

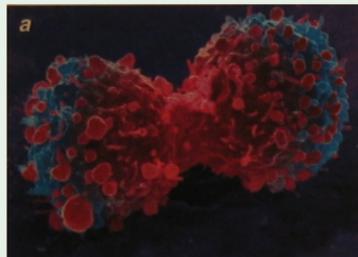
Problématiques statistiques en cancérologie

- 1 Problématiques statistiques en cancérologie
 - Cancers et altérations du génome
 - Données génomiques et problèmes statistiques

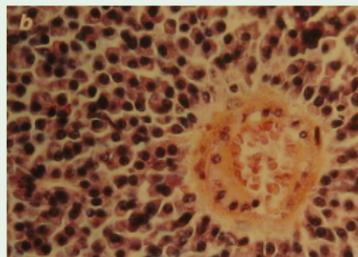
Dysfonctionnement des mécanismes de régulation

Caractéristiques des cellules tumorales

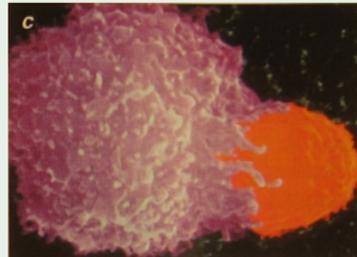
Hanahan & Weinberg, 2000



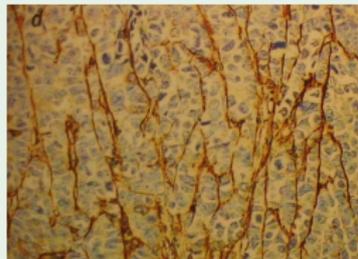
auto-suffisance en facteurs de croissance



insensibilité aux signaux anti-croissance



pas d'apoptose (= mort programmée)



angiogénèse



pouvoir de réplication illimité



invasion des tissus, métastases

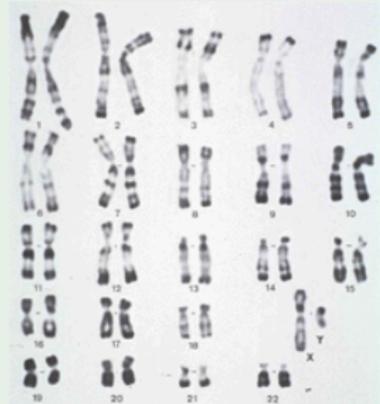
Acquises via la forte **instabilité génétique** des cellules tumorales

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

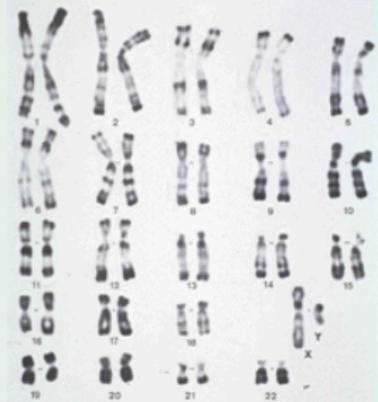


Le cancer, une maladie des gènes

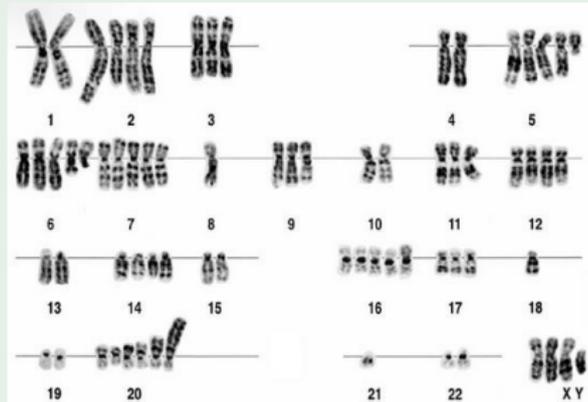
Tous les cancers s'accompagnent de désordres génétiques

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

caryotype normal



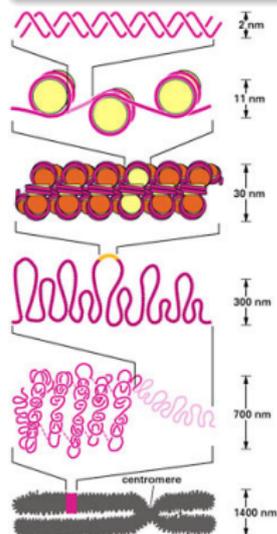
caryotype d'une tumeur



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 20 μ m)



Différents degrés de compaction

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

hélice de nucléosomes ($\delta = 30\text{nm}$)

boucles ancrées sur un échafaudage protéique

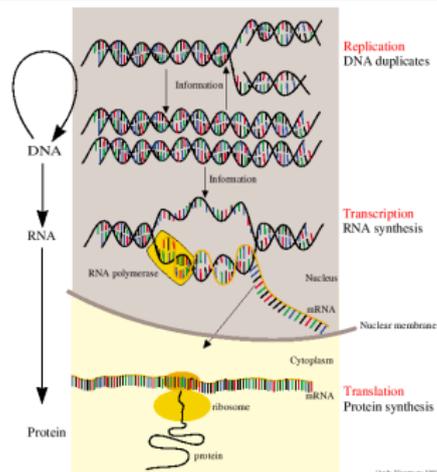
bras de chromosome : super-hélice

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

génomique : 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

- 1 Problématiques statistiques en cancérologie
 - Cancers et altérations du génome
 - Données génomiques et problèmes statistiques

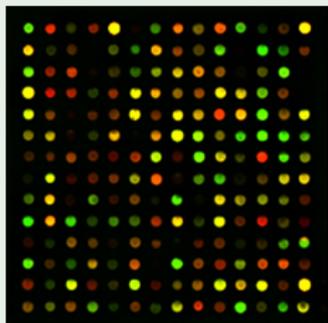
Technologie : puces à ADN

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

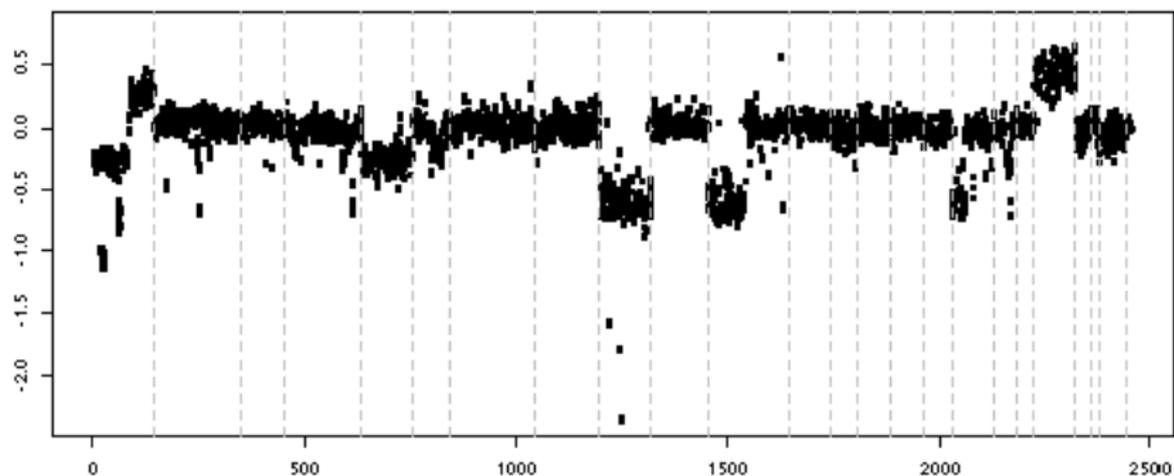


Caractéristiques de ces techniques

- 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

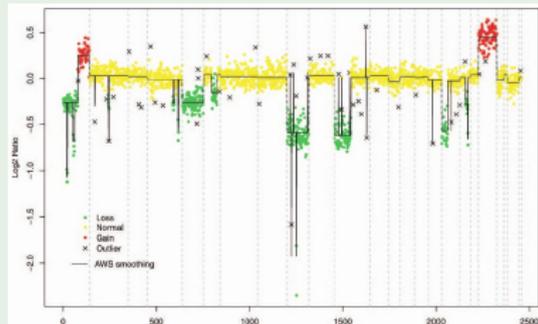
Cartographie des variations du nombre de copies d'ADN

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 :

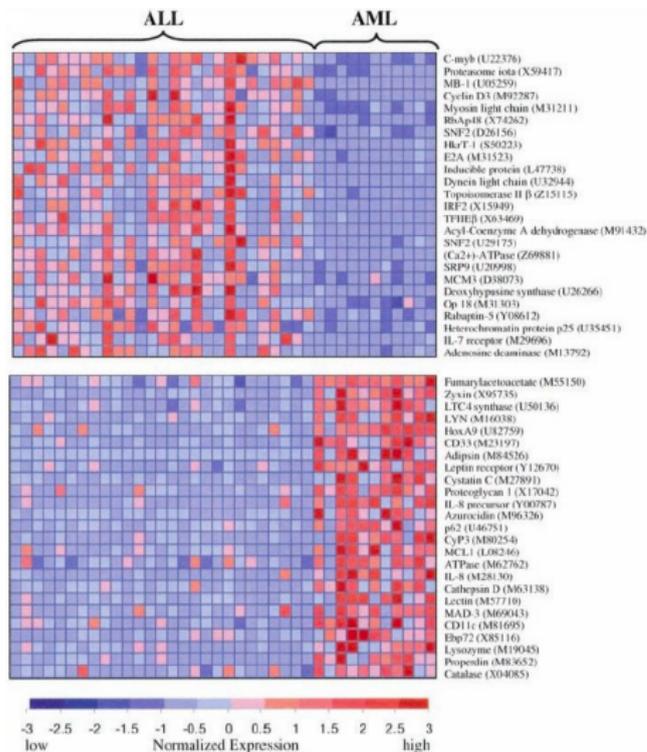
- 1 détection des ruptures
- 2 attribution d'un statut aux régions obtenues : gain, normal, perte

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

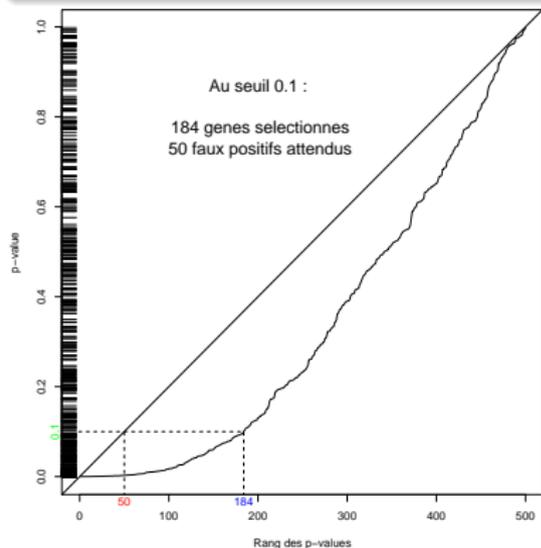
- ① extraction du signal biologique
- ② **tests d'expression différentielle**
- ③ classification non supervisée
- ④ classification supervisée
- ⑤ prédiction de la survie
- ⑥ inférence de réseaux de régulation

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

Voir cours de JP Vert au second semestre

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

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

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

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

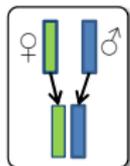
4 Segmentation de données de nombre de copies d'ADN (total)

- Recours aux modèles de rupture
- Exemples d'approches proposées

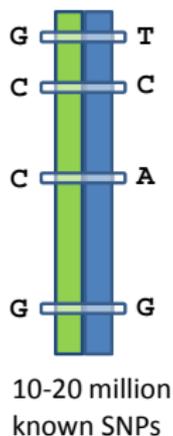
5 Estimation du nombre de copies d'ADN

- Détection : utilisation conjointe de C et DH
- Étiquetage : cellules normales et ploïdie

Génotypes dans un chromosome diploïde

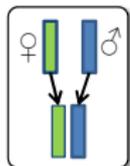


Single nucleotide polymorphism

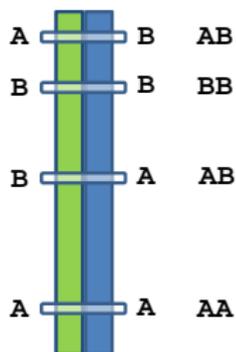


slide: H. Bengtsson.

Génotypes dans un chromosome diploïde



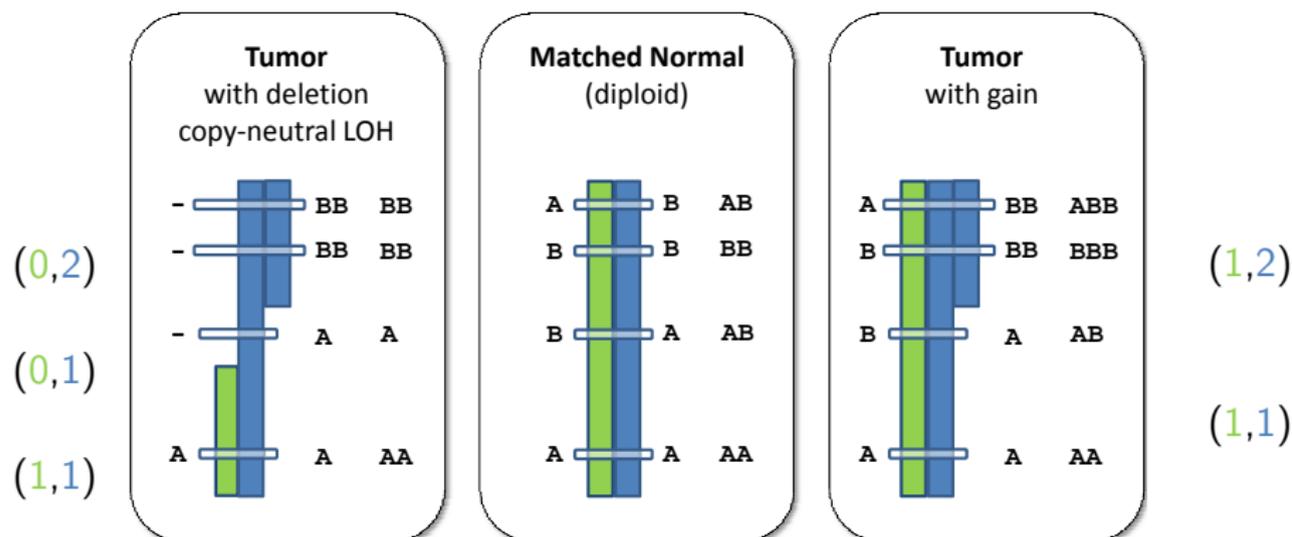
Single nucleotide polymorphism



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

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

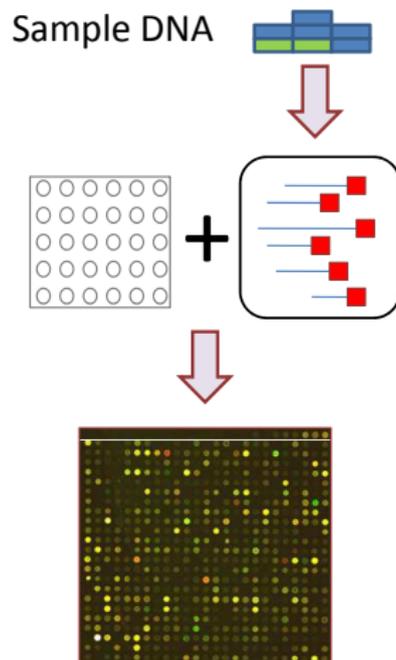
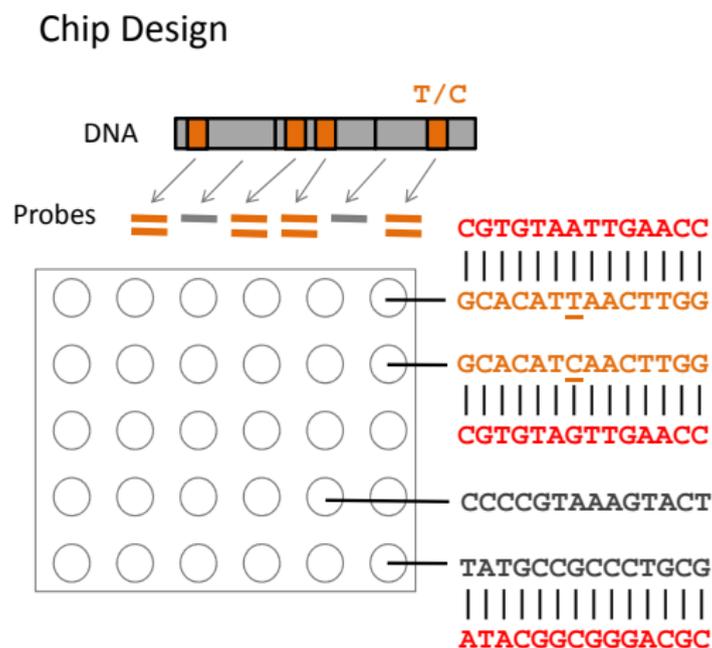
4 Segmentation de données de nombre de copies d'ADN (total)

- Recours aux modèles de rupture
- Exemples d'approches proposées

5 Estimation du nombre de copies d'ADN

- Détection : utilisation conjointe de C et DH
- Étiquetage : cellules normales et ploïdie

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

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

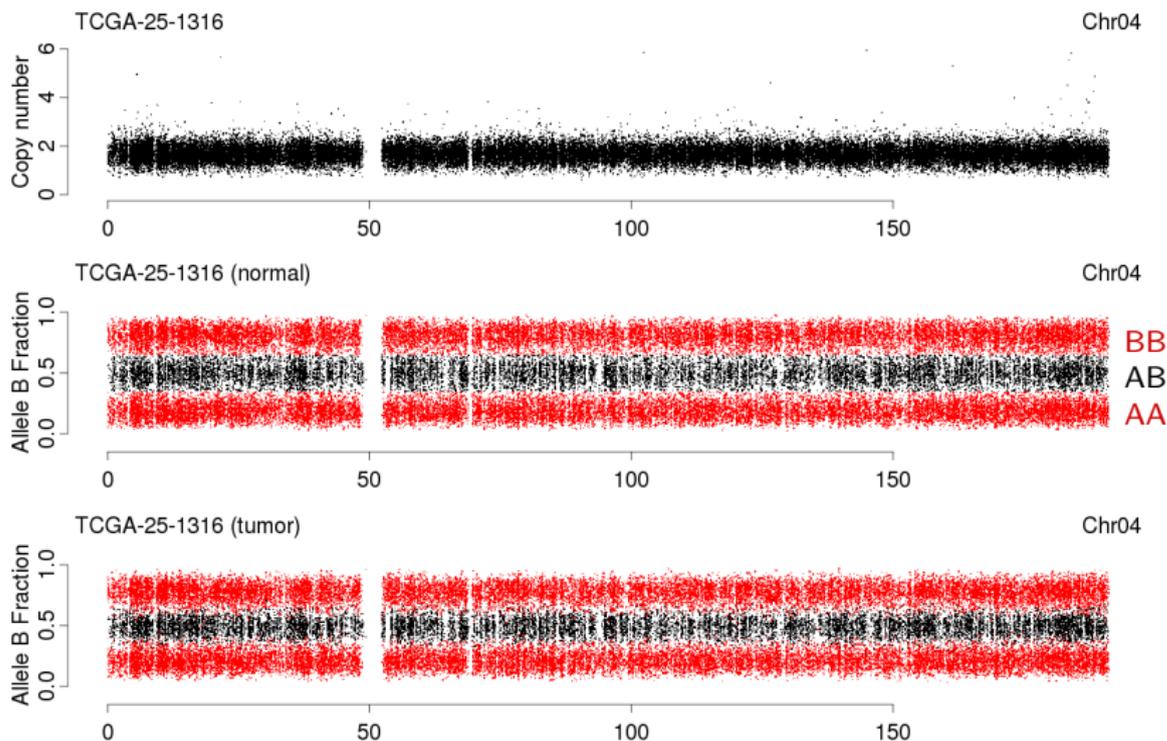
4 Segmentation de données de nombre de copies d'ADN (total)

- Recours aux modèles de rupture
- Exemples d'approches proposées

5 Estimation du nombre de copies d'ADN

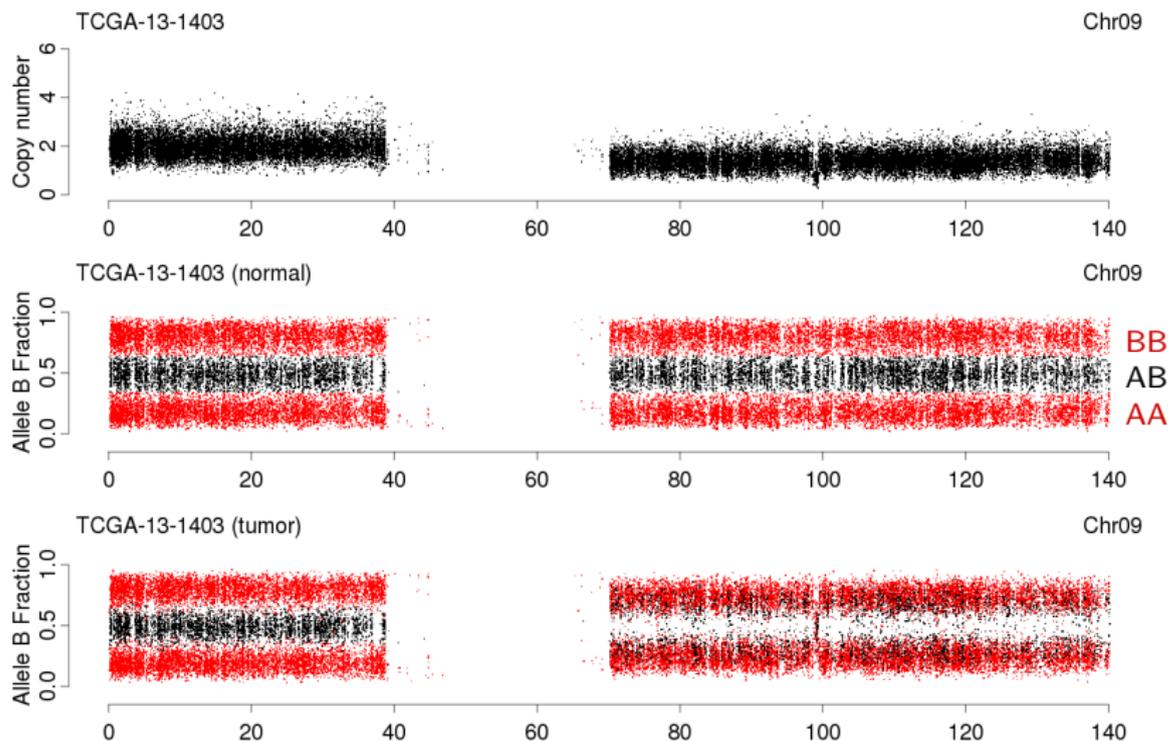
- Détection : utilisation conjointe de C et DH
- Étiquetage : cellules normales et ploïdie

Région normale : (1,1)



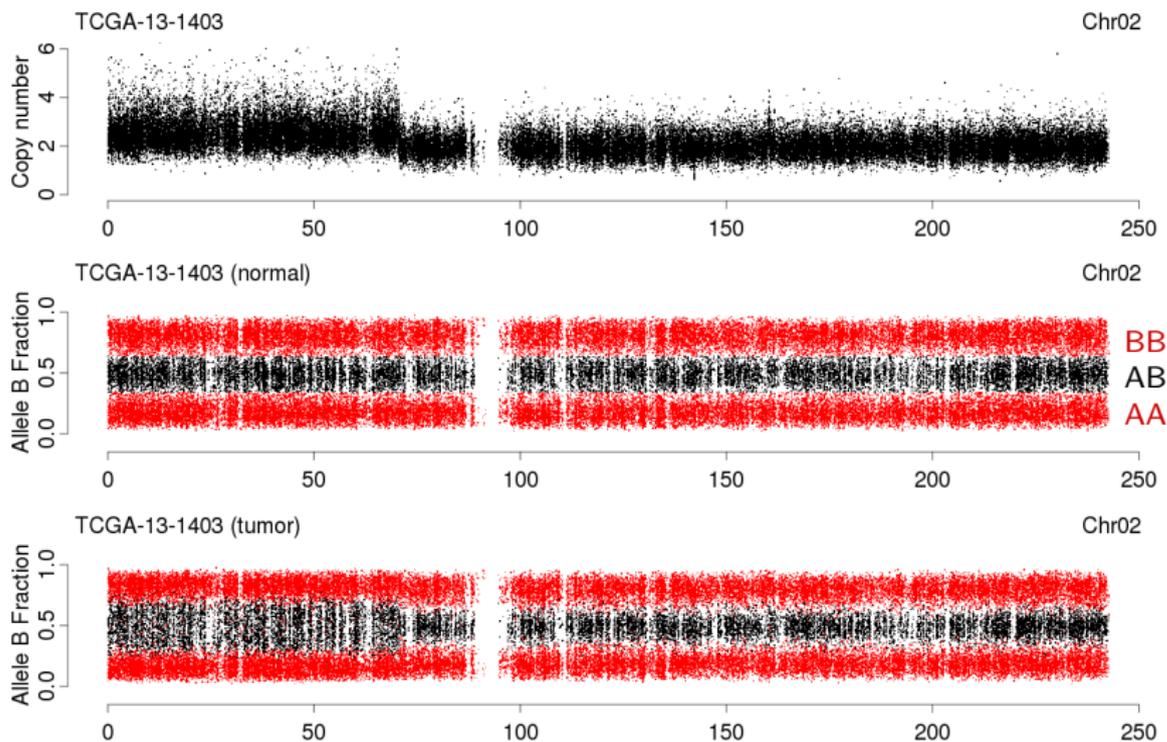
Les SNPs homozygotes dans l'échantillon normal sont en rouge.

