Improved copy number analysis

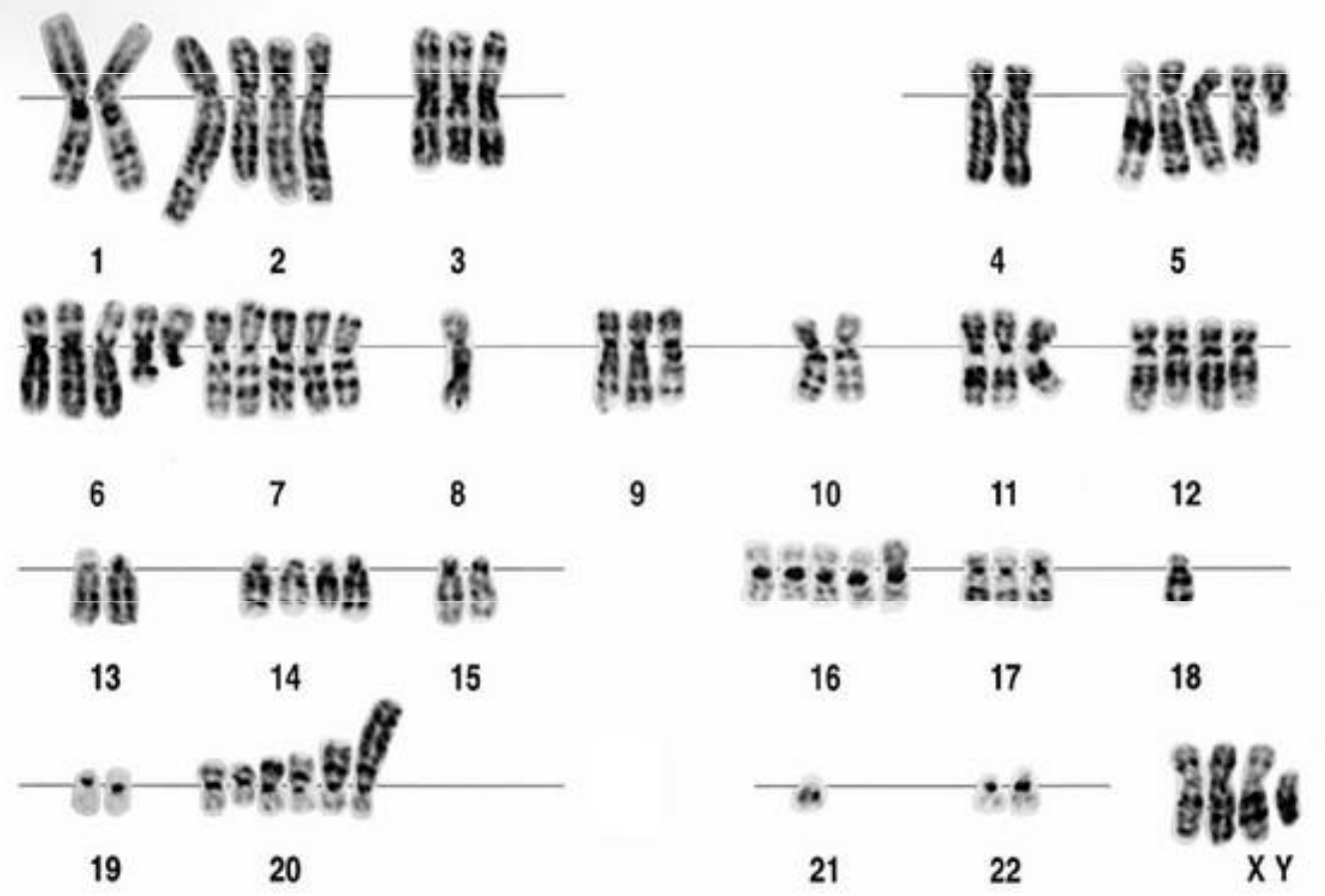
- What is copy number analysis?
- Why is it hard to detect some events?
- Statistical methods that improve detection

Henrik Bengtsson

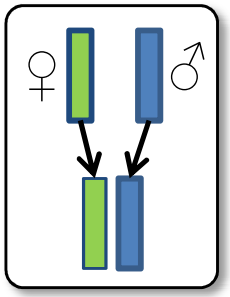
PhD Mathematical Statistics, MSc Computer Science
Dept of Statistics, University of California, Berkeley

UCSF, October 13, 2009

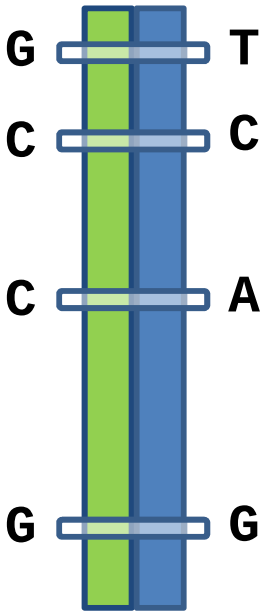
How do tumor cells differ from normal cells?



Genotypes in a diploid chromosome

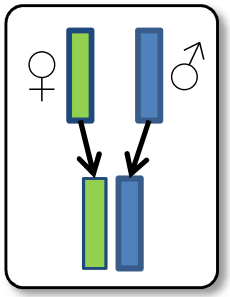


Single nucleotide polymorphism

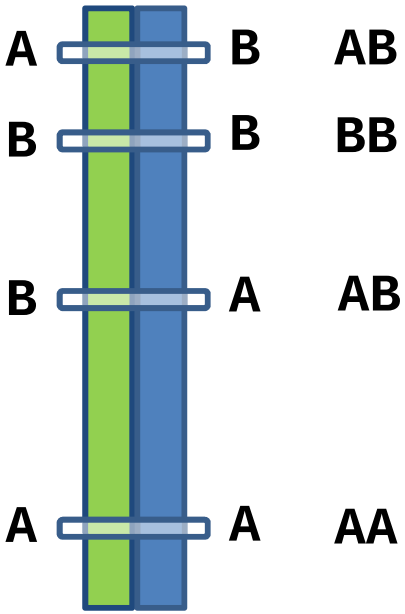


10-20 million
known SNPs

Genotypes in a diploid chromosome

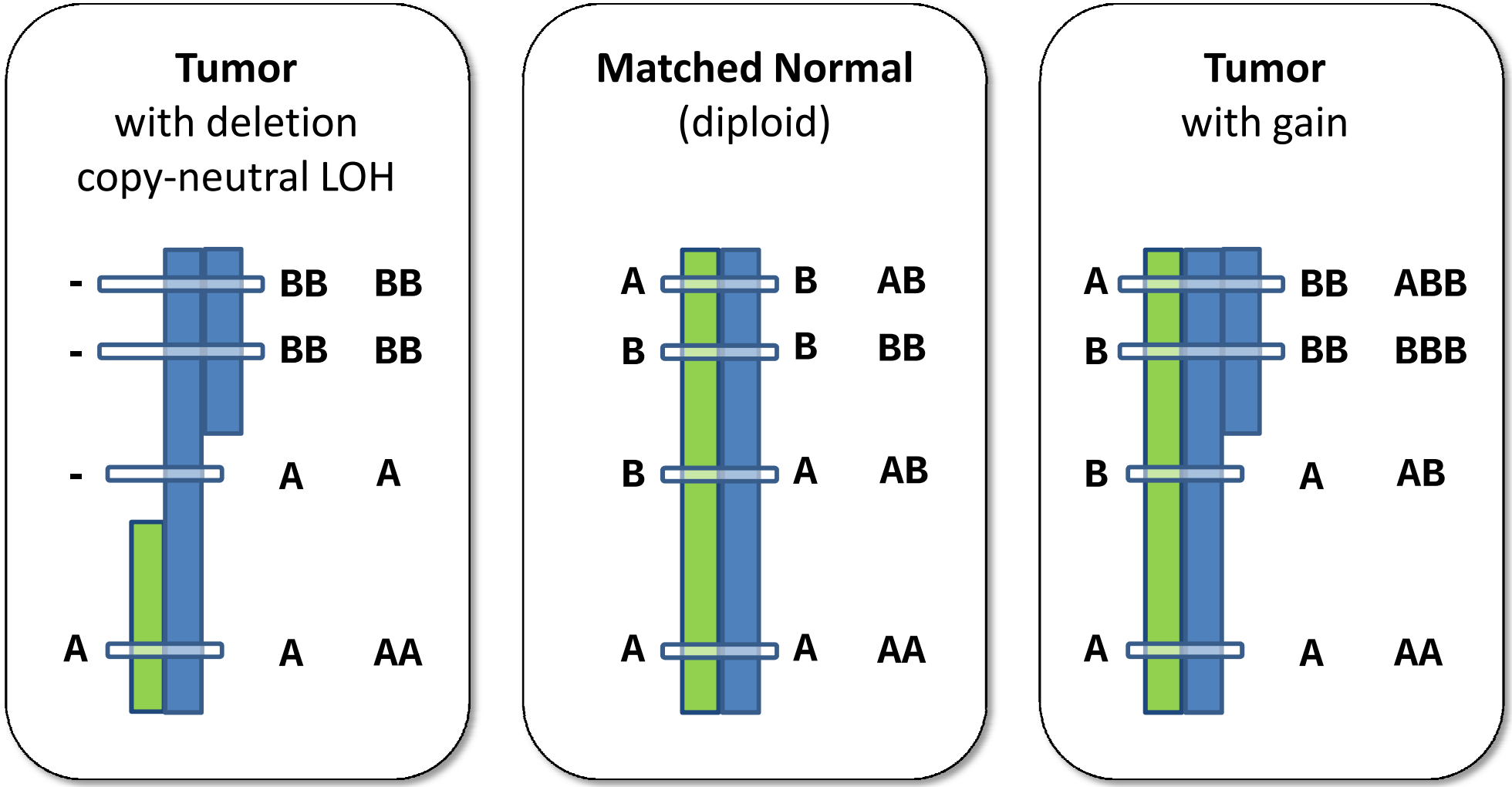


Single nucleotide polymorphism

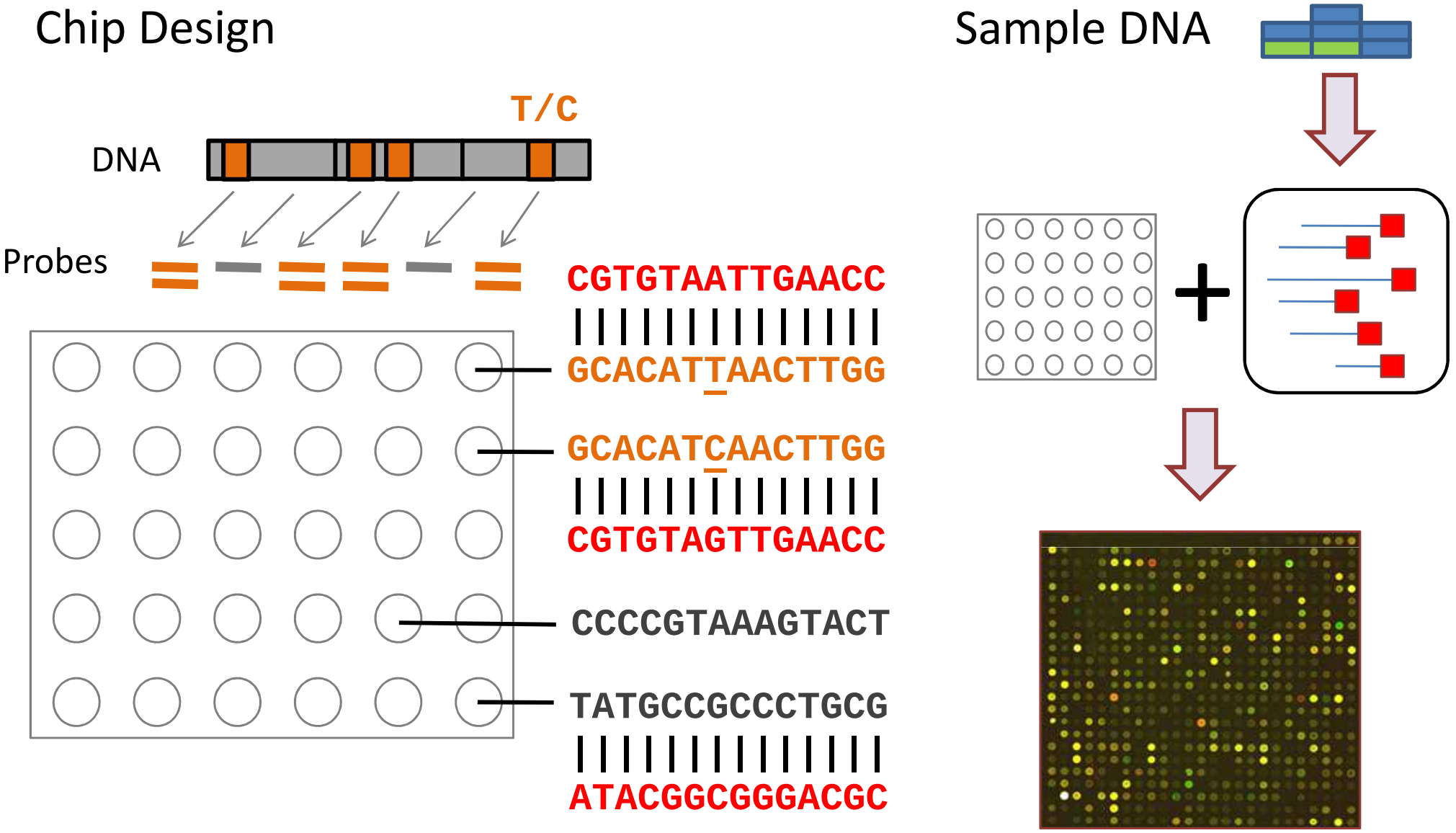


10-20 million
known SNPs

Genotypes and copy numbers in a tumor



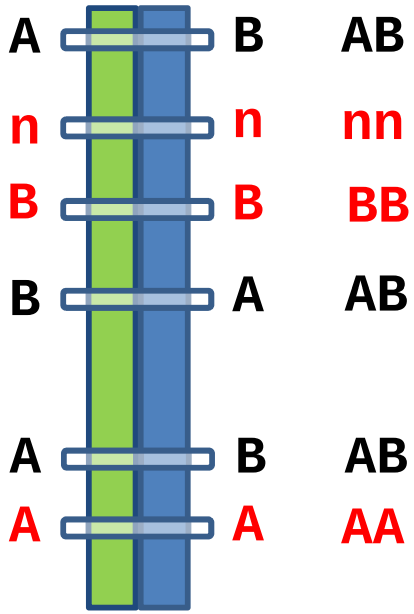
Technology:
Copy number and genotyping microarrays



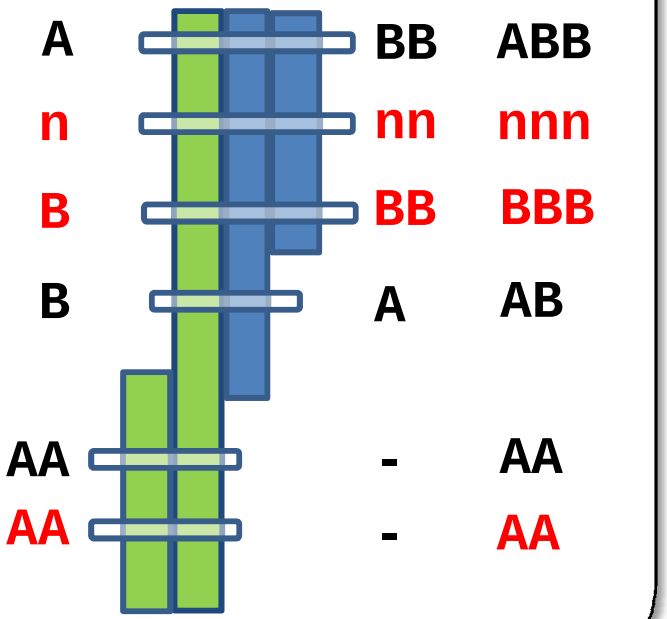
Only heterozygous SNPs (AB in normal) are informative for Loss of Heterozygosity (LOH)

Homozygous SNPs only provides total CNs just like CN probes

Matched Normal
(diploid)



