Biostatistics (2006), **7**, 3, *pp.* 399–421
doi:10.1093/biostatistics/kxj015
Advance Access publication on January 9, 2006

A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations

DAVID A. ENGLER

Department of Biostatistics, Harvard University, 655 Huntington Avenue, Boston, MA 02115, USA
engler@fas.harvard.edu

GAYATRY MOHAPATRA

Massachusetts General Hospital, Department of Pathology, CNY-7015, 149 13th Street, Charlestown, MA 02129, USA

DAVID N. LOUIS

Molecular Neuro-Oncology Laboratory and Molecular Pathology Unit, Departments of Pathology, Cancer Center and Neurosurgical Service, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

REBECCA A. BETENSKY*

Department of Biostatistics, Harvard University, 655 Huntington Avenue, Boston, MA 02115, USA
betensky@hsph.harvard.edu

SUMMARY

DNA sequence copy number has been shown to be associated with cancer development and progression. Array-based comparative genomic hybridization (aCGH) is a recent development that seeks to identify the copy number ratio at large numbers of markers across the genome. Due to experimental and biological variations across chromosomes and hybridizations, current methods are limited to analyses of single chromosomes. We propose a more powerful approach that borrows strength across chromosomes and hybridizations. We assume a Gaussian mixture model, with a hidden Markov dependence structure and with random effects to allow for intertumoral variation, as well as intratumoral clonal variation. For ease of computation, we base estimation on a pseudolikelihood function. The method produces quantitative assessments of the likelihood of genetic alterations at each clone, along with a graphical display for simple visual interpretation. We assess the characteristics of the method through simulation studies and analysis of a brain tumor aCGH data set. We show that the pseudolikelihood approach is superior to existing methods both in detecting small regions of copy number alteration and in accurately classifying regions of change when intratumoral clonal variation is present. Software for this approach is available at http://www.biostat.harvard.edu/∼betensky/papers.html.

*To whom correspondence should be addressed.

1. INTRODUCTION

DNA sequence copy number changes in tumor cells are associated with cancer development and progression (Forozan *et al.*, 2000; Kallioniemi *et al.*, 1994; Tirkkonen *et al.*, 1998). Copy number changes are typically manifest as losses or gains of chromosomes or chromosomal regions. A chromosomal loss in a tumor is classically associated with underexpression of genes whose activity prevents tumor development, so called tumor suppressor genes. Copy number gains, on the other hand, are often associated with overexpression of genes that promote cell growth, so called oncogenes. Detection and mapping of such gains and losses in tumor genomes will ultimately lead to the identification of critical genes associated with these diseases, and eventually to improved therapeutic approaches.

Initial attempts to identify regions of copy number variation included karyotyping and directed molecular genetic studies. The inability of these methods to characterize chromosomal changes in detail (karyotyping) or to examine large regions of the genome (directed molecular methods) led to the development of more global screening approaches, such as comparative genomic hybridization (CGH) (Kallioniemi *et al.*, 1992). CGH entails the simultaneous hybridization of differentially labeled test DNA and reference DNA to normal chromosomal spreads. Test DNAs are obtained from tumors and reference DNA is typically obtained from lymphocytes of a healthy individual. Two separate fluorescent dyes (usually red-fluorescent dye Cy5 for the reference sample and green-fluorescent dye Cy3 for the tumor sample) are used to label probes for hybridization to metaphase chromosomes, thus allowing a fluorescence intensity ratio to be calculated for each approximate location on a particular chromosome.

Array-based comparative genomic hybridization (aCGH) is a recent modification of CGH that provides greater resolution by using microarrays of DNA fragments rather than metaphase chromosomes (Pinkel *et al.*, 1998; Snijders *et al.*, 2001). These arrays can be generated with different types of DNA preparations. One method uses bacterial artificial chromosomes (BACs), each of which consists of a 100- to 200-kilobase DNA segment. Other arrays are based on cDNAs (Pollack *et al.*, 1999, 2002) or oligonucleotide fragments (Lucito *et al.*, 2000). As in CGH analysis, the resultant map of gains and losses is a result of calculating fluorescence ratios. By arraying large numbers of DNA sequences, however, one can potentially use aCGH to determine gains and losses with high resolution (e.g. at an individual gene level) across the entire genome (e.g. with arrays that 'tile' the whole genome).

The determination of patterns of genomic gains and losses translates into a number of possible uses in cancer diagnosis and management. For instance, in a group of patients diagnosed with the same pathological type of cancer, genetic subtyping can predict markedly different responses to chemotherapies and offer powerful prognostic information. aCGH could therefore prove to be a powerful tool for investigating associations between genetic alterations and pathological diagnosis, response to individual therapies, and clinical outcome. For example, a recent study has as its focus the evaluation of copy number gains and losses in tissue taken from malignant gliomas, a common type of human brain tumor (Mohapatra *et al.*, 2006). Figure 1 displays the aCGH data for four tumors from this study.

A number of statistical methods have been proposed for the analysis of aCGH data. Such methods have taken one of the two approaches. One approach prespecifies the types of underlying, but unobserved, copy number alteration events (e.g. gain, loss, no-change). The second approach makes no such assumptions, but attempts to identify locations of $\log_2$ ratio mean change (i.e. breakpoints) and to estimate the value of those means. A major limitation of all currently proposed methods is their analysis of each chromosome, of each hybridization, separately. That is, current methods do not utilize the entire available data set to estimate the features that are shared in common among chromosomes and hybridizations. Additionally, some methods fail to account for dependence between clones. Finally, the interpretation of results obtained from many existing methods does not emerge naturally from the original analysis, but requires a second stage of analysis and additional assumptions.

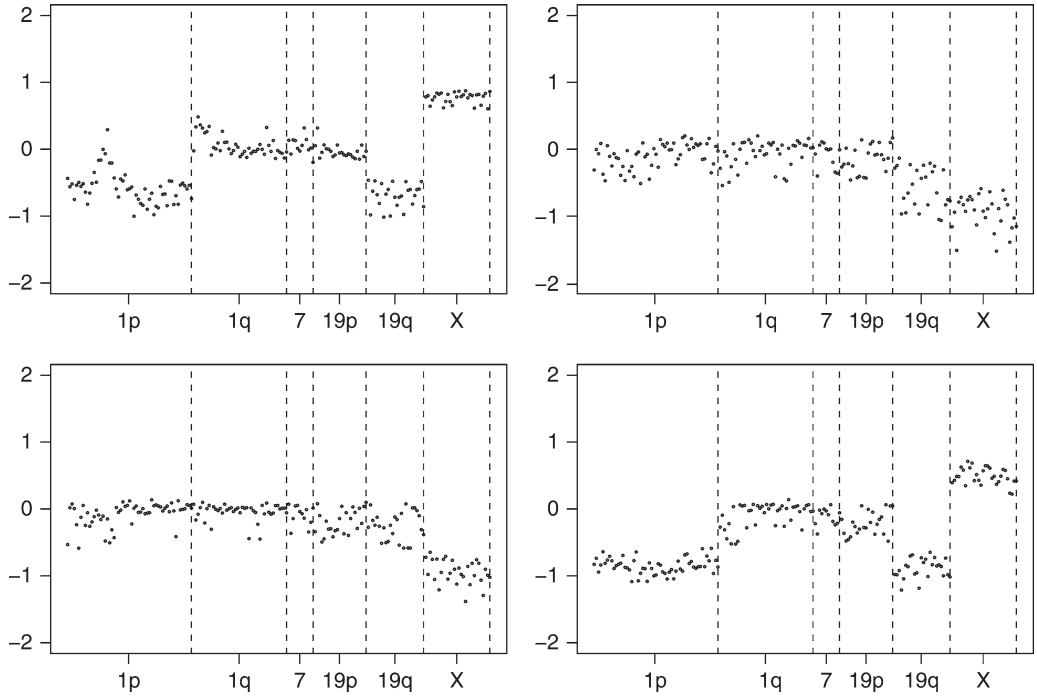

Fig. 1. The $\log_2$ fluorescence ratios from four oligodendroglioma samples.

In this paper, we propose a joint analysis of aCGH data, via a Gaussian mixture model with a local hidden Markov dependence structure, that fully exploits the entire data set to efficiently estimate the common features of the genetic alterations. At the same time, our approach accommodates the variability within chromosomes, between chromosomes, and between hybridizations. For computational feasibility, we use a pseudolikelihood function as the basis for estimation. As it is based on a probability model, our approach produces quantitative assessments of the likelihood of various genetic events at each clone, via posterior probabilities, which aid in interpretation of results. In addition, the method produces a graphical display of this information. In both real and simulated data sets, the proposed pseudolikelihood approach appears to be more sensitive than existing methods to small segments of true copy number change. Additionally, simulation studies show that when variation in gain and loss levels exists within hybridizations, the proposed method more effectively classifies regions of change.

In Section 2, we provide a brief outline of aCGH data characteristics. Section 3 reviews and assesses existing methods of aCGH data analysis. In Section 4, we present the theoretical framework for our method along with proposed methods for identification of gains and losses and for identification of copy number breakpoint locations. We apply our proposed method to the glioma study and present the results in Section 5. Section 6 presents the results of several simulation studies. We conclude with a discussion in Section 7.

2. aCGH DATA CHARACTERISTICS AND ASSUMPTIONS

It is of interest to estimate the true sequence of underlying copy number ratios that generated the observed sequence of fluorescence ratios. In the idealized setting in which the reference DNA does not contain

any copy number alterations, the genomic regions of copy number gains will have ratios greater than or equal to $3/2$ (i.e. $\log_2(3/2) \geqslant 1.58$) and regions of copy number loss should have ratios equal to $1/2$ (i.e. $\log_2(1/2) = -1$). Logarithms of the ratios are commonly used because the ratios are dependent on absolute magnitude and are often highly skewed; logged intensities often provide a better sense of the true variation (Amaratunga and Cabrera, 2004). Allowing for random error, often assumed to be normal with mean zero, the underlying ratio of the true copy number of each test (tumor) clone to the reference clone can then be estimated given the observed data.

A variety of biological and experimental factors cause the observed $\log_2$ ratios to deviate from their theoretical values. For one, the copy number of the reference sample might not be two (Iafrate *et al.*, 2004). Also, the test sample is not pure; the tumor sample is typically contaminated with some normal cells. In the event of true copy number alteration, the presence of normal cells (exhibiting no copy number change) lowers the absolute magnitude of the $\log_2$ ratios. Deviation from the expected ratios might also be due to intratumoral clonal variation (Okada *et al.*, 2003). That is, the mean $\log_2$ ratio magnitudes corresponding to a particular copy number alteration event (e.g. single copy number gain) may vary both within and between chromosomes in a given tumor.

