Statistical Analysis of Single Nucleotide Polymorphism Microarrays in Cancer Studies

Stanford Biostatistics Workshop

Pierre Neuvial
with Henrik Bengtsson and Terry Speed

Department of Statistics, UC Berkeley

September 30, 2010

Outline

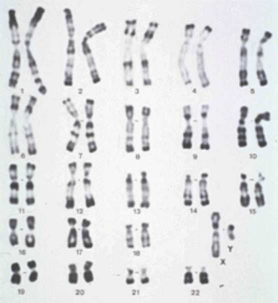
1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data
2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation
3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both C and DH
   - Calling: influence of tumor purity, ploidy, and signal saturation

Outline

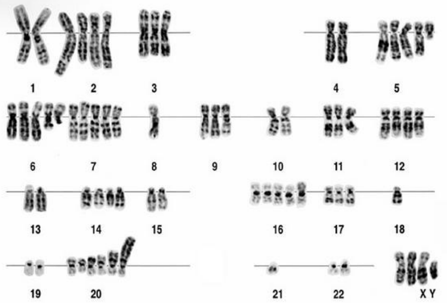
1. Genotying microarrays in cancer research
    - DNA copy number changes in cancer cells
    - Genotyping microarray data
2. TumorBoost: improved power to detect CN changes
    - Method: taking advantage of SNP effects
    - Results: improved signal to noise ratio
    - ROC evaluation
3. Challenges for detecting and calling of copy number events
    - Detecting copy number changes from both C and DH
    - Calling: influence of tumor purity, ploidy, and signal saturation

Genomic changes at the DNA level are hallmarks of cancer

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



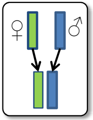
Normal karyotype



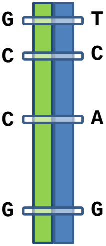
Tumor karyotype

Our goal: identify CN changes to improve characterization, classification, and treatment of cancers

Genotypes in a diploid chromosome

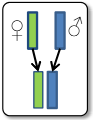


Single nucleotide polymorphism

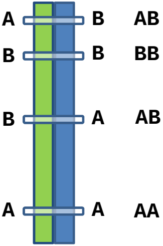


10-20 million
known SNPs

Genotypes in a diploid chromosome

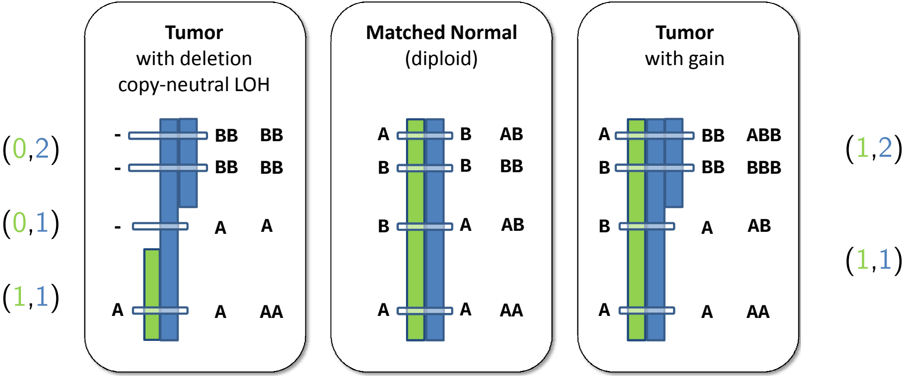


Single nucleotide polymorphism



10-20 million
known SNPs

Genotypes and copy numbers in a tumor



Parental, minor and major copy numbers

Parental copy numbers at genomic locus j: (m_j, p_j), the **unobserved** number of maternal and paternal chromosomes at j.

Copy number state at genomic locus j

$$CN = (C_{1j}, C_{2j})\,,$$

where $C_{1j} = \min(m_j, p_j)$ and $C_{2j} = \max(m_j, p_j)$.

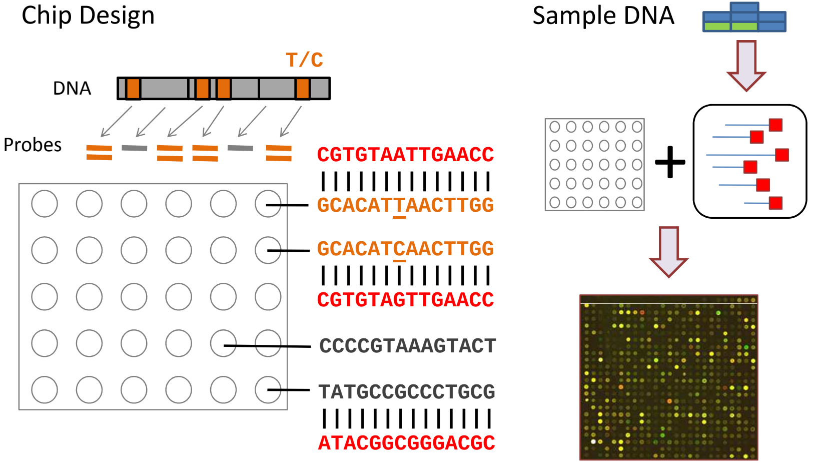
Minor (C_1) and major (C_2) copy numbers:

- characterize the above CN events in cancers
- can be estimated from SNP arrays

Outline

1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data
2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation
3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both C and DH
   - Calling: influence of tumor purity, ploidy, and signal saturation

Technology: Copy number and genotyping microarrays



(C_1, C_2) can be estimated from SNP arrays

For SNP j in sample i, observed signal intensities can be summarized as (θ, β), where $\theta_{ij} = \theta_{Aij} + \theta_{ijB}$ and $\beta_{ij} = \theta_{ijB}/\theta_{ij}$.

Total copy numbers

$$C_{ij} = 2\frac{\theta_{ij}}{\theta_{Rj}} = C_{1ij} + C_{2ij}$$

Decrease in heterozygosity

$$DH_{ij} = 2\,|\beta_{ij} - 1/2| = \frac{C_{2ij} - C_{1ij}}{C_{2ij} + C_{1ij}}$$

Notes:

- DH only defined for SNPs that were **heterozygous in the germline**
- Both dimensions are needed to understand what is going on:
  - Copy neutral LOH: $CN = (0, 2)$, normal total copy number
  - Balanced duplication: $CN = (2, 2)$, allelic balance

The Cancer Genome Atlas (TCGA)

"Accelerate our understanding of the molecular basis of cancer"

- 20 tumor types: brain (glioblastoma multiforme), ovarian, breast, lung, leukemia (AML)...
- Large studies: 500 tumor-normal pairs for each tumor type
- Data levels: DNA copy number, gene expression, DNA methylation
- Platforms: microarray and sequencing

For SNP arrays: identify **copy number changes**: (C, DH) or (C_1, C_2):

1. **detection**: finding regions
2. **classification** labeling regions

Data shown in this presentation: high-grade serous ovarian adenocarcinoma (OvCa).

