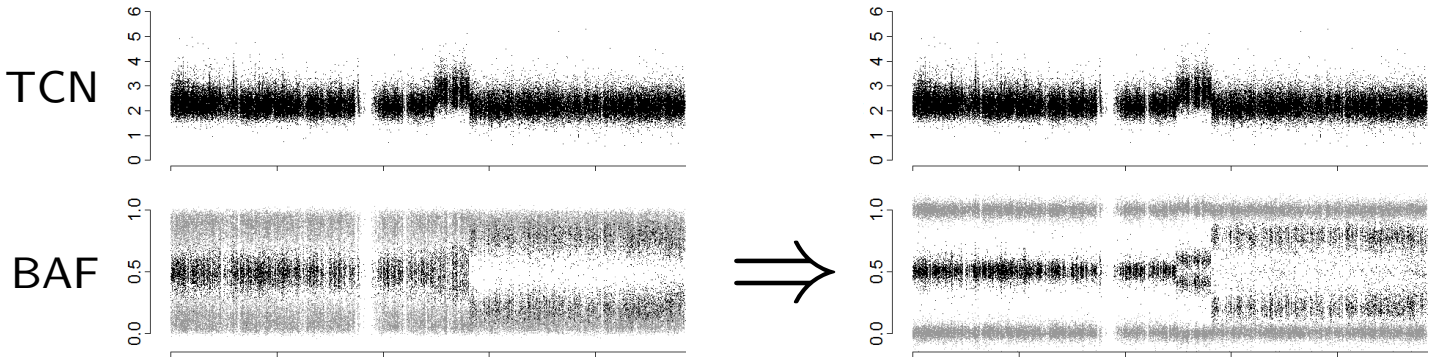
Greatly improved allele-specific tumor copy numbers with DNA microarrays when a matched normal is available

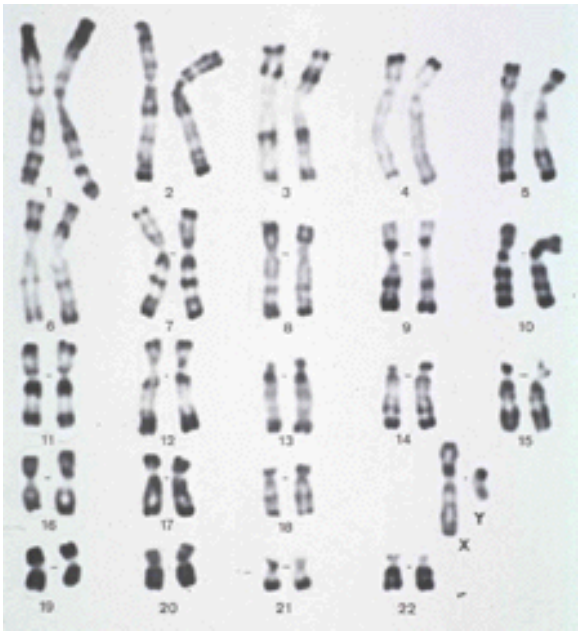
Pierre Neuvial, Henrik Bengtsson, Terry Speed

Department of Statistics, UC Berkeley

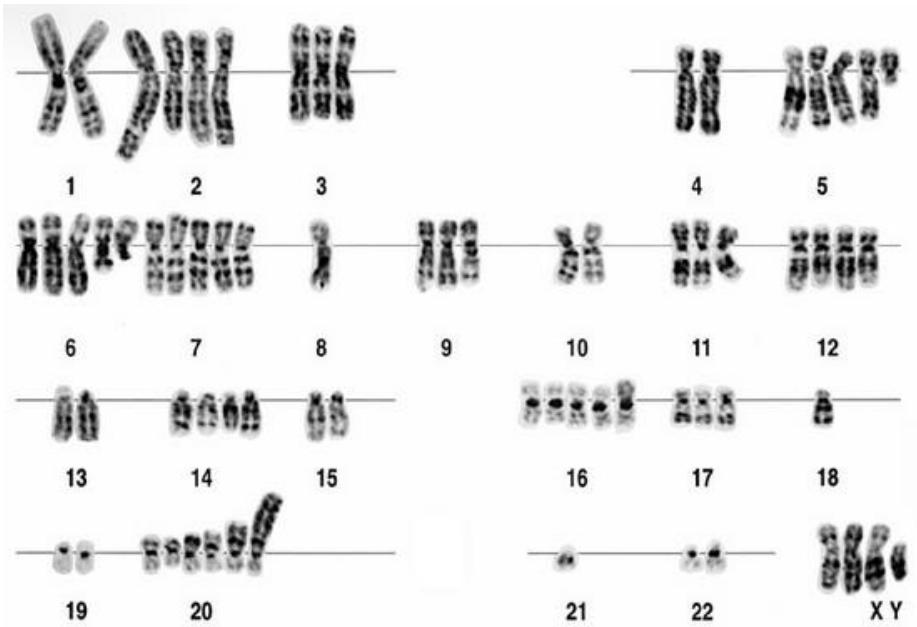


Genomic changes at the DNA level are hallmarks of cancer

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



Normal karyotype



Tumor karyotype

Goal: identify CN changes in cancer to improve their characterization, classification, and treatment

Tool: genotyping microarrays (aka SNP arrays)

