

# Non Parametric Methods for Genomic Inference

Peter J. Bickel

*Statistics, University of California, Berkeley, USA.*

Nathan Boley

*Statistics, University of California, Berkeley, USA*

James B. Brown

*Applied Science & Technology, University of California, Berkeley, USA*

Haiyan Huang

*Statistics, University of California, Berkeley, USA*

Nancy R. Zhang

*Statistics, Stanford University, USA*

**Summary.** Large-scale statistical analysis of data sets associated with genome sequences plays an important role in modern biology. A key component of such statistical analyses is the computation of p-values and confidence bounds of statistics that operate along the genome. Currently such computation is commonly achieved through ad hoc simulation measures. The method of randomization, which is at the heart of these simulation procedures, can significantly affect the resulting statistical conclusions. Most simulation schemes introduce a variety of hidden assumptions regarding the nature of the randomness in the data, resulting in a failure to capture biologically meaningful relationships. To address the need for a method of assessing the significance of observations within large scale genomic studies, where there often exists a complex dependency structure between observations, we propose a unified solution built upon a data subsampling approach. We propose a piecewise stationary model for genome sequences and show that the subsampling approach gives correct answers under this model. We illustrate the method on two simulation studies and on a real data example from the ENCODE project.

## 1. Introduction

### 1.1. Background

This paper grew out of a number of examples arising in data coming from the ENCODE Pilot Project (Birney et al., 2007), which is composed of multiple, diverse experiments performed on a targeted 1% of the human genome. Computational analyses of this data are aimed at revealing new insights about how the information coded in the DNA blueprint is turned into functioning systems in the living cell. Variations of some of the methods described here have been applied at various places in that paper, as well as in Margulies et al. (2007), for assessing significance and computing confidence bounds for statistics that operate along a genomic sequence. The background of these methods is described in cookbook form in the supplements to those papers, and it is the goal of this paper to present them rigorously and to develop some necessary refinements.

Essentially, we will argue that, in making inference about statistics computed from “large” stretches of the genome, in the absence of real knowledge about the evolutionary

path which led to the genome in question, the best we can do is to model the genome by a piecewise stationary ergodic random process. The variables of this process be base pair composition or some other local features, such as various annotated functional elements.

In the purely stationary case some of the types of questions that we will address, such as tests for independence of point processes, confidence bounds for expectations of local functions, goodness of fit of models, have been considered extensively. However, inference for piecewise stationary models appears not to have been investigated. With the advent of enormous amounts of genomic data all sorts of inferential questions have arisen. The proposed model may be the only truly nonparametric approach to the genome, although just as in ordinary nonparametric statistics there are many possible ways of carrying out inference.

Our methods are based on a development of the resampling schemes of Politis and Romano (1994), Politis, Romano, and Wolf (1999) and the block bootstrap methods of Künsch (1989). As we shall see, in many situations, Gaussian approximations can replace these schemes. But in these situations, as with the ordinary bootstrap, we believe that a block bootstrap approach is valuable for the following reasons:

- (a) Letting the computer do the approximation is much easier.
- (b) Some statistics, such as tests of the Kolmogorov Smirnov type, are functions of stochastic processes to which a joint Gaussian approximation applies. Then, limiting distributions can only be computed by simulation.
- (c) The bootstrap distributions of our statistics show us whether the approximate Gaussianity we have invoked for the “true” distribution of these statistics is in fact warranted. This visual confirmation is invaluable.

This paper is organized as follows. We begin with some concrete examples from the ENCODE data as well as other types of genomic data in Section 1.2, and proceed with a motivated description of our model in Section 2. Our methods are discussed both qualitatively and mathematically in Sections 3 and 4. Sections 5 contain results from simulation studies and real data analysis. Finally, an appendix with proofs of theorems stated in Sections 3 and 4 completes the paper.

Additionally, the statistics and methods discussed in this paper have been implemented in several computing languages and are available for download at <http://encode.dyndns.org/>. Each of these implementations runs in  $n\log(n)$  time, where  $n$  is the number of instances of the more frequent feature. On a desktop PC (Intel Core Duo 3Ghz and 2Gb RAM) the Python version takes over 1000 samples per second for features on the order of  $10^4$  instances.

## 1.2. *Motivating Examples*

We start with several fundamental questions in genomic studies.

- (a) **Association of functional elements in the human genome.** In the analysis of the human genome, a natural problem of interest is the association among the different functional sites/features annotated on the DNA sequences. Its biological motivation comes from the common belief that significant physical overlapping of functional sites in the genome suggests biological constraints or relationships. In the ENCODE project, to understand the possible functional roles of the constrained sequences that

are conserved across multiple species, overlap between the constrained sequences and several experimental annotations, such as the 5'UTR, RxFragments, pseudogenes and coding sequences (CDSs), have been evaluated using the method discussed in this paper. It was found that most experimental annotations are significantly different from a random expectation for overlaps with the constrained sequences (Birney, E. et al., 2007). An illustrative example from The ENCODE Project (Birney, E. et al., 2007) is detailed in Section 5.1.

- (b) **Cooperativity between transcription factor binding sites.** In some situations, it is interesting to study the associations between neighboring functional sites that do not necessarily overlap. For instance, it is known that transcription factors often work cooperatively and their binding sites (TFBS) tend to occur in clusters. Consequently, an effective method to identify interacting transcription factors has been to evaluate the significance of cooccurrences of their binding sites in a local genomic region (7; 8). This study has the same formulation as the above ENCODE examples given a functional site defined as follows: for a TFBS of length  $l$  at position  $i$ , we define the region  $(i - m, i + l + m)$  as a functional site. Then two overlapping functional sites is equivalent to two neighboring TFBSs with interdistance less than  $2m$ , and the methods discussed in this paper for evaluating the significance of overlapping functional features can be applied, though we leave this application which involves considering statistics of the K-S type to a later paper.
- (c) **Correlating DNA copy number with expression.** Recent technology has made it possible to assay both the DNA copy number and expression level at very fine scale along the genome. It has been known for some time that in cancer cells, regions of the genome ranging from  $< 1$  Mb to entire chromosomes may be amplified or deleted. However, to what extent does the DNA copy number change affect the expression level of the genes in the region? The global effect of copy number on expression level can be investigated by a model similar to that proposed for the ENCODE data. Specifically, the first feature could represent regions of DNA amplification, while the second feature represents regions of amplified expression. Then, the question of “does DNA copy number influence expression” can be phrased in terms of the significance of overlap between the two features. In a more detailed model, one may ask whether the DNA copy number amplification has affected the genes of a specific pathway. In this case, the second feature can be defined as regions of amplified expression for genes of that pathway.

As we have seen in these examples, a common question asked in many applications is the following: Given the position vectors of two features in the genome, e.g. “conservation between species” and “transcription start sites”, and a measure of relatedness between features, e.g. base or region percentage overlap; how significant is the observed value of the measure? How does it compare with that which might be observed “at random”?

The essential challenge in the statistical formulation of this problem is the appropriate modeling of randomness of the genome, since we observe only one of the multitudes of possible genomes that evolution might have produced for our and other species.

How have such questions been answered previously? Existing methods employ varied ways to simulate the locations of features within genomes, but all center around the uniformity assumption of the features' start positions: The features must occur homogeneously in the studied genome region, e.g. Blakesley et al (2004) and Redon et al. (2006). This assumption ignores the natural clumping of features as well as the non-stationarity of genome

sequences. Clumping of features is quite common along the genome due to either the feature's own characteristic, e.g. transcription factor binding sites (TFBSs) tend to occur in clusters, or the genome's evolutionary constraints, e.g. conserved elements are often found in dense conservation neighborhoods. Ignoring these natural properties could result in misleading conclusions.

We propose a model of the features along the genome which we view as “nonparametric” as possible. Its biological motivation is plausible, as we partially demonstrate in several real examples.

*How about changing the above paragraph to this:* In this paper, we suggest a piecewise stationary model for the genome (see section 2), and based on it, propose a method to infer the relationships between features which we view as “nonparametric” as possible (see sections 4.2 and 4.4). The biological motivations are plausible, as we partially demonstrate in several real examples later.

## 2. The Piecewise Stationary Model

We postulate the following for the observed genomes or stretches of genomes:

- (a) They can be thought of as a concatenation of number of regions, each of which is homogenous in a way we describe below.
- (b) Features that are located very far from each other on the average have little to do with each other.
- (c) The number of such homogeneous regions is small compared to the total length of the observed genomes that we consider.

These assumptions are motivated by earlier studies of DNA sequences, which show that there are global shifts in base composition, but that certain sequence characteristics are locally un-changing. One such characteristic is the GC content. Bernardi et al. (1985) coined the term “isochore” to denote large segments (of length greater than 300 Kb) that have fairly homogeneous base composition, and especially, constant GC composition. Even earlier evidence of segmental DNA structure can be found in chromosomal banding in polytene chromosomes in drosophila, visible through the microscope, that result from underlying physical and chemical structure. These banding patterns are stable enough to be used for the identification of chromosomes and for genetic mapping, and are physical evidence for a block stationarity model for the GC content of the genome.

The experimental evidence for segmental genome structure and the increasing availability of DNA sequence data have inspired attempts to computationally segment DNA into statistically homogeneous regions. The paper by Braun and Müller (1998) offers a review of statistical methods developed for detecting and modeling the inhomogeneity in DNA sequences. There have been many attempts to segment DNA sequences by both base composition (Fu and Curnow (1990), Churchill (1989,1992), Li et al (2002)) and chemical characteristics (Li et al. (1998)). Most of these computational studies concluded that a model that assumes block-wise stationarity gives a significantly better fit to the data than stationary models (see, for example, the conclusions of two very different studies by Fickett, Torney, and Wolf (1992) and Li et al. (1998)).

A subtle issue in the definition of “homogeneity” is the scale at which the genome is being analyzed. Inhomogeneity at the kilobase resolution, for example, might be “smoothed

out” in an analysis at the megabase level. The level of resolution is a modeling issue that must be considered carefully with the goal of the analysis in mind.

Implicit in our formulation is an “ergodic” hypothesis. We want probabilities to refer to the population of potential genomes. We assume that the statistics of the genome we have mimic those of the population of genomes. This is entirely analogous to the ergodic hypothesis that long term time averages agree with space averages for trajectories of dynamic systems.

In mathematical terms, the block stationarity model assumes that we observe a sequence of random variables  $\{X_1, \dots, X_n\}$  positioned linearly along the genomic region of interest.  $X_k, k = 1, \dots, n$ , may be base composition, or some other measurable feature. We assume that there exist integers  $\tau = \tau^{(n)} = (\tau_0, \dots, \tau_U)$ , where  $0 = \tau_0 < \tau_1 < \dots < \tau_U = n$ , such that the collections of variables,  $\{X_{\tau_i}, \dots, X_{\tau_{i+1}}\}$  are separately stationary for each  $i = 0, \dots, U - 1$ . We let  $n_i = \tau_i - \tau_{i-1}$  be the length of the  $i$ -th region, and let there be  $U$  such regions in total. For convenience, we introduce the mapping

$$\pi : \{1, \dots, n\} \rightarrow \{(i, j) : 1 \leq i \leq U, 1 \leq j \leq n_i\}$$

which relates the relabeled sequence,  $\{X_{ij} : 1 \leq i \leq U, 1 \leq j \leq n_i\}$  to the original sequence  $\{X_1, \dots, X_n\}$ . We write  $\pi = (\pi_1, \pi_2)$  with  $\pi(k) = (i, j)$  if and only if  $k = \tau_i + j$ . We will use the notations  $X_{ij}$  and  $X_k$  interchangeably throughout this paper.

For any  $k_1, k_2$ , let  $\mathcal{F}_{k_1}^{k_2}$  be the  $\sigma$ -field generated by  $X_{k_1}, \dots, X_{k_2}$ . Define  $m(k)$  to be the standard Rosenblatt mixing number (Dedecker et al, 2007),

$$m(k) = \sup\{|\mathbb{P}(AB) - \mathbb{P}(A)P(B)| : A \in \mathcal{F}_1^l, B \in \mathcal{F}_{l+k}^n, 1 \leq l \leq n - k\}.$$

Then, assumptions 1-3 stated at the beginning of this section translate to the following:

- A1. The sequence  $\{X_1, \dots, X_n\}$  is piecewise stationary. That is,  $\{X_{ij} : 1 \leq j \leq n_i\}$  is a stationary sequence for  $i = 1, \dots, U$ .
- A2. There exists constants  $c$  and  $\beta > 0$  such that  $m(k) \leq ck^{-\beta}$  for all  $k$ .
- A3.  $U/n \rightarrow 0$ .

An immediate and important consequence of A1-A3 is that for any fixed small  $k$ , if we define  $W_1 = (X_1, \dots, X_k), W_2 = (X_{k+1}, \dots, X_{2k}), \dots, W_m = (X_{n-k+1}, \dots, X_n)$ , where  $m = n/k$ , then  $\{W_1, \dots, W_m\}$  also obey A1-A3. This is useful, for example, in the region overlap example considered in the next section.

The remarkable feature of these assumptions, which are more general to our knowledge than any made heretofore in this context, is that they still allow us to conduct most of the statistical inference of interest. Not surprisingly, these assumptions lead to more conservative estimates of significance than any of the previous methods.

In this paper, based on the piecewise stationary model, we propose a combined segmentation-block bootstrap method, in which segmentation parameters governing scale are chosen first and then the size of the subsample is chosen based on stability criteria. ....I prefer to move this paragraph to section 4 since we have not really mentioned block-bootstrap method yet.

### 3. Linear Statistics and Gaussian Approximation

Under the above piecewise stationary model, we consider the distribution of a class of linear statistics of interest. As an illustration, we consider the ENCODE data examples, and suppose that we are interested in base pair overlap between Feature A and Feature B. We can represent base pair overlap by defining

$$\begin{aligned} I_k &= 1 \text{ if position } k \text{ belongs to Feature A and 0 otherwise,} \\ J_k &= 1 \text{ if position } k \text{ belongs to Feature B and 0 otherwise.} \end{aligned}$$

We can then define  $X_k = I_k J_k$  to be the indicator that position  $k$  belongs to both Feature A and Feature B. Then, for the  $n = 30$  Megabases of the ENCODE regions, the mean base pair overlap is equal to

$$\bar{X} = \sum_{k=1}^n X_k / n.$$

Similarly, if we consider the raw region overlap, we can let  $X_k = I_k J_k (1 - I_{k+1} J_{k+1})$  since the boundary of a region is marked by a position  $k$  which belongs to both features followed by a position which belongs only to one or neither of the features. Then, the quantity of interest is again  $\bar{X}$ . We focus our attention on statistics that can be expressed as a function of the mean of  $\mathbf{g}(X_i)$ , where  $\mathbf{g}$  is some well behaved  $d$ -dimensional vector function to be characterized in later sections. By the flexible definition of  $\mathbf{g}$ , this encompasses a wide class of situations.

