Biostatistics (2005), **6**, 1, *pp.* 45–58
doi: 10.1093/biostatistics/kxh017

A method for calling gains and losses in array CGH data

PEI WANG
Department of Statistics, Stanford University, CA, 94305, USA
wp57@stanford.edu

YOUNG KIM, JONATHAN POLLACK
Department of Pathology, Stanford University, CA, 94305, USA

BALASUBRAMANIAN NARASIMHAN
Department of Statistics, Stanford University, CA, 94305, USA

ROBERT TIBSHIRANI
Departments of Health, Research & Policy, and Statistics, Stanford University, CA, 94305, USA

SUMMARY

Array CGH is a powerful technique for genomic studies of cancer. It enables one to carry out genome-wide screening for regions of genetic alterations, such as chromosome gains and losses, or localized amplifications and deletions. In this paper, we propose a new algorithm 'Cluster along chromosomes' (CLAC) for the analysis of array CGH data. CLAC builds hierarchical clustering-style trees along each chromosome arm (or chromosome), and then selects the 'interesting' clusters by controlling the False Discovery Rate (FDR) at a certain level. In addition, it provides a consensus summary across a set of arrays, as well as an estimate of the corresponding FDR. We illustrate the method using an application of CLAC on a lung cancer microarray CGH data set as well as a BAC array CGH data set of aneuploid cell strains.

Keywords: Array CGH; CLAC; Cluster; DNA copy number; FDR.

1. INTRODUCTION

Genomic DNA copy number alterations are key genetic events in the development and progression of human cancers (Lengauer *et al.*, 1998). The technique of microarray comparative genomic hybridization (array CGH) enables one to screen genome-wide for all possible regions with DNA copy number alterations. With more and more array CGH data sets emerging, there is a need for efficient algorithms that automatically select regions of gains and losses and at the same time provide some estimate of error for this selection process.

Some work has already been done to address this question. The authors in Pollack *et al.* (2002) devised a threshold method of selecting genes with extreme log ratios outside certain threshold bounds, and provided an estimate of the FDR for the result. In another paper (Hodgson *et al.*, 2001), the distribution of the log ratio of the DNA copy numbers was treated as a mixture of three Gaussian distributions and an MLE solution was implemented. Moreover, Cheng *et al.* (2003) described a regression-based statistical method to test for altered copy numbers. However, a common drawback of the above methods is that

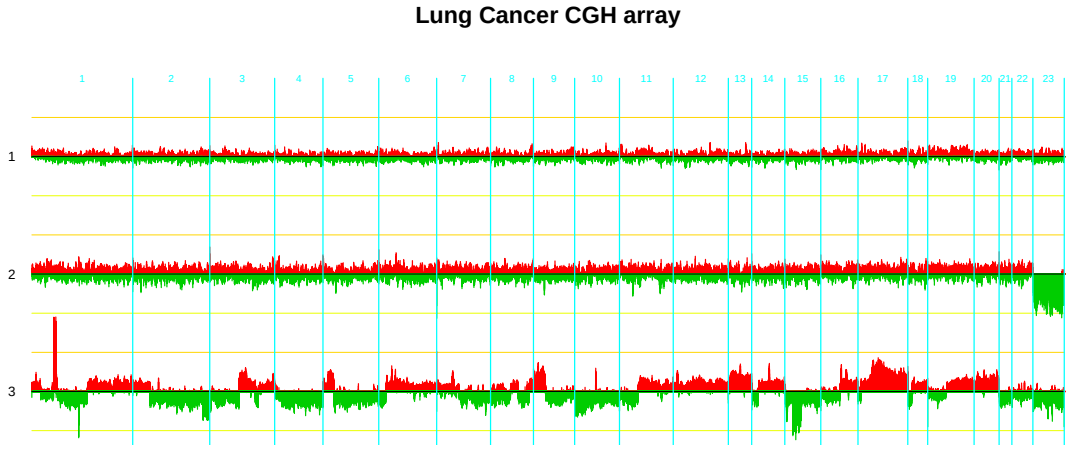

Fig. 1. The result of CGH on cDNA microarrays. The picture shows the result of three CGH arrays (with moving average of 5). The top one (1) is the CGH array result of a normal sample. The middle one (2) is the result of a XY/XX setup. The picture at the bottom (3) shows a cell line of lung cancer. In each array, the horizontal direction is the genome order from $1pter$ to $23qter$. The vertical blue lines indicate the end point of each chromosome; the height of the red and green lines represent the $\log_2\left(\frac{\text{red light intensity}}{\text{green light intensity}}\right)$ for the genes/clones. The red color means the log ratio for that gene/clone is positive, while the green color means the log ratio is negative. The width between two horizontal yellow lines is exactly 1.

the spatial relationship among genes on the genome has not been taken into consideration—an important factor for DNA copy number alterations.

Some other methods have taken into account the spatial factor, such as the break point model proposed in Jong *et al.* (2003), which is similar to the model used in the Matlab tool CGH-Plotter (Autio *et al.*, 2003). The idea of these two methods is to divide the genome to K regions, and then assign each region to one of the states of {normal, gain, loss}. One difficulty with this kind of method is that the value of K is always hard to determine. In addition, the computational expense in looking for the best 'break points' might be high when K is large.

In another study (Snijders *et al.*, 2003), the authors used an unsupervised Hidden Markov Model methodology to identify copy number transitions on the chromosomes. Though it does not provide a way to estimate the FDR, the procedure we propose in this paper would be applicable to their method too.

In this paper, we describe a very simple but efficient method—'Cluster Along Chromosomes' (CLAC)—which can accurately pick out the signal regions by capturing the underlying spatial structure of the genomic alterations. The idea is to build a hierarchical cluster-style tree along each chromosome arm, such that the gain/loss regions are separated into different branches. Then for each node of the tree, three statistics are considered: the height of this node in the tree, the size of the sub-tree, and the mean value of the leaves of the sub-tree. Then the nodes corresponding to the gain/loss regions are selected based on the joint distribution of these three statistics. The FDR is estimated by approximating the null distribution with some normal–normal array hybridizations. The FDR for the consensus summary across different samples is also estimated.

