Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2012)
74, *Part* 5, *pp.* 773–797

Robust detection and identification of sparse segments in ultrahigh dimensional data analysis

T. Tony Cai, X. Jessie Jeng and Hongzhe Li

University of Pennsylvania, Philadelphia, USA

[Received October 2010. Final revision November 2011]

Summary. Copy number variants (CNVs) are alternations of DNA of a genome that result in the cell having less or more than two copies of segments of the DNA. CNVs correspond to relatively large regions of the genome, ranging from about one kilobase to several megabases, that are deleted or duplicated. Motivated by CNV analysis based on next generation sequencing data, we consider the problem of detecting and identifying sparse short segments hidden in a long linear sequence of data with an unspecified noise distribution. We propose a computationally efficient method that provides a robust and near optimal solution for segment identification over a wide range of noise distributions. We theoretically quantify the conditions for detecting the segment signals and show that the method near optimally estimates the signal segments whenever it is possible to detect their existence. Simulation studies are carried out to demonstrate the efficiency of the method under various noise distributions. We present results from a CNV analysis of a HapMap Yoruban sample to illustrate the theory and the methods further.

Keywords: DNA copy number variant; Next generation sequencing data; Optimality; Robust segment detector; Robust segment identifier

1. Introduction

Structural variants in the human genome (Sebat *et al.*, 2004; Feuk *et al.*, 2006), including copy number variants (CNVs) and balanced rearrangements such as inversions and translocations, play an important role in the genetics of complex disease. CNVs are alternations of deoxyribonucleic acid (DNA) of a genome that result in the cell having less or more than two copies of segments of the DNA. CNVs correspond to relatively large regions of the genome, ranging from about one kilobase to several megabases, that are deleted or duplicated. Analysis of CNVs in developmental and neuropsychiatric disorders (Feuk *et al.*, 2006; Walsh *et al.*, 2008; Stefansson *et al.*, 2008; Stone *et al.*, 2008) and in cancer (Diskin *et al.*, 2009) has led to the identification of novel disease-causing mutations, thus contributing important new insights into the genetics of these complex diseases. Changes in DNA copy number have also been highly implicated in tumour genomes; most are due to somatic mutations that occur during the clonal development of the tumour. The copy number changes in tumour genomes are often referred to as copy number aberrations. In this paper, we focus on the CNVs from the germ line constitutional genome where most of the CNVs are sparse and short (Zhang *et al.*, 2009).

CNVs can be discovered by cytogenetic techniques, array comparative genomic hybridization (Urban *et al.*, 2006) and by single-nucleotide polymorphism arrays (Redon *et al.*, 2006). The emerging technologies of DNA sequencing have further enabled the identification of CNVs by next generation sequencing (NGS) in high resolution. NGS can generate millions of short

Address for correspondence: X. Jessie Jeng, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 207 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA.
E-mail: xjeng@upenn.edu

 1369–7412/12/74773

sequence reads along the whole human genome. When these short reads are mapped to the reference genome, both distances of paired end data and read depth (RD) data can reveal the possible structure variations of the target genome (for reviews, see Medvedev *et al.* (2009) and Alkan *et al.* (2011)). The mapping distances between pair ends of reads provide better power to detect small to medium-size insertions and deletions or CNVs (Medvedev *et al.*, 2009; Alkan *et al.*, 2011). In this approach, two paired reads are generated at an approximately known distance in the donor genome and pairs mapping at a distance that is substantially different from the expected length, or with anomalous orientation, suggest structural variants. Methods based on the mapping distances often involve finding the clusters of reads that show anomalous mapping (Chen *et al.*, 2009). Instead of mapping the short reads onto the reference genome, one can also perform whole genome *de novo* assembly and align the *de novo* assemblies of two genomes to identify gaps and segmental rearrangements in pairwise alignments (Li *et al.*, 2011). However, this approach requires data from two genomes.

Another important source of information that is useful for inferring CNVs from reads alignment is the RD. The RD data are generated to count the number of reads that cover a genomic location or a small bin along the genome which provide important information about the CNVs that a given individual carries (Shendure and Ji, 2008; Medvedev *et al.*, 2009). When the genomic location or bin is within a deletion, we expect to observe a smaller number of read counts or lower mapping density than the background RD. In contrast, when the genomic location or bin is within an insertion or duplication, we expect to observe a larger number of read counts or higher mapping density. Therefore, these RDs can be used to detect and identify the CNVs. The RD data provide more reliable information for large CNVs and CNVs flanked by repeats, where accurate mapping reads is difficult. The RD data also provide information on CNVs based on the targeted sequences where only targeted regions of the genome are sequenced (Nord *et al.*, 2011).

In this paper, we consider the problem of CNV detection and identification based on the RD data from NGS. Several methods have been developed for such RD data. Yoon *et al.* (2009) developed an algorithm for RD data to detect CNVs, where they converted the read count of a window into a Z-score by subtracting the mean of all windows and dividing by the standard deviation and identified the CNVs by computing upper tail and lower tail probabilities by using a normality assumption on the RD data. The windows are then selected on the basis of the extreme values of these probabilities controlling for the genomewide false positive rates. Abyzov *et al.* (2011) developed an approach first to partition the genome into a set of regions with different underlying copy numbers by using the mean shift technique and then to merge signal and call CNVs by performing t-tests. Xie and Tammi (2009), Chiang *et al.* (2009) and Kim *et al.* (2010) developed methods for CNV detection based on RD data when pairs of samples are available. The basic idea underlying these two methods is to convert the counts data into ratios and then to apply existing copy number analysis methods developed for array comparative genomic hybridization data such as the circular binary segmentation (CBS) (Olshen *et al.*, 2004) for CNV detection. Methods have also been developed for CNV detections in cancer cells (Ivakhno *et al.*, 2010; Miller *et al.*, 2011) based on the RD data.

One common feature of these existing methods is to make a certain parametric distribution assumption on the RD data. However, the distribution of the RD data is in general unknown owing to the complex process of sequencing. Some recent literature assumes a constant read sampling rate across the genome and a Poisson distribution or negative binomial distribution for the read counts data (Xie and Tammi, 2009; Cheung *et al.*, 2011). However, owing to guanine–cytosine content, mappability of sequencing reads and regional biases, genomic sequences obtained through high throughput sequencing are not uniformly distributed across the genome

and therefore the counts data are likely not to follow a Poisson distribution (Li *et al.*, 2010; Miller *et al.*, 2011; Cheung *et al.*, 2011). The feature of the NGS data also changes with advances in sequencing technologies. To analyse such data, classical parametric methods do not work well. It is crucial for these methods to specify the distribution of their test statistics, which depends on the data distribution. A misspecified data distribution can lead to a complete failure of these methods. Although some data distributions can be estimated by non-parametric methods, popular non-parametric methods such as permutation are often computationally expensive and not feasible for ultrahigh dimensional data. Therefore, robust methods that are adaptive to unknown data distributions and computationally efficient at the same time are greatly needed. The goal of the present paper is to develop such a robust procedure for CNV identification based on NGS data and to study its properties.

In this paper, we assume that a long linear sequence of noisy data $\{Y_1, \ldots, Y_n\}$ is modelled as

$$Y_i = \alpha_i + \xi_i, \qquad \alpha_i = \begin{cases} \mu_j, & i \in I_j \text{ for some } j \in \{1, \ldots, q\}, \\ 0, & \text{otherwise}, \end{cases} \quad 1 \leqslant i \leqslant n, \tag{1}$$

where $q = q_n$ is the number of signal segments, possibly increasing with n, $I_1, \ldots, I_q$ are disjoint intervals representing signal segments, $\mu_1, \ldots, \mu_q$ are unknown constants and $\xi_1, \ldots, \xi_n$ are independent, identically distributed (IID) random errors with median 0. Here positive μ_i implies duplication or insertion and a negative mean implies deletion. Let $\mathbb{I} = \{I_1, \ldots, I_q\}$ denote the set of all segment intervals. For the problem of CNV detection based on the NGS data, Y_i is the guanine–cytosine content-adjusted RD counts at genomic location or bin i, which can be regarded as continuous when coverage of the genome is sufficiently high, e.g. greater than 20 (Yoon *et al.*, 2009; Abyzov *et al.*, 2011). This model describes the phenomenon that some signal segments are hidden in the n noisy observations. The number, locations, mean values of the segments and the distribution of the random errors are unknown.