First, we consider vector linear statistics of the form  $\mathbf{T}_n(\mathbf{X}) = n^{-1} \sum_{k=1}^n \mathbf{g}(X_k)$ . We introduce the following notation:

$$E[\mathbf{T}_n] \equiv \boldsymbol{\mu} \equiv \sum_{i=1}^U f_i \boldsymbol{\mu}_i,$$

where

$$\begin{aligned} \boldsymbol{\mu}_i &\equiv E[\mathbf{g}(X_{i1})], \\ f_i &\equiv n_i / n, \end{aligned}$$

and

$$\Sigma_n \equiv \text{Var}(n^{\frac{1}{2}} \mathbf{T}_n) = \sum_{i=1}^U f_i C_i(n f_i), \quad (3.1)$$

where

$$C_i(m) = C_{i0} + 2 \sum_{e=1}^m C_{ie} \left( 1 - \frac{(e-1)}{m} \right)$$

and

$$C_{i0} \equiv \text{Varg}(X_1), \quad C_{ie} \equiv \text{Cov}(\mathbf{g}(X_{i1}), \mathbf{g}(X_{i(l+1)})) . \quad (3.2)$$

In Theorem 3.1 below, we prove asymptotic Gaussianity of  $\mathbf{T}_n$  given a few more technical assumptions:

$$\text{A4. } \frac{1}{n} \sum_{i: n_i \leq l} n_i \rightarrow 0 \text{ for all } l < \infty.$$

A5.  $\forall i, |\mathbf{g}|_\infty \leq C < \infty$ .

A6.  $0 < \varepsilon_0 \leq \|\Sigma_n\| \leq \varepsilon_0^{-1}$ , for all  $n$ , where  $\|\cdot\|$  is a matrix norm.

In particular, A4 implies that the contribution of “small regions” to the overall statistic must not be too large.

**Theorem 3.1.** *Under conditions A1-A6,*

$$n^{\frac{1}{2}} \Sigma_n^{-\frac{1}{2}} (\mathbf{T}_n - \boldsymbol{\mu}) \Rightarrow \mathcal{N}(\mathbf{0}, J) \quad (3.3)$$

where  $J$  is the  $d \times d$  identity.

If we have estimates  $\hat{\boldsymbol{\tau}}$  of  $\boldsymbol{\tau}$  which are consistent in a suitably uniform sense, then estimates of  $C_{ie}$ ,  $C_i(m)$  using  $\hat{\boldsymbol{\tau}}$  in place of  $\boldsymbol{\tau}$  are also consistent. However, simply plugging these estimates into (3.2) does not yield consistent estimates of  $\sigma^2$ , as is well known for the stationary case. Some regularization is necessary. We do not pursue this direction but prefer to approach the problem from a resampling point of view – see next section.

In many cases, the statistics of interest are not linear. For example, in the analysis of the ENCODE data a more informative statistic is the %bp overlap defined as

$$B = \frac{\bar{X}}{D}, \quad (3.4)$$

where

$$D = \sum_{k=1}^n I_k$$

is the total base count of feature  $A$ . The same applies to the % regional overlaps. A standard delta method computation shows that the standard errors of  $B$  can be approximated as follows: Let  $\mu(D)$  and  $\mu(\bar{X})$  be respectively the expectation of  $D$  and  $\bar{X}$ . Then,

$$\frac{\bar{X}}{D} - \frac{\mu(\bar{X})}{\mu(D)} \approx \frac{\bar{X} - \mu(\bar{X})}{\mu(D)} - \mu(\bar{X}) \frac{(D - \mu(D))}{\mu^2(D)},$$

and hence we can approximate  $\frac{\bar{X}}{D}$  by a Gaussian variable with mean  $\frac{\mu(\bar{X})}{\mu(D)}$  and variance

$$\sigma^2(B) \approx \frac{\sigma^2(\bar{X})}{\mu^2(D)} + \frac{\mu^2(\bar{X})}{\mu^4(D)} \sigma^2(D) - 2 \frac{\mu(\bar{X})}{\mu^3(D)} \text{cov}(\bar{X}, D), \quad (3.5)$$

where  $\sigma^2(B)$ ,  $\sigma^2(\bar{X})$ ,  $\sigma^2(D)$  are the corresponding variances and  $\text{Cov}(\bar{X}, D)$  denotes the covariance. In doing inference, we can use the approximate Gaussianity of  $B$  with  $\sigma^2(B)$  estimated using the above formula with regularized sample moments replacing the true moments.

We also note that goodness of fit or equality of population test statistics, such as Kolmogorov-Smirnov and many others, can be viewed as functions of empirical distributions, which themselves are infinite dimensional linear statistics, and we expect, but have not proved, that the methods discussed in this paper and the underlying theories apply to those cases as well.

#### 4. Subsampling Based Methods

Under the piecewise stationary model, we propose a subsampling based approach, particularly, a combined segmentation-block bootstrap method to conduct statistical inference. In the method, the segmentation parameters governing scale are chosen first and then the size of the subsample is chosen based on stability criteria. The segmentation procedure, as we discussed, is motivated by the heterogeneity of large-scale genomic sequences. The insight of the block bootstrap method is to take into account the local genomic structure, such as natural clumping of features, when conducting statistical inference. We have explicitly demonstrated the advantages of using segmentation and block bootstrap by simulation studies in section 5.

##### 4.1. The Stationary Block Bootstrap Method

###### 4.1.1. Review of Results for the Case of $U = 1$

For completeness we recall the following basic algorithm of Politis and Romano (1994) to obtain an estimate of the distribution of the statistic  $\mathbf{T}_n(X_1, \dots, X_n)$  under the assumption that the sequence  $X_1, \dots, X_n$  is stationary (i.e.,  $U = 1$ ).

**Algorithm 4.1.** a) Given  $L \ll n$  choose a number  $N$  uniformly at random from  $\{1, \dots, n - L\}$ .

b) Given the statistic  $\mathbf{T}$ , as above, compute

$$\mathbf{T}_L(X_{N+1}, \dots, X_{N+L}) \equiv \mathbf{T}_{L1}^* .$$

c) Repeat  $B$  times without replacement to obtain  $\mathbf{T}_{L1}^*, \dots, \mathbf{T}_{LB}^*$ .

d) Estimate the distribution of  $\sqrt{n}(\mathbf{T}_n - \mu)$  by the empirical distribution  $\mathcal{L}_B^*$  of

$$\left\{ \sqrt{\frac{n}{L}} [\mathbf{T}_{Lj}^* - \mathbf{T}_n(X_1, \dots, X_n)], \quad 1 \leq j \leq B \right\} .$$

By Theorem 4.2.1 of Politis, Romano and Wolf (1999),

$$\mathcal{L}_B^* \implies \mathcal{N}_d(\mathbf{0}, \Sigma) . \quad (4.1)$$

in probability if (3.3) holds and if  $\frac{L}{n} \rightarrow 0$ . As usual, convergence of  $\mathcal{L}_B^*$  in law in probability simply means that if  $\rho$  is any metric for weak convergence on  $R^d$  then  $\rho(\mathcal{L}_B^*, \mathcal{L}) \xrightarrow{P} 0$ .

Since all variables we deal with are in  $L_2$  we take  $\rho$  to be the Mallows metric,

$$\rho_M^2(F, G) = \min \{ E_P(U - V)^2 : P \text{ such that } U \sim F, V \sim G \} .$$

Useful properties of  $\rho_M$  are:

- a)  $\rho_M^2(\sum \pi_i F_i, \sum \pi_i G_i) \leq \sum \pi_i \rho_M^2(F_i, G_i)$  for all  $\pi_i \geq 0$ ,  $\sum \pi_i = 1$  and
- b) If  $F = F_1 * \dots * F_m$ ,  $G = G_1 * \dots * G_m$ , that is  $F$  and  $G$  are distributions of sums of  $m$  independent variables, then  $\rho_M^2(F, G) \leq \sum_{i=1}^m \rho_M^2(F_i, G_i)$ .

For convenience, when no confusion is possible we will write  $\rho_M(V, W)$  for  $\rho_M(F, G)$  for random variables  $V \sim F$ ,  $W \sim G$ .

4.1.2. *Performance of the block bootstrap method in the piecewise stationary model when  $U > 1$ .*

We turn to the analogue of Theorem 4.2.1 in Politis, Romano and Wolf (1999) for  $U > 1$ . We consider a vector linear statistic, for which the true distribution was described in Section 3. Here, we ask how Algorithm 4.1, which assumes stationarity, performs in this nonstationary context. In general, it does not give correct confidence bounds but is conservative, sometimes exceedingly so. We sketch the issues in Theorem 4.2 below, for simplicity letting  $g$  be the one dimensional identity function  $g(x) = x$ . Let

$$\tau^2 = U^{-1} \sum_{i=1}^U (\mu_i - \bar{\mu})^2$$

$$\bar{X}_i \equiv n_i^{-1} \sum_{j=1}^{n_i} X_{ij} \quad \bar{X} \equiv n^{-1} \sum_{k=1}^n X_k = \sum_{i=1}^U f_i \bar{X}_i$$

Also let

$$n_i^* \equiv \text{Cardinality of } S_i \equiv \{k : k \in [N, N+L], \pi_1(k) = i\}$$

and

$$\bar{X}_i^* = 1(n_i^* \neq 0) \sum_j \{X_{ij} : j \in S_i\} / n_i^*,$$

$$\bar{X}_L^* = \sum_{i=1}^U f_i^* \bar{X}_i^*, \quad \text{where } f_i^* \equiv \frac{n_i^*}{L}.$$

We introduce one assumption that is obviously needed for any analysis of the block or segmented resampling bootstraps

$$\text{A7. } L \rightarrow \infty$$

and two other assumptions that are used only in this part of the paper and are thus given a different numbering

$$\text{B1. } \frac{L}{n} \rightarrow 0$$

$$\text{B2. } \frac{LU}{n} \rightarrow 0$$

**Theorem 4.2.** *Let  $\mathcal{L}_n$  be the distribution which assigns mass  $f_i$  to  $(\mu_i - \mu)$ ,  $1 \leq i \leq U$  and write  $C_i$  for  $C_i(nf_i)$ . Suppose assumptions A1-A5, and A7 hold.*

$$(i) \text{ If B2 holds, } \rho_M(\bar{X}_L^* - \bar{X}, \mathcal{L}_n) \xrightarrow{P} 0$$

(ii) If

$$\sum_{i=1}^U f_i (\mu_i - \mu)^2 = o(L^{-1}) \tag{4.2}$$

and B1 holds, then

$$\rho_M[\sqrt{L}(\bar{X}_L^* - \bar{X}), \sum_{i=1}^U f_i \mathcal{N}(0, C_i)] \xrightarrow{P} 0$$

(iii) If (4.2) and B1 hold and

$$\sum_{i=1}^U f_i 1(|\Sigma_n - C_i| \geq \varepsilon) \rightarrow 0 \quad (4.3)$$

for all  $\varepsilon > 0$ , then

$$\rho_M(\sqrt{L}(\bar{X}_L^* - \bar{X}), \mathcal{N}(0, \Sigma_n)) \xrightarrow{P} 0.$$

The implications of Theorem 4.2 are as follows. If equation (4.2) doesn't hold then  $\bar{X}_L^* - \bar{X}$  doesn't converge in law at scale  $L^{-\frac{1}{2}}$  so that it doesn't reflect the behaviour of  $L^{\frac{1}{2}}(\bar{X}_L - \mu)$  at all. This is a consequence of inhomogeneity of the segment means. Evidently in this case, confidence intervals of the percentile type for  $\mu$ ,  $[\bar{X} + c_n(\alpha), \bar{X} + c_n(1 - \alpha)]$  where  $c_n(\alpha)$  is the  $\alpha$  quantile of the distribution of  $\bar{X}_L^* - \bar{X}$ , will have coverage probability tending to 1, since  $c_n(\alpha)$  and  $c_n(1 - \alpha)$  do not converge to 0 at rate  $L^{-\frac{1}{2}}$  as they should by Theorem 4.2. If B2 does not hold we have to consider the possibility that  $[N, N+L]$  covers  $K_N$  consecutive segments, whose total length is  $o(n)$ , such that the average over all such blocks is close to  $\mu$ . However, in the absence of a condition such as (4.2) or mutual cancellation of  $\mu_i^*$  the scale of  $\bar{X}_L^*$  will be larger than  $L^{-1/2}$ . These issues will be clarified by the proof of Theorem 4.2 in the appendix. We note also that (4.2) can be weakened to requiring that the mean of blocks of consecutive segments whose total length is small compared to  $n$  are close to  $\mu$  to order  $o(L^{-1/2})$ . But our statement makes the issues clear. Finally, note that B2 holds automatically if the number of segments  $U$  is bounded and if B1 holds.

If (4.2) does hold but (4.3) doesn't, then  $\sqrt{L}(\bar{X}_L^* - \bar{X})$  converges in law to the Gaussian mixture  $\sum_{i=1}^U f_i \mathcal{N}(0, C_i)$ .

The mixture of Gaussians is more dispersed in a rough sense than a Gaussian with the same variance, which is,

$$\sigma_n^2 = \sum_{i=1}^U f_i C_i.$$

See Andrews and Mallows(1974). Especially note that, if  $W$  has the mixture distribution and  $V$  is the Gaussian variable with the same variance, then

$$E e^{tW} = \sum f_i e^{-\frac{t^2}{2} c_i} \geq e^{-\frac{t^2}{2} \sum f_i C_i} = E e^{tV}$$

by Jensen's inequality. This suggests the same phenomenon for tail probabilities eventually.

The overdispersion here, which leads to conservativeness that is not as extreme as in case (i), is due to inequality of the variances from segment to segment. Finally, if (4.3) holds then the segments have essentially the same mean and variance and the stationary block bootstrap works.

A mark of either (4.2) or (4.3) failing is a lack of Gaussianity in the distribution of  $\bar{X}_L^* - \bar{X}$ . This was in fact observed at some scales in the ENCODE project, which led us to crudely segment on biological grounds with reasonable success. However, the correct solution, which we now present in this paper, is to estimate the segmentation and appropriately adjust the bootstrap.

## 4.2. A Segmentation Based Block Bootstrap Method

We saw in the previous section that the naïve block bootstrap method that was designed for the stationary case breaks down when the sequence follows a piecewise stationary model. We propose a stratified block bootstrap strategy, which stratifies the subsample based on a “good” segmentation of the sequence which is estimated from the data. We first state the block bootstrap method, and then in Section 4.2.3 give minimal conditions on the estimated segmentation for consistency of the block bootstrap method. In Section 4.3 we discuss possible segmentation methods.

### 4.2.1. Description of Algorithm

Assume that we are given a segmentation  $\mathbf{t} = (0 = t_0, t_1, \dots, t_{m+1} = n)$ , where  $m$  is the number of regions in  $\mathbf{t}$ . Assume that the total size  $L$  of the subsample is pre-chosen. We define a stratified block bootstrap scheme as follows.

