Published in final edited form as:

Biometrics. 2016 September ; 72(3): 815–826. doi:10.1111/biom.12478.

Sequential Model Selection based Segmentation to Detect DNA Copy Number Variation

Jianhua Hu^{1,*}, Liwen Zhang^{2,**}, and Huixia Judy Wang^{3,***}

¹Department of Biostatistics, UT M. D. Anderson Cancer Center, Houston, Texas, U.S.A

²School of Economics, Shanghai University, Shanghai, China

³Department of Statistics, George Washington University, Washington D.C., U.S.A

Summary

Array-based CGH experiments are designed to detect genomic aberrations or regions of DNA copy-number variation that are associated with an outcome, typically a state of disease. Most of the existing statistical methods target on detecting DNA copy number variations in a single sample or array. We focus on the detection of group effect variation, through simultaneous study of multiple samples from multiple groups. Rather than using direct segmentation or smoothing techniques, as commonly seen in existing detection methods, we develop a sequential model selection procedure that is guided by a modified Bayesian information criterion. This approach improves detection accuracy by accumulatively utilizing information across contiguous clones, and has computational advantage over the existing popular detection methods. Our empirical investigation suggests that the performance of the proposed method is superior to that of the existing detection methods, in particular, in detecting small segments or separating neighboring segments with differential degrees of copy-number variation.

Keywords

Array-based CGH; Bayesian information criterion; Copy-number variation; Segmentation; Sequential model selection

1. Introduction

High-throughput bioinformatics technologies, such as array-based comparative genomic hybridization (aCGH) experiments (Snijders et al., 2001), are designed to measure genome-wide regions of DNA copy-number variation (CNV). Typically the normal DNA copy number for humans is two for all of the autosomes. However, the copy number in a region of a genome can be altered by the development and progression of cancer or other diseases. In aCGH cancer studies, DNA samples isolated from both cancerous and normal cells are labeled with two distinct fluorescent dyes and are then hybridized to a microarray that was

¹Corresponding author: Jianhua Hu. ²Co-Corresponding author: Liwen Zhang.

*jhu@mdanderson.org

**lzhang7399@gmail.com

***judywang@gwu.edu

previously spotted with DNA sequences which map to chromosomal regions within the human genome. Data analysis of the scanned array is typically implemented on the logarithm-transformed ratio of the intensities of the two fluorochrome at each spot compared with the known physical chromosomal location on the microarray.

The existing methods are generally designed to analyze a single array at a time for detection of copy-number variations. Segmentation methods are one important class of such methods. This class includes Tibshirani and Wang (2007), Guha et al. (2008), Lai et al. (2008), and others. A widely used method in this class is the circular binary segmentation (CBS) algorithm, which was proposed by Olshen et al. (2004). Niu and Zhang (2012) proposed a Screening and Ranking algorithm (SaRa) to detect the multiple change points in a single array. They used the neighborhood information of each position to check its probability being a change point. It is computationally efficient with complexity $O(n)$. However, the accuracy of the approach is dependent on the window size selection which is quite challenging. Up to date, the assumption of definite segments with discrete boundaries was adopted by a lot of researchers (Pinkel and Albertson, 2005). Additionally, we observed the phenomenon of discrete segments of copy-number variation along the chromosomes in our analysis of published data, which is discussed in Section 5. Our experience and observations motivated us to assume a sequence of discrete segments of DNA copy-number variation along the chromosomes when developing our proposed statistical method.

Identifying common/recurrent regions shared by the samples in a common disease group is biologically meaningful, since the identified regions could more likely correspond to important genes associated with the disease. Rouveirol et al. (2006) used “recurrent region” to define such sequence of adjacent clones with aberrations shared by samples in a common group. A review of methods for detecting recurrent CNV regions is available in Rueda et al. (2010).

We also note that comparison of samples from multiple groups has been largely neglected. It is indeed an important area since biologists are not only interested in identifying the common aberrant regions for samples from one group, but also interested in identifying regions where two groups differ in copy number in many aCGH studies. For the problem of a single group study, researchers generally applied existing aforementioned algorithms to each individual sample independently, and then combine the analyzed individual profiles to identify common aberrant regions; see Bendor et al. (2007), Ylipaa et al. (2008), and Klijn et al. (2008), among others. More recent development on multi-sample analysis includes a proposal of sum of chi-square statistics to combine individual samples (Zhang et al., 2010; Siegmund et al., 2011) borrowing the essential idea of the single-sample approach in (Olshen et al., 2004), and a proportion adaptive segment selection procedure to detect both the rare and common copy number variants (Jeng et al., 2013). A simple false discovery rate approach (Efron and Zhang, 2011) was also proposed to control for multiple testing.

For the comparison of two groups, Willenbrock and Fridlyand (2005), and Huang et al. (2007) implemented segmentation methods on a single index (group mean difference) at each clone location along the chromosomes. This implementation procedure does not make good use of the data across patient samples and clones, and thus may lose detection

sensitivity. Wang and Hu (2011) developed a penalized regression approach with a fused adaptive lasso penalty and determined the nonrandom aberrant genomic segments by assessing the significance through bootstrap. However, their method cannot detect within-group segments and tends to over-estimate the number of differential segments.

In this paper, we propose a new approach in which the log-ratios across samples and clones are modeled by a linear regression model and the challenge of detecting regions of copy-number variation is converted into a model selection problem. Its innovation is threefold: First, different from the typical model selection that aims to select the optimal set of explanatory variables, the proposed method is to select the parsimonious set of segments, each of which shares the common mean intensity in a group. Second, we propose a sequential model selection procedure such that the information across the contiguous clones can be accumulatively used to improve detection accuracy. Third, the new method is capable of simultaneously detecting the common aberrant regions within each group and identifying the differential regions between multiple groups. It is evident by our empirical investigation that the proposed method has superior performance in terms of segment detection, particularly good at detecting small segments or neighboring segments with differential degrees of copy-number variation. The proposed method also has a computational advantage over the existing methods because it requires computation only on the order of the number of clones, whereas the MSCBS method of Zhang et al. (2010), as an example, requires computation on the order of the number of clones times the sample size.

We consider Bayesian information criterion (BIC, Schwartz (1978)) type of methods for model selection. BIC has the desirable consistency property in the sense that it selects the true model with probability approaching one if the true model is in the class of candidate models. This consistency has been formally discussed when the likelihood function is correctly specified (Nishii, 1984). However, the nice property holds only with the fixed number of parameters (Shao, 1997; Shi and Tsai, 2002). More recently, Wang et al. (2009) modified BIC in the case of diverging number of parameters by including a function of the total number of observations as a scaling factor in the penalty term, which preserves model selection consistency. We follow this same idea to guide the automatic model selection procedure for segment detection.

