Statistical methods for genomic data analysis

DNA copy number analyses

Pierre Neuvial

http://neuvial.ensae.net

Laboratoire Statistique et Génome
Université d'Evry-Val d'Essonne, UMR CNRS 8071 - USC INRA

Centrale Paris — 2011/2012

DNA copy number analyses

1. Genotyping microarrays in cancer studies
   - DNA copy number changes in cancers
   - Genotyping microarray data
2. Extracting biological information
   - Pre-processing : making signals comparable across samples
   - Post-processing : total copy numbers
   - Post-processing : allelic ratios
3. Segmentation of DNA copy number profiles
   - The need for breakpoint detection methods
   - Existing approaches : examples
   - Multi-sample or cross-platform segmentation
4. Estimating DNA copy numbers
   - Joint use of C and DH for detection
   - Calling : influence of tumor purity, ploidy, and signal saturation

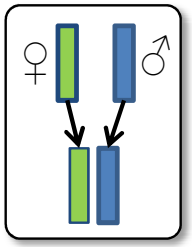
References


P. Neuvial, H. Bengtsson et T. P. Speed (2011).
Statistical analysis of single nucleotide polymorphism microarrays in cancer studies. In H. H.-S. Lu, B. Schölkopf, and Z. Hongyu, editors,
Handbook of Statistical Bioinformatics, Springer Handbooks of Computational Statistics. Springer, 1st edition, 2011.

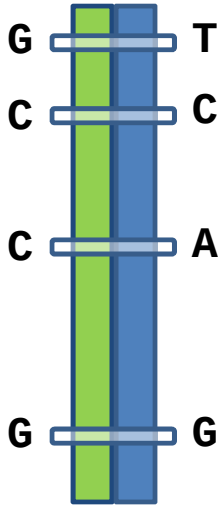

N. R. Zhang (2010)
DNA copy number profiling in normal and tumor genomes. In J. Feng, W. Fu, and F. Sun, editors,
Frontiers in Computational and Systems Biology, pages 259–281. Springer-Verlag.

DNA copy number analyses

Genotypes in a diploid chromosome



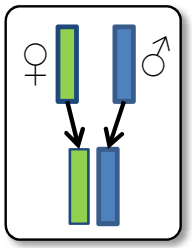
Single nucleotide polymorphism



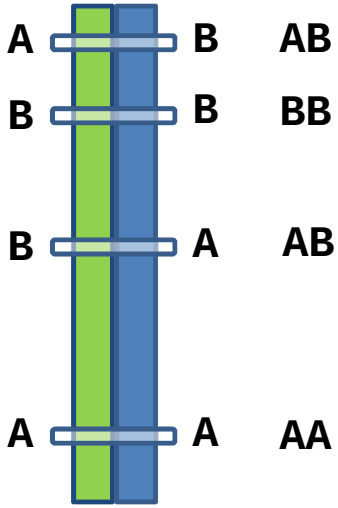
10-20 million
known SNPs

slide: H. Bengtsson.

Genotypes in a diploid chromosome



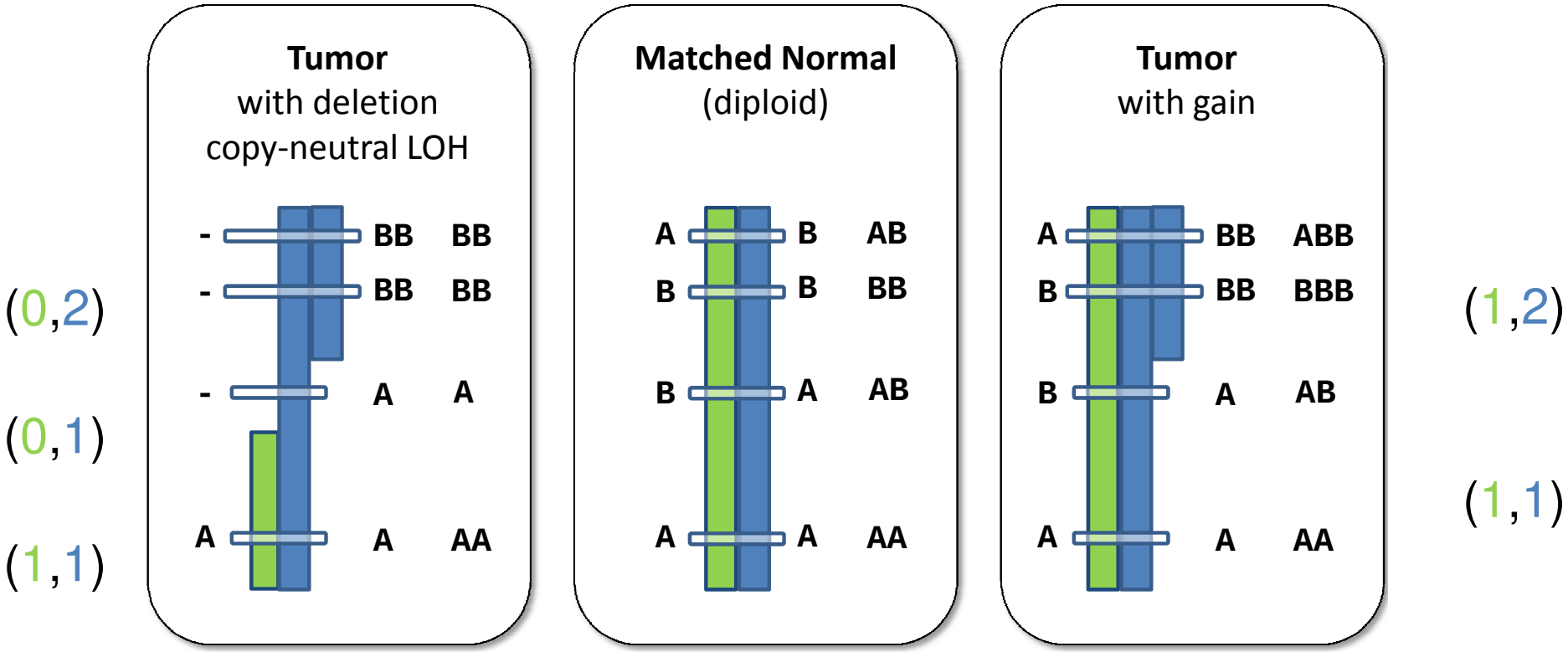
Single nucleotide polymorphism



10-20 million
known SNPs

slide: H. Bengtsson.

Genotypes and copy numbers in a tumor



slide: H. Bengtsson.

Parental, minor and major copy numbers

Parental copy numbers at genomic locus j : (m_j, p_j), the **unobserved** number of maternal and paternal chromosomes at j.

Copy number state at genomic locus j

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

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

Minor (C_1) and major (C_2) copy numbers :

- characterize the above CN events in cancers
- can be estimated from SNP arrays

Parental copy numbers (CN)

The number of copies of each parental chromosome.
Notation : $CN = (C_1, C_2)$, with $C_1 \leq C_2$.
In a region of no genomic alteration : $CN = (1, 1)$

Genotyping microarrays quantify

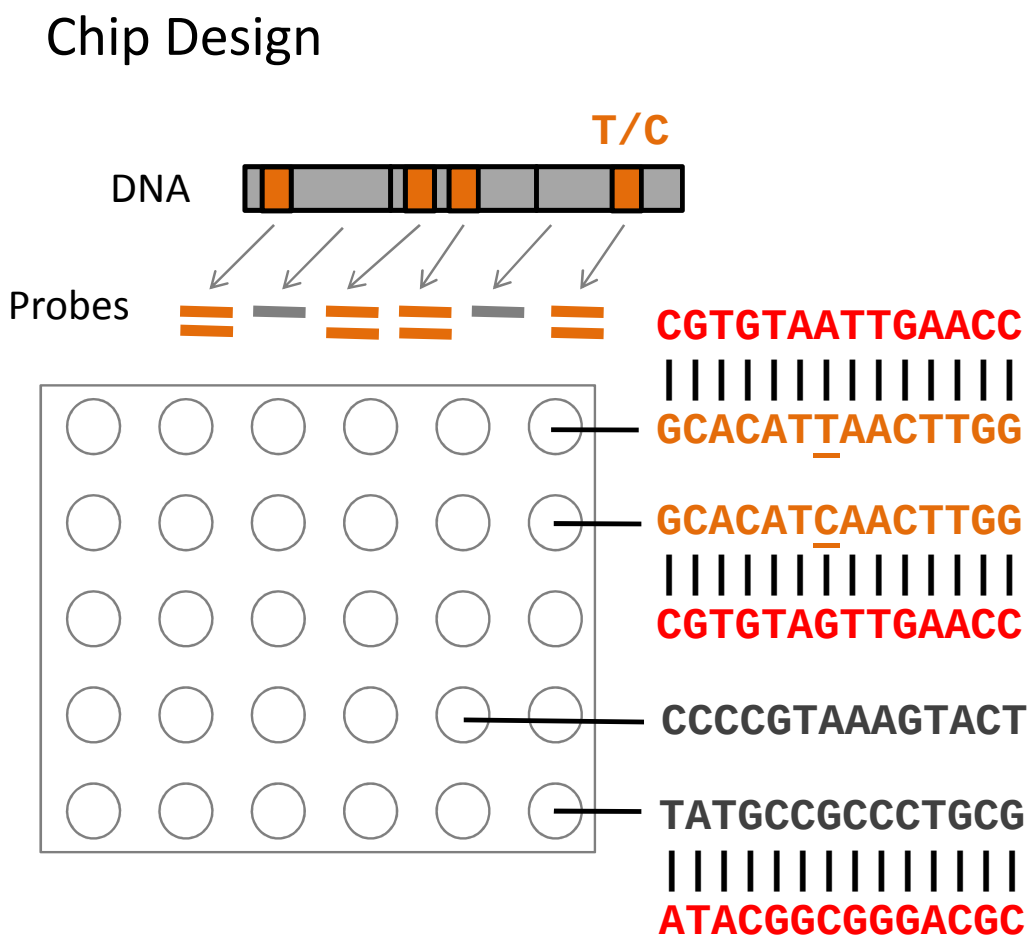
1. total copy number : $TCN = C_1 + C_2$
2. alleleic composition, which is related to $\frac{C_1}{C_1 + C_2}$

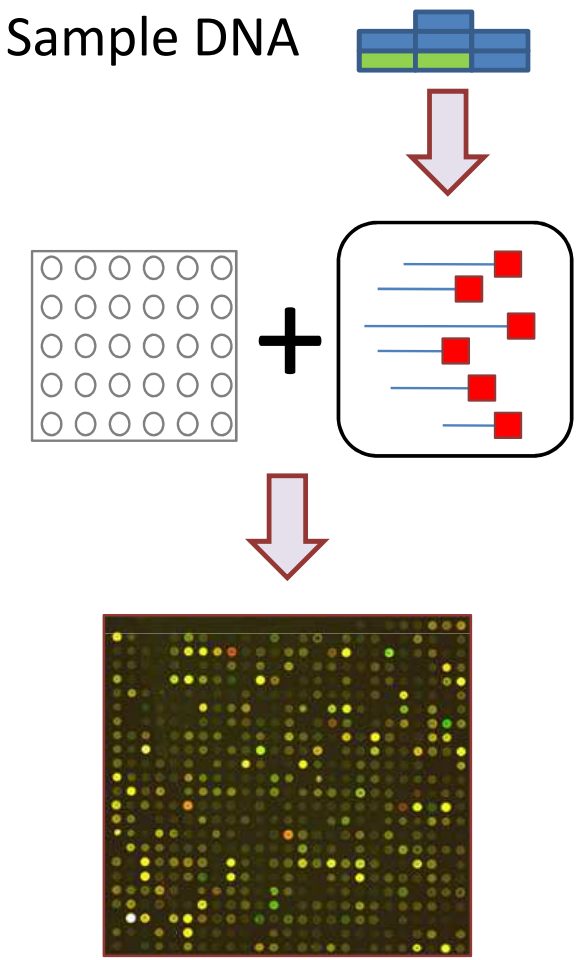
Both quantities are needed to understand what is happening :

- Copy neutral LOH : $CN = (0, 2)$
- Balanced duplication : $CN = (2, 2)$

DNA copy number analyses

Copy number and genotyping microarrays





slide: H. Bengtsson.

(C_1, C_2) can be estimated from SNP arrays

For SNP j in sample i, observed signal intensities can be summarized as (θ, β), where $\theta_{ij} = \theta_{Aij} + \theta_{ijB}$ and $\beta_{ij} = \theta_{ijB}/\theta_{ij}$.

Total copy numbers

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

Decrease in heterozygosity

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

Notes :

- *DH* only defined for SNPs that were **heterozygous in the germline**
- Both dimensions are needed to understand what is going on :
  - Copy neutral LOH : $CN = (0, 2)$, normal total copy number
  - Balanced duplication : $CN = (2, 2)$, allelic balance

The Cancer Genome Atlas (TCGA)

"Accelerate our understanding of the molecular basis of cancer"

- 20 tumor types : brain (glioblastoma multiforme), ovarian, breast, lung, leukemia (AML)...
- Large studies : 500 tumor-normal pairs for each tumor type
- Data levels : DNA copy number, gene expression, DNA methylation
- Platforms : microarray and sequencing

For SNP arrays : identify **copy number changes** : (C, DH) or (C_1, C_2) :

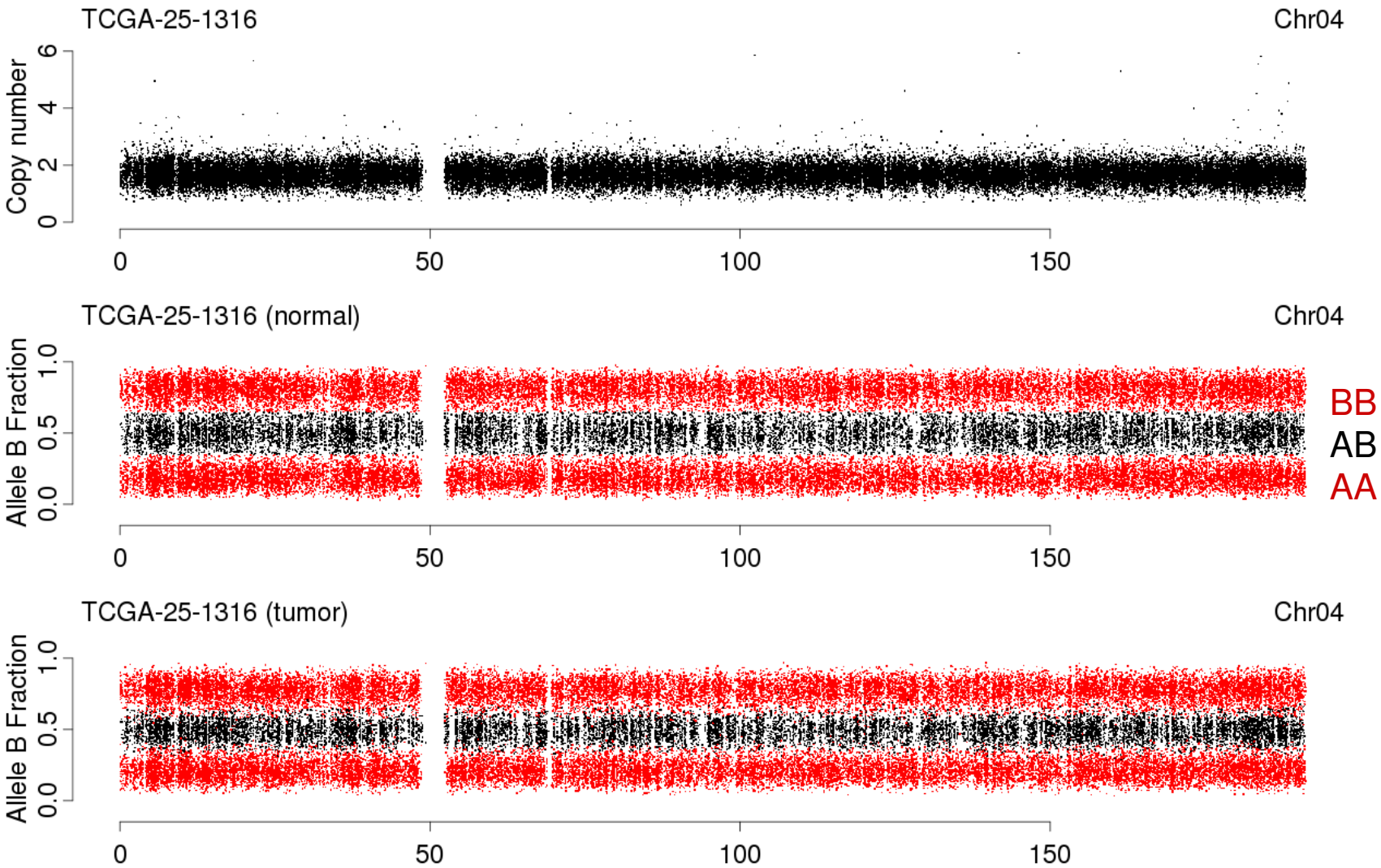
1. **detection** : finding regions
2. **classification** labeling regions

Data shown in this presentation : high-grade serous ovarian adenocarcinoma (OvCa).

DNA copy number analyses

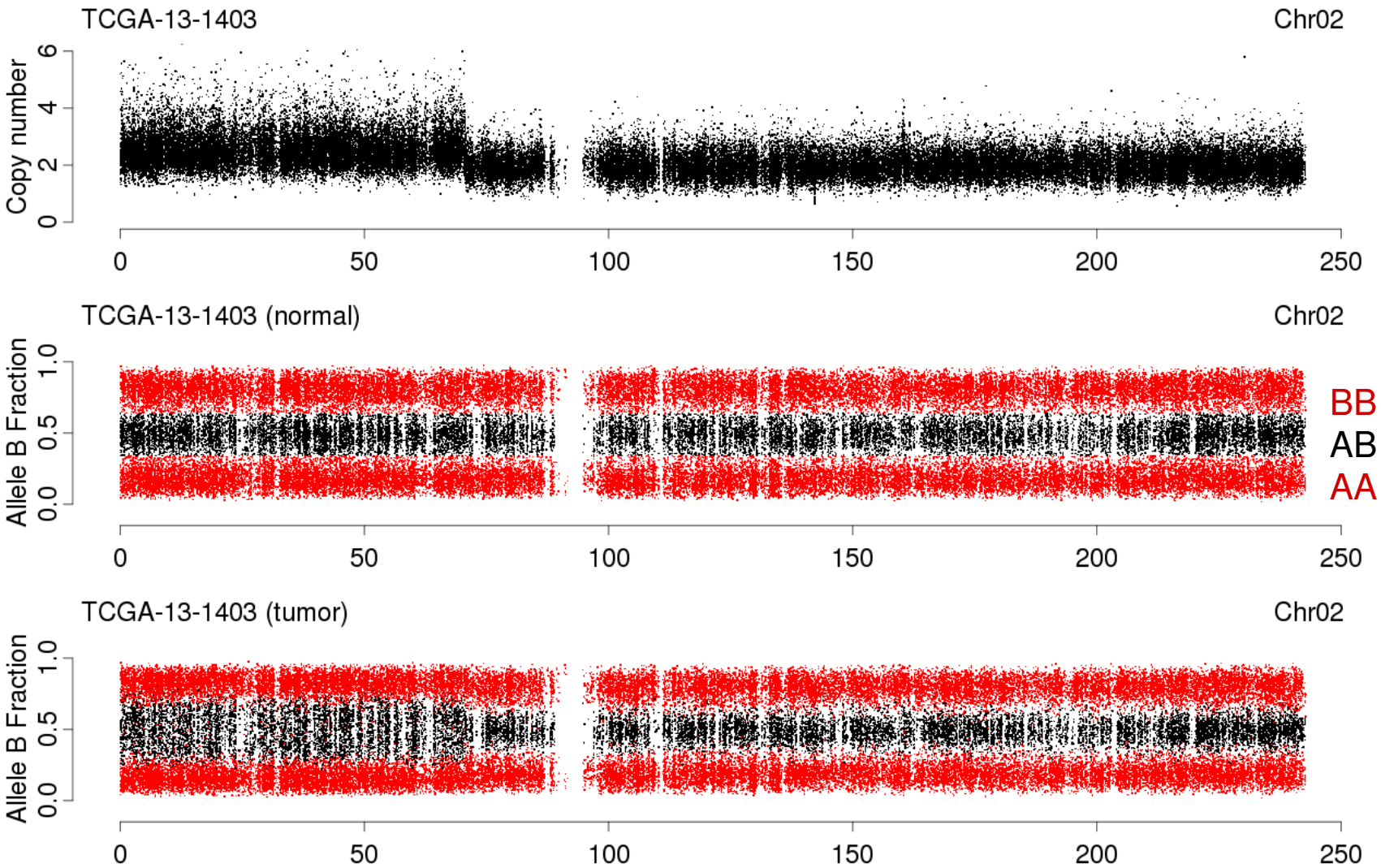
1. Genotyping microarrays in cancer studies
   - DNA copy number changes in cancers
   - Genotyping microarray data
2. Extracting biological information
   - Pre-processing : making signals comparable across samples
   - Post-processing : total copy numbers
   - Post-processing : allelic ratios
3. Segmentation of DNA copy number profiles
   - The need for breakpoint detection methods
   - Existing approaches : examples
   - Multi-sample or cross-platform segmentation
4. Estimating DNA copy numbers
   - Joint use of C and DH for detection
   - Calling : influence of tumor purity, ploidy, and signal saturation

