Résultat plus réaliste !

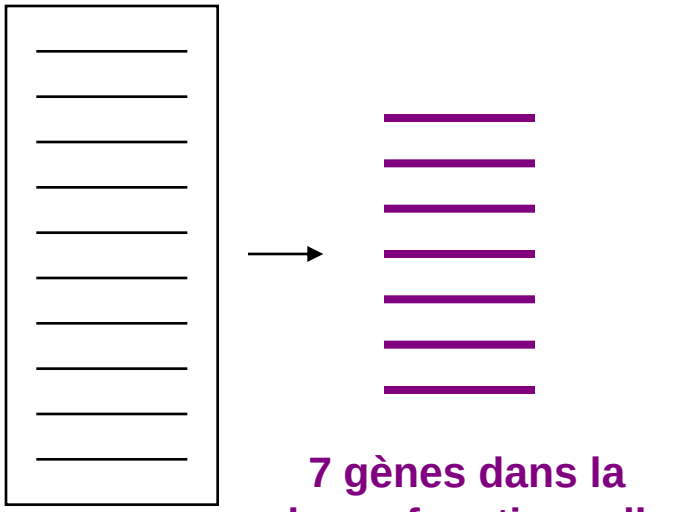
→ Conclusion :

- Il faut choisir soigneusement sa méthode d'analyse (en essayer plus d'une !)
- Les résultats doivent être croisés avec d'autres types de données (données bibliographiques, annotation fonctionnelle, bases de données…)

Analyse fonctionnelle des groupes de gènes issus de la classification

✓ Un groupe de gène est dit « fonctionnellement enrichi » en une fonction biologique si la proportion de gènes dans le groupe connus pour être impliqués dans cette fonction biologique excède le nombre attendu par le hasard.

✓ **Analyse d'une classe fonctionnelle (ex: « biosynthèse des ribosomes »)**



Liste de 10 gènes

7 gènes dans la classe fonctionnelle

f_{obs} **= 7/10 = 70%**

Sur l'ensemble du génome on sait que 215 gènes sur 6000 appartiennent à la classe fonctionnelle étudiée

f_{ref} **= 215/6000 = 3,5%**

→ Il existe différents outils WEB qui calculent la probabilité d'observer une fréquence f_{obs} par le hasard (compte tenu de la valeur de f_{ref}).

→ Plus cette probabilité est faible, plus le groupe est « fonctionnellement enrichi ».

Quelques outils « d'analyse fonctionnelle »

✓ Les différents outils disponibles diffèrent par :

- La manière dont est calculée la probabilité d'observer une fréquence f_{obs} par le hasard (loi hypergéométrique, comparaison de fréquences, etc.).
- La base de données d'annotation fonctionnelle (GO, MIPS, etc.)

FUNSPEC :

http://funspec.med.utoronto.ca/

GOMiner :

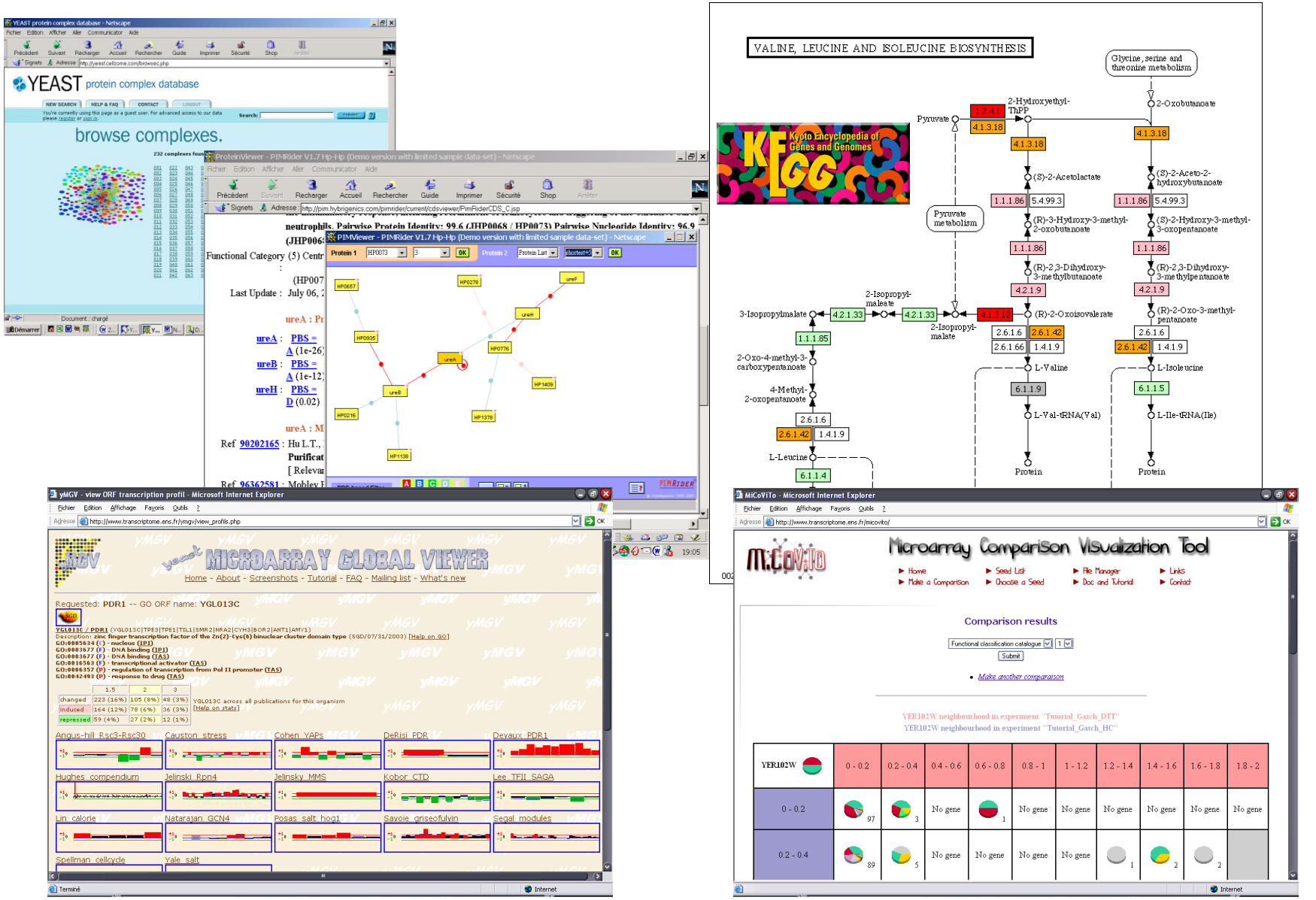
http://discover.nci.nih.gov/gominer/

GO Term Finder :

http://db.yeastgenome.org/cgi-bin/GO/goTermFinder

Et bien d'autres …

Utilisez les outils sur Internet !



Plus d'informations ?