On detecting and calling DNA copy number alterations in cancer samples from genotyping microarrays

Pierre Neuvial

Department of Statistics, UC Berkeley

Outline

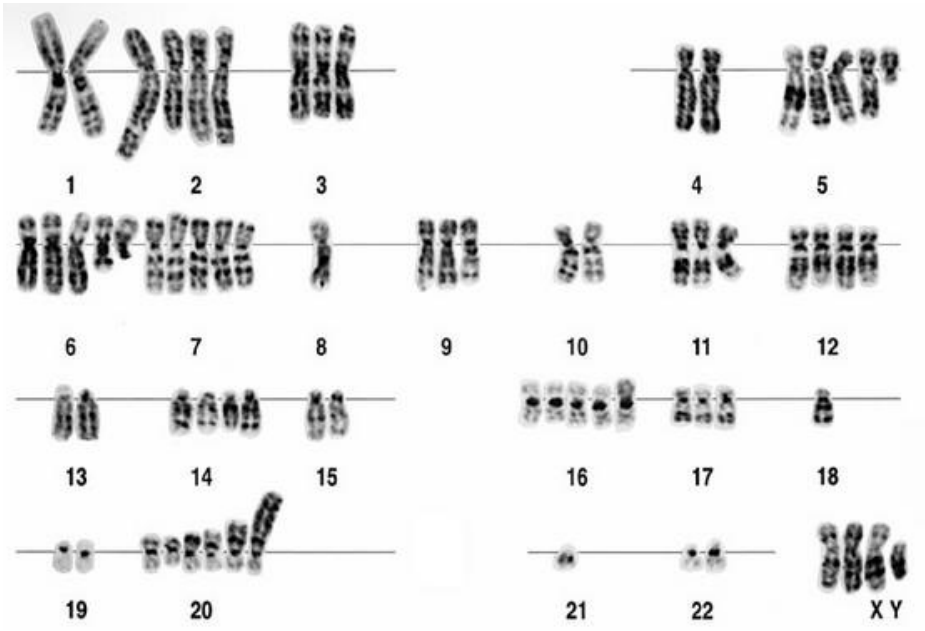
1. Background and motivation
2. Normalizing each SNP of a single tumor/normal pair
   - Motivation: taking advantage of SNP effects
   - Results: improved signal to noise ratio of allelic signals
3. Detection: is it better to use AR or TCN ?
   - Detecting copy number changes from TCN and AR
   - Comparing detection power of TCN and AR
4. Calling: influence of purity and ploidy
   - Purity and ploidy
   - Thoughts for calling copy number states

Outline

Genomic changes at the DNA level are hallmarks of cancer

We inherited 23 paternal and 23 maternal chromosomes, mostly identical.

Normal karyotype



Tumor karyotype

Our goal: identify CN changes to improve characterization, classification, and treatment of cancers

Parental, minor and major copy numbers

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

Copy number state at genomic locus j

$$(\underline{\gamma}_j, \overline{\gamma}_j),$$

where

$$\begin{cases} \underline{\gamma}_j & = \min(m_j, p_j) \\ \overline{\gamma}_j & = \max(m_j, p_j) \end{cases}$$

Copy numbers states of interest in cancer

- amplification of small regions
- recurrent gains or losses across samples
- Loss of Heterozygosity (LOH)

	Deletion	Neutral	Gain
Loss of Heterozygosity	(0,1)	(0,2)	$(0, M)$ with $M \geq 3$
Heterozygosity	(0,0)	(1,1)	(m, M) with $1 \leq m < M$

CN states as the conjunction of information regarding total copy number (columns) and heterozygosity (rows).

Minor and major copy numbers characterize these CN events in cancers

Genotyping microarrays (SNP arrays)

Single Nucleotide Polymorphisms (SNPs)

Genomic loci (single base positions) of variation across individuals.
Variants are called alleles and arbitrarily labeled A and B

SNP arrays quantify

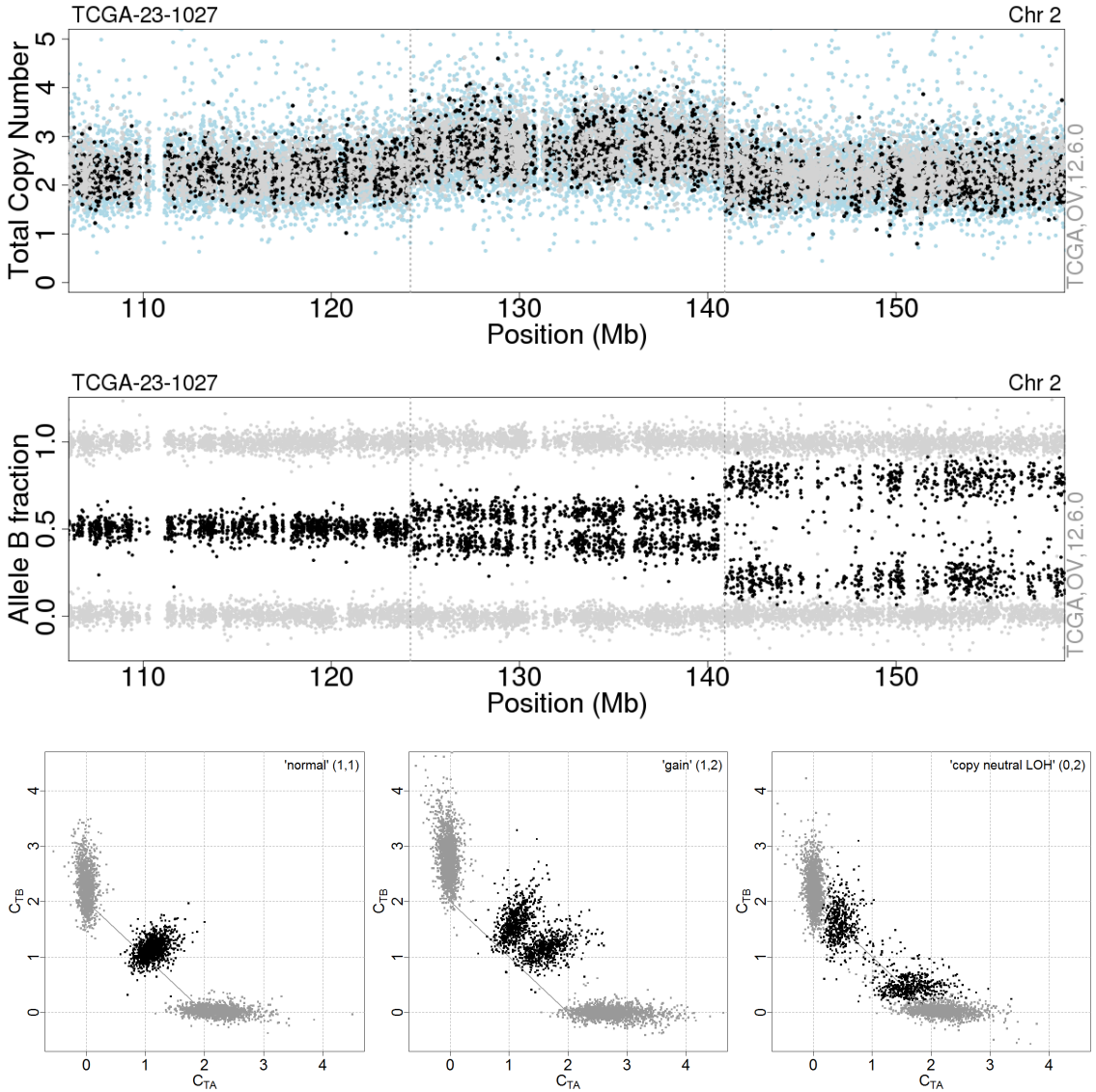
- allelic copy numbers (C_A, C_B) at $\sim 10^6$ SNPs
- total copy numbers at non-SNP locations

The data are generally summarized by a 2d vector (C, β):

- Total Copy Numbers (TCN) : $C = C_A + C_B$
- Allelic Ratios (AR): $\beta = C_B/(C_A + C_B)$

Minor and major copy numbers can be estimated from SNP arrays

What SNP array data look like



Statistical questions

Identification of two types of CN changes:

1. Variation in total copy numbers
2. Allelic Imbalance (AI)

Identification means detection (finding regions) and calling (labelling regions).