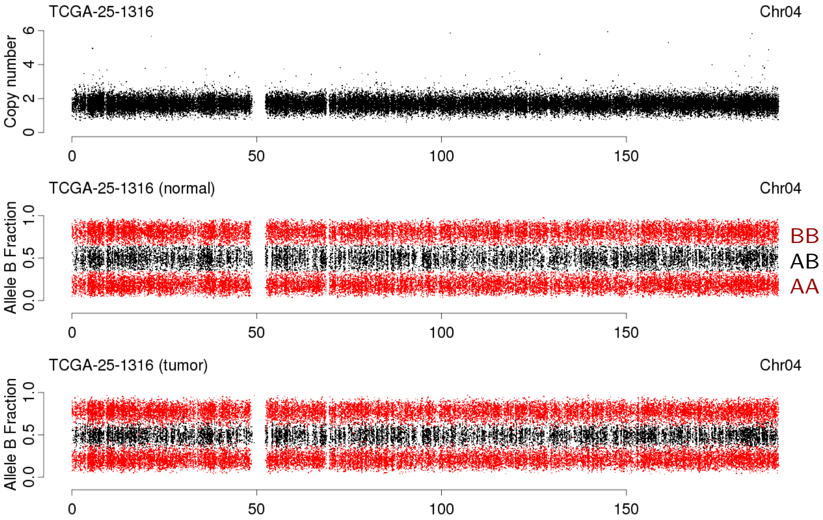
Outline

1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data

2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation

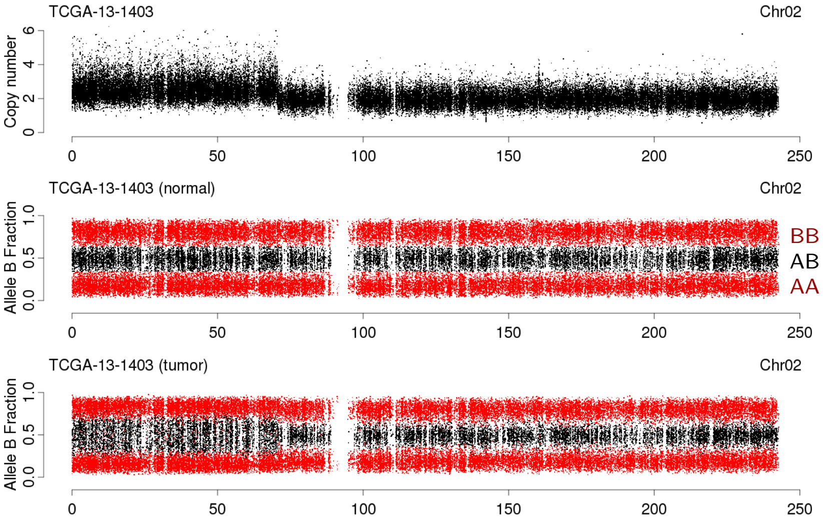
3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both C and DH
   - Calling: influence of tumor purity, ploidy, and signal saturation

No copy number change: (1,1)



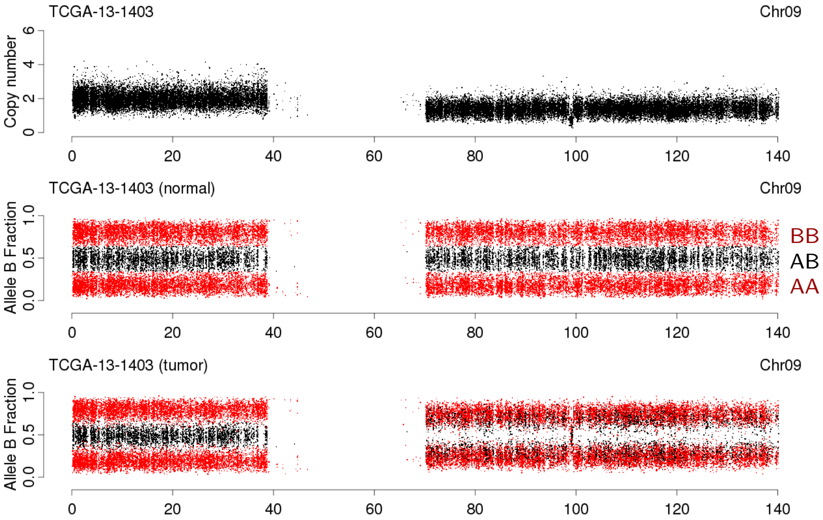
Homozygous SNPs in the normal sample are highlighted in red.

Gain: (1, 2)



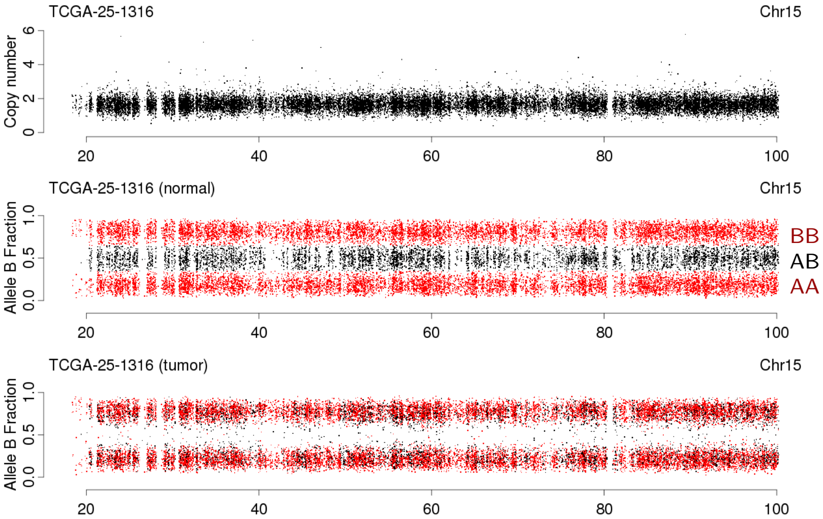
Homozygous SNPs in the normal sample are highlighted in red.

Deletion: (0, 1)



Homozygous SNPs in the normal sample are highlighted in red.

Copy number neutral LOH: (0, 2)



Homozygous SNPs in the normal sample are highlighted in red.

