Biometrika (0000), **00**, 0, *pp*. 1–18
© 2008 Biometrika Trust
Printed in Great Britain

Advance Access publication on 00 Month 0000

Detecting Simultaneous Change-points in Multiple Sequences

BY NANCY R. ZHANG, DAVID O. SIEGMUND

Department of Statistics, Stanford University, 390 Serra Mall, Stanford, California, 94305-4065

nzhang@stanford.edu dos@stat.stanford.edu

AND HANLEE JI

CCSR 1115 - Division of Oncology, 269 Campus Drive, Stanford University School of Medicine, Stanford, California, 94305-4065

hanleeji@stanford.edu

AND JUN Z. LI

Department of Human Genetics, 5789 Med Sci II, 1241 E. Catherine St. SPC 5618, Ann Arbor, Michigan, 48109-5618

junzli@umich.edu

SUMMARY

We discuss the problem of detecting local signals that occur at the same location in multiple one dimensional noisy sequences, with particular attention to relatively weak signals that may occur in only a fraction of the sequences. We propose simple scan and segmentation algorithms based on the sum of the chi-square statistics for each individual sample, which is equivalent to the generalized likelihood ratio for a model where the errors in each sample are independent. The simple geometry of the statistic allows us to derived accurate analytic approximations to the significance level of such scans. The formulation of the model is motivated by the biological problem of detecting recurrent DNA copy number variants in multiple samples. We show using replicates and parent-child comparisons that pooling data across samples results in more accurate detection of copy number variants. We also apply the multisample segmentation algorithm to the analysis of a cohort of tumor samples containing complex nested and overlapping copy number aberrations, for which our method gives a sparse and intuitive cross-sample summary.

Some key words: Change-point detection, DNA copy number, Meta-analysis, Scan statistics, Segmentation, Boundary crossing

1. INTRODUCTION

We study in this paper the statistical problem of detecting local signals that occur at the same location in multiple noisy sequences. This inquiry is motivated by current problems in biology, where high-throughput genomic profiles are collected for cohorts of biological samples, and it may be of interest to pool data across samples to boost power for detecting simultaneously occurring signals.

We start by describing a few motivating applications. The primary focus of this paper is the detection of DNA copy number variants. DNA copy number variants are gains and losses of segments of chromosomes, and comprise an important class of genetic variation. Various laboratory techniques have been developed to measure the DNA copy number. These measurements are taken at a set of probes, each mapping to a specific location in the genome. The recorded data for each probe is usually a log transform of the ratio of the copy number measurement at that probe in the given sample versus its expected value, often computed from a set of population controls. The data thus produced are a set of linear profiles, one for each biological sample in the study. The goal in analyzing such data is often to find shared copy number variants across samples. We focus on this application in detail later in this paper, and provide an indepth literature review in Section 3.

Such simultaneous scans also arise in the analysis of other types of genomic profiling data, for example, data from genomic tiling microarrays. High-density genomic tiling microarrays cover a complete genome with densely tiled probes. These arrays can be used to assay in an unbiased manner multiple types of activity on the genome, including transcription, DNA-protein-binding, and chromatin modification. The earliest of the vast literature on this subject include Selinger et al. (2000) and Kapranov et al. (2002). As for copy number data, tiling array data are often collected for multiple samples in one study. It is also frequently of interest to detect common regions of activity, and to pool data across samples to locate weak signals (Piccolboni, 2008; Huber et al., 2006).

A third example is the meta-analysis of genetic linkage studies. Whole genome linkage scans seek to identify genetic regions that may contain susceptibility genes for diseases or other traits of interest. Often, several linkage studies with modest sample sizes are reported, with differing results for the same genomic region. This is not surprising, since the power of detection by individual studies is often modest. Wise et al. (1999) and Badner & Gershon (2002) proposed statistical criteria for the simultaneous analysis of multiple genome scans.

All of these motivating applications involve situations where a simultaneous scan for a shared signal across multiple linear profiles can potentially improve robustness and power by pooling information across profiles. Within individual profiles, the signal of interest, as well as the noise structure, may vary across applications. In this paper, we examine the specific problem of detecting a shared abrupt shift in mean when the noise within each profile is assumed to be independent and identically distributed Gaussian. The mean shift model can be directly applied to the detection of copy number variants. With modifications for correlated errors and probe-level effects, the methods can potentially also apply to transcription profiling using tiling arrays. The meta-analysis of multiple linkage studies can be viewed in similar light, but would need to account for the diversity of study designs. All of these applications have their own set of idiosyncracies that must be factored into the models, but we hope to convey themes common to simultaneous scan statistics that extend across applications.

We propose a simple scan procedure based on summing the chi-square statistics across samples. This is equivalent to the generalized likelihood ratio statistic for a model where the errors in each sample are independent. We provide accurate approximations to the false positive rate of such scans, which adjust for simultaneous testing.

In treating the specific problem of DNA copy number analysis, we show using a data set containing technical replicates and parent-child trios that conducting a simultaneous scan across samples allows higher detection accuracy. For the detection of multiple, possibly nested variant intervals, we propose a recursive algorithm that extends the conceptual foundations of the circular binary segmentation algorithm (Olshen et al., 2004), which was shown in the comparative evaluations of Lai et al. (2005) and Willenbrock & Fridlyand (2005) to perform well in single sample scans. We illustrate the segmentation algorithm on a set of tumor samples containing a complex region of nested aberrations, and make comparisons to existing hidden Markov model approaches to this problem.

2. METHODS

2·1. *Model Formulation*

The observed data is a two dimensional array $\{y_{it} : \quad i = 1, \ldots, N, \quad t = 1, \ldots, T\}$, where y_{it} is the data point for the i-th profile at location t, N is the total number of profiles, and T is the total number of locations. We assume that for each i, the random variables $y_i = \{y_{it} : \quad t = 1, \ldots, T\}$ are mutually independent and Gaussian with mean values μ_{it} and variances σ_i^2. Under the null hypothesis, the means for each profile are identical across locations. Under the alternative hypothesis of a single changed interval, there exist integer values $1 \leq \tau_1 < \tau_2 \leq T$ and a set of profiles $\mathcal{J} \subset \{1, \ldots, N\}$, such that for $i \in \mathcal{J}$,

$$\mu_{it} = \mu_i + \delta_i I_{\{\tau_1 < t \leq \tau_2\}}, \tag{1}$$

where the δ_i are non-zero constants and μ_i is the baseline mean level for profile i. Under the alternative hypothesis we refer to $(\tau_1, \tau_2]$ as a variant interval and $\mathcal{J}$ as the set of carriers associated with the interval. If the alternative hypothesis is true, we are interested primarily in detecting this situation and in estimating the endpoints of the variant interval, and secondarily in determining the carriers. Figure 1 shows a hypothetical data set containing $N = 4$ profiles and $T = 100$ data points per profile. In the applications we consider, N is usually in the tens to thousands, and T is usually in the hundreds of thousands.

