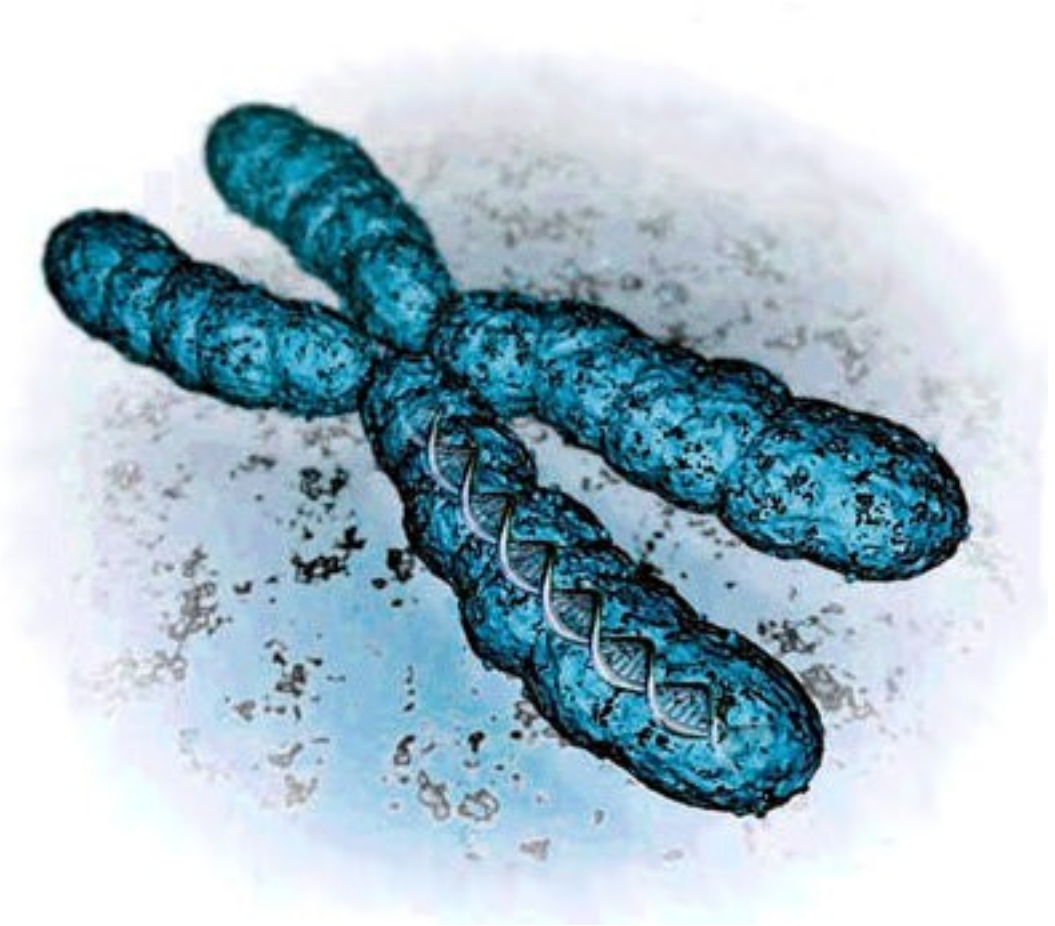
Sylvain Mareschal [@etu.univ-rouen.fr]

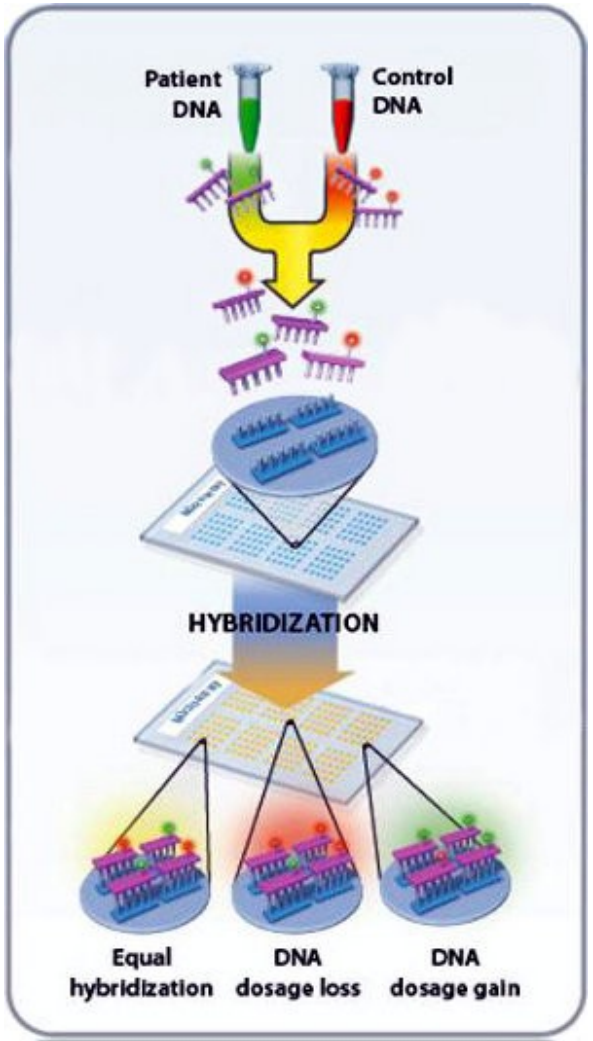
CGH-array



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



Theisen 2008

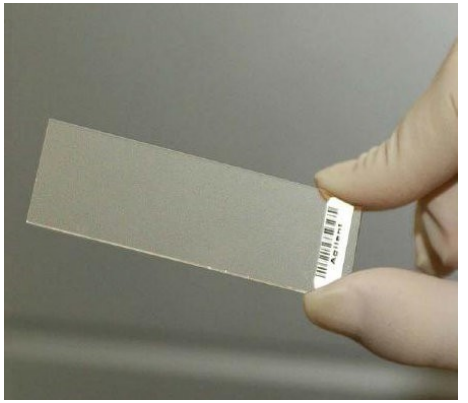
Fragmentation du génome
Enzymes de restriction
Sonication