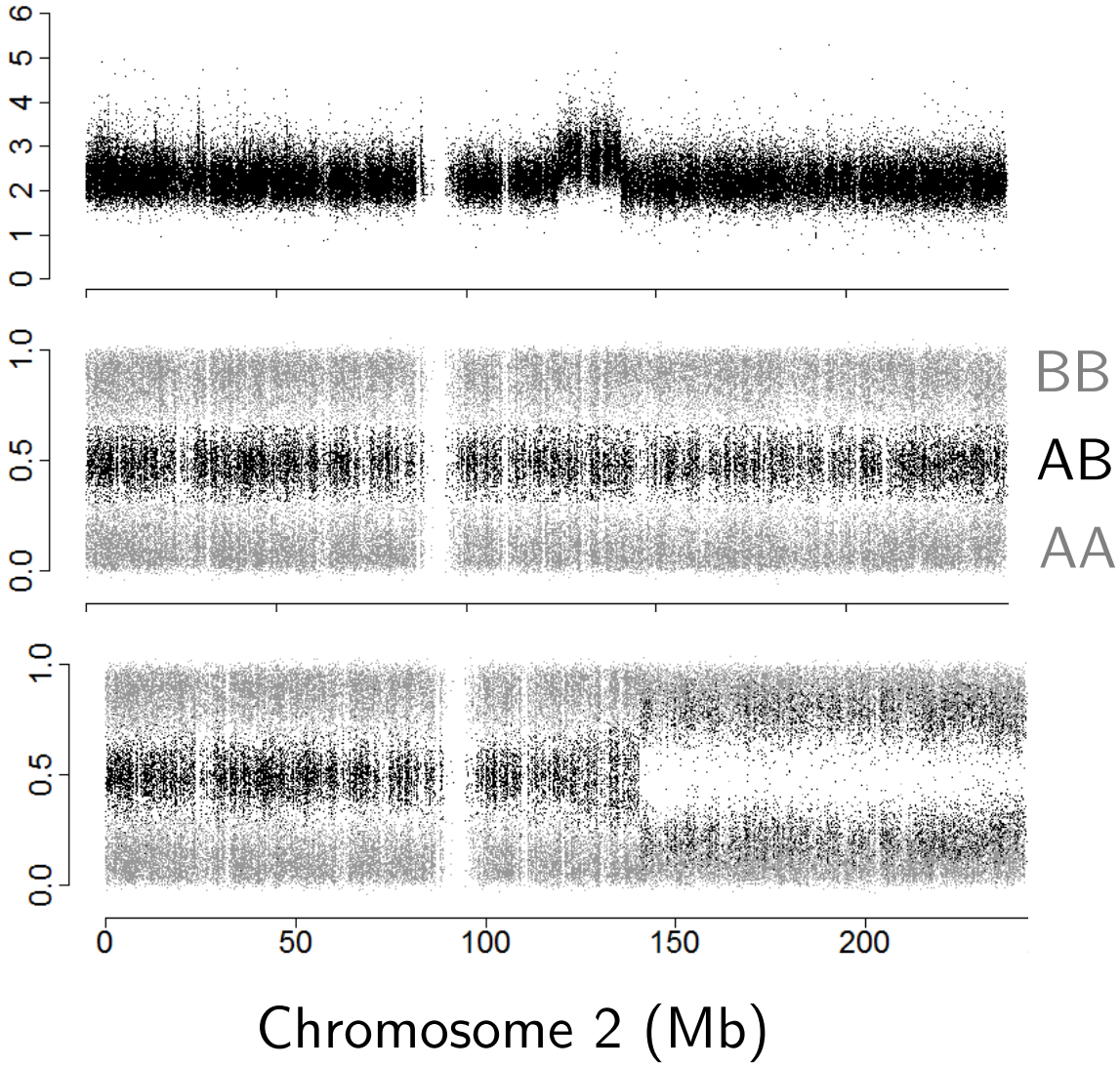
Allelic copy number measurements are noisy

Genomic signals from a tumor-normal pair of SNP arrays after preprocessing (CRMAv2)

Total copy number (TCN)

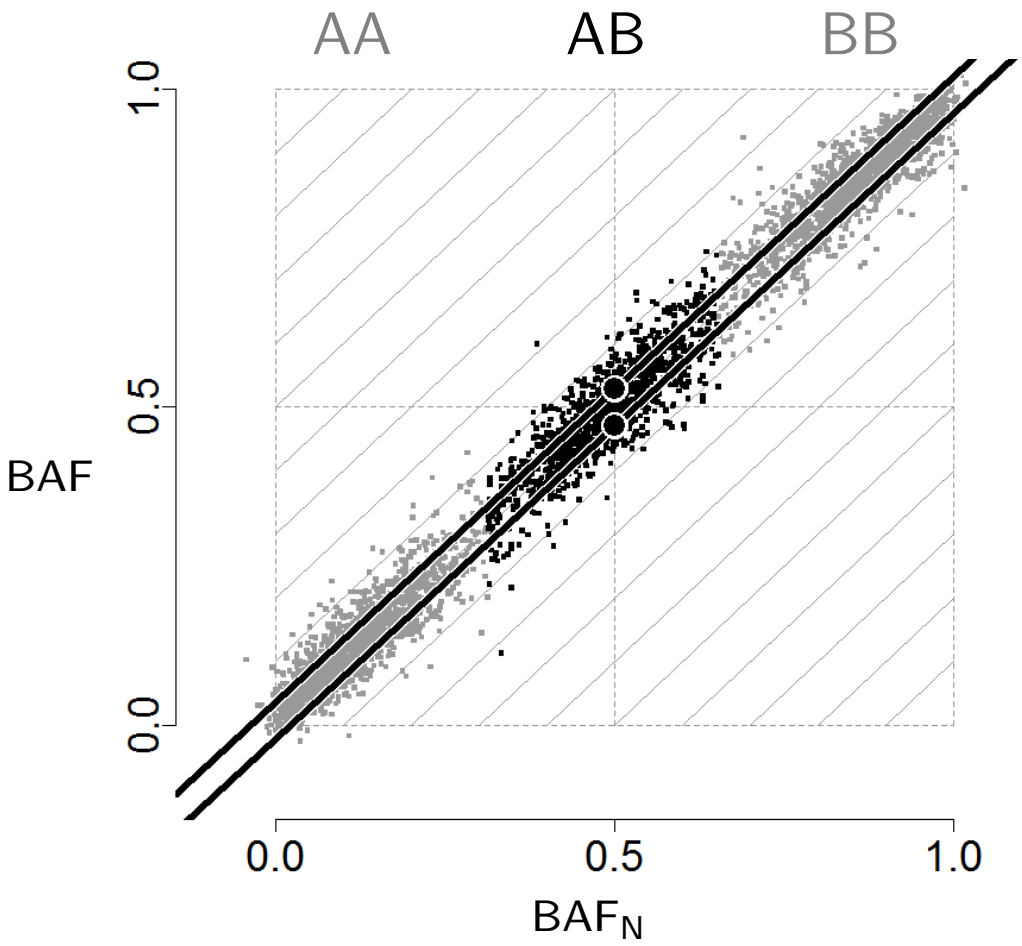
Allele B fraction, normal (BAF_N)

Allele B fraction, tumor (BAF)



