Submitted to the Annals of Applied Statistics

THE SCREENING AND RANKING ALGORITHM FOR CHANGE-POINTS DETECTION IN MULTIPLE SAMPLES

By Chi Song*, Xiaoyi Min* and Heping Zhang

Yale University

The chromosome copy number variation (CNV) is the deviation of genomic regions from their normal copy number states, which may associate with many human diseases. Current genetic studies usually collect hundreds to thousands of samples to study the association between CNV and diseases. CNVs can be called by detecting the change-points in the means from sequences of measurements with noise. Although multiple samples are of interest, the majority of the available CNV calling methods are single sample based. Only a few multiple sample methods were proposed. They all used scan statistics similar to the circular binary segmentation (CBS) algorithm that is computationally expensive, and were designed toward either common or rare change-points detection. In this paper, we propose a novel multiple sample method by adaptively combining the scan statistic of the screening and ranking algorithm (SaRa), which is computationally efficient and able to detect both common and rare change-points. We prove that asymptotically this method can find the true change-points with certainty and show in theory that multiple sample methods are superior to single sample methods when shared change-points are of interest. Additionally, we give extensive simulation studies and a real data application to examine the performance of our proposed method.

1. Introduction. The chromosome copy number refers to the number of copies of a genomic deoxyribonucleic acid (DNA) region. In the human genome, except for the sex chromosomes, the DNA copy numbers are normally two, with one copy from the mother and the other copy from the father. Copy number variation (CNV) can therefore be defined as the deviation from the "normal" copy number for a region of genomic DNA, which includes both duplication and deletion. In general, CNVs can be either generated from *de novo* mutation or inherited from the ascendants. *De novo* CNVs can possibly be long in length and unique for different individuals. For example, cancer CNVs as a type of *de novo* CNVs can span as long as a whole chromosome, and can be very heterogeneous across different patients.

*These authors contributed equally to this work.

Primary 62P10; secondary 92D10, 62F35

Keywords and phrases: change-point detection, multi-sample inference, adaptive Fisher's method

Inherited CNVs, on the contrary, are generally short in length, shared by many people, and aligned well across samples. Recent studies have shown that the CNVs can play important roles in human diseases. For example, *de novo* CNVs are found to be strongly associated with diseases such as autism (Sebat et al., 2007) and cancer (Pollack et al., 2002); while inherited CNVs are shown to be associated with Crohn's disease (McCarroll et al., 2008) and the resistance to HIV (Gonzalez et al., 2005). To study the association of CNV and human diseases, it is critical to identify CNV regions in each sample collected under the study. Over the last decade, tremendous amount of efforts have been made to study the CNVs by utilizing the high-throughput technologies, such as array-comparative genomic hybridization (aCGH), single-nucleotide polymorphism (SNP) array, and next-generation sequencing (NGS). Because the data produced by these technologies inevitably contain noise, various statistical methods have been proposed and applied to call CNV regions from the noisy data.

1.1. *Statistical model.* Regardless of the technology or platform, the copy number calling problem can be formulated in the following way. Given N samples and T markers, the raw CNV intensities are measured for each sample on all the markers. Denote the intensities measured for sample i by $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,T})^T$ for $1 \leq i \leq N$. We assume

$$\mathbf{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, \quad 1 \leq i \leq N, \tag{1.1}$$

where $\boldsymbol{\mu}_i = (\mu_{i,1}, \mu_{i,2}, \ldots, \mu_{i,T})^T$ is a piecewise constant mean vector for the intensities of sample i, and the errors $\boldsymbol{\epsilon}_i \sim \text{MVN}(\mathbf{0}, \sigma_i^2 \mathbf{I})$. Then τ is a change-point for sample i if $\mu_{i,\tau} \neq \mu_{i,\tau+1}$. We can further assume that for sample i, there are J_i change-points which we denote by $0 < \tau_{i,1} < \tau_{i,2} < \cdots < \tau_{i,J_i} < T$. The goal is to estimate all the change-points $\boldsymbol{\theta}_i = \{\tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,J_i}\}$ for each sample i. Then CNV regions can be called between these change-points.

Moreover, we denote the collection of all the change-points in all the samples as $\boldsymbol{\theta} = \{\tau_1 < \cdots < \tau_J\}$ and let $\delta_{i,j} = \mu_{i,\tau_j+1} - \mu_{i,\tau_j}$ be the mean change at point τ_j for sample i. For each change-point τ_j, we say that sample i is a "carrier" when $\delta_{i,j} \neq 0$. Note that estimating change-points for individual samples is equivalent to estimating $\boldsymbol{\theta}$ and identifying individual carriers of each change-point. The method that we propose in this paper is based on this strategy.

1.2. *Current methods.* Currently, various methods have been proposed for the CNV calling problem. These methods can be categorized into the single sample methods and the multiple sample methods according to the

strategies that they take. The single sample methods, on the one hand, assume that the change-points in different samples are completely irrelevant and independent, and apply the CNV calling algorithm to each of the individual samples repeatedly. The multiple sample methods, on the other hand, assume that certain change-points may be shared by a proportion of the samples, and call the shared change-points by utilizing information from multiple samples.

Because of the complexity of the multiple sample problem, most of the current methods focus on a single sample. Yao (1988) and Yao and Au (1989) proposed to search for the combination of change-points that minimize a BIC score, and they showed the consistency of the estimates. Circular binary segmentation (CBS) algorithm (Olshen et al., 2004; Venkatraman and Olshen, 2007) recursively finds segments with changed means, and is one of the commonly used single sample methods. The scan statistic of CBS given as follows is also widely adopted by other methods. Let $S_{i,t}$ be the partial sums of sequence $\mathbf{Y}_i$ (i.e. $S_{i,t} = \sum_{j=1}^{t} Y_{i,j}$), $\bar{Y}_i = S_{i,T}/T$, and $\hat{\sigma}_i^2 = \sum (Y_{i,j} - \bar{Y}_i)^2/T$. The scan statistic of CBS algorithm on a region (s,t) is defined as:

$$U_i(s,t) = \frac{(S_{i,t} - S_{i,s})/(t-s) - (S_{i,T} - S_{i,t} + S_{i,s})/(T-t+s)}{\hat{\sigma}_i \sqrt{1/(t-s) + 1/(T-t+s)}}. \tag{1.2}$$

The CBS algorithm is based on the test statistic $U_{i,C} = \max_{1 \le s < t \le T} U_i(s,t)$. Another approach uses the ℓ_1 penalization methods in order to introduce sparsity to the segment means or the differences in these means (Huang et al., 2005; Tibshirani and Wang, 2008). Niu and Zhang (2012) demonstrated that local information is more efficient than the global information for high-throughput data for change-points detection. They also proposed a new screening and ranking algorithm (SaRa) using the scan statistic

$$D_i(t,h) = \frac{1}{h}\Big(\sum_{k=1}^{h} Y_{i,t-k+1} - \sum_{k=1}^{h} Y_{i,t+k} \Big), \tag{1.3}$$

where h is the bandwidth. Because $D_i(t,h)$ is calculated from only the local information within the $2h$ window, the complexity of the algorithm is linear in the length of the sequence T. Other than the change-point models, other models such as the Hidden Markov Model (HMM) are also applied to the CNV problem. For example, PennCNV (Wang et al., 2007) and Birdsuite (Korn et al., 2008) are two of the most popular HMM methods. Because this paper focuses on the change-point models, we will not discuss the other models in detail.

It was pointed out that different people can share CNV regions (Zhang et al., 2010). In terms of the change-point model, some of the change-points

are shared by different samples, and hence several multiple sample methods have been developed to find those shared change-points. Zhang et al. (2010) proposed taking the sum of squares of the scan statistics from individual samples to find the common change-points. Siegmund, Yakir and Zhang (2011) further extended the method by using a weighted sum of squares statistic, which increases the power for rare change-point detection when prior information on the carrier proportions is available. Instead of using the sum-based statistics, Jeng, Cai and Li (2013) summarized the scan statistics based on the higher criticism method which can detect both common and rare CNVs (Cai, Jeng and Jin, 2011). It is noteworthy that the major difference among these multiple sample methods is the way that multiple scan statistics are combined. The scan statistics used by those methods for individual samples, however, are virtually the same as the statistic $U_i(s,t)$ of the CBS algorithm.

1.3. *Motivations.* Despite the success of the aforementioned methods, there are several aspects of them that either need to be addressed or could be improved upon. Firstly, the multiple sample methods that we reviewed all use the same scan statistics as in the CBS algorithm which is based on global information. As a result, these methods might be statistically powerful for using all the available information, but they all suffer from higher computational complexities, especially when applied to the high-throughput genomic data. To overcome this computational burden, we propose a generalization of SaRa to accommodate multiple samples. The proposed method enjoys similar computational efficiency and statistical properties as the single sample SaRa.

Secondly, we noted that the available methods for combining multiple scan statistics are either suitable for finding common change-points but not powerful in finding rare ones (in terms of the proportion of carriers), or vice versa, or rely on prior knowledge or assumption of the carrier proportion. Thus, it is desirable to develop a unified method that is robust to the change-point proportion and does not require any prior knowledge or assumption. For this purpose, we propose an adaptive Fisher's method which adaptively combines the scan statistics according to their likeliness of being from a change-point carrier. We show that, regardless of the carrier proportion, this method has a good power of finding change-points.

