HHS Public Access

Author manuscript

Ann Appl Stat. Author manuscript; available in PMC 2017 December 01.

Published in final edited form as:

Ann Appl Stat. 2016 December ; 10(4): 2102–2129. doi:10.1214/16-AOAS966.

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

Chi Song[*],
Ohio State University

Xiaoyi Min[*], and
Georgia State University

Heping Zhang
Yale University

Abstract

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 mean for sequences of array-based intensity measurements. Although multiple samples are of interest, the majority of the available CNV calling methods are single sample based. Only a few multiple sample methods have been proposed using scan statistics that are computationally intensive and 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 is able to detect both common and rare change-points. We prove that asymptotically this method can find the true change-points with almost certainty and show in theory that multiple sample methods are superior to single sample methods when shared change-points are of interest. Additionally, we report extensive simulation studies to examine the performance of our proposed method. Finally, using our proposed method as well as two competing approaches, we attempt to detect CNVs in the data from the Primary Open-Angle Glaucoma Genes and Environment study, and conclude that our method is faster and requires less information while our ability to detect the CNVs is comparable or better.

Keywords and phrases

change-point detection; multi-sample inference; adaptive Fisher's method

1. Introduction

The chromosome copy number refers to the number of copies of a genomic deoxyribonucleic acid (DNA) region in a DNA mixture, relative to a control sample or a population control. In a human genome, except for the sex chromosomes, the DNA copy

[*]These authors contributed equally to this work.

numbers are normally two, with one copy from mother and the other copy from 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* mutations or inherited from 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 (Lengauer, Kinzler and Vogelstein, 1998), and can be very heterogeneous across different patients (Mermel et al., 2011). Inherited CNVs, on the contrary, are generally short in length, shared by many people, and aligned well across samples (Zhang et al., 2010). Recent studies have shown that 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 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. Over the last decade, high-throughput technologies such as array-comparative genomic hybridization (aCGH), single-nucleotide polymorphism (SNP) array, and next-generation sequencing (NGS) have been used to detect CNVs (Carter, 2007; Alkan, Coe and Eichler, 2011). Because the data produced by these technologies inevitably contain noise, various statistical methods have been proposed and applied to call CNV regions from noisy data. We mainly focus on detecting CNV from array-based data in this paper and briefly discuss the extension to NGS data in the Discussion section.

1.1. Statistical model

Regardless of the technology or platform, CNV detection can be formulated in the following way. Given N samples and T markers, raw copy number 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 $1 \le i \le N$. We assume

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i, \quad 1 \le i \le N, \quad (1.1)$$

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 $\varepsilon_i \sim \mathrm{MVN}(\mathbf{0}, \sigma_i^2 \mathbf{I})$. We call τ a change-point for sample i if $\mu_{i,\tau} \neq \mu_{i,\tau+1}$. For sample i, we denote its J_i change-points by $0 < \tau_{i,1} < \tau_{i,2} < \cdots < \tau_{i,J_i} < T$. By estimating all of the change-points $\boldsymbol{\theta}_i = \{\tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,J_i}\}$ for each sample i, CNV regions can be called between these change-points.

We denote the collection of change-points in all 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. Our proposed method 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 single sample methods and multiple sample methods according to their strategies. Single sample methods, on the one hand, simply apply a CNV calling algorithm to each individual sample repeatedly. Multiple sample methods, on the other hand, assume that certain change-points may be shared by a proportion of the samples, and call these shared change-points using information from multiple samples.

Because of the complexity of analyzing multiple samples together, most current methods focus on a single sample. Yao (1988) and Yao and Au (1989) proposed to search for the combination of change-points that minimizes a BIC score, and they showed the consistency of their estimates. Another approach uses 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). Circular binary segmentation (CBS) algorithm (Olshen et al., 2004; Venkatraman and Olshen, 2007) uses a strategy of recursively finding segments with changed means in a sequence. It is based on the following scan statistic: for a region (s, t),

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

where $S_{i,t}$ is the partial sum of sequence $\mathbf{Y}_i$ (i.e. $S_{i,t}=\sum_{j=1}^{t}Y_{i,j}$), $\overline{Y}_i = S_{i,T}/T$, and $\hat{\sigma}_i^2=\sum(Y_{i,j}-\overline{Y}_i)^2/T$. The region with the highest $U_i(s, t)$ is further scrutinized. Note that CBS uses global information to detect change-points. Niu and Zhang (2012) demonstrated that local information is more efficient than global information for high-throughput data for change-points detection. They proposed a screening and ranking algorithm (SaRa) using the following scan statistic

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

for $1 \le t \le T$, where h is a bandwidth parameter. Because $D_i(t, h)$ is calculated from local information within a $2h$ window, the complexity of this algorithm is linear in T. This algorithm was refined by Xiao, Min and Zhang (2015) and further studied theoretically by Hao, Niu and Zhang (2013). In addition to change-point models, other models such as Hidden Markov Model (HMM) are also applied to CNV detection. For example, PennCNV (Wang et al., 2007) and Birdsuite (Korn et al., 2008) are the two most popular HMM methods. Due to space limitation, we do not discuss these models in detail. In Section 5, we examine the performance of PennCNV in a real data analysis.

Zhang et al. (2010) noted that different people can share CNV regions. In the framework of change-point model, this means some of the change-points are shared by multiple samples.

Based on this idea, several multiple sample methods have been developed to find shared change-points. Zhang et al. (2010) proposed taking the sum of squared scan statistics from individual samples to find common change-points. Siegmund, Yakir and Zhang (2011) further extended this method by using a weighted sum of squares statistic, which increases the power for rare change-point detection when prior information on carrier proportions is available. Instead of using these sum-based statistics, Jeng, Cai and Li (2013) summarized the scan statistics based on 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 these methods for individual samples, however, are virtually the same as the CBS scan statistic. Alternatively, Vert and Bleakley (2010) considered a group LASSO approach for detecting shared change-points in multiple samples. Fan et al. (2015) also used a penalized likelihood approach but assumed Laplace distribution for the observed sequences to detect change-points in either mean or variance.

1.3. Motivations

Despite the success of the aforementioned methods, several aspects of them need to be addressed or could be improved. First, the multiple sample methods that we reviewed all use the CBS scan statistic which is based on global information. In real data, it is most likely that there exists more than one region of change, and a global statistic may contain data points that are irrelevant and increase heterogeneity, and hence lose power. In addition, these methods tend to suffer from higher computational complexities, especially when applied to high-throughput genomic data. To overcome this computational burden and potentially enhance the power, 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.

Second, we note that most 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 carrier proportion and does not require any prior knowledge or assumption. To this end, Jeng, Cai and Li (2013) proposed to use higher criticism method, which enjoys good theoretical properties and could detect any “detectable” shared variants with any carrier proportion. However, we found that the power of this method in CNV detection is low. 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 report that, regardless of the carrier proportion, this method has a good power of finding change-points.

