Bayesian joint estimation of CN and LOH aberrations

IWPACBB09 - Salamanca, June 10-12 2009

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

paola@idsia.ch

[1] Dalle Molle Institute for Artificial Intelligence (IDSIA), Manno-Lugano, Switzerland

[2] Laboratory of Experimental Oncology, IOSI, Bellinzona, Switzerland

[3] Dipartimento di Matematica, Università degli Studi di Milano, Milano, Italy

[4] RSISE @ ANU and SML @ NICTA, Canberra, ACT, 0200, Australia

GENOTYPING 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:

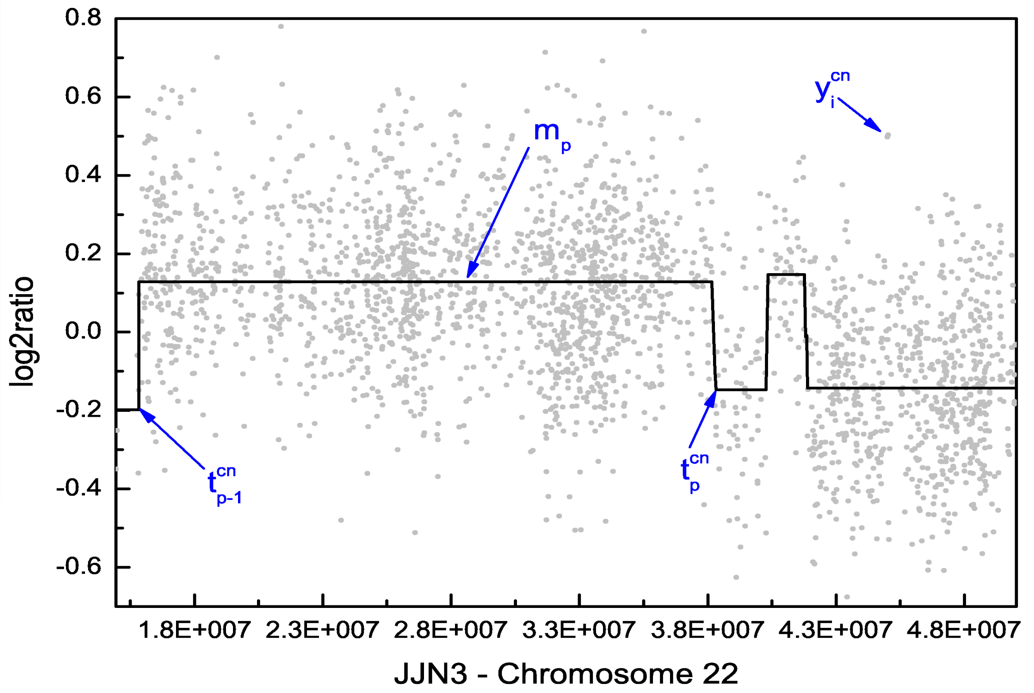
$$\left.\begin{array}{c} AB \end{array}\right\} \quad \textbf{Heterozygosity} \text{ or Het}$$

$$\left.\begin{array}{c} AA \\ BB \end{array}\right\} \quad \textbf{Homozygosity} \text{ or Hom}$$

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

COPY NUMBER DATA

- **DNA copy number** (**CN**) = for a given genomic region, is the number of copies of DNA of that region (normal CN = 2) $\Rightarrow$ we can divide the genome in regions of constant CN (usually a $\log_2$ratio scale is used)



DNA ABERRATIONS

- Type of aberrations regarding genotyping and copy number data:
  - **amplification** (CN>4) $\Rightarrow \{Z = 2\}$
  - **gain** (CN=3,4) $\Rightarrow \{Z = 1\}$
  - **loss** (CN=1) $\Rightarrow \{Z = -1\}$
  - **homozygous deletion** (CN=0) $\Rightarrow \{Z = -2\}$
  - **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)