Perte d'une copie : (0, 1)



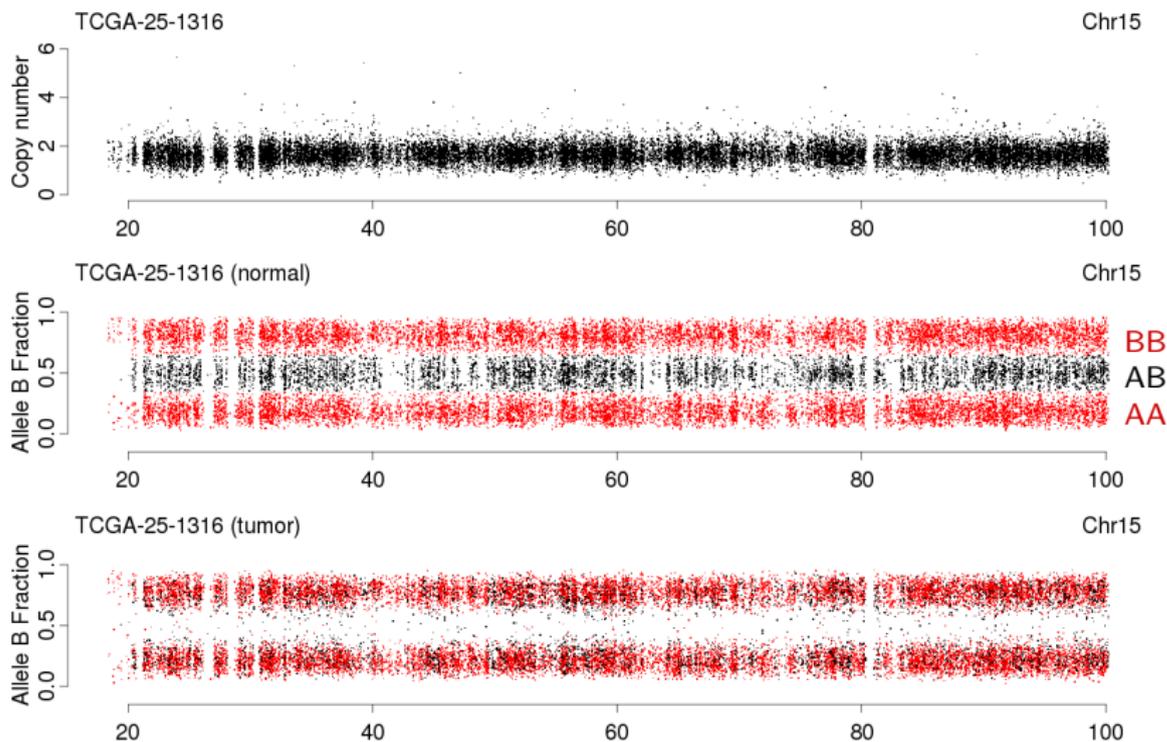
Les SNPs homozygotes dans l'échantillon normal sont en rouge.

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.

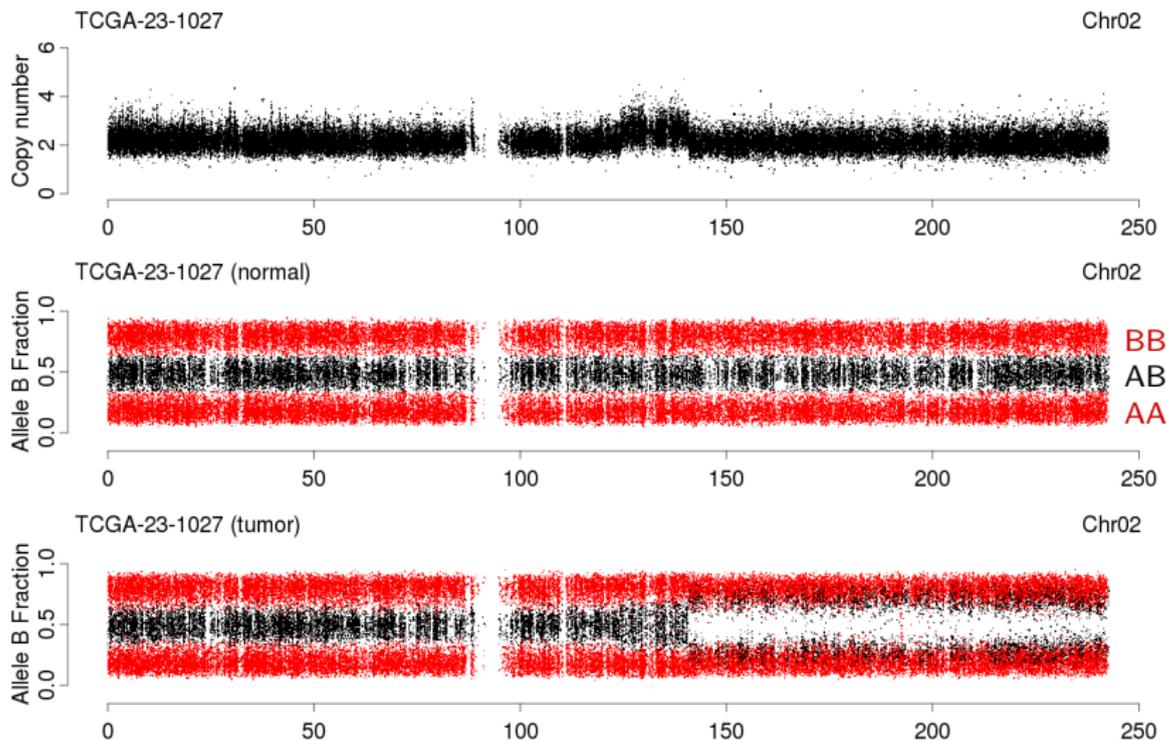
Présence de cellules normales

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.

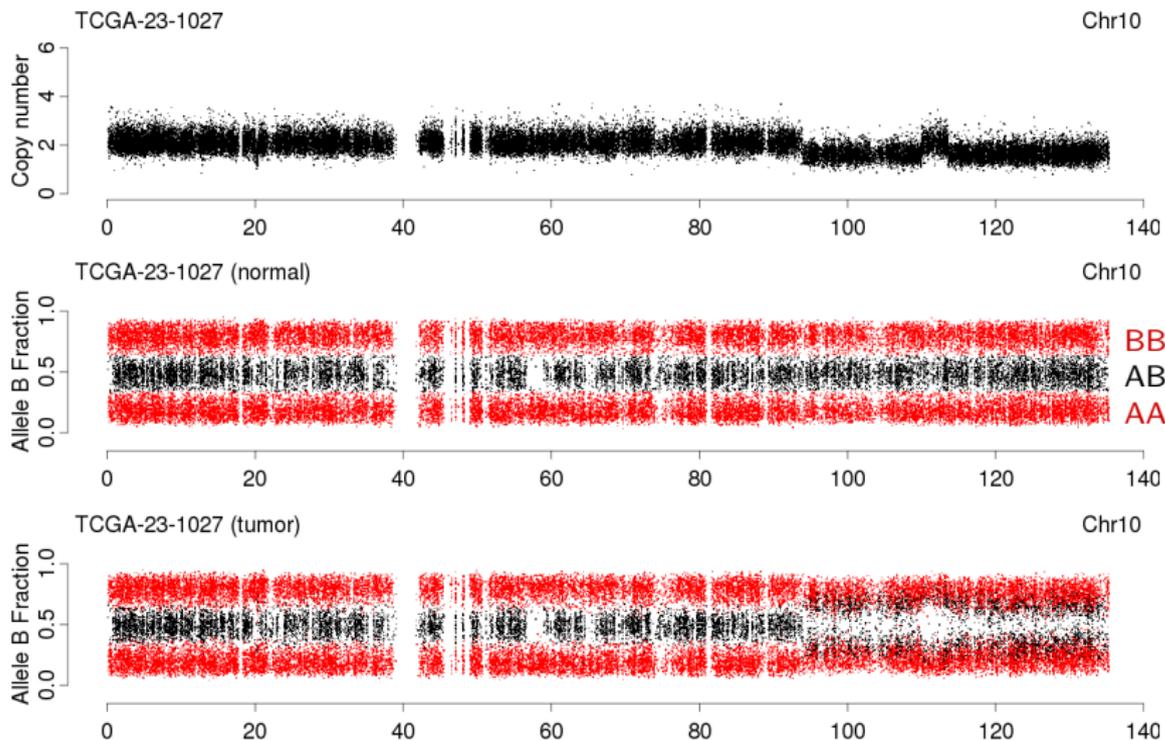
En présence de cellules normales, les ratios alléliques des SNPs hétérozygotes **se rapprochent de $1/2$** .

Normal, gain, isodisomie



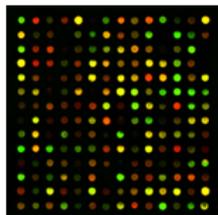
Les SNPs homozygotes dans l'échantillon normal sont en rouge.

Normal, perte, isodisomie

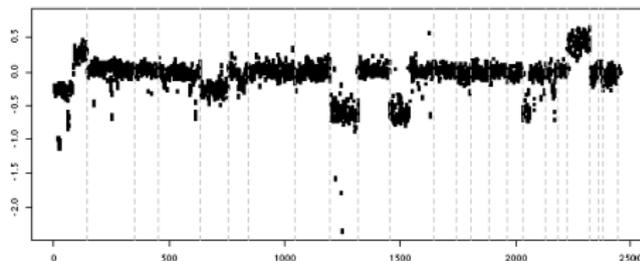


Les SNPs homozygotes dans l'échantillon normal sont en rouge.

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

- 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
- 5 Estimation du nombre de copies d'ADN
 - Détection : utilisation conjointe de C et DH
 - Étiquetage : cellules normales et ploïdie

Copy-numbers by Robust Microarray Analysis (CRMA)

Une méthode de pre-processing applicable échantillon par échantillon

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

<i>Calibrating & normalizing for hybridization artifacts</i>	<ol style="list-style-type: none"> 1. Offset and Allelic crosstalk calibration 2. Probe-sequence normalization
<i>Summarization of technical replicates</i>	<ol style="list-style-type: none"> 1. CN loci have one probe  2. Robust averaging of replicated SNPs probes 
<i>Normalizing for assay artifacts</i>	<ol style="list-style-type: none"> 1. PCR fragment-length normalization 2. GC-content normalization
<i>Total and Allele-specific copy numbers</i>	$(C_A, C_B), C = C_A + C_B$

slide: H. Bengtsson.

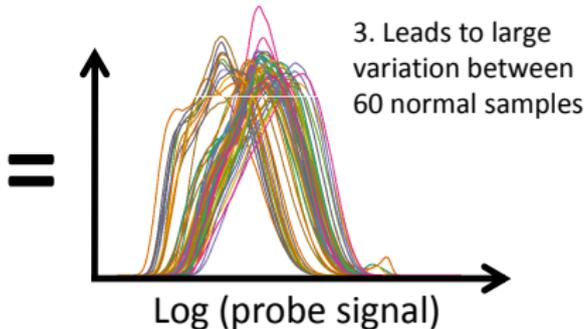
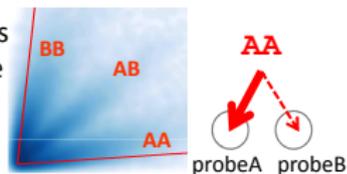
Explication de la variation systématique entre puces

Décalage (offset) du scanner et hybridation croisée entre allèles

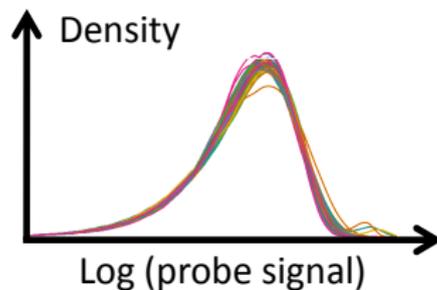
1. The scanner's shifts all probe signals (offset)



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



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



slide: H. Bengtsson.

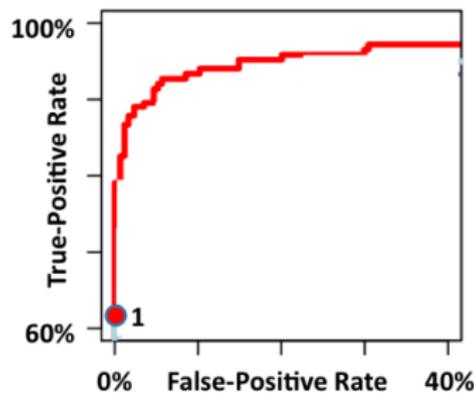
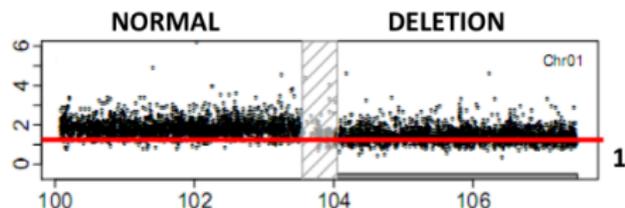
Évaluation ROC

Pour un échantillon donné

- 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



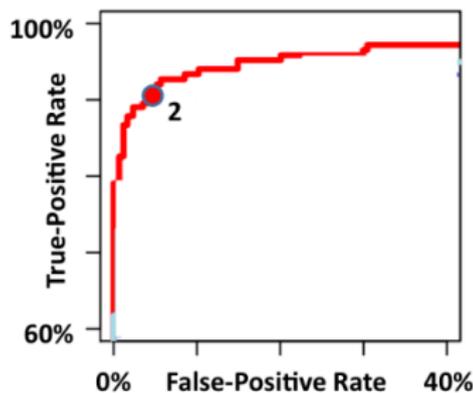
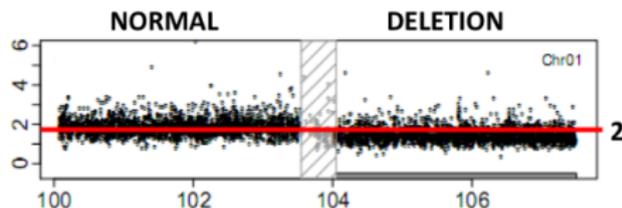
Évaluation ROC

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



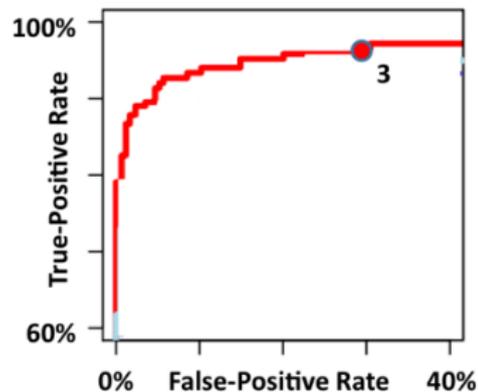
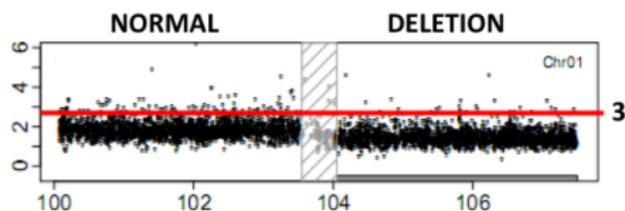
Évaluation ROC

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)

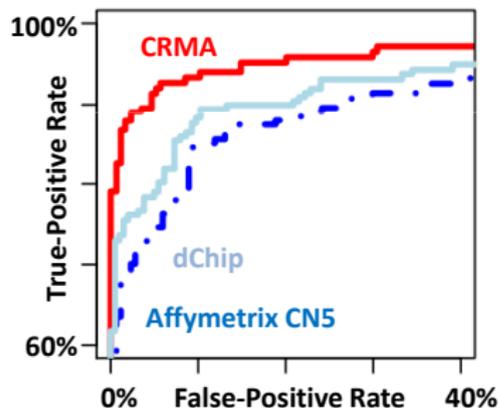
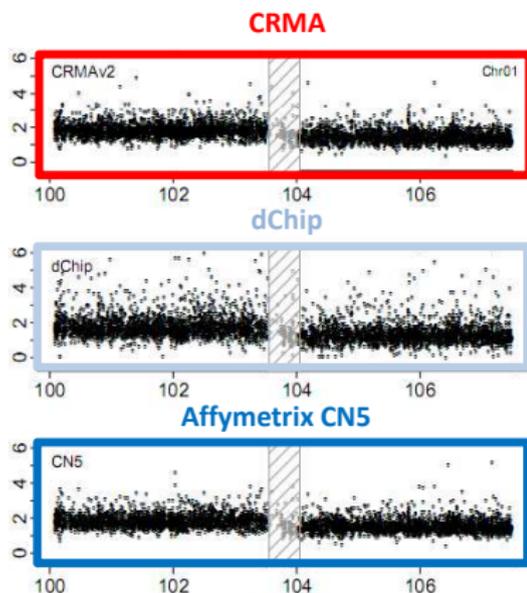
Pour chaque valeur du seuil τ :

- 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.

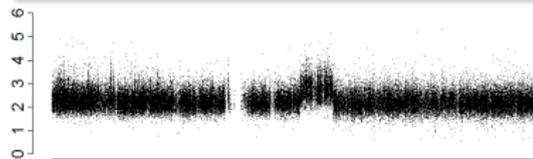
- 2 Puces SNP en cancérologie
 - Changements de nombre de copies d'ADN dans les cancers
 - Données de puces SNP
- 3 Extraction de l'information biologique
 - Pre-processing : des signaux comparables entre échantillons
 - Post-processing : nombre de copies totaux
 - Post-processing : ratios alléliques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

Motivation : rapport signal/bruit le long du génome

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

Nombre de copies total

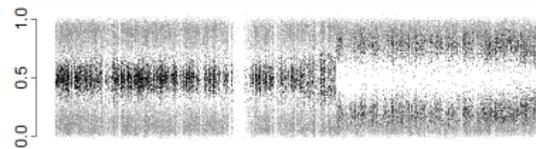
$$\begin{aligned} C_{ij} &= 2 \frac{\theta_{ij}}{\theta_{Rj}} \\ &= C_{1ij} + C_{2ij} \end{aligned}$$



Choix de la référence R ?