Section 2 provides some background information for the array CGH experiment. Section 3 describes the CLAC algorithm, while Section 4 outlines details of how to estimate the FDR. Applications of CLAC to 48 CGH arrays of lung cancer as well as four BAC arrays of aneuploid cell strains are shown in Section 5. In the Appendix, we discuss the choices of tuning parameters for the method.

2. ARRAY CGH

Normally a human cell contains two copies of each of the 22 non-sex chromosomes: when genetic alterations occur, the DNA copy numbers will differ from two. Array Comparative Genomic Hybridization (array-CGH) is an approach for genome-wide scanning of differences in DNA copy numbers. In a typical experiment, a tumor sample labeled red (Cy5) is hybridized to a reference normal sample labeled green (Cy3). For each gene/clone (one spot on the microarray chips), a scanner reports the ratio of the red light intensity to the green light intensity, which corresponds to the ratio of the DNA copy number of the gene in the tumor sample to that of the normal sample. A more elaborate introduction to array-CGH can be found in Pinkel *et al.* (1998). Figure 1 illustrates the output data from some CGH experiments using cDNA microarrays.

Our goal in this paper is to objectively identify regions of gains and losses in such CGH array outputs. We describe the method next.

(Note that, we use the term 'genes/clones' to refer spots on microarray chips, while there are some other studies using term 'loci(locus)' instead, especially for BAC arrays. In addition, sometimes we may just use 'gene' instead of 'gene/clone' for simplicity.)

3. THE CLAC ALGORITHM

3.1 *Cluster formation*

The CLAC algorithm uses a variation of a standard agglomerative clustering algorithm, a bottom-up strategy that generates a binary tree to represent the similarities in the data. Agglomerative clustering algorithms begin with every observation representing a singleton cluster. At each of the $n - 1$ steps (n = total number of objects) the closest two (least dissimilar) clusters are merged into a single cluster, producing one less cluster at the next higher level (see e.g. Hastie *et al.*, 2001, p. 475).

When we try to separate the regions of gains and losses based on the array output (Figure 1), the problem here is very similar to a clustering problem. So building a hierarchical clustering tree is a reasonable way to understand the underlying data structure. This is illustrated in Figure 2.

Although the performance of the method would not be affected whether it considers chromosome arms separately or not, due to the size of the arrays, we work with chromosome arms for microarray CGH, and with whole chromosomes for BAC array CGH.

Building a clustering tree along one chromosome arm (or one chromosome) differs in two ways from standard agglomerative clustering. First, the order of the genes on the chromosome is fixed, i.e. the order of the leaves of the tree is fixed. So only adjacent clusters are joined together when the tree is generated from the bottom-up. This makes the algorithm $O(n)$, while the original agglomerative clustering algorithm is $O(n^2)$. Second, the 'similarity' between two clusters no longer refers to the spatial distance but to the similarity of the array measurements ($\log_2(\frac{\text{red}}{\text{green}})$) between the two clusters. Therefore we introduce a statistic called 'relative difference' (`rd`) to measure this similarity.

Suppose we have n genes(clones) on one chromosome arm. The $\log_2$ ratios for these n genes(clones) are $\{x_1, x_2, \ldots, x_n\}$. Define relative difference (`rd`) for two contiguous genes(clones) as

$$\texttt{rd}(x_i, x_{i+1}) = \frac{|x_i - x_{i+1}|}{|x_i| + |x_{i+1}| + |x_i + x_{i+1}|}. \tag{3.1}$$

The denominator $|x_i| + |x_{i+1}| + |x_i + x_{i+1}|$ in the above definition gives some advantage to gene pairs with relatively larger absolute values ($|x_i| + |x_{i+1}|$), while having the same signs ($|x_i + x_{i+1}|$).

Also define the distance between two contiguous clusters $C_i = \{i_1, i_2, \ldots, i_k\}$ and $C_j =$

$\{j_1, j_2, \ldots, j_k\}$ as

$$\texttt{rd}_{\text{nearby}}(C_i, C_j) = \texttt{rd}(x_{i_k}, x_{j_1}), \tag{3.2}$$

Or

$$\texttt{rd}_{\max}(C_i, C_j) = \max\{\texttt{rd}(x_{i_t}, x_{j_s}) | i_t \in C_i, j_s \in C_j\}, \tag{3.3}$$

where C_i is the left neighbor of C_j, i.e. $j_1 = i_k + 1$. Since the definition of $\texttt{rd}_{\max}$ here is very similar to complete linkage in the hierarchical clustering literature, we still refer to $\texttt{rd}_{\max}$ as complete linkage, while referring to $\texttt{rd}_{\text{nearby}}$ as nearby linkage.

We summarize the clustering procedure on one chromosome arm as follows:

Clustering procedure along one chromosome arm
1. Begin with n clusters with one gene(clone) in each cluster.
2. Merge the two adjacent clusters with the smallest rd.
3. Repeat Step 2 until one big cluster is obtained.

The tree structure for a simulated example is shown in Figure 2. In this example, the log 2 ratio values for a chromosome segment with 150 genes/clones are simulated as follows:

$$\begin{aligned} \text{Amplified region} &: g_i \sim N(0.7, 0.3), i \in A = \{10, 11, \ldots, 60\} \\ \text{Deleted region} &: g_i \sim N(-0.7, 0.3), i \in D = \{90, 91, \ldots, 140\} \\ \text{Noise region} &: g_i \sim N(0, 0.3), i \overline{\in} (A \cup D). \end{aligned}$$

In addition, an average smoothing of window-size 5 has been performed on this chromosome segment before the CLAC tree is built on it. In the Figure, the height of a node in the tree represents the rd between the left branch and right branch of that node. We can see that the region with copy number gains/losses has joined the tree at a rather smaller rd level than the noisy parts.