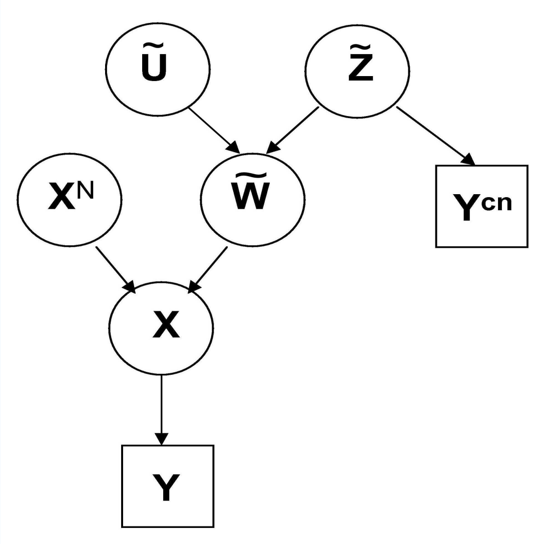
GOAL

- 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)

$\Rightarrow$ 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)

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

THE MODEL



$$\begin{aligned}
\mathbf{Y} &= \text{genotypes detected by the microarray} \\
&\quad (Y_i \in \mathbb{Y} = \{Het,\, NHet,\, NoCall\}) \\
\mathbf{X} &= \text{true genotypes in cancer cells} \\
&\quad (X_i \in \mathbb{X} = \{Het,\, Hom\}) \\
\mathbf{X}^N &= \text{true genotypes in normal cells} \\
&\quad (X_i^N \in \mathbb{X}) \\
\widetilde{\mathbf{W}} &= \text{genotyping \& CN aberrations} \\
\widetilde{\mathbf{Z}} &= \text{CN aberrations} \\
\widetilde{\mathrm{U}} &= \text{occurrence of IBD/UPD} \\
\mathbf{Y}^{cn} &= \text{raw CN data} \\
&\Rightarrow \text{for each interval } p, \\
&\quad \{W_p = w\} = \{Z_p = z,\, U_p = u\}
\end{aligned}$$

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

DEFINITION OF THE PRIORS (1)

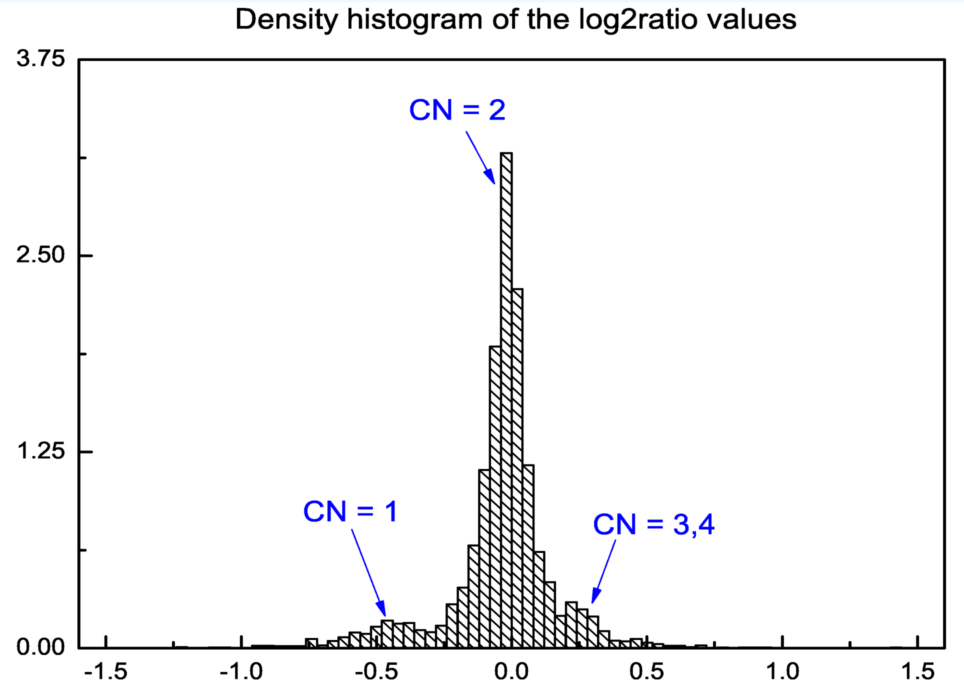
- $\mathrm{P}(X_i^N = Het)$ on the basis of the microarray annotation file
- for $\mathrm{P}(\widetilde{U}_i = 1)$, we tried two values 0.001 and 0.0001, on the basis of the estimations obtained using the data in *Bacolod et al. (2008)* and *The International HapMap Consortium (2007)*
- the priors of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

$$\begin{aligned} P(\mathbf{T} = \mathbf{t} \mid K = k) &= \text{uniform} \\ P(K = k) &\propto 1/k^2 \end{aligned}$$

