*Harvard University*
Harvard University Biostatistics Working Paper Series

*Year* 2006 *Paper* 24

# Bayesian Hidden Markov Modeling of Array CGH Data

Subharup Guha[*] Yi Li[†]
Donna Neuberg[‡]

[*]Harvard School of Public Health, sguha@hsph.harvard.edu
[†]Harvard University and Dana Farber Cancer Institute, yili@jimmy.harvard.edu
[‡]Harvard School of Public Health, neuberg@hsph.harvard.edu

# Bayesian Hidden Markov Modeling of Array CGH Data

Subharup Guha, Yi Li, and Donna Neuberg

## Abstract

Genomic alterations have been linked to the development and progression of cancer. The technique of Comparative Genomic Hybridization (CGH) yields data consisting of fluorescence intensity ratios of test and reference DNA samples. The intensity ratios provide information about the number of copies in DNA. Practical issues such as the contamination of tumor cells in tissue specimens and normalization errors necessitate the use of statistics for learning about the genomic alterations from array-CGH data. As increasing amounts of array CGH data become available, there is a growing need for automated algorithms for characterizing genomic profiles. Specifically, there is a need for algorithms that can identify gains and losses in the number of copies based on statistical considerations, rather than merely detect trends in the data.

We adopt a Bayesian approach, relying on the hidden Markov model to account for the inherent dependence in the intensity ratios. Posterior inferences are made about gains and losses in copy number. Localized amplifications (associated with oncogene mutations) and deletions (associated with mutations of tumor suppressors) are identified using posterior probabilities. Global trends such as extended regions of altered copy number are detected. Since the posterior distribution is analytically intractable, we implement a Metropolis-within-Gibbs algorithm for efficient simulation-based inference. Publicly available data on pancreatic adenocarcinoma, glioblastoma multiforme and breast cancer are analyzed, and comparisons are made with some widely-used algorithms to illustrate the reliability and success of the technique.

# Bayesian Hidden Markov Modeling of Array-CGH Data

Subharup Guha, Yi Li and Donna Neuberg[*]

October 6, 2006

[*]Subharup Guha is Research Fellow, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115 (sguha@hsph.harvard.edu); Yi Li is Associate Professor, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115 (yili@hsph.harvard.edu); and Donna Neuberg is Senior Lecturer, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115 (neuberg@hsph.harvard.edu). This material is based upon work supported by the National Institute of Health under Award No. R01CA95747. The first author thanks Professors Steven MacEachern, Louise Ryan and David Harrington for many insightful comments that helped improve the focus of the paper.

*Keywords*: Amplifications; Cancer; Deletions; DNA; Copy number; Genomic alterations; Intensity ratios; MCMC; Tumor.

**Abstract**

Genomic alterations have been linked to the development and progression of cancer. The technique of Comparative Genomic Hybridization (CGH) yields data consisting of fluorescence intensity ratios of test and reference DNA samples. The intensity ratios provide information about the number of copies in DNA. Practical issues such as the contamination of tumor cells in tissue specimens and normalization errors necessitate the use of statistics for learning about the genomic alterations from array-CGH data. As increasing amounts of array CGH data become available, there is a growing need for automated algorithms for characterizing genomic profiles. Specifically, there is a need for algorithms that can identify gains and losses in the number of copies based on statistical considerations, rather than merely detect trends in the data.

We adopt a Bayesian approach, relying on the hidden Markov model to account for the inherent dependence in the intensity ratios. Posterior inferences are made about gains and losses in copy number. Localized amplifications (associated with oncogene mutations) and deletions (associated with mutations of tumor suppressors) are identified using posterior probabilities. Global trends such as extended regions of altered copy number are detected. Since the posterior distribution is analytically intractable, we implement a Metropolis-within-Gibbs algorithm for efficient simulation-based inference. Publicly available data on pancreatic adenocarcinoma, glioblastoma multiforme and breast cancer are analyzed, and comparisons are made with some widely-used algorithms to illustrate the reliability and success of the technique.

# 1 INTRODUCTION

**The genomics of cancer.** The normal DNA of human females has two copies of the entire genomic code because there are 23 matched pairs of chromosomes. Human males have 22 matched pairs of non-sex (or *autosomal*) chromosomes and an unmatched pair of sex chromosomes. Hence the *copy number* of normal male DNA is two for the autosomal chromosomes. The ends of the chromosomes are called the *telomeres*. The telomere corresponding to the short arm of a chromosome is called the $p$ telomere, while the one corresponding to the long arm is called the $q$ telomere.

Human cells can be classified into *somatic* (or *body*) cells and *germ* cells. Barring a few exceptions like red blood cells, muscle cells and brain cells, the life cycle of somatic cells consists of a period of growth followed by cell division through mitosis. Cells must satisfy certain "quality control checks" before they can progress to a subsequent stage of the cycle. These checks ensure that the cells develop normally, that defects are repaired and that DNA is correctly copied during mitosis. Two kinds of genes play very important roles in the regulation procedure: proto-oncogenes and tumor-suppressors. Proto-oncogenes encourage the body cells to grow and divide, pushing them through the quality control check points. Tumor-suppressors tend to hold the cells back, inhibiting mitosis when there are cell defects, and signaling the cells to die when their lifespans have ended or when there are cell defects that cannot be repaired. Further details about the relevant biology for this problem are given in Pasternak (1999).

Occasionally, proto-oncogenes may mutate into oncogenes. The mutations are propagated to new cells through mitosis. Oncogenes duplicate themselves through several stages of mitosis so that cells end up with multiple copies of oncogenes. Oncogenes have a dominant effect on the cell function, causing the cells to divide at a rapid rate and resulting in the development of tumors. Tumors may also develop due to mutations in tumor-suppressors that cause them to become non-functional and allow the proto-oncogenes to play a dominant role. Tumor-suppressor mutations eventually result in the loss of one or both copies of the gene. A *deletion* is the loss of both copies in a genomic region.

A single mutation is usually not enough to trigger cancer. A number of complex biological events occur before a person acquires the phenotype of cancer. An example, but not a necessary condition, is the ability of tumor cells to metastasize making the tumor malignant. Not all the cells in a tumor specimen necessarily exhibit the same kind of genomic alteration. Additionally, there is a lot of variation among individuals. As the disease progresses, there are larger-scale changes in tumor DNA because of the breakdown of quality control in cell division.

*Copy number changes*, or alterations in the number of copies in tumor DNA, are therefore closely associated with the development and progression of cancer. A number of methods are currently available to detect genomic changes. *Karyotyping* views the chromosomes through a microscope during the metaphase stage of the cell cycle. This technique covers the entire genome but has low resolution because only the changes spanning large regions of the DNA, such as missing chromosomes, monosomies (loss of single copies) and trisomies (gain of additional copies of chromosomes) can be detected by this method. At the other end of the spectrum, *molecular genetic studies* are capable of single base pair resolution. Since the genome consists of approximately 3 billion bases, this technique cannot be used in the absence of prior knowledge to identify the DNA regions associated with a disease. Researchers must rely on other methods to first identify candidate loci involved in the disease pathogenesis.

**Array-CGH.** Comparative Genomic Hybridization (CGH) has emerged as a powerful technique because it combines relatively high resolution of a few million bases with the ability to span the entire genome in a single experiment (Kallioniemi et al. 1992). Fragmented DNA from a test sample is labeled with fluorochrome (typically Cy3) and is mixed with normal DNA that is identically fragmented but labeled using a different (typically Cy5). The normal and tumor DNA fragments are simultaneously hybridized to a normal metaphase spread. Image analysis yields data consisting of fluorescence intensity ratios along the genomes of the test and reference DNA samples. The more recently developed array-CGH techniques (Solinas-Toldo et al., 1997; Pinkel et al., 1998; Snijders et al., 2001; Pinkel and Albertson, 2005) hybridize the DNA fragments or "clones" to mapped array fragments rather than metaphase chromosomes. CGH arrays that rely on BAC (bacterial artificial chromosome) clones have a resolution of the order of 1 Mb (one million base pairs). Oligonucleotide and cDNA arrays (Pollack et al., 1999; Brennan et al., 2004) provide a higher resolution of 50–100 kb (1 kb = thousand base pairs). As with all microarray-based techniques, the fluorescence intensity ratios have to be normalized as part of a pre-processing step to correct for non-biological sources of error such as intensity fluctuations, background noise and fabrication artifacts (Brown et al., 2001; McLachlan et al., 2004). Refer to Khojasteh et al. (2005) for a comparison of different normalization methods for array-CGH data.

Array-CGH intensity ratios (equivalently, their transformation on the $\log_2$ scale) provide much useful information about genome-wide changes in copy number. Imagine an idealized situation where all the cells in a tumor specimen have identical genomic alterations and are uncontaminated by cells from surrounding normal tissue. In the absence of normalization or measurement errors, the normal (or *copy-neutral*) clones would correspond to a $\log_2$ ratio of zero because the normal and tumor DNA fragments both have two copies. The log-intensity ratios of single copy losses would be exactly $\log_2 1/2 = -1$ and those of single copy gains would be $\log_2 3/2 = 0.58$. Multiple copy gains or *amplifications*, often associated with oncogenes, would correspond to data belonging to the sequence: $\log_2 4/2, \log_2 5/2, \ldots$. Losses of both copies or deletions, often associated with tumor-suppressor mutations, would correspond to a value of $-\infty$. In this hypothetical situation, the genomic alterations can be easily deduced from the data without statistical techniques.

For comparison with the above idealized scenario, Figure 1 plots the normalized $\log_2$ ratios of breast cancer specimen S0034 analyzed by Snijders et al. (2001). The data are available from Web Table J at `http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html` Although relatively clean by array-CGH standards, the data highlight some of the issues that necessitate the use of statistical methods. For example, even after accounting for measurement error, the $\log_2$ ratios differ considerably from the theoretical values. In particular, the numbers are typically shrunk toward zero. This is caused by several factors including contamination of the tumor sample with normal cells. There is a more subtle effect of the zero varying slightly from chromosome to chromosome due to normalization errors. There is also an obvious dependence among the intensity ratios of neighboring clones.