Normalization of the $\log_2$ ratios is typically conducted in an attempt to adjust for sources of systematic variation. Since these effects are often not known or measured, most aCGH methodologies incorporate global normalization techniques, centering the data about the sample mean or median for a given hybridization (Fridlyand *et al.*, 2004). An alternative approach is available when a tumor type has not been shown to have copy number gain or loss in a given chromosome by other laboratory methods. In this case, that chromosome can be used for normalization by using its median $\log_2$ ratio to center all $\log_2$ ratios on that hybridization. More refined approaches have been suggested as well; for example, Dudoit *et al.* (2002) and Yang *et al.* (2002) proposed methods of normalization that adjust for spatial effects on the array. Such methods, however, may or may not be appropriate for aCGH data. In summary, normalization remains imperfect and accurate estimation of the copy number is unlikely. It is assumed, however, that changes in the observed, normalized $\log_2$ ratios correspond directly to changes in the true copy numbers.

3. CURRENT METHODS OF SINGLE-CHROMOSOME ANALYSIS

All current aCGH analysis methods treat each chromosome of each hybridization separately. This is motivated by the presumed independence of chromosomes and hybridizations. Also, this acknowledges that the magnitudes of the $\log_2$ ratios for loss or gain may vary across chromosomes and hybridizations (e.g. Fridlyand *et al.*, 2004). Current methods can be classified as those that prespecify the number and type of gain and loss events on a chromosome or as those that identify regions of common $\log_2$ ratio magnitude. In this section, we review these methods and summarize their limitations.

3.1 *Current methods: prior specification of underlying events*

One class of analytic approaches is predicated on *a priori* specification of the underlying, but unobserved, copy number event types. To date, all such methods have specified three possible events: gain, loss, and no-change. One example of this is a threshold approach (e.g. Pollack *et al.*, 1999; Weiss *et al.*, 2003; Aguirre *et al.*, 2004). This involves estimation of the variability of the $\log_2$ ratios that correspond to no genetic alteration using normal–normal hybridizations. Assuming, then, that $\log_2$ ratios corresponding to no genetic alteration in the tumor are normally distributed with mean zero, regions of gain and loss are identified as those clones with $\log_2$ ratios in the upper and lower tails of the distribution. A second example is a full mixture model approach in which all parameters governing the distributions of the $\log_2$ ratios

associated with each of the different copy number events are estimated. For example, Hodgson *et al.* (2001) assumed that the data from a given chromosome for a given hybridization are distributed as a three-component (loss, no-change, and gain) Gaussian mixture. As in the simpler threshold approach based on the normal–normal hybridizations, the variability of the no-change event is the parameter of interest. The mixture model approach differs from the threshold approach in that normal–normal hybridizations are not required. Both approaches assume independence between clones.

3.2 *Current methods: segmentation*

The second class of analytic approaches does not prespecify the underlying copy number events on a given chromosome, but rather focuses on the identification of segments of common $\log_2$ ratio mean. These segmentation methods seek to identify locations of $\log_2$ ratio mean change (i.e. change points or breakpoints) and to estimate the values of those means.

Olshen *et al.* (2004) proposed a circular binary segmentation (CBS) algorithm that identifies the change points through successive comparison of segments of the chromosome and with evaluation of local significance via permutation. This is followed by a pruning algorithm to control the number of change points identified. Their method of permutation assumes independence between clones. Picard *et al.* (2004) proposed a likelihood-based method to identify the change points for the sequence of $\log_2$ ratios. This method assumes the ratios on a given segment to be independent and normally distributed with a common, unspecified, mean. They employed a dynamic programming approach to determine the location of change points. They also incorporated penalization to limit the number of estimated change points. Hupe *et al.* (2004) likewise presented a Gaussian-based likelihood approach (GLAD). Change points are identified through the use of a piecewise constant regression model that employs adaptive weights smoothing in which the errors are assumed to be independent and normally distributed. To explicitly account for the physical dependence between clones, Fridlyand *et al.* (2004) proposed a discrete-state hidden Markov model (HMM) approach. This model assumes that given the genetic states at all 'previous' clones, the genetic state at a given clone depends only on the true state at the immediately previous clone. Of note, the 'states' in the Fridlyand *et al.* (2004) approach are not underlying copy number events such as gain and loss (as described in Section 3.1), but are segments of common mean. A change in state corresponds to a breakpoint. Their approach differs from the CBS (Picard *et al.*, 2004) and GLAD approaches in that instead of allowing copy number segments to assume any mean value, the maximum number of distinct mean values on each chromosome is required to be preselected. They estimate the number of mean levels on each chromosome through a penalized likelihood approach (Akaike Information Criteria or Bayesian Information Criteria) followed by an algorithm that potentially merges adjacent segments. Following segmentation, they identify genetic features such as focal aberrations and amplifications (low- and high-level alteration within a segment involving a small number of clones, respectively) and outliers. Wang *et al.* (2005) proposed the use of an agglomerative clustering technique (cluster along chromosomes) to identify change points. Their technique iteratively clusters clones, and subsequently groups of clones, according to a measure of relative difference between adjacent clones/clusters. They classify clusters of two or more clones as 'interesting' according to the number of clones in the cluster, the mean $\log_2$ ratio value of the cluster, and the value of the relative difference between adjacent clones within the cluster.

All these segmentation methods provide breakpoint locations but do not identify the associated genomic alterations as gains or losses. Because a primary objective of aCGH analysis is to identify regions of copy number gain and loss, follow-up methods have been proposed for this from segmentation results. First, some authors have used a nonparametric estimate of the standard deviation to identify a global threshold for categorizing segments (Paris *et al.*, 2004; Rossi *et al.*, 2005). The threshold is obtained by calculating the median absolute deviation for each identified segment and using the median of the median absolute deviations (MMAD) as an estimate of the standard deviation. Clones on segments whose

magnitude exceeds a multiple of the MMAD are classified as gains and clones on segments whose magnitude is less than a multiple of the MMAD are classified as losses.

A second approach to gain and loss identification based on segmentation results entails the combination of identified segments across chromosomes and a subsequent establishment of a no-change baseline. Hupe *et al.* (2004) outline a method, GLADmerge, for combining segments obtained from GLAD, first within and then across chromosomes through hierarchical clustering in which clusters of segments are identified from the resultant dendrograms. The super-cluster with magnitude closest to zero is labeled as the no-change baseline. The remaining clusters are labeled as gain or loss. Willenbrock and Fridlyand (2005) proposed an alternative approach, MergeLevels, which can be applied to a number of different segmentation methods (e.g. CBS, HMM). Segments on a given hybridization are combined (assigned the same magnitude or level) if (1) the distributions of the $\log_2$ ratios on each of the two segments are not significantly different from one another or (2) the difference in segment mean magnitudes falls below a dynamically determined threshold. Following segment combination, the segment level with magnitude closest to zero is identified as the baseline. All other levels are identified as regions of copy number gains or losses.

3.3 *Current methods: limitations*

A major limitation of all currently proposed methods is their separate analysis of each chromosome and each hybridization. These methods do not utilize all the available information to estimate the features that are shared in common among chromosomes and hybridizations. For this reason, they are likely to be highly inefficient.

An additional limitation of some methods is their failure to account for dependence between clones. Problems that arise from assumptions of independence are particularly apparent in methods that prespecify the number and type of underlying copy number events (i.e. Section 3.1). In such approaches, individual clones whose absolute magnitude exceeds a given magnitude are classified as gains or losses regardless of the behavior of adjacent clones. Hence, such methods can result in a high misclassification rate.

The methods in Section 3.1 may also suffer from numerical difficulties in the estimation procedure if the prespecified copy number events such as gain or loss occur infrequently in a particular chromosome. To avoid these difficulties, subjective input on a chromosome by chromosome basis would be required, which is also an undesirable feature. Another limitation of these methods is that they do not account for intratumoral clonal variation. The use of thresholds entails the assumption that levels of gain and loss are constant across the chromosome. If, therefore, there is a range of levels of gain and loss due to clonal variation within the chromosome, the estimated levels of gain and loss might be biased toward zero (i.e. no-change).

Segmentation methods (Section 3.2) also have a number of limitations. First, many of the proposed segmentation methods require the use of *ad hoc* algorithms to prune or combine segments. The sensitivity of segmentation methods is often dependent upon user-defined values in such algorithms. Second, interpretation of segmentation results is problematic. That is, once the change points are identified and associated mean $\log_2$ ratio levels are estimated, it is not at all clear which segments of common mean represent real genetic alterations (i.e. copy number gain or loss) and which are simply due to experimental variability. The thresholding (MMAD) or pruning and combination (GLADmerge and MergeLevels) second-stage methods have been developed to overcome this drawback. However, these methods must be employed subsequent to the initial segmentation and are not natural extensions of the original modeling. In fact, classification of gain and loss that is accomplished through these secondary methods seems to ignore the variability both within and between chromosomes that motivated the initial use of segmentation (see Fridlyand *et al.*, 2004). Such methods do appear to have difficulty when there is variation in copy number levels between chromosomes (see Section 6).

4. PSEUDOLIKELIHOOD APPROACH

In this section, we outline the theoretical framework for our estimation procedure. We additionally present methods for classification of gains and losses and for identification of copy number breakpoint locations.

4.1 *Notation and pseudolikelihood function*

Let $x_{im} = (x_{im1}, \ldots, x_{imJ})$ denote the $\log_2$ ratios for hybridization i on chromosome m for clones $1, \ldots, J$, where J may vary across chromosomes and hybridizations. Assuming independence across chromosomes and hybridizations, the likelihood of the data is

$$\prod_i \prod_m P(x_{im}|\boldsymbol{\lambda}) = \prod_i \prod_m P(x_{im1}, \ldots, x_{imJ}|\boldsymbol{\lambda}), \tag{4.1}$$

where $\boldsymbol{\lambda}$ denotes the array of parameters governing the distribution of x_{im}.

It is plausible that the dependence between clones is local and that clones that are separated by sufficient distances can be reasonably regarded as independent. Of course, specific patterns of dependence across the genome are complex. For example, in examining patterns of recombination, Gabriel *et al.* (2002) found that haplotypes (particular combinations of alleles) on different blocks can be considered independent, whereas those on the same block cannot. However, due to the complexity of such blocks (e.g. varying length of blocks, differences in patterns across populations), incorporation of these patterns into a model is difficult.