The problem of segment detection and identification can be separated into two steps. The first step is to detect the existence of such signal segments, i.e. to test

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

and the second step is to identify the locations of the segments if there are any. A procedure called likelihood ratio selection (LRS) (Jeng *et al.*, 2010) has recently been developed to treat the above problem in the case of Gaussian noise. Under the Gaussian and certain sparsity assumptions, LRS has been shown to be optimal in the sense that it can separate the signal segments from noise whenever the segments are identifiable. However, when the noise distribution is heavy tailed, LRS may fail and provide a large number of misidentifications. To tackle this difficulty, in the present paper we introduce a computationally efficient method called the robust segment identifier (RSI), which provides a robust and near optimal solution for segment identification over a wide range of noise distributions. As an illustration, we generate 1000 observations based on a Cauchy $(0, 1)$ distribution and set the signal segment at $[457 : 556]$ with a positive mean. Fig. 1 compares LRS with the RSI. In this example, LRS fails to work at all by identifying too many false segments, whereas the RSI, in contrast, provides a good estimate of the signal segment even when the noise distribution is unknown and heavy tailed.

A key step of the RSI is a local median transformation, where the original observations are first divided into T small bins with m observations in each bin and then the median values of the data in these bins are taken as a new data set. The central idea is that the new data set can be well approximated by Gaussian random variables for a wide collection of error distributions. After the local median transformation, existing detection and identification methods that are designed for Gaussian noise can then be applied to the new data set. Brown *et al.* (2008) and

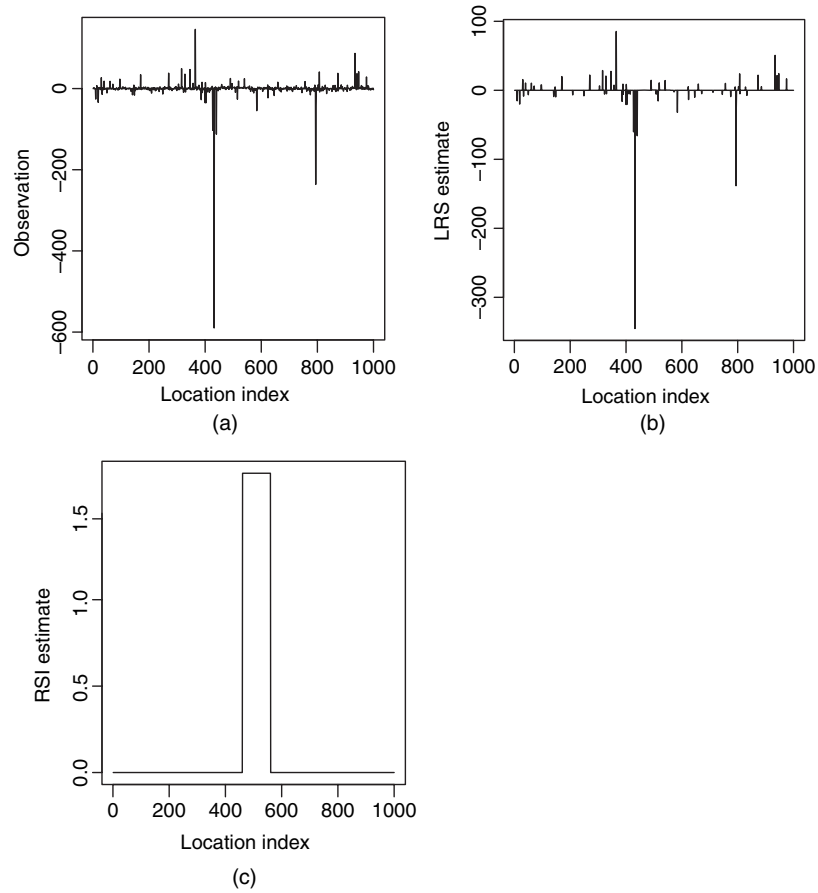

Fig. 1. Effects of long-tailed error distribution on segment identification: (a) data with Cauchy noise and a signal segment at [457 : 556]; (b) intervals identified and estimated interval means by LRS; (c) interval identified and estimated means by the RSI

Cai and Zhou (2009) used local medians to turn the problem of non-parametric regression with unknown noise distribution into a standard Gaussian regression. Here we use the local median transformation for signal detection and identification that is robust over a large collection of error distributions, including those that are heavy tailed. Local median transformations or other local smoothing procedures have been applied in analysis of microarray data for data normalization (Quackenbush, 2002).

To elucidate the effect of data transformation in the simplest and cleanest way, we begin by considering the detection part of our problem, which is to test H_0 against H_1. We propose a robust segment detector (RSD), which applies the generalized likelihood ratio test (GLRT) to the transformed data. Arias-Castro *et al.* (2005) showed that the GLRT is an optimal procedure for detecting a single segment with constant length in the Gaussian noise setting. We find here that the RSD is an near optimal procedure for the transformed data when the bin size m is properly chosen. The key condition for the RSD to be successful is that some segment I_j has its mean and length roughly satisfying $\mu_j\sqrt{|I_j|} > \sqrt{\log(n)}/\{h(0)\sqrt{2}\}$ and the noise density satisfies the Lipschitz condition, where $h(0)$ is the noise density at 0. Clearly, a larger value of $h(0)$, which corresponds to a more concentrated noise distribution, results in a more relaxed condition. This

result agrees with our intuition that detection of signal segments should be easier if the noise distribution is more concentrated.

The RSD can detect the existence of signal segments with unknown noise distribution. However, it does not tell where the signal segments are. For segment identification, we propose a procedure called the RSI, which first transforms the data by binning and taking local medians, and then applies the LRS procedure on the transformed data. Unlike the RSD, which searches through all possible intervals after the data transformation, the RSI utilizes the short segment structure and considers only short intervals with length less than or equal to L, where L is some number that is much smaller than T. Furthermore, the data transformation step significantly reduces the dimension from n to T. These together make the RSI a computationally efficient method to handle ultrahigh dimensional data. It is shown that the RSI provides robust identification results for a large collection of noise distributions, and it is a near optimal procedure for the transformed data when m and L are properly chosen.

The rest of the paper is organized as follows. Section 2 introduces the data transformation technique and the RSD. The robust segment identification procedure RSI is proposed and its theoretical properties are studied in Section 3. The numerical performance of the RSI is investigated in Section 4 by using simulations and is compared with the performances of LRS and CBS. We then present results in Section 5 from an analysis of sequence data of one individual from the 1000 genomes project (`http://www.1000genomes.org/`). We conclude with a discussion in Section 6. The proofs are detailed in Appendix A.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Data transformation and robust detection

In this section, we first introduce the local median transformation to tackle the problem of an unknown and possibly heavy-tailed noise distribution. After the transformation the data can be well approximated by Gaussian random variables. A robust segment detection procedure is then developed to separate H_0 from H_1 reliably over a wide range of noise distributions.

2.1. Local median transformation

Let $Y_1, \ldots, Y_n$ be a sequence of observed data generated from model (1) with an unknown noise distribution. We assume that there are q sparse and short signal segments in the observed data and that the number of observations n is very large. The goal is to detect and identify these q segments. To do so, we first equally divide the n observations into $T = T_n$ groups with $m = m_n$ observations in each group. Define the set of indices in the kth group as $J_k = \{i : (k-1)m + 1 \leqslant i \leqslant km\}$ and generate the transformed data set as

$$X_k = \text{median}\{Y_i : i \in J_k\}, \qquad 1 \leqslant k \leqslant T. \tag{2}$$

Set

$$\eta_k = \text{median}\{\xi_i : i \in J_k\}, \qquad 1 \leqslant k \leqslant T; \tag{3}$$

then the medians X_k can be written as

$$X_k = \theta_k + \eta_k, \qquad 1 \leqslant k \leqslant T, \tag{4}$$

where

$$\theta_k = \begin{cases} \mu_j, & J_k \subseteq I_j \text{ for some } I_j, \\ \mu_k^* \in [0, \mu_j], & J_k \cap I_j \neq \emptyset \text{ for some } I_j \text{ and } J_k \not\subseteq I_j, \\ 0, & \text{otherwise.} \end{cases}$$

After the local median transformation, the errors ξ_i in the original observations are re-represented by η_k. The main idea is that η_k can be well approximated by Gaussian random variables for a wide range of noise distributions. Specifically, we assume that the distribution of ξ_i is symmetric about 0 with density function h satisfying $h(0) > 0$ and

$$|h(y) - h(0)| \leqslant Cy^2 \tag{5}$$

in an open neighbourhood of 0. This assumption is satisfied, for example, by the Cauchy distribution, the Laplace distribution and the t-distributions, as well as the Gaussian distribution. A similar assumption was introduced in Cai and Zhou (2009) in the context of non-parametric function estimation. The distributions of η_k are approximately normal. This can be precisely stated in the following lemma.