No copy number change : (1,1)



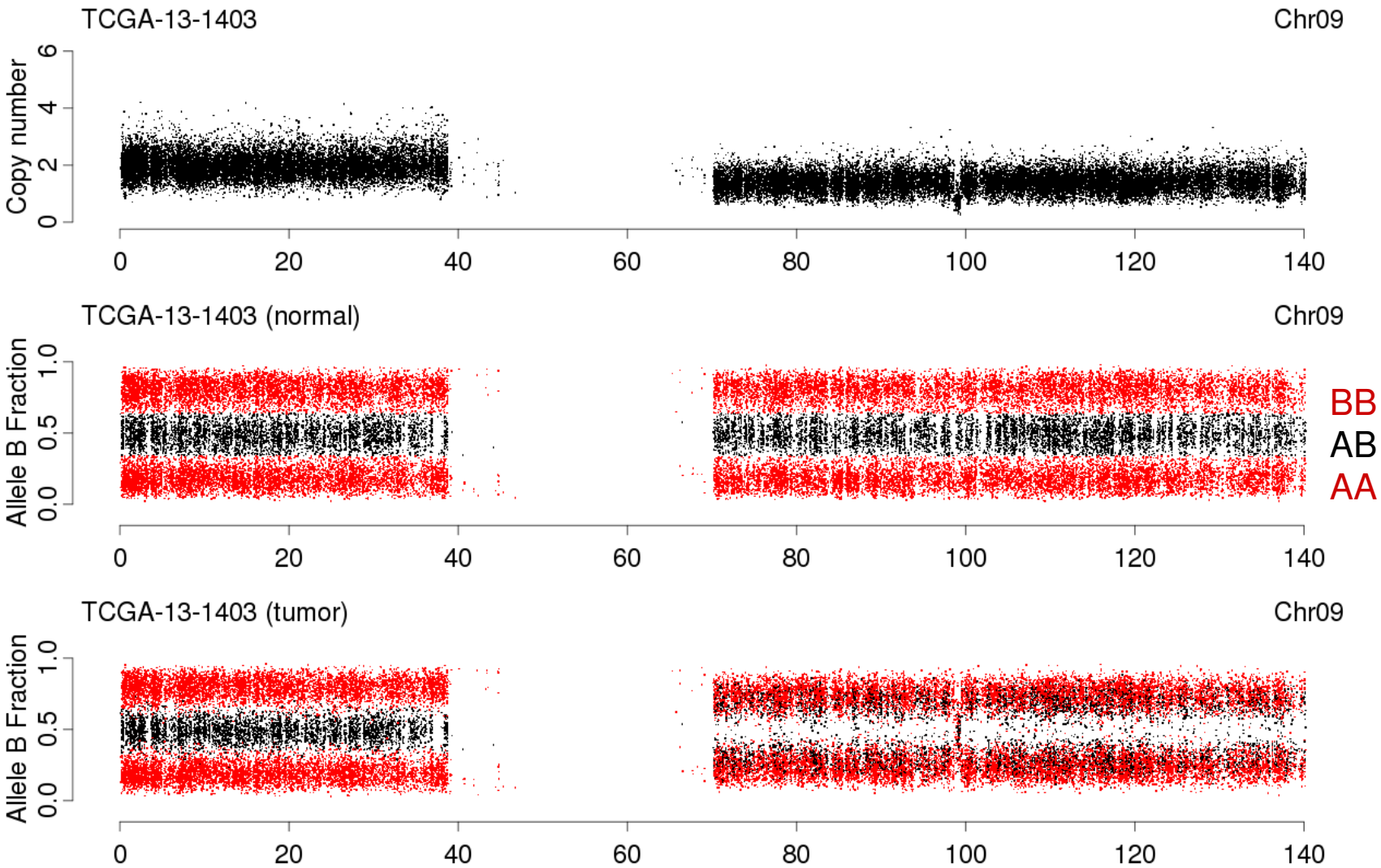
Homozygous SNPs in the normal sample are highlighted in red.

Gain : (1, 2)



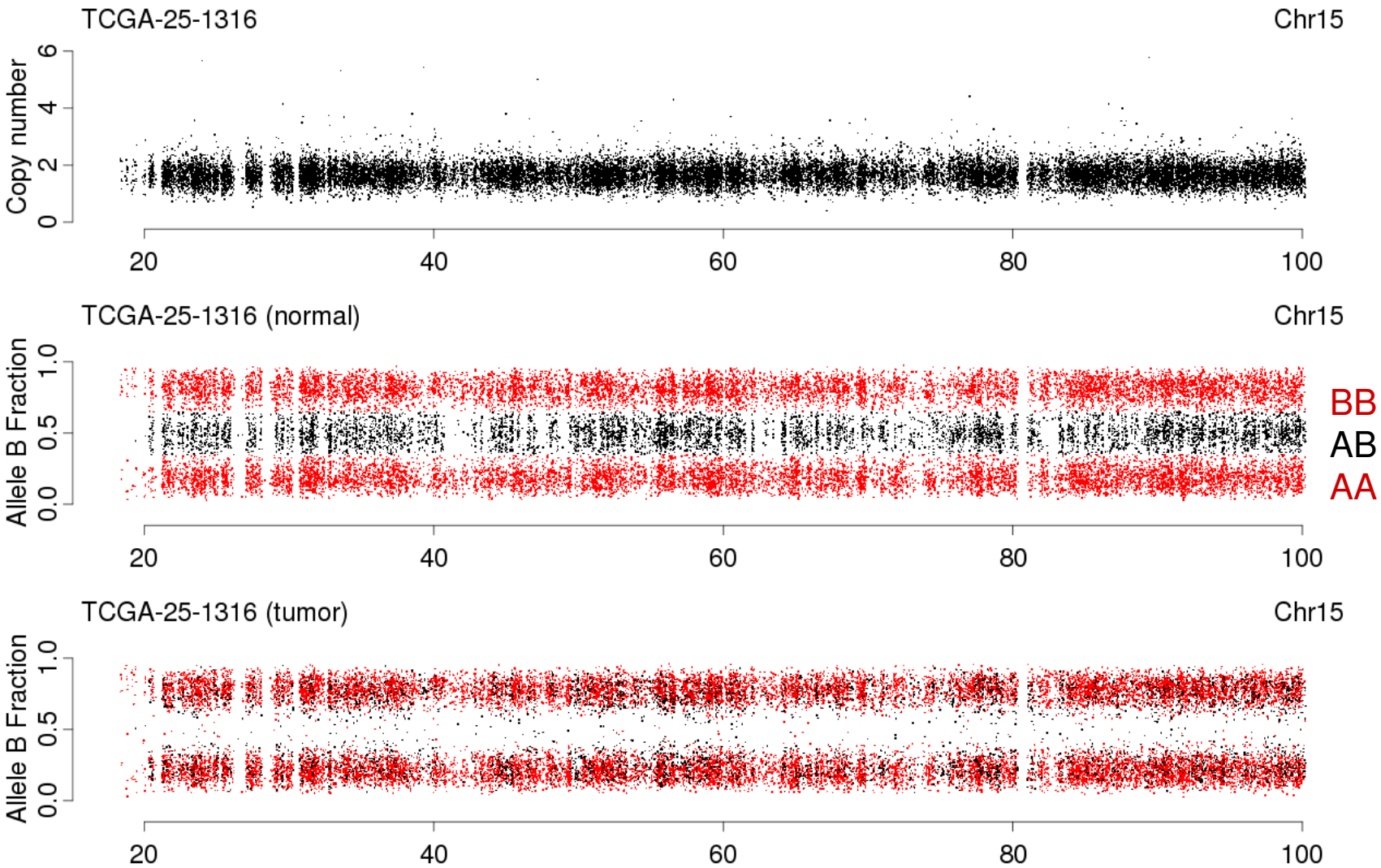
Homozygous SNPs in the normal sample are highlighted in red.

Deletion : (0, 1)



Homozygous SNPs in the normal sample are highlighted in red.

Copy number neutral LOH : (0, 2)



Homozygous SNPs in the normal sample are highlighted in red.

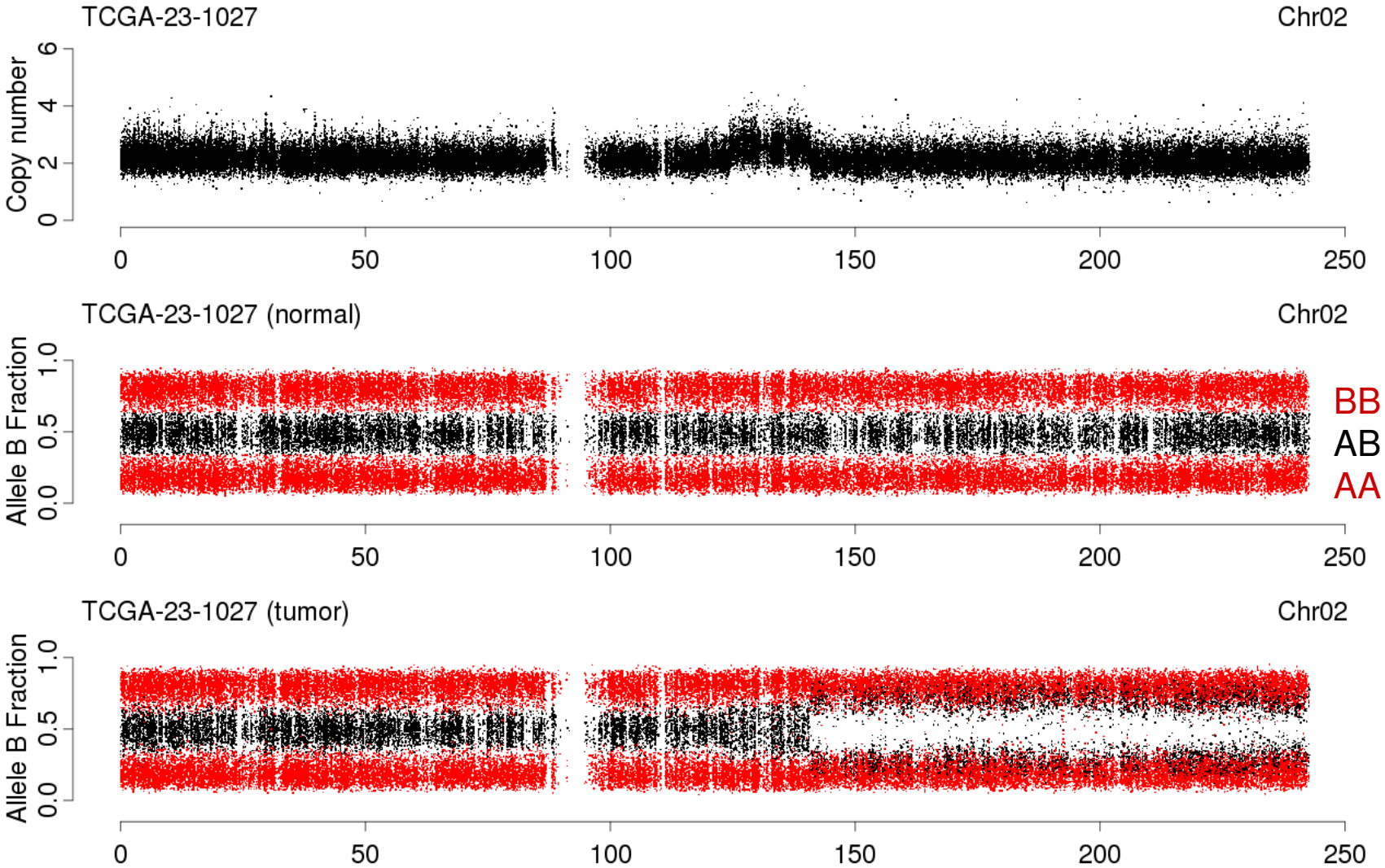
Tumor purity/normal contamination

In practice what we call tumor samples are actually **a mixture of tumor and normal cells.**

The ones just shown have the largest fraction of tumor cells in the data set.

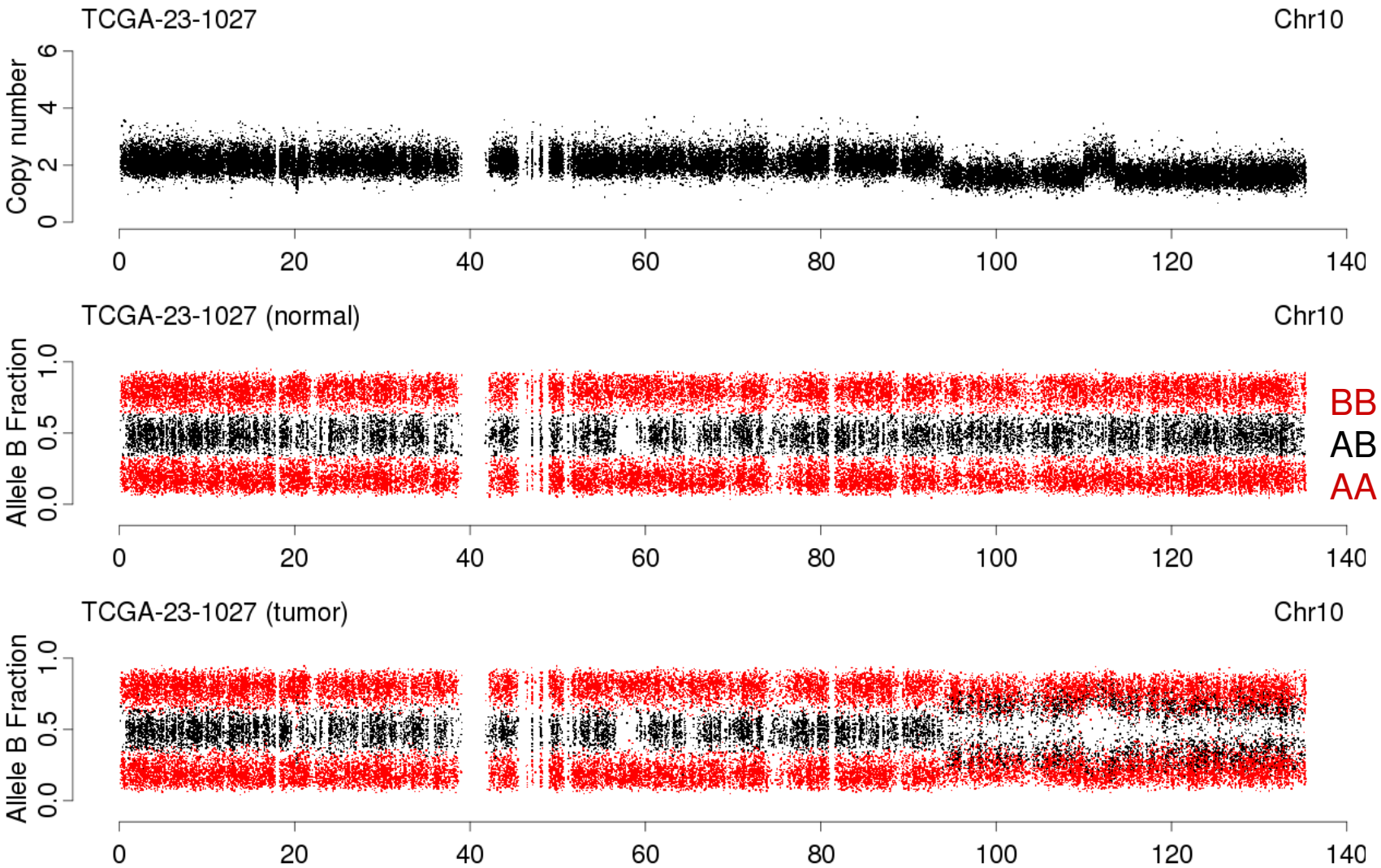
In presence of normal contamination allele B fractions for heterozygous SNPs are **shrunk toward 1/2**.

Normal, gain, copy neutral LOH



Homozygous SNPs in the normal sample are highlighted in red.

Normal, deletion, copy neutral LOH



Homozygous SNPs in the normal sample are highlighted in red.

Normalization : definition and objective

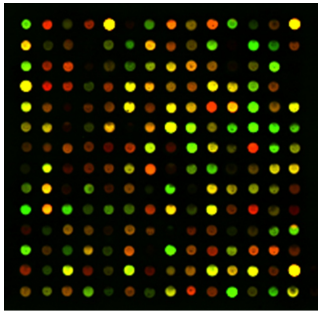
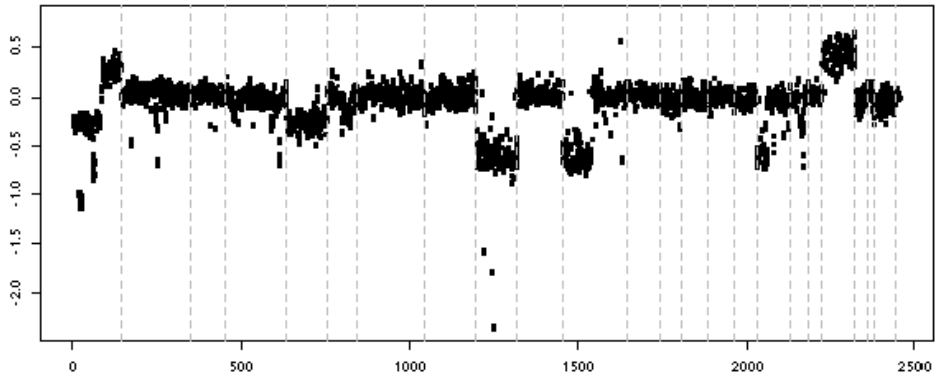


Image analysis

$\Rightarrow$

Normalization



Motivation : substantial experimental variability

- lack of reproducibility between experiments
- each step of a microarray contains potential sources of bias

Goal : increasing the signal to noise ratio

- differentiate biological variability from experimental artifacts
- making data coming from different experiments comparable

DNA copy number analyses

Copy-numbers by Robust Microarray Analysis (CRMA)

A single sample preprocessing method

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

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

slide: H. Bengtsson.

