Published in final edited form as:

Biometrics. 2011 June ; 67(2): 353–362. doi:10.1111/j.1541-0420.2010.01457.x.

Identification Of Differential Aberrations in Multiple-sample Array CGH Studies

Huixia Judy Wang* and

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A

Jianhua Hu

Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA

Summary

Most existing methods for identifying aberrant regions with array CGH data are confined to a single target sample. Focusing on the comparison of multiple samples from two different groups, we develop a new penalized regression approach with a fused adaptive lasso penalty to accommodate the spatial dependence of the clones. The nonrandom aberrant genomic segments are determined by assessing the significance of the differences between neighboring clones and neighboring segments. The algorithm proposed in this article is a first attempt to simultaneously detect the common aberrant regions within each group, and the regions where the two groups differ in copy number changes. The simulation study suggests that the proposed procedure outperforms the commonly used single-sample aberration detection methods for segmentation in terms of both false positives and false negatives. To further assess the value of the proposed method, we analyze a data set from a study that identified the aberrant genomic regions associated with grade subgroups of breast cancer tumors.

Keywords

Array CGH; Change point; Common aberration; Copy number aberration; Fused lasso; Median regression; Segmentation

1. Introduction

Array comparative genomic hybridization (CGH) is a powerful technique for measuring genomic aberrations involving DNA copy number gains and losses (Pinkel et al., 1998; Snijders et al., 2001). Such copy number aberrations (CNAs) can lead to abnormal mRNA transcript levels and result in cellular malfunctions. The detection of CNA regions in a set of tumor samples can help biologists identify genes involved in the genesis and progression of cancer. In a typical CGH experiment, DNA strands from a test and a reference sample are hybridized on the same array consisting of thousands of clones with known genomic locations. After hybridization, the log ratio between the test and the reference intensities is computed on each clone. Because the reference sample is often assumed to have no CNAs, significantly positive log ratios indicate copy number gains, and significantly negative log ratios indicate copy number losses in the test sample.

*wang@stat.ncsu.edu.

6. Supplementary Materials

The proof of Theorem 1 and assumptions A1–A3 referenced in Section 2.2 are given in the Supplementary Materials, and are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

Existing methods for identifying genomic regions with CNAs have focused on the analysis of a single target sample. One popular algorithm is the circular binary segmentation (CBS) method proposed by Olshen et al. (2004). The CBS method uses a permutation test to detect change points by recursively splitting each contiguous segment until no significant splits can be found. Other segmentation methods include hidden Markov modeling (Guha et al., 2008, Fridlyand et al., 2004), a clustering-based method (Wang et al., 2005), a Gaussian-likelihood-based approach (Hupe et al., 2004), and a Bayesian approach (Lai, Xing and Zhang, 2008). Taking into account the spatial dependence of CGH data, Eilers (2003), Huang et al. (2005), and Tibshirani and Wang (2007) employed the penalized least squares method by shrinking the distances between signals at adjacent clone locations. Using similar strategies, Eilers and Menezes (2005), and Li and Zhu (2007) considered smoothing the quantiles of the log ratio data. These approaches focused on visualization rather than change point detection. Willenbrock and Fridlyand (2005), and Lai et al. (2005) considered both sensitivity and specificity when comparing various approaches and determined that the CBS method had superior performance for change point detection.

Little attention has been given to the comparison of samples from multiple groups. Due to population heterogeneity, individual samples may have subject-specific CNAs that would represent “passenger” alterations, that is, random somatic events without pathological relevance (Shah, 2008). Therefore, it is of biological interest to identify genomic regions that are common or recurrent among individuals with the same disease, as they are more likely to contain disease-critical genes. Adopting the definition in Rouveirol et al. (2006), we use “recurrent region” to refer to a sequence of adjacent clones that have common aberrations for samples from the same group. Focusing on one group with multiple samples, some authors proposed applying existing aberration-calling algorithms independently to the individual samples, and then combining the discretized profiles to identify recurrent aberrant regions; see Diskin et al. (2006), Rouveirol et al. (2006), Lipson et al. (2006), BenDor et al. (2007), Ylipaa et al. (2008), and Klijn et al. (2008), among others. Rueda et al. (2009) provided a comprehensive review of methods for detecting recurrent CNA regions under different scenarios. For the comparison of two groups, Willenbrock and Fridly (2005), and Huang et al. (2007) suggested another two-step procedure whereby multiple samples are first summarized into a single index at each clone location (e.g., by the clone-wise group mean differences), and then segmentation is performed on the summarized indices. Though straightforward and easy to implement, these two-step procedures do not take into account the variations within and across clones at the same time, and may lose accuracy and sensitivity in detecting aberrant regions of interest.

