Harvard University
Harvard University Biostatistics Working Paper Series

Year 2007 *Paper* 73

Assessing Population Level Genetic Instability via Moving Average

Samuel McDaniel* Rebecca Betensky†
Tianxi Cai‡

*Harvard University
†Harvard University, betensky@hsph.harvard.edu
‡Harvard University, tcai@hsph.harvard.edu

Assessing Population Level Genetic Instability via Moving Average

Samuel McDaniel, Rebecca Betensky and Tianxi Cai*

Department of Biostatistics, Harvard University, Boston, MA 02115

**email: tcai@hsph.harvard.edu*

Abstract

Tumoral tissues tend to generally exhibit aberrations in their DNA sequence of copy numbers which are associated with the development and progression of cancer. Consequently interests lie in identifying the true underlying sequence of copy numbers along the entire genome. The analysis of array-based Comparative Genomic Hybridization data seeks to establish this. To address some of the shortfalls of existing methods, including strong model assumptions, lack of sampling variability of estimators, and the assumption that clones are independent, we propose a simple graphical approach to assess population-level genetic alterations over the entire genome based on moving average. Covariates are incorporated through a possibly mis-specified *working* model and sampling variabilities of estimators are approximated using a resampling method that is based on perturbing observed processes. Our proposal, which is applicable to part, an entire or multiple chromosomes, is illustrated by applying it to datasets from two separate studies.

Keywords: *aCGH data, Moving average, Perturbation method, Gaussian process, Genomic data.*

1 Introduction

Genetic analyses of a variety of cancers have suggested that losses and gains in the DNA sequence of copy numbers are associated with cancer detection and development (Pinkel et al., 2005). The potential association between genetic instability and development of cancer has resulted in a preponderance of research. A chromosomal loss in a tumor cell typically results in the under-expression of tumor suppressor genes whose activity prevent tumor development, while copy number gains tend to result in the over expression of proto-oncogenes that promote tumor growth (Heiskanen et al., 2000). Array-based Comparative Genomic Hybridization (aCGH) assay offers a high-throughput approach to compare the DNA copy numbers of genetic materials of tumor and reference samples across the whole genome. Test (tumoral) and reference (normal) samples are respectively treated with red and green fluorescent dye and then mixed. The combined sample is subsequently hybridized to microchips with probes each corresponding to a location-specific clone of the genome and covering the entire genome (Olshen et al., 2004). At each location, the copy number alterations are measured by the $\log_2$ ratio of the fluorescence intensities of the two colors. Reference sample is typically diploid and so, in the ideal setting, in the absence of contaminated cells, its DNA would have a *normal* copy number of 2. On the other hand, regions of copy number loss in the tumoral sample would have a copy numbers of 1 (corresponding to a loss of heterozygosity), while genomic regions of copy number gain would have copy numbers of 3 (haploid duplication) or more (polyploid duplication). This translates into $\log_2$ ratios of $-1, 0$ and ≥ 0.58, corresponding to copy number loss, no-change, and gain respectively. It is important to note that, due to both biological and experimental reasons, the observed $\log_2$ ratios

may deviate from these theoretical values (REFERENCE?)

By identifying and characterizing new oncogenes (amplified regions) and tumor suppressor genes (deleted regions), we can better understand the molecular mechanisms associated with the initiation, progression, and metastasis of the cancer. This may consequently lead to more effective treatment selection and patient care. To analyze aCGH data, various statistical methods have been developed to identify regions with genetic aberrations that are associated with different types of tumor. A common approach is to determine threshold values for defining gains and losses accounting for data variability. For example, Pollack et al. (1999, 2002), Weiss et al. (2003) and Aguirre et al. (2004) estimated the variability of the $\log_2$ ratios that correspond to no genetic alteration under the assumption that $\log_2$ ratios corresponding to no genetic alteration in the tumor are normally distributed with mean zero. A three-component mixture model approach was considered by Hodgson et al. (2001) where the components correspond to copy number loss, gain and no-change. These methods require the normality assumption and independence between clones. Another commonly used approach considers segments of common $\log_2$ ratio means and aims to identify change-points of the means. For example, Olshen et al. (2004) proposed a circular binary segmentation (CBS) algorithm that identifies the change points through successive comparison of segments of the chromosome. Local significance was evaluated via permutation tests and a pruning algorithm was used to control the number of change points. Picard et al. (2004) identified change points for the sequence of $\log_2$ ratios using a penalized likelihood-based approach. These method also require the assumption of independence between clones. More recently, various methods have been developed to incorporate the possible dependence

between clones. Fridlyand et al. (2004) proposed a discrete-state Hidden Markov Model (HMM) approach assuming that given the genetic states at all previous locations, the genetic state at a given location depends only on the true state at the immediately previous location. The "states" correspond to segments of common mean and a change in state corresponds to a change-point. The number of mean levels on each chromosome was determined by an AIC criterion. Engler et al. (2006) accounted for the dependence of the data by using a pseudolikelihood function to fit a Gaussian mixture model with a Hidden Markov structure. Rueda et al. (2006) employed a rather computationally intensive method based on a Hidden Markov Model without pre-specifying the number of states. Other Hidden Markov models have been examined by, for example, Shah et al. (2006) and Guha et al. (2006).

Most existing methods are derived under relatively strong model assumptions. Violation of these assumptions may lead to incorrect conclusions about the association between the genetic instability and clinical outcomes of interest. Also, existing literature focuses primarily on obtaining point estimation of the parameters of interest without accounting for the sampling variability in such estimators. Furthermore, a majority of these methods do not account for possibly inflated type I error rate while making simultaneous inference along the entire genome. With advancement of technology, the number of locations available may become increasingly large. Traditional methods such as the Bonferroni adjustment or controlling for the false discovery rate (Benjamini and Hochberg, 1995) may be too conservative to detect regions of interest when there is high correlation between adjacent regions. To overcome such difficulties, we propose a simple graphical approach to assessing population-level genetic alterations over the entire genome

based on a moving average technique. To summarize the population level genetic instability, we consider average instability levels across various regions of interest and propose non-parametric estimates without requiring the commonly employed normality assumptions. When there are covariates that may affect the genetic instability, we propose the use of possibly mis-specified *working* models to approximate the association between the $\log_2$ ratios and the covariates and develop procedures for making inference about the average covariate effect without requiring the models to hold. In section 2, we provide a general framework for making inference about the population level genetic instability as well as the average covariate effect on the genetic instability. We illustrate the proposed procedures using an aCGH dataset from a Meningioma study in section 3. Some discussions are given in section 4.

