



Analyse de données de nombre de copies d'ADN

Segmentation de données génomiques

Pierre Neuvial

Centre National de la Recherche Scientifique
Institut de Mathématiques de Toulouse
Équipe Statistique et Optimisation

Université Paul Sabatier, UMR CNRS 5219
<http://math.univ-toulouse.fr/~pneuvial>

ENSAI 3A-SV – 2021-2022

Analyse de données de nombre de copies d'ADN

- 2 Données de nombres de copies d'ADN en cancérologie
- 3 Extraction de l'information biologique
- 4 Modèle statistique pour la segmentation
- 5 Heuristiques pour la segmentation
- 6 Application aux données de puces SNP

Analyse de données de nombre de copies d'ADN

2 Données de nombres de copies d'ADN en cancérologie

- Changements de nombre de copies d'ADN dans les cancers
- Exemple des données de puces SNP

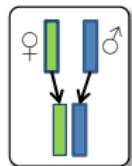
3 Extraction de l'information biologique

4 Modèle statistique pour la segmentation

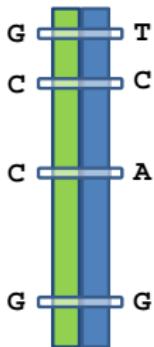
5 Heuristiques pour la segmentation

6 Application aux données de puces SNP

Génotypes dans un chromosome diploïde



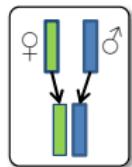
Single nucleotide polymorphism



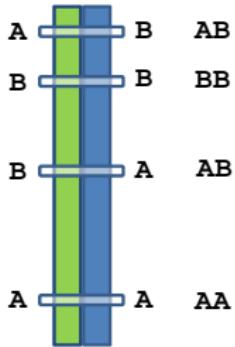
10-20 million
known SNPs

slide: H. Bengtsson.

Génotypes dans un chromosome diploïde



Single nucleotide polymorphism

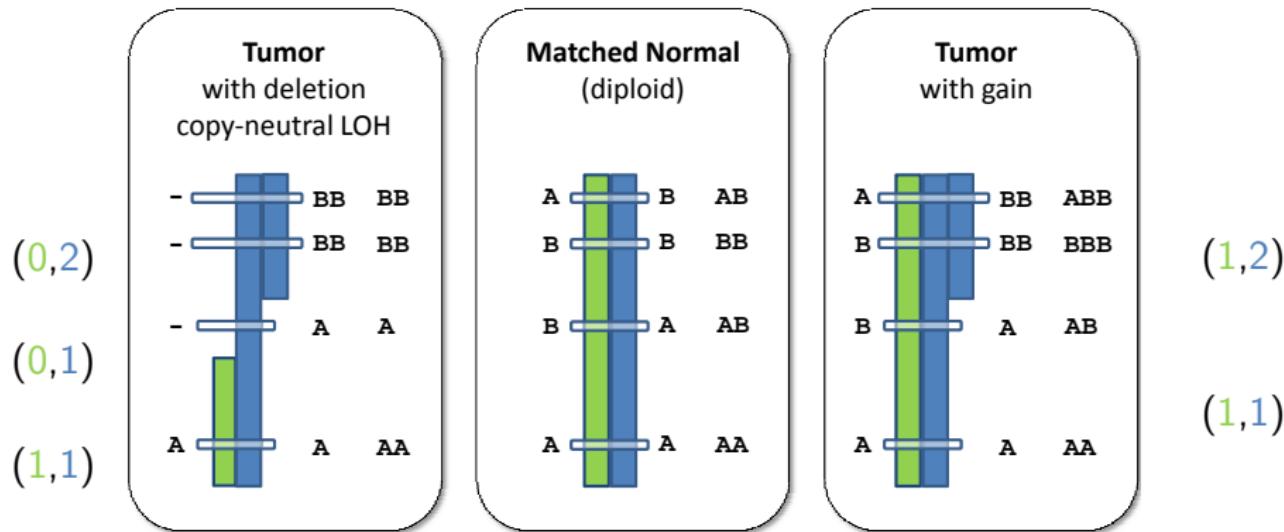


10-20 million
known SNPs

Only two possible letters at a given position !

slide: H. Bengtsson.

Genotypes et nombre de copies d'ADN : aneuploïdie



slide: H. Bengtsson.

Nombre de copies parentaux, majeur et mineur

Nombres de copies parentaux au locus j : (m_j, p_j) : nombre **non-observé** de copies provenant de la mère et du père en j .

Etat du nombre de copies en j

$$CN = (\textcolor{brown}{C}_{1j}, \textcolor{blue}{C}_{2j}),$$

où $\textcolor{brown}{C}_{1j} = \min(m_j, p_j)$ et $\textcolor{blue}{C}_{2j} = \max(m_j, p_j)$.

Les nombres de copies mineur ($\textcolor{brown}{C}_1$) et majeur ($\textcolor{blue}{C}_2$) :

- caractérisent les altérations d'intérêt dans les cancers
- peuvent être estimés à l'aide des données de puces SNP

Analyse de données de nombre de copies d'ADN

2 Données de nombres de copies d'ADN en cancérologie

- Changements de nombre de copies d'ADN dans les cancers
- Exemple des données de puces SNP

3 Extraction de l'information biologique

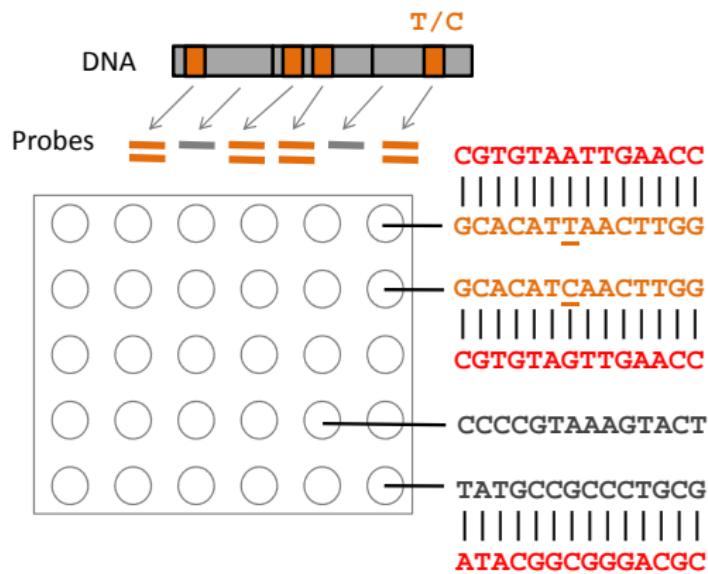
4 Modèle statistique pour la segmentation

5 Heuristiques pour la segmentation

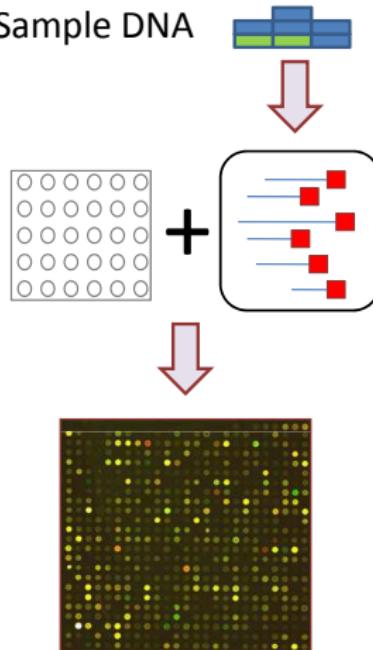
6 Application aux données de puces SNP

Technologie : puces SNP et nombre de copies d'ADN

Chip Design



Sample DNA



slide: H. Bengtsson.

(C_1, C_2) peut être estimé à partir de données SNP

Pour le **SNP** j dans l'échantillon i , les signaux observés sont résumés par (θ, β) , où $\theta_{ij} = \theta_{ijA} + \theta_{ijB}$ et $\beta_{ij} = \theta_{ijB}/\theta_{ij}$.

Nombre de copies total

$$\begin{aligned} C_{ij} &= 2 \frac{\theta_{ij}}{\theta_{Rj}} \\ &= \textcolor{blue}{C_{1ij}} + \textcolor{blue}{C_{2ij}} \end{aligned}$$

Diminution d'hétérozygotie

$$\begin{aligned} DH_{ij} &= 2 |\beta_{ij} - 1/2| \\ &= \frac{\textcolor{blue}{C_{2ij}} - \textcolor{blue}{C_{1ij}}}{\textcolor{blue}{C_{2ij}} + \textcolor{blue}{C_{1ij}}} \end{aligned}$$

Notes :

- DH n'est défini que pour les SNP **hétérozygotes dans la lignée germinale**
- Les deux dimensions sont nécessaires à l'interprétation :
 - ▶ Isodisomie (Copy neutral LOH) : $CN = (\textcolor{blue}{0}, \textcolor{blue}{2})$: deux copies au total
 - ▶ Duplication équilibrée : $CN = (\textcolor{blue}{2}, \textcolor{blue}{2})$, ratio allélique normal

The Cancer Genome Atlas (TCGA)

“Accelerate our understanding of the molecular basis of cancer”

- 20 types de cancers, dont cerveau (glioblastoma multiforme), ovaire, sein, poumon, leucémies.
- Études à grande échelle : 500 paires tumeur/normal pour chaque type de cancer
- Niveaux d'étude : nombre de copies d'ADN, expression des gènes et petits ARNs, méthylation de l'ADN
- Plate-formes : puces à ADN et grand séquençage

Pour les données de puces SNP, : identifier les **changements de nombre de copies d'ADN** : (C, DH) ou (C_1, C_2) :

- ① **détection** : trouver les regions
- ② **classification** étiqueter les regions

Ici : illustration sur des données de cancer de l'ovaire.

Analyse de données de nombre de copies d'ADN

2 Données de nombres de copies d'ADN en cancérologie

- Changements de nombre de copies d'ADN dans les cancers
- Exemple des données de puces SNP

3 Extraction de l'information biologique

- Pre-processing : des signaux comparables entre échantillons
- Post-processing : nombre de copies totaux
- Post-processing : ratios alléliques

4 Modèle statistique pour la segmentation

- Limites des approches directes
- Modèles de rupture
- Solution exacte par programmation dynamique

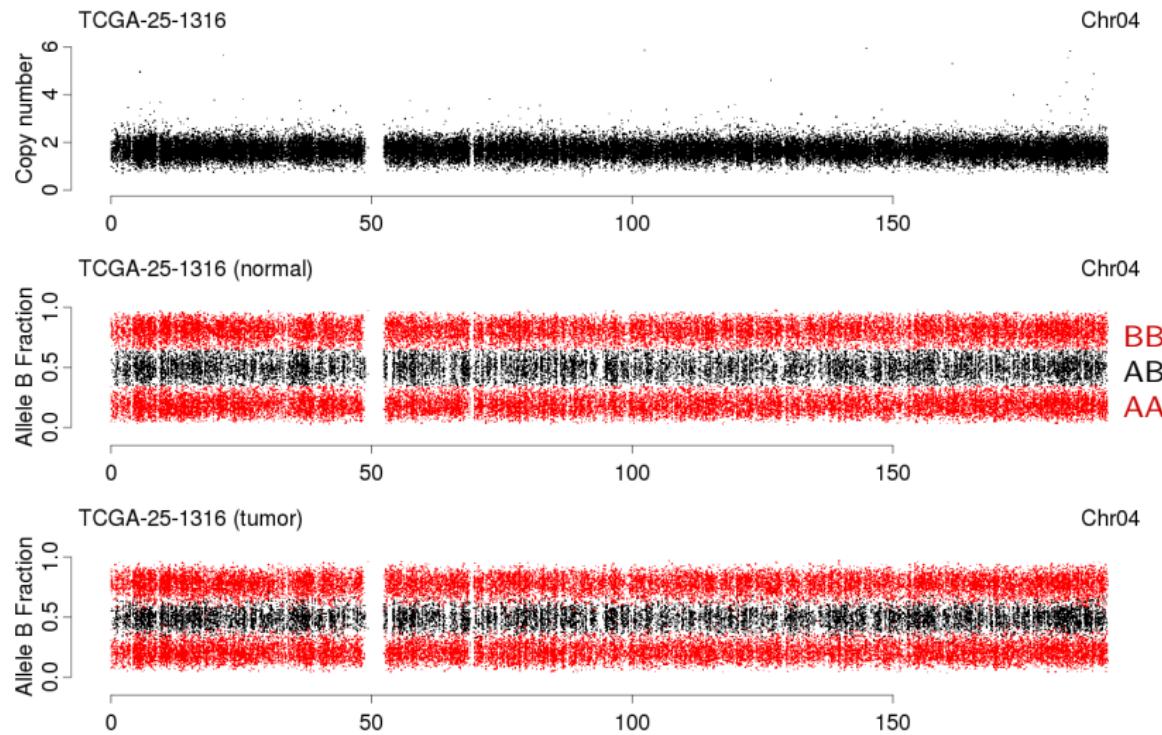
5 Heuristiques pour la segmentation

- Segmentation binaire récursive (circulaire)
- Relaxation convexe : fused lasso

6 Application aux données de puces SNP

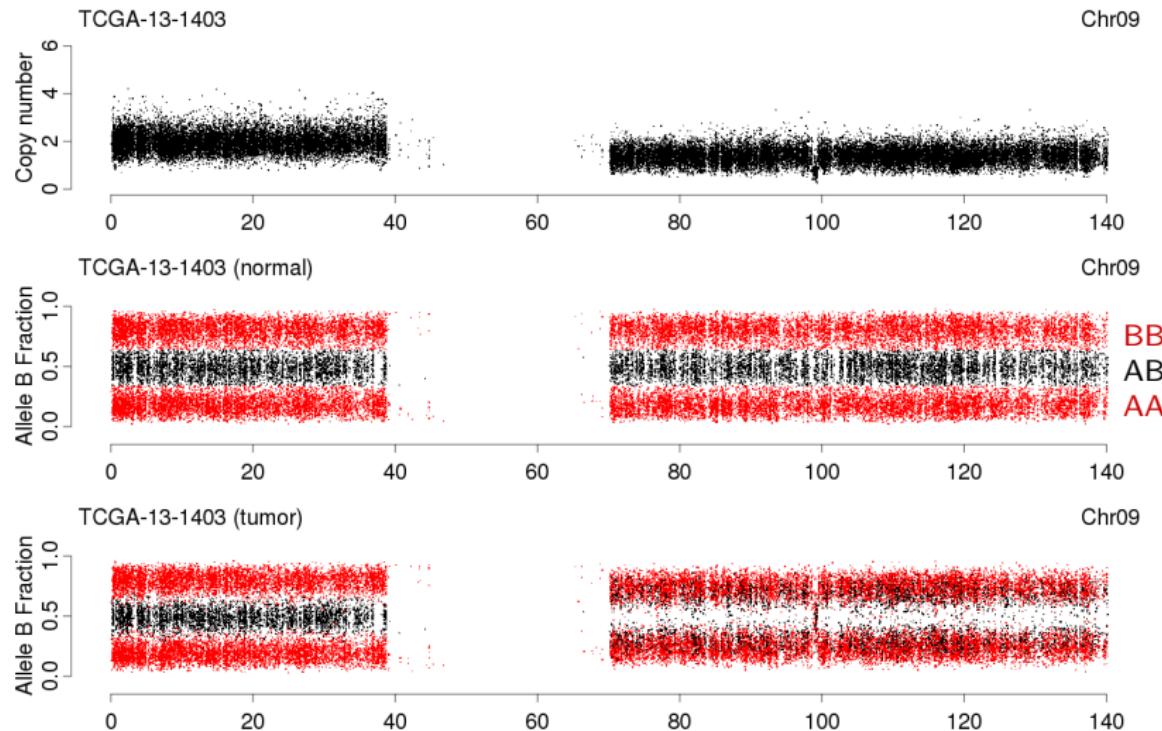
- Extension au problème de segmentation conjointe
- Construction de jeux de données à réponse connue

Région normale : (1,1)



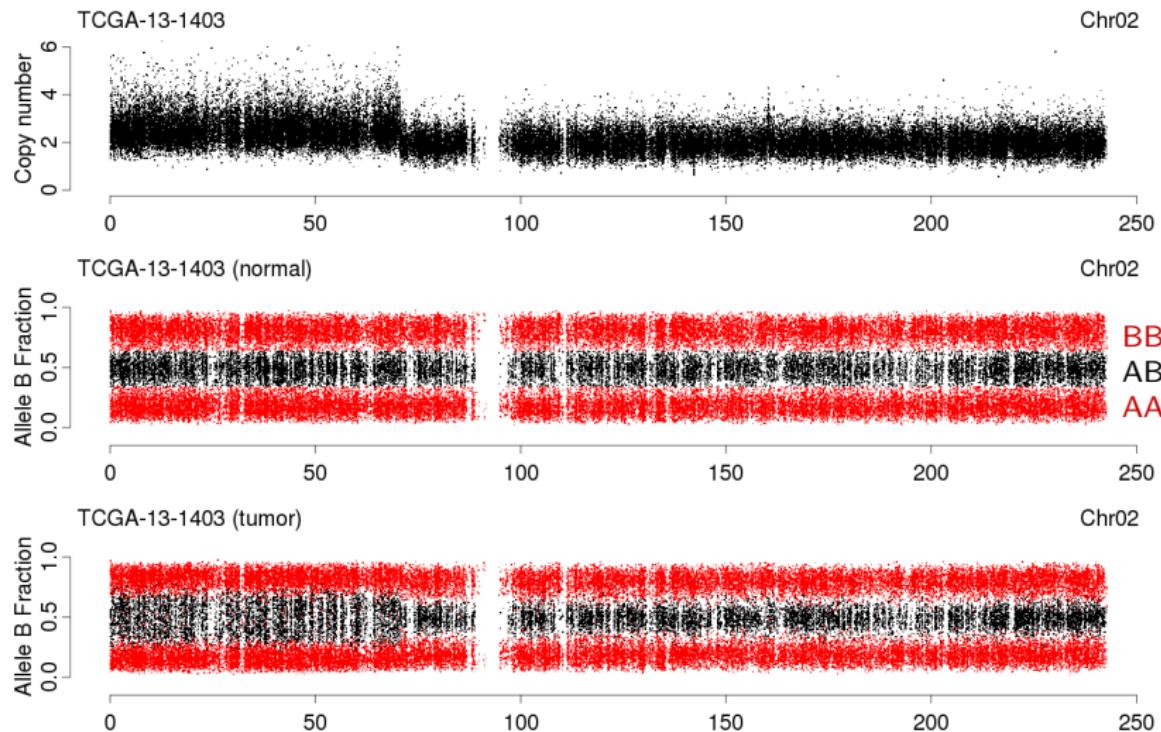
Les SNPs homozygotes dans l'échantillon normal sont en rouge.