Finally, an important issue with regard to multiple sample methods is whether they provide any improvement over single sample methods in detecting shared CNVs. To address this question, Zhang et al. (2010) and Siegmund, Yakir and Zhang (2011) concluded through simulations and real data analyses that cross-sample scans perform better than single sample scans. In this paper, we provide both theoretical and numerical comparisons between our

proposed method and single sample methods, which further confirm that the power of multiple sample methods is higher than that of a single sample method.

Section 2 presents our method in detail, and Section 3 provides its theoretical properties. We demonstrate its performance via simulation in Section 4, and we analyze a real dataset in Section 5.

2. Method

2.1. SaRa for a single sample

First, we review the SaRa method proposed by Niu and Zhang (2012). For a single sample i, given a band-width 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 collection 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 no change-point exists in window $(t - h + 1, t + h)$, $D_i(t, h) \sim N(0, \frac{2}{h}\sigma_i^2)$. Therefore, we can define a standardized scan statistic as

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

where $\hat{\sigma}_i$ is an estimate of σ_i. By assuming that the number of change-points in sample i, $J_i \ll T$, 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 across samples to identify shared change-points, we need to combine single sample statistics for all samples. A natural choice is to take the sum of squares of $\tilde{D}_i(t, h)$ across samples as in Zhang et al. (2010) and define a multiple sample scan statistic

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

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

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

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

The two methods above combine the scan statistics $\tilde{D}_i(t, h)$ directly. We can also combine 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

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

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

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

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

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

where

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

Because both common and rare CNVs have been found to be associated with many human diseases (McCarroll and Altshuler, 2007), a desirable CNV detection method should be powerful for both types of CNVs. Therefore, we need a combining method which is sensitive to change-points with different carrier proportions. While the sum of squares statistic is easy to implement, it is good in capturing only change-points that are shared by many samples. Conversely, the higher criticism statistic can 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 low in practical applications. Although the weighted sum of squares statistic can detect both common and rare change-

points, it depends on a tuning parameter π_0 whose choice relies on prior assumptions of the change-points. 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 ordered *p*-values so that only those 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-weigh the non-carriers, but it is time consuming and involves exhaustive search for the weights. Yu et al. (2009) and Zhang, Chen and Pfeiffer (2013) considered a similar adaptive rank truncated product statistic of the *p*-values, but they rely on either permutations or numerical integration to decide the significance level. 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, $X_i(t,h) \overset{iid}{\sim} \mathrm{EXP}(1)$, 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 $\xi_i(t,h) \overset{iid}{\sim} \mathrm{EXP}(1)$ 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 $w(k,i) = \min(1, i/k)$. We standardize $V_i(t,h)$ 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

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

In CNV detection, we are mainly interested in detecting signals arising from shifted means. Therefore, we can consider only the smaller p-values and the one-sided tests that the p-values are less than their expected values. Moreover, genetic data for CNV detection, especially SNP array data, are prone to artifacts including guanine-cytosine (GC) content, batch effects, and bad probes on the chips. Outliers caused by these artifacts could lead to false discoveries. Considering these issues, the adaptive Fisher's statistic can also be defined as

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

where a tuning parameter n_0 specifies that at least n_0 observations are combined so that the statistic is more robust to outliers. Similarly, we could modify (2.6) into

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

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 $W^{AF}(t,h)$. We now extend the SaRa method for multiple samples using these methods. Let $\{W(t,h): t=1,\ldots,T\}$ be the sequence of combined statistics using any of the six combining methods with a bandwidth h. Then we can find the local maximizers of this sequence, and select a subset of the local maximizers by thresholding, as done in 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 collection of local maximizers* $\mathcal{M} = \{t : W(t, h) > W(t', h), \forall t' \in (t - h, t + h)\}$.

4. *Given a threshold* λ, *estimate the shared change-points as a subset of* $\mathcal{M}$, $\hat{\boldsymbol{\theta}} = \{\hat{\tau}_1 < \hat{\tau}_2 < \cdots < \hat{\tau}_{\hat{J}}\} \subseteq \mathcal{M}$, *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 1: In the calculation of $\tilde{D}_i(t, h)$, for $Y_{i,k}$ with $k < 1$ or $k > T$, use $\bar{Y}_i$ instead. This happens only when t is near either end of a sequence.

Remark 2: 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 $T \gg h$, we can simulate the null distribution of $W(t, h)$ using any length T' that satisfies $T' \gg h$. Let $\hat{F}(\cdot)$ be the simulated empirical distribution function of $W(t, h)$, where t is a local maximizer. Given a significance level α, the threshold can be calculated as $\lambda = \hat{F}^{-1}(1 - \alpha)$. Alternatively, we can also find λ as the $(1 - \alpha')$ quantile of the observed $W(t, h)$'s on the local maximizers for different values of α'.

2.3.1. Multiple-bandwidth SaRa—Genomic CNVs are different in size, ranging from one SNP site to the entire chromosome. Because we do not know the sizes of the CNVs to detect, there is no one bandwidth that fits all CNVs. The selection of bandwidth h may affect the result depending on the distance between adjacent change-points. As described by Niu and Zhang (2012), a large 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. Consider a set of B bandwidths $\mathbf{h} = \{h_1 < h_2 < \cdots < h_B\}$. With bandwidth h_b, we can estimate a set of change-points $\hat{\boldsymbol{\theta}}^{(b)}$. Then the candidates for shared change-points are estimated by $\hat{\boldsymbol{\theta}} = \cup_{b=1}^{B} \hat{\boldsymbol{\theta}}^{(b)}$. Because different bandwidths may yield change-points with slightly different positions, some change-points in $\hat{\boldsymbol{\theta}}$ may be redundant. To resolve this issue, we keep the corresponding change-point and drop the other change-point when two change-points detected by two different bandwidths are close to each other (e.g., the distance between them is less than the shorter bandwidth), as we "trust" the longer bandwidth. Moreover, some change-points with small mean shifts may not be reliable. 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 summary scan statistics. Consequently, we do not know which individuals carry a particular change. Hence, it is necessary and useful to identify the 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 local means along the chromosome, making it difficult to differentiate a real change-point from a shift caused by trends. This can be

resolved by thresholding as follows for a given sample i and candidate change-points $\hat{\boldsymbol{\theta}} = \{\hat{\tau}_1 < \hat{\tau}_2 < \cdots < \hat{\tau}_{\hat{\mathcal{J}}}\}$.