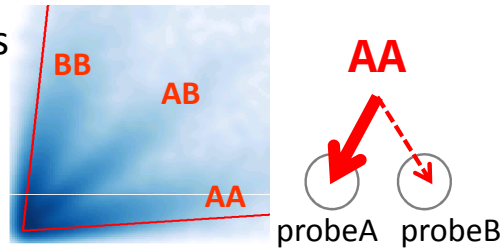
Explanation of systematic variation across arrays

Scanner offset and cross-hybridization between alleles

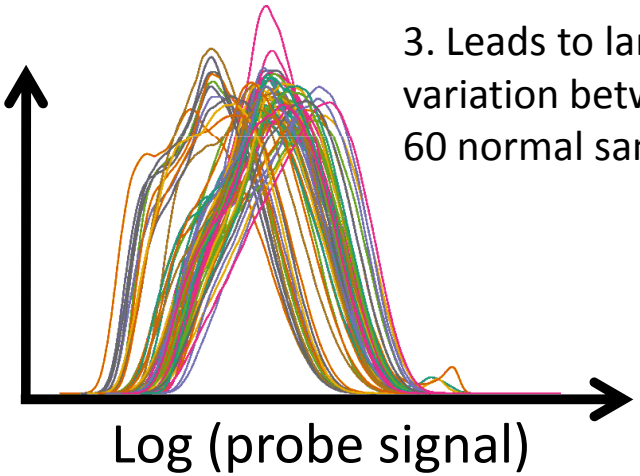
1. The scanner's shifts all probe signals (offset)

+

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

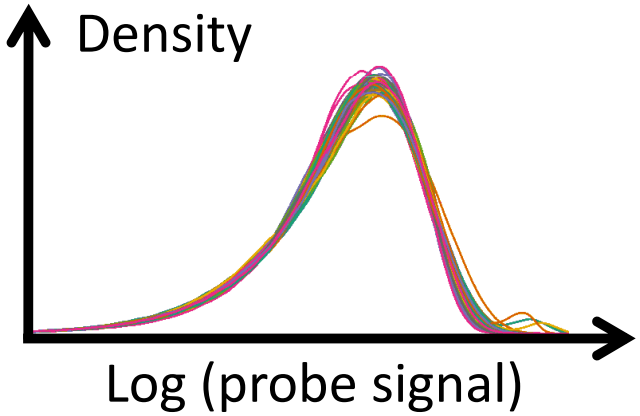


=



3. Leads to large variation between 60 normal samples

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



slide: H. Bengtsson.

ROC evaluation

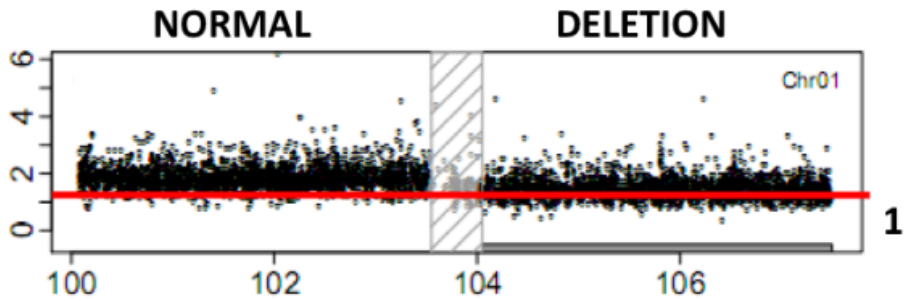
For a given sample :

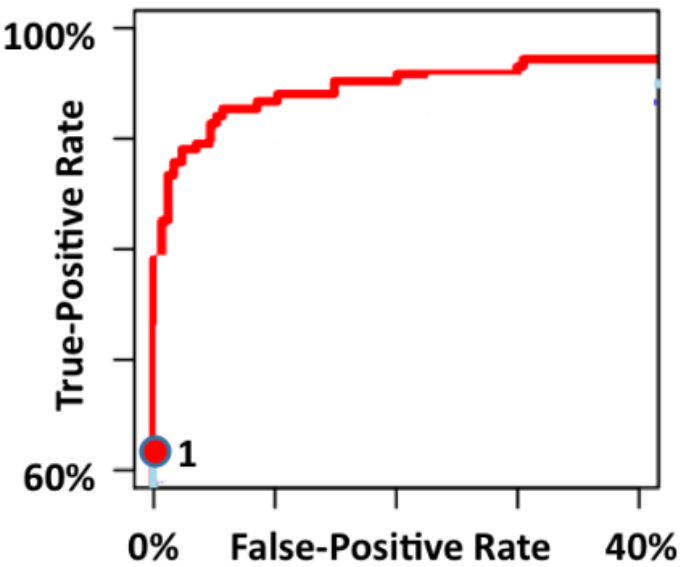
- find a clear change point
- label flanking regions, e.g. NORMAL (1,1) and DELETION (0,1)
- choose one reference state and one state to call

For each value of a threshold τ :

- Call SNPs below τ a DELETION
- Count number of true and false DELETIONs.

ROC curve is built by adjusting τ.





ROC evaluation

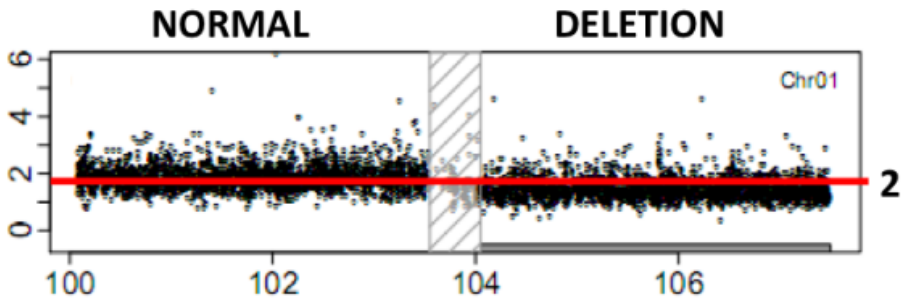
For a given sample :

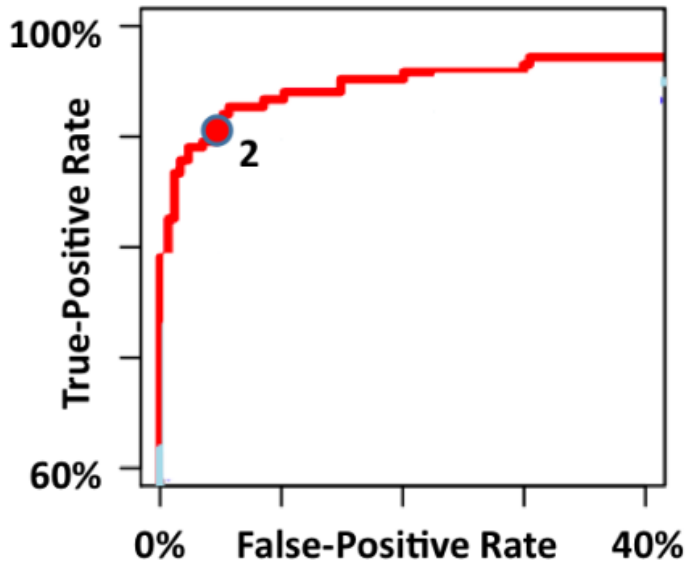
- find a clear change point
- label flanking regions, e.g. NORMAL (1,1) and DELETION (0,1)
- choose one reference state and one state to call

For each value of a threshold τ :

- Call SNPs below τ a DELETION
- Count number of true and false DELETIONs.

ROC curve is built by adjusting τ.





ROC evaluation

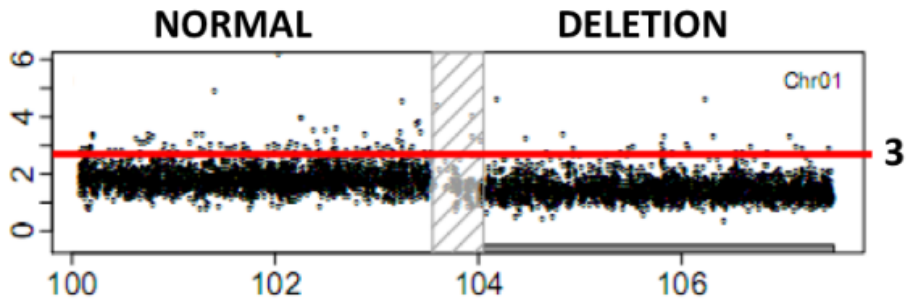
For a given sample :

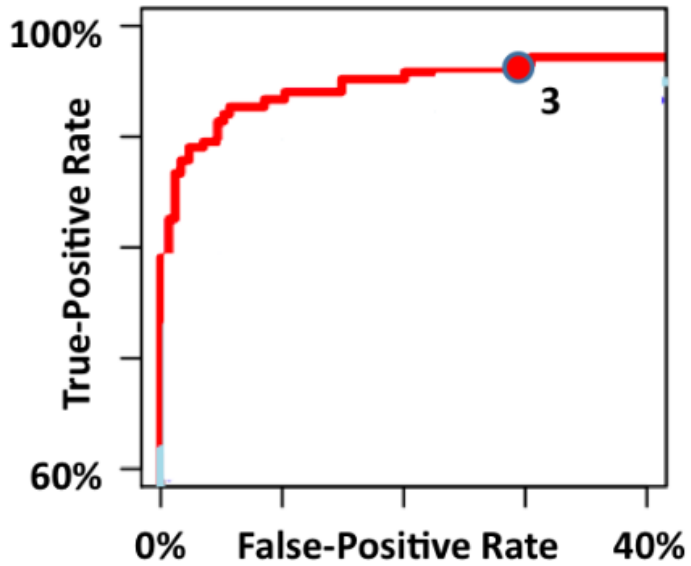
- find a clear change point
- label flanking regions, e.g. NORMAL (1,1) and DELETION (0,1)
- choose one reference state and one state to call

For each value of a threshold τ :

- Call SNPs below τ a DELETION
- Count number of true and false DELETIONs.

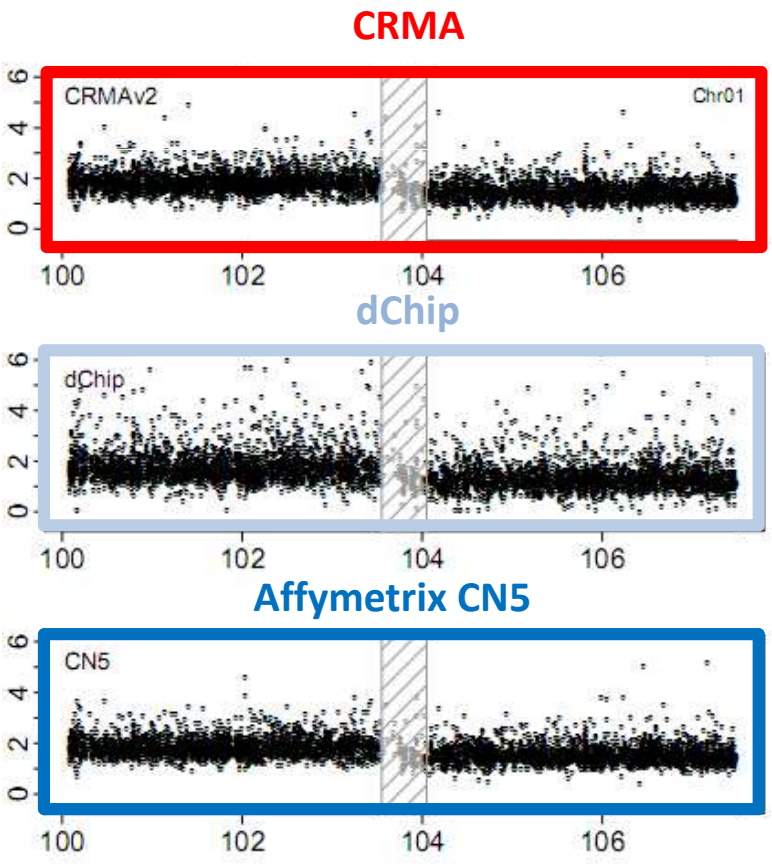
ROC curve is built by adjusting τ.

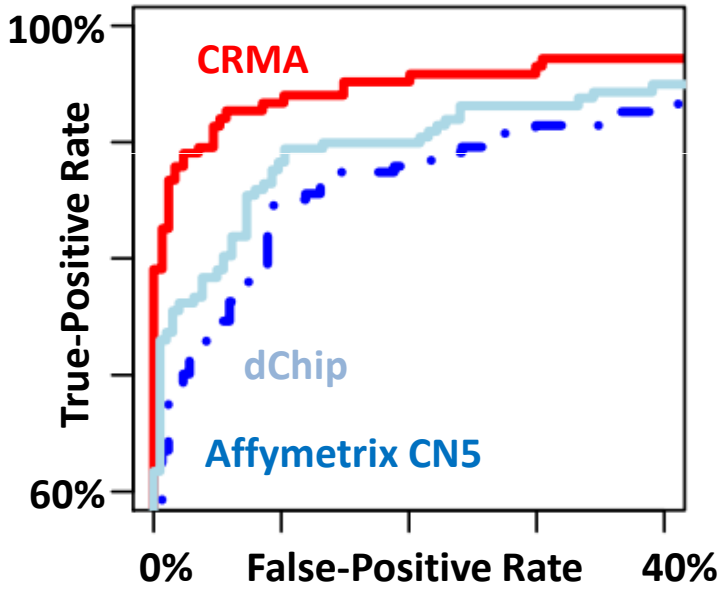




CRMA does better than multi-array methods

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





Data set:

- Tumor-normal pairs (HCC1143).
- 68 hybridizations, Affymetrix 6.0

Preprocessing:

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

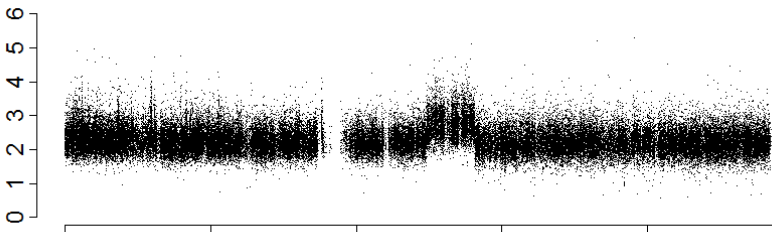
slide: H. Bengtsson.

Motivation : signal to noise ratio along the genome

For SNP j in sample i, observed signals are summarized by (θ, β), where $\theta_{ij} = \theta_{Aij} + \theta_{ijB}$ and $\beta_{ij} = \theta_{ijB}/\theta_{ij}$.

Total copy number

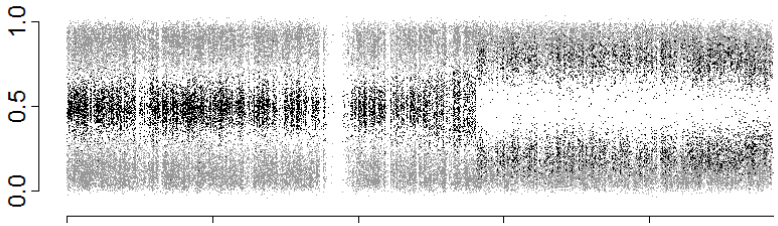
$$\begin{aligned} C_{ij} &= 2\frac{\theta_{ij}}{\theta_{Rj}} \\ &= C_{1ij} + C_{2ij} \end{aligned}$$



Choice of reference R ?

Decrease in heterozygosity

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



Low signal to noise ratio

DNA copy number analyses

Choosing a reference

Example : breast cancer cell lines.

36 experiments, in 3 batches.

Possible choices of a reference for a given experiment

1. 192 “normal” samples from another lab
2. the set of 36 samples from the same lab
3. the experiments of the same batch

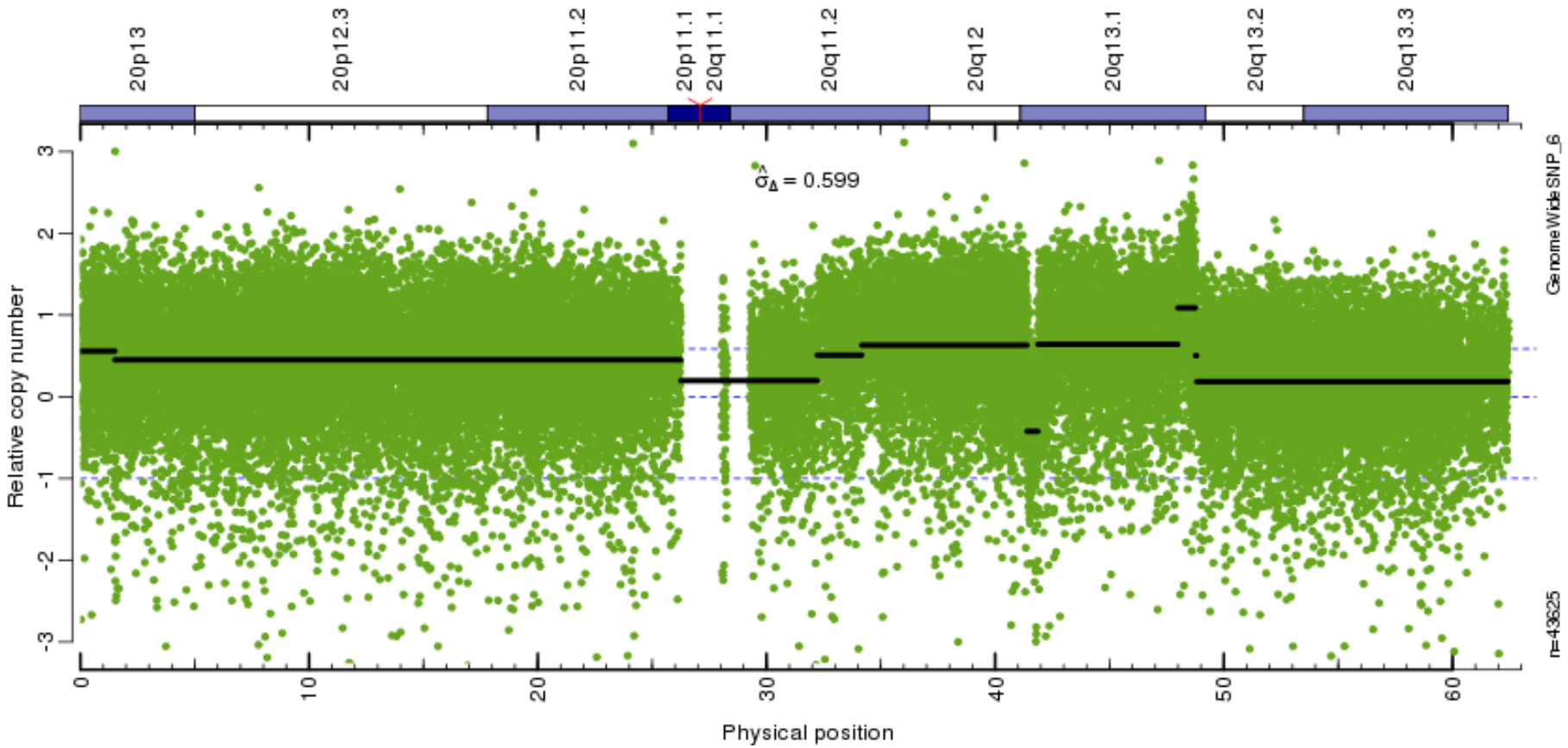
Quantification of the noise level :

$$\widehat{\sigma}_{\Delta} = \frac{1}{\sqrt{2}} \cdot \Phi^{-1}(3/4) \cdot \underset{j}{\text{median}} \left(\left| z_j - \underset{j'}{\text{median}}(z_{j'}) \right| \right)$$

where z_j are first order differences between DNA copy numbers

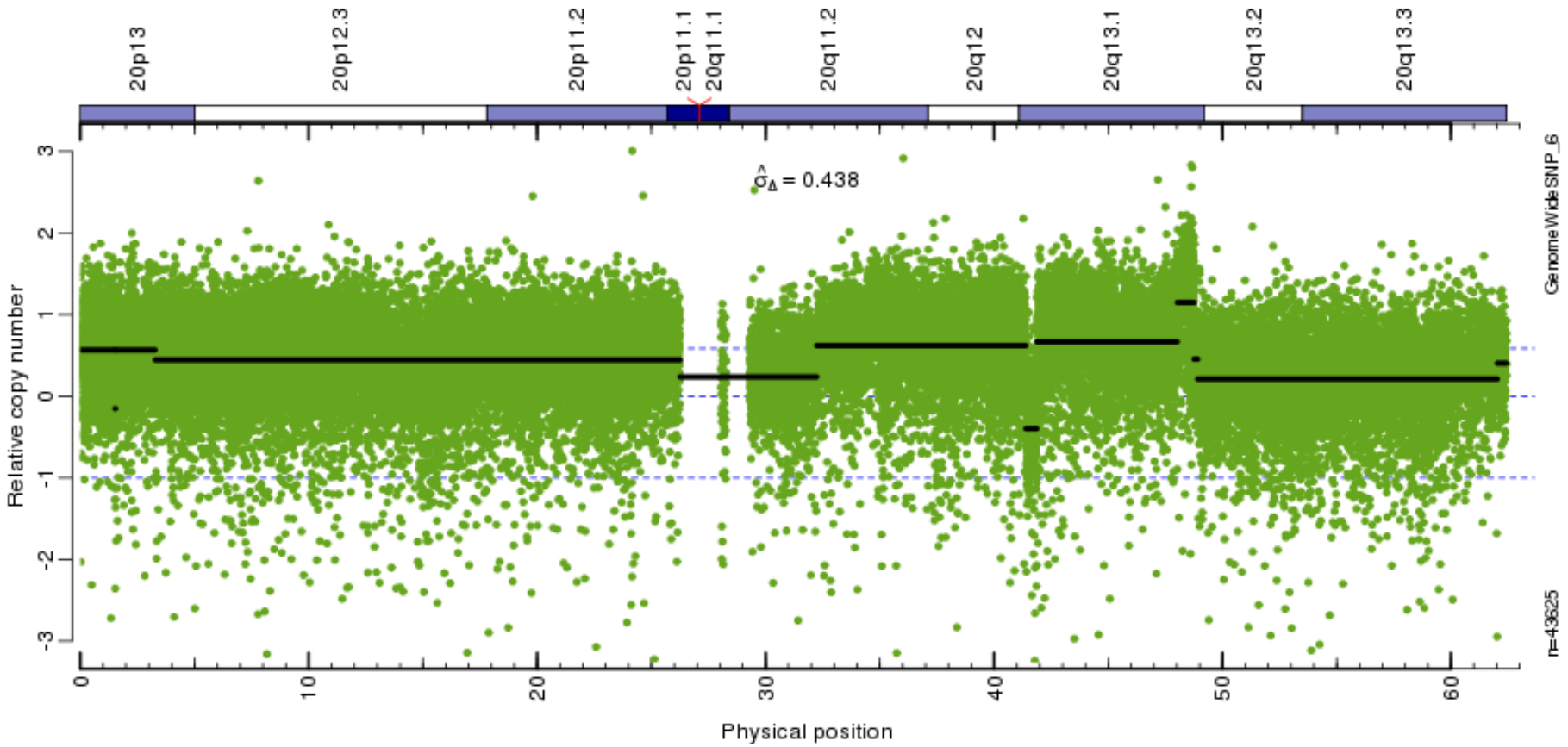
Choosing a reference

Different lab (n=192)