Perte d'une copie : (0, 1)



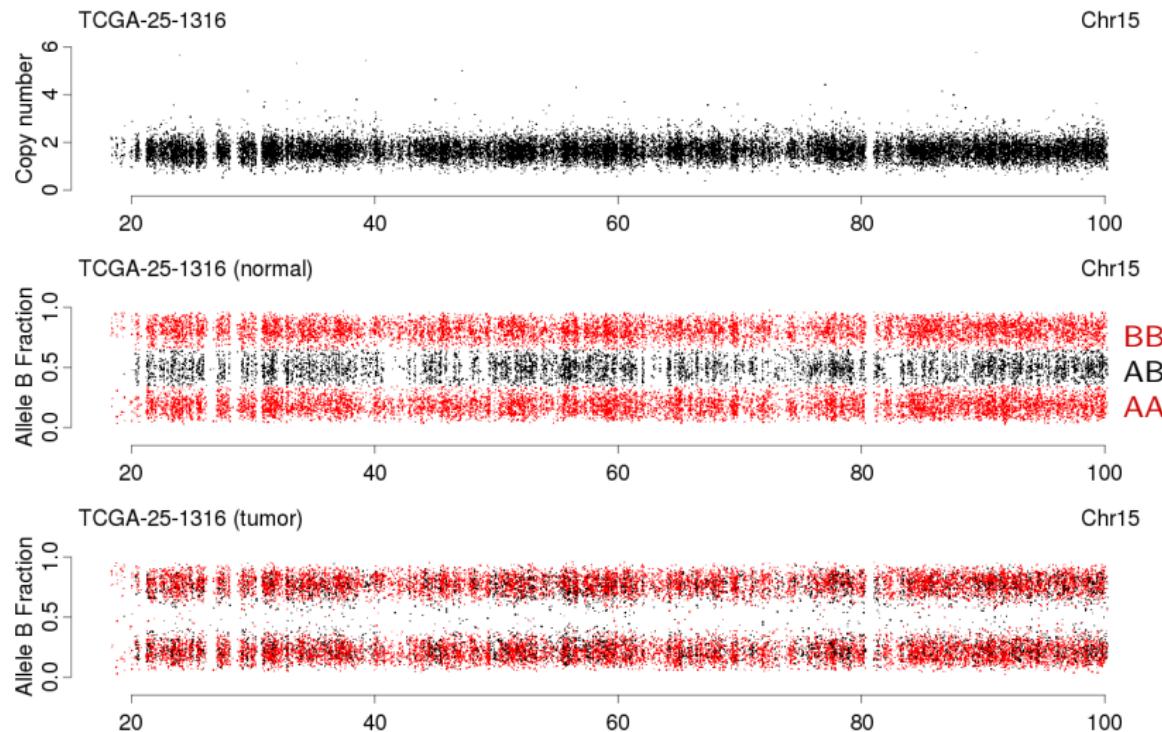
Les SNPs homozygotes dans l'échantillon normal sont en rouge.

Gain d'une copie : (1, 2)



Les SNPs homozygotes dans l'échantillon normal sont en rouge.

Isodisomie : (0, 2)



Les SNPs homozygotes dans l'échantillon normal sont en rouge.

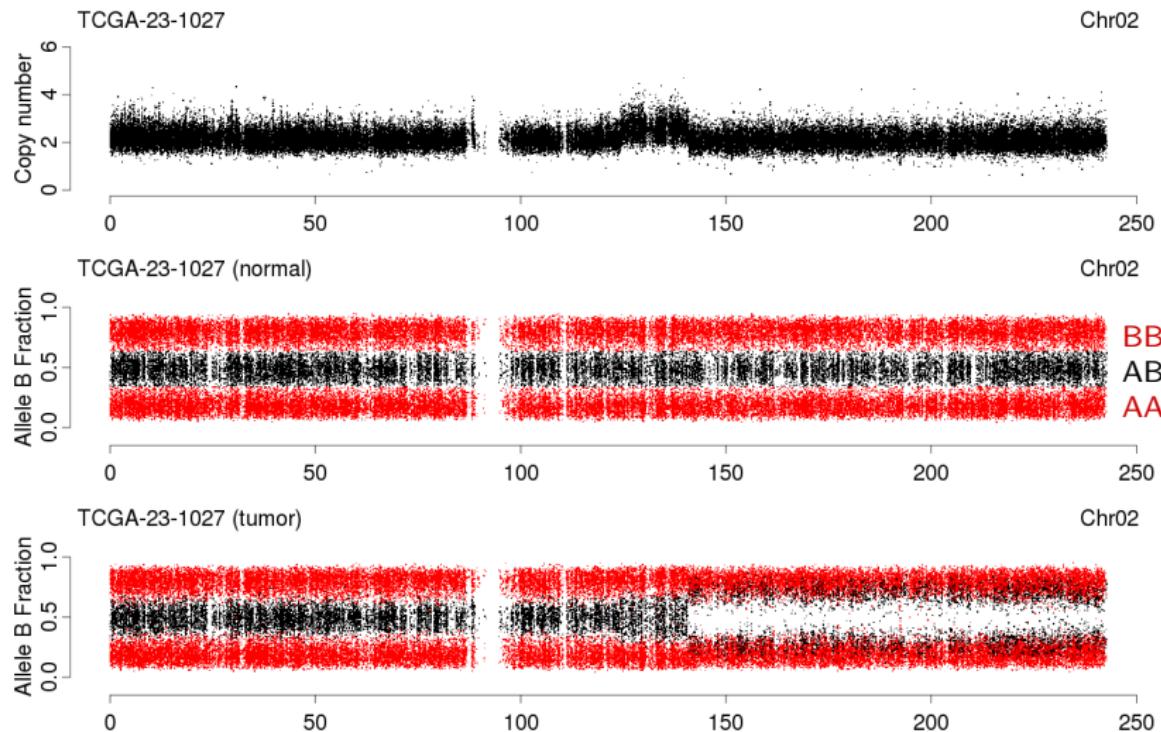
Présence de cellules normales

En pratique, les “échantillons tumoraux” contiennent en fait un **mélange de cellules tumorales et normales.**

Les exemples ci-dessus sont ceux pour lesquels la proportion de cellules tumorales est la plus élevée de tout le jeu de données.

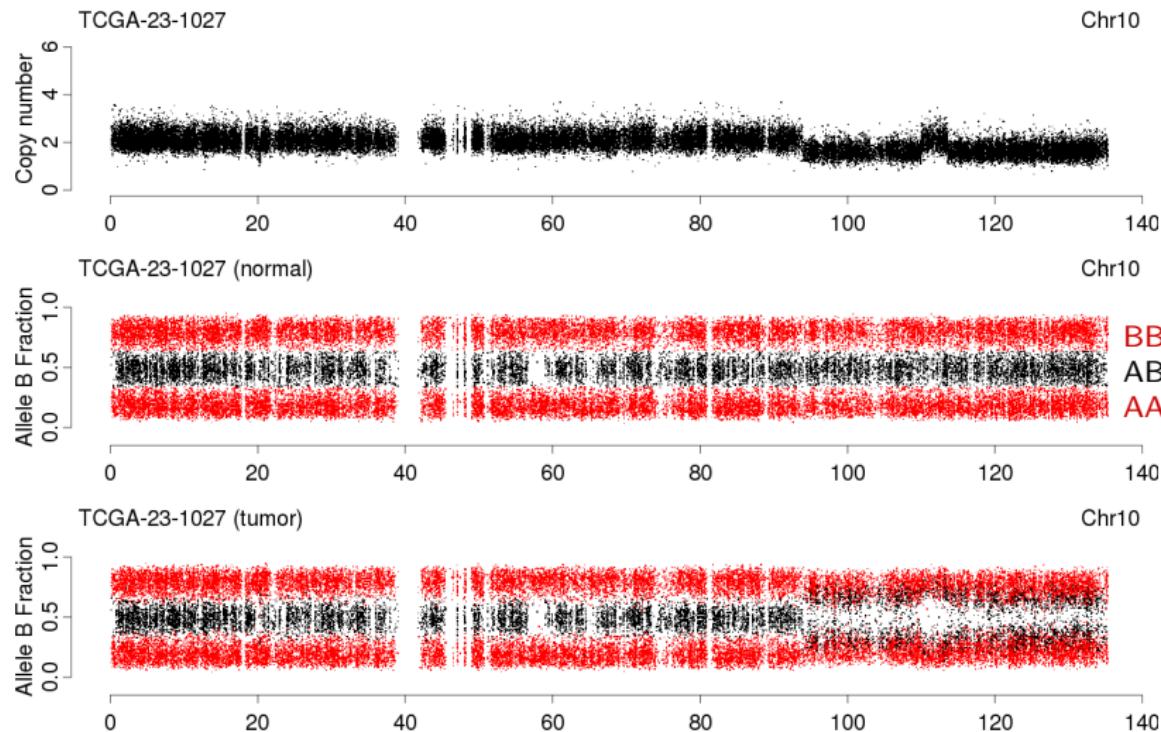
En présence de cellules normales, les ratios alléliques des SNPs hétérozygotes **se rapprochent de 1/2.**

Normal, gain, isodisomie



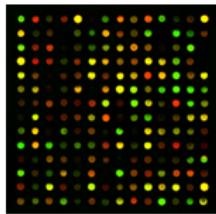
Les SNPs homozygotes dans l'échantillon normal sont en rouge.

Normal, perte, isodisomie

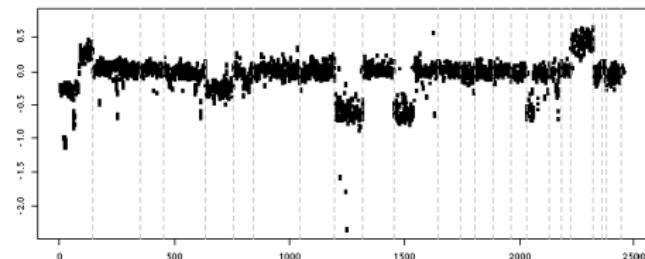


Les SNPs homozygotes dans l'échantillon normal sont en rouge.

Normalisation : définition et objectifs



Analyse d'image
⇒
Normalisation



Motivation : grande variabilité expérimentale

- faible reproductibilité des expériences
- autant de sources de biais que d'étapes expérimentales

Objectif : augmenter le rapport signal sur bruit

- distinguer variabilité biologique et artefacts expérimentaux
- rendre les données de plusieurs expériences comparables

Analyse de données de nombre de copies d'ADN

2 Données de nombres de copies d'ADN en cancérologie

3 Extraction de l'information biologique

- Pre-processing : des signaux comparables entre échantillons
- Post-processing : nombre de copies totaux
- Post-processing : ratios alléliques

4 Modèle statistique pour la segmentation

5 Heuristiques pour la segmentation

6 Application aux données de puces SNP

Copy-numbers by Robust Microarray Analysis (CRMA)

Une méthode de pre-processing applicable échantillon par échantillon

For each Affymetrix array ($i = 1, 2, 3, \dots, 10000$) independently:

<i>Calibrating & normalizing for hybridization artifacts</i>	1. Offset and Allelic crosstalk calibration 2. Probe-sequence normalization
<i>Summarization of technical replicates</i>	1. CN loci have one probe 2. Robust averaging of replicated SNPs probes
<i>Normalizing for assay artifacts</i>	1. PCR fragment-length normalization 2. GC-content normalization
<i>Total and Allele-specific copy numbers</i>	$(C_A, C_B), C = C_A + C_B$

slide: H. Bengtsson.

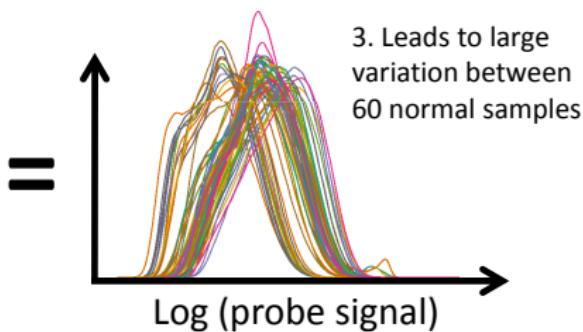
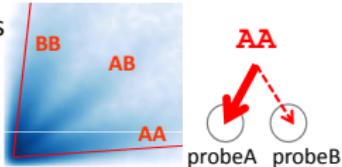
Explication de la variation systématique entre puces

Décalage (offset) du scanner et hybridation croisée entre allèles

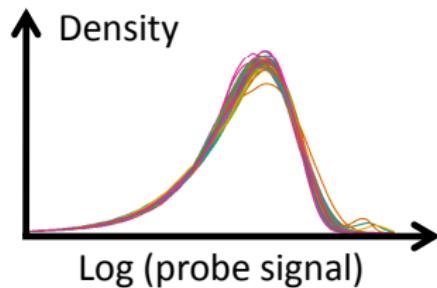
1. The scanner's shifts all probe signals (offset)



2. Cross-hybridization causes signal to leak between allele A and allele B



4. Calibration for both removes a majority of artifacts between samples



slide: H. Bengtsson.

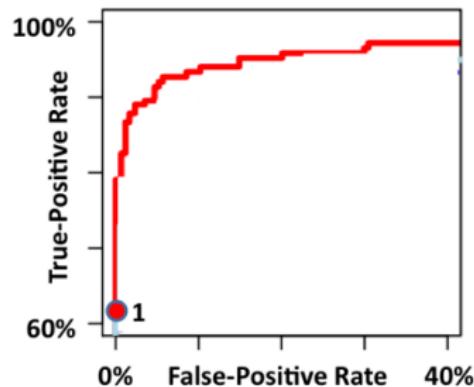
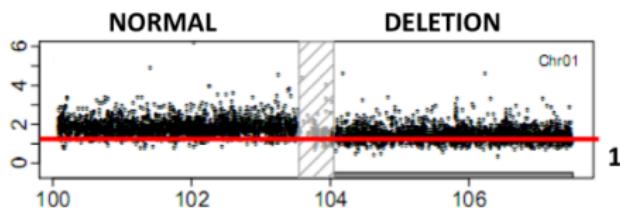
Évaluation ROC

Pour un échantillon donné

- ➊ identifier une rupture claire
- ➋ étiqueter les régions voisines, par ex. NORMAL (1,1) et DELETION (0,1)
- ➌ choisir un état de référence (NORMAL) et un état à identifier (DELETION)

Pour chaque valeur du seuil τ :

- Un SNP sous le seuil τ est appelé DELETION
- Le nombre d'erreurs définit un point sur la courbe ROC



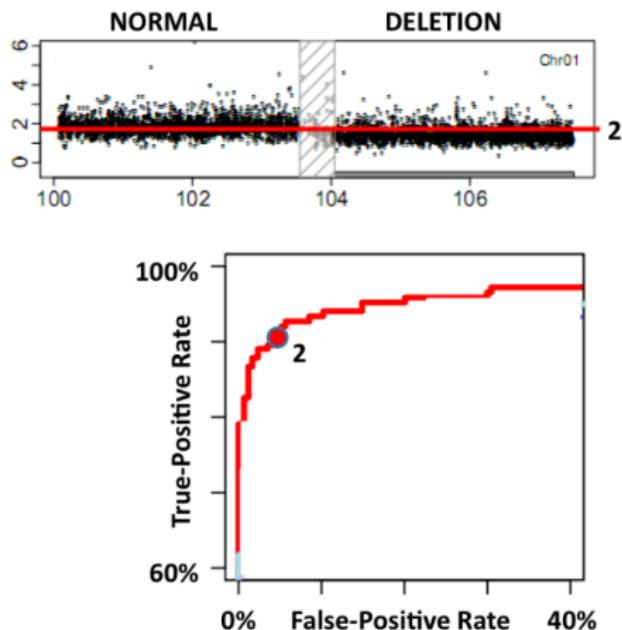
Évaluation ROC

Pour un échantillon donné

- ➊ identifier une rupture claire
- ➋ étiqueter les régions voisines, par ex. NORMAL (1,1) et DELETION (0,1)
- ➌ choisir un état de référence (NORMAL) et un état à identifier (DELETION)

Pour chaque valeur du seuil τ :

- Un SNP sous le seuil τ est appelé DELETION
- Le nombre d'erreurs définit un point sur la courbe ROC



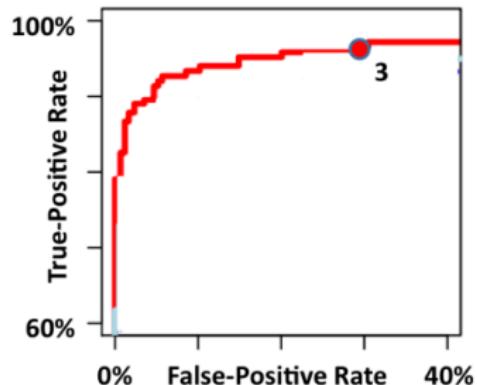
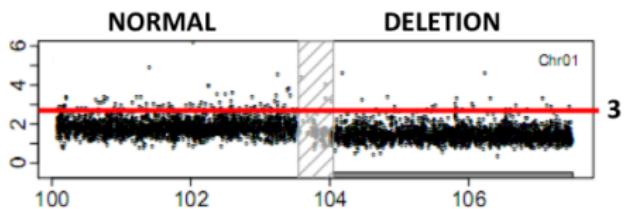
Évaluation ROC

Pour un échantillon donné

- ➊ identifier une rupture claire
- ➋ étiqueter les régions voisines, par ex. NORMAL (1,1) et DELETION (0,1)
- ➌ choisir un état de référence (NORMAL) et un état à identifier (DELETION)

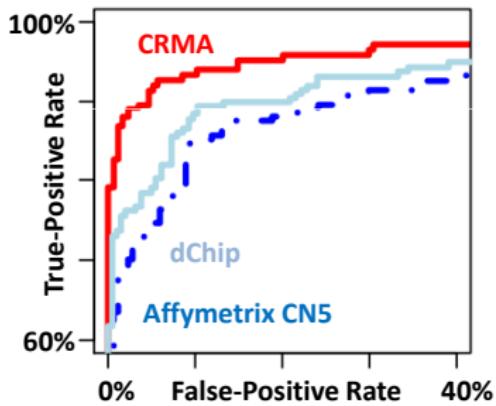
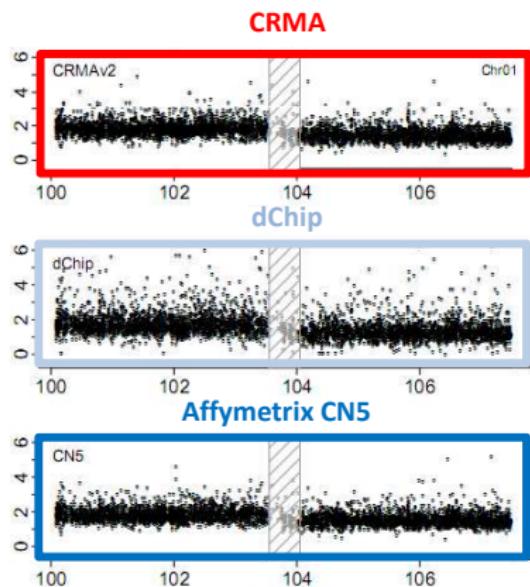
Pour chaque valeur du seuil τ :

- Un SNP sous le seuil τ est appelé DELETION
- Le nombre d'erreurs définit un point sur la courbe ROC



CRMA fait mieux que les méthodes multi-puces

Bengtsson *et al*, *Bioinformatics*, 2008 et Bengtsson *et al*, *Bioinformatics*, 2009



Data set:

- Tumor-normal pairs (HCC1143).

