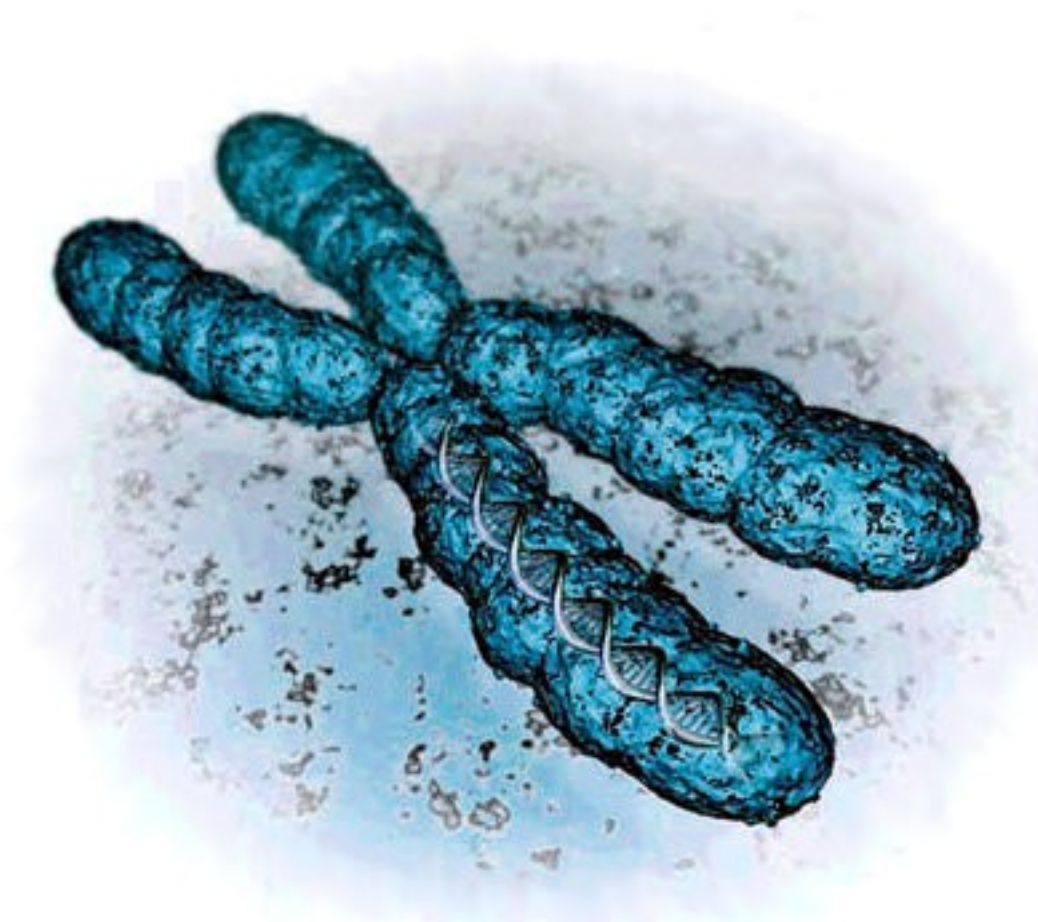


Sylvain Mareschal [[@etu.univ-rouen.fr](mailto:@etu.univ-rouen.fr)]

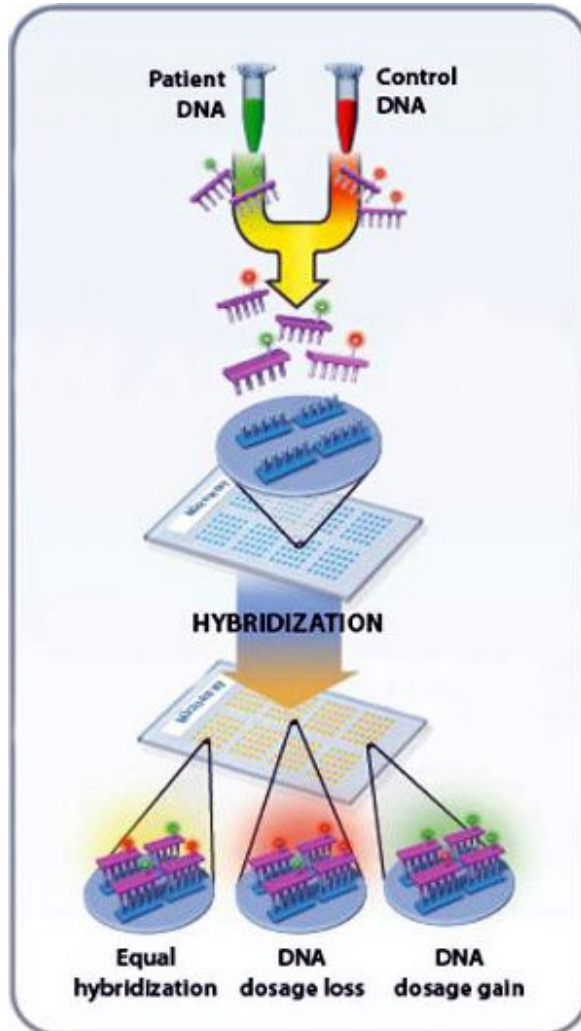
# CGH-array



Van de Wiel et al (2011) Brief Bioinform. Jan;12(1):10-21.  
*Preprocessing and downstream analysis of microarray DNA copy number profiles.*

# 1. Théorie

## 1.1. Principe



### Fragmentation du génome

Enzymes de restriction  
Sonication

### Hybridations compétitives

Un ADN cible (vert)  
Un ADN référence (rouge)

### Localement sur le génome

Historiquement BAC ( $10^2$  à  $10^3$  pour hg)  
60-mer généralement ( $10^5$  à  $10^6$  actuellement)  
Réparties sur le génome ou la région cible

### Deux scans successifs

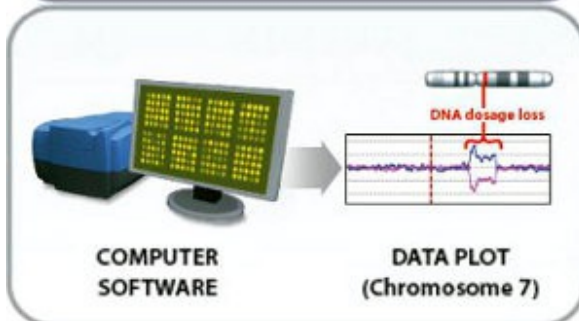
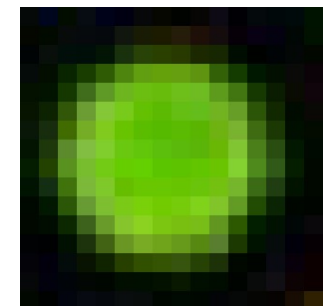
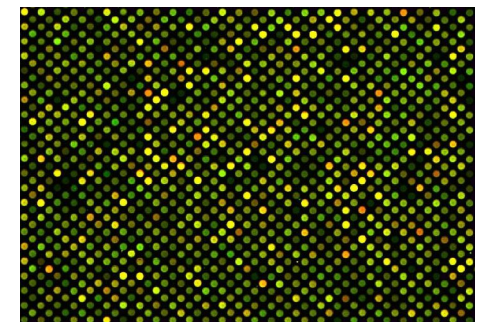
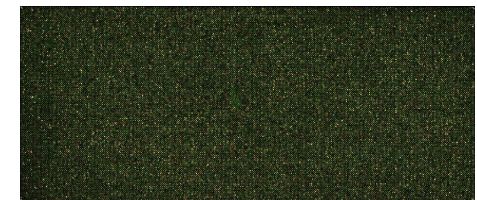
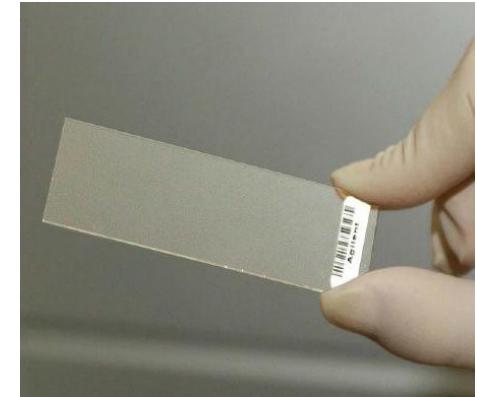
Fluorescences verte et rouge  
Identification grâce aux sondes de position  
Qualité (sondes contrôles, pixels *outliers*)

