Statistics and Its Interface Volume 3 (2010) 211–221

Nonparametric tests for longitudinal DNA copy number data

Ke Zhang[†] and Haiyan Wang[*]

Array comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) array data are becoming commonly available for scientists to study genetic mechanisms involved in complex biological processes. Such data typically contain a large number of probes observed repeatedly over time. Due to cost concerns, the number of replicates is often very limited. Effective hypothesis testing tools need to take into account the high dimensionality and small sample sizes. In this paper, we present a set of nonparametric hypothesis testing theory to test for main and interaction effects related to a large number of probes for longitudinal DNA copy number data from aCGH or SNP arrays. The asymptotic distributions of the test statistics are obtained under a realistic model setup that allows distribution-free robust inference in presence of temporal correlations for heteroscedastic high dimensional low sample size data. They provide a flexible tool for a wide range of scientists to accelerate novel gene discovery such as identification of genome regions of aberration to control tumor progression. Simulations and applications of the new methods to DNA copy number aberration from Wilm's tumor relapse study are presented.

AMS 2000 subject classifications: Primary 62P10, Secondary 62G10, 62G35.

Keywords and phrases: Repeated measures, Nonparametric statistics, Hypothesis testing, DNA copy number aberration, High dimensional data analysis.

1. INTRODUCTION

Tumor cells usually undergo dramatic chromosome changes resulting in gain or loss of DNA copy numbers. For normal tissues, most human DNA segment has two copies with the exception of the sex chromosome. However, the DNA of tumor cells is often subject to translocation, amplification, and deletion, which leads to DNA copy number abnormality. High throughput array comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) array technologies have made it possible to simultaneously examine DNA copy numbers at thousands or millions of sites of a genome. Fluorescence-labeled DNA probes are designed to bind to specific chromosomal locations that are distributed throughout a human genome with high resolution. Genomic DNA fragments from tumor and normal reference samples are hybridized with the probes, and the log ratios of fluorescent intensities between tumor and normal samples are calculated. If both samples contain the same quantity of DNA copies, the expected log ratio is equal to 0. In contrast, when the tumor sample has a gain or loss of DNA copy number, the expected log ratio is greater or smaller than 0, respectively. Therefore, the log ratio data can be used as a raw copy number, and chromosome aberration can be detected by analyzing the raw copy number. Although copy number data provide rich information, they also raise challenges for statistical and computational methods.

We are often interested in the DNA copy number aberrations for chromosome segments. For example, each chromosome has two arms, p and q, that are connected by a centromere. Chromosome rearrangement often causes one arm to be translocated, duplicated or lost. The copy number changes of a chromosome arm will affect thousands of probes located in it. In recent years, many statistical methods have been used to identify the regions of aberration when the signal intensities on different chips are from independent samples. These include hidden Markov models (Fridlyand et al. [2004]), circular binary segmentation (CBS) based on change point analysis (Olshen et al. [2004]), Bayesian models (Daruwala et al. [2004]), and regression (Tibshirani and Wang [2008]). Most of these methods assume a specific distribution for the (log ratio of) intensities of the probes such as normal, log-normal or Poisson distribution in addition to the fact that they are only designed to work with independent samples.

To gain a better understanding of tumor development and progression, some recent cancer studies have been carried out to investigate the dynamic changes of genomic DNA by monitoring the experiments longitudinally. Lai et al. [2007] found increasing genomic instability during premalignant neoplastic progression in an aCGH study. The DNA samples were collected in three distinct stages of molecular evolution for each patient. Such a longitudinal study was also conducted to leukemia or lymphoma patients, whose tumor cells can be relatively easily collected. For instance,

[*] Corresponding author.

[†] Research was partially supported by NIH grant 2P20RR016471-09, North Dakota INBRE.

Mullighan et al. [2008] investigated the relapse of acute lymphoblastic leukemia (ALL) by measuring DNA copy number from patient blood or bone marrow samples at both diagnosis and relapse. Another application of longitudinal copy number study is tumor cell line research. It is known that after dozens of generations, the genome of a tumor cell line will be substantially different from the original one due to accumulation of DNA changes. For instance, the scientists in Abbott Laboratories have collected DNA samples from 5 continuous splits of small cell lung carcinoma (SCLC) cell lines, in order to identify the maximum number of splits that preserves characteristics of the original SCLC tumor (unpublished data). In short, longitudinal DNA copy number research has drawn increasing attention in cancer research, and it will continue to provide rich information of genomic variation during disease development.

Appropriately incorporating the correlation due to the repeated measurements from the same subject can increase the power of statistical analysis. Current literature on longitudinal array data analysis are mainly focused on a univariate test for an individual probe based on expression profile, followed by false discovery rate (FDR) adjustment for multiple comparisons. To detect the abnormal genomic region, it is necessary to conduct a statistical test of equal copy number for all the probes within the region. Due to the large number of parameters involved in the model, classical methods such as linear mixed effects models and generalized estimating equations have limited applications in DNA copy number data analysis. Tsai and Qu [2008] performed hypothesis testing for a class of genes by applying a quadratic inference function (QIF) to account for within-gene correlation. Similar to GEE, QIF requires the number of estimating equations greater than the number of parameters.

When there are only two observations per probe, a paired test is often considered by an applied scientist for each probe. We remark that a paired-observation model is only a special case of a longitudinal model in that the paired observations can be treated as two repeated observations from the same probe. A paired test is more powerful than a corresponding test that assumes the two dependent sets of observations are independent. However, the paired test is not more powerful than a test from a longitudinal model with appropriate covariance structure even in the traditional sense. In fact, the paired test either requires the difference of the paired observations to follow Gaussian distribution, or requires large sample sizes. Neither of these two conditions are reasonable for the setting of this manuscript because real copy numbers are typically integers plus multiple sources of noises.

In this manuscript, we propose a set of nonparametric statistical tests to assess if a DNA region has copy number aberration using longitudinal aCGH or SNP array copy number data. Due to the small sample sizes, the tests proposed in this manuscript borrow strength from considering copy number observations from many probes without assuming that they have a common distribution. The proposed methods have a number of advantages. First, no distributional assumptions are required for raw copy numbers. Secondly, the method works effectively for a large number of probes with low replications. Thirdly, heteroscedastic within and between-probe correlations are taken into account. Fourthly, unbalanced designs with general model set-up are used to allow for flexible and realistic modeling. Lastly, the computation is fast comparing to probe-by-probe test method because in our model all probes in a genomic region are tested simultaneously by fitting into one single test statistic.

The outline of this manuscript is as follows. In section 2, we present the study design and the model specification. The test statistics and their asymptotic distributions are provided in section 3. Simulation studies are presented in section 4. In section 5, we use our methods to investigate the genetic basis of Wilms' tumor relapse.