In this paper, we develop a unified approach to identify regions with differential CNAs between groups as well as regions with common CNAs within each group. The proposed approach models the CGH data from multiple samples directly, and thus is able to account for both within-clone and between-clone variations. We assume that the CNAs are consistent within groups but potentially different between groups. In addition, as the underlying biological process is discrete, we model the CGH data from the same group as a series of discrete segments with unknown boundaries. We propose to smooth the multiple arrays by fitting a median regression model with a fused adaptive lasso penalty, which shrinks the baseline and group effects toward piecewise constants. After smoothing, we employ a block bootstrap procedure to detect the change points that divide the genome into segments so that clones within neighboring segments will have distinct effects.

In Section 2, we describe the median smoothing and segmentation procedure, and provide the theoretical properties of the proposed penalized estimators. In Section 3, we conduct a simulation study to compare the performance of the proposed method and some existing methods. In Section 4, we apply the proposed method to data from a breast cancer study

conducted at the National Cancer Institute, the goal of which was to determine copy number aberrations associated with grade 1 and grade 3 breast cancer tumors. We provide some final remarks in Section 5.

2. Proposed method

2.1 Smoothing based on penalized median regression

Let y_{ij} denote the log ratio of the i th array (subject) and the j th clone, where $i = 1, \dots, n$ and $j = 1, \dots, G$. There are a total of $N = nG$ observations. Suppose the first n_1 arrays belong to group 1, the last $n_2 = n - n_1$ arrays belong to group 2, and the clones $(1, \dots, G)$ are ordered by their physical locations on a chromosome. We assume the following model:

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

where z_i is the group indicator with $z_i = 0$ for $i = 1, \dots, n_1$, $z_i = 1$ for $i = n_1 + 1, \dots, n$, e_{ij} are the random errors with median zero, μ_j is the baseline effect (i.e., the median of group 1) and β_j is the group effect on clone j .

For CGH data, because of the physical dependence of neighboring clones, the errors e_{ij} tend to be locally correlated among the clones (Huang et al., 2005). The spatial dependence of CGH data is also exhibited in the effects μ_j and β_j in the sense that they tend to be piecewise constants except in the regions where the signals change abruptly. Employing a fused lasso approach similar to that of Tibshirani and Wang (2007), we propose to estimate μ_j and β_j by minimizing

$$\sum_{i=1}^n \sum_{j=1}^G |y_{ij} - \mu_j - \beta_j z_i| + \lambda \sum_{j=2}^G w_{j1} |\mu_j - \mu_{j-1}| + \lambda \sum_{j=2}^G w_{j2} |\beta_j - \beta_{j-1}|, \quad (2)$$

where λ is a nonnegative regularization parameter, and w_{j1} and w_{j2} are the adaptive weights. In (2), we discourage changes in the neighboring clones by penalizing $|\mu_j - \mu_{j-1}|$ and $|\beta_j - \beta_{j-1}|$. The influence of the penalty is controlled by the parameter λ . For $\lambda = 0$, we obtain the ordinary median estimates of μ_j and β_j ; and as $\lambda \rightarrow \infty$, $\mu_j \rightarrow \mu_1$ and $\beta_j \rightarrow \beta_1$ for all $j = 2, \dots, G$. Define $\tilde{\mu}_j$ and $\tilde{\beta}_j$ as some initial consistent estimators of μ_j and β_j . Throughout our numerical studies, we take $\tilde{\mu}_j$ and $\tilde{\beta}_j$ as the ordinary median estimates obtained with $\lambda = 0$. We set $w_{j1} = \max\{|\tilde{\mu}_j - \tilde{\mu}_{j-1}|, 10^{-5}\}^{-1}$ and $w_{j2} = \max\{|\tilde{\beta}_j - \tilde{\beta}_{j-1}|, 10^{-5}\}^{-1}$ for $j = 2, \dots, G$. Instead of applying the same penalty to all clones, the adaptive lasso assigns larger penalties to the adjacent clones that have similar expressions; hence the differences of the effects at these clones are shrunk more toward zero. Numerical studies in Section 3 suggest that the fused adaptive lasso leads to more accurate segmentation and fewer false discoveries of CNA regions, when compared to the counterpart with unadaptive weights.