This paper is organized as follows. We introduce the proposed method and model selection procedure in Section 2. The finite sample performance of the proposed method is investigated through a simulation study in Section 3. We apply the proposed method on a myeloma cancer study (Carrasco et al., 2006) in Section 4 and provide some concluding remarks in Section 5.

2. Proposed approach

2.1 Model formulation

In an aCGH experiment, let y_{ij} denote the log₂-ratios of the fluorescence intensities between the tumor sample of the i th subject and the reference sample on the j th clone. Assume that the first n_1 and the last n_2 subjects belong to two different groups (e.g. wild-type versus

mutant samples). Our main objective is to identify the clone regions of the change points within each group simultaneously.

Let g_i index the group membership with 0 and 1 corresponding to the wild-type and mutant-type samples, respectively. We consider the following model

$$y_{ij} = \alpha_i + \mu_j + \beta_j g_i + e_{ij}, \quad (1)$$

for $i = 1, \dots, n$ and $j = 1, \dots, p$, where α_i is the individual effect for the i th subject, μ_j and β_j are the baseline and group effects of the j th clone, respectively, and e_{ij} are independent random errors with mean 0 and variance of σ_j^2 . Denote $\eta_j = \mu_j + \beta_j$, and we note that the effects on the j th clone for the reference and tumor samples are μ_j and η_j respectively. Denote $\mathbf{Y} = (y_{11}, \dots, y_{1p}, \dots, y_{n1}, \dots, y_{np})^T$ as the $N \times 1$ response vector with $N = np$ where $n = n_1 + n_2$. We can rewrite model (1) in the matrix form

$$\mathbf{Y} = \mathbf{\alpha} + \mathbf{X}\boldsymbol{\theta} + \mathbf{e}, \quad (2)$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} \otimes \mathbf{I}_p & \mathbf{0} \\ \mathbf{1}_{n_2} \otimes \mathbf{I}_p & \mathbf{1}_{n_2} \otimes \mathbf{I}_p \end{pmatrix}, \mathbf{\alpha} = (\alpha_1, \dots, \alpha_n)^T \otimes \mathbf{1}_p, \boldsymbol{\theta} = (\mu_1, \dots, \mu_p, \beta_1, \dots, \beta_p)^T,$$

$\mathbf{e} = (e_{11}, \dots, e_{1p}, \dots, e_{n1}, \dots, e_{np})^T$, $\mathbf{0}$ is the zero matrix with $n_1 p$ rows and p columns, $\mathbf{1}_n$ denotes the $n \times 1$ vector consisting of 1's, and $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. Let $d_0 = 0$ and $d_{H+1} = T$. The mathematical problem is to identify H change-point clone locations $\{d_h; h = 1, \dots, H\}$ such that $\mu_j = \mu_{(h)}$, $\eta_j = \eta_{(h)}$ if $d_h < j < d_{h+1}$ for $h = 0, 1, \dots, H$, where $\mu_{(h)} = \mu_{(h+1)}$ or $\eta_{(h)} = \eta_{(h+1)}$ for $h = 1, 2, \dots, H$.

Rather than a direct segmentation algorithm such as the circular binary segmentation method of Olshen et al. (2004), we will approach this problem from the viewpoint of model selection. Distinct from the typical model selection problem of obtaining an optimal set of nontrivial covariates, our challenge is to select a set of change-point clone locations such that the clones in each segment identified by two contiguous change-point locations have the common mean effect. Most existing segmentation approaches are designed to identify the segments of a single array with different location parameters, μ . In contrast, we focus on studies with multiple samples, and aim to identify segments for each group. Note that H ranges from 1, where all the clones have the common group effect, to p , where each clone forms one segment. Our task is to identify a parsimonious model with a small H that can adequately capture the information in the data.

2.2 Criterion

We index each candidate model by a p -vector $\boldsymbol{\nu} = (\gamma_1, \dots, \gamma_p)^T$, where $\gamma_j = h$ if clone j belongs to the segment h , $h = 1, \dots, H$. For instance, for a chromosome with total 10 clones, $\boldsymbol{\nu} = (1, 1, 1, 1, 1, 2, 3, 3, 3, 3)^T$ indicates that the first 5 clones belong to segment 1, the 6th clone forms segment 2, and the last 4 clones belong to segment 3. Let $\boldsymbol{\xi}_{\boldsymbol{\nu}}$ denote the vector of parameters in the candidate model and $d_{\boldsymbol{\nu}}$ denote the total number of distinct parameters in the candidate model, excluding the subject-specific effects α_j , which measures the complexity of the model. In the previous example, $\boldsymbol{\xi}_{\boldsymbol{\nu}} = (\mu_1, \dots, \mu_1, \mu_2, \mu_3, \dots, \mu_3, \beta_1, \dots, \beta_1, \beta_2, \beta_3, \dots, \beta_3)^T$, the number of change points $H = 2$, and $d_{\boldsymbol{\nu}} = 2(H + 1) = 6$.

We adopt the following BIC type of method to select the final model via minimizing the objective function

$$LBIC(\boldsymbol{\nu}) = L_2(\hat{\boldsymbol{\xi}}_{\boldsymbol{\nu}}) + \lambda_N d_{\boldsymbol{\nu}} \log(N), \quad (3)$$

where $\hat{\boldsymbol{\xi}}_{\boldsymbol{\nu}}$ is the minimizer of the weighted L_2 distance

$$L_2(\boldsymbol{\xi}_{\boldsymbol{\nu}}) = \sum_{k=1}^N \frac{(y_k - \alpha_k - \mathbf{x}_k^T \boldsymbol{\xi}_{\boldsymbol{\nu}})^2}{w_k}, \quad (4)$$

where y_k , α_k and $\mathbf{x}_k$ are the k th rows of $\mathbf{Y}$, $\boldsymbol{\alpha}$ and $\mathbf{X}$, respectively. Herein, $\lambda_N = c \log\{\log(N)\}$ (Wang et al., 2009), where N is the total number of clone intensity observations. Note that (4) is reduced to the traditional BIC with $\lambda_N = 1$. Our empirical investigation suggests that the range between 0.3 and 0.4 for the value of c is sensible. Here the weights $\{1/w_k\}$ are incorporated to account for the unequal variation of signal intensities among the samples at different clones. In practice, the w_k can be estimated from a small-bias model, such as the full model without affecting the consistency property of the model selection procedure. For y_k corresponding to the j th clone and using the notations in (1), hereafter, we adopt

$$w_k = \frac{1}{n} \sum_{i=1}^n (y_{ij} - \hat{\alpha}_i - \hat{\mu}_j - \hat{\beta}_j g_i)^2, \quad (5)$$

where $\hat{\alpha}_i$, $\hat{\mu}_j$ and $\hat{\beta}_j$ are the estimates of the parameters in the full model where each clone forms an individual segment. In this application, we treat $\{w_k\}$ as fixed to reduce the computational complexity.

