Biostatistics Advance Access published January 5, 2011

Biostatistics (2011), **0**, 0, *pp.* 1–16
doi:10.1093/biostatistics/kxq076

Joint segmentation, calling, and normalization of multiple CGH profiles

FRANCK PICARD*

Laboratoire de Biometrie et Biologie Evolutive, UMR CNRS 5558 - Univ. Lyon 1, F-69622, Villeurbanne, France and Projet BAMBOO, INRIA Rhône-Alpes, F-38330 Montbonnot Saint-Martin, France
franck.picard@univ-lyon1.fr

EMILIE LEBARBIER
UMR AgroParisTech/INRA MIA 518, Paris, France and LBBE, UMR INRA 518 MIA, Paris, France

MARK HOEBEKE
Laboratoire, Statistique et Génome, UMR CNRS 8071-INRA 1152-UEVE F-91000 Evry, France

GUILLEM RIGAILL
Department of Translational Research, Institut Curie, 26 rue d'Ulm, Paris, 75248, France, UMR AgroParisTech/INRA MIA 518, Paris, France and UMR INRA 518 MIA, Paris, France

BABA THIAM
EQUIPPE, Univ. Lille 3, BP 149, 59653 Villeneuve d'Ascq, France, UMR AgroParisTech/INRA MIA 518, Paris, France and UMR INRA 518 MIA, Paris, France

STÉPHANE ROBIN
UMR AgroParisTech/INRA MIA 518, Paris, France and UMR INRA 518 MIA, Paris, France

SUMMARY

The statistical analysis of array comparative genomic hybridization (CGH) data has now shifted to the joint assessment of copy number variations at the cohort level. Considering multiple profiles gives the opportunity to correct for systematic biases observed on single profiles, such as probe GC content or the so-called "wave effect." In this article, we extend the segmentation model developed in the univariate case to the joint analysis of multiple CGH profiles. Our contribution is multiple: we propose an integrated model to perform joint segmentation, normalization, and calling for multiple array CGH profiles. This model shows great flexibility, especially in the modeling of the wave effect that gives a likelihood framework to approaches proposed by others. We propose a new dynamic programming algorithm for break point positioning, as well as a model selection criterion based on a modified bayesian information criterion

*To whom correspondence should be addressed.

proposed in the univariate case. The performance of our method is assessed using simulated and real data sets. Our method is implemented in the R package cghseg.

Keywords: Array CGH; Calling; Dynamic programming; Joint segmentation; Wave effect.

1. INTRODUCTION

Cancer bioinformatics has received enormous attention in the past 10 years, and studying the structure of cancer genomes has been a productive research direction. Linking chromosomal aberrations and cancer is far from new: oncogenes and tumor suppressor genes are known to be frequently amplified or deleted, leading to DNA copy imbalances. In the late 1990s, the microarray comparative genomic hybridization (CGH) technology has allowed the investigation of copy number changes at the genome scale in one experiment (Snijders *and others*, 2001). To date, statistical efforts have mainly focused on the recovery of the segmental structure by segmentation and the hidden discrete copy number values from the raw data with a "calling" step, at the single-sample level. More than 30 methods have been published on the subject, and reviews concerning array CGH data analysis are now available (Park, 2008; Van de Wiel *and others*, 2010). The proposed statistical frameworks range from break point detection (Olshen *and others*, 2004; Picard *and others*, 2005; Rancoita *and others*, 2009) to smoothing (Hupe *and others*, 2004; Ben-Yaacov and Eldar, 2008) and hidden Markov models (Marioni *and others*, 2006; Stjernqvist *and others*, 2007). Existing methods have already been compared, and one consistent result is that segmentation methods perform best for the analysis of array CGH data (Lai *and others*, 2005; Willenbrock and Fridlyand, 2005).

As the array CGH (aCGH) technology becomes more popular, biologists now face the problem of analyzing profiles associated with several patients simultaneously. Even though break point detection can easily be achieved at the single-patient level, new modeling and computational challenges arise at the multi-patient level. Many questions need to be addressed such as the joint analysis of chromosomal alterations for a set of profiles (Pique-Regi *and others*, 2009; Van de Wiel *and others*, 2009), the detection of recurrent alterations within this set (Rouveirol *and others*, 2006; Shah, 2008; Rueda and Diaz-Uriarte, 2009; Robin and Stefanov, 2009) and the clustering of patients according to their CGH profile (Van Wieringen *and others*, 2008; Liu *and others*, 2006). In this paper, we address the first task that involves 3 subtasks: (i) segmentation, (ii) calling, and (iii) normalization. Each of these tasks could be achieved on each profile separately, but their efficiency and sensitivity is expected to be improved by joint analysis.

(i) The efficiency of the segmentation approach is based on the use of dynamic programming (DP) (Picard *and others*, 2005). However, a drawback of this algorithm is that its complexity is $O(Kn^2)$, with n being the number of markers and K the number of segments. Consequently, segmenting multiple profiles raises a major computational issue. In this work, we propose a trick to use DP on multiple profiles, whose complexity is reduced thanks to a second layer of DP.

(ii) The calling step consists in the assignment of copy number values to probes to determine which probes are in the "deleted," "amplified," or "normal" state for instance. One limitation of pure segmentation methods is that these do not give information about the copy number values. "Merging" steps have been proposed to cluster segments into groups of homogeneous copy number values. These strategies are based on statistical tests (Willenbrock and Fridlyand, 2005) or on clustering (Van de Wiel *and others*, 2007; Picard *and others*, 2007). Willenbrock and Fridlyand (2005) showed that this downstream step was of "paramount importance" when using segmentation for aCGH. But the merging step only constitutes a second-stage procedure, whereas segmentation can also learn from the calling step in a unified model to gain in power in the detection of breaks that correspond to changes in copy number values (Picard *and others*, 2007). Considering multiple profiles gives the opportunity to perform global calling for the whole data set since the average signal associated to each copy number change is likely to be common across profiles.

(iii) By normalization, we refer to a step that removes or accounts for possible artifacts of the aCGH technology. Performing a joint analysis constitutes an opportunity to correct for this bias that is shared by all signals measured on the same type of arrays. The origins of this bias is unclear, but a consensus exists on its link with GC content (Carter, 2007; Pique-Regi *and others*, 2009). It can be viewed as a heterogeneity between hybridization intensities that would be observed even when dealing with DNA without aberration. When considering one profile only, correcting this bias is dangerous since there exists an aliasing between copy number changes and wavy patterns. Consequently, this correction may be suitable for single copy number variation (CNV) profiles, but not for cancer profiles for which aberrations could be smoothed as well. When considering multiple profiles, a calibration set can be used to estimate this wave bias and to remove it from the data (Van de Wiel *and others*, 2009). However, when no calibration set is available, this bias can be modeled for by adding a correction term within the segmentation model. In this article, we propose a unified statistical framework to correct for those effects. The model we propose can be viewed as a generalization of those of Pique-Regi *and others* (2009) and Van de Wiel *and others* (2009) by the integration of the calling method within the segmentation model. We also propose a general normalization strategy that may include probe-specific bias correction or account for any exogenous covariate such as GC-content.

