NIH Public Access
Author Manuscript
IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2014 September 04.

Published in final edited form as:
IEEE/ACM Trans Comput Biol Bioinform. 2009 ; 6(4): 529–541. doi:10.1109/TCBB.2008.129.

A Statistical Change Point Model Approach for the Detection of DNA Copy Number Variations in Array CGH Data

Jie Chen and
Department of Mathematics and Statistics, University of Missouri-Kansas City, 5100 Rockhill Road, Kansas City, MO 64110. chenj@umkc.edu

Yu-Ping Wang
School of Computing and Engineering, University of Missouri-Kansas City, 5100 Rockhill Road, Kansas City, MO 64110. wangyup@umkc.edu

Abstract

Array comparative genomic hybridization (aCGH) provides a high-resolution and high-throughput technique for screening of copy number variations (CNVs) within the entire genome. This technique, compared to the conventional CGH, significantly improves the identification of chromosomal abnormalities. However, due to the random noise inherited in the imaging and hybridization process, identifying statistically significant DNA copy number changes in aCGH data is challenging. We propose a novel approach that uses the mean and variance change point model (MVCM) to detect CNVs or breakpoints in aCGH data sets. We derive an approximate p-value for the test statistic and also give the estimate of the locus of the DNA copy number change. We carry out simulation studies to evaluate the accuracy of the estimate and the p-value formulation. These simulation results show that the approach is effective in identifying copy number changes. The approach is also tested on fibroblast cancer cell line data, breast tumor cell line data, and breast cancer cell line aCGH data sets that are publicly available. Changes that have not been identified by the circular binary segmentation (CBS) method but are biologically verified are detected by our approach on these cell lines with higher sensitivity and specificity than CBS.

Keywords

Statistical hypothesis testing; aCGH microarray data; gene expression; DNA copy numbers; CNVs

1 Introduction

Cancer development and genetic disorders often result in chromosomal DNA copy number changes. The conventional comparative genomic hybridization (CGH) technique [1] has proved to be useful in producing a map of DNA sequence copy number at chromosomal DNA locations. The modification of conventional CGH using microarrays is called array CGH (aCGH). This technology offers high resolution and is useful for genome-wide studies

© 2009 IEEE

For information on obtaining reprints of this article, please send tcbb@computer.org, and reference IEEECS Log Number TCBB-2008-02-0029.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

of copy number change [2]. It is known that some forms of cancer are caused by somatic or inherited mutations in oncogenes and tumor suppressor genes [3]. Therefore, identification of these genes will (at least partially) facilitate the development of medical diagnostic tools and treatment regimes for cancer and other genetic diseases. In aCGH experiments, differentially labeled sample and reference DNA are hybridized to DNA microarrays ([2], [4]) and the ratios of the fluorescence intensities of the test and reference samples (usually denoted by *T/R*) on biomarkers along the chromosome are obtained as aCGH profiles [5]. Since the reference sample is assumed or chosen to have no copy number changes, markers whose normalized test sample intensities are significantly higher (or lower) than the normalized reference sample intensities correspond to DNA copy number gains (or losses) in the test sample at those locations ([2], [4], [6]). Because of the imaging and hybridization noise, the detection of copy number changes in aCGH data becomes challenging. Consequently, a statistical analysis of aCGH data, usually in the form of $\log_2 T/R$, for copy number changes is needed for detecting significant copy number change positions.

Several approaches exist for the study of copy number changes using aCGH data. The copy number alterations in mouse islet carcinomas were studied using aCGH technique in [7] and the resulting aCGH data were analyzed for DNA copy number changes by a finite Gaussian mixture model with three components (a “no change” component, a “loss” component, and a “gain” component). The parameters of the mixture Gaussian model include the proportion of each component, the mean, and the variance of each component. The estimation of these model parameters was carried out by two steps: the visual estimation and the least squares estimation. However, the significance of their findings has not been accessed statistically. To study DNA copy number change in breast tumors and breast cancer cell lines, an aCGH data set was obtained [8] and analyzed by pairwise t-tests for classes of genes, correlation coefficients between DNA copy number and mRNA level, and linear regression model among the tumors. A threshold analysis method to identify DNA copy number changes based on aCGH data was also proposed [9]. However, there was a lack of statistical assessment of such thresholds. Another method called representational oligonucleotide microarray analysis (ROMA) was developed [10], for the detection of copy number aberrations in cancer and normal humans. Profiles of a primary breast cancer sample (CHTN159) and a breast cancer cell line compared with a normal male reference (SK-BR-3) were obtained using ROMA and the mean ratios of the profiles were analyzed for the detection of DNA copy number changes. A practical software solution to the analysis of DNA copy number data using the Matlab toolbox, CGH-Plotter, was provided in [11]. The CGH-Plotter is a three-stage procedure that includes filtering, k-means clustering, and dynamic programming. A method called Chromosomal Aberration Region Miner (CHARM) was proposed [5] to identify segmental aneuploidies in gene expression and aCGH data. Recently, the single nucleotide polymorphism (SNP) array analysis was also used to study copy number changes in primary human lung carcinoma specimens and cell lines [12] in which the hidden Markov model (HMM) was applied to the SNP data and copy number changes that are vital to the development of lung cancer were identified. An unsupervised HMM approach was implemented in [13] to map the clones into states, which represent the underlying copy number of the group of clones, and therefore, possible copy number changes were identified. The Adaptive Weights Smoothing (AWS) procedure was

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

introduced in [14] to estimate copy number gains and losses on simulated data and some publicly available reference aCGH data sets. This approach provided better estimations on the changes with an optimal choice of the parameters in the algorithm [14]. However, the choice of the optimal parameters was either empirical or through other extensive computational algorithms, and this feature of the AWS procedure made it impractical for application. Picard et al. [15] proposed a method for obtaining the maximum likelihood estimates of the breakpoints, the mean and variance of the segments for a given number of segments, and an adaptive criterion to estimate the number of segments in aCGH data. It was an alternative to the estimation of DNA copy number changes but there was no statistical assessment (in terms of p-value) on how good their estimation was.

Although the aforementioned studies used some statistical estimation methods in analyzing aCGH data and gene expression data for copy number changes, the validation of those methods is not completely satisfactory. Several methods mentioned above were compared in [16], and it was shown in [16] that the comparisons of these methods are difficult due to a possibly suboptimal choice of parameters in these methods. Nevertheless, such comparisons reveal general characteristics [16] that are helpful to the biological investigator who needs to analyze copy number changes. It is concluded in [16] that a statistical analysis of copy number changes should essentially address two important points: One is how to estimate the loci, where the DNA copy number has changed, and the other is how good the estimation is in terms of giving the probability of an observed significance (or simply the p-value) for the estimated locus of a change. Therefore, in our opinion, an appropriate statistical change point model ([17]) is most suitable for analyzing DNA copy number data. From statistical point of view, a change point is defined as a point (either an index or a spatial location) before which a random sequence follows a distribution with certain parameter(s)), and after which the random sequence follows another distribution (or the same distribution as before but with different parameter(s)). Statistical change point analysis can be rooted back to the 1950s [18]. A Bayesian estimator of the current mean for a priori uniform distribution on the whole real line was derived in [19] with a quadratic loss function. The exact and asymptotic distribution of the test statistic for testing a single change in the mean of a sequence of normal random variables was obtained in [20]. The null distributions in the cases of known and unknown variances of a single change in the mean were derived as well [21]. The mean and variance change point problem for normal distributions was first studied in [22]. The asymptotic null distribution of the likelihood procedure statistic [23] for the simultaneous change in the mean vector and covariance of a sequence of normal random vectors was studied in [24]. Other works related to change point(s) problem(s) can be found in the literature of statistical change point analysis [25], [26], [27], [28], [29], [30], [31], [32].

