Normalization of allelic tumor signals from one tumor/normal pair of genotyping microarrays

UC Berkeley Statistics and Genomics Seminar

Pierre Neuvial
with Henrik Bengtsson and Terry Speed

Department of Statistics, UC Berkeley

October 1$^{\mathrm{st}}$, 2009

Outline

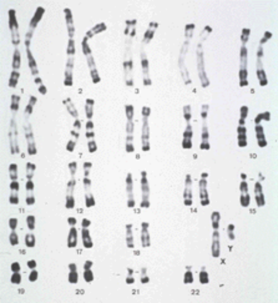
1. Genotyping microarrays in cancer research
2. Normalizing each SNP of a single tumor/normal pair
3. Improved power to detect CN changes
4. Conclusions

Outline

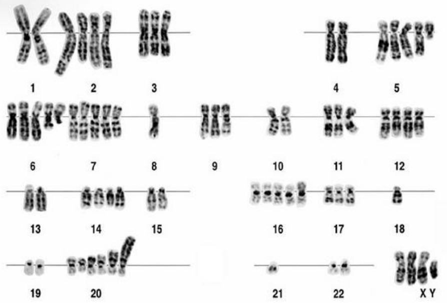
1. Genotyping microarrays in cancer research
   - DNA copy number analyses for cancer research
   - Genotyping microarray data
2. Normalizing each SNP of a single tumor/normal pair
   - SNP effects
   - Proposed normalization
3. Improved power to detect CN changes
   - Allele B fractions along the genome
   - Comparing before and after
   - ROC evaluation
4. Conclusions
   - All we need is a tumor/normal pair
   - Further thoughts

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

Parental copy numbers (PCN)

The number of copies of each parental chromosome.
Notation: $PCN = (C_1, C_2)$, with $C_1 \leq C_2$.
In a region of no genomic alteration : $PCN = (1, 1)$

Genotyping microarrays quantify

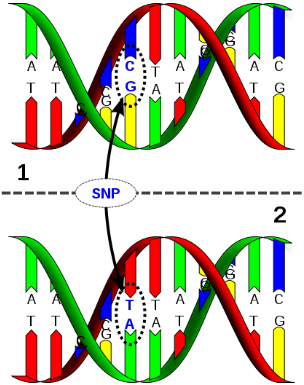
1. total copy number : $TCN = C_1 + C_2$
2. alleleic composition, which is related to $\frac{C_1}{C_1+C_2}$

Both quantities are needed to understand what is happening:

- Copy neutral LOH: $PCN = (0, 2)$
- Balanced duplication: $PCN = (2, 2)$

Single Nucleotide Polymorphisms (SNPs)

SNP: a locus where two different DNA letters can be observed.
These two alleles are noted “A” and “B”. Genotyping microarrays quantify the corresponding amount of DNA in sample i at SNP j: as $(\theta_{ijA}, \theta_{ijB})$.



C/T heterozygosity.

Heterozygous SNPs are informative to identify changes in allelic compostion, using

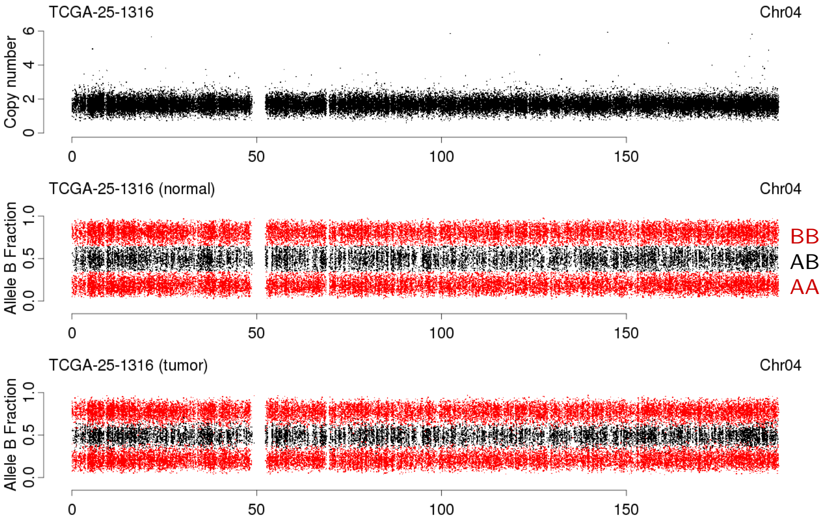
$$[\text{Allele B fraction}]: \beta_{ij} = \frac{\theta_{ijB}}{\theta_{ijA} + \theta_{ijB}}$$