- 68 hybridizations, Affymetrix 6.0

Preprocessing:

- CRMA v2** only two arrays.
- Affymetrix CN5** and **dChip** used all 68 arrays.

slide: H. Bengtsson.

2 Données de nombres de copies d'ADN en cancérologie

- Changements de nombre de copies d'ADN dans les cancers
- Exemple des données de puces SNP

3 Extraction de l'information biologique

- Pre-processing : des signaux comparables entre échantillons
- Post-processing : nombre de copies totaux
- Post-processing : ratios alléliques

4 Modèle statistique pour la segmentation

- Limites des approches directes
- Modèles de rupture
- Solution exacte par programmation dynamique

5 Heuristiques pour la segmentation

- Segmentation binaire récursive (circulaire)
- Relaxation convexe : fused lasso

6 Application aux données de puces SNP

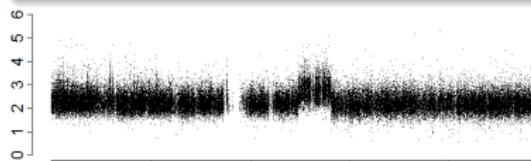
- Extension au problème de segmentation conjointe
- Construction de jeux de données à réponse connue
- Application

Motivation : rapport signal/bruit le long du génome

Pour le SNP j dans l'échantillon i , les signaux observés sont résumés par (θ, β) , où $\theta_{ij} = \theta_{ijA} + \theta_{ijB}$ et $\beta_{ij} = \theta_{ijB}/\theta_{ij}$.

Nombre de copies total

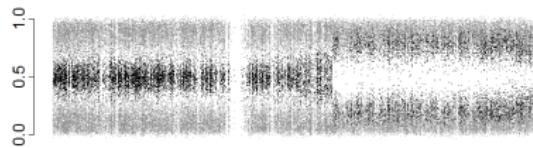
$$\begin{aligned} C_{ij} &= 2 \frac{\theta_{ij}}{\theta_{Rj}} \\ &= C_{1ij} + C_{2ij} \end{aligned}$$



Choix de la référence R ?

Diminution d'hétérozygotie

$$\begin{aligned} DH_{ij} &= 2 |\beta_{ij} - 1/2| \\ &= \frac{C_{2ij} - C_{1ij}}{C_{2ij} + C_{1ij}} \end{aligned}$$



Rapport signal sur bruit faible

Analyse de données de nombre de copies d'ADN

2 Données de nombres de copies d'ADN en cancérologie

3 Extraction de l'information biologique

- Pre-processing : des signaux comparables entre échantillons
- Post-processing : nombre de copies totaux
- Post-processing : ratios alléliques

4 Modèle statistique pour la segmentation

5 Heuristiques pour la segmentation

6 Application aux données de puces SNP

Choix d'une référence

Exemple concret : lignées cellulaires de cancer du sein.

On dispose de 36 expériences, réparties en trois lots.

Choix de référence possibles pour une expérience donnée

- ① 192 échantillons "normaux" d'un autre laboratoire
- ② l'ensemble des 36 échantillons
- ③ les expériences du même lot

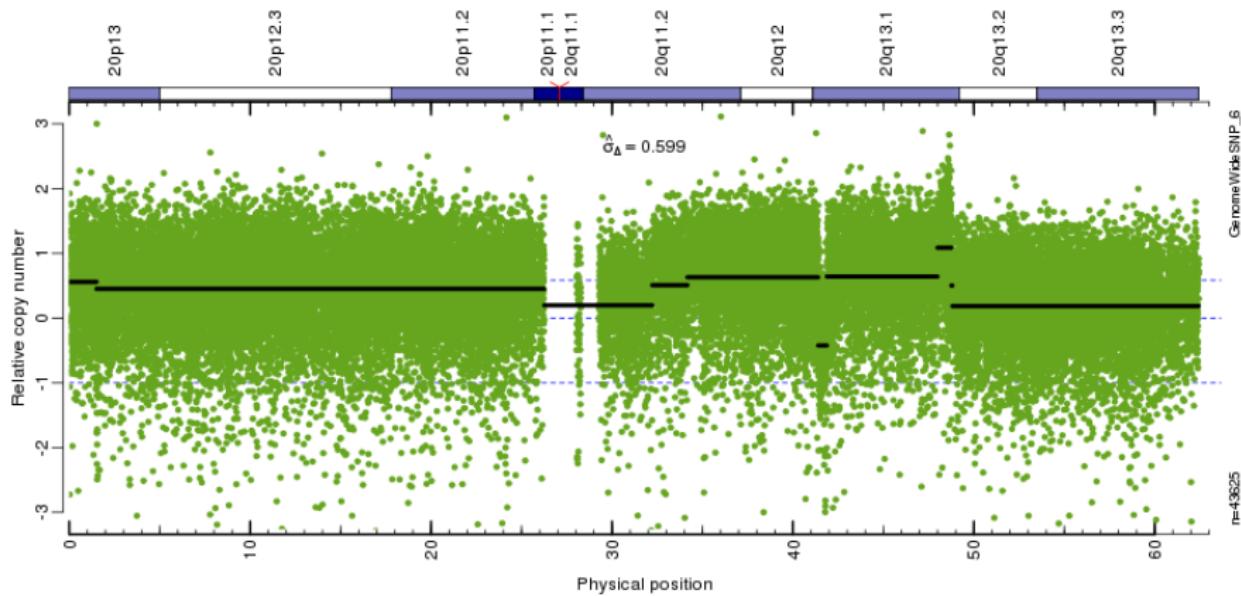
Quantification du niveau de bruit :

$$\widehat{\sigma}_\Delta \propto \text{median} \left(\left| z_j - \text{median}_{j'}(z_{j'}) \right| \right)$$

où les z_j sont les différences successives de nombre de copies d'ADN

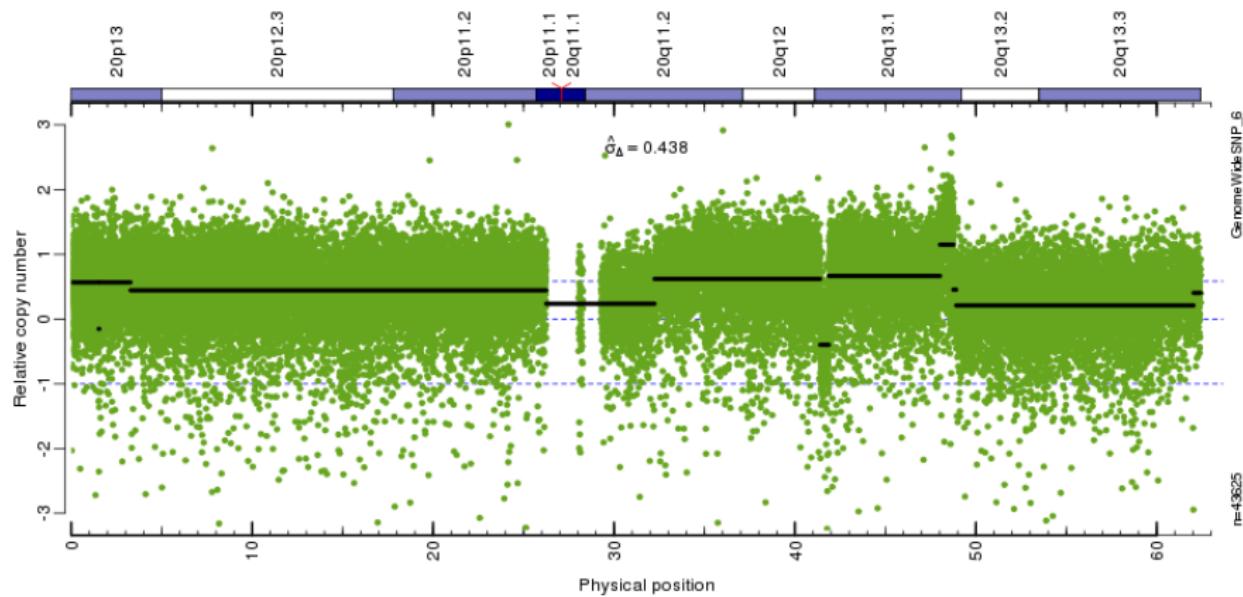
Choix d'une référence

Laboratoire différent ($n=192$)



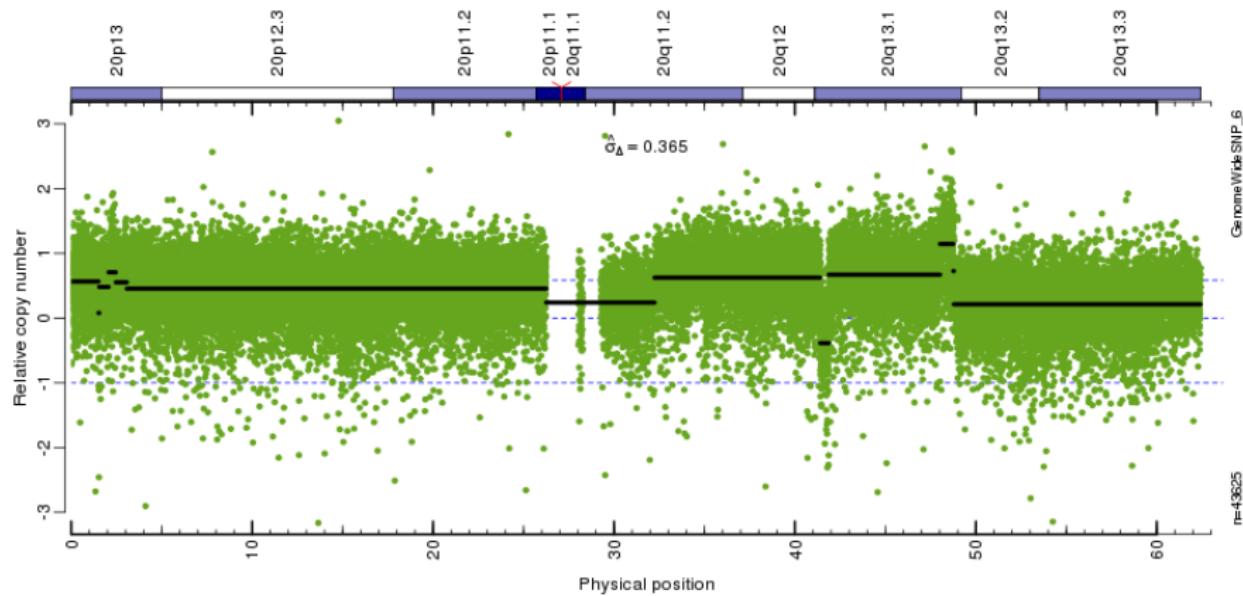
Choix d'une référence

Même laboratoire, tous lots d'expériences confondus ($n=36$)



Choix d'une référence

Même laboratoire, même lot d'expériences (n=22)



Analyse de données de nombre de copies d'ADN

2 Données de nombres de copies d'ADN en cancérologie

3 Extraction de l'information biologique

- Pre-processing : des signaux comparables entre échantillons
- Post-processing : nombre de copies totaux
- Post-processing : ratios alléliques

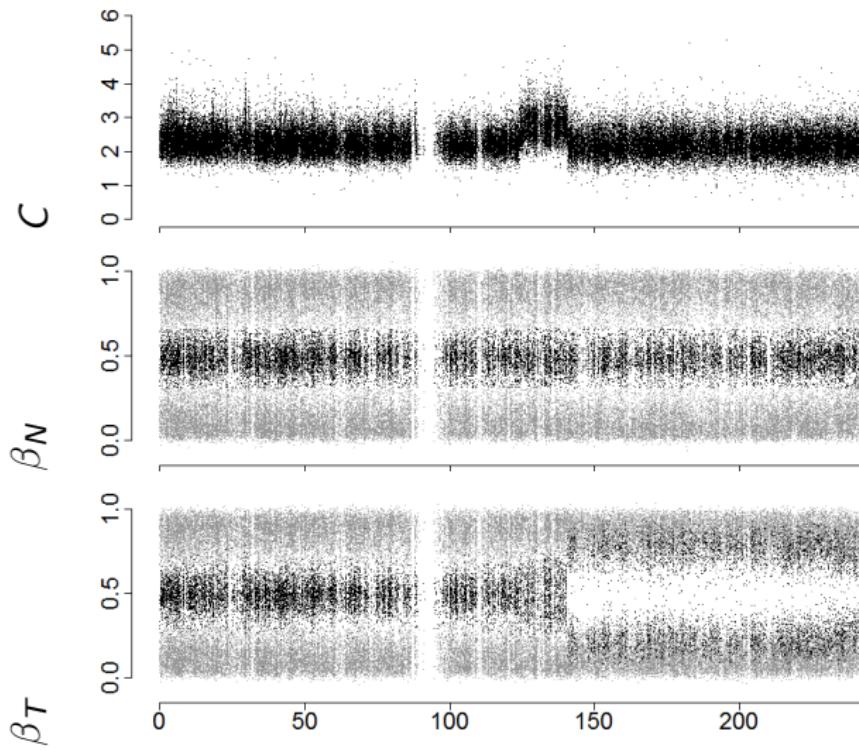
4 Modèle statistique pour la segmentation

5 Heuristiques pour la segmentation

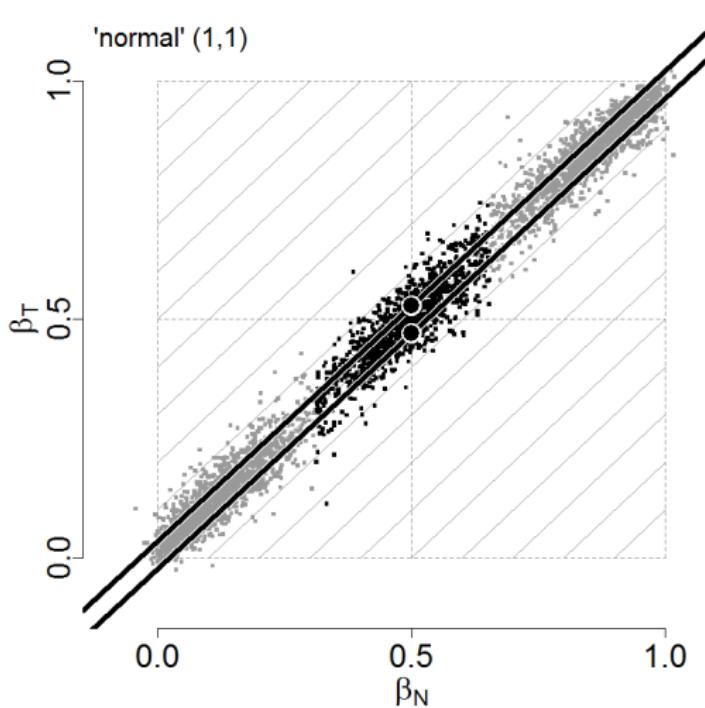
6 Application aux données de puces SNP

Les ratios alléliques sont bruités

Exemple après pre-processing suivant la méthode CRMAv2



Effet SNP dans une région “normale” de la tumeur

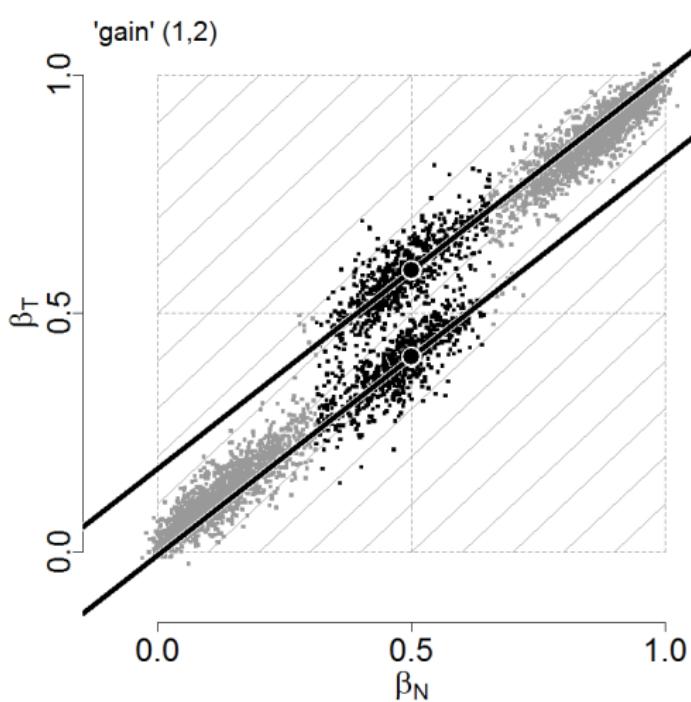


- Au lieu de trois points et $(0,0)$, $(\frac{1}{2}, \frac{1}{2})$ et $(1,1)$, on observe trois groupes ; la déviation est appelée **effet SNP** :

$$\delta_{ij} = \beta_{ij} - \mu_{ij}$$

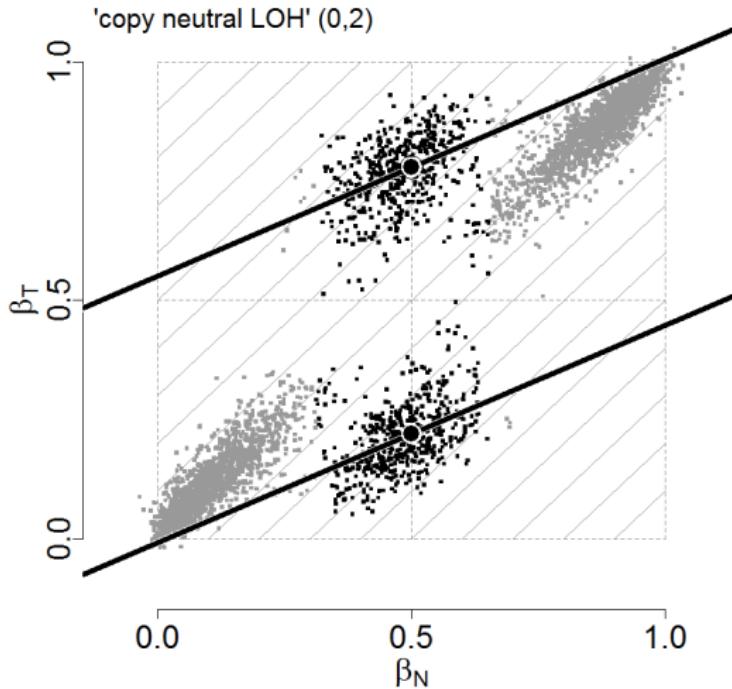
- δ est très similaire dans la tumeur et le normal

Effet SNP dans une région de gain d'une copie d'ADN



- Les groupes homozygotes n'ont pas changé
- Le groupes hétérozygote se sépare en deux et tourne

Effet SNP dans une région d'isodisomie



- Les groupes homozygotes n'ont pas changé
- Les groupes hétérozygotes tournent encore plus