A mean change point model (MCM) was recently applied [33] to detect DNA copy number changes that were observed in the gene expression experiment on Dermatofibrosarcoma Protuberans. A circular binary segmentation (CBS) method was later proposed [6] to identify DNA copy number changes in aCGH data based on the MCM proposed in [20]. The CBS approach [6] is among a few methods that gives both the estimation of the locus, where the copy number change takes place, and the p-value of the significance of the finding. The p-value in [6], however, is only obtained by a permutation method and takes a long computation time when the sequence is long (which is the case for high-density array data).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Hence, the method in [6] has the slowest computational speed [15]. A recent result in [34] has improved the computational speed of CBS. A Bayesian HMM approach was proposed for the analysis of aCGH data using extensive computational techniques [35]. Most recently, a computational approach using a Bayesian segmentation modeling of aCGH data was introduced, aiming at giving good estimation and confidence intervals of the DNA aberration regions [36].

In this paper, we propose to use a different approach, the MVCM ([22], [29]), for the detection of DNA copy number changes in aCGH and gene expression data. Our proposed approach emphasizes both estimation and hypothesis testing regarding the change point(s) in the DNA copy number data. Our MVCM approach allows researchers to do both a genome-wide search [41] for copy number changes and a chromosome-wide search for copy number changes. We also compare our model with the CBS method in [6]. As pointed out in [16], there are basically two types of methods for analyzing aCGH copy number data: One is the estimation-oriented approach and the other is the statistical inference approach, which emphasizes both the estimation and the significance of the detected change. Our method and the CBS method [6] fall into the second category. Among a number of different methods, we just compared the proposed MVCM with the MCM of [6] because they are more methodologically comparable. Our results showed that by adding the variance component in the change parameter, we can detect copy number changes with fewer false positives, and we can directly use the p-value formula to approximate the statistical significance of the detected copy number changes with ease. That is, our model outperforms the CBS and gives a faster computation of the approximate p-value for the change identified without going through time-consuming permutation calculation in [6] and [34].

2 The Multiple Mean and Variance Change Point Model for aCGH Data

In the aCGH data, $\log_2 T/R = 0$ indicates no copy number change at that locus while $\log_2 T/R < 0$ (or > 0) signifies a deletion (or duplication) in the test sample at that locus. However, due to various random noise, which occurs largely during the experimental and image processing stages, the $\log_2 T/R$ becomes a random variable. Ideally, this random variable is assumed to follow a Gaussian distribution of mean 0 and constant variance σ^2. Then, deviations from the constant parameters (mean and variance) presented in $\log_2 T/R$ data may indicate a copy number change. Hence, the key to identifying true DNA copy number changes (breakpoints) becomes the problem of how to identify multiple changes (breakpoints) in the parameters of a normal distribution based on the observed sequence of $\log_2 T/R$.

Let X_i denote the normalized $\log_2 T_i/R_i$ at the ith locus along the chromosome, then $\{X_i\}$ is considered to be a sequence of normal random variables taken from $N\left(\mu_i, \sigma_i^2\right)$, respectively, for $i = 1; \ldots; n$ ([6], [7]). Typically, the DNA copy number changes were analyzed using the MCM ([6], [33]) with a fixed variance in the distributions of the sequence $\{X_i\}$. Since the aCGH technology may not guarantee the aCGH data to have a constant variance [7] due to some uncontrollable errors during hybridization and because of the mean and variance change observed in the aCGH data of SK-BR-3 (see Fig. 6 and [10]),

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

we therefore propose to analyze the DNA copy number changes using the MVCM for the sequence $\{X_i\}$.

The multiple DNA copy number changes can be defined as testing the null hypothesis in the mean and variance parameters in the sequence of log intensity ratios $\{X_i\}$:

$$\begin{aligned} H_0: \quad & \mu_1 = \mu_2 = \cdots = \mu_n = \mu \quad \text{and} \\ & \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2 \end{aligned} \tag{1}$$

versus the alternative:

$$\begin{aligned} H_1: \quad & \mu_1 = \cdots = \mu_{k_1} \neq \mu_{k_1+1} = \cdots = \mu_{k_2} \neq \cdots \neq \mu_{k_{q-1}} \\ & = \cdots = \mu_{k_q} \neq \mu_{k_q+1} \cdots = \mu_n \quad \text{and} \\ & \sigma_1^2 = \cdots = \sigma_{k_1}^2 \neq \sigma_{k_1+1}^2 = \cdots = \sigma_{k_2}^2 \neq \cdots \neq \sigma_{k_{q-1}}^2 \\ & = \cdots = \sigma_{k_q}^2 \neq \sigma_{k_q+1}^2 \cdots = \sigma_n^2, \end{aligned} \tag{2}$$

where μ and σ^2 are the unknown common mean and variance under the null hypothesis (1); $1 < k_1 < k_2 < \ldots < k_q < n$; q is the unknown number of change points and k_1, k_2, … k_q are the unknown change points' positions (loci), respectively, under the alternative hypothesis (2). The null hypothesis (1) refers to the claim that there are no changes in the mean and variance parameters of the distribution from which the sequence $\{X_i\}$ is drawn, and the alternative hypothesis (2) indicates that there are several changes (q changes) in the mean and variance parameters.

2.1 A Binary Segmentation Procedure for Searching Multiple Change Points

For detecting multiple change points, an effective method ([20], [17]) is the binary segmentation procedure (BSP) proposed in [37]. It searches the first significant change point in a sequence, then breaks the original sequence into two subsequences: one before the first significant change point (including the change point) and the other after the first significant change point. Thereafter, the procedure tests the two subsequences separately for a change point. The process is repeated until no further subsequences have change points. The collection of change point locations found at the end is denoted by $\left\{\hat{k}_1, \hat{k}_2, \ldots, \hat{k}_q\right\}$ and the estimated total number of points is consequently q.

Using this BSP, we just need to focus on how to detect the single change (the most significant one) each time and repeat the searching scheme of BSP to get all the significant changes. Specifically, to identify all significant DNA copy number changes in the MVCM as specified in (1) and (2), we first turn to the searching of a single change by testing the null hypothesis (1) versus the following new alternative hypothesis:

$$\begin{aligned} H_1: \quad & \mu_1 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n \quad \text{and} \\ & \sigma_1^2 = \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_n^2, \end{aligned} \tag{3}$$

where k, $1 < k < n$, is the unknown position of the single change at each stage. The strategy is to convert the testing of multiple change points in the MVCM given by (1) and (2) into several stages of hypothesis testing of (1) versus the single change point hypothesis (3) in the MVCM. At each stage, failure to the rejection of H_0 (1) at a given significance level α

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

indicates no change in the DNA copy number sequence and then the search scheme stops at this stage. The rejection of H_0 (1) or acceptance of the alternative hypothesis (3) at a given significance level α at certain stage indicates that a significant change in the DNA copy number sequence is found and the search scheme of BSP continues to the next stage until no more significant changes are found in a later stage. The key for testing the null hypothesis (1) versus the alternative (or research) hypothesis (3) is to obtain a test statistic and its null distribution. This will be discussed in the next section.

2.2 The Schwarz Information Criterion and the Test Statistic

We use the Schwarz information criterion (SIC)-based approach [38], proposed in [29], to test (1) against (3). In general, SIC is defined as

$$SIC = -2 log L\left(\hat{\theta}\right) + k log n,$$

where $L\left(\hat{\theta}\right)$ is the maximum likelihood function of a model, k is the number of parameters to be estimated, and n is the sample size. The information criterion principle for model selection is to choose the model with minimal SIC as the best possible model. The SIC approach is a likelihood-based approach with the penalty term added for possible model overparameterization [29]. It converts hypothesis testing into a model selection process in which the null hypothesis H_0 in (1) corresponds to a model of no change in the sequence and the alternative hypothesis H_1 in (3) corresponds to models with change locations specified in each model. Specifically, due to the requirement for the existence of the maximum likelihood estimator (MLE) of the variance component, we can only detect changes located from the 2nd to the $(n - 1)$th in the sequence $\{X_i\}$. Therefore, the alternative hypothesis H_1 in (3) actually corresponds to $n - 3$ change point models with change located at $2, \ldots, n - 2$, respectively.

