Single Tumor-Normal Pair Parent-Specific Copy Number Analysis

Henrik Bengtsson

Department of Epidemiology & Biostatistics, UCSF

with: **Pierre Neuvial**

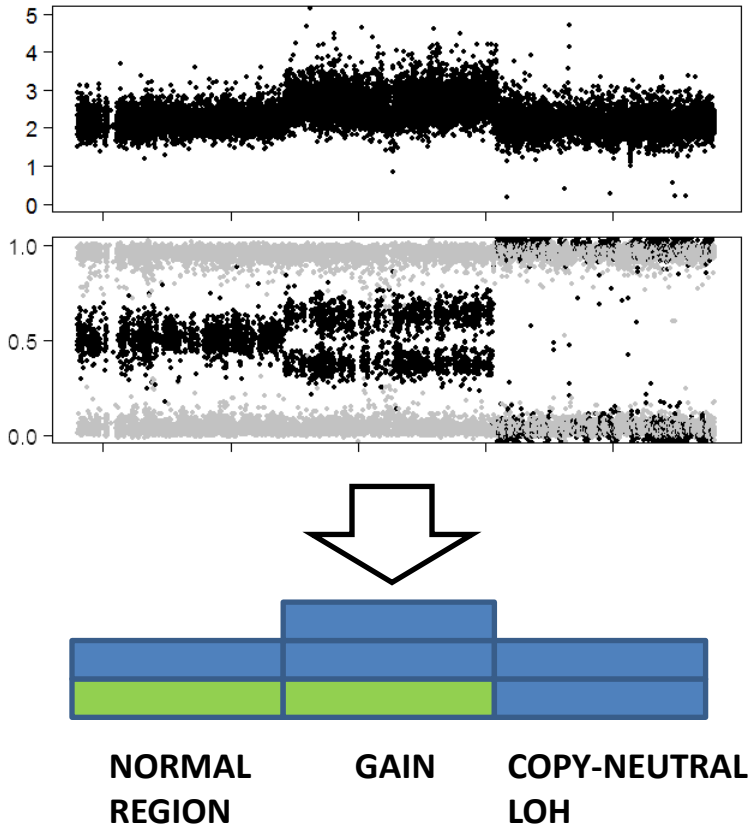
Adam Olshen

Richard Olshen

Venkatraman Seshan

Terry Speed

Paul Spellman



Thanks to: **TCGA, NCI, NHI**

Slides at http://aroma-project.org/

Paired PSCBS

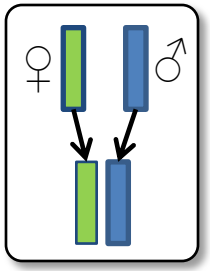
Parent-specific copy numbers from
a single tumor-normal pair of SNP arrays

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

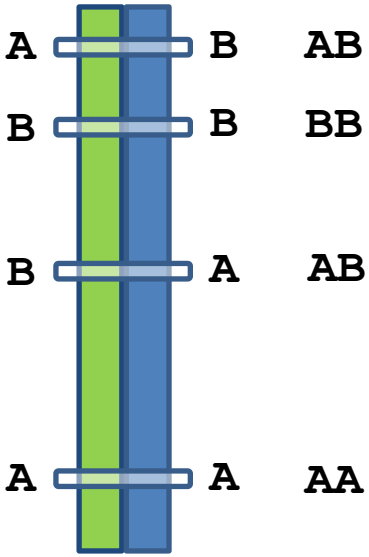
-- H Bengtsson, P Neuvial, TP Speed, TumorBoost: Normalization of allele-specific tumor copy numbers from one single tumor-normal pair of genotyping microarrays, BMC Bioinformatics 2010.