2. MODEL SPECIFICATION AND DEPENDENCE AMONG PROBES

Before we can develop inference, it is necessary to consider the dependence among the probes. Two probes of two very close genomic locations may be highly correlated. On the other hand, two probes on two different haplotype blocks that are far away in genomic locations may be less or not correlated. That is, block-wise independence may be a natural assumption for copy numbers from different probes of the same subject. However, such independence structure requires manual partition of the genome into blocks. As we would like our tests to be applicable to any DNA segments of reasonable size with least user intervention, we choose a more general and flexible dependence assumption. Specifically, we first align all probes from the same subject along their relative genomic locations and let X_{ijk} be the jth measurement of the raw copy number of the ith probe from subject k, $i = 1, \ldots, I$; $j = 1, \ldots, J$; $k = 1, \ldots, n_i$. The number of probes is large, whereas the number of time points and the number of subjects are bounded. The design is assumed to be either balanced or unbalanced, in that the number of replications may vary for different probes. If all copy number data come from the same version of chips, the design is balanced; otherwise, the design could be unbalanced. For instance, properly normalized Affymetrix SNP arrays have been used in DNA copy number research recently (cf. Rigaill et al. [2008]; Redon et al. [2006]; Barnes et al. [2008]; Winchester et al. [2009] and the references therein). If the data have both Affymetrix 100K and 250K arrays, the probes shared in both 100K and 250K arrays have more replications than the probes that only exist in 250K arrays. Though normal or lognormal distributions are often used for fluorescent intensities, it has been vigorously argued as to whether it is adequate to fit real image data with a well-defined distribution (Kerr et al. [2000]; Konishi [2004]).

For all probes from the same subject at a specific time point, we assume the copy numbers $\{X_{ijk}, i = 1, \ldots\}$ satisfy an α-mixing condition with mixing coefficient

$$\alpha_m = \sup_{i, A\in\sigma(X_{i_1jk}, i_1\le i), B\in\sigma(X_{i_2jk}, i_2\ge i+m)} |P(A \bigcap B) - P(A)P(B)| = O(m^{-5}),$$

where $\sigma(X_{i_1jk}, i_1 \le i)$ is the σ-field generated by the variables $\{X_{i_1jk}, i_1 \le i\}$. This α-mixing condition was first introduced by Rosenblatt [1956]. It implies that the magnitude of correlation between two observations for two probes that are far apart in their genomic locations decreases in general as their genomic distance increases without assuming a specific parametric form for the correlation matrix (cf. Billingsley [1995]).

Denote $\mathbf{X}_k$ to be the $I \times J$ matrix of copy numbers from the kth subject with entry for the ith row and jth column being X_{ijk}. Each column of $\mathbf{X}_k$ is the vector of copy numbers for all probes observed at a single time point. Each row of $\mathbf{X}_k$ contains the repeatedly measured copy numbers at all time points for one probe. Denote the true mean copy number for the i^{th} probe at the j^{th} time point by $\mu_{ij} = E(X_{ijk})$. Note that we are interested in finding the probes that have copy number aberrations for a whole group of diseased patients in cancer studies so that the genes encoded on these regions where these probes are located can serve as therapeutic targets for treatment and drug development. Therefore, subject specific copy numbers are not of interest. The mean copy numbers μ_{ij} can be decomposed to yield the i^{th} probe effect α_i, j^{th} time effect β_j, and probe by time interaction effect γ_{ij}:

$$\mu = I^{-1}J^{-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}, \qquad \alpha_i = J^{-1}\sum_{j=1}^{J}\mu_{ij} - \mu,$$

$$\beta_j = I^{-1}\sum_{i=1}^{I}\mu_{ij} - \mu, \qquad \gamma_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j.$$

The decomposition above implies the following natural constraints for these effects: $\sum_{i=1}^{I}\alpha_i = \sum_{j=1}^{J}\beta_j = \sum_{i=1}^{I}\gamma_{ij} = \sum_{j=1}^{J}\gamma_{ij} = 0$. Even though the effects can be defined through univariate notation, the error term $\varepsilon_{ijk} = X_{ijk} - \mu_{ij}$ contains multiple sources of variations including the random subject effect, random subject by time interaction effect, and measurement error. Let $\boldsymbol{\epsilon}_k$ be the $I \times J$ error matrix with the (i, j) element ε_{ijk}. Then the dependence among probes and among repeated measurements renders all elements in $\boldsymbol{\epsilon}_k$ being correlated. Putting all terms together leads to the model for all copy numbers from the same subject:

$$\mathrm{vec}(\mathbf{X}_k') = \mu \cdot \mathbf{1}_{IJ} + \boldsymbol{\alpha} \bigotimes \mathbf{1}_J + \mathbf{1}_I \bigotimes \boldsymbol{\beta} + \mathrm{vec}(\boldsymbol{\gamma}') + \mathrm{vec}(\boldsymbol{\epsilon}_k'),$$

where vec is the vector-operator of a matrix, $\mathbf{1}_I$ is an I-dimensional vector of ones, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$, $\boldsymbol{\gamma}$ is an $I \times J$ matrix with (i, j) element γ_{ij}, and $\bigotimes$ is the kronecker product. The unknown covariance matrix of $\mathrm{vec}(\boldsymbol{\epsilon}_k')$ can be written as

$$\Sigma_{IJ\times IJ} = \begin{pmatrix} \Omega_1 & V_{12} & \cdots & V_{1I} \\ V_{21} & \Omega_2 & \cdots & V_{2I} \\ \vdots & \vdots & \vdots & \vdots \\ V_{I1} & V_{i2} & \cdots & \Omega_I \end{pmatrix},$$

where Ω_i, $i = 1, \ldots, I$, are $J \times J$ symmetric matrix with (j, j') element $\sigma_{iijj'} = \mathrm{Cov}(X_{ijk}, X_{ij'k})$; V_{mn}, $m, n = 1, \ldots, I$, are $J \times J$ symmetric matrix with (j, j') element $\sigma_{mnjj'} = \mathrm{Cov}(X_{mjk}, X_{nj'k})$.

Assume the copy numbers from different subjects are independent. The model in this paper is appropriate for both continuous and discrete data. Thus it can model both the raw copy number data or integer copy number derived from data preprocessing. The unknown covariance matrix Σ is completely unstructured except for the α mixing condition. The sub-covariance matrix Ω_i for probe i can be different for different probes. This is more reasonable since the covariance of the copy numbers for a probe at different time points is possibly dependent on their locations on a chromosome. Furthermore, experiments of biological time course study are often not evenly spaced in time. Therefore, the same correlation structure may not be appropriate. Note that the subject random effect is included in our formulation though it is not specifically written in the model.

For a DNA segment, we first consider the probe by time interaction effect, $H_0 : \gamma_{ij} = 0$, for all i, j. This evaluates if all probe profiles are parallel. If the segment contains some probes whose copy numbers significantly vary over time, then tumor progression may be associated to those probes. If there is no significant interaction effect detected, we consider testing the null hypothesis of no copy number aberration for a whole group of probes, $H_0 : \alpha_i = 0$, for all i. The test can be applied to detect the local DNA copy number changes in a given genome region so that DNA segmentation can be done to partition the whole genome into amplified, deleted, and normal regions.

We remark that in both hypotheses, the parameters under the null lie in a low dimensional space but those under the alternatives lie in a high dimensional space. Meanwhile, the number of nuisance parameters (the dimension of the covariance matrix Σ) go to infinity as the number of probes increases. Due to this reason, the maximum likelihood estimators for the parameters may not be consistent as the total number of parameters approaches infinity while the number of replications stays fixed, and even when the estimators are consistent it may still fail to be efficient (cf. Neymann and Scott [1948]; Fan and Lin [1998]; Li et al. [2003] and the references therein). Most of the nonparametric approaches is unapplicable since they require large n_i or J. New methods need to be developed for an effective inference in the current setting.

The following notation will be used throughout the manuscript: $\overline{X}_{ij\cdot} = n_i^{-1}\sum_{k=1}^{n_i} X_{ijk}$, $\widetilde{X}_{i\cdot\cdot} = J^{-1}\sum_{j=1}^{J}\overline{X}_{ij\cdot}$, $\widetilde{X}_{\cdot j\cdot} = I^{-1}\sum_{i=1}^{I}\overline{X}_{ij\cdot}$, $\widetilde{X}_{\dots} = I^{-1}\sum_{i=1}^{I}\widetilde{X}_{i\cdot\cdot}$. Denote $n(i,i')$ to be the number of subjects such that the observed copy numbers for both probes i and i' from each subject are available.

