Bayesian Joint Estimation of CN and LOH Aberrations*

Paola M.V. Rancoita[1,2,3], Marcus Hutter[4], Francesco Bertoni[2], and Ivo Kwee[1,2]

[1] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Manno, Switzerland
paola@idsia.ch
[2] Laboratory of Experimental Oncology, Oncology Institute of Southern Switzerland, Bellinzona, Switzerland
[3] Dipartimento di Matematica, Università degli Studi di Milano, Milano, Italy
[4] RSISE, ANU and SML, NICTA, Canberra, ACT, 0200, Australia

Abstract. SNP-microarrays are able to measure simultaneously both copy number and genotype at several single nucleotide polymorphism positions. Combining the two data, it is possible to better identify genomic aberrations. For this purpose, we propose a Bayesian piecewise constant regression which infers the type of aberration occurred, taking into account all the possible influence in the microarray detection of the genotype, resulting from an altered copy number level. Namely, we model the distributions of the detected genotype given a specific genomic alteration and we estimate the hyper-parameters used on public reference datasets.

Keywords: Bayesian regression, piecewise constant function, change point problem, DNA copy number estimation, LOH estimation.

1 Introduction

Single nucleotide polymorphisms (SNPs) are single base-pair locations in the genome where the nucleotide can assume two possible values among the four bases (thymine, adenine, cytosine, guanine). Since we have two copies of each chromosome, at each SNP corresponds a pair of nucleotides (called **alleles**), whose genotype can be AA, BB or AB, denoting with A and B the two possible values that the nucleotide can assume. We can divide the genotypes into two classes: **Heterozygosity** or **Het** (i.e., AB) and **Homozygosity** or **Hom** (i.e., AA or BB).

Several diseases are due changes in genotype and/or in DNA copy number, CN (i.e. number of copies of DNA, which is normally two). The former aberrations are often displayed by unusual long stretches of homozygous SNPs, called LOH regions (i.e. loss of heterozygosity). The latter aberrations can be divided as high amplification (CN>4), gain (CN=3, 4), loss (CN=1) and homozygous deletion (CN=0). We call these classes **copy number events**.

* Research supported by Swiss National Science Foundation grant 205321-112430; Oncosuisse grants OCS-1939-8-2006 and OCS - 02296-08-2008; Cantone Ticino Ticino in rete grant; Fondazione per la Ricerca e la Cura sui Linfomi (Lugano, Switzerland).

S. Omatu et al. (Eds.): IWANN 2009, Part II, LNCS 5518, pp. 1109–1117, 2009.

SNP-based microarrays are able to measure simultaneously both the DNA CN and the genotype at each SNP position. In this way, we can observe the "abnormalities" of the genome regarding both CN and genotype, and combine them for a better identification of the events occurred. For example, at a deletion of one copy of a chromosomal segment, we usually detect a long stretch of homozygous SNPs (since the microarray is unable to distinguish between the presence of only one allele and the presence of two equal alleles), but, in general, the same genotype can also occur for other reasons, such as uniparental disomy (when two equal alleles are inherited from the same parent). In this situation, the knowledge of both types of data can lead to the correct interpretation of the phenomenon. Another example is when an amplified genomic segment is present: if one of the two copies of the segment is high amplified, then, even if there are heterozygous SNPs, all SNPs of the region will be likely detected as homozygous, because the DNA quantity of one allele is much higher than the other one. In this case, the integration of both types of data may better identify the dosage of the DNA aberration.

In literature, many methods have been developed for the estimation of the CN profile (such as [9], [10] and references therein) and others for the discovery of LOH regions (such as [3], [8]). Nevertheless, only one method has been developed for the integration of these two types of data and it uses an HMM model [12].

Here, we propose a method which estimates simultaneously the CN event profile and the abnormal stretches of homozygous SNPs, using both genotyping and CN data. Our model appears more complete than the one in [12], since the latter does not distinguish regions with high amplification of DNA from gained ones, and regions with the deletion of both alleles from those with the loss of only one allele.

2 Biological Model

In order to integrate the information given by CN and genotyping data, we need to clarify the random variables involved in the model and the relationship among them.