The penalization parameter λ can be selected by minimizing the modified Schwartz information criterion (SIC; Schwarz, 1978; Koenker et al., 1994):

$$\text{SIC}(\lambda) = \log \left(\sum_{ij} |y_{ij} - \hat{\mu}_j - \hat{\beta}_j z_i| \right) + d \log N / (2N), \quad (3)$$

where $(\hat{\mu}_j, \hat{\beta}_j)$ are the estimated coefficients, and d is a measure of the complexity of the fitted model with the penalization parameter λ . In our implementation, we take

$$d=2+\sum_{j=2}^G\{I(\hat{\mu}_j \neq \hat{\mu}_{j-1})+I(\hat{\beta}_j \neq \hat{\beta}_{j-1})\}.$$

2.2 Asymptotic property

In this section, we establish the oracle property (Fan and Li, 2001) of the fused adaptive lasso estimators $\hat{\mu}_j$ and $\hat{\beta}_j$. We first reparameterize and let $\delta_j = \mu_j - \mu_{j-1}$ and $\Delta_j = \beta_j - \beta_{j-1}$, $j = 1, \dots, G$, where $\mu_0 = \beta_0 = 0$. Let $\theta = (\delta_1, \dots, \delta_G, \Delta_1, \dots, \Delta_G)^T$ denote the $(2G)$ -dimensional

unknown parameter vector. Define the design matrix as $X = \begin{pmatrix} 1_{n_1} & 0 \\ 1_{n_2} & 1_{n_2} \end{pmatrix} \otimes B$, where B is a $G \times G$ lower triangular matrix with 1 on and below the diagonal. Let x_{ij} be the $\{(i-1)G+j\}$ -th row of X . Therefore, (2) is equivalent to

$$\sum_{i=1}^n \sum_{j=1}^G |y_{ij} - x_{ij}^T \theta| + \lambda \sum_{k \in \{1, G+1\}} w_k |\theta_k|, \quad (4)$$

where $w_k = |\tilde{\theta}_k|^{-1}$ with $\tilde{\theta}_k$ as some initial consistent estimator of θ . Denote $\theta_{k,0}$ as the true coefficients, $\mathcal{A} = \{k: \theta_{k,0} \neq 0, k = 1, \dots, 2G\}$. We write $d_{\mathcal{A}} = (d_k, k \in \mathcal{A})^T$ for any vector $d = (d_1, \dots, d_{2G})^T$, and $C_{\mathcal{A}\mathcal{A}}$ as the sub-matrix of C with both row and column indices in $\mathcal{A}$ for any $(2G) \times (2G)$ matrix C . The fused adaptive lasso estimator $\hat{\theta}$ enjoys the following oracle property.

Theorem 1—Suppose assumptions A1–A3 as spelled out in the Supplementary Materials hold, as $n \rightarrow \infty$, we have

1. consistency in selection: $Pr(\{k: \hat{\theta}_k \neq 0, k = 1, \dots, 2G\} = \mathcal{A}) \rightarrow 1$;
2. asymptotic normality: $\sqrt{n}(\hat{\theta}_{\mathcal{A}} - \theta_{\mathcal{A},0}) \rightarrow N(0, \Sigma)$,

where

$$\Sigma = \lim_{n \rightarrow \infty} H_{\mathcal{A}\mathcal{A}}^{-1} V_{\mathcal{A}\mathcal{A}} H_{\mathcal{A}\mathcal{A}}^{-1}, \quad H_{\mathcal{A}\mathcal{A}} = n^{-1} \sum_{ij} x_{ij\mathcal{A}} x_{ij\mathcal{A}}^T f_{ij}(0), \quad V_{\mathcal{A}\mathcal{A}} = (4n)^{-1} \sum_{ij} x_{ij\mathcal{A}} x_{ij\mathcal{A}}^T + n^{-1} \sum_{i,j \neq j'} x_{ij\mathcal{A}} x_{ij'\mathcal{A}}^T \{Pr(e_{ij} < 0, e_{ij'} < 0)\},$$

and f_{ij} is the density function of e_{ij} .

2.3 Determination of change points and genomic segments

We first describe the procedure for identifying the genomic segments associated with the group effect. Note that in (4), a nontrivial Δ_j corresponds to a location at which the group effect is changed. To assess the significance of each Δ_j for $j = 2, \dots, G$, we employ the following block bootstrap procedure.

1. Given the observed data $\{y_{ij}\}$, compute the fused adaptive lasso estimates $\hat{\mu}_j$ and $\hat{\beta}_j$ by minimizing (2), and obtain the centered data $\tilde{y}_{ij} = y_{ij} - \hat{\mu}_j - \hat{\beta}_j z_i$.
2. Resample the centered data $\tilde{y}_{ij}$ with replacements by treating each sample as a unit to obtain the bootstrap data y_{ij}^* .
3. Fit the penalized median regression with the bootstrap data to obtain the fused adaptive lasso estimates $\hat{\beta}_j^*$ and thus $\Delta_j^* = \hat{\beta}_j^* - \hat{\beta}_{j-1}^*$, $j = 2, \dots, G$. To save the computational cost, we fix λ at the optimal value chosen by SIC based on the original data.