-- 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

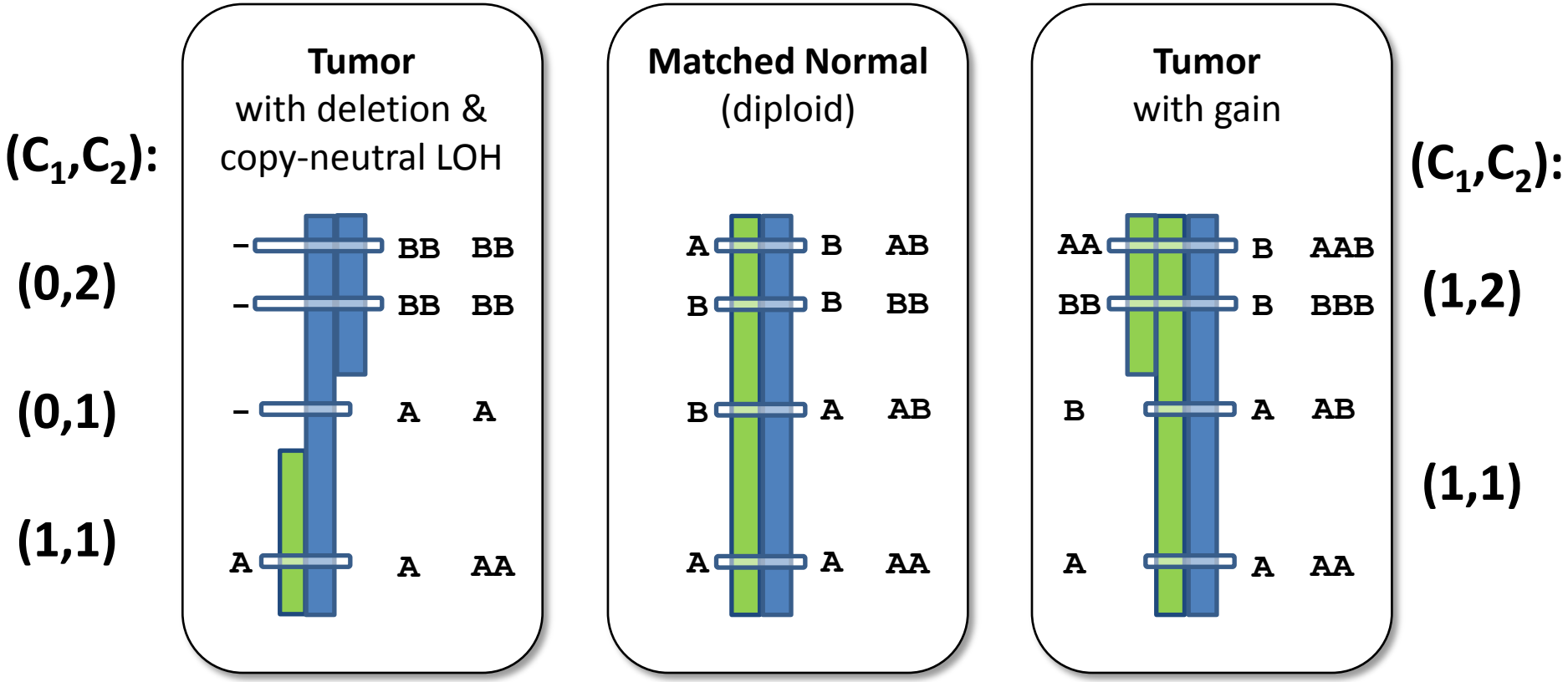


Single nucleotide polymorphism



10-20 million
known SNPs

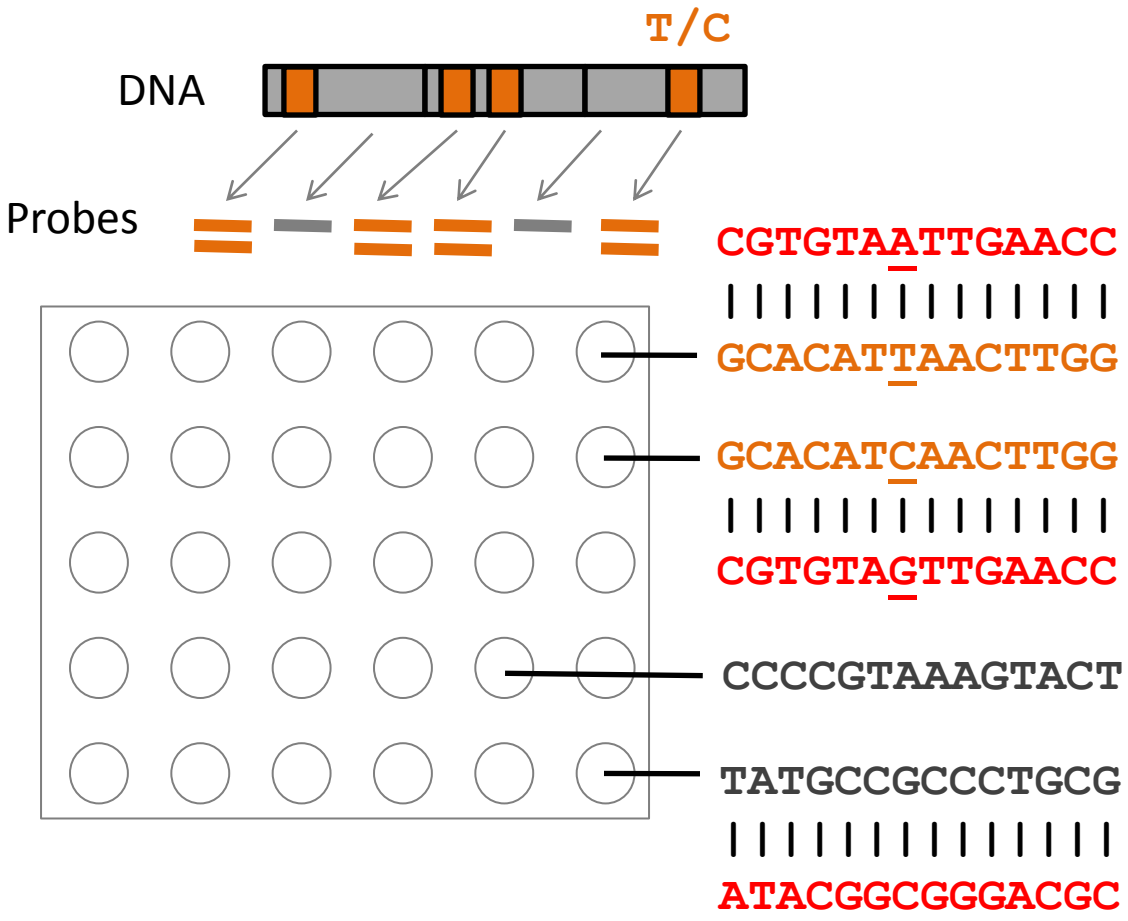
Genotypes and total copy numbers reflect the parent-specific copy numbers



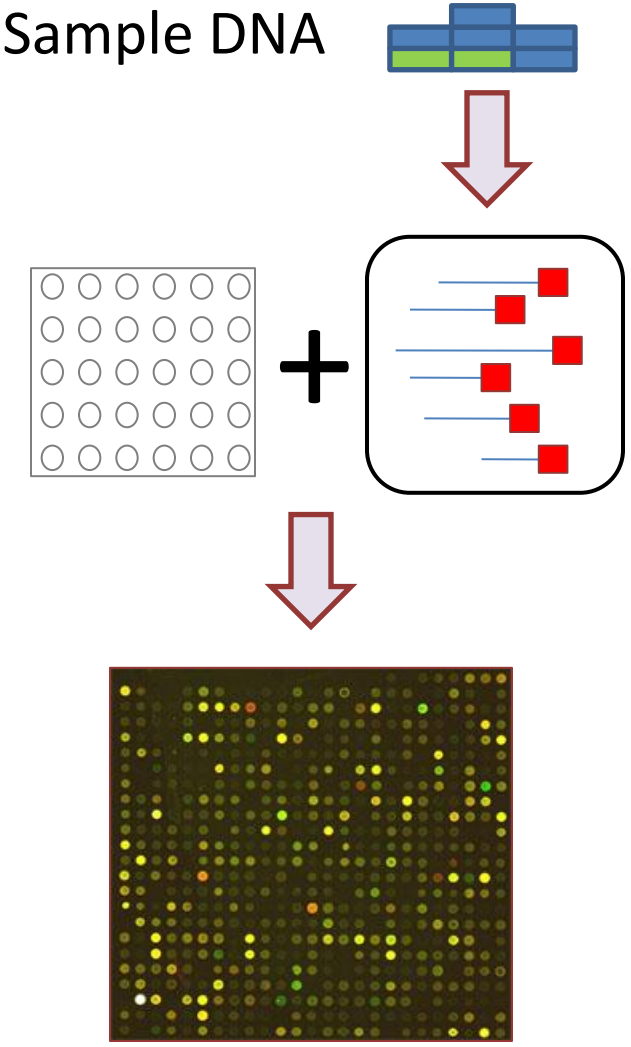
* Occam's razor: Minimal number of events has occurred.

SNP microarrays quantify total and allele-specific copy numbers

Chip Design



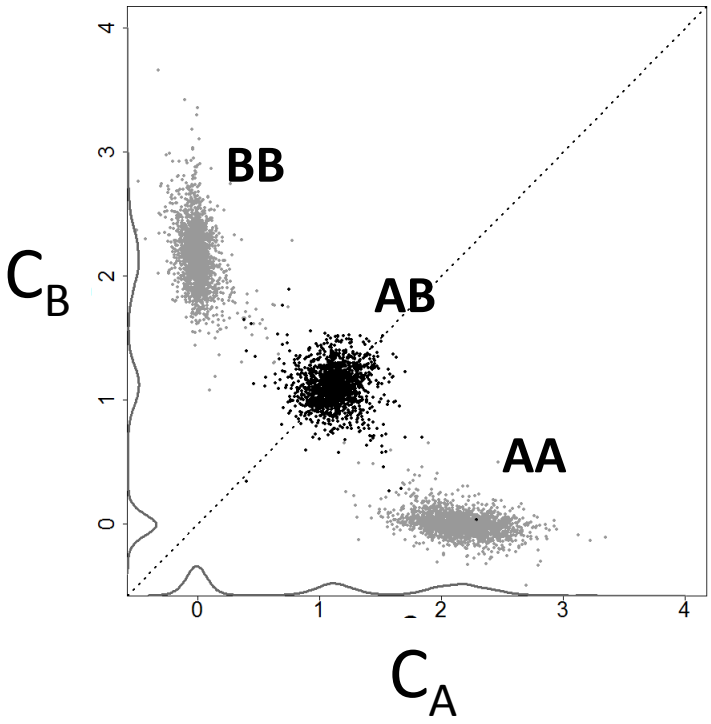
Sample DNA



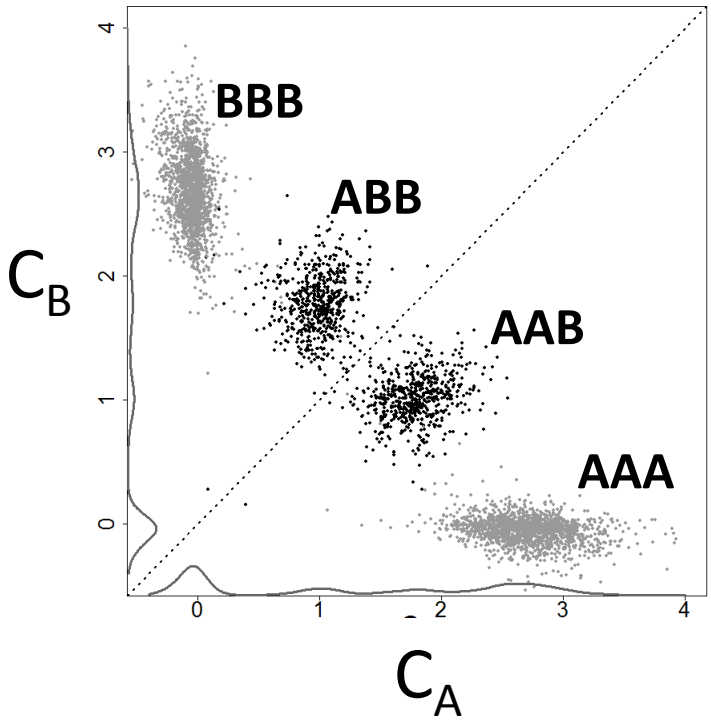
Together the SNPs of a region indicate the parent-specific copy numbers

1 individual, many SNPs

NORMAL (1,1)



GAIN (1,2)