Diminution d'hétérozygotie

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



Rapport signal sur bruit faible

Analyse de données de nombre de copies d'ADN

- 2 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
- 5 Estimation du nombre de copies d'ADN
 - Détection : utilisation conjointe de C et DH
 - Étiquetage : cellules normales et ploïdie

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

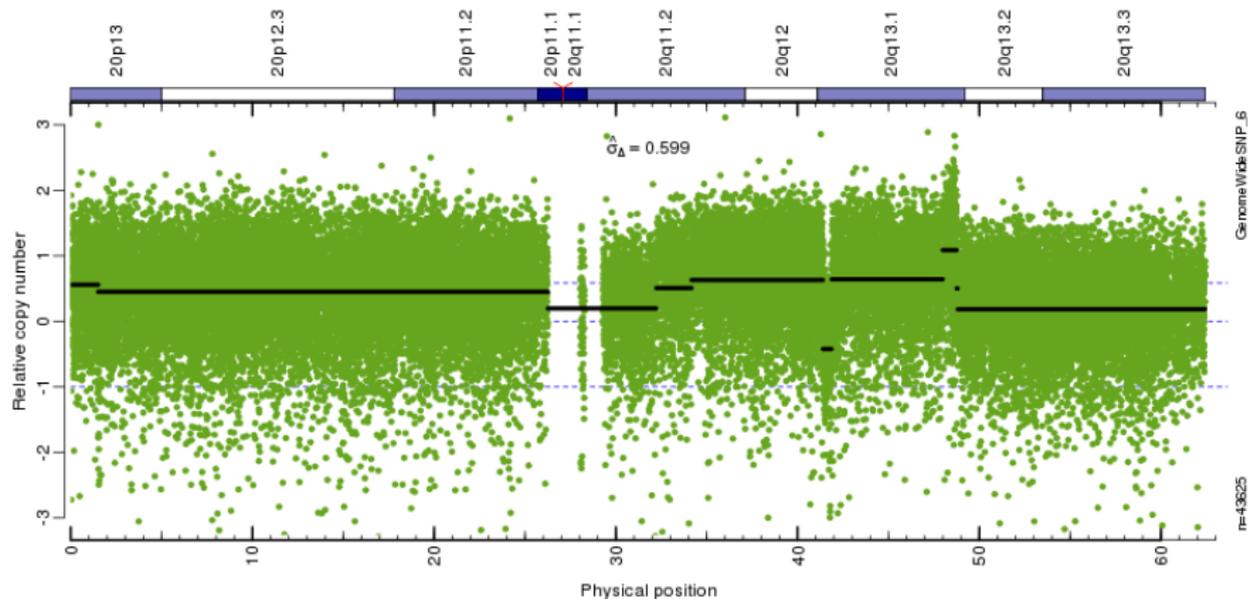
Quantification du niveau de bruit :

$$\hat{\sigma}_{\Delta} \propto \operatorname{median}_j \left(\left| z_j - \operatorname{median}_{j'}(z_{j'}) \right| \right)$$

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

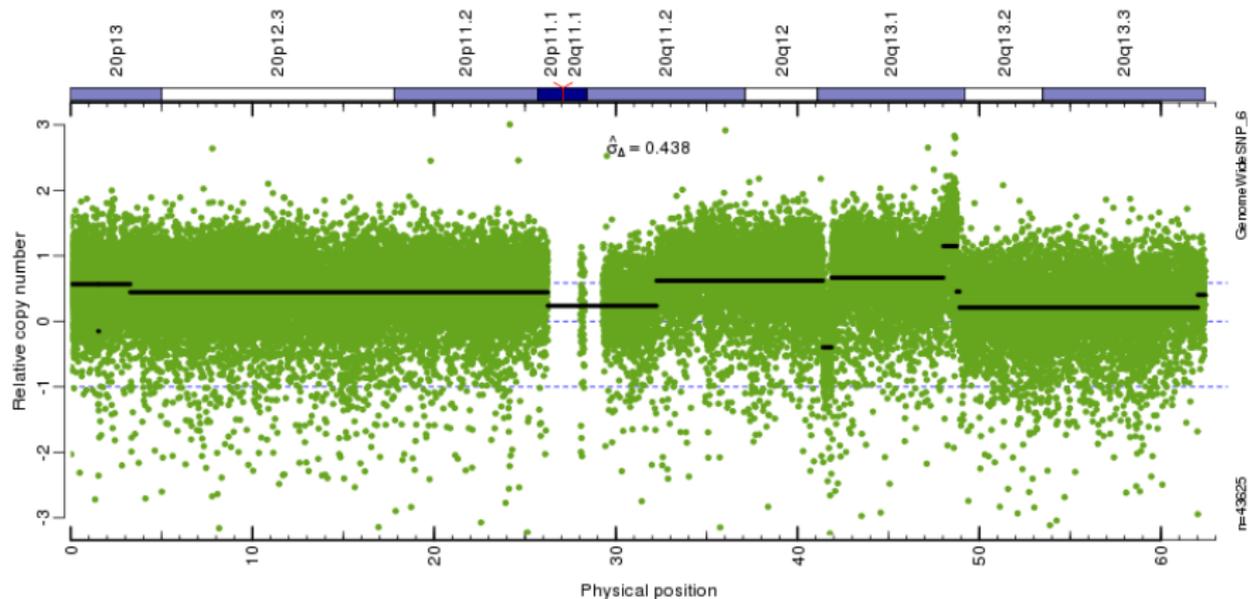
Choix d'une référence

Laboratoire différent (n=192)



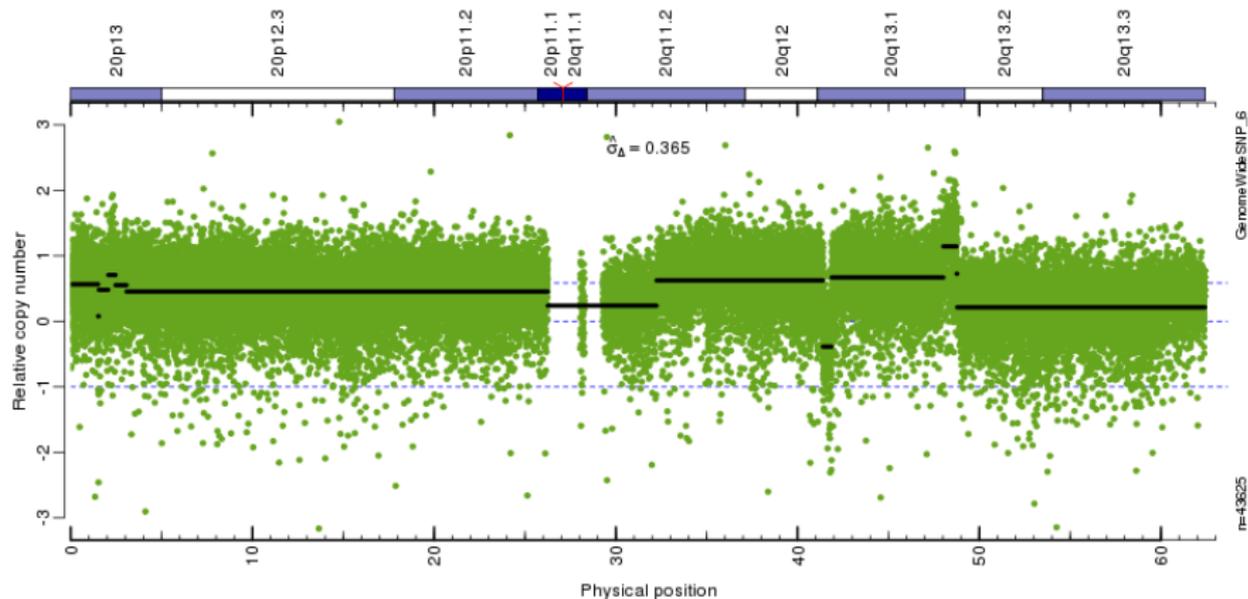
Choix d'une référence

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



Choix d'une référence

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



Analyse de données de nombre de copies d'ADN

2 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)

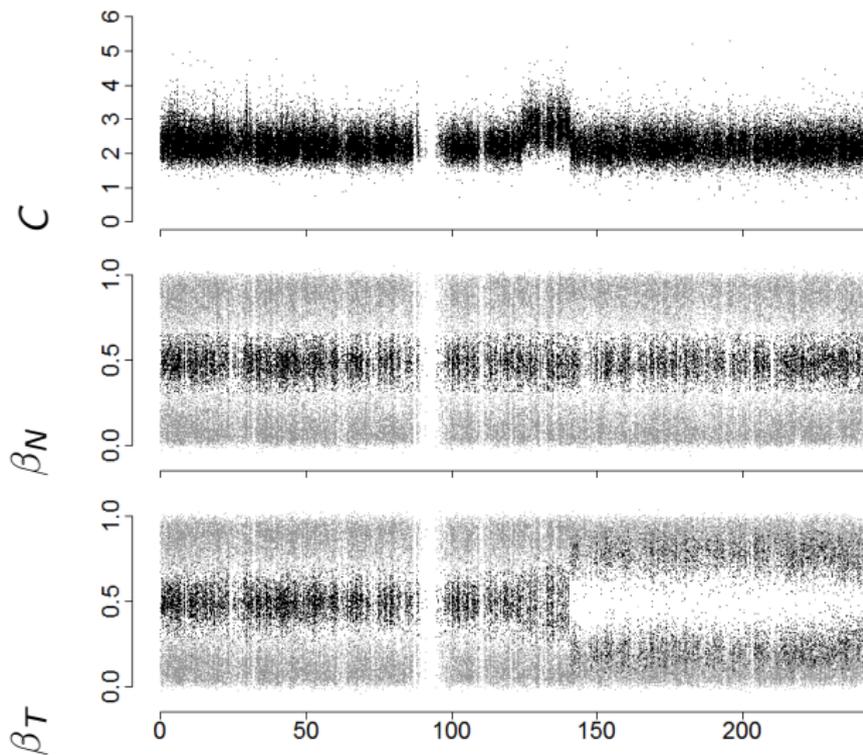
- Recours aux modèles de rupture
- Exemples d'approches proposées

5 Estimation du nombre de copies d'ADN

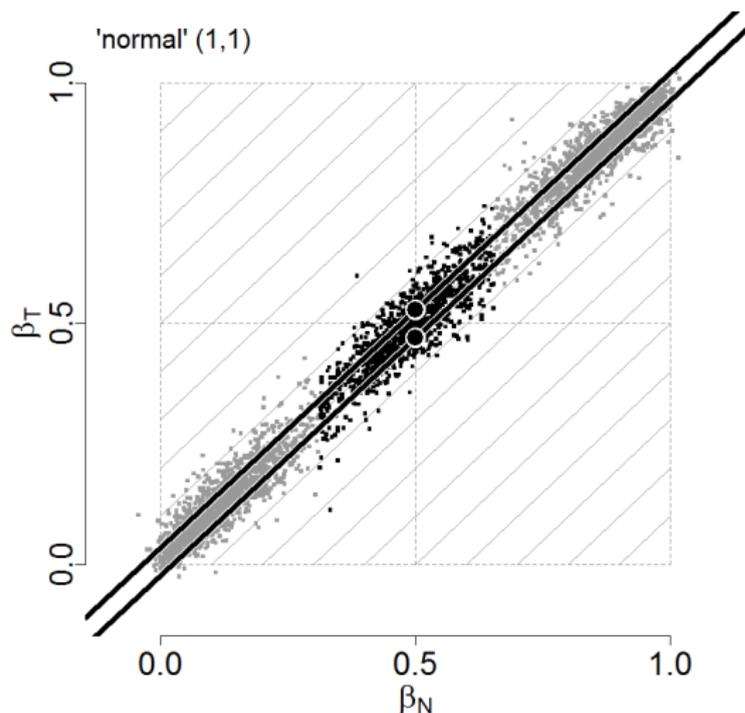
- Détection : utilisation conjointe de C et DH
- Étiquetage : cellules normales et ploïdie

Les ratios alléliques sont bruités

Exemple après pre-processing suivant la méthode CRMAv2



Effet SNP dans une région "normale" de la tumeur

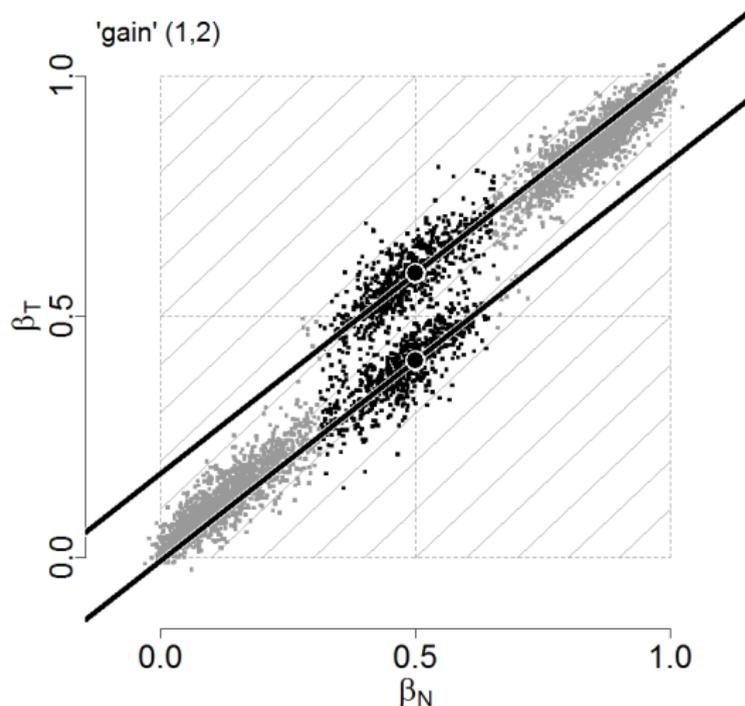


- Au lieu de trois points et $(0,0)$, $(\frac{1}{2}, \frac{1}{2})$ et $(1,1)$, on observe trois groupes; la déviation est appelée **effet SNP** :

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

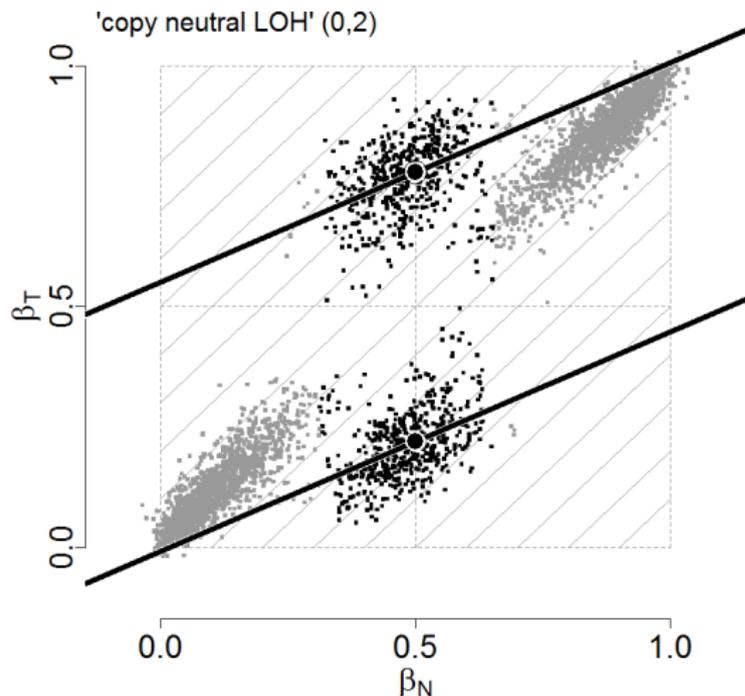
- δ est très similaire dans la tumeur et le normal

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



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

Effet SNP dans une région d'isodisomie



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

TumorBoost : normalisation d'une paire tumeur/normal

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

Idée

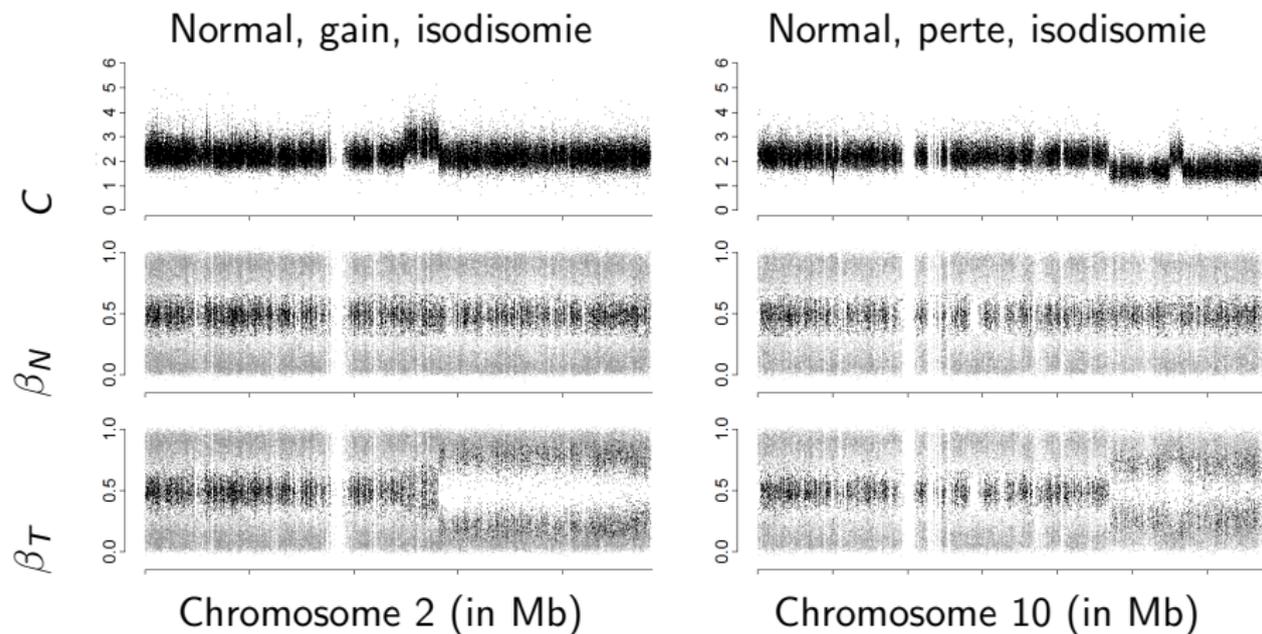
- 1 l'effet SNP est similaire dans le normal et dans la tumeur
- 2 le normal est relativement simple à analyser (seulement trois génotypes attendus)

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

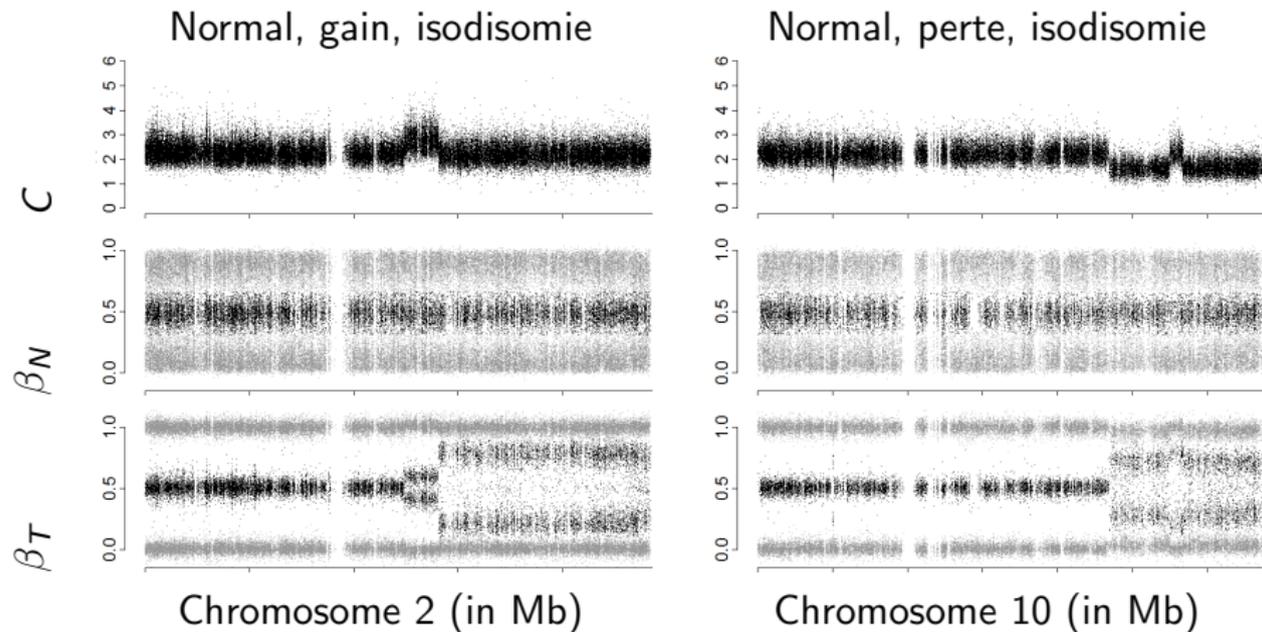
Caractéristiques de la méthode

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

Signaux génomiques avant normalisation

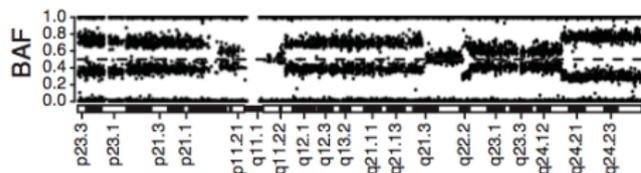


Signaux génomiques après normalisation

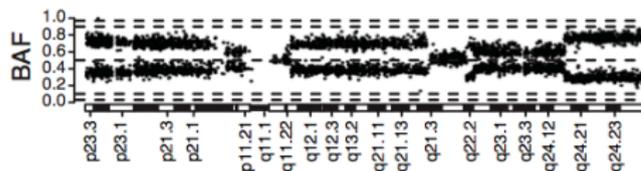


Détection de ruptures à partir des ratios alléliques

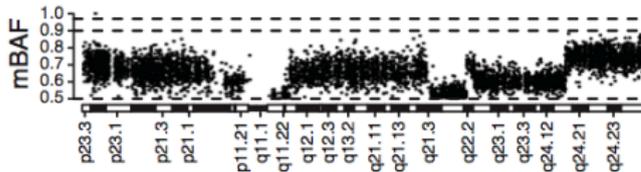
D'après Staaf *et al*, Genome Biology, 2008



ratios alléliques : β



ratios alléliques pour les SNPs hétérozygotes



Diminution d'hétérozygotie pour les SNPs hétérozygotes :

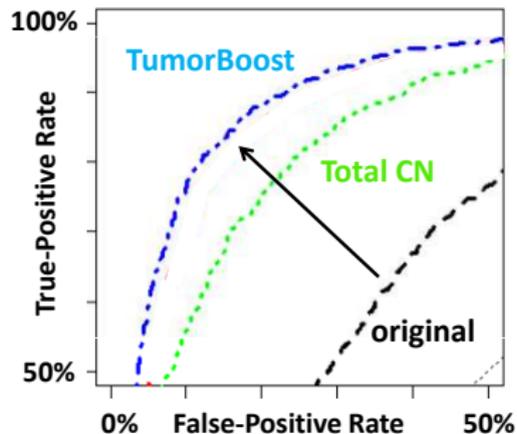
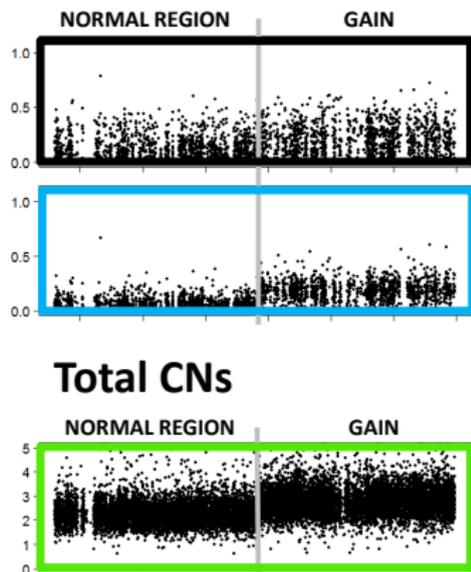
$$\rho = |\beta - 1/2| = DH/2$$

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

Allelic imbalance



slide: H. Bengtsson.