4. Repeat (2) and (3) K times and obtain a set of bootstrap estimates $\{\Delta_{jk}^*, k = 1, \dots, K\}$.

We calculate the p -values for the observed $\hat{\Delta}_j$ by $K^{-1} \sum_{k=1}^K I(|\Delta_{jk}^*| \geq |\hat{\Delta}_j|)$, $j = 2, \dots, G$. We declare that a clone j is a change point associated with the group effect if its corresponding p -value is smaller than p^* . For a fixed p -value cutoff p^* , we follow the approach of Fan et al. (2004) and Huang et al. (2005) to estimate the false discovery rate (FDR) by

$$\widehat{\text{FDR}} = \frac{p^* \times \text{total number of clones}}{\text{number of clones whose } p\text{-value are less than } p^*}.$$

The p -value cutoff p^* is chosen so that the FDR is controlled at a desirable level α , say, 5%. Suppose that the p -values are smaller than p^* at clones $c_1, \dots, c_{m-1}$, then m segments will be formed by clones $[1, c_1 - 1]$, $[c_1, c_2 - 1]$, $\dots$, $[c_{m-1}, c_G]$, so that the clones within each segment will have the same group effect.

The next step is to test the significance of the group effects within each segment in order to identify the regions where two groups differ in CNAs. For each identified segment, we apply the rank score test of Wang and He (2007), which is robust and able to account for the correlation among clones within the same subject. Herein we assume a common intra-subject correlation between any two clones from one segment. In our context, the rank score test is based on the signs of residuals obtained by fitting the median regression model under the null hypothesis of no group effects.

Following a similar procedure, we can also identify the segments associated with the baseline effects μ_j , that is, the segments that are common to samples from group 1. The common segments for group 2 can then be formed by the change points that are associated with either group 1 or the group effect. For instance, suppose group 1 has two segments formed by clones $[1, 20]$ and $[21, 50]$, and the group effect is associated with two segments formed by clones $[1, 25]$ and $[26, 50]$. Group 2 will then have three segments formed by clones $[1, 20]$, $[21, 25]$ and $[26, 50]$. We can then apply the rank score test on each segment to assess the significance of the intercept effects, and to identify the group-specific regions that have significant copy number aberrations.

2.4 Computation

Hereafter, we will refer to the proposed algorithm based on the fused adaptive lasso penalization as the FAL method. The procedure is very easy to implement with commonly used software. Given a fixed tuning parameter λ , the minimization of (2) can be formulated as a linear programming problem and solved efficiently. Denote $Y = (y_{ijk})$ as the response vector, and $\tilde{Y} = (Y^T, 0_{2G-2}^T)^T$ as the augmented response vector, where 0_n denotes an n -vector of zeros. Appending the penalty term in (2) to the design matrix, we obtain the augmented design matrix

$$\tilde{X} = \begin{pmatrix} 1_{n_1} \otimes I_G & 0 \\ 1_{n_2} \otimes I_G & 1_{n_2} \otimes I_G \\ \lambda W_1 D & 0 \\ 0 & \lambda W_2 D \end{pmatrix}, \quad \text{where } D = \text{diff}(I_G) = \begin{pmatrix} 1 & -1 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix},$$

$W_1 = \text{diag}(w_{21}, \dots, w_{G1})$, $W_2 = \text{diag}(w_{22}, \dots, w_{G2})$, and I_G is the $G \times G$ identity matrix. For a given λ , the penalized estimator of $(\mu_1, \dots, \mu_G, \beta_1, \dots, \beta_G)^T$ can be obtained by regressing $\tilde{Y}$ on $\tilde{X}$ at the median. The dimension of the design matrix may appear daunting, but $\tilde{X}$ is very sparse, consisting of only $\{(n_1 + 2n_2 + 4)G - 4\}$ nonzero elements. We fit the median regression model using the sparse Frisch-Newton interior-point algorithm, implemented by the “rq.fit.sfn” function in the R package *quantreg*. The algorithm uses the sparse matrices adopted from Koenker and Ng (2003). The computation time is roughly proportional to the number of nonzero elements in the design matrix (Koenker and Ng, 2005). The computation of our implemented method is quite fast. For example, using R (version 2.8.1) in a PC with 4.0 GB of RAM, the algorithm took 88 seconds on average to complete the computation for one simulated data set from Case 2 in Section 3 with $n_0 = n_1 = 10$ and $G = 1000$, including tuning parameter selection and segmentation based on 500 bootstrap samples.

The proposed median-based procedure can also be extended to a general quantile level τ to identify regions with CNA in the tails of the intensity distribution. Such CNA regions, for instance those shared by 25% of a group of patients, are of biological interest in practice. For FAL at a general quantile level $0 < \tau < 1$, we need to replace the L_1 loss function $|u|$ in the first term of (2)–(4) by the quantile check function $\rho_\tau(u) = u\{\tau - I(u < 0)\}$. The optimization can still be formulated as a linear programming problem, and thus be solved by using the sparse interior point algorithm. The real data example in Section 4 indicates that the study of tails can help identify interesting CNA regions that may be smoothed away at the median.

