An integrated Bayesian model for genotyping and copy number data

Paola M.V. Rancoita[1,2], M. Hutter[3,4], F. Bertoni[2], I. Kwee[1,2]

[1]IDSIA, Manno, Switzerland,
[2]IOSI, Bellinzona, Switzerland,
[3]ANU, Canberra, Australia,
[4]NICTA, Canberra, Australia

Summary

- We derive a new method for the joint estimation of CN events and IBD/UPD regions
- It takes into account all errors in the microarray genotyping measurements, due to CN aberrations
- The goodness of our model is supported by the results on real data

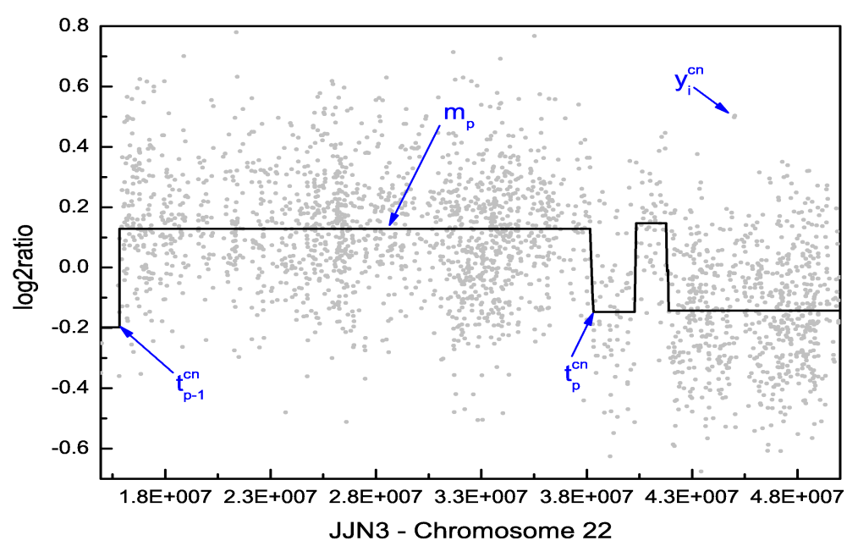
Genotyping and copy number data

- **Single nucleotide polymorphism** (**SNPs**) = single base-pair location in the genome where the nucleotide can assume two possible values among the four bases (T, A, C, G)
- We have two copies of each chromosome ⇒ at each SNP corresponds a pair of nucleotides:

$$\left.AB\right\}\ \textbf{Heterozygosity}\text{ or Het}\qquad \left.\begin{matrix}AA\\BB\end{matrix}\right\}\ \textbf{Homozygosity}\text{ or Hom}$$

where A and B are the two possible values of the SNP

- **DNA copy number** (**CN**) = for a given genomic region, is the number of copies of DNA of that region (normal CN = 2) ⇒ we can divide the genome in regions of constant CN, i.e. is a piecewise constant function of $\hat{k}^{cn}$ intervals with boundaries $\underline{t}^{cn}$=$(0 = \hat{t}^{cn}_0, \hat{t}^{cn}_1, \dots, t^{cn}_{k_0} = n)$ and levels of the segments $\underline{m} \in \mathbb{R}^{\hat{k}^{cn}}$ (usually a log$_2$ratio scale is used)



- Type of aberrations regarding genotyping and copy number data:
  - **amplification** (CN>4) ⇒ $\{Z = 2\}$
  - **gain** (CN=3,4) ⇒ $\{Z = 1\}$
  - **loss** (CN=1) ⇒ $\{Z = -1\}$
  - **homozygous deletion** (CN=0) ⇒ $\{Z = -2\}$
  - **Loss of heterozygosity** (**LOH**) with normal copy number, i.e. unusual long stretches of homozygous SNPs due to uniparental disomy or autozygosity (called **IBD/UPD regions**)

  where Z is the r.v. which represents the CN aberration occurred ($\{Z = 0\}$ is the normal CN)

