

Genome analysis

Detection of DNA copy number alterations using penalized least squares regression

Tao Huang¹, Baolin Wu⁵, Paul Lizardi^{2,3} and Hongyu Zhao^{1,4,*}

¹Department of Epidemiology and Public Health, ²Department of Pathology, ³Department of Molecular Biochemistry and Biophysics, ⁴Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA and ⁵Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, Minneapolis, MN 55455, USA

Received on June 2, 2005; revised on August 18, 2005; accepted on August 24, 2005

Advance Access publication August 30, 2005

ABSTRACT

Motivation: Genomic DNA copy number alterations are characteristic of many human diseases including cancer. Various techniques and platforms have been proposed to allow researchers to partition the whole genome into segments where copy numbers change between contiguous segments, and subsequently to quantify DNA copy number alterations. In this paper, we incorporate the spatial dependence of DNA copy number data into a regression model and formalize the detection of DNA copy number alterations as a penalized least squares regression problem. In addition, we use a stationary bootstrap approach to estimate the statistical significance and false discovery rate.

Results: The proposed method is studied by simulations and illustrated by an application to an extensively analyzed dataset in the literature. The results show that the proposed method can correctly detect the numbers and locations of the true breakpoints while appropriately controlling the false positives.

Availability: <http://bioinformatics.med.yale.edu/DNACopyNumber>

Contact: hongyu.zhao@yale.edu

Supplementary Information: <http://bioinformatics.med.yale.edu/DNACopyNumber>

1 INTRODUCTION

Genomic DNA copy number alterations are characteristic of many human diseases including cancer, where deletions and amplifications of DNA can contribute to alterations in the expression of tumor-suppressor genes and oncogenes, respectively. Therefore, the identification of DNA copy number alterations is important in understanding the genesis and progression of human cancers (Lengauer *et al.*, 1998). Various techniques and platforms have been developed for genome-wide analysis of DNA copy number, such as comparative genomic hybridization (CGH) (Kallioniemi *et al.*, 1992), array-based comparative genomic hybridization (aCGH) (Pinkel *et al.*, 1998; Snijders *et al.*, 2001), representational difference analysis (RDA) (Lisitsyn *et al.*, 1993) and commercially available single nucleotide polymorphism (SNP) arrays (Zhao *et al.*, 2004).

The goal of the analysis of DNA copy number data is to partition the whole genome into segments where copy numbers change

between contiguous segments, and subsequently to quantify the copy number in each segment. Therefore, identifying the exact locations of copy number changes is fundamentally important to the analysis of DNA copy number data. Many statistical methods have been developed to address this issue. Jong *et al.* (2003) developed a genetical local search algorithm to best localize the breakpoints along the chromosome. Olshen *et al.* (2004) proposed a modified binary segmentation procedure, called circular binary segmentation (CBS), to look for two breakpoints at a time by considering the segment as a circle. Fridlyand *et al.* (2004) used an unsupervised hidden Markov model (HMM) approach to classify each chromosome into different states representing different copy numbers. Wang *et al.* (2005) proposed a hierarchical clustering algorithm, called ‘cluster along chromosomes’ (CLAC), to select interesting clusters by controlling the false discovery rate (FDR, Benjamini and Hochberg, 1995; Storey, 2002). Hsu *et al.* (2005) used a wavelets approach to denoising the data to uncover the true copy number changes. Lai and Zhao (2005) used the *t*-test to detect copy number alterations by aggregating information from replicated arrays. More recently Price *et al.* (2005) applied dynamic programming to search for breakpoints, and Picard *et al.* (2005) further combined dynamic programming with penalized likelihood to identify breakpoints.

In this paper, we propose a novel approach to assess DNA copy number alterations based on the penalized least squares method. Let us consider an array CGH profile, and denote Y_i as the log₂ ratio of the intensities of the red over green channels of marker i on a chromosome where the red and green channels measure the intensities of the cancer and normal samples. We further assume that the observed Y_i is a realization of the true relative copy number β_i at marker i plus a random noise,

$$Y_i = \beta_i + \epsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where n is the number of markers on a given chromosome. Note that the copy number data are ordered by the locations of the markers and have spatial dependence due to the physical dependence of nearby markers. In fact, the spatial dependence of the copy number data is exhibited in both signals β_i and noises ϵ_i . The signals β_i have spatial dependence because the true copy numbers of the nearby markers are the same except in the regions where the copy numbers change abruptly. In Figure 1, we illustrate the spatial dependence of

*To whom correspondence should be addressed.