"Noise" comes from a reproducible SNP effect

BAF: tumor vs normal in a region of no CN change in the tumor



BB

AB

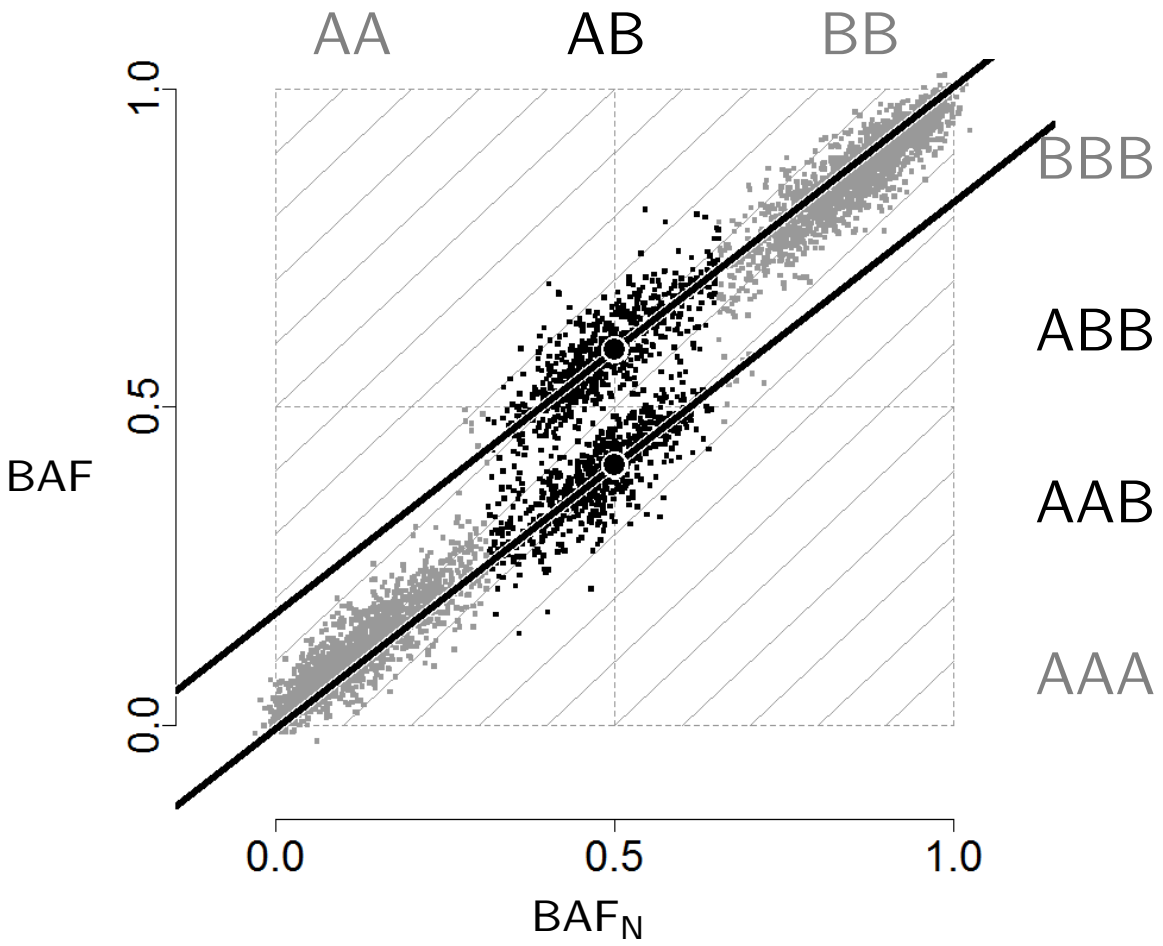
AA

- 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

“Noise” comes from a reproducible SNP effect

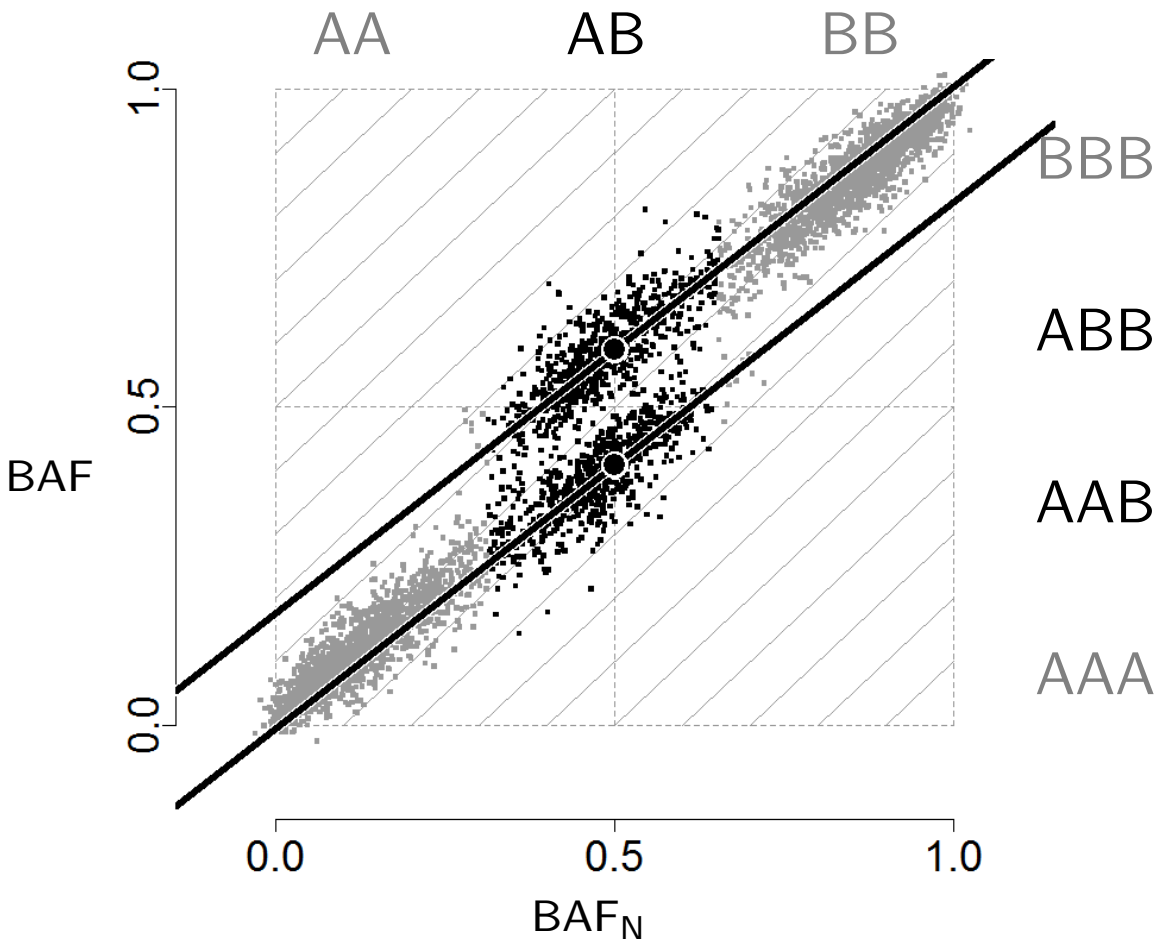
BAF: tumor vs normal in a region where the tumor has a gain



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

"Noise" comes from a reproducible SNP effect

BAF: tumor vs normal in a region where the tumor has a gain



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

In practice, copy number regions are not known in advance !

