International Journal of Biomedical Data Mining
ISSN: 2090-4924

Wu and Zhu, Biomedical Data Mining 2015, 4:1
DOI: 10.4172/2090-4924.1000116

Research Article Open Access

A Bayesian Analysis of Copy Number Variations in Array Comparative Genomic Hybridization Data

Xiaowei Wu* and Hongxiao Zhu

Department of Statistics, Virginia Tech, 250 Drillfield Drive, Blacksburg, VA, 24061, USA

Abstract

Array Comparative Genomic Hybridization (CGH) has been widely used for detecting genomic copy number variations (CNVs). The central goal of array CGH data analysis is to accurately detect homogeneous regions of log intensity ratios which represent relative changes in DNA copy number. Various methods have been proposed in recent years. Most methods, however, do not consider correlations of neighboring probe measurements, and are usually designed for analysis at single sample level rather than detecting common or recurrent CNVs among multiple samples. We propose a Bayesian segment-based approach for efficient analysis of array CGH data. The proposed method is based on simple assumptions but is general enough to accommodate various spatial correlations among probe measurements. It also allows for multiple samples with recurrent CNVs, therefore is able to borrow strength across samples. In contrast to another probe-based approach developed in the same Bayesian framework, the segment-based approach parameterizes the mean log intensity ratios in a more appropriate way, which leads to a posterior sampling scheme based on reversible-jump Markov chain Monte Carlo. We perform a simulation study to compare these two approaches and the commonly-used circular binary segmentation method and Bayesian hidden Markov model method. The segment-based approach achieves better estimation accuracy and higher computational efficiency compared to the probe-based approach, and also provides improved results compared to the other two methods, especially for data with relatively low signal to noise ratio and high correlation. The segment-based approach is further applied to the Corriel cell lines data and Pancreatic Adenocarcinoma data.

Keywords: Copy number; Intensity ratios; MCMC; Reversible jump

Introduction

Array-based comparative genomic hybridization (CGH) is a high throughput technique that simultaneously measures relative changes in DNA copy number at thousands of genomic loci [1,2]. In array CGH experiments, test (tumor) and reference (normal) DNA samples are labeled by different fluorochrome and hybridized onto an array containing genomic clones. The resulting fluorescence intensity ratios are recorded according to the physical location of the corresponding probes on the genome, and further normalized and transformed to $\log_2$ scale to indicate genome-wide changes in copy number. The log intensity ratios therefore indicate distinct copy number states such as copy neutral, copy losses and copy gains. Ideally (without tissue contamination, measurement errors, etc.), in copy neutral regions, both test and reference DNA samples have two copies hence the log intensity ratio is $\log_2(2/2) = 0$. Similarly, in regions of single-copy loss, single-copy gain and double-copy gain, the corresponding log intensity ratios are $\log_2(1/2) = -1$, $\log_2(3/2) = 0.58$ and $\log_2(4/2) = 1$, respectively. In this study, we mainly consider these four copy number states. Multiple-copy (greater than 2) gains or amplifications can be included in the same manner if needed, and double-copy losses or deletions can be easily detected without using statistical techniques since their corresponding log intensity ratio is $\log_2(0/2) = -\infty$. In practice, the log intensity ratios do not exactly follow the theoretical values due to various experimental and biological reasons. Also there is usually a non-negligible dependence among the log intensity ratios of adjacent probes. As an example, Figure 1 shows the normalized log intensity ratios for breast cancer specimen S0034 [1].

The purpose of array CGH data analysis is to identify homogeneous regions of high or low intensity ratios, i.e., copy number variations (CNVs) in the genome. From the statistical point of view, the analysis involves estimating two types of parameters: the location of homogeneous regions and the copy number states of these regions. There is an extensive literature on the analysis of array CGH data. For example, Hodgson et al. [3] applied a hybrid least-square and maximum likelihood method to fit a mixture Gaussian distribution to a histogram of log intensity ratios. Olshen et al. [4] and Venkatraman and Olshen [5] used circular binary segmentation (CBS) algorithms to divide the genome into regions of equal copy number, based on a well-developed change point detection theory using hypothesis testing [6]. Wang et al. [7] and Picard et al. [8] adopted segmentation/clustering algorithms to select clusters with regions of genetic alterations. Smoothing techniques and Wavelet methods are proposed by Eilers and de Menezes [9] and Hsu et al. [10]. There are also hidden Markov model (HMM) based approaches which model the dependence across genome location by transition between hidden states [11,12]. In addition, Bayesian procedures have been studied for single and multiple change point problems [13-18], and have been applied to CGH data analysis [19-24]. In a recent work of Baladandayuthapani et al. [25], Bayesian functional mixed models are proposed which treat the CGH arrays as functions and accommodate subject-wise random effects. In general, most available methods can be categorized into two groups: empirical methods and model-based methods. When using empirical methods, one usually performs data processing through two steps: smoothing (or denoising) and thresholding. Though easy to implement, these methods usually lack statistical power. On the other hand, when using

***Corresponding author:** Xiaowei Wu, Department of Statistics, Virginia Tech, 250 Drillfield Drive, Blacksburg, VA, 24061, USA, Tel: 540-231-0023; E-mail: xwwu@vt.edu

Received July 22, 2015; **Accepted** September 01, 2015; **Published** September 25, 2015

Citation: Wu X, Zhu H (2015) A Bayesian Analysis of Copy Number Variations in Array Comparative Genomic Hybridization Data. Biomedical Data Mining 4: 116. doi:10.4172/2090-4924.1000116

Copyright: © 2015 Wu X, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

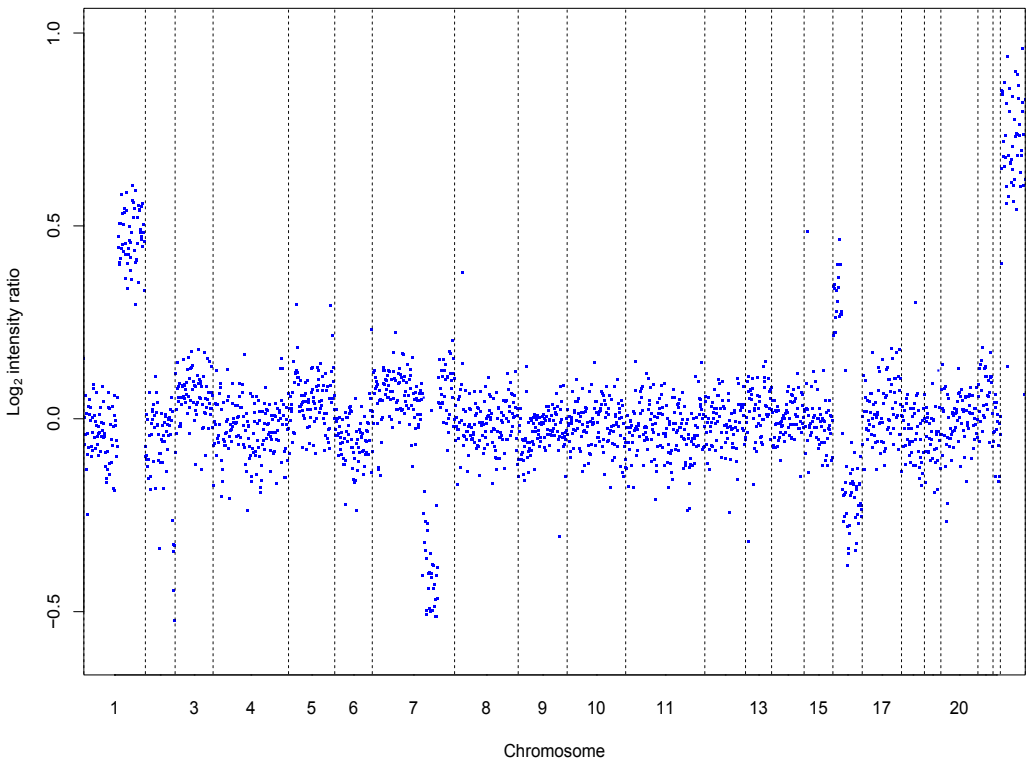

Figure 1: An example of array CGH profile. Blue dots: normalized log intensityratios of breast cancer specimen S0034 [1]; Vertical black dashedlines: borders between chromosomes.

model-based methods, one aims at fitting data with suitable statistical models and making inference on model parameters. These methods are able to provide statistical significance for determining CNVs, but they are usually complicated and computationally intensive compared to empirical methods. From the modeling perspective, many available methods are designed for single sample analysis. For recurrent CNV (i.e., multiple samples sharing the same CNV information), these methods cannot borrow strength across samples. Furthermore, most of them assume independence across probes, or independence across segments, without considering the spatial correlations across genomic locations.

In this paper, we propose a Bayesian segment-based approach for array CGH data analysis. This approach is based on simple assumptions and is computationally efficient. It is also general enough to accommodate both among-probe correlations and multiple samples with recurrent CNVs. Under the same model assumptions, there exists another probe-based approach in a similar Bayesian framework. The proposed segment-based approach differs from the probe-based approach in the parameterization of the mean log intensity ratios, which leads to a different posterior sampling scheme. We perform a simulation study to evaluate their performance under a range of signal to noise ratios (SNRs). Simulation results show that the segment-based approach achieves better estimation accuracy and higher computational efficiency than the probe-based approach. We also compare the segment-based approach to the commonly-used CBS and Bayesian HMM methods using simulated data with low SNR and high correlation. The result shows that under certain circumstances when CBS and HMM may not work well, the segment-based approach still achieves relatively reasonable estimates of the CNVs. The segment-based approach is further applied to the Corriel cell lines data and Pancreatic Adenocarcinoma data and shows good performance.