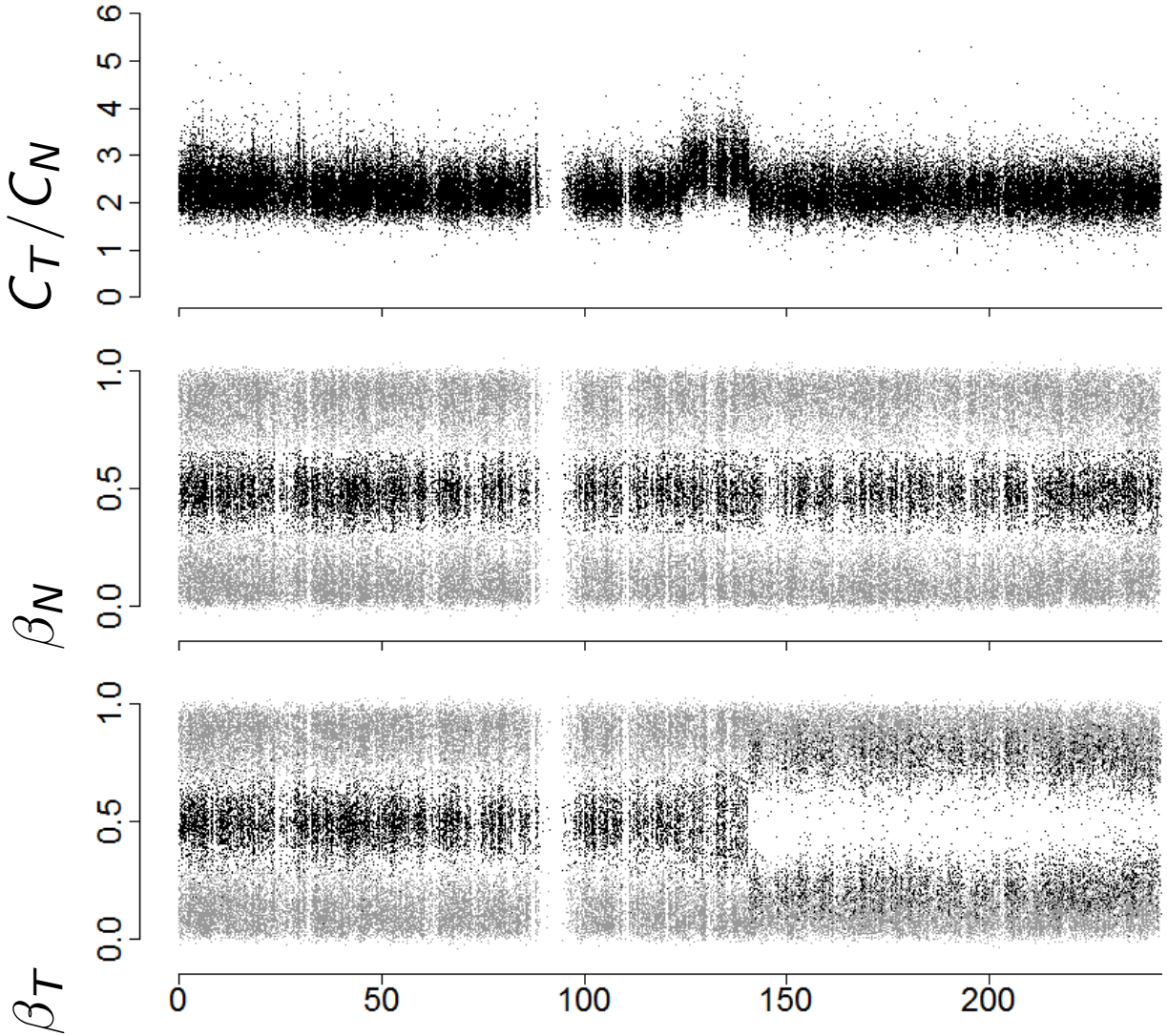
Outline

Outline

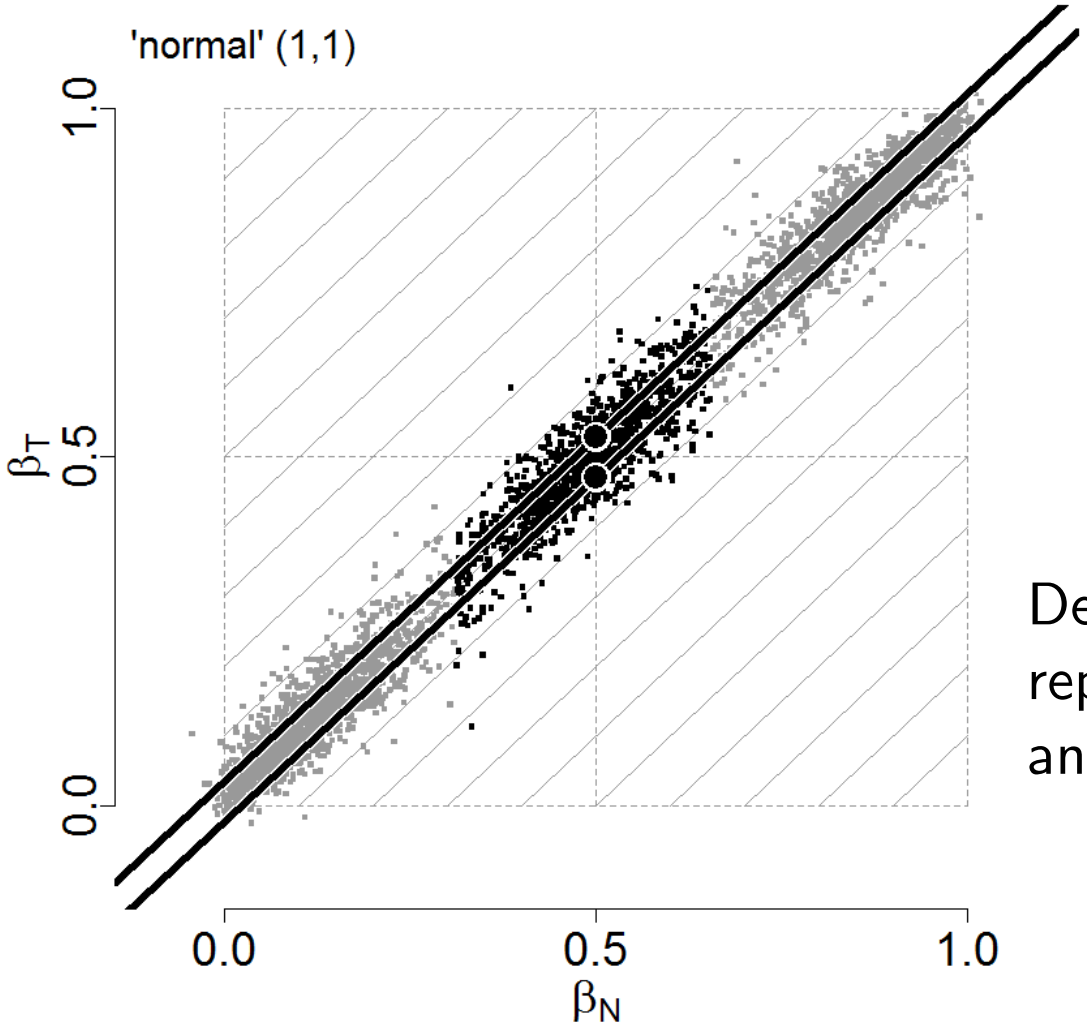
Outline

Raw genomic signals

After preprocessing using the CRMAv2 method



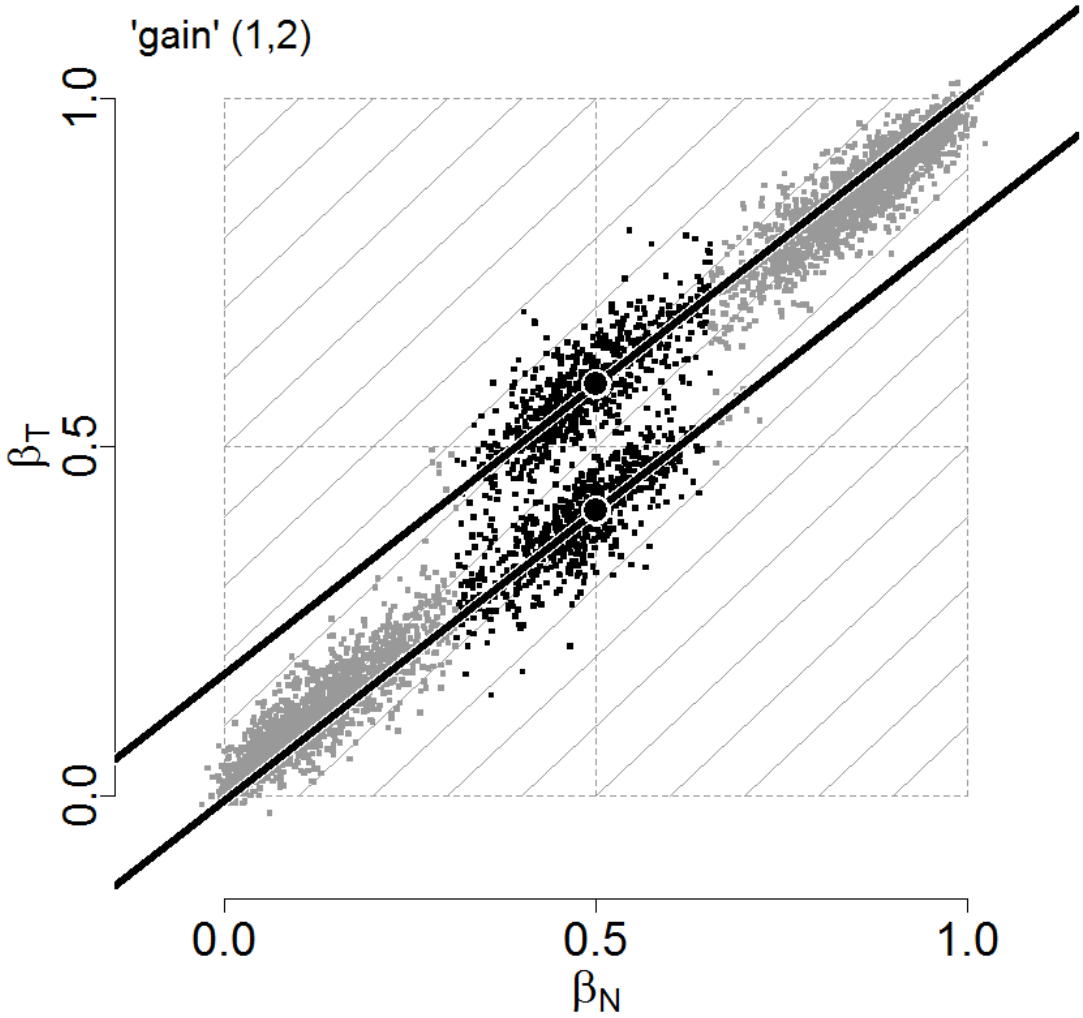
SNP effect in a region of no CN change in the tumor



- Expected: (0,0), $(\frac{1}{2}, \frac{1}{2})$, (1,1)
- Observed: elongated clusters

