Biostatistics (2004), **5**, 4, *pp.* 557–572
doi: 10.1093/biostatistics/kxh008

Circular binary segmentation for the analysis of array-based DNA copy number data

ADAM B. OLSHEN, E. S. VENKATRAMAN
Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, USA
olshena@mskcc.org

ROBERT LUCITO, MICHAEL WIGLER
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

SUMMARY

DNA sequence copy number is the number of copies of DNA at a region of a genome. Cancer progression often involves alterations in DNA copy number. Newly developed microarray technologies enable simultaneous measurement of copy number at thousands of sites in a genome. We have developed a modification of binary segmentation, which we call *circular binary segmentation*, to translate noisy intensity measurements into regions of equal copy number. The method is evaluated by simulation and is demonstrated on cell line data with known copy number alterations and on a breast cancer cell line data set.

Keywords: Array CGH; Binary segmentation; Change-point; ROMA.

1. INTRODUCTION

The DNA copy number of a region of a genome is the number of copies of genomic DNA. In humans the normal copy number is two for all the autosomes. Variations in copy number are common in cancer and other diseases. These variations are a result of genomic events causing discrete gains and losses in contiguous segments of the genome. For this reason, efforts have been made over the last ten years to make whole genome copy number maps from a single study. Technologies to accomplish this have included comparative genomic hybridization (CGH) (Kallioniemi *et al.*, 1992) and representational difference analysis (RDA) (Lisitsyn *et al.*, 1993). In order to increase the resolution of the resulting maps, both techniques have been modified for use with microarrays, the laboratory techniques for which are similar to cDNA gene expression experiments. Each microarray consists of thousands of genomic targets or probes, which we will refer to as *markers*, that are spotted or printed on a glass surface. In a copy number experiment a DNA sample of interest, called the *test sample*, and a diploid *reference* sample are differentially labelled with dyes, typically Cy3 and Cy5, and mixed. This combined sample is then hybridized to the microarray and imaged which results in test and reference intensities for all the markers.

The modification of conventional CGH to obtain high resolution data is called array CGH (aCGH) (Pinkel *et al.*, 1998; Snijders *et al.*, 2001). Here the genomic targets are bacterial artificial chromosomes (BACs), which are large segments of DNA, typically 100–200 kilobases. Representational oligonucleotide microarray analysis (ROMA) (Lucito *et al.*, 2000, 2003) is the high resolution version of RDA. In ROMA, the test and reference samples are based on *representations*, (Lisitsyn *et al.*, 1993), which are subsets of a genome. To create representations, genomic DNA is first shattered using an enzyme. The DNA pieces of

proper size, less than 1.2 kilobases, are then selectively amplified by PCR. Importantly, the positions of shearing and pieces that amplify are the same every time. Typically, a representation contains less than 5% of the original sample. The reduction in complexity of a representation compared to the original sample leads to a reduction in hybridization to the wrong marker, which is termed *cross-hybridization*. Thus the DNA segments on a ROMA array can be much smaller than for other types of copy number arrays. A third technique to estimate copy number is to simply employ the same cDNA arrays used for gene expression studies (Pollack *et al.*, 1999, 2002).

The data from array based copy number experiments are the test and reference sample intensities for each marker. Since we assume that the reference sample does not have any copy number aberrations, markers with normalized test intensities significantly greater than the reference intensities are indicative of copy number gains in the test sample at those positions. Similarly, significantly lower intensities in the test sample are signs of copy number losses. The statistical methods for analyzing copy number data are thus aimed at identifying locations of gains or losses of copy number.

The most common method of analysis for these data is to identify gains and losses using thresholds, such as in Weiss *et al.* (2003). These thresholds are often based on the variability of data from experiments where the test sample and the reference sample are the same normal tissue. Sometimes, the data are first smoothed via local averaging. A variant of the typical analysis can be seen in Pollack *et al.* (2002). Here, the data were smoothed and a statistic was calculated for each marker in normal–normal experiments by averaging over an optimally determined window size. Then, a threshold was determined for gains or losses based on the false discovery rate (Benjamini and Hochberg, 1995). A model-based approach for array copy number data is due to Hodgson *et al.* (2001). They fit a three-component normal mixture model to mouse islet tumor data. In this model, there is one component for 'decreased' copy number, one for 'normal' copy number, and one for 'increased' copy number. Autio *et al.* (2003) developed CGH-Plotter which combines filtering, 3-means clustering and dynamic programming to split CGH data into three groups as above. A promising approach is due to Snijders *et al.* (2003). They developed heuristic methods for fitting a Gaussian hidden Markov model to array copy number data. Linn *et al.* (2003) used a change-point model on RNA expression data to obtain the maximum likelihood estimate of the location of a copy number change, which they then compared to estimates from array DNA copy number data.

The model underlying our work is that gains or losses of copy number are discrete. These aberrations occur in contiguous regions of the chromosome that often cover multiple markers up to whole chromosome arms or chromosomes. In addition, the array copy number data can be noisy, so that some markers will not reflect the true copy number in the test sample. Therefore, we seek a method to split the chromosomes into regions of equal copy number that accounts for the noise in the data. We propose a modification of binary segmentation (Sen and Srivastava, 1975) that we call *circular binary segmentation* (CBS) for this purpose. Our method is novel in that it provides a natural way to segment a chromosome into contiguous regions and bypasses parametric modeling of the data with its use of a permutation reference distribution.

The rest of the paper is organized as follows. In Section 2 we show the relationship between the identification of aberrant genomic regions and change-point problems and introduce the CBS methodology. Results of using the approach of Section 2 to array CGH cell line data with known aberrations are covered in Section 3. In Section 4 results are shown from application to ROMA data from 23 breast cancer cell lines. In Section 5 we study the accuracy of the CBS method via simulation. We summarize our results and discuss future directions in Section 6.

2. CHANGE-POINT METHODS

We will now show the connection between estimating the locations of regions with aberrant DNA copy numbers and the change-point detection problem. This makes change-point methods a natural framework

to approach the analysis of array DNA copy number data. Let $X_1, X_2, \ldots$ be a sequence of random variables. An index ν is called a change-point if $X_1, \ldots, X_\nu$ have a common distribution function F_0 and $X_{\nu+1}, \ldots$ have a different common distribution function F_1 until the next change-point (if one exists). Shaban (1980) and Basseville (1988) provide extensive reviews of change-point problems and methods.