The rest of the paper is organized as follows. In Section 2 we introduce the basic framework, present details of the probe-based Bayesian approach and the segment-based Bayesian approach, and describe the corresponding MCMC algorithms. Section 3 shows the comparison results using simulated data. Section 4 demonstrates the success of the segment-based approach using publicly available Corriel cell lines data and Pancreatic Adenocarcinoma data. We conclude with a brief discussion in Section 5.

Methods

The basic framework

Suppose that the target DNA sequence has L probes. Let $\boldsymbol{x} = (x_1,\ldots x_L)^T$ be the normalized log intensity ratios. In our modeling scheme, we assume that $\boldsymbol{x}$ is a realization of a random vector $\boldsymbol{X}$, which is of dimension L, with mean $\boldsymbol{\mu} = (\mu_1,\ldots \mu_L)^T$ and covariance matrix Σ. Here μ_j is the mean log intensity ratio of probe j, $1\leq j \leq L$ and it takes value from a vector of theoretical levels $\{-1, 0, 0.58, 1\}$. Multiple-copy gains can be included in the vector of theoretical levels if needed. For a more general case where the theoretical levels are unknown, our Bayesian approach is still applicable with slight extension. More details are discussed in Section 5. The mean vector $\boldsymbol{\mu}$ contains information of copy number variations in the DNA sequence and is assumed to be piecewise constant. The covariance matrix Σ indicates the correlation structure of the probe intensity ratios.

We further make the assumption that $\boldsymbol{X}$ has a certain multivariate distribution. A variety of multivariate distributions can be used for the likelihood. In this study, we focus on the multivariate normal distributions for computational convenience. This gives the likelihood

$$f(\mathbf{x}\,|\,\boldsymbol{\mu},\Sigma)=(2\pi)^{-\frac{L}{2}}\,|\,\Sigma\,|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Note that this likelihood is only for a single sample x. If multiple independent samples $x_1,x_2,\ldots$exist, the likelihood can be simply replaced by a product of likelihoods from each sample. The advantage of multiple sample analysis will be demonstrated through simulations later on in Section 3. The parametric forms of the covariance matrix Σ can also be varied depending on the biological information and the experimental condition. One choice we will adopt is the first-order autoregressive form with correlation ρ and marginal variance ^2, i.e.,

$\Sigma=\sigma^2\begin{pmatrix}1&\rho&\cdots&\rho^{L-1}\\ \rho&1&\cdots&\rho^{L-2}\\ \vdots&\vdots&\ddots&\vdots\\ \rho^{L-1}&\rho^{L-2}&\cdots&1\end{pmatrix}$. Under this setting, the likelihood becomes

$$f(\mathbf{x}\,|\,\boldsymbol{\mu},\sigma^2,\rho)=(2\pi\sigma^2)^{-\frac{L}{2}}(1-\rho^2)^{-\frac{L-1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^TW(\mathbf{x}-\boldsymbol{\mu})\right\},\qquad(1)$$

where

$$W=\frac{1}{1-\rho^2}\begin{pmatrix}1&-\rho&0&\cdots&0&0&0\\ -\rho&1+\rho^2&-\rho&\cdots&0&0&0\\ 0&-\rho&1+\rho^2&\cdots&0&0&0\\ \vdots&\vdots&\vdots&\ddots&\vdots&\vdots&\vdots\\ 0&0&0&\cdots&1+\rho^2&-\rho&0\\ 0&0&0&\cdots&-\rho&1+\rho^2&-\rho\\ 0&0&0&\cdots&0&-\rho&1\end{pmatrix}$$

is the inverse of the correlation matrix. For the auto-regressive correlation assumption, analytic form of its inverse and determinant exists, which significantly simplifies the computation of the likelihood.

Our goal is to accurately estimate the mean vector $\boldsymbol{\mu}$ based on the observed sample(s), treating other parameters as nuisance. For this purpose, we adopt Bayesian method. Under the basic framework, different parameterizations of $\boldsymbol{\mu}$ can be used when setting up the priors, which leads to different posterior sampling schemes. We introduce the following two approaches in detail.

Probe-based approach

In the probe-based approach, we assume a simple discrete prior distribution for each component of $\boldsymbol{\mu}$, with probability mass at the theoretical levels, i.e. the prior of the *i*-th component μ_i, $1\le i\le L$ is

$$\pi(\mu_i\,|\,\mathbf{p})=p_1\cdot 1_{\{\mu_i=-1\}}+p_2\cdot 1_{\{\mu_i=0\}}+p_3\cdot 1_{\{\mu_i=0.58\}}+p_4\cdot 1_{\{\mu_i=1\}},$$

where $1_{\{.\}}$ is an indicator function and $\boldsymbol{p}=(p_1,\ldots p_4)^T$. Letting μ_i be i.i.d *a priori* (note that this independence is only for priors), we can write the prior of $\boldsymbol{\mu}$ as

$$\pi(\boldsymbol{\mu}\,|\,\mathbf{p})=\left[p_1\cdot 1_{\{\mu_i=-1\}}+p_2\cdot 1_{\{\mu_i=0\}}+p_3\cdot 1_{\{\mu_i=0.58\}}+p_4\cdot 1_{\{\mu_i=1\}}\right]^L.$$

We assume an inverse Gamma prior for σ^2, and a Beta prior for ρ (in most situations, we consider non-negative correlation only), i.e.,

$$\pi(\sigma^2\,|\,a,b)=\frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)}\exp\left(-\frac{b}{\sigma^2}\right),\sigma^2>0,$$

$$\pi(\rho\,|\,\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\rho^{\alpha-1}(1-\rho)^{\beta-1},0\le\rho<1.$$

The σ^2 and ρ are *a priori* mutually independent and independent of $\boldsymbol{\mu}$. Based on this set up, we can write the joint posterior as

$$\pi(\boldsymbol{\mu},\sigma^2,\rho\,|\,\mathbf{x})\propto(2\pi\sigma^2)^{-\frac{L}{2}}(1-\rho^2)^{-\frac{L-1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^TW(\mathbf{x}-\boldsymbol{\mu})\right\}$$
$$\cdot(\sigma^2)^{-(a+1)}\exp\left(-\frac{b}{\sigma^2}\right)\rho^{\alpha-1}(1-\rho)^{\beta-1}$$
$$\cdot\left[p_1\cdot 1_{\{\mu_i=-1\}}+p_2\cdot 1_{\{\mu_i=0\}}+p_3\cdot 1_{\{\mu_i=0.58\}}+p_4\cdot 1_{\{\mu_i=1\}}\right]^L.$$

We further factorize the joint posterior as $\pi(\boldsymbol{\mu},\sigma^2,\rho\,|\,\mathbf{x})=\pi(\boldsymbol{\mu},\rho\,|\,\mathbf{x})\cdot\pi(\sigma^2\,|\,\boldsymbol{\mu},\rho,\mathbf{x})$. It is easy to see that the conditional posterior of σ² is again an inverse Gamma distribution:

$\pi(\sigma^2\,|\,\boldsymbol{\mu},\rho,\mathbf{x})\sim IG(\tilde{a},\tilde{b})$, where

$\tilde{a}=a+\frac{L}{2},\tilde{b}=b+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^TW(\mathbf{x}-\boldsymbol{\mu})$. Integrating out σ² analytically from the joint posterior, we get the marginalized joint posterior for $(\boldsymbol{\mu},\rho)$:

$$\pi(\boldsymbol{\mu},\rho\,|\,\mathbf{x})\propto\rho^{\alpha-1}(1-\rho)^{\beta-1}(1-\rho^2)^{-\frac{L-1}{2}}\left[b+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^TW(\mathbf{x}-\boldsymbol{\mu})\right]^{-\left(a+\frac{L}{2}\right)}\qquad(2)$$
$$\cdot\left[p_1\cdot 1_{\{\mu_i=-1\}}+p_2\cdot 1_{\{\mu_i=0\}}+p_3\cdot 1_{\{\mu_i=0.58\}}+p_4\cdot 1_{\{\mu_i=1\}}\right]^L.$$

Based on the above description of the joint and marginalized posterior distributions, we design a hybrid MCMC algorithm for posterior sampling. Detailed steps are described as follows:

Step 1: Set initial values for $\boldsymbol{\mu}$, σ^2 and ρ.

Step 2: Conditioning on the current value of ρ, update μ_i sequentially $\pi(\mu_i\,|\,\boldsymbol{\mu}_{(-i)},\rho,\mathbf{x}),i=1,\cdots,L$ using Gibbs sampler, where $\mu_{(-i)}$ is from the vector of $\boldsymbol{\mu}$ with the ith component removed. Note that the conditional posterior of μ_i is discrete, taking values in {−1, 0, 0.58, 1} and the probabilities can be computed based on (2). The order of updating μ_i can be randomized.

Step 3: Conditioning on the current value of $\boldsymbol{\mu}$, update ρ from the conditional posterior $\pi(\rho|\boldsymbol{\mu},\boldsymbol{x})$ using Metropolis-Hastings update. For the proposal distribution, we adopt a random walk proposal based on the logit transform of ρ, i.e., proposing a new value $\tilde{\rho}$ from $\text{logit}(\tilde{\rho})\,|\,\text{logit}(\rho)\sim N(\text{logit}(\rho),\nu^2)$, where ν is the step-size.

Step 4: Conditioning on the current values of $\boldsymbol{\mu}$ and ρ, sample σ^2 based on conditional posterior $\pi(\sigma^2|\boldsymbol{\mu},\rho,\boldsymbol{x})$.

Repeat 2~4 until reaching the pre-specified maximum number of iterations.