3. TEST STATISTICS AND THEIR ASYMPTOTIC DISTRIBUTIONS

First, we consider the probe and time interaction effects. Significant interactions indicate that the copy numbers of a group of probes on a DNA segment change differently over time. The null hypothesis is

$$H_0(AB) : \text{all } \gamma_{ij} = 0, \text{ for } i = 1, \dots, I, \text{ and } j = 1, \dots, J.$$

As explained in previous section, this hypothesis tests against high dimensional alternatives. We use the difference of two quadratic forms, $T_{AB} - E_{AB}$, as the test statistic so that the two quadratic forms have matching expectation under the null hypothesis with the heteroscedastic model, where

$$T_{AB} = \sum_{i,j} \frac{(\overline{X}_{ij\cdot} - \widetilde{X}_{i\cdot\cdot} - \widetilde{X}_{\cdot j\cdot} + \widetilde{X}_{\dots})^2}{(I-1)(J-1)},$$

$$E_{AB} = \sum_{i=1}^{I}\sum_{k=1}^{n_i} \frac{\left[\sum_j^J (X_{ijk} - \overline{X}_{ij.})^2 - J(\overline{X}_{i\cdot k} - \overline{X}_{i..})^2\right]}{I(J-1)n_i(n_i-1)}.$$

The asymptotic distribution of the test statistic is given in the next theorem for fixed sample sizes and a large number of probes.

Theorem 3.1. *Assume X_{ijk} have finite 16th central moment. Further, assume for all j, k, $\{X_{ijk}, i = 1, \dots\}$ is an α-mixing sequence with mixing coefficient $\alpha_m = O(m^{-5})$. Then under $H_0(AB)$,*

$$\frac{\sqrt{I}(T_{AB} - E_{AB})}{\tau_{AB}} \xrightarrow{d} N(0,1) \text{ as } I \to \infty,$$

where

$$\tau_{AB} = \frac{2}{I(J-1)^2}\sum_{i,i'}^{I} \frac{n(i,i')[n(i,i')-1]}{n_i^2(n_i-1)^2} \times \left[\sum_{j,j_1}^{J}\sigma_{ii'jj_1}^2 + \frac{1}{J^2}\left(\sum_{j,j_1}^{J}\sigma_{ii'jj_1}\right)^2 - \frac{2}{J}\sum_{j,j_1,j_2}^{J}\sigma_{ii'jj_1}\sigma_{ii'jj_2}\right],$$

provided that τ_{AB} is bounded away from 0 as $I \to \infty$.

To apply the test statistics to real data, we need estimates of the asymptotic variances. The following proposition gives a consistent estimate for τ_{AB}.

Proposition 3.2. *For all i, i', assume $n(i,i') \geq 4$. Let*

$$\widehat{\sigma}_{ii'}(j, j', j_1, j_1') = \sum_{k_1 \neq k_2 \neq k_3 \neq k_4}^{n(i,i')} \frac{(X_{ijk_1} - X_{ijk_2})(X_{i'j'k_1} - X_{i'j'k_2})}{4n(i,i')[n(i,i')-1]} \times \frac{(X_{ij_1k_3} - X_{ij_1k_4})(X_{i'j_1'k_3} - X_{i'j_1'k_4})}{[n(i,i')-2][n(i,i')-3]},$$

and

$$\widehat{\tau}_{AB} = \sum_{|i-i'| \leq I^h}^{I} \frac{2n(i,i')[n(i,i')-1]}{I(J-1)^2 n_i^2(n_i-1)^2}\left[\sum_{j,j_1}^{J}\widehat{\sigma}_{ii'}(j, j_1, j, j_1) + \sum_{j,j_1,j',j_1'}^{J}\frac{\widehat{\sigma}_{ii'}(j, j_1, j', j_1')}{J^2} - \sum_{j,j_1,j_2}^{J}\frac{2\widehat{\sigma}_{ii'}(j, j_1, j, j_2)}{J}\right],$$

for some $1/5 < h < 1$. Then under the assumptions of Theorem 3.1, $\widehat{\tau}_{AB}$ is a consistent estimator of τ_{AB} as $I \to \infty$.

The proofs of Theorem 3.1 and Proposition 3.2 are given in the appendix.

Next, we consider to test whether copy number aberration exists in a given genomic region. Under the null hypothesis $H_0(A)$: *all* $\alpha_i = 0$, *for* $i = 1, \dots, I$, there is no copy number difference within the DNA segment of interest, where I is the total number of probes located in this DNA segment.

Similar to the test of no interaction effects, we use $T_A - E_A$ as the test statistic, where

$$T_A = \frac{1}{I-1}\sum_{i=1}^{I}\sum_{j=1}^{J}(\widetilde{X}_{i\cdot\cdot} - \widetilde{X}_{\dots})^2,$$

$$E_A = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j,j_1}^{J}\sum_{k=1}^{n_i}\frac{(X_{ijk} - \overline{X}_{ij.})(X_{ij_1k} - \overline{X}_{ij_1.})}{n_i(n_i-1)}.$$

The definition of T_A is different from that of ANOVA in that unweighted averages are used instead of weighted averages. The definition of E_A is different from that of the traditional MSE in that the within-subject correlation over time is taken into account. Such modification ensures the method to be valid for unbalanced sample sizes. and heteroscedasticity. The next theorem gives the asymptotic distribution of this test statistic for small sample sizes and a large number of probes.

Theorem 3.3. *For testing $H_0(A)$: all $\alpha_i = 0$, assume X_{ijk} have finite 16th central moment. Further, for all j, k, assume $\{X_{ijk}, i = 1, \dots\}$ is an α-mixing sequence with mixing coefficient $\alpha_m = O(m^{-5})$. Assume the τ_A defined below is bounded away from 0,*

$$\tau_A = \frac{1}{IJ^2}\sum_{i,i'}^{I}\frac{2n(i,i')[n(i,i')-1]}{n_i^2(n_i-1)^2}\left(\sum_{j,j_1}^{J}\sigma_{ii'jj_1}\right)^2.$$

Then under $H_0(A)$, $\sqrt{I}(T_A - E_A)/\tau_A \xrightarrow{d} N(0,1)$, as $I \to \infty$. A consistent estimator for τ_A is

$$\widehat{\tau}_A = \frac{1}{IJ^2}\sum_{|i-i'|\le I^h}^{I} \frac{2n(i,i')[n(i,i')-1]}{n_i^2(n_i-1)^2}\sum_{j,j_1,j',j_1'}^{J}\widehat{\sigma}_{ii'}(j,j_1,j',j_1'),$$

where $\widehat{\sigma}_{ii'}(j,j_1,j',j_1')$ is defined in Proposition 3.2.

Note that the tests given above are only applicable to a DNA segment that has at least 50 probes since they need $I \geq 50$ to have reliable type I error estimates for a variety of distributions (Zhang [2008]). On the other hand, this provides one of our advantages compared to other algorithms in that this will allow us to quickly screen the whole genome with only one or a few tests. If the number of probes on a chromosome region is less than 50, we do not recommend our tests. Instead, traditional parametric model/method may be applied with the tradeoff of possible violation on distributional assumptions.