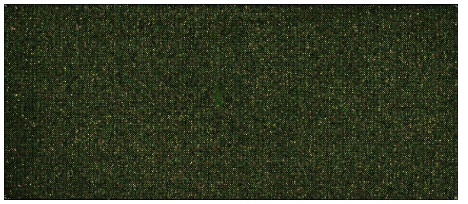
Hybridations compétitives
Un ADN cible (vert)
Un ADN référence (rouge)

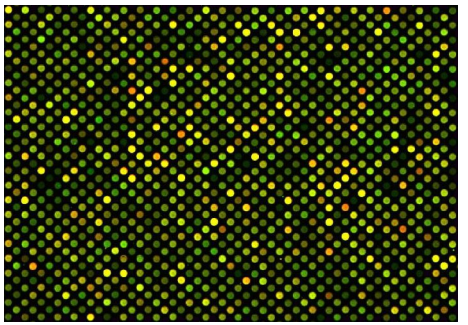
Localement sur le génome
Historiquement BAC (10^2 à 10^3 pour hg)
60-mer généralement (10^5 à 10^6 actuellement)
Réparties sur le génome ou la région cible

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

Calcul des log-ratios
$M = \log_2(F_{vert} / F_{rouge})$
$M > 0$: amplification
$M < 0$: délétion



