

Transcriptome

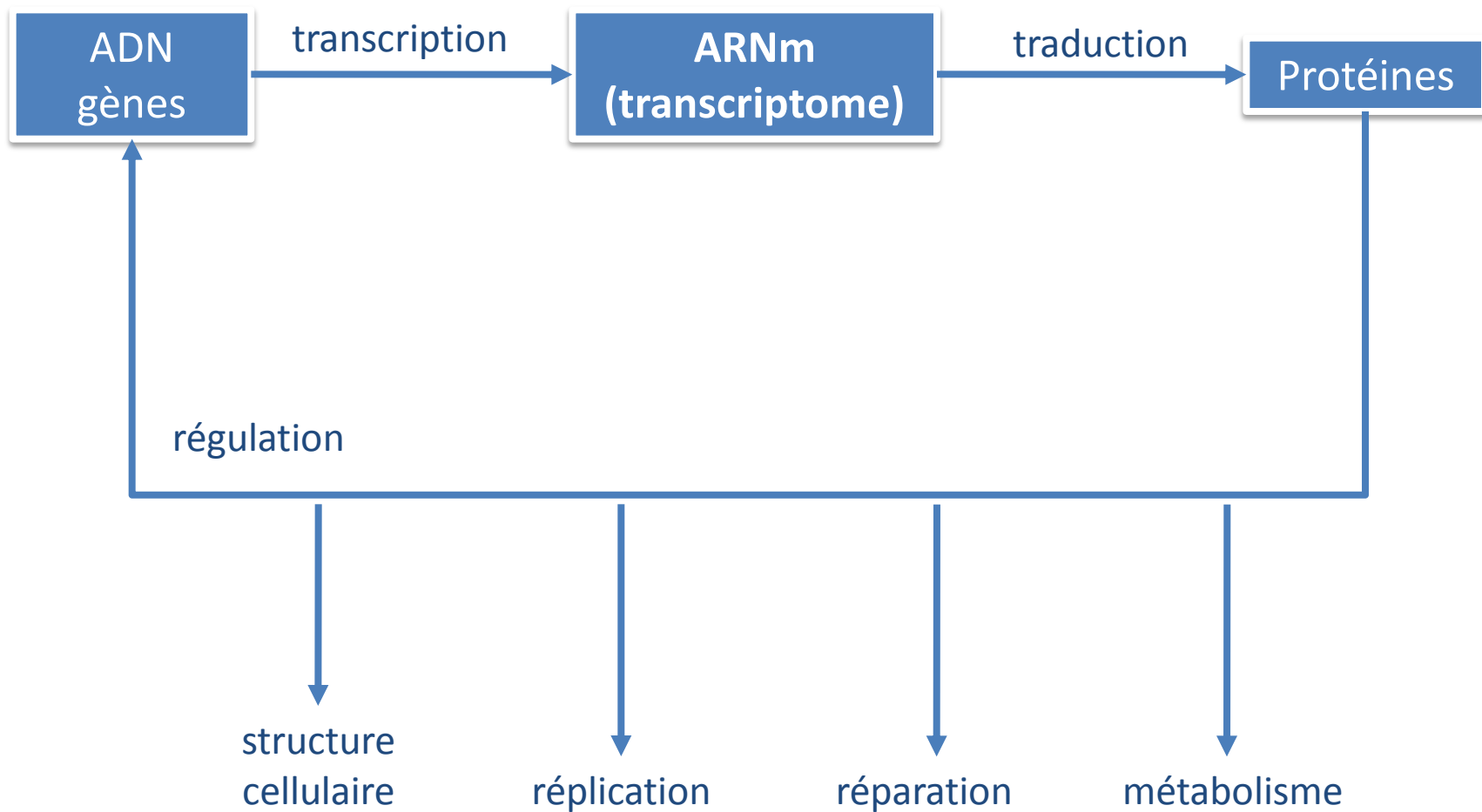
- Transcriptome : ensemble des ARNm ou transcrits présents dans une cellule ou une population de cellules dans des conditions données.
- Plan
 - Introduction
 - Acquisition des données
 - Description des données
 - Transformation, normalisation et filtrage
 - Analyse des données de transcriptome
 - Gènes différentiellement exprimés
 - Gènes co-exprimés
 - Interprétation
 - Caractérisation d'un ensemble de gènes

Applications

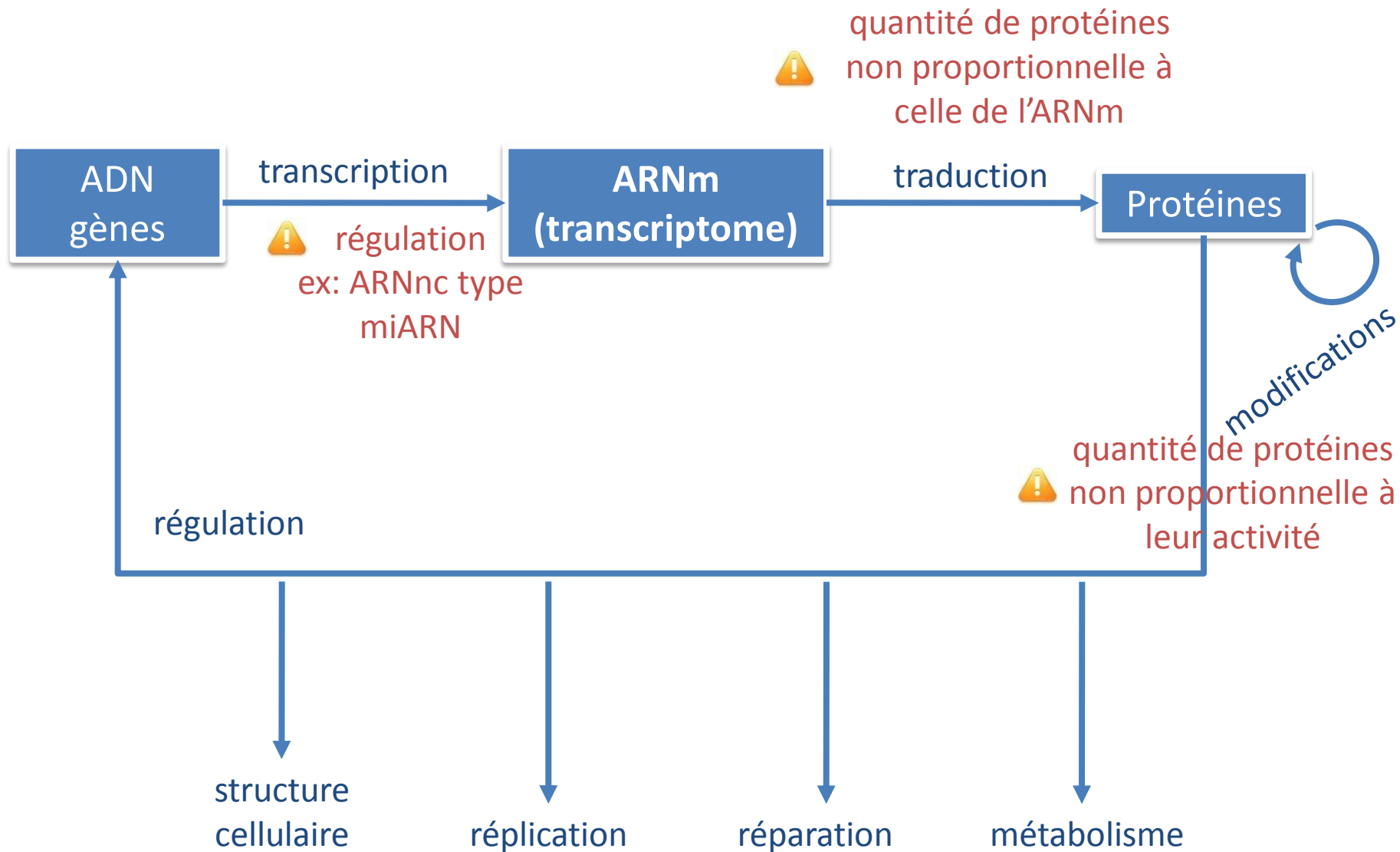
Accès au niveau d'expression de milliers de gènes simultanément

- Indication sur la **fonction** des gènes ou implication des gènes dans des **processus biologiques**
- **criblage** antérieur à des expérimentations plus ciblées, plus longues et plus coûteuses
- Reconstruction de réseaux de régulation (cinétique)
- Exemples
 - Traitement chimique, antibiotique, ... : gènes de résistance, processus biologique (ex: transformation et compétence) , toxicité
 - Tissus sain vs. tissus malade
 - cancer : oncogènes et gènes suppresseurs, diagnostique clinique et traitement adapté
 - Organes différents : gènes spécifiques et « gènes de ménage »
 - Différents stades de développement : gènes impliqués au cours des différentes phases

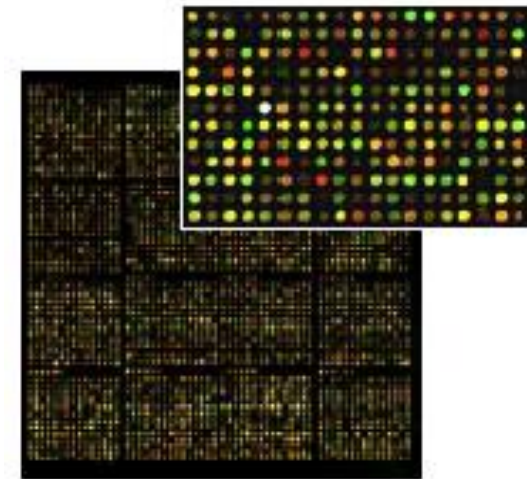
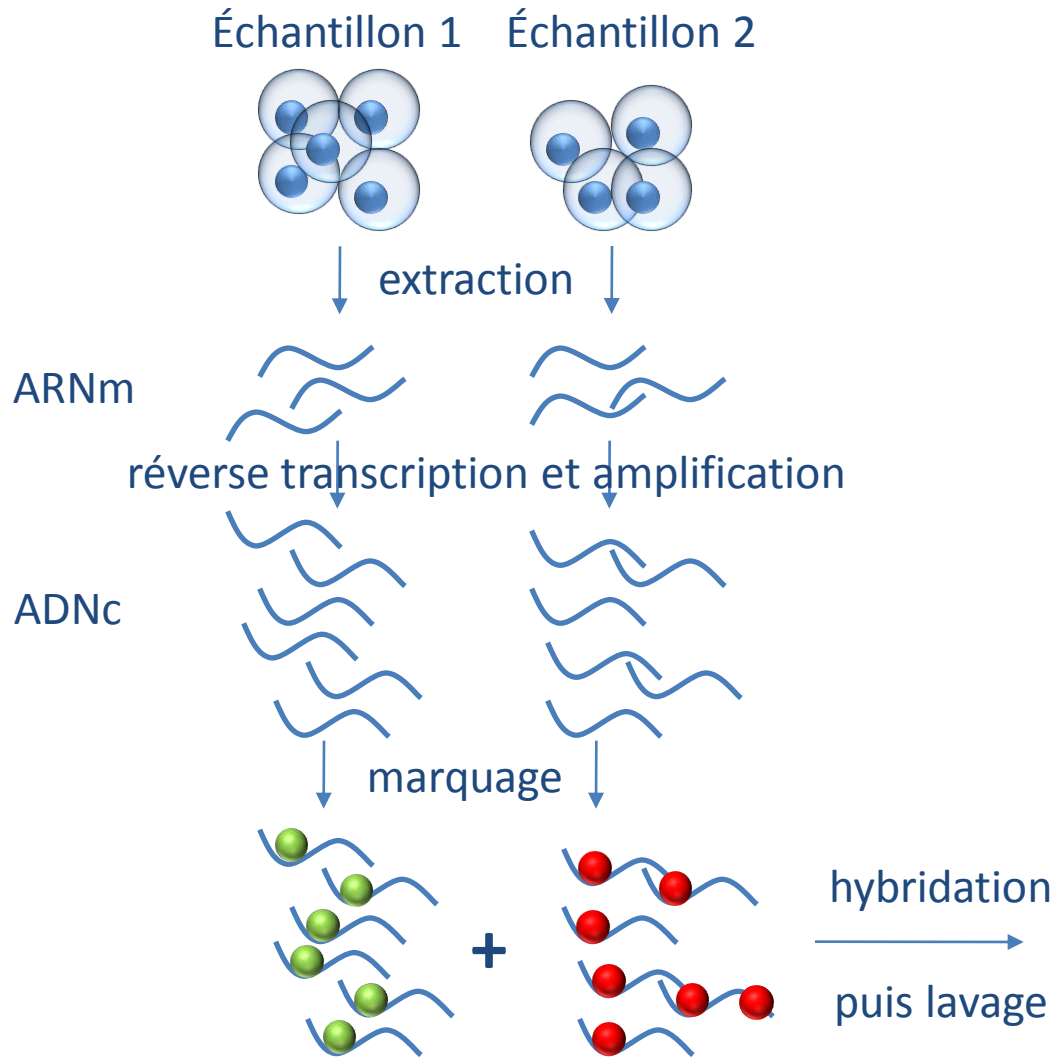
Contexte



Contexte



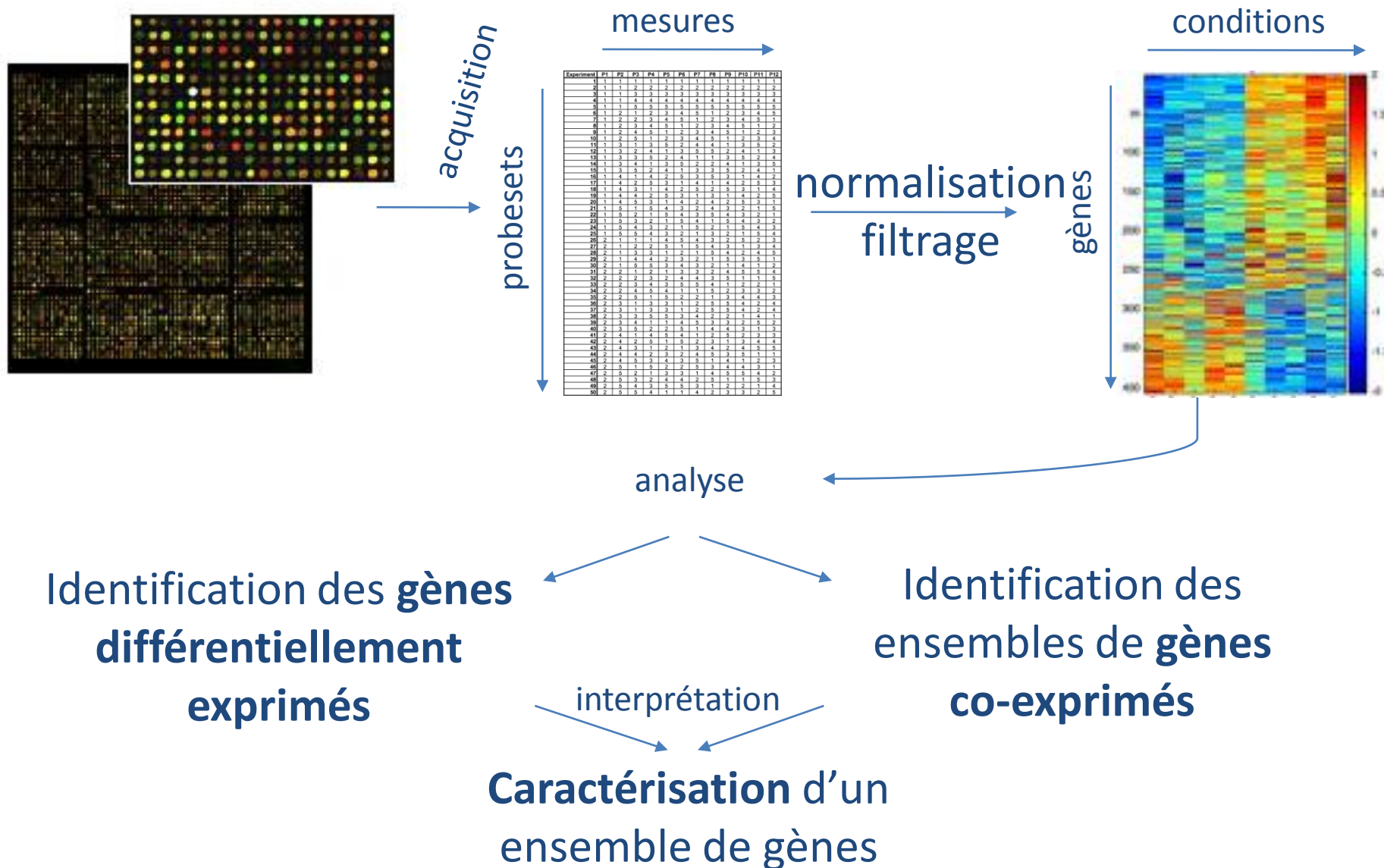
Acquisition des données



↑ scan



Analyse et interprétation des données

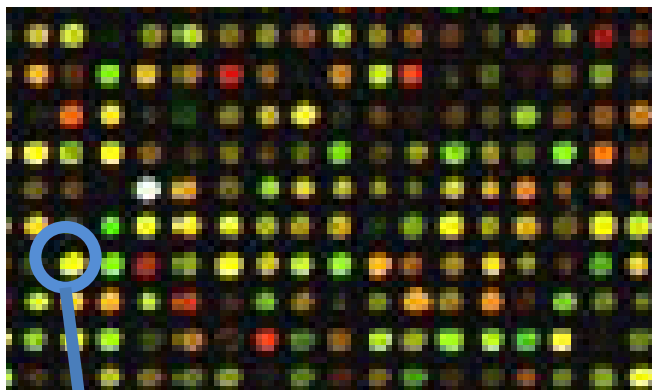


Données de transcriptome

- Accès au niveau d'expression de milliers de gènes simultanément
 - Intensité de fluorescence par spot
 - proportionnelle à la quantité d'ADN hybridé
 - **abondance relative des transcrits** : ratio (quantification absolue encore difficile)

Mesure du niveau d'expression

- ♦ échantillon 1 = fluorochrome vert (Cy3)
- ♦ échantillon 2 = fluorochrome rouge (Cy5)



2 canaux :

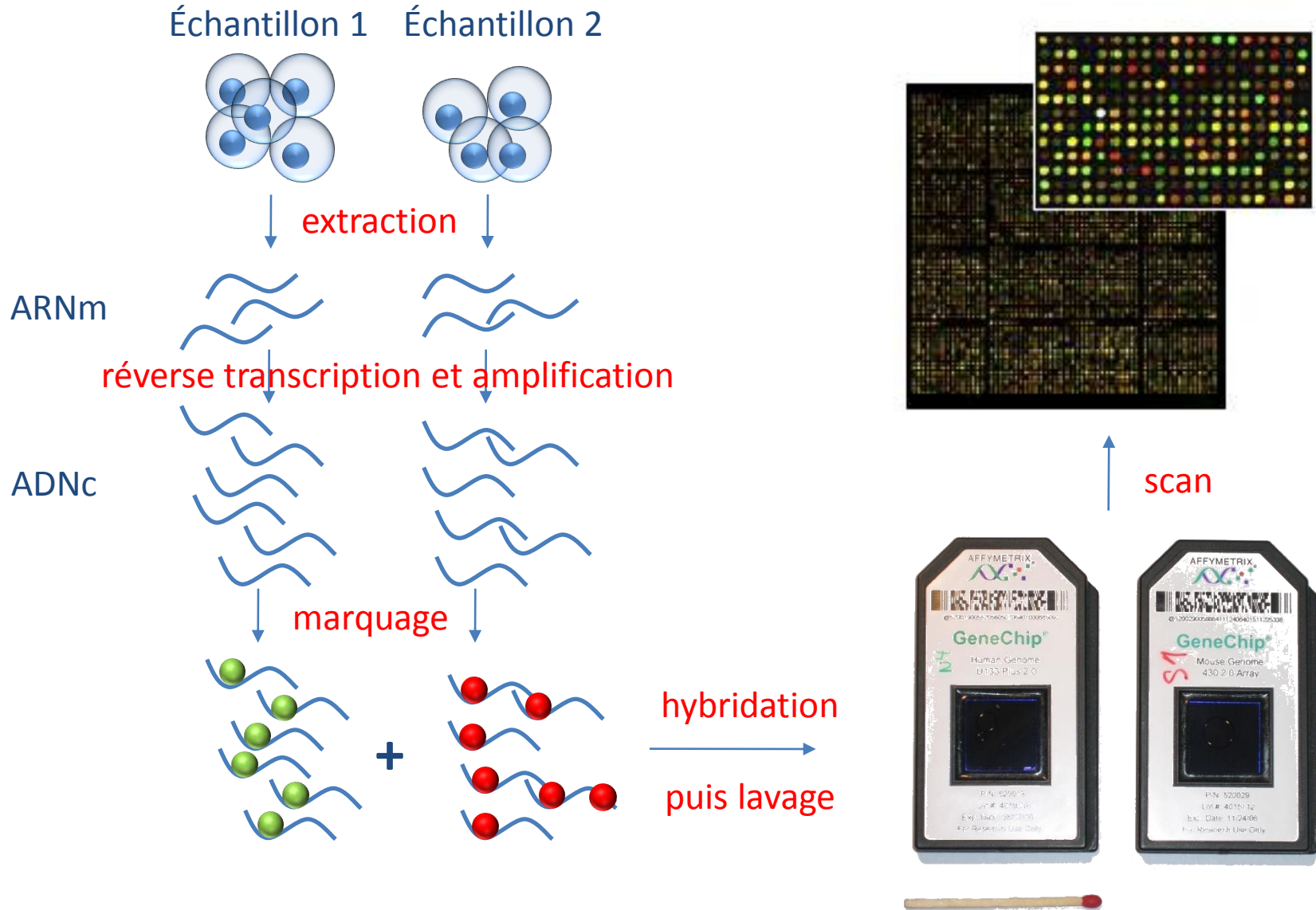
- intensité de vert
- intensité de rouge

