Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2011, Article ID 860732, 15 pages
doi:10.1155/2011/860732

Research Article

A New Normalizing Algorithm for BAC CGH Arrays with Quality Control Metrics

Jeffrey C. Miecznikowski,[1,2] Daniel P. Gaile,[1,2] Song Liu,[1,3] Lori Shepherd,[1,2] and Norma Nowak[4,5]

[1] *Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, USA*
[2] *Department of Biostatistics, Roswell Park Cancer Institute, Buffalo, NY 14263, USA*
[3] *Department of Bioinformatics, Roswell Park Cancer Institute, Buffalo, NY 14263, USA*
[4] *Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY 14263, USA*
[5] *Department of Biochemistry, University at Buffalo, Buffalo, NY 14214, USA*

Correspondence should be addressed to Jeffrey C. Miecznikowski, jcm38@buffalo.edu

Received 1 July 2010; Revised 23 November 2010; Accepted 18 December 2010

Academic Editor: Adewale Adeyinka

The main focus in pin-tip (or print-tip) microarray analysis is determining which probes, genes, or oligonucleotides are differentially expressed. Specifically in array comparative genomic hybridization (aCGH) experiments, researchers search for chromosomal imbalances in the genome. To model this data, scientists apply statistical methods to the structure of the experiment and assume that the data consist of the signal plus random noise. In this paper we propose "SmoothArray", a new method to preprocess comparative genomic hybridization (CGH) bacterial artificial chromosome (BAC) arrays and we show the effects on a cancer dataset. As part of our R software package "aCGHplus," this freely available algorithm removes the variation due to the intensity effects, pin/print-tip, the spatial location on the microarray chip, and the relative location from the well plate. removal of this variation improves the downstream analysis and subsequent inferences made on the data. Further, we present measures to evaluate the quality of the dataset according to the arrayer pins, 384-well plates, plate rows, and plate columns. We compare our method against competing methods using several metrics to measure the biological signal. With this novel normalization algorithm and quality control measures, the user can improve their inferences on datasets and pinpoint problems that may arise in their BAC aCGH technology.

1. Introduction

Pin-tip microarray technology was invented in the early 1990s [1]. The technology has grown tremendously, and now there are numerous types of probes and target elements. Target elements can include genes, oligonucleotides, or bacterial artificial chromosomes (BACs) and new microarray chips can contain on the order of a hundred thousand probes. Due to the technology, the signal obtained is a combination of the biological signal and technological signal. Specifically, we will focus on BAC CGH microarrays. Array-based comparative genomic hybridization (aCGH) technology is similar to cDNA arrays and is an extension from conventional CGH that is used to identify and quantify DNA copy number changes across the genome in a single experiment [2]. The advantages of aCGH include high-resolution and high-throughput measurement capability allowing for more quantitative analysis of the genomic aberrations. A thorough introduction to the design and manufacture of microarrays is provided in [3] while [4] provides an introduction to the statistical issues in analyzing microarray datasets.

In BAC aCGH, the probes corresponding to locations on a genome are cloned (grown) in a bacterial culture and then arrayed to a glass slide. BAC aCGH technology can be employed to discover markers in diseases as in [5–8] and for detecting genomic imbalances in cancers as described in [9–20]. In BAC aCGH studies, the markers for cancer are often discovered by comparing the signal at

a given chromosome loci between the tumor sample and a control sample. Specifically, researchers often examine the logarithm (base 2) of the ratio of the tumor sample to the control sample ($\log T/C$). This value will allow researchers to determine the presence of an imbalance in copy number for a given marker between the tumor sample (T) and the control sample (C).

It is necessary to normalize the raw $\log T/C$ values before subsequent analysis to determine regions of chromosomal imbalance. Normalization procedures have been recognized as necessary for microarray experiments, and a recent search on PubMed (http://pubmed.org) reveals over 500 references to articles on microarray normalization. For two-channel microarrays, some of the normalization algorithms similar to the proposed analysis in this paper are cited in [21, 22]. Specifically, with regards to aCGH datasets, recent normalization algorithms are proposed in [23–27] and in R software packages [28, 29]. Our normalization approach for the $\log T/C$ data will be similar to the approach in [23] but will feature several important differences in the estimation procedures for the technical effects.

Specifically, the goal of this paper is to isolate the biological signal in the $\log T/C$ by removing the technological signal via the novel "SmoothArray" normalizing process. The technological signal is composed of three major components: (1) signal due to the intensity of each scanning channel, (2) signal due to spatial (array) location, and (3) signal due to the spotting technology. We remove each signal sequentially thus isolating the biological signal. We compare our procedure against other normalization procedures and use several metrics to measure the improvements of our method. Although our results can be applied to any print-tip microarray setting, our examples were obtained from the Roswell Park Cancer Institute (RPCI) aCGH microarray facility. Note our software is written in the R programming language [30] and is freely available at [31].

2. Materials and Methods

2.1. The Arrayer Procedure. In the RPCI aCGH microarray facility, differentially labeled total genomic DNA from a "test" and a "reference" cell population are cohybridized to the BAC clones. After hybridization, a GenePix Axon scanner generates two images of the array at the wavelengths of light corresponding to the two dyes (Cy3 and Cy5). The images are processed to generate a single number corresponding to each sample (dye) for each spot on the array. For the RPCI facilities, GenePix is currently used to perform the image processing. The resulting ratio of the fluorescent intensities at a location on the chromosomes is approximately proportional to the ratio of the copy numbers of the corresponding DNA sequences in the test and reference genomes. A traditional experiment describing tumor extraction, preparation, and so on is described in [32].

For our BAC aCGH studies, the RPCI 19 K BAC array was utilized containing ~19,000 BAC clones (probes) that were chosen by virtue of their STS content, paired BAC end-sequence, and association with heritable disorders and cancer. Reference and test sample genomic DNA (1 μg each) were individually fluorescent labeled using the BioArray CGH Labeling System (Enzo Life Sciences) as described in [33]. The hybridized BAC-based aCGH slides were scanned using a GenePix 4200AL Scanner (Molecular Devices) to generate high-resolution (5 μm) images for both Cy3 (test) and Cy5 (control) channels.

In bacterial artificial chromosome (BAC) aCGH technology, the target DNA elements are physically arrayed in a two-dimensional grid on a chemically modified glass slide. Note that the BAC clones are stored in freezers on a total of 51 plates (384 wells per plate). A potentially large source of technical variation may be present because of these plates.

Another source of potential variation is due to the pin array process of printing the BAC clones on the glass microarray slide. For our data, the 48 pins in the arrayer are arranged in a 12×4 matrix structure, approximately 4.5 mm on center, so that they transport the probes to the slide where each pin fills one region or "grid" of the array. The spots are approximately 80 μm in diameter, with respective centers 150 μm apart from each other to ensure no overlap between spots.

The array has the spots laid out in a 116×348 array of 40368 spots. More specifically, each of the grids within the array (corresponding to pin number) has dimensions 29×29 thus there are 841 spots per grid (pin). The array's spot locations are consecutively labeled row-wise within each pin, first numbering within Pin 1 (1–841), followed by the spots within Pin 2 (842–1682), and so forth. Thus, the spot location values range from 1 to 40368. Due to this geometry, each BAC clone is repeated on the array; in other words, each clone has two spots on each array. Further, note that there are $384/48 = 8$ spots per grid per plate. Since $8 \times 51 \times 2 = 816$ and each grid has 841 spots, there are $841 - 816 = 25$ blank spots in each grid. Each plate is used twice, as each spot is replicated within a grid on the array. Put another way, this procedure can produce the intensity levels of $40,368/2 - 25 \times 48 = 18984$ BAC clones per array arranged in a two-dimensional array on the slide that accommodates up to 40368 spots. The remaining $25 \times 48 = 1200$ spot locations remain unused and are, therefore, not considered in this analysis (note that however, these unused spots may contain valuable information regarding the background and laser scanner settings).

2.2. Data Acquisition. The data summary gives the intensity readings from the Cy3- and Cy5- labeled genetic material for each spot, as produced by the image processing software. Since our dataset is related to oncology research, we will refer to the data in terms of the tumor channel (T) and the control channel (C). For spot i on the array, we will be interested in the logarithmic ratio of the tumor channel (T_i) to the control channel (C_i). In other words, we will focus on $M_i \equiv \log_2 T_i/C_i$ for a given spot i. We will define the vector M as the collection of M_i values for a given sample. From [34] there are three major sources of possible systematic variation in M which are a consequence of the experimental procedure and do not contribute to differential expression. The first source of variation is due to the intensity effects. This bias is evident from Figure 2 and is a noted source of variation in

Table 1: Each step in the "SmoothArray" normalization process for CGH microarray experiments. Each step is detailed in Section 2.3.

Bias	Normalization method	Description
Intensity	Global Loess	Fit M according to intensity values A
Spatial	Spatial Kernel smoother	Fit a spatial smoother to M values after CBS [37]
Spotting process	Median	Compute a median to estimate the pin, plate, plate row, and plate column effects

two-channel microarrays [35]. In the data shown in Figure 2, this bias predominately produces a curvature shape where the lower intensity probes tend to have a large log T/C value.

The second source of variation is the physical layout on the glass slide; one can imagine that there are spatial effects across the slide (caused, e.g., by the way the dye-labeled material is hybridized to the slide) which would manifest as a pattern of row and/or column effects if the data were analyzed as 348×116 array. The third source of variation stems from the 384 well plates which are the source of the spots on the glass slide; one can imagine that there are effects which are localized to one (or more) specific plates which would appear as localized effects on the glass slide. As stated in [23, 36], this bias "may be caused by the fact that different clones that are produced in the plates might have experienced slightly different physical conditions during the polymerase chain reaction (PCR) or in subsequent purification steps."