All SNPs are informative to identify changes in total copy number, using

$$[\text{Total copy number}]: C_{ij} = 2\frac{\theta_{ij}}{\theta_{Rj}},$$

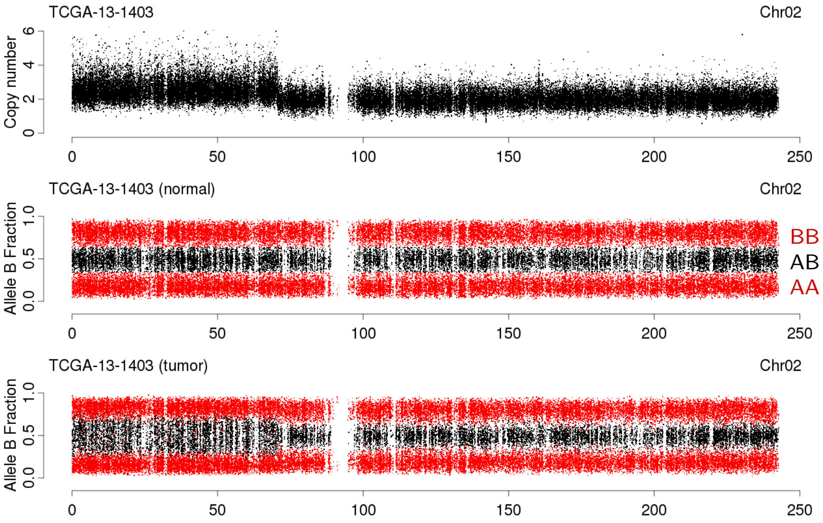
where $\theta_{ij} = \theta_{ijA} + \theta_{ijB}$.

No copy number change: (1,1)



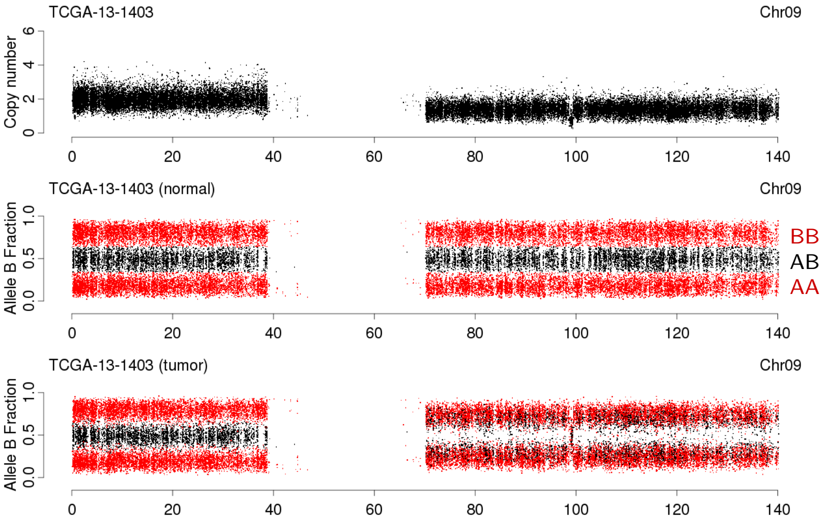
Homozygous SNPs in the normal sample are highlighted in red.

Gain: (1, 2)



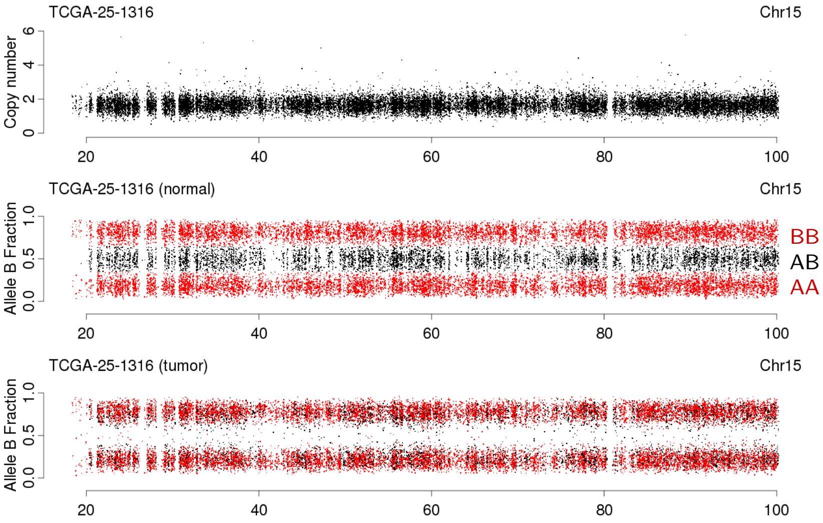
Homozygous SNPs in the normal sample are highlighted in red.