Algorithm: Carrier identification:

1. *Set* $\hat{\mathcal{J}}_i = \hat{\mathcal{J}}$ *and* $\hat{\tau}_{i,j} = \hat{\tau}_j$ *for* $j = 1, \ldots, \hat{\mathcal{J}}_i$. *Denote* $\hat{\boldsymbol{\theta}}_i = \{\hat{\tau}_{i,j}, j = 1, \ldots, \hat{\mathcal{J}}_i\}$.
2. *Let* $\hat{\tau}_{i,0} = 0$ *and* $\hat{\tau}_{i,\hat{\mathcal{J}}_i+1} = T$. *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* $0 \le j \le \hat{\mathcal{J}}_i$.
3. *Calculate the estimated jump size at each change-point* $d_{i,j} = m_{i,j} - m_{i,(j-1)}$ *for* $1 \le j \le \hat{\mathcal{J}}_i$.
4. *Find the change-point with the smallest absolute jump size*

$$j^* = \arg\min_{1 \le j \le \hat{\mathcal{J}}_i} |d_{i,j}|.$$

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

Remark 3: The choice of γ_i should be based on the particular dataset and scientific application. When technical replicates are available, γ_i can be determined based on the proportion of detections that can be verified. We illustrate this approach in Section 5.

Remark 4: 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 consider two questions regarding the proposed method: 1. Can we rely on this method to detect shared CNVs in multiple samples? 2. Since most current methods are single-sample-based, are multiple sample methods really advantageous to single sample methods to justify their use? To address these questions, we discuss two theoretical properties of multi-sample SaRa. First, we prove a sure coverage property as sample size N increases. This property guarantees to the users that when sample size is large, our method can detect shared CNVs in multiple samples with a high probability. Second, in the framework of our method, we compare the use of multiple samples versus one sample at a time for CNV detection. We show that multiple sample methods have higher asymptotic power in detecting shared change-points or CNVs and should be used instead of single sample methods. Admittedly, these asymptotic analyses may not be applicable to quantify

precisely the actual gain in power. Therefore, we rely on simulation studies to compare different methods in Section 4.

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 $\delta_{i,j}$ is 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 setting corresponds to a practical scenario that the platform for genotyping is fixed, and the locations of the underlying CNVs are also fixed. For each shared CNV, its carriers 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, so without loss of generality, we set them all equal to 1. Moreover, following Niu and Zhang (2012), we call a point t h-flat if the interval $(t - h, t + h)$ contains no change-point. 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 $\hat{\boldsymbol{\theta}} \pm h \equiv \cup_{j=1}^{\hat{J}} (\hat{\tau}_j - h, \hat{\tau}_j + h)$.

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 maximizer, 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 shared change-point. This is equivalent to using the maximum statistic of $\{\tilde{D}_i(t,h)\}_{i=1}^N$ and calling t a change-point 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 this naïve single sample method detecting a true change-point tends to 1 at a rate slower than 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 calls 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 5: We note 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.

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 certain proportion of 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* $1 \le i \le \lceil N\pi^* \rceil$, *sample* $Y_{i,j} \overset{iid}{\sim} N(0,1)$ *if* $1 \le j \le T/2$, *and sample* $Y_{i,j} \overset{iid}{\sim} N(\delta,1)$ *if* $T/2 < j \le T$. *Here,* $\lceil \cdot \rceil$ *is the ceiling function.*

3. *For* $\lceil N\pi^* \rceil < j \le N$, *sample* $Y_{i,j} \overset{iid}{\sim} N(0,1)$ *for* $1 \le j \le T$.

Different combining statistics were considered (for W^{WSum}, we set $\pi_0 = 0.01$ and $\pi_0 = 0.1$ because in real applications, the carrier proportion cover a large range, and we do not know what π_0 value is the best; for W^{HC} and W^{AF}, $n_0 = 4$ was used). We correctly detect the change-point 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. We also counted the number of local maximizers that fall out of $50 - h$ and $50 + h$ and exceeeds the threshold as the number of false discoveries. We checked the number of false discoveries of different methods to ensure that they are at the same level so that our comparison of power is fair (see details in Supplement Figure 2). 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 among 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 where a majority (99%) come from non-carriers. On the contrary, using the maximum test statistic as an example of 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. For weighted sum of squares statistic with $\pi_0 = 0.01$, even though the prior information is correctly specified, its power is slightly lower than that of the adaptive Fisher's method. As expected, the performance of weighted sum statistic with $\pi_0 = 0.1$ lies between the sum of squares and weighted sum with $\pi_0 = 0.01$.

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, Fisher's statistic, and weighted sum with $\pi_0 = 0.1$ have the best statistical power. Adaptive Fisher's statistic and Stouffer's statistic perform similarly with slightly lower power. Weighted sum statistic with $\pi_0 = 0.01$, 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 $h = 10$. To see the comparison more clearly, we calculate the relative power of different methods as their original power divided by the power of adaptive Fisher's method, such that the relative power of adaptive Fisher is always 1. A relative power greater than 1 means the method is more powerful than adaptive Fisher, and vise versa. The relative power is shown in Figure 2(b). Similar to our previous observation, maximum statistic, higher criticism, and weighted sum of squares statistic with $\pi_0 = 0.01$ 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. Adaptive Fisher's statistic enjoys competitive statistical power no matter π^* is small or large. Weighted sum of squares with $\pi_0 = 0.1$ is the closest competitor. It has the highest statistical power when the real carrier proportion is between 0.07 and 0.18, but is suboptimal in detecting rare change-points. In real applications, because rare change-points are more difficult to detect and we do not know the true carrier proportion for each change-point, we decided to use adaptive Fisher's statistic rather than the weighted sum of squares for our algorithm. To illustrate how the adaptive Fisher's statistic works, we show in Figure 3 its average peak positions and those of 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 $h = 20$ 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 does not perform well when π^* gets large.

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* $1 \le i \le 1000$ *and* $1 \le t \le 500$, *and set* $\mu_{i,t}$ *to* 0 *except for the following change-regions in their carriers.*
   - **a.** *Region 1:* $28 \le t \le 54$ *(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:* $116 \le t \le 130$ *(length is* 15*), set* $\mu_{i,t} = \delta_2 = -1.92$ *if sample i is a carrier,*

the carrier proportion π_2 = 0.05.

c. *Region 3:* $222 \le t \le 306$ *(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* $\mathbf{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$ *for* $1 \le i \le 1000$*, where* $\boldsymbol{\varepsilon}_i \sim MVN(\mathbf{0}, \mathbf{I})$.

Figure 4 displays five representative examples of individual sequences, with a total of 3 pairs of change-points: one shared by sequences 2 and 4 (positions 27 and 54), one shared by sequences 3, 4, and 5 (positions 115 and 130), and one unique to sequence 5 (positions 221 and 306). 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, α = 0.001 was used when determining λ, and $\gamma_i = 2\hat{\sigma}_i \sqrt{2/h}$ was used for each bandwidth *h*).

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. Because single sample methods may not detect the same change-point at the same location for different samples, we grouped close-by change-points if they are within 3 markers. 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 a change-point as a true positive provided that it falls within 5 markers of the true position. From these tables, we can see that our proposed method performed slightly better than multiple-sample CBS in terms of sensitivity and specificity among the five methods.