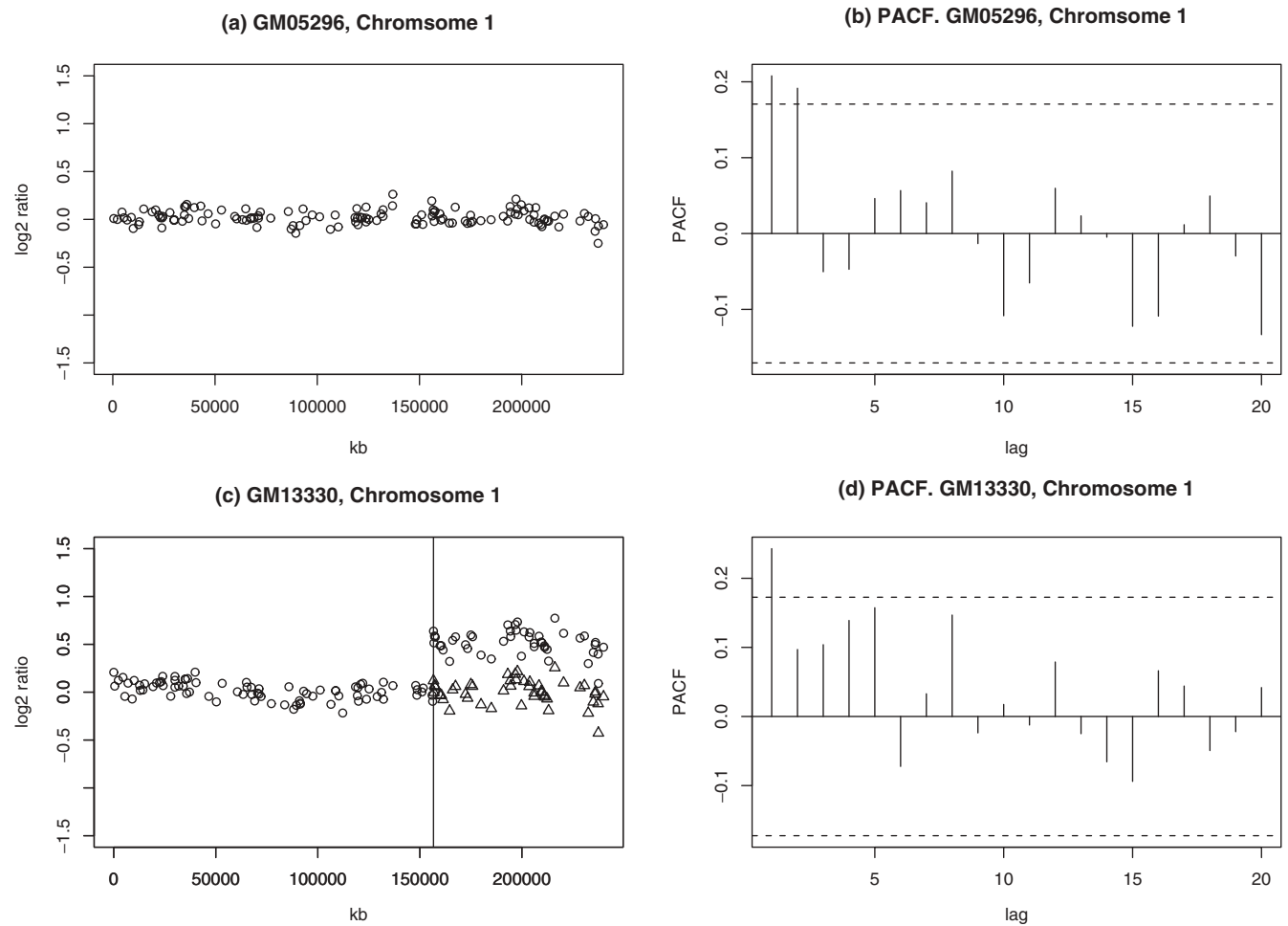


Fig. 1. Spatial dependence of noises in DNA copy number data. GM05296 and GM13330 are two cell lines from the Coreil data set. (a) A normal chromosome 1 without alterations. (b) The corresponding partial auto correlation function of chromosome 1 on GM05296. (c) A chromosome 1 with alterations. Centering the dots (those on the right side of the vertical line) around zero yields the triangles. (d) The partial auto correlation function of the centered data.

noises on chromosome 1 of two cell lines from the Coriel dataset (Snijders *et al.*, 2001). Similar to the analysis of time-series data, we use the partial auto-correlation function (PACF, Brockwell and Davis, 1996) to characterize the spatial dependence, though the markers are not equally spaced along the chromosome. Figure 1(a) and (b) depict a normal chromosome 1 (GM05296) and its corresponding PACF. Figure 1(c) depicts an abnormal chromosome 1 (GM13330). After centering the altered segment (the dots on the right side of the vertical line) to have mean zero, Figure 1(d) depicts the PACF of the centered data. Figure 1(b) and (d) demonstrate the spatial dependence of noise in copy number data. It is desirable and necessary to utilize both types of spatial dependences of the signal and noise in statistical inference.