2 Procedures for Making Inference about the Covariate Effect on Genetic Instability

Let $X_{ik}(t)$ denote the $\log_2$aCGH of the ith subject at location t of the kth measurement and let $\mathbf{Z}_{ik}$ denote the $p \times 1$ baseline covariate of the ith subject corresponding to the kth measurement, for $k = 1, ..., K$, $i = 1, ..., n$ and $t \in \mathcal{T}$. Assume that K is a fixed constant and n sets of clustered observations $\{X_{i1}(t), \cdots, X_{iK}(t), t \in \mathcal{T}, \mathbf{Z}_{i1}, ..., \mathbf{Z}_{iK}\}$ are independent and identically distributed. Without loss of generality, let $\mathcal{T} = [0, \tau]$.

We are interested in assessing the population level genetic instability within various patient populations indexed by the covariate information $\mathbf{Z}$ and comparing the genetic instability between

different patient populations. To this end, we propose to approximate the association between the covariate $\mathbf{Z}$ and $X(t)$ through the marginal regression *working* model

$$X_{ik}(t) = g\{\boldsymbol{\gamma}_\mu(t)'\vec{\mathbf{Z}}_{ik}\} + h\{\boldsymbol{\gamma}_\sigma(t)'\vec{\mathbf{Z}}_{ik}\}\epsilon_{ik}, \tag{1}$$

where ϵ_{ik} is a mean zero random variable, $\vec{\mathbf{Z}} = (1, \mathbf{Z}')'$, $g(\cdot)$ and $h(\cdot)$ are pre-specified strictly increasing functions with $h(\cdot) > 0$, $\boldsymbol{\gamma}_\mu(t) = (\alpha_\mu(t), \boldsymbol{\beta}_\mu(t)')'$ and $\boldsymbol{\gamma}_\sigma(t) = (\alpha_\sigma(t), \boldsymbol{\beta}_\sigma(t)')'$. Thus, $\boldsymbol{\beta}_\mu(t)$ quantifies the average effect of $\mathbf{Z}$ on the population mean level of genetic instability at location t and $\boldsymbol{\beta}_\sigma(t)$ quantifies the average effect of $\mathbf{Z}$ on the population variation of genetic instability at location t. In the simple setting where interest lies in assessing the population average genetic instability, we may employ the null model with no covariates and thus summarize it by $\mu(t) = g\{\alpha_\mu(t)\}$. If $\mathbf{Z}$ is an index of cancer sub-type, then one may summarize the difference between the cancer sub-types based on $\boldsymbol{\beta}_\mu(t)$ and $\boldsymbol{\beta}_\sigma(t)$.

To obtain an estimate for the regression parameter $\boldsymbol{\gamma}_\mu(t)$, we consider the following simple estimating equation

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \vec{\mathbf{Z}}_{ik}\left\{X_{ik}(t) - g(\boldsymbol{\gamma}'\vec{\mathbf{Z}}_{ik})\right\} = 0$$

Let $\widehat{\boldsymbol{\gamma}}_\mu(t)$ denote the solution to the above estimating equation. An estimate of $\boldsymbol{\gamma}_\sigma(t)$ may be obtained as $\widehat{\boldsymbol{\gamma}}_\sigma(t)$, the solution to

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \vec{\mathbf{Z}}_{ik}\left[\left\{X_{ik}(t) - g(\widehat{\boldsymbol{\gamma}}_\mu(t)'\vec{\mathbf{Z}}_{ik})\right\}^2 - h(\boldsymbol{\gamma}'\vec{\mathbf{Z}}_{ik})^2\right] = 0$$

It follows from the same argument as given in Tian et al. (2007) that $\widehat{\boldsymbol{\gamma}}(t) = (\widehat{\boldsymbol{\gamma}}_\mu(t)', \widehat{\boldsymbol{\gamma}}_\sigma(t)')'$ is

always convergent to $\boldsymbol{\gamma}_0(t) = (\boldsymbol{\gamma}_{\mu 0}(t)', \boldsymbol{\gamma}_{\sigma 0}(t)')'$, where $\boldsymbol{\gamma}_0(t)$ is the unique solution to

$$E\begin{pmatrix} \sum_{k=1}^{K} \vec{\mathbf{Z}}_{1k} \left\{ X_{1k}(t) - g(\boldsymbol{\gamma}_\mu' \vec{\mathbf{Z}}_{1k}) \right\} \\ \sum_{k=1}^{K} \vec{\mathbf{Z}}_{1k} \left[\left\{ X_{1k}(t) - g(\boldsymbol{\gamma}_\mu' \vec{\mathbf{Z}}_{1k}) \right\}^2 - h(\boldsymbol{\gamma}_\sigma' \vec{\mathbf{Z}}_{1k})^2 \right] \end{pmatrix} = 0,$$

regardless of the adequacy of the working model (1). $\boldsymbol{\gamma}_{\mu 0}(t)$ and $\boldsymbol{\gamma}_{\sigma 0}(t)$ are the respective true values of $\boldsymbol{\gamma}_\mu(t)$ and $\boldsymbol{\gamma}_\sigma(t)$ in (1) when it holds; and is a valid summary of the average effect of $\mathbf{Z}$ on $X(t)$ even when the working model (1) fails to hold. We next develop inference procedures about the average effects without requiring model (1) to hold.

To make inference about $\boldsymbol{\gamma}_0(t)$ at a given location t, we note that $n^{\frac{1}{2}}\{\widehat{\boldsymbol{\gamma}}(t) - \boldsymbol{\gamma}_0(t)\}$ is asymptotically equivalent to $n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{U}_i(t)$ and converges in distribution to a multivariate normal with mean 0 and covariance matrix $\boldsymbol{\Sigma}_\gamma(t) = E\{\mathbf{U}_i(t)^{\otimes 2}\}$, where $\mathbf{U}_i(t) = [\mathbf{U}_{\mu i}(t)', \mathbf{U}_{\sigma i}(t)']'$,

$$\mathbf{U}_{\mu i}(t) = \sum_{k=1}^{K} \mathbb{A}_\mu^{-1}(t) \vec{\mathbf{Z}}_{ik} \left\{ X_{ik}(t) - g(\boldsymbol{\gamma}_{\mu 0}(t)' \vec{\mathbf{Z}}_{ik})) \right\}$$

$$\mathbf{U}_{\sigma i}(t) = \sum_{k=1}^{K} \mathbb{A}_\sigma^{-1}(t) \left(\mathbb{A}_{\mu\sigma}(t) \mathbf{U}_{\mu i}(t) + \vec{\mathbf{Z}}_{ik} \left[\left\{ X_{ik}(t) - g(\boldsymbol{\gamma}_{\mu 0}(t)' \vec{\mathbf{Z}}_{ik}) \right\}^2 - h(\boldsymbol{\gamma}_{\sigma 0}(t)' \vec{\mathbf{Z}}_{ik})^2 \right] \right)$$

