

# Journée des utilisateurs des plateformes puces à ADN de Nantes et Rennes

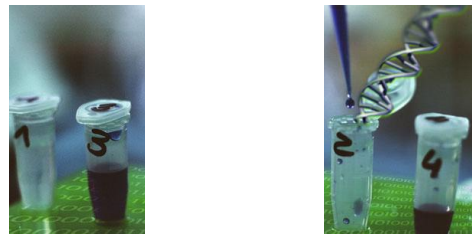
Analyse des données de  
transcriptome

# Plan

- Introduction : Principe des puces à ADN
- Conception expérimentale
- Normalisation
- Filtrage



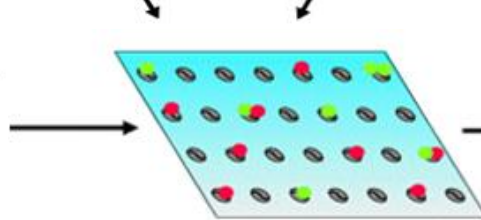
# Principe des puces à ADN



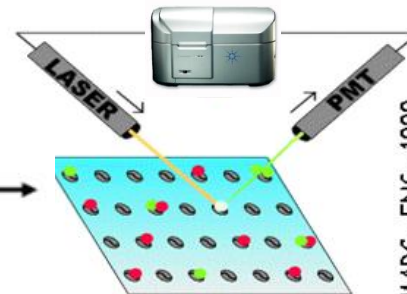
Extraction des ARN



Transcription inverse des ARNm



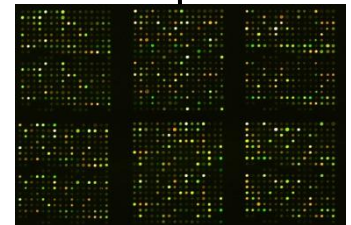
Hybridation



Lecture

Données brutes

Analyse d'image



P. MARC - ENS - 1999

# Types de puces

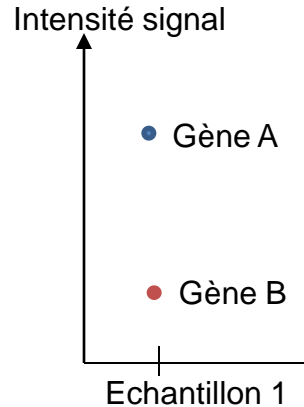
Type	Technologie	Fabrication	Nb ech/lame
Maison	Bi-couleur	Dépot	2 (test-ref)
Agilent	- Mono-couleur - Bi-couleur : « Pseudo mono-couleur »	Synthèse in situ	4 8
Affymetrix	Mono-couleur	Synthèse in situ	1
Illumina	Mono-couleur	Billes	12
Nimblegen	Mono-couleur	Synthèse in situ	multiple



# Conception Expérimentale

Une puce mesure le taux d'expression des gènes dans un échantillon

Constat d'ignorance : on ne connaît pas le taux d'expression « normal » d'un gène



Gène A a un taux d'expression plus fort que Gène B

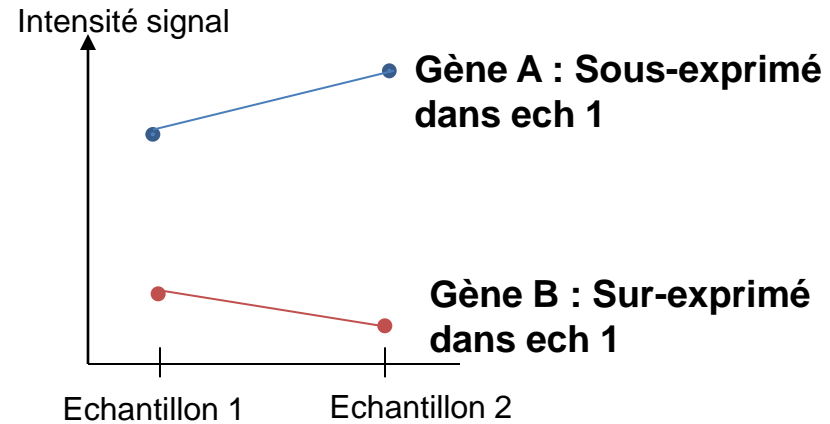
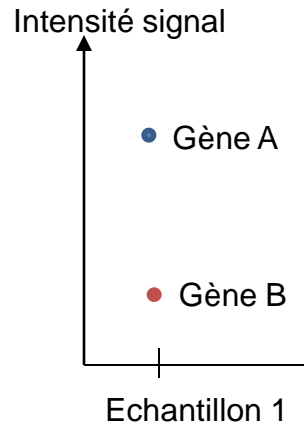
- Une faible expression d'un gène peut avoir une incidence biologique importante

- Pour un autre, il faut un niveau d'expression plus élevé pour atteindre un effet biologique

Ex : Protéines ribosomales toujours en quantité importante dans la cellule : gènes avec un signal fort quelques soient les conditions



# Conception Expérimentale



Dans l'absolu, Gène A a un taux d'expression plus fort que Gène B

Mais...

Gène A sous-exprimé dans échantillon 1  
Gène B sur-exprimé dans échantillon 1

**L'IMPORTANT est de regarder l'expression des gènes à travers les échantillons**

# Conception Expérimentale

- Le choix des échantillons est très important  
→ C'est lui qui permettra de répondre à la question posée.
- Des conditions différentes et bien choisies permettront d'identifier des gènes différentiels entre ces conditions  
→ Notion de classes d'échantillons

**Importance de bien définir les classes des échantillons**



# Conception Expérimentale

## Types d'expérience

Type d'expérience	Question posée
Différentiel	Quels gènes différentient ma condition A de ma condition B ? → Recherche de biomarqueurs
Cinétique	Peut-on trouver une cinétique dans l'expression de certains gènes ?
Recherche de sous-classes	Ai-je plusieurs classes d'échantillons ? Quels gènes me permettent de distinguer ces classes ? Ex : gravité d'une pathologie

# Conception Expérimentale

= Organiser l'expérience pour répondre efficacement aux questions posées compte tenu des contraintes expérimentales et matérielles

Pour limiter les confusions d'effets :

- Biais expérimentateur
- Biais du jour d'hybridation
- Biais d'amplification
- Biais de marquage