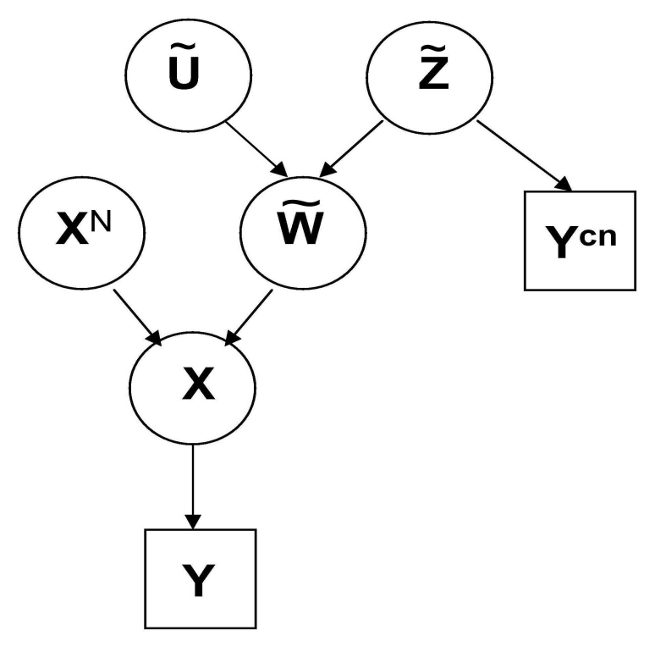
SNP microarray

- SNP microarrays are able to measure simultaneously genotyping and copy number data
- Microarray technology is not able to distinguish between the loss of one allele (e.g. A) or an Homozygosity (e.g. AA)

  ⇒ Integration of the two types of data to better identifies the aberrations (e.g. it possible to distinguish between IBD/UPD and loss or between gain and high amplification)

  ⇒ Bayesian regression to estimate the piecewise constant profile of the aberrations $\underline{\widetilde{W}}$=$(\widetilde{W}_1, \dots, \widetilde{W}_n)$ at n SNP loci. The profile consists of k_0 intervals, with boundaries $0 = t^0_0 < t^0_1 < \dots < t^0_{k_0-1} < t^0_{k_0} = n$, so that $\widetilde{W}_{t^0_{p-1}+1} = \dots = \widetilde{W}_{t^0_p} =: W_p$, for all $p = 1, \dots, k_0$.

The model



$$\begin{aligned}
\underline{Y} &= \text{vector of the SNP genotypes detected by the microarray } (Y_i \in \mathbb{Y}),\\
&\quad \text{where } \mathbb{Y} = \{Het,\, NHet,\, NoCall\} \text{ and } NHet = \text{not Het}\\
\underline{X} &= \text{vector of the true SNP genotypes in cancer cells } (X_i \in \mathbb{X}),\\
&\quad \text{where } \mathbb{X} = \{Het,\, Hom\}\\
\underline{X}^N &= \text{vector of the true SNP genotypes in normal cells } (X^N_i \in \mathbb{X})\\
\underline{\widetilde{Z}} &= \text{vector of the CN aberrations}\\
\underline{\widetilde{U}} &= \text{vector of the occurrence of IBD/UPD}\\
\underline{Y}^{cn} &= \text{vector of the raw CN data}\\
&\Rightarrow \text{for each interval } p,\ \{W_p = w\} = \{Z_p = z,\, U_p = u\}
\end{aligned}$$

$\mathrm{P}(\widetilde{y}_i | \widetilde{w}_i, x^N_i)$ estimated on two public datasets (*Zhao et al. (2004)*, *Forconi et al. (2008)*)

The priors & the posterior

The priors are defined as following:

- $\mathrm{P}(X^N_i = Het)$ set on the basis of the microarray annotation file
- for $\mathrm{P}(\widetilde{U}_i = 1)$, we tried two values 0.001 and 0.0001, on the basis of the estimations obtained using the data in *Bacolod et al. (2008)* and *The International HapMap Consortium (2007)*
- the priors of K and $\underline{T}$ are similar to mBPCR (*Rancoita et al. (2009)*):