In array copy number studies the data to be analyzed are naturally ordered by the marker location along the chromosome of interest. The data are the test and reference intensities for each marker (denoted I_{tm} and I_{rm} respectively for marker m). These intensities are related to the DNA copy number in the test and the reference samples, C_{tm} and C_{rm}, respectively. This relationship can be modeled as $I_{tm} = \beta_{tm} C_{tm}(1 + \epsilon)$ and $I_{rm} = \beta_{rm} C_{rm}(1 + \epsilon)$, where the parameters ($\beta$) depend on factors such as sample amplification, probe affinity and labeling and the ϵ are random errors. Note that the normalization of data used to correct for the factors above, centers the log ratio of the intensities around zero. This makes the copy numbers unidentifiable without some additional modeling, since, for example, the normalized data from diploid and triploid test samples will appear similar. However, the location of the log ratio of the intensities changes whenever $\beta_{tm} C_{tm}/(\beta_{rm} C_{rm})$ changes and thus the corresponding marker indices are the change-points we want to detect. Since the reference sample is assumed to have no abnormalities and the log ratio of the β is assumed to be constant, all the change-points correspond to changes in the copy numbers of the test sample.

The array data to be used for change-point detection are the log ratio of normalized intensities indexed by the marker locations. Observe that there may be multiple change-points in a given chromosome, each corresponding to a change in the copy number in the test sample. Our goal is to identify all the change-points which will then partition the chromosome into segments where copy numbers are constant. Once the chromosome is partitioned we can estimate the copy numbers of the segments with the help of additional information such as the ploidy of the chromosome. This will provide the locations of copy number aberrations.

Let $X_1, \ldots, X_n$ be the log ratios of the intensities, which are indexed by the locations of the n markers being studied and let $S_i = X_1 + \cdots + X_i$, $1 \leqslant i \leqslant n$, be the partial sums. When the data are normally distributed with a known variance (without loss of generality 1), the likelihood ratio statistic for testing the null hypothesis that there is no change against the alternative that there is exactly one change at an unknown location i (Sen and Srivastava, 1975) is given by $Z_B = \max_{1 \leqslant i < n} |Z_i|$, where

$$Z_i = \{1/i + 1/(n-i)\}^{-1/2} \{S_i/i - (S_n - S_i)/(n-i)\}.$$

The null hypothesis of no change is rejected if the statistic exceeds the upper αth quantile of the null distribution of Z_B and the location of the change-point is estimated to be i such that $Z_B = |Z_i|$. Sen and Srivastava derived the critical value to be used for the test by Monte Carlo simulations. It can be computed quickly using the approximation for the tail probabilities of the test statistic given by Siegmund (1986). The *binary segmentation procedure* applies the test recursively until no more changes are detected in any of the segments obtained from the change-points already found.

The binary segmentation procedure was shown to be consistent under suitable regularity conditions (Vostrikova, 1981). If the variance is unknown the procedure can be extended with a good estimate of it derived from the data. Note that in this case the statistics Z_i is replaced by the corresponding t-statistic and the overall statistic to test for a change is the maximum of these absolute t. Since the binary segmentation procedure is based on a test to detect a single change, a potential problem with it is that it cannot detect a small changed segment buried in the middle of a large segment (Venkatraman, 1992). We propose the following modification of the binary segmentation procedure to address this problem.

This problem with the binary segmentation procedure is due to the fact that it looks for only one change-point at a time. Levin and Kline (1985) proposed a statistic to test for no change against the epidemic or square wave alternative with two change-points. (In the square wave alternative the mean up

to the first change and after the second are assumed to be the same.) If we consider the segment to be spliced at the two ends to form a circle, the likelihood ratio test statistic for testing the hypothesis that the arc from $i+1$ to j and its complement have different means is given by

$$Z_{ij} = \{1/(j-i) + 1/(n-j+i)\}^{-1/2}\{(S_j - S_i)/(j-i) - (S_n - S_j + S_i)/(n-j+i)\}.$$

Our modification of the binary segmentation procedure, which we call *circular binary segmentation* (CBS), is based on the statistic $Z_C = \max_{1 \leqslant i < j \leqslant n} |Z_{ij}|$. Note that Z_C allows for both a single change ($j = n$) and the epidemic alternative ($j < n$). As before, we declare a change if the statistic exceeds an appropriate threshold level based on the null distribution. This critical value when the X_i are normal can again be computed using Monte Carlo simulations or the approximation given by Siegmund (1986) for the tail probability. Once the null hypothesis is rejected the change-point(s) is (are) estimated to be i (and j) such that $Z_C = |Z_{ij}|$ and the procedure is applied recursively to identify all the changes. Other change-point detection schemes such as one based on the Schwartz criterion (Yao, 1988) could also be used for the analysis of array copy number data.

An issue that can arise with the CBS procedure is the edge effect in the estimation of the change-points. That is, if the i and j that correspond to the maximal statistic are such that either i is 'close' to 1 or j is 'close' to n, then there might be only one true change instead of the two changes suggested by the data. We undo a change if the data do not support it as follows. First, we test whether the data support i to be a viable change-point for the segment $X_1, \ldots, X_j$ and undo the change at i if it is not a viable change-point. A similar test is performed for j. Note this is testing for a binary split. Since it is difficult to determine whether a change-point is 'close' to the boundary based just on the values of i and j, we currently perform this test on all change-points derived from ternary splits, that is, splits that result in three different pieces.

The reference distributions used so far were derived using the normality of the data. We can generalize the procedure to non-normal data by generating a reference distribution using a permutation approach as follows. Under the null hypothesis of no change-point in the data, the X_i are identically distributed. Let $X_1^*, \ldots, X_n^*$ be a random permutation of the data and let $Z_C^* = \max |Z_{ij}^*|$ be the statistic derived as above from the permuted data. The threshold value can be chosen to be the upper αth quantile of the permutation distribution given by the Z_C^*. Since the significance level α used for the test is small we need a very large number of permutations (P) for the estimation of p-value (on the order of 10 000). Considerable computational efficiency can be achieved by stopping the permutation procedure once the number of $Z_C^* > Z_C$ exceeds αP. Note that α is the type I error in testing for a change in a single segment with no change-points. Since the procedure tests for changes recursively on all resulting sub-segments the probability of finding spurious change-points is a function of the number of true change-points and could be larger than α. Since the true number of change-points is unknown we do not correct for this multiple testing problem.

The permutation approach is computationally intensive. A modification is sometimes needed for large data sets. Our solution is to divide the data into K overlapping windows $W_k, k = 1, \ldots, K$, of (approximately) equal size and search for change-points within each. The number of windows K depends on the window size and the overlap. The overall test statistic is defined as $Z_C = \max_k Z_k$ where Z_k is the maximum statistic for the data in the window W_k. The permutation process is repeated as above, but with the new, faster maximization procedure.

