Enhanced power for segmenting parent-specific copy numbers

Henrik Bengtsson

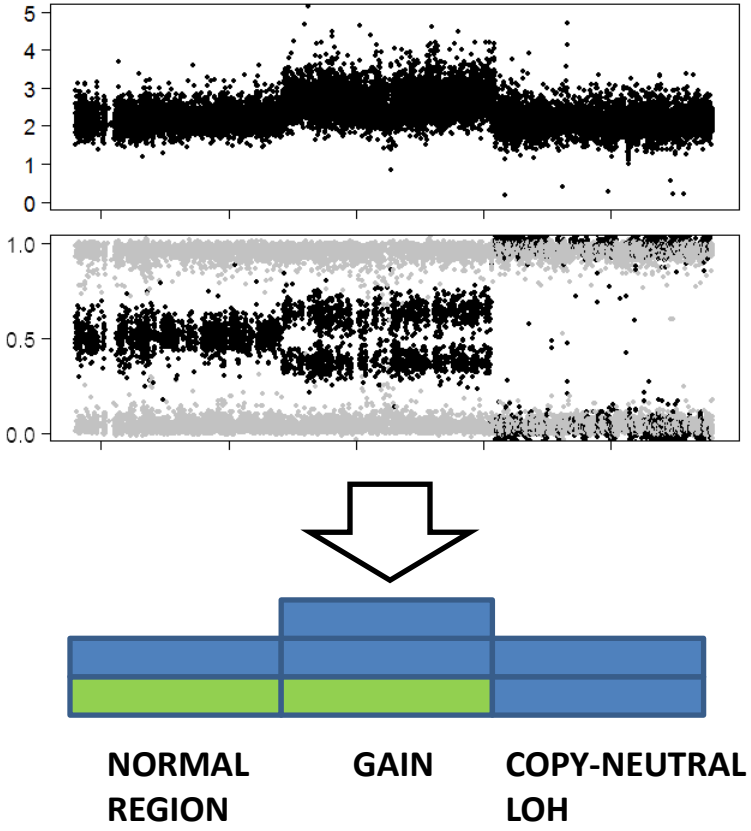
Department of Epidemiology & Biostatistics, UCSF

with

Pierre Neuvial (USA & France)

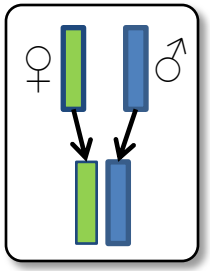
Terry Speed (USA & Australia)

Angel Rubio, Maria Ortiz, Ander Aramburu (Spain)

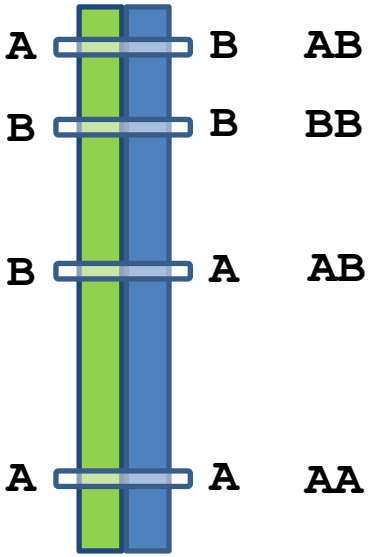


ENAR, Miami, FL, March 22, 2011 (25 min)

Genotypes are observed at single loci

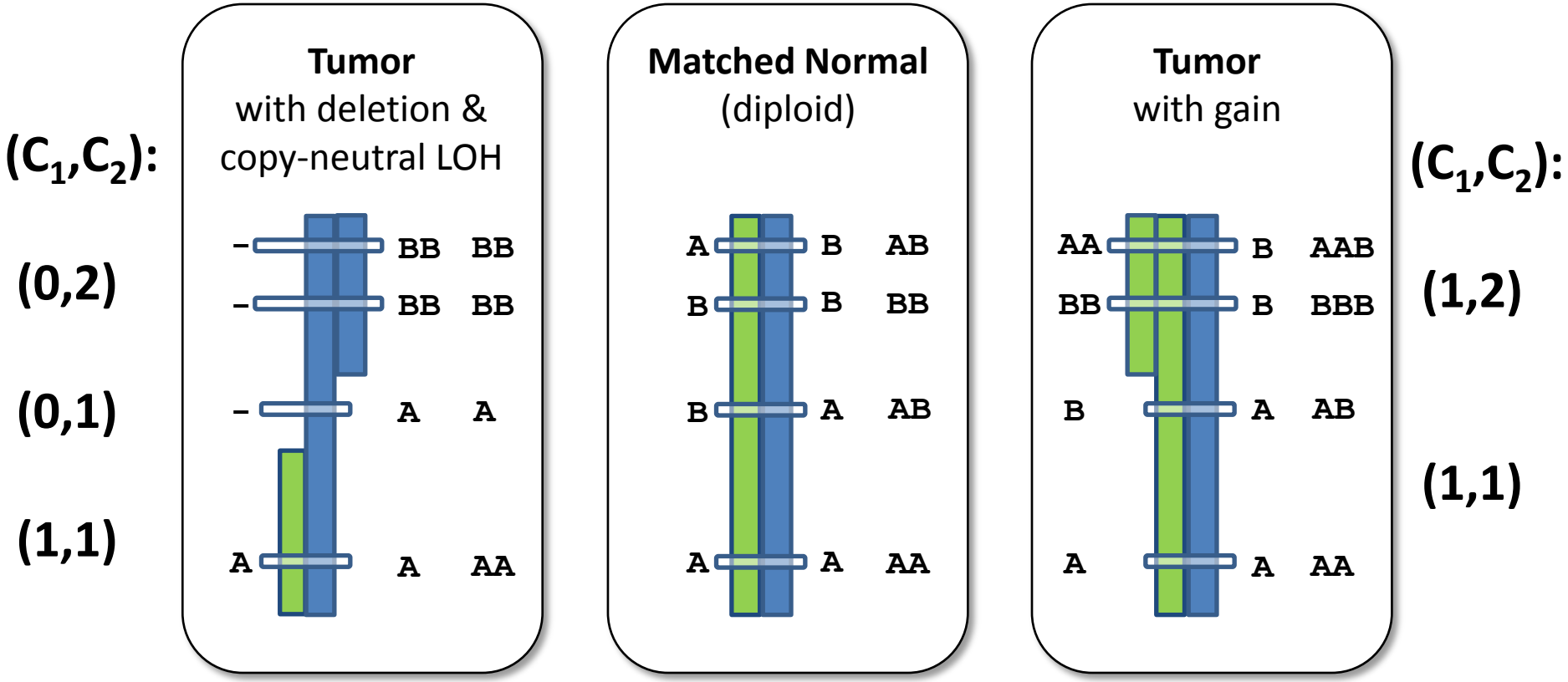


Single nucleotide polymorphism



10-20 million
known SNPs

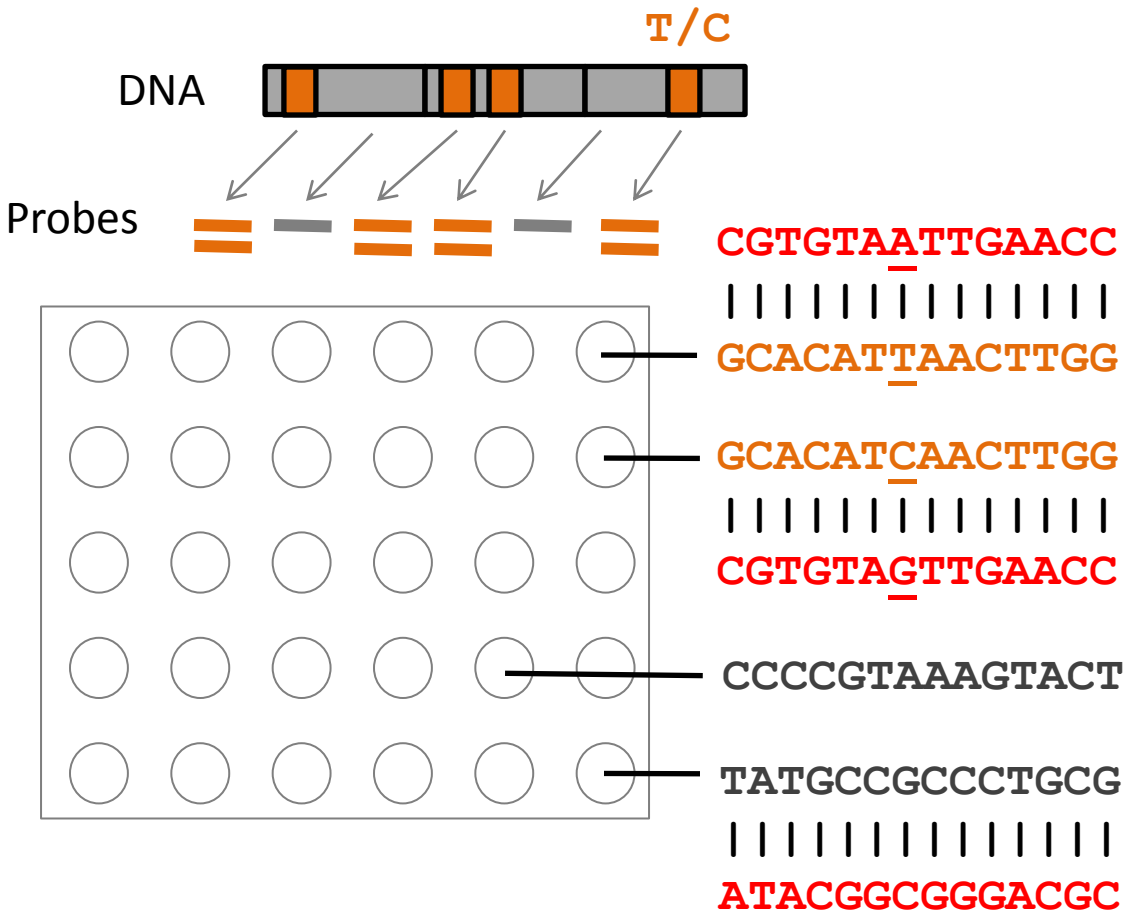
Genotypes and total copy numbers reflect the parent-specific copy numbers



