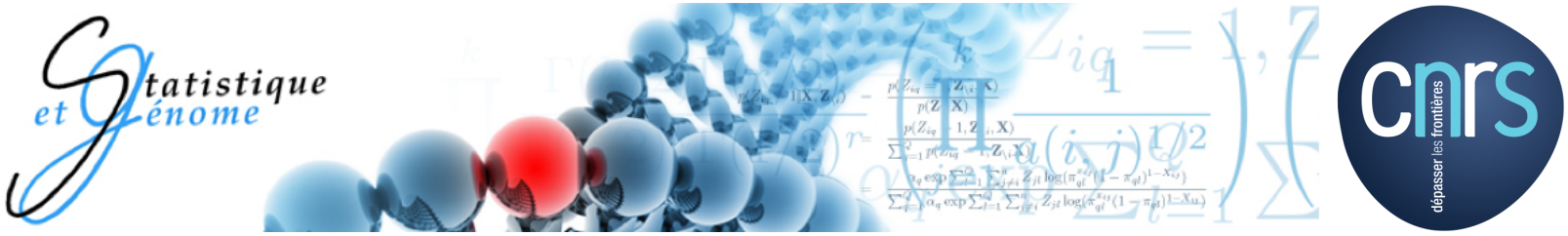


Méthodes statistiques pour l'analyse de données génomiques

Pierre Neuvial

Laboratoire de Mathématiques et Modélisation d'Evry
Équipe Statistique et Génome
Université d'Evry-Val-d'Essonne, UMR CNRS 8071 - USC INRA
`http://stat.genopole.cnrs.fr/~pneuvial`

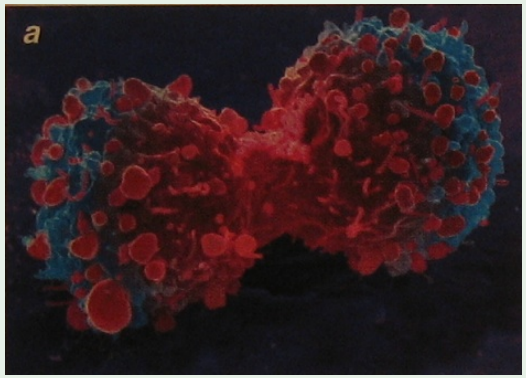
M2 BIBS Orsay — 2014/2015

Problématiques statistiques en cancérologie

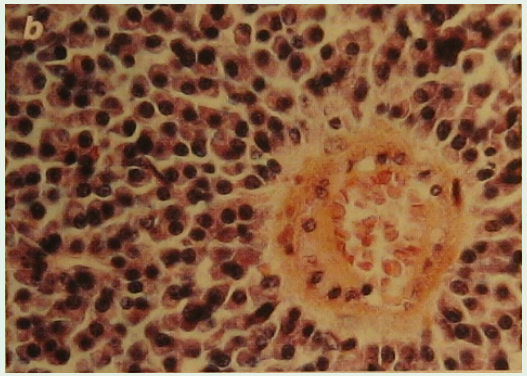
Dysfonctionnement des mécanismes de régulation

Caractéristiques des cellules tumorales

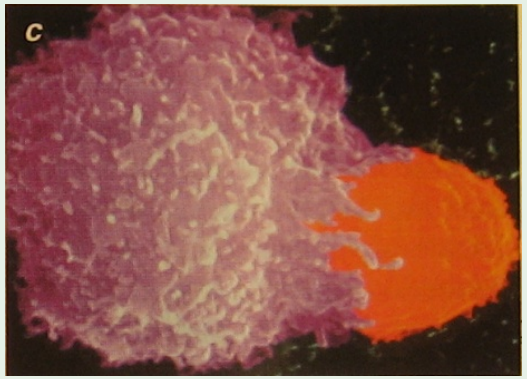
Hanahan & Weinberg, 2000



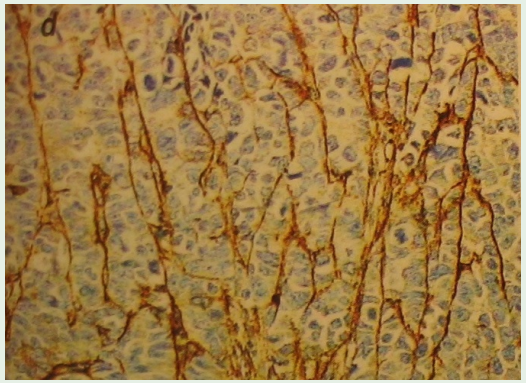
auto-suffisance en facteurs de croissance



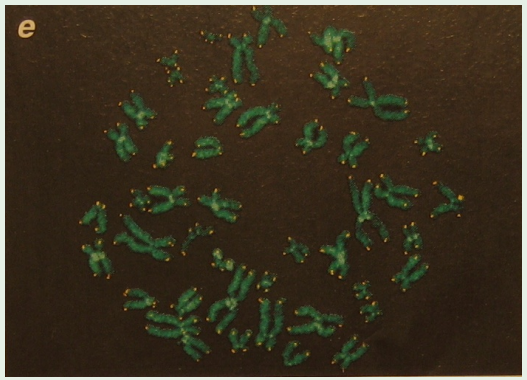
insensibilité aux signaux anti-croissance



pas d'apoptose (= mort programmée)



angiogénèse



pouvoir de réplication illimité

invasion des tissus, métastases

Acquises via la forte **instabilité génétique** des cellules tumorales

Le cancer, une maladie des gènes

Tous les cancers s'accompagnent de désordres génétiques

- formes familiales : existence de gènes de prédisposition
- formes sporadiques : accumulation d'événements génétiques

caryotype normal

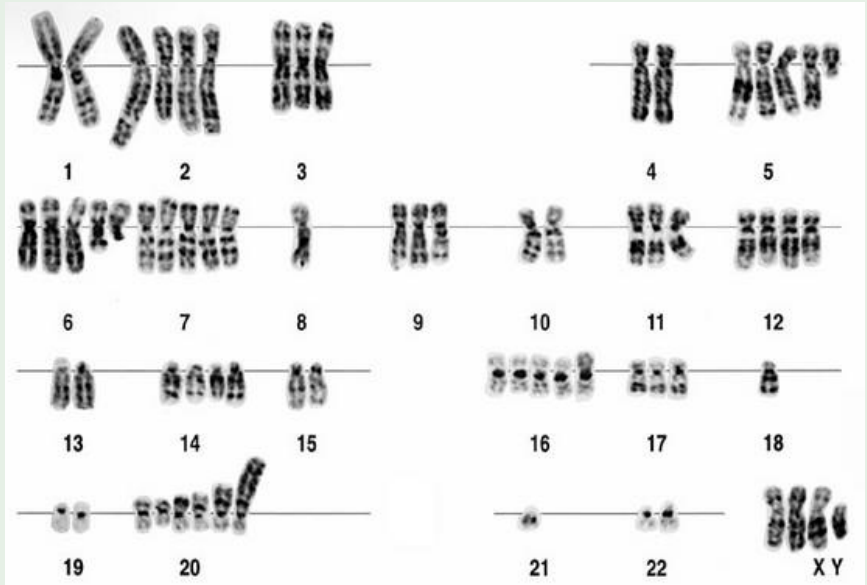
Le cancer, une maladie des gènes

Tous les cancers s'accompagnent de désordres génétiques

- formes familiales : existence de gènes de prédisposition
- formes sporadiques : accumulation d'événements génétiques

caryotype normal

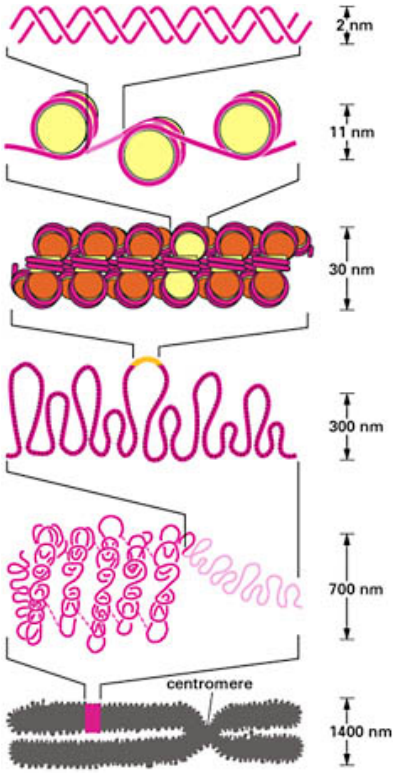
caryotype d'une tumeur



Chromosomes et ADN

Compaction de l'information

- le chromosome, structure de stockage de l'ADN
- 2m d'ADN dans chaque cellule (de diamètre 10 à 20μm)



Différents degrés de compaction

nucléosomes : enroulement de 100 bp autour de protéines, les histones

hélice de nucléosomes ($\delta = 30nm$)

boucles ancrées sur un échafaudage protéique

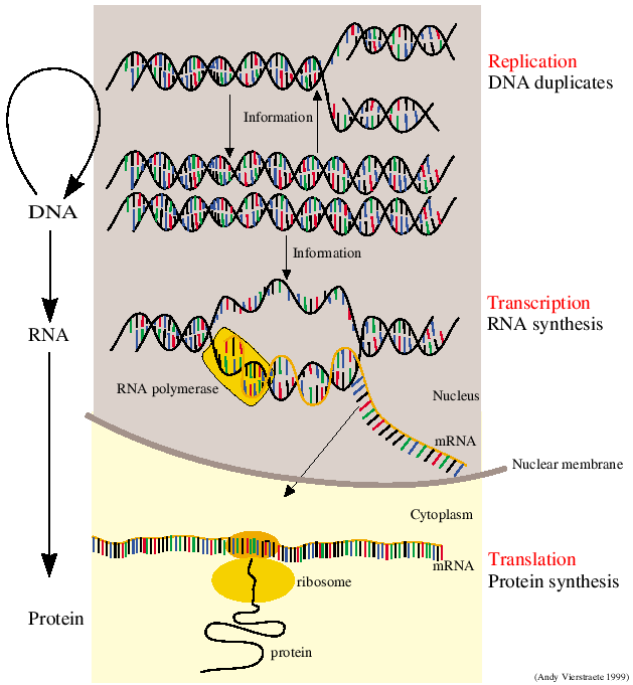
bras de chromosome : super-hélice

Dogme central de la biologie moléculaire

Mécanismes de stockage, de réplication, et d'expression de l'information biologique

Les protéines, unités fonctionnelles essentielles de la cellule

- assurent les principales fonctions cellulaires
- produites par les **gènes**



Dans le noyau

ADN→ADN réplication

ADN→ARN transcription

Hors du noyau

ARN→ protéine traduction

Questions d'intérêt

Questions biologiques

- mécanismes de l'oncogénèse (formation des cancers)
- inférence de fonction des gènes impliqués

Questions cliniques

- aide au diagnostic (types de cancers)
- prédiction du pronostic (survie, métastases)
- prédiction de la réponse au traitement

Niveaux d'étude de la génomique fonctionnelle

génome : analyse de séquences, altérations (épi)génétiques

transcriptome : expression des gènes et des petits ARN

protéome : étude des fonctions et interactions des protéines

Problématiques statistiques en cancérologie

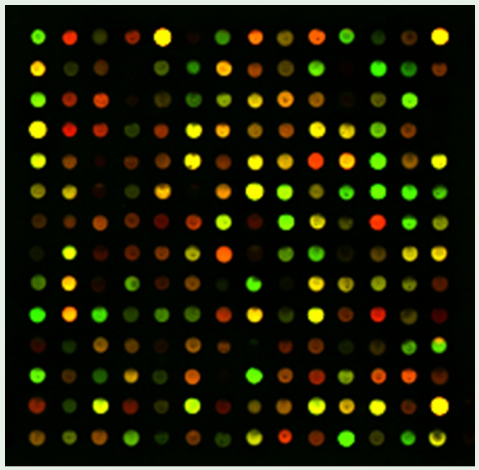
Technologie : puces à ADN

Mesure parallèle de nombreuses variables biologiques

Utilisation des propriétés physico-chimiques de l'ADN :

- l'ADN de sondes connues est fixé à un support
- des molécules d'ADN cibles s'hybrident de façon spécifique à leur complémentaire

biopuce scannée

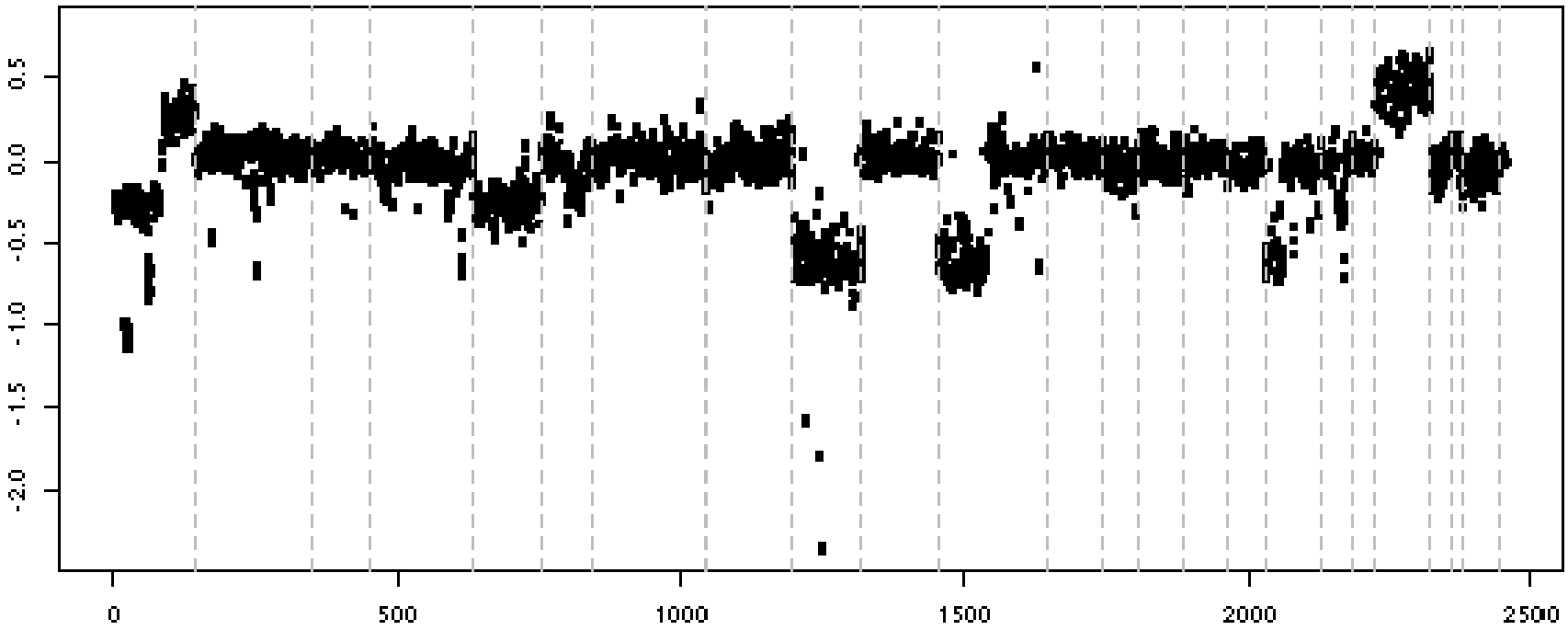


Caractéristiques de ces techniques

- expériences coûteuses (temps et argent)
- données de grande dimension ($n \ll p$)
- variabilité expérimentale importante

Altérations du nombre de copies d'ADN

Cartographie des variations du nombre de copies d'ADN



Données de puces à hybridation génomique comparative (CGH) pour un patient atteint de cancer de la vessie.
Ici : 2400 marqueurs pour chaque patient

Altérations du nombre de copies d'ADN

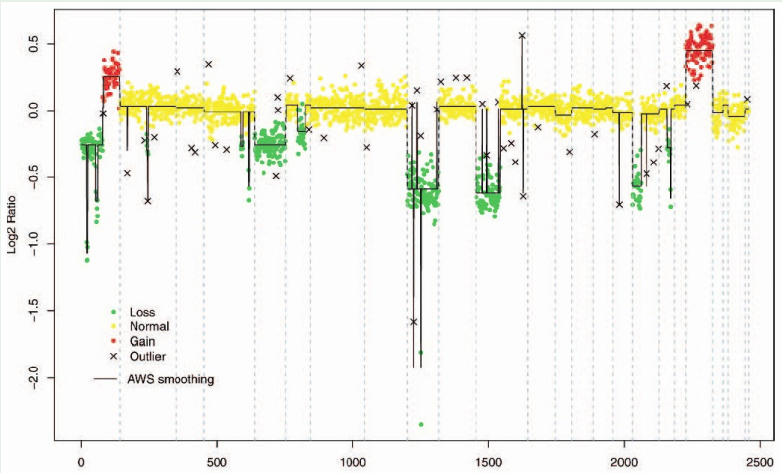
Cartographie des variations du nombre de copies d'ADN

Types d'analyses statistiques

1. **extraction du signal biologique**
2. **détection de ruptures**
3. **étiquetage des régions altérées**
4. recherche d'altérations récurrentes
5. classification non supervisée
6. classification supervisée

Altérations du nombre de copies d'ADN

Objectif : discrétisation des données



Étapes :

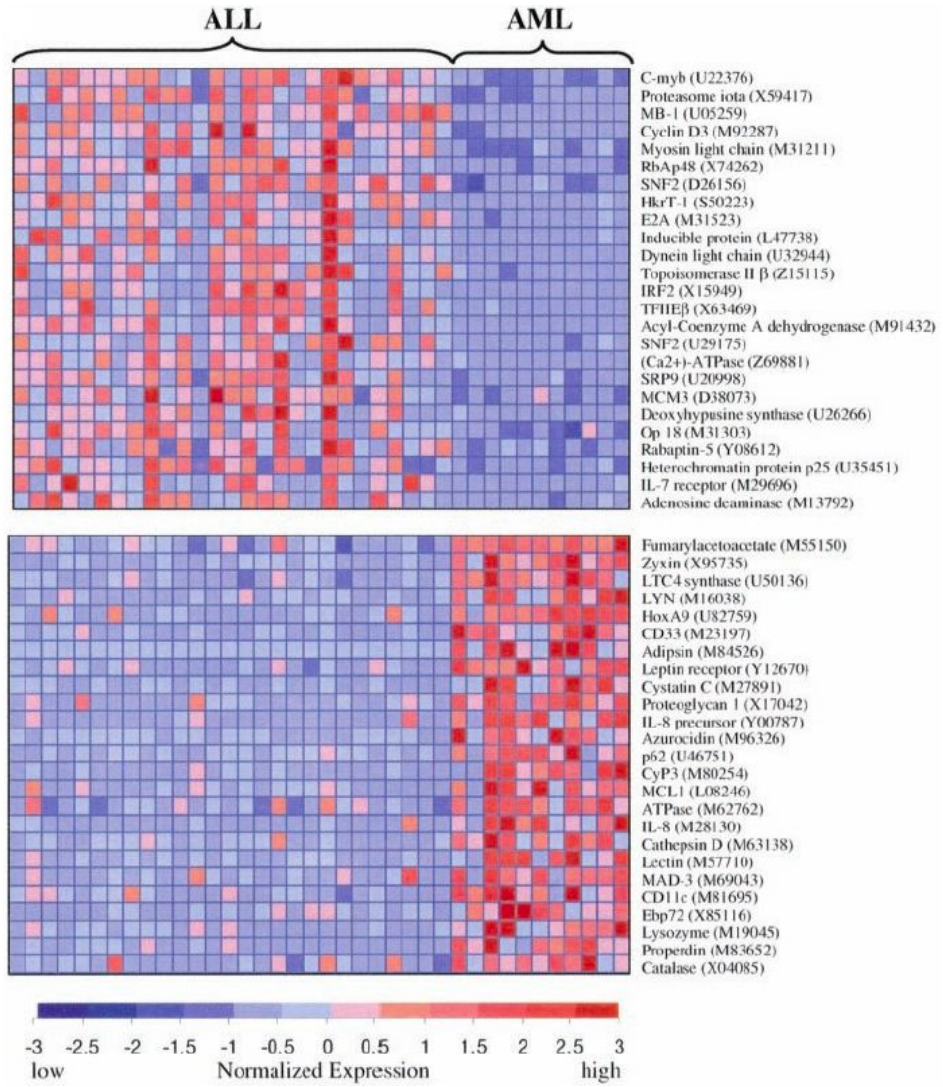
1. détection des ruptures
2. attribution d'un statut aux régions obtenues : gain, normal, perte

Enjeux

- choix d'un modèle statistique
- prise en compte de la dépendance spatiale
- choix d'un nombre de ruptures : sélection de modèles
- efficacité de l'algorithme

Altérations des niveaux d'expression

Quantification du nombre de transcrits



Niveau d'expression de 50 gènes (choisis parmi 3051) chez 27+11 patients atteints de deux types de leucémie.

Altérations des niveaux d'expression

Quantification du nombre de transcrits

Types d'analyses statistiques

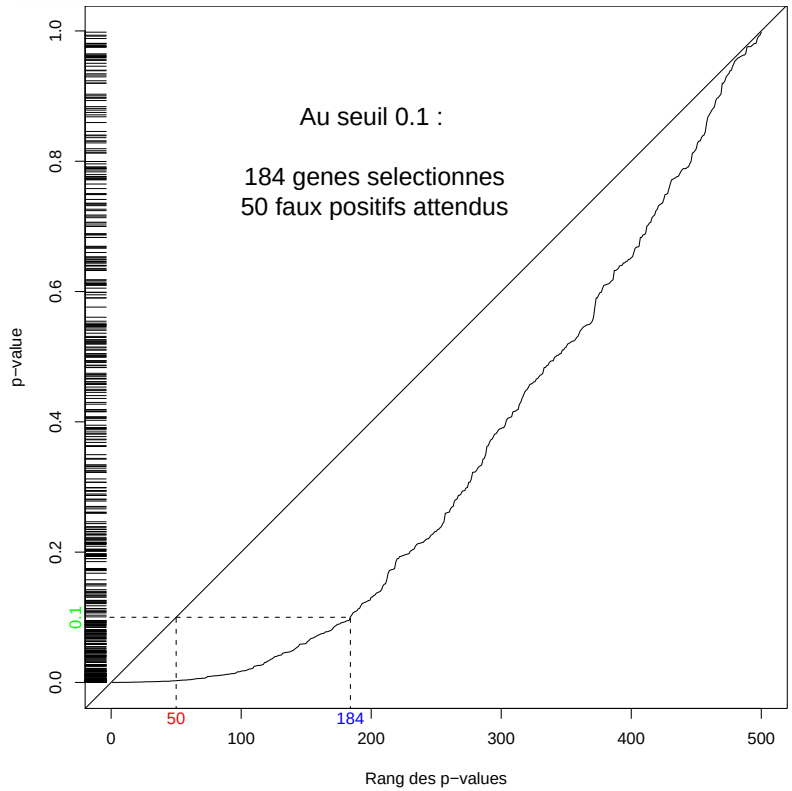
1. extraction du signal biologique
2. **tests d'expression différentielle**
3. classification non supervisée
4. classification supervisée
5. prédiction de la survie
6. inférence de réseaux de régulation

Expression différentielle et tests multiples

Objectif : identifier les gènes dont l'expression diffère entre deux groupes

Utilisation de la théorie des tests statistiques

- calcul d'une statistique de test (pour chaque gène)
- contrôle du taux de faux positifs et/ou de faux négatifs



Enjeux

- choix d'une statistique de test
- problèmes de tests multiples
- interprétation des résultats

Classification et régression en génomique

Prédiction d'une variable clinique ou biologique

- distinguer les individus sains de ceux atteints par une maladie
- prédire la réponse thérapeutique ou la survie
- inférer un réseau d'interactions entre gènes

Approche classique en classification/régression

1. apprentissage d'une règle sur un ensemble d'entraînement
2. application de cette règle à un nouvel échantillon

Enjeux en génomique

- nécessité d'une sélection de variables
- importance d'une méthode de validation objective
- interprétabilité du prédicteur

III. Classification et régression en génomique

9. Prédiction de classe (classification supervisée)
   - Introduction
   - Exemples de méthodes de classification

10. Questions statistiques soulevées
   - Estimation des performances
   - Réduction de dimension pour la classification
   - Sélection de variables par pénalisation

11. Performance, stabilité, interprétabilité des signatures moléculaires
   - Construction d'une signature moléculaire
   - Méthodes comparées
   - Critères pour l'évaluation d'une signature moléculaire
   - Résultats

Objectifs du cours

Des problématiques apparemment classiques en statistique :

- détection de ruptures
- tests d'hypothèses
- classification, régression

Spécificités des données génomiques

- grande dimension : bien plus de variables que d'observations
- complexité : issues d'expériences biologiques pointues
- structuration : réalité biologique sous-jacente

Besoin de méthodes statistiques tirant parti de ces spécificités

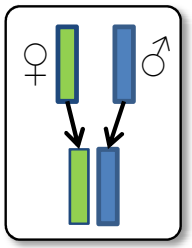
- modèles statistiques rigoureux
- résultats interprétables
- algorithmes efficaces

Analyse de données de nombre de copies d'ADN

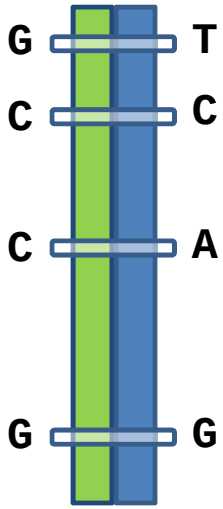
2. Puces SNP en cancérologie
   - Changements de nombre de copies d'ADN dans les cancers
   - Données de puces SNP
3. Extraction de l'information biologique
   - Pre-processing : des signaux comparables entre échantillons
   - Post-processing : nombre de copies totaux
   - Post-processing : ratios alléliques
4. Segmentation de données de nombre de copies d'ADN (total)
   - Recours aux modèles de rupture
   - Exemples d'approches proposées
   - Segmentation multi-échantillons ou multi plate-forme
5. Estimation du nombre de copies d'ADN
   - Détection : utilisation conjointe de C et DH
   - Étiquetage : cellules normales et ploïdie

Analyse de données de nombre de copies d'ADN

Génotypes dans un chromosome diploïde



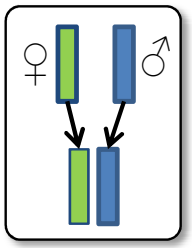
Single nucleotide polymorphism



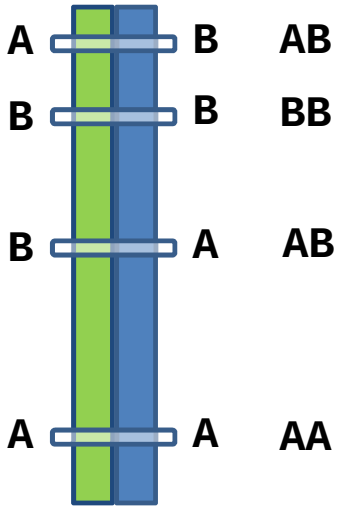
10-20 million
known SNPs

slide: H. Bengtsson.

Génotypes dans un chromosome diploïde



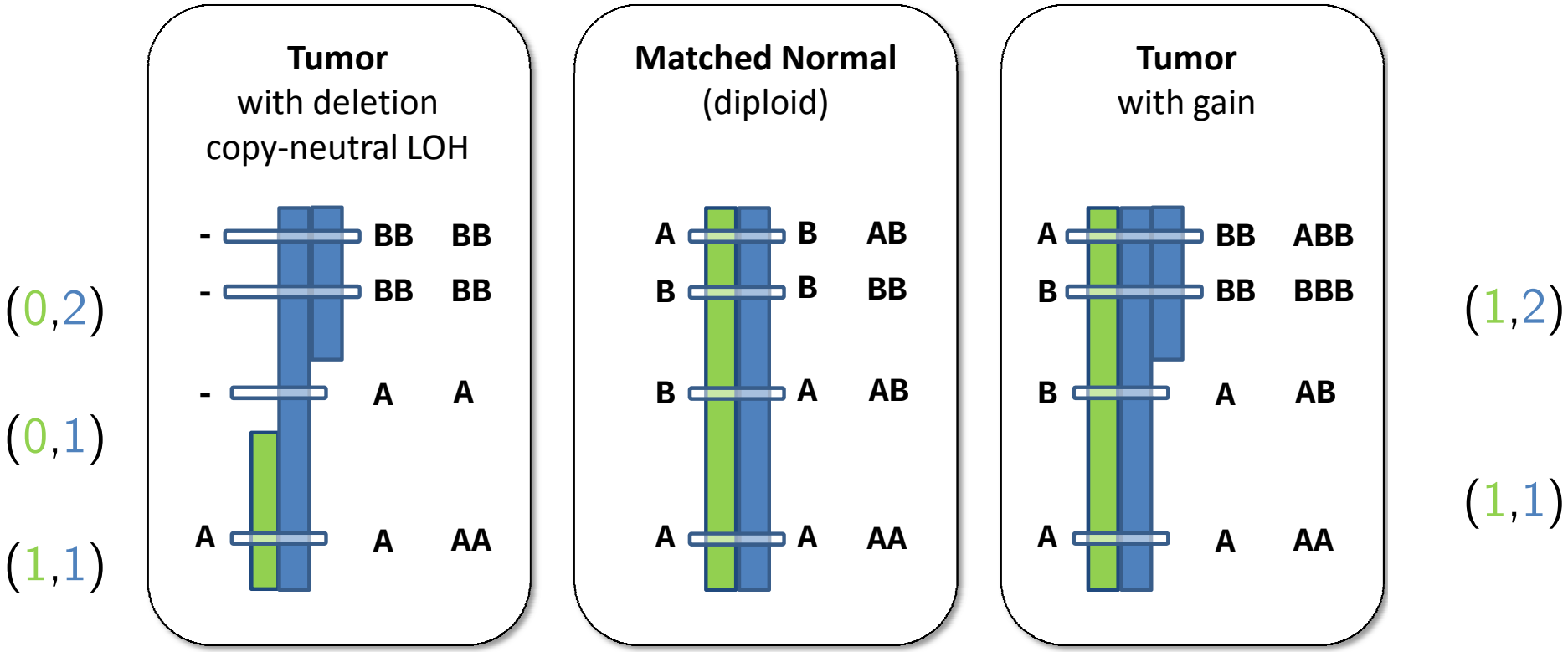
Single nucleotide polymorphism



10-20 million
known SNPs

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

$$CN = (C_{1j}, C_{2j}),$$

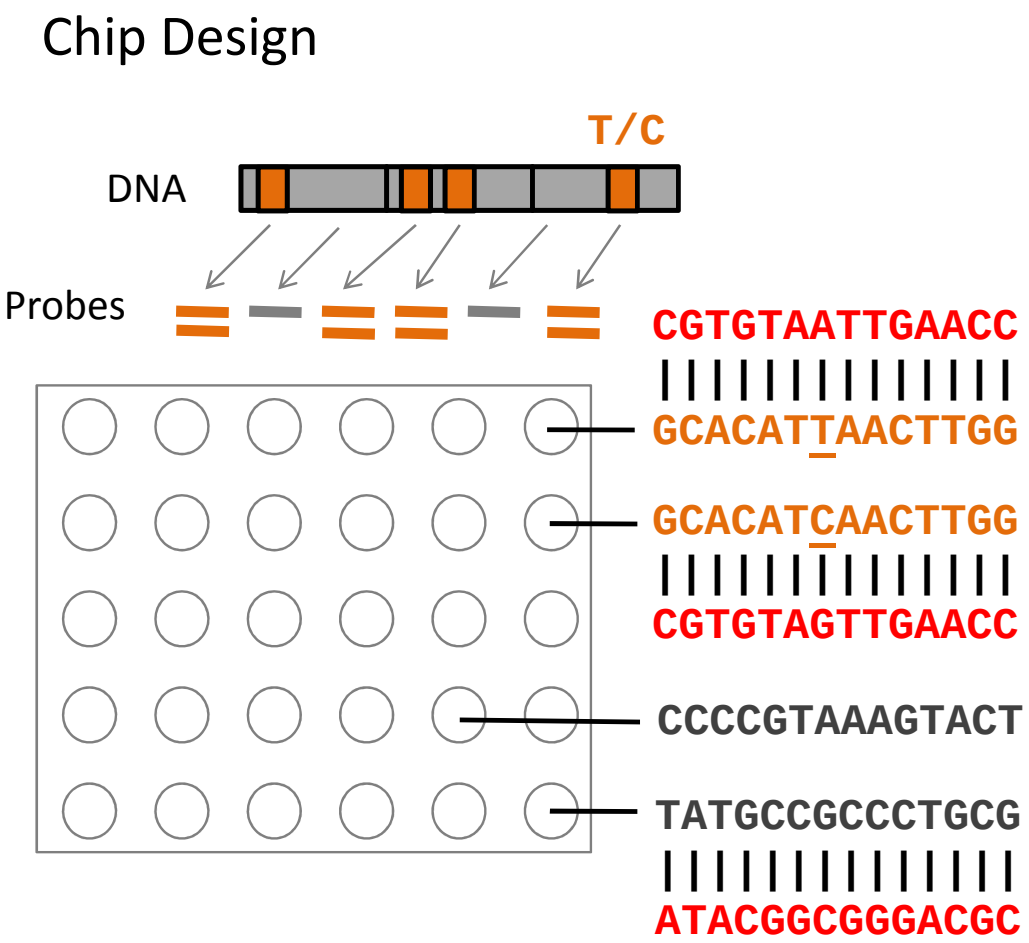
où $C_{1j} = \min(m_j, p_j)$ et $C_{2j} = \max(m_j, p_j)$.