Usually, CN data are used in a $\log_2$ratio scale, where the ratio is defined with respect to a normal reference dataset. Therefore, the estimation procedures commonly estimate the CN profile as a piecewise constant function (i.e. the genome is divided in regions of constant CN), where the levels assume real values. For the purpose of our model, we estimate this profile by mBPCR [10]. The estimated profile consists of $\hat{k}^{cn}$ intervals with boundaries $\hat{\mathbf{t}}^{cn}=(0 = \hat{t}_0^{cn}, \hat{t}_1^{cn}, \ldots, \hat{t}_{k_0}^{cn} = n)$ and levels of the segments $\widehat{\mathbf{m}} \in \mathbb{R}^{\hat{k}^{cn}}$. Let us call $\widetilde{Z}_i$ the random variable which represents a CN event at SNP i. It assumes values: -2 (homozygous deletion), -1 (loss), 0 (normal), 1 (gain), 2 (high amplification). Since we estimate the CN profile with a piecewise constant function, we can also divide the genome in regions of constant CN event. We denote with Z_p the copy number event of the p^{th} interval.

In the past, long stretches of homozygous SNPs without copy number change (copy-neutral LOH) were considered as a consequence of an uniparental disomy event (UPD). Recently, long homozygous segments have also been detected

in genomes of normal individuals within the same population, supporting the hypothesis that some LOH regions might represent autozygosity (e.g. [6]). A relationship between some tumors and both types of aberrant events has been shown (e.g. [1], [2]). In normal situation, there are two copies of each chromosome (apart from the sex ones), called **homologues**, and each of them is inherited from a different parent. UPD occurs when both homologues of a part of a chromosome are inherited from only one parent. The UPD event can happen during the meiosis or mitosis and, in cancer cells, it can occur when an homologue of a part of a chromosome is lost and the remaining is duplicated. Instead, the autozygosity describes a situation where the alleles are identical by descendent (IBD), because they are inherited from a common ancestor. Therefore, IBD and UPD can be detected because they appear as a long sequence of homozygous SNPs with normal CN and with a low probability to occur. We define $\widetilde{U}_i$ as the random variable which represents the presence of IBD/UPD at SNP i (this event can occur only if $\widetilde{Z}_i$=0). We define p_{upd}=P($\widetilde{U}_i = 1$), for $i = 1, \ldots, n$, i.e. the probability of IBD/UPD at any SNP.

We denote the vector of the aberration events at n SNP loci with $\widetilde{\mathbf{W}}$=($\widetilde{W}_1$, $\ldots, \widetilde{W}_n$). Each component i of the vector assumes values: -3 ($\widetilde{Z}_i$=0 and $\widetilde{U}_i$=1), -2 ($\widetilde{Z}_i$=-2), -1 ($\widetilde{Z}_i$=-1), 0 ($\widetilde{Z}_i$=0 and $\widetilde{U}_i$=0), 1 ($\widetilde{Z}_i$=1), 2 ($\widetilde{Z}_i$=2). From the previous discussion, we can divide the genome in intervals corresponding to the same aberration event, i.e the profile of the aberrations consists of k_0 intervals, with boundaries $0 = t_0^0 < t_1^0 < \ldots < t_{k_0-1}^0 < t_{k_0}^0 = n$, so that $\widetilde{W}_{t_{p-1}^0+1} = \ldots = \widetilde{W}_{t_p^0} =: W_p$, for all $p = 1, \ldots, k_0$.

SNP microarray technology is unable to distinguish among a homozygosity due to the presence of two equal alleles or the one due to the loss or high amplification of one allele. Hence, the presence of heterozygosity can ensure that the CN is normal or gained with a high probability (we assume no difference in the genotyping detection in presence of normal or gained CN), while the homozygosity can be due to different events. Moreover, in case of homozygous deletion (i.e. deletion of both alleles), the microarray should detect a "NoCall" at the corresponding SNP positions, but this is very rare analyzing clinical samples, because, for example, the DNA sample often contains also a percentage of DNA of normal cells. Nevertheless, in some cases the information given by the "NoCall" genotypes can be useful to distinguish between the loss of one allele or a homozygous deletion. Therefore, three different genotyping data are present in the biological model: the vector of true genotype in normal cells ($\mathbf{X}^N$), the one of true genotype in "cancer" cells ($\mathbf{X}$), which is the due to CN changes or IBD/UPD, and the vector of the genotypes detected by the microarray ($\mathbf{Y}$). The components of the first two random vectors can assume only values in $\mathbb{X}$={Het, Hom} and we assume that they are independently distributed as Bernoulli random variables. On the other hand, the components of $\mathbf{Y}$ can assume values in $\mathbb{Y}$={$NoCall$, Het, $NHet$ (i.e. Not Heterozygosity)}.