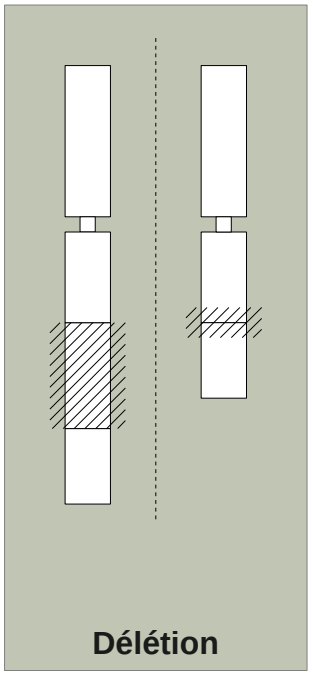




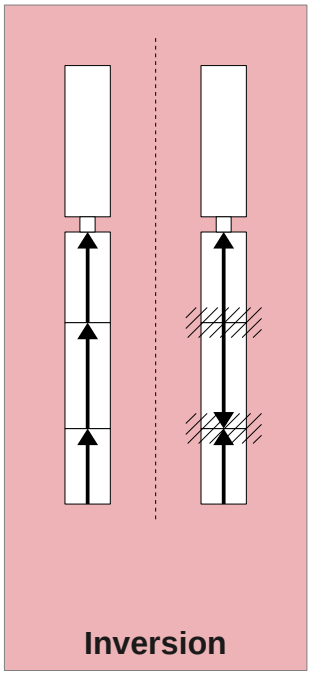
Agilent 44K
Martin Figeac

1. Théorie

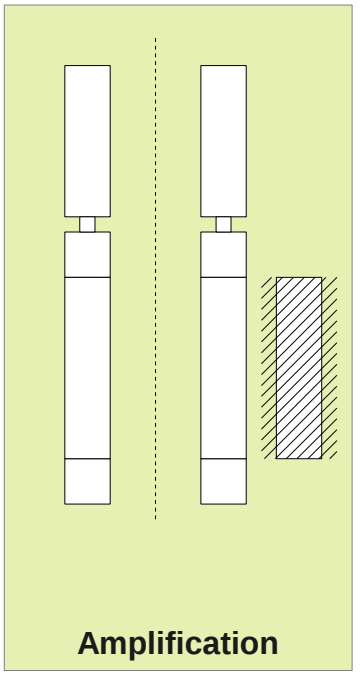
1.2. Anomalies visibles



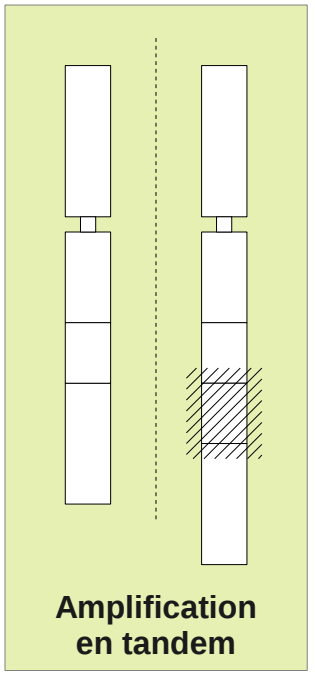
Délétion



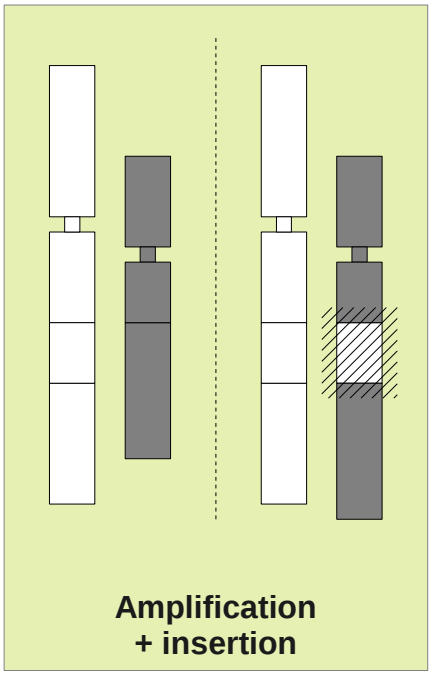
Inversion



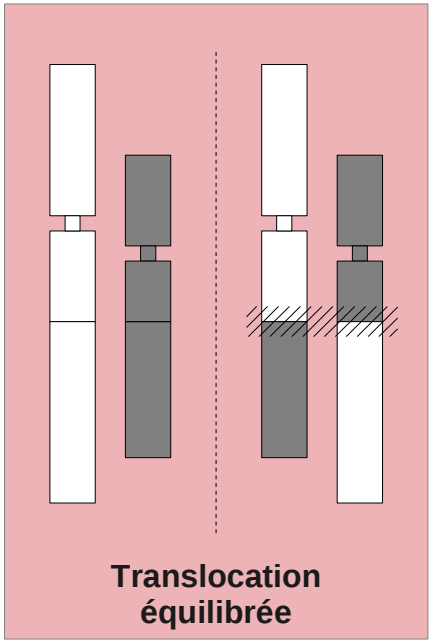
Amplification



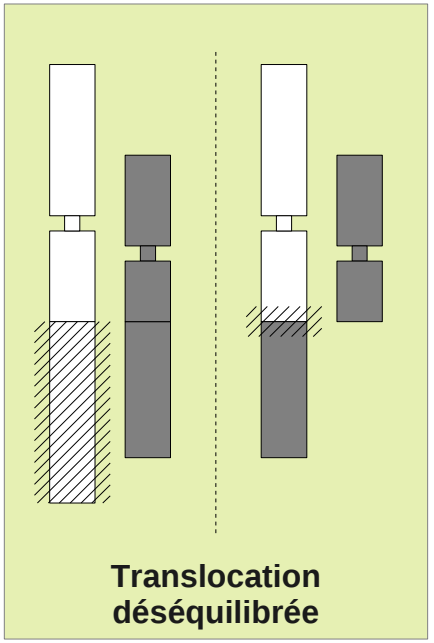
Amplification en tandem



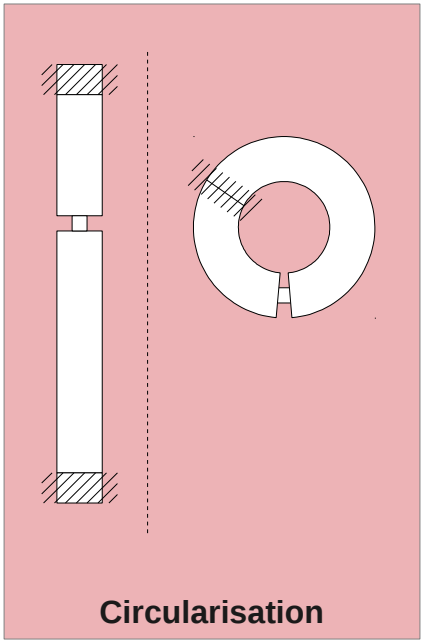
Amplification + insertion



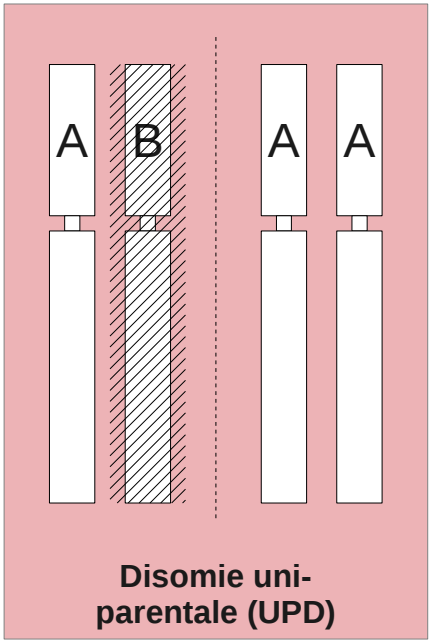
Translocation équilibrée



Translocation déséquilibrée



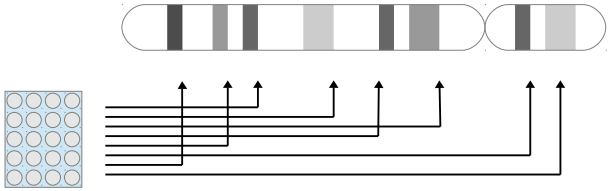
Circularisation



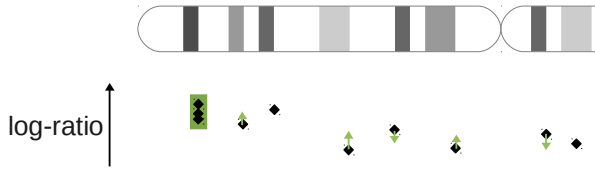
Disomie uni-parentale (UPD)