Through simulations (as seen in Section 3), we find that the probe-based approach usually produces spurious local spikes in the estimation of $\boldsymbol{\mu}$, especially in single-sample analysis. This is because it assumes *a priori* independence for μ_i at each probe and does not take into account homogeneous regions formed by the probes. One way of fixing this is to use the window-based approach, i.e., dividing the target sequence into small windows with fixed width, and updating $\boldsymbol{\mu}$ window by window instead of probe by probe. Depending on whether the probes within each window are constrained to have the same mean log intensity ratio *a priori*, there are two different cases. If no *a priori* constraint is added, we simply update μ_i's window-wisely rather than probe-wisely, which is equivalent to the probe-based approach but with higher efficiency in MCMC. If such a constraint "$\mu_i=\mu_j$, for *i*, *j* in the same window" is added *a priori*, then in the multivariate normal model, $\boldsymbol{\mu}$ is parameterized as $\boldsymbol{\mu}=(w_1,\ldots,w_m)^T$ where w_j is a vector of repeated μ_j's and m = ⌈*L*/window size⌉. This is equivalent to assuming

a fixed-window segmentation in the prior, or adding an additional dimension reduction step in modeling. The reduced dimension makes the MCMC more efficient, and removes some of the small spikes caused by singleton noise in the estimation. However, it also shrinks the support of the prior distribution because only those $\boldsymbol{\mu}$'s that satisfy such fixed-window segmentation constraint are allowed. Therefore it is crucial to choose appropriate window size, otherwise the MCMC will not converge to the true $\boldsymbol{\mu}$ and will result in wrong estimates. From the above discussion, we see that the window-based approach has limited improvements, and it has difficulty to determine the correct window size. We therefore propose a new segment-based approach in Section 2.3, which is shown to have better estimation accuracy and higher computational efficiency.

Segment-based approach

The probe-based approach assumes that the μ_i's are *a priori* independent, which does not reflect regional features of CNVs. Therefore the posterior estimates can be sensitive to local aberrations (spurious spikes) unless there are multiple samples to borrow strength from. We now propose a segment-based approach which does take into account the regional features in the prior. The segment-based approach is based on a different parameterization of $\boldsymbol{\mu}$. We assume that the L probes form n segments (homogeneous regions). Here n is a hyper-parameter in our Bayesian scheme. Note that some expressions such as priors and posteriors in this subsection are conditional on n, but, for convenience this condition may be omitted from notation in the following context. We assume that probes in each region share the same mean log intensity ratio, and two adjacent regions have different means. Denote the length of segment j by z_j, with mean log intensity ratio $m_j, j = 1,\ldots,n$. We have $z_j \in \mathbb{N}$, $z_j \le L - n + 1$, $\sum_{j=1}^{n} z_j = L, m_j \in \{-1, 0, 0.58, 1\}$ and $m_j \ne m_{j+1}$, $1 \le j < n$. We write

$$\begin{aligned}\boldsymbol{\mu} &= (\underbrace{m_1,\cdots,m_1}_{z_1},\underbrace{m_2,\cdots,m_2}_{z_2},\cdots,\underbrace{m_n,\cdots,m_n}_{z_n})^T \\ &= (m_1\mathbf{1}_{z_1}^T, m_2\mathbf{1}_{z_2}^T,\cdots,m_n\mathbf{1}_{z_n}^T)^T \\ &= \mathbf{m}\otimes\mathbf{1}_{\mathbf{z}},\end{aligned}$$

where $\boldsymbol{m} = (m_1, m_2, \ldots, m_n)^T$, $\mathbf{1}_{\mathbf{z}}$ is a vector consisting of sub-vectors of 1's with length $z_j, j = 1,\ldots,n$, and $\otimes$ is the operator of vector expansion. With this parameterization, we rewrite the likelihood as follows:

$$f(\mathbf{x}\mid\mathbf{z},\mathbf{m},\sigma^2,\rho) = (2\pi\sigma^2)^{-\frac{L}{2}}(1-\rho^2)^{-\frac{L-1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{m}\otimes\mathbf{1}_{\mathbf{z}})^T W(\mathbf{x}-\mathbf{m}\otimes\mathbf{1}_{\mathbf{z}})\right\}.$$

For simplicity, we assume a discrete uniform prior for z. It is clear that to separate L probes into n segments subject to a condition that the length of each segment is a positive integer and all sum to L, there are $\binom{L-1}{n-1}$ possible configurations. The discrete uniform prior means

$$\pi(\mathbf{z}) = \frac{1}{\binom{L-1}{n-1}} = \frac{(n-1)!(L-n)!}{(L-1)!}.$$

We assume that each mj follows a discrete prior distribution supported at the four theoretical levels, i.e.,

$$\pi(m_j\mid\mathbf{p}) = p_1\cdot 1_{\{m_j=-1\}} + p_2\cdot 1_{\{m_j=0\}} + p_3\cdot 1_{\{m_j=0.58\}} + p_4\cdot 1_{\{m_j=1\}}$$

Note that the prior of z plays an important role in a MCMC sampler that involves dimension change. Improperly selected prior may increase computational complexity, or result in slow mixing of the Markov chain. For example, if we use a rescaled multinomial prior, i.e., $\pi(\mathbf{z}\mid\mathbf{q}) = \frac{L!}{d\cdot z_1!\cdots z_n!}q_1^{z_1}\cdots q_n^{z_n}$, the computation will be difficult because the rescaling constant d needs to be computed for each configuration. Another example is the "length-proportional" prior: $\pi(z_i = x) = \frac{x}{L}$. Although it seems reasonable, this prior causes difficulty when calculating π(z), because the z_i's are not independent a priori. In practice, we found the discrete uniform prior easy to use and have good performance.

The m_j's are not a priori independent due to the constraint that m_j's on the neighboring segments can not be equal. Therefore we cannot simply write the prior of m as the product of the priors of its components. Nevertheless, Markov property induced by the constraint $m_j \ne m_{j+1}$, $1 \le j < n$ gives

$$\pi(\mathbf{m}\mid\mathbf{p}) = \pi(m_1\mid\mathbf{p})\pi(m_2\mid m_1,\mathbf{p})\cdots\pi(m_n\mid m_{n-1},\mathbf{p}).$$

Assuming independence of z and m, the prior of μ becomes

$$\pi(\boldsymbol{\mu}\mid\mathbf{p}) = \pi(\mathbf{z})\cdot\pi(\mathbf{m}\mid\mathbf{p}).$$

Other priors are set to be the same as in the probe-based approach. We may thereby write the joint posterior as

$$\begin{aligned}\pi(\mathbf{z},\mathbf{m},\sigma^2,\rho\mid\mathbf{x}) \propto{}& (2\pi\sigma^2)^{-\frac{L}{2}}(1-\rho^2)^{-\frac{L}{2}}\exp\left\{-\frac{\quad}{\quad}(\mathbf{x}-\mathbf{m}\otimes\mathbf{1}_{\mathbf{z}})\quad(\mathbf{x}-\mathbf{m}\otimes\mathbf{1}_{\mathbf{z}})\right\} \\ &\cdot(\sigma^2)^{-(\ +1)}\exp\left(-\frac{\quad}{\quad}\right)\cdot\rho^{\alpha-1}(1-\rho)^{\beta-1} \\ &\cdot\frac{(n-1)!(L-n)!}{(\quad 1)!}\cdot\pi(m_1\mid\mathbf{p})\pi(m_2\mid m_1,\mathbf{p})\quad\pi(m_n\mid m_{n\ 1},\mathbf{p}).\end{aligned}$$

Similarly as in the probe-based method, we can integrate out σ^2 and obtain the marginalized joint posterior:

$$\begin{aligned}\pi(\mathbf{z},\mathbf{m},\rho\mid\mathbf{x}) \propto{}& \rho^{\alpha-1}(1-\rho)^{\beta-1}(1-\rho^2)^{-\frac{L-1}{2}}\left[b+\frac{1}{2}(\mathbf{x}-\mathbf{m}\otimes\mathbf{1}_{\mathbf{z}})^T W(\mathbf{x}-\mathbf{m}\otimes\mathbf{1}_{\mathbf{z}})\right]^{-\left(a+\frac{L}{2}\right)} \\ &\cdot\frac{(n-1)!(L-n)!}{(L-1)!}\pi(m_1\mid\mathbf{p})\pi(m_2\mid m_1,\mathbf{p})\cdots\pi(m_n\mid m_{n-1},\mathbf{p}).\end{aligned}$$

In the segment-based approach, since the number of segments n, which determines the dimension of the parameter space, is not known *a priori*, we implement reversible jump in our MCMC. The algorithm is carefully designed following the reversible jump MCMC principles [26], therefore the ergodicity is guaranteed. Details of the algorithm is listed as follows:

Step 1: Set initial values for $\boldsymbol{z}$, $\boldsymbol{m}$, ρ and σ^2.

Step 2: Conditioning on the current value of ρ, update $\boldsymbol{z}$ and $\boldsymbol{m}$ from the conditional posterior either sequentially or simultaneously, depending on the type of move. There are three types of proposals corresponding to different cases of dimension change, each happening with a certain probability. The details are described further on.

Step 3: Conditioning on the current values of $\boldsymbol{z}$ and $\boldsymbol{m}$, update ρ using Metropolis-Hastings based on the conditional posterior $\pi(\rho|\boldsymbol{z},\boldsymbol{m},\boldsymbol{x})$.

Step 4: Conditioning on the current values of $\boldsymbol{z}$, $\boldsymbol{m}$ and ρ, update σ^2 from inverse gamma distribution.

Repeat 2~4 until reaching the pre-specified maximum number of iterations.

Denote the parameter subspace when n takes value k as $\mathbb{C}_k$, $n_{\min} \le k \le n_{\max}$. In step 2, there are three types of proposals corresponding to the cases of "no change", "increase" and "decrease" in dimensionality, described as follows:

1. $\mathbb{C}_k \to \mathbb{C}_k$. This update involves no dimension change. The update of $\boldsymbol{z}$ and $\boldsymbol{m}$ can be done sequentially.