Also there are potential effects due to the 384 well-plate rows (16 rows) and 384 well plate columns (24 columns). Note that the localization is complicated because of the arrayer procedure described above; recall the complex numbering scheme. There may also be a source of variation due to the pins themselves. One can easily imagine that the pins vary in shape, head size, or some other property that causes the observations to vary from quadrant to quadrant on the array. Equally, one can imagine a serial (in time) correlation among the observations caused by, for example, the pins not being adequately cleaned between successive dips into the wells on the plates. This is not intended to be an exhaustive list of possible sources of systematic variation, but rather simply a short list of obvious possibilities. The key point here, and in all subsequent analysis, is that we assume a random spatial distribution of the probes on the microarray chip. In other words, we assume that there is no correlation between a probe's genomic location and its spatial coordinates on the array. This random assumption is required in order for our normalization method to preserve the biological signal present in the chip.

With this layout and the potential sources of technical variation, we can define the "SmoothArray" algorithm to preprocess BAC CGH arrays. Note the goal of our "SmoothArray" algorithm will be to remove the technical variation in the log T/C values. The data used to demonstrate the process consist of 219 head and neck tumor samples obtained from the RPCI microarray facility.

2.3. The SmoothArray Algorithm. The "SmoothArray" process consists of the steps described in Table 1 applied sequentially to M, where M denotes the vector of M_i spot values for a specific array. Let A denote the vector of A_i values, where

$$A_i \equiv \log_2(T_i \times C_i), \quad i = 1, 2, 3, \ldots, 37968. \tag{1}$$

A is the vector of logarithm products for the two channels in an CGH microarray experiment. In all subsequent analysis, we will use a typical sample from a dataset designed to examine head and neck tumors as obtained from the RPCI microarray facility.

The M vector of values from the scanner represents the input values for the "SmoothArray" algorithm. Figure 1 represents the input data for "SmoothArray." Figure 1 shows the raw dataset (M values) in the form of an array image and a genomic plot. Note that since each BAC clone is printed twice on the array, the genomic plot is obtained by averaging the two M_i values for each BAC clone. The representation of M values and the ranked M values based on their location on the array image is a good way to view the spatial bias present on an array. The plot shown in Figure 1(a) is commonly referred to as an M-XY plot. By design, each sample is hybridized against a sex mismatch hence, as seen in the genomic plot (Figure 1(b)), there is a gain (increase) in the ratio on the X chromosome mimicking a single-copy gain. The data is not centered around 0 in Figure 1(b) indicating a bias in the log T/C values.

2.3.1. Intensity Effect Step. The first step in the "SmoothArray" algorithm employs a loess smoother on the M_i values using the A_i values as the explanatory variable. In short, a loess smoother fits a polynomial surface determined by the set of explanatory variables. We consider this a global operation since the entire set of probes from an array is used in the loess fitting function, regardless of position on the array or genome. Fitting is by (weighted) least squares using the "loess" function in the R package *stats* [30]. The result from this operation is a set of fitted values, $\mathrm{GL}(M_i)$, according to the A_i values, where we denote the loess function as GL(). The fitted values, $\mathrm{GL}(M_i)$, represent the intensity bias present at spot i due to the probe intensity A_i. By subtracting the fitted values $\mathrm{GL}(M_i)$ we account for this technological bias. Hence, we carry forth the M' vector to the next step, where M' is a vector consisting of

$$M_i' = M_i - \mathrm{GL}(M_i), \quad i = 1, 2, 3, \ldots, 37968. \tag{2}$$

M_i' represents the signal for spot i after removing the technological signal due to the product of the intensities from the two channels. The goal in the next step is to remove the spatial bias present in the aCGH technology.

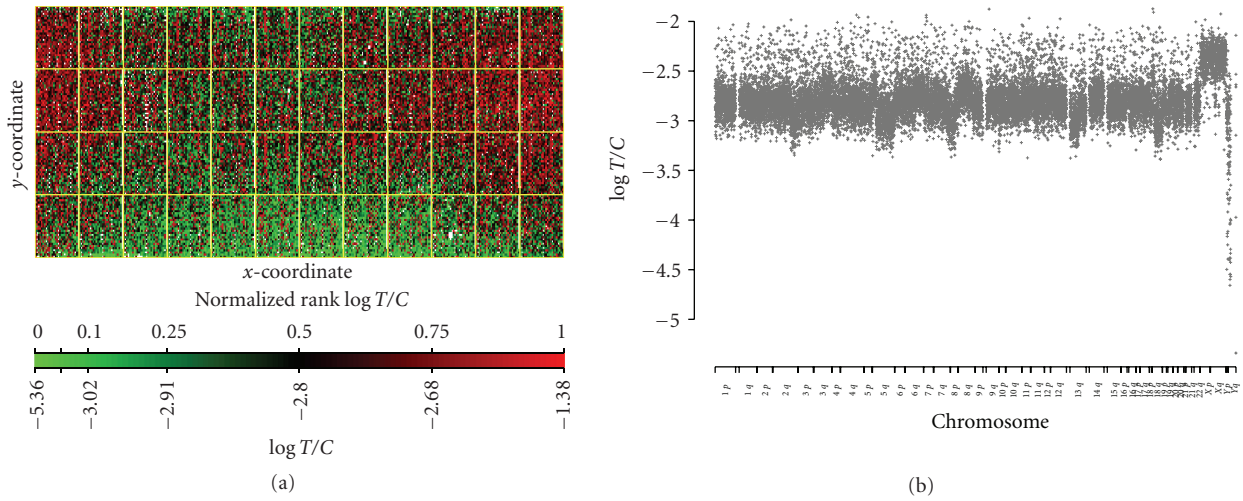

Figure 1: Raw Data: (a) Image of log T/C signal arranged in an image format corresponding to location on the microarray. (b) The aCGH log T/C data as arranged in chromosomal order. Since each BAC clone is spotted twice on the array, the genomic plot is created by averaging the two log T/C values for each BAC clone. This data corresponds to the raw normalization method where no normalization is performed.

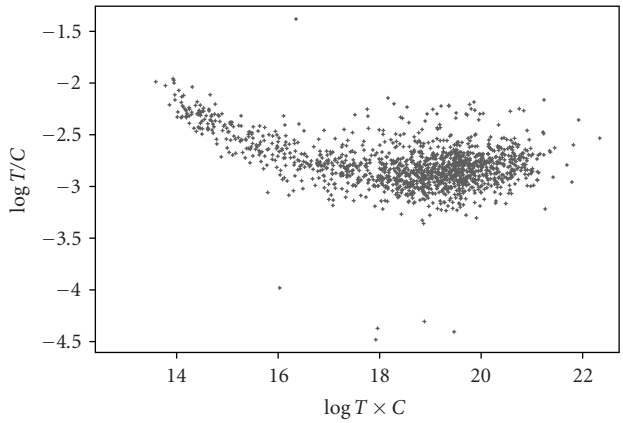

Figure 2: Intensity effect: the estimated loess fit (red line) using the log T/C as the response with log $T \times C$ as the explanatory variable.

2.3.2. Spatial Effect Step. The next step in the algorithm removes the technological noise due to the spatial location in the array. It is reasonable to expect nearby spots to be correlated with each other due to the reagents process and the hybridization process in microarray technology. The goal in this step is to accurately determine the spatial pattern present in the array and thus remove it. The data depicted in Figure 3 is the starting point for the spatial smoothing step.

Rather than perform the spatial smoothing on the data depicted on Figure 3, we will modify our approach to ensure we preserve the biological signal. The biological signal present in aCGH data can be captured using the circular binary segmentation (CBS) algorithm described in [37]. In short, the CBS algorithm can be applied to cluster the M' values into segments of estimated equal copy number according to their location on the genome [37]. After CBS, each probe i is a member of a specific segment where we will denote CBS(M_i') as the log T/C group mean for the segment containing probe i. We apply the CBS function (denoted by CBS()) to the values in Figure 3 and compute the residuals from the CBS operation as $M' - \text{CBS}(M')$. The spatial two-dimensional kernel density smoother is applied to the residuals from the CBS operation. Note that, by subtracting the CBS group we are, in a sense, removing the genome/biological signal to ensure that our kernel density estimate of the spatial signal has a minimal amount of biological signal and is only modeling the technical variation. This is one of the key ways in which our method differs from the algorithm described in [23]. In short, we apply the kernel density smoother to the vector $M' - \text{CBS}(M')$. The two-dimensional kernel density smoothing is performed using the function "smooth.2d" in the *fields* library in R [38]. The smoothing parameter (bandwidth) is chosen via a cross-validation (CV) procedure. Namely, a random subset is removed from the data and the surface is fit. After fitting a surface, the absolute value loss (L1 loss) or the sum of squares loss (L2 loss) for the random subset is computed. The value of the smoothing parameter that yields the smallest sum of squares is used as the optimal value in the kernel smoothing algorithm. After estimating the spatial technical variation (Figure 4(a)), we compute the resulting set of log T/C values with the spatial variation removed. In other words for spot i, we compute

$$M_i'' = M_i' - \text{KS}(M_i' - \text{CBS}(M_i')), \tag{3}$$

where KS represents the kernel smoothing function. M_i'' represents the log T/C value for spot i after the technical signal due to intensity and the spatial location has been removed.