**Il faut pouvoir distinguer un effet artéfactuel  
d'un effet biologique**



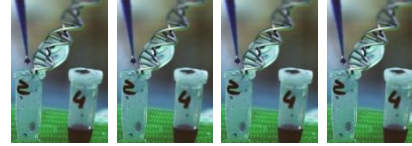
# Conception Expérimentale : Ex 1

Expérimentateur 1

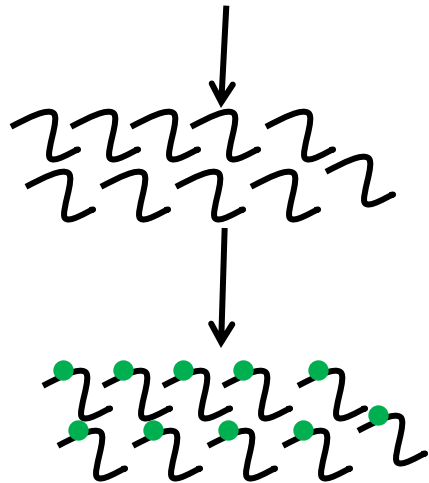


Echantillons condition **A**

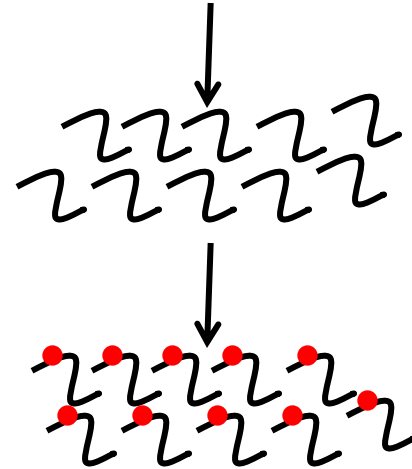
Expérimentateur 2



Echantillons condition **B**



Extraction des ARN



Marquage

La différence d'expression des gènes entre la condition A et la condition B est due à un effet biologique ou à un effet expérimentateur ?

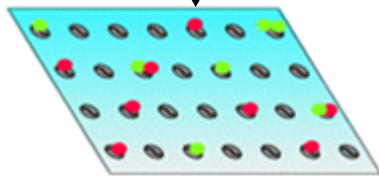
**Avec une telle conception, ON NE PEUT PAS REpondre !!**

# Conception Expérimentale : Ex 2

Expérimentateur 1

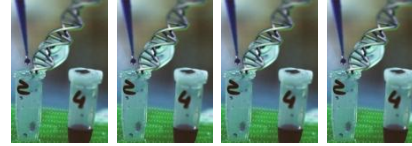


Echantillons condition **A**

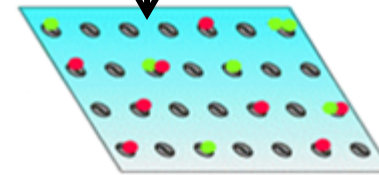


Jour J1

Expérimentateur 1



Echantillons condition **B**



Jour J2

Extraction  
Marquage

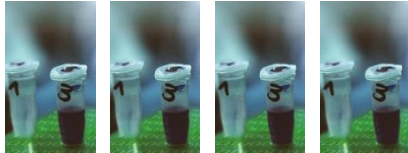
Hybridation

La différence d'expression des gènes entre la condition A et la condition B est due à un effet biologique ou à un effet jour d'hybridation ?

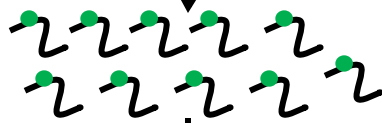
**Avec une telle conception ON NE PEUT PAS REpondre !!**

# Conception Expérimentale

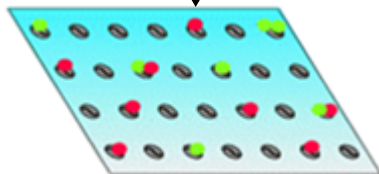
Expérimentateur 1



Echantillons condition **A**

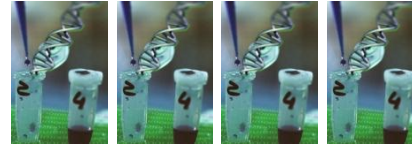


Extraction (J1)  
Marquage (J2) A + B

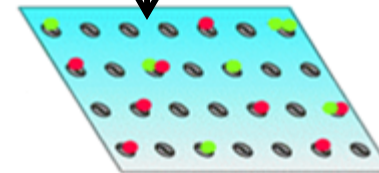


Hybridation J3  
Mélange A + B

Expérimentateur 1



Echantillons condition **B**



Il faut limiter les confusions en « randomisant » les étapes

Ainsi on peut dissocier un effet artéfactuel et un effet biologique

# Conception Expérimentale

- Nombre d'échantillons :
  - Triplicats indépendants au minimum par classe ,
  - 20 échantillons minimum au total.

Plus il y a de réplicats biologiques, plus les tests statistiques sont robustes.

- Critères d'inclusion des échantillons :
  - 100-300 ng d'ARNt
  - RIN > 8 (Bioanalyseur Agilent)

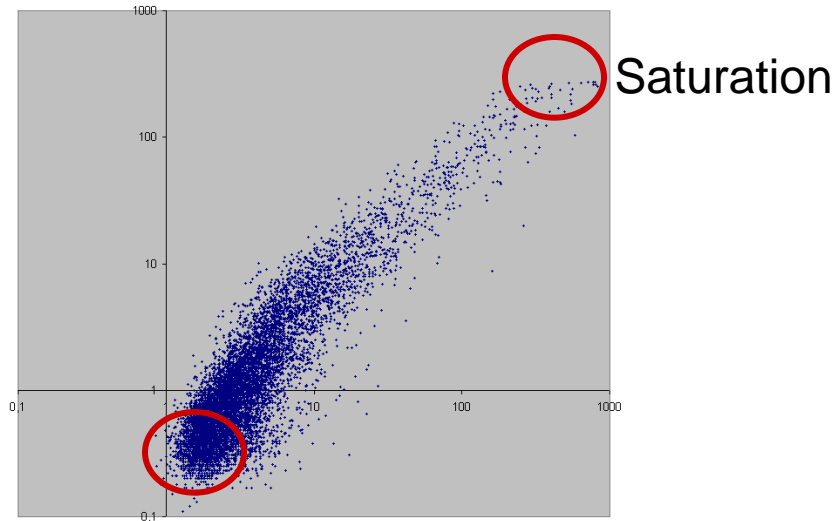


# Normalisation : Pourquoi ?

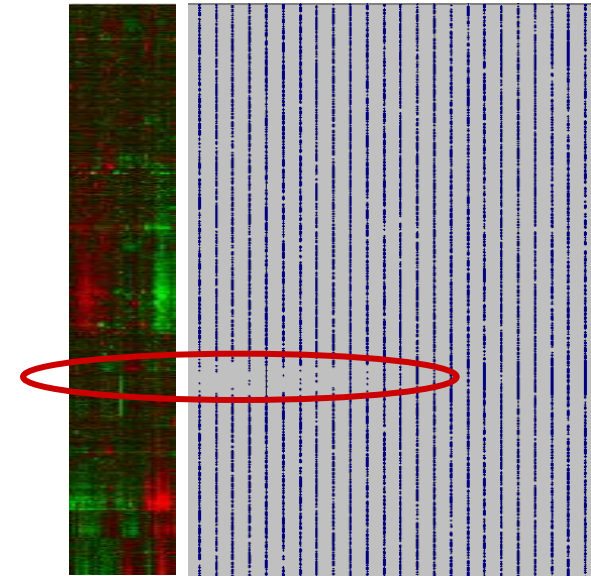
- Pour corriger les effets liés à l'expérimentation :
  - Qualité du dépôt (puces maison)
  - Qualité de l'ARN
  - Qualité du marquage
  - Qualité de l'hybridation
  - Qualité du lavage
  - Différences d'incorporation des fluorochromes
  - Différences à la lecture
  - etc...