**Algorithm 4.3.** For  $i = 1, \dots, m$ , let  $\lambda_i = \lambda_i(t) = \lceil (t_i - t_{i-1})L/n \rceil$ . We use the notation  $X_{i;l}$  to denote the block of length  $l$  starting at  $i$ :

$$X_{i;l} = (X_i, \dots, X_{i+l-1}).$$

Then, for each subsample,

Draw integers  $\mathbf{N} = \{N_1, \dots, N_m\}$ , with  $N_i$  chosen uniformly from  $\{(t_{i-1} + 1, \dots, t_i - \lambda_i(t) + 1)\}$ , and let

$$X^* = (X_1^*, \dots, X_m^*) = (X_{N_1; \lambda_1(t)}, \dots, X_{N_m; \lambda_m(t)}).$$

Repeat the above  $B$  times to obtain  $B$  subsamples:  $X^{*,1}, \dots, X^{*,B}$ .

To obtain a confidence interval for  $\boldsymbol{\mu}$ , we assume that  $T_n$  has approximately a  $N(\boldsymbol{\mu}, \Sigma_n/n)$  distribution as in the previous section. For each subsample drawn as described in Algorithm 4.3, compute the statistic  $T_L^{*,b} = T_L^{*,b}(\mathbf{t}) = T_L(X^{*,b})$ . Form the sampling estimate of variance,

$$\hat{\Sigma}_n \equiv \frac{L}{B} \sum_{b=1}^B (T_L^{*,b} - \bar{T}_L^*)' (T_L^{*,b} - \bar{T}_L^*), \quad (4.4)$$

where  $\bar{T}_L^* \equiv \sum_{b=1}^B T_L^{*,b} / B$ . We can now proceed to estimate the confidence interval for  $T_n$  in a sequence of standard ways. For example, in the univariate case where  $\sigma_n^2 \equiv \Sigma_n$  is a scalar:

- Use  $\bar{X} \pm z_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}}$ , where  $z_\beta$  is the  $\beta$ th quantile of  $N(0, 1)$ , for a  $1 - \beta$  confidence interval.
- Efron’s percentile method: Let  $\bar{X}_{(1)}^* < \dots < \bar{X}_{(B)}^*$  be the ordered  $\bar{X}^{*,b}$ , then use  $[\bar{X}_{([B\alpha/2])}^*, \bar{X}_{([B(1-\alpha/2)])}^*]$  as a  $1 - \alpha$  confidence interval.
- Use a studentized interval (Efron (1981)) or Efron’s (1987) *BCA* method, see Hall (1993) for an extensive discussion.

Although the theory for (c) giving the best coverage approximation has not been written down, as it has been for the ordinary bootstrap, we expect it to continue to hold. Evidently, these approaches can be applied not only to vector linear statistics like  $T_n$  but also to smooth functions of vector linear statistics.

This algorithm assumes a given segmentation  $\mathbf{t}$ , which should be set to some good estimate  $\hat{\boldsymbol{\tau}}^{(n)} = \{0 = \hat{t}_0, \hat{t}_1, \dots, \hat{t}_m = n\}$  of the true change points  $\boldsymbol{\tau}^{(n)}$ , which estimates the true changepoints  $\boldsymbol{\tau}^{(n)}$ . In order for the algorithm to perform well, a good segmentation is critical unless the sequence is already reasonably homogeneous. In Section 4.2.3 below, we will state a few assumptions on the data determined segmentation  $\hat{\boldsymbol{\tau}}^{(n)}$  which would enable us to act as if the segmentation were known and state Theorem 4.5 to that effect.

#### 4.2.2. Consistency with True Segmentation

Under the hypothetical situation where the segmentation  $\mathbf{t}$  assumed in Algorithm 4.3 is the true segmentation, then the algorithm can be easily shown to be consistent. Here we state the result, which will be proved in Appendix.

First, we state a stronger version of the assumption *B1*, which requires that the square of the subsample size  $L = L_n$  to be  $o(n)$ :

$$\text{A8. } L_n^2/n \rightarrow 0.$$

Then, the consistency of Algorithm 4.3 given the true segmentation is given in the following theorem.

**Theorem 4.4.** *If assumptions A1-A8 hold, then*

$$L_n^{1/2} \Sigma_n^{-1/2} [T_{L_n}^*(\tau_n) - T_n] \Rightarrow N(0, J) \quad (4.5)$$

in probability.

#### 4.2.3. Consistency with Estimated Segmentation

Let  $\hat{\boldsymbol{\tau}} = \hat{\boldsymbol{\tau}}^{(n)} = (\hat{\tau}_1^{(n)}, \dots, \hat{\tau}_{\hat{U}_n}^{(n)})$  be a segmentation of the sequence  $X_1, \dots, X_n$ , which is determined from the data, and which is intended to estimate the true change-points  $\boldsymbol{\tau} = \boldsymbol{\tau}^{(n)}$ . We will state conditions on  $\hat{\boldsymbol{\tau}}$  such that the statistic obtained from Algorithm 4.3 based on  $\hat{\boldsymbol{\tau}}$  is close to the statistic obtained from the same algorithm based on the true segmentation  $\boldsymbol{\tau}$ . This can be stated formally as follows. For any segmentation  $\mathbf{t}$ , let  $\mathbf{X}^*(\mathbf{t})$  be a subsample drawn according to Algorithm 4.3 based on  $\mathbf{t}$ . Let  $F_{n,\mathbf{t}}^*(\cdot)$  be the distribution of  $\sqrt{L}\{T[\mathbf{X}^*(\mathbf{t})] - E^*T[\mathbf{X}^*(\mathbf{t})]\}$  conditioned on  $X_1, \dots, X_n$  and  $\mathbf{t}$ . Then, the desired property on the estimated segmentation  $\hat{\boldsymbol{\tau}}$  is that

$$\rho_M^2[F_{n,\hat{\boldsymbol{\tau}}^{(n)}}^*, F_{n,\boldsymbol{\tau}^{(n)}}^*] \rightarrow_p 0, \quad \text{as } n \rightarrow \infty \quad (4.6)$$

where  $\rho_M^2$  is the Mallow's metric described in Section 4.1.1. That is, for inferential purposes,  $T[\mathbf{X}^*(\hat{\boldsymbol{\tau}})]$  has approximately the same distribution as  $T[\mathbf{X}^*(\boldsymbol{\tau})]$ . Then, since we have shown in Section 4.2.2 that

$$\rho_M^2[F_{n,\boldsymbol{\tau}^{(n)}}^*, \Phi(\Sigma_n)] \rightarrow_p 0,$$

where  $\Phi(\Sigma_n)$  is the Gaussian distribution with mean 0 and variance  $\Sigma_n$ , (4.6) implies that

$$\sqrt{L_n}\Sigma_n^{-1}\{T[\mathbf{X}^*(\hat{\boldsymbol{\tau}}^{(n)})] - E^*T[\mathbf{X}^*(\boldsymbol{t})]\} \rightarrow N(0, J).$$

Let  $\hat{n}_i = \hat{\tau}_{i+1}^{(n)} - \hat{\tau}_i^{(n)}$ . We first state conditions on the estimated segmentation which guarantee 4.6.

A9.  $\frac{\hat{U}_n}{n} \rightarrow 0$ ,

A10.  $\frac{1}{n} \sum_{i:\hat{n}_i \leq k} \hat{n}_i \rightarrow 0$  for all  $k < \infty$ ,

A11.  $\frac{L_n}{n} \sum_{i=1}^{U_n} \min_{1 \leq j \leq \hat{U}_n} |\tau_i - \hat{\tau}_j| \rightarrow_p 0$ .

Assumptions A9 and A10 for  $\hat{\boldsymbol{\tau}}^{(n)}$  mirror assumptions A3 and A4 for  $\boldsymbol{\tau}^{(n)}$ . Assumption A11 is a consistency criterion: As the data set grows, the total discrepancy in the estimation of  $\boldsymbol{\tau}^{(n)}$  by  $\hat{\boldsymbol{\tau}}^{(n)}$  must be small.

**Theorem 4.5.** *Under assumptions A1-A11, (4.6) holds.*

The proof is given in the appendix. There are trivial extensions of this theorem to smooth functions of vector means, which are, in fact, needed but simply cloud the exposition.

Theorem 4.5 implies that confidence intervals based on subsamples

$$\{\mathbf{X}^{*,j}(\hat{\boldsymbol{\tau}}^{(n)}) : j = 1, \dots, B\}$$

constructed by Algorithm 4.3 conditional on  $\hat{\boldsymbol{\tau}}^{(n)}$  are consistent, as long as  $\hat{\boldsymbol{\tau}}^{(n)}$  satisfies A9-A11. Here is the formal statement of this fact in the one dimensional case, where  $\hat{\sigma}_n^2$  replaces  $\hat{\Sigma}_n$  and  $\boldsymbol{g}$  is the identity.

**Corollary 4.6.** *Under assumptions A1-A11,*

(a) *Let  $\hat{\sigma}_n^2$  be the block bootstrap estimate of variance defined in (4.4), then*

$$P(\bar{X} - z_{1-\alpha/2}\hat{\sigma}_n/\sqrt{n} < \mu < \bar{X} + z_{1-\alpha/2}\hat{\sigma}_n/\sqrt{n}) \rightarrow_p 1 - \alpha.$$

(b) *The confidence interval estimated by Efron's percentile method is consistent. That is,*

$$P([\bar{X}_{([n\alpha/2])}^* < \mu < \bar{X}_{([n(1-\alpha/2)])}^*]) \rightarrow_p 1 - \alpha.$$

### 4.3. Segmentation Methods

The objective of the segmentation step is to divide the original data sequence  $X_1, \dots, X_n$  into approximately homogeneous regions so that the variance estimated in Algorithm 4.3 approximates the true variance of  $T_n$ . A segmentation into regions of constant mean guarantees that Algorithm 4.3 gives consistent variance estimates. Therefore, we focus here on the segmentation of  $X$  into regions of constant mean.

In our simulation and data analysis, we used the dyadic segmentation approach, which we motivate and describe here using the simple case of  $g$  identity. First consider the simple case where  $X_1, \dots, X_n$  are independent with variance 1. In testing the null hypothesis

$$H_0 : E[X_i] = \mu,$$

versus the alternative  $H_A$  that there exists  $1 < \tau < n$  such that  $E[X_i] = \mu_1$  for  $i < \tau$  and  $E[X_i] = \mu_2 \neq \mu_1$  for  $i \geq \tau$ , one can show that the following is the generalized likelihood ratio test:

$$\text{Reject } H_0 \text{ if } \max_{1 < j < n} nM(j) > c,$$

where

$$M(j) = \frac{j}{n} (\bar{X}_{1:j} - \bar{X}_{1:n})^2 + \frac{n-j}{n} (\bar{X}_{j+1:n} - \bar{X}_{1:n})^2. \quad (4.7)$$

The maximum likelihood estimate of the change-point parameter  $\tau$  is the value that maximizes  $M(j)$ .

In our original problem of variance estimation, our proof of Theorem 4.5 in the appendix shows that, in the case where there is one true change in mean at  $\tau$ , the difference between the variance estimated by block bootstrap with block length  $L$  given no segmentation (i.e.  $\mathbf{t}^{(n)} = \{0, n\}$ ) and the variance estimated by Algorithm 4.3 conditioned on a change-point at  $\tau$  is  $LM(\tau) + o_p(1)$ . Hence, segmenting at  $\hat{\tau} = \arg \max_j M(j)$  is optimal in the sense that  $\hat{\tau}$  is the change-point estimate that minimizes the asymptotic error of the block bootstrap variance estimate. This fact does not require the independence assumption, and is true for any stationary sequence. Thus, if we knew that there were only one change-point, and if the goal of the segmentation is to obtain the best stratified variance estimate, then the best place to segment is  $t$ . The block bootstrap variance estimate, given the segmentation  $\{0, t, n\}$ , would be

$$\begin{aligned} V(t) &= \left(\frac{t}{n^2}\right) \sum_{i=1}^{t-tL/n} (\bar{X}_{i:i+tL/n} - \bar{X}_{1:t})^2 \\ &+ \left(\frac{n-t}{n^2}\right) \sum_{i=t+1}^{n-(n-t)L/n} (\bar{X}_{i:i+(n-t)L/n} - \bar{X}_{t+1:n})^2. \end{aligned} \quad (4.8)$$

The Dyadic segmentation procedure recursively applies the above logic, as described below.

**Algorithm 4.7.** Fix minimum region length  $0 < L_s < n$  and threshold  $b > 0$ . Initialize  $\mathbf{t} = \{t_0 = 0, t_1 = n\}$ . Repeat:

- (a) For  $i = 1, \dots, |\mathbf{t}| - 1$ , let  $M^{(i)}(j)$  and  $V^{(i)}(j)$  be respectively the processes (4.7) and (4.8) computed on the subsequence  $X_{t_{i-1}+1}, \dots, X_{t_i}$ . If  $t_i - t_{i-1} > 2L_s$  then let  $t'_i = \arg \max_{t_{i-1}+L_s < j < t_i - L_s} M^{(i)}(j)$ ,  $B_i = M^{(i)}(t'_i)$ , and  $V_i = V^{(i)}(t'_i)$ . Otherwise, let  $B_i = 0$ ,  $V_i = \infty$ .
- (b) let  $\lambda_i = L(t_i - t_{i-1})/n$ , and

$$J_i = 1 \left( \frac{(t_i - t_{i-1})B_i}{\sqrt{V_i \hat{\lambda}_i}} > b \right).$$

If  $\prod_i J_i = 0$ , stop and return  $\mathbf{t}$ .

- (c) Let  $i^* = \arg \max_i B_i$ , and  $t^{\text{new}} = t'_{i^*}$ .
- (d) Let  $\mathbf{t} = \mathbf{t} \cup t^{\text{new}}$ , reordered so that  $t_i$  is monotonically increasing in  $i$ .

Given the current segmentation, each step of the recursion in Algorithm 4.7 proceeds as follows: In step 1,  $M^{(i)}(j)$ , the between group sum of squares, and  $V^{(i)}(j)$ , the block bootstrap variance estimates, are computed for each segment  $[t_{i-1} + 1, t_i]$ .  $B_i$  is the maximum squared difference in mean for segment  $i$ ,  $t'_i$  is the change-point estimate that achieves this maximum, and  $\hat{\lambda}_i V_i$  is the estimate of variance given a change-point at  $t'_i$ . In computing  $B_i$  and  $V_i$  we do not allow segmentations that create a region with length less than  $L_s$ . In step 2, we normalize the statistic  $(t_i - t_{i-1})B_i$  by the estimated standard deviation  $\sqrt{\hat{\lambda}_i V_i}$ . If this normalized statistic is below the boundary  $b$  for every subsegment, then the recursion stops and returns the current segmentation. Otherwise, in step 3, the optimal location to segment  $t^{(new)}$  is chosen to be the cut that maximizes the decrease in error of the block bootstrap variance estimate, conditioned on the fact that it had passed the thresholding in step 2. In step 4, this new change-point  $t^{(new)}$  is added to the current segmentation  $\mathbf{t}$ . The normalization by  $V_i$  in step 2 is optional, and requires an appropriate choice  $L = L_b$  of the block bootstrap sample size. Often it is easier to choose this parameter after the segmentation is given. If a ball park value of  $L_b$  is not available, then the normalization by  $V_i$  can be omitted, in which case the parameter  $b$  in step 3 should be set to 0. This would be equivalent to stopping the segmentation only when the next optimal cut will violate the minimum region length  $L_s$ . In the example of Section 5.1 we did not use this option, thus decoupling the choice of  $L_s$  from that of  $L_b$ .

