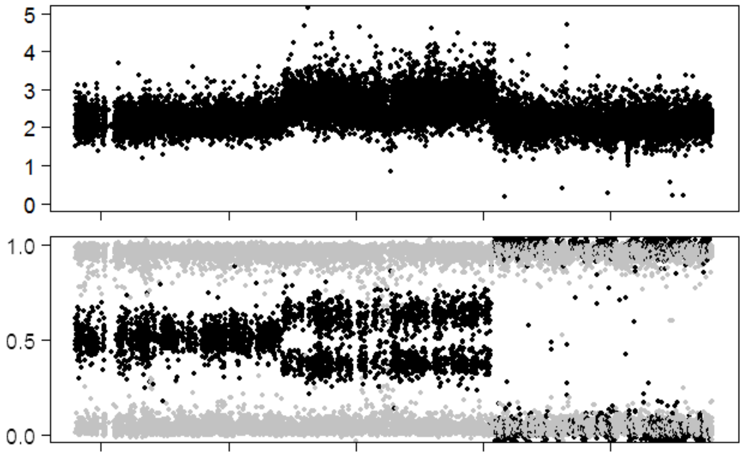
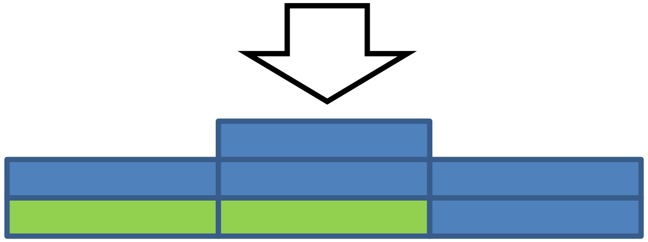
Single Tumor-Normal Pair
Parent-Specific Copy Number Analysis

Henrik Bengtsson

Department of Epidemiology & Biostatistics,
UCSF

with: **Pierre Neuvial**, Berkeley/CNRS
Adam Olshen, UCSF
Richard Olshen, Stanford
Venkatraman Seshan, MSKCC
Terry Speed, Berkeley/WEHI
Paul Spellman, LBNL/OHSU





NORMAL REGION **GAIN** **COPY-NEUTRAL LOH**

“This presentation has been modified from its original version...“

The content of the slides was formatted to fit the upper 3/4 of the screen at IPAM, so that also the audience in the back would be able to see all of it.

Paired PSCBS

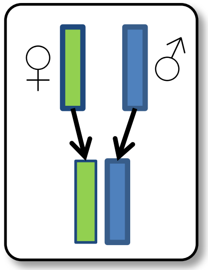
Parent-specific copy numbers from
a single tumor-normal pair of SNP arrays

1. Tumor-normal pair
2. Genotype normal
3. Normalize tumor using normal
4. Segment tumor CNs in two steps
5. Estimate PSCNs within segments
6. Call segments

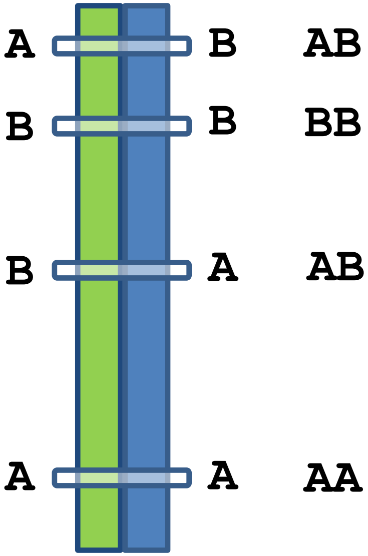
-- H Bengtsson, P Neuvial, TP Speed, TumorBoost: Normalization of allele-specific tumor copy numbers from one single tumor-normal pair of genotyping microarrays, BMC Bioinformatics 2010.

-- AB Olshen, H Bengtsson, P Neuvial, PT Spellman, RA Olshen, VE Seshan, Parent-specific copy number in paired tumor-normal studies using circular binary segmentation, Bioinformatics 2011.

Genotypes are observed at single loci

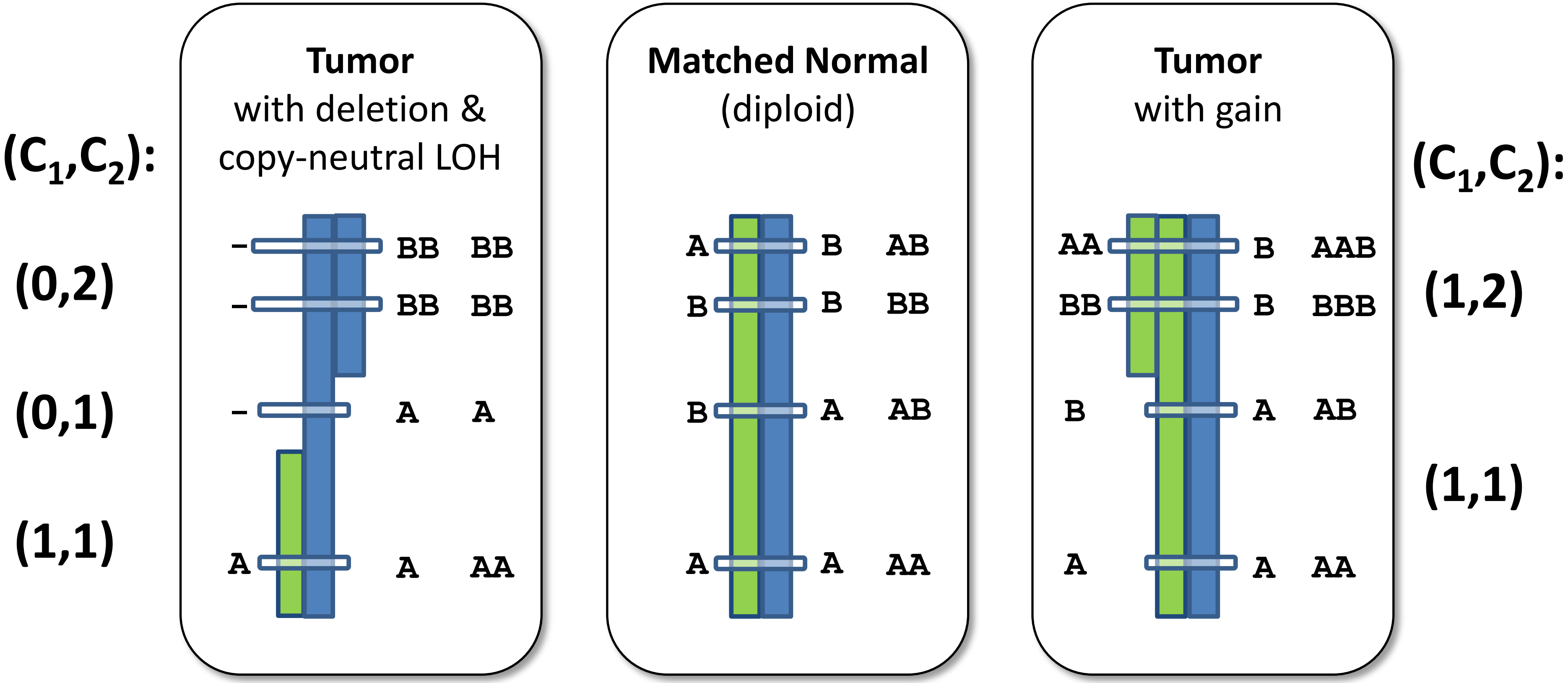


Single nucleotide polymorphism



10-20 million known SNPs

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



(C_1,C_2):

(0,2)

(0,1)

(1,1)

Tumor
with deletion &
copy-neutral LOH

– BB BB

– BB BB

– A A

A A AA

Matched Normal
(diploid)

A B AB

B B BB

B A AB

A A AA

Tumor
with gain

AA B AAB

BB B BBB

B A AB

A A AA

(C_1,C_2):

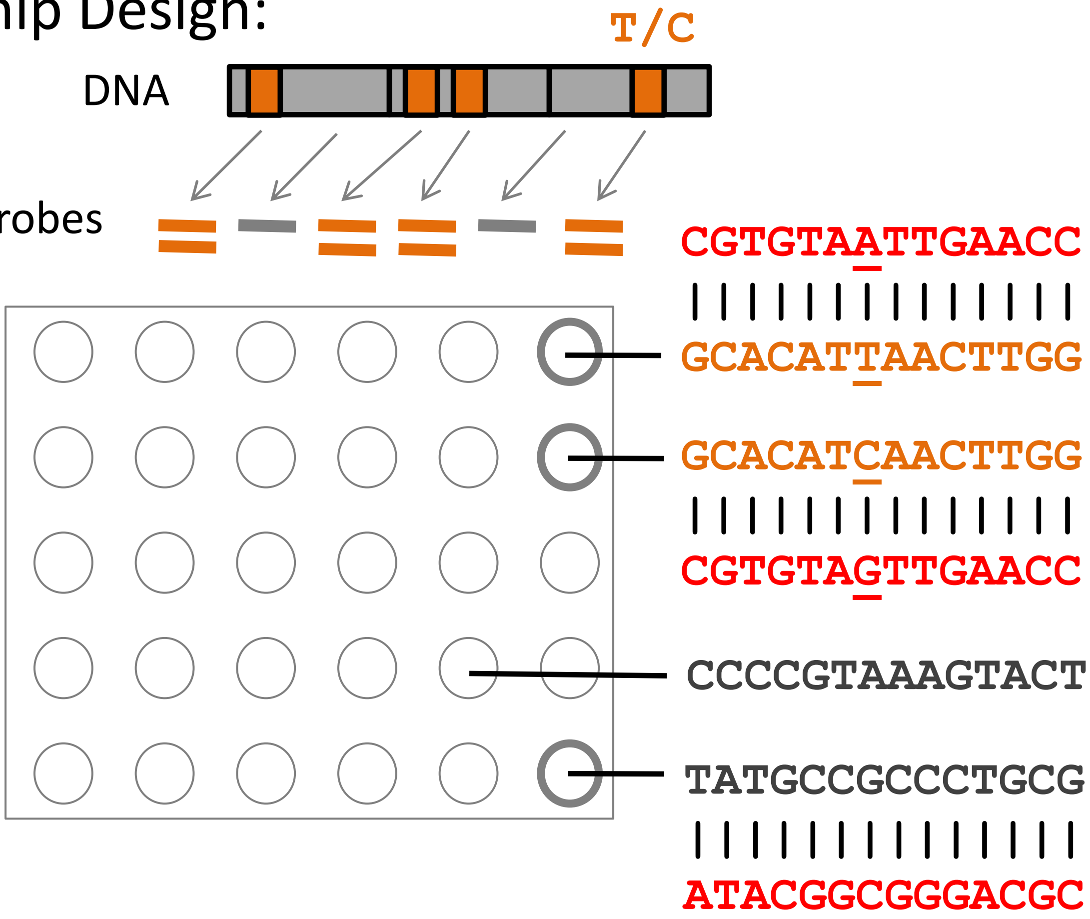
(1,2)

(1,1)

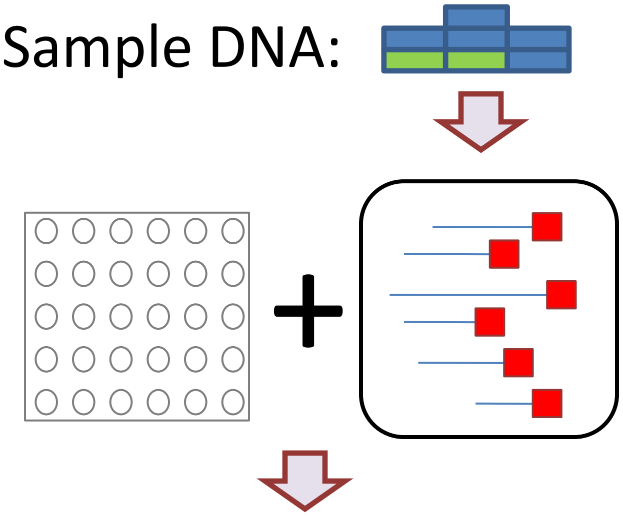
* Occam's razor: Minimal number of events has occurred.