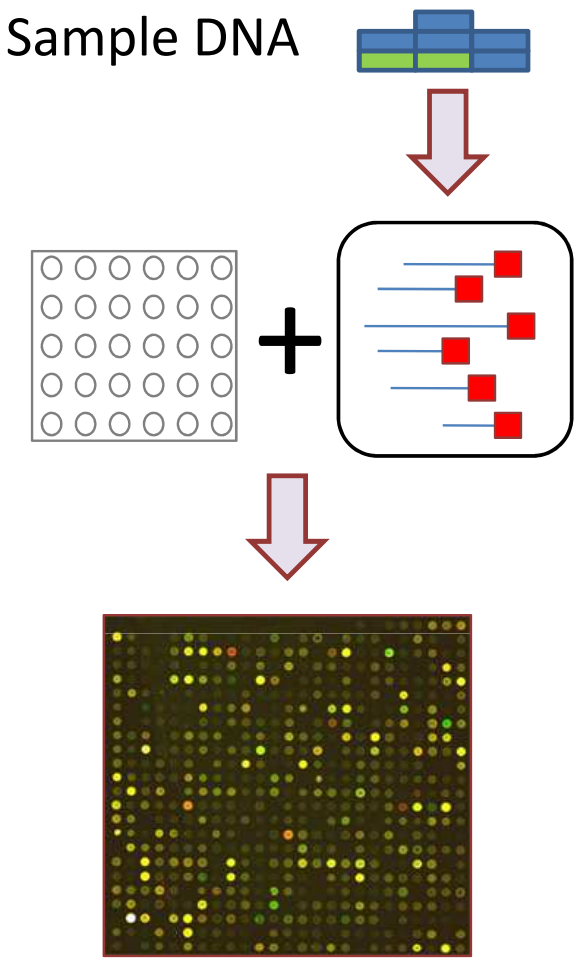
Les nombres de copies mineur (C_1) et majeur (C_2) :

- 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

Technologie : puces SNP et nombre de copies d'ADN





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

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

Diminution d'hétérozygotie

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

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) : $CN = (0, 2)$: deux copies au total
  - Duplication équilibrée : $CN = (2, 2)$, 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

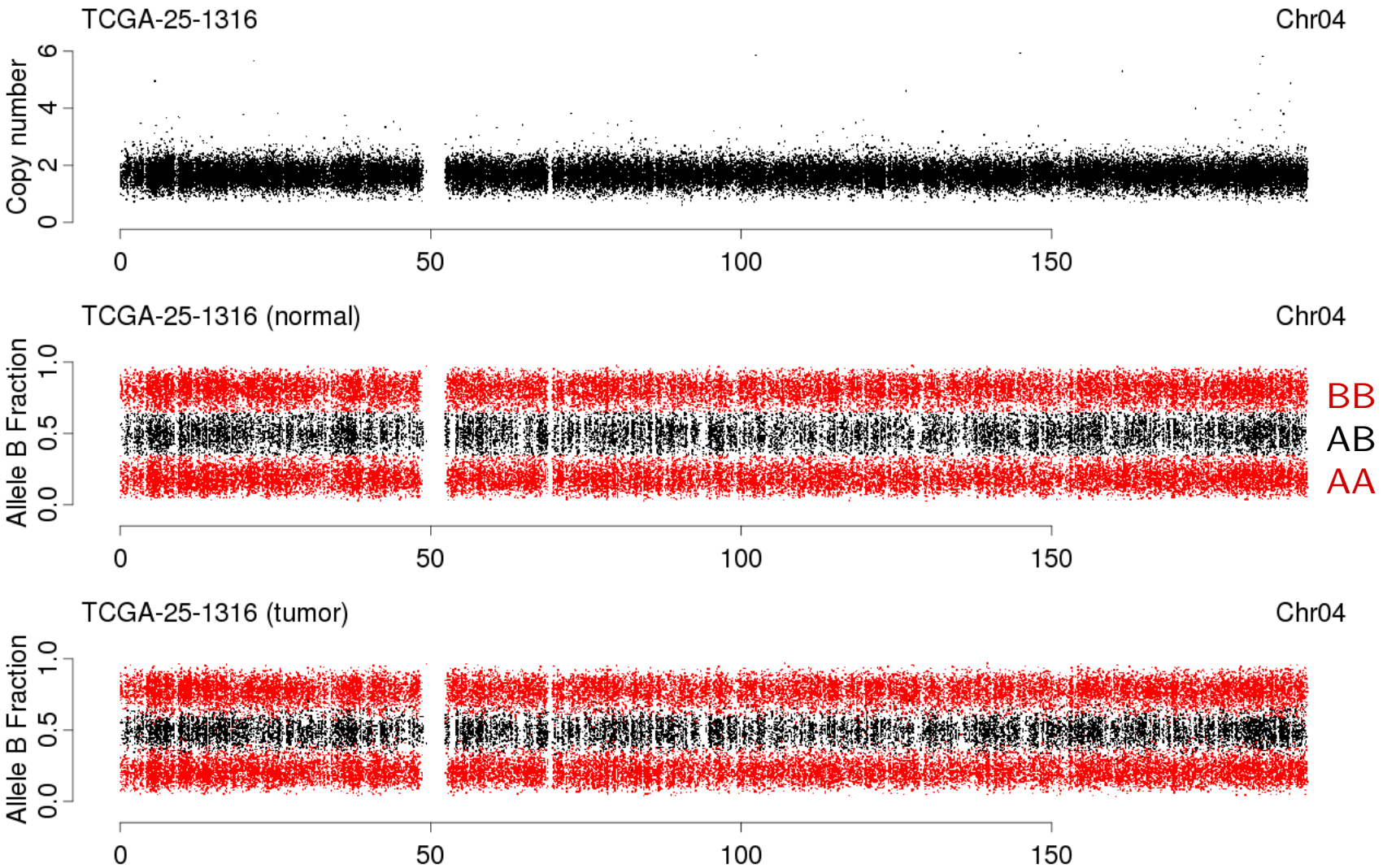
Pour les données de puces SNP, : identifier les **changements de nombre de copies d'ADN** : (C, DH) ou (C_1, C_2) :

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

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

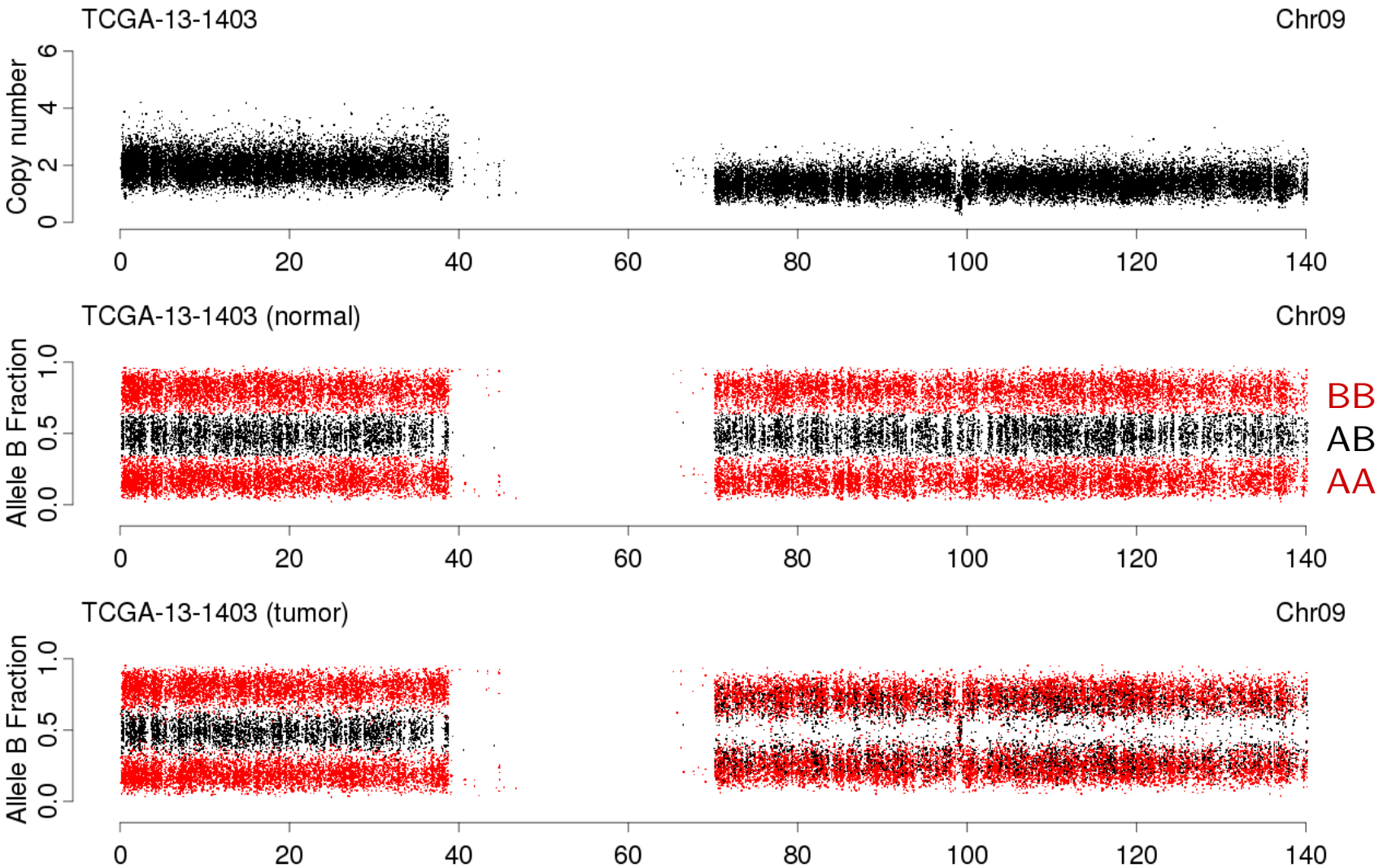
Analyse de données de nombre de copies d'ADN

Région normale : (1,1)



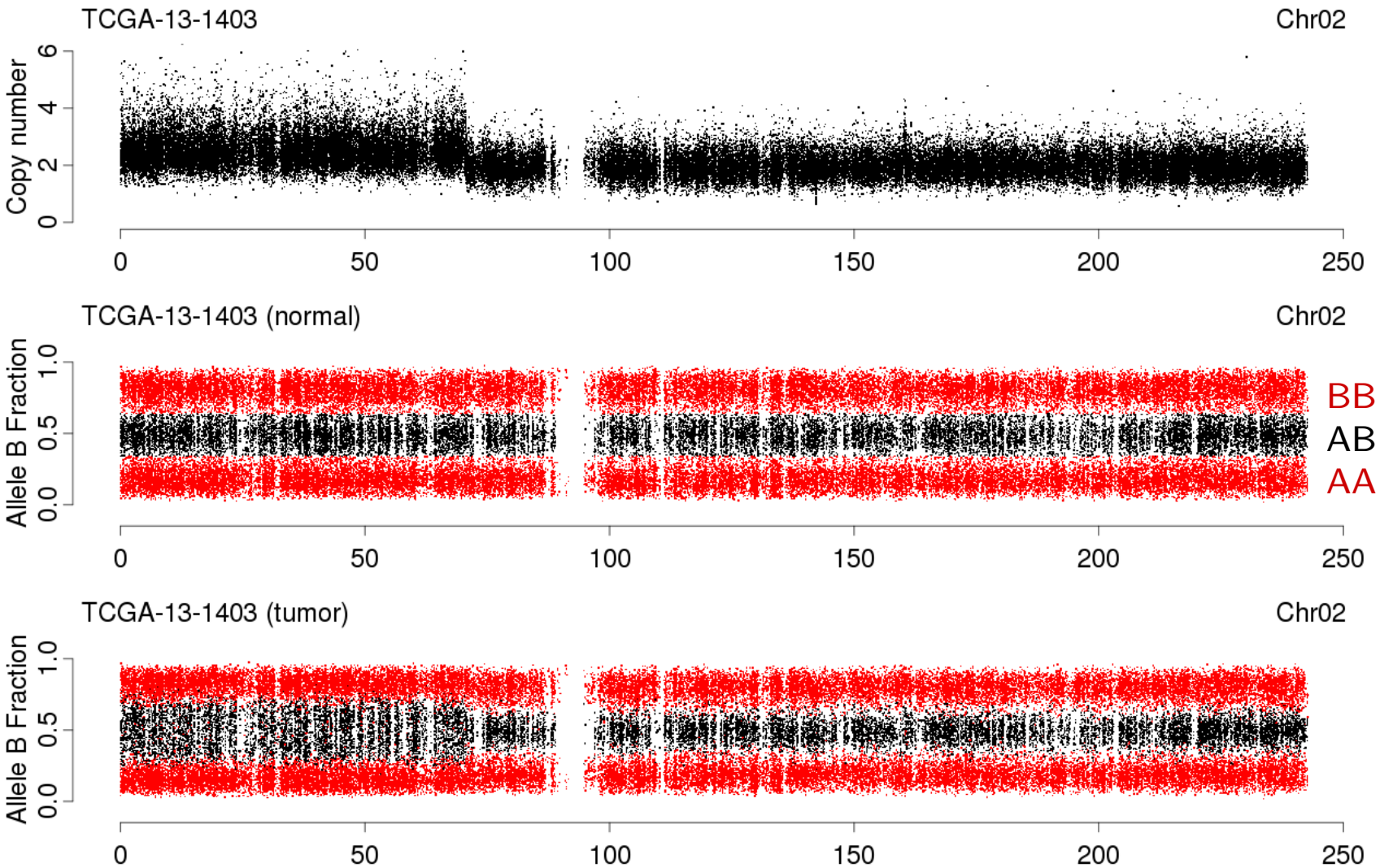
Les SNPs homozygotes dans l'échantillon normal sont en rouge.

Perte d'une copie : (0, 1)



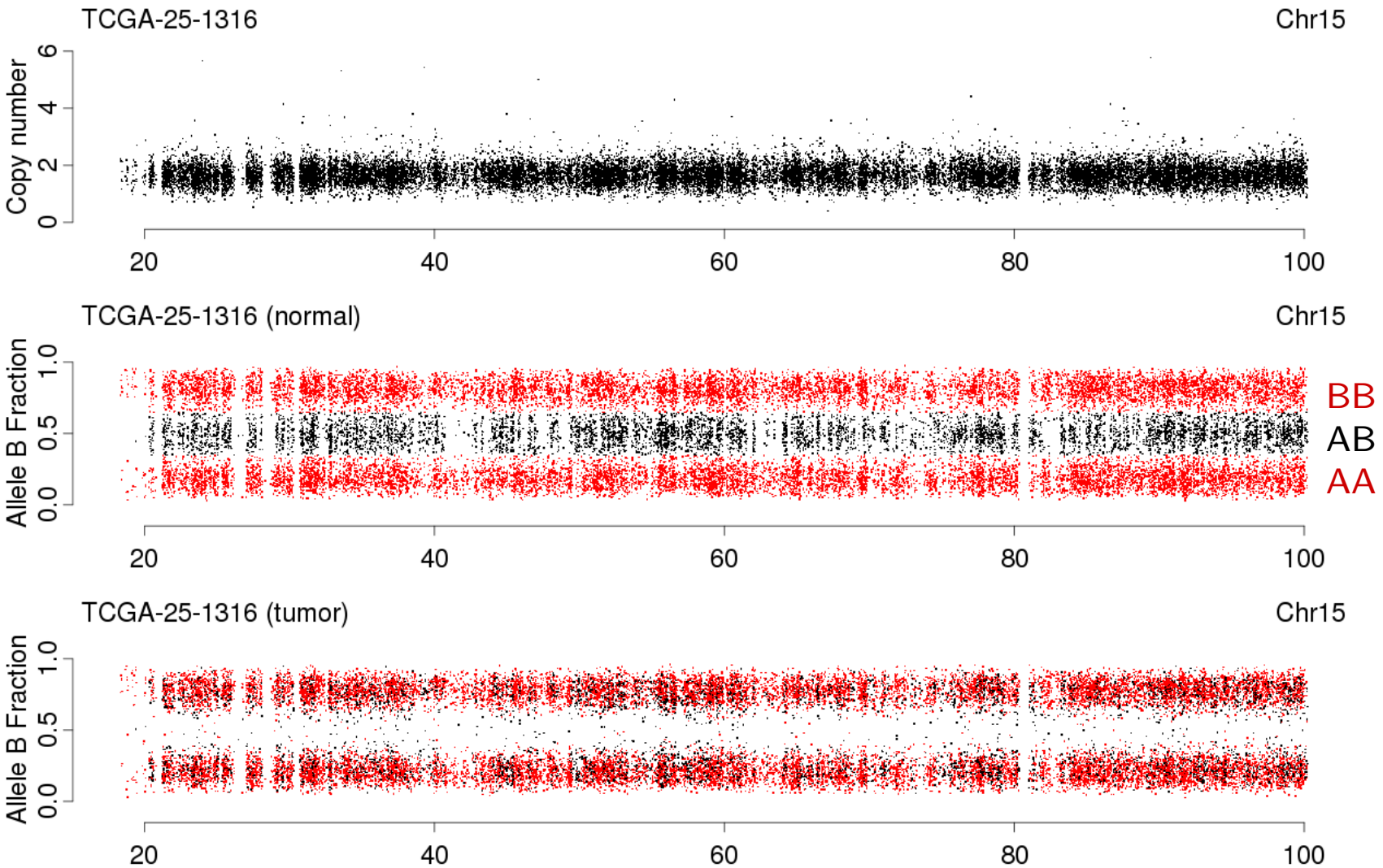
Les SNPs homozygotes dans l'échantillon normal sont en rouge.

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.

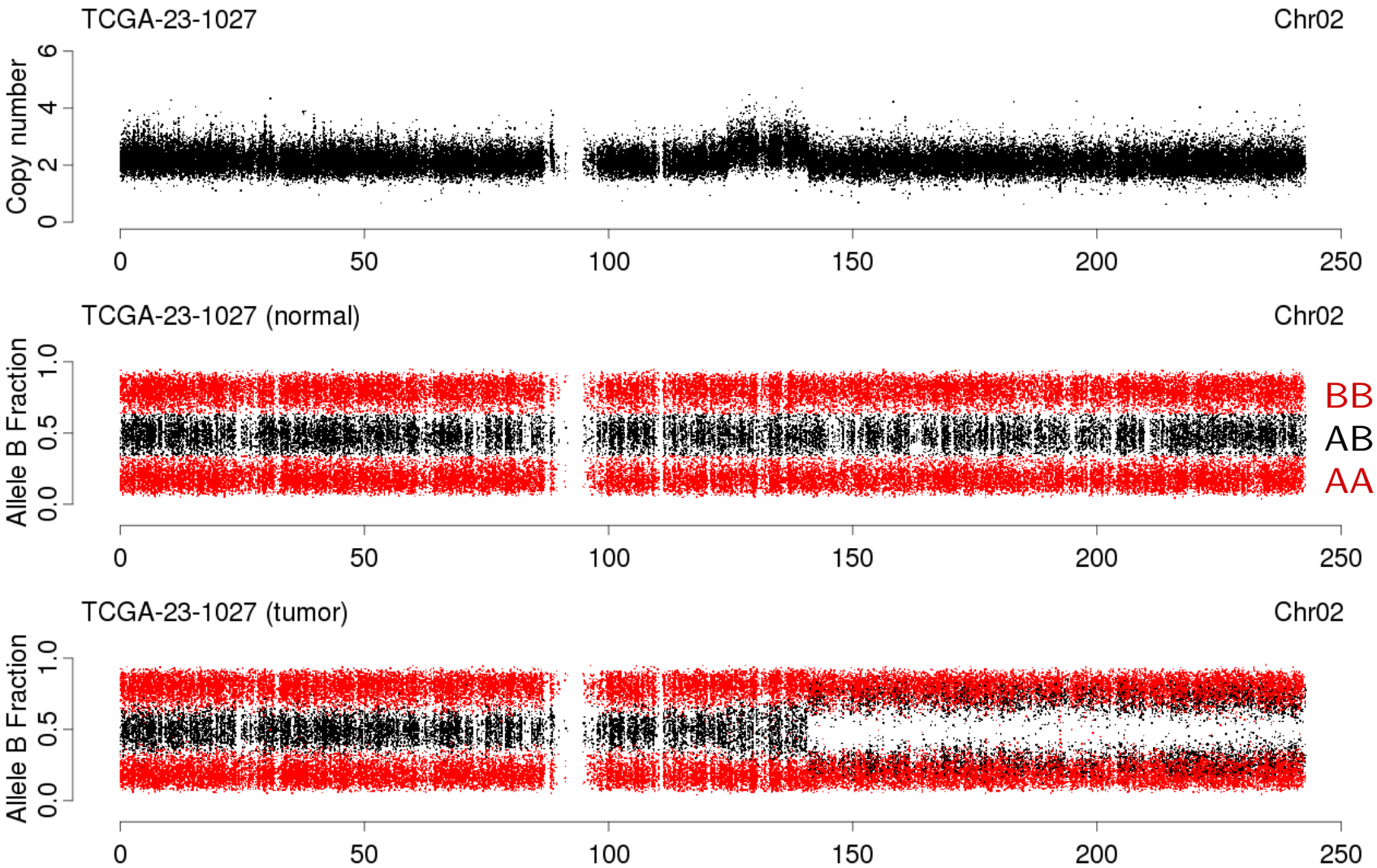
Présence de cellules normales

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.

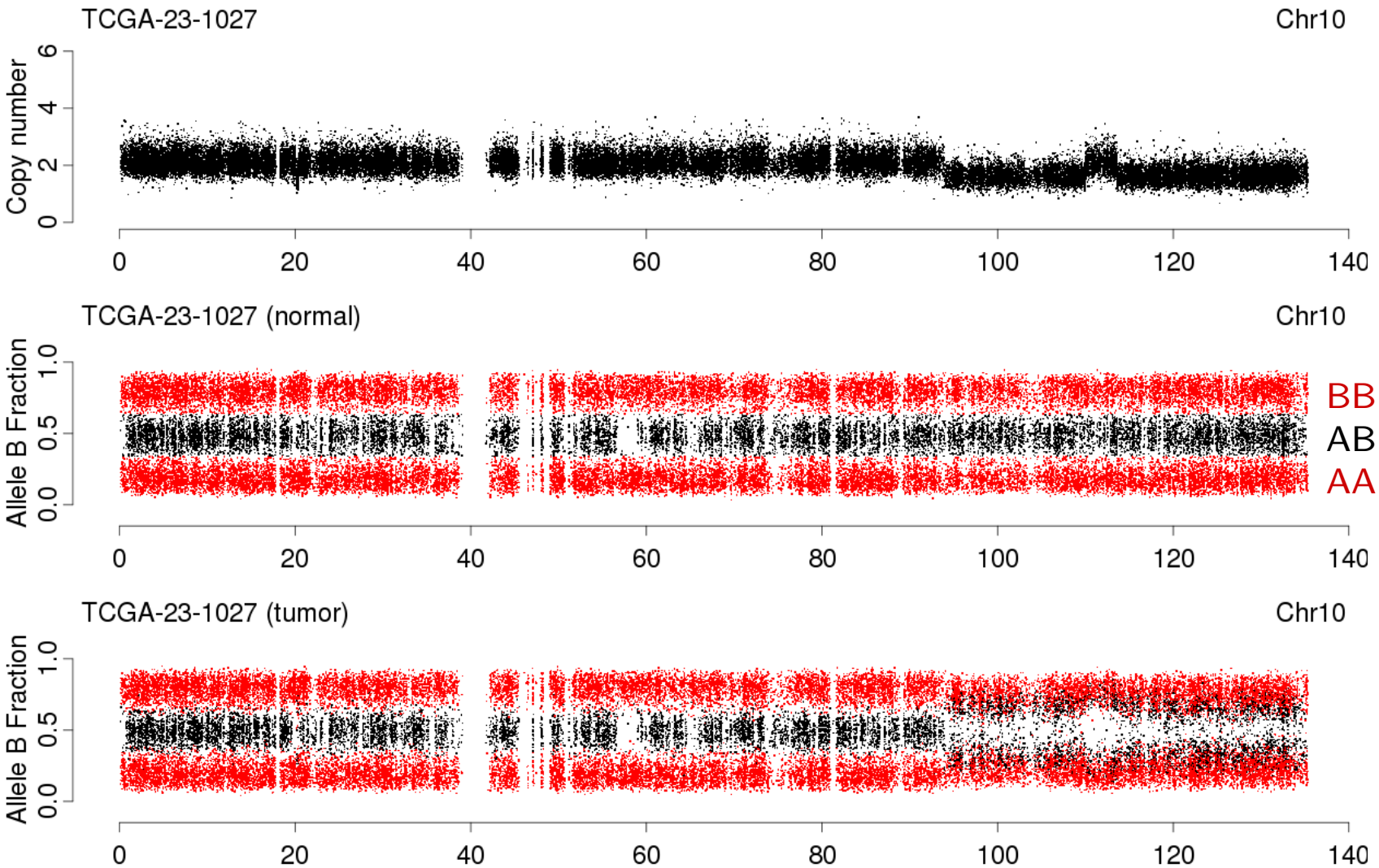
En présence de cellules normales, les ratios alléliques des SNPs hétérozygotes **se rapprochent de 1/2**.

Normal, gain, isodisomie



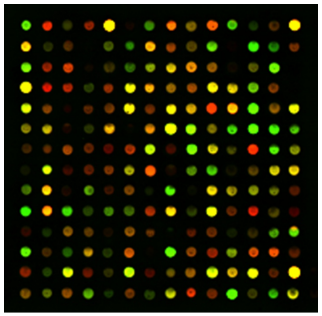
Les SNPs homozygotes dans l'échantillon normal sont en rouge.

Normal, perte, isodisomie



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

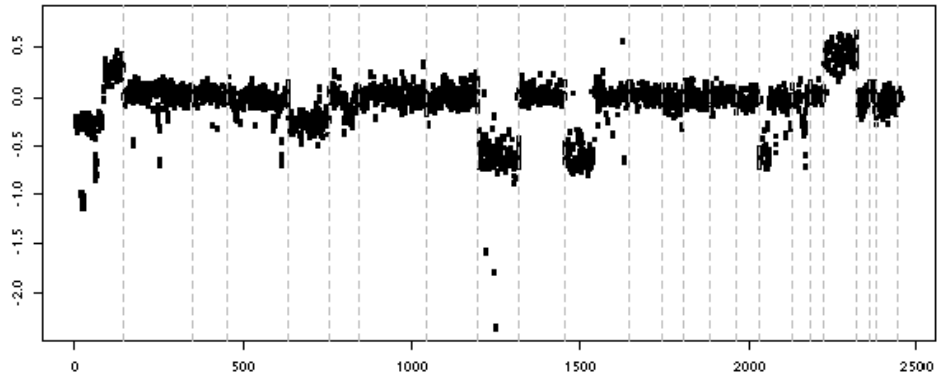
Normalisation : définition et objectifs



Analyse d'image

$\Rightarrow$

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

Copy-numbers by Robust Microarray Analysis (CRMA)

Une méthode de pre-processing applicable échantillon par échantillon

For each Affymetrix array (*i = 1,2,3,…,10000*) independently:

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

slide: H. Bengtsson.

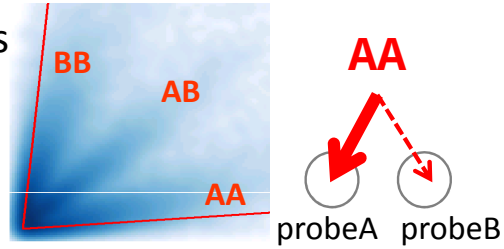
Explication de la variation systématique entre puces

Décalage (offset) du scanner et hybridation croisée entre allèles

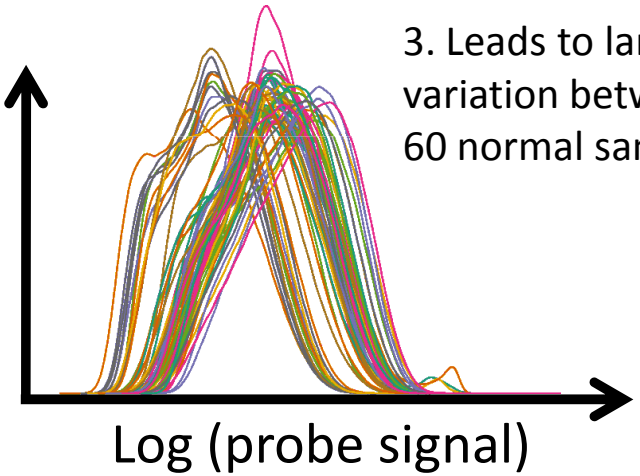
1. The scanner's shifts all probe signals (offset)

+

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

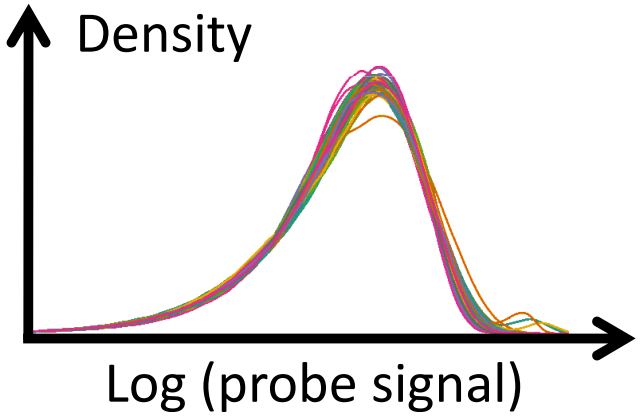


=



3. Leads to large variation between 60 normal samples

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



slide: H. Bengtsson.

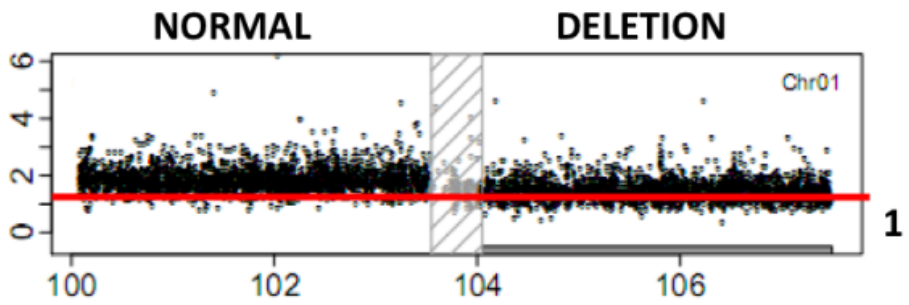
Évaluation ROC

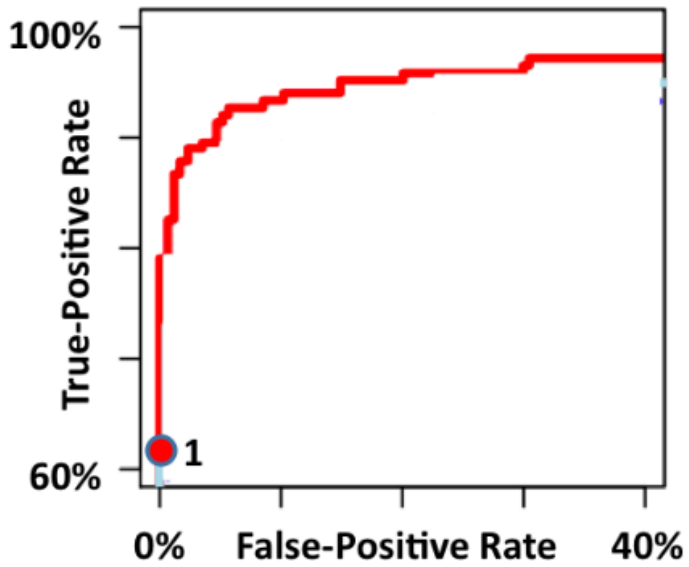
Pour un échantillon donné

1. identifier une rupture claire
2. étiqueter les régions voisines, par ex. NORMAL (1,1) et DELETION (0,1)
3. 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





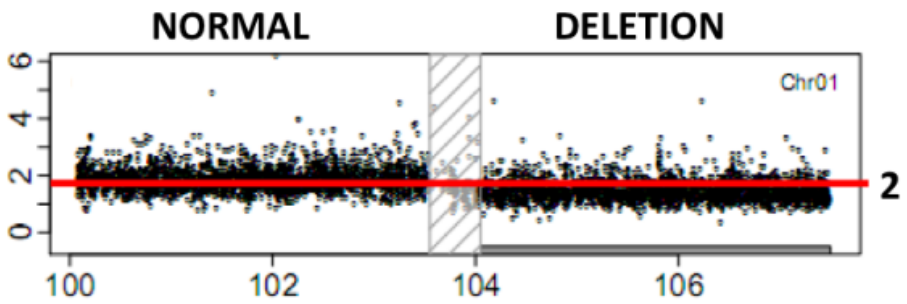
Évaluation ROC

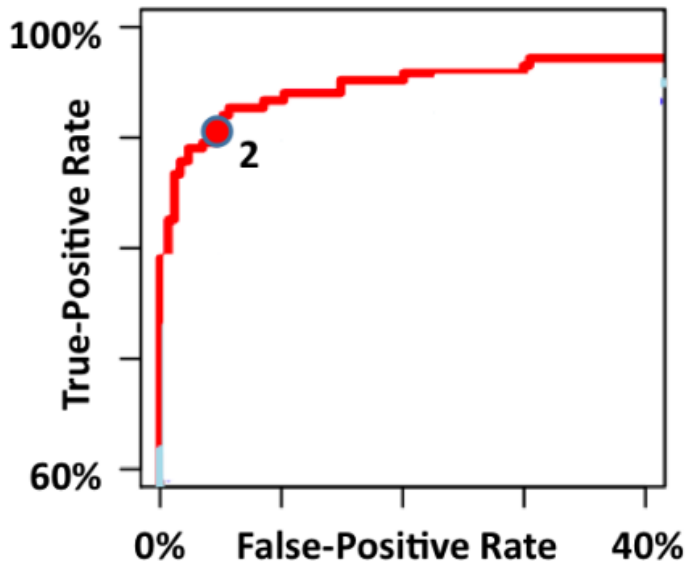
Pour un échantillon donné

1. identifier une rupture claire
2. étiqueter les régions voisines, par ex. NORMAL (1,1) et DELETION (0,1)
3. 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





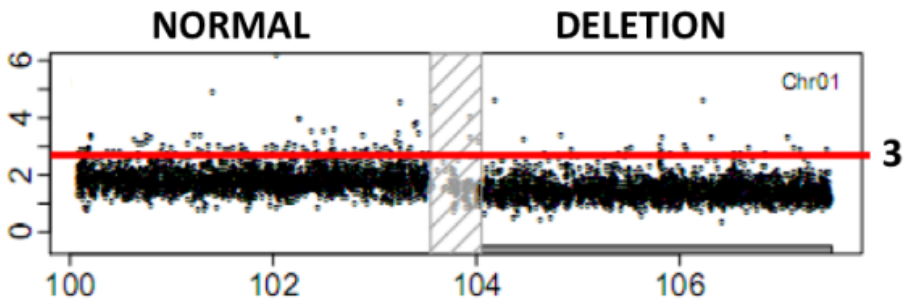
Évaluation ROC

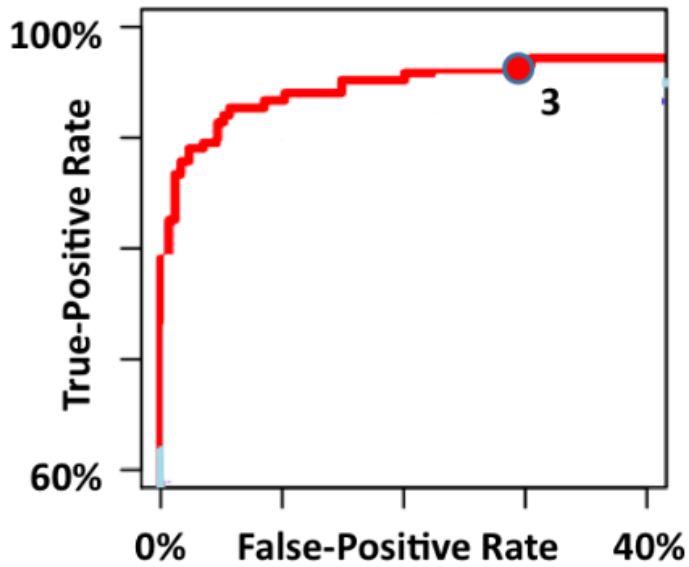
Pour un échantillon donné

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

Pour chaque valeur du seuil τ :

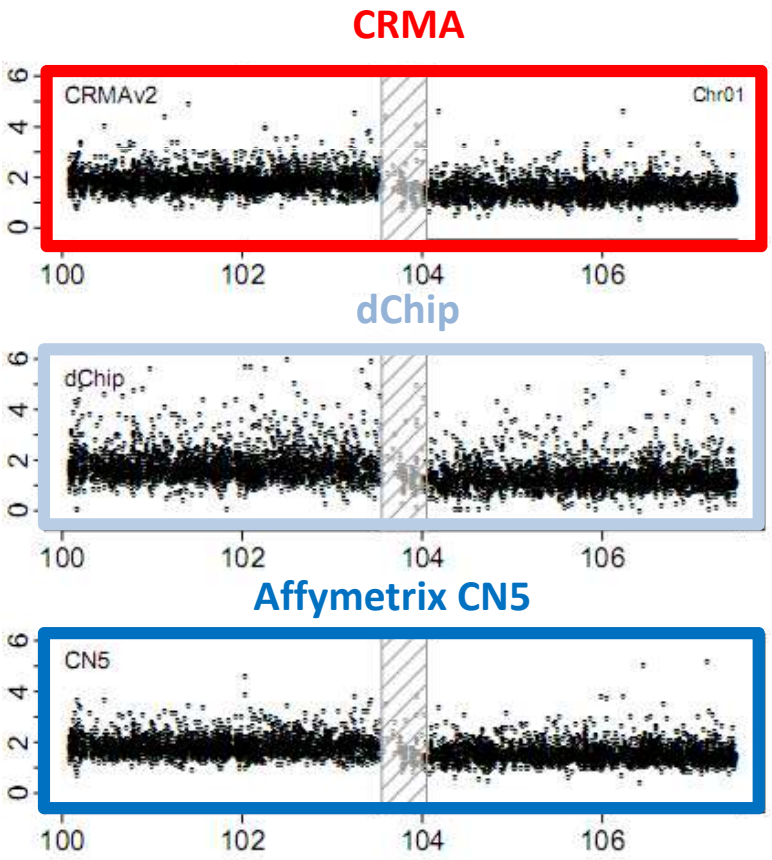
- Un SNP sous le seuil τ est appelé DELETION
- Le nombre d'erreurs définit un point sur la courbe ROC

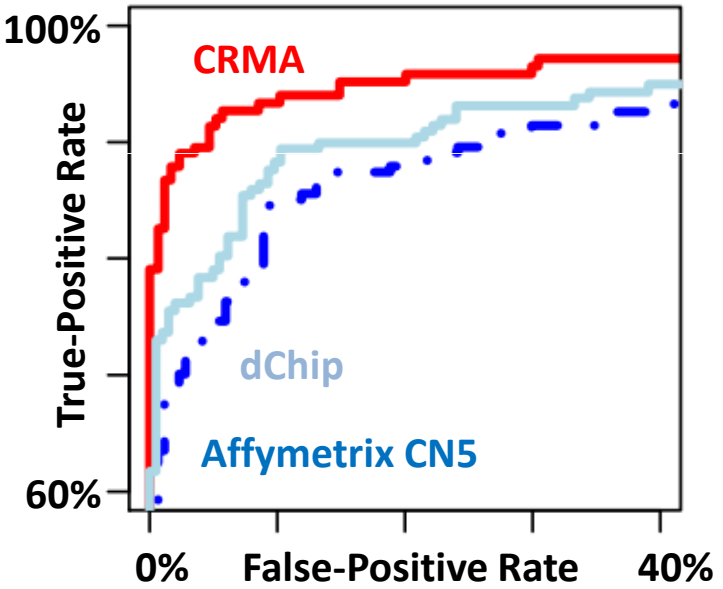




CRMA fait mieux que les méthodes multi-puces

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.

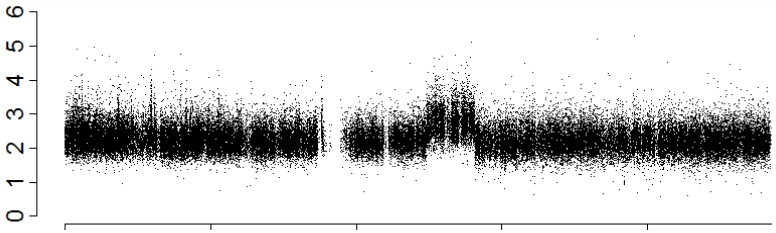
slide: H. Bengtsson.