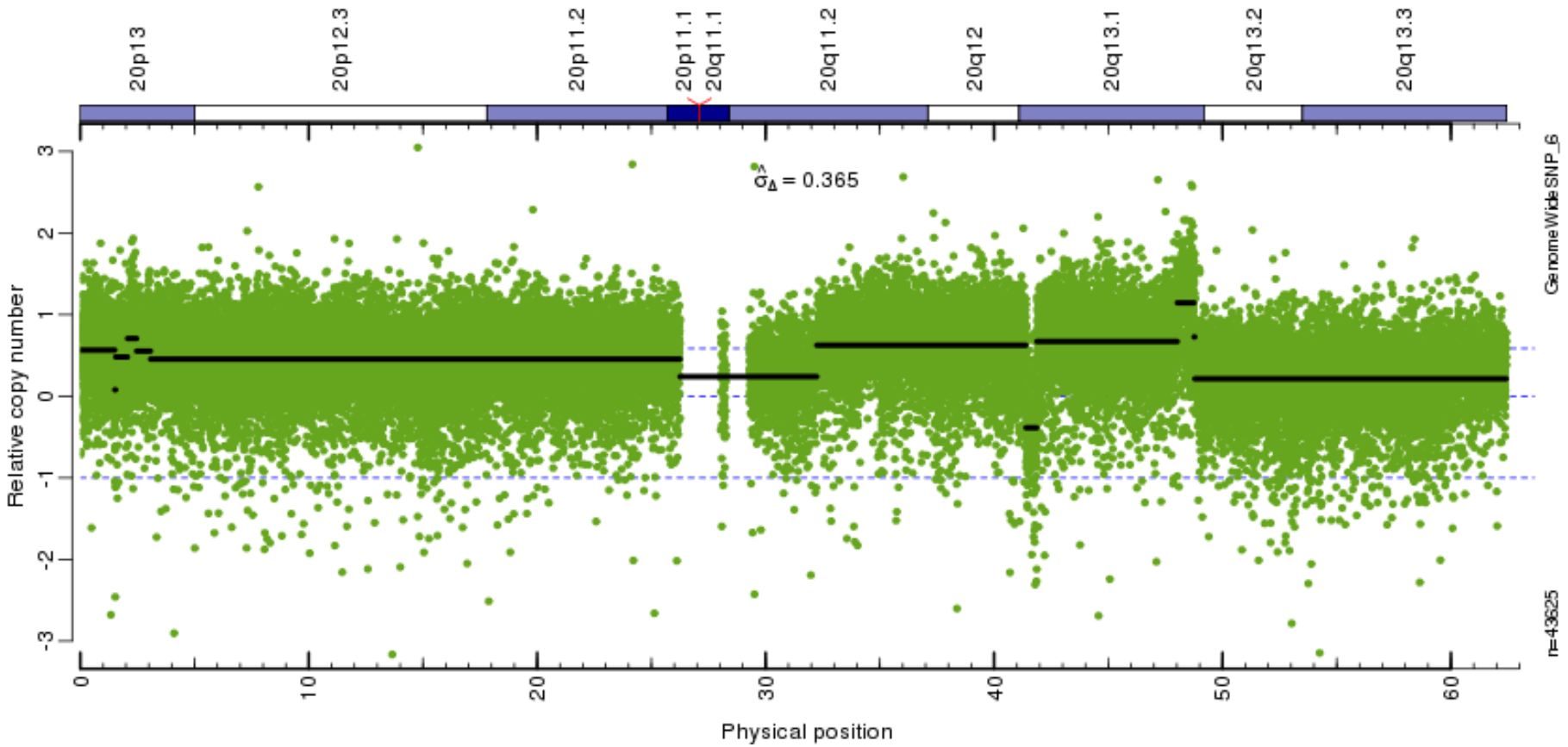
Choosing a reference

Same lab, all batches (n=36)



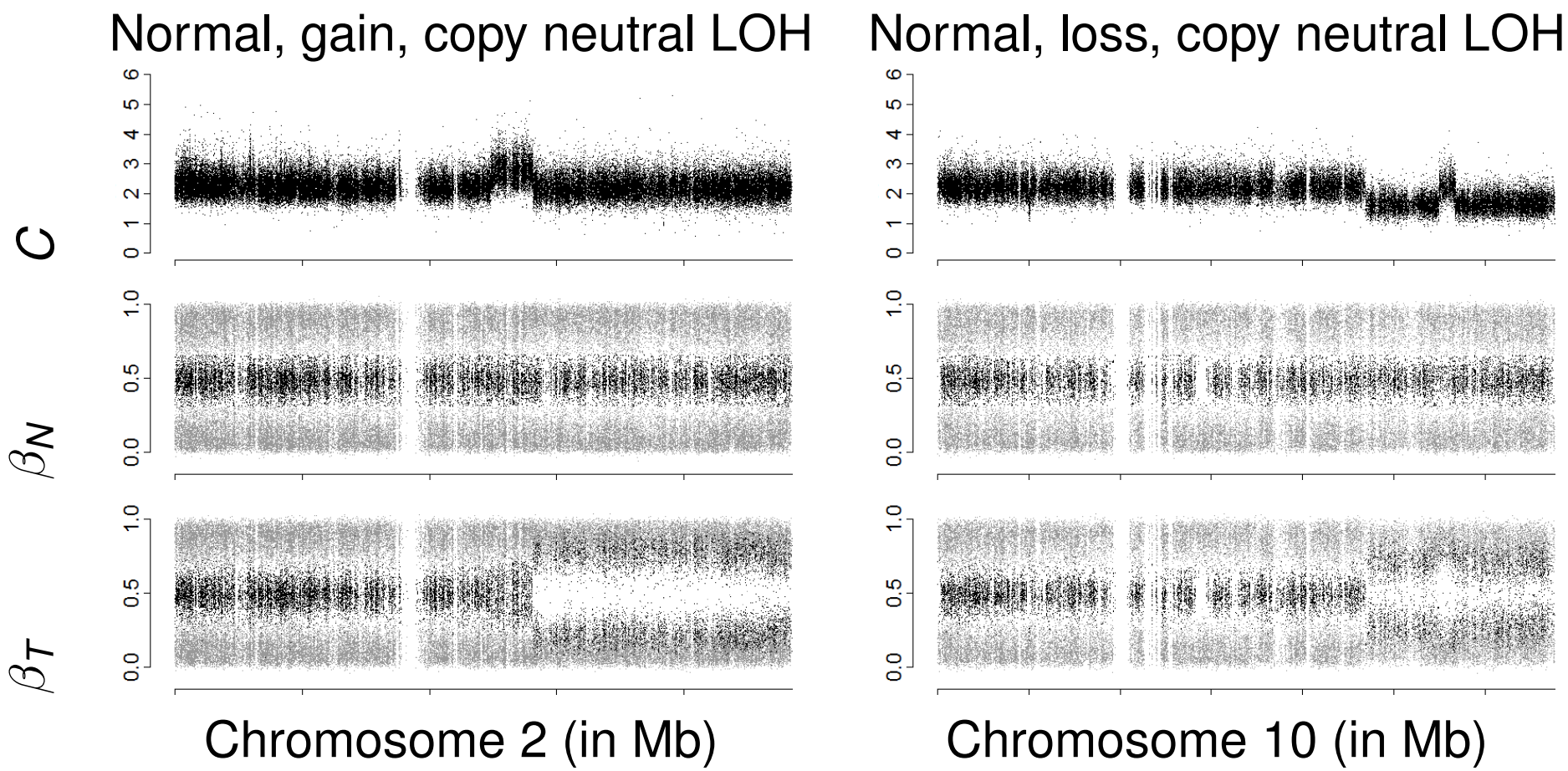
Choosing a reference

Same lab, same batch (n=22)