- ⚠ 1 spot = ensemble d'oligonucléotides
 - ♦ tous les mêmes
 - ♦ variations de séquence
 - ♦ plusieurs séquences spécifiques d'un gène
- ⚠ des spots différents peuvent correspondre au même gène
- ⚠ un spot peut correspondre à plusieurs gènes

Données de transcriptome

- De **nombreuses sources d'erreur et de variabilité**
 - Variabilité biologique
 - Population de cellules ou patients/tissus différents
 - Variabilité technique
 - Étape d'amplification
 - Incorporation des fluorochromes
 - Bruit (artefacts, bruit de fond)
 - Données manquantes (mesures absentes pour certains réplicats)
 - Erreur, ex : Saturation
 - du scanner pour les fluorochromes
 - de la plaque pour la radioactivité
 - du spot sur la puce

Acquisition des données



Données de transcriptome

- Solution : réplicats & traitement statistique
 - ⚠ Nombre de réplicats augmente la fiabilité des résultats
 - Réplicat biologique & réplicat technique

Réplicats & validation

- Motivation
 - Variabilité des mesures
 - 2 expériences de puces avec les mêmes paramètres produisent des résultats (légèrement) différents
 - estimer l'erreur non systématique associée à une mesure
 - évaluer le niveau de variabilité des mesures

Réplicats & validation

- Nombre et nature des réplicats dépendent des objectifs de l'étude
 - réplikat technique : plusieurs extraits d'un même échantillon
 - ex: dye swap
 - variabilité due au bruit expérimental
 - réplikat biologique : échantillons différents
 - provenant d'expériences menées en parallèle
 - ex: population de cellules, patient différent
 - variabilité « naturelle » d'un système

Réplicats

Réplicats
biologiques

Expérience

Réplicat 1

Réplicat 2

Réplicats
techniques

Extrait 1

Extrait 2

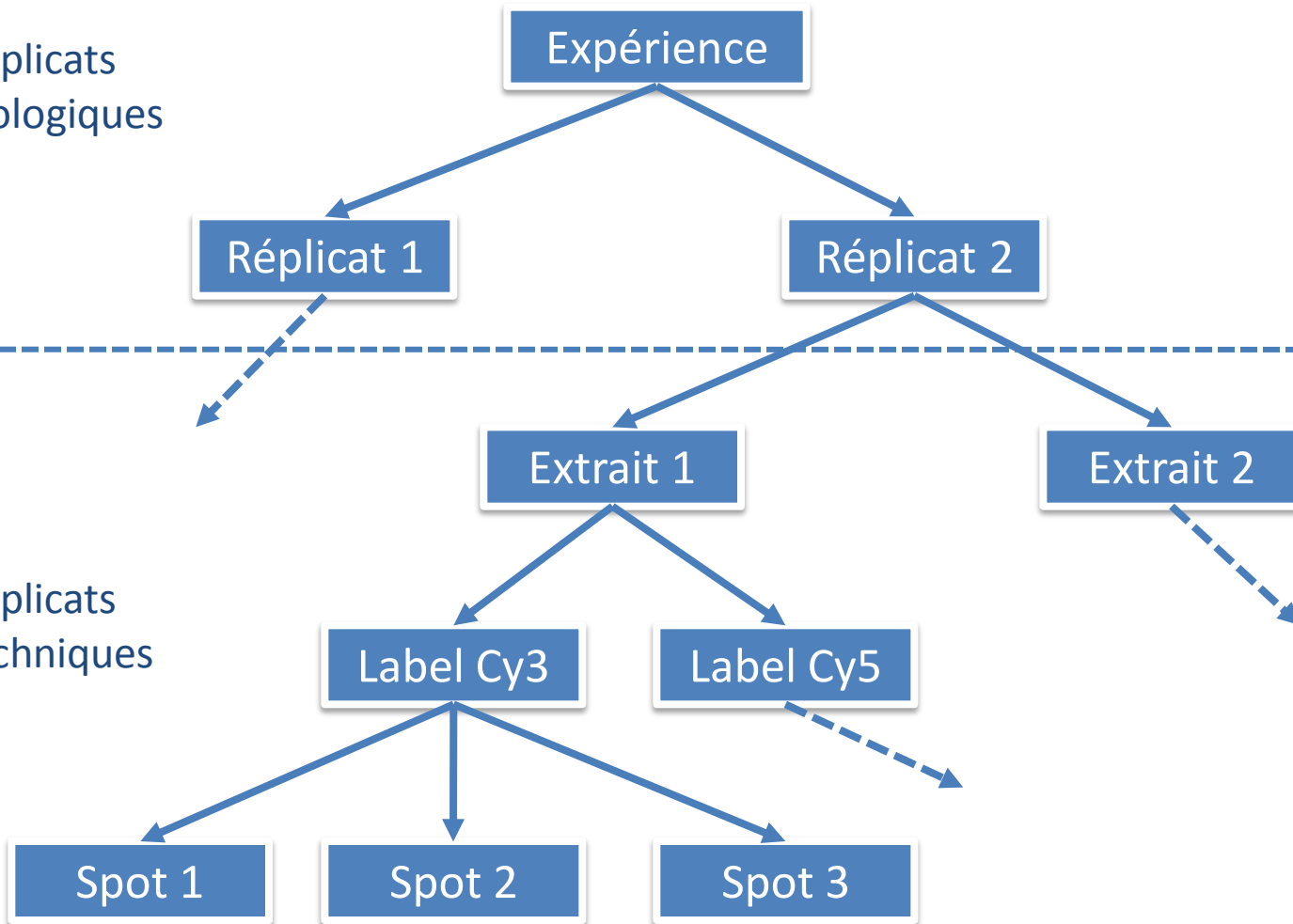
Label Cy3

Label Cy5

Spot 1

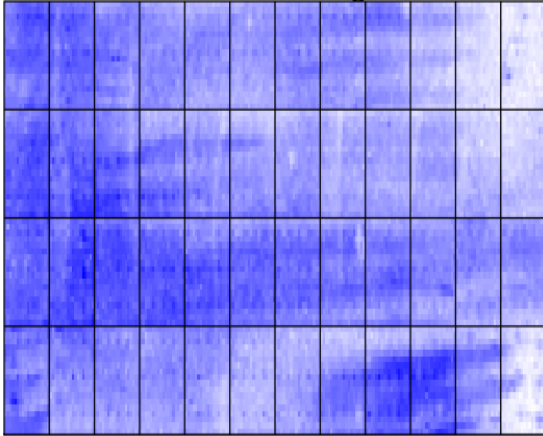
Spot 2

Spot 3



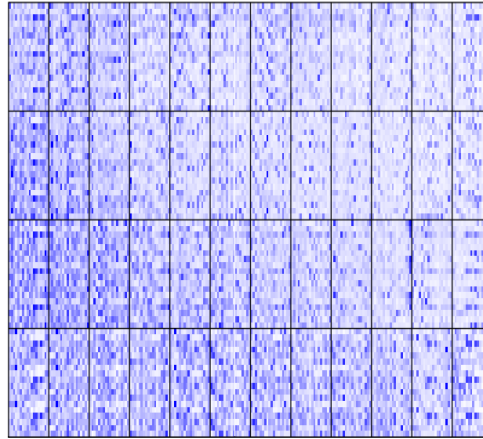
Exemples d'hybridation

300768 Red bg



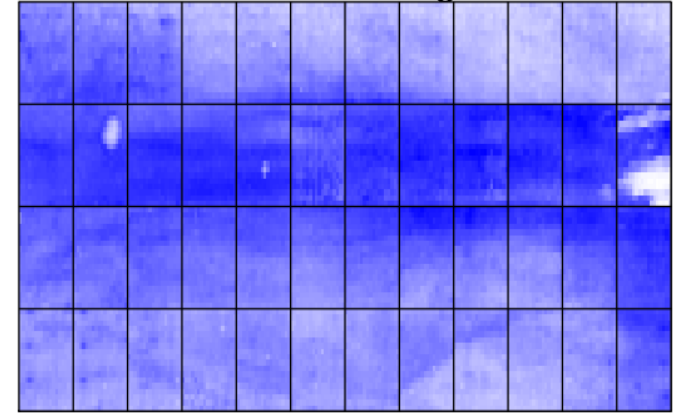
z-range 6.5 to 10.2 (saturation 6.5, 10.2)

300768 Red



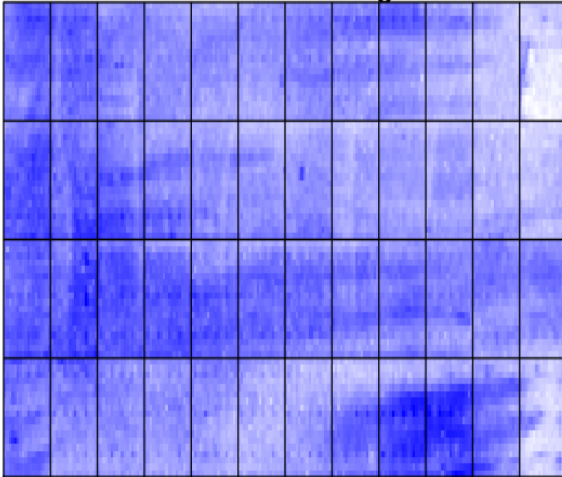
z-range 6.7 to 15.9 (saturation 6.7, 15.9)

227839 Red bg



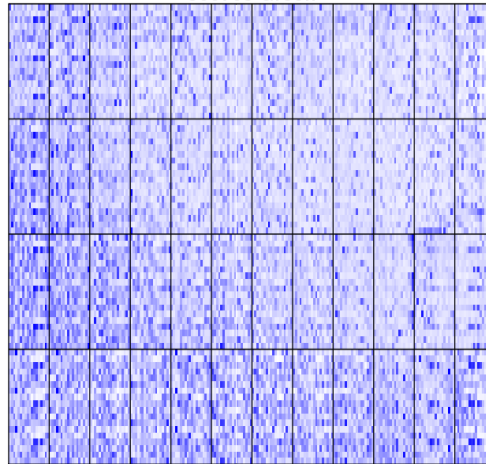
z-range 6.6 to 10.1 (saturation 6.6, 10.1)

300768 Green bg



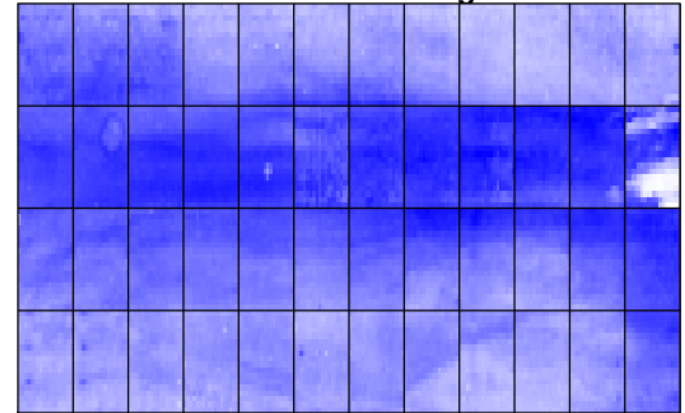
z-range 7.1 to 10.4 (saturation 7.1, 10.4)

300768 Green



z-range 7.2 to 15.9 (saturation 7.2, 15.9)

227839 Green bg

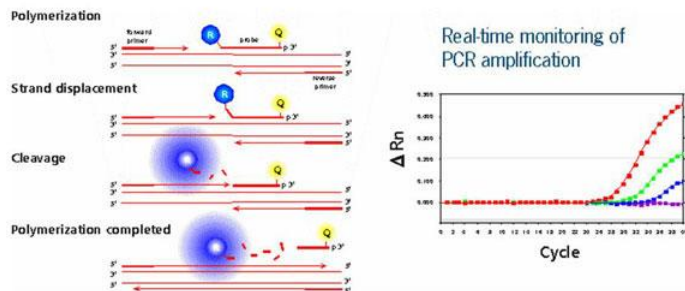


z-range 6.4 to 10.1 (saturation 6.4, 10.1)

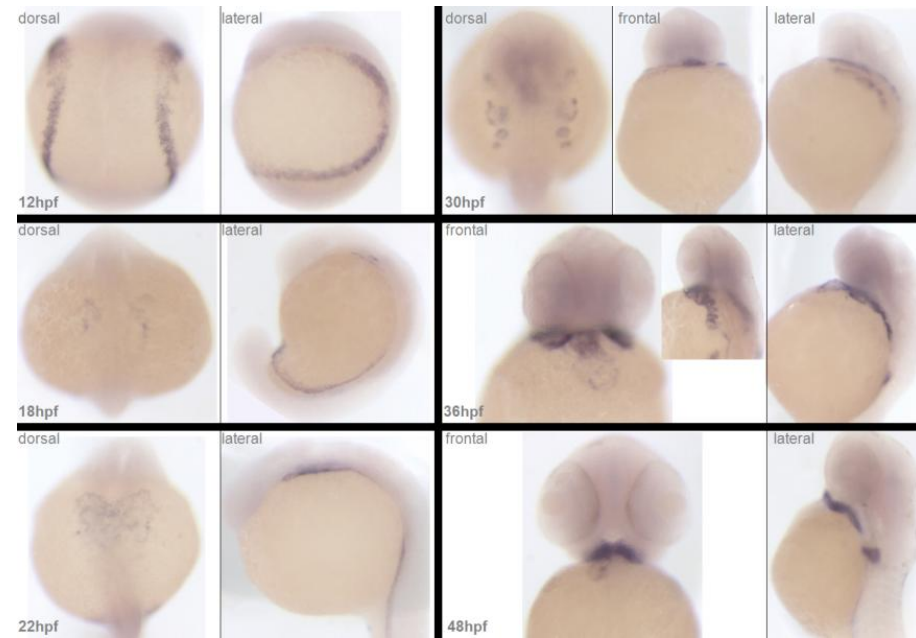
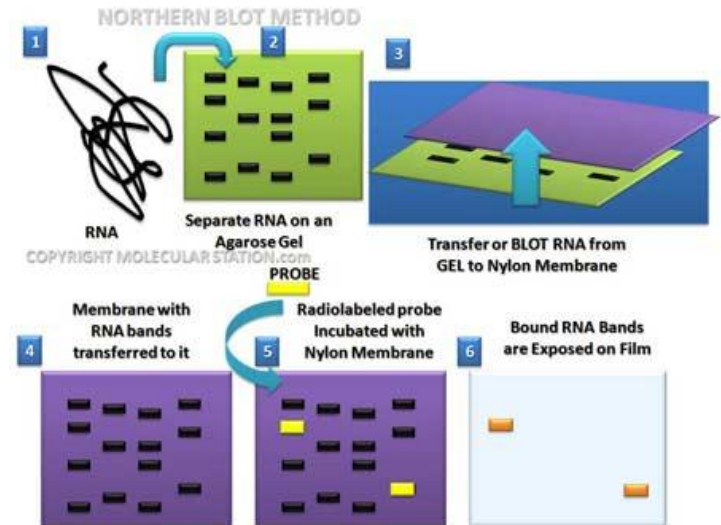
Validation

- Validation
 - directe : Northern blot, qPCR, TaqMan
 - indirecte : hybridation *in situ*, Western blot

TaqMan system



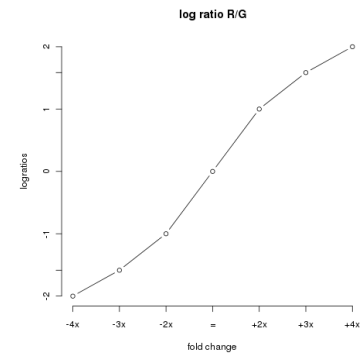
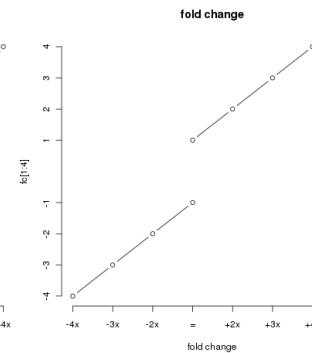
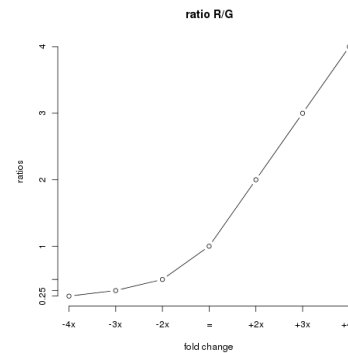
<http://documents.plant.wur.nl/>



Transformation des données

- Données initiales
 - valeurs des intensités pour les différentes conditions/canaux
- Variations de l'expression

- ratios $T = I_{\text{rouge}}/I_{\text{verte}}$
 - ⚠️ effet multiplicatif :
 - 1: pas de changement
 - 2: 2x plus exprimé
 - 0.5: 2x moins exprimé



- fold change :
 - $-1/T$ lorsque $T < 1$ (ex: 0.5 donne -2)
 - difficultés pour les analyses mathématiques dues à la discontinuité entre -1 et 1
- transformation logarithmique
 - mesure continue
 - $\log_2(0.5) = -1$; $\log_2(1) = 0$; $\log_2(2) = 1$

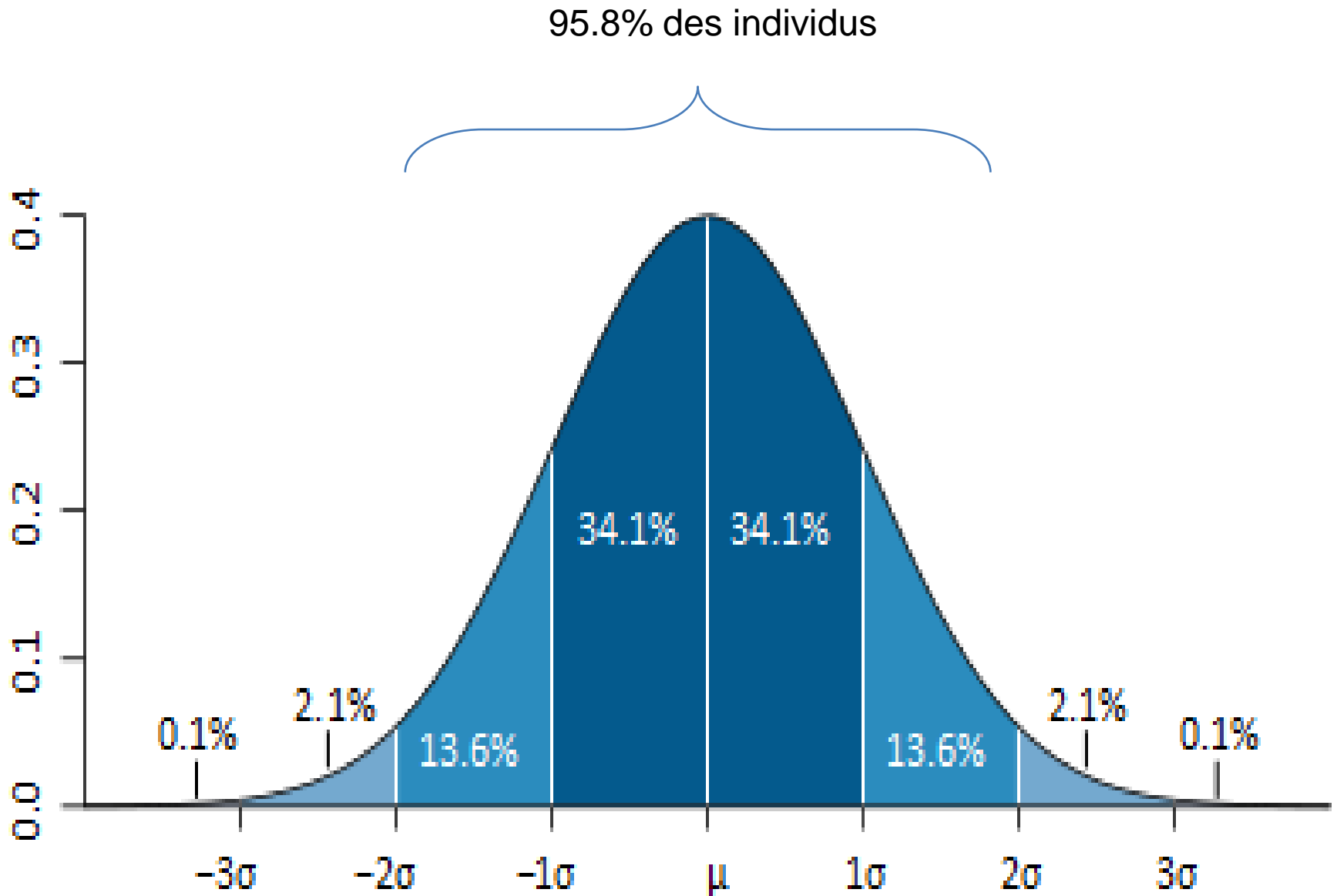
Filtrage

- Motivation
 - valeurs de (trop) faible intensité
 - non exprimé ?
 - valeur manquante (problème sur la puce) ?
 - outlier (valeurs aberrantes)