2.3.3. Spotting Process Effect Step. At this point, we remove the technological signal due to the spotting of the array. The technical issues of the spotting procedure noted in [22, 34, 39] demonstrate a solution for CGH microarray

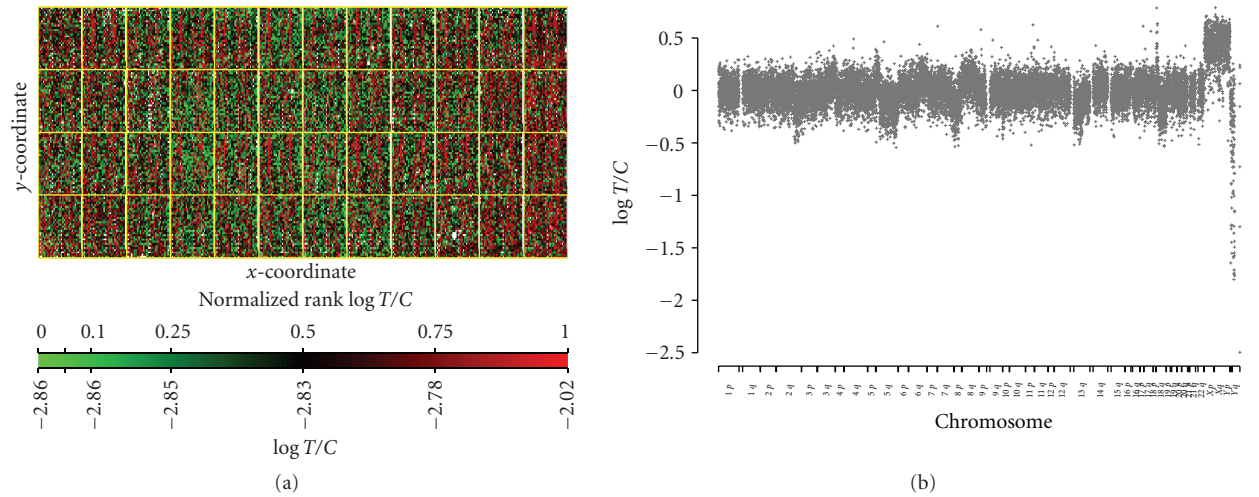

Figure 3: Intensity correction results: The aCGH data from Figure 1 after removing the intensity effects shown as a function of (a) location on the array (b) and genomic location. The genomic profile was created by averaging the signal for the two replicates of each BAC clone.

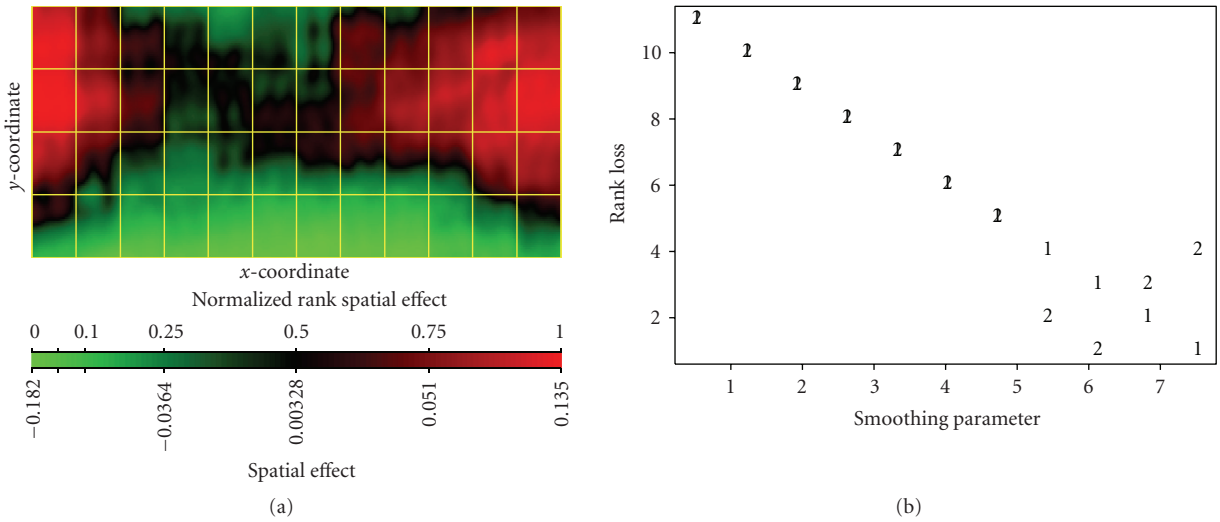

Figure 4: Spatial effect: (a) the estimated spatial effect as determined by a kernel density smoother. (b) The ranked cross-validation error as a function of the smoothing parameter (bandwidth) for the kernel density smoother. The cross-validation error is shown as an absolute error loss ("1") and as a squared error loss ("2").

chips. Similar to determining the kernel smoothed spatial surface, we determine the pin, plate, plate row, plate column and repetition effects for the vector M'' populated with M_i'' for spot i. Similar to the estimation of the spatial effect, we will estimate the spotting process effects on the dataset where we preserve the biological signal by employing the CBS algorithm. We apply the CBS function to the M'' data and then compute the residuals by subtracting the CBS segment mean. The residuals are used to estimate the effects of the spotting process. We compute a median of the residuals for each pin, plate, plate row, plate column, and repetition value and use that as our estimate for the spotting process effect. We obtain the final set of log T/C values representing the remaining biological signal by subtracting the effect of the pin, plate, plate row, plate column and repetition effect from M'', that is, the vector of log T/C values after accounting for intensity and spatial effects.

2.4. Other Normalization Methods. To assess our "SmoothArray" procedure, we compare it with five other normalization methods. Specifically, we examine raw, grid loess, background subtraction, global median normalization, and quantile normalization. In the raw normalization method, we use the raw log ratios without performing any normalization. In a grid loess procedure, within each pin grid, a local polynomial regression fit [40] is performed on the M_i values using the A_i values as the explanatory variables. The normalized values from this

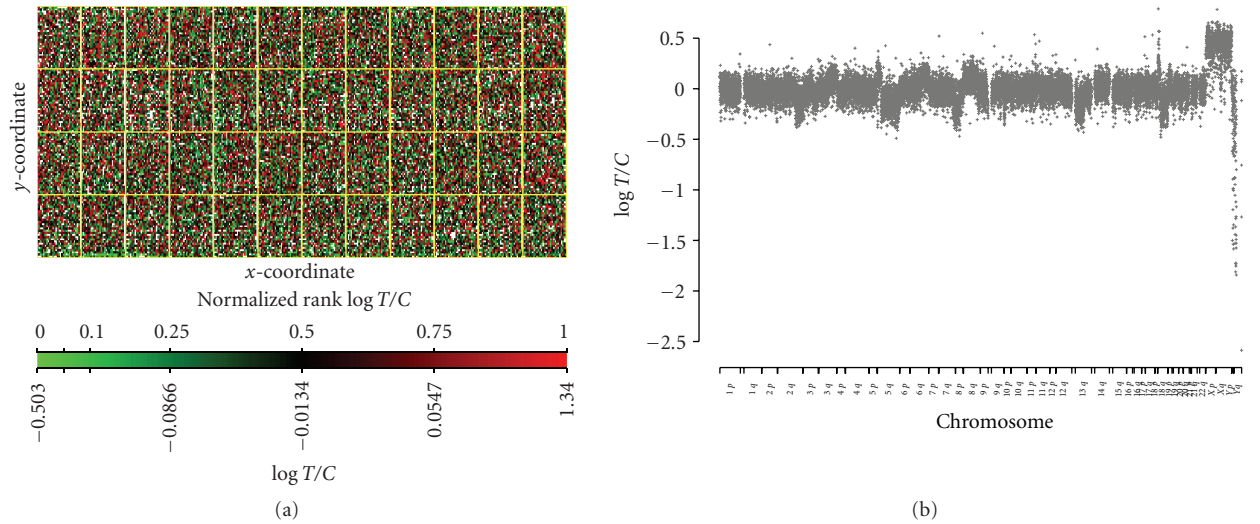

FIGURE 5: Spatial correction results: (a) The array image after spatial correction (and intensity correction). (b) The genome plot of log T/C after spatial correction (and intensity correction).

procedure are obtained by taking the difference between the raw values (M_i) and the fitted loess values. The loess fitting is done using the "loess" function in the R package *stats* [30]. The background subtraction method is employed by subtracting the estimated background intensity from the estimated foreground intensity of each spot before taking the logarithm ratios. The global median normalization procedure is performed by estimating the median log ratio on each array and then subtracting this value from the log ratios on the array. With a global median normalization procedure, the empirical (sample) median of the normalized log ratios for each array is zero. The quantile normalization procedure is enacted according to the procedure outlined in [41]. In this procedure the goal is to impose the same empirical distribution of log ratios to each array. A mean array is created by taking the sample mean across the sorted values of the log ratios in each array. Then the distribution of the log ratios in the other arrays in the experiment is matched to the empirical distribution of log ratios in this hypothetical mean array. In the Results Section we compare these normalization methods with the "SmoothArray" normalization method.

