Change-Point Analysis of Paired Allele-Specific Copy Number Variation Data

YINGLEI LAI

ABSTRACT

The recent genome-wide allele-specific copy number variation data enable us to explore two types of genomic information including chromosomal genotype variations as well as DNA copy number variations. For a cancer study, it is common to collect data for paired normal and tumor samples. Then, two types of paired data can be obtained to study a disease subject. However, there is a lack of methods for a simultaneous analysis of these four sequences of data. In this study, we propose a statistical framework based on the change-point analysis approach. The validity and usefulness of our proposed statistical framework are demonstrated through the simulation studies and applications based on an experimental data set.

Key words: statistics, DNA arrays, genetic variation.

1. INTRODUCTION

WITH THE CURRENT MICROARRAY TECHNOLOGIES, it is feasible to collect different types of genome-wide high-throughput data (Cancer Genome Atlas Research Network, 2008). The recent single nucleotide polymorphism (SNP) based arrays enable us to collect genome-wide allele-specific copy number variation data (Assie et al., 2008). In addition to DNA copy number information measured by the well-developed array-based comparative genomic hybridization (array-CGH) technology, genotype information can also be measured by the recent SNP-based arrays. This improvement allows us to understand the allelic imbalance for genome-wide genotype variations as well as copy number variations (Van Loo et al., 2010). These two different types of information are usually called the B Allele Frequency (BAF) data and the Log R Ratio (LRR) data. Since the introduction of this technology, many statistical methods have also been published for the analysis of allele-specific copy number variation (CNV) data (Dellinger et al., 2010). Most of these methods are developed based on either hidden Markov models (Wang et al., 2007; Liu et al., 2010; Chen et al., 2011) or change-point analysis (Olshen et al., 2011). It has been demonstrated that more efficient analysis results can be obtained when both LRR and BAF data are considered (Wang et al., 2007; Li et al., 2009). (Notice that we also use the term “allele-specific CNV” to distinguish CNV data collected in early years [Zhao et al., 2004]. These data have been collected mostly based on the array-CGH technology by which the allelic imbalance information cannot be collected. The background about “allele-specific CNV” has been briefly described in Assie et al. [2008].

For cancer studies, it is common to collect allele-specific CNV data for paired samples (Staaf et al., 2008). For a study disease subject, the data from a paired normal and tumor samples can be collected. Since both samples are from the same subject, this approach enables us to discover chromosomal changes that are

potentially important to a disease progression. To our best knowledge, the existing analysis methods for the paired allele-specific CNV data are based on the difference (or ratio) data from the paired samples (Olshen et al., 2011). Statistical methods for a simultaneous change-point analysis of paired allele-specific CNV data have not been well addressed in the literature. However, as we discussed below, such an analysis can be useful to generate informative discoveries.

With paired allele-specific CNV data from a study disease subject, a simultaneous partition of two same chromosomes for the normal and tumor samples can generate more biologically informative results. We can detect allele-specific copy number variations on a chromosome for either the normal or the tumor sample. The allele-specific copy number variations on a chromosome for both samples can also be detected. Since both samples have been collected from the same disease subject, the detection of these variations can be informative for us to understand the molecular changes during the disease progression. To further understand these molecular changes, it is necessary to investigate the same chromosomal regions for the paired samples. If we perform a partition process for the paired normal and tumor samples separately, it can be difficult to draw a clear conclusion since the partitioned chromosomal regions and the related estimates for different samples can be inconsistent. Therefore, a simultaneous partition is necessary for the analysis of paired allele-specific CNV data.

To detect multiple change-points, we need a statistical test for the significance evaluation (p -value) of a change-point. For a simultaneous change-point analysis of paired sample data for a given chromosome, we will specify the related hypothesis as no changes for either the normal or the tumor sample. (The F -test described below is a typical example.) Since this null hypothesis is more specific than the frequently used null hypothesis for difference data, we expect an improved detection power. A further improved detection power is expected if we can simultaneously analyze the genotype variation (BAF) data and the copy number variation (LRR) data collected for a paired sample.

There are several statistical issues when we perform a change-point analysis for allele-specific CNV data. There is usually a considerable dependence between the data from the paired normal and tumor samples from the same study disease subject. There is also a considerable dependence among the data from adjacent SNPs. For the experimental allele-specific CNV data (including both BAF and LRR data), it is usually difficult to specify any theoretical probability distributions for the accurate calculation of p -values. Furthermore, an efficient approach is necessary for the simultaneous analysis of BAF and LRR data.

The issue of multiple hypothesis testing (or multiple comparisons) is also a critical concern when we analyze genome-wide allele-specific CNV data. The goal of our change-point analysis is to partition a given chromosome into smaller consecutive and non-overlapped regions such that the data from any two adjacent regions are significantly different. We have recently proposed a modified dynamic programming algorithm for a global optimization (Lai, 2011). The number of multiple comparisons (also the time complexity) is up to the cubic order (in term of the number of SNPs). Instead of a global optimization approach, we can consider the recursive partition approach that has been used in the circular binary segmentation (CBS) procedure (Olshen et al., 2004). Then, the number of multiple comparisons can be up to the cubic order. We can also consider the recursive combination approach that has been suggested by Wang et al. (2005) and the number of multiple comparisons is simply the linear order. We have shown that the performance of these three approaches can be comparable when the sample size (number of observations at each time point or marker) is relatively small (Lai, 2011). Since the number of SNPs for a chromosome is usually large and it is critical to maintain a satisfactory control of false positives, we propose to use the recursive combination approach. (Notice that the number of multiple comparisons will be much larger for the change-point analysis of deep sequencing data.)

Our solution to the change-point analysis of paired allele-specific CNV data is a novel statistical framework. In the following, we first give a brief summary before the technical details of our proposed statistical framework. Then, we use a published experimental data set to illustrate our approach. We also conduct simulation studies to confirm our choice of test statistics and also the calculation and approximation of p -values. The simulation data are generated based on the experimental data. Then, we focus on the experimental data for chromosomes 20 and 21 for the applications of our approach.

2. METHODS

2.1. Summary

For the allele-specific CNV data of a chromosomal region (or a whole chromosome), our goal is to partition the region into smaller consecutive and non-overlapped regions such that the data from any two

adjacent regions are significantly different at a given level α_C (for the claim of a detection of significant change-point). The statistical significance is evaluated based on both the LRR data and the BAF data. As we have explained above, we propose to use the hierarchical clustering approach to the partition of chromosomal region. Among the feasible partitions, we select the one that minimizes the sum of squared errors (SSE) for the allelic copy numbers.