This model is motivated by the analysis of DNA copy number data, for which we provide more background in Section 3. In that application, each profile is usually a different biological sample, with the locations referring to positions along chromosomes. The change-points τ_1, τ_2 demarcate changes in copy number. Empirical evidence suggest that the baseline means and sample variances differ substantially across samples, and that for a given copy number variant the shifts in mean differ across carriers. The two histograms in Figure 2a,b show the sample means $\bar{y}_{i,\tau_1:\tau_2} = (y_{i,\tau_1+1} + \cdots + y_{i,\tau_2})/(\tau_2 - \tau_1)$ within a given variant interval for a set of 62 samples described in Section 3·2, among which only a subset are carriers. The values of the sample means for carriers are marked by triangles. The locations of the triangles vary over a wide range, which motivates the allocation of a separate δ_i for each carrier at any given copy number variant.

In many applications, there are usually multiple variant intervals defined by different τ_1 and τ_2, and $\mathcal{J}$. In DNA copy number data, the magnitude of change differs widely across different changed intervals for any given sample. Figure 2c,d presents empirical evidence. For each of the samples $i = 1, 2$, a histogram of $\{y_{it} : t = 1, \ldots, T\}$ is plotted. The triangles mark the magnitudes of change for the detected change-points in that sample that were validated by the procedure described in Section 3·2. The locations of the triangles vary substantially, which motivates the estimation of a separate mean shift for each interval $(\tau_1, \tau_2]$. We describe our test statistics first for the simple case where there is at most one variant interval. Then, we build on these test statistics to obtain segmentation algorithms for cases where multiple variant intervals can occur.

2·2. *The Sum-of-Chisquares Statistic*

We begin by reviewing a method for the analysis of a single profile, where temporarily we suppress the dependence of our notation on the profile indicator i. For $\{y_t : \ t = 1, \ldots, T\}$, let $S_t = y_1 + \ldots + y_t$, $\bar{y}_t = S_t/t$, and $\hat{\sigma}^2 = T^{-1} \sum_1^T (y_t - \bar{y}_T)^2$. Change-point detection in a single sequence has been reviewed by Zacks (1983) and Bhattacharya (1994). Recently, Olshen et al. (2004) used likelihood ratio based statistics for analysis of DNA copy number data, and Zhang & Siegmund (2007) proposed a related model selection criterion for estimating the number of change-points. The statistic used by Olshen et al. (2004) is

$$\max_{s,t} U^2(s,t), \tag{2}$$

where

$$U(s,t) = \hat{\sigma}^{-1}\{S_t - S_s - (t-s)\bar{y}_T\}/[(t-s)\{1-(t-s)/T\}]^{1/2}, \tag{3}$$

and the max is taken over $1 \le s < t \le T,\ \ t-s \le T_0$. Here $T_0 < T$ is an assumed upper bound on the length of the variant interval, which in some contexts may be much smaller than T.

If the error standard deviation σ were known and used in place of $\hat{\sigma}$ in (3), (2) would be the likelihood ratio statistic. In practice σ must be estimated. Since T is usually large in typical applications, we shall for theoretical developments treat σ as known. Then we can without loss of generality set $\sigma = 1$. Numerical studies suggest that this is a reasonable simplification.

Now consider the model (1) for the original problem involving N sequences. To test the null hypothesis H_0 that $\mu_{it} = \mu_i$ for all t and all $i = 1, \ldots, N$ versus the alternative H_A that there exist values of $\tau_1 < \tau_2$ for which some δ_i are not zero, a direct generalization of (2) is $\max_{s<t} Z(s,t)$, where

$$Z(s,t) = \sum_{i=1}^{N} U_i^2(s,t) \tag{4}$$

and $U_i(s,t)$ is the sequence specific statistic defined as in (3) for the ith sequence. As in the single profile case, if the variances were known, (4) would be the generalized log likelihood ratio statistic for testing H_0 versus H_A. For each fixed $s < t$, the null distribution of $Z(s,t)$ is approximately χ^2 with N degrees of freedom. Large values of $\max_{s<t} Z(s,t)$ are evidence against the null hypothesis. If the null hypothesis is rejected, the maximum likelihood estimate of the location of the variant interval is $(s^*, t^*) = \text{argmax}_{s,t} Z(s,t)$.

2·3. *Approximations for the Significance Level*

We now describe an analytic approximation to the significance level for scan statistics of the form (4), which accounts for the simultaneous testing of multiple hypotheses that are dependent through the overlap of adjacent scanning windows. The approximation gives a fast and computationally simple way of controlling the false positive rates.

To describe the approximation, let f_N be the chi-square density with N degrees of freedom. Let $\nu(x)$ be the overshoot function defined in Siegmund (1985, p. 85), a simple approximation of which is $\nu(x) \approx [(2/x)\{\Phi(x/2) - 1/2\}]/\{(x/2)\Phi(x/2) + \varphi(x/2)\}$, where φ and Φ are respectively the standard Gaussian density and distribution function.

Then the significance level of the scan (4) using threshold b^2 is

$$\text{pr}\left(\max_{\substack{0<s<t<T \\ c_1T<t-s<c_2T}} Z_{s,t} > b^2\right) \approx .5b^4(1 - \frac{N-1}{b^2})^3 f_N(b^2) \tag{5}$$

$$\int_{c_1}^{c_2} \frac{1}{u^2(1-u)} \nu^2\left[\frac{b\{1-(N-1)/b^2\}}{\{Tu(1-u)\}^{1/2}}\right] du.$$

For $N = 1$, (5) is the approximation given for a single sequence in Siegmund (1992). The derivation method given there can be generalized to the case $N > 1$, but the simple direct generalization does not include the factor $1 - (N-1)/b^2$, which adjusts for the discrepancy between a sphere in N dimensions and its tangent hyperplane at a point. This discrepancy can be quite important when N is of the order of b^2, which is frequently the case for our applications. Some details are given in the appendix.

We used Monte Carlo simulations to test the accuracy of (5). The approximation is very accurate at moderate to small p-values, at all values of N. Detailed figures are given in the supplement.