The two parameters required by Algorithm 4.7 are  $L_s$  and  $b$ . The choice for  $L_s$  is discussed in Section 4.5. The choice of  $b$  can be guided by the fact that under the null hypothesis, if  $L$  were chosen appropriately, then  $(t_i - t_{i-1})M^{(i)}(j)/[V^{(i)}(j)\hat{\lambda}_i]^{1/2}$  is a pivot with approximate distribution  $\chi_1^2$ . Asymptotic approximations for the family-wise error rate have been derived in the case of independent sequences (James et al., 1987). In the case of dependent sequences a crude Bonferroni adjustment can be applied to adjust for multiple testing. We also used the formulas given in James et al. (1987) to get a crude cutoff, which seems to work reasonably in practice.

Algorithm 4.7 belongs to the class of dyadic segmentation algorithms for detection of change-points, the consistency of which are studied by Vostrikova (1981). These algorithms are greedy procedures that avoid the search over all possible segmentations, which would be computationally intractable. They have been applied successfully to various settings in biology, including segmentation of GC content (Li et al., 2002) and the analysis of DNA copy number data (Olshen et al., 2004).

The consistency of Algorithm 4.3 requires conditions A9-A11 to be satisfied by the estimated segmentation. The key condition is A11 which defines a consistency criterion on the segmentation. Consistency of dyadic segmentation has been proved in Vostrikova (1981) for sequences that satisfy the following conditions:

- (a) Let  $\epsilon_t = X_t - E[X_t]$ , then  $\|\epsilon_t\|^2$  is a submartingale and  $E\|\epsilon_t\|^2 < ct^\beta$ ,  $c > 0$ ,  $\beta < 2$ .
- (b) The number of regions is fixed and the region sizes are of order  $n$ , i.e.

$$\tau_n = (nr_1, \dots, nr_U), \quad 0 < r_1 < \dots < r_U.$$

It is easy to verify that condition 1 is satisfied by the piecewise stationary model due to the mixing condition A2. Condition 2 is more stringent than our assumptions A3 and A4, under which  $U_n$  is allowed to increase with  $n$ . The consistency of dyadic segmentation for the case of  $U_n \rightarrow \infty$  has been explored in Venkatraman (1992), who gave asymptotic conditions on the rejection threshold and on the sizes of the regions to ensure consistency

under the assumption of an independent Gaussian sequence. However, these conditions are hard to verify in practice, and we believe that for our applications in genomics the more stringent condition of Vostrikova (1981) gives a reasonable approximation. Previous studies on segmenting the genome based on features such as the GC content (Fu and Curnow (1990), Li et al. (2002)) have used this finite regions assumption to achieve reasonable results.

The dyadic segmentation procedure uses information from the entire sequence to call the first change, and then recursively uses all of the information from each subsegment to call each successive change in that segment. An alternative is to use pseudo-sequential procedures, which are sequential (online) schemes that have been adapted for change-point detection when the entire sequence of a fixed length is completely observed. The basic idea of this class of methods is to do a directional scan starting at one end of the sequence. Every time a change-point is called, the observations prior to the change-point are ignored and the process starts over to look for the next change after the previously detected change-point. Specifically, let  $\hat{\tau}_0 = 0$ , and given  $\hat{\tau}_1, \dots, \hat{\tau}_k$ ,

$$\hat{\tau}_{k+1} = \inf\{l > \hat{\tau}_k : S(X_{\hat{\tau}_k}, X_{\hat{\tau}_{k+1}}, \dots, X_{\hat{\tau}_l}) > b\},$$

where  $S$  is a suitably defined change-point statistic and  $b$  is a pre-chosen boundary. The estimates from pseudo-sequential schemes depend on the direction in which the data is scanned. Thus, while they may be suitable for, say, timeseries data, they may not be natural for segmentation of genomic data, which in most cases does not have an obvious directionality. The consistency of pseudo-sequential procedures has been studied by Venkatraman (1992), who gave conditions on  $b = b_n$  and  $\hat{\tau}^{(n)}$  for consistency of  $\hat{\tau}^{(n)}$  under the setting that  $X_i$  are independent Gaussian with changing means.

#### 4.4. Testing the Null Hypothesis of No Associations

Here we extend the results in section 4.2 to testing null hypothesis of no association using non-linear statistics. As we discussed in Section 1.2, the inference problem typically posed in high-throughput genomics is that of association of two features. In terms of our framework we have two 0-1 processes  $\{I_k\}_{k=1, \dots, n}$  and  $\{J_k\}_{k=1, \dots, n}$  both defined on a segment of length  $n$  of the genome. We assume that the joint process  $\{I_k, J_k\}$  is piecewise stationary and mixing and want to test the hypothesis that the two point processes  $\{I_k\}_{k=1, \dots, n}$  and  $\{J_k\}_{k=1, \dots, n}$  are independent. We have studied two fairly natural test statistics in ENCODE, the ‘‘percent basepair overlap’’,

$$O_n = \frac{\sum_{k=1}^n I_k J_k}{\sum_{k=1}^n I_k},$$

and the ‘‘regional overlap,’’ which we define as

$$R_n = \frac{\sum_{k=1}^n I_k J_k (1 - I_{k-1} J_{k-1})}{\sum_{k=1}^n I_k (1 - I_{k-1})},$$

with large values of these statistics indicating dependence. The major problem we face in constructing a test is what critical values  $o_{n\alpha}, r_{n\alpha}$  we should specify so that

$$P_{H_0}[O_n \geq o_{n\alpha}] \approx \alpha, \tag{4.9}$$

and similarly for  $R_n$ . Here  $H_0$  is the hypothesis that the vectors  $(I_1, \dots, I_n)^T$  and  $(J_1, \dots, J_n)^T$  are independent.

We aim for statistics based on  $O_n, R_n$  (respectively) which are asymptotically Gaussian with mean 0 under  $H_0$ , the hypothesis that the vectors  $(I_1, \dots, I_n)^T$  and  $(J_1, \dots, J_n)^T$  are independent. The following approximations suggest what such statistics should be for  $O_n$ .

$$\begin{aligned} E_{H_0} O_n &\approx \frac{E_{H_0} \sum_{k=1}^n I_k J_k}{E_{H_0} \sum_{k=1}^n I_k} \\ &\approx \frac{\frac{1}{n} \sum_{k=1}^n E_{H_0} I_k E_{H_0} J_k}{\frac{1}{n} \sum_{k=1}^n E_{H_0} I_k} \end{aligned}$$

For a single stationary regime,  $U = 1$ ,

$$E_{H_0} O_n \approx \frac{E_{H_0} I_1 E_{H_0} J_1}{E_{H_0} I_1} = E_{H_0} J_1 .$$

More generally

$$E_{H_0} O_n \approx \frac{\sum_{i=1}^U \lambda_i E_{H_0}^{(i)}(U) E_{H_0}^{(i)}(J)}{\sum_{i=1}^U \lambda_i E_{H_0}^{(i)}(U)} \equiv \mu$$

where  $E_{H_0}^{(i)}$  is the expectation for the  $i$ th segment under  $H$ .

The same heuristics show that if

$$\bar{I}_i \equiv \frac{1}{n_i} \sum_{k=1}^{n_i} I_{ik}, \quad \bar{J}_i \equiv \frac{1}{n_i} \sum_{k=1}^{n_i} J_{ik}$$

where, as usual,  $I_{ik}, J_{ik}$  correspond to the  $k$ th observation in the  $i$ th segment, then

$$E_{H_0} \frac{\sum_{i=1}^U \lambda_i \bar{I}_i \bar{J}_i}{\sum_{i=1}^U \lambda_i \bar{I}_i} \approx \mu \quad (4.10)$$

also. Similarly,

$$E_{H_0}(U_n) \approx E_{H_0} \left\{ \frac{\sum_{i=1}^U \lambda_i (\bar{I}_i \bar{J}_i - \bar{\bar{I}}_i \bar{\bar{J}}_i)}{\sum_{i=1}^U \lambda_i (\bar{I}_i - \bar{\bar{I}}_i)} \right\} \quad (4.11)$$

$$\begin{aligned} \text{where } \bar{\bar{I}}_i &= \frac{1}{(n_i - 1)} \sum_{k=2}^{n_i} I_{ik} I_{i(k-1)} \\ \bar{\bar{J}}_i &= \frac{1}{(n_i - 1)} \sum_{k=2}^{n_i} J_{ik} J_{i(k-1)} . \end{aligned}$$

We can apply this principle more generally to statistics which are functions of sums of products of  $I$ 's and  $J$ 's evaluated at the same positions. We proceed with construction of test statistics and estimation of null distributions. In view of (4.10) our test statistic based on  $O_n$  is

$$T_n^O \equiv n^{\frac{1}{2}} (O_n - \tilde{J}_n) \quad (4.12)$$

where

$$\tilde{J}_n \equiv \left( \sum_{i=1}^U \hat{\lambda}_i \hat{I}_i \hat{J}_i \right) / \frac{1}{n} \sum_{j=1}^n \hat{I}_j$$

$$\text{where } \hat{\lambda}_i = \lambda_i(\hat{\mathbf{t}}), \quad \hat{I}_i = n_i^{-1}(\hat{\mathbf{t}}) \sum_{j=\hat{t}_{i-1}+1}^{\hat{t}_i} I_j \quad (4.13)$$

with  $\hat{J}_i$  similarly defined. Here is the algorithm.

**Algorithm 4.8.** *We do the following.*

- (a) *Pick at random without replacement two starting points,  $K_1$  and  $K_2$ , of blocks of length  $L$  from  $\{1, \dots, n - L\}$ .*
- (b) *Let  $(I_{K_1+1}, \dots, I_{K_1+L})^T$  and  $(J_{K_1+1}, \dots, J_{K_1+L})^T$ ,  $(I_{K_2+1}, \dots, I_{K_2+L})^T$  and  $(J_{K_2+1}, \dots, J_{K_2+L})^T$  be the two sets of two feature indicators. Consider  $O_n$  with  $R_n$  being treated analogously.*
- (c) *Form*

$$\begin{aligned} \overline{IJ}_{nL}^{*1} &\equiv \frac{1}{L} \sum_{l=1}^L I_{K_1+l} J_{K_2+l} \\ \bar{I}_{nL}^{*1} &\equiv \frac{1}{L} \sum_{l=1}^L I_{K_1+l} \\ \overline{IJ}_{nL}^{*2} &\equiv \frac{1}{L} \sum_{l=1}^L I_{K_2+l} J_{K_1+l} \end{aligned}$$

and define  $\bar{I}_{nL}^{*2}$ ,  $\bar{J}_{nL}^{*1}$ ,  $\bar{J}_{nL}^{*2}$  analogously. Let

$$\begin{aligned} F_{nL}^* &\equiv \frac{1}{2} \left( \frac{\overline{IJ}_{nL}^{*1}}{\bar{I}_{nL}^{*1}} + \frac{\overline{IJ}_{nL}^{*2}}{\bar{I}_{nL}^{*2}} \right) \\ T_{nL}^* &\equiv F_{nL}^* - \bar{J}_{nL}^* \end{aligned}$$

where

$$\bar{J}_{nL}^* = \frac{1}{2} (\bar{J}_{nL}^{*1} + \bar{J}_{nL}^{*2})$$

and  $\bar{I}_{nL}^*$  is defined analogously. Let  $F_{nLb}^*$ ,  $\bar{IJ}_{nLb}^{*1}$  etc. be obtained by choosing  $(K_{1b}, K_{2b})$ ,  $b = 1, \dots, B$  independently as usual.

- (d) *We use the following  $c_{nL\alpha}$  as a critical value for  $O_n$  at level  $\alpha$ ,*

$$c_{nL\alpha} = \bar{J}_n + \left( \frac{2L}{n} \right)^{\frac{1}{2}} T_{nL(B(1-\alpha))}^*,$$

where  $T_{nL(1)}^* \leq \dots \leq T_{nL(B)}^*$  are the ordered  $T_{nLb}^*$  and  $[\cdot]$  denotes integer part and  $\bar{J}_n = \frac{1}{n} \sum_{k=1}^n J_k$ .

- (e) *If the sequence is piecewise stationary with estimated segments  $j = 1, \dots, s$  as in Section 4.3, we draw independently  $B$  sets of starting points,  $K_{11}^{(j)}, \dots, K_{1B}^{(j)}$  and  $K_{21}^{(j)}, \dots, K_{2B}^{(j)}$ , of blocks of length  $\hat{\lambda}_j L$  from each segment  $i = 1, \dots, j$  when each pair is drawn at random without replacement. Here  $\sum_{i=1}^U \hat{\lambda}_i = 1$  and  $\hat{\lambda}_i$  is proportional to the length of*

estimated segment  $i$ . Then piece  $T_{nLb}^*$  together as follows. Let

$$\begin{aligned}\overline{IJ}_{nLb}^{*1i} &= \frac{1}{L\widehat{\lambda}_i} \sum_{l=1}^{\widehat{\lambda}_i} I_{iK_{1b+l}} J_{iK_{2b+l}} \\ \bar{I}_{nLb}^{*1i} &= \frac{1}{L\widehat{\lambda}_i} \sum_{l=1}^L I_{iK_{1b+l}} \\ &\text{etc} \\ \bar{F}_{nLb}^* &= \sum_{i=1}^{\hat{U}} \widehat{\lambda}_i \left( \frac{\overline{IJ}_{nLb}^{*1i}}{\bar{I}_{nLb}^{*1i}} + \frac{\overline{IJ}_{nLb}^{*2i}}{\bar{I}_{nLb}^{*2i}} \right).\end{aligned}$$

Then,

$$T_{nLb}^* = F_{nLb}^* - \tilde{J}_{nLb}^*,$$

where

$$\tilde{J}_{nLb}^* = \frac{\sum_{i=1}^{\hat{U}} (\bar{I}_{nLb}^{*i})(\bar{J}_{nLb}^{*i})\widehat{\lambda}_i}{\sum_{i=1}^{\hat{U}} (\bar{I}_{nLb}^{*i})\widehat{\lambda}_i}$$

with  $\bar{I}_{nLb}^{*i} = \bar{I}_{nLb}^{*1i} + \bar{I}_{nLb}^{*2i}$ . The critical value is,

$$\tilde{J}_n + \left(\frac{2L}{n}\right)^{\frac{1}{2}} T_{nL(B(1-\alpha))}^*,$$

as before.