Filtrage

- Valeurs de faible intensité
 - les valeurs dépassant légèrement le bruit de fond ont plus de chance d'être imprécises ou de mauvaise qualité
- Filtrage : on élimine les valeurs inférieures à
 - $I_{\text{médiane}} + 2 \times \sigma(\text{bruit de fond})$
 - $I_{\text{moyenne}} + 2 \times \sigma(\text{bruit de fond})$

Distribution normale

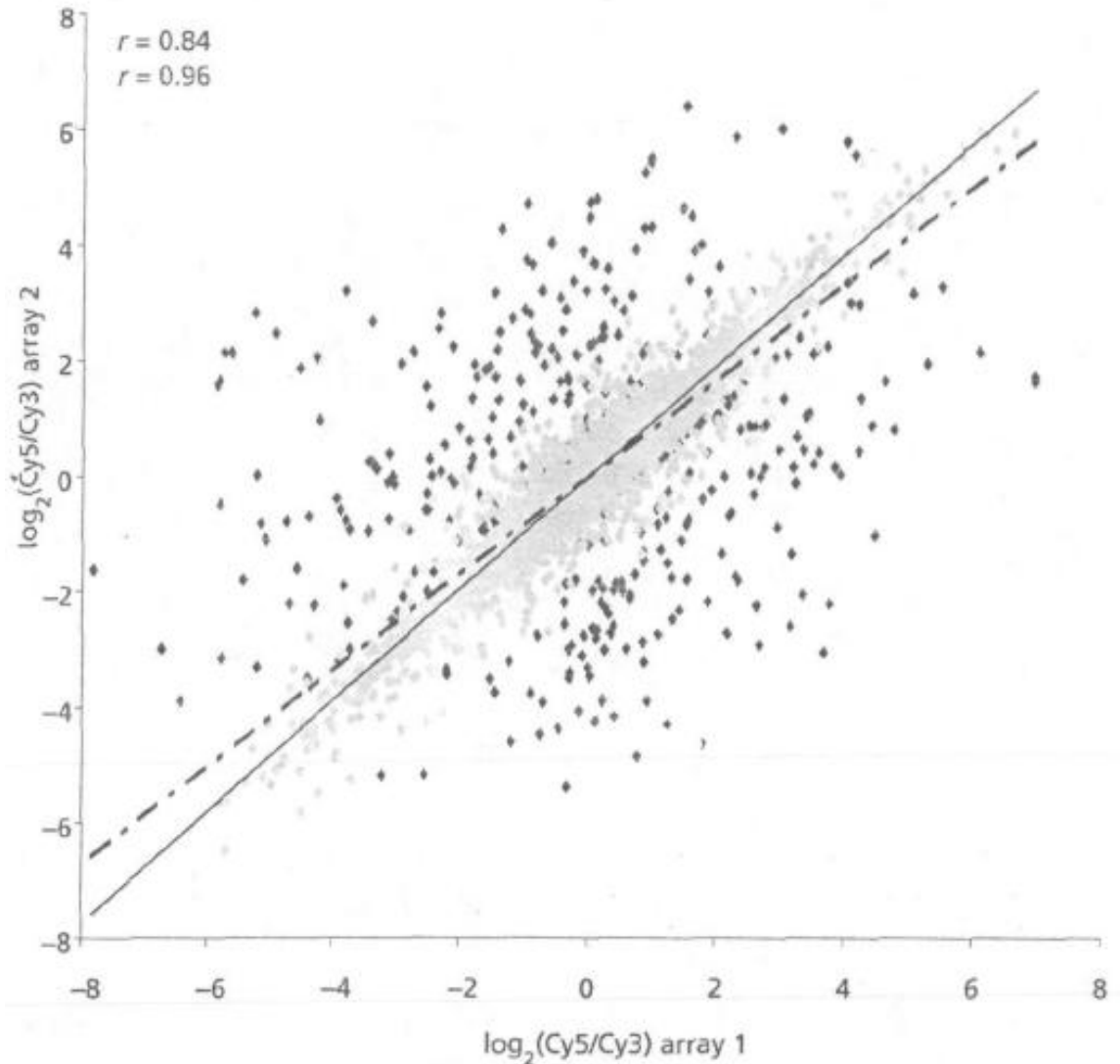


Filtrage des outliers (dye swap)

- Variabilité des réplicats
- Exemple: 2 échantillons A et B
- 1^{ère} expérience A rouge (Cy5) et B vert (Cy3)
- pour le i-ème gène on a
$$T_{1i} = \frac{R_{1i}}{G_{1i}} = \frac{A_{1i}}{B_{1i}}$$
- 2^{ème} expérience (dye swap) A vert et B rouge
- pour le i-ème gène on a
$$T_{2i} = \frac{R_{2i}}{G_{2i}} = \frac{B_{2i}}{A_{2i}}$$
- on attend $(T_{1i} * T_{2i}) = \left(\frac{A_{1i}}{B_{1i}} * \frac{B_{2i}}{A_{2i}} \right) = 1$ équivalent à $\log_2 (T_{1i} * T_{2i}) = 0$

Filtrage des outliers (2 réplicats)

- Moyenne et écart-type
 - Inspection manuelle afin d'identifier le spot aberrant
 - suppression des spots



Normalisation

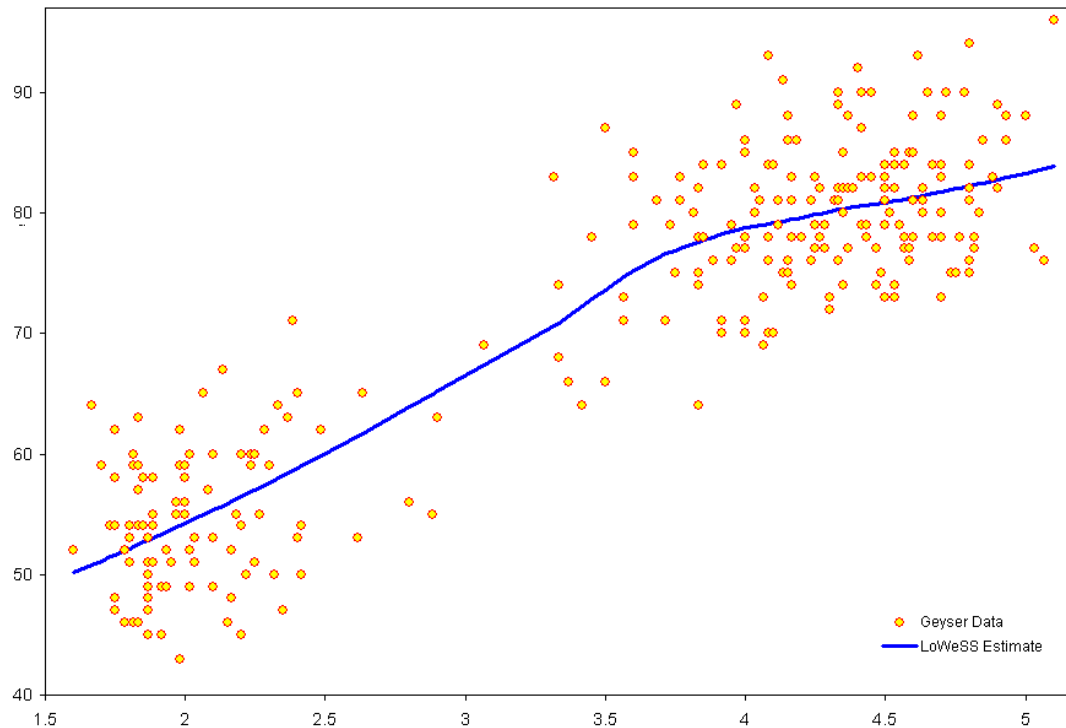
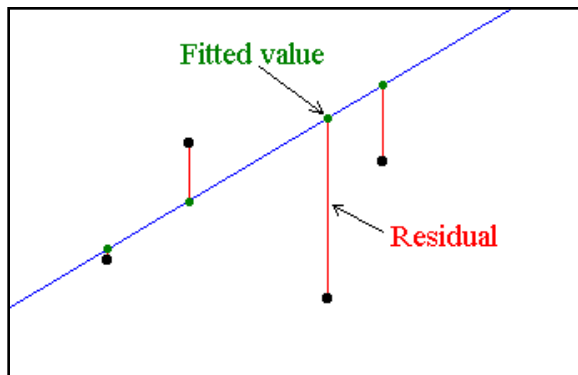
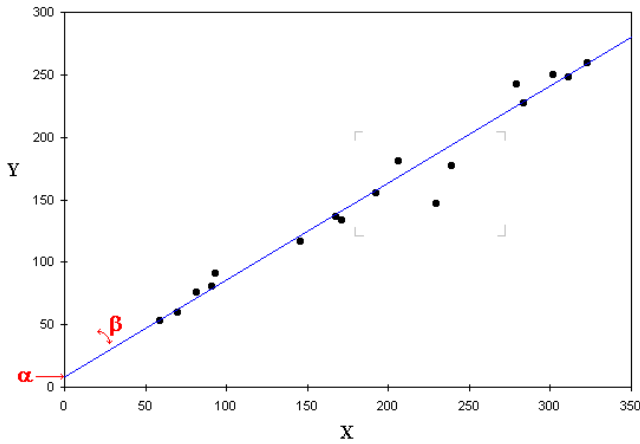
- Motivations
 - quantité d'ARN différentes dans les échantillons
 - efficacité de la détection de fluorescence
 - biais systématiques, artefacts
 - ex: pour un même échantillon marqué vert et rouge, le $\log_2(\text{ratio})$ est rarement 0.
- Normalisation: transformation des données pour corriger ces effets.

Normalisation

- Approches:
 - ensemble de contrôle
 - soit gènes de ménage, soit exogène
 - (sous-)ensemble des intensités sur la puce
 - suppose que la plupart des gènes ont le même niveau d'expression
- Nombreuses méthodes:
 - intensité totale
 - centrage sur la moyenne des log
 - régression linéaire
 - lowess

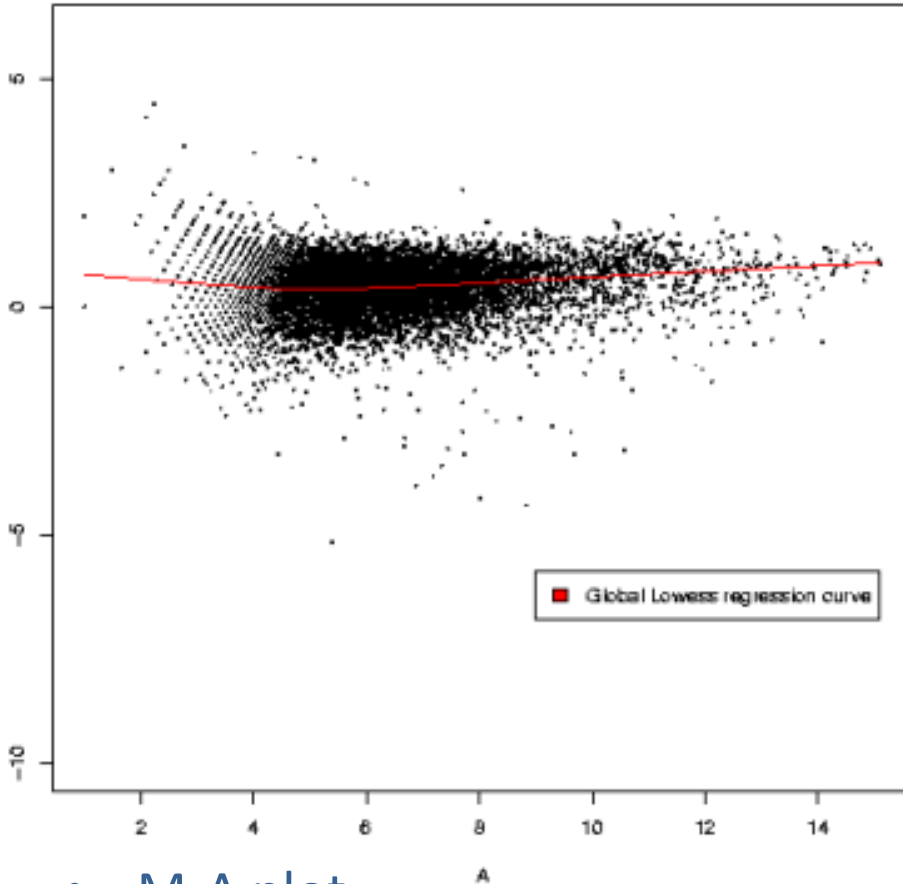
Lowess

- Régression linéaire
- ◆ **Locally weighted scatter plot smoothing**

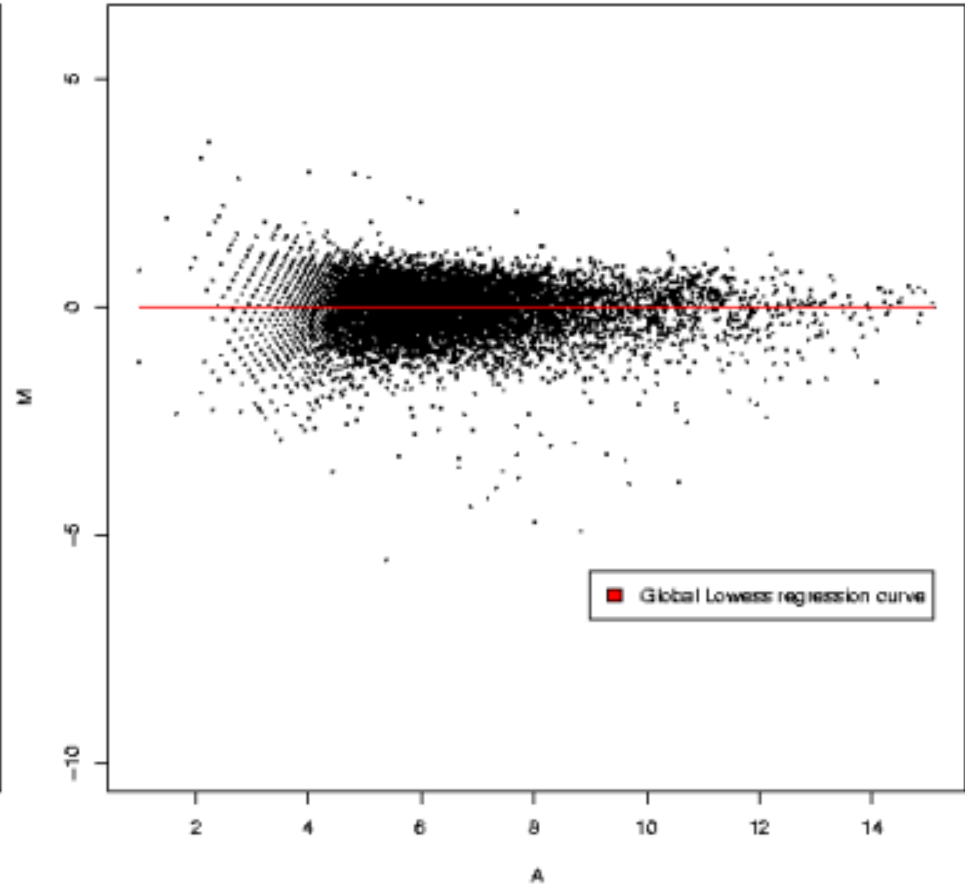


Normalisation lowess

MA-plot before normalisation (excluding filtered spots)



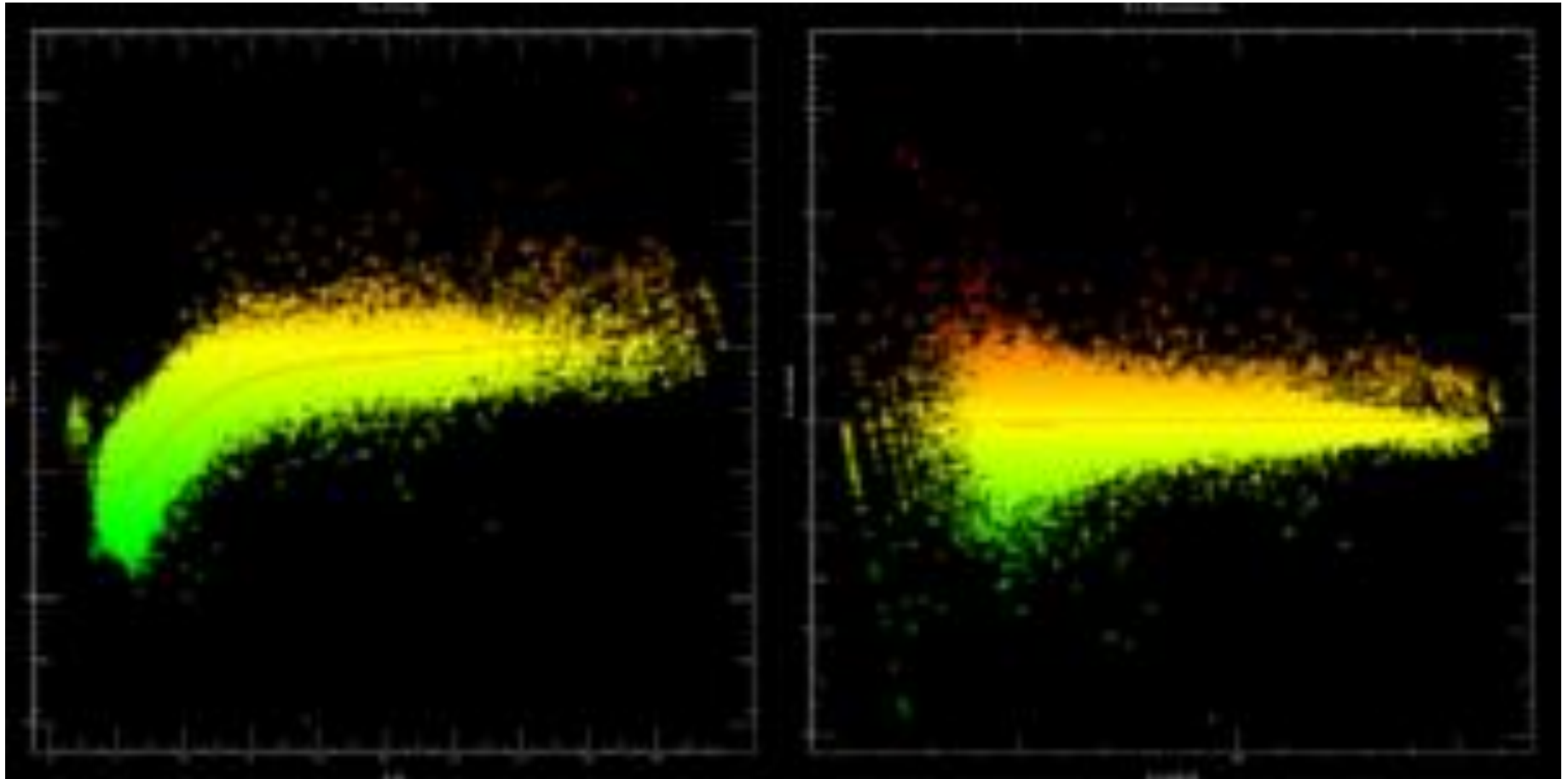
MA-plot after normalisation with lowess global curve (excluding filtered spots)