Lastly but not the least, despite the many multiple sample methods that have been proposed, a justification is needed yet missing to support their use instead of the single sample methods. In this paper, we provide both theoretical and numerical comparisons between our proposed method and

a single sample method, which show that the power of multiple sample methods is indeed higher than that of a single sample method.

This paper is organized as follows. In Section 2, we will present our method in detail. The theoretical properties of the proposed method are discussed in Section 3. The numerical results will be presented in Section 4, and a real data analysis example will be given in Section 5.

2. Method.

2.1. *SaRa for a single sample.* First, let us revisit the SaRa method proposed by Niu and Zhang (2012). For a single sample i, given a bandwidth h, the scan statistic $D_i(t,h)$ can be calculated for every position t from (1.3). Define t as a local maximizer if $|D_i(t,h)| \geq |D_i(t',h)|$ for all $t' \in (t-h, t+h)$. Let $\mathcal{LM}_i$ be the set of all local maximizers found for sample i. Then the change-points for sample i can be estimated as $\tilde{\boldsymbol{\theta}}_i = \{\tilde{\tau}_{i,1} < \tilde{\tau}_{i,2} < \cdots < \tilde{\tau}_{i,\tilde{J}_i}\} \subseteq \mathcal{LM}_i$ by a thresholding rule

$$|D_i(\tilde{\tau}, h)| > \lambda_i.$$

The threshold λ_i can be obtained asymptotically or from the simulated null distribution.

For any t, if there is no change-point in window $(t-h+1, t+h)$, it is easy to know that $D_i(t,h) \sim N(0, \frac{2}{h}\sigma_i^2)$. Therefore, we can define the standardized scan statistic as

$$\tilde{D}_i(t,h) = \sqrt{\frac{h}{2\hat{\sigma}_i}} D_i(t,h), \tag{2.1}$$

where $\hat{\sigma}_i$ is an estimate of σ_i. By assuming that the number of change-points in sample i, $J_i << T$, the estimation of $\hat{\sigma}_i$ is trivial. For example, we can use the sample standard deviation of $\mathbf{Y}_i$ as $\hat{\sigma}_i$.

2.2. *Combining test statistics from multiple samples.* In order to combine information from multiple samples to help identify shared change-points, we need to combine the single sample statistics across multiple samples. A natural choice is to take the sum of squares of $\tilde{D}_i(t,h)$ across samples as Zhang et al. (2010) did and define the multiple sample scan statistic

$$W^{Sum}(t,h) = \sum_{i=1}^{N} \tilde{D}_i^2(t,h). \tag{2.2}$$

The weighted sum of squares (Siegmund, Yakir and Zhang, 2011) is an alternative method, for which we define

$$W^{WSum}(t,h) = \sum_{i=1}^{N} w_{\pi_0}[\tilde{D}_i^2(t,h)]\tilde{D}_i^2(t,h), \tag{2.3}$$

where $w_{\pi_0}(x) = \exp(x/2)/[(1-\pi_0)/\pi_0 + \exp(x/2)]$, and π_0 is the carrier proportion that is assumed known.

The two methods above combine the scan statistics $\tilde{D}_i(t,h)$ directly. We can also define combining statistics based on the p-values $p_i(t,h) = 2\{1 - \Phi[|\tilde{D}_i(t,h)|]\}$ or their order statistics $p_{(i)}(t,h)$ in ascending order. Traditional methods include Fisher's method (Fisher, 1925) defined as

$$W^{Fisher}(t,h) = -\sum_{i=1}^{N} \log p_i(t,h), \tag{2.4}$$

and Stouffer's method (Stouffer et al., 1949)

$$W^{Stouffer}(t,h) = \sum_{i=1}^{N} \Phi^{-1}[1 - p_i(t,h)]. \tag{2.5}$$

The higher criticism statistic (Donoho and Jin, 2004; Cai, Jeng and Jin, 2011) can be defined as

$$W^{HC}(t,h) = \max_{1\le i\le N} |HC_i(t,h)|, \tag{2.6}$$

where

$$HC_i(t,h) = \sqrt{N}\frac{i/N - p_{(i)}(t,h)}{\sqrt{p_{(i)}(t,h)[1-p_{(i)}(t,h)]}}.$$

In practice, we are interested in finding change-points that are shared by either many of the samples or just a few of the samples; in other words, we would like to find both commonly and rarely occurring change-points. The sum of squares statistic, on the one hand, is naïve and easy to implement, but only good in capturing change-points that are shared by many samples. The higher criticism statistic, on the other hand, is able to detect rare change-points; however, because it is based on an adaptively chosen single order statistic, its power for detecting common change-points with a limited sample size is lower than the sum of squares statistic in practical applications. Although the weighted sum of squares statistic can detect both common

and rare change-points, it depends on the tuning parameter π_0 whose choice relies on the prior assumptions of the change-points. The Fisher's method is well-known for being powerful and asymptotically Bahadur optimal (Littell and Folks, 1971, 1973). However, when the change-points are rare, the statistical power of Fisher's method will be compromised by the non-carriers. The same problem also exists for Stouffer's method. Therefore, we propose a new summary statistic, which can detect both common and rare change-points and does not require prior knowledge or assumption.

The idea of our approach is to adaptively combine the p-values so that only the ones that most likely come from the carriers are combined. In the same spirit, Li and Tseng (2011) proposed an adaptively weighted Fisher's statistic to down-weight the non-carriers, but it is time consuming and involves exhaustive search for the weights. We propose a more concise adaptive Fisher's statistic as follow. For given t and h, let

$$X_i(t,h) = -\log p_i(t,h),$$

and

$$X_{(i)}(t,h) = -\log p_{(i)}(t,h).$$

We first define

$$V_i(t,h) = \sum_{j=1}^{i} X_{(j)}(t,h).$$

Under the null hypothesis, $X_i(t,h) \overset{iid}{\sim} \text{EXP}(1)$, and $X_{(1)}(t,h) \geq \cdots \geq X_{(N)}(t,h)$ are the decreasing ordered statistics. Let $X_{(N+1)}(t,h) = 0$ and $\xi_i(t,h) = i[X_{(i)}(t,h) - X_{(i+1)}(t,h)]$ for $1 \leq i \leq N$. It can be shown that $\xi_i(t,h) \overset{iid}{\sim} \text{EXP}(1)$ under the null. Thus

$$V_i(t,h) = \sum_{j=1}^{i}\sum_{k=j}^{N} \xi_k(t,h)/k = \sum_{k=1}^{N} w(k,i)\xi_k(t,h),$$

where $w(k,i) = \min(1, i/k)$. The standardized $V_i(t,h)$ can be calculated as

$$\tilde{V}_i(t,h) = \frac{V_i(t,h) - \sum_{k=1}^{N} w(k,i)}{\sqrt{\sum_{k=1}^{N} w^2(k,i)}}.$$

Our proposed adaptive Fisher's statistic for multiple samples is defined as

$$W^{AF}(t,h) = \max_{1\leq i\leq N} |\tilde{V}_i(t,h)|.$$

REMARK 1. *In the CNV detection problem, we are mainly interested in detecting signals arising from shifted means and/or increased variances. In other words, we believe that the p-values for signals are smaller than their null distribution. Therefore, the adaptive Fisher's statistic can also be defined as follow to (i) consider only the smaller half of the p-values and (ii) test whether p-values are smaller than expected,*

$$W^{AF}(t,h) = \max_{n_0 \le i \le N/2} \tilde{V}_i(t,h), \tag{2.7}$$

where n_0 is a tuning parameter to stabilize the statistic. Similarly, we could modify (2.6) into

$$W^{HC}(t,h) = \max_{n_0 \le i \le N/2} HC_i(t,h). \tag{2.8}$$

For the reason stated above, we apply (2.7) and (2.8) for CNV detection.

2.3. *SaRa for multiple samples.* In the previous section, we defined six scan statistics including $W^{Sum}(t,h)$, $W^{WSum}(t,h)$, $W^{Fisher}(t,h)$, $W^{Stouffer}(t,h)$, $W^{HC}(t,h)$ and $W^{AF}(t,h)$ for the multiple sample problem. By using one of those six statistics, we now extend the SaRa method for multiple samples. Let $\{W(t,h) : t = 1,\ldots,T\}$ be the sequence of combined statistics using any of the six combining methods with some bandwidth h. Then we can find the local maximizers of the sequence, and select a subset of the local maximizers by thresholding, which is the same technique used in the SaRa for single samples. The detailed algorithm is described as below.