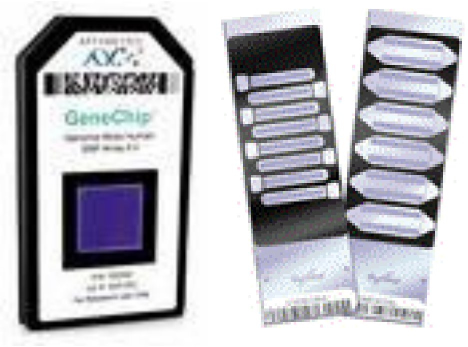
SNP microarrays quantify total and allele-specific copy numbers

Chip Design:



Sample DNA:



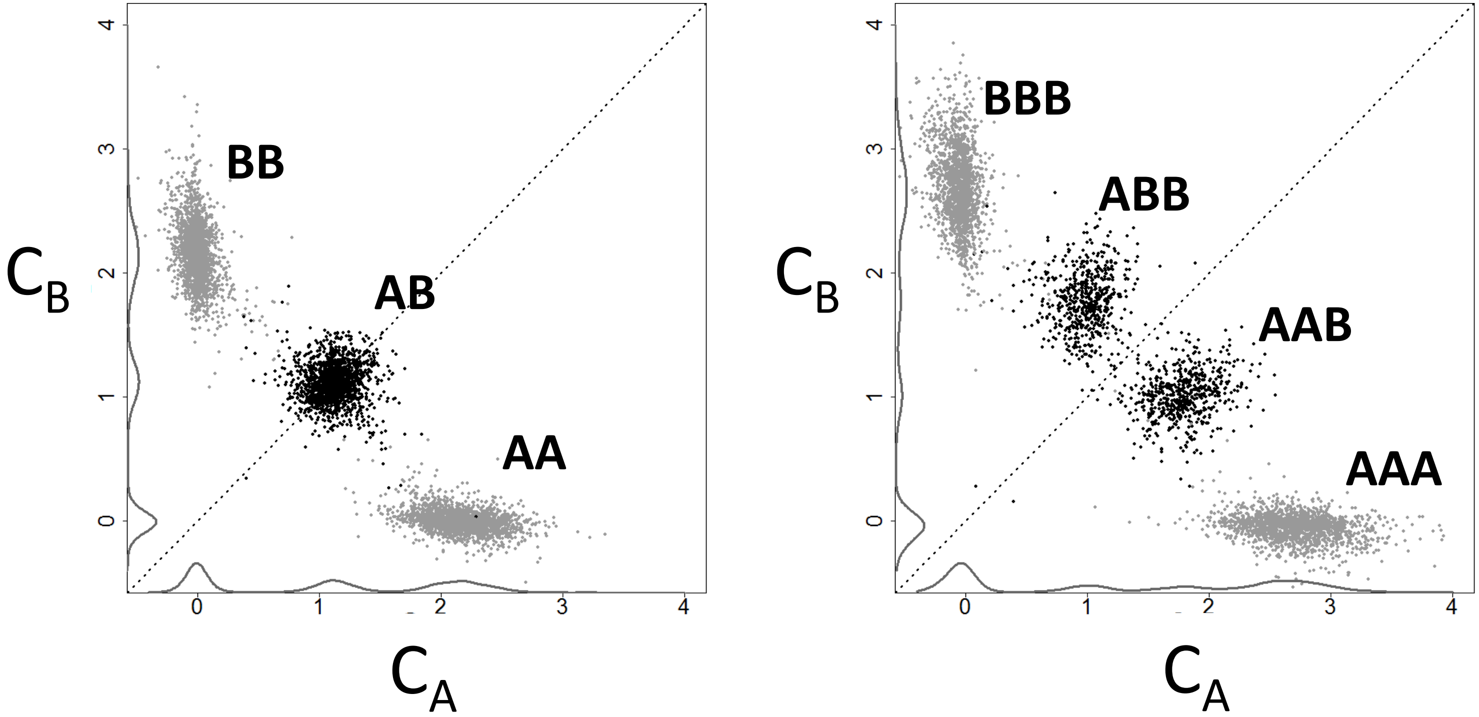


Together the SNPs of a region indicate the parent-specific copy numbers

Total CN: $C = C_A + C_B$

NORMAL (1,1)

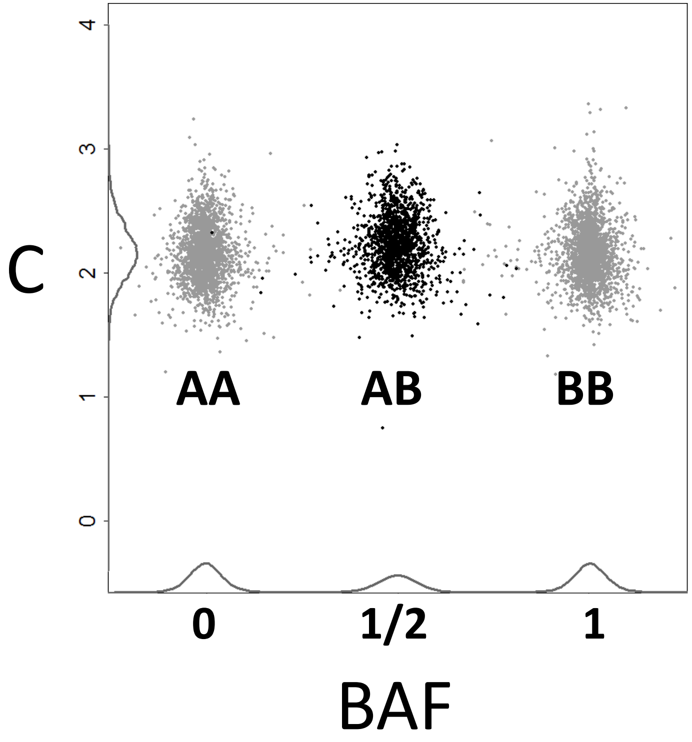
GAIN (1,2)



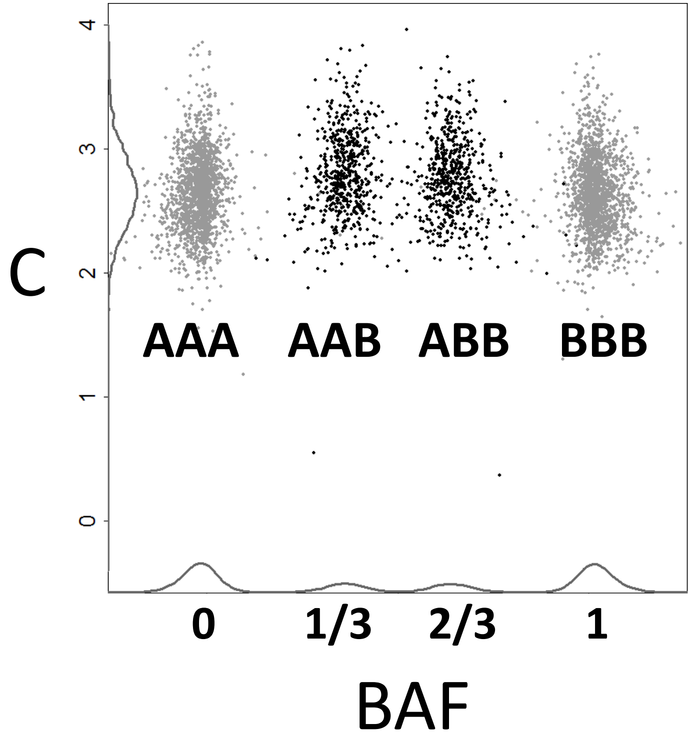
(1 individual, many SNPs, 2 different regions)

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

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

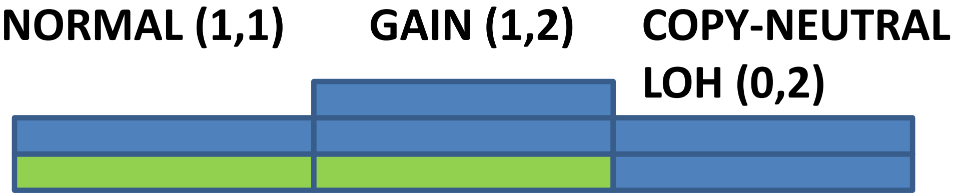


NORMAL (1,1)



GAIN (1,2)

Total CNs and BAFs reflect the underlying parent-specific CNs



NORMAL (1,1)

GAIN (1,2)

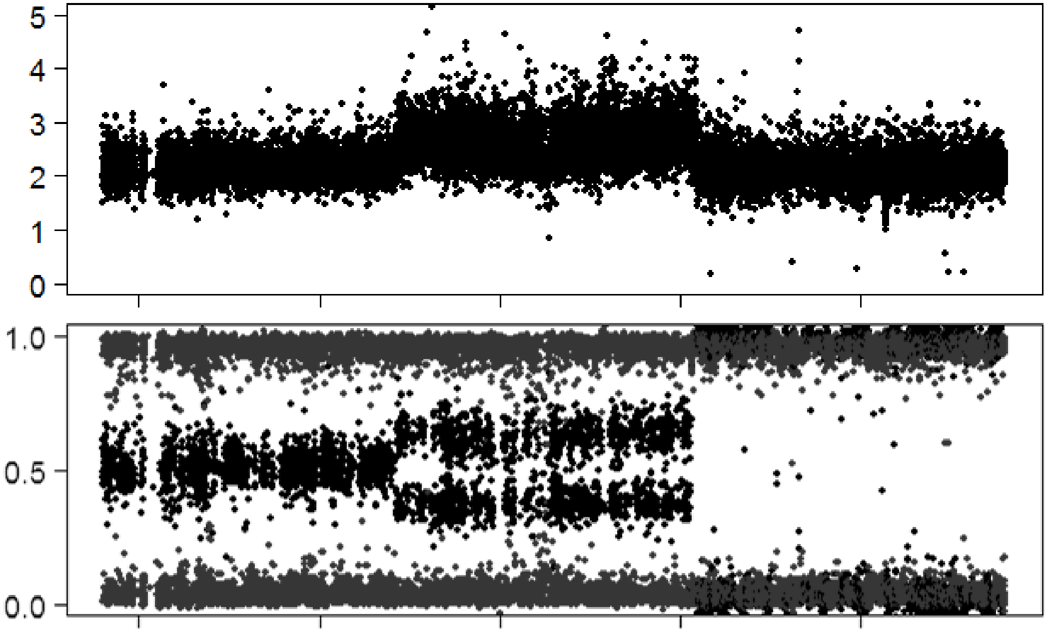
COPY-NEUTRAL LOH (0,2)

Total CN:

$C = C_A + C_B$

Allele B Fraction:

$\beta = C_B / C$



← CN=3

← CN=2

← 100% B:s

← 50% B:s

← 0% B:s

Matched tumor-normals

- With a matched normal it is easier!

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

- Also, much greater SNRs

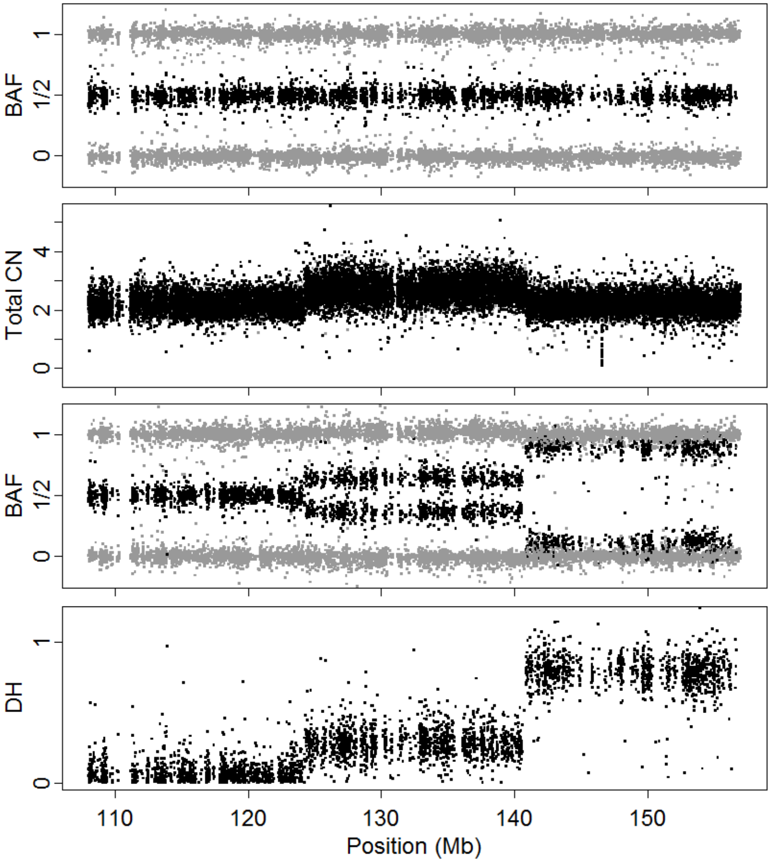
Heterozygous SNPs (not homozygous) are informative for PSCNs