### Calcul des log-ratios

$$M = \log_2(F_{\text{vert}} / F_{\text{rouge}})$$

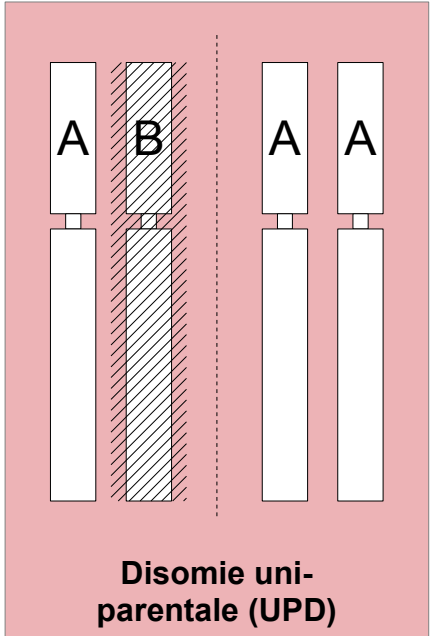
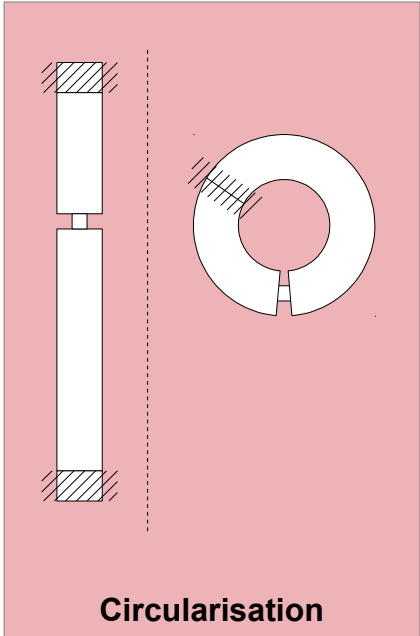
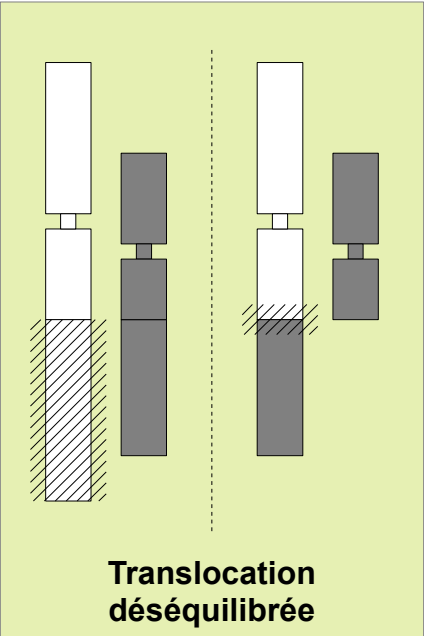
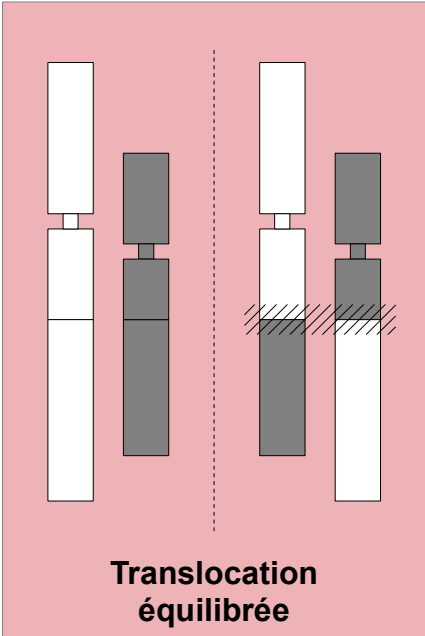
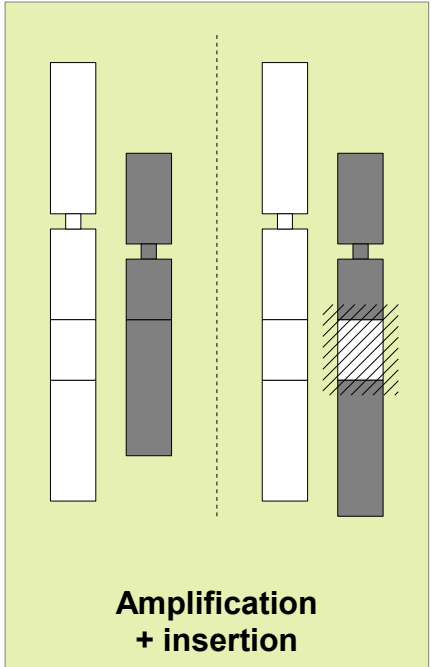
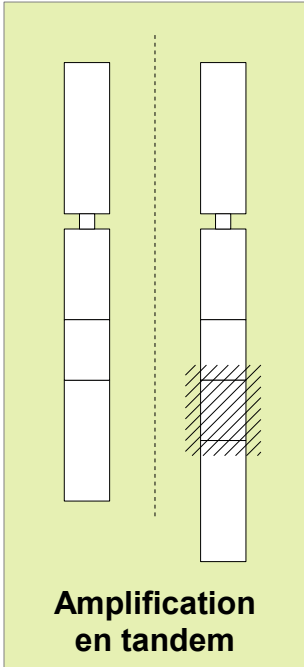
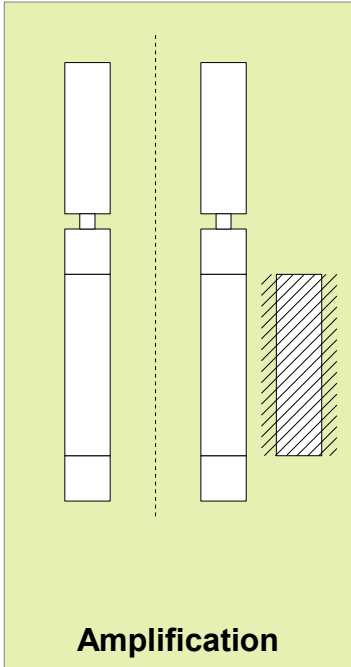
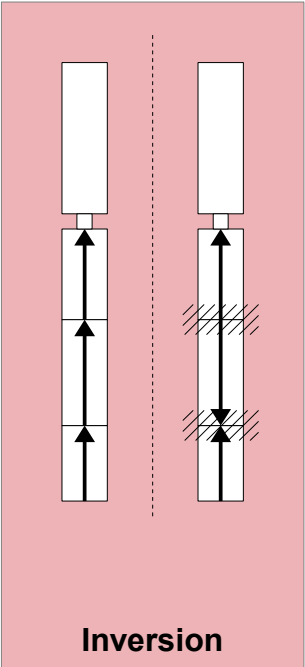
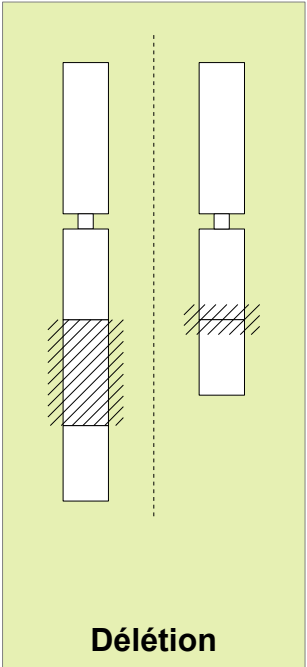
$M > 0$  : amplification

