[Frontiers in Bioscience 9, 540-549, January 1, 2004]

STATISTICAL METHODS FOR DETECTING GENOMIC ALTERATIONS THROUGH ARRAY-BASED COMPARATIVE GENOMIC HYBRIDIZATION (CGH)

Yuedong Wang [1] and Sun-Wei Guo [2]

[1] Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106, and [2]Department of Pediatrics and Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Materials and methods
3.1. Sample preparations
3.2. Normalization
3.3. Standard t-test
3.4. Smoothed t-test
3.5. HAS and SHAS methods
3.6. False discovery rate
4. Results
4.1. Genomic alterations associated with endometriosis
4.2. Simulation studies
5. Discussion
6. Acknowledgments
7. References

1. ABSTRACT

Array-based comparative genomic hybridization (ABCGH) is an emerging high-resolution and high-throughput molecular genetic technique that allows genome-wide screening for chromosome alterations associated with tumorigenesis. Like the cDNA microarrays, ABCGH uses two differentially labeled test and reference DNAs which are cohybridized to cloned genomic fragments immobilized on glass slides. The hybridized DNAs are then detected in two different fluorochromes, and the significant deviation from unity in the ratios of the digitized intensity values is indicative of copy-number differences between the test and reference genomes. Proper statistical analyses need to account for many sources of variation besides genuine differences between the two genomes. In particular, spatial correlations, the variable nature of the ratio variance and non-Normal distribution call for careful statistical modeling. We propose two new statistics, the standard t-statistic and its modification with variances smoothed along the genome, and two tests for each statistic, the standard t-test and a test based on the hybrid adaptive spline (HAS). Simulations indicate that the smoothed t-statistic always improves the performance over the standard t-statistic. The t-tests are more powerful in detecting isolated alterations while those based on HAS are more powerful in detecting a cluster of alterations. We apply the proposed methods to the identification of genomic alterations in endometrium in women with endometriosis.

2. INTRODUCTION

CGH is a powerful molecular genetic technique that allows genome-wide screening for chromosome alterations associated with tumorigenesis (1). Characterization of these alterations is important in diagnosis and prognosis in cancer, and in delineation of molecular genetic mechanisms underlying tumorigenesis. With CGH, cryptic gains and/or losses of chromosomal regions in the tissue samples of interest can be detected, without prior knowledge of specific regions of interest. However, the wider application of CGH has been hampered by its two limitations due to the use of high-quality metaphase chromosome preparations as hybridization targets: the limited resolution, typically 3-10 Mb (2, 3) and the low throughput (4, 5). The recently emerged ABCGH, or matrix-based CGH as it is sometimes called, overcomes these limitations (6-9), and is increasingly becoming the method of choice for high-throughput, high-resolution screening of genomic alterations (9).

Similar to the spotted cDNA microarrays (10), ABCGH uses two differentially labeled test (unknown sample to be analyzed) and reference (known to be genomically normal) DNAs which are cohybridized, under in situ suppression hybridization conditions, to cloned genomic fragments with known physical locations, spotted and immobilized on glass slides. The hybridized DNAs are then detected in two different fluorochromes, and the ratios of the digitized intensity values in the hybridized patterns of the DNAs onto the cloned fragments are indicative of copy-number differences between the test and the reference genomes.

Besides genuine differences between the two genomes, however, stochastic fluctuations, measurement errors or other errors of unknown origins, and consistent,

region-specific variations caused by differences in hybridization characteristics of the incorporated fluorochromes and by local variation in chromosomal structures, can all cause the ratio to deviate from unity (11). Therefore, the detection of genomic alterations using ABCGH requires proper statistical analysis of the intensity values from the two fluorochromes.

For conventional CGH, a calibration process is usually invoked, in which reference versus reference hybridizations are performed to gauge the normal range of ratio variations (12). The ratios of the test-reference hybridizations, at each chromosomal segment where the ratio is calculated, are then compared with, say, the two standard deviations (SD) outside the mean, obtained from the calibration, and a gain or loss is declared if the ratio is above or under the 2 SDs (presumably the nominal 95% confidence bounds without multiple comparison adjustment) (13). Sometimes a pair of fixed, global thresholds, say, 1.15 and 0.85 (14, 15), are used in lieu of 2 SDs.

Recognizing the variable nature of the variance of the mean ratio within and between reference:reference hybridizations and possible inequality of variances of mean ratio between the test:reference and reference:reference experiments, a t-like statistic incorporating reference:reference and test:reference variations to detect genomic alterations segment by segment were proposed (16). This method, however, assumes that the ratio of the variances of test:reference ratio means and of reference:reference ratio means is constant across the whole genome, which may not be true. In addition, correlation in the estimated variances and the spatial correlation of ratios in the neighboring segments are completely ignored.

In contrast to the enormous efforts devoted to the development of the ABCGH and to the analysis of cDNA microarray data, the statistical analysis of ABCGH data has received scant attention. This is somewhat surprising, considering the great potential that ABCGH espouses (4, 7). Most applications of ABCGH use either fixed, global thresholds, e.g. 0.85 for loss and 1.15 for gain in relative copy number (4, 5, 17), or 2 SDs (18). The t-statistic proposed in (16) for conventional CGH could also be used (14). However, the use of fixed thresholds has little, if any, statistical validity since hybridization efficiency, and thus variation, can vary with chromosomal structures (16). The 2-SD method and the t-statistics method completely ignore spatial correlations between neighboring clones, which can be prominent in ABCGH data, especially with high density ABCGH, since, once a clone exhibits alteration, its neighboring clones also tend to have alterations (20). As these methods virtually ignore information embedded in the neighboring clones, they are less efficient. In addition, they depend on strong assumptions about the variance of the intensity ratio, which may not be entirely valid or justifiable. The method proposed in (20) takes account the spatial correlations by assuming a fixed-width window with the same distance-dependent correlation, but this may be too rigid since, first, correlations may vary with chromosomal structures and thus locations, and, second, the size of genomic alterations may vary. The specification of the width of correlation in (20) is somewhat arbitrary.