TumorBoost : normalisation d'une paire tumeur/normal

H. Bengtsson et al, *BMC Bioinformatics* (2010)

Idée

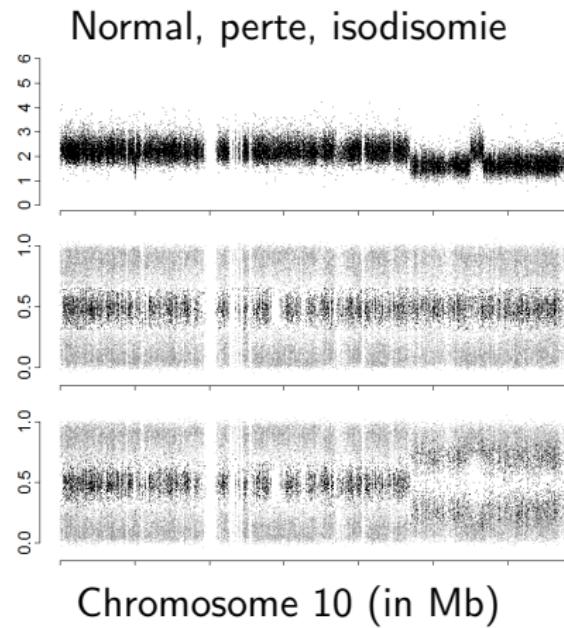
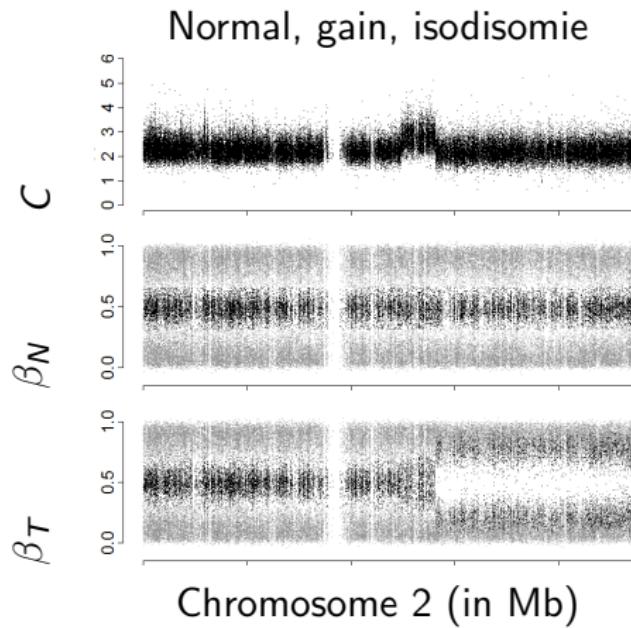
- ① l'effet SNP est similaire dans le normal et dans la tumeur
- ② le normal est relativement simple à analyser (seulement trois génotypes attendus)

⇒ Pour chaque SNP, on “soustrait” des ratios tumoraux l'effet SNP estimé grâce à l'échantillon normal

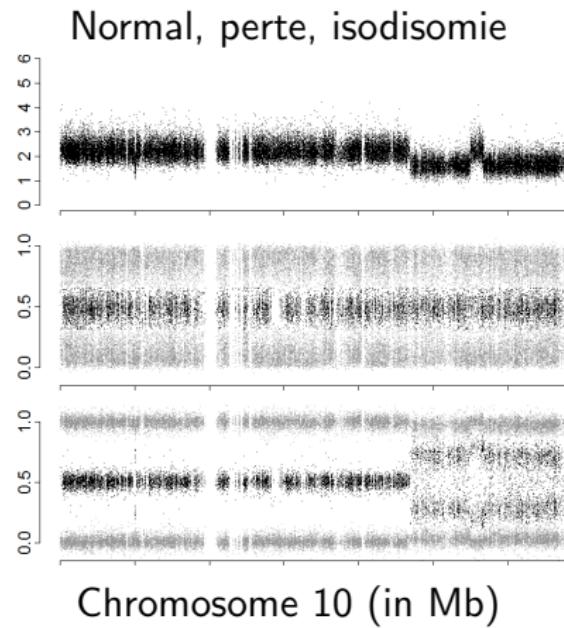
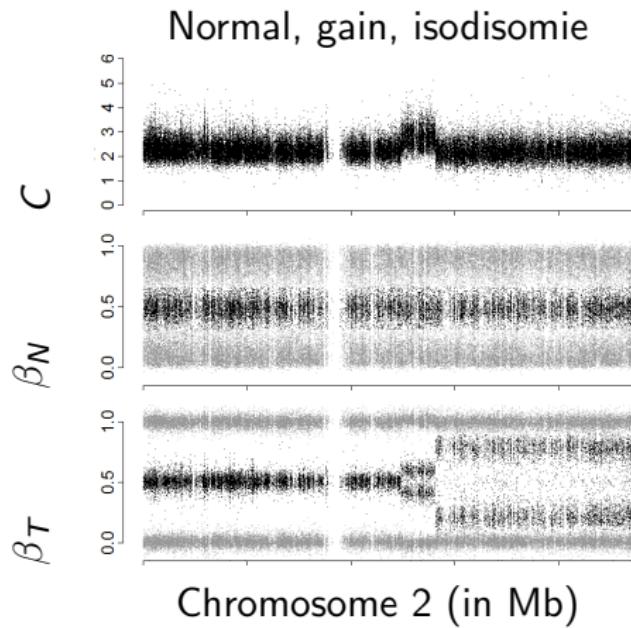
Caractéristiques de la méthode

- Pas besoin de connaître les régions à l'avance
- La normalisation est effectuée SNP par SNP
- Chaque paire tumeur/normal est analysée séparément

Signaux génomiques avant normalisation

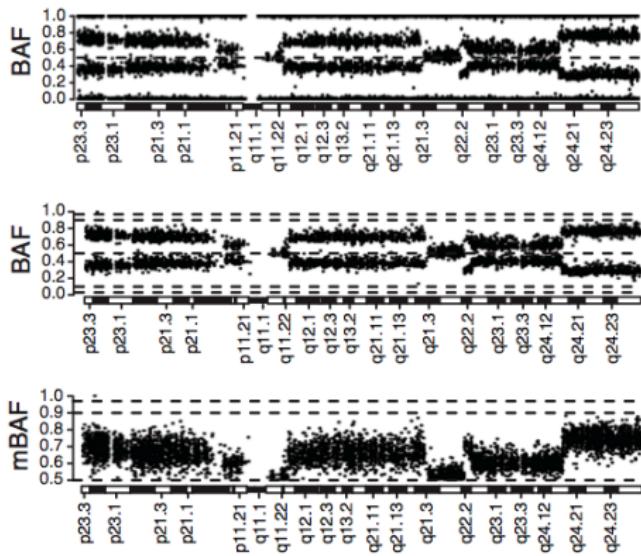


Signaux génomiques après normalisation



Détection de ruptures à partir des ratios alléliques

D'après Staaf *et al*, Genome Biology, 2008



ratios alléliques : β

ratios alléliques pour les SNPs hétérozygotes

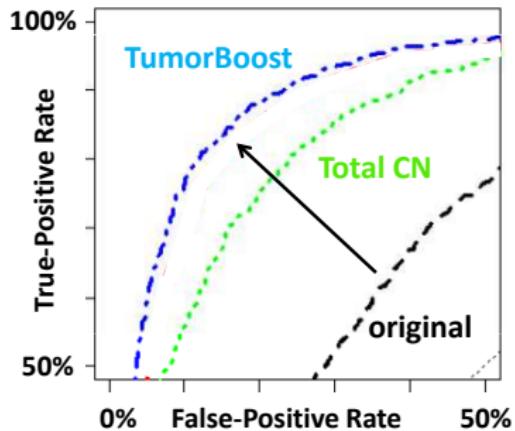
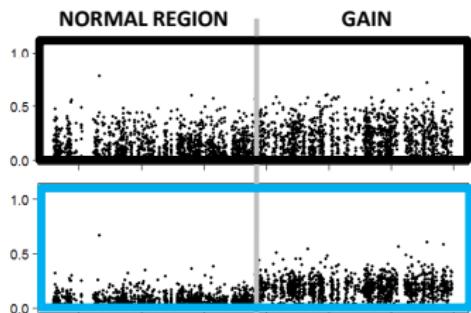
Diminution d'hétérozygotie pour les SNPs hétérozygotes :

$$DH = 2|\beta - 1/2|$$

DH a un seul mode et peut donc être segmenté facilement
On utilise la même évaluation ROC que pour les signaux totaux

Résultat : Meilleure détectabilité des ruptures

Allelic imbalance



slide: H. Bengtsson.

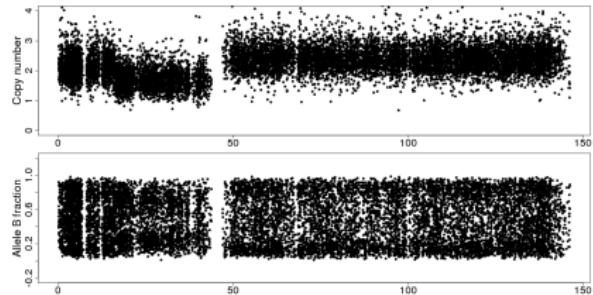
Sans échantillon normal apparié : CalMaTe

Ortiz-Estevez *et al*

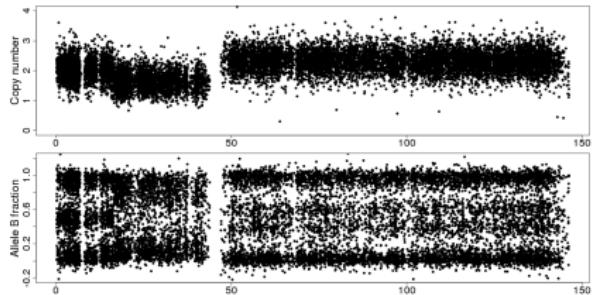
Pour chaque SNP :

- Estimer une fonction de calibration des signaux observés aux génotypes grâce à un ensemble d'échantillons de référence
- Calibrer les échantillons test à partir de cette fonction

Avant CalMaTe



Après CalMaTe



Analyse de données de nombre de copies d'ADN

2 Données de nombres de copies d'ADN en cancérologie

3 Extraction de l'information biologique

4 Modèle statistique pour la segmentation

- Limites des approches directes
- Modèles de rupture
- Solution exacte par programmation dynamique

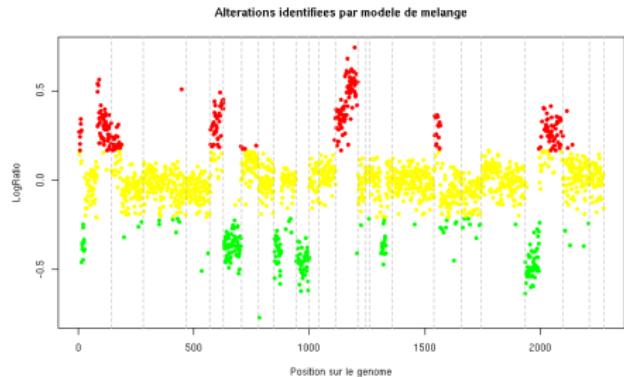
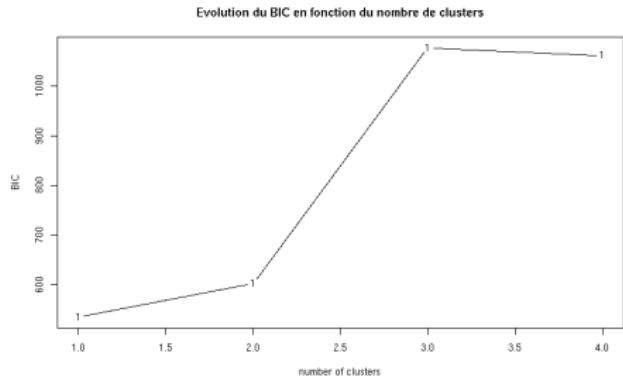
5 Heuristiques pour la segmentation

6 Application aux données de puces SNP

Modèles de mélange

Méthode

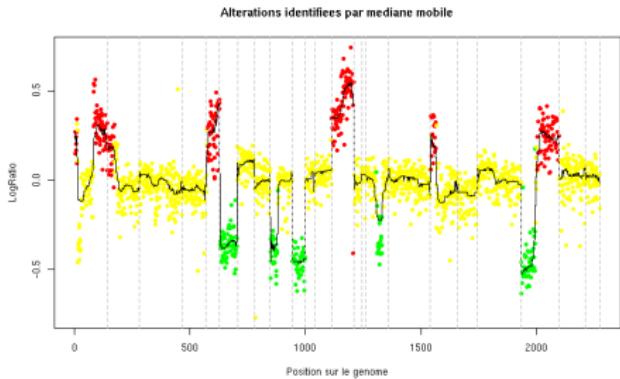
- à K fixé, estimation d'un modèle de mélange par l'algorithme EM
- choix de K à l'aide d'une pénalisation de type BIC



Méthodes de lissage

Méthode

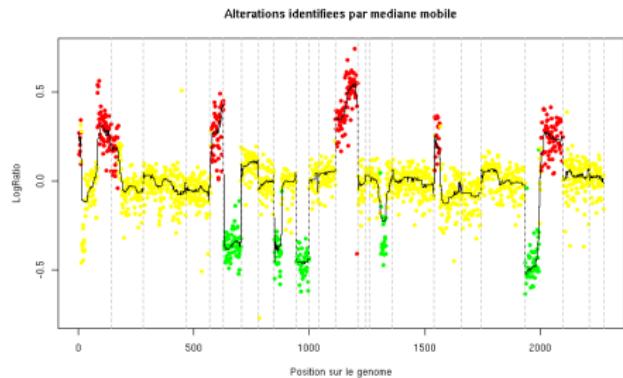
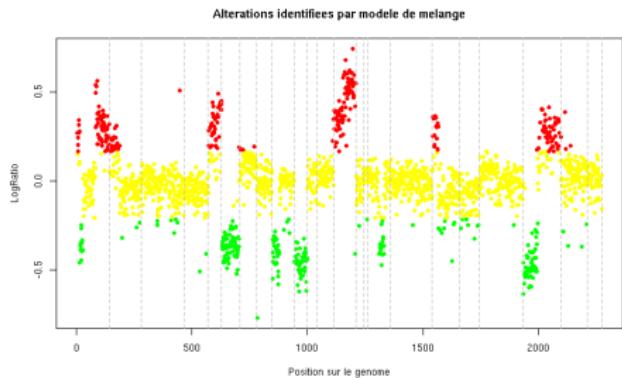
- calcul d'une médiane mobile autour de chaque locus
- découpage du signal obtenu en plusieurs classes



Paramètres

- diamètre de la fenêtre mobile
- nombre de classes
- seuils à appliquer sur le signal lissé

Nécessité de méthodes plus fines



Une “bonne méthode” doit combiner :

- une prise en compte de la **dimension du génome**
- la possibilité de détecter des **ruptures**

Analyse de données de nombre de copies d'ADN

2 Données de nombres de copies d'ADN en cancérologie

3 Extraction de l'information biologique

4 Modèle statistique pour la segmentation

- Limites des approches directes
- Modèles de rupture
- Solution exacte par programmation dynamique

5 Heuristiques pour la segmentation

6 Application aux données de puces SNP

Modèles de rupture

Notations

- $\mathcal{J} = 1 \dots J$: loci génomiques
- $\gamma = (\gamma_j)_{j=1 \dots J}$: vrais nombres de copies d'ADN
- $\mathbf{c} = (c_j)_{j=1 \dots J}$: observations

Hypothèses

- ruptures : $\mathbf{t}(K) = (t_k)_{0 \leq k \leq K}$, vecteur ordonné avec $t_0 = 1$ et $t_K = J$
- nombres de copies d'ADN de niveau région $\mathbf{\Gamma} = (\Gamma_k)_{1 \leq k \leq K}$

tels que $\gamma_j = \Gamma_k$; $\forall j \in [t_{k-1}, t_k)$, $\forall k \in \{1, \dots, K\}$..

On observe donc $c_j = \Gamma_{k(j)} + \varepsilon_j$, avec $k(j) = \max\{k, t_k \leq j\}$, où les erreurs $(\varepsilon_j)_{j=1 \dots J}$ sont iid et en général supposées de loi $\mathcal{N}(0, \sigma^2)$

Estimation dans le modèle de rupture Gaussien

Log-vraisemblance du modèle

$$\ell(K, 1 : J) = -\frac{J}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^K \sum_{j=t_{k-1}}^{t_k} (c_j - \Gamma_{k(j)})^2.$$

A t_k fixés, $\widehat{\Gamma}_{k(j)}^{EMV} = \frac{1}{t_k - t_{k-1}} \sum_{j=t_{k-1}}^{t_k} c_j$

Maximiser ℓ revient à résoudre le problème d'optimisation :

$$\min_{(\gamma_j)_{1 \leq j \leq J}} \sum_{j=1}^J (c_j - \gamma_j)^2 \quad \text{s.c.} \quad \sum_{j=1}^{J-1} \mathbf{1}_{\gamma_{j+1} \neq \gamma_j} \leq K$$

Deux problèmes statistiques

Le **nombre** et la **position** des ruptures sont inconnus :

- problème de sélection de modèle : choix de K
- problème de segmentation : localisation des ruptures parmi les $\binom{K-1}{J-1}$ possibles $\binom{K-1}{J-1} = O(J^{K-1})$: une recherche exhaustive est impossible dans des situations réalistes. $\binom{10^5}{50} = 3.2 \times 10^{185}$

Stratégie classique :

- ➊ pour $1 \leq k \leq K_{\max}$, trouver un “bon” modèle à k ruptures
- ➋ estimer k à l'aide d'un critère de choix de modèle

Dans ce cours : on s'intéresse au problème de **segmentation**

Quelques unes des approches existantes

- Programmation dynamique (DP)
- Segmentation binaire récursive (CART)
- Segmentation binaire récursive circulaire (CBS)
- Fused Lasso
- Modèles de Markov cachés (HMM)

Compromis entre performances statistiques et efficacité algorithmique ?

Analyse de données de nombre de copies d'ADN

2 Données de nombres de copies d'ADN en cancérologie

3 Extraction de l'information biologique

4 Modèle statistique pour la segmentation

- Limites des approches directes
- Modèles de rupture
- Solution exacte par programmation dynamique

5 Heuristiques pour la segmentation

6 Application aux données de puces SNP

Programmation dynamique

Picard *et al.*, 2005

$V(k, j_1 : j_2)$: log-vraisemblance du meilleur modèle à k segments entre j_1 et j_2

Programmation dynamique : additivité de la log-vraisemblance

- ➊ Calculer $V(1, j_1 : j_2)$ pour tous (j_1, j_2) tels que $1 \leq j_1 < j_2 \leq J$
- ➋ Passer de $V(k, \cdot)$ à $V(k + 1, \cdot)$ en remarquant que

$$V(k + 1, j_1 : j_2) = \max_{h \in [j_1, j_2]} V(k, j_1 : h) + V(1, (h + 1) : j_2)$$

La complexité en temps pour détecter K ruptures passe de $O(J^K)$ à $O(KJ^2)$, pour une complexité en espace de $O(J^2)$.