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

$C = C_A + C_B$

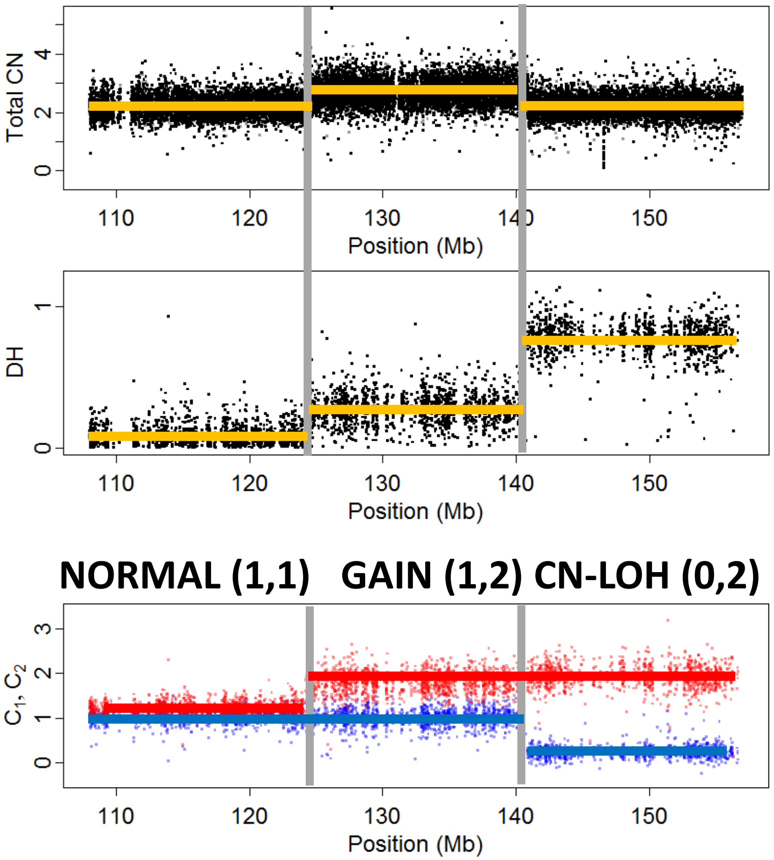
Decrease in Heterozygosity

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

Per-segment PSCNs (C_1,C_2):

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

$C_2 = C - C_1$



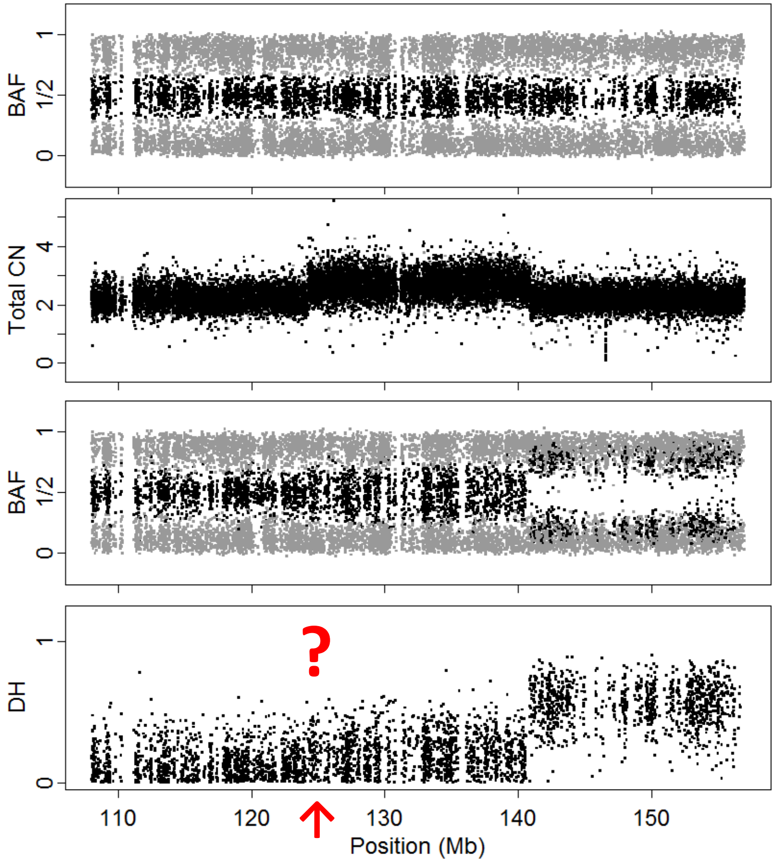
avg(all loci)

avg(hets only)

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)

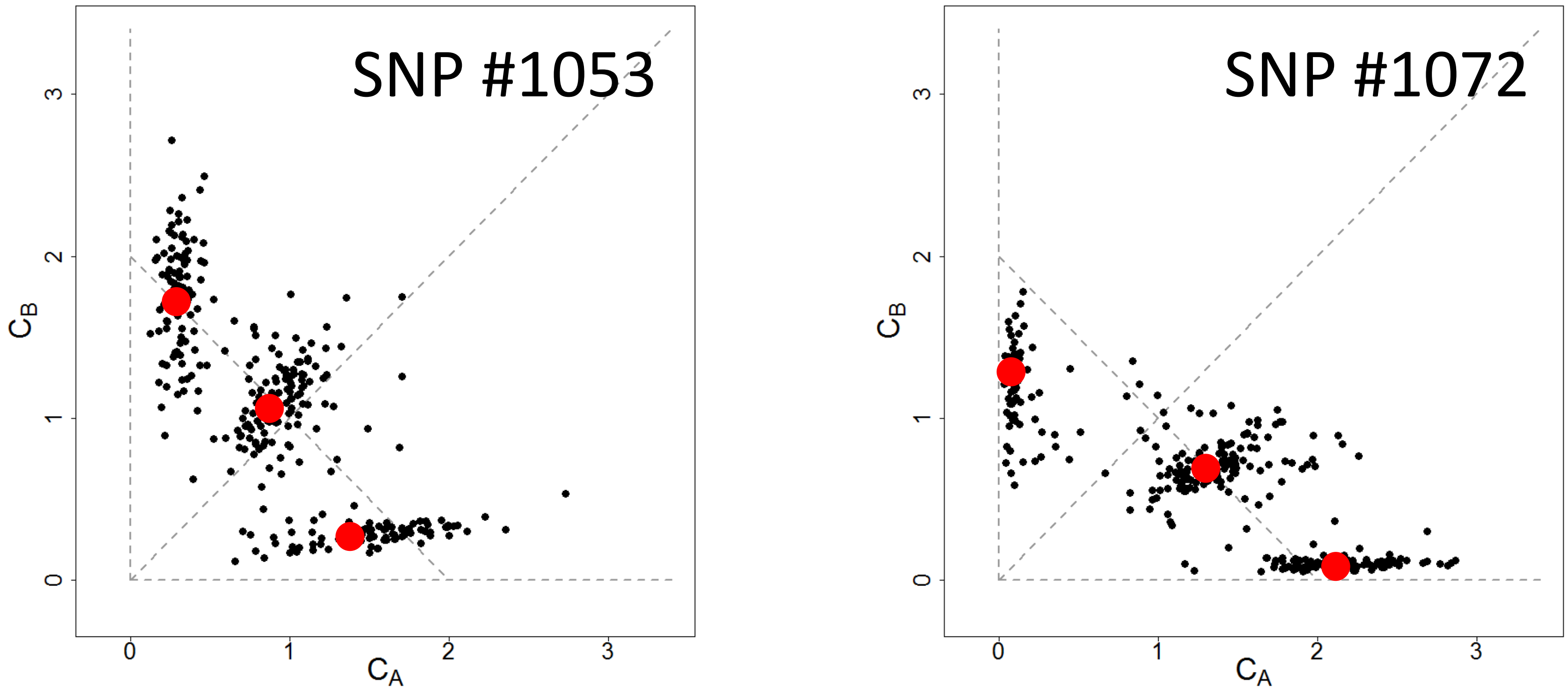
More: http://www.aroma-project.org/

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 per SNP:

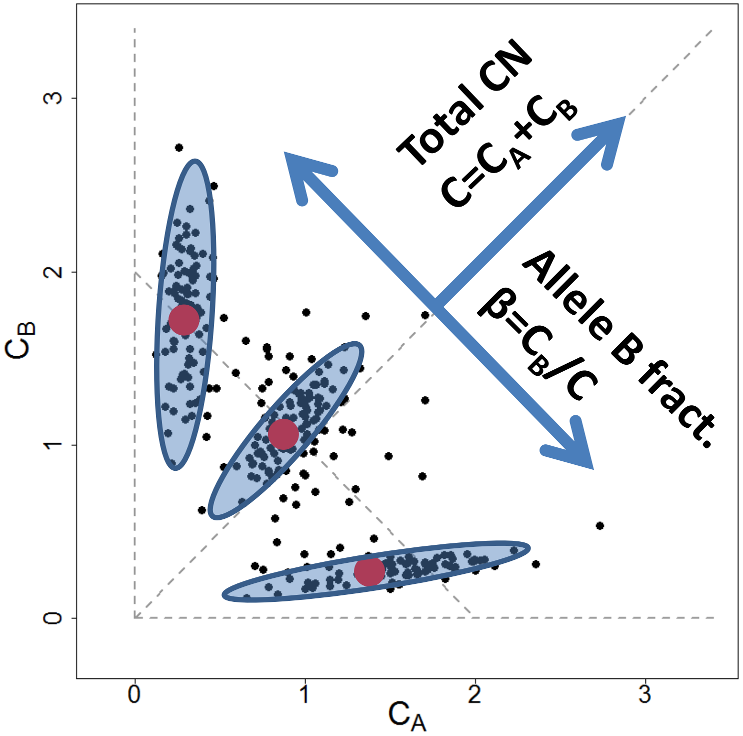
Systematic effects… **…are SNP specific!**



Allele B fractions (BAFs): The bias is greater than the noise

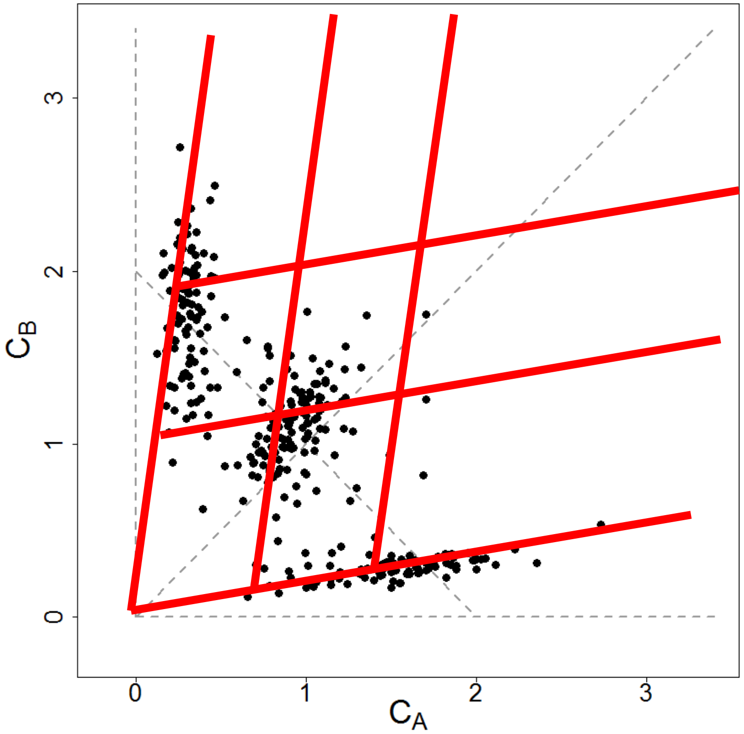
Example: (C_A,C_B) for <u>310 samples</u> per SNP.

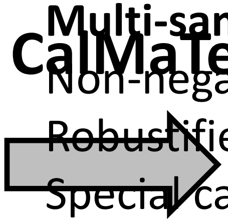
TCN: between 2 arrays. BAF: within array.