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

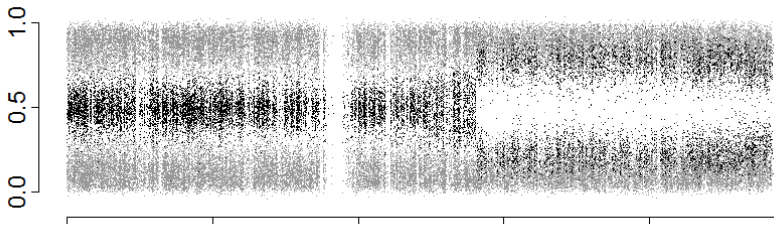
$$\begin{aligned} C_{ij} &= 2\frac{\theta_{ij}}{\theta_{Rj}} \\ &= C_{1ij} + C_{2ij} \end{aligned}$$



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

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

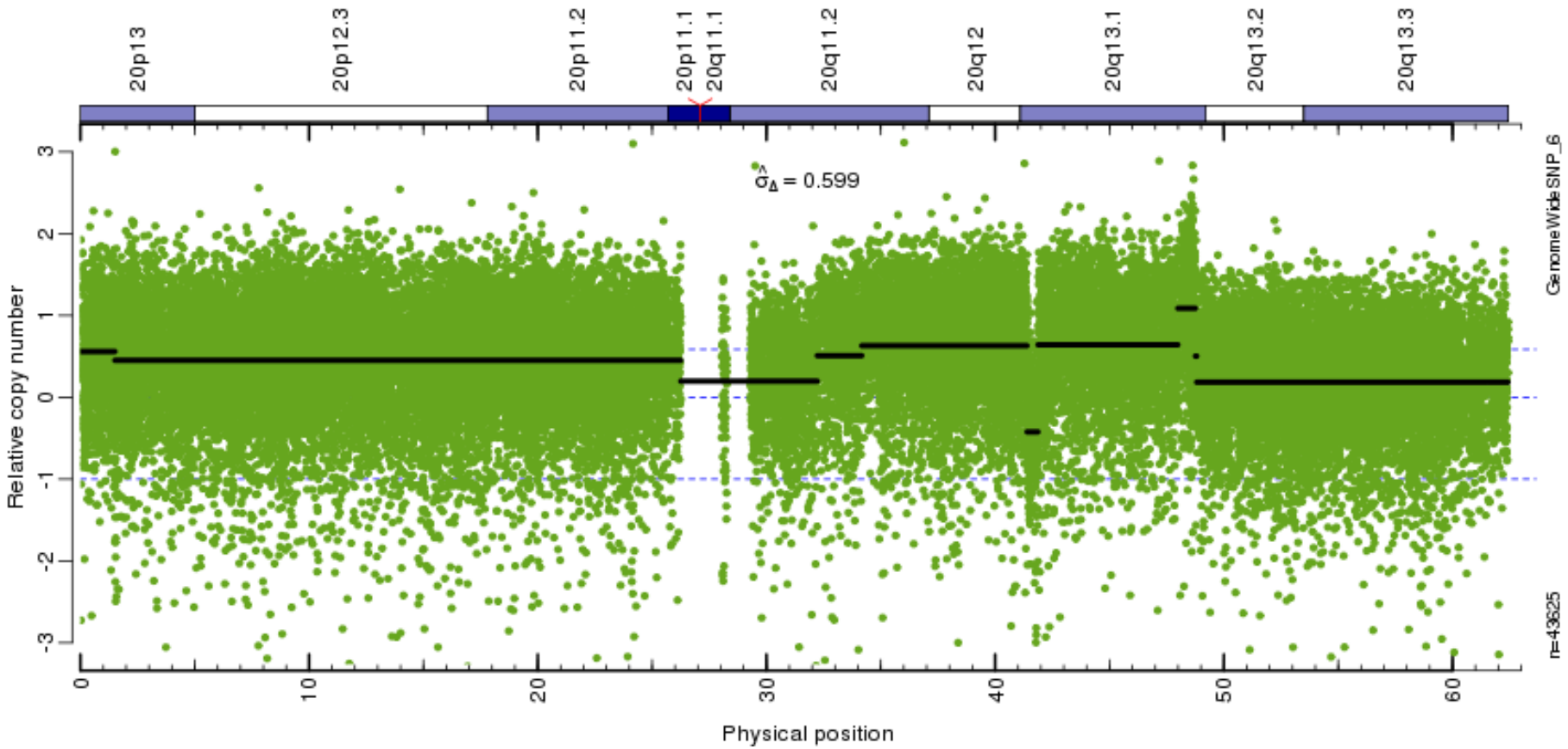
Quantification du niveau de bruit :

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

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

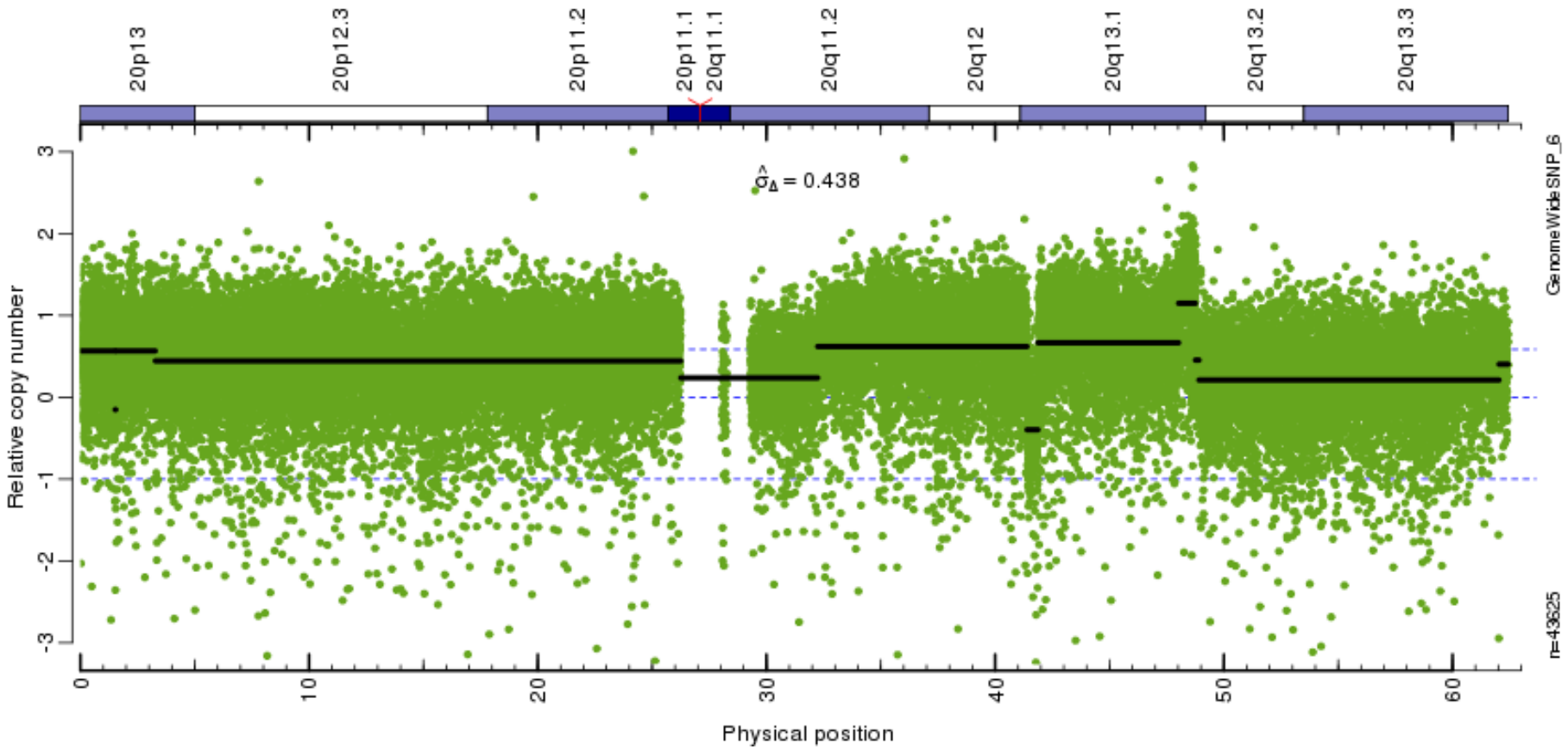
Choix d'une référence

Laboratoire différent (n=192)



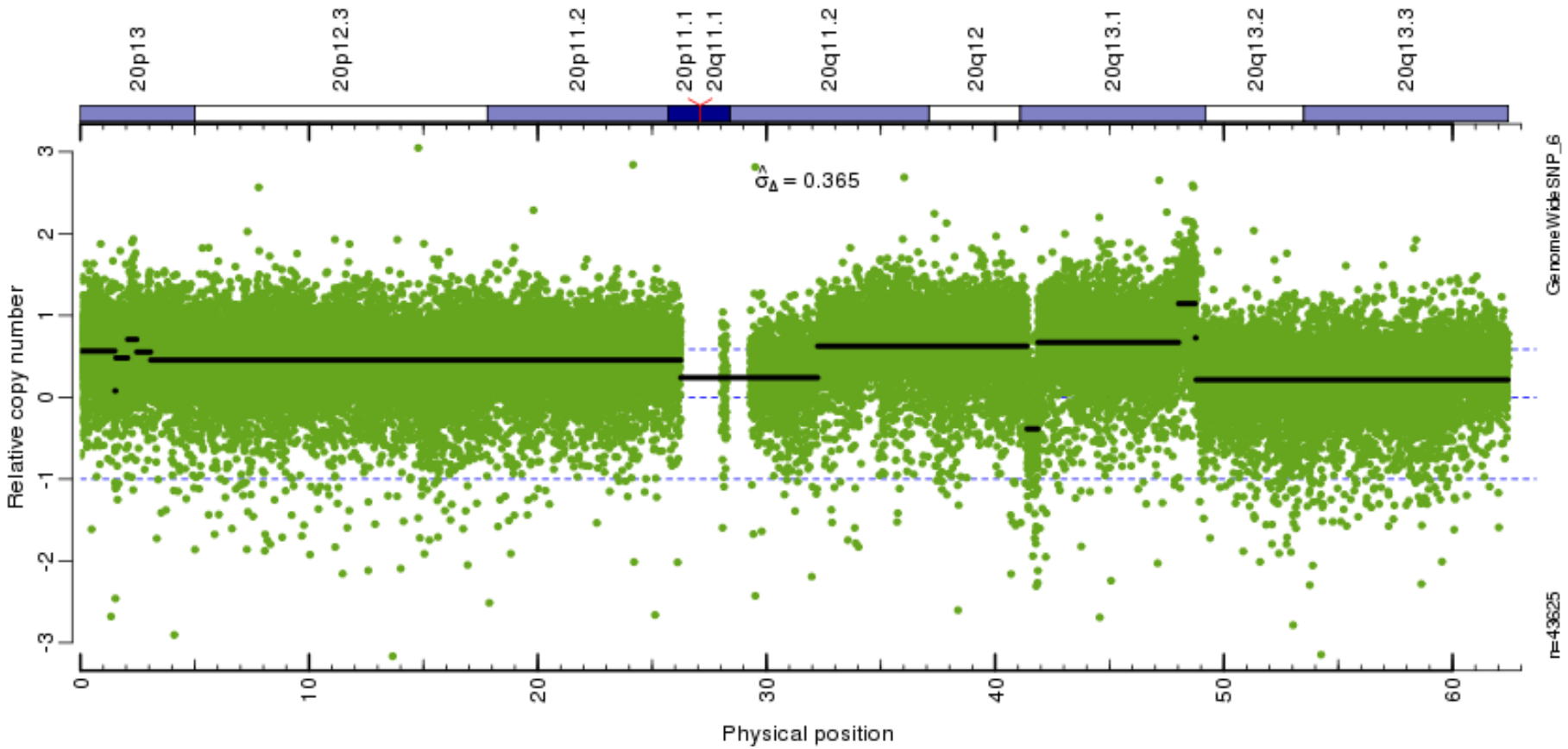
Choix d'une référence

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



Choix d'une référence

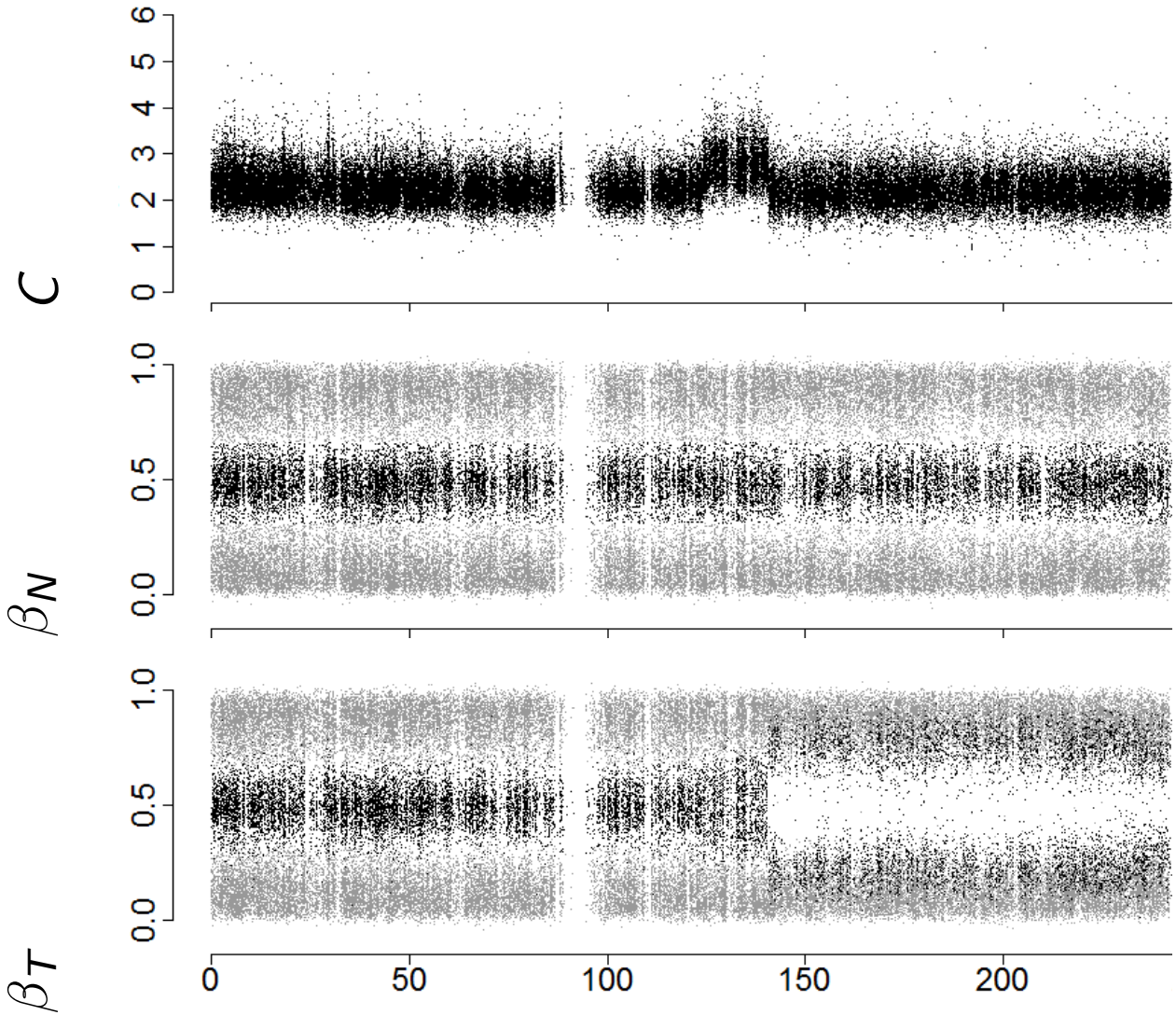
Même laboratoire, même lot d'expériences (n=22)



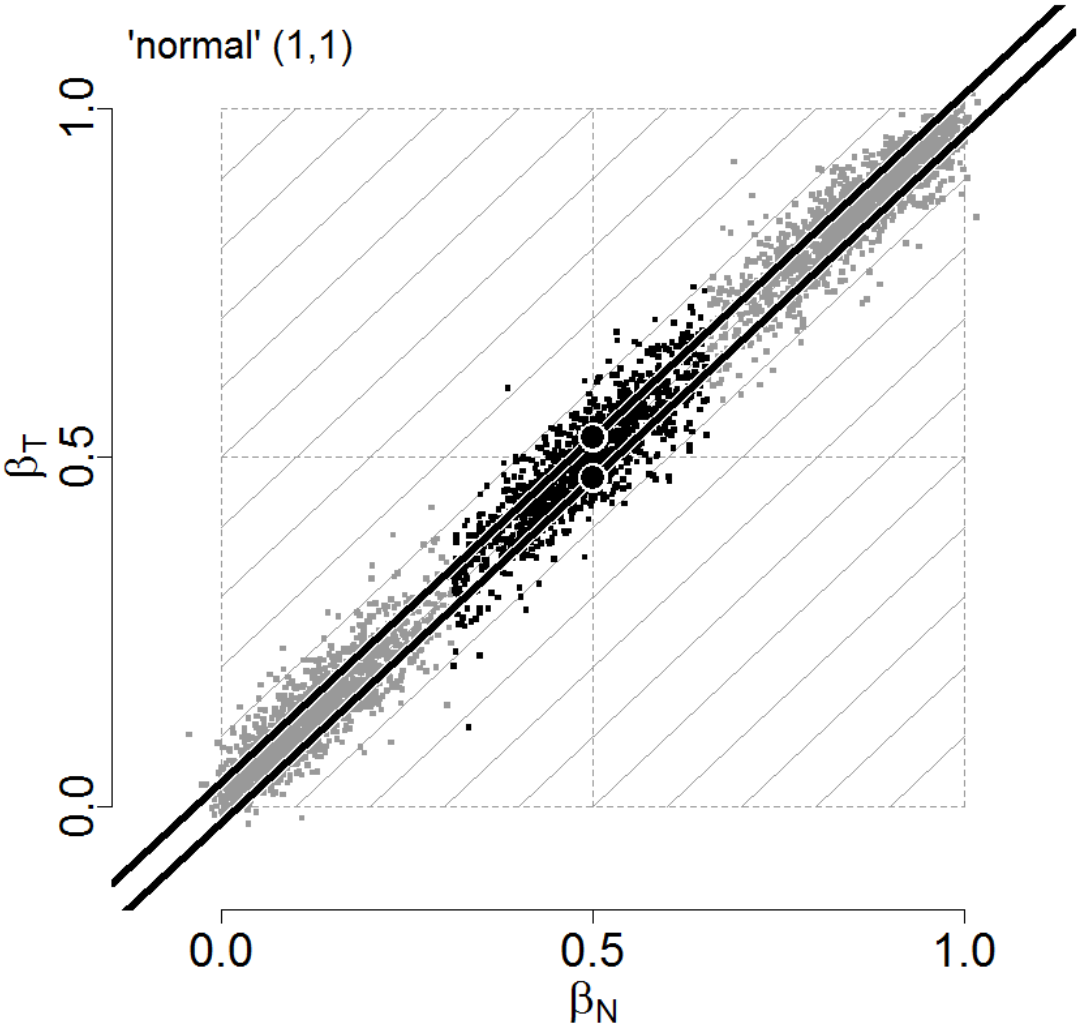
Analyse de données de nombre de copies d'ADN

Les ratios alléliques sont bruités

Exemple après pre-processing suivant la méthode CRMAv2



Effet SNP dans une région "normale" de la tumeur

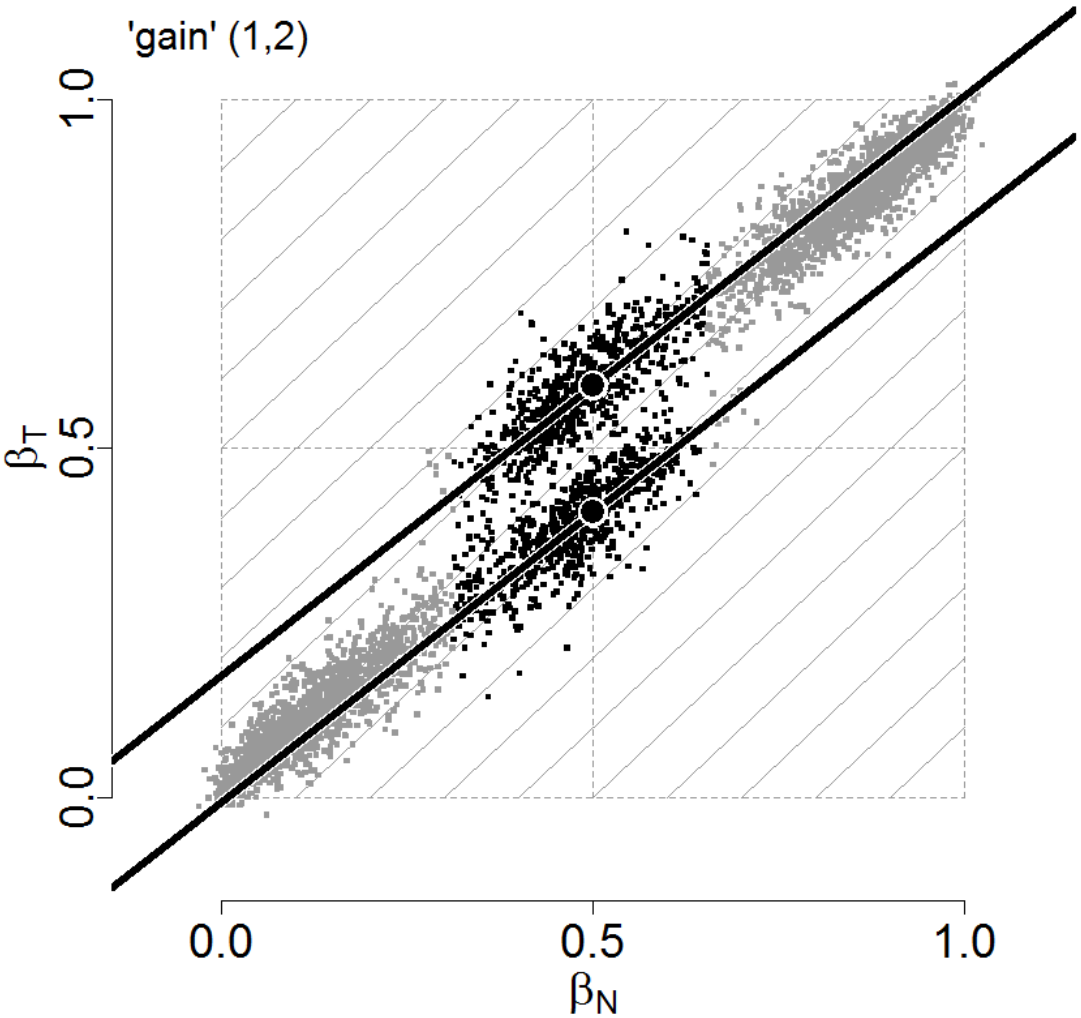


- Au lieu de trois points et (0,0), $(\frac{1}{2}, \frac{1}{2})$ et (1,1), on observe trois groupes ; la déviation est appelée **effet SNP** :

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

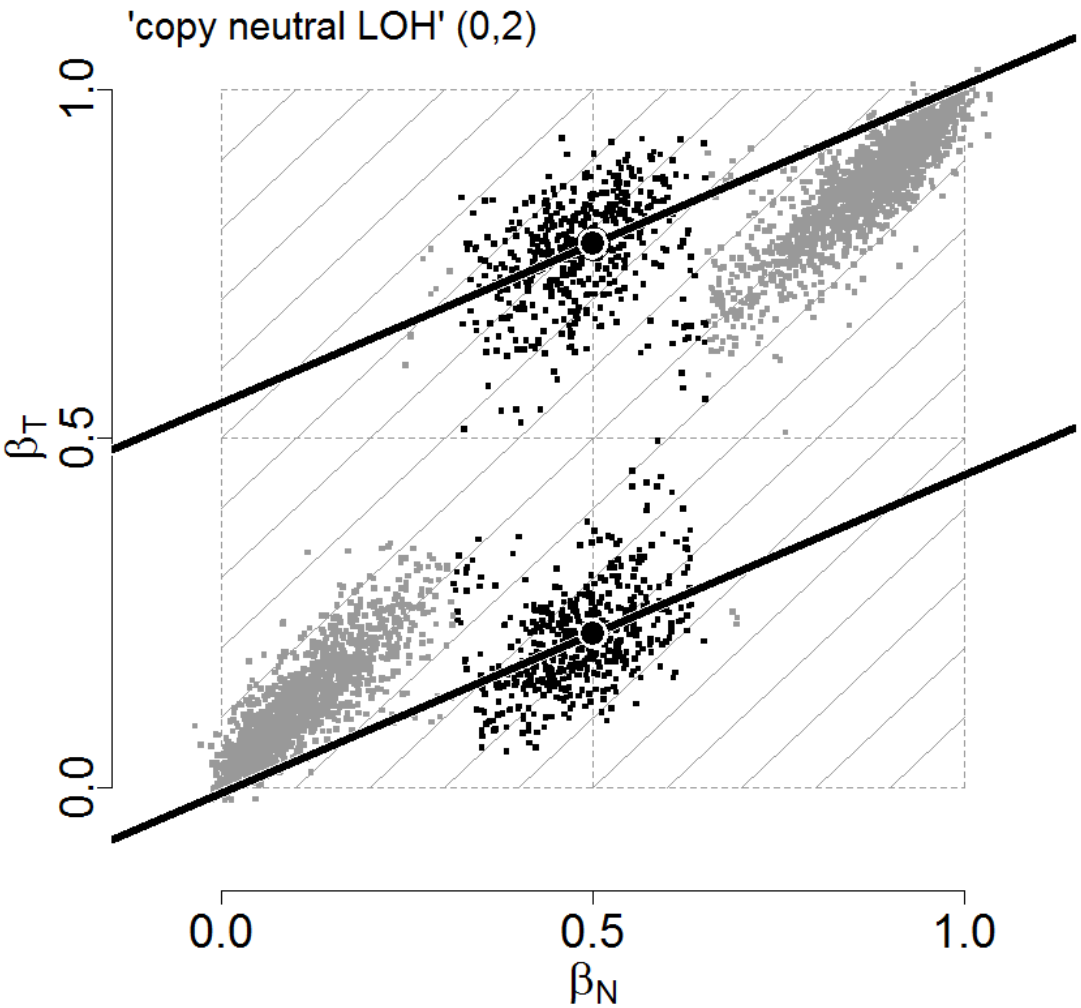
- δ est très similaire dans la tumeur et le normal

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

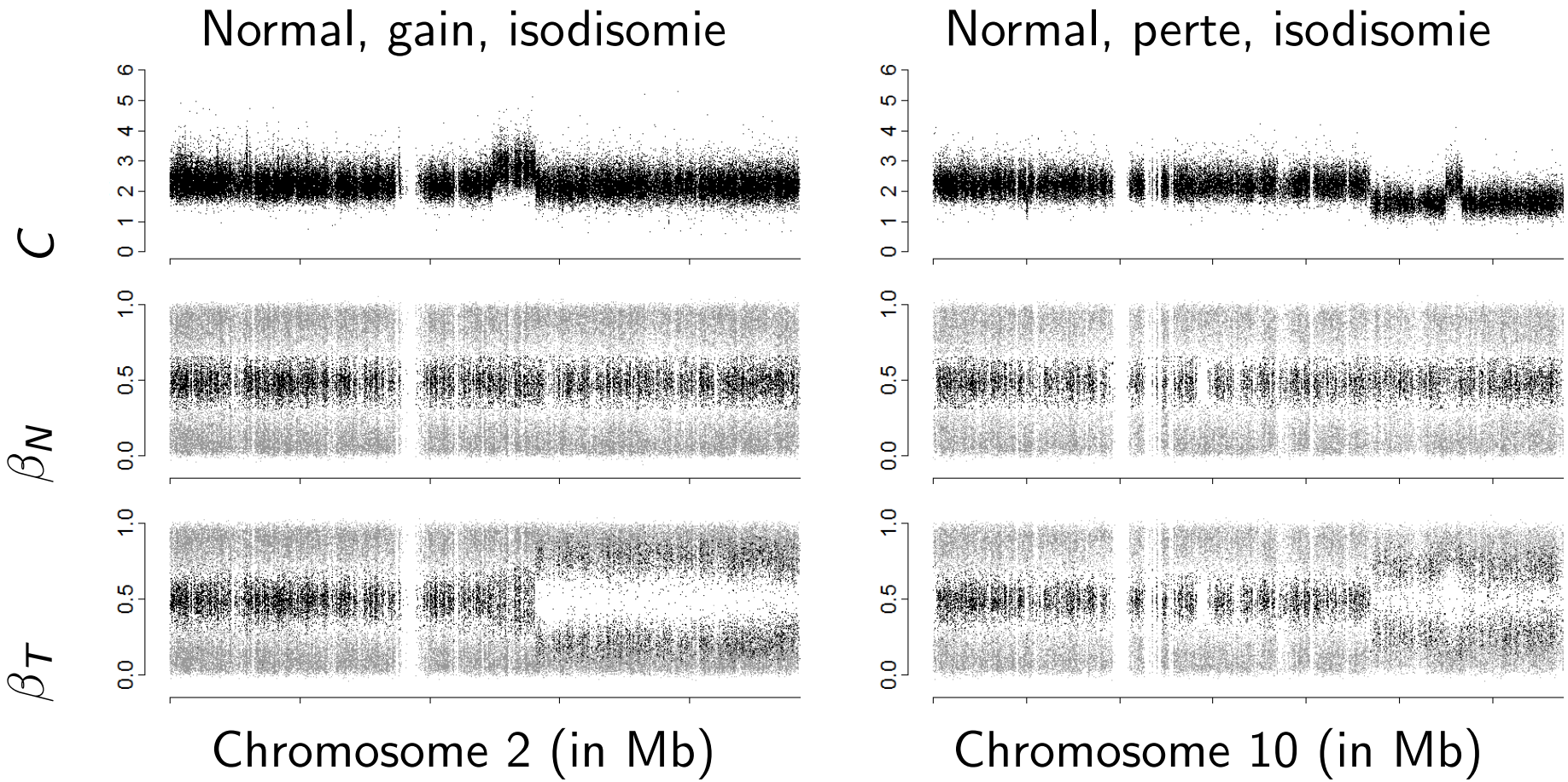
1. l'effet SNP est similaire dans le normal et dans la tumeur
2. le normal est relativement simple à analyser (seulement trois génotypes attendus)

$\Rightarrow$ Pour chaque SNP, on "soustrait" des ratios tumoraux l'effet SNP estimé grâce à l'échantillon normal

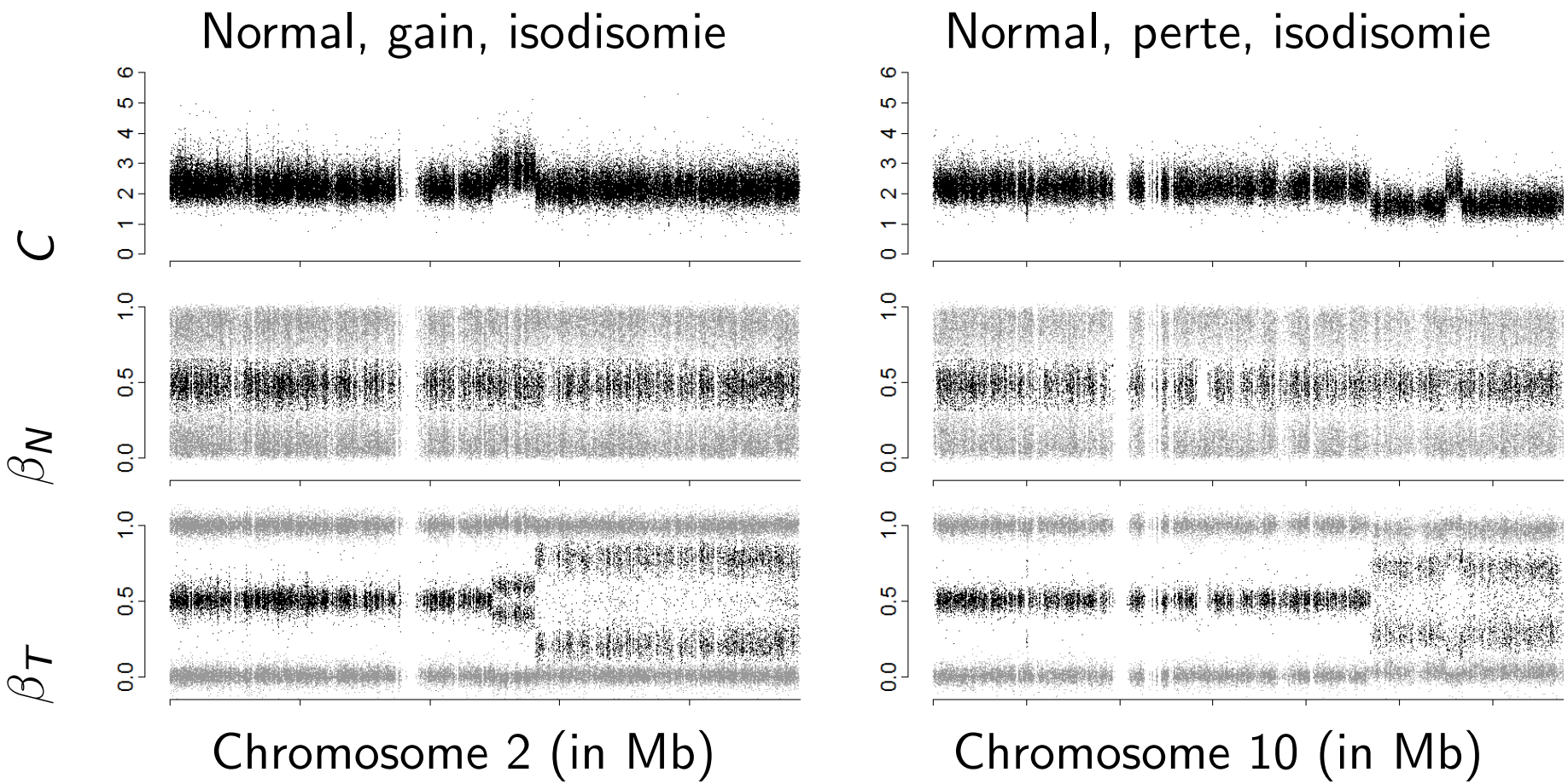
Caractéristiques de la méthode

- 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

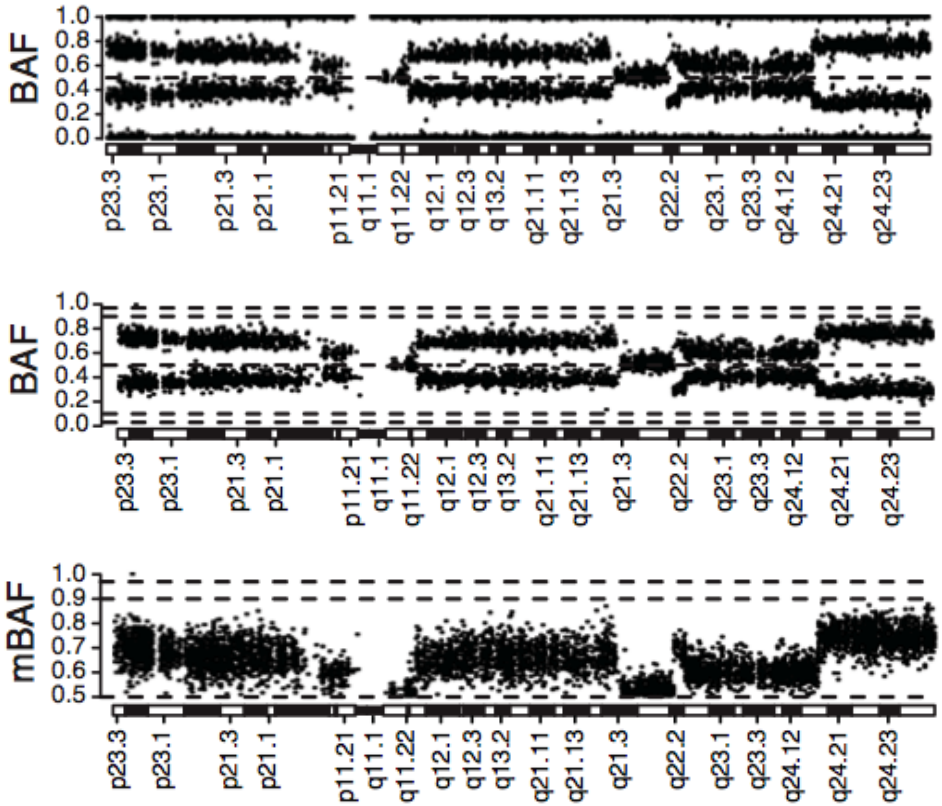


Signaux génomiques après normalisation



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

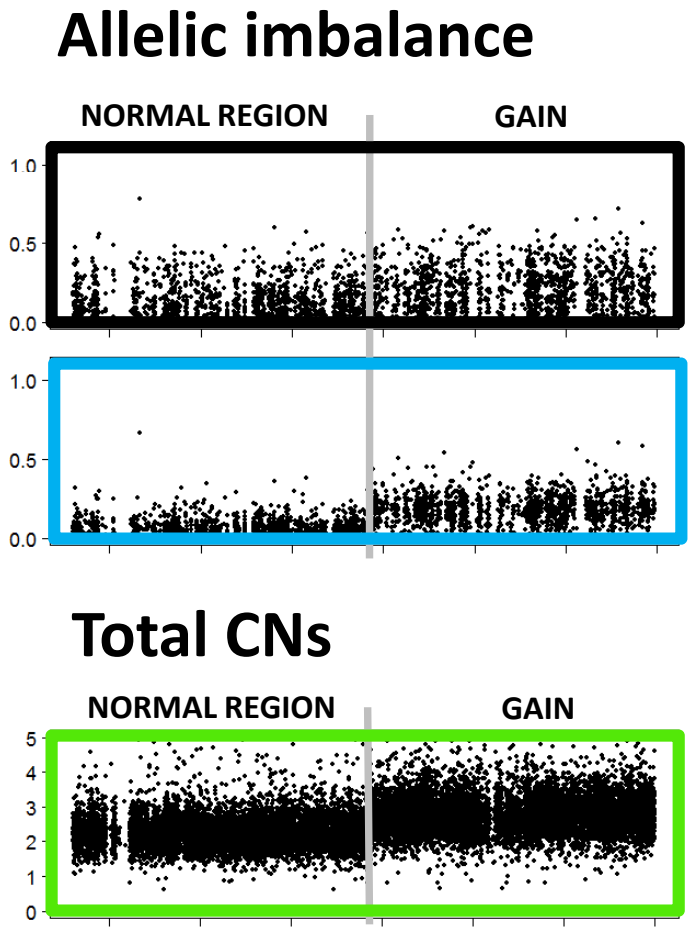
Diminution d'hétérozygotie pour les SNPs hétérozygotes :

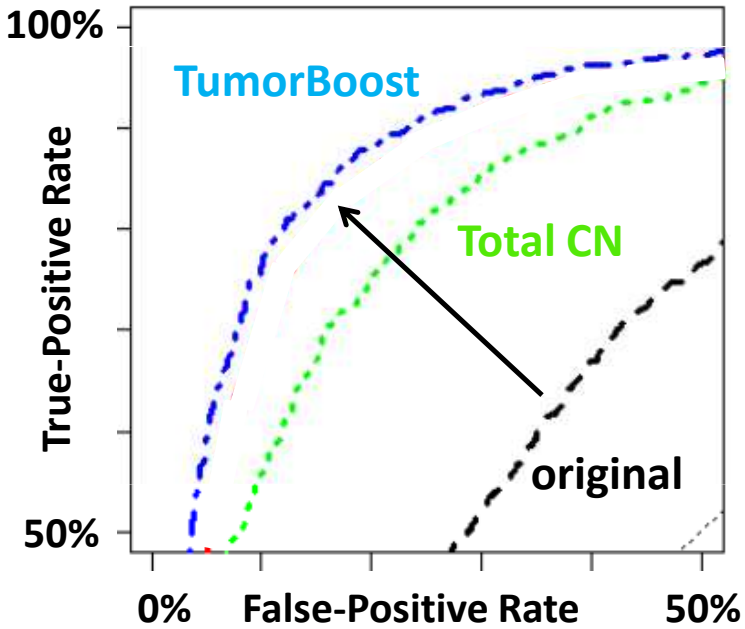
$$\rho = |\beta - 1/2| = DH/2$$

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





slide: H. Bengtsson.