On the subject of joint CGH analysis, Korn *and others* (2008) and Wu *and others* (2009) perform segmentation and calling. Shah *and others* (2009) follow the same 2 goals, with a more sophisticated model, combining hidden Markov models and mixtures. Pique-Regi *and others* (2009) propose an iterative procedure, based on a sparse Bayesian learning approach, that both performs segmentation and normalization. Van de Wiel *and others* (2009) address the same issues, using regularized least squares with a calibration set. Among other questions addressed in the literature, we also mention the assessment of the significance of detected alterations, which holds for the segment call but does not provide a formal clustering. This is considered in Van de Wiel *and others* (2009). Zhang *and others* (2010) propose a statistical test to assess the existence of a CNV at a given position, based on pooled aCGH coming from different platforms. Finally, we also mention that each of the 3 tasks raises model selection issues, segmentation being probably the most crucial. Zhang and Siegmund (2007) recently proposed a modified bayesian information criterion (BIC) criterion for segmentation models that we generalize to the mutiple profiles case.

In this article, we first present the general model in Section 2. We then address the computational issues raised by multiple aCGH segmentation in Section 2. Several corrections (normalizations) are then proposed in Section 3. The mixture model for the calling step is presented in Section 4, and a model selection criterion is derived in the following section. We finally assess the efficiency of our procedure through a simulation study and show what can be learned from the estimated background intensity (Section 6). We also show how the method can be used to increase the performance of other segmentation algorithms such as CBS of Olshen *and others* (2004). Our procedure is illustrated on the hapmap data set in the final section. The cghseg R package integrates all the presented methods as well as the former segmentation procedures published in Picard *and others* (2005).

2. Joint segmentation model

2.1 *Notations for segmentation models*

Since segmentation models have been shown to be efficient for the analysis of single profiles, our first objective is to propose their generalization to multiple profiles. Then $Y_i(t)$ will denote the log-ratio measured at position t for patient i, $\mathbf{Y}_i$ will denote the single profile for patient $i = 1, \ldots, I$ of size n_i. Some samples may show missing values so that the size of each signal may differ between samples. Generalizing the framework proposed by Picard *and others* (2005), we suppose that the mean of profile $\mathbf{Y}_i$ is subject to $k_i - 1$ abrupt changes at break points $\{t_k^i\}$ (with convention $t_0^i = 0$ and $t_{k_i}^i = n_i$) and is constant

between 2 break points within the interval $]t^i_{k-1}, t^i_k]$. In the following, we denote by $K = \sum_i^I k_i$ the total number of segments across profiles and $N = \sum_i^I n_i$ the total number of observations. Thus, we consider the following model:

$$\forall t \in]t^i_{k-1}, t^i_k], \quad Y_i(t) = \mu_{ik} + E_i(t),$$

where $E_i(t)$ stands for a Gaussian white noise with variance σ^2. In order to use the matricial formulation of linear models, we introduce the incidence matrix of break points denoted by $\mathbf{T}_{[N\times K]} = \text{Bloc}_{[n_i\times k_i]}[\mathbf{T}_i]$ with $\mathbf{T}_i = \text{Bloc}[\mathbb{1}_{n^i_{k_i}}]$ being the incidence matrix of break points in profile i, and with $n^i_k = t^i_k - t^i_{k-1} + 1$ being the length of segment k for profile i. We also introduce notation $\boldsymbol{\mu}_{[K\times 1]} = [\mu_{ik}]$. Then our model is $\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{E}$, where $\mathbf{Y}_{[N\times 1]}$ stands for the observed data, and where $\mathbf{E}$ is centered Gaussian with diagonal covariance matrix $\sigma^2\mathbf{I}$.

2.2 *Using DP for joint segmentation*

In this section, the total number of segments K is fixed and the goal is to find the best joint segmentation into K segments according to the maximum likelihood criterion as in the case of single segmentation (Picard *and others*, 2005). The selection of K will be studied in Section 5. For this purpose, DP is the computational key ingredient. However, the question of computational efficiency is asked when considering I profiles because DP complexity is quadratic with the size of the data which is of order nI. We propose a computational trick to reduce this burden when segmenting multiple profiles.

The minimization problem resumes to finding $\{\hat{\mathbf{T}}, \hat{\boldsymbol{\mu}}\}$ such that

$$\{\hat{\mathbf{T}}, \hat{\boldsymbol{\mu}}\} = \underset{\{\mathbf{T},\boldsymbol{\mu}\}}{\text{argmin}}\, \text{RSS}_K(\mathbf{T}, \boldsymbol{\mu}),$$

with $\text{RSS}_K(\mathbf{T}, \boldsymbol{\mu})$ the residual sum of squares of a segmentation model with K segments such that

$$\begin{aligned}\text{RSS}_K(\boldsymbol{\mu}, \mathbf{T}) = \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu}\|^2 &= \sum_{i=1}^{I}\sum_{k=1}^{k_i} \text{RSS}^i_k(\boldsymbol{\mu}_i, \mathbf{T}_i) \\ &= \sum_{i=1}^{I}\sum_{k=1}^{k_i}\sum_{t\in]t^i_{k-1},t^i_k]} (Y_i(t) - \mu_{ki})^2.\end{aligned}$$

When dealing with multiple profiles, this minimization must be done under an additional constraint that is, $\sum_i k_i = K$. The computational trick we propose is based on the following breakdown:

$$\min_{\{\mathbf{T},\boldsymbol{\mu}\}} \text{RSS}_K(\mathbf{T}, \boldsymbol{\mu}) = \min_{k_1+\ldots+k_I=K} \left\{ \sum_{i=1}^{I} \min_{\mathbf{T}_i,\boldsymbol{\mu}_i} \text{RSS}^i_{k_i}(\mathbf{T}_i, \boldsymbol{\mu}_i) \right\}.$$

Since the RSS is additive according to the profiles and to the number of segments, we propose a double-stage DP to solve this optimization problem. Let us introduce a new notation to explain the core of the algorithm and denote by $\widehat{\mathbf{T}}^i(k_i)$ the set of optimal breaks with k_i segments for profile i.

Stage 1. The first step consists in finding all optimal break points for each profile for $k_i = 1, \ldots, k_{\max}$ segments: $\widehat{\mathbf{T}}^i(k_i)$. $k_{\max}$ corresponds to the maximum number of segments for one profile (to be set by the user). This first stage is done using classical DP described in Picard *and others* (2005).

Stage 2. The second step consists in the allocation of the optimal number of segments to each profile. We aim at determining the optimal sequence $\widehat{k}_1, \ldots, \widehat{k}_I$, such that $K = \sum_i \widehat{k}_i$. We denote by

$\mathrm{RSS}_K(\widehat{\mathbf{T}}^1(k_1),\ldots,\widehat{\mathbf{T}}^I(k_I))$ the total sum of squares for a model with K segments spread over I profiles, each having k_i segments. This step is solved using recursion: $\forall i \in [1:I]$,

$$\{\widehat{k}_1,\ldots,\widehat{k}_i\} = \underset{k_1+\ldots+k_i=K}{\operatorname{argmin}} \mathrm{RSS}_K(\widehat{\mathbf{T}}^1(k_1),\ldots,\widehat{\mathbf{T}}^i(k_i))$$
$$= \underset{k'+k''=K}{\operatorname{argmin}} \left\{ \underset{k'_1+\ldots+k'_i=k'}{\mathrm{RSS}_{k'}}(\widehat{\mathbf{T}}^1(k'_1),\ldots,\widehat{\mathbf{T}}^{i-1}(k'_{i-1})) + \mathrm{RSS}^i_{k''}(\widehat{\mathbf{T}}^i(k'')) \right\}.$$

At the end of this double-stage DP, we have the optimal break point positions for the optimal number of segments in each profile.