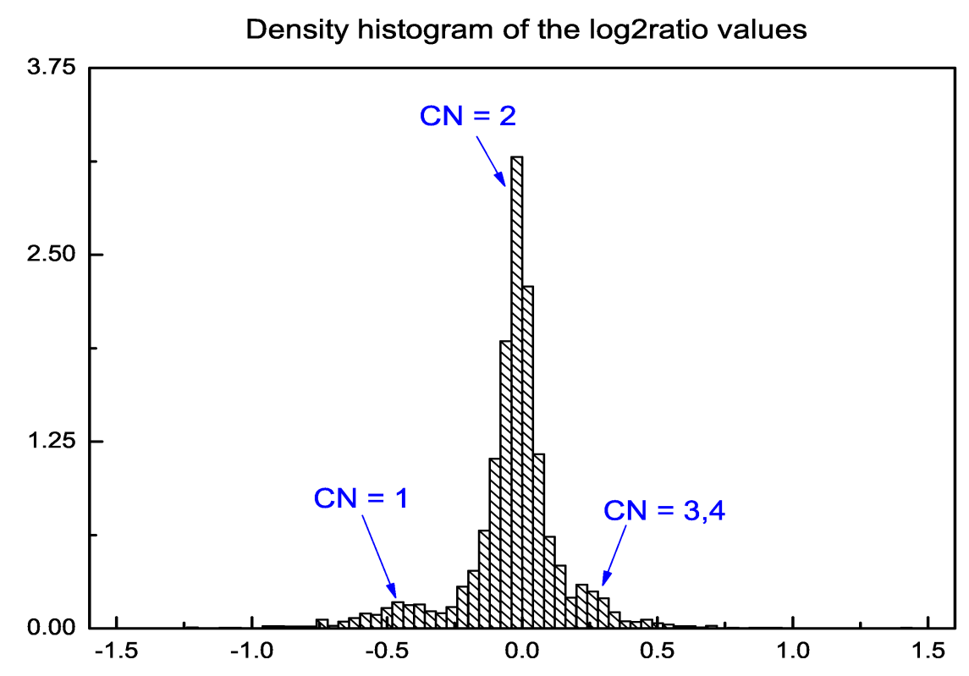
$$\begin{aligned}
P(\underline{T} = \underline{t} \,|\, K = k) &= \binom{n-1}{k-1}^{-1}, \qquad \underline{t} \in \mathbb{T}_{k,n}\\
P(K = k) &= \frac{k_{\max}+1}{k_{\max}}\frac{1}{k(k+1)}, \qquad k \in \mathbb{K} = \{1, \dots, k_{\max}\}
\end{aligned}$$

- $\mathrm{P}(Z_p = z)$ derived from the mBPCR estimated profile of CN data (we need to map the continuous log$_2$ratio values into the classes of CN aberrations):

$$\begin{aligned}
\mathrm{P}(Z_p = 2) &= \mathrm{P}\left(M_p \geq \hat{\mu}_4 + 3\hat{\sigma}_4 \,|\, cn\right)\\
\mathrm{P}(Z_p = 1) &= \mathrm{P}\left(\hat{\mu}_2 + 3\hat{\sigma}_2 < M_p \leq \hat{\mu}_4 + 3\hat{\sigma}_4 \,|\, cn\right)\\
\mathrm{P}(Z_p = 0) &= \mathrm{P}\left(\hat{\mu}_2 - 3\hat{\sigma}_2 < M_p \leq \hat{\mu}_2 + 3\hat{\sigma}_2 \,|\, cn\right)\\
\mathrm{P}(Z_p = -1) &= \mathrm{P}\left(\hat{\mu}_1 - 3\sigma_1 < M_p \leq \hat{\mu}_2 - 3\hat{\sigma}_2 \,|\, cn\right)\\
\mathrm{P}(Z_p = -2) &= \mathrm{P}\left(M_p \leq \hat{\mu}_1 - 3\hat{\sigma}_1 \,|\, cn\right),
\end{aligned}$$

where:

$$\begin{aligned}
cn &= \text{all the information regarding the copy number data}\\
M_p &= \text{CN value in the } p^{th} \text{ interval, } M_p \sim \mathcal{N}(\widehat{m}_p,\, \hat{V}_p)\\
(\widehat{m}_p,\, \hat{V}_p) &= \text{posterior mean and variance of } M_p \text{ estimated by mBPCR}\\
(\hat{\mu}_{cn},\, \hat{\sigma}^2_{cn}) &= \text{estimated mean and variance of the normal distribution}\\
&\quad \text{corresponding to } CN = cn
\end{aligned}$$



From the model, the posterior of $\underline{\widetilde{W}}$ is:

$$p(\underline{\widetilde{w}} \,|\, \underline{y},\, \underline{t}^0,\, k_0) \propto \prod_{p=1}^{k_0} \prod_{i=t^0_{p-1}+1}^{t^0_p} \sum_{x \in \mathbb{X}} p(y_i \,|\, X^N_i = x,\, w_p) \mathrm{P}(X^X_i = x) p(w_p),$$