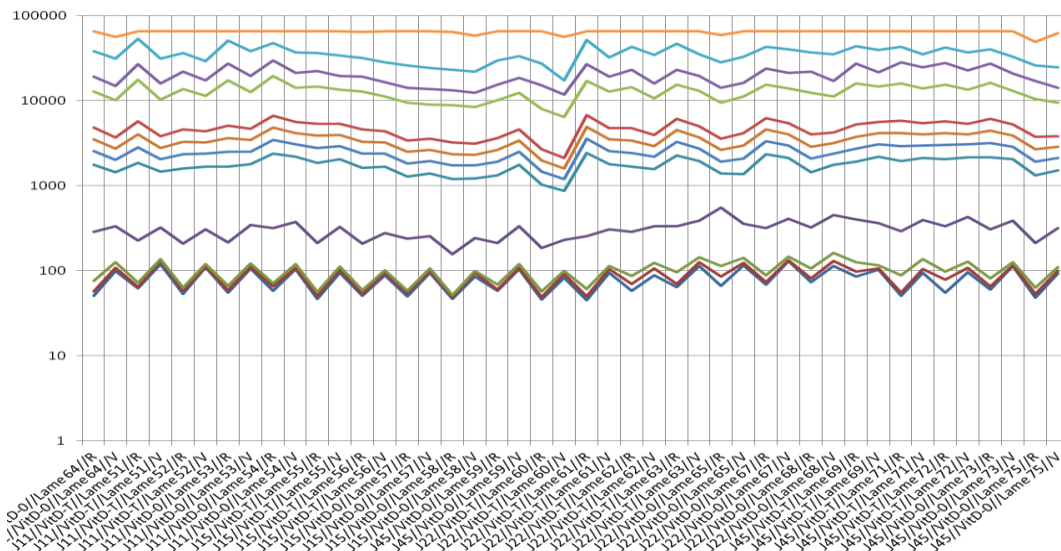
# Normalisation : correction des biais



Bruit de fond



Biais de blocs

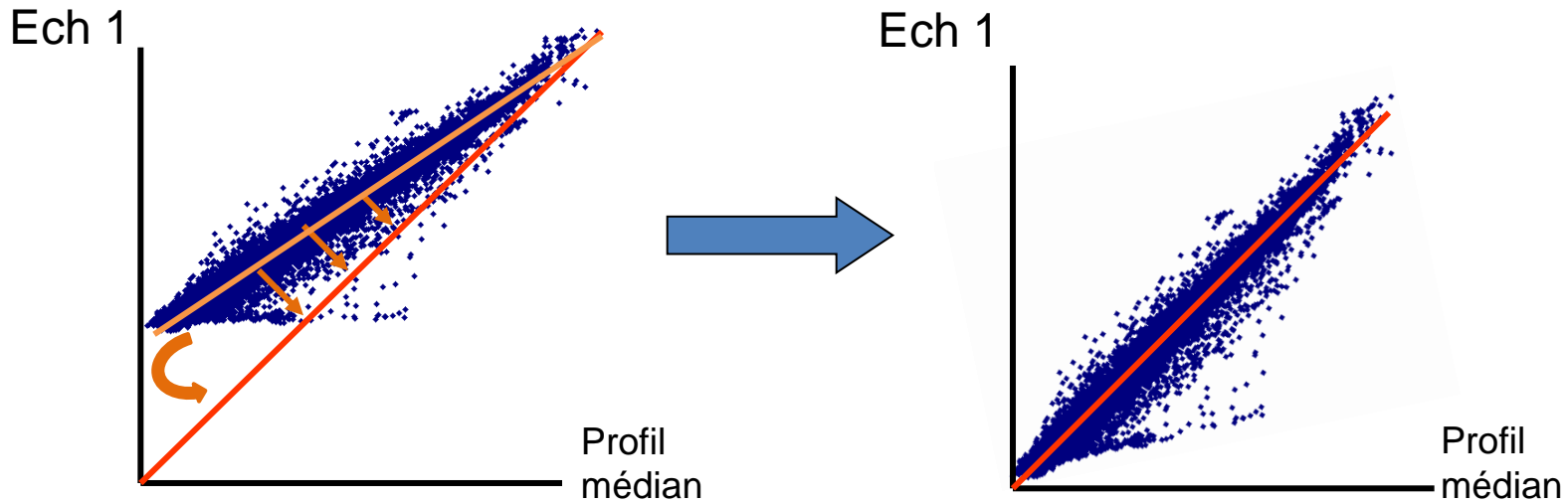


Distribution inégale du signal de chaque échantillon

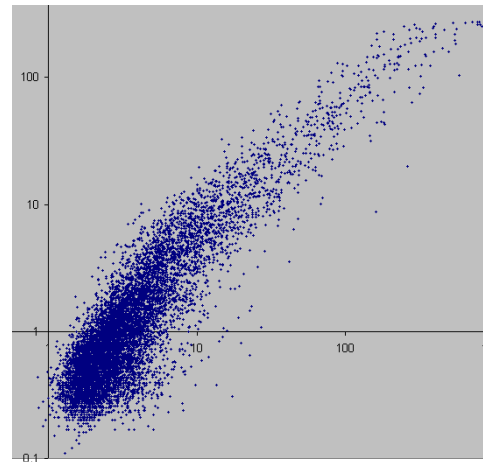