Lemma 1. Assume expressions (1) and (5), and transformation (4); then η_k can be written as

$$\eta_k = \frac{1}{2h(0)\sqrt{m}} Z_k + \frac{1}{\sqrt{m}} \zeta_k, \tag{6}$$

where $Z_k \sim^{\mathrm{IID}} N(0, 1)$ and ζ_k are independent and stochastically small random variables satisfying $E(\zeta_k) = 0$, and can be written as

$$\zeta_k = \zeta_{k1} + \zeta_{k2}$$

with

$$E(\zeta_{k1}) = 0 \quad \text{and} \quad E(|\zeta_{k1}|^l) \leqslant C_l m^{-l}, \tag{7}$$

$$P(\zeta_{k2} = 0) \geqslant 1 - C \exp(-am) \tag{8}$$

for some $a > 0$ and $C > 0$, and all $l > 0$.

The proof of this lemma is similar to that of proposition 1 in Brown *et al.* (2008) and that of proposition 2 in Cai and Zhou (2009), and is thus omitted. The key fact is that η_k can be well approximated by $Z_k/\{2h(0)\sqrt{m}\}$, which follows $N[0, 1/\{4h^2(0)m\}]$, so, after the data transformation in expression (4), existing methods for Gaussian noise can be applied to $X_k, 1 \leqslant k \leqslant T$. It will be shown that, by properly choosing the bin size m, a robust procedure can be constructed to detect the signal segments reliably. We note that the noise variance for the transformed data, $1/\{4h^2(0)m\}$, can be easily estimated and the estimation error does not affect the theoretical results. So we shall assume that $h(0)$ is known in the next section. Estimation of $h(0)$ is discussed in Section 3.

2.2. Robust segment detection

Our first goal is signal detection, i.e. we wish to test $H_0 : \mathbb{I} = \emptyset$ against $H_1 : \mathbb{I} \neq \emptyset$. When the noise distribution is Gaussian, the GLRT, which applies a thresholding procedure on the extreme value of the likelihood ratio statistics of all possible intervals, has been proved to be an optimal procedure in Arias-Castro *et al.* (2005). However, the threshold that is used by the GLRT may perform poorly on non-Gaussian data. We propose the RSD, which applies a similar

procedure to the transformed data, and we show that the RSD provides robust results over a wide range of noise distributions satisfying inequality (5). For simplicity of presentation, we assume that $\mu_i > 0$, for $i = 1, \ldots, q$. When both positive and negative signal segments exist, a simple modification is to replace the relevant quantities by their absolute values.

The RSD procedure can be described as follows. After the local median transformation, define for any interval $\tilde{I}$

$$X(\tilde{I}) = \sum_{k \in \tilde{I}} X_k / \sqrt{|\tilde{I}|}, \tag{9}$$

and threshold

$$\lambda_n = \sqrt{\{2 \log(n)\}} / \{2 h(0) \sqrt{m}\}. \tag{10}$$

The RSD rejects H_0 when $\max_{\tilde{I} \in \mathbb{J}_T} \{X(\tilde{I})\} > \lambda_n$, where $\mathbb{J}_T$ is the collection of all possible intervals in $\{1, \ldots, T\}$.

Note that the threshold λ_n is chosen by analysing the distribution of $X(\tilde{I})$ under the null hypothesis H_0. By equation (6), we have

$$X(\tilde{I}) = \frac{Z(\tilde{I})}{2 h(0) \sqrt{m}} + \frac{\zeta(\tilde{I})}{\sqrt{m}} \tag{11}$$

under hypothesis H_0, where

$$Z(\tilde{I}) = \sum_{k \in \tilde{I}} Z_k / \sqrt{|\tilde{I}|}$$

and

$$\zeta(\tilde{I}) = \sum_{k \in \tilde{I}} \zeta_k / \sqrt{|\tilde{I}|}.$$

Since ζ_k are stochastically small random variables according to lemma 1, $\max_{\tilde{I} \in \mathbb{J}_T} \{\zeta(\tilde{I})\}$ should be much smaller than $\max_{\tilde{I} \in \mathbb{J}_T} \{Z(\tilde{I})\}$ for large m. The following lemma provides the asymptotic bounds for both $\max_{\tilde{I} \in \mathbb{J}_T} \{\zeta(\tilde{I})\}$ and $\max_{\tilde{I} \in \mathbb{J}_T} \{Z(\tilde{I})\}$.

Lemma 2. Assume expressions (1) and (5), and transformation (4) with $m = \log^{1+b}(n)$ for some $b > 0$. For the collection $\mathbb{J}_T$ of all the possible intervals in $\{1, \ldots, T\}$, we have

$$P\left[\max_{\tilde{I} \in \mathbb{J}_T} \{\zeta(\tilde{I})\} > \frac{a \sqrt{\log(T)}}{m}\right] \leqslant \frac{C}{\sqrt{\log(T)}} T^{-C}, \tag{12}$$

for some $a > 4$ and $C > 0$, and

$$P[\max_{\tilde{I} \in \mathbb{J}_T} \{Z(\tilde{I})\} > \sqrt{\{2 \log(T)\}}] \leqslant \frac{C}{\sqrt{\log(T)}}. \tag{13}$$

Lemma 2 and equation (11) imply that the threshold on $\max_{\tilde{I} \in \mathbb{J}_T} \{X(\tilde{I})\}$ should be approximately that of $\max_{\tilde{I} \in \mathbb{J}_T} \{Z(\tilde{I})\} / \{2 h(0) \sqrt{m}\}$, which is $\sqrt{\{2 \log(T)\}} / \{2 h(0) \sqrt{m}\}$. We set the threshold slightly more conservatively. The following theorem shows the control of familywise type I error.

Theorem 1. Assume expressions (1) and (5), and transformation (4) with $m = \log^{1+b}(n)$ for some $b > 0$. For the collection $\mathbb{J}_T$ of all the possible intervals in $\{1, \ldots, T\}$,

$$P_{H_0}[\max_{\tilde{I} \in \mathbb{J}_T} \{X(\tilde{I})\} > \lambda_n] \leqslant \frac{C}{\sqrt{\log(T)}} \to 0, \qquad T \to \infty.$$

The bound $C/\sqrt{\log(T)}$ converges to 0 quite slowly. To control the familywise type I error better, we can increase λ_n a little to $\sqrt{\{2(1+\varepsilon_n)\log(n)\}}/\{2h(0)\sqrt{m}\}$ for some $\varepsilon_n = o(1)$. This small increase does not change the theoretical results in this paper.

When there are segmental signals that are sufficiently strong, the RSD with threshold λ_n can successfully detect their existence while controlling the familywise type I error. This is shown in the following theorem.

Theorem 2. Assume expressions (1) and (5), and transformation (4) with $m = \log^{1+b}(n)$ for some $b > 0$. If there is some segment $I_j \in \mathbb{I}$ that satisfies

$$|I_j|/m \to \infty \tag{14}$$

and

$$\mu_j\sqrt{|I_j|} \geqslant \sqrt{\{2(1+\varepsilon)\log(n)\}}/\{2h(0)\} \tag{15}$$

for some $\varepsilon > 0$, then the RSD has the sum of the probabilities of type I and type II errors going to 0.

Condition (14) guarantees that the difference between $|I_j|/m$ and the cardinality of $\{J_k : J_k \subseteq I_j\}$ is negligible. Condition (15) shows the requirement for the signal strength of some segment I_j, which is a combined effect of μ_j and $|I_j|$. This condition is easier to satisfy for a bigger $h(0)$, which corresponds to a more concentrated noise distribution. This agrees with our intuition that the detection of signal segments should be easier if the observation noises are more concentrated.

Next we characterize the situation when the RSD cannot have asymptotically vanishing type I and type II errors. In fact, in this situation, no testing procedure works.

Theorem 3. Assume expressions (1) and (5), and transformation (4) with $m = \log^{1+b}(n)$ for some $b > 0$. If $\log|\mathbb{I}| = o\{\log(n)\}$, and for all segments $I_j \in \mathbb{I}$,

$$\log|I_j| = o\{\log(n)\}, \tag{16}$$

$$\mu_j\sqrt{|I_j|} \leqslant \sqrt{\{2(1-\varepsilon)\log(n)\}}/\{2h(0)\} \tag{17}$$

for some $\varepsilon > 0$, then no testing procedure constructed on the transformed data $X_1, \ldots, X_T$ has the sum of the probabilities of type I and type II errors going to 0.

The results in theorems 2 and 3 imply that the RSD is a near optimal procedure to detect short signal segments based on $X_1, \ldots, X_T$. It can successfully separate H_0 and H_1 on the basis of the transformed data whenever there is some testing procedure that can do so.