Instead of using one sample or the difference (or ratio) between two samples, we intend to keep both samples in our analysis. Then, we have to consider the dependence between two samples since they are from the same subject. A traditional approach is to fit a mixed effects model with some random effects so that the dependence between two samples can be considered. However, this increases the computing burden significantly since each model fitting requires a time-consuming numerical optimization. For our analysis, it is computationally crucial that the test procedure is simple and fast. Therefore, we propose to use the simple analysis of variance (ANOVA) approach but also use the permutation procedure for p -value calculation. Furthermore, we suggest a novel method to approximate extremely small p -value so that the global false positive control can be well addressed.

Before the description of technical details, we give the necessary mathematical notations. We use U to denote the LRR variable (Staaf et al., 2008) and u_{ij} to denote the LRR observation at location i for sample j , where subjects $j = 1$ and $j = 2$ are the paired normal and abnormal samples, respectively. Similarly, we use V to denote the BAF variable (Staaf et al., 2008) and v_{ij} to denote the observation at location i for sample j . Since SNP genotypes are not phased, a simple transformation of BAF variable is frequently used in practice: $\tilde{v}_{ij} = |v_{ij} - 0.5|$ (Staaf et al., 2008).

2.2. A change-point test

Let Y be either the LRR variable or the transformed BAF variable (see above for detail), we assume a simple regression model for its observations within a block of contiguous SNPs at locations $S = \{k+1, k+2, \dots, k+m_1+m_2\}$:

$$y_{ij} = \alpha_j + \epsilon_{ij}; \quad i = k+1, k+2, \dots, k+m_1+m_2; \quad j = 1, 2. \quad (1)$$

Notice that k is a reasonable integer value so that S can be a possible block on a given chromosome. (The symbols m_1 and m_2 are used later as the numbers of SNPs [sample sizes] within two adjacent blocks [sample groups].) In this model, α_1 and α_2 are the mean levels for the paired normal and abnormal samples, respectively. ϵ_{ij} is the noise. We assume zero for the mean and a common variance for each ϵ_{ij} . (Notice that we do not assume any independence among different ϵ 's. Furthermore, the distribution of noise is not assumed and it is not necessarily normal, which is clearly the case for BAF data.) In this study, we still use the sample mean to estimate $\hat{\alpha}_j = \sum_{i=k+1}^{k+m_1+m_2} y_{ij} / (m_1+m_2)$ and then calculate $SSE_0 = \sum_{j=1}^2 \sum_{i=k+1}^{k+m_1+m_2} (y_{ij} - \hat{\alpha}_j)^2$. (SSE means the sum of squared errors.)

Consider a partition that split the block S into two smaller adjacent blocks: $S_1 = \{k+1, k+2, \dots, k+m_1\}$ and $S_2 = \{k+m_1+1, k+m_1+2, \dots, k+m_1+m_2\}$. For this partition, we also consider a simple regression model for Y :

$$y_{ij} = \alpha_j + \delta(i > k+m_1)\beta_j + \epsilon_{ij}; \quad i = k+1, k+2, \dots, k+m_1+m_2; \quad j = 1, 2. \quad (2)$$

In this model, for the SNPs in S_1 , α_1 and α_2 are the mean levels for the paired normal and abnormal samples, respectively; for the SNPs in S_2 , $\alpha_1 + \beta_1$ and $\alpha_2 + \beta_2$ are the mean levels for the paired normal and abnormal samples, respectively. $\delta(\cdot)$ is the indicator function that $\delta(\text{TRUE}) = 1$ and $\delta(\text{FALSE}) = 0$. Again, we still use the sample mean to estimate $\hat{\alpha}_j = \sum_{i=k+1}^{k+m_1} y_{ij} / m_1$, $\hat{\beta}_j = \sum_{i=k+m_1+1}^{k+m_1+m_2} y_{ij} / m_2 - \hat{\alpha}_j$ and calculate $SSE_1 = \sum_{j=1}^2 \sum_{i=k+1}^{k+m_1+m_2} (y_{ij} - \hat{\alpha}_j - \delta(i > k+m_1)\hat{\beta}_j)^2$.

A change-point can be claimed when the partition reduces the SSE significantly, i.e. the difference between SSE_0 and SSE_1 is significantly large. Statistically, we test the hypothesis $H_0: Y$ in block S follows the model (1) vs. $H_1: Y$ in block S follows the model (2). For the test statistic, we still consider the traditional F -test (Casella and Berger, 2002):

$$F = \frac{(SSE_0 - SSE_1)/2}{SSE_1/(m_1+m_2-4)}. \quad (3)$$

Under the assumption that ϵ 's are independently and identically distributed (i.i.d.) and they are normally distributed with mean zero and a common variance, it is well-known that F follows an F -distribution with

degrees of freedom 2 and $m_1 + m_2 - 4$. However, in this study, the observations from adjacent SNPs can be correlated and the noise distributions are usually not normal, then the theoretical F -distribution is not appropriate for calculating the p -value. Therefore, the widely used permutation procedure is our choice for the p -value calculation.

2.3. Mean and variance changes

For the LRR data, there is no data transformation. It is well known that the above F -test evaluates the change of population means. However, a transformation is performed before we analyze the BAF data. Analyzing the transformed data is actually related to the test of variance change. In a simple one-way ANOVA, if we transform the data to be the absolute deviation to the sample mean or sample median, then the traditional F -test will be the well-known Levene's test or Brown-Forsythe's test, respectively. This approach has been widely used to evaluate the homogeneity of variance (Levene, 1960; Brown and Forsythe, 1974; Gastwirth et al., 2009). Similarly, for the transformed BAF data, the F -test specified in Equation (3) can also be considered as a test for the homogeneity of variance. If we replace 0.5 with either the sample mean or sample median in the transformation formula for the BAF data, then the F -test specified in Equation (3) is an extension of Levene's test or Brown-Forsythe's test, respectively. It is necessary to understand which test is more powerful for the experimental BAF data. We will consider a simulation study in which we sample from the experimental data randomly. Since a relatively large proportion of data show clear changes of BAF patterns, we consider this a valid approach to the comparison of different tests. Furthermore, we will also include the F -test (Equation 3) for the original BAF data to evaluate changes of population means (although we do not expect to frequently observe any clear changes of means for the BAF data).

We will also conduct a similar simulation study to understand the test power for the experimental LRR data. Although it is common not to evaluate changes of variances for the LRR data (since we do not expect to frequently observe any clear changes of population variances for the LRR data), we still intend to confirm that testing the homogeneity of variance is not necessary.