3. Simulation study

Our simulation mimics the squamous cell carcinoma (SCC) data set in Snijders et al. (2001), which contains the log-ratio intensities of 14 TP53-mutant and 61 wild-type samples of primary squamous cell carcinomas for 1979 clones. The scientific goal is to identify the genomic regions with CNAs that are associated with the TP53 mutation. In this simulation study, we assume there are two groups, each consisting of 10 arrays. The log ratio data of G equally spaced clones are generated from the model (1). To generate e_{ij} , we first obtain the estimated residuals $\hat{e}_{ij}$ by subtracting the clone-wise median effects of each group from the log-ratio intensities y_{ij} of the SCC data. Then we randomly select G clone locations, and take 10 samples from each of the two groups. Finally, the residuals from the 20 samples are shuffled at each clone location to make sure that they carry no group-specific information.

To assess the performance of FAL in both small and large data sets, we consider two cases with $G = 200$ in Case 1 and $G = 1000$ in Case 2. Case 1 contains 10 contiguous segments including 5 with nonzero group effects. Case 2 contains 11 segments, 6 of which have nonzero group effects. The specific choices of μ_j and β_j in these two cases are given in Table 1. In real tumor samples, the CNA segments usually have different starting/ending points in different tumor samples. Therefore, we randomly perturb the data in Case 2 such that the starting/ending points of segment 5 are shifted by up to 3 clones in all the samples from group 2. For each case, we conduct 100 simulations.

We compare the following methods: the proposed fused adaptive lasso method (FAL); the fused lasso method without adaptive weights (FL); the circular binary segmentation (CBS) method by Olshen et al. (2004); the smooth segmentation method (SMS) by Huang (2007), and the fused quantile regression approach (LZ) of Li and Zhu (2007). To implement SMS, we use the function “FDRcgh” in the R package *smoothseg* with default options based on Wilcoxon test. For CBS and LZ that are designed for single arrays, segmentation is performed on the summarized indices, the clone-wise median differences between the two groups. For both FAL and FL, we focus the analysis on the median.

We first look at the Receiver Operator Characteristic (ROC) curves to compare the sensitivity and specificity of different methods for detecting the change points. There are a total of 9 change points for Case 1, and 10 for Case 2. Figure 1 shows the number of true positives (TP) against the number of false positives (FP), averaged over 100 simulations. The ROC curves of FAL, FL, LZ and CBS are constructed by varying the p -value cutoffs, where the p -values are obtained from bootstrap for the first three methods. Since SMS can not be directly used for change point detection, we calculate the differences of the smoothed clone-wise Wilcoxon test statistic values at adjacent clones, and identify the clones with differences exceeding a given threshold value as change points. The ROC curve of SMS is obtained by varying the test statistic threshold value. Figure 1 suggests that the SMS method performs worse than all the other methods at the beginning, but then catches up when threshold is reduced ($FP \geq 50$). In general, the FL method yields smaller bootstrap p -values and thus detects more change points than FAL with the same p -value cutoff. For instance, with a small p -value cutoff of 10^{-7} , on average FL yields 7.7 TPs and 10 FPs, while FAL yields 6.3 TPs and 3.6 FPs in Case 1. Therefore, for case 1 with a smaller number of clones, FL is more sensitive and it can detect more change points without losing much on the false positives. However, in Case 2 with $G = 1000$, the loss of FL on the false positives is more pronounced and it gives much worse ROC curves than FAL. Considering both cases, FAL is the most favorable method for detecting change points.

Next we assess the performance of different methods on segmentation. The bottom panel of Table 1 summarizes the frequencies of different methods for successfully identifying each segment out of the 100 simulations. We count it as a successful segmentation if the starting and end locations of an estimated segment differ from the true segment by less than 2 clone locations. Since SMS smoothes out the segments, we consider two adjacent segments distinct if SMS identifies that one segment has significant group effects but the other not. To perform segmentation with the LZ method, we use bootstrap to assess the significance of jumps at contiguous clones, and estimate FDR following the same procedure as described in Section 2.3. The number of bootstrap samples is set as 500 for FAL, FL and LZ. As the CBS method does not account for the multiple test problems, we choose a p -value threshold of $0.05/G$ for the individual test. The FDR threshold is set as 0.05 for the other four methods. Segmentation is more challenging in Case 2, as it involves relatively shorter segments (for instance, segment 7 with only 5 clones). For visual presentation, Figures 2–3 depict the segmentation results of two typical data sets from Cases 1 and 2, respectively. In each plot, the vertical lines separate the true 10 segments. In plots (a)–(c), the horizontal bars are the estimated segment-wise group effects of the segments (marked by brackets []) identified by FAL, FL and CBS, respectively, and the thicker horizontal bars in (a)–(b) correspond to the detected segments with significant group effects. We point out that the CBS method does not provide any results on the significance of the group effects. Plot (d) shows the estimated clone-wise group effects from the LZ method. Plot (e) shows the segmentation results from the SMS method with the bigger solid dots corresponding to the detected clones that have significant group effects.