ALGORITHM. *SaRa for multiple samples:*

1. *Given a bandwidth h, calculate individual scan statistics $\tilde{D}_i(t,h)$ using (2.1), for $1 \le t \le T$ and $1 \le i \le N$.*
2. *Calculate the summary scan statistic $W(t,h)$ using (2.2), (2.3), (2.4), (2.5), (2.8), or (2.7).*
3. *Find the set of local maximizers $\mathcal{LM} = \{t : W(t,h) > W(t',h), \forall t' \in (t-h, t+h)\}$.*
4. *Given a threshold λ, estimate the shared change-points as a subset of $\mathcal{LM}$, $\hat{\boldsymbol{\theta}} = \{\hat{\tau}_1 < \hat{\tau}_2 < \cdots < \hat{\tau}_{\hat{J}}\} \subseteq \mathcal{LM}$, that satisfies $W(\hat{\tau}_j, h) > \lambda$ for $1 \le j \le \hat{J}$, where $\hat{J}$ is the number of estimated shared change-points.*

REMARK 2. *In the calculation of $\tilde{D}_i(t,h)$, if $Y_{i,k}$ with $k < 1$ or $k > T$ is referred to, use $\bar{Y}_i$ instead. This only happens when t is near either end of a sequence.*

REMARK 3. *To determine the threshold λ, we can simply simulate the null distribution of $W(t,h)$ by assuming that $\mathbf{Y}_i \overset{iid}{\sim} MVN(\mathbf{0},\mathbf{I})$ for $1 \le i \le N$. Because $W(t,h)$ is calculated locally and $T >> h$, we can simulate the null distribution of $W(t,h)$ using any length T' that satisfies $T' >> h$. Let $\hat{F}(\cdot)$ be the simulated empirical distribution function of $W(t,h)$, where t is a local maximizer (or any point in the sequence). Given a significance level α, the threshold can be calculated as $\lambda = \hat{F}^{-1}(1-\alpha)$.*

2.3.1. *Multiple-bandwidths SaRa.* The selection of bandwidth h may affect the result. As described by Niu and Zhang (2012), a larger h may increase the statistical power. However, if h is too large such that more than one change-points are included in the window, the algorithm will yield unreliable results. In practice, we use multiple bandwidths to ease this difficulty. Given a handful of bandwidths $\mathbf{h} = \{h_1 < h_2 < \cdots < h_B\}$, where B is the number of different bandwidths, we can estimate the change-points using each of the bandwidth in $\mathbf{h}$, and get $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}, \ldots, \hat{\boldsymbol{\theta}}^{(B)}$. Then the candidates for the shared change-points are estimated by $\hat{\boldsymbol{\theta}} = \bigcup_{b=1}^{B} \hat{\boldsymbol{\theta}}^{(b)}$. However, since different bandwidths may yield change-points with slightly different positions, some change-points in $\hat{\boldsymbol{\theta}}$ may be redundant. Besides, the included change-points may not be substantiated. Such points will be excluded as described in Section 2.3.2.

2.3.2. *Change-point carrier identification.* Recall that the shared change-points are detected through the summary scan statistics. Consequently, we do not know which individuals carry a particular change. Hence, it is necessary and useful to identify those carriers of a given change-point. A simple approach is to test the means on two sides of a candidate change-point, but as discussed by Zhang et al. (2010), the existence of trends that are unrelated to the change-point could cause slight shifts in the local means along the chromosome, making it difficult to differentiate the real change-point from a shift caused by a trend. This can be resolved by thresholding as follows for a given sample i and the candidate change-points $\hat{\boldsymbol{\theta}} = \{\hat{\tau}_1 < \hat{\tau}_2 < \cdots < \hat{\tau}_{\hat{J}}\}$.

ALGORITHM. *Carrier identification:*

1. *Set $\hat{J}_i = \hat{J}$ and $\hat{\tau}_{i,j} = \hat{\tau}_j$ for $j = 1, \ldots, \hat{J}_i$. Denote $\hat{\boldsymbol{\theta}}_i = \{\hat{\tau}_{i,j}, j = 1, \ldots, \hat{J}_i\}$.*
2. *Let $\hat{\tau}_{i,0} = 0$ and $\hat{\tau}_{i,\hat{J}_i+1} = T$. Calculate the segment means $m_{i,j} = \frac{\sum_{t=\hat{\tau}_{i,j}+1}^{\hat{\tau}_{i,(j+1)}} Y_{i,t}}{\hat{\tau}_{i,(j+1)} - \hat{\tau}_{i,j}}$ for $0 \le j \le \hat{J}_i$.*
3. *Calculate the estimated jump size at each change-point $d_{i,j} = m_{i,j} - m_{i,(j-1)}$ for $1 \le j \le \hat{J}_i$.*

4. *Find the smallest absolute jump size* $j^* = \arg\min_{1\le j\le \hat{J}_i} |d_{i,j}|$. *If* $|d_{i,j^*}|$ *is less than a pre-specified threshold* η, *remove the* j^**-th change-point by replacing* $\hat{\boldsymbol{\theta}}_i$ *with* $\hat{\boldsymbol{\theta}}_i \backslash \{\hat{\tau}_{i,j^*}\}$ *and replacing* $\hat{J}_i$ *with* $\hat{J}_i - 1$, *and then repeat the procedure from step 2; otherwise, estimate all the individual change-points for sample* i *by* $\hat{\boldsymbol{\theta}}_i$.

REMARK 4. *We suggest using* $\eta = 2\hat{\sigma}_i\sqrt{2/h}$ *in practice. For multiple-bandwidths SaRa, we suggest using* $\eta = 2\hat{\sigma}_i\sqrt{2/\min\{\mathbf{h}\}}$.

REMARK 5. *If no individual carrier is identified for a particular change-point, we will remove this change-point from the shared set* $\hat{\boldsymbol{\theta}}$, *further improving the precision and reliability of* $\hat{\boldsymbol{\theta}}$.

3. Statistical properties. In this section, we show the theoretical properties of multiple sample SaRa from two aspects. First, we prove the sure coverage property as the sample size N increases. That is to say, the union of h-neighborhoods of the elements in $\hat{\boldsymbol{\theta}}$ estimated as in Section 2.3 covers all the true change-points in $\boldsymbol{\theta}$ with probability tending to one. Second, we show that SaRa for multiple samples is advantageous to a single sample method in the sense that its asymptotic power for detecting change-points is higher.

Throughout this section, we assume that the sequence length T and the set of change-points $\boldsymbol{\theta} = \{\tau_1, \ldots, \tau_J\}$ are fixed. For convenience in notation, we denote $\tau_0 = 0$ and $\tau_{J+1} = T$, and we let $L = \min_{1\le j\le J+1}(\tau_j - \tau_{j-1})$. Recall that we denote as $\delta_{i,j}$ the mean change of sample i at τ_j. Here, we assume for simplicity that, for each $1 \le j \le J$, $\delta_{1,j}, \ldots, \delta_{N,j}$ are independent and

$$\delta_{i,j} \begin{cases} = 0, & \text{with prob. } (1-\pi_j), \\ \sim N(\Delta_j, (\eta_j^*)^2), & \text{with prob. } \pi_j, \end{cases}$$

where $\pi_j > 0$, Δ_j, and $(\eta_j^*)^2$ are fixed and assumed known. This corresponds to a practical scenario that the carriers of a common CNV constitute a certain proportion of the population and the mean intensity change in the CNV region may vary for each carrier. We also assume that $\sigma_1^2, \ldots, \sigma_N^2$ are known, then without loss of generality, we can assume that they are all equal to 1. Moreover, following Niu and Zhang (2012), we call a point t h-flat if there is no change-point in the interval $(t-h, t+h)$. Then we have

$$\tilde{D}_i(t,h) \sim \begin{cases} N(0,1), & \text{if } t \text{ is } h\text{-flat}, \\ (1-\pi_j)N(0,1) + \pi_j N(-\Delta_j\sqrt{h/2}, \eta_j^2), & \text{if } t = \tau_j, \end{cases}$$

where $\eta_j^2 \equiv (\eta_j^*)^2 + 1$.

THEOREM 1. *Using SaRa for multiple samples with any of the following combining methods:* W^{Sum}, W^{WSum}, W^{Fisher}, $W^{Stouffer}$, W^{HC}, *and* W^{AF}, *there exist suitable* h *and* λ *such that the estimated change-points* $\hat{\boldsymbol{\theta}}$ *satisfy*

$$\lim_{N\to\infty} P(\{\hat{J} = J\} \cap \{\boldsymbol{\theta} \subset \hat{\boldsymbol{\theta}} \pm h\}) = 1,$$

where $\hat{\boldsymbol{\theta}} \pm h \equiv \bigcup_{j=1}^{\hat{J}} (\hat{\tau}_j - h, \hat{\tau}_j + h)$.

The previous theorem states that a threshold λ exists to ensure the sure coverage property of SaRa for multiple samples. However, the choice of such a threshold depends on the underlying truth which is generally unknown. Therefore, in practice, the threshold is usually chosen so that at a flat-point or at a local miximizer, the scan statistic goes above the threshold with a certain probability, say α. We show in the next theorem that the "power" of detecting a true change-point, in other words the probability that the scan statistic at a true change-point exceeds this threshold, tends to 1 pretty fast.

In comparison, we consider a naïve single sample procedure that calls change-points in single samples first and then combines the obtained change-points in all the samples. In other words, for some λ^*, whenever $|\tilde{D}_i(t,h)| > \lambda^*$ for any i, we claim that t is a change-point for sample i and thus a common change-point. This is equivalent to using the maximum statistic of $\{\tilde{D}_i(t,h)\}_{i=1}^N$ and claiming that there is a change-point at t when $\max_i |\tilde{D}_i(t,h)| > \lambda^*$. Note that due to multiplicity, controlling the false positive rate for individual samples is not enough. Instead, we need to choose λ^* such that $P(\max_i |\tilde{D}_i(t,h)| > \lambda^*) = \alpha$ for an h-flat point t. We show in the following theorem that the power of detecting a true change-point tends to 1 at a rate slower than the multiple sample methods.