The remainder of the article is organized as follows. In the Methods section, we incorporate the spatial dependence of signals into model (1) and formalize it as a penalized least squares regression problem. In addition, we consider the spatial dependence of noises and use a stationary bootstrap approach to estimating P -value and FDR. In the Results section, the proposed methods

are evaluated by a simulation study and illustrated by an application to an extensively studied dataset. We summarize the results and discuss future research directions in the Discussion section.

2 METHODS

2.1 Penalized least squares regression

The signals β_i have spatial dependence due to the physical dependence of nearby markers. Intuitively, $\sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i|$ provides a measurement of the smoothness of the parameters β_i , which essentially reflects the spatial dependence of the signals. Hence, we propose to estimate β_i by

$$\arg \min \sum_{i=1}^n (Y_i - \beta_i)^2, \quad \text{subject to } \sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i| < s,$$

where s is a tuning parameter. The global smoothness of the parameters is controlled by the constraint $\sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i| < s$. Set $\beta_0 = 0$, and define $\mu_j = \beta_j - \beta_{j-1}$, which can be interpreted

as the jump between the $(j-1)$ th and j th markers. Then model (1) can be transformed into the following model

$$Y_i = \sum_{j=1}^i \mu_j + \epsilon_i, \quad i = 1, \dots, n. \quad (2)$$

The parameters μ_j can be estimated by

$$\arg \min \sum_{i=1}^n \left(Y_i - \sum_{j=1}^i \mu_j \right)^2, \quad \text{subject to } \sum_{j=2}^n |\mu_j| < s. \quad (3)$$

Problem (3) is a penalized least squares regression with L_1 penalty, and also referred to as Lasso regression (Tibshirani, 1996; Efron *et al.*, 2004) in model selection. Henceforth, we call the proposed method the Lasso-based (LB) method. For a fixed s , the solution for problem (3) can be obtained by using standard quadratic programming. Efron *et al.* (2004) demonstrated that Lasso belongs to a more generalized model selection algorithm, called ‘Least Angle Regression’ (LARS), which is a less aggressive version of forward stepwise regression. LARS exploits the geometry of the algorithm, and requires the same order of computational effort as ordinary least squares. A simple modification of the LARS algorithm calculates the entire path of Lasso solutions (for all values of s), which are piecewise linear functions of s .

The Lasso regression results in a soft thresholding rule which shrinks some coefficients to zero (Donoho and Johnstone, 1994). The tuning parameter s controls the sparsity of the solution. The smaller the s , the more sparse the solution and more parameters are shrunk to zero. Sparsity is desirable from both statistical and biological points of view, which can reduce the complexity of the model and also requires that the copy number of the altered region has to be substantially larger than a threshold. (We note that it is possible to utilize the SCAD penalty proposed by Fan and Li, 2001, Antoniadis and Fan, 2001, which also has the sparsity property, though we do not pursue them here further.)

Choosing s is like choosing the number of breakpoints. Denote A as the active set of breakpoints. As s is varied from 0 to infinity, Lasso algorithm adds or removes one breakpoint at a time from A . Correctly choosing s is crucial because if s is too large, the LB method may detect more false positives; and if the selected s is too small, the LB method may not detect all the true breakpoints. In this paper, we empirically choose s in the following way:

- (1) Starting with all coefficients equal to zero, we have $s=0$ and $A = \emptyset$.
- (2) When one additional covariate (i.e. a breakpoint) is added into the model, the corresponding breakpoint will partition a particular segment into two subsegments. If both of the following conditions (a) and (b) are satisfied, s is updated as the L_1 norm of coefficients of the current model. Otherwise, we stop.
 - (a) The difference of the means of the two subsegments is >0.35 .
 - (b) At least one of the subsegments has a mean >0.35 .

The threshold .35 is chosen because the absolute \log_2 ratio is .35 when the copy number is increased or decreased by about half,

$\log_2(2.55/2) = -\log_2(1.57/2) = 0.35$. In practice, we repeat step 2 up to ten times, because the number of true breakpoints is small and Lasso adds them into A in the first few steps.

In model (2), the parameter μ_j can be interpreted as the jump between the $(j-1)$ th and j th markers. A significant μ_j corresponds to a true breakpoint at which the copy number is changed. In the next section, we propose a stationary bootstrap approach to test the significance of the parameters.