Tumor
with gain



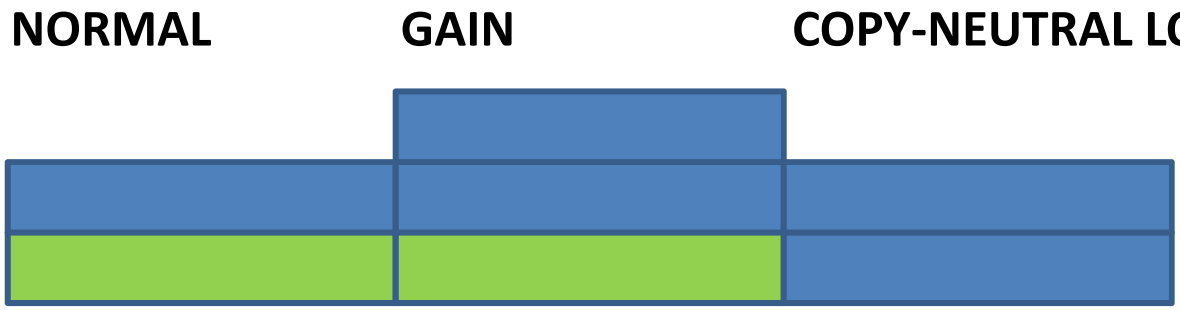
	Informative for Total CN	Informative for "LOH"
A B AB	☑	☑
n n nn	☑	☐
B B BB	☑	☐
B A AB	☑	☑
A B AB	☑	☑
A A AA	☑	☐

Allelic Imbalance =ΔSNP / Total CN
How much a heterozygous SNP has changed

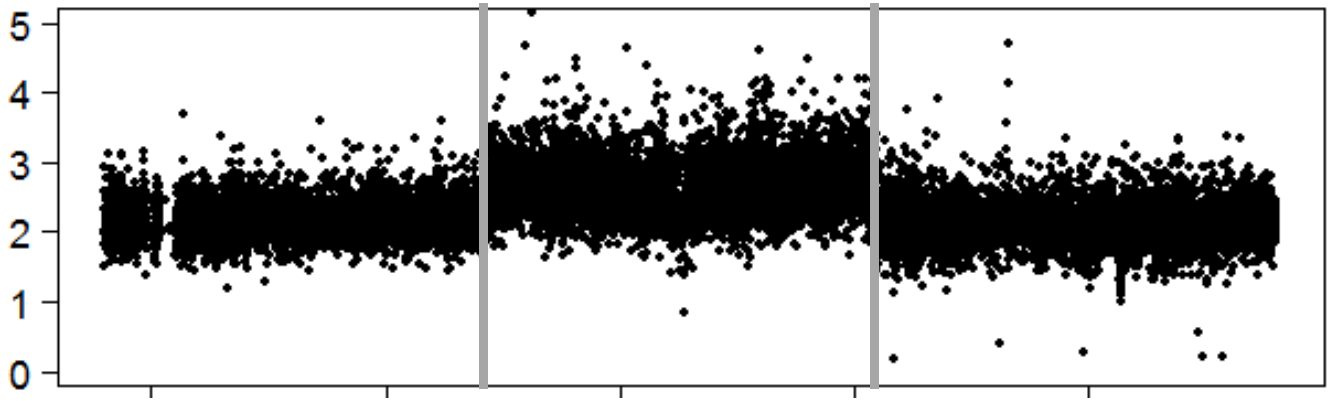
	Normal	Tumor	ΔSNP (A or B)	Total CN	Allelic Imbalance	Examples
Deletion (LOH)	AA **AB** BB	A **A / B** B	**1**	1 **1** 1	- **1/1 = 100%** -	100% = LOH A tumor suppressor gene disappears
Copy-neural LOH	AA **AB** BB	AA **AA / BB** BB	**1+1=2**	2 **2** 2	- **2/2 = 100%** -	
Gain	AA **AB** BB	AAA **AAB / ABB** BBB	**1**	3 **3** 3	- **1/3 = 33%** -	
Large Gain	AA **AB** BB	AAAAAA **AAAAAB / ABBBBB** BBBBBB	**4**	6 **6** 6	- **4/6 = 67%** -	An oncogene gets expressed

All SNPs and CN probes are informative for total copy numbers.

Observed signals in three regions



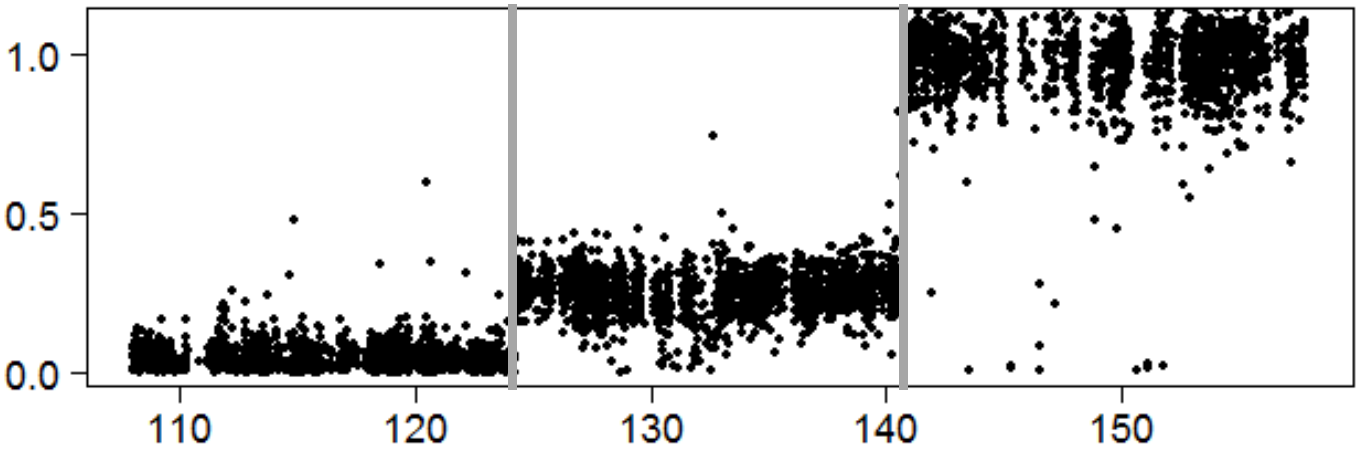
Total copy number



← CN=2

Allelic imbalance

"change from AB"

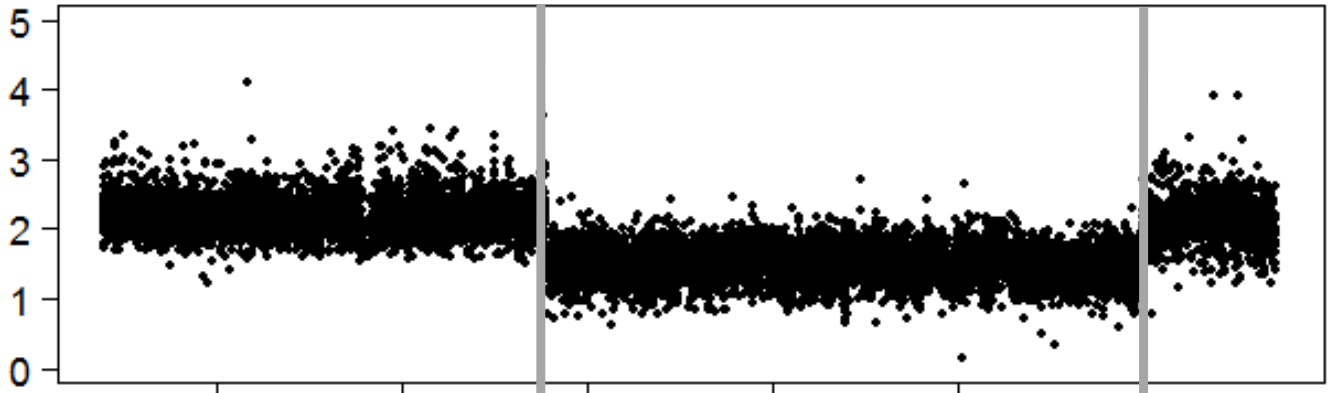


← LOH

no ← change

Observed signals in three other regions

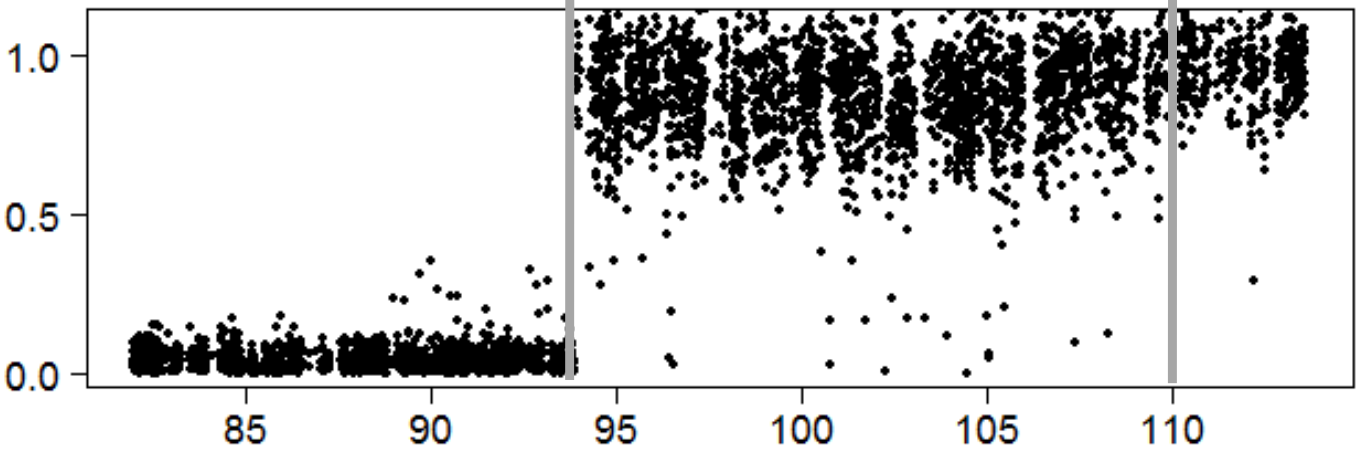
Total copy number



← CN=2

Allelic imbalance

"change from AB"



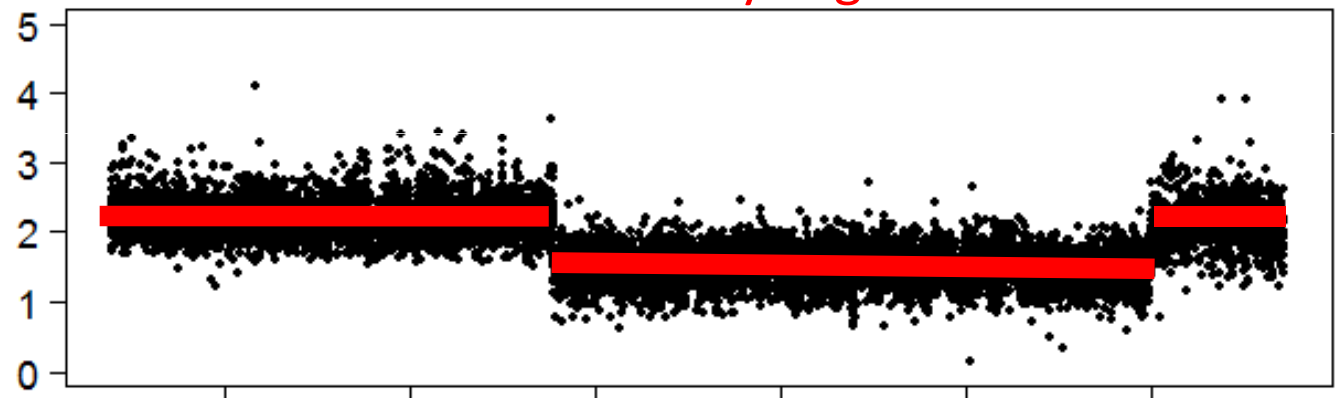
← LOH

no ← change

Segmentation methods identify regions of constant copy numbers and allelic imbalances