THEOREM 2. *(a) Use SaRa for multiple samples with any of the following combining methods:* W^{Sum}, W^{WSum}, W^{Fisher}, $W^{Stouffer}$, W^{HC}, *and* W^{AF}, *and choose the threshold* λ *such that for an* h*-flat point* t *we have* $P(W(t,h) > \lambda) = \alpha$ *with a specific level* α. *Then for any* $j = 1, \ldots, J$, $P(W(\tau_j, h) > \lambda)$ *tends to* 1 *at least at an exponential rate in* N.

(b) Use the single sample procedure that claims a common change-point at t *when* $\max_i |\tilde{D}_i(t,h)| > \lambda^*$ *where* λ^* *is chosen such that* $P(\max_i |\tilde{D}_i(t,h)| > \lambda^*) = \alpha$ *for an* h*-flat point* t. *Then for any* $j = 1, \ldots, J$, $P(\max_i |\tilde{D}_i(\tau_j, h)| > \lambda^*) \to 1$ *as* $N \to \infty$ *but with a rate slower than the exponential rate in* N.

REMARK 6. *We shall point out that the convergence rate for the single sample method in Part (b) of the theorem depends on* η_j^2. *The convergence is slower for smaller* η_j^2. *At the extreme case when* $\eta_j^2 = 1$, *i.e. when the mean*

changes for carriers of a change-point are fixed, the convergence gets much slower.

The two theorems in this section only discuss the scenario that, when subject i is a carrier of τ_j, $\tilde{D}_i(\tau_j, h)$ has an increased variance ($\eta_j^2 \geq 1$) and/or a changed mean. Some of the results can be generalized to the scenario that η_j^2 might be less than 1. For example, the results for W^{HC} and W^{AF} still hold, whereas the other multiple sample methods could work but need to be modified such that both increased and decreased mean in the scan statistics can be detected. However, for the single sample method, the results are very different. The power of detecting τ_j as in Theorem 2 falls below α when $\eta_j^2 < 1$, suggesting that the single sample method is quite sensitive to the model assumptions. This is another reason why the multiple sample methods are favorable to the single sample method.

4. Numerical result.

4.1. *Power for detecting a single change-point.* To study the power of SaRa for multiple samples, we simulated simple datasets with only one change-point shared by a proportion of the samples. The datasets were simulated in the following procedure.

1. *Let N be the number of samples, T be the length of the sequence, δ be the jump size, and π^* be the proportion of samples that carry the change-point.*
2. *For $1 \leq i \leq \lceil N\pi^* \rceil$, sample $Y_{i,j} \overset{iid}{\sim} N(0,1)$ if $1 \leq j \leq T/2$, and sample $Y_{i,j} \overset{iid}{\sim} N(\delta, 1)$ if $T/2 < j \leq T$. Here, $\lceil \cdot \rceil$ is the ceiling function.*
3. *For $\lceil N\pi^* \rceil < j \leq N$, sample $Y_{i,j} \overset{iid}{\sim} N(0,1)$ for $1 \leq j \leq T$.*

Different combining scan statistics were considered (for W^{WSum}, we set $\pi_0 = 0.01$; for W^{HC} and W^{AF}, $n_0 = 4$ was used). A shared change-point was called when at least one local maximizer of the scan statistics falls between $50-h$ and $50+h$, and exceeds the 99% quantile of the null distribution of the local maximizers. The simulation results are a summary of 1000 replications.

To demonstrate how the power changes according to N when detecting both rare and common change-points, we simulated two scenarios with $N \in \{100, 200, \ldots, 1000\}$. For the rare change-point scenario, we set $\pi^* = 0.01$, $\delta = 1$, and $h = 20$; for the common change-point scenario, we set $\pi^* = 0.2$, $\delta = 0.5$, and $h = 10$. The parameters were selected to enhance the differences between methods.

Figure 1(a) compares the power of different methods for detecting a rare change-point with carrier proportion $\pi^* = 0.01$. As expected, the sum of

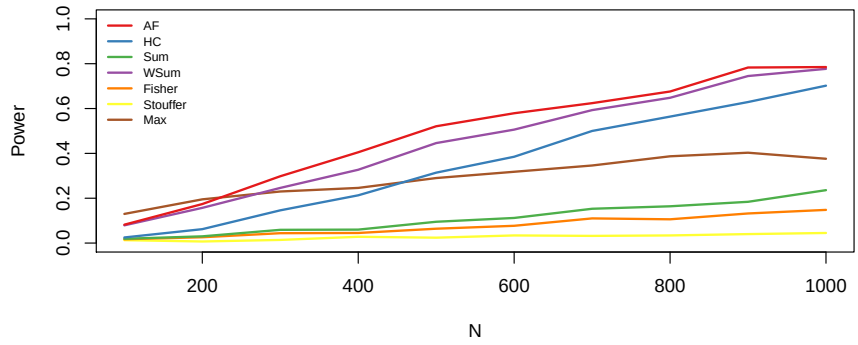
squares statistic, Fisher's statistic, and Stouffer's statistic have the lowest statistical power, because they combine all of the scan statistics which contain a large proportion (99%) of noises. On the contrary, using the maximum test statistic as an extension of the single sample methods as described in Section 3 enjoys a reasonable statistical power. However, its power increases very little as N increases, because only the single strongest test statistic is used, which is a waste of information. This result is consistent with the theoretical conclusion of Theorem 2. Similar to the observation in Jeng, Cai and Li (2013), the higher criticism statistic has a relatively good statistical power in detecting rare signals, and the power increases as N increases. Our proposed adaptive Fisher's statistic performs the best among the methods under comparison. Even though the prior information $\pi_0 = 0.01$ is correctly specified for the weighted sum of squares statistic, its power is slightly lower than that of the adaptive Fisher's method.

Figure 1(b) compares those methods in terms of the power for detecting a common change-point with carrier proportion $\pi^* = 0.2$. As expected, the sum of squares statistic and Fisher's statistic have the best statistical power. Our adaptive Fisher's statistic and Stouffer's statistic perform similarly with slightly lower power. Weighted sum statistic, higher criticism statistic, and maximum statistic have the lowest power. Similar to the rare change-point case, the maximum statistic does not benefit much from the increase in the sample size.

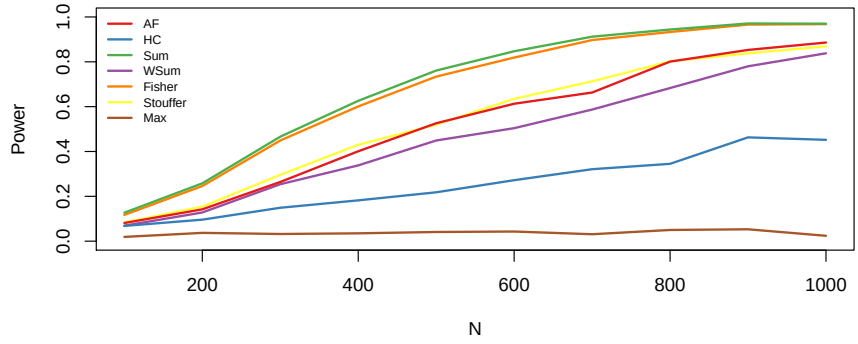
To display the power of different methods as π^* changes, we also simulated data using $\pi^* \in \{0, 0.01, 0.02, \ldots, 0.25\}$, $N = 100$, and $\delta = 1$. Moreover, to illustrate how the adaptive Fisher's statistic and higher criticism statistic adapt to different carrier proportions, we calculated the peak positions of these two statistics as π^* changes, which are the maximizer indices of equations (2.7) and (2.8) divided by N.

Figure 2(a) shows the power of different methods with bandwidth $h = 10$. Similar to our previous observation, maximum statistic, higher criticism, and weighted sum of squares statistic only perform well for small π^*, whereas the sum of squares and Fisher's statistics only perform well for large π^*. Stouffer's statistic performs good only when π^* gets close to 0.25, which is due to its well-known property of "robustness" against a few "outliers". Only our adaptive Fisher's statistic enjoys competitive statistical power no matter π^* is small or large. To illustrate how the adaptive Fisher's statistic works, we show in figure 2(b) the average peak positions of our adaptive Fisher's statistic and the higher criticism statistic, which can be interpreted as the proportions of scan statistics that contribute to the combined statistics. We can see that the proportion of scan statistics that contributes to the

adaptive Fisher's statistic tends to increase as π^* increases. This trend is even stronger when the larger bandwidth $h = 20$ is used. On the contrary, the higher criticism method tends to select a much smaller proportion of p-values to combine, which can partly explain why it is under-powered and does not perform well enough when π^* gets large.