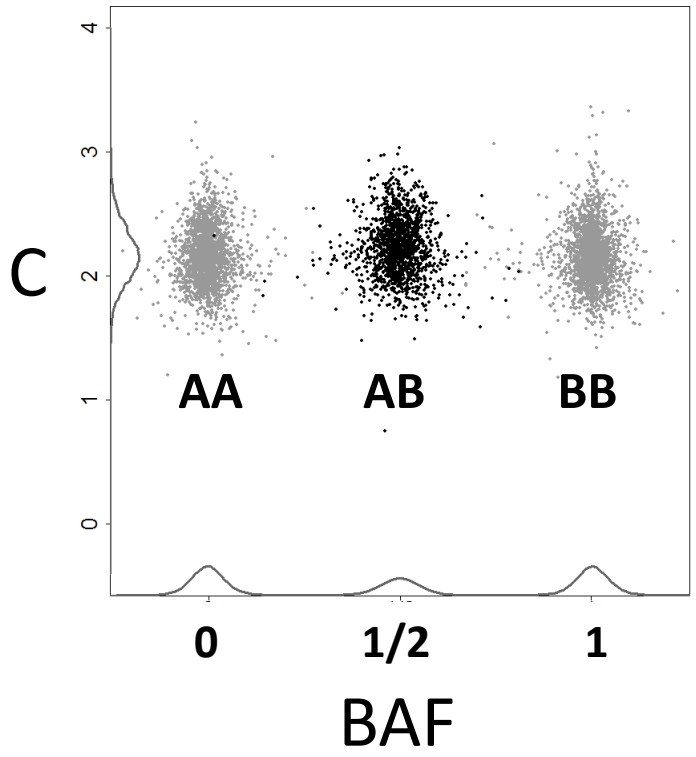


Total CN: $C = C_A + C_B$

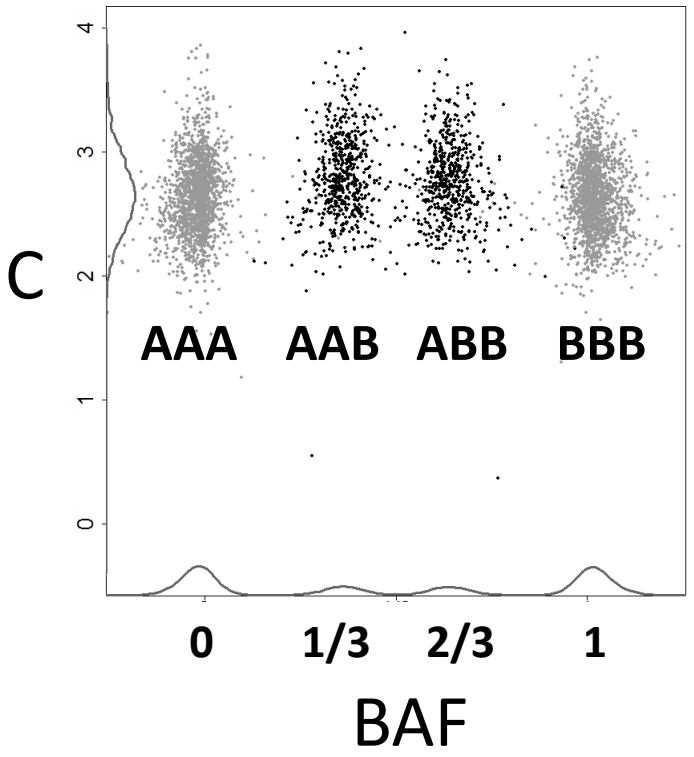
Total CNs and allele B fractions are easier to work with than ASCNs

1 individual, many SNPs, same 2 regions:

NORMAL (1,1)

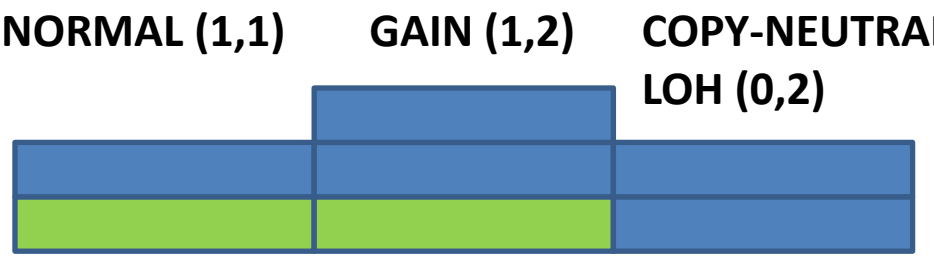


GAIN (1,2)

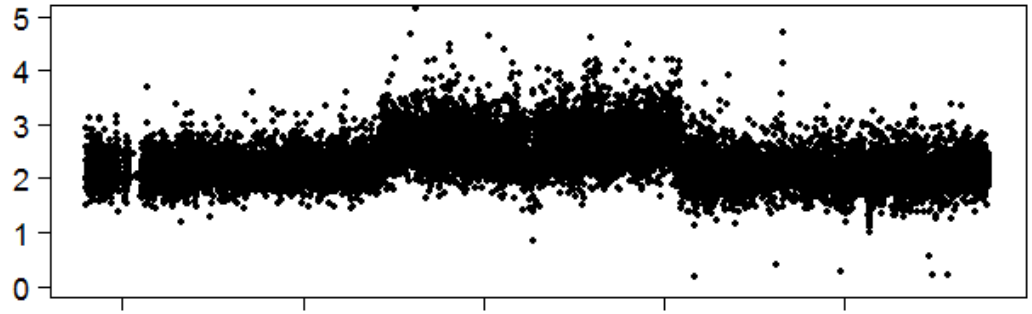


Total CN: $C = C_A + C_B$ BAF: $\beta = C_B / C$

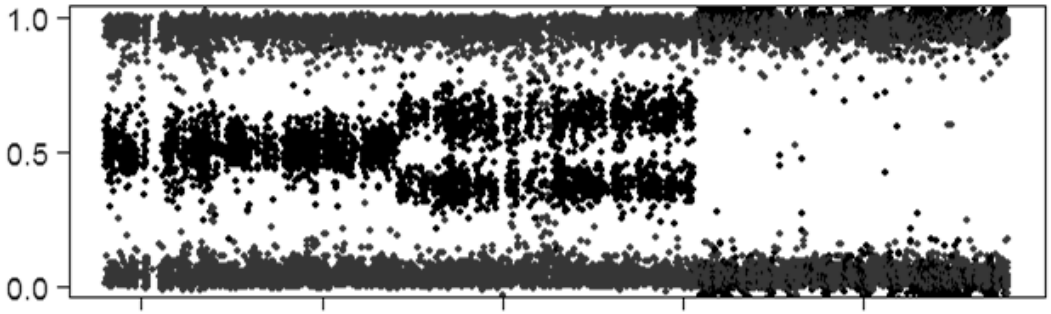
Total CNs and BAFs reflect the underlying parent-specific CNs



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...

- Also, much greater SNRs

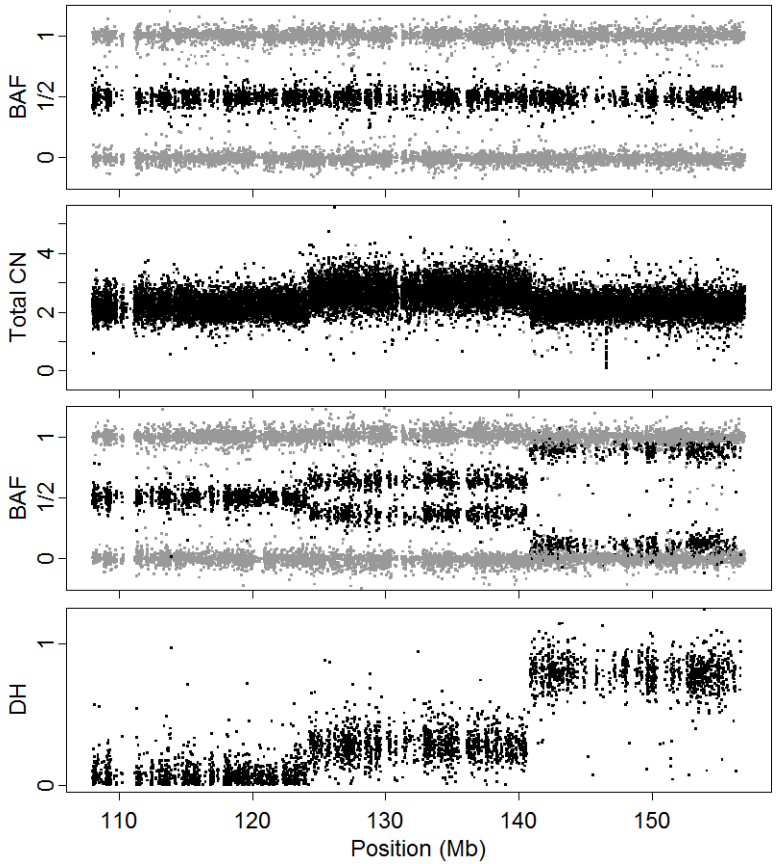
Heterozygous SNPs (not homozygous) are informative for PSCNs