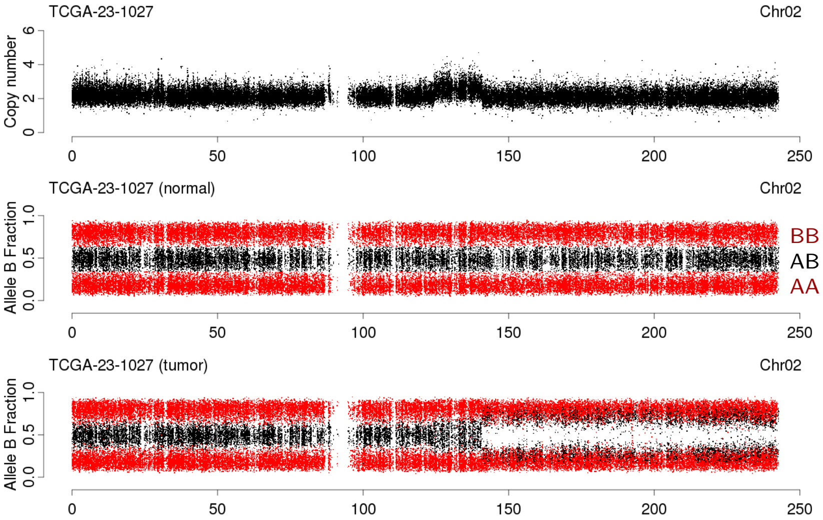
Tumor purity/normal contamination

In practice what we call tumor samples are actually **a mixture of tumor and normal cells.**

The ones just shown have the largest fraction of tumor cells in the data set.

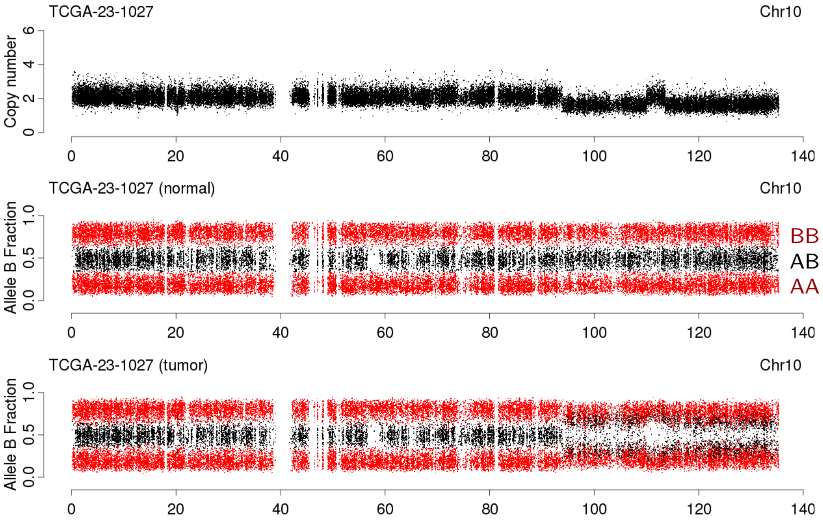
In presence of normal contamination allele B fractions for heterozygous SNPs are **shrunk toward 1/2**.

Normal, gain, copy neutral LOH



Homozygous SNPs in the normal sample are highlighted in red.

Normal, deletion, copy neutral LOH



Homozygous SNPs in the normal sample are highlighted in red.

Outline

1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data

2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation

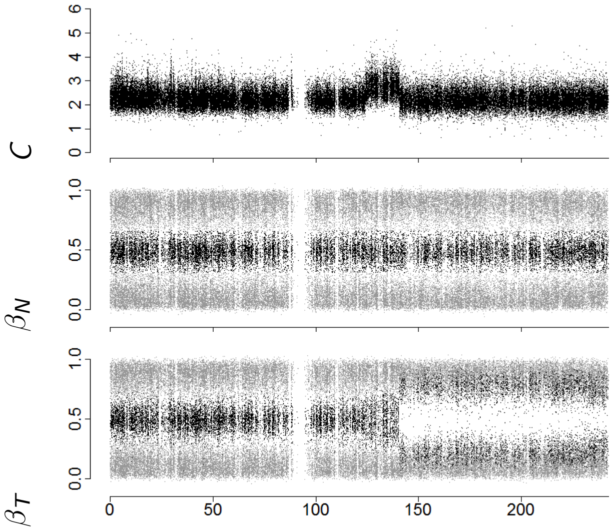
3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both C and DH
   - Calling: influence of tumor purity, ploidy, and signal saturation

Outline

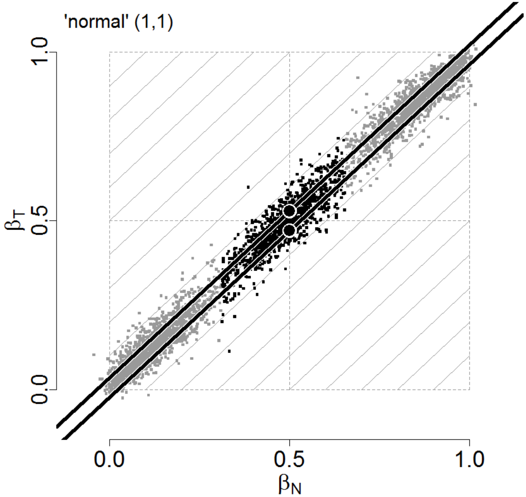
1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data
2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation
3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both C and DH
   - Calling: influence of tumor purity, ploidy, and signal saturation

Raw genomic signals: allelic ratios are noisy

After preprocessing using the CRMAv2 method



SNP effect in a region of no CN change in the tumor

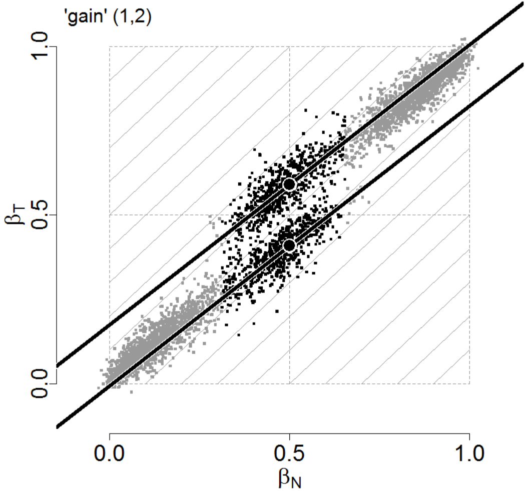


- Instead of three points at (0,0), $(\frac{1}{2}, \frac{1}{2})$ and (1,1), we have three clusters; the observed deviation is a *SNP effect*:

$$\delta_{ij} = \beta_{ij} - \mu_{ij}$$

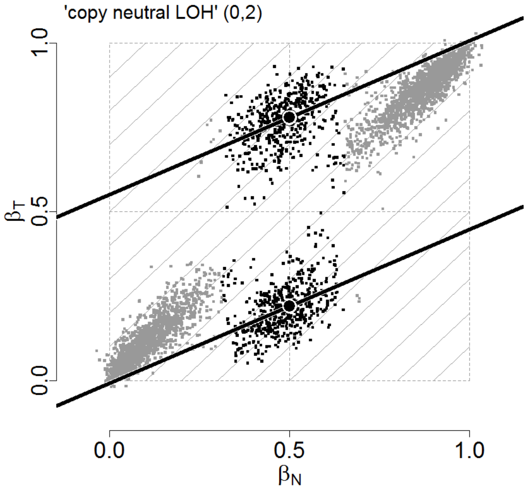
- δ is quite reproducible between the normal and the tumor

SNP effect in a region where tumor has a gain



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

SNP effect in a region where tumor is CNNLOH



- Homozygous clusters are similar as before
- Heterozygous cluster is even more tilted

Overview of the TumorBoost method

Idea

1. the SNP effect is reproducible between tumor and normal
2. in the normal the truth is easier to infer because we only expect three true allele B fractions, corresponding to genotypes AA, AB, BB.