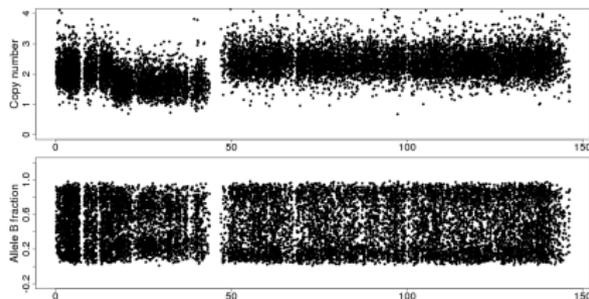
Cas où on ne dispose pas d'un échantillon normal apparié : CalMaTe

Ortiz-Estevéz *et al*

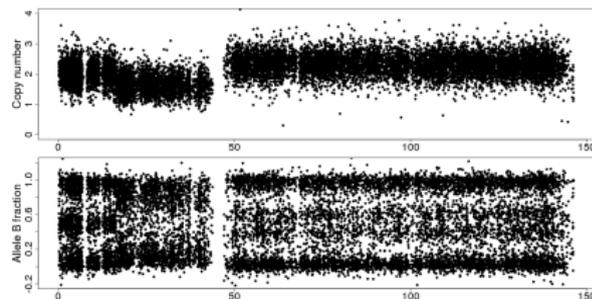
Pour chaque SNP :

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

Avant CalMaTe



Après CalMaTe



Analyse de données de nombre de copies d'ADN

2 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

5 Estimation du nombre de copies d'ADN

- Détection : utilisation conjointe de C et DH
- Étiquetage : cellules normales et ploïdie

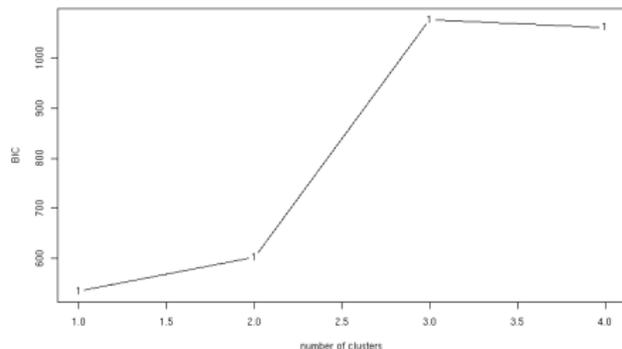
Limites des approches directes

Modèles de mélange

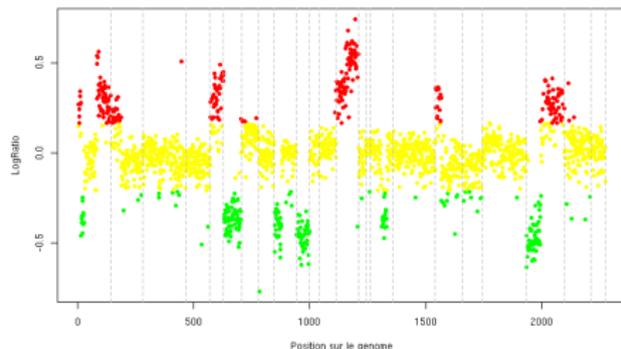
Méthode

- 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

Evolution du BIC en fonction du nombre de clusters



Alterations identifiées par modèle de mélange



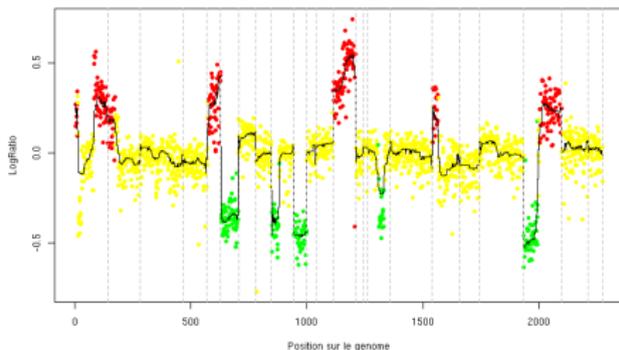
Limites des approches directes

Méthodes de lissage

Méthode

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

Alterations identifiées par médiane mobile



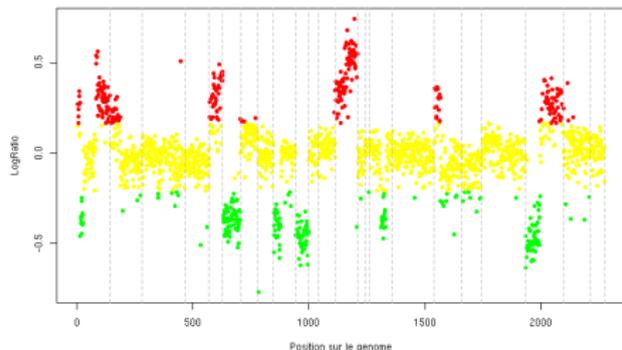
Paramètres

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

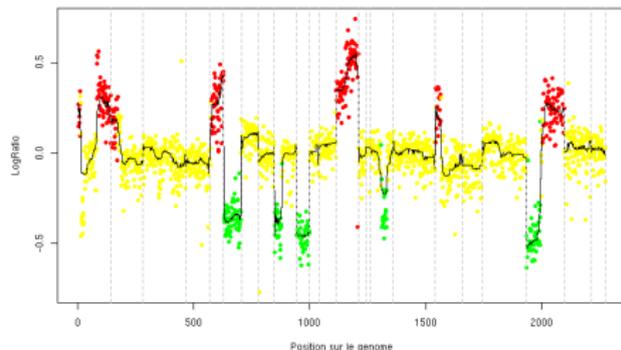
Limites des approches directes

Nécessité de méthodes plus fines

Alterations identifiées par modèle de mélange



Alterations identifiées par médiane mobile



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

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

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

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

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 $\bar{\ell}(K, \cdot) = \ell(K, \cdot) - \beta \text{pen}(K)$, avec $\text{pen}(K)$ croissante en K
- choix de $\text{pen}(K)$ 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 homoscedastique : $\sigma, (\Gamma_k)_{k=1\dots K}$

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 $J \gg K$ et propose une méthode d'estimation adaptative de β

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

5 Estimation du nombre de copies d'ADN

- Détection : utilisation conjointe de C et DH
- Étiquetage : cellules normales et ploïdie

Grands types d'approches

Segmentation proprement dite

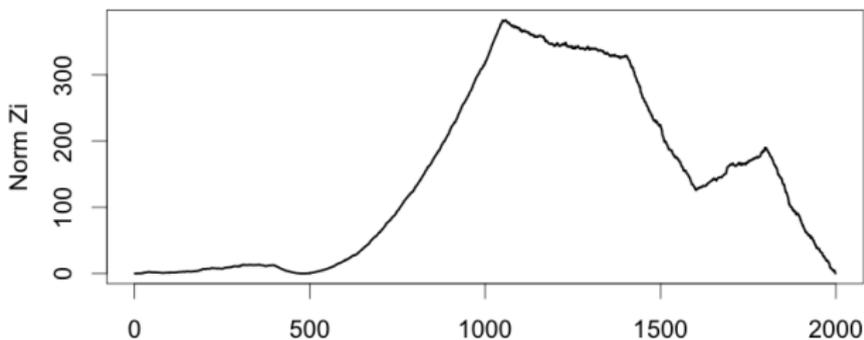
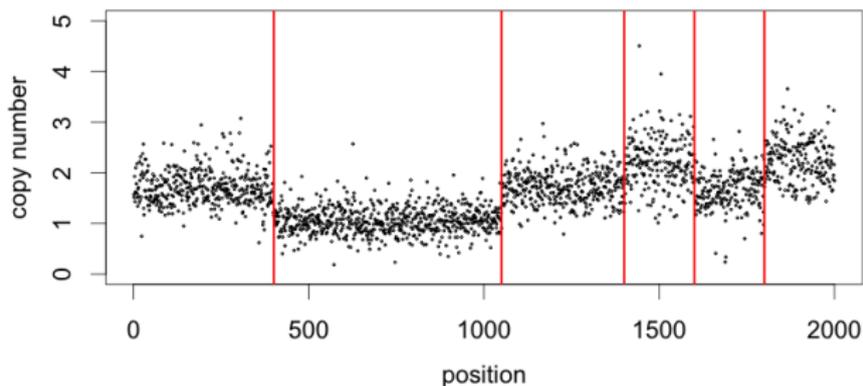
- Segmentation binaire réursive (CART)
- Segmentation binaire réursive circulaire (CBS)
- Programmation dynamique (DP)
- Fused Lasso

Segmentation et étiquetage conjoints

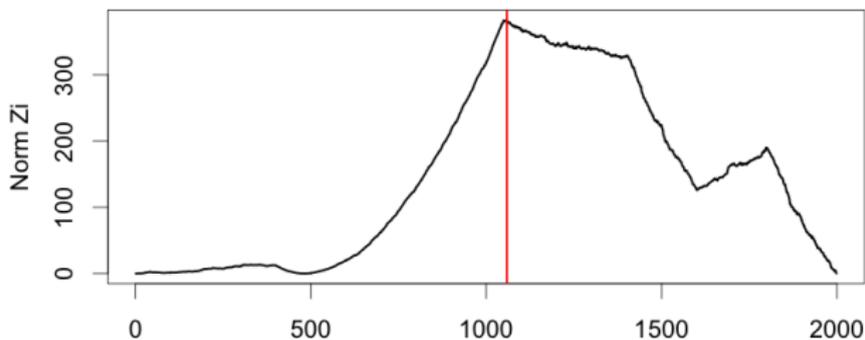
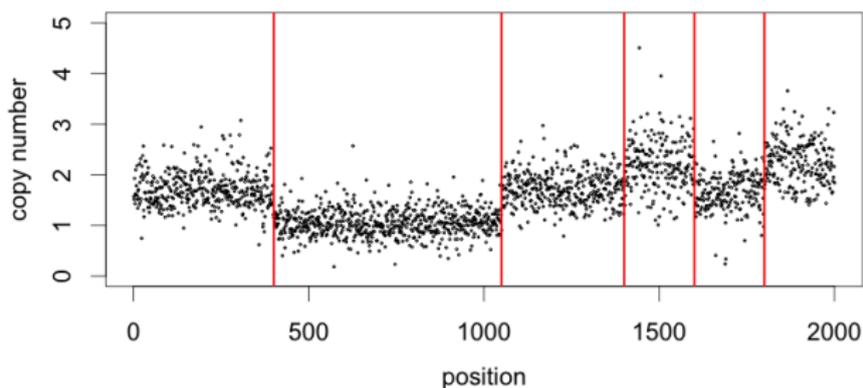
- Modèles de Markov Cachés (HMM)

Compromis entre performances statistiques et efficacité algorithmique ?

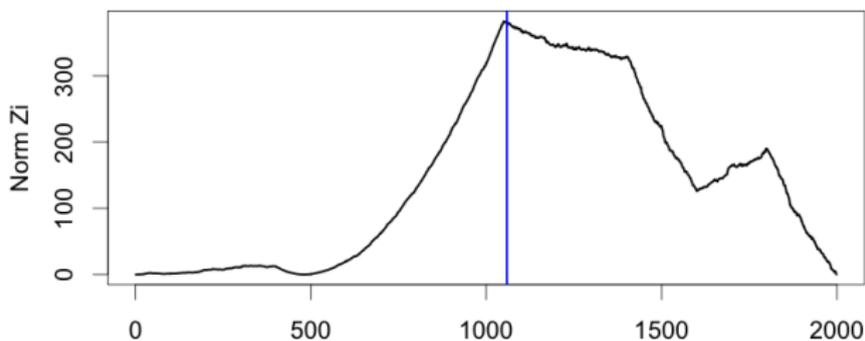
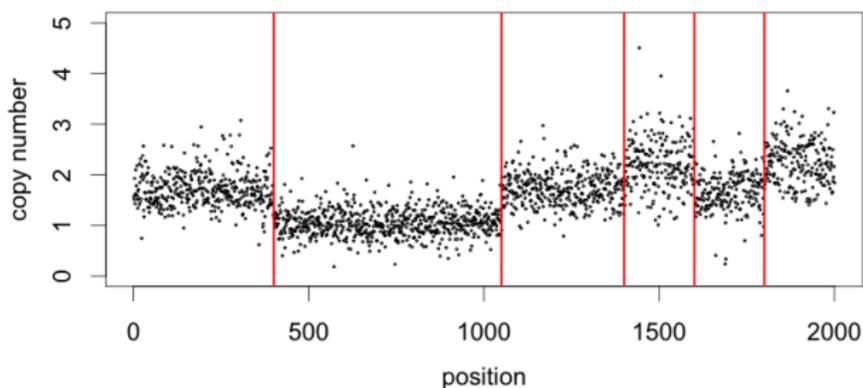
Heuristiques : segmentation binaire récursive



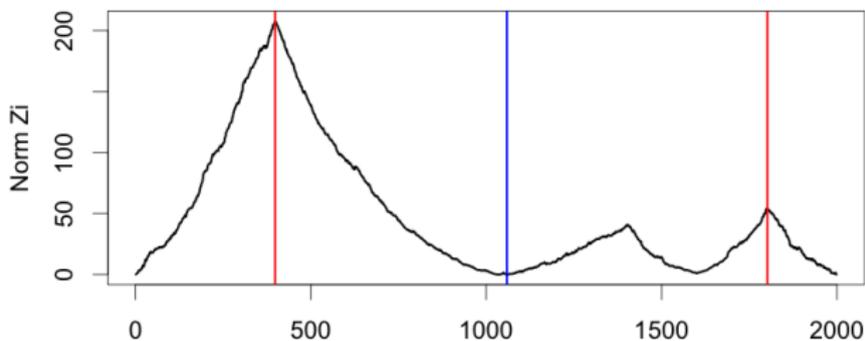
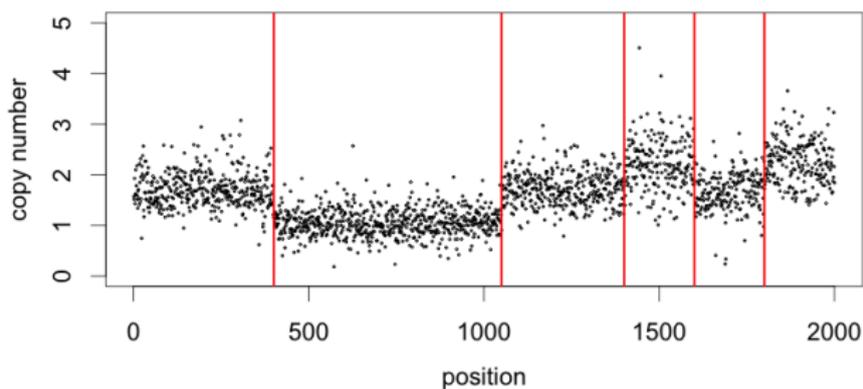
Heuristiques : segmentation binaire récursive



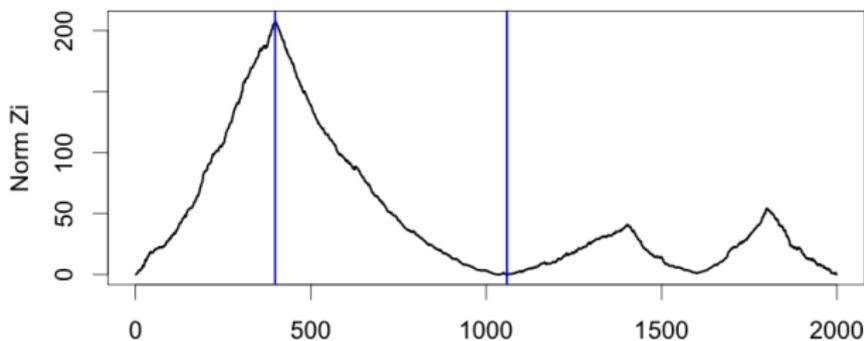
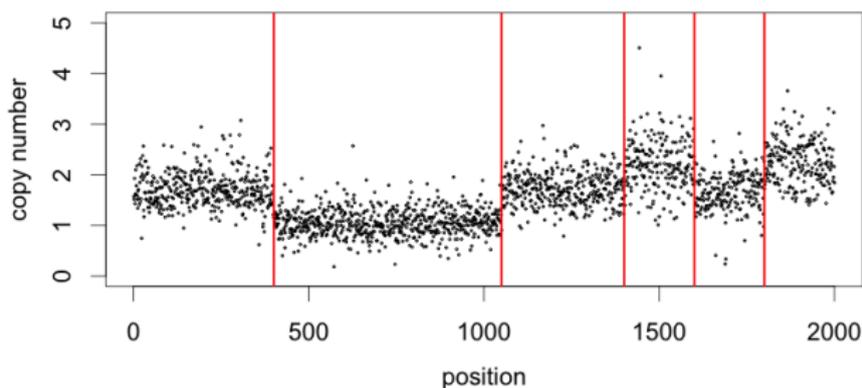
Heuristiques : segmentation binaire récursive



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** :

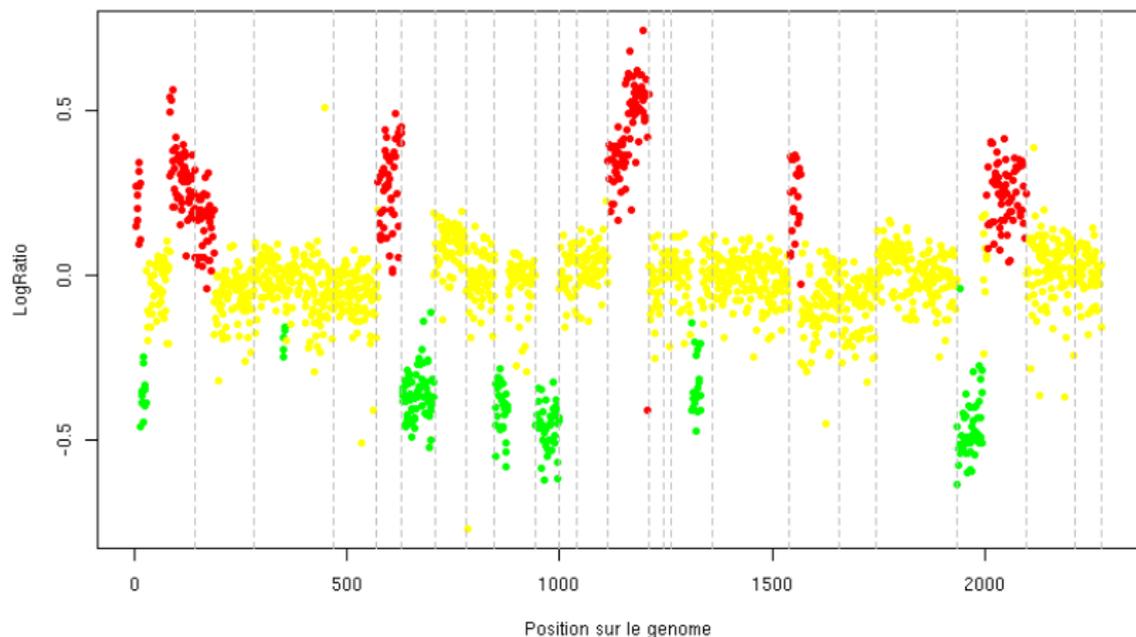
$$Z_{uv} = \frac{1}{\sqrt{\frac{1}{v-u} + \frac{1}{J-(v-u)}}} \times \left[\frac{S_v - S_u}{v-u} - \frac{S_J - (S_v - S_u)}{J-(v-u)} \right]$$

- 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)

Alterations identifiées par Circular Binary Segmentation



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

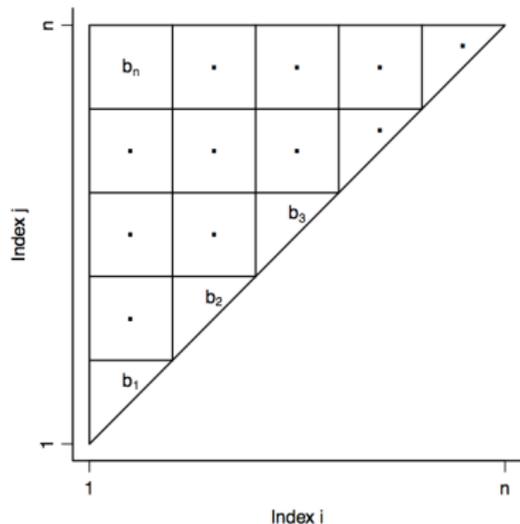
Nouveau mode de calcul des p -values

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

Heuristiques : segmentation binaire récursive

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

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



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

Illustration : V. E. Seshan

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)

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

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 :

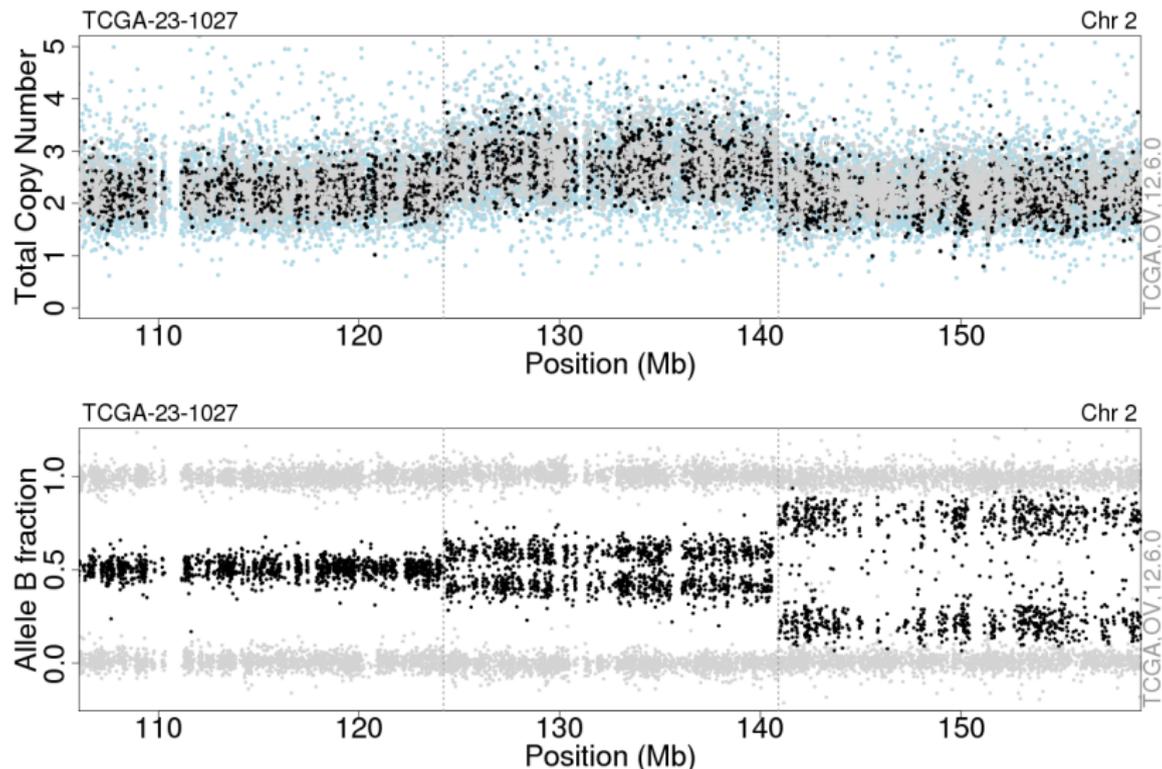
- ➊ recherche très rapide mais avec des faux positifs
- ➋ élagage, ou recherche exhaustive sur les solutions restantes