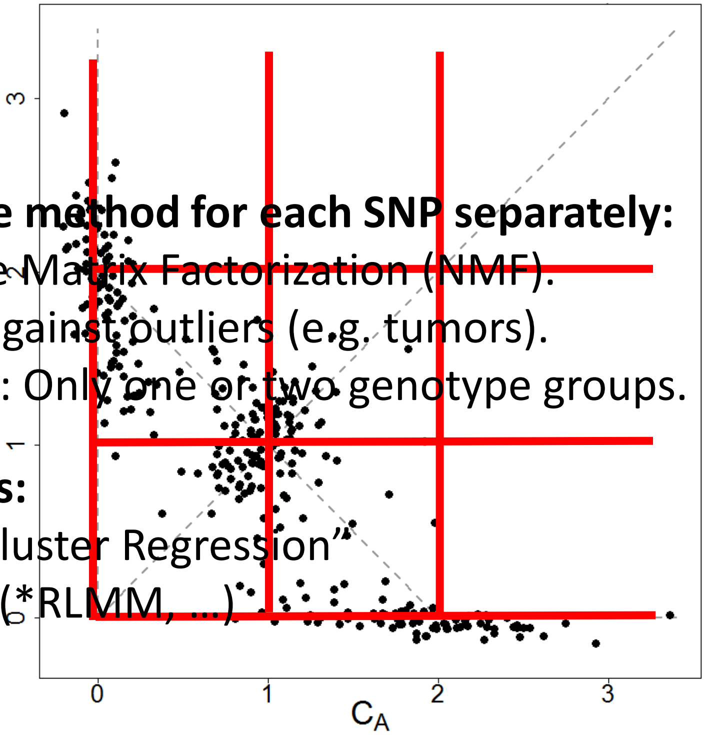


SNP #1053

Multi-sample model: (one per SNP)
Fit affine transform across samples





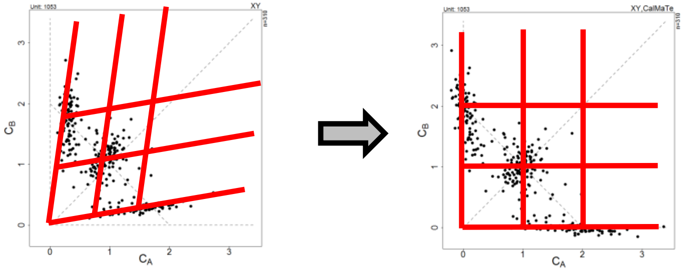


CalMaTe

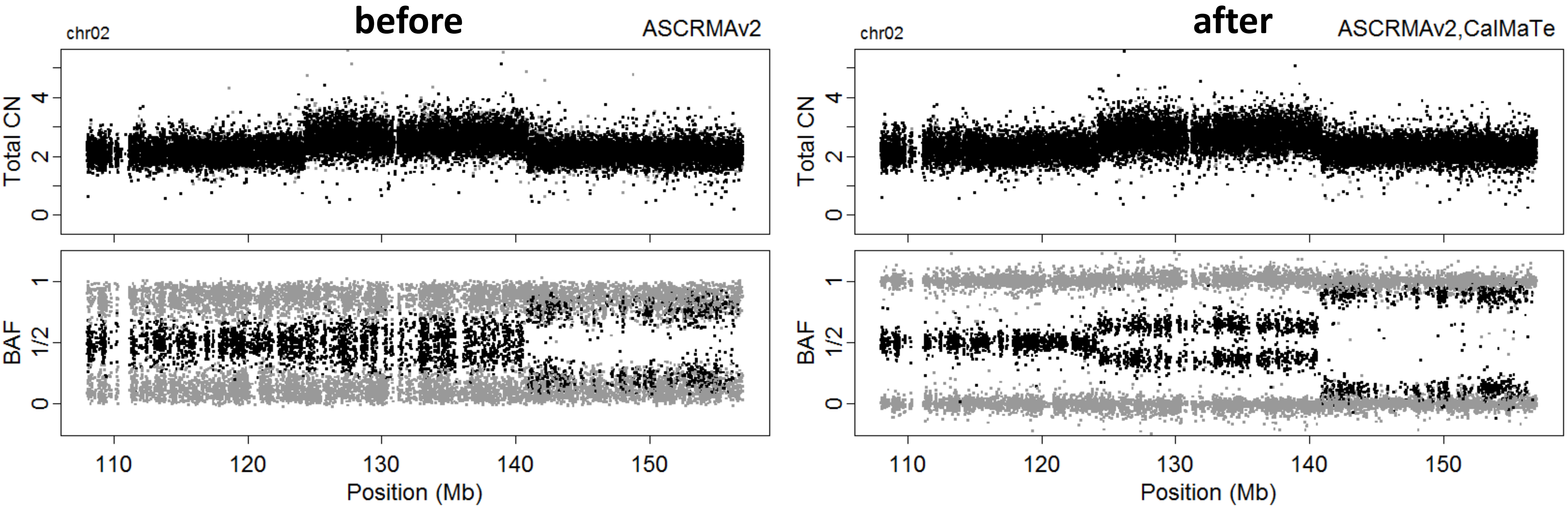
Multi-sample method for each SNP separately:
Non-negative Matrix Factorization (NMF).
Robustified against outliers (e.g. tumors).
Special cases: Only one or two genotype groups.

Related ideas:
Illumina's "Cluster Regression"
CRLMM CNs (*RLMM, ...)
...

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



Estimate & Backtransform
Repeat for all 1,000,000 SNPs



The above is the chromosomal plot for one sample of the 310 samples.

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)

More: 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.

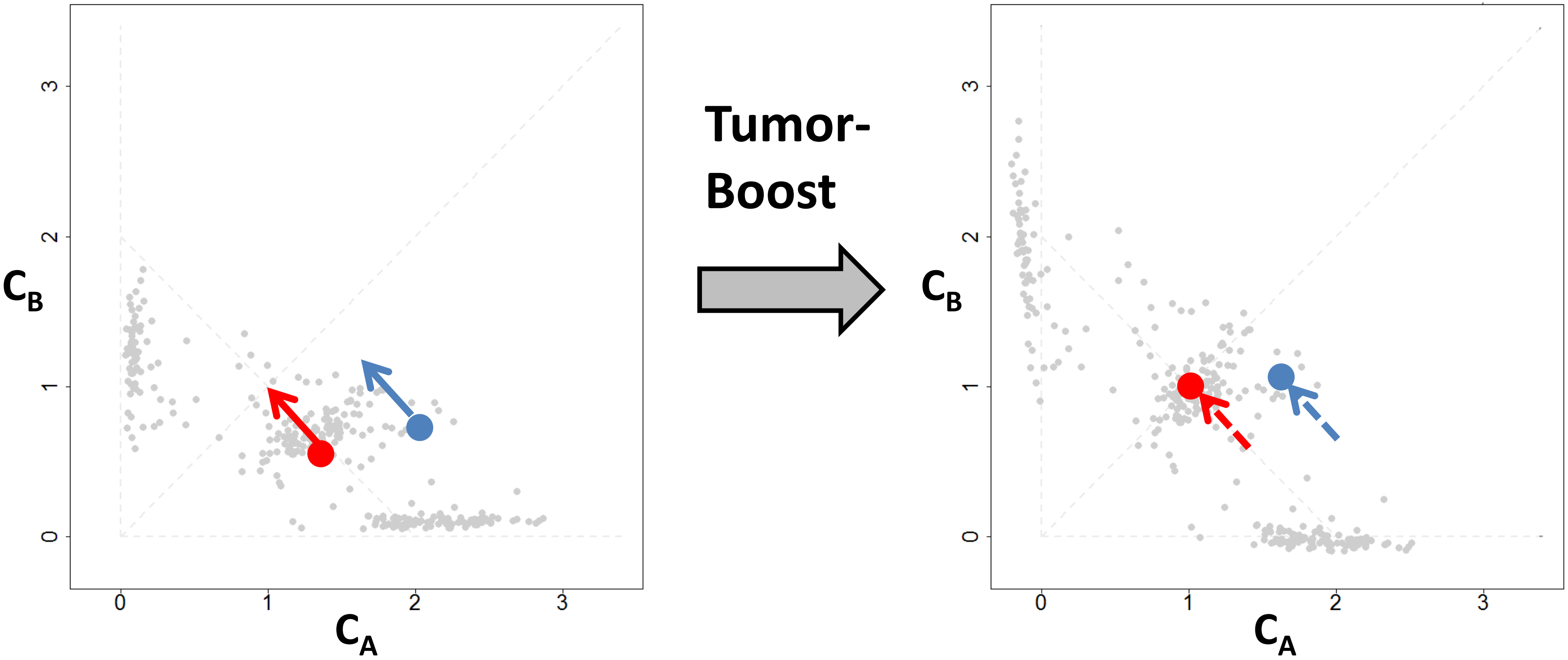
The tumor “should be” close to its normal

When we have only a single tumor-normal pair:

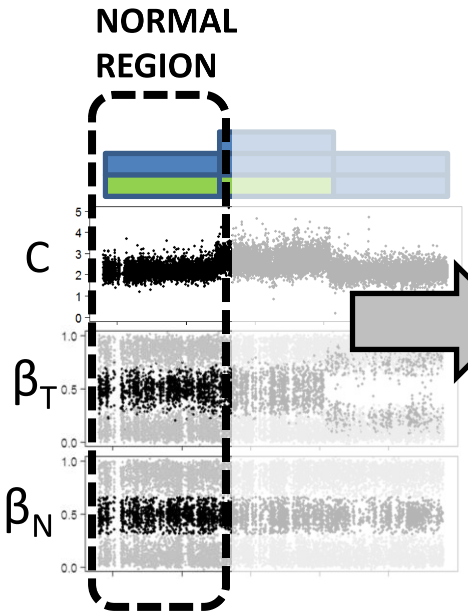
(i) **Normal** should be at e.g. (1,1) …so lets move it there!

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

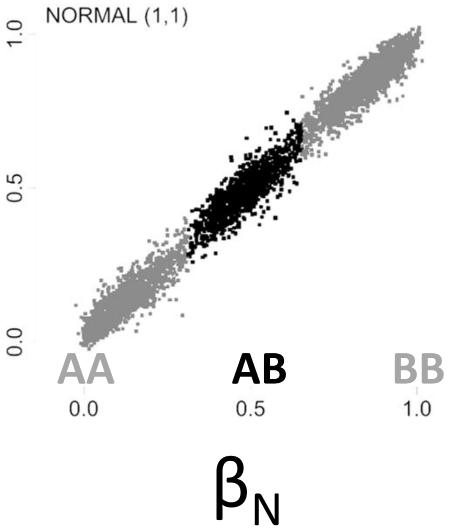
One SNP,
a tumor-normal pair



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



β_T



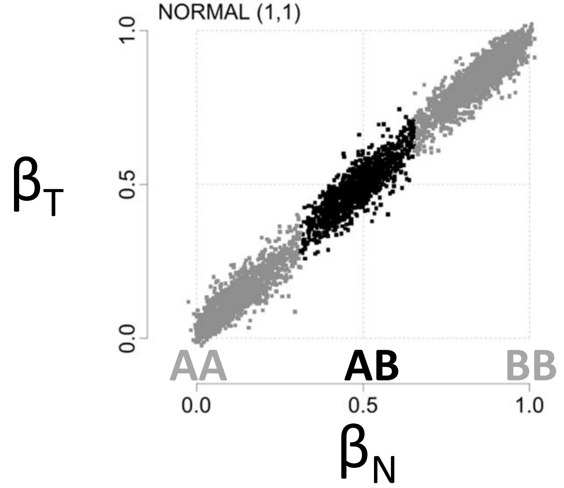
For each genotype:

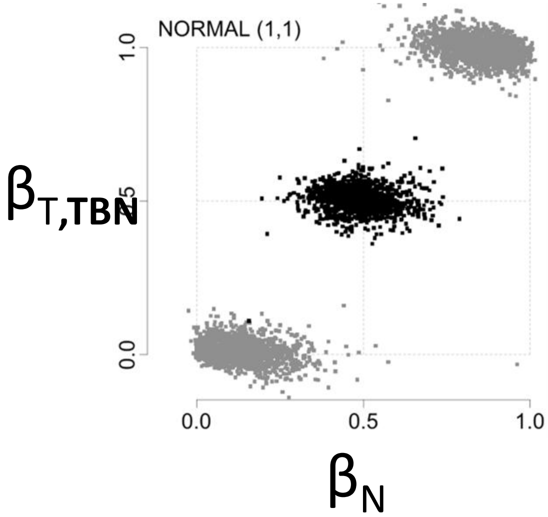
$\mathrm{Cor}(\beta_T, \beta_N) \approx 1$

A shared SNP effect:

systematic variation

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:

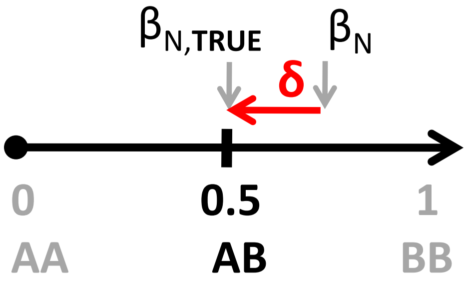
SNP effect

$\delta = \beta_N - \beta_{N,TRUE}$

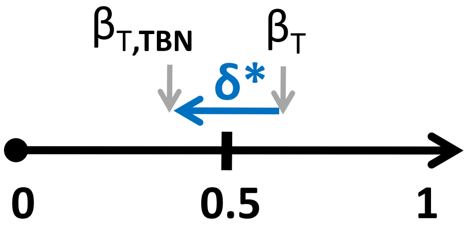
Remove from tumor:

$\beta_{T,TBN} = \beta_T - \delta^*$

1. Estimate SNP effect in the normal and its genotypes



2. Remove SNP effect from the tumor



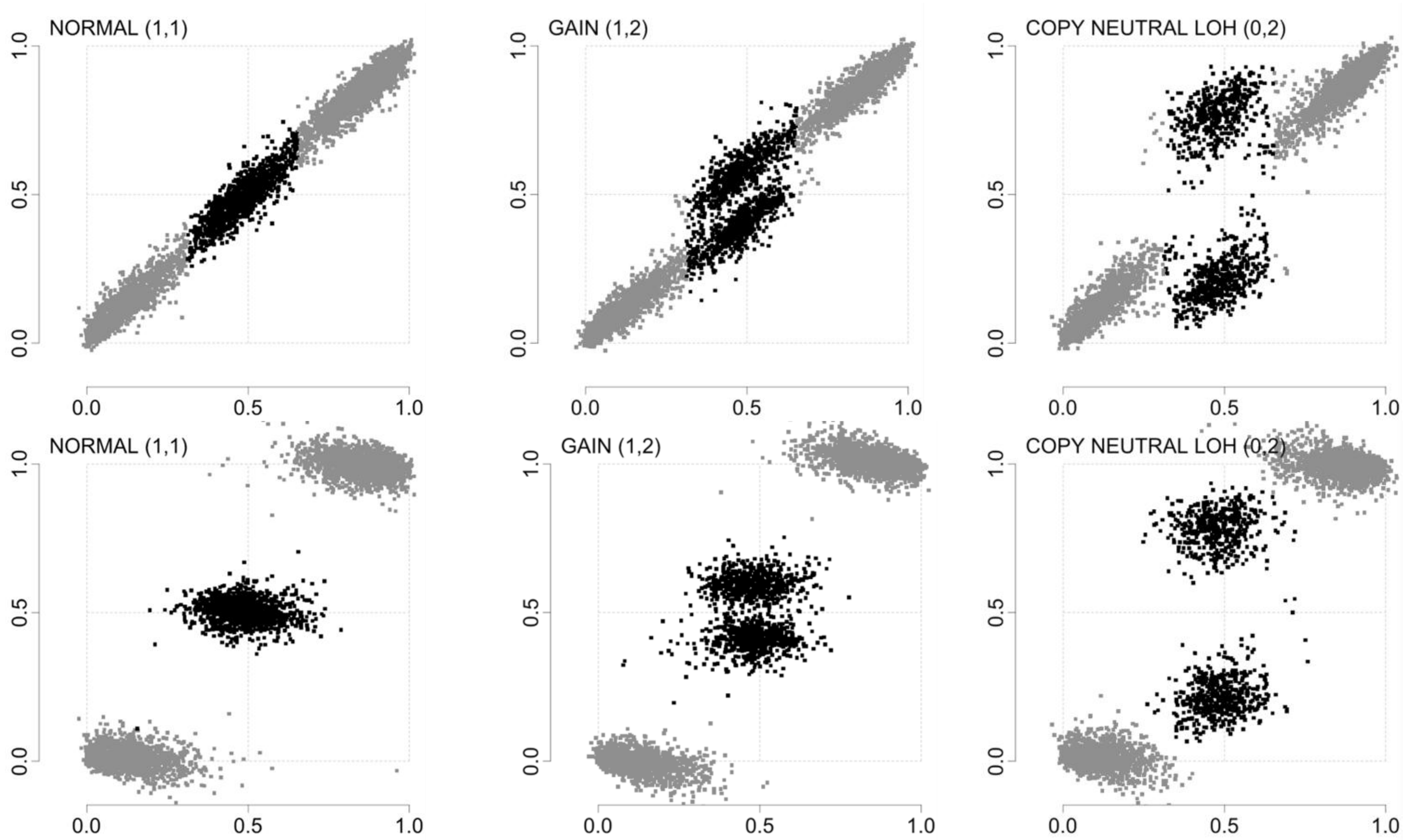
3. Repeat for all SNPs.

TumorBoost removes the SNP effects from the tumor (only)

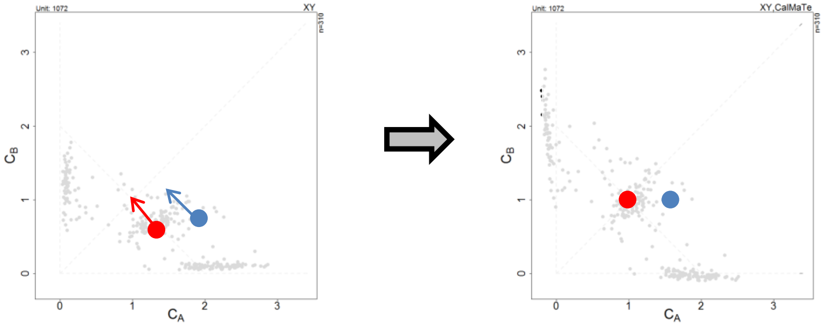
NORMAL (1,1) **GAIN (1,2)** **CN-LOH (0,2)**

Before:

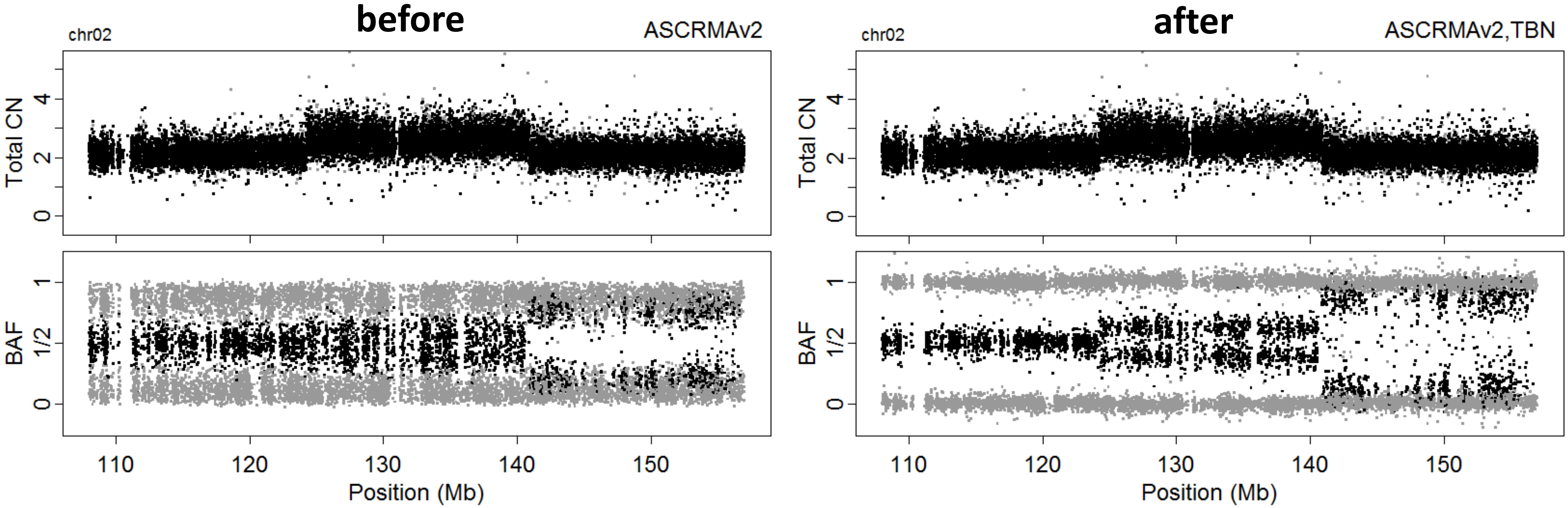
After:



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



Estimate & Backtransform
Repeat for all 1,000,000 SNPs



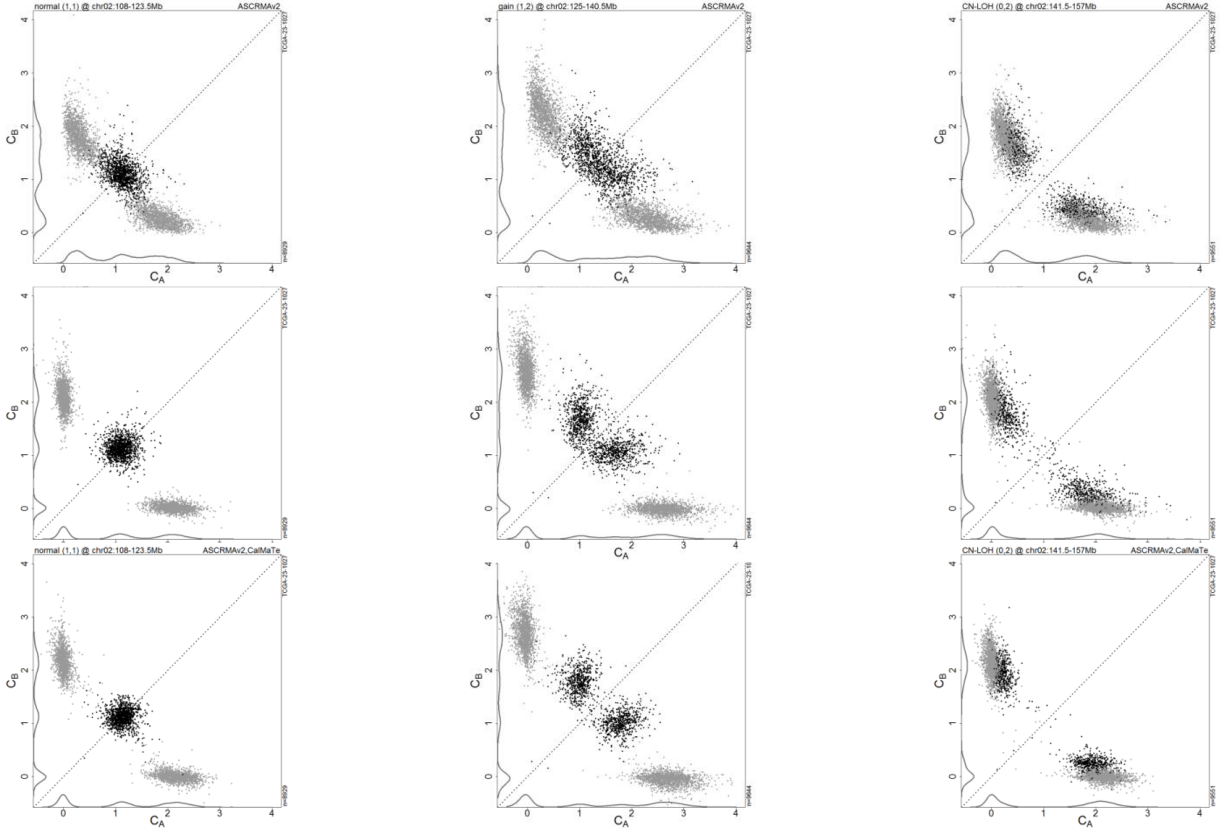
TumorBoost => more distinct (C_A,C_B) - key for PSCN segmentation

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

Original:

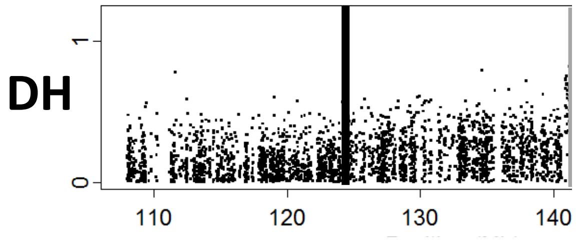
TumorBoost:
- single-pair
- tumor-normals
- normal is
  not corrected

CalMaTe:
- multi-sample

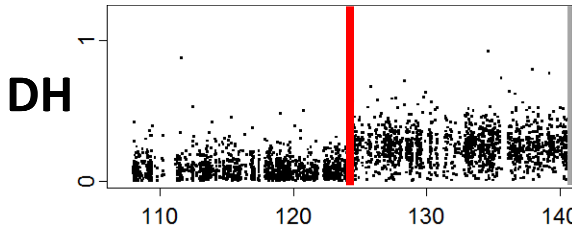


TumorBoost and CalMaTe significantly improve power to detect change points !