$M < 0$  : délétion



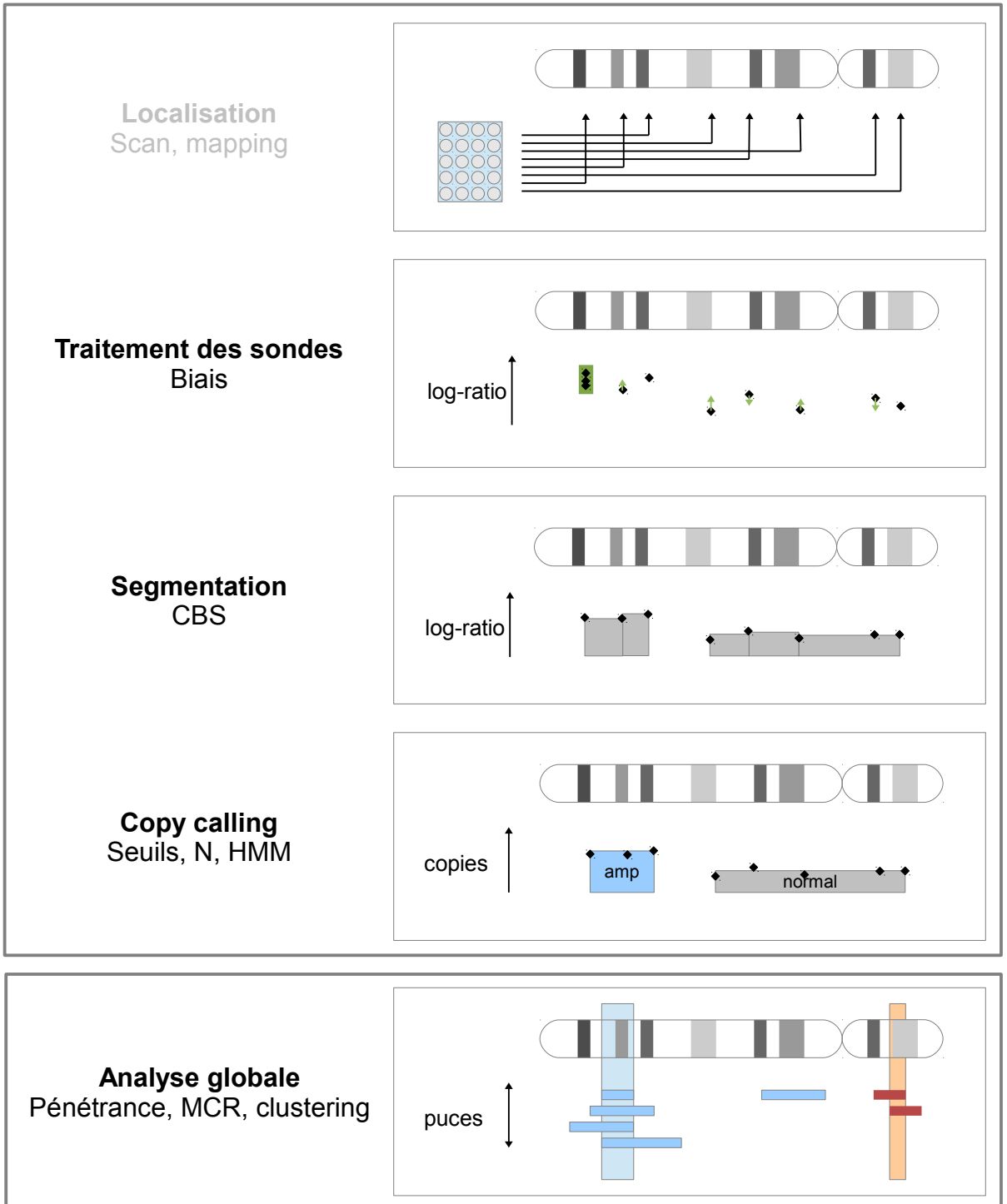
# 1. Théorie

## 1.2. Anomalies visibles



# 1. Théorie

## 1.3. Workflow



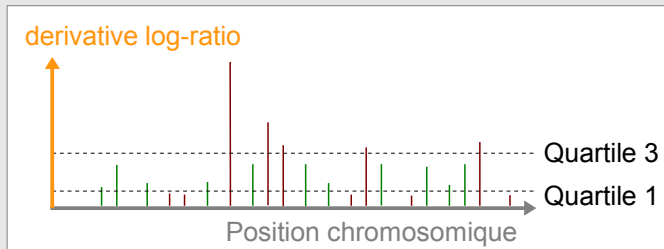
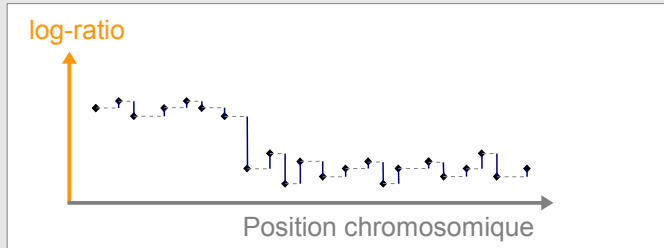
## 2. Pratique - les biais

### 2.1. Biais de sondes - puces en général

#### Mesure de la qualité de la puce

##### Derivative Log Ratio Spread (DLRS)

$$DLRS = \text{IQR}(\text{diff}(M)) / (1,349 * \text{sqrt}(2))$$



Excellent < 0,2 < Good < 0,3 < Poor

Kincaid et al (2007) US 2007/0031883 A1 application patent 2007  
Analyzing CGH data to identify aberrations.

##### Median Absolute Deviation (MAD)

$$\text{MAD} = \text{median}(M - \text{median}(M))$$

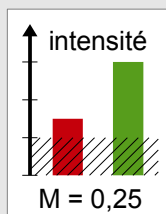
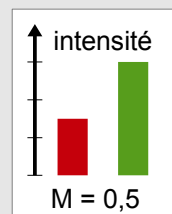
Van de Wiel et al (2011) Brief Bioinform. Jan;12(1):10-21.  
*Preprocessing and downstream analysis of microarray DNA copy number profiles.*

#### Bruit de fond

##### Soustraction

Possible mais peu utilisée  
Effet sur les intensités faibles

##### Filtration éventuelle



#### Biais d'intensité

##### Centralisation

Communément admis  
Médiane / Moyenne inadaptées  
 $M' = M - \text{mode}(M)$   
Attention aux polyploïdes

##### LOWESS

$$A_{\text{amplification}} > A_{\text{normal}} > A_{\text{deletion}}$$

Risqué sur ADN complexes (perte d'altérations)

##### LOWESS sur sondes normales

Même principe, en excluant les altérations

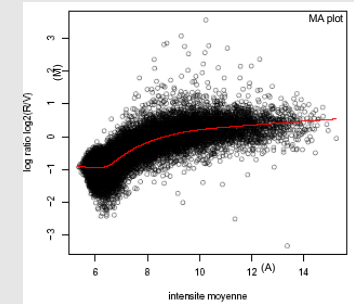
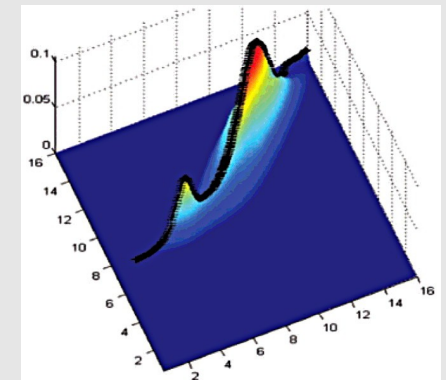
Revient à analyser les données durant la pré-analyse !