DEFINITION OF THE PRIORS (2)

$\mathrm{P}(Z_p = z)$ derived from the mBPCR estimated profile:

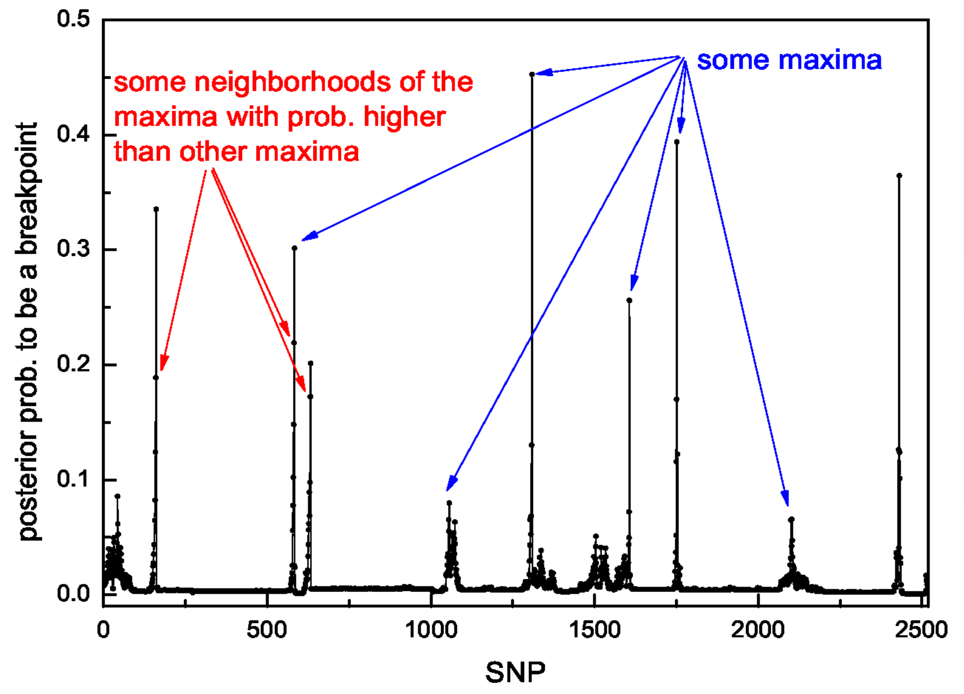
$$\begin{aligned}
\mathrm{P}(Z_p = 2) &= \mathrm{P}\left(\mu_p \geq \hat{\mu}_4 + 3\hat{\sigma}_4 \,|\, cn\right) \\
\mathrm{P}(Z_p = 1) &= \mathrm{P}\left(\hat{\mu}_2 + 3\hat{\sigma}_2 < \mu_p \leq \hat{\mu}_4 + 3\hat{\sigma}_4 \,|\, cn\right) \\
\mathrm{P}(Z_p = 0) &= \mathrm{P}\left(\hat{\mu}_2 -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



THE ESTIMATION: METHOD 1

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

$\widehat{\mathbf{T}}_{BinErrAk}$ consists of the $\hat{k}_{01}$ positions which have the highest posterior probability to be a breakpoint (p_i) $\Rightarrow$ possible problems



THE ESTIMATION: METHOD 2

- estimate the number of the segments and the breakpoints with, respectively, the number of peaks and the locations of their maxima ($\mathbf{W}$ 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).

  $\Rightarrow$ corresponding estimators $\widehat{K}_{Peaks,thr_1,thr_2}$ and $\widehat{\mathbf{T}}_{Peaks,thr_1,thr_2}$ (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

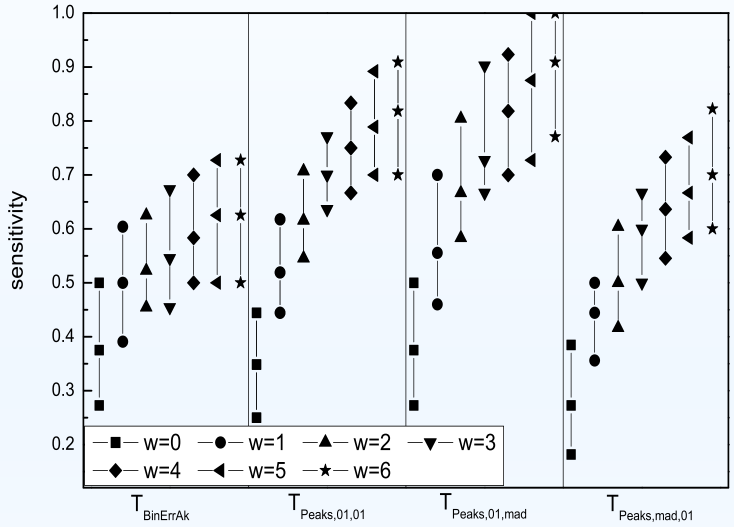
$$\begin{aligned} 01 &= \max(0.01, \text{quantile of } \mathbf{p} \text{ at } 0.95) \\ mad &= median(\mathbf{p}) + 3 * mad(\mathbf{p}) \end{aligned}$$