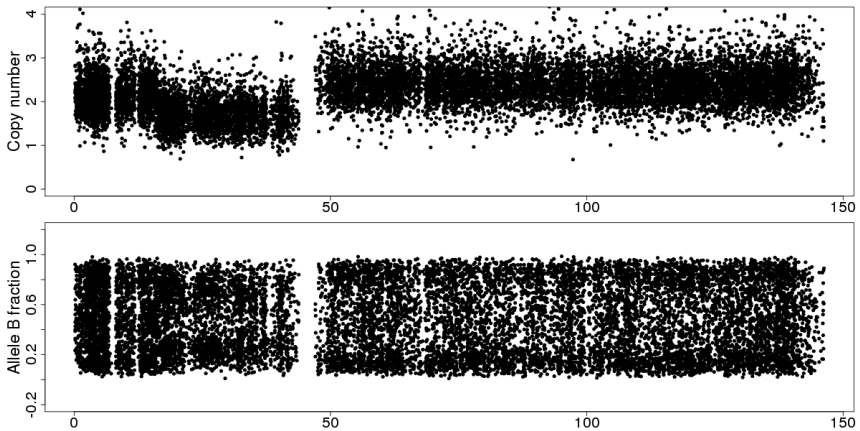
Cas où on ne dispose pas d'un échantillon normal apparié : CalMaTe

Ortiz-Estevez *et al*

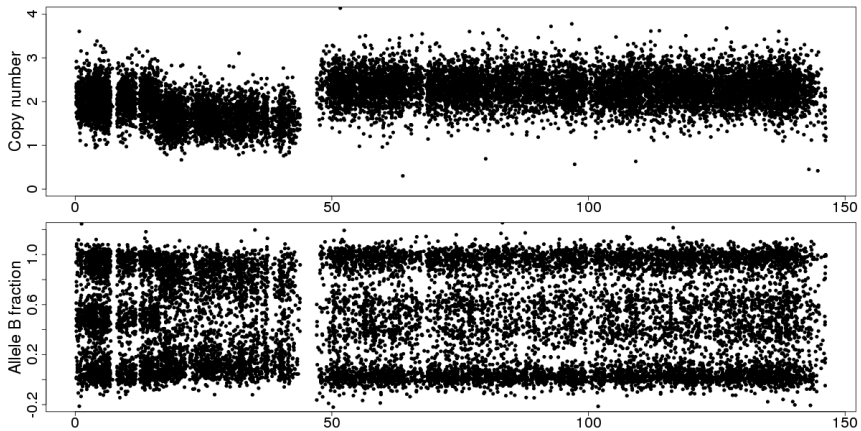
Pour chaque SNP :

- 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



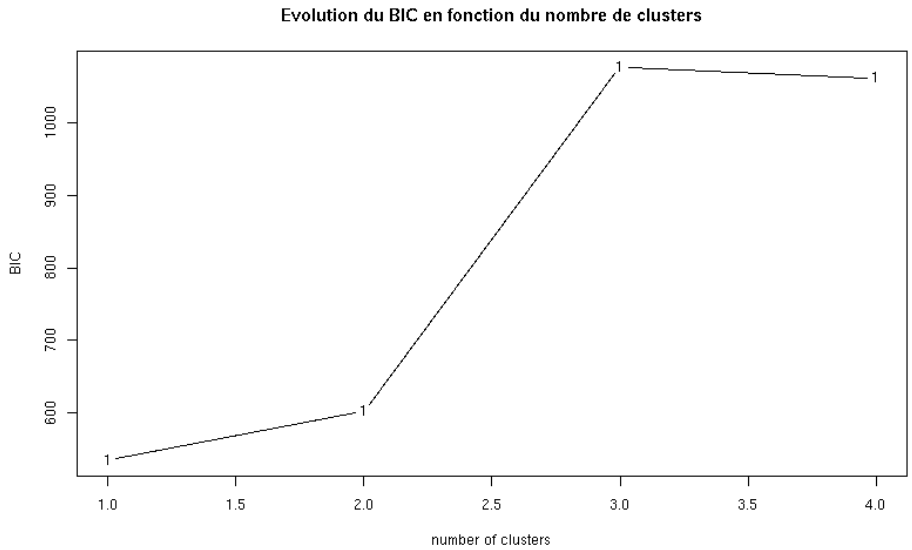
Analyse de données de nombre de copies d'ADN

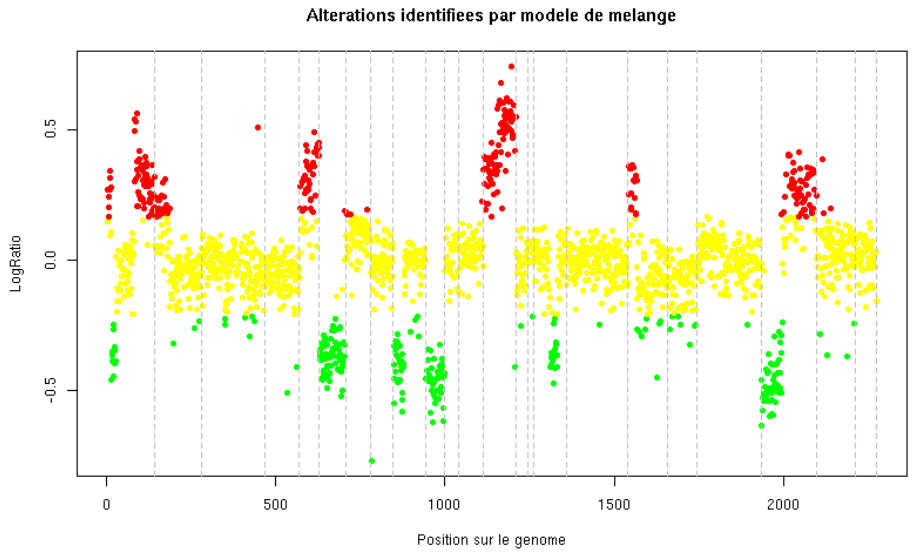
Limites des approches directes

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



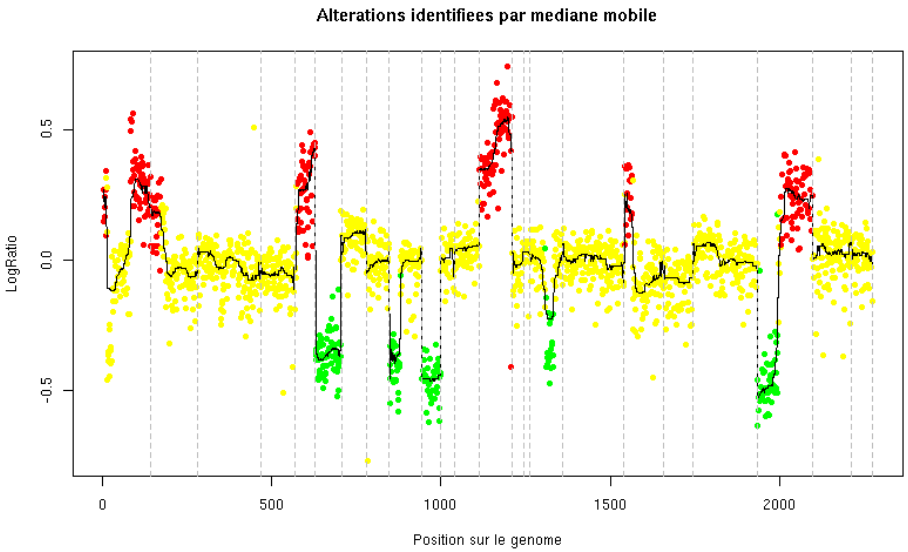


Limites des approches directes

Méthodes de lissage

Méthode

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

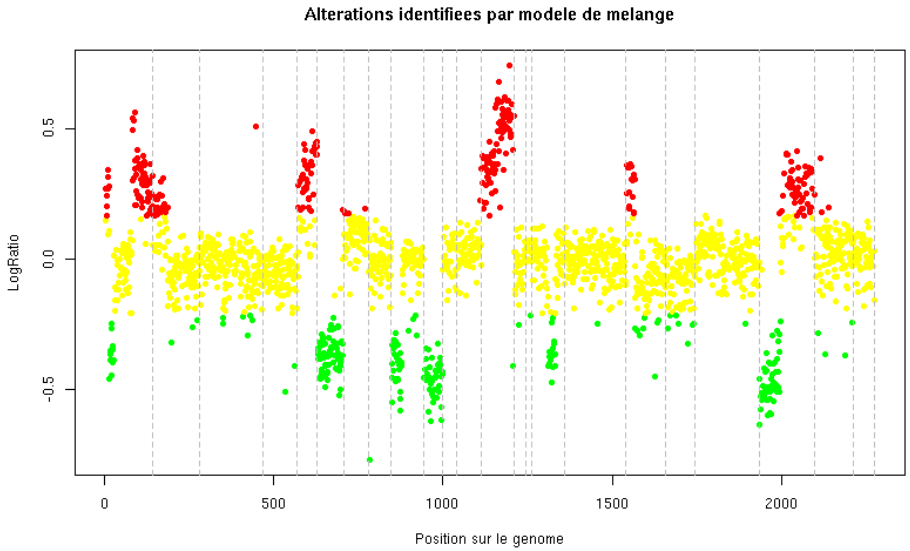


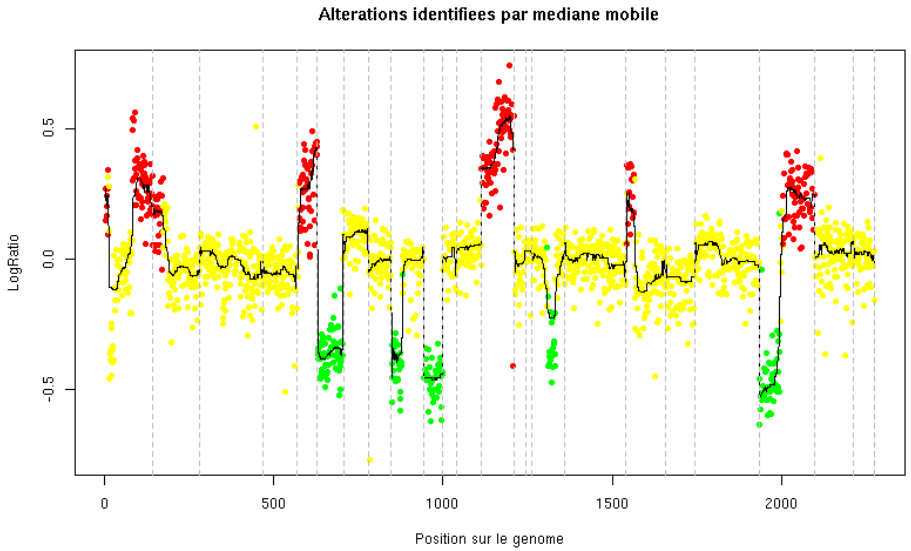
Paramètres

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

Limites des approches directes

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**

Modèles de rupture

Notations

- $\mathcal{J} = 1 \dots J$: loci génomiques
- $\boldsymbol{\gamma} = (\gamma_j)_{j=1\dots J}$: 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 $\boldsymbol{\Gamma} = (\Gamma_k)_{1 \leq k \leq K}$

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

Maximisation de la log-vraisemblance

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

$$\widehat{\Gamma_{k(j)}}^{EMV} = \frac{1}{t_k - t_{k-1}}\sum_{j=t_{k-1}}^{t_k} c_j$$

En pratique, le nombre et la position des ruptures sont inconnus

- sélection de modèle : choix de K
- combinatoire : localisation des ruptures parmi les $\binom{K-1}{J-1}$ possibles

$\binom{J-1}{K-1} = O(J^{K-1})$: une recherche exhaustive est impossible dans des situations réalistes. $\binom{10^5}{50} = 3.2 \times 10^{185}$.

Approches proposées

Choix de modèles : pénalisation de la vraisemblance $\ell(K, \cdot)$

- vraisemblance pénalisée $\bar{\ell}(K, \cdot)) = \ell(K, \cdot) - \beta pen(K)$, avec $pen(K)$ croissante en K
- choix de $pen(K)$ selon le nombre de paramètres à estimer
- choix usuels de β : $\beta = 1$ (AIC), $\beta = \frac{1}{2}\log(J)$

Exploration des partitions possibles

- Solutions exactes par programmation dynamique
- Heuristiques
  - segmentation binaire récursive
  - segmentation binaire récursive circulaire
  - relaxations convexes (fused lasso)

Pénalisation dans les modèles de rupture

Approche classique pour le choix de $pen(K)$ (AIC, BIC)

Nombre de paramètres **continus** indépendants à estimer :

- $pen(K) = 2K$ pour le cas hétéroscédastique : $(\sigma_k, \Gamma_k)_{k=1\dots K}$
- $pen(K) = K$ pour le cas homoscédastique : $\sigma, (\Gamma_k)_{k=1\dots K}$

Prise en compte de la complexité du modèle

Paramètres **discrets** supplémentaires : localisations des $K-1$ points de rupture

- Lebarbier (2005) : ajout d'un terme en $\log(\frac{J}{K})$ (complexité des états visités) : $pen(K) = 2K(c_1 + c_2 \log(\frac{J}{K}))$ pour $\mathcal{M}_1$
- Lavielle (2005) : conserve $pen(K) = 2K$ pour $J >> K$ et propose une méthode d'estimation adaptative de β

Analyse de données de nombre de copies d'ADN

Grands types d'approches

Segmentation proprement dite

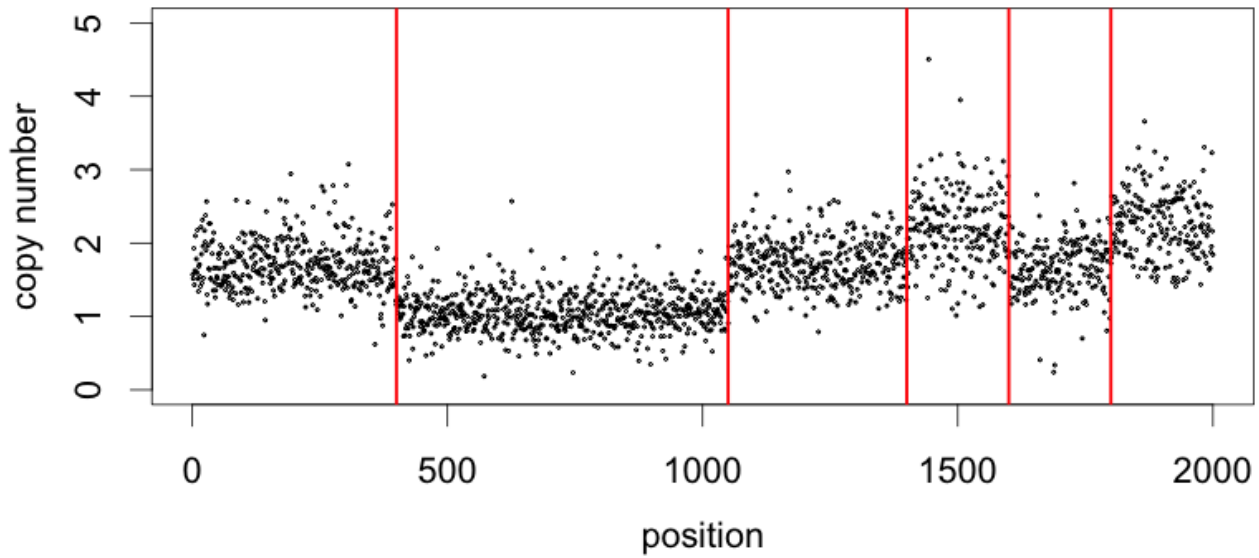
- Segmentation binaire récursive (CART)
- Segmentation binaire récursive circulaire (CBS)
- Programmation dynamique (DP)
- Fused Lasso

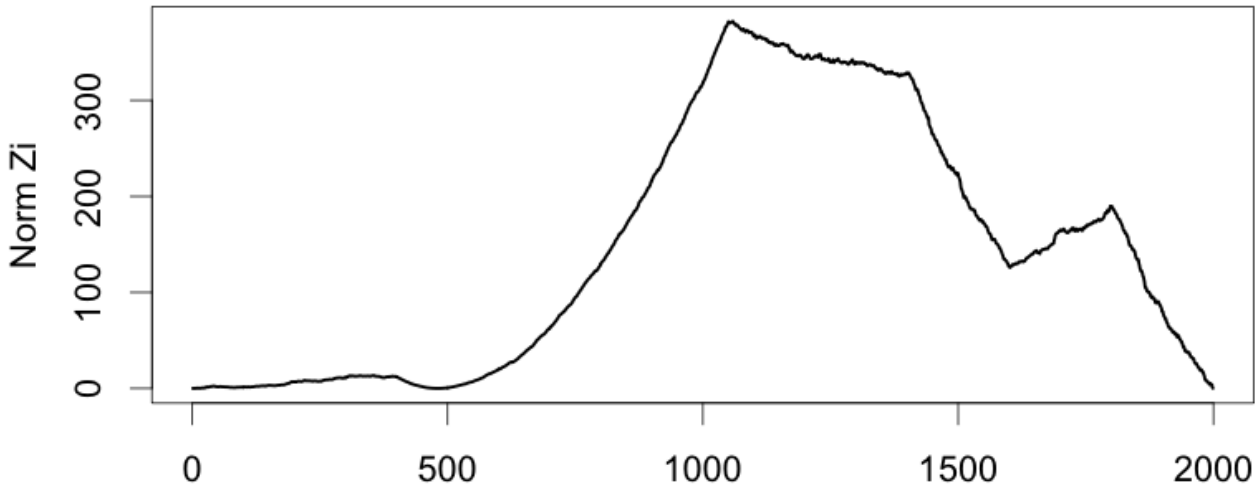
Segmentation et étiquetage conjoints

- Modèles de Markov Cachés (HMM)

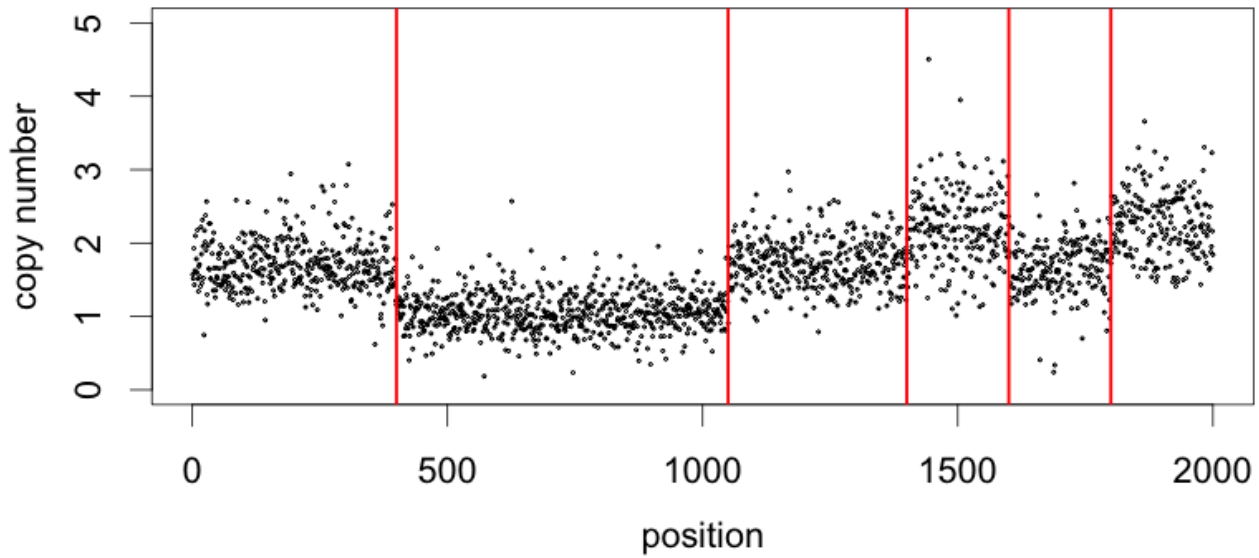
Compromis entre performances statistiques et efficacité algorithmique ?

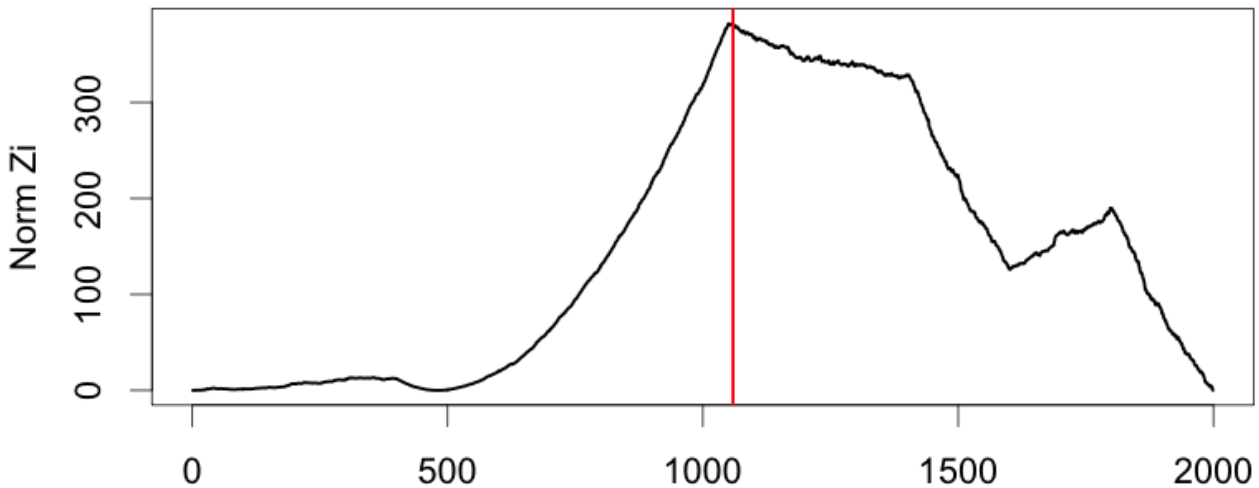
Heuristiques : segmentation binaire récursive



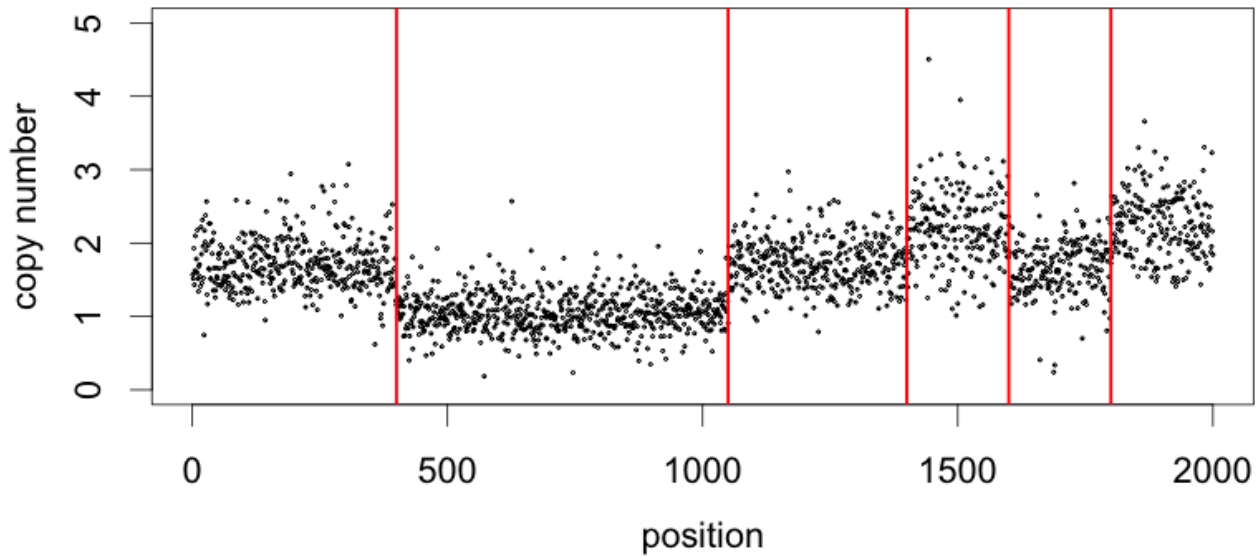


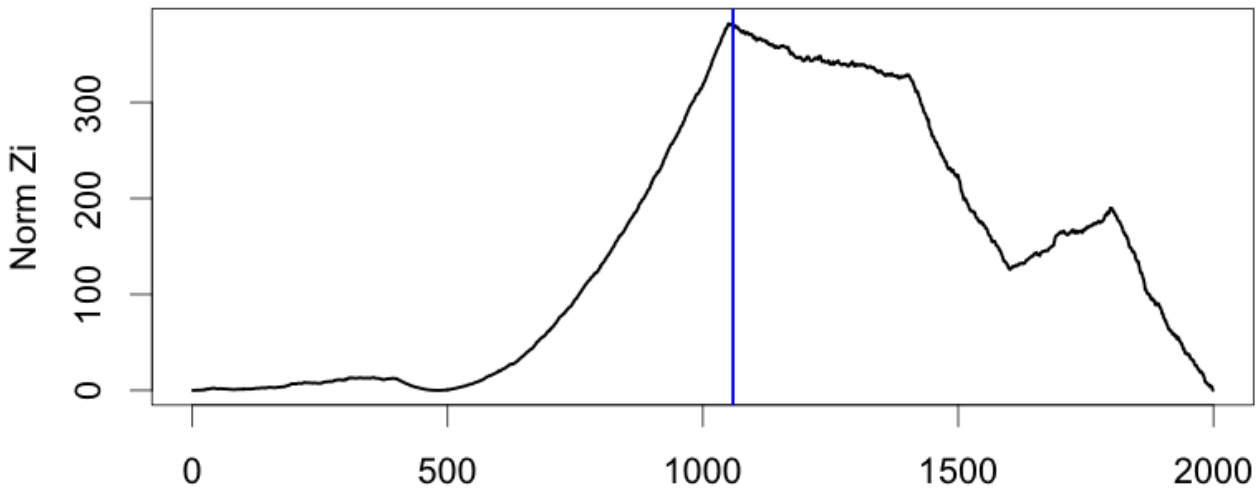
Heuristiques : segmentation binaire récursive



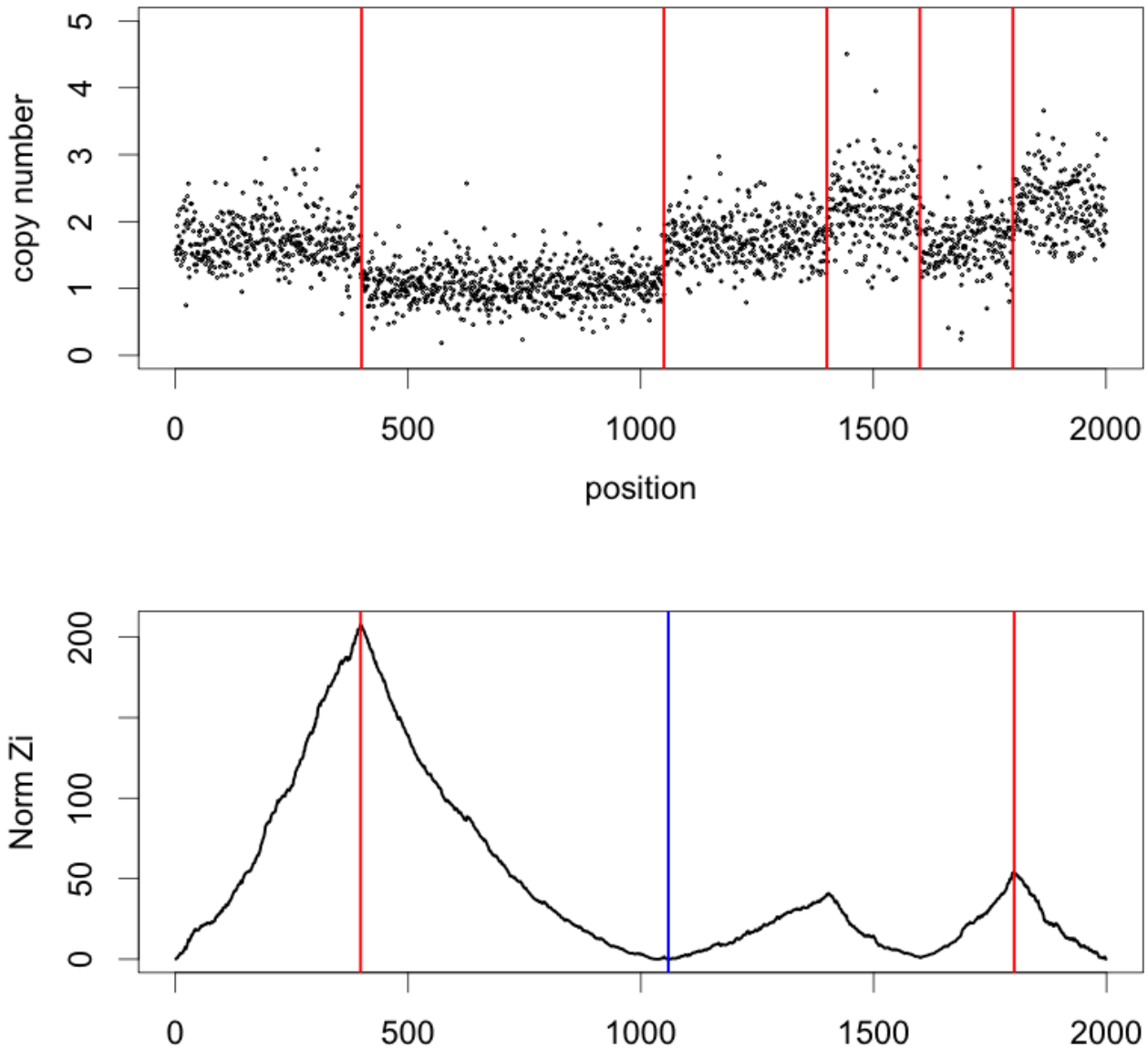


Heuristiques : segmentation binaire récursive

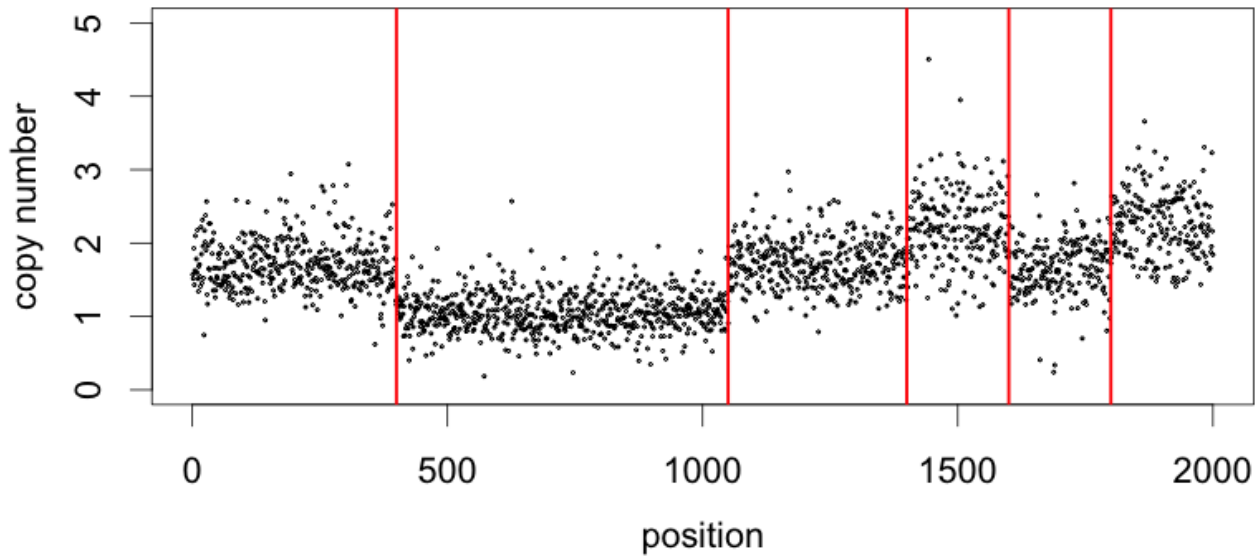


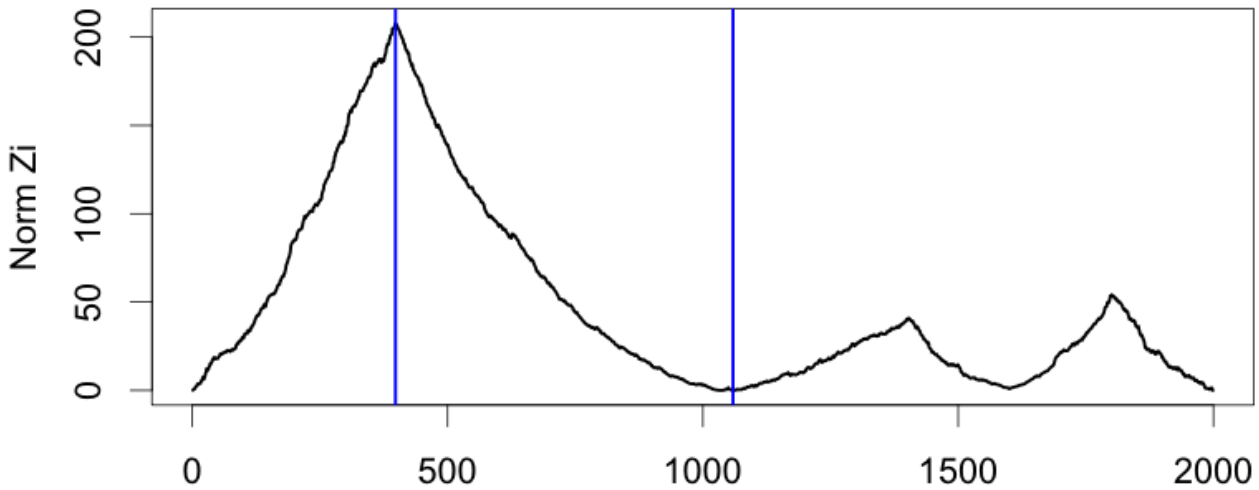


Heuristiques : segmentation binaire récursive



Heuristiques : segmentation binaire récursive





Heuristiques : segmentation binaire récursive circulaire

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

Principe de la segmentation binaire : modèle gaussien

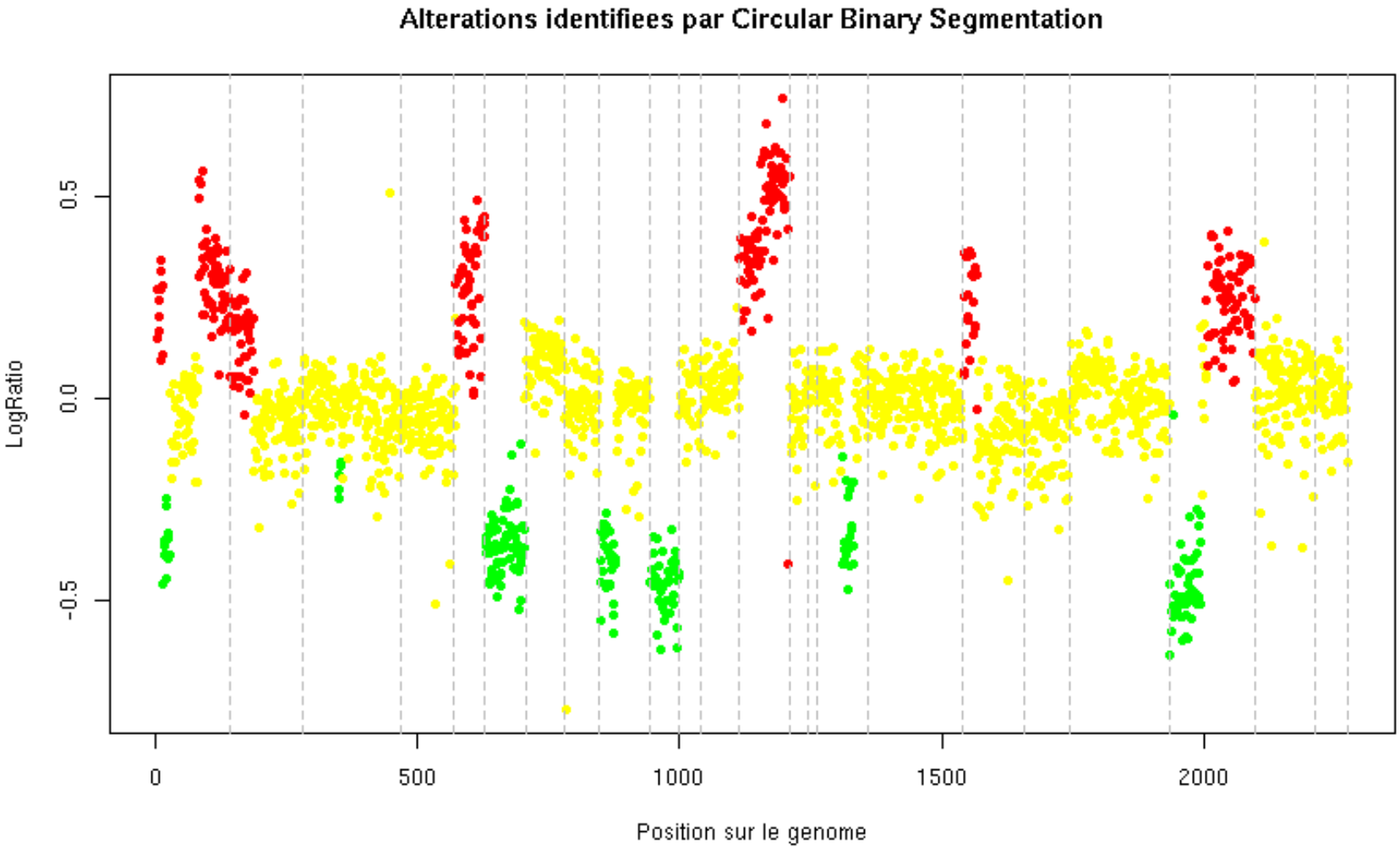
- comparaison de sommes partielles : $S_u = \sum_{j=1}^{u} y_j$
- statistique de test (LR) : $Z_u = \frac{1}{\sqrt{\frac{1}{u}+\frac{1}{J-u}}} \times \left[\frac{S_u}{u} - \frac{S_J - S_u}{J-u}\right]$
- sous l'hypothèse d'absence de rupture, $Z_u \sim \mathcal{N}[0, \sigma^2]$

Adaptations aux données de nombre de copies d'ADN

- 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**

Heuristiques : segmentation binaire récursive

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



Heuristiques : 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 ($O(Jlog(J))$)