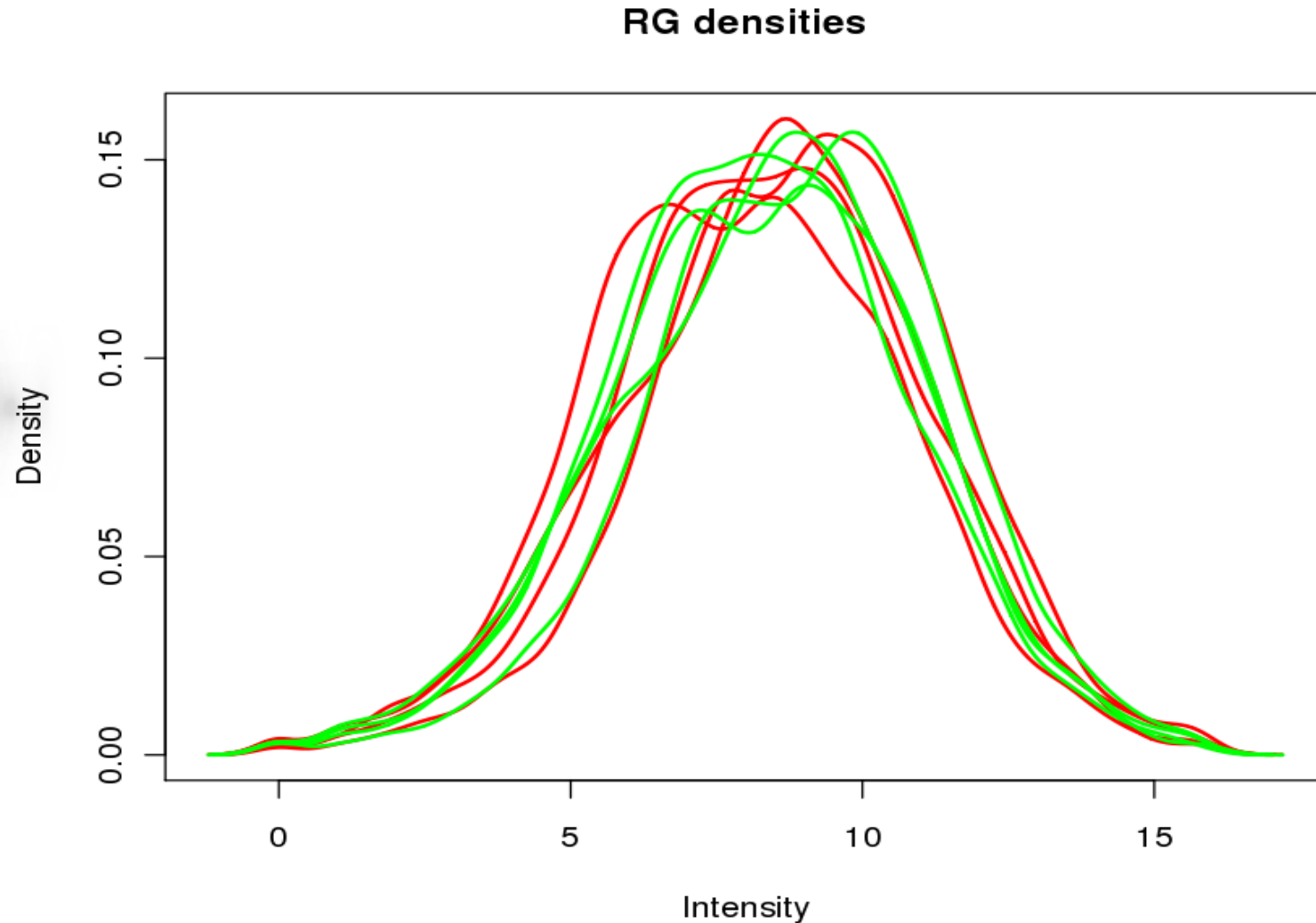
- M-A plot

- M: ordonnées, ratios des intensités. $\log_2 R - \log_2 G = \log_2 (R/G)$
- A: abscisses, moyenne des intensités du spot. $\frac{1}{2} (\log_2 R + \log_2 G)$

Normalisation

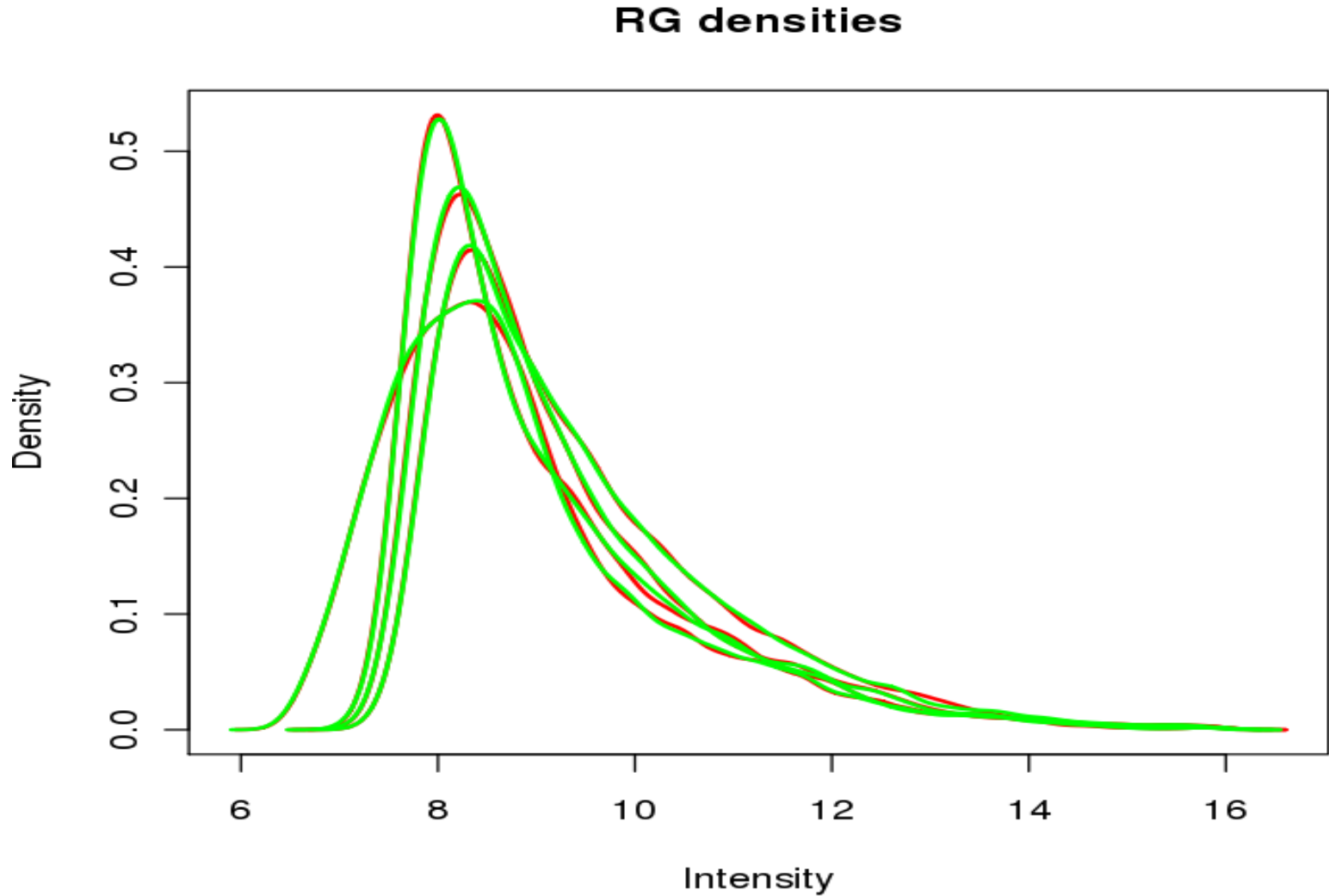


Avant normalisation



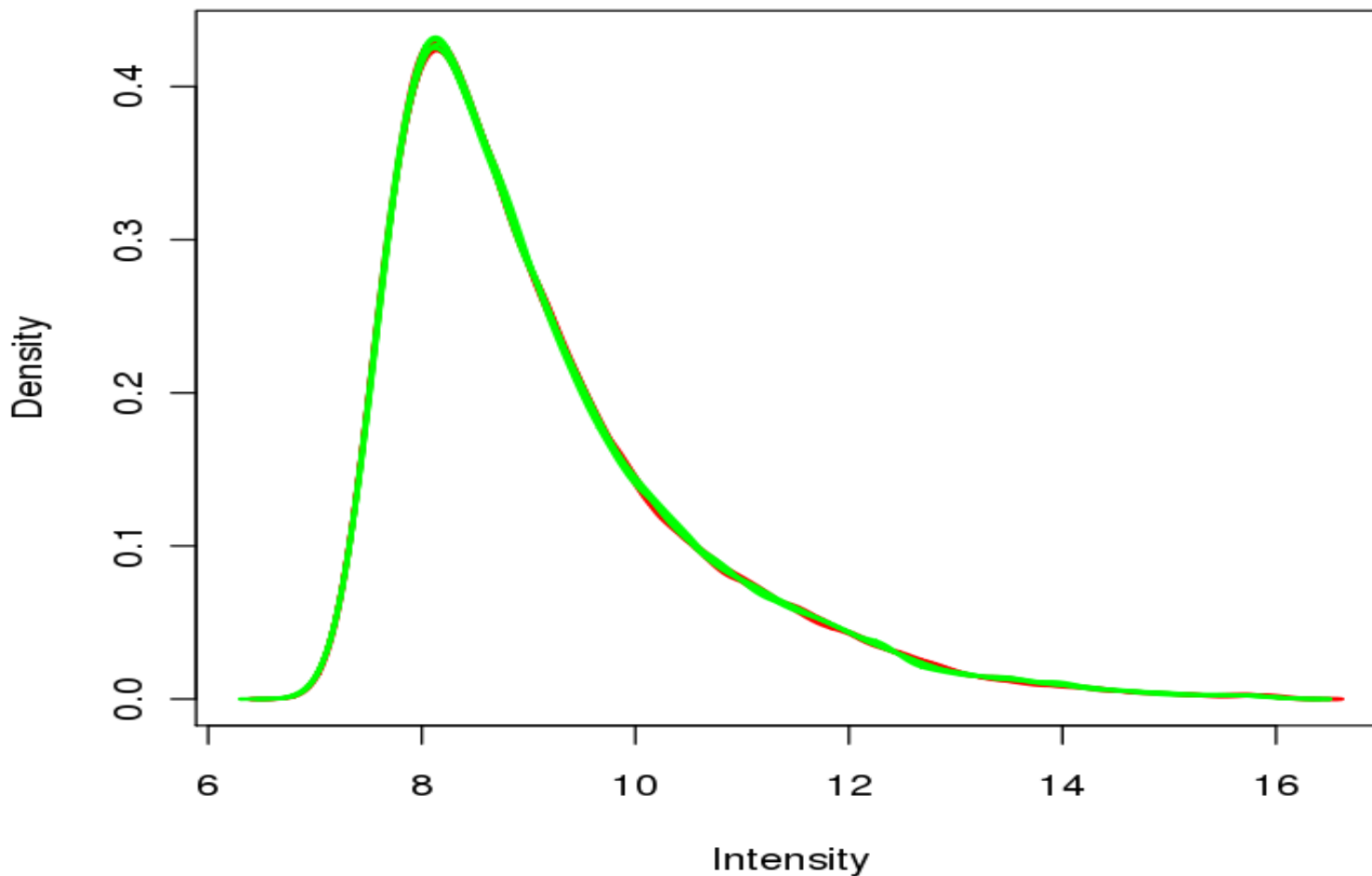
Distribution des intensités rouges et vertes sur 4 hybridations

Après normalisation intra-puces

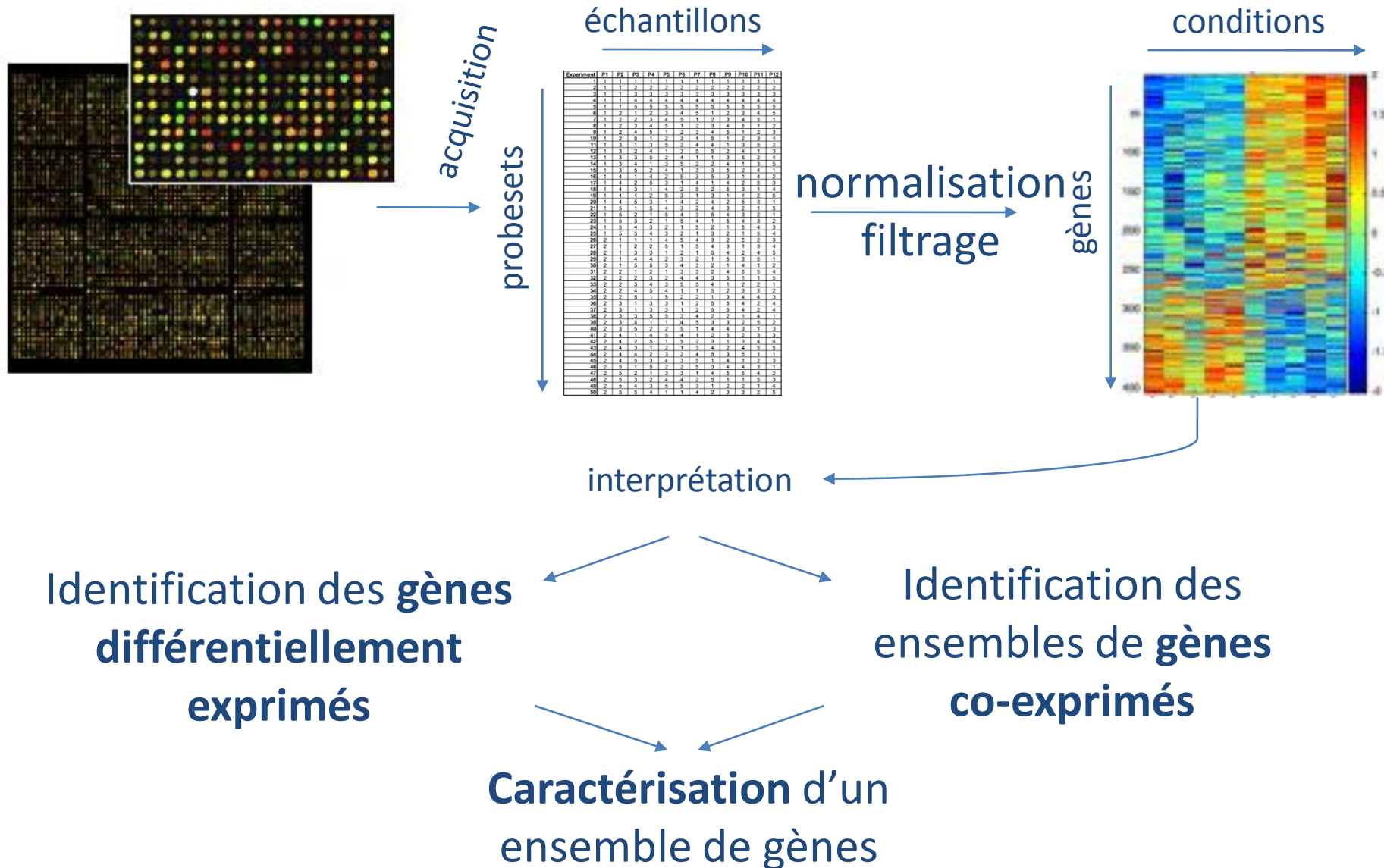


Après normalisation inter-puces

RG densities

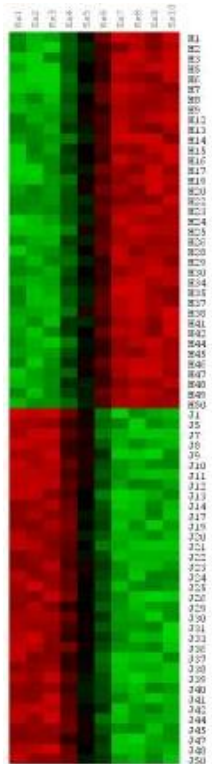


Analyse et interprétation des données



Gènes différentiellement exprimés

- Motivation
 - Gènes activés (induits) ou inactivés (réprimés) dans certaines conditions expérimentales/environnementales
- Identification des gènes différentiellement exprimés
 - Fold change
 - Modèles statistiques

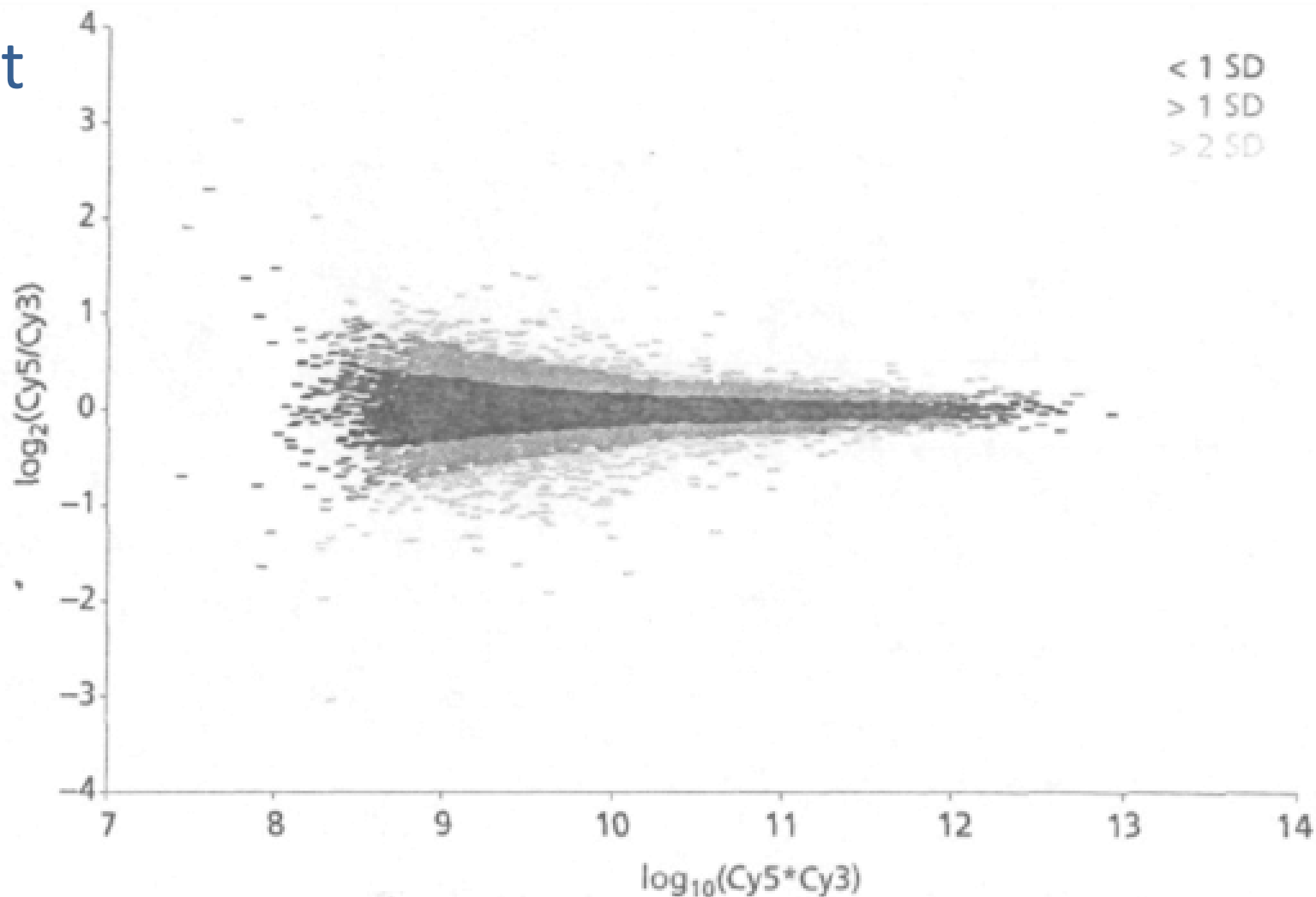


Gènes différentiellement exprimés

- Fold-change
 - seuil au-delà duquel un gène est considéré comme différentiellement exprimé
 - Ex :
 - 2x plus ou 2x moins exprimé
 - s'écarte de plus de 2x l'écart type
- ⚠ Pas un test statistique, pas de niveau de confiance
- ⚠ Ne tient pas compte de la variance au sein des réplicats