Table 1 suggests that CBS has difficulty identifying short segments such as segments 7–10 in Cases 1–2, and segments 2–3 in Case 2. Because of its continuous segment assumption, SMS cannot distinguish neighboring segments with group effects in the same direction but with different degrees, for instance, segments 2 and 3 in Cases 1–2, and segments 9 and 10 in Case 2; see Figure 3(e). By using summarized indices instead of data from multiple samples, LZ loses sensitivity for detecting aberrant segments in both cases. In general, FL performs inferior to FAL, and it generates more false short segments. For the example data set shown in Figure 2(b), FL mistakenly divides segment 2 into three distinct segments. Compared with the other four methods, FAL is more successful in identifying the true segments and the segments with nonzero group effects, with only slight shifts of the change

point locations. In addition, FAL is more powerful in separating neighboring segments that have different, especially low-level, chromosomal aberrations (for instance, segments 7–9). Recall that in Case 2, the underlying region of segment 5 is shifted by a few clones in some samples. Results from Table 1 indicates that all methods are quite robust against such slight location shifts.

4. Empirical Data Analysis

We apply FAL to a breast cancer data to assess its practical performance. The experiment was conducted by the National Cancer Institute, enrolling invasive breast cancer patients with ductal carcinoma in situ. The raw data sets can be downloaded from the National Center for Biotechnology Information (accession no. GSE7882). For array CGH, amplified Promega Normal Male DNA (Promega, WI, USA) was used as the reference sample. Each CGH array contains measurements for 36,288 clones on 24 chromosomes. More information on the data set was provided by Balleine et al. (2008). The data set consists of 13 cases with grade 1, and 10 cases with grade 3 invasive cancer. The histopathologic grade has been widely accepted as a powerful indicator of prognosis in breast cancer, and it was shown to be associated with survival times (Elston and Ellis, 1991). In this section, we identify the genomic regions with differential CNAs between these two grade subgroups.

By controlling FDR at level 0.05, FAL at the median identifies segments with differential CNAs on 4 chromosomes: 8, 11, 16 and 17. More specifically, grade 1 tumors have more gains at chromosome arms 11p15 and 17q21-q23, while grade 3 tumors involve more frequent losses at 8p23-p21, 11q22-q25, and 17p13-p12. Hereafter, we focus on the study of chromosomes 8 and 17, which were shown to contain particularly influential regions that distinguish between grade 1 and grade 3 associated lesions in Balleine et al. (2008). As a comparison, we also apply the CBS method, and the FAL method at the lower and upper quartiles $\tau = 0.25$ and $\tau = 0.75$. The CBS method is based on the differences of the clone-wise medians of the two subgroups, and the p -value cutoff is chosen as $0.05/G$ for CBS, where $G = 1157$ for chromosome 8 and $G = 1890$ for chromosome 17. We didn't include the LZ and SMS methods as both were designed for smoothing rather than for segmentation.

Figure 4 shows the genomic profiles of chromosomes 8 and 17. The top row shows the grade-associated segments identified by CBS, and the bottom three rows show the segments identified by FAL at three quartiles, respectively, where the thicker horizontal bars are for the segments with significant grade-associated effects. The circle points correspond to the clone-wise median differences between the two subgroups in the top two rows, and clone-wise first and third quartile differences in the bottom two rows.

On chromosome 8, FAL at the median detects five segments that are formed by clone locations [1,91], [126,155], [156, 165], [240, 255] and [256, 280]. However, the CBS method fails to separate segments 2 and 3, and segments 4 and 5 when they have differential CNAs in the same direction but with different degrees, which agrees with the observations from the simulation study. More specifically, CBS classifies clones 240–345 to be on the same segment. For further examination, we perform the rank score test on regions [240, 255], [256, 280] and [281, 345] separately, and obtain the p -values of 0.002, 0.001 and 0.008, respectively. This suggests that only regions [240, 255] and [256, 280] are significantly associated with grade after multiple test adjustment across the 11 segments identified by FAL, and thus are deserving of more attention. By comparing the results from FAL at three quartiles, we notice that region [126, 165] has significantly negative grade effects only at the median. An examination of data from two groups suggests that the distribution of grade 3 tumors from clones in this region is heavily skewed to the right, and the large variation in the right tail makes FAL fail to detect the group effect at the third

quartile. Suppose we define a clone with log-ratio intensity greater than 1 as addition, and less than -1 as deletion, there are a total of 20 additions and 2 deletions in the grade 1 group, while there are 4 additions and 5 deletions in the grade 3 group. The FAL at the first quartile fails to detect the group effect in this region mainly because of the small difference of cases with deletions between the two groups.