- K-mean clustering      Staaf et al (2007) BMC Genomics.
- Segmentation              Van Houte et al (2010) Bioinformatics.

##### Densité 2D

Régression de la crête  $d(R, V)$   
Essentiellement des sondes normales (majoritaires)

Chen et al (2008) Bioinformatics.  
Aug 15;24(16):1749-56.



Caroline Bérard

#### Normalisation inter-array

##### Variance, quantiles

Hypothèse de distributions similaires déraisonnable  
A éviter, sauf si profils équivalents (rare)

## 2. Pratique - les biais

### 2.2. Biais de sondes - CGH spécifiques

#### Vagues

##### Phénomène mal connu

+ ou - fort selon la quantité d'ADN  
Lié à la composition en GC

#### WACA

Leprêtre et al (2010) NAR Apr;38(7):e94.  
*Waved aCGH: to smooth or not to smooth.*

#### 5 LOWESS successives

- %GC(sonde +/- 150kb)
- %GC(sonde +/- 500kb)
- %GC(sonde)
- %GC(fragment génomique ciblé)
- taille(fragment génomique ciblé)

#### Calculs préalables nécessaires

Calculer les biais par sonde, par design  
Modification selon le protocole utilisé

#### Implémentation peu confortable

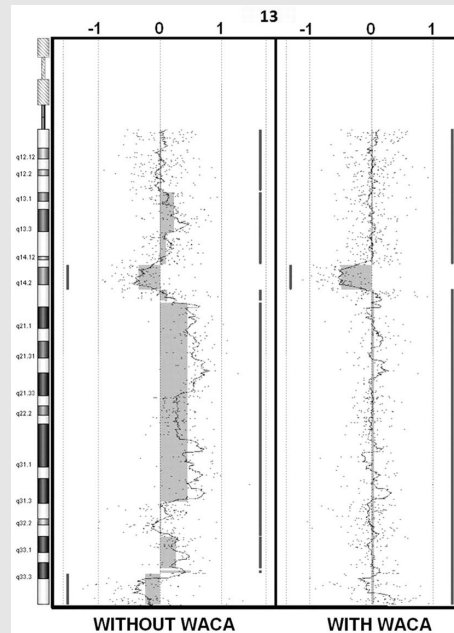
RReportGenerator, Agilent, pas de calcul des biais  
Voir cghRA.probes

#### Plaques de calibration

Van de Wiel et al (2009) Bioinformatics. May 1;25(9):1099-104  
*Smoothing waves in array CGH tumor profiles.*

#### Régression sur données d'hybridation normal / normal

Nécessite une série du même design  
Nombreux biais corrigés



Leprêtre et al 2009

#### Biais spatial

##### CGH : positions aléatoires

Pas de lien spatial puce / chromosome  
Répartition homogène attendue des M

##### Artefact local

Traces de doigt, de colle

##### Gradient

Mélange de réactifs hétérogène  
Lames mal immergées

##### Canal spécifique ou non

Rarement identique sur les 2 canaux  
Impact sur le log-ratio à surveiller

##### Impact fort

Augmente le bruit sur tout le génome  
Perturbe la modélisation

##### Corriger les artefacts locaux

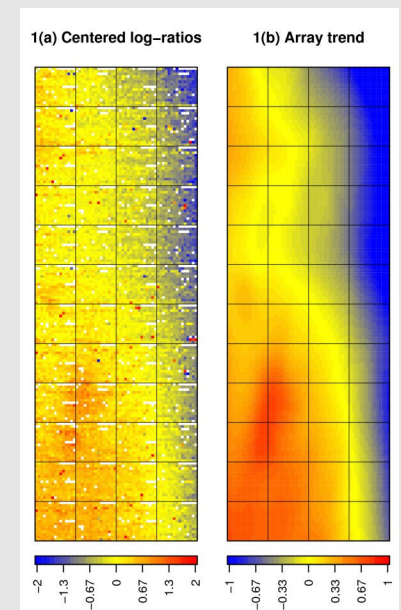
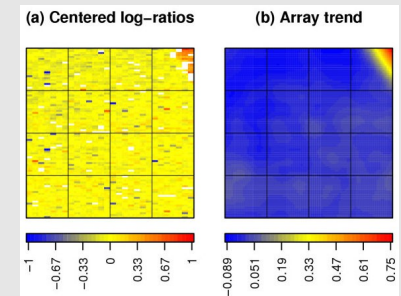
Seule solution : exclure les sondes  
- Délimitation manuelle  
- *Clustering spatial* (MANOR)

##### Corriger les gradients

Régression envisageable  
- Plus simple : exclure la puce  
- Corrigé indirectement par la print-tip LOWESS  
- 2D LOWESS (MANOR)

##### Bioconductor 'MANOR'

Neuval et al (2006) BMC  
Bioinformatics. May 22;7:264.  
*Spatial normalization of array-CGH data.*



Neuval et al 2006

## 2. Pratique - les biais

### 2.3. Biais d'expérience

#### Cellularité

##### ADN cible rarement homogène

Nombreuses sous-populations cellulaires  
Une altération peut n'en toucher qu'une partie

##### Effet sur les log-ratios

$$M_{\text{observé}} = \text{moyenne}(M_{\text{chaque cellule}})$$

Effet variable selon les proportions

##### Cellules normales

Non concernées par les altérations (2 copies)

Resserrent M autour de 0

Exemple d'une délétion :

0% normales	$\log(1/2) = -1$
10 % normales	$\log((1/2)*90\% + (2/2)*10\%) = -0,86$
20 % normales	$\log((1/2)*80\% + (2/2)*20\%) = -0,74$
50 % normales	$\log((1/2)*50\% + (2/2)*50\%) = -0,42$

##### Sous-populations tumorales

CNV propres à une ou plusieurs sous-populations  
Analyse complexe (fréquent dans les tumeurs solides)  
Modélisation du nombre de copies risquée

##### Effacer les cellules normales

Estimation cytologique de  $p_{\text{tum}}$

$$M = \log_2\left(\left(\frac{i_{\text{cible}}}{i_{\text{réf}}} * p_{\text{tum}}\right) + \left(\frac{i_{\text{réf}}}{i_{\text{réf}}} * (1 - p_{\text{tum}})\right)\right)$$

$$M' = \log_2\left(\frac{i_{\text{cible}}}{i_{\text{réf}}}\right)$$

$$M' = \log_2\left(\frac{2^M - 1 + p_{\text{tum}}}{p_{\text{tum}}}\right)$$

##### Éviter les sous-populations

- Micro-dissection

- *Single-cell* (micromanipulation, *Whole Genome Amplification*)

Hannemann et al (2011) PLoS One. 6(11):e26362

*Quantitative high-resolution genomic analysis of single cancer cells.*

#### Anomalies constitutionnelles

##### Attention à l'ADN de référence

Habituellement : pool d'ADN normaux  
→ variations constitutionnelles

##### Polymorphismes importants

> 200 CNV (76 loci) sur 20 individus, jusqu'à 100 kb

Sebat et al (2004) Science. Jul 23;305(5683):525-8.