As a close approximation to the truth and for computational simplicity, we propose a local dependence model. In the setting of spatial data that also exhibited local dependence, Besag (1975) suggested the use of a pseudolikelihood that exploited this dependence structure and treated data points separated by a given distance as independent. Besag proved consistency of the maximum pseudolikelihood estimates, and thus provided a theoretical foundation for the approach. Efficiency of the method was examined by Lindsay (1988). Several authors employed a similar approach in situations of bivariate association (e.g. Clayton, 1978; Oakes, 1986). Recently, Cox and Reid (2004) derived the conditions under which consistent estimators are obtained when using marginal and bivariate probabilities to form estimating functions from a 'pairwise likelihood'.

We likewise propose the use of a pseudolikelihood as an approximation to the full likelihood (4.1). In contrast to many previous applications of pseudolikelihood that used a pairwise (bivariate) likelihood, we construct a pseudolikelihood as the product of overlapping triples of $\log_2$ ratios: $\prod_i \prod_m \prod_{j=2}^{J-1} P(x_{im(j-1)}, x_{imj}, x_{im(j+1)}|\boldsymbol{\lambda})$. Note that each $\log_2$ ratio contributes to the pseudolikelihood three times: once as x_{j-1}, once as x_j, and once as x_{j+1}.

As suggested by previous authors, a Gaussian finite-mixture model provides a natural formulation of the distribution of the $\log_2$ ratios:

$$\prod_i \prod_m \prod_{j=2}^{J-1} \sum_{\mathbf{C}} P(x_{im(j-1)}, x_{imj}, x_{im(j+1)}|\boldsymbol{\lambda}, \mathbf{C}) P(\mathbf{C}|\boldsymbol{\lambda}), \tag{4.2}$$

where the vector $\mathbf{C}$ contains the set of unobserved states (C_{j-1}, C_j, and C_{j+1}) corresponding to the $\log_2$ ratios, $x_{im(j-1)}$, x_{imj}, and $x_{im(j+1)}$. For aCGH data, possible states include events such as copy number gain, copy number loss, and no-change (no gain and no loss). The summation over $\mathbf{C}$, then, represents the summation over all possible state combinations for these three $\log_2$ ratios. Although the underlying state of a given $\log_2$ ratio is obviously fixed, regardless of whether the data point is at position $j-1$, j, or $j+1$, this is not enforced in (4.2). This is due to our introduction of the mixture model at the level of the trivariate probability distributions, rather than at the level of the full joint distribution. We have done this

to avoid high-dimensional numerical integration and examine its impact on parameter estimation in our simulation studies (Section 6).

The array $\boldsymbol{\lambda}$ in (4.2) is expanded from the array in (4.1) to include the parameters governing the mixture distribution, including initial-state probabilities and transition probabilities (e.g. Rabiner, 1989). We assume that it is constant across all hybridizations, chromosomes, and clones. We also assume that the $\log_2$ ratios corresponding to a particular state are normally distributed with a state-specific mean and a common variance. Hence, for a Gaussian three-state model consisting of copy number loss (L), no-change (0), and copy number gain (G), $\boldsymbol{\lambda}$ includes the mean and variance parameters governing the distributions of $\log_2$ ratios (μ_L, μ_0, μ_G, σ), the vector of initial-state probabilities ($\boldsymbol{\pi}$), and the array of transition probabilities (A).

Due to intratumoral clonal variation, it is not reasonable to assume that the mean levels of gain or loss are constant within or between chromosomal segments. To account for the variation in mean levels of $\log_2$ ratios corresponding to a particular copy number event, C, we assume that each small neighborhood of three clones has a random mean level, γ_C, and that the $\log_2$ ratios generated from that particular state, C, in that neighborhood have mean, γ_C, and variance σ^2. We assume that the γ_C are drawn from a normal distribution with mean μ_C and variance σ_C^2. For identifiability, we assume that the random effects, γ_L, γ_0, and γ_G, follow truncated normal distributions such that $\gamma_L < -\varepsilon < \gamma_0 < \varepsilon < \gamma_G$. In our analyses of real and simulated data, our estimation algorithm converged for $\varepsilon \in (0.1, 0.4)$ and the results were not sensitive to the actual choice of ε within that range. Under this random-effects model, the pseudolikelihood for the finite-mixture model (4.2) is modified to be

$$\prod_i \prod_m \prod_{j=2}^{J-1} \sum_{\mathbf{C}} \int_{\gamma} \frac{P(x_{im(j-1)}, x_{imj}, x_{im(j+1)}|\boldsymbol{\lambda}, \mathbf{C}, \boldsymbol{\gamma}) P(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathbf{C}) P(\mathbf{C}|\boldsymbol{\lambda})}{P(\gamma_L < -\varepsilon) P(|\gamma_0| < \varepsilon) P(\gamma_G > \varepsilon)} \, \mathrm{d}\boldsymbol{\gamma}, \tag{4.3}$$

where the random vector $\boldsymbol{\gamma}$ contains the random means for loss, no-change, and gain (γ_L, γ_0, γ_G). The parameter array $\boldsymbol{\lambda}$ is now expanded to include the parameters governing the distributions of the random effects: $\mu_L, \sigma_L, \mu_0, \sigma_0, \mu_G, \sigma_G$.

We assume that conditional on the underlying states, $\mathbf{C}$, and random mean levels γ, the x_{imj} are independent. Hence,

$$P(x_{im(j-1)}, x_{im(j)}, x_{im(j+1)}|\boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{C}) = \prod_{r=-1}^{1} P(x_{im(j+r)}|\boldsymbol{\lambda}, C_r, \boldsymbol{\gamma}) = \prod_{r=-1}^{1} \frac{1}{\sigma} \phi\left(\frac{x_{im(j+r)} - \gamma_{c_r}}{\sigma}\right).$$

We assume also that the random-state means, γ, are mutually independent and independent of $\mathbf{C}$, so that

$$\begin{aligned} P(\gamma|\boldsymbol{\lambda}, \mathbf{C}) &= P(\gamma_L, \gamma_0, \gamma_G|\boldsymbol{\lambda}, \mathbf{C}) = P(\gamma_L|\boldsymbol{\lambda}) P(\gamma_0|\boldsymbol{\lambda}) P(\gamma_G|\boldsymbol{\lambda}) \\ &= \frac{1}{\sigma_L} \phi\left(\frac{\gamma_L - \mu_L}{\sigma_L}\right) \frac{1}{\sigma_0} \phi\left(\frac{\gamma_0 - \mu_0}{\sigma_0}\right) \frac{1}{\sigma_G} \phi\left(\frac{\gamma_G - \mu_G}{\sigma_G}\right). \end{aligned}$$

Finally, we assume that the unobserved states follow a one-step Markov process, so that

$$\begin{aligned} P(C_{j-1}, C_j, C_{j+1}|\boldsymbol{\lambda}) &= P(C_{j-1}|\boldsymbol{\pi}, \mathbf{A}) P(C_j|C_{j-1}, \boldsymbol{\pi}, \mathbf{A}) P(C_{j+1}|C_j, \boldsymbol{\pi}, \mathbf{A}) \\ &= \pi_{c_{j-1}} a_{c_j, c_{j-1}} a_{c_{j+1}, c_j}, \end{aligned}$$

where $a_{c,c'}$ denotes the probability of transitioning from state C' to state C and π_c denotes the marginal probability of state C.

Following some algebra and a change of variables, closed-form evaluation of the integral in (4.3) is possible, and the pseudolikelihood is reexpressed as

$$\prod_i \prod_m \prod_{j=2}^{J-1} \sum_C \frac{a_{c_{j+1},c_j} a_{c_j,c_{j-1}} \pi_{c_{j-1}} d_c(x_{im(j-1)}, x_{imj}, x_{im(j+1)}, \boldsymbol{\lambda}, \varepsilon)}{\Phi\left(\frac{-\varepsilon-\mu_L}{\sigma_L}\right) 2\left(1-\Phi\left(\frac{\varepsilon-\mu_0}{\sigma_0}\right)\right)\left(1-\Phi\left(\frac{\varepsilon-\mu_G}{\sigma_G}\right)\right)}, \tag{4.4}$$

where $d_c(\cdot)$ is a function specific to each possible triple of hidden state combinations, $\mathbf{C}$, associated with x_{j-1}, x_j, x_{j+1}. It is a function of the data, the value of ε, and the parameters, μ_L, μ_0, μ_G, σ, σ_L, σ_0, σ_G, and σ. The model parameters can be estimated by maximizing (4.4) using a simple iterative optimization technique, such as Newton–Raphson.

4.2 *Posterior probabilities and classification of clones*

Following the estimation of parameters, it is of interest to estimate the posterior probabilities of the underlying states given the observed data. One possible approach is to calculate $P(C_{imj} = c|x_{imj})$. However, this may be inaccurate for outlying $\log_2$ ratios. An alternative approach that may diminish the influence of outliers is to smooth the data by calculating the probability of a given state conditional on the corresponding and adjacent $\log_2$ ratios:

$$P(C_{imj} = c|x_{im(j-1)}, x_{imj}, x_{im(j+1)}). \tag{4.5}$$

Since $P(x_{j-1}, x_j, x_{j+1}|C_j = c)$ is equal to

$$\sum_{C_{j-1}} \sum_{C_{j+1}} P(x_{j-1}, x_j, x_{j+1}|C_{j-1}, C_j = c, C_{j+1}) P(C_{j-1}, C_{j+1}|, C_j = c),$$

the probability in (4.5) can be rewritten, by Bayes' formula, as

$$\frac{\sum_{C_{j-1}} \sum_{C_{j+1}} P(x_{j-1}, x_j, x_{j+1}|C_{j-1}, C_j = c, C_{j+1}) P(C_{j-1}, C_{j+1}|C_j = c) P(C_j = c)}{\sum_{C_{j-1}} \sum_{C_j} \sum_{C_{j+1}} P(x_{j-1}, x_j, x_{j+1}|C_{j-1}, C_j, C_{j+1}) P(C_{j-1}, C_j, C_{j+1})}.$$

This quantity is calculated based on the assumptions of the model in conjunction with the parameter estimates. Individual clones could then be assigned to the state corresponding to the highest of the three 'classification probabilities':

$$P(C_j = L|x_{j-1}, x_j, x_{j+1}), \quad P(C_j = 0|x_{j-1}, x_j, x_{j+1}), \quad P(C_j = G|x_{j-1}, x_j, x_{j+1}).$$

In addition to the posterior probability given in (4.5), two alternative posterior probabilities could be calculated for each clone j: $P(C_j = c|x_{j-2}, x_{j-1}, x_j)$ and $P(C_j = c|x_j, x_{j+1}, x_{j+2})$. These suggest an additional 'smoothing' of the data by averaging the three posterior probabilities and, likewise, an additional option for classification via the averaged probabilities. In our analysis of the glioma study and in our simulation studies, we classified individual clones using the averaged posterior probabilities.