To apply the above tests, it is not necessary to predetermine which region to apply the proposed test. If there are some units or blocks (such as chromosome arms) of initial interest, the tests can be applied to each of these blocks. Otherwise, we recommend to recursively apply the tests and partition the genome following the test-based partition algorithm in von Borries and Wang [2009]. Briefly, we start with all probes on the genome and apply the test. If the test is significant, we partition the genome into blocks and test each block (a block can be a chromosome, a chromosome arm, or for simplicity, half of the block that was previously tested) until the partition can not proceed due to small block size.

4. SIMULATION STUDIES

In this section, we compare the proposed nonparametric test (NPT) with some commonly-used methods in terms of type I error rate and power analysis. Hidden Markov models (HMM) and circular binary segmentation (CBS) have been widely used for copy number variation (CNV) identification (Fridlyand et al. [2004], Olshen et al. [2004], Lai et al. [2007]). They process DNA copy number data sample by sample. Consequently, correlation between multiple samples can not be taken into account and there is no guidelines available regarding how to combine the results from different samples. On the other hand, linear mixed effects models (LME) and generalized estimating equations (GEE) are the most commonly used methods for correlated data. Therefore, we first compare NPT with LME and GEE for data generated with unstructured correlations among the repeated measurements from the same probe. It is difficult to specify a reasonable correlation structure among data from different probes of the same subject. To avoid bias, we generate data by resampling from real copy numbers such that correlations among the probes are inherited for our generated data. All calculations and simulations are implemented with R programming language. HMM and CBS were conducted with R packages *aCGH* and *DNAcopy*, and LME and GEE calculations were conducted with R packages *nlme* and *geepack*. We first present the comparisons with LME and GEE.

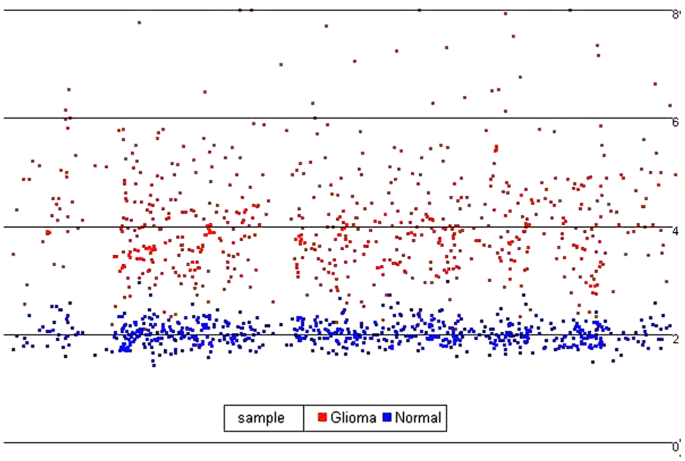

Figure 1. The plots of DNA copy numbers in chromosome 7q of normal and glioma samples. Red color denotes the copy numbers of glioma SNPs, and blue color denotes the copy numbers of normal SNPs. The x axis showed the relative genomic positions of each SNP on chromosome 7q.

For data generation, previous work has shown that amplification of chromosome 7q is associated with glioma tumor (Maher [2006]). We acquired the copy numbers of 3,000 probes that represent single nucleotide polymorphism (SNP) in chromosome 7q from Affymetrix 100K SNP arrays for both a healthy person and a glioma patient. A scatter plot of the copy numbers is given in Figure 1 with blue color for the healthy person and red color for the patient (please see the online version for colored Figure 1). The glioma sample has 7q amplification with a mean copy number 4.4. The normal sample has a mean value 2.05. The 3,000 copy numbers from the healthy patient provide the population for resampling under the null hypothesis. The 3,000 copy numbers from the glioma patient are to be used as the population for resampling under the alternatives.

We sample the copy numbers for 100 SNPs with an unbalanced design and create 5 repeated measures (time points) through introducing within-probe correlation as below.

- Firstly, at the jth time point, 6 replications were obtained for one fifth of the SNPs through resampling, and 4 for the remaining four fifths (denote as $X_{i,j}$). This creates unbalanced data such that some SNPs have 6 replications while others have 4 replications per time point.
- Next, an unstructured within-probe correlation was introduced iteratively. Suppose for SNP i, the correlation between the jth and (j+1)th time points is $\rho_{i,j}$. Given

a copy number $X^*_{i,j}$ for the jth time point, the random copy number $X^*_{i,j+1}$ of the (j+1)th time point can be generated by

$$X^*_{i,j+1} = \rho_{i,j} X^*_{i,j} + (1 - \rho_{i,j}) X_{i,j+1},$$

where $X^*_{i,1} = X_{i,1}$ and $\rho_{i,j} \sim \text{Uniform}(0.5, 1)$. Since $\rho_{i,j}$ are not constant, the correlation structure is not AR(1).

- Each data set generated contains a majority of SNPs from normal 7q (under H_0), and a small proportion, p_0, of SNPs from the glioma 7q (under H_a). We let p_0 ranges from 0 to 1% for the test of no probe effect, and from 0 to 4% for the test of no interaction effect. That is, each data set is mainly composed of normal copy numbers contaminated with a small proportion of glioma data.

For each contamination proportion of glioma copy numbers in the sampling data, we estimated the power curves of NPT, LME, and GEE at 0.05 level for the effects of SNP, and SNP by time interaction (Figure 2). For the SNP effect, the proposed method (NPT) had the fastest convergence rate to 1. At 0.9% contamination, the estimated power of NPT is 98.6%, whereas the power of LME is 56.2%, and that of GEE is only 29.5%. For the SNP by time interaction, the power of NPT outperformed the other two methods when there was at least 0.9% contamination. With 4% of contamination, NPT had a power of 96.8%, LME of 63.6%, and GEE of 89.0%.

For HMM and CBS, we generated 1,000 independent samples with 100 SNPs per sample. Under the null hypothesis, all copy numbers for 100 SNPs are resampled from normal 7q data, whereas a small percentage p_0 of them are from the glioma 7q under the alteratives. We let p_0 range from 0 to 50%. The power of HMM and CBS are estimated by the percentage of samples that are detected to contain copy number variations. For HMM, the proportions of detections are 45.9%, 70.8%, 56%, 30.8%, and 7% for $p_0 = 0, 5\%, 10\%, 20\%, 50\%$ respectively. That is, the type I error of HMM is unacceptably high while the power is low. This is because HMM calculation relies on the standard deviation of the copy number data. When p_0 is small, the standard deviation is small since most data come from normal SNPs that have a copy number that equals 2. Thus, HMM tends to identify focal copy number variations such that the type I error rate is high. When p_0 is large ($p_0 = 50\%$), the standard deviation is large, which results in low powers of HMM. The proportion of detections for CBS is .009%, 54.9%, 61.6%, 59.8%, and 39.2% for $p_0 = 0, 5\%, 10\%, 20\%, 50\%$ respectively. The type I error rate of CBS is acceptable under the null hypothesis of no copy number variation, but the power is relatively low comparing to NPT (Figure 2). The advantage of NPT compared to HMM and CBS lies in the fact that it can use multiple samples to increase power without assuming distributional assumptions whereas HMM and CBS are only for one sample and require specific parametric assumptions (such as Gaussian and Markov property for HMM) that may not hold for real data.