Fold change, seuil variable

- fenêtre dans laquelle on considère l'écart-type
- R-I plot



Modèles statistiques

- t -test
 - 2 conditions
- Analyse de variance (ANOVA)
 - >2 conditions
- Bayésiens, modèles de mélange (mixture models), ...

t-test

- But : déterminer si un gène est différentiellement exprimé entre 2 conditions
- Motivation :
 - Le niveau d'expression du gène est mesuré dans les 2 conditions en faisant n réplicats
 - ex : $R_1, R_2 \dots$ et G_1, G_2, \dots
 - Si le gène n'est pas différentiellement exprimé, la moyenne des ratios d'expression du gène vaut 1
 - $\overline{R} = \overline{G} \ ?$
 - two sample *t*-test permet de déterminer si les valeurs observées proviennent de distributions ayant la même moyenne

t-test

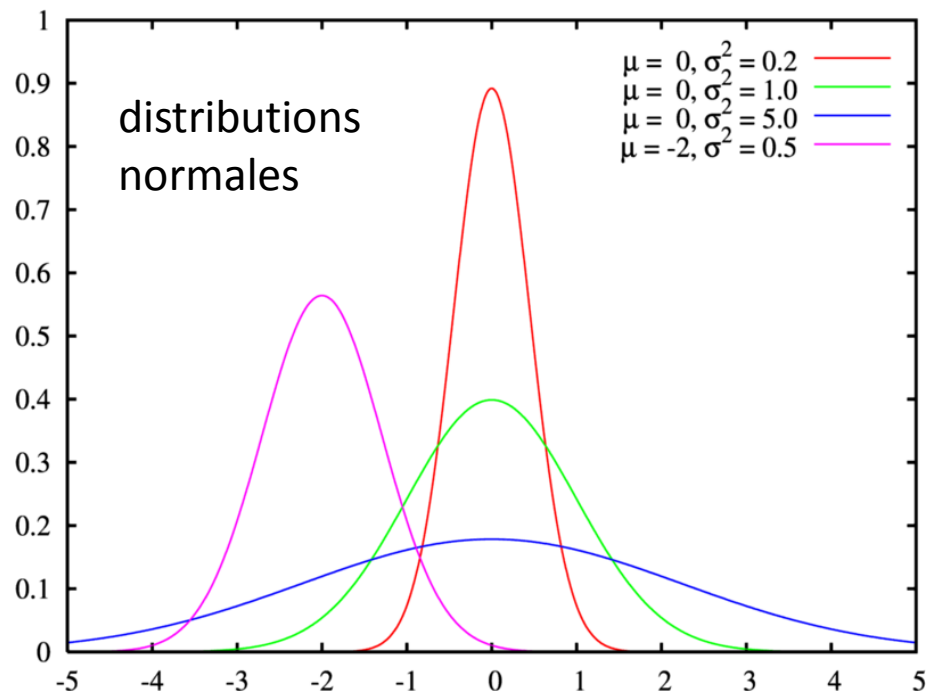
- Échantillons : R_1, R_2, R_3 et G_1, G_2, G_3
- On suppose que les mesures proviennent de distributions normales $N_R(\mu_R, \sigma_R^2)$ et $N_G(\mu_G, \sigma_G^2)$
- Erreur standard à la

moyenne :

$$ESM = \sigma / \sqrt{n}$$

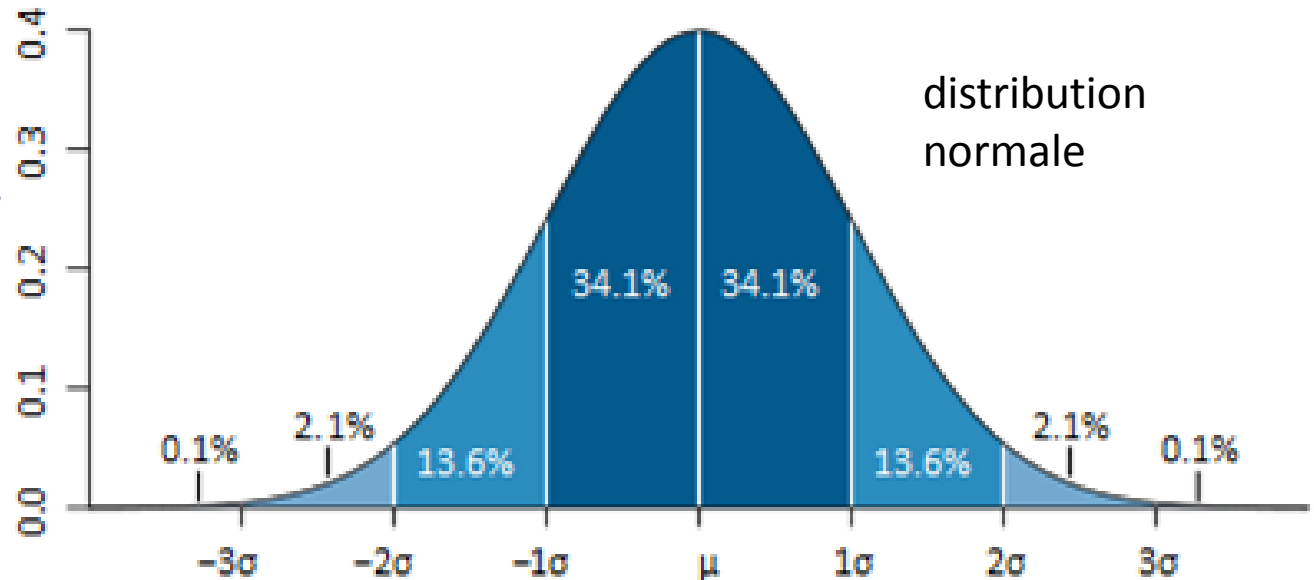
- Erreur standard de la différence des moyennes :

$$ESDM = \sqrt{ESM_R^2 + ESM_G^2}$$



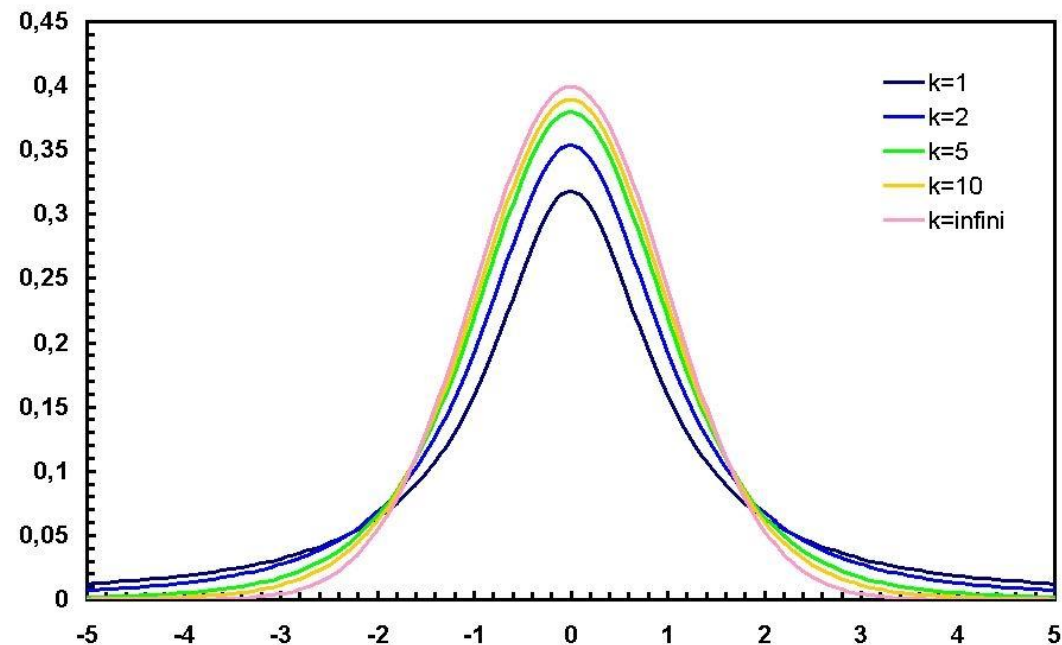
t-test

- Pour des mesures qui suivent une loi normale, on a ~95% de chances de rester à moins de 2σ de la moyenne réelle
- Formule :
$$t = \frac{\overline{R} - \overline{G}}{ESDM}$$
- d'où t doit être compris entre -2 et 2



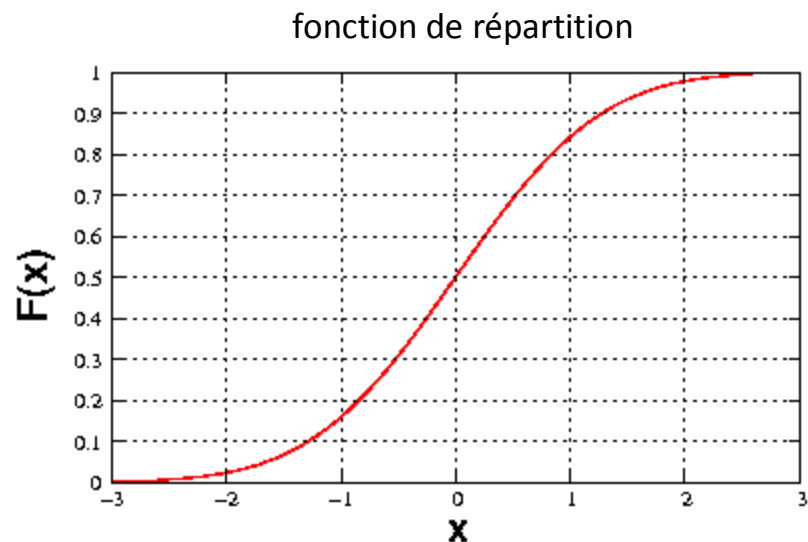
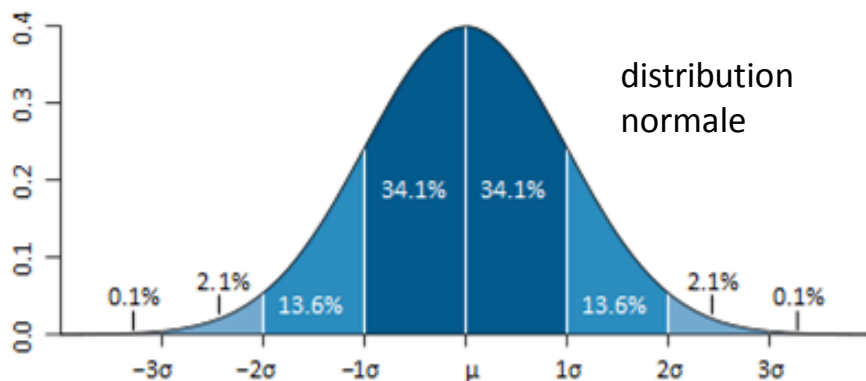
t-test

- H_0 : les valeurs observées proviennent de distributions ayant la même moyenne (le gène n'est pas différentiellement exprimé)
- Calcul de $t = \frac{\overline{R} - \overline{G}}{ESDM}$
- Probabilité donnée par la loi de probabilité de t



p -valeur

- p -valeur : probabilité d'obtenir un résultat au moins aussi extrême
- probabilité : $p(x = X)$
- p -valeur : $p(x > X)$



- exemple avec un dé à 6 faces
 - $p(3) = 1/6$
 - $p(\text{au moins } 3) = p(3) + p(4) + p(5) + p(6) = 4/6$
 - $p(\text{moins de } 3) = 2/6$

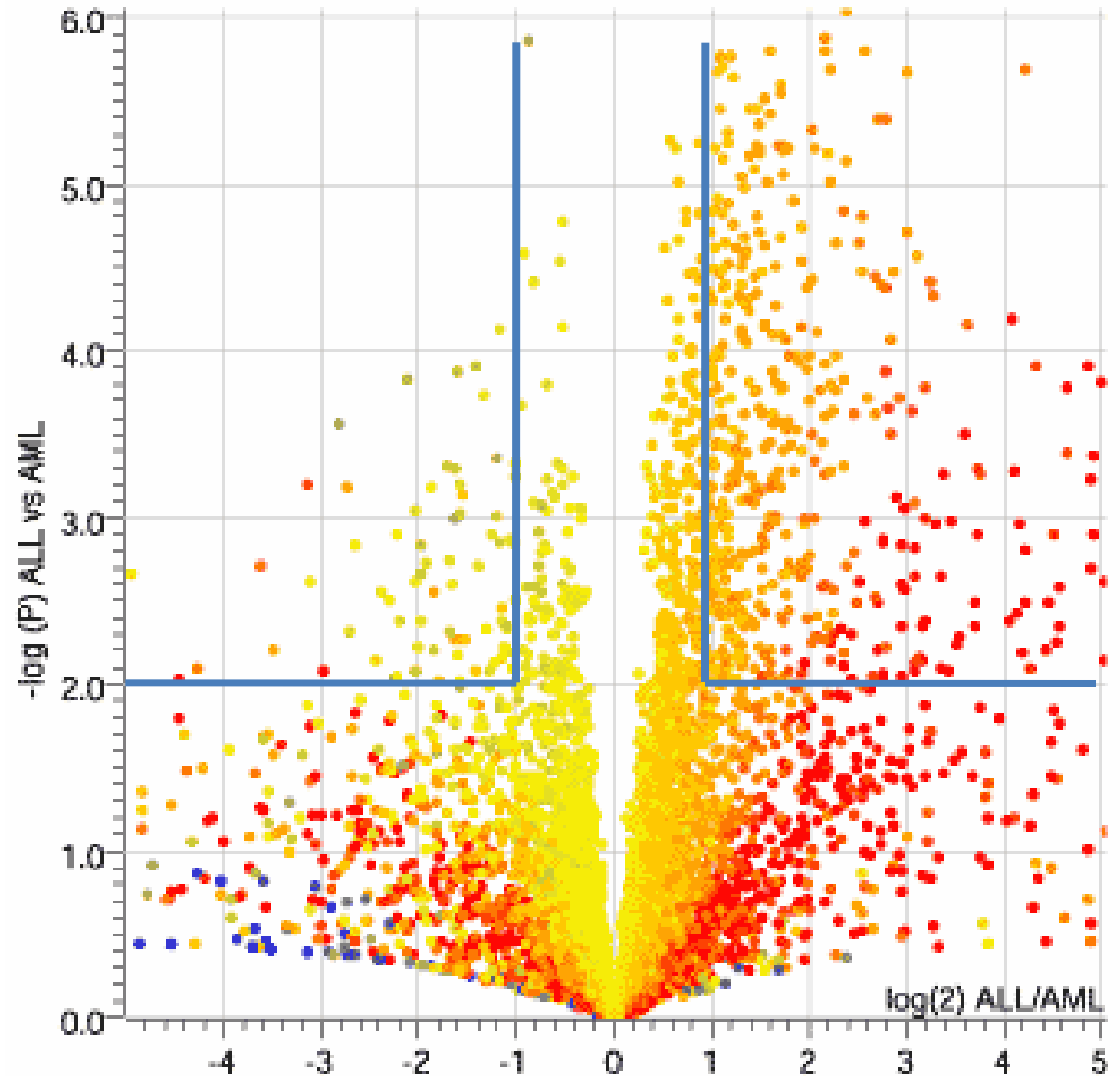
t-test

- Application
 - R : contrôle
 - G : traitement

	R_1	R_2	G_1	G_2	p -value
267627_at	57	6	45.5	38.6	0.7504
267628_at	441.8	431.5	347.2	355.2	0.0072
267629_at	226.5	205.6	148.2	132.9	0.0343
267630_at	1142.6	1080.7	1019.8	1018.6	0.2055
267631_at	77.7	58	84.4	57.4	0.8734

Volcano plot

- fold change vs. p -valeur (t -test ou autre)



ANOVA

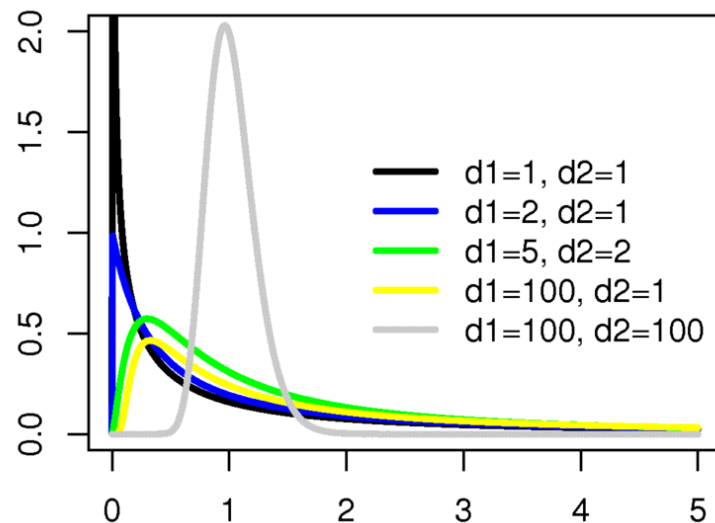
- Hypothèse testée : les moyennes des différentes conditions sont égales
- Variabilité inter-classes
i.e. variabilité entre les conditions
- Variabilité intra-classe
i.e. variabilité observée à l'intérieur de chaque condition

$$F = \frac{\frac{S_{\text{int } er}}{c - 1}}{\frac{S_{\text{int } ra}}{N - c}}$$

- Remarque: pour 2 conditions, cela équivaut au *t*-test

$$S_{\text{int } er} = \sum_{i=1}^c r_i (\bar{T}_i - \bar{T})^2$$

$$S_{\text{int } ra} = \sum_{i=1}^c \sum_{j=1}^{r_i} (T_{ij} - \bar{T}_i)^2$$



Tests multiples

- H_0 : le gène g a un niveau d'expression constant
- seuil α typique de 5% *i.e.* g est considéré comme différentiellement exprimé si $p\text{-valeur}(g) \leq 0.05$
- Idée : plus on augmente le nombre de tests, plus on a de chances de décider qu'un gène est différentiellement exprimé alors qu'il ne l'est pas
- combien de faux positifs et de faux négatifs ?