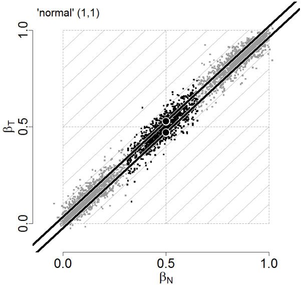
⇒ For each SNP, we estimate the SNP effect in the normal hybridization, and “subtract” it from the tumor.

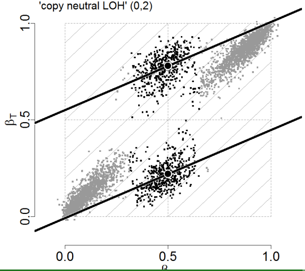
Features

- No need to know copy number regions in advance
- Normalization is performed for each SNP separately
- Only one tumor/normal pair required

H. Bengtsson, P. Neuvial, T.P. Speed
TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics* (2010) 11:245.

Proposed normalization strategy





Estimate the SNP effect in the normal sample as

$$\hat{\delta}_{Nj} = \beta_{Nj} - \hat{\mu}_{Nj} \, ,$$

where $\hat{\mu}_{Nj} \in \{0, 1/2, 1\}$ is the normal genotype

For homozygous SNPs ($\hat{\mu}_{Nj} \in \{0, 1\}$):

$$\tilde{\beta}_{Tj} = \beta_{Tj} - \beta_{Nj} + \hat{\mu}_{Nj}$$

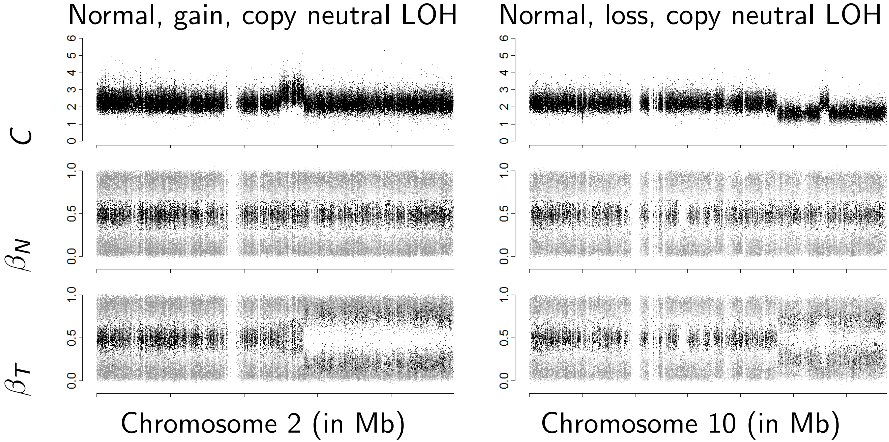
For heterozygous SNPs ($\hat{\mu}_{Nj} 1/2$):

$$\tilde{\beta}_{Tj} = \begin{cases} \frac{1}{2} \cdot \frac{\beta_{Tj}}{\beta_{Nj}} & \text{if } \beta_{Tj} < \beta_{Nj} \\ 1 - \frac{1}{2} \cdot \frac{1-\beta_{Tj}}{1-\beta_{Nj}} & \text{otherwise} \end{cases}$$

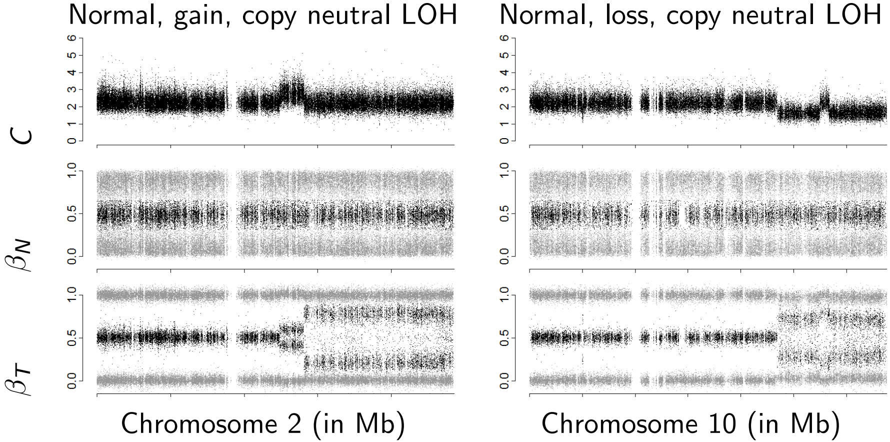
Outline

1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data
2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation
3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both *C* and *DH*
   - Calling: influence of tumor purity, ploidy, and signal saturation