2.4. Permutation p -value

To calculate the p -value for an F -test, we use the computer-intensive permutation approach. The permutation test has been widely used in practice (Good, 2005) and its procedure is described as below:

1. Permute $\{1, 2, \dots, m_1 + m_2\}$ randomly to obtain $\{l_1, l_2, \dots, l_{m_1 + m_2}\}$;
2. Construct $Y^* = \{y_{l_i, j}\}$ that is a permuted version of Y ;
3. Calculate the F -test (Equation 3) with Y^* as the observed data;
4. Repeat the above 3 steps R times and obtain a set of $\{F_1^*, F_2^*, \dots, F_R^*\}$;
5. Report the p -value = (number of $F^* \geq F$)/ R , where F is the F -test value based on the original data.

Essentially, we break the association between Y and the partition by re-assigning SNPs into S_1 and S_2 randomly. (The standard permutation procedure is used in this study.)

2.5. Extreme p -value approximation

One disadvantage of permutation p -values is that their precision is usually relatively low. For example, the computing burden will be high if we require a precision level of 10^{-6} for extremely small permutation p -values. A critical issue in genomics studies is the control of false positives since a large number of comparison tests are conducted simultaneously (Dudoit et al., 2003). To address the issue of multiple comparison adjustment, it is necessary to calculate extremely small p -values at a high precision level. However, this is usually not feasible in practice: the required computing burden may not be affordable if we perform a huge number of permutations for each of a large number of comparison tests.

Therefore, in addition to the above permutation procedure, we also need an approximation procedure to calculate extremely small p -values at a high precision level. Venkatraman and Olshen (2007) have developed an approximation method for the circular binary segmentation (CBS) procedure (Olshen et al. 2004). Since the change-points are detected based on an F -type test statistic in this study, we propose a different approximation method. It is based on the fact that an F -distribution is left-bounded and also the inverse of an F -distributed random variable is still F -distributed with the degrees of freedom switched. The procedure uses a linear or quadratic approximation to approximate permutation p -values $< 1/R$.

- If the permutation p -value $\geq 1/R$, then no approximation needed;
- Otherwise, let $t = 1/F$ and denote $\Phi_X(\cdot)$ the cumulative distribution function (c.d.f.) of a random variable X and $\Phi_X^{-1}(\cdot)$ the corresponding inverse function.
 - Linear approximation:
 - * $b = 1/\max(F_1^*, F_2^*, \dots, F_R^*)$;
 - * $a = 1/\Phi_{F_{2,m1+m2-4}}^{-1}(1 - 1/R)$;
 - * $\beta = b/a$;
 - * $p_L = \Phi_{F_{m1+m2-4,2}}(z)$, where $z = t/\beta$.
 - Quadratic approximation
 - *let $1/b$ be the top K permuted F values from $\{F_1^*, F_2^*, \dots, F_R^*\}$;
 - * $a = \{1/\Phi_{F_{2,m1+m2-4}}^{-1}(1 - k/R) : k = 1, 2, \dots, K\}$;
 - * $\beta = \frac{(\sum_{k=1}^K a^3)(\sum_{k=1}^K a^2 b) - (\sum_{k=1}^K a^4)(\sum_{k=1}^K ab)}{(\sum_{k=1}^K a^3)^2 - (\sum_{k=1}^K a^2)(\sum_{k=1}^K a^4)}$;
 - * $\gamma = \frac{(\sum_{k=1}^K ab) - \beta(\sum_{k=1}^K a^2)}{\sum_{k=1}^K a^3}$;
 - * $p_Q = \Phi_{F_{m1+m2-4,2}}(z)$, where $z = (\sqrt{\beta^2 + 4\beta\gamma t} - \beta)/(2\gamma)$.
- If p_Q is not a real value, or $p_Q \geq 1/R$ or $p_Q < p_L$, then report p_L ; otherwise, report p_Q .

Essentially, we approximate the p -value of an observed F value through a theoretical F -distribution with a predicted z . We first build either a linear model $b = f(a) = \beta a + \epsilon$ or a quadratic model $b = f(a) = \beta a + \gamma a^2 + \epsilon$; then we determine z as $f^{-1}(1/F)$. Since there is no guarantee that p_Q is always a real value or $p_Q \geq 1/R$, we will need to report p_L for these situations. Notice that p_L is simply predicted based on the largest permuted F values from $\{F_1^*, F_2^*, \dots, F_R^*\}$ since this setting guarantees $p_L < 1/R$. Furthermore, since we prefer a conservative p -value approximation, we will report p_L if $p_Q < p_L$. (In this study, we set $K = 10$ when $R = 200$.)

2.6. Joint test for LRR and BAF data

There are two types of information contained in an allele-specific CNV data set: the Log R Ratio (LRR) data and the B Allele Frequency (BAF) data. The LRR data are the measurements of DNA copy numbers and the BAF data are the measurements of minor allele proportions. Therefore, with the allele-specific CNV data for a given chromosome, we can analyze two types of changes: the change of DNA copy number and the change of allele heterozygosity. Since a joint analysis of both changes can increase the detection power, we conduct the following hypothesis testing in a general description: H_0 , where there are neither changes of DNA copy number nor changes of allele heterozygosity, versus H_1 , where there are some changes of either DNA copy number or allele heterozygosity (or both).

Statistically, we screen sequentially all the SNPs on a given chromosome to detect regions (blocks of SNPs) with significant changes. A chromosomal region will be identified if the data in this region are significantly different from the data in the next adjacent region. Consider the comparison between two adjacent blocks of SNPs: $S_1 = \{k+1, k+2, \dots, k+m_1\}$ and $S_2 = \{k+m_1+1, k+m_1+2, \dots, k+m_1+m_2\}$. For the LRR data, a p -value (from the permutation procedure or the approximation procedure, see above for details) can be obtained based on the F -type test described above. We use p_{LRR} and F_{LRR} to denote this p -value and the corresponding F -test value. Similarly, we can obtain p_{BAF} and F_{BAF} for the transformed BAF data.