The *SIC* corresponding to the no change model specified by H_0, denoted by $SIC(n)$, is obtained as ([29]):

$$\begin{aligned} SIC(n) = & -2 log L_0\left(\hat{\mu}, \hat{\sigma}^2\right) + 2 log n \\ = & n log 2\pi + n log \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + n \\ & + (2 - n) log n, \end{aligned} \qquad (4)$$

where $L_0\left(\hat{\mu}, \hat{\sigma}^2\right)$ is the maximum likelihood function with respect to H_0, and $\hat{\mu}$ and $\hat{\sigma}^2$ are the MLEs of μ and σ^2 under H_0, respectively, which are found to be:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

Corresponding to H_1, there are $n - 3$ change point models, and the *SIC* for each model, denoted by $SIC(k)$ for fixed k, $2 \le k \le n - 2$, is obtained as:

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

$$\begin{aligned} SIC\left(k\right)= & -2logL_1\left(\hat{\mu}_1,\hat{\mu}_n,\hat{\sigma}_1^2,\hat{\sigma}_n^2\right)+4logn \\ = & nlog2\pi+klog\hat{\sigma}_1^2+\left(n-k\right)log\hat{\sigma}_n^2+n \\ & +4logn, \end{aligned} \tag{5}$$

where $L_1\left(\hat{\mu}_1,\hat{\mu}_n,\hat{\sigma}_1^2,\hat{\sigma}_n^2\right)$ is the maximum likelihood function under H_1, and

$$\begin{aligned} \hat{\mu}_1= & \bar{X}_k=\frac{1}{k}\sum_{i=1}^{k}X_i,\hat{\sigma}_k^2=\frac{1}{k}\sum_{i=1}^{k}\left(X_i-\bar{X}_k\right)^2, \\ \hat{\mu}_n= & \bar{X}_{n-k}=\frac{1}{n-k}\sum_{i=K+1}^{n}X_i,\hat{\sigma}_{n-k}^2=\frac{1}{n-k}\sum_{i=k+1}^{n}\left(X_i-\bar{X}_{n-k}\right)^2, \end{aligned}$$

are the MLEs for μ_1, σ_1^2, μ_n, and σ_n^2, respectively.

According to the principle of information criterion, if $SIC(n) < \min_{2\leq k-2}SIC(k)$, there is no change in the sequence. Otherwise, there is a change in the sequence and the change point position k is estimated by $\hat{k}$ such that

$$SIC\left(\hat{k}\right)=\min_{2\leq k\leq n-2}SIC\left(k\right). \tag{6}$$

Using this method, the change is easily located. The algorithm is fast in identifying change point estimates.

Due to random disturbance during the process of aCGH experiments, aCGH data do contain fluctuations that may not indicate a copy number change. To address this issue, Chen and Gupta [28] proposed to view

$$\Delta_n=\min_{2\leq k\leq n-2}\left[SIC\left(k\right)-SIC\left(n\right)\right]$$

as a statistic, and hence, used the null distribution of Δ_n to make formal inference decision regarding H_0 given by (1) versus H_1 given by (3). The exact null distribution of Δ_n still remains unknown so far. However, the asymptotic null distribution of a function of Δ_n is ([22], [29])

$$\begin{aligned} \lim_{n\to\infty} & P\left[a\left(logn\right)-\left(2logn-\Delta_n\right)^{1/2}-b\left(logn\right)\leq x\right] \\ & =exp\left(-2e^{-x}\right), \end{aligned}$$

where $a(\log n) = (2 \log \log n)^{1/2}$ and $b(\log n) = 2 \log \log n + \log \log \log n$. Using the above asymptotic null distribution, we can then obtain an approximate p-value for testing H_0 given by (1) versus H_1 given by (3). Specifically, the approximate p-value for rejecting the null hypothesis H_0 given by (1) in favor of H_1 given by (3) (hence, estimating the change location as $\hat{k}$ is obtained as

$$p-value=1-exp\left\{-2exp\left[b\left(logn\right)-a\left(logn\right)\lambda_n^{1/2}\right]\right\}, \quad (7)$$

where

$$\lambda_n=2logn-\Delta_n.$$

Equation (7) above gives the analytic formula for calculating the approximate p-value of the test of "no change" versus "one change." This p-value is the observed significance of the test based on the observed sample information. When this p-value is less than a prespecified significance level a, the null hypothesis of "no change" is rejected.

As mentioned before, in order for the MLE to exist, we only detect changes located from the 2nd to the (n − 1)th position in the sequence $\{X_i\}$. This requirement helps to avoid the edge effect of the CBS algorithm discussed in [6]. We summarize our proposed MVCM approach into five steps. As a note, our approach handles the edge effect as stated in step 5 below.

The Algorithm of the MVCM Approach:

1. Calculate the *SIC*(*n*) and *SIC*(*k*) according to (4) and (5), respectively, for the whole sequence.

2. Obtain the estimate $\hat{k}$, denoted by $\hat{k}_1$, of the locus, where the DNA copy number has possibly changed according to (6).

3. Calculate the p-value according to (7). If the p-value is less than a prespecified significance level a at $\hat{k}_1$ obtained in step 2, then $\hat{k}_1$ is the first significant change identified and go to step 4; otherwise, there is no significant copy number change in the sequence and the search stops here.

4. Break the original sequence at $\hat{k}_1$ into two subsequences: the first sequence includes the first through the $\hat{k}_1$th observations of the original sequence and the second sequence contains the rest of the observations starting from the $\left(\hat{k}_1+1\right)$ observation. Repeat steps 1-3 for each of the two subsequences until no further changes are found.

5. For the first and last observations (boundary points) in the sequence of $\{X_i\}$, we do a final check by first calculating a 99.7 percent confidence interval,

$$I_1=\left(\overline{X}_{\hat{k}}-3\hat{\sigma}^2_{\hat{k}},\overline{X}_{\hat{k}}+3\hat{\sigma}^2_{\hat{k}}\right)$$

for the first piece of the sequence, and a 99.7 percent confidence interval,

$$I_2=\left(\overline{X}_{n-\hat{k}}-3\hat{\sigma}^2_{n-\hat{k}},\overline{X}_{\hat{k}}+3\hat{\sigma}^2_{n-\hat{k}}\right)$$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

for the second piece of the sequence. If $X_1 \in I_1$, the first observation is a change; otherwise, it is not. Similarly, if $X_n \in I_2$, the last observation is a change; otherwise, it is not.

3 Simulation Studies

The approximate p-value given by (7) becomes more precise when the size of the whole sequence is large enough as it is derived from the limiting null distribution (see Section 2). To ensure that the approximate p-value obtained through the asymptotic null distribution is good enough for finite sample sizes, we performed several simulations for moderate sample size *n* (the whole sequence size) of 30, 40, 50, and 60 for the scenarios of the true change being located at the front (the 10th observation in the sequence), at the center (the $\frac{n}{2}$ th observation), and at the end (the (n-7)th observation) of the sequence. Using aCGH data on the fibroblast cell lines [39] as benchmark data sets, we observed that the segments before and after a detected change point mostly have mean difference ranging from 0.36 to 0.7 (or larger), a standard deviation difference ranging mostly from 0.05 to 0.2, and a coefficient of variation $CV\left(CV=100s/\bar{x}\right)$ ranging from 9 to 15,031. We therefore investigated the cases when the mean, the standard deviation, and the CV are within the above-mentioned ranges. Specifically, we simulated a normal sequence that is with mean 0 and standard deviation 0.05 before the change point and is of values δ_i after the change points, respectively, for i = 1, … 9, where $\delta_i = (\mu_i, \sigma_i)$ takes

$$(0.4,0.12),(0.41,0.13),(0.42,0.14),(0.43,0.15),(0.44,0.16),(0.45,0.17),(0.46,0.18),(0.47,0.19),$$

and (0.48, 0.2), for i = 1, …, 9. These choices of δ_i result in an increasing order of CV: 30, 32, 33, 35, 36, 37, 39, 40, and 42. Each simulation is carried out 1,000 times. The relative frequency f_1 that $\hat{k}$ equals the true change location and the relative frequency f_2 that the p-value is less than 0.01 at $\hat{k}$ are summarized in Table 1.