First, conditioning on current values of m and ρ, update z based on conditional posterior using Metropolis-Hastings. The candidate sample $\tilde{\mathbf{z}}$ can be proposed in different ways, e.g., (1) independent proposal from e.g. a multinomial distribution, (2) discrete random walk on hyperplane $\sum_{j=1}^{n} z_j = L$, as a simplified case we may randomly choose two segments z_i, z_j from $\boldsymbol{z}$ and let $\tilde{z}_i = \lceil u(z_i + z_j) \rceil, \tilde{z}_j = \lfloor (1-u)(z_i + z_j) \rfloor$, where u is a unif(0,1) random variable. With independent proposal, the proposal ratios are easy to compute but it can cause low acceptance rate and the chain may mix slow. In our algorithm, we use the simplified discrete random walk proposal and choose z_i, z_j to be neighboring segments. We may also consider using non-symmetric random walk proposals, such as $p(\tilde{\mathbf{z}} \mid \mathbf{z}) \sim \text{multinomial}(L, \frac{\mathbf{z}}{L})$. Using this type of proposal involves re-evaluation of the proposal ratio in each step, and therefore introduces additional burden in computation.

Second, conditioning on current values of $\boldsymbol{z}$ and ρ, update $\boldsymbol{m}$ based on conditional posterior using Metropolis-Hastings. Similarly, there are also different ways, (1) independent sequential proposal where we may update the whole $\boldsymbol{m}$ vector in one Metropolis-Hastings step by proposing a new vector $\tilde{\boldsymbol{m}}$ independent of current values of $\boldsymbol{m}$, under the constraint that $m_j \neq m_{j+1}$, for all $1 \leq j < n$. (2) only update mi, mj corresponding to the z_i, z_j updated using the simplified discrete random walk. The proposal is a discrete uniform distribution whose support should satisfy the constraint $\tilde{m}_j$ that does not equal to the neighbors m_{j-1}, m_{j+1}. In our algorithm, we use method (2) and let the proposal independent of current m. For example, if $m_{j-1} = 0$, $m_j = 1$, $m_{j+1} = -1$, then $\tilde{m}_j = 1$ or 0.58 with probability 0.5.

2. $\mathbb{C}_k \to \mathbb{C}_{k+1}$. This update is called "split".

We first generate an auxiliary random variable $u \sim \text{Beta}(a_u, b_u)$, then randomly choose a segment, say z_j, and split it into two segments: $\tilde{z}_{ja} = \lceil uz_j \rceil, \tilde{z}_{jb} = \lfloor (1-u)z_j \rfloor$. The split of m_j into $\tilde{m}_{ja}$ and $\tilde{m}_{jb}$ is subject to the constraint $m_j \neq m_{j+1}$, $1 \leq j < n$. The proposed values of $\tilde{m}_{ja}$ and $\tilde{m}_{jb}$ is independent of the splitting strategy of z_j. But the update of z_j and mj has to be performed simultaneously in one Metropolis-Hastings step. Denote the proposal ratio $R = \frac{p(m_j \mid \tilde{m}_{ja}, \tilde{m}_{jb})}{p(\tilde{m}_{ja}, \tilde{m}_{jb} \mid m_j)}$. The appropriate acceptance probability for split is obtained by min{1, A}, where

$$A = \frac{\pi(\tilde{z}_{ja}, \tilde{z}_{jb}, \tilde{m}_{ja}, \tilde{m}_{jb} \mid, \mathbf{z}_{(-j)}, \mathbf{m}_{(-j)}, \rho, x)\pi(n+1)p(\mathcal{C}_k \mid \mathcal{C}_{k+1})}{\pi(z_j, m_j \mid \mathbf{z}_{(-j)}, \mathbf{m}_{(-j)}, \rho, x)\pi(n)p(\mathcal{C}_{k+1} \mid \mathcal{C}_k) \cdot q(u)} \left| \frac{\partial(\tilde{z}_{ja}, \tilde{z}_{jb})}{\partial(z_j, u)} \right| R.$$

Remarks: (1) For the prior of hyperparameter n, we set

$$\pi(n = k) = \frac{1}{n_{max} - n_{min} + 1}, k \in [n_{min}, n_{max}].$$

(2) The probabilities of choosing merge or split are set to be equal, i.e., $p(\mathcal{C}_k \mid \mathcal{C}_{k+1}) = p(\mathcal{C}_{k+1} \mid \mathcal{C}_k) = q, 0 \leq q \leq 0.5$ for all k. Clearly, q = 0 corresponds to sampling within the current subspace and q = 0.5 corresponds to always switching. In practice we set q = 0.35. (3) We can see that $\left| \frac{\partial(\tilde{z}_{ja}, \tilde{z}_{jb})}{\partial(z_j, u)} \right| \approx z_j$. The proposal of m is again discrete uniformly distributed with support satisfying the constraint, by symmetry of the proposal distribution, R = 1.

3. $\mathbb{C}_k \to \mathbb{C}_{k+1}$. This update is called "merge". We randomly choose a neighboring pair of segments and merge them into one segment. The acceptance probability for merge is min{1, B} where B is simply the inverse of A. Note in this step, we shall set $u = \frac{\tilde{z}_{ja}}{z_j}$.

Simulation Study

An easy single-sample simulation example

Simulation of array CGH profile data is done by superimposing a sequence of Gaussian noise to a pre-specified piecewise constant log intensity ratio signal. We first consider an easy case with high SNR. We set the number of probes $L = 100$. The piecewise constant log intensity ratio signal consists of 3 aberrations with amplitude 1, 0.58 and −1, each with width 25 (wide), 15 (medium) and 5 (narrow) respectively. The Gaussian noise is multivariate normal with mean 0, standard deviation 0.18, correlation 0.3. Panel A of Figure 2 shows an example of the simulated data. The autocorrelation plot of the Gaussian noise is shown in panel B of Figure 2. In our Bayesian analysis, we set prior parameters $a = 2.1$, $b = 0.02$, $\alpha = 1.5$ and $\beta = 1.5$, other parameters $v = 0.1$ and $q = 0.35$. We set initial value $\rho^{(0)} = 0.5$, and set the first 50 probes of $\boldsymbol{\mu}^{(0)}$ with log intensity value −1 and the second 50 probes with log intensity value 0. The length of the Markov chain is 4×10^4. After discarding a burn-in period of length 3×10^4, we obtain estimates for

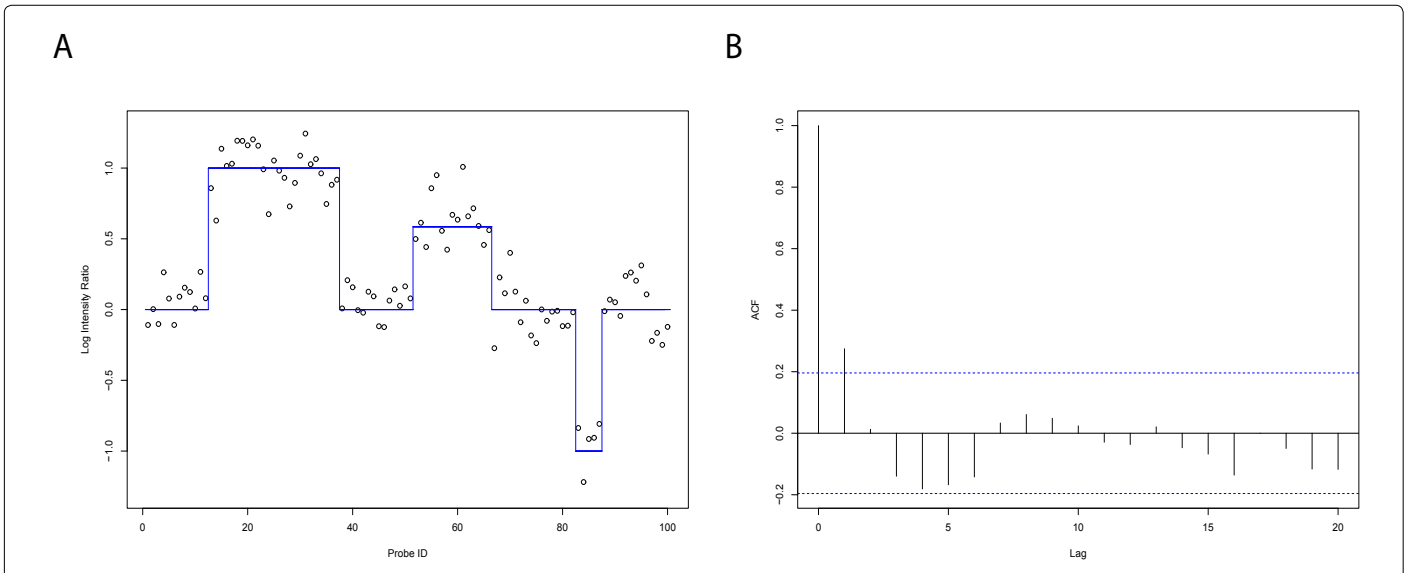

Figure 2: A: An easy example of simulated data. Black circles: observed log intensity ratios x; Blue solid line: true μ. B: Autocorrelation plot of x-μ.

$\boldsymbol{\mu}$, ρ and σ^2 from their posterior samples. Figure 3 shows the estimated $\boldsymbol{\mu}$ by using the probe-based and segment-based approaches. We see clearly that under this setting, the probe-based approach produced spurious spikes and misclassified a few probes, whereas the segment-based approach obtained a perfect estimate. For better illustration, we also draw trace plots for the posterior samples of ρ and σ^2 in Figure 4. It is clearly seen that the segment-based approach achieves better estimations than the probe-based approach. In addition, panel E of Figure 4 gives the trace plot of posterior n samples in the segment-based approach, which illustrates the changing of dimensionality of the parameter space.