There are two additional modifications to the basic procedure to make it more appropriate for array DNA copy number data. The first is to smooth outliers before segmenting. Outliers can be caused either by technical errors in an experiment or by aberrant copy number in a region covering only a single marker. The smoothing region for each i is given by $i - R, \ldots, i, \ldots, i + R$, where R is a small integer (say 2 to 5). Let m_i be the median of the data in the smoothing region and let $\hat{\sigma}$ be the standard deviation of the entire data. If the observation X_i is the maximum or the minimum of all the observations in the

smoothing region we find j in the smoothing region closest to it. If the distance from X_i to X_j exceeds $L\hat{\sigma}$ we replace X_i with $m_i + \text{sign}(X_i - X_j)M\hat{\sigma}$. The values we use for L and M are 4 and 2, respectively.

The second modification is because, for reasons that are not totally understood, there are local trends in the data that are not indicative of real copy number changes. This can lead to the identification of change-points that are not biologically meaningful. Therefore, we use a 'pruning' procedure like in CART (Breiman *et al.*, 1984) to eliminate some of them. Suppose there are C change-points after CBS. The sum of squared deviations of data points in segments around their segment average can be represented by $SS(C)$. (This is equivalent to the error sum of square in one way ANOVA.) We then compute $SS(1), \ldots, SS(C-1)$, which are the sum of squares corresponding to the best set of change-points of sizes 1 to $C-1$, choosing only among the change-points previously identified. Then $c' = \min\{c : [SS(c)/SS(C) - 1] < \gamma\}$, where γ is some pre-specified constant (such as 0.05 or 0.10). The change-points are those that led to $SS(c')$.

3. Array CGH example

We applied the CBS methodology with a permutation-based reference distribution to the aCGH data featured in Snijders *et al.* (2001). (These data are freely available for download at `http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html`.) The data consisted of single experiments on 15 fibroblast cell lines. Each array contained 2276 mapped BACs spotted in triplicate. The variable used for analysis was the normalized average of the log base 2 test over reference ratio, as processed by the authors.

There were either one or two alterations in each cell line as identified by spectral karyotyping. Of these, all the alterations for six cell lines covered whole chromosomes and thus would not be identified by our methodology. Therefore, we limited our analysis to the other nine cell lines. For those lines, we tested for change-points one chromosome at a time. As there is a multiple comparison issue from examining 23 chromosomes, we examined our procedure with the α values 0.01 and 0.001. Results can be found in Table 1. The data from a typical cell line experiment, specifically from cell line GM05296, can be seen in Figure 1.

Notably, both α levels lead to identification of the same regions for the chromosomes that were truly altered. Of the 15 altered regions, 12 were found. Of those not found, chromosome 9 on GM03563 had only two altered points among 139, so in the permutations it was not unlikely to find the two altered points together. For chromosome 12 on GM01535, the region of alteration is represented by only one point and single altered points cannot be found when using a permutation reference. Finally, for chromosome 15 on GM07081, our result is consistent with Snijders *et al.* (2001) in that no evidence of an alteration is seen in the aCGH data. Therefore, our methodology found everything that it should have.

Our methodology also found a number of changes not detected by spectral karyotyping. We are calling these 'false positives', although some may be real and not detectable by spectral karyotyping. The number of false positive chromosomes ranged from 0 to 8, with averages of 4.1($SD = 2.6$) for $\alpha = 0.01$ and 1.8(2.1) for $\alpha = 0.001$. Most of the false positives were a result of what appeared to be local trends in the data, examples of which can be found in Figure 2, which shows the cell line GM03563. Note that these local trends were often in the same locations across cell lines suggesting that there may be a biological reason for them. The segmentation procedure detects change points as it approximates the local trend by a step function, thus leading to the 'false positives'.

4. ROMA example

Our methods were also applied to unpublished ROMA (Lucito *et al.*, 2003) experiments on 23 breast cancer cell lines. In this case, the labeled samples were hybridized to a slide containing 9820 unique probes

Table 1. *Results from applying CBS to nine cell lines with known copy number alterations. 'Yes' means the alteration was found for the particular cell line and chromosome at the given α level, while 'No' means that it was not. For GM07081/15, the asterisk is because there was no evidence in the array data of an alteration. 'False' is the number of chromosomes for the cell line where change-points were found that do not have known alterations*

Cell Line/Chrom.	$\alpha = 0.01$	$\alpha = 0.001$
GM03563/3	Yes	Yes
GM03563/9	No	No
GM03563/False	8	5
GM05296/10	Yes	Yes
GM05296/11	Yes	Yes
GM05296/False	3	0
GM01750/9	Yes	Yes
GM01750/14	Yes	Yes
GM01750/False	1	0
GM03134/8	Yes	Yes
GM03134/False	3	1
GM13330/1	Yes	Yes
GM13330/4	Yes	Yes
GM13330/False	8	5
GM01535/5	Yes	Yes
GM01535/12	No	No
GM01535/False	2	0
GM07081/7	Yes	Yes
GM07081/15	No*	No*
GM07081/False	1	0
GM13031/17	Yes	Yes
GM13031/False	5	3
GM01524/6	Yes	Yes
GM01524/False	6	2

that were each 70 bases long, with each probe spotted only once. Probes were mapped based on the draft human genome sequence. Each cell line was hybridized twice, once with the test sample labeled with Cy3 and the reference labeled with Cy5, and once with the two dyes swapped. This *dye-swapping* negates probe-specific bias in favor of Cy3 or Cy5. The arrays were imaged using the program Genepix. To show the robustness of our method, no spots were eliminated. The log base 2 test over references intensities were normalized by subtracting off the log ratio that corresponded to a *lowess* (Cleveland, 1979) fit of the log ratio to the average of the test and reference log intensities, as suggested by Yang *et al.* (2002). This normalization was undertaken in each of the 16 sub-arrays of each array. The normalized log ratios from