3.2 *Cluster selection*

After we build the hierarchical tree, we need to decide which 'clusters' are 'interesting.' We consider all $n - 1$ clusters corresponding to nodes of the tree at all heights (a cluster consists of at least two genes/clones).

In standard hierarchical clustering, the dendrogram is often cut at a fixed height, and the clusters formed above that height are selected. However, in our application, the choice of clusters is more complex as we must take into account different aspects of the quality of clusters.

We examine three properties of each node/cluster:

1. rd: The rd of this node in the tree (the biggest rd for nearby gene pairs in the cluster, for the cases of nearby linkage).
2. size: The size of the subtree with this node as the root (the number of genes in the cluster).
3. meanvalue: The mean value of the leaves of the subtree (the mean value of the log ratio for genes in the cluster).

To simplify, we transform $\{\texttt{size}_i\}$ monotonically into [0, 1] by defining

$$\texttt{lsize}_i = \frac{\log(\texttt{size}_i)}{\max\{\log(\texttt{size}_i)\}} \tag{3.4}$$

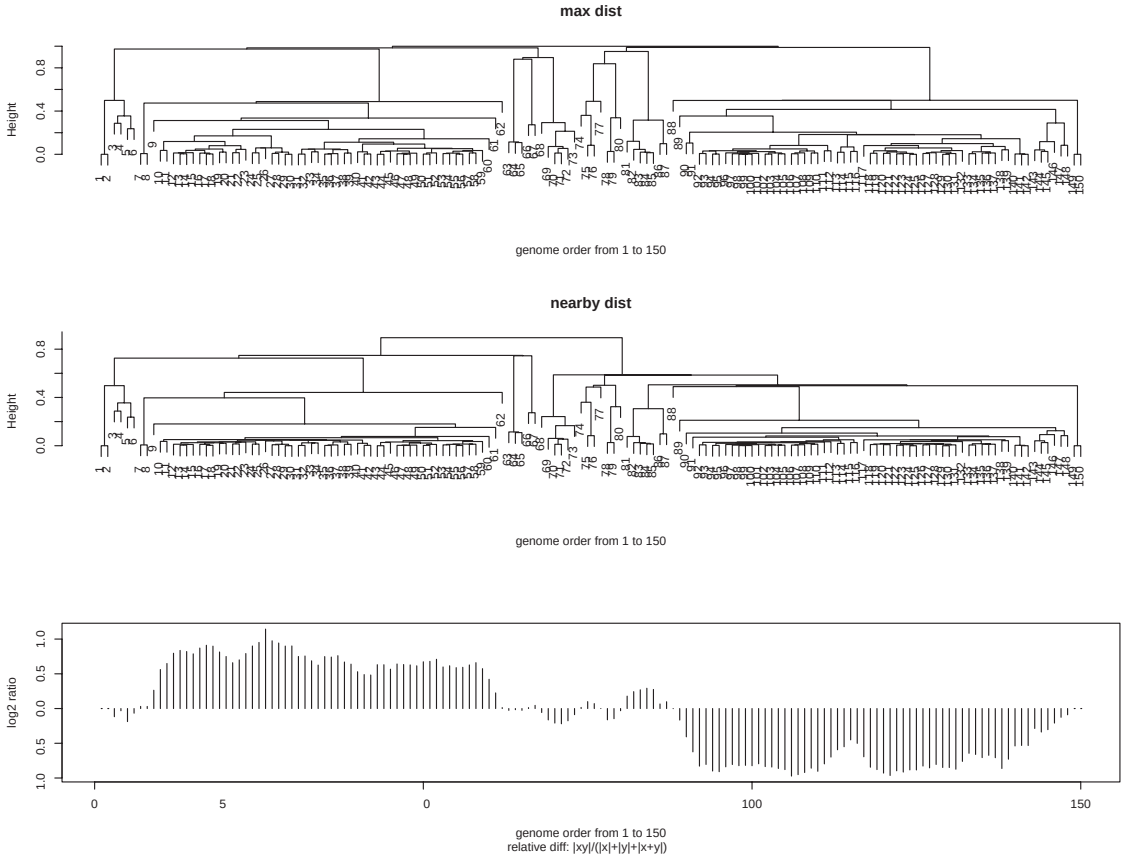

Fig. 2. Tree structure for a simulated example. The bottom picture shows the $\log_2$ ratio of CGH measurements for this 150 length chromosome arm. The tree at the top is built based on the $\texttt{rd}_{\max}$ defined in (3.3), while the tree in the middle is built based on the $\texttt{rd}_{\text{nearby}}$ defined in (3.2).

Figure 3 shows two examples of the empirical joint distribution of (`lsize`, `rd`) and (`lsize`, `meanvalue`).

Now we can use (`size`, `rd`, `meanvalue`) to select 'interesting' regions. There are two different kinds of interesting regions. The first kind is characterized as a big spike, which is always a small region with extremely large or small log ratio values. The second kind is the consistent gain/loss region, whose log ratios might not deviate away from 0 very much, but tend to stay positive(gain) or negative(loss) in the whole region.

