Biometrika (2013), **100**, 1, *pp.* 157–172
doi: 10.1093/biomet/ass059
Advance Access publication 4 December 2012

Simultaneous discovery of rare and common segment variants

By X. JESSIE JENG

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.

xjeng@mail.med.upenn.edu

T. TONY CAI

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

tcai@wharton.upenn.edu

and HONGZHE LI

Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, U.S.A.

hongzhe@upenn.edu

Summary

Copy number variant is an important type of genetic structural variation appearing in germline DNA, ranging from common to rare in a population. Both rare and common copy number variants have been reported to be associated with complex diseases, so it is important to identify both simultaneously based on a large set of population samples. We develop a proportion adaptive segment selection procedure that automatically adjusts to the unknown proportions of the carriers of the segment variants. We characterize the detection boundary that separates the region where a segment variant is detectable by some method from the region where it cannot be detected. Although the detection boundaries are very different for the rare and common segment variants, it is shown that the proposed procedure can reliably identify both whenever they are detectable. Compared with methods for single-sample analysis, this procedure gains power by pooling information from multiple samples. The method is applied to analyse neuroblastoma samples and identifies a large number of copy number variants that are missed by single-sample methods.

Some key words: DNA copy number variant; Information pooling; Population structural variant.

1. Introduction

Copy number variant is a type of DNA structural variation that results in the genome having abnormal numbers of copies of DNA segments. Copy number variants correspond to relatively large regions of the genome that have been deleted or duplicated on certain chromosomes (Zhang et al., 2009). Copy number variants can be inherited or caused by de novo mutations and have been shown to be associated with complex diseases such as cancer (Diskin et al., 2009). Such associations can involve both rare and common variants. Since recent genome-wide association studies have shown that common variants can explain only a small fraction of heritabilities of complex phenotypes, genetic studies of rare variants, including rare copy number variants, have become even more important.

An important problem is to identify all the copy number variants in the human genome, including both the rare and common ones, in order to have a complete variant catalog for future association and population genetics analysis. While efficient procedures have been developed for identifying variants in a long sequence of genome-wide observations, they mostly focus on identification based on data from a single sample. Important examples include the optimal likelihood ratio selection method (Jeng et al., 2010), the hidden Markov model-based method (Wang et al., 2007) and change-point based methods (Olshen et al., 2004). To identify the recurrent copy number variants that appear in multiple samples, some type of post-processing is often used. This type of procedure first identifies copy number variants based on individual samples and then selects regions with highly recurrent variants. One problem with such an approach is that the power for identifying the recurrent variants does not improve with the increase in the size of samples. The locations of a recurrent copy number variant mostly overlap across samples, so identification power can be improved if information from multiple samples can be efficiently pooled. In addition, most variants from the germline constitutional genome have a range of less than 20 single nucleotide polymorphisms (Zhang et al., 2009) in a typical Illumina 660K chip. Many of these short variants cannot be identified even by the optimal method based on data from a single sample (Jeng et al., 2010). Efficiently pooling information from multiple samples can greatly benefit the discovery of short variants that are missed in single-sample analysis.

Methods for simultaneously detecting rare and common copy number variants based on a large set of population samples have not been fully developed. Zhang et al. (2010) introduced a method for detecting simultaneous change-points in multiple sequences that is effective only for detecting the common variants. Siegmund et al. (2010) extended their method by introducing a prior variant frequency that needs to be specified. No rigorous power studies were given in these papers. For common variants, the power of identification can be increased by summing up the test statistics over all the samples (Zhang et al., 2010). This approach, however, fails for the rare copy number variant identification because the information of the few signals can be diluted greatly. For rare copy number variants, methods based on outliers of test statistics over all the samples can be more efficient (Siegmund et al., 2010). There is a need for a unified and theoretically justifiable approach that can identify both rare and common copy number variants simultaneously.

In this paper, we propose a proportion adaptive segment selection procedure, which is optimally adaptive to the unknown proportions of the carriers of the segment variants. At its core is an efficient scanning algorithm based on a test statistic that is sensitive to both the rare and common segment variants. To study the theoretical properties of this procedure, we first characterize the detection boundary that separates the region where a segment variant is detectable by some method from the region where it cannot be detected by any method. The results show that the detection boundaries are very different for the rare and common segment variants. Despite the significant differences, it is shown that this adaptive procedure can simultaneously identify both the rare and common segment variants whenever they are detectable. This procedure automatically adapts to the unknown proportions of the carriers of the segment variants.

Compared with single-sample analysis, the proposed adaptive procedure gains power by pooling information from multiple samples; compared with other information pooling methods, it provides a unified approach to identify a wide range of copy number variants. In addition to DNA copy number analysis, the proposed method can also be used for applications where the detection of recurrent signal segments is of interest. One example is to detect linear objects in images with multiple looks where information pooling also sheds light on the discovery of common and subtle objects.

2. Statistical model and method

2·1. *Statistical model for multisample copy number variation analysis*

Suppose there are N linear sequences or samples of noisy data and that each sequence has T observations. Let X_{it} be the observed data for the ith sample at the tth location. If there are no signal variations, X_{it} varies around zero for any i and t. Suppose that at certain non-overlapping segments or subintervals $I_1, \ldots, I_q$, some samples have elevated or dropped means from the baseline and others do not. We call the samples that carry the variation the carriers. Denote the collection of the non-overlapping segments by $\mathbb{I} = \{I_1, \ldots, I_q\}$, the carrier proportion at segment I_k in the population by π_k, and the magnitude of the segment for sample i by A_{ik}. Then an observation for sample $i \in \{1, \ldots, N\}$ at location $t \in \{1, \ldots, T\}$ can be modelled as

$$X_{it} = A_{ik} 1_{\{t \in I_k\}} + Z_{it}, \quad I_k \in \mathbb{I}, \tag{1}$$

with

$$A_{ik} \sim (1 - \pi_k)\delta_0 + \pi_k N(\mu_k, \tau_k^2), \tag{2}$$

where δ_0 is a point mass at zero, $\mu_k \neq 0$, and $Z_{it} \sim N(0, \sigma_i^2)$. The noise variance σ_i^2 for sample i can be easily estimated when T is large and the signal segments are sparse in the linear sequence of data for sample i. For example, the robust median absolute deviation estimate can be applied. Without loss of generality, we assume $\sigma_i^2 = 1$ in the theoretical analysis. All of the other parameters $I_k, \pi_k, \mu_k, \tau_k$ $(k = 1, \ldots, q)$ are unknown. From this model, if t is not in any signal segment, X_{it} is Gaussian noise following $N(0, \sigma_i^2)$. If t is in a signal segment I_k, then

$$X_{it} \sim (1 - \pi_k)N(0, \sigma_i^2) + \pi_k N(\mu_k, \sigma_i^2 + \tau_k^2). \tag{3}$$

This Gaussian mixture is both heterogenous and heteroscedastic. The τ_k of the second component represents the additional variability introduced by the different magnitudes of signal segments in the population.