Comparison between probe-based and segment-based approaches under various settings

When comparing different array CGH analysis methods using simulation, one should be cautious not to simulate data in favor of certain model assumptions. Otherwise the conclusion may be misleading. Since our two approaches are based on the same multivariate normal model, it is possible for us to perform a comparison of their performance. In general, there are several criteria [27] to evaluate the goodness of estimation of the log intensity ratio $\boldsymbol{\mu}$. For instance, misclassification rate (MCR), defined as $\frac{1}{L}\cdot$Number of non zero elements in$\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}$; false positive rate (FPR), defined as misclassification rate in copy-neutral area; true positive rate (TPR), defined as 1 – (misclassification rate in aberration area); and mean squared error (MSE), defined as $\frac{1}{L}\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2^2$. In real data analysis when the true log intensity ratio is unknown, $\frac{1}{L}\|\hat{\boldsymbol{\mu}}-x\|_2^2$ can be used instead.

Table 1 lists the performance of the two Bayesian approaches under different parameter settings. By repeating the above simulation 100 times, we obtained the average MCR, FPR and TPR in small σ^2 (0.15^2), large σ^2 (0.3^2) and small ρ (0.1), large ρ (0.6) for both approaches in both single-sample and double-sample analysis. We set the length of the Markov chain to be 10^5 with a burn-in period 9×10^4. To reduce sample autocorrelations, a thinning process was done by keeping every 10th simulated draw from the posterior samples. This simulation study was done on a desktop computer with Intel Core2Duo E7500 CPU and 3G RAM. The main program to perform MCMC and estimation is coded in C and compiled into a dynamic library which can be called by R. The code is available upon request from the authors. The average computing time for simulating 10^5 posterior samples is also listed in the table. From the single-sample analysis result, we see that under all settings, the segment-based approach outperforms the probe-based approach in both estimation accuracy and computational efficiency. In particular, we notice that the segment-based approach runs much faster than the probe-based approach, though the latter uses a simpler MCMC algorithm. This is because in the segment-based approach, the dimension of the parameter space is in the order of n, and the update of $\boldsymbol{z}$ and $\boldsymbol{m}$ is through Metropolis-Hastings, which updates one segment per iteration; whereas in the probe-based approach, the dimension of the parameter space is L, and the update of $\boldsymbol{\mu}$ is through Gibbs sampler, which updates L times sequentially in each iteration, moreover, the posterior distribution for each μ_i needs to be recomputed in each iteration. The latter certainly takes much longer time since the number of probes L is large and $L >> n$. In the simulation, the estimation accuracy turns out to be sensitive to the parameter σ^2 as it controls the noise level. Large σ^2 implies a low SNR, which leads to high noise data and makes the estimation difficult. The other parameter ρ affects the estimation accuracy differently for the two approaches. For the probe-based approach, when noise level is high, many probes are misclassified, therefore the effect of correlation is dominated by the effect of low SNR and the MCR does not change much (0.6109 versus 0.6175); in the low noise case when most probes are classified correctly, as seen from the high TPR (>0.8), large ρ tends to magnify the effect of high SNR, hence it decreases the FPR (0.1115 versus 0.0402) and the MCR (0.1389 versus 0.0786). For the segment-based approach, regardless of the noise level, large ρ tends to form spurious segments which are possible to be incorrectly detected, hence it increases the FPR (0.0007 versus 0.0285 for $\sigma^2 = 0.15^2$, 0.0224 versus 0.0689 for $\sigma^2 = 0.3^2$) and the MCR (0.0207 versus 0.0458 for $\sigma^2 = 0.15^2$, 0.0602 versus 0.1357 for $\sigma^2 = 0.3^2$). We also include the double-sample analysis result in the table. Analysis of recurrent CNVs in multiple independent samples is done by simply replacing the likelihood (1) with a product of likelihoods

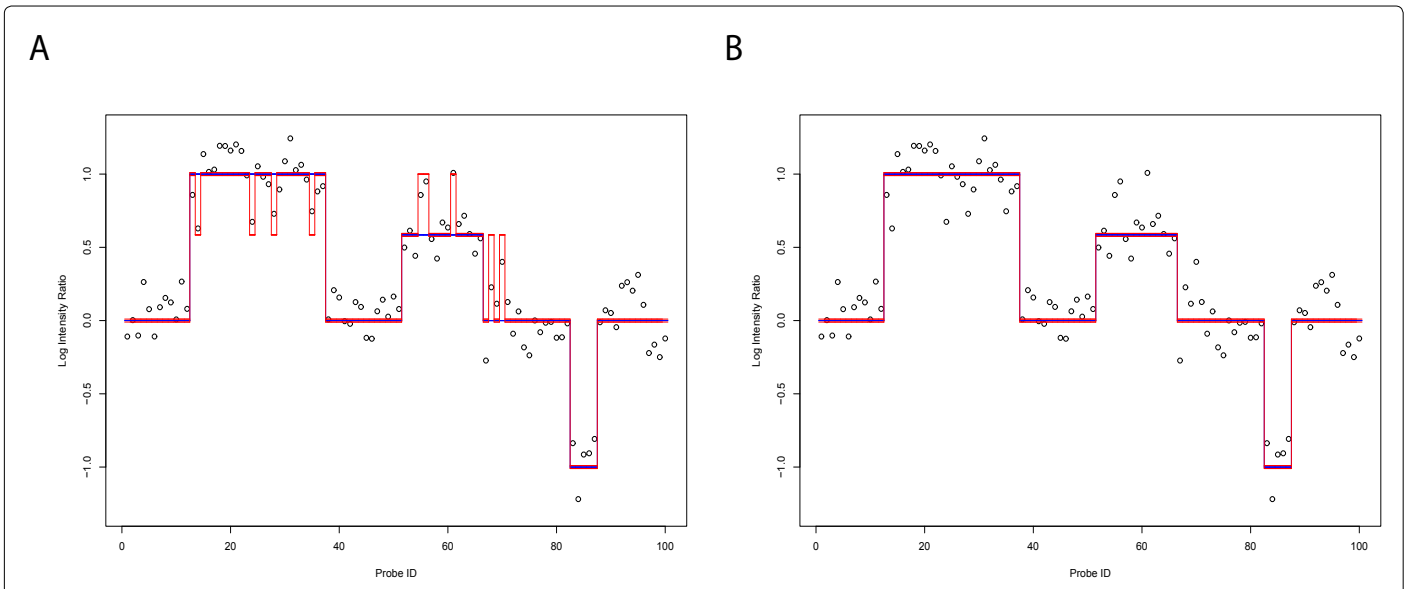

Figure 3: Estimated μ for the simulated data example. Black circles: observed log intensity ratios x; Blue solid line: true μ; Red solid line: estimated μ. A: probe-based approach; B: segment-based approach.

from each sample. From Table 1, we see that in double-sample analysis, the method gains more accurate estimates especially in the high noise case.

Comparison with CBS and HMM in single-sample analysis

To better illustrate the strength of the segment-based approach, we compare it to the commonly-used CBS and Bayesian HMM methods using a more complex (low SNR and high correlation) single-sample data from the simulation. It is of no doubt that the segment-based approach should perform better since the simulated data is in favor of our multivariate normal model assumptions, nevertheless, this comparison is useful because it shows that, under certain circumstances when CBS and HMM may not work well, the segment-based approach still achieves relatively reasonable estimates of the CNVs. Here we set $L = 100$ and 2 aberrations in the data: amplitude equals 1 with width 25 and amplitude equals 0.58 with width 20. The Gaussian noise has standard deviation 0.5 and correlation 0.4. The length of the Markov chain is 5×10^4 with a burn-in period of length 4.5×10^4. The other settings are the same as in Section 3.1. Figure 5 panel A shows the simulated data with true mean log intensity ratios, panels B, C and D gives the estimation results by CBS, Bayesian HMM and the segment-based approach, respectively. From panels B and D, we see that both CBS and the segment-based approach identify the first CNV correctly. However, due to the high noise level and strong correlation, CBS fails to detect the second CNV and produces a false positive estimation for the last three probes, whereas the segment-based approach is able to detect the starting change point of the second CNV though it also fails to detect the ending change point for the same reason. Bayesian HMM method, on the other hand, is shown to be too sensitive to the outliers under this setting, as seen from the false positive detections on state 1 (amplitude equals −1), 3 (amplitude equals 0.58) and 4 (amplitude equals 1). We see that although HMM takes into account correlation, it seems not appropriate to analyze such data where correlation only exists in the superposing Gaussian noise.

(1)For $\sigma^2 = 0.15$, $\rho = 0.1$		Average MCR	Average FPR	Average TYR	Running time (sec/10^5 iter)
probe-based	Single-sample analysis	0.1389	0.1115	0.8276	33.65
	Double-sample analysis	0.1112	0.0967	0.8711	48.88
Segment-based	Single-sample analysis	0.0207	0.0007	0.9549	0.97
	Double-sample analysis	0.0099	0	0.9780	1.20
(2)For $\sigma^2 = 0.15$, $\rho = 0.6$		**Average MCR**	**Average FPR**	**Average TYR**	**Running time (sec/10^5 iter)**
probe-based	Single-sample analysis	0.0786	0.0402	0.8744	33.61
	Double-sample analysis	0.0850	0.0771	0.9053	48.64
segment-based	Single-sample analysis	0.0458	0.0285	0.9331	0.97
	Double-sample analysis	0.0609	0.0545	0.9313	1.20
(3) For $\sigma^2 = 0.32$, $\rho = 0.1$		**Average MCR**	**Average FPR**	**Average TYR**	**Running time (sec/10^5 iter)**
probe-based	Single-sample analysis	0.6109	0.6525	0.4400	35.20
	Double-sample analysis	0.1634	0.1169	0.7798	49.10
segment-based	Single-sample analysis	0.0602	0.0224	0.8936	0.97
	Double-sample analysis	0.0595	0.0076	0.8771	1.21
(4)For $\sigma^2 = 0.32$, $\rho = 0.6$		**Average MCR**	**Average FPR**	**Average TPR**	**Running time (sec/10^5 iter)**
probe-based	Single-sample analysis	0.6175	0.6927	0.4744	34.46
	Double-sample analysis	0.2643	0.2149	0.6753	49.30
segment-based	Single-sample analysis	0.1357	0.0689	0.7827	0.97
	Double-sample analysis	0.0399	0.0116	0.9256	1.20

Table 1: Performance of two Bayesian approaches under different parameter settings.