The big spike regions correspond to the nodes with big |`meanvalue`|. We choose a cutoff as

$$|\texttt{meanvalue}| > \min(1, 3 \cdot \text{sd}). \tag{3.5}$$

Here 'sd' is the standard deviation of all the $\log 2$ ratio measurements from this array. The '1' in the equation refers to two DNA copy gain ($\log_2(\frac{4}{2}) = 1$) or one DNA copy loss ($\log_2(\frac{1}{2}) = -1$). Since we are focussing at short segments of genome here (amplifications or deletions with a size more than ∼1000kB would be captured as regions of consistent gains/losses), one or two extreme noise measurements would

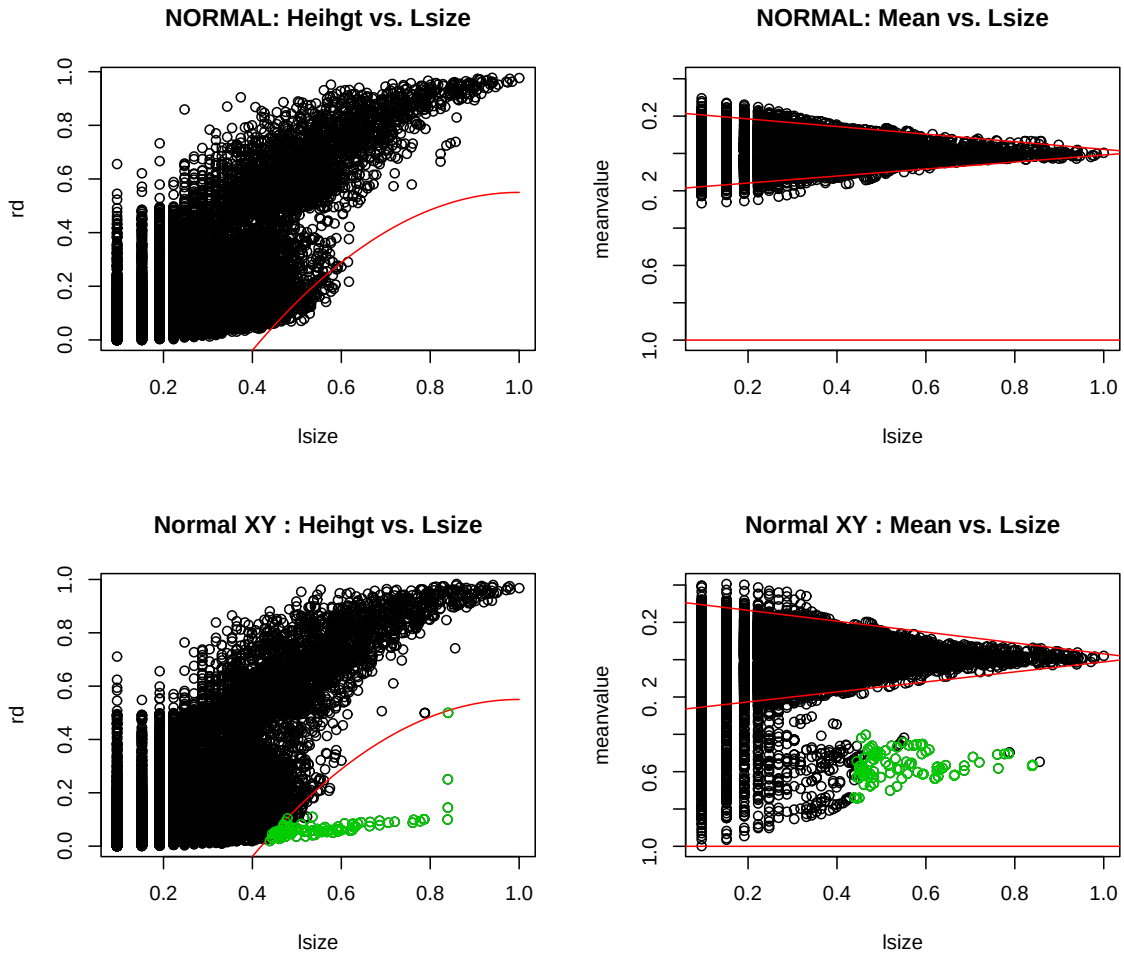

Fig. 3. Empirical joint distribution of (`lsize`, `rd`) and (`lsize`, `meanvalue`). The upper two plots are for the normal–normal array hybridization, while the lower two plots are for the XY/XX array hybridization. The red curves/lines represent the selecting rule (the `FDR` is set to be 1%). The points outside the band of $\texttt{meanvalue} = \pm 1$ in the (`lsize`, `meanvalue`) figure are selected as big-spike regions (no such point in this figure). The nodes that are both below the red curve in the (`lsize`, `rd`) plot, and outside the red triangular of the (`lsize`, `meanvalue`) plot, are selected as consistent gain/loss regions. The green points represent the selected nodes.

affect the |`meanvalue`| greatly. Thus the conservative cutoff $\min(1, 3 \cdot \text{sd})$ helps to eliminate wrong calls caused by random noise.

The second kind of regions—consistent gain/loss regions—correspond to the nodes with bigger `size`, smaller `rd`, and with |`meanvalue`| not too small.

Nodes with big `size` and small `rd` lie in the lower right corner of the plane of (`rd`, `lsize`). Observing the curved shape of the null distribution, we settled on the rule

$$\texttt{rd} < a \cdot (1 - \texttt{lsize})^2 + b \tag{3.6}$$

for (`rd`, `size`). Any node above the curve has too large an `rd` for its `size`, so it is not considered to

represent a gain/loss region.

In the plane of (`mean value`, `lsize`), the null distribution has a clear triangular shape, so we use two lines to represent this triangular. Any node lying between these two lines has too small a `mean value` for its `size`, so it is not considered to represent a gain/loss region either. This rule can be denoted as

$$(\texttt{meanvalue} > -k_1 \cdot \texttt{lsize} + b_1)|(\texttt{meanvalue} < k_2 \cdot \texttt{lsize} - b_2), \tag{3.7}$$

where b, a, k_1, k_2, b_1 and b_2 are parameters. The values of b, k_1, k_2 and b_1 are fixed, as detailed in the Appendix. The parameter a is a tuning parameter: varying a changes the false discovery rate as discussed in Section 4.