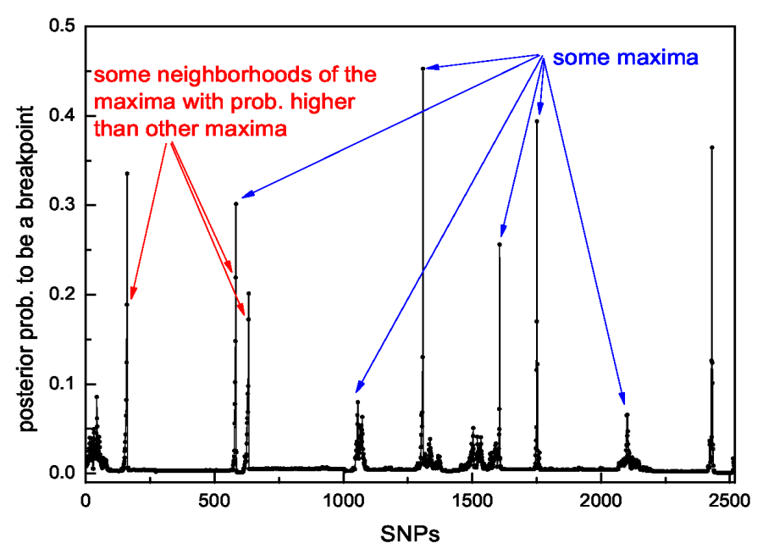
The estimation

- Method 1 (similar to mBPCR):

$$\begin{aligned}
\widehat{K}_{01} &= \arg\max_{k \in \mathbb{K}} p(k \,|\, \underline{Y},\, cn),\\
\widehat{\underline{T}}_{BinErrAk} &= \arg\max_{\underline{t}' \in \mathbb{T}_{\widehat{k},n}} \mathrm{E}\left[\sum_{q=1}^{\hat{k}-1}\sum_{p=1}^{k_0-1} \delta_{t'_q,\, t^0_p} \,\middle|\, \underline{Y},\, cn\right]\\
\widehat{W}_p &= \arg\max_{w} \mathrm{P}(W_p = w \,|\, \underline{Y},\, \hat{\underline{t}},\, \hat{k},\, cn), \qquad p = 1, \dots, \hat{k}
\end{aligned}$$

- $\widehat{\underline{T}}_{BinErrAk}$ consists of the $\hat{k}_{01}$ positions which have the highest posterior probability to be a breakpoint p_i

  ⇒ problem: we could take some points in the neighborhood of the higher maxima of $\underline{p}$ and not some maxima with a lower probability



- Method 2: estimate the number of the segments and the breakpoints with, respectively, the number of peaks and the locations of their maxima ($\underline{W}$ estimated as previously)
- It uses two thresholds: one for the determination of the peaks (thr_1) and one for the definition of the values close to zero (thr_2).

  ⇒ corresponding estimators $\widehat{K}_{Peaks,thr_1,thr_2}$ and $\widehat{\underline{T}}_{Peaks,thr_1,thr_2}$ (the method is denoted with $(thr_1,\, thr_2)$)
- Paired thresholds selected on the basis of results obtained on simulations: $(01, 01)$, $(mad, 01)$, $(01, mad)$, where