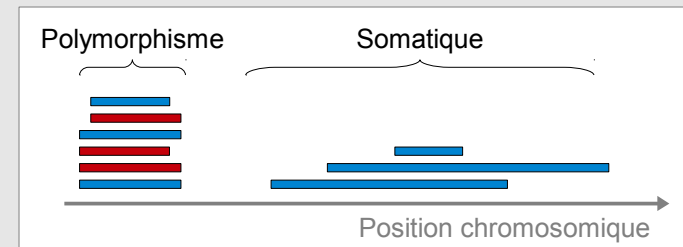
*Large-scale copy number polymorphism in the human genome.*

##### Solution 1 : filtration à l'échelle de la série

Régions généralement courtes (< 100 kb)

Très fort recouvrement d'une puce à l'autre

Amplifiées et délétées généralement



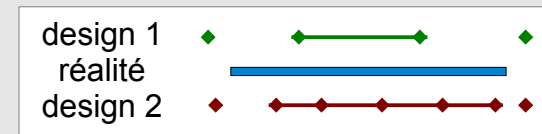
##### Solution 2 : filtration selon des données publiques

Database of Genomic Variants

Répertoire de CNV constitutionnels

Méthodes et designs très hétérogènes (CGH-BAC, NGS ...)

Recouvrement jamais parfait



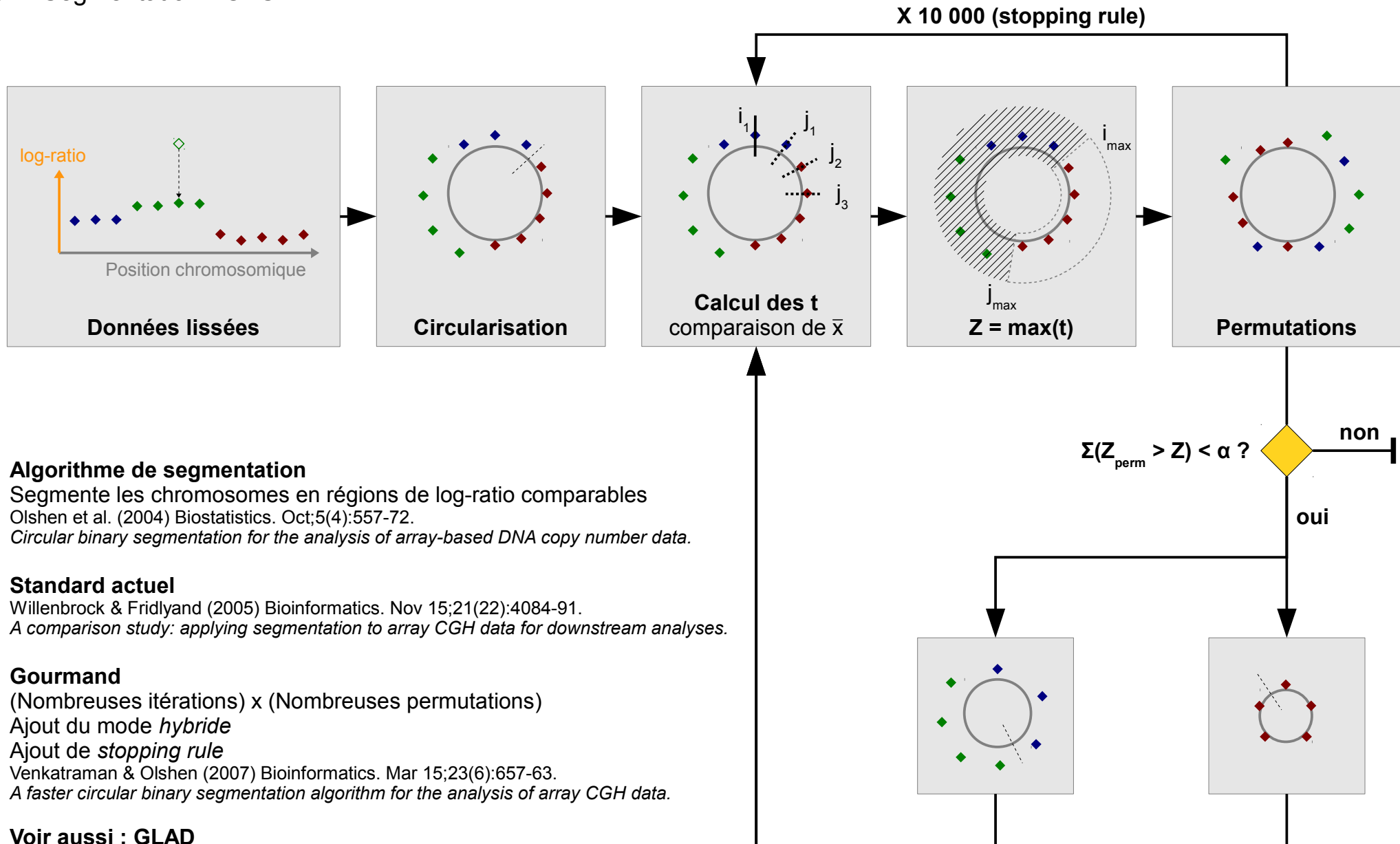
##### Solution 3 : utiliser l'ADN normal correspondant

Anomalie invisible si présente dans la tumeur et la référence (ratio)

Simplifie grandement l'analyse en oncologie

# 3. Pratique - la puce

## 3.1. Segmentation - CBS



### Algorithme de segmentation

Segmente les chromosomes en régions de log-ratio comparables

Olshen et al. (2004) Biostatistics. Oct;5(4):557-72.

*Circular binary segmentation for the analysis of array-based DNA copy number data.*

### Standard actuel

Willenbrock & Fridlyand (2005) Bioinformatics. Nov 15;21(22):4084-91.

*A comparison study: applying segmentation to array CGH data for downstream analyses.*

### Gourmand

(Nombreuses itérations) x (Nombreuses permutations)

Ajout du mode *hybride*

Ajout de *stopping rule*

Venkatraman & Olshen (2007) Bioinformatics. Mar 15;23(6):657-63.

*A faster circular binary segmentation algorithm for the analysis of array CGH data.*

### Voir aussi : GLAD

Hupé et al (2004) Bioinformatics. Dec 12;20(18):3413-22

*Analysis of array CGH data: from signal ratio to gain and loss of DNA regions.*



# 3. Pratique - la puce

## 3.2. Calling - au plus simple

### Problème trivial en théorie

Délétion homozygote  $M_{do} = \log_2(0/2) \rightarrow -\text{inf}$

Délétion hétérozygote  $M_{dt} = \log_2(1/2) = -1$

Normal  $M_n = \log_2(2/2) = 0$

Amplification  $M_{dt} = \log_2(3/2) = 0,585$

Double amplification  $M_{dt} = \log_2(4/2) = 1$

### Nombreux biais en pratique

Qualité d'ADN, manipulation, mesure, cellularité ...

### Utilisation de seuils

Avec 2 seuils, perte de la distinction du nombre de copies  
Choix délicat

### Seuils unitaires

Seuils :  $0,8 < \text{ratio} < 1,2 \rightarrow -0,32 < M < 0,26$

Basé sur l'observation d'hybridations normal / normal

**Généralisation très risquée**

*Weiss et al., J Pathol 2003*

### Seuils quantiles

Considère les 50% centraux ( $Q_1 \rightarrow Q_3$ ) comme normaux

Seuils +/- 1 copie :  $\text{mean}(M_{\text{normal}}) \pm 4 \cdot \text{sd}(M_{\text{normal}})$