# Normalisation : régression

## Régression linéaire : correction linéaire

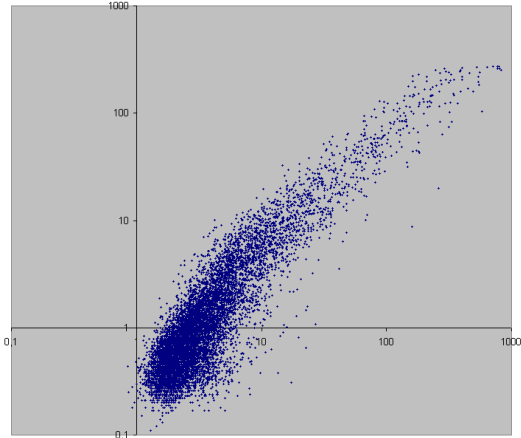


**Ne corrige pas les effets non linéaires**

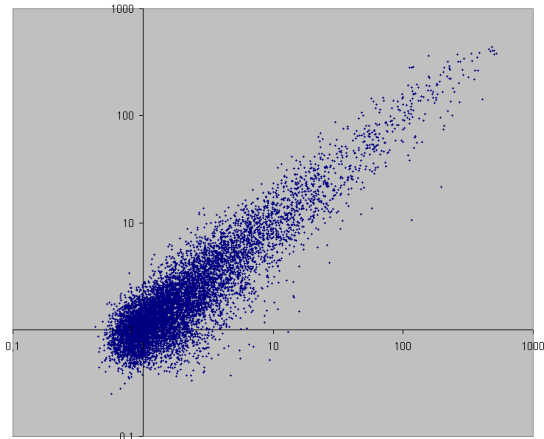


# Normalisation : Lowess

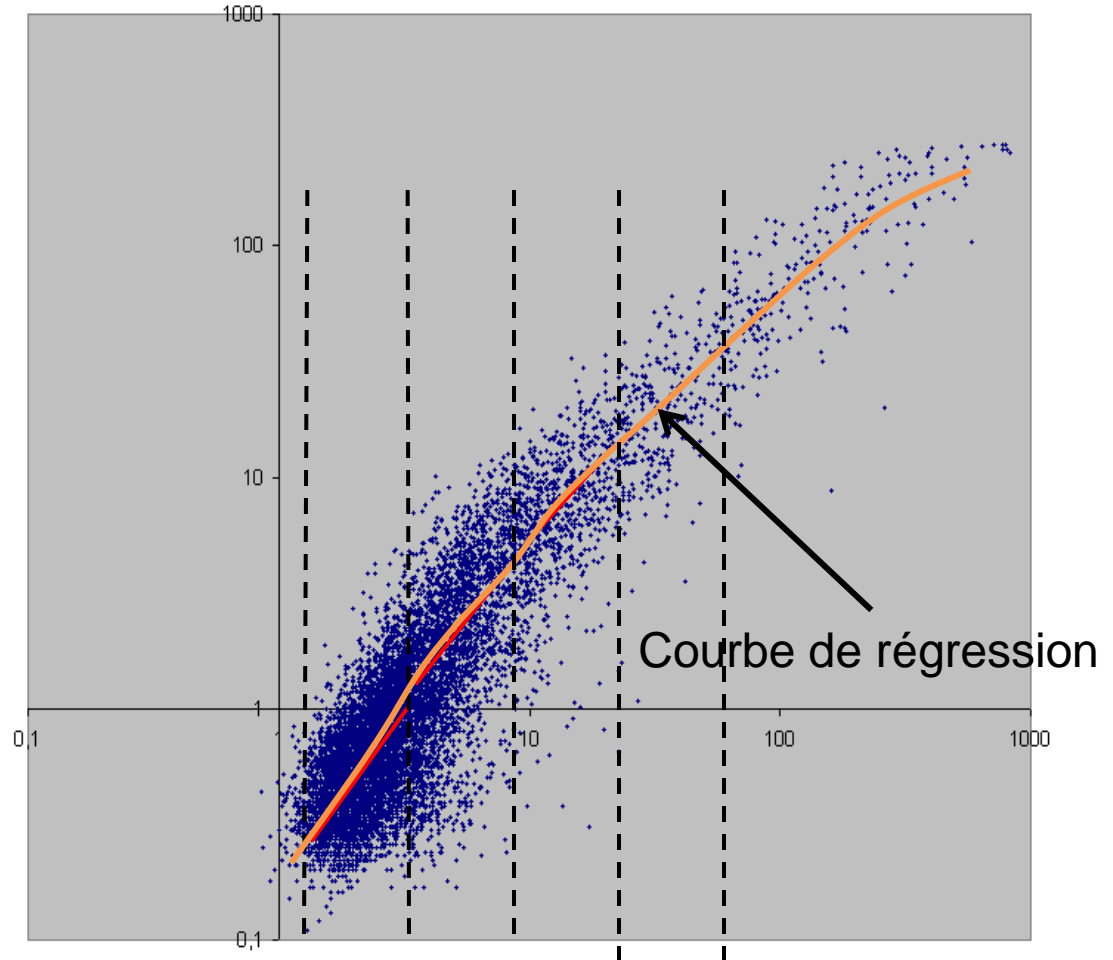
Lowess (Yang et al. 2002 NAR 30:e15)