2·4. *Search Algorithm for Multiple Variant Intervals*

In general the data may contain several, possibly nested, variant intervals. We now describe algorithms for detecting multiple change-points that are shared across samples. In motivating the algorithms, it is useful to distinguish between two scenarios: In the first, the variant intervals are short and reasonably well separated. For example, in the analysis of DNA copy number data collected from normal tissue samples, the copy number variants usually involve changes of small magnitude over short segments that are well separated along the genome. In this case detection of all variant intervals can be achieved in a single step as implemented in the multi-sample scan algorithm below. The carriers of the variant intervals are not identified, although they are often obvious from visual inspection of the data.

In the second scenario, the variant intervals cover a substantial portion of the sequences being analyzed, and changes may be overlapping or nested. An example is DNA copy number data collected from cancer samples, where somatic aberrations often span entire chromosomes and do not align as neatly across samples. In these cases the more complex multi-sample circular binary segmentation algorithm, which involves a recursion, works better. The multi-sample circular binary segmentation algorithm is conceptually similar to the iterative circular binary segmentation procedure proposed by Olshen *et al.* (2004) for segmentation of a single sequence. For multiple sequences in the course of the recursion we implicitly identify putative carriers of the variant intervals. We discuss below possible solutions of this auxiliary problem.

Algorithm (Multi-sample Scan). Fix a global significance level α, a maximum window size $T_0 < T$, and an overlap fraction $0 < f < 1$.

1. For each $\{(s,t): \quad 1 \le s < t \le T, \quad t-s < T_0\}$, compute $z_{s,t,\text{obs}}$, the observed value of $Z(s,t)$, and let $p_{s,t} = \text{pr}(Z_{\max} > z_{s,t,\text{obs}})$ denote the global p-value associated with $z_{s,t,\text{obs}}$.
2. Let $\mathcal{S} = \{(s,t) : p_{s,t} < \alpha\}$. Rank the pairs in $\mathcal{S}$ from smallest p-value to largest.
3. Starting from the first element in $\mathcal{S}$, if it overlaps by more than f with any of the segments ranked before it in $\mathcal{S}$, eliminate it from $\mathcal{S}$.

The set of variant intervals reported would be the final set $\mathcal{S}$. □

Algorithm (Multi-sample Circular Binary Segmentation). Fix the global significance level α, parameter p, and a maximum window $T_0 < T$. We denote by $Y_{h:k}$ the matrix $\{y_{i,t} : \quad i = 1,\ldots,N, \quad t = h,\ldots,k\}$.

1. Initialize $T_1 = 1$ and $T_2 = T$.
2. Compute

$$Z_{\max} = \max_{\substack{T_1 \le s < t \le T_2 \\ 1 \le t-s \le T_0}} \{Z(s,t)\}.$$

   Let (s^*, t^*) be the maximizing interval.
3. If the p-value of $Z_{\max}$, as computed using the approximations in Section 2·3, is less than α, then for each $(u,v) \in \{(T_1, s^*-1),\ (s^*,t^*),\ (t^*+1, T_2)\}$, do:

a. Determine which samples carry the variation, as described below. For all $t = u, \dots, v$, if a sample carries the variation, let $\hat{y}_{i,t} = \bar{y}_{i,u:v}$, and for the other samples let $\hat{y}_{i,t} = \bar{y}_{i,T_1:T_2}$. Let $Y'_{u:v} = Y_{u:v} - \hat{Y}_{u:v}$, where $\hat{Y}_{u:v}$ is the matrix $\{\hat{y}_{i,t} : \ i = 1, \dots, N, \ t = u, \dots, v\}$.

b. Repeat steps 2-3 for $T_1 = u$, $T_2 = v$ and the newly normalized $Y'_{u:v}$. □

This second algorithm is understandably slower than the multi-sample scan, because every time a changed segment is found within (u, v), the entire interval must be re-scanned in the next step of the recursion. The second algorithm is, however, as fast as separately applying circular binary segmentation to each of the individual sequences. If $T_0 = T$, then both algorithms, as stated, are $O(NT^2)$ in running time. The computation time of both can be improved to $O(NT \log T)$ using a recursive algorithm similar to binary search.

When a variant interval $(s, t]$ is identified across samples, it is often of interest to determine its carriers. This is in fact a necessary part of Step 3(a) of the multi-sample segmentation algorithm. In many cases the identification of carriers is obvious by visual inspection, but in other cases this poses a difficult auxiliary problem. It is natural to identify as carriers those samples whose interval specific statistic $U^2_{i,s,t}$ falls above a suitable threshold, so there is some statistical evidence indicating that this particular sample and interval have a variant mean value, although the evidence might not by itself be statistically significant after accounting for multiple testing.

In copy number data, there are sometimes long, small shifts in mean due to experimental artifacts. These can pass the test of the preceding paragraph, but the shifts are so small that they are of no scientific interest. These artifacts motivate an addition of a second part to our thresholding rule based on the standardized absolute difference in mean (or median) between points inside $(s, t]$ and the entire sample. These considerations have also been used by others, e.g., Willenbrock & Fridlyand (2005); Lai et al. (2005).

For the reasons given above, in applications to copy number data we found that a combination of both types of thresholding gives the best empirical results. Thus, if a multi-sample scan identifies a variant interval at $(s, t]$, we declare that the ith sample is a carrier if both of the following two conditions hold: The absolute difference in mean (or median) between values inside the interval and for the entire sample is greater than $\delta_\mu \hat{\sigma}_i$, and the nominal p-value of the sequence and interval specific chi-square statistic, $U_i^2(s, t)$, is less than δ_{χ^2}.

In Section 3, we choose the thresholds δ_μ and δ_{χ^2} based on performance on a set of validation data described in Section 3·2. To choose these thresholds when validation data are not available, the classification rules are functions of two quantities: the effect size defined as the shift in mean divided by the standard deviation, and length of the interval. Figure 3 shows the region in the effect size by interval length plane where a sample would be classified as a carrier. For copy number variants longer than $L = \delta_\mu / c$, where c^2 is the $1 - \delta_{\chi^2}$ quantile of χ^2_1 distribution, the absolute mean threshold rule is in effect, and for those variants shorter than L, the chi-square threshold is in effect. Thus, δ_{χ^2} can be chosen first. Then, δ_μ can be chosen based on a minimum shift in mean that would be scientifically interesting.

The curves in Figure 3 are computed using values of δ_μ and δ_{χ^2} that work well on the validation data set. The figure also shows the detection curve for a single sample scan of the entire genome containing 500,000 Illumina probes at a maximum window size of 200 and global p-value of 0.01. The area between the two detection boundaries are those effect size by interval length combinations that are missed in a single sample scan, but that might be detectable in a multi-sample scan through the pooling of information across samples. See the following section for examples.

These classification rules are designed specifically for analysis of DNA copy number data. For other types of data, different rules for identifying the carriers, perhaps incorporating problem specific knowledge and objectives, may be appropriate.