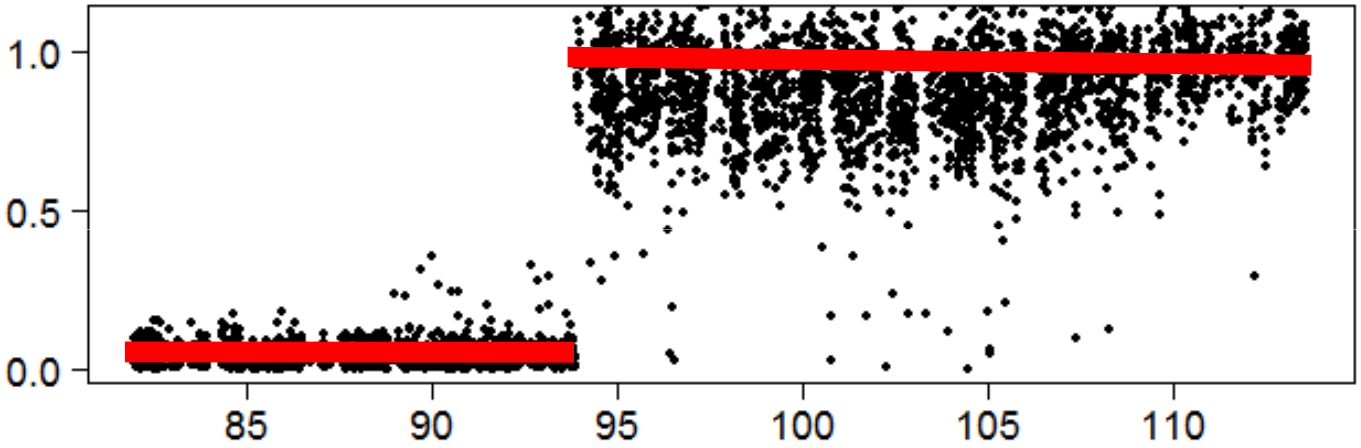
CBS: Circular Binary Segmentation

Total copy number



← **CN=2**

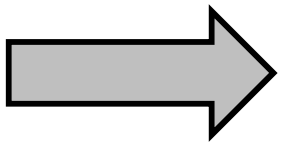
Allelic imbalance



← **LOH**

← **no change**

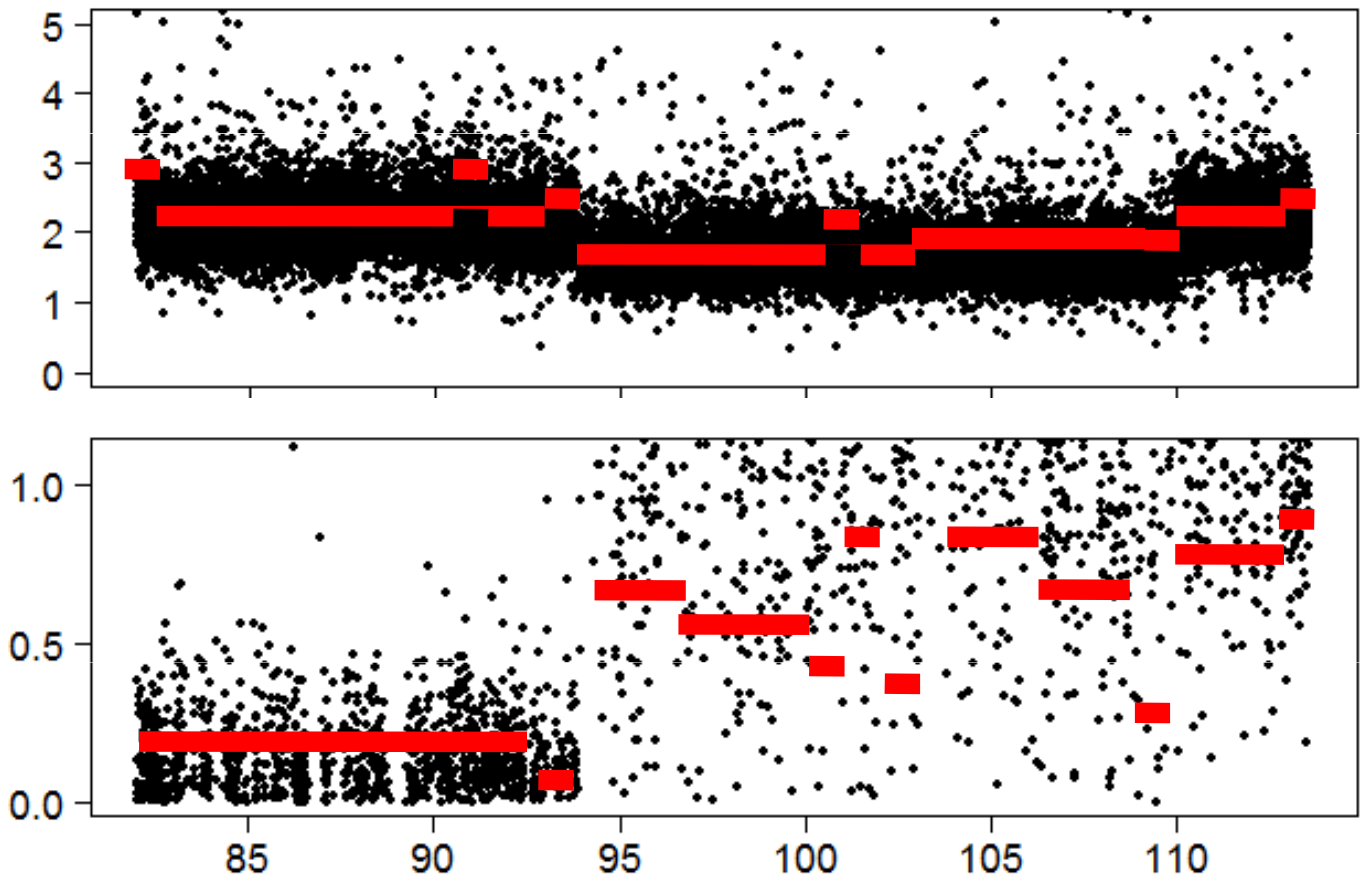
Final call



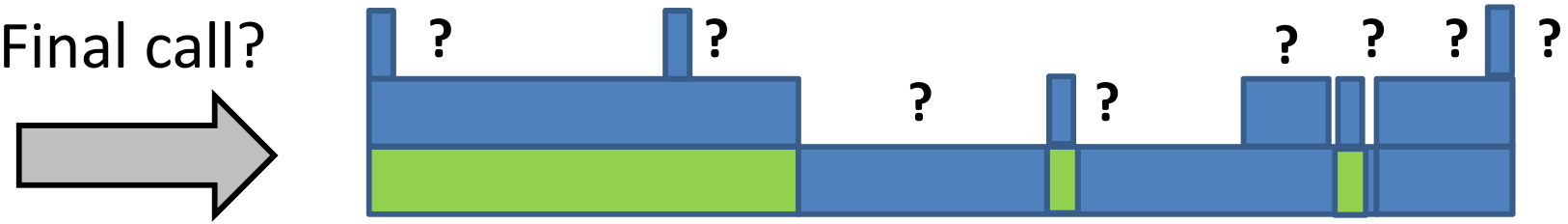
Reality can be much noisier leading to missed regions and falsely discovered regions

Total copy number

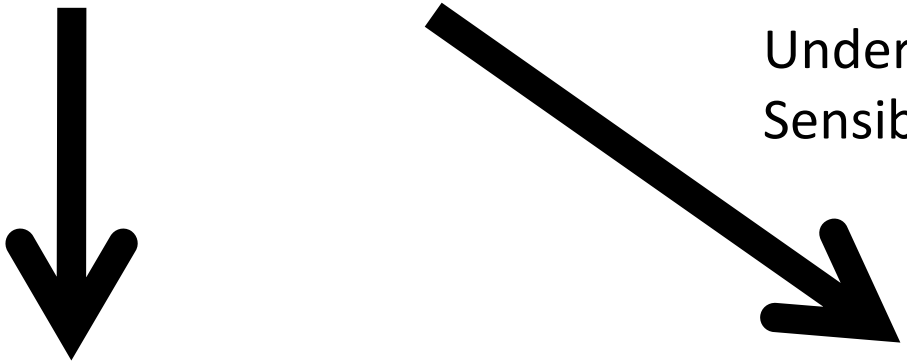
Allelic imbalance



Final call?



Signal to noise ratio



Understanding the technology
Sensible preprocessing

POOR:
- low power
- large false positive rate

GREAT:
- large power
- small false positive rate

Collaborative sharing of anonymous samples from the same facility greatly improves SNRs

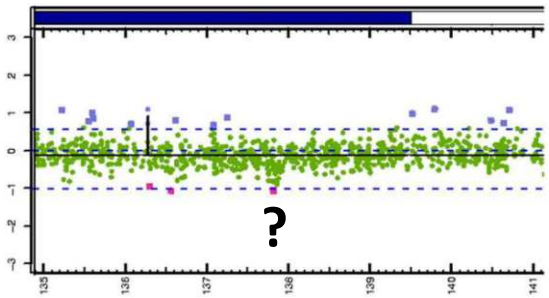
Acute Myeloid Leukemia study:

3 individuals / microarrays

Reference set: (C=2·tumor/pool of ref's)

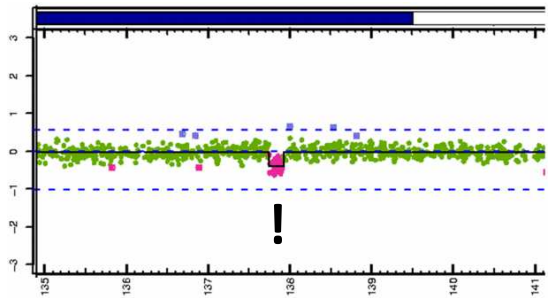
270 HapMap ("gold standard")

σ=0.24



11 anonymous(!) "in-house"

σ=0.12



We found 5x more regions using the much smaller in-house reference set:

Chr	Length	Type	HapMap	In-house
9	1.0	gain		X
20	5.2	loss		X
13	10.8	gain		X
10	26.8	loss		X
5	34.4	loss		X
4	48.0	gain		X
14	22.3	gain	X	X
6	37.0	loss		X
6	37.0	loss		X
3	38.2	loss		X
3	39.1	loss		X
11	21.4	loss		X
14	153.1	gain	X	X
14	153.1	gain	X	X
22	225.1	gain		X
13	297.9	loss		X
8	171.5	loss		X
14	411.5	loss		X
23	2,697.0	loss		X
23	2,697.0	gain	poorly	X
11	32,485.5	loss	X	X
21	37,006.6	gain		X
			5	25

Preprocessing methods that increase the signal-to-noise ratios

CRMA:

A better Affymetrix preprocessing method.

MSCN:

Integrate total CNs from multiple platforms.

TumorBoost:

Improved allelic imbalances of a tumor given a matched normal.

CRMA

Better total copy numbers for all Affymetrix chip types

H. Bengtsson, P. Wirapati, T.P. Speed
A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6.
Bioinformatics, 2009.

H. Bengtsson, R.A. Irizarry, B. Carvalho, T.P. Speed
Estimation and assessment of raw copy numbers at the single locus level. Bioinformatics, 2008.

H. Bengtsson, O. Hössjer
Methodological study of affine transformations of gene expression data with proposed robust non-parametric multi-dimensional normalization method. BMC Bioinformatics, 2006.

H. Bengtsson, G. Jönsson, J. Vallon-Christersson
Calibration and assessment of channel-specific biases in microarray data with extended dynamical range. BMC Bioinformatics, 2004.

Software: aroma.affymetrix, aroma.light

Copy-numbers by Robust Microarray Analysis (CRMA) - a single-array preprocessing method

For each Affymetrix array (*i = 1,2,3,…,10000*) independently:

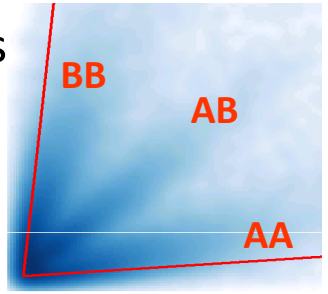
Calibrating & normalizing for hybridization artifacts	1. Offset and Allelic crosstalk calibration 2. Probe-sequence normalization
Summarization of technical replicates	1. CN loci have one probe 2. Robust averaging of replicated SNPs probes 
Normalizing for assay artifacts	1. PCR fragment-length normalization 2. GC-content normalization
Total and Allele-specific copy numbers	(C_A, C_B), $C = C_A + C_B$

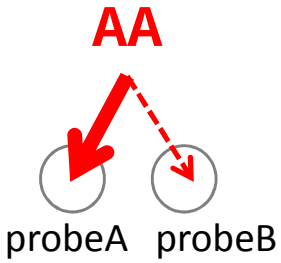
Systematic variation across arrays is due to scanner offset and allelic cross-hybridization

1. The scanner's shifts all probe signals (offset)

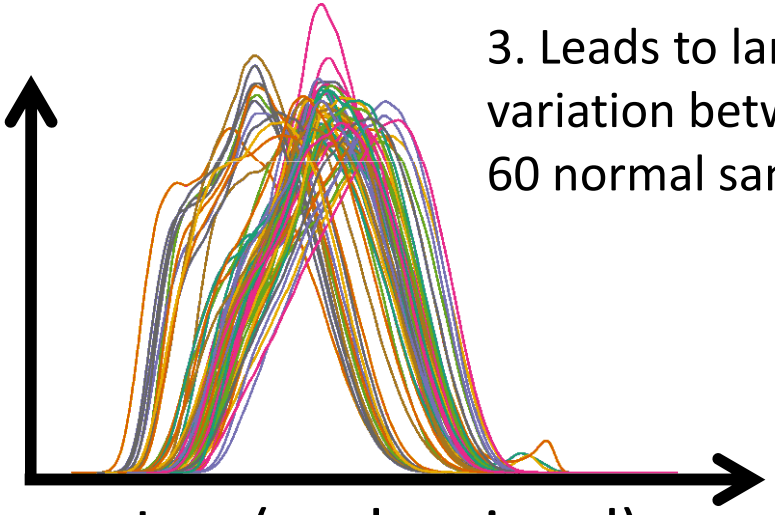
+

2. Cross-hybridization causes signal to leak between allele A and allele B



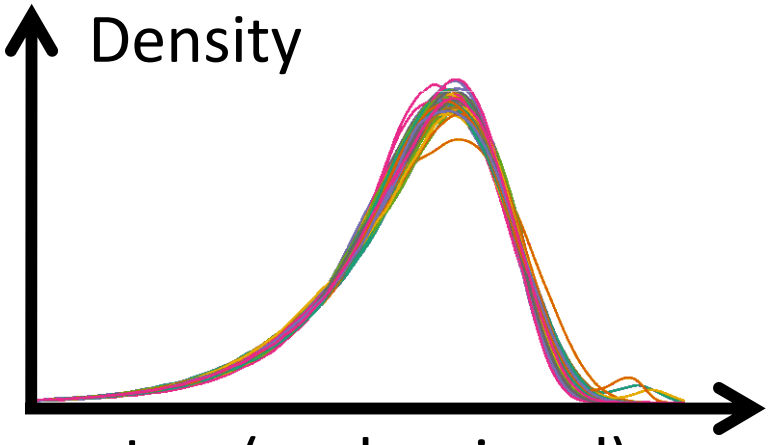


=



3. Leads to large variation between 60 normal samples

4. Calibration for both removes a majority of artifacts between samples



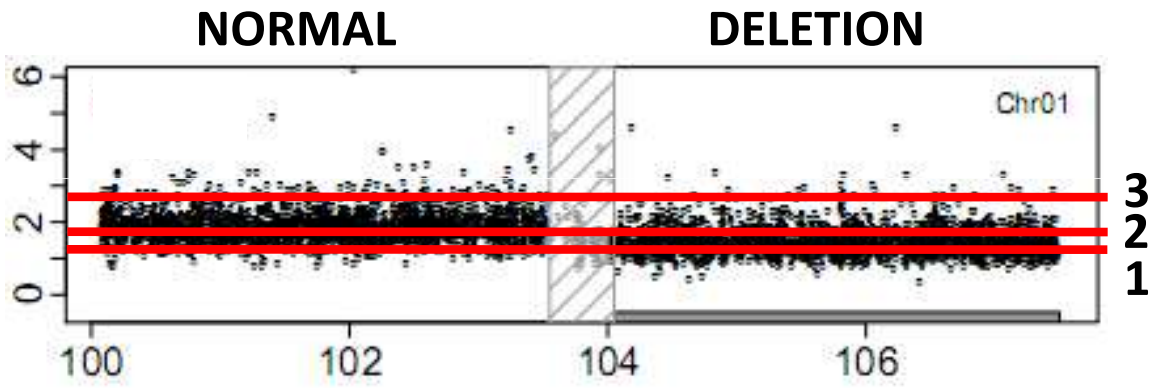
Comparing CRMA with other methods using Receiver Operator Characteristics (ROC)

Data:

(1) Pick a random sample.

(2) Find a clear CN change point.

(3) Two CN states:
- NORMAL.
- DELETION – the one to call.



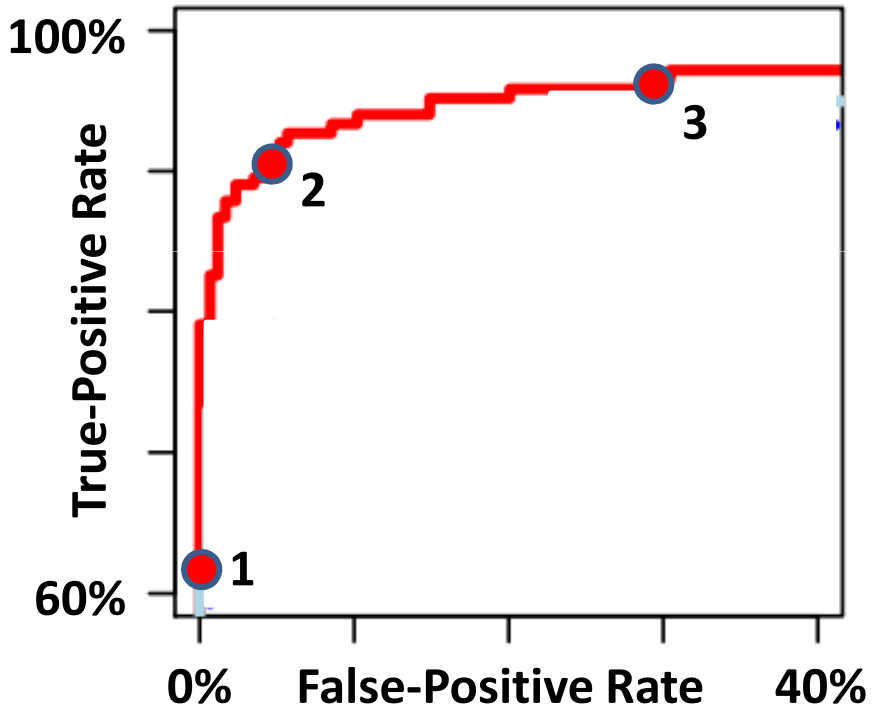
ROC assessment:

(1) Use a **threshold**.