1. Théorie
1.3. Workflow

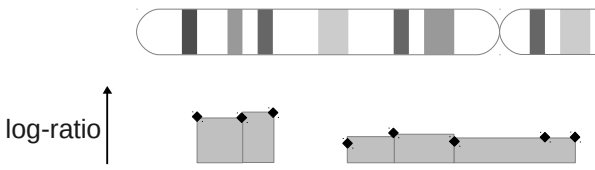
Localisation
Scan, mapping



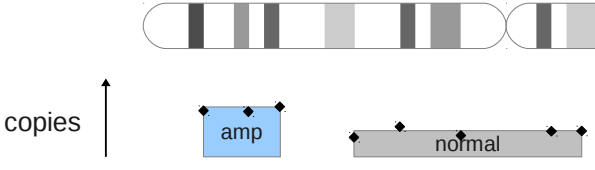
Traitement des sondes
Biais



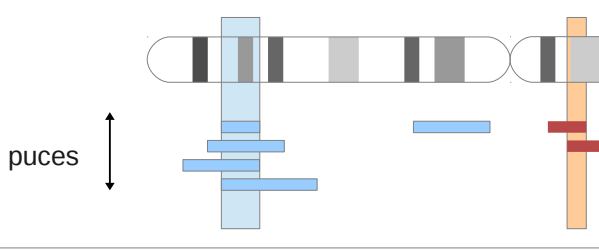
Segmentation
CBS



Copy calling
Seuils, N, HMM



Analyse globale
Pénétrance, MCR, clustering

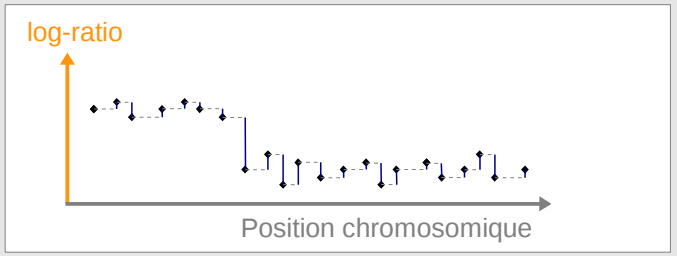


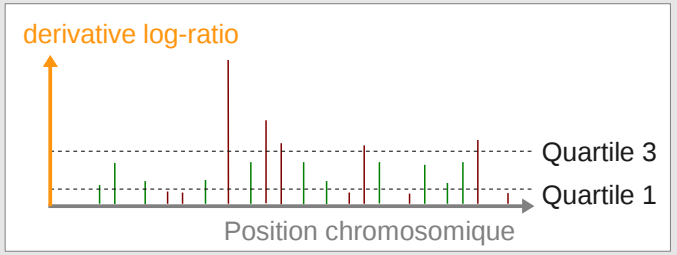
2. Pratique - les biais

2.1. Biais de sondes - puces en général

Mesure de la qualité de la puce

Derivative Log Ratio Spread (DLRS)
DLRS = IQR(diff(M)) / (1,349 * sqrt(2))

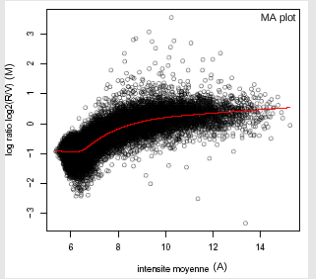




Excelent < 0,2 < Good < 0,3 < Poor
Kincaid et al (2007) US 2007/0031883 A1 application patent 2007
Analyzing CGH data to identify aberrations.

Median Absolute Deviation (MAD)
MAD = median(M – 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.

Biais d'intensité



Caroline Bérard

Centralisation
Communément admis
Médiane / Moyenne inadaptées
M' = M - mode(M)
Attention aux polyploïdes

LOWESS
$A_{amplification} > A_{normal} > A_{deletion}$
Risqué sur ADN complexes (perte d'altérations)

LOWESS sur sondes normales
Même principe, en excluant les altérations

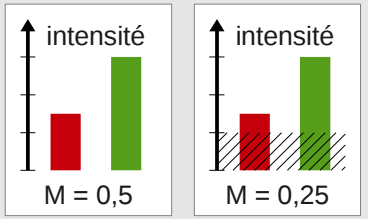
K-mean clustering
Staaf et al (2007) BMC Genomics. Oct 22;8:382.
Normalization of array-CGH data: influence of copy number imbalances.

Segmentation
Van Houte et al (2010) Bioinformatics. May 15;26(10):1366-7
CGHnormaliter: a Bioconductor package for normalization of array CGH data with many CNAs.

Densité 2D
Chen et al (2008) Bioinformatics. Aug 15;24(16):1749-56.
A probe-density-based analysis method for array CGH data: simulation, normalization and centralization.

Bruit de fond

Soustraction
Possible mais peu utilisée
Effet sur les intensités faibles



Filtration éventuelle

Normalisation inter-array

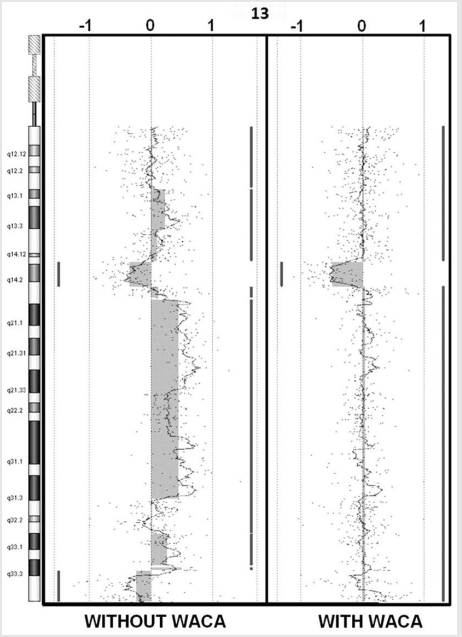
Variance, quantiles
Hypothèse de distributions similaires déraisonnable
A éviter, sauf si profils équivalents (rare)

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



Leprêtre et al 2009

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

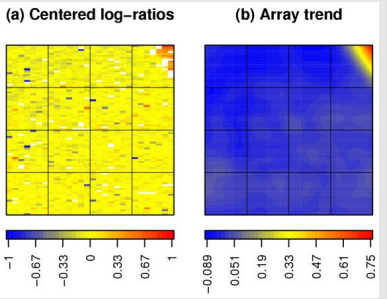
Plaques de calibration
Van de Wiel et al (2009) Bioinformatics. May 1;25(9):1099-104
Smoothing waves in array CGH tumor profiles.

Régression sur données d'hybridation normal / normal
Nécessite une série du même design
Nombreux biais corrigés

Biais spatial

CGH : positions aléatoires
Pas de lien spatial puce / chromosome
Répartition homogène attendue des M

Artefact local
Traces de doigt, de colle

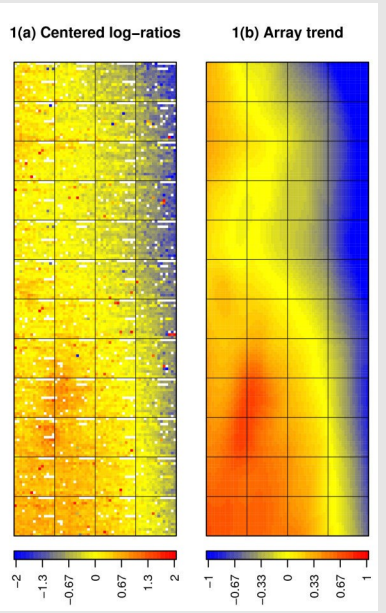


Gradient
Mélange de réactifs hétérogène
Lames mal immergées

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)



Neuvial et al 2006

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

Bioconductor 'MANOR'
Neuvial et al (2006) BMC
Bioinformatics. May 22;7:264.
Spatial normalization of array-CGH data.