Our goal is to detect the existence of recurrent segment variants across samples; and to identify the locations of the segments. Precisely, we wish to first test

$$H_0 : \mathbb{I} = \emptyset, \quad H_1 : \mathbb{I} \neq \emptyset,$$

and if H_0 is rejected, detect each $I_k \in \mathbb{I}$. The segment variants can be classified as rare and common based on their carrier proportions in the population. Specifically, we say that

$$I_k \text{ corresponds to a rare variant if } \pi_k \leqslant N^{-1/2}, \tag{4}$$

$$I_k \text{ corresponds to a common variant if } \pi_k > N^{-1/2}. \tag{5}$$

The separation boundary $N^{-1/2}$ is frequently seen in large-sample theory. See, for example, Cai et al. (2011) for a similar classification of recurrent signals. For the common variants with $\pi_k > N^{-1/2}$, classical large-sample theory implies that methods based on the sample mean are efficient. For the rare variants with $\pi_k \leqslant N^{-1/2}$, classical results cannot be applied and new theoretical and methodological developments are needed.

2·2. *Proportion adaptive segment selection procedure*

We now introduce the proportion adaptive segment selection procedure, which performs an efficient scan over long linear sequences of data based on a test statistic that is sensitive to both the rare and common segment variants. The procedure utilizes the short-segment structure of the signals by considering only intervals of length at most L, where $L \ll T$. Denote the set of these intervals by $\mathbb{J}_{T,N}(L)$. The choice of L should satisfy the condition

$$L \geqslant \bar{s}, \tag{6}$$

where $\bar{s}$ is the length of the longest signal segments. This condition is easily satisfied when signal segments are short, as often seen in copy number variants.

For any interval $J \in \mathbb{J}_{T,N}(L)$, we calculate the standardized sum of observations in J for each sample i as

$$X_{J,i} = \sum_{t \in J} X_{it}/|J|^{1/2} \quad (i = 1, \ldots, N), \tag{7}$$

where $|J|$ denotes the length of the interval J. By (1) and the assumption $\sigma_i^2 = 1$, $X_{J,i} \sim N(0, 1)$ under H_0. When J overlaps with some signal segment, $X_{J,i}$ follows a heterogeneous and heteroscedastic Gaussian mixture according to (3). Specifically, when $J = I_k$ for some $I_k \in \mathbb{I}$,

$$X_{I_k,i} \sim (1 - \pi_k)N(0, \sigma_i^2) + \pi_k N(\mu_k |I_k|^{1/2}, \sigma_i^2 + \tau_k^2). \tag{8}$$

The mean of the second component includes the value of jump size μ_k and length of the segment variant at I_k.

Based on the $X_{J,i}$ statistic, we pool information from multiple samples by calculating the extreme value of the standardized ordered p-values of $X_{J,i}$ $(i = 1, \ldots, N)$. The p-values of $X_{J,i}$ are two-sided, and the ordered p-values are denoted by $p_{J,(1)} \leqslant \cdots \leqslant p_{J,(N)}$. The standardized ordered p-values are defined as

$$W_{J,(i)} = N^{1/2} \frac{i/N - 2p_{J,(i)}}{\{2p_{J,(i)}(1 - 2p_{J,(i)})\}^{1/2}} \quad (i = 1, \ldots, N). \tag{9}$$

Since the p-values are uniformly distributed under H_0, the $W_{J,(i)}$ comprise a standardized uniform empirical process and its extreme value has a well studied distribution (Shorack & Wellner, 2009, pp. 596–600). We use the extreme value $V_N(J) = \max_{\alpha_0 \leqslant i \leqslant N/2} W_{J,(i)}$ for some small α_0 as our test statistic. If J overlaps with some signal segment, the distribution of the test statistic deviates to the positive side. An interval is selected if its test statistic passes a certain threshold and achieves a local maximum. Since the signal intervals are not known, this procedure examines all the overlapping intervals of length $\leqslant L$ and chooses the intervals that achieve a local maximum of the extreme values $\{V_N(J), J \in \mathbb{J}_{T,N}(L)\}$. The detailed algorithm is given as follows.

Step 1. For each long sequence of data $\{X_{it} : t = 1, \ldots, T\}$, standardize the data by subtracting the sample median and dividing by the median absolute deviation estimate of σ_i.

Step 2. Set the maximum interval length L and denote by $\mathbb{J}_{T,N}(L)$ the collection of the intervals with length less than or equal to L.

Step 3. For any given interval $J \in \mathbb{J}_{T,N}(L)$, calculate $X_{J,i}$ as in (7) and the two-sided p-values of $X_{J,i}$ as $p_{J,i}$ =pr$\{N(0, 1) > |X_{J,i}|\}$ $(i = 1, \ldots, N)$.

Step 4. Order the p-values as $p_{J,(1)} \leqslant \cdots \leqslant p_{J,(N)}$ and calculate the standardized empirical process of the p-values as $W_{J,(i)}$ in (9).

Step 5. Calculate the test statistic for each interval $J \in \mathbb{J}_{T,N}(L)$ as

$$V_N(J) = \max_{\alpha_0 \leqslant i \leqslant N/2} W_{J,(i)}, \tag{10}$$

for some small $\alpha_0 > 1$ and pick candidate set

$$\mathbb{I}^{(1)} = \{J \in \mathbb{J}_{T,N}(L) : V_N(J) > \lambda_{T,N}\} \tag{11}$$

for some threshold $\lambda_{T,N}$. If $\mathbb{I}^{(1)} \neq \emptyset$, we reject the null hypothesis, set $j = 1$, and proceed to the following steps.

Step 6. Let $\hat{I}_j = \arg\max_{J \in \mathbb{I}^{(j)}} V_N(J)$, and update $\mathbb{I}^{(j+1)} = \mathbb{I}^{(j)} \backslash \{J \in \mathbb{I}^{(j)} : J \cap \hat{I}_j \neq \emptyset\}$.

Step 7. Repeat Steps 6–7 with $j = j + 1$ until $\mathbb{I}^{(j)}$ is empty.

Step 8. Define the collection of selected intervals as $\hat{\mathbb{I}} = \{\hat{I}_1, \hat{I}_2, \ldots\}$ and identify the signal segments as all the elements in $\hat{\mathbb{I}}$.

After the test statistic $V_N(J)$ is calculated for each interval $J \in \mathbb{J}_{T,N}(L)$, a threshold $\lambda_{T,N}$ is set based on the distribution of $V_N(J)$ under H_0. Since all the intervals in $\mathbb{J}_{T,N}(L)$ are considered, the threshold $\lambda_{T,N}$ needs to adjust for multiple testing, so that the familywise Type I error is controlled at a desired level. Section 3·1 provides a detailed discussion on setting $\lambda_{T,N}$ theoretically or by simulations.

Steps 6–7 find all the local peaks in the candidate set $\mathbb{I}^{(1)}$. Intuitively, if the signal segments are well separated, the test statistic $V_N(I_k)$ of a signal segment I_k is larger than those of other intervals overlapping I_k, so that the local peaks provide good estimates of the signal segments.

Remark 1. The tuning parameter α_0 in (10) is used to stabilize the procedure and to better control the familywise error with finite samples. This parameter excludes the endpoints of the standardized uniform empirical process, where the process diverges. The rationale is that the convergence of the extreme values without truncation is much slower than the convergence of a truncated version. By choosing $\alpha_0 > 1$, $V_N(J)$ can be more stable in finite samples, which also leads to a smaller threshold on $V_N(J)$ to control the overselection, and higher power for detecting signals with small intensity.

Remark 2. The test statistic $V_N(J)$ is closely related to some other test statistics based on a standardized uniform empirical process, such as the Anderson & Darling (1952) statistic and higher criticism (Donoho & Jin, 2004). However, the setting is different here and the adaptivity of $V_N(J)$ to rare and common segment variants is an interesting new discovery.