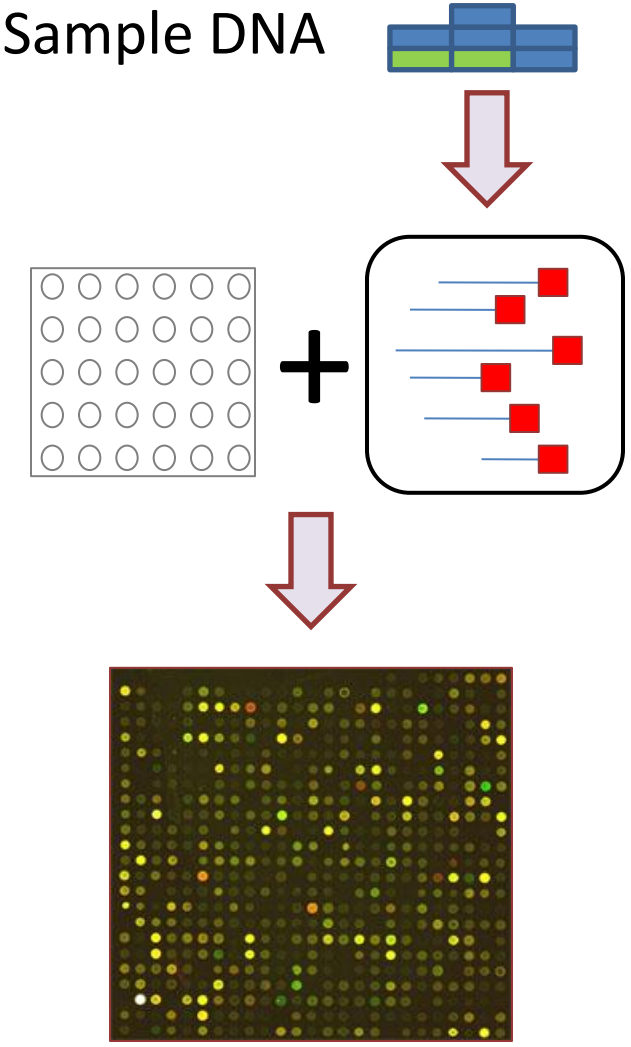
* Occam's razor: Minimal number of events has occurred.

SNP microarrays quantify total and allele-specific copy numbers

Chip Design



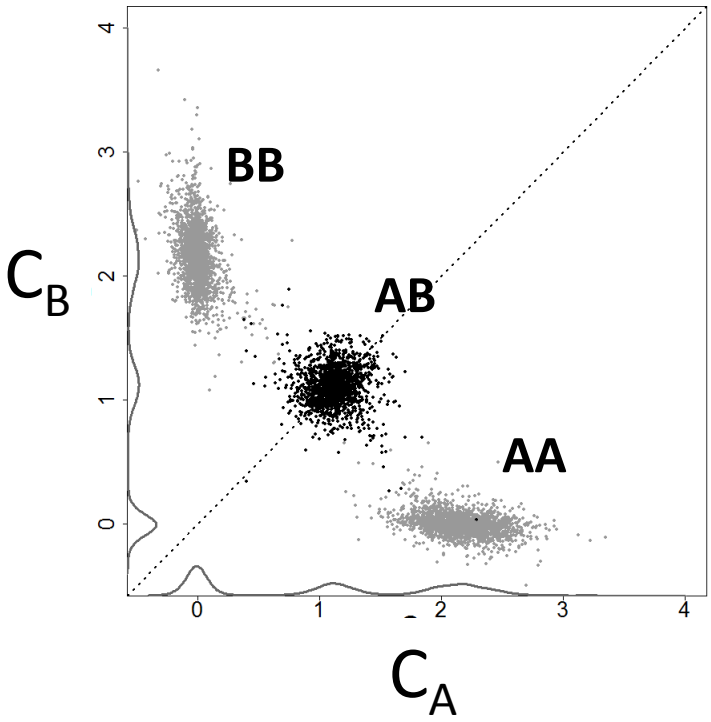
Sample DNA



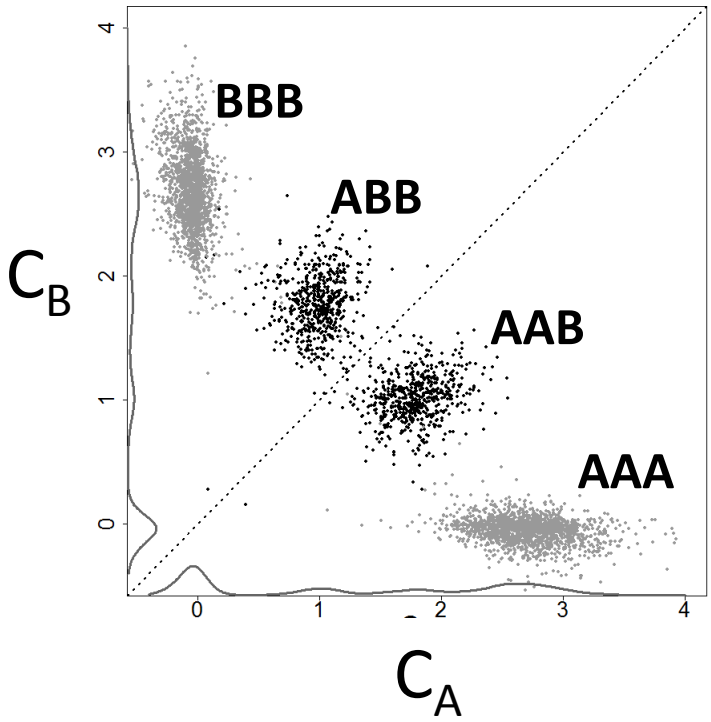
Together the SNPs of a region indicate the parent-specific copy numbers

1 individual, many SNPs

NORMAL (1,1)



GAIN (1,2)