We make a practical assumption that fewer than 50% clones are expected to show aberrations. For computational consideration, we take the sample median across clones to be

the estimate of subject baseline effect α_j throughout the model selection procedure described in the following section.

2.3 Sequential model selection procedure

Note that there are a total number of 2^p candidate models. Clearly, an exhaustive search is not feasible, as p is often in the order of thousands in aCGH studies. A sequential procedure that utilizes the physical location ordering of the clone among the chromosome is stated as follows.

Assume that the clones c_j ($j = 1, \dots, p$) are ordered in a sequence with the physical position increasing as we read the location from the left to the right.

1. We start by considering whether the k^{th} left-most clone c_k is a change-point location. To decide whether c_k belongs to the segment s_1 formed by clones $[c_1, c_{k-1}]$, we conduct the following hypotheses

$$\begin{aligned} &H_0: \mu_k = \mu_{(s_1)} \text{ and } \eta_k = \eta_{(s_1)} \\ \text{vs. } &H_a: \mu_k \neq \mu_{(s_1)} \text{ or } \eta_k \neq \eta_{(s_1)}, \quad (6) \end{aligned}$$

which state that k is a change point under the alternative hypothesis and not under the null. For example, if $k = 5$, the model configuration is $\mathbf{v} = (1, 1, 1, 1, 1, 2, 3, 4, \dots, p-k+1)$ under H_0 and $(1, 1, 1, 1, 2, 3, 4, 5, \dots, p-k+2)$ under H_a . Denote the objective function (3) under models H_0 and H_a as LBIC_0 and LBIC_1 , respectively. If $\text{LBIC}_0 - \text{LBIC}_1 \leq 0$, then c_k belongs to the segment s_1 and c_k is not a change-point; otherwise, $d_1 = c_k$.

2. If c_{j-1} is not a change-point location, we determine whether the next clone c_j can join the previous segment s_h . Only clone c_j and all the clones on the segment s_h are involved in the parameter estimation and evaluation of the difference in the objective function (3) between the two models under H_0 : $\mu_j = \mu_{(s_h)}$ and $\eta_j = \eta_{(s_h)}$ versus H_a : $\mu_j \neq \mu_{(s_h)}$ or $\eta_j \neq \eta_{(s_h)}$. If c_{j-1} is a change point, we move to clone c_{j+k-1} to examine whether it can join segment s_{h+1} formed by clones $[c_j, c_{j+k-2}]$. In this case, only clones $[c_j, c_{j+k-1}]$ are considered in model evaluation. This procedure is sequentially implemented until reaching the right-most clone c_p .
3. Up to this point, only a single clone location at a time has been considered for merging with the previous segments. Our empirical investigation indicates that this would result in the discovery of small segments due to the sensitivity of the algorithm to the single clone location. Therefore, we further refine the segmentation results by determining in a sequential manner whether the detected neighboring segments can be combined based on the objective function (3). To diminish the impact of search directions on the change point detection, we switch the search directions in the following circular manner. If the merging direction at steps 1–2 is from the left to the right, the next stage of segment re-merging is conducted

from the right to the left, and the direction switch continues until no modifications can be made. Similarly if the initial searching order at Steps 1–2 is from the right to the left, the re-merging will be examined from the left to the right, and continue the direction switch until convergence.

Remark 1—The discussion thus far focuses on simultaneous detection of the change points within each group. Another important problem in real data applications is to identify group-difference-associated segments (e.g., disease or not), whereas we are interested in identifying H change-point clone locations $\{d_h; h = 1, \dots, H\}$ such that $\beta_j = \beta_{(h)}$ if $d_h < j < d_{h+1}$ for $h = 0, 1, \dots, H$, where $\beta_{(h)} = \beta_{(h+1)}$ for $h = 1, 2, \dots, H$. For this purpose, we can modify the hypotheses in (6) as

$$M_0: \beta_k = \beta_{(s_1)} \quad \text{vs.} \quad M_1: \beta_k \neq \beta_{(s_1)}, \quad (7)$$

and implement the procedure similarly as steps 1–3.

Remark 2(a)—In principle, the algorithm can start from any clone positions. We examine two scenarios for illustration: (i) starting from the k^{th} left-most towards the right end; (ii) starting from the k^{th} right-most clone towards the left end. Our empirical studies manifest only slight variation between two cases of different starting points, with the circular re-merging procedure. Alternatively, we can consider multiple starting points, among which the one maximizing the likelihood can be chosen as the final solution.

Remark 2(b)—The rationale for starting from the k^{th} clone away from an end in step (1) is that we believe the copy-number variation segments that are truly biologically relevant should cover at least $k - 1$ clones because the aCGH data provide a dense map of the whole genome. The choice of k is rather motivated from biological perspective in terms of the smallest size of a copy-number variation region that can be claimed scientifically meaningful. In our empirical study, we used $k = 6$. The principle of the proposed procedure is invariant to the choice of k .

Remark 2(c)—The procedure we propose, which we call the model selection segmentation (MSS) method, has several computational advantages. First, the proposed sequential procedure only requires the order of p steps, in contrast to a computational complexity in the order of $np \log p$ for the MSCBS algorithm. Second, at each step of the proposed method, the computation involves the examination of only the current clone c_j (or segments described in step (2)) and the clones on the segment where the previous clone c_{j-1} is located (or the previous segment), which reduces the computational cost. Third, the method does not require computationally expensive bootstrap/permutation procedures.

3. Simulation study

The performance of the proposed method is illustrated by five cases, two of which intend to mimic a real aCGH experiment. In each case, 100 data sets are generated. We compare our

proposed model selection segmentation (MSS) method with three existing methods: the multiple-sample circular binary segmentation (MSCBS) method of Zhang et al. (2010), the proportion adaptive segment selection method (PASS) of Jeng et al. (2013), and the fused adaptive lasso selection method (FAL) of Wang and Hu (2011). Both MSCBS and PASS methods are applicable to multiple arrays, and have been shown to have superior performance over several other popular methods. PASS and MSCBS methods are implemented using R code provided at <https://sites.google.com/site/xingejeng/> and http://statweb.stanford.edu/~nzhang/web_msscan/, respectively. The R code for implementing FAL is available at <http://home.gwu.edu/~judywang/research/software/>.

We consider three variants of the proposed method: MSS-LTR, MSS-RTL and MSS-UN, which correspond to the weighted left-to-right segmentation detection, weighted right-to-left segmentation, and unweighted left-to-right procedures, respectively. In comparison to the other methods in Table 1 and Figures 1, 2 and 3, the results of MSS methods are reported at $c = 0.35$. Furthermore, we examine the sensitivity of the BIC criterion by considering $c = (0.3, 0.35, 0.4)$ and show the results in Table 2.