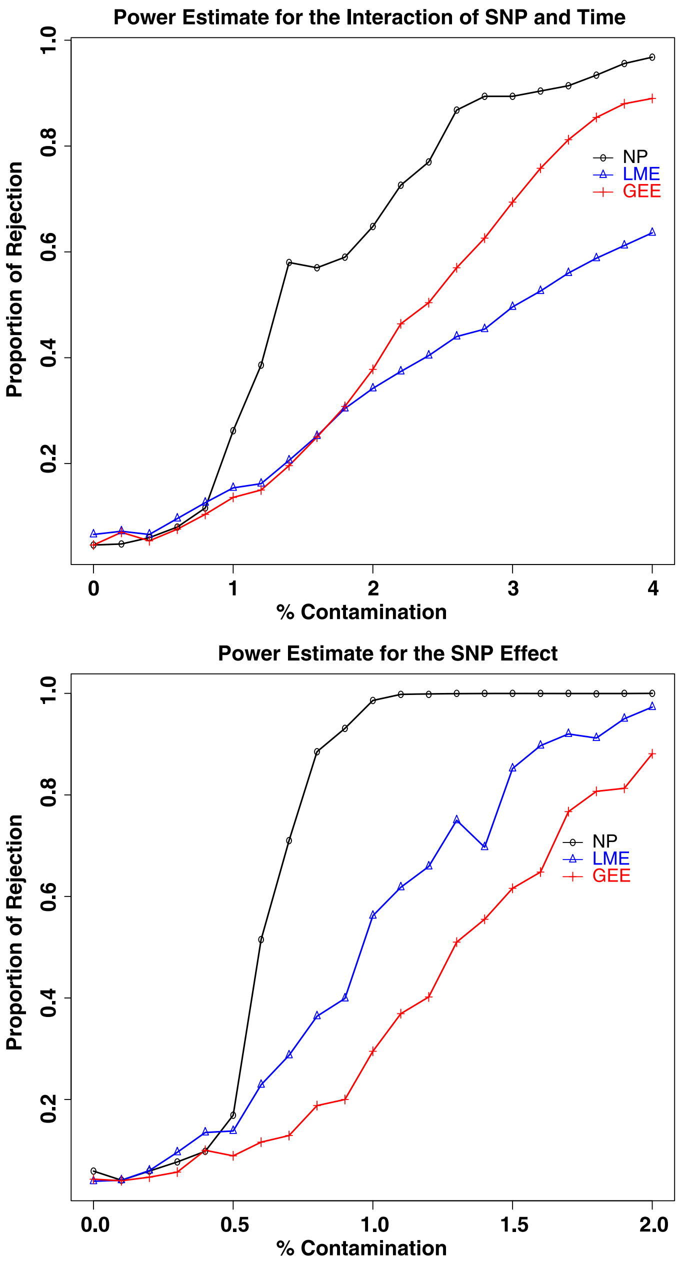

Figure 2. The estimated power for SNP by time interaction effect (top panel) and SNP effect (bottom panel) as the percent of contamination increases for an unbalanced design with unstructured correlations.

5. STUDY OF WILMS' TUMOR RELAPSE

Wilms' tumor typically occurs in children's kidney. Although the percentage of patients who survive at least five years is above 90%, 15% of patients will suffer from tumor relapse. Genetic aberrations such as loss of heterozygosity and chromosome copy number changes have been found to be associated with the tumor relapse (Grundy et al. [2005]; Yuan et al. [2005]). Recently, longitudinal studies have been conducted to identify biomarkers that are responsible for tumor progression and recurrence.

Natrajan et al. [2007] carried out aCGH experiments for 10 Wilms' tumor patients with relapse. The aCGH samples were conducted at both diagnosis and relapse for each patient. They used Breakthrough Breast Cancer Human CGH 4.6K 1.1.2 arrays that consist of 4,179 Bacterial Artificial Chromosome (BAC) clones. The BAC clones serve as probes for measuring the genomic DNA copy number. In their report, 29 chromosome regions were identified that may have copy number alterations responsible for Wilms' tumor relapse. However, their paired T-test for comparing the copy numbers between diagnosis and relapse did not discover any significant genomic regions. The reproducibility of their analysis was low, and the claimed biomarkers may not be useful for predicting the potential relapse of new Wilms' patients. In fact, only 6 of the 29 regions were found in 2 of the 10 patients in their paper. Motivated by the need to redo the analysis with improved power, we acquired the raw aCGH data and conducted analysis with the following steps.

We first performed quality control and normalization for the raw data. As female and male have different number of sex chromosomes, we removed X and Y chromosomes from the data to avoid confounding effect of imbalance. The raw data were adjusted to a baseline by subtracting the median background signal. In the experiment, each probe was labeled with two fluorescent dyes, Cy5 and Cy3. The fluorescent intensity ratio of Cy5/Cy3 were used as input data. The Cy5/Cy3 ratio were subject to quantile normalization across all samples (Bolstad et al. [2003]). The processed data had a median copy number of 2 and a standard deviation of 0.04 for each sample. They were used for subsequent analysis.

As discussed in section 1, a first goal of copy number study is usually to detect the gain or loss of a chromosome arm because it is often the unit of genomic mutation and translocation activity. For instances, Lu et al. [2002] found that the gain of chromosome 1q is associated with relapse of Wilms' tumor. We applied our proposed methods to each chromosome arm for probe and probe×time interaction effects. Out of 44 autosomal chromosome arms, 5 arms, 13p, 14p, 15p, 21p, and 22p, had no probes in the arrays. For the other arms, the minimum number of probes was 84, and the maximum number was 699. Table 1 lists the chromosome arms that have p-values less than 0.05 for the tests. After Bonferroni correction, only 6 arms showed significant probe×time interaction at a familiy-wise error rate of 0.05 (totally 16 arms significant if no Bonferroni correction). This implies that the copy numbers of some of the probes in these arms varied between diagnosis and relapse. Of the remaining chromosome arms, two arms showed some weak evidence of probe effects (8p and 21q).

*Table 1. Summary of p-values that are less than 0.05 calculated by NPT methods for each chromosome arm. 'Chr.' refers to the chromosome number. The ones labeled with * are significant at 0.05 level after Bonferroni correction*

Chr.	p arm		q arm	
	probe	probe×time	probe	probe×time
2	2.55×10^{-3}	2.72×10^{-3}		
3	6.69×10^{-3}	2.47×10^{-4} *	0.014	0.010
5	5.07×10^{-9} *	1.21×10^{-11} *	2.38×10^{-4} *	9.93×10^{-5} *
6				0.034
7	2.43×10^{-8} *	0.023		0.031
8	0.041			
9	0.036	0.036		
11			5.05×10^{-8} *	4.35×10^{-7} *
12		0.016		
15				0.038
16			5.54×10^{-5} *	6.84×10^{-7} *
17				0.015
18	2.55×10^{-14} *	1.77×10^{-15} *		
21			1.59×10^{-3}	
22				0.013

In our analysis, 26 chromosome arms (those in Table 1 that do not have p-values provided) were not detected for any effect even with weak evidence. Further analysis should be conducted for these arms to evaluate if an arm is amplified or lost at both time points. This is important because the desired biomarkers for predicting relapse should show a consistent pattern between diagnosis and relapse. If a genetic event only occurs in one of the two time points, its association with tumor recurrence is not clear. We calculated the mean copy number for each arm. Unfortunately, none of these mean values were abnormally higher or lower than 2.

For the rest of the arms, at least weak evidence from the samples indicates that some effects exist. However, for the purpose of identifying prognostic biomarkers, chromosome arms with probe×time interaction alone are not enough although the interaction may indicate important genetic regulation mechanisms. The findings here can not be verified from the real time PCR result in Natrajan et al. [2007] as they did not report a list of alterations shown by their real time PCR analysis. Further biological studies are necessary to confirm the alterations that we report here.