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
$M_{observé}$ = moyenne($M_{chaque\ cellule}$)
Effet variable selon les proportions

Cellules normales
Non concernées par les altérations (2 copies)
Log-ratios tendent vers 0
Exemple d'une délétion :

0% normales	log2(1/2) = -1
10 % normales	log2((1/2)*90% + (2/2)*10%) = -0,86
20 % normales	log2((1/2)*80% + (2/2)*20%) = -0,74
50 % normales	log2((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 anapathologique des proportions
$M_{corrigé} = \log_2((2^M - 1 + p_{tumorale}) / p_{tumorale})$

Éviter les sous-populations
- Micro-dissection
- *Single-cell* (micromanipulation, *Whole Genome Amplification*)

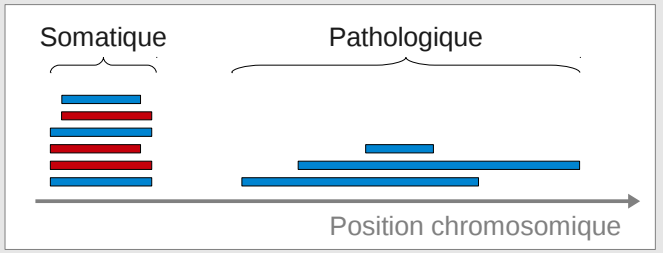
Hannemann et al (2011) PLoS One. 6(11):e26362
Quantitative high-resolution genomic analysis of single cancer cells.

Anomalies constitutionnelles

Attention à l'ADN de référence
Pool d'ADN : 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.

Détection 1 : à l'échelle de la série
Régions généralement courtes (< 100 kb)
Très fort recouvrement d'une puce à l'autre
Amplifié et délété très fréquemment

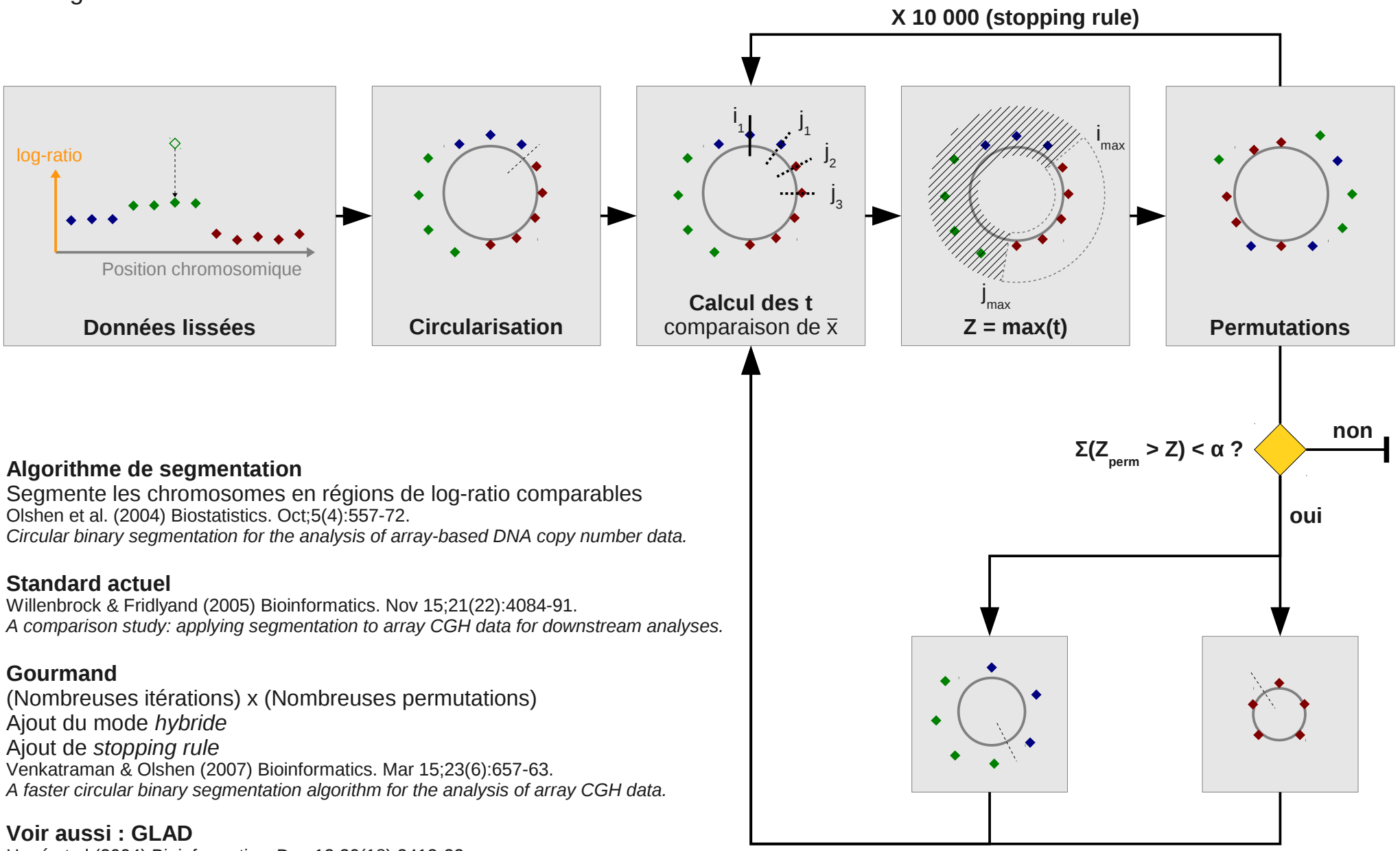


Détection 2 : DGV
Database of Genomic Variants (projects.tcag.ca/variation)
Répertoire de CNV constitutionnels
Données très hétérogènes (CGH-BAC, NGS …)
Recouvrement jamais parfait

Solution : utiliser l'ADN normal correspondant
Normal / cible préférable lorsque possible (cancer)
Peut pénaliser la modélisation (moins d'altérations)

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.