TumorBoost estimates and subtracts this "noise"

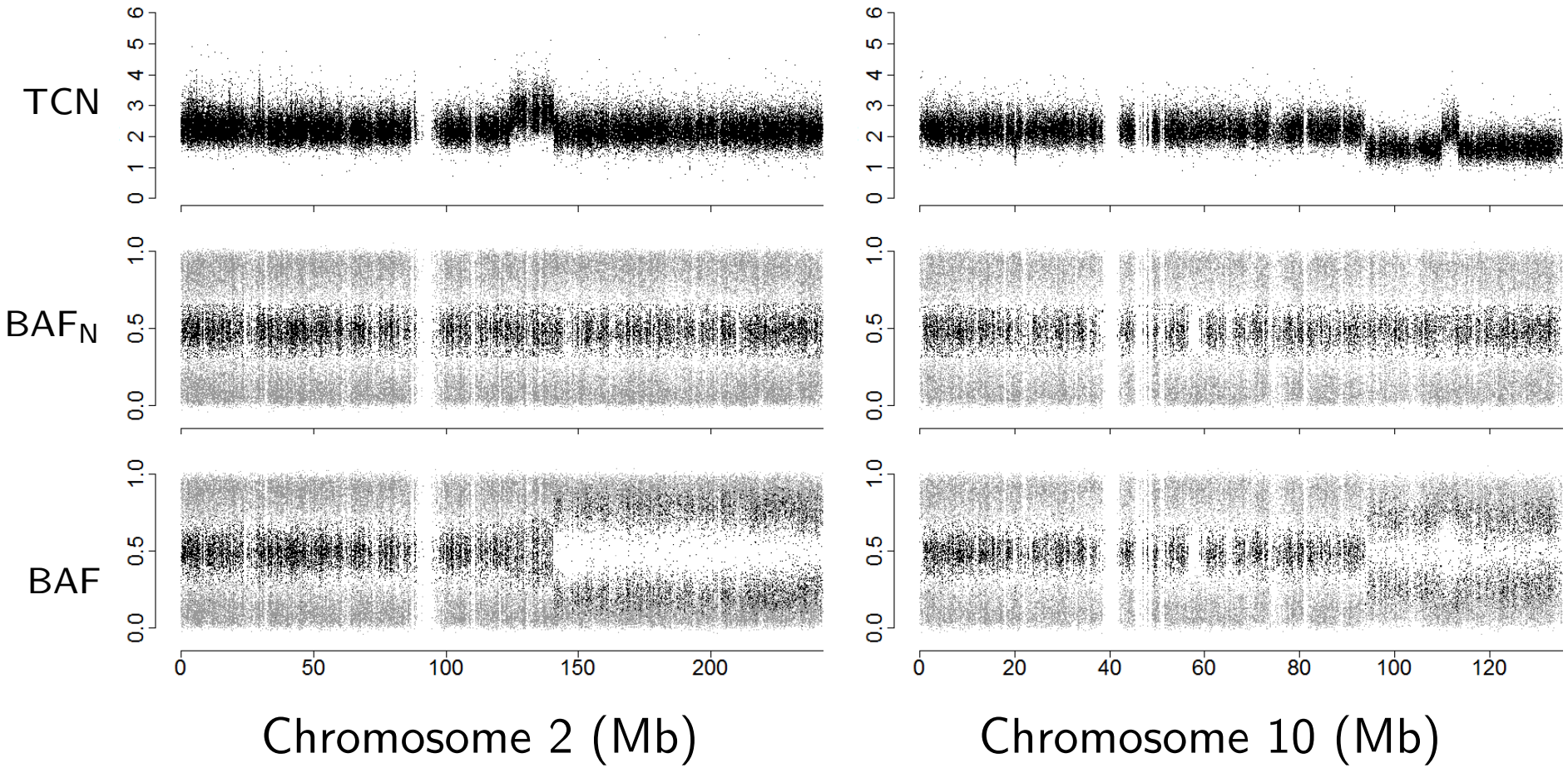
Idea

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

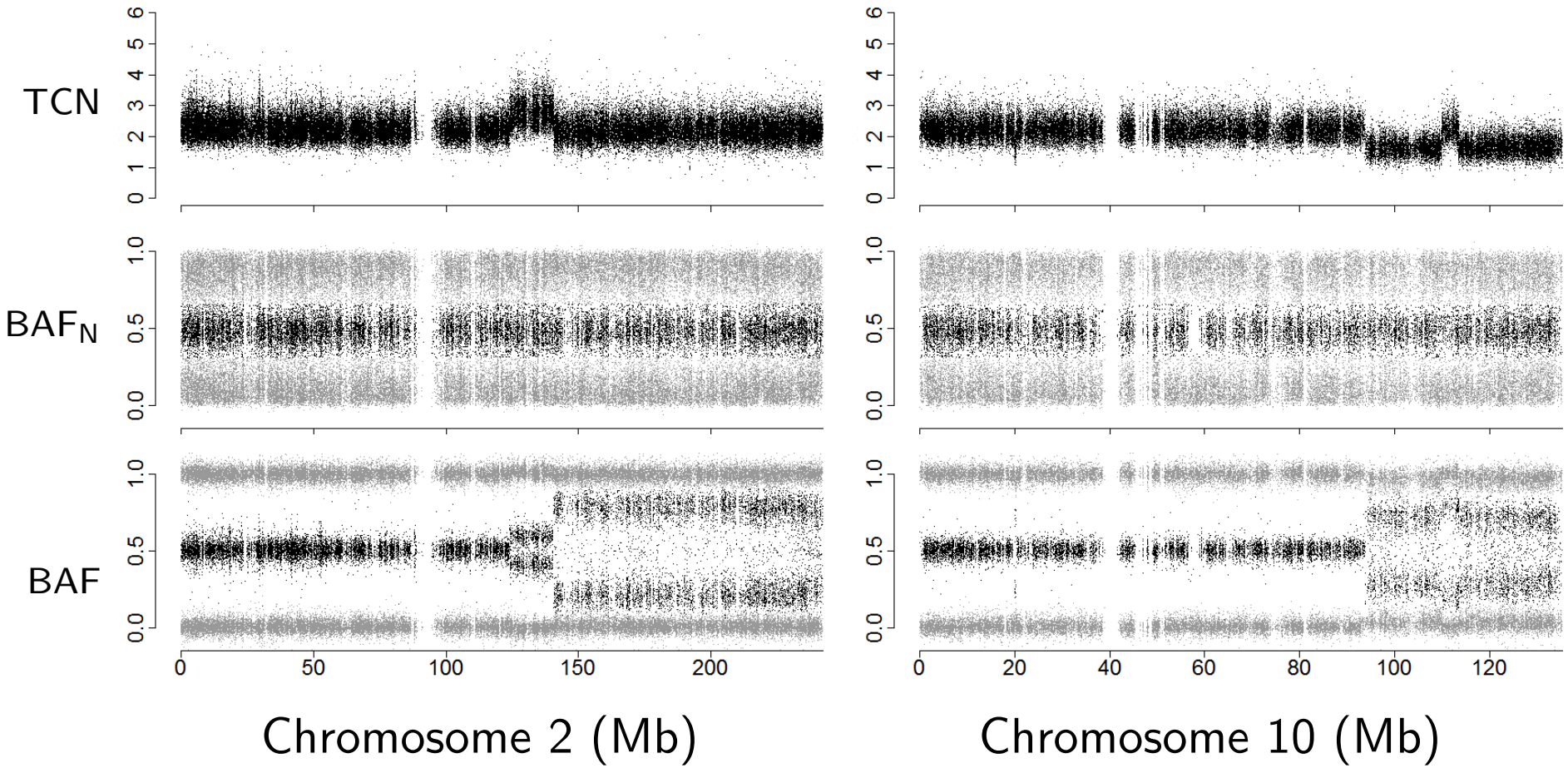
$\Rightarrow$ For each SNP, we estimate the SNP effect in the normal hybridization, and "subtract" it from the tumor.

H. Bengtsson, P. Neuvial, T. P. Speed, *BMC Bioinformatics*, 2010, **11**:245

Genomic signals before normalization



Genomic signals after normalization