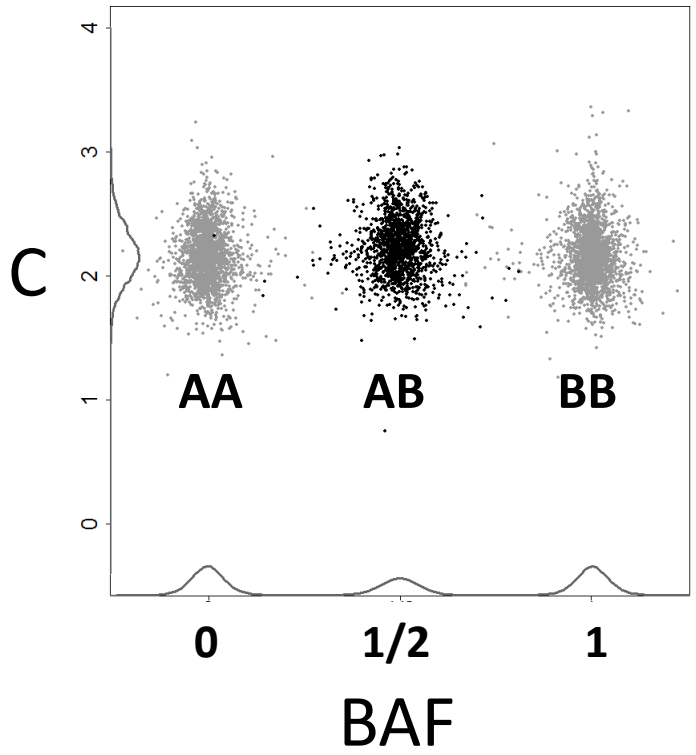


Total CN: $C = C_A + C_B$

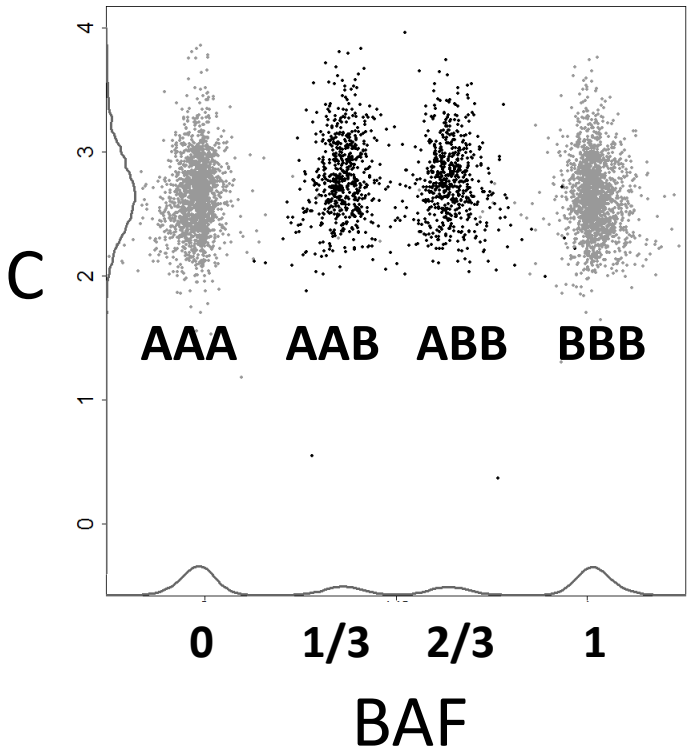
Total CNs and allele B fractions are easier to work with than ASCNs

1 individual, many SNPs, same 2 regions:

NORMAL (1,1)

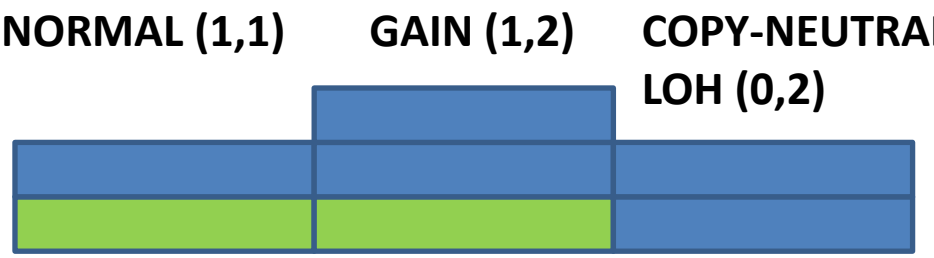


GAIN (1,2)

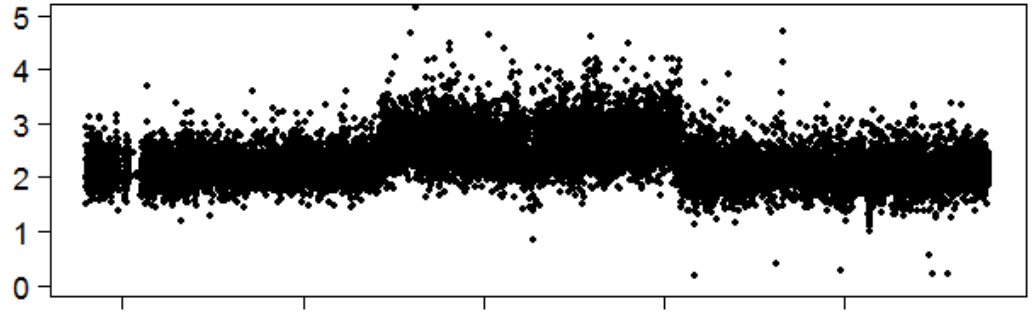


Total CN: $C = C_A + C_B$ BAF: $\beta = C_B / C$

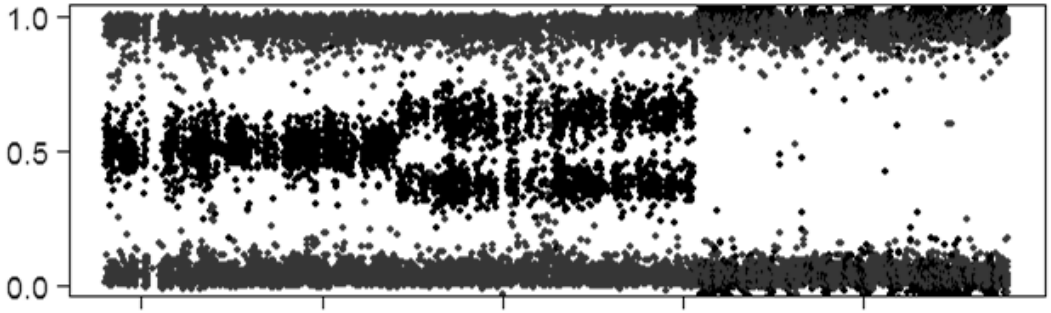
Total CNs and BAFs reflect the underlying parent-specific CNs



Total CN:
$C = C_A + C_B$



Allele B Fraction:
$\beta = C_B / C$



Matched tumor-normals

- With a matched normal it is easier!

…because we can genotype the normal and find the heterozygous SNPs...

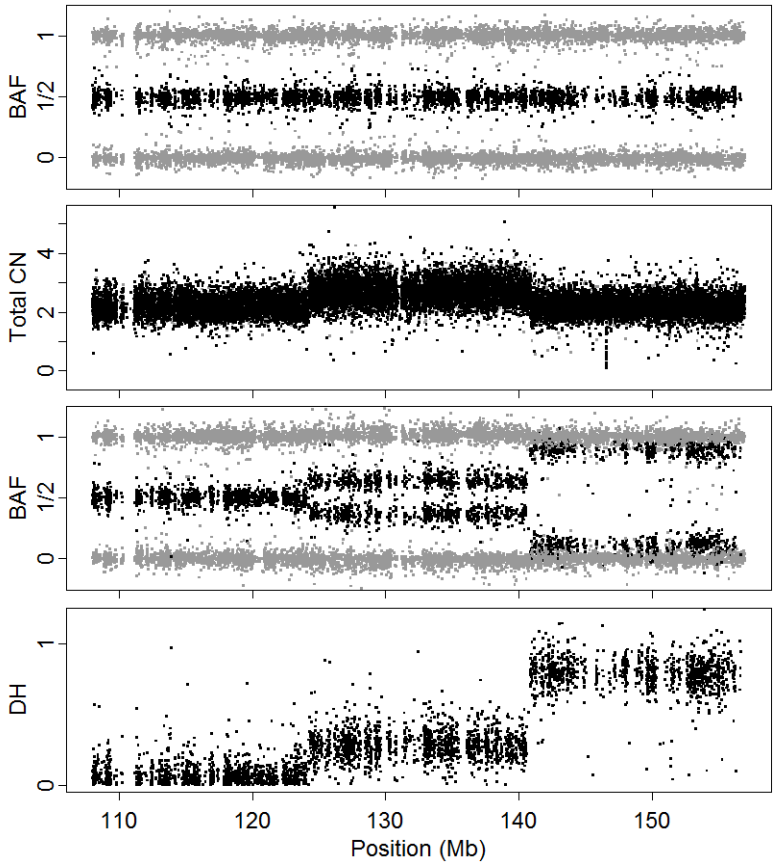
Heterozygous SNPs (not homozygous) are informative for PSCNs

1. **Genotypes (AA,AB,BB)**
from BAFs of a matched normal

2a. **Total CNs**
$C = C_A + C_B$

2b. **Tumor BAFs**
$\beta = C_B / C$

3. **Decrease in Heterozygosity**
$\rho = 2*|\beta - 1/2|$; hets only



← BB
← **AB**
← AA

(all loci)

(SNPs only)

(hets only)

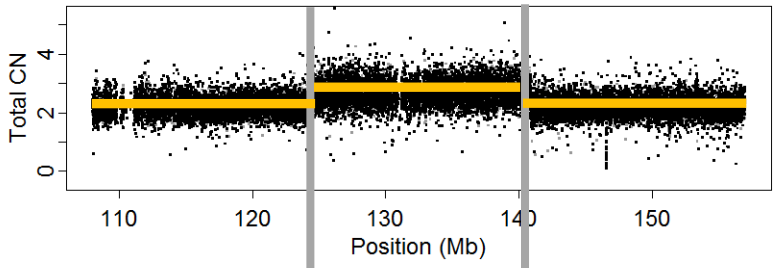
Total CNs & DHs segmentation gives us PSCN regions and estimates