Complexity in time. The first stage corresponds to the segmentation of individual profiles into $k_{\max}$ segments each, with complexity $\mathrm{O}(n^2 I k_{\max})$. The complexity of the second stage is $\mathrm{O}((I k_{\max})^2 \times I)$ which makes the overall complexity of order $\mathrm{O}(I n^2 k_{\max} + k^2_{\max} I^3)$. Assuming that the major term is n (which is consistent with the ever increasing density of aCGH), the second term remain negligible and the complexity becomes $\mathrm{O}(I k_{\max} n^2)$. This complexity should be compared with the one of DP applied to the complete data set with $N = In$ points into $K_{\max} = I k_{\max}$ segments, that is, $\mathrm{O}(K_{\max} N^2) = \mathrm{O}(I^3 k_{\max} n^2)$. The 2-stage DP therefore reduces the complexity with a factor I^2.

3. Integrative normalization

The interest in considering many profiles is that if a systematic bias is observed for every profile, considering the joint analysis can help in its correction. In the following, we will denote by $b(t)$ this bias at position t (for probes $t = 1,\ldots,n$), and we suppose that it is present and constant accross profiles, such that the segmentation model becomes

$$\forall t \in]t^i_{k-1}, t^i_k], \quad Y_i(t) = \mu_{ik} + b(t) + E_i(t).$$

Then we use a unified matricial formulation such that $\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{X}\mathbf{b} + \mathbf{E}$, with $\mathbf{X} = (\mathbf{I}_n,\ldots,\mathbf{I}_n)^T$ (I blocks) which spreads the common fixed effect $\mathbf{b} = (b(t_1),\ldots,b(t_n))^T$ over the I patients. In the following, we propose 2 ways to model this bias.

3.1 *Probe effect*

A first model consists in considering a linear model where $b(t) = \beta_t$ stands for a probe effect, or a reference hybridization intensity as proposed by Pique-Regi *and others* (2009). This modeling would consider that if a probe shows a systematic bias, it would be detected by this effect. This is the simplest correction that could be made on the data. Parameters $\{\beta_t\}_t$ can be estimated using an iterative least squares algorithm such that $\widehat{\beta}_t^{[h+1]} = \sum_{i=1}^I (Y_i(t) - \widehat{\mu}_{ik}^{[h]})/I$, and breaks are updated with DP on $\mathbf{Y} - \mathbf{X}\widehat{\mathbf{b}}^{[h+1]}$.

3.2 *Smoothing with splines*

One criticism that can be made to this position-effect model is that it does not account for spatial correlations suggested by the "wave" pattern. An alternative would be to model this bias by a smooth function. This is why we model $b(t)$ in a nonparametric fashion such that $\mathbf{b}$ is a functional part of the model. We get a semiparametric model with $\mathbf{T}\boldsymbol{\mu}$ its parametric part (break points). The nonparametric part is handled by a functional basis such as a spline basis to determine the shape of this bias function. The idea is to introduce some control on the regularity of the bias function. In this context, the fit complexity of

function $\mathbf{b} = (b(t_1), \ldots, b(t_n))$ can be controlled by a regularization strategy using the classical second derivative–based penalty, such that

$$\min_{\mathbf{T},\boldsymbol{\mu},\boldsymbol{\theta}} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda I \int [b''(t)]^2 \mathrm{d}t \right\}. \tag{3.1}$$

It is well known that the solution of this minimization problem is given by a spline basis (Hastie *and others*, 2001). Here, we avoid the knot selection problem by using the maximal set of knots $(t_1, \ldots, t_n)$. By analogy, we denote by $\{\mathbf{W}\}_{jk} = W_j(t_k)$ a n-dimensional set of natural spline functions such that $\mathbf{b} = \mathbf{W}\boldsymbol{\theta}$. Then criterion (3.1) becomes

$$\min_{\mathbf{T},\boldsymbol{\mu},\boldsymbol{\theta}} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu} - \mathbf{X}\mathbf{W}\boldsymbol{\theta}\|_2^2 + \lambda I \boldsymbol{\theta}^T \boldsymbol{\Omega} \boldsymbol{\theta} \right\}, \tag{3.2}$$

with $\boldsymbol{\Omega}_{ij} = \int W_i''(t) W_j(t) \mathrm{d}t$. The solution of this minimization is given by

$$\widehat{\boldsymbol{\theta}} = \{\mathbf{W}^T\mathbf{W} + \lambda\boldsymbol{\Omega}\}^{-1}\mathbf{W}^T(\mathbf{X}^T\widetilde{\mathbf{Y}}/I),$$

where $\mathbf{X}^T\widetilde{\mathbf{Y}}/I$ represents the average segmentation residuals at each position. Then break points are updated using an unbiased version of the signal $\widetilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\mathbf{b}}^{[h]}$. As for constant λ, it is calibrated using cross validation. From a methodological point of view, other basis could be used. In the supplementary material available at *Biostatistics* online, we show how the wavelet basis can be used as well.

3.3 *Correction for GC content*

In the last step, we adjust for the GC content of probes. GC correction has been shown to improve the signal to noise ratio in some platforms (Nannya *and others*, 2005) and can be performed using a simple quadratic regression scheme:

$$\forall t \in]t_{k-1}^i, t_k^i], \quad Y_i(t) = \mu_{ik} + b(t) + \alpha_1 \mathrm{GC}_t + \alpha_2 \mathrm{GC}_t^2 + E_i(t),$$

where (α_1, α_2) can be estimated by an iterative least squares algorithm as proposed for the "position" effect.

4. MULTIVARIATE CALLING

The principle of the segmentation/clustering model proposed in Picard *and others* (2007) is to integrate in the segmentation model that different segments with the same underlying copy number should share the same mean signal. Let's suppose we observe P distinct states, then the mean of each segment should lie in a restricted set $\mathbf{m} = \{m_1, \ldots, m_P\}$, where m_1 is the inferred mean from all segments in cluster 1 that share the same copy number. This is why we introduce a random classification matrix $\mathbf{C}_{[K\times P]}$ that gives the state for each segment with $\{C_{kp}^i = 1\}$ if segment k of profile i is in cluster p. Then the joint segmentation/clustering model is such that

$$\{C_{kp}^i = 1\}, \quad \forall t \in]t_{k-1}^i, t_k^i], \quad Y_i(t) = m_p + b(t) + E_i(t)$$

or equivalently $\mathbf{Y} = \mathbf{TCm} + \mathbf{Xb} + \mathbf{E}$. Term $\mathbf{TCm}$ gives information about breaks position ($\mathbf{T}_{[N\times K]}$), state of segments ($\mathbf{C}_{[K\times P]}$), and mean corresponding to each state ($\mathbf{m}_{[P\times 1]}$), whereas the bias term $\mathbf{b}$ is still common across profiles.