3. Analysis of DNA Copy Number Data

3·1. *Literature Review and Data Pre-processing*

DNA copy number variants are an important class of genetic variation, recently reviewed in Scherer et al. (2007), that may underlie a broad spectrum of human traits and diseases (Perry et al., 2007; Hollox et al., 2007). While there are many published methods for segmentation of copy number data, most deal with samples one at a time and emphasize data from tumor samples (Fridlyand et al., 2004; Olshen et al., 2004; Daruwala et al., 2004; Xing et al., 2007; Wang et al., 2005; Picard et al., 2005; Hsu et al., 2005; Engler et al., 2006; Wen et al., 2006; Broët & Richardson, 2006; Hupé et al., 2004; Lai et al., 2007; Tibshirani & Wang, 2008). However, since copy number variants can be inherited and are often shared across individuals, we would like to scan all samples simultaneously to detect shared copy number variants and to obtain a sparse multi-sample summary that can serve as the overall molecular signature for the cohort of samples.

In this paper, we focus on the de novo detection of inherited copy number variants. Since these variants are often population level polymorphisms due to a single mutation event in the history of the cohort, the break points should be exactly shared between samples that contain the same variant. They are usually relatively short and often involve only single copy changes. Thus, the signal within each sample is weak, and a joint analysis across samples has the potential to boost power.

Existing approaches for cross-sample analysis of DNA copy number fall into three categories: (I) Post-segmentation methods (Diskin et al., 2006; Newton et al., 1998; Newton & Lee, 2000; Rouveirol et al., 2006) segment each sample separately, reducing them to categorical vectors indicating regions of amplification, deletion, or normal copy number. Then, the samples are aligned, and a statistical model (Newton et al. (1998); Newton & Lee (2000)) or permutation based approach (Diskin et al. (2006)) is used to identify regions of shared variation. A hidden Markov model based approach is proposed in Wang et al. (2008), where the change-points are not assumed to be shared across samples. The output of Wang et al. (2008) is a plot by location in each of the samples of the posterior probability of variation. While Wang et al. (2008) focused on the analysis of cancer data, the authors mention that a shared change-point model would be desirable for the detection of inherited copy number variants, and they note the substantial computational task inherent in a satisfactory hidden Markov model approach for this problem. (II) Shah et al. (2007) used a multi-layer hierarchical hidden Markov model to segment all samples simultaneously. This method involves restrictive assumptions on the way that copy number changes are shared across samples. For example, it assumes that all carriers of a given copy number variant must have a change in the same direction. This is often not the case in copy number data from normal samples, as seen in the example in Section 3·3. It also assumes that all deletions or gains for a given sample have the same underlying mean, which is shown in Figure 2(c,d) to be inappropriate for our data. (III) The interval scores method of Lipson et al. (2006) uses a statistic similar to $Z(s,t)$ but without the squares. Like Shah et al. (2007), this method focuses only on common deletions and common amplifications, and is not suitable for detection of inherited copy number variants which often have both types of carriers at a given locus. The paper is mainly algorithmic and proposes useful approximate methods for fast search for high scoring intervals, which are quite different from the two algorithms we propose.

We will show evidence below that it can be beneficial to pool data across samples during the initial segmentation step. In contrast to existing cross-sample methods, our approach can computationally handle thousands of samples simultaneously, relies on less restrictive model assumptions, and involves easily comprehended tuning parameters.

Data measuring copy number contain well documented artifacts, which should be removed by pre-processing. One artifact is local trends, which were first noted in the statistics literature by Olshen et al. (2004). These local trends correlate with GC content (Bengtsson et al., 2008) and manifest themselves as local low magnitude shifts in mean that are reproducible across samples. In our experience, the local trends from Affymetrix and Illumina platforms processed on normal samples can be well estimated by the first or first and second principal components of the matrix of y values. Hence we normalize the data by reducing it to the residuals of its projection on the first 2 principal components.

Still another artifact is badly behaving individual probe sets, which give observations that are consistently quite different from background. Hence, to ameliorate the effect of probe sets that are consistently poorly performing, we also standardize each probe set to have median 0 and inter-quartile range 1. (This does not eliminate the effect of outliers, which are also present. See below.)

3·2. *Detection Accuracy of Inherited Copy Number Variants*

We assess the accuracy of our detection method on a set of 62 Illumina 550K Beadchips. The experiments were performed on DNA samples extracted from lymphoblastoid cell lines derived from healthy individuals, and were used as part of the Quality Assessment panel in a genomewide association study recently carried out at the Stanford Human Genome Center. The 62 samples represent 10 sets of trios consisting of a child and his/her two parents, and 16 pairs of technical replicates for 16 independent DNA samples.

To assess detection accuracy, we compare copy number variants identified for the two technical replicates of the same individual and those identified for the child with those identified for the parents. It is not possible to estimate type 1 and type 2 error rates from the data, but it is possible to define other measures of accuracy. Specifically, we define inconsistency of detections of copy number variants in individual samples as follows: In replicates, if a detected variant in one of the replicate pairs is not detected in the second sample of the pair, the variant is considered inconsistent. In this case, either the detection is a false positive or there is a false negative in the other sample. In trios, if a detected variant in the child is not detected in at least one of the parents, it is considered inconsistent. In this case, neglecting the rare event that the detection represents a de novo mutation, either the detection made in the child is a false positive or there is a false negative in one or both of the parents. In this way, detections made in the child samples and in the replicate sample pairs can be classified as consistent or inconsistent. The detections made in the parent samples are used only to validate the detections made in the child samples, and are not counted towards the total number of detections. Detection accuracy is thus assessed by plotting the number of consistent versus inconsistent detections, and different methods can be compared in such a plot. As described in the previous Section, after a copy number variant is found at a location $(s, t]$, one still needs to identify the carriers, and this affects the level of consistency. For example, if all of the samples are classified as "changed" at all variant locations, then there would be no inconsistencies. The preceding section describes practical thresholding solutions for carrier identification.

Figure 4 shows the results for different settings of the carrier detection thresholds. The horizontal axis is the number of total detections and the vertical axis is the number of inconsistent detections. For example, if a variant interval is found, and 5 child or replicate samples are determined to be carriers, it contributes 5 detections to the total. If 2 of those detections are not validated, then that adds 2 to the number on the vertical axis. In the parent child trios, a parent can validate a child but not vice versa. Figure 4 also plots the results obtained by segmenting each sample individually using circular binary segmentation, the curve shows the results for different p-value thresholds. We can see by comparing the multi-sample segmentation and circular binary segmentation that pooling information across samples does indeed improve accuracy. For example, out of 3000 copy number variant calls made by single sample segmentation, 1500 are inconsistent, whereas multi-sample scanning makes about 5000 total calls for 1500 inconsistent calls.