To model the distribution of $\mathbf{Y}$, we take into account all the variability that can affect it, such as PCR amplification, the presence of different cancer cell subpopulations or normal cells, copy number changes (in particular, homozygous

deletion, loss and high amplification). The polymerase chain reaction (PCR) amplification is a biological process used to amplify the sequences of DNA, before hybridizing them on the microarray. Given the true value of the genotype and the CN event at each position, we consider the genotyping data points $\mathbf{Y}$ as independent, since their values depend only on both noise and genotyping detection errors. Hence, for each component of $\mathbf{Y}$, we define $\mathrm{P}(Y_i = y \,|\, X_i^N = x,\, \widetilde{W}_i = w)$, for $y \in \mathbb{Y}$, $x \in \mathbb{X}$, w=-3, -2, -1, 0, 2. For example, the probability $\mathrm{P}(Y_i = Het \,|\, X_i^N = Het,\, \widetilde{W}_i = -2)$ takes into account the error of the genotyping detection due to the presence of different types of normal and/or cancer cell subpopulations or to PCR amplification, while $\mathrm{P}(Y_i = NHet \,|\, X_i^N = Het,\, \widetilde{W}_i = 2)$ considers the error due to the amplification of only one allele. Since we suppose no difference in the genotype detection given a normal or gained CN, $\mathrm{P}(Y_i = y \,|\, X_i^N = x,\, \widetilde{W}_i = 1)$=$\mathrm{P}(Y_i = y \,|\, X_i^N = x,\, \widetilde{W}_i = 0)$.

From the model, given k_0 and $\mathbf{t}^0$, the posterior distribution of $\widetilde{\mathbf{W}}$ is

$$p(\widetilde{\mathbf{w}} \,|\, \mathbf{y},\, \mathbf{t}^0,\, k_0) \propto \prod_{p=1}^{k_0} \prod_{i=t_{p-1}^0+1}^{t_p^0} \sum_{x \in \mathbb{X}} p(y_i \,|\, X_i^N = x,\, w_p) \mathrm{P}(X_i^N = x) p(w_p), \qquad (1)$$

where the prior of $\mathbf{W}$ derives from the ones of $\mathbf{Z}$ and $\mathbf{U}$.

Remaining Z Prior Definition. While the estimated levels of the $\log_2$ratio profile are continuous variables, Z classifies the CN events considering (as it is) the CN as a discrete variable. Then, the major problem in the definition of the prior for Z consists in mapping the continuous values of the levels into the discrete values of Z, i.e. in defining a partition of the $\log_2$ratio values such that each interval corresponds to a particular CN event.

Usually, the histogram of the estimated $\log_2$ratio values shows a multimodal density with peaks corresponding to $CN = 1$, $CN = 2$ and $CN = 3, 4$. Similarly to [7], we modeled it as a mixture of three normal distributions. Estimated the parameters of the density, we can map the $\log_2$ratio values into the copy number event classes, using the confidence interval around the peaks of the multimodal density. Therefore, for each $p = 1, \ldots, \hat{k}^{cn}$, we define the prior distribution of Z_p as:

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

where we denote with cn all the information regarding the copy number (both raw data and estimated profile with mBPCR), M_p is the random variable representing the copy number value in the p^{th} interval, and $(\hat{\mu}_{cn},\, \hat{\sigma}^2_{cn})$ are, respectively, the estimated mean and variance of the normal distribution corresponding to $CN = cn$. From the mBPCR model, given cn, the conditional

posterior distribution of any M_p is $\mathcal{N}(\widehat{m}_p, \hat{V}_p)$, where $(\widehat{m}_p, \hat{V}_p)$ are, respectively, the posterior mean and variance of M_p estimated by mBPCR.

Hyper-Parameters Estimation. The hyper-parameters of our model are: $P(X_i^N = Het)$, for $i = 1, \ldots, n$, p_{upd} and all $P(Y_i = y \mid X_i^N = x, \widetilde{W}_i = w)$, for $y \in \mathbb{Y}$, $x \in \mathbb{X}$, w=-3, -2, -1, 0, 2.