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)

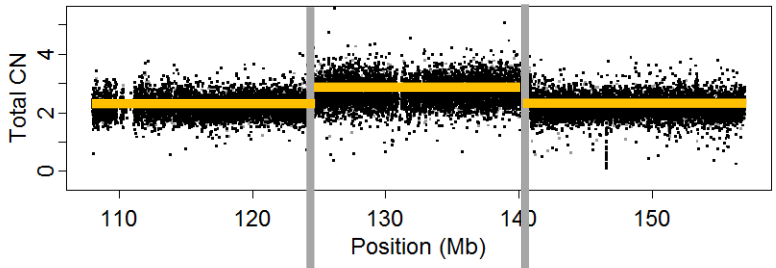
Total CNs & DHs segmentation gives us PSCN regions and estimates

(i) Find change points

(ii) Estimate mean levels

Total CNs

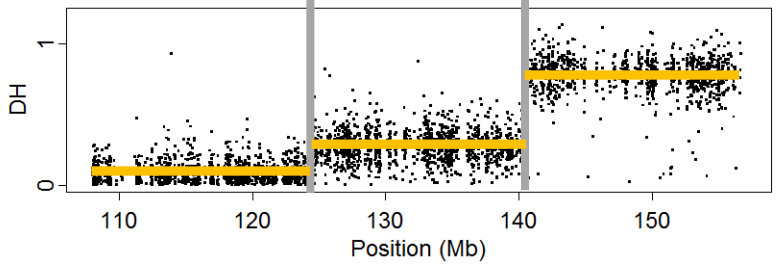
$C = C_A + C_B$



avg(all loci)

Decrease in Heterozygosity

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



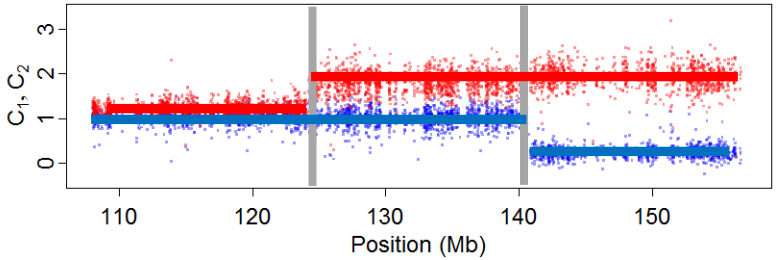
avg(hets only)

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

Per-segment PSCNs (C_1,C_2):

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

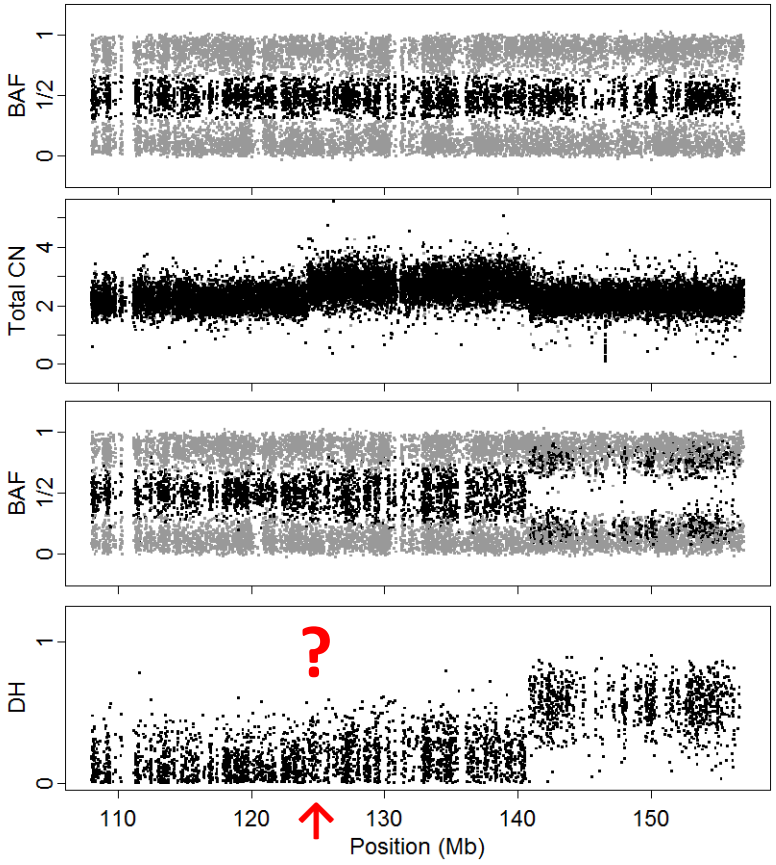
$C_2 = C - C_1$



avg(all loci) * avg(hets only)

It is hard to infer PSCNs reliably when signals are noisy

Actual data:



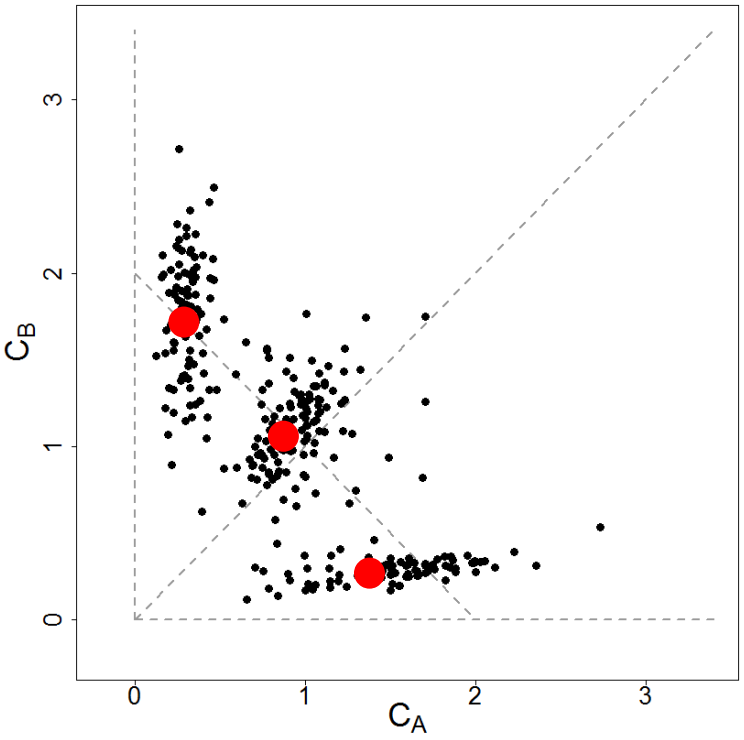
Segmentation may fail…

Let's improve this…

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

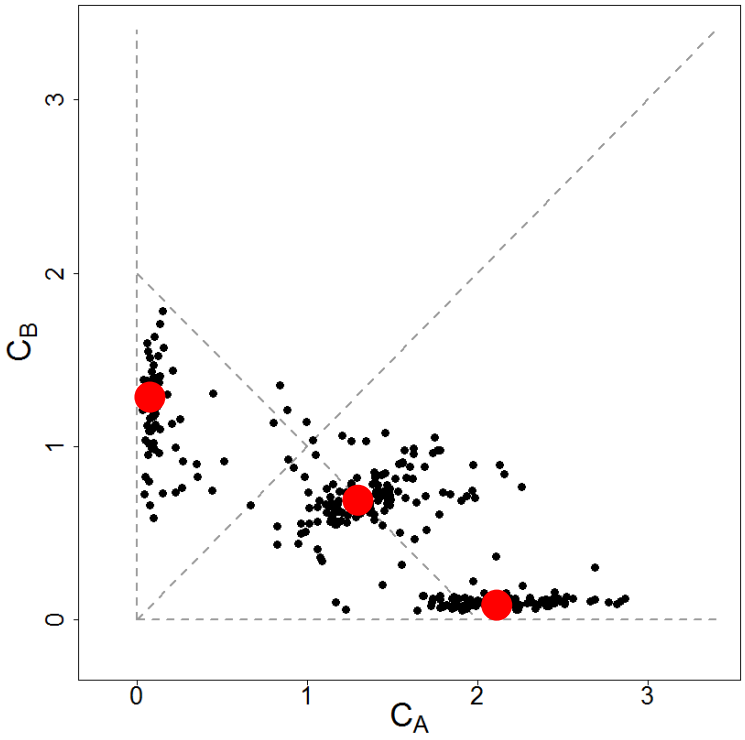
Example: (C_A,C_B) for 310 samples per SNP:

Systematic effects…



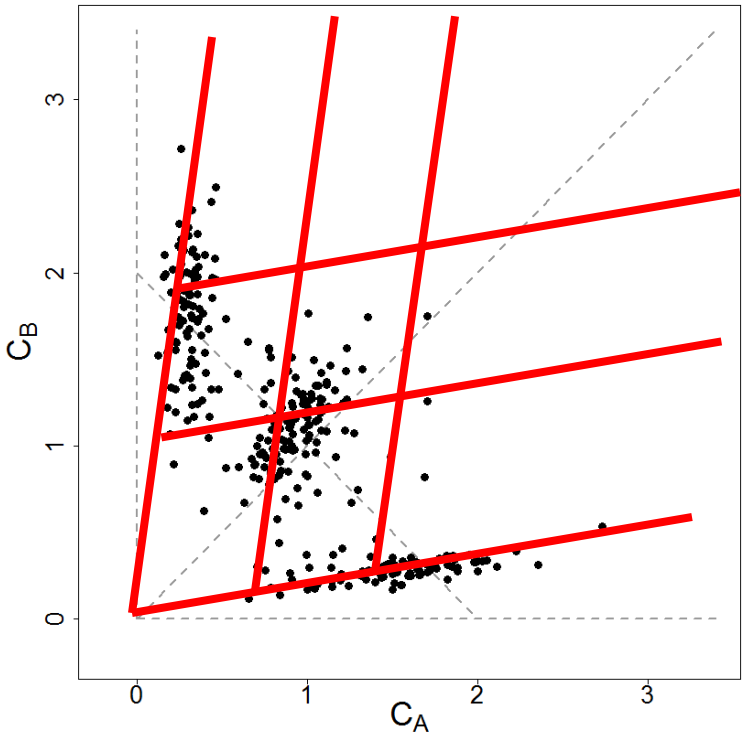
SNP #1053

…are SNP specific!

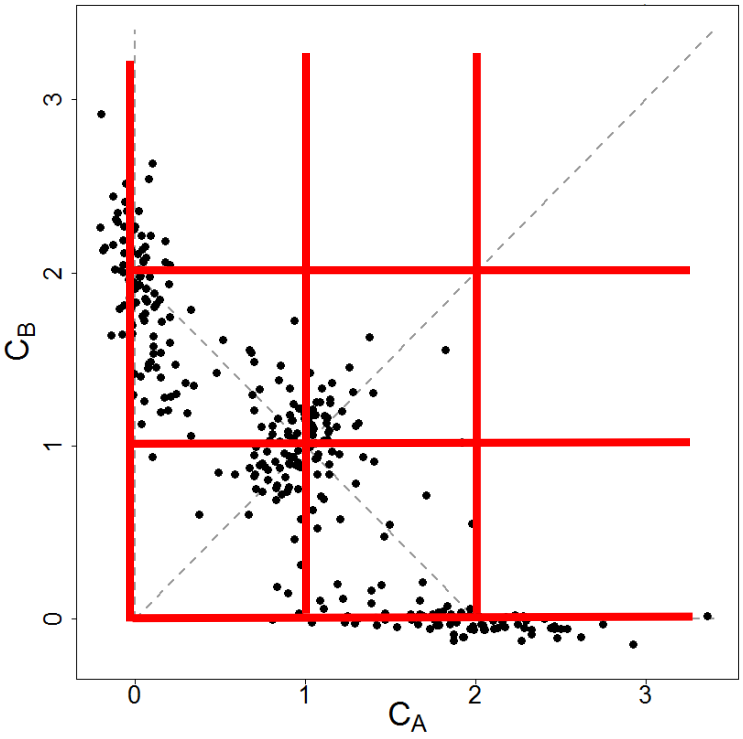


SNP #1072

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



CalMaTe

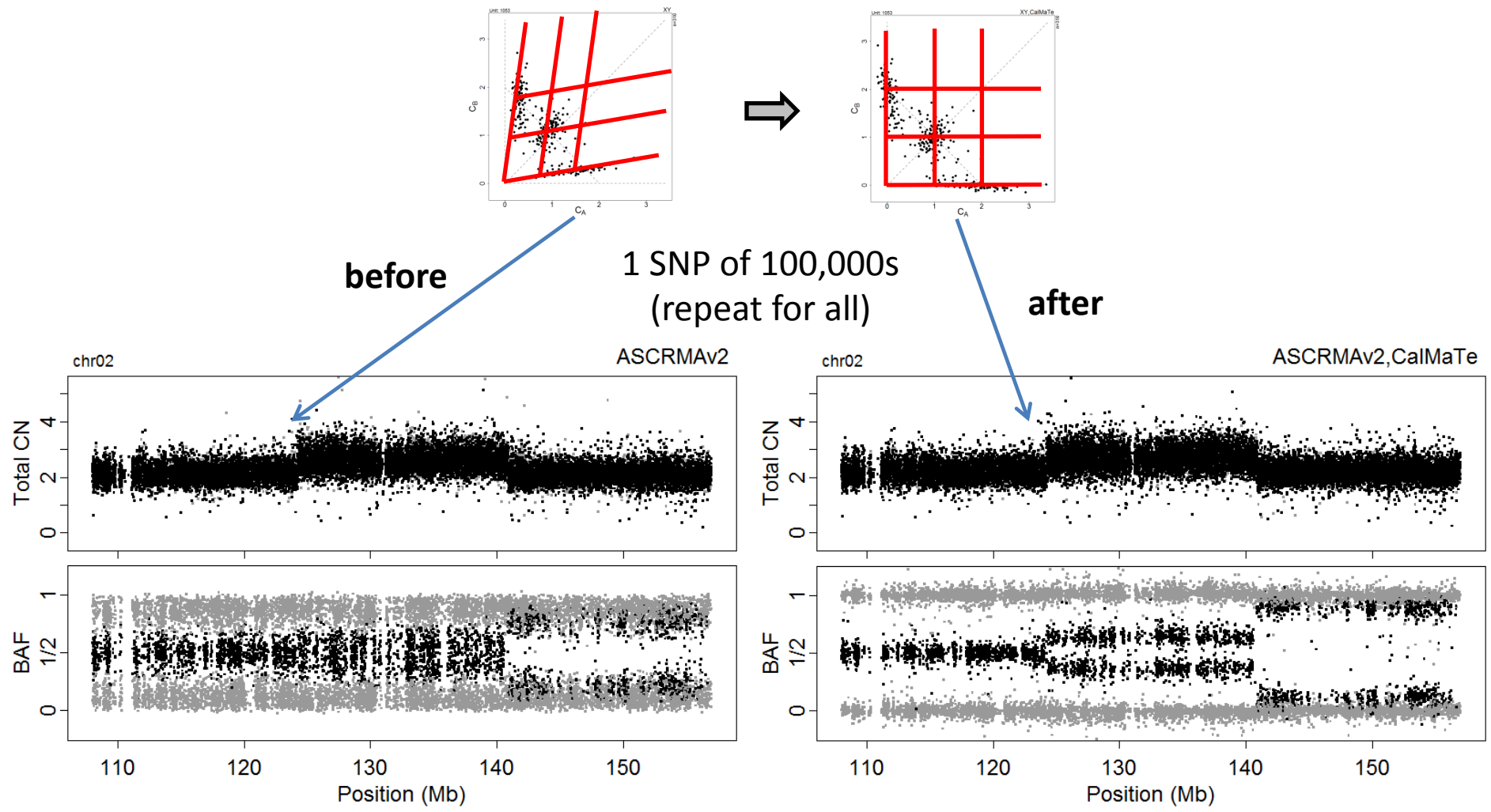


Multi-sample method for each SNP separately:
Non-negative Matrix Factorization (NMF).
Robustified against outliers (e.g. tumors).
Special cases: Only one or two genotype groups.

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

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).

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.

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.