A substantial fraction of the detections are inconsistent. From visual inspection, we believe that most of the inconsistencies involve variant intervals of length one caused by low quality probe sets, which produce outliers that were not removed by the pre-processing described above. To reduce the influence of a individual probe sets, previous studies have placed a lower bound on the length of a variant interval (e.g. 10 in Jakobsson et al. (2008)). Although allowing copy number variants covering only one single nucleotide polymorphism creates many inconsistent calls, insertions and deletions that cover only one single nucleotide polymorphism in fact also make up the majority of the consistent calls. Consequently we find it preferable to flag these putative variant intervals and try to eliminate the false positives by closer examination of the data. For example a putative copy number variant that involves only one single nucleotide polymorphism in a single sequence may well be an outlier, and in any case our scientific interest is in polymorphims having some minimal frequency in the population.

3·3. *Example Analysis of a Complex Region*

As is documented in the Database of Genomic Variants (Iafrate et al., 2004), chromosome 22 contains a complex region of nested deletions at cytoband 22q11, which has several different variants in the human population. Many of the 62 samples we described in Section 3·2 carry this variant region, as is clearly noticeable in the heatmap in Figure 5. Since this variant interval contains nested changes, the circular binary segmentation algorithm is preferred to the scan algorithm for its analysis.

We consider only the first 2000 single nucleotide polymorphisms (SNPs) mapping to chromosome 22, which are shown completely in the top panel of Figure 5. We applied the multi-sample circular binary segmentation algorithm to this region with parameters $T_0 = T = 2000$, $\alpha = 0.001$, $\delta_\mu = 1.5$, and $\delta_{\chi^2} = 0.001$. The segmentation is shown in the lower panel of Figure 5. There are 3 visually noticeable variant regions. The first region is from SNP 416 to SNP 442, which corresponds to positions 17,017-17,368 kilobases (Kb). Compared to the cohort mean, there are both gains and losses in this region. The second region spans SNPs 996 to 1329 (20706 to 21549 Kb), and contains several layers of nested deletions with change-points at SNPs 1167, 1217, 1309, 1321 corresponding to chromosome positions 20996, 21110, 21379, and 21436 Kb. These nested variants have been previously identified using Affymetrix SNP-arrays (McCarroll et al., 2006), paired-end mapping (Kidd et al., 2008), and were found in other data taken from normal populations (Iafrate et al., 2004). Comparing the top and bottom panels of Figure 5, we see that the recursive algorithm reconstructs this complex region quite well. The third visible copy number variant is SNPs 1830-1880 (at 23986-24234 Kb), where there are at least 3 copy number levels. All of the copy number estimates in the child and replicate samples for these three variant regions are validated.

The hidden Markov model based method of Shah et al. (2007), when applied to this region, did not identify any of the three copy number variant regions. This presumably is a consequence of the modeling assumptions, which do not allow simultaneous deletions and insertions, and which require all deletions to be of the same magnitude. However, one should also acknowledge that the method of Shah et al. (2007) is designed for a different purpose. While our method aims to detect shared variant intervals and provide sparse summaries of a set of samples, Shah et al. (2007) is designed to find regions where a large fraction of samples experience changes in the same direction.

4. DISCUSSION

The proposed scan statistic is based on summing a chi-square change-point statistic across sequences. The simple geometry of the statistic allowed us to derive accurate analytic approximations to the significance level of such scans. The algorithms we proposed for detecting multiple change-points and identifying the carriers rely on 4 parameters. These are the global significance level α and T_0 for identifying the variant intervals, and δ_μ and δ_{χ^2} for identifying the carriers. The procedure is very robust to variation in T_0, so it can be specified conservatively. All of these parameters are easy to interpret and affect the results in a simple transparent way, so they can be easily modified to suit different scientific conditions.

The formulation we have chosen was motivated by the success of Olshen et al. (2004) in their analysis of copy number data in single samples. It is doubtful that any one approach can be optimal in problems of this complexity, and it would be useful to extend other single sample methods to deal with multiple samples. A useful version of hidden Markov models would be particularly welcome. There is one multi-sample method (Shah et al., 2007) for which there is readily available software. However, the model makes quite different assumptions from ours, and is aimed at different goals. Its running time is also forbiddingly long for even moderately large amounts of data. It would be interesting to make a more systematic comparison of these methods along the lines of Lai et al. (2007) for single samples.

We are studying two alternative methods. One is a multi-sequence version of the Bayes Information Criterion for model selection that we used for single sequence analysis (Zhang and Siegmund, 2006). This has the potential to identify variant intervals and carriers in a unified analysis. For a wide range of parameter settings it seems to identify carriers by using what amounts to the δ_{χ^2} part of our criterion. A second method, motivated by the expectation that only a small subset of the samples will exhibit variants at any particular location, is to use a weighted sum of chi-squares statistic that favors strong evidence from a subset of samples over weak evidence from all samples. Preliminary results indicate that both of these methods are promising.

We have also tried our methods on cancer data, and have found that they perform satisfactorily, although the main advantage of cross sample analyses seems to be found in studying inherited copy number variants, since their foot print is typically much shorter and weaker. The main potential advantage for cancer data is to provide a relatively clean overall signature for downstream analysis of related samples.

5. ACKNOWLEDGEMENTS

This research was partially supported by the National Institutes of Health and by the National Science Foundation.

APPENDIX

Proof of (2.5)

We indicate briefly here modifications of the proof of (2.5) in the one dimensional case required for the proof when N is of comparable order of magnitude as b^2.

In the one dimensional case the proof involves considering a union of events of the form that $Z_{s_0,t_0} > b$, but $Z_{s_0+s,t_0+t} < b$ for certain values of (s,t) that are small compared to (s_0,t_0). For these small values it is shown by taking one term of a Taylor series expansion that $b(Z_{s_0+s,t_0+t} - Z_{s_0,t_0})$ behaves like the sum of two independent random walks, one indexed by s, the other by t. After determining the (conditional on $Z_{s_0,t_0} \sim b$) mean and variance of these random walks the multiple testing correction in the form of the integral in (2.5) follows from renewal theory, as demonstrated by Siegmund (1992, Lemma 4).

The marginal distribution of $Z_{s,t}$ is χ with N degrees of freedom. Let $f_N(x)$ denote the χ^2 density with N degrees of freedom. From a straightforward approximation for large b of $\text{pr}\{Z_{s,t} \in b + dx/b\}$, in which we do not neglect N/b^2 even though b is assumed large, we find that the factor e^{-x} that multiplies $2f_N(b^2)$ in the one dimensional case now becomes $\exp[-x\{1-(N-1)/b^2\}]$.