$\mathbb{A}_\mu(t) = \sum_{k=1}^{K} E\{\dot{g}(\boldsymbol{\gamma}_{\mu 0}(t)' \vec{\mathbf{Z}}_{1k}) \vec{\mathbf{Z}}_{1k}^{\otimes 2}\}$, $\mathbb{A}_\sigma(t) = 2 \sum_{k=1}^{K} E\{\dot{h}(\boldsymbol{\gamma}_{\sigma 0}(t)' \vec{\mathbf{Z}}_{1k}) h(\boldsymbol{\gamma}_{\sigma 0}(t)' \vec{\mathbf{Z}}_{1k}) \vec{\mathbf{Z}}_{1k}^{\otimes 2}\}$, $\mathbb{A}_{\mu\sigma}(t) = 2 \sum_{k=1}^{K} E[\{X_{1k}(t) - g(\boldsymbol{\gamma}_{\mu 0}(t)' \vec{\mathbf{Z}}_{1k})\} \dot{g}(\boldsymbol{\gamma}_{\mu 0}(t)' \vec{\mathbf{Z}}_{1k}) \vec{\mathbf{Z}}_{1k}^{\otimes 2}]$, and for any vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}'$. For any specific location t, one may construct confidence intervals for $\boldsymbol{\gamma}_{\mu 0}(t)$ and $\boldsymbol{\gamma}_{\sigma 0}(t)$ based on this large sample normal approximation.

In practice, it is often of interest to simultaneously assess the population level instability across a region of t. Unfortunately, it is unclear whether $n^{\frac{1}{2}}\{\widehat{\boldsymbol{\gamma}}(t) - \boldsymbol{\gamma}_0(t)\}$ converges as a process in t due to the unknown correlation structure between $X(t)$ and $X(s)$. As a result, $\sup_{t \in \mathcal{T}} |n^{\frac{1}{2}}\{\widehat{\boldsymbol{\gamma}}(t) - \boldsymbol{\gamma}_0(t)\}|$ may not converge in distribution and thus it is difficult to construct confidence bands

for $\boldsymbol{\gamma}_0(t)$. To overcome such a difficulty, we propose to take the moving average approach by assessing the *average* genetic stability across various small regions of interest. Specifically, we consider to make joint inference about

$$\boldsymbol{\Gamma}_0(t) = \frac{1}{2b_0}\int_{t-b_0}^{t+b_0} \boldsymbol{\gamma}_0(s)ds, \quad \text{for } t \in [b_0, \tau - b_0]\,,$$

where b_0 is a pre-specified positive constant half window width. $\boldsymbol{\Gamma}_0(t)$ is essentially a smoothed version of $\boldsymbol{\gamma}_0(t)$ and summarizes the average covariate effect on the genetic instability in the region $s \in [t - b_0, t + b_0]$. Based on $\widehat{\boldsymbol{\gamma}}(t)$, one may obtain a simple plug-in estimate for $\boldsymbol{\Gamma}_0(t)$,

$$\widehat{\boldsymbol{\Gamma}}\,(t) = \frac{1}{2b_0}\int_{t-b_0}^{t+b_0} \widehat{\boldsymbol{\gamma}}(s)ds \tag{2}$$

We show in the appendix that under mild regularity conditions, the process $\widehat{\mathcal{W}}(t) = n^{\frac{1}{2}}\{\widehat{\boldsymbol{\Gamma}}(t) - \boldsymbol{\Gamma}_0(t)\}$ converges weakly in to a zero-mean Gaussian process, $\mathcal{W}(t)$.

To approximate the distribution of $\widehat{\mathcal{W}}(t)$, we propose to use the simple perturbation resampling method similar to what has been considered in Cai et al. (2000) and Park et al. (2003). Let $\mathcal{N}_{1...n} = \{\mathcal{N}_i, i = 1, ..., n\}$ denote n i.i.d copies of standard normal random variables generated independent of the data. For any given set of $\mathcal{N}_{1...n}$, let

$$\widehat{\mathcal{W}}^*(t) = \frac{n^{\frac{1}{2}}}{2b_0}\sum_{i=1}^{n}\int_{t-b_0}^{t+b_0} \widehat{\mathbf{U}}_i(t)\mathcal{N}_i$$

where $\widehat{\mathbf{U}}_i(t)$ is obtained by replacing all the theoretical quantities in $\mathbf{U}_i(t)$ by their empirical counterparts. It is not difficult to show that conditional on the data, $\widehat{\mathcal{W}}^*(t)$ converges weakly to the Gaussian process $\mathcal{W}(t)$. Therefore, one may approximate the distribution of $\widehat{\mathcal{W}}(t)$ by the conditional distribution of $\widehat{\mathcal{W}}^*(t)$.

In practice we can approximate the distribution of $\widehat{\mathcal{W}}(t)$ by generating a large number, M say, independent samples of $\{\mathcal{N}_i, i = 1 \ldots, n\}$. For the mth sample, we obtain a realization $\widehat{\mathcal{W}}^*_{(m)}(t)$ of $\widehat{\mathcal{W}}^*(t)$, $m = 1, \ldots, M$. At any given location t and for any constant vector $\mathbf{c}$, we may construct $100(1-\alpha)\%$ point-wise confidence interval for $\mathbf{c}'\mathbf{\Gamma}_0(t)$ as

$$\mathbf{c}'\widehat{\Gamma}(t) \pm z_\alpha \widehat{\sigma}(t), \qquad \text{where} \quad \widehat{\sigma}^2(t) = \frac{1}{M}\sum_{m=1}^{M}\left\{\boldsymbol{a}'\widehat{\mathcal{W}}^*_{(m)}(t)\right\}^2$$

and z_α is the $100(1-\alpha/2)$th percentile of the standard normal. Furthermore, $100(1-\alpha)\%$ simultaneous confidence interval for $\{\boldsymbol{a}'\mathbf{\Gamma}_0(t), b_0 \le t \le \tau - b_0\}$ may be obtained as

$$\mathbf{c}'\widehat{\Gamma}(t) \pm s_\alpha \widehat{\sigma}(t),$$

where s_α is the $100(1-\alpha)$th percentile of

$$\left\{\sup_{t\in[b_0,\tau-b_0]}|\boldsymbol{a}'\widehat{\mathcal{W}}^*_{(1)}(t)/\widehat{\sigma}(t)|, \cdots, \sup_{t\in[b_0,\tau-b_0]}\left|\boldsymbol{a}'\widehat{\mathcal{W}}^*_{(M)}(t)/\widehat{\sigma}(t)\right|\right\}$$