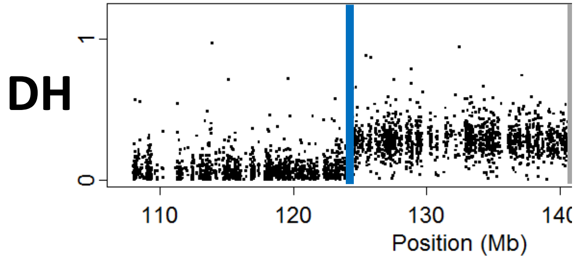
Original

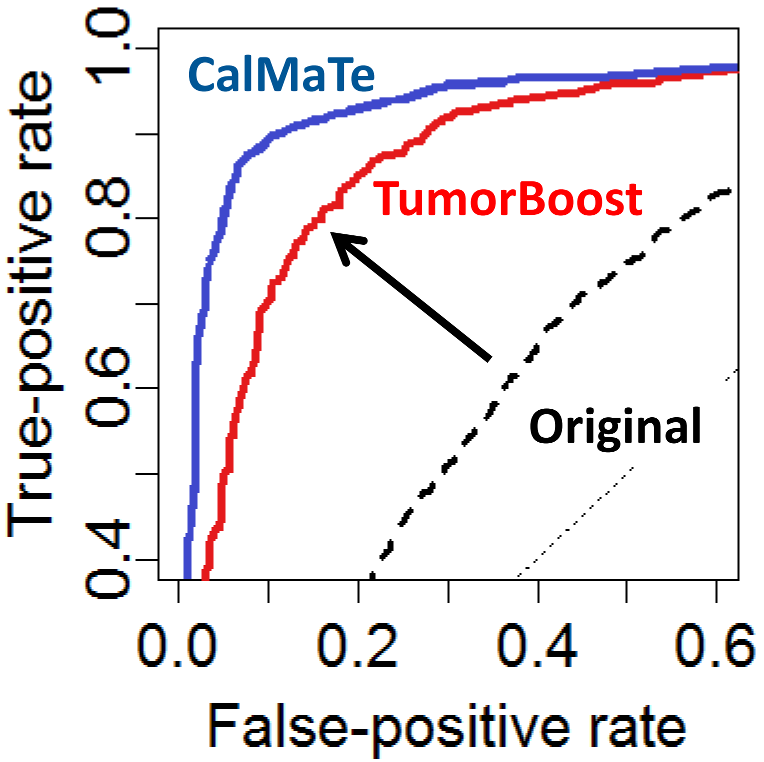


TumorBoost (single pair)



CalMaTe (multi-sample)





Assessment: 1 sample, 1 change point

Paired PSCBS

Parent-specific copy numbers from
a single tumor-normal pair of SNP arrays

1. Tumor-normal pair
2. Genotype normal
3. Normalize tumor using normal
4. CBS segment tumor: (a) TCN, then (b) DH
5. Estimate PSCNs within segments
6. Call segments

Total CNs & DHs segmentation gives us PSCN regions and estimates

(i) Find change points

(ii) Estimate mean levels

Total CNs

$C = C_A + C_B$

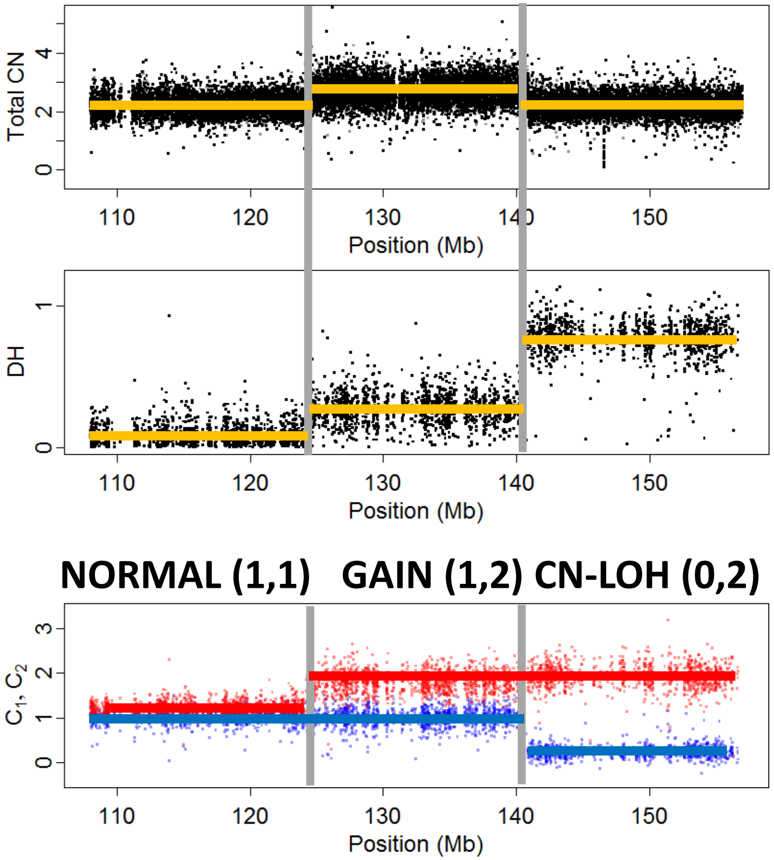
Decrease in Heterozygosity

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

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)

avg(all loci) *
avg(hets only)

Calling allelic balance and LOH

Calling allelic balance:

- Null: $C_1 = C_2$ (equivalent to DH = 0)
- DH is estimated with bias near 0, so we need offset Δ_{AB} in test.
- Reject null if α:th percentile of bootstrap-estimated DH - $\Delta_{AB} > 0$.
- How do we choose Δ_{AB}?

Calling LOH:

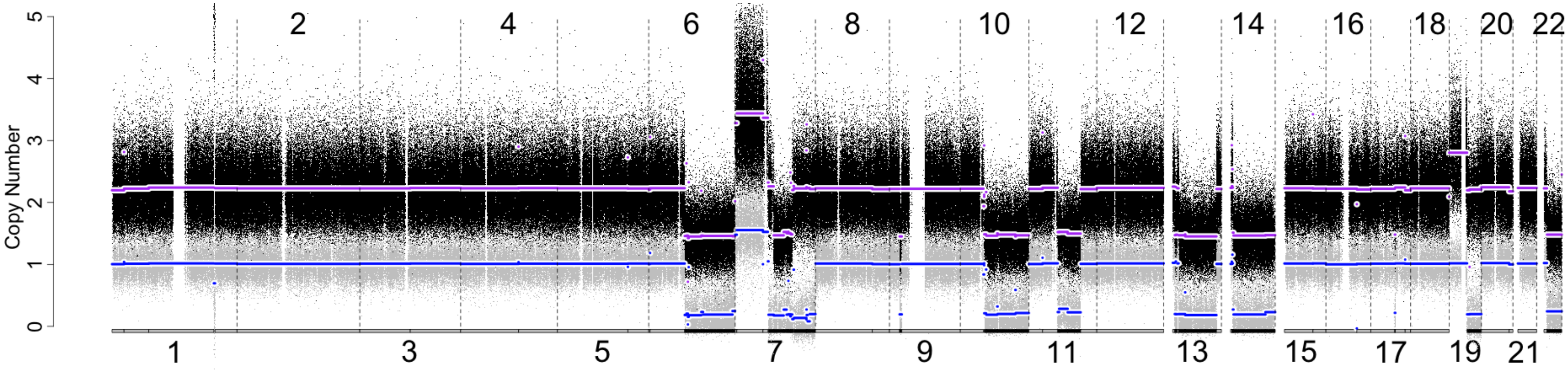
- Null: $C_1 > 0$ ("not in LOH")
- C_1 is estimated with bias due to background (e.g. normal contamination), so we need offset Δ_{LOH} in test.
- Reject null if (1-α):th percentile of bootstrap-estimated $C_1 - \Delta_{LOH} < 0$.
- How do we choose Δ_{LOH}?

Results

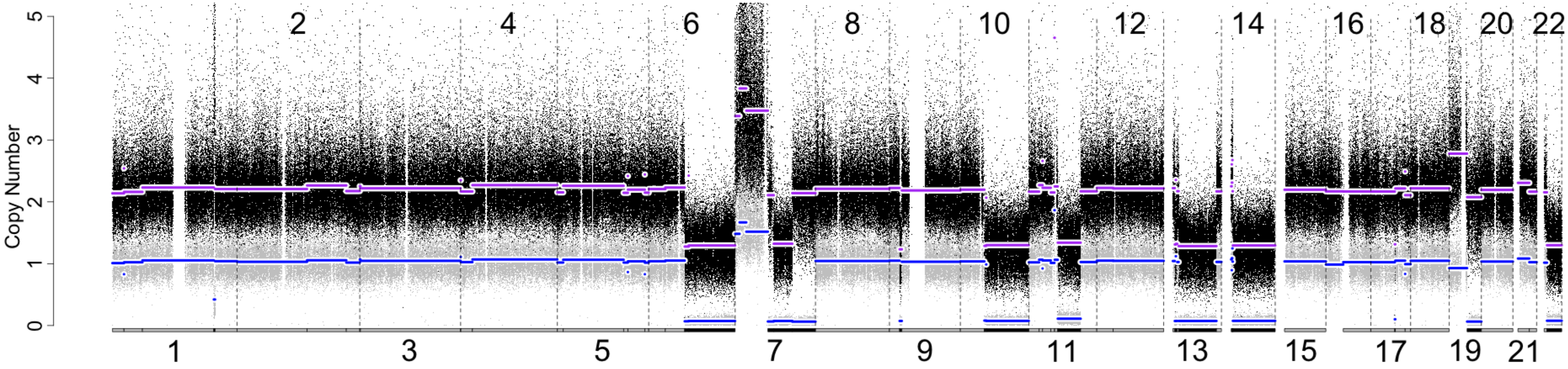
PSCBS works with any SNP array
- similar results on Affymetrix and Illumina

!

Affymetrix GenomeWideSNP_6



Illumina HumanHap550



Other methods exists
e.g. Paired BAF segmentation

Paired BAF (Staaf et al., 2008) is a paired.

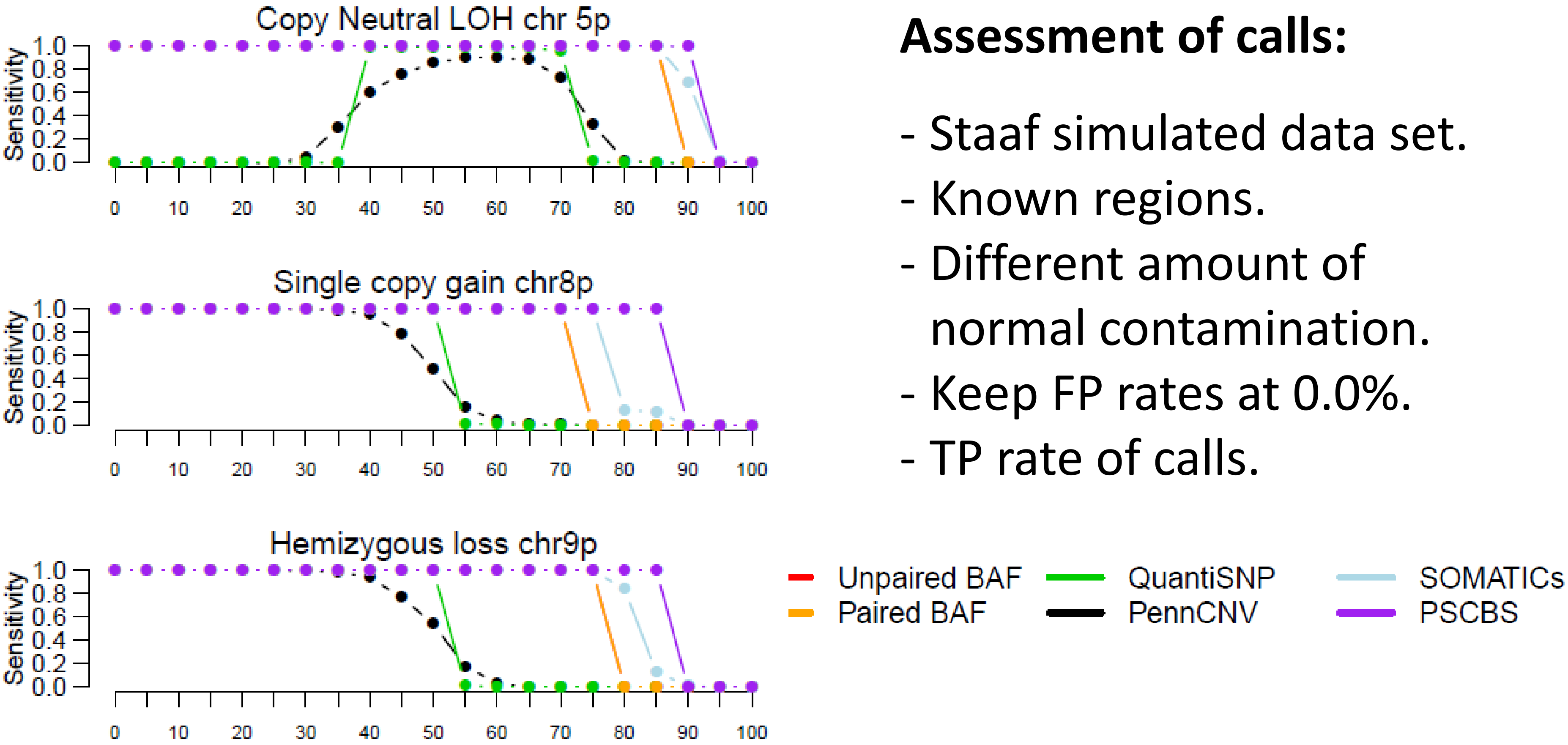
Algorithm:

1. Genotype normal sample
2. Drop homozygote SNPs
3. Segment “mirrored BAF” (like DH)
4. Estimate parent-specific copy numbers

Paired PSCBS performs very well compared to other PSCN methods !