We use a modified version of the expectation–maximization algorithm to estimate the maximum likelihood parameters adapted from Picard *and others* (2007). Briefly, the E-step is used to assess $\{\tau_p^{ik}\}$ the "posterior" probabilities of membership to clusters for each segment using a classical Bayes rule, and the means $\{m_p\}_p$ corresponding to each state are estimated in the M-step such that

$$\tau_p^{ik[h+1]} = \frac{\pi_p^{[h]}\phi(\mathbf{Y}_k^i; m_p^{[h]}, \sigma^{2[h]})}{\sum_\ell \pi_\ell^{[h]}\phi(\mathbf{Y}_k^i; m_\ell^{[h]}, \sigma^{2[h]})},$$

$$m_p^{[h+1]} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{k_i}\tau_p^{ik[h+1]}\sum_{t\in]t_{k-1}^i,t_k^i]}(Y_i(t) - b^{[h]}(t))}{\sum_{i=1}^{I}\sum_{k=1}^{k_i} n_k^i \tau_p^{ik[h+1]}}.$$

Here, π_p stands for the "prior" proportion of cluster p, $\mathbf{Y}_k^i$ stands for the vector of observations within segment k of sample i (of size n_k^i), and $\phi(\bullet)$ stands for the Gaussian density.

To update $b(t)$ in the probe-effect model (β_t), the maximum likelihood estimator is $\beta_t^{[h+1]} = \sum_{i,p}\tau_p^{ik(t)[h]}(Y_i(t) - m_p^{[h+1]})/I$ which corresponds to the weighted residuals at position t after segmentation/clustering. When considering the semiparametric model, the projection to the spline basis is done on the same weighted residuals.

5. Model selection

As discussed by many authors, segmentation models raise a difficult issue in terms of model selection. The question has been studied in the single profile context (Picard *and others*, 2005; Zhang and Siegmund, 2007), and the work should be done for joint segmentation. As mentioned by Zhang and Siegmund (2007), the discrete nature of the break points make the use of the classical BIC criterion not theoretically justified. Thus, they proposed a powerful framework to circumvent this difficulty by considering a continuous-time version of the model. In this setting, they derive an efficient modification of the BIC for single profile segmentation which we generalize to the joint segmentation.

The generalization of the criterion is based on the following remark: segmenting I profiles of lengths $(n_i)_i$ into K segments is equivalent to segmenting a single profile with length N ($N = \sum_i n_i$) into K segments with I break points being fixed. Then the generalization is based on a new definition of the prior distribution of the break points (denoted by $f(\tau)$ in Zhang (2005)). Here, $K - I$ break points are spread on $[0, N]$ with uniform probabilities $f(\tau) = (K-I)!/N^{(K-I)}$. Since there is no change but the prior distribution in the construction of the criterion, the complete derivation of the generalization follows the proof of Theorem 2.2 in Zhang (2005), and the new criterion for joint segmentation becomes

$$\underset{\text{JointSeg}}{\text{mBIC}}(K) = \left(\frac{N-K+1}{2}\right)\log\left[1 + \frac{SS_{\text{bg}}(\widehat{\mathbf{T}})}{SS_{\text{wg}}(\widehat{\mathbf{T}})}\right] + \log\left[\frac{\Gamma\left(\frac{N-K+1}{2}\right)}{\Gamma\left(\frac{N+1}{2}\right)}\right]$$
$$+\frac{K}{2}\log(SS_{\text{all}}) - \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{k_i}\log\widehat{n}_k^i + \left(\frac{1}{2} - (K-I)\right)\log(N),$$

where SS_{bg}, SS_{wg}, and SS_{all} stand for the between-group, within-group, and total sum of squares, respectively. $SS_{\text{bg}} = \sum_{i=1}^{I}\sum_{k=1}^{k_i}\widehat{n}_k^i(\bar{Y}_{ik} - \bar{Y})^2$, $SS_{\text{all}} = \sum_{i=1}^{I}\sum_{t=1}^{n_i}(Y_i(t) - \bar{y})^2$, $SS_{\text{wg}} = SS_{\text{all}} - SS_{\text{bg}}$, with $\widehat{n}_k^i$ is the length of segment k in profile i ($\widehat{n}_k^i = \widehat{t}_k^i - \widehat{t}_{k-1}^i + 1$) and $\bar{Y}_{ik} = \sum_{t=\widehat{t}_{k-1}^i+1}^{\widehat{t}_k^i} Y_i(t)/\widehat{n}_k^i$.

Then the model selection strategy consists in selecting the number of segments K that maximizes this criterion. Note that to reduce the computational time of the search of this optimum, we use the golden

search algorithm that avoids the computation of this criterion of all possible values of K. This algorithm is shortly described in the supplementary material available at *Biostatistics* online.

When joint calling is performed, we generalize the criterion given in Theorem 3 by Zhang and Siegmund (2007) with the same new prior distribution for break points that gives

$$\underset{\text{JointSeg/Clust}}{\text{mBIC}}(K, P) = \left(\frac{N-P+1}{2}\right)\log\left[1+\frac{SS_{\text{bg}}(\widehat{\mathbf{T}}, \widehat{\mathbf{m}})}{SS_{\text{wg}}(\widehat{\mathbf{T}}, \widehat{\mathbf{m}})}\right] + \log\left[\frac{\Gamma\left(\frac{N-P+1}{2}\right)}{\Gamma\left(\frac{N+1}{2}\right)}\right] + \frac{P}{2}\log(SS_{\text{all}}) - \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{k_i}\sum_{p=1}^{P}\log \widehat{n}_k^i(p) + \left(\frac{1}{2}-(K-I)\right)\log(N),$$

where $\widehat{n}_k^i(p)$ is the length of segment k in profile i that belongs to cluster p.

When normalization is performed, we use the same criterion since the dimension of the added effects does not depend on the number of segments. Consequently, the penalty term remains unchanged but the quality of fit is considered in the likelihood of the model.

6. SIMULATION STUDY

6.1 *Profiles without common variations*

Motivations. We first use simulations to assess the bias and precision of our methods for various signal configurations. We also propose to compare the different normalization strategies and to compete our method with CBS of Olshen *and others* (2004) like in other studies (Van de Wiel *and others*, 2009). We use the mergeLevels procedure of Willenbrock and Fridlyand (2005) to perform calling in the context of CBS. The R function can be downloaded at http://www.cbs.dtu.dk/∼hanni/aCGH/. To establish a fair comparison, we also propose to combine CBS with the wave corrections proposed above: recursive split and wave normalization are performed iteratively until stabilization of the segmentation. Our procedure has been implemented in the cghseg R package which is freely available (soon deposited on the CRAN).