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

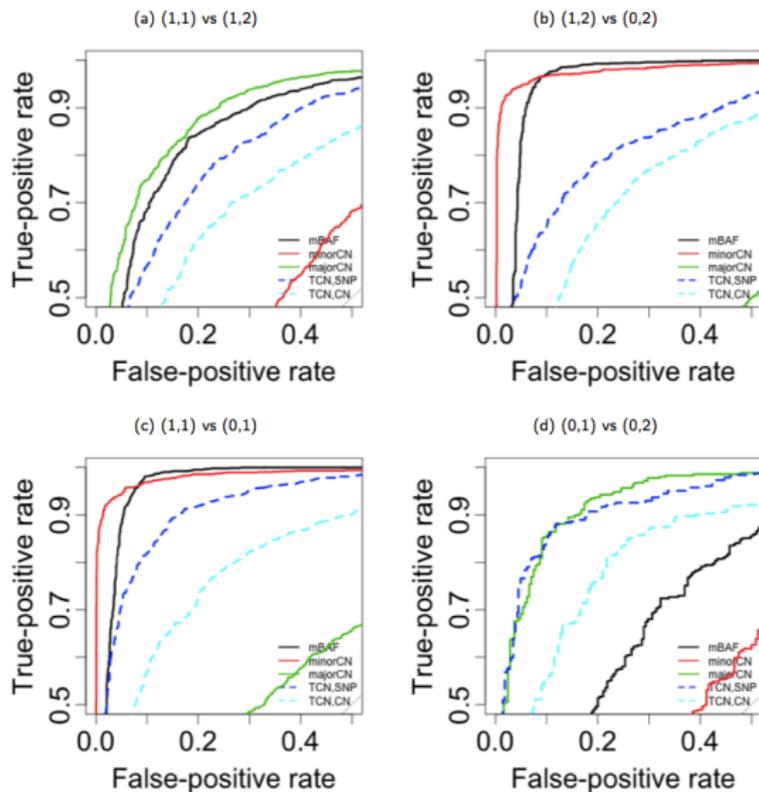
Analyse de données de nombre de copies d'ADN

- 2 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
- 5 **Estimation du nombre de copies d'ADN**
 - **Détection : utilisation conjointe de C et DH**
 - Étiquetage : cellules normales et ploïdie

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

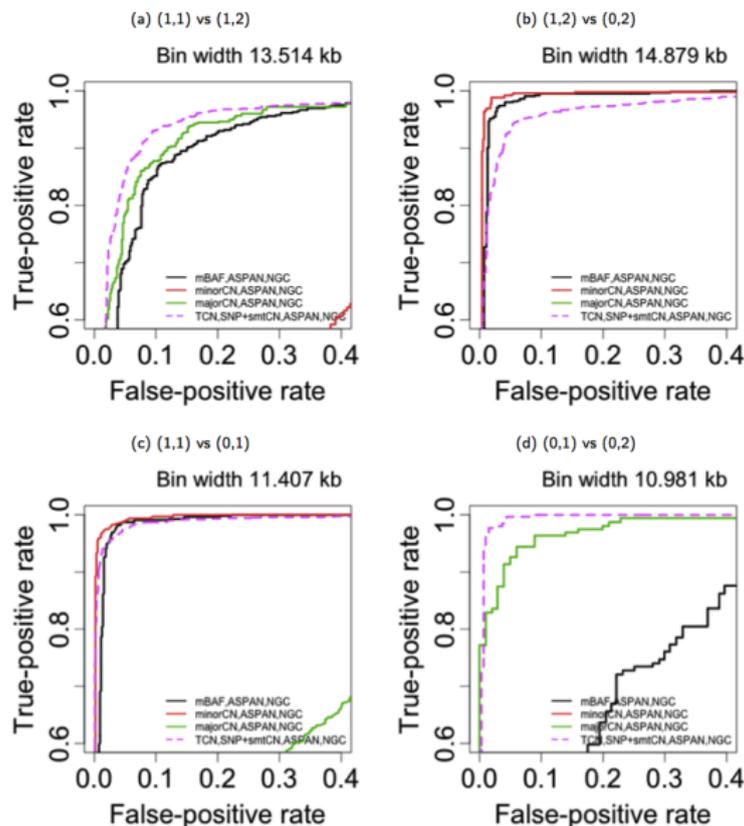
	Toutes sondes	sondes CN	sondes SNP
Effectif	1,856,069	946,705	909,364
Proportion	100%	51%	49%

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

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

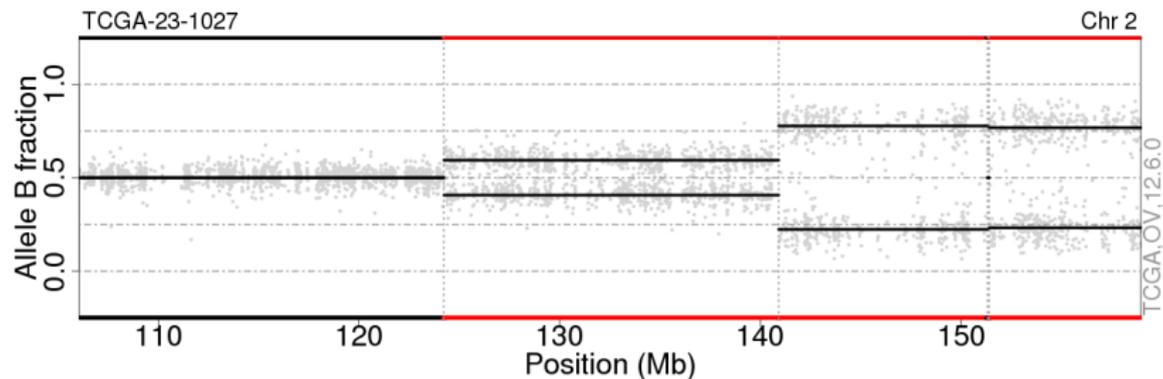
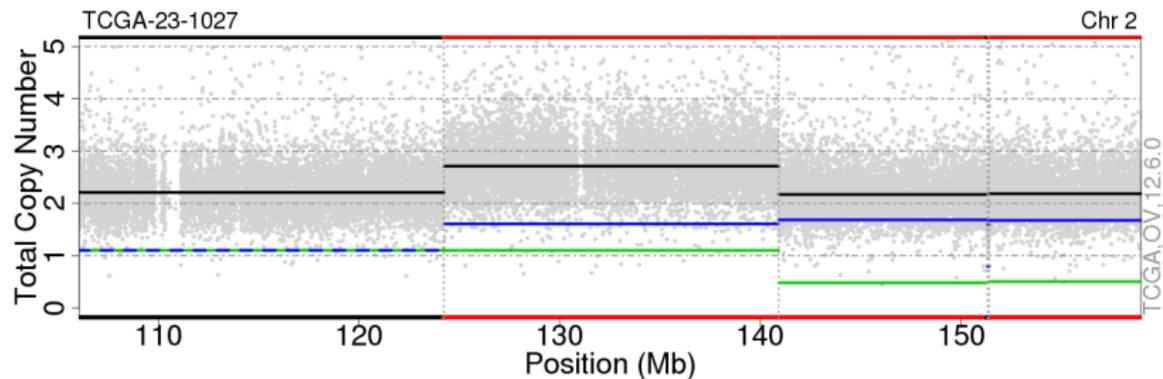
4 Segmentation de données de nombre de copies d'ADN (total)

- Recours aux modèles de rupture
- Exemples d'approches proposées

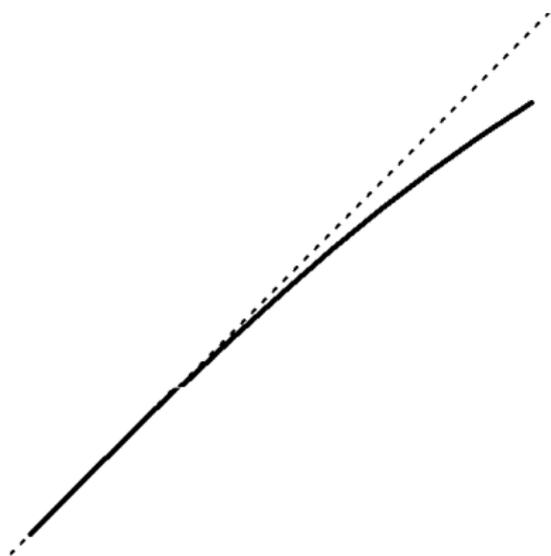
5 Estimation du nombre de copies d'ADN

- Détection : utilisation conjointe de C et DH
- Étiquetage : cellules normales et ploïdie

Les nombres de copies d'ADN ne sont pas calibrés



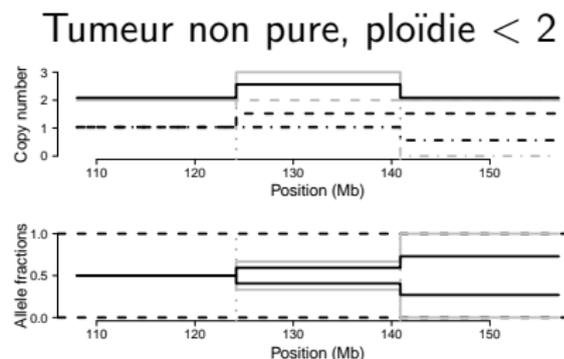
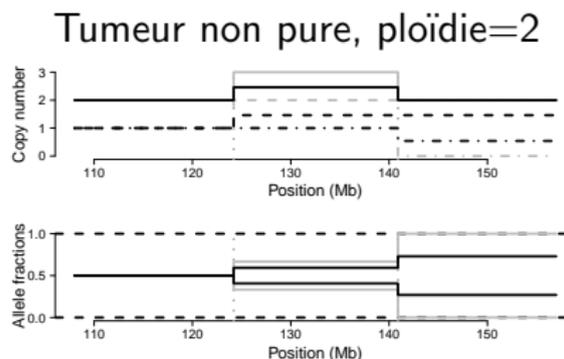
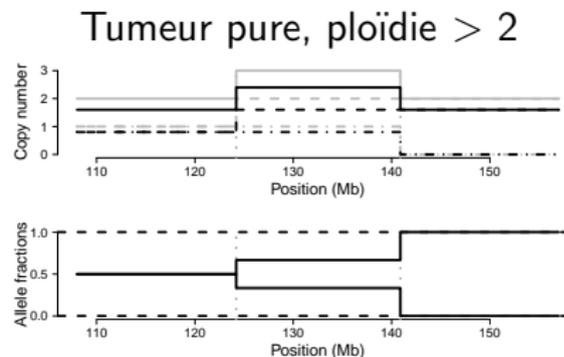
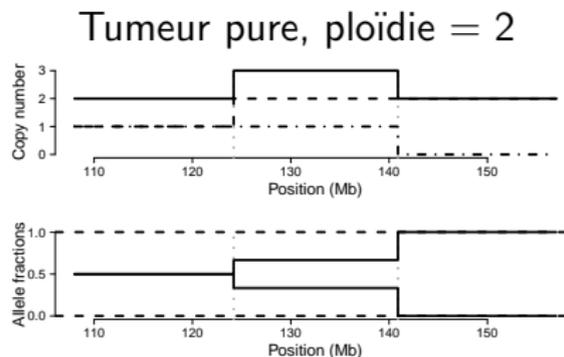
Non-calibration : saturation du signal



$$c = f(\gamma) < \gamma$$

f est inconnue

Non-calibration : ploïdie et cellules normales



Cellules normales, ploïdie, et saturation du signal

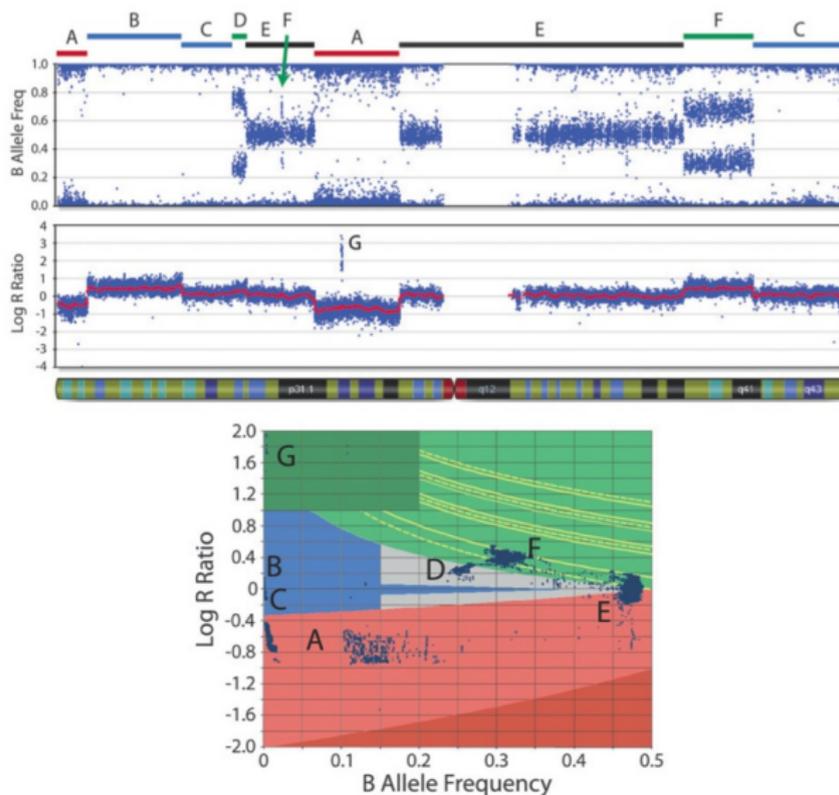
Causes de non-calibration

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

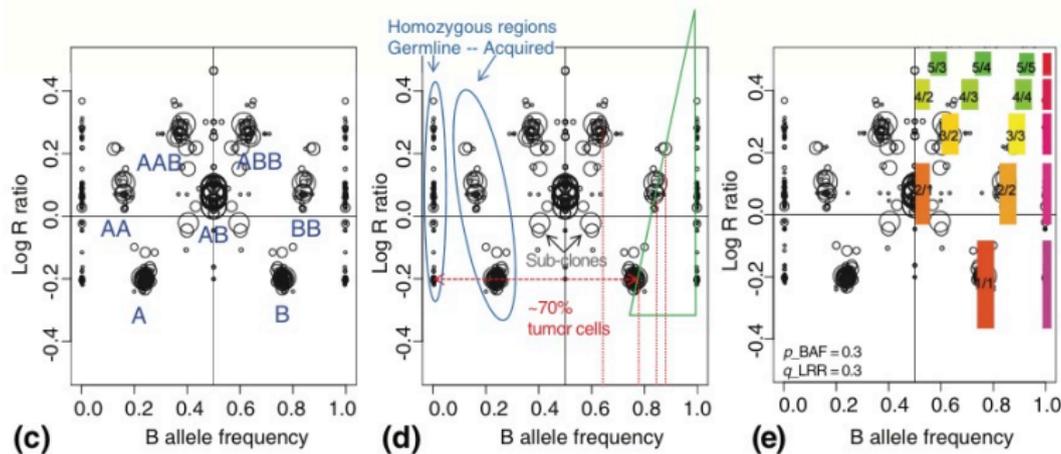
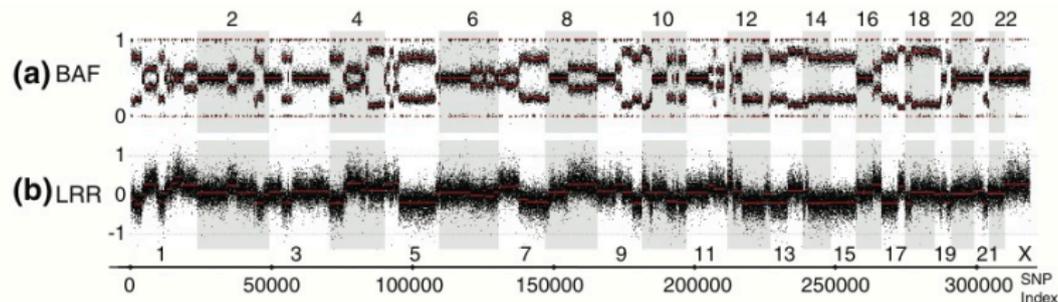
Remarques

- 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

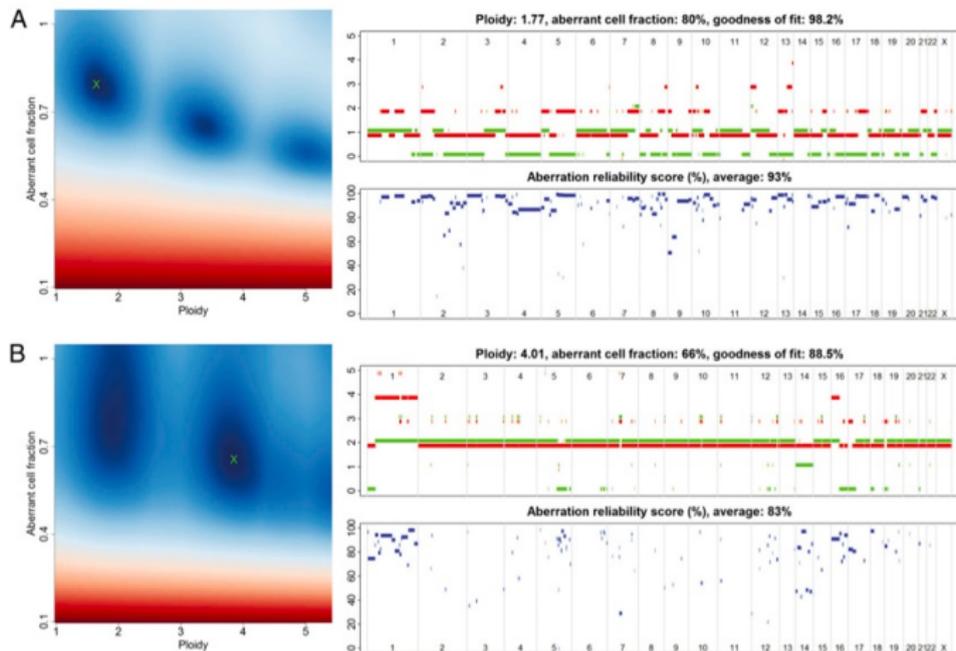


Fig. 1. ASCAT profiles and their calculation. Two examples are given: (A) a tumor with ploidy close to $2n$ and (B) a tumor with ploidy close to $4n$. (Left) ASCAT first determines the ploidy of the tumor cells ψ , and the fraction of aberrant cells ρ . This procedure evaluates the goodness of fit for 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 (conjointe) performantes existent
- 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. Rigaiil and B. Thiam and S. Robin.

Joint segmenation, calling and normalization of multiple CGH profiles.
Biostatistics, 2011.



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

Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays.
PLoS Comput Biol, 2011.

Références : encore des méthodes de segmentation



G. Rigaill.

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



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

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



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

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



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

Stochastic segmentation models for array-based comparative genomic hybridization data analysis.

Biostatistics, (2008)



Z. Harchaoui and C. Lévy-Leduc.

Catching change-points with lasso.

Advances in Neural Information Processing Systems, 2008.