3 Examples

3.1 Meningioma Study

We now evaluate the population level genetic instability based on an aCGH dataset from the Meningioma study, conducted at the Massachusetts General Hospital. Meningiomas are brain tumors that represents approximately 15% of all primary brain tumors. Most of these tumors are benign but it is not unusual for them to be malignant. The dataset consists of 72 subjects classified as sporadic solitary meningioma and profiles their copy number $\log_2$ ratios at various locations (or clones) across chromosomes 1 to 23. The pathological subclassification lists 34

subjects as being benign, 25 atypical and 13 malignant cases. Goals of this study include the development of a useful graphical method of displaying the data, the identification of regions of loss or gain that are shared by patients with this type of tumor as well as to assess differences between the pathological subtypes.

First we evaluate the population level genetic instability, based on the mean $\log_2$aCGH ratio level, on chromosome 1 for the *benign* tumor subtype. The $\log_2$aCGH ratio at a total of 1339 locations were recorded for chromosome 1. The population mean of the $\log_2$aCGH ratio, $\gamma_\mu(t) = E\{X_i(t)\}$, can be estimated empirically by $\widehat{\gamma}_\mu(t) = n^{-1}\sum_{i=1}^{n} X_i(t)$. Here, each subject has one observation with $K = 1$ and for simplicity, we drop the subscript k. The raw mean process is shown in Figure 1(a). We propose to quantify the population level genetic instability based on $\Gamma_\mu(t) = (2b_0)^{-1}\int_{t-b_0}^{t+b_0}\gamma_\mu(s)ds$ which can be estimated by $\widehat{\Gamma}_\mu(t) = (2b_0)^{-1}\int_{t-b_0}^{t+b_0}\widehat{\gamma}_\mu(s)ds$. In Figure 1(b), we show the estimated process $\{\widehat{\Gamma}_\mu(t);\, b_0 \le t \le \tau - b_0\}$ along with their 95% simultaneous confidence bands. In this and other figures in this example, the shaded regions on the upper level and the lower level represent the regions that the population version of the displayed process is significantly greater or less than 0 (respectively). In this and subsequent analyses, for the moving averages, we choose a priori, overlapping fixed windows of size $b_0 = 10$ and 95% confidence bands are based on perturbing the original process $M = 500$ times. Since the simultaneous confidence intervals cover 0 for most of the regions, we conclude that there appears to be little genetic instability at chromosome 1 for patients with benign tumor.

Next we compare the population average genetic instability at chromosome 1 among the three

tumor subtypes. To this end, we consider the simple regression model

$$E\{X_i(t)\,|Z_{i(1)}, Z_{i(2)}\} = \beta_0(t) + \beta_1(t)\,Z_{i(1)} + \beta_2(t)\,Z_{i(2)},$$

where $Z_{i(1)} = 1$ if subject i has an atypical tumor and 0 otherwise, and $Z_{i(2)} = 1$ if subject i has a malignant tumor and 0 otherwise. The reference group, corresponding to $Z_{i(1)} = Z_{i(2)} = 0$, consists of the patients classified as 'benign'. Figure 2 presents the point and simultaneous interval estimates of $\Gamma_{\beta_1}(t) = (2b_0)^{-1}\int_{t-b_0}^{t+b_0}\beta_1(s)ds$ and $\Gamma_{\beta_2}(t) = (2b_0)^{-1}\int_{t-b_0}^{t+b_0}\beta_2(s)ds$.

The results suggest that on average, the level of genetic instability at chromosome 1 is higher for atypical tumor, and is even more extreme for malignant tumor, when compared to benign tumor. For both atypical and malignant tumors, there appears to be regions of deletion at locations $0 \sim 700$ and regions of amplification at locations $700 \sim 1200$.

To assess the genetic instability across the entire genome, we obtained point estimates of $\Gamma_{\beta_1}(\cdot)$ and $\Gamma_{\beta_2}(\cdot)$ for all 23 chromosomes and obtained simultaneous confidence intervals adjusting for all the locations considered. Note that the moving average is only applied within chromosomes. The left panel in Figure 3 shows the moving average process related to $\Gamma_{\beta_1}(\cdot)$ and the corresponding graph relating to $\Gamma_{\beta_2}(\cdot)$ are displayed in the right panel. In general, there appears to be differential level of genetic instability for the three tumor types. The difference in genetic instability between the benign tumor and the atypical tumor is most significant at chromosomes 1, 14 and 23. The difference is also apparent in non-trivial regions at chromosomes 2, 5, 6, 7, 11, 15, 16, 17, 19, 20 and 22. The difference between the benign tumor and malignant tumor has a similar pattern, but is more extreme except for chromosome 22. In addition, the difference is apparent at chromosomes 8, 9, 10, 12 and 18.

3.2 Glioma Study

The second example is from a glioma study (Mohapatra et al., 2006) which consists of 47 patients diagnosed with two subtypes of glimoas: 22 were diagnosed with oligodendroglioma (OLIGO) and 25 were diagnosed with glioblastoma multiforme (GBM). Gliomas are the most commonly diagnosed primary brain tumors, accounting for approximately 45-50% of all primary brain tumors. Among high-grade gliomas, OLIGOs have a more favorable prognosis than GBMs (Kleihues and Cavenee, 2000). GBMs are resistant to most available therapies, while OLIGOs are often chemosensitive. Currently, gliomas are classified according to defined histological features characteristic of the presumed normal cell of origin. Such methods, however, may lack diagnostic accuracy and reproducibility in many cases (Nutt et al., 2003). To develop more objective approaches to glioma classification, recent investigations have focused on molecular genetic analyses (e.g. Sasaki et al. , Burger et al. 11). We are interested in comparing genetic profiles between these two sub-types of gliomas based on the aCGH data, available on chromosomes 1,7 and 19, for subjects in this study. Until recently, aCGH could only be reproducibly performed on frozen tissue samples and with significant tissue amounts. For brain tumors however, paraffin-embedded tissue blocks from small stereotactic biopsies may be the only tissue routinely available. The development of methods to analyze formalin-fixed, paraffin-embedded material therefore has the potential to impact molecular diagnosis in a significant way. For each subject in this study, both frozen tissues and FFPE tissues are available. We are also interested in comparing results of copy number $\log_2$ ratio measurements between frozen and FFPE materials. To perform such