Case 1—The data is generated from model (1), where the random errors e_{ij} 's are randomly sampled from $N(0, 0.5^2)$. We consider 1000 clones and 10 samples in each group. A total of 10 contiguous segments with distinct mean effects are formed by clones [1, 10], [11, 25], [26, 45], [46, 340], [341, 360], [361, 660], [661, 675], [676, 695], [696, 705], and [706, 1000], respectively. From the left to the right along the physical locations, the mean effects of the first and second groups corresponding to the 10 segments are $\mu_j = 0, 1.36, 0.60, 0, -0.45, 0, 0.75, 1.38, 3.78, 0$ and $\eta_j = 0, 0, 0.52, 0, 0.45, 0, 0.68, 1.42, 3.82, 0$, respectively. This data generating mechanism allows for evaluating the performance of the methods subject to various lengths of the nontrivial-group-effect segments between 10 and 20 and various effect magnitudes of segments between 0.45 and 3.82.

For each simulated data set, segments identified by a method is considered correct if a change point is detected within the vicinity of the boundary between two true contiguous segments. Over 100 simulations, we report the frequency of accurate detection for each true change point (TP), the frequency of the number of incorrectly identified change points (FP), the average number of false positives (MFP), and the average computing time in seconds (Time) for analyzing a simulated data set by different methods in all the cases in Table 1. The results for Case 1 is in the top panel. Note that errors in this case are homoscedastic, so it is not a surprise that the two weighted MSS methods perform similarly with MSS-UN. Overall, the MSS methods outperform both PASS and MSCBS. Specifically, PASS misses the true change points 340 and 360 most of the times, and MSCBS misses the true change points 10 and 340. Furthermore, both of them, especially MSCBS, produce much more false positives than the MSS methods. The result in the rightmost column also suggests that MSS methods tend to be more computationally efficient than PASS and MSCBS.

For a visual demonstration, we show the results of various methods for a simulated data set in Figure 1. The mean effects of the clones for the tumor and reference groups are indicated by the yellow and green dots in all the panels, respectively. Meanwhile, we label the detected segments for the tumor and reference groups as red and blue horizontal bars. The true

change point locations are indicated by vertical lines. We observe that the three MSS methods show similar results; three variants are able to detect all the true segments. In contrast, PASS can not distinguish small segments, such as [26, 45], [341, 355] and [696, 705], and MSCBS detects a lot of false segments than the other methods.

Case 2—The model is similar to that in Case 1, except that heteroscedastic errors are considered in Case 2. The random errors are set to be $e_{ij} = (1 + 1.2X_j)e_{ij}$, where $X_j \sim U[0, 1]$ and $e_{ij} \sim N(0, 0.5^2)$. The simulation results are presented in Table 1. The weighted MSS methods show clear advantage over the un-weighted version in terms of true segment detection, especially at the change points 25, 45, 340 and 360. Similar to Case 1, MSS methods, particularly MSS-LTR, overall outperform PASS and MSCBS.

Case 3—In our third simulation case, we intend to mimic real an aCGH experiment. The example data set consists of formalin-fixed tissue samples of primary oral squamous cell carcinomas (Snijders et al., 2001), available at <http://www.cbs.dtu.dk/~hanni/aCGH/>. This data set contains 14 TP53 mutant samples and 61 wild-type samples with log-ratio expression intensities available at 1979 clone locations along the chromosomes. The scientific goal is to detect the genomic regions with copy-number variations that are associated with TP53 mutant status.

To generate the residuals in each simulation, we first randomly select 10 samples from each of the two groups (TP53 mutant or not) and subtract the mean group intensity from the individual expression intensities in each of the two groups at each clone location. We then randomly select 1000 clone locations with no missing data and perturb the centered residual values among 20 samples at each location to make sure the residuals did not carry any group-specific information. We generate a total of 11 segments. We intentionally make this case more difficult than Case 1 by considering important segments with a length as short as 5. Specifically, the segments defined in order from the left-most to the right-most clone positions are 1 to 100, 101 to 110, 111 to 120, 121 to 430, 431 to 450, 451 to 745, 746 to 750, 751 to 760, 761 to 770, 771 to 800, and 801 to 1000. The corresponding segment-wise baseline intensity, μ_j , is arbitrarily set to 0, -0.65, -0.95, 0, 0.47, 0, 1.06, 0, 0.65, 1.42, 0, while the second group effect η_j is corresponding to 0, 0, 0.75, 0, -0.53, 0, -0.99, 0, 0.82, 1.52, 0. The results of 100 simulations are shown in the panel of Case 2 of Table 1. The results indicate that PASS fairly frequently misses the true change points 110, 430, 450, 745, 750, 770. In contrast, MSS and MSCBS have better performance, whereas MSCBS is inferior to the MSS methods in distinguishing 750 and 770, wherein 750 is a change point of the shortest segment. Again, the proposed method shows clear advantage in computation and false positives over PASS and MSCBS.

The results of all these methods in one simulation are shown in Figure 2. As in Figure 1, the mean effects of the clones for the two groups are indicated by yellow and green dots in all the panels, with the vertical lines corresponding to the locations of true segment change points. We can see that the MSS methods starting from both left and right points can detect all the true segments. In contrast, both the PASS and MSCBS methods miss the shortest segment [746, 750] with a nontrivial effect, whereas the former also fails to distinguish the

short segment [761, 770]. In addition, the MSCBS method detects more false positives than the other two methods.

In Cases 1–3, we focus on detecting the change points within the tumor or reference group of multiple samples. These cases demonstrate the detection accuracy and computing advantages of the proposed MSS method over several popular methods. In Cases 4 and 5 considering different numbers of clones, we intend to detect segments defined on group differences (e.g., diseased versus normal), which is often of interest in disease association studies. We investigate the FAL method (Wang and Hu, 2011) and MSS methods.

Case 4—We take the mean segment intensities of the two groups, respectively, as (0, −0.65, −0.95, 0, 0.87, 0, 1.06, 0, 0.1, 2.2, 0) and (0, 0, 0.75, 0, −1.05, 0, −0.99, 0, 0.82, 0.95, 0). This results in the mean group differences in the corresponding segments as (0, 0.65, 1.7, 0, −1.92, 0, −2.05, 0, 0.72, −1.25, 0). The residuals are generated similar as that in Case 3. MSS-LTR is able to detect almost all the true change points in each run, while FAL fails to detect change points 100, 745, 750, and 760 in multiple data sets. In particular, FAL scarcely detects the shortest segment [745, 750], indicating its lower capability of short segment detection. We also notice that MSS outperforms FAL in terms of false segment detections in terms of both FP and MFP. In addition, implementing MSS is much faster than FAL, as shown in the last column of Table 1.