Avant Normalisation



Après Normalisation

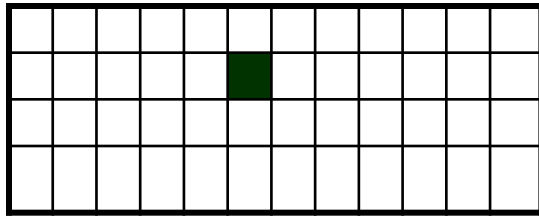


Régression locale  
par fenêtre glissante

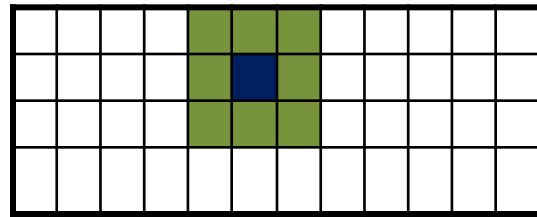
# Normalisation : Lowess

## Print-tip Lowess

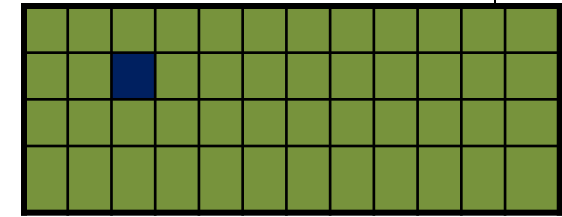
Pin



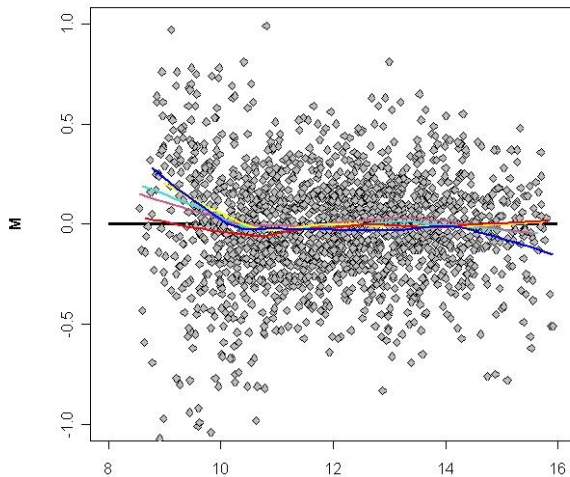
Proximal



Array

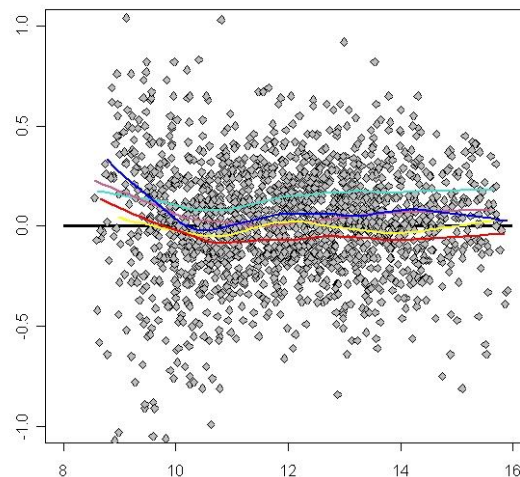


Pin



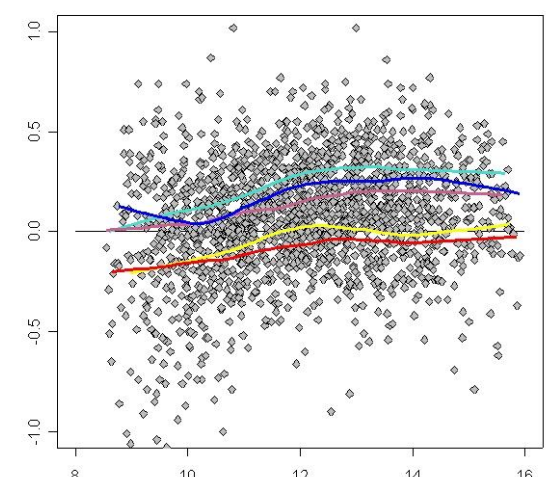
A

Proximal

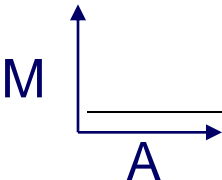


A

Array



A

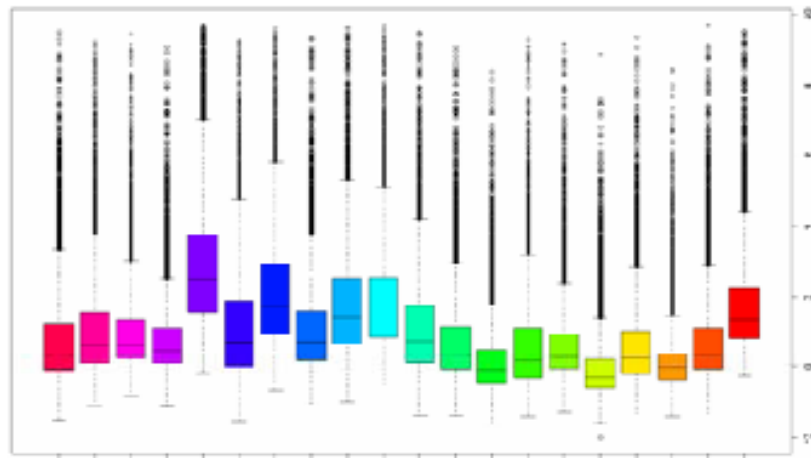


Plateforme Puces à ADN

(Le Meur, N. et al., Nucl. Ac. Res., 2004)

# Normalisation globale (inter-array)

- Centrage médian des échantillons :



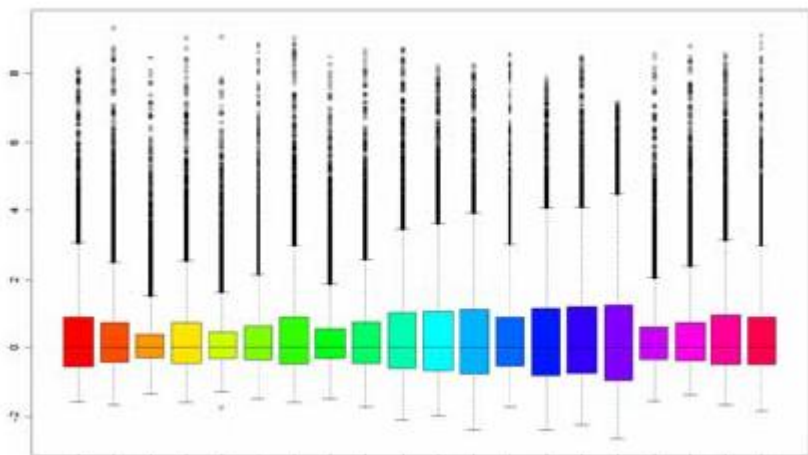
Données brutes



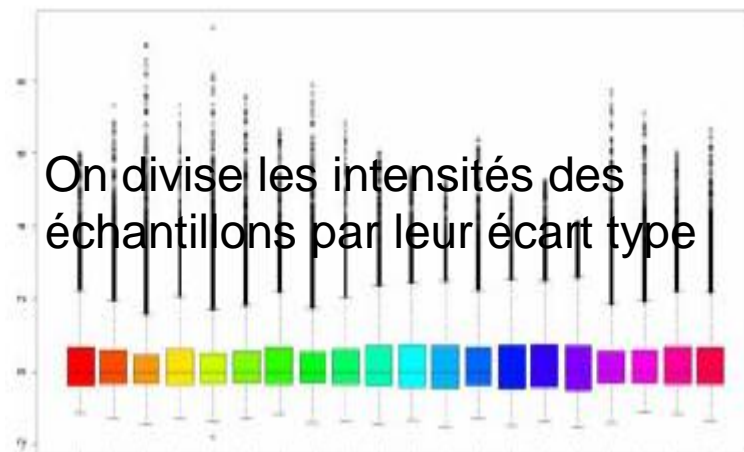
Données centrées sur la médiane

# Normalisation globale

- Réduction des données :  
Correction de la dispersion entre les puces



Données centrées



Données centrées réduites

# Normalisation par Quantile

	Sample 1	Sample 2	Sample 3	
Gene 1	20	10	350	
Gene 2	100	500	200	
Gene 3	300	400	30	