Since LRR and BAF are two different types of measurements, also based on our data exploration, it is reasonable to assume that F_{LRR} and F_{BAF} are independent. Then, we can use the product of two p -values ($p_{LRR} \times p_{BAF}$) as the joint test. We denote $p_{LRR,BAF}$ the p -value of the joint test. For a joint test score t , the corresponding p -value can be theoretically calculated as:

$$p_{LRR,BAF} = \Pr(p_{LRR} \times p_{BAF} \leq t) = t - t \ln(t). \quad (4)$$

Notice that both p_{LRR} and p_{BAF} are uniformly distributed under the null hypothesis. Then, it is well known that the probability density function (p.d.f.) for the product of two standard uniform distributions is simply $-\ln(t)$, and the corresponding cumulative density function (c.d.f.) is $t - t \ln(t)$.

2.7. A clustering procedure

Since the hierarchical clustering has been widely used in practice (Hastie et al., 2009), it is not necessary to discuss its details here. When we cluster the SNPs on a given chromosome, a special feature is that only

adjacent SNPs can be clustered. Therefore, the required computing time can be much reduced (compared to the general hierarchical clustering procedure). Before the clustering procedure, we need to provide a significance level α_C to determine feasible combinations. Two adjacent blocks are considered a feasible combination if their $p_{LRR,BAF} \geq \alpha_C$. Our clustering procedure is described as below:

- At the beginning, each marker is a partitioned block.
- For each pair of adjacent blocks, calculate the corresponding $p_{LRR,BAF}$ and SSE (see below for the calculation of SSE).
- **while** there are still feasible combinations **do**
 - Among the feasible combinations, identify the one with the smallest SSE and update the list of blocks.
 - Combine the corresponding two blocks to form a new larger block.
 - Update the $p_{LRR,BAF}$ and SSE for the new block and its adjacent blocks. (There are two adjacent blocks if the new block is in the middle of the list, and only one if the new block is on the boundary of the list.)
- Return the final list of partitioned blocks.

Since LRR and BAF are two different types of information, it is difficult to define a rigorous optimization function (e.g., SSE) that uses LRR and BAF data directly. However, the allelic copy numbers are comparable data and they can be used to define a rigorous optimization function. The sum of squared errors (SSE) used in the above procedure is calculated based on the allelic copy numbers. For any SNP, using the LRR and BAF observations from either sample, we can calculate the corresponding allelic copy numbers A and B as described in Chen et al. (2011). Then, we have paired A_{ij} and B_{ij} for the SNP at location i and for sample j . The SSE for a block $S = \{k+1, k+2, \dots, k+m\}$ is simply calculated as:

$$SSE(k+1, k+2, \dots, k+m) = \sum_{j=1}^2 \sum_{i=k+1}^{k+m} [(A_{ij} - \bar{A}_j)^2 + (B_{ij} - \bar{B}_j)^2],$$

where $\bar{A}_j = \sum_{i=k+1}^{k+m} A_{ij}/m$ and $\bar{B}_j = \sum_{i=k+1}^{k+m} B_{ij}/m$. Notice that k is a reasonable integer value so that S can be a possible block on a given chromosome.

2.8. False discovery rate and α_C

In a genomics study, it is usually preferred to report discoveries with a reasonable control of false discovery rate (FDR), which evaluates the proportion of false positives among the claimed positives (Benjamini and Hochberg, 1995). Then, we can determine the significance level α_C (an important threshold value to identify feasible combinations in the above clustering procedure) based on a given FDR level (e.g., 5%). In this study, each detected change-point is a discovery. The number of claimed positives is simply the number of detected change-points based on the original data. The number of false positives can be estimated through a global permutation procedure. Each time, we re-assign the SNP locations randomly and re-run the hierarchical clustering procedure. After many repetitions (e.g., 100 global permutations), we calculate the average number of detected change-points based on the globally permuted data. The corresponding FDR can be calculate for any given α_C . Then, α_C can be determined when a satisfactory FDR (e.g., <10%) is observed.

2.9. Family-wise error rate and α_C

If any chromosome shows some strong change patterns, then we may detect the corresponding change-points with highly significant p -values. In such a situation, it is preferred to report the conservative family-wise error rate (FWER). The widely used Bonferroni correction for the FWER calculation is simply to divide a p -value by the number of tested hypotheses (Wright, 1992). Based on our clustering procedure, an upper bound for the number of tested hypotheses is $(L-1) + 2(L-3) + 1$, where L is the number of SNPs on the given chromosome and it is usually a number much larger than 1000. Then, a threshold value of $0.05/[(L-1) + 2(L-3) + 1]$ for the α_C will guarantee FWER <0.05.

3. RESULTS

3.1. Experimental data

We chose a published allele-specific CNV data set for our simulation and application studies. The data set was collected for studying cancer cells (Staaf et al., 2008), and it is freely available with GEO accession

number GSE11976. Since the volume of data is huge for each paired cell lines, the analysis results are usually lengthy even for the data for just one chromosome. Due to space limitation, we only selected the paired normal and tumor cell lines UC513_B and UC513_M1 (urothelial carcinoma) for our data analysis. The data used for our simulation studies are based on the experimental data from all chromosomes. However, for our application studies, we only focused on the analysis for chromosomes 20 and 21.

3.2. Simulation studies

As we explained above, we sample the experimental data randomly to generate our simulation data. For each simulation repetition (we repeat 1000 times), we generate a pair of two-sample LRR and BAF data sets. The simulation procedure for the simulated data with original order is described as follows:

1. Use the Poisson distribution with mean 30 to generate the sample size m (m is also the number of SNPs per block and we observe a range 10-55 in our simulations).
2. Sample a chromosome region with $2m$ SNPs randomly (those SNPs with a “NC” or “no call” genotype value from either normal or tumor cell line are excluded from the selection). The LRR and BAF data of these $2m$ SNPs are our simulated data. The data of the first m SNPs are assigned into sample-group one and the data of the rest m SNPs are assigned into sample-group two.
3. Generate 20000 random permutations based on the index set $\{1, 2, \dots, 2m\}$. These permuted indices are used to calculate permutation p -values.

Furthermore, we also generate the simulated data with randomized order by an additional random location permutation of these sampled $2m$ SNPs in Step2 of the above simulation procedure (before assigning the SNPs into two sample-groups).

3.2.1. p -value distributions. Although we used the widely used permutation procedure to calculate p -values (the approximation of extremely small p -values is also based on the permuted test scores), it is still necessary to check whether our permutation p -values are correct since the LRR/BAF data from adjacent SNPs can be considerably correlated. One approach that has been frequently used in practice is to test whether the p -value distribution under a null hypothesis is uniformly distributed. (Notice that, for this issue, it is not necessary to consider the approximation of extreme p -values.)