The RSD is robust over a wide range of noise distributions when assumption (5) is satisfied. However, in the special case when the Gaussian assumption holds for the noise distribution, the GLRT procedure that is specifically designed for this case is more efficient. The GLRT rejects H_0 if $\max_{\tilde{I}\in\mathbb{J}_n}\{Y(\tilde{I})\} > \sqrt{\{2\log(n)\}}$, where $\mathbb{J}_n$ is the collection of all possible intervals in $\{1, \ldots, n\}$ and $Y(\tilde{I}) = \Sigma_{i\in\tilde{I}} Y_i/\sqrt{|\tilde{I}|}$. We have the following proposition.

Proposition 1. Assume expression (1) and $\xi_i \sim N(0,1)$. If there is some segment $I_j \in \mathbb{I}$ that satisfies

$$\mu_j\sqrt{|I_j|} \geqslant \sqrt{\{2(1+\varepsilon)\log(n)\}} \tag{18}$$

for all $\varepsilon > 0$, then the GLRT built on the original Y_i has the sum of the probabilities of type I and type II errors going to 0. However, if $\log(q) = o\{\log(n)\}$, and, for all segments $I_j \in \mathbb{I}$,

$$\log |I_j| = o\{\log(n)\}, \tag{19}$$

$$\mu_j \sqrt{|I_j|} \leqslant \sqrt{\{2(1-\varepsilon)\log(n)\}}, \tag{20}$$

for some $\varepsilon > 0$, then no testing procedure has the sum of the probabilities of type I and type II errors going to 0.

This proposition generalizes theorems 2.1 and 2.2 in Arias-Castro *et al.* (2005). By comparing proposition 1 with theorems 2 and 3, we can see the exact power loss due to the local median transformation if the noise distribution is known to be Gaussian. When $\xi_i \sim N(0, 1)$, condition (5) is satisfied with $h(0) = 1/\sqrt{2(\pi)} \approx 0.4$. If we use the transformed data, the detection bound of $\mu_j \sqrt{I_j}$ is $\sqrt{\{2\log(n)\}}/\{2h(0)\} \approx 1.25\sqrt{\{2\log(n)\}}$, where $\sqrt{\{2\log(n)\}}$ is the corresponding bound for the original data. Therefore, the power loss is due to the stronger condition on $\mu_j \sqrt{I_j}$. However, a significant advantage of the RSD is that it automatically adapts to a wide range of unknown noise distributions, whereas the GLRT procedure specifically designed for the Gaussian case may fail completely if the noise distribution is heavy tailed.

3. Robust segment identification

In this section we turn to segment identification, which is to locate each $I_j \in \mathbb{I}$ when the alternative hypothesis is true. Recall the model $y_i = \alpha_i + \xi_i$, where $\alpha_i = \mu_j \mathbf{1}_{\{i \in I_j\}}$ for some $I_j \in \mathbb{I}$. In this section, we define $\underline{s} = \min_{I_j \in \mathbb{I}} |I_j|$, $\bar{s} = \max_{I_j \in \mathbb{I}} |I_j|$ and $\underline{d} = \min_{I_j \in \mathbb{I}} \{\text{distance between } I_j \text{ and } I_{j+1}\}$, and assume

$$\begin{aligned} \underline{s} &\geqslant \log^2(n), \\ \log(\bar{s}) &= o\{\log(n)\}, \end{aligned} \tag{21}$$

which means that the lengths of the signal segments are neither too long nor too short. Examples of such segments include those that have $|I_j| = \log^a(n)$, $a \geqslant 2$. Further, assume that

$$\log(q) = o\{\log(n)\}, \tag{22}$$

which implies that the number of signal segments is less than n^b for any $b > 0$.

To identify each $I_j \in \mathbb{I}$, a computationally efficient method, called the RSI, is proposed. The RSI first transforms data by expression (4) and then applies a similar procedure to LRS to the transformed data. The selected intervals have their statistics $X(\tilde{I})$ defined in expression (9) pass certain thresholds and achieve local maxima. The RSI is computationally efficient even for ultrahigh dimensional data. The transformation step significantly reduces the dimension by a factor of m. Further, the second step utilizes the short segment structure and considers only intervals with length less than or equal to L/m, where L is some number much smaller than n.

The ideal threshold for the RSI should be the same as that of the RSD, which is λ_n defined in expression (10). However, $h(0)$ is usually unknown in practice and needs to be estimated. By lemma 1, $1/\{2h(0)\sqrt{m}\}$ is approximately the standard deviation σ of the transformed noise η_k, which can be estimated accurately when signals are sparse. One such robust estimator is the median absolute deviation estimator:

$$\hat{\sigma} = \frac{\text{median}|X_k - \text{median}(X_k)|}{0.6745}. \tag{23}$$

Therefore, we can set the data-driven threshold for the RSI at

$$\lambda_n^* = \hat{\sigma}\sqrt{\{2\log(n)\}}. \tag{24}$$

The algorithm for the RSI for a fixed L can be stated as follows.

Step 1: transform the data by expression (4). Let $\mathbb{J}_T(L)$ be the collection of all possible subintervals in $\{1, \ldots, T\}$ with interval length L/m or less.
Step 2: let $j = 1$. Define $\mathbb{I}^{(j)} = \{\tilde{I} \in \mathbb{J}_T(L) : X(\tilde{I}) > \lambda_n^*\}$, where $X(\tilde{I})$ and λ_n^* are defined as in equations (9) and (24).
Step 3: let $I_j^* = \arg\max_{\tilde{I} \in \mathbb{I}^{(j)}} X(\tilde{I})$ and update $\mathbb{I}^{(j+1)} = \mathbb{I}^{(j)} \backslash \{\tilde{I} \in \mathbb{I}^{(j)} : \tilde{I} \cap I_j^* \neq \emptyset\}$.
Step 4: repeat step 3 with $j = j + 1$ until $\mathbb{I}^{(j)}$ is empty.
Step 5: for each I_j^* generated above, let l_j be the (first element in $I_j^* - 1) \times m + 1$ and r_j be the last element in $I_j^* \times m$, and denote $\hat{I}_j = \{l_j, \ldots, r_j\}$.

Denote the collection of selected intervals as $\hat{\mathbb{I}} = \{\hat{I}_1, \hat{I}_2, \ldots\}$. If $\hat{\mathbb{I}} \neq \emptyset$, we reject the null hypothesis and identify the signal segments by all the elements in $\hat{\mathbb{I}}$. Note that the above RSI procedure is designed for positive signal segments ($\mu_j > 0$). When both positive and negative signal segments exist, a simple modification is to replace the $X(\tilde{I})$ in steps 2 and 3 with $|X(\tilde{I})|$.

We now show that, with m and L chosen properly, the RSI consistently estimates the segmental signals if they are sufficiently strong. Define the dissimilarity between any pair of $\hat{I} \in \hat{\mathbb{I}}$ and $I \in \mathbb{I}$ as

$$D(\hat{I}, I) = 1 - |\hat{I} \cap I| / \sqrt{(|\hat{I}||I|)}. \tag{25}$$

It is clear that $0 \leqslant D(\hat{I}, I) \leqslant 1$ with $D(\hat{I}, I) = 1$ indicating disjointedness and $D(\hat{I}, I) = 0$ indicating complete identity. The following theorem presents estimation consistency of the RSI for $\mathbb{I}$.

Theorem 4. Assume expressions (1), (5), (21) and (22), and transformation (4) with $m = \log^{1+b}(n)$ for $0 < b < 1$. Suppose that $\hat{\sigma}$ is an n^{γ}-consistent estimator of the standard deviation of η_k for some $\gamma > 0$, and L satisfies

$$\begin{gathered} \bar{s} \leqslant L < \underline{d}, \\ \log(L) = o\{\log(n)\}. \end{gathered} \tag{26}$$

If condition (15) is satisfied for all $I_j \in \mathbb{I}$, then the RSI is consistent for $\mathbb{I}$ in the sense that

$$P_{H_0}(|\hat{\mathbb{I}}| > 0) + P_{H_1}[\max_{I_j \in \mathbb{I}} \min_{\hat{I}_j \in \hat{\mathbb{I}}} \{D(\hat{I}_j, I_j)\} > \delta_n] \to 0 \tag{27}$$

for some $\delta_n = o(1)$.

The asymptotic result (27) essentially says that both the probability of having at least one false positive result and the probability of some signal segments not being matched well by any of the selected intervals are asymptotically small. Condition (26) provides some insights on the selection of L. The range $[\bar{s}, \underline{d})$ is very large when signal segments are relatively short and rare as in the applications that we are interested in. The second part $\log(L) = o\{\log(T)\}$ is satisfied, if, for instance, $L = \log^a(T)$ for $a \geqslant 0$. More discussions and some sensitivity study on L for the original LRS procedure in the Gaussian case can be found in Jeng *et al.* (2010).