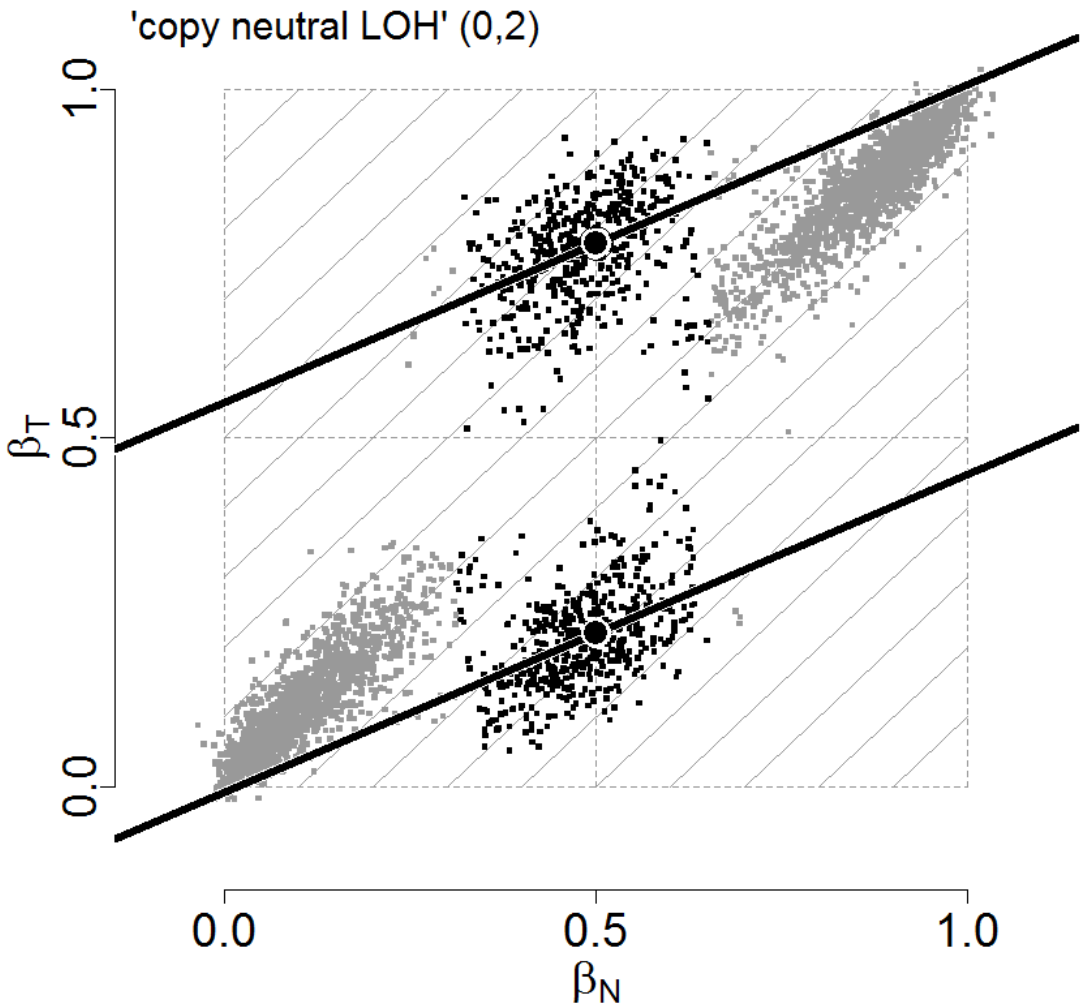
Deviation : a *SNP effect*, quite reproducible between the normal and the tumor

SNP effect in a region where tumor has a gain



- Homozygous clusters are similar as before
- Heterozygous cluster is split in two, and tilted

SNP effect in a region where tumor is CNNLOH



- Homozygous clusters are similar as before
- Heterozygous cluster is even more tilted

Overview of the TumorBoost method

Idea

1. the SNP effect is reproducible between tumor and normal
2. truth is easy to infer in the normal: three genotypes AA, AB, BB.

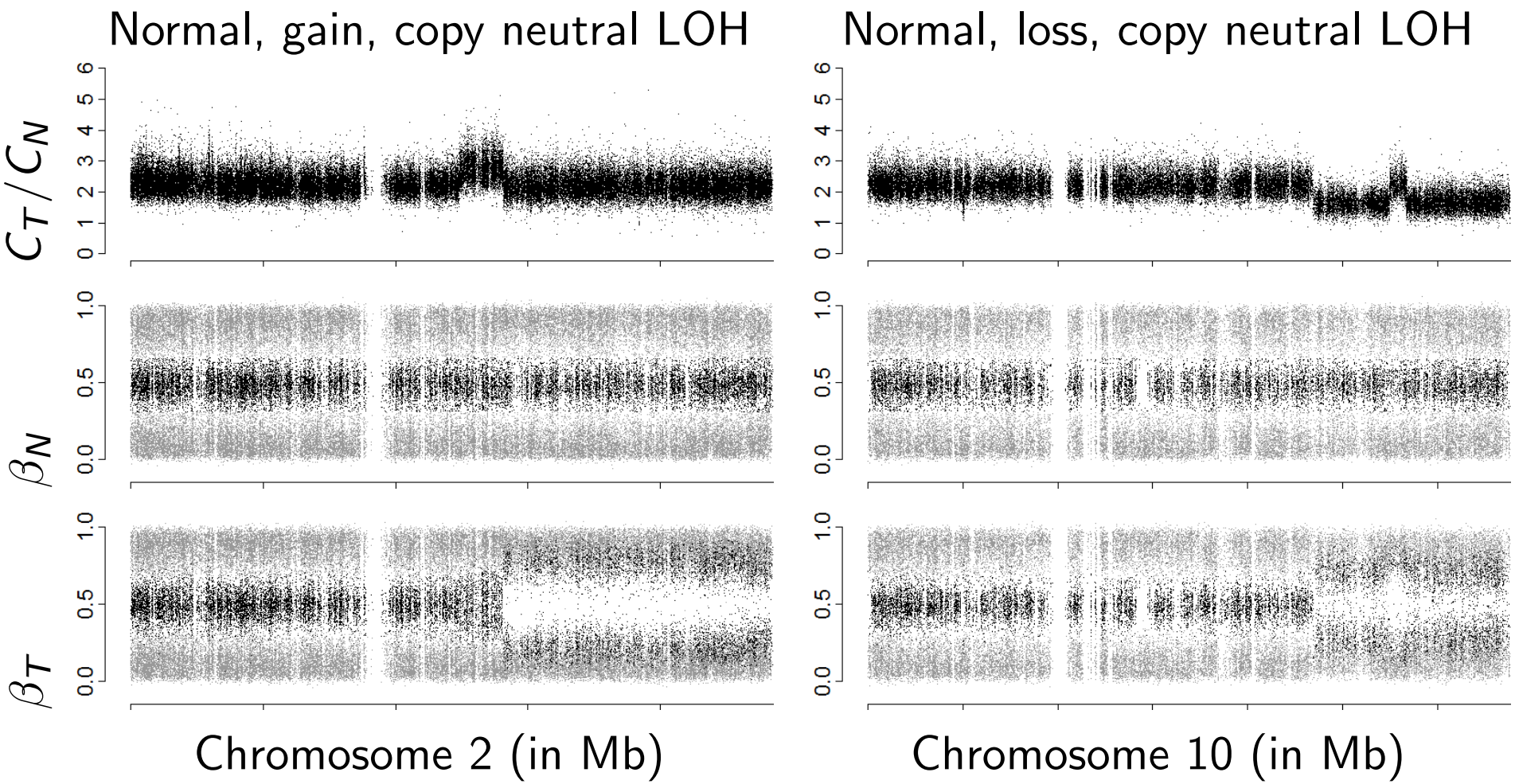
$\Rightarrow$ For each SNP, we estimate the SNP effect in the normal hybridization, and "subtract" it from the tumor.