We explored chromosome 8p and 21q that showed only weak evidence of a probe effect. Significant probe effect suggests some regions in the two arms have a gain or loss of DNA copies. By calculating the mean value of each probe with the measures from both diagnosis and relapse, we found four regions with abnormal copy numbers. The results were summarized in Table 2. Chromosome region 8p21.3 was found to have a DNA deletion. Two genes are encoded in this region, INTS10 and LPL. INTS10 is a subunit of RNA polymerase. Reduced expression level of RNA polymerase could lead to abnormal expression of many other genes.

Table 2. Summary of the copy number alterations detected for both primary and relapse tumors

Genomic region	Gene	Function
8p21.3	INTS10	RNA transcription
	LPL	lipoprotein
21q21.1	CR614803	NA
	NCAM2	NA
21q21.3	CYYR1	NA
21q22.3	NX1	anti-viral response
	NX2	GTPase
	TMPRSS2	Serine protease

Thus, it is a potential oncogenesis gene. LPL is responsible for lipoprotein uptake, and was reported to be associated with prostate cancer (Narita et al. [2004]). Chromosome 21q11.1 and 21q11.3 loss may affect the expression of genes CR614803, NCAM2, and CYYR1. However, the gene functions and their relevancy with cancer is not clear currently. The loss of 21q22.3 was associated with functions of 3 genes, NX1, NX2, and TMPRSS2. NX1 is responsible for anti-viral reaction; NX2 is a subunit of GTPase; TMPRSS2 belongs to the serine protease family. Both GTPase and serine protease are involved in a number of fundamental gene regulation pathways. The four selected regions overlapped with 2 copy number alterations reported by Natrajan et al. [2007].

6. SUMMARY AND DISCUSSION

Longitudinal DNA copy number studies can provide unique insights into the genetic abnormalities involved in disease development and progression. However, there are a number of challenges faced in statistical inference. Researchers often use over-simplified analysis methods that are not able to provide sufficient statistical power and justification. In this paper, we provided a set of robust hypothesis testing tools based on a more realistic model setup and distribution-free non-parametric statistics. Both continuous and discrete response variables are allowed in the model. The proposed method basically includes a large number of probes in one model and conducts hypothesis tests. By handling all probes in a genomic region simultaneously instead of performing analysis for each individual probe, the proposed method has significant gain over commonly used traditional methods in two senses: (1) The large number of probes provide asymptotic power for our tests. (2) There is a significant reduction in the number of tests required to screen the whole genome comparing to traditional methods. This is true even after taking into account possible recursive partitioning. In addition, the proposed method only requires consistent estimates of the asymptotic variance instead of estimating all parameters for each probe leading to a large number of unknown parameters and extensive computation (such as the transition probability matrix for each copy number state in Hidden Markov Models). In summary, in this article a convenient set of tests are provided to analyze longitudinal copy number data with small sample sizes, and therefore are expected to have potentially broad applications in genomic screenings.

Some studies of tumor genomes are interested in detecting focal CNVs of small region, such as of size 1,000 base pairs. We remark that detection of such focal CNVs depends on the resolution of the chips used to produce copy number data. As of the date of this manuscript, the densest array for copy number study is the Affymetrix Genome-Wide Human SNP Array 6.0 that features a total of 1.8 million markers (among which 946,000 are copy number probes and 906,600 are SNP markers, see the fact sheet at http://www.affymetrix.com/support/technical/datasheets/genomewide_snp6_datasheet.pdf). The human genome contains approximately 3 billion base pairs. Hence the average inter-marker distance over all 1.8 million SNP and copy number markers combined is about 1,667 base pairs. Therefore, current array technology does not allow us to consider 1kb CNVs detection with the proposed method. As technologies continue to advance, it may be possible to have denser arrays available in the future and detection of focal CNVs can be conducted with the proposed method.

When there are only two time points, paired analysis tends to be used due to its simplicity. As mentioned in the introduction, the paired analysis can only be applied to an individual probe. Due to the small sample size limitation, the paired test statistic has a large variance and therefore is lacking power to detect copy number variations. On the other hand, the proposed method includes all probes of interest in one model and takes into account the between-probe correlations to increase its power. We conducted a small simulation study in which we generated random data for 100 probes with 5 replications at each of 2 time points from normal distribution with standard deviation 1 and correlation 0.5. Under the null hypothesis where all data have a mean 0, the type I error rate for the proposed test is 0.043, and that for paired analysis is 0.048. Under the alternative hypothesis, the mean was 0 for 80% of probes, and 1 for the remaining 20%. The paired test has only 11.6% power to detect the probe effect, whereas the proposed method has 100% power. This is a simple comparison in favor of the proposed method over a paired analysis.

A last discussion of the proposed method is that it is designed to detect CNVs frequently shared in the disease population because such CNVs are representative for elucidating common pathological features that can be used for cancer treatment therapy. If the interest is in detecting unique CNVs in an individual, we advise the user to seek alternative methods as the proposed method requires replications.

ACKNOWLEDGEMENTS

We are grateful to the anonymous referees whose insightful comments have led to significant improvement of the manuscript.

APPENDIX: SOME TECHNICAL PROOFS

Here we give a sketch proof of Theorem 3.1. Theorem 3.3 can be proved following similar technical arguments. The proof involves first finding a projection of T_{AB} onto the space $span\{\mathbf{X}_i = (X_{i11}, \ldots, X_{iJn_i})', i = 1, \ldots, I\}$. The $P_{AB}(\varepsilon)$ in Lemma A.1 is the projection of T_{AB}. Then the test statistic is shown to be asymptotically equivalent to $\sqrt{I}(P_{AB}(\varepsilon) - E_{AB})$ whose asymptotic distribution can be easily derived.

Lemma A.1. *Under the settings and assumptions of Theorem 3.1 and under $H_0(AB)$, we have*

$$\sqrt{I}(T_{AB} - P_{AB}(\varepsilon)) \xrightarrow{p} 0 \ \textit{as}\ I \to \infty,$$

where $P_{AB}(\varepsilon) = \frac{1}{I(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{\varepsilon}_{ij.} - \tilde{\varepsilon}_{i..})^2$.

Proof. Under $H_0(AB)$, we can write

$$\begin{aligned}
&P_{AB} - T_{AB}(\varepsilon)\\
&= \frac{2}{I(I-1)(J-1)} \sum_{i \neq i_1}^{I} \sum_{j=1}^{J} (\bar{\varepsilon}_{ij.} - \tilde{\varepsilon}_{i..})(\bar{\varepsilon}_{i_1 j.} - \tilde{\varepsilon}_{i_1..})\\
&= \frac{2}{I(I-1)(J-1)} \left\{ \sum_{i \neq i_1}^{I} \sum_{j=1}^{J} \bar{\varepsilon}_{ij.} \bar{\varepsilon}_{i_1 j.} - J \sum_{i \neq i_1}^{I} \widetilde{\varepsilon}_{i..} \widetilde{\varepsilon}_{i_1..} \right\}\\
&= \frac{2}{I(I-1)(J-1)} \left\{ \sum_{i \neq i_1}^{I} \sum_{j=1}^{J} \bar{\varepsilon}_{ij.} \bar{\varepsilon}_{i_1 j.} - \sum_{i \neq i_1}^{I} \sum_{j=1}^{J} \sum_{j_1=1}^{J} \frac{\bar{\varepsilon}_{ij.} \bar{\varepsilon}_{i_1 j_1.}}{J} \right\}
\end{aligned}$$