For the experimental allele-specific CNV data, based on our knowledge, we do not expect to frequently observe any clear changes of means for the BAF data or any clear changes of variances for the LRR data within a relatively short chromosomal region. Therefore, when we sample the experimental data randomly to generate the simulated data with original order, it is reasonable to assume no change of means as the null hypothesis for the BAF data and also to assume no change of variance as the null hypothesis for the LRR data. Then, a variance change test for the simulated LRR data or a mean change test for the simulated BAF data is expected to be uniformly distributed. This can be statistically tested by the widely used Kolmogorov-Smirnov (K-S) one-sample test (against the theoretical standard uniform distribution). The corresponding Kolmogorov-Smirnov test p -values given in Table 1 are all larger than the widely used

TABLE 1. EVALUATING UNIFORMITY AND CORRELATIONS FOR DIFFERENT TEST STATISTICS

Test statistic	K-S		Spearman
	LRR	BAF	
F -test (original data)	0.010	0.093	0.780
F -test (transformed data)	0.312	~ 0	0.791
Extended Levene’s test	0.757	~ 0	0.969
Extended Brown-Forsythe’s test	0.636	~ 0	0.624

The column of “Test statistic” lists different test statistics described in the text. The columns of “K-S” give the p -values based on the one-sample Kolmogorov-Smirnov test (against the theoretical standard uniform distribution). The column of “LRR” gives the p -values based on the permutation test p -values for the randomly sampled LRR data, and the column of “BAF” gives the p -values based on the permutation test p -values for the randomly sampled BAF data. The column of “Spearman” gives the p -values based on the Spearman’s rank correlation. The correlations are calculated based on two lists of permutation test p -values for the randomly sampled paired LRR and BAF data. All the p -values are two-sided. A p -value of ~ 0 means that it is extremely small ($< 10^{-15}$). All the p -values in the table are based on the simulated data with original order.

$\alpha = 5\%$ threshold level, which supports that the p -values can be correctly calculated through the permutation procedure.

3.2.2. Power of different test statistics. For each simulation repetition, both the LRR and BAF data are sampled from the same chromosomal locations. Furthermore, for each test statistic, the same permutations are used for the p -value calculation. (Notice that, for the power comparison, it is still not necessary to consider the approximation of extreme p -values.) Therefore, all the p -values from different test statistics on both LRR and BAF data are comparable. The power of these test statistics can be compared through their empirical cumulative distribution functions (ECDF). A test statistic is relatively more powerful if its ECDF curve is relatively higher. A diagonal line $y = x$ represents the corresponding null hypothesis. If any ECDF curves are close to the diagonal line, then it is not necessary to consider the related test statistics in an experimental data analysis.

For the simulated data with original order, Figure 1a, b shows these ECDF curves. For the LRR data, the ECDF curves for the variance change tests are all close to the diagonal line, which suggests that it is not necessary to consider any variance changes in an experimental data analysis (this has also been confirmed in Table 1). The ECDF curve for the mean change test is clearly above the diagonal line, which suggests that this type of test is necessary for the LRR data analysis. For the BAF data, the ECDF curve for the mean change test is close to the diagonal line, which suggests that it is not necessary to consider any mean changes in an experimental data analysis (this has also been confirmed in Table 1). All the ECDF curves for the variance change tests are clearly above the diagonal line. The F -test based on Equation (3) is clearly more powerful than the extended Levene's test and Brown-Forsythe's test. Therefore, we choose the F -test based on Equation (3) for the transformed BAF data analysis.

3.2.3. Correlation between p_{LRR} and p_{BAF} . According to our discussion above, since the simulated LRR and BAF data are paired and all the permutation p -values are comparable, it is feasible to evaluate the correlation between p_{LRR} and p_{BAF} for any mean/variance change test. (Again, for this issue, it is not necessary to consider the approximation of extreme p -values.) This can be statistically tested by the widely used Spearman's rank correlation test. For the simulated data with original order, the corresponding test p -values given in Table 1 are all larger than 50%, which clearly supports the assumption of zero correlation between p_{LRR} and p_{BAF} . Therefore, it is reasonable to use our proposed joint test to analyze LRR and BAF data together.

3.2.4. Join test p -value. To further confirm the validity of the theoretical p -value formula (Equation 4), using the simulated data with randomized order (so that both the simulated LRR and BAF data are consistent with the null hypothesis), we perform the one-sample Kolmogorov-Smirnov test to compare the empirical distribution of $p_{LRR} \times p_{BAF}$ to the corresponding theoretical distribution. The two-sided p -value is 0.355, which supports the validity of the theoretical p -value formula. (The two-sided p -value based on the Spearman's rank correlation between p_{LRR} and p_{BAF} is 0.239. When comparing the empirical distribution of either p_{LRR} or p_{BAF} to the standard uniform distribution, we obtain the two-sided one-sample Kolmogorov-Smirnov test p -value 0.387 or 0.301, respectively.)

3.2.5. Approximated extreme p -values. In our analysis of experimental data, we cannot afford a large number (e.g., 20000 in our simulation studies) of permutations for each test. Instead, we usually can only afford a relatively small number (e.g., 200 in our applications) of permutations and then we use our

FIG. 1. The empirical cumulative distribution function curves for different tests. **(a)** ECDF curves based on the LRR data. **(b)** ECDF curves based on the BAF data. In each plot, the gray dotted line is the diagonal line $y = x$.

FIG. 2. Comparison between the approximated permutation p -values and accurate permutation p -values. (a) LRR data based comparison. (b) Transformed BAF data based comparison. In each plot, the gray dots represent the permutation p -values without any approximation; the black dots represents the approximated p -values for the y -axis. The gray dotted line is the diagonal line $y = x$. The black dashed, dotted, and dot-dashed lines represent α -level at 0.001, 0.0005, and 0.0001, respectively.

proposed approximation procedure to calculate extremely small p -values. To evaluate whether our proposed procedure can give a satisfactory approximation, we use the simulated data with original order to compare the approximated p -values (based on 200 permutations) with the exact permutation p -values (based on 20000 permutations). Figure 2a,b shows an overall consistent pattern. These approximated p -values (<0.005) are represented by the black dots. We can use the McNemar's test to evaluate the consistency between these two paired lists of p -values. For both types of p -values, if their proportions of p -values less than a cutoff value are not significantly different, then we can conclude that two types of p -values are consistent. We choose three cutoff values: 0.001, 0.0005, and 0.0001. For the simulated LRR data, there are only a few p -values less than these three cutoff values and the McNemar's test gives 0.134, 0.617, and >0.999 significance, respectively. For the simulated BAF data, there are many p -values less than these three cutoff values and the McNemar's test gives >0.999 , >0.999 , and 0.773 significance, respectively. The Spearman's rank correlation coefficients are 0.511 (p -value = 0.043) and 0.718 (p -value ~ 0) for the paired approximated (<0.005) versus exact permutation LRR and BAF p -values, respectively. These test results suggest that a satisfactory approximation of extremely small p -values can be achieved by our proposed approximation procedure.