Features

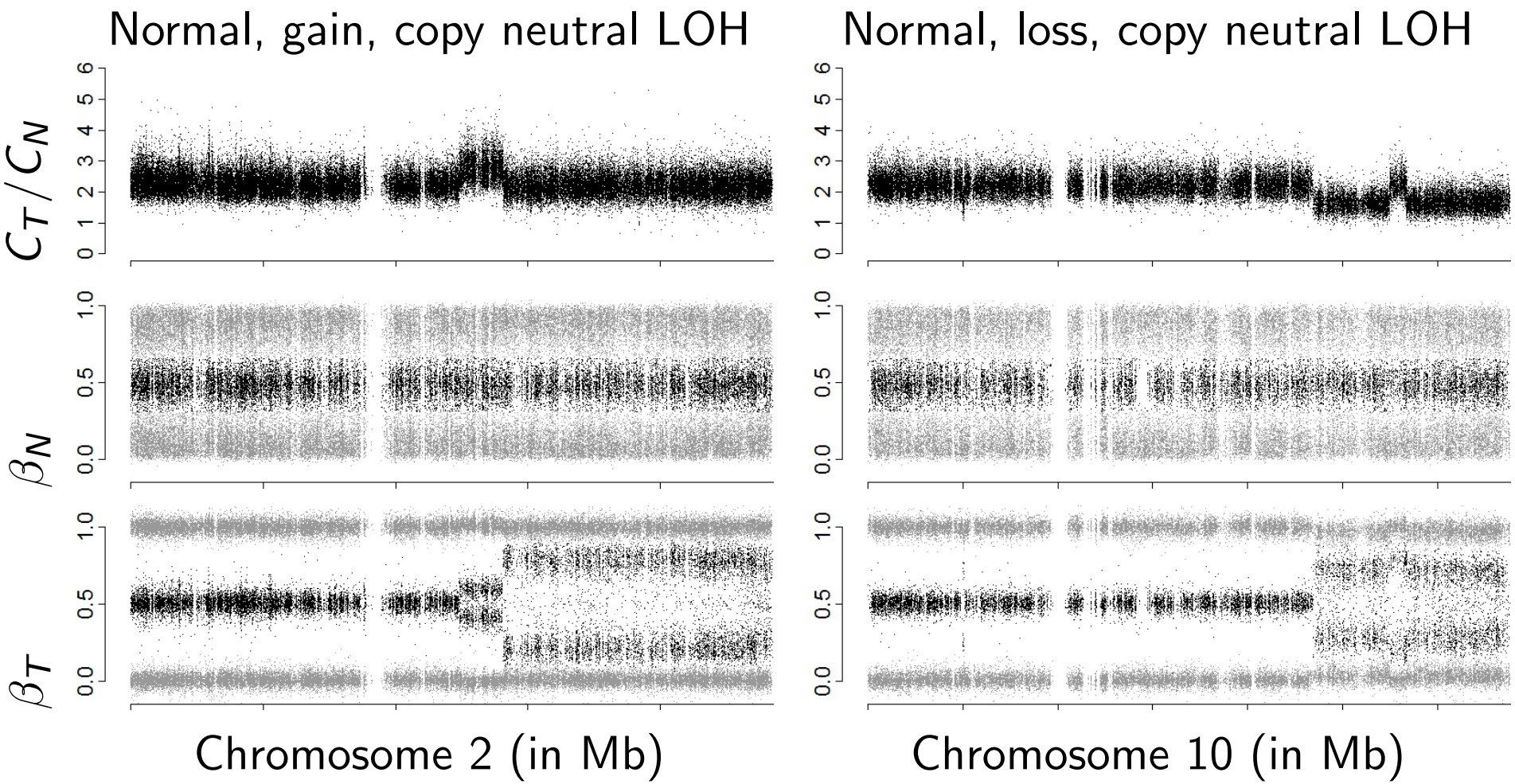
- we don't need to know copy number regions in advance
- normalization is performed for each SNP separately
- it only requires one tumor/normal pair