As increasing amounts of array-CGH data become available, there is a need for automated algorithms for characterizing the genomic profiles. A number of well-known methods strive to fulfil this need. For example, Pollack et al. (2002) propose a threshold method for identifying clones having extreme value of emissions. Cheng et al. (2003) discuss a regression-based test for altered copy numbers. Hodgson et al. (2001) use a normal mixture of three components to model the observed emissions. Olshen et al. (2004) develop a variation of binary segmentation to identify chromosomal segments with altered copy numbers. Fridlyand et al. (2004) apply an unsupervised hidden Markov model. Wang et al. (2005) build hierarchical clustering-style trees along each chromosome and select interesting clusters by controlling the False Discovery Rate. Jong et al. (2003) propose a break point model to segment the clones. Eilers and de Menezes (2005) apply quantile smoothing method, while Huang et al. (2005) use penalized least squares regression and Hsu et al. (2005) apply wavelets. Hupe et al. (2004) rely on a likelihood function with adaptively determined weights using a smoothed version of the data. Picard et al. (2005) use a penalized likelihood function. Myers et al. (2004) apply an edge filter to detect the segments. Lingjaerde et al. (2005) perform smoothing using the signs of neighboring data values, inspecting the width and magnitude of the segments to detect regions of copy number change.

A recent paper by Lai et al. (2005) makes comparisons of some of the above algorithms using real and simulated data. In evaluating the algorithms, Lai and co-authors comment that "a particularly helpful feature for future implementations of some algorithms would be to estimate the statistical significance of the detected copy number changes and then rank them accordingly." They point out that only two algorithms (those of Wang et al., 2005 and Lingjaerde et al., 2005) can actually detect copy number changes based on statistical significance. Both methods rely on false discovery rates.

In Section 2, we develop a statistical framework for detecting copy number gains and losses, identifying localized amplifications and deletions, and partitioning tumor DNA into regions of relatively stable copy number. We rely on the hidden Markov model (HMM) to account for the dependence between neighboring clones. We adopt a Bayesian approach, assuming informative priors for the model parameters that are flexible enough to allow Bayesian learning. Since the posterior distribution is analytically intractable, Section 3 develops a framework for simulation-based posterior inference. In Section 4, we demonstrate the success of the technique using publicly available data. Section 4.4 compares the proposed Bayesian HMM with some of the existing algorithms using the framework of Lai et al.~(2005).

Unlike the HMM of Fridlyand et al., which is purely a segmentation method, the likelihood function of Section 2.1 allows the use of objective decision rules based on posterior probabilities to detect copy number alterations. Unlike most of the existing array-CGH methods, the biologist is not required to subjectively decide, after the algorithm's output has been obtained, plausible thresholds for identifying changes in the number of DNA copies. The proposed framework allows the use of the simple classification scheme of Section 3.1, which is motivated by biological considerations and which makes the algorithm output easy to interpret. Section 5 uses simulation studies to compare the Bayesian HMM with alternative techniques for analyzing array-CGH data.

# 2 BAYESIAN HIDDEN MARKOV MODEL

## 2.1 Likelihood function

Since the propensity for genomic alterations varies across the chromosomes, we allow each chromosome to have a distinct set of parameters. For a given chromosome, let $L_1, \ldots, L_n$ represent the mapped clones or DNA fragments arranged from the $p$-telomere to the $q$-telomere. Let $Y_k$ denote the normalized $\log_2$ ratio observed at clone $L_k$.

As mentioned earlier, the aim of the analysis is to learn about genome-wide changes in copy number from the data. A key innovation that directly achieves this goal is a latent variable called the *copy number state* $s_k$ associated with each clone $L_k$, where $k = 1, \ldots, n$. The variable $s_k$ takes values in the set $\{1, 2, 3, 4\}$. The value $s_k = 1$ represents a copy number loss at $L_k$ that could be either a single copy loss or a deletion; $s_k = 2$ represents the copy-neutral state; $s_k = 3$ represents a single copy gain; $s_k = 4$ represents an amplification (i.e. multiple copy gain) at $L_k$. The parameters of interest that summarize the copy number changes on the chromosome are $s_1, \ldots, s_n$.

For $j = 1, \ldots, 4$, we define $\mu_j$ as the expected $\log_2$ ratio of all clones $L_k$ for which $s_k = j$. For example, the expected $\log_2$ ratio of single copy gains is $\mu_3$. The theoretical value of $\mu_3$ is 0.58, but as mentioned earlier, the actual value could be different for many reasons, e.g. contamination of tumor samples with normal tissue. Although the $\mu_j$'s are unknown parameters, the biological interpretation associated with the state space of $s_k$ allows us to assume the ordering: $\mu_1 < \mu_2 < \mu_3 < \mu_4$. Conditional on the copy number states, the normalized $\log_2$ ratios are assumed to be distributed as $Y_k \overset{indep}{\sim} N(\mu_{s_k}, \sigma^2_{s_k})$, where $k = 1, \ldots, n$.

We model the dependence of the neighboring clones using a hidden Markov model (Rabiner, 1989; MacDonald and Zuchchini, 1997; Durbin et al., 1998). For any $m$ indices for which $1 \leq k_1 \leq \ldots \leq k_m \leq n$, a Markov model for the copy number states assumes that $\Pr\left[s_{k_m} \mid s_1, \ldots, s_{k_{m-1}}\right] = \Pr\left[s_{k_m} \mid s_{k_{m-1}}\right]$. The hidden Markov model (HMM) assumes that the conditional probabilities of neighboring clones is $\Pr\left[s_{k+1} \mid s_k\right] = a_{s_k s_{k+1}}$ where $\boldsymbol{A} = ((a_{ij}))$ is the matrix of stationary transition probabilities. We assume that the elements of $\boldsymbol{A}$ are strictly positive. The hidden Markov process is then aperiodic, irreducible and its four states are positive recurrent. Transition matrix $\boldsymbol{A}$ has a unique stationary distribution, denoted by $\pi_{\boldsymbol{A}} = (\pi_{\boldsymbol{A}}(1), \pi_{\boldsymbol{A}}(2), \pi_{\boldsymbol{A}}(3), \pi_{\boldsymbol{A}}(4))$, where $\pi_{\boldsymbol{A}}(i)$ is strictly positive for state $i = 1, \cdots, 4$ (Karlin and Taylor, 1981). We also assume that $s_1$, the copy number state of the first clone, is distributed as $\pi_{\boldsymbol{A}}$. Together with the hidden Markov assumption, this uniquely determines the joint likelihood of a given sequence $s_1, \ldots, s_n$. The chromosome-specific hyperparameters are therefore the transition probability matrix $\boldsymbol{A}$, means $\{\mu_1, \mu_2, \mu_3, \mu_4\}$ and error variances $\{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2\}$.

## 2.2 Priors

The Bayesian approach assumes priors for all unknown parameters. Since the copy number states defined in Section 2.1 have a well-defined meaning, this facilitates the use of informative priors based on our knowledge of array-CGH data. For example, we know that the mean $\mu_1$ of copy number losses cannot be a positive number, although individual $\log_2$ ratios that correspond to copy number losses could be. Independent priors are assumed for the chromosome-specific parameters. This results in independent posteriors for all the chromosomes. The marginal posterior $[s_1, \ldots, s_n \mid Y_1, \ldots, Y_n]$ is of interest. As with many Bayesian applications, the marginal posterior cannot be analytically computed and so simulation-based techniques are necessary. While analyzing HMMs, a key issue is label switching (refer to Scott, 2002 for a discussion). This is an identifiability issue where the likelihood is invariant under arbitrary permutations of the state space labels, resulting in inefficient exploration of the posterior by simulation. The likelihood of Section 2.1 avoids this problem by assuming order constraints. Specifically, the constraint $\mu_1 < \mu_2 < \mu_3 < \mu_4$ is violated on permutating the labels.

Let $X \sim F \cdot I(c < X < d)$ imply that $X$ has the distribution $F$ restricted to the interval $(c, d)$ with the density suitably rescaled to make it a random variable. For the mean $\mu_1$ corresponding to copy number losses, we assume the prior $\mu_1 \sim N\left(-1, \tau_1^2\right) \cdot I\left(\mu_1 < -\epsilon\right)$ where $\epsilon > 0$. We comment below on the choice of $\epsilon$. For the copy-neutral state, we assume $\mu_2 \sim N\left(0, \tau_2^2\right) \cdot I\left(-\epsilon < \mu_2 < \epsilon\right)$. For single copy gains, we assume $\mu_3 \sim N\left(0.58, \tau_3^2\right) \cdot I\left(\epsilon < \mu_3 < 0.58\right)$, and for multiple copy gains, we assume $[\mu_4 \mid \mu_3, \sigma_3] \sim N\left(1, \tau_4^2\right) \cdot I\left(\mu_4 > \mu_3 + 3\sigma_3\right)$. These informative priors were chosen as follows. For $\mu_2$ and $\mu_3$, the means of the untruncated distributions are set equal to the theoretical values for pure samples. For $\mu_1$ $(\mu_4)$, the untruncated distribution is centered at the theoretical value for a loss (gain) of one copy. The lower endpoint of the support of $\mu_4$ is chosen to be $3\sigma_3$ units away from $\mu_3$ so that a small fraction of single copy gains are erroneously classified as multiple copy gains. The results are not sensitive to choices of $\tau_1$, $\tau_2$ and $\tau_3$ belonging to the interval $[0.5, 2]$. Setting $\tau_4 \leq 2$ guarantees sufficiently high prior probability to large values of $\mu_4$ associated with high-level amplifications. We set $\tau_1 = \tau_2 = \tau_3 = 1$ and set $\tau_4 = 2$ in Sections 4 and 5.

Unlike a threshold-based approach for detecting changes in copy number, the constant $\epsilon$ determines the boundaries for the means $\mu_j$ rather than for the $\log_2$ ratios. These boundaries are not the same as threshold levels for detecting gains and losses. In fact, our assumptions allow *positive* log-intensity ratios for copy number losses, especially with large measurement errors, although $\mu_1$ itself cannot exceed $-\epsilon$. In our analyses of actual array-CGH data, we have found the results to be robust to choices of $\epsilon$ in the range $[0.05, 0.15]$. This is shown by verified in Section 5.2. For all our analyses, we set $\epsilon = 0.1$.