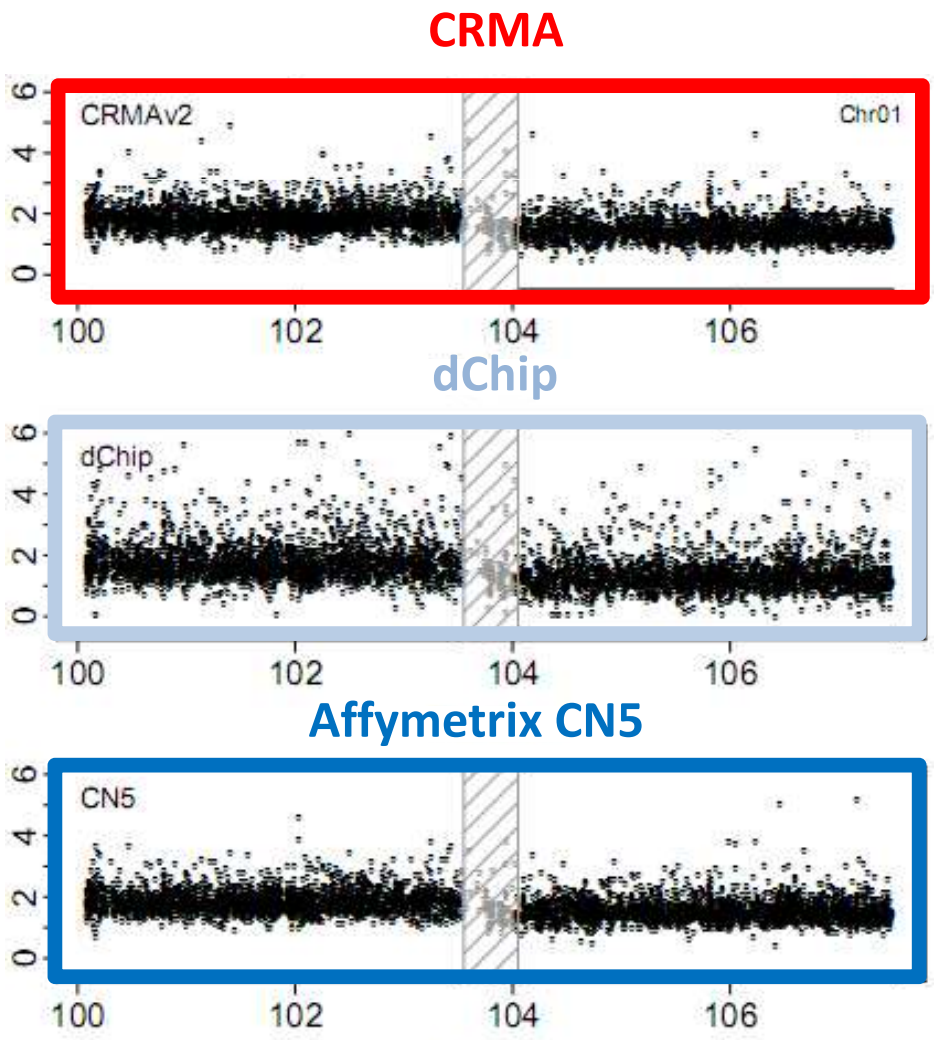
(2) Call SNPs below a DELETION.

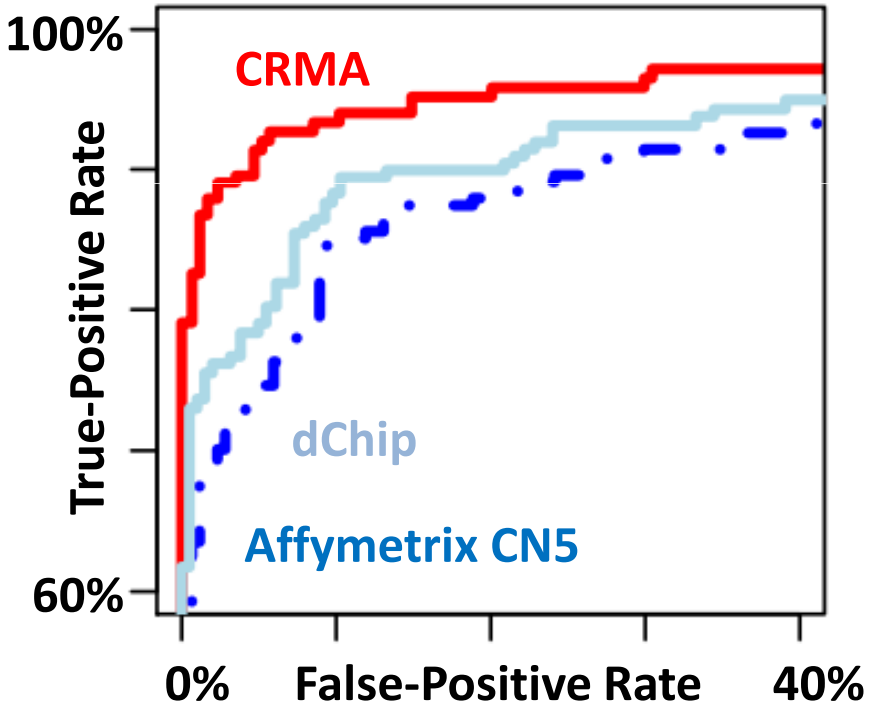
(3) Count number of true and false DELETIONs.

(4) Adjust threshold up and down.



Result: Single-sample CRMA outperforms existing multi-array methods





Data set:

- Tumor-normal pairs (HCC1143).
- 68 hybridizations, Affymetrix 6.0

Preprocessing:

- **CRMA v2** only two arrays.
- **Affymetrix CN5** and **dChip** used all 68 arrays.

MSCN

Combining copy numbers from multiple platforms and labs

H. Bengtsson, A. Ray, P. Spellman and T.P. Speed

A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. Bioinformatics, 2009.

Software: aroma.cn, aroma.tcga

TCGA - The Cancer Genome Atlas project

"Accelerate our understanding
of the molecular basis of cancer"

Multi-center project:

LBL, Broad, Washington University, MSKCC, Harvard, Stanford, UCSF, MD Anderson, …

Tumor types:

GBM, ovarian, lung cancer, …

Large studies:

500 tumor-normal pairs for each tumor type

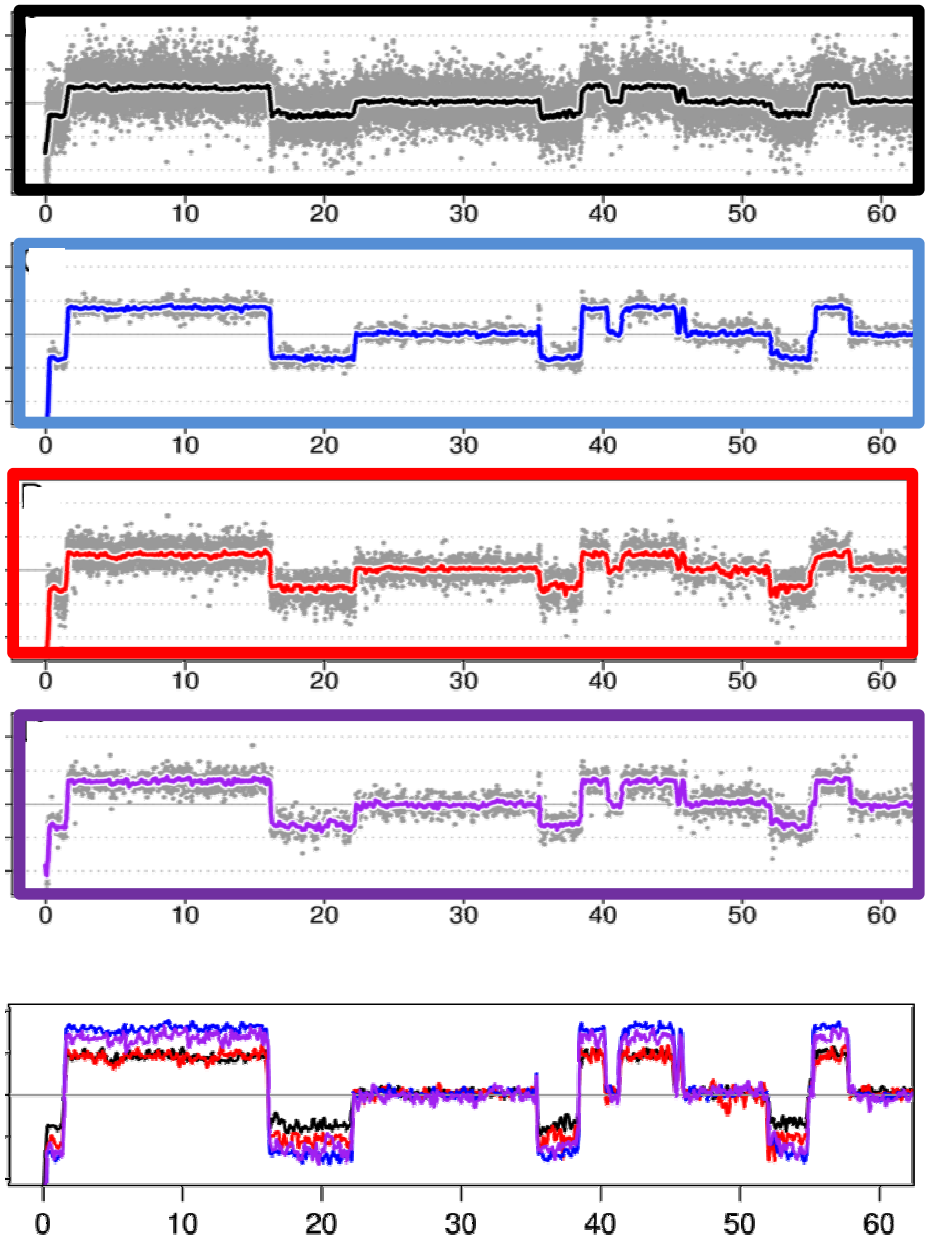
Four centers/platforms produce copy numbers How to merge?

A. Broad
Affymetrix
GenomeWideSNP_6
(n=1.8 · 10^6)

B. MSKCC
Agilent
HG-CGH-244A
(n=0.25 · 10^6)

C. Stanford
Illumina
HumanHap550
(n=0.55 · 10^6)

D. Harvard
Agilent
HG-CGH-244A
(n=0.25 · 10^6)



Problem:
Non-linear relationship between platforms

True CN:
x (unknown)

Smooth CN for platform s=1,2,3,4:
$y^{(s)} = f^{(s)}(x) + \text{noise}$

Smoothed pair (s,t):
$(y^{(s)}, y^{(t)}) = (f^{(s)}(x), f^{(t)}(x)) + \text{noise}$

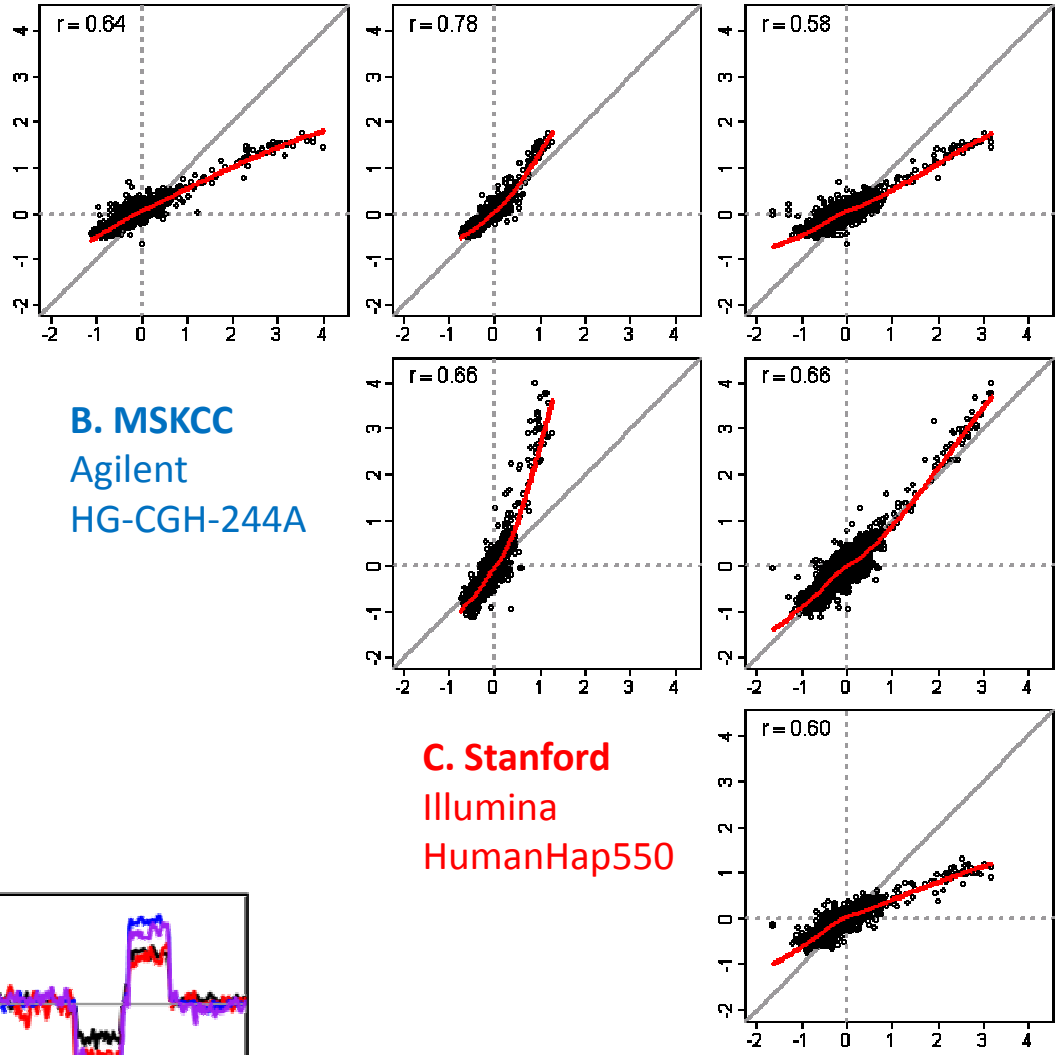
A. Broad
Affymetrix
GenomeWideSNP_6

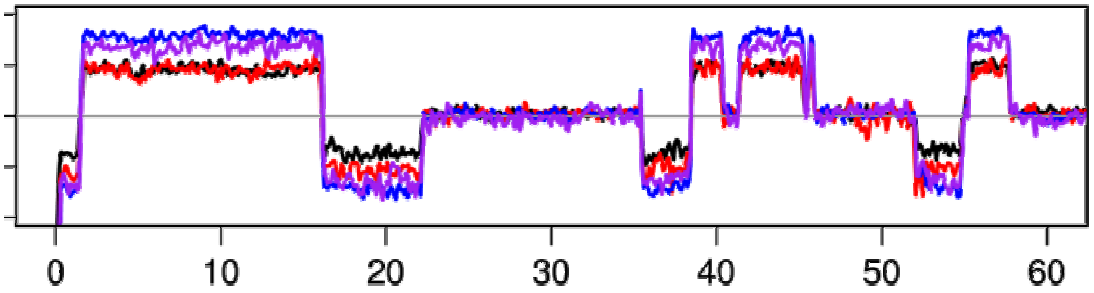
B. MSKCC
Agilent
HG-CGH-244A

C. Stanford
Illumina
HumanHap550

D. Harvard
Agilent
HG-CGH-244A

all chromosomes (n=26,640)





Principal Curves

Multiplatform data in R^4 (4 platforms):

True CN:

x (an unknown scalar)

Smoothed CNs:

$\underline{y} = (y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)})^T$

Unknown transformation:

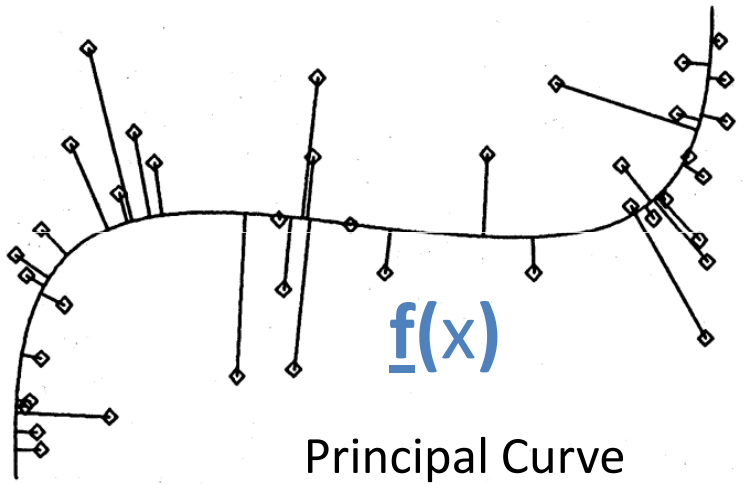
$\underline{f}(x) = (f^{(1)}(x), f^{(2)}(x), f^{(3)}(x), f^{(4)}(x))^T$

Noise:

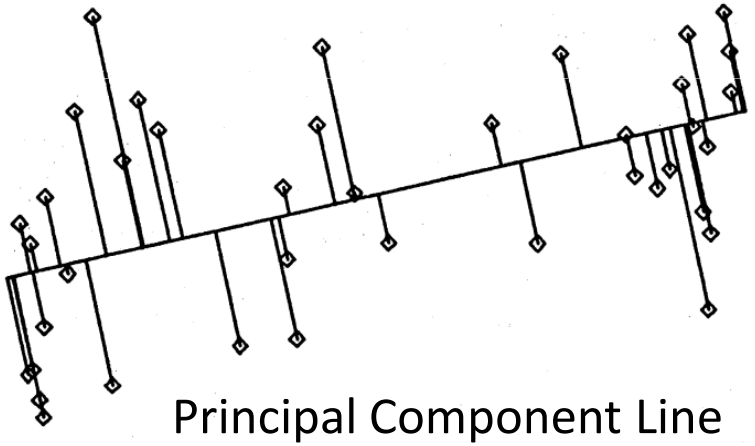
$\underline{\varepsilon} = (\varepsilon^{(1)}, \varepsilon^{(2)}, \varepsilon^{(3)}, \varepsilon^{(4)})^T$

Vector model:

$\underline{y} = \underline{f}(x) + \underline{\varepsilon}$



Principal Curve



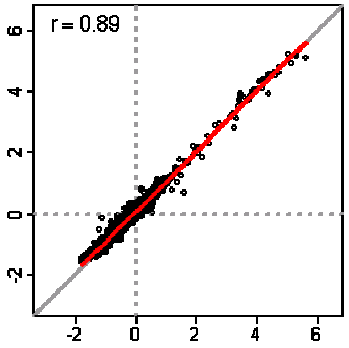
Principal Component Line

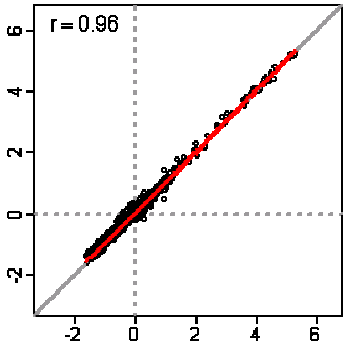
Hastie et al., Principal Curves, JASA, 1989

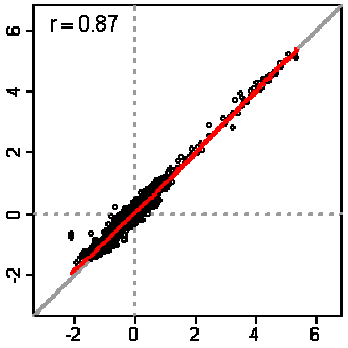
Result:
Linear relationship after back-transformation

all chromosomes (n=26,640)

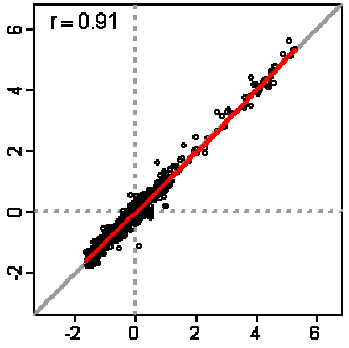
A. Broad
Affymetrix
GenomeWideSNP_6

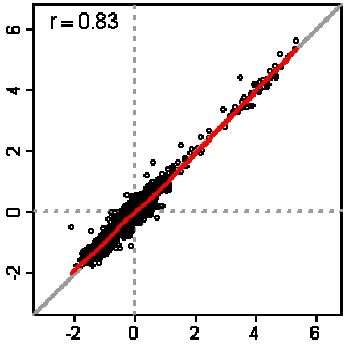




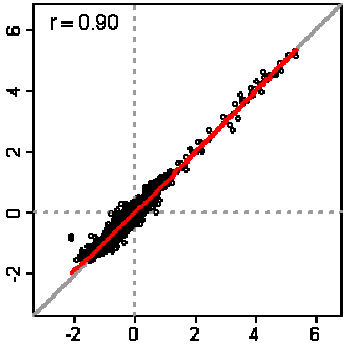


B. MSKCC
Agilent
HG-CGH-244A

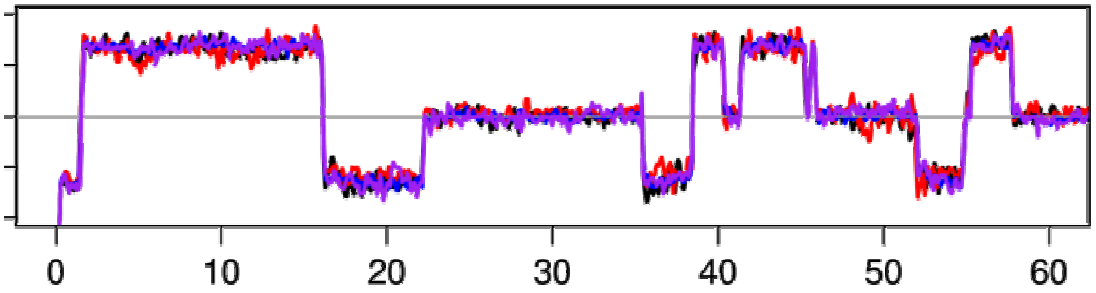




C. Stanford
Illumina
HumanHap550



D. Harvard
Agilent
HG-CGH-244A



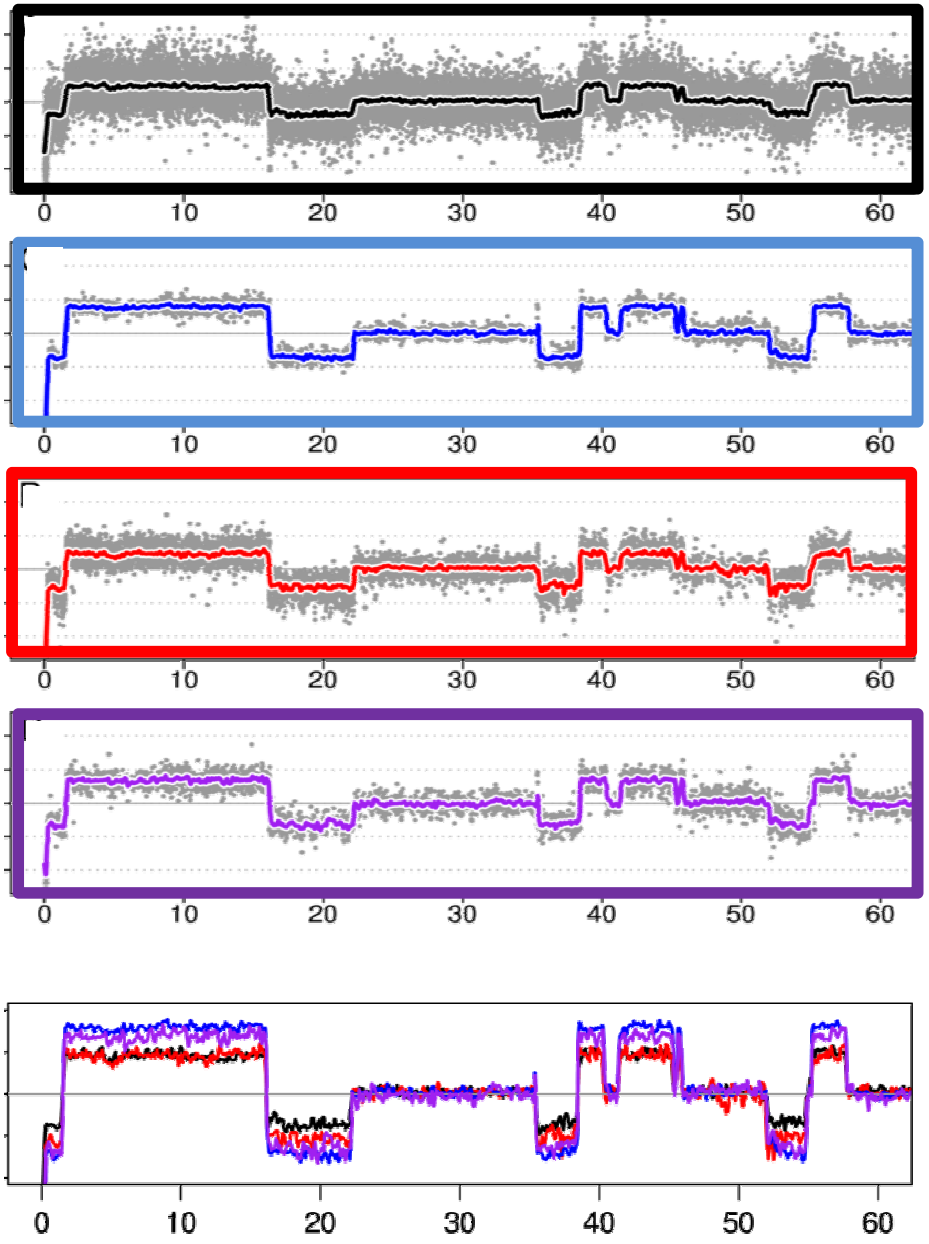
Apply the back-transformation on probe data

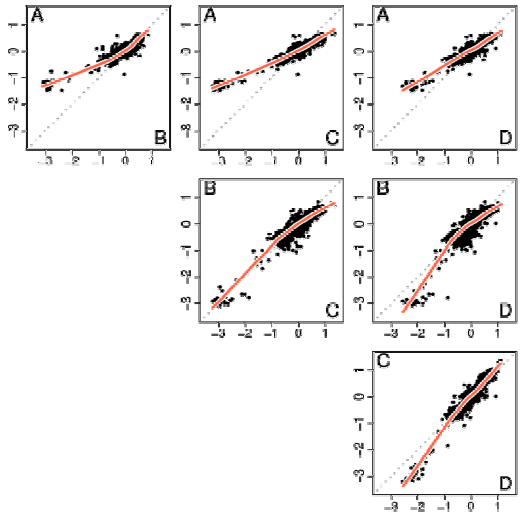
A. Broad
Affymetrix
GenomeWideSNP_6
(n=1.8 · 10^6)

B. MSKCC
Agilent
HG-CGH-244A
(n=0.25 · 10^6)

C. Stanford
Illumina
HumanHap550
(n=0.55 · 10^6)

D. Harvard
Agilent
HG-CGH-244A
(n=0.25 · 10^6)





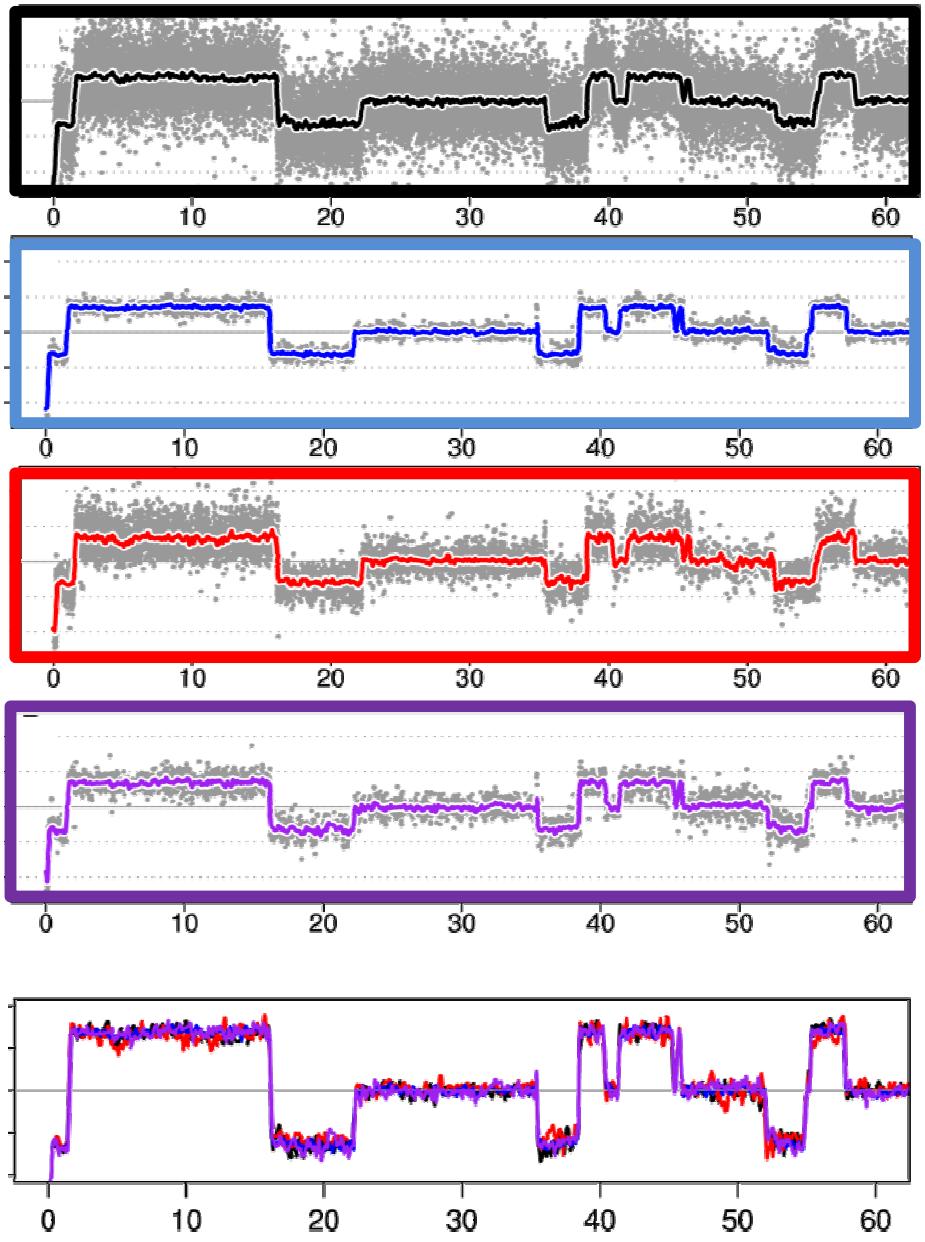
Result:
Platforms agree on the copy numbers

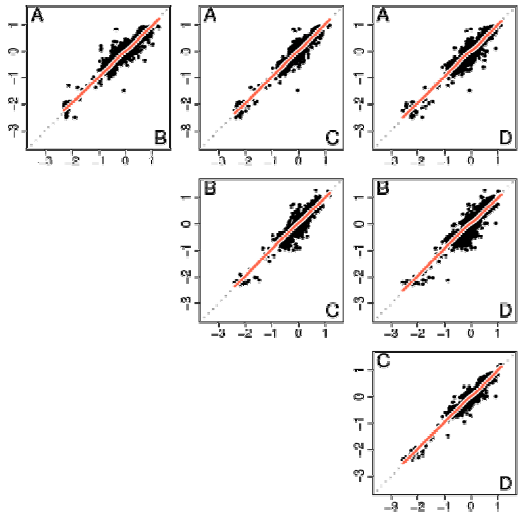
A. Broad
Affymetrix
GenomeWideSNP_6
($n=1.8 \cdot 10^6$)

B. MSKCC
Agilent
HG-CGH-244A
($n=0.25 \cdot 10^6$)

C. Stanford
Illumina
HumanHap550
($n=0.55 \cdot 10^6$)

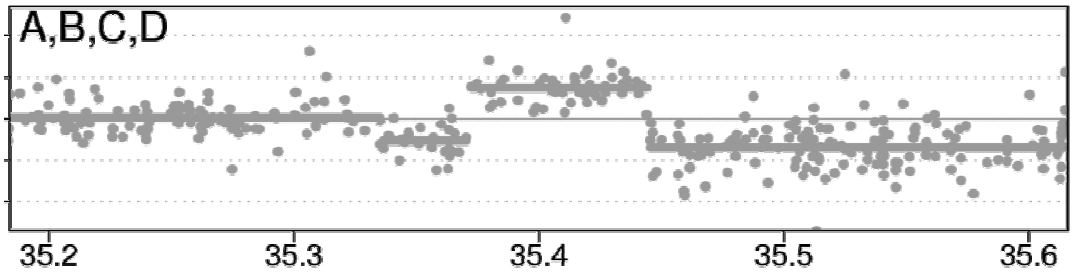
D. Harvard
Agilent
HG-CGH-244A
($n=0.25 \cdot 10^6$)