3.3. Applications

It is not computationally feasible to calculate F -test p -values based on a large number of permutations. Therefore, we need to consider both permuted and approximated p -values in our analysis of experimental data. Before the approximation of extremely small p -values, we consider the following number of permutations for each F -test: if the number of all possible permutations is not larger than 300, then we will use all these permutations for the p -value calculation; otherwise, we will randomly generate 200 permutations for the p -value calculation. For a given partition of block, if the number of possible permutations is less than 100, then we will simply report one for the p -value since it is not necessary to consider such a significance level for the multiple comparison adjustment.

To illustrate our proposed method for analyzing paired allele-specific CNV data, we report the analysis results for the chromosomes 20 and 21 based on the paired normal and tumor cell lines UC513_B and UC513_M1 (urothelial carcinoma). We first use $\text{FWER} < 0.05$ for the detection of highly significant change-points. Table 2 lists the partitioned blocks for these two chromosomes with the corresponding statistics. The related change-points are described in Figures 3 and 4 (black solid vertical lines). To explore more discoveries, we also use $\text{FDR} < 0.1$ to detect additional change-points for potential copy number variation regions. The results are also given in Figures 3 and 4 (gray dashed vertical lines).

Both the FWER and FDR based α_C 's are selected for the joint test p -values, which are calculated according to Equation (4). Conversely, we can use a numerical procedure like the R-function uniroot to find the $p_{LRR} \times p_{BAF}$ threshold value $\tilde{\alpha}_C$ that corresponds to a required α_C . If any p_{LRR} or p_{BAF} is below that threshold value, then the corresponding change-point can be detected by just one type of data (either LRR or BAF data). In Table 2, we list these p -values (p_{LRR} or p_{BAF}) in bold font. Otherwise, a change-point must be detected due to some relatively moderate changes in both LRR and BAF data.

TABLE 2. CHANGE-POINT TEST AND ESTIMATION RESULTS FOR THE CHROMOSOMES 20 AND 21 BASED ON A PAIRED ALLELE-SPECIFIC COPY NUMBER VARIATION EXPERIMENTAL DATA SET

Block index	Estimated mean					
	LRR		Transformed BAF		p-value (F-test)	
	Normal	Tumor	Normal	Tumor	LRR	BAF
Chromosome 20						
1	0.009	0.172	0.328	0.348		
2	0.008	0.282	0.431	0.439	6.60E-04	1.57E-06
3	0.076	0.342	0.260	0.287	3.00E-02	1.15E-11
4	0.035	0.205	0.334	0.343	1.40E-05	5.00E-03
5	-0.053	0.092	0.495	0.494	5.00E-03	1.15E-05
6	0.090	0.295	0.194	0.214	1.72E-04	5.63E-15
7	0.033	0.169	0.325	0.338	9.83E-04	5.37E-05
8	-0.023	0.124	0.476	0.476	4.23E-03	6.23E-21
9	-0.014	0.114	0.338	0.345	6.10E-01	2.52E-22
10	0.012	0.844	0.300	0.461	2.32E-64	4.66E-05
11	-0.007	0.044	0.335	0.333	1.66E-82	3.76E-04
12	-0.031	0.047	0.407	0.405	2.05E-01	3.54E-07
13	0.040	0.045	0.336	0.339	2.80E-04	3.20E-07
Chromosome 21						
1	0.038	0.304	0.353	0.352		
2	0.113	0.417	0.337	0.339	5.13E-12	2.25E-01
3	0.045	0.243	0.320	0.334	4.49E-22	3.05E-01
4	0.055	0.277	0.497	0.495	5.50E-01	1.72E-07
5	0.054	0.283	0.305	0.321	9.80E-01	1.67E-08
6	-0.022	0.154	0.497	0.495	3.25E-04	1.34E-15
7	0.011	0.162	0.339	0.353	3.45E-01	3.20E-13
8	0.016	0.055	0.440	0.457	3.63E-17	6.80E-48

FWER < 0.05 is the criterion for change-point detection. The indices for the partitioned blocks are listed in the first column. For the partitioned blocks, columns 2–5 give the estimated means for the LRR and transformed BAF data of the normal and tumor cell lines. For the LRR or the transformed BAF data, the last two columns give the change-point test p -values between the current block and its previous block.

The detected change-points based on FWER < 0.05 and FDR < 0.1 are both presented in Figures 3 and 4. The results include those clearly visible changes as well as many less visible changes. It is interesting to observe that many detected changes are mostly from BAF data. Due to space limitation, we only discuss the details for these detections with FWER < 0.05. The results for chromosomes 20 and 21 are separately described below. Notice that we apply our method to the transformed BAF data. For the estimated means for the transformed BAF data, a value close to 0 means that the corresponding chromosomal region is highly heterozygous; a value close to 0.5 means that the corresponding chromosomal region is highly homozygous, which can be a sign of loss of heterozygosity (LOH). For the LRR data, the estimated means are directly related to the chromosomal copy numbers.

3.3.1. Chromosome 20. For chromosome 20, there are 8136 SNPs for our analysis (SNPs with a “NC,” or “no call,” genotype value from either normal or tumor cell line are excluded). The threshold α_C values are 2.05E-6 and 9.84E-4 for FWER < 0.05 and FDR < 0.1, respectively. For FWER < 0.05, we detect 12 change-points (or 13 partitioned regions) as presented in Table 2 and Figure 3. The threshold $\tilde{\alpha}_C$ value is 1.21E-7. In Table 2, there are 2 LRR p -values and 4 BAF p -values below this threshold value (6 change-points can be detected by just one type of data). The rest change-points are detected based on the simultaneous changes of LRR and BAF data. For the partitioned blocks of SNPs, the estimated means for normal LRR data are mostly close to zero, while the estimated means for tumor LRR data show an alternate increasing-decreasing pattern. The estimated means for both normal and tumor BAF data (transformed) also show an alternate increasing-decreasing pattern. For either Block 3 or Block 6, the estimated means for both normal and tumor BAF data (transformed) are clearly overall low (but still clearly larger than 0 for a