The simulations indicate that the proposed method can pick up changes in the center or at the end of the sequence with very high powers (at least 96 percent), even when the sample size is moderate; and the power increases as the sample size increases. When the change is located at the beginning of the sequence and the sample size is 30, the power of the method is low (from 47 to 70.7 percent). However, when the sample size increases, the power of detecting a change at the beginning of the sequence increases dramatically (up to 99.9 percent). As the aCGH data are usually of large sample sizes, the proposed method can confidently be applied to the identification of DNA copy number changes.

4 Applications to aCGH Data

We have used our proposed method to analyze several publicly available array CGH data sets.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4.1 Analysis of 15 Fibroblast Cell Lines

In Snijders et al. [39], aCGH experiments were performed on 15 fibroblast cell lines and the normalized averages of the $\log_2 T_i/R_i$ (based on triplicate) along positions on each chromosome were available at the fibroblast cell lines data Web site [40]. There were either one or two alterations in each of the 15 cell lines as indicated by karyotyping [6]. As pointed out before, our method allows us to do both a genome-wide search [41] for copy number changes and a chromosome-wide search for copy number changes. For the nine cell lines (GM01524, GM01535, GM01750, GM03134, GM03563, GM05296, GM07081, GM13031, and GM13330), as changes occur within a particular chromosome, we did a chromosome-wide search for copy number changes for each of the 23 chromosomes within each cell line using our method. As the number of changes is unknown, we did not control the multiple comparison error but we preset the significance level σ to a small value such as 0.01 and 0.001 (just like the ones used in [6]). Our approach identified significant changes in the log ratios for each of the cell lines; and the locations of the changes along with their corresponding p-values were listed in Table 2.

It turned out that our method can identify the changes with fewer false positives compared to the CBS method [6]. Furthermore, the changes identified by our MVCM method match with the copy number variations found through the spectral karyotyping (see [39, Table 1] for the karyotyping results). The findings of our method, the MVCM with BSP, and the MCM with CBS were given in Table 3 for comparison purpose. It should be noted that, in Table 3, “Yes” means the change was found by the specific method (MCM with CBS or MVCM with BSP) for the known alteration verified by spectral karyotyping in (Snijders et al. [39]) on the specific chromosome in the cell line at the given α level; “No” means the change was not found by a specific method, but was identified by spectral kayotyping; and “Number of false positives” gives the number of changes found by the specific method for a cell line while there were no known alterations actually found by spectral karyotyping [6].

Further, we have calculated the specificity (where specificity = number of true negatives/(number of true negatives) number of false positives)) and sensitivity (where sensitivity = number of true positives/(number of true positives) number of false negatives)) for all cases. The specificities and sensitivities were also given in Table 3. Graphical illustration of the change detected by our method for chromosome 5 of GM01535 is given in Fig. 1 and the associated SIC values for this chromosome are given in Fig. 2. The red arrow in Fig. 2 denotes the significant minimum SIC value occurring at locus 86 and that location then corresponds to the red circled observation in Fig. 1 indicating the change point. We also illustrated the changes identified on chromosome 10 of GM05296 in Fig. 3. In light of Tables 2 and 3, we can conclude that our proposed method performed better than the CBS of [6] in terms of a smaller false positive rate, a higher sensitivity, and a higher specificity.

We also did a genome-wide search for copy number change on the six cell lines: GM00143, GM02948, GM03576, GM04435, GM07408, and GM10315, whose changes are known to be on a whole chromosome arm [6, 36] by spectral karyotyping and these changes could not be detected by the CBS method provided in [6]. In this case, we can still use the proposed method to search for DNA copy number changes based on the genome data (23

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

chromosomes altogether). Our findings on these six cell lines are summarized in Table 4. Specifically, we identified changes on the whole arm of chromosome 18 of the cell line GM00143 (see Fig. 4), on the whole arm of chromosome 13 of cell line GM02948, on the whole arm of chromosome 20 of GM07408, on the whole arm of chromosome 22 of GM10315, on the whole arm of chromosome 2 of GM03576, and on the whole arm of chromosome 16 of GM04435 (see Table 4).

These discoveries are in line with the results of spectral karyotyping [39]. We missed one change on GM03576, GM04425, and GM07408, respectively. Overall, the proposed method is more effective in searching for DNA copy number changes than the compatible CBS method. The proposed method can do a chromosome-wide or genome-wide search for DNA copy number changes.

4.2 Analysis of aCGH Data of Two Breast Tumors

Snijders et al. [39] also used aCGH to detect DNA copy number gains, losses, and amplifications for two breast tumors noted as S0034 and S1514. They found low-level gains and losses on both tumors, as well as high-level amplifications in both tumors. Using the proposed method for breast tumor S0034, we identified low-level gains and losses (with *p-value* < 0.001) on chromosome 1, end of chromosome 2, on chromosomes 7 and 8, and the high-level amplification on the whole arm of chromosome 23 (the X chromosome) with a p-value of 0.0000. For breast tumor S1514, we identified low-level gains and losses (significant with *p-value* < 0.001) changes on chromosomes 3 and 4, at the end of chromosomes 5 and 13, and at the beginning of chromosome 15. We also identified the high-level amplification on chromosome 20 with p-value of 0.0000 (see Fig. 5). All of our findings on these two breast tumor cell lines are consistent with the conclusions validated with karyotyping in Snijders et al. [39]. We missed only two changes on each tumor in this analysis. These false negatives are due to the fact that the segment at which the changes are located is short and the approximate p-values are not significant enough.

4.3 Analysis of aCGH Data of a Breast Cancer Cell Line

Lucito et al. [10] developed the ROMA method for the detection of DNA copy number changes in cancer and normal humans. They assembled three ROMA data sets, namely CHTN159, SK-BR-3, and Pygmy, at 10,000 and 85,000 resolution. We analyzed the ROMA data (see Web site [42]), SK-BR-3, which is a breast cancer cell line compared with a normal reference. We identified changes on the highly turbulent chromosome 8 [9], and changes on the somewhat less active chromosomes 5, 17, and X as specified in [9]. Changes on the other 19 chromosomes of this breast cancer cell line were also identified (see Fig. 6 for chromosome X of SK-BR-3). All changes on these chromosomes of SK-BR-3 were validated using a “representation”-based experiment in [10]. Moreover, the gains and losses of DNA copy numbers on several chromosomes of SK-BR-3 (for example, see Fig. 6) indicate that the mean and variance of the log ratios have been changed. This observation further justifies the appropriateness of using MVCM.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

5 Discussion and Conclusion

In this paper, we propose to use the MVCM approach to study the DNA copy number changes in aCGH data sets. The approximate p-value of identifying a change using SIC method is given and the procedure to detect all changes in the data is carried out using BSP.

The advantage of using the MVCM model is that it leads to fewer change points than that of MCM as MCM tends to divide large segments into smaller pieces so that the homogenous variance assumption for all segments can be met ([15]). Therefore, the MVCM model has the potential to give fewer false positives than MCM. Adding the variance component in the change point analysis will improve the estimation of the change point location even if just the mean shifts greatly. This is because in the MVCM model, the variances under the alternative hypothesis are estimated for each subsequence without pooling all subsequences (with possible different means) together, while in MCM, the homogeneous variance under the alternative hypothesis is estimated by pooling all subsequences having different means together. Depending on the test statistic used under each model, the false negative rate needs to be assessed for each model when one decides which model to use for the aCGH data at hand (see the simulation results in Section 3). Using either MVCM or MCM also depends on the biological experiment in which the scientists may have prior knowledge on whether there are potential variance changes. In that case, the MVCM model is proposed as an alternative to MCM when possible variance changes exist in the sequence.

Simulation studies and applications of the proposed method on aCGH data sets of several cell lines indicate that the proposed method has very high power (or low false negative) in identifying DNA copy number changes. For a change in a shorter sequence (small n) of a chromosome, the SIC value can indicate the position of the change although the corresponding p-value may not be significant. This is due to the fact that the sample size of that segment is small and the p-value is calculated based on large sample property. This is the only limitation of our method. However, this limitation can be overcome by finding the p-value using simulation on a permutation method, which will be explored in the future. Furthermore, the confidence intervals given at the end of Section 2 can be used to identify any changes in the boundaries.