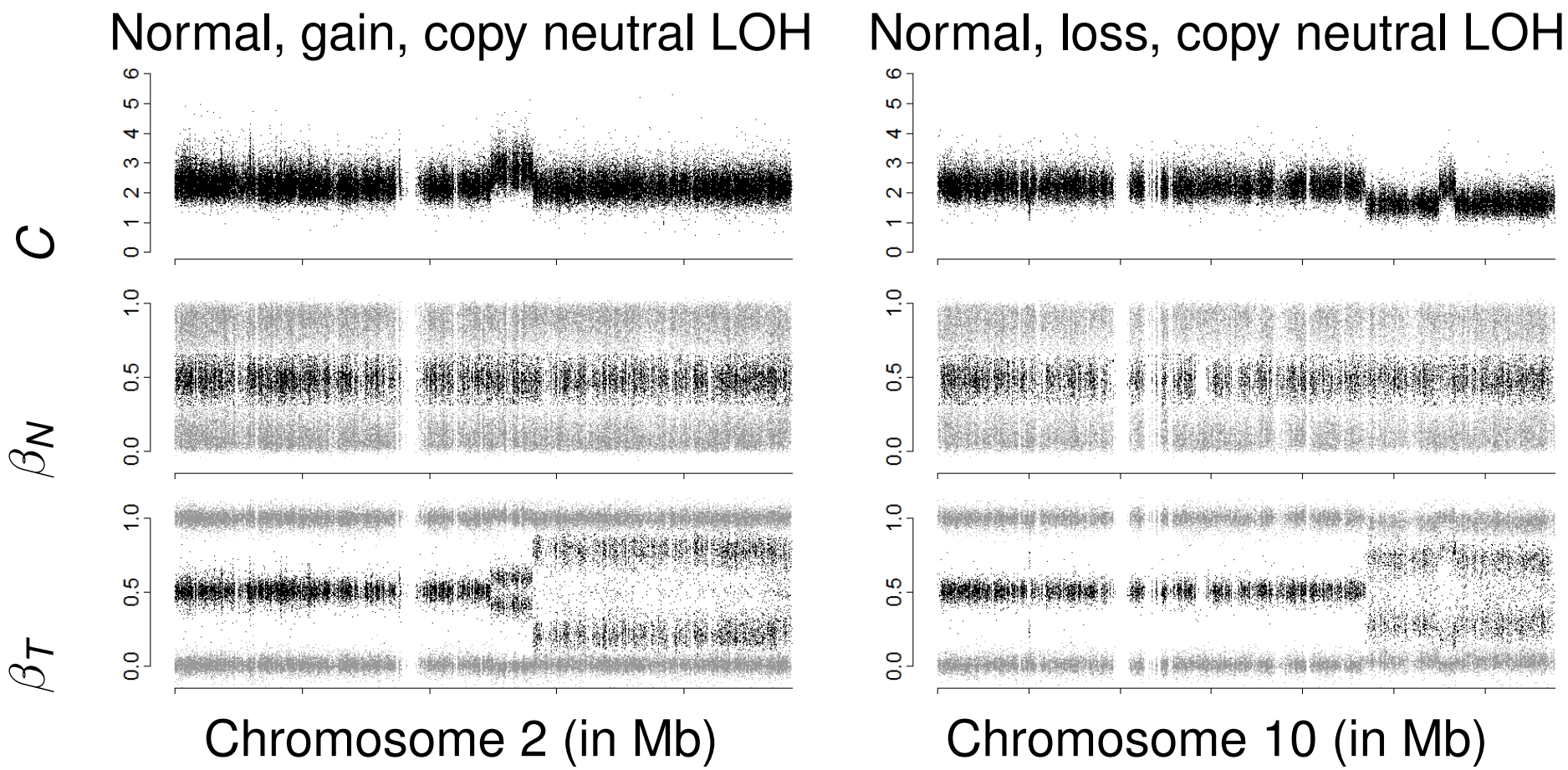


DNA copy number analyses

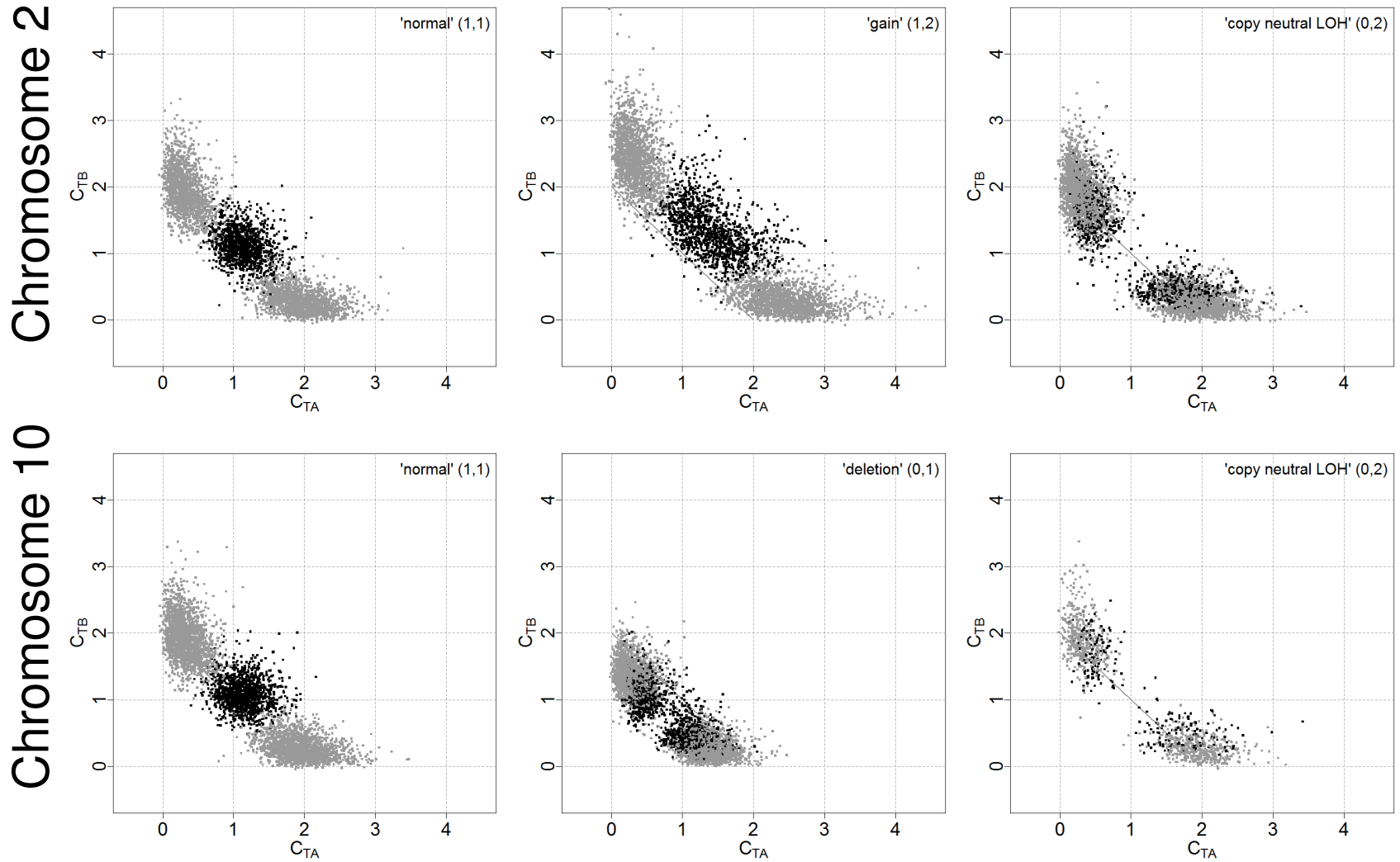
Genomic signals before normalization



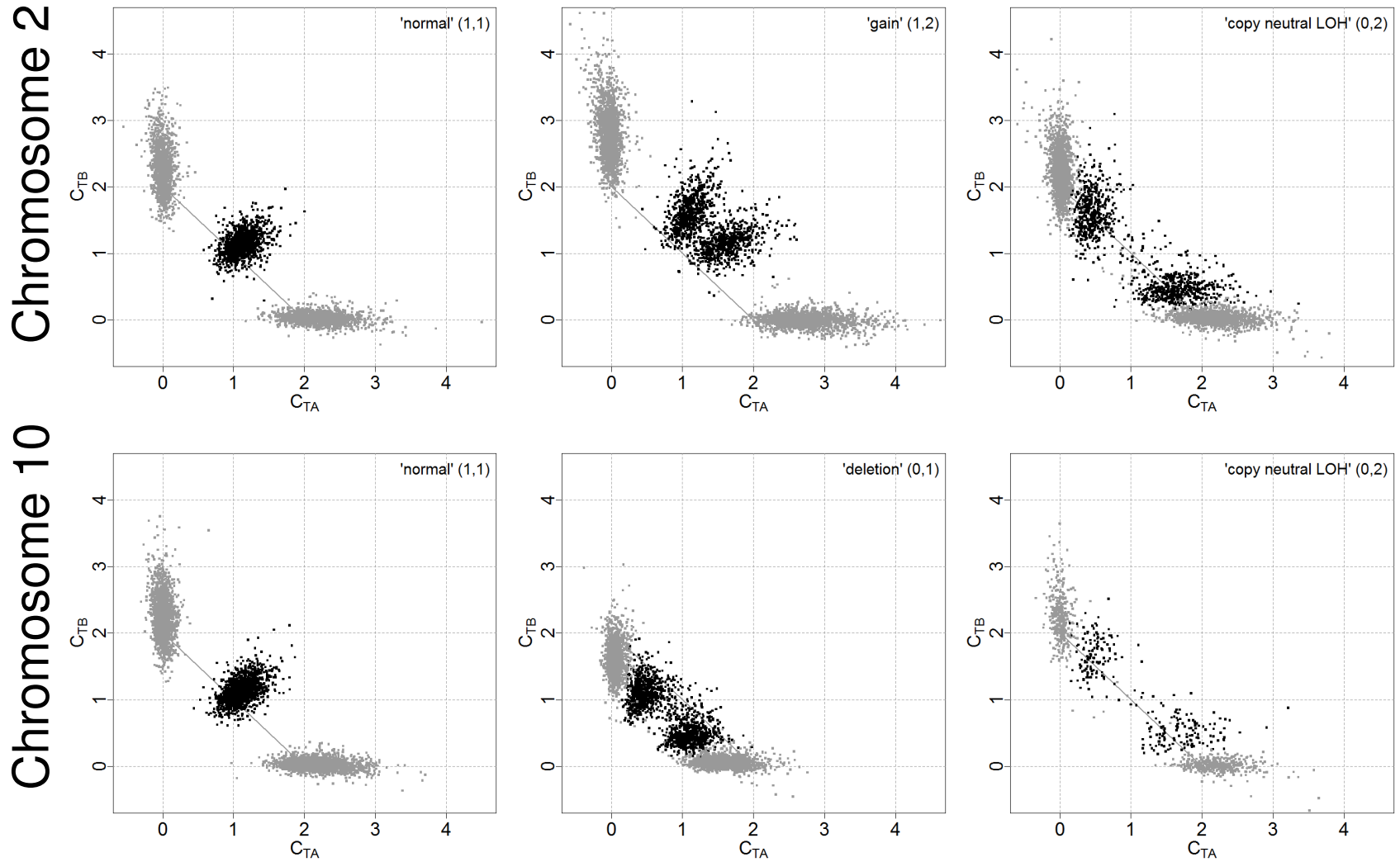
Genomic signals after normalization



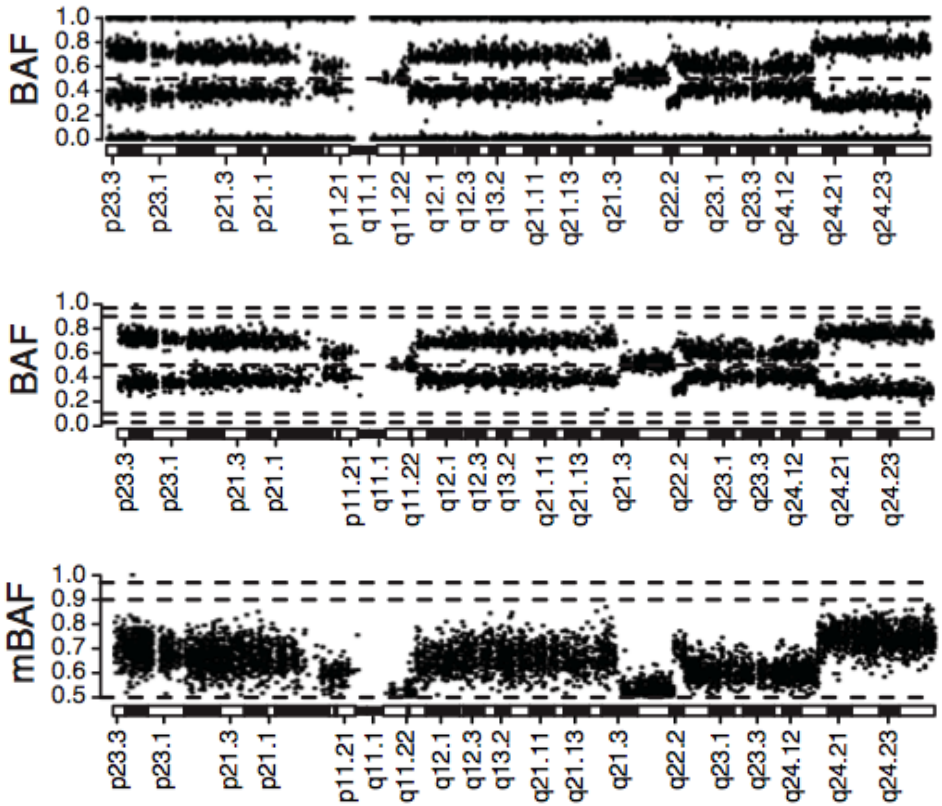
ASCNs before normalization



ASCNs after normalization



Detecting changes in allele B fractions



allele B fractions : β

allele B fractions for heterozygous SNPs

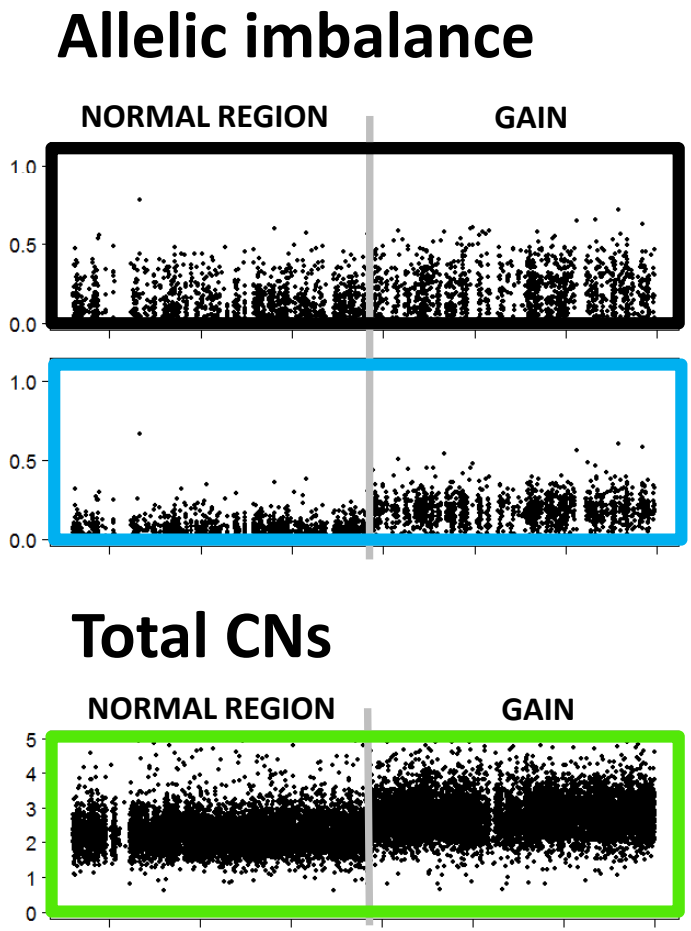
"mirrored" allele B fractions for heterozygous SNPs :

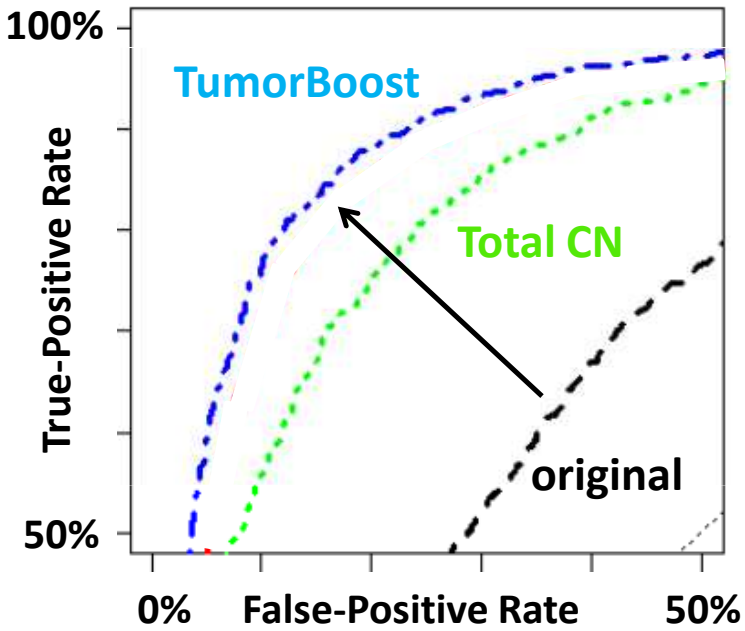
$$\rho = |\beta - 1/2| = DH/2$$

For heterozygous SNPs *DH* has a single mode : it can be segmented.

We use ROC analysis to assess how **separated** two regions on each side of a known change point in *DH* are.

Result : Better detection of allelic imbalances





slide: H. Bengtsson.

Complete preprocessing for a single tumor/normal pair

Available from aroma.cn and aroma.affymetrix at : `http://aroma-project.org`

- Normalization and locus-level summarization using CRMAv2 (Bengtsson et al, 2009) for the normal and the tumor sample separately
- "Naive" genotyping of the normal sample : threshold density of β
- TumorBoost normalization (Bengtsson et al, 2010)

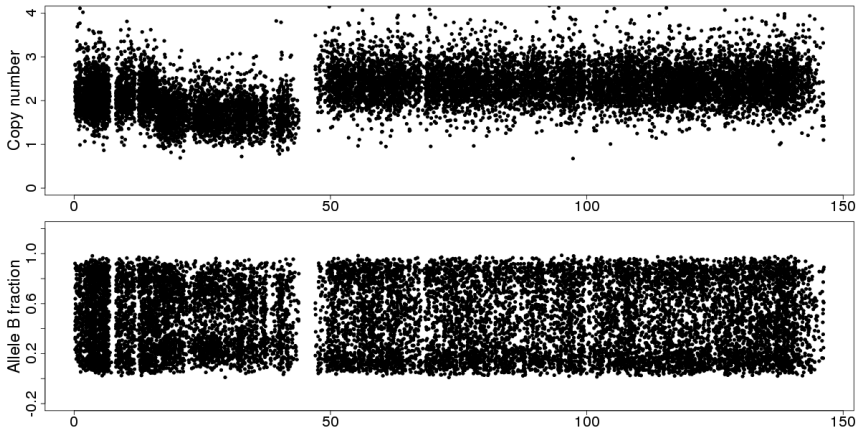
Note : genotyping errors can be taken care of by smoothing or using confidence scores.

What if no matched normal is available ? CalMaTe

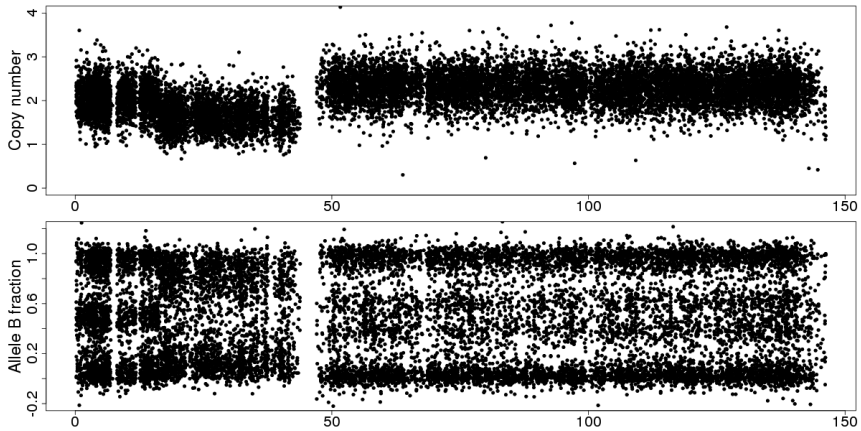
For each SNP :

- Estimate a calibration function (from observed signals to genotypes) using a set of reference samples
- Back-transform test samples

Before CalMate normalization



After CalMate normalization



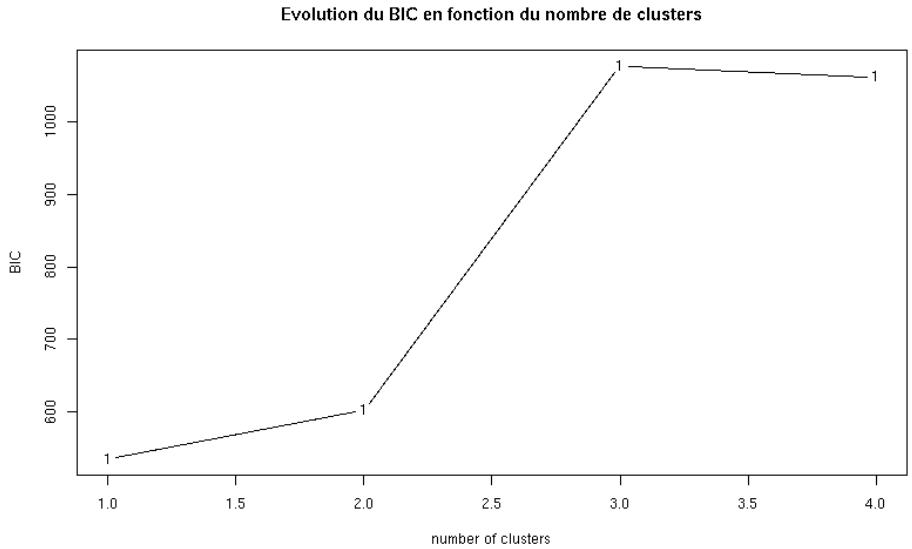
DNA copy number analyses

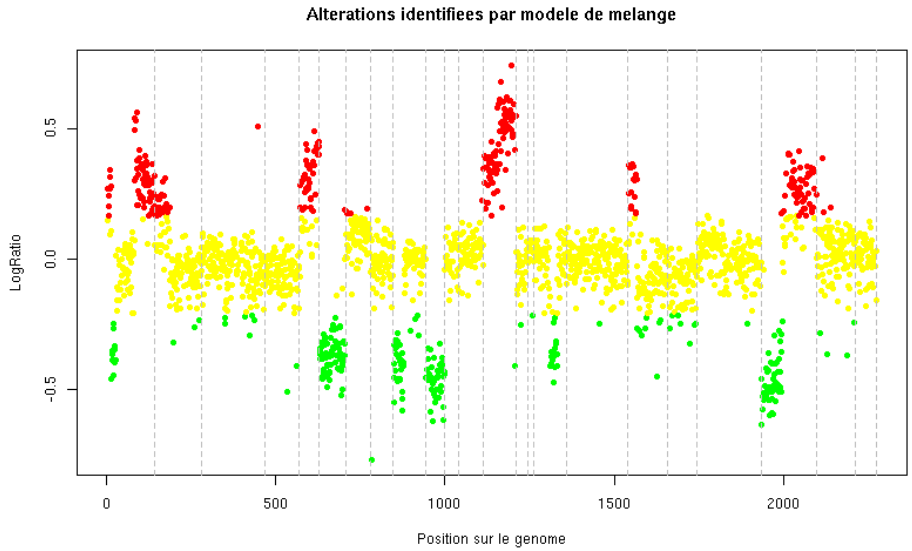
Limitation of direct approaches

Mixture models

Method

- for a fixed K, estimation of a mixture model by the EM algorithm
- choice of K : penalization such as BIC



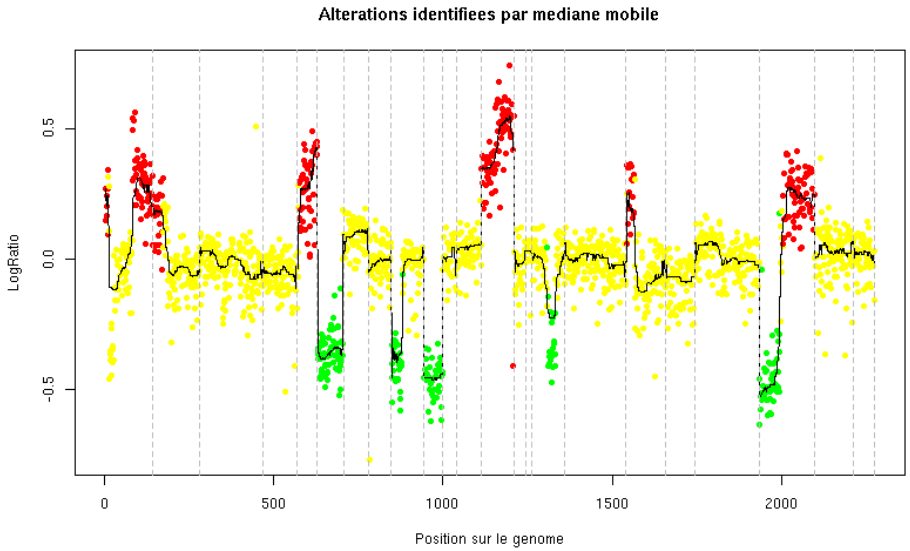


Limitation of direct approaches

Smoothing methods

Method

1. moving average around each locus
2. threshold-based segmentation

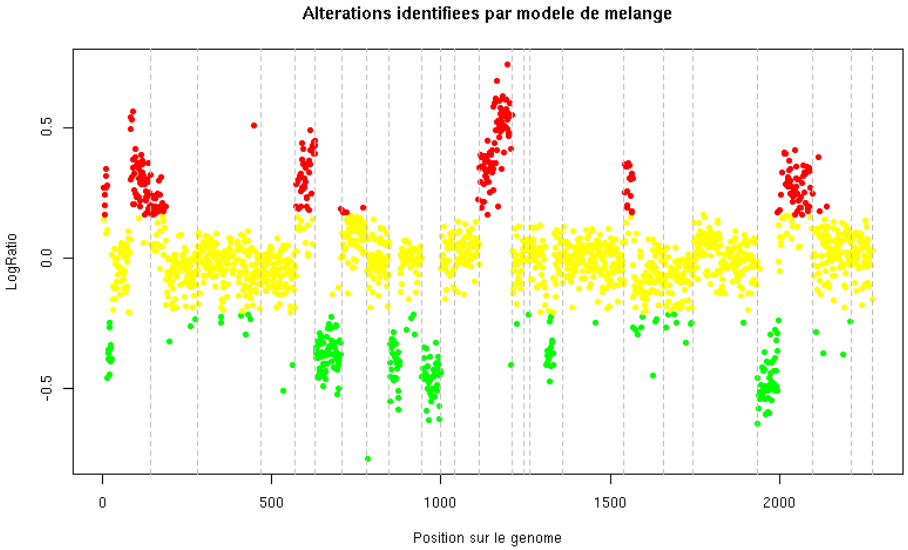


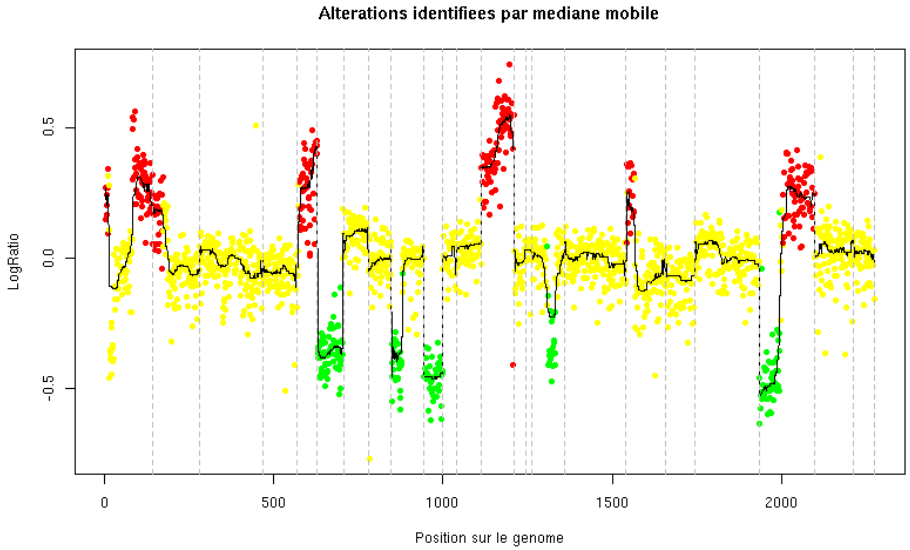
Paramerts

- window size
- number of classes
- thresholds to be applied to the smoothed signal

Limitation of direct approaches

The need for finer methods





A "good method" needs to :

- take the **genome** into account
- be able to detect **abrupt changes** (or breakpoints)

Change point detection methods

Notation

- $\mathcal{J} = 1 \dots J$: genomic loci
- $\boldsymbol{\gamma} = (\gamma_j)_{j=1\dots J}$: true DNA copy numbers
- $\mathbf{c} = (c_j)_{j=1\dots J}$: observations

Assumptions

- change points : $\mathbf{t}(K) = (t_k)_{0 \leq k \leq K}$, ordered vector with $t_0 = 1$ and $t_K = J$
- region-level DNA copy numbers $\boldsymbol{\Gamma} = (\Gamma_k)_{1 \leq k \leq K}$

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

We thus observe $c_j = \Gamma_{k(j)} + \varepsilon_j$, with $k(j) = max\{k, t_k \leq j\}$, where errors $(\varepsilon_j)_{j=1\dots J}$ are iid and generally assumed to be distributed as $\mathcal{N}(0, \sigma^2)$

Estimation in the Gaussian segmentation model

Log-likelihood maximization

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

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

In practice : number of change points and their position are unknown

- model selection : choosing K
- combinatorics : change point location among all $\binom{K-1}{J-1}$

$\binom{K-1}{J-1} = O(J^{K-1})$: exhaustive search is unfeasible in realistic situations : $\binom{50}{10^5} = 3.2 \times 10^{185}$.