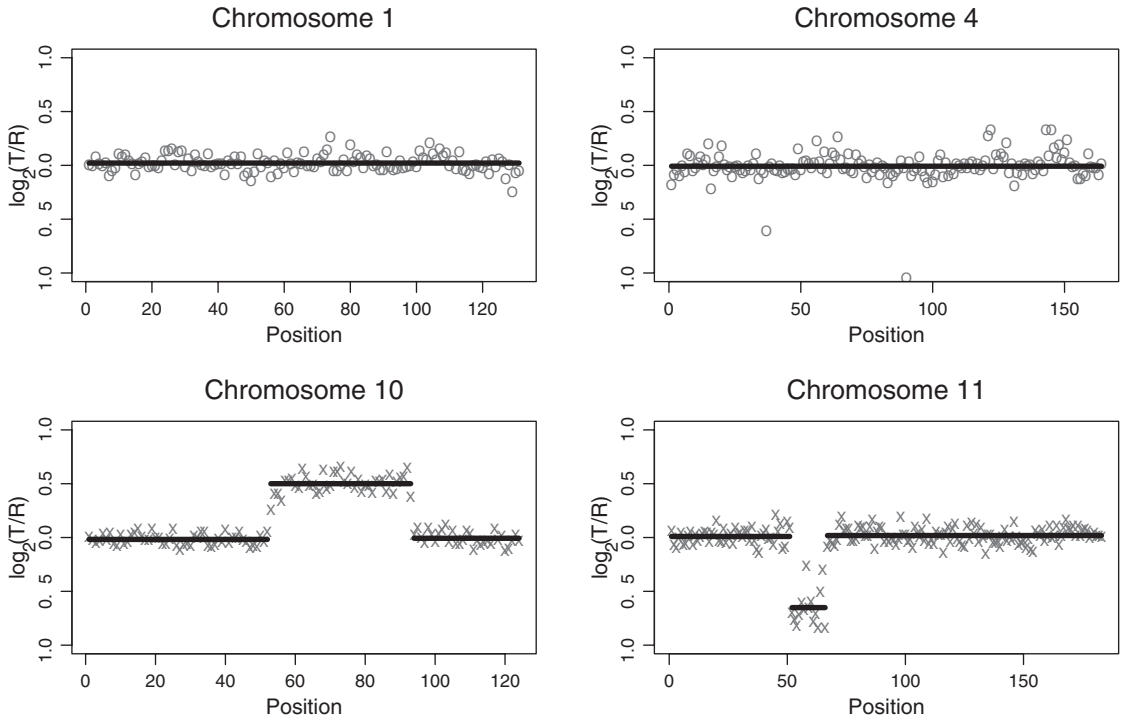

Fig. 1. A CBS analysis of the fibroblast cell line GM05296, which has known alterations only on chromosomes 10 and 11. The points are normalized log ratios, and the lines are the mean values among points in segments obtained by CBS.

each dye-swap were averaged before segmentation. The α level for CBS was fixed at 0.01.

The results from applying CBS to these data are shown in Figure 3. Since we do not have external confirmation of these results, our interpretation is necessarily modest. We first focus on a region of chromosome 17 near 40 MB where the ERB-B2 (HER2NEU) gene resides. The ERB-B2 gene is important because it is amplified in 10–40% of breast cancers (Menard *et al.*, 2002) and the drug Herceptin can be used to treat ERB-B2-amplified cancers. CBS found change-points in the ERB-B2 region in five of the cell lines. The cell line in the seventh row and first column of Figure 3 is particularly interesting in this region. Note that the ratios for some of the probes in the region appear to be at the normal level, but CBS is still able to define a likely aberrant region. In addition, CBS helps to define the altered region in the cases where ERB-B2 is amplified.

Another noteworthy aspect of these data is that there were eight cell lines with no ERB-B2 alteration where there appeared to be a copy number gain or loss in a whole arm of the chromosome. One would expect that the change-point would be found right at the centromere. Since, in most cases, the change-points were found three probes after the centromere, it is likely that those three probes are mis-mapped to the wrong side of the centromere. Thus the combination of the high-resolution ROMA data and the CBS algorithm was helpful in identifying these likely errors in the genome sequence.

Figure 4 shows results from the application of CBS to a whole breast cancer cell line. It can be seen that the sorted means of segments separated into plateaus. It is reasonable to assume that each plateau corresponds to a particular copy number, although what that copy number is remains unclear without additional information because the ploidy of the cell line is unknown. In addition, note the high degree of

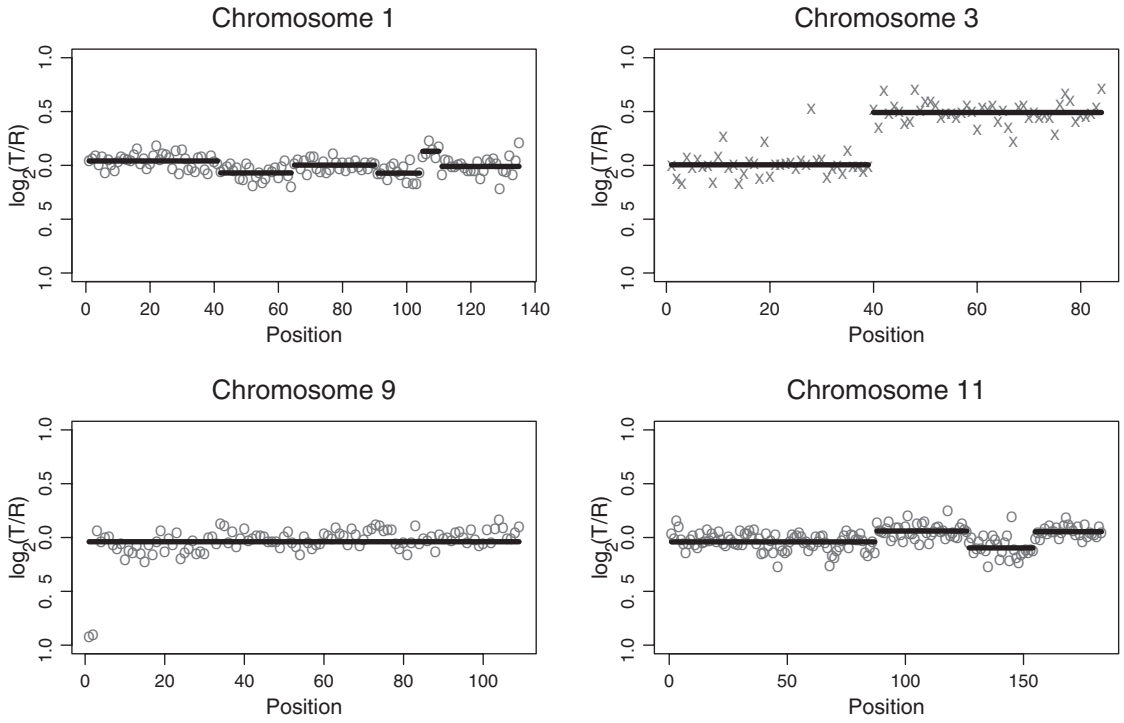

Fig. 2. A CBS analysis of the fibroblast cell line GM03563, which has known alterations only on chromosomes 3 and 9. Note that change-points found on chromosome 1 and chromosome 11 appear to be because of local trends in the data, and a change-point is missed on chromosome 9 because only two points among 139 are altered.

overlap of points that are part of segments in different plateaus. This overlap highlights the weakness of threshold-based methods.