Note that for each i and j, $\bar{\varepsilon}_{ij.}$ is the average of finite number of independent terms. Apply Theorem 5.2 of Bradley [2005], we know that the process defined by Boreal functions of finite number of independent α-mixing processes is still an α-mixing process with the same mixing coefficient. Therefore, for each j, $\{\bar{\varepsilon}_{ij.}, i = 1, \ldots\}$ is an α-mixing process with mixing coefficient $\alpha_m = O(m^{-5})$ as the lag $m \to \infty$. By Lemma 2.1 of Wang and Akritas [2010], we know that $E(\sum_{i \neq i_1}^{I} \bar{\varepsilon}_{ij.} \bar{\varepsilon}_{i_1 j.})^2 = O(I^2)$ and $E(\sum_{i \neq i_1}^{I} \bar{\varepsilon}_{ij.} \bar{\varepsilon}_{i_1 j_1.})^2 = O(I^2)$ for each j, j_1. Since J stays finite, we have $\sqrt{I}(T_{AB} - P_{AB}(\varepsilon)) \xrightarrow{p} 0$ under $H_0(AB)$ as $I \to \infty$. □

Proof of Theorem 3.1. Lemma A.1, we need only to consider the asymptotic distribution of $Q_{AB}(\varepsilon) = \sqrt{I}(P_{AB}(\varepsilon) - E_{AB})$ under $H_0(AB)$.

With some algebra, we can write

$$\text{(A1)} \qquad Q_{AB}(\varepsilon) = \frac{1}{\sqrt{I}(J-1) n_i (n_i - 1)} \sum_{i}^{I} H_i,$$

where

$$H_i = \sum_{k \neq k_1}^{n_i} \left[\sum_{j}^{J} \varepsilon_{ijk} \varepsilon_{ijk_1} - \frac{1}{J} \sum_{j, j_1}^{J} \varepsilon_{ijk} \varepsilon_{ij_1 k_1} \right].$$

Therefore, $E[Q_{AB}(\varepsilon)] = 0$. The number of common subjects $n(i, i')$, whose copy numbers for both probes i and i' from the each subject are observed, contributes to the variance of $Q_{AB}(\varepsilon)$. It follows that

$$\begin{aligned}
&Var(Q_{AB}(\varepsilon))\\
&= \sum_{i}^{I} \sum_{i'}^{I} \sum_{k \neq k_1}^{n(i,i')} \frac{2I^{-1}(J-1)^{-2}}{n_i^2 (n_i - 1)^2} \Bigg[\sum_{j,j'}^{J} E(\varepsilon_{ijk} \varepsilon_{i'j'k}) E(\varepsilon_{ijk_1} \varepsilon_{i'j'k_1})\\
&\quad + \frac{1}{J^2} \sum_{j, j_1, j', j_1'}^{J} E(\varepsilon_{ijk} \varepsilon_{i'j'k}) E(\varepsilon_{ij_1 k_1} \varepsilon_{i'j_1' k_1})\\
&\quad - \frac{2}{J} \sum_{j, j_1, j_2}^{J} E(\varepsilon_{ijk} \varepsilon_{i'j_1 k}) E(\varepsilon_{ijk_1} \varepsilon_{i'j_2 k_1}) \Bigg]\\
&= \frac{2}{I(J-1)^2} \sum_{i,i'}^{I} \frac{n(i,i')[n(i,i') - 1]}{n_i^2 (n_i - 1)^2} \Bigg[\sum_{j,j'}^{J} \sigma_{ii',jj'}^2\\
&\quad + \frac{1}{J^2} \sum_{j, j_1, j', j_1'}^{J} \sigma_{ii',jj'} \sigma_{ii',j_1 j_1'} - \frac{2}{J} \sum_{j, j_1, j_2}^{J} \sigma_{ii',jj_1} \sigma_{ii',jj_2} \Bigg].
\end{aligned}$$

The $Var(Q_{AB}(\varepsilon))$ is finite since the double summation over i and i' is of order $O(I)$ by part (a) of Lemma 2.1 in Wang and Akritas [2010]. To show the asymptotic normality of $Q_{AB}(\varepsilon)$, note that H_i in (A1) are Boreal functions of finite number of random variables. Therefore, $\{H_i, i = 1, \ldots\}$ is an α-mixing process with the same mixing coefficient as $\{\varepsilon_{ijk}, i = 1, \ldots\}$. Hence, the Central Limit Theorem of Wang and Akritas [2010] for α-mixing process (part (b) of Lemma 2.1) can be applied to Q_{AB} if we can show $\limsup_i E(H_i^{16}) < \infty$.

By Hölder's Inequality,

$$\begin{aligned}
E(H_i^{16}) \leq n_i^{15} (n_i - 1)^{15} \sum_{k \neq k_1}^{n_i} 2^{15} \Bigg[&J^{15} \sum_{j}^{J} E(\varepsilon_{ijk})^{16} E(\varepsilon_{ijk_1})^{16}\\
&+ J^{14} \sum_{j, j_1}^{J} E(\varepsilon_{ijk})^{16} E(\varepsilon_{ij_1 k_1})^{16} \Bigg].
\end{aligned}$$

With given assumption $E(\varepsilon_{ijk})^{16} < \infty$, we know that $E(H_i^{16}) < \infty$ for all i. This completes the proof. □

Proof of Proposition 3.2.

$$\begin{aligned}
&|\widehat{\tau}_{AB} - \tau_{AB}|\\
&\quad \leq \sum_{|i - i'| \leq I^h}^{I} \frac{2n(i,i')[n(i,i') - 1]}{I(J-1)^2 n_i^2 (n_i - 1)^2}
\end{aligned}$$

$$\times \left[\sum_{j,j_1}^{J} |\widehat{\sigma}_{i,i'}(j,j_1,j,j_1) - \sigma^2_{ii'jj_1}| \right. \tag{A2}$$

$$+ \frac{1}{J^2} \sum_{j,j_1,j',j'_1}^{J} |\widehat{\sigma}_{i,i'}(j,j_1,j',j'_1) - \sigma_{ii'jj_1}\sigma_{ii'j'j'_1}| \tag{A3}$$

$$\left. + \frac{2}{J} \sum_{j,j_1,j_2}^{J} |\widehat{\sigma}_{i,i'}(j,j_1,j,j_2) - \sigma_{ii'jj_1}\sigma_{ii'jj_2}| \right] \tag{A4}$$

$$+ \frac{2n(i,i')[n(i,i')-1]}{I(J-1)^2 n_i^2(n_i-1)^2} \sum_{|i-i'|>I^h}^{I} \left| \sum_{j,j_1}^{J} \sigma^2_{ii'jj_1} \right. \tag{A5}$$

$$\left. + \frac{1}{J^2} \left(\sum_{j,j_1}^{J} \sigma_{ii'jj_1} \right)^2 - \frac{2}{J} \sum_{j,j_1,j_2}^{J} \sigma_{ii'jj_1}\sigma_{ii'jj_2} \right| . \tag{A6}$$

Apply inequality $\sigma_{ii'jj_1} \leq C\alpha^{1/2}_{|i-i'|}$ (see Billingsley [1995]), where C is some finite constant, the terms in (A5) and (A6) are bounded by

$$\frac{2n(i,i')[n(i,i')-1]}{I(J-1)^2 n_i^2(n_i-1)^2} \sum_{|i-i'|>I^h}^{I} J^2 4C\alpha_{|i-i'|}$$

$$\leq 8I^{-1} \sum_{i=1}^{I} \sum_{m=I^h}^{I} J^2 C I^{-5h} = O(I^{1-5h}) \to 0 \text{ for } h > 1/5.$$