4. FDR ESTIMATION

4.1 *FDR for one array*

The false discovery rate is defined in Benjamini & Hochberg (1995) as

$$\mathsf{FDR} = E\ (\text{proportion of rejected } H_i \text{ that are actually true}) \tag{4.1}$$

with the proportion equaling zero if nothing is rejected. More precisely, let

$$R = \text{number of genes/clones selected,}$$
$$V = \text{number of genes/clones that are selected but from } H_0,$$

then

$$\mathsf{FDR} = E\left(\frac{V}{R} \cdot 1_{\{R>0\}}\right) = E\left(\frac{V}{R}\middle| R > 0\right) P(R > 0). \tag{4.2}$$

Here, the null hypothesis for gene/clone i is

$$H_0 = \text{ Gene/clone } i \text{ does not belong to any gain/loss region.}$$

Although we do not have independent H_0 for each gene in our problem, we still can use

$$\widehat{\mathsf{FDR}} = \frac{\text{number of genes picked in the normal array (under the same criteria)}}{\text{number of genes picked in the tumor array}} \tag{4.3}$$

as an estimator for FDR (Benjamini & Hochberg, 1995; Tusher *et al.*, 2001; Storey, 2002; Efron & Tibshirani, 2002).

Therefore, for each tumor array, when we decrease the parameter a in equation (3.6) from 1 to 0, the corresponding $\widehat{\mathsf{FDR}}$ increases from 0 to 1. Then we pick the a which makes the $\widehat{\mathsf{FDR}}$ first cross over at some certain level, for example 1%. Thus, if there are m selected genes under this a, we will have the confidence that more than $0.99m$ genes among the m selected ones should be true significant.

However, to get a reliable approximation of the H_0 distribution, some well matched reference normal/normal hybridization arrays are needed. Here, 'well matched' means two arrays are produced under the same experimental conditions and of the same quality (e.g. spots are filtered under the same criteria).

For an array of normal/normal hybridization, the measurements can be deemed as some random noise fluctuating around zero. Thus, the joint distribution of (`rd`, `mean value`, `lsize`) is affected mainly by the standard deviation (SD) and the average fluctuation ($= \sum_{i=1}^{N} |x_i - x_{i+1}|/N$) of the array. Since the

algorithm adjusts the SD of each tumor array according to the SD of the normal arrays, the ratio of the average fluctuation over SD is the only key factor. After we studied a few normal/normal hybridization arrays from different samples of different experiments (different batch, different labs ...), we find that this ratio is mainly affected by experiment conditions. The differences between the ratios are smaller than 0.02 for arrays from the same batch/lab/time. In other words, the normal arrays from the same experiment condition as the tumor arrays provide reliable approximation for H_0, while the biological difference of different normal samples would not matter significantly.

4.2 FDR *for the consensus summary*

Normally, more than one sample (i.e. different patients diagnosed with the same cancer) would be collected in an array CGH study. The more arrays in which one gene demonstrates an alteration, the more likely this gene is a potential oncogenic or tumor suppressor gene. Therefore, after we select gain/loss regions for each single array, it is always helpful to look at the consensus result across all samples. For this we need to have an estimator of FDR for the consensus summary.

Here we use another version of FDR: positive False Discovery Rate (Storey, 2002). Compared with equation (4.1), pFDR is defined as

$$p\mathtt{FDR} = E\left(\frac{V}{R}\middle| R > 0\right) = \mathtt{FDR}/P(R > 0). \tag{4.4}$$

pFDR and FDR would be similar to each other when $P(R > 0)$ is close to 1, i.e. the event that no significant genes are selected is rare. With Theorem 1 of Storey (2002), the pFDR of our problem can be denoted as

$$p\mathtt{FDR}_K = P(\text{gene is from } H_0 | \text{gene is selected in at least } K \text{ arrays}). \tag{4.5}$$

Suppose the total number of arrays is m, the total number of genes in one array is N. Denote the jth gene in the array as g_j, and the set of selected genes for the ith array as S_i. In addition, pFDR for the selected result of the ith array is

$$p\mathtt{fdr}_i = P(g \in H_0 | g \in S_i), \tag{4.6}$$

where g is any random picked gene in the array. It follows that

$$p_i = P(g \in S_i | g \in H_0) = \frac{p\mathtt{fdr}_i \cdot P(g \in S_i)}{P(g \in H_0)}. \tag{4.7}$$

One assumption we make here is that given $g_j \in H_0$, $g_j \in S_{i_1}$ is independent with $g_j \in S_{i_2}$. This is quite reasonable if all the samples (patients) are assumed to be independent from one another.

Suppose there are K_j arrays in which g_j demonstrates an alteration, i.e. $K_j = \#\{S_i : g_j \in S_i\}$. Denote

$$A_k = \{g_j : K_j \geqslant k\}, \tag{4.8}$$

then we have

$$P(g \in H_0 | g \in A_k) = \frac{P(g \in H_0)}{P(g \in A_k)} P(g \in A_k | g \in H_0). \tag{4.9}$$

When $p_i = p$ for all $i = 1, 2, \ldots, m$, it is easy to see that

$$P_k = P(g \in A_k | g \in H_0) = P(\text{binomial}(m, p) \geqslant k), \tag{4.10}$$

while if p_i are not equal across samples, P_k is still computable, though a little complex.

Therefore, the $p\mathtt{FDR}$ for the consensus summary can be estimated by

$$\widehat{p\mathtt{FDR}_k} = \widehat{P}(g_j \in H_0 | g_j \in A_k) \tag{4.11}$$

$$= \frac{\widehat{P}(g_j \in H_0)\widehat{P}_k}{\widehat{P}(g_j \in A_k)} \approx \frac{1 \cdot \widehat{P}_k}{\#\{A_k\}/N}. \tag{4.12}$$

$\widehat{P}_k$ can be calculated by using $\widehat{p\mathtt{fdr}_i} \approx \widehat{\mathtt{fdr}_i}$ of each sample, and $\widehat{P}(g_j \in S_i) = \#\{S_i\}/N$.