4.3 *Identification of breakpoints*

Ultimately it is also of interest to identify the locations at which copy number transitions are likely to have occurred. Locations of such breakpoints are, in part, visually identifiable through examination of plots of probabilities described in Section 4.2. Through inspection of such plots, researchers are provided with a measure of the strength of evidence there is for each copy number transition.

Results from the pseudolikelihood method could also be used to more formally identify these breakpoints. One approach entails the calculation of the probabilities that, conditional on the data triplet x_{j-1}, x_j, and x_{j+1}, the state of the clone C_j differs from either of the states of the adjacent clones C_{j-1} and C_{j+1}. These are given by

$$P(C_{j-1} \neq C_j \neq C_{j+1} | x_{j-1}, x_j, x_{j+1}),$$
$$P(C_{j-1} = C_j \neq C_{j+1} | x_{j-1}, x_j, x_{j+1}),$$
$$P(C_{j-1} \neq C_j = C_{j+1} | x_{j-1}, x_j, x_{j+1}).$$

The maximum of the three probabilities could be plotted on the graphical display or it could be used in conjunction with a threshold to identify high-probability breakpoint locations.

This approach for breakpoint detection might not detect a breakpoint between contiguous segments of low-level copy gain and high-level copy gain if there is approximately equal evidence for both (i.e. the posterior probabilities are comparable). In this case, the algorithm would identify the union of the contiguous segments as a single segment of gain. If it were important to identify breakpoints of contiguous segments of this sort, the model could be expanded to include a fourth state of high-level gain. Alternatively, a breakpoint detection algorithm that simultaneously examined the posterior probabilities of state changes along with posterior mean estimates of the local mean level of gain, i.e. γ_G, could be developed.

4.4 *Software*

Analyses were performed using the R software package (http://www.r-project.org). The R implementation of the pseudolikelihood approach presented in this paper is available at http://www.biostat.harvard.edu/~betensky/papers.html.

5. GLIOMA STUDY

Classification of human gliomas based on molecular genetic alterations has already achieved clinical relevance (Cairncross *et al.*, 1998; Smith *et al.*, 2000, 2001; Sasaki *et al.*, 2002; Nutt *et al.*, 2003; van den Bent *et al.*, 2003). Among the major subtypes of gliomas, oligodendrogliomas are distinguished by their sometimes remarkable sensitivity to chemotherapy. While no clinical or histopathologic feature of these tumors allows accurate prediction of their response to chemotherapy, allelic loss of chromosome 1p is a strong predictor of chemosensitivity, and combined loss of chromosome 1p and 19q is statistically significantly associated with both chemosensitivity and longer recurrence-free survival after chemotherapy (Cairncross *et al.*, 1998). Loss of 1p also appears to have prognostic importance in low-grade oligodendrogliomas (Sasaki *et al.*, 2002; van den Bent *et al.*, 2003). Tumors without 1p loss are associated with more aggressive behavior. A recent study (Mohapatra *et al.*, 2006) was undertaken to evaluate the genetic features of oligodendroglial tumors at a much finer resolution than had been used in previous studies, which used the older, low-throughput techniques of loss of heterozygosity and fluorescence *in situ* hybridization (FISH). The 28 gliomas included 10 grade II oligodendrogliomas, 9 anaplastic oligodendrogliomas, 1 grade II oligoastrocytoma, and 8 anaplastic oligoastrocytomas.

While the ultimate goal of the glioma study is to identify novel regions of loss or gain that are shared by patients with oligodendroglial tumors, the first step is to determine regions of loss or gain within individual patients. For each of the 28 subjects in the study, we obtained aCGH data from BAC arrays containing clones from chromosomes 1, 7, 19, and X. We compared the results of our proposed method to results obtained using the MergeLevels approach proposed by Willenbrock and Fridlyand (2005) based on segments identified by the CBS change point detection method of Olshen *et al.* (2004).

Prior to analysis, we normalized the data. Due to the high percentage of loss and low percentage of gain in the data, median-centered normalization was not effective and it did not center the no-change regions at 0. The experiment included three normal–normal hybridizations, which could be used for normalization. However, due to the variability across individuals, this also was inadequate. As patients with oligodendroglial brain tumors typically do not exhibit genetic alterations on chromosome 1q, we centered each hybridization about the median of its 1q $\log_2$ ratios.

Using the normalized data, we estimated the parameters of the model using our proposed pseudolikelihood function (4.4) and we calculated the averaged posterior probabilities. We then assigned each clone to the state corresponding to the highest of its three posterior probabilities. We used the truncating value, $\varepsilon = 0.20$; the results were not sensitive to the precise value of ε within (0.1, 0.4).

Figures 2 and 3 contain the data from four tumors and analysis results from the proposed method along with those obtained from the MergeLevels-CBS method. The probabilities of loss provided by the pseudolikelihood method are represented by the vertical light-gray lines, extending from the bottom of the figure upward, and their values are the associated values on the left axis. The probabilities of gain are represented by the vertical dark-gray lines, extending from the top of the figure downward, and their values are calculated as one minus the associated value on the left axis (or simply their lengths, measured using the scale on the left axis). The probabilities of no-change are represented by the white lines, in the middle of the figure, between the light- and dark-gray lines. Their values are calculated as one minus the probabilities of loss and gain (or simply their lengths, measured using the scale on the left axis). For example, in Figure 2(a), the pseudolikelihood method yields posterior probabilities of loss that are close to 1 for all clones on chromosome 1p. The posterior probabilities of gain for one of the segments on chromosome 1q is close to 0.2 (i.e. the dark-gray lines extend downward to about 0.8). The right axis in the figures refers to the $\log_2$ scale. The vertical dashed lines demarcate the chromosomes. The horizontal thick dashed lines drawn through the data points on each chromosome represent the mean $\log_2$ levels for the segments estimated by the MergeLevels-CBS method. The original CBS segmentation results are denoted by the horizontal thin solid lines. For the tumors in Figures 2 and 3, we also formally identified breakpoints using the method outlined in Section 4.3: locations at which the maximum of the three probabilities exceeded a threshold of 0.6 were identified as breakpoints. Breakpoints in both figures are denoted by black tick marks extending below the 0 probability level of the plots.

For the hybridizations depicted in Figure 2, there is general agreement among all three methods. For the hybridization shown in Figure 2(a), loss is detected by all three methods at 1p and 19q. All three methods find no-change on 1p and 19q for the hybridization shown in Figure 2(b). One difference between the two plots is instructive. The tumor displayed in Figure 2(a) contains an apparent small region of low-level gain on 1q, as does the tumor in Figure 2(b) on 19p. These regions are flagged by the pseudolikelihood method (dark-gray downward spikes) and the corresponding quantitative assessments of the probability of gain are low: 0.30–0.40. The CBS method identifies a change point demarcating the relatively large segment of transition on 1q in Figure 2(a), but does not identify the smaller segment on 19p in Figure 2(b). Since 1q has been used as a basis for normalization, it should be noted that either any gains on this arm should be viewed as suspect or the justification for the normalization should be revisited. Nevertheless, it is difficult to assess from the CBS method the extent to which those identified segments should be viewed as true alterations or as segments due to chance variation. The MergeLevels-CBS method does not identify a gain at either location.

The results shown in Figure 3 illustrate an increased sensitivity of the pseudolikelihood method relative to the CBS and MergeLevels-CBS methods with regard to small regions of apparent within-chromosome change. Figure 3(a) illustrates a case for which there appears to be copy number alteration on a small region of chromosome 7 (i.e. at the epidermal growth factor receptor [EFGR] gene). EGFR amplification for that case was, in fact, confirmed by FISH. The pseudolikelihood method identifies this region exactly. In contrast, the CBS method in effect averages this small region of amplification with the