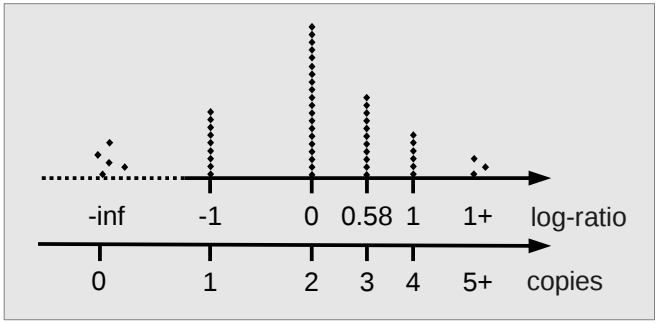
Voir aussi : GLAD
Hupé et al (2004) Bioinformatics. Dec 12;20(18):3413-22
Analysis of array CGH data: from signal ratio to gain and loss of DNA regions.

3. Pratique - la puce

3.2. Calling - au plus simple

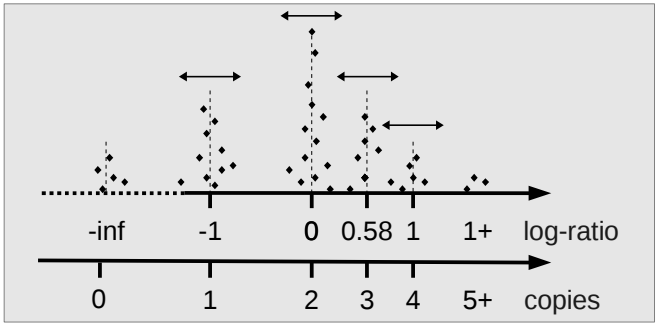
Problème trivial en théorie

Délétion homozygote	$M_{do} = \log_2(0/2) \rightarrow -inf$
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 ADN, manipulation, mesure ...



Utilisation de seuils

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

Choix délicat

Seuils unitaires

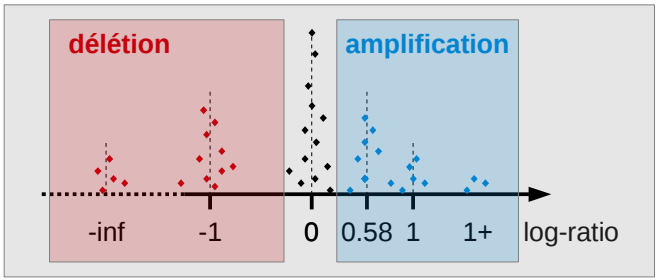
Seuils : 0,8 < ratio < 1,2 → -0,32 < M < 0,26

Basé sur l'observation d'hybridations normal / normal

Généralisation très risquée

Weiss et al (2003) J Pathol. Jul;200(3):320-6.

Determination of amplicon boundaries at 20q13.2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization.



Seuils quantiles

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

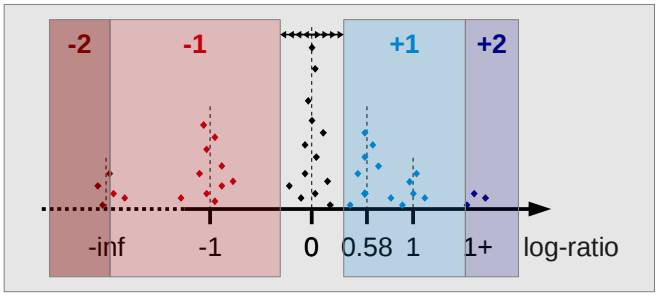
Seuils +/- 1 copie : mean(M_{normal}) +/- 4*sd(M_{normal})

Seuils +/- 2 copies : 3^{e} et 97^{e} percentiles

Hypothèse forte sur la distribution des log-ratios

Aguirre et al. (2004) PNAS Jun 15;101(24):9067-72.

High-resolution characterization of the pancreatic adenocarcinoma genome.



Seuils assistés

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

Corrélation entre techniques pas toujours évidente

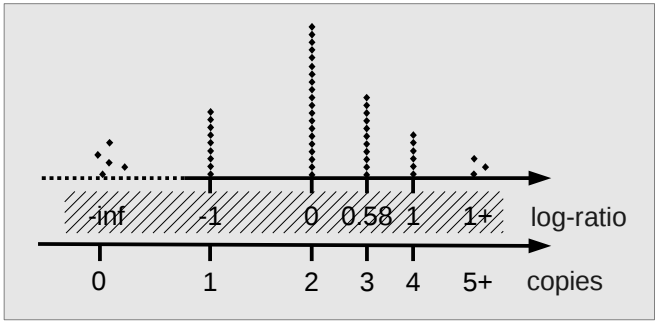
Coût supplémentaire

3. Pratique - la puce

3.3. Calling - modèles de mélange gaussien

Problème trivial en théorie

Délétion homozygote	$M_{do} = \log_2(0/2) \rightarrow -inf$
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	$M_{dt} = \log_2(4/2) = 1$

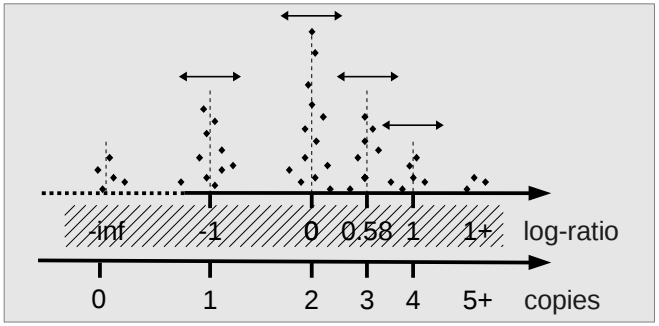


Nombreux biais en pratique

Qualité ADN, manipulation, mesure ...

Modélisation gaussienne

Variation à ce modèle simple : bruit de fond

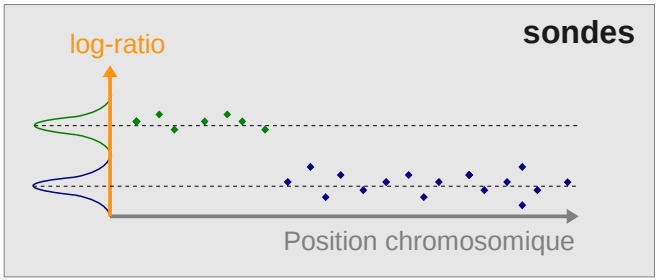


Application directe aux sondes

Estimation difficile, peu utilisé

Picard et al (2005) BMC Bioinformatics. Feb 11;6:27

A statistical approach for array CGH data analysis.