Recall theorem 3, which shows that, when signals are very sparse, no testing procedure based on the transformed data can separate H_0 and H_1 if, for all $I_j \in \mathbb{I}$,

$$\mu_j \sqrt{|I_j|} \leqslant \sqrt{\{2(1-\varepsilon)\log(n)\}} / \{2h(0)\}.$$

This implies that the RSI is a near optimal identification procedure for the transformed data when m and L are properly chosen. In other words, the RSI consistently estimates the signal segments whenever it is possible to detect their existence on the basis of the transformed data.

4. Simulation studies

We evaluate the finite sample behaviour of the RSI through simulation studies and compare its performance with the performances of LRS and another popular procedure: CBS (Olshen *et al.*, 2004).

4.1. Performance of robust segment identifier under various noise distributions

We generate ξ_i from a set of t-distributions with degrees of freedom 1, 3 and 30, where $t(1)$ is the standard Cauchy distribution, which has heavy tails. As the degree of freedom increases, the tails become thinner, and the t-distribution approaches the standard normal distribution. Nevertheless, the t-distributions satisfy the general assumption in expression (5). We set the sample size at $n = 5 \times 10^4$, the number of segments at $|\mathbb{I}| = 3$, the lengths of the segments at $|I_1| = 100$, $|I_2| = 40$ and $|I_3| = 20$, and the signal mean for all segments at $\mu = 1.0, 1.5, 2.0$. We randomly select three locations for the segments and generate the data from

$$Y_i = \alpha + \xi_i,$$

$i = 1, \ldots, n$, where $\alpha = \mu$ if i is in some signal segment, and $\alpha = 0$ otherwise.

We apply the RSI on Y_i with $m = 20$, $L = 120$ and $\lambda_n = \hat{\sigma}\sqrt{\{2 \log(T)\}}$, where $\hat{\sigma}$ is calculated as in equation (23). Fig. 2 shows the histograms of the original data $Y_i \in [-60, 60]$ with $t(1)$ noise and $\mu = 1$, and that of the transformed data X_k. Clearly, even though the original distribution is far from being Gaussian, the transformed data are close to being normally distributed. In this case, $m = 20$ is sufficiently large to stabilize the noise. More discussions on the choice of m are given in Section 4.3.

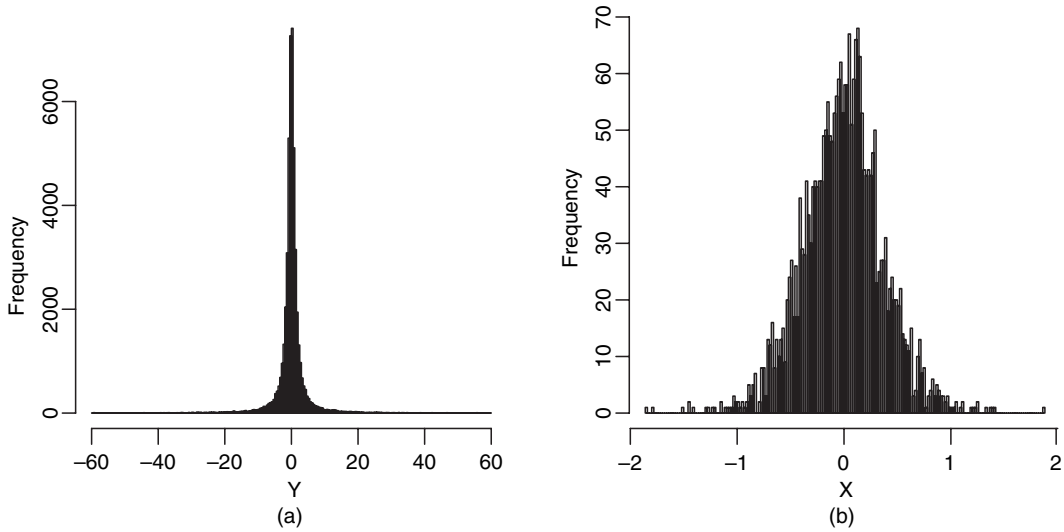

Fig. 2. Simulated data: (a) histogram of the original data Y_i with $t(1)$ noise and $\mu = 1$; (b) histogram of the transformed data X_k

Table 1. Simulation results: medians of D_j and $\#O$ for the RSI with $m = 20$ and $L = 6$†

Distribution	μ	$D_{1(\lvert I_1\rvert=100)}$	$D_{2(\lvert I_2\rvert=40)}$	$D_{3(\lvert I_3\rvert=20)}$	$\#O$
$t(1)$	1.0	0.080 (0.015)	1.000 (0.026)	1.000 (0.000)	2.000 (0.330)
	1.5	0.087 (0.003)	0.184 (0.017)	1.000 (0.000)	2.000 (0.260)
	2.0	0.087 (0.009)	0.150 (0.020)	0.423 (0.220)	2.000 (0.140)
$t(3)$	1.0	0.087 (0.005)	1.000 (0.270)	1.000 (0.000)	0.000 (0.000)
	1.5	0.060 (0.009)	0.175 (0.029)	1.000 (0.000)	0.000 (0.000)
	2.0	0.050 (0.008)	0.150 (0.016)	0.293 (0.019)	0.000 (0.000)
$t(30)$	1.0	0.070 (0.014)	1.000 (0.32)	1.000 (0.000)	0.000 (0.000)
	1.5	0.065 (0.012)	0.175 (0.021)	1.000 (0.245)	0.000 (0.000)
	2.0	0.050 (0.010)	0.175 (0.019)	0.250 (0.028)	0.000 (0.000)

†In Tables 1–3, the estimated standard errors based on the bootstrap appear in parentheses.

As used in Jeng *et al.* (2010), the accuracy of identification of the RSI is measured by two quantities, D_j and $\#O$, where D_j measures how well the signal segment I_j is estimated, and $\#O$ counts the number of overselections. In detail, for each signal segment I_j, define

$$D_j = \min_{\hat{I} \in \hat{\mathbb{I}}} \{D(\hat{I}, I_j)\},$$

where $D(\hat{I}, I_j)$ is defined in equation (25). Obviously, smaller D_j represents better matching between I_j and some estimate $\hat{I} \in \hat{\mathbb{I}}$, and $D_j = 0$ if and only if $I_j = \hat{I}$. Define

$$\#O = \#\{\hat{I} \in \hat{\mathbb{I}} : \hat{I} \cap I_j = \emptyset, \forall j = 1, \ldots, q\}.$$

So $\#O$ is a non-negative integer, and $\#O = 0$ if there are no overselected intervals. According to theorem 4, $\mu_j \sqrt{|I_j|}$ should be at least $\sqrt{\log(n)}/\{\sqrt{2}\, h(0)\} \approx 7.307$ for segment I_j to be consistently estimated by the RSI in this example.

We repeat the simulations 50 times to calculate D_j and $\#O$. The medians of $D_1, \ldots, D_q$ and $\#O$ are reported in Table 1 with estimated standard errors. To estimate the standard errors of the medians, we generate 500 bootstrap samples out of the 50 replication results, and then calculate a median for each bootstrap sample. The estimated standard error is the standard deviation of the 500 bootstrap medians. Table 1 shows that the RSI provides quite accurate estimates for any of the signal segments when μ is sufficiently large. In all the cases, the overselection error is controlled very well. It also shows that larger μ is needed for shorter segments, which agrees with our theoretical results. More importantly, the results are very stable over different noise distributions.

4.2. Comparison with likelihood ratio selection and circular binary segmentation

In the second set of simulations, we compare the performance of the RSI with that of the original LRS and CBS under various noise distributions. All the parameters are chosen the same as in the previous simulations except that μ is fixed at 2.0. Further, the maximum interval length L for the original LRS is set at 45. Table 2 shows that, in the cases of $t(1)$ and $t(3)$, LRS has lower power and a large number of overselections, whereas the RSI remains very stable. In contrast, CBS fails to select any true intervals. When the noise distribution is $t(30)$, which is very close to Gaussian, both LRS and CBS outperform the RSI with better power and better identification of the signal segments. However, the RSI still performs reasonably well and has no overselection. This agrees with our theoretical results in Section 2.2.