(i) Find change points

(ii) Estimate mean levels

Total CNs

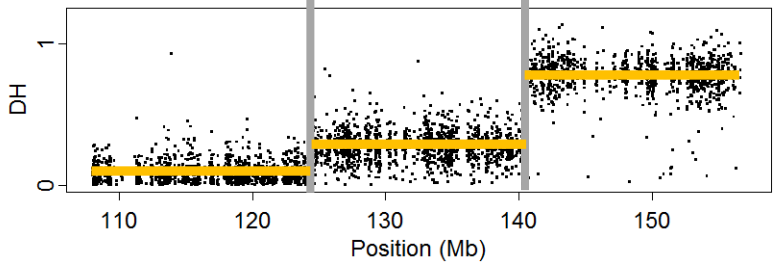
$C = C_A + C_B$



avg(all loci)

Decrease in Heterozygosity

$\rho = 2*|\beta - 1/2|$; hets only



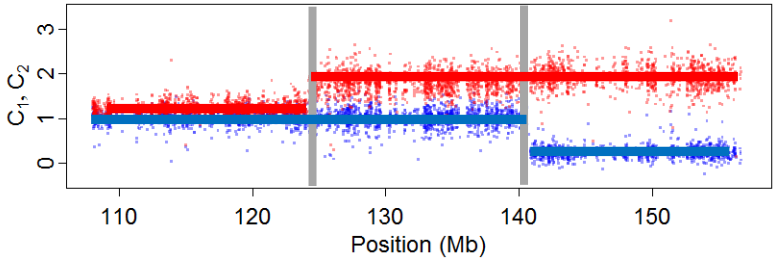
avg(hets only)

NORMAL (1,1) GAIN (1,2) CN-LOH (0,2)

Per-segment PSCNs (C_1,C_2):

$C_1 = 1/2 * (1- \rho) * C$

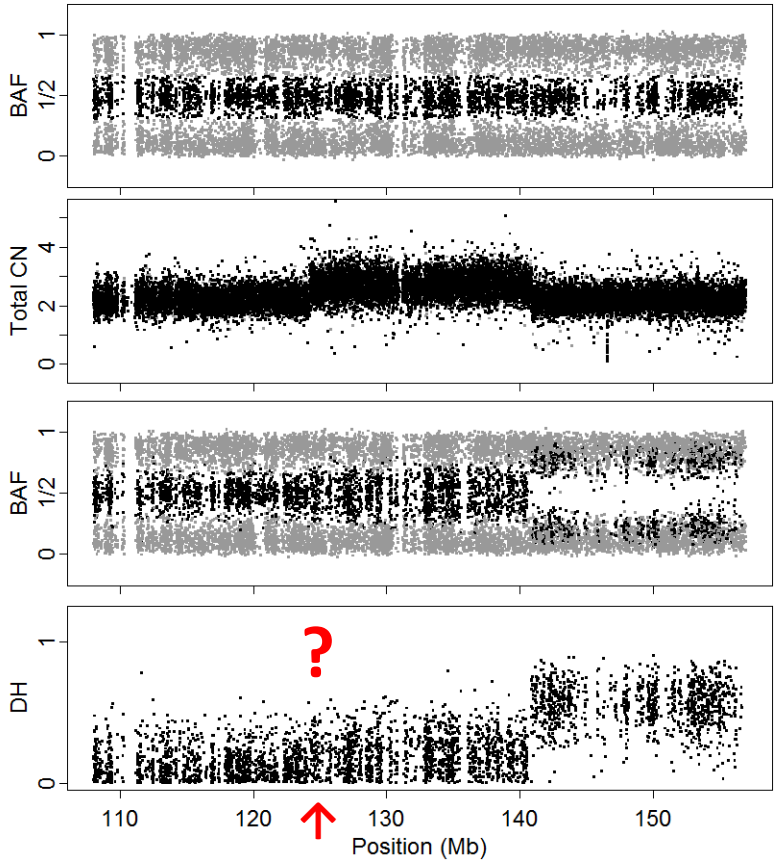
$C_2 = C - C_1$



avg(all loci) * avg(hets only)

It is hard to infer PSCNs reliably when signals are noisy

Actual data:



Segmentation may fail…

Let's improve this…

CalMaTe

Better allele-specific copy numbers
in tumors without matched normals
by borrowing across many samples

Features:

- Multiple (> 30) samples.
- Any SNP microarray platform.
- Bounded memory usage (< 1GB of RAM)

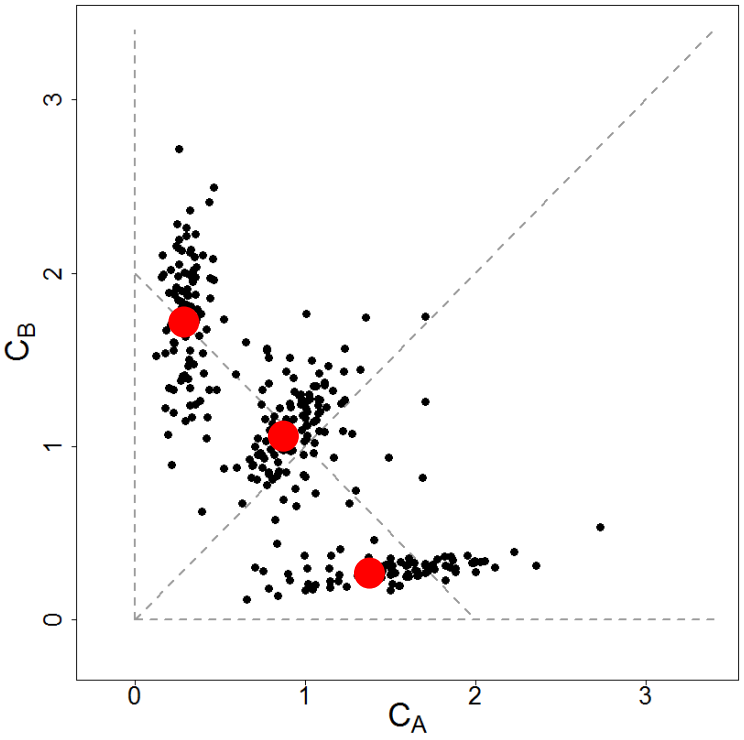
Available: http://www.aroma-project.org/

M Ortiz-Estevez, A. Aramburu, H. Bengtsson, P. Neuvial, & A. Rubio.
A calibration method to improve allele-specific copy number estimates from SNP microarrays (submitted).

The noise is due to SNP-specific effects that we can estimate and remove

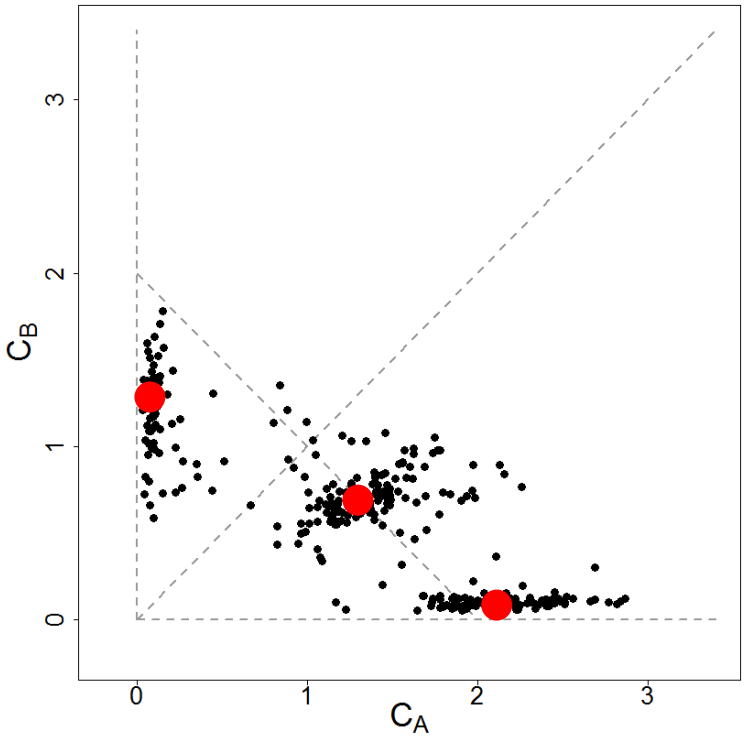
Example: (C_A,C_B) for 310 samples, one SNP:

Systematic effects…



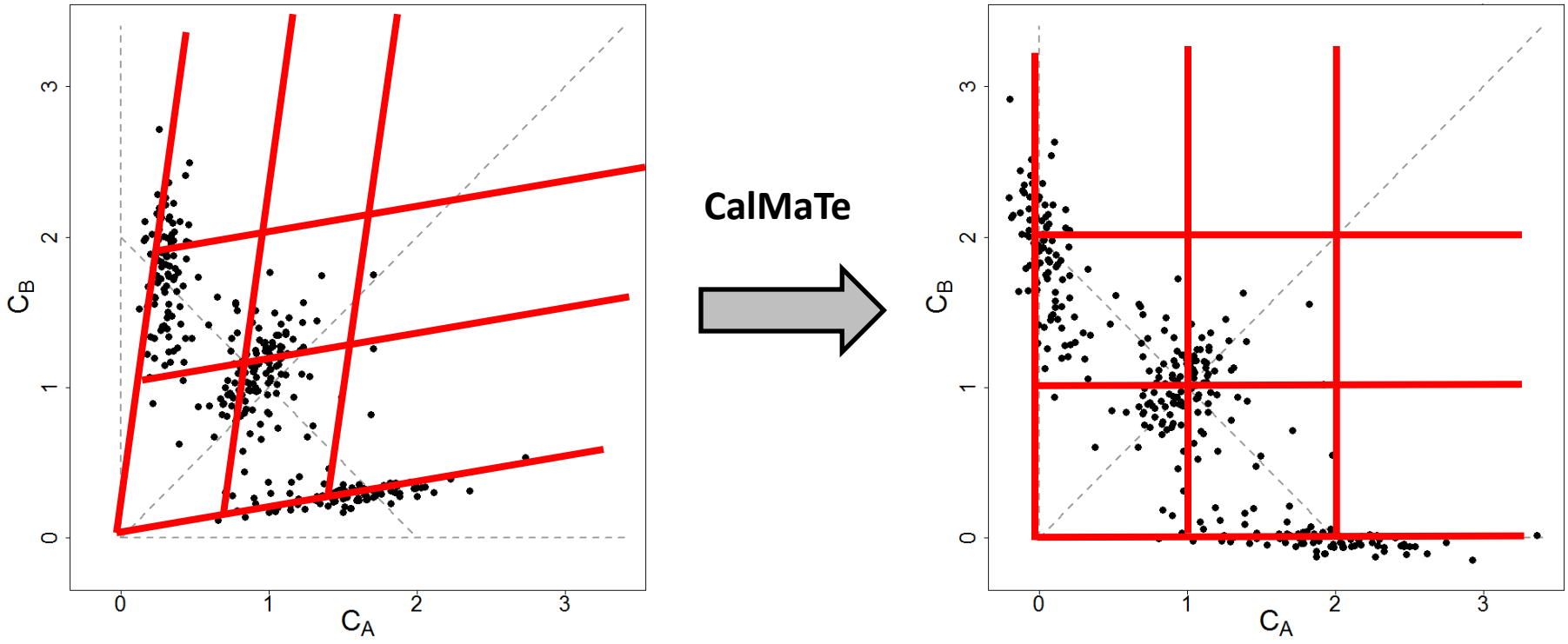
SNP #1053

…are SNP specific!



SNP #1072

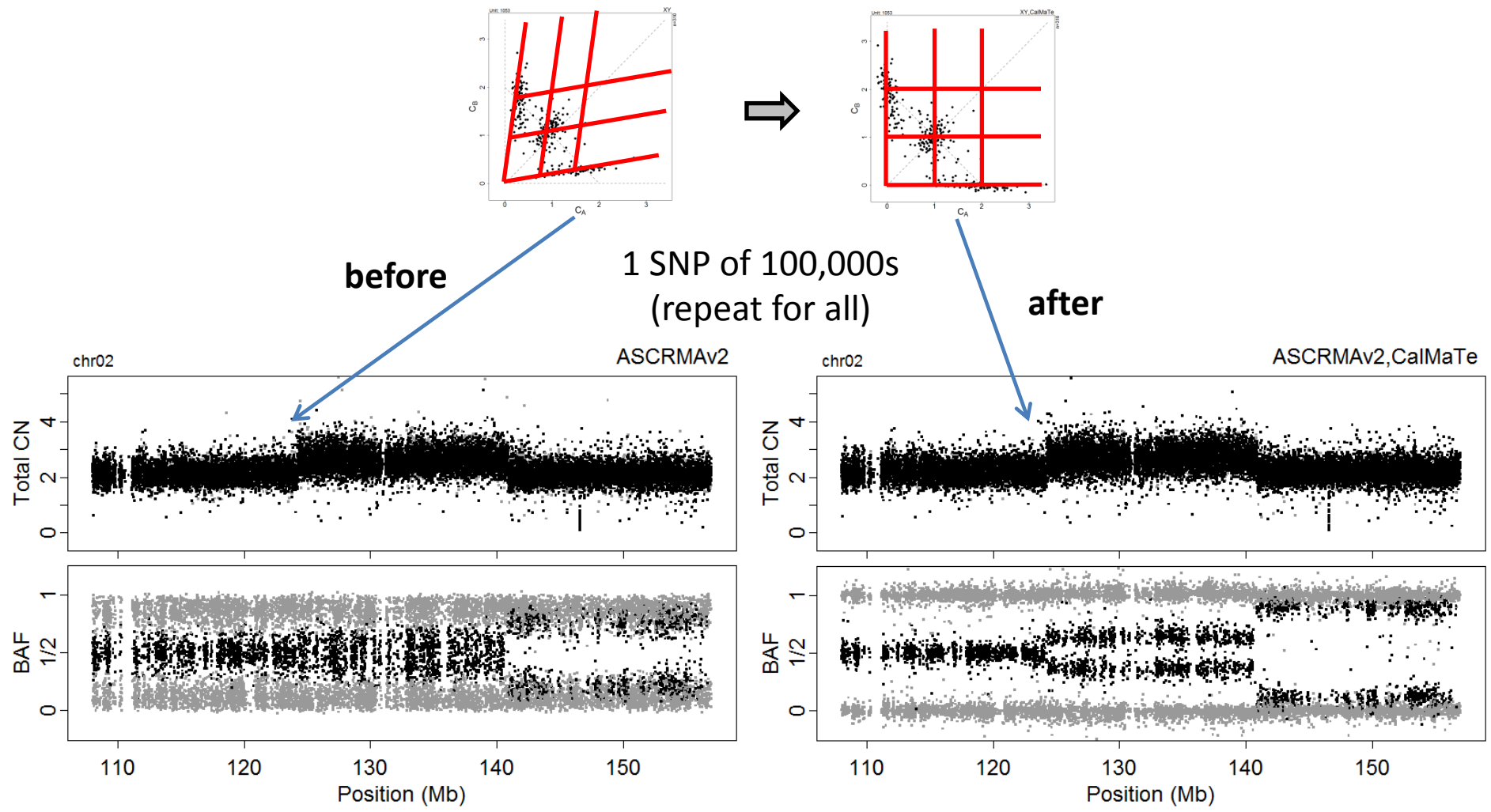
Fit affine transform (one per SNP) across samples and back-transform



Multi-sample method for each SNP separately:
Non-negative Matrix Factorization (NMF).
Robustified against outliers (e.g. tumors).
Special cases: Only one or two genotype groups.

Related methods/ideas:
- Illumina's "Cluster Regression"
- CRLMM CNs (*RLMM, …)
- …

Improved SNR of BAFs (and total CNs) when removing SNP-specific variation



TumorBoost

Better allele-specific copy numbers
in tumors with matched normals

Requirements:

- Matched tumor-normal pairs.
- A single pair is enough.
- Any SNP microarray platform.
- Bounded memory usage (< 1GB of RAM)

Available: http://www.aroma-project.org/

H. Bengtsson, P. Neuvial, T.P. Speed
TumorBoost: Normalization of allele-specific tumor copy numbers from one single tumor-normal pair of genotyping microarrays, BMC Bioinformatics, 2010.

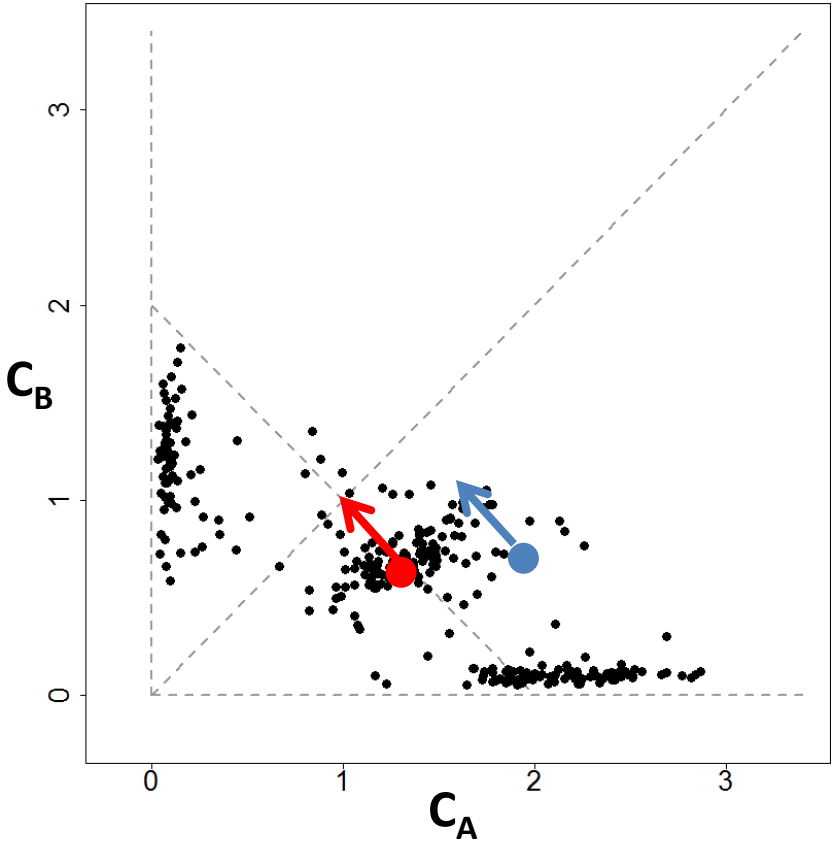
The tumor “should be” close to its normal