Données brutes

	Sample 1	Sample 2	Sample 3	Mediane
Quantile 1	20	10	30	20
Quantile 2	100	400	200	200
Quantile 3	300	500	350	350

Trier chaque échantillon en ordre croissant  
Calculer la médiane par ligne

	Sample 1	Sample 2	Sample 3	Mediane
Quantile 1	20	20	20	20
Quantile 2	200	200	200	200
Quantile 3	350	350	350	350

Remplacer la valeur de chaque ligne par la médiane

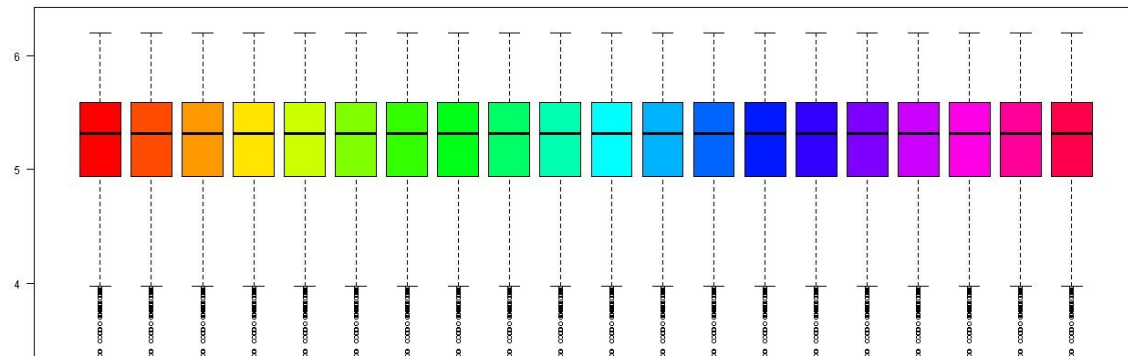
	Sample 1	Sample 2	Sample 3	Mediane
Gene 1	20	20	350	20
Gene 2	200	350	200	200
Gene 3	350	200	20	350

Restaurer l'ordre initial



# Normalisation par Quantile

**Rend les distributions identiques :** Chaque échantillon contiendra le même jeu de valeurs.



## Avantages :

- Rapide
- Efficace

## Problème :

Peut donner du poids à certaines sondes faibles

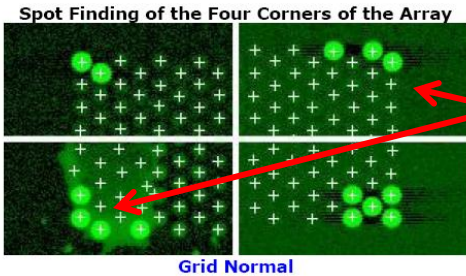
# Filtrage

- Des échantillons : détection des lames défectueuses
- Des gènes : flags (saturation, BG, forme)
- Bruit de fond : soustraction
- Filtrage sur les valeurs : intensité, Fold-change, CV



# Filtrage des échantillons

- QC report des logiciels d'analyse d'image

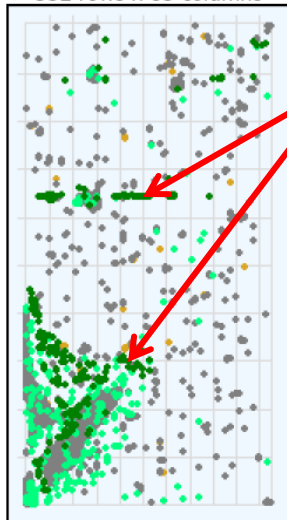


Bruit de fond

Grid Normal

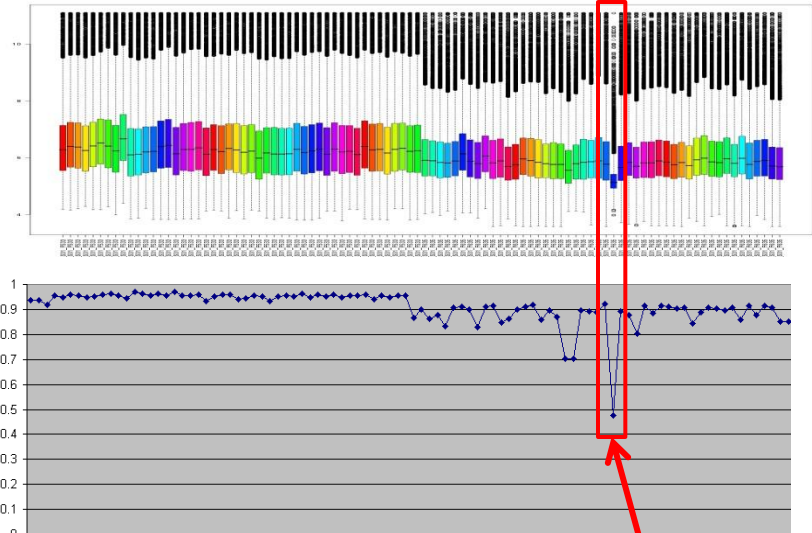
Non Uniform	194	368
Population	281	1833

Spatial Distribution of All Outliers on the Array  
532 rows x 85 columns

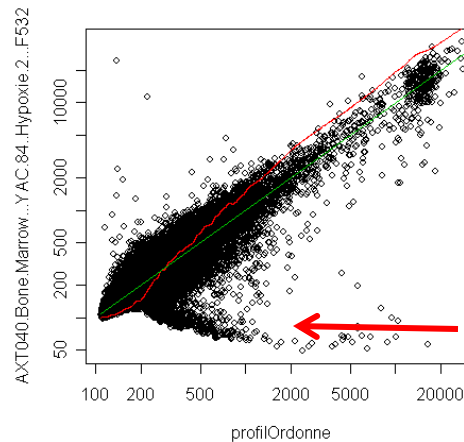


Spots non uniformes

- Statistiques descriptives



Echantillon différent des autres



Scatter plot anormal



# Filtrage des sondes

- Flags : Saturation, forme, homogénéité donnés par le logiciel d'analyse d'images
- Bruit de fond : soustraction