Simulation set-up. We use the same simulation setting as Pique-Regi *and others* (2009). We fix the number of profiles at $I = 20$ and the length of the profiles at $n = 500$ to reduce the number of free parameters. The true signal $\boldsymbol{\mu}^0$ takes values in $\{-1, 0, \log_2(3/2)\}$, and we set the average number of segments per profile at $\bar{k} = 5$. The bias function is modeled as $b^0(t) = \tau \sin\left(\frac{2\pi t}{100}\right) + F(t)$, with $F(t) \sim \mathcal{N}(0, \tau^2)$ to account for both sinusoidal waves and noise without spatial structure (Pique-Regi *and others*, 2009). The measurement noise is supposed to be i.i.d. $\mathcal{N}(0, \sigma^2)$. To account for noise balance between measurement error and background intensity, we use $\lambda = \sigma/\tau$, a noise ratio parameter that equals 2 in Pique-Regi *and others* (2009) ($\sigma = 1$, $\tau = 0.5$). We explore more complex configurations with $\lambda \in \{1; 1.5; 2\}$. Finally, we define the signal to noise ratio as $\text{SNR} = \|\boldsymbol{\mu}^0\|^2/(nI\sigma^2)$.

Examples of simulated trajectories are given in the Figure 1 of the supplementary material available at *Biostatistics* online. Each configuration is simulated 20 times. To assess the performance of each method, we use the following criteria:

- the bias and precision of the model selection procedure to estimate the number of segments per profile using the bias $\text{Bias}(k_i)$ and the mean square error $\text{MSE}(k_i)$,
- the performance of break point positioning: given P the number of detected breaks, N the number of positions that are not breaks, FP the number of detected breaks that do not correspond to true breaks, and FN the number of true breaks that are not detected, we use the false discovery rate ($\text{FDR} = 1 - \text{TP/P}$) and false-negative rate ($\text{FNR} = \text{FN/N}$).

- the quality of profiles estimation with the mean square error $\text{MSE}(\boldsymbol{\mu}) = \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^0\|^2$,
- the quality of the bias estimation with the mean square error $\text{MSE}(\mathbf{b}) = \|\widehat{\mathbf{b}} - \mathbf{b}^0\|^2$.

Results. Results should be interpreted sequentially since model selection determines the number of segments that determines the quality of the subsequent estimators (position of break points and mean parameters). Figure 1 shows that the proposed model selection criterion underestimates the number of segments.

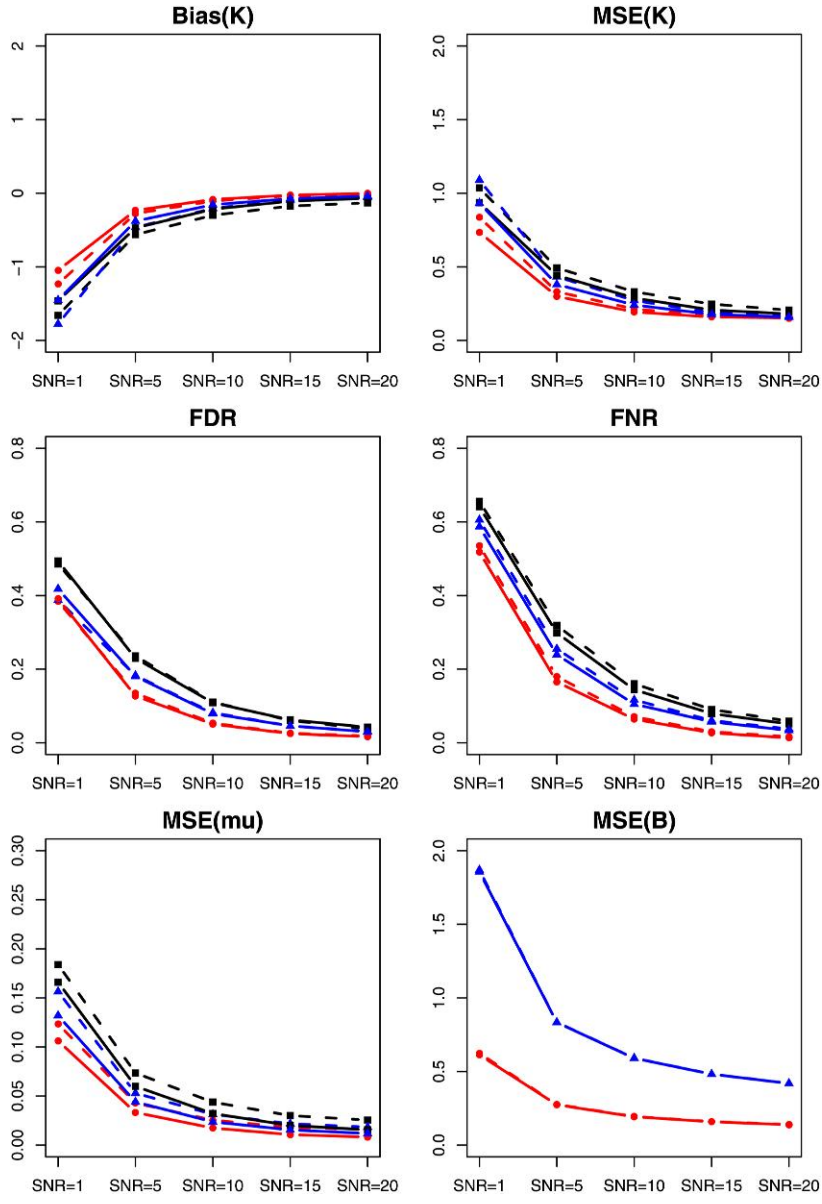

Fig. 1. Estimation performance in terms of bias and MSE according to SNR, calling, and normalization when using joint segmentation. Dashed line: without calling, solid line: with calling, ■ without normalization, ● (red) with position effect correction, and ▲ (blue) spline-based normalization.

For instance, when $\bar{k} = 5$ and SNR $= 1$, 3 segments are selected per profile on average. However, when looking at one realization of such profiles (Figure 1 of supplementary Material available at *Biostatistics* online, top panel), one may prefer to avoid false-positive break points when the SNR is low. This suggests that our heuristic criterion for model selection is conservative on average. On the contrary, the splitting strategy proposed in CBS is positively biased: when no correction is applied too many segments are detected whatever the SNR that does not increase the power of the procedure in terms of break point positioning (in terms of FDR and FNR, Figure 2).