The fact that the test statistic $V_N(J)$ is able to capture the signal information of both the rare and common variants can be illustrated by simulation. In this example, the sample size $N = 1600$, so that the separating value for the carrier proportion is $N^{-1/2} = 2{\cdot}5\%$. The data are generated from models (1)–(2) with $T = 10\,000$ and $\sigma_i^2 = 1$ $(i = 1, \ldots, N)$. Two locations are randomly selected for segment variants with length $|I_k| = 10$ for $k = 1, 2$. The first location has a rare variant with $\pi_1 = 1\%$ and $\mu_1 = 1$; the second location has a common variant with $\pi_1 = 50\%$ and $\mu_1 = 0{\cdot}1$. Fix $\tau_k = 0{\cdot}7$ for $k = 1, 2$. Figure 1 shows $W_{I_1,(i)}$ and $W_{I_2,(i)}$ $(i = \alpha_0, \ldots, N)$ with $\alpha_0 = 4$. The samples are ordered by their p-values of $X_{I_k,i}$, and the $W_{I_k,(i)}$ statistics are plotted in

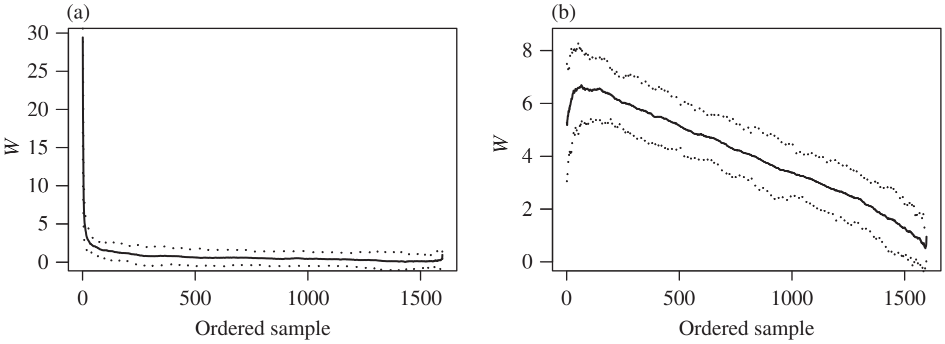

Fig. 1. Illustrations of $W_{I_1,(i)}$ and $W_{I_2,(i)}$ $(i = \alpha_0, \ldots, N)$. (a): mean ± median absolute deviation of $W_{I_1,(i)}$ over 100 replications for a rare segment variant with relatively large magnitude. (b): mean ± median absolute deviation of $W_{I_2,(i)}$ over 100 replications for a common segment variant with small magnitude.

the same order. The signal information shows up at different places in these two plots. In Fig. 1(a), $W_{I_1,(i)}$ reaches the peak at the left end, where the small number of large signals locate; whereas in Fig. 1(b), $W_{I_2,(i)}$ reaches the peak to the right of the left end, where the information of many small signals lumps up. According to (10), the test statistics $V_N(I_1)$ and $V_N(I_2)$ capture the peaks in these two cases respectively. This is essentially the reason for the strong detection power of the proportion adaptive segment selection for both the rare and common segment variants.

3. Optimal adaptivity of the proportion adaptive segment selection

3·1. *Familywise error control*

In this section, we show that under H_0, the proportion adaptive segment selection procedure with a theoretical threshold asymptotically controls the familywise error. The theoretical threshold is constructed based on the limiting distribution of $V_N(J)$ as $N \to \infty$ under H_0. Define

$$a_N = (2 \log \log N)^{1/2}, \quad b_N = 2 \log \log N.$$

Then $a_N V_N(J) - b_N$ converges to a nondegenerate random variable (Shorack & Wellner, 2009, p. 600). Since all TL intervals in $\mathbb{J}_{T,N}(L)$ are considered, the theoretical threshold is defined as

$$\lambda_{T,N} = \{C_0 \log(TL) + b_N\}/a_N \tag{12}$$

for some $C_0 > 1$. The following theorem shows that the procedure asymptotically controls the familywise error for any fixed T.

Theorem 1. *Assume models* (1)–(2). *The candidate set* $\mathbb{I}^{(1)}$ *constructed in* (11) *with* $\lambda_{T,N}$ *defined in* (12) *is empty with high probability under* H_0. *More specifically, we have*

$$\text{pr}(\mathbb{I}^{(1)} \neq \emptyset) \leqslant C_1(TL)^{-(C_0-1)}$$

for any fixed T *and all sufficiently large* N, *where* $C_1 > 0$ *is a constant and* $C_0 > 1$ *is defined in* (12).

The proof of Theorem 1 is given in the Appendix. The proof for the familywise error when T increases with N is more involved due to the calculation of the convergence rate for the extreme value of the standardized empirical process. A detailed proof is outside the scope of this paper.

Table 1. *Means and standard deviations (in parentheses) of the data-driven thresholds for V_N over 100 replications*

	$\alpha_0 = 1$	$\alpha_0 = 2$	$\alpha_0 = 4$	$\alpha_0 = 7$	$\alpha_0 = 10$	$\alpha_0 = 13$
$\#O = 5$	47·7 (10·9)	12·3 (1·6)	6·9 (0·4)	5·6 (0·3)	5·2 (0·2)	5·0 (0·2)
$\#O = 2$	75·5 (26·7)	15·2 (2·3)	7·7 (0·7)	6·0 (0·3)	5·5 (0·3)	5·3 (0·3)
$\#O = 0$	178·9 (163·3)	24·2 (8·9)	9·8 (2·0)	7·1 (1·0)	6·3 (0·6)	5·9 (0·5)

$\#O$, number of overselected intervals; α_0, number of observations left out.

The convergence is slow in general, mainly due to the end points of the interval [0, 1] as shown in Shorack & Wellner (2009), § 16.1. By choosing $\alpha_0 > 1$, the test statistic $V_N(J)$ is more stable and its familywise error better controlled under H_0. While the precise choice of α_0 in practice is difficult, simulation results in Table 1 can provide useful guidance.

Although Theorem 1 shows that the Type I error can be controlled when T and N are sufficiently large, in finite sample situations the convergence of $V_N(J)$ as $N \to \infty$ can be slow for small α_0. In addition, it is difficult to choose the constants C_0 and C_1 in setting the threshold. In simulations and real data analysis, we suggest using simulations to determine a data-driven threshold to control the number of overselections. Section 4·1 presents the details.

3·2. *Detection power of the proportion adaptive segment selection*

Under the alternative hypothesis, the proposed procedure asymptotically selects either the true signal segments or short intervals overlapping the true segments, whenever the signal segments are detectable. We characterize the detection boundary that separates the region where a segment variant is detectable by some method from the region where it cannot be detected by any methods. If a method can reliably detect a segment variant whenever it is detectable, we say that the method is optimal. If a method applies a unified approach to optimally detect both rare and common segment variants without using the information of their carrier proportions and other unknown parameters, we say the method is optimally adaptive to the carrier proportions and other unknown parameters in the model.