Génère des valeurs négatives

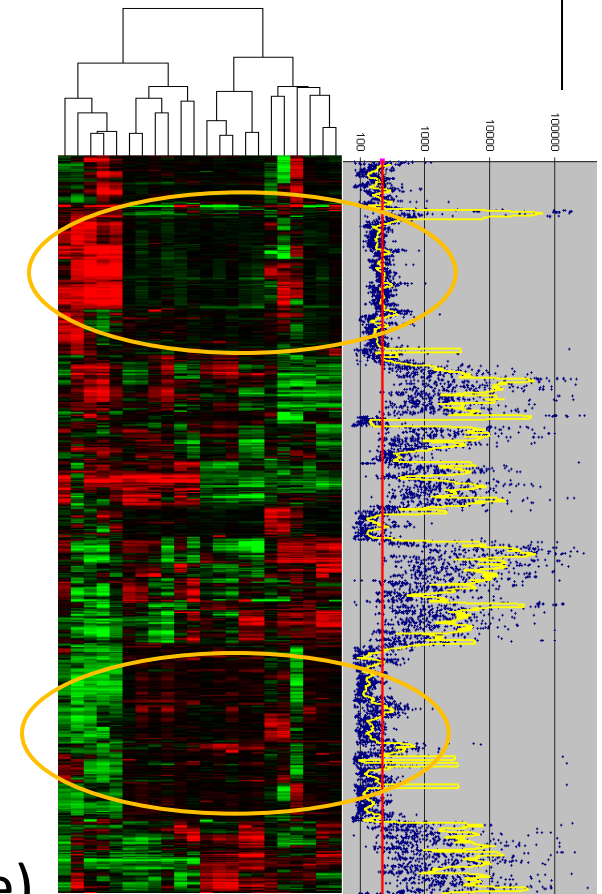


Valeurs manquantes dans la matrice  
Pose problème pour les tests et le clustering

# Filtrage des sondes de valeurs faibles

Les sondes faibles ne sont pas fiables  
et peuvent donc se retrouver  
différentiellement exprimées

- Estimation du bruit de fond
  - sur la distribution
  - ou utilisation des contrôles négatifs
- Pourcentage des valeurs par sonde  $\leq$  bruit de fond  
(sur la totalité des échantillons ou par classe)



Bruit de fond différentiel

**Elimination de la sonde si trop de valeurs faibles**

# Filtrage des sondes de faibles variations

Elles n'apportent aucune information.

Les éliminer permet de diminuer le nombre de tests statistiques donc de diminuer le nombre de faux positifs.

Exemples de paramètres pour le filtrage :

- Ecart type  $\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$
- Coefficient de variation  $CV = \frac{\sigma}{\mu}$



# Filtrage des sondes avec trop de valeurs manquantes

- **Regroupement anarchique des sondes au clustering car vecteurs trop petits**

Gène 1	100	<del>80</del>	<del>null</del>	<del>13</del>	<del>null</del>	<del>56</del>	<del>null</del>	<del>67</del>	<del>45</del>	78
Gène 2	56	<del>null</del>	<del>78</del>	<del>null</del>	<del>28</del>	<del>null</del>	<del>23</del>	<del>null</del>	<del>null</del>	30

2 mesures pour le calcul de la corrélation

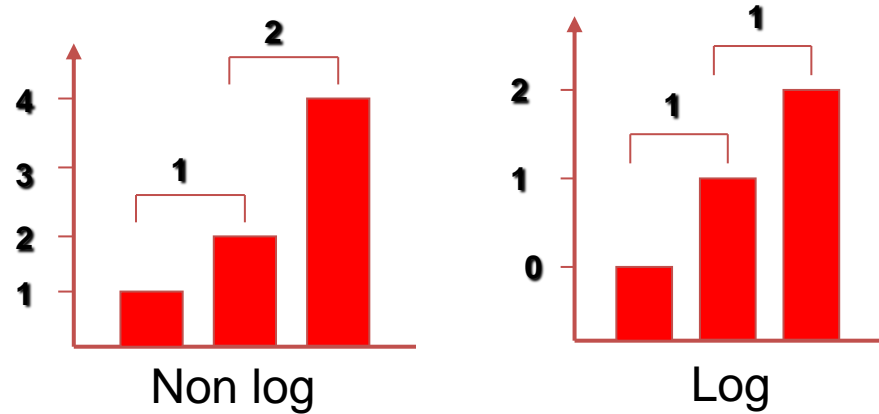
Coefficient de corrélation = 1

- **Trop peu de valeurs pour les tests statistiques : diminution de la robustesse**

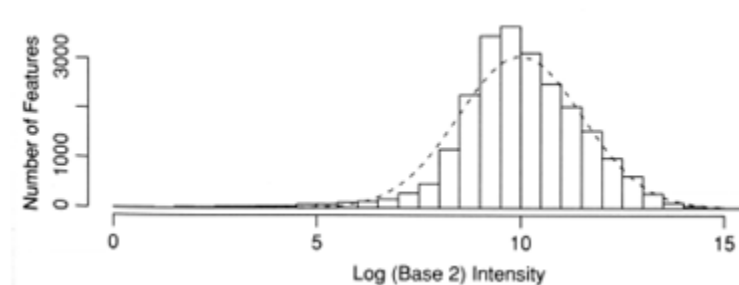
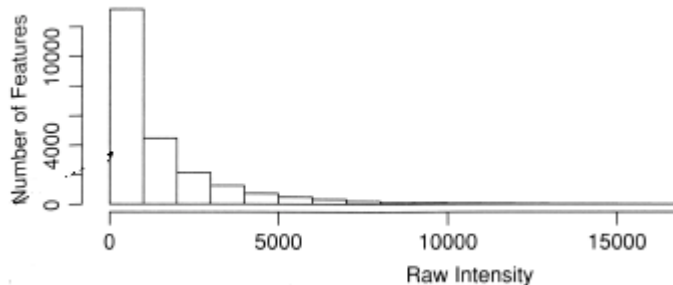


# Transformation logarithmique

- Donne le même poids à une augmentation ou une diminution d'un facteur 2



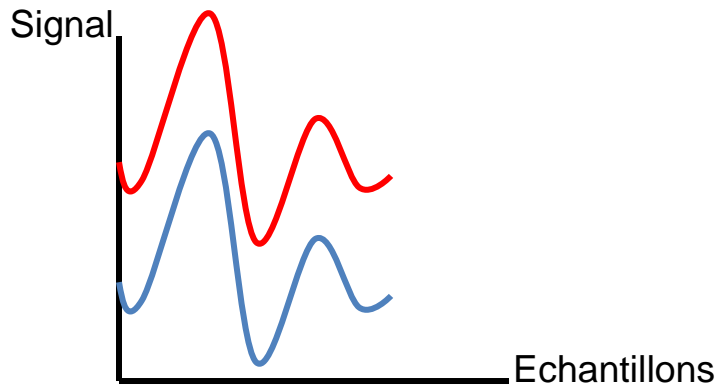
- Après transformation log, la distribution des valeurs se rapproche d'une gaussienne