3. Results

As a tool to quantify the results at each step of "SmoothArray" in terms of reduction of noise we employed the median absolute deviation (MAD) on the X chromosome. The signal was estimated by the MAD on the X chromosome since the X chromosome, by design, was always altered by the virtue of the sex-mismatched controls. Further, we do not expect there to be any disease-specific imbalances to occur on the X chromosome. The MAD as calculated on the X chromosome acts as a good measure of the performance of "SmoothArray" and other preprocessing algorithms since the genomic log T/C data is ultimately used to call regions of genomic imbalance, and this measure is based on a subset of the genomic data. Also note that none of the "SmoothArray" steps require knowledge of the genomic location for a given spot. Hence by using this metric to evaluate "SmoothArray" we are assured that the improvement (reduction) in this measurement must be the result of the removal of technical variation and not biological variation. For the chosen array in our experiment, the MAD on the X chromosome prior to applying our algorithm is 0.087.

3.1. Intensity Effect Results. Figure 2 shows the loess fit as a function of the intensity (A). The intensity bias in this case is depicted by a convex curve (red line) that shows that probes with a small intensity are more likely to have a large log T/C value. Figure 3 shows the resulting M' vector as function of location on the array (Figure 3(a)) and genomic location (Figure 3(b)). For the genomic plot (Figure 3(b)) we averaged the two probes for each BAC clone in order to obtain a value for the loci. The MAD for the X chromosome after this step is 0.102, indicating a slight increase from the starting MAD on the X chromosome.

3.2. Spatial Effect Results. The results from studying the spatial effects for this data are provided in Figures 4 and 5. Figure 4(a) is the spatial effect image as estimated from a kernel density smoother when using L2 loss as opposed to L1 loss. Figure 4(b) shows the ranked cross-validation (CV) values as estimated using an absolute error or L1 loss ("1" line) and using a squared error or L2 loss ("2" line). Note that, the smoothing parameter (bandwidth) for our kernel density smoother changes slightly when using L2 loss versus L1 loss. Note that the L2 loss is the default loss function in our algorithm. The M'' data is shown via array location in Figure 5(a) and genomic profile in Figure 5(b). Contrasting Figure 5(a) with the data in Figure 1(a), it is clear that the technical spatial variation has been removed. In fact, the MAD for the X chromosome after removing the spatial effect is 0.075, indicating a reduction from the MAD prior to this step.

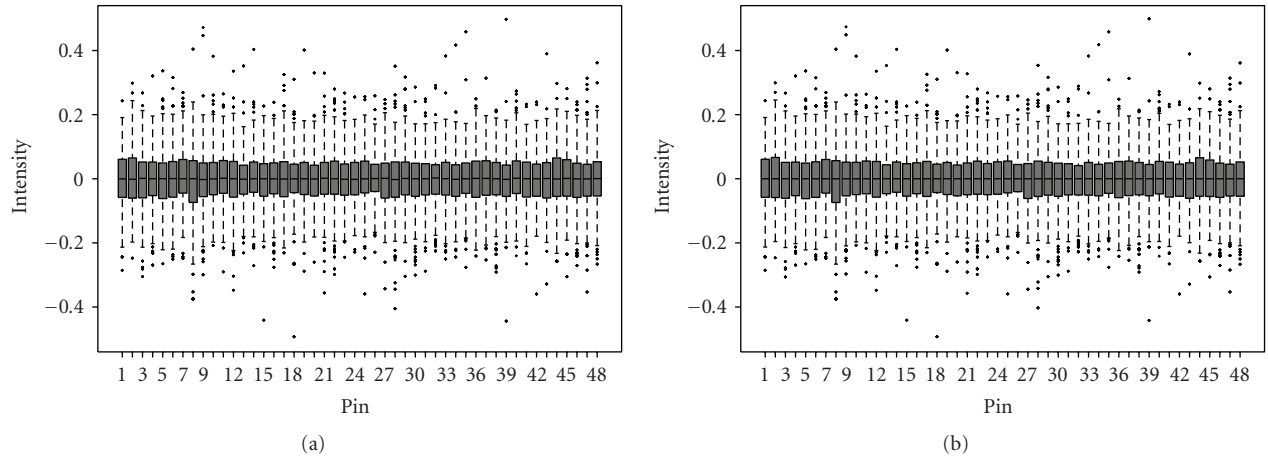

Figure 6: Spotting process: Pin: The box-plot showing the distribution of log T/C values for each *pin* (a) before correction and (b) after correction by median subtraction.

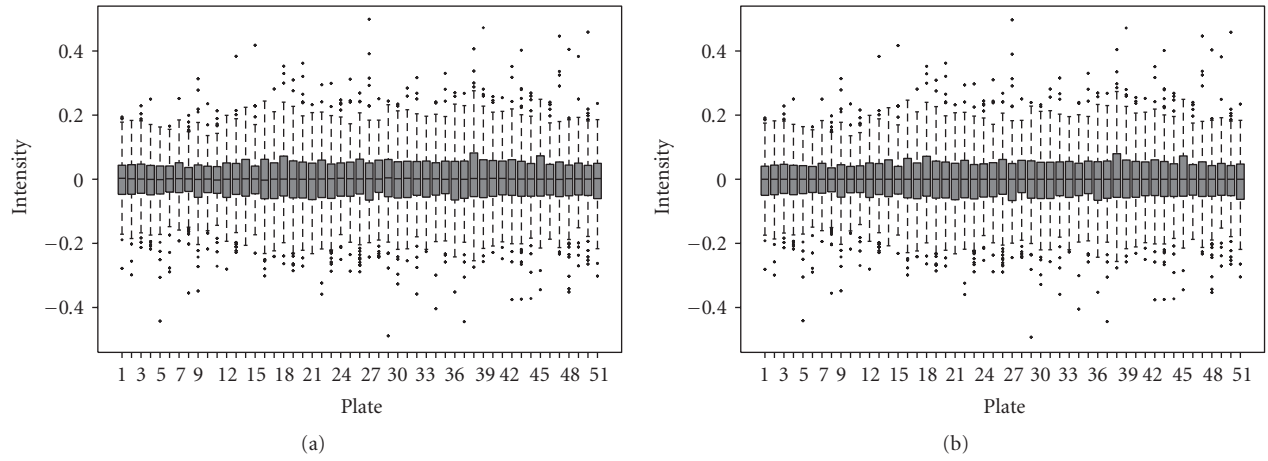

Figure 7: Spotting process: plate: the box plot showing the distribution of log T/C values for each *plate* (a) before correction and (b) after correction by median subtraction.

3.3. Spotting Process Effect Results. Figures 6, 7, 8 and 9 show the side by-side box plots of each of the spotting process effects before and after their removal. The final dataset after the "SmoothArray" process is displayed in an array location image (Figure 10(a)) and genomic profile (Figure 10(b)). From a closer analysis in Figures 6(a)–9(a), there is not an obvious outlier in terms of the spotting process effects after accounting for the intensity and spatial bias. The MAD for the X chromosome after removing the spotting process effect is 0.069, which represents a reduction when compared to the MAD score of 0.075 prior to accounting for this effect.

3.4. Overall Results. The initial MAD for our sample is 0.087. After applying the "SmoothArray" algorithm, our MAD is 0.069. This 20 percent reduction of noise indicates that subsequent aCGH calls of gains and losses should be improved due to the "SmoothArray" process. Figure 10(c) demonstrates this reduction in noise as a function of location on the chromosome. The raw genomic profile is represented in grey points, while the values after "SmoothArray" are represented in red.

As a global measure of the normalization methods, we also examined the MAD for the X chromosome for each of the 219 samples in the experiment designed to examine biomarkers in head and neck tumors. The median MAD prior to the "SmoothArray" process (raw normalization) is 0.1465, while the median MAD across the 219 samples after "SmoothArray" is 0.1139. Thus for this experiment, the "SmoothArray" process provides approximately a 23 percent reduction in noise. Table 2 displays the median MAD across the 219 samples for each of the competing normalization methods while Figure 11 shows the distribution of the MAD for each method. From Table 2 and Figure 11, we see that "SmoothArray" provides the optimal reduction in noise.

We also examined each of the normalization methods by using a nonparametric one-sample t test to examine the significance of the log ratios on the X chromosome. For each probe, we obtain a P value measuring the significance from