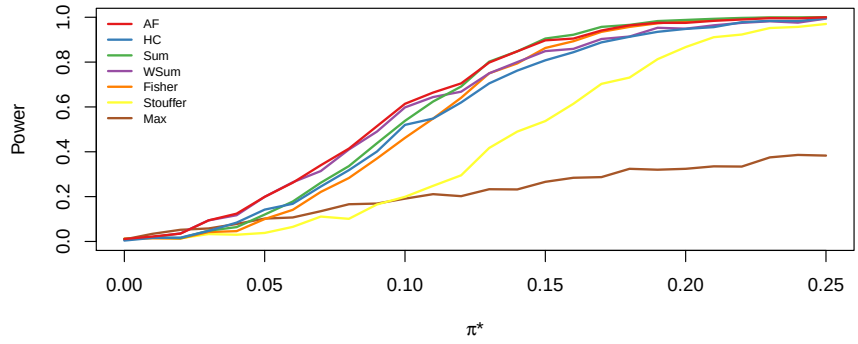


(a) Power for detecting a single rare change-point.

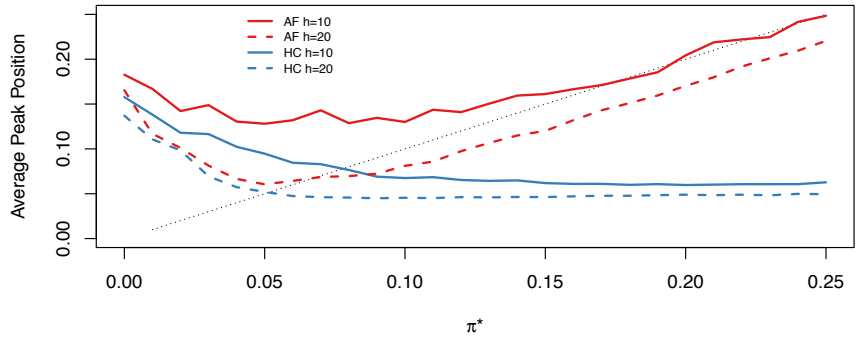


(b) Power for detecting a single common change-point.

FIG 1. *Power of different methods for detecting a single rare or common change-point as* N *changes from* 100 *to* 1000. *In (a), a single rare* ($\pi^* = 0.01$) *change-point was simulated and detected using* $\delta = 1$ *and* $h = 20$*; in (b), a single common* ($\pi^* = 0.2$) *change-point was simulated and detected using* $\delta = 0.5$ *and* $h = 10$.



(a) Power of different methods as π^* changes.



(b) Average adaptive peaks of adaptive Fisher's method and higher criticism as π^* changes.

FIG 2. *Simulation result for single change-point detection as* π^* *changes from* 0 *to* 0.25. $N = 100$ *and* $\delta = 1$ *were used for the simulation. The powers of the seven combining methods (with* $h = 10$*) are compared in (a). The average adaptive peak position of adaptive Fisher's statistic and higher criticism statistic are compared in (b), where the dotted line shows the true proportion of sample carriers.*

4.2. *Simulation with multiple changes.*

4.2.1. *Data without trend.* We further simulated data from a more realistic model to compare our method and some existing ones. In each of the 1000 replications, we simulated a dataset of 500 SNPs and 1000 samples.

The detailed simulation procedure is described below.

1. *First, simulate the mean signal* $\boldsymbol{\mu}_i$ *without noise. For* $1 \leq i \leq 1000$ *and* $1 \leq t \leq 500$*, and set* $\mu_{i,t}$ *to* 0 *except for the following change-regions in their carriers.*
   (a) *Region 1:* $28 \leq t \leq 54$ *(length is* 27*), set* $\mu_{i,t} = \delta_1 = 2.58$ *if sample* i *is a carrier, the carrier proportion* $\pi_1 = 0.02$*.*
   (b) *Region 2:* $116 \leq t \leq 130$ *(length is* 15*), set* $\mu_{i,t} = \delta_2 = -1.92$ *if sample* i *is a carrier, the carrier proportion* $\pi_2 = 0.05$*.*
   (c) *Region 3:* $222 \leq t \leq 306$ *(length is* 85*), set* $\mu_{i,t} = \delta_3 = 1.74$ *if sample* i *is a carrier, the carrier proportion* $\pi_3 = 0.1$*.*
2. *Add random noise to the mean signal. Simulate* $\mathbf{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i$ *for* $1 \leq i \leq 1000$*, where* $\boldsymbol{\epsilon}_i \sim MVN(\mathbf{0}, \mathbf{I})$*.*

Figure 3 displays five representative examples of individual sequences, and there are a total of 6 unique and shared change-points. We compared five methods: a fast implementation of CBS (fast-CBS) from Venkatraman and Olshen (2007), CBS with *post hoc* subset selection for the change-points using BIC (CBS-SS), multiple-bandwidth SaRa for single samples (m-SaRa), multiple-sample CBS (Zhang et al., 2010), and our proposed method (multiple-sample m-SaRa, $\alpha = 0.001$ was used when determining λ).

Table 1 presents the number of shared change-points detected by each of the five methods. Multiple-sample CBS and our method correctly detected exactly 6 change-points in all replications. Tables 2 provides the details of the performance for each method (by row) in detecting each change-point (by column). Tables 2(a) and 2(b) offer the average numbers of true and false positives for each of the six change-points, respectively. Because the single sample methods do not detect the change-point positions as accurately as the multiple sample methods, for the single sample methods, we treat the change-point as if it is a true positive provided that it falls in a small neighborhood of the true position. From these tables, we can see that our proposed method performed the best in terms of both sensitivity and specificity among the five methods.

4.2.2. *Data with trend.* We now evaluate how the change-point detection could be affected by underlying trends in the data that are unrelated to change-points. To this end, we simulated data by introducing a systematic trend. Otherwise, the rest of the simulation procedure was the same as that in Section 4.2.1. Specifically, we simulated $\mathbf{Y}_i$ by adding both random noise

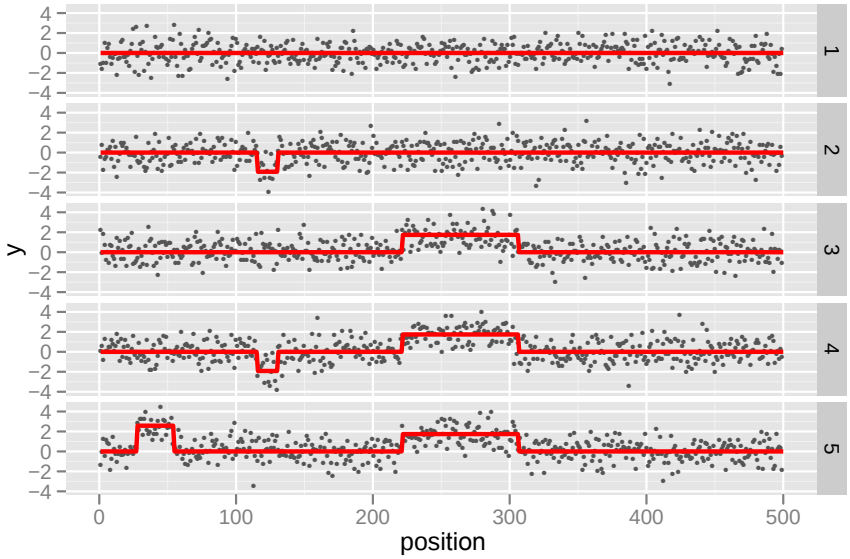


FIG 3. *The simulated data with no trend. Five samples are shown. The mean signals without noise are shown by bold red lines.*

TABLE 1
The number of shared change-points detected for the simulation with no trend.

	Number of change-points				
Method	≤ 5	**6**	7	8	> 8
fast CBS	0	**481**	395	108	16
CBS-SS	0	**524**	376	90	10
m-SaRa	0	**0**	0	0	1000
Multiple-sample CBS	0	**1000**	0	0	0
Multiple-sample m-SaRa	0	**1000**	0	0	0

and trend as follows

$$Y_{i,t} = \mu_{i,t} + 0.12\sin(2\pi t/96 + \psi) + 0.24\sin(2\pi t/240 + \phi_i) + \epsilon_{i,t},$$

where $\psi \sim U(0, 2\pi)$, $\phi_i \sim U(0, 2\pi)$, and $\epsilon_{i,t} \sim N(0, 1)$. In this model, the term $0.12\sin(2\pi t/96+\psi)$ is the trend shared by all samples, and $0.24\sin(2\pi t/240+\phi_i)$ is the trend unique for each sample. Five representative sequences are given in Figure 4. Similar to Table 1, Table 3 presents the number of shared change-points detected. We can see that the trends exercised great impact on the single sample methods. Multiple-sample CBS also yielded a notable number of false change-points. In fact, we considered various scenarios with

TABLE 2
True and false positives grouped by the change-points (CP1-CP6) for the simulation with no trend.

(a) Average number of true positives.

	CP1	CP2	CP3	CP4	CP5	CP6
Number of Carriers	20	20	50	50	100	100
fast CBS	19.9	19.9	47.6	47.6	94.5	94.5
CBS-SS	19.9	19.9	47.5	47.6	94.5	94.5
m-SaRa	19.8	19.8	47.3	47.2	92.0	91.8
Multiple-sample CBS	20.0	20.0	50.0	50.0	100.0	100.0
Multiple-sample m-SaRa	20.0	20.0	50.0	50.0	100.0	100.0

(b) Average number of false positives.