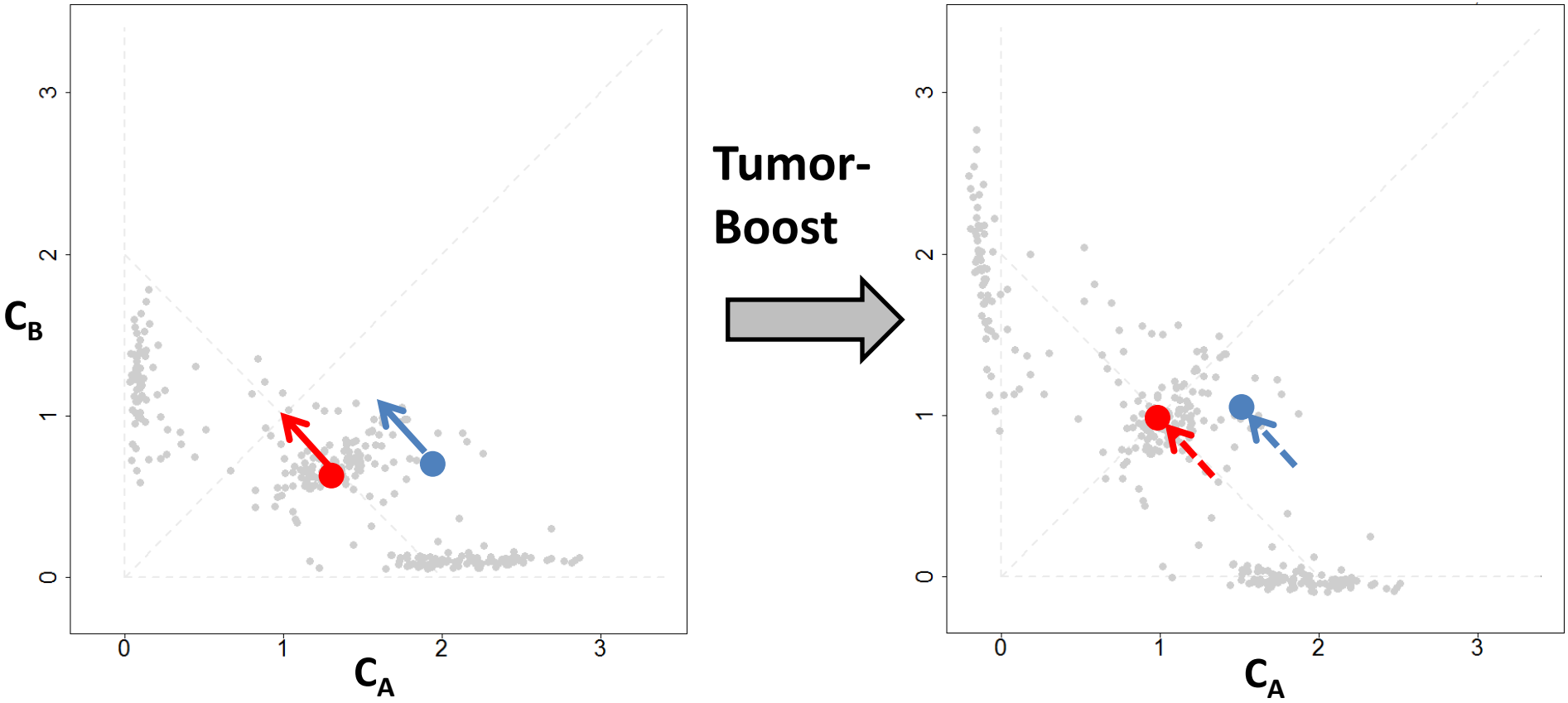
The tumor “should be” close to its normal