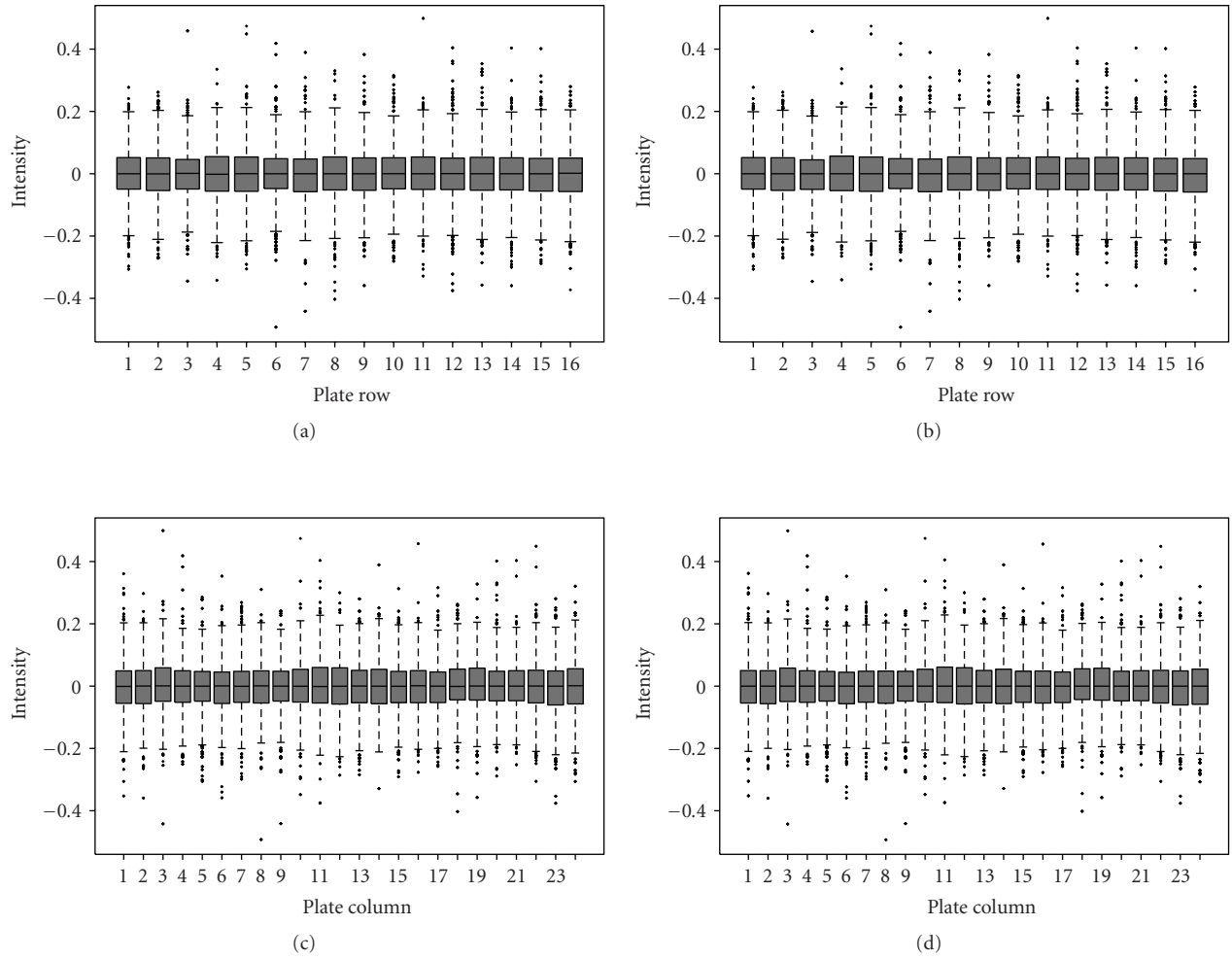

Figure 8: Spotting process: plate row/column: the box-plot showing the distribution of $\log T/C$ values for each *plate row* (a) before correction and (b) after correction by median subtraction and for each *plate column* (c) before correction and (d) after correction by median subtraction.

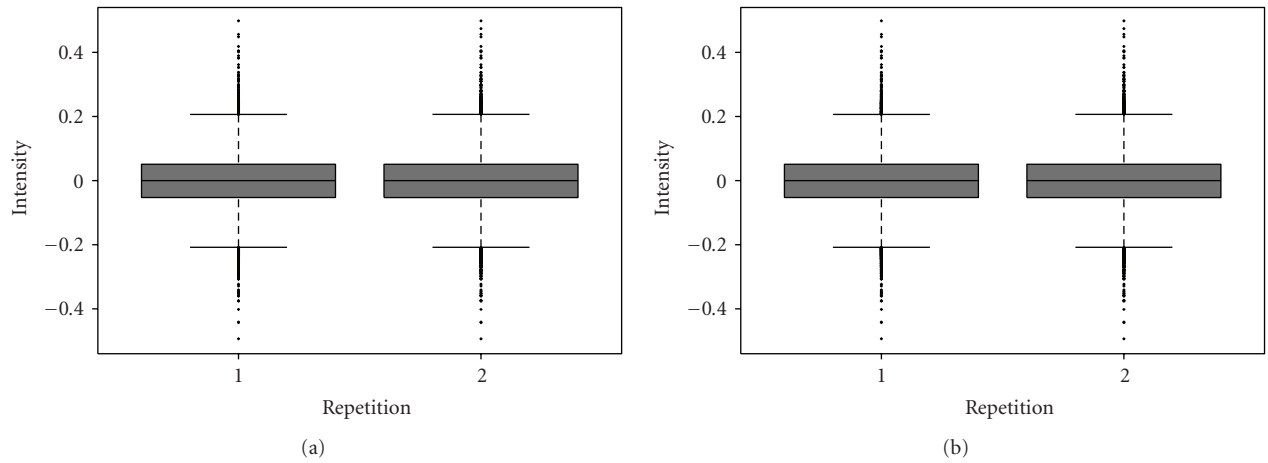

Figure 9: Repetition: the box plot showing the distribution of $\log T/C$ values for each repetition of the BAC clones (a) before correction and (b) after correction. In the RPCI microarray facility, each BAC clone is spotted twice on the array.

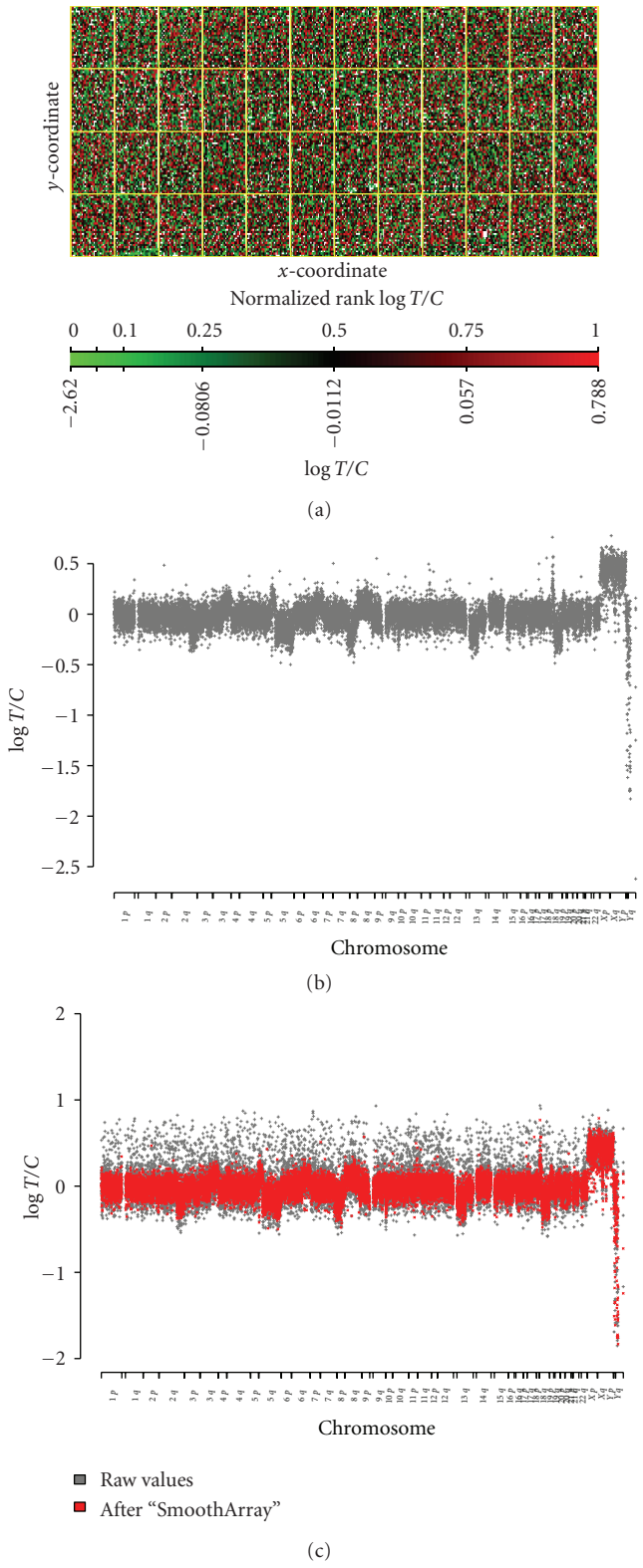

Figure 10: Final Results: The final log T/C values from "SmoothArray" as shown via (a) array image and (b) genomic plot. (c) A genomic plot showing the (median centered) raw genomic profile (grey points) with the same genomic profile after applying the "SmoothArray" algorithm (red points). Clearly there is a significant reduction in variation using the red profile compared to the grey profile.

a mean of 0 for each log ratio. Since each sample is hybridized against a sex-mismatched control, we expect each log ratio on the X chromosome to have a significantly small P value. For this analysis, we employed the Wilcoxon Rank Sum test (a nonparametric test) for each log ratio. The percentage of P values less than .05 is shown in Table 3. The cutoff of.05 was chosen since, on a univariate level, this would indicate a single copy gain or loss. Using this metric, we see that the "SmoothArray" normalization compares favorably to the grid loess, and global median normalization methods.

As a further comparison of each normalization scheme, the use of M-XY plots allows the users a visual metric to compare the different normalization methods. Figure 12 allows the user to determine if any spatial abnormalities are present on the array. From examining Figure 12, we see that the other normalization methods all appear to have residual spatial artifacts.

4. Discussion

Through a series of sequential steps we have developed an algorithm called "SmoothArray" which normalizes the logarithmic ratios from a CGH- based microarray platform. This normalization removes three major sources of technological signal. The technological signal due to the intensity effects is removed first. Secondly the signal due to the spatial location on the microarray is accounted for and removed. Lastly the signal due to the spotting process is removed. Each of these sources of signal is a well-documented problem in aCGH literature [23]. Throughout the algorithm our philosophy was to employ parsimonious and straightforward approaches to correct for the technical effects at each step.