For the measurement error precisions, we assume the priors $\sigma_j^{-2} \sim \text{gamma}\,(1,1) \cdot I(\sigma_j^{-2} > 6)$ for $j = 1, 2, 3$, and $\sigma_4^{-2} \sim \text{gamma}\,(1,1)$. For the states $j = 1, 2, 3$, the assumption $\sigma_j^{-2} > 6$ is equivalent to $\sigma_j < 0.41$. This assumption is mild because typical array-CGH data suggest much lower within-group variability for the states 1, 2 and 3. The support of $\sigma_4^{-2}$ is not bounded below because state 4 is an aggregation of multiple copy gains which usually results in a higher within-group variability (i.e. smaller precision).

We assume independent Dirichlet priors on $\Re^4$ for the rows of the stochastic matrix $\boldsymbol{A}$, since this distribution has the set of all probability 4-tuples as its support. That is, with $\boldsymbol{a}_i$ denoting the $i^{th}$ row of matrix $\boldsymbol{A}$, we assume that $\boldsymbol{a}_i \overset{indep}{\sim} \mathcal{D}_4\left(\theta_{i1}, \theta_{i2}, \theta_{i3}, \theta_{i4}\right)$ where $i = 1, \cdots, 4$ and the constants $\{\theta_{ij}\}$ are positive. As shown in Section 5.2, the results are not affected by the choices of $\theta_{ij}$ that are small in comparison to $n$. We fixed the $\theta_{ij}$'s equal to one in Sections 4 and 5.

The above priors are found to work consistently well for array-CGH data. They are flexible enough to allow Bayesian learning and information sharing across the clones. We find in Sections 4 and 5 that the posterior inference is reliable and sensitive to the characteristics of the data.

# 3 CHARACTERIZING ARRAY-CGH PROFILES

We rely on simulation-based methods for inference because the posterior distribution cannot be investigated by mathematical analysis or numerical integration. An efficient Metropolis-within-Gibbs algorithm for generating posterior samples of the parameters is given in the Appendix. The algorithm generates the parameters in blocks conditional on the remaining parameters and the data. The transition matrix $\boldsymbol{A}$ is generated using an independent-proposal Metropolis-Hastings algorithm. The copy number states are simulated by a stochastic version of the forward-backward algorithm (Chib, 1996; Robert, Ryden and Titterington, 1999) that mixes faster than a Gibbs sampler (refer to Scott, 2002). The remaining model parameters are generated by Gibbs sampling. The algorithm has been implemented using R and will soon be publicly available.

## 3.1 Classification scheme

The generated copy number states represent draws from the marginal posterior of interest, $[s_1, \ldots, s_n \mid Y_1, \ldots, Y_n]$. For each MCMC draw, the generated states are inspected and, possibly non-exclusively, classified as *focal aberrations*, *transition points*, *amplifications*, *outliers* and *whole chromosomal changes.* In the following discussion, *altered state* refers to a copy number state which is different from 2:

1. **Focal aberrations** represent localized regions of altered copy number: *(i)* a single clone not belonging to a telomere having an altered state different from its neighbors, *(ii)* two clones belonging to a telomere sharing a common altered state different from that of the third clone from the telomere, or *(iii)* two or

more adjacent clones mapped within 5 Mb having a common altered state different from their neighbors. Focal aberrations are used to detect transition points and outliers (defined below).

2. **Transition points** can be regarded as a property of the $n - 1$ inter-clonal spaces on the chromosome. An inter-clonal space is a transition point if it borders on two large regions associated with different copy number states. In contrast, focal aberrations represent small regions of altered copy number. A transition point is an inter-clonal space for which both of these conditions hold: *(i)* it is not adjacent to a telomere, and *(ii)* after excluding all focal aberrations on the chromosome, the neighboring clones on both sides of the inter-clonal space have different copy number states.

   *Transition points are different from segments.* Transition points differ from "segments" defined by the CBS algorithm of Olshen et al. (2004), an outstanding algorithm (refer to Lai et al., 2005) for analyzing array-CGH data. The CBS algorithm segments clones regardless of their spacing on the chromosome. A transition point, on the other hand, is associated with large-scale regions of gains and losses, and is declared only when the width of the altered region exceeds 5 Mb. For example, five contiguous clones that are highly amplified would generally be identified as a segment by the CBS algorithm (although there are examples in Section 4 where the procedure ignores obvious amplification and deletions to control the false positive rate). In contrast, if these five clones are located within 5 Mb, the Bayesian HMM algorithm labels them as focal aberrations rather than identify them as a separate region.

3. **High-level amplifications.** A clone for which $s_k = 4$.

4. **Outliers.** An outlier is a focal aberration satisfying: *(i)* $s_k = 1$ and $(Y_k - \mu_1)/\sigma_1 < -2$, or *(ii)* $s_k = 3$ and $(Y_k - \mu_3)/\sigma_3 > 2$. Type-*(i)* outliers could be associated with mutations on tumor suppressors and are labeled as **deletions**. Type-*(ii)* outliers may be associated with oncogene mutations.

5. **Whole chromosomal changes.** The entire chromosome is identified as gained or lost if all the clones except the focal aberrations have altered copy number states.

## 3.2 Posterior inference

For a given clone, the classification scheme of Section 3.1 results in a Bernoulli variable for each MCMC iterate and type of genomic alteration. For example, the $k^{th}$ clone is classified as a focal aberration ("1") for some

MCMC draws and as "0" for the remaining draws. The probability that this Bernoulli variable equals one is the posterior probability that clone $L_k$ is a focal aberration. For a sufficiently large number of MCMC samples, the average of these binary outcomes is a simulation-consistent estimate of the posterior probability. Therefore, we declare clone $L_k$ to be a focal aberration if this posterior probability exceeds 0.5, which is the Bayes decision rule corresponding to a 0-1 loss function (Berger, 1985, pp. 164). A similar method is used to identify deletions. Whole chromosomal changes correspond to a common Bernoulli outcome for all $n$ clones. A chromosomal alteration is declared if the posterior probability of a chromosome-wide alteration exceeds 0.5.

High-level amplifications could be detected by a similar method. However, a more efficient method is available as a by-product of the forward-backward algorithm, which computes the conditional probability that $s_k = 4$ given the hyperparameters and the data. Averaging these conditional probabilities over the MCMC sample gives a simulation-consistent estimate of the posterior probability that clone $L_k$ is a high-level amplification.

We have noticed a potential problem with identifying transition points based on the marginal posterior probabilities of the inter-clonal gaps. We recommend detecting the change points based on the configuration of change points having the highest *joint* posterior probability. Formally, let us write the configuration of change points as $\boldsymbol{\nu}(\boldsymbol{s}) = (g_1, \ldots, g_{n-1})$, where $g_j$ equals one if the $j^{th}$ inter-clonal gap is a change point, and equals zero otherwise. Notice that the mapping from $\boldsymbol{s}$ to $\boldsymbol{\nu}(\boldsymbol{s})$ is many-one. The posterior distribution of $\boldsymbol{\nu}(\boldsymbol{s})$ is maximized to compute $\boldsymbol{\nu}^*$, the configuration having the highest posterior probability. A simulation-consistent estimate of $\boldsymbol{\nu}^*$ is computed using the MCMC sample and is used to detect the transition points.

Summary tables and plots that are of direct interest to the biologist can now be constructed. Large-scale and localized regions of copy number change identified by the Bayesian HMM algorithm can be important tools for identifying candidate genes associated with cancer.

# 4 ILLUSTRATIONS

## 4.1 Pancreatic adenocarcinoma data

Pancreatic adenocarcinoma is among the most lethal of cancers. The disease is characterized by a high level of genomic instability from the earliest stages of the disease (Gisselsson et al., 2000 and 2001; van Heek et al., 2002). Genomic changes identified in the progression of the disease include early-stage mutations in the oncogene

*KRAS* and later-stage losses of the tumor supressors $p16^{INK4A}$, $p53$ and $SMAD4$ (Bardeesy and DePinho, 2002). Using a variety of techniques ranging from karyotype analyses, CGH and loss of heterozygosity mapping, frequent gains and losses have been mapped to regions on chromosomes 3–13, 17, 18, 21 and 22 (Johansson et al., 1992; Solinas-Toldo et al., 1996; Mahlamaki et al., 1997 and 2002; Seymour et al., 1994, among many others).

Aguirre et al. (2004) studied the array CGH profiles of 24 pancreatic adenocarcinoma cell lines and 13 primary tumor specimens. In that paper, the profiles were individually analyzed using the CBS algorithm of Olshen et al. (2004), which segments the data and computes the within-segment means but does not detect gains or losses. The CBS algorithm was first run on the unnormalized $\log_2$ ratios to obtain the distribution of the within-segment means. The tallest mode of the distribution was subtracted from the data to compute the normalized $\log_2$ ratios, which are available at `http://genomic.dfci.harvard.edu/ array_cgh.htm`. Setting thresholds in an ad-hoc manner, Aguirre et al. (2004) and declared normalized $\log_2$ ratios greater than 0.13 in magnitude as copy number changes (gains or losses), greater than 0.52 as high-level amplifications, and less than $-0.58$ as deletions. They also defined objective criteria for comparing the copy number alterations of individual array-CGH profiles. These criteria were applied to analyze the 37 tumor samples and to identify 54 frequently altered *minimal common regions* (MCRs) associated with pancreatic adenocarcinoma. In a subsequent study, candidate genes located within the MCRs were confirmed by the analysis of expression profiles.

We applied the Bayesian HMM algorithm to analyze these data and made comparisons with the CBS procedure. The complete set of results are presented in the supplementary materials. Throughout, the Bayesian HMM is found to perform reliably and compare favorably with the CBS procedure. We discuss a few examples here. Our primary reference for the MCRs associated with pancreatic cancer is Aguirre et al. (2004).

The upper left panel of Figure 2 displays the result for chromosome 8 of specimen 30. The green horizontal lines represent the within-segment means computed by the CBS algorithm. The vertical lines correspond to the transition points identified by the Bayesian HMM. We find that both algorithms picked up the overall trend in the data. However, while the end-user (often a biologist with relatively little statistical training) decides whether or not the CBS algorithm's within-segment means correspond to copy number changes, the Bayesian HMM automatically identified the first region as primarily copy-neutral and the second region as consisting of mainly single-copy gains.