5. Real data study

5.1 *cDNA microarray CGH*

In the Lung Cancer study carried out by Young Kim and Jonathan Pollack (Department of Pathology, Stanford University; unpublished), array CGH hybridization experiments were performed to 48 lung cancer cell lines.

cDNA microarrays were obtained from the Stanford Functional Genomics Facility and contained PCR-amplified cDNAs representing 25 736 different mapped human genes (i.e. UniGene clusters). The average clone spacing is 60kB, though the spacing is more (less) in gene-rich (gene-poor) regions of chromosomes. The reference sample was sex-matched normal leukocyte DNA.

After raw fluorescence intensities for array spots were extracted, the data were preprocessed as follows. First, background-subtracted fluorescence ratios were normalized for each array by setting the average fluorescence log ratio for all array elements equal to 0; second, reliably measured genes were selected, whose average fluorescence intensity for the normal DNA reference channel was at least 1.4-fold above background; third, map positions for arrayed cDNA clones were assigned using the NCBI genome assembly, accessed through the UCSC genome browser (July 2003 freeze); in the end, for genes represented by multiple arrayed cDNAs, average fluorescence ratios were reported.

Then we applied both the CLAC method and the threshold method (Pollack *et al.*, 2002) to this data set. The results for chromosome 1 of an example array (the third tumor array in Figure 1) are shown in Figure 4. Clearly, the threshold method only picks out the few strongest signals, while not being able to recognize any contiguous region of potential amplification or deletion. In contrast, CLAC method successfully identifies both the localized amplification (the big spike region of $1p$) and the possible chromosome arm gain ($1q$).

Moreover, Figure 5 illustrates the result of CLAC for all the three array experiments shown in Figure 1. We can see that for the XY/XX hybridization, CLAC has accurately identified the one X chromosome copy loss region. The result of the tumor array is also quite impressive. It agrees very well with visual assessment of one expert pathologist (Pollack).

Among the 48 cell lines, there are two different types of lung cancer; one is called 'small cell lung cancer (SCLC)' while the other is referred to as 'non-SCLC'. The summary plot over these two different types is shown in Figure 6. CLAC identifies both common and distinct regions of gain/loss for SCLC and NSCLC specimens.

5.2 *BAC array CGH*

To further assess the performance of CLAC, we apply it to a BAC array data set with known DNA change numbers (Snijders *et al.*, 2001). We use the four arrays in Web Table E (GM00143, GM01750,GM01524,GM01535), downloaded at `http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html`.

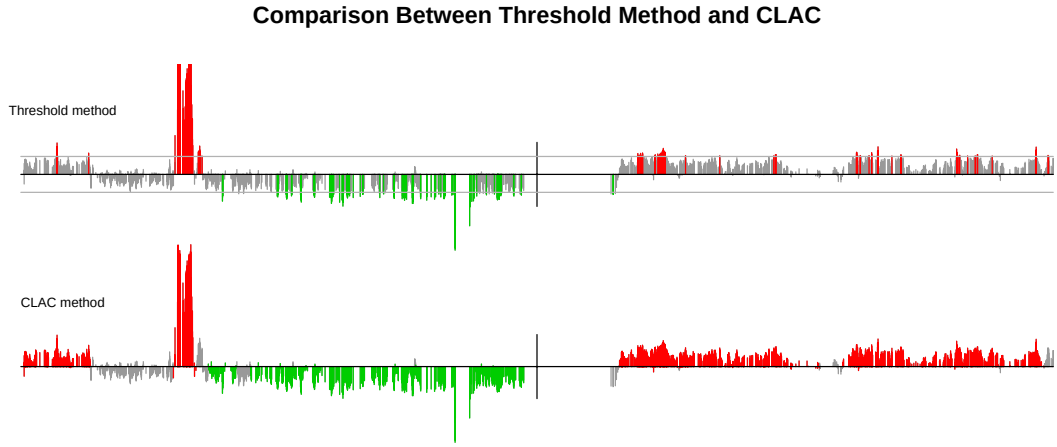

Fig. 4. Comparison of threshold method with CLAC. The picture shows the results of both CLAC method and the threshold method on chromosome 1 of an example array (the third tumor array in Figure 1, also with moving average of 5). The selected gain/loss regions are plotted in red and green, while the other regions are in gray. The black vertical bar in the middle illustrates the location of the centromere of this chromosome. Threshold method: the two horizontal gray lines stands for the upper cut bound 0.28 and the lower cut bound −0.306; the corresponding FDR is 0.04. CLAC method: the corresponding FDR is 0.009.

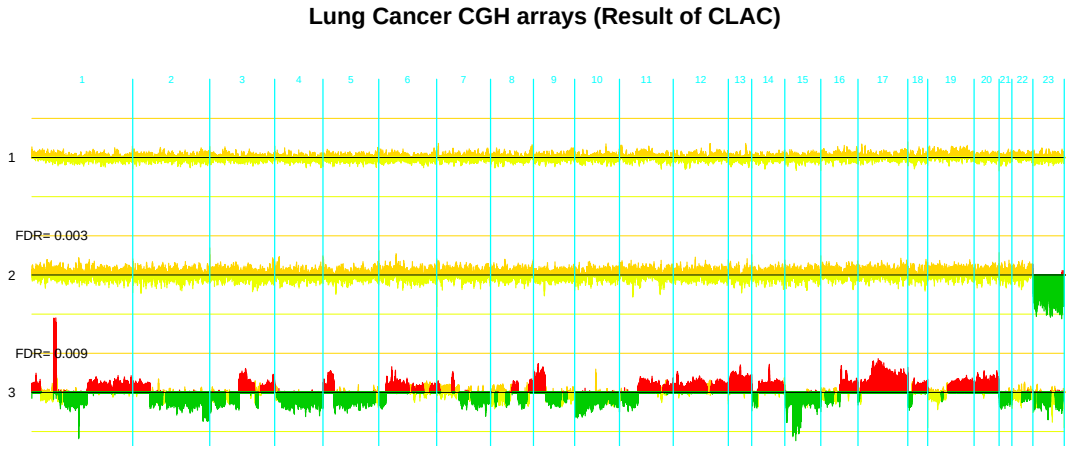

Fig. 5. Result of CLAC. The selected gain/loss regions are plotted in red and green, while the other regions are in orange and lime green.