The statistic corresponding to  $R_n$  is

$$T_n^R = n^{\frac{1}{2}}(R_n - \hat{E}(R_n)) \quad (4.14)$$

where

$$\hat{E}(R_n) = \frac{\sum_{i=1}^{\hat{U}} \widehat{\lambda}_i (\hat{I}_i \hat{J}_i - \hat{I}_i \hat{J}_i)}{\sum_{i=1}^{\hat{U}} \widehat{\lambda}_i (\hat{I}_i - \hat{I}_i)}$$

where  $\hat{I}_i, \hat{J}_i$  are defined by (4.13) and, if we write  $\hat{n}_i$  for  $n_i(\hat{t})$ ,

$$\begin{aligned}\hat{I}_i &\equiv \hat{n}_i^{-1} \sum_{l=2}^{\hat{n}_i L/n} I_{\hat{t}_{i-1+l}} I_{\hat{t}_{i-1+l-1}} \\ \hat{J}_i &\equiv \hat{n}_i^{-1} \sum_{l=2}^{\hat{n}_i L/n} J_{\hat{t}_{i-1+l}} J_{\hat{t}_{i-1+l-1}}\end{aligned} \quad (4.15)$$

The proof of the following theorem is given in the Appendix.

**Theorem 4.9.** *If  $\mathcal{L}_0, P_0$  denote distributions under the hypothesis of independence and (A1)-(A11) hold, then*

$$1. \mathcal{L}_0(T_n^O) \implies \mathcal{N}(0, \sigma_0^2)$$

2. With probability tending to 1,

$$\mathcal{L}_0^*(T_{n,L}^{O*}) \implies \mathcal{N}(0, \sigma_0^2)$$

3.  $P_0[T_n^O \geq (\frac{2L}{n})^{\frac{1}{2}} \hat{q}_{1-\alpha}^0] \rightarrow \alpha$  where  $\hat{q}_{1-\alpha}^0$  is the  $[(1-\alpha)B]$ th of  $T_{nLb}^{O*}$ ,  $1 \leq b \leq B$ .

**Note.** The same results hold with  $T_n^O$  replaced by  $T_n^R$ ,  $T_{n,L}^{O*}$  replaced by  $T_{n,L}^{R*}$  etc. and the vector having added the components  $(\frac{1}{n} \sum_{i=2}^n I_i I_{i-1} J_i J_{i-1}, \frac{1}{n} \sum_{i=2}^n I_i I_{i-1}, \frac{1}{n} \sum_{i=2}^n J_i J_{i-1})$ .

#### 4.5. Choice of Segment Size $L_s$

Two tuning parameters appear in our procedure in addition to  $b$  appearing in the segmentation scheme.  $L_s$  is the smallest allowed size of a “stationary” piece after segmentation. It essentially determines the scale of the segmentation, which we view as an application context dependent quantity that users need to control. The reason is that stationarity is a matter of scale. To put it concretely, consider the situations where  $I_j$ ,  $j = 1, \dots, n$  are simply the base pair nucleotides  $A, C, G, T$  and consider the scale of a large gene of length  $n$ . Then, it seems natural that the exons and introns correspond to consecutive stationary regimes. However, suppose we now move our scale to a gene rich genomic region of length  $N$ . Now, it is the genes themselves and the intergenic regions which correspond to an at least initial segmentation.

This dependence of segmentation on scale has a natural intuitive consequence. Consider a statistic such as base pair overlap of two features. As one increases the region size  $n$  in which one wishes to declare significant overlap, the standard deviation of the statistic, which is  $O(n^{-1/2})$ , decreases, and p-values decrease. However, if, as one would expect, the region over which  $n$  increases becomes homogeneous on a larger scale, coarser segmentation would then be called for. This, as we have noted, necessarily increases the standard deviation of the statistic, and from that point of view significance becomes more difficult to achieve.

Put another way, it is not impossible to think of the whole genome itself as being stationary on a large scale, and that we can hierarchically segment the genome in many ways so that each large subsegment is stationary. For instance, a natural initial segmentation is to chromosomes.

Finally, we put it in mathematical terms going the other way from inhomogeneity to homogeneity. Start with a sequence of independent (say) bernoulli variables  $X_1, X_2, \dots, X_n$ , with  $X_j$  being Bernoulli( $p_j$ ). If the  $p_j$  are arbitrary, the only segmentation we can perform is the useless trivial one, where each  $X_j$  is its own segment. But, now if suppose that we tell ourselves that  $p_j$ ,  $1 \leq j \leq n/2$  are drawn i.i.d. from  $U(0, 1/2)$  and for  $n/2 + 1 \leq j \leq n$  from  $U(1/2, 1)$ , we suddenly just have two segments to consider.

Thus,  $L_s$  in our view needs to be treated as the smallest scale on which homogeneity is expected. Note that these considerations are not limited to testing. They also govern confidence intervals, as discussed in section 4.2.3.

#### 4.6. Choice of $L_b$ , the subsample size

We believe that the best way to choose  $L_b$ , after segmentation has been estimated, is so that the resulting bootstrap distribution of the statistics is as stable as possible and  $L_s$  is large but  $\ll n$ . We advocate but do not analyze further the following proposal put forward in  $m$ -out-of- $n$  subsampling by Bickel, Götze, and van Zwet (1997) and analyzed in detail by Götze and Rackauskas (2001) and Bickel and Sakov (2005):

- (a) Let  $\bar{X}_n^*(L)$  be the statistic computed from bootstrap sample drawn with blocks of length  $L$ . Compute the block bootstrap distribution  $\mathcal{L}_{L_k}$  for the statistic

$$\sqrt{L_k}(\bar{X}_n^*(L_k) - \bar{X}_n)$$

and  $L_k = \rho^k n$ , where  $\rho < 1$  and  $k = 1, 2, \dots, K$ .

- (b) Compute a distance  $d^*(k)$  between  $\mathcal{L}_{L_k}$  and  $\mathcal{L}_{L_{k-1}}$ .  
 (c) Choose  $L_b = L_{k_0^*}$ , where  $k_0^* = \arg \min d^*(k)$ .

In continuing work with Götze, van Zwet, we are in the process of trying to show that, under mild conditions, as  $n \rightarrow \infty$  we have  $L_b \rightarrow \infty$ ,  $L_b/n \rightarrow 0$ . More significantly, we expect that in a fashion analogous to Götze and Rackauskas (2001) and Bickel and Sakov (2005), under restrictive conditions and for suitable choice of distance,  $L_b$  yields an estimate which is as good as possible in the following sense: If  $\mathcal{L}_m$  is the actual distribution of  $\sqrt{m}(\bar{X}_m - \mu)$ ,  $d(k)$  is the distance between  $\mathcal{L}_m$  and  $\mathcal{L}_{L_k}$ , and  $k_0 = \arg \min_k d(k)$ , then

$$\frac{d(k_0^*)}{d(k_0)} \rightarrow_p c.$$

Thus,  $L_{k_0^*} = \rho^{k_0^*} n$  yields performance of the same order as  $\rho^{k_0} n$ .

## 5. Simulation and Data Studies

### 5.1. Simulation Study I

In this section, we perform a simple simulation study to demonstrate the power of block-bootstrapping method in the situation that features are naturally clustered. We simulated a binary sequence  $x_1 x_2 \dots x_n$  with  $n = 10K$  by the following Markovian model:

$$P(x_1 = 1) = \frac{p_0}{2}, P(x_i = 1) = \frac{p_0}{2} + (1 - p_0) \frac{\sum_{k=i-w}^{i-1} x_k}{w} \text{ for } i = 2, \dots, n, \quad (5.1)$$

where  $w$  is the order of the Markov model, or intuitively, the size of the dependent window, and  $p_0$  indicates the level of dependencies. The smaller  $p_0$  is, the stronger the dependence between the neighboring positions is. We define the following two types of features at position  $i$  in the sequence:

- Feature I: the occurrence of sequence 11100 starting at position  $i$
- Feature II: the occurrence of more than six 1's in the next 10 consecutive positions including the current position  $i$ .

From model (5.1), the feature II will occur in clusters in the sequence. The overlap between the two types of features can be measured by the Statistic

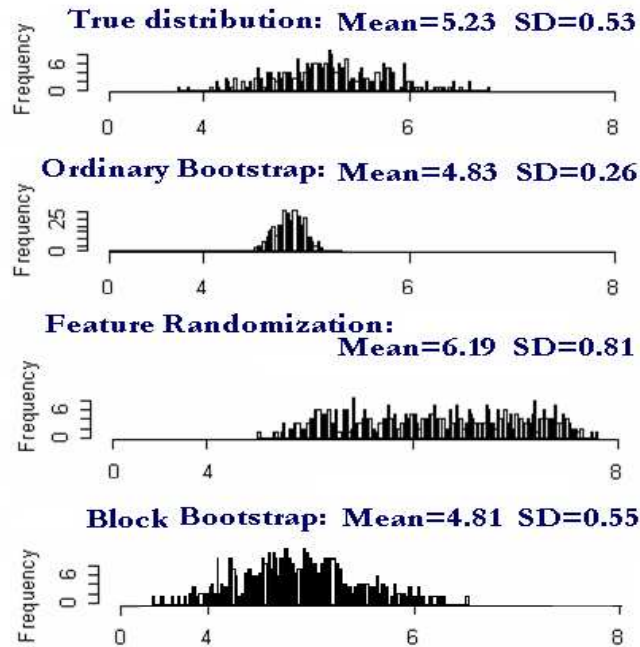
$$S = \frac{\sum_{k=1}^n I_k J_k}{\sum_{k=1}^n I_k}$$

with  $I_k, J_k$  being binary and indicating the occurrences of sites of Type I and II respectively.

Figure 1 shows the distribution of  $S$  estimated through different ways:

- The true distribution is the empirical distribution of estimated  $S$  from 10000 random sequences generated under model (5.1)

- The Ordinary Bootstrap distribution is derived by performing a base-by-base uniform sampling of the sequence  $x_1x_2\dots x_n$  to construct 10000 sequences of length  $n$ .
- The Feature Randomization distribution is derived by keeping features of type I fixed and randomizing uniformly the start positions of the features of type II to construct 10000 sequences of length  $n$ .
- The Block Bootstrap distribution is derived by drawing independent samples of blocks of length  $L = 40$  and stringing the blocks together to construct 10000 sequences of length  $n$ .



**Fig. 1.** Comparison of different bootstrapping schemes

We see that block bootstrap produces more reliable estimates on the variance of  $S$  compared to the naive methods: the ordinary bootstrapping and feature randomization. Both naive methods ignore the dependencies between positions and thus fails to take into account the natural clumps of the feature II. This is the key reason for the poor performance of the two naive methods.

## 5.2. Simulation Study II

Our second simulation study examines the case where the sequence is generated from a piecewise stationary model where there is more than one homogeneous region. As before,

we consider the problem of estimating the percentage of base pair overlap between two features, and compare the performance of four strategies:

- (a) feature randomization,
- (b) naive block bootstrap from unsegmented sequence,
- (c) block bootstrap from sequence segmented using the true change-points, and
- (d) block bootstrap from sequence segmented using the change-points estimated by binary segmentation.

In our simulation model, we generate  $X_t, Y_t$  independently from a Neyman-Scott process characterized as follows:

- (a) Clusters centers occur along the sequence according to a Poisson process of rate  $\lambda_i$  in region  $i$ .
- (b) The number of features in each cluster follows Poisson distribution with mean  $\alpha$ .
- (c) The start of features are located at a geometric distance (mean  $\mu$ ) from the cluster center.
- (d) The features are generated with length that is geometric with mean  $\beta$ .
- (e) Overlap between features generated using steps 1-4 are ignored.

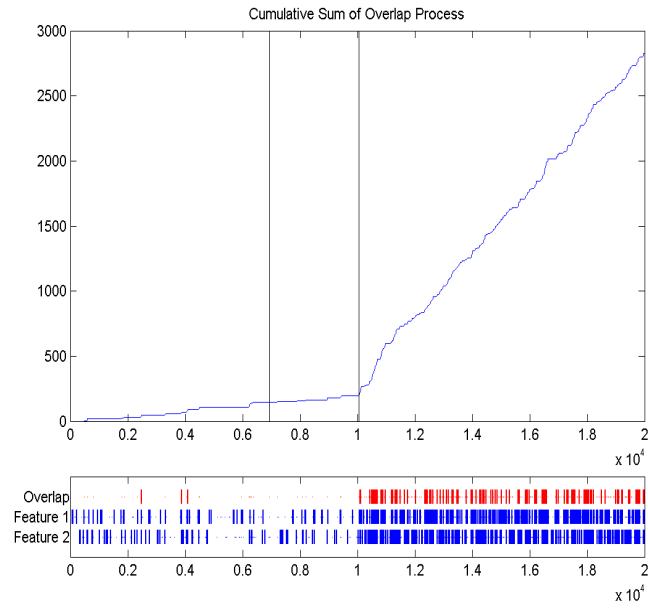
For simplicity we let there be only 2 homogeneous regions, each of length  $T = 10000$ . Consider the setting where the parameters for the two regions have the following values:  $(\lambda_1, \alpha_1, \mu_1, \beta_1) = (0.01, 10, 10, 5)$  and  $(\lambda_2, \alpha_2, \mu_2, \beta_2) = (0.02, 10, 10, 5)$ . Figure 2 shows a simulated example, where features A and B are plotted as well as their overlap. Figure 2 also shows the cumulative sum and the segmentation. Figure 3 shows respectively the histograms of the estimated distribution of the overlap statistic  $\bar{X}^*$  centered and scaled. It is clear that the feature randomization underestimates the standard deviation, whereas naive block bootstrap without segmentation gives a mixture distribution with long tails. Strategy 3, which subsamples assuming the true changepoint at  $\tau$  is known, gives the correct distribution as expected. Strategy 4, which uses the estimated change-point, reassuringly gives a very similar distribution to Strategy 3. Table 1 gives the standard deviation estimates.