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.

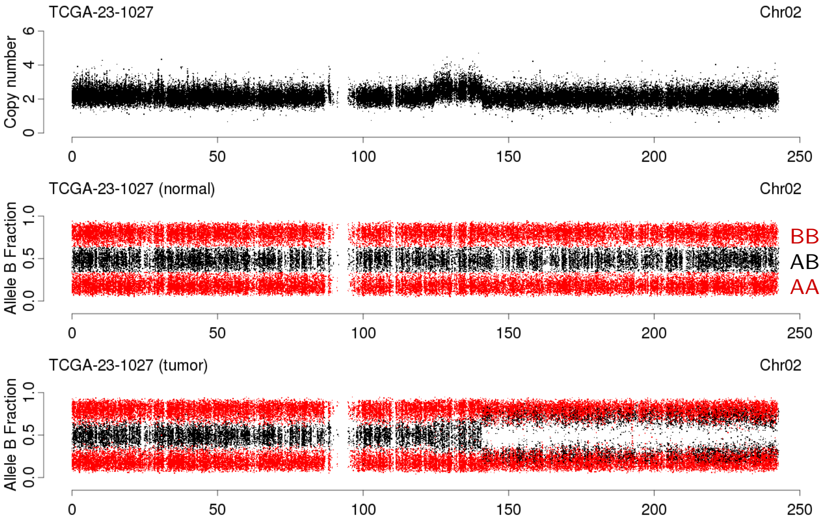
Tumor purity

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.

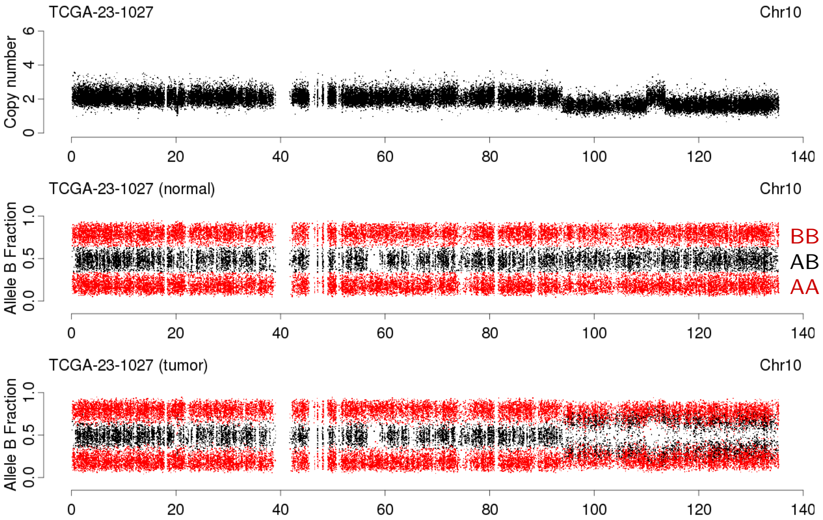
We'll see that in presence of normal contamination allele B fractions for heterozygous SNPs are shrunk toward1/2.

Normal, gain, copy neutral LOH



Homozygous SNPs in the normal sample are highlighted in red.

Normal, deletion, copy neutral LOH

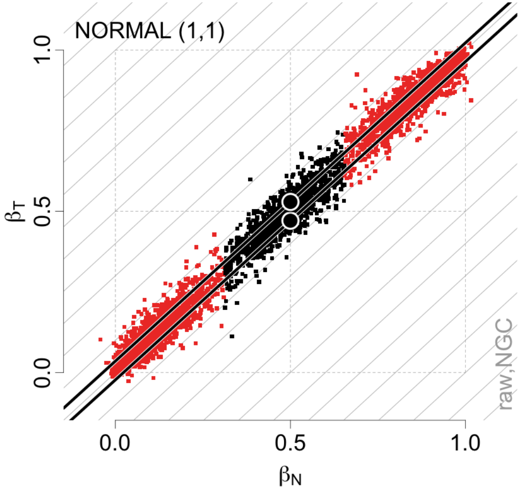


Homozygous SNPs in the normal sample are highlighted in red.