Table 2. Simulation comparisons of the RSI, LRS and CBS, where both homogeneous and heterogeneous noises are considered†

	Results for RSI		*Results for LRS*		*Results for CBS*	
	$D_{2(\lvert I_2\rvert=40)}$	$\#O$	$D_{2(\lvert I_2\rvert=40)}$	$\#O$	$D_{2(\lvert I_2\rvert=40)}$	$\#O$
$t(1)$	0.163 (0.024)	2 (0.2)	0.340 (0.054)	3882 (6.6)	1.000 (0.000)	0 (0.0)
$t(3)$	0.125 (0.028)	0 (0.0)	0.025 (0.006)	467 (4.4)	1.000 (0.000)	0 (0.0)
$t(30)$	0.125 (0.018)	0 (0.0)	0.000 (0.001)	2 (0.0)	0.006 (0.006)	0 (0.0)
$\tau=0.5$	0.125 (0.015)	2 (0.4)	0.013 (0.005)	37 (3.1)	0.180 (0.006)	4 (0.6)
$\tau=1.0$	0.113 (0.022)	12 (0.6)	0.000 (0.006)	227 (6.1)	1.000 (0.010)	10 (1.1)
$\tau=1.5$	0.125 (0.016)	26 (0.8)	0.000 (0.006)	461 (10.9)	1.000 (0.000)	8 (1.1)

†Homogeneous noise is generated from the t-distribution with degrees of freedom 1, 3 and 30. Heterogeneous noise is generated from a mixture of $N(0,1)$ and $N(0,\sigma^2)$, where $\sigma\sim$ gamma$(2,\tau)$. μ is fixed at 2.0.

We next consider the case when the errors have heterogeneous variances along the genome. The baseline noise is generated from $N(0,1)$. We randomly select 100 intervals with length 50 and generate heterogeneous noise in these intervals. In each interval, the noises follow $N(0,\sigma^2)$, where σ is generated from gamma$(2,\tau)$ with $\tau=0.5,1,1.5$. Note that the noise variances are constant within an interval but different for different intervals. The bottom half of Table 2 shows a comparison of the three procedures. The RSI still has the best overall performance. It results in smaller numbers of overselections than LRS and better power than CBS. However, as the noise variance increases, the RSI can result in more overselections.

Computation is more expensive for LRS, especially for the $t(1)$ and $t(3)$ cases, because a large number of intervals pass the threshold of LRS. In contrast, the RSI is computationally much more efficient because the data transformation step regularizes the noise and also reduces the dimension from n to T.

4.3. Choice of m

The third set of simulations evaluates the effect of the bin size m on the performance of the RSI. We use the same simulation setting as in the first set of simulations except that μ is fixed at 2.0; m takes values of 10, 20 and 40.

Table 3 shows that there is a trade-off between the power and overselection when m varies.

Table 3. Simulation results: effect of bin size m on the performance of the RSI†

Distribution		$D_{1(\lvert I_1\rvert=100)}$	$D_{2(\lvert I_2\rvert=40)}$	$D_{3(\lvert I_3\rvert=20)}$	$\#O$
$t(1)$	10	0.035 (0.009)	0.10 (0.018)	0.184 (0.033)	19.000 (0.850)
	20	0.087 (0.009)	0.15 (0.020)	0.423 (0.220)	2.000 (0.140)
	40	0.101 (0.006)	0.25 (0.056)	1.000 (0.024)	0.000 (0.000)
$t(3)$	10	0.030 (0.004)	0.088 (0.015)	0.150 (0.033)	1.000 (0.220)
	20	0.050 (0.008)	0.150 (0.016)	0.293 (0.019)	0.000 (0.000)
	40	0.087 (0.006)	0.293 (0.041)	1.000 (0.250)	0.000 (0.000)
$t(30)$	10	0.020 (0.007)	0.075 (0.008)	0.150 (0.018)	0.000 (0.000)
	20	0.050 (0.010)	0.175 (0.019)	0.250 (0.028)	0.000 (0.000)
	40	0.105 (0.008)	0.293 (0.035)	1.000 (0.094)	0.000 (0.000)

†μ is fixed at 2.

Smaller m results in better power but more overselections, especially when the noise distribution has a heavy tail. The greater power is due to finer binning, which preserves the signal information better. In contrast, when m is large, it is possible that none of the original observations in segments with length less than m is kept in the transformed data, such as the case when $m = 40$ and $I_3 = 20$. However, if m is too small, the Gaussian approximation of the transformed noise is not sufficiently accurate to overcome the effect of the heavy-tailed noise on segment selection, which in turn leads to more overselections in the case of the $t(1)$ distribution and $m = 10$.

5. Application to identification of copy number variants based on the next generation sequencing data

To demonstrate our proposed methods, we analyse the short reads data on chromosome 19 of a HapMap Yoruban female sample (NA19240) from the 1000 genomes project. Let Y_i denote the guanine–cytosine content adjusted number of short reads that cover the base pair position i in the genome, for $i = 1, \ldots, n$ where n is very large. After the short reads have been mapped to the reference human DNA sequences, we obtain the RD data at $n = 54361060$ genomic locations. Fig. 3(a) shows the RD for the first 10000 observations of the data set. The median of the count number over all n sites is 30.

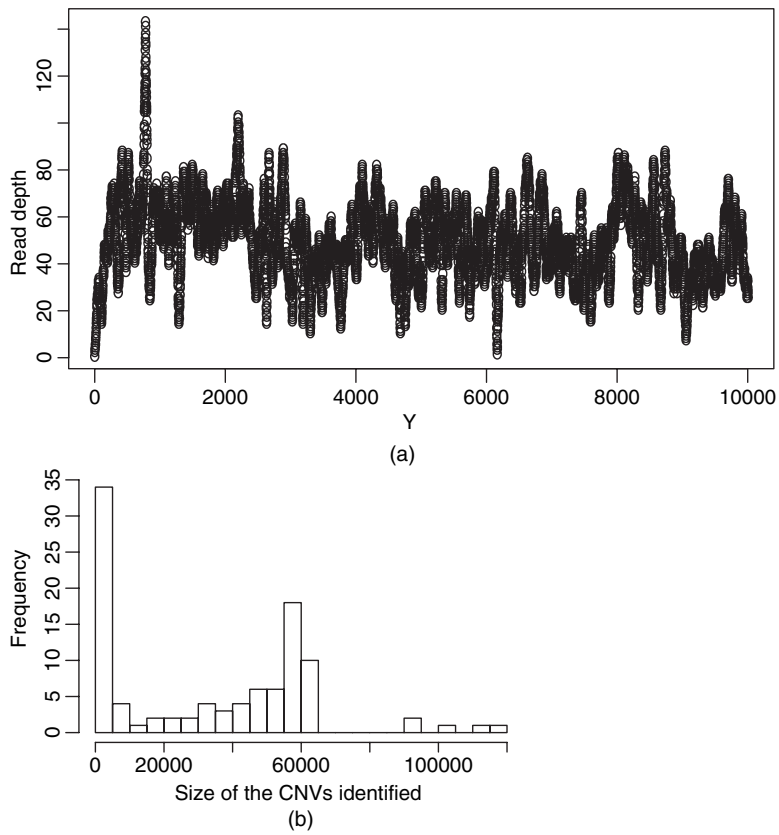

Fig. 3. Analysis of the chromosome 19 data of individual NA19240 from the 1000 genomes project: (a) scatter plot of the first 10000 observations; (b) histogram of the sizes of the CNVs identified in base pairs

The statistical challenges for CNV detection based on NGS data include both ultrahigh dimensionality of the data that requires fast computation and the unknown distribution of the RDs data. A close examination of our data shows that the variance of the data is much larger than its mean, indicating that the standard Poisson distribution cannot be used for modelling these RD data.

We apply the proposed RSI with $m = 400$ and $L = 60000$, which assumes that the CNVs based on our preprocessed data are less than 60000 base pairs (BPs). This is sensible since typical CNVs include multikilobase deletions and duplications (McCarroll and Altshuler, 2007; Medvedev *et al.*, 2009). We chose $L = 60000$ partially because of computational considerations. If the true CNVs are longer than the maximum allowable interval length, these intervals are often divided into several contiguous segments; we can then simply perform some post-processing to merge these segments into longer ones. The RSI selected 115 CNV segments, ranging from 400 to 75991 BPs in sizes; among these 24 are contiguous segments. After merging these contiguous segments, we obtained 101 CNVs, ranging from 400 to 125440 BPs in size with a median size of 38860 BPs. See Fig. 3(b) for the distributions of the sizes of CNVs identified. There are eight CNVs of size 400 BPs, four CNVs of size 800 BPs and four CNVs of size 1200 BPs. These small CNVs are identified since we did not set a lower limit on the sizes of the CNVs.

To visualize the CNVs that are identified by the RSI, Fig. 4 shows six CNVs identified, including two duplications, two deletions and two regions with the shortest CNVs. It is clear that these identified regions indeed represent the regions with different RDs from their neighbouring regions. Examinations of the other CNV regions that were identified also show that these regions contain more or fewer reads than their neighbouring regions, further indicating the effectiveness of the RSI procedure in identifying the CNVs.