Assessment of calls:

- Staaf simulated data set.
- Known regions.
- Different amount of normal contamination.
- Keep FP rates at 0.0%.
- TP rate of calls.



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

Preprocessing:

- Affymetrix: ASCRMAv2 (single-array) [aroma.affymetrix]
- Illumina: <elsewhere>

Normalization of ASCNs:

- Single tumor-normal pair: TumorBoost [aroma.light, aroma.cn]
- Multiple samples: CalMaTe [CalMaTe]

PSCN segmentation:

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

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

Conclusions

Paired PSCBS w/ TumorBoost:

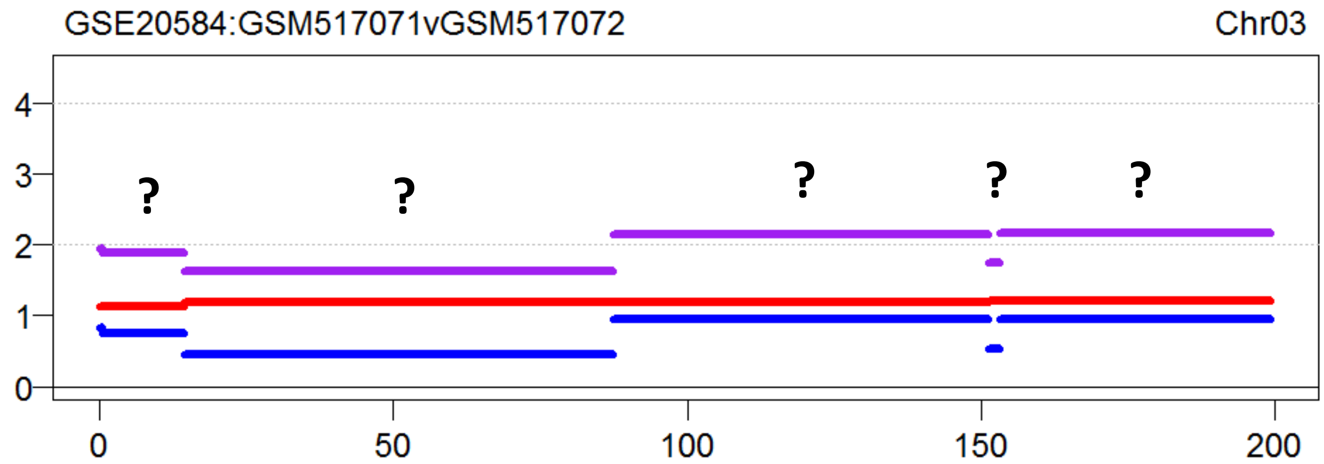
- High quality tumor PSCNs
- Single tumor-normal pair
- No external references needed
- Any SNP microarray technology
- Algorithms is fast and bounded in memory

Future:

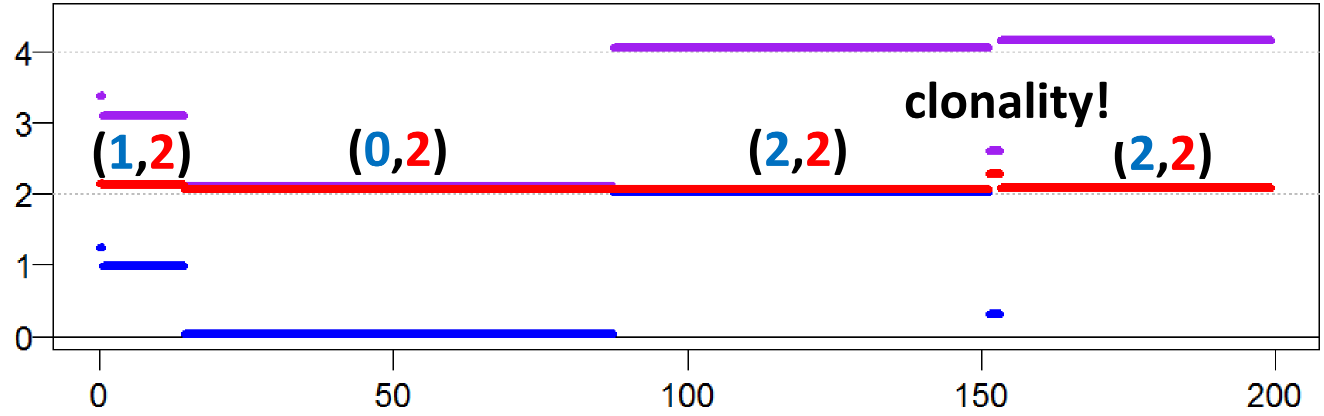
- Non-paired PSCBS
- Calibration of PSCN states (e.g. clonality & ploidy)…

Next: We need to calibrate (C1,C2) before calling! (ongoing work with Pierre Neuvial)

Before



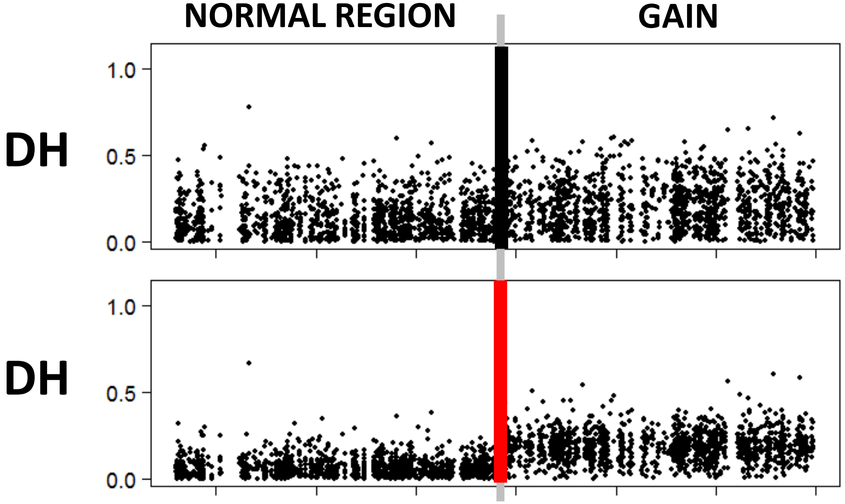
After



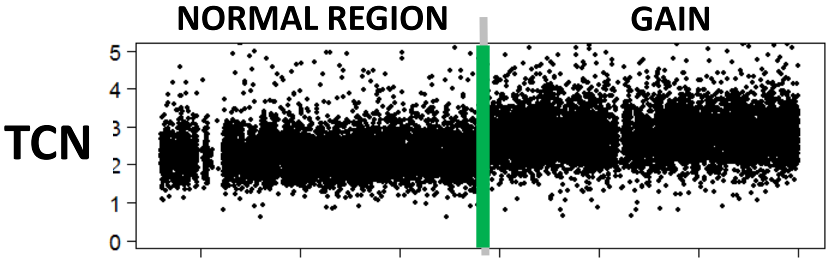
Extra slides

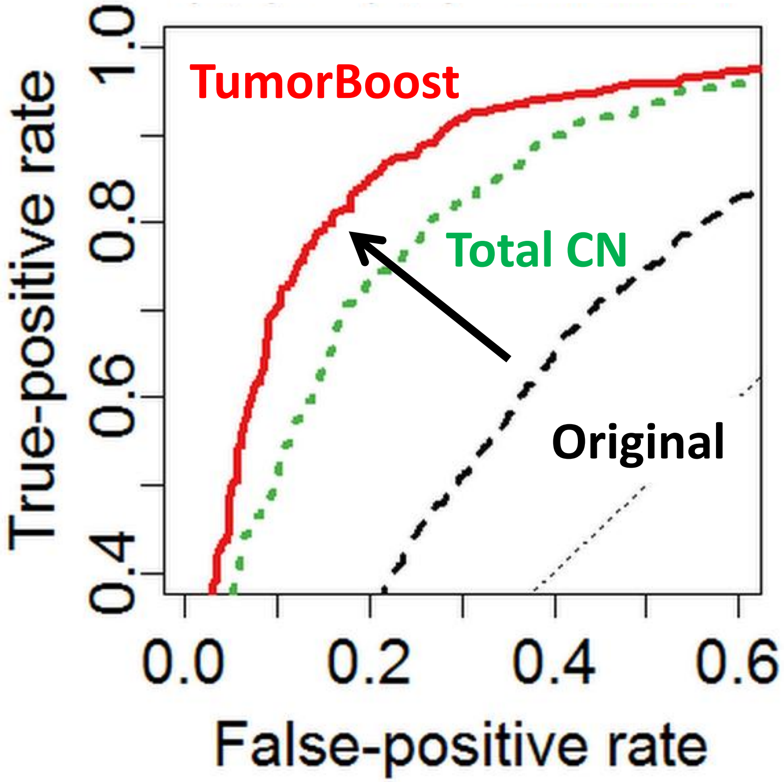
The power to detect a change point varies with type of change!

Decrease in Heterozygosity



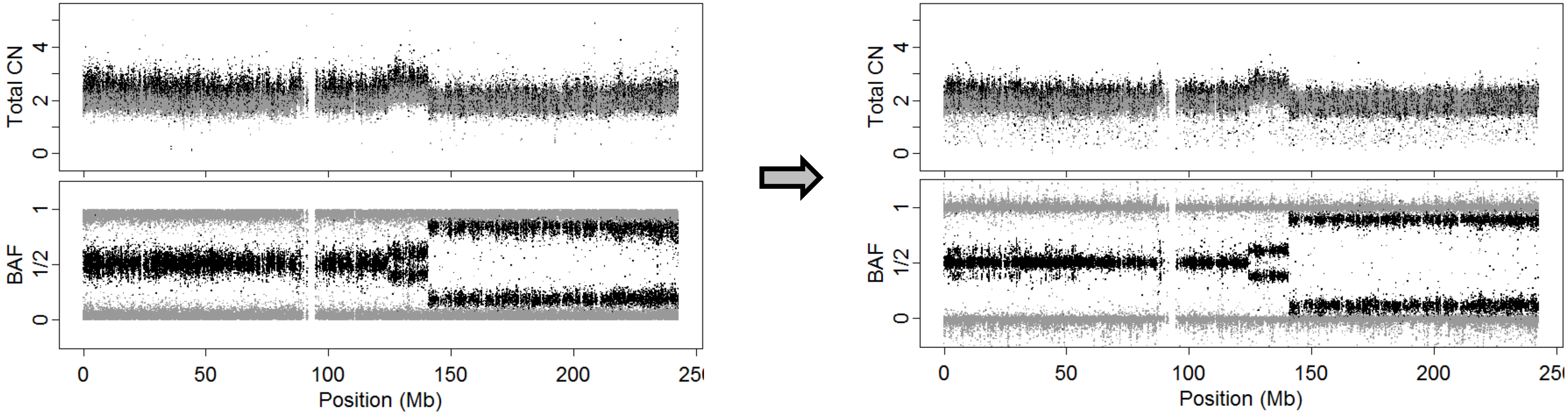
Total CNs



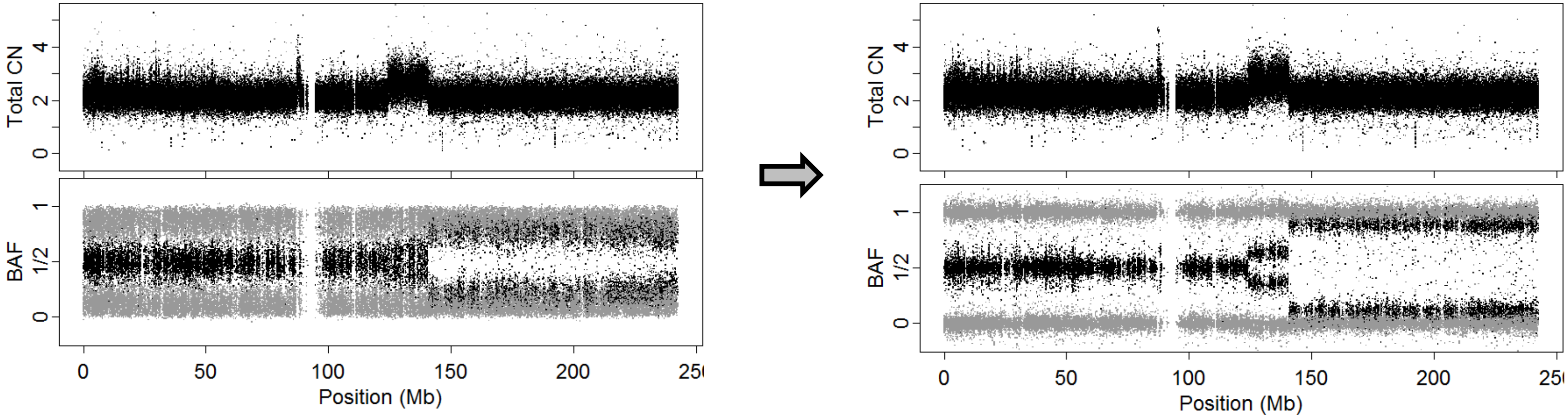


The reason why Illumina is “better” is because they do this calibration - Affymetrix does not.