analyses, we consider the model

$$X_{ik}(t) = \beta_0(t) + \beta_1(t)Z_{ik(1)} + \beta_2(t)Z_{ik(2)} + \exp\left\{\alpha_0(t) + \alpha_1(t)Z_{ik(1)} + \alpha_2(t)Z_{ik(2)}\right\}\varepsilon_{ik} \quad (3)$$

for $k = 1, 2$ and $n = 1, ..., 47$. Here, $X_{ik}(t)$ is the $\log_2$ ratio of the copy numbers at location t obtained using the kth tissue type of the ith subject i, $Z_{ik(1)}$ is the tissue type indicator with $Z_{ik(1)} = 1$ for FFPE tissues and 0 otherwise, $Z_{ik(2)}$ is the indicator of tumor sub-type with $Z_{ik(2)} = 1$ for GBM tumors and 0 otherwise. Estimates of $\boldsymbol{\beta}(t)$ and $\boldsymbol{\alpha}(t)$ in model (3) are obtained using data on each of the three chromosomes. These are used to generate point estimates and simultaneous interval estimates for the processes $\Gamma_{\beta_1}(t) = (2b_0)^{-1}\int_{t-b_0}^{t+b_0}\beta_1(s)ds$, $\Gamma_{\beta_2}(t) = (2b_0)^{-1}\int_{t-b_0}^{t+b_0}\beta_2(s)ds$, $\Gamma_{\exp(\alpha_1)}(t) = (2b_0)^{-1}\int_{t-b_0}^{t+b_0}\exp(\alpha_1(s))ds$ and $\Gamma_{\exp(\alpha_2)}(t) = (2b_0)^{-1}\int_{t-b_0}^{t+b_0}\exp(\alpha_2(s))ds$. These processes capture the local average covariate effects on the population mean, and variance genetic instability processes. The results are shown in Figure 4. In (a) and (c) of Figure 4, the shaded regions on the upper level and the lower level represent the regions that the processes $\Gamma_{\beta_1}(\cdot)$ $\Gamma_{\beta_2}(\cdot)$ are significantly greater or less than 0 (respectively), whereas in (b) and (d), the shaded regions represent regions where the processes $\Gamma_{\exp(\alpha_1)}(\cdot)$ and $\Gamma_{\exp(\alpha_2)}(\cdot)$ are significantly greater or less than 1, respectively. We observe that compared to the frozen prepared sample, on average, the paraffin prepared sample tend to have a greater mean and variance and the difference in mean effects is significant in certain regions on all three chromosomes. Comparing gbm to oligo, the results suggest that the gbm has a greater effect on the population mean genetic instability over the three chromosomes achieving statistical significance in regions of chromosomes 1 and 19 and in a spiked region on chromosome 7. The effect of tumor subtype on the variability seem to vacillate across the chromosomes.

4 Discussion

We developed a new procedure for assessing the population level genetic instability and the effect of covariates on the genetic instability with array CGH data. The use of the moving average technique allows for making simultaneous assessment across the entire genome without requiring to specify the unknown complex correlation structure between regions. When assessing the association between covariates and the population level genetic instability, we proposed the use of *working models* to approximate the association. However, our procedures for making inference about the average covariate effect are valid even if the fitted working models fail to hold.

To ensure the validity of the simultaneous inference, the half window width parameter b_0 is a pre-determined constant. For our analyses we have chosen b_0 to be set at 10 and for array CGH data we find that this is rather informative and allows the moving average process to somewhat resemble the original (raw) process with sufficient smoothness to ensure the validity of the simultaneous inference about the moving average process.

5 Appendix

Let $t \in \mathcal{T} = [0, \tau]$ be such that $n^{1/2}\{\widehat{\boldsymbol{\gamma}}(t) - \boldsymbol{\gamma}_0(t)\}$ is asymptotically equivalent to a sum of independent and identically distributed terms, $n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{U}_i(t)$. Here, we assume that $\mathbf{U}_i(t)$ has total variation bounded by a constant. It is not difficult to see that under the location-scale working model and with the assumption that $X_{ik}(t)$ has total variation bounded by a constant, this condition is satisfied. In this section, we show that for $b_0 > 0$, and $t \in \mathcal{T}' = [b_0, \tau - b_0]$, the process

$$\widehat{\mathcal{W}}(t) = n^{1/2}\{\widehat{\boldsymbol{\Gamma}}(t) - \boldsymbol{\Gamma}_0(t)\}$$

where $\boldsymbol{\Gamma}_0(t) = (2b_0)^{-1} \int_{t-b_0}^{t+b_0} \boldsymbol{\gamma}_0(s)d(s)$, $\widehat{\boldsymbol{\Gamma}}(t) = \frac{1}{2b_0} \int_{t-b_0}^{t+b_0} \widehat{\boldsymbol{\gamma}}(s)ds$ converges weakly to a zero mean Gaussian process $\mathcal{W}(t)$ with continuous sample paths and covariance matrix function $\Upsilon(s,t) = E\{\boldsymbol{L}_1(t)\boldsymbol{L}_1(s)'\}$, where $\boldsymbol{L}_1(r) = (2b_0)^{-1} \int_{r-b_0}^{r+b_0} \boldsymbol{U}_1(u)d(u)$ and $s, t \in \mathcal{T}'$.

To this end, We first note that

$$\widehat{\mathcal{W}}(t) = n^{1/2}\{\widehat{\boldsymbol{\Gamma}}(t) - \boldsymbol{\Gamma}_0(t)\} = \frac{n^{1/2}}{2b_0} \int_{t-b_0}^{t+b_0} \{\widehat{\boldsymbol{\gamma}}(s) - \boldsymbol{\gamma}_0(s)\}d(s)$$

which is asymptotically equivalent to $n^{-1/2}(2b_0)^{-1} \sum_{i=1}^{n} \int_{t-b_0}^{t+b_0} \boldsymbol{U}_i(s)d(s)$. For brevity, let $\boldsymbol{L}_i(t) = \int_{t-b_0}^{t+b_0} \boldsymbol{B}_i(s)d(s)$ where $\boldsymbol{B}_i = 2b_0^{-1}\boldsymbol{U}_i$ so that $\widehat{\mathcal{W}}(t)$ is asymptotically equivalent to $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{L}_i(t)$ for $t \in \mathcal{T}'$. To show the desired convergence, we need to verify the conditions (i) - (v) of the Functional Central Limit Theorem of Pollard (1990, Ch.10). Since there exists M such that $||\boldsymbol{B}_i(t)|| \leq M$ for all $i = 1, \cdots, n$ and $t \in [0, \tau]$, we can rewrite $\boldsymbol{B}_i(t)$ as $\boldsymbol{B}_i^+(t) - \boldsymbol{B}_i^-(t)$ where both $\boldsymbol{B}_i^+(t)$ and $\boldsymbol{B}_i^-(t)$ are positive functions bounded by M. Writing $\boldsymbol{L}_i(t) = \int_0^{t+b_0} \boldsymbol{B}_i^+(u)du - \int_0^{t-b_0} \boldsymbol{B}_i^+(u)du - \int_0^{t+b_0} \boldsymbol{B}_i^-(u)du + \int_0^{t-b_0} \boldsymbol{B}_i^-(u)du$, each term has pseudo-dimension at most 1