Seuils +/- 2 copies : 3<sup>e</sup> et 97<sup>e</sup> percentiles

**Hypothèse forte sur la distribution des log-ratios**

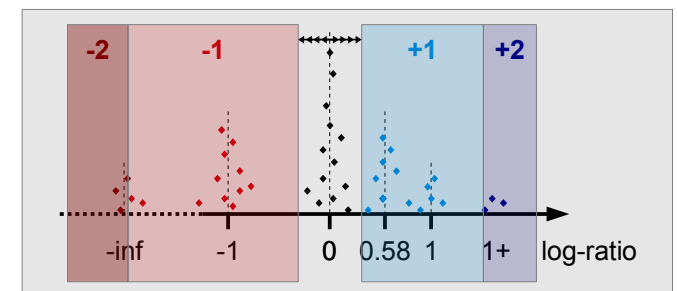
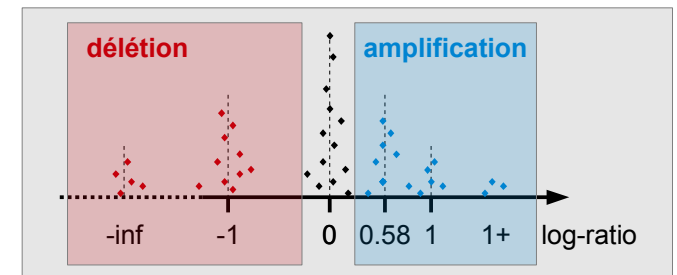
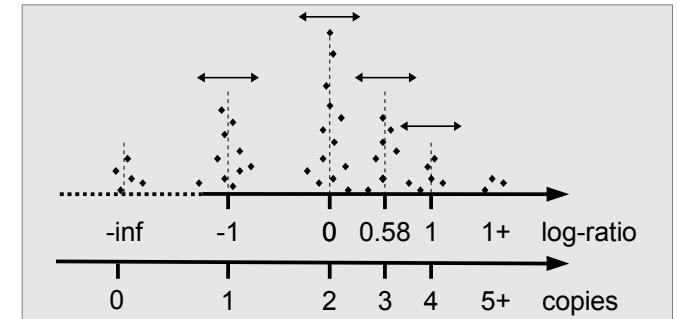
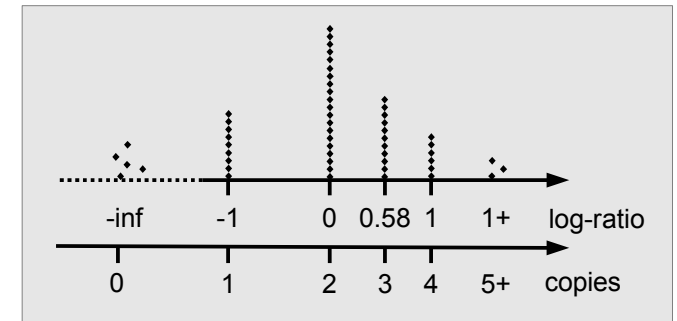
*Aguirre et al., PNAS 2004*

### Seuils assistés

Observation du log-ratio d'anomalies larges confirmées (FISH, caryotype ...)

Corrélation entre techniques pas toujours évidente

Coût supplémentaire



# 3. Pratique - la puce

## 3.3. Calling - modèles de mélange gaussien

### Problème trivial en théorie

Délétion homozygote  $M_{do} = \log_2(0/2) \rightarrow -\text{inf}$

Délétion hétérozygote  $M_{dt} = \log_2(1/2) = -1$

Normal  $M_n = \log_2(2/2) = 0$

Amplification  $M_{dt} = \log_2(3/2) = 0,585$

Double amplification  $M_{dt} = \log_2(4/2) = 1$

### Nombreux biais en pratique

Qualité d'ADN, manipulation, mesure, cellularité ...

### Modélisation gaussienne

Variation à ce modèle simple : bruit de fond

### Application directe aux sondes

Estimation difficile, peu utilisé

*Picard et al., BMC Bioinformatics 2005*

### Application aux segments : CGHcall

*Van de Wiel et al., Bioinformatics 2007*

### 6 états autorisés

$$M_{\text{probe,segment,sample}} \sim N(\mu_{\text{segment,sample}}, \sigma^2)$$
$$\mu_{\text{segment,sample}} \sim \sum (p_{\text{etat}} * N(\mu_{\text{etat}}, \sigma^2_{\text{etat}}))$$

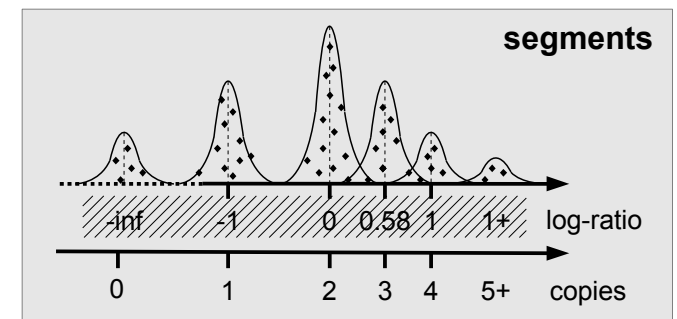
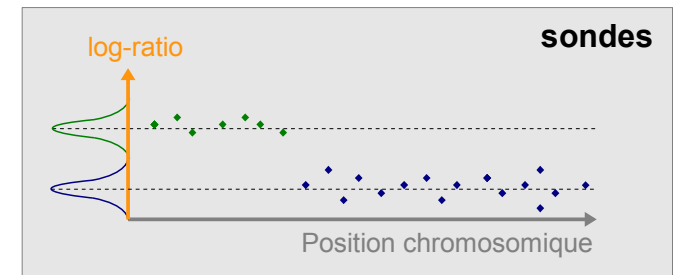
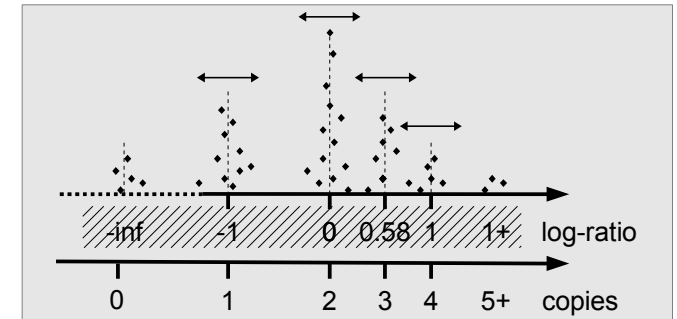
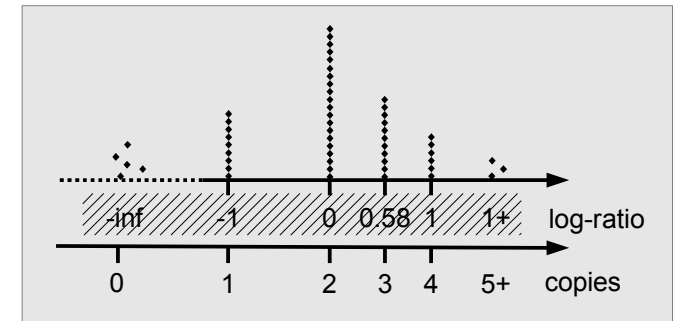
### Moyennes exactes inconnues

Biais de cellularité, fluorochrome ...