Tests multiples en génomique

- Exemples d'applications
- Définition de critères de risque adaptés

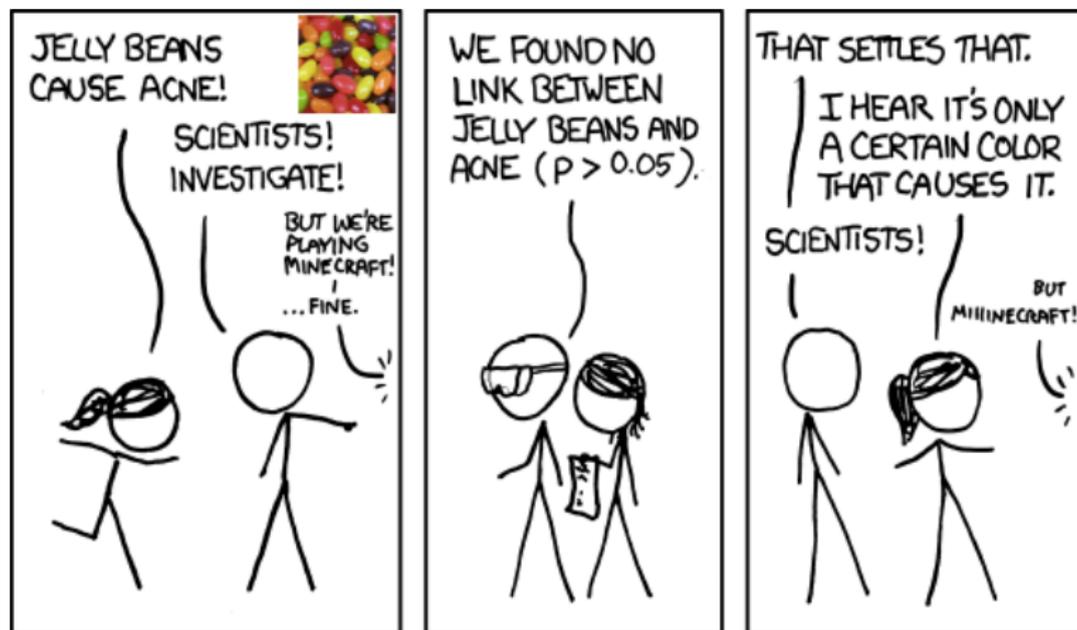
6 Propriétés asymptotiques du FDP

- Cadre d'analyse
- Loi asymptotique du FDP et loi des p -valeurs en 0
- Loi asymptotique du FDP et loi des p -valeurs en 1

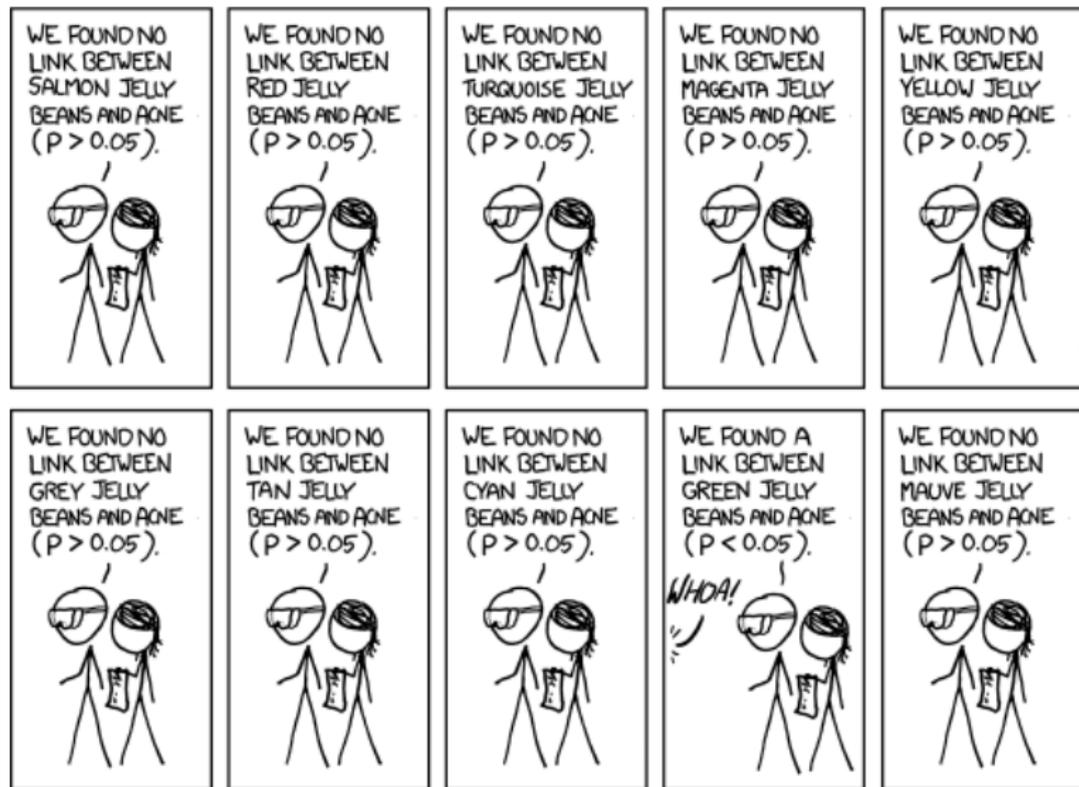
7 Interprétation des résultats d'analyse différentielle

- Enrichissement d'ensembles de gènes
- Tests multivariés

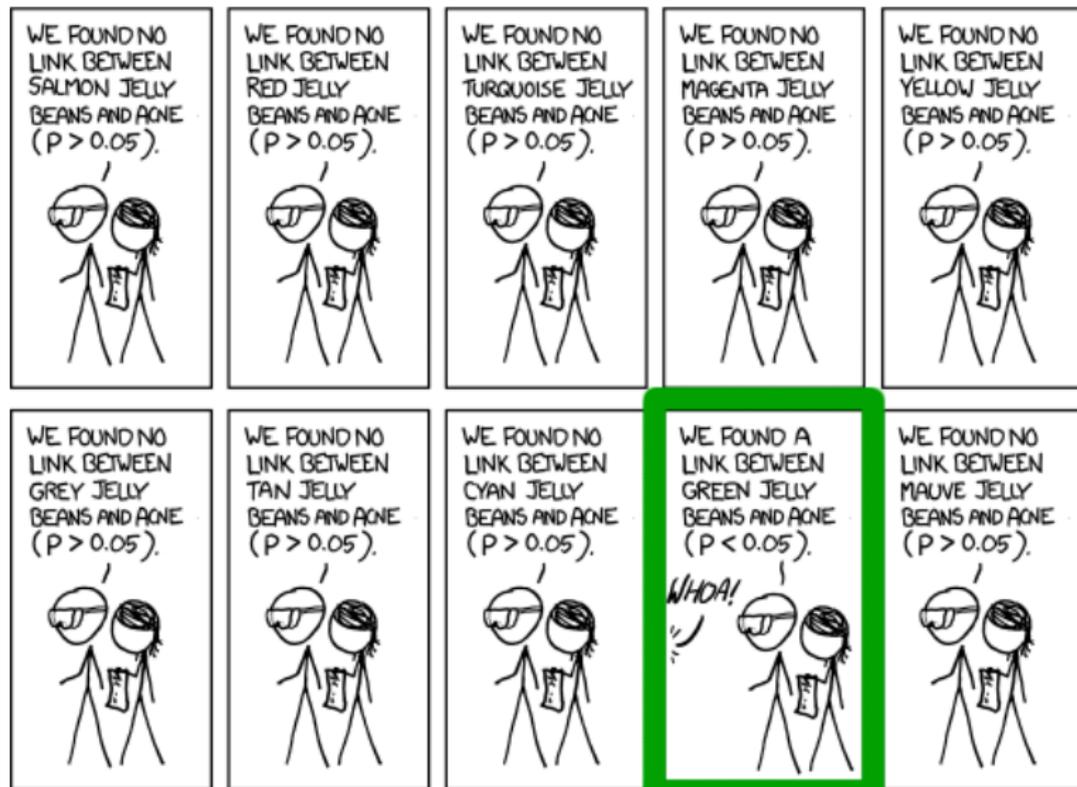
A multiple testing joke



A multiple testing joke



A multiple testing joke



Tests multiples en génomique

- Exemples d'applications
- Définition de critères de risque adaptés

6 Propriétés asymptotiques du FDP

- Cadre d'analyse
- Loi asymptotique du FDP et loi des p -valeurs en 0
- Loi asymptotique du FDP et loi des p -valeurs en 1

7 Interprétation des résultats d'analyse différentielle

- Enrichissement d'ensembles de gènes
- Tests multivariés

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

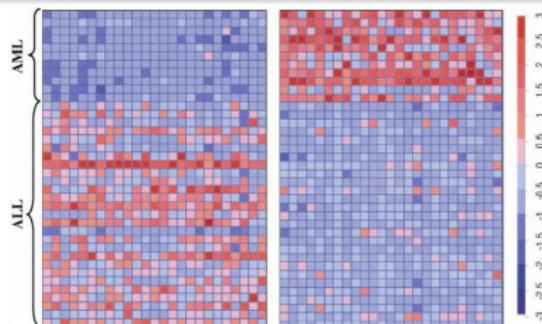
Matrice de données d'expression (données de Golub)

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

Mise en évidence du problème

Contrôler le niveau individuel des tests est insuffisant

- On effectue un test par gène $\Rightarrow m$ p -values (p_1, \dots, 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

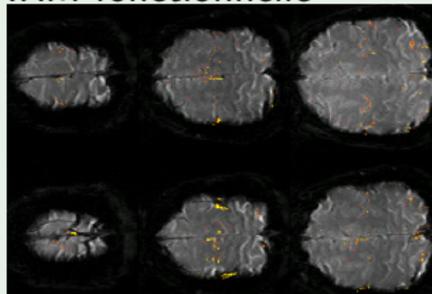
Analyse de données génomiques

- Inférence de réseaux d'expression génique à partir de corrélations
- Spectrométrie de masse
- Analyse différentielle de profils de nombre de copies d'ADN
- Corrélation entre nombre de copies d'ADN et expression

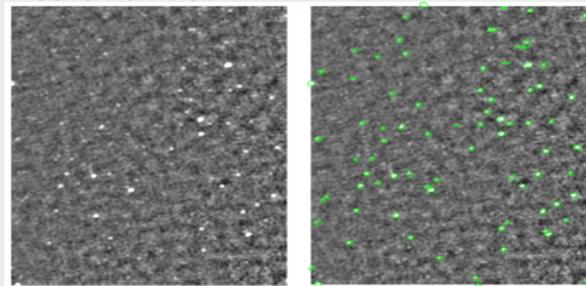
Autres domaines d'application

(images : C. Genovese)

IRM fonctionnelle



Astronomie



Tests multiples en génomique

- Exemples d'applications
- Définition de critères de risque adaptés

6 Propriétés asymptotiques du FDP

- Cadre d'analyse
- Loi asymptotique du FDP et loi des p -valeurs en 0
- Loi asymptotique du FDP et loi des p -valeurs en 1

7 Interprétation des résultats d'analyse différentielle

- Enrichissement d'ensembles de gènes
- Tests multivariés

Référence



E. Roquain. (2011)

Type I error rate control in multiple testing : a survey with proofs.
Journal de la Société Française de Statistique 152(2) :3-38 (2011)
<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**

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 : $FWER = P(V > 0)$

k -Family-wise error rate $k - FWER = P(V > k)$

False Discovery Proportion $FDP = \frac{V}{R \vee 1}$

False Discovery Rate $FDR = E(FDP)$

γ -quantile du FDP $P(FDP > \gamma)$

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 : $FWER = P(V > 0)$

k -Family-wise error rate $k - FWER = P(V > k)$

False Discovery Proportion $FDP = \frac{V}{R \vee 1}$

False Discovery Rate $FDR = E(FDP)$

γ -quantile du FDP $P(FDP > \gamma)$

Contrôle du Family-Wise Error Rate (FWER)

Méthodes "single-step"

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

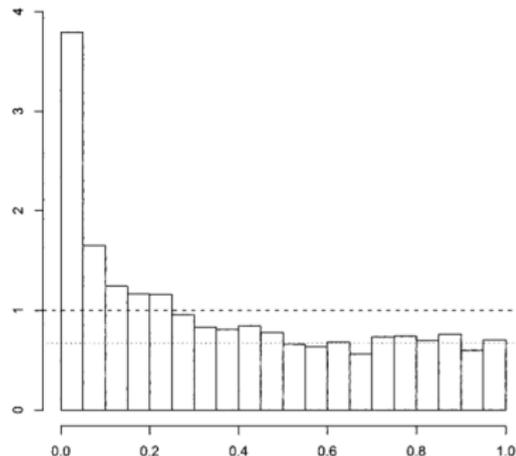
Auteur	seuil t	type de contrôle
Sidak	$1 - (1 - \alpha)^{1/m}$	suppose l'indépendance des gènes calculé sous l'hypothèse $m_0 = m$
Bonferroni	α/m	pas d'hypothèse d'indépendance des gènes très stringent (conservateur)

Variantes et améliorations

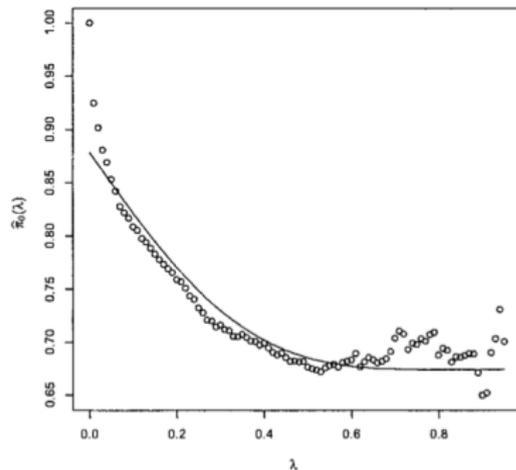
- incorporation d'une estimation de $\pi_0 = m_0/m$
- prise en compte des dépendances entre hypothèses

Estimation de π_0

Méthode dite de “Storey” (Schweder et Spjøtvoll, 1982)



Estimation



Compromis biais/variance

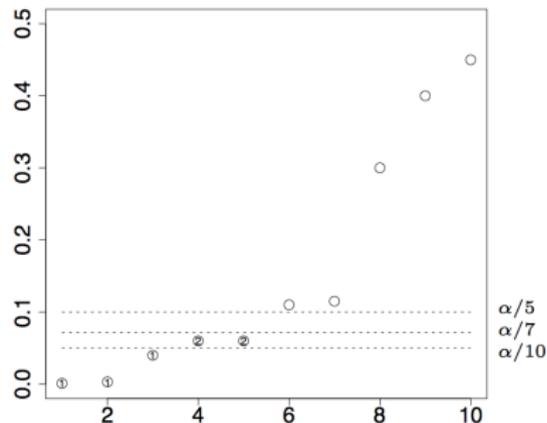
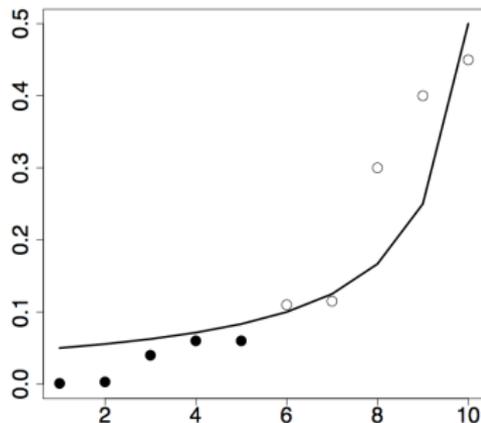
graphiques de Storey & Tibshirani, PNAS 2003.

Estimation de π_0

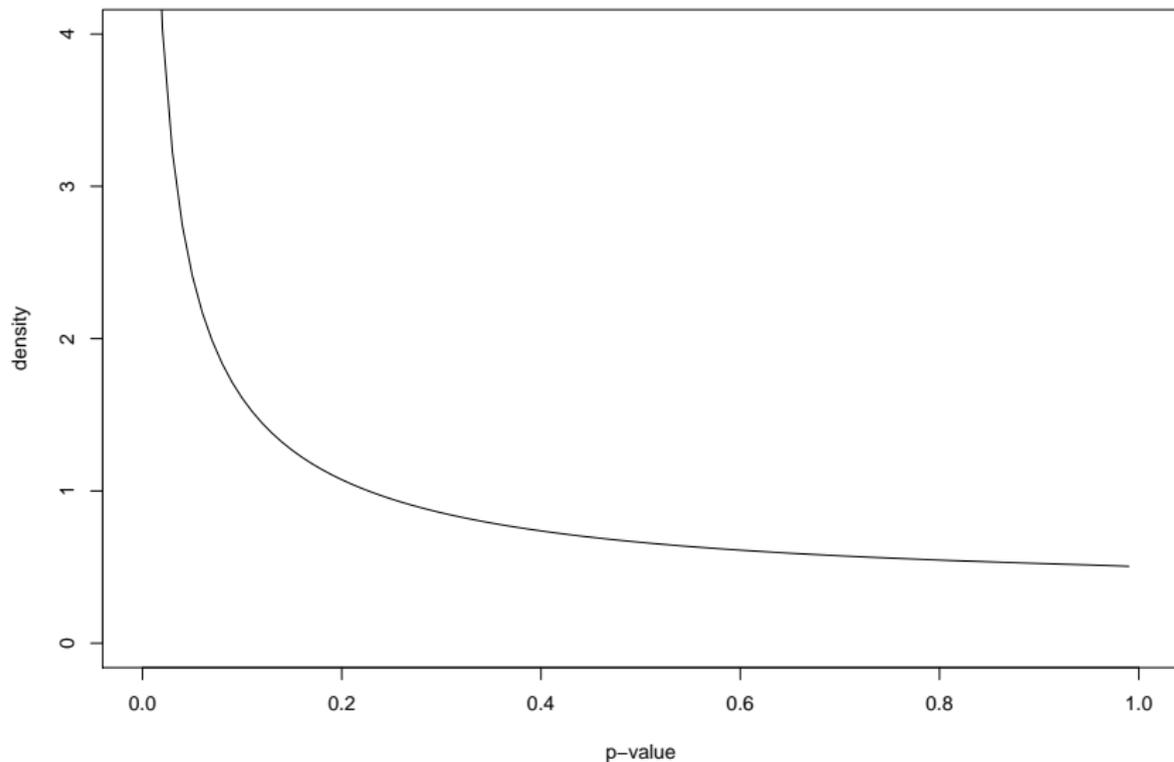
Une méthode "step-down" (Holm, 1979)

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

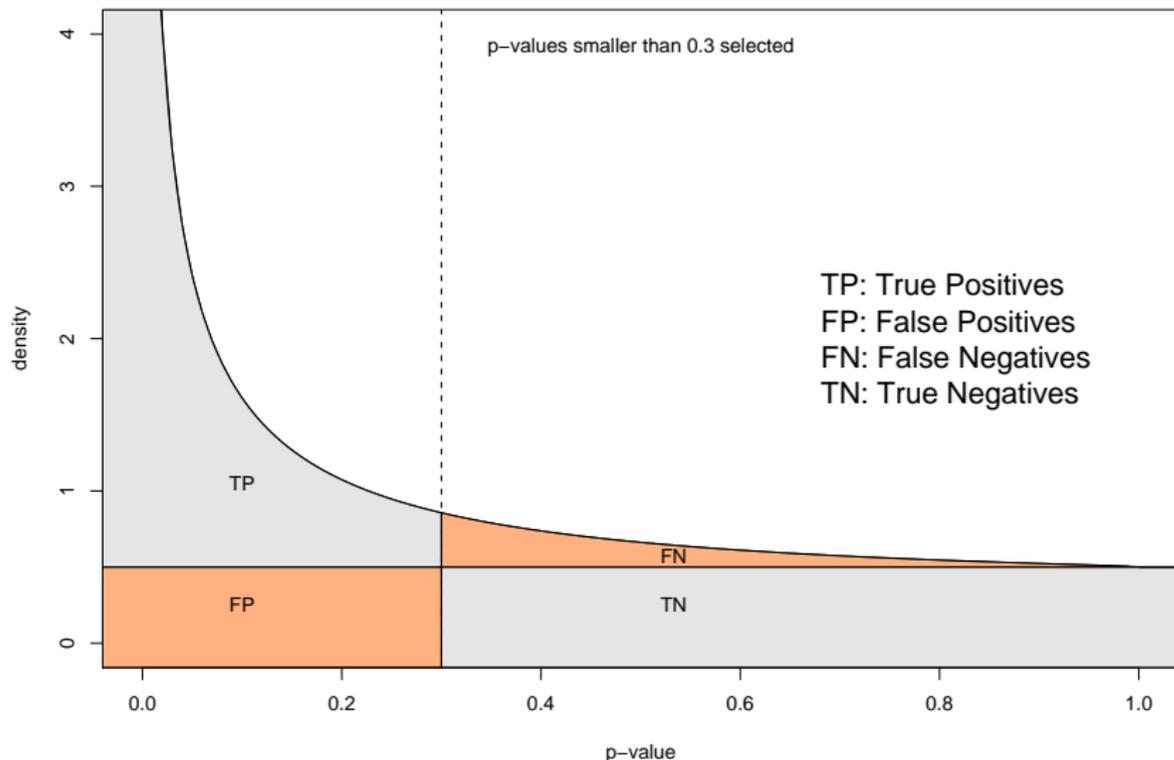
- 1 p -values ordonnées : $p_{(1)} \leq \dots \leq p_{(m)}$
- 2 On définit $\hat{I} = \text{Max} \{i | \forall k \leq i, p_{(k)} \leq \alpha / (m - k + 1)\}$
- 3 On rejette les p -values inférieures à $\alpha / (m - \hat{I} + 1)$



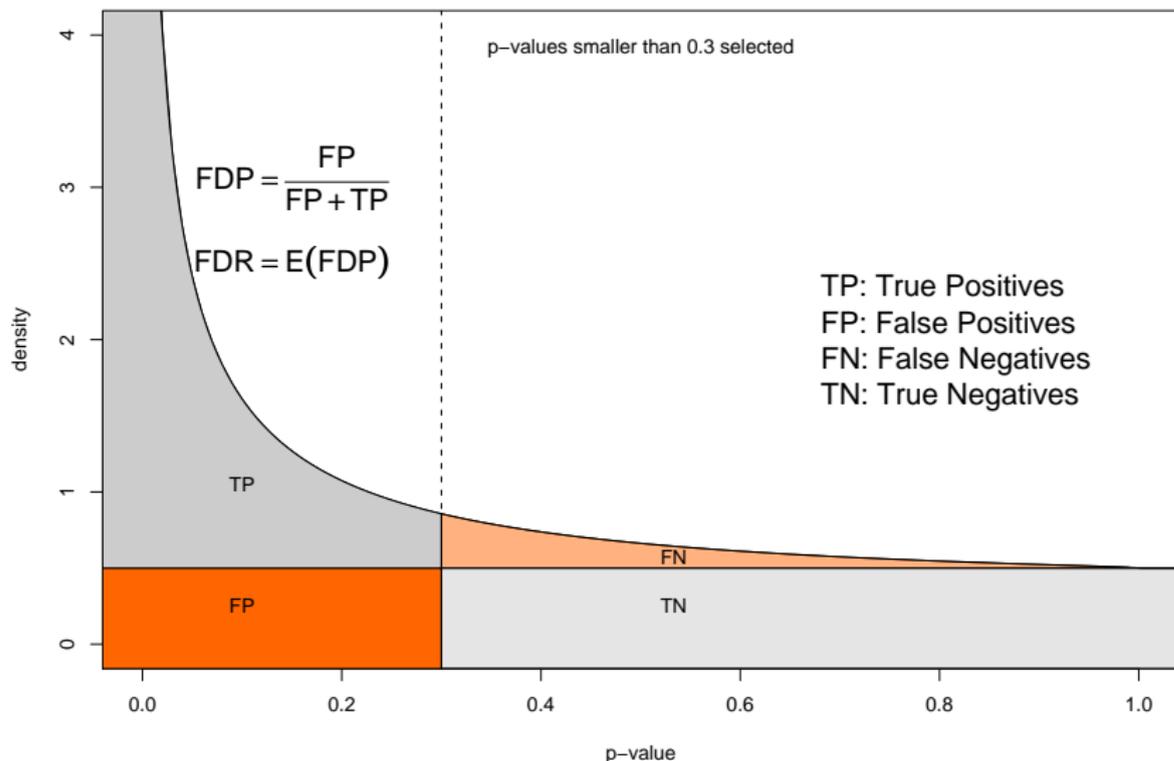
False Discovery Proportion et False Discovery Rate



False Discovery Proportion et False Discovery Rate



False Discovery Proportion et False Discovery Rate

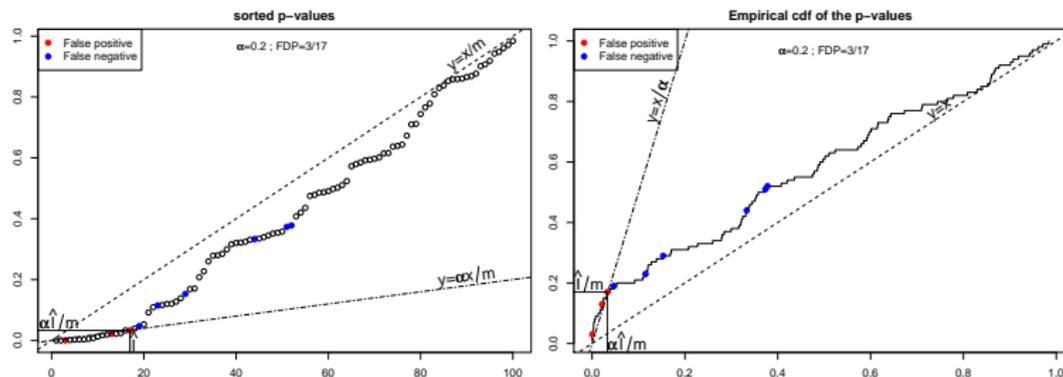


La procédure de Benjamini et Hochberg (1995)

Définition et interprétation graphique

Description de la procédure au niveau α

- 1 p -values ordonnées : $P_{(1)} \leq \dots \leq P_{(m)}$
- 2 On définit $\hat{l} = \text{Max} \{k | P_{(k)} \leq \alpha \frac{k}{m}\}$
- 3 On rejette les p -values inférieures à $P_{(\hat{l})} = \alpha \frac{\hat{l}}{m}$



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

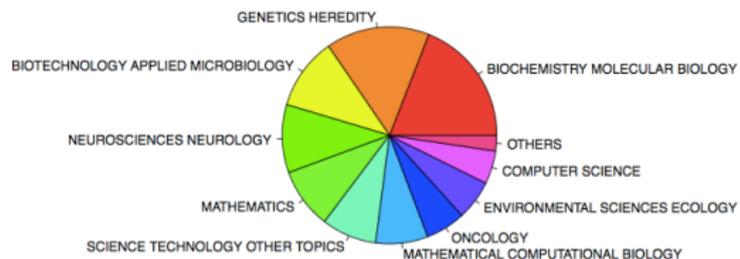
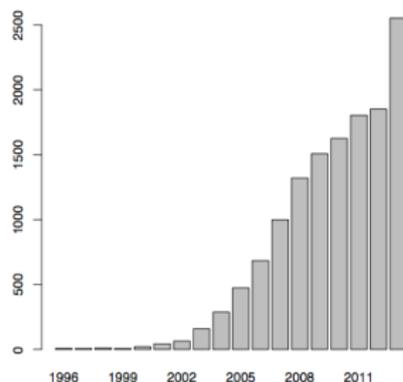
- soit indépendantes
- soit positivement associées (PRDS)

Améliorations de l'approche de Benjamini et Hochberg

- contrôle avec hypothèses de dépendance entre les gènes
- estimation de π_0 pour un contrôle plus fin du FDR

Bibliométrie

Citations de l'article BH95 en 2013



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	$n_2 = 27$

Résultats (avec un test de Wilcoxon pour chaque gène)

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

- Bonferroni : 87
- Sidak, Holm : 92
- Benjamini-Hochberg : 679
- sans correction : 1055

Adaptation à la densité de signal ?