The purpose of this comparison is not to show that the segment-based approach is superior to the other two, but to illustrate that the segment-based approach may achieve reasonable estimates when other benchmark methods do not work, especially when the data satisfy the multivariate normal assumption and show low SNR and high correlation. We also see that the performance of a method may highly depend on the property of the data. A method may perform well for some datasets but fail for others, and there is no panacea in array CGH data analysis. Therefore, it is very important for researchers to investigate the data in detail before applying suitable methods.

Performance of the segment-based approach in the case of large L and n

We have shown through simulations that the segment-based approach is effective and efficient compared to the probe-based approach. In practice the analysis of array CGH data is often done on the chromosome level, involving several hundreds to thousands of probes. Therefore we need to consider the large L case. For example, in the Pancreatic Adenocarcinoma dataset which will be discussed later on in Section 4.2, chromosome 1, the longest one, contains 1,339 probes, and the average number of probes over all chromosomes is 517. Large L then leads to possible large n. Some methods may not be able to analyze such long DNA sequences or may suffer from heavy computational burden when L and/or n are large, due to their complex underlying model assumptions or implementing algorithms. For this reason, we design a single-sample simulation study to check the feasibility of our segment-based approach when L and n are relatively large.

We set the number of probes $L = 800$ and the number of segments $n = 41$. Starting from the second segment, every other segment is set to be an aberration with amplitude randomly selected from 1, 0.58 and −1. The width of the even segments is drawn independently from discrete unif (5, 25), and for the odd segments, the width is drawn from a multinomial distribution with equal probabilities. All the other parameters are set to be the same as in Section 3.1, except that now the Markov chain is set to have 8×10^6 iterations with a burn-in period of length 7.9×10^6, and is further thinned in every 10 posterior samples. Initial values are set in the same manner, with the first half of the probes starting from log intensity value −1 and the second half starting from 0.

The simulation is repeated for 100 times. Figure 6 shows one of the simulated data and its estimated $\boldsymbol{\mu}$ by using the segment-based approach. We see that in this particular simulation, the 11th (with 9 probes) and the 19th (with 13 probes) aberration segments were misclassified. For this simulation study, the segment-based approach gives an average MCR of 0.052, an average FPR of 0.0075 and an average TPR of 0.882. The average running time for total 8×10^6 iterations is 251 seconds.

It is clear that the computing time scale of the segment-based

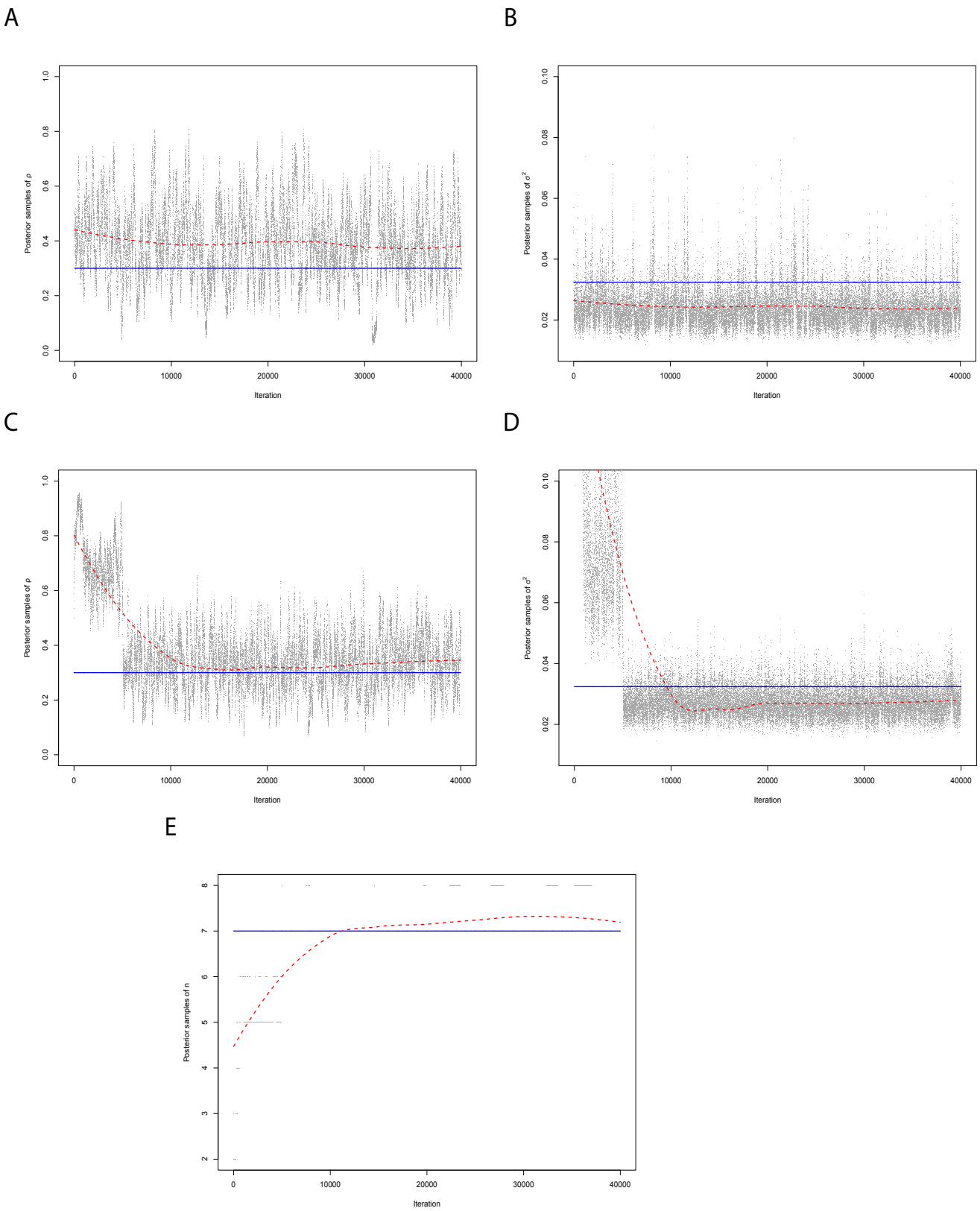

Figure 4: Posterior sample trace plots. Gray dots: posterior samples in 4 × 10^4iterations; Blue solid line: true value; Red dashed line: loess curve of the trace.A: posterior ρ samples obtained by probe-based approach; B: posterior σ^2samplesobtained by probe-based approach; C: posterior ρ samples obtained by segment-based approach; D: posterior σ^2samples obtained by segment-based approach; E:posterior n samples obtained by segment-based approach.

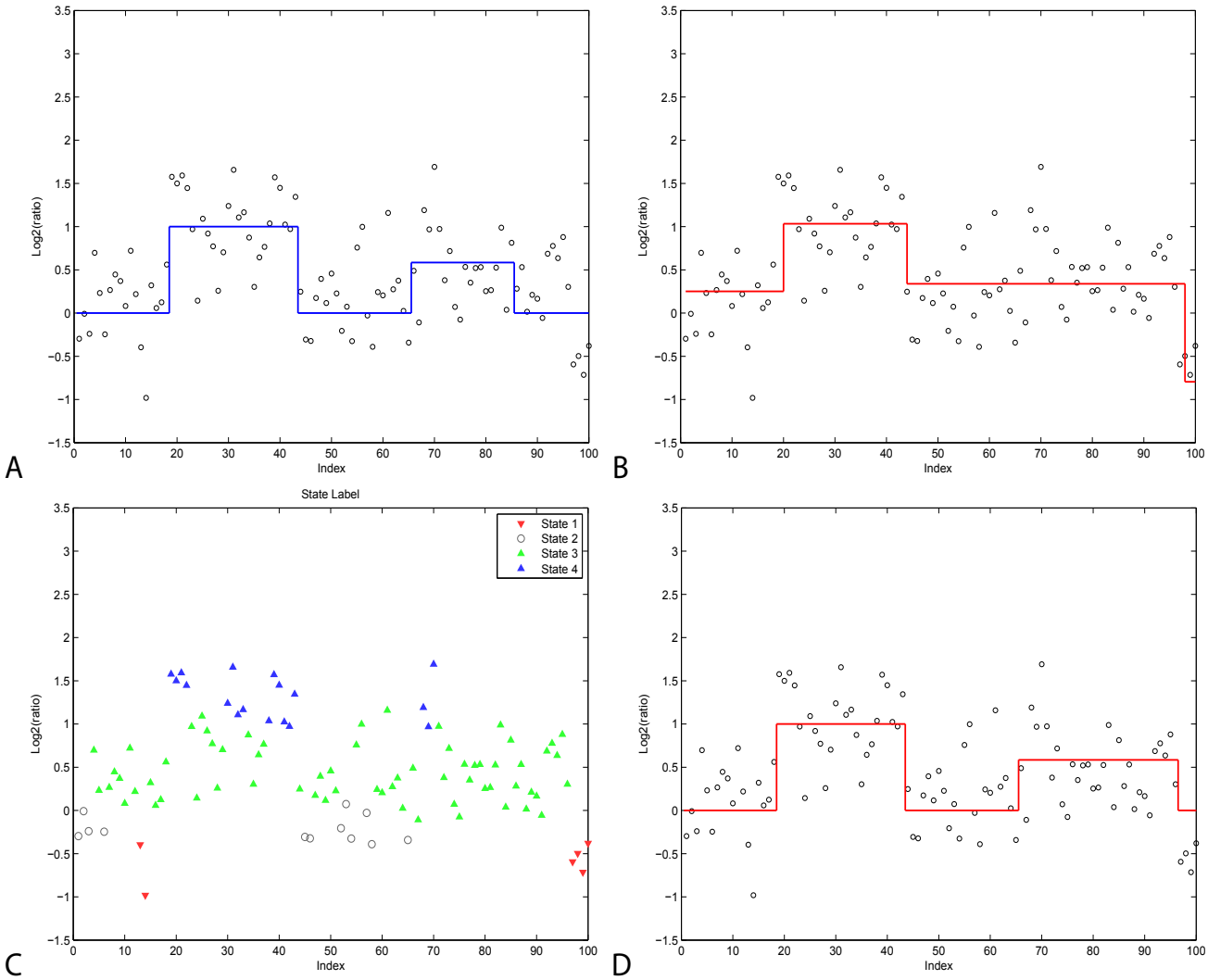

Figure 5: Comparison with CBS and HMM using simulated data with low SNR and high correlation. A: simulated data with true mean log intensity ratios; B: estimation result by CBS; C: estimation result by Bayesian HMM; D: estimation result by segment-based approach.

approach depends highly on the number of segments n. For longer sequence of probes with many segments, say, $n > 200$, we suggest to divide the whole sequence into several pieces and analyze each one separately. By doing so we could achieve better mixing for relatively short sequences and save the total computing time.