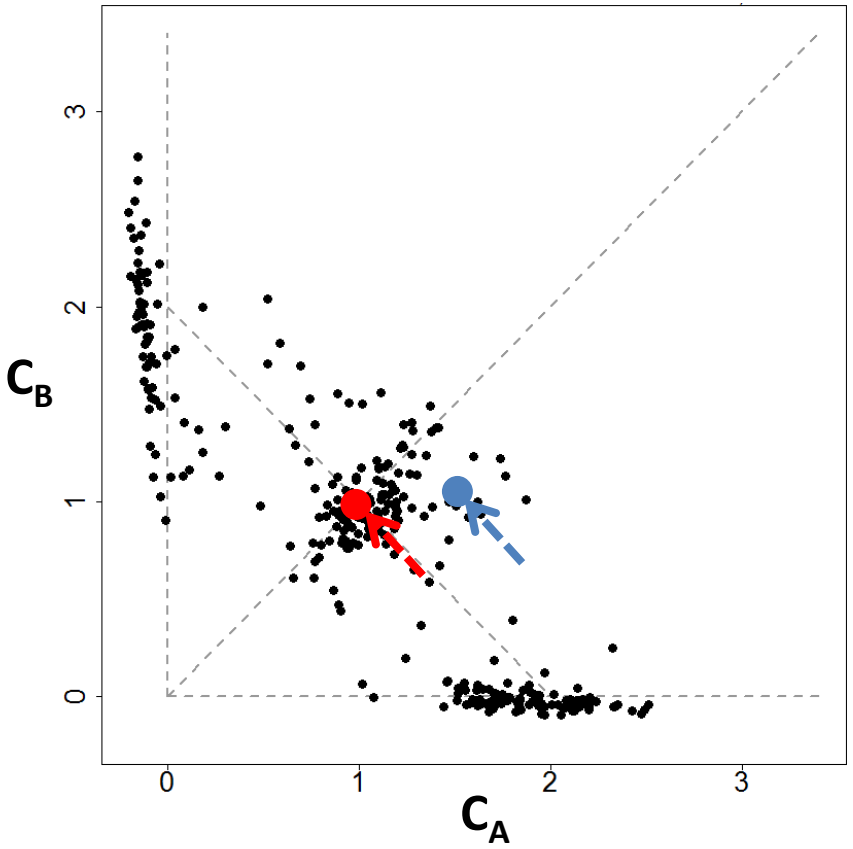
When we have only a single tumor-normal pair:

(i) **Normal** should be at (1,1) …so lets move it there!

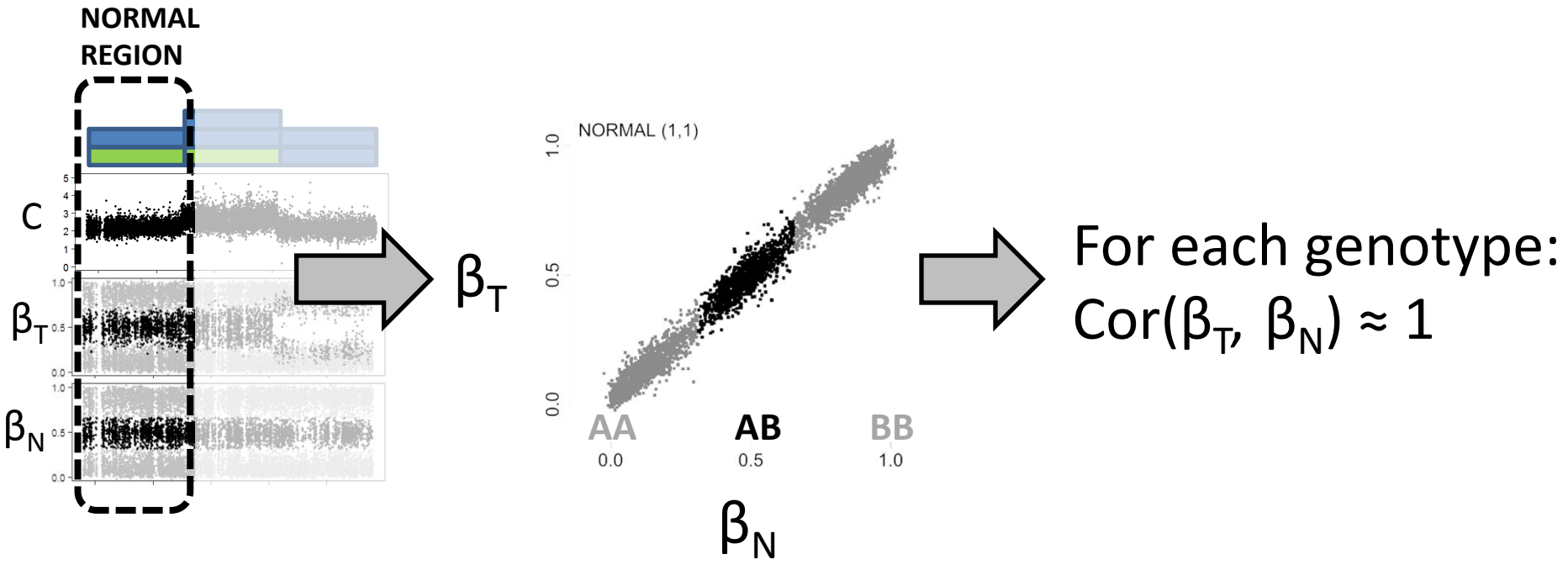
(ii) Adjust the **tumor** in a “similar” direction.

One SNP, many samples

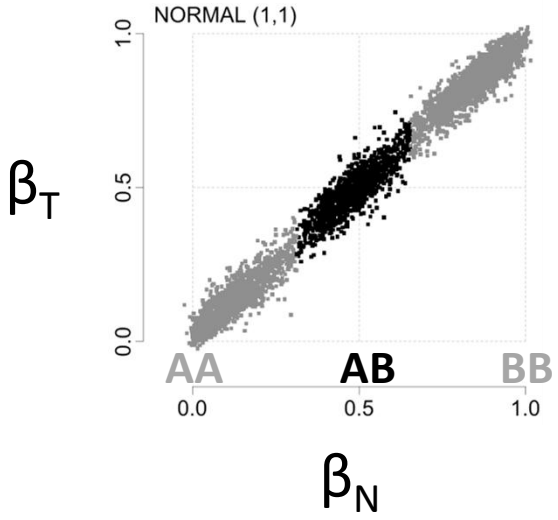


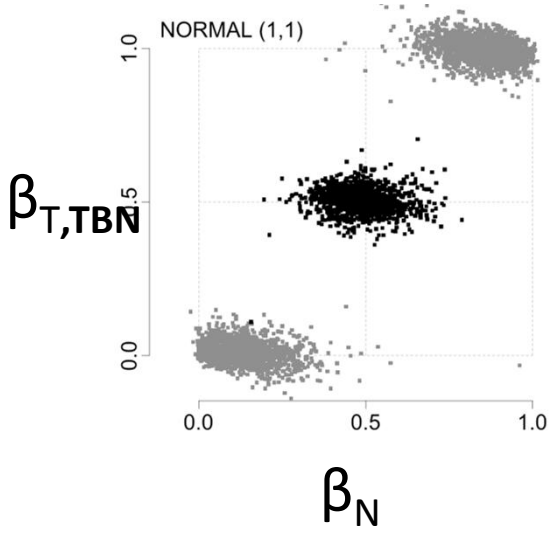


The tumor “should be” close to the normal;
- data strongly agree!



The SNP effect can be estimated & removed for each SNP independently!