Heuristiques : segmentation binaire récursive

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

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

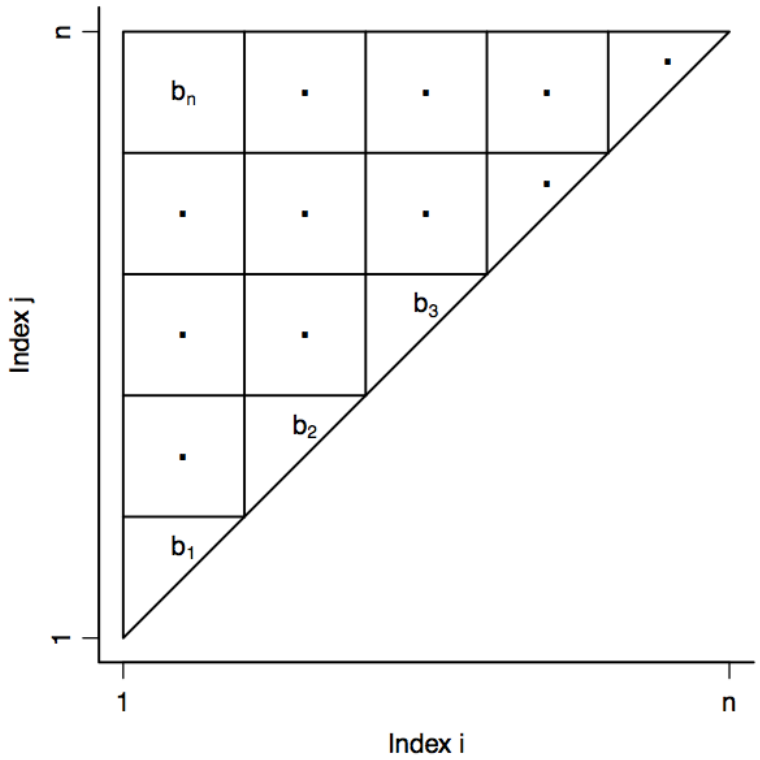


Illustration : V. E. Seshan

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

Solution exacte par programmation dynamique

Picard *et al*, 2005

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

Programmation dynamique : additivité de la log-vraisemblance

1. Calculer $V(1, j_1 : j_2)$ pour tous (j_1, j_2) tels que $1 \leq j_1 < j_2 \leq J$
2. Passer de $V(K, \cdot)$ à $V(K+1, \cdot)$ en remarquant que

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

La complexité passe de $O(J^K)$ à $O(KJ^2)$

Passage à l'échelle (Rigaill, 2010)

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

Relaxations convexes

Résolution d'un problème approché mais plus facile algorithmiquement

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

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

Complexité : $O(J^2)$

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 : $O(K^3 + JK^2)$

Conclusions sur les modèles de rupture

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

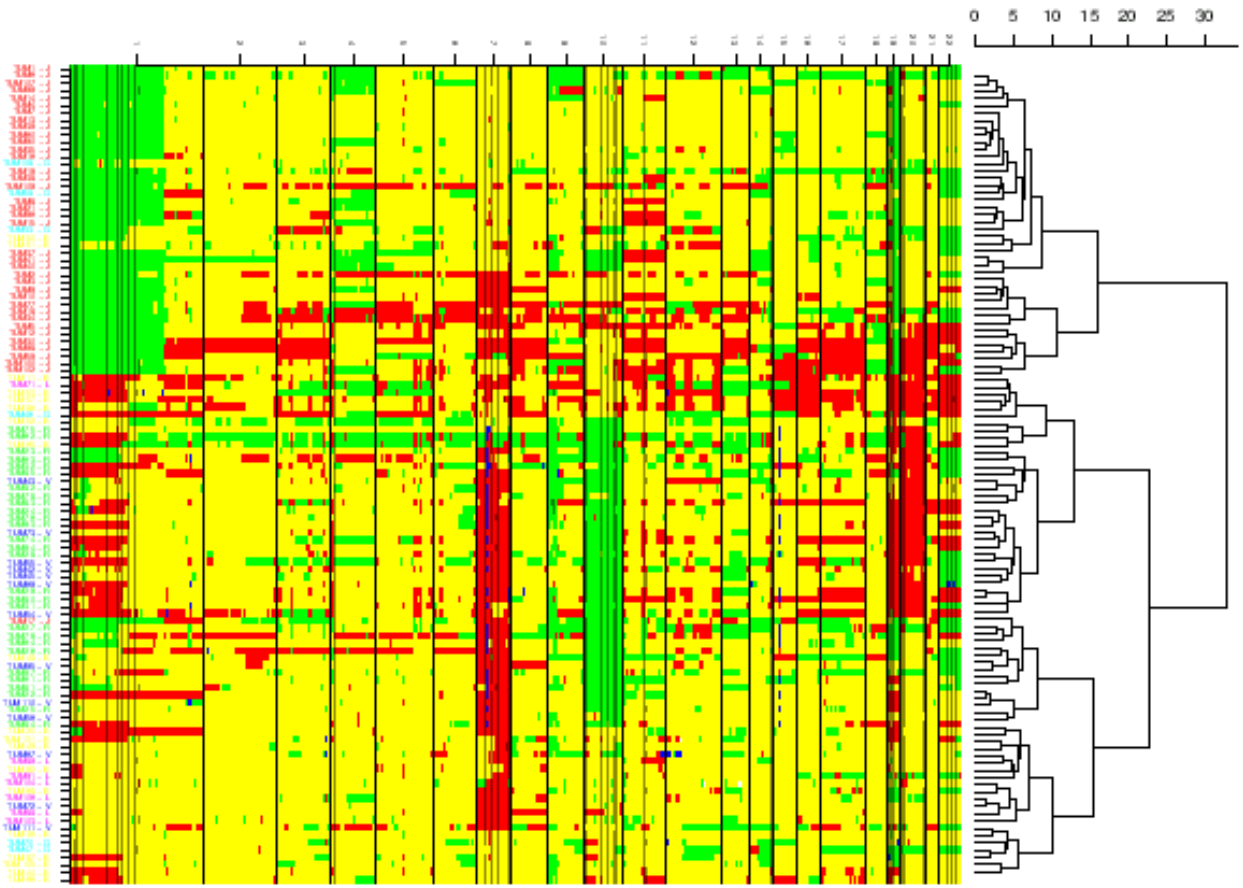
Possibilité de procéder en deux étapes :

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

Bénéfices attendus d'une segmentation multi-profils



Méthodes existantes :

- Extension de CBS (Zhang et al, Bioinformatics, 2010)
- Extension des méthodes de type Lasso (Bleakley and Vert, NIPS 2010)

Segmentation binaire récursive multi-profils

Zhang et al, Biometrika, 2010

Statistique de test pour le profil n (variance connue) :

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

Statistique de test pour N profils (variance connue)

$$Z_{uv}^{[N]} = \sum_{n=1}^{N} Z_{uv}^n$$

En l'absence de cassure, $Z_{uv}^{[N]} \sim \chi^2(N)$ (asymptotiquement)

Contrôle du taux d'erreur

Approximation de $\mathrm{P}\left[\max_{1 \leq u < v \leq J, c_1 J < v-u < c_2 J} Z_{uv}^{[N]} > b^2\right]$

Segmentation multi-plate-forme : motivations

Approches existantes :

- Intégration avant segmentation (Bengtsson et al, Bioinform., 2009)
- Extension de CBS (Zhang et al, Bioinform., 2010)

Intégration avant segmentation : méthode

Bengtsson et al, Bioinformatics, 2009

Multiplatform data in R^4 (4 platforms):

True CN:

x (an unknown scalar)

Smoothed CNs:

$\underline{y} = (y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)})^T$

Unknown transformation:

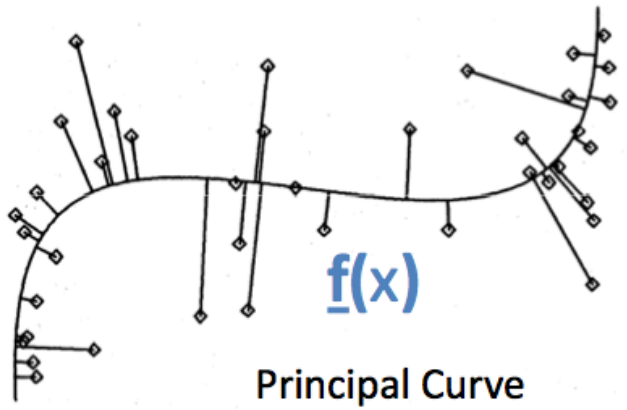
$\underline{f}(x) = (f^{(1)}(x), f^{(2)}(x), f^{(3)}(x), f^{(4)}(x))^T$

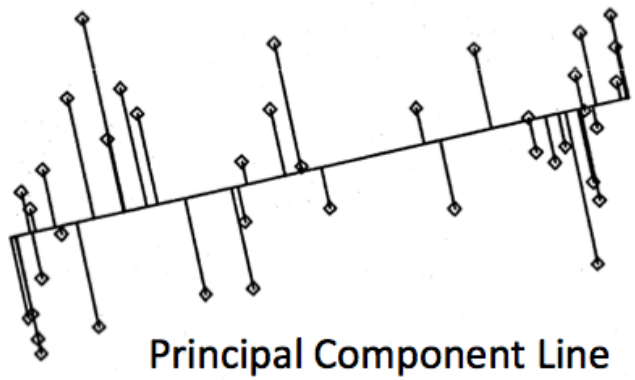
Noise:

$\underline{\varepsilon} = (\varepsilon^{(1)}, \varepsilon^{(2)}, \varepsilon^{(3)}, \varepsilon^{(4)})^T$

Vector model:

$\underline{y} = \underline{f}(x) + \underline{\varepsilon}$





Hastie et al., Principal Curves, JASA, 1989

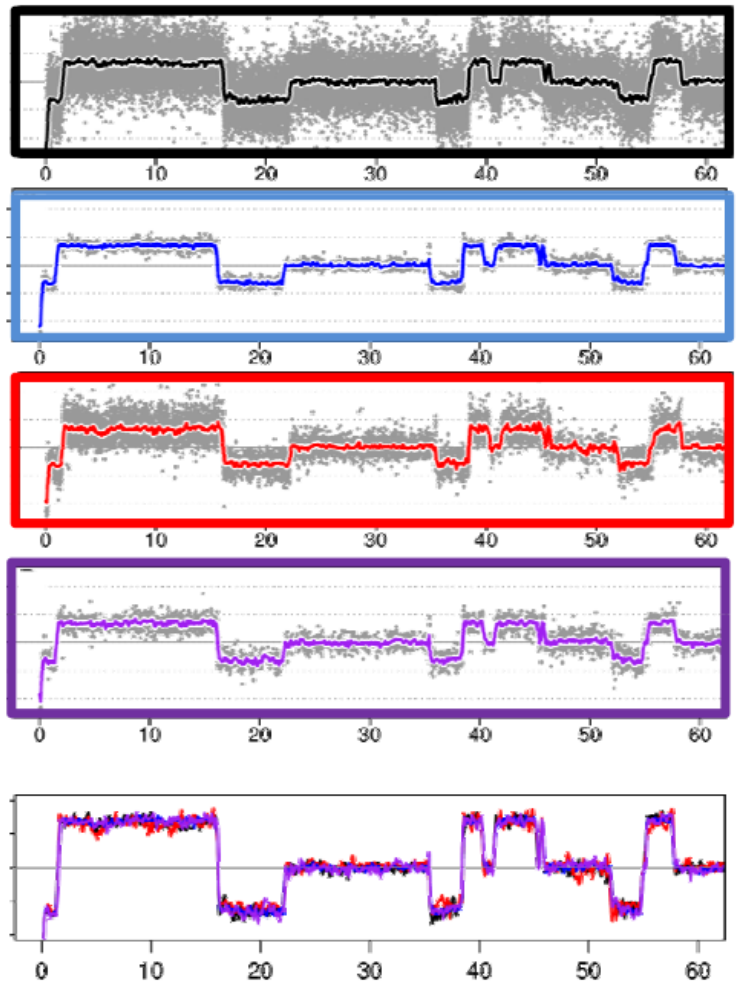
Intégration avant segmentation : résultats

A. Broad
Affymetrix
GenomeWideSNP_6
($n=1.8 \cdot 10^6$)

B. MSKCC
Agilent
HG-CGH-244A
($n=0.25 \cdot 10^6$)

C. Stanford
Illumina
HumanHap550
($n=0.55 \cdot 10^6$)

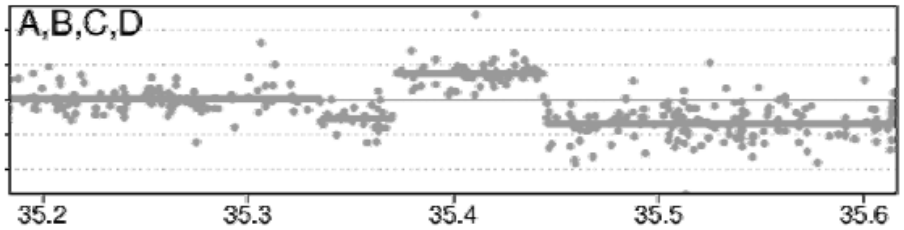
D. Harvard
Agilent
HG-CGH-244A
($n=0.25 \cdot 10^6$)



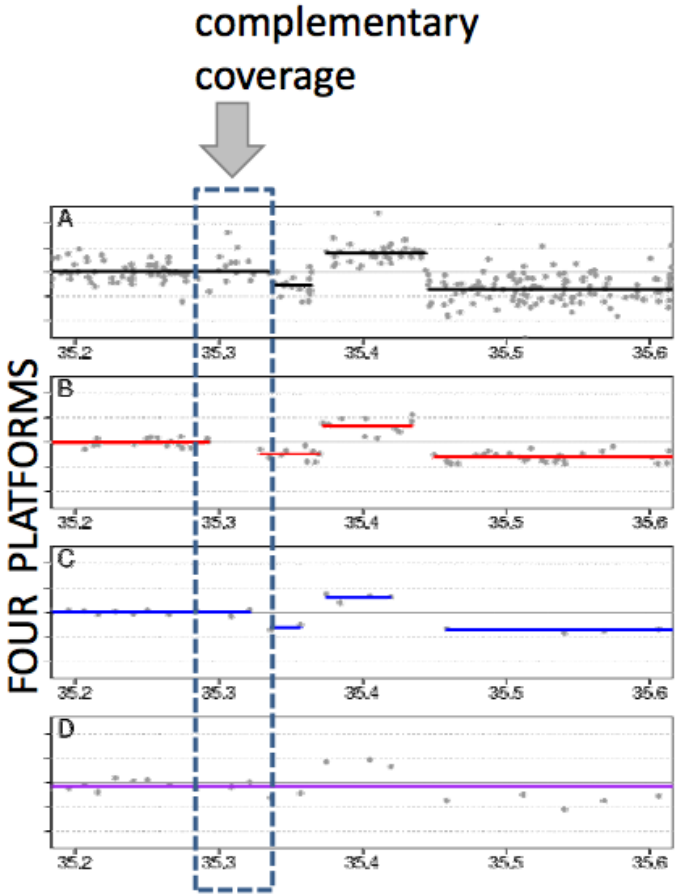
slide: H. Bengtsson.

Intégration avant segmentation : résolution

Combining normalized data:



1. Greater power to detect CN changes
2. More precise locations.
3. Greater resolution.
4. Greater and complementary coverage.

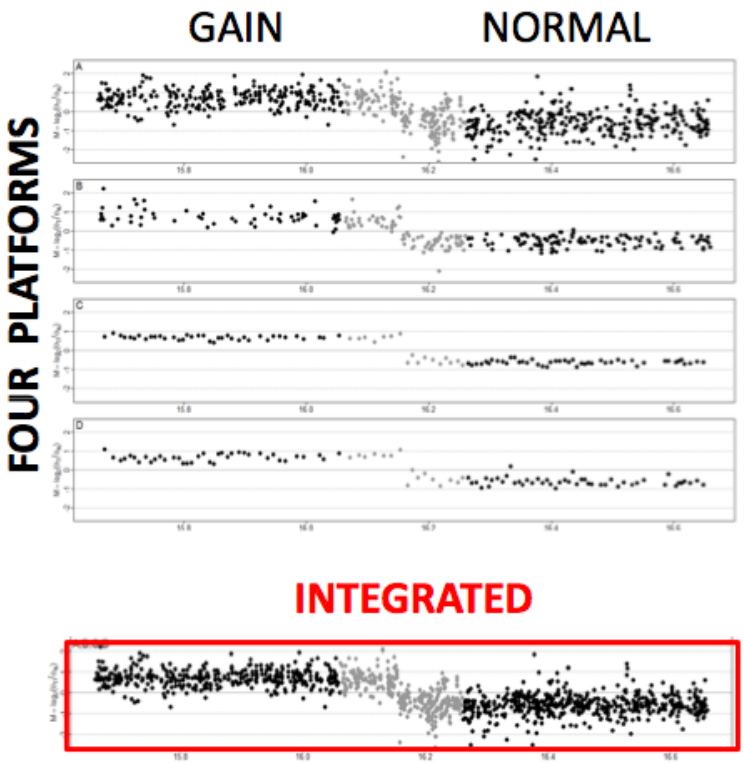


slide: H. Bengtsson.

Intégration avant segmentation : performances

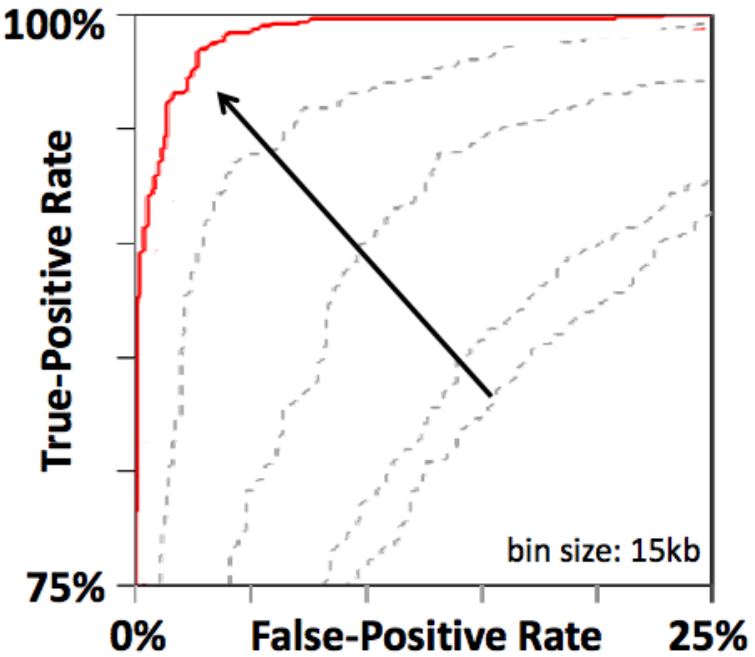
Data:

(1) Pick a random sample.
(2) Find a clear CN change point.
(3) Two CN states: GAIN and NORMAL.



Assessment via ROC:

(1) Quantify how well we can call GAIN:s from NORMAL:s.



Repeat:

Repeat the above for several change points.

slide: H. Bengtsson.

Segmentation binaire récursive multi-plate-forme

Zhang et al, Bioinformatics, 2010

Statistique de test pour la plate-forme n :

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

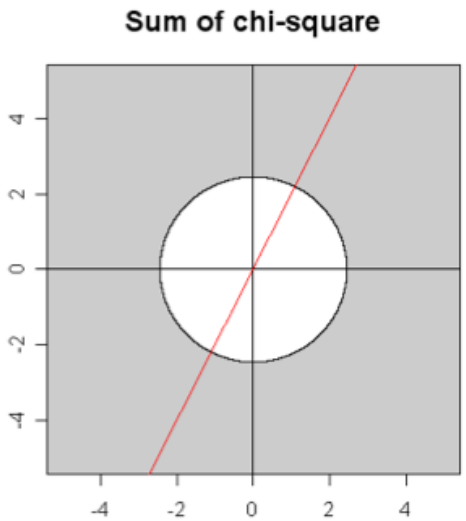
Statistique de test pour N plate-formes

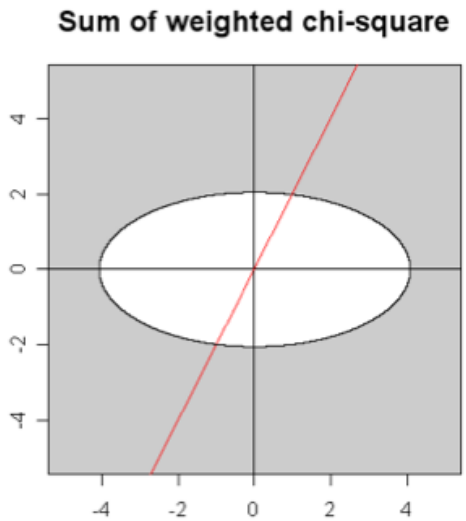
$$Z_{uv}^{[N]} = \frac{\left[\sum_{n=1}^N \delta_{uv}^k Z_{uv}^n\right]^2}{\sum_{n=1}^N (\delta_{uv}^k)^2}$$

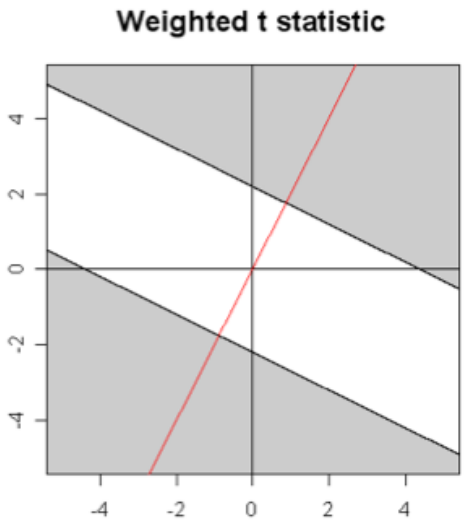
Choix d'un nombre de ruptures à l'aide d'un critère BIC modifié.

Segmentation binaire récursive multi-plate-forme

Intuition pour la statistique de test





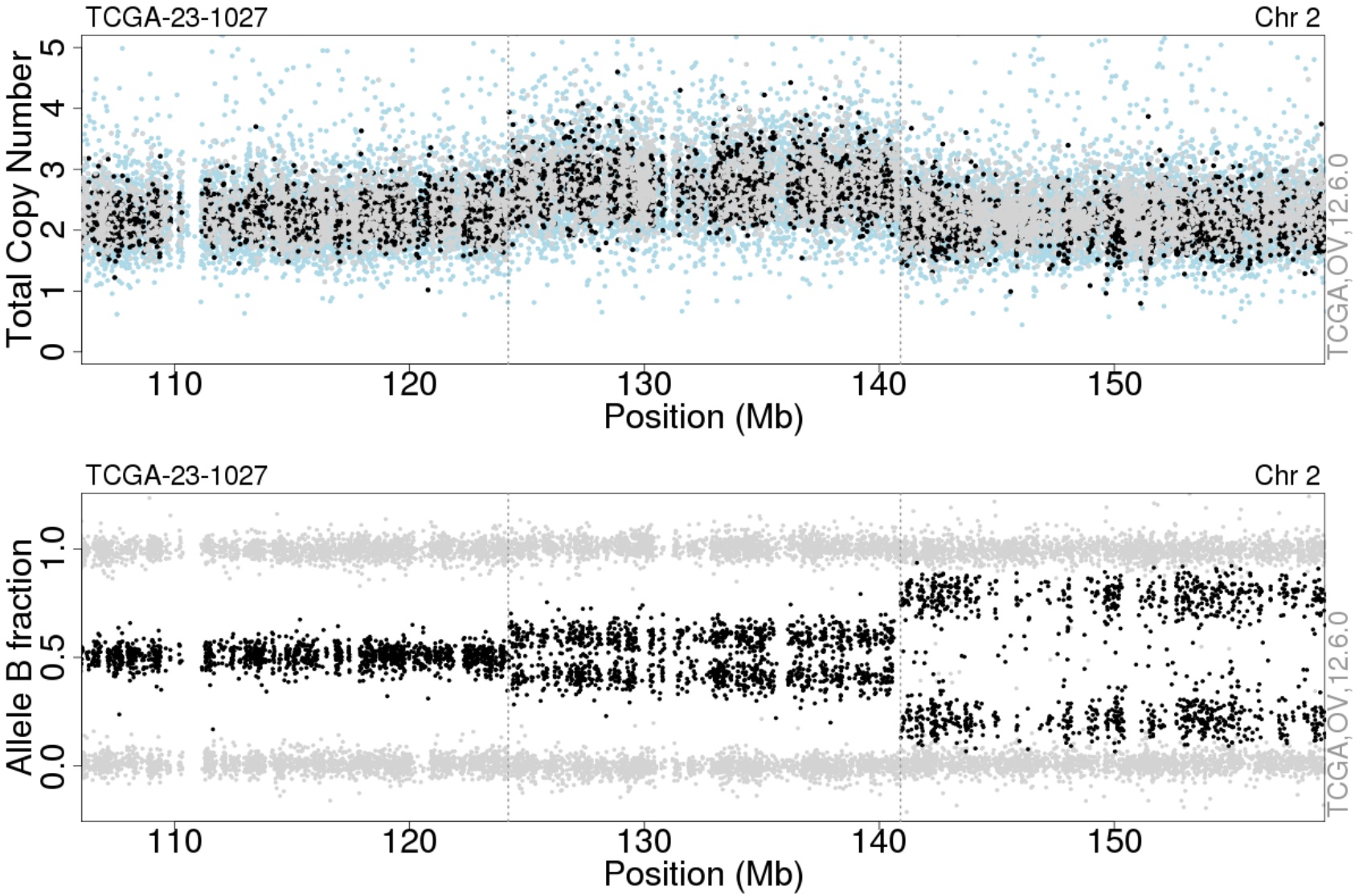


Zo

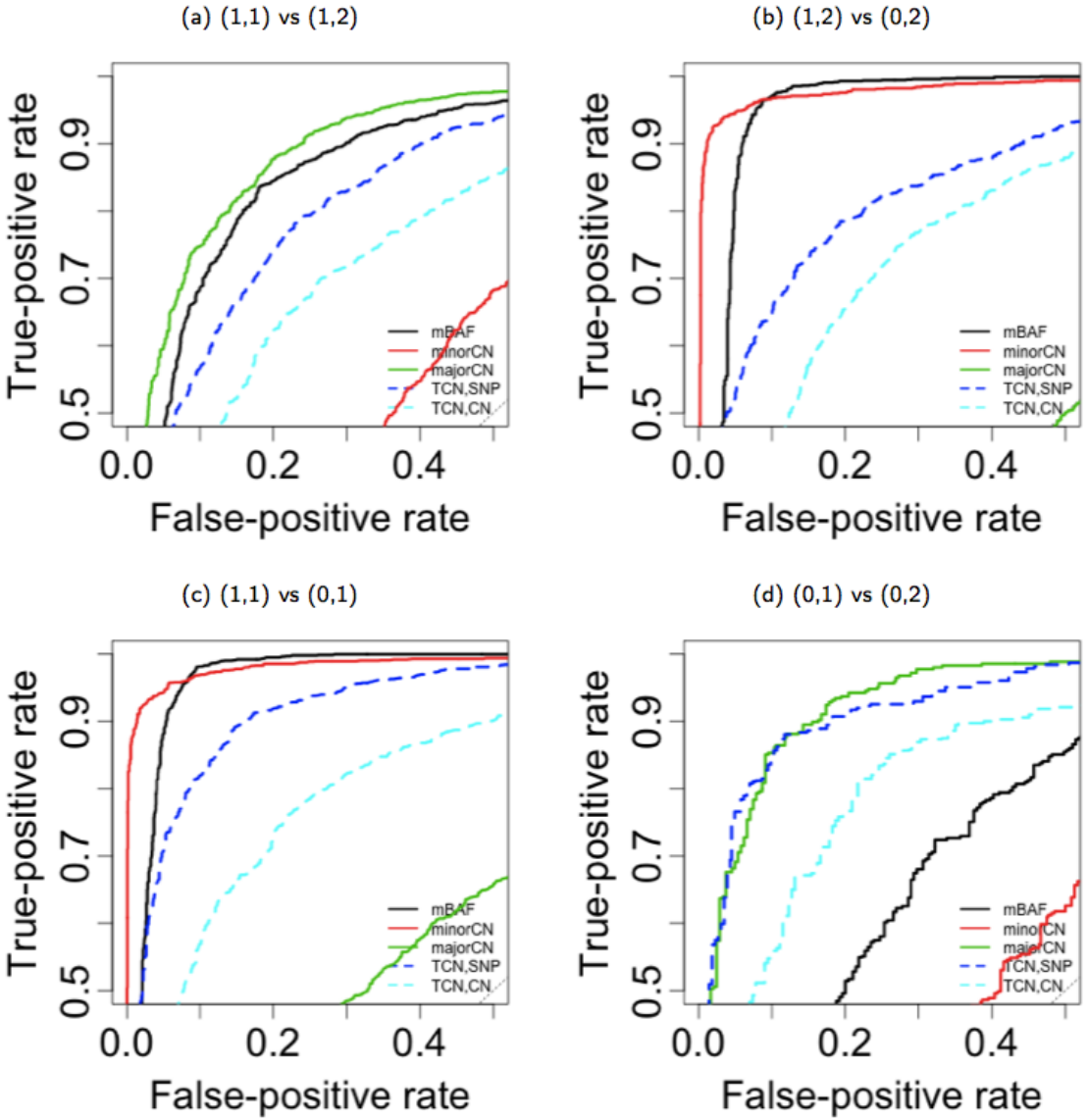
d'acceptation (blanc) et de rejet (gris) pour une statistique bi-dimensionnelle construite à partir de deux statistiques uni-dimensionnelles.

Analyse de données de nombre de copies d'ADN

Les ruptures sont présentes dans les deux dimensions



DH est plus puissant que *C* locus par locus



Plus de sondes informatives pour C que pour DH

Affymetrix GenomeWideSNP_6

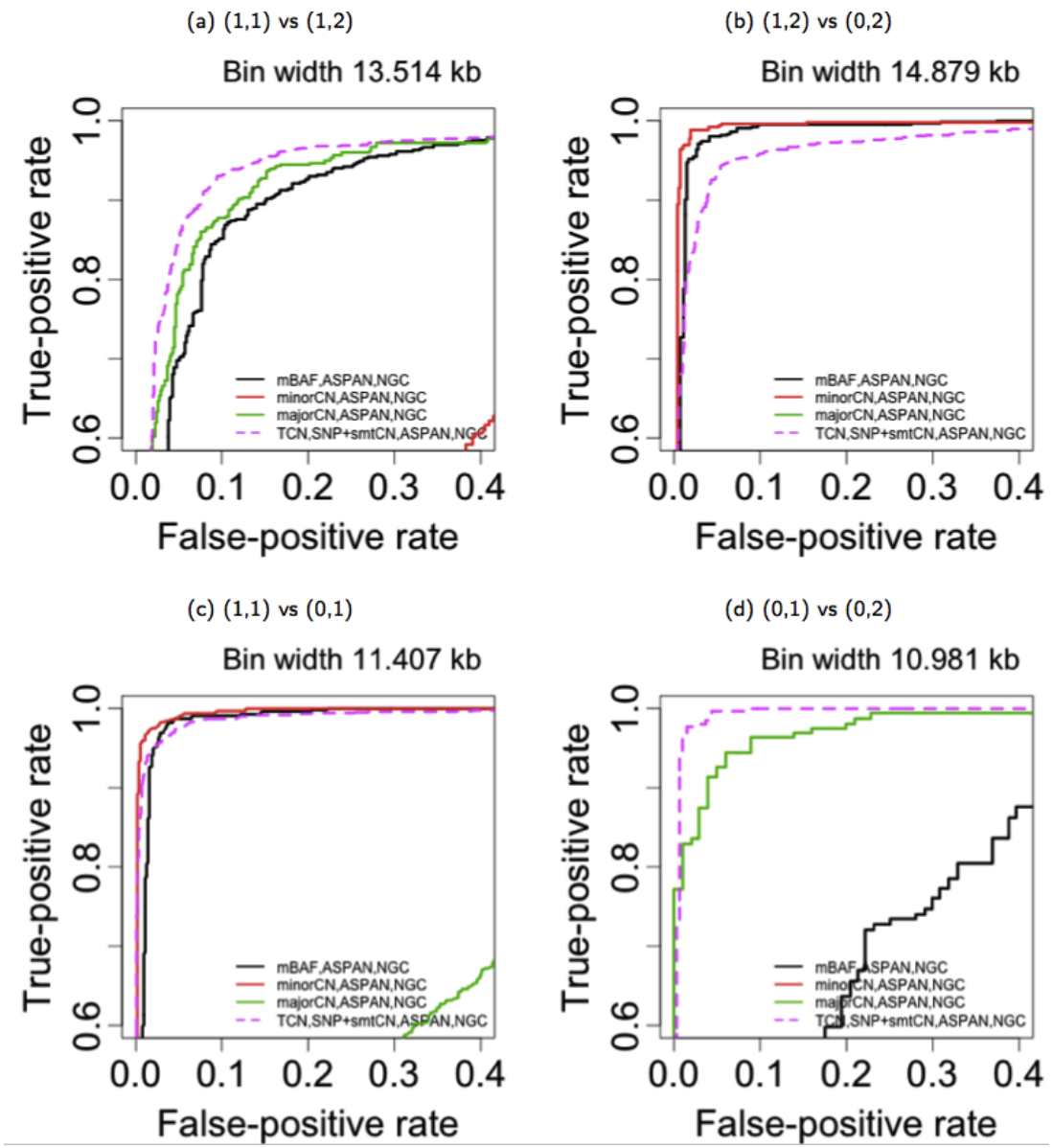
	Toutes sondes	sondes CN	sondes SNP
Effectif	1,856,069	946,705	909,364
Proportion	100%	51%	49%

Répartition par type de sonde

	Tous les SNPs	AA	AB	BB
Effectif	1,856,069	326,500	251,446	331,418
Proportion	100%	18%	**14%**	18%

Nombre de SNP par génotype pour l'échantillon TCGA-23-1027

Puissance de détection comparable à résolution fixée



Etat de l'art pour la segmentation conjointe

Méthode	Temps	# dims
Programmation dynamique (DP)		
[Rigaill et al.(2010)]	$n\log(n)$	1
[Picard et al. (2005)]	$d \cdot K \cdot n^2$	≥ 1
Fused Lasso / LARS		
[Harchaoui and Lévy-Leduc(2008)]	$K \cdot n$	1
[Bleakley and Vert (2011)]	$d \cdot K \cdot n$	≥ 1
Segmentation binaire récursive (CART)		
[Gey and Lebarbier (2008)]	$dn\log(K)$	≥ 1
Segmentation binaire circulaire récursive (CBS)		
[Olshen AB et al. (2004)]	$n\log(n)$	1
[Zhang et al.(2010), Olshen AB et al. (2011)]	$d \cdot n^2$	≥ 1
Modèles de Markov Cachés (HMM)		
[Lai et al.]	n^2	1
[Chen et al. (2011)]	n^2	2

Segmentation binaire

Cas $d = 1$

- Tester $\mathcal{H}_0$: "No breakpoint" contre $\mathcal{H}_1$: "Exactly one breakpoint"
- Statistique du rapport de vraisemblance $\max_{1 \leq i \leq n} |Z_i|$

$$Z_i = \frac{\left(\frac{S_i}{i} - \frac{S_n - S_i}{n-i} \right)}{\sqrt{\frac{1}{i} + \frac{1}{n-i}}},$$

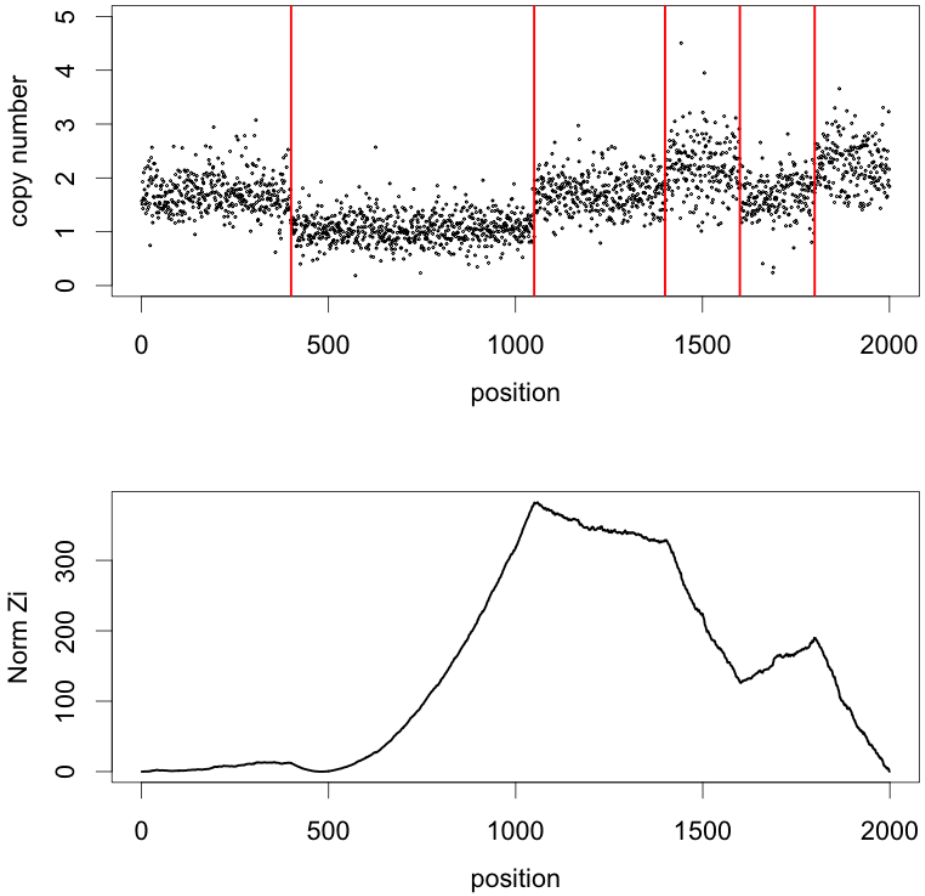
où $S_i = \sum_{1 \leq l \leq i} y_l$.