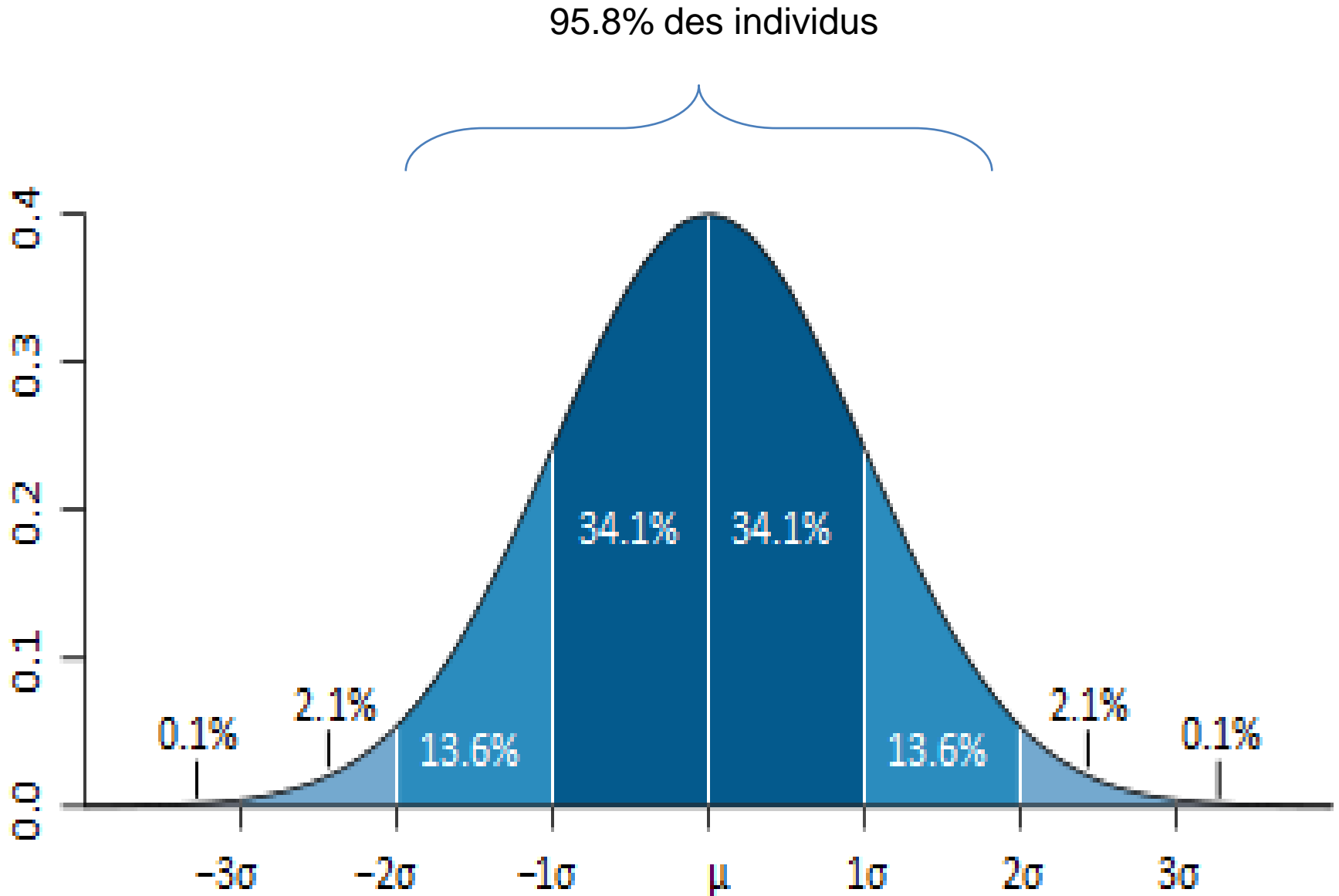
Erreurs de 1^{ère} et 2^{ème} espèce

- Erreur de 1^{ère} espèce (Type 1 error) :
 - probabilité α de rejeter H_0 alors qu'elle est vraie
 - probabilité de décider qu'un gène est diff. exprimé alors qu'il ne l'est pas
 - faux positif
- Erreur de 2^{ème} espèce (Type 2 error) :
 - probabilité β d'accepter H_0 alors qu'elle est fausse
 - probabilité de décider qu'un gène n'est pas diff. exprimé alors qu'il l'est
 - faux négatif

Situation	Décision	
	accepter H_0	rejeter H_0
H_0 vraie	$1-\alpha$	α
H_0 fausse (diff. expr.)	β	$1-\beta$

- Conséquence :
 - En testant les 20 000 gènes de la puce avec $\alpha = 5\%$
 - 200 gènes ont une p-valeur comprise entre 0.01 et 0.05
 - on s'attend à obtenir au moins 200×0.01 faux positifs soit >2 gènes qui ne sont en réalité **pas** différentiellement exprimés

Distribution normale



Correction pour tests multiples

- False Discovery Rate (FDR) *Benjamini & Hochberg '95*
- Principe : ajuster le seuil α en fonction des résultats observés (p -valeurs obtenues)
- m tests ayant des p -valeurs $P_1..P_m$ triées par ordre croissant
- Pour un seuil α trouver le plus grand k tel que

$$P_k \leq \frac{k}{m} \alpha$$

et déclarer les gènes $1..k$ différentiellement exprimés

Application de la FDR

- gène g différentiellement exprimé si

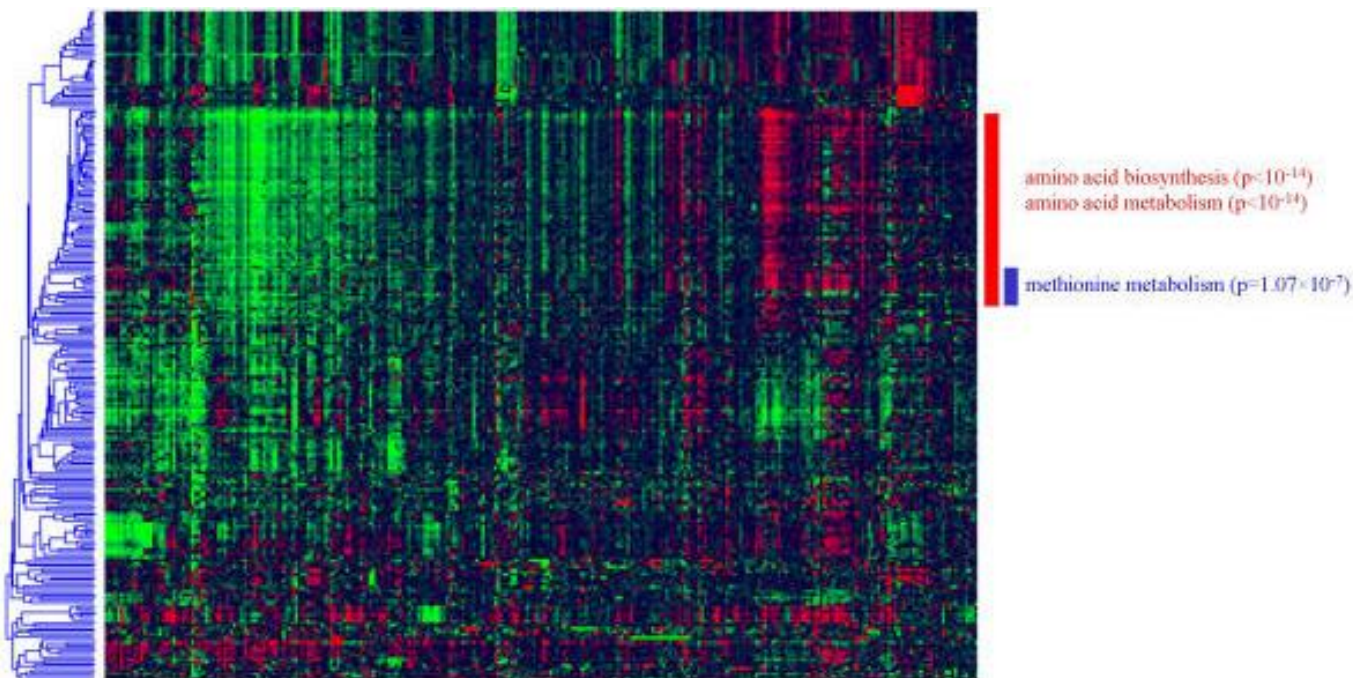
$$P_k \leq \frac{k}{m} \alpha$$

	R_1	R_2	G_1	G_2	p -value	$\alpha * k/m$
267628_at	441.8	431.5	347.2	355.2	0.0072	0.01
267629_at	226.5	205.6	148.2	132.9	0.0343	0.02
267630_at	1142.6	1080.7	1019.8	1018.6	0.2055	0.03
267627_at	57	6	45.5	38.6	0.7504	0.04
267631_at	77.7	58	84.4	57.4	0.8734	0.05

un gène est déclaré différentiellement exprimé
pour $\alpha = 0.05$

Gènes co-exprimés

- Motivation : les gènes ayant des profils d'expression similaires sont potentiellement co-régulés et participent à un même processus biologique
- But : regrouper les gènes impliqués dans un même processus biologique



Qu'est-ce que le clustering ?

- analyse de clustering
 - regroupement des objets en clusters
- un cluster : une collection d'objets
 - similaires au sein d'un même cluster
 - dissimilaires aux objets appartenant à d'autres clusters
- classification non supervisée : pas de classes prédéfinies
- Applications typiques
 - afin de mieux comprendre les données
 - comme prétraitement avant d'autres analyses

Principales approches

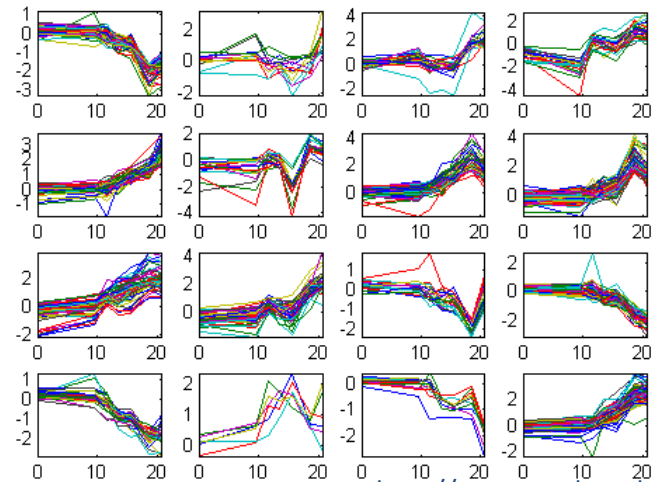
- partitionnement
 - partitionne les objets et évalue les partitions (les ensembles)
 - ex: *k*-means
- hiérarchique
 - décomposition hiérarchique d'ensembles d'objets
- densité
 - basée sur une fonction de densité ou de connectivité
- grille
 - basée sur une structure de granularité à plusieurs niveaux
- basée sur un modèle
 - construction d'un modèle pour chaque cluster
- ...

Gènes co-exprimés

- Profils d'expression
- Mesure de similarité entre 2 profils :
Coefficient de corrélation de Pearson
 - -1 : corrélation négative
 - 0 : indépendance
 - 1 : corrélation positive
- Clustering des profils
 - Ensembles de gènes ayant des profils d'expression similaires

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)\sigma_X \sigma_Y}$$

K-Means Clustering of Profiles



Structures de données

- Matrice de données :
les profils d'expression

$$\begin{array}{c}
 \text{conditions expérimentales} \\
 \left[\begin{array}{ccccc}
 x_{11} & \dots & x_{1f} & \dots & x_{1p} \\
 \dots & \dots & \dots & \dots & \dots \\
 x_{i1} & \dots & x_{if} & \dots & x_{ip} \\
 \dots & \dots & \dots & \dots & \dots \\
 x_{n1} & \dots & x_{nf} & \dots & x_{np}
 \end{array} \right]
 \end{array}
 \begin{array}{c}
 \\
 \\
 \text{gènes} \\
 \\
 \\
 \\
 \end{array}$$

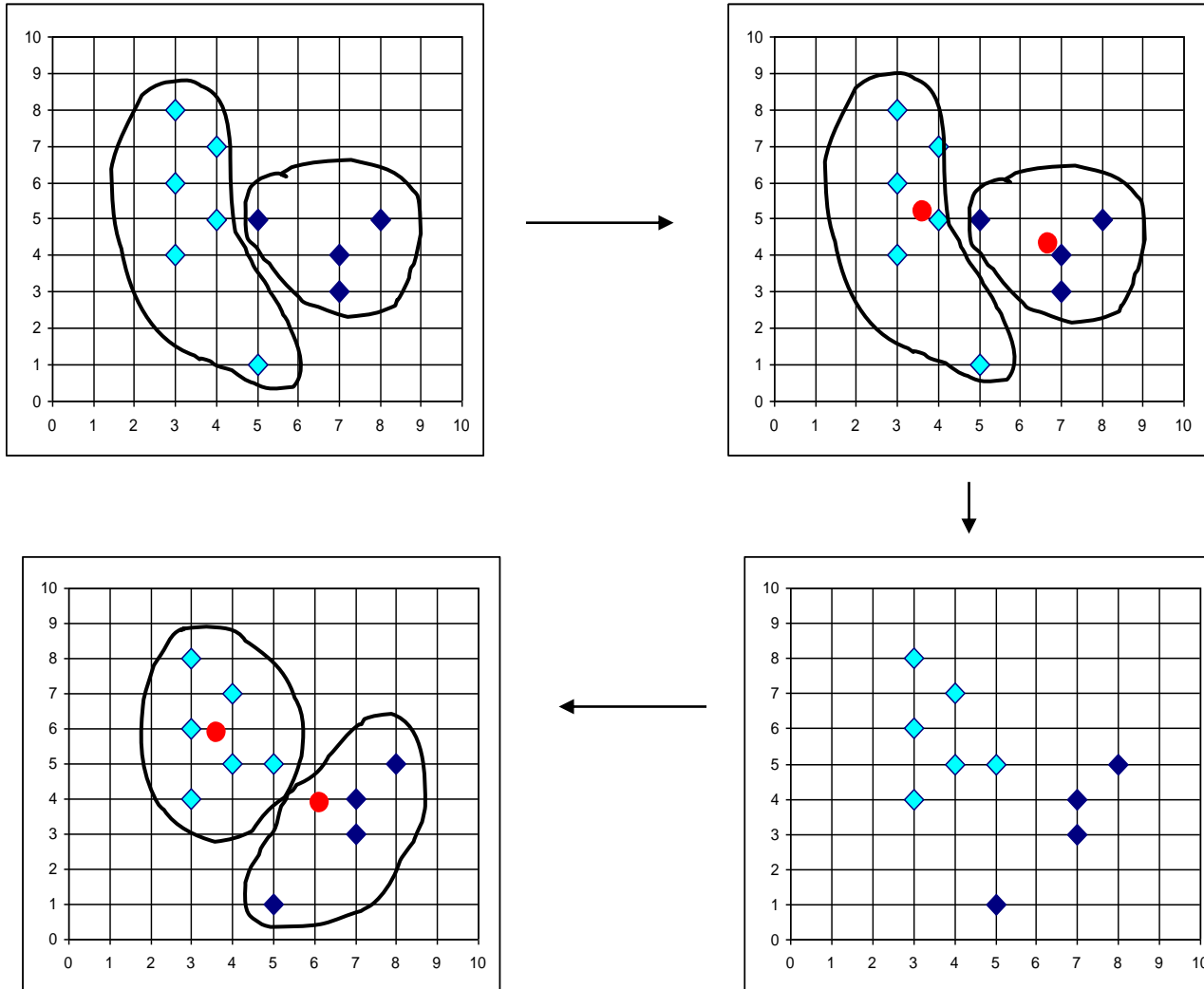
- Matrice de distance
(ou dissimilarité) :
pour chaque paire de
gènes, le coefficient
de corrélation

$$\begin{array}{c}
 \text{gènes} \\
 \left[\begin{array}{ccccc}
 0 & & & & \\
 d(2,1) & 0 & & & \\
 d(3,1) & d(3,2) & 0 & & \\
 \vdots & \vdots & \vdots & & \\
 d(n,1) & d(n,2) & \dots & \dots & 0
 \end{array} \right]
 \end{array}
 \begin{array}{c}
 \\
 \\
 \\
 \text{gènes} \\
 \\
 \\
 \\
 \end{array}$$

k-means

- 4 étapes
 1. Partitionne les objets en k ensembles non vides
 2. Calcule le centroïde de chaque partition/cluster
 3. Assigne à chaque objet le cluster dont le centroïde est le plus proche
 4. boucle en 2, jusqu'à ce les clusters soient stables.