Outline

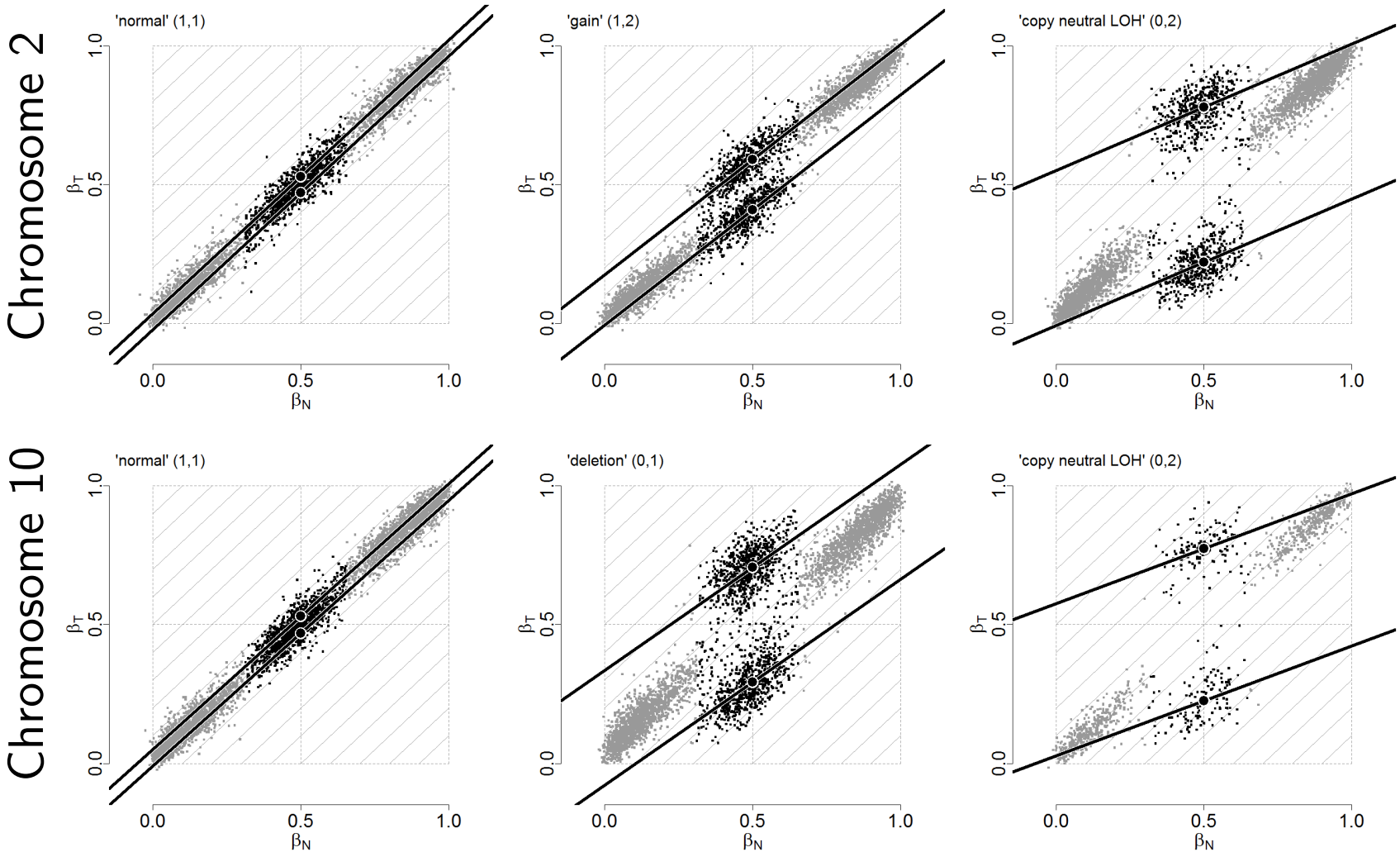
Genomic signals before normalization



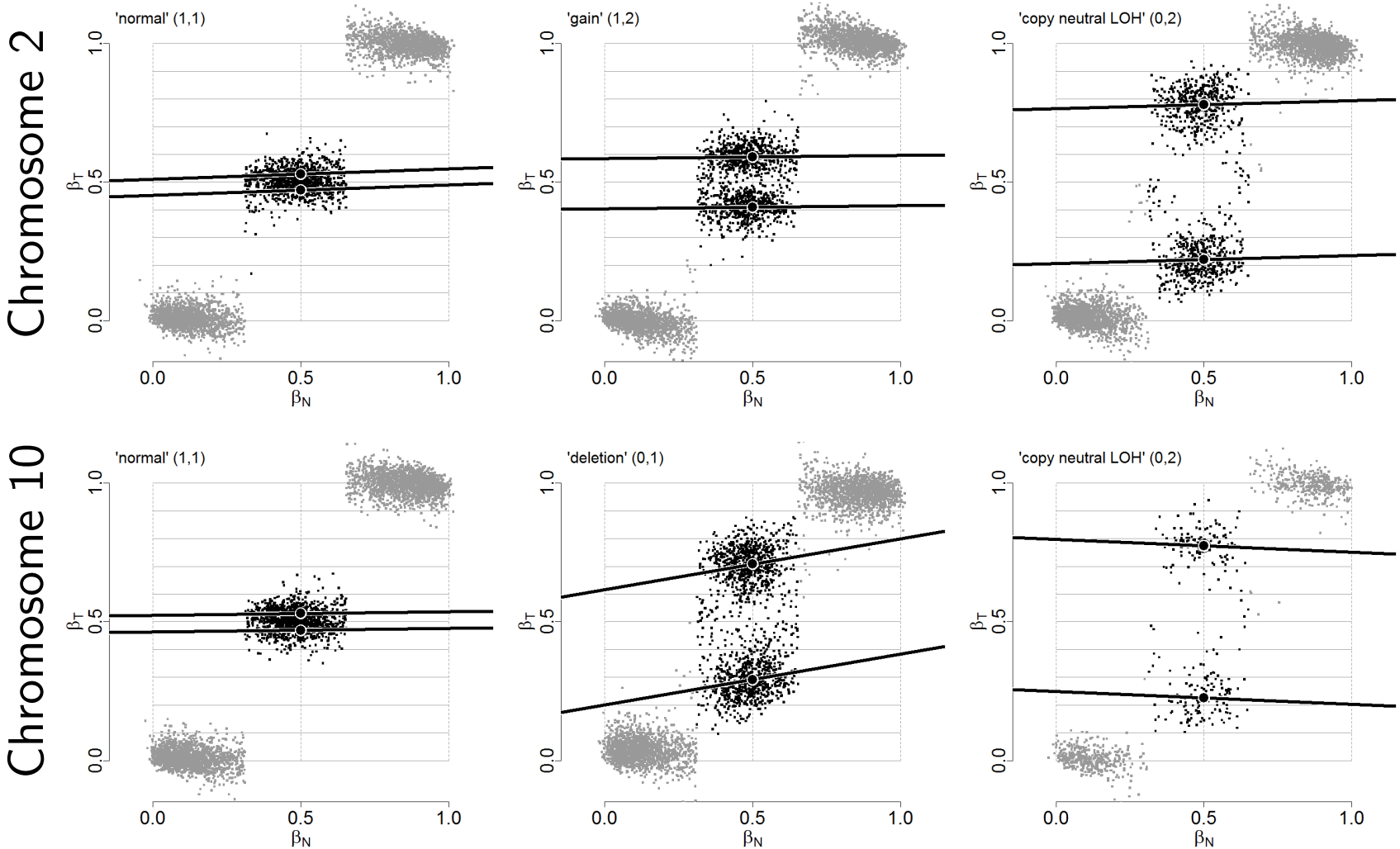
Genomic signals after normalization



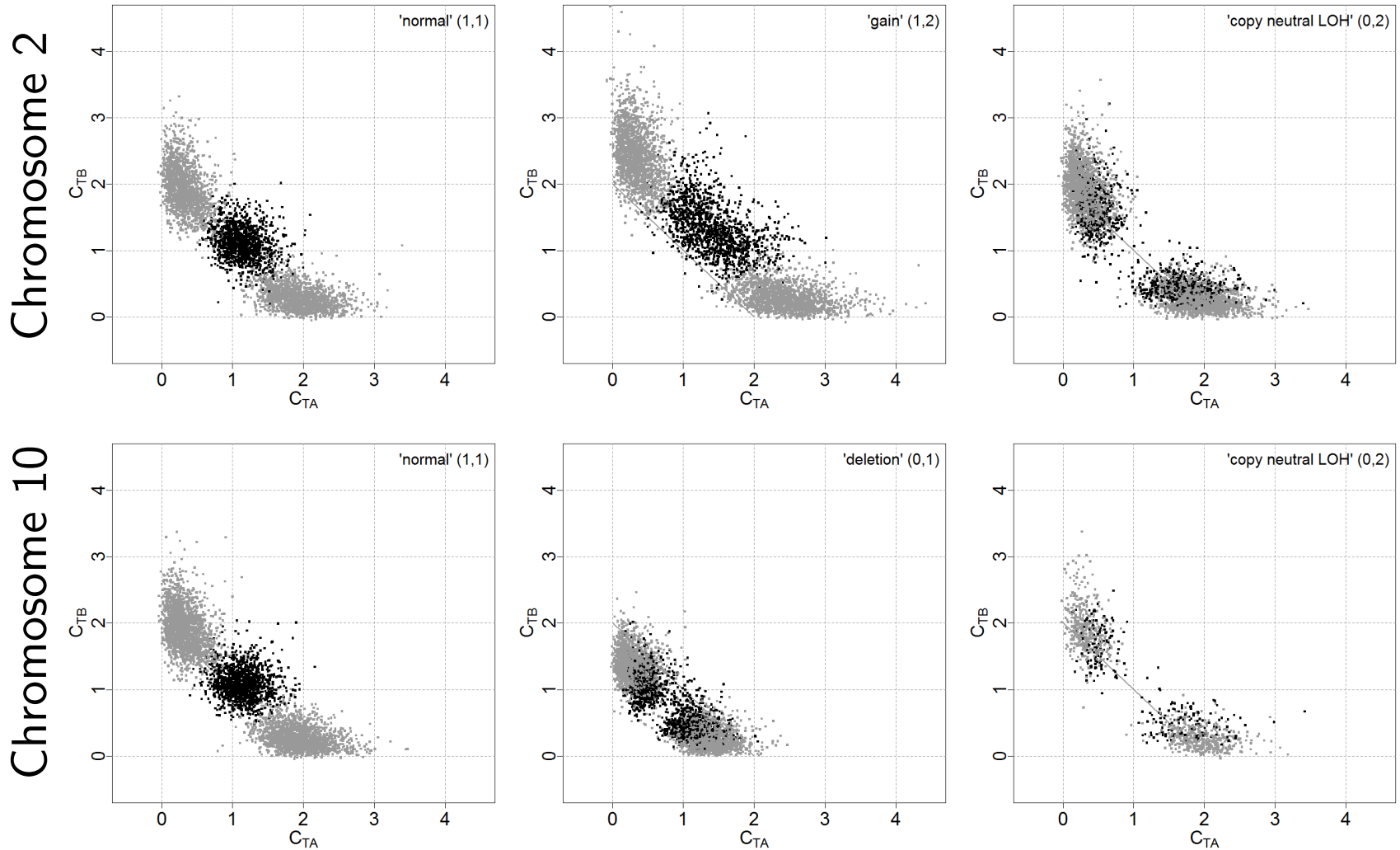
Allele B fractions before normalization



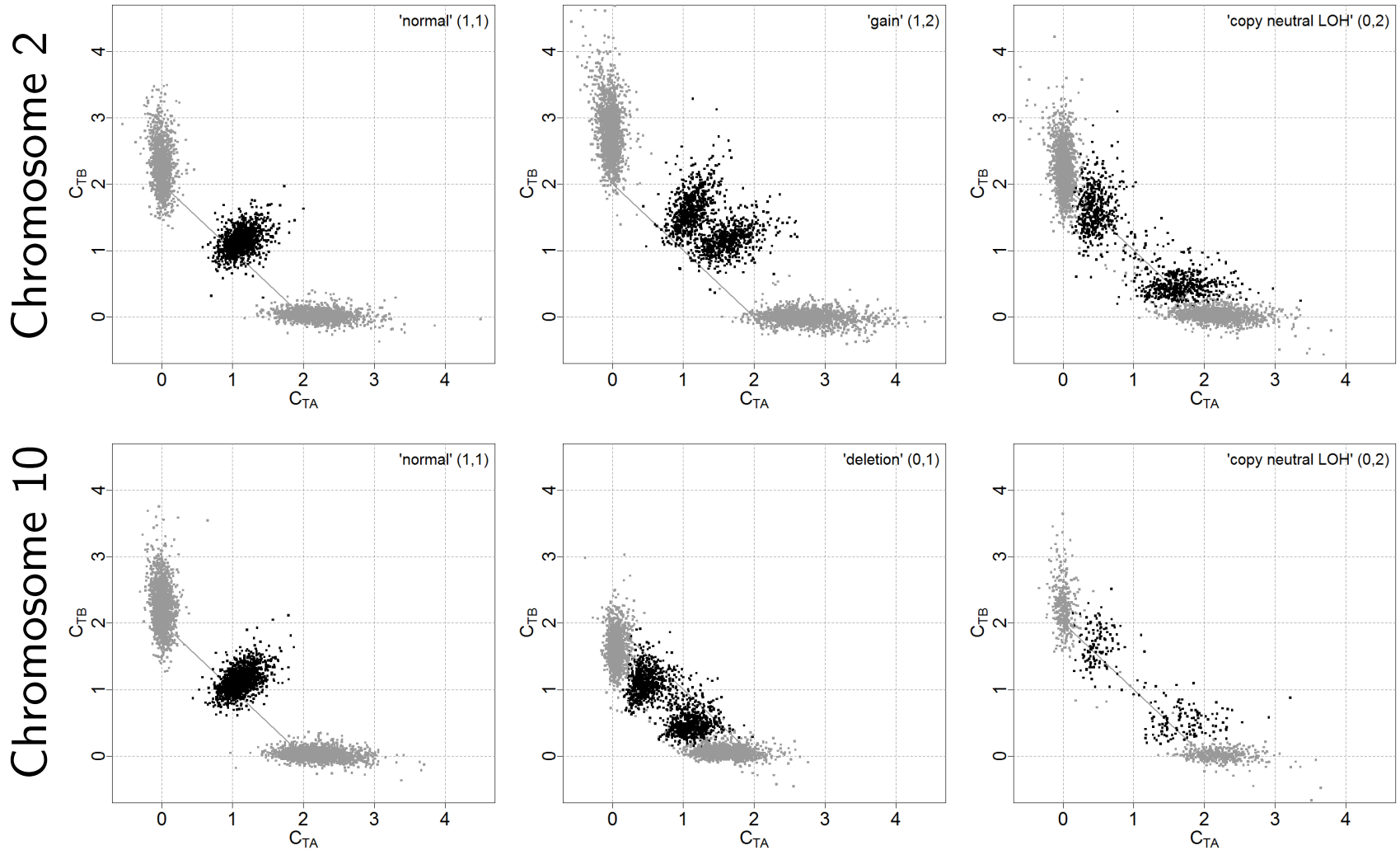
Allele B fractions after normalization



ASCNs before normalization



ASCNs after normalization



Complete preprocessing for a single tumor/normal pair

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