Proposed approaches

Model selection : penalization of the likelihood $\ell(K, \cdot)$

- penalized likelihood $\bar{\ell}(K, \cdot)) = \ell(K, \cdot) - \beta pen(K)$, with $pen(K)$ increasing in K
- choice of $pen(K)$ depends on the number of parameters to estimate
- usual choices for β : $\beta = 1$ (AIC), $\beta = \frac{1}{2}\log(n)$ (BIC)

Exploring the space of possible partitions

- Heuristics : genetic algorithm or circular binary segmentation
- Exact solutions by dynamic programming
- Convex relaxation using Lasso-type approaches

DNA copy number analyses

Heuristics : genetic algorithm

Jong *et al*, 2003

Local search for a fixed K

- initialization : choice of K change points
- iteration : move a change point left or right if likelihood increases

Genetic algorithm (iterative)

initialization : population of N segmentations

1. choice of two “parents” at random
2. generation of two “offsprings”be recombination of parental profiles
3. local search for each offspring
4. replacement of the less adapted individuals among N by the two offsprings

Heuristics : circular binary segmentation

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

Binary segmentation in Gaussian models

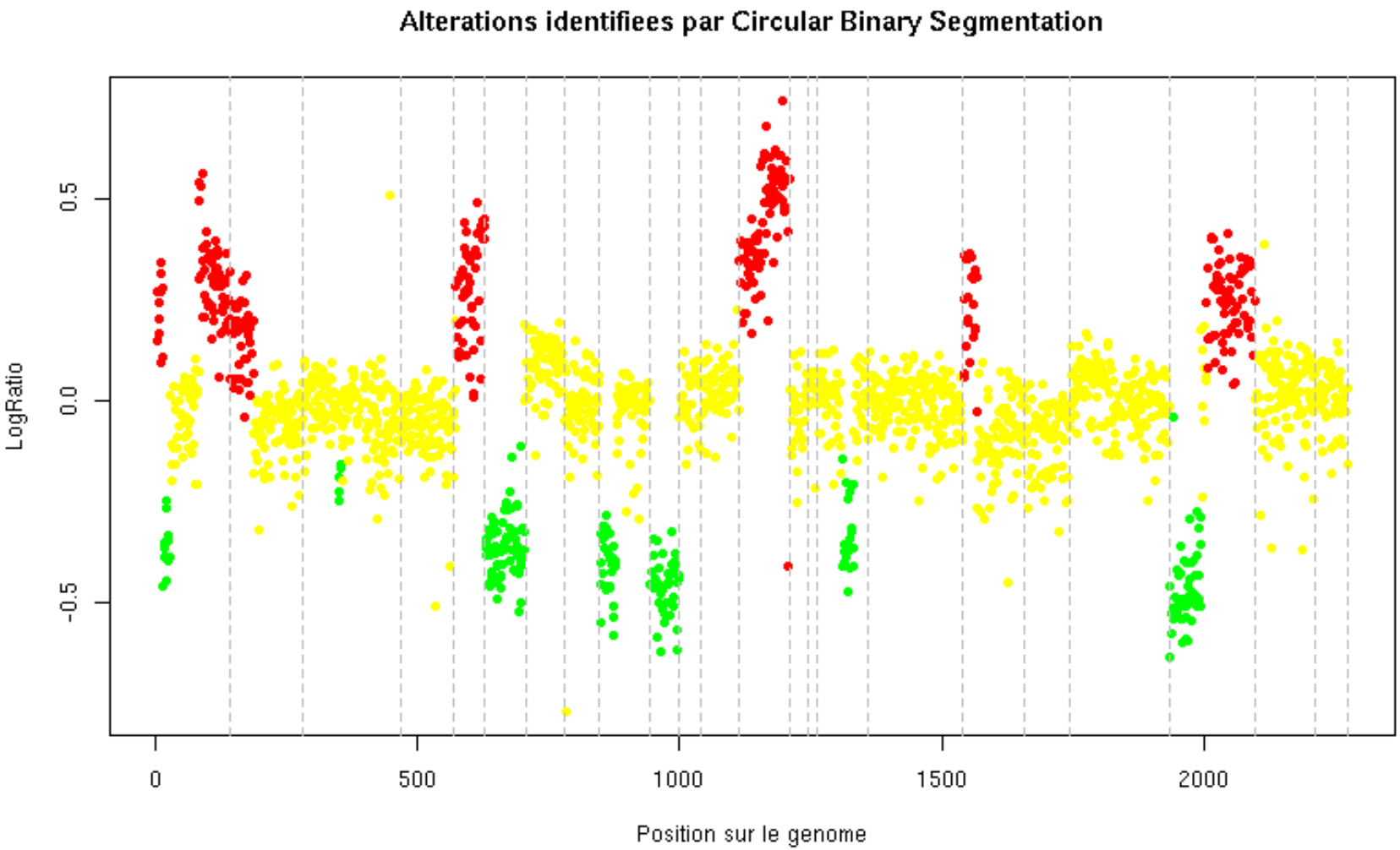
- comparison of partial sums : $S_u = \sum_{j=1}^{u} y_j$
- test statistic (LR) : $Z_u = \frac{1}{\sqrt{\frac{1}{u}+\frac{1}{J-u}}} \times \left[\frac{S_u}{u} - \frac{S_J - S_u}{J-u} \right]$
- assuming no change point, $Z_u \sim \mathcal{N}[0, \sigma^2]$

Adaptations to DNa copy number data

- detection of **nested segments :**
  $$Z_{uv} = \frac{1}{\sqrt{\frac{1}{v-u}+\frac{1}{J-(v-u)}}} \times \left[\frac{S_v - S_u}{v-u} - \frac{S_J - (S_v - S_u)}{J-(v-u)} \right]$$
- detection de **several segments** by recursion
- calculation of a *p*-value using **permutations**

Heuristics : circular binary segmentation

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



Exact solution by dynamic programming

Picard *et al*, 2005

Dynamic programming : additivity of the likelihood

1. Calculate $\ell(1, j_1 : j_2)$ for any (j_1, j_2) such that $1 \leq j_1 < j_2 \leq J$
2. Go from $\ell(K, \cdot)$ to $\ell(K+1, \cdot)$ noting that

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

Complexity decreases from $O(J^K)$ to $O(KJ^2)$

Choosing the number of change points

Penalization proposed by Lavielle (2005) :
boils down to choosing $\hat{K}$ as an inflexion point of $K \mapsto \ell(K, 1 : J)$

Convex relaxations

Solves a different, but computationally simpler problem

Adaptation of the "Fused Lasso" (Tibshirani and Wang, 2007)

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

Complexity : $O(J^2)$

Simplification (Harchaoui and Lévy-Leduc, 2008)

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

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

Conclusions on change point models

Tradeoff between accuracy and computing time.

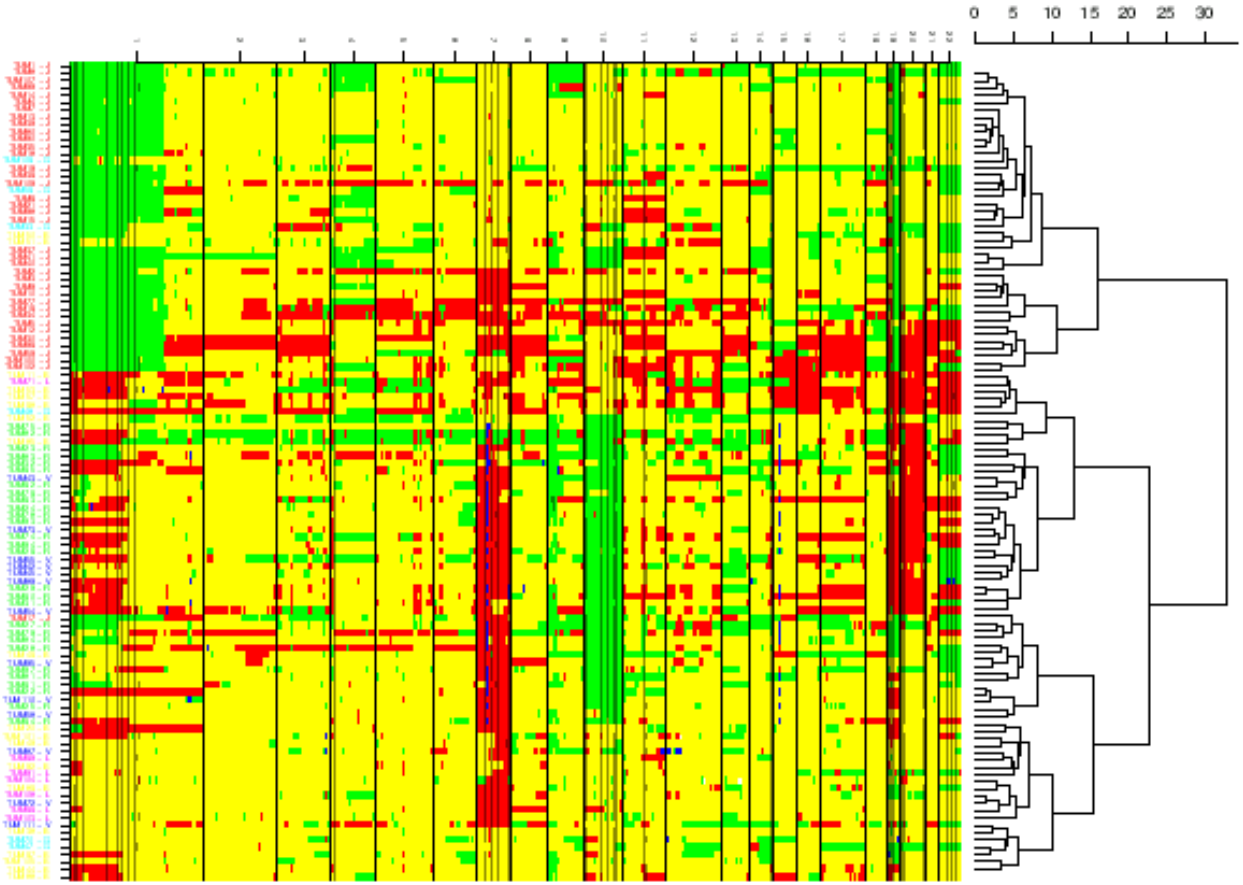
One possibility is to have two steps :

1. fast search with false positives
2. pruning or exhaustive search on remaining change points

Requires minimizing the number of **false negatives** during first step

DNA copy number analyses

Expected benefits of a multi-profile segmentation



Existing methods : :

- Extension of CBS (Zhang et al, Biometrika, 2010)
- Extension of Lasso-like methods (Bleakley and Vert, NIPS 2010)

Multi-profile circular binary segmentation

Zhang et al, Biometrika, 2010

Test statistic for the n^{th} profile (known variance) :

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

Test statistic for N profiles (known variance)

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

If no change point, $Z_{uv}^{[N]} \sim \chi^2(N)$ (asymptotically)

Error rate control

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

Multiplatform segmentation : motivations

Existing approaches :

- Integration before segmentation (Bengtsson et al, Bioinform., 2009)
- Extension of CBS (Zhang et al, Bioinform., 2010)

Integration before segmentation : method

Bengtsson et al, Bioinformatics, 2009

Multiplatform data in R^4 (4 platforms):

True CN:

x (an unknown scalar)

Smoothed CNs:

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

Unknown transformation:

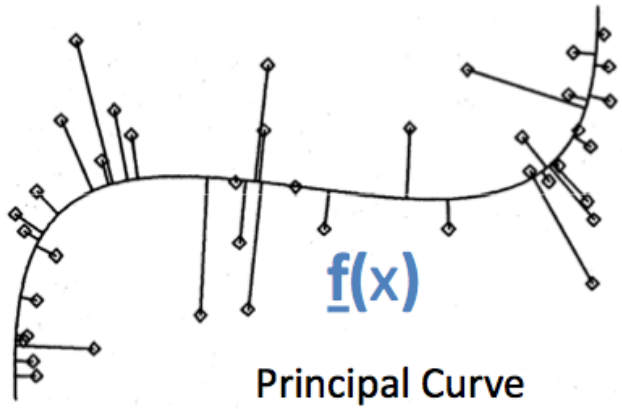
$\underline{f}(x) = (f^{(1)}(x), f^{(2)}(x), f^{(3)}(x), f^{(4)}(x))^T$

Noise:

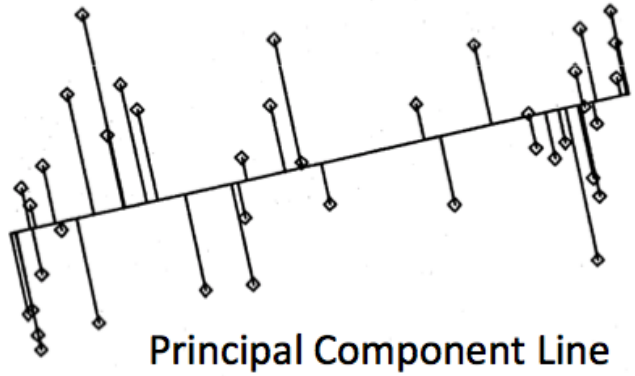
$\underline{\varepsilon} = (\varepsilon^{(1)}, \varepsilon^{(2)}, \varepsilon^{(3)}, \varepsilon^{(4)})^T$

Vector model:

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



Principal Curve



Principal Component Line

Hastie et al., Principal Curves, JASA, 1989

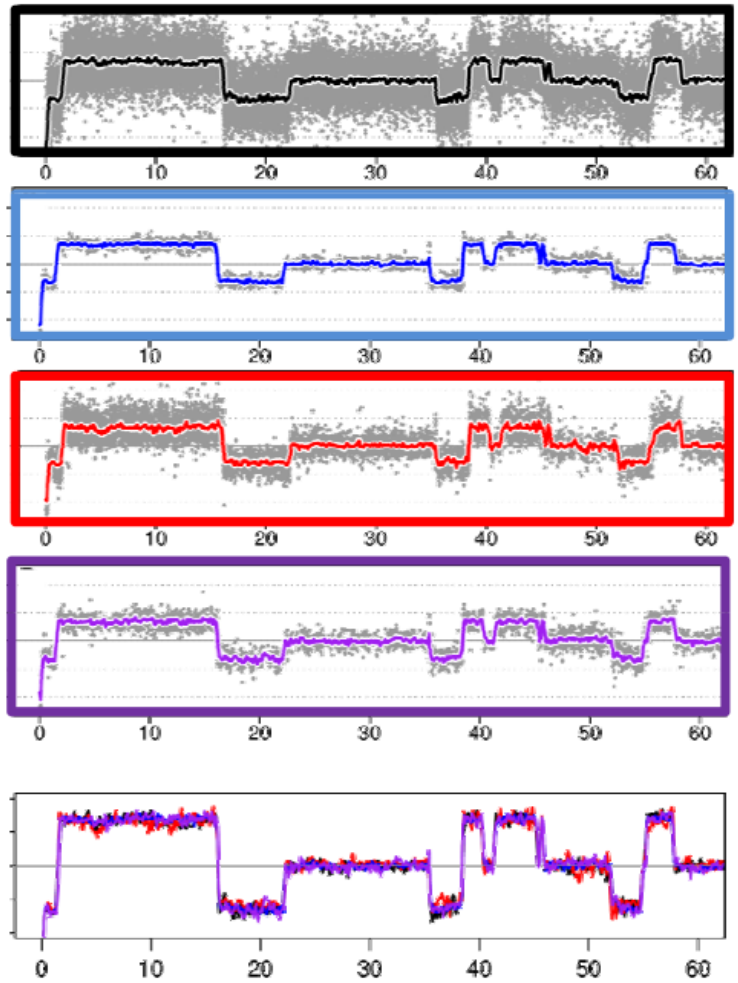
Integration before segmentation : results

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

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

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

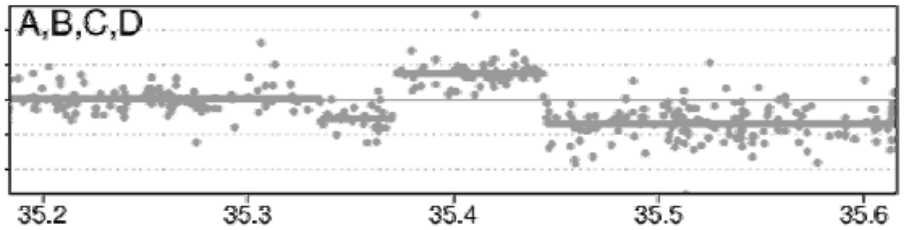
D. Harvard
Agilent
HG-CGH-244A
($n=0.25 \cdot 10^6$)



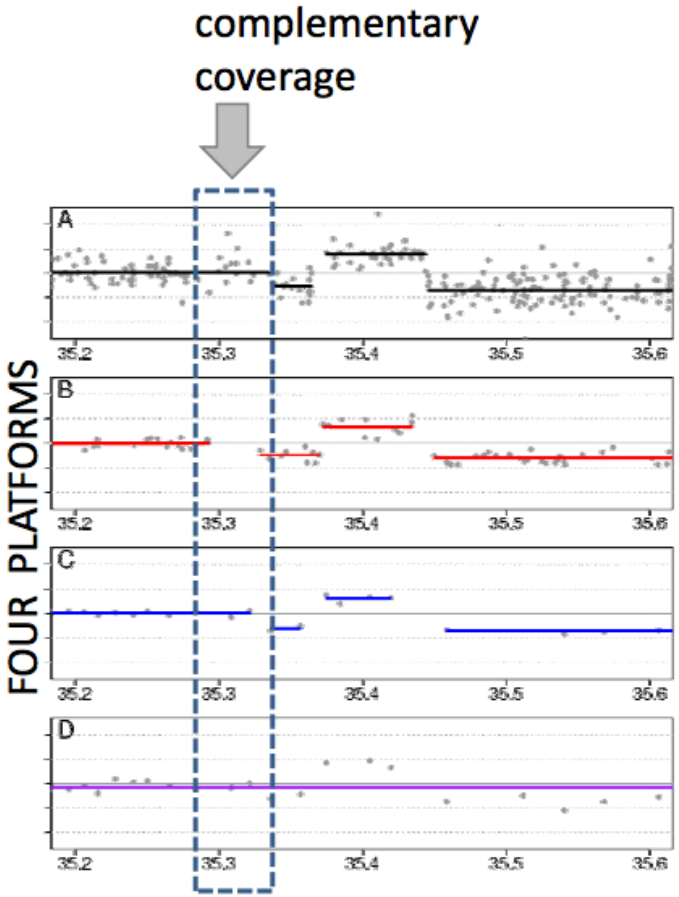
slide: H. Bengtsson.

Integration before segmentation : resolution

Combining normalized data:



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

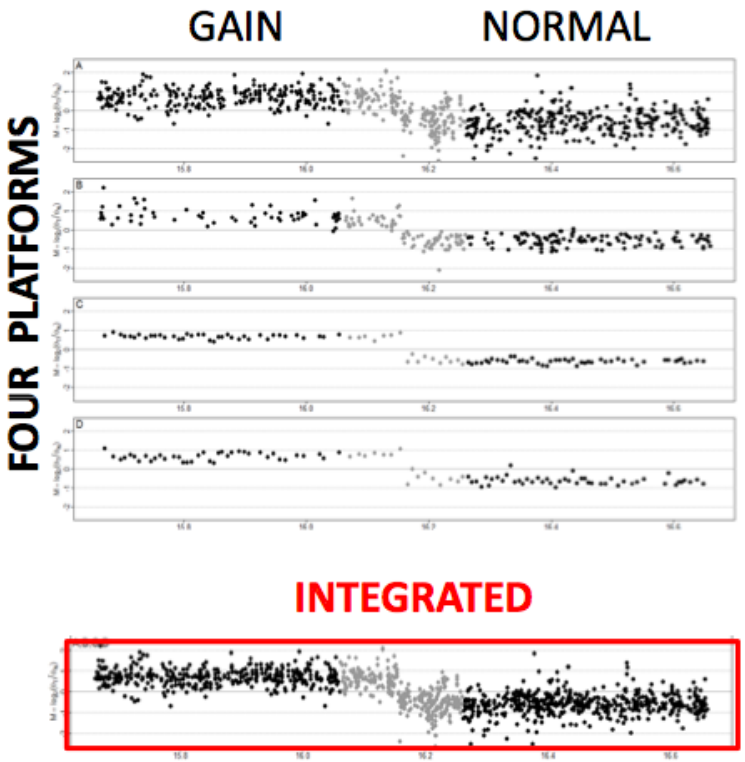


slide: H. Bengtsson.