Figure 3 presents the results of various methods for a simulated data set. The mean group difference of the clones are indicated by the green dots in all panels, where the detected segments of group difference are represented by the red bars. MSS methods are observed to detect all the true segments and very few false segments. In general, MSS shows better performance than FAL, which fails to detect the segments [111, 120] and [761, 770] in this simulated data set, for example.

Case 5—This case is to mimic much higher dimensional bioinformatics data which is often encountered in recent biomedical studies (Duan et al., 2013 and Tan et al., 2014). We consider 20,000 clones and 10 samples in each group with 11 contiguous segments. The segments from the left to the right are [1, 4000], [4001, 4100], [4101, 4200], [4201, 10000], [10001, 10050], [10051, 18000], [18001, 18020], [18021, 18040], [18041, 18060], [18061, 18100], [18101, 20000]. The corresponding segment-wise intensities of groups 1 and 2 are set to (0, 0.95, 1.20, 0, −0.66, 0, −0.58, 0, −0.36, −0.98, 0) and (0, 0, −0.99, 0, 0.60, 0, 0.49, 0, 0.55, 0.75, 0), respectively. This gives the segment-wise differential group effects of (0, −0.95, −2.19, 0, 1.26, 0, 1.07, 0, 0.91, 1.73, 0). The random errors e_{ij} are generated from $N(0, 0.1^2)$.

The simulation results presented in the bottom panel of Table 1 show that our proposed MSS method also works for the ultra-high dimensional bioinformatics data. Similar to Case 3, MSS method outperforms FAL in terms of both segment detection accuracy and computing speed. We note that MSS-UN method runs slower than the two weighted methods because that the former one often detects much longer segments in the quite homogeneous data than the latter two, for which simple algebras such as summations of values within a segment are computationally inefficient in high-dimensional cases.

Moreover, it is interesting to investigate the impact of the value of c in the BIC criterion on segment detection. We show the results of all the five cases at several c values between 0.3 and 0.4 with the MSS-LTR procedure in Table 2. It is intuitive that the number of detected segments decrease along the values of c since stronger penalty is imposed in the BIC criterion. Overall, the results are reasonably similar among different c values.

4. Multiple myeloma study

Multiple myeloma is characterized by clonal proliferation of plasma cells in the bone marrow. High-resolution aCGH genomic profiling of multiple myeloma patients were performed as described in Carrasco et al. (2006). This data set contains 38 relapse-free and 26 relapsed patient samples, and 16,097 clones located on all the 24 chromosomes for each patient sample. The primary interest is to identify copy-number variation associated with the relapse status. Equivalently, we intend to identify clone segments, each of which has a common degree of differential expression between the two groups of relapsed and relapse-free patients.

We apply the MSS, PASS, and MSCBS methods to the multiple myeloma data set. For the MSS method, we consider identifying group-different-associated segments by focusing on hypotheses (7) and we only include the result of MSS-LTR since the simulation showed that the initial searching direction has little impact on segmentation. The penalty parameter c in the BIC criterion takes the value of 0.35, as in the simulation studies. Since PASS and MSCBS can only be used to identify segments within one group, we apply them to the relapsed and relapse-free groups separately. We do not include the FAL method since it cannot be applied in the presence of missing observations.

By applying the MSS method, we detect 19 change points across all the chromosomes, three of which reside on chromosome 1 with the change point locations of 849, 914 and 1,163. In addition, we detect four change points on chromosomes 4, 13, 14, and 17, with the corresponding clone locations 3,707, 10,285, 10,635, and 12,805, respectively. The results appear to be coherent with the findings of Smetana et al. (2014) that multiple myeloma is likely associated with CNV regions on these chromosomes. Hereafter, we focus our study on chromosome 1, which is recognized for its frequent occurrence of abnormalities that are associated with multiple myeloma.

We focus our further investigation on chromosome 1. We notice that PASS and MSCBS detect many and small segments, 147 and 1210, respectively. In contrast, MSS divides the large region [1, 1662] into 4 segments, whose importance can be ranked according to group effect tests based on the aCGH profiling (see Table 3(a)). With this information available, scientists can conduct focused and efficient study on the most important segments when the resource is limited. It is interesting to further explore the biological relevance of the 4 segments with significant group effect. We conduct an integrated gene expression microarray data analysis, as CNV functions to alter expression of resident genes (Tonon et al., 2005). The gene expression data is of Affymetrix H133Plus2.0 Genechip platform in which the annotation file contains the genomic positions of each gene. The RNA samples of the same set of 64 patients are extracted from their bone marrow-derived plasma cells prior

to any treatments. This data set is available in NCBI's Gene Expression Omnibus through GEO Series accession number GSE4452.

We first implement quantile normalization (Bolstad et al., 2003) to make the arrays comparable. It is followed by performing the ordinary two-sample t-test between the group of relapsed patients and the group of relapse-free patients for all the 4,179 genes on chromosome 1 at the logarithm-transformed scale of gene expression intensities. The task is to identify the genes which have significantly differential expression between the relapsed and relapse-free groups residing on these 4 segments. Table 3(b) reports the 14 genes with p-value no larger than 0.005. The column "rank" contains the ranks of the listed genes among all the genes on chromosome 1 according to the significance of the t-test in the decreasing order. The column "segment" contains the indexes of 1, 2 and 4 indicating segments [1163, 1662], [914, 1162] and [1, 854], respectively. The column "gene name" contains the gene symbols. The column "alternation" contains the minus and plus symbols, representing down-regulation and up-regulation in the group of relapsed patients. The column "location" contains the genomic regions of the genes on chromosome 1.

Our further investigation manifests that segments 1, 2, and 4 contain 7, 6, and 1 interesting genes, respectively. The first gene *RBM8A* is interesting since it is the most differentially expressed and down-regulated gene on chromosome 1. This gene encodes ribosomal protein 8A and has been found to be down-regulated in multiple myeloma tumors (Carrasco et al., 2006). It also has been documented to be associated with lymph node metastasis in patients with cervical cancer (Kim et al., 2008). In addition, Salicioni et al. (2000) identified the conserved residues in the RBM8 protein family that are likely to contact RNA in a protein-RNA complex and discovered that *RBM8A* interact with gene *OVCA1* which is a candidate for the breast and ovarian tumor suppressor gene. Our study suggests that this gene may play an important role in causing relapse of multiple myeloma patients.