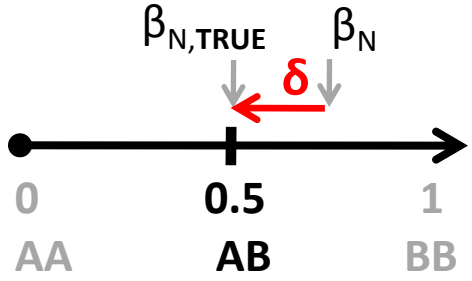
Observed:
Allele B fractions
$\beta_N \in [0,1]$
$\beta_T \in [0,1]$

Genotype calls (AA,AB,BB):
$\beta_{N,TRUE} \in \{0, 0.5, 1\}$

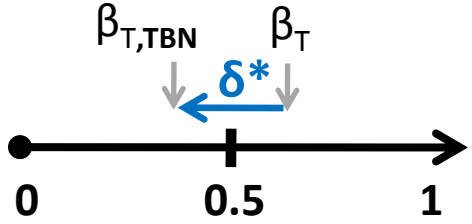
Estimate from normal:
SNP effect
$\delta = \beta_N - \beta_{N,TRUE}$

Remove from tumor:
$\beta_{T,TBN} = \beta_T - \delta^*$

1. Estimate SNP effect in the normal and its genotypes

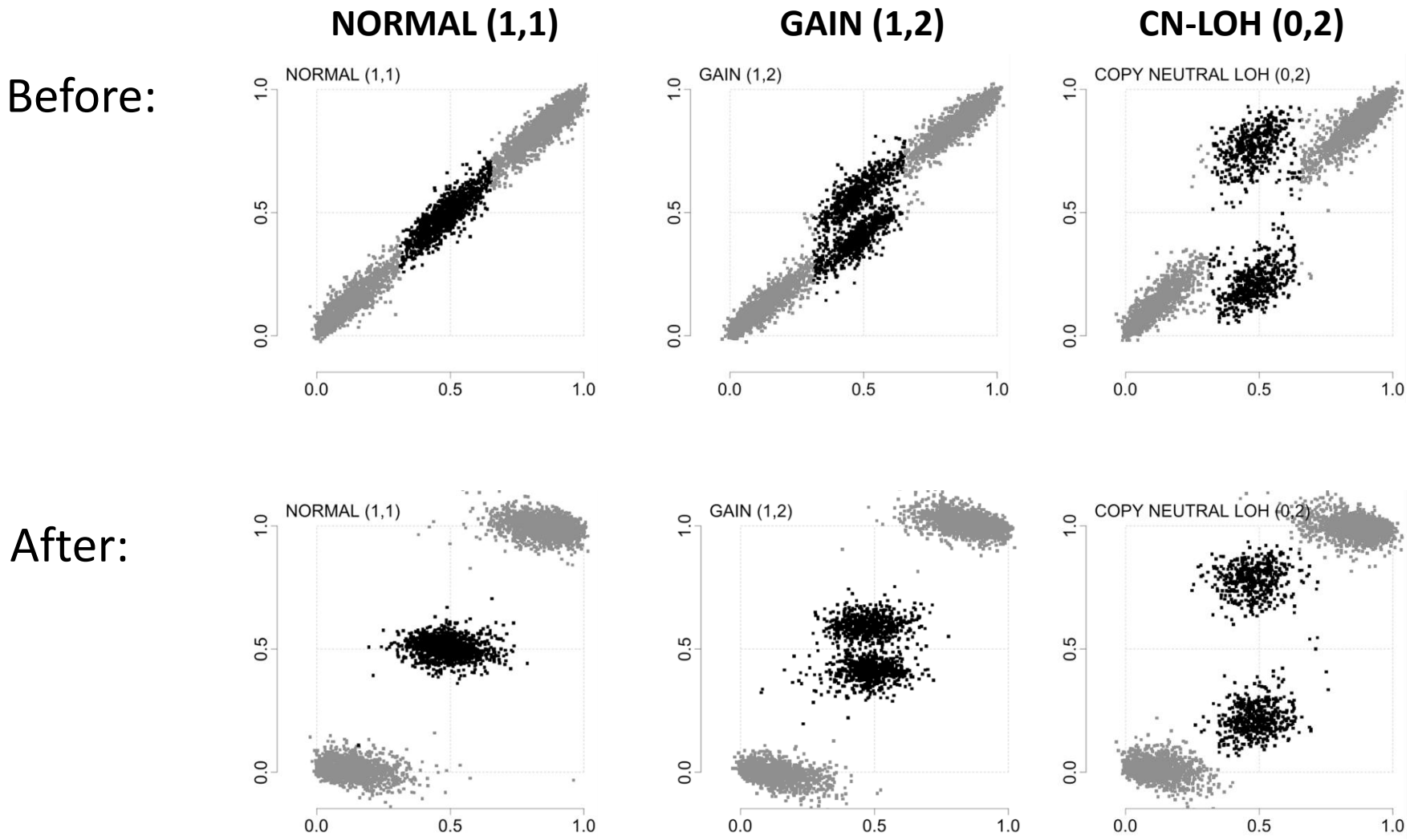


2. Remove SNP effect from the tumor

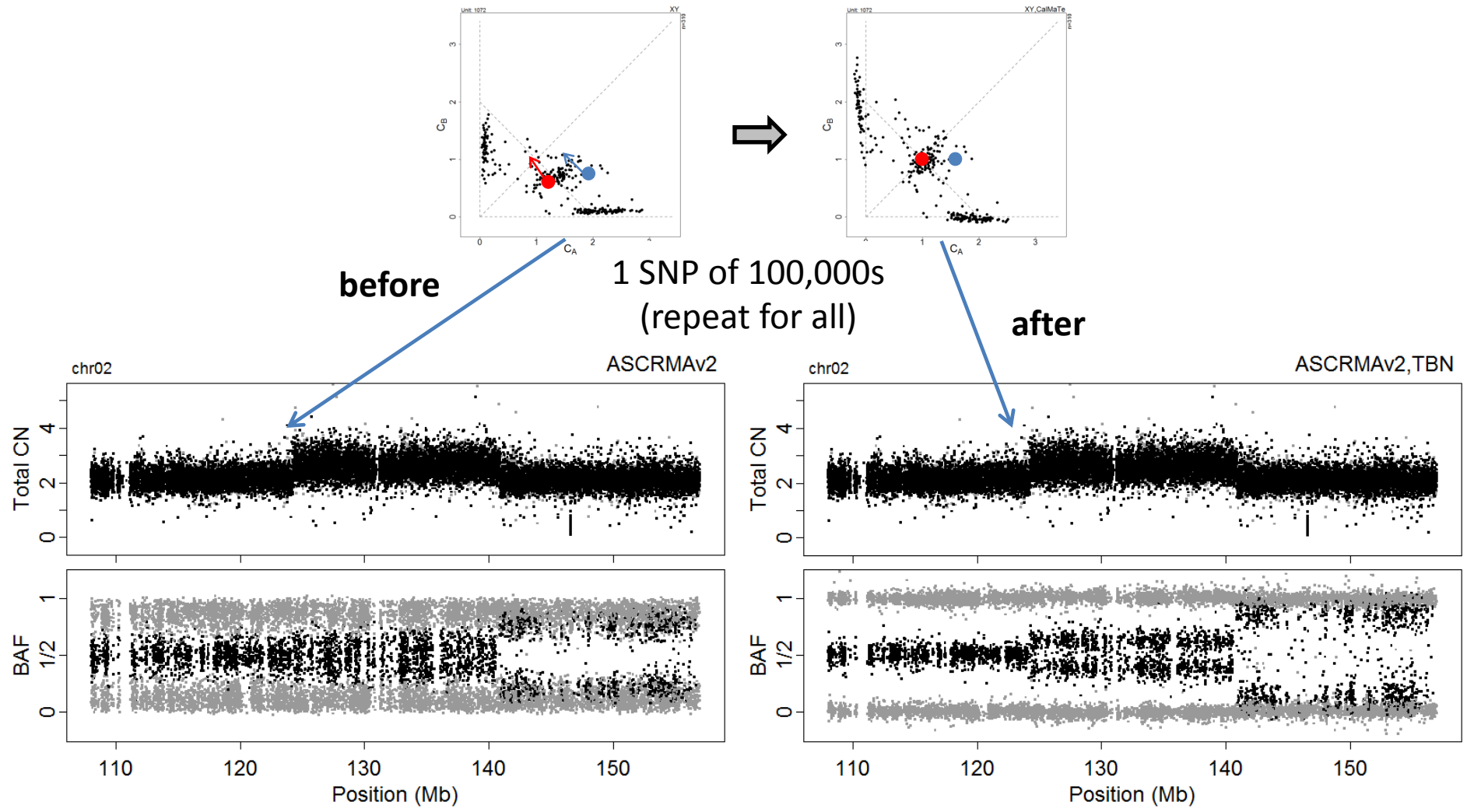


3. Repeat for all SNPs.

TumorBoost removes the SNP effects from the tumor (only)