Programmation dynamique élaguée

pruned DP algorithm (pDPA)

- Rigaill, 2010
- Programmation dynamique “élaguée” : solution exacte en $O(KJ \log(J))$ (complexité moyenne)
- Ne fonctionne que pour un signal uni-dimensionnel

Pruned Exact Linear Time (PELT)

- Killick *et al*, 2011
- Complexité linéaire ($O(J)$) pour trouver la meilleure solution en un nombre de ruptures non connu à l'avance

Analyse de données de nombre de copies d'ADN

- 2 Données de nombres de copies d'ADN en cancérologie
- 3 Extraction de l'information biologique
- 4 Modèle statistique pour la segmentation
- 5 Heuristiques pour la segmentation
 - Segmentation binaire récursive (circulaire)
 - Relaxation convexe : fused lasso
- 6 Application aux données de puces SNP

Segmentation binaire récursive

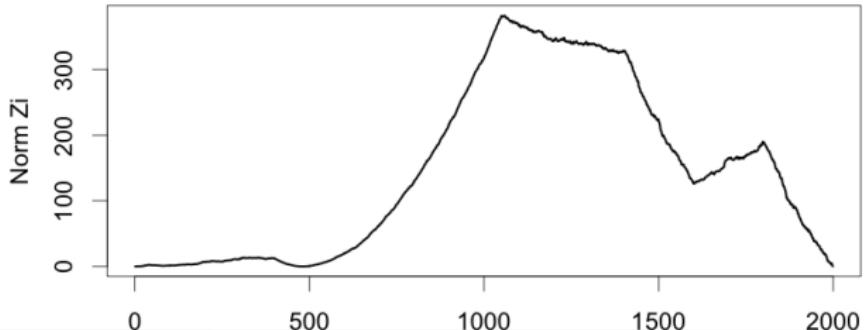
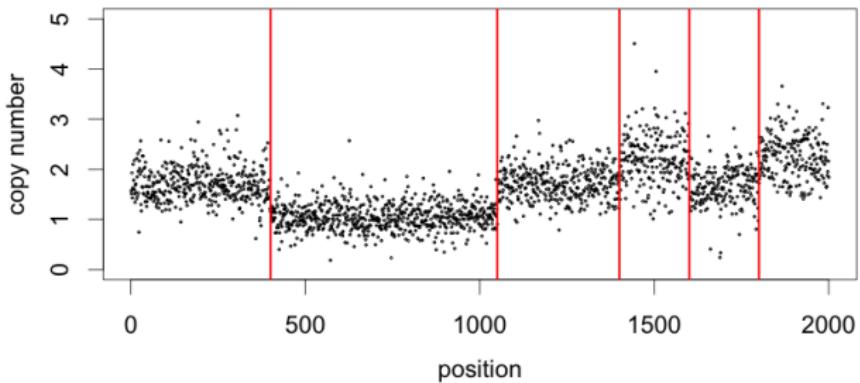
- Tester l'hypothèse \mathcal{H}_0 : "Pas de point de cassure" contre \mathcal{H}_1 : "Exactlement un point de cassure"
- Statistique du rapport de vraisemblance $\max_{1 \leq j \leq J} |Z_j|$

$$Z_j = \frac{\left(\frac{S_j}{j} - \frac{S_J - S_j}{J-j} \right)}{\sqrt{\frac{1}{j} + \frac{1}{J-j}}},$$

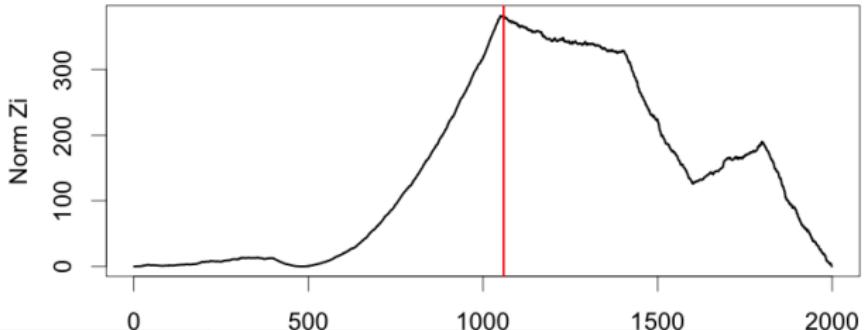
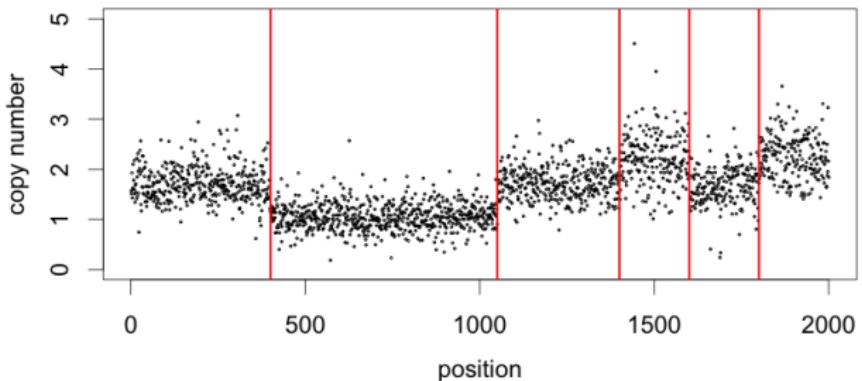
où $S_j = \sum_{1 \leq i \leq j} c_i$.

- point de cassure candidat : $\arg \max_i |Z_j|$
- complexité de cette étape : $O(J)$

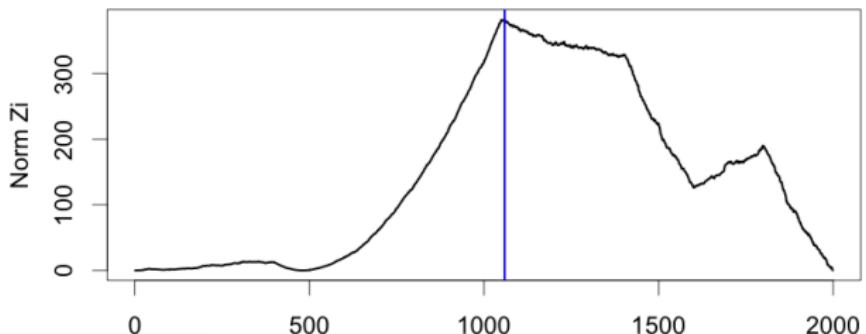
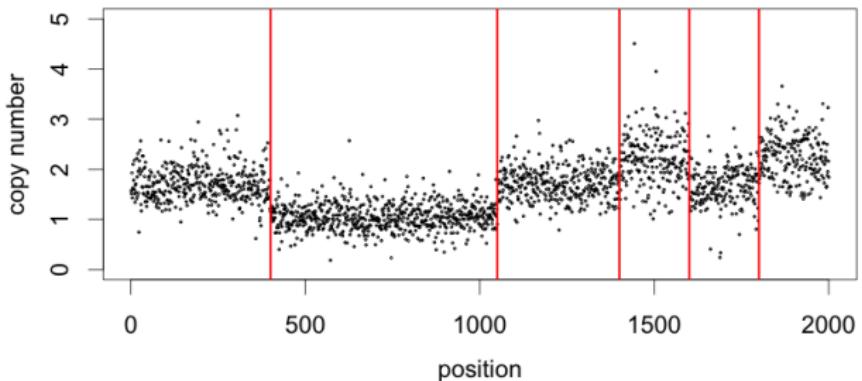
Segmentation binaire récursive



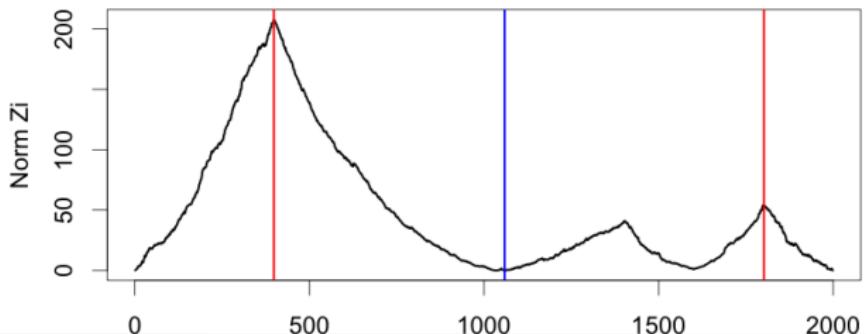
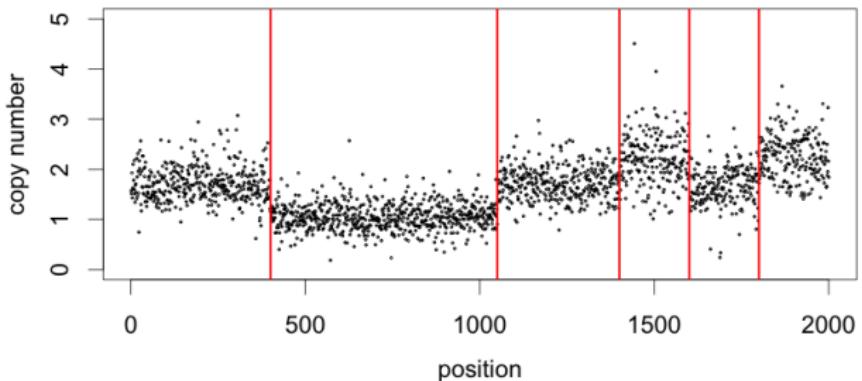
Segmentation binaire récursive



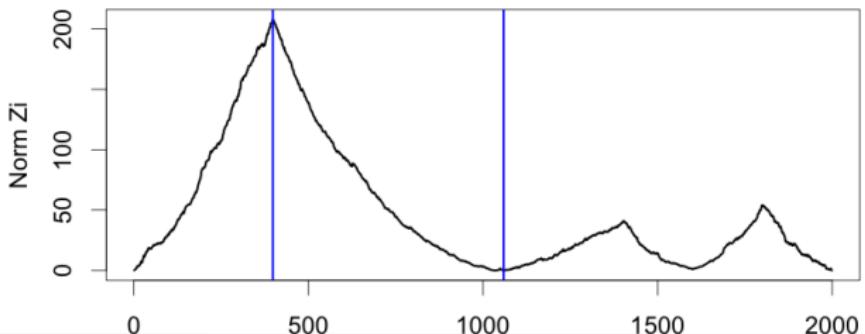
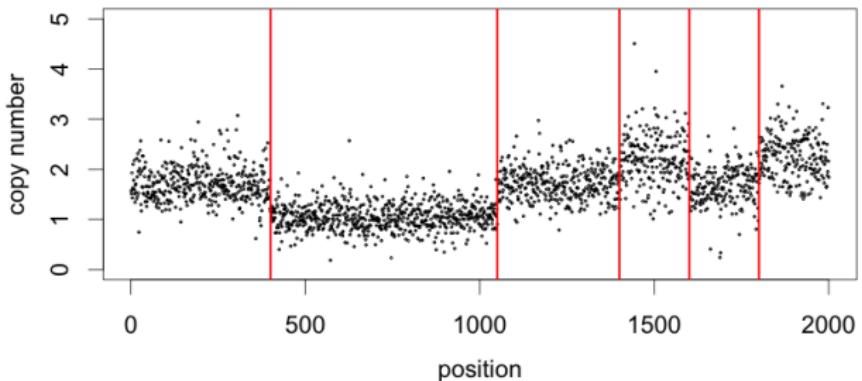
Segmentation binaire récursive



Segmentation binaire récursive



Segmentation binaire récursive



Segmentation binaire récursive circulaire

Circular Binary Segmentation (Olshen *et al*, 2004)

Motivation : la segmentation binaire n'est pas *a priori* adaptée pour détecter un signal en créneau.

- détection de **segments imbriqués** :

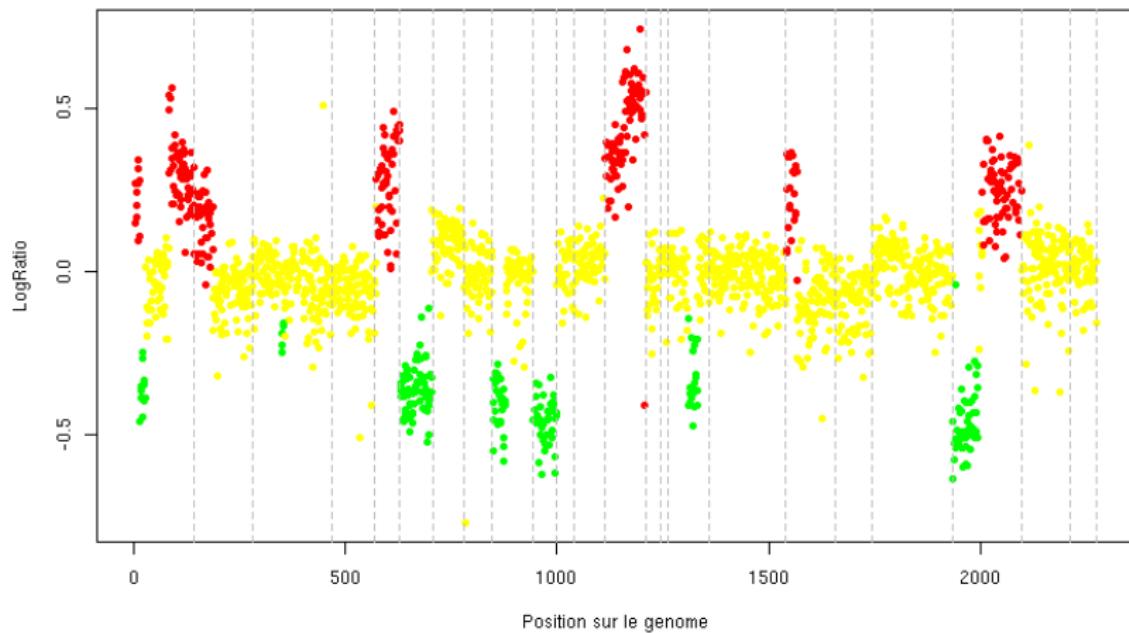
$$Z_{uv} = \frac{1}{\sqrt{\frac{1}{v-u} + \frac{1}{J-(v-u)}}} \times \left[\frac{S_v - S_u}{v-u} - \frac{S_J - (S_v - S_u)}{J - (v-u)} \right]$$

- détection de **plusieurs segments** par récursion
- calcul d'une *p*-value à l'aide de **permutations**

Segmentation binaire récursive

Circular Binary Segmentation (Olshen *et al*, 2004)

Alterations identifiées par Circular Binary Segmentation



Segmentation binaire récursive

Passage à l'échelle (Venkatraman and Olshen, 2007)

Nécessité d'un algorithme plus rapide

- algorithme quadratique en le nombre de loci
- nouvelles puces à haute résolution

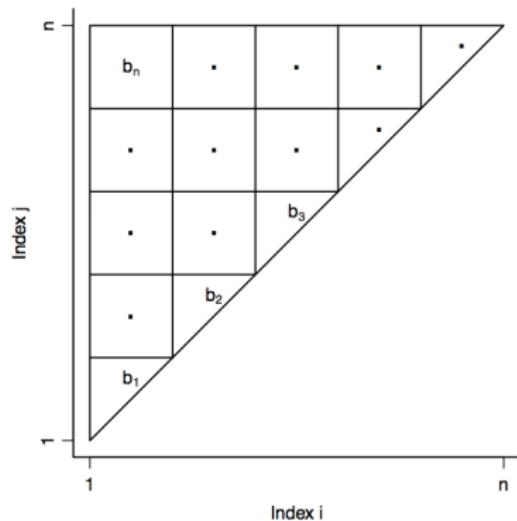
Nouveau mode de calcul des *p*-values

- calcul exact si peu de loci, approché sinon
- arrêt rapide si locus très probablement significatif
- algorithme quasi linéaire ($O(J \log(J))$)

Segmentation binaire récursive

Algorithme en $O(J \log(J))$ (Venkatraman and Olshen, 2007)

Idée : diviser $\{1, \dots, J\}$ en blocs de taille $\sqrt{J} \times \sqrt{J}$



- seulement **J blocs** au total !
- borne sur la statistique maximale dans chaque bloc
- la plupart des blocs sont **élagués** sans recherche exhaustive

Illustration : V. E. Seshan

Analyse de données de nombre de copies d'ADN

- 2 Données de nombres de copies d'ADN en cancérologie
- 3 Extraction de l'information biologique
- 4 Modèle statistique pour la segmentation
- 5 Heuristiques pour la segmentation
 - Segmentation binaire récursive (circulaire)
 - Relaxation convexe : fused lasso
- 6 Application aux données de puces SNP

Principe de la relaxation convexe

Rappel :

Maximiser ℓ revient à résoudre le problème d'optimisation :

$$\min_{(\gamma_j)_{1 \leq j \leq J}} \sum_{j=1}^J (c_j - \gamma_j)^2 \quad \text{s.c.} \quad \sum_{j=1}^{J-1} \mathbf{1}_{\gamma_{j+1} \neq \gamma_j} \leq K$$

Ce n'est pas un problème convexe ! Stratégie "classique" : remplacer la contrainte ℓ_0 par une contrainte ℓ_1 .

$$\min_{(\gamma_j)_{1 \leq j \leq J}} \sum_{j=1}^J (c_j - \gamma_j)^2 \quad \text{s.c.} \quad \sum_{j=1}^{J-1} |\gamma_{j+1} - \gamma_j| \leq C$$

Estimation dans les modèles linéaires avec $n \ll p$

Modèle : $Y = X\beta + u$

- Y : observations (vecteur de taille n)
- X : variables (matrice de taille $n \times p$)
- β : paramètre à estimer (vecteur de taille p)
- u : résidus (vecteur de taille n)

On s'intéresse ici notamment au cas où $n \ll p$

Limites de la méthode traditionnelle (MCO)

- ➊ $\hat{\beta}_j^0$ peu biaisés mais variants
- ➋ difficulté d'interprétation : aucun $\hat{\beta}_j^0$ n'est strictement nul

Par ailleurs, pas de solution unique dans le cas où $n < p$

Régression Ridge

Hoerl & Kennard, In Encyclopedia of Statistical Sciences (1988). *Ridge regression.*

Principe : ajout d'une contrainte sur $\|\beta\|_2$

On résout le programme d'optimisation sous contrainte

$$\underset{\beta}{\text{Min}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Avantages et inconvénients : ridge regression vs RFE

	stabilité	interprétabilité
Ridge regression	++	-
Recursive Feature Elimination	-	++

⇒ Peut-on faire mieux à la fois en stabilité et interprétabilité ?