4.2.2. Data with trend related to GC content—Signal intensities measured by SNP arrays are often prone to genomic waves. Diskin et al. (2008) found that these waves are related to genomic GC content. The correlation between the intensities and local genomic GC content can be either positive or negative, and the magnitude of the genomic wave is related to the DNA quantity loaded in the SNP array experiment. In other words, different samples may share the same wavy pattern, but the magnitude of these waves are different and often related to the batch of the experiments. Although these waves can be partially adjusted by regressing the observed intensities on the local GC content, it is not guaranteed that they can be completely removed. For example, selecting the bandwidth to calculate the local GC content is not trivial - a large bandwidth may result in insufficient adjustment, whereas a small bandwidth may fail to capture the local GC content and overfit the

adjustment model. Therefore, we believe that it is still beneficial for the CNV calling algorithm to be robust to local trend related to GC contend and batch effect.

To examine our method, we simulated data with a same wavy pattern in all the samples, but the magnitudes of the waves are different and randomly given. The rest of the simulation was the same as in Section 4.2.1. Specifically, we simulated $\mathbf{Y}_i$ by adding local trend and random noise as follows:

$$Y_{i,t}=\mu_{i,t}+a_i[\sin(2\pi t/96+\psi)+2\sin(2\pi t/240+\phi)]+\varepsilon_{i,t},$$

where $\psi \sim U(0, 2\pi)$, $\phi \sim U(0, 2\pi)$, $a_i \sim U(-0.15, 0.15)$, and $\varepsilon_{i,t} \sim N(0, 1)$. In this model, the wavy pattern was composed of two sine signals with different periods and amplitudes, which mimics the GC content. The overall magnitude of the wave a_i is uniformly distributed between −0.15 and 0.15 for each sample. Table 3 shows the number of shared change-points detected. We can see that the single sample methods were all greatly impacted by the trend introduced. Multiple-sample CBS was less affected, but still yielded a fair amount of false change-points. Our multiple-sample m-SaRa method, however, still performed robustly and detected all 6 true change-points in 998 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 the SaRa scan statistic look for sharp mean change using local information, which makes it immune to the influence of a large scale trend. In addition to this setting, we simulated two more scenarios to test our robustness towards trends with different patterns among individuals as well as dependent measurement error (see Supplement Sections 1 and 2 for details).

4.2.3. Data with imperfectly aligned change-points—In real data, even though the CNV can be shared across a proportion of samples, the change-points can be slightly different in each carrier. To evaluate our method under this situation, we simulated data with imperfectly aligned change-points. The simulation procedure we adopted is the same as in Section 4.2.1, except that we added a random shift up to 3 SNPs to each change-point in each sample. The probabilities that a change-point shift by 1, 2, and 3 SNPs from the original location is 30%, 20% and 10% respectively, which leaves the probability of having no shift in the change-point at 40%. We set the maximal shift to be 3 because if the change-points differ by more than 3 SNPs, it is more appropriate to consider them as different change-points. Table 4 presents the number of shared change-points detected. We can see that the single sample methods often detected more than 6 change-points in the 1000 replications; whereas multiple sample methods including multiple-sample CBS and multiple-sample m-SaRa always detected 6 change-points.

5. Real data analysis

We examined the performance of our method by applying it to the genetic data from the Primary Open-Angle Glaucoma Genes and Environment (GLAUGEN) study (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000308.v1.p1). This dataset contains 1363 samples from 1343 individuals, including 20 pairs of technical

replicates. A total of 657, 366 markers were genotyped using Illumina Human660W-Quad v1 A chips. The detail steps on data pre-processing and implementation of our method are as follows:

1. *The observed Log R Ratio sequences were adjusted by the local GC content as suggested by* Diskin et al. (2008) *using the program "genomic wave.pl" in the PennCNV package.*
2. *The adjusted sequences were each centered at* 0.
3. *Multiple-sample m-SaRa were run using bandwidths 5, 10, and 15. The cutoff* λ *was determined following the approaches in Remark 2, first as the* 99.99*-th percentile of the simulated null distribution of* $W(t, h)$ *at local maximizers (*$\alpha = 0.0001$*), and then as the* 50*-th percentile from the observed values of* $W(t, h)$ *at the local maximizers (*$\alpha' = 0.5$*).*
4. *Different values for the cutoff on the mean differences were tested:* $\gamma_i = k\hat{\sigma}_i$ *where* $k = 0.1, 0.2, \ldots, 3$.

Note that even after adjusting for the GC content, microarray data could still be affected by artifacts. For example, outliers may arise as a result of bad probes. Furthermore, the normality assumption on the errors are generally violated. For these issues, one may consider applying further adjustments on the data, and the approaches include median polishing and quantile normalization (Xiao, Min and Zhang, 2015).

We used the same approach as Zhang et al. (2010) and Siegmund, Yakir and Zhang (2011) to assess detection accuracy. Specifically, we first applied the proposed method to all 1363 samples without the information on replications. Then we compared the detected copy number variants, or more precisely the detected change-points, for each pair of technical replicates. We defined inconsistent detections in a pair of samples as the variants detected in one but not the other sample of this pair. The remaining detected variants were called consistent detections. The proportion of consistent detections was calculated for each pair of technical replicates, and these proportions were then used to measure the performance of our method and other CNV detection methods. Note that high values in these proportions do not necessarily imply high accuracy. For example, they all equal 1 if all the samples are identified as carriers of every change-point by letting γ_i be 0 in our method. Therefore, we contrasted them with baseline proportions of consistent detections in 1000 randomly selected pairs of samples. Although it is not clear what the true baseline proportion should be for a random pair of samples, we expect that a good detection method should give higher proportions of consistent detections in replicate pairs than in random pairs.

We display the results for the 11, 244 markers on chromosome 22 in Figure 5 and Table 5. In Figure 5(a), we plot the total number of detections out of the 20 replicate pairs under different values of γ_i, and we also plot the median proportion of consistent detections in Figure 5(b). When γ_i's were low, the total number of detections was very high. Meanwhile, the proportions of consistent detections were high both for replicate pairs and for random

pairs, most likely because most samples were kept as carriers for each change-point due to low γ_i's, which rendered the detections unreliable. As γ_i's increased, more detections were filtered out, and the proportions of consistent detections dropped. The median proportion of consistent detections in random pairs became quite stable after $\gamma_i/\hat{\sigma}_i$ got greater than 1, which is a suitable range to choose γ_i's. For replicate pairs, the median proportion began increasing when $\gamma_i/\hat{\sigma}_i$ got greater than 0.8, and eventually climbed to 1 for $\gamma_i/\hat{\sigma}_i$ greater than 2.