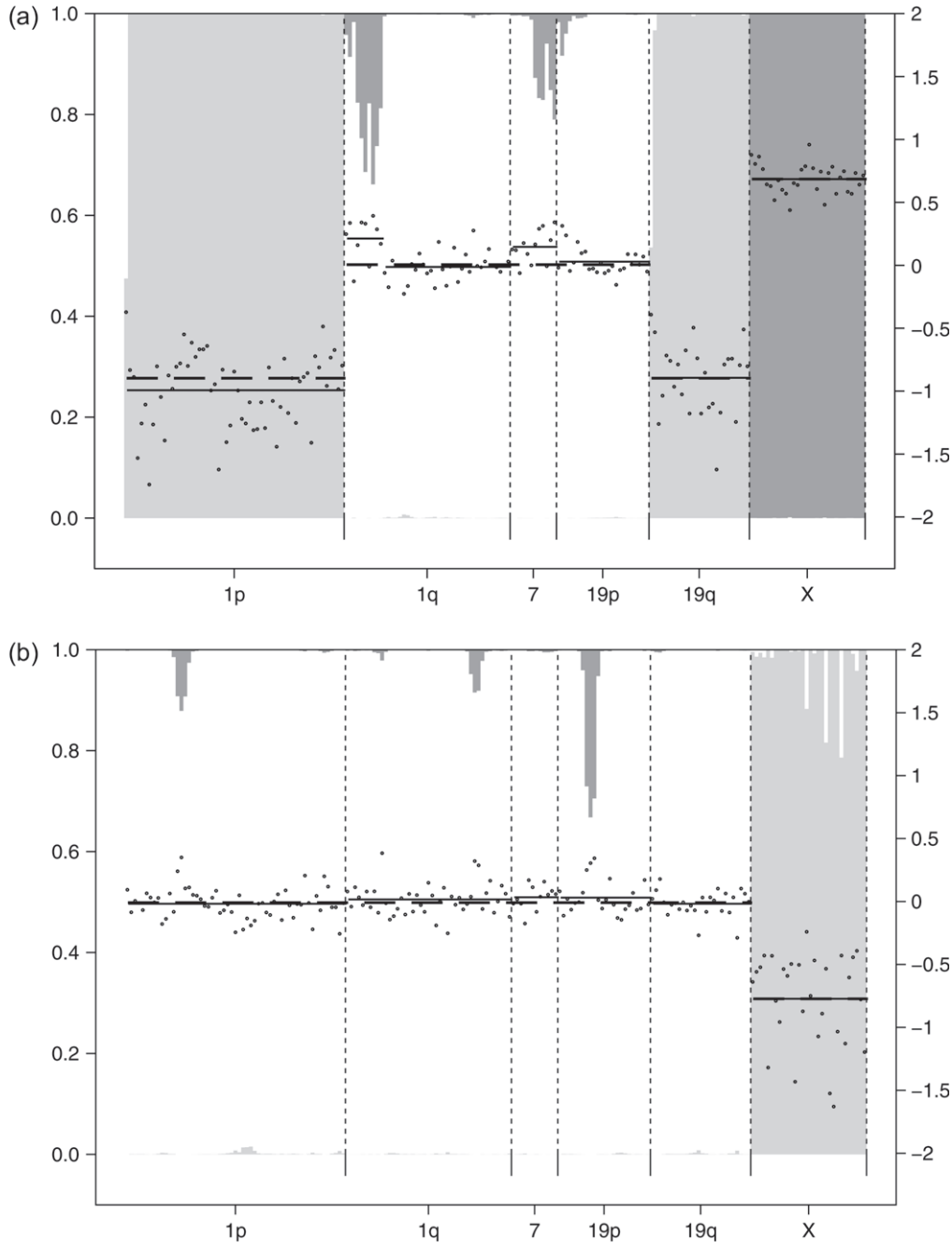

Fig. 2. Two examples in which all three methods agree in their identification of segments of loss and gain. (a) Loss is identified at both 1p and 19q by all methods. (b) No-change is detected at both 1p and 19q by all methods. The axis on the right is on the probability scale. The axis on the left is on the $\log_2$ scale. The thin solid and thick dashed horizontal lines represent CBS and MergeLevels-CBS segments, respectively. The dark- and light-gray vertical bars indicate the posterior probabilities of gain and loss, respectively. Breakpoints identified by the pseudolikelihood method are denoted by the tick marks extending below the 0 probability level in each plot.

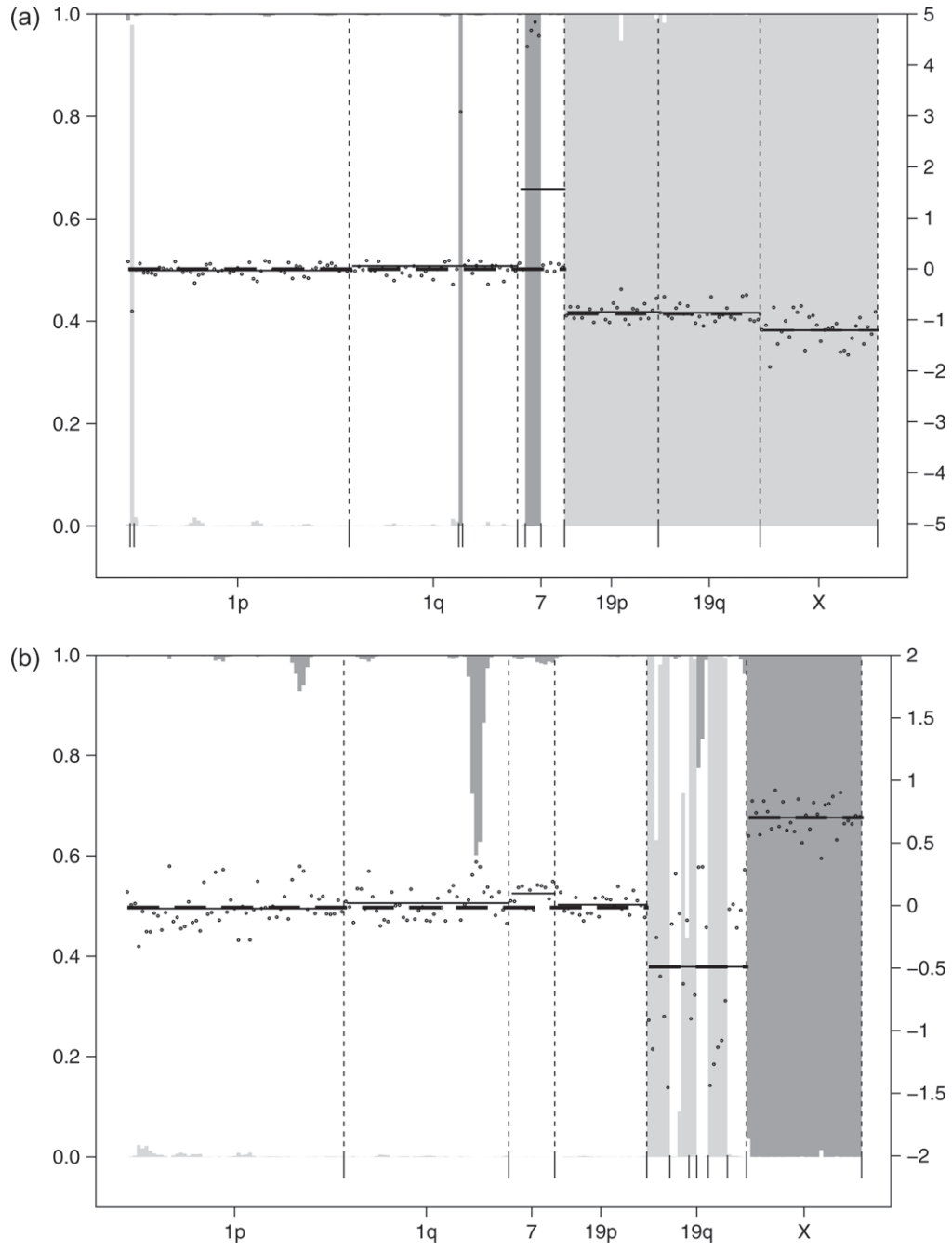

Fig. 3. Two examples that demonstrate the pseudolikelihood's increased sensitivity to within-chromosome change in comparison to CBS and MergeLevels-CBS. (a) A small region of within-chromosome amplification on 7 is identified by the pseudolikelihood method. In contrast, the CBS method identifies a lower level of gain on the entire chromosome. MergeLevels-CBS does not recognize any gain on 7. (b) Loss is identified on regions of 19q by the pseudolikelihood method. The entire arm is categorized as loss by both CBS and MergeLevels-CBS. Breakpoints identified by the pseudolikelihood method are denoted by the tick marks extending below the 0 probability level in each plot.

normal copy numbers from the remaining part of the chromosome and identifies a lower level of transition across the entire chromosome. The MergeLevels-CBS method does not categorize any of the clones on chromosome 7 as copy number gains. Figure 3(b) similarly illustrates the identification of multiple regions of loss at 19q by the pseudolikelihood method versus identification of loss of the entire chromosome arm by CBS and MergeLevels-CBS.

6. SIMULATION STUDY

We conducted simulation studies to investigate the performance of the pseudolikelihood method under a variety of conditions. We completely specified the parameters of the model from which we generated both the state vector, C, and the $\log_2$ ratios. Each of the 5000 simulated data sets contained 15 hybridizations, each hybridization consisted of five chromosomes, and each chromosome consisted of 50 clones. For each simulated data set, we calculated misclassification rates (see Section 4.2) for the pseudolikelihood method by comparing the classification results to the truth, i.e. C. We also analyzed the data using the MergeLevels-CBS approach. We compared the MergeLevels-CBS classification results to the truth (C) to obtain its misclassification rate. We then averaged the misclassification rates over all 5000 simulated data sets. We did not study the MergeLevels results using segments identified by other approaches such as the HMM method of Fridlyand *et al.* (2004). Likewise, we did not compare the results to other methods such as GLADmerge. We made the choice because results of these methods have been shown to be similar to those obtained using the MergeLevels-CBS approach (Willenbrock and Fridlyand, 2005). In fact, these authors found that among these methods, the MergeLevels-CBS approach 'has the best operational characteristics in terms of its sensitivity and false discovery rate for breakpoint detection.'

We initially selected parameter values for the data generation process that were similar to those that we estimated for the glioma data. However, while these data exhibited variation in the mean levels of gain and loss, there were very few copy number transitions; instead, there were whole chromosome losses. Hence, we increased the transition probabilities to yield data with more genetic alteration and with small regions of loss and gain. The resultant simulated data sets exhibited extensive noise due to both mean level variation and a large number of state transitions. Figure 4 depicts a single realization of a simulated high-transition hybridization. True copy number losses are denoted with circles and gains are denoted with triangles. We used $\varepsilon = 0.2$ in all the analyses.

Simulation results for the high-transition data are listed in Tables 1 and 2. Table 1 contains the parameter estimates. The model parameters that are not listed in Table 1 are functions of those that are specified. Table 2 lists the average misclassification rates of both the pseudolikelihood and MergeLevels-CBS analyses. False positives are gains or losses identified by the respective method that were not actually present. False negatives are true gains or losses that were missed by the method. Results in Tables 1 and 2 under the heading 'HM' (for hybridization-specific means) were obtained for data that were generated under a model in which a single mean for each of the states of loss, gain, and no-change (i.e. a single γ_L, γ_G, and γ_0) was generated for each hybridization. Given the true state vector C, and the generated state means, the $\log_2$ ratios were then generated. The results under the heading 'CM' (for chromosome-specific means) were obtained for data that were generated under a model in which different realizations of the state means, γ_L, γ_G, and γ_0, were generated for each chromosome. Lastly, the results under the heading 'HV' (for hybridization-specific variances) were obtained from data that were generated under a model in which the variability of the $\log_2$ ratios varied across hybridizations. Specifically, we allowed σ to vary normally across hybridizations with mean 0 and standard deviation 0.045. These data violate the assumption of the pseudolikelihood approach that the variability of the $\log_2$ ratios is constant across hybridizations. We examined this violation to assess the robustness of our approach because, in reality, some hybridizations are 'noisier' than others.