Real Data Analysis

Corriel cell lines

We apply the segment-based approach to the Corriel cell lines dataset [1]. Genomic alterations in this dataset were previously characterized by cytogenetics, as shown in Table 1 on the above website, therefore it can be used as a "gold standard" to evaluate array CGH data analyzing methods. The data have been normalized to the genome-wide median log intensity ratio. As an example, we show the estimation result of profile GM05296 chromosome 10 in Figure 7, panel A. We see that the segment-based approach detects a trisomy region from probe 54 to 94, corresponding to the 10q21–10q24 region on chromosome 10, which matches the karyotypes presented in Table 1. Figure 7, panel B gives another example of profile GM13330 from chromosome 1 to chromosome 5, at region 1q32–5q34, which involves more than one CNVs. We see that the segment-based approach detects two CNVs, one trisomy region from probe 38 to 84, corresponding to the 1q25–1$qter$ region on chromosome 1, the other from probe 385 to 401, corresponding to the 4q35–4$qter$ region on chromosome 4. Both match the karyotypes.

Pancreatic adenocarcinoma data

We also implement the segment-based approach to the Pancreatic Adenocarcinoma dataset, which contains data from Aguirre et al. [28] and is available in MATLAB Bioinformatics toolbox. This dataset includes array CGH profiles of 24 Pancreatic Adenocarcinoma cell lines and 13 primary tumor specimens. As an example, we apply the segment-based approach to analyze sample 10, chromosome 19. The estimation result is shown in Figure 8. We see that the segment-based approach detects three CNVs, in which the first one seems more like a single-probe amplification with log intensity ratio 2. In panel A, we use the usual possible copy number states {−1, 0, 0.58, 1}, so we can

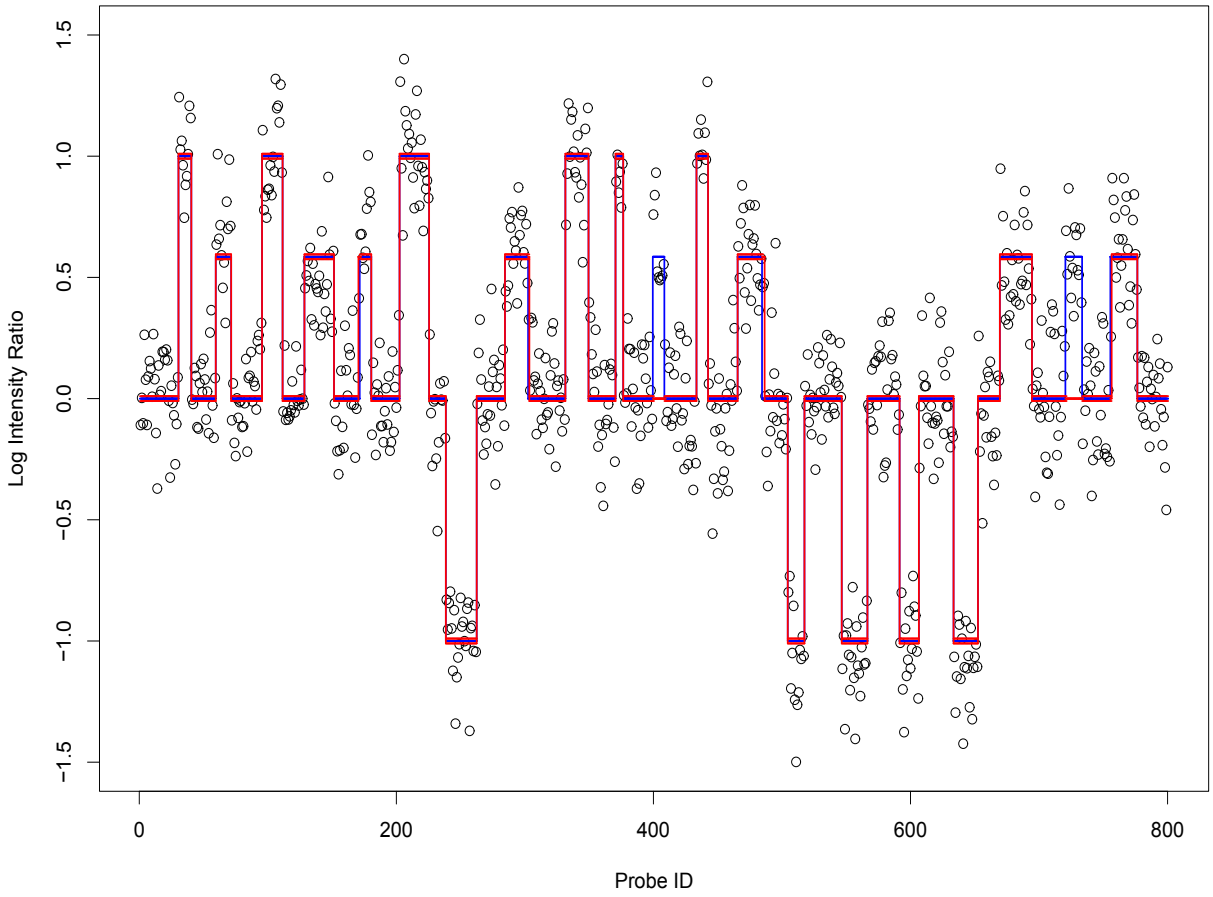

Figure 6: A simulated example of large L and n, and the estimate by segment-based approach. Black circles: observed log intensity ratios x; Blue solid line: true μ; Red solid line: estimated μ.

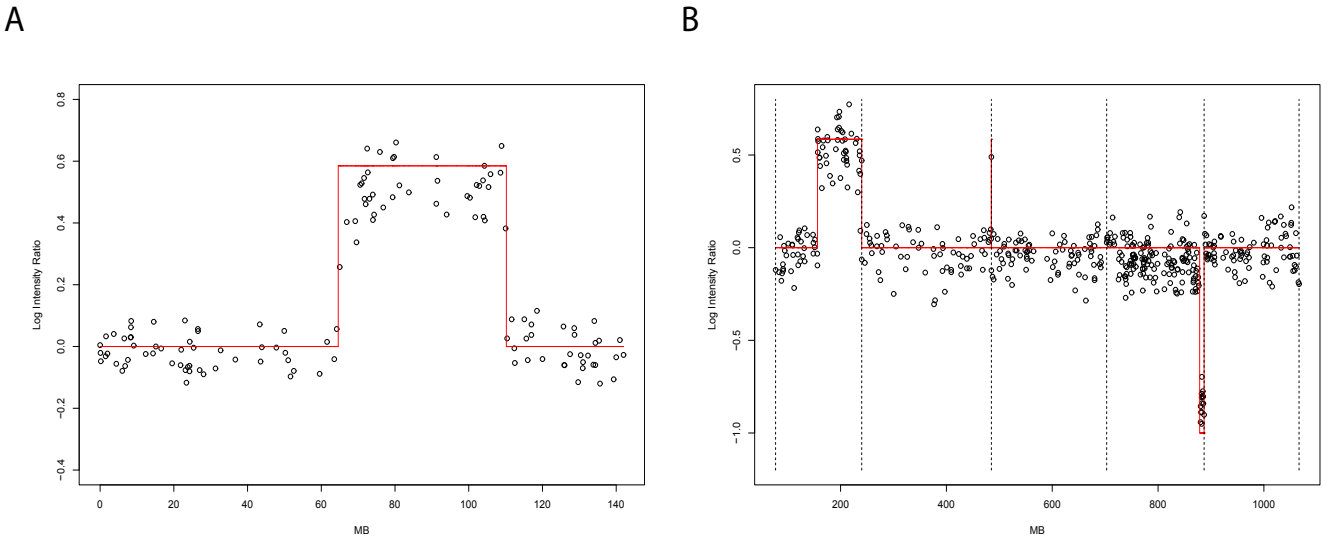

Figure 7: Application to Corriel cell lines data. Black circles: observed log intensity ratios (normalized); Red solid lines: estimated mean log intensity ratios. The x axis displays the distance from the p telomere. A: profile GM05296 chromosome 10; B: profile GM13330 1q32-5q34, vertical black dashed lines indicate borders between chromosomes.