Outline

1. Genotyping microarrays in cancer research
   - DNA copy number analyses for cancer research
   - Genotyping microarray data
2. Normalizing each SNP of a single tumor/normal pair
   - SNP effects
   - Proposed normalization
3. Improved power to detect CN changes
   - Allele B fractions along the genome
   - Comparing before and after
   - ROC evaluation
4. Conclusions
   - All we need is a tumor/normal pair
   - Further thoughts

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

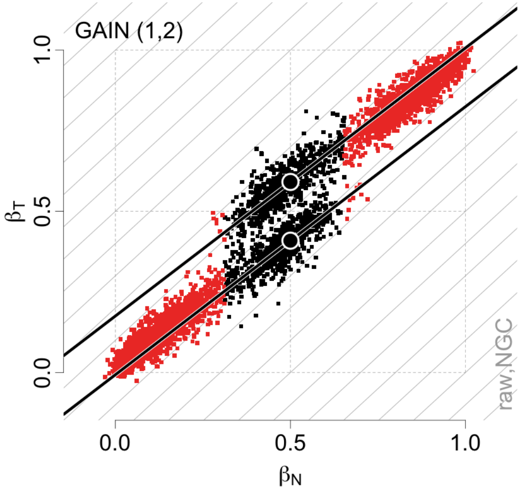


- Instead of three points at (0,0), $(\frac{1}{2}, \frac{1}{2})$ and (1,1), we have three clusters; the observed deviation is a *SNP effect*:

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

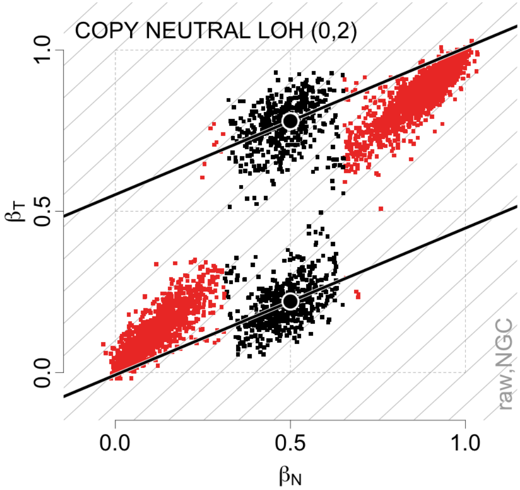
- δ is quite reproducible between the normal and the tumor

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 method

Idea

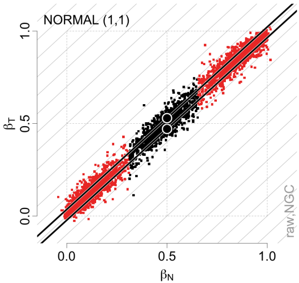
1. the SNP effect is reproducible between tumor and normal
2. in the normal the truth is easier to infer because we only expect three true allele B fractions, corresponding to genotypes AA, AB, BB.

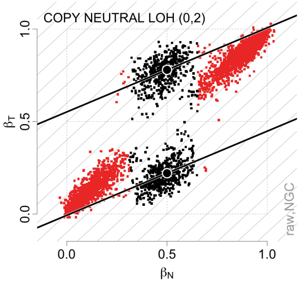
⇒ For each SNP, we estimate the SNP effect in the normal hybridization, and "subtract" it from the tumor.

Remarks

- we don't need to know copy number regions in advance !
- this is done for each SNP separately
- it only requires one tumor/normal pair