2.2 Estimation of P -value and FDR

There is no parametric testing method for Lasso regression due to the complexity of the algorithm. In addition, we do not specify a parametric distribution of the random noises in model (2). In the absence of parametric tests, permutation and bootstrap are two possible methods to infer the significance of the parameters. However, the presence of the spatial dependence in the DNA copy number data invalidates the exchangeability of the data and the simple global permutation is inappropriate in this context. In this paper, we propose to use the stationary bootstrap method for statistical inferences. We first assume that the Y_i follow the same distribution under the null hypothesis of no copy number alterations. If we further assume that the Y_i follow a weakly dependent stationary process, then we can resample the true null distribution using the stationary bootstrap method proposed by Politis and Romano (1994). The stationary bootstrap method resamples ‘blocks of blocks’ of observations of random length, where the length of each block follows a geometric distribution. More specifically, we first pick one observation randomly from the original observations, say Y_{i_1} . Then with probability θ we pick a new observation randomly from the original observations, say Y_{i_2} . Note that i_1 could be equal to i_2 . Or with probability $1-\theta$ we pick the ‘next’ observation following current observation, and that is equivalent to picking Y_{i_1+1} . Note that the ‘next’ observation following Y_n is Y_1 . This procedure is repeated n times to resample a new set of observations from the original observations. With the above stationary bootstrap method, we propose the following procedure to calculate the P -value and FDR.

- Step 1: Given the original observations $\{Y_i\}_{i=1}^n$, denote $\hat{\mu}_i$ as the Lasso solution to (3). All non-zero $\hat{\mu}_i$ indicate putative breakpoints which partition a chromosome into segments. For each segment, center Y_i around zero by subtracting the average copy number of the segment, and denote the centered data as $\{Y_i^*\}_{i=1}^n$.
- Step 2: Given the centered data $\{Y_i^*\}_{i=1}^n$, resample N sets of observations using the above stationary bootstrap method.
- Step 3: For a set of bootstrapped observations, say the k th, denote $\{\hat{\mu}_{ik}^*\}_{i=1}^n$ as the Lasso solution to (3).

Note that the same s is used in Steps 1 and 3 in order to derive the correct distribution. Given a large number of resamplings, say $N=1000$, the distribution of μ_i under the null hypothesis can be approximated by the marginal distribution of $\{\hat{\mu}_{i1}^*, \dots, \hat{\mu}_{iN}^*\}$. The P -value for the observed $\hat{\mu}_i$ can be estimated by

$$\hat{p}_i = \frac{\#\{|\hat{\mu}_{ik}^*| > |\hat{\mu}_i|, k=1, \dots, N\}}{N} + \frac{\#\{|\hat{\mu}_{ik}^*| = |\hat{\mu}_i|, k=1, \dots, N\}}{2N}, \quad i = 1, \dots, n, \quad (4)$$