	CP1	CP2	CP3	CP4	CP5	CP6
fast CBS	0.3	0.3	0.3	0.3	0.2	0.2
CBS-SS	0.2	0.2	0.2	0.2	0.2	0.2
m-SaRa	2.9	4.1	4.7	5.1	5.0	5.1
Multiple-sample CBS	2.7	2.3	2.8	2.7	1.3	1.2
Multiple-sample m-SaRa	0.7	0.1	0.3	0.3	0.0	0.0

different period and magnitude of the trends. Not surprisingly, the performance of multiple-sample CBS, and of course the single sample methods, became worse. Fortunately, our multiple-sample m-SaRa procedure performed robustly, detecting all 6 true change-points in 997 out of 1000 replicates. An intuitive explanation is that the CBS scan statistic uses global information and thus cannot distinguish between a large scale trend and a real changed region, whereas SaRa scan statistic look for sharp mean change using local information, which makes it immune from the influence of a large scale trend.

TABLE 3
The number of shared change-points detected for the simulation with trend.

	Number of change-points				
Method	≤ 5	**6**	7	8	> 8
fast CBS	0	**0**	0	0	1000
CBS-SS	0	**0**	0	0	1000
m-SaRa	0	**0**	0	0	1000
Multiple-sample CBS	3	**749**	231	8	9
Multiple-sample m-SaRa	0	**996**	4	0	0

4.2.3. *Data with dependent errors.* Even though our method and most other methods assume independent errors, we would like to simulate situations with correlated errors and find out how robust the methods are when

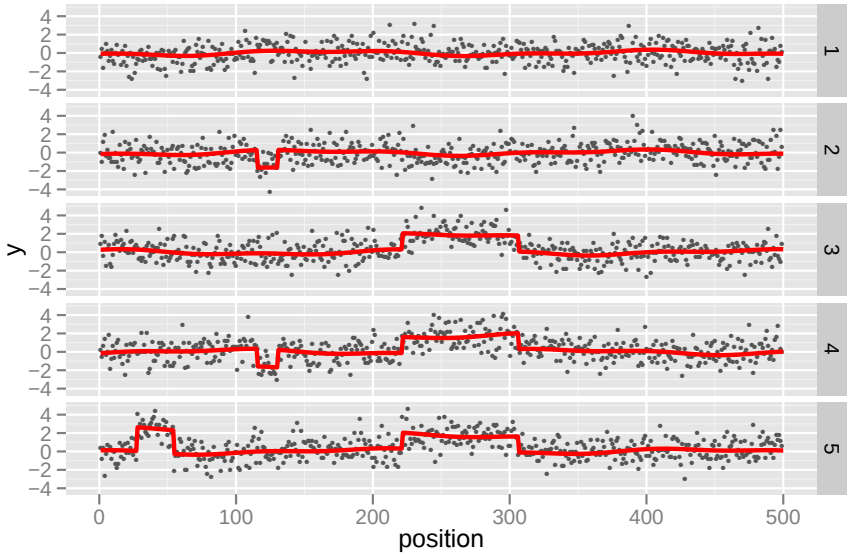


FIG 4. *The simulated data with trend. Five samples are shown. The mean signals without noise are shown by bold red lines.*

the independent error assumption is violated.

To evaluate the effect of dependent errors, we simulated data using the same mean model as specified in Section 4.2.1. Instead of assuming $\boldsymbol{\epsilon}_i \sim \text{MVN}(0, \mathbf{I})$, we simulated $\boldsymbol{\epsilon}_i$ from a moving average model of order 20, with all the 20 parameters equal 0.01. Similarly, Table 4 shows the number of shared change-points detected. Compared with Table 1, we notice that the correlated errors increased the false positive in single sample methods but did not affect the performance of the multiple sample methods. An explanation similar to Section 4.2.2 applies here, because the effects of correlated errors on each sample is similar to random short trends.

TABLE 4
The number of shared change-points detected for the simulation with dependent errors.

	Number of change-points				
Method	≤ 5	**6**	7	8	> 8
fast CBS	0	**31**	93	229	647
CBS-SS	0	**137**	312	314	237
m-SaRa	0	**0**	0	0	1000
Multiple-sample CBS	0	**1000**	0	0	0
Multiple-sample m-SaRa	0	**1000**	0	0	0

5. Real data analysis. To demonstrate the usage of our proposed method, we applied our method to the Children's Hospital of Philadelphia (CHOP) CNV data (Shaikh et al., 2009). The raw Log R Ratio (LRR) data for $2,026$ healthy children from CHOP was downloaded from dbGaP. To obtain biologically meaningful results, we mapped the probes to the most current genome build GRCh37.p10. Multiple-sample m-SaRa were ran using bandwidth of 5, 10, and 15, and $\alpha = 10^{-4}$ was used to determine the cutoff λ.

At the end, we identified $36,518$ unique change-points, among which 79.3% are shared by multiple individuals. Figure 5 shows the distribution of the carrier proportions of the identified change-points. About 94.1% of the change-points are carried by less than 5% of the individuals. To further confirm the length of identified CNV regions, we calculated the number of SNP markers between each pair of two consecutive change-points identified in each individual. The distribution of the number of markers are shown in Figure 6. Note that if we assume that the CNV regions are separated by normal regions, about half of the regions between consecutive change-points are CNV regions that should span only small numbers of SNP markers because the individuals are all healthy subjects. As expected, 50.2% of the regions between our identified change-points span across no more than 30 SNP markers.

6. Discussion. Although CNV has been studied for more than a decade, multiple sample based calling methods had not been proposed until recent years. In practice, single sample methods are still dominating. This is partly due to the lack of systematical research comparing multiple sample methods and single sample methods. In this study, we have shown that in terms of shared change-point detection, single sample methods are equivalent to taking the most significant statistic across samples, which is under-powered and sometimes does not work. Therefore, to achieve biologically meaningful detection power, specificity has to be sacrificed in single sample method, which inevitably increases the number of false positives. This approach is a waste of information across samples, especially with the growth of studies with large sample sizes. To the contrary, multiple sample methods combine evidences from multiple samples to detect shared change-points, which boosts the statistical power and hence reduces the false positives. Theoretically, we have proven that the power of multiple sample methods always converges to 1 at an exponential rate in the number of samples, which is faster than single sample methods. This can also be seen in our simulation.

Instead of using the CBS-like scan statistic, we used the SaRa scan statis-

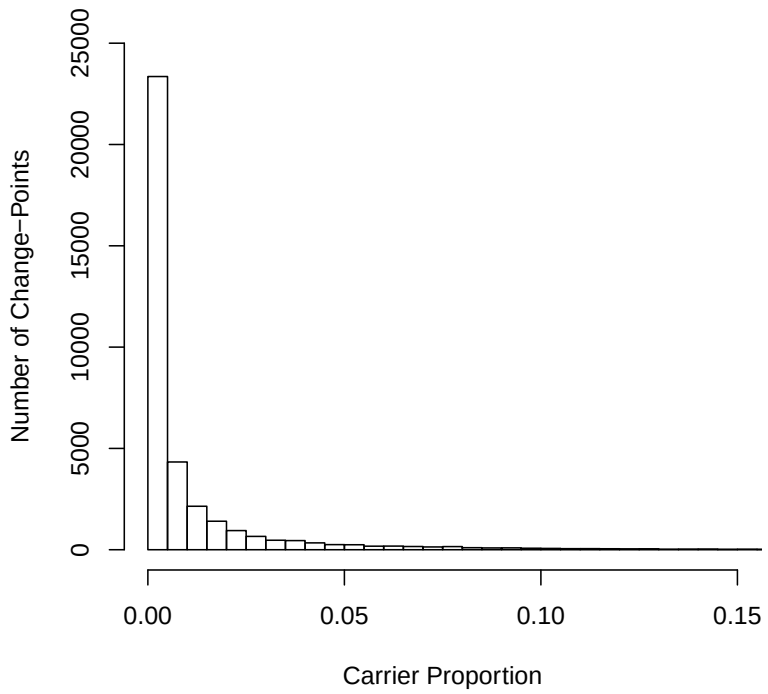
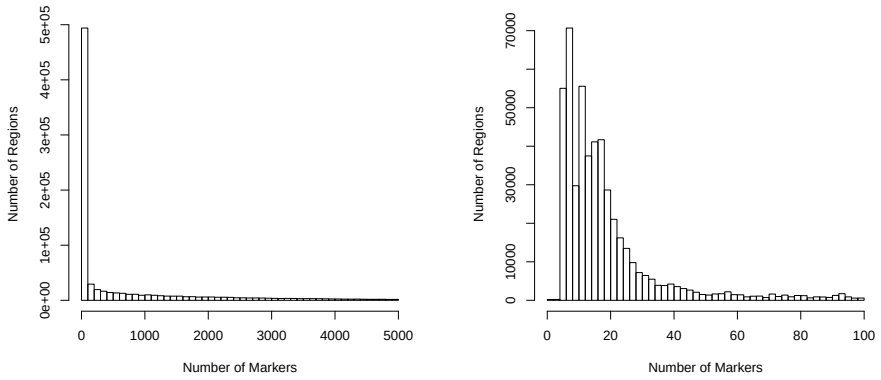


FIG 5. *The histogram of the carrier proportion of individual change-points identified in CHOP data.*

tic in our method. The SaRa statistic utilizes the local information, which can significantly speed up the computation. Because SaRa scan statistic uses a moving window, the computation complexity is linear in the number of markers T. Sorting is also needed in combining multiple samples using adaptive Fisher's method, thus the overall complexity of our proposed method is $\mathrm{O}(TN \log N)$. In practice $T >> N$, our method is much more computationally efficient than other competing methods whose computation complexities are as least $\mathrm{O}(NT^2)$ or $\mathrm{O}(NT \log T)$.