From a statistical standpoint, the identification of genomic alterations with ABCGH appears to be more challenging than the identification of differentially expressed genes in cDNA microarrays, since, besides all the data normalization issues and the account for various sources of variations, the former may also need to take account of spatial correlations, which could be safely ignored in the latter. This spatial information also is important in mapping where the genomic alterations are, in addition to the characterization of the alterations (e.g. loss or gain).

In analysis of ABCGH data, several considerations are in order. First, less restrictive assumptions on variance are preferable, since the variance may vary according to the chromosomal structures and thus locations of the clones. Second, spatial correlations should be properly accounted for, which also should increase statistical efficiency and improve precision in estimation. This would also translate into requirement for less calibration samples because of increased efficiency. Third, since the nature of variation in variance may vary from laboratory to laboratory due to considerable variations in the execution of ABCGH experiments, less distributional assumption on the ratio would be preferable. Lastly, robustness to outliers and the minimization of the dominating effect of clones with very small variance would be desirable.

In this article we propose two methods for detecting genomic alterations. We consider two statistics: the standard t-statistic and its modification with variances smoothed along the genome.

For each of these two statistics, we propose two methods to detect clones with significant alterations: the standard t-test and a test based on the hybrid adaptive spline (HAS). As a statistical method for function estimation, HAS has the ability to handle a wide variety of shapes and spatial inhomogeneities (21). It is an objective approach that allows data to dictate the shape of a function. Since alterations typically occur in local regions, the expectation of a ratio profile along the genome equals zero except in some regions where alterations occur. Spatially adaptive, HAS was created to handle spatial inhomogeneity as in ABCGH data. We also propose a bootstrap method within the HAS framework to assess statistical significance and false discovery rate. For the ease of exposition, we refer these four methods as STNT, SMOT, HAS and SHAS methods in the remaining of this paper.

3. MATERIALS AND METHODS

3.1. Sample preparations

We conducted an experiment to identify genomic alterations, if any, in the endometrial tissues in women with endometriosis. Nine endometrial tissue samples were taken from five patients with endometriosis (cases) and four

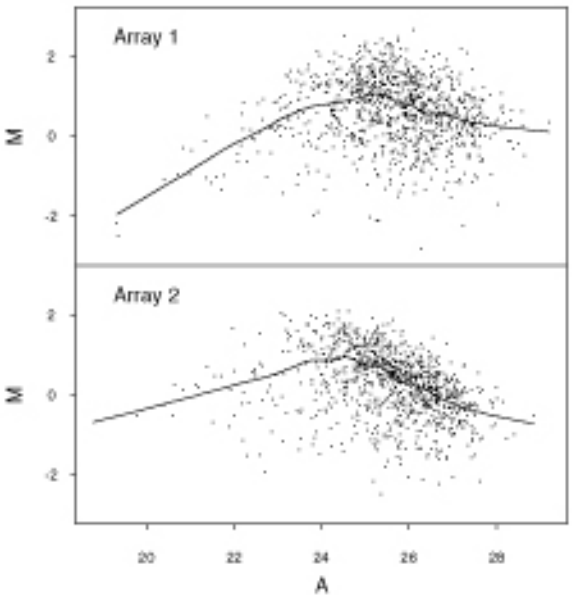

Figure 1. M vs. A plot for two arrays in a case sample. The solid lines are the *lowess* fits with $f = 0.2$.

apparently normal women (controls) who underwent elective abortions at area clinics in Milwaukee. One additional sample from a normal placenta was also used to serve as a control. The research protocol has been approved by the Froedtert Hospital and the Medical College of Wisconsin IRB.

Five :m histological sections of formalin-fixed, paraffin-embedded tissue samples were mounted on uncharged, non-coated glass slides and stained with H&E reagents (Fisher Healthcare, Houston, TX). Epithelial cells were harvested using laser capture microdissection (Arcturus, Mountainview, CA), and DNA was isolated using Phenol/Chroloform extraction methods. All DNA samples and reference DNA (Promega) were then amplified using DOP-PCR (22). The PCR products were purified by Zymo purification columns (Zymo Research, Orange, CA). Before hybridization, DNA was quantified by the PicoGreen assay (Molecular Probes, Eugene, OR) according to the manufacturer's instructions.

Human BAC arrays from the Spectral Genomics (Houston, TX) were used for this study. The arrays contain 995 BAC *clones*, spanning the whole human genome. Each chip has two identical *panels*, left and right, each containing the 995 clones. The hybridization was done according to the manufacturer's instructions. Each DNA sample was co-hybridized with the reference DNA in two arrays, with one array switching the dyes. All arrays were scanned using a ScanArray 5000 (GSI Lumonics, Billerica, MA). Images were processed using Matarray image-quality assurance software (23).

3.2. Normalization

We define the *group* variable *g*, with *g*=1 or 2 corresponding to the case and control sample, respectively. To ease exposition, we denote n_1 (n_2) as the number of case (control) samples, and *C* as the number of clones. In our experiment, $n_1 = n_2 = 5$, *C*=995. For all clones, their physical locations, in terms of chromosome number and the *distance* (denote as x_c), in cM, from the p-end of the chromosome, are completely known.

For a particular combination of *group g*, *sample s* in *group g*, *array a*, *panel p* and *clone c*, let T_{gsapc} and R_{gsapc} denote the measured fluorescence intensities for the test and reference samples respectively. For *g*=1,2, $s = 1,...,n_g$, $a = 1,2$, $p = 1,2$, and $c = 1,...,C$, we define $M_{gsapc} = \log_2(T_{gsapc} / R_{gsapc})$ and $A_{gsapc} = \log_2(T_{gsapc} \cdot R_{gsapc})/2$.

All M-A plots showed a strong pattern commonly seen in cDNA microarray data (24). All samples have similar patterns as those in Figure 1. Two arrays in the same sample usually have the same pattern. Thus the dye-swap design is not self-normalizing as defined in (24) (note that we use ratios of test and reference rather than red and green). We applied the same normalization procedure based on *lowess* as in (24) to each array. Specifically, let l_{gsac} be the *lowess* fit to the M-A plot for *array* a of *sample* s in *group g*. Then the normalized *M* is defined as $N_{gsapc} = M_{gsapc} - l_{gsac}$. The normalization was done at the *array* level since hybridization was performed for each array and the patterns of two panels within an array are very similar.