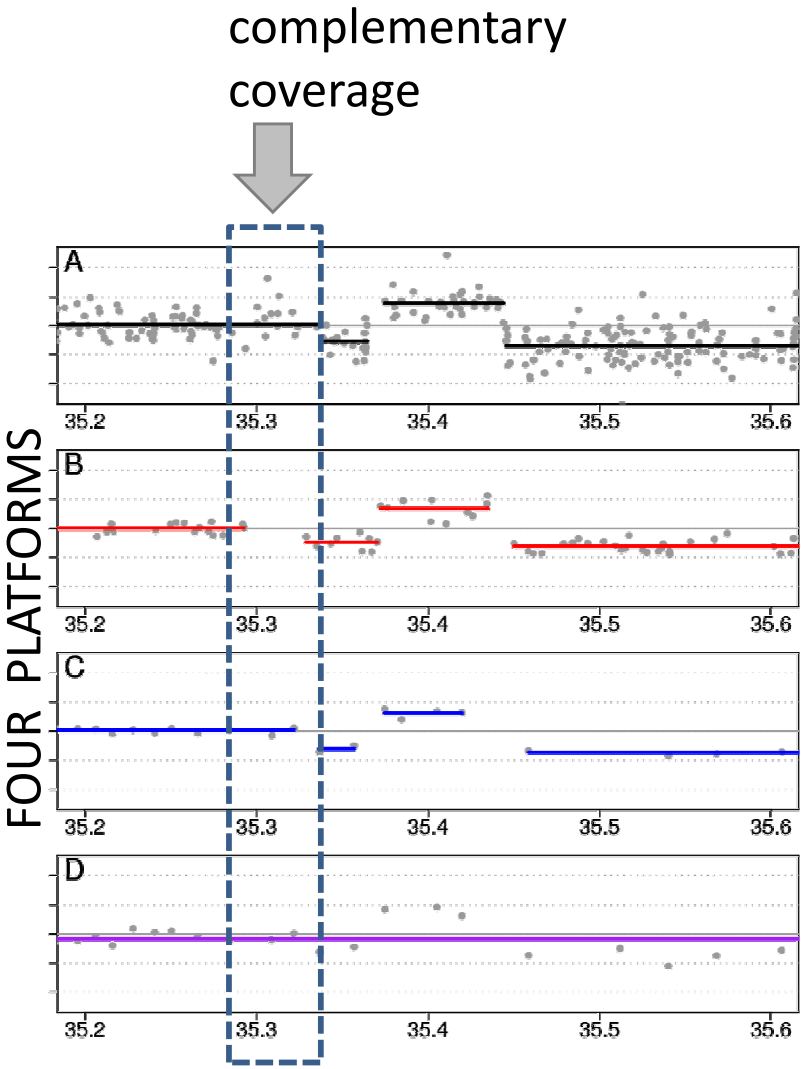


Combining standardized copy numbers - enhanced detection of events

Combining normalized data:



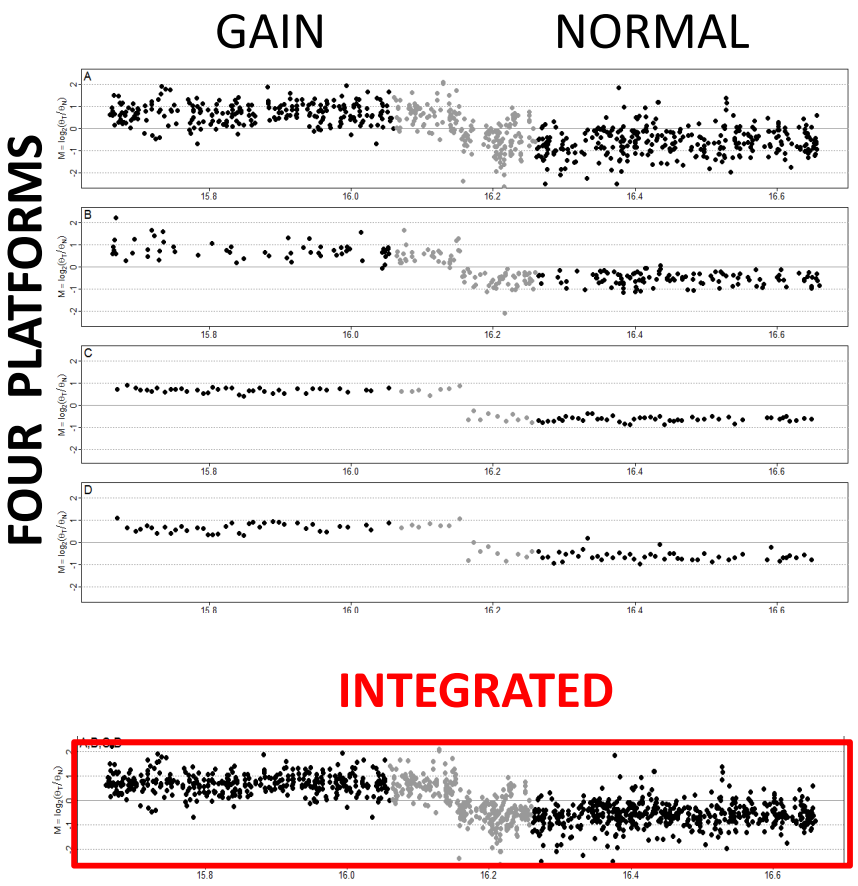
1. Greater power to detect CN changes
2. More precise locations.
3. Greater resolution.
4. Greater and complementary coverage.



Result:
More true regions and fewer false regions

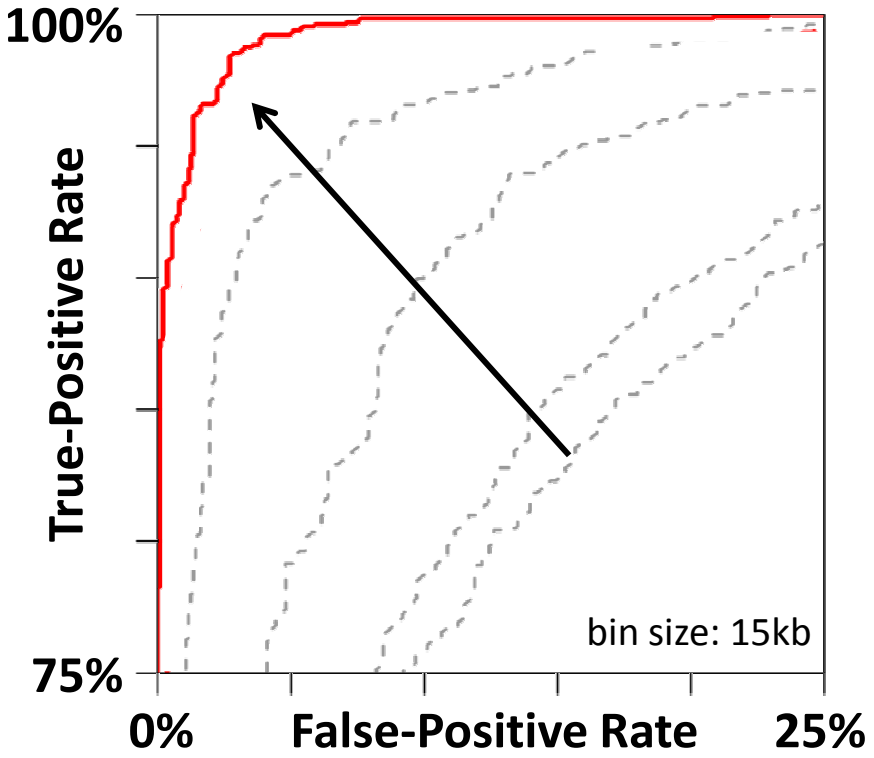
Data:

(1) Pick a random sample.
(2) Find a clear CN change point.
(3) Two CN states: GAIN and NORMAL.



Assessment via ROC:

(1) Quantify how well we can call GAIN:s from NORMAL:s.



Repeat:

Repeat the above for several change points.

TumorBoost

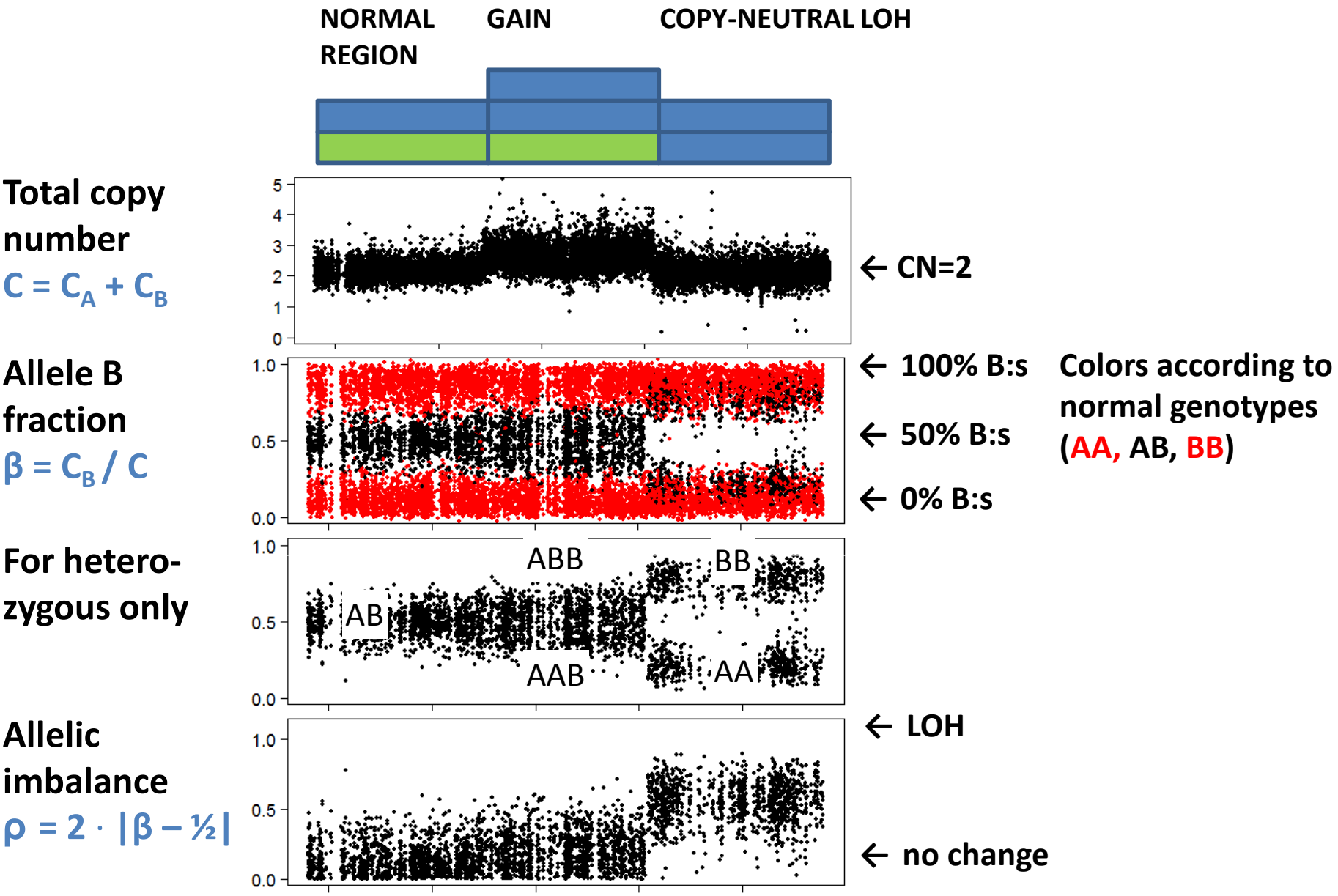
Better allele-specific copy numbers in tumors with matched normals

H. Bengtsson, P. Neuvial, T.P. Speed
TumorBoost: Normalization of allele-specific tumor copy numbers from one single tumor-normal pair of genotyping microarrays. (to be submitted)

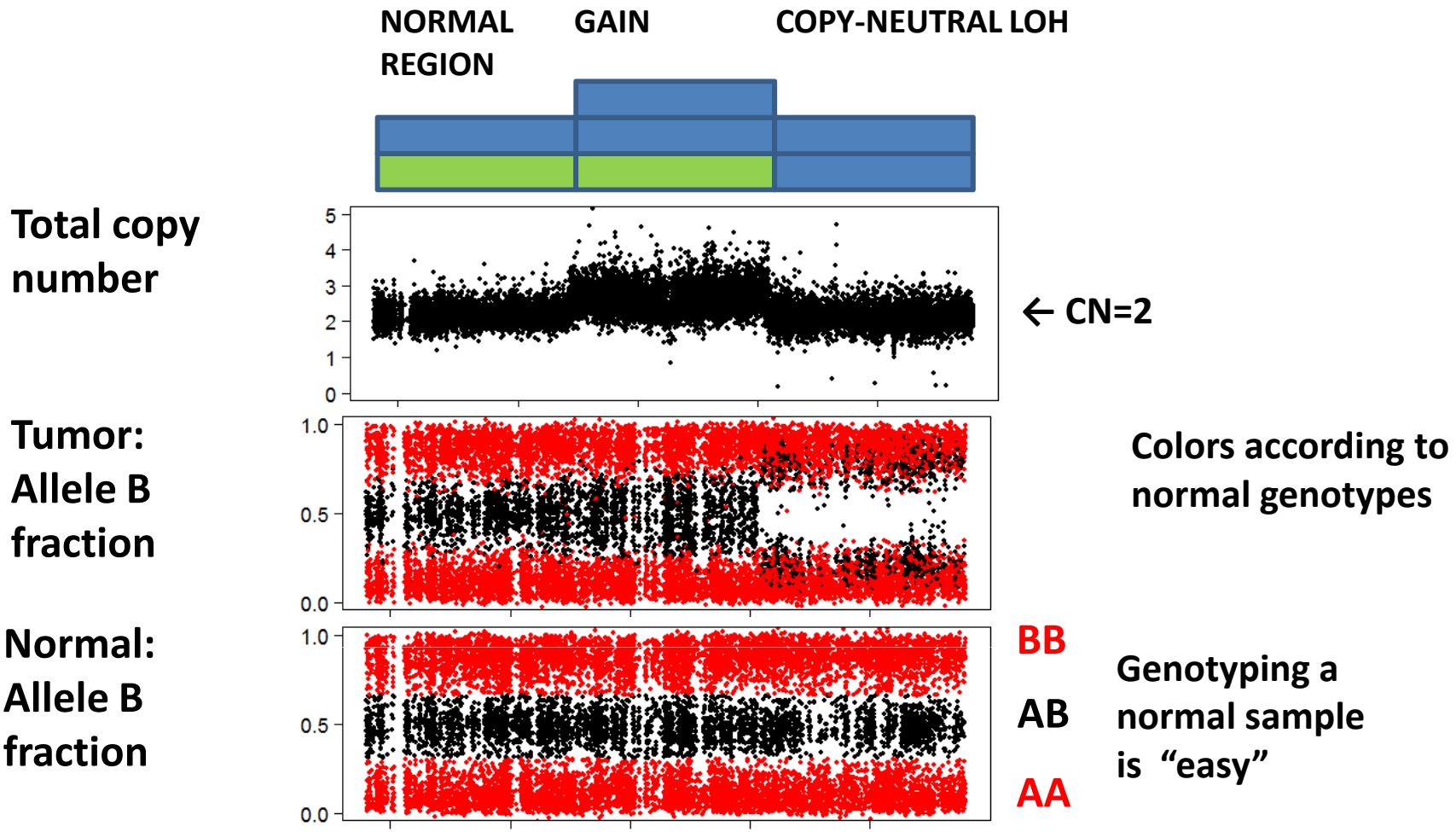
B. Carvalho, H. Bengtsson, T.P. Speed, R.A. Irizarry
Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data, Biostatistics, 2006.