### Estimation des paramètres

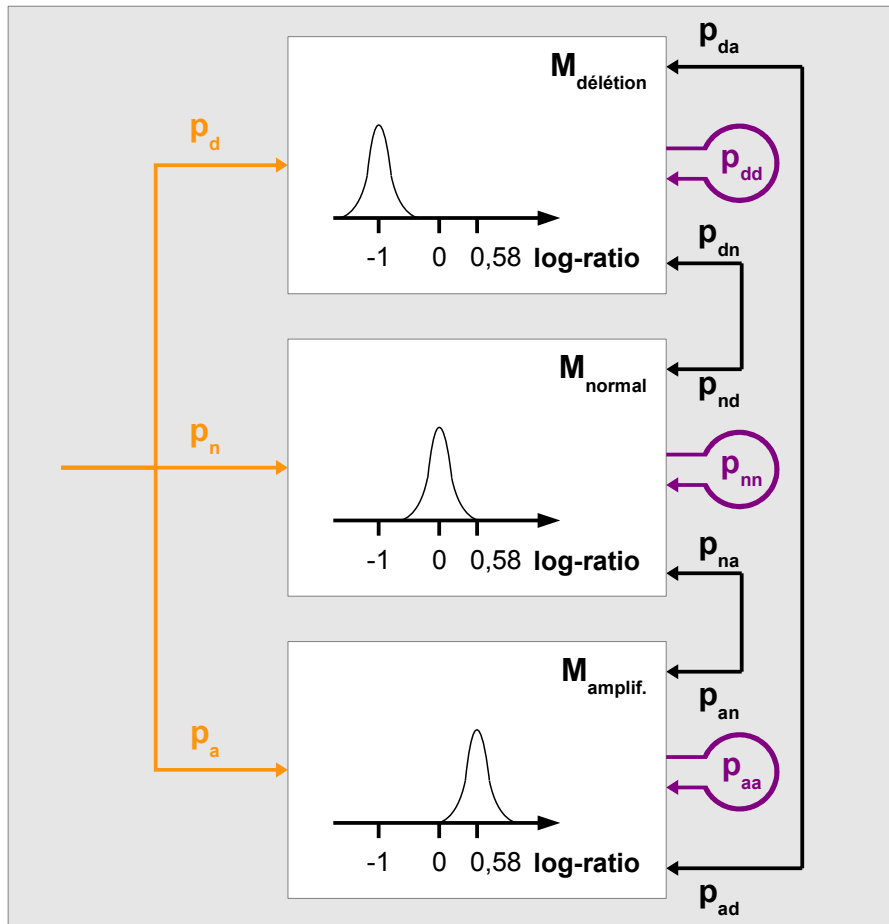
Expectation-Maximization

Modèle commun à la série (suppose les cellularités identiques !)



### 3. Pratique - la puce

#### 3.4. Calling - modèles de Markov cachés



#### Nombre restreint d'états pour une sonde

Délétée, normale ou amplifiée

Un modèle gaussien pour le log-ratio de chaque état

#### Sondes le long d'un chromosome = suite d'états

Le plus probable : même état que la sonde précédente

Possibilité de changer d'état (point de cassure)

	$M_d$	$M_n$	$M_a$
$M_d$	$p_{dd}$	$p_{dn}$	$p_{da}$
$M_n$	$p_{nd}$	$p_{nn}$	$p_{na}$
$M_a$	$p_{ad}$	$p_{an}$	$p_{aa}$

#### État initial aléatoire

$M_d$	$p_d$
$M_n$	$p_n$
$M_a$	$p_a$

#### Problème bien connu

Estimation des modèles et probabilités de transition par algorithme EM

Reconstruction de la chaîne d'états par algorithme de Viterbi (1967)

#### Bioconductor 'aCGH'

Fridlyand et al (2004) J. Multivar. Anal.;90:132-153.

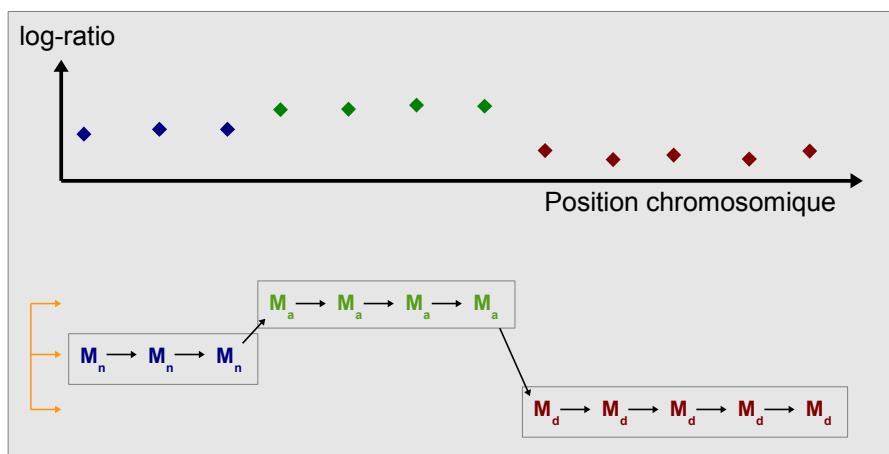
*Hidden Markov models approach to the analysis of array CGH data.*

#### Bioconductor 'snapCGH' (BioHMM)

Prise en compte des distances inter-sondes

Marioni et al (2006) May 1;22(9):1144-6

*BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data.*



# 4. Pratique - l'étude

## 4.1. Anomalies récurrentes

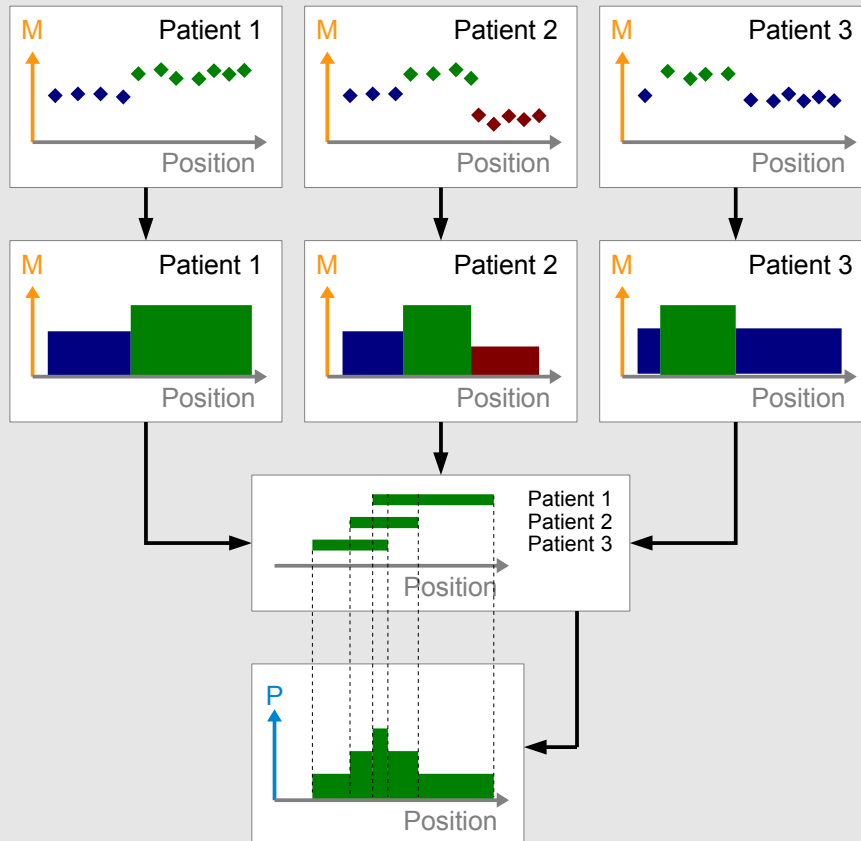
### Pénétrance

#### Proportion de la série concernée

Définie en chaque point du génome

Définie pour chaque type d'altération (amplification, délétion ...)

$$P_{\text{délétion,point}} = \frac{n_{\text{patients délétés,point}}}{n_{\text{patients étudiés}}}$$



#### Met en évidence de larges régions d'intérêt

Bras ou chromosomes entiers

Seuil subjectif

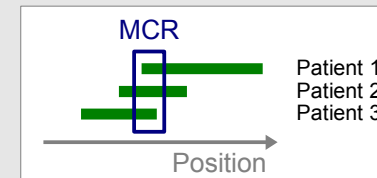
### Minimal Common Regions

#### Région communément altérée dans la série

Recouvrement entre les CNV de puces différentes

Pas de définition consensuelle

#### Littéralement : région commune minimale



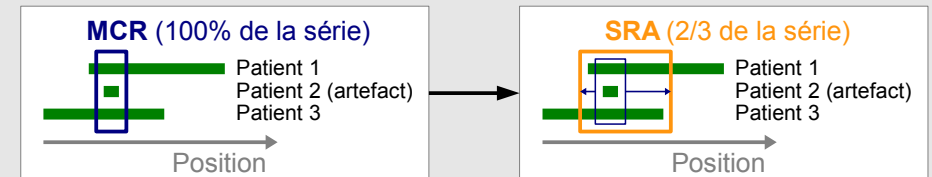
#### Définition un peu extrême

Sensible aux artefacts (micro-altérations dues à une sonde)

→ étendre en diminuant la similarité requise

Lenz et al (2008) PNAS Sep 9;105(36):13520-5

*Molecular subtypes of DLBCL arise by distinct genetic pathways.*



→ étendre en comblant les gaps

Aguirre et al (2004) PNAS. 101(24):9067-9072.