with the same envelope $M\tau$. Therefore $\{\boldsymbol{L}_i(t), i = 1, \cdots, n\}$ is Euclidean and hence manageable. Condition (ii), the existence of limiting variance covariance matrix follows from law of large numbers. It follows from law of large numbers that the limiting covariance matrix of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{L}_i(t)$ converges to $E\{\boldsymbol{L}_1(t)\boldsymbol{L}_1(s)'\}$. Since $||\boldsymbol{B}_i(t)|| \leq M$, we can choose the envelope $F_{ni} = 2Mb_0$ for $\boldsymbol{L}_i(t)$. It is then obvious that conditions (iii) and (iv)

$$\lim_{n\to\infty} \sum_{i=1}^{n} E(F_{ni}^2) < \infty\,, \quad \sum_{i=1}^{n} E(F_{ni}^2; F_{ni} > \epsilon) \to 0\,,$$

hold. Condition (v) is trivial since $\rho_n(s,t) \equiv \rho(s,t)$. Therefore $\widehat{\mathcal{W}}(t) = n^{1/2}\{\widehat{\boldsymbol{\Gamma}}(t) - \boldsymbol{\Gamma}_0(t)\}$ converges in distribution to the same mean zero Gaussian process with covariance matrix function $\Upsilon(s,t) = E\left\{(2b_0)^{-2} \int_{t-b_0}^{t+b_0} \boldsymbol{U}_1(u)du \int_{s-b_0}^{s+b_0} \boldsymbol{U}_1(u)'du\right\}$.

6 References

Aguirre, A., Brennan, C., Bailey, G., Sinha, R., Feng, B., Leo, C., Zhang, Y., Zhang, J., Gans, J., Bardeesy, N., Cauwels, C., Cordon-Cardo, C., Redston, M., Depinho, R., AND Chin, L. (2004). High-resolution characterization of the pancreatic adenocarcinoma genome. *PNAS* **24**, 9067–9072.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Ser. B*, **57**, 289–300.

Cai, T., Wei, L.J., and Wilcox, M. (2000). Semi-Paramteric Regression Analysis for Clustered Failure Time Data. *Biometrika*, **87**, 867–878.

Engler, D. A., Mohaptra, G., Louis, D. N. and Betensky, R. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, **7(3)**, 399–421

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

Guha, S., Li, Y. and Neuberg, D. (2006). Bayesian hidden markov modeling of array cgh data. *Harvard University Biostatistics Working Paper Series, 24*.

Heiskanen, M. A., Bittner, M. L., Chen, Y., Khan, J., Adler, K. E., Trent, J. M., and Meltzer, P. S. (2000). Detection of gene amplification by genomic hybridization to cdna microarrays. *Cancer Res*, **60(4)**, 799–802.

Hodgson, G., Hager, J. H., Volik, S., Hariono, S., Wernick, M., Moore, D., Albertson, D. G., Pinkel, D., Collins, C., Hanahan, D., and Gray, J. W. (2001). Genome scanning with array CGH deliniates regional alternatives in mouse islet carcinomas. *Nat. Genet.* **29**, 459–464.

Hupe, P., Stransky, N., Thiery, J. P., Radvanyi, F., Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413–3422.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y.,

Scherer, S. W., Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951.

Lin, D.Y., Wei, L.J. and Ying, Z. (2002). Model-checking techniques based on cumulative residuals, *Biometrics* **58**, 1-12.

Misra, A., Pellarin, M., Nigro, J., Smirnov, I., Moore, D., Lamborn, K. R., Pinkel, D., Albertson, D. G., and Feuerstein, B. G. (2005). Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. *Clin Cancer Res*, **11(8)**, 2907–2918.

Marioni, J. C., Thorne, N. P., and Tavar ´ e, S. (2006). Biohmm: A heterogeneous hidden markov model for segmenting array cgh data. *Bioinformatics*, **22(9)**, 1144–1146.

Okada, Y., Hurwitz, E. E., Esposito, J. M., Brower, M. A., Nutt, C. L., Louis, D. N. (2003). Selection pressures of TP53 mutation and microenvironmental location in uence epidermal growth factor receptor gene amplication in human glioblastomas. *Cancer Res.* **63**, 413–416.

Olshen, A. B., Venkatraman, E. S., Lucito, R., Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **4**, 557–572.

Park, Y. and Wei, L. J. (2003). Estimating Subject-Specific Survival Functions under the Accelerated Failure Time Model. *Biometrika* **90**, 717–723.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J. J. (2005). A statistical approach for array cgh data analysis. *BMC Bioinformatics*, **6**, 27.

Pinkel, D., and D. G. Albertson, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nat Genet*, **37 Suppl**, S11–S17.

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

Pollack, J. R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Borresen-Dale, A., and Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alternation in the transcriptional program of human breast tumors. *PNAS* **99**, 12963–12968.

Pollard, D. (1990). Empirical Processes: Theory and Applications. *Hayward, CA: Institute of Mathematical Statistics*

Rueda, O. M. and Diaz-Uriarte, R. (2006). A flexible statistical method for detecting genomic copy-number changes using Hidden Markov Models with reversible jump MCMC. *COBRA Preprint Series*

Shah, S. P., Xuan, X., Deleeuw, R. J., Khojasteh, M., Lam, W. L., Ng, R., and K. P. Murphy, K. P. (2006). Integrating copy number polymorphisms into array cgh analysis using a robust hmm. *Bioinformatics*, **22(14)**, e431–e439.