The segment variants can be characterized into the rare and common groups by (4) and (5). We calibrate π_k as

$$\pi_k = N^{-\beta_k}, \quad 0 \leqslant \beta_k < 1. \tag{13}$$

If $1/2 \leqslant \beta_k < 1$, I_k corresponds to a rare variant, and if $0 \leqslant \beta_k < 1/2$, I_k corresponds to a common variant. Extremely rare variants with carrier proportion at the order of N^{-1} are not considered here. For a fixed I_k and a sample i, the sufficient statistic $X_{I_k,i}$ defined in (7) follows a Gaussian mixture distribution as in (8). Since the mean of the nonnull component has absolute value $|\mu_k||I_k|^{1/2}$, we calibrate $|\mu_k||I_k|^{1/2}$ for rare and common variants respectively as

$$|\mu_k||I_k|^{1/2} = (2r_k \log N)^{1/2}, \quad r_k > 0, \qquad 1/2 \leqslant \beta_k < 1, \tag{14}$$

$$|\mu_k||I_k|^{1/2} = N^{-r_k}, \quad r_k \geqslant 0, \qquad 0 \leqslant \beta_k < 1/2. \tag{15}$$

For a rare variant, the carrier proportion is so small that the signal intensity, which is represented by $|\mu_k||I_k|^{1/2}$, must be sufficiently large to make the variant detectable; whereas for a common variant, the carrier proportion is so large that a small signal intensity can be amplified to make the detection successful. This is the reason for the different calibrations of $|\mu_k||I_k|^{1/2}$. These calibrations are similar to those of signal intensity in Cai et al. (2011), where detection of Gaussian mixtures is considered.

We find the detection boundary for the rare and common variants, respectively. Let

$$m(\tau_k) = \min\left(\frac{1+\tau_k^2}{4}, \frac{1}{1+\tau_k^2}\right).$$

If $1/2 \leqslant \beta_k < 1$, define

$$\rho^+(\beta_k,\tau_k) = \begin{cases} [1-\{(1+\tau_k^2)(1-\beta_k)\}^{1/2}]^2, & 1-m(\tau_k) < \beta < 1, \\ (1-\tau_k^2)_+(\beta_k - 1/2), & 1/2 \leqslant \beta_k < 1-m(\tau_k); \end{cases}$$

and if $0 \leqslant \beta_k < 1/2$, define

$$\rho^-(\beta_k,\tau_k) = \begin{cases} 1/2-\beta_k, & \tau_k = 0, \\ \infty, & \tau_k > 0. \end{cases}$$

The following proposition shows that $\rho^+(\beta_k,\tau_k)$ and $\rho^-(\beta_k,\tau_k)$ are the separating lines between the detectable and undetectable regions for the rare and common segment variants.

Proposition 1. *Assume models* (1)–(2) *and* (6). *Suppose I_1 is known and π_1 and $\mu_1|I_1|^{1/2}$ are calibrated as in* (13), (14)*, and* (15). *If $r_1 > \rho^+(\beta_1,\tau_1)$ for $1/2 \leqslant \beta_1 < 1$, or if $r_1 < \rho^-(\beta_1,\tau_1)$ for $0 \leqslant \beta_1 < 1/2$, there exists a consistent test for $H_0^*:\pi_1 = 0$ vs $H_1^*:\pi_1 > 0$ for which the sum of Type I and Type II error probabilities tends to 0 as $N \to \infty$. If $r_1 < \rho^+(\beta_1,\tau_1)$ for $1/2 \leqslant \beta_1 < 1$, or if $r_1 > \rho^-(\beta_1,\tau_1)$ for $0 \leqslant \beta_1 < 1/2$, a consistent test does not exist.*

Proposition 1 is an extension of Theorems 2.1–2.4 in Cai et al. (2011), where, based on a random sample $\{Y_1,\ldots,Y_N\}$, the problem is to test

$$H_0^{(N)}: Y_i \sim N(0,1), \quad H_1^{(N)}: Y_i \sim (1-\epsilon)N(0,1) + \epsilon N(A,\sigma^2),$$

where ϵ, A, and σ^2 are unknown parameters. Here the information in I_1 for sample i is summarized by the sufficient statistic $X_{I_1,i}$, then for I_1, the sufficient statistics of the N observations, $X_{I_1,1},\ldots,X_{I_1,N}$, are treated as a random sample. It is easy to see that $X_{I_1,i} \sim N(0,1)$ under H_0^* and $X_{I_1,i} \sim (1-\pi_1)N(0,1) + \pi_1 N(\mu_1|I_1|^{1/2}, 1+\tau_1)$ under H_1^*. Therefore, the mixture model of $X_{I_1,i}$ under H_1^* is a special case of the mixture model of Y_i under $H_1^{(n)}$ with $\sigma^2 > 1$. So the detection boundary for the segment variant at I_1 can be derived in a similar way as that in Cai et al. (2011). We omit the proof here.

The detection boundary can be used as a benchmark to evaluate the performance of a method theoretically. Our problem is more difficult than detecting Gaussian mixtures at a fixed interval, because the locations $I_1,\ldots,I_q$ are unknown. The proportion adaptive segment selection procedure first pools information across samples for all intervals in $\mathbb{J}_{N,T}(L)$ and then searches through these intervals to detect segment variants. The following theorem states that as long as the length of the sequence T is not too large compared to the sample size N, any segment variant in the detectable region is included with a high probability in the candidate set of the proportion adaptive selection, implying that the proportion adaptive segment selection procedure with the theoretical threshold is an asymptotically optimal procedure for detecting segment variants. Furthermore, the implementation of the proportion adaptive segment selection does not require the information of $\{q,\pi_k,I_k,\mu_k,\tau_k : k = 1,\ldots,q\}$. Therefore, the procedure is asymptotically optimally adaptive to all the unknown parameters in the model.

THEOREM 2. *Assume models (1)–(2) and (6). Suppose* $\mathbb{I} \neq \emptyset$, *and for any* $I_k \in \mathbb{I}$, *calibrate* π_k *and* $\mu_k|I_k|^{1/2}$ *as in (13), (14), and (15). In addition, assume* $N^C \gg \log T$ *for any* $C > 0$ *and* $\alpha_0 = o(N^C)$ *for any* $C > 0$. *Then, if* $r_k > \rho^+(\beta_k, \tau_k)$ *for* $1/2 \leqslant \beta_k < 1$ *or if* $r_k < \rho^-(\beta_k, \tau_k)$ *for* $0 \leqslant \beta_k < 1/2$, *we have*

$$\mathrm{pr}(I_k \in \mathbb{I}^{(1)}) \geqslant 1 - CN^{-C(r_k,\beta_k,\tau_k)} \to 1,$$

where

$$C(r_k, \beta_k, \tau_k) = \begin{cases} 1 - \beta_k - (1 - r_k^{1/2})^2/(1 + \tau_k^2), & \tau_k \geqslant 1,\ 1/2 < \beta < 1, \\ 1 - \beta_k - (1 - r_k^{1/2})^2/(1 + \tau_k^2), & \tau_k < 1,\ 1 - m(\tau_k) \leqslant \beta_k < 1, \\ \min\{1 - \beta_k - (1 + \tau_k^2) r_k/(1 - \tau_k^2)^2, & \\ 1 - 2\beta_k + 2r_k/(1 - \tau_k^2)\}, & \tau_k < 1,\ 1/2 < \beta_k < 1 - m(\tau_k), \\ 1 - 2\beta_k - 2r_k, & 0 \leqslant \beta_k < 1/2,\ \tau_k = 0, \\ 1 - 2\beta_k, & 0 \leqslant \beta_k < 1/2,\ \tau_k > 0. \end{cases} \tag{16}$$

Clearly, the convergence rate is larger for smaller β_k, which corresponds to a larger carrier proportion. The interval I_k being in $\mathbb{I}^{(1)}$ implies that either I_k itself or a short interval overlapping with I_k is selected in the final collection $\hat{\mathbb{I}}$. In applications, follow-up studies rarely just look at the selected intervals but rather examine small regions covering the selected intervals to verify the exact locations of signal segments.