Application aux segments : CGHcall

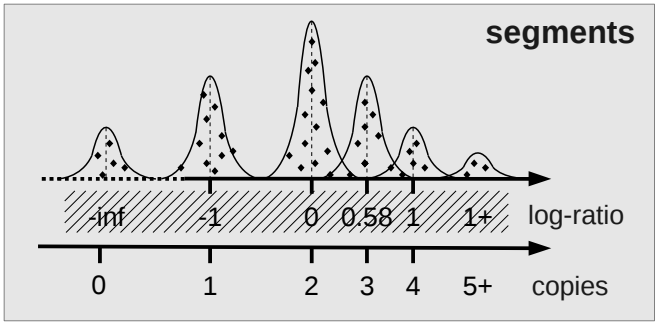
Van de Wiel et al (2007) Bioinformatics. Apr 1;23(7):892-4.

CGHcall: calling aberrations for array CGH tumor profiles.

6 états autorisés

$M_{probe,segment,sample} \sim N(\mu_{segment,sample}, \sigma^2)$

$\mu_{segment,sample} \sim \Sigma(p_{etat} * N(\mu_{etat}, \sigma^2_{etat}))$



Moyennes exactes inconnues

Biais de cellularité, fluorochrome ...

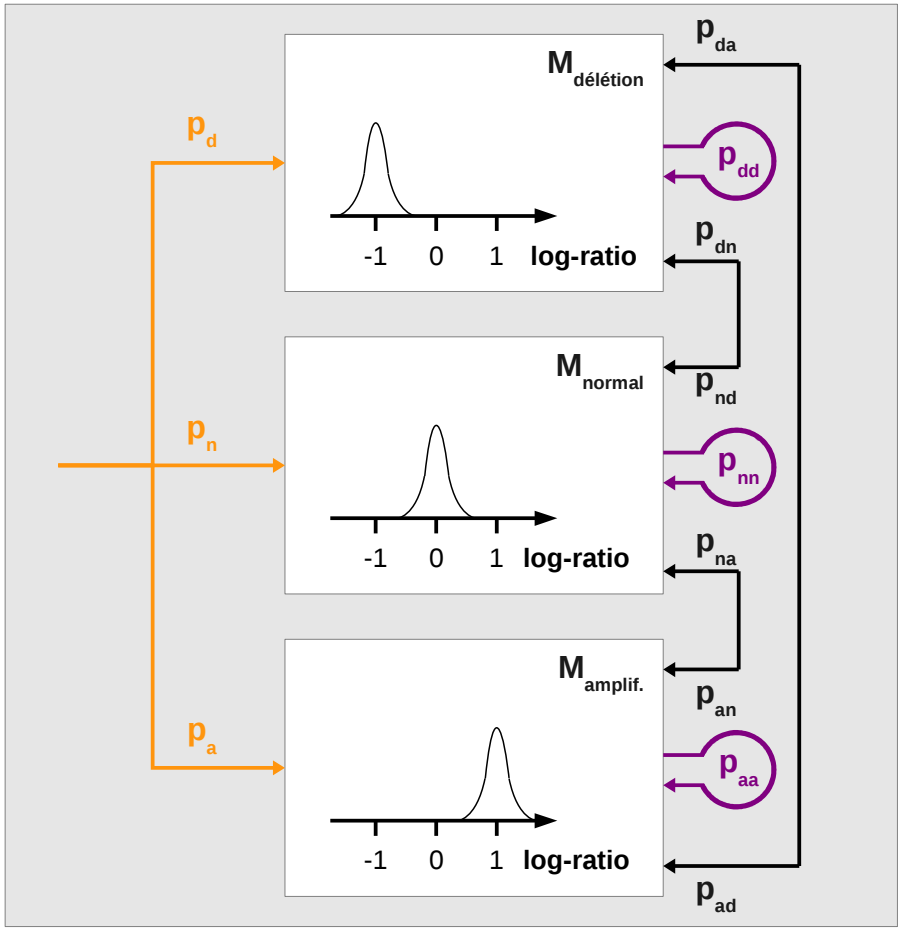
Estimation des paramètres

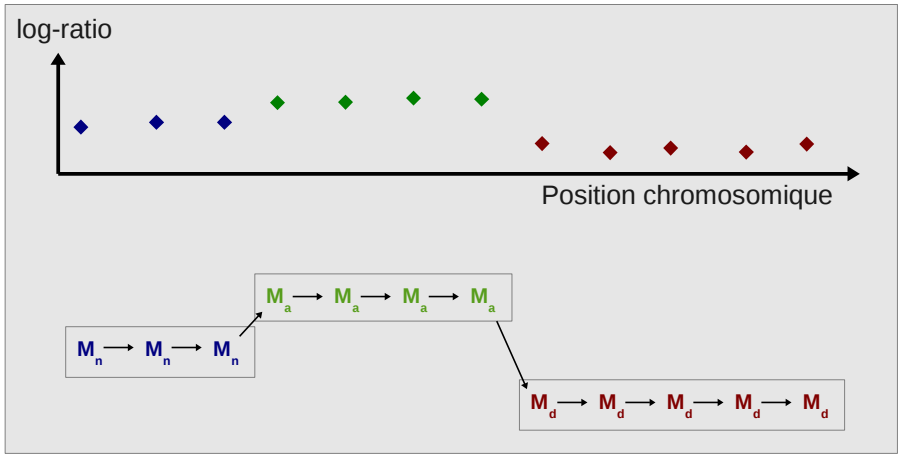
Expectation-Maximization

Modèle commun à la série (!!!) → cghRA

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

Chromosome : suite d'états
Le plus probable : même état que la sonde précédente
Possibilité de changer d'état

	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.