In the upper right panel of Figure 2, the CBS procedure declared the first set of high intensity ratios on chromosome 12 of specimen 6 as two separate segments. This is because the CBS procedure identifies trends in the data. The Bayesian HMM, on the other hand, is motivated from the perspective of copy number change. It declared these clones as high-level amplifications and therefore as a single region. The next set of clones having lower $\log_2$ ratios were identified as focal aberrations because they are localized changes less than 2 Mb in width. The two amplified regions detected by the Bayesian HMM correspond to the two minimal common regions (MCRs) on chromosome 12 associated with copy number gains (see Table 1 of Aguirre et al.) The first MCR contains the KRAS2 gene, point mutations of which occur in more than 75% of pancreatic cancer cases (Almoguera et al., 1988). The CBS algorithm failed to detect the second MCR.

The bottom left panel of Figure 2 displays the profile for chromosome 17 of specimen 13. The region from 17p13.3 to 17q11.1 (10.36 Mb to 12.8 Mb) contains the tumor supressors *p*53 and $MKK4$. Mutations on the gene *p*53 are found in at least 50% of pancreatic adenocarcinoma cases (Caldas et al., 1994). The single probe corresponding to this region was easily detected by the Bayesian HMM as a deletion. In contrast, the CBS algorithm effectively declared the *entire* chromosome as copy-neutral.

The bottom right panel presents the array-CGH profile of chromosome 18 of specimen 2. The Bayesian HMM algorithm detected an outlier associated with a copy number loss around 48 Mb. The outlier corresponds to the $SMAD4$ tumor suppressor gene located at 18q21, a mutation on which is associated with pancreatic cancer (Bardeesy and DePinho, 2002). Aguirre and co-authors mention that the CBS procedure completely missed the well-established association with the $SMAD4$ gene, even though it was clearly visible in several specimens of the data set.

The CBS procedure often ignores obvious single-probe aberrations to control the False Discovery Rate. Such errors can be misleading, because subsequent gene validation involves experimental techniques that are much more expensive than CGH. For this reason, single-probe aberrations that are frequently observed across tumor specimens provide one of the most cost-effective avenues for further research about the underlying causes of cancer. There are many other instances of the differences between the CBS and Bayesian HMM algorithms. For example, the MCR from 68.27 to 68.85 Mb on chromosome 12 maps to highly amplified clones in 34 out of 37 specimens (see the supplementary materials). In every case, the Bayesian HMM declared them as high-level amplifications, but the CBS procedure detected only the amplification in specimen 8. The Bayesian HMM also outperformed the CBS algorithm in detecting the mutation on gene FEZ1 in specimen 26, and of the genes OZF and AKT2 in specimen 6.

The results demonstrate that the Bayesian HMM is effective not only in detecting global trends, but also highly localized changes in copy number. This feature is important in identifying genes associated with cancer (e.g. $SMAD4$ in the foregoing example) on which the point mutations do not become large-scale genomic changes as the disease progresses. The algorithm has potential for use as a diagnostic tool during the early stages of cancer.

### 4.2 Corriel cell lines

The Corriel cell line is widely regarded a "gold standard" data set and analyzed in Snijders et al. (2001). The data, normalized to the genome-wide median $\log_2$ ratio, are available in Web Tables E–H at `http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html`. A table of known karyotypes is presented in Web Table I on the same website. We compared these cytogenically mapped alterations with the profiles produced by our algorithm and verified that the results match in all the specimens. For example, for cell line GM05296, Web Table I reports a trisomy at 10q21–10q24 and a monosomy at 11p12–11p13. The array-CGH profile for chromosomes 10 and 11 of cell line GM05296 are displayed in Figure 3. The regions of gain and loss identified by the Bayesian HMM match the karyotypes presented in Web Table I. We omit the results for the other cell lines for brevity.

### 4.3 Breast cancer data

A useful feature of the Bayesian approach is that posterior probability plots can be created for the different kinds of genomic alterations. These plots provide a "bird's eye view" of the copy number alterations. They are useful in identifying genomic regions associated with the disease. The procedure can be easily automated for a large number of genomic profiles. To illustrate, we analyzed the breast cancer data given in Snijders et al. (2001). The data were normalized by centering to the genome-wide median $\log_2$ ratios. The posterior probability plot for specimen S1514 is displayed in Figure 4. There are several high-level amplifications on chromosome 20 and deletions on chromosomes 13 and 14. Consistent with Figure 4, a region of high-level amplifications is seen on the array-CGH profile of chromosome 20 in Figure 5.

### 4.4 Comparisons with some existing methods

Using the Glioblastoma Multiforme data of Bredel el al. (2005), Lai et al. (2005) evaluated 11 array-CGH algorithms based on segment detection as well as smoothing. The data was normalized using the Limma package (Smyth, 2004) and are available at `http://www.chip.org/~ppark/Supplements/Bioinformatics05b.html`. Graphical summaries of the results are presented in that paper as Figures 3 and 4. Sample GBM31 (Figure 3 of Lai et al., 2005) exhibits low signal-to-noise ratio. There is a large region of losses on chromosome 13. Lai and co-authors found that the algorithms CGHseg of Picard et al. (2005), GLAD of Hupe et al. (2004), CBS of Olshen et al. (2004) and GA of Jong et al. (2003) segmented chromosome 13 into two regions and detected the region of copy number loss. Smoothing-based methods like lowess, the quantreg algorithm of Eilers and de Menezes (2005) and wavelet algorithm of Hsu et al. (2005) were sensitive to local trends but were less effective in detecting global trends. The HMM algorithm of Fridlyand et al. (2004) did not find any segments.

We followed an identical evaluation procedure to compare the Bayesian HMM with the afore-mentioned methods. Figure 6 displays the result for sample GBM31. The partitioned regions are the same as those identified by the CGHseg, CBS, GLAD and GA algorithms. Local changes in the number of copies, identical to those collectively detected by the GLAD and CGHseg algorithms, are marked as high-level amplifications (▲) and deletions (▼).

The second data set investigated in Lai et al. (2005) is a fragment of chromosome 7 from sample GBM29 (refer to Figure 4 of that paper). The data show some high $\log_2$ intensity ratios around the EGFR locus. The algorithms CGHseg, quantreg, GLAD, wavelet and GA separated the data into three distinct amplification regions. The algorithms CBS, CLAC and ACE (Lingjaerde et al., 2005) detected two distinct regions instead of three. ChARM (Myers et al., 2004) grouped all the high $\log_2$ intensity ratios into a single region. The HMM algorithm of Fridlyand et al. (2004) did not detect the amplifications.

Figure 7 displays the results for the Bayesian HMM algorithm. The high $\log_2$ ratios are identified as high-level amplifications (▲). Unlike the algorithms investigated in Lai et al. (2005), the single clone having a highly negative value is detected by the algorithm and marked as a deletion. The amplifications are identified as focal aberrations, rather than as separate regions, because both clusters are less than 5 Mb in width.

We find that the Bayesian HMM algorithm combines the strength of the smoothing-based algorithms in detecting local features with the strength of the segmentation-based methods in detecting global trends. The reliability of the procedure is especially impressive with noisy data.

# 5 SIMULATION STUDIES

## 5.1 Comparison with non-Bayesian HMM and CBS algorithms

The frequentist analysis matching the foregoing Bayesian procedure estimates the hyperparameters of the likelihood using the Baum-Welch EM algorithm, iteratively incrementing the likelihood until relative changes in the hyperparameters become sufficiently small. Conditional on the estimated hyperparameters, the Viterbi algorithm then computes the *aposteriori* most likely sequence of states $s_1, \ldots, s_n$. This technique is not identical as the non-Bayesian HMM of Fridlyand et al. (2004). In particular, the latter technique does not assign biological meanings to the latent states and cannot directly detect changes in copy number.

To find the global maximum in the 20-dimensional hyperparameter space, the EM algorithm has to be run from several starting points. For typical array-CGH data, each run often requires hundreds of iterations to converge. Because of this, the computational costs associated with the frequentist and Bayesian analyses are often comparable. When `R` is used as the computing platform, the CBS algorithm is considerably faster than either method. However, all three approaches are computationally feasible and have negligible costs compared to the many months of experimental effort required to process the tumor specimens.

The non-Bayesian array-CGH profiles for the Section 4.1 data are presented in the supplementary materials. A detailed comparison with the Bayesian profiles reveals that the two procedures often gave similar results. However, there are many profiles for which the answers are noticeably different. Examples of such chromosome–specimen pairs include $(5, 2)$, $(5, 7)$, $(12, 10)$, $(7, 13)$, $(15, 13)$, $(5, 19)$, $(18, 31)$ and $(19, 34)$. Two of the profiles are displayed in Figure 9. The non-Bayesian hyperparameter estimates correspond to a greater value of the likelihood function than the Bayes estimates in all these examples. However, the Bayesian profiles look more reasonable.

We performed a simulation study of the differences between the methods. For each of the afore-mentioned chromosome–specimen pairs, we obtained signal-to-noise ratios that were typical of array-CGH data by setting the hyperparameters equal to the Bayes estimates. We then generated the underlying copy number states and data for $n = 200$ clones. The Bayesian and non-Bayesian HMMs were applied to infer the latent copy number states. The procedure was independently replicated 100 times. Table 1 displays the percentage of correctly labeled copy number states for the two methods. The Bayesian HMM outperforms the non-Bayesian HMM in all the cases.

Using eight *randomly* selected chromosome–specimen pairs, but an otherwise identical simulation strategy, Table 2 compares the CBS algorithm with the Bayesian and non-Bayesian HMMs. The method used by Aguirre et al. (2004) was applied to declare copy number gains and losses for the CBS algorithm. The Bayesian HMM outperforms the CBS algorithm, often substantially, in seven cases. The difference is inconclusive in one case. In six out of eight cases, the Bayesian HMM outperforms the non-Bayesian HMM, with the difference being inconclusive in one case. These results provide significant evidence in favor of the Bayesian HMM.