We note the similarity of our "SmoothArray" process with the preprocessing defined in [23]. Specifically, both methods remove noise due to intensity effects, spatial effects, and plate effects. However, we have several novel additional features that distinguishour method. Namely, we employ the CBS algorithm throughout our "SmoothArray" algorithm to ensure that we preserve the biological signal, in other words, to ensure that we only remove the technical variation present in the dataset. Although minor, we use a loess rather than the lowess method as in [23] to remove the intensity bias. We also employ a kernel density smoother with bandwidth chosen via cross-validation to account for the spatial bias rather than a 11×11 window median smoother employed in [23]. Further, we remove the effects due to the pin arrayer procedure and additionally the plate row, plate column, and repetition effect. By comparing these results on the X chromosome we show the results of our "SmoothArray" algorithm.

Also, our "SmoothArray" algorithm has the flexibility to explore the cross-validation steps using an absolute (L1) loss function or squared error (L2) loss function. Note that, currently we have a background option in aCGHplus which allows the user to subtract a weighted version of the background. Future work will explore employing a step that takes into account the background in preprocessing CGH BAC arrays. The background image may come into consideration within "SmoothArray" via two ways (1) by using the blank spots at the end of each grid (see Section 2.1)

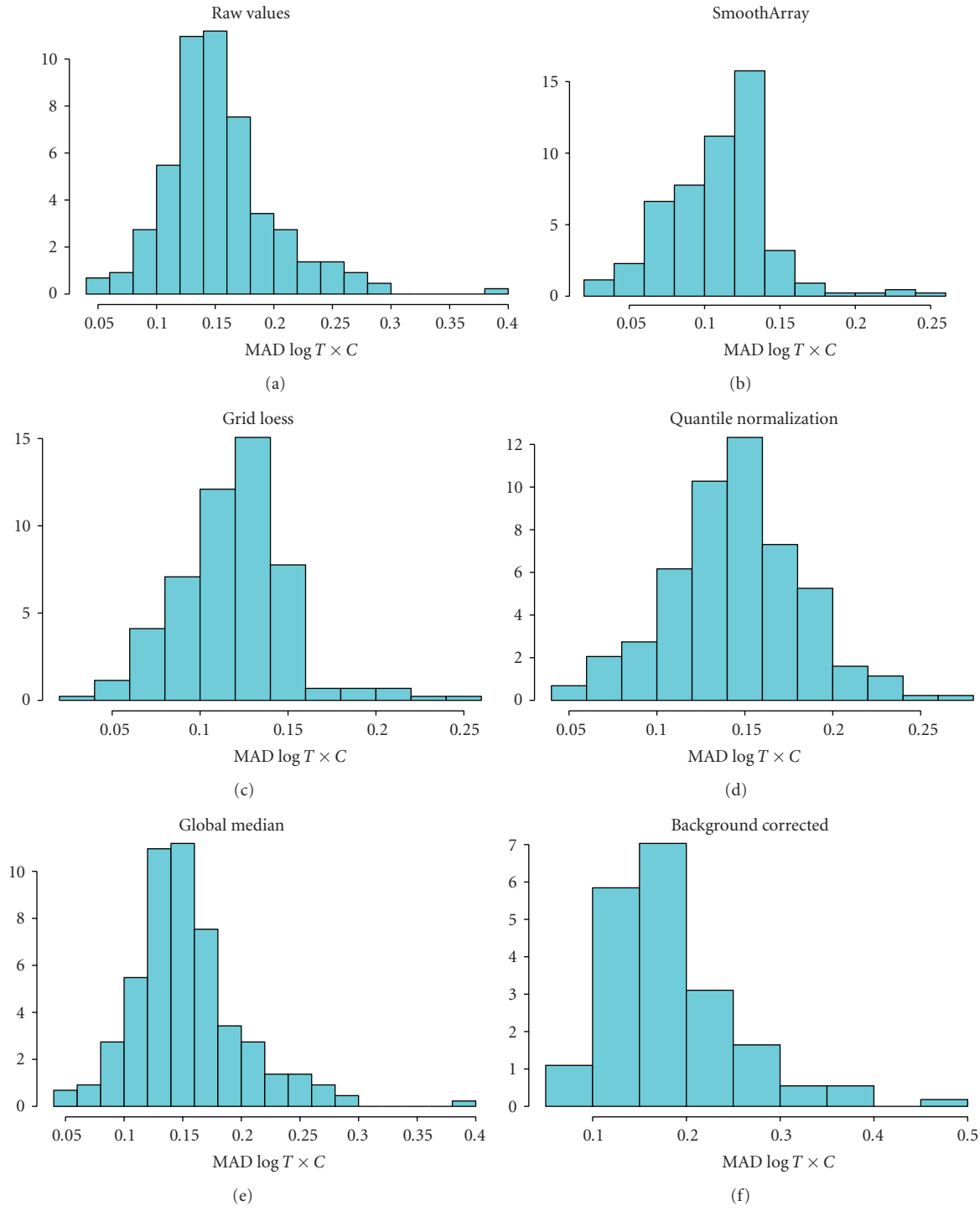

Figure 11: Distribution of the MAD: for each normalization method, the distribution of the MAD values for the log ratio of each probe on the X chromosome. The median for each of the distributions is summarized in Table 2. According to Table 2, the smallest median MAD is obtained using the "SmoothArray" algorithm.

Table 2: Table comparing the median MAD values on the X chromosome for each of the normalization methods. From this metric, we see that the "SmoothArray" normalization procedure provides the optimal noise reduction for probes on the X chromosome.

Raw	"SmoothArray"	Grid loess	Quantile normalization	Global median	Background subtraction
0.1465	0.1139	0.1215	0.1468	0.1465	0.1675

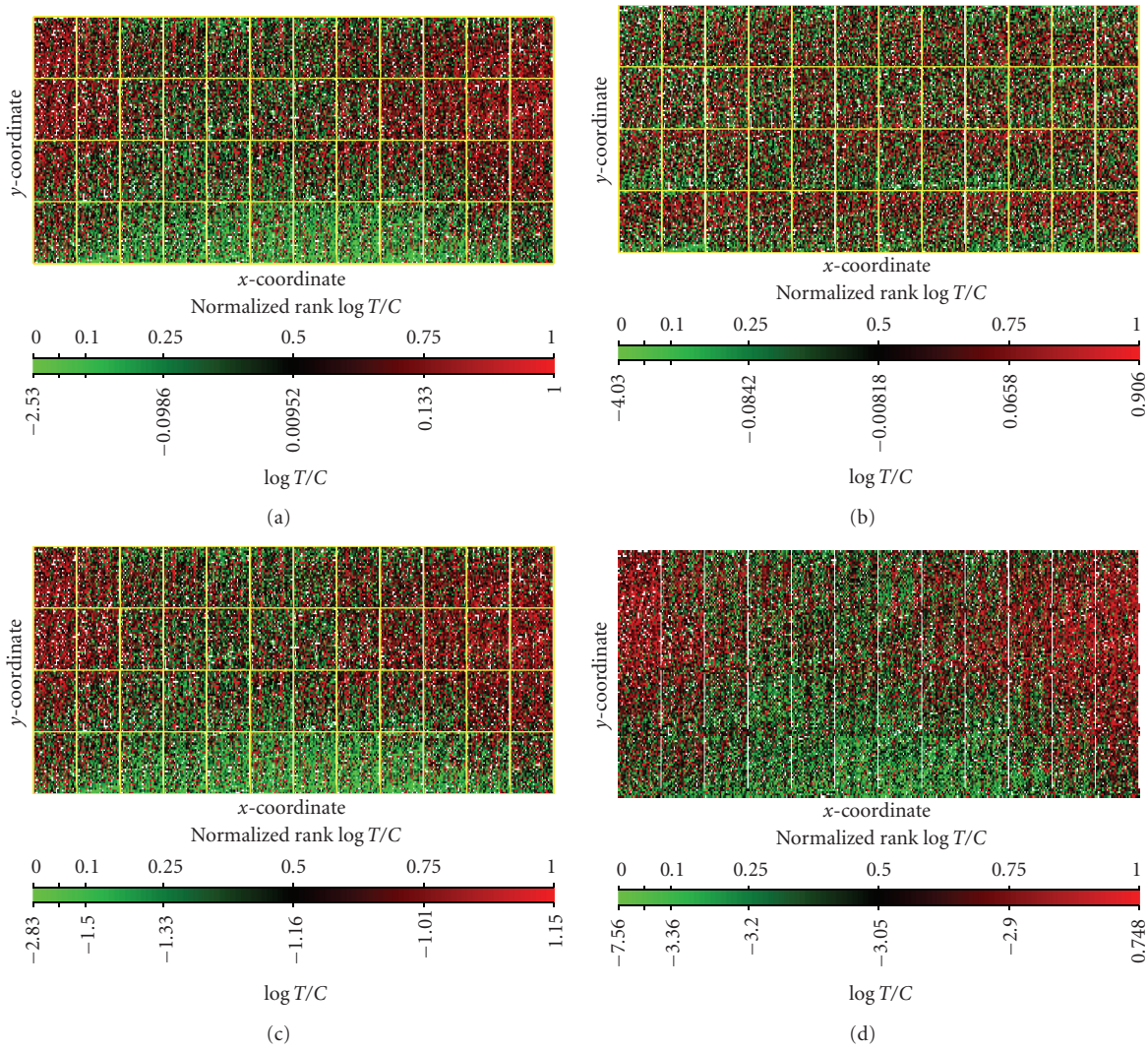

Figure 12: Other methods: the final log T/C values from other normalizing methods, namely, (a) global median subtraction, (b) grid loess normalization, (c) quantile normalization, and (d) background subtraction. From a visual inspection, there appear to be spatial artifacts that are not removed by these normalization methods.