3.3. Standard t-test (STNT)

After normalization, we define, for $1,...,n_1$ and $c = 1,...,C$,

$$z_{sc} = \frac{1}{4}\sum_{a=1}^{2}\sum_{p=1}^{2} N_{1sapc} - \frac{1}{4n_2}\sum_{s=1}^{2}\sum_{a=1}^{2}\sum_{p=1}^{2} N_{2sapc}$$
$$= \bar{N}_{1s\cdot\cdot c} - \bar{N}_{2s\cdot\cdot c}$$

the differences of mean ratios between a single case sample and n_2 control samples.

Our goal is to detect clones with the expectation of z_{sc} differ from zero. Since clones with alterations may vary from patient to patient, combining all case samples may obscure altered clones shared only by one or two patients. Consequently, we attempted to detect clones with alterations for each case sample separately.

For fixed *g*, *s* and *c*, we have a 2×2 factorial design for *array* and *panel*. For a given case sample *s*, we assume the model

$$N_{1sapc} = \mu_{1sc} + \alpha_{1sac} + \varepsilon_{1sapc}, \quad (3.1)$$
$$a = 1,2; p = 1,2,$$

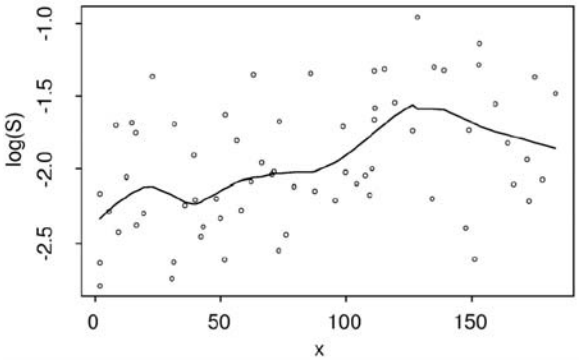

Figure 2. $\log(S)$ vs x plot for chromosome 6 in a case sample. The circles are estimates based on ANOVA models. The solid line is the *lowess* estimate of h in model (3.6).

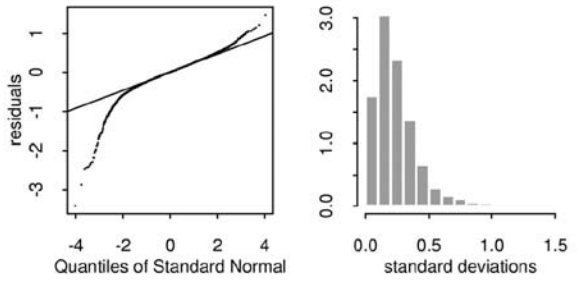

Figure 3. Left: normal probability plot of residuals. Right: histogram of estimated residual standard deviations.

where μ_{1sc} represents the common mean for all *arrays* and *panels* in *sample s*, α_{1sac} represents *array* (dye) effect, and ε_{1sapc} are random errors with mean 0 and variance σ^2_{1sc}. Since the dye effect is canceled out due to the dye-swap design, we have $\bar{N}_{1s\cdot\cdot c} = \mu_{1sc} + \bar{\varepsilon}_{1s\cdot\cdot c}$. For n_2 control samples, we assume that $\bar{N}_{2s\cdot\cdot c}(s = 1,...,n_2)$ are independent and identically distributed with mean μ_{2c} and variance σ^2_{2c}. It is easy to see that $E(z_{sc}) = \mu_{1sc} - \mu_{2sc}$ and $Var(z_{sc}) = \sigma^2_{1sc}/4 + \sigma^2_{2c}/n_2$. $E(z_{sc}) \neq 0$ if and only if there is an alteration at *clone c*. We estimated σ^2_{1sc} and σ^2_{2c} by

$$s^2_{1sc} = \sum_{a=1}^{2}(N_{1sa1c} - N_{1sa2c})^2/4 \quad (3.2)$$

and

$$s^2_{2c} = \sum_{s=1}^{n_2}(\bar{N}_{2s\cdot\cdot c} - \bar{N}_{2\cdot\cdot c})^2/(n_2 - 1) \quad (3.3)$$

respectively. Denoting $S_{sc} = \sqrt{s^2_{1sc}/4 + s^2_{2c}/n_2}$, the standard t-statistic

$$t_{sc} = z_{sc}/S_{sc} \quad (3.4)$$

Using Satterthwaite method, we then approximated the null distribution of the t-statistic t_{sc} by a t-distribution with degree of freedom

$$r_{sc} = \frac{(s^2_{1sc}/n_1 + s^2_{1c}/n_2)^2}{s^4_{1sc}/n_1^2(n_1 - 2) + s^4_{2c}/n_2^2(n_2 - 1)} \quad (3.5)$$

Note that this t-test is different from the one in (16) where certain relationships between the variances of the test and the control samples had been assumed.

3.4. Smoothed t-test (SMOT)

Estimates of variances (3.2) and (3.3) can be unreliable when sample size is small as in our experiment. These estimates are highly variable (Figure 2).

From a modeling perspective, chromosomes can be practically viewed as continuous, and the clones, with known physical locations, are observation points interspersed along the chromosomes. Within each chromosome, it is reasonable to assume that the variance is a smooth function of clone locations. Specifically, for a fixed case sample s and chromosome, we assume that

$$\log(S_{sc}) = h(x_c) + e_c. \quad (3.6)$$

where h is a smooth function of the *distance* of *clone c*, x_c. We fit model (3.6) using the robust *lowess* method with 30% of the data used for smoothing at each position. A typical fit is shown in Figure 2. We then defined a modified t-like statistic

$$u_{sc} = z_{sc}/\exp(\hat{h}(x_c)). \quad (3.7)$$

Replacing standard errors by their smoothed estimates also reduces the effect of outliers and prevents clones with very small variances from dominating the result. We applied the same *lowess* smoothing to the approximated degrees of freedom in (3.5). We then approximated the null distribution of the t-like statistic u_{sc} by a t-distribution with degree of freedom equals to the smoothed degree of freedom.