LASSO

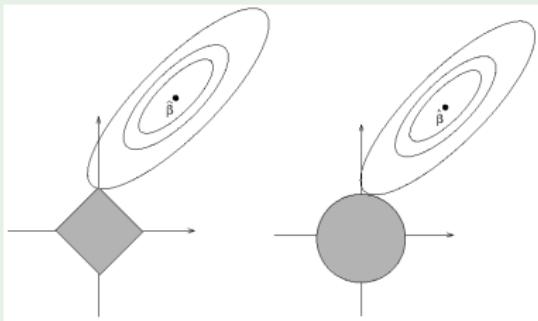
Tibshirani, JRSS (B), 1996. *Regression shrinkage and selection via the lasso.*

Principe : ajout d'une contrainte sur $\|\beta\|_1$

On résout le programme d'optimisation sous contrainte

$$\underset{\beta}{\text{Min}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

LASSO vs ridge regression



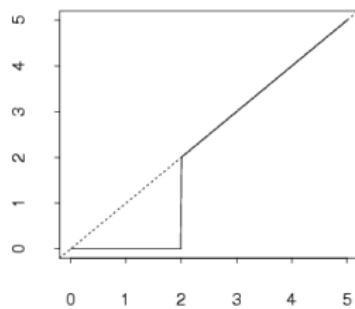
Avantages

- Bonne interprétabilité : certains $\hat{\beta}_j$ sont **nuls**
- Bonne stabilité : les autres sont moins variants que les $\hat{\beta}_j^{OLS}$ ("shrinkage")

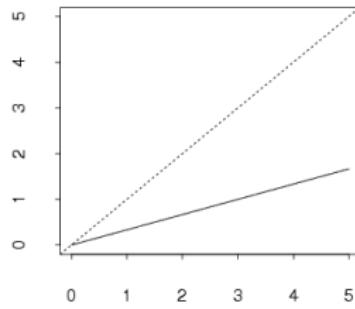
Comparaison des trois méthodes

Interprétation géométrique : $\hat{\beta} = f(\beta)$ (cas orthonormal)

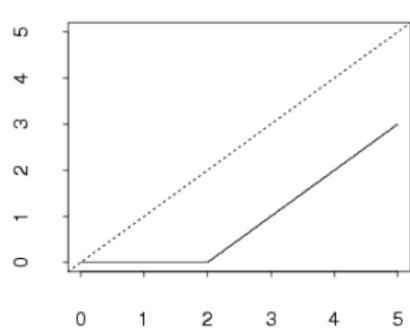
Subset



Ridge



Lasso



$$\hat{\beta}_j^0 \mathbf{1}_{\hat{\beta}_j^0 > \gamma}$$

$$\frac{1}{1 + \gamma} \hat{\beta}_j^0$$

$$signe(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+$$

$\hat{\beta}_j^0$ est l'estimateur des MCO

Relaxations convexes

Résolution d'un problème approché mais plus facile algorithmiquement

Adaptation du "Fused Lasso" (Tibshirani and Wang, 2007)

$$\min_{(\gamma_j)_{1 \leq j \leq J}} \sum_{j=1}^J (c_j - \gamma_j)^2 \quad \text{s.c.} \quad \sum_{j=1}^{J-1} |\gamma_{j+1} - \gamma_j| \leq v \text{ et } \sum_{j=1}^J |\gamma_j - 2| \leq u$$

Simplification (Harchaoui and Lévy-Leduc, 2008)

$$\min_{(\gamma_j)_{1 \leq j \leq J}} \sum_{j=1}^J (c_j - \gamma_j)^2 \quad \text{s.c.} \quad \sum_{j=1}^{J-1} |\gamma_{j+1} - \gamma_j| \leq v$$

Complexité : $O(JK)$

Conclusions sur les modèles de rupture

Arbitrage entre exactitude de la réponse au problème posé, et temps de calcul.

Possibilité de procéder en deux étapes :

- ① recherche très rapide mais avec des faux positifs
- ② élagage, ou recherche exhaustive sur les solutions restantes

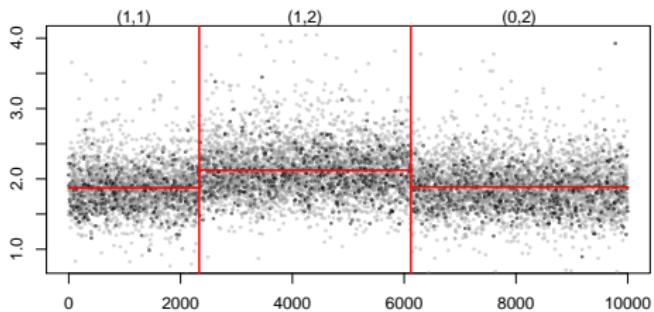
Nécessité de minimiser le taux de **faux négatifs** à la première étape

Analyse de données de nombre de copies d'ADN

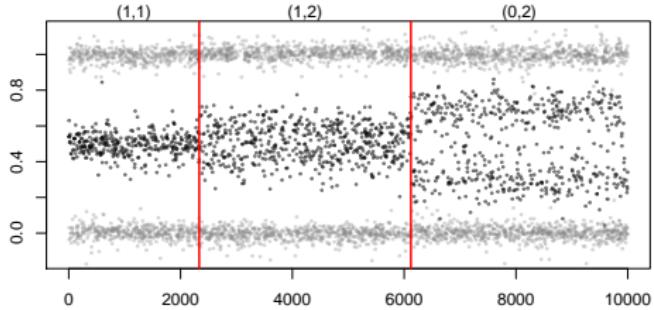
- 2 Données de nombres de copies d'ADN en cancérologie
- 3 Extraction de l'information biologique
- 4 Modèle statistique pour la segmentation
- 5 Heuristiques pour la segmentation
- 6 Application aux données de puces SNP
 - Extension au problème de segmentation conjointe
 - Construction de jeux de données à réponse connue
 - Application

Données de puces SNP en cancérologie

Nombres de copies totaux (C)

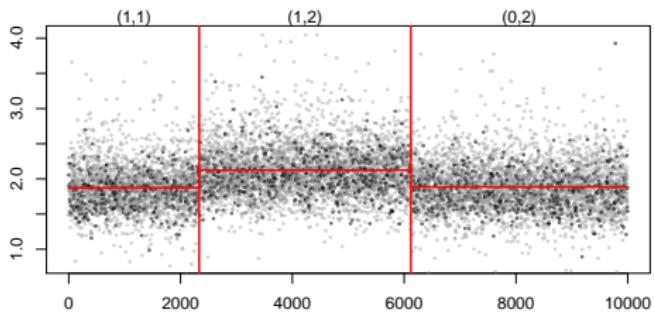


Ratios alléliques (β)

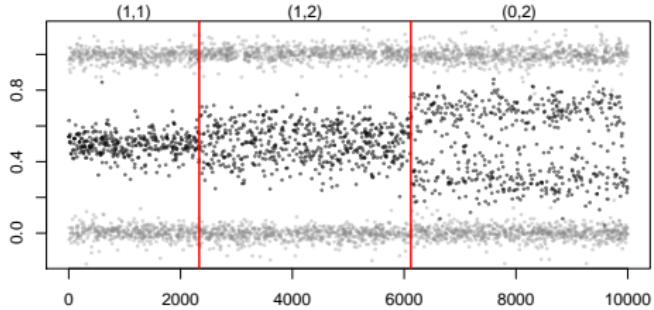


Données de puces SNP en cancérologie

Nombres de copies totaux (C)

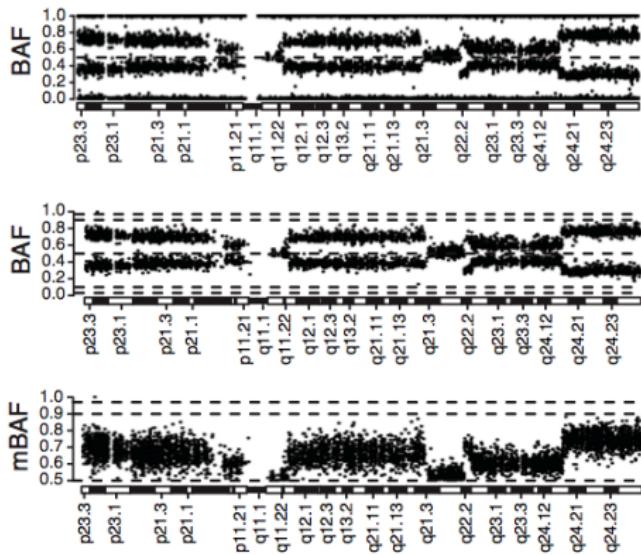


Ratios alléliques (β)



Rappel : détection de ruptures à partir des ratios alléliques

D'après Staaf *et al*, Genome Biology, 2008



ratios alléliques : β

ratios alléliques pour les SNPs hétérozygotes

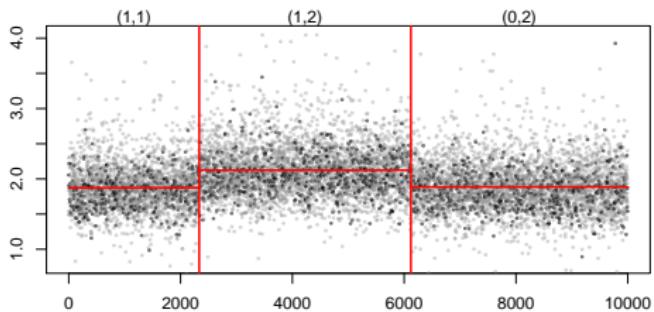
Diminution d'hétérozygotie pour les SNPs hétérozygotes :

$$DH = 2|\beta - 1/2|$$

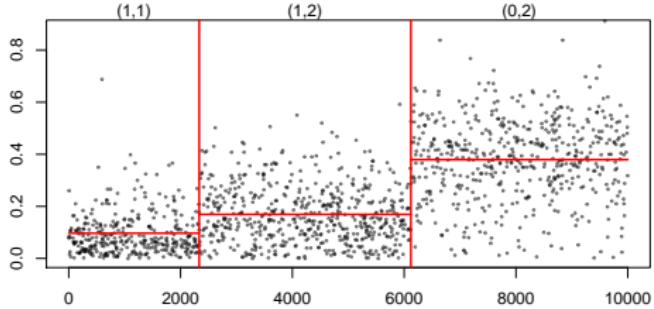
DH a un seul mode et peut donc être segmenté facilement

Données de puces SNP en cancérologie

Nombres de copies totaux (C)



Diminution d'hétérozygotie (DH)



Quelques méthodes de segmentation (conjointe)

Méthode	Temps	# dims
Dynamic programming (DP)		
[Rigaill et al.(2010)]	$J \log(J)$	1
[Picard et al. (2005)]	$d \cdot K \cdot J^2$	≥ 1
Fused Lasso		
[Harchaoui and Lévy-Leduc(2008)]	$K \cdot J$	1
[Bleakley and Vert (2011)]	$d \cdot K \cdot J$	≥ 1
Recursive binary segmentation (RBS/CART)		
[Gey and Lebarbier (2008)]	$dJ \log(K)$	≥ 1
Circular binary segmentation (CBS)		
[Olshen AB et al. (2004)]	$J \log(J)$	1
[Olshen AB et al. (2011)]	$J \log(J)$	2
[Zhang et al.(2010)]	$d \cdot J^2$	≥ 1
Hidden Markov Models (HMM)		
[Lai et al.]	J^2	1
[Chen et al. (2011)]	J^2	2

Questions

- Les méthodes qui exploitent les deux dimensions du signal sont-elles toujours plus performantes que les méthodes “1d” correspondantes ?
- La programmation dynamique est-elle toujours la meilleure ?

En théorie (dans le modèle Gaussien), les réponses sont évidentes. Est-ce vrai en pratique ?

Besoins

- Schéma d'évaluation des performances qui permet de répondre à ces questions
- Paramètres biologiques influençant ces performances

Analyse de données de nombre de copies d'ADN

- ② Données de nombres de copies d'ADN en cancérologie
- ③ Extraction de l'information biologique
- ④ Modèle statistique pour la segmentation
- ⑤ Heuristiques pour la segmentation
- ⑥ Application aux données de puces SNP
 - Extension au problème de segmentation conjointe
 - Construction de jeux de données à réponse connue
 - Application

Proposed approach

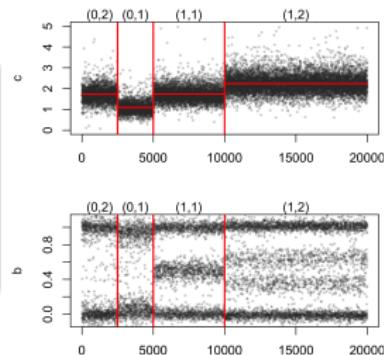
Limitations of existing approaches

- simulation models : hard to get biological insight
- dilution series : few regions
- automatically annotated data sets : depend on a segmentation method
- manually annotated data sets : SNR cannot be tuned

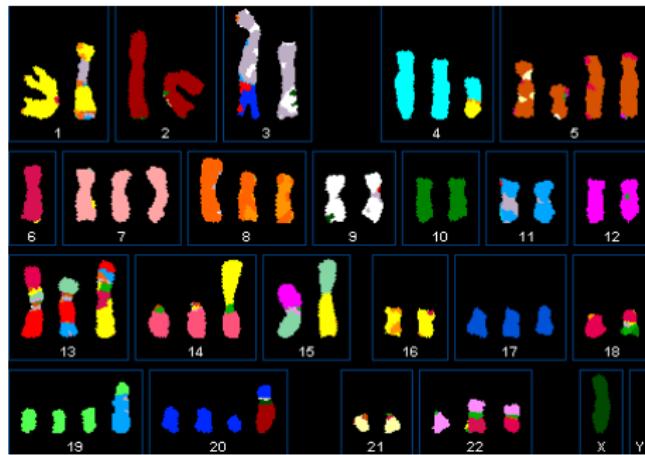
Ingredients for the proposed approach

- ① breakpoint positions : $(t_k)_{k=1 \dots K}$
- ② copy-number state labels : $(\Gamma_k)_{k=1 \dots K+1}$
- ③ signal : resampled from real data

This requires real data with known “truth”



Lung cancer cell line NCI-H1395

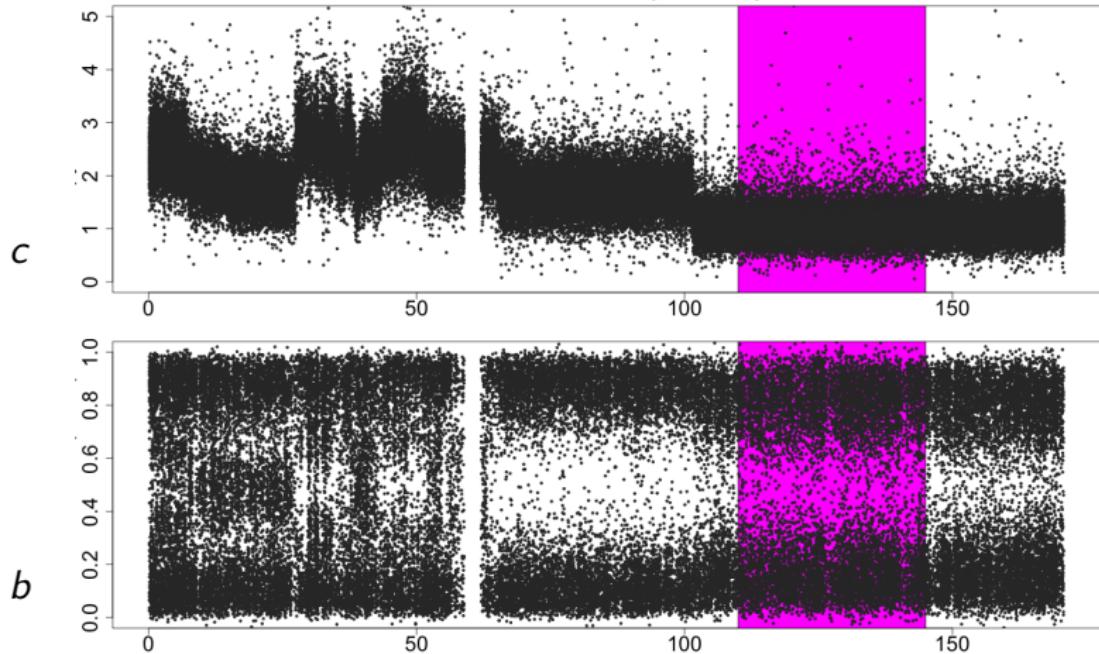


from :

<http://www.path.cam.ac.uk/~pawefish/LungCellLineDescriptions/NCI-H1395.html>

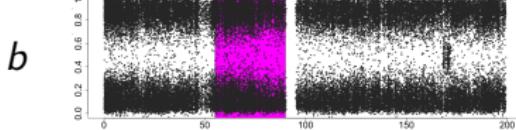
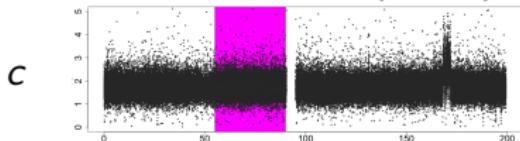
Real data annotation : NCI-H1395, chr 6

Loss of one copy (Chr 6)

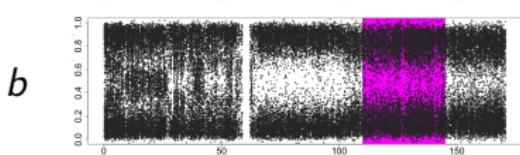
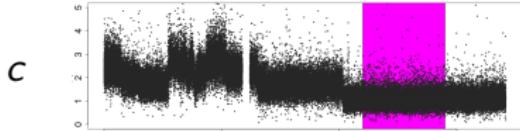


Real data annotation : NCI-H1395

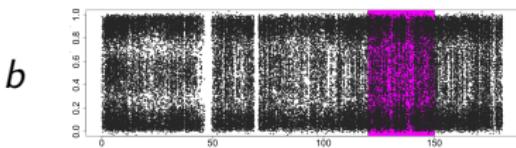
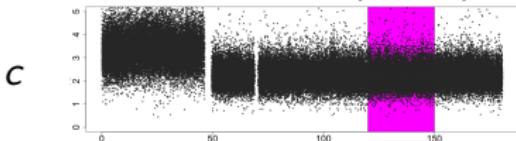
Copy-neutral LOH (Chr 3)



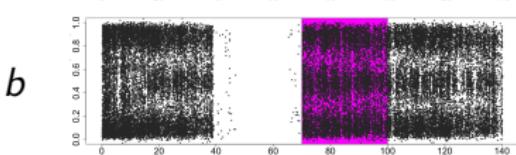
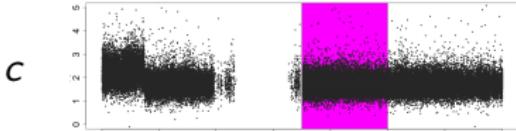
Loss of one copy (Chr 6)