In addition, to take account of the large number of dimensions in which $b\{Z(s_0+s,t_0+t) - Z(s_0,t_0)\}$ can vary, we consider not a one term, but a two term Taylor series expansion of the increments $b(Z_{s_0+s,t_0+t} - Z_{s,t})$. We can by spherical symmetry assume without loss of generality that all the coordinates of the vector $(U_1(s_0,t_0),\ldots,U_N(s_0,t_0))'$ are zero except for the first one. The expansion of $b(Z_{s_0+s,t_0+t} - Z_{s,t})$ contains linear terms in the first coordinate direction in the form of the sum of two random walks indexed by s and t, with (negative) means and variances proportional to $b^2/[2(t_0-s_0)\{1-(t_0-s_0)/T\}]$. In addition there are independent quadratic terms in the $N-1$ orthogonal directions with means proportional to $(N-1)/[2(t_0-s_0)\{1-(t_0-s_0)/T\}]$ and variances proportional to $(N-1)/[(t_0-s_0)\{1-(t_0-s_0)/T\}]^2$. Asymptotically important values of $t_0 - s_0$ are of order b^2, so stochastic fluctuations of the quadratic terms are negligible. The consequence of adding $(N-1)/[2(t_0-s_0)\{1-(t_0-s_0)/T\}]$ to the means of the random walks is that both the exponential under the integral and the drift of the local random walks are modified by the same correction factor: $1-(N-1)/b^2$, while the variances of the local random walks remain unchanged. Calculation now shows that Lemma 4 of Siegmund (1992) applies again to yield (2.5), which now contains the modifying $1-(N-1)/b^2$.

REFERENCES

BADNER, J. & GERSHON, E. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Molecular Psychiatry* **7**, 405–411.

BENGTSSON, H., IRIZARRY, R., CARVALHO, B. & SPEED, T. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759–767.

BHATTACHARYA, P. (1994). Some aspects of change-point analysis. In *Change-point Problems, IMS Monograph 23*, E. Carlstein, H. Muller & D. Siegmund, eds. Institute of Mathematical Statistics, pp. 28–56.

BROËT, P. & RICHARDSON, S. (2006). Detection of gene copy number changes in cgh microarrays using a spatially correlated mixture model. *Bioinformatics* **22**, 911–918.

DARUWALA, R. S., RUDRA, A., OSTRER, H., LUCITO, R., WIGLER, M. & MISHRA, B. (2004). A versatile statistical analysis algorithm to detect genome copy number variation. *Proc Natl Acad Sci U S A* **101**, 16292–16297.

DISKIN, S. J., ECK, T., GRESHOCK, J., MOSSE, Y. P., NAYLOR, T., STOECKERT JR., C. J., WEBER, B. L., MARIS, J. M. & GRANT, G. R. (2006). Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments. *Genome Research* **16**, 1149–1158.

ENGLER, D., MOHAPATRA, G., LOUIS, D. & BETENSKY, R. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridications. *Biostatistics* **7**, 399–421.

FRIDLYAND, J., SNIJDERS, A., PINKEL, D., ALBERTSON, D. G. & JAIN, A. (2004). Application of hidden markov models to the analysis of the array-cgh data. *Journal of Multivariate Analysis* **90**, 132–153.

HOLLOX, E. J. J., HUFFMEIER, U., ZEEUWEN, P. L. J. M. L., PALLA, R., LASCORZ, J., RODIJK-OLTHUIS, D., VAN DE KERKHOF, P. C. M. C., TRAUPE, H., DE JONGH, G., MARTIN, REIS, A., ARMOUR, J. A. L. A. & SCHALKWIJK, J. (2007). Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet* **40**, 23–25.

Hsu, L., Self, S., Grove, D., Randolph, T., Wang, K., Delrow, J., Loo, L. & Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6**, 211–226.

Huber, W., Toedling, J. & Steinmetz, L. M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**, 1963–1970.

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

Iafrate, J. A., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. & Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics* **36**, 949–951.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, R. J., Vanliere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A. & Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. & Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, A. N., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., Mckernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., Mccarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R. & Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64.

Lai, T. L., Xing, H. & Zhang, N. R. (2007). Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics* **9**, 290–307.

Lai, W. R., Johnson, M. D., Kucherlapati, R. & Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics* **21**, 3763–3770.

Lipson, D., Aumann, Y., Ben-Dor, A., Linial, N. & Yakhini, Z. (2006). Efficient calculation of interval scores for dna copy number data analysis. *Journal of Computational Biology* **13**, 215–228.

McCarroll, S., Hadnott, T., Perry, G., Sabeti, P., Zody, M., Barrett, J., Dallaire, S., Gabriel, S., Lee, C., Daly, M., Altshuler, D. & The International HapMap Consortium (2006). Common deletion polymorphisms in the human genome. *Nature Genetics* **38**, 86–92.

Newton, M., Gould, M., Reznikoff, C. & Haag, J. (1998). On the statistical analysis of allelic-loss data. *Statistics in Medicine* **17**, 1425–1445.

Newton, M. & Lee, Y. (2000). Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics* **56**, 1088–1097.

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.

Perry, G. H. H., Dominy, N. J. J., Claw, K. G. G., Lee, A. S. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A. A., Mountain, J. L. L., Misra, R., Carter, N. P. P., Lee, C. & Stone, A. C. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics* **39**, 1256 – 1260.

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

Piccolboni, A. (2008). Multivariate segmentation in the analysis of transcription tiling array data. *Journal of Computational Biology* **15**, 845–856.

Rouveirol, C., Stransky, N., Hupé, P., La Rosa, P., Viara, E., Barillot, E. & Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-cgh data. *Bioinformatics* **22**, 849–856.

Scherer, S., Lee, C., Birney, E., Altshuler, D., Eichler, E., Carter, N. & Hurles, M. (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39**, S7–S15.

Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M., Richmond, C. S., Blattner, F. R., Lockhart, D. J. & Church, G. M. (2000). Rna expression analysis using a 30 base pair resolution escherichia coli genome array. *Nature Biotechnology* **18**, 1262–1268.

Shah, S. P., Lam, W. L., Ng, R. T. & Murphy, K. P. (2007). Modeling recurrent dna copy number alterations in array cgh data. *Bioinformatics* **23**, 450–458.

Siegmund, D. (1992). Tail approximations for maxima of random fields. In *Probability Theory: Proceedings of the 1989 Singapore Probability Conference*, L. H. Chen, K. Choi, K. Yu & J.-H. Lou, eds. deGruyter, pp. 147–58.