The second and third most differentially expressed and up-regulated ones among the 14 genes are *CRB1* and *KIAA1383* both residing on segment 2. Research has demonstrated that *CRB1* is susceptible to mutations and alternative splicings that are directly associated with various diseases. For example, Mehalow et al. (2003) discovered that mutations within this gene causes human retinal diseases including retinitis pigmentosa and Leber's congenital amaurosis. Interestingly, *KIAA1383* appears to be the genomic contig of multiple myeloma tumor-associated protein 2. Gene *PCANAP6* that is ranked the fifteenth and resides on segment 1 was known to regulate prostate cancer associated protein 6. Gene *PADI1* residing on segment 1 encodes a member of the peptidyl arginine deiminase family of enzymes and was shown to be associated with formation of oral squamous cell carcinoma (Chen et al., 2008). *LOC54499* encodes Putative membrane protein and its up-regulation in multiple myeloma was supported by Largo et al. (2006). Gene *C1orf25* was also documented to be up-regulated in prostate cancer. The ranked eighth gene *ARHGEF2* is a Rho/Rac guanine nucleotide exchange factor and was known to play an important role in tumor cell invasion and cancer metastasis (Lu et al., 2006). The other gene *SDHC* that encodes proteins involved in energy production pathways has been discovered to have increased expression in multiple myeloma in another study Fabris et al. (2007). Gene *IRTA1* encoding cell surface receptors homologous to the Fc was also documented to be up-regulated in myeloma cell lines. In

addition, gene *NCSTN*, encoding a Type I transmembrane glycoprotein that is an integral component of the multimeric gamma-secretase complex, has been shown to be associated with hepatocellular carcinoma progression.

In summary, most of the detected genes are either directly associated with multiple myeloma or other cancers. This finding is important because the past research (Tonon et al., 2005) has suggested that multiple myeloma shares common mechanisms of disease pathogenesis with other unrelated cancers. Further laboratory based validation procedures are desired in identifying the roles of these gene candidates in causing relapse in multiple myeloma patients.

5. Conclusion

We have proposed MSS, a method based on model selection to detect regions of DNA copy-number variation that are associated with a phenotype. We propose modeling the original data in order to conveniently deal with a wide range of phenotypes, such as multiple groups or continuous variables. The proposed sequential procedure also enables us to accumulatively borrow information across contiguous clones to improve detection accuracy. The weighting scheme we have adopted in the objective function takes into consideration the unequal clone-wise variation. The proposed method is also more computationally efficient because it requires iterations only in the order of the number of clones, and does not require time-consuming permutation procedures. Our empirical studies indicate that the proposed method has superior performance in terms of detecting small segments and neighboring segments with differential degrees of CNVs.

In this paper, we focus on detecting the common CNV regions for samples in one disease group. In practice, subject-specific CNVs may exist for individual samples (arrays) due to population heterogeneity, which would be random somatic events without pathological relevance (Shah, 2008). For instance, only a small proportion of samples have CNVs in one segment, or the change points have small subject-specific shifts. To accommodate the first scenario, we may extend our proposed method by adopting the quantile loss function in Koenker (2005) to identify changes in either the upper or lower quantiles of y_{ij} across subjects in one group. To accommodate the second scenario, we can first apply the proposed method to identify common segments, and then search in the nearby few clones for each subject to identify subject-specific change points; this approach is feasible in cases with few change points or small number of subjects. This interesting research direction certainly deserves further investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the referee, the associate editor, and the editor for their constructive comments and suggestions which have led to significant improvement of the article. This research is partially supported by the National Science Foundation through grants DMS-0706818, DMS-0706963 and DMS-1149355, and by the National Institutes of Health through grants R01 RGM080503A, R21CA129671, and NCI CA97007.

References

- Ahn T, Lee E, Huh N, Park T. Personalized identification of altered pathways in cancer using accumulated normal tissue data. *Bioinformatics*. 2014; 30:i422–i429. [PubMed: 25161229]
- Bendor, A.; Lipson, D.; Tsalenko, A.; Reimers, M.; Baumbusch, L.; Barrett, M.; Weinstein, J.; BorresenDale, A.; Yakhini, Z. Framework for identifying common aberrations in DNA copy number data[C]//Research in Computational Molecular Biology. Springer; Berlin Heidelberg: 2007. p. 122-136.
- Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*. 2003; 19:185–193. [PubMed: 12538238]
- Carrasco DR, Tonon G, Huang Y, Zhang Y, Sinha R, Feng B, Stewart JP, Zhan F, Khatry D, Protopopova M, Protopopov A, Sukhdeo K, Hanamura I, Stephens O, Barlogie B, Anderson KC, Chin L, Shaughnessy JD, Brennan C, Depinho RA. High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. *Cancer cell*. 2006; 9:313–325. [PubMed: 16616336]
- Chen C, Mendez E, Houck J, Fan W, Lohavanichbutr P, Doody D, Yueh B, Futran ND, Upton M, Farwell DG, Schwartz SM, Zhao LP. Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer Epidemiological Biomarkers Prevention*. 2008; 17:2152–2162.
- Efron B, Zhang NR. False discovery rates and copy number variation. *Biometrika*. 2011; 98:251–271.
- Duan J, Zhang JG, Deng HW, Wang YP. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PloS one*. 2013; 8:e59128. [PubMed: 23527109]
- Fabris S, Ronchetti D, Agnelli L, Baldini L, Morabito F, Biciato S, Basso D, Todoerti K, Lombardi L, Lambertenghi-Deliliers G, Neri A. Transcriptional features of multiple myeloma patients with chromosome 1q gain. *Leukemia*. 2007; 21:1113–1116. [PubMed: 17315022]
- Guha S, Li Y, Neuberger D. Bayesian Hidden Markov Modeling of Array CGH Data. *Journal of the American Statistical Association*. 2008; 103:485–497. [PubMed: 22375091]
- Huang J, Gusnanto A, O'Sullivan K, Staaf J, Borg A, Pawitan Y. Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics*. 2007; 23:2463–2469. [PubMed: 17660206]
- Jeng XJ, Cai TT, Li H. Simultaneous discovery of rare and common segment variants. *Biometrika*. 2013; 100:157–172. [PubMed: 23825436]
- Kim T, Choi J, Kim W, Choi C, Lee J, Bae D, Son D, Kim J, Park B, Ahn G, Cho E, Kim B. Gene expression profiling for the prediction of lymph node metastasis in patients with cervical cancer. *Cancer Science*. 2008; 99:31–38. [PubMed: 17986283]
- Klijn C, Holstege H, de Ridder J, Liu X, Reinders M, Jonkers J, Wessels L. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic acids research*. 2008; 36:e13–e13. [PubMed: 18187509]
- Koenker, R. Quantile regression. New York: Cambridge University Press; 2005.
- Lai TL, Xing H, Zhang N. Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics*. 2008; 9:290–307. [PubMed: 17855472]
- Largo C, Alvarez S, Saez B, Blesa D, Martin-Subero JJ, Gonzalez-Garcia I, Brieva JA, Dopazo J, Siebert R, Calasanz MJ, Cigudosa JC. Identification of overexpressed genes in frequently gained/amplified chromosome regions in multiple myeloma. *Haematologica*. 2006; 91:184–191. [PubMed: 16461302]
- Lu H, Knutson KL, Gad E, Disis ML. The Tumor Antigen Repertoire Identified in Tumor-Bearing Neu Transgenic Mice Predicts Human Tumor Antigens. *Cancer Research*. 2006; 66:9754–9761. [PubMed: 17018635]
- Lu T, Lai L, Tsai M, Chen P, Hsu C, Lee J, Hsiao C, Chuang E. Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. *PloS one*. 2011; 6:e24829. [PubMed: 21935476]
- Lu T, Hsiao C, Lai L, Tsai M, Hsu C, Lee J, Chuang E. Identification of regulatory SNPs associated with genetic modifications in lung adenocarcinoma. *BMC Research Notes*. 2015; 8:92. [PubMed: 25889623]