The terms in (A2)–(A4) can be shown to converge to zero in probability by showing that their second moment is asymptotically negligible. We show one of them here:

$$E \left\{ \sum_{|i-i'|\leq I^h}^{I} \frac{2n(i,i')[n(i,i')-1]}{I(J-1)^2 n_i^2(n_i-1)^2} \right.$$

$$\left. \times \sum_{j,j_1}^{J} |\widehat{\sigma}_{i,i'}(j,j_1,j,j_1) - \sigma^2_{ii'jj_1}| \right\}^2$$

$$\leq \sum_{|i-i'|\leq I^h}^{I} \sum_{|i_1-i'_1|\leq I^h}^{I} \sum_{j,j_1}^{J} \sum_{j',j'_1}^{J} \frac{4}{I^2(J-1)^4}$$

$$\times \frac{E|(\widehat{\sigma}_{i,i'}(j,j_1,j,j_1) - \sigma^2_{ii'jj_1})(\widehat{\sigma}_{i_1,i'_1}(j,j_1,j,j_1) - \sigma^2_{i_1i'_1jj_1})|}{n_i^2(n_i-1)^2 n(i,i')^{-1}[n(i,i')-1]^{-1} n_{i_1}(n_{i_1}-1)}$$

$$\leq \sum_{|i-i'|\leq I^h}^{I} \sum_{|i_1-i'_1|\leq I^h}^{I} \frac{2n(i,i')[n(i,i')-1]}{I^2(J-1)^4 n_i^2(n_i-1)^2}$$

$$\times \sum_{j,j_1}^{J} \sum_{j',j'_1}^{J} 8C\alpha_{|i-i'|}\alpha_{|i_1-i'_1|} = o(1),$$

where the last equality is due to the fact that $\sum_{i=1}^{I}\sum_{i'=1}^{I} \alpha_{|i_1-i'_1|} = O(I)$ and $h < 1$. Applying similar proof to the terms in (A3) and (A4), this will complete the proof. □

Received 23 April 2009

REFERENCES

Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D., and Hurles, M. (2008). A robust statistical method for case-control association testing with copy number variation. *Nature Genetics*, 40:1245–1252.

Billingsley, P. (1995). *Probability and Measure*. Third Edition, Wiley, New-York. MR1324786

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93.

Bradley, R. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:1549–5787. MR2178042

Daruwala, R., Rudra, A., Ostrer, H., Lucito, R., Wigler, M., and Mishra, B. (2004). A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl. Acad. Sci. USA*, 101:16292–16297.

Fan, J. and Lin, S. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.*, 93:1007–1021. MR1649196

Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D., and Jain, A. (2004). Hidden Markov models approach to the analysis of array CGH data. *J. Multivariate Anal.*, 90:132–153. MR2064939

Grundy, P. E., Breslow, N. E., Li, S., Perlman, E., Beckwith, J. B., Ritchey, M. L., Shamberger, R. C., Haase, G. M., D'Angio, G. J., Donaldson, M., Coppes, M. J., Malogolowkin, M., Shearer, P., Thomas, P. R., Macklis, R., Tomlinson, G., Huff, V., and Green, D. M. (2005). Loss of heterozygosity for chromosomes 1p and 16q is an adverse prognostic factor in favorable-histology Wilms tumor: a report from the National Wilms Tumor Study Group. *J. Clin. Oncol.*, 23(29):7312–21.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, 7(6):819–37.

Konishi, T. (2004). Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. *BMC Bioinformatics*, 5:5.

Lai, L. A., Paulson, T. G., Li, X., Sanchez, C. A., Maley, C., Odze, R. D., Reid, B. J., and Rabinovitch, P. S. (2007). Increasing genomic instability during premalignant neoplastic progression revealed through high resolution array-CGH. *Genes Chromosomes Cancer*, 46(6):532–42.

Li, H., Lindsay, B., and Waterman, R. (2003). Efficiency of projected score methods in rectangular array asymptotics. *J. R. Statist. Soc. B*, 65:191–208. MR1959821

Lu, Y. J., Hing, S., Williams, R., Pinkerton, R., Shipley, J., and Pritchard-Jones, K. (2002). Chromosome 1q expression profiling and relapse in Wilms' tumour. *Lancet*, 360(9330):385–6.

Maher, E. A. (2006). Marked genomic difference characterize primary and secondary globalstoma subtypes and identify two distinct molecular and clinical secondary glioblastoma entities. *Cancer Res.*, 66(23):11502–11513.

Mullighan, C. G., Phillips, L. A., Su, X., Ma, J., Miller, C. B., Shurtleff, S. A., and Downing, J. R. (2008). Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*, 322(5906):1377–80.

Narita, S., Tsuchiya, N., Wang, L., Matsuura, S., Ohyama, C., Satoh, S., Sato, K., Ogawa, O., Habuchi, T., and Kato, T. (2004). Association of lipoprotein lipase gene polymorphism with risk of prostate cancer in a Japanese population. *Int. J. Cancer*, 112(5):872–6.

Natrajan, R., Little, S. E., Sodha, N., Reis-Filho, J. S., Mackay, A., Fenwick, K., Ashworth, A., Perlman, E. J., Dome, J. S., Grundy, P. E., Pritchard-Jones, K., and Jones, C. (2007). Analysis by array CGH of genomic changes associated with the progression or relapse of Wilms' tumour. *J. Pathol.*, 211(1):52–9.

Neymann, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32. MR0025113

Olshen, A., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5:557–572.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

Rigaill, G., Hupé, P., Almeida, A., Rosa, P. L., Meyniel, J.-P., Decraene, C., and Barillot, E. (2008). ITALICS: an algorithm for normalization and DNA copy number calling for affymetrix SNP arrays. *Bioinformatics*, 24:768–774.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA*, 42:43–47. MR0074711

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused LASSO. *Biostatistics*, 9(1):18–29.

Tsai, G.-F. and Qu, A. (2008). Testing the significance of cell-cycle patterns in time-course microarray data using nonparametric quadratic inference functions. *Comput. Stat. Data Anal.*, 52(3):1387–1398. MR2422743

von Borries, G. and Wang, H. (2009). Partition clustering of high dimensional low sample size data based on p-values. *Computational Statistics and Data Analysis*, 53:3987–3998.

Wang, H. and Akritas, M. (2010). Inference from heteroscedastic functional data. *J. Nonparametr. Stat.*, 22(2):149–168. MR2549431

Winchester, L., Yau, C., and Ragoussis, J. (2009). Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics & Proteomics*, 8:353–366.

Yuan, E., Li, C. M., Yamashiro, D. J., Kandel, J., Thaker, H., Murty, V. V., and Tycko, B. (2005). Genomic profiling maps loss of heterozygosity and defines the timing and stage dependence of epigenetic and genetic events in Wilms' tumors. *Mol. Cancer Res.*, 3(9):493–502.

Zhang, K. (2008). *Inference of Nonparametric Hypothesis Testing on High Dimensional Longitudinal Data and Its Application in DNA Copy Number Variation And Microarray Data Analysis.* Ph.D. Dissertation, Kansas State University.

Ke Zhang
Department of Pathology
School of Medicine and Health Sciences
University of North Dakota
Grand Forks, ND 58202
E-mail address: kzhang@medicine.nodak.edu

Haiyan Wang
Department of Statistics
Kansas State University
Manhattan, KS 66506
E-mail address: hwang@ksu.edu