For chromosome 17, FAL at the median suggests that the two groups have differential CNAs in segments [44, 80] and [332, 386]. The rank score test gives p -values of 0.002 and 0.001 for these two segments, respectively. Similar to the results on chromosome 8, CBS fails to separate the two segments from their neighbors. Another noteworthy observation is that FAL at the third quartile identifies the segment [1160, 1387] as being associated with a positive group effect. Using arbitrary values 1 and -1 as cutoffs for addition and deletion, we find 44 cases with additions and 226 cases with deletions in the grade 1 group in this segment, while there are 408 additions and 118 deletions in the grade 3 group. This indicates that grade 3 tumors have more frequent gains than grade 1 tumors, and this difference is reflected in the upper quantiles and detected by FAL at the third quartile. However, the two grade groups tend to have no significant differences at either median or the lower quartile. This example suggests that using FAL at tail quantiles can help discover interesting CNA regions that may be overlooked by comparing only the centers of the intensity distributions.

5. Discussion

We have developed a new unified approach for analyzing CGH data with multiple samples from two groups. In contrast to the existing methods based on summarized indices, our approach models the original data, and thus is able to account for variations from both within and across clones. As discussed in Rueda et al. (2009), there are different scenarios of recurrent regions. Our proposed method operates under the paradigm that the recurrent regions are shared by most subjects from the same group. In practice, the change points may vary for different individuals. However, our simulation study suggests that the proposed procedure is robust to random shifts of breakpoints for some subjects. Our empirical studies suggest that the proposed method has superior performance for detecting short segments and those with low-level copy number aberrations, and for distinguishing neighboring segments with differential degrees of copy number aberrations. In addition, analysis at the tail quantiles can help identify interesting regions that might be smoothed away at the median due to population heterogeneity.

In this article, we focused on the comparison of two groups, but the proposed method can be extended to multiple groups by using multiple predictors to index the groups with one of the groups as the reference, in a way similar to ANOVA models. In addition, the fused lasso idea can be extended to analyze CGH data with continuous phenotype to identify genomic segments with dissimilar phenotype effects. The proposed fused estimation was obtained by assuming a working independence structure for residuals from neighboring clones. We showed that the resulting estimation still enjoys the oracle property. Our empirical experience suggested that the within-clone variations may differ at different clone locations in some CGH studies. More efficient estimation might be achieved by incorporating the appropriate weights to account for the intra-subject correlation structure and the heterogeneity in clones (Wang, 2009). Such extensions are beyond the scope of this paper and deserve further study.

Acknowledgments

This research work is partially supported by the National Science Foundation, grants DMS-07-06963 and DMS-0706818, the National Institutes of Health, grants R01RGM080503A, R21CA129671 and the National

Cancer Institute CA97007. The authors would like to thank two anonymous reviewers, the associate editor and the editor for constructive comments and suggestions that helped improve the paper significantly.