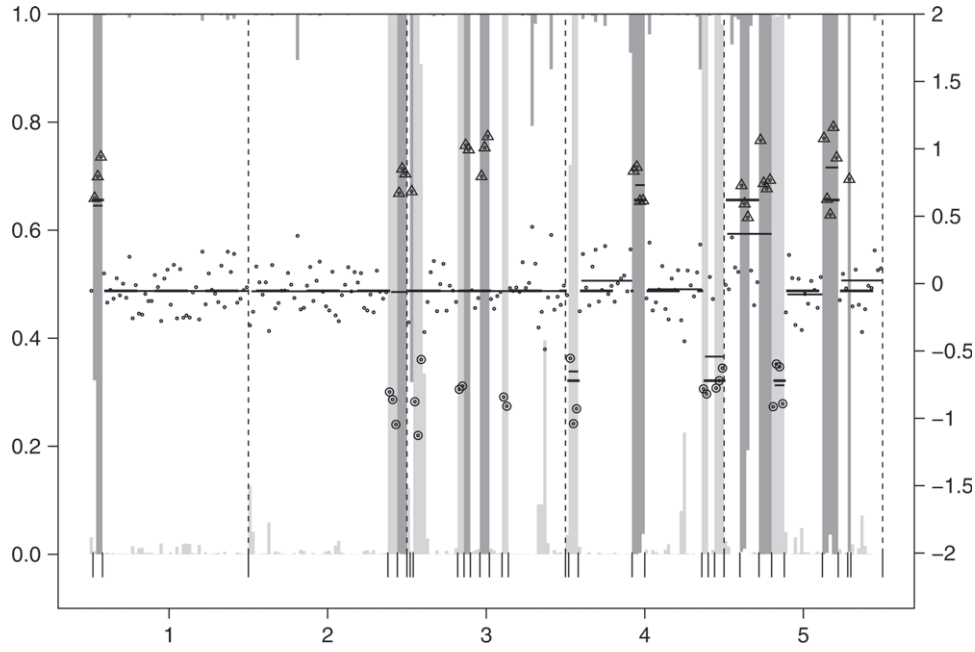

Fig. 4. An example of a simulated hybridization generated using high-transition probabilities along with results of pseudolikelihood, original CBS analysis results (thin solid horizontal lines), and MergeLevels-CBS results (thick dashed horizontal lines). Clones at which a true copy number loss exists are denoted by a circle around the data point. Clones at which a true copy number gain exists are similarly denoted by a triangle. Breakpoints identified by the pseudolikelihood method are also identified for both tumors by the tick marks extending below the 0 probability level of each figure.

Table 1. *Parameter estimates (MSEs) for high-transition data from simulation study*[†]

Parameter	True value	HM[‡]	CM[§]	HV[¶]
μ_L	−0.755	−0.754 (4.5×10^{-3})	−0.752 (1.9×10^{-3})	−0.734 (5.3×10^{-3})
μ_0	0.000	0.000 (3.0×10^{-4})	0.000 (7.6×10^{-5})	0.000 (3.1×10^{-4})
μ_G	0.613	0.637 (4.4×10^{-3})	0.631 (1.7×10^{-3})	0.622 (4.1×10^{-3})
σ^2	0.032	0.033 (3.7×10^{-6})	0.033 (3.2×10^{-6})	0.032 (1.3×10^{-5})
σ_L^2	0.050	0.045 (4.5×10^{-4})	0.048 (2.0×10^{-4})	0.049 (4.4×10^{-4})
σ_0^2	0.005	0.003 (4.5×10^{-6})	0.003 (3.6×10^{-6})	0.003 (5.2×10^{-6})
σ_G^2	0.051	0.038 (4.4×10^{-4})	0.041 (1.9×10^{-4})	0.040 (3.7×10^{-4})
π_L	0.080	0.076 (1.0×10^{-4})	0.076 (7.7×10^{-5})	0.082 (1.5×10^{-4})
π_G	0.080	0.080 (1.5×10^{-4})	0.081 (1.0×10^{-4})	0.085 (1.7×10^{-4})
a_{G0}	0.038	0.036 (3.7×10^{-5})	0.037 (2.5×10^{-5})	0.041 (5.8×10^{-5})
a_{LG}	0.100	0.097 (3.4×10^{-4})	0.099 (3.1×10^{-4})	0.101 (3.8×10^{-4})
a_{0L}	0.400	0.413 (1.3×10^{-3})	0.412 (1.2×10^{-3})	0.418 (1.4×10^{-4})

[†]Study consisted of 5000 repetitions of 15 simulated hybridizations. Each hybridization consisted of five chromosomes with 50 clones apiece.

[‡]HM: data sets in which state means (γ values) varied across hybridizations.

[§]CM: data sets in which state means varied across chromosomes.

[¶]HV: data sets in which the variance of the $\log_2$ ratios (σ) varied across hybridizations.

Table 2. *Pseudolikelihood and MergeLevels-CBS misclassification rates for high-transition data from simulation study*

Data set	Error type	Pseudolikelihood	MergeLevels-CBS
HM[†]	False positive	0.0098	0.0113
HM	False negative	0.0324	0.1340
HM	Total	0.0422	0.1453
CM[‡]	False positive	0.0100	0.1341
CM	False negative	0.0322	0.1145
CM	Total	0.0422	0.2486
HV[§]	False positive	0.0173	0.0150
HV	False negative	0.0304	0.1335
HV	Total	0.0477	0.1485

[†]HM: data sets in which state means (γ values) varied across hybridizations.
[‡]CM: data sets in which state means varied across chromosomes.
[§]HV: data sets in which the variance of the $\log_2$ ratios (σ) varied across hybridizations.

We also examined the performance of both methods on data sets with fewer copy number transitions, i.e. low-transition data. Figure 5 displays a single realization of a simulated low-transition hybridization. Tables 3 and 4 contain the results of these simulation studies.

Overall, the pseudolikelihood method resulted in accurate parameter estimation for both high- and low-transition data (see Tables 1 and 3). Not surprisingly, the mean squared errors (MSEs) are slightly larger for the HV data sets than for the HM and CM data sets. Also, our procedure appears to perform slightly better on the CM data than on the HM data with regard to parameter estimation.

The pseudolikelihood method also performs well with regard to classification of clones. For the high-transition data (e.g. Figure 4) in which there is substantial chromosomal instability, the false-negative rate of the pseudolikelihood approach is roughly a third of that of the MergeLevels-CBS approach in all three settings. With regard to the false-positive rate, the performance of the two methods is similar in the HM and HV settings. In the CM settings, however, when copy number mean levels vary across chromosomes within a hybridization, the pseudolikelihood method has a much lower false-positive rate than the MergeLevels-CBS approach. For low-transition data (e.g. Figure 5) in which there is greater chromosomal stability and fewer small copy number transitions, the false-negative error rates for both methods are similar in all three settings. False-positive rates are also similar in the HM and HV settings. Again, however, in the CM setting, the pseudolikelihood method has a much lower false-positive rate than the MergeLevels-CBS approach. Also of note, the misclassification rates of pseudolikelihood method for the HV data are similar to those of the HM and CM data, for both the high- and low-transition models. This again suggests that the pseudolikelihood model is robust to departures from the assumption of constant variance.

Two separate shortcomings of the MergeLevels-CBS approach seem to be apparent in these results. First, the high false-negative error rates for the high-transition data suggest that the MergeLevels-CBS approach is less sensitive to small regions of copy number alteration than the pseudolikelihood approach. It appears that through combination of segments across chromosomes, small regions of change (which are numerous in high-transition data) are combined with larger regions of no-change. Hence, true changes are missed, resulting in a high false-negative rate. Second, the high false-positive error rates in the CM setting suggest that the MergeLevels-CBS approach has difficulty in accounting for variability in mean $\log_2$ ratio values for gain and loss for both high- and low-transition data and often does not combine segments from the same genetic alteration type. Thus, when the no-change level is then identified, a number of true no-change segments are misclassified as gains or losses, resulting in a high false-positive rate.

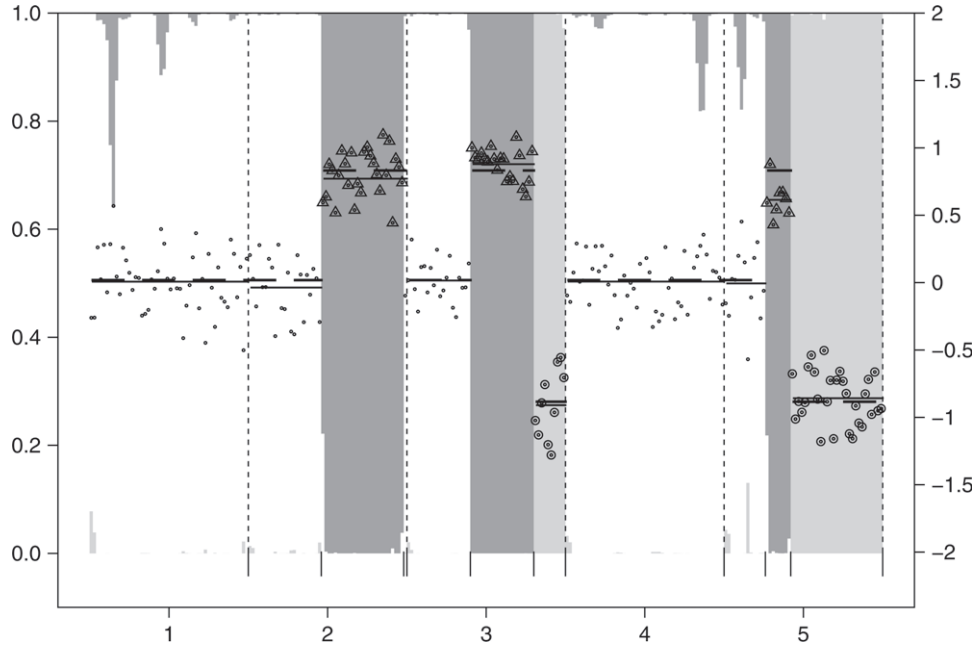

Fig. 5. An example of a simulated hybridization generated using low-transition probabilities along with results of pseudolikelihood, original CBS analysis results (thin solid horizontal lines), and MergeLevels-CBS results (thick dashed horizontal lines). Clones at which a true copy number loss exists are denoted by a circle around the data point. Clones at which a true copy number gain exists are similarly denoted by a triangle. Breakpoints identified by the pseudolikelihood method are also identified for both tumors by the tick marks extending below the 0 probability level of each figure.