To estimate the set of conditional probabilities $\{P(Y_i = y \mid X_i^N = x, \widetilde{W}_i = w)$, $y \in \mathbb{Y}$, $x \in \mathbb{X}$, w=-2, -1, 0, 2$\}$, we needed paired normal-cancer samples, since they are related to the probability of detecting a certain genotype in a cancer cell, given the corresponding genotype in a normal cell of the same patient and under some CN event. Hence, we used some breast cancer cell line samples of [13], suitable for our purpose. Instead, to estimate $\{P(Y_i = y \mid X_i^N = x, \widetilde{W}_i = -3)$, $y \in \mathbb{Y}$, $x \in \mathbb{X}\}$, we used 11 IBD/UPD regions previously found by us on 5 samples of patients with Hairy Cell Leukemia [5] and on the B-cell lymphoma cell line KARPAS-422 (unpublished). All regions were detected by dChip [3]. Their width was between 3Mb and 100Mb (covering from 300 to 9800 SNPs), so that they were large enough to be really considered IBD/UPD regions. In both cases, we used a maximum likelihood estimation.

Regarding the prior probability of heterozygosity of each SNP i, $P(X_i^N = Het)$, we set it as the estimated probability of heterozygosity contained in the annotation file of the microarray used. In our application in Section 4, it is the GeneChip Human Mapping 250K NspI (Affymetrix, Santa Clara, CA, USA).

We did not have a suitable dataset to estimate the frequency of an IBD/UPD event (p_{upd}). In order to understand at least the order of magnitude of this parameter, we considered two studies on IBD regions: [1] and [6]. Using the data of the former paper (only the normal samples), we could estimate $p_{upd} \approx 1.7 \cdot 10^{-3}$. Instead, with the data of the latter, we estimated $p_{upd} \approx 1.5 \cdot 10^{-3}$, by considering all regions greater than 1Mb, while $p_{upd} \approx 1.46 \cdot 10^{-4}$, by considering only the regions greater than 3Mb. The differences in the estimations are due to the different resolutions of the technology used (in fact, in the former the number of SNPs used was 58,960, while in the latter was 3,107,620) and to the minimum length allowed for these regions. The wider the regions are, the higher is the probability that the regions represent "abnormalities" and the lower becomes the probability of their occurrence (so that p_{upd} is lower). In Section 4, we will try two values: $p_{upd} = 10^{-3}$ and $p_{upd} = 10^{-4}$.

3 Estimation Procedure

To estimate the piecewise constant profile of the aberration events, we used a Bayesian piecewise constant regression similar to mBPCR [10]. The prior distributions of the number of segments and the boundaries are defined as: $P(K = k) = \frac{k_{\max}+1}{k_{\max}} \frac{1}{k(k+1)}$, for $k \in \mathbb{K} = \{1, \ldots, k_{\max}\}$, and $P(\mathbf{T} = \mathbf{t} \mid K = k) = \binom{n-1}{k-1}^{-1}$, for $\mathbf{t} \in \mathbb{T}_{k,n}$, where $\mathbb{T}_{k,n}$ is a subspace of $\mathbb{N}_0^{k+1}$ such that $t_0 = 0$, $t_k = n$ and

$t_q \in \{1, \ldots, n-1\}$ for all $q = 1, \ldots, k-1$, in an ordered way and without repetitions. The estimators of the number of segments k_0 and the boundaries $\mathbf{t}^0$ are:

$$\widehat{K}_{01} = \arg\max_{k \in \mathbb{K}} p(k \,|\, \mathbf{Y},\, cn), \tag{3}$$

$$\widehat{\mathbf{T}}_{BinErrAk} = \arg\max_{\mathbf{t}' \in \mathbb{T}_{\hat{k},n}} \mathrm{E}\left[\sum_{q=1}^{\hat{k}-1} \sum_{p=1}^{k_0-1} \delta_{t'_q,\, t^0_p} \,\middle|\, \mathbf{Y},\, cn \right]. \tag{4}$$

Essentially, $\widehat{\mathbf{T}}_{BinErrAk}$ consists of the $\hat{k}_{01}$ positions which have the highest posterior probability to be a breakpoint. The difference with mBPCR is in the prior and in the estimation of the number of segments. Instead of using a uniform prior and an estimator which minimizes the posterior expected squared error, we consider a prior similar to $1/k^2$ and an estimator which minimizes the 0-1 error, in order to reduce the FDR in case of few segments.

Another difference with respect to mBPCR is in the level estimation. While in the CN model the levels were continuous random variables, now they assume categorical values. Hence, they are estimated separately (as before) with the MAP estimator instead of the posterior expected value,

$$\widehat{W_p} = \arg\max_{w=-3,-2,-1,0,1,2} \mathrm{P}(W_p = w \,|\, \mathbf{Y},\, \underline{\hat{t}},\, \hat{k},\, cn), \qquad p = 1, \ldots, \hat{k}, \tag{5}$$

where $\hat{\mathbf{t}}$ and $\hat{k}$ are estimates of, respectively, $\mathbf{t}^0$ and k_0. To compute the estimation, we used a dynamic program similar to the one used for mBPCR.