Biographies

Jie Chen received the BS degree in applied mathematics from Chongqing University, China, in 1985; the MS degree in applied statistics from the University of Akron, Ohio, in 1990; and the PhD degree in statistics from Bowling Green State University, Ohio, in 1995. She is currently a professor of statistics in the Department of Mathematics and Statistics at

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the University of Missouri-Kansas City. Her research interests are in the areas of change point analysis, model selection criteria, applied statistics, statistical genetics, and microarray gene expression modeling. She is the leading author of the book *Parametric Statistical Change Point Analysis* (Birkhaüser, 2000).

Yu-Ping Wang received the BS degree in applied mathematics from Tianjin University, China, in 1990, and the MS degree in computational mathematics and the PhD degree in communications and electronic systems from Xi'an Jioatong University, China, in 1993 and 1996, respectively. After his graduation, he had visiting positions at the Center for Wavelets, Approximation and Information Processing of the National University of Singapore and Washington University Medical School in St. Louis. From 2000 to 2003, he worked as a senior research engineer at Perceptive Scientific Instruments, Inc., and then Advanced Digital Imaging Research, LLC, Houston, Texas. In the Fall of 2003, he returned to academia as an assistant professor of computer science and electrical engineering at the University of Missouri-Kansas City. His research interests lie in the interdisciplinary biomedical imaging and bioinformatics areas, where he has more than 90 publications. He has served on numerous program committees and review panels. He was a guest editor for the *Journal of VLSI Signal Processing Systems* on a special issue on genomic signal processing and is a member of Machine Learning for Signal Processing technical committee of the IEEE Signal Processing Society. He is a senior member of the IEEE.

References

[1]. Kallioniemi A, Kallioniemi O-P, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors. Science. 1992; 258:818–821. [PubMed: 1359641]

[2]. Pinkel D, Seagraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo W-L, Chen C, Zhai Y, Zhai Y, Dairkee S, Ljjung B-M, Gray JW, Albertson D. High Resolution Analysis of DNA Copy Number Variation Using Comparative Genomic Hybridization to Microarrays. Nature Genetics. 1998; 20:207–211. [PubMed: 9771718]

[3]. Lucito R, West J, Reiner A, Alexander D, Esposito D, Mishra B, Powers S, Norton L, Wigler M. Detecting Gene Copy Number Fluctuations in Tumor Cells by Microarray Analysis of Genomic Representations. Genome Research. 2000; 10:1726–36. [PubMed: 11076858]

[4]. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO. Genome-Wide Analysis of DNA Copy-Number Changes Using cDNA Microarrays. Nature Genetics. 1999; 23:41–46. [PubMed: 10471496]

[5]. Myers CL, Dunham MJ, Kung SY, Troyanskaya OG. Accurate Detection of Aneuploidies in Array CGH and Gene Expression Microarray Data. Bioinformatics. 2004; 20:3533–3543. [PubMed: 15284100]

[6]. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data. Biostatistics. 2004; 5:557–572. [PubMed: 15475419]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

[7]. Hodgson G, Hager JH, Volik S, Hariono S, Wernick M, Moore D, Nowak N, Albertson DG, Pinkel D, Collins C, Hanahan D, Gray JW. Genome Scanning with Array CGH Delineates Regional Alterations in Mouse Islet Carcinomas. Nature Genetics. 2001; 29:459–464. [PubMed: 11694878]

[8]. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO. Microarray Analysis Reveals a Major Direct Role of DNA Copy Number Alteration in the Transcriptional Program of Human Breast Tumors. Proc. Nat'l Academy of Sciences USA. 2002; 99:12963–12968.

[9]. Weiss MM, Snijders AM, Kuipers EJ, Ylstra B, Pinkel D, Meuwissen SGM, van Diest PJ, Albertson DG, Meijer GA. Determination of Amplicon Boundaries at 20q13.2 in Tissue Samples of Human Gastric Adenocarcinomas by High-Resolution Microarray Comparative Genomic Hybridization. The J. Pathology. 2003; 200:320–326.

[10]. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West JA, Rostan S, Nguyen KC, Powers S, Ye KQ, Olshen A, Venkatraman E, Norton L, Wigler M. Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. Genome Research. 2003; 13:2291–2305. [PubMed: 12975311]

[11]. Autio R, Hautaniemi S, Kauraniemi P, Yli-Harja O, Astola J, Wolf M, Kallioniemi A. CGH-Plotter: MATLAB Toolbox for CGH-data Analysis. Bioinformatics. 2003; 19:1714–1715. [PubMed: 15593402]

[12]. Zhao X, Weir BA, La Framboise T, Lin M, Beroukhim R, Garraway L, Beheshti J, Lee JC, Naoki K, Richards WG, Sugarbaker D, Chen F, Rubin MA, Janne PA, Girard L, Minna J, Christiani D, Li C, Sellers WR, Meyerson M. Homozygous Deletions and Chromosome Amplifications in Human Lung Carcinomas Revealed by Single Nucleotide Polymorphism Array Analysis. Cancer Research. 2005; 65:5561–5570. [PubMed: 15994928]

[13]. Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN. Hidden Markov Models Approach to the Analysis of Array CGH Data. J. Multivariate Analysis. 2004; 90:132–153.

[14]. Hupé P, Stransky N, Thiery J, Radvanyi F, Barillot E. Analysis of Array CGH Data: From Signal Ratio to Gain and Loss of DNA Regions. Bioinformatics. 2004; 20:3413–3422. [PubMed: 15381628]

[15]. Picard F, Robin S, Lavielle M, Vaisse C, Daudin J. A Statistical Approach for Array CGH Data Analysis. BMC Bioinformatics. 2005; 6 article 27.

[16]. Lai WR, Mark D, Johnson MD, Raju Kucherlapati R, Park PJ. Comparative Analysis of Algorithms for Identifying Amplifications and Deletions in Array CGH Data. Bioinformatics. 2005; 21:3763–3770. [PubMed: 16081473]

[17]. Chen, J.; Gupta, AK. Parametric Statistical Change Point Analysis. Birkhaüser; 2000.

[18]. Page ES. A Test for a Change in a Parameter Occurring at an Unknown Point. Biometrika. 1955; 42:523–527.

[19]. Chernoff H, Zacks S. Estimating the Current Mean of a Normal Distribution Which Is Subject to Change in Time. Annals of Math. Statistics. 1964; 35:999–1018.

[20]. Sen A, Srivastava MS. On Tests for Detecting a Change in Mean. Annals of Statistics. 1975; 3:98–108.

[21]. Worsley KJ. On the Likelihood Ratio Test for a Shift in Location of Normal Populations. J. Am. Statistical Assoc. 1979; 74:365–367.

[22]. Horváth L. The Maximum Likelihood Methods for Testing Changes in the Parameters of Normal Observations. Annals of Statistics. 1993; 21:671–680.

[23]. Lehmann, EL. Testing Statistical Hypotheses. second ed.. Wiley; 1986.

[24]. Chen J, Gupta AK. Likelihood Procedure for Testing Change Points Hypothesis for Multivariate Gaussian Model. Random Operators and Stochastic Equations. 1995; 3:235–244.

[25]. Yao Q. Tests for Change-Points with Epidemic Alternatives. Biometrika. 1993; 80:179–191.

[26]. Yao Y-C. Estimating the Number of Change-Points via Schwarz' Criterion. Statistics and Probability Letters. 1988; 6:181–189.

[27]. Gupta AK, Chen J. Detecting Changes of Mean in Multidimensional Normal Sequences with Application to Literature and Geology. Computational Statistics. 1996; 11:211–221.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

[28]. Chen J, Gupta AK. Testing and Locating Variance Change Points with Application to Stock Prices. J. Am. Statistical Assoc. 1997; 92:739–747.

[29]. Chen J, Gupta AK. Change Point Analysis of a Gaussian Model. Statistical Papers. 1999; 40:323–333.

[30]. Chen J, Gupta AK. On Change Point Detection and Estimation. Comm. Statistics-Simulation and Computation. 2001; 30:665–697.

[31]. Chen J, Gupta AK. Information-Theoretic Approach for Detecting Change in the Parameters of a Normal Model. Math. Methods of Statistics. 2003; 12:116–130.