The Bayesian HMM is found to benefit from the informative priors of Section 2.2. Prior knowledge about array-CGH helps the procedure distinguish between competing sets of hyperparameter values that are almost equally plausible under the likelihood but not under the posterior. For example, consider the frequently encountered situation where there are very few $\log_2$ ratios are assigned to one or more copy number state. In such a situation, the likelihood alone may be unable to distinguish between the matching non-Bayesian HMM and a model having fewer than four states. This results in likelihood-based estimates where one or more of the $\mu_j$'s are approximately equal. Because of the well-defined meanings assigned to the four states of the HMM, the sequence of copy number states assigned by the non-Bayesian model often seem incorrect in such cases. The Bayesian approach is more robust in such situations. The informative priors prevent even states having very few probes and $\log_2$ ratios having a considerable amount of overlap due to high measurement error from being classified as a common state. For some data, a model having fewer states than four may be better-fitting than the proposed model. However, the states might not have a simple biological interpretation in terms of copy number change. The detection of copy number gains and losses, which is one of the main goals of the analysis, may also be less straightforward.

Several examples in Section 4.1 suggest that the Bayesian HMM is better than the CBS algorithm in detecting amplifications that are localized to a small number of probes. This advantage is of practical importance, because single-probe amplifications frequently occurring across specimens are often the focus of future, more expensive gene validation studies. To investigate the difference by a controlled simulation, we independently generated 25 data sets using the following procedure: *(i)* Fifty out of $n = 200$ clones were randomly chosen to be amplifications having a mean signal of 2 on the $\log_2$ scale. *(ii)* The remaining clones were assumed to be copy-neutral with a mean signal of zero. *(iii)* The data were generated by adding Gaussian noise with a standard deviation of 0.1 to these means.

The high signal-to-noise ratio ($SNR$) of 20 is atypical of array-CGH data. The percentage of amplified probes (25%) is also very high. However, in spite of these features that simplify the detection of copy number change, the CBS algorithm failed to detect *any* amplification. The Bayesian HMM on the other hand, correctly identified *all* the amplifications. Unsurprisingly for such a high $SNR$, the false discovery rate of the Bayesian HMM was zero for all the data sets and the average true discovery rate exceeded 99%.

### 5.2 Prior sensitivity

The preceding analyses assumed that $\epsilon = 0.1$ for the supports of the $\mu_j$'s (refer to Section 2.2) and that $\theta_{ij} = 1$ for the priors of the transition matrix rows, where $i = 1, \ldots, 4$ and $j = 1, \ldots, 4$. To alleviate concerns that the results are sensitive to the choice of $\epsilon$, we generated 100 data sets with $n = 500$ clones each. For each data set, the true means $\mu_1, \ldots, \mu_4$ were uniformly generated from narrow intervals centered respectively at $-0.5$, 0, 0.5 and 1. The standard deviations $\sigma_j$ were uniformly generated in the interval $[0.2, 0.25]$ which is typical of noisy array-CGH data. The true transition matrices were simulated as follows. For row 2 corresponding to the copy-neutral state, the off-diagonal elements were uniformly generated in the intervals $[0.01, 0.02]$; for the remaining rows, the off-diagonal elements were uniformly generated in the intervals $[0.02, 0.05]$. These nine elements uniquely determined the row-stochastic transition matrix. For $k = 1, \ldots, 500$, the copy number states $s_k$ were then generated and the data were obtained by adding Gaussian noise to the means $\mu_{s_k}$.

For $\epsilon$ belonging to a grid of points in the interval $[0.05, 0.15]$, the Bayesian HMM was used to analyze each simulated data set. The posterior expectations of the means $\mu_j$, the true discovery rates and false discovery rates were found to be robust to the choice of $\epsilon$. Figures 8 plots the estimates of $\mu_1, \ldots, \mu_4$ for three randomly chosen data sets as $\epsilon$ varies. The flatness of the lines provides evidence of the lack of sensitivity to $\epsilon \in [0.05, 0.15]$. The results were also found to be robust to $\{\theta_{ij}\}_{i,j}$ that were small compared to $n$.

# 6 CONCLUSIONS

We propose a Bayesian hierarchical approach relying on a hidden Markov model for analyzing array-CGH data. The informative priors allow Bayesian learning from the data. One of the strengths of the fully automated approach is the ability to detect copy number changes like gains, losses, amplifications, outliers and transition points based on the posterior. Summaries of the array-CGH profiles are generated. The profiles can then be compared across individuals to identify the genomic alterations involved in the disease pathogenesis.

The examples of Section 4 demonstrate the reliability of the Bayesian HMM. The sensitivity of the algorithm to individual probes often allows us to find candidate genes that are missed by other algorithms. The performance of the algorithm is impressive not only for the "gold-standard" Corriel cell lines but also for the Glioblastoma data set of Bredel el al. (2005) having high measurement error. Combined with the results presented in Lai et al. (2005), the latter analysis reveals a very favorable comparison with outstanding algorithms like those of Picard et al. (2005) and Olshen et al. (2004). Section 5 compares the Bayesian HMM and alternative algorithms using controlled simulations. The results confirm the accuracy of the approach.

A strength of the Bayesian HMM is that it relies on essentially *no* tuning parameters. Unlike many other algorithms (see Lai et al., 2005), the user is only required to input the normalized $\log_2$ ratios. This is a convenient feature for the end-user with little or no statistical training. In all our analyses, we have used the default parameterizations specified in Section 2.2. Certain features of the Bayesian HMM may be changed to produce a different result. Possible features include the constant $\epsilon$ in the prior specification of the means $\mu_j$ and the constants $\theta_{ij}$ in the transition matrix priors in Section 2.2. However, the simulation study in Section 5.2 and our own experience with the algorithm indicate that the results are robust to variations in these quantities. The informative priors for the means $\mu_j$ substantially influence the results, as we find in Section 5.1 on comparing the Bayesian HMM with the matching non-Bayesian model. However, the order constraints on the $\mu_j$'s and the biological meanings assigned to $s_k \in \{1, 2, 3, 4\}$ allow the specification of priors that work consistently well across different data sets. For this reason, we recommend using the default parameterizations of the Bayesian HMM for most array-CGH applications.

# REFERENCES

Aguirre A. J., Brennan C., Bailey G., Sinha R., Feng B., Leo C., Zhang Y., Zhang J., Gans J. D., Bardeesy N., Cauwels C., Cordon-Cardo C., Redston M. S., DePinho R. A. and Chin L. (2004). High-resolution characterization of the pancreatic adenocarcinoma genome. *Proceedings of the National Academy of Sciences USA* **101**, 9067–9072.

Almoguera, C., Shibata, D., Forrester, K., Martin, J., Arnheim, N., Perucho, M. (1988). Most human carcinomas of the exocrine pancreas contain mutant c-K-ras genes. *Cell* **53**, 549–554.

Bardeesy, N. and DePinho, R. A. (2002). Pancreatic cancer biology and genetics. *Nature Reviews Cancer* **2**, 897-909.

Brennan, C., Brennan, C., Zhang, Y., Leo, C., Feng, B., Cauwels, C., Aguirre, A. J., Kim, M., Protopopov, A., and Chin, L. (2004). High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Research* **64**, 4744-4748.

Brown, C. S., Goodwin, P. C. and Sorger P. K. (2001). Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences USA* **98**, 8944–8949.

Caldas, C., Hahn, S. A., da Costa, L. T., Redston, M. S., Schutte, M., Seymour, A. B., Weinstein, C. L., Hruban, R. H., Yeo, C. J., Kern, S. E. (1994). Frequent somatic mutations and homozygous deletions of the p16 (MTS1) gene in pancreatic adenocarcinoma. *Nature Genetics* **8**, 27–32.

Cheng, C., Kimmel, R., Neiman, P. and Zhao, L. P. (2003). Array rank order regression analysis for the detection of gene copy-number changes in human cancer. *Genomics* **82**, 122–129.

Chib, S. (1996). Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models. *Journal of Econometrics* **75**, 79–97.

Durbin, R., Eddy, S., Krogh, A., and Michison, G. (1998). Biological Sequence Analysis. *Cambridge University Press*, 1998.

Eilers, P. H. C. and de Menezes, R. X. (2005). Quantile smoothing of array CGH data. *Bioinformatics* **21**, 1146-1153.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., Jain, A. N. (2004). Application of Hidden Markov Models to the analysis of the array CGH data. *Journal of Multivariate Analysis* **90**, pp. 132–153.

Gisselsson, D., Pettersson, L., Hoglund, M., Heidenblad, M., Gorunova, L., Wiegant, J., Mertens, F., Dal Cin, P., Mitelman, F. and Mandahl, N. (2000). Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proceedings of the National Academy of Sciences, USA* **97**, 5357-5362.

Gisselsson, D., Jonson, T., Petersen, A., Strombeck, B., Dal Cin, P., Hoglund, M.,Mitelman, F., Mertens, F. and Mandahl, N. (2001). Telomere dysfunction triggers extensive DNA fragmentation and evolution of complex chromosome abnormalities in human malignant tumors. *Proceedings of the National Academy of Sciences, USA* **98**, 12683-12688.

van Heek, N. T., Meeker, A. K., Kern, S. E., Yeo, C. J., Lillemoe, K. D., Cameron, J. L., Offerhaus, G. J., Hicks, J. L., Wilentz, R. E., Goggins, M. G., et al. (2002). Telomere Shortening Is Nearly Universal in Pancreatic Intraepithelial Neoplasia. *American Journal Of Pathology* **161**, 1541-1547.

Hodgson, G., Hager, J., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D., Pinkel, D., Collins, C. et al., (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **929**, 459-464.

Huang, T., Wu, B., Lizardi, P., and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics* **21**, 3811 – 3817.

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.

Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L. and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6**, 211-226.

Johansson, B., Bardi, G., Heim, S., Mandahl, N., Mertens, F., Bak-Jensen, E., Andren- Sandberg, A. and Mitelman, F. (1992). Nonrandom chromosomal rearrangements in pancreatic carcinomas. *Cancer* **69**, 1674-1681.

Jong, K., Marchiori, E., Vaart, A., Ylstra, B., Weiss, M. and Meijer, G. (2003). Chromosomal breakpoint detection in human cancer. In *Applications of evolutionary computing: Evolutionary computation and bioinformatics*, Vol. 2611, Springer, pp. 54–65.

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.

Karlin, S. and Taylor, H. M. (1975). A First Course in Stochastic Processes, $2^{nd}$ edition. Academic Press, New York.