Method	Standard Error Estimate	Fold change from true value
True value	1.2e-002	–
Uniform shuffle	3.6e-003	0.3
Subsample, no segmentation	1.7e-002	1.4
Subsample, true segmentation	1.0e-002	0.84
Subsample, estimated segmentation	9.6e-003	0.80

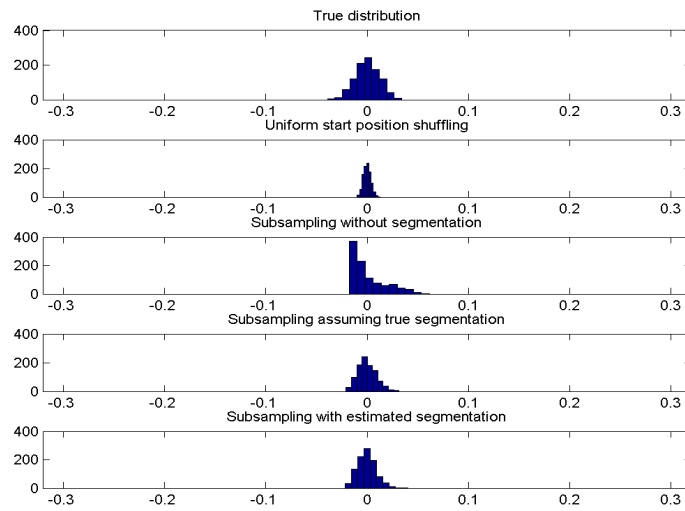
**Table 1.** Estimates of standard error by four sampling strategies in simulation study 2.

### 5.3. Association of Non-coding ENCODE annotations and Constrained Sequences.

Here we present a real example of the study of association between “constrained sequences” and “all non-exonic annotations” from the ENCODE project, limited to the 1.87Mbp ENCODE Pilot Region ENm001, also known as the CFTR locus. The constrained sequences are those highly conserved between human and the 14 mammalian species studied and sequenced by the ENCODE consortium. Evidence for evolutionary constraint in these non-coding annotations is of particular interest, as, for the most part these annotations do not



**Fig. 2.** Example of one instance from simulation model 2. Top plot shows cumulative sum and estimated segmentation.



**Fig. 3.** Comparison of different bootstrapping schemes

come with proposed biological functions, but rather constitute biochemically “active sites” of the DNA molecules in vivo. Enrichment of evolutionary constraint at these sites implies that the biochemical assays employed by the ENCODE consortium are capable of identifying biologically functional elements. We tested the association of this set of non-coding annotations and constrained elements using the base pair overlap statistic  $O_n$  in Section 4.3. We interpret the lack of association for two features as, given sequence composition as observed, are the assignments (by Nature) of feature A and feature B to individual bases made independently. We derive the significance of the observed statistic under this null hypothesis following the method proposed in Section 4.3.

As we discussed, we have several issues to deal with:

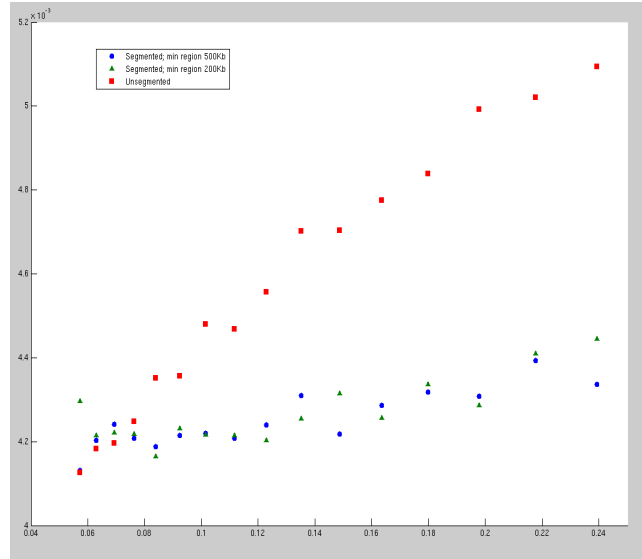
- i) How do we segment? That is, what statistic(s) do we use for segmentation?
- ii) Is segmentation necessary or is the region sufficiently homogeneous?
- iii) If we segment, what  $L_s$ ?
- iv) Given a segmentation, what  $L_b$  is appropriate?

Here are our methods:

- a) The simplest choice for i) and the one we followed was to segment according to both numerator and denominator: intersect partitions and enforce an  $L_s$  bound. Given our theory, this should ensure homogeneity in the mean of our statistic.
- b) Although strictly speaking ii) and iii) can be combined, we experimented a bit to also see if the theory of Section 4.1 was borne out in practice.
- c) We did not use the  $V$  statistic and thus only had to choose  $L_s$ . Again, we experimented with  $L_s = 500Kb$  to preserve as much genomic structure as possible, and  $L_s = 200Kb$  to ensure we had not undersegmented.

If the sequence were sufficiently homogeneous, we could forgo the initial segmentation step. Figure 4 shows an estimate of variance (with the appropriate renormalization) for a reasonable range of  $L_b$ , both before and after segmentation using the algorithm in Section 4.3. Two trends are clearly evident. First, segmentation greatly reduces the estimated variance. As we discussed in Section 4.1.2, inhomogeneity of the sequence causes an inflated estimate of variance. If the data were homogeneous, segmentation should not change the variance estimate. Thus, the fact that the estimated variances drop after segmentation for such a large range of  $L_b$ 's suggests that the data is inhomogeneous. Secondly, and more importantly, the estimated variance has significantly sharper increase with increasing  $L_b$  in the unsegmented data. This is evidence of inhomogeneity in mean across this Encode region: Underlying shifts in mean, if ignored, can be mistaken for spurious long range auto-correlation, which also implicitly runs against our assumption. In either case, as Theorem 4.2.1 suggests, we would be overly conservative. Thus, a preliminary exploration of the data convinces us that this Encode region is inhomogeneous and segmentation is necessary.

To segment the data, we applied the method in Section 4.3 to both features A and B, and then combined the segmentation. In segmenting each feature, we experimented with minimum segment lengths  $L_s$  of 200 and 500 Kb. Before subsampling, we combined the segmentations of A and B by taking a union of the change-points. This created regions



**Fig. 4.** Estimated  $\sigma_n$  as a function of  $L_b$  for 10,000 samples

with length less than  $L_s$ . However the total length of these regions comprise  $< 0.1\%$  of the total Encode region, and were left out of the remaining analyses.

200Kb and 500Kb gave 3 and 5 segments respectively. Table 2 gives the results for 500Kb. What is fairly surprising, but comforting is that over the whole broad range of  $L_b$  considered, the estimated SD of the statistic under the null was essentially flat after segmentation.

Flat here means that variability was within a Monte Carlo SD for the 10,000 replications we used. We expect that this phenomena will disappear as we increase the region size considered and the number of bootstrap replications.

We found that there is still substantial deviation from Gaussianity in both the segmented and unsegmented case for  $0.05 < L_b < 0.25$ , both in the tails, as detected by the Shapiro-Wilk test, and in the body of the distribution under Lilliefors test. As we discussed in Section 4.5, the definition of stationarity depends on the scale at which we view the genome. This suggests that our segmentation still does not take care of inhomogeneity in the variance. Hence, as we have mentioned, if we use the variance for the Gaussian approximation our results are still conservative.

The scientific conclusion of this example is that, indeed, there is strong association since the z value is over 9 SDs. But we note that for practical purposes although the effect of segmentation in variance is in the correct direction, the net effect is small.

$\rho = 0.99$			
Fraction ENCODE	basepair overalp ( $O_n$ )		
$(L_b/n)$	IQ Statistic	$SD_n$	z-score (of Observed Overlap)
0.239	0.0318	0.00509	8.47
0.217	0.0103	0.00502	8.60
0.197	0.0332	0.00499	8.65
0.179	0.0101	0.00483	8.92
0.163	0.0072	0.00477	9.04
0.148	0.0173	0.00470	9.18
0.135	0.0049	0.00470	9.18
0.122	0.0189	0.00455	9.47
0.111	0.0073	0.00446	9.66
0.101	0.0048	0.00447	9.64
0.092	0.0165	0.00435	9.91
0.083	0.0063	0.00435	9.92
0.076	0.0144	0.00424	10.16
0.069	0.0047	0.00419	10.29
0.063	0.00097	0.00418	10.32
0.057	n/a	0.00412	10.46

**Table 2:** Comparison of block-bootstrap distributions, Unsegmented Case:  $\rho^\beta n$  vs.  $\rho^{\beta+1} n$ 

$\rho = 0.99$			
Fraction ENCODE	basepair overalp ( $O_n$ )		
$(L_b/n)$	IQ Statistic	$SD_n$	z-score (of Observed Overlap)
0.239	0.0430	0.00433	9.95
0.217	0.0265	0.00439	9.82
0.197	0.0081	0.00430	10.02
0.179	0.0257	0.00431	10.00
0.163	0.0173	0.00428	10.07
0.148	0.0272	0.00421	10.23
0.135	0.0166	0.00430	10.02
0.122	0.0187	0.00423	10.18
0.111	0.0114	0.00421	10.26
0.101	0.0249	0.00422	10.23
0.092	0.0057	0.00421	10.25
0.083	0.0300	0.00418	10.31
0.076	0.0110	0.00420	10.26
0.069	0.0134	0.00424	10.18
0.063	0.0126	0.00420	10.27
0.057	n/a	0.00413	10.45

**Table 3:** Comparison of block-bootstrap distributions, Segmented, min region 500Kb:  $\rho^\beta n$  vs.  $\rho^{\beta+1} n$

Fraction ENCODE ( $L_b/n$ )	$\rho = 0.99$ basepair overlap ( $O_n$ )		
	IQ Statistic	$SD_n$	z-score (of Observed Overlap)
0.239	0.0267	0.00444	9.71
0.217	0.0150	0.00441	9.79
0.197	0.0237	0.00428	10.07
0.179	0.0154	0.00433	9.96
0.163	0.0337	0.00425	10.14
0.148	0.0158	0.00431	10.01
0.135	0.0271	0.00425	10.14
0.122	0.0155	0.00420	10.27
0.111	0.0225	0.00421	10.24
0.101	0.0218	0.00421	10.24
0.092	0.0168	0.00423	10.20
0.083	0.0150	0.00416	10.37
0.076	0.0223	0.00421	10.24
0.069	0.0140	0.00422	10.23
0.063	0.0179	0.00421	10.24
0.057	n/a	0.00429	10.05

**Table 4:** Comparison of block-bootstrap distributions, Segmented, min region 200Kb:  $\rho^\beta n$  vs.  $\rho^{\beta+1}n$

## 6. Appendix

### 6.1. Proof of Theorem 3.1

For simplicity we have  $d = 1$  and  $\mathbf{g}$  the identity. The general case follows by the Cramér-Wold device.

$$\text{Var}(S_n/\sqrt{n}) = A_n + B_n,$$

where

$$A_n = \frac{1}{n} \sum_{a=1}^n \sum_{b=1}^n \text{Cov}[X_a, X_b] I(\pi_1(a) = \pi_1(b)),$$

$$B_n = \frac{1}{n} \sum_{a=1}^n \sum_{b=1}^n \text{Cov}[X_a, X_b] I(\pi_1(a) \neq \pi_1(b)).$$

We will show that

$$B_n = o(I) \tag{6.1}$$

The theorem then follows from, for example, Corollary 1, page 142 of Herndorff (1984).

PROOF OF (6.1): We first note that, by A2 we have the standard bound

$$|\text{Cov}[X_a, X_{a+k}]| \leq C^2 m(k), \tag{6.2}$$

and by A2, since  $\beta > 1$ ,

$$\sum_{k \geq k_0} m(k) \leq c \sum_{k \geq k_0} k^{-\beta} \rightarrow 0 \quad \text{as } k_0 \rightarrow \infty.$$

Thus, for all  $\epsilon > 0$ , exists  $k_1(\epsilon)$  such that for all  $k \geq k_1(\epsilon)$ ,

$$\frac{1}{n} \sum_{a=1}^n \sum_{b=1}^n |\text{Cov}[X_a, X_b]| I(\pi_1(a) \neq \pi_1(b), |a - b| > k) \leq \epsilon. \quad (6.3)$$

Now, by A4, since

$$\sum_{a=1}^n I(\pi_1(a + k) > \pi_1(a) + 1) \leq \sum_{i: n_i \leq k} n_i = o(n),$$

by A2 and (6.2) we have,

$$\begin{aligned} & \frac{1}{n} \sum_{a=1}^n \sum_{b=1}^n |\text{Cov}[X_a, X_b]| I(\pi_1(a) \neq \pi_1(b), |a - b| \leq k) \\ & \leq \frac{C^2}{n} \sum_{a=1}^n \sum_{b=1}^n I(\pi_1(b) = \pi_1(a) + 1, |a - b| \leq k) + o(1). \end{aligned} \quad (6.4)$$

The first term on right hand side of (6.4) is bounded by

$$\frac{C^2}{n} \sum_{i=1}^U \sum_{j=n_i-k+1}^{n_i} 1 \leq 2C^2 U k / n. \quad (6.5)$$

Thus by A3, the above expression is  $o(1)$ . Combining (6.3-6.5) we obtain (6.1). Evidently,  $A_n = \sigma_n^2$ , and thus Theorem 4.2 follows.

## 6.2. Proof of Theorem 4.2.

(i) We use  $P^*$  throughout here for the randomization measure. By B2,

$$P^*[\pi_1(N) = \pi_1(N + L - 1)] \rightarrow 1 \quad (6.6)$$

since the complementary event can happen iff  $\pi_1(N) = i, \pi_2(N) = n_i - L + 1, \dots, n_i$  for some  $i$  and that probability is bounded by  $\sum_{i=1}^U \frac{L}{n}$ . But given  $\pi_1(N) = \pi_1(N + L - 1)$ ,  $\bar{X}_L^*$  is then a draw from the finite population  $\{L^{-1} \sum_{j=k}^{k+L-1} X_{ij} : 1 \leq k \leq n_i - L + 1\}$ . By A5 and A7 and Dedeker et al. (2004)

$$\max\{\rho_M(L^{-1} \sum_{j=k}^{k+L-1} X_{ij}, \mu_i) : 1 \leq k \leq n_i - L + 1, 1 \leq i \leq U\} \xrightarrow{P} 0 \quad (6.7)$$

and (i) follows from 6.6, 6.7 and property b) of  $\rho_M$ .

(ii) It is enough to show that

$$\rho_M\left(\mathcal{L}^*(\sqrt{L}(\bar{X}_L^* - \bar{X}) \mid \pi_1(N) = i), \mathcal{N}(0, C_i)\right) \xrightarrow{P} 0$$

for each  $i$ . But by Dedeker et al (2004),

$$\rho_M\left(L^{-\frac{1}{2}} \sum \left\{ (X_k - \mu_i) : \pi_1(k) = i, j \leq \pi_2(k) \leq j + L - 1 \right\}, \mathcal{N}(0, C_i)\right) \rightarrow 0$$

for each  $i$  uniformly in  $j$  by the stationarity of  $\{X_k : \pi_1(k) = i\}$  as  $L \rightarrow \infty$ . But

$$\sqrt{L}(\bar{X}_L^* - \bar{X}) = \sqrt{L}(\bar{X}_L^* - \mu) + o_P(1)$$

by B1. Further, as we have shown,

$$\rho_M(\sqrt{L}(\bar{X}_L^* - \mu_{\pi_1(N)}), \mathcal{N}(0, C_i)) \xrightarrow{P} 0$$

and finally,

$$L \sum_{i=1}^U f_i (\mu_i - \mu)^2 = o(1)$$

by (??), and again the result follows by Property b) of  $\rho_M$ .