[32]. Chen J, Gupta AK. Statistical Inference of Covariance Change Points in Gaussian Model. Statistics. 2004; 38:17–28.

[33]. Linn SC, West RB, Pollack JR, Zhu S, Hernandez-Boussard T, Nielsen TO, Rubin BP, Patel R, Goldblum JR, Siegmund D, Botstein D, Brown PO, Gilks CB, van de Rijn M. Gene Expression Patterns and Gene Copy Number Changes in Dermatofibrosarcoma Protuberans. Am. J. Pathology. 2003; 163:2383–2395.

[34]. Venkatraman ES, Olshen AB. A Faster Circular Binary Segmentation Algorithm for the Analysis of Array CGH Data. Bioinformatics. 2007; 23:657–663. [PubMed: 17234643]

[35]. Guha, S.; Li, Y.; Donna Neuberg, D. Bayesian Hidden Markov Modeling of Array CGH Data. Harvard Univ. Biostatistics Working Paper Series; Oct. 2006 Working Paper 24

[36]. Lai TL, Xing H, Zhang N. Stochastic Segmentation Models for Array-Based Comparative Genomic Hybridization Data Analysis. Biostatistics. 2008; 9:290–307. [PubMed: 17855472]

[37]. Vostrikova LJU. Detecting Disorder in Multidimensional Random Processes. Soviet Math.-Doklady. 1981; 24:55–59.

[38]. Schwarz G. Estimating the Dimension of a Model. Annals of Statistics. 1978; 6:461–464.

[39]. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Alberston DG. Assembly of Microarrays for Genome-Wide Measurement of DNA Copy Number. Nature Genetics. 2001; 29:263–264. [PubMed: 11687795]

[40]. The Fibroblast Cell Lines Data. 2009. http://www.nature.com/ng/journal/v29/n3/full/ng754.html

[41]. Chen, J.; Wang, Y. Detection of DNA Copy Number Changes Using Statistical Change Point Analysis; Proc. 2006 IEEE Int’l Workshop Genomic Signal Processing and Statistics; 2006; p. 11-12.

[42]. ROMA Data Website. 2009. http://roma.cshl.org/human.html

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

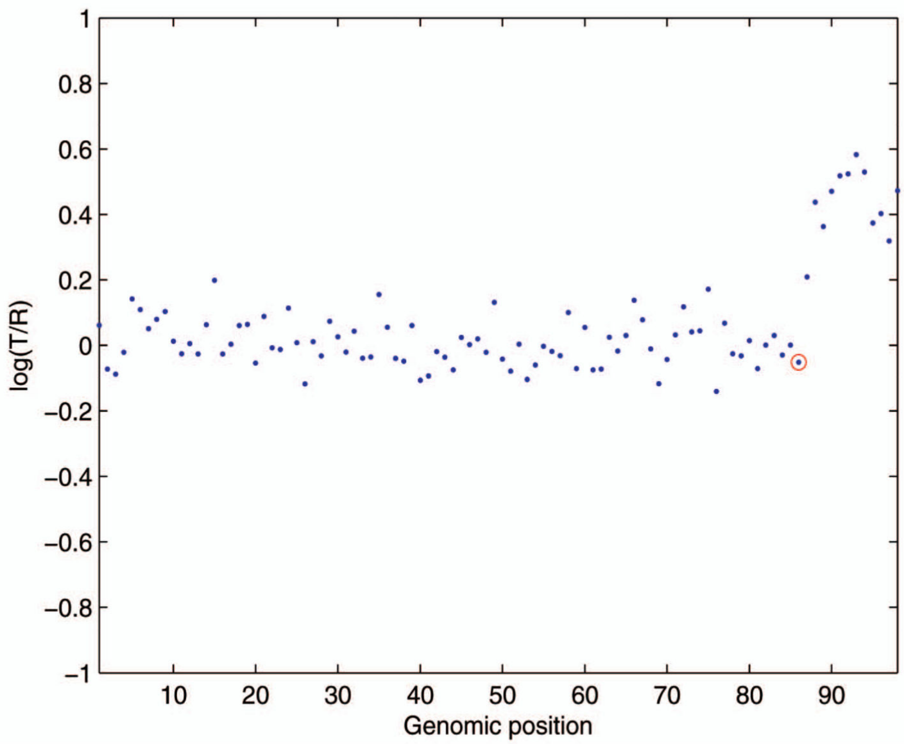

Fig. 1.

Chromosome 5 of the fibroblast cell line GM01535 [39]: In all figures, a red circle indicates a significant DNA copy number change point such that the segment before this red circle (inclusive of the red circle) is different from the successor segment after the red circle (exclusive of the red circle).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

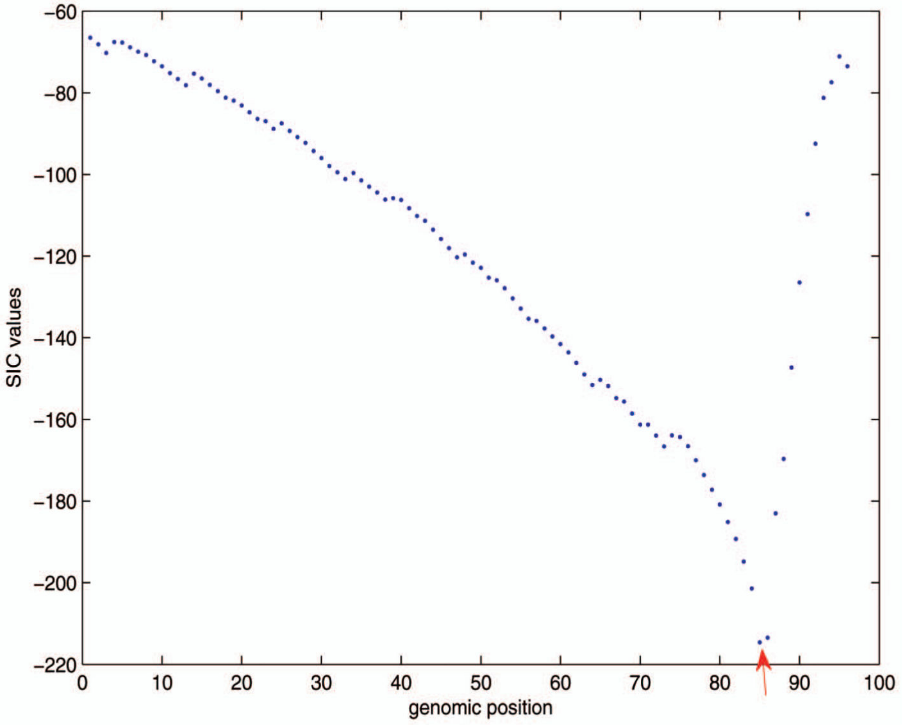

Fig. 2.
SIC plot for chromosome 5 of the fibroblast cell line GM01535 [39].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

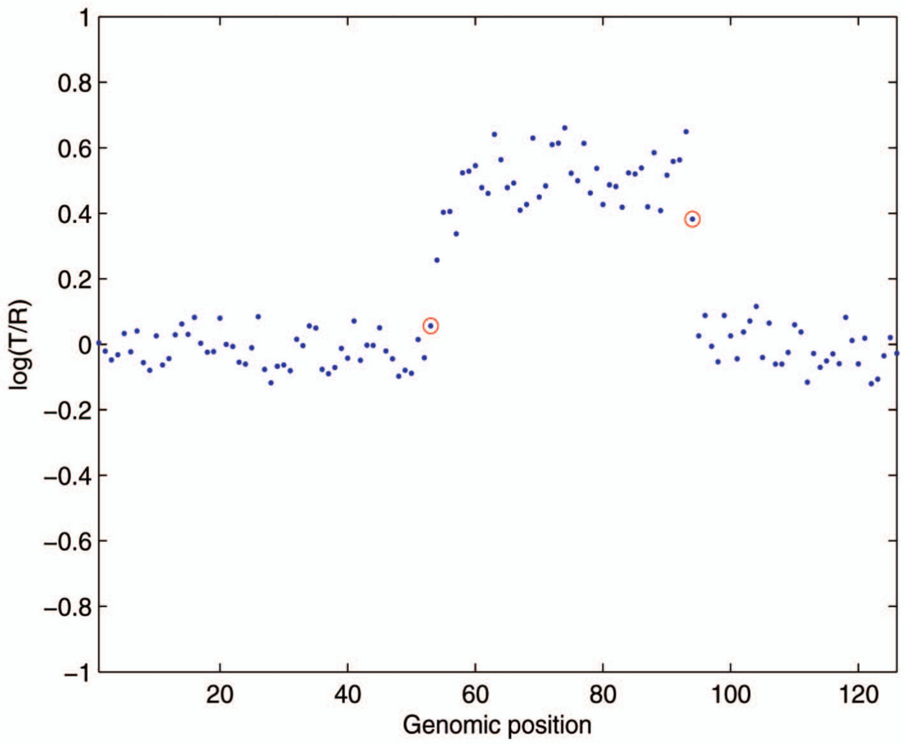

Fig. 3.