A subsequent question is: what values should we use for γ_i's? In general, this depends on how noisy the data are, the true mean shifts for different copy number changes, and the researcher's preference in the trade-off between true positive and false negative detections. For this dataset, $\gamma_i = 1.2\hat{\sigma}_i$ is a plausible threshold because the proportion of consistent detections was briefly stable as $\gamma_i/\hat{\sigma}_i$ is between 1.2 and 1.4, which suggested that the true positives were being removed along with the false negatives for increasing γ_i's in this region. The proportion of consistent detections in random pairs was very stable until $\gamma_i/\hat{\sigma}_i =$ 1.7, so this is another potential cutoff. One could also use $\gamma_i = 2\hat{\sigma}_i$ if only the most reliable detections are wanted, but this tends to retain only a small number of change-points with the largest mean shifts, which in practice are the changes of two or more copies.

We applied two competing methods: multiple-sample CBS and PennCNV to the same data after pre-processing and compare their detection accuracies with our method in Table 5. The median proportion of consistent detections in the 20 pairs of replicates and in 1000 randomly selected pairs are summarized. We present only the results for γ_i's that led to a similar number of total detections to the competing methods so the results are comparable. When $\gamma_i/\hat{\sigma}_i = 1.2$, our method detected a similar number of variants to PennCNV. The median proportion of consistent detections in replicate pairs by our method was higher than PennCNV, and the proportion in random pairs by our method was also slightly higher. Note that our method only used the LRR values, whereas PennCNV also used the B-allele frequencies, yet our method gave similar or slightly better performance comparing to PennCNV. When $k = 0.5$, our method had a similar number of detected variants to multiple-sample CBS, the proportion of consistent detections in replicate pairs was slightly higher than multiple-sample CBS, but the proportion in random pairs was also higher than multiple-sample CBS. These results, however, seem to be worse than the results from our method with $k = 1.2$ and PennCNV.

We also performed our method with an additional bandwidth $h = 2$ in light of the fact that many CNV regions might be short and cannot be captured when the smallest bandwidth is 5. As displayed in Table 5, the accuracy was improved when the total number of detection was smaller, but it was compromised when the total number of detection was larger. Nevertheless, we detected more shorter CNV regions as illustrated in Figure 6. Figures 6(a) and 6(b) present the distribution of the lengths of regions (in numbers of markers) between consecutive change-points detected by our method (with $k = 1.2$) using 3 bandwidths ($h = 5$, 10, 15) and 4 bandwidths ($h = 2$, 5, 10, 15), respectively. Since most of these regions are short, we only plot those with no more than 60 markers. Here, we considered chromosomes 1–22 in all 1363 samples. These two figures indicate that with the extra bandwidth 2, the

number of detected regions with more than 10 markers did not change much, but many more shorter regions were detected, especially those with 5 or fewer markers.

In Table 6, we compare the running time that each method used to identify CNVs/change-points from the 11, 244 markers on chromosome 22 for 1, 363 samples. The computation was performed on a Windows workstation equipped with 2 Inter(R) Xeon(R) E5645 processors (12 cores in total) and 48GB RAM though we did not apply parallelization for any method. Our proposed method was much faster than the competing methods, which confirmed its advantage of having lower computational complexity. In fact, our method took about 87 minutes to finish scanning the 640, 663 markers on chromosomes 1–22 for the 1, 363 samples, which is the size of a typical GWAS study. Note that for our method, obtaining the threshold λ from the simulated null distribution takes additional time and is computationally intensive. We could use the quantile in the observed values instead, which does not cost extra time. As can be seen in Figure 5 and Table 5, this threshold gave very similar results to the simulated threshold.

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 systematic evaluation of multiple sample methods and single sample methods. In this study, we have demonstrated 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 does not utilize information across samples, especially with the growth of studies with large sample sizes. Conversely, 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 is validated by our simulation.

Instead of using the CBS scan statistic, we employed the SaRa scan statistic in our method. The SaRa statistic utilizes 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 $O(TN \log N)$. In practice $T \gg N$, our method is much more computationally efficient than other competing methods whose computation complexities are at least $O(NT^2)$ or $O(NT \log T)$.

We should note that despite the simplicity and speed of SaRa, the selection of 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 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 multi-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 multi-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 demonstrated 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 the real data analysis, we noticed that the proposed statistic might be over-sensitive by picking too many candidates for common change-points when we used the threshold λ from the simulated null distribution. This might be due to the noise in the data as well as the departure from the normal assumption in our model. In this regard, we can decide the threshold empirically as suggested in the paper, or other distribution assumptions can be used for the LRR sequences. Alternatively, we can consider using mean, median, or other quantiles in (2.7) as one of the referees suggested. This would make W^{AF} more robust to outliers caused by artifacts in the data, but the sensitivity might be compromised. Further study is needed to address the pros and cons of these alternatives.

Recent developments in NGS methods have allowed us to analyze DNA sequences at a much finer level. CNVs can also be detected by scanning for change-points in sequences of read depths (Alkan, Coe and Eichler, 2011). For this type of data, our proposed method could be extended naturally and has the advantage of being computationally efficient, but it faces several challenges. First, read depth sequences are count data and correlated, so new distributional assumptions are needed, and the properties of the SaRa statistic need to be re-evaluated. Second, the p-values that we combine in the adaptive Fisher's method might also be discrete as the SaRa statistics are discrete, which makes evaluation of the significance level more difficult. These are important issues for future research.

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 has compared the single sample methods and multiple sample methods theoretically and numerically.

Acknowledgments

This work is supported by grant R01DA016750 from the National Institute on Drug Abuse and initially submitted while the first two authors were postdoctoral associates at Yale University. We thank the editor, an Associate Editor, and two anonymous referees for their helpful comments and suggestions. Funding support for the GLAUGEN study was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004728). The GLAUGEN study is one of the genome-wide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH GEI (U01 HG04424). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000308.v1.p1. R package for multiple-sample CBS was downloaded from http://penncnv.openbioinformatics.org/en/latest/.

References

Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nature Reviews Genetics. 2011; 12:363–376.

Cai TT, Jeng XJ, Jin J. Optimal detection of heterogeneous and heteroscedastic mixtures. Journal of the Royal Statistical Society: Series B. 2011; 73:629–662.

Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nature genetics. 2007; 39:S16–S21. [PubMed: 17597776]

Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Research. 2008; 36:e126. [PubMed: 18784189]

Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures. The Annals of Statistics. 2004; 32:962–994.

Fan Z, Dror RO, Mildorf TJ, Piana S, Shaw DE. Identifying localized changes in large systems: Change-point detection for biomolecular simulations. Proceedings of the National Academy of Sciences. 2015; 112:1–6.