In general, the boundary estimator $\widehat{\mathbf{T}}_{BinErrAk}$ is an estimator with a high sensitivity, but also a medium FDR. The vector of the posterior probabilities to be a breakpoint, for all the points in the sample, (called $\mathbf{p}$) represents a multimodal function with the maxima at the breakpoint positions, but often in a neighborhood of each maximum there are other positions with high probability because of the uncertainty. Hence, if we take the first k_0 points with the highest probability (definition of $\widehat{\mathbf{T}}_{BinErrAk}$), we could take some points in the neighborhood of the higher maxima and not some maxima with a lower probability. To improve the estimation, since commonly the function shows clearly the positions of the true breakpoints in correspondence to the maxima, we thought to estimate, at the same time, both the number of the segments and the breakpoints with, respectively, the number of peaks and the locations of their maxima. The problem of the determination of the peaks is numerical and we made an algorithm to find them, which basically 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). We will denote the corresponding estimators with $\widehat{K}_{Peaks,thr_1,thr_2}$ and $\widehat{\mathbf{T}}_{Peaks,thr_1,thr_2}$.

We considered several pairs of thresholds and, on the basis of the results obtained on simulations (see [11]), we selected ($\hat{K}_{Peaks,01,01}$, $\hat{\mathbf{T}}_{Peaks,01,01}$), ($\hat{K}_{Peaks,01,mad}$, $\hat{\mathbf{T}}_{Peaks,01,mad}$) and ($\hat{K}_{Peaks,mad,01}$, $\hat{\mathbf{T}}_{Peaks,mad,01}$), where 01= $\max(0.01, \text{quantile of } \mathbf{p} \text{ at } 0.95)$, mad=$median(\mathbf{p})$+3∗$mad(\mathbf{p})$ and $mad(.)$ is the median absolute deviation. All the thresholds used were derived from different definitions of which probability values are to be considered significant.

4 Application on Real Data

The real data we used were paired samples of patients affected by chronic lymphocytic leukemia (CLL), which then transformed in diffuse large B-cell lymphoma (DLBCL), see [4]. For two patients we had also a third sample. In general, samples coming from the same patient should present the same IBD/UPD regions (the germ line ones) for the majority of the genome. Hence, we used them to evaluate the IBD/UPD detection of our method. Moreover, in [4] they also estimated the copy number of some genomic regions with FISH technique (fluorescent in situ hybridization) and we used them to evaluated the CN event estimation. For the estimation, we considered the estimators ($\hat{K}_{Peaks,01,01}$, $\hat{\mathbf{T}}_{Peaks,01,01}$), ($\hat{K}_{Peaks,01,mad}$, $\hat{\mathbf{T}}_{Peaks,01,mad}$) and ($\hat{K}_{Peaks,mad,01}$, $\hat{\mathbf{T}}_{Peaks,mad,01}$) and, as probability of IBD/UPD, either $p_{upd} = 10^{-4}$ or $p_{upd} = 10^{-3}$.

The sample of a patient can contain also a subpopulation of normal cells and other subpopulations of tumor cells bearing different gene lesions. Moreover, we observed that the $\log_2$ratio values corresponding to normal, gain, loss regions are sufficiently well separated only when we look at the CN changes born in at least 60% of the cells. As a consequence, the aim of our algorithm was to detect the aberrations present in at least 60% of the cells to ensure that the identified aberrations were true and not due to the noise of microarray data.

On the samples considered for the comparison, we had a total of 133 regions estimated by FISH technique. Regarding the 17 detectable aberrations (aberrations in at least 60% of the cells), 2 gains were not identified by all versions of the method, because the estimated $\log_2$ratio ($\sim$0.14) was lower than the threshold for the gains ($\sim$0.17). All versions found 3 of the 26 CN events not detectable and another was discovered by ($\hat{K}_{Peaks,01,01}$, $\hat{\mathbf{T}}_{Peaks,01,01}$) and ($\hat{K}_{Peaks,01,mad}$, $\hat{\mathbf{T}}_{Peaks,01,mad}$) with $p_{upd} = 10^{-3}$ and ($\hat{K}_{Peaks,mad,01}$, $\hat{\mathbf{T}}_{Peaks,mad,01}$) with $p_{upd} = 10^{-4}$. Only in 2/90 normal segments, all estimators discovered an aberration. In general, the samples used for microarray and FISH are not exactly the same, hence the percentage of cells which carry the aberrations can be different and a discordance between the two techniques is possible.