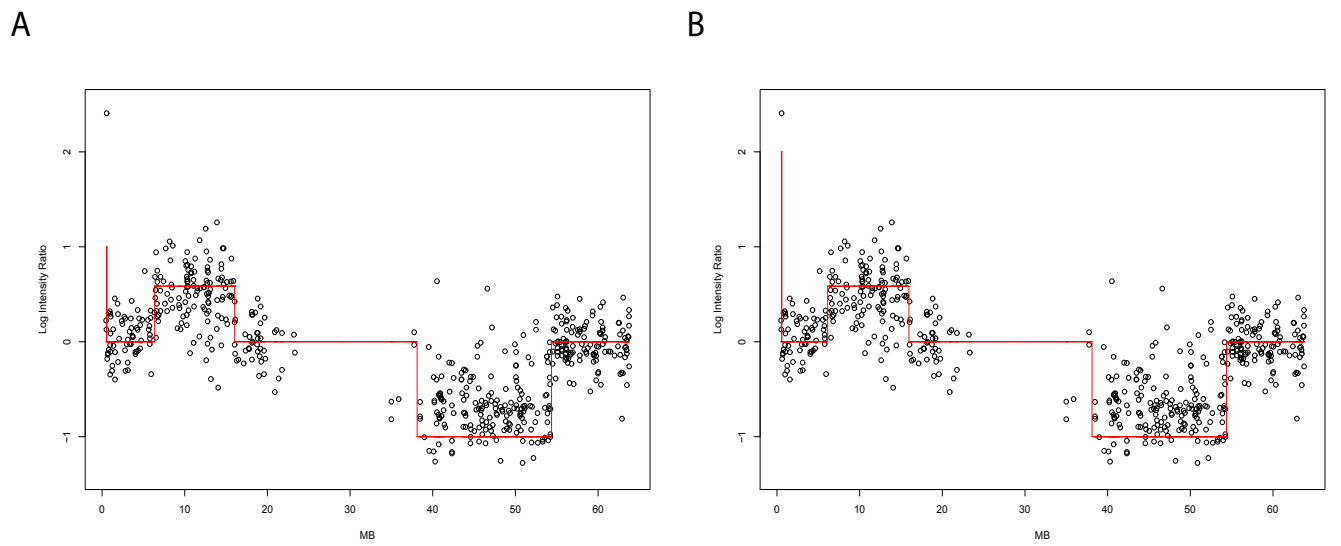

Figure 8: Application to Pancreatic Adenocarcinoma data, sample 10, chromosome 19. Black circles: observed log intensity ratios (normalized); Red solid lines: estimated mean log intensity ratios. The x axis displays the distance from the p telomere. A: use possible copy number states {-1, 0, 0.58, 1}, B: use possible copy number states {-1, 0, 0.58, 1, 2}.

only see that the first CNV hits log intensity ratio 1. After adjusting the possible copy number states to include copy number state 2, we obtain a reasonable result, as shown in panel B.

Discussion

The estimation of the mean log intensity ratios **μ** involves two problems: the location of homogeneous regions and the copy number states of these regions. Some methods, such as CBS [4], concentrate on detecting the boundaries of the regions (change points). On the other hand, other methods such as the HMM methods [11,12] and the Bayesian probe-based approach, estimate the intensity ratio of each probe first, then group the probes with the same ratios to form homogeneous regions. Perhaps a more effective way is to integrate the region homogeneity constraint into estimation of the intensity ratios, i.e., to solve the two estimation problems simultaneously. The Bayesian segment-based approach is proposed for this reason, and is shown through simulations to have better estimation accuracy and higher computational efficiency. In summary, our Bayesian segment-based approach and the multivariate normal model have the following advantages:

1. Efficiency in computation. From the likelihood expression (1), we see that for certain covariance structures, the computation of the likelihood does not involve matrix operations (note the special form of matrix W) hence the MCMC sampling is fast. In particular, the segment-based approach is much faster than the probe-based approach although the latter is based on a more restricted prior assumption.

2. Flexibility in modeling various covariance structures of the data. The current first-order autoregressive covariance structure assumes homogeneous correlation and marginal variance. It can be easily adjusted to other parametric forms, such as blockwise homogeneous correlation (i.e., independence between regions), blockwise homogeneous marginal variance or even more general structures. For example, we might specify $\rho(u) = exp(-\alpha u^c)$ where u is the unit (probe order) difference, α and c are tuning parameters.

3. Gain in efficiency when analyzing recurrent CNVs in multiple samples. Analysis of recurrent CNVs in multiple independent samples is done by replacing the likelihood (1) with a product of likelihoods from each sample, while the rest derivations remain the same. By borrowing strength from independent samples, this model is shown to be more powerful and robust.

4. The proposed method can also be integrated with some other array CGH analysis methods to improve performance. For example, the output of change point detection, smoothing or clustering methods provides rough information about region boundaries and may be used as the initial values in our MCMC algorithm.

We also notice that most methods analyze array CGH data using the order of the probes, not using their actual position in the genome. When the spatial dependence is determined by the position instead of the order of the probes, the covariance structure of the log intensity ratio turns out to be totally different so that it is difficult for many existing methods, such as the commonly used HMM method, to make the corresponding adjustment. Our proposed method may handle this issue with only a slight modification to the covariance matrix unit from the probe order to the probe position in the genome.

Another important issue in real data analysis is that the mean log intensity ratios do not exactly follow the theoretical levels. We can modify our segment-based approach to handle arbitrary copy number states. One way is to determine the possible states through a pre-processing scan or using other change-point detection methods. When the total number of theoretical levels is known, an alternative way is to treat the unknown theoretical levels as hyper parameters, set appropriate prior distributions and then proceed as proposed before. The additional hierarchy of hyper parameters also provides estimation for the theoretical levels. The alternative adjustment is shown to be effective through simulations (results not shown).

Finally, although the segment-based approach is motivated by the analysis of array CGH data, it is a general framework that can be used to solve many similar problems, especially when the data show

a "sequential clustering" property, i.e., different regions of a spatial sequence can be clustered into several categories based on certain rules. Such type of data exist widely in many real problems, especially in genetics and bioinformatics area. A good example is the problem of haplotype block identification and haplotype phasing. We expect that the proposed method will help researchers to analyze such data effectively and efficiently as well.

Acknowledgements

This work was supported by Virginia Tech's Open Access Subvention Fund.

References

1. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, et al. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. Nature Genetics 29: 263-264.
2. Pinkel D, Segraves R, Sudar S, Clark S, Poole I, et al. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to DNA microarrays. Nature Genetics 20: 207-211.
3. Hodgson G, Hager JH, Volik S, Hariono S, Wernick M, et al. (2001) Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. Nature Genetics 29: 459-464.
4. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557-572.
5. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657-663.
6. Sen A, Srivastava MS (1975) On tests for detecting a change in mean. Annals of Statistics 3: 98-108.
7. Wang P, Kim Y, Pollack J, Narasimhan B, Tibshirani R (2004) A method for calling gains and losses in array CGH data. Biostatistics 6: 45-58.
8. Picard F, Robin SE, Lebarbier E, Daudin JJ (2007) A segmentation/ clustering model for the analysis of array CGH data. Biometrics 63: 758-766.
9. Eilers PHC, de Menezes RX (2005) Quantile smoothing of array CGH data. Bioinformatics 21: 1146-1153.
10. Hsu L, Self SG, Grove D, Randolph T, Wang K, et al. (2005) Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics 6: 211-226.
11. Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN (2004) Hidden markov models approach to the analysis of array CGH data. Journal of Multivariate Analysis 90: 132-153.
12. Guha S, Li Y, Neuberg D (2008) Bayesian hidden markov modeling of array CGH data. Journal of the American Statistical Association 103: 485-497.
13. Carlin B, Gelfand A, Smith AFM (1992) Hierarchical bayesian analysis of change-point problems. Applied Statistics 41: 389-405.
14. Inclan C (1993) Detection of multiple changes of variance using posterior odds. Journal of Business & Economic Statistics 11: 289-300.
15. Stephens DA (1994) Bayesian retrospective multiple-changepoint identication. Applied Statistics 43: 159-178.
16. Chib S (1998) Estimation and comparison of multiple change-point models. Journal of Econometrics 86: 221-241.
17. Barry D, Hartigan J (1993) A bayesian analysis for change point problems. Journal of the American Statistical Association 88: 309-319.
18. Hutter M (2007) Exact Bayesian regression of piecewise constant functions. Bayesian Analysis 2: 635-664.
19. Rancoita PM, Hutter M, Bertoni F, Kwee I (2009) Bayesian DNA copy number analysis. BMC Bioinformatics 10: 10.
20. Broet P, Richardson S (2006) Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. Bioinformatics 22: 911-918.
21. Erdman C, Emerson JW (2008) A fast bayesian change point analysis for the segmentation of microarray data. Bioinformatics 24: 2143-2148.
22. Pique-Regi R, Monso-Varona J, Ortega A, Seeger CR, Triche JT, et al. (2008) Sparse representation and bayesian detection of genome copy number alterations from microarray data. Bioinformatics 24: 309-318.
23. Lai LT, Xing H, Zhang N (2008) Stochastic segmentation models for array-based comparative genomic hybridization data analysis. Biostatistics 9: 209-307.
24. Tai YC, Mark NK, Witte JS (2010) Segmentation and estimation for SNP microarrays: a Bayesian multiple change-point approach. Biometrics 66: 675-683.
25. Baladandayuthapani V, Ji Y, Talluri R, Nieto-Barajas LE, Morris JS (2010) Bayesian random segmentation models to identify shared copy number aberrations for array CGH data. Journal of the American Statistical Association 492: 1358-1375.
26. Green PJ (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82: 711-732.
27. Lai WR, Johnson MD, Raju K, Park P (2005) Comparative analysis of algorithms for identifying ampli cations and deletions in array CGH data. Bioinformatics 21: 3763-3770.
28. Aguirre AJ, Brennan C, Bailey G, Sinha R, Feng B, et al. (2004) High resolution characterization of the pancreatic adenocarcinoma genome. Proceedings of the National Academy of Sciences 24: 9067-9072.