Fisher, RA. Statistical methods for research workers. Edinburgh: 1925.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science. 2005; 307:1434–1440. [PubMed: 15637236]

Hao N, Niu YS, Zhang H. Multiple change-point detection via a screening and ranking algorithm. Statistica Sinica. 2013; 23:1553–1572. [PubMed: 24489450]

Huang T, Wu B, Lizardi P, Zhao H. Detection of DNA copy number alterations using penalized least squares regression. Bioinformatics. 2005; 21:3811–3817. [PubMed: 16131523]

Jeng XJ, Cai TT, Li H. Simultaneous discovery of rare and common segment variants. Biometrika. 2013; 100:157–172. [PubMed: 23825436]

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature Genetics. 2008; 40:1253–1260. [PubMed: 18776909]

Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature. 1998; 396:643–649. [PubMed: 9872311]

Li J, Tseng GC. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. The Annals of Applied Statistics. 2011; 5:994–1019.

Littell RC, Folks JL. Asymptotic optimality of Fisher's method of combining independent tests. Journal of the American Statistical Association. 1971:802–806.

Littell RC, Folks JL. Asymptotic optimality of Fisher's method of combining independent tests II. Journal of the American Statistical Association. 1973:193–194.

McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nature genetics. 2007; 39:S37–S42. [PubMed: 17597780]

McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nature genetics. 2008; 40:1107–1112. [PubMed: 19165925]

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G, et al. GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. 2011

Niu YS, Zhang H. The screening and ranking algorithm to detect DNA copy number variations. The Annals of Applied Statistics. 2012; 6:1306–1326. [PubMed: 24069112]

Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004; 5:557–572. [PubMed: 15475419]

Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Børresen-Dale AL, Brown PO. 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. 2002; 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. Strong association of de novo copy number mutations with autism. Science. 2007; 316:445–449. [PubMed: 17363630]

Siegmund D, Yakir B, Zhang NR. Detecting simultaneous variant intervals in aligned sequences. The Annals of Applied Statistics. 2011; 5:645–668.

Stouffer, SA.; Suchman, EA.; Devinney, LC.; Star, SA.; Williams, RM, Jr. The American soldier: adjustment during army life. Princeton Univ. Press; 1949.

Tibshirani R, Wang P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. Biostatistics. 2008; 9:18–29. [PubMed: 17513312]

Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007; 23:657–663. [PubMed: 17234643]

Vert, J.; Bleakley, K. Fast detection of multiple change-points shared by many signals using group LARS. In: Lafferty, JD.; Williams, CKI.; Shawe-Taylor, J.; Zemel, RS.; Culotta, A., editors. Advances in Neural Information Processing Systems. Vol. 23. Curran Associates, Inc; 2010. p. 2343-2351.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, Hakonarson H, Bucan M. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research. 2007; 17:1665–1674. [PubMed: 17921354]

Xiao F, Min X, Zhang H. Modified screening and ranking algorithm for copy number variation detection. Bioinformatics. 2015; 31:1341–1348. [PubMed: 25542927]

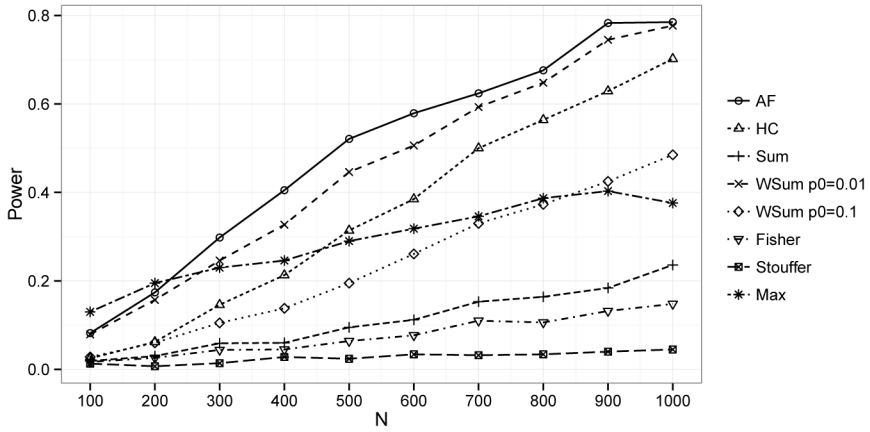
Yao YC. Estimating the number of change-points via Schwarz' criterion. Statistics & Probability Letters. 1988; 6:181–189.

Yao YC, Au ST. Least-squares estimation of a step function. Sankhyā: The Indian Journal of Statistics, Series A. 1989; 51:370–381.

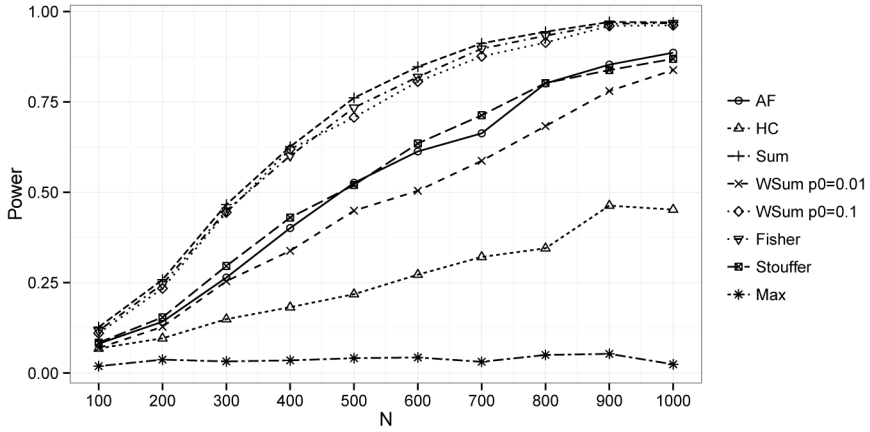
Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P, Chatterjee N. Pathway analysis by adaptive combination of P-values. Genetic epidemiology. 2009; 33:700–709. [PubMed: 19333968]

Zhang S, Chen HS, Pfeiffer RM. A combined p-value test for multiple hypothesis testing. Journal of Statistical Planning and Inference. 2013; 143:764–770.

Zhang NR, Siegmund DO, Ji H, Li JZ. Detecting simultaneous changepoints in multiple sequences. Biometrika. 2010; 97:631–645. [PubMed: 22822250]



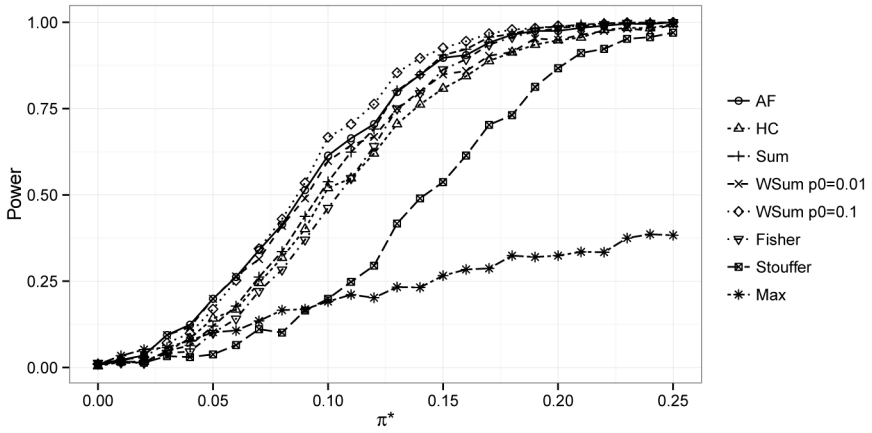
(a) Power for detecting a single rare change-point.