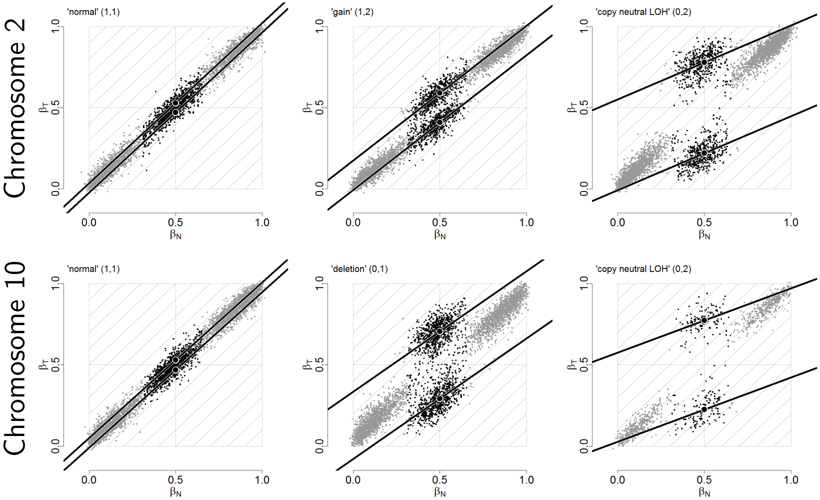
Genomic signals before normalization



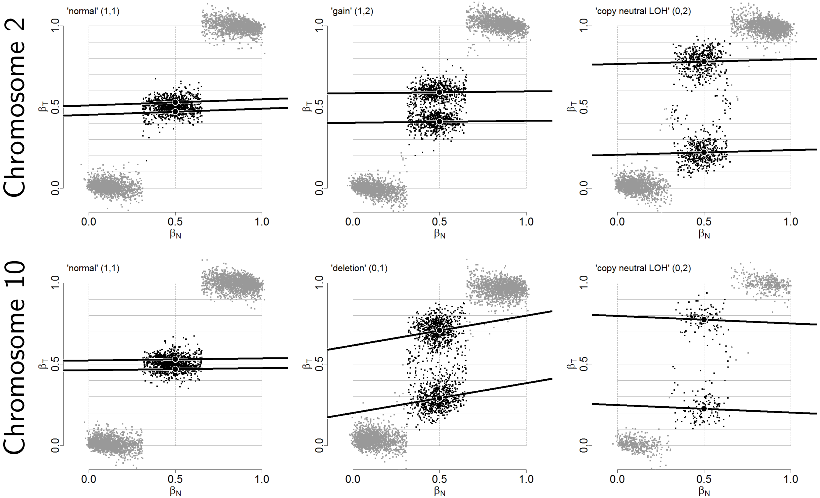
Genomic signals after normalization



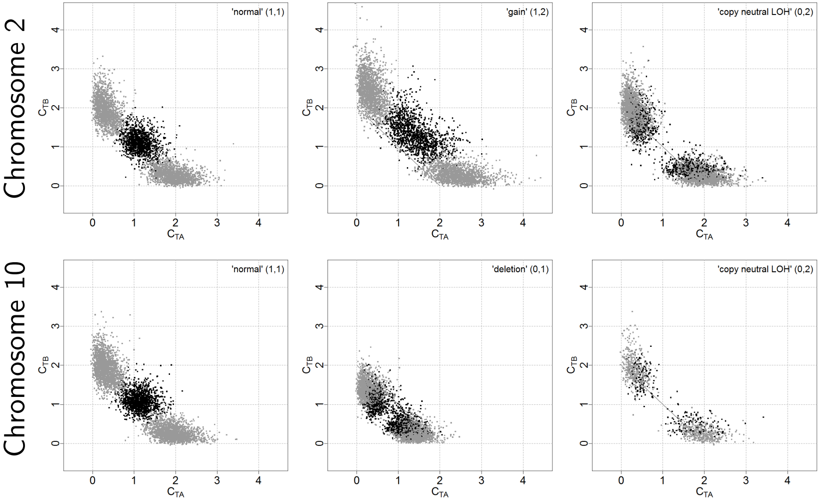
Allele B fractions before normalization



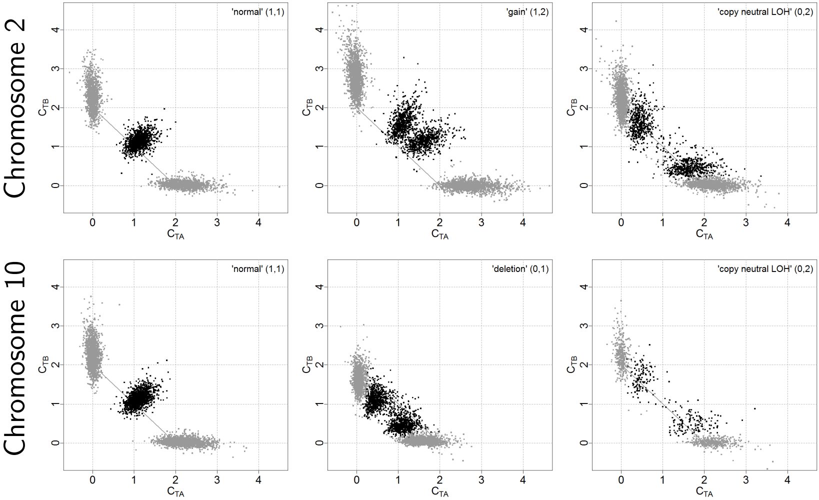
Allele B fractions after normalization



ASCNs before normalization



ASCNs after normalization



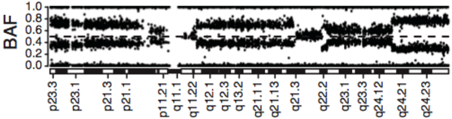
Outline

1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data

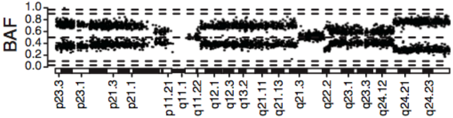
2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation

3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both C and DH
   - Calling: influence of tumor purity, ploidy, and signal saturation

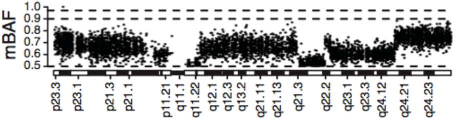
Detecting changes in allele B fractions



allele B fractions: β



allele B fractions for heterozygous SNPs



"mirrored" allele B fractions for heterozygous SNPs:

$$\rho = |\beta - 1/2| = DH/2$$

For heterozygous SNPs *DH* only has one mode so it can be segmented.

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

ROC evaluation

Available from aroma.cn.eval at: `http://aroma-project.org`

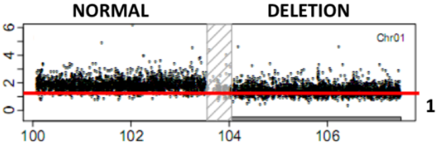
For a given sample:

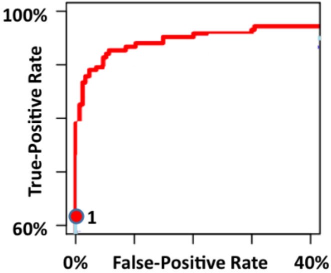
- find a clear change point
- label flanking regions, e.g. NORMAL (1,1) and DELETION (0,1)
- choose one reference state and one state to call

For each value of a threshold τ:

- Call SNPs below τ a DELETION
- Count number of true and false DELETIONs.

ROC curve is built by adjusting τ.





ROC evaluation

Available from aroma.cn.eval at: `http://aroma-project.org`

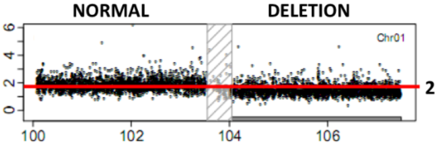
For a given sample:

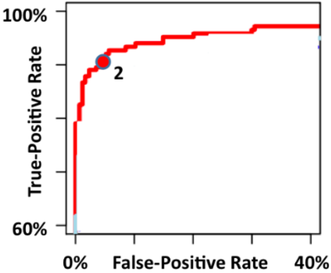
- find a clear change point
- label flanking regions, e.g. NORMAL (1,1) and DELETION (0,1)
- choose one reference state and one state to call

For each value of a threshold τ:

- Call SNPs below τ a DELETION
- Count number of true and false DELETIONs.

ROC curve is built by adjusting τ.