Si $d > 1$: la statistique devient $\max_{1 \leq i \leq n} \|Z_i\|_2^2$

Segmentation binaire récursive (RBS/CART)

Complexité : $O(d \cdot n \cdot \log(K))$

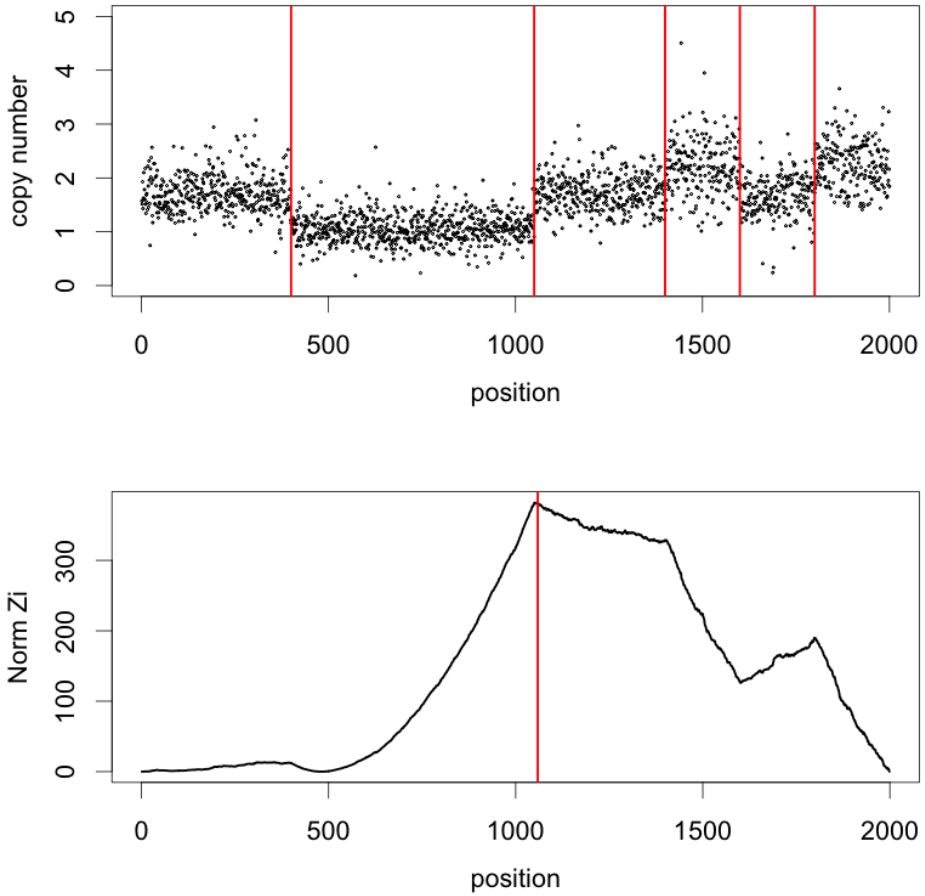
- Premier breakpoint
- Pour chaque i : on calcule Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$



Segmentation binaire récursive (RBS/CART)

Complexité : $O(d \cdot n \cdot \log(K))$

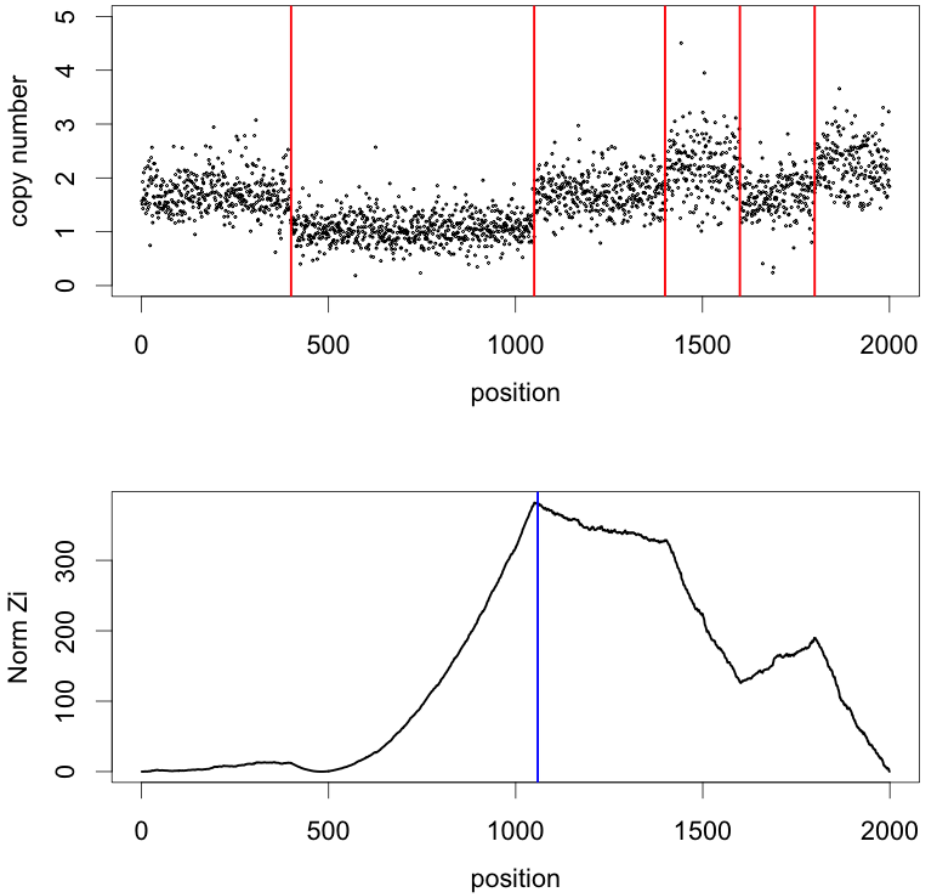
- Premier breakpoint
- Pour chaque i : on calcule Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$



Segmentation binaire récursive (RBS/CART)

Complexité : $O(d \cdot n \cdot \log(K))$

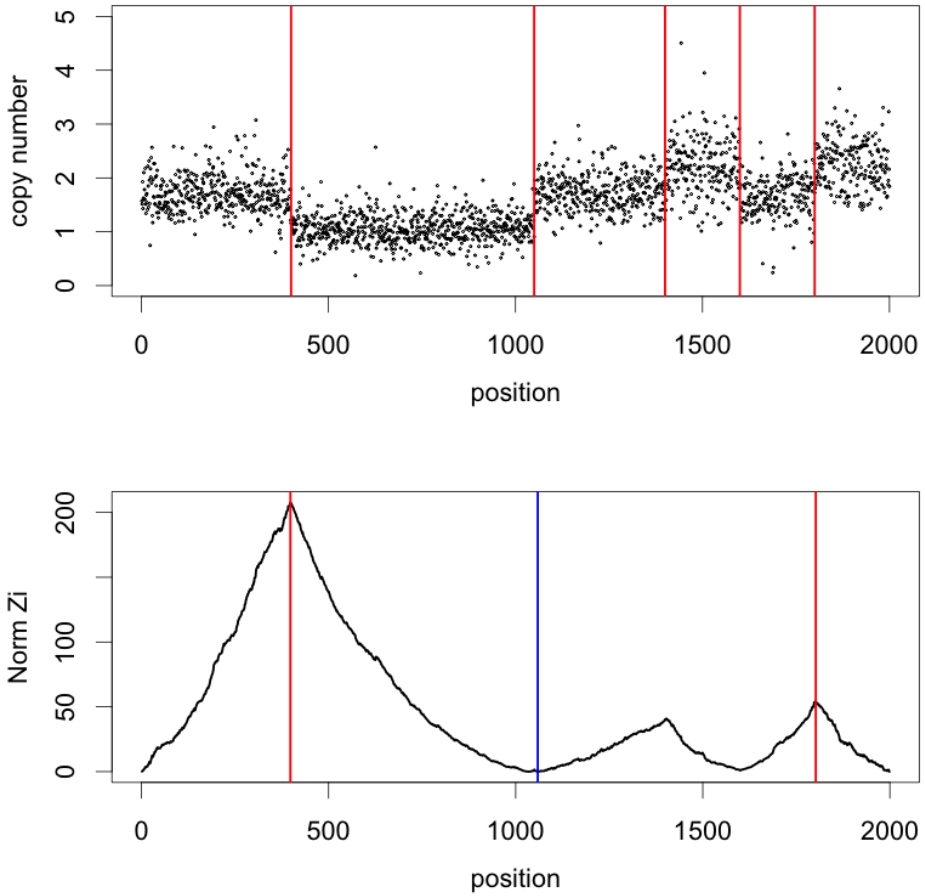
- Premier breakpoint
- Pour chaque i : on calcule Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$



Segmentation binaire récursive (RBS/CART)

Complexité : $O(d \cdot n \cdot \log(K))$

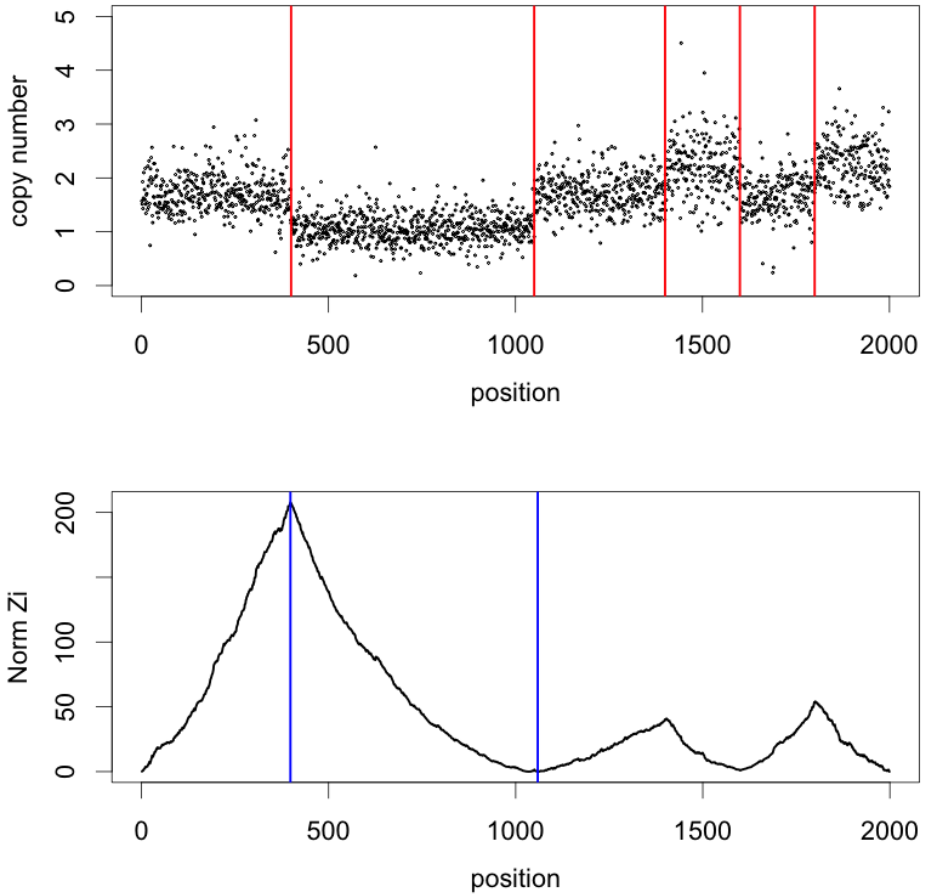
- Premier breakpoint
- Pour chaque i : on calcule Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$



Segmentation binaire récursive (RBS/CART)

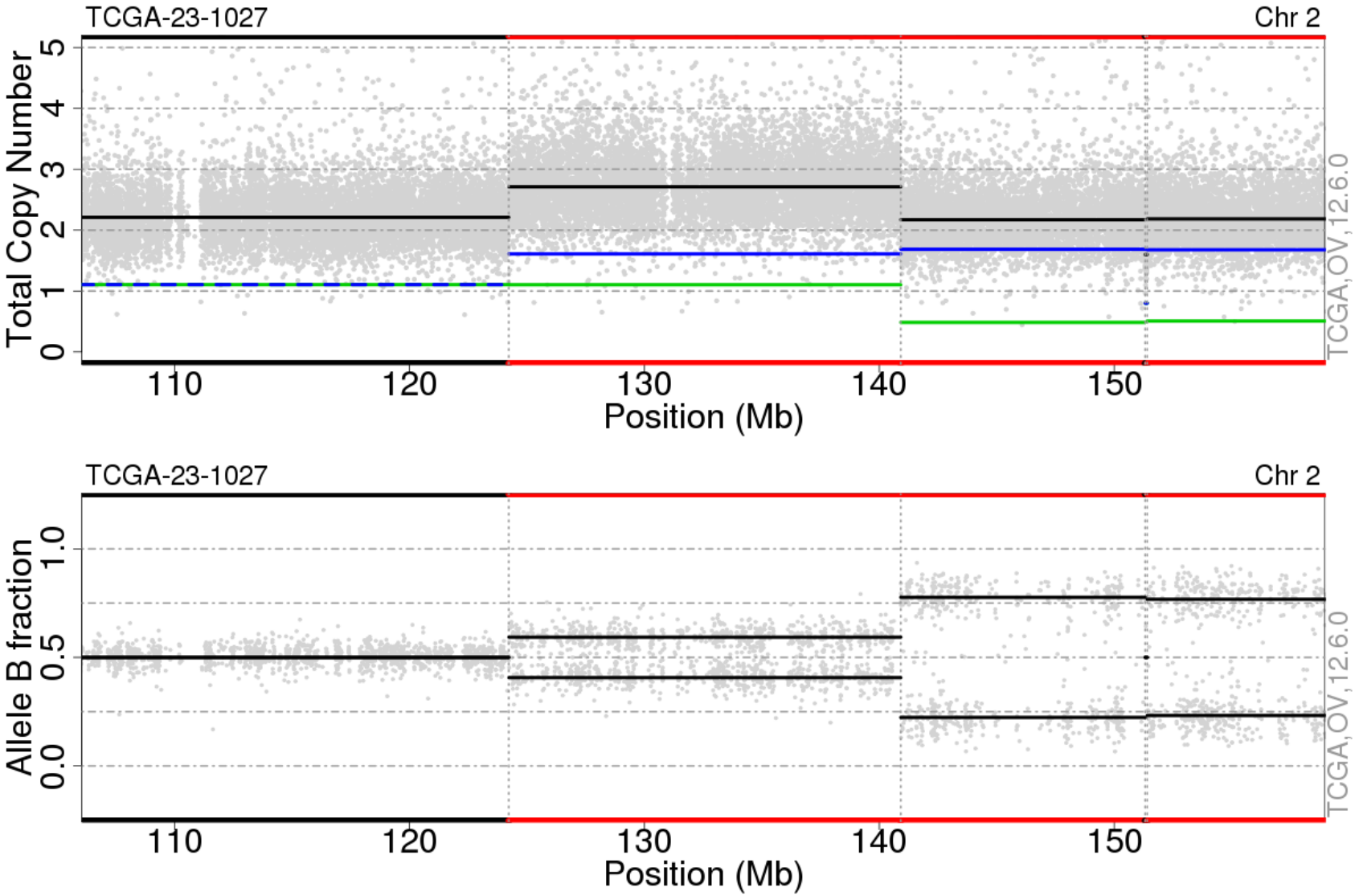
Complexité : $O(d \cdot n \cdot \log(K))$

- Premier breakpoint
- Pour chaque i : on calcule Z_i :
  $b_1 = \arg\max_{1 \leq i \leq n} \|Z_i\|_2^2$

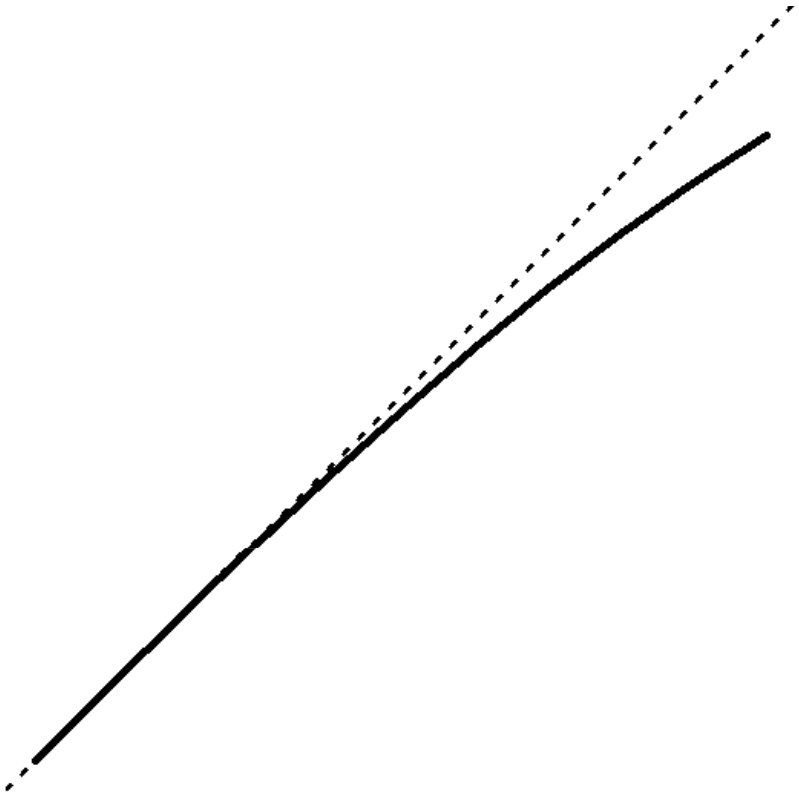


Analyse de données de nombre de copies d'ADN

Les nombres de copies d'ADN ne sont pas calibrés



Non-calibration : saturation du signal

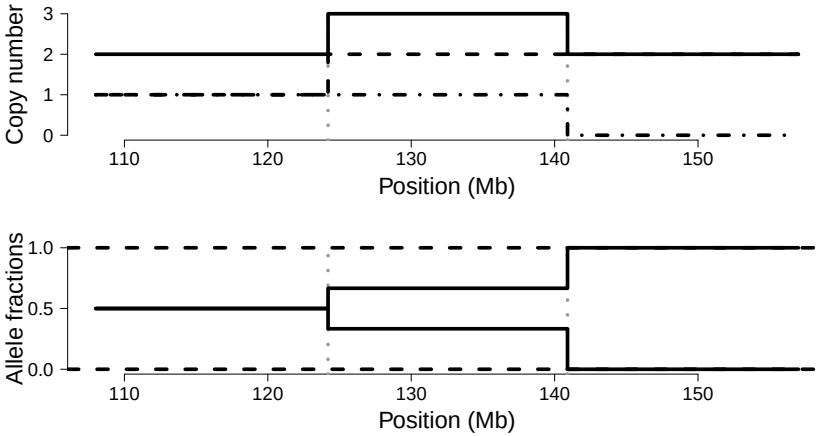


$$\mathbf{c} = f(\gamma) < \gamma$$

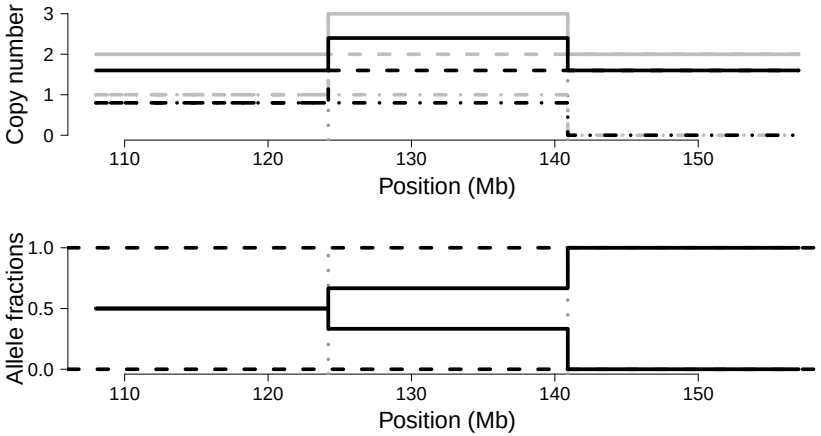
f est inconnue

Non-calibration : ploïdie et cellules normales

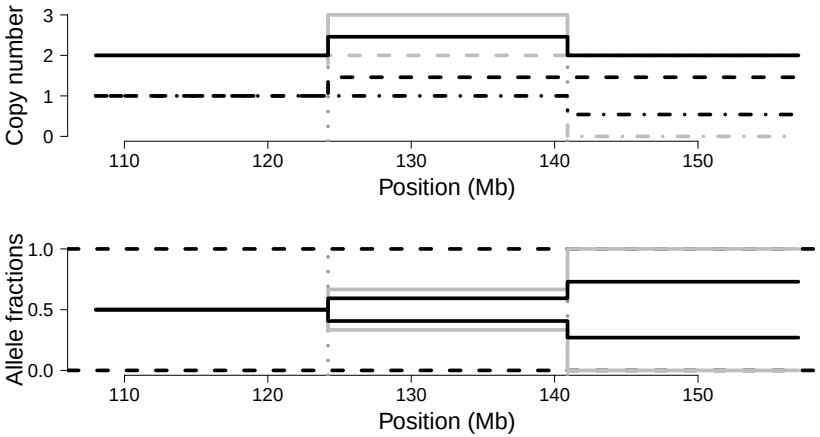
Tumeur pure, ploïdie = 2



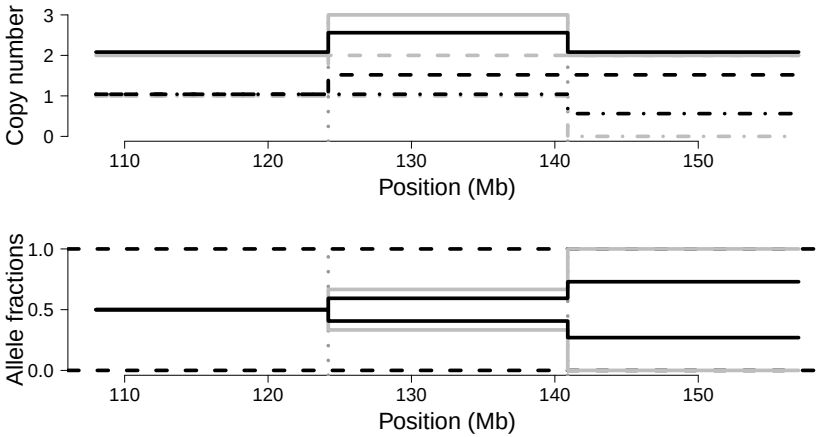
Tumeur pure, ploïdie > 2



Tumeur non pure, ploïdie=2



Tumeur non pure, ploïdie < 2



Cellules normales, ploïdie, et saturation du signal

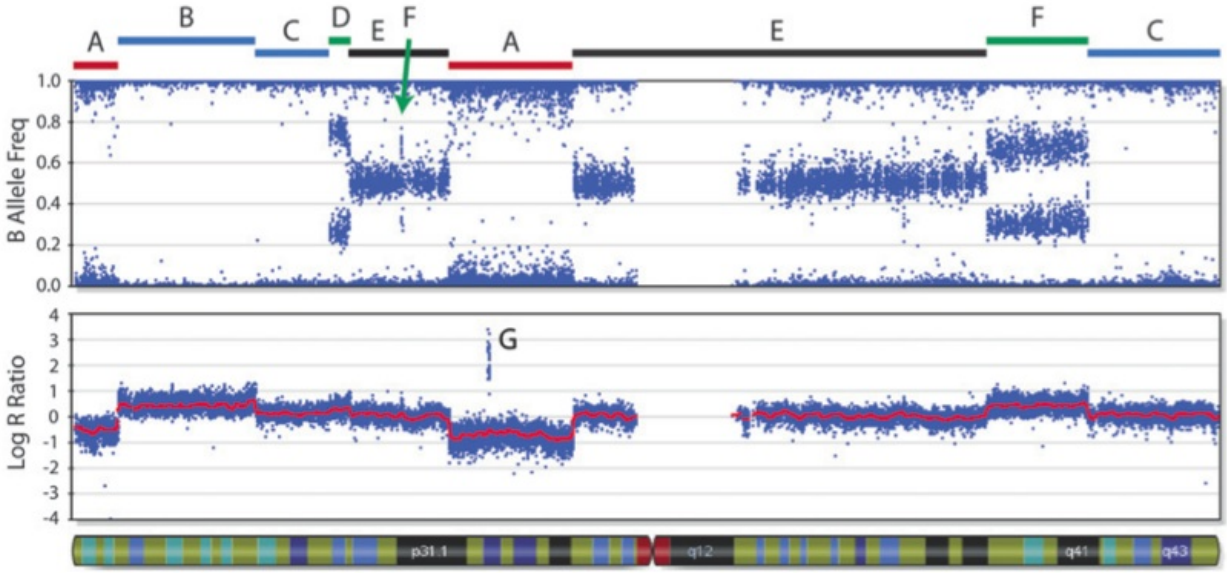
Causes de non-calibration

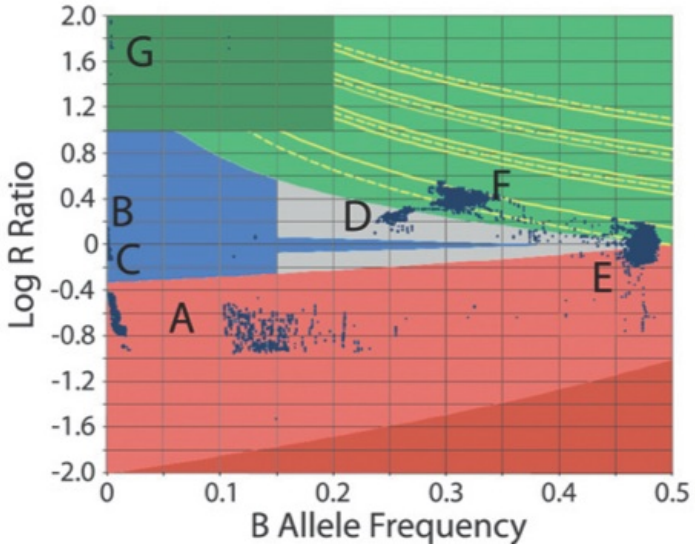
- saturation du signal
- présence de cellules normales dans l'échantillon "tumor sample"
- ploïdie : la quantité totale d'ADN est fixée par l'expérience

Remarques

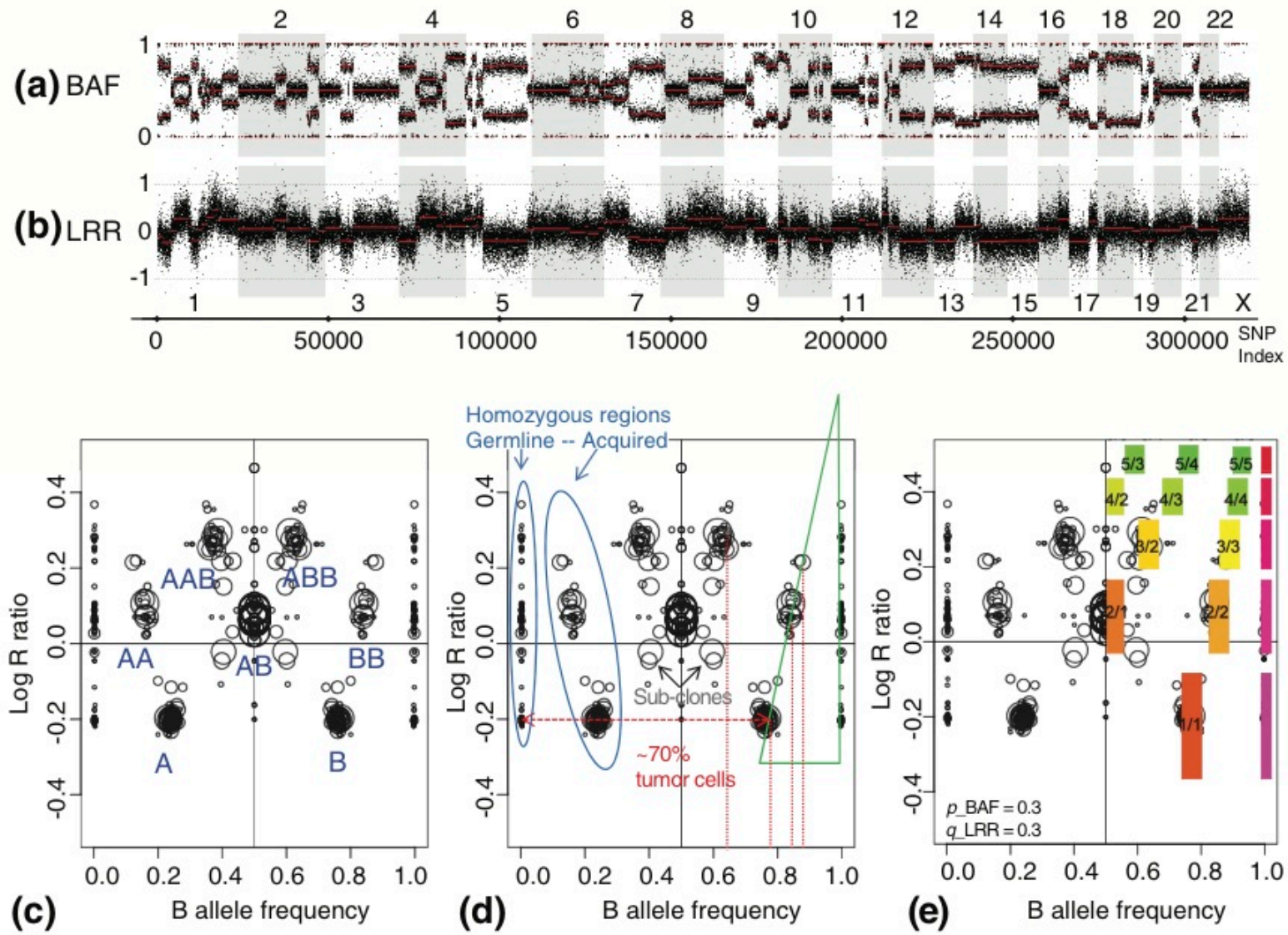
- la ploïdie de la tumeur **n'est pas identifiable**
- proportion de cellules normales et ploïdie de la tumeur sont des propriétés biologiques de l'échantillon
- la saturation du signal est un artefact de l'expérience
- sous le tapis : hétérogénéité tumorale

OverUnder : Attiyeh et al, Genome Research, 2009





GAP : Popova et al, Genome Biology, 2009



ASCAT : Van Loo et al, PNAS, 2010

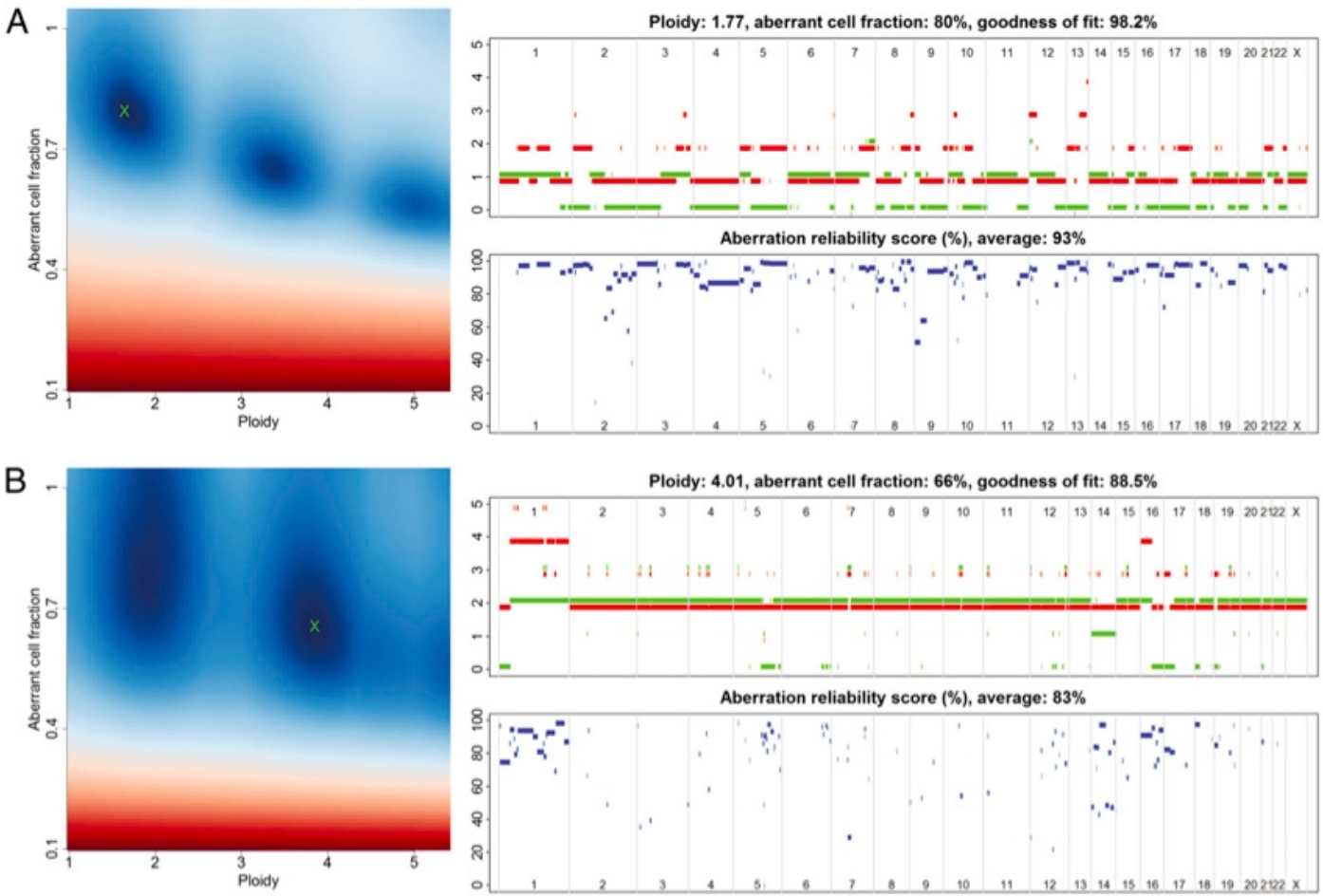


Fig. 1. ASCAT profiles and their calculation. Two examples are given: (*A*) a tumor with ploidy close to 2n and (*B*) a tumor with ploidy close to 4n. (*Left*) ASCAT first determines the ploidy of the tumor cells ψ_t and the fraction of aberrant cells ρ. This procedure evaluates the goodness of fit for a grid of possible values for both parameters (blue, good solution; red, bad solution; detailed in *Materials and Methods*). On the basis of this goodness of fit, the optimal solution is selected (green cross). Using the resulting tumor ploidy and aberrant cell fraction, an ASCAT profile is calculated (*Upper Right*), containing the allele-specific copy number of all assayed loci [copy number on the *y* axis vs. the genomic location on the *x* axis; green, allele with lowest copy number; red, allele with highest copy number; for illustrative purposes only, both lines are slightly shifted (red, down; green, up) such that they do not overlap; only probes heterozygous in the germline are shown]. Finally, for all aberrations found, an aberration reliability score is calculated (*Lower Right*).

Conclusions

- L'étape d'extraction du signal biologique est cruciale
- Des méthodes de segmentation performantes existent
- Elles doivent encore être adaptées à la segmentation jointe
- Ploïdie et présence de cellules normales sont des paramètres biologiques qui compliquent l'étiquetage des régions
- La plupart des méthodes sont implémentées en R ou Matlab

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.

Referérences : 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.

Tests multiples en génomique

6. Problèmes de tests multiples

7. Définition de mesures de risque adaptées

8. Interprétation des résultats d'analyse différentielle
   - Enrichissement d'ensembles de gènes
   - Tests multivariés sur des graphes

Référence

E. Roquain. (2011)
Type I error rate control in multiple testing : a survey with proofs.
Journal de la Société Française de Statistique (to appear)
http://hal.archives-ouvertes.fr/hal-00547965

Ingrédients d'un problème de tests multiples

- Observation du **comportement** d'un grand nombre de candidats
- Connaissance du comportement des candidats sous une **hypothèse** nulle

Exemple canonique : analyse différentielle de données d'expression

Matrice de données d'expression (données de Golub)

Niveaux d'expression de $m = 3051$ gènes pour $n = 38$ échantillons de deux types de leucémie :

AML	Acute Myeloblastic Leukemia	$n_1 = 11$
ALL	Acute Lymphoblastic Leukemia	$n_2 = 27$

Objectif

Trouver des gènes
différentiellement exprimés
entre AML et ALL

Exemple canonique : analyse différentielle de données d'expression

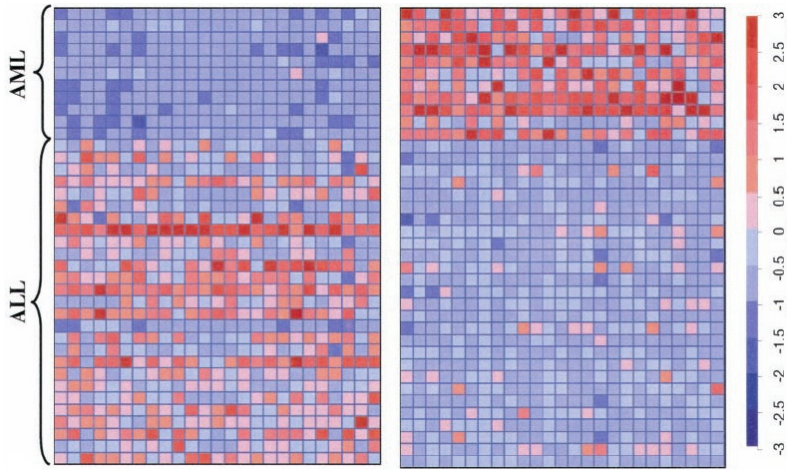
Matrice de données d'expression (données de Golub)

Niveaux d'expression de $m = 3051$ gènes pour $n = 38$ échantillons de deux types de leucémie :

AML	Acute Myeloblastic Leukemia	$n_1 = 11$
ALL	Acute Lymphoblastic Leukemia	$n_2 = 27$

Objectif

Trouver des gènes **différentiellement exprimés** entre AML et ALL



Le recours à la théorie des tests statistiques

Objectif : identifier les gènes dont l'expression diffère entre deux groupes

Approche gène par gène

- définition d'une statistique de test
  exemple : différence entre les moyennes d'expression des deux groupes
- contrôle des erreurs de type I (faux positifs) et/ou II (faux négatifs)

Problèmes spécifiques aux données de grande dimension

- n petit : le choix de la statistique de test est crucial
- m grand : problèmes de tests multiples

Tests d'hypothèses simples : vocabulaire

Notations pour l'étude de l'expression différentielle du gène g

Notions de base

$\mathcal{H}_0$	hypothèse nulle	absence d'expression différentielle de g
$\mathcal{H}_1$	hypothèse alternative	expression différentielle de g
$T_n(X)$	statistique de test	mesure empirique de l'expression différentielle
p	probabilité critique	probabilité que $T_n(X)$ soit supérieure $T_n(X_g)$ sous l'hypothèse $\mathcal{H}_0$

lorsque $\mathcal{H}_0$ est vraie (g n'est pas différentiellement exprimé)

risque α	erreur de type I	probabilité de rejeter $\mathcal{H}_0$
α	niveau	probabilité d'accepter $\mathcal{H}_0$

lorsque $\mathcal{H}_0$ est fausse (g est différentiellement exprimé)

risque β	erreur de type II	probabilité d'accepter $\mathcal{H}_0$
$1-\beta$	puissance du test	probabilité de rejeter $\mathcal{H}_0$

Statistiques de test usuelles

Exemple de test paramétrique : test de Student

- Idée : supposer les deux classes gaussiennes, et comparer les moyennes
- Statistique de test : différence des moyennes rapportée à la variabilité
- Sous $\mathcal{H}_0$, T_n suit une loi de Student, connue

Exemple de test non paramétrique : test de Wilcoxon

- Idée : rassembler les deux échantillons, et ranger le vecteur obtenu
- Statistique de test : somme des rangs d'une des classes
- Sous $\mathcal{H}_0$, T_n suit une loi multinomiale de paramètres connus

Mise en évidence du problème

Contrôle le niveau individuel des tests est insuffisant

- On effectue un test par gène $\Rightarrow$ m p-values $(p_1, \ldots p_m)$
- En seuillant au niveau α, on attend $m\alpha$ faux positifs

$\Rightarrow$ le nombre d'erreurs **dépend du nombre de tests**. Nécessité d'ajuster le seuil en fonction de m pour rendre le résultat interprétable.