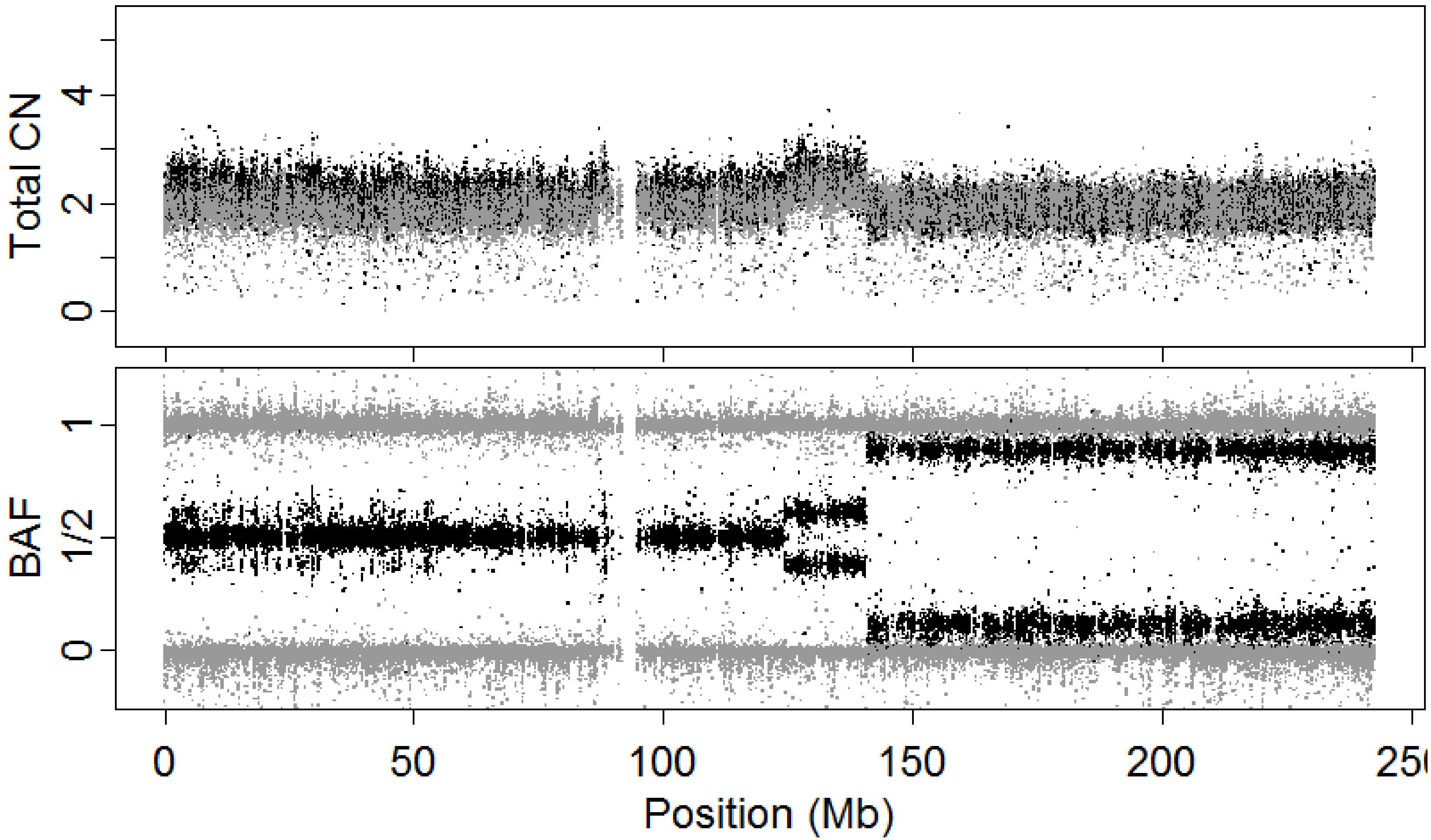
Illumina (Human1M-Duo):



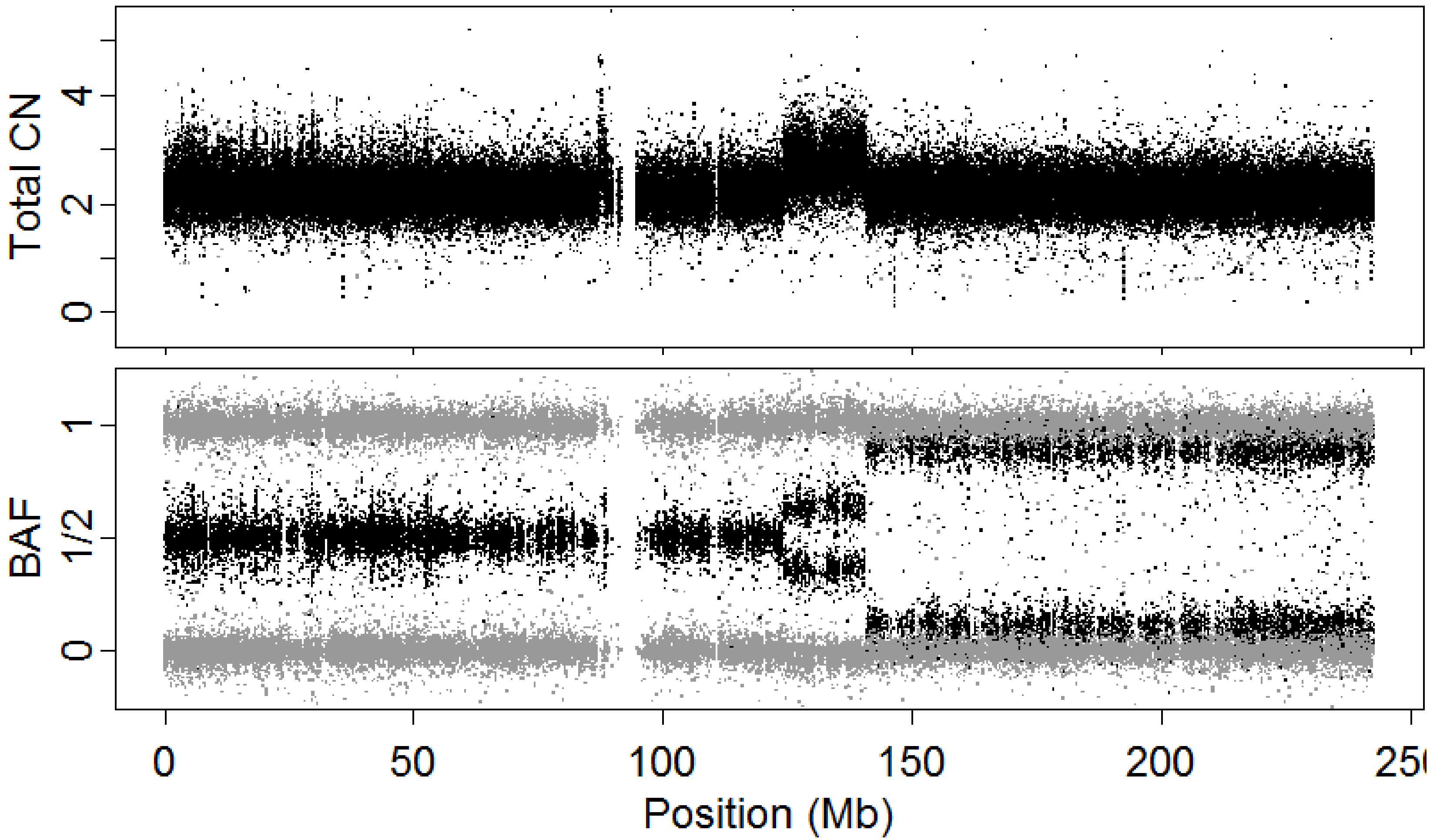
Affymetrix (GenomeWideSNP_6):



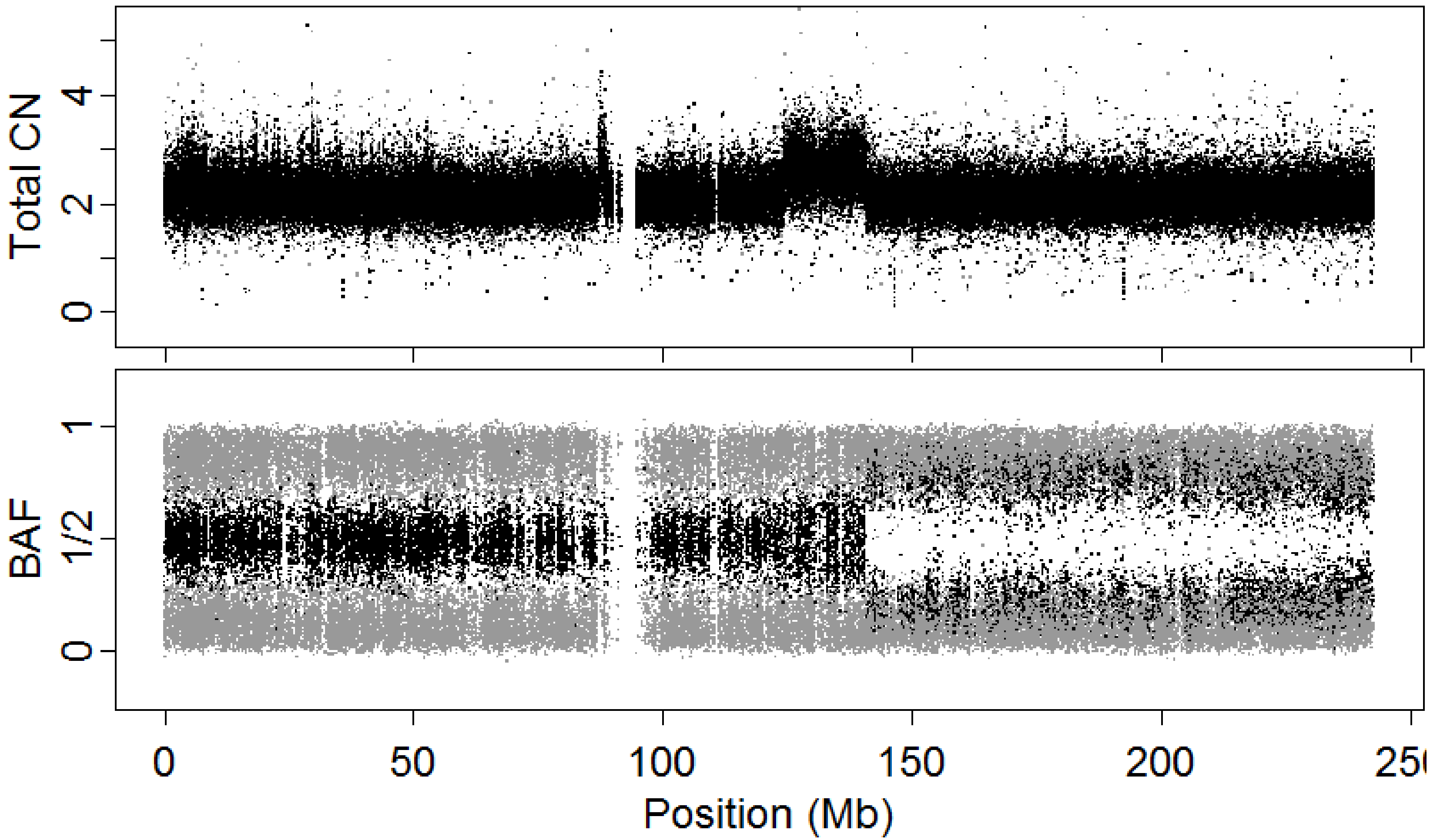
Illumina and Affymetrix have similar noise levels after CalMaTe.



Illumina and Affymetrix have similar noise levels after CalMaTe



Illumina and Affymetrix have similar noise levels after CalMaTe



PSCNs can be estimated at each SNP if we know which SNPs are heterozygous

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

4. SNP-specific (C_1, C_2):
$C_1 = 1/2 * (1 - \rho) * C$
$C_2 = C - C_1$