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

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

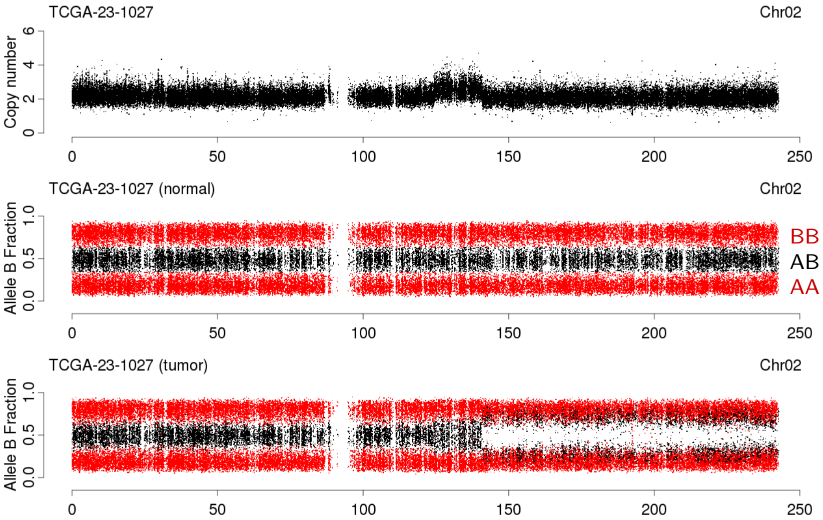
For heterozygous SNPs ($\hat{\mu}_{Nj} 1/2$):

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

Outline

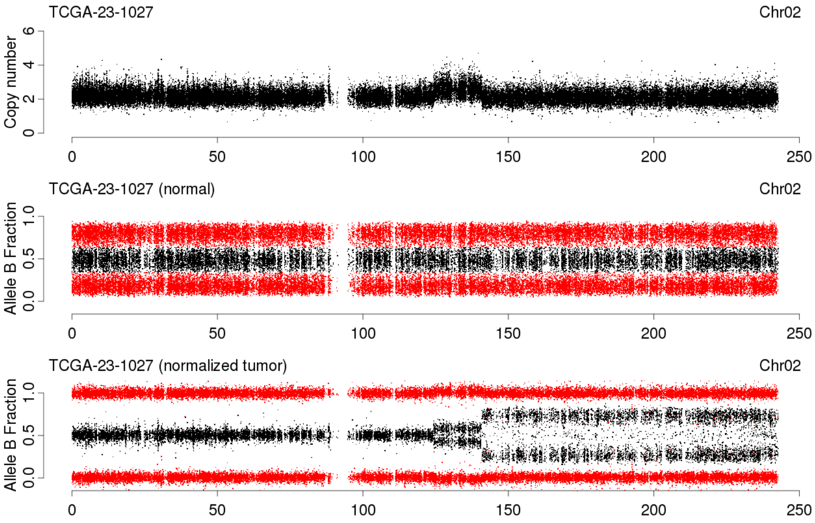
1. Genotyping microarrays in cancer research
   - DNA copy number analyses for cancer research
   - Genotyping microarray data
2. Normalizing each SNP of a single tumor/normal pair
   - SNP effects
   - Proposed normalization
3. Improved power to detect CN changes
   - Allele B fractions along the genome
   - Comparing before and after
   - ROC evaluation
4. Conclusions
   - All we need is a tumor/normal pair
   - Further thoughts

Normal, gain, copy neutral LOH before normalization



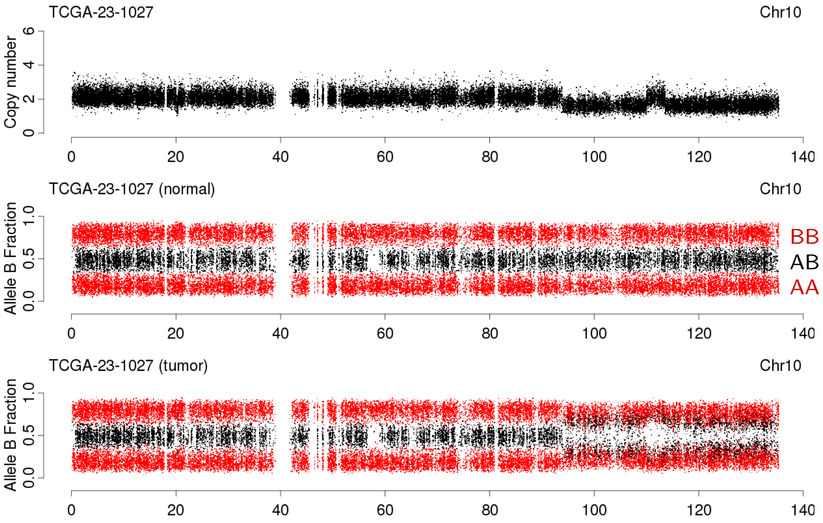
Homozygous SNPs in the normal sample are highlighted in red.