For the evaluation of the IBD/UPD region detection, we considered the only two patients with three samples (see Table 1). For the first patient, we found $\sim 78\%$ IBD/UPD regions exactly equal in all three samples and in total we could validate $\sim 95 - 98\%$ regions (considering also the regions exactly equal in at least two samples and the overlapping segments). For the second, we discovered $\sim 19 - 25\%$ equal IBD/UPD regions and validated $\sim 74\%$ regions. In both cases, almost all the remaining segments were smaller than 1Mb. The differences between the results of the two patients were partially due to the difference in the noise of the samples.

In conclusion, the three estimators behaved similarly and equally well on the real data used. Moreover, with both values of p_{upd} we often detected the same breakpoints for the IBD/UPD regions, but generally with $p_{upd} = 10^{-3}$ we discovered a higher number of regions and even smaller ones. Thus, $p_{upd} = 10^{-4}$ could be preferred in order to have more realistic IBD/UPD regions.

Table 1. Results regarding the IBD/UPD region detection, obtained on two patients using the three pair of estimators ($\hat{K}_{Peaks,01,01}$, $\hat{\mathbf{T}}_{Peaks,01,01}$), ($\hat{K}_{Peaks,01,mad}$, $\hat{\mathbf{T}}_{Peaks,01,mad}$) and ($\hat{K}_{Peaks,mad,01}$, $\hat{\mathbf{T}}_{Peaks,mad,01}$) and, as probability of IBD/UPD, either $p_{upd} = 10^{-4}$ or $p_{upd} = 10^{-3}$.

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

5 Conclusions

We propose a new algorithm for the joint estimation of CN events and IBD/UPD regions, in order to better identify these types of genomic aberrations. Our method consists in a Bayesian piecewise constant regression, which takes into account the errors in the genotyping measurements of microarrays, due to the aberrations affecting the CN. Moreover, differently from the only other method present in literature (i.e., [12]), it considers all the CN events biologically relevant. The goodness of our model is supported by the results obtained on real data. Therefore, our method can be very useful, for example, in cancer research, to find genomic mutations that characterize the disease.

References

1. Bacolod, M.D., et al.: The Signatures of Autozygosity among Patients with Colorectal Cancer. Cancer Research 68, 2610–2621 (2008)
2. Bea, S., et al.: Uniparental disomies, homozygous deletions, amplifications and target genes in mantle cell lymphoma revealed by integrative high-resolution whole genome profiling. Blood (2008)
3. Beroukhim, R., et al.: Inferring Loss-of-Heterozygosity from Unpaired Tumors Using High-Density Oligonucleotide SNP Arrays. PLOS Computational Biology 2, 323–332 (2006)

4. Bertoni, F., et al.: Genome wide-DNA profiling of Richter's syndrome-diffuse large B-cell lymphoma (RS-DLBCL): differences with de novo DLBCL and possible mechanisms of transformation from chronic lymphocytic leukemia (CLL). Blood (ASH annual meeting abstracts) 112(11), 720 (2008)
5. Forconi, F., et al.: High density genome-wide DNA profiling reveals a remarkably stable profile in hairy cell leukaemia. British Journal of Haematology 141, 622–630 (2008)
6. The international HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–862 (2007)
7. Hodgson, G., et al.: Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. Nature Genetics 29, 459–464 (2001)
8. Newton, M.A., Lee, Y.: Inferring the Location and Effect of Tumor Suppressor Genes by Instability-Selection Modelling of Allelic-Loss Data. Biometrics 56, 1088–1097 (2000)
9. Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M.: Circular Binary Segmentation for the Analysis of Array-based DNA Copy Number Data. Biostatistics 4, 557–572 (2004)
10. Rancoita, P.M.V., Hutter, M., Bertoni, F., Kwee, I.: Bayesian DNA copy number analysis. BMC Bioinformatics 10(10) (2009)
11. Rancoita, P.M.V., Hutter, M., Bertoni, F., Kwee, I.: An integrated Bayesian analysis of genotyping and copy number data (in preparation)
12. Scharpf, R.B., Parmigiani, G., Pevsner, J., Ruczinski, I.: Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. Annals of Applied Statistics 2, 687–713 (2008)
13. Zhao, X., et al.: An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. Cancer Research 64, 3060–3071 (2004)