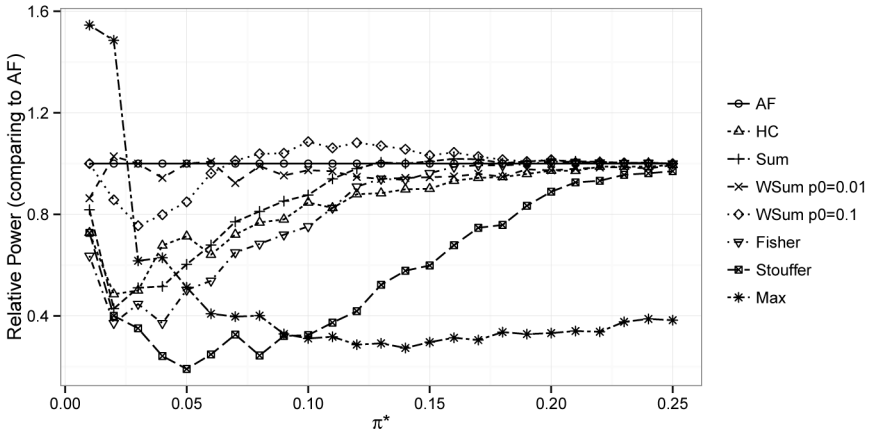
(b) Power for detecting a single common change-point.

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) Relative power of different methods as π^* changes.

Fig 2.

Simulation result for single change-point detection as $\boldsymbol{\pi}^*$ 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). To make the differences clear enough to see, the relative power of different methods comparing to adaptive Fisher's method is plotted in (b), where the relative power is calculated as the original power divided by the power of adaptive Fisher's method.

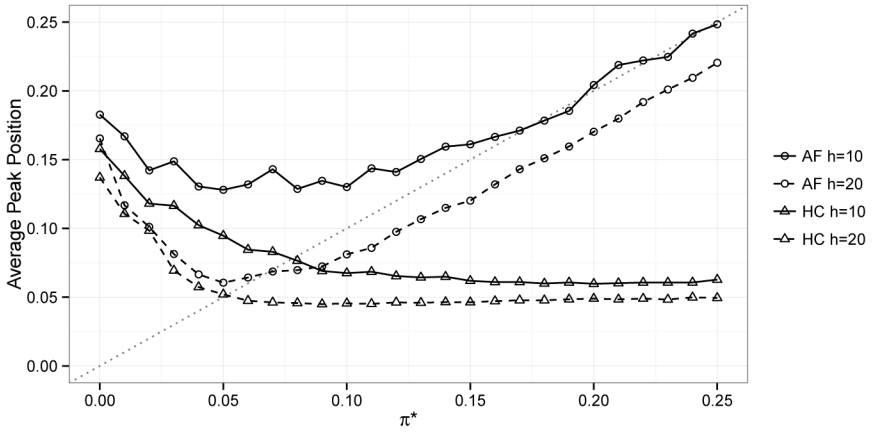


Fig 3.

Average adaptive peak position of adaptive Fisher's statistics and higher criticism statistic, where the dotted line shows the true proportion of sample carriers.

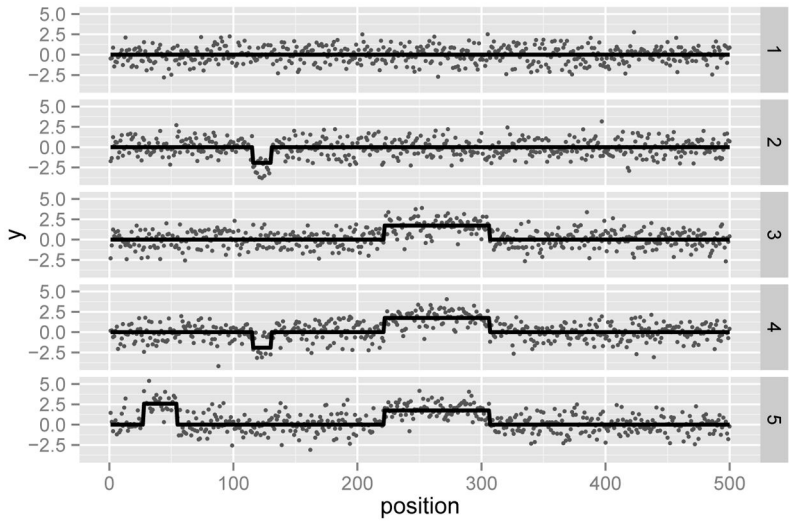
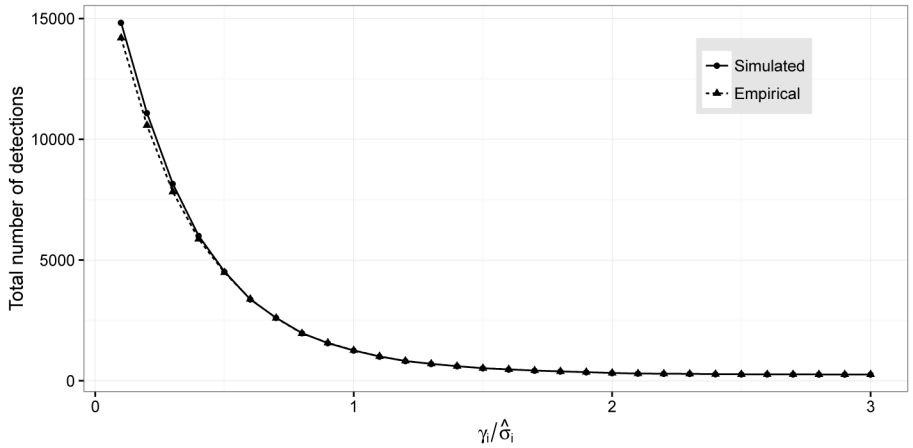
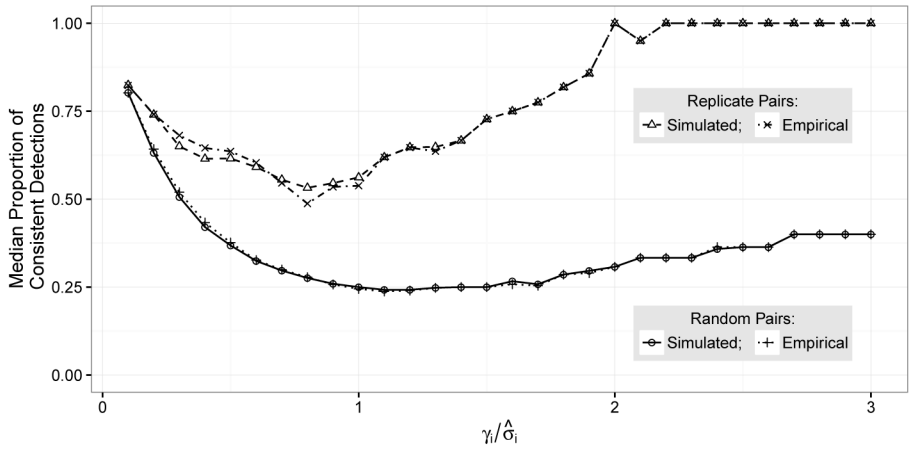


Fig 4.