Table 3. *Parameter estimates (MSEs) for low-transition data from simulation study*[†]

Parameter	True value	HM[‡]	CM[§]	HV[¶]
μ_L	−0.755	−0.750 (6.4×10^{-3})	−0.751 (4.7×10^{-3})	−0.739 (7.6×10^{-3})
μ_0	0.000	0.000 (3.0×10^{-4})	0.000 (7.6×10^{-5})	0.000 (3.2×10^{-4})
μ_G	0.613	0.630 (6.9×10^{-3})	0.629 (4.3×10^{-3})	0.620 (6.1×10^{-3})
σ^2	0.032	0.033 (2.1×10^{-6})	0.033 (1.7×10^{-6})	0.033 (1.4×10^{-5})
σ_L^2	0.050	0.043 (6.1×10^{-4})	0.046 (4.3×10^{-4})	0.044 (5.8×10^{-4})
σ_0^2	0.005	0.003 (4.3×10^{-6})	0.003 (3.2×10^{-6})	0.003 (5.0×10^{-6})
σ_G^2	0.051	0.036 (5.8×10^{-4})	0.039 (3.8×10^{-4})	0.036 (5.2×10^{-4})
π_L	0.080	0.078 (4.9×10^{-4})	0.079 (4.7×10^{-4})	0.081 (4.8×10^{-4})
π_G	0.080	0.080 (4.6×10^{-4})	0.081 (4.6×10^{-4})	0.083 (5.1×10^{-4})
a_{G0}	0.004	0.005 (4.8×10^{-6})	0.005 (4.7×10^{-6})	0.007 (1.7×10^{-5})
a_{LG}	0.020	0.021 (1.6×10^{-4})	0.021 (7.0×10^{-5})	0.023 (1.0×10^{-4})
a_{0L}	0.039	0.059 (5.8×10^{-4})	0.058 (5.8×10^{-4})	0.069 (1.3×10^{-3})

[†]Study consisted of 5000 repetitions of 15 simulated hybridizations. Each hybridization consisted of five chromosomes with 50 clones apiece.
[‡]HM: data sets in which state means (γ values) varied across hybridizations.
[§]CM: data sets in which state means varied across chromosomes.
[¶]HV: data sets in which the variance of the $\log_2$ ratios (σ) varied across hybridizations.

Table 4. *Pseudolikelihood and MergeLevels-CBS misclassification rates for low-transition data from simulation study*

Data set	Error type	Pseudolikelihood	MergeLevels-CBS
HM[†]	False positive	0.0046	0.0124
HM	False negative	0.0173	0.0155
HM	Total	0.0219	0.0279
CM[‡]	False positive	0.0046	0.1148
CM	False negative	0.0177	0.0142
CM	Total	0.0223	0.1290
HV[§]	False positive	0.0085	0.0123
HV	False negative	0.0176	0.0160
HV	Total	0.0261	0.0283

[†]HM: data sets in which state means (γ values) varied across hybridizations.
[‡]CM: data sets in which state means varied across chromosomes.
[§]HV: data sets in which the variance of the $\log_2$ ratios (σ) varied across hybridizations.

Table 5. *Comparison of single-chromosome analysis versus simultaneous analysis of multiple chromosomes for high-transition data from simulation study*[†]

Data set	Parameters	Error type	Single[‡]	Multiple[§]
I[¶]	$\sigma^2 = 0.032$	False positive	0.0319	0.0145
	$(\sigma_L^2, \sigma_0^2, \sigma_G^2) = (0.05, 0.005, 0.05)$	False negative	0.0283	0.0304
	$(\mu_L, \mu_G) = (-0.755, 0.613)$	Total	0.0602	0.0449
II[‖]	$\sigma^2 = 0.064$	False positive	0.0432	0.0327
	$(\sigma_L^2, \sigma_0^2, \sigma_G^2) = (0.05, 0.005, 0.05)$	False negative	0.0488	0.0493
	$(\mu_L, \mu_G) = (-0.755, 0.613)$	Total	0.0920	0.0820
III[#]	$\sigma^2 = 0.032$	False positive	0.0442	0.0232
	$(\sigma_L^2, \sigma_0^2, \sigma_G^2) = (0.10, 0.010, 0.10)$	False negative	0.0353	0.0395
	$(\mu_L, \mu_G) = (-0.755, 0.613)$	Total	0.0795	0.0627
IV[††]	$\sigma^2 = 0.032$	False positive	0.0290	0.0192
	$(\sigma_L^2, \sigma_0^2, \sigma_G^2) = (0.05, 0.005, 0.05)$	False negative	0.0616	0.0617
	$(\mu_L, \mu_G) = (-0.500, 0.500)$	Total	0.0906	0.0809

[†]Study consisted of 500 simulated hybridizations. Each hybridization consisted of five chromosomes with 50 clones apiece.
[‡]Five chromosomes analyzed separately.
[§]Five chromosomes analyzed simultaneously.
[¶]Data simulated using parameters Table 1.
[‖]Same as I, except variability (σ^2) of the $\log_2$ ratios was doubled.
[#]Same as I, except variability of the random effects (γ_L, γ_0, γ_G) was doubled.
[††]Same as I, except magnitudes of gain and loss means (μ_G and μ_L) were decreased.

We lastly conducted a small simulation study to assess whether the simultaneous analysis of chromosomes and hybridizations conducted by the pseudolikelihood approach is indeed advantageous relative to single-chromosome analyses and to assess the performance of the model in small sample sizes (i.e. single chromosome consisting of 50 clones). Using the same methods as above, we generated data for five

Table 6. *Comparison of single-chromosome analysis versus simultaneous analysis of multiple chromosomes for low-transition data from simulation study*[†]

Data set	Parameters	Error type	Single[‡]	Multiple[§]
I[¶]	$\sigma^2 = 0.032$	False positive	0.0364	0.0143
	$(\sigma_L^2, \sigma_0^2, \sigma_G^2) = (0.05, 0.005, 0.05)$	False negative	0.0144	0.0139
	$(\mu_L, \mu_G) = (-0.755, 0.613)$	Total	0.0508	0.0282
II[‖]	$\sigma^2 = 0.064$	False positive	0.0362	0.0282
	$(\sigma_L^2, \sigma_0^2, \sigma_G^2) = (0.05, 0.005, 0.05)$	False negative	0.0239	0.0269
	$(\mu_L, \mu_G) = (-0.755, 0.613)$	Total	0.0601	0.0551
III[#]	$\sigma^2 = 0.032$	False positive	0.0464	0.0223
	$(\sigma_L^2, \sigma_0^2, \sigma_G^2) = (0.10, 0.010, 0.10)$	False negative	0.0232	0.0240
	$(\mu_L, \mu_G) = (-0.755, 0.613)$	Total	0.0696	0.0463
IV[††]	$\sigma^2 = 0.032$	False positive	0.0365	0.0141
	$(\sigma_L^2, \sigma_0^2, \sigma_G^2) = (0.05, 0.005, 0.05)$	False negative	0.0375	0.0379
	$(\mu_L, \mu_G) = (-0.500, 0.500)$	Total	0.0740	0.0520

[†]Study consisted of 500 simulated hybridizations. Each hybridization consisted of five chromosomes with 50 clones apiece.
[‡]Five chromosomes analyzed separately.
[§]Five chromosomes analyzed simultaneously.
[¶]Data simulated using parameters Table 3.
[‖]Same as I, except variability (σ^2) of the $\log_2$ ratios was doubled.
[#]Same as I, except variability of the random effects ($\gamma_L, \gamma_0, \gamma_G$) was doubled.
[††]Same as I, except magnitudes of gain and loss means (μ_G and μ_L) were decreased.

chromosomes. First, we analyzed each of the five chromosomes separately and calculated the misclassification rates using total numbers of misclassifications across the five analyses. Then we analyzed the five chromosomes jointly and calculated the misclassification rate. We repeated this using 500 simulations.

The results for high- and low-transition data, all generated under the CM model with the same parameter values as in Tables 1 and 3, are contained in Tables 5 and 6, respectively. We then varied some of these parameters to obtain three additional scenarios: (1) larger σ^2 (variability of the $\log_2$ ratios), (2) larger $\sigma_L^2, \sigma_0^2, \sigma_G^2$ (variability of the random effects, $\gamma_L, \gamma_0, \gamma_G$), and (3) decreased magnitudes of μ_L and μ_G (mean levels for gain and loss).

In this small study, we found that for both high- and low-transition data, the performance of the pseudolikelihood method is improved through the simultaneous analysis of multiple chromosomes. For high-transition data, the single-chromosome analysis total misclassification rate is roughly 50% larger than that of the simultaneous chromosome analysis, and for low-transition data, it is roughly twice as large. This difference is due to the fact that single chromosomes are much more informative for high-transition data than for low-transition data and thus the advantage of the simultaneous analysis is greater for low-transition data. The advantage in error rate of the simultaneous analysis is entirely due to a decrease in false-positive rate. This occurs because the single-chromosome analysis gives greater weight to $\log_2$ outliers than does the simultaneous analysis and hence results in a higher false-positive rate. The multiple-chromosome analysis corrects for this but, in doing so, misses a few copy number changes that are real, and thus, its false-negative rate is not diminished relative to the single-chromosome analysis. Even in the single-chromosome analysis, however, total misclassification rates are still an improvement over those obtained through the use of the MergeLevels-CBS approach (see Tables 2 and 4). Hence, the approach is advantageous even in small sample situations.

7. DISCUSSION

We have proposed an analytic approach for aCGH data that exploits features that are shared in common among chromosome and hybridizations, while allowing for variation among these same units. We have shown that this approach is an improvement over currently used methods both in identifying small regions of copy number gain and loss and in classifying regions of change when intratumoral clonal variation is present. Furthermore, we have shown that the method does in fact borrow strength across chromosomes and that by utilizing all available data, results in improved identification of copy number alterations. Finally, the pseudolikelihood method yields easily interpretable graphical output, allowing researchers to identify regions of possible copy number gain and loss and to understand their associated probabilities.

While we did allow for variability in the mean levels for loss, gain, and no-change, we did not allow for variability in the transition or state probabilities, i.e. the π values and a values. In reality, however, it is suspected that certain areas on a given chromosome and certain chromosomes are more susceptible to change in copy number than others (Gabriel *et al.*, 2002). Additionally, there is probably an interactive effect between individuals and chromosomal susceptibility to copy number change. In future work, we will investigate the feasibility of treating these probabilities as random effects within the modeling framework that we have proposed in this paper, thereby allowing for this likely variability. It will be of interest to ascertain whether this added flexibility further improves the error rates.