References

- Balleine RL, Webster LR, Davis S, Salisbury EL, Palazzo JP, Schwartz GF, Cornfield DB, Walker RL, Byth K, Clarke CL, Meltzer PS. Molecular grading of ductal carcinoma in situ of the breast. *Clinical Cancer Research*. 2008; 14:8244–8252. [PubMed: 19088042]
- 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. *Proceedings of RECOMB '07*. 2007; 4453:122–136.
- Diskin S, Eck T, Greshock J, Mosse Y, Naylor T, Stoeckert C, Weber B, Maris J, Grant G. STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Research*. 2006; 16:1149–1158. [PubMed: 16899652]
- Eilers PHC. A perfect smoother. *Analytical Chemistry*. 2003; 75:3631–3636. [PubMed: 14570219]
- Eilers PHC, Menezes RX. Quantile smoothing of array CGH data. *Bioinformatics*. 2005; 21:1146–1153. [PubMed: 15572474]
- Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*. 1991; 19:403–410. [PubMed: 1757079]
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Fan J, Tam P, Woude GV, Ren Y. Normalization and analysis of cDNA microarrays using withinarray replications applied to neuroblastoma cell response to a cytokine. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101:1135–1140. [PubMed: 14739336]
- Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN. Application of hidden Markov models to the analysis of the array CGH data. *Journal of Multivariate Analysis*. 2004; 90:132–153.
- Guha S, Li Y, Neuberg D. Bayesian hidden Markov modeling of array CGH data. *Journal of the American Statistical Association*. 2008; 103:485–497.
- 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]
- Huang T, Wu B, Lizardi P, Zhao H. Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*. 2005; 21:3811–3817. [PubMed: 16131523]
- Hupe P, Stransky N, Thiery JP, 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]
- 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. [PubMed: 18187509]
- Koenker R, Ng P, Portnoy S. Quantile smoothing splines. *Biometrika*. 1994; 81:673–680.
- Koenker R, Ng P. SparseM: a sparse matrix package for R. *Journal of Statistical Software*. 2003; 8(6)
- Koenker R, Ng P. A Frisch-Newton algorithm for sparse quantile regression. *Journal Acta Mathematicae Applicatae Sinica (English Series)*. 2005; 21:225–236.
- Lai TL, Xing H, Zhang N. Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics*. 2008; 9:290–307. [PubMed: 17855472]
- Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*. 2005; 21:3763–3770. [PubMed: 16081473]
- Li Y, Zhu J. Analysis of array CGH data for cancer studies using fused quantile regression. *Bioinformatics*. 2007; 23:2470–2476. [PubMed: 17644559]
- Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhini Z. Efficient calculation of interval scores for DNA copy number data analysis. *Journal of Computational Biology*. 2006; 13:215–228. [PubMed: 16597236]
- 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, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*. 1998; 20:207–211. [PubMed: 9771718]
- 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. Computation of recurrent minimal genomic alterations from array-CGH data. *Current Bioinformatics*. 2009 in press.
- Schwarz 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]
- 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]
- Tibshirani R, Wang P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*. 2007; 9:18–29. [PubMed: 17513312]
- Wang H. Inference on quantile regression for heteroscedastic mixed models. *Statistica Sinica*. 2009; 19:1247–1261.
- Wang H, He X. Detecting differential expressions in GeneChip microarray studies: a quantile approach. *Journal of the American Statistical Association*. 2007; 102:104–112.
- Wang P, Kim Y, Pollack J, Balasubramanian N, Tibshirani R. A method for calling gains and losses in array CGH data. *Biostatistics*. 2005; 6:45–58. [PubMed: 15618527]
- 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 the 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP 2008)*; St. Julians, Malta. Mar 2008; 2008. p. 1199-1204.

Figure 1. The ROC curves of different methods for detecting the change points associated with the group effect in Cases 1–2. The TP and FP are the total number of true and false change points identified, respectively. The ideal TP is 9 for Case 1 and 10 for Case 2.

Figure 2.

The segmentation results of different methods for a data set in Case 1. The true segments are separated by the vertical lines, and the segments identified by FAL, FL and CBS are marked by brackets [] in (a)–(c).

Figure 3.

The segmentation results of different methods for a data set in Case 2. The true segments are separated by the vertical lines, and the segments identified by FAL, FL and CBS are marked by brackets [] in (a)–(c).

Figure 4.

The segmentation results from CBS and FAL at the three quartiles on chromosomes 8 and 17. The top two are for median, and the bottom two are for the first and the third quartiles, respectively. The horizontal bars are the estimated segment-wise group effects. The thicker horizontal bars indicate segments with significant grade effects.

The top panel shows the simulation design in Cases 1 and 2, where loc.start and loc.end are the locations of the beginning and the end of a segment. The bottom panel summarizes the frequencies of different methods for identifying contiguous segments among the 100 simulations.

Table 1

	Segment <i>j</i>										
	1	2	3	4	5	6	7	8	9	10	11
Simulation design											
Case 1											
loc.start	1	11	26	46	76	101	151	166	181	191	/
loc.end	10	25	45	75	100	150	165	180	190	200	/
μ_j	0	0.25	0.45	0.5	0.5	0	0.43	0.5	1	-1	/
β_j	0	-0.15	-0.25	0	0.1	0	-0.1	0	0.1	0	/
Case 2											
loc.start	1	101	111	121	431	451	746	751	761	771	801
loc.end	100	110	120	430	450	745	750	760	770	800	1000
μ_j	0	0.25	0.45	0.5	0.5	0	0.425	0.5	1	0.5	-1
β_j	0	-0.2	-0.35	0	0.15	0	-0.15	0	0.15	0.245	0
Frequency of correct identification of each segment											
Case 1											
FAL	94	75	82	84	68	49	83	81	61	65	4
FL	88	15	81	31	74	13	90	40	95	89	19
CBS	93	73	76	74	58	25	40	26	32	34	/
LZ	84	30	73	28	66	12	75	35	56	46	/
SMS	10	0	0	33	76	73	74	62	58	78	/
Case 2											
FAL	87	56	68	50	77	19	58	70	54	55	95
FL	24	4	20	12	86	0	0	17	5	10	42
CBS	79	20	25	78	83	14	12	11	39	41	100
LZ	18	0	3	0	83	0	0	15	29	56	7
SMS	1	0	0	41	70	3	3	4	0	0	59