The simulated data with no trend. Five samples are shown. The mean signals without noise are shown by bold black lines.



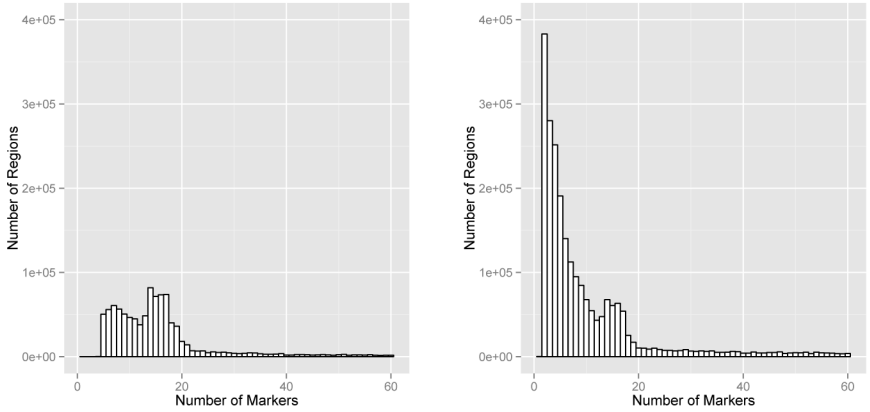
(a) Number of total detections.



(b) Proportion of consistent detections.

Fig 5.

Results of the real data analysis by the proposed method as $\gamma_i/\hat{\sigma}_i$ changes from 0.1 to 3. λ was determined using two approaches: first as the 99.99-th percentile of the simulated null distribution of $W(t, h)$ on local maximizers (denoted as "Simulated"), and then as the 50-th percentile from the corresponding observed distribution (denoted as "Empirical"). The total number of detections in the 20 replicate pairs is given in (a). The median proportions of consistent detections in the 20 replicate pairs and in 1000 randomly selected pairs of samples are given in (b).



(a) Regions detected with $h = 5, 10, 15$ (b) Regions detected with $h = 2, 5, 10, 15$

Fig 6.

The histograms of the number of markers between change-points detected in the GLAUGEN data. Regions with no more than 60 markers were shown. The results based on three bandwidths (h = 5, 10, 15) are shown in (a). The results based on four bandwidths (h = 2, 5, 10, 15) are shown in (b).

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

Table 2

True and false positives grouped by the change-points (CP1-CP6) for the simulation with no trend. Standard errors are shown in parentheses.

(a) Average number of true positives.

Number of Carriers	CP1 20	CP2 20	CP3 50	CP4 50	CP5 100	CP6 100
fast CBS	19.9(0.3)	19.9(0.3)	47.6(1.5)	47.6(1. 5)	94.5(2.3)	94.5(2.2)
CBS-SS	19.9(0.3)	19.9(0.3)	47.5(1.5)	47.6(1. 5)	94.5(2.3)	94.5(2.2)
m-SaRa	19.8(0.4)	19.8(0.4)	47.3(1.6)	47.2(1. 6)	92.0(2.7)	91.8(2.6)
Multiple-sample CBS	20.0(0.0)	20.0(0.0)	50.0(0.1)	50.0(0. 1)	100.0(0.0)	100.0(0.0)
Multiple-sample m-SaRa	20.0(0.0)	20.0(0.0)	50.0(0.1)	50.0(0. 1)	100.0(0.0)	100.0(0.0)

(b) Average number of false positives.

	CP1	CP2	CP3	CP4	CP5	CP6
fast CBS	0.3(0.6)	0.3(0.6)	0.3(0.5)	0.3(0.5)	0.2(0.5)	0.2(0.5)
CBS-SS	0.2(0.4)	0.2(0.5)	0.2(0.5)	0.2(0.5)	0.2(0.5)	0.2(0.5)
m-SaRa	2.9(1.7)	4.1(2.0)	4.7(2.2)	5.1(2.2)	5.0(2.2)	5.1(2.2)
Multiple-sample CBS	2.7(1.1)	2.3(1.1)	2.8(1.0)	2.8(1.0)	1.3(0.3)	1.2(0.3)
Multiple-sample m-SaRa	0.2(0.5)	0.1(0.3)	0.3(0.6)	0.3(0.6)	0.0(0.0)	0.0(0.0)

Table 3

The number of shared change-points detected for the simulation with trend related to GC content.

	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	41	**332**	242	203	182
Multiple-sample m-SaRa	0	**998**	2	0	0

Table 4

The number of shared change-points detected for the simulation with trend related to GC content.

	Number of change-points				
Method	**≤ 5**	**6**	**7**	**8**	**> 8**
fast CBS	0	**52**	206	364	378
CBS-SS	0	**64**	237	371	328
m-SaRa	0	**0**	0	0	1000
Multiple-sample CBS	0	**1000**	0	0	0
Multiple-sample m-SaRa	0	**1000**	0	0	0

Table 5

The median proportions of consistent detections in 20 pairs of technical replicates and in 1000 random pairs. S = Simulated; E = Empirical.

		Median Proportion		
Method	**Settings**	**Replicate Pairs**	**Random Pairs**	**Total Detections**
Multiple-sample m-SaRa (h = 5, 10, 15)	S, k = 1.2	0.648	0.242	820
	E, k = 1.2	0.646	0.240	824
	S, k = 0.5	0.616	0.369	4521
	E, k = 0.5	0.636	0.377	4483
Multiple-sample m-SaRa (h = 2, 5, 10, 15)	S, k = 1.6	0.667	0.222	858
	E, k = 1.6	0.636	0.205	922
	S, k = 0.7	0.507	0.290	4962
	E, k = 0.7	0.457	0.275	5940
Multiple-sample CBS	p_0 = 1	0.616	0.311	4912
	p_0 = 0.1	0.618	0.311	4918
	p_0 = 0.01	0.594	0.305	4865
PennCNV		0.558	0.2	903

Table 6

Running time for CNV detection on chromosome 22 in 1, 363 samples.

	Method		
	Multiple-sample m-SaRa	**Multiple-sample CBS**	**PennCNV**
Time	66 sec.	~ 200, 000 sec.	~ 5, 000 sec.