Even with a single tumor-normal pair, we can greatly improve the SNR

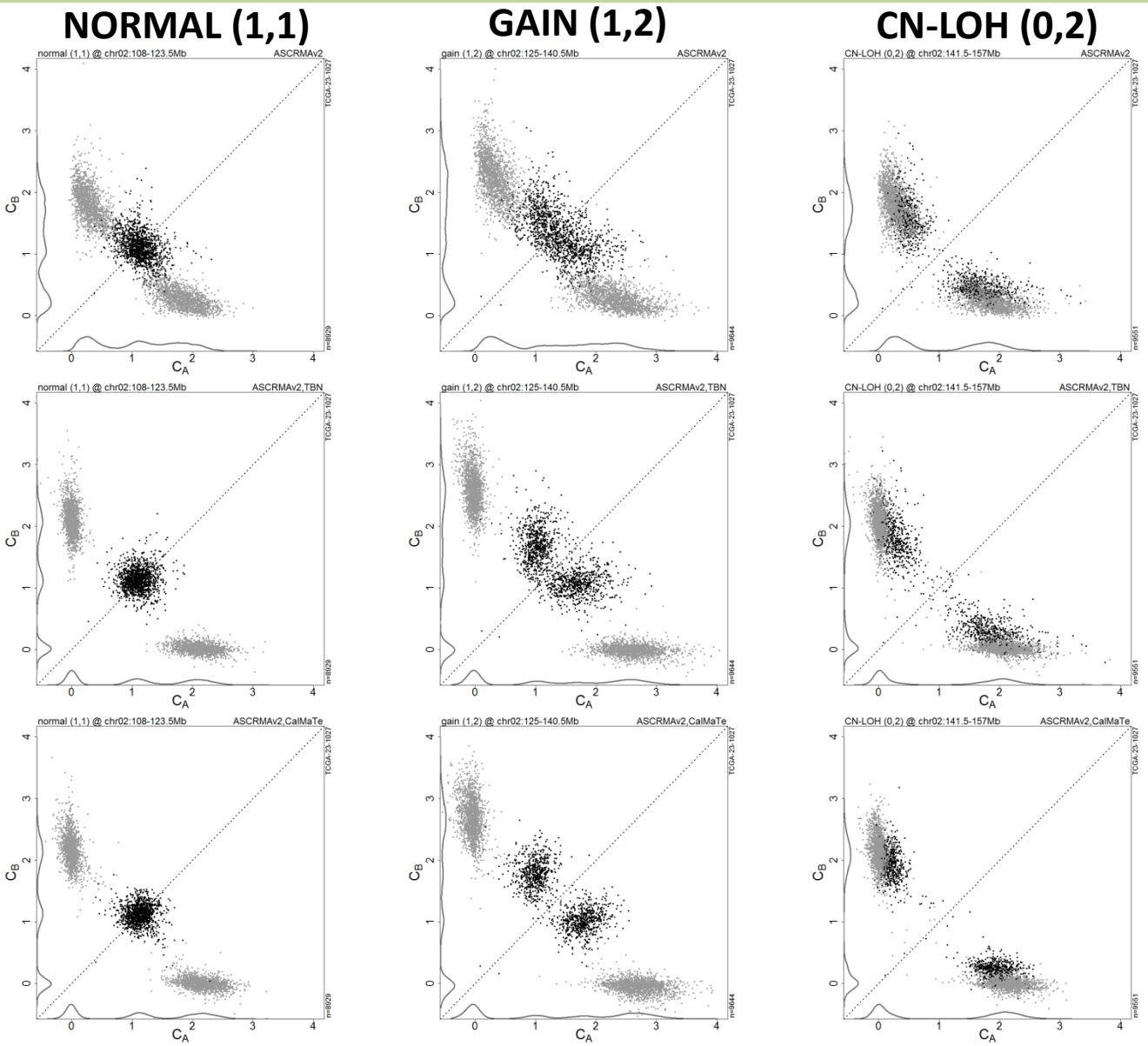


TumorBoost / CalMaTe => more distinct (C_A,C_B) - key for PSCN segmentation

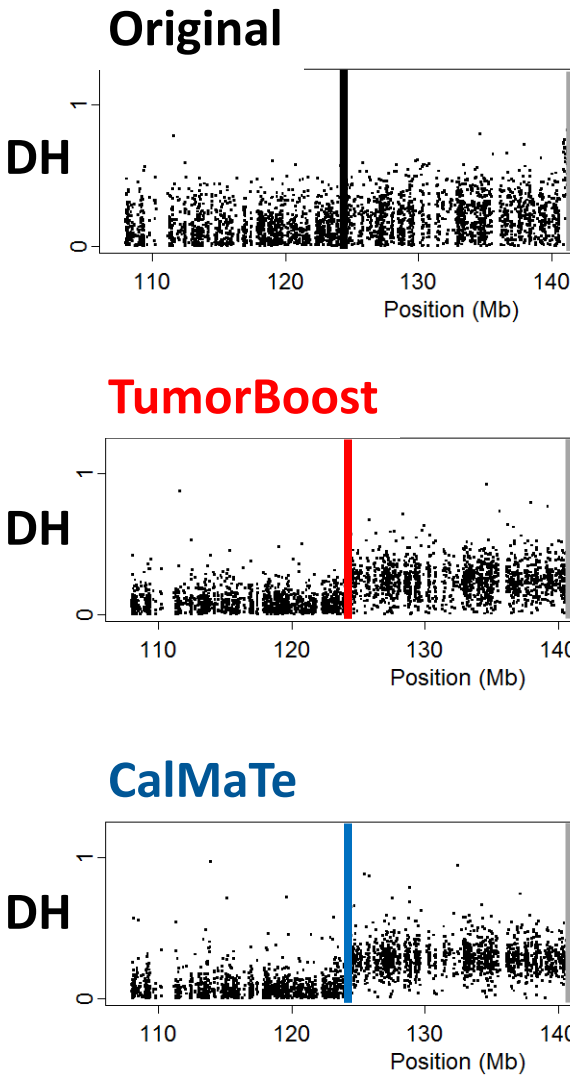
Original:

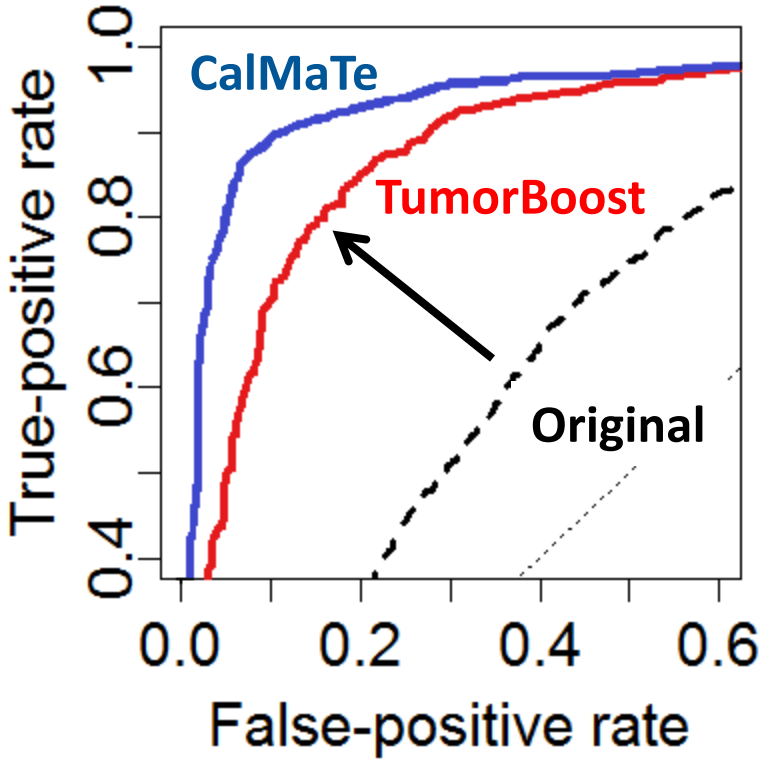
TumorBoost:
- single-pair
- tumor-normals
- normal is not corrected

CalMaTe:
- multi-sample



TumorBoost and CalMaTe significantly improve power to detect change points





One sample, one change point

Methods are available now (www.aroma-project.org)

Preprocessing:

- Affymetrix: ASCRMAv2 (single-array)
- Illumina: <elsewhere>

Normalization of ASCNs:

- Single tumor-normal pair: TumorBoost
- Multiple samples: CalMaTe

PSCN segmentation:

- Single tumor-normal pair: Paired PSCBS
- No matched normals: <we're working on it>

Everything is bounded in memory (< 1GB of RAM)

The End

Noise in PSCN signals is due to SNP-specific effects, which can be removed if we have:

- a large set of samples, or
- a matched normal.

=> Better PSCN segmentation!

Acknowledgments

Pratyaksha Wirapati (Swiss Institute of Bioinformatics)
Ken Simpson, Mark Robinson (WEHI, Australia)
James Bullard, Kasper Hansen (UC Berkeley; John Hopkins)
Adam Olshen (UCSF)
NCI, NHI, TCGA