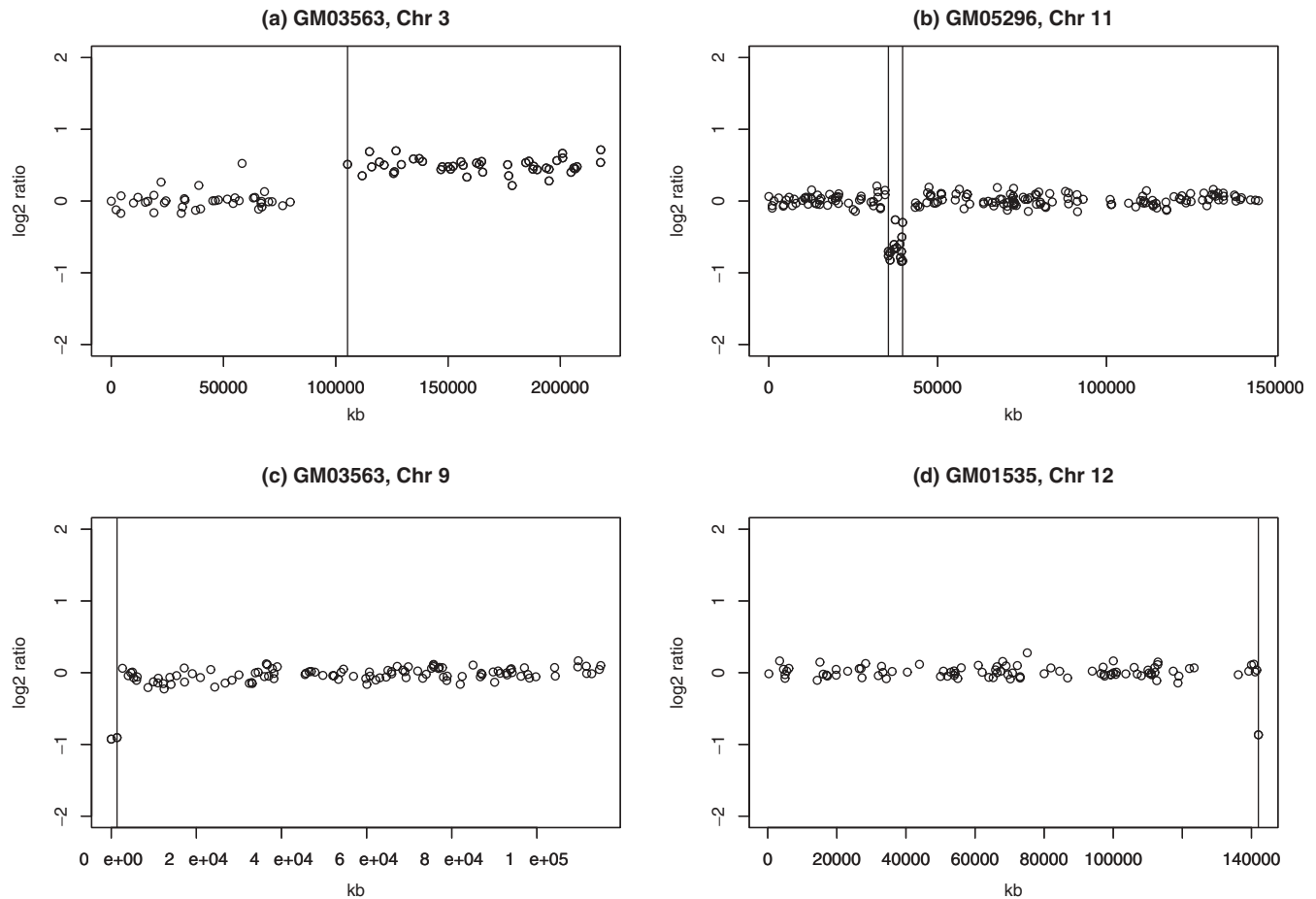


Fig. 2. Examples of the application of the LB method to the Coriel cell lines. (a) contains a copy gain at one side of the chromosome; (b) contains a copy deletion in the middle of the chromosome; (c) and (d) contain only a single or two altered points at the telomeric ends; where HMM defines them as focal aberration, and CBS cannot identify them.

where # represents the number of elements in a set. As for the estimate of FDR, for a given cutoff value, say $p=0.005$, it can be estimated by

$$\widehat{\text{FDR}} = \frac{p \times \text{total number of markers}}{\text{number of markers whose P-values are less than } p}.$$

As pointed out by Fan *et al.* (2004), this FDR estimation method is equivalent to the Benjamini and Hockberg (1995) method with the empirical control of the FDR.

A practical issue of the stationary bootstrap approach is the selection of θ . Note that the stationary bootstrap method becomes the classic bootstrap method (sample with replacement) when the $\{Y_i\}_{i=1}^n$ are independent. Politis and Romano (1994) pointed out that the selection of θ is essentially a ‘smoothing’ problem, and it is difficult to choose θ optimally. For real data analysis and simulations in the next section, we tried a number of values ($\theta=0.05, 0.1, 0.25, 0.35, 0.5$) to test the robustness of the LB method to the selection of θ . The results showed that the LB method is robust to the choice of θ (results not shown). The results presented in the next section were obtained using $\theta=0.25$.

3 RESULTS

3.1 Application to BAC array

To evaluate the LB method, we first apply it to a BAC array dataset¹ with experimentally tested DNA copy number alterations (Snijders *et al.*, 2001). The dataset was also used by Olshen *et al.* (2004), Fridlyand *et al.* (2004), Wang *et al.* (2005), Hsu *et al.* (2005) and others to evaluate their methods. The dataset consists of single experiments on 15 fibroblast cell lines. Each array contains measurements for 2700 BACs spotted in triplicates. There were 15 chromosomes with partial alterations and 8 whole chromosomal alterations. All but one of these alterations were confirmed by spectral karyotyping [Chromosome 15 on GM07801, see Figure 3(h)].

The LB method identified all 14 partial chromosomal alterations confirmed by Snijders *et al.* (2001), and no chromosomal alteration for chromosome 15 on GM07801, given the cutoff P -value 0.005. In comparison with the LB method, HMM of Fridlyand *et al.* (2004) and CBS of Olshen *et al.* (2004) could not detect narrow regions at