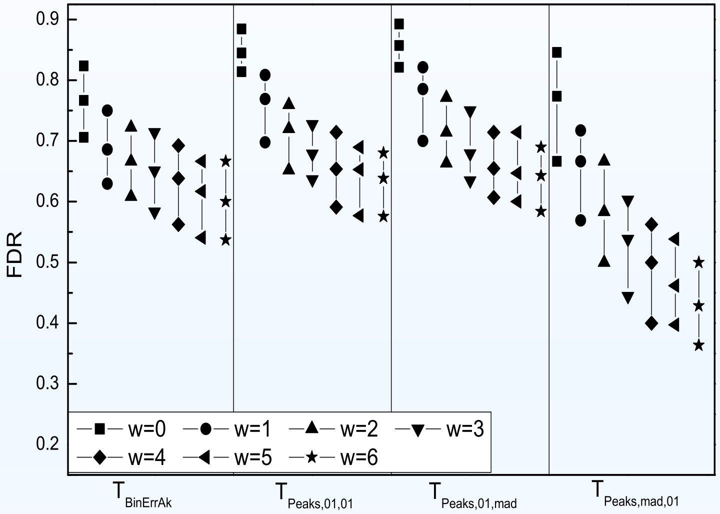
SIMULATIONS: DESCRIPTION

- 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 $\mathbf{t}^0$): each sample is a raw profile coming from the prior definition of $\mathbf{X}^N$ 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 $\mathbf{Z}$ ($\mathrm{P}(Z_p = z) =: q^z$)

	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

SIMULATIONS: BREAKPOINT ESTIMATION





$\Rightarrow$ Method 2 has higher sensitivity and similar or lower FDR.

SIMULATIONS: CN ABERRATION DETECTION

- best result, - worst result

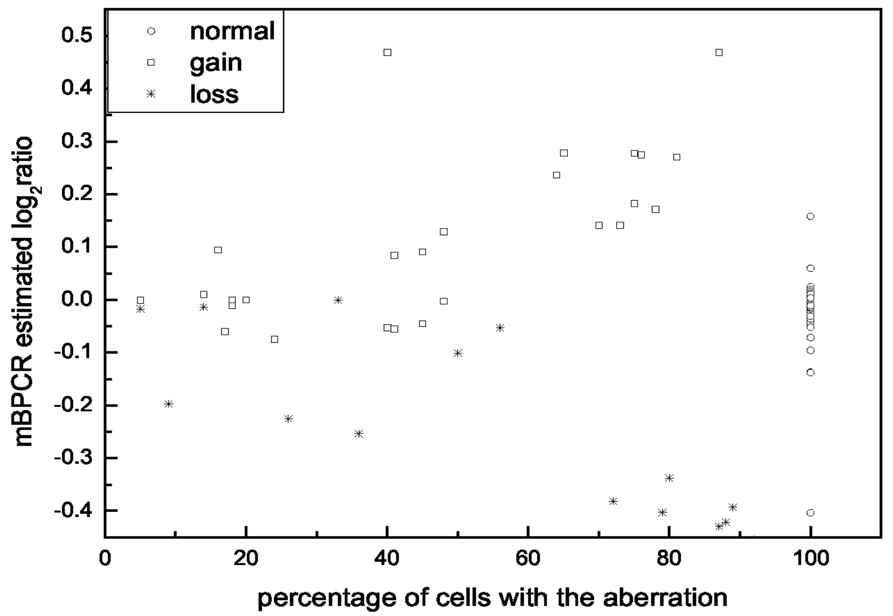
method	sum 0-1 err	SSQ	$\sqrt{SSQ/n}$
method 1	421.79	1226.59	0.70
$(01,\, 01)$	109.39	286.15	0.34
$(01,\, mad)$	109.39	286.15	0.34
$(mad,\, 01)$	111.75	283.77	0.34