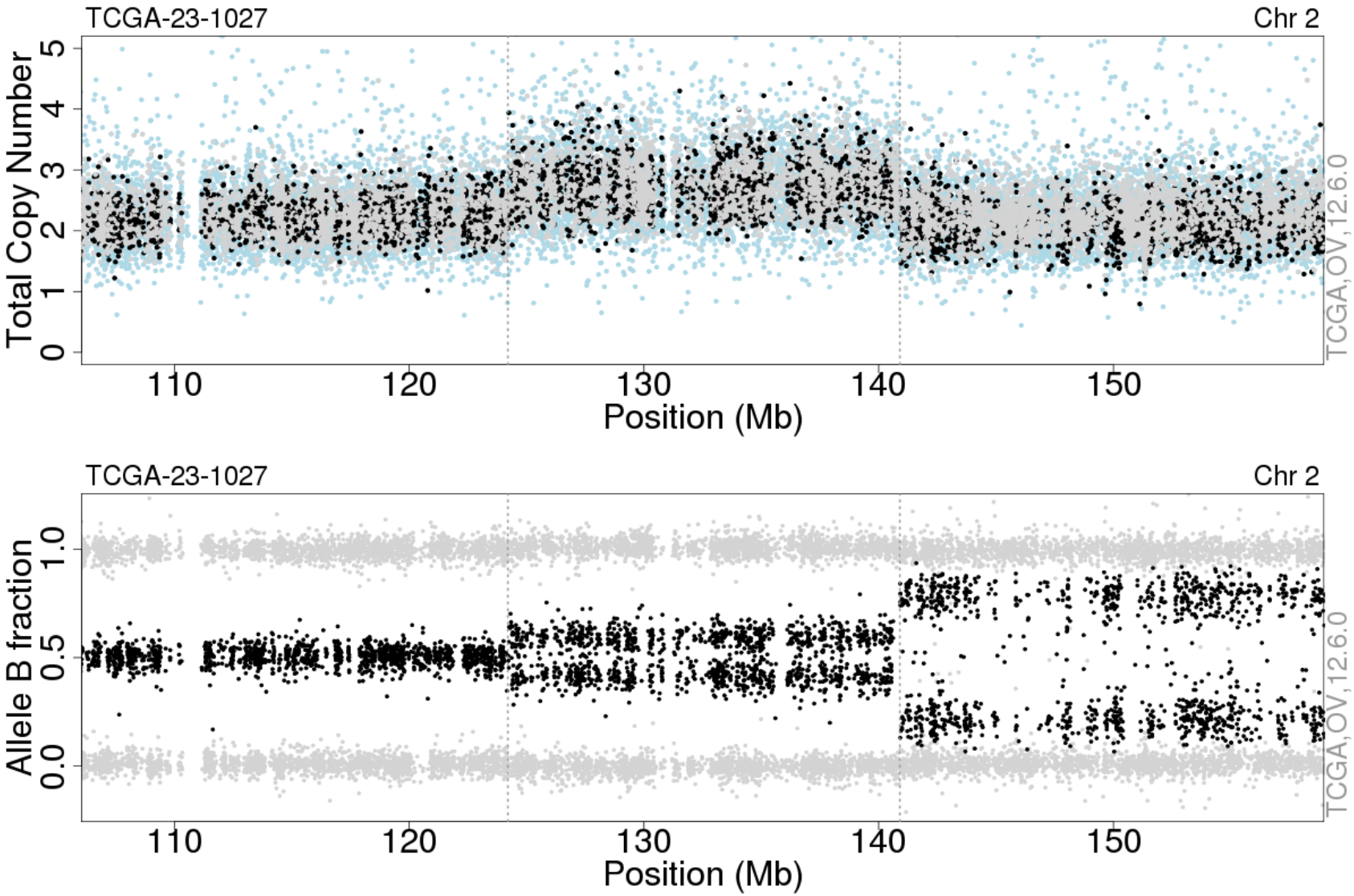
1. normalization and locus-level summarization using CRMAv2 (Bengtsson et al, 2009) for the normal and the tumor sample separately
2. naive genotyping of the normal sample: thresholding the density of β
3. TumorBoost normalization (Bengtsson et al, 2010)

Outline

Outline

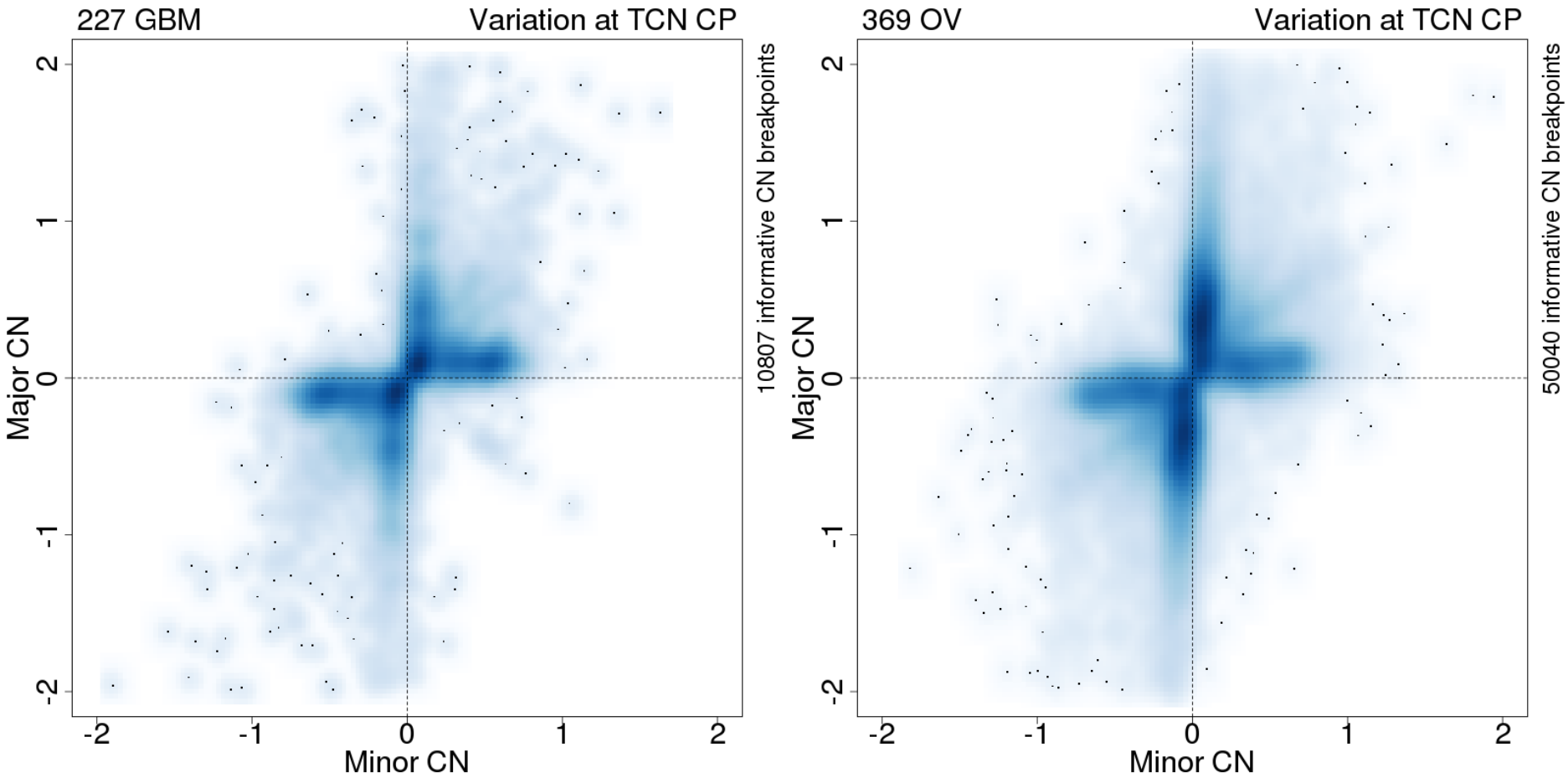
Changes often occur in either minor or major, not both

Looking at one sample



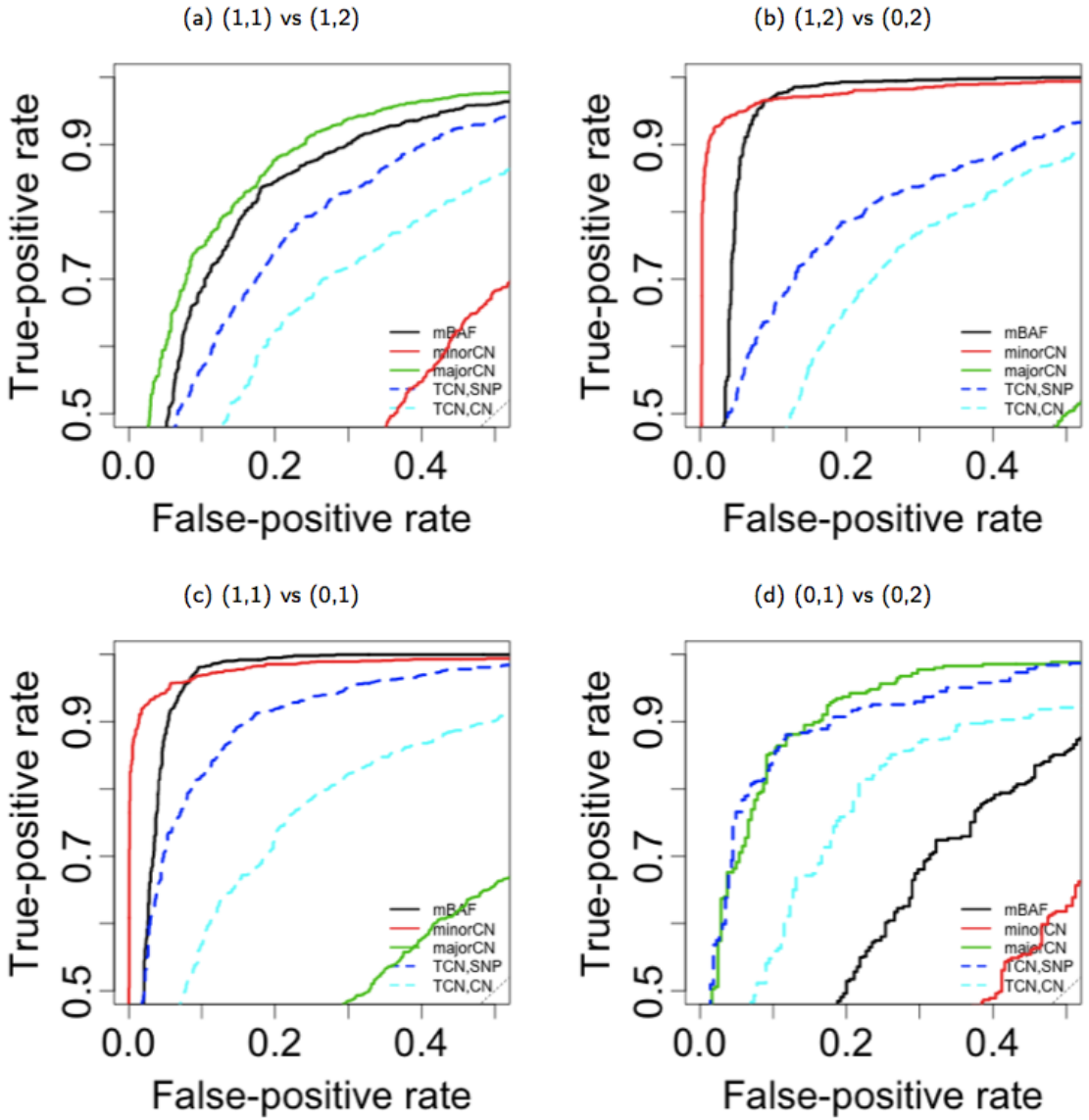
Changes often occur in either minor or major, not both