¹Download at http://www.nature.com/ng/journal/v29/n3/supinfo/ng754_S1.html

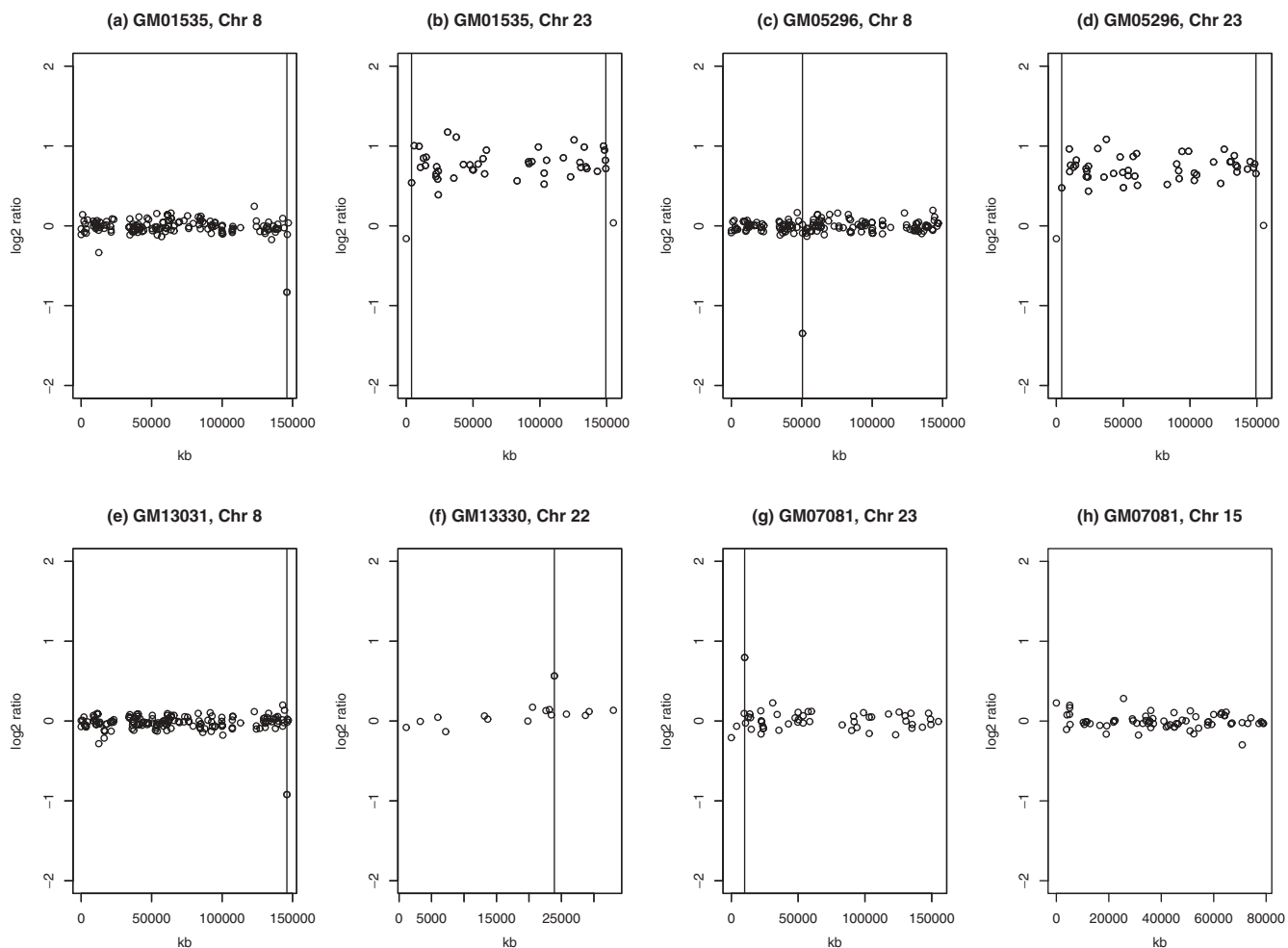


Fig. 3. (a)–(g) show the additionally detected chromosomes with putative alterations. HMM identified the same altered regions except (c). (h) Chromosome 15 on GM07081, which does not contain any altered region confirmed by spectral karyotyping.

the telomeric ends (Chromosome 9 on GM03563 and Chromosome 12 on GM01535). The HMM method defined these regions as focal aberrations and placed them into an abnormal state. Figure 2 illustrates four chromosomes identified by the LB method among the 14 chromosomes. In addition, the LB method found seven other chromosomes with alterations which have not been confirmed by spectral karyotyping. Five of them are single marker alterations. The other two are sex chromosomes 23 (GM01535 and GM07081) which showed whole chromosome alterations except the single marker in the telomeric regions. These additional identifications are summarized in Figure 3 (a)–(g). They could be false positives, or represent real DNA copy number alterations which are undetectable due to the low resolution of spectral karyotyping. In fact, HMM of Fridlyand *et al.* (2004) also detected a number of unconfirmed single-marker aberrations. When we ran HMM on these seven chromosomes, it detected the same breakpoints as the LB method. CBS of Olshen *et al.* (2004) could not detect single-marker aberrations, but still found a number of chromosomes with alterations which were not confirmed by spectral karyotyping. The authors argued that these additional identifications were a result of local trends in the data.