We should note that despite the simplicity and speed of SaRa, the selection of the bandwidth h is nontrivial: too small an h may reduce the statistical power, whereas too large an h may miss the short CNVs. A similar problem also haunts other single sample methods. Specifically, short CNV regions are hard to detect since the statistical evidence is relatively



(a) Regions with ≤ 5000 markers. (b) Regions with ≤ 100 markers.

FIG 6. *The histogram of the number of SNP markers between change-points identified in CHOP data. Regions with* ≤ 5000 *markers are shown on the left, and regions with* ≤ 100 *markers are shown on the right.*

weak. Thus, the false positive rate usually has to be sacrificed to detect these short regions. Some *ad hoc* methods have been proposed to solve this problem. For example, in Birdsuite, a program called Canary can detect common short CNVs by using prior knowledge. This solution is, however, platform-specific and cannot work when the prior knowledge is lacking. This problem is greatly alleviated in multiple-sample SaRa. Because we have shown in theory that the statistical power of multiple-sample SaRa converges to 1 as the number of samples increases, a large h is no longer crucial to get decent statistical power given enough samples. In multiple sample SaRa, we recommend h be selected as large as possible provided that the biological interests are accommodated. For example, the median distance between adjacent markers is below 700 bases in Affymetrix Genome-Wide Human SNP Array 6.0. Using this platform, h should be set ≤ 15 to study CNVs longer than 10k bases.

Furthermore, we proposed a novel adaptive Fisher's method which combines p-values while adapting to the proportions of true signals. We have shown by simulation that this statistic is powerful regardless of the proportion of true signals among the combined p-values. Another advantage is that the sums of the transformed order p-values are standardized using their theoretical means and variances, which saves computation time by avoiding a double permutation procedure.

In conclusion, we proposed a new change-point calling method which utilizes information from multiple samples. The SaRa scan statistic is used to make this method computationally efficient and robust against long range trends in the data. The novel adaptive Fisher's statistic enables the method to accommodate both rare and common change-points. It should also be noted that this work is the first that compared the single sample methods and multiple sample methods theoretically and numerically.

APPENDIX A: PROOF OF THEOREM 1

Let $h = L/2$, then with arguments similar to Lemma 3 in Niu and Zhang (2012), for any combining statistic $W(t,h)$, it suffices to show that there exists λ such that as $N \to \infty$,

$$P\big(W(t,h) < \lambda\big) \to 1, \quad \text{for any } h\text{-flat point } t, \tag{A.1}$$

$$P\big(W(\tau_j,h) < \lambda\big) \to 0, \quad \text{for any } j \in \{1,\dots,J\}. \tag{A.2}$$

We first prove (A.1) and (A.2) for W^{Sum}, W^{WSum}, W^{Fisher}, and $W^{Stouffer}$ by proving them for a general class of combining statistics of the following form

$$W(t,h) = \frac{1}{N}\sum_{i=1}^{N} f\big[\tilde{D}_i(t,h)\big], \tag{A.3}$$

where $f(\cdot)$ is a function satisfying conditions (C1) and (C2) specified below.

Let $Z_0 \sim N(0,1)$ and Z_j follows the same distribution as $\tilde{D}_i(\tau_j,h)$ (for $j = 1,\dots,J$).

(C1) $E|f(Z_0)|^3$ and $E|f(Z_j)|^3$ $(j = 1,\dots,J)$ exist.
(C2) $\nu_0 \equiv Ef(Z_0) < \nu_j \equiv Ef(Z_j)$ for any j.

It can be shown that W^{Sum}, W^{WSum}, W^{Fisher}, and $W^{Stouffer}$ all follow (A.3) and satisfy (C1). They also satisfy (C2) when $\eta_j^2 > 1$ or when $\eta_j^2 = 1$ and $\Delta_j \neq 0$.

Next, we show that (A.1) and (A.2) hold for (A.3) with the two conditions specified above. We let $\zeta = \min_j(\nu_j - \nu_0)/2$, $\lambda = \nu_0 + \zeta$, and $\zeta_j = \zeta/\sqrt{\mathrm{Var}[f(Z_j)]}$ for $j = 0,\dots,J$. For (A.1), note that for an h-flat point t,

$$P\big(W(t,h) < \lambda\big) = P\left(\frac{W(t,h) - \nu_0}{\sqrt{\mathrm{Var}[f(Z_0)]/N}} < \sqrt{N}\zeta_0\right),$$

which can be approximated by the corresponding normal probability with the difference being controlled by the non-uniform Berry-Esseen bound (see

e.g. Theorem 14 on Page 125 of Petrov (1975)) as follows.

$$\begin{aligned} & P\big(W(t,h) < \lambda\big) \\ \geq\ & \Phi\big(\sqrt{N}\zeta_0\big) - \frac{C \cdot E|f(Z_0) - \nu_0|^3}{\sqrt{N}\big\{\mathrm{Var}[f(Z_0)]\big\}^{3/2}\big(1 + \sqrt{N}\zeta_0\big)^3} \\ \geq\ & 1 - \frac{1}{\sqrt{2N\pi}\zeta_0}\exp\Big(-\frac{1}{2}N\zeta_0^2\Big) - \frac{C \cdot E|f(Z_0) - \nu_0|^3}{N^2\zeta^3} \\ \to\ & 1, \text{ as } N \to \infty, \end{aligned}$$

where C is an absolute constant.

Similarly, since $\nu_j > \lambda$,

$$\begin{aligned} & P\big(W(\tau_j,h) < \lambda\big) \\ \leq\ & \Phi\big(-\sqrt{N}\zeta_j\big) + \frac{C \cdot E|f(Z_j) - \nu_j|^3}{\sqrt{N}\big\{\mathrm{Var}[f(Z_j)]\big\}^{3/2}\big(1 + \sqrt{N}\zeta_j\big)^3} \\ \leq\ & \frac{1}{\sqrt{2N\pi}\zeta_j}\exp\Big(-\frac{1}{2}N\zeta_j^2\Big) + \frac{C \cdot E|f(Z_j) - \nu_j|^3}{N^2\zeta^3} \\ \to\ & 0, \text{ as } N \to \infty. \end{aligned}$$

Therefore, we prove the sure coverage property for W^{Sum}, W^{WSum}, W^{Fisher}, and $W^{Stouffer}$.

For W^{HC}, the sure coverage property follows directly from Theorem 7 in Cai, Jeng and Jin (2011) which showed that, in our setting, letting $\lambda = \sqrt{2(1+\rho)\log\log N}$ for a positive number ρ guarantees (A.1) and (A.2).

For adaptive Fisher's method, we consider $f^{Fisher}(\cdot) = -\log\{2[1 - \Phi(|\cdot|)]\}$. We verified (C1) and (C2) for f^{Fisher} before and continue to use the notation ν_j $(j = 0. \ldots, J)$ and ζ. Recall that for a flat-point t, $W^{AF}(t,h) = \max_{1\leq i\leq N}|\tilde{V}_i(t,h)|$ where for each $1 \leq i \leq N$, $\tilde{V}_i(t,h)$ is a standardized sum of independent exponential random variables. In this case, let $\lambda = \sqrt{N}\zeta$ and we have

$$\begin{aligned} P\big(W^{AF}(t,h) < \sqrt{N}\zeta\big) &= P\Big(\bigcap_{i=1}^{N}\big\{-\sqrt{N}\zeta < \tilde{V}_i(t,h) < \sqrt{N}\zeta\big\}\Big) \\ &\geq 1 - \sum_{i=1}^{N}\big[1 - P\big(-\sqrt{N}\zeta < \tilde{V}_i(t,h) < \sqrt{N}\zeta\big)\big], \end{aligned}$$

where according to Result 23 on Page 132 of Petrov (1975), for an absolute

constant C,

$$\begin{aligned} & P\big(-\sqrt{N}\zeta < \tilde{V}_i(t,h) < \sqrt{N}\zeta\big) \\ \geq\ & 1 - \sqrt{\frac{2}{N\pi\zeta^2}}\exp\Big(-\frac{1}{2}N\zeta^2\Big) - \frac{2C}{\big(1+\sqrt{N}\zeta\big)^3}\cdot\frac{\sum_{k=1}^N w^3(k,i)}{\big[\sum_{k=1}^N w^2(k,i)\big]^{3/2}} \\ \geq\ & 1 - \sqrt{\frac{2}{N\pi\zeta^2}}\exp\Big(-\frac{1}{2}N\zeta^2\Big) - \frac{2C}{\big(1+\sqrt{N}\zeta\big)^3}. \end{aligned}$$

Therefore,

$$\begin{aligned} & P\big(W^{AF}(t,h) < \sqrt{N}\zeta\big) \\ \geq\ & 1 - \sqrt{\frac{2N}{\pi\zeta^2}}\exp\Big(-\frac{1}{2}N\zeta^2\Big) - \frac{2NC}{\big(1+\sqrt{N}\zeta\big)^3} \\ \to\ & 1, \quad \text{as } N\to\infty. \end{aligned}$$

On the other hand, for each $j \in \{1,\ldots,J\}$,

$$\begin{aligned} & P\big(W^{AF}(\tau_j,h) > \sqrt{N}\zeta\big) \\ \geq\ & P\big(|\tilde{V}_N(\tau_j,h)| > \sqrt{N}\zeta\big) \\ =\ & 1 - P\big(-\sqrt{N}\zeta < \tilde{V}_N(\tau_j,h) < \sqrt{N}\zeta\big) \\ =\ & 1 - P\Big(-\zeta < \frac{1}{N}\sum_{i=1}^N f^{Fisher}\big[\tilde{D}_i(\tau_j,h)\big] - \nu_0 < \zeta\Big) \\ \to\ & 1, \quad \text{as } N\to\infty. \end{aligned}$$

This concludes our proof of the sure coverage property for the adaptive Fisher's method.