3.5. HAS and SHAS methods

The standard t-test ignores possible spatial correlations which may make them less efficient. In addition, the random errors may not follow a normal distribution. For each gene c, we fit the following model to the control samples

$$N_{2sapc} = \mu_{2c} + \alpha_{2sac} + \beta_{sac} + (\alpha\beta)_{sac} + \varepsilon_{2sapc}, \quad (3.8)$$
$$s = 1,...,n_2; a = 1,2; p = 1,2,$$

where β_{sc} and $(\alpha\beta)_{sac}$ represent random main effect of the control sample s and interactions between *sample* and *array*, and ε_{2sapc} 's are random errors. The normal probability plot of the residuals (Figure 3, left panel) indicates that the distribution has heavy tails. We also calculated residual standard deviations using formula (3.2) for all combinations of subjects and genes. The histogram of residual standard deviations on the right panel of Figure 3 indicates that there is a large variation among error variances.

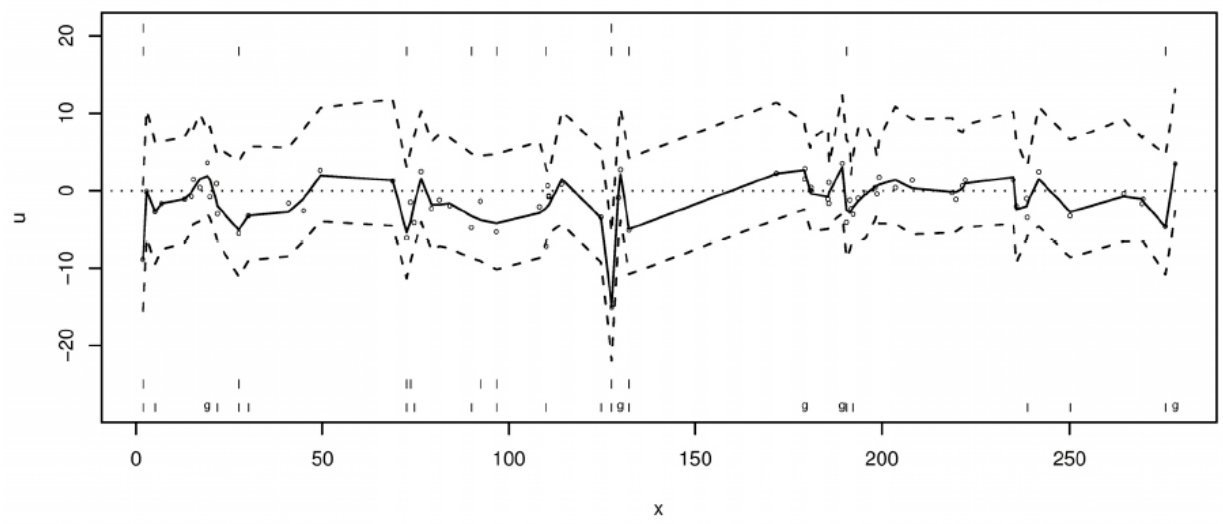

Figure 4. Profile of the smoothed t-like statistic for chromosome 1 in one case sample. Circles are the *u* statistics. The solid line is the HAS fit. Two dotted lines are 95% bootstrap confidence intervals with *B*=5000 bootstrap samples. Locations with significant alteration at 95% (99%) level based on the SHAS and SMOT methods are marked in the first and second rows respectively at the bottom (top) of the plot. The letter *l* represents a loss and the letter *g* represents a gain.

We now consider the reduced statistic t_{si} and u_{si} as functions of the distance x_i. Figure 4 shows the profile of the statistic u for chromosome 1 in a case sample. Our goal is to detect regions of the profile that deviate significantly from zero; the clones in these regions are then classified as having statistically significant alterations (gains or losses).

For a fixed case sample s and chromosome, let y_i equal to the standard t-statistic t_{si} (corresponds to the HAS method) or the smoothed t-like statistic u_{si} (corresponds to the SHAS method). We assume that

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1,...,n, \tag{3.9}$$

where n is the number of observed clones in this chromosome, f is a smooth function of x_i and ε_i's are random errors with mean zero and variance σ^2. For simplicity we transformed the variable x into the interval [0,1]. Suppose that $f \in W_2[0,1]$ (25), where

$$W_2[0,1] = \left\{ f : f' \text{ abs. cont.}, \int_0^1 (f'')^2 dx < \infty \right\}.$$

The conventional cubic smoothing spline is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx, \tag{3.10}$$

where the first part is the least square, the second part is a penalty for the roughness of the function, and λ is a smoothing parameter which controls the trade-off between the goodness-of-fit and the smoothness of the estimate. The minimizer of (3.10) can be represented by

$$f_\lambda(x) = d_1\phi_1(x) + d_2\phi_2(x) + \sum_{i=1}^{n} c_i\xi_i(x), \tag{3.11}$$

where $\phi_1(x) = 1, \phi_2(x) = x - 0.5,$ and

$$\xi_i(x) = \int_0^{\min(x_i, x)} (x_i - u)(x - u)du .$$

Since the smoothness (complexity) of the function in the whole interval is controlled by a single smoothing parameter λ, $f_\lambda(x)$ may over-smooth in regions where *f* is rough and under-smooth in regions where *f* is smooth. The solution (3.11) uses all distinct design points as knots. One usually needs more knots in regions where the shape of *f* is complex and less knots in regions where the shape of *f* is simple. Thus a subset of diligently placed knots can improve a spline estimate for a spatial inhomogeneous f such as the profile of u statistics. Since the true function f is unknown, we need an objective method to select knots based on data. The HAS procedure proposed in (21) is a powerful method for this purpose. It selects a subset of bases from $\{\xi_1(x),...,\xi_n(x)\}$ as follows:

1. *initialization*: set the maximum number of bases q ($q \geq 2$) and the inflated degrees of freedom (IDF). Start with *k*=2 and two bases $\{\phi_1(x), \phi_2(x)\}$;

2. forward stepwise selection: for $k = 3,...,q,$ choose the *k*th basis $\xi_{i_k}(x)$ to maximize the reduction in the residual sum of squares (RSS);

3. *optimal number of bases*: choose $k \geq 2$ as the minimizer of the generalized cross-validation (GCV) score

$$\text{GCV}(k) = \text{RSS}/(1 - (2 + (k-2) \times \text{IDF})/n)^2;$$

4. *backward elimination*: perform backward elimination to the selected bases. Decide the final number of bases by the Akaike Information Criteria (AIC);

5. fit: fit a standard or ridge regression model to the final selected bases.