Chromosome 10 of the fibroblast cell line GM05296 [39].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

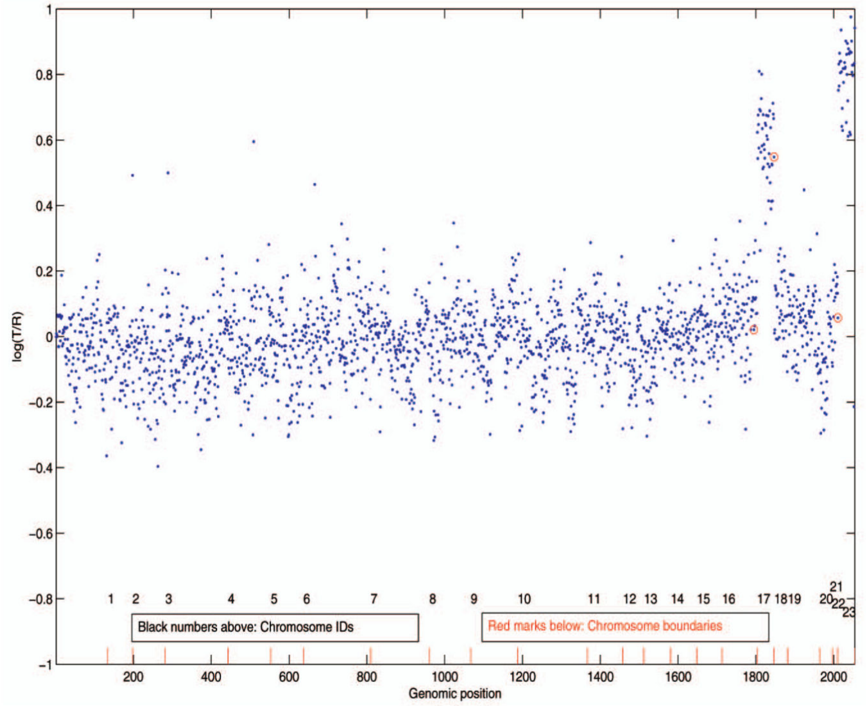

Fig. 4.

Genome of the fibroblast cell line GM00143 [39].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

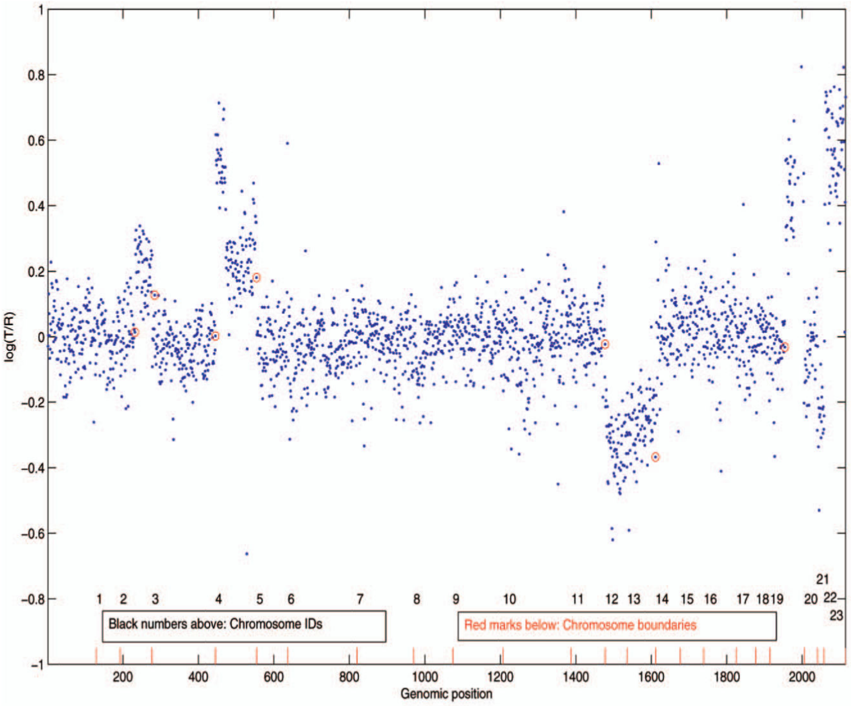

Fig. 5.

Genome of the breast tumor S1514 [39].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

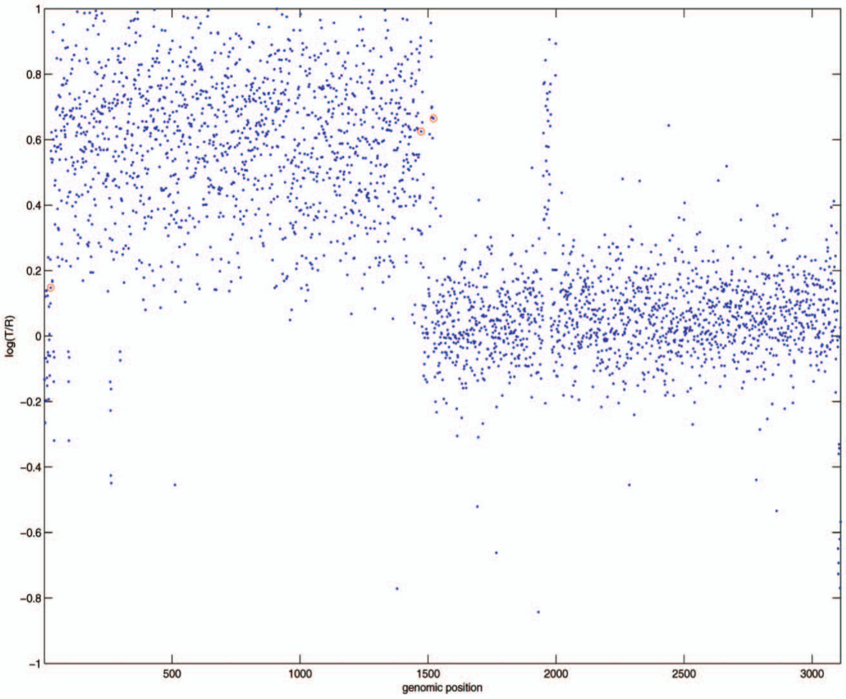

Fig. 6.
Chromosome X of the breast cancer cell line SK-BR-3 [10].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 1