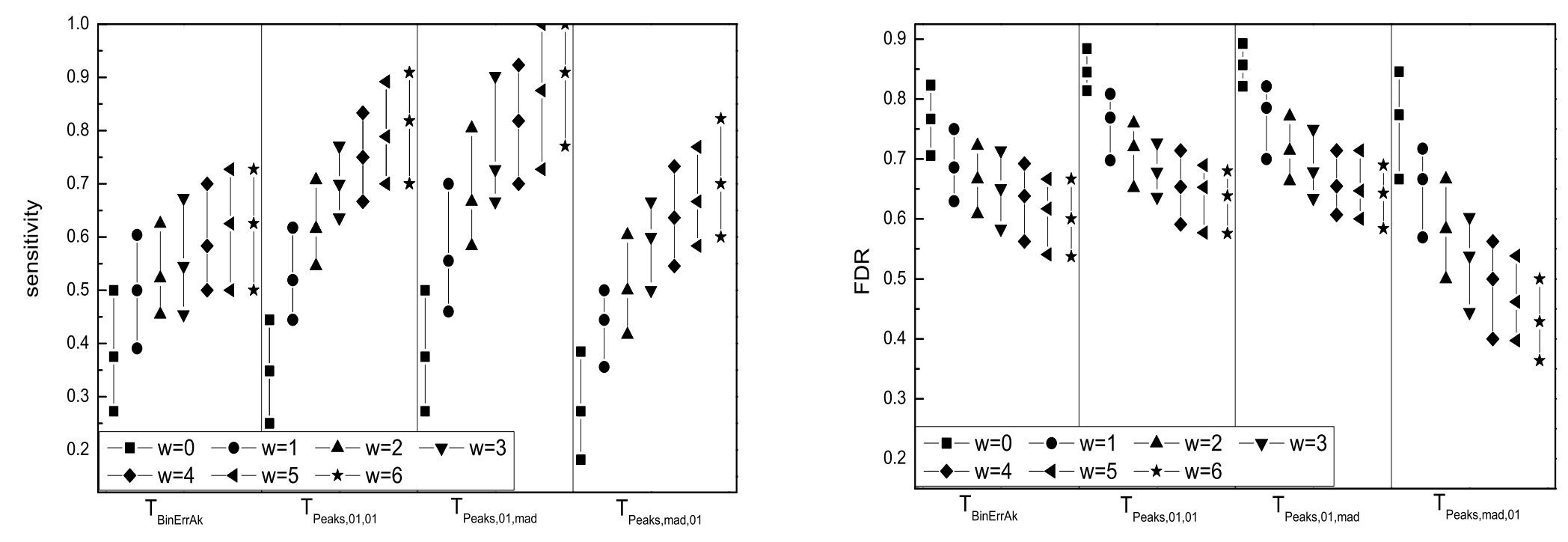
$$\begin{aligned}
01 &= \max(0.01, \text{quantile of } \underline{p} \text{ at } 0.95)\\
mad &= median(\underline{p}) + 3 * mad(\underline{p})
\end{aligned}$$

Some results on simulations

- Aberrations not considered in the simulations:
  - gain (because it does not influence the genotype detection)
  - IBD/UPD (difficult to simulate realistically)
- Simulated dataset (100 samples with fixed k_0 and $\underline{t}^0$): each sample is a raw profile coming from the prior definition of $\underline{X}^N$ given by the annotation file for the SNPs of chr. 22 in the Affymetrix GeneChip Mapping 250K Array ($n = 2520$) and the following prior definition of $\underline{Z}$ ($\mathrm{P}(Z_p = z) =: q^z$)

	segment														
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV
q^1	0	0.1	0	0.1	0.5	0.1	0	0	0.1	0.5	0	0.1	0.5	0.1	0
q^0	0.1	0.6	0.1	0.6	0.4	0.6	0.1	0.1	0.6	0.4	0.1	0.6	0.4	0.6	0.1
q^{-1}	0.6	0.3	0.6	0.3	0.1	0.3	0.6	0.4	0.3	0.1	0.6	0.3	0.1	0.3	0.6
q^{-2}	0.3	0	0.3	0	0	0	0.3	0.5	0	0	0.3	0	0	0	0.3

- Some results on breakpoint estimation:



⇒ Method 2 has higher sensitivity and similar or lower FDR.

- Some results CN aberration estimation (- best result, - worst result):

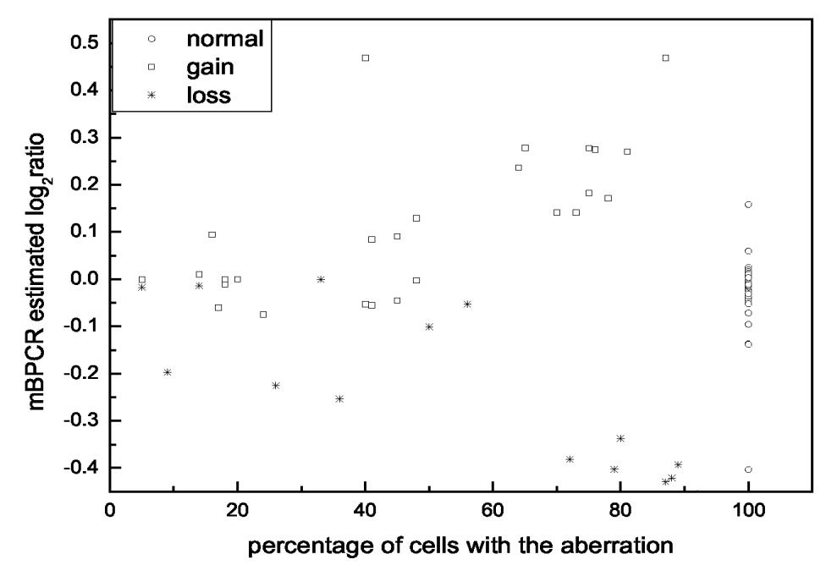
method	sum 0-1 err	SSQ	$\sqrt{SSQ/n}$
method 1	421.79	1226.59	0.70
$(01,\, 01)$	109.39	286.15	0.34
$(01,\, mad)$	109.39	286.15	0.34
$(mad,\, 01)$	111.75	283.77	0.34