Looking across samples



Outline

AR has greater detection power than TCN at a single locus



More informative probes for TCN than AR

Affymetrix GenomeWideSNP_6

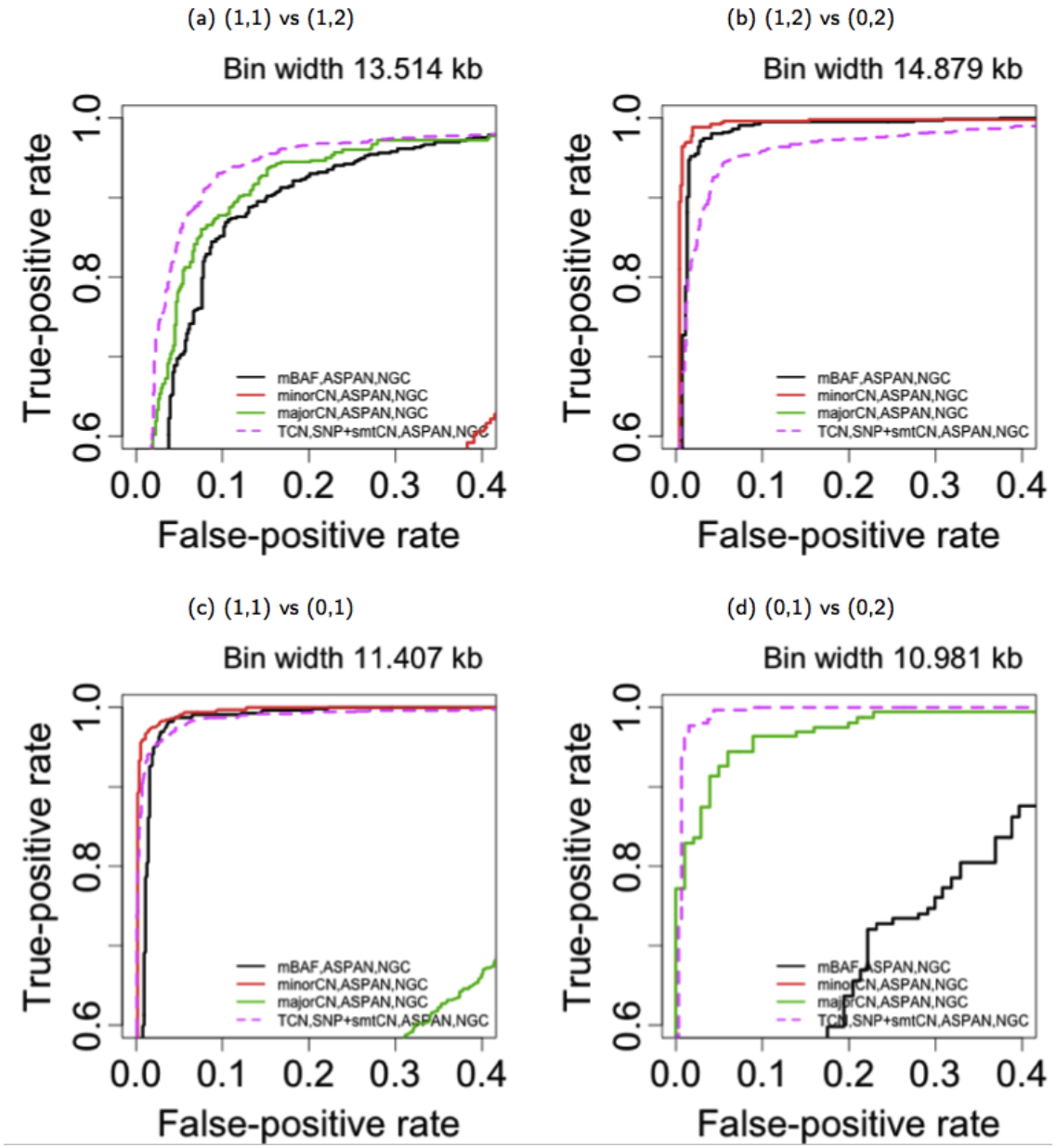
	All units	CN units	SNP units
Frequency	1,856,069	946,705	909,364
Proportion	100%	51%	49%

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

Rejoinder: similar detection power at a fixed resolution



The need for a truly two-dimensional segmentation method

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

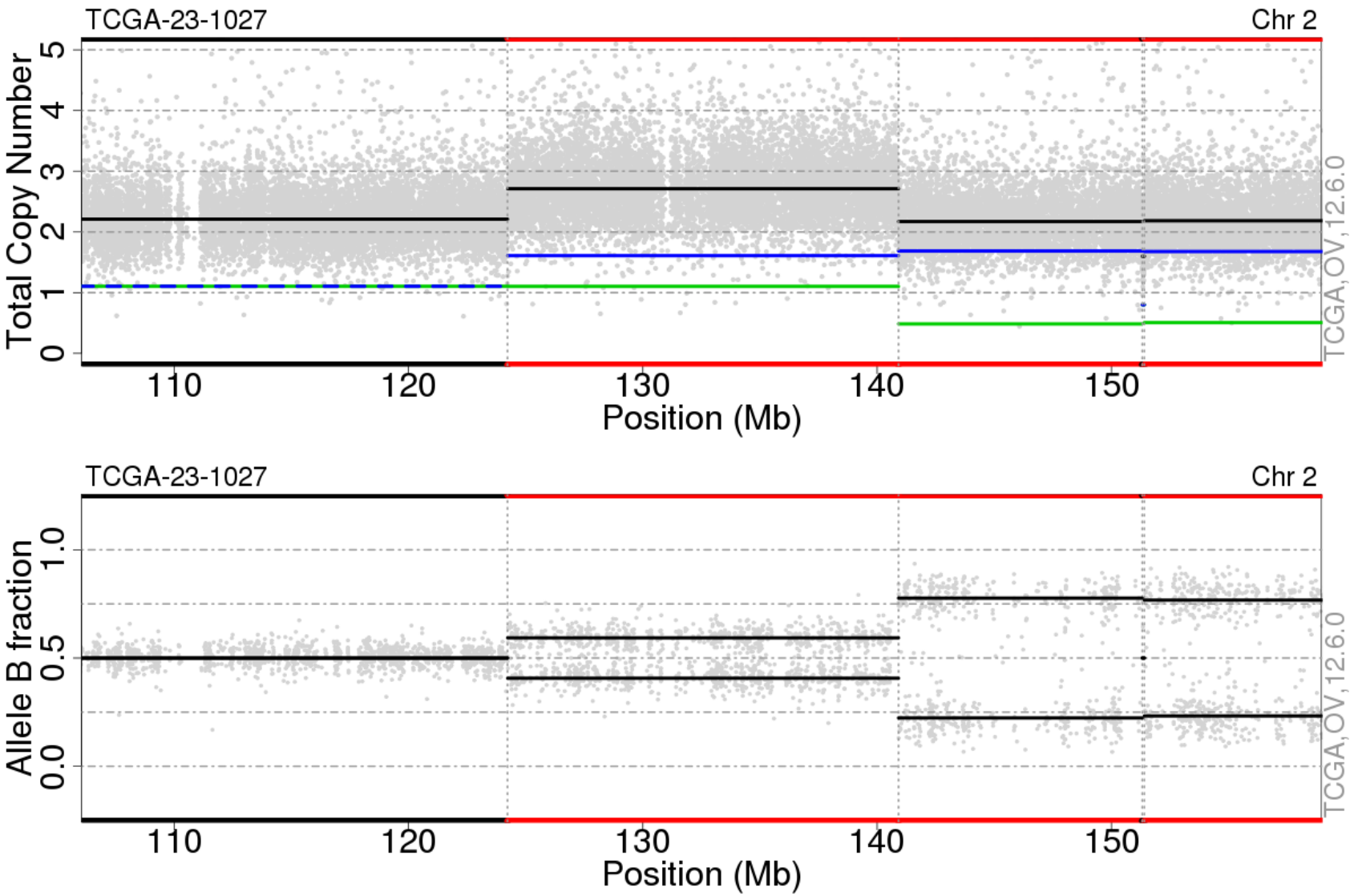
Challenges for a truly 2d segmentation method

- A two-dimensional signal
- Only heterozygous SNPs can be used to detect CN changes from AR
- Bias in the estimation of allelic imbalances
- AR are not Gaussian

Outline

Outline

Copy numbers are not calibrated



What you get isn't quite what you want.

Purity, ploidy, and a scaling factor

Why copy numbers are not calibrated

- non purity: presence of normal cells in the "tumor sample"
- ploidy: the total amount of DNA is fixed by the assay
- a scaling factor: the previous point is not quite true in practice

$$C_{ij} = \frac{\eta_i}{\lambda_i}\phi_j\gamma_{ij} + \varepsilon_{ij}$$

- hybridization i, probe j
- ϕ_j : affinity of probe j
- η_i: scaling factor
- λ_i: ploidy
- γ_{ij}: true copy number for (i, j)
- ε_{ij}: error term

A model

For a tumor/normal pair:

$$\begin{cases} C_{Nj} &= \frac{\eta_N}{\lambda_N}\phi_j\gamma_{Nj} + \varepsilon_{Nj} \\ C_{Tj} &= \frac{\eta_T}{\lambda_T}\phi_j\gamma_{Tj} + \varepsilon_{Tj} \end{cases}$$

Assuming a fraction κ of normal cells in the "tumor sample",

$$\gamma_{Tj} = (1-\kappa)\gamma^{\star}_{Tj} + \kappa\gamma_{Nj}$$

where $\gamma^{\star}_{Tj}$ is the number of copies of pure tumor. To cancel probe affinities (unknown), we usually work with $\hat{\gamma}_{Tj} = 2C_{Tj}/C_{Nj}$:

$$\hat{\gamma}_{Tj} = \frac{\eta_T}{\eta_N} \cdot \frac{\lambda_N}{\lambda_T}\left(2(1-\kappa)\frac{\gamma^{\star}_{Tj}}{\gamma_{Nj}} + 2\kappa\right)$$

Outline

What can we estimate ?