Each array contains measurements for around 2700 BAC clones. Since there is no reference normal array available, we make three ‘pseudo normal’ arrays from Web Table F (GM02948,GM03134,GM03563,GM03576) by throwing away the measurements outside 97% quantiles. Comparing the result (Figure 7) with the true copy number alterations (Web Table I), we can see that CLAC accurately picks out all the trisomic chromosomal regions. However, the FDR is overestimated here, which is due to using ‘pseudo normal’ arrays instead of true normal arrays. Again, in Figure 7, we can see that there are quite a few other noise spikes which would fail the threshold-type methods.

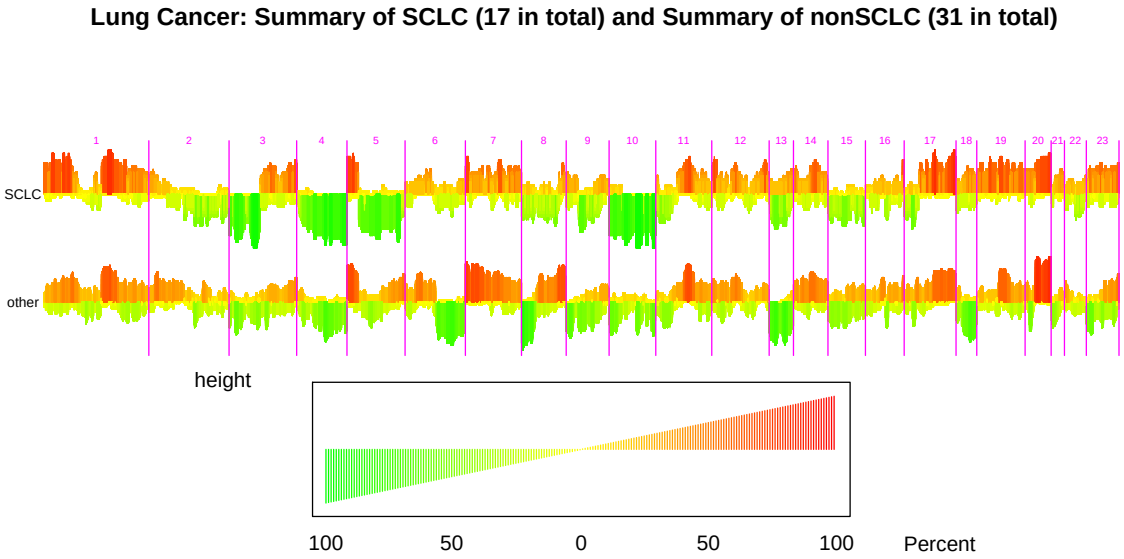

Fig. 6. Consensus summary plot for the two groups. The picture at the top is the summary across 17 SCLC cell lines. The picture at the bottom is the summary across 31 non-SCLC cell lines. The heights and the color of the vertical lines represent the percentage of samples in which the corresponding genes have DNA copy number alteration.

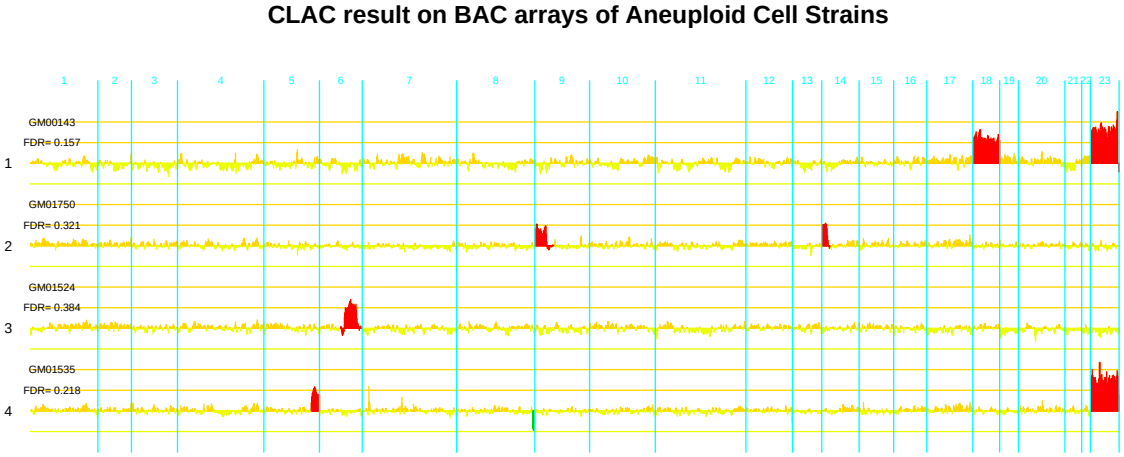

Fig. 7. Result of CLAC on BAC arrays (with moving average of 3). The selected gain/loss regions are plotted in red and green, while the other regions are in orange and lime green.

6. Discussion

The CLAC method is a potentially useful technique for calling gains and losses in array CGH data. It is conceptually simple and automatic, allowing estimation of its false discovery rate over an entire array or collection of arrays. This last property is important, since the chance of false positive calls is large when scanning large numbers of sites.

The type of approach used here—spatial clustering followed by node selection—may be useful in

other genomic problems where a signal is measured along a physical axis. One example is protein mass spectroscopy, in which the intensity of proteins is estimated at many different mass/charges values.