5. SIMULATIONS

In this section we will present the results from the Monte Carlo simulations we conducted to evaluate the performance of the CBS algorithm. The data to be segmented were generated from the model $x_i = \mu_i + \epsilon_i$, $1 \leqslant i \leqslant n$, where n is the sample size, μ is the mean and ϵ the error term which is distributed as $N(0, \sigma^2)$. We used a permutation reference distribution to obtain the p-value of the segmentation procedure with values smaller than 0.01 $(= \alpha)$ being considered significant.

In the first set of simulations the mean was set to be $\mu_i = c\sigma\mathcal{I}\{l < i \leqslant l+k\}$, where $\mathcal{I}$ is the indicator function and the parameters c, l and k control the change in the mean, the location of the change and the width of the changed segment, respectively. For these simulations we chose the value of c from $\{2, 3, 4\}$, l from $\{0, \lfloor(n-k)/2\rfloor\}$ and k from $\{2, 3, 4, 5\}$. The two values for l correspond to the location of the changed segment being the edge and the center of the data, with the correct number of change-points being 1 in the first case and 2 in the second. The CBS algorithm was designed to overcome a shortcoming of binary segmentation which is that it cannot detect a narrow changed segment buried in the middle of a wide segment. Hence in this set of simulations we ran both the procedures to compare their performance. The number of change-points detected from segmenting 1000 simulated data sets are summarized in Table 2.

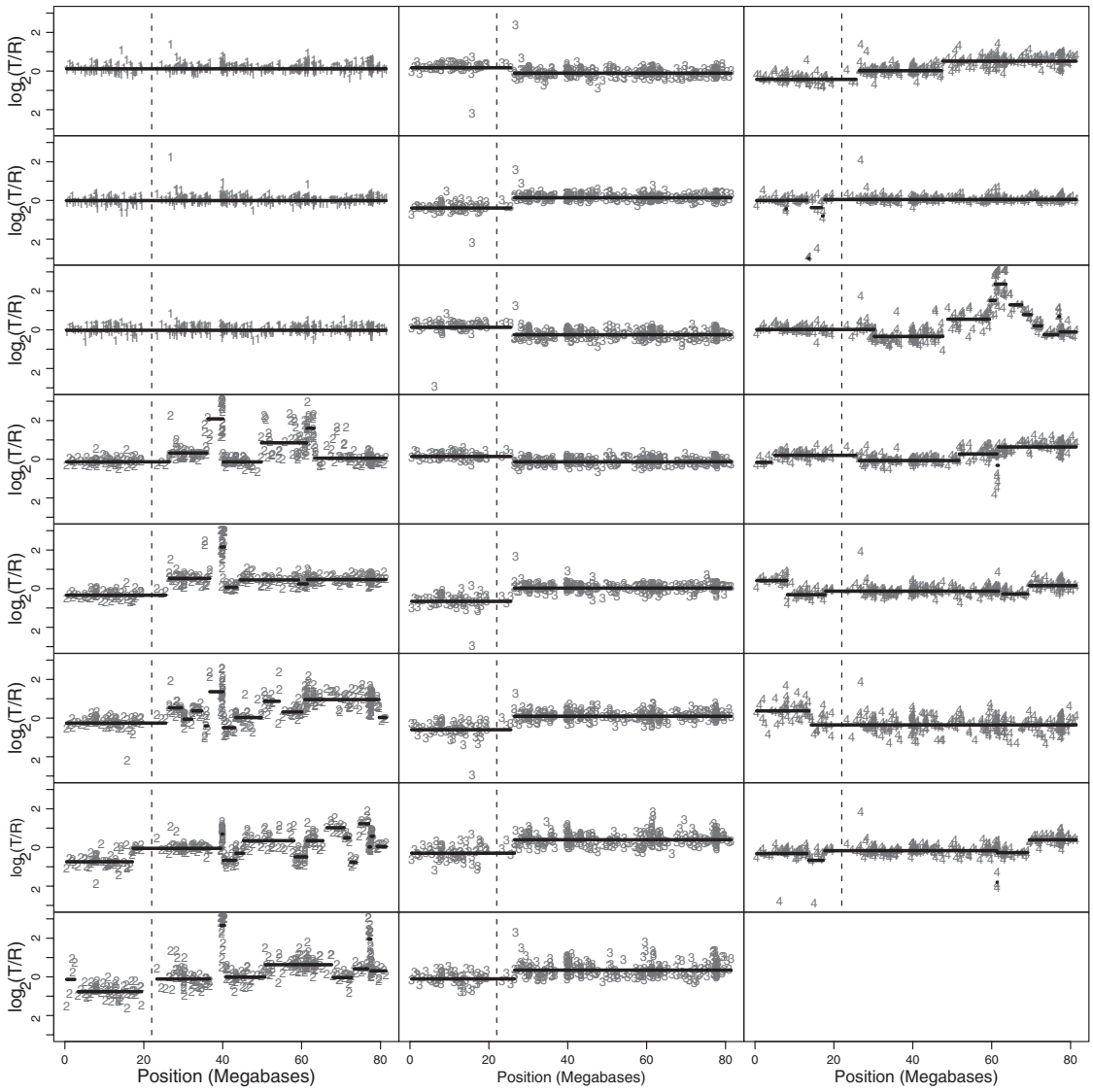

Fig. 3. CBS applied to 23 breast cancer cell lines using the ROMA technology. The plotting symbol for cell lines with no changes is 1, the symbol for those that appear to have ERB-B2 amplifications (near 40 MB) is 2, the symbols for those that appear to have whole-arm alteration is 3, and the symbol for those with other changes is 4. The dashed line is at the centromere.

The simulations show that the estimated number of changes exceeded the true number a maximum of 3% of the times (median: 1.75%; range: 0–3%). Even though it exceeds 1%, the excess is reasonable and is consistent with the multiple testing issue discussed in Section 2. The segmentation procedures have low power to detect a change when the difference in means is small or if the width of the changed segment is small. The proportion of data sets in which the estimated number of change-points equals the true number increases as either c or k increases, except when the binary segmentation procedure was used to detect