Software: aroma.cn, aroma.affymetrix

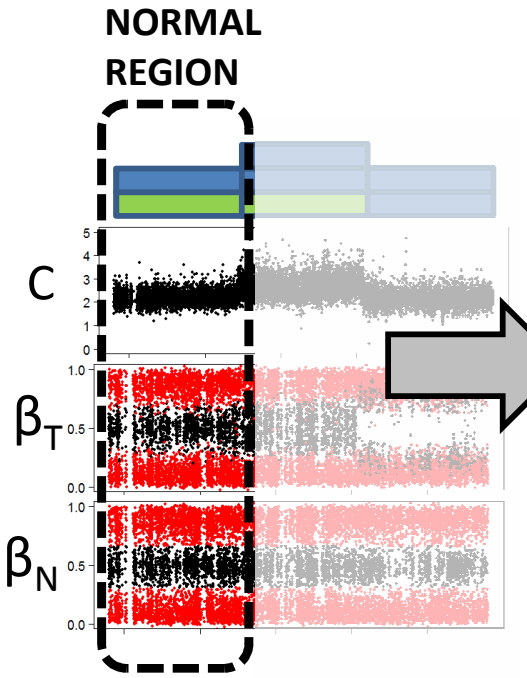
Observed Allelic Imbalances via Allele B fractions



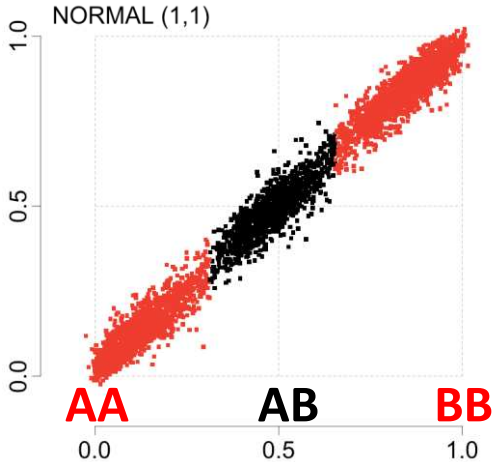
Allele B fractions in a tumor and matched normal



Exploratory data analysis: Tumor-normal Allele B fraction pairs



β_T

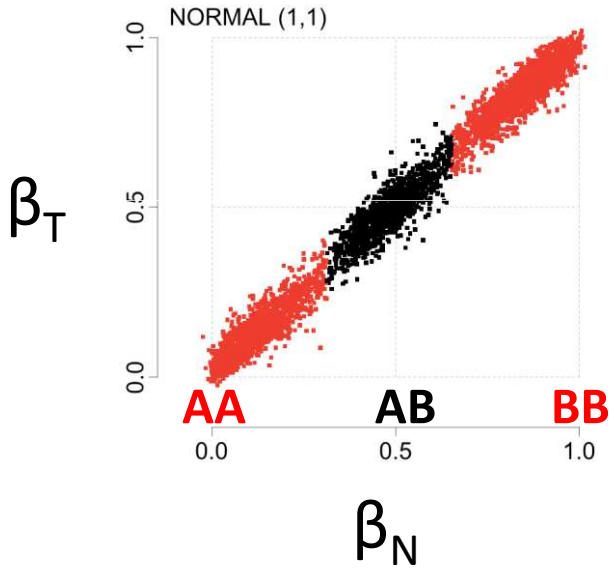


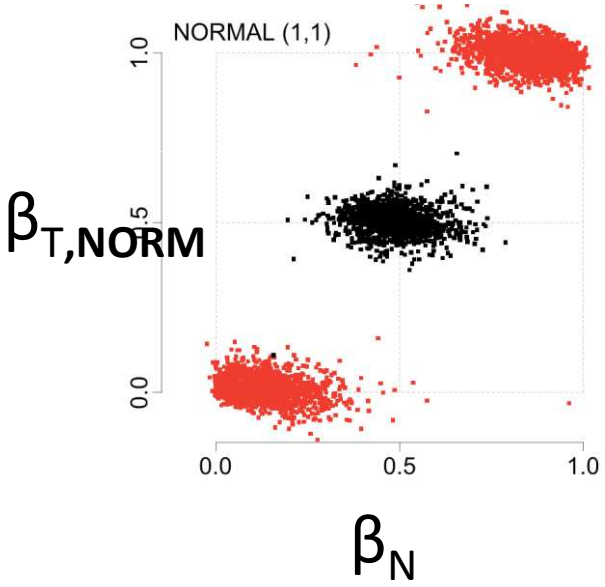
β_N

For each genotype:
$Cor(\beta_T, \beta_N) \approx 1$

A <u>shared SNP effect</u>:
systematic variation

Estimation of the SNP effect





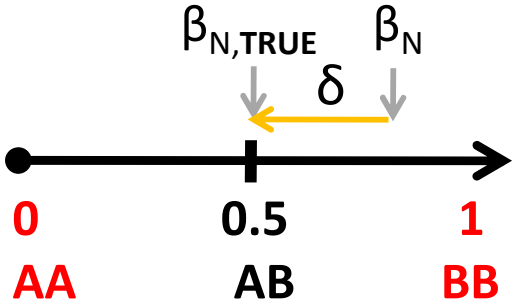
Observed/Called
Allele B fractions
$\beta_N \in [0,1]$
$\beta_T \in [0,1]$

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

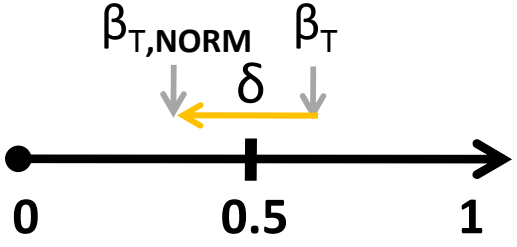
"Estimate"
SNP effect
$\delta = \beta_N - \beta_{N,TRUE}$

"Normalize"
$\beta_{T,NORM} = \beta_T - \delta$

1. Estimate SNP effect in the normal and its genotypes

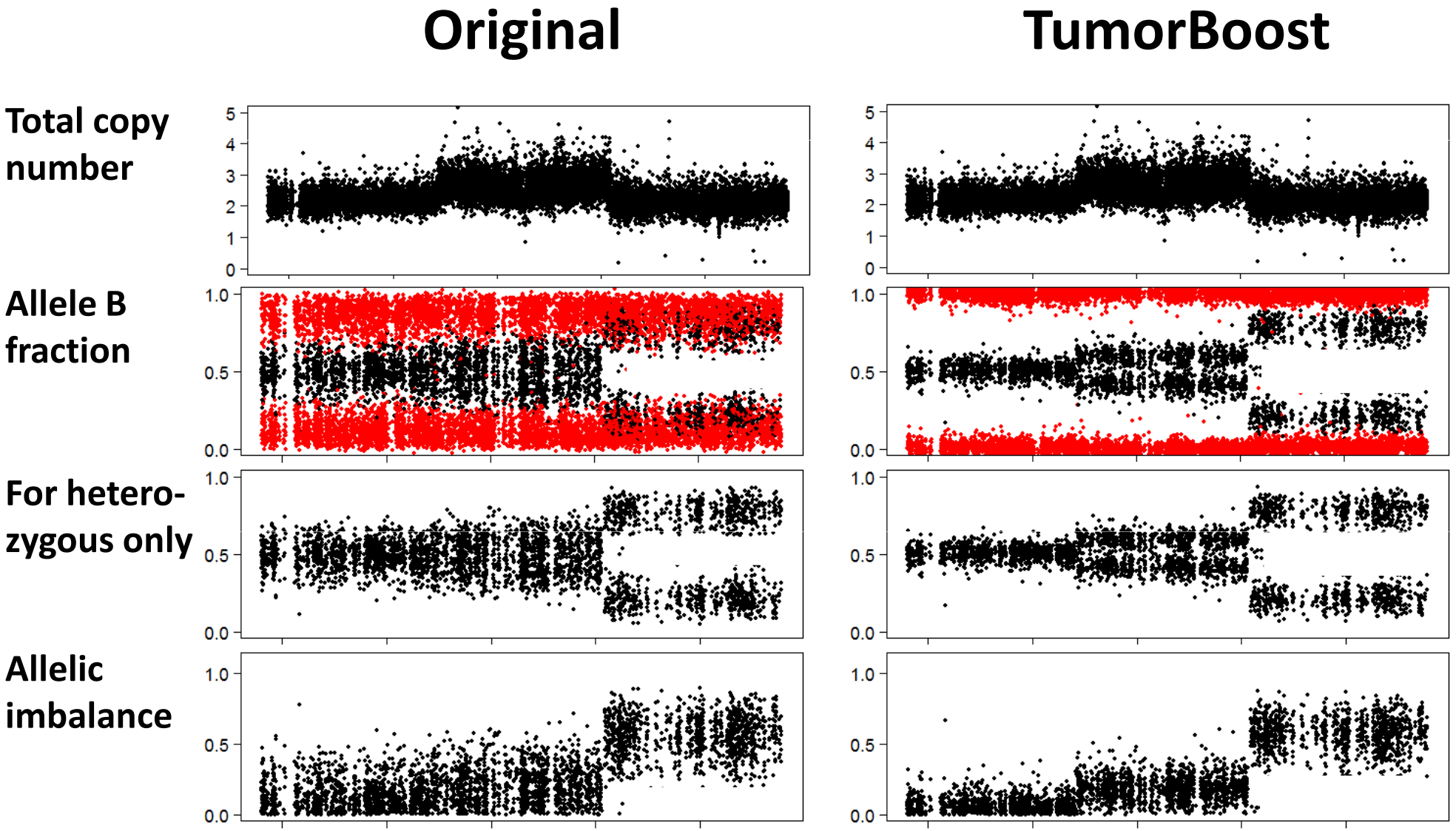


2. Remove SNP effect from the tumor

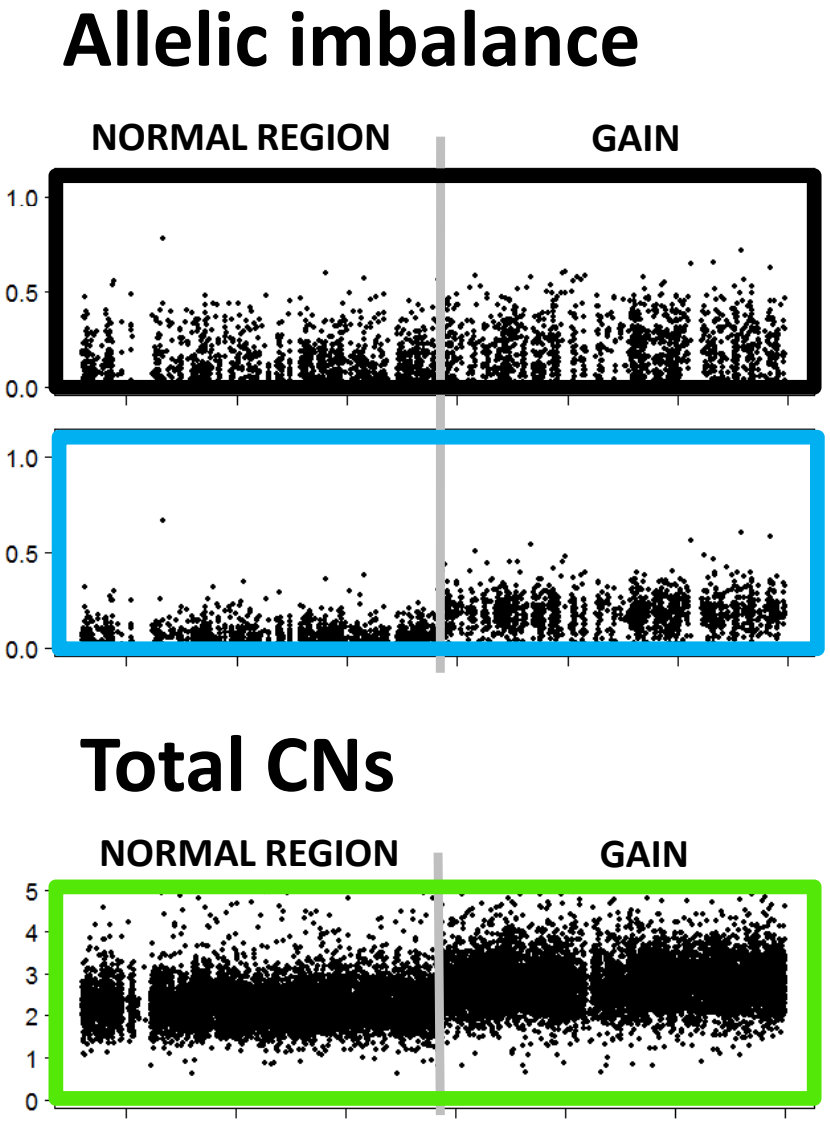


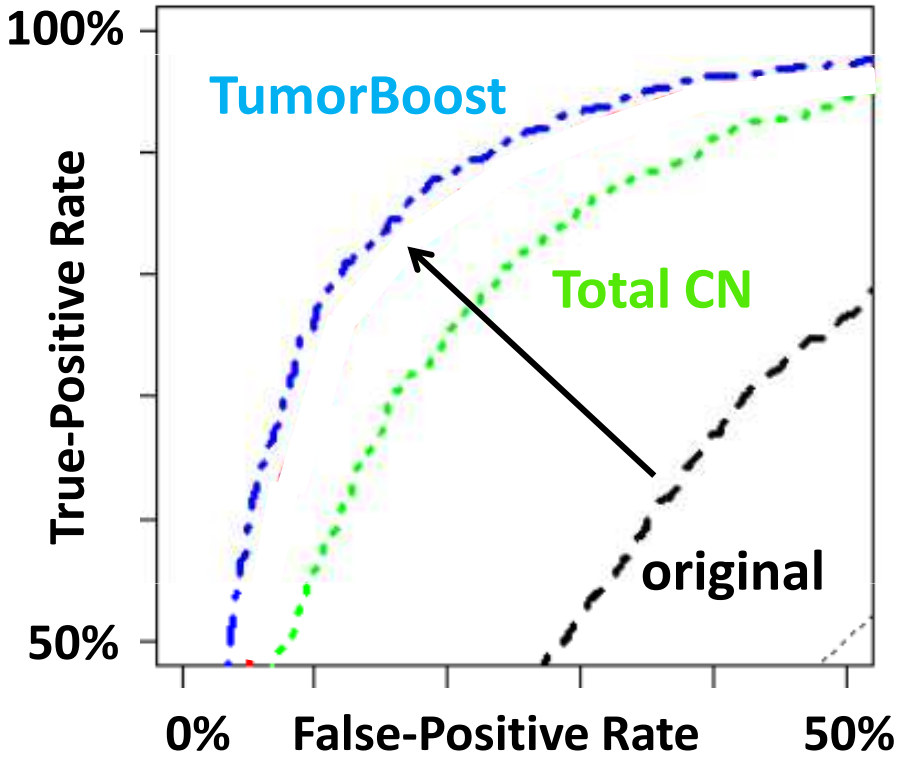
3. Repeat for all SNPs.

Before and after TumorBoost normalization



Result:
Better detection of allelic imbalances





Summary

Problem-driven methodologies:

- Biological knowledge (CN events, …)
- Assay and Technology
- Computationally challenging
- Sustainable methods

Interface of computation, biology and statistics

Acknowledgments

UC Berkeley

James Bullard

Sandrine Dudoit

Kasper Hansen

Pierre Neuvial

Elizabeth Purdom

Terry Speed

LBL

Amrita Ray

Paul Spellman

WEHI , Melbourne

Mark Robinson

Hamish Scott, Catherine Carmichael

Ken Simpson

Gordon Smyth

Swiss Institute of Bioinformatics

Pratyaksha Wirapati

John Hopkins

Benilton Carvalho

Rafael Irizarry

Broad Institute

Gaddy Getz, Scott Carter

Murdoch Childrens Research Institute, Melbourne

Howard Slater, Damien Bruno

Andrew Sinclair, Katarina Bell

Stockholm University

Ola Hössjer

Oncology, Lund University

Åke Borg, Göran Jönsson, Johan Vallon-Christersson, Johan Staaf

Affymetrix

Ben Bolstad, Simon Cawley, Jim Veitch

aroma-project.org:

- aroma.affymetrix
  CRMA, FIRMA
- aroma.cn
  TumorBoost

Selected publications

H Bengtsson, A Ray, P Spellman, TP Speed, *A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods*. Bioinformatics, 2009.

H Bengtsson, P Wirapati, TP Speed, *A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6.* Bioinformatics, 2009.

H Bengtsson, RA Irizarry, B Carvalho, TP Speed, *Estimation and assessment of raw copy numbers at the single locus level.* Bioinformatics, 2008.