The IDF is used to account for the added flexibility in adaptively selected bases. In general a good choice of IDF is 1.2 (16). For ABCGH data, we found that this choice of IDF sometimes under- or over-estimates the number of bases. Therefore we added the backward elimination step to the original HAS procedure. Our experiences suggest that the combination of a smaller IDF (1 or 1.1) with the backward elimination step provides better fits. We also found that the ridge regression step in the original HAS procedure can lead to over-smoothing for ABCGH data. Thus we added the standard regression using a numerically stable procedure as an option in the last step.

We used the following bootstrap procedure to construct confidence intervals and calculate p-values. Denote the HAS estimates of *f* and σ as $\hat{f}$ and $\hat{\sigma}$ respectively. We first generated a bootstrap sample

$$y_i^* = \hat{f}(x_i) + \varepsilon_i^*, \quad i = 1, ..., n$$

where ε_i^* are sampled with replacement from residuals. Denote the HAS estimates of f and σ based on the bootstrapped sample as $\hat{f}^*$ and $\hat{\sigma}^*$ respectively. Let $D_i^* = (\hat{f}^*(x_i) - \hat{f}(x_i))/\hat{\sigma}^*$. Repeat this process *B* times and denote $D_i^*(b)$ as the D_i^* statistic based on the *b*th bootstrapped sample. Let be the lower and upper $d_{\alpha/2}$ percentile of $\{D_i^*(b), b = 1, ..., B\}$, respectively. Then the $(1-\alpha)100\%$ bootstrap confidence interval is $(\hat{f}(x_i) - d_{1-\alpha/2}\hat{\sigma}, \hat{f}(x_i) - d_{\alpha/2}\hat{\sigma})$ (26). Regions with zero outside the confidence intervals are classified as ones with statistically significant alterations.

We can also calculate the p-values as

$$p_i = \#\left\{b : |D_i^*(b)| > |\hat{f}(x_i)|/\hat{\sigma}\right\}/B$$

Then clones with $p_i \le \alpha$ are significant at level α.

3.6. False Discovery Rate

We used the False Discovery Rate (FDR) to circumvent the problem of multiple comparisons (27). For a fixed case sample and clone *c*, $c = 1, ..., C$, let p_c be the p-value for the hypothesis of no alteration at clone c based on a method such as standard t-test, smoothed t-test or HAS. Let

$$i_\alpha = \arg\max_i \left\{ p_{(i)} \le \frac{i}{C}\frac{\alpha}{p_0} \right\},$$

where $p_{(i)}, i = 1, ..., C,$ are ordered values of p_i for all clones in this case sample, and p_0 is the proportion of clones without alteration. Then we classify clones with p-values less or equal to $p_{(i_\alpha)}$ as significant. This should guarantee FDR $\le \alpha$. Since p_0 is typically unknown, we assumed $p_0 = 1$ in our calculations, yielding more conservative results.

4. RESULTS

We applied the proposed methods to the endometrial tissues data described in the previous section. We also performed simulation studies to evaluate the performance of these four methods. In this section we show partial results of the data analyses and results from simulation studies.

4.1. Genomic alterations associated with endometriosis

We use chromosome 1 in one case sample for illustration. The same analyses were performed for all chromosomes except Y chromosome in all five case samples. The complete results will be reported elsewhere. Simulation studies (see below) indicate that the smoothed t-statistic performs better than the standard t-statistic. Thus we present results based on the SMOT and SHAS methods only.

The profile of smoothed t-like statistics and its HAS fit are shown in Figure 4. Residual plots indicate that the random errors did not follow a normal distribution.

Figure 5 shows p-values and rejection regions. The SMOT method concludes that 23 clones have significant alterations (18 losses and 5 gains) at 0.05 level since there are 23 points below the dotted line marked as α =0.05. Similarly, 10 of these 23 clones are significant at the level of 0.01, and 16 are significant with FDR ≤ 0.1. The SHAS detected fewer alterations: 6 clones have significant alterations (all losses) at 0.05 level, 2 clones at the level of 0.01 and only 1 clone with FDR ≤ 0.1.

Preliminary validation studies based on analysis of heterozygosity (LOH) and real-time PCR have confirmed genomic alterations detected by SHAS for this data set (data not shown), suggesting that the results based on SHAS are reliable.

4.2. Simulation studies

We conducted simulations to evaluate and compare four proposed methods: standard t-test (STNT), smoothed t-test (SMOT), HAS to the standard t-statistics in (3.4) (HAS) and HAS to the smoothed t-statistics in (3.7) (SHAS). We selected simulation parameters as close to the real data as possible. Specifically, we had $n_1 = 1$ (one case sample), $n_2 = 5$ (five control samples), two arrays and two panels. We used the same design points x_c's as in chromosome 2 with 93 clones. We generated observations for the case sample from model (3.1) with nine different signal strengths which are combinations of three shapes and three strengths. Specifically, we considered the following functions for the mean μ_{11c} in (3.1): $\mu_{11c} = \delta$ if $c \in \mathrm{S}$, or