Mills *et al.* (2011) recently reported a map of CNVs based on whole genome DNA sequencing data from 185 human genomes from the 1000 genomes project, where both the RDs and the pair end mapping distances were used for CNV identifications. Among the methods that are applied in Mills *et al.* (2011), three were based on the RDs data as we used in our data set. These three methods identified a total of 438, 332 and 615 CNVs on chromosome 19, on the basis of the data from all 185 samples. Out of the 101 CNVs we identified for one single sample NA19240, 76 of them overlap with the CNVs that were reported in the CNV map of Mills *et al.* (2011), indicating high sensitivity of our method in detecting the CNVs on the basis of the RD data since our CNVs calls are based on data from only a single sample (NA19240). We cannot make a direct comparison on CNV calls for this particular sample since the sample level CNV calls are not available from Mills *et al.* (2011).

6. Conclusion and further discussion

Motivated by CNV analysis based on the RD data generated by the NGS technology, we have studied the problem of segment detection and identification from an ultralong linear sequence of data with an unknown noise distribution. We have developed the robust detection procedure RSD and the robust segment identification procedure RSI, which are adaptive over a wide range of noise distributions and are near optimal for the transformed data. The RSI procedure has been applied to identify the CNVs based on the NGS data of a Yoruban female sample from the 1000 genomes project. The CNV regions identified all show deviated RDs when compared with the neighbouring regions. The key ingredient of our approaches is a local median transformation of the original data. This not only provides the basis for our algorithm and theoretical development but also saves a large amount of computational cost by greatly reducing the data

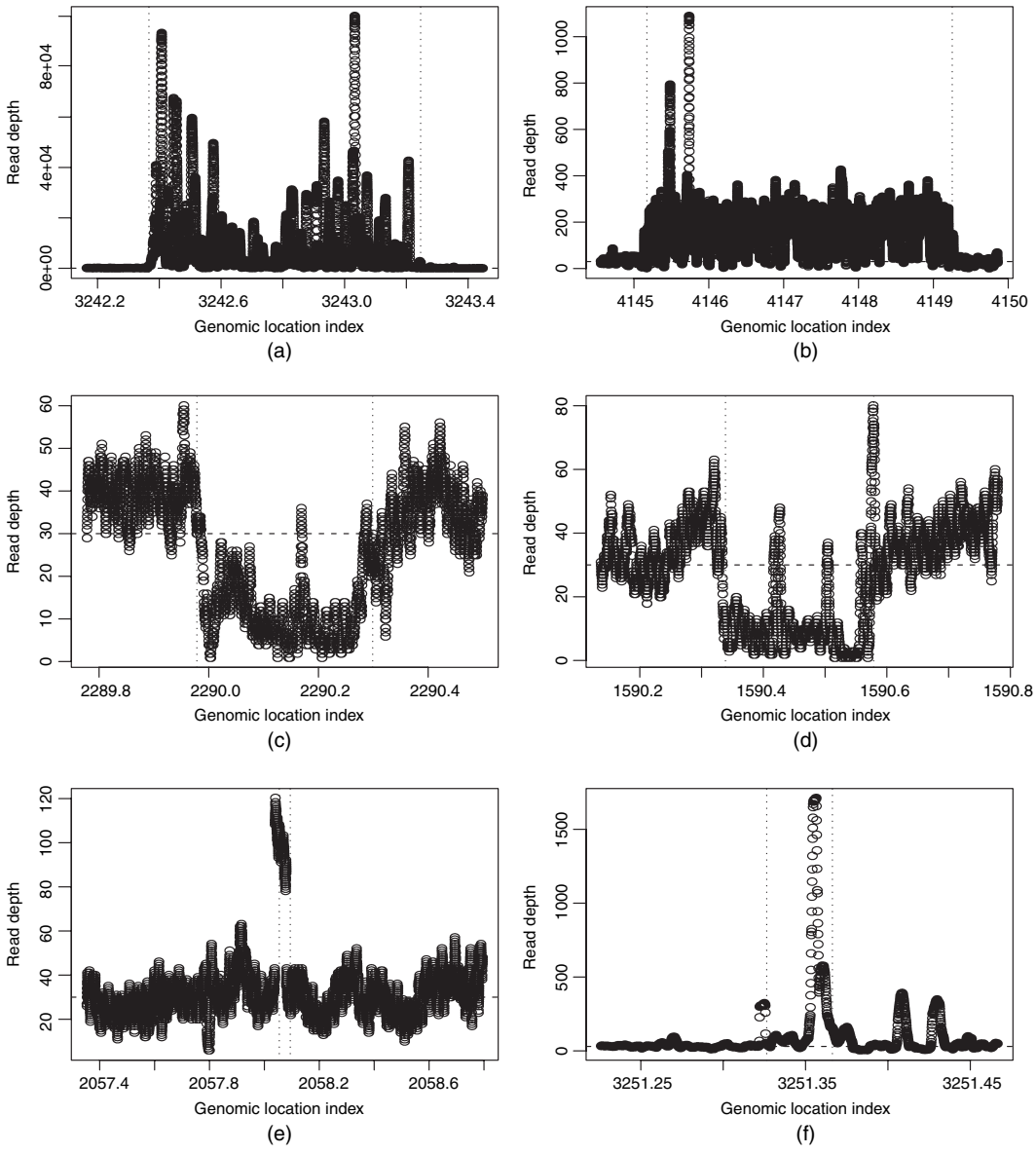

Fig. 4. Examples of CNVs identified on chromosome 19 of individual NA19240 from the 1000 genomes project (for each plot, the x-axis is the genome location in BPs per 10000) (-------, median count of 30; ⋮, estimated CNV boundaries): (a), (b) duplications, regions with the highest scores; (c), (d) deletions, regions with the smallest scores; (e), (f) the two shortest CNVs identified

dimension, which makes it particularly appealing for analysing the ultrahigh dimensional NGS data. As increasingly more NGS data are becoming available, we expect to see more applications of the RSI procedure in CNV identifications.

Model (1) does not require a specification of the noise distribution. However, we assume that the noise is IID, which can be violated for the RD data. The noise distribution of the RD data is directly related to the uncertainty and errors that are inherent in the sequencing process and

is the result of complex processing of noisy continuous fluorescence intensity measurements known as base calling. Bravo and Irizarry (2010) showed that the complexity of the base calling discretization process results in reads of widely varying quality within sequence samples and this variation in processing quality results in infrequent but systematic errors. Such errors can lead to violation of the assumption of IID data for the RD data. Although such errors can have a great effect on analysis of single-nucleotide polymorphisms and rare genetic variants, their effect on the RD data distribution and CNV identification is not clear. Our simulations (Table 2) showed that, when the noise has heterogeneous variances, the RSI can still identify the correct CNVs unless the variance is very large.

The methods that are presented in this paper can be extended to more general settings, where one only needs to assume that the density function h of the noise satisfies

$$\int_{-\infty}^{0} h(y) = \tfrac{1}{2}, \qquad h(0) > 0 \qquad \text{and } h(y) \text{ is Lipschitz at } y = 0.$$

Obviously, this assumption is more general than expression (5) (Brown *et al.*, 2008). To accommodate this more general assumption, a larger m is needed for the robust methods that are developed in this paper. Our methods can also be extended to detect and identify general geometric objects in two-dimensional settings (Arias-Castro *et al.*, 2005; Walther, 2010). Other interesting extensions include identification of CNVs shared between multiple individuals based on the NGS data and to test for CNV associations. Our methods provide the basic tools for these extensions to consider structured signals with unknown and possibly heavy-tailed noise distributions.

Acknowledgements

Jeng and Li were supported by National Institutes of Health grants ES009911 and CA127334. The research of Tony Cai was supported in part by National Science Foundation Focused Research Group grant DMS-0854973. We thank Dr Mingyao Li for providing the sequence data from the 1000 genomes project.

Appendix A: Proofs

In this appendix we present the proofs for lemma 2 and theorems 1–4.

A.1. Proof of lemma 2

Condition (13) has been proved in Arias-Castro *et al.* (2005). We only need to show result (12).