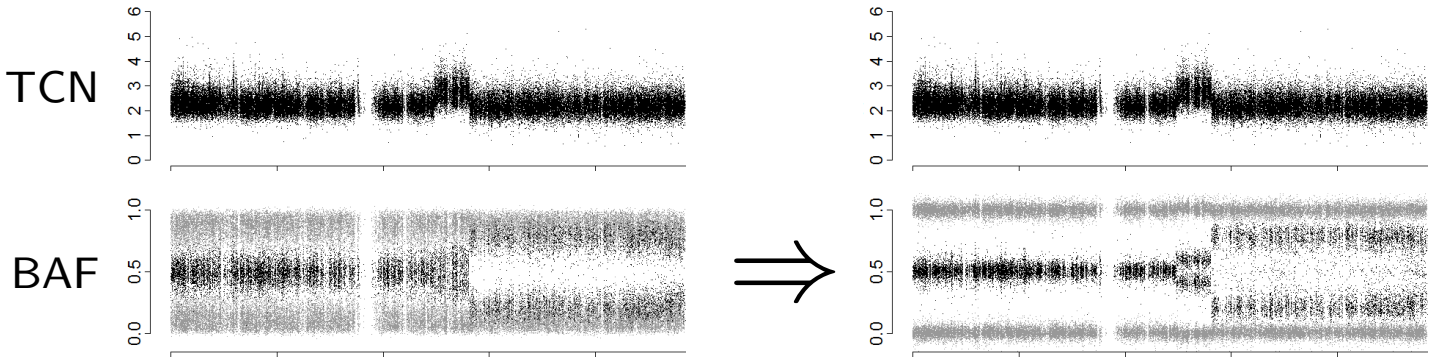


Conclusion

Features:

- no need to know copy number regions in advance
- normalization is performed for each SNP separately
- only requires one tumor/normal pair
- applicable to any SNP array technology (Affymetrix, Illumina)
- applicable after any preprocessing method

Results:

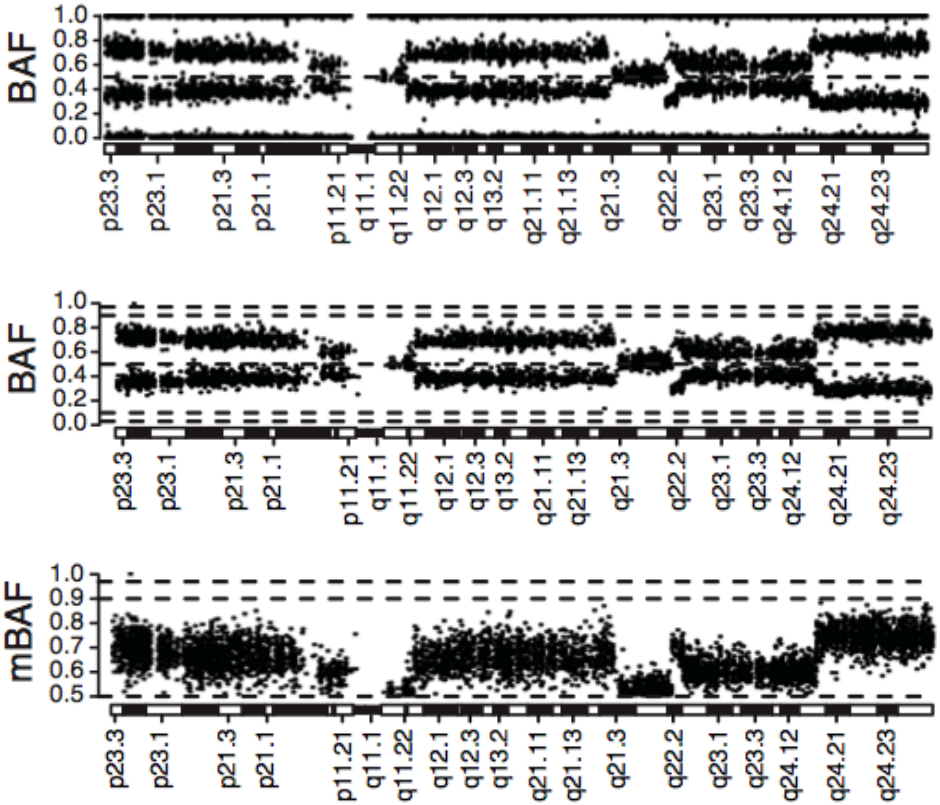


Open source R package: http://aroma-project.org/

Appendix

Detecting changes in allele B fractions

(Figure from Staaf *et al* (2006))