Furthermore, we did not allow for the variability of the observed $\log_2$ ratios to vary across individuals. However, for both biological and experimental reasons, it is likely that it does. Nonetheless, in our simulations we found that it may not be necessary to build this into our estimation procedure as the performance of our procedure is robust to changes in variability.

A major contribution of our estimation procedure is its ability to conduct simultaneous analysis of an entire experiment. One question that arises in this regard is that of consistency of the estimates. In our simulation of single chromosomes consisting of 50 clones each (a very small sample size given current arrays) in which gains and losses were often absent, the method performed well, assigning very low probabilities of gain and loss. The robustness of this procedure is, in part, due to the use of the truncation parameter, ε, which assists in the identifiability of the three state means. Nonetheless, despite these promising results, it may be advantageous to include a penalty term in the pseudolikelihood function to obtain consistency in certain situations. Analogous to the use by Cox and Reid (2004) of the univariate likelihood in their penalty for the pairwise likelihood function, we might use the pairwise likelihood in the penalty for our trivariate pseudolikelihood. This is a topic of interest for future research.

An additional major contribution of our method is its output of quantitative assessments of the likelihood of the various genetic alterations at each clone. This is not provided by the competing segmentation methods of Olshen *et al.* (2004) and Fridlyand *et al.* (2004). An obvious and important extension of our procedure will be the derivation of confidence intervals for the posterior probabilities of the different genetic states. This will be possible through the framework of generalized estimation equations. Also, we will need to develop a clear graphical display for our results that incorporates the estimates of variability afforded by the confidence intervals. Finally, we have proposed a model-based assessment of breakpoint locations, either as probabilities or as thresholded outcomes. Further research remains to be done on the optimal method for identification of these breakpoints.

ACKNOWLEDGMENTS

This work was supported in part by National Institutes of Health grants NS048005, CA075971, CA106695, and CA121884.

REFERENCES

AGUIRRE, A., BRENNAN, C., BAILEY, G., SINHA, R., FENG, B., LEO, C., ZHANG, Y., ZHANG, J., GANS, J., BARDEESY, N. *et al.* (2004). High-resolution characterization of the pancreatic adenocarcinoma genome. *Proceedings of the National Academy of Sciences of the United States of America* **24**, 9067–9072.

AMARATUNGA, D. AND CABRERA, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. Hoboken, NJ: Wiley.

BESAG, J. E. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

CAIRNCROSS, J. G., UEKI, K., ZLATESCU, M. C., LISLE, D. K., FINKELSTEIN, D. M., HAMMOND, R. R., SILVER, J. S., STARK, P. C., MACDONALD, D. R., INO, Y. *et al.* (1998). Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *Journal of the National Cancer Institute* **90**, 1473–1479.

CLAYTON, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.

COX, D. R. AND REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–737.

DUDOIT, S., YANG, Y., CALLOW, M. AND SPEED, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

FOROZAN, F., MAHLAMAKI, E. H., MONNI, O., CHEN, Y., VELDMAN, R., JIANG, Y., GOODEN, G. C., ETHIER, S. P., KALLIONIEMI, A. AND KALLIONIEMI, O. P. (2000). Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data. *Cancer Research* **60**, 4519–4525.

FRIDLYAND, J., SNIJDERS, A., PINKEL, D., ALBERTSON, D. AND JAIN, A. (2004). Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* **90**, 132–153.

GABRIEL, S. B., SCHNAFFNER, S. F., NGUYEN, H., MOORE, J. M., ROY, J., BLUMENSTIEL, B., HIGGENS, J., DEFELICE, M., LOCHNER, A., FAGGART, M. *et al.* (2002). The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

HODGSON, G., HAGER, J. H., VOLIK, S., HARIONO, S., WERNICK, M., MOORE, D., ALBERTSON, D. G., PINKEL, D., COLLINS, C., HANAHAN, D. *et al.* (2001). Genome scanning with array CGH delineates regional alternatives in mouse islet carcinomas. *Nature Genetics* **29**, 459–464.

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.

IAFRATE, A. J., FEUK, L., RIVERA, M. N., LISTEWNIK, M. L., DONAHOE, P. K., QI, Y., SCHERER, S. W. AND LEE, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics* **36**, 949–951.

KALLIONIEMI, A., KALLIONIEMI, O. P., PIPER, J., TANNER, M., STOKKE, T., CHEN, L., SMITH, H. S., PINKEL, D., GRAY, J. W. AND WALDMAN, F. M. (1994). Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 2156–2160.

KALLIONIEMI, A., KALLIONIEMI, O. P., SUDAR, D., RUTOVITZ, D., GRAY, J. W., WALDMAN, F. AND PINKEL, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821.

LINDSAY, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221–239.

LUCITO, R., WEST, J., REINER, A., ALEXANDER, J., ESPOSITO, D., MISHRA, B., POWERS, S., NORTON, L. AND WIGLER, M. (2000). Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Research* **10**, 1726–1736.

MOHAPATRA, G., BETENSKY, R. A., MILLER, E. R., CAREY, B., GAUMONT, L. D., ENGLER, D. A. AND LOUIS, D. N. (2006). Glioma test array for use with formalin-fixed, paraffin-embedded tissue: array comparative genomic hybridization correlates with loss of heterozygosity and fluorescence in situ hybridization. *The Journal of Molecular Diagnostics* (in press).

NUTT, C. L., MANI, D. R., BETENSKY, R. A., TAMAYO, P., CAIRNCROSS, J. G., LADD, C., POHL, U., HARTMANN, C., MCLAUGHLIN, M. E., BATCHELOR, T. T. *et al.* (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research* **63**, 1602–1607.

OAKES, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika* **73**, 353–361.

OKADA, Y., HURWITZ, E. E., ESPOSITO, J. M., BROWER, M. A., NUTT, C. L. AND LOUIS, D. N. (2003). Selection pressures of TP53 mutation and microenvironmental location influence epidermal growth factor receptor gene amplification in human glioblastomas. *Cancer Research* **63**, 413–416.

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* **4**, 557–572.

PARIS, P. L., ANDAYA, A., FRIDLYAND, J., JAIN, A. N., WEINBERG, V., KOWBEL, D., BREBNER, J. H., SIMKO, J., WATSON, J. E. V., VOLIK, S. *et al.* (2004). Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. *Human Molecular Genetics* **13**, 1303–1313.

PICARD, F., ROBIN, S., LAVIELLE, M., VAISSE, C. AND DAUDIN, J. (2004). A statistical approach for CGH microarray data analysis. *RR5139*. Institut National de la Recherche en Informatique et en Automatique.

PINKEL, D., SEGRAVES, R., SUDAR, D., CLARK, S., POOLE, I., KOWBEL, D., COLLINS, C., KNO, W. L., CHEN, C., ZHAI, Y. *et al.* (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–211.

POLLACK, J. R., PEROU, C. M., ALIZADEH, A. A., EISEN, M. B., PERGAMENSCHIKOV, A., WILLIAMS, C. F., JEFFREY, S. S., BOTSTEIN, D. AND BROWN, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* **23**, 41–46.

POLLACK, J. R., SORLIE, T., PEROU, C. M., REES, C. A., JEFFREY, S. S., LONNING, P. E., TIBSHIRANI, R., BOTSTEIN, D., BORRESEN-DALE, A. AND BROWN, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alternation in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12963–12968.

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the Institute of Electrical and Electronics Engineers* **77**, 257–286.

ROSSI, M. R., GAILE, D., LADUCA, J., MATSUI, S. I., CONROY, J., MCQUAID, D., CHERVINSKY, D., EDDY, R., CHEN, H. S., BARNETT, G. H. *et al.* (2005). Identification of consistent novel submegabase deletions in low-grade oligodendrogliomas using array-based comparative genomic hybridization. *Genes, Chromosomes & Cancer* **44**, 85–96.

SASAKI, H., ZLATESCU, M. C., BETENSKY, R. A., JOHNK, L. B., CUTONE, A. N., CAIRNCROSS, J. G. AND LOUIS, D. N. (2002). Histopathological-molecular genetic correlations in referral pathologist-diagnosed low-grade "oligodendroglioma." *Journal of Neuropathology and Experimental Neurology* **61**, 58–63.

SMITH, J. S., PERRY, A., BORELL, T. J., LEE, H. K., O'FALLON, J., HOSEK, S. M., KIMMEL, D., YATES, A., BURGER, P. C., SCHEITHAUER, B. W. *et al.* (2000). Alterations of chromosome arms 1p and 19q as predictors of survival in oligodendrogliomas, astrocytomas, and mixed oligoastrocytomas. *Journal of Clinical Oncology* **18**, 636–645.

SMITH, J. S., TACHIBANA, I., PASSE, S. M., HUNTLEY, B. K., BORELL, T. J., ITURRIA, N., O'FALLON, J. R., SCHAEFER, P. L., SCHEITHAUER, B. W., JAMES, C. D. *et al.* (2001). PTEN mutation, EGFR amplification, and outcome in patients with anaplastic astrocytoma and glioblastoma multiforme. *Journal of the National Cancer Institute* **93**, 1246–1256.

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

TIRKKONEN, M., TANNER, M., KARHU, R., KALLIONIEMI, A., ISOLA, J. AND KALLIONIEMI, O. P. (1998). Molecular cytogenetics of primary breast cancer by CGH. *Genes, Chromosomes & Cancer* **21**, 177–184.

VAN DEN BENT, M. J., LOOIJENGA, L. H., LANGENBERG, K., DINJENS, W., GRAVELAND, W., UYTDEWILLIGEN, L., SILLEVIS SMITT, P. A., JENKINS, R. B. AND KROS, J. M. (2003). Chromosomal anomalies in oligodendroglial tumors are correlated with clinical features. *Cancer* **97**, 1276–1284.

WANG, P., YOUNG, K., POLLACK, J., NARASIMHAN, B. AND TIBSHIRANI, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics* **6**, 45–58.

WEISS, M. M., SNIJDERS, A. M., KUIPERS, E. J., YLSTRA, B., PINKEL, D., MEUWISSEN, S. G. M., VAN DIEST, P. J., ALBERTSON, D. G. AND MEIJER, G. A. (2003). Determination of amplicon boundaries at 20q13.2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization. *The Journal of Pathology* **200**, 320–326.

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

YANG, Y. H., DUDOIT, S., LUU, P., LIN, D. M., PENG, V., NGAI, J. AND SPEED, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, e15.

[*Received September 27, 2005; revised December 5, 2005; accepted for publication January 4, 2006*]