Normal, gain, copy neutral LOH after normalization



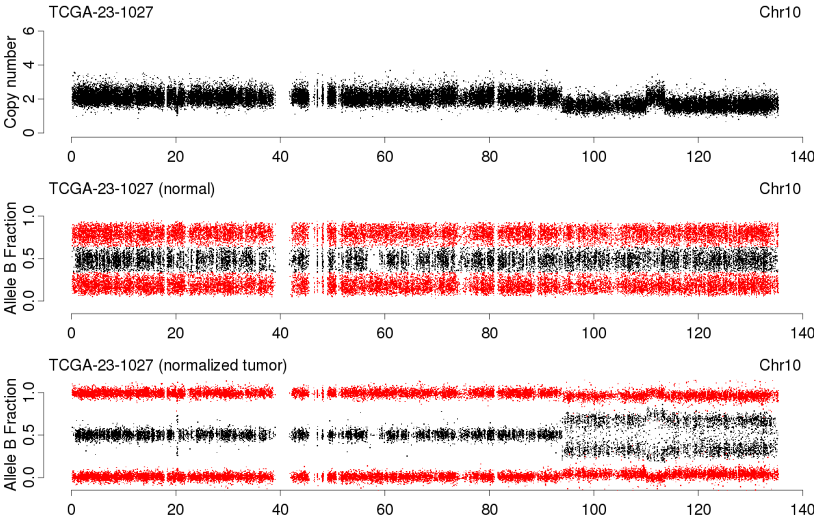
Homozygous SNPs in the normal sample are highlighted in red.

Normal, deletion, copy neutral LOH before normalization



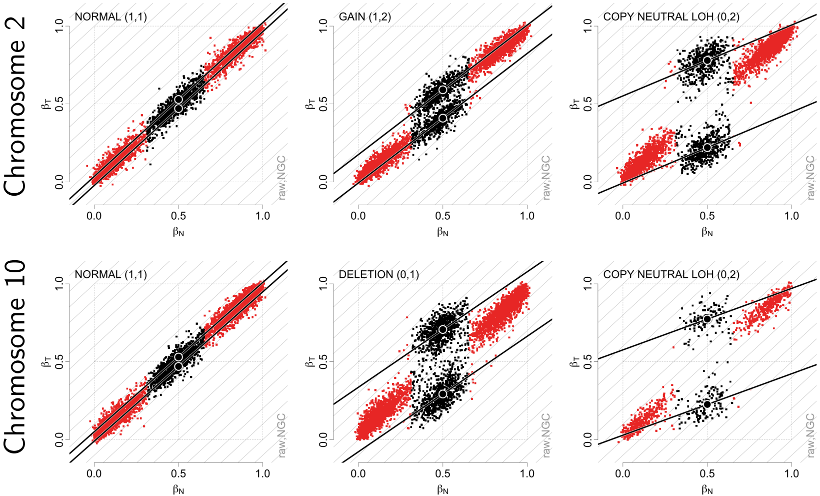
Homozygous SNPs in the normal sample are highlighted in red.

Normal, deletion, copy neutral LOH after normalization

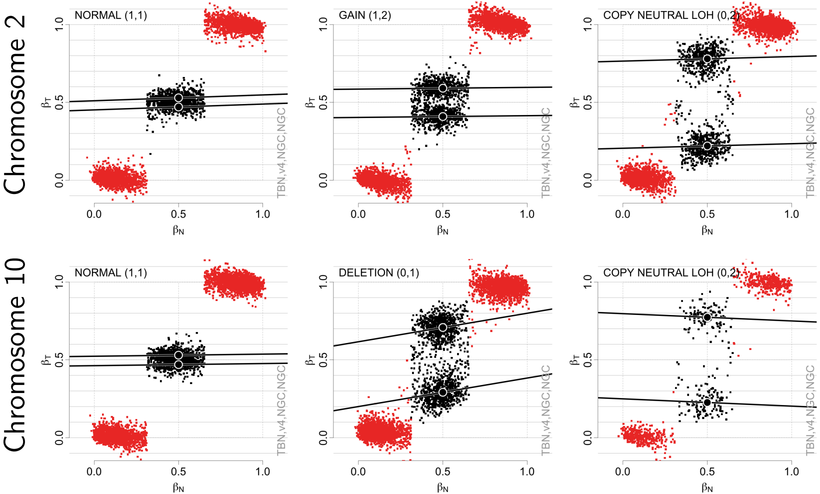


Homozygous SNPs in the normal sample are highlighted in red.