Table 1. Comparison of the performance of three methods for data generated under model (5). For each method, the first row lists the average number of detected breakpoints along with the estimated standard deviations; the second row lists the average number of detected breakpoints within and beyond 2 markers of the true breakpoints, respectively.

Method	1000 markers	500 markers	250 markers
LB	5.30 (0.66) 5.10, 0.20	5.50 (0.89) 5.35, 0.15	5.90 (1.02) 5.65, 0.25
HMM	10.85 (7.28) 4.95, 5.90	7.60 (2.68) 4.95, 2.65	5.75 (1.07) 5.10, 0.65
CBS	40.60 (6.71) 5.00, 35.60	13.95 (4.62) 5.00, 8.95	5.85 (1.14) 5.00, 0.85

3.2 Simulation

In this section, we investigate the performance of the LB method through simulations. Suppose there are 1000 markers equally spaced along a chromosome, under the null hypothesis of no

Table 2. Comparison of the performance of three methods for models (6) and (7).

Method	Independent model number of markers			AR(1) model number of markers		
	1000	500	250	1000	500	250
LB	5.95 (0.89) 5.70, 0.25	6.15 (0.88) 6.05, 0.10	6.20 (1.28) 6.10, 0.10	5.55 (0.89) 5.35, 0.20	6.00 (0.79) 5.75, 0.25	6.10 (1.17) 5.95, 0.15
HMM	6.10 (1.68) 5.15, 0.95	6.05 (1.43) 5.25, 0.80	5.45 (1.23) 5.10, 0.35	11.75 (7.18) 5.25, 6.50	6.50 (2.48) 5.00, 1.50	6.10 (2.71) 5.00, 1.10
CBS	5.00 (0.00) 4.95, 0.05	5.05 (0.22) 5.00, 0.05	5.00 (0.00) 5.00, 0.00	16.60 (6.80) 4.90, 11.70	5.85 (1.50) 5.00, 0.85	5.00 (0.00) 5.00, 0.00

DNA copy number alterations, the log2 ratios of these 1000 markers are simulated from an AR(2) model as follows:

$$\epsilon_i = \alpha_1 \epsilon_{i-1} + \alpha_2 \epsilon_{i-2} + e_i, \quad i = 1, \dots, 1000, \quad (5)$$

where $(\alpha_1, \alpha_2) = (0.6, 0.2)$ and $e_i \sim N(0, 0.1^2)$. We then assume that there are three altered regions along the chromosome which correspond to quadruploid, triploid and monoploid states, respectively. More specifically, the true log2 ratios of 1000 markers are generated as follows:

$$Y_i = \mu_i + \epsilon_i, \quad i = 1, \dots, 1000,$$

where the μ_i are defined in the following table,

i	1–100	101–150	151–450	451–600	601–900	901–1000
μ_i	0	1	0	0.585	0	-1

To consider the density of markers along the chromosome, we create two subsets containing 500 and 250 markers, respectively, by drawing one marker from every two and four markers. We then simulate 20 datasets from the above model, and apply the LB, HMM and CBS methods to estimate the numbers and locations of the breakpoints. For the LB method, we choose the cutoff P -values as 0.001, 0.002 and 0.004, respectively, to control the number of false positives to be at most 1. For the CBS method, we choose the same cutoff P -values, and also use the ‘prune’ procedure to reduce the number of false positives. For the HMM method, we use the Akaike information criterion (AIC) and set the minimum difference of merging two states to be 0.35. Table 1 presents the simulation results.

For each method, the first row lists the average number of detected breakpoints (and the corresponding estimated standard deviation). It shows that the LB method performs quite robustly for different number of markers. CBS tends to detect more false positives when the density of markers becomes higher. In fact, Olshen *et al.* (2004) showed that CBS might detect false positives even when the density of markers and the noise level are low. Though HMM can accurately estimate the number of hidden states, it might not correctly partition the chromosome. It tends to detect more false positives when the signal-to-noise ratio is low. To further evaluate the accuracy of localizing the breakpoints, we also calculate the average number of detected breakpoints within and beyond two markers of the true breakpoints, and list the results in the second row. It shows that the breakpoints detected by the LB method are very close to the true breakpoints. (More details are provided in the Supplementary website.)

Furthermore, we run two other similar simulations to study the dependence assumption on the three methods. More specifically, instead of simulating ϵ_i from an AR(2) model, we generate them either independently

$$\epsilon_i = e_{i0}, \quad i = 1, \dots, 1000, \quad (6)$$

or from an AR(1) model:

$$\epsilon_i = .6\epsilon_{i-1} + e_{i1}, \quad i = 1, \dots, 1000, \quad (7)$$

where $e_{i0} \sim N(0, 0.154^2)$ and $e_{i1} \sim N(0, 0.123^2)$. Note the standard deviation of ϵ_i is the same for these three models. Table 2 summarizes the results, which show that the LB method is quite robust to the dependence assumption. In contrast, the HMM and CBS methods are sensitive to the dependence assumption, and tend to detect more false positives when the dependence between the nearby markers becomes stronger.