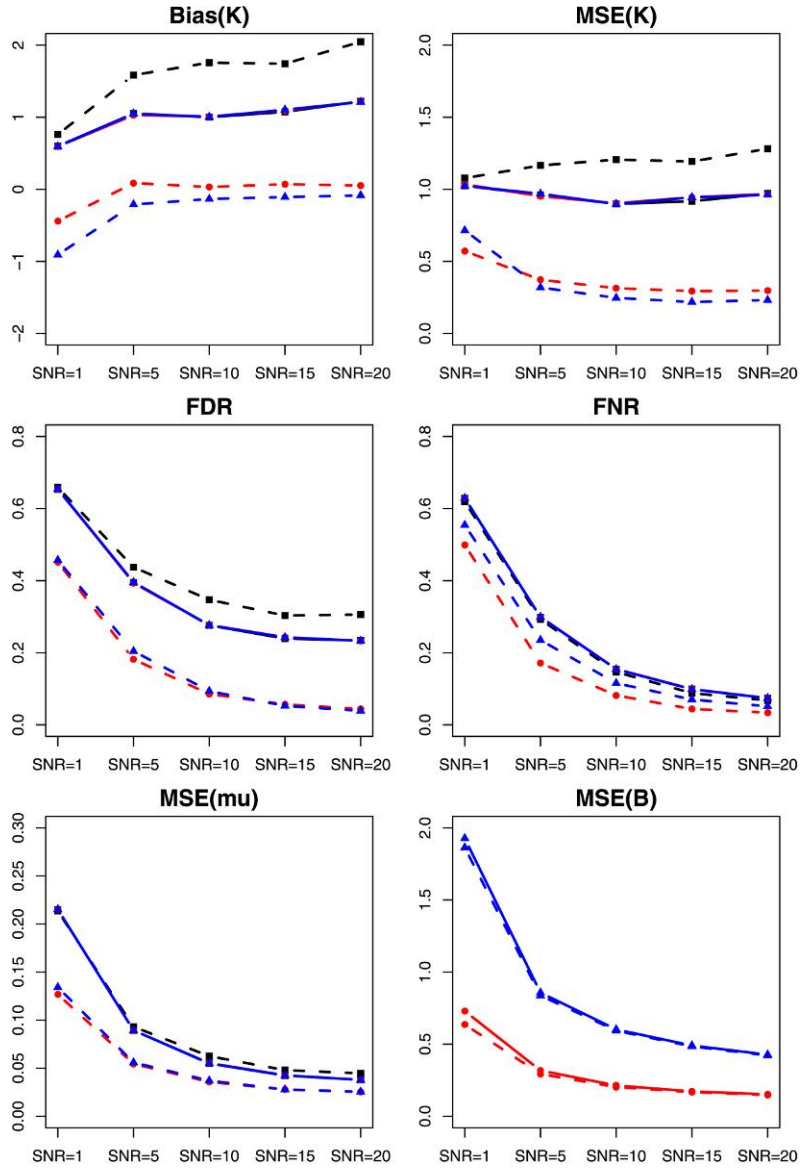

Fig. 2. Estimation performance in terms of bias and MSE according to SNR, calling, and normalization when using CBS. Dashed line: without calling, solid line: with calling, ■ without normalization, ● (red) with position effect correction, and ▲ (blue) spline-based normalization.

Normalization increases the performance of both joint segmentation and CBS whatever the difficulty of the simulation (Figures 1 and 2). The estimation error on the number of segments is decreased leading to smaller FDR and FNR. Normalization leads also to a gain in power by separating segments from the waves of the background intensity. However, the effect of calling is different in both methods: it increases the power of joint segmentation as described in the single profile case (Willenbrock and Fridlyand, 2005), but it does not increase the performance of CBS. Let us recall that the calling procedure proposed in the mergeLevels function is profile-specific, whereas our calling procedure integrates all profiles. Consequently, aliasing between common waves across profiles and specific mergeLevels aggregation decreases the global performance of calling when it is performed sample by sample. This trend is not true for joint segmentation, which shows that segmentation calling and normalization should be performed globally.

Then we examine the differences between normalization methods. Background intensities are supposed to be made of a sinusoidal trend plus noise, and the 3 proposed methods have very specific behaviors. The "position effect" catches both trend and noise, whereas the spline captures the sinusoidal trend only. The position effect corresponds to the strongest possible correction as it takes all information on the segmentation residuals, whereas splines are performant to recover smooth functions by denoising. This behavior is illustrated in the Figure 2 of the supplementary material available at *Biostatistics* online. Finally, the position-effect correction seems to be the best normalization method whatever the segmentation method.

6.2 *Profiles sharing common variations*

Motivations. In this second set of simulations, we consider the case where a given proportion of patients share some aberrations. Since the background intensity is a common effect across profiles, aliasing is likely to occur between background and common variations, especially when a high proportion of profiles are concerned by the variation. We use the setting proposed by Pique-Regi *and others* (2009) for this purpose. We consider that a proportion $\pi = (25, 50, 75, 100)\%$ of profiles shares 3 segments with length 1, 10, and 100, respectively. Each break point is sampled in $\mathcal{U}[-\eta, \eta]$ with $\eta = (0, 5, 10, 20)$, meaning that when $\eta = 0$, every break point is at the same location. SNR is defined as previously. Aliasing is uncovered when studying the estimated background corrected by the true reference trend, that is, $\widehat{F}(t) = \widehat{b}(t) - \sin(2\pi t/100)$ (bias of the background estimator).

Results. Figure 3 shows that jumps are captured by the background intensity when all profiles share the same aberrations. This trend is more important when the SNR is low, with small aberrations, and when the break positions match exactly ($\eta = 0$). This corresponds difficult configurations with strong aliasing between common waves and individual jumps. Figure 3 also shows the advantage of using a joint segmentation procedure compared with individual splits (like in CBS). As aliasing between common aberrations and common background intensity will be present in any joint procedure, the advantage of using joint segmentation lies in the joint control of the segmentation and background correction. When a systematic trend is present in the data (Figure 3 bottom panel with strictly common breaks) using joint segmentation allows us to capture it, whereas CBS (with normalization) does not globally control the shape captured by the background intensity when common breaks are present. Consequently, we need to develop to assess the part of shared aberrations that is embedded in the background intensity estimate.

Finding outliers in the background intensity. Our proposal is to detect common aberrations as exceptional values in the background intensity estimates. To do so, we take advantage of the 2 best normalization procedures. The spline method estimates the smooth trend of the background intensity, whereas the position effect β_t captures this trend plus a random noise that is linked to the background variance (Figure 2 of the supplementary Material available at *Biostatistics* online). This motivates the definition of a "corrected"

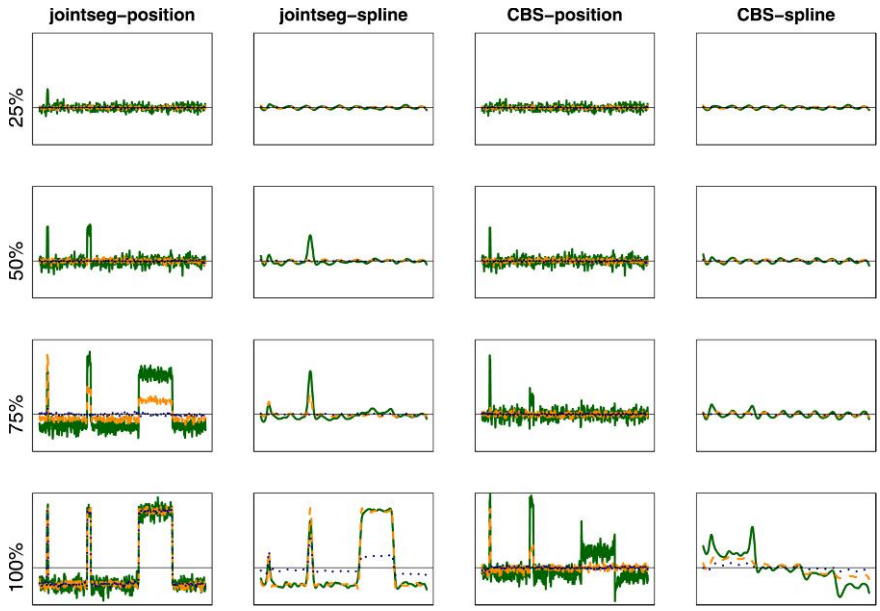

Fig. 3. Bias of the background estimator for proportions of patients sharing the segments at the exact same position ($\eta = 0$). *x*-axis: genomic position, *y*-axis: signal intensity (the plain horizontal line corresponds to the zero level). Only 3 values of SNR are represented: plain line (green) SNR = 1, dashed line (orange) SNR = 5, and dot line (blue) SNR = 20.