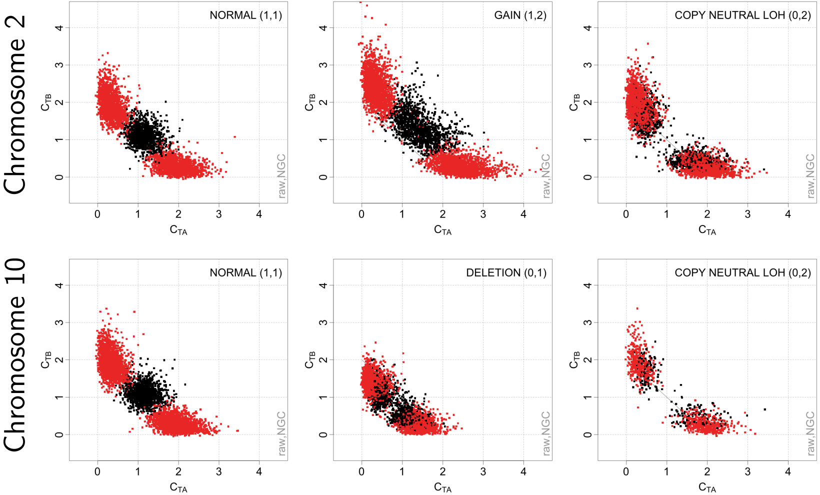
Allele B fractions before normalization



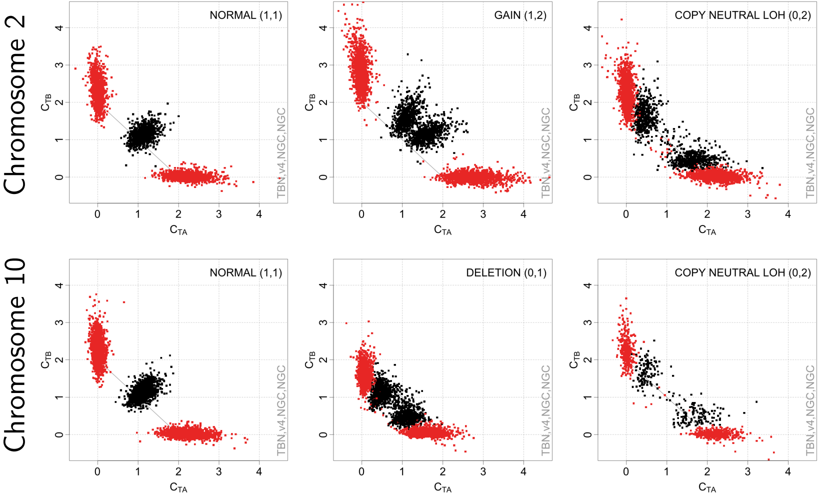
Allele B fractions after normalization



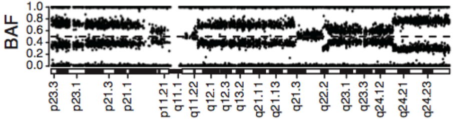
ASCNs before normalization



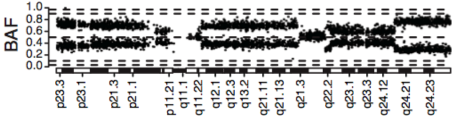
ASCNs after normalization



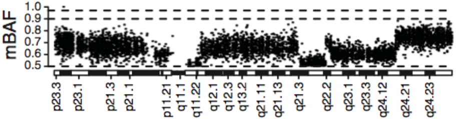
Detecting changes in allele B fractions



allele B fractions: β



allele B fractions for heterozygous SNPs



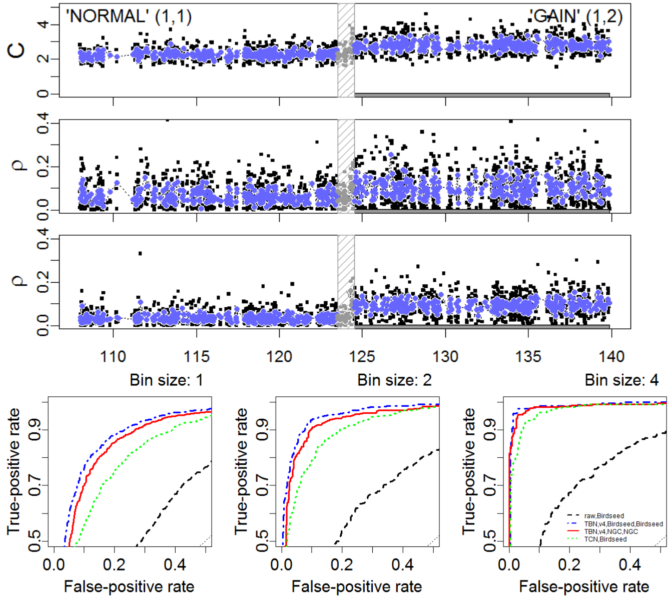
"mirrored" allele B fractions for heterozygous SNPs:

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

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

We use ROC analysis to assess how well two regions on each side of a known change point in ρ *separate*.

ROC analysis: from (1,1) to (1,2)



Outline

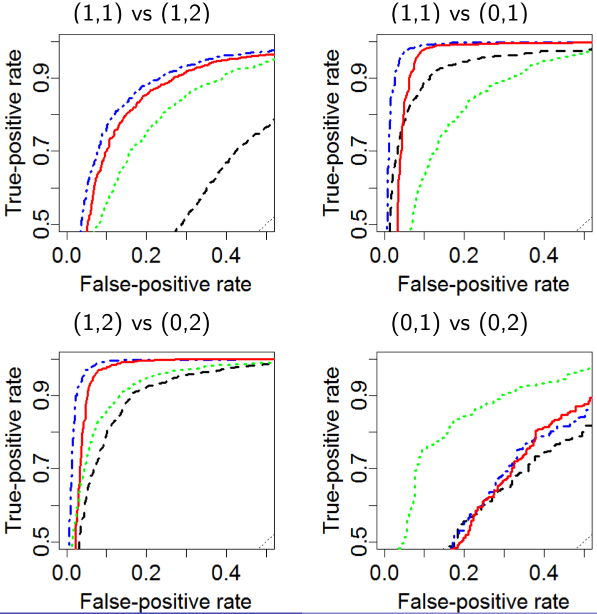
1. Genotyping microarrays in cancer research
   - DNA copy number analyses for cancer research
   - Genotyping microarray data
2. Normalizing each SNP of a single tumor/normal pair
   - SNP effects
   - Proposed normalization
3. Improved power to detect CN changes
   - Allele B fractions along the genome