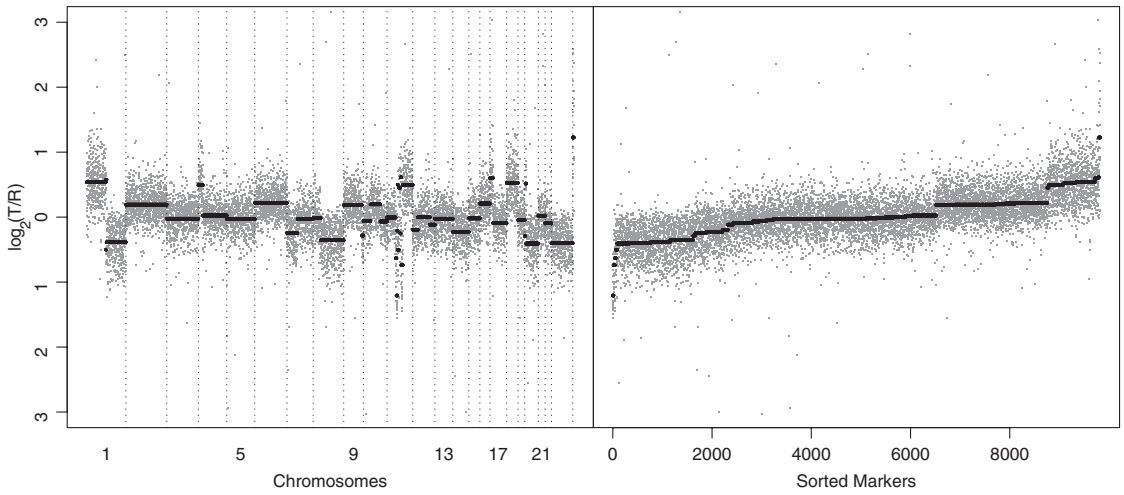

Fig. 4. A CBS analysis of a whole breast cancer cell line. The left panel shows markers arranged by chromosome, while the right shows the markers sorted by the mean of the corresponding segment. The points are the actual normalized log ratios, while the lines are the segment means. Each plateau in the segment means implies a different copy number in the test sample.

a changed segment in the middle. The simulation results are a clear demonstration of the inability of the binary segmentation to find a narrow aberrant region in the middle of a chromosome; no change was detected in over 99% of the data sets with the changed segment in the middle. The CBS procedure gains this ability by trading some of its power to detect a changed segment on the edge. Note that the power of both procedures increases more rapidly when the difference in means c increases than when the width k of the segment increases since change in c is equivalent to a change in the square root of k. Finally, the exact changed segment is more easily identified when the difference in the means increases than when the width of the segment increases. The exact estimation requires that the test statistic (Z_{ij} for CBS and Z_i for binary segmentation) is maximized when the segment edges are the true change-points. This happens when the segments are clearly separated, which is likely only when the two means are far apart.

A second set of simulations were performed with data sets simulated based on the CBS fit to chromosome 11 of a real ROMA breast cancer experiment. There were 497 markers in chromosome 11 with six change-points estimated at 137, 224, 241, 298, 307, 331 and the average log-ratios of intensities within segments given by:

i	1-137	138–224	225–241	242–298	299–307	308–331	332–497
$f(i)$	-0.18	0.08	1.07	-0.53	0.16	-0.69	-0.16

As earlier the data were generated using the model $x_i = \mu_i + \epsilon_i$ where μ is the mean and ϵ the error distributed as $N(0, \sigma^2)$. In these simulations a local trend component was incorporated into the mean in order to study its effect on segmentation giving the mean to be $\mu_i = f(i) + 0.25\sigma \sin(a\pi i)$. The noise parameter σ was set to be one of 0.1 or 0.2, and the trend parameter a was set to be one of 0, 0.01 or 0.025 corresponding to no trend and local trends with long and short periods respectively. Figure 5 shows typical data sets constructed by this model. The CBS algorithm was used to detect change-points in the simulated data using a permutation reference distribution with a p-value cutoff of 0.01. The change-point search

Table 2. *Counts of the number of change-points observed when applying CBS and binary segmentation to 1000 data sets of 250 points simulated from the Gaussian distribution. The columns under the heading 'Exact' provide the number of cases in which the exact number (1 for edge and 2 for center) and locations of the change-points are observed. Here k is the width of the changed segment and c is the number of standard deviations between the two means. The large font corresponds to the CBS results and the small font corresponds to the binary segmentation results. Each data set had one elevated region ranging from 2–5 points, and the elevated region varied from 2–4 SDs above the mean. The elevated region was all the way to one edge of the data set or at the exact center of the data set. The α for the simulation was* 0.01

		Change-points (edge)					Change-points (center)				
k	*c*	0	1	2	3-4	# Exact	0	1	2	3-4	# Exact
	2	980	11	8	1	5	968	0	32	0	9
		762	238	0	0	170	990	10	0	0	0
2	3	832	159	4	5	128	821	0	175	4	174
		297	699	4	0	607	992	8	0	0	0
	4	430	556	5	9	518	405	0	583	12	516
		34	957	9	0	900	995	5	0	0	0
	2	874	115	7	4	81	857	0	141	2	79
		539	458	3	0	294	992	8	0	0	0
3	3	348	635	8	9	538	330	0	654	16	496
		76	914	10	0	754	994	6	0	0	0
	4	35	947	2	16	891	23	0	954	23	847
		1	989	10	0	925	995	5	0	0	0
	2	720	261	15	4	192	689	0	307	4	159
		334	662	4	0	439	992	8	0	0	0
4	3	115	863	10	12	716	97	0	883	20	648
		12	979	9	0	802	994	5	1	0	0
	4	3	977	5	15	918	0	0	978	22	867
		0	990	10	0	931	996	3	1	0	0
	2	531	439	23	7	297	511	0	481	8	232
		192	801	7	0	516	991	7	2	0	0
5	3	24	954	6	15	818	19	0	961	20	692
		1	988	11	0	842	994	5	1	0	0
	4	0	982	4	14	937	0	0	981	19	877
		0	989	11	0	943	997	1	1	1	1

was undertaken either over all data points or over overlapping windows of size 100 that had 75% overlap. The change-points detected thus were pruned using a sum of squares threshold γ of 0.05. In addition to the number of change-points detected the following distance measure was used to assess the accuracy of the procedure. Let $\nu_1 < \cdots < \nu_k$ be the true change-points and $\hat{\nu}_1 < \cdots < \hat{\nu}_{\hat{k}}$ be the estimated change-points where $\hat{k}$ is the number of change-points detected. The distance measure D is defined as $\max_{1 \leqslant i \leqslant k} |\nu_i - \hat{\nu}_i|$ for a data set with true number of changes (i.e. $\hat{k} = k$) and undefined otherwise. The number of change-points detected and the median and range of D are shown in Table 3 for the unpruned procedure and Table 4 for the pruned procedure. These results are based on 100 replicate simulations.