ROC evaluation

Available from aroma.cn.eval at: `http://aroma-project.org`

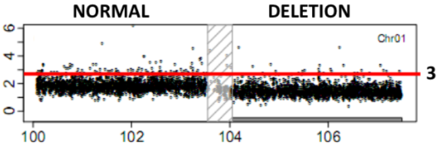
For a given sample:

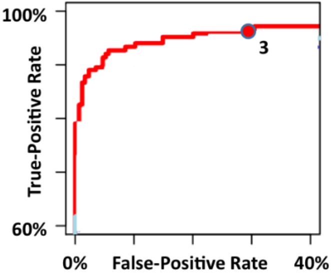
- find a clear change point
- label flanking regions, e.g. NORMAL (1,1) and DELETION (0,1)
- choose one reference state and one state to call

For each value of a threshold τ:

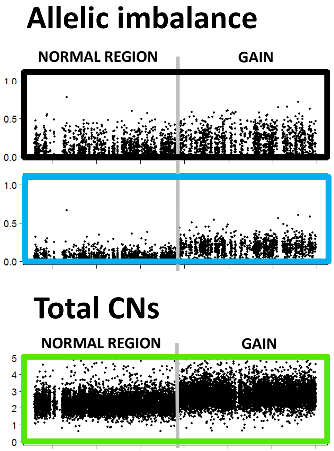
- Call SNPs below τ a DELETION
- Count number of true and false DELETIONs.

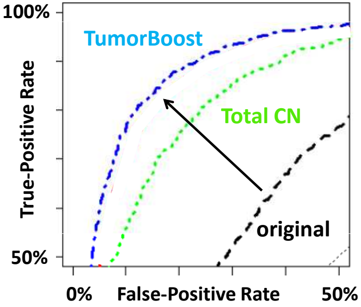
ROC curve is built by adjusting τ.





Result: Better detection of allelic imbalances





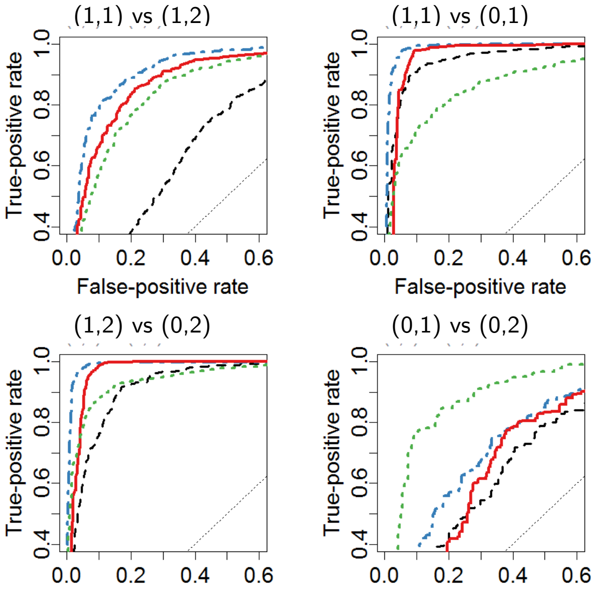
Complete preprocessing for a single tumor/normal pair

Available from aroma.cn and aroma.affymetrix at: `http://aroma-project.org`

- Normalization and locus-level summarization using CRMAv2 (Bengtsson et al, 2009) for the normal and the tumor sample separately
- "Naive" genotyping of the normal sample: threshold density of β
- TumorBoost normalization (Bengtsson et al, 2010)

Note: genotyping errors can be taken care of by smoothing or using confidence scores.

Observed power to detect changes

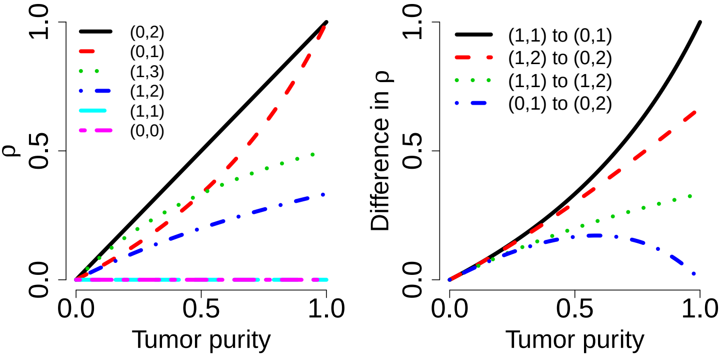


Legend:
Total copy number
Raw allele B fractions
Normalized β (naive)
Normalized β (Birdseed)

- TCN is consistent across change points
- β is not !

Expected power to detect changes

C varies from one unit in all change points just shown
For DH — and thus (C_1, C_2) — it's more complicated:



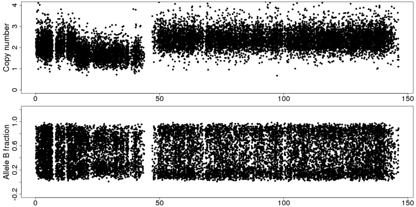
The expected improvement depends on the type of change point and on normal contamination.

What if no matched normal is available? CalMaTe

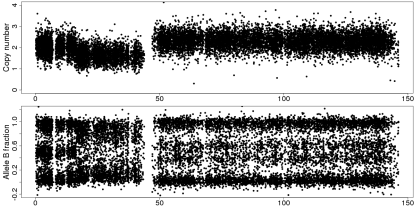
For each SNP:

- Estimate a calibration function (from observed signals to genotypes) using a set of reference samples
- Back-transform test samples

Before CalMate normalization



After CalMate normalization



Outline

1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data

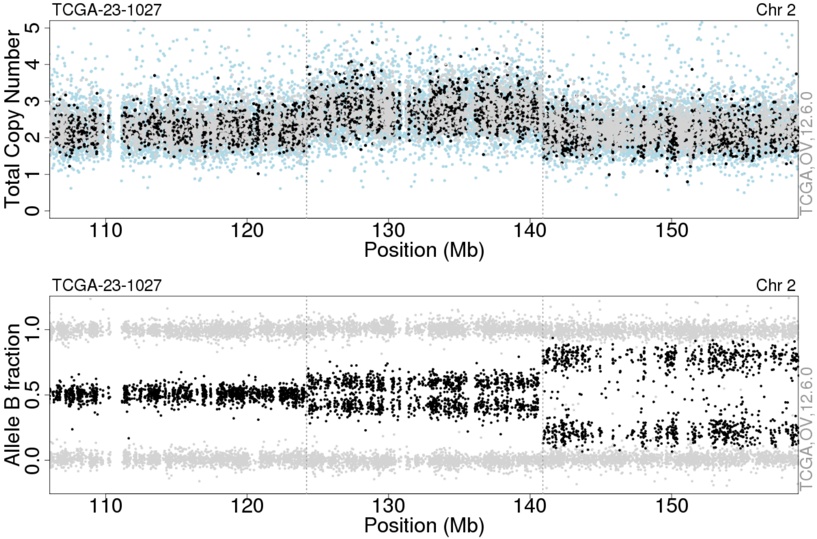
2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation

3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both C and DH
   - Calling: influence of tumor purity, ploidy, and signal saturation

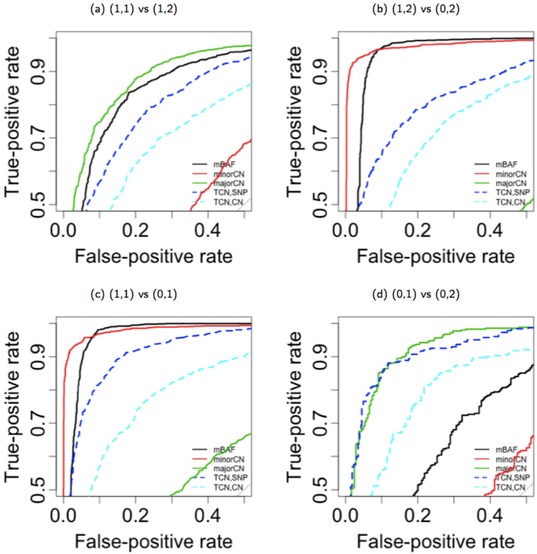
Outline

1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data
2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation
3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both C and DH
   - Calling: influence of tumor purity, ploidy, and signal saturation

Changes can be reflected in both dimensions



DH has greater detection power than C at a single locus



(a) (1,1) vs (1,2)

(b) (1,2) vs (0,2)

(c) (1,1) vs (0,1)

(d) (0,1) vs (0,2)

More informative probes for *C* than *DH*

Affymetrix GenomeWideSNP_6

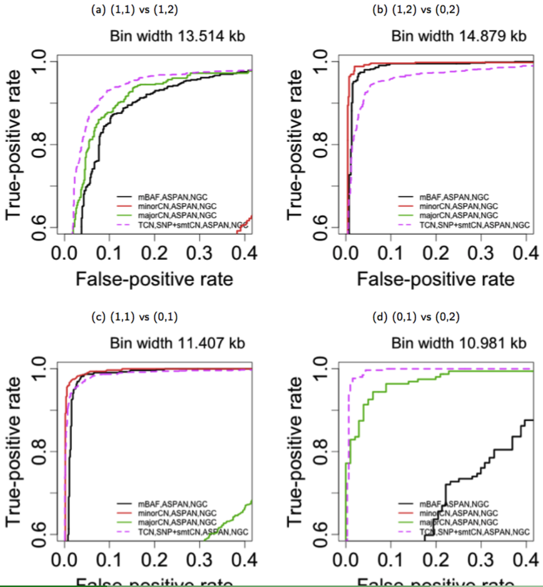
	All units	CN units	SNP units
Frequency	1,856,069	946,705	909,364
Proportion	100%	51%	49%

Unit types

	All units	AA	AB	BB
Frequency	1,856,069	326,500	251,446	331,418
Proportion	100%	18%	**14%**	18%

SNPs by genotype call for sample TCGA-23-1027

Similar detection power at a fixed resolution



(a) (1,1) vs (1,2)

(b) (1,2) vs (0,2)

(c) (1,1) vs (0,1)

(d) (0,1) vs (0,2)

Need for a truly joint dimensional segmentation method

- Most methods segment only *one* of *C* and *DH*
- Some use two-way segmentation: Olshen *et al*, [PSCBS]
- A handful are truly two-dimensional :
  - Chen *et al*, [pscn]
  - Greenman *et al*, Biostat., 2010, [PICNIC]
  - Sun *et al*, NAR, 2009, [genoCNA]

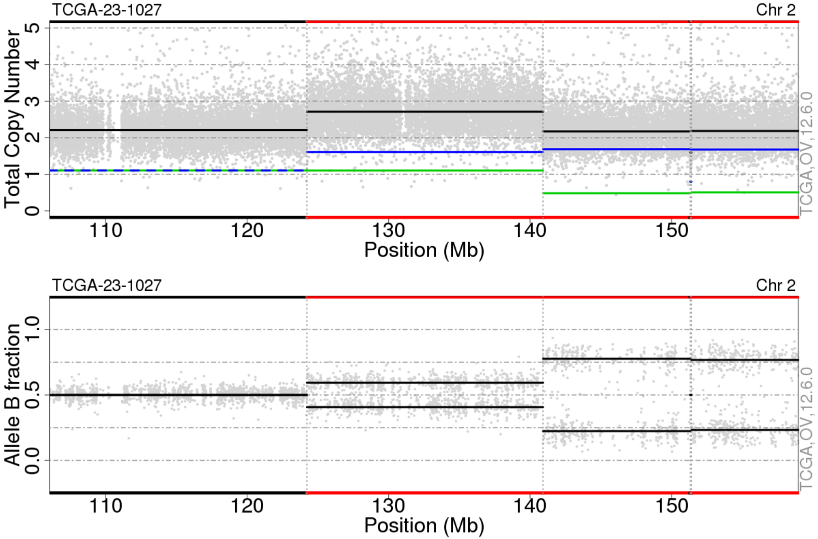
Challenges for a truly joint segmentation method

- A two-dimensional signal
- Only heterozygous SNPs can be used to detect CN changes from *DH*
- Bias in the estimation of *DH*
- *DH* is not Gaussian

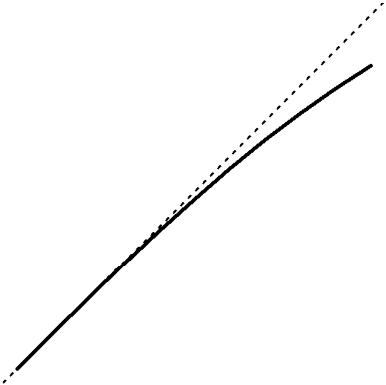
Outline

1. Genotyping microarrays in cancer research
   - DNA copy number changes in cancer cells
   - Genotyping microarray data
2. TumorBoost: improved power to detect CN changes
   - Method: taking advantage of SNP effects
   - Results: improved signal to noise ratio
   - ROC evaluation
3. Challenges for detecting and calling of copy number events
   - Detecting copy number changes from both *C* and *DH*
   - Calling: influence of tumor purity, ploidy, and signal saturation