background intensity $\widehat{F}(t) = \widehat{\beta}_t - \widehat{b}_{\text{spline}}(t)$, whose outliers are likely to be common break points (like in Figure 3). This strategy is relevant only if we make the hypothesis that technological sources of artifact vary smoothly as a function of physical position. Then a natural estimator of the background variance τ is $\widehat{\tau}^2 = \sum_{t=1}^{n} \widehat{F}(t)^2/n$. This estimator shows excellent performance in terms of bias and MSE (not shown). Finally, we perform a position-wise test to compute *P*-values such that $p_v(t; \widehat{\tau}) = 1 - 2 \times \Phi(|\widehat{F}(t)|/\widehat{\tau})$, using an FDR adjustment with respect to the number of positions. This defines the excess of signal observed in the background compared with what would be expected in the background noise $\mathcal{N}(0, \tau^2)$. This procedure is illustrated in the following application section.

7. APPLICATIONS

In this last section, we provide an illustration of our procedure on the hapmap data set of Redon *and others* (2006) that can be downloaded at http://www.sanger.ac.uk/humgen/cnv/. We use the WGTP array CGH data and focus on chromosome 22. The method to detect common variants described above is illustrated in Figure 4. The spline-based normalization method enables to highlight the trend that exists in the segmentation residuals provided by the position-effect method. This represents an estimation of the wave effect in the background. Thus by substracting this trend, we can test whether exceptional values are observed in the background. The result is presented in Figure 4 (bottom panel), and all exceptional values correspond to confirmed CNVs (Table 2 of the supplementary Material available at *Biostatistics* online). Then we compare the performance of the joint segmentation procedure with CBS complemented with normalization. We use the database of genomic variants http://projects.tcag.ca/variation/ to validate the identified segments. We define a true positive as a position on the array that is declared to be a CNV

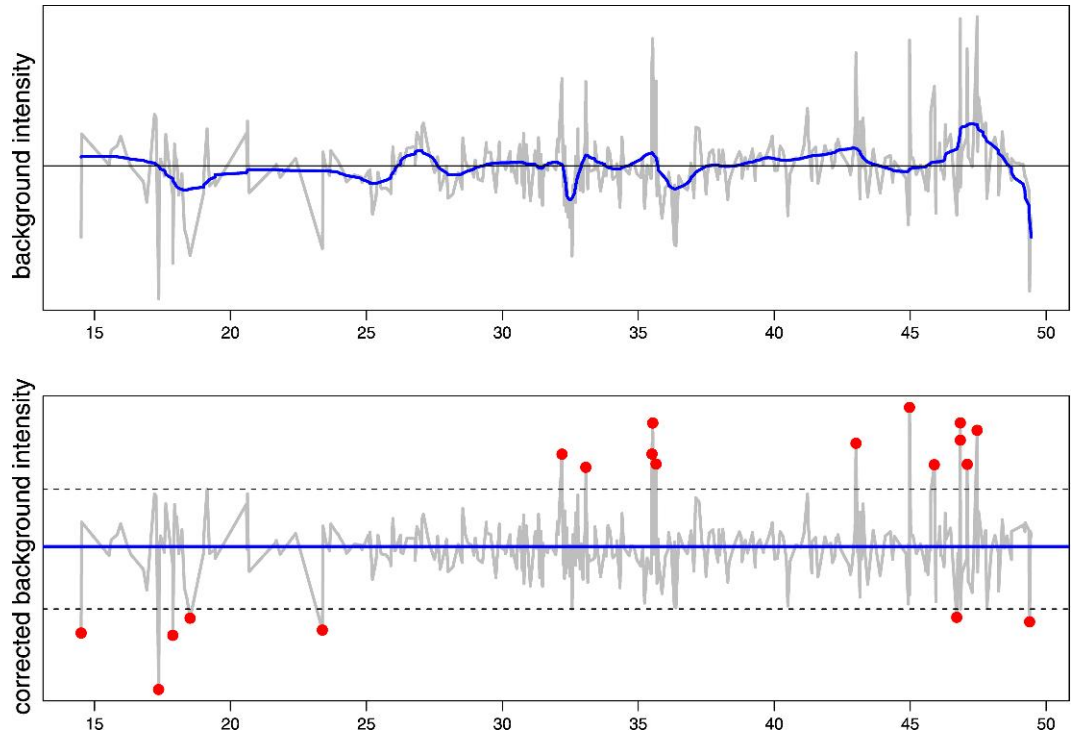

Fig. 4. Top panel: background intensity on chromosome 22 (gray) and estimated trend by the spline method (blue line). Bottom panel: normalized background intensity and exceptional values in the background (red points) with respect to empirical Gaussian quantiles based on an estimate of the background intensity $\widehat{\tau}$. All points are confirmed CNV.

at least in one profile and which has been confirmed in the database. A true negative is a position on the array that is never declared to be a CNV and which is not referenced in the database. This gives the following rates: $\text{TPR}_{\text{joint}} = 29\%$, $\text{TPR}_{\text{CBS}} = 38\%$, $\text{FPR}_{\text{joint}} = 0.9\%$, and $\text{FPR}_{\text{CBS}} = 52\%$. This important FPR for CBS is induced by the high level of segmentation that makes every position a potential FP due to aliasing with the wave effect. Lastly, we compare the speed of execution of both methods (Table 2 of the supplementary material available at *Biostatistics* online). While there is an advantage for CBS that is known to be very fast, the effect of normalization is very important for both CBS and joint segmentation as many iterations are required to identify a correct background intensity. However, the method can be run in parallel for every chromosome, which decreases the global computational burden.

8. Conclusion

In this article, we propose a statistical method for the joint analysis of multiple CGH profiles. This method is a generalization of the segmentation framework that has already shown excellent performance for single array analysis. Joint analysis is an opportunity to perform better calling (as the levels are learned on all profiles) and allows the estimation of a background intensity which catches the wave effect observed in many studies. We propose the first quantitative simulations to assess the performance of our method in the case of multiple profiles, which is inspired from Lai *and others* (2005) and Pique-Regi *and others* (2009).