In order to see the power gain of the proportion adaptive segment selection by pooling information from multiple samples, we consider the situation when only one sequence of data with length T is available. In this situation, a theoretically optimal likelihood ratio selection has been developed in Jeng et al. (2010). For the likelihood ratio selection to successfully detect a signal segment I_1, the condition on $\mu_1|I_1|^{1/2}$ is $\mu_1|I_1|^{1/2} \geqslant \{2(1 + \epsilon_n)\log T\}^{1/2}$ for some $\epsilon_n = o(1)$. This condition is in general stronger than the condition on $\mu_1|I_1|^{1/2}$ in Theorem 2, which is $\mu_1|I_1|^{1/2} \geqslant C(\log N)^{1/2}$ for $1/2 \leqslant \beta_1 < 1$ or $\mu_1|I_1|^{1/2} \geqslant N^{-C}$ for $0 \leqslant \beta_1 < 1/2$, when T is much larger than N. In high-throughput copy number variation data analysis, T is usually above 500 000 and N mostly under 1000. Significant power gains can be achieved especially for detecting common variants.

4. SIMULATION STUDIES

4·1. *Choice of* α_0

The parameter α_0 in (10) of the algorithm determines how many end points of the empirical process of the p-values are left out from the test statistic V_N. By choosing $\alpha_0 > 1$, V_N can be more stable with finite samples, which also leads to a smaller threshold on V_N to control over-selection, and higher power for detecting small-intensity signals. To demonstrate this, Table 1 shows the mean and standard deviation of the data-driven threshold based on 100 replications of simulated data. In each replication, we generate 400 sequences, and each sequence has 5000 observations generated from the standard normal distribution. We apply our procedure with $L = 6$. The data-driven threshold is defined as the smallest threshold to guarantee that there is no more than a prespecified number of intervals in $\hat{\mathbb{I}}$ that do not overlap with any of the segments in $\mathbb{I}$. The value of α_0 has a great effect on the threshold and therefore on the power of the test statistic V_N. Since $\alpha_0 - 1$ samples with the extreme p-values are left out from the test statistic, the proposed procedure cannot be very effective in identifying extremely rare copy number variants.

Table 2. *Empirical power* (%) *and standard error* (*in parentheses*) *of the proportion adaptive segment selection and the single-sample method over* 100 *replications*

	$\mu = 0{\cdot}5$	$\mu = 0{\cdot}7$	$\mu = 0{\cdot}9$	$\mu = 1{\cdot}1$
PASS	21 (4·0)	54 (4·9)	89 (3·3)	100 (0·0)
LRS	22 (4·2)	29 (4·7)	38 (4·9)	59 (4·9)

PASS, proportion adaptive segment selection; LRS, single-sample method of Jeng et al. (2010).

4·2. *Improvement over the single-sample method*

In this section, simulation studies are carried out to investigate the numerical performance of the proportion adaptive segment selection and to compare it with other methods. In the following simulations, we set $N = 400$, $T = 5000$, and $\sigma_i^2 = 1$ for each sample i.

We begin by considering testing H_0 against H_1. The power gain of information pooling is shown by comparing the performance of the proportion adaptive segment selection with the single-sample method in Jeng et al. (2010), which scans through all the intervals of length at most L for each sample and calculates the sufficient statistic for each interval as in (7). The single-sample method rejects H_0 if the extreme value of all the sufficient statistics has absolute value greater than $\{2\log(NTL)\}^{1/2}$. This is derived from the distribution theory of the extreme value under H_0. The single-sample method does not utilize the information that the locations of a recurrent variant across samples are mostly overlapping.

To assess the Type I error, we generate each $X_{it} \sim N(0, 1)$ and calculate the empirical Type I error of the proportion adaptive segment selection with $L = 6$ and $\alpha_0 = 10$ and the single-sample method. The empirical Type I error is defined as the percentage of replications in which some interval is selected under H_0. We observe empirical Type I errors of 0·081 and 0·097, respectively, both with standard error 0·009. To assess the power of detecting the segments, one segment I is randomly selected and each X_{it} in that segment is generated from models (1)–(2) with $|I| = |I_1| = 5$, $\tau = \tau_1 = 1$, $\pi = \pi_1 = 0{\cdot}1$, and $\mu = \mu_1 =$ 0·5, 0·7, 0·9 and 1·1. The empirical power based on 100 replications is defined as the percentage of replications in which some interval in $\hat{\mathbb{I}}$ overlaps with the segment I. Table 2 shows that proportion adaptive selection outperforms the single-sample method, resulting in much higher power for a wide range of μ values.

The estimated standard errors of the empirical power over 100 replications are also included in Table 2. To estimate the standard error of the medians, we generate 500 bootstrap samples from the 100 replication results, then calculate a median for each bootstrap sample. The standard error is the standard deviation of the 500 bootstrap medians. The standard errors are in general small for all the simulations in § § 4·2–4·4.

4·3. *Effects of segment length and signal variance*

Further simulations are provided to demonstrate the effect of segment length and signal variance on the proportion adaptive segment selection. We randomly select three locations for segment variants with different segment length, $|I_1| = 4$, $|I_2| = 9$, and $|I_3| = 16$. The other parameters are set as $\mu_k = 1$, $\pi_k = 0{\cdot}05$, and $\tau_k = 2{\cdot}5$, 1·5 and 0·0 for $k = 1, 2, 3$. Table 3 shows the estimation accuracy and the control of overselection for the proportion adaptive selection with $L = 20$ and $\alpha_0 = 10$. The estimation accuracy for signal segment I_k is demonstrated by the dissimilarity measure

$$D_k = \min_{\hat{I}_j \in \hat{\mathbb{I}}} \{1 - |\hat{I}_j \cap I_k| / (|\hat{I}_j||I_k|)^{1/2}\},$$

Table 3. *Medians and standard errors (in parentheses) of the dissimilarity measure D_k ($k = 1, 2, 3$) and the number of overselections for the proportion adaptive segment selection over* 100 *replications. The lengths of the intervals are $|I_1| = 4$, $|I_2| = 9$ and $|I_3| = 16$*

τ	D_1	D_2	D_3	$\#O$
2·5	1 (0·3)	0·18 (0·02)	0·10 (0·01)	2 (0·2)
1·5	1 (0·1)	0·10 (0·02)	0·06 (0·01)	2 (0·0)
0·0	1 (0·0)	0·05 (0·02)	0·03 (0·01)	2 (0·3)

$\#O$, number of overselected intervals.

where $\hat{\mathbb{I}}$ denotes the collection of intervals selected by proportion adaptive segment selection. Apparently, $D_k \in [0, 1]$ and smaller values of D_k correspond to a greater overlap of I_k with some intervals in $\hat{\mathbb{I}}$. Table 3 shows that longer segment length and/or smaller signal variance result in better identification of the segments by proportion adaptive segment selection.