APPENDIX B: PROOF OF THEOREM 2

For Part (a), we can show that W^{Sum}, W^{WSum}, W^{Fisher}, and $W^{Stouffer}$ all follow (A.3) and satisfy Condition (C3) specified below.

Let $Z_0 \sim N(0,1)$ and Z_j follows the same distribution as $\tilde{D}_i(\tau_j,h)$ (for $j=1,\ldots,J$).

(C3) The moment generating function of the random variable $f(Z_j)$, namely $M_j(\beta) \equiv E\{\exp[\beta f(Z_j)]\}$ exists on a neighborhood B_0 of 0 for $j = 0,1,\ldots,J$.

Note that (C1) and (C2) still holds for these combining methods. We can define λ respectively for each combining method as in Theorem 1 that

satisfies (A.1). This threshold is more conservative than what is required by the current theorem, so it suffices to show that with this threshold, (A.2) holds with a rate of convergence exponential in N.

Under Condition (C3), Cramer-Chernoff Theorem gives

$$\lim_{N\to\infty} \frac{1}{N} \log P\big(W(\tau_j, h) < \lambda\big) = -I_j(\lambda), \tag{B.1}$$

where $I_j(x) = \sup_{\beta\in B_0}[\beta x - \log M_j(\beta)]$. Since $\lambda \neq \nu_j$, $I_j(\lambda) > 0$. Therefore, the claim is true for W^{Sum}, W^{WSum}, W^{Fisher}, and $W^{Stouffer}$.

To prove the claim for the higher criticism method, note that for each j, we can find $0 < x_j < 1$ such that $P(p_i(\tau_j, h) < x_j) \neq x_j$. With arguments similar to those in Cai, Jeng and Jin (2011),

$$\begin{aligned} & P\big(W^{HC}(\tau_j, h) \le \sqrt{2(1+\rho)\log\log N}\big) \\ = \; & P\bigg(\sup_{0<x<1} \sqrt{N}\Big|\frac{\frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\{p_i(\tau_j,h)<x\}} - x}{\sqrt{x(1-x)}}\Big| \le \sqrt{2(1+\rho)\log\log N}\bigg) \\ \le \; & P\bigg(\Big|\frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\{p_i(\tau_j,h)<x_j\}} - x_j\Big| \le \sqrt{\frac{2x_j(1-x_j)(1+\rho)\log\log N}{N}}\bigg). \end{aligned}$$

Since $E(\mathbf{1}_{\{p_i(\tau_j,h)<x_j\}}) = P(p_i(\tau_j, h) < x_j) \neq x_j$, we can apply the Cramer-Chernoff Theorem again and obtain an upper bound converging to zero exponentially fast as in (B.1).

For the adaptive Fisher's method, with the same notation as Theorem 1,

$$P\big(W^{AF}(\tau_j, h) < \sqrt{N}\zeta\big) \le P\Big(\nu_0 - \zeta < \frac{1}{N}\sum_{i=1}^N f^{Fisher}\big[\tilde{D}_i(\tau_j, h)\big] < \nu_0 + \zeta\Big).$$

Then since $E\{f^{Fisher}[\tilde{D}_i(\tau_j, h)]\} > \nu_0 + \zeta$, we can again show it converges to zero at a rate at least exponential in N.

For Part (b), we use notation $A_N \sim B_N$ to denote that $\lim_{N\to\infty} A_N/B_N = 1$ for two sequences $\{A_N\}$ and $\{B_N\}$. We first find the order of λ^*. For an h-flat point t, we have

$$P\big(\max_i |\tilde{D}_i(t, h)| > \lambda^*\big) = 1 - \{1 - 2[1 - \Phi(\lambda^*)]\}^N = \alpha.$$

This suggests that $\lim_{N\to\infty} \lambda^* = \infty$ and thus by Mill's ratio $1 - \Phi(\lambda^*) \sim \frac{1}{\lambda^*}\phi(\lambda^*)$. With simple calculation, the formula above leads to $\frac{1}{\lambda^*}\phi(\lambda^*) \sim -\log(1-\alpha)/2N$. Solve for λ^*, we can show that $\lambda^* \sim \sqrt{2\log N}$, or more precisely,

$$\lim_{N\to\infty} (\lambda^*)^2 - 2\log N + \log\log N + \log\big[\pi \log^2(1-\alpha)\big] = 0.$$

Second, we find the asymptotic power of detecting change-point τ_j, which is

$$\begin{aligned} & P\big(\max_i |\tilde{D}_i(\tau_j, h)| > \lambda^*\big) \\ &= 1 - \Big\{1 - 2(1-p_j)\big[1-\Phi(\lambda^*)\big] - p_j\Big[1 - \Phi\Big(\frac{\lambda^* - \psi_j}{\eta_j}\Big) - \Phi\Big(\frac{-\lambda^* - \psi_j}{\eta_j}\Big)\Big]\Big\}^N \\ &\equiv 1 - (1 - P_1 - P_2)^N, \end{aligned}$$

where $\psi_j = -\Delta_j\sqrt{h/2}$. Note that

$$P_2 \sim \frac{C}{\lambda^*}\exp\Big[\frac{-(\lambda^*)^2 + 2|\psi_j|\lambda^*}{2\eta_j^2}\Big],$$

where C denotes a generic constant. Since $\eta_j^2 \geq 1$, $P_2/P_1 \to \infty$, as $N \to \infty$,

$$\begin{aligned} & \log\big[1 - P\big(\max_i |\tilde{D}_i(\tau_j, h)| > \lambda^*\big)\big] \\ \sim\ & -N\frac{C}{\lambda^*}\exp\Big[\frac{-(\lambda^*)^2 + 2|\psi_j|\lambda^*}{2\eta_j^2}\Big] \\ \sim\ & -C\cdot N^{1-\frac{1}{\eta_j^2}}(\log N)^{-\frac{1}{2}+\frac{1}{2\eta_j^2}}\exp\Big(\frac{|\psi_j|}{\eta_j^2}\sqrt{2\log N}\Big), \end{aligned}$$

which tend to $-\infty$ as $N \to \infty$ but slower than $-N$. Thus, the theorem is proved.

ACKNOWLEDGEMENTS

This work is supported by the grant ??. The CHOP Control Copy Number Variation dataset was supplied by the Center for Applied Genomics at the Children's Hospital of Philadelphia through dbGaP accession number phs000199.v1.p1.

REFERENCES

Cai, T. T., Jeng, X. J. and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B* **73** 629–662.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32** 962–994.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J. et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307** 1434–1440.

Huang, T., Wu, B., Lizardi, P. and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics* **21** 3811–3817.

Jeng, X. J., Cai, T. T. and Li, H. (2013). Simultaneous discovery of rare and common segment variants. *Biometrika* **100** 157–172.

Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K. et al. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40** 1253–1260.

Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* **5** 994–1019.

Littell, R. C. and Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association* 802–806.

Littell, R. C. and Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests II. *Journal of the American Statistical Association* 193–194.

McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P., Zody, M. C., Hall, J. L., Brant, S. R., Cho, J. H. et al. (2008). Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature genetics* **40** 1107–1112.

Niu, Y. S. and Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *The Annals of Applied Statistics* **6** 1306–1326.

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

Petrov, V. V. (1975). *Sums of independent random variables. Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

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

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. et al. (2007). Strong association of de novo copy number mutations with autism. *Science* **316** 445–449.

Shaikh, T. H., Gai, X., Perin, J. C., Glessner, J. T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L. K., D'arcy, M. et al. (2009). High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome research* **19** 1682–1690.

Siegmund, D., Yakir, B. and Zhang, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. *The Annals of Applied Statistics* **5** 645–668.

Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. and Williams Jr, R. M. (1949). *The American soldier: adjustment during army life.* Princeton Univ. Press.

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

Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23** 657–663.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H. and Bucan, M. (2007). PennCNV: An integrated hidden Markov model de-

signed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17** 1665–1674.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters* **6** 181–189.

Yao, Y.-C. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A* **51** 370–381.

Zhang, N. R., Siegmund, D. O., Ji, H. and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97** 631–645.

E-mail: heping.zhang@yale.edu