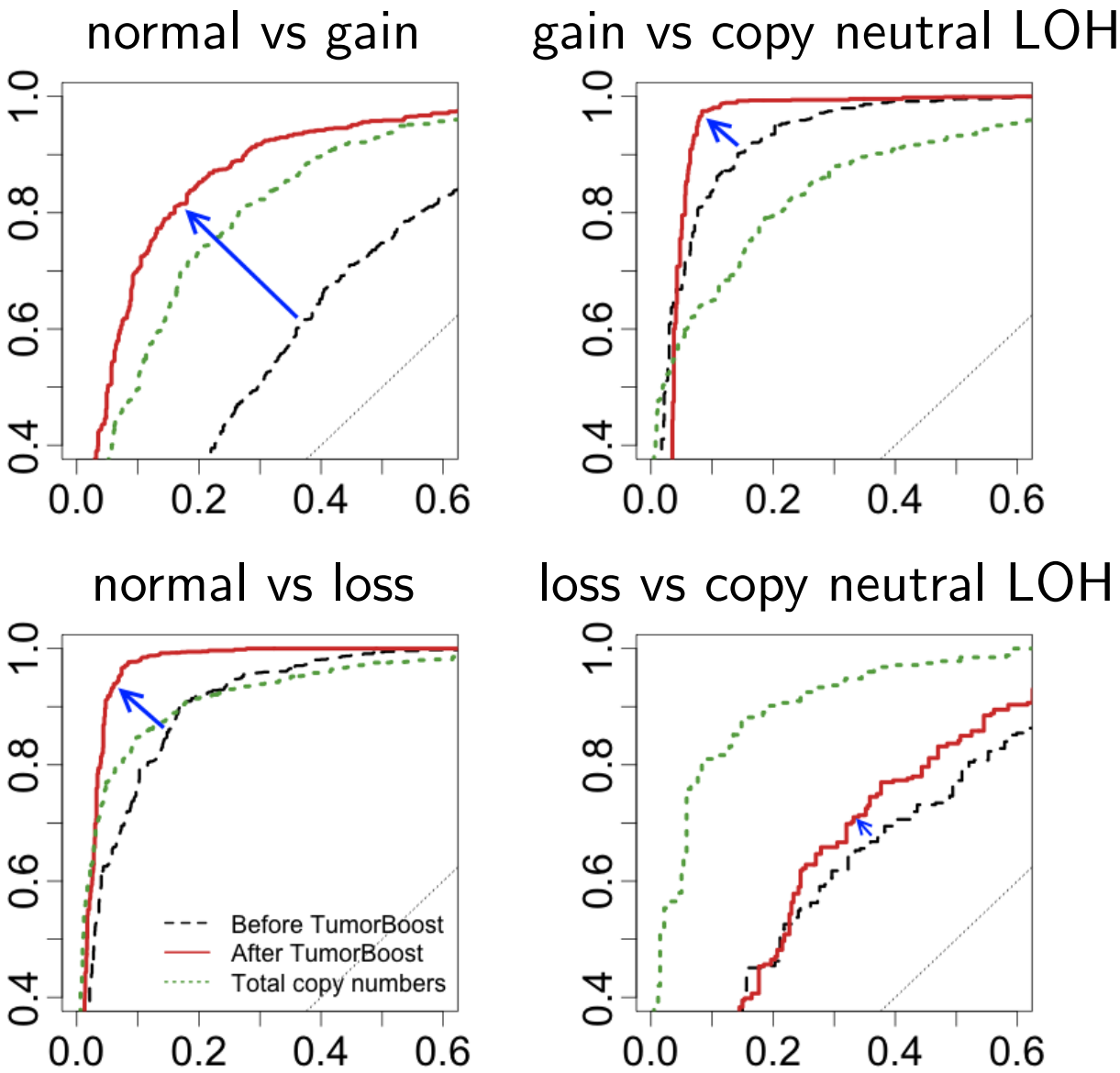
allele B fractions

allele B fractions for heterozygous SNPs

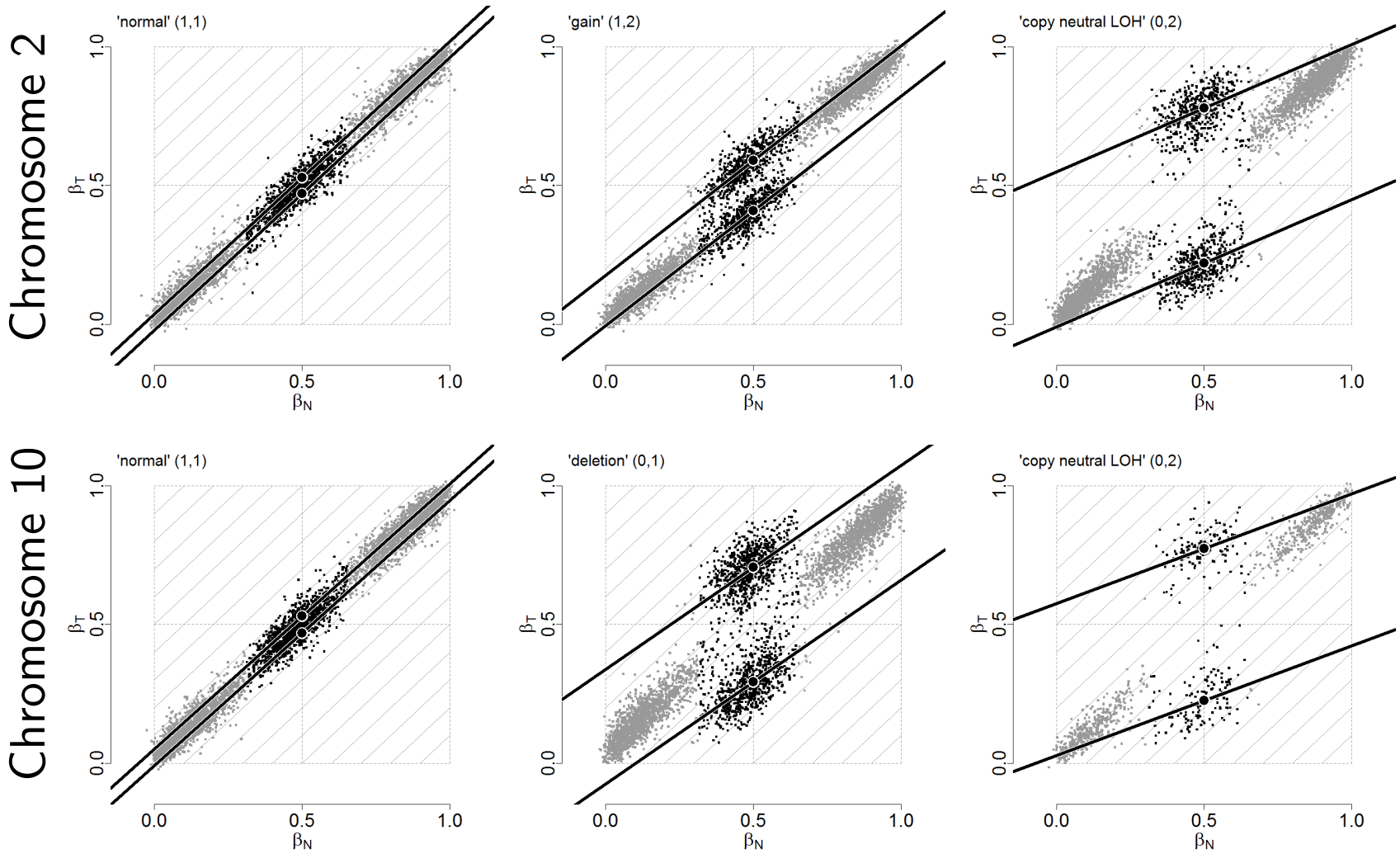
"mirrored" allele B fractions for heterozygous SNPs:
$mBAF = |BAF - 1/2|$