Khojasteh M., Lam W. L., Ward R. K., MacAulay C. (2005). A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics* **6**, 274.

Lai, W., Johnson, M. J., Kucherlapati, R., Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763-3770.

Lengauer, C., Kinzler, K. and Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature* **396**.

Lingjaerde, O. C., Baumbusch, L. O., Liestol, K., Glad, I. K. and Borresen-Dale, A.-L. (2005). CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* **21**, 8218-22.

Myers, C. L., Dunham, M. J., Kung, S. Y. and Troyanskaya, O. G. (2004). Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics* **20**, 3533-3543.

MacDonald, I.L. and Zuchchini, W. (1997). Hidden Markov and Other Models for Discrete-value Time Series. Boca Raton: Chapman & Hall, Inc.

McLachlan, G. J., Do, K.-A. and Ambroise C. (2004). Analyzing Microarray Gene Expression Data. Hoboken, New Jersey: John Wiley & Sons, Inc.

Mahlamaki, E. H., Barlund, M., Tanner, M., Gorunova, L., Hoglund, M., Karhu, R. and Kallioniemi, A. (2002). Frequent amplification of 8q24, 11q, 17q, and 20q-specific genes in pancreatic cancer. *Genes Chromosomes Cancer* **35**, 353-358.

Mahlamaki, E. H., Hoglund, M., Gorunova, L., Karhu, R., Dawiskiba, S., Andren-Sandberg, A., Kallioniemi, O. P. and Johansson, B. (1997). Comparative genomic hybridization reveals frequent gains of 20q, 8q, 11q, 12p, and 17q, and losses of 18q, 9p, and 15q in pancreatic cancer. *Genes Chromosomes Cancer* **20**, 383-391.

Murphy, K. M., Brune, K. A., Griffin, C., Sollenberger, J. E., Petersen, G. M., Bansal, R., Hruban, R. H. and Kern, S. E. (2002). Evaluation of candidate genes MAP2K4, MADH4, ACVR1B, and BRCA2 in familial pancreatic cancer: deleterious BRCA2 mutations in 17

Olshen A. B., Venkatraman E. S., Lucito R., Wigler M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **4**, 557–572.

Pasternak, J. J. (1999). An introduction to human molecular genetics: mechanism of inherited diseases. Fitzgerald Science Press, Bethesda, MD.

Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, J.-J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6**, 27.

Pinkel, D. and Albertson, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* **37**, Suppl. 11-17.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W., Chen, C., Zhai, Y. et al., (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**.

Pollack, J.R., Perou C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S.

S., Botstein, D., Brown, P. O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* **23**, 41-46.

Pollack, J., Sorlie, T., Perou, C., Rees, C., Jeffrey, S., Lonning, P., Tibshirani, R., Botstein, D., Borresen-Dale, A. and Brown, P. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences, USA* **99**, 12963–12968.

Rabiner, L.R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286.

Robert, C. P., Ryden T., and Titterington D. M. (1999). Convergence controls for MCMC algorithms, with applications to hidden Markov chains. *Journal of Statistical Computing and Simulation* **64**, 327–355.

Scott, S. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association* **97**, 337–351.

Seymour, A. B., Hruban, R. H., Redston, M., Caldas, C., Powell, S. M., Kinzler, K. W., Yeo, C. J. and Kern, S. E. (1994). Allelotype of pancreatic adenocarcinoma. *Cancer Research* 54, 2761-2764.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D., Albertson, D. G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29**.

Solinas-Toldo, S. et al. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20**, 399-407.

Wang, P., Kim, Y., Pollack, J., Balasubramanian, N. and Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics* **6**, pp. 45–58.

# APPENDIX

## An MCMC algorithm

The following algorithm is independently run for each chromosome to generate an MCMC sample for the chromosomal parameters. We group the model parameters into four blocks, namely, $B_1 = \boldsymbol{A}$, $B_2 = (s_1, \ldots, s_n)$,

$B_3 = (\mu_1, \mu_3, \mu_4)$, and $B_4 = \left(\sigma^2, \sigma_4^2\right)$. The starting values of the parameters are generated from the priors. The algorithm iteratively generates each of the four blocks conditional on the remaining blocks and the data. Let $B_1^{(v-1)}, \ldots, B_4^{(v-1)}$ denote the values of the blocks at the $(v-1)^{st}$ iteration. In the next iteration, the blocks are generated as follows:

**Updating block $B_1$.** The transition matrix is generated using a Metropolis-Hastings step because the normalizing constant of the full conditional cannot be computed in closed form. This step makes independent proposals from a distribution that closely approximates the full conditional of the transition matrix. The proposal is accepted or rejected with a probability that compensates for the approximation. Typically, most of the Metropolis-Hastings proposals are accepted. Using the the copy number states generated at iteration $v-1$, we compute the number of transitions from state $i$ to state $j$, denoted by $u_{ij}^{(v)} = \sum_{k=1}^{n-1} I\left(s_k^{(v-1)} = i, s_{k+1}^{(v-1)} = j\right)$, where $i, j = 1, \ldots, 4$. We generate a proposal $\boldsymbol{C}$ for the transition matrix from the distributions $[\boldsymbol{c}_i \mid \boldsymbol{Y}, B_{-1}] \sim \mathcal{D}_3\left(1 + u_{i1}^{(v)}, 1 + u_{i2}^{(v)}, 1 + u_{i3}^{(v)}, 1 + u_{i4}^{(v)}\right)$, where row $i = 1, \ldots, 4$, and $B_{-1}$ denotes the blocks, $\{B_2, B_3, B_4\}$. The proposal ignores the marginal distribution of state $s_1$ and so it differs from the full conditional of the transition matrix. To compensate for this, we accept the proposal (in other words, set $\boldsymbol{A}^{(v)}{=}\boldsymbol{C}$) with probability $\beta$, where $\beta = \min\left\{1, \pi_{\boldsymbol{C}}(s_1^{(v-1)})/\pi_{\boldsymbol{A}^{(v-1)}}(s_1^{(v-1)})\right\}$, and otherwise reject the proposal (in other words, set $\boldsymbol{A}^{(v)}{=}\boldsymbol{A}^{(\boldsymbol{v-1})}$). As defined earlier, $\pi_{\boldsymbol{D}}(s)$ denotes the probability of state $s$ under the stationary distribution of a given transition matrix $\boldsymbol{D}$.

**Updating block $B_2$.** The copy number states are generated by a stochastic version of the forward-backward algorithm. We compute the distribution $[s_n \mid B_{-2}, Y_1, \ldots, Y_n]$ at the beginning of the backward step. We generate $s_n$ from this distribution. The backward step is continued to compute and generate a draw the distribution $[s_{n-1} \mid s_n, B_{-2}, Y_1, \ldots, Y_n]$. The sequence of computing and generating a draw from $[s_k \mid s_{k+1}, B_{-2}, Y_1, \ldots, Y_n]$ is iterated for $k = n-2$ down to $k = 1$. This produces a sample from the joint distribution $[s_1, \ldots, s_n \mid B_{-2}, Y_1, \ldots, Y_n]$.

**Updating block $B_3$.** For $s = 1, \ldots, 4$, let $\delta_{0s}$ be the center of the untruncated normal distribution in the prior specification of $\mu_s$. Compute the sums $n_s = \sum_{k=1}^{n} I\left(s_k^{(v)} = s\right)$, averages $\bar{Y}_s = \frac{1}{n_s}\sum_{k=1}^{n} Y_k \cdot I(s_k^{(v)} = s)$, precisions $\theta_s^2 = \tau_s^{-2} + \left(\sigma_s^{(v-1)}/\sqrt{n_s}\right)^{-2}$ and weighted means $\gamma_s = \frac{1}{\theta_s^2}\left[\delta_{0s} \cdot \tau_s^{-2} + \bar{Y}_s \cdot \left(\frac{\sigma_s^{(v-1)}}{\sqrt{n_s}}\right)^{-2}\right]$. For $s = 1, \ldots, 4$, generate $\left[\mu_s^{(v)} \mid Y, B_{-3}\right] \sim N\left(\gamma_s, \theta_s^{-2}\right) \cdot I_s$, where the intervals $I_s$ denotes the support of the $\mu_s$ (see prior specification).

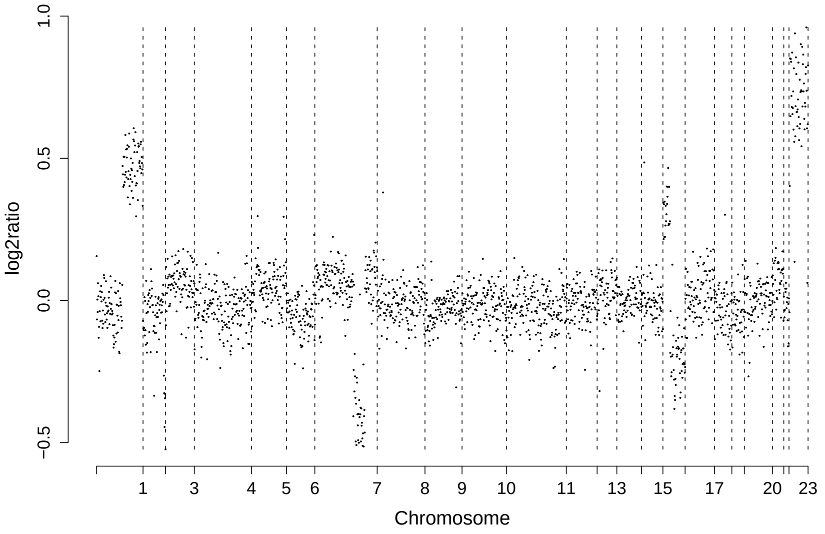

Figure 1: Normalized copy number ratios of a comparison of DNA from cell strain S0034 (Snijders et al., 2001) with normal DNA. The BACs are ordered by position in the genome beginning at 1p and ending at Xq. The vertical bars indicate borders between chromosomes.