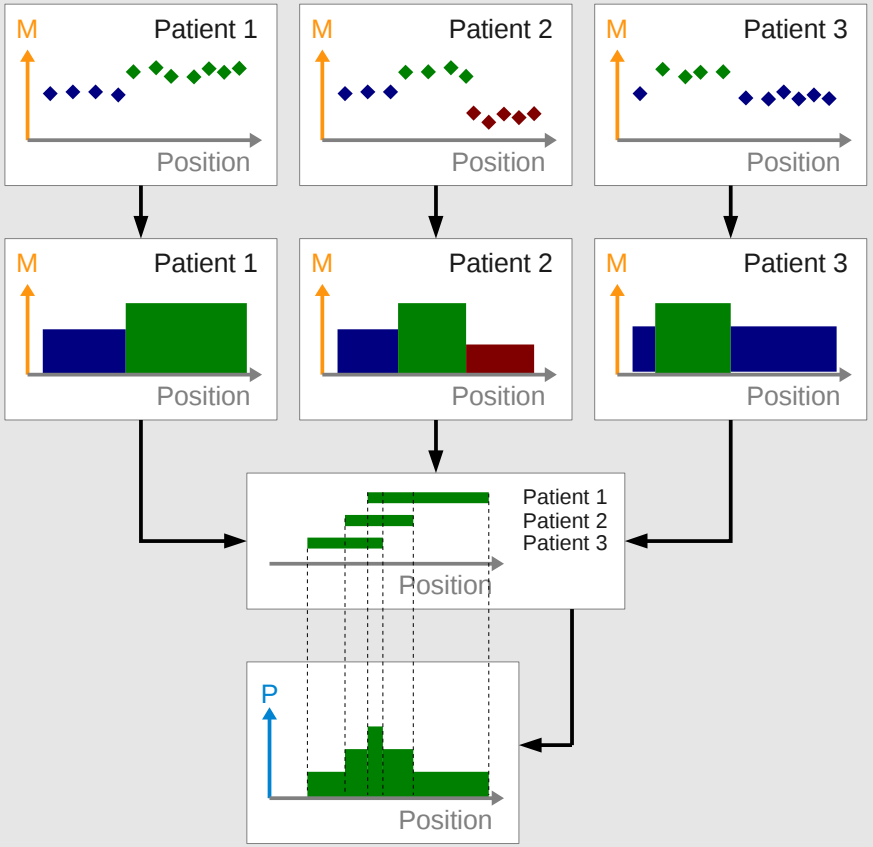
Bioconductor 'snapCGH' (BioHMM)
Prise en compte des distances inter-sondes
Marioni et al (2006) May 1;22(9):1144-6
BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data.

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 ...)
$P_{délétion,point} = n_{patients\ délétés,point} / n_{patients\ étudiés}$

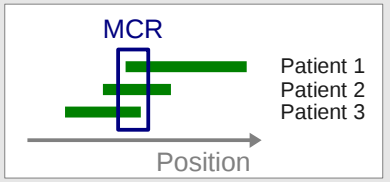


Met en évidence de larges régions d'intérêt
Bras ou chromosomes entiers
Seuil subjectif

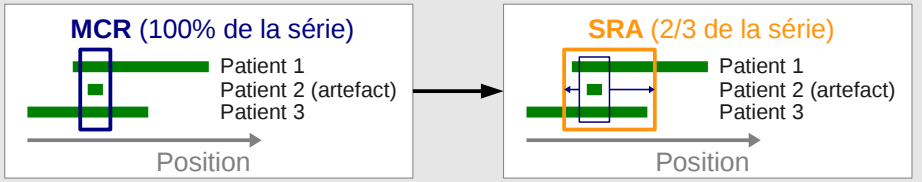
Minimal Common Regions

Région communément altérée dans la série
Recouvrement entre les CNV de puces différentes
Pas de définition consensuelle

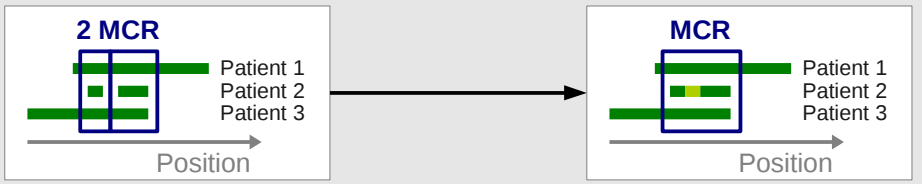
Littéralement : région commune minimale



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.



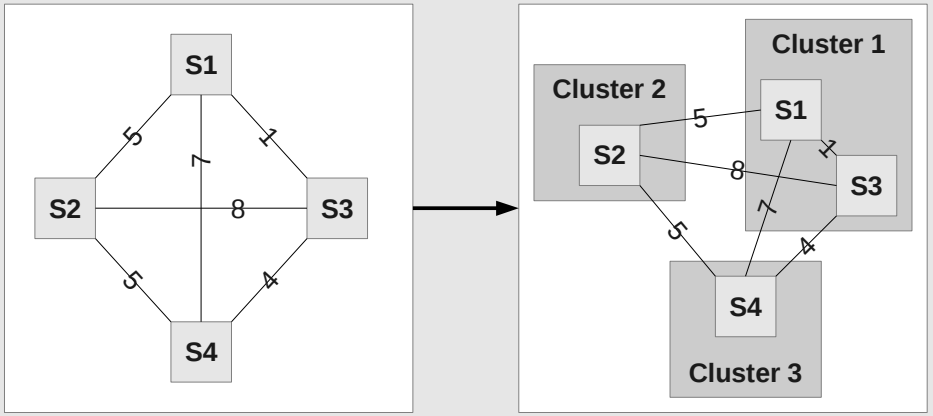
4. Pratique - l'étude

4.2. 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.

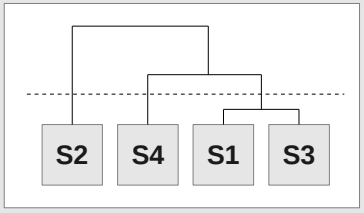
Distance numérique entre chaque paire d'échantillons



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 – mean(M)) / sd(M)
- Différence de copies (après modélisation)

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