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

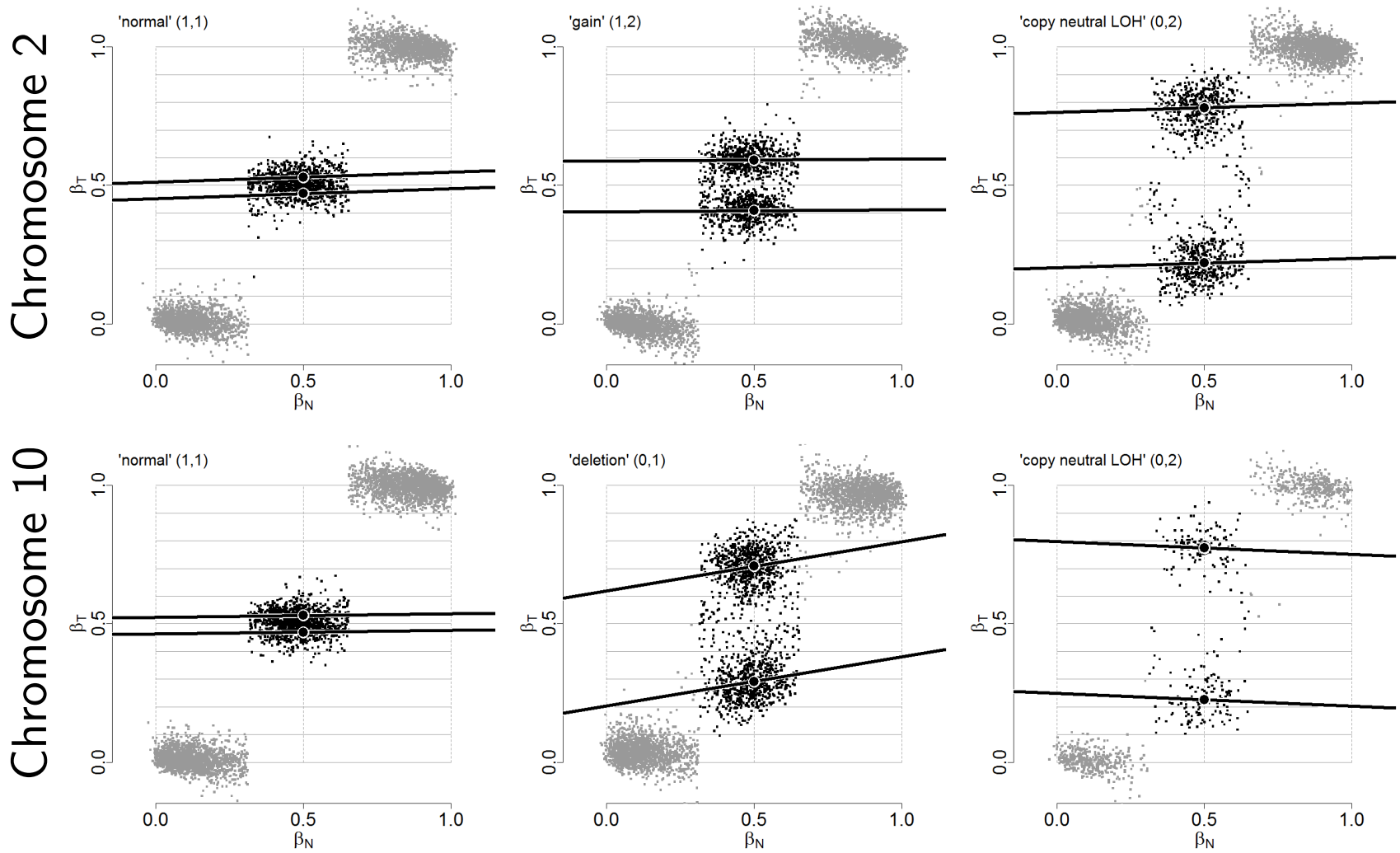
ROC analysis: TP rate vs FP rate



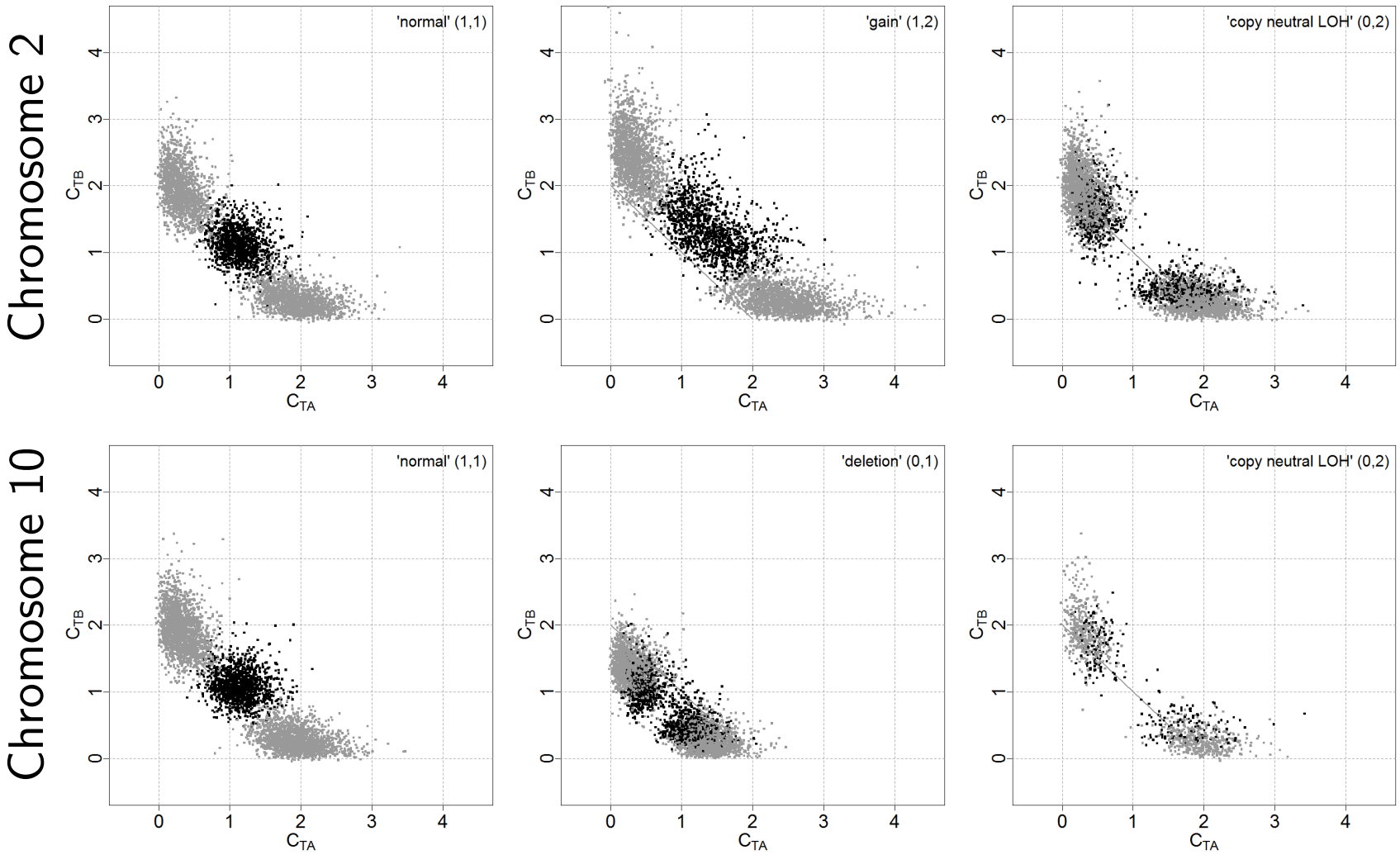
Allele B fractions before normalization: (BAF_T, BAF_N)



Allele B fractions after normalization: (BAF_T, BAF_N)