Procédures de tests multiples

- Définition d'un risque qui dépend de l'**ensemble** des hypothèses
- Construction d'un algorithme permettant de contrôler ce risque
- On considèrera ici des algorithmes reposant sur les p-values individuelles

Applications autres que l'analyse différentielle

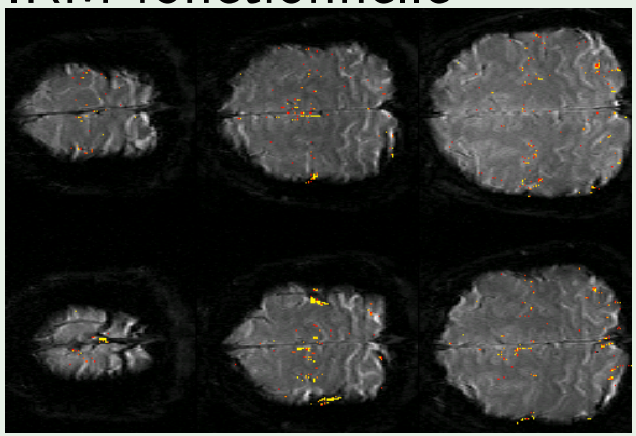
Analyse de données génomiques

- Inférence de réseaux génétiques à partir de corrélations
- Spectrométrie de masse
- Etudes d'association pangénomique (GWAS)
- Corrélation entre nombre de copies d'ADN et expression

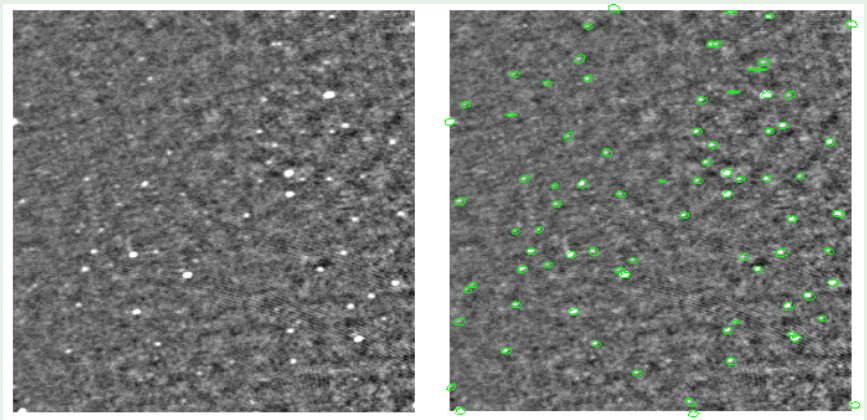
Autres domaines d'application

(images : C. Genovese)

IRM fonctionnelle



Astronomie



Hypothèses sur la loi des p-values sous $\mathcal{H}_0$

Exemple : localisation Gaussienne, tests unilatéraux

- On observe $(X_i)_{1\leq i\leq m} \sim \mathcal{N}(\mu, \Sigma)$, avec $\Sigma_{ii} = 1$
- On teste $\mathcal{H}_0^i : \mu_i \leq 0$ contre $\mathcal{H}_1^i : \mu_i > 0$

Si on définit les p-values : $p_i = \bar{\Phi}(X_i)$, où $\bar{\Phi}(x) = P(\mathcal{N}(0,1) > x)$, **alors** on a

$$\begin{aligned} P_{\mathcal{H}_0^i}(p_i \leq t) &= P_{\mathcal{H}_0^i}(X_i \geq \bar{\Phi}^{-1}(t)) \\ &\leq P_{\mathcal{H}_0^i}(X_i - \mu_i \geq \bar{\Phi}^{-1}(t)) = t \end{aligned}$$

Domination stochastique (SD) par la loi $\mathcal{U}[0,1]$

On supposera que les p-values satisfont $P_{\mathcal{H}_0^i}(p_i \leq t) \leq t$

Mesures de risque de type I pour les tests multiples

Issues possibles d'une procédure de tests multiples

	acceptations	rejets	
vraies nulles	U	V	m_0
vraies alternatives	S	T	$m - m_0$
	$m - R$	R	m

Mesures de risque usuelles

Family-wise error rate :	$FWER = P(V \geq 1)$
k-Family-wise error rate	$k - FWER = P(V \geq k)$
False Discovery Proportion	$FDP = V/max(R, 1)$
False Discovery Rate	$FDR = E(FDP)$
γ-quantile du FDP	$P(FDP > \gamma)$

Contrôle du Family-Wise Error Rate (FWER)

Principe

- idée contrôle du risque (I) associé aux m tests
- $FWER$ = probabilité d'avoir au moins un faux positif

Exemples de méthodes de contrôle du FWER au niveau α

Auteur	seuil t	type de contrôle
Bonferroni	α/m	pas d'hypothèse d'indépendance des gènes calculé sous l'hypothèse $m_0 = m$
Šidàk	$1-(1-\alpha)^{1/m}$	suppose l'indépendance des gènes calculé sous l'hypothèse $m_0 = m$

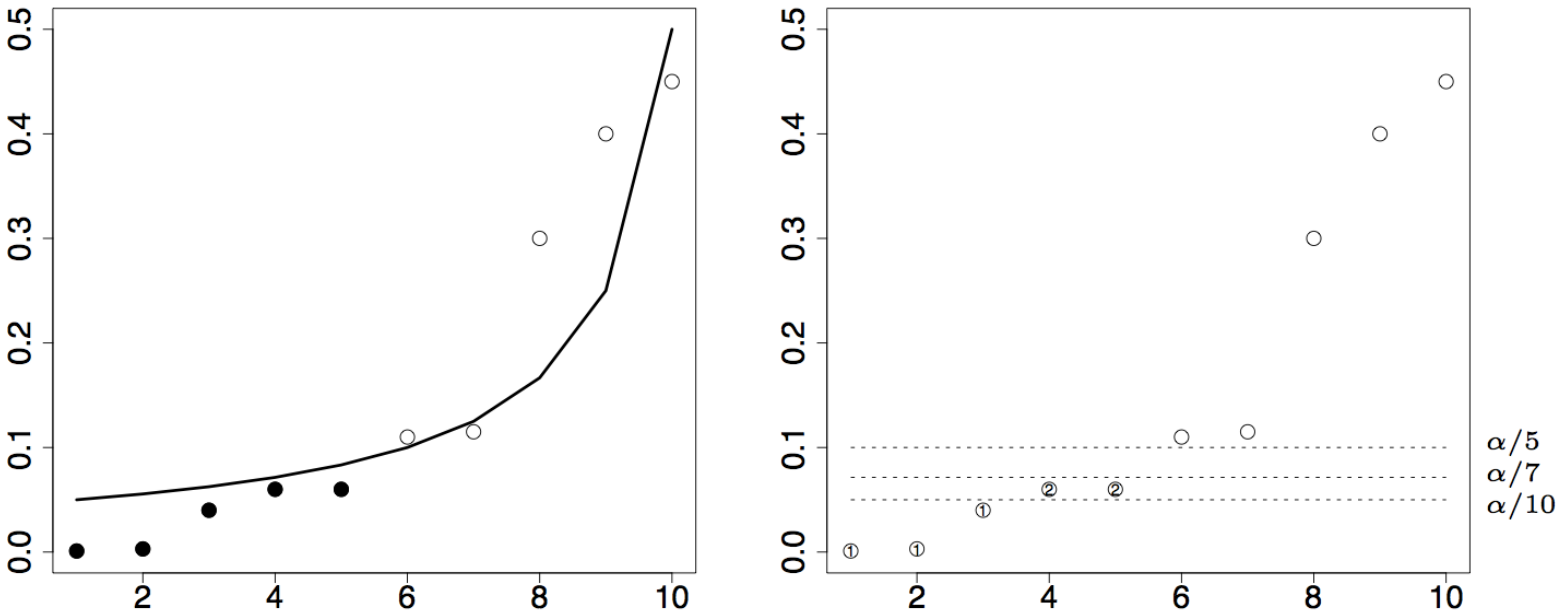
Variantes et améliorations

- incorporation directe d'une estimation de $\pi_0 = m_0/m$: "single step"
- incorporation indirecte d'une minoration de $\pi_0 = m_0/m$: "step down"

Exemple de procédure "step-down" (Holm, 1979)

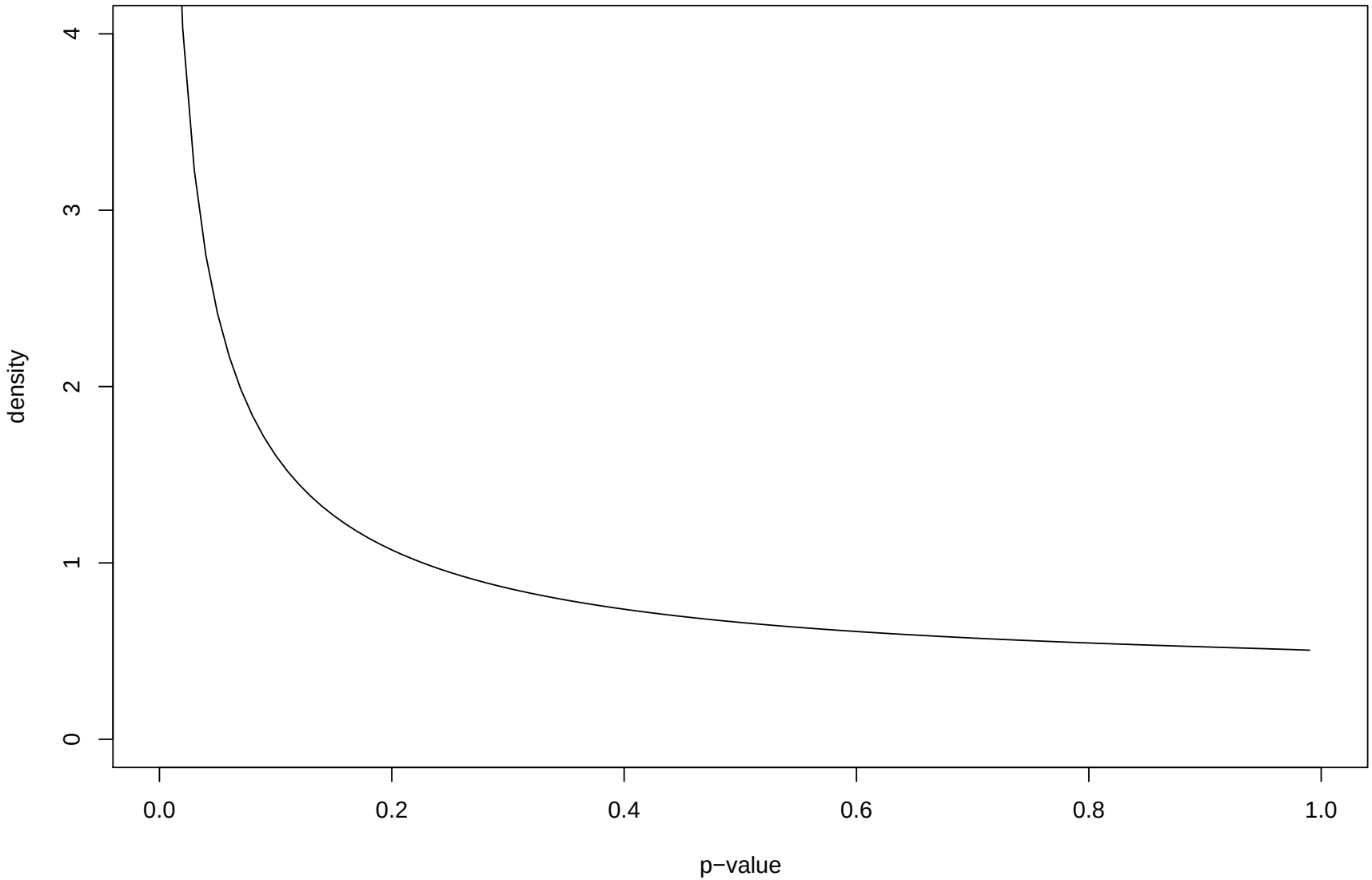
Description de la procédure au niveau α (figure de gauche)

1. p-values ordonnées : $p_{(1)} \leq \ldots \leq p_{(m)}$
2. On définit $\widehat{I} = Max\left\{i | \forall k \leq i, p_{(k)} \leq \alpha/(m-k+1)\right\}$
3. On rejette les p-values inférieures à $\alpha/(m-\widehat{I}+1)$

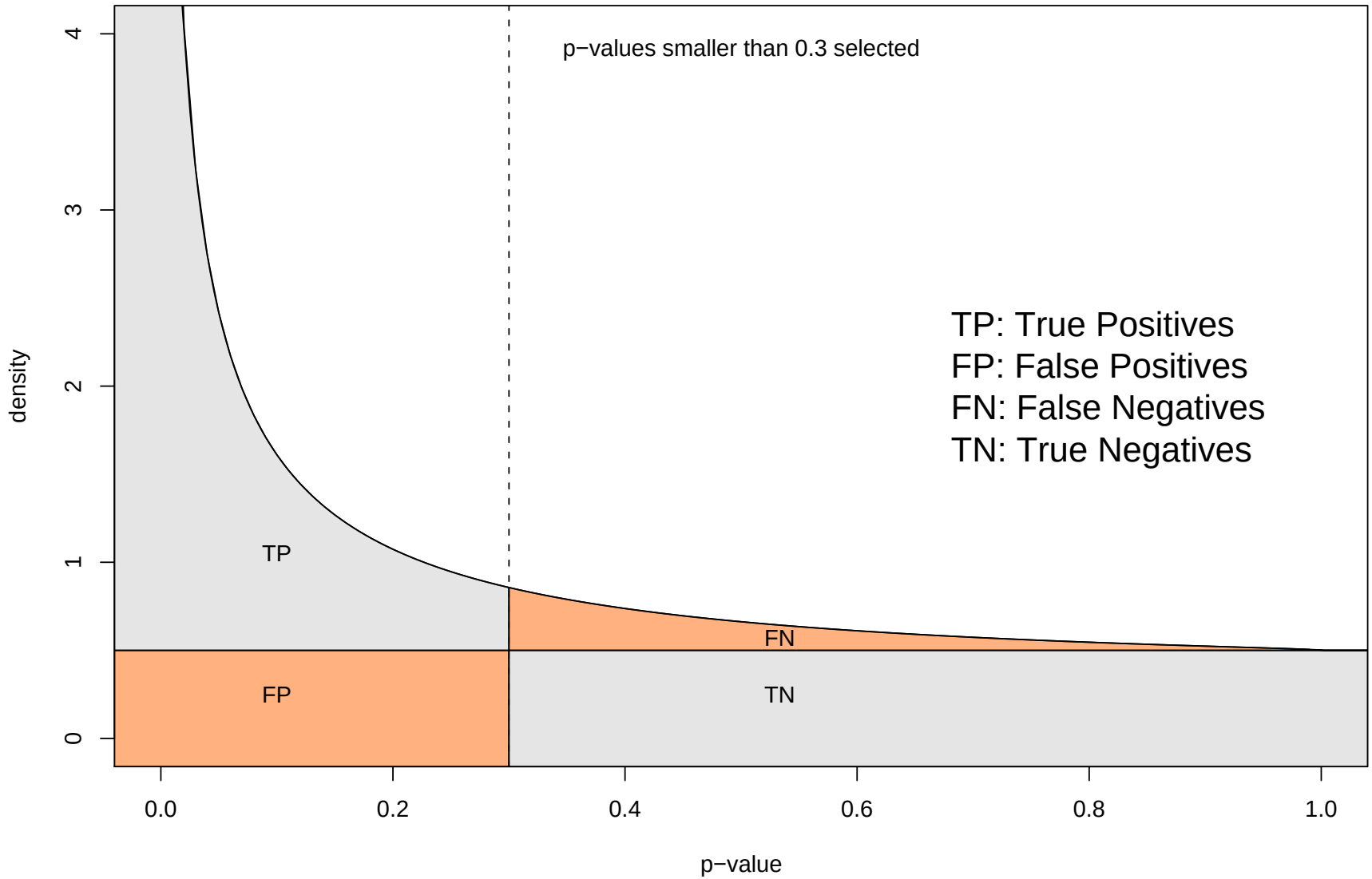


Interprétation (à droite) : raffinements successifs d'un majorant de π_0.

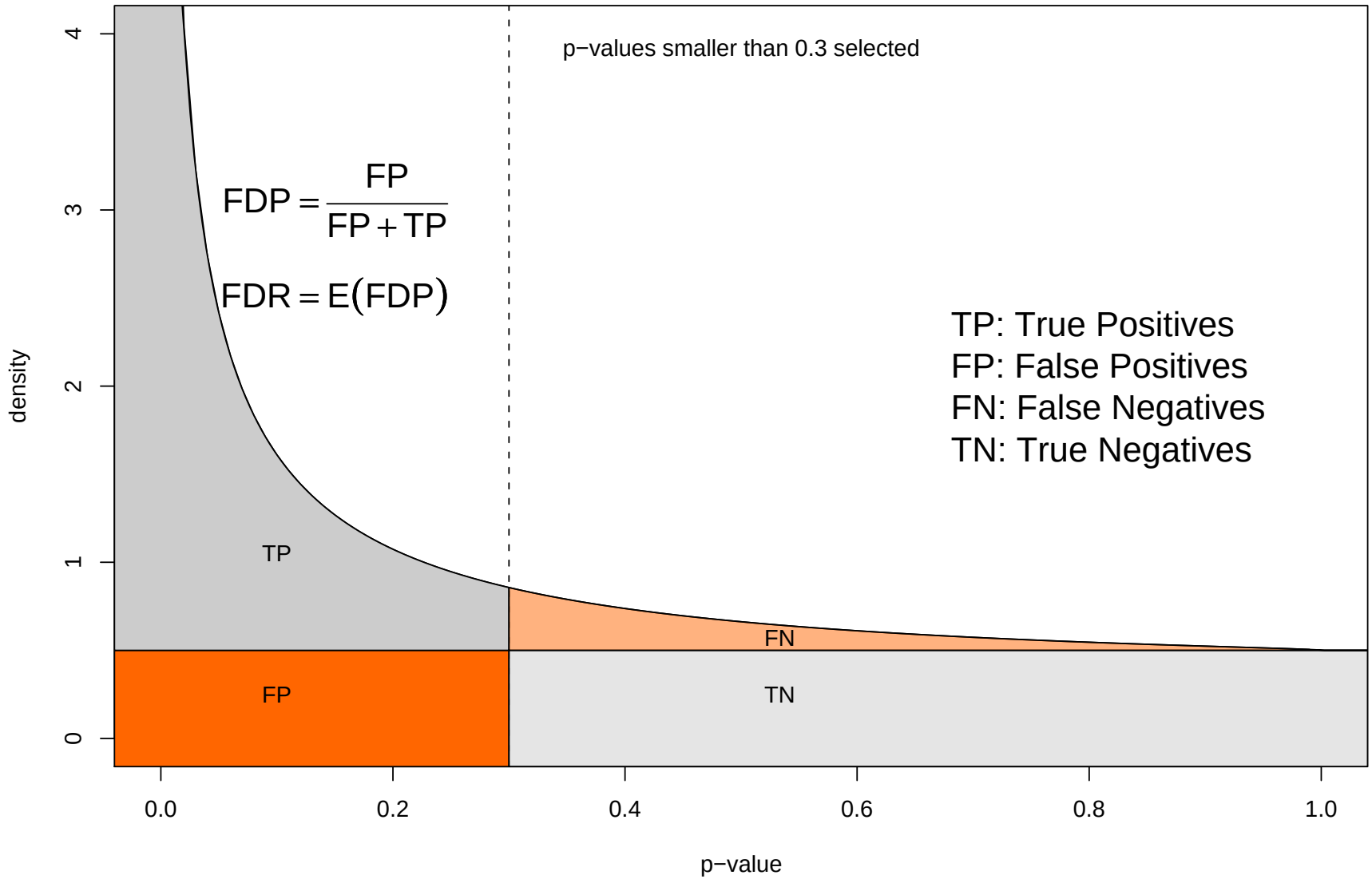
False Discovery Proportion et False Discovery Rate



False Discovery Proportion et False Discovery Rate



False Discovery Proportion et False Discovery Rate

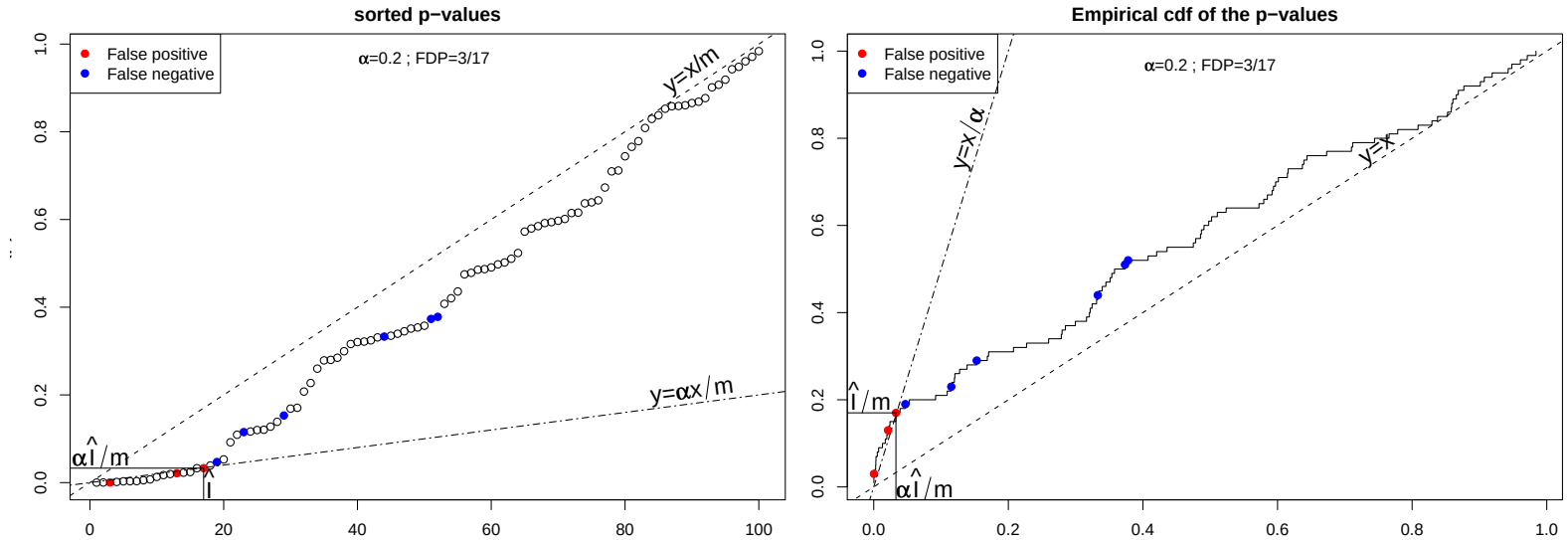


La procédure de Benjamini et Hochberg (1995)

Définition et interprétation graphique

Description de la procédure au niveau α

1. p-values ordonnées : $p_{(1)} \leq \ldots \leq p_{(m)}$
2. On définit $\widehat{l} = Max\left\{k | p_{(k)} \leq \alpha \frac{k}{m}\right\}$
3. On rejette les p-values inférieures à $\alpha \widehat{l}/m$



La procédure de Benjamini et Hochberg (1995)

Propriétés

Contrôle du FDR au niveau $\pi_0\alpha$ si les p-valeurs tirées sous $\mathcal{H}_0$ sont

- soit indépendantes
- soit positivement associées (PRDS)

Améliorations de l'approche de Benjamini et Hochberg

- contrôle dans d'autres cadres de dépendance entre les gènes
- estimation de π_0 pour un contrôle plus fin du FDR

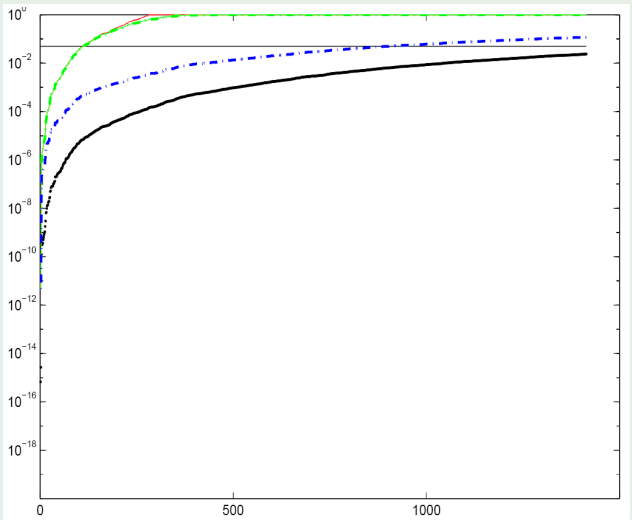
FDR et FWER : Comparaison sur des données réelles

Matrice de données d'expression (données de Golub)

Niveaux d'expression de $m = 3051$ gènes pour $n = 38$ échantillons de deux types de leucémie :

AML	Acute Myeloblastic Leukemia	$n_1 = 11$
ALL	Acute Lymphoblastic Leukemia	$n_2 = 27$

Distribution des p-values ajustées (d'après S. Robin)



Nombre de gènes sélectionnées pour $\alpha = 0.05$:

- Bonferroni : 111
- Sidak : 113
- Benjamini-Hochberg : 903
- p-values brutes : 1887

Tests multiples en génomique

Motivation

Interprétation des résultats d'analyse différentielle

Approche classique en 3 étapes

1. Evaluation du degré d'expression différentielle entre 2 groupes
2. Choix d'un seuil de significativité (FDR)
   $\Rightarrow$ **liste** de gènes différentiellement exprimés
3. Recherche d'enrichissement de cette liste en annotations

Limitations

Les résultats dépendent du choix du seuil de significativité
Du point de vue biologique, ce seuil reste arbitraire

Propositions

1. Rechercher l'enrichissement d'**ensembles** de gènes d'intérêt
2. Faire directement des tests multivariés

Tests multiples en génomique

Principales étapes d'une analyse d'enrichissement

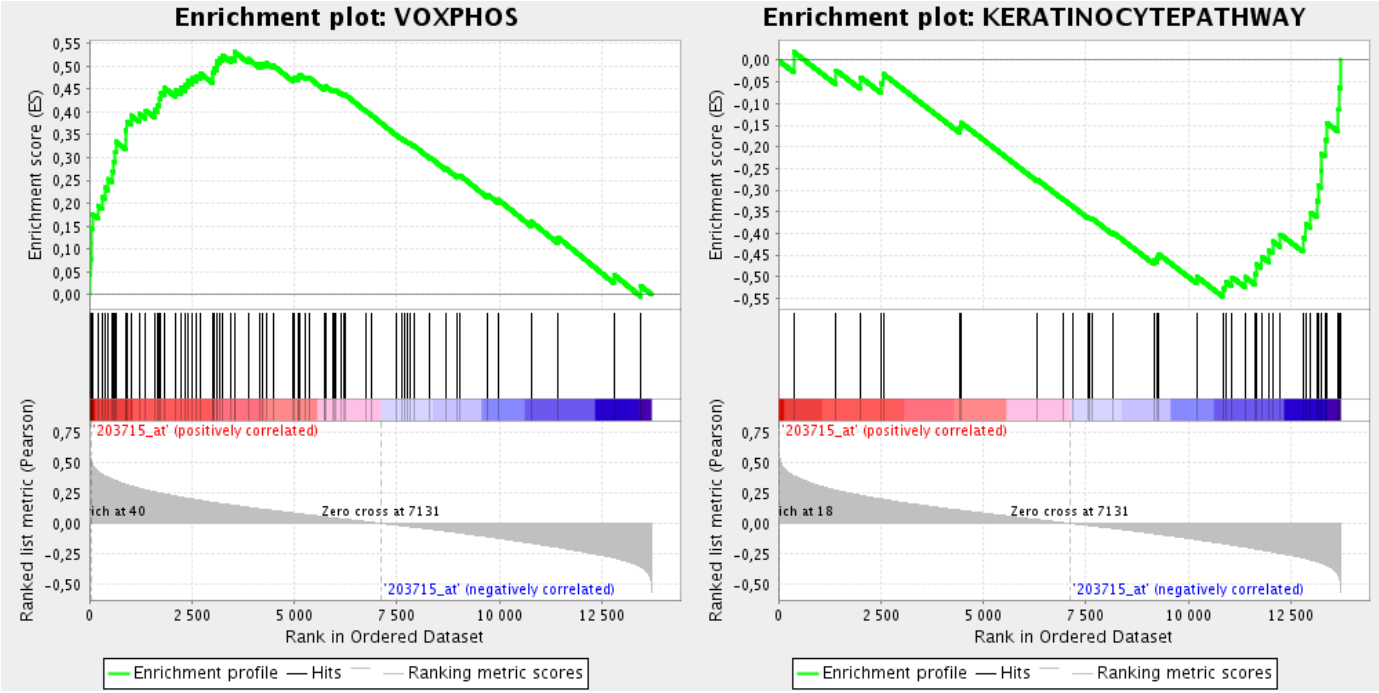
1. Mesure d'expression différentielle
2. Score d'enrichissement
3. Hypothèse testée
4. Estimation de la significativité
5. Correction des tests multiples

Score d'enrichissement d'un ensemble de gènes E

Comparer la distribution des mesures d'association des gènes de E et $\bar{E}$

Exemples de scores

- test de localisation paramétrique : T-test
- test de localisation non paramétrique : test de Wilcoxon
- test d'adéquation : test de Kolmogorov-Smirnov



Définition d'une hypothèse nulle et d'une *p*-valeur

d'après Goeman et Bühlmann, *Bioinformatics*, 2007

Définition d'une hypothèse nulle

Tests compétitifs l'expression différentielle de E est comparée à celle de son complémentaire

Tests intrinsèques l'expression différentielle de E est évaluée indépendamment des gènes qui ne sont pas dans E

Les tests compétitifs rejettent plus d'hypothèses.

Définition d'une *p*-valeur

Échantillonnage des gènes : les gènes de E sont supposés iid

Échantillonnage des individus : les patients sont supposés iid

Les gènes ne sont justement pas indépendants : c'est pour cela qu'on teste des **ensembles** de gènes

Estimation de la significativité

Méthodes paramétriques

1. On suppose que sous $\mathcal{H}_0$, le score d'enrichissement suit une loi connue ($\mathcal{N}(0,1)$)
2. On **calcule** une p-value à l'aide de cette loi

Méthodes non paramétriques

La distribution des scores d'enrichissement, est **estimée** à l'aide de rééchantillonnages.

Contrôle des tests multiples avec dépendance

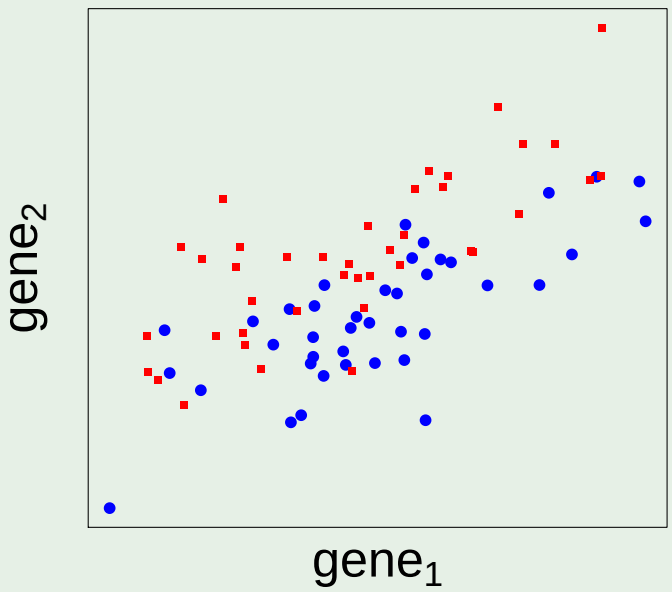
FWER Westfall-Young, Bonferroni

FDR Benjamini-Yekutieli, Yekutieli-Benjamini, ...

Tests multiples en génomique

Limitations des méthodes d'enrichissement

Exemple : 2 dimensions



Tests univariés et tests multivariés

- Aucun des deux gènes pris individuellement n'est différentiellement exprimé
- En deux dimensions une différence entre les deux classes apparaît

Test multivariés sur des graphes

Test multivarié sur un ensemble de gènes E

- On suppose que les échantillons suivent une loi **normale multivariée** dans l'espace de E
- On teste $\mu_1 = \mu_2$ avec la statistique de T^2 Hotelling

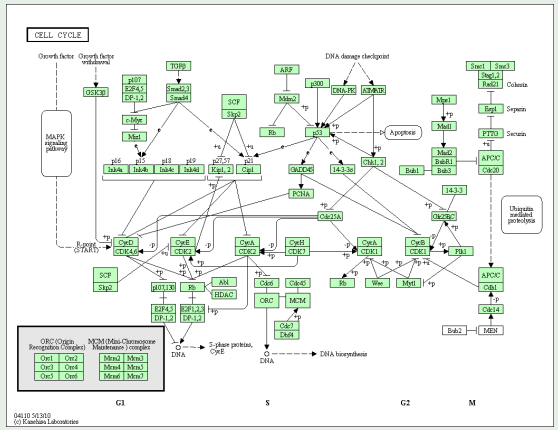
\+ Certaines limitations de l'analyse d'enrichissement sont levées

\- **Perte de puissance très rapide quand la dimension augmente**

Test structuré

On utilise une information *a priori* sur $\mu_2 - \mu_1$ pour **réduire la dimension** et **gagner en puissance**.

Un *a priori* possible : $\mu_2 - \mu_1$ est "cohérent" avec un réseau de régulation connu.

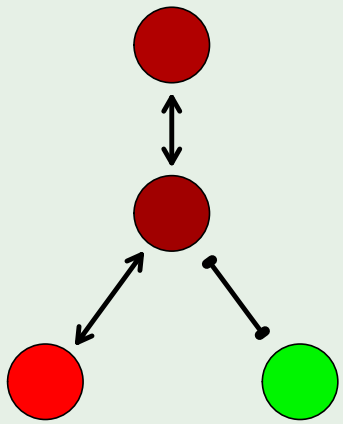


Profils d'expression : des fonctions sur des graphes

Une fonction f sur un graphe $\mathcal{G}$ **associe une valeur réelle à chaque noeud** du graphe.

Exemples

- Expression des gènes dans le réseau (x_i),
- Moyenne des expressions des patients par classe $(\bar{x}_1)$,
- Difference entre ces moyennes $(\bar{x}_1 - \bar{x}_2)$.



Analyse harmonique sur des graphes

Décomposition de Fourier d'un graphe

- Les coefficients $\tilde{f}_i$ sont les projection de f sur les vecteurs propres de $\mathcal{L}$:

$$\tilde{f}_i \triangleq u_i^\top f, \; i = 1, \ldots, |\mathcal{V}|.$$

- Transformée inverse :

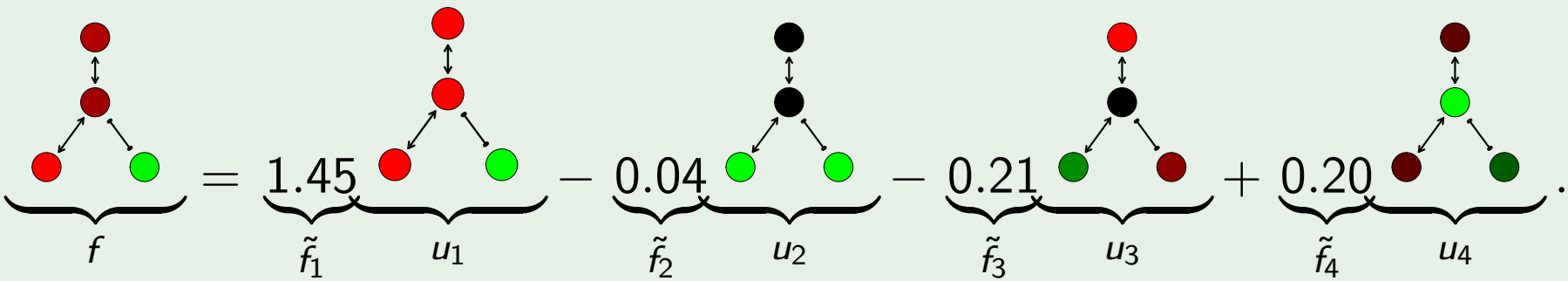
$$f = \sum_{i=1}^{|\mathcal{V}|} \tilde{f}_i u_i.$$

f et $\tilde{f}$ sont **deux façons duales d'écrire la même fonction** :

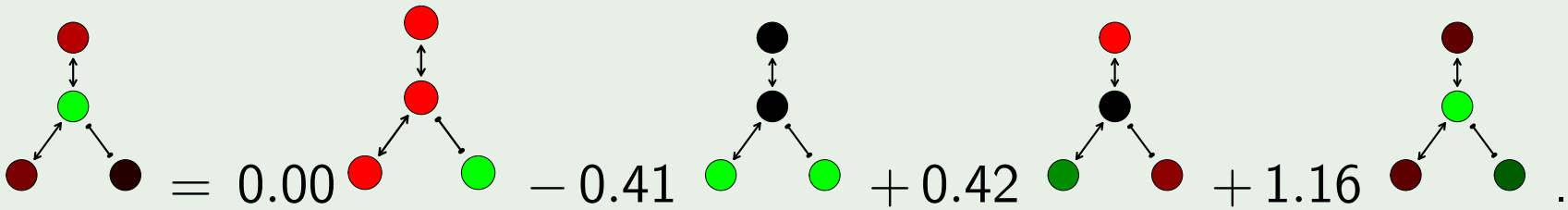
- Comme un ensemble de valeurs pour les noeuds f_i (*par ex.* les différences d'expression),
- Comme les coefficients $\tilde{f}_i$ de la décomposition du Fourier du graphe.