The software 'CGH-Miner' provides both the R-package and an Excel add-in for CLAC algorithm, which is available at `http://www-stat.stanford.edu/~wp57/CGH-Miner`

ACKNOWLEDGMENTS

We would like to thank two editors and two referees for comments that improved this manuscript. Pei Wang was partially supported by the Stanford Graduate Fellowship. Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183. Jonathan Pollack was supported in part by NIH grant R01CA97139.

APPENDIX

In this section, we discuss the choice of parameters b, k_1, k_2, b_1,and b_2.

First we discuss the choice of the parameter b which determines the position of the quadratic cut line for the joint distribution of (`rd`, `lsize`). Note that the distribution of `rd` does not depend on the scale of the noise of each array. If we recall the definition

$$\texttt{rd}(x_i, x_j) = \frac{|x_i - x_j|}{|x_i| + |x_j| + |x_i + x_j|}, \tag{A.1}$$

it is easy to see that

$$\texttt{rd}(\alpha \cdot x_i, \alpha \cdot x_j) = \texttt{rd}(x_i, x_j). \tag{A.2}$$

Consequently, the choice of b can be independent of the noise level of each single array. Therefore, we use a fixed value 0.55 for b in our later study. The meaning of 0.55 can be understood as (without loss of generality, suppose $|x_i| > |x_j|$)

$$\texttt{rd}(x_i, x_j) > 0.55 \Longrightarrow x_i \cdot x_j < 0, \text{ and } |x_j| > \frac{|x_i|}{10}; \tag{A.3}$$

in other words, for the chosen gain/loss regions, we would not tolerate a noise level that is more than $1/10$ of the single level.

As to k_1, k_2, b_1,and b_2, we first fit lines

$$l_{up} : y = k_1^0 \cdot x + b_1^0,\ l_{low} : y = k_2^0 \cdot x + b_1^0 \tag{A.4}$$

to the boundary of the empirical joint distribution of (`meanvalue`, `lsize`) of the normal array (Figure 3), which is an approximation to the null distribution of (`meanvalue`, `lsize`). However, for different arrays, the null distributions of `meanvalue` might be different, which depends on the scale of noise level of each array. So we need to consider

$$\alpha = \frac{sd(\text{noise of the tumor array})}{sd(\text{noise of the normal array})}. \tag{A.5}$$

To the normal–normal array hybridization, all the measurements other than 0 can be deemed as noise, so

$$\widehat{sd}(\text{noise}) = sd(\text{array}). \tag{A.6}$$

To estimate the sd(noise) for a tumor array, we use the following procedure. First cut all the clustering trees of the array at a height of 0.5, and then pick the clusters with `size` > 15 and $|$`meanvalue`$| < 1$. The genes in these clusters can be deemed as the 'normal' region of this array. Actually, from the definition of `rd`, it is easy to see that

$$x : y > 0 \Leftrightarrow \texttt{rd}(x, y) < 0.5.$$

Therefore, the cutoff `rd` < 0.5 separates the genome into segments which are either all positive or all negative. The measurements of noise regions always fluctuate around 0 frequently, while, for gain/loss regions, measurements tends to stay positive/negative more consistently. Thus those segments with small size should be good representations of noise regions. This can also be seen in Figure 2. Then, deeming these small segments as 'pseudo normal' regions, we can calculate the estimator of sd(noise) for the array.

It follows that

$$\widehat{sd}(\text{noise}) = sd(\text{picked genes in 'normal' region}), \tag{A.7}$$

and

$$(k_1, b_1, k_2, b_2) = \alpha \cdot (k_1^0, b_1^0, k_2^0, b_2^0). \tag{A.8}$$

REFERENCES

AUTIO, R., HAUTANIEMI, S., KAURANIEMI, P., YLI-HARJA, O., ASTOLA, J., WOLF, M. AND KALLIONIEMI, A. (2003). Cgh-plotter, Matlab toolbox for CGH data analysis. *Bioinformatics* **19**, 1714–1715.

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 B* 289–300.

CHENG, C., KIMMEL, R., NEIMAN, P. AND ZHAO, L. P. (2003). Array rank order regression analysis for the detection of gene copy-number changes in human cancer. *Genomics* **82**, 122–129.

EFRON, B. AND TIBSHIRANI, R. (2002). Microarrays, empirical bayes methods, and false discovery rates. *Genetic Epidemiology*.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. New York: Springer.

HODGSON, G., HAGER, J., VOLIK, S., HARIONO, S., WERNICK, M., MOORE, D., NOWAK, N., ALBERTSON, D., PINKEL, D., COLLINS, C. *et al.*, (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **29**, 491.

JONG, K., MARCHIORI, E., VAART, A., YLSTRA, B., WEISS, M. AND MEIJER, G. (2003). Chromosomal breakpoint detection in human cancer. **2611**, 54–65.

LENGAUER, C., KINZLER, K. AND VOGELSTEIN, B. (1998). Genetic instabilities in human cancers. *Nature* **396**.

PINKEL, D., SEGRAVES, R., SUDAR, D., CLARK, S., POOLE, I., KOWBEL, D., COLLINS, C., KUO, W., CHEN, C., ZHAI, Y. *et al.*, (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**.

POLLACK, J., SORLIE, T., PEROU, C., REES, C., JEFFREY, S., LONNING, P., TIBSHIRANI, R., BOTSTEIN, D., BORRESEN-DALE, A. AND BROWN, P. (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**, 12963–12968.

SNIJDERS, A., FRIDLYAND, J., MANS, D., SEGRAVES, R., JAIN, A., PINKEL, D. AND ALBERTSON, D. (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* 263–264.

STOREY, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society* 479–498.

TUSHER, V., TIBSHIRANI, R. AND CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences, USA* **98**, 5116–5121.

[*Received February 9, 2004; revised April 23, 2004; accepted for publication May 21, 2004*]