4·4. *Simultaneous discovery of rare and common segment variants*

We demonstrate the adaptivity property of the proportion adaptive segment identification by comparing its performance with two recently published methods. One method developed by Zhang et al. (2010), which pools information through the sum of observations over all the samples, is expected to work well for common signals with weak signal intensity. Another method developed by Siegmund et al. (2010) with the prior probability of carrier fixed at $p_0 = 0{\cdot}01$ pools information through the outliers of observations over all the samples and is more efficient for detecting rare signals with strong signal intensity. Since these methods control the genome-wide Type I error at a given level and our method does not provide a fixed-level Type I error control, their parameters are tuned to obtain a comparable overselection of intervals. Specifically, we choose $\alpha_0 = 10$ and a data-driven threshold at 5·52 to control the number of overselected intervals fewer than two.

The simulations are repeated 100 times. The parameters used are $N = 400$, $T = 5000$, $q = 5$, $L = 6$ and $|I_k| = 5$, $\tau_k = 0$ for $k = 1, \ldots, q$. We consider two different scenarios. The first scenario considers rare segment variants with a large signal intensity where μ is fixed at 1 and the π_k vary from 0·04 to 0·08. The second scenario considers common segment variants with a small signal intensity where μ is fixed at 0·3 and the π_k vary from 0·4 to 0·8. We compare the performances of these methods in terms of overselection and empirical power in Table 4. The numbers of overselections are comparable for all these methods. The method of Zhang et al. (2010) performs best for identifying the common segment variants, while that of Siegmund et al. (2010) with $p_0 = 0{\cdot}01$ performs best for rare ones. The performance of the proportion adaptive segment selection lies between these two methods, demonstrating its adaptability and good power for identifying both rare and common segment variants simultaneously. We finally compare the results from the combined method of applying Zhang et al. (2010) and Siegmund et al. (2010) with $p_0 = 0{\cdot}01$ together with the median number of overselections of two. The proportion adaptive procedure results in slightly lower power than the combined method.

5. Application to neuroblastoma samples

We apply the proportion adaptive segment selection to a sample of 674 neuroblastoma cases that were collected as part of a large-scale genome-wide association study of neuroblastoma (Diskin et al., 2009). For each sample, over 600 000 single nucleotide polymorphism markers

Table 4. *Empirical power* (%) *and median of the number of overselection* $\#O$ *for proportion adaptive segment selection and methods of* Zhang et al. (2010) *and* Siegmund et al. (2010) *over* 100 *replications. The standard errors appear in parentheses*

Rare and strong signal

$\mu = 1$	$\pi_1 = 0{\cdot}04$	$\pi_2 = 0{\cdot}05$	$\pi_3 = 0{\cdot}06$	$\pi_4 = 0{\cdot}07$	$\pi_5 = 0{\cdot}08$	$\#O$
PASS	35 (4·8)	46 (4·8)	66 (4·7)	86 (3·6)	91 (2·8)	2·5 (0·4)
MSCP1	20 (3·9)	34 (4·7)	60 (5·1)	73 (4·5)	86 (3·3)	1 (0·0)
MSCP001	46 (4·9)	56 (4·9)	77 (4·5)	85 (3·5)	95 (2·2)	2 (0·4)
COMB	45 (5·1)	57 (4·8)	70 (4·3)	82 (3·8)	93 (2·5)	2 (0·1)

Common and weak signal

$\mu = 0{\cdot}3$	$\pi_1 = 0{\cdot}4$	$\pi_2 = 0{\cdot}5$	$\pi_3 = 0{\cdot}6$	$\pi_4 = 0{\cdot}7$	$\pi_5 = 0{\cdot}8$	$\#O$
PASS	8 (2·6)	19 (3·8)	21 (4·1)	41 (5·0)	54 (5·2)	2 (0·4)
MSCP1	11 (3·3)	25 (4·4)	36 (4·7)	58 (4·9)	69 (4·6)	1 (0·1)
MSCP001	11 (3·3)	9 (2·9)	12 (3·1)	24 (4·1)	24 (4·2)	2 (0·2)
COMB	12 (3·3)	21 (4·0)	26 (4·4)	45 (5·0)	57 (5·2)	2 (0·5)

PASS, proportion adaptive segment selection with $\alpha_0 = 10$; MSCP1, methods of Zhang et al. (2010); MSCP001, method of Siegmund et al. (2010) with p_0 =0·01; COMB, combined MSCP1 and MSCP001.

were genotyped using the Illumina genotype platform and the log R-ratio data were obtained. In order to account for possible wave-effect or local effects, we performed similar processing as in Siegmund et al. (2010) to obtain the normalized data by subtracting the sample median and regressing on the first principal component. In our analysis, we considered only data from chromosome 1, which includes $T = 40\,929$ log R-ratios.

In our analysis, we choose $L = 20$ and $\alpha_0 = 4$ to allow the selection of the copy number variant with four or more carriers. We then use simulations to determine the threshold for V_N to control the number of overselections to zero. Specifically, 674 samples and 40 929 observations are simulated 50 times from a standard normal distribution. The mean and standard deviation of the simulated threshold are 12·57 and 2·98. With the threshold set at 12·57, our proportion adaptive procedure resulted in selection of 335 copy number variants with three or more markers, including 171 copy number variants with three markers, and 100 copy number variants with 4 markers, and 11 copy number variants with 10 or more markers. The median size of the copy number variants identified is 4165 bps with a range of 462 to 1 038 000 bps. Figure 2 shows likelihood ratio statistics and the data plots for six copy number variants that we identified, demonstrating different characteristics. The first two plots show two common copy number variants detected in these 674 neuroblastoma samples, where the first copy number variant with 8 markers overlaps with the 7-marker copy number variant that was showed to be associated with the risk of neuroblastoma in Diskin et al. (2009). The second 3-marker copy number variant is very common and is also validated by Redon et al. (2006). The third and fourth plots show two rare copy number variants that were detected by the proportion adaptive segment selection, where only a few samples show large likelihood ratio statistics. These two copy number variants were also validated by Redon et al. (2006). These results indicate that the proportion adaptive segment selection can indeed detect both rare and common copy number variants.

Since the identification of the short copy number variants is more susceptible to local wave effects or other artifacts of the data, we should interpret the copy number variants of three or four markers with caution and focus the following comparison on the 64 identified copy number variants of five or more markers. Among these 64 copy number variants, 30 overlap with the copy number variants in the database of genomic variants (Zhang et al., 2006). This database includes only the relatively common copy number variants identified in healthy human cases. To further