Assuming $\gamma_{Nj} = 2$ we get

$$\hat{\gamma}_{Tj} = \frac{\eta}{\lambda}\left((1-\kappa)\gamma^{\star}_{Tj} + 2\kappa\right)$$

where $\eta = \frac{\eta_T}{\eta_N}$ and $\lambda = \frac{\lambda_T}{\lambda_N}$.

- we can estimate η by comparing the average genome-wide total copy number over to 1.
- purity influences the absolute difference between successive CN
- ploidy influences the global scale

For ploidy and purity we need more assumptions.

Existing methods typically assume no normal contamination: [OverUnder], [PICNIC] or diploidy: [genoCNA]. [GAP] deals with both.

Estimating κ and λ

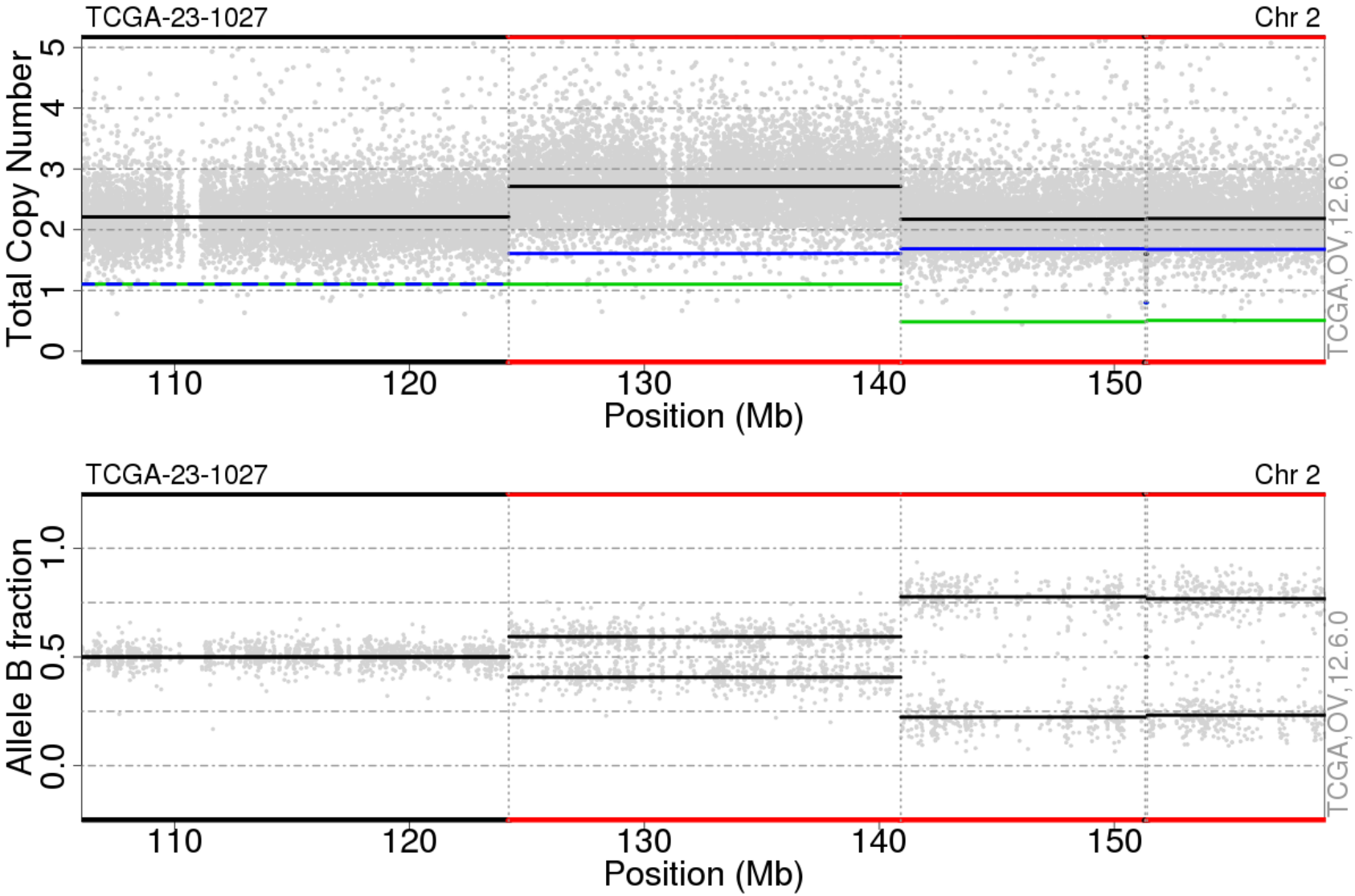
- Assuming most change points correspond to one unit of either major or minor CN, one can estimate

$$\frac{\eta}{\lambda}(1-\kappa)$$

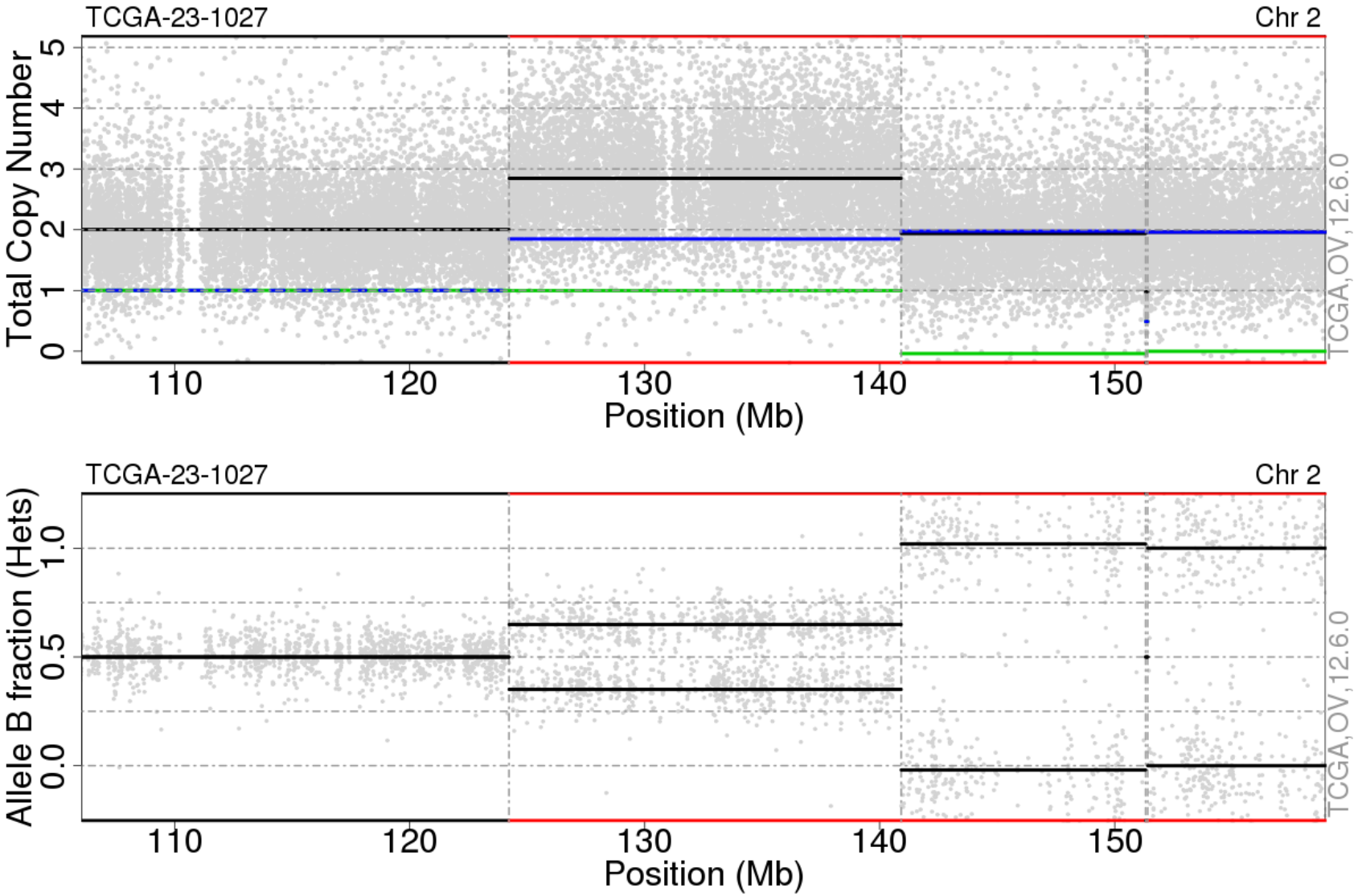
- Assuming that the mode of TCN with no allelic imbalance corresponds to the normal, one can estimate

$$\frac{2\eta}{\lambda}$$

Before calibration



After calibration



Issues

- we are making several assumptions to estimate κ and λ
- non linearity: $TCN = 0, 1, 2, 3, 4, ...$ are not equally well calibrated
- bias in the estimation of AI
- changes in the germline are not accounted for and could break our assumptions

Further thoughts

- calling change points before calling regions ?
- one of major and minor can be enforced to be constant

Thanks

- Henrik Bengtsson
- Terry Speed
- Nancy R. Zhang