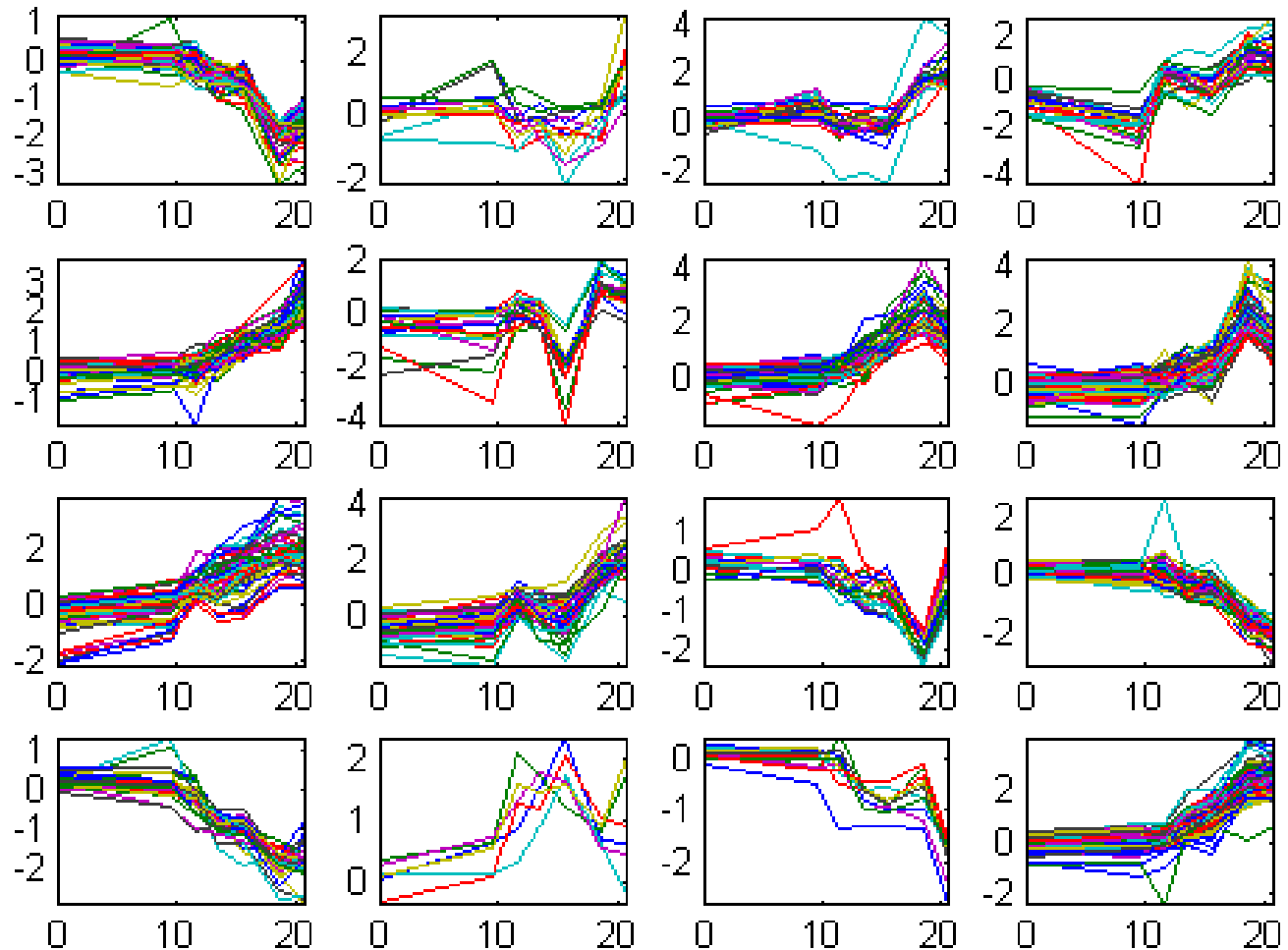
k-means, exemple



k -means, application

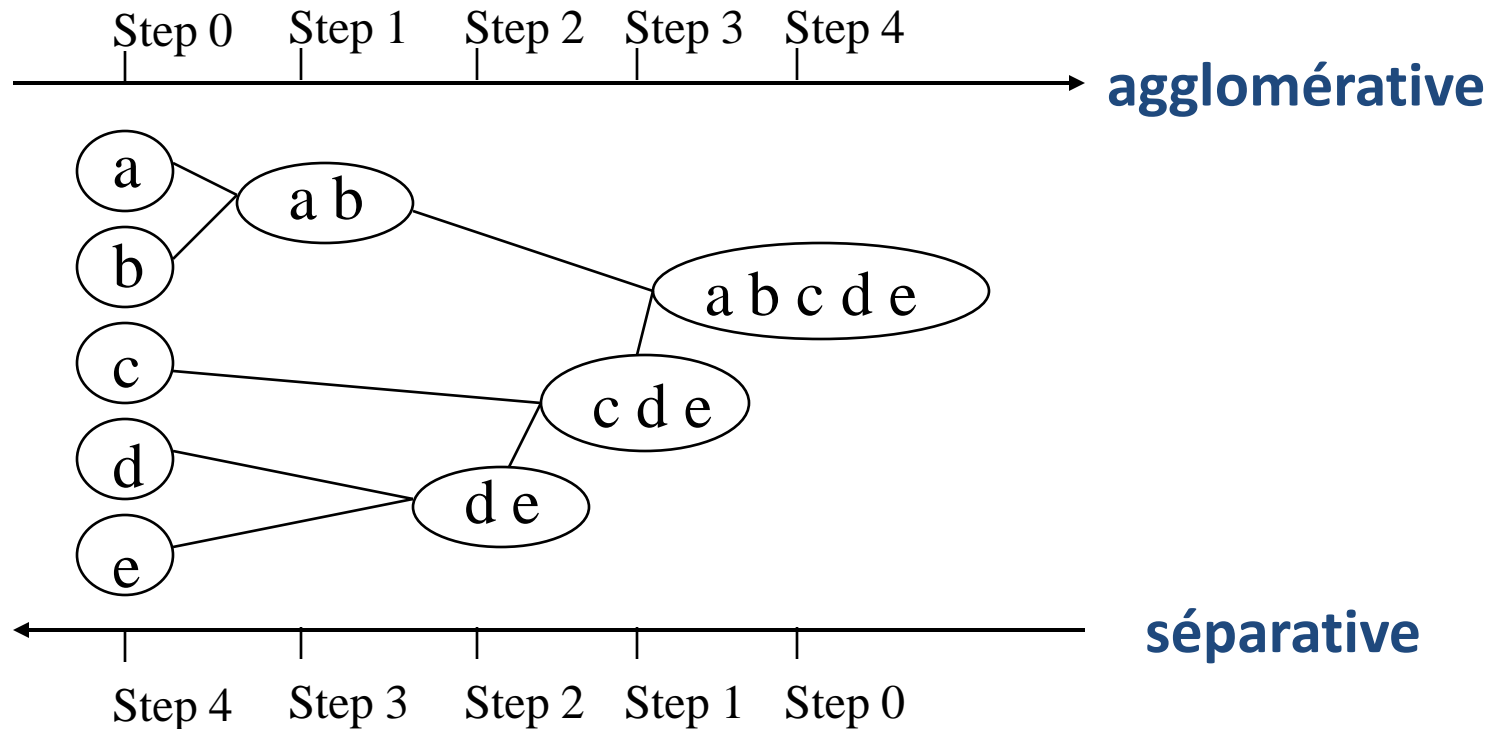
- $k = 16$ clusters

K-Means Clustering of Profiles



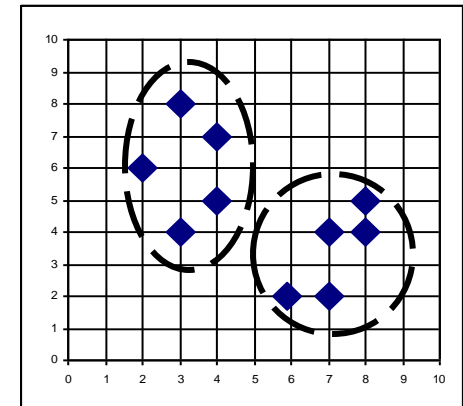
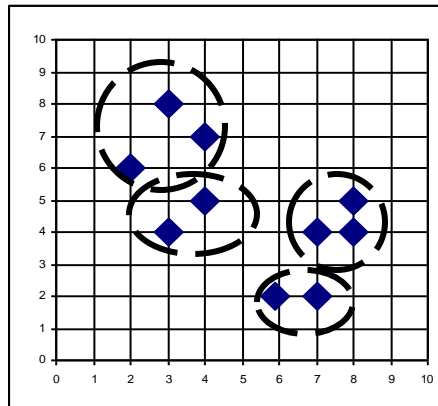
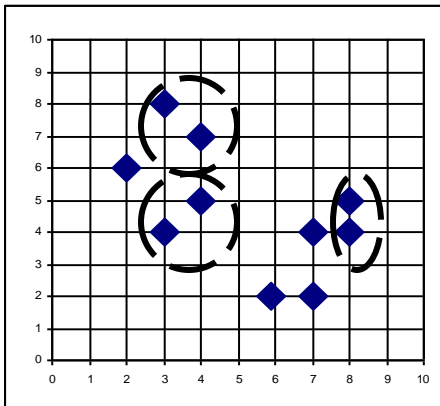
Clustering hiérarchique

- Utilisation d'une matrice de distance : ne nécessite pas de spécifier le nombre de clusters



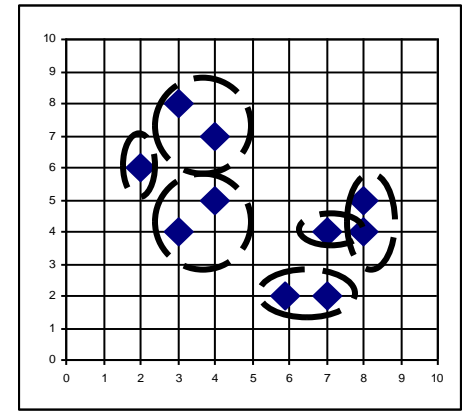
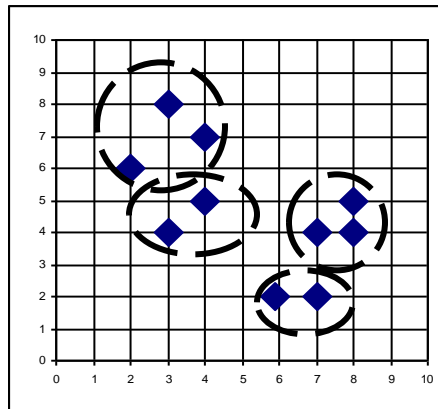
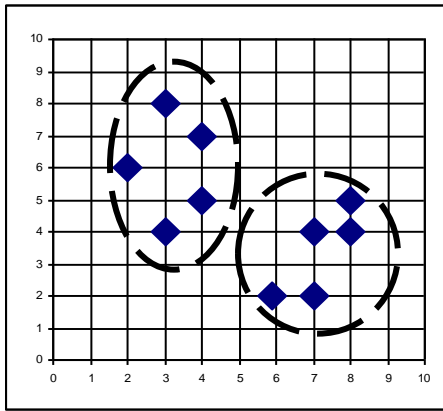
AGNES (Agglomerative Nesting)

- Utilise une matrice de dissimilarité
- Fusionne les nœuds les moins dissimilaires



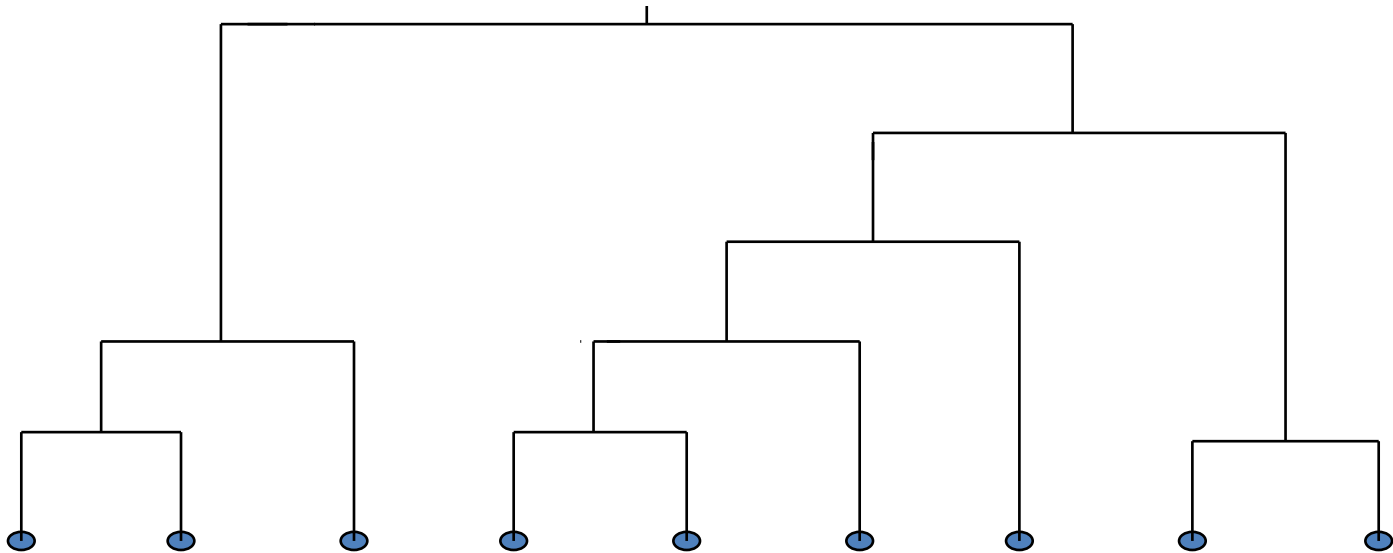
DIANA (Divisive Analysis)

- Inverse d'AGNES



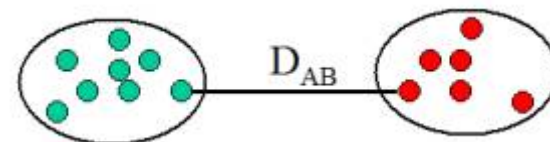
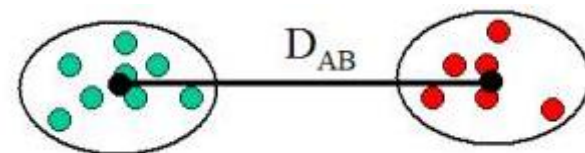
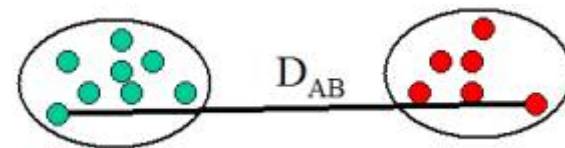
Dendrogramme : clusters fusionnés hiérarchiquement

- Décompose les données en plusieurs niveaux imbriqués de partitionnement
- Un clustering est obtenu en coupant le dendrogramme au niveau choisi

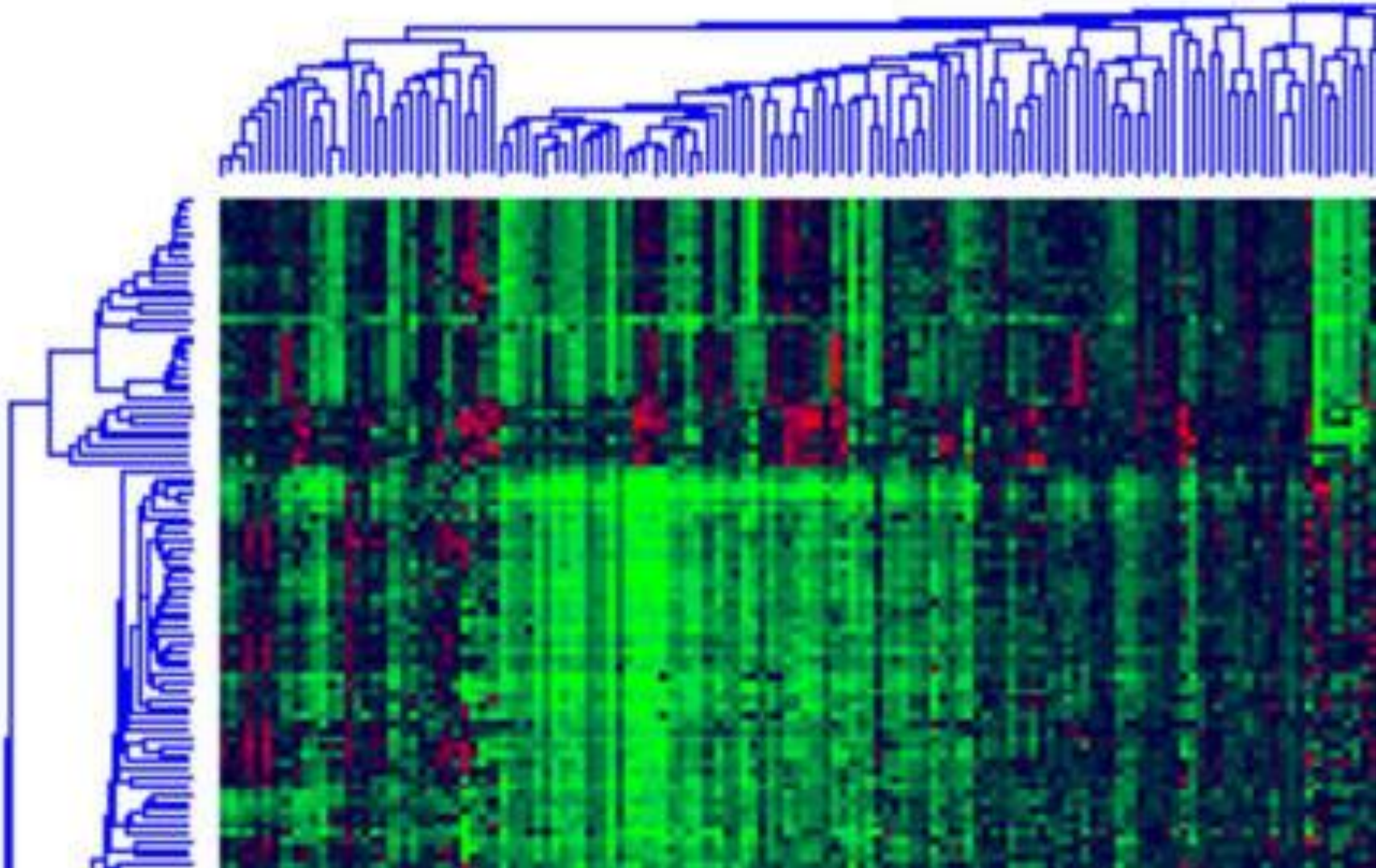


Mesures de similarité entre 2 clusters

- complete linkage
 - plus petite similarité/plus grande distance entre toutes les paires de gènes entre 2 clusters
- average linkage
 - similarité moyenne entre les paires de gènes
- single linkage
 - plus grande similarité/plus petite distance entre 2 gènes de 2 clusters



Clustering hiérarchique



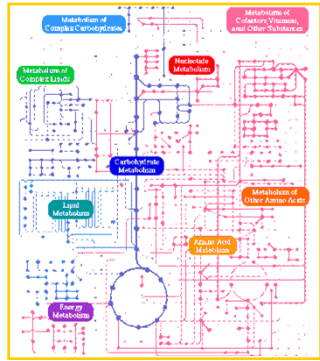
Caractérisation d'une liste de gènes

- Motivation
 - jusqu'à plusieurs milliers de gènes co-exprimés ou différentiellement exprimés
 - analyse « manuelle » impossible
- Principe
 - Rechercher les caractéristiques communes aux gènes
 - Surreprésentation statistique

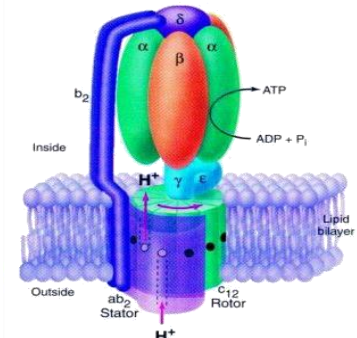
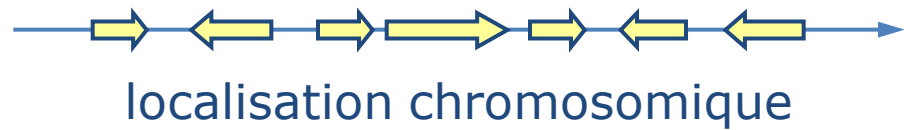
Exemple

P07245	Amino-acid biosynthesis	ATP-binding	Complete proteome	Cytoplasm
Q03677	Amino-acid biosynthesis	Cell membrane	Complete proteome	Dioxygenase
P22768	Amino-acid biosynthesis	Arginine biosynthesis	ATP-binding	Complete proteome
P05150	Amino-acid biosynthesis	Arginine biosynthesis	Complete proteome	Cytoplasm
P04076	Amino-acid biosynthesis	Arginine biosynthesis	Complete proteome	Lyase
Q01217	Amino-acid biosynthesis	Arginine biosynthesis	Complete proteome	Kinase
P18544	Amino-acid biosynthesis	Aminotransferase	Arginine biosynthesis	Complete proteome
P08566	Amino-acid biosynthesis	Aromatic amino acid biosynthesis	ATP-binding	Complete proteome
P14843	Amino-acid biosynthesis	Aromatic amino acid biosynthesis	Complete proteome	Phosphoprotein
P49089	Amino-acid biosynthesis	Asparagine biosynthesis	Complete proteome	Glutamine amidotransferase
P49090	Amino-acid biosynthesis	Asparagine biosynthesis	Complete proteome	Glutamine amidotransferase
P22801	Amino-acid biosynthesis	Aminotransferase	Branched-chain amino acid biosynthesis	Complete proteome

Sources de données



voies métaboliques



complexes protéiques

ensembles de gènes

Nucleic Acids Research Advance Access published May 28, 2008

Nucleic Acids Research Advance Access published May 28, 2008

Nucleic Acids Research Advance Access published May 28, 2008
Nucleic Acids Research, 2008, 1-8
 doi:10.1093/nar/gkn225

ENDEAVOUR update: a web resource for gene prioritization in multiple species

Léon-Charles Tranchevent¹, Roland Barriot¹, Shi Yu¹, Steven Van Vooren¹, Peter Van Loo^{1,2,3}, Bert Coessens¹, Bart De Moor¹, Stein Aerts^{3,4} and Yves Moreau^{1,*}

¹Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, ²Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB Leuven, ³Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and ⁴Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven (Belgium)

Received February 7, 2008; Revised April 30, 2008; Accepted May 7, 2008

ABSTRACT
 Endeavour (<http://www.esat.kuleuven.be/endeavour>) web: this web site is free and open to all users and there is no login requirement) is a web resource for the prioritization of candidate genes. Using a training set of genes known to be involved in a biological process of interest, our approach consists of (i) inferring several models (based on various genomic data sources), (ii) applying each model to the candidate genes to rank those candidates against the profile of the known genes and (iii) merging the several rankings into a global ranking of the candidate genes. In the present

BACKGROUND
 With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, converting genomic data into biological knowledge to identify genes involved in a particular process or disease remains a major challenge. Nevertheless, there is much evidence to suggest that functionally related genes often cause similar phenotypes (1-5). To identify which genes are responsible for which phenotype, association studies and linkage analyses are often used, resulting in large lists of candidate genes. In

co-citation



domaines protéiques

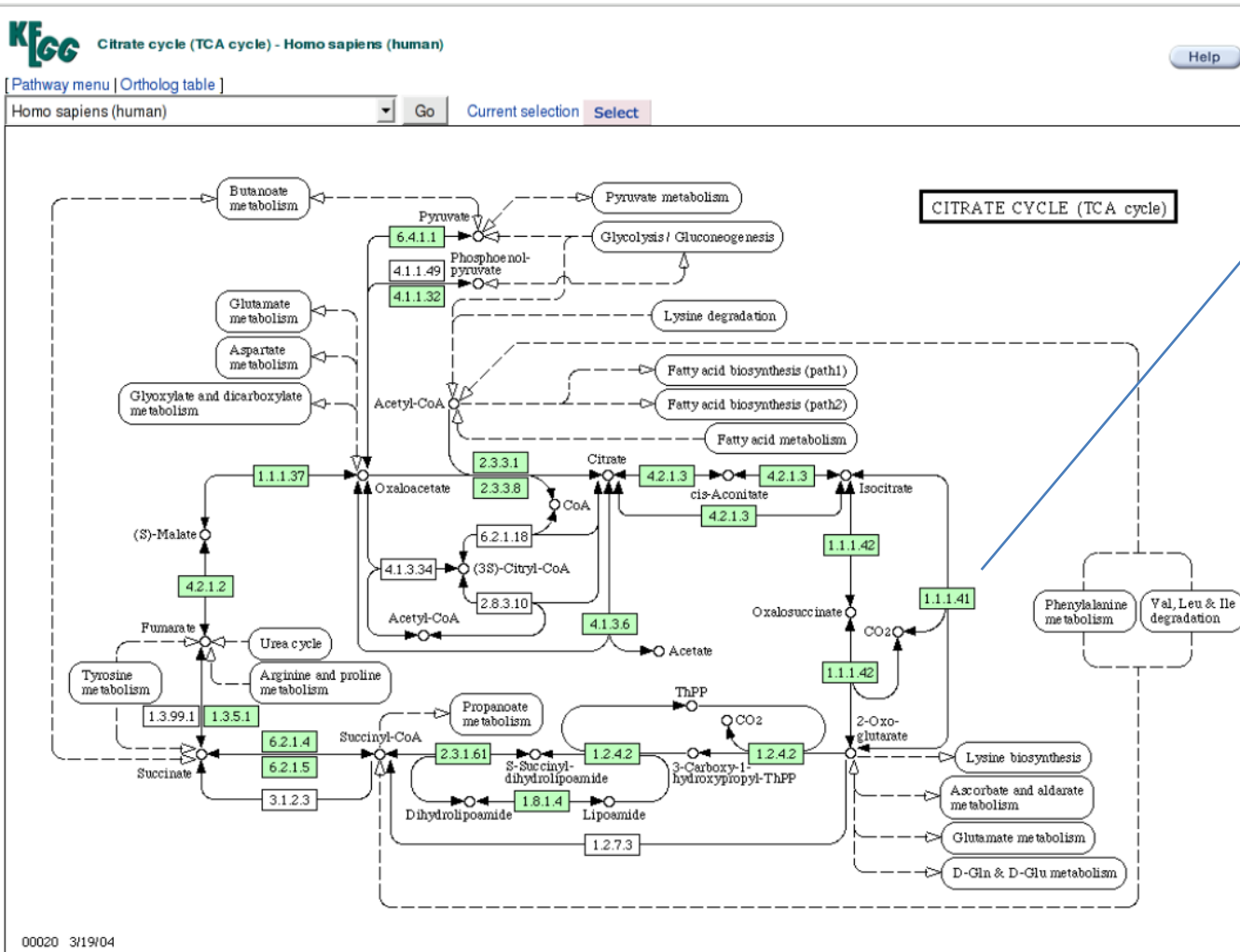


Gene Ontology

KEGG Pathways

- Classification de processus biologiques
 1. Metabolism
 1. Carbohydrate Metabolism
 - Glycolysis / Gluconeogenesis
 - Citrate cycle (TCA cycle)
 - ...
 2. ...
 2. Genetic Information Processing
 3. Environmental Information Processing
 4. Cellular Processes
 5. Human Diseases
 6. Drug Development