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

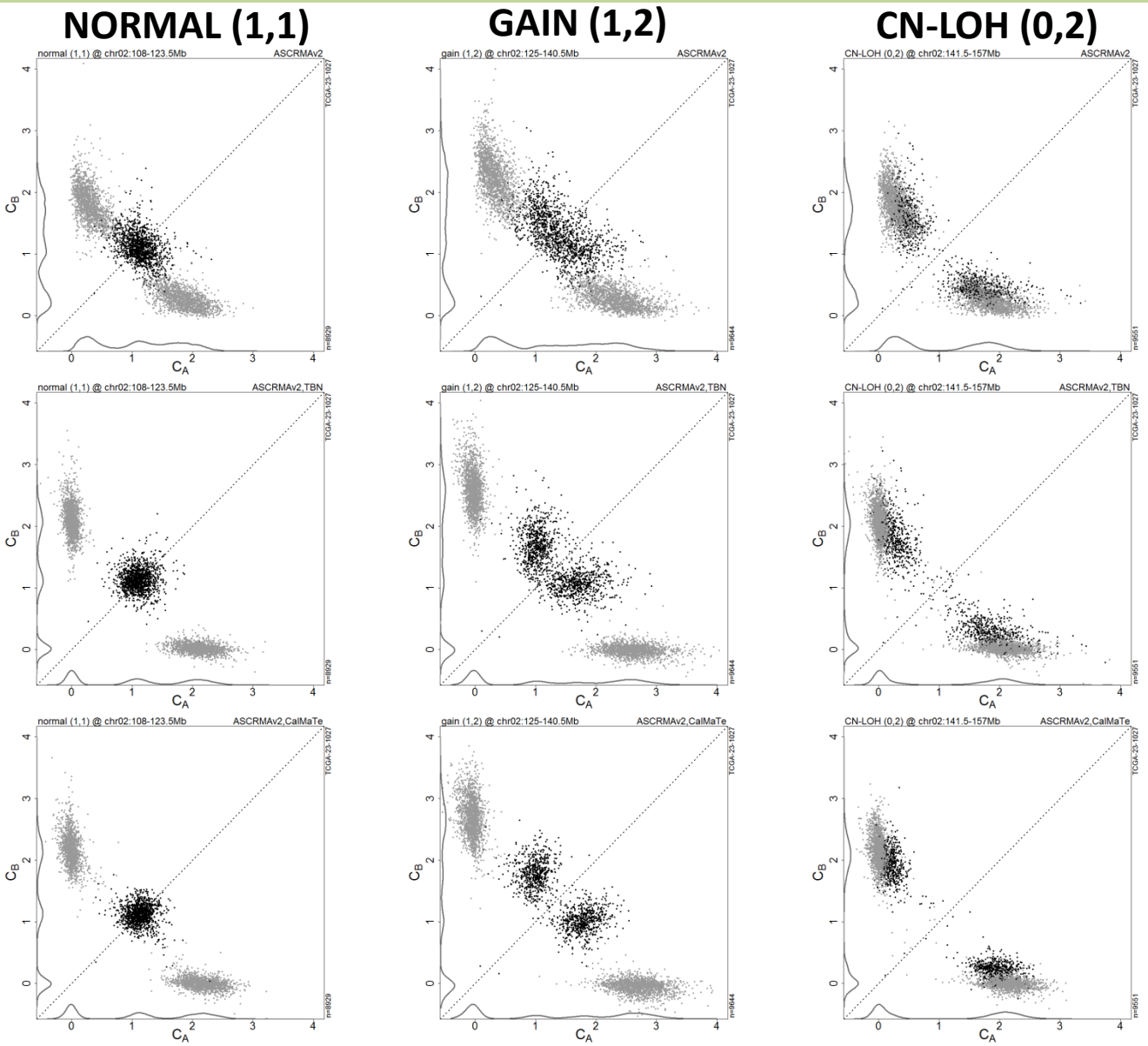


TumorBoost => more distinct (C_A,C_B) - key for PSCN segmentation

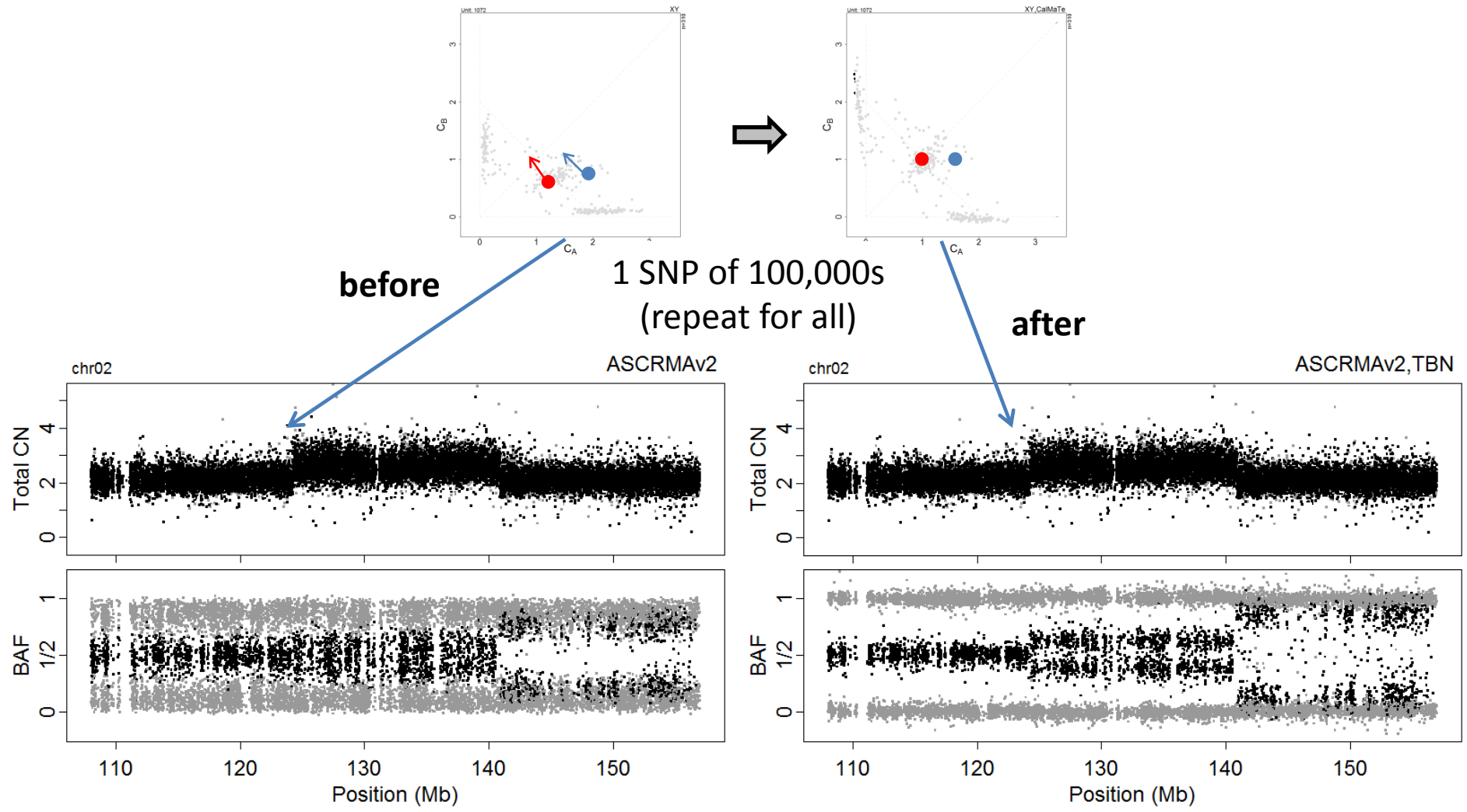
Original:

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

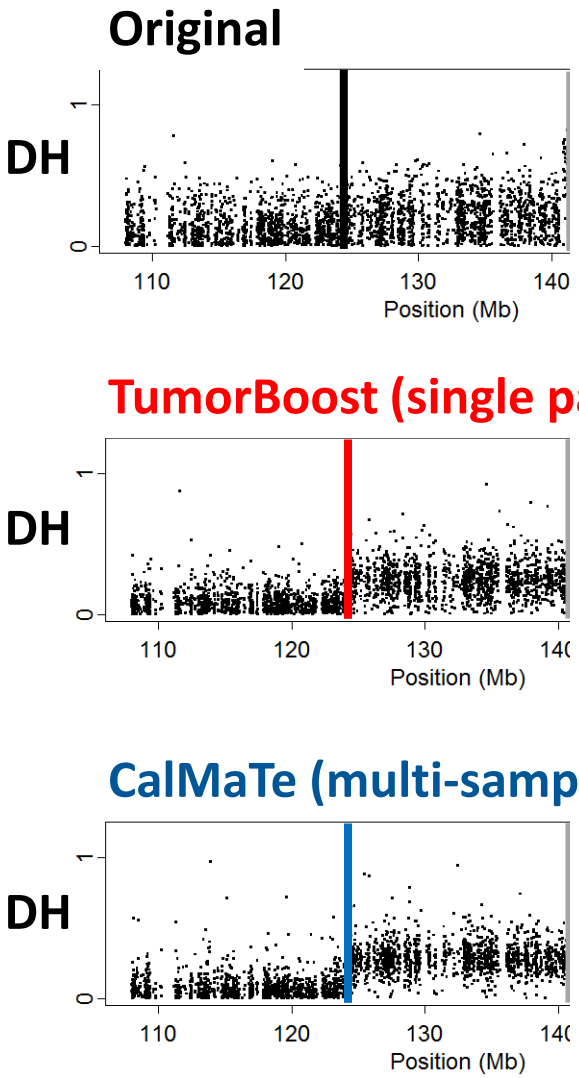
CalMaTe:
- multi-sample

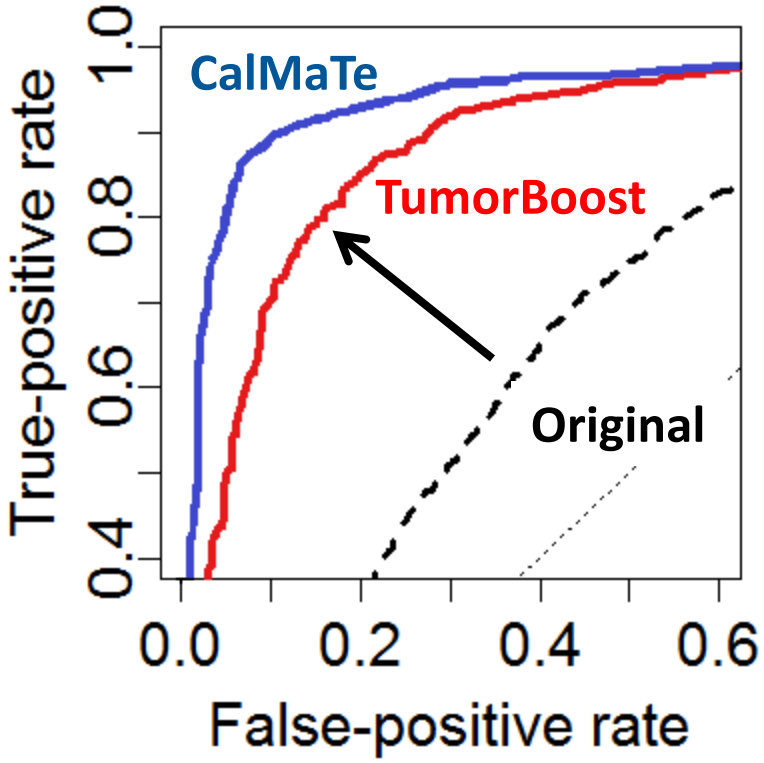


Even with a single tumor-normal pair, we can greatly improve the SNR



TumorBoost significantly improves power to detect change points





One sample, one 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

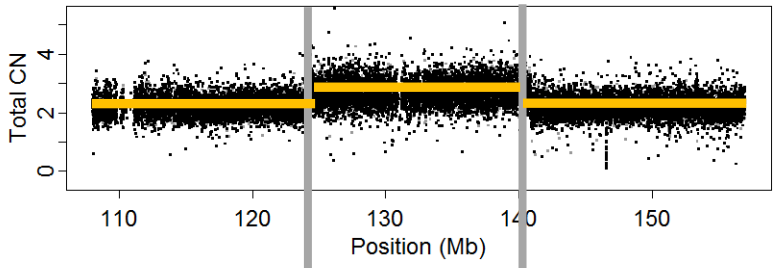
Total CNs & DHs segmentation gives us PSCN regions and estimates

(i) Find TCN change points, then extra DH ones

(ii) Estimate mean levels

Total CNs

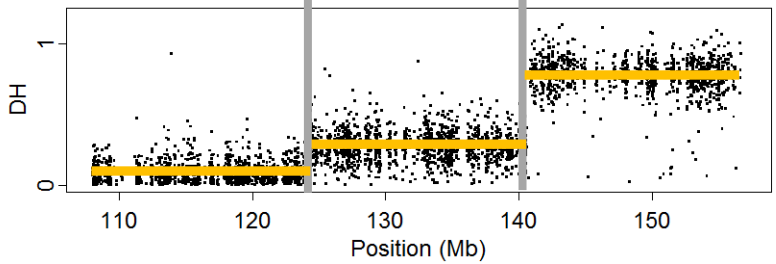
$C = C_A + C_B$



avg(all loci)