Simulation Results

δ_i	*CV*	**loci**	***n* = 30**		***n* = 40**		***n* = 50**		***n* = 60**	
			f_1	f_2	f_1	f_2	f_1	f_2	f_1	f_2
(.4, .12)	30	Front	0.980	0.707	0.968	0.946	0.970	0.992	0.978	0.999
		Center	0.981	0.970	0.970	1.000	0.982	1.000	0.974	1.000
		End	0.968	0.977	0.977	1.000	0.972	1.000	0.981	1.000
(.41, .13)	32	Front	0.973	0.676	0.975	0.915	0.973	0.982	0.970	0.996
		Center	0.969	0.968	0.979	1.000	0.972	1.000	0.979	1.000
		End	0.981	0.978	0.971	1.000	0.970	0.999	0.976	1.000
(.42, .14)	33	Front	0.969	0.614	0.965	0.890	0.955	0.972	0.967	0.992
		Center	0.969	0.956	0.972	1.000	0.977	1.000	0.971	1.000
		End	0.976	0.971	0.974	0.998	0.967	1.000	0.975	1.000
(.43, .15)	35	Front	0.973	0.592	0.954	0.839	0.973	0.944	0.958	0.983
		Center	0.960	0.972	0.966	1.000	0.962	1.000	0.977	1.000
		End	0.971	0.980	0.973	0.999	0.968	1.000	0.958	1.000
(.44, .16)	36	Front	0.966	0.570	0.961	0.839	0.963	0.943	0.965	0.977
		Center	0.968	0.964	0.968	1.000	0.970	1.000	0.964	1.000
		End	0.967	0.982	0.973	0.999	0.968	1.000	0.963	1.000
(.45, .17)	37	Front	0.956	0.539	0.951	0.782	0.962	0.928	0.953	0.965
		Center	0.948	0.965	0.957	1.000	0.961	1.000	0.959	1.000
		End	0.957	0.986	0.957	1.000	0.951	1.000	0.952	1.000
(.46, .18)	39	Front	0.945	0.505	0.954	0.759	0.954	0.900	0.958	0.942
		Center	0.956	0.965	0.954	1.000	0.966	1.000	0.955	1.000
		End	0.967	0.990	0.961	1.000	0.947	1.000	0.965	1.000
(.47, .19)	40	Front	0.944	0.528	0.946	0.750	0.937	0.882	0.951	0.935
		Center	0.942	0.960	0.952	1.000	0.946	1.000	0.959	1.000
		End	0.951	0.989	0.961	1.000	0.940	1.000	0.953	1.000

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

δ_i	*CV*	**loci**	***n* = 30**		***n* = 40**		***n* = 50**		***n* = 60**	
			f_1	f_2	f_1	f_2	f_1	f_2	f_1	f_2
(.48, .20)	42	Front	0.949	0.470	0.952	0.726	0.939	0.864	0.934	0.926
		Center	0.944	0.964	0.945	1.000	0.953	1.000	0.941	1.000
		End	0.957	0.990	0.961	1.000	0.951	1.000	0.949	1.000

In table, f_1 is the relative frequency that the estimate $\check{k}$ equals the true change location and f_2 is the relative frequency that the p-value, given in (7) of the MVCM approach, is less than 0.01 at $\check{k}$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE

Loci of the Copy Number Changes Found Using MVCM in the Nine Fibroblast Cell Lines [39]

Cell line/Chromosome	locus	p-value	$(\bar{x}, s)$ in each segment
GMO1524/6	49th	2.3724×10^{-4}	(0.0006, 0.0923) (0.3600,0.2643)
GM01535/5	86th	1.7252×10^{-5}	(0.0068, 0.0709) (0.4335, 0.1058)
GMO 1750/9	24th	2.5317×10^{-5}	(0.3966, 0.1159) (−0.0155, 0.0996)
GMO 1750/14	11th	5.6560×10^{-5}	(.4886, 0.0694) (0.0220, 0.0790)
GM03134/8	95th	2.0743×10^{-6}	(−0.0128, 0.0903) (−0.2241, 0.4726)
GM03563/3	39th	1.9396×10^{-5}	(0.0052, 0.1251) (0.4944, 0.1034)
GM03563/9	2nd	3.9577×10^{-5}	(−0.9156, 0.0150) (−0.0228, 0.0810)
GM05296/10	53rd	1.0229×10^{-5}	
	94th	1.0018×10^{-5}	(−0.0165, 0.0527) (0.5002, 0.0886) (−0.0076, 0.0631)
GM05296/11	51st	1.8824×10^{-5}	
	67th	2.1324×10^{-6}	(0.0121, 0.0735) (−0.6159, 0.2224) (0.0180, 0.0738)
GM07081/7	69th	3.9350×10^{-8}	(0.4553, 0.1407) (0.0061, 0.0776)
GM13031/17	55th	9.9718×10^{-4}	
	64th	3.4882×10^{-3}	(0.0339, 0.0798) (−0.5008, 0.2673) (0.0437, 0.1566)
GM13330/1	82nd	2.0249×10^{-7}	(0.0180, 0.0901) (0.5179, 0.1255)
GM13330/4	150th	6.4653×10^{-9}	(−0.0687, 0.0969) (−0.8389, 0.0655)

In table, the loci of significant changes found by our approach are listed along with their p-values. In the last column of the table, we list the sample mean and sample standard deviation before and after the locus of the change, and the sample mean $\bar{x}$ is listed as the first number in the parenthesis and the sample standard deviation s is the second number in the parenthesis for each segment.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 3

Comparison of the Changes Found Using MCM and MVCM on the Nine Fibroblast Cell Lines [39]

Cell line/Chromosome	**MCM with CBS**	**of Olshen *et al.***	**MVCM with BSP**	**of Chen and Wang**
	$\alpha = 0.01$	**$\alpha = 0.001$**	**$\alpha = 0.01$**	**$\alpha = 0.001$**
GMO 1524/6	Yes	Yes	Yes	Yes
Number of false positives	6	2	0	0
Specificity	72.7%	90.9%	100%	100%
Sensitivity	100%	100%	100%	100%
GMO 1535/5	Yes	Yes	Yes	Yes
GM01535/12	No	No	No	No
Number of false positives	2	0	1	1
Specificity	90.5%	100%	95.2%	95.2%
Sensitivity	50%	50%	50%	50%
GMO 1750/9	Yes	Yes	Yes	Yes
GMO 1750/14	Yes	Yes	Yes	Yes
Number of false positives	1	0	0	0
Specificity	95.2%	100%	100%	100%
Sensitivity	100%	100%	100%	100%
GM03134/8	Yes	Yes	Yes	Yes
Number of false positives	3	1	1	0
Specificity	86.4%	95.5%	95.5%	100%
Sensitivity	100%	100%	100%	100%
GM03563/3	Yes	Yes	Yes	Yes
GM03563/9	No	No	Yes	Yes
Number of false positives	8	5	0	0
Specificity	61.9%	76.2%	100%	100%
Sensitivity	50%	50%	100%	100%
GM05296/10	Yes	Yes	Yes	Yes
GM05296/11	Yes	Yes	Yes	Yes
Number of false positives	3	0	1	0
Specificity	88%	100%	96%	100%
Sensitivity	100%	100%	100%	100%
GM07081/7	Yes	Yes	Yes	Yes
GM07081/15	No	No	No	No
Number of false positives	1	0	0	0
Specificity	95.2%	100%	100%	100%
Sensitivity	50%	50%	50%	50%
GM13031/17	Yes	Yes	Yes	Yes
Number of false positives	5	3	1	1

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Cell line/Chromosome	**MCM with CBS** **α = 0.01**	**of Olshen *et al.*** **α = 0.001**	**MVCM with BSP** **α = 0.01**	**of Chen and Wang** **α = 0.001**
Specificity	79.2%	87.5%	95.8%	95.8%
Sensitivity	100%	100%	100%	100%
GM13330/1	Yes	Yes	Yes	Yes
GM13330/4	Yes	Yes	Yes	Yes
Number of false positives	8	5	0	0
Specificity	61.9%	76.2%	100%	100%
Sensitivity	100%	100%	100%	100%

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 4

Copy Number Changes Found Using MVCM on the Genome of the Other Six Fibroblast Cell Lines [39

Cell line/Chrom.	change	Cell line/Chrom.	change
GM00143/18	Yes	GM02948/13	Yes
Number of false positives	1	Number of false positives	0
Specificity	87.5%	Specificity	100%
Sensitivity	100%	Sensitivity	100%
GM07408/20	Yes	GM10315/22	Yes
Number of false positives	1	Number of false positives	0
Specificity	75%	Specificity	100%
Sensitivity	100%	Sensitivity	100%
GM03576/2	Yes	GM04435/16	Yes
GM03 576/21	No	GM04435/21	No
Number of false positives	1	Number of false positives	1
Specificity	75%	Specificity	83.3%
Sensitivity	50%	Sensitivity	50%

The notations are the same as in Table 3 and the significance level is 0.001.