H Bengtsson, O Hössjer, *Methodological study of affine transformations of gene expression data with proposed robust non-parametric multi-dimensional normalization method.* BMC Bioinformatics, 2006.

B Carvalho, H Bengtsson, TP Speed, RA Irizarry, *Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data,* Biostatistics, 2006.

A Bengtsson, H Bengtsson, *Microarray image analysis: background estimation using quantile and morphological filters*, BMC Bioinformatics, 2006.

H Bengtsson, G Jönsson, J Vallon-Christersson, *Calibration and assessment of channel-specific biases in microarray data with extended dynamical range*. BMC Bioinformatics, 2004.

Submitted

E Sutton, P Thomas, …, H Bengtsson, …, A Sinclair, *A novel role for SOX3 in mouse and human XX male sex reversal*, 2009.

C Carmichael, E Wilkins, H Bengtsson, …, H Scott, *Poor prognosis in familial acute myeloid leukemia with combined biallelic CEBPA mutations and downstream events affecting the ATM, FLT3 and CDX2 genes,* 2009.

H Bengtsson, P Neuvial, TP Speed, *TumorBoost: Normalization of allele-specific tumor copy numbers from one single tumor-normal pair of genotyping microarrays.* (to be submitted).

Selected software

aroma-project.org:

- aroma.affymetrix
  CRMA, FIRMA
- aroma.cn
  MSCN, TumorBoost
- aroma.cn.eval
  Assessment of copy number results
- aroma.tcga
  TCGA pipelines
- aroma.light
  Efficient and robust estimators

Work was supported by:

NCI/TCGA

American-Scandinavian Foundation
Wenner-Gren Foundation

Solander Foundation
Lennander Foundation

STINT
Blanceflor Boncompagni-Ludovisi

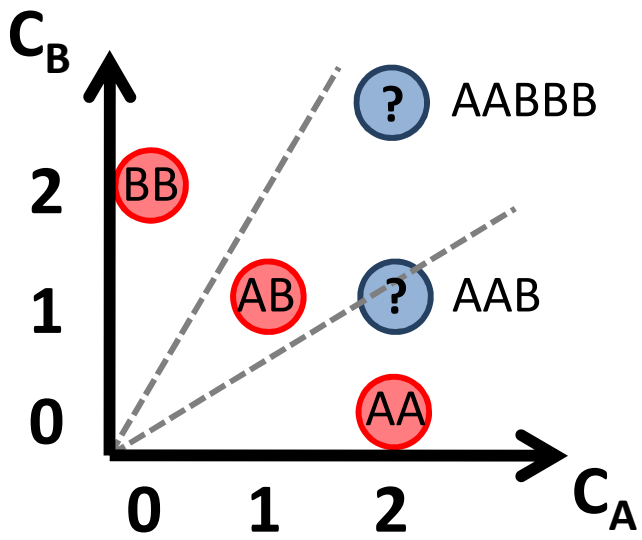
The Fulbright Commission

Extra slides

Tumor copy numbers and allelic imbalances are not discrete

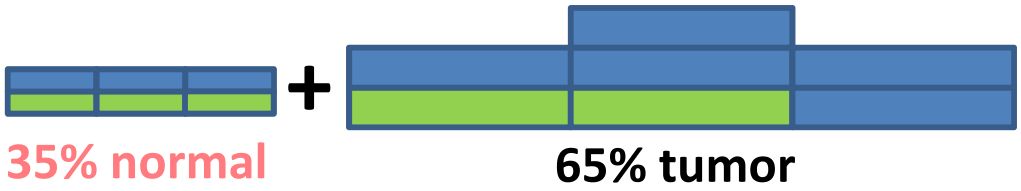
1. Genotyping algorithms are not designed for tumors

- Calls only AA, AB, BB:
  - AAB: called as AA or AB
- Bad at calling gains, e.g. AABBB
- Poor when data is noisy
  - No concept of segmentation



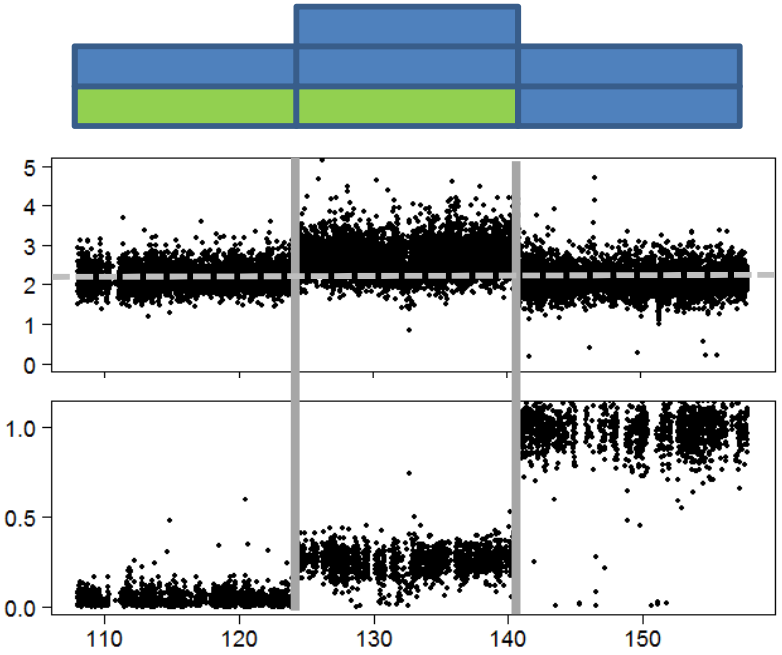
2. Tumor samples are often contaminated with normal cells

- E.g. 65% tumor + 35% normal
- Normal contamination shrinks:
  - the total CN toward 2 copies
  - the allelic imbalance toward 0

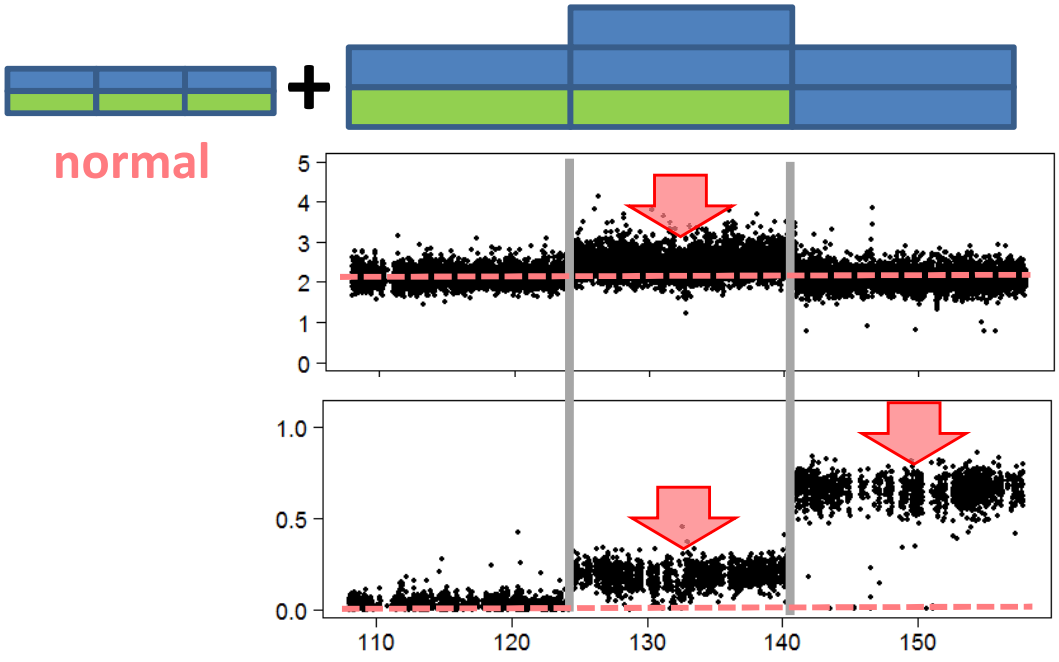


We use continuous signals to measure tumors since they are contaminated with normal cells

100% tumor

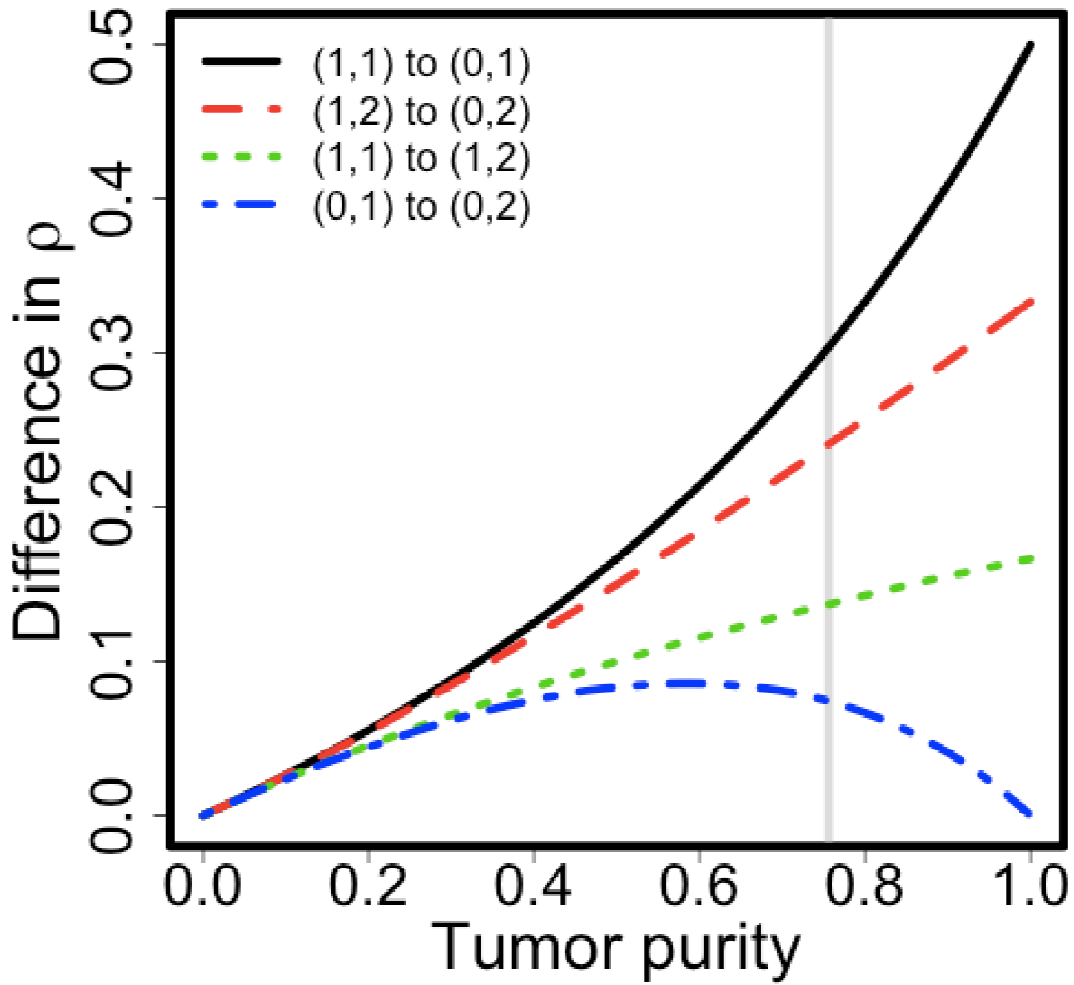


65% tumor and 35% normal

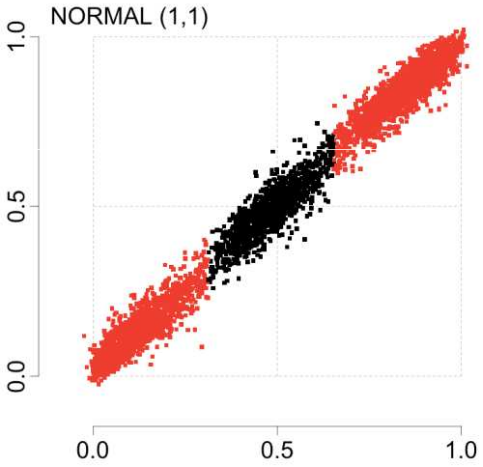


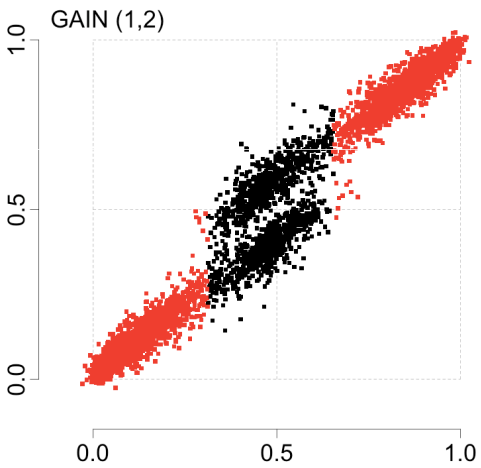
The signals become more and more normal

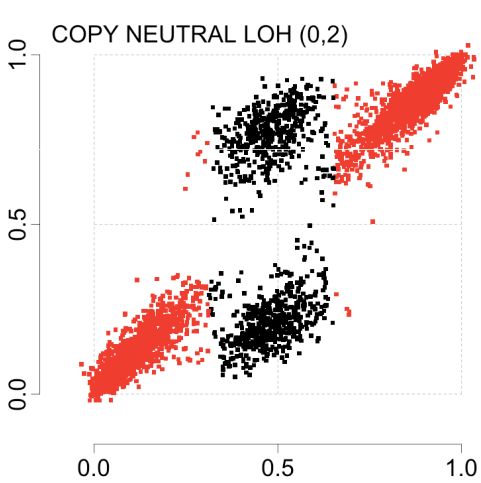
Power to detect Allelic Imbalances

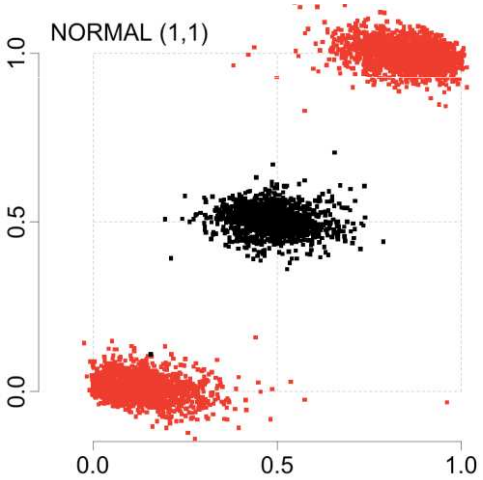


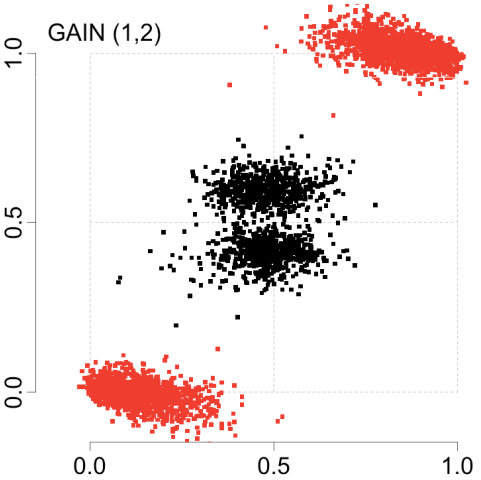
Before and after TumorBoost

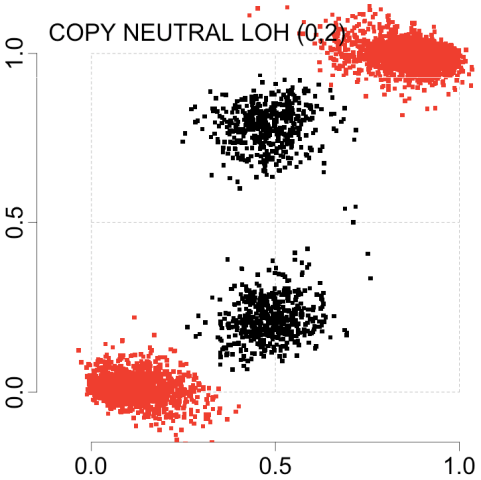












Observed Allelic Imbalances via Allele B fractions