Tian, L., Cai, T., Goetghebeur, E. and Wei, L. J. (2007). Model Evaluation Based on the Distribution of Estimated Absolute Prediction Error. *http://www.bepress.com/harvardbiostat/paper3*

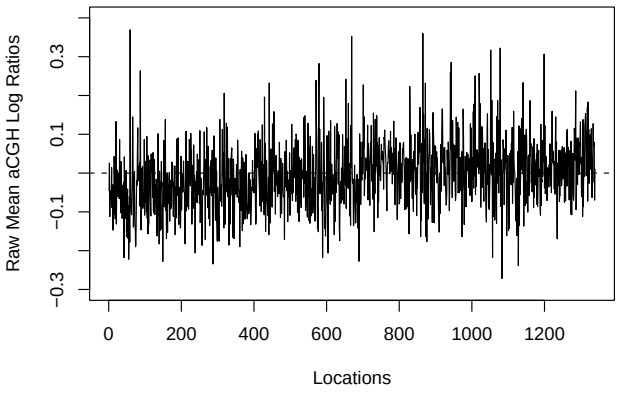
Veltman, J. A., Fridlyand, J., Pejavar, S., Olshen, A. B., JKorkola, J. E., DeVries, S., Carroll, P., Kuo, W. L., Pinkel, D., Albertson, D., Cordon-Cardo, C., Jain, A. N., and F. M. Waldman, F. M. (2003). Array-based comparative genomic hybridization for genome-wide screening of dna copy number in bladder tumors. *Cancer Res*, **63(11)**, 2872–2880.

Wang, P., Young, K., Pollack, J., Narasimham, B., AND Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics* **6**, 45–58.

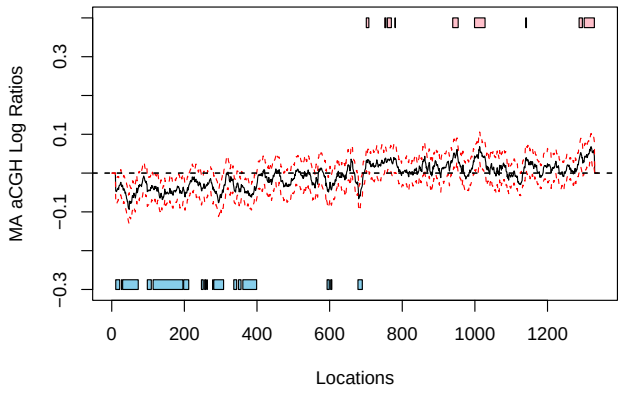
Weiss, M. M., Snijders, A. M., Kuipers, E. J., Ylstra, B., Pinkel, D., Meuwissen, S. G. M., Van Diest, P. J., Albertson, D. G., AND Meijer, G. A. (2003). Determination of amplicon boundaries at 20q13.2 in tissue samples of human gastric adenocarcinomas by high-resolution microarray comparative genomic hybridization. *The Journal of Pathology* **200**, 320–326.

Willenbrock, H. and Fridlyand, J. (2005). A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.

Figure 1: The Raw and moving average mean $\log_2$aCGH ratio levels (solid curve) and 95% simultaneous confidence bands (dashed lines) on Chromosome 1 for the benign tumor patients.

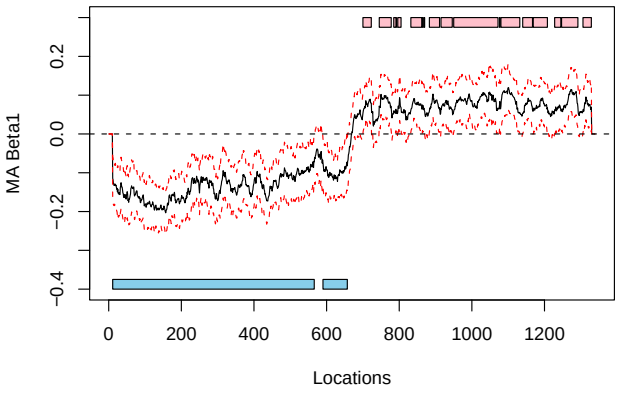


(a) The $\widehat{\gamma}_{\mu}(t)$ (raw mean) process

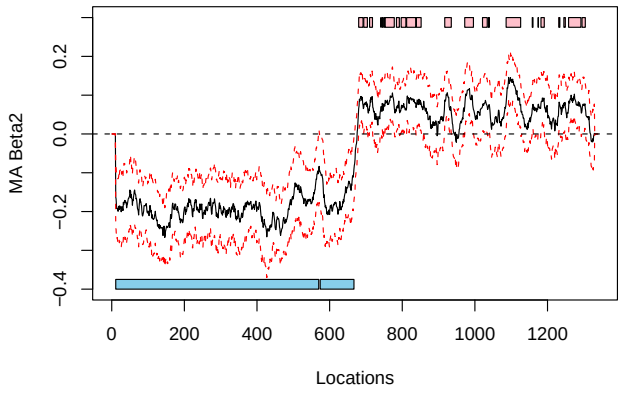


(b) The $\widehat{\Gamma}_{\mu}(t)$ process with 95% confidence bands

Figure 2: The moving average mean difference in $\log_2$aCGH ratio levels (solid curve) with 95% simultaneous confidence bands (dashed lines) at Chromosome 1 (a) between 'atypical' and benign; and (b) between 'malignant' and benign.



(a) $\widehat{\Gamma}_{\beta_1}(\cdot)$



(b) $\widehat{\Gamma}_{\beta_2}(\cdot)$

Figure 3: The moving average mean difference in $\log_2$aCGH ratio levels (solid curve) and 95% simultaneous confidence bands (dashed lines) across all chromosomes (a) between atypical and benign; and (b) between malignant and benign.