Copy numbers are not calibrated



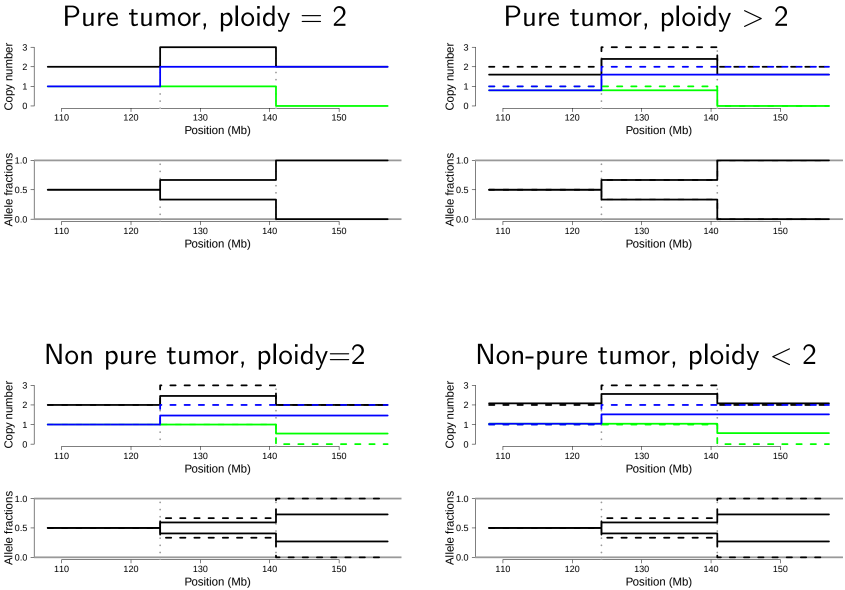
Non-calibrated signals: signal saturation



$$C_{\text{obs}} = f(C_{\text{true}}) < C_{\text{obs}}$$

f is unknown

Non-calibrated signals: ploidy and purity



Pure tumor, ploidy $= 2$

Pure tumor, ploidy > 2

Non pure tumor, ploidy=2

Non-pure tumor, ploidy < 2

Purity, ploidy, and signal saturation

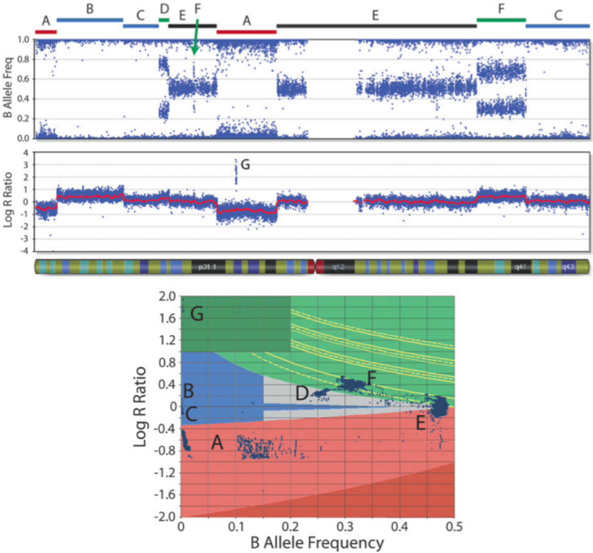
Why copy numbers are not calibrated

- signal saturation
- non purity: presence of normal cells in the "tumor sample"
- ploidy: the total amount of DNA is fixed by the assay

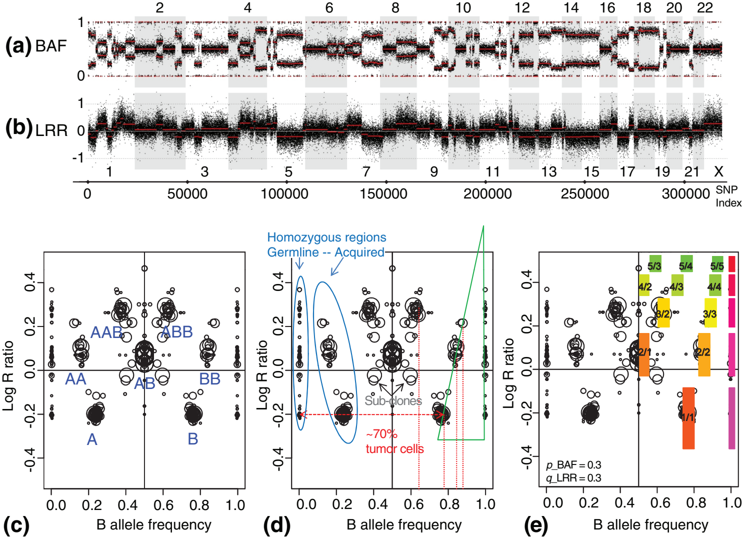
Remarks

- ploidy is **not identifiable**
- purity and ploidy are biological properties of the sample
- signal saturation is an artifact from the assay
- under the rug: tumor heterogeneity

OverUnder: Attiyeh et al, Genome Research, 2009



GAP: Popova et al, Genome Biology, 2009



ASCAT: Van Loo et al, PNAS, 2010

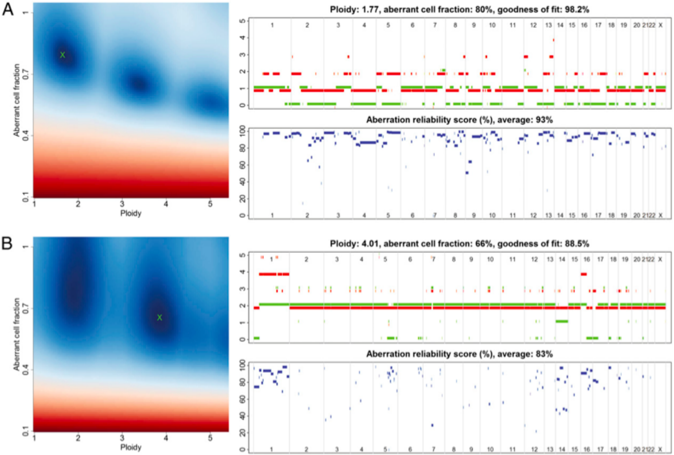


Fig. 1. ASCAT profiles and their calculation. Two examples are given: (*A*) a tumor with ploidy close to 2n and (*B*) a tumor with ploidy close to 4n. (*Left*) ASCAT first determines the ploidy of the tumor cells ψ_t and the fraction of aberrant cells ρ. This procedure evaluates the goodness of fit for a grid of possible values for both parameters (blue, good solution; red, bad solution; detailed in *Materials and Methods*). On the basis of this goodness of fit, the optimal solution is selected (green cross). Using the resulting tumor ploidy and aberrant cell fraction, an ASCAT profile is calculated (*Upper Right*), containing the allele-specific copy number of all assayed loci [copy number on the *y* axis vs. the genomic location on the *x* axis; green, allele with lowest copy number; red, allele with highest copy number; for illustrative purposes only, both lines are slightly shifted (red, down; green, up) such that they do not overlap; only probes heterozygous in the germline are shown]. Finally, for all aberrations found, an aberration reliability score is calculated (*Lower Right*).

Comments on existing approaches

- What about Affymetrix data ?
- Choice between candidate solutions
- Perform ad hoc correction for saturation
- Tumor heterogeneity ?

Thanks

- **Henrik Bengtsson**
- Adam Olshen
- Maria Ortiz
- Angel Rubio
- Venkat E. Seshan
- Terry Speed
- Paul Spellman
- Nancy R. Zhang

Funding: The Cancer Genome Atlas