Table 3 shows that unpruned CBS found the correct number of change-points in at least 52% of the simulations with false positive detection more likely. The number of false positives appear to be nominal

Table 3. *Counts of the number of change-points observed when applying CBS to data simulated from the step-function f from Section 5. If 'Window' is 'No', a search over all points was undertaken to find the best change-point. If 'Window' is 'Yes', the search was only over overlapping window (window size* $= 100$*; overlap* $= 75\%$*). If 'Trend' is 'Long',* $a = 0.01$*. If Trend is 'Short',* $a = 0.025$*. The true number of change-points was six. Distances are the maxima of the minimum distances from the* i*th observed change-point to the* i*th true change-point in every simulation. The median and range was then computed over 100 simulations*

			Number of change-points							Max. distance	
σ	Window	Trend	5	6	7	8	9	10	11	Median	Range
0.1	No	None	0	96	1	2	1	0	0	0	0–4
0.1	Yes	None	0	92	5	3	0	0	0	0	0–5
0.2	No	None	0	91	6	2	1	0	0	1	0–11
0.2	Yes	None	9	78	10	3	0	0	0	1	0–28
0.1	No	Short	0	82	5	10	2	1	0	0	0–6
0.1	Yes	Short	0	73	8	15	2	2	0	0	0–6
0.2	No	Short	0	66	23	8	3	0	0	1	0–23
0.2	Yes	Short	7	58	17	11	4	3	0	2	0–28
0.1	No	Long	0	71	7	18	4	0	0	0	0–5
0.1	Yes	Long	0	75	12	9	2	2	0	0	0–6
0.2	No	Long	0	68	11	17	2	1	1	1	0–28
0.2	Yes	Long	18	52	20	6	2	2	0	2	0–46

Table 4. *As Table 3, except that the change-points have been pruned*

			Number of change-points							Max. distance	
σ	Window	Trend	5	6	7	8	9	10	11	Median	Range
0.1	No	None	0	99	1	0	0	0	0	0	0–4
0.1	Yes	None	0	100	0	0	0	0	0	0	0–5
0.2	No	None	0	99	1	0	0	0	0	1	0–11
0.2	Yes	None	9	91	0	0	0	0	0	1	0–28
0.1	No	Short	0	94	5	1	0	0	0	0	0–6
0.1	Yes	Short	0	94	4	2	0	0	0	0	0–6
0.2	No	Short	0	95	5	0	0	0	0	1	0–23
0.2	Yes	Short	8	84	6	2	0	0	0	1	0–31
0.1	No	Long	0	91	9	0	0	0	0	0	0–5
0.1	Yes	Long	0	94	4	1	1	0	0	0	0–6
0.2	No	Long	0	90	8	1	1	0	0	2	0–40
0.2	Yes	Long	20	75	4	1	0	0	0	1	0–46

when there is no trend in the mean function and the longer the period of the local trend the larger the number of false positives. This is the expected behavior since the longer the period of local trend the easier it is to approximate it by a non-constant step function. Larger noise in the data makes it more difficult to detect a change-point and estimate its location correctly. This can be seen in the numbers of change-points and the median and range of D used to assess the accuracy of the procedure. The windowing scheme was

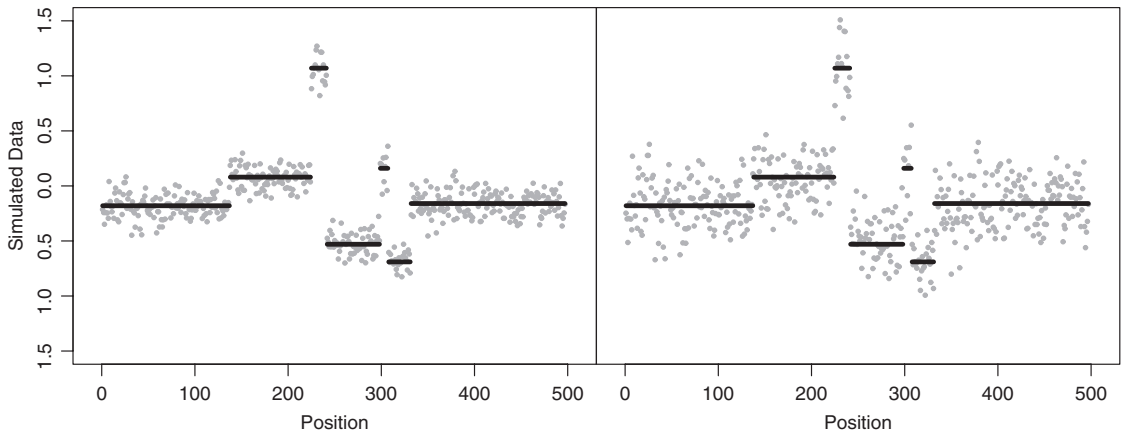

Fig. 5. Example simulation data sets. The lines are the step function. The left panel is for the data set when $\sigma = 0.1$ and no local trend. The right panel is for the same data set except $\sigma = 0.2$ and the local trend has a long period.

devised to ease the computational burden which grows as a square of the sample size. Windowing appeared to have little impact when the noise was low, but it led to a change-point being missed occasionally when noise was high. Similarly, there is a small drop in the accuracy of the estimated locations of the change-points as seen in the medians and ranges of D. Thus the computational gains from windowing appear to come at the cost of a small drop in the accuracy of the procedure.

Table 4 shows that pruning greatly improves the accuracy of the procedure by substantially reducing the false positives caused by the local trends in the data. This can be seen in the larger number of times the correct number of changes are estimated with minimal detrimental effect on D. Pruning also occasionally removes a false positive change in a case with the correct number of change-points leading to the appearance of a missed change-point. Overall, these simulations show that the CBS procedure with pruning is a desirable method to analyze copy number data.

6. Discussion

We have developed a variant of binary segmentation that we call circular binary segmentation or CBS for identifying genomic alterations in array copy number experiments. We applied our procedure to copy number data from aCGH experiments on 15 fibroblast cell lines. The CBS algorithm identified all the expected alterations detected through spectral karyotyping of the cell lines. We also applied our procedure to ROMA data from 23 breast cancer cell lines. While there is no biological verification for all the changes detected, the procedure found alterations in the ERB-B2 region of chromosome 17 in 5 of the 23 cases which is consistent with the known rate of abnormality in breast cancer in this region. Finally, we showed through a series of simulations that the procedure performs well in identifying changes and estimating their locations, especially when detecting narrow regions of change of the square wave type.

Even though the step-function model is appropriate for copy number data, we have seen fluctuations in the log intensity ratio that are not due to copy number changes. We call these local trends since these fluctuations exhibit a similar pattern across cell line case and believe that they may have a biological reason. These local trends can lead to false positive detection of change-points. We developed a pruning component to our procedure to address this problem and showed through simulations that it achieves the goal of removing most false-positive change-points. We are currently exploring methods to estimate

local trends from data across cell lines so that it would be possible to subtract out local trends before segmenting.