4 DISCUSSION

In this paper, we have proposed a new approach, the LB method, to assess DNA copy number alterations along the chromosome. The LB method was applied to an aCGH dataset, and was able to detect all the alterations confirmed by spectral karyotyping. Through simulations we demonstrated that the LB method can correctly infer the numbers and locations of the true breakpoints while appropriately controlling the false positives.

The LB method is conceptually simple and easy to interpret, and may be useful in other genomic problems where the data have spatial dependence structure, such as tiling arrays (Bertone *et al.*, 2004). It showed better performance in comparison with the HMM and CBS methods. The LB method is quite robust when the marker density becomes higher and the spatial dependence of noises becomes stronger, where the HMM and CBS methods tend to detect more false positives. In fact, CBS does not utilize the spatial information and assumes that markers are independent, and HMM has no inference feature and simply partitions the genome based on the hidden state of the markers.

The LB method estimates the parameters through Lasso regression, and performs well for a moderate dataset. When the number of markers is large, we divide the data into a number of overlapping windows of equal size and progress the LB method within each to facilitate the computation of Lasso regression. All the numerical analyses are done using R, and the program can be downloaded at <http://bioinformatics.med.yale.edu/DNACopyNumber>. It took 9 CPU minutes to analyze a chromosome with 1000 markers on a Dell Pentium4 PC.

In the future, we would like to incorporate the distance between markers into model (2), as the current model implicitly assumes that the markers are equally spaced along the chromosome. We would also like to refine the selection of tuning parameter s as well as the dependence parameter θ in the stationary bootstrap in order to reduce the false positives and apply the LB method to even larger datasets, such as 50K and 100K SNP array data. Another important issue to be considered is extending the proposed model to efficiently utilize the replicate information as studied in Lai and Zhao (2005).

ACKNOWLEDGEMENTS

We thank two reviewers for their constructive comments. This work was supported in part by NIH grants GM59507 and CA99135, and NSF grant DMS 0241160.

Conflict of Interest: none declared.

REFERENCES

- Antoniadis, A. and Fan, J. (2001) Regularization of wavelet approximations. *J. Am. Stat. Assoc.*, **96**, 939–967.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.
- Bertone, P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
- Brockwell, P.J. and Davis, R.A. (1996) *Introduction to Time Series and Forecasting*. Springer, NY.
- Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **87**, 425–455.
- Efron, B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.
- Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**, 1348–1360.
- Fan, J. *et al.* (2004) Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Natl Acad. Sci. USA*, **101**, 1135–1140.
- Fridlyand, J. *et al.* (2004) Application of hidden Markov models to the analysis of the array CGH data. *J. Mult. Anal.*, **90**, 132–153.
- Hsu, L. *et al.* (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.
- Jong, K., Marchiori, E., Vaart, A., Ylstra, B., Weiss, M. and Meijer, G. (2003) Chromosomal breakpoint detection in human cancer. In Raidl, G.R., Cagnoni, S., Cardalda, J.J.R., Corne, D.W., Gottlieb, J., Guillot, A., Hart, E., Johnson, C.G., Marchiori, E., Meyer, J.A. and Middendorf, M. (eds), *Applications of Evolutionary Computing: evolutionary computation and bioinformatics*, Springer Verlag, University of Essex, England, UK, **Vol. 2611**, pp. 54–56.
- Kallioniemi, A. *et al.* (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.
- Lai, Y.L. and Zhao, H.Y. (2005) A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data. *Comp. Bio. Chem.*, **29**, 90–98.
- Lengauer, C. *et al.* (1998) Genetic instabilities in human cancers. *Nature*, **396**, 643–649.
- Lisitsyn, N. *et al.* (1993) Cloning the differences between two complex genomes. *Science*, **259**, 946–951.
- Olshen, A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
- Picard, F. *et al.* (2005) A statistical approach for array CGH data analysis. *Bioinformatics*, **6**, 27.
- Pinkel, D. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.
- Politis, D.N. and Romano, J.P. (1994) The stationary bootstrap. *J. Amer. Stat. Assoc.*, **89**, 1303–1313.
- Price, T.S. *et al.* (2005) SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res.*, **33**, 3455–3464.
- Snijders, A.M. *et al.* (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, **29**, 263–264.
- Storey, J. (2002) A direct approach to false discovery rate. *J. Roy. Stat. Soc. B*, **64**, 479–498.
- Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO. *J. Roy. Stat. Soc. B*, **58**, 267–288.
- Wang, P. *et al.* (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.
- Zhao, X.J. *et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism array. *Cancer Res*, **64**, 3060–3071.