Decrease in Heterozygosity

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



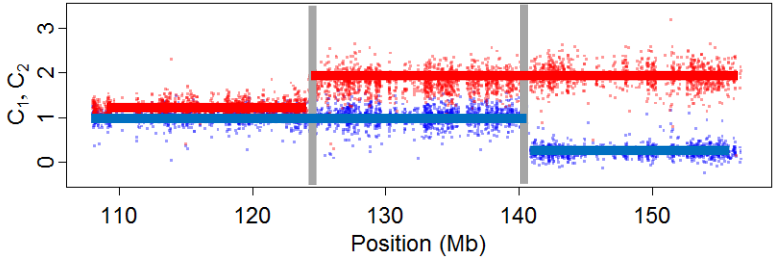
avg(hets only)

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)

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:

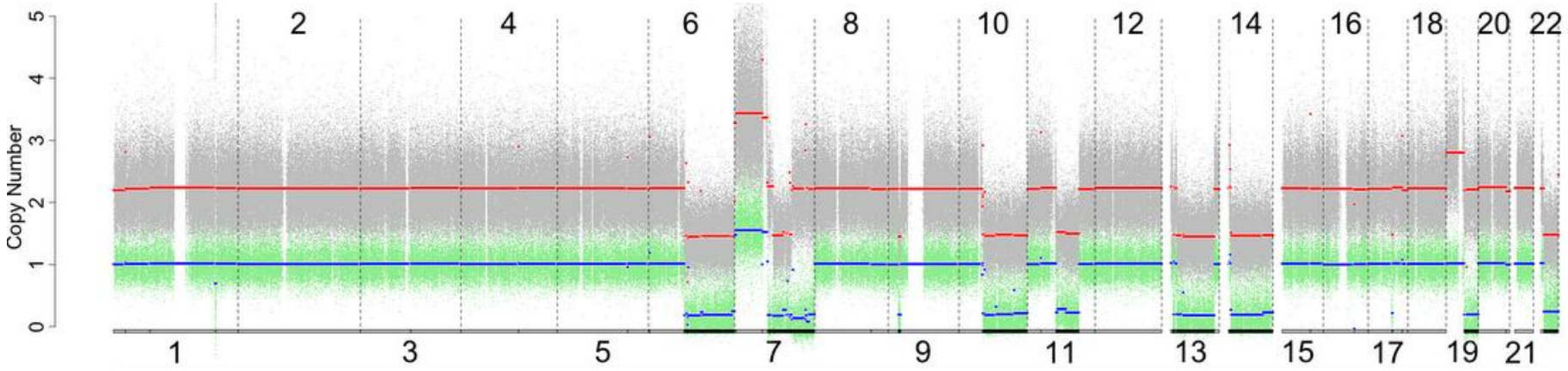
- Null: $C_1 > 0$ (“not in LOH”)
- C_1 is estimated with bias due to background (e.g. normal contamination), so we need offset Δ_{LOH} in test.
- Reject null if (1-α):th percentile of bootstrap-estimated $C_1 - \Delta_{LOH} < 0$.
- How do we choose Δ_{LOH}?

Results

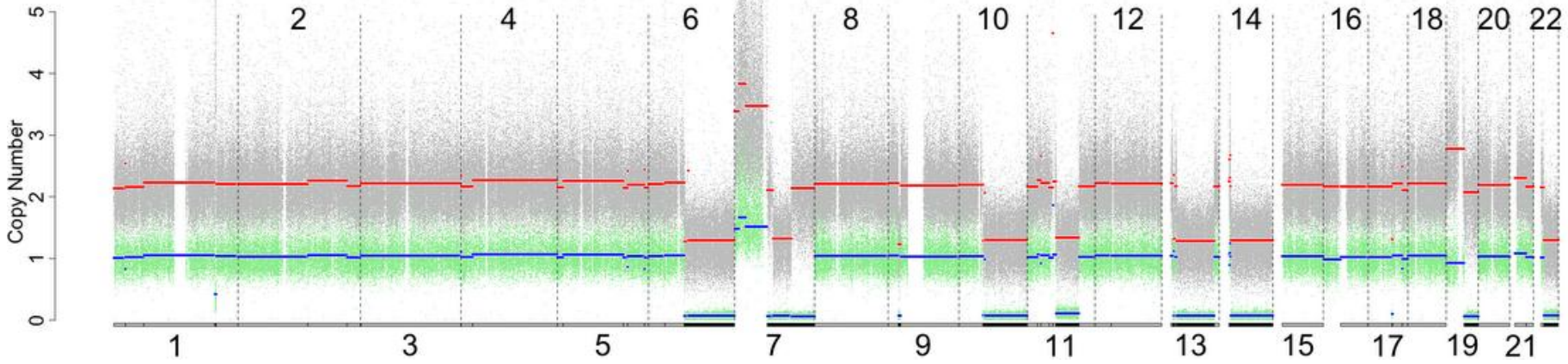
PSCBS works with any SNP array
- similar results on Affymetrix and Illumina

!

Affymetrix GenomeWideSNP_6



Illumina HumanHap550



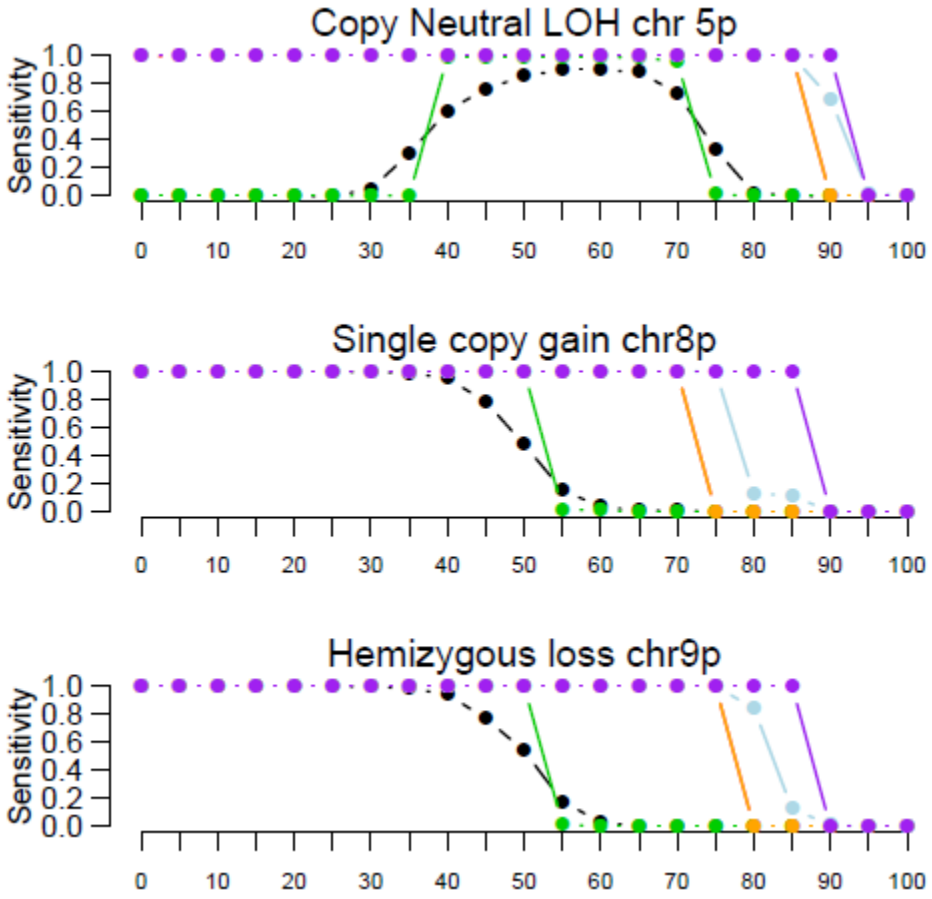
Other methods exists
e.g. Paired BAF segmentation

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.

Conclusions

Paired PSCBS:

- High quality tumor PSCNs
- Single tumor-normal pair
- No external references needed
- Any SNP microarray technology
- Algorithm is fast and bounded in memory
- R package ‘PSCBS’

Future:

- Add PSCBS pipeline to the aroma-project.org
- Non-paired PSCBS
- Calibration of PSCN states (e.g. “purity” & “ploidy”)