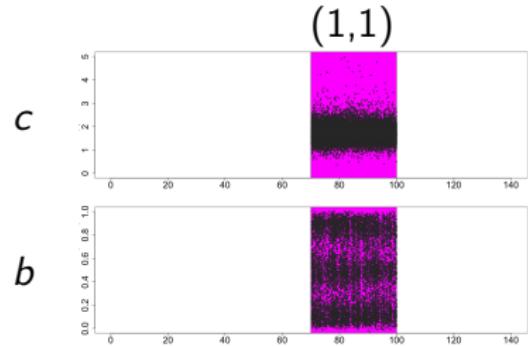
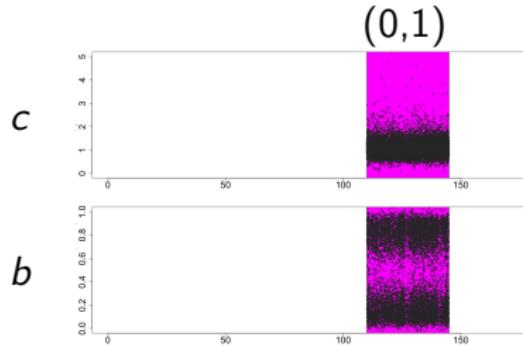
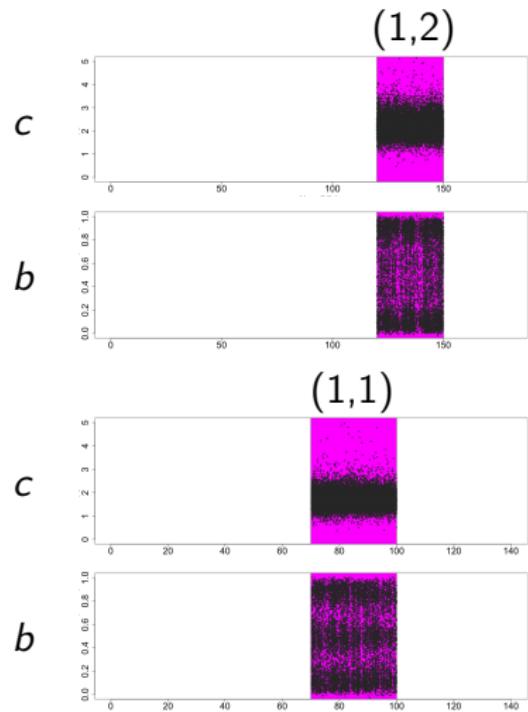
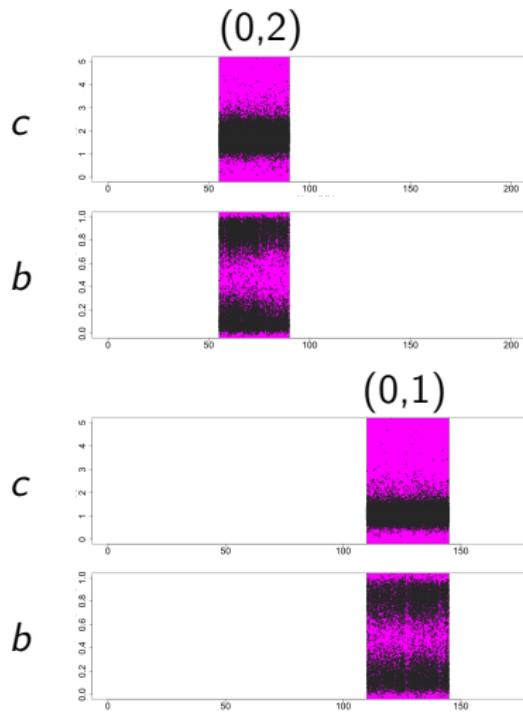
Gain of one copy (Chr 5)



Normal (Chr 9)

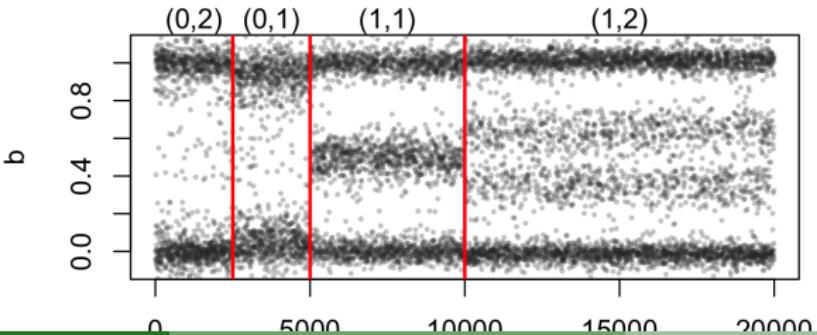
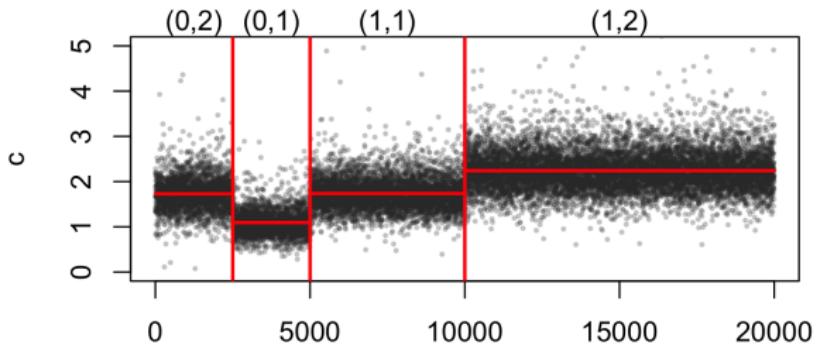


Real data annotation : NCI-H1395



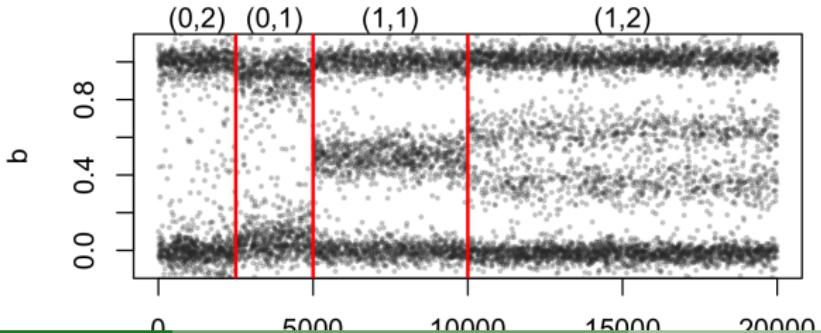
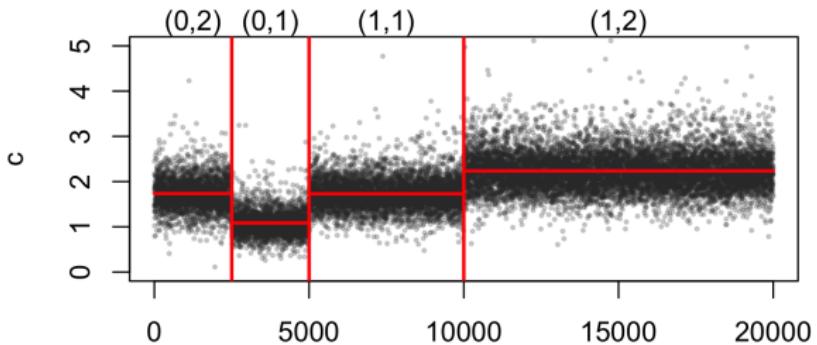
Synthetic data generation

Example : data set 1, 100% tumor cells



Synthetic data generation

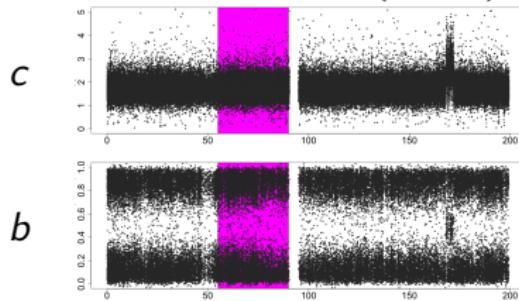
Example : data set 1, 100% tumor cells (same "truth")



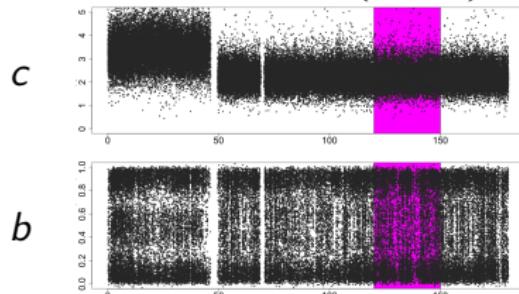
Real data annotation : NCI-H1395

100% tumor cells

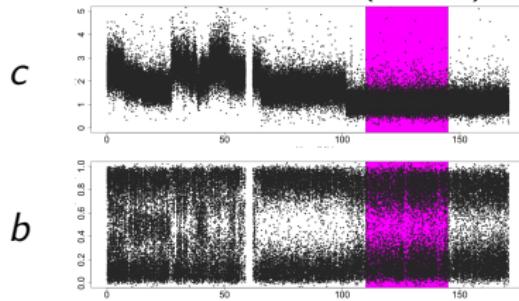
Copy-neutral LOH (Chr 3)



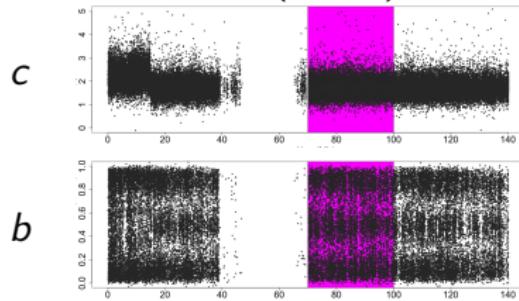
Gain of one copy (Chr 5)



Loss of one copy (Chr 6)



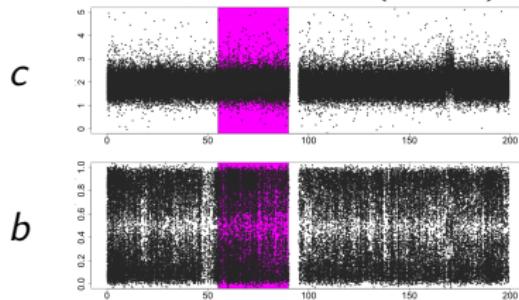
Normal (Chr 9)



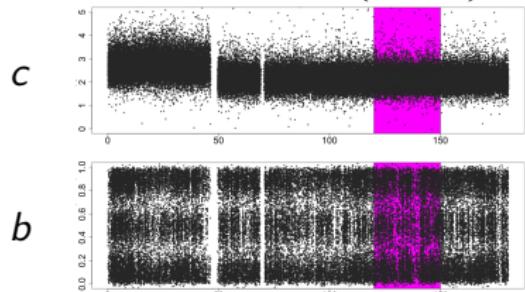
Real data annotation : NCI-H1395

70% tumor cells (using annotation from the 100% data set !)

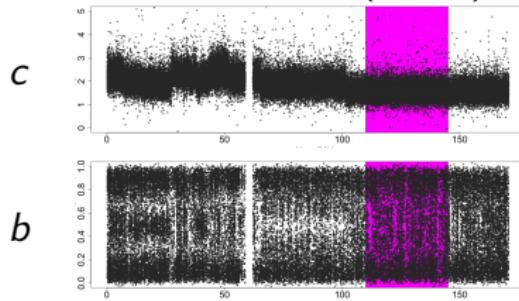
Copy-neutral LOH (Chr 3)



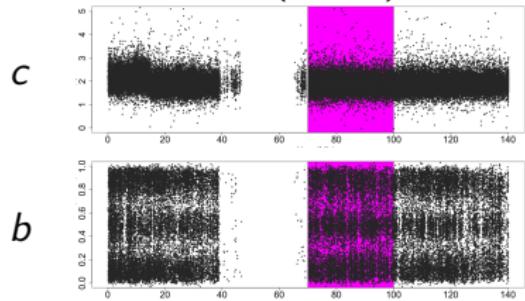
Gain of one copy (Chr 5)



Loss of one copy (Chr 6)



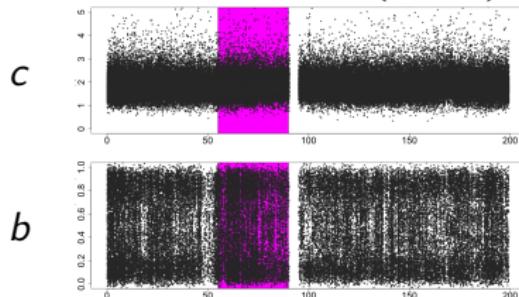
Normal (Chr 9)



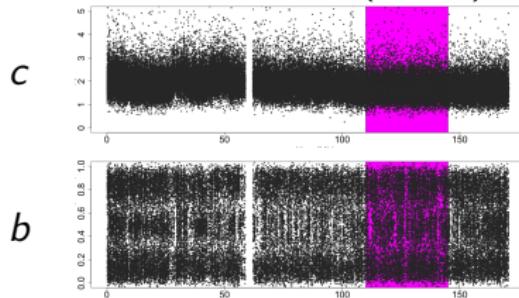
Real data annotation : NCI-H1395

50% tumor cells (using annotation from the 100% data set !)

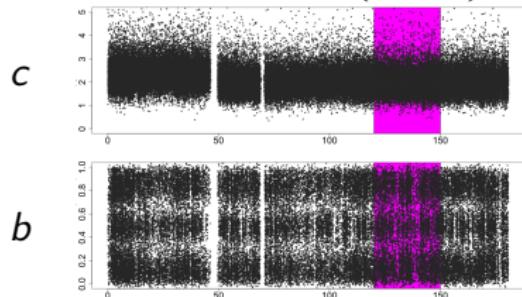
Copy-neutral LOH (Chr 3)



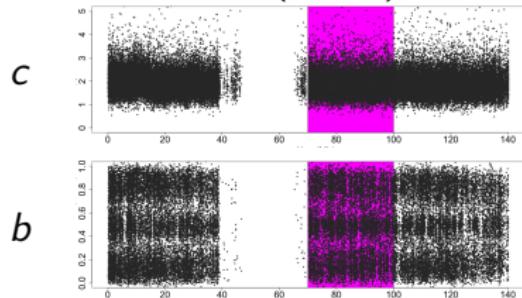
Loss of one copy (Chr 6)



Gain of one copy (Chr 5)



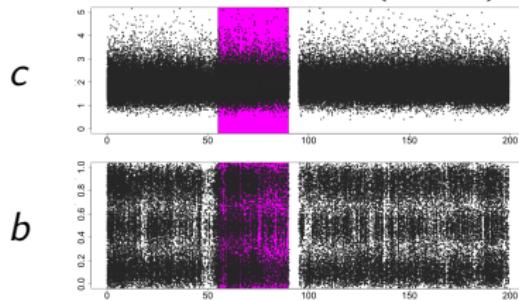
Normal (Chr 9)



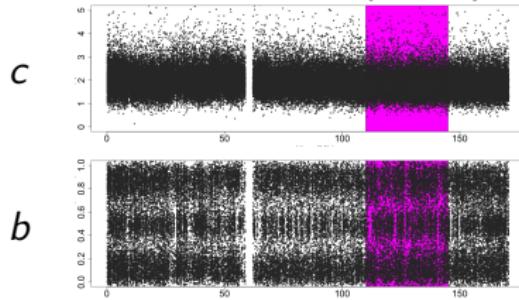
Real data annotation : NCI-H1395

30% tumor cells (using annotation from the 100% data set !)

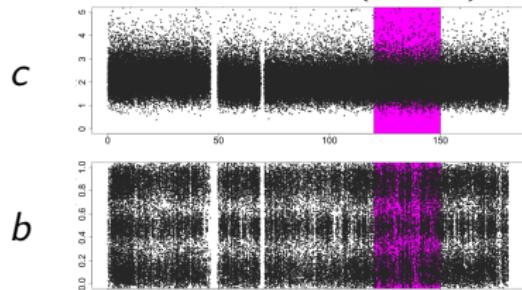
Copy-neutral LOH (Chr 3)



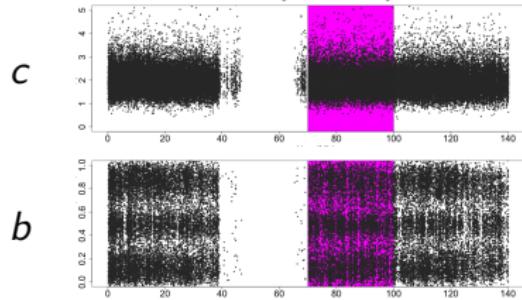
Loss of one copy (Chr 6)



Gain of one copy (Chr 5)

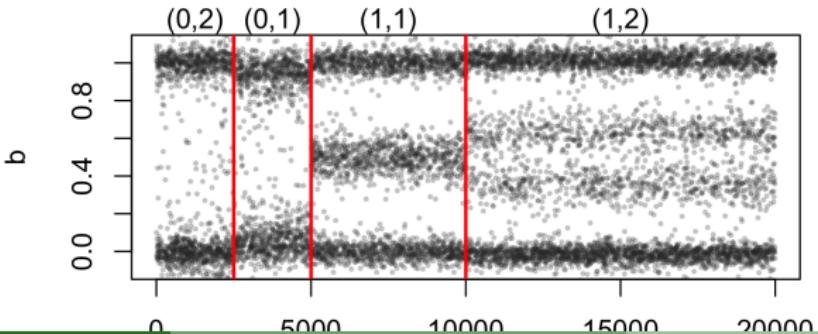
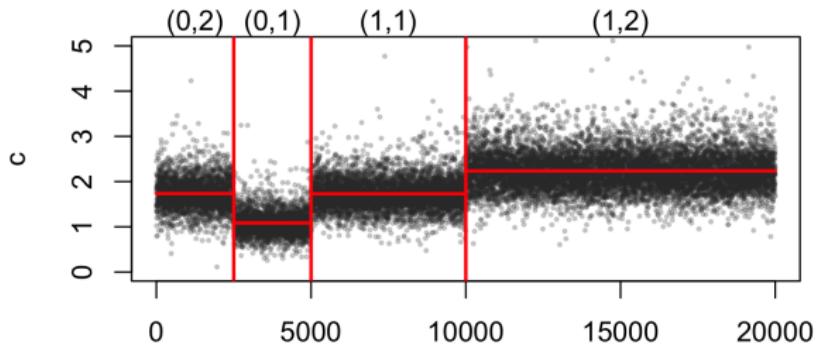


Normal (Chr 9)



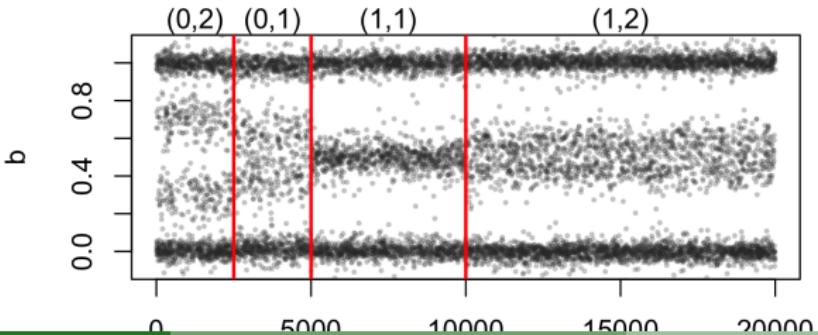
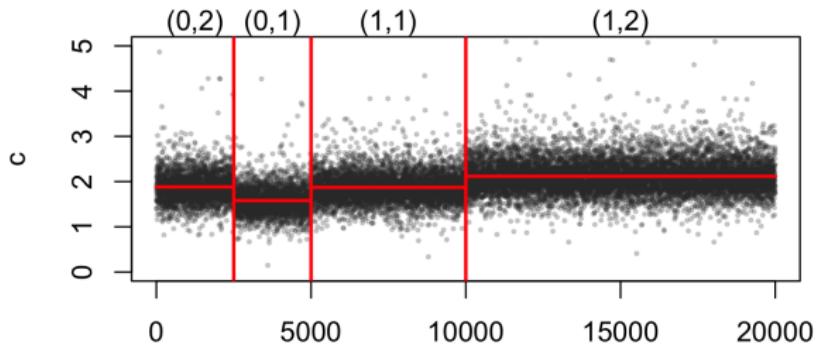
Signal-to-noise ratio can be controlled