Integration before segmentation : performances

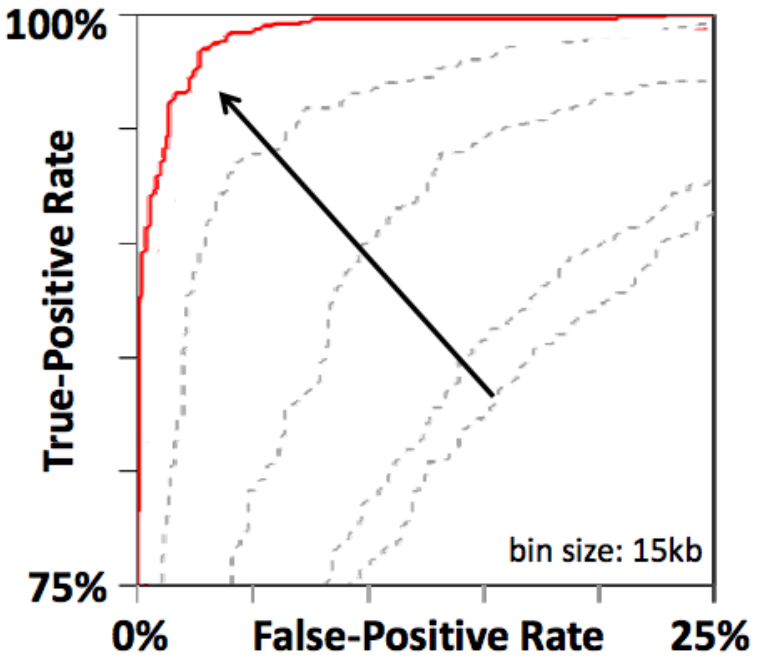
Data:

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



Assessment via ROC:

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



Repeat:

Repeat the above for several change points.

slide: H. Bengtsson.

Multi-platform circular binary segmentation

Zhang et al, Bioinformatics, 2010

Test statistic for platform n :

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

Tets statistic for N platforms

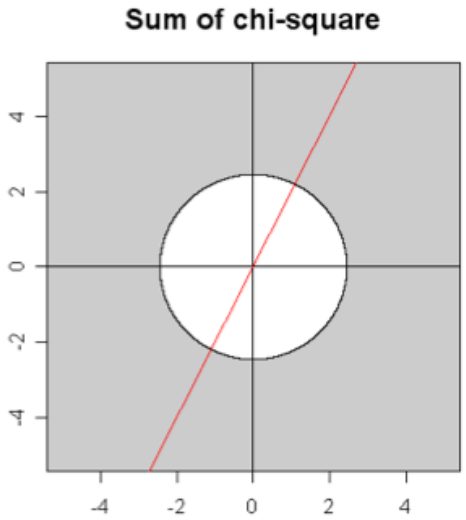
$$Z_{uv}^{[N]} = \frac{\left[\sum_{n=1}^{N} \delta_{uv}^{k} Z_{uv}^{n}\right]^2}{\sum_{n=1}^{N} (\delta_{uv}^{k})^2}$$

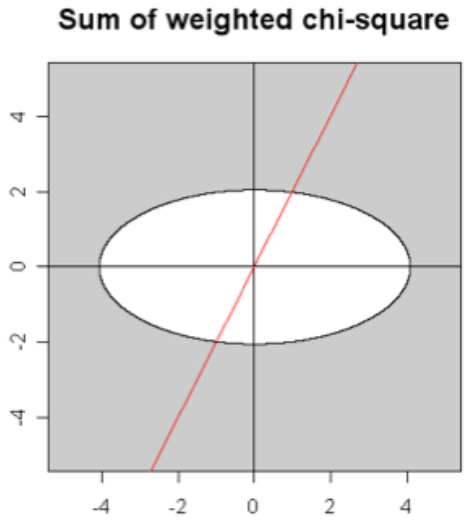
+ modified BIC criterion to choose the number of breakpoints.

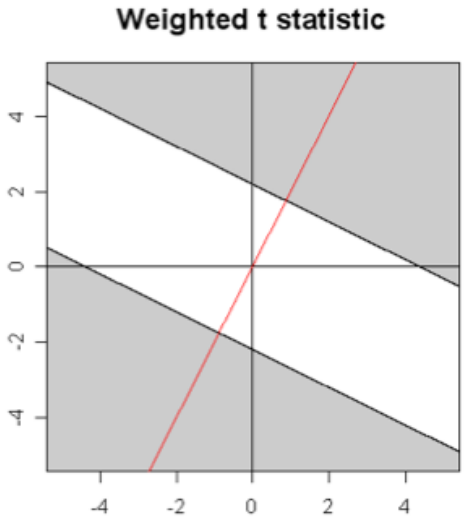
Multi-platform circular binary segmentation

Intuition for the test statistics

Rejection regions (gray) for a two-dimensional statistic built from two one-dimensional statistics.

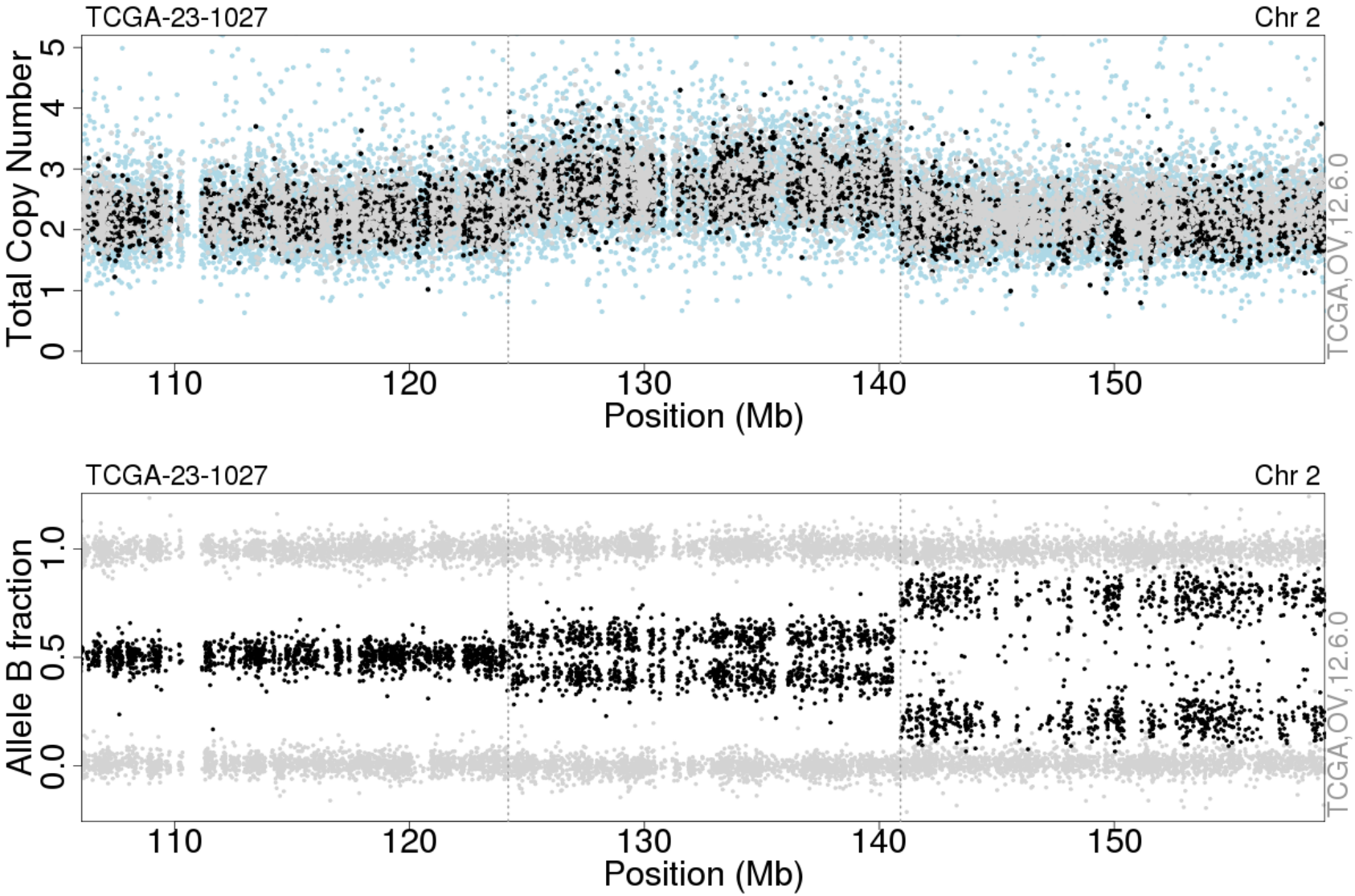




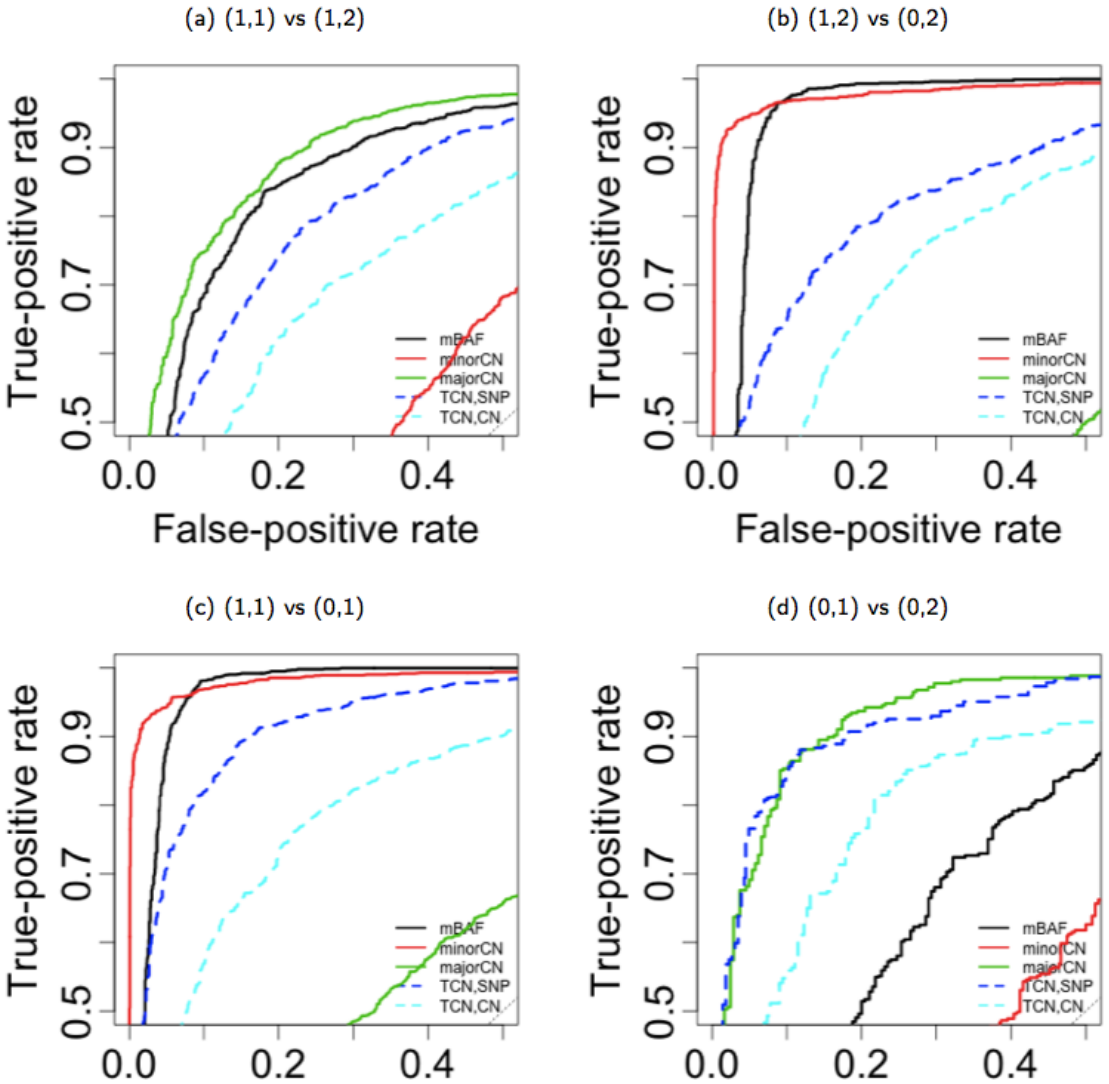


DNA copy number analyses

Changes can be reflected in both dimensions



DH has greater detection power than *C* at a single locus



More informative probes for *C* than *DH*

Affymetrix GenomeWideSNP_6

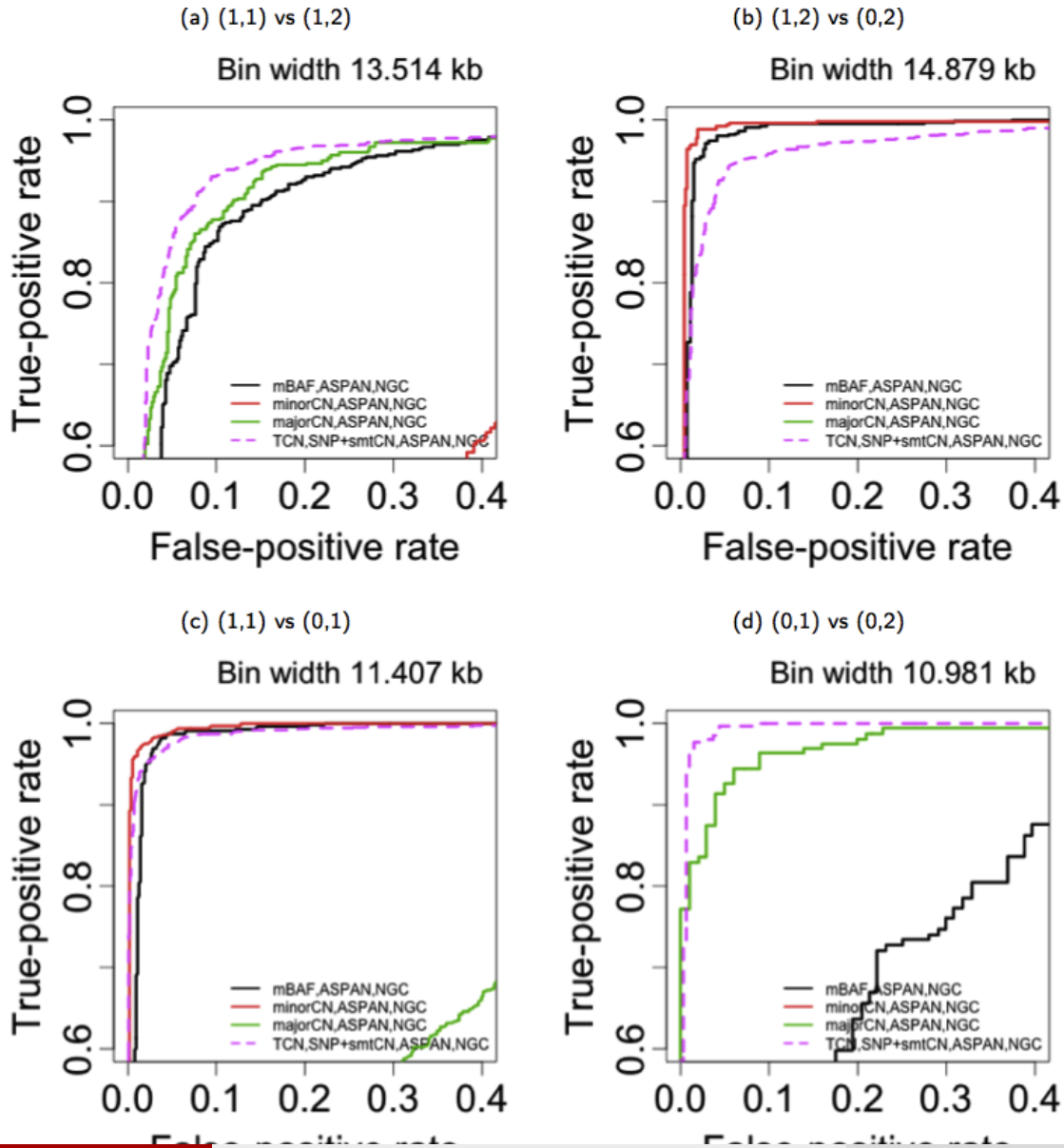
	All units	CN units	SNP units
Frequency	1,856,069	946,705	909,364
Proportion	100%	51%	49%

Unit types

	All units	AA	AB	BB
Frequency	1,856,069	326,500	251,446	331,418
Proportion	100%	18%	**14%**	18%

SNPs by genotype call for sample TCGA-23-1027

Similar detection power at a fixed resolution



Need for a truly joint dimensional segmentation method

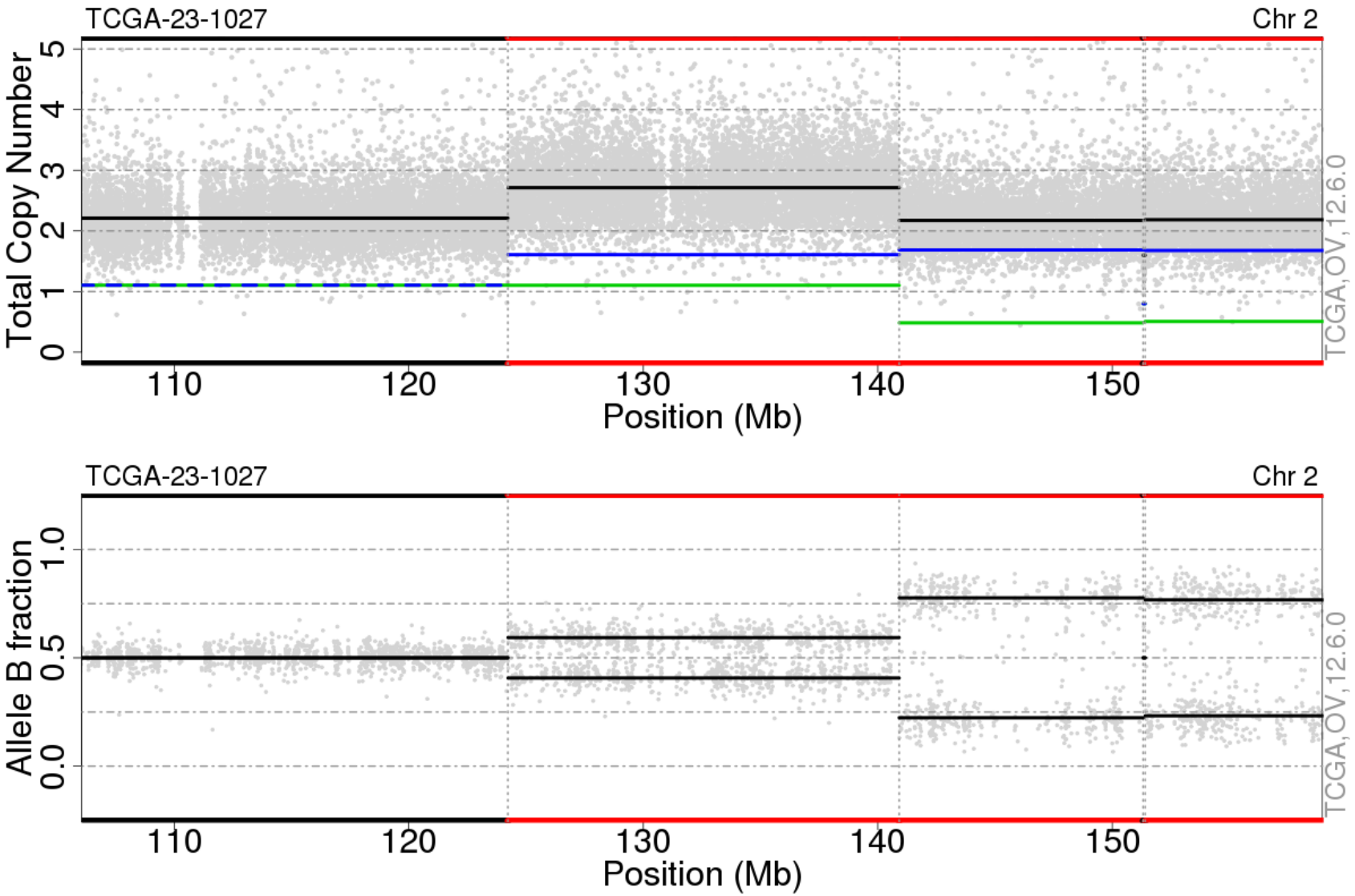
- Most methods segment only *one* of *C* and *DH*
- Some use two-way segmentation : Olshen *et al*, [PSCBS]
- A handful are truly two-dimensional :
  - Chen *et al*, [pscn]
  - Greenman *et al*, Biostat., 2010, [PICNIC]
  - Sun *et al*, NAR, 2009, [genoCNA]