*High-resolution characterization of the pancreatic adenocarcinoma genome.*



# 4. Pratique – l'étude

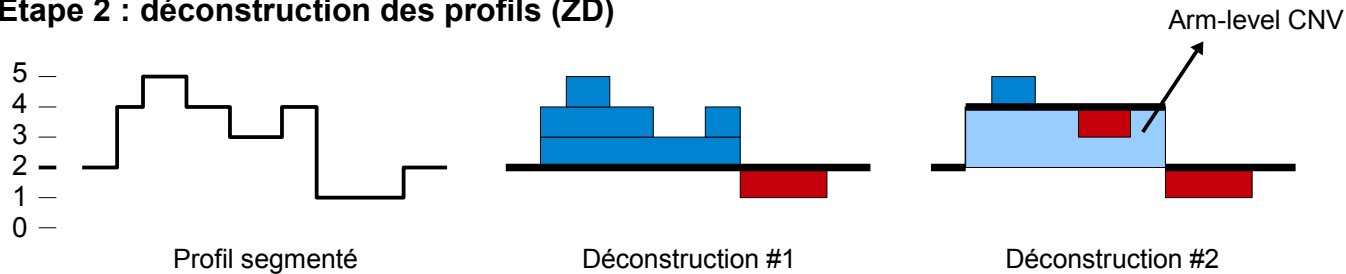
## 4.2. Analyse intégrée : GISTIC 2

Mermel et al., Genom Biol 2011

GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.

**Étape 1 : pre-processing** (voir diapo précédentes)

**Étape 2 : déconstruction des profils (ZD)**



**Étape 3 : calcul des scores par sonde**

- $G \sim \Sigma$  (amplitude x fréquence)
- Relocalisation aléatoire des sondes de chaque patient (permutations)  
Aucune altération récurrente attendue ( $H_0$ )
- $p = f(G_{\text{permuté}} > G_{\text{réel}})$

**Étape 4 : liaison des CNV aux pics ( $G_{\text{max}}$ )**

Définition de « pics » (maxima locaux de  $G$ , > seuil)  
Attribution des CNV observés au pic qu'ils chevauchent

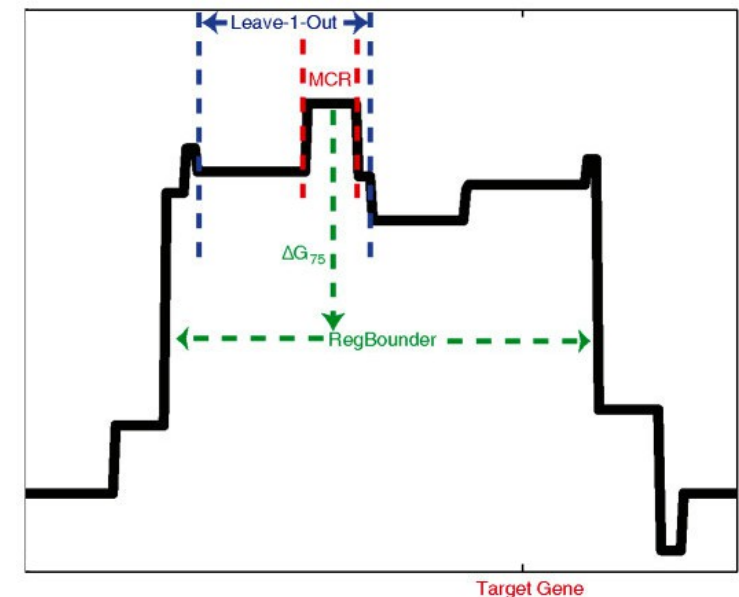
**Étape 5 : extension des pics (RegBouncer)**

Objectif :  $\gamma\%$  de chances de contenir le gène driver  
→  $\gamma$  élevé = régions très élargies

Modélisation de  $G$  par relocalisation aléatoire des segments  
Aucune altération récurrente attendue ( $H_0$ )

**En pratique**

Outil populaire en oncologie (version 1 surtout)  
Implémenté MATLAB, disponible en stand-alone  
Nécessite des données segmentées normalisées, éviter le copy calling préalable



Target Gene

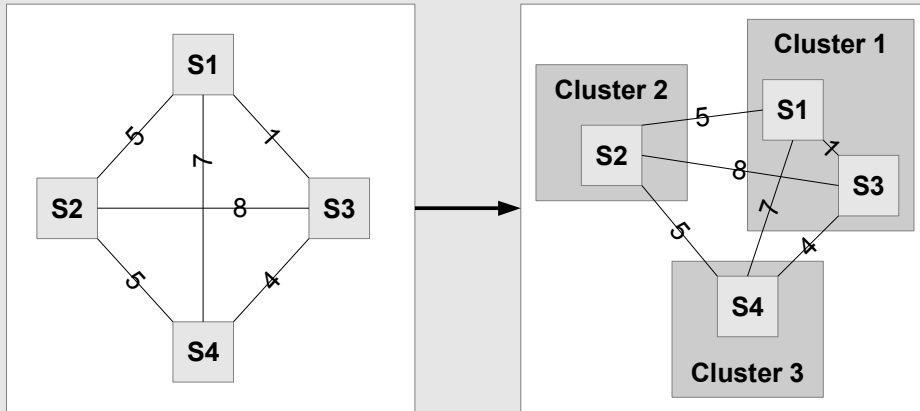
# 4. Pratique - l'étude

## 4.3. Classification des échantillons

### Clustering basé sur les distances

Liu et al (2006) Bioinformatics. Aug 15;22(16):1971-8  
*Distance-based clustering of CGH data.*

#### Distance numérique entre chaque paire d'échantillons



#### Plusieurs échelles possibles

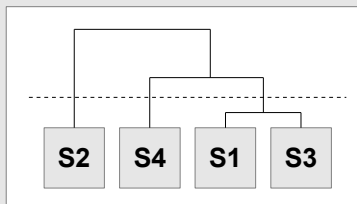
- Sondes : régions larges privilégiées
- Segments : plus stable mais plus complexe

#### Plusieurs distances possibles

- Différence de Log-ratio
- Différence de Z-Score :  $(M - \text{mean}(M)) / \text{sd}(M)$
- Différence de copies (après modélisation)
- Index de Jaccard (altéré VS non-altéré)

#### Plusieurs analyses des distances possibles

- K-means (Liu et al 2006)
- Clustering hiérarchique (nombreux algorithmes d'agrégation)



### Autres approches

#### Modèles de Markov

Shah et al (2009) Bioinformatics. Jun 15;25(12):i30-8  
*Model-based clustering of array CGH data.*

#### Analyse en Composante Principale

#### Sujet peu abordé

Approches variées  
 Ni consensus, ni études comparatives  
 Implémentations pas toujours disponibles