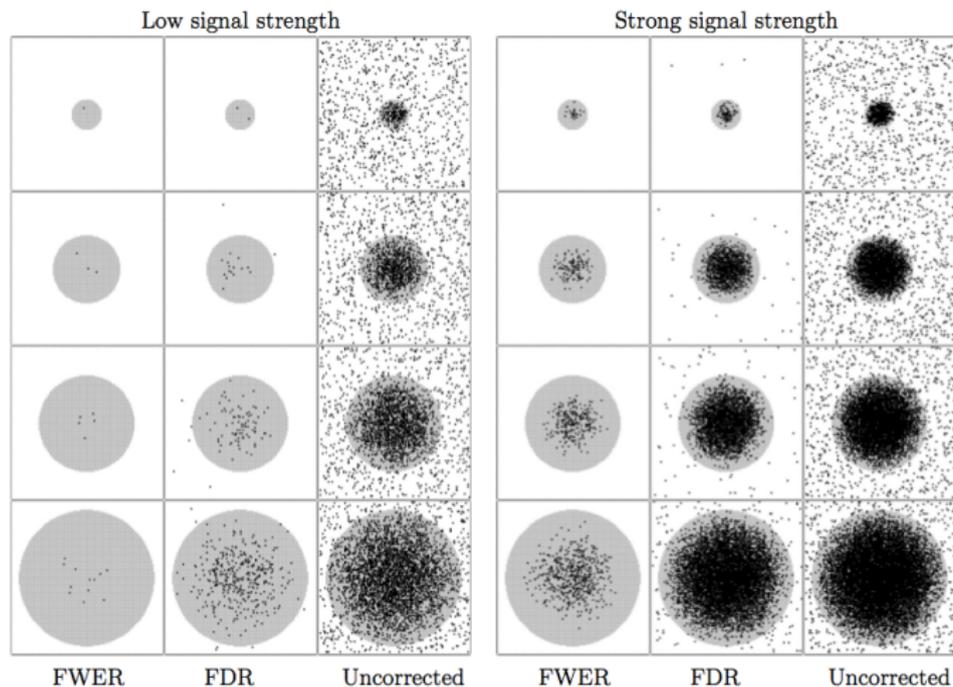
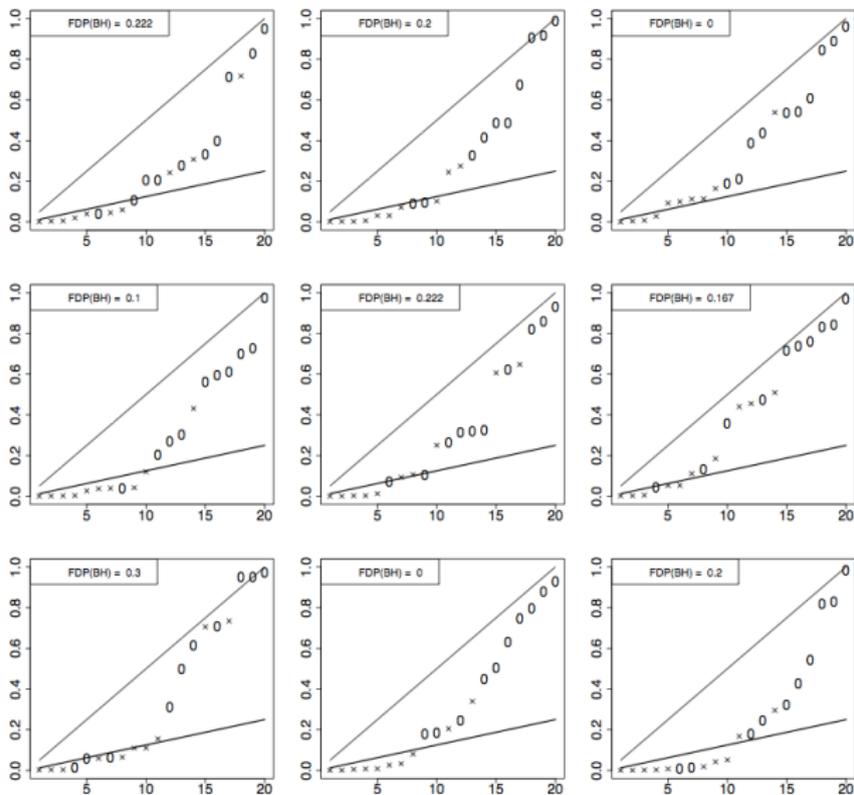


Illustration : E. Roquain

FDR et FDP



Tests multiples en génomique

- Exemples d'applications
- Définition de critères de risque adaptés

6 Propriétés asymptotiques du FDP

- Cadre d'analyse
- Loi asymptotique du FDP et loi des p -valeurs en 0
- Loi asymptotique du FDP et loi des p -valeurs en 1

7 Interprétation des résultats d'analyse différentielle

- Enrichissement d'ensembles de gènes
- Tests multivariés

Modèle de mélange

Notation et hypothèses

- \mathcal{H}_0 : hypothèse nulle (par ex. absence d'expression différentielle)
- \mathcal{H}_1 : hypothèse alternative (par ex. expression différentielle)
- m : nombre d'hypothèses testées
- π_0 : proportion (fixée) d'hypothèses nulles vraies

Hypothèses

- $(P_i)_{1 \leq i \leq m}$ sont indépendentes
- $P_i | \mathcal{H}_0 \sim \mathcal{U}(0, 1)$, $P_i | \mathcal{H}_1 \sim G_1$

On a alors, $(P_i)_{1 \leq i \leq m} \stackrel{iid}{\sim} G : x \mapsto \pi_0 x + (1 - \pi_0) G_1(x)$

- G est continue, croissante et strictement concave, i.e.
 $g = \pi_0 + (1 - \pi_0)g_1$ est décroissante

Procédures de tests multiples et fonctions de seuil

Définition d'une procédure de tests multiples (PTM)

Une collection $\mathcal{M} = (\mathcal{M}_m)_{m \in \mathbb{N}}$ telle que $\mathcal{M}_m : [0, 1]^m \rightarrow [0, 1]$ rejette les hypothèses i vérifiant

$$P_i \leq \mathcal{M}_m(P_1, \dots, P_m)$$

pour tout vecteur de p -values de dimension $m : (P_1, \dots, P_m)$

Définition d'une fonction de seuil

La procédure \mathcal{M} admet pour fonction de seuil $\mathcal{T} : D[0, 1] \rightarrow [0, 1]$ si et seulement si

$$\forall m \in \mathbb{N}, \mathcal{M}_m(P_1, \dots, P_m) = \mathcal{T}(\widehat{\mathbb{G}}_m)$$

Le FDP : processus stochastique d'un seuil aléatoire

- Comment le FDP se **concentre**-t-il autour du FDR ?
- Étude du processus stochastique $(FDP(t))_{0 < t < 1}$: Genovese & Wasserman (Ann. Stat., 2004), Storey, Taylor & Siegmund (JRSS B, 2004)
- Quid du FDP **effectivement atteint par une PTM donnée** ?

$$FDP(\mathcal{T}(\hat{\mathbb{G}}_m))$$

Idée principale

- Établir des propriétés asymptotiques (en m) du FDP atteint grâce à la méthode du Delta fonctionnelle
- Comprendre comment ces propriétés dépendent de la **procédure** elle-même, et de la **distribution des p -values**

Exemples de fonction de seuil

Procédure BH95

Benjamini & Hochberg, JRSS B, 1995

$$\mathcal{T}^{\text{BH95}}(F) = \sup \{u \in [0, 1], F(u) \geq u/\alpha\}$$

La procédure BH95 est conservatrice

Si les p -values sont indépendantes, la procédure BH95 au niveau α donne

$$FDR \leq \pi_0 \alpha \leq \alpha$$

La procédure Sto02(λ)

Storey, JRSS B, 2002

Appliquer la procédure BH95 au niveau $\alpha/\hat{\pi}_0(\lambda)$, où

$$\hat{\pi}_0(\lambda) = \frac{1 - \hat{\mathbb{G}}_m(\lambda)}{1 - \lambda}$$

Formalisme utilisé pour l'étude asymptotique

Deux grands types de raffinements de la procédure BH95

substitution : application de BH95 au niveau $\alpha/\hat{\pi}_0$, où $\hat{\pi}_0$ est un estimateur de π_0

adaptatives : comparaison de $\hat{\mathbb{G}}_m$ à une **courbe de rejet** non linéaire

Formalisme : fonctions de seuil

Pour BH95 et les procédures de substitution (plug-in) :

$$\mathcal{T}(F) = \sup\{u \in [0, 1], F(u) \geq u/\mathcal{A}(F)\}$$

- BH95 $\mathcal{A} = \alpha$
- Sto02(λ) $\mathcal{A}(F) = \alpha \frac{1-\lambda}{1-F(\lambda)}$

Tests multiples en génomique

- Exemples d'applications
- Définition de critères de risque adaptés

6 Propriétés asymptotiques du FDP

- Cadre d'analyse
- Loi asymptotique du FDP et loi des p -valeurs en 0
- Loi asymptotique du FDP et loi des p -valeurs en 1

7 Interprétation des résultats d'analyse différentielle

- Enrichissement d'ensembles de gènes
- Tests multivariés

Théorème de Donsker

Theorem (Donsker)

Si les p -valeurs sont indépendantes, alors

$$\textcircled{1} \quad \sqrt{m} \left(\begin{pmatrix} \widehat{G}_{0,m} \\ \widehat{G}_{1,m} \end{pmatrix} - \begin{pmatrix} G_0 \\ G_1 \end{pmatrix} \right) \rightsquigarrow \begin{pmatrix} Z_0 \\ Z_1 \end{pmatrix} \quad \text{on } [0, 1], \text{ où } Z_0 \text{ et } Z_1 \text{ sont de}$$

processus Gaussiens indépendants tels que $Z_0 \stackrel{(d)}{=} \frac{1}{\sqrt{\pi_0}} \mathbb{B}$ et

$Z_1 \stackrel{(d)}{=} \frac{1}{\sqrt{1-\pi_0}} \mathbb{B} \circ G_1$ et \mathbb{B} est un pont Brownien standard sur $[0, 1]$.

$$\textcircled{2} \quad \sqrt{m}(\widehat{G}_m - G) \rightsquigarrow Z \quad \text{on } [0, 1], \text{ où } Z = \pi_0 Z_0 + (1 - \pi_0) Z_1 \text{ est un processus Gaussien à trajectoires continues ayant pour fonction de covariance}$$

$$\mathbb{E}[Z(s)Z(t)] = \pi_0 \gamma_0(s, t) + (1 - \pi_0) \gamma_0(G_1(s), G_1(t)),$$

où γ_0 est la fonction de covariance de \mathbb{B} , i.e. $\gamma_0 : (s, t) \mapsto s \wedge t - st$.

Résultat général

Theorem (Loi asymptotique de FDP_m pour la procédure \mathcal{T})

Soit \mathcal{T} une fonction de seuil, et $\tau^* = \mathcal{T}(G)$. Si \mathcal{T} est différentiable au sens de Hadamard en G , alors

$$\sqrt{m} \left(FDP_m(\mathcal{T}(\hat{G}_m)) - \frac{\pi_0 \tau^*}{G(\tau^*)} \right) \rightsquigarrow X,$$

où

$$X = \frac{\pi_0}{G(\tau^*)} \left(\mathbb{Z}_0(\tau^*) - \frac{\tau^*}{G(\tau^*)} \mathbb{Z}(\tau^*) + \left(1 - \frac{\tau^* g(\tau^*)}{G(\tau^*)} \right) \dot{\mathcal{T}}_G(\mathbb{Z}) \right),$$

- valable quelle que soit la forme de la fonction de seuil
- contrôle asymptotique du FDR au niveau α dès que $\frac{\pi_0 \tau^*}{G(\tau^*)} \leq \alpha$

Application aux procédures étudiées

Forme de la dérivée au sens de Hadamard

Si

$$\mathcal{T}(F) = \mathcal{U}(F, \mathcal{A}(F))$$

où $\mathcal{U}(F, \alpha) = \sup\{u \in [0, 1], F(u) \geq r_\alpha(u)\}$, alors pour tout $H \in D[0, 1]$,

$$\dot{\mathcal{T}}_G(H) = \nabla_F \mathcal{U}_{(G, \mathcal{A}(G))}(H) + \dot{\mathcal{A}}_G(H) \nabla_\alpha \mathcal{U}_{(G, \mathcal{A}(G))}$$

Cas particuliers

- pour les procédures adaptatives, $\dot{\mathcal{A}}_G = 0$
- pour les procédures de plug-in, $\dot{\mathcal{A}}_G$ fait typiquement apparaître le point utilisé pour estimer π_0 (e.g. λ for Sto02)

Théorème de la limite centrale pour le seuil BH95

Theorem (Procédure BH95)

Soient $\alpha^* = 1/g(0)$ et $\tau^* = \mathcal{T}^{\text{BH95}}(G)$. Si $\alpha > \alpha^*$, alors

$$\sqrt{m} \left(\text{FDP}_m^{\text{BH95}} - \pi_0 \alpha \right) \rightsquigarrow \mathcal{N} \left(0, (\pi_0 \alpha)^2 \frac{1 - \tau^*}{\tau^*} \right)$$

Lien avec la criticité

α^* est la **valeur critique** identifiée par Chi (Ann. Stat., 2007) :

- Si $\alpha < \alpha^*$, le nombre de (vraies) découvertes est borné lorsque le nombre d'hypothèses augmente ;
- Si $\alpha > \alpha^*$, la proportion de découvertes converge en probabilité vers une valeur strictement positive : $\tau^* = \mathcal{T}(G)$.

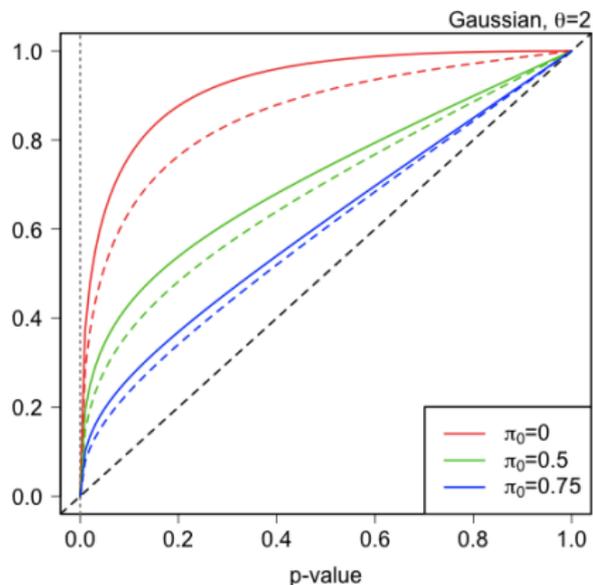
Criticité dans les modèles de localisation

Fonction de répartition des p -valeurs

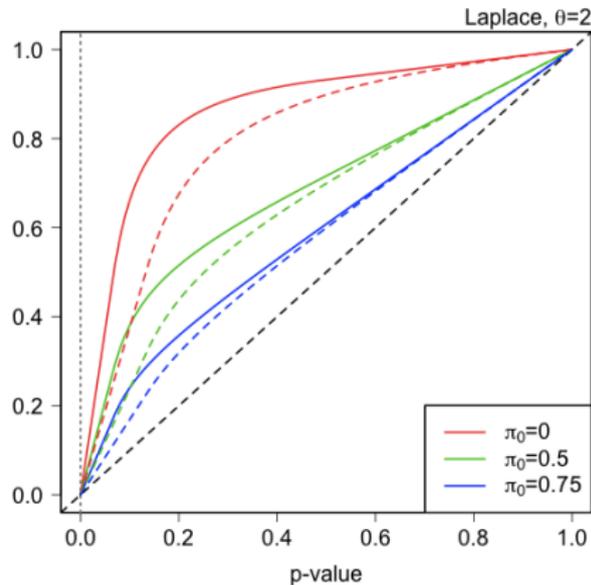
La statistique de test suit f_0 sous \mathcal{H}_0 et $f_1 : t \mapsto f_0(t - \theta)$ sous \mathcal{H}_1

Gaussien : $f_0(t) = \frac{1}{2\pi} e^{-t^2/2}$

Laplace : $f_0(t) = \frac{1}{2} e^{-|t|}$



Situation non critique



Situation critique

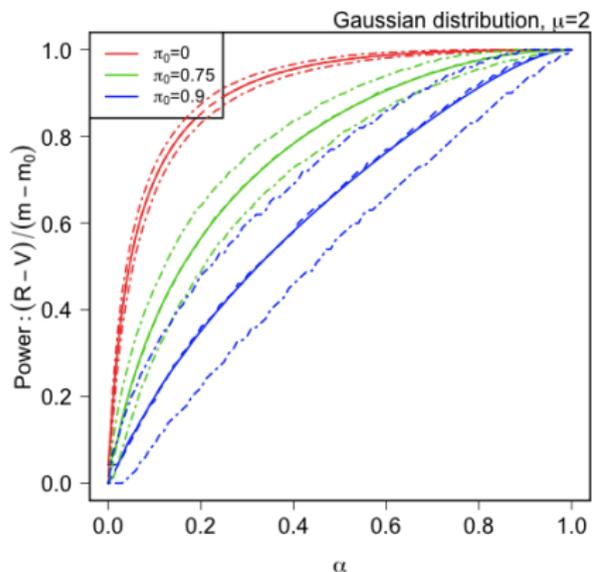
Criticité dans les modèles de localisation

Puissance de la procédure BH95 (non observable)

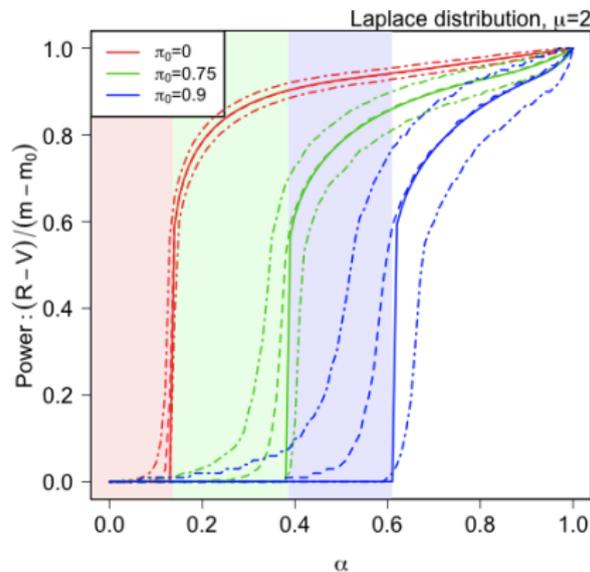
La statistique de test suit f_0 sous \mathcal{H}_0 et $f_1 : t \mapsto f_0(t - \theta)$ sous \mathcal{H}_1

Gaussien : $f_0(t) = \frac{1}{2\pi} e^{-t^2/2}$

Laplace : $f_0(t) = \frac{1}{2} e^{-|t|}$



Situation non critique



Situation critique

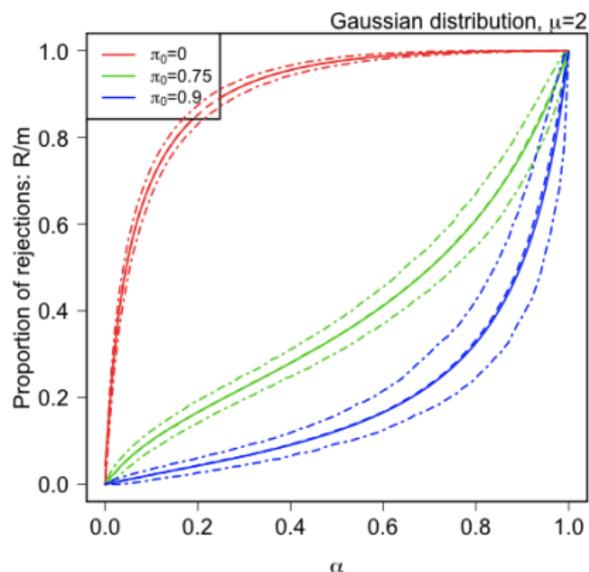
Criticité dans les modèles de localisation