**Updating block** $B_4$**.** For $j = 1, \ldots, 4$, compute $n_j = \sum_{k=1}^{n} I\left(s_k^{(v)} = j\right)$ and $V_j = \sum_{k=1}^{n}\left(Y_k - \mu_{s_k}^{(v)}\right)^2 \cdot I\left(s_k^{(v)} = j\right)$. Generate

$$\left[\sigma_j^{(v)} \mid Y, B_{-4}\right] \sim \left[\text{gamma}\left(1 + \frac{n_j}{2}, \epsilon + \frac{V_j}{2}\right)\right]^{-0.5}.$$

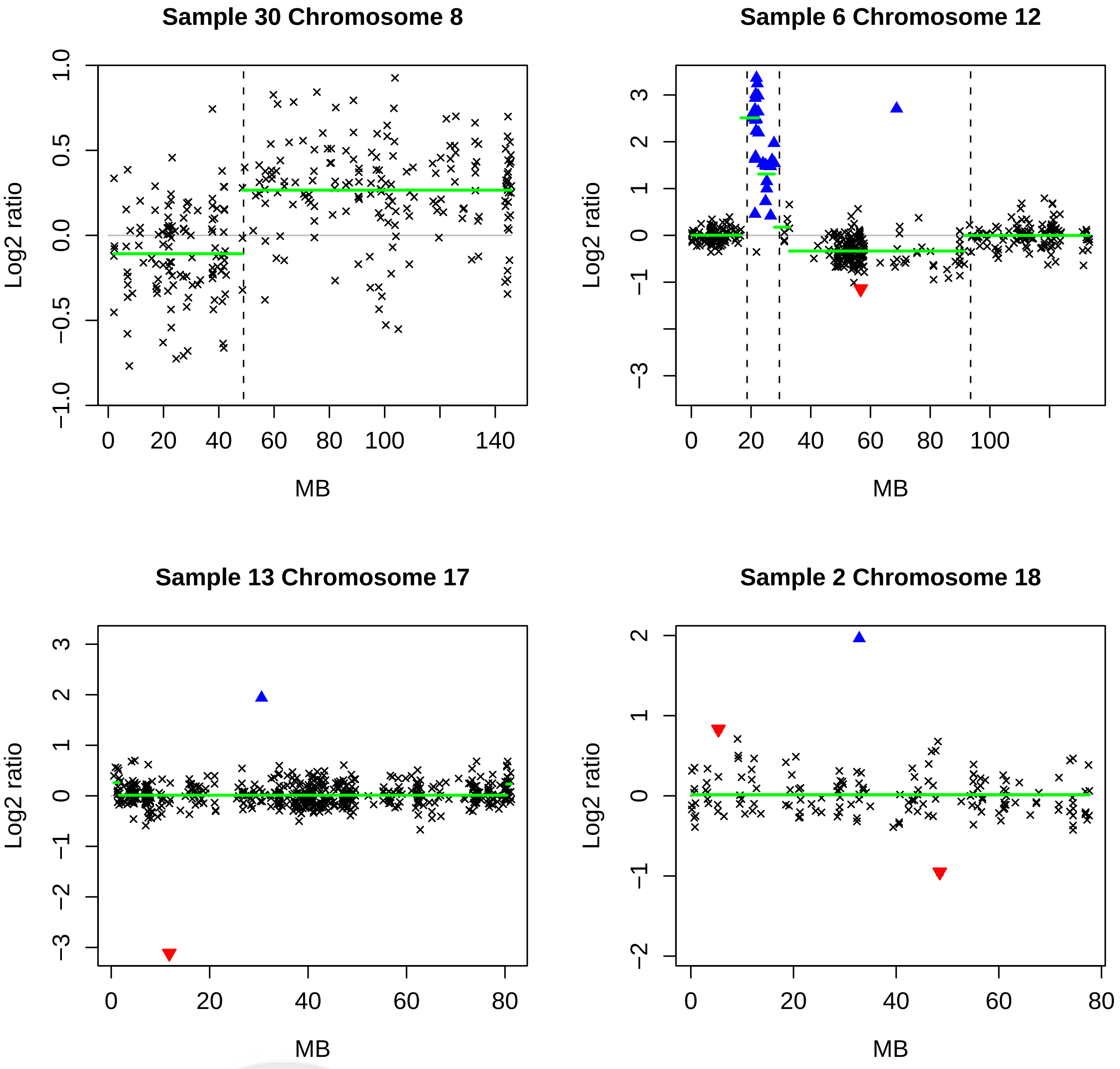

Figure 2: Array-CGH profiles of some pancreatic cancer specimens. In each panel, the clonal distance in Mb from the *p* telomere has been plotted on the x-axis. High-level amplifications and outliers are respectively indicated by ▲ and ▼. The broken vertical lines represent transition points. For comparison, the green lines display the segment means computed by the CBS algorithm. See Section 4.1 for further discussion.

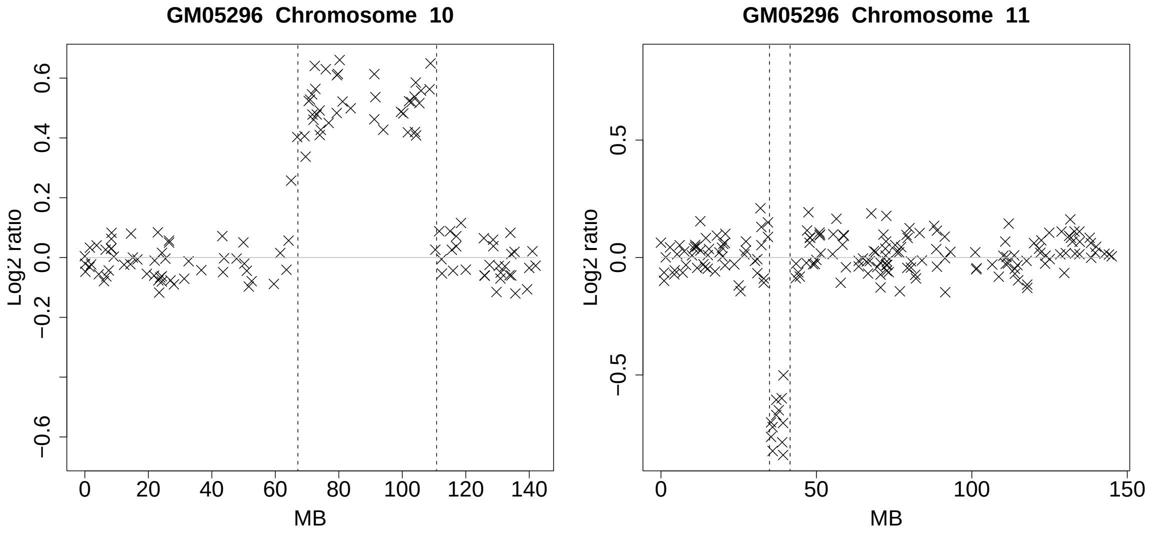

Figure 3: Array-CGH profile of chromosomes 10 and 11 of Corriel cell strain GM05296. The x-axis displays the clonal distance from the $p$ telomere in Mb. The broken vertical lines represent transition points.

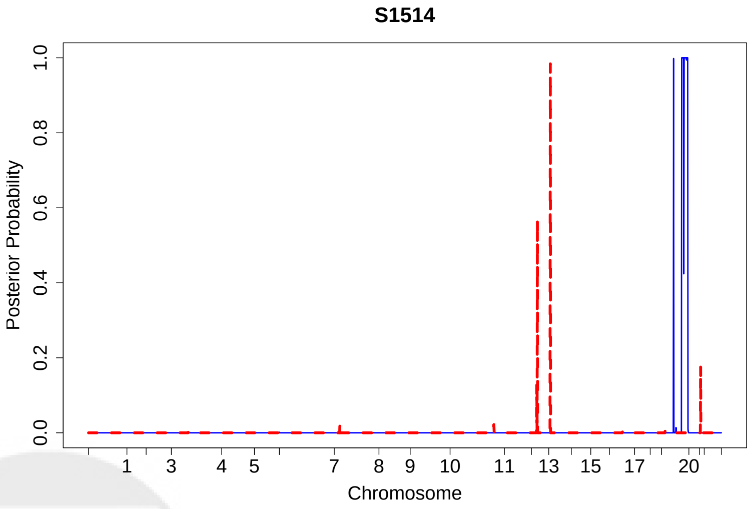

Figure 4: Posterior probabilities of genomic alterations for specimen S1514. The solid line represents high-level amplifications while the dashed line corresponds to deletions. The numbers on the horizontal axis represent the q telomere of the chromosomes. The BACs are ordered by position in the genome beginning at 1p and ending at Xq.

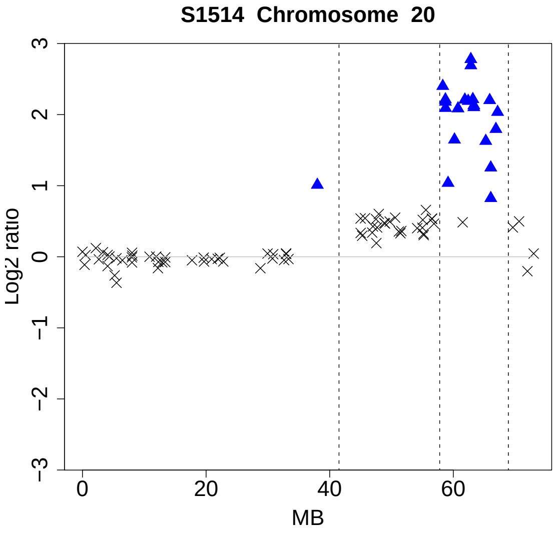

Figure 5: Array-CGH profile of chromosome 20 of S1514. The x-axis represents clonal distance in Mb from the *p* telomere. The broken vertical lines represent transition points. High-level amplifications are shown using ▲.