FIG. 3. Partition of chromosome 20 based on our proposed change-point analysis. **(a)** Normal LRR data, gray dotted horizontal lines represent the values $(-1, 0, 1)$. **(b)** Tumor LRR data, gray dotted horizontal lines represent the values $(-1, 0, 1)$. **(c)** Normal BAF data, gray dotted horizontal lines represent the values $(0, 0.5, 1)$. **(d)** Tumor BAF data, gray dotted horizontal lines represent the values $(0, 0.5, 1)$. In each plot, gray dots represent the experimental data for SNPs. The change-points detected based on $\text{FWER} < 0.05$ are shown as black solid vertical lines; the corresponding block numbers are also given on the bottom of figure. The change-points detected based on $\text{FDR} < 0.1$ are shown as gray dashed vertical lines. The common x -axis gives the physical position of SNPs.

sign of completely heterozygous), and the estimated means for the previous block (2 or 5) are clearly high and close to 0.5 (a sign of LOH). Interestingly, the estimated means for normal LRR data in both Blocks 3 and 6 are visibly increased (although still relatively low), and this can also be observed for the corresponding estimated means for the tumor LRR data. Another clear sign of LOH for both normal and tumor samples is from Block 8, in which the LRR changes are quite moderate. Block 10 is highly interesting: there is a significant tumor LRR increase (while the normal LRR data remain stable and close to 0), and there is a significant tumor BAF (transformed) increase (while the normal BAF data remain stable). These observations suggest a strong chromosomal amplification only on a single strand.

3.3.2. Chromosome 21. For chromosome 21, there are 5113 SNPs for our analysis after removing SNPs with a “NC,” or “no call,” genotype value from either normal or tumor cell line. The threshold α_C values are $3.26\text{E-}6$ and $5.45\text{E-}4$ for $\text{FWER} < 0.05$ and $\text{FDR} < 0.1$, respectively. For $\text{FWER} < 0.05$,

FIG. 4. Partition of chromosome 21 based on our proposed change-point analysis. **(a)** Normal LRR data, gray dotted horizontal lines represent the values $(-1, 0, 1)$. **(b)** Tumor LRR data, gray dotted horizontal lines represent the values $(-1, 0, 1)$. **(c)** Normal BAF data, gray dotted horizontal lines represent the values $(0, 0.5, 1)$. **(d)** Tumor BAF data, gray dotted horizontal lines represent the values $(0, 0.5, 1)$. In each plot, gray dots represent the experimental data for SNPs. The change-points detected based on $\text{FWER} < 0.05$ are shown as black solid vertical lines, the corresponding block numbers are also given on the bottom of figure. The change-points detected based on $\text{FDR} < 0.1$ are shown as gray dashed vertical lines. The common x -axis gives the physical position of SNPs.

we detect 7 change-points (or 8 partitioned regions) as presented in Table 2 and Figure 4. The threshold $\tilde{\alpha}_C$ value is $1.98\text{E-}7$. In Table 2, it is clear that all the change-points can be detected by just one type of data. It is interesting to observe that both the LRR and BAF changes are less than the threshold value for the last change-points. For the related last block (Block 8), the estimated means show a significant tumor LRR decrease (while the normal LRR remain stable and close to 0) and a significant increase for both normal and tumor transformed BAF data (the LOH trend is clearer for tumor BAF data). There are two blocks (4 and 6) with a significantly sign of LOH. Both the normal and tumor LRR patterns remain stable for Block 4 but there is a moderate LRR decrease for Block 6. For Block 2, the estimated means for both normal and tumor LRR data are significantly increased. Both Figure 4 and Table 2 show that the tumor sample is more homozygous than the normal sample in Blocks 3, 5, 7, and 8.

4. DISCUSSION

In this study, we proposed a statistical framework to analyze paired allele-specific copy number variation (CNV) data. In order to detect change-points simultaneously for four sequences of data (paired LRR and paired BAF data), we proposed to use the traditional F -type test for our analysis. Since the dependence between the data from a paired samples and also the dependence among the data from adjacent SNPs were not negligible, we proposed to calculate the test p -values based on the widely used permutation procedure. We also confirmed that the p -values can be correctly calculated through the permutation procedure. Since the test on the transformed BAF data can also be considered as a test for the homogeneity of variance, we conducted a simulation based comparison of different test statistics to confirm that the proposed F -test is a more powerful choice for our analysis. We also confirmed that it is reasonable to assume an independence between the LRR data based F -test and the BAF data based F -test, and the theoretical formula is valid for calculating the joint test p -value. Furthermore, we confirmed that our proposed extreme p -value approximation is appropriate. To maintain the control of false positives at a satisfactory level, we proposed to use a clustering procedure for the partition of SNPs. We also discussed the issues of FDR and FWER control. For the applications of our proposed approach, we considered a published data set. The allele-specific CNV data for chromosomes 20 and 21 from a paired normal/tumor cell lines were analyzed and interesting chromosomal aberration regions were detected by our proposed approach.

One clear advantage of our proposed statistical framework is that changes from either normal or tumor samples (or both) can be detected (since the paired data are analyzed simultaneously). For example, in our analysis results for chromosome 21 (Table 2), we can simply calculate the differences between the paired LRR estimates. The difference is 0.266 for Block 1 and 0.304 for Block 2, and these two differences are relatively close. Notice that the related change-point has been identified just based on the LRR difference. This change-point would be much less significant if we would only analyze the difference (or ratio) data between the paired samples. Therefore, with an appropriate p -value calculation method, we can detect more change-points by a simultaneous analysis of paired CNV data. Another clear advantage of our proposed statistical framework is that the detection power can be clearly improved by the joint analysis of LRR and BAF data. For example, in our analysis results for chromosome 20 (Table 2), there are many change-points detected based on the simultaneous changes of LRR and BAF data.

A follow-up analysis after the partition of SNPs (detection of change-points) is to estimate the underlying DNA copy numbers and also the related genotypes for each partitioned chromosomal region. For a given chromosomal region, this analysis can be achieved through a well-developed statistical framework. For example, the observed BAF data represent a mixture of different types of homozygous and heterozygous genotypes. Similarly, the observed LRR data also represent a mixture of different DNA copy numbers since either a normal or a tumor sample may contain different types of cells with different DNA copy numbers. One frequently used approach is to implement the E-M algorithm (McLachlan and Krishnan, 1997) for a specific mixture model. A related technical difficulty is how to model the dependence among adjacent SNPs. Another related technical challenge is how to borrow information from different partitioned blocks so that an efficient estimation can be achieved. Although some of these issues have been addressed in the current literature, it is still necessary to pursue a comprehensive research so that a satisfactory solution can be proposed. Furthermore, it would be ideal if this follow-up analysis can be integrated with the change-point analysis (partition of chromosomal regions) so that a more efficient solution could be achieved.