Tibshirani, R. & Wang, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics* **9**, 18–29.

Wang, H., Veldink, J. H., Ophoff, R. A. & Sabatti, C. (2008). Markov models for inferring copy number variations from genotype data on illumina platforms. *Technical Report, Dept. of Statistics, University of California at Los Angeles* .

WANG, P., KIM, Y., POLLACK, J., NARASIMHAN, B. & TIBSHIRANI, R. (2005). A method for calling gains and losses in array-cgh data. *Biostatistics* **6**, 45–58.

WEN, C., WU, Y., HUANG, Y., CHEN, W., LIU, S., JIANG, S., JUANG, J., LIN, C., FANG, W., HSIUNG, C. & CHANG, I. (2006). A bayes regression approach to array-cgh data. *Statistical Applications in Molecular Biology* **5**.

WILLENBROCK, H. & FRIDLYAND, J. (2005). A comparison study: applying segmentation to arraycgh data for downstream analyses. *Bioinformatics* **21**, 4084–4091.

WISE, L., LANCHBURY, J. & LEWIS, C. (1999). Meta-analysis of genome scans. *Annals of Human Genetics* **63**, 263–272.

XING, B., GREENWOOD, C. M. T. M. & BULL, S. B. B. (2007). A hierarchical clustering method for estimating copy number variation. *Biostatistics* **8**, 632–653.

ZACKS, S. (1983). Survey of classical and bayesian approaches to the change-point problem: Fixed sample and sequential procedures in testing and estimation. In *Recent Advances in Statistics*. Academic Press, pp. 245–269.

ZHANG, N. & SIEGMUND, D. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63**, 22–32.

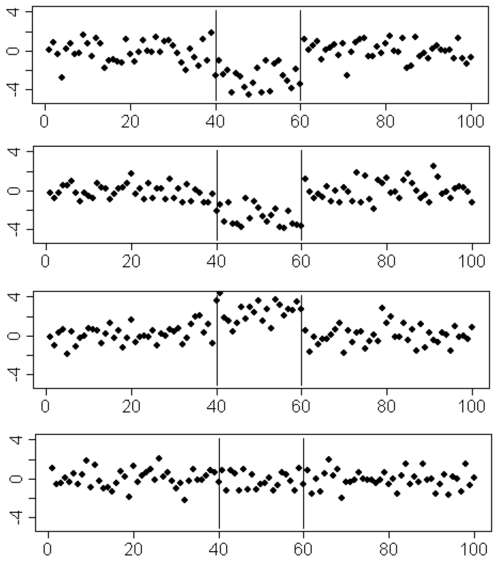

Fig. 1. Simulated data containing $N = 4$ profiles and $T = 100$ observations per profile. Horizontal axis is order of the data points and vertical axis is the y values. The two vertical lines delineate a changed segment, in which the top two samples have a lower mean, the third sample has a higher mean, and the fourth sample has no change.

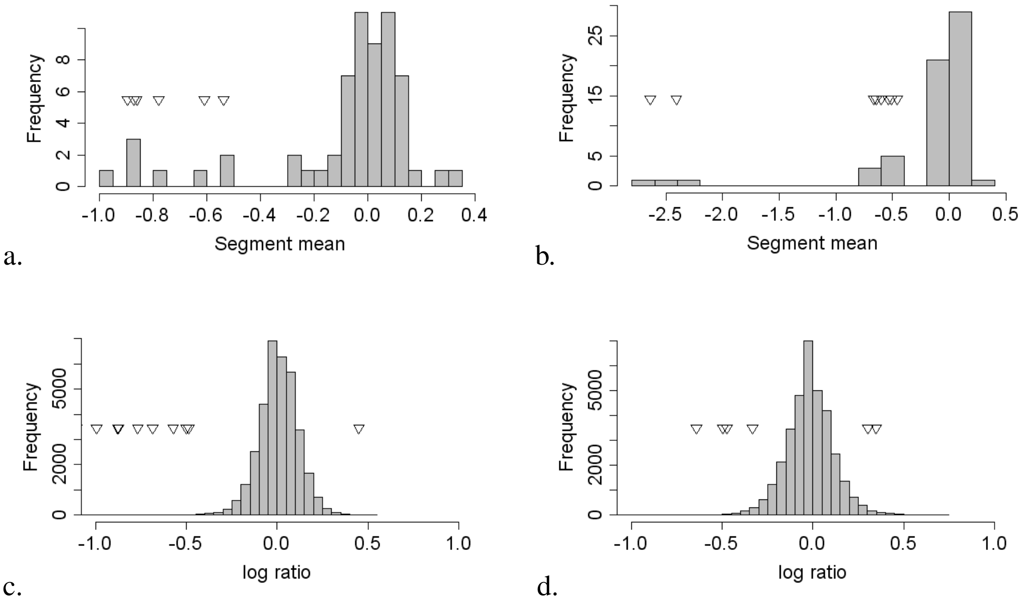

Fig. 2. Histograms (a,b) show the distribution across samples of the mean log ratio within two copy number variant regions. There are 62 samples, so each histogram represents the counts for 62 numbers. Both variants are deletion polymorphisms. The triangles show the estimated mean log ratio within those segments for the validated carriers among the samples. Observe that the triangles have a wide spread in values, suggesting that the model needs a separate mean shift for each sample within the same copy number variant. Figures (c,d) are histograms for $\{y_{it} : \quad t = 1, \dots, T\}$ for two different samples. The triangles show the estimated values of $\delta_i(\tau_1, \tau_2)$ for validated variant intervals τ_1, τ_2 on chromosome 5 for that sample. Observe again that the triangles have a wide spread in values, suggesting that the shift in mean is different across variant intervals within the same sample.