- Mehalow AK, Kameya S, Smith RS, Hawes NL, Denegre JM, Young JA, Bechtold L, Haider NB, Tepass U, Heckenlively JR, Chang B, Naggert JK, Nishina PM. CRB1 is essential for external limiting membrane integrity and photoreceptor morphogenesis in the mammalian retina. *Human Molecular Genetics*. 2003; 12:2179–2189. [PubMed: 12915475]
- Nishii R. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*. 1984; 12:758–765.
- Niu YS, Zhang H. The screening and ranking algorithm to detect DNA copy number variations. *The Annals of Applied Statistics*. 2012; 6:1306–1326. [PubMed: 24069112]
- 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]
- Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*. 2005; 37(Suppl):S11–7. [PubMed: 15920524]
- Rouveirol C, Stransky N, Hupe P, Rosa PL, Viara E, Barillot E, Radvanyi F. Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*. 2006; 22:849–856. [PubMed: 16434445]
- Rueda OM, Diaz-Uriarte R. Finding Recurrent Copy Number Alteration Regions: A Review of Methods. *Current Bioinformatics*. 2010; 5:1–17.
- Salicioni AM, Xi M, Vanderveer LA, Balsara B, Testa JR, Dunbrack RL Jr, Godwin AK. Identification and structural analysis of human RBM8A and RBM8B: two highly conserved RNA-binding motif proteins that interact with OVCA1, a candidate tumor suppressor. *Genomics*. 2000; 69:54–62. [PubMed: 11013075]
- Schwartz G. Estimating the dimension of a model. *The Annals of Statistics*. 1978; 6:461–464.
- Shah SP. Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenetic and genome research*. 2008; 123:343–351. [PubMed: 19287173]
- Shao J. An asymptotic theory for linear model selection. *Statistica Sinica*. 1997; 7:221–264.
- Shi P, Tsai CL. Regression model selection a residual likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2002; 64:237–252.
- Siegmund DO, Yakir B, Zhang NR. Detecting simultaneous variant intervals in aligned sequences. *The Annals of Applied Statistics*. 2011; 5:645–668.
- Smetana J, Frohlich J, Zaoralova R, Vallova V, Greslikova H, Kupska R, Nemec P, Mikulasova A, Almasi M, Pour L, Adam Z, Sandecka V, Zahradová L, Hajeck R, Kuglik P. Genome-wide screening of cytogenetic abnormalities in multiple myeloma patients using array-CGH technique: a Czech multicenter experience. *BioMed research international*. 2014:209–670.
- 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, Albertson DG. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*. 2001; 29:263–264. [PubMed: 11687795]
- Siegmund D, Yakir B, Zhang NR. Detecting simultaneous variant intervals in aligned sequences. *The Annals of Applied Statistics*. 2011; 5:645–668.
- Tan R, Wang Y, Kleinstein SE, Liu YZ, Zhu XL, Guo HZ, Jiang QH, Allen AS, Zhu MF. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Human Mutation*. 2014; 35:899–907. [PubMed: 24599517]
- Tibshirani R, Wang P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*. 2007; 9:18–29. [PubMed: 17513312]
- Tonon G, Wong KK, Maulik G, Brennan C, Feng B, Zhang Y, Khatry DB, Protopopov A, You MJ, Aguirre AJ, Martin ES, Yang Z, Ji H, Chin L, Depinho RA. High-resolution genomic profiles of human lung cancer. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102:9625–9630. [PubMed: 15983384]
- Wang H, Hu J. Identification of differential aberrations in multiple-sample array CGH studies. *Biometrics*. 2011; 67:353–362. [PubMed: 20618310]
- Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2009; 71:671–683.
- Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*. 2005; 21:4084–4091. [PubMed: 16159913]

- Ylipaa, A.; Nykter, M.; Kivinen, V.; Hu, L.; Cogdell, D.; Hun, K.; Zhang, W.; Yli-Harja, O. Finding common aberrations in array CGH data. Proceedings of 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP 2008); St. Julians, Malta. Mar 2008; 2008. p. 1199-1204.
- Zhang NR, Siegmund DO, Ji H, Li J. Detecting simultaneous change-points in multiple sequences. *Biometrika*. 2010; 97:631–645. [PubMed: 22822250]

Figure 1.

Simulation Case 1: the vertical lines indicate the true change-point locations. The MSS-LTR and the MSS-unweighted starting from the left-most point, the MSS-RTL starting from the right-most point, PASS and MSCBS methods are shown in the upper-left, middle-left, upper-right, middle-right and lower-left panels, respectively. The mean effects of the clones for the tumor and reference groups are indicated by the yellow and green dots in all the panels, respectively. The detected segments for the tumor and reference groups are labeled by red and blue horizontal bars.

Figure 2.

Simulation Case 3: the vertical lines indicate the true change-point locations. The MSS-LTR and the MSS-unweighted starting from the left-most point, the MSS-RTL starting from the right-most point, PASS and MSCBS methods are shown in the upper-left, middle-left, upper-right, middle-right and lower-left panels, respectively. The mean effects of the clones for the tumor and reference groups are indicated by the yellow and green dots in all the panels, respectively. The detected segments for the tumor and reference groups are labeled by red and blue horizontal bars.

Figure 3.

Simulation Case 4: The vertical lines indicate the true change-point locations. The MSS-LTR and the MSS-unweighted starting from the left-most point, the MSS-RTL starting from the right-most point, FAL methods are shown in the upper-left, lower-left, upper-right and lower-right panels, respectively. The mean group difference of the clones are indicated by the green dots in all panels. The detected segments are labeled by red horizontal bars.

Table 1

Segmentation results of different methods in Cases 1–5. The TP and FP are the total number of true and false change points identified, respectively. The MFP is the mean number of false positives. Time is the average computing time in seconds for analyzing one simulated data set.