(iii) It is enough to show that

$$\rho_M\left(\sum f_i \mathcal{N}(0, C_i), \mathcal{N}(0, \Sigma_n)\right) \rightarrow 0.$$

But

$$\rho_M(\mathcal{N}(0, C_i), \mathcal{N}(0, \Sigma_n)) = O(|C_i - \Sigma_n|)$$

and again

$$\rho_M^2\left(\sum f_i \mathcal{N}(0, C_i), \sum f_i \mathcal{N}(0, \Sigma_n)\right) \leq \sum f_i \rho_M^2(\mathcal{N}(0, C_i), \mathcal{N}(0, \Sigma_n)) = o(1).$$

□

### 6.3. Proof of Theorem 4.4

Let  $\bar{g}_i = \sum_{j=1}^{n_i} g(X_{ij})/n_i$ . By Theorem 4.2.1 of Politis, Romano and Wolf (1999), if  $f_i L \rightarrow \infty$ , then

$$(f_i L)^{1/2} [T_{f_i L}(X_{N_i; f_i L}) - \bar{g}_i] \Rightarrow N(0, C_i(f_i L))$$

in law in probability. Since by the stratified block bootstrap algorithm, the terms on the left above are independent across  $i$  under  $P^*$ , for every  $\epsilon > 0$  there exists  $U(\epsilon)$  independent of  $n$  such that

$$E^* \left| L^{1/2} \sum_{i=U(\epsilon)+1}^U f_i [T_{f_i L}(X_{N_i; f_i L}) - \bar{g}_i] \right|^2 \leq c^2 \sum_{i=U(\epsilon)+1}^U f_i \leq \epsilon$$

and hence,

$$E^* \left| L^{1/2} (T_L^* - T_n) - L^{1/2} [T_L^*(U(\epsilon)) - T_n(U(\epsilon))] \right|^2 \leq \epsilon$$

where

$$T_L^*(U(\epsilon)) = \sum_{i=1}^{U(\epsilon)} f_i T_{f_i L}(X_{N_i; f_i L}) \quad (6.8)$$

$$T_n(U(\epsilon)) = \sum_{i=1}^{U(\epsilon)} f_i \bar{g}_i. \quad (6.9)$$

The result then follows if

$$\min\{nf_i : 1 \leq i \leq U(\epsilon)\} \rightarrow \infty$$

for all  $\epsilon > 0$ . This is implied by assumption A4.

#### 6.4. Proof of Theorem 4.5

We first define some notation. Let  $\hat{R}_i = \{\hat{\tau}_{i-1} + 1, \dots, \hat{\tau}_i\}$  be the region between  $\hat{\tau}_{i-1}$  and  $\hat{\tau}_i$ , and

$$k_i = \sum_{j=1}^{U_n} I_{\{\hat{\tau}_{i-1} < \tau_j < \hat{\tau}_i\}} + 1$$

be the number of stationary regions within  $\hat{R}_i$ . Let  $\tau_{i,0} = \hat{\tau}_i$  and for  $j = 1, \dots, k_i - 1$ ,  $\tau_{i,j} = \min\{j : \tau_j > \tau_{i,j-1}\}$ . Thus  $\tau_{i,j}$  is the  $j$ -th true change-point in region  $\hat{R}_i$ , and let

$$\tilde{\tau} = \tilde{\tau}^{(n)} = \tau^{(n)} \bigcup \hat{\tau}^{(n)} = \{\tau_{i,j} : i = 1, \dots, U_n, j = 1, \dots, k_i\}.$$

Define  $R_{ij} = \{\tau_{i,j-1} + 1, \dots, \tau_{i,j}\}$  to be the  $j$ -th stationary region in  $\hat{R}_i$ . Then, define

$$\begin{aligned} f_{ij} &= |R_{ij}|/n, \\ \hat{f}_i &= |\hat{R}_i|/n, \\ \lambda_{ij} &= f_{ij}L, \\ \hat{\lambda}_i &= \hat{f}_iL. \end{aligned}$$

We use the notation  $X_{s;t} = \{X_s, \dots, X_{s+t-1}\}$  to denote the block of size  $t$  starting at  $s$ , and  $\bar{X}_{s;t}$  to be the mean of this block.

Consider a subsample  $X^*(\tilde{\tau}) = \{X_{N_{i,j}, \lambda_{ij}} : i = 1, \dots, \hat{U}_n; j = 1, \dots, k_i\}$  by Algorithm 4.3 conditioned on  $\tilde{\tau}$ . Thus for each  $i, j$ ,  $N_{i,j} \sim \text{Uniform}(R_{ij})$ . Let

$$\tilde{\mu}_{n,i} = \sum_{j=1}^{k_i} \frac{f_{ij}}{\hat{f}_i} \bar{X}_{N_{i,j}, \lambda_{ij}} \quad (6.10)$$

$$\tilde{\mu} = \sum_i \hat{f}_i \tilde{\mu}_{n,i}. \quad (6.11)$$

Then, because of assumptions (A10-A11),  $\tilde{\tau}$  satisfies all of the conditions of Theorem 4.4 and hence as a direct consequence of that theorem,

$$\sqrt{L_n} \sigma_n^{-1} [\tilde{\mu}_n - \bar{X}_n] \rightarrow N(0, 1). \quad (6.12)$$

Thus, proving Theorem 4.5 is equivalent to proving

$$\rho_M^2[F_n^*(\hat{\tau}^{(n)}), F_n^*(\tilde{\tau}^{(n)})] \rightarrow_p 0. \quad (6.13)$$

The subsample drawn by Algorithm 4.3 conditional on  $\hat{\tau}$  is

$$X^*(\hat{\tau}) = \{X_{N_i, \hat{\lambda}_i} : i = 1, \dots, \hat{U}_n\}.$$

We couple  $X^*(\hat{\tau})$  to  $X^*(\tilde{\tau})$  by letting

$$N_i = N_{ij} \quad \text{with probability } f_{ij}/\hat{f}_i, \quad j = 1, \dots, k_i$$

for each  $i = 1, \dots, \hat{U}_n$ . One can verify that by this construction,  $N_i \sim \text{Uniform}(\hat{R}_i)$  as required by Algorithm 4.3. We let  $\hat{\mu}_{n,i} = \bar{X}_{N_i, \hat{\lambda}_i}$  and  $\hat{\mu}_n = \sum_{i=1}^{\hat{U}_n} \hat{f}_i \hat{\mu}_{n,i}$ .

Let  $F_{n,i}^*(\hat{\tau})$  be the distribution of  $\sqrt{L}(\hat{\mu}_{n,i} - \bar{X}_{\tau_{i-1};n_i})$  and  $F_{n,i}^*(\tilde{\tau})$  be the distribution of  $\sqrt{L}(\tilde{\mu}_{n,i} - \bar{X}_{\tau_{i-1};n_i})$  conditional on  $X_1, \dots, X_n$  and  $\hat{\tau}^{(n)}$ . Since under this conditioning,  $\{\tilde{\mu}_{n,i}\}$  and  $\{\hat{\mu}_{n,i}\}$  are each sets of independent random variables, by property (b) of the Mallow's metric,

$$\begin{aligned} \rho_M^2[F_n^*(\hat{\tau}), F_n^*(\tilde{\tau})] &\leq \sum_{i=1}^{\hat{U}_n} \hat{f}_i^2 \rho_M^2[F_{n,i}^*(\hat{\tau}), F_{n,i}^*(\tilde{\tau})] \\ &\leq L \sum_{i=1}^{\hat{U}_n} \hat{f}_i^2 \mathbb{E}^*[\hat{\mu}_{n,i} - \tilde{\mu}_{n,i}]^2 \\ &= L \sum_{i=1}^{\hat{U}_n} \hat{f}_i^2 \text{Var}^*[\hat{\mu}_{n,i} - \tilde{\mu}_{n,i}]. \end{aligned}$$

The second inequality above is due to the definition of Mallow's metric, and the following equality is due to the fact that  $\mathbb{E}^*[\hat{\mu}_{n,i} - \tilde{\mu}_{n,i}] = o_p(1)$ . Hence, to show (6.13) it is sufficient to show that

$$\sum_{i=1}^{\hat{U}_n} \hat{f}_i \hat{\lambda}_i \text{Var}^*(\tilde{\mu}_{n,i} - \hat{\mu}_{n,i}) \rightarrow_p 0. \quad (6.14)$$

We now study the terms  $\hat{\lambda}_i \text{Var}^*(\tilde{\mu}_{n,i} - \hat{\mu}_{n,i})$ . For clarity, we first derive an explicit formula for this term when  $\hat{U}_n = 1$  and  $U_n = 2$ , and then generalize to the case  $\hat{U}_n > 2$ ,  $U_n > 1$ . Under this simple scenario, the estimated segmentation contains no change-points, but there is one true change-point at  $\tau$ . In this specific case we simplify our notation to let  $N_1$  be uniformly drawn from  $\{1, \dots, \tau\}$  and  $N_2$  be uniformly drawn from  $\{\tau + 1, \dots, n - L\}$ , and let  $f_1 = \tau/n$  and  $f_2 = 1 - f_1$ . Then, define

$$\begin{aligned} A_i^1 &= f_1 L \bar{X}_{i;f_1 L}, \\ A_i^2 &= f_2 L \bar{X}_{i;f_2 L}, \\ B_i^1 &= (1 - f_1) L \bar{X}_{i;(1-f_1)L}, \\ B_i^2 &= (1 - f_2) L \bar{X}_{i;(1-f_2)L}. \end{aligned}$$

The subsample drawn by Algorithm 4.3 assuming  $\hat{\tau}$  would simply be one block of length  $L$  with starting index  $N$  uniformly distributed on  $\{1, \dots, n\}$ . Since  $L/n \rightarrow 0$ , we can ignore the effects at the edges of the sequence and thus by the coupling of  $N$  to  $N_1$  and  $N_2$ ,

$$\hat{\mu} = [(A_{N_1}^1 + B_{N_1}^1)(1 - J) + (A_{N_2}^2 + B_{N_2}^2)J]/L + o(1),$$

where  $J \sim \text{Bernoulli}(f_2)$  is the indicator variable for the event  $\{N = N_2\}$ . The coupled subsample which assumes knowledge of  $\tau$  would have mean

$$\tilde{\mu} = (A_{N_1}^1 + A_{N_2}^2)/L + o(1),$$

and hence,

$$\hat{\mu} - \tilde{\mu} = [(B_{N_1}^1 - A_{N_2}^2)(1 - J) + (B_{N_2}^2 - A_{N_1}^1)J]/L.$$

Since

$$\text{Var}^*(\hat{\mu} - \tilde{\mu}) = \mathbb{E}^*[\text{Var}^*(\hat{\mu} - \tilde{\mu}|J)] + \text{Var}^*[\mathbb{E}^*(\hat{\mu} - \tilde{\mu}|J)],$$

we will examine each of the two terms in the above decomposition separately:

$$\begin{aligned}
 \mathbb{E}^*(\hat{\mu} - \tilde{\mu}|J = 0) &= \mathbb{E}^*(B_{N_1}^1 - A_{N_2}^2)/L \\
 &= \frac{1}{L} \left[ \sum_{N_1=1}^{\tau-L} \sum_{i=N_1+f_1L}^{N_1+L} X_i - \sum_{N_2=\tau+1}^{n-L} \sum_{i=N_2}^{N_2+f_2L} X_i \right] \\
 &= f_2[\bar{X}_{1:\tau} - \bar{X}_{\tau+1:n}] + o(1).
 \end{aligned}$$

Similarly,

$$\mathbb{E}^*(\hat{\mu} - \tilde{\mu}|J = 1) = f_1[\bar{X}_{1:\tau} - \bar{X}_{\tau+1:n}].$$

It is easy to check by combining the above equations that  $\mathbb{E}^*[\hat{\mu} - \tilde{\mu}] = 0$ , and hence,

$$\begin{aligned}
 L\text{Var}^*[\mathbb{E}^*(\hat{\mu} - \tilde{\mu})] &= L(f_1f_2^2 + f_2f_1^2)[\bar{X}_{1:\tau} - \bar{X}_{\tau+1:n}]^2 \\
 &= f_1f_2L[\bar{X}_{1:\tau} - \bar{X}_{\tau+1:n}]^2.
 \end{aligned} \tag{6.15}$$

Now consider the term  $\mathbb{E}^*[\text{Var}^*(\hat{\mu}_n - \tilde{\mu}_n|J)]$ .

$$\begin{aligned}
 \text{Var}^*[\hat{\mu} - \tilde{\mu}|J = 0] &= \text{Var}^*(B_{N_1}^1 - A_{N_2}^2)/L^2 \\
 &= [\text{Var}^*(B_{N_1}^1) + \text{Var}^*(A_{N_2}^2) - 2\text{Cov}^*(A_{N_2}^2, B_{N_1}^1)]/L^2.
 \end{aligned}$$

By independence of  $N_1$  and  $N_2$ ,  $\text{Cov}^*(A_{N_2}^2, B_{N_1}^1) = 0$ .

$$\begin{aligned}
 \text{Var}^*(B_{N_1}^1) &= f_2L\hat{\sigma}_1^2(f_2L), \\
 \text{Var}^*(A_{N_2}^2) &= f_2L\hat{\sigma}_2^2(f_2L),
 \end{aligned}$$

thus

$$\text{Var}^*[\hat{\mu} - \tilde{\mu}|J = 0] = f_2[\hat{\sigma}_1^2(f_2L) + \hat{\sigma}_2^2(f_2L)]/L,$$

and similarly,

$$\text{Var}^*[\hat{\mu} - \tilde{\mu}|J = 1] = f_1[\hat{\sigma}_1^2(f_1L) + \hat{\sigma}_2^2(f_1L)]/L,$$

and therefore

$$L\mathbb{E}^*[\text{Var}^*(\hat{\mu} - \tilde{\mu}|J)] = f_1f_2[\hat{\sigma}_1^2(f_1L) + \hat{\sigma}_1^2(f_2L) + \hat{\sigma}_2^2(f_1L) + \hat{\sigma}_2^2(f_2L)]. \tag{6.16}$$

Now we generalize (6.15) and (6.16) to the case where  $U_n > 2$ ,  $\hat{U}_n = 1$ . For  $j = 1, \dots, U_n$ , let  $N_j$  be uniformly distributed on  $\{\tau_{j-1} + 1, \dots, \tau_j\}$  and

$$\begin{aligned}
 A_{N_i}^i &= f_iL\bar{X}_{N_i;f_iL} \\
 B_{N_i}^i &= (1 - f_i)L\bar{X}_{N_i;(1-f_i)L}.
 \end{aligned}$$