Table 3: Table comparing the percentage of P values less than .05 on the X chromosome. From this metric, we see that the "SmoothArray" procedure performs favorably compared to the grid loess and global median normalization methods.

Raw	"SmoothArray"	Grid loess	Quantile normalization	Global median	Background subtraction
100%	93%	90%	100%	89%	100%

or (2) by using images obtained from the GenePix scanner where segmentation algorithms are applied to determine a background signal for each spot.

By examining the spot process effect we can employ violin plots to examine the quality control for each array-design variable pin, plate, plate row, plate column, repetition. Note that the repetition effect acts as a surrogate measure for a potential time effect. That is, the time elapsed in spotting the probes on the glass slide is represented by the "repetition" variable, since the second spot for each BAC clone is not spotted until all other BAC clones have been spotted once.

Violin plots have numerous references in current statistical literature as a way of combining the information available from local density estimates with the basic summary statistics inherent in standard box plots. Combining the box plot and the density trace on a single plot, comparing the distributions of several variables via violin plots, is a great tool for CGH microarrays [42].

For the RPCI aCGH lab, there are three flags used to determine the quality of the spotted probes. Firstly, the spot can be flagged for having a low signal-to-noise value. The signal-to-noise value is determined by taking the mean value of the pixels in the signal and dividing them by the standard

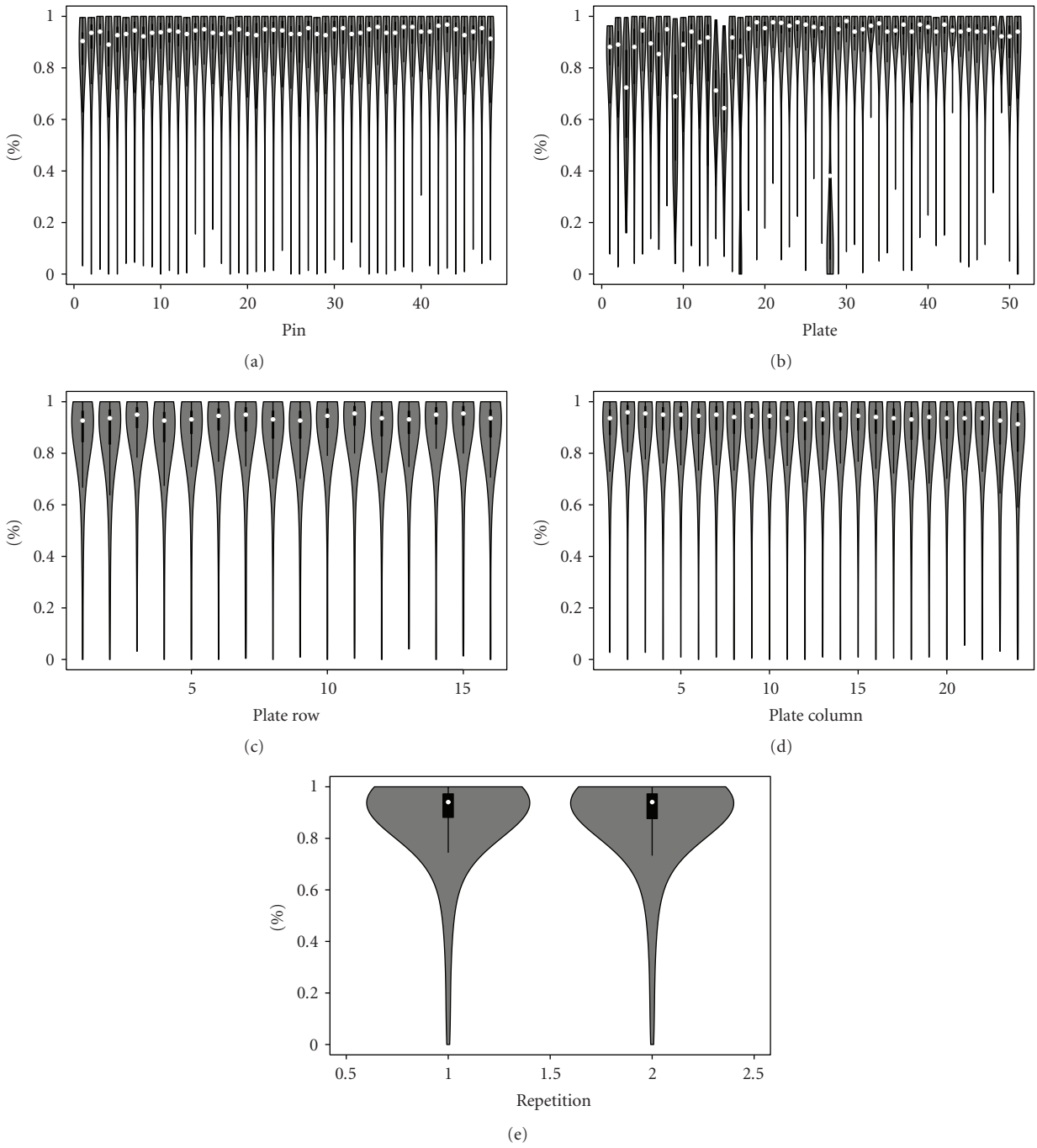

Figure 13: Violin plots for quality control: violin plots for quality control in the BAC aCGH spotting process. Each spot is evaluated according to the hybridization quality control. There were ~200 samples comprising the experiment, and the percentage (over the number of samples) of times that each spot passed quality control was recorded. The violin plot in (a) shows the distribution of this percentage for each pin. The median is shown in white. The quality control for probes is assessed according to the (b) 384-well plates, (c) plate row, (d) plate column, and (e) repetition number for the BAC clones.

deviation of the background. If this value is too low (<2.5) then the spot is of poor quality. Secondly, the spot can be manually flagged by the user to determine poor quality, or thirdly, the spot can be flagged of poor quality because of a dim signal in one of the channels. Dimness is determined by having a mean signal value under a prescribed cutoff in one of the channels. With this set of flags used for poor spot quality, for a given set of samples (experiment), we can compute the rate for each spot being flagged with a QC problem. Then we can examine the distribution of this percentage via violin plots across each spotting process variable. Figure 13(a) shows the distribution of the spot

percentages across each *pin*, where the spot percentage is the number of times the given spot passed quality control divided by the number of samples in the experiment (219 samples). Similarly, Figure 13(b) shows the distribution of the percentage of BAC clones from each *plate* that passed the quality control. By studying Figure 13, it is shown that Pins 1 and 3 may be suspect in terms of quality control, while several plates have a larger frequency of quality control problems. Specifically, Plate 28 consists of BAC clones that are consistently flagged for quality control problems. By examining Figures 13(c), 13(d), and 13(e) for this experiment, there does not appear to be any obvious problems affecting the plate row, plate column, or repetition number for the BAC clones. The concept of violin plots for quality control can further be extended to other commonly reported spot variables such as background mean, background standard deviation, and other potential outlier flags.

For future work, we plan to examine another quality control measure relevant to the BAC clones. Due to the nature of the BAC clones and the updates to the human genome, it is possible that the BAC clones could be mismapped from their position on the genome. Mismapped BAC clones can manifest themselves as appearing as an outlier when viewed via their genomic profile. Using a mixture model approach, we plan to subset the number of BAC clones under consideration based on estimating the probability for a given BAC clone to be mismapped. This approach shows great promise based on our early attempts at modeling mismapped BAC clones.

A key component in preprocessing aCGH data lies in understanding the subsequent analysis steps. With our preprocessed data, the next step in the aCGH BAC analysis pipeline involves characterizing the genome in terms of detecting regions of chromosomal copy number variations (gains and losses). The softwares and algorithms designed for this analysis include CGHcall [43] and other breakpoint detection methods, for example, [44–46]. A slightly different approach allows the researcher to analyze each chromosomal arm rather than examining within each arm for chromosomal breakpoints [47]. This approach allows the researcher to characterize a chromosome in terms of overall imbalance (with confidence) rather than focus on specific regions of gains and losses.

Our "SmoothArray" method clearly shows improvement in reducing the noise for a dataset of 219 samples designed to study head and neck tumors. Further, when using several quality control metrics, our method performs favorably to five other competing normalization methods described in Materials and Methods. For future work, we plan to extend our comparisons to quantify the amount of improvement over other competing pre-processing methods such as those in [23–27]. Experiments and comparisons such as those employed in [32, 48, 49] can be used to assess the performance and determine the best analysis routes for identifying genomic imbalances in BAC aCGH datasets. This future study would also compare the subsequent algorithms that assess gains and losses across the genome.

5. Conclusion

This paper proposes a novel algorithm to preprocess BAC CGH arrays. This novel method compares favorably against several other normalization measures when evaluated using several quality control metrics. For this study, we focused on data obtained from the RPCI microarray facility on a study of ~200 head and neck tumor samples. For this experiment, our algorithm reduced the noise by approximately 23 percent. By removing the technological noise due to the intensity effect, spatial effect, and spotting process, the resulting data has reduced noise and is suitable for subsequent analyses to determine chromosomal regions of gains and losses. The "SmoothArray" method also offers the user the option to examine several quality control figures which allows the researcher to pinpoint problems that may arise in the spotting process. This software is freely available at [31].