KEGG Pathways



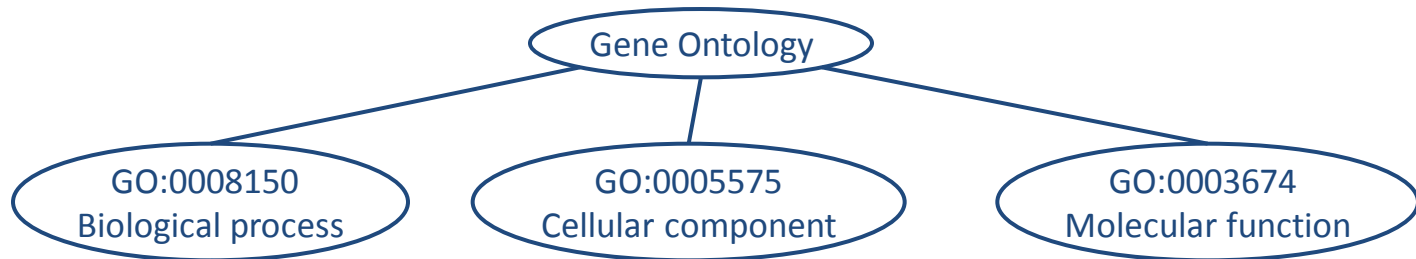
code d'enzyme
(EC number)

Ensemble des gènes codant pour les enzymes impliquées dans un pathway

Gene Ontology

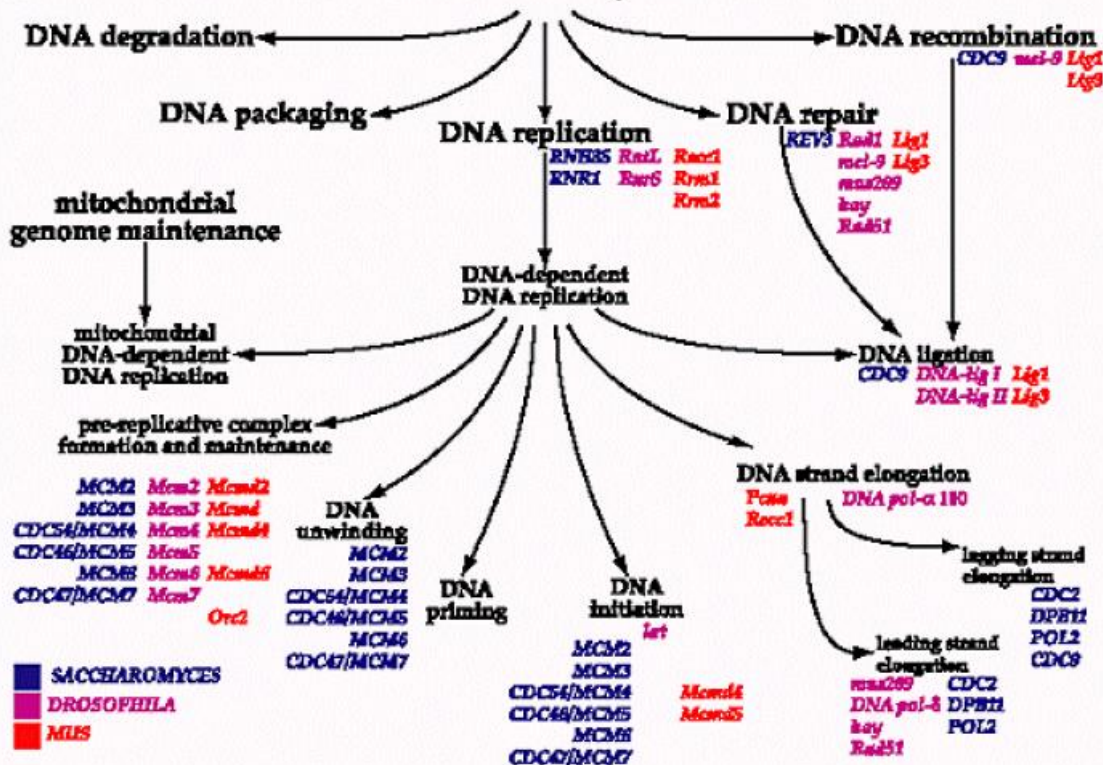
- Vocabulaire contrôlé : le même terme pour parler de la même chose
- Ensemble de termes (définitions) reliés par des relations de type est-un ou fait-parti-de
- Trois ontologies:
 - Biological process
 - Molecular function
 - Cellular component

Gene Ontology



a

DNA metabolism



- à chaque terme correspond un ensemble de gènes annotés avec ce terme ou un plus spécifique

Mots-clés Uniprot/Swissprot

- à chaque mot-clé correspond un ensemble de protéines annotées avec ce mot-clé

```


DR EMBL; M73748; AAA39866.1; -; mRNA.
DR EMBL; M96646; AAA37724.1; -; mRNA.
DR EMBL; AJ250246; CAB58997.1; -; mRNA.
DR EMBL; AJ297944; CAC16152.1; -; mRNA.
DR EMBL; AY115493; AAM66761.1; -; Genomic_DNA.
DR EMBL; AK158855; BAE34695.1; -; mRNA.
DR EMBL; BC026551; AAH26551.1; -; mRNA.
DR Ensembl; ENSMUSG00000028583; Mus musculus.
DR KEGG; mmu:14726; -.
DR MGI; MGI:103098; Pdpn.
DR ArrayExpress; Q62011; -.
DR RSPD-ProtExp; IOM20239; -.
DR GO; GO:0030175; C:filopodium; IDA.
DR GO; GO:0030027; C:lamellipodium; IDA.
DR GO; GO:0005886; C:plasma membrane; IDA.
DR GO; GO:0001726; C:ruffle; IDA.
DR GO; GO:0000902; P:cellular morphogenesis; IDA.
DR GO; GO:0030324; P:lung development; IMP.
DR GO; GO:0001946; P:lymphangiogenesis; IMP.
DR GO; GO:0051272; P:positive regulation of cell motility; IDA.
DR InterPro; IPR008783; Podoplanin.
DR PANTHER; PTHR16861; Podoplanin; 1.
DR Pfam: PF05808; Podoplanin; 1.
KW Cell shape; Developmental protein; Direct protein sequencing;
KW Glycoprotein; Membrane; Sialic acid; Signal; Transmembrane.
FT SIGNAL 1 22 Potential.
FT CHAIN 23 172 Podoplanin.
FT /FTid=PRO_0000021352.
FT TOPO_DOM 23 141 Extracellular (Potential).
FT TRANSMEM 142 162 Potential.
FT TOPO_DOM 163 172 Cytoplasmic.
.....
FT CONFLICT 29 31 EDD -> KNN (in Ref. 2).
FT CONFLICT 38 39 GD -> EN (in Ref. 1).
SQ SEQUENCE 172 AA; 18233 MW; C035ED251918CE6F CRC64;
MNTVPLVLFVW LGSVWFWD SA QGGTIGVNED DIVTPGTGDG MVPPGIEDKI TTTGATGGLN
ESTGKAPLVP TQRRERGKPP LEELSTSAT S DHDHREH EST TTVKVVTSHS VDKKTSHPNR
DNAGDETQTT DKKDGLPVVT LVGIIVGVLL AIGFVGGIFI VVMKKISGRF SP



```

//

Domaines protéiques

- InterPro intègre les principales banques de domaines (Pfam, ProSite, SMART)
- à un domaine correspond un ensemble de protéines

EMBL-EBI  All Databases

Databases Tools EBI Groups Training Industry About Us Help Site Index  

EBI > Databases > InterPro

Jump to: [InterProScan](#) [Databases](#) [Documentation](#) [FTP site](#) [Help](#) [Advanced search](#)

InterPro: IPR000254 Cellulose-binding region, fungal

Protein matches

UniProtKB Matches: 504 proteins

Overview: [sorted by AC](#), [sorted by name](#), [of known structure](#), [proteins with splice variants](#)
 Detailed: [sorted by AC](#), [sorted by name](#), [of known structure](#), [proteins with splice variants](#)
 Table: [For all matching proteins](#), [of known structure](#)
[Architectures](#)
[Accession List](#)

Accession IPR000254 CBD_fun

Type Domain

Database	ID	Name	Proteins
Pfam	PF00734	CBM_1	487
PROSITE pattern	PS00562	CBM1_1	417
PROSITE profile	PS51164	CBM1_2	480
SMART	SM00236	fCBD	454
SuperFamily	SSF57180	CBD_fun	485

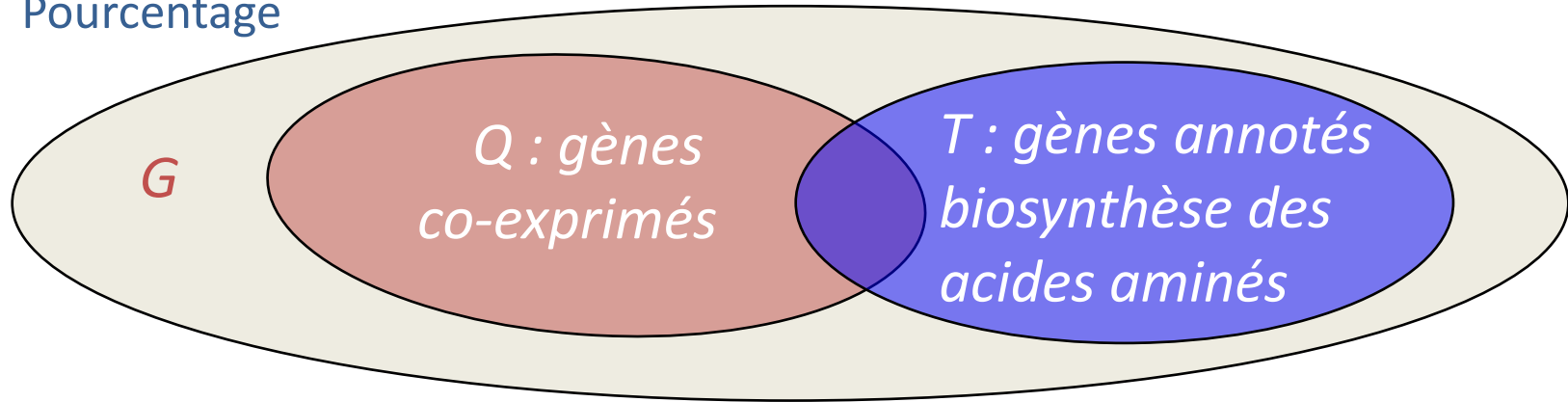
Signatures

GO Term annotation

Process	GO:0005975 carbohydrate metabolic process
Function	GO:0004553 hydrolase activity, hydrolyzing O-glycosyl compounds GO:0030248 cellulose binding
Component	GO:0005576 extracellular region

Test de surreprésentation

- Loi binomiale
- χ^2
- Pourcentage



- Loi hypergéométrique : probabilité d'avoir au moins le nombre d'éléments communs observé entre 2 échantillons issus d'une même population
 - test de surreprésentation

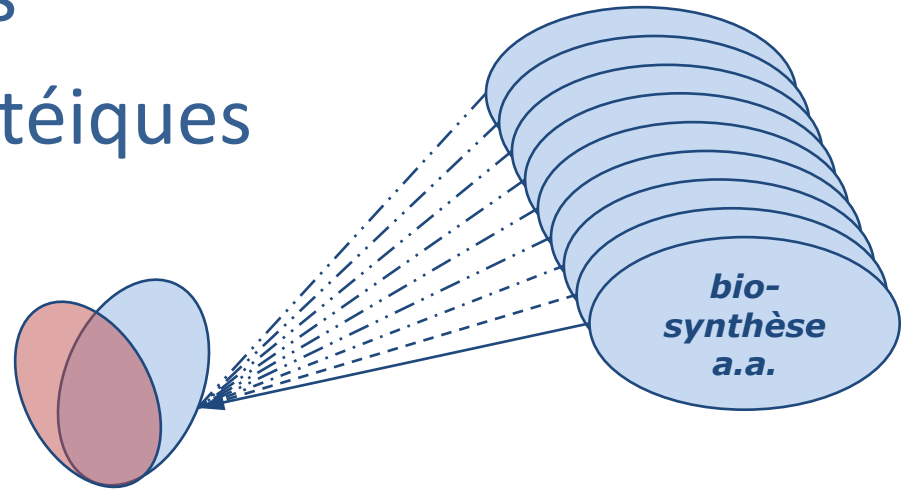
$$p - \text{valeur} (c, t, q, g) = \sum_{k=c}^{\min(q, t)} \frac{\binom{t}{k} \binom{g-t}{q-k}}{\binom{g}{q}}$$

avec

- $g = |G|$: nombre total de gènes
- $q = |Q|$: nombre de gènes co-exprimés
- $t = |T|$: nombre de gènes annotés biosynthèse des a.a.
- $c = |Q \cap T|$: nombre de gènes communs

Recherche de caractéristiques communes

- Annotations Gene Ontology
- Domaines protéiques
- Complexes multi-protéiques
- Voies métaboliques
- ...

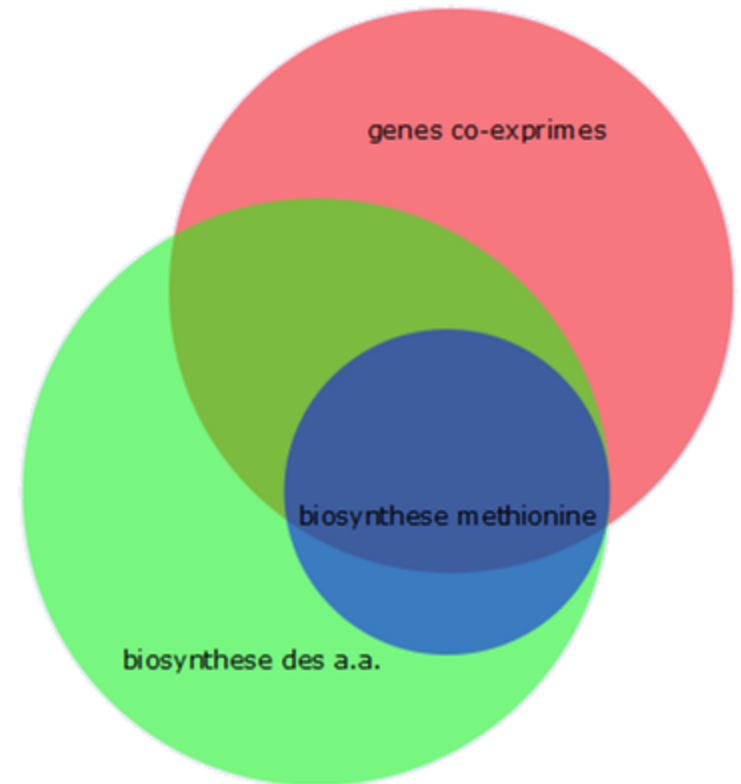


- Correction pour tests multiples (FDR ou autre)
- On conserve les caractéristiques statistiquement significatives

Visualisation

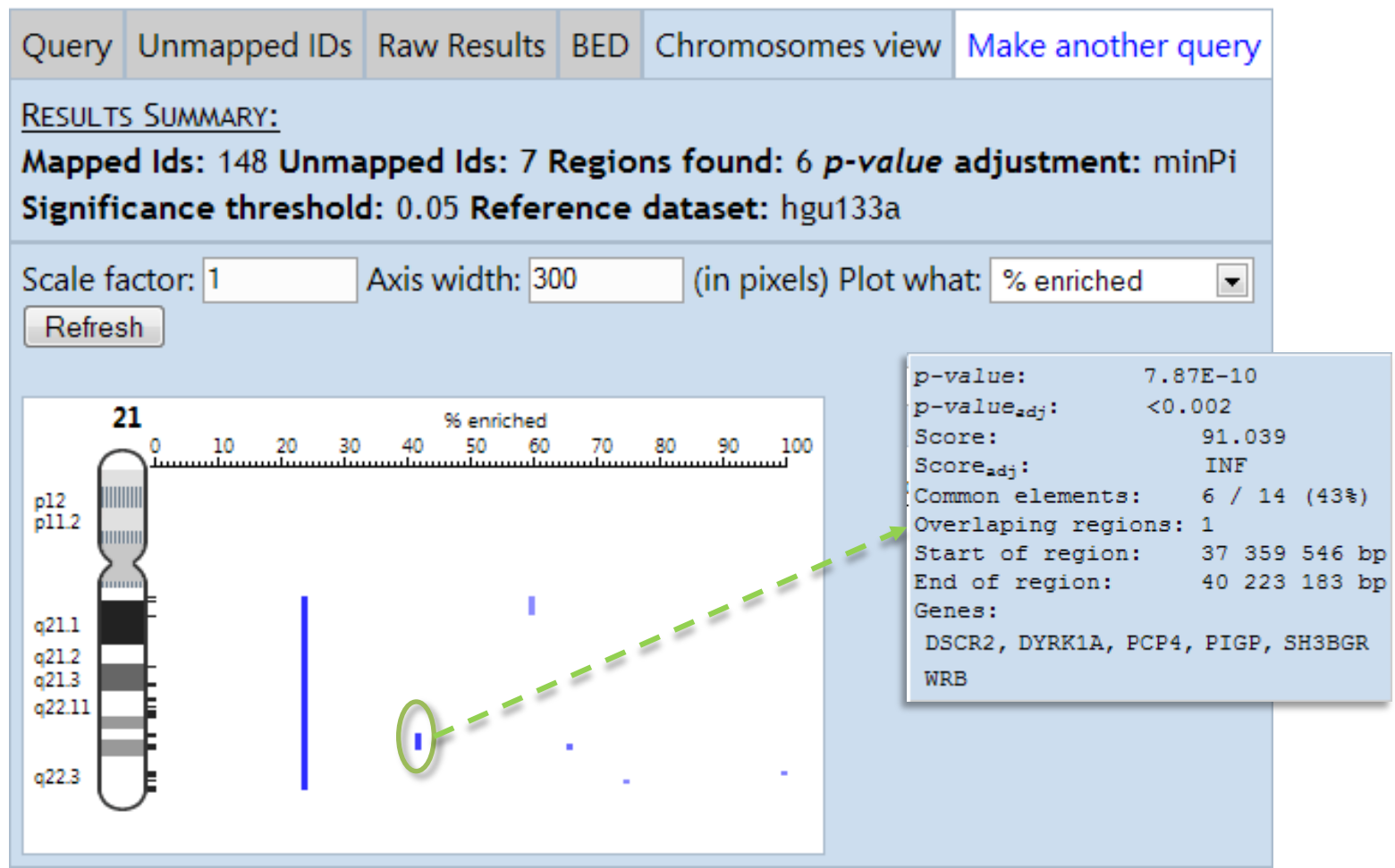
- Diagramme de Venn
 - Aire proportionnelle à la taille des ensembles
 - Chevauchement proportionnel aux gènes communs
 - possible pour un petit nombre d'ensembles

Diagramme de Venn



Application

- Gènes différentiellement exprimés dans le cerveau des patients atteints du syndrome de Down (trisomie 21)



Communauté, standards et banques de données

- Microarray Gene Expression Data (MGED) society
- MIAME (Minimum Information About a Microarray Experiment)
 - interprétation non ambiguë
 - reproductibilité
- MGED (MicroArray Gene Expression Data)
 - MAGE-ML (Markup Language): format d'échange
 - MAGE-OM (Object Model)
 - MGED Ontology: vocabulaire contrôlé
- Entrepôts
 - GEO (Gene Expression Omnibus) au NCBI
 - ArrayExpress
 - SMD (Stanford Microarray Database)