The number of computations needed to obtain the test statistic used in CBS is a function of the square of the sample size. Since our test procedure is based on a permutation reference distribution, these computations must be repeated thousands of times to accurately estimate the upper tail probability of the reference distribution. We have developed a windowing method that reduces this computational burden. We showed here through simulations that it has minimal effect on the accuracy of the procedure and have applied it to data from arrays that contained 85 000 markers (Lucito *et al.*, 2003). We are devising additional methods to further speed up the computations in order to analyze even larger data sets.

Once the change-points have been estimated it is of interest to estimate the copy numbers of the test sample in every region. One possibility that operates on the output of the CBS procedure is presented in Lucito *et al.* (2003). As demonstrated in Figure 4, the CBS procedure segments array copy number data into regions whose means are consistent across chromosomes. These plateaus in the plot of the segment means reflect different copy number states in the test sample. It is not possible to know the true copy numbers in each of these states without additional data acquired using another technique.

The software used in this paper was written in R and Fortran and is freely available at `http://www.mskcc.org/biostat/~olshena/research/`.

ACKNOWLEDGEMENTS

The authors would like to thank an associate editor and a referee for helpful comments. They would also like to thank David Siegmund for his input.

This work was supported by a grant to E. S. Venkatraman from the NCI (CA73848). The work was also supported by grants to Michael Wigler from the National Institutes of Health and NCI (CA078544; CA45508); 1 in 9: The Long Island Breast Cancer Action Coalition; Lillian Goldman and the Breast Cancer Research Foundation; Marks Family Foundation; The Miracle Foundation. Michael Wigler is an American Cancer Society Research Professor.

REFERENCES

AUTIO, R., HAUTANIEMI, S., KAURANIEMI, P., YLI-HARAJA, O., ASTOLA, J., WOLF, M. AND KALIONIEMI, A. (2003). CGH-Plotter: MATLAB toolbox fo CGH-data analysis. *Bioinformatics* **19**, 1714–1715.

BASSEVILLE, M. (1988). Detecting changes in signals and systems—a survey. *Automatica* **24**, 309–326.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. AND STONE, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

HODGSON, G., HAGER, J. H., VOLIK, S., HARIONO, S., WERNICK, M., MOORE, D., NOWAK, N., ALBERTSON, D. G., PINKEL, D., COLLINS, C. *et al.*, (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **29**, 459–464.

KALLIONIEMI, A., KALLIONIEMI, O.-P., SUDAR, D., RUTOVITZ, D., GRAY, J. W., WALDMAN, F. AND PINKEL, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821.

LEVIN, B. AND KLINE, J. (1985). The CUSUM test of homogeneity with an application in spontaneous abortion epidemiology. *Statistics in Medicine* **4**, 469–488.

LINN, S. C., WEST, R. B., POLLACK, J. R., ZHU, S., HERNANDEZ-BOUSSARD, T., NIELSEN, T. O., RUBIN, B. P., PATEL, R., GOLDBLUM, J. R., SIEGMUND, D. *et al.*, (2003). Gene expression patterns and gene copy number changes in dermatofibrosarcoma protuberans. *American Journal of Pathology* **163**, 2383–2395.

LISITSYN, N., LISITSYN, N. AND WIGLER, M. (1993). Cloning the differences between two complex genomes. *Science* **259**, 946–951.

LUCITO, R., HEALY, J., ALEXANDER, J., REINER, A., ESPOSITO, D., CHI, M., RODGERS, L., BRADY, A., SEBAT, J., TROGE, J., WEST, J. A. *et al.*, (2003). Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Research* **13**, 2291–2305.

LUCITO, R., WEST, J., REINER, A., ALEXANDER, D., ESPOSITO, D., MISHRA, B., POWERS, S., NORTON, L. AND WIGLER, M. (2000). Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Research* **10**, 1726–1736.

MENARD, S., TAGLIABUE, E., CAMPIGLIO, M. AND PUPA, S. M. (2002). Role of HER2 gene overexpression in breast carcinoma. *Journal of Cellular Physiology* **182**, 150–162.

PINKEL, D., SEAGRAVES, R., SUDAR, D., CLARK, S., POOLE, I., KOWBEL, D., COLLINS, C., KUO, W.-L., CHEN, C., ZHAI, Y., ZHAI, Y., DAIRKEE, S., LJJUNG, B.-M., GRAY, J. W. AND ALBERTSON, D. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–211.

POLLACK, J. R., PEROU, C. M., ALIZADEH, A. A., EISEN, M. B., PERGAMENSCHIKOV, A., WILLIAMS, C. F., JEFFREY, S. S., BOTSTEIN, D. AND BROWN, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics* **23**, 41–46.

POLLACK, J. R., SORLIE, T., PEROU, C. M., REES, C. A., JEFFREY, S. S., LONNING, P. E., TIBSHIRANI, R., BOTSTEIN, D., BORRESEN-DALE, A. L. AND BROWN, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences USA*, **99**, pp. 12963–12968.

SEN, A. AND SRIVASTAVA, M. S. (1975). On tests for detecting a change in mean. *Annals of Statistics* **3**, 98–108.

SHABAN, S. A. (1980). Change-point problem and two phase regression: an annotated bibliography. *International Statistical Review* **48**, 83–93.

SIEGMUND, D. (1986). Boundary crossing probabilities and statistical applications. *Annals of Statistics* **14**, 361–404.

SNIJDERS, A. M., FRIDLYAND, J., MANS, D. A., SEGRAVES, R., JAIN, A. N., PINKEL, D. AND ALBERSTON, D. G. (2003). Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene* **22**, 4370–4379.

SNIJDERS, A. M., NOWAK, N., SEGRAVES, R., BLACKWOOD, S., BROWN, N., CONROY, J., HAMILTON, G., HINDLE, A. K., HUEY, B., KIMURA, K. *et al.*, (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29**, 263–264.

VENKATRAMAN, E. S. (1992). Consistency results in multiple change-point situations. *Technical report*, Department of Statistics. Stanford University.

VOSTRIKOVA, L. J. (1981). Detecting 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady* **24**, 55–59.

YANG, Y. H., DUDOIT, S., LUU, P., LIN, D., PENG, V., NGAI, J. AND SPEED, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, e15.

WEISS, M. M., SNIJDERS, A. M., KUIPERS, E. J., YLSTRA, B., PINKEL, D., MEUWISSEN, S. G. M., VAN DIEST, P. J., ALBERTSON, D. G. AND MEIJER, G. A. (2003). Determination of amplicon boundaries at 20q13.2 in tissue

samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization. *The Journal of Pathology* **200**, 320–326.

YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' Criterion. *Statistics and Probability Letters* **6**, 181–189.

[*Received December 5, 2003; first revision March 9, 2004; second revision March 19, 2004*]