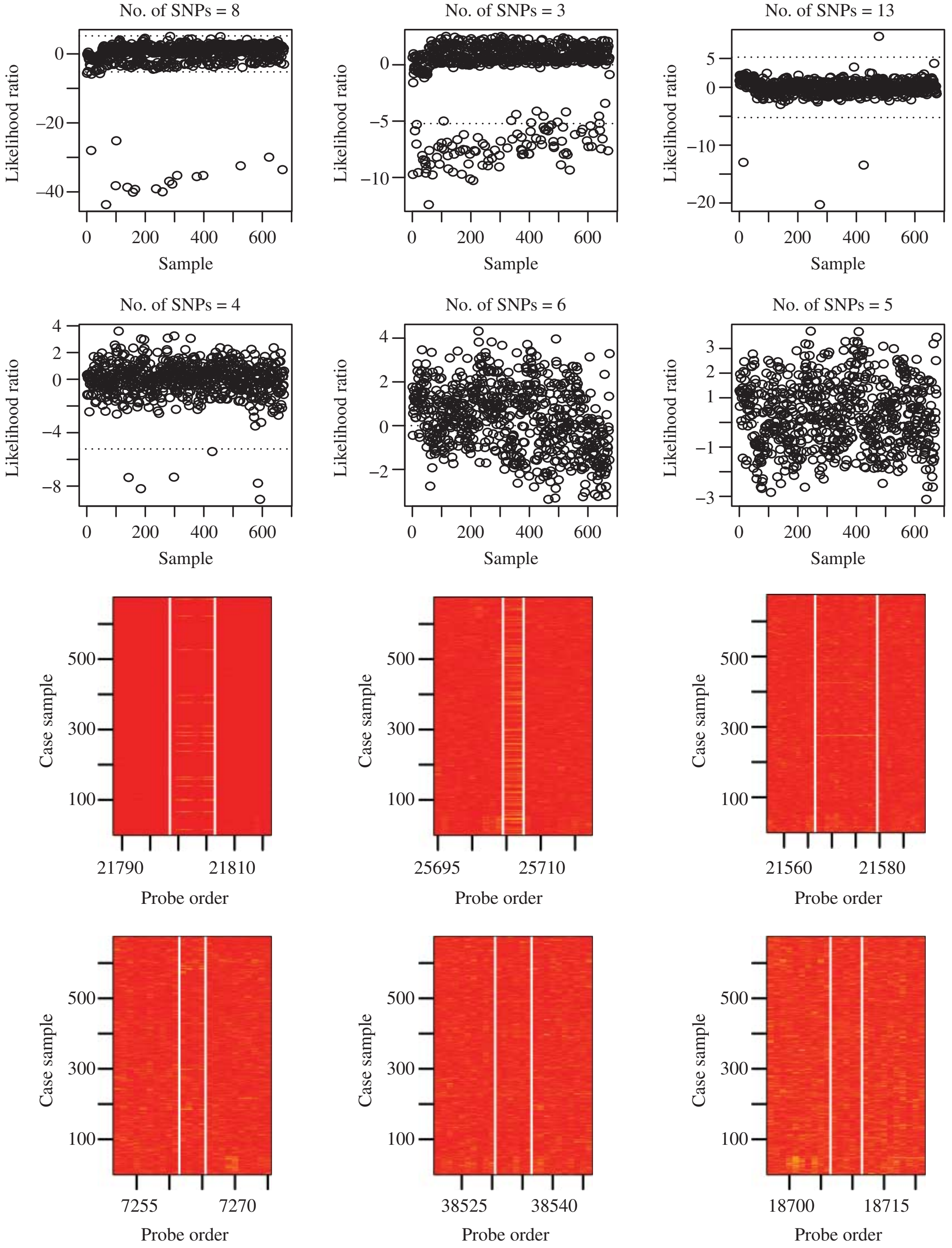

Fig. 2. Examples of the copy number variants identified. Top two panels are the likelihood ratio statistic for each sample, the bottom two panels show the heatmap of absolute value of the observed log R-ratio values for the markers within and around the copy number variants identified (vertical white lines).

demonstrate the power of the proportion adaptive selection, we also performed single-sample copy number identification using the optimal likelihood ratio selection procedure of Jeng et al. (2010). Among the 64 copy number variants, 20 of them did not reach the theoretical threshold of $\{2\log(TL)\}^{1/2} = 5{\cdot}22$ in any of the 674 samples, indicating loss of power of detecting the copy number variants based on the single-sample analysis. Of these 20 copy number variants missed by the single-sample analysis, ten overlap with the copy number variants in the genomic variants database (Zhang et al., 2006). These copy number variants were reported in Redon et al. (2006). As an example, the fourth and fifth panels of Figure 2 show the likelihood ratio statistics and the observed log R-ratios for two copy number variants identified by the proportion adaptive selection, but no samples pass the theoretical threshold value for single-sample analysis.

Acknowledgement

This research is supported by the National Institutes of Health and the National Science Foundation. We thank the reviewer and associate editor for many insightful comments.

Appendix

Proof of Theorem 1

Based on results for the extreme values of a standardized uniform empirical process, we have for any t free of N,

$$\text{pr}\{a_N V_N(J) - b_N \leqslant t\} \rightarrow \exp(-e^{-t}),$$

as $N \rightarrow \infty$, which implies

$$\text{pr}\{a_N V_N(J) - b_N \leqslant t\} \geqslant \exp(-C_1 e^{-t})$$

for any fixed t and sufficiently large N, where C_1 is a constant. This combined with the choice of $\lambda_{T,N}$ implies that

$$\text{pr}\{a_N V_N(J) - b_N \leqslant a_N \lambda_{T,N} - b_N\} = \text{pr}\{a_N V_N(J) - b_N \leqslant C_0 \log(TL)\} \geqslant \exp\{-C_1 (TL)^{-C_0}\}.$$

Therefore,

$$\begin{aligned}
\text{pr}(\mathbb{I}^{(1)} \mp \emptyset) &= \text{pr}\{\text{there exists a } J \in \mathbb{J}_{T,N}(L) \text{ such that } V_N(J) > \lambda_{T,N}\} \\
&\leqslant TL[1 - \text{pr}\{a_N V_N(J) - b_N \leqslant a_N \lambda_{T,N} - b_N\}] \\
&\leqslant TL[1 - \exp\{-C_1 (TL)^{-C_0}\}] \\
&\leqslant TLC_1 (TL)^{-C_0} \\
&= C_1 (TL)^{-(C_0 - 1)},
\end{aligned}$$

where the third inequality uses the fact that $e^{-x} \geqslant 1 - x$. The result follows by the condition $C_0 > 1$.

Proof of Theorem 2

By the calibration of π_k and the condition on α_0, the carrier proportion remains $N^{-\beta_k}$ when the $(\alpha_0 - 1)$ smallest p-values are excluded in (10). By the construction of $\mathbb{I}^{(1)}$, it is enough to show that for any $I_k \in \mathbb{I}$

with $r_k > \rho^+(\beta_k, \tau_k)$ for $1/2 < \beta_k < 1$ or $r_k > \rho^-(\beta_k, \tau_k)$ for $0 < \beta_k < 1/2$,

$$\mathrm{pr}\{V_N(I_k) \leqslant \lambda_{T,N}\} \leqslant CN^{-C(r_k,\beta_k,\tau_k)} \rightarrow 0. \tag{A1}$$

Defining the standardized empirical process as

$$W_{N,I_k}(t) = N^{1/2}\frac{\bar{F}_{N,I_k}(t) - \bar{\Phi}(t)}{[\bar{\Phi}(t)\{1-\bar{\Phi}(t)\}]^{1/2}},$$

where $\bar{\Phi}(t)$ is the survival function of a standard normal random variable and

$$\bar{F}_{N,I_k}(t) = \frac{1}{N}\sum_{i=1}^{N} 1(X_{I_k,i} > t),$$

we can rewrite $V_N(I_k)$ defined in (10) as $V_N(I_k) = \sup_{-\infty<t<\infty} W_{N,I_k}(t)$. For any fixed t,

$$E\{W_{N,I_k}(t)\} = \frac{N^{1/2}\pi_k\left[\bar{\Phi}\left\{\frac{t-\mu_k|I_k|^{1/2}}{(1+\tau_k^2)^{1/2}}\right\} - \bar{\Phi}(t)\right]}{[\bar{\Phi}(t)\{1-\bar{\Phi}(t)\}]^{1/2}}, \quad \mathrm{var}\{W_{N,I_k}(t)\} = \frac{\bar{F}(t)\{1-\bar{F}(t)\}}{\bar{\Phi}(t)\{1-\bar{\Phi}(t)\}}. \tag{A2}$$