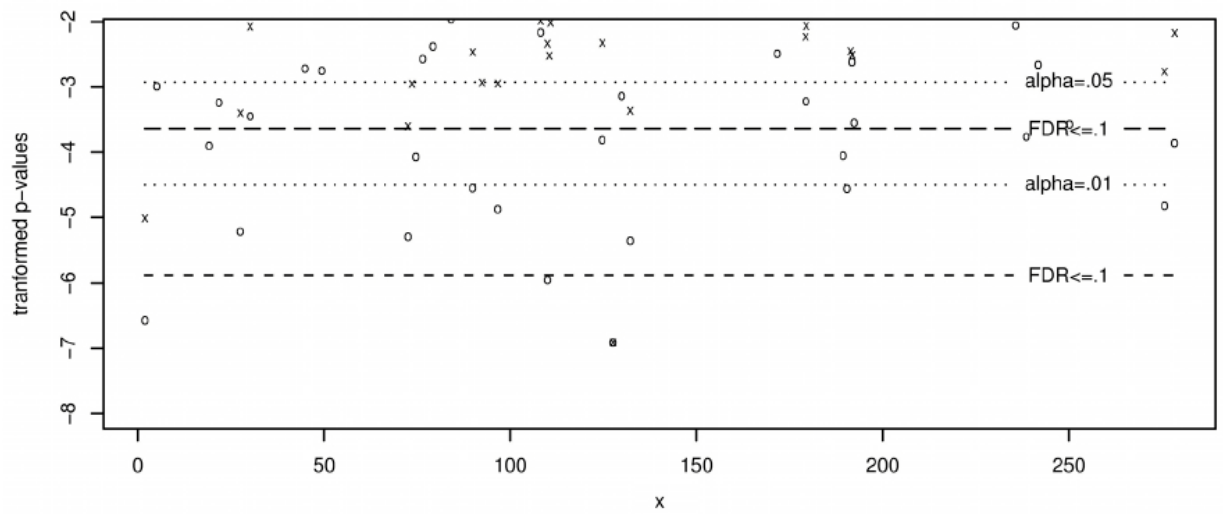

Figure 5. Transformed p-values, $\log((p+a)/(1-p+a))$ with a=0.001, from the SMOT method (circles) and the SHAS method (crosses). We only show p-values which are smaller than 0.1. Regions below two dotted lines from the bottom up represent rejection regions with α =0.01 and α =0.05 respectively. Regions below the long and short dashed lines represent rejection regions with FDR ≤ 0.1 for the SMOT and SHAS methods respectively.

0 if $c \in S^c$ where S is the set of clones with alterations, and δ represents the strength of the signal. We considered three choices of S: $S=\{50\}, S=\{49,50,51\}$ and $S=\{48,49,50,51,52\}$ which correspond, respectively, to a single, three consecutive and five consecutive clones with alterations, and three choices of δ: $\delta=-0.3, \delta=-0.5$ and $\delta=-0.7$ which correspond to weak, medium and strong signals. In reality the random errors ε_{11apc} in (3.1) may not follow a normal distribution. In fact, the left panel of Figure 3 indicates that the distribution has heavy tails. To mimic the real data, we generated ε_{11apc} as random samples from the collection of all residuals from fitting model (3.8) to all clones. For each clone, five observations as control samples were generated. Again, the distribution may not be normal. To be realistic, we collected all sets $\square_c = \{\bar{N}_{2s\cdot c} - \bar{N}_{2\cdot\cdot c}, s=1,...,5\}$ from five control samples in the real data. Note that each set $\square_c$ is centered and clones with less than five control samples due to missing data are excluded. For each clone, we then generated a set of five control samples as a random sample from the collection of all sets $\square_c$. We used 500 bootstrap samples in the computation of bootstrap p-values for the HAS and SHAS methods. The following results are based on these 500 replications.

Each method returns a vector of p-values which we denote as $p_c, c=1,...,93$. For a cutoff value α, we calculated

$$\text{type I error} = \#\{c \in S^c, p \leq \alpha\}/\#S^c$$

and

$$\text{power} = \#\{c \in S, p_c \leq \alpha\}/\#S$$

where # denotes the cardinality of the set.

We then computed the type I error and power by averaging over 500 replications. Figure 6 shows plots of power vs type I errors for a set of cutoff values such that the type I error is between 0.001 and 0.05. These curves show potential performance of different methods. Viewed from left panel to the right, we can see how power increases with increasing signal strength, as expected. Because curves based on the smoothed t-like statistics are always above those based on the standard t-statistics, we can conclude that the smoothed t-statistics always improve the performance. It should be noted that since we made no assumption on the functional form of the variance such as model (3.6) in the simulation, we expect greater improvement when (3.6) is true. Since the distributions are heavily tailed, smoothing (shrinking) variances tend to make the t-statistic more robust to outliers. HAS performs considerably worse than the t-test in detecting a single alteration. However, the powers of the HAS and SHAS methods increase as the cluster of alteration increases while the power of the t-tests remains virtually unchanged. Therefore, while the t-tests are more powerful in detecting an isolated alteration as expected, the HAS methods are more powerful in detecting a cluster of alterations. This also suggests that as the density of the CGH array increases, the HAS methods will perform even better than the t-like statistics since spatial correlation increases as the density of the array increases.

Note that Figure 6 only indicates potential performance of the proposed methods. Since, in practice, the null distributions for all four methods are only