We investigate the cases where common breaks are present or not, and we show that much information can be learned from the background intensity estimate in terms of common break points. We do not assess the question of recurrent aberrations. Our method rather corresponds to a first step toward the proper calling of multiple profiles that may be used by downstream procedures like presented in Rouveirol *and others* (2006) or in Robin and Stefanov (2009). Modeling perspectives of this work will concern the integration of possible heteroskedasticity between profiles to study heterogeneous sample contamination for instance. Batch effects could also be introduced if they were observed on several samples and if their correction was based on linear models. cghseg is not limited to CGH arrays and could be also used on single nucleotide polymorphism (SNP) arrays. However, the model should be enriched to perform allele-specific segmentation as proposed by Bengtsson *and others* (2008).

The 2 main research directions concern the algorithmic and the model selection parts. For the first one, it appears that even if the 2-stage DP procedure reduces the computational burden, the complexity remains in n^2 that limits the use of cghseg to very large signals (last generation of SNP arrays for instance). A future version of the method will integrate recent work on DP for very large signals that is under investigation (Rigaill, 2010). For the statistical part, the main direction is to derive a model selection criterion that will integrate the functional part of the model. In the first approximation, we proposed to neglect this part as the dimension of the model does not depend on the background intensity. However, it would be interesting to study in details the potential interactions between the jump process observed in each profile and the smooth background function that is common to all profile. This part is currently under investigation as well.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://www.biostatistics.oxfordjournals.org.

ACKNOWLEDGMENT

Conflict of Interest: None declared.

FUNDING

Institut National du Cancer (INCA) to G.R.

REFERENCES

BENGTSSON, H., IRIZARRY, R., CARVALHO, B. AND SPEED, T. P. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759–767.

BEN-YAACOV, E. AND ELDAR, Y. C. (2008). A fast and flexible method for the segmentation of aCGH data. *Bioinformatics* **24**, 139–145.

CARTER, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics* **39**, S16–S21.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

HUPE, P., STRANSKY, N., THIERY, J. P., RADVANYI, F. AND BARILLOT, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413–3422.

KORN, J. M., KURUVILLA, F. G., MCCARROLL, S. A., WYSOKER, A., NEMESH, J., CAWLEY, S., HUBBELL, E., VEITCH, J., COLLINS, P. J., DARVISHI, K. *and others* (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40**, 1253–1260.

LAI, W. R., JOHNSON, M. D., KUCHERLAPATI, R. AND PARK, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763–3770.

LIU, J., MOHAMMED, J. J. AND CARTER, R. S., KAHVECI, T. AND BAUDIS, M. (2006). Distance-based clustering of CGH data. *Bioinformatics* **22**, 1971–1978.

MARIONI, J. C., THORNE, N. P. AND TAVARE, S. (2006). BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics* **22**, 1144–1146.

NANNYA, Y., SANADA, M., NAKAZAKI, K., HOSOYA, N., WANG, L., HANGAISHI, A., KUROKAWA, M., CHIBA, S., BAILEY, D. K., KENNEDY, G. C. *and others* (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Research* **65**, 6071–6079.

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. AND WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572.

PARK, P. J. (2008). Experimental design and data analysis for array comparative genomic hybridization. *Cancer Investigation* **26**, 923–928.

PICARD, F., ROBIN, S., LAVIELLE, M., VAISSE, C. AND DAUDIN, J.-J. (2005). A statistical approach for CGH microarray data analysis. *BMC Bioinformatics* **6**, 27.

PICARD, F., ROBIN, S., LEBARBIER, E. AND DAUDIN, J.-J. (2007). A segmentation/clustering model of the analysis of array CGH data. *Biometrics* **63**, 758–766.

PIQUE-REGI, R., ORTEGA, A. AND ASGHARZADEH, S. (2009). Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. *Bioinformatics* **25**, 1223–1230.

RANCOITA, P. M., HUTTER, M., BERTONI, F. AND KWEE, I. (2009). Bayesian DNA copy number analysis. *BMC Bioinformatics* **10**, 1–19.

REDON, R. ISHIKAWA, S., FITCH, K. R., FEUK, L., PERRY, G. H., ANDREWS, T. D., FIEGLER, H., SHAPERO, M. H., CARSON, A. R. AND CHEN, W. *and others* (2006). Global variation in copy number in the human genome. *Nature* **444**, 444–454.

RIGAILL, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. *Technical Report*. arXiv:1004.0887v1.

ROBIN, S. AND STEFANOV, V. T. (2009). Simultaneous occurrences of runs in independent Markov chains. *Methodology and Computing in Applied Probability* **11**, 267–275.

ROUVEIROL, C., STRANSKY, N., HUPÉ, PH., LA ROSA, PH., VIARA, E., BARILLOT, E. AND RADVANYI, F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics* **22**, 849–856.

RUEDA, O. M. AND DIAZ-URIARTE, R. (2009). Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously. *BMC Bioinformatics* **10**, 308.

SHAH, S. P. (2008). Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenetic and Genome Research* **123**, 343–351.

SHAH, S. P., CHEUNG, K. J., JOHNSON, N. A., ALAIN, G., GASCOYNE, R. D., HORSMAN, D. E., NG, R. T. AND MURPHY, K. P. (2009). Model-based clustering of array CGH data. *Bioinformatics* **25**, i30–i38.

SNIJDERS, A. M., NOWAK, N., SEGRAVES, R., BLAKWOOD, S., BROWN, N., CONROY, J., HAMILTON, G., HINDLE, A. K., HUEY, B., KIMURA, K. *and others* (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29**, 263–264.

STJERNQVIST, S., RYDEN, T., SKOLD, M. AND STAAF, J. (2007). Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics* **23**, 1006–1014.

VAN DE WIEL, M., BROSENS, R., EILERS, P. H. C., KUMPS, C., MEIJER, G. A., MENTEN, B., SISTERMANS, E., SPELEMAN, F., TIMMERMAN, M. E. AND YLSTRA, B. (2009). Smoothing waves in array CGH tumor profiles. *Bioinformatics* **25**, 1099–1104.

VAN DE WIEL, M. A., KIM, K. I., VOSSE, S. J., VAN WIERINGEN, W. N., WILTING, S. M. AND YLSTRA, B. (2007). CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics* **23**, 892–894.

VAN DE WIEL, M. A., PICARD, F., VAN WIERINGEN, W. N. AND YLSTRA, B. (2010). Preprocessing and downstream analysis of microarray DNA copy number profiles. *Briefings in Bioinformatics*.

VAN WIERINGEN, W. N., VAN DE WIEL, M. AND YLSTRA, B. (2008). Weighted clustering of called aCGH data. *Biostatistics* **9**, 484–500.

WILLENBROCK, H. AND FRIDLYAND, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084–4091.

WU, L. Y., CHIPMAN, H. A., BULL, S. B., BRIOLLAIS, L. AND WANG, K. (2009). A Bayesian segmentation approach to ascertain copy number variations at the population level. *Bioinformatics* **25**, 1669–1679.

ZHANG, N. (2005). Change-point detection and sequence alignment: statistical problems of genomics, [PhD. Thesis]. Standford, CA: Standford University.

ZHANG, N. R., SENBABAOGLU, Y. AND LI, J. Z. (2010). Joint estimation of DNA copy number from multiple platforms. *Bioinformatics* **26**, 153–160.

ZHANG, N. R. AND SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63**, 22–32.

[*Received May 4, 2010; revised October 4, 2010; accepted for publication November 12, 2010*]