   - Comparing before and after
   - ROC evaluation
4. Conclusions
   - All we need is a tumor/normal pair
   - Further thoughts

Complete preprocessing for a single tumor/normal pair

- 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: thresholding the density of β
- TumorBoost normalization

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

Observed power to detect changes


(1,1) vs (1,2)

(1,1) vs (0,1)

(1,2) vs (0,2)

(0,1) vs (0,2)

Legend:

Total copy number

Raw allele B fractions

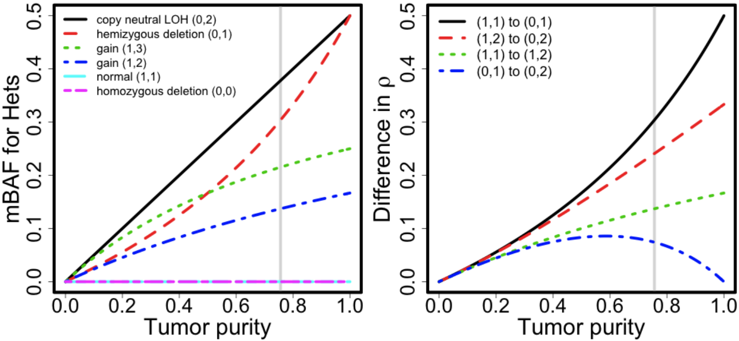
Normalized β (naive)

Normalized β (Birdseed)

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

Expected power to detect changes

CN varies from one unit in all change points just shown
For ASCN it's more complicated:



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

Yet to be solved

- When a matched normal is not available
- Two-dimensional segmentation methods
- Estimation of tumor purity
- Estimation of (unphased) parental CNs: (C_1, C_2)
- Integration of ASCN estimates from two different platforms (Affy and Illumina)

Thanks

- Henrik Bengtsson
- Terry Speed