TP															FP			MFP	Time									
10															25	45	340	360	660	675	695	705	1	2	3	>3		
Case 1																												
MSS-LTR	100	99	100	95	89	100	100	100	100	100	100	100	100	100	15	2	0	0	0.19	10.85								
MSS-RTL	100	100	100	85	87	100	100	100	100	100	100	100	100	100	18	5	0	0	0.28	10.94								
MSS-UN	100	100	100	93	85	100	100	100	100	100	100	100	100	100	18	2	0	0	0.22	11.57								
PASS	99	66	92	28	28	99	85	65	100	35	29	8	2	1.25	98.02													
MSCBS	0	100	100	1	100	100	100	100	100	0	0	0	100	83.43	19.87													
Case 2	10	25	45	340	360	660	675	695	705	1	2	3	>3															
MSS-LTR	100	80	71	55	54	94	95	100	100	41	31	8	1	1.31	6.76													
MSS-RTL	99	74	61	35	33	80	86	100	100	22	36	30	8	2.16	6.78													
MSS-UN	100	58	43	17	14	92	71	100	100	37	30	10	4	1.44	7.33													
PASS	94	71	84	67	70	94	64	62	100	2	10	13	75	4.60	86.93													
MSCBS	2	97	100	8	92	100	99	100	100	0	0	0	100	81.00	15.71													
Case 3	100	110	120	430	450	745	750	760	770	800	1	2	3	>3														
MSS-LRT	100	100	100	100	100	100	100	100	100	3	0	0	0	0.03	11.23													
MSS-RTL	99	100	100	100	100	100	100	100	100	6	0	0	0	0.06	10.93													
MSS-UN	100	100	100	100	100	100	100	100	100	14	2	1	0	0.21	8.64													
PASS	100	68	100	75	75	61	53	100	62	100	17	0	0	0.17	88.50													
MSCBS	100	100	100	100	100	89	54	100	92	100	0	0	0	191.95	41.73													
Case 4	100	110	120	430	450	745	750	760	770	800	1	2	3	>3														
MSS-LTR	100	100	100	100	100	100	99	100	100	19	28	11	7	1.41	6.38													
MSS-RTL	100	100	100	100	100	100	100	100	99	24	28	13	6	1.50	6.04													
MSS-UN	100	100	100	100	100	100	99	100	100	20	25	14	11	1.61	5.47													
FAL	79	100	100	100	100	70	30	93	100	0	0	1	99	10.22	29.20													

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	TP										FP			MFP	Time
	4000	4100	4200	10000	10050	18000	18020	18040	18060	18100	1	2	3		
Case 5													>3		
MSS-LTR	100	100	100	100	100	100	100	100	100	100	3	23	5	53	279.49
MSS-RTL	100	100	100	100	100	100	100	100	100	100	6	28	7	43	280.90
MSS-UN	100	100	100	100	100	100	100	100	100	100	5	8	0	2	720.53
FAL	100	100	100	100	100	100	100	100	100	100	0	0	0	100	522.56

Table 2

Segmentation results of MSS-LTR at different values of penalty parameter c in the BIC criterion.

Case	c	TP						FP			MFP		
		10	25	45	340	360	660	675	695	705	1	2	3
1													
	0.30	100	99	100	97	94	100	100	100	100	17	3	0
	0.23										0	0	0
	0.35	100	99	100	95	89	100	100	100	100	15	2	0
											0	0	0
	0.19												
	0.40	100	99	100	91	84	100	100	100	100	19	4	0
											0	0	0
	0.27												
2													
	0.30	10	25	45	340	360	660	675	695	705	1	2	3
	>3												
	0.30	100	86	80	62	61	97	96	100	100	42	24	6
											3	3	1.21
	0.35	100	80	71	55	54	94	95	100	100	41	31	8
											1	8	1
	0.19												
	0.40	99	74	62	40	43	92	93	100	100	33	30	14
											2	2	1.43
	0.27												
3													
	0.30	100	110	120	430	450	745	750	760	770	800	1	2
	>3												
	0.30	100	100	100	100	100	100	100	100	100	5	1	0
											0	0	0.07
	0.35	100	100	100	100	100	100	100	100	100	3	0	0
											0	0	0.03
	0.40	100	100	100	100	100	100	100	100	100	2	0	0
											0	0	0.02
	0.02												
4													
	0.30	100	110	120	430	450	745	750	760	770	800	1	2
	>3												
	0.30	100	100	100	100	100	100	99	100	100	16	19	20
											31	20	31
	0.35	100	100	100	100	100	100	100	100	100	19	28	11
											7	7	1.41
	0.40	100	100	100	100	100	100	99	100	100	22	10	6
											1	6	1
	0.65												
5													
	0.30	4000	4100	4200	10000	10050	18000	18020	18040	18060	18100	1	2
	>3												
	0.30	100	100	100	100	100	100	100	100	100	1	3	4
											91	3	9.46
	0.35	100	100	100	100	100	100	100	100	100	3	23	5
											53	5	3.55
	0.40	100	100	100	100	100	100	100	100	100	6	30	3
											20	3	1.66
	1.66												

Table 3

Results from the multiple myeloma data analysis. (a): the estimated group effects and p-values from testing the significance of group effects for the four segments detected by the MSS method on chromosome 1; (b): information of genes residing on the three segments of chromosome 1 with p-values $j = 0.005$, where the p-values are from the two-sample t-test for testing the presence of differential gene expression.

(a)			
rank	segment	group effect	p-value
1	[1163, 1662]	0.160	8.72×10^{-172}
2	[914, 1162]	0.148	1.03×10^{-69}
3	[855, 913]	0.190	2.11×10^{-18}
4	[1, 854]	-0.019	8.41×10^{-13}

(b)			
rank	segment	gene name	alteration location
1	4: [1, 854]	<i>RBM8A</i>	- chr1q12
4	2: [914, 1162]	<i>CRB1</i>	+ chr1q31-q32.1
6	2: [914, 1162]	<i>KIAA1383</i>	+ chr1q42.2
8	2: [914, 1162]	<i>ARHGEF2</i>	+ chr1q21-q22
13	2: [914, 1162]	<i>SDHC</i>	+ chr1q21
15	1: [1163, 1662]	<i>PCANAP6</i>	+ chr1q32.1
18	1: [1163, 1662]	<i>FLJ11752</i>	+ chr1q24.2
19	1: [1163, 1662]	<i>BLZF1</i>	- chr1q24
20	2: [914, 1162]	<i>IRTA1</i>	+ chr1q21
21	1: [1163, 1662]	<i>PADI1</i>	+ chr1p36.13
22	1: [1163, 1662]	<i>LOC54499</i>	- chr1q22-q25
23	1: [1163, 1662]	<i>CLorf25</i>	+ chr1q25.2
24	2: [914, 1162]	<i>NCSTN</i>	+ chr1q22-q23
26	2: [914, 1162]	<i>ZFP67</i>	- chr1q22