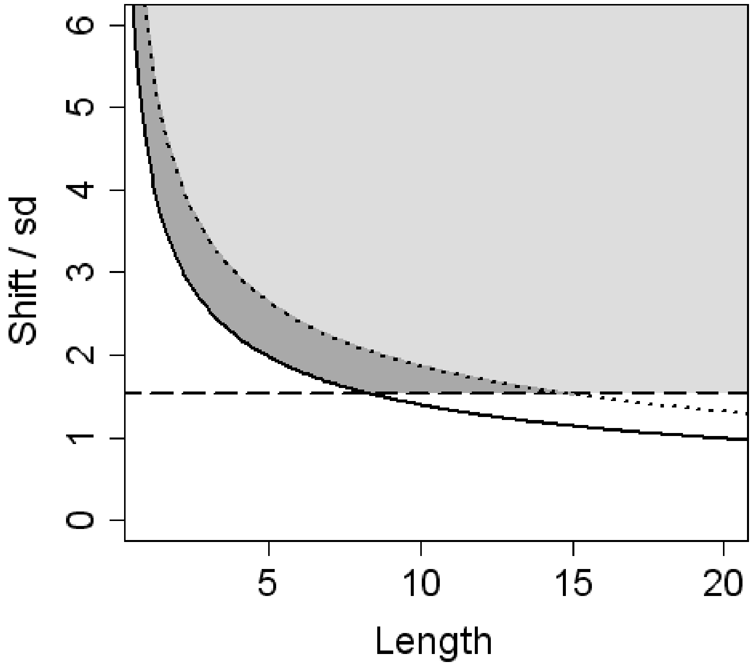

Fig. 3. The horizontal axis is the length of the segment and the vertical axis is the ratio of the shift in mean versus error standard deviation. Solid line shows the rejection boundary for a multi-sample scan with $N = 100$ samples. The dotted line is the rejection boundary for a single sample scan with $T = 500,000$ data points, $T_0 = 200$, and global p-value of 0.01. The dashed line shows the threshold $\delta_\mu = 1.5$. The light gray region shows the values of segment length $(\tau_2 - \tau_1)$ and effect size (δ_i/σ_i) that are classified as carrier for a detected variant interval. This region is determined by setting $\delta_{\chi^2} = 10^{-5}$ and $\delta_\mu = 1.55$. The dark gray region between the two boundaries contain those values that are missed in a single sample scan, but may be detectable in a multi-sample scan.

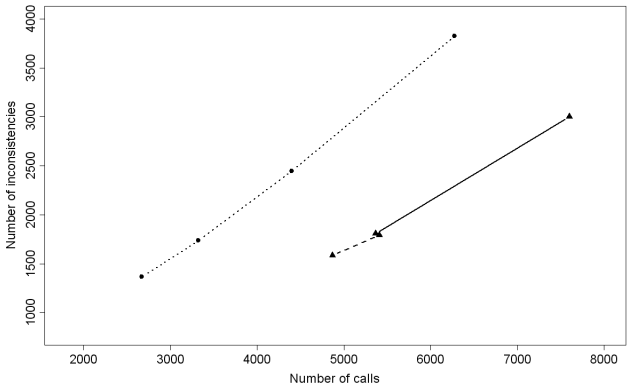

Fig. 4. Comparison of a single sample method (Olshen et al. 2003) with multiple sample scan on the quality assessment panel described in Section 3·2. The dotted line shows the total number of calls versus the number of inconsistent calls for results obtained using the single sample algorithm. The solid and dashed lines show the same information for results obtained using the multi-sample scanning algorithm. The global significance value is 10^{-3}. The sample calling thresholds for the multi-sample scan are $\delta_\mu \in \{0.2, 0.4\}$ and $\delta_{\chi^2} = 10^{-5}$ (solid line) and 10^{-8} (dashed line).

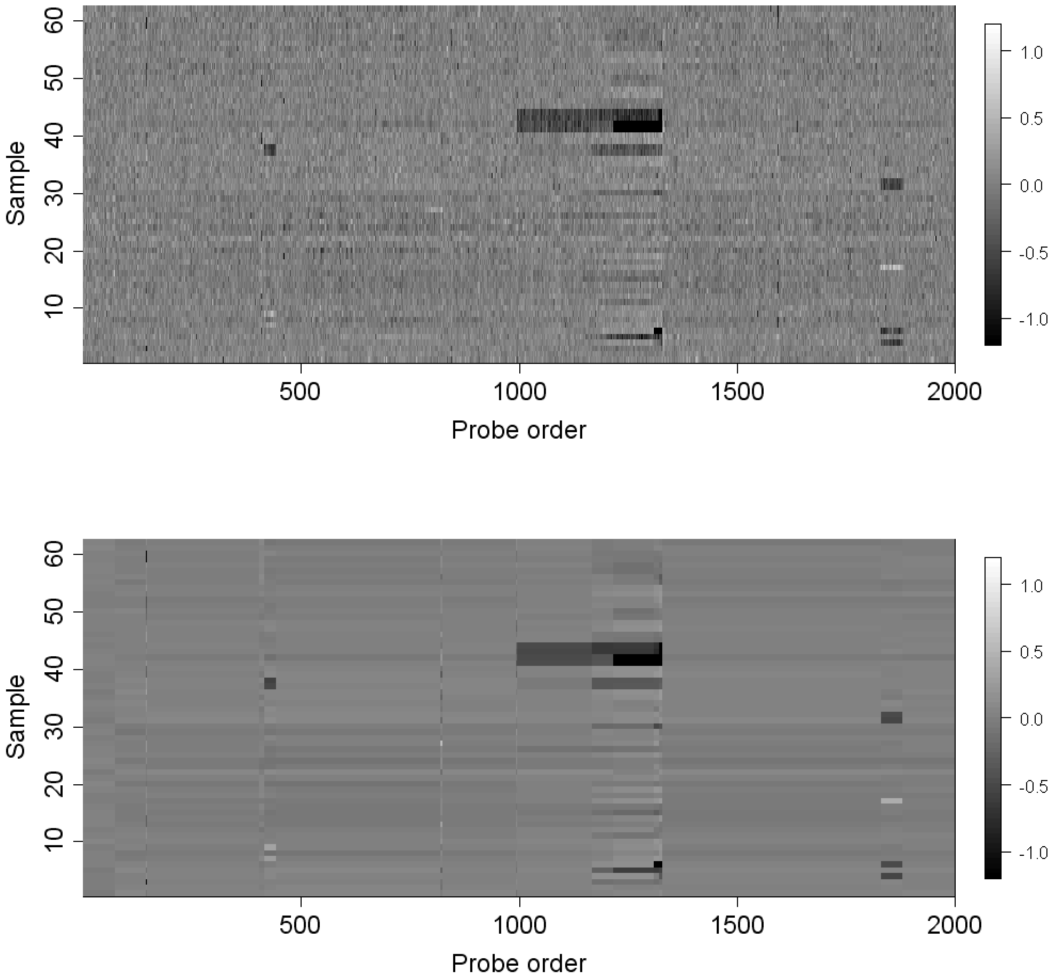

Fig. 5. Example 2000 marker region in cytoband 22q11 containing a complex copy number variant with nested deletions across 62 samples described in Section 3·3. Each row is a sample, and each column is a marker. The markers are ordered by their position along the chromosome. The grayscale shows the log intensity ratio, as indicated by the gradient bar on the right. Top panel shows the normalized, unsegmented data. Bottom panel shows segmentation given by the multi-sample circular binary segmentation algorithm.

[*Received Month* 0000. *Revised Month* 0000]