5. APPENDIX

To address the validity of permutation p -values (since the data correlation structure can be complex), we have assumed no change of means as the null hypothesis for the BAF data and also no change of variance as the null hypothesis for the LRR data within a relatively short chromosomal region. In our simulation studies, we have set 60 ($m = 30$ per sample group) as the average sample size. Furthermore, it is necessary to understand whether this assumption holds for a longer chromosomal region. Therefore, we repeat our simulation analysis for additional values of $m = 15, 45, 60, 90, 180$, and 360. The analysis results from this comprehensive simulation study are summarized in Figure 5.

Figure 5a shows that the permutation p -values for the variance test of LRR data can remain uniformly distributed when $m = 15, 30$, and 45. When $m \geq 90$, all the Kolmogorov-Smirnov test p -values are clearly

FIG. 5. Summary p -values for different simulation sample size. **(a)** The one-sample Kolmogorov-Smirnov test p -values (against the theoretical standard uniform distribution) based on the LRR data. **(b)** The one-sample Kolmogorov-Smirnov test p -values (against the theoretical standard uniform distribution) based on the BAF data. **(c)** The Spearman's rank correlation test p -values. The correlations are calculated based on two lists of permutation test p -values (based on the same type of test) for the randomly sampled paired LRR and BAF data. All the p -values are two-sided, and they are based on the simulated data with original order. In each plot, the gray solid line represents p -value = 0.05. A point with 10^{-6} value means that the related p -value is less than or equal to 10^{-6} . The black solid, dashed, dotted, and dash-dotted curves represent the F -test with no data transformation, the F -test with a data transformation, the extended Levene's test, and the extended Brown-Forsythe's test, respectively.

significant. Based on our knowledge, it is common to observe a clear variance change when there is a highly significant mean change. For longer chromosomal regions, it is likely to observe significant mean changes. Then, it is also likely to observe variance changes. This can make the related permutation p -value distribution non-uniform. Figure 5b shows that the permutation p -values for the mean test of BAF data can remain uniformly distributed for a relatively long chromosomal region. Both Figures 5a and 5b suggest that the mean changes for LRR data and the variance changes for BAF data exist even within a relatively short chromosomal region. Furthermore, Figure 5c shows that the correlation between p_{LRR} and p_{BAF} (each pair are based on the same type of test) can remain insignificant for a relatively long chromosomal region. The correlation between p_{LRR} for the F -test with no data transformation (mean test) and p_{BAF} for the F -test with a data transformation (variance test) can also remain overall insignificant. For $m = 15, 30, 45, 60, 90, 180$, and 360 , we observe Spearman's rank correlation test p -values $0.066, 0.183, 0.037, 0.296, 0.049, 0.166$, and 0.223 , respectively. All the above observations further support that the p -values can be correctly calculated through the permutation procedure and it is reasonable to use our proposed joint test to analyze LRR and BAF data together.

ACKNOWLEDGMENTS

I am grateful to Professor Fengzhu Sun for his helpful comments and suggestions. This work was supported by the National Institutes of Health (grant GM092963).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Assié, G., LaFramboise, T., Platzer, P., et al. 2008. SNP arrays in heterogeneous tissue: highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples. *Am. J. Hum. Genet.* 82, 903–915.
- Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.

- Brown, M.B., and Forsythe, A.B. 1974. Robust tests for equality of variances, *J. Am. Stat. Assoc.* 69, 364–367.
- Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068.
- Casella, G., and Berger, R.L. 2002. *Statistical Inference*, 2nd ed. Duxbury, Pacific Grove, CA.
- Chen, H., Xing, H., and Zhang, N.R. 2011. Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. *PLoS Comput. Biol.* 7, e1001060.
- Dellinger, A.E., Saw, S.M., Goh, L.K., et al. 2010. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res.* 38, e105.
- Dudoit, S., Shaffer, J.P., and Boldrick, J.C. 2003. Multiple hypothesis testing in microarray experiments. *Stat. Sci.* 18, 71–103.
- Gastwirth, J.L., Gel, Y., and W. Miao 2009. The impact of Levene's test of equality of variances on statistical theory and practice. *Stat. Sci.* 24, 343–360.
- Good, P. 2005 *Permutation, Parametric, Bootstrap Tests of Hypotheses*. Springer, New York.
- Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York.
- Lai, Y. 2011. On the adaptive partition approach to the detection of multiple change-points. *PLoS ONE* 6, e19754.
- Levene, H. 1960. Robust tests for equality of variances, 278–292. In: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, Palt Alto, CA.
- Li, W., Lee, A., and Gregersen, P.K. 2009. Copy-number-variation and copy-number-alteration region detection by cumulative plots. *BMC Bioinform.* 10, S67.
- Liu, Z., Li, A., Schulz, V., et al. 2010. MixHMM: inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. *PLoS ONE*, 5, e10909.
- McLachlan, G.J., and Krishnan, T. 1997. *The EM Algorithm and Extensions*. John Wiley & Sons, Inc., New York.
- Olshen, A.B., Venkatraman, E.S., Lucito, R., et al. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557–572.
- Olshen, A.B., Bengtsson, H., Neuvial, P., et al. 2011. Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics* 27, 2038–2046.
- Staaf, J., Lindgren, D., Vallon-Christersson, J., et al. 2008. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol.* 9, R136.
- Van Loo, P., Nordgard, S.H., Lingjærde, O.C., et al. 2010. Allele-specific copy number analysis of tumors. *Proc. Nat. Acad. Sci. USA* 107, 16910–16915.
- Venkatraman, E.S., and Olshen, A.B. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23, 657–663.
- Wang, K., Li, M., Hadley, D., et al. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17, 1665–1674.
- Wang, P., Kim, Y., Pollack, J., et al. 2005. A method for calling gains and losses in array CGH data. *Biostatistics* 6, 45–58.
- Wright, S. P. 1992 Adjusted p-values for simultaneous inference. *Biometrics* 48, 1005–1013.
- Zhao, X., Li, C., Paez, J.G., et al. 2004. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* 64, 3060–3071.

Address correspondence to:

Dr. Yinglei Lai
Department of Statistics and Biostatistics Center
The George Washington University
801 22nd Street N.W.
Washington, DC 20052

E-mail: ylai@gwu.edu