Proportion d'hypothèses rejetées par BH95 (observable)

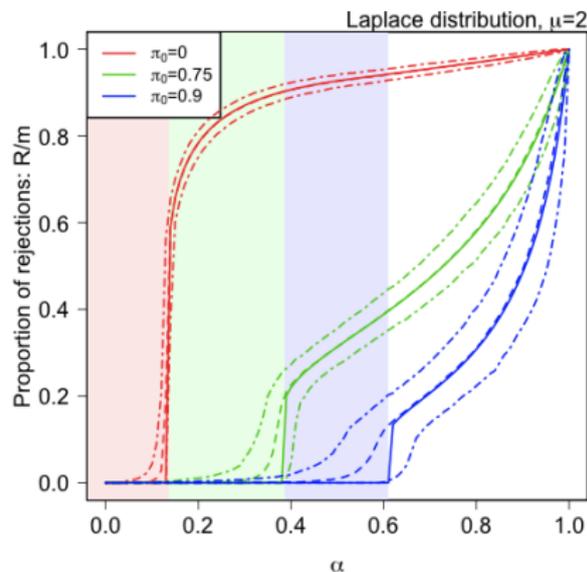
La statistique de test suit f_0 sous \mathcal{H}_0 et $f_1 : t \mapsto f_0(t - \theta)$ sous \mathcal{H}_1

Gaussien : $f_0(t) = \frac{1}{2\pi} e^{-t^2/2}$

Laplace : $f_0(t) = \frac{1}{2} e^{-|t|}$

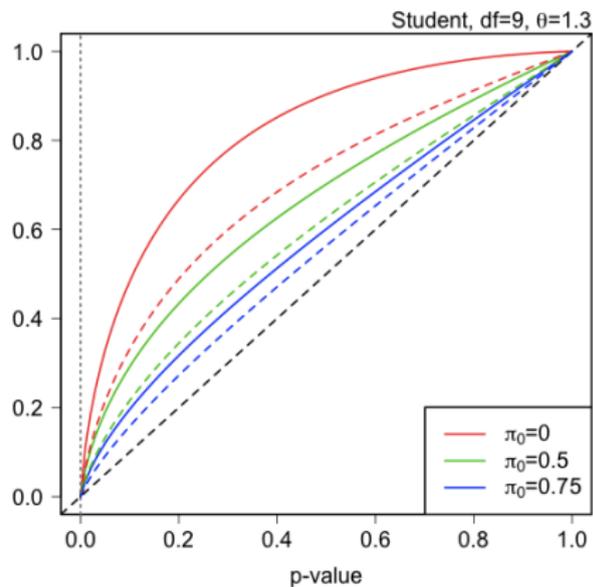
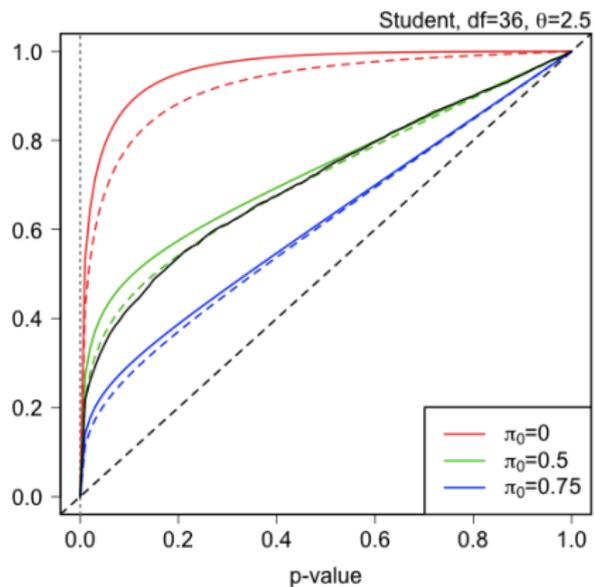


Situation non critique

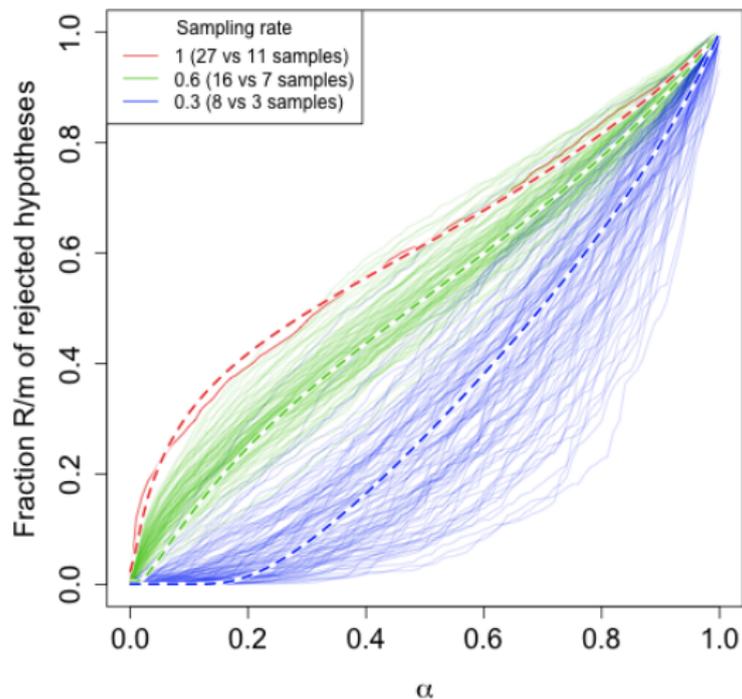


Situation critique

Criticité pour les tests de Student



Criticité : illustration sur des données réelles



Théorème de la limite centrale pour le seuil $\text{Sto02}(\lambda)$

$$\mathcal{T}^{\text{Sto02}(\lambda)}(F) = \sup \left\{ u \in [0, 1], F(u) \geq \frac{u}{\alpha} \frac{1 - F(\lambda)}{1 - \lambda} \right\}$$

Theorem (Procédure $\text{Sto02}(\lambda)$)

Soit $\bar{\pi}_0(\lambda) = \frac{1 - G(\lambda)}{1 - \lambda}$. Si $\alpha > \bar{\pi}_0(\lambda) \alpha^*$, alors

$$\sqrt{m} \left(\text{FDP}_m^{\text{Sto02}(\lambda)} - \frac{\pi_0}{\bar{\pi}_0(\lambda)} \alpha \right) \rightsquigarrow X^{\text{Sto02}(\lambda)},$$

où $X^{\text{Sto02}(\lambda)}$ suit une loi normale centrée, dont la variance dépend de α , $\mathcal{T}^{\text{Sto02}(\lambda)}(G)$ et λ .

Conséquences des résultats

Applications directes

- Intervalles de confiance asymptotique du $FDP(\mathcal{T}(\hat{\mathbb{G}}_m))$
- Comparaisons de puissance entre procédures

Extension au cas Gaussian equi-corrélé

Delattre & Roquain, 2011

Cadre : $(X_i)_{1 \leq i \leq m} \sim \mathcal{N}(\mu, \Sigma)$, avec $\Sigma_{ii} = 1$ et $\Sigma_{ij} = \rho_m$ pour $i \neq j$

- Même résultats que ci-dessus si $\rho_m \rightarrow 0$ à vitesse $1/\sqrt{m}$ ou plus rapidement
- Vitesse de convergence $\rho_m^{1/2}$ si $\rho_m \rightarrow 0$ plus lentement que $1/\sqrt{m}$

⇒ La dépendance entre hypothèses testées perturbe la concentration du FDP autour de son espérance

Tests multiples en génomique

- Exemples d'applications
- Définition de critères de risque adaptés

6 Propriétés asymptotiques du FDP

- Cadre d'analyse
- Loi asymptotique du FDP et loi des p -valeurs en 0
- Loi asymptotique du FDP et loi des p -valeurs en 1

7 Interprétation des résultats d'analyse différentielle

- Enrichissement d'ensembles de gènes
- Tests multivariés

Idée : rendre l'estimateur de Storey's consistant

On suppose que π_0 est identifiable. Pour $0 < \lambda < 1$,

$$\hat{\pi}_0(\lambda) = \frac{1 - \hat{\mathbb{G}}_m(\lambda)}{1 - \lambda}$$

converge à vitesse $1/\sqrt{m}$ vers $\frac{1-G(\lambda)}{1-\lambda}$.

⇒ Pour estimer π_0 sans biais asymptotiquement, faire tendre $\lambda \rightarrow 1$ quand $m \rightarrow +\infty$.

Proposition (Loi asymptotique de $\hat{\pi}_0(1 - h_m)$)

On suppose $g(1) = 0$ (lié à l'identifiabilité de π_0). Soit $(h_m)_{m \in \mathbb{N}}$ une suite strictement positive telle que $h_m \rightarrow 0$ et $mh_m \rightarrow +\infty$ lorsque $m \rightarrow +\infty$.

Alors

$$\sqrt{mh_m} (\hat{\pi}_0(1 - h_m) - \mathbb{E}[\hat{\pi}_0(1 - h_m)]) \rightsquigarrow \mathcal{N}(0, \pi_0).$$

Choix de fenêtre optimale pour l'estimateur de Storey

Theorem

Si g est C^k en 1, avec $g^{(l)}(1) = 0$ pour $0 \leq l < k$, et $g^{(k)}(1) \neq 0$, alors :

- ① La fenêtre optimale en termes d'erreur quadratique moyenne est donnée par $h_m^*(k) = C_k m^{-\frac{k}{2k+1}}$, où C_k est une constant explicite qui dépend de k , π_0 , et $g^{(k)}(1)$;
- ② Pour tout η_m tendant vers 0, si $h_m(k) = h_m^*(k)\eta_m^2$, on a

$$m^{\frac{k}{2k+1}} \eta_m (FDP_m - \alpha) \rightsquigarrow \mathcal{N} \left(0, \frac{\alpha^2 C_k}{\pi_0} \right).$$

Idée de la preuve

- L'estimateur de Storey est un cas particulier d'**estimateur à noyau**
- La régularité de g en 1 détermine donc son biais asymptotique

Convergence lente même dans des cas simples

Localisation Gaussienne, tests unilatéraux

Modèle

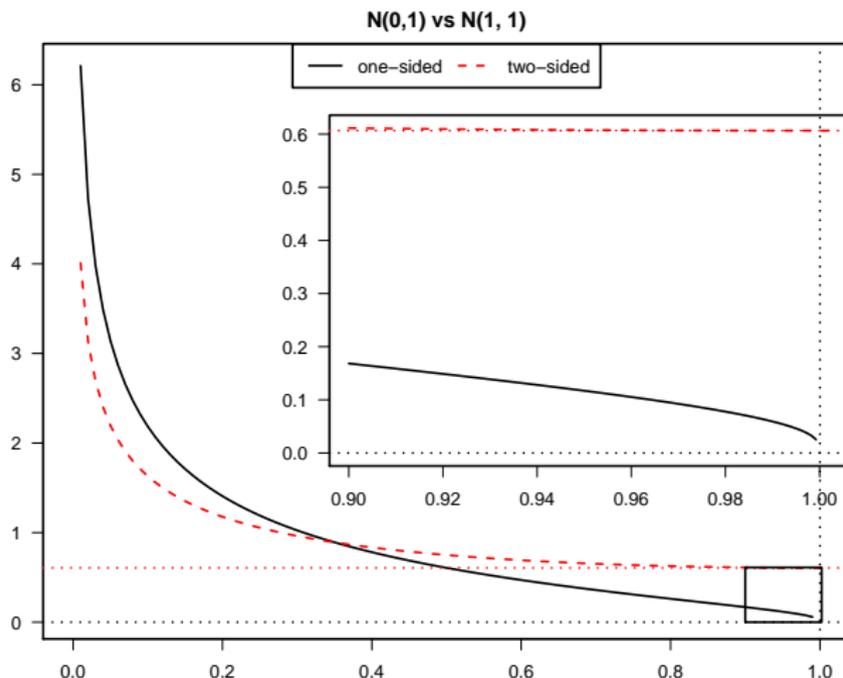
- Statistique de test suit $f_0 : t \mapsto \frac{1}{2\pi} e^{-t^2/2}$ sous \mathcal{H}_0 et $f_1 : t \mapsto f_0(t - \theta)$ sous \mathcal{H}_1
- Densité des p -values sous $\mathcal{H}_1 : g_1(x) = \exp\left(-\frac{\mu^2}{2} - \mu\Phi^{-1}(x)\right)$, où $\Phi(x) = P(\mathcal{N}(0, 1) \leq x)$

Propriétés

- $\lim_{x \rightarrow 0} 1/g_1(x) = 0$: pas de criticité
- $g_1(1) = 0$: π_0 est identifiable
- g_1 n'est pas même dérivable en 1 : convergence lente

Convergence lente même dans des cas simples

Localisation Gaussienne, tests unilatéraux



Localisation Gaussienne : densité sous \mathcal{H}_1 des p -values unilatérales et

Conclusions de l'étude asymptotique

- Le FDR étant l'espérance du FDP, la **concentration** du FDP est importante
- Le formalisme des fonctions de seuil permet d'obtenir la loi asymptotique du FDP d'un grand nombre de procédures de tests multiples :
 - ▶ La criticité dépend de la loi des p -valeurs
 - ▶ Le niveau de FDR effectivement atteint dépend de la procédure
 - ▶ La vitesse de convergence du FDP vers le FDR est $1/\sqrt{m}$ sous indépendance mais peut être bien plus lente en présence de corrélations
- La régularité de g en 1 conditionne la qualité d'estimation de π_0 , donc la possibilité d'un contrôle **exact** du FDR
- Dans les applications génomiques, cette régularité peut être mauvaise

Tests multiples en génomique

- Exemples d'applications
- Définition de critères de risque adaptés

6 Propriétés asymptotiques du FDP

- Cadre d'analyse
- Loi asymptotique du FDP et loi des p -valeurs en 0
- Loi asymptotique du FDP et loi des p -valeurs en 1

7 Interprétation des résultats d'analyse différentielle

- Enrichissement d'ensembles de gènes
- Tests multivariés

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)
⇒ **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

- Exemples d'applications
- Définition de critères de risque adaptés

6 Propriétés asymptotiques du FDP

- Cadre d'analyse
- Loi asymptotique du FDP et loi des p -valeurs en 0
- Loi asymptotique du FDP et loi des p -valeurs en 1

7 Interprétation des résultats d'analyse différentielle

- Enrichissement d'ensembles de gènes
- Tests multivariés

Principales étapes d'une analyse d'enrichissement

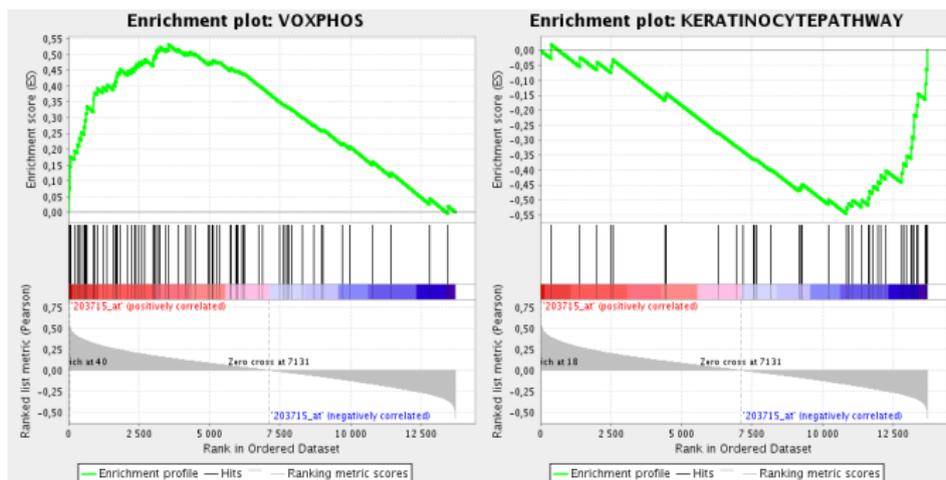
- 1 Mesure d'expression différentielle
- 2 Score d'enrichissement
- 3 Hypothèse testée
- 4 Estimation de la significativité
- 5 Correction des tests multiples

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



Tests multiples en génomique

- Exemples d'applications
- Définition de critères de risque adaptés

6 Propriétés asymptotiques du FDP

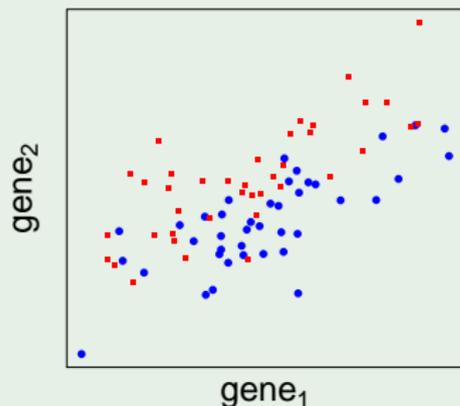
- Cadre d'analyse
- Loi asymptotique du FDP et loi des p -valeurs en 0
- Loi asymptotique du FDP et loi des p -valeurs en 1

7 Interprétation des résultats d'analyse différentielle

- Enrichissement d'ensembles de gènes
- Tests multivariés

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

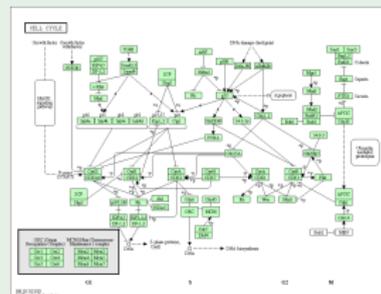
Test multivarié sur un ensemble de gènes E

- 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.

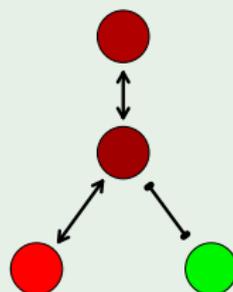


Profils d'expression : des fonctions sur des graphes

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),
- Différence 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} :

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

- 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)

$$f = 1.45 \underbrace{\text{graph}}_{\tilde{f}_1} u_1 - 0.04 \underbrace{\text{graph}}_{\tilde{f}_2} u_2 - 0.21 \underbrace{\text{graph}}_{\tilde{f}_3} u_3 + 0.20 \underbrace{\text{graph}}_{\tilde{f}_4} u_4 .$$

Exemple 2 (shift irrégulier)

$$f = 0.00 \text{ graph} - 0.41 \text{ graph} + 0.42 \text{ graph} + 1.16 \text{ graph} .$$

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}_k^2 dans l'espace de Fourier restreint aux k premiers coefficients

Intuition

Le shift en distribution est cohérent avec un graphe connu



Concentré dans les basses fréquences



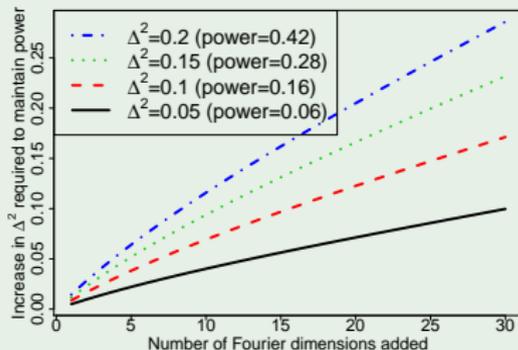
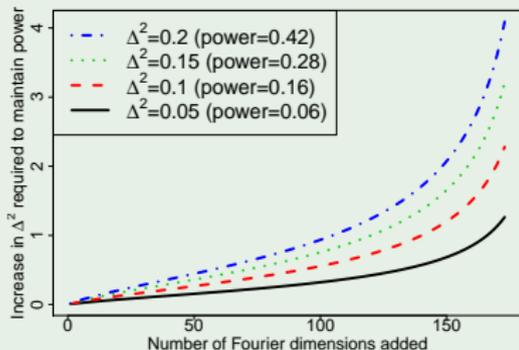
Le test dans les premiers coefficients de Fourier est plus puissant

Gain en puissance

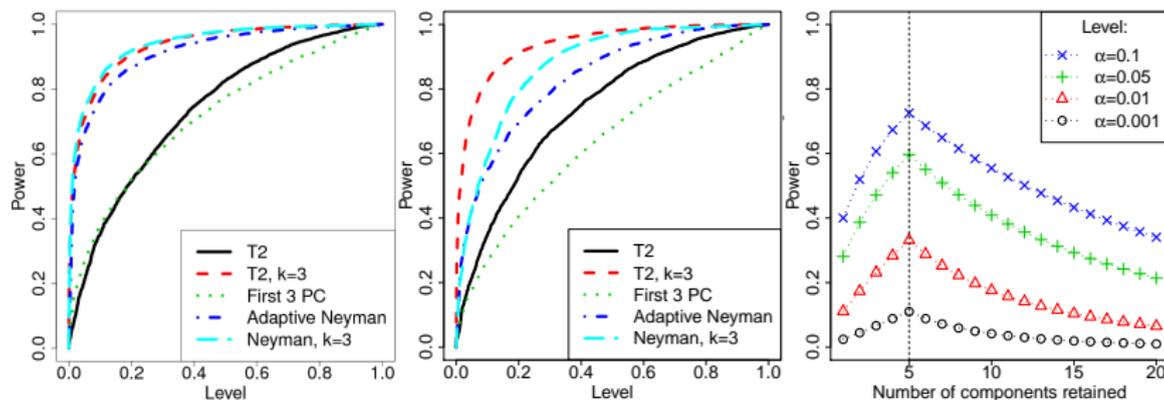
Lemma

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

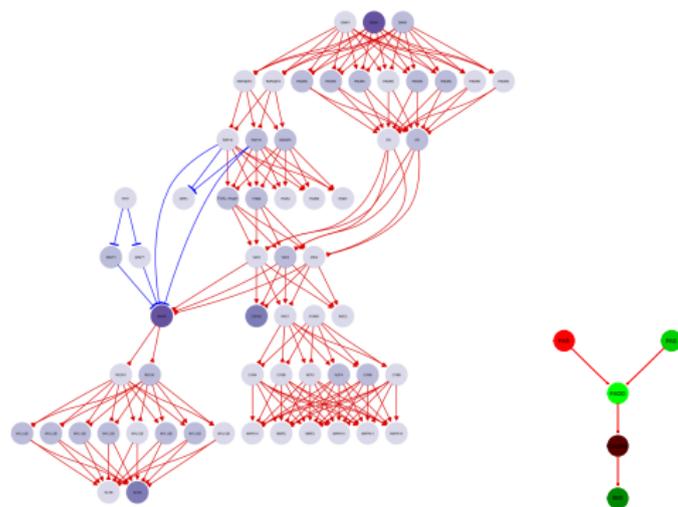


Gain en puissance : données simulées



- 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 premiers 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