The key step of the proof is to find a t value such that $W_{N,I_k}(t) > \lambda_{T,N}$ with large probability. Define

$$t_k^* = \begin{cases} \min\{2(2r_k \log N)^{1/2}/(1-\tau_k^2),\, (2\log N)^{1/2}\}, & \tau_k < 1,\, 1/2 < \beta_k < 1, \\ (2\log N)^{1/2}, & \tau_k \geqslant 1,\, 1/2 < \beta_k < 1, \\ 1, & 0 \leqslant \beta_k < 1/2, \end{cases} \tag{A3}$$

When $\tau_k < 1$, $1/2 \leqslant \beta_k < 1$ and $r_k > \rho^+(\beta_k, \tau_k)$, we have $m(\tau_k) = (1+\tau_k^2)/4$ and $\beta_k \geqslant 1 - m(\tau_k)$ if and only if

$$2(2r_k \log N)^{1/2}/(1-\tau_k^2) \geqslant (2\log N)^{1/2}.$$

Then,

$$t_k^* = \begin{cases} (2\log N)^{1/2}, & \tau_k < 1,\, 1 - m(\tau_k) \leqslant \beta_k < 1, \\ 2(2r_k \log N)^{1/2}/(1-\tau_k^2), & \tau_k < 1,\, 1/2 < \beta_k < 1 - m(\tau_k). \end{cases} \tag{A4}$$

Applying calibrations of $\mu_k|I_k|^{1/2}$ for $1/2 \leqslant \beta_k < 1$ and $0 \leqslant \beta_k < 1/2$ respectively, we have,

$$E\{W_{N,I_k}(t_k^*)\} \sim \begin{cases} CN^{1/2-\beta_k}\bar{\Phi}\left\{\frac{t_k^*-(2r_k\log N)^{1/2}}{(1+\tau_k^2)^{1/2}}\right\}/\{\bar{\Phi}(t_k^*)\}^{1/2}, & 1/2 < \beta_k < 1, \\ CN^{1/2-\beta_k+r_k}, & 0 \leqslant \beta_k < 1/2,\, \tau_k = 0, \\ CN^{1/2-\beta_k}, & 0 \leqslant \beta_k < 1/2,\, \tau_k > 0. \end{cases}$$

Combining the above with (A3), (A4) and the fact that

$$\bar{\Phi}\left\{\frac{(2q\log N)^{1/2} - (2r_k\log N)^{1/2}}{(1+\tau_k^2)^{1/2}}\right\} \sim \frac{C}{(\log N)^{1/2}} N^{-(q^{1/2}-r_k^{1/2})^2/(1+\tau_k^2)},$$

we have

$$E\{W_{N,I_k}(t_k^*)\} \sim \begin{cases} \frac{C}{(\log N)^{1/2}} N^{1-\beta_k-(1-r_k^{1/2})^2/(1+\tau_k^2)}, & \tau_k \geqslant 1,\, 1/2 < \beta < 1, \\ \frac{C}{(\log N)^{1/2}} N^{1-\beta_k-(1-r_k^{1/2})^2/(1+\tau_k^2)}, & \tau_k < 1,\, 1 - m(\tau_k) \leqslant \beta_k < 1, \\ \frac{C}{(\log N)^{1/2}} N^{r_k/(1-\tau_k^2)-(\beta_k-1/2)}, & \tau_k < 1,\, 1/2 < \beta_k < 1 - m(\tau_k), \\ CN^{1/2-\beta_k-r_k}, & 0 \leqslant \beta_k < 1/2,\, \tau_k = 0, \\ CN^{1/2-\beta_k}, & 0 \leqslant \beta_k < 1/2,\, \tau_k > 0. \end{cases}$$

Then it can be shown that $E\{W_{N,I_k}(t_k^*)\} \geqslant CN^C$ for some $C > 0$ given $r_k > \rho^+(\beta_k, \tau_k)$ for $1/2 \leqslant \beta_k < 1$ or $r_k < \rho^-(\beta_k, \tau_k)$ for $0 \leqslant \beta_k < 1/2$. Further, by condition $N^C \gg \log T$ for any $C > 0$, we have

$$E\{W_{N,I_k}(t_k^*)\} \geqslant CN^C \gg \lambda_{T,N}. \tag{A5}$$

By Chebyshev's inequality, (A2), and (A5),

$$\mathrm{pr}\{W_{N,I_k}(t_k^*) \leqslant \lambda_{T,N}\} \leqslant C\frac{\mathrm{var}\{W_{N,I_k}(t_k^*)\}}{[E\{W_{N,I_k}(t_k^*)\}]^2} \leqslant \frac{C\bar{F}(t_k^*)}{N\pi_k^2\left[\bar{\Phi}\left\{\frac{t_k^*-\mu_k|I_k|^{1/2}}{(1+\tau_k^2)^{1/2}}\right\} - \bar{\Phi}(t_k^*)\right]^2}.$$

Applying (A3) and condition $r_k > \rho^+(\beta_k, \tau_k)$ for $1/2 \leqslant \beta_k < 1$ or $r_k < \rho^-(\beta_k, \tau_k)$ for $0 \leqslant \beta_k < 1/2$ to the above, we have

$$\mathrm{pr}\{W_{N,I_k}(t_k^*) \leqslant \lambda_{T,N}\} \leqslant CN^{-C(r_k,\beta_k,\tau_k)} \to 0, \tag{A6}$$

where $C(r_k, \beta_k, \tau_k)$ is as in (16). By combining (A6) and the fact that

$$\mathrm{pr}\{V_N(I_k) \leqslant \lambda_{T,N}\} \leqslant P_{H_1}\{W_{N,I_k}(t_k^*) \leqslant \lambda_{T,N}\},$$

(A1) is verified.

References

Anderson, T. W. & Darling, D. A. (1952). Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Ann. Math. Statist.* **23**, 193–212.

Cai, T. T., Jeng, X. J. & Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Statist. Soc.* B **73**, 629–62.

Diskin, S. J., Hou, C., Glessner, J. T., Attiyeh, E. F., Laudenslager, M., Bosse, K., Cole, K., Moss, Y., Wood, A. Lynch, J. E. et al. (2009). Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987–91.

Donoho, D. & Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962–94.

Jeng, X. J., Cai, T. T. & Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *J. Am. Statist. Assoc.* **105**, 1156–66.

Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–72.

Redon, R., Ishikawa, S., Fitch, K., Feuk, L., Perry, G., Andrews, T., Fiegler, H., Shapero, M., Carson, A. Chen, W. et al. (2006). Global variation in copy number in the human genome. *Nature* **444**, 444–54.

Shorack, G. R. & Wellner, J. A. (2009). *Empirical Processes with Applications to Statistics*. Philadelphia: SIAM.

Siegmund, D. O., Yakir, B. & Zhang, N. R. (2010). Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Statist.* **5**, 645–68.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J. T., Grant, S. F. A., Hakonarson, H. & Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–74.

Zhang, F., Gu, W., Hurles, M. & Lupski, J. (2009). Copy number variation in human health, disease and evolutions. *Annu. Rev. Genomics Human Genet.* **10**, 451–81.

Zhang, J., Feuk, L., Duggan, G. E., Khaja, R. & Scherer, S. W. (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.* **115**, 205–14.

Zhang, N. R., Siegmund, D. O., Ji, H. & Li, J. (2010). Detecting simultaneous change-points in multiple sequences. *Biometrika* **97**, 631–45.

[*Received September* 2011. *Revised August* 2012]