Acknowledgments

The data used for this paper was obtained under Grant nos. 5R01CA113882-04NIH (Nowak PI) and P30CA016056 (Trump PI) while the research for J. C. Miecznikowski was partially funded under both grants.

References

[1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[2] A. M. Snijders, N. Nowak, R. Segraves et al., "Assembly of microarrays for genome-wide measurement of DNA copy number," *Nature Genetics*, vol. 29, no. 3, pp. 263–264, 2001.

[3] F. Falciani, *Microarray Technology Through Applications*, Routledge, Routledge, UK, 2007.

[4] R. Gentleman, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, New York, NY, USA, 2005.

[5] D. Lugtenberg, A. P. M. De Brouwer, T. Kleefstra et al., "Chromosomal copy number changes in patients with non-syndromic X linked mental retardation detected by array CGH," *Journal of Medical Genetics*, vol. 43, no. 4, pp. 362–370, 2006.

[6] N. Miyake, O. Shimokawa, N. Harada et al., "BAC array CGH reveals genomic aberrations in idiopathic mental retardation," *American Journal of Medical Genetics*, vol. 140, no. 3, pp. 205–211, 2006.

[7] P. Stankiewicz and A. L. Beaudet, "Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation," *Current Opinion in Genetics and Development*, vol. 17, no. 3, pp. 182–192, 2007.

[8] R. Ullmann, G. Turner, M. Kirchhoff et al., "Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation," *Human Mutation*, vol. 28, no. 7, pp. 674–682, 2007.

[9] D. G. Albertson, B. Ylstra, R. Segraves et al., "Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene," *Nature Genetics*, vol. 25, no. 2, pp. 144–146, 2000.

[10] G. Hodgson, J. H. Hager, S. Volik et al., "Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas," *Nature Genetics*, vol. 29, no. 4, pp. 459–464, 2001.

[11] J. R. Pollack, T. Sørlie, C. M. Perou et al., "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 of the United States of America*, vol. 99, no. 20, pp. 12963–12968, 2002.

[12] D. G. Albertson, "Profiling breast cancer by array CGH," *Breast Cancer Research and Treatment*, vol. 78, no. 3, pp. 289–298, 2003.

[13] D. G. Albertson, C. Collins, F. McCormick, and J. W. Gray, "Chromosome aberrations in solid tumors," *Nature Genetics*, vol. 34, no. 4, pp. 369–376, 2003.

[14] C. S. Hackett, J. G. Hodgson, M. E. Law et al., "Genome-wide array CGH analysis of murine neuroblastoma reveals distinct genomic aberrations which parallel those in human tumors," *Cancer Research*, vol. 63, no. 17, pp. 5266–5273, 2003.

[15] C. Garnis, B. P. Coe, L. Zhang, M. P. Rosin, and W. L. Lam, "Overexpression of LRP12, a gene contained within an 8q22 amplicon identified by high-resolution array CGH analysis of oral squamous cell carcinomas," *Oncogene*, vol. 23, no. 14, pp. 2582–2586, 2004.

[16] J. A. Veltman, J. Fridlyand, S. Pejavar et al., "Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors," *Cancer Research*, vol. 63, no. 11, pp. 2872–2880, 2003.

[17] L. W. M. Loo, D. I. Grove, E. M. Williams et al., "Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes," *Cancer Research*, vol. 64, no. 23, pp. 8541–8549, 2004.

[18] D. Pinkel and D. G. Albertson, "Array comparative genomic hybridization and its applications in cancer," *Nature Genetics*, vol. 37, no. 6, pp. S11–S17, 2005.

[19] M. R. Rossi, J. Conroy, D. McQuaid, N. J. Nowak, J. T. Rutka, and J. K. Cowell, "Array CGH analysis of pediatric medulloblastomas," *Genes Chromosomes and Cancer*, vol. 45, no. 3, pp. 290–303, 2006.

[20] A. Idbaih, Y. Marie, C. Lucchesi et al., "BAC array CGH distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas," *International Journal of Cancer*, vol. 122, no. 8, pp. 1778–1786, 2008.

[21] M. E. Futschik and T. Crompton, "OLIN: optimized normalization, visualization and quality testing of two-channel microarray data," *Bioinformatics*, vol. 21, no. 8, pp. 1724–1726, 2005.

[22] Y. Xiao, M. R. Segal, and H. Y. Yee, "Stepwise normalization of two-channel spotted microarrays," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, p. 1117, 2005.

[23] M. Khojasteh, W. L. Lam, R. K. Ward, and C. MacAulay, "A stepwise framework for the normalization of array CGH data," *BMC Bioinformatics*, vol. 6, no. 1, article 274, 2005.

[24] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin, "A statistical approach for array CGH data analysis," *BMC Bioinformatics*, vol. 6, no. 1, article 27, 2005.

[25] P. Neuvial, P. Hupé, I. Brito et al., "Spatial normalization of array-CGH data," *BMC Bioinformatics*, vol. 7, no. 1, article 264, 2006.

[26] J. Staaf, G. Jönsson, M. Ringnér, and J. Vallon-Christersson, "Normalization of array-CGH data: influence of copy number imbalances," *BMC Genomics*, vol. 8, no. 1, article 382, 2007.

[27] H. Huang, N. Nguyen, S. Oraintara, and AN. Vo, "Array CGH data modeling and smoothing in Stationary Wavelet Packet Transform domain," *BMC Genomics*, vol. 9, supplement 2, p. S17, 2008.

[28] J. Fridlyand and P. Dimitrov, *aCGH: Classes and functions for Array Comparative Genomic Hybridization Data*, 2009, R package version 1.22.0.

[29] P. Neuvial and P. Hupe, "MANOR: CGH Micro-Array NORmalization," r package version 1.18.0., 2009, http://bioinfo.curie.fr/projects/manor/index.html.

[30] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[31] D. Gaile, L. Shepherd, and J. Miecznikowski, "aCGHplus," r package version 2.4.1, 2009, http://sphhp.buffalo.edu/biostat/research/software/acghplus/index.php.

[32] N. J. Nowak, J. Miecznikowski, S. R. Moore et al., "Challenges in array comparative genomic hybridization for the analysis of cancer samples," *Genetics in Medicine*, vol. 9, no. 9, pp. 585–595, 2007.

[33] D. Miliaras, J. Conroy, S. Pervana, S. Meditskou, D. McQuaid, and N. Nowak, "Karyotypic changes detected by comparative genomic hybridization in a stillborn infant with chorioangioma and liver hemangioma," *Birth Defects Research Part A—Clinical and Molecular Teratology*, vol. 79, no. 3, pp. 236–241, 2007.

[34] K. Sellers, J. Miecznikowski, and W. Eddy, "Removal of systematic variation in genetic microarray data," Tech. Rep. 779, Carnegie Mellon University, 2004.

[35] P. Stafford, *Methods in Microarray Normalization*, CRC, London, UK, 2008.

[36] S. K. Watson, R. J. deLeeuw, A. S. Ishkanian, C. A. Malloff, and W. L. Lam, "Methods for high throughput validation of amplified fragment pools of BAC DNA for constructing high resolution CGH arrays," *BMC Genomics*, vol. 5, no. 1, article 6, 2004.

[37] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, 2004.

[38] D. Nychka, "Fields: tools for spatial data," r package version 2.0., 2009, http://www.cgd.ucar.edu/stats/Software/Fields.

[39] J. Miecznikowski, D. Gaile, J. Conroy, and N. Nowak, "Quality control metrics for array comparative genomic hybridization data," Tech. Rep. 0606, Department of Biostatistics, University at Buffalo, Buffalo, NY, USA, 2006.

[40] W. Cleveland, E. Grosse, and W. Shyu, "Local regression models," in *Statistical Models in S*, pp. 309–376, Chapman & Hall, Boca Raton, Fla, USA, 1992.

[41] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[42] J. L. Hintze and R. D. Nelson, "Violin plots: a box plot-density trace synergism," *American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.

[43] M. A. van de Wie, K. I. Kim, S. J. Vosse, W. N. van Wieringen, S. M. Wilting, and B. Ylstra, "CGHcall: calling aberrations for array CGH tumor profiles," *Bioinformatics*, vol. 23, no. 7, pp. 892–894, 2007.

[44] K. Jong, E. Marchiori, G. Meijer, A. V. D. Vaart, and B. Ylstra, "Breakpoint identification and smoothing of array comparative genomic hybridization data," *Bioinformatics*, vol. 20, no. 18, pp. 3636–3637, 2004.

[45] P. Wang, Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani, "A method for calling gains and losses in array CGH data," *Biostatistics*, vol. 6, no. 1, pp. 45–58, 2005.

[46] Y. Nannya, M. Sanada, K. Nakazaki et al., "A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays," *Cancer Research*, vol. 65, no. 14, pp. 6071–6079, 2005.

[47] D. P. Gaile, E. D. Schifano, J. C. Miecznikowski, J. J. Java, J. M. Conroy, and N. J. Nowak, "Estimating the arm-wise false discovery rate in array comparative genomic hybridization experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, article 32, 2007.

[48] W. R. Lai, M. D. Johnson, R. Kucherlapati, and P. J. Park, "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data," *Bioinformatics*, vol. 21, no. 19, pp. 3763–3770, 2005.

[49] H. Willenbrock and J. Fridlyand, "A comparison study: applying segmentation to array CGH data for downstream analyses," *Bioinformatics*, vol. 21, no. 22, pp. 4084–4091, 2005.