	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

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

APPLICATION TO 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



ESTIMATION OF 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, equal to the one found in the same region of the paired sample
- simply using the prior thresholds, we detected 3 more aberrations, but 4 normal regions were seen as aberrations
- Remark: a slight discordance with FISH measurements is possible, because the samples used are not exactly the same

IBD/UPD 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

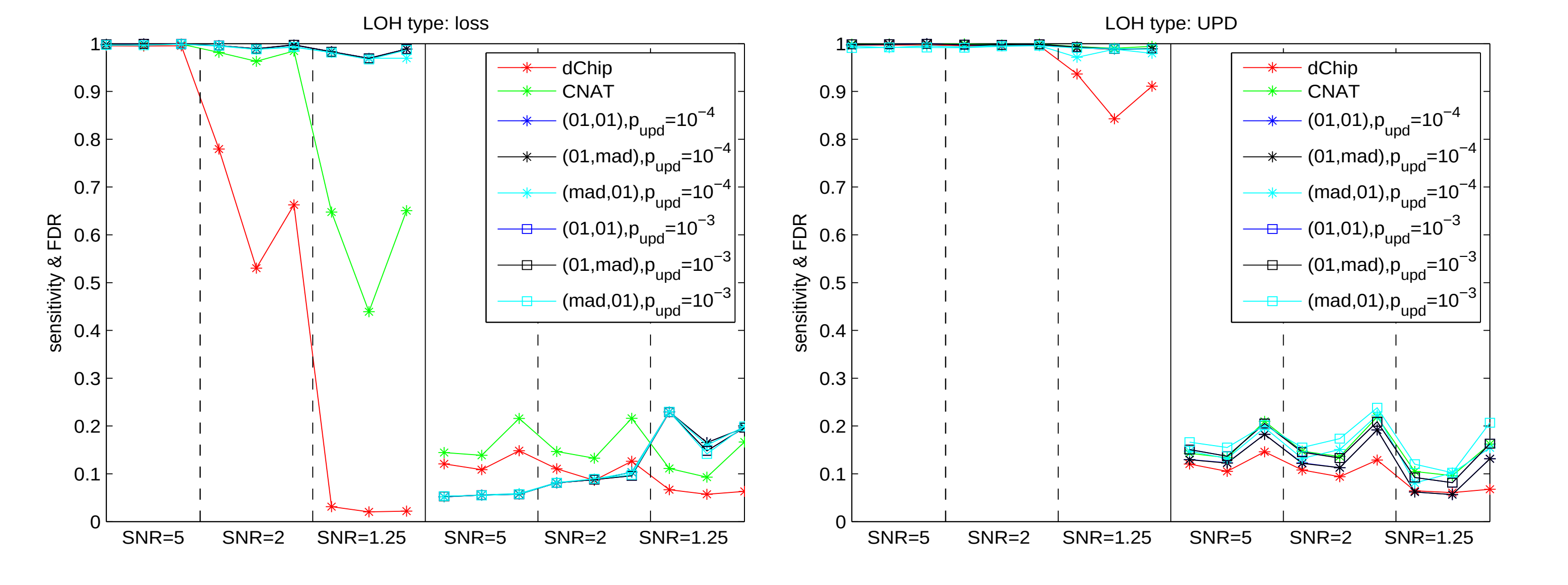
$\Rightarrow$ The 3 estimators behaved similarly and equally well on real data

SUMMARY & 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.
- All the proposed final versions of the method behave similarly.

ONGOING WORK

- Since the parameters related to the $NoCall$ detection depend on the noise of the sample, we are finding a solution to adjusting them in dependency to the noise.
- We are making comparisons among our method and two well-known methods for LOH estimation: dChip and CNAT. For example (artificial data from *Wu et al. (2009)*):



THANKS TO:

- M. Hutter

  RSISE @ ANU and SML @ NICTA, Canberra, ACT, 0200, Australia

- F. Bertoni

  Laboratory of Experimental Oncology, IOSI, Bellinzona, Switzerland

- I. Kwee

  Laboratory of Experimental Oncology, IOSI, Bellinzona, Switzerland
  IDSIA, Manno-Lugano, Switzerland

THANK YOU
FOR YOUR ATTENTION!