Analyse harmonique sur des graphes

Exemple 1 (shift régulier)



Example 2 (shift irrégulier)



Remarque

Les fonctions **régulières** ont de **grands** coeff. au **début** du spectre

Test dans l'espace du graphe ou l'espace de Fourier

Remarques

- Les tests dans les deux espaces sont équivalents : $T^2 = \tilde{T}^2$
- Plus généralement :

  T^2 calculée après **filtrage** des fréquences de rang supérieur à k

  $=$

  $\tilde{T}^2_k$ dans l'espace de Fourier restreint aux k premiers coefficients

Intuition

Le shift en distribution est cohérent avec un graphe connu

$\Downarrow$

Concentré dans les basses fréquences

$\Downarrow$

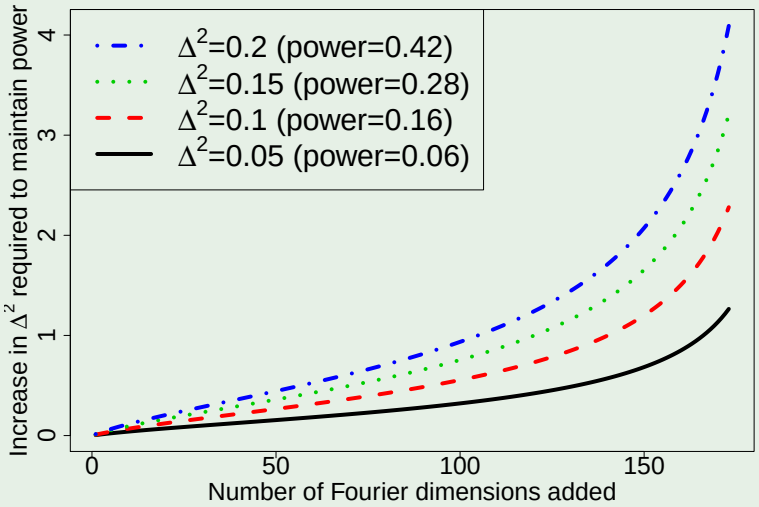
Le test dans les premiers coefficients de Fourier est plus puissant

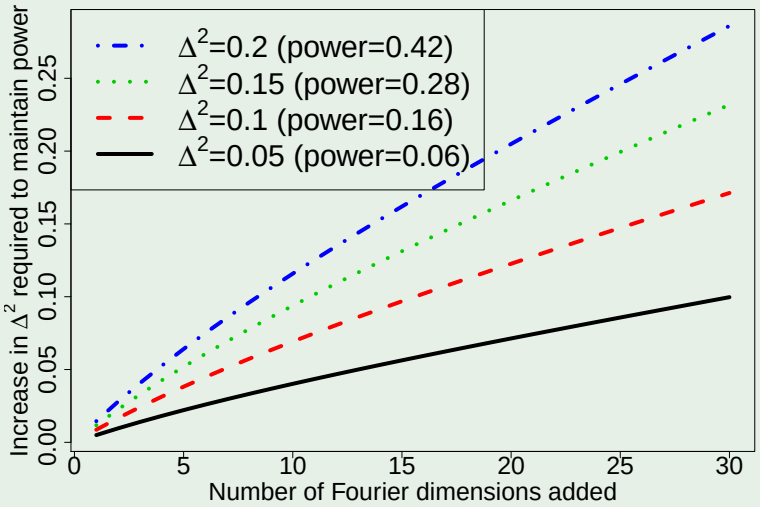
Gain en puissance

Lemma

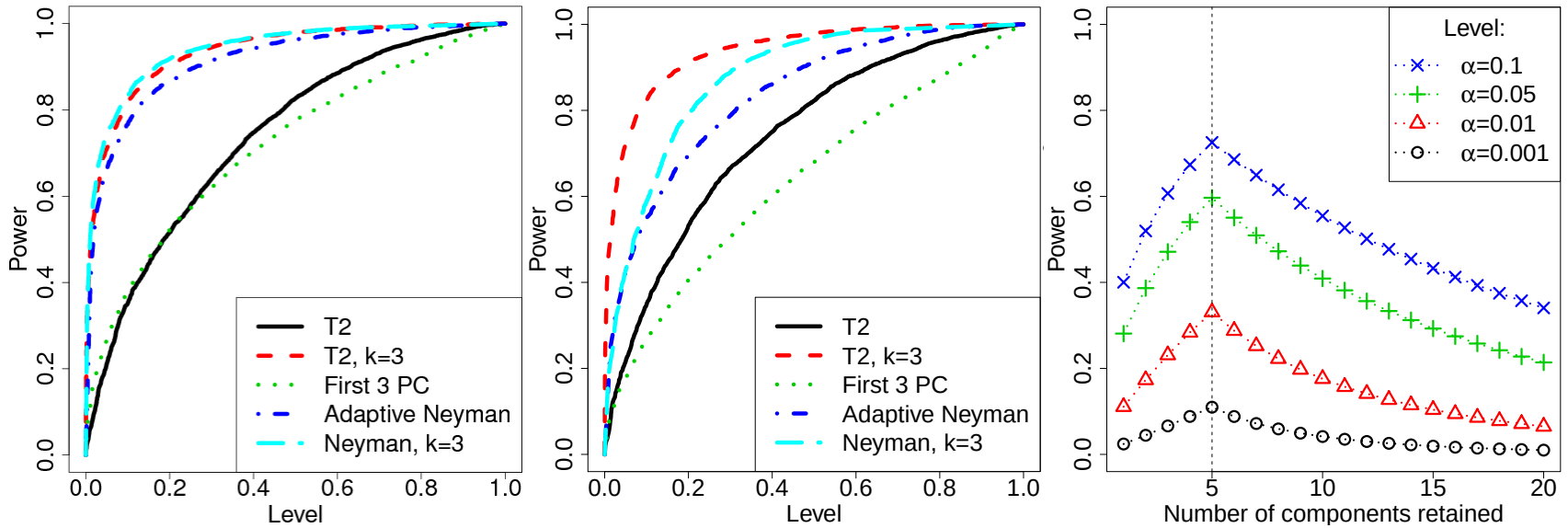
Pour tout niveau et tout nombre de coefficients de Fourier, **préserver la puissance** *du test du T^2 dans l'espace de Fourier suite à l'ajout d'un coefficient nécessite une* **augmentation strictement positive** *du shift entre les deux groupes*

Illustration



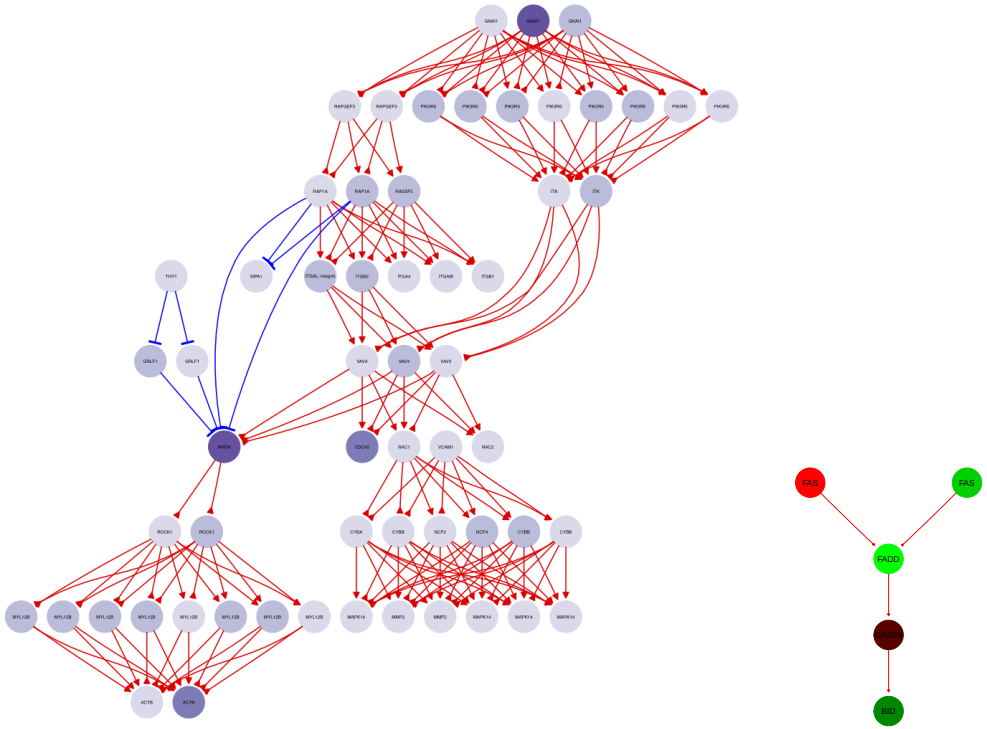


Gain en puissance : données simulées



- Courbes ROC pour la détection d'un shift régulier pour différentes statistiques de test : covariance diagonale (gauche) et diagonale par blocs (milieu)
- Droite : puissance du test T^2 dans l'espace de Fourier : shift uniformément réparti parmi les 5 premiès coefficients de Fourier

Données de cancer du sein et données KEGG



Difference entre l'expression moyenne des patients résistant et non-résistants au Tamoxifen, pour des gènes présents dans deux réseaux de régulation KEGG

Classification et régression en génomique

9. Prédiction de classe (classification supervisée)
   - Introduction
   - Exemples de méthodes de classification

10. Questions statistiques soulevées
    - Estimation des performances
    - Réduction de dimension pour la classification
    - Sélection de variables par pénalisation

11. Performance, stabilité, interprétabilité des signatures moléculaires
    - Construction d'une signature moléculaire
    - Méthodes comparées
    - Critères pour l'évaluation d'une signature moléculaire
    - Résultats

Classification et régression en génomique

Contexte

Problématique

Prédiction d'une variable clinique ou biologique à partir de données génomiques (nombre de copies d'ADN ou expression par exemple) :

- distinguer les individus sains de ceux atteints par une maladie
- prédire la réponse thérapeutique ou la survie
- inférence de réseaux de régulation

Principe

1. apprentissage d'une règle sur un ensemble d'entraînement
2. application de cette règle à un nouvel échantillon

Régression : la variable à prédire est quantitative

Classification : la variable à prédire est qualitative

Construction d'un prédicteur

Cadre statistique et notations

- $\mathcal{P}$: loi de probabilité sur $(\mathcal{X}, \mathcal{Y})$ (inconnue)
- Observations $(X_i, Y_i)_{1 \leq i \leq n}$: n-échantillon de loi $\mathcal{P}$
- $t : \mathcal{X} \mapsto \mathcal{Y}$: *prédicteur*
- $\ell : \mathcal{Y}^2 \to \mathbb{R}$: fonction de perte
- $R : t \mapsto \mathbb{E}_{\mathbb{P}} \ell(t(X), Y)$: risque du prédicteur t associé à ℓ

Exemple : fonction de perte 0-1 : $\ell_{0-1} : (s, t) \mapsto \mathbf{1}_{s \neq t}$

- Risque associé : probabilité de mauvaise classification
- $\eta : x \mapsto \mathbb{P}(Y = 1 | X = x)$ minimise ce risque
  - En régression, on cherche à estimer la **fonction de régression** η
  - En classification, on cherche à estimer le **classifieur de Bayes** $t_b : x \mapsto \mathbf{1}_{\eta(x) > \frac{1}{2}}$

Prédiction de données génomiques

Propriétés souhaitables pour un prédicteur

- performances
- stabilité/robustesse
- interprétabilité
- vitesse d'apprentissage

Problèmes spécifiques aux données de grande dimension

p grand : besoin de sélectionner les variables

n petit : nécessité de méthodes de validation appropriées

Ces questions sont illustrées ici pour le problème de **classification** mais on peut transposer au problème de régression

Classification et régression en génomique

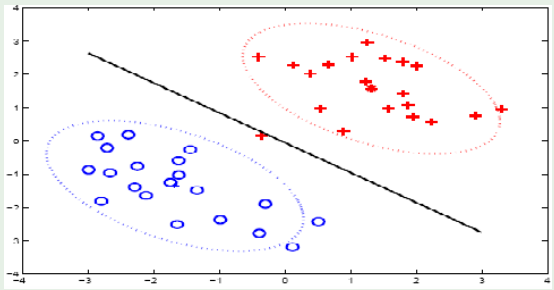
Analyse discriminante (de Fisher)

Un exemple de méthode paramétrique

Hypothèse : lois conditionnelles des classes gaussiennes

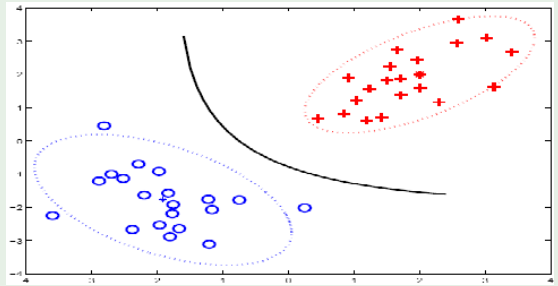
- estimation des lois *a posteriori* des classes
- attribution à un nouvel individu de la classe de proba *a posteriori* maximale

Discrimination linéaire



- variances identiques
- frontière linéaire en l'expression des gènes

Discrimination quadratique



- variances non identiques
- frontière quadratique en l'expression des gènes

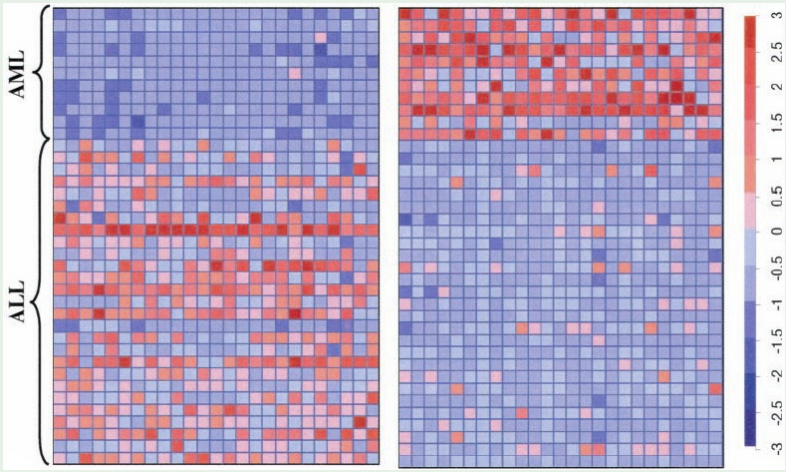
k plus proches voisins (kNN)

Un exemple de méthode non paramétrique

Principe

- aucune hypothèse sur la loi *a priori* des classes
- on attribue à un nouvel individu la classe majoritaire parmi ses k plus proches voisins

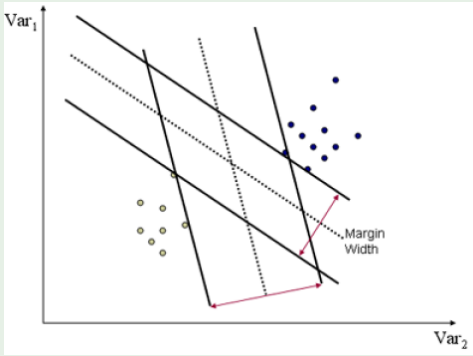
Application : votes pondérés avec seuillage (Golub et al, 1999)



- poids des k voisins fonction de sa distance au nouvel échantillon
- attribution d'une classe ssi la *force de prédiction* excède un seuil prédéfini

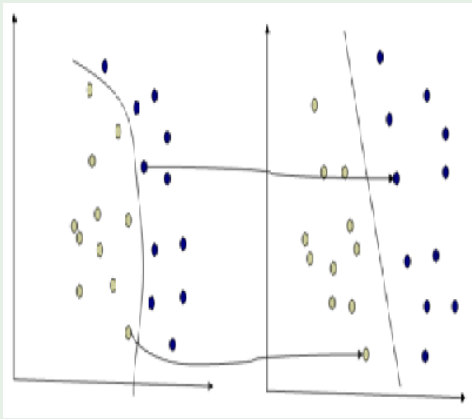
Support Vector Machine (SVM)

Principe : méthode de séparation linéaire



- idée : séparer les données par un hyperplan (comme pour LDA)
- trouver l'hyperplan qui maximise la marge de séparation des données

Extension au cas de données non linéairement séparables



- espace de représentation intermédiaire où les données sont linéairement séparables
- pas besoin de calculer les nouvelles coordonnées pour réaliser l'estimation

Classifieurs à forte marge

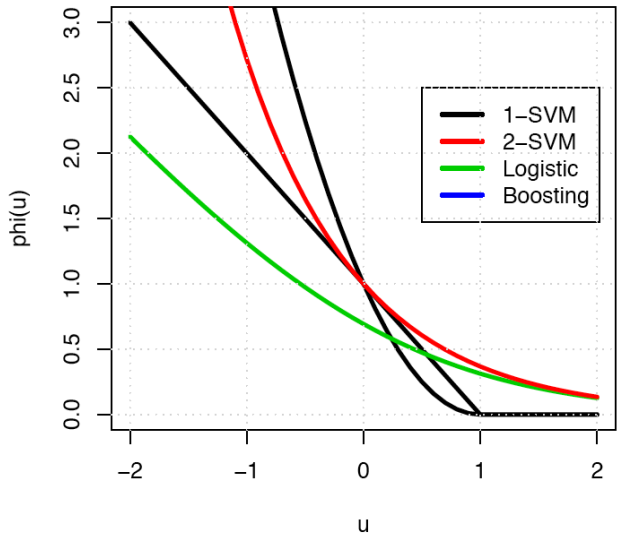
Perte quadratique : $l(t(X), y) = \mathbb{P}\left((y - t(X))^2\right)$

Marge et fonctions de perte (cas où $\mathcal{Y} = \{-1, 1\}$)

- **Marge** du classifieur t pour le couple (X, y) : $yt(X)$
- On cherche un classifieur à forte marge
- Fonctions de perte : $l(t(X), y) = \phi\left(yt(X)\right)$ où ϕ est décroissante, convexe

Choix de ϕ

Méthode	$\phi(u)$
Régression logistique	$\log\left(1 + e^{-u}\right)$
Support vector machine (1-SVM)	$\max\left(1 - u, 0\right)$
Support vector machine (2-SVM)	$\max\left(1 - u, 0\right)^2$
Boosting	e^{-u}



Classification et régression en génomique

Estimation du taux d'erreur

Taux d'erreur empirique associé au classifieur $\hat{t}$

- $EER(\hat{t}) = \frac{1}{n} \sum_{i=0}^{n} \mathbf{1}_{\hat{t}_n(X_i) \neq Y_i}$
- estimateur biaisé (optimiste) si estimé sur les données d'apprentissage
- surajustement : si $n < p$ on peut trouver un modèle suffisamment compliqué pour que $ERR(\hat{t}) = 0$

Méthodes de validation envisageables

- Validation un **échantillon test indépendant** de l'échantillon d'apprentissage
  = impossible pour de petites tailles d'échantillon
- Validation croisée, en particulier **leave one out**

Estimation du taux d'erreur

Taux d'erreur empirique associé au classifieur $\hat{t}$

- $EER(\hat{t}) = \frac{1}{n} \sum_{i=0}^{n} \mathbf{1}_{\hat{t}_n(X_i) \neq Y_i}$
- estimateur biaisé (optimiste) si estimé sur les données d'apprentissage
- surajustement : si $n < p$ on peut trouver un modèle suffisamment compliqué pour que $ERR(\hat{t}) = 0$

Méthodes de validation envisageables

- Validation un **échantillon test indépendant** de l'échantillon d'apprentissage
  = impossible pour de petites tailles d'échantillon
- Validation croisée, en particulier **leave one out**

Validation croisée

Principe

- définir une partition de l'échantillon L de départ en K éléments de même cardinal r : $(L_1, \ldots L_K)$
- pour chaque élément k de $(1, \ldots K)$:
  - construire le classifieur sur $L - L_k$
  - estimer l'erreur $\hat{e}_k$ grâce à l'échantillon L_k
- erreur estimée : moyenne des $\hat{e}_k$: $\hat{e} = \frac{1}{K} \sum_{k=1}^{K} \hat{e}_k$
- cas particulier : *leave one out* ($r = 1$, et $K = n$)

Propriétés de cet estimateur de l'erreur de classification

- quand r augmente, $Var(\hat{e})$ diminue mais le biais augmente
  = compromis biais-variance
- temps de calcul conséquent

Classification et régression en génomique

Classification et régression en génomique

Objectifs et méthodes

Pourquoi ?

- éviter le surapprentissage
- augmenter la robustesse du classifieur
- améliorer l'interprétabilité biologique du classifieur
- alléger l'étape de construction du classifieur

Approches

- Construction de nouvelles variables synthétiques
  $\Rightarrow$ méthodes de **réduction de dimension**
- Sélection de variables **avant** la classification
  $\Rightarrow$ méthodes de filtrage et d'emballage (*wrapper*)
- Sélection de variables **pendant** la classification
  $\Rightarrow$ approches par pénalisation

Construction de nouvelles variables synthétiques

Méthode supervisée ou non supervisée

Principe

- rechercher les directions les plus informatives de l'espace des variables
- ne garder que quelques composantes
- méthodes supervisées (PLS : moindres carrés partiels) ou non (PCA : Analyse en Composantes principales)

Exemples de méthodes

PCA : directions orthogonales maximisant la variabilité du nuage de points

PLS : directions maximisant la covariance entre les variables explicatives et une variable expliquée

Sélection de variables avant la classification

Recours à des heuristiques

Impossibilité de tester tous les sous-groupes de variables
$\Rightarrow$ Recours à des méthodes non exhaustives

Méthodes de filtrage

- le pouvoir prédictif de chaque gène est testé séparément
- on retient les gènes les plus discriminants
- mal adapté aux variables corrélées

Méthodes d"'emballage" (wrapper methods, p.ex. RFE)

- algorithme itératif de sélection (ascendante, descendante)
- sélection guidée par les performances du classifieur
- coûteux en temps de calcul

Classification et régression en génomique

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)

1. $\hat{\beta}_j^0$ peu biaisés mais variants
2. difficulté d'interprétation : aucun $\hat{\beta}_j^0$ n'est strictement nul

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

Ridge regression

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	–	++

$\Rightarrow$ Peut-on faire mieux à la fois en stabilité et interprétabilité ?

LASSO

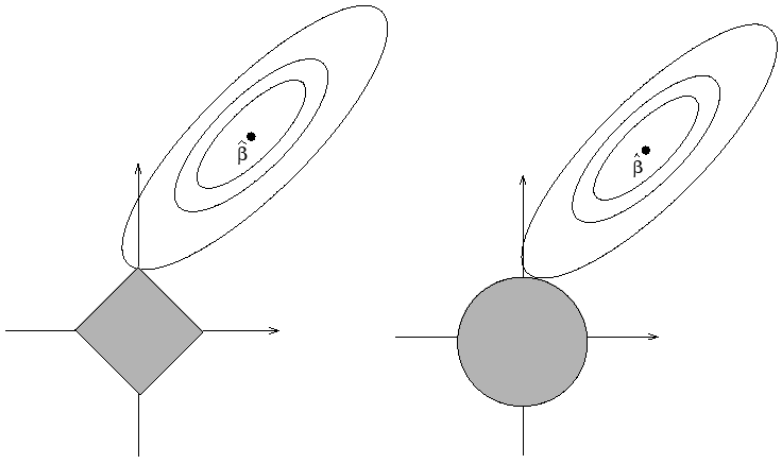
Tibshirani, JRSS (B), 1996. *Regression shrinkage and selection via the lasso*.

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



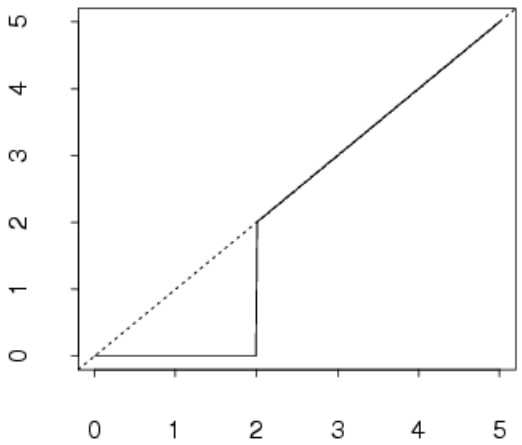
Avantages

- 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

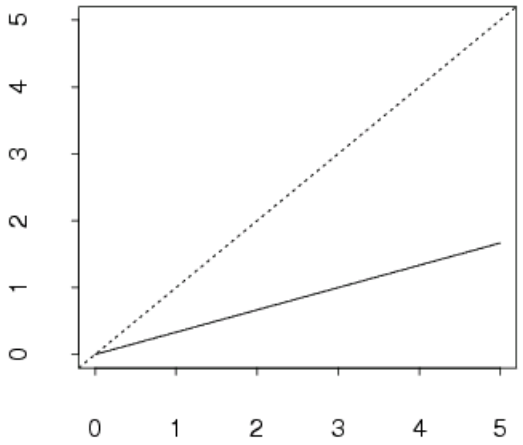
Interprétation géométrique : $\hat{\beta} = f(\beta)$ (cas orthonormal)

Subset



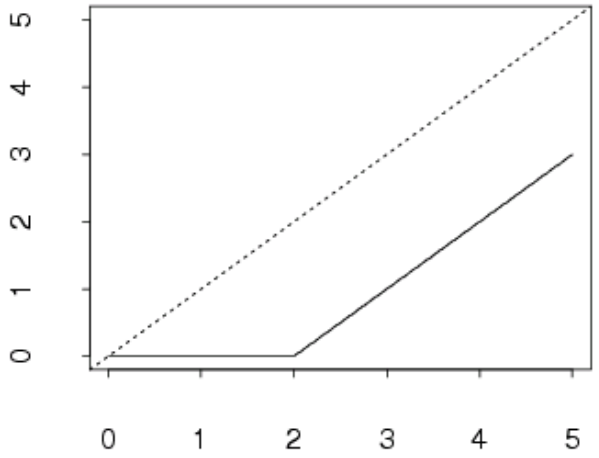
$\hat{\beta}_j^0 \mathbf{1}_{\hat{\beta}_j^0 > \gamma}$

Ridge



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

Lasso



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

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

Elastic nets

Zho & Hastie, JRSS (B), 2005. *Regularization and variable selection via the elastic net*.

Inconvénients de LASSO

- nombre de variables sélectionnées : $p^\star < n$
- LASSO sélectionne arbitrairement une variable parmi un groupe de variables très corrélées

Elastic nets : combiner ridge regression et LASSO

$$\underset{\beta}{\text{Min}} \, ||y - X\beta||_2^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2$$

- on peut obtenir strictement plus de n prédicteurs
- problème d'optimisation strictement convexe
  $\Rightarrow$ les variables corrélées sont conservées
  Estimation des paramètres par double validation croisée

Elastic nets vs LASSO : illustration

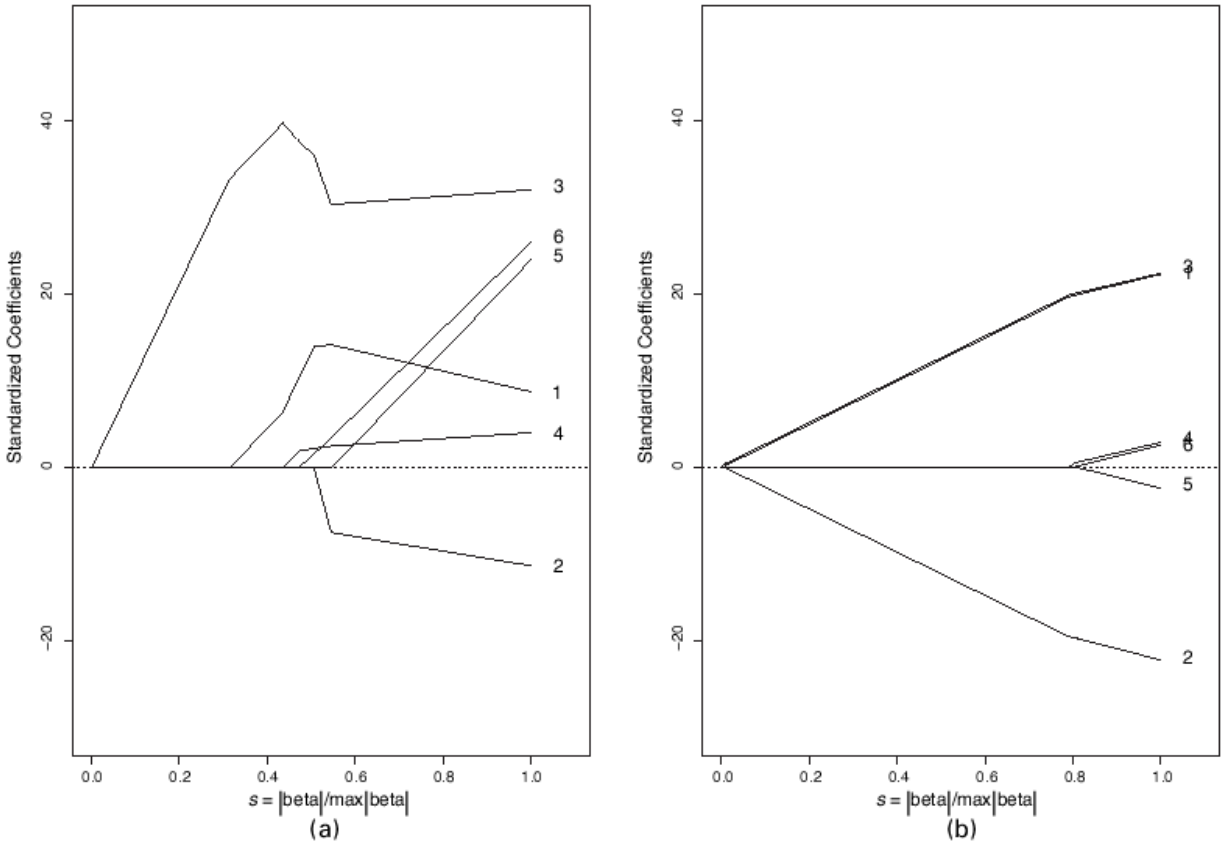


Fig. 5. (a) Lasso and (b) elastic net ($\lambda_2 = 0.5$) solution paths: the lasso paths are unstable and (a) does not reveal any correction information by itself; in contrast, the elastic net has much smoother solution paths, while clearly showing the 'grouped selection'—$\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ are in one 'significant' group and $\mathbf{x}_4$, $\mathbf{x}_5$ and $\mathbf{x}_6$ are in the other 'trivial' group; the decorrelation yields the grouping effect and stabilizes the lasso solution

Classification et régression en génomique

Référence

A.-C. Haury, P. Gestraud & J.-P. Vert. (2011)
The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures.
Technical Report, Mines ParisTech/Institut Curie/INSERM U900
http://hal.archives-ouvertes.fr/hal-00559580

Classification et régression en génomique

Signature moléculaire

Médecine personnalisée

Prise en compte des spécificités de chaque patient et de sa maladie :

- risque de rechute
- réponse au traitement

Objectif : adapter un traitement aux spécificités de chaque patient

Principe des signatures moléculaires

- Recherche de marqueurs biologiques grâce à des expériences de puces à ADN
- Essais cliniques pour valider le caractère prédictif de ces marqueurs

Motivation : très faible intersection entre signatures moléculaires inférées sur des jeux de données distincts

Exemple du test Mammaprint

Laura J. van 't Veer *et al.* (2002)
Gene expression profiling predicts clinical outcome of breast cancer.
Nature, 415(6871) :530-536.

Breast cancer patients with the same stage of disease can have markedly different **treatment responses** *and* **overall outcome**. *The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumours according to their clinical behaviour.*

Chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third ; however, **70-80% of patients receiving this treatment would have survived without** *it. None of the signatures of breast cancer gene expression reported to date allow for* **patient-tailored therapy strategies**.

Here we used **DNA microarray analysis** *on primary breast tumours of 117 young patients, and applied* **supervised classification** *to identify a gene expression signature strongly predictive of a short interval to distant metastases ("poor prognosis" signature) (...).*

Exemple du test Mammaprint

Utilisation

- utilisé aux États-Unis et dans plusieurs pays européens pour évaluer le risque de rechute dans les cancers du sein
- les résultats du test entrent en compte dans la décision du médecin de prescrire une chimiothérapie à la patiente

Méthode décrite dans l'article

1. *selection of discriminating candidate genes by their correlation with the category*
2. *determination of the optimal set of reporter genes using a leave-one-out cross validation procedure*
3. *prognostic or diagnostic prediction based on the gene expression of the optimal set of reporter genes*

Classification et régression en génomique

Signatures moléculaires

Filtrage

- test de Student
- test de Wilcoxon
- distance de Bhattacharyya
- entropie relative

"Emballage" (wrapper)

- SVM RFE
- sélection ascendante gloutonne

Pénalisation

- Lasso
- Elastic net

Stabilisation des signatures par ré-échantillonnages

Idée

- Pour $b = 1 \dots B$, stockage du vecteur $(r_j^b)_{1 \leq j \leq p}$ des rangs des variables pour une ré-échantillonnage d'indice b
- Pour chaque variable $j = 1 \dots p$, calcul d'un score : $S_j = \frac{1}{B} \sum_{b=1}^{B} f(r_j^b)$
- Les variables sélectionnées sont les s variables de plus grand score

Choix de la fonction d'agrégation f

moyenne : $f(r) = r$

sélection par stabilité : $f(r) = \mathbf{1}_{r \leq s}$

exponentielle : $f(r) = \exp(-r/s)$

Classification et régression en génomique

Performances des classifieurs obtenus

Construction de plusieurs classifieurs à partir d'une signature

1. plus proches centroïdes (NC)
2. 9-plus proches voisins (KNN)
3. SVM linéaire
4. analyse discriminant linéaire (LDA)
5. Bayes naïf (BAYES)

Évaluation

Aire sous la courbe ROC (AUC) dans deux situations :

- identification de la signature et 10-fold CV du classifieur sur le même jeu de données
- identification de la signature sur un jeu de données et 10-fold CV du classifieur sur un autre

Stabilité de la signature

Évaluation

Étant données deux signatures S_1 et S_2 de même taille s, on regarde la fraction de gènes communs :

$$\frac{|S_1 \cap S_2|}{s}$$

Création des deux signatures

perturbation douce : sous-échantillonnage avec 80% de recoupement

perturbation dure : sous-échantillonnage avec 0% de recoupement

entre jeux : deux jeux de données différents

Interprétabilité et stabilité fonctionnelle de la signature

Idée : Utilisation d'annotations biologiques

Test hypergéométrique pour évaluer la sur-représentation de termes Gene Ontology (GO, processus biologiques) dans la signature

Interprétabilité

Nombre de termes GO sur-représentés pour un seuil de FDR (BH) à 5%

Stabilité fonctionnelle

1. Génération de deux signatures (comme ci-dessus)
2. Mesure de la similarité entre les 10 termes GO les plus sur-représentés

Classification et régression en génomique

Résultats

Performances

- Sélectionner les gènes au hasard ne fonctionne pas si mal...
- Peu de différences entre méthodes de classification
- Les plus proches centroïdes sont assez bons partout
- La stabilisation des signatures n'apporte pas grand chose

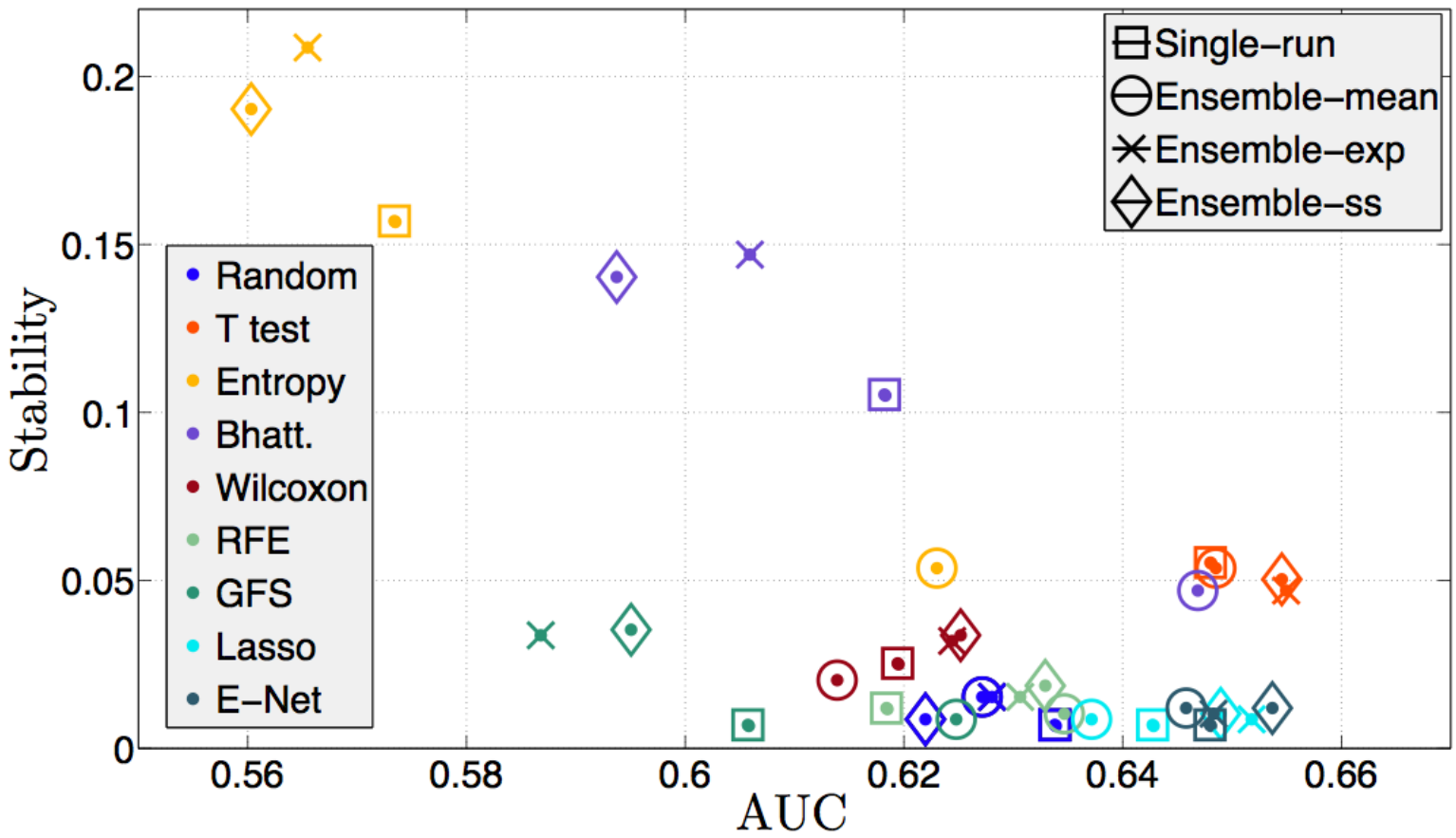
Stabilité

- Les méthodes de filtrages sont supérieures
- Préférer sélection par stabilité ou sélection exponentielle

Interprétabilité

- Les méthodes de filtrages sont très supérieures
- Peu d'apport de la stabilisation des signatures

Arbitrage entre performance et stabilité



"The t-test seems to give the best stability among methods with good accuracy"