Then

$$\hat{\mu} = \frac{1}{L} \sum_{i=1}^{U_n} I_{\{J=i\}}(A_{N_i}^i + B_{N_i}^i),$$

where  $J$  is multinomial with  $P(J = i) = f_i$ ,  $i = 1, \dots, U_n$ . The corresponding coupled statistic is

$$\tilde{\mu} = \frac{1}{L} \sum_{i=1}^{U_n} A_{N_i}^i,$$

and hence

$$\hat{\mu} - \tilde{\mu} = \frac{1}{L} \sum_{i=1}^{U_n} I_{\{J=i\}} \left( B_{N_i}^i - \sum_{j \neq i} A_{N_j}^j \right).$$

Similar to the  $U_n = 2$  case, by computing the first and second moments of  $A_{N_i}^i$  and  $B_{N_i}^i$  conditional on the observed sequence, we have the corresponding equations to (6.15) and (6.16) for the  $U_n > 2$ ,  $\hat{U}_n = 1$  case:

$$\begin{aligned} L \mathbb{V}\text{ar}^*[\mathbb{E}^*(\hat{\mu} - \tilde{\mu}|J)] &= \sum_{i=1}^{U_n} f_i \mathbb{E}^*(\hat{\mu} - \tilde{\mu}|J=i) \\ &= \sum_{i=1}^{U_n} f_i (1-f_i)^2 L \left[ \sum_{j \neq i} \frac{f_j}{1-f_i} (\bar{X}_{R_j} - \bar{X}_{R_i}) \right]^2 + o_p(1). \end{aligned} \quad (6.17)$$

$$L \mathbb{E}^*[\mathbb{V}\text{ar}^*(\hat{\mu} - \tilde{\mu}|J)] = \sum_{i=1}^{U_n} f_i (1-f_i) \left[ \hat{\sigma}_i^2((1-f_i)L) + \sum_{j \neq i} \frac{f_j}{1-f_i} \hat{\sigma}_j^2(f_i L) \right] + o_p(1) \quad (6.18)$$

Finally, generalizing (6.17) and (6.18) to the  $\hat{U}_n > 1$  case, we have, for each region  $i$ , an independent variable  $J_i$  defined as for  $J$  above, taking values in  $1, \dots, k_i$ , giving us:

$$\hat{\lambda}_i \mathbb{V}\text{ar}^*[\mathbb{E}^*(\hat{\mu}_{n,i} - \tilde{\mu}_{n,i}|J_i)] = \hat{\lambda}_i \sum_{j=1}^{k_i} \frac{f_{ij}}{\hat{f}_i} \left(1 - \frac{f_{ij}}{\hat{f}_i}\right)^2 \left[ \sum_{k \neq j} \frac{f_{ik}}{\hat{f}_i - f_{ij}} (\bar{X}_{R_{ik}} - \bar{X}_{R_{ij}}) \right]^2,$$

and

$$\begin{aligned} &\hat{\lambda}_i \mathbb{E}^*[\mathbb{V}\text{ar}^*(\hat{\mu}_{n,i} - \tilde{\mu}_{n,i}|J_i)] \\ &= \sum_{j=1}^{k_i} \frac{f_{ij}}{\hat{f}_i} \left(1 - \frac{f_{ij}}{\hat{f}_i}\right) \left[ \hat{\sigma}_{ij}^2((\hat{f}_i - f_{ij})L) + \sum_{k \neq j} \frac{f_{ik}}{\hat{f}_i - f_{ij}} \hat{\sigma}_{ik}^2(f_{ij}L) \right], \end{aligned}$$

where for any  $R \subseteq \{1, \dots, n\}$ ,  $\bar{X}_R = \sum_{i \in R} X_i / |R|$  and for any  $i, j$ , and  $l$   $\hat{\sigma}_{ij}^2(l) = l \sum_{i \in R_{ij}} (\bar{X}_{i;l} - \bar{X}_{R_{ij}})^2 / n_{ij}$ . By assumptions A5 and A6, and summing the above quantities over  $i = 1, \dots, \hat{U}_n$ , we have

$$\begin{aligned} \sum_{i=1}^{\hat{U}_n} f_i \hat{\lambda}_i \mathbb{V}\text{ar}^*(\hat{\mu}_{n,i} - \tilde{\mu}_{n,i}) &\leq C \sum_{i=1}^{\hat{U}_n} f_i \sum_{j=1}^{k_i} \left[ \hat{\lambda}_i \frac{f_{ij}}{\hat{f}_i} \left(1 - \frac{f_{ij}}{\hat{f}_i}\right)^2 + \frac{f_{ij}}{\hat{f}_i} \left(1 - \frac{f_{ij}}{\hat{f}_i}\right) \right] \\ &\leq C \sum_{i=1}^{\hat{U}_n} \sum_{j=1}^{k_i} [L \min(f_{ij}, \hat{f}_i - f_{ij}) + \min(\hat{f}_i, \hat{f}_i - f_{ij})] \\ &\leq 2C \frac{L+1}{n} \sum_{i=1}^{\hat{U}_n} \min_{1 \leq j \leq \hat{U}_n} |\tau_i - \hat{\tau}_j|. \end{aligned}$$

By assumption A11, the above converges in probability to 0, and thus (6.14) holds.  $\square$

### 6.5. Proof of Theorem 4.10

We begin with the known stationary case. Evidently, we need only apply the delta method to

$$(\overline{IJ}_n, \bar{I}_n, \bar{J}_n)$$

where  $\overline{IJ}_n \equiv \frac{1}{n} \sum_{k=1}^n I_k J_k$  and  $\bar{I}_n \equiv \frac{1}{n} \sum_{k=1}^n I_k$ ,  $\bar{J}_n$  as defined. The result follows from Theorem 3.1. We continue with part 2. We require first a series of couplings.

Since we are working under  $H_0$ ,  $\{I_k\}$  and  $\{J_k\}$  are independent. Conditional on  $K_1, K_2$  we construct for each  $M$  a joint distribution of  $(I_n^c, J_n^c)$  with the following properties.

1. Let  $S_{11} \equiv \{k : k < K_1 < K_2, K_2 > K_1 + M + L\}$ ,  $S_{12} = \{k : k > K_2 > K_1 + M + L\}$  and  $S_{21}, S_{22}$  with  $K_1, K_2$  interchanged. Let  $S_a = \cup_{b=1}^2 S_{ab}$ . If  $|K_1 - K_2| \leq M + L$  let  $S_a \equiv S_b \equiv \phi$ . Else let  $(I_k^c, J_k^c)$ ,  $k \leftarrow S_1$  be independent of  $(I_k^c, J_k^c)$ ,  $k \in S_2$ .
2. The marginal distributions of  $\{I_k^c\}$  and  $\{I_k\}$  and  $\{J_k^c\}$  and  $\{J_k\}$  are the same respectively.
3. The conditional distribution of  $\{(I_k^c, J_k^c) : k \notin S_1 \cup S_2\}$  given  $\{(I_k^c, J_k^c) = (\varepsilon_{k1}, \varepsilon_{k2}) : k \in S_1 \cup S_2\}$  and  $\{(I_k, J_k) : 1 \leq k \leq n\}$  is that of  $\{(I_k, J_k) : k \notin S_1 \cup S_2$  given  $(I_k, J_k) = (\varepsilon_{k1}, \varepsilon_{k2}) : k \leftarrow S_1 \cup S_2\}$ . Since  $\{I_k\}, \{J_k\}$  are independent  $\{I_k^c\}$  and  $\{J_k^c\}$  are also.
4.  $P[I_k \neq I_k^c \text{ or } J_k \neq J_k^c \text{ for any } k \in S_1 \cup S_2] \leq m(M)$ .  
This construction is possible by Strassen's (1965) Theorem.
5. If  $|K_1 - K_2| \leq M + L$ , then make  $\{I_k^c\}\{J_k^c\}$  be independent of each other and of  $\{I_k, J_k\}$ .

If we adjoin  $(K_1, K_2)$  we see that we have constructed a coupling which depends only on  $M$  and which has properties 1–4 given  $K_1, K_2$ . We now define statistics  $\overline{IJ}^{*1c}$ ,  $\bar{J}^{*1c}$  etc. (suppressing dependence on  $n, L, M$ ) defined on  $(I^c, J^c)$  and bootstrap versions  $\overline{IJ}^{*1c}$  etc. as well.

We want to argue generically that if  $|g| \leq c < \infty$  and the assumptions of the theorem hold and

$$S_L \equiv \frac{1}{L} \sum_{l=1}^L [g(I_{K_1+l}, J_{K_2+l}) + g(I_{K_2+l}, J_{K_1+l})]$$

$$S_L^c = \frac{1}{L} \sum_{l=1}^L [g(I_{K_1+l}, J_{K_2+l}^c) + g(I_{K_2+l}, J_{K_1+l}^c)]$$

then,

$$|S_L - S_L^c| = o_P(L^{-\frac{1}{2}})$$

as  $L, n \rightarrow \infty$ ,  $L = o(n)$ ,  $M = o(n)$  uniformly in  $g$  as above. But,

$$P[|S_L - S_L^c| \geq \varepsilon L^{-\frac{1}{2}}, \quad |K_1 - K_2| \leq M + L] \leq \frac{M + L}{n} \quad (6.19)$$

$$P[|S_L - S_L^c| \geq \varepsilon L^{-\frac{1}{2}}, \quad |K_1 - K_2| > M + L] \leq m(M) \quad (6.20)$$

Plugging in  $g(I, J) = IJ$ ,  $I, J$  we see that in the stationary case,

$$T_{nL}^* = T_{nL}^{*c} + o_P(L^{-\frac{1}{2}})$$

where  $T_{nL}^{*c}$  is defined by replacing  $I, J$  by  $I^c, J^c$  throughout in Algorithm 4.9. Now we can appeal to Theorem 4.1 to establish part 2 of Theorem 4.10 once we note that for  $T_{nL}^c$  given  $K_1 - K_2 > M + L$ ,  $\bar{I}\bar{J}^{*1c}, \bar{I}\bar{J}^{*2c}, \bar{I}^{*1c}, \bar{I}^{*2c}, \bar{J}^{*1c}, \bar{J}^{*2c}$  are mutually independent so that the asymptotic covariance of the two components of  $T_{nL}^{*c}$  and  $J_{nL}^{*c}$  is 0.

The generalization to the case of a fixed known segmentation satisfying the assumptions of Theorem 4.3 is straightforward by coupling the component corresponding to each segment and computing the variance of the difference using the independence of the bootstrap statistics for different segments. Finally the case of unknown segmentation can be dealt with by reducing the unknown segmentation case to the known one using Theorem 4.4 and then applying the result for known segmentation, Theorem 4.5. The result follows.  $\square$

## References

- [1] Andrews, D. and Mallows, C. (1974). Scale mixtures of normal distributions. JRSS (B) 26: 99-102
- [2] Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. Science 228, 953-958.
- [3] Bickel, P.J. and Sakov A. (2005). On the Choice of  $m$  in the  $m$  out of  $n$  Bootstrap and its Application to Confidence Bounds for Extreme Percentiles, Statistica Sinica, to appear.
- [4] Bickel, P.J., Gotze, F. and van Zwet, W.R. (1997). Resampling fewer than  $n$  observations: gains, losses, and remedies for losses. Statist. Sinica. no.1 1-31. Empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995) (1997)
- [5] Birney, E. et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 447, 799-816.
- [6] Blakesley, R.W. et al. (2004). An intermediate grade of finished genomic sequence suitable for comparative analyses Genome Res. 14:2235-2244.
- [7] Kang K, Chung JH, Kim J.(2009). Evolutionary Conserved Motif Finder (ECMFinder) for genome-wide identification of clustered YY1- and CTCF-binding sites. Nucleic Acids Res. 37(6):2003-13.
- [8] Yu H, Yoo AS, Greenwald I.(2004). Cluster Analyzer for Transcription Sites (CATS): a C++-based program for identifying clustered transcription factor binding sites. Bioinformatics 20(7):1198-200.
- [9] Braun, J. and Muller, H.-G. (1998). Statistical Methods for DNA Sequence Segmentation. Statistical Science. 13(2):142-162.
- [11] Churchill, G.A. (1989). Stochastic Models for heterogeneous genome sequences. Bulletin of Mathematical Biology. 51:79-94
- [11] Churchill, G.A. (1992). Hidden markov chains and the analysis of genome structure. Computers in Chemistry. 16:107-115.

- [12] Dedecker, J., Doukhan, P., Lang, G., Leon R., J.R., Louhichi, S., Prieur, C. (2007). Weak dependence: with examples and applications. Lecture Notes in Statistics, 190. Springer, New York.
- [13] Efron, B. (1981). Nonparametric standard errors and confidence intervals. With discussion and a reply by the author. *Canad. J. Statist.* 9, no. 2, 139–172.
- [14] Fickett, J.W., Torney, D.C., Wolf D.R. (1992). Base compositional structure of genomes, *Genomics*, 13:1056-1064.
- [15] Fu, Y.-X. and Curnow, R.-N. (1990). Maximum likelihood estimation of multiple change-points. *Biometrika*, 77:563-573.
- [16] Gotze, F. and Rackauskas, A. (2001). Adaptive choice of bootstrap sample sizes. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of IMS Lecture Notes Monogr. Ser., pages 286-309. Inst. Math. Statist., Beachwood, OH, 2001.
- [17] James, B., James, K. L. and Siegmund, D. Tests for a change-point. *Biometrika*, 74, 71-84 (1987).
- [18] Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* 17, 1217-1241.
- [19] Li W, Stolovitzky G, Bernaola-Galván P, Oliver JL. (1998). Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Research*. 1998 8(9):916-28.
- [20] Li,W., Pedro Bernaola-Galván, Fatameh Haghighi, Ivo Grosse: Applications of Recursive Segmentation to the Analysis of DNA Sequences. *Computers & Chemistry* 26(5): 491-510 (2002)
- [21] Margulies, E.H. et al. (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome *Genome Res.* 17: 760-774.
- [22] Morgan, James N. and John A. Sonquist (1963), Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58:415-435.
- [23] 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*. 5:557-572.
- [24] Politis, D. and Romano, J. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics* 22, 2031–2050.
- [25] Politis, D., Romano, J., and Wolf, M. (1999). *Subsampling*. Springer-Verlag: New York.
- [26] Redon, R. et al. (2006). Global variation in copy number in the human genome. *Nature*. 444, 444 - 454
- [27] Sen, A. and Srivastava, M.S. (1975). On Tests for Detecting Change in Mean. *The Annals of Statistics*, Vol. 3, No. 1, 98-108.

- [28] Strassen, V. (1965) The existence of probability measures with given marginals, *Ann. Math. Stat.* 423-439. MR 31:1693
- [29] Thisted, R. and Efron, B. (1987) Did Shakespeare write a newly-discovered poem? *Biometrika* 74, no. 3, 445-455.