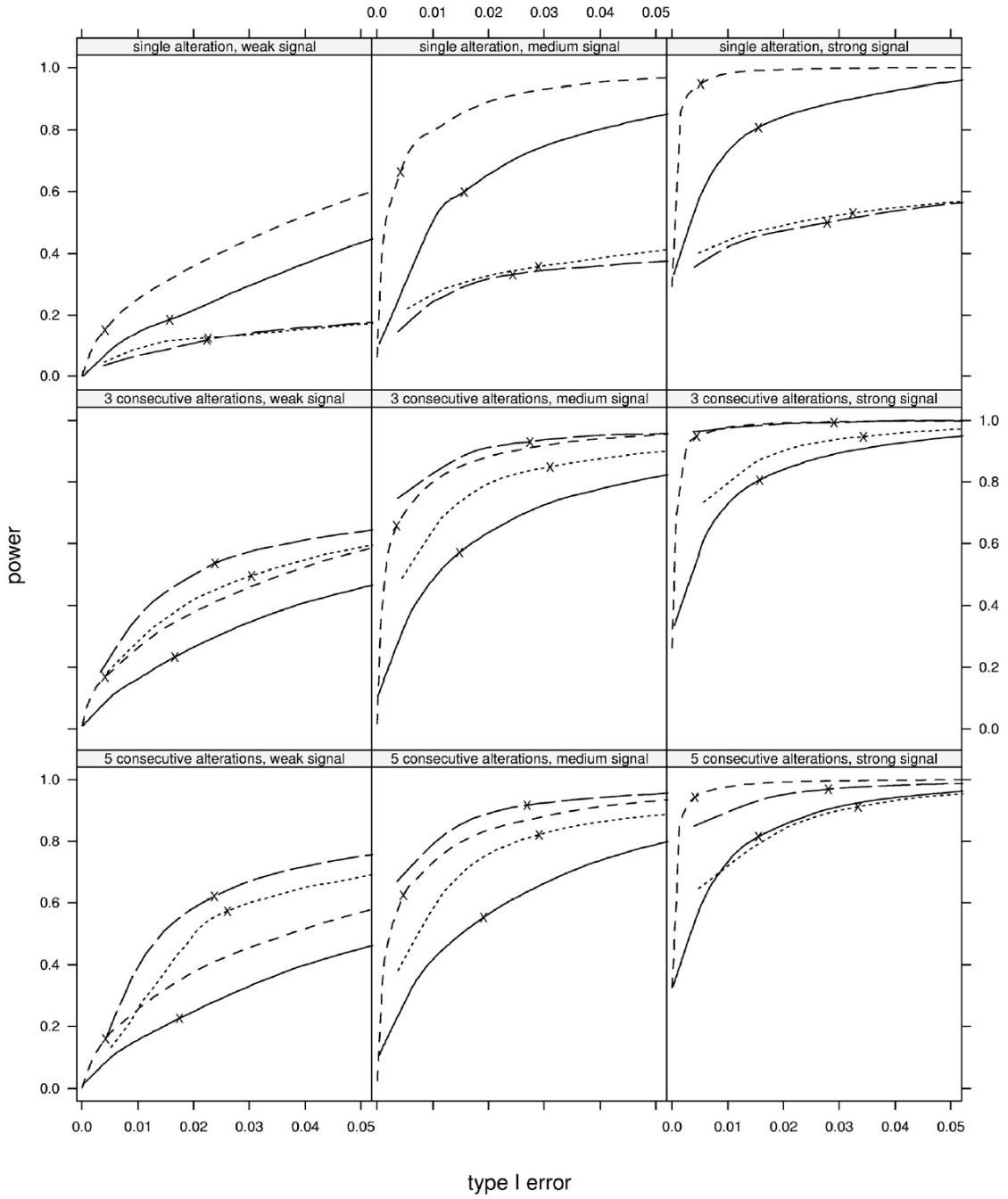

Figure 6. Comparison of four different methods: STNT plotted as solid lines, SMOT plotted as short dashed lines, HAS plotted as dotted lines, and SHAS plotted as long dashed lines. The actual type I errors with α =.05 are marked as crosses on each curve. Simulation settings are shown in the stripes.

approximated, the identified statistically significant regions based on these approximations may not be accurate. Thus the type I error actually committed at level α may be different from the nominal value α . From the actual type I error with α =.05 marked on each curve, we can see that all methods are conservative in the sense that they all commit smaller type I errors than the nominal value of α =0.05. The SMOT method tends to be very conservative with type I errors always smaller than 0.01. Consequently, its potential power may not be realized in practice. As can be seen in the middle panel in the third row, for example, the actual power of the SHAS method (0.922) is much larger than that of the SMOT method (0.628) even though the two curves are close to each other. From the left panel of the first row, we can see that the actual powers at α =.05 level are similar for all four methods even though

the SMOT method could achieve much more. Clearly, finding more accurate approximations to the null distribution warrants further research.

5. DISCUSSION

Surpassing conventional CGH in throughput and resolution, the emerging ABCGH also has the promise of becoming an automated tool in cytogenetic diagnosis and delineation of tumorigenesis (9). Compared with tremendous efforts in developing the ABCGH technology, the statistical methodology for detection of genomic alterations based on ABCGH data is still in its infancy.

Sharing with the cDNA microarrays many issues such as normalization and statistical identification, the ABCGH data also pose an additional challenge of high spatial correlations between neighboring clones.

Our proposed methods for the identification of genomic alterations using ABCGH have several advantages. First, they require much less restrictive assumptions on the variances of the test:reference ratios. Second, by smoothing variances along the genome, the smoothed t-like statistics are more robust to outliers. This is especially important for experiments with relatively small number of control samples and/or ratios follow a distribution with heavy tails. Note that some outliers had been removed from the control samples for data analysis and simulation. Therefore, the outlier problem may be more challenging in practice than in our simulation. Third, by incorporation of neighboring data with high correlations, HAS is more efficient and robust in detecting clusters of alterations. It handles nicely the inhomogeneous "curvature"of the ratio profiles along the genome. Our inference procedure based on bootstrap dose not require the normality assumption. In view of the observation that there are consistent, region-specific variations in ratio profiles, which may be caused by differences in hybridization characteristics of the incorporated fluorochromes and by local variation in chromosome structures (such as telomeres or centromeres) (16), and, in particular, the functional form of the variation, as a function of clone locations, is typically unknown, the HAS and SHAS are well suited for the ABCGH data.

With the further progression of the Human Genome Project, more clones will soon be available for ABCGH. This increasing density of ABCGH will further boost its power to detect smaller genomic alterations that might be responsible for the phenotype of interest. At the same time, the higher density also would increase spatial correlations in the neighboring clones. As the density increases, we expect that performance of the 2 SD method and the t-statistics method will become progressively worse. In contrast, the HAS method should perform better as resolution become higher. In addition, as the method makes no distributional assumptions, it is also well-suited for automated analysis.

6. ACKNOWLEDGMENTS

This research was supported by grants GM58533 (YW) and GM56515 (SWG) from the National Institute of Health and a grant from the Endomtriosis Association. We would like to thank Dr. Yan Wu for performing the ABCGH experiments, Drs. Andre Balla, Zainab Basir and Estil Strawn for providing clinical materials and preparation of tissue samples, and Dr. Xujing Wang for pre-processing the image data.