	sensitivity				FDR			
method	Z=2	Z=0	Z=-1	Z=-2	Z=2	Z=0	Z=-1	Z=-2
method 1	0.681	0.932	0.968	0.555	0.017	0.047	0.306	0.025
$(01,\, 01)$	0.896	0.983	0.961	0.946	0.043	0.031	0.068	0.020
$(01,\, mad)$	0.896	0.983	0.961	0.946	0.043	0.031	0.068	0.020
$(mad,\, 01)$	0.889	0.984	0.963	0.942	0.038	0.026	0.075	0.023

⇒ Method 2 best estimates the profile
(best paired threshold: $(01,\, 01)$, $(01,\, mad)$).

Applications on real data

- Data: paired samples of patients affected by chronic lymphocytic leukemia (CLL), which then transformed in diffuse large B-cell lymphoma (DLBCL) (*Bertoni et al. (2008)*). Of two patients, we had three samples.
- **detectable CN aberrations** = the ones born by at least 60% of cells in the sample



- Evaluation of the estimation of the CN aberrations: comparison with the estimated CN of some genomic regions with FISH (fluorescent in situ hybridization), which gives also the percentage of cells bearing the aberration
  - 15/17 detectable aberrations found by all estimators
  - 3/26 not detectable aberrations found by all estimators and another by $(01,\, 01)$ and $(01,\, mad)$ with $p_{upd} = 10^{-3}$ and $(mad,\, 01)$ with $p_{upd} = 10^{-4}$
  - in only 2/90 normal segments, all estimators discovered an aberration
  - Remark: a slightly discordance between the 2 techniques is possible, because the samples used are not exactly the same
- Evaluation of the IBD/UPD region detection: comparison of the regions found in the 3 samples of 2 patients

Patient 1:						
	$p_{upd} = 10^{-4}$			$p_{upd} = 10^{-3}$		
types of regions	$01, 01$	$01, mad$	$mad, 01$	$01, 01$	$01, mad$	$mad, 01$
distinct (total)	413	413	414	494	492	519
equal (%)	0.79	0.79	0.78	0.78	0.78	0.77
equal in 2 samples (%)	0.15	0.15	0.20	0.15	0.15	0.18
overlapping (%)	0.03	0.03	0.01	0.02	0.02	0.03
validated (%)	**0.98**	**0.98**	**0.98**	**0.95**	**0.95**	**0.98**
remaining (%)	0.02	0.02	0.02	0.05	0.05	0.02
% of remaining < 1Mb	0.80	0.80	0.88	0.93	0.92	1.00
Patient 2:						
distinct (total)	441	441	454	580	580	618
equal (%)	0.21	0.21	0.25	0.19	0.19	0.24
equal in 2 samples (%)	0.02	0.02	0.03	0.03	0.03	0.02
overlapping (%)	0.50	0.50	0.47	0.51	0.51	0.50
validated (%)	**0.73**	**0.73**	**0.74**	**0.74**	**0.74**	**0.76**
remaining (%)	0.27	0.27	0.26	0.26	0.26	0.24
% of remaining < 1Mb	0.88	0.88	0.89	0.91	0.91	0.93

⇒ the 3 estimators behaved similarly and equally well on real data

Summary and conclusions

- Our method is a new algorithm for the joint estimation of CN events and IBD/UPD regions, which takes into account the errors in the genotyping measurements of microarrays, due to the aberrations affecting the CN.
- Differently from the only other method present in literature (i.e., Scharpf et al. (2008)), it considers all the CN events biologically relevant.
- The goodness of our model is supported by the results obtained on simulated and real data.