Challenges for a truly joint segmentation method

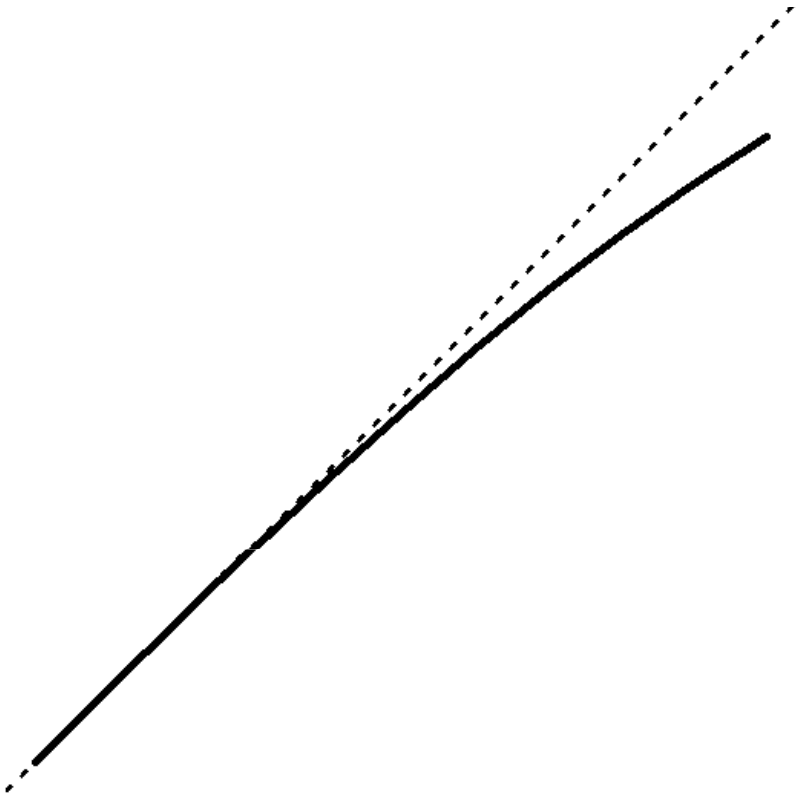
- A two-dimensional signal
- Only heterozygous SNPs can be used to detect CN changes from *DH*
- Bias in the estimation of *DH*
- *DH* is not Gaussian

DNA copy number analyses

Copy numbers are not calibrated



Non-calibrated signals : signal saturation

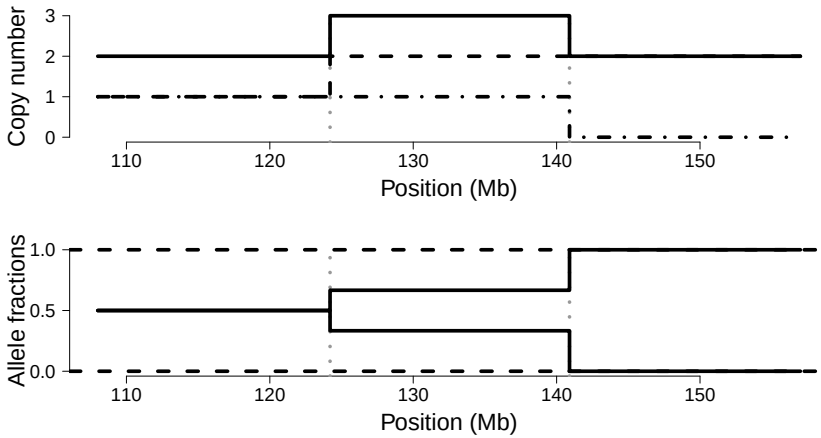


$$C_{\text{obs}} = f(C_{\text{true}}) < C_{\text{obs}}$$

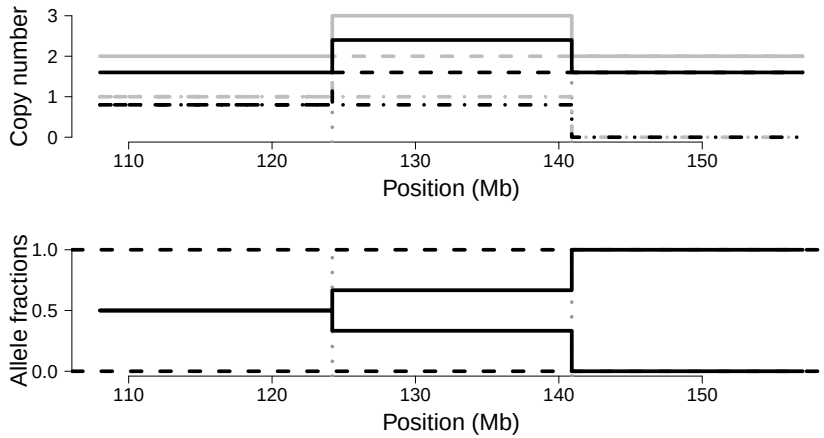
f is unknown

Non-calibrated signals : ploidy and purity

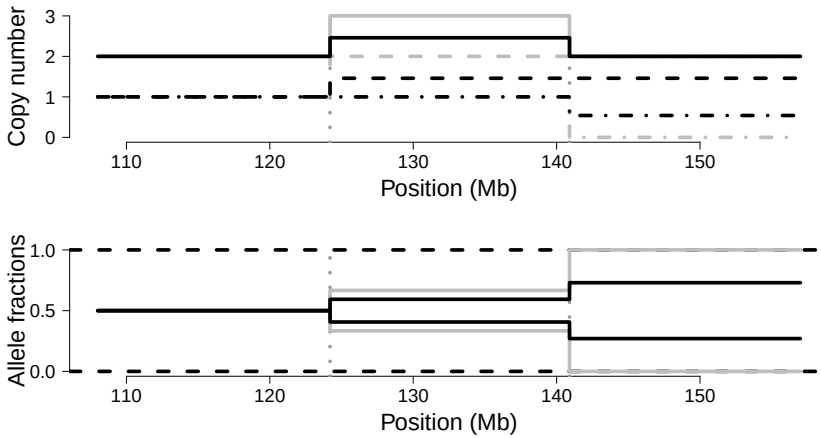
Pure tumor, ploidy = 2



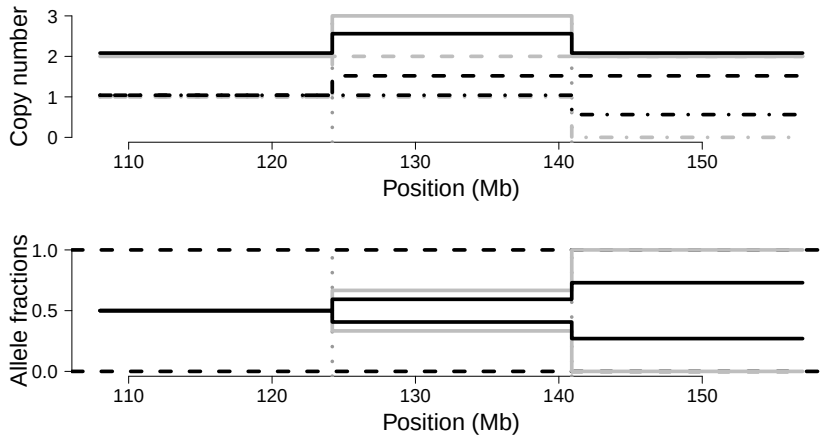
Pure tumor, ploidy > 2



Non pure tumor, ploidy=2



Non-pure tumor, ploidy < 2



Purity, ploidy, and signal saturation

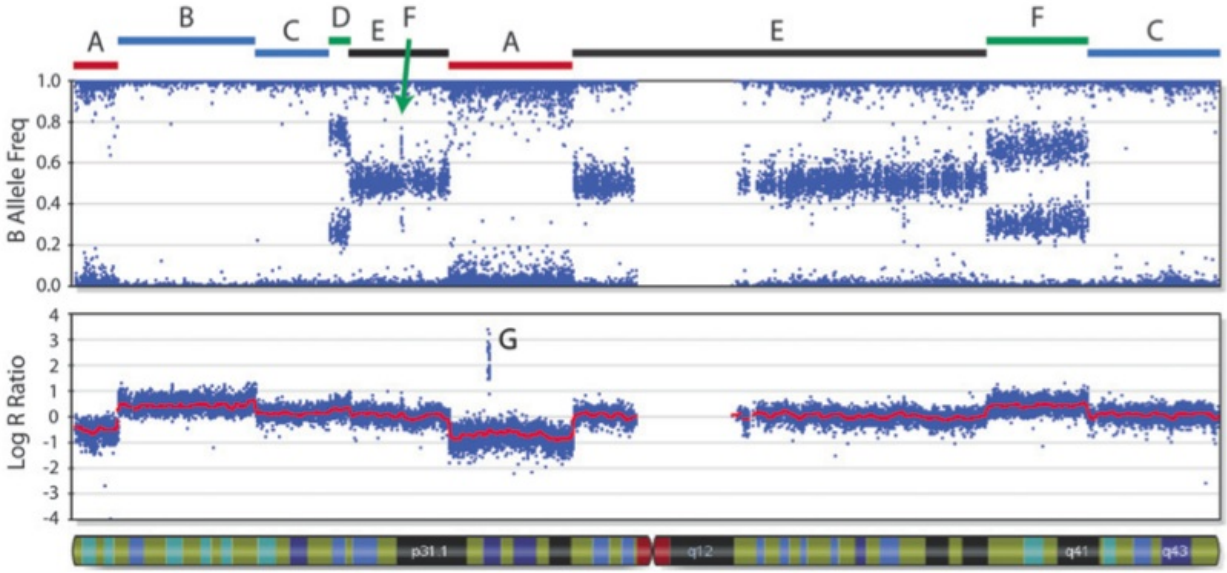
Why copy numbers are not calibrated

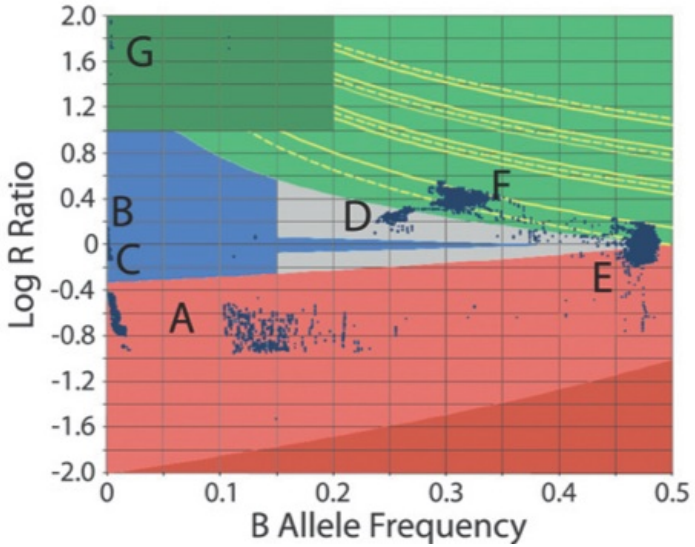
- signal saturation
- non purity : presence of normal cells in the “tumor sample”
- ploidy : the total amount of DNA is fixed by the assay

Remarks

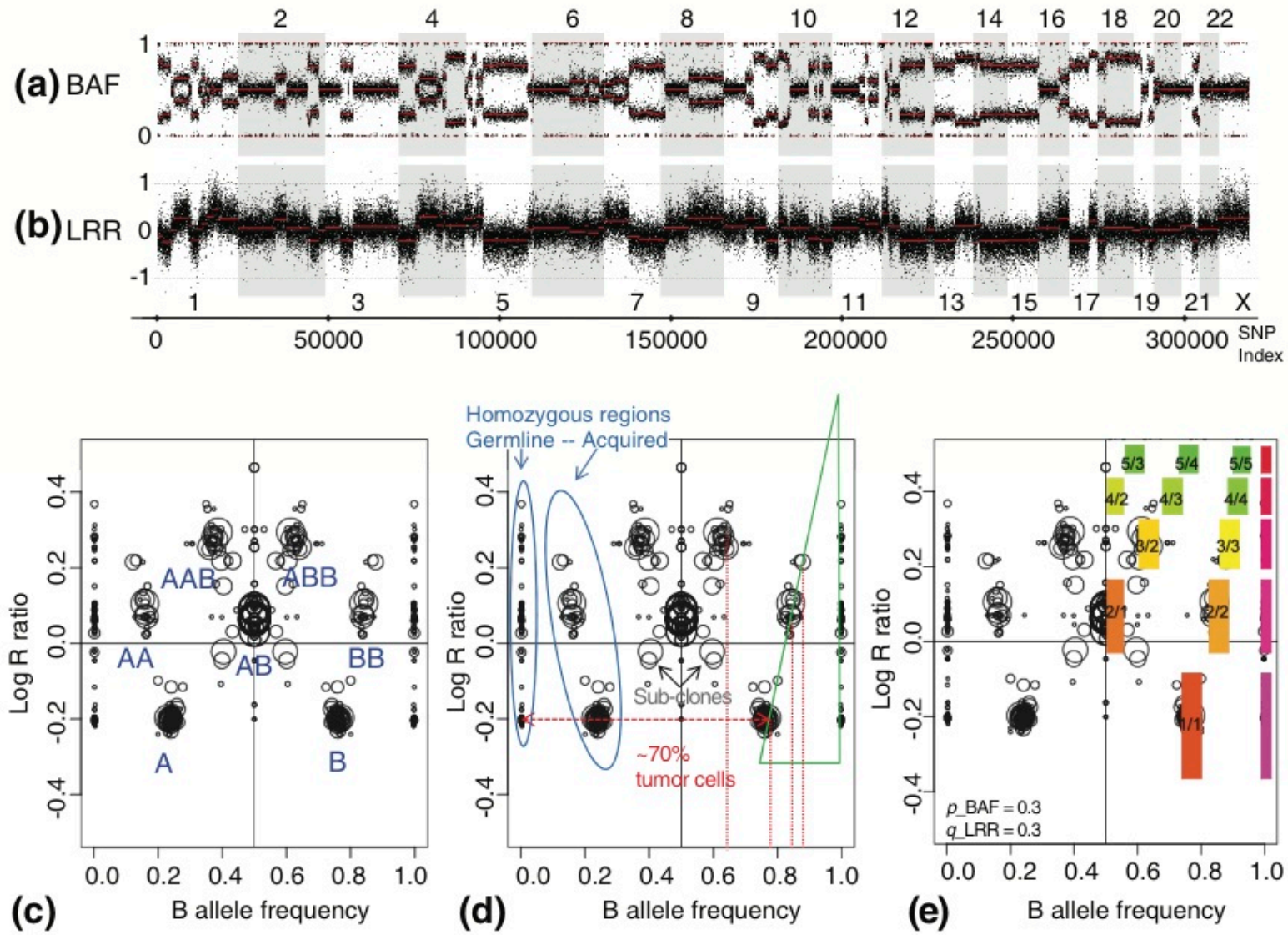
- ploidy is **not identifiable**
- purity and ploidy are biological properties of the sample
- signal saturation is an artifact from the assay
- under the rug : tumor heterogeneity

OverUnder : Attiyeh et al, Genome Research, 2009





GAP : Popova et al, Genome Biology, 2009



ASCAT : Van Loo et al, PNAS, 2010

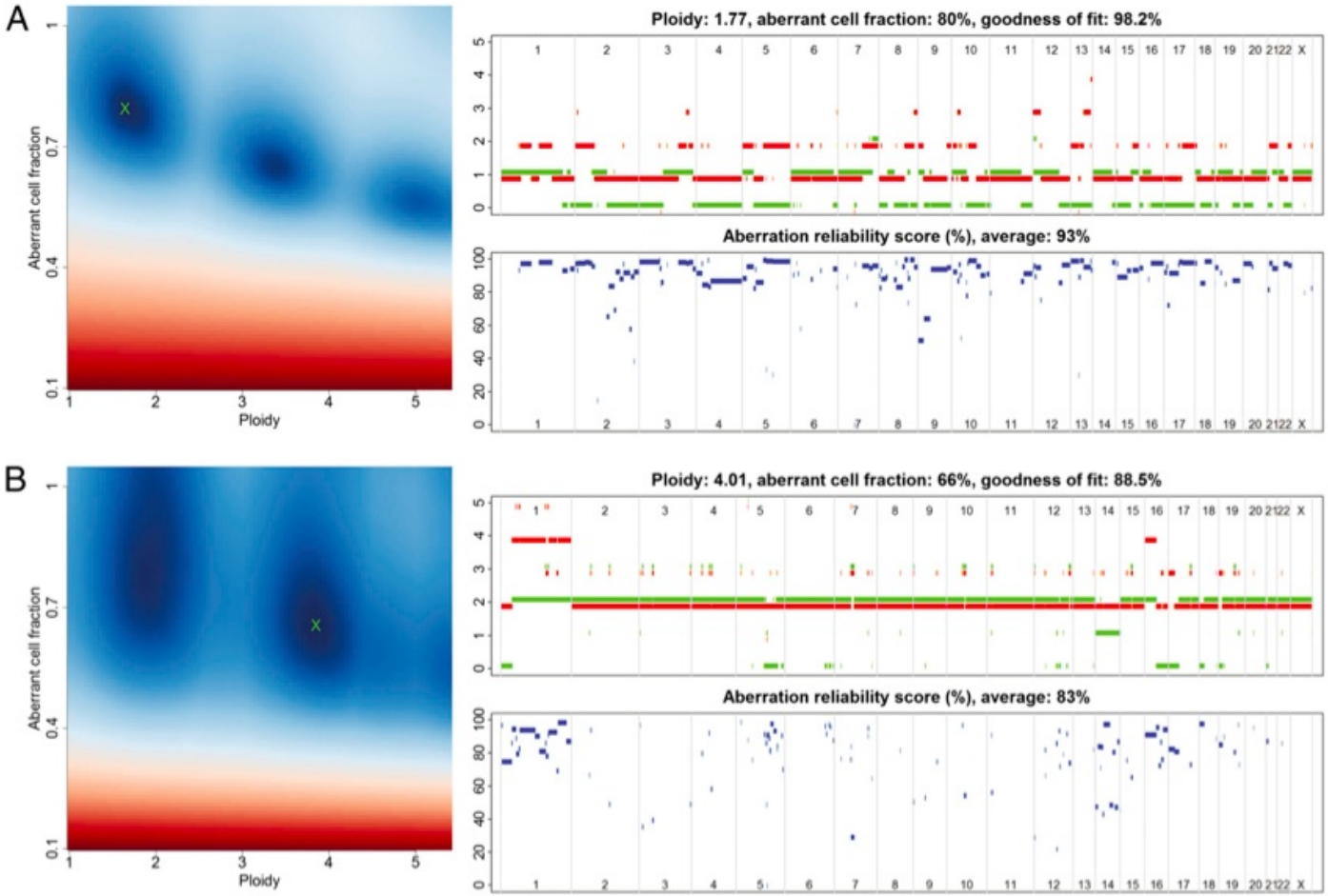


Fig. 1. ASCAT profiles and their calculation. Two examples are given: (*A*) a tumor with ploidy close to 2n and (*B*) a tumor with ploidy close to 4n. (*Left*) ASCAT first determines the ploidy of the tumor cells ψ_t and the fraction of aberrant cells ρ. This procedure evaluates the goodness of fit for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

Comments on existing approaches

- What about Affymetrix data ?
- Choice between candidate solutions
- Perform ad hoc correction for saturation
- Tumor heterogeneity ?

Conclusions

- Extracting biological information is crucial
- Joint segmentation methods exist
- They have to be adapted to joint segmentation
- Ploidy and the presence of tumor cells complicate region calling
- Most methods are implemented in R and Matlab