Example : data set 1, 100% tumor cells



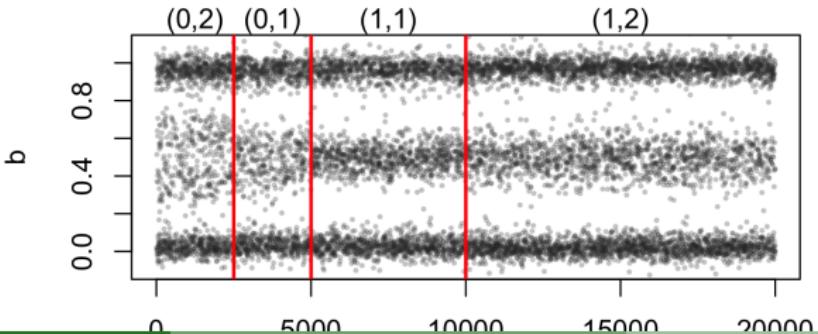
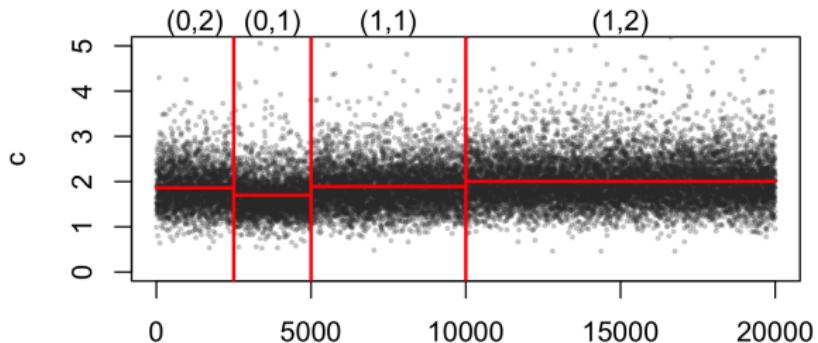
Signal-to-noise ratio can be controlled

Example : data set 1, 70% tumor cells (same “truth”)



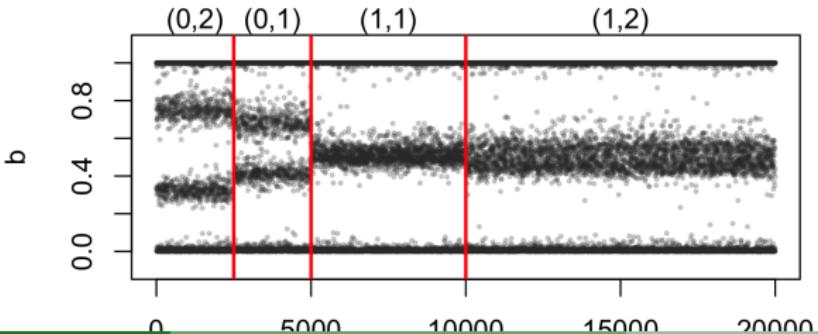
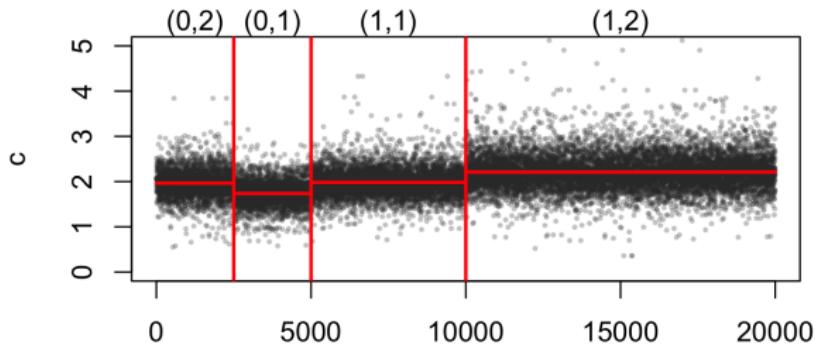
Signal-to-noise ratio can be controlled

Example : data set 1, 50% tumor cells (same “truth”)



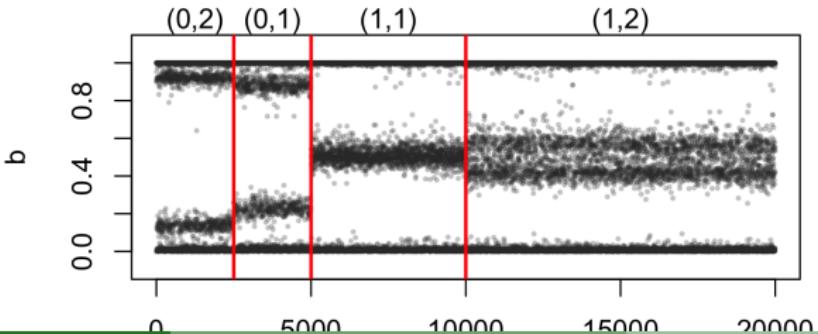
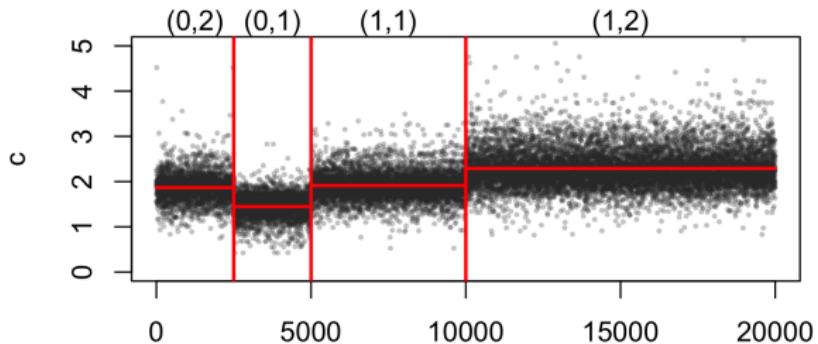
Signal-to-noise ratio can be controlled

Example : data set 2, 50% tumor cells (same “truth”)



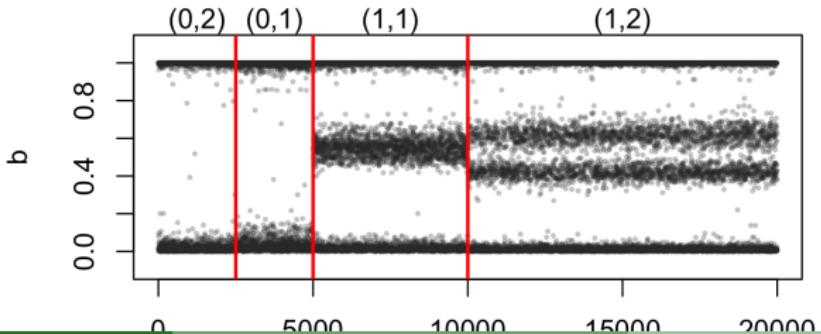
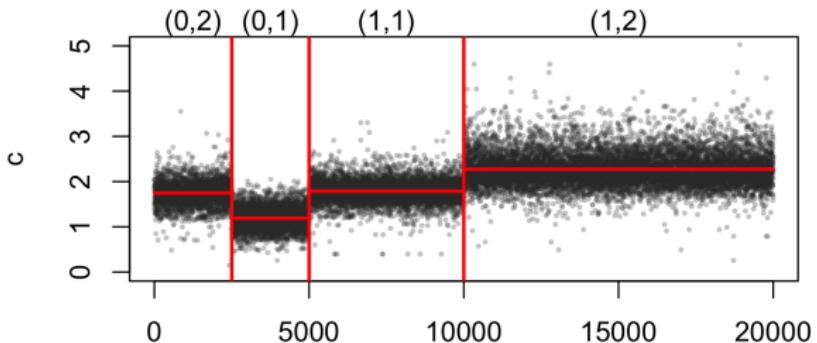
Signal-to-noise ratio can be controlled

Example : data set 2, 79% tumor cells (same "truth")



Signal-to-noise ratio can be controlled

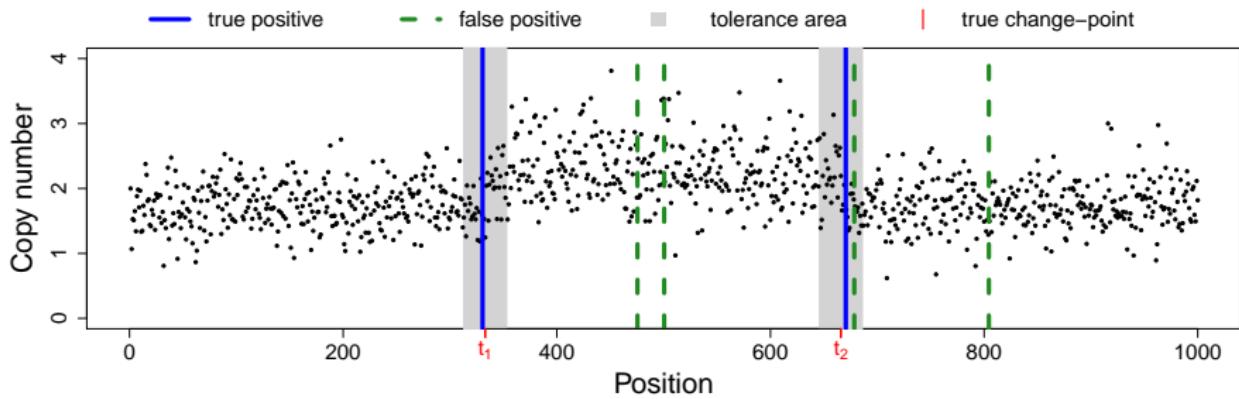
Example : data set 2, 100% tumor cells (same "truth")



Analyse de données de nombre de copies d'ADN

- 2 Données de nombres de copies d'ADN en cancérologie
- 3 Extraction de l'information biologique
- 4 Modèle statistique pour la segmentation
- 5 Heuristiques pour la segmentation
- 6 Application aux données de puces SNP
 - Extension au problème de segmentation conjointe
 - Construction de jeux de données à réponse connue
 - Application

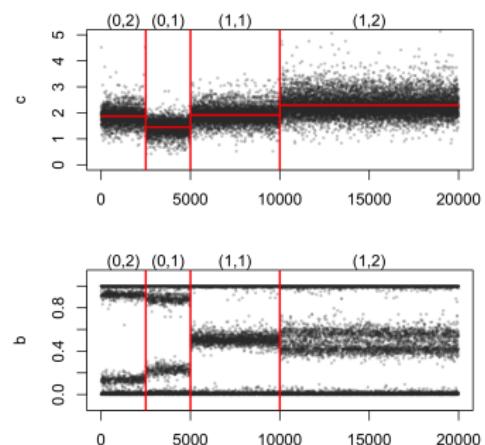
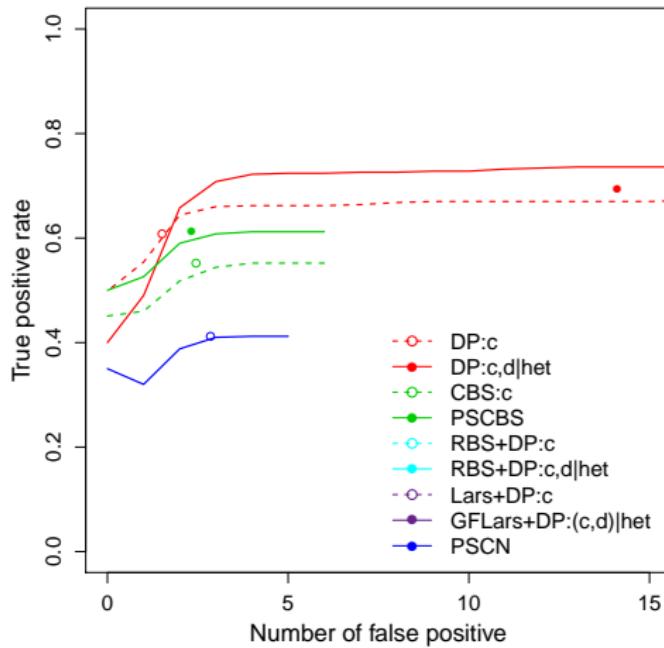
Defining true and false positives



- two breakpoints at t_1 and t_2
- TP=2, FP=4

Taking both dimensions into account helps

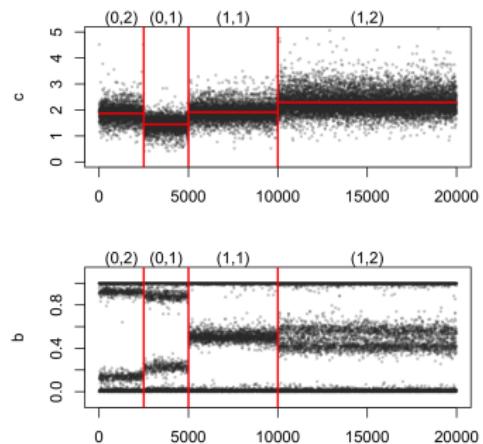
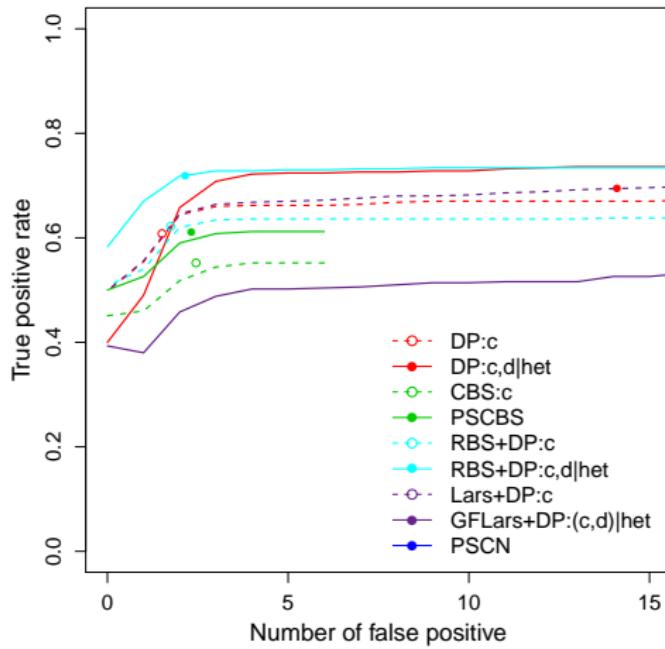
100 profiles, $n = 5000$, $K = 5$, purity = 79%, precision = 1



2d methds \gg 1d methods

Taking both dimensions into account helps... or not

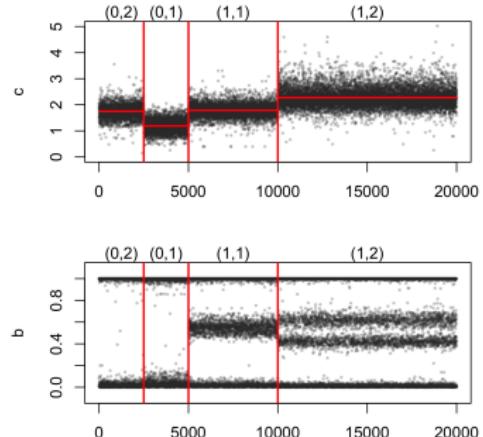
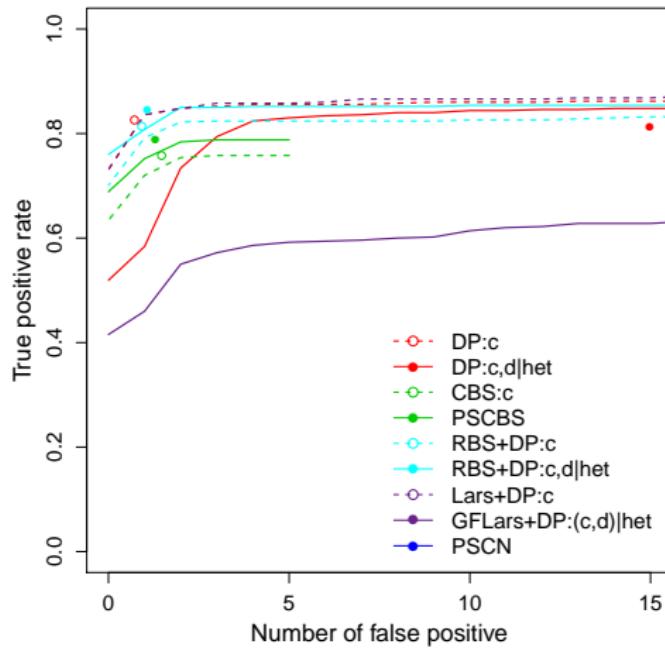
100 profiles, $n = 5000$, $K = 5$, purity = 79%, precision = 1



- 2d version of Fused Lars outperformed by 1d counterpart !
- reason : missing values in ' d ' not handled

Influence of the proportion of normal cells

100 profiles, $n = 5000$, $K = 5$, purity = 100%, precision = 1



Method ranking depends on %
of normal cells

Conclusions

- Compromis statistique/algorithme/biologie
- Méthodes très variées : programmation dynamique, segmentation binaire (CART), fused lasso
- Importance de l'évaluation des performances
- Importance de la reproductibilité des résultats (reproducible research)

Références : méthodes de segmentation

 K. Bleakley and J.-P. Vert.

The group fused lasso for multiple change-point detection.

Technical report, Mines ParisTech, 2011.

 Olshen AB et al.

Parent-specific copy number in paired tumor-normal studies using circular binary segmentation

Bioinformatics, (2011).

 S. Gey and E. Lebarbier.

Using CART to Detect Multiple Change Points in the Mean for Large Sample

Technical report, *Statistics for Systems Biology research group*, 2008.

 F. Picard and E. Lebarbier and M. Hoebeke and G. Rigaill and B. Thiam and S. Robin.

Joint segmenation, calling and normalization of multiple CGH profiles.

Biostatistics, 2011.

 Chen, H., Xing, H. and Zhang, N.R.

Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays.

PLoS Comput Biol, 2011.

Références : encore des méthodes de segmentation



G. Rigaill.

Pruned dynamic programming for optimal multiple change-point detection.
Technical report, <http://arXiv.org/abs/1004.0887>, 2010.



Olshen AB, Venkatraman ES, Lucito R, Wigler M.

Circular binary segmentation for the analysis of array-based DNA copy number data.
Biostatistics, (2004).



Zhang, Nancy R. and Siegmund, David O. and Ji, Hanlee and Li, Jun Z.

Detecting simultaneous changepoints in multiple sequences.
Biometrika, (2010)



Lai, Tze Leung and Xing, Haipeng and Zhang, Nancy

Stochastic segmentation models for array-based comparative genomic hybridization data analysis.
Biostatistics, (2008)



Z. Harchaoui and C. Lévy-Leduc.

Catching change-points with lasso.

Advances in Neural Information Processing Systems, 2008.