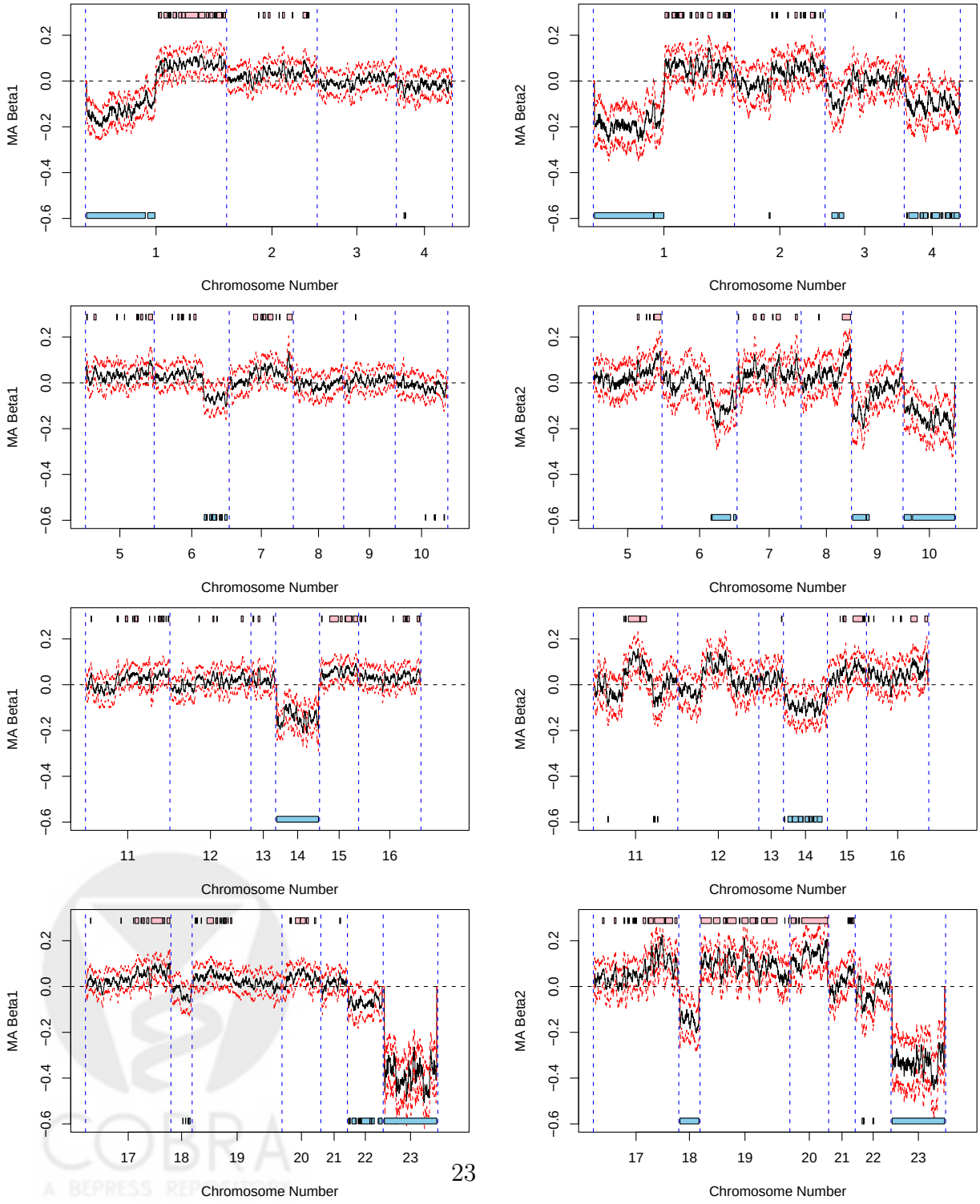
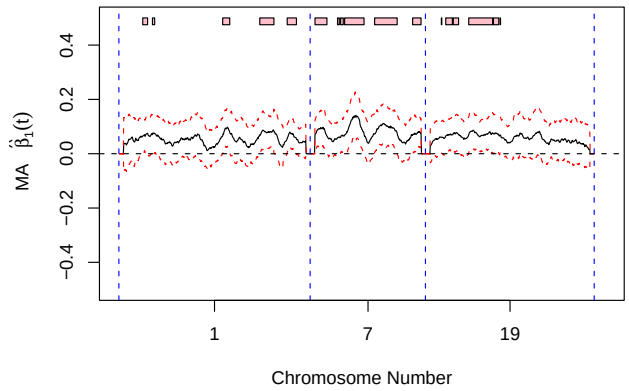
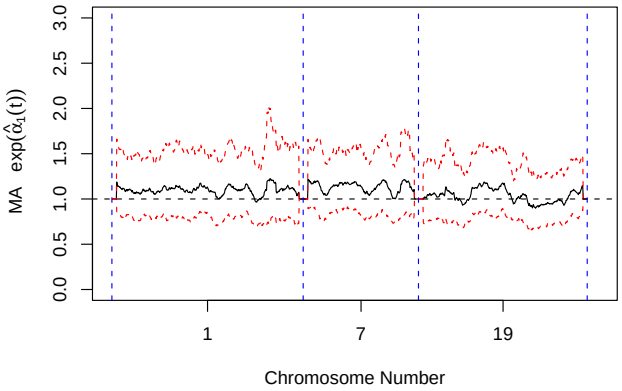


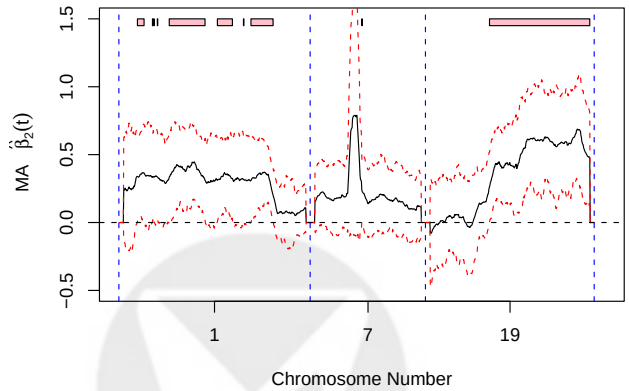
Figure 4: The moving average processes (solid curve) and 95% simultaneous confidence bands (dashed lines) showing local average covariate effects on (a) the population mean genetic instability process and (b) the population variance genetic instability process across the chromosomes between frozen (froz) and formalin-fixed, paraffin-embedded (FFPE) treated methods and between tumor types glioblastoma multiforme (gbm) and oligodendroglioma (oligo).



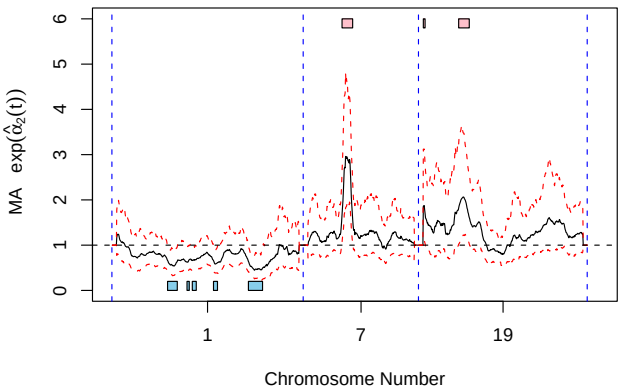
(a) $\widehat{\Gamma}_{\beta_1}(\cdot)$: FFPE (vs froz) mean effect



(b) $\widehat{\Gamma}_{\exp(\alpha_1)}(\cdot)$: FFPE (vs froz) relative variance effect



(c) $\widehat{\Gamma}_{\beta_2}(\cdot)$: gbm (vs oligo) mean effect



(d) $\widehat{\Gamma}_{\exp(\alpha_2)}(\cdot)$: gbm (vs oligo) relative variance effect