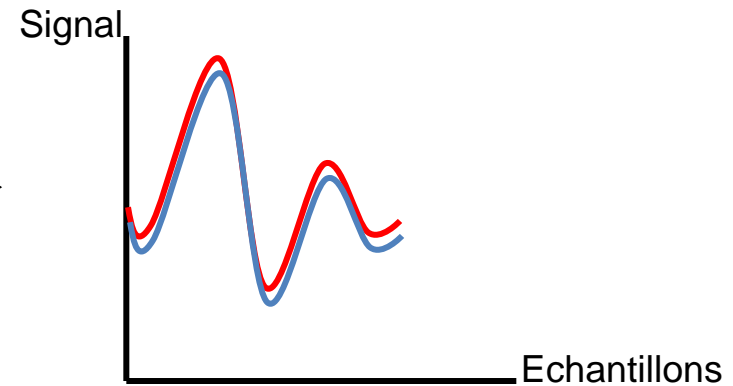
- Evite de favoriser les valeurs extrêmes

# Transformation des données

- Centrage médian des gènes :



Données brutes ou normalisées



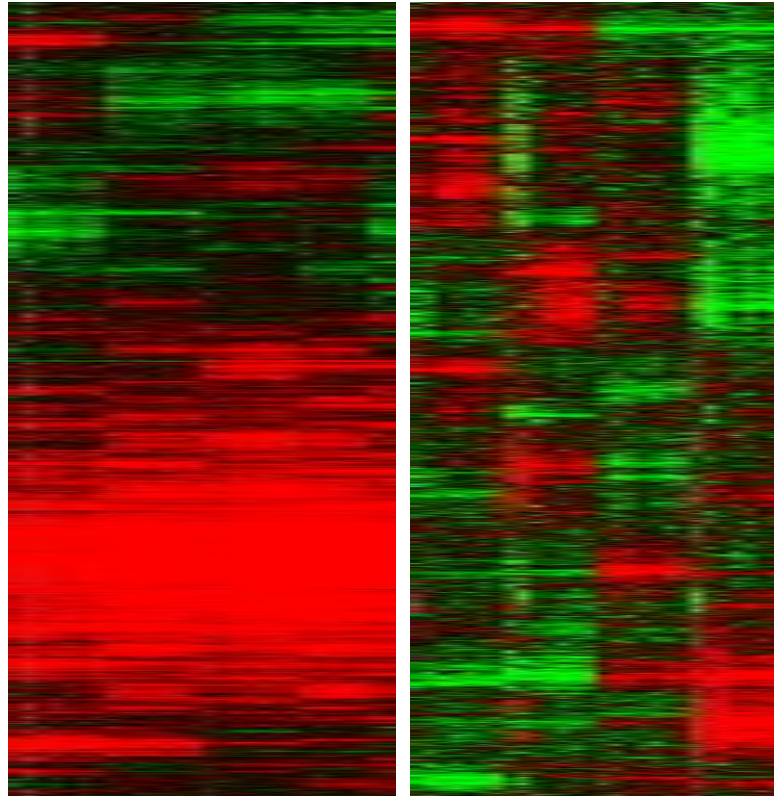
Données centrées sur la médiane

Indispensable pour la classification hiérarchique

# Transformation des données

- Centrage médian des gènes

Si l'on classe les données telles quelles, on observe que certains gènes sont faiblement exprimés, d'autres fortement.



Ce qui nous intéresse c'est de savoir si un gène s'exprime différemment dans certains échantillons.

Le centrage médian des gènes permet de visualiser de tels changements.

C'est aussi un constat d'ignorance: le niveau « normal » d'un gène, c'est la médiane

# Outils

- Excel
- Scripts R :
  - Lowess
  - Packages Bioconductor (ex : limma) : fonctions pour la normalisation et le filtrage
- Outils en ligne :
  - TIGR MeV : <http://www.tm4.org/mev/>
  - GEPAS : <http://gepas.bioinfo.cipf.es/>
- Outils intégrés payants :
  - Genespring
  - Partek

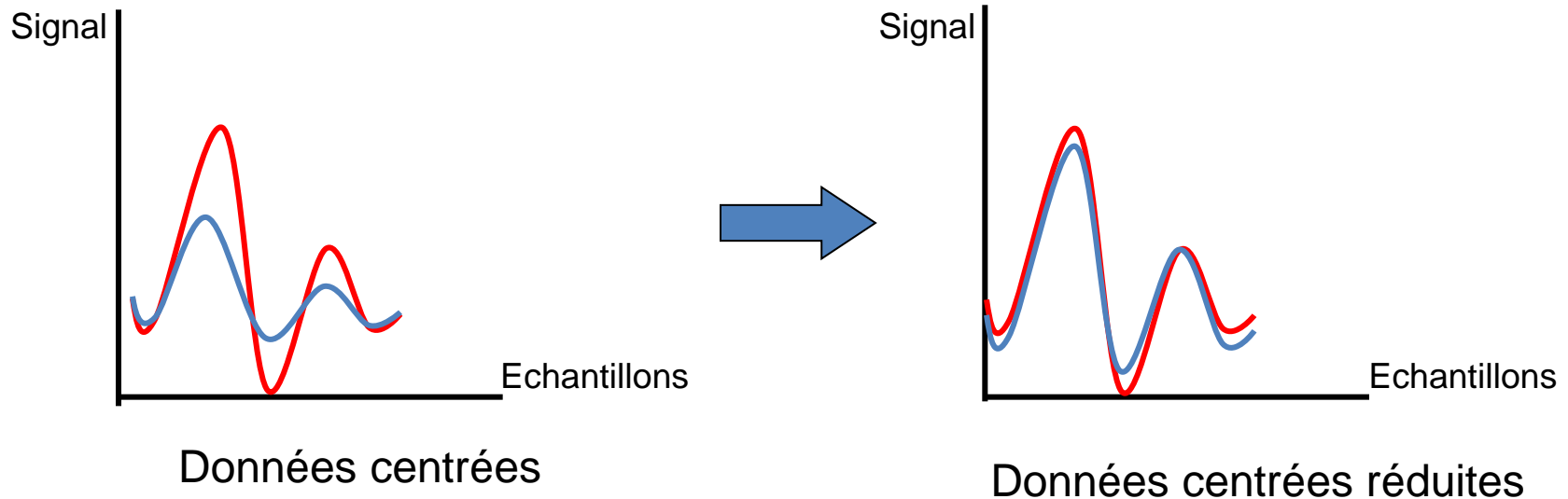


# Des Questions ??



# Transformation des données

- Réduction des données :  
**Correction de la dispersion entre les gènes**



Ainsi on compare les variations d'expression sans tenir compte des amplitudes