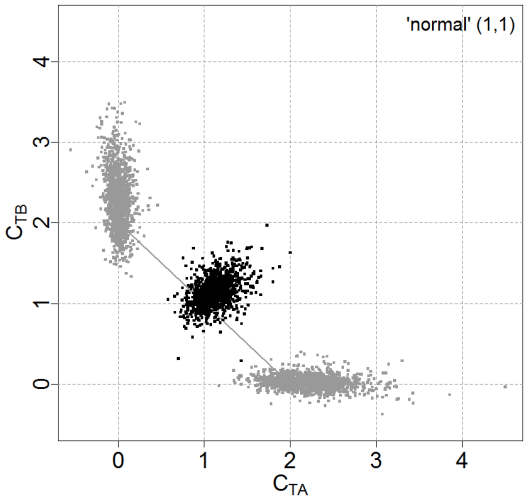


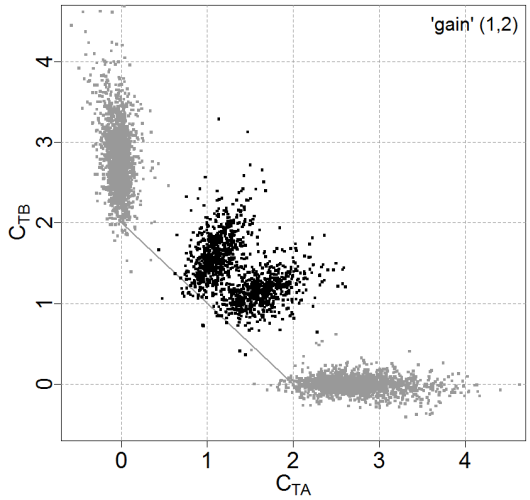
ASCNs before normalization: (C_{TB}, C_{TA})

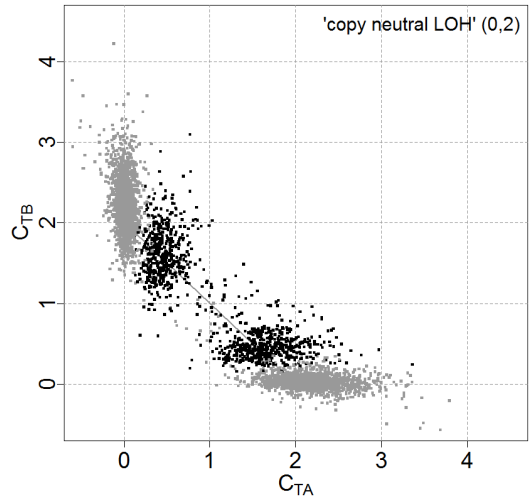


ASCNs after normalization: (C_{TB}, C_{TA})

Chromosome 2







Chromosome 10