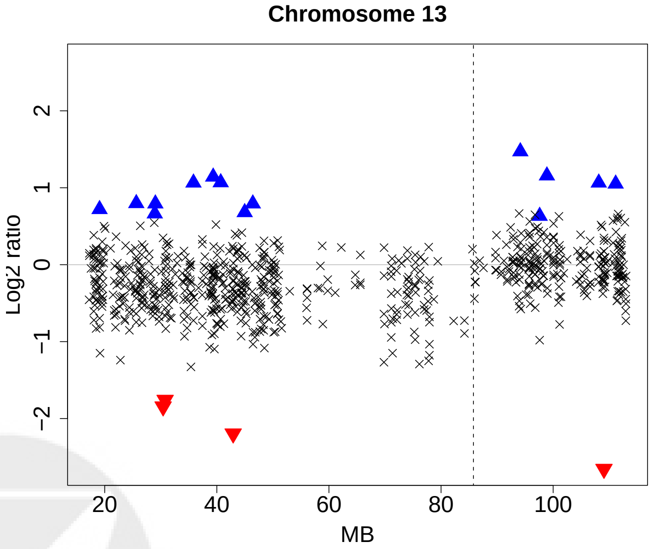

Figure 6: Array-CGH profile of chromosome 13 of GBM31. The clonal distance in Mb from the *p* telomere is plotted on the x-axis. High-level amplifications and outliers are respectively indicated using ▲ and ▼. The broken vertical line represents a transition point.

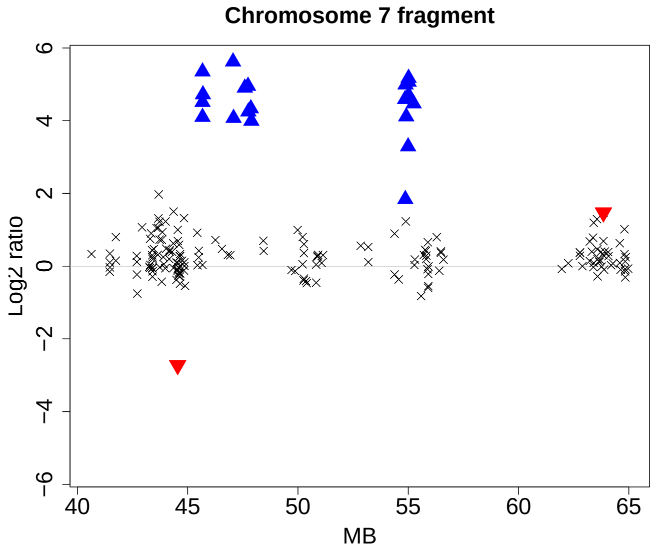

Figure 7: Partial array-CGH profile of chromosome 7 of GBM29. The clonal distance in Mb from the $p$ telomere is plotted on the x-axis. High-level amplifications and outliers are respectively indicated using ▲ and ▼.

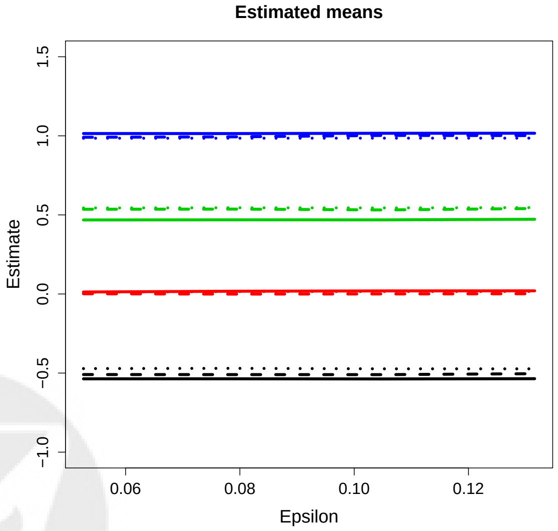

Figure 8: Estimated means $\hat{E}[\mu_j|\boldsymbol{Y}]$ for three independently generated data sets (shown by solid, dashed and dotted lines) plotted against $\epsilon$.

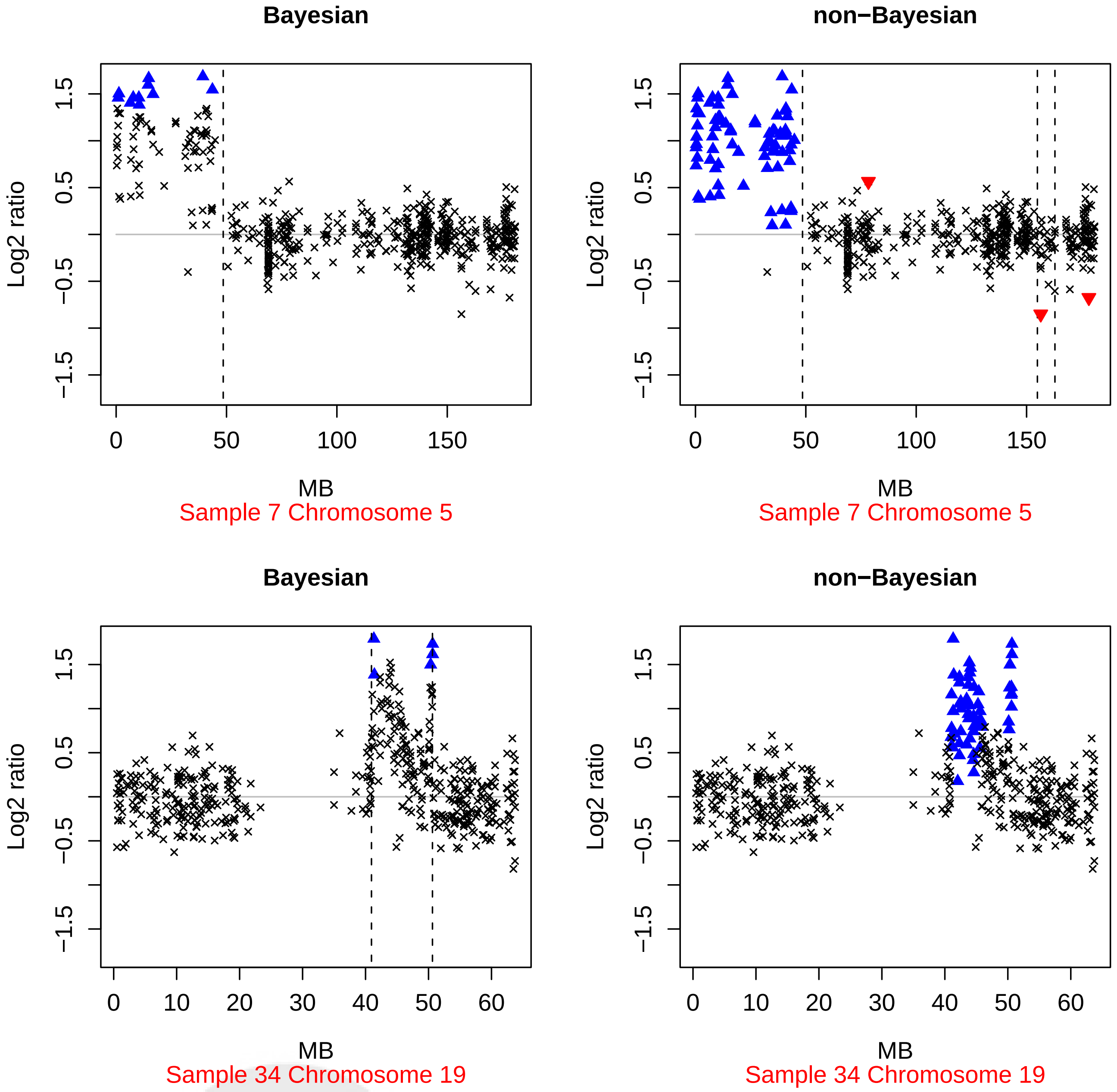

Figure 9: Examples from Section 4.1 where the Bayesian and non-Bayesian array-CGH profiles are different. The upper panels correspond to chromosome 5 of sample 7 and the lower panels correspond to chromosome 19 of sample 34. The clonal distance in Mb from the $p$ telomere has been plotted on the x-axis. High-level amplifications and outliers are indicated using ▲ and ▼ respectively. The broken vertical lines represents transition points.

| **Source** | | **Bayesian HMM** | | **Non-Bayesian HMM** | |
|---|---|---|---|---|---|
| Chromosome | Specimen | % accuracy | SE | % accuracy | SE |
| 5 | 2 | 94.81 | 0.789 | 86.89 | 1.685 |
| 5 | 7 | 91.99 | 1.188 | 81.44 | 1.942 |
| 12 | 10 | 95.22 | 0.390 | 89.08 | 1.378 |
| 7 | 13 | 92.41 | 1.019 | 80.09 | 2.333 |
| 15 | 13 | 92.42 | 1.322 | 82.55 | 1.649 |
| 5 | 19 | 88.02 | 2.189 | 73.09 | 2.873 |
| 18 | 31 | 84.95 | 2.512 | 71.17 | 2.448 |
| 19 | 34 | 88.13 | 2.000 | 72.10 | 2.124 |

Table 1: Estimated percentages of correctly discovered copy number states for the Bayesian and non-Bayesian methods, along with the estimated standard errors. The estimates were based on 100 independently generated data sets. The first two columns specify the chromosome and specimen numbers of the Section 4.1 data set whose the estimated hyperparameters were used to generate the data. See the text for an explanation.

| **Source** | | **Bayesian HMM** | | **Non-Bayesian HMM** | | **CBS** | |
|---|---|---|---|---|---|---|---|
| Chromosome | Specimen | % accuracy | SE | % accuracy | SE | % accuracy | SE |
| 13 | 33 | 94.38 | 1.203 | 72.01 | 2.634 | 67.72 | 3.512 |
| 19 | 4 | 88.20 | 1.129 | 87.94 | 0.534 | 75.36 | 1.726 |
| 14 | 1 | 87.35 | 1.893 | 76.47 | 1.834 | 86.70 | 0.426 |
| 12 | 17 | 80.84 | 1.736 | 76.11 | 1.453 | 44.12 | 1.791 |
| 1 | 24 | 40.64 | 2.512 | 54.31 | 1.460 | 35.37 | 2.470 |
| 3 | 35 | 96.03 | 0.239 | 72.06 | 2.509 | 92.43 | 0.488 |
| 23 | 12 | 74.31 | 3.417 | 65.2 | 2.420 | 58.08 | 3.311 |
| 15 | 34 | 90.79 | 2.164 | 68.3 | 2.798 | 55.22 | 4.175 |

Table 2: Estimated percentages of correctly discovered copy number states for the Bayesian and non-Bayesian methods, along with the estimated standard errors. The estimates were based on 100 independently generated data sets. The first two columns specify the chromosome and specimen numbers of the Section 4.1 data set whose the estimated hyperparameters were used to generate the data. See the text for an explanation.