7. REFERENCES

1. Kallioniemi A, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258, 818-821 (1992)

2. Bentz M., A. Plesch, S. Stilgenbauer, H. Dohner, P. Lichter: Minimal sizes of deletions detected by comparative genomic hybridization. *Genes Chromosomes Cancer 21*, 172-175 (1988)

3. Kirchhoff M., T. Gerdes, J. Maahr, H. Rose, M. Bentz, H. Dohner H. & C. Lundsteen: Deletions below 10 megabasepairs are detected in comparative genomic hybridization by standard reference intervals. *Genes Chromosomes Cancer 25*, 410-413 (1999)

4. Veltman J.A., E. F. Schoenmakers, B. H. Eussen, I. Janssen, G. Merkx, B. van Cleef, C. M. van Ravenswaaij, H. G. Brunner, D. Smeets & A. G. van Kessel: High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am J Hum Genet. 70*, 1269-1276 (2002)

5. Wessendorf S., B. Fritz, G. Wrobel, M. Nessling, S. Lampel, D. Goettel, M. Kuepper, S. Joo, T. Hopman, F. Kokocinski, H. Dohner, M. Bentz, C. Schwaenen & P. Lichter: Automated screening for genomic imbalances using matrix-based comparative genomic hybridization. *Lab Invest. 82*, 47-60 (2002)

6. Solinas-Toldo S., S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Dohner, T. Cremer & P. Lichter: Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer 20*, 399-407 (1997)

7. Pinkel D., R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray & D. G. Albertson: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet. 20*, 207-211 (1998)

8. Pollack J. R., C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, P. O. Brown: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet. 23*, 41-46 (1999)

9. Snijders A. M., N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel & D. G. Albertson:

Assembly of microarrays for genome-wide measurement of DNA copy number**.** *Nat Genet. 29***,** 263-264 (2001)

10. Schena M., D. Shalon, R. W. Davis & P. O. Brown: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science 270*, 467-470 (1995)

11. Piper J., D. Rutovitz, D. Sudar, A. Kallioniemi, O. P. Kallioniemi, F. M. Waldman, J. W. Gray & D. Pinkel: Computer image analysis of comparative genomic hybridization. *Cytometry 19*, 10-26 (1995)

12. Cher M.L., G. S. Bova, D. H. Moore, E. J. Small, P. R. Carroll, S. S. Pin, J. I. Epstein, W. B. Isaacs & R. H. Jensen: Genetic alterations in untreated metastases and androgen-independent prostate cancer detected by comparative genomic hybridization and allelotyping. *Cancer Res. 56*, 3091-3102 (1996)

13. du Manoir S., O. P. Kallioniemi, P. Lichter, J. Piper, P. A. Benedetti, A. D. Carothers, J. A. Fantes, J. M. Garcia-Sagredo, T. Gerdes & M. Giollant: Hardware and software requirements for quantitative analysis of comparative genomic hybridization. *Cytometry 19*, 4-9 (1995)

14. Lundsteen C., J. Maahr, B. Christensen, T. Bryndorf, M. Bentz, P. Lichter & T. Gerdes: Image analysis in comparative genomic hybridization. *Cytometry 19*, 42-50 (1995)

15. Schleger C., N. Arens, H. Zentgraf, U. Bleyl & C. Verbeke: Identification of frequent chromosomal aberrations in ductal adenocarcinoma of the pancreas by comparative genomic hybridization (CGH). *J Pathol. 191*, 27-32 (2000)

16. Moore D.H., M. Pallavicini, M. L. Cher & J. W. Gray: A t-statistic for objective interpretation of comparative genomic hybridization (CGH) profiles. *Cytometry 28*,183-90 (1997)

17. Bruder C.E., C. Hirvela, I. Tapia-Paez, I. Fransson, R. Segraves, G. Hamilton, X. X. Zhang, D. G. Evans, A. J. Wallace, M. E. Baser, J. Zucman-Rossi, M. Hergersberg, E. Boltshauser, L. Papi, G. A. Rouleau, G. Poptodorov, A. Jordanova, H. Rask-Andersen, L. Kluwe, V. Mautner, M. Sainio, G. Hung, T. Mathiesen, C. Moller, S. M. Pulst, H. Harder, A. Heiberg, M. Honda, M. Niimura, S. Sahlen, E. Blennow, D. G. Albertson, D. Pinkel, J. P. Dumanski: High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH. *Hum Mol Genet. 10*, 271-282 (2001)

18. Wilhelm M., J. A. Veltman, A. B. Olshen, A. N. Jain, D. H. Moore, J. C. Presti Jr, G. Kovacs & F. M. Waldman: Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Res. 62*, 957-960 (2002)

19. Piper J., S. Stegenga, E. Pestova, H. Marble, M. Lucas, K. Wilber & W. King: In *Third Euroconference on Quantitative Molecular Cytogenetics*, Rosenn, Sweden, July 4-6, 109-114 (2002)

20. Carothers A.D.: A likelihood-based approach to the estimation of relative DNA copy number by comparative genomic hybridization. *Biometrics 53*, 848-856 (1997)

21. Luo, Z. & G. Wahba: Hybrid Adaptive Splines. *J. Am. Stat. Assoc. 92*, 107-116 (1997)

22. Telenius H., N. P. Carter, C. E. Bebb, M. Nordenskjold, B. A. Ponder & A. Tunnacliffe: Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics 13*, 718-725 (1992)

23. Wang X., S. Ghosh & S. W. Guo: Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.* , E75-5 (2001)

24. Yang Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, T. P. Speed: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic
variation. *Nucleic Acids Res 30*, E15 (2002)

25. Wahba, G.: Spline Models for Observational Data. SIAM, Philadelphia (1990)

26. Wang, Y. & G. Wahba: Bootstrap Confidence Intervals for Smoothing Spline Estimates and Their Comparison to Bayesian Confidence Intervals. *J. Statist. Comput. Simul*. 51, 263-279 (1995)

27. Benjamini, Y.D. & Y. Hochberg: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal. Stat. Soc. B 57*, 289-300 (1995)

Abbreviations: ABCGH: array-based CGH, BAC: bacterial artificial chromosomes, CGH: comparative genomic hybridization, cM: centi-Morgan, DOP-PCR: degenerate oligonucleotide primed-polymer chain reactions, FDR: false discovery rate, GCV: generalized cross-validation, HAS: hybrid adaptive spline, IDF: inflated degrees of freedom, Mb: Mega base-pairs

Key Words: Array-based comparative genomic hybridization, genomic alterations, hybrid adaptive splines, microarrays, statistical methods, t-statistic, Review

Send correspondence to: Yuedong Wang, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106, Tel: 805-893-4870, Fax: 805-893-2334, E-mail: yuedong@pstat.ucsb.edu