Decompose $\zeta(\tilde{I})$ as $\zeta(\tilde{I}) = \zeta_1(\tilde{I}) + \zeta_2(\tilde{I})$, where $\zeta_1(\tilde{I}) = \Sigma_{k\in\tilde{I}}\zeta_{k1}/\sqrt{|\tilde{I}|}$ and $\zeta_2(\tilde{I}) = \Sigma_{k\in\tilde{I}}\zeta_{k2}/\sqrt{|\tilde{I}|}$. Then

$$P\{|\zeta(\tilde{I})| > x\} \leqslant P\{|\zeta_1(\tilde{I})| > x/2\} + P\{|\zeta_2(\tilde{I})| > x/2\}. \tag{28}$$

Let $A_k = m\zeta_{k1}$; then, by expression (7) in lemma 1, $E(A_k) = 0$ and $E(|A_k|^l) \leqslant C_l$ for any $l > 0$. According to lemma 2 in Zhou (2006), there is some positive constant ε' such that, for any $0 \leqslant x \leqslant \varepsilon'$ and interval $\tilde{I}$,

$$P(|\sum_{k\in\tilde{I}} A_k/\sqrt{|\tilde{I}|}| > x) \leqslant \bar{\Phi}(x)\exp\{O(1/\sqrt{|\tilde{I}|})\},$$

where $\bar{\Phi}$ is the survival function of a standard normal random variable. Then we have

$$P\left\{|\zeta_1(\tilde{I})| > \frac{x}{2}\right\} \leqslant P\left(\left|\sum_{k\in\tilde{I}} \frac{A_k}{\sqrt{|\tilde{I}|}}\right| > \frac{xm}{2}\right) \leqslant C\,\bar{\Phi}\left(\frac{xm}{2}\right) \leqslant \frac{C}{xm}\exp\left(-\frac{x^2m^2}{8}\right), \tag{29}$$

where the last step is by Miller's inequality. In contrast, $|\zeta_2(\tilde{I})| > x/2$ implies that there is some ζ_{k2} such that $|\zeta_{k2}| > x/(2\sqrt{|\tilde{I}|})$; then

$$P\{|\zeta_2(\tilde{I})| > x/2\} \leqslant T\, P\{|\zeta_{k2}| > x/(2\sqrt{|\tilde{I}|})\} \leqslant CT\exp(-am) \leqslant Cn^{-C\log^b(n)}, \tag{30}$$

where the second inequality is by expression (8) and the last inequality is by the choice of m. Combining expressions (28)–(30) we have

$$P\{|\zeta(\tilde{I})| > x\} \leqslant \frac{C}{xm}\exp\left(-\frac{x^2m^2}{8}\right) + Cn^{-C\log^b(n)}, \tag{31}$$

and consequently

$$P[\max_{\tilde{I}\in\mathbb{J}_T}\{\zeta(\tilde{I})\} > x] \leqslant T^2\, P\{|\zeta(\tilde{I})| > x\} \leqslant \frac{C}{xm}\exp\left\{2\log(T) - \frac{x^2m^2}{8}\right\} + Cn^{-C\log^b(n)}.$$

Therefore, result (12) follows by letting $x = a\sqrt{\log(T)/m}$ for some $a > 4$.

A.2. Proof of theorem 1

Decompose $P_{H_0}[\max_{\tilde{I}\in\mathbb{J}_T}\{X(\tilde{I})\} > \lambda_n]$ into two terms:

$$\begin{aligned} P_{H_0}[\max_{\tilde{I}\in\mathbb{J}_T}\{X(\tilde{I})\} > \lambda_n] &= P_{H_0}[\max_{\tilde{I}\in\mathbb{J}_T}\{X(\tilde{I})\} > \lambda_n,\ \max_{\tilde{I}\in\mathbb{J}_T}\{\zeta(\tilde{I})\} \leqslant 5\sqrt{\log(T)/m}] \\ &\quad + P_{H_0}[\max_{\tilde{I}\in\mathbb{J}_T}\{X(\tilde{I})\} > \lambda_n,\ \max_{\tilde{I}\in\mathbb{J}_T}\{\zeta(\tilde{I})\} > 5\sqrt{\log(T)/m}]. \end{aligned}$$

By expressions (11) and (13), the first term

$$\begin{aligned} P_{H_0}\left[\max_{\tilde{I}\in\mathbb{J}_T}\{X(\tilde{I})\} > \lambda_n,\ \max_{\tilde{I}\in\mathbb{J}_T}\{\zeta(\tilde{I})\} \leqslant \frac{5\sqrt{\log(T)}}{m}\right] &\leqslant P\left[\max_{\tilde{I}\in\mathbb{J}_T}\{Z(\tilde{I})\} > \sqrt{\{2\log(n)\}} - \frac{10\,h(0)}{m}\sqrt{\log(T)}\right] \\ &\leqslant P[\max_{\tilde{I}\in\mathbb{J}_T}\{Z(\tilde{I})\} > \sqrt{\{2\log(T)\}}] \leqslant \frac{C}{\sqrt{\log(T)}}. \end{aligned}$$

In contrast, it is easy to show that the second term is bounded by $CT^{-C}/\sqrt{\log(T)}$ by using result (12). The result follows by combining the upper bounds of the two terms.

A.3. Proof of theorem 2

Since theorem 1 implies that the type I error of the RSD goes to 0, all we need to show is that

$$P_{H_1}[\max_{\tilde{I}\in\mathbb{J}_T}\{X(\tilde{I})\} \leqslant \lambda_n] \to 0. \tag{32}$$

Suppose that under hypothesis H_1 segment $I_j \in \mathbb{I}$ satisfies conditions (14) and (15). Define $\tilde{I}_j$ to be the collection of the index of J_k such that $J_k \subseteq I_j$, i.e.

$$\tilde{I}_j = \{k : J_k \subseteq I_j\}; \tag{33}$$

then

$$|\tilde{I}_j| \geqslant \lfloor |I_j|/m - 1\rfloor > |I_j|/m - 2, \tag{34}$$

and, for each $k \in \tilde{I}_j$, $\theta_k = \mu_j$. This combined with equations (4) and (6) implies that

$$X_k = \mu_j + \frac{Z_k}{2h(0)\sqrt{m}} + \frac{\zeta_k}{\sqrt{m}}, \qquad k \in \tilde{I}_j,$$

which further implies that

$$X(\tilde{I}_j) = \mu_j\sqrt{|\tilde{I}_j|} + \frac{Z(\tilde{I}_j)}{2h(0)\sqrt{m}} + \frac{\zeta(\tilde{I}_j)}{\sqrt{m}}.$$

By expressions (34), (14) and (15), we have

$$\mu_j\sqrt{|\tilde{I}_j|} \geqslant \frac{\mu_j\sqrt{|I_j|}}{\sqrt{m}}\sqrt{\left(1-\frac{2m}{|I_j|}\right)} \geqslant \frac{\mu_j\sqrt{|I_j|}}{\sqrt{m}}\sqrt{\left\{1-\frac{\varepsilon}{2(1+\varepsilon)}\right\}} \geqslant \frac{\sqrt{\{2(1+\varepsilon/2)\log(n)\}}}{2h(0)\sqrt{m}}.$$

Then

$$\begin{aligned} P_{H_1}\{X(\tilde{I}_j) \leqslant \lambda_n\} &\leqslant P[Z(\tilde{I}_j) \leqslant \sqrt{\{2\log(n)\}} - \sqrt{\{2(1+\varepsilon/2)\log(n)\}} - 2h(0)\zeta(\tilde{I}_j)] \\ &\leqslant P\left[N(0,1) \leqslant -\sqrt{\{2\log(n)\}}\left\{\sqrt{(1+\varepsilon/2)} - 1 - \frac{h(0)\varepsilon\sqrt{2}}{m}\right\}\right] + P\left\{\zeta(\tilde{I}_j) < -\frac{\varepsilon\sqrt{\log(n)}}{m}\right\} \\ &\leqslant Cn^{-C\varepsilon^2}, \end{aligned} \tag{35}$$

where the last inequality is by Miller's inequality and inequality (31). Expression (32) follows directly.

A.4. Proof of theorem 3

Let $\bar{s} = \max_{I_j \in \mathbb{I}} \lceil |I_j|/m \rceil$ and $\bar{I}_j = \{k : J_k \cap I_j \neq \emptyset\}$. Assume that each $\bar{I}_j$ is in $\{l_j\bar{s}+1, \ldots, (l_j+1)\bar{s}\}$ for some l_j, where $\tilde{I}_j$ is defined in equation (33). We show that no testing procedure has both type I and type II errors going to 0 under this situation. This is enough to show that no procedure has both type I and type II errors going to 0 without this assumption. Let

$$W_l = (X_{l\bar{s}+1} + \ldots + X_{(l+1)\bar{s}})/\sqrt{\bar{s}}, \qquad l = 0, \ldots, \lfloor T/\bar{s} \rfloor - 1;$$

then W_l can be rewritten as

$$2h(0)W_l\sqrt{m} = \theta_l' + Z_l' + 2h(0)\zeta_l',$$

where $Z_l' \sim^{\mathrm{IID}} N(0,1)$, $\zeta_l' = (\zeta_{l\bar{s}+1} + \ldots + \zeta_{(l+1)\bar{s}})/\sqrt{\bar{s}}$, and $\theta_l' = 0$ at all except at most $|\mathbb{I}|$ positions where $\theta_l' \leqslant \sqrt{\{2(1-\varepsilon)\log(n)\}}$ by expression (17).

By the well-known relationship between the L_1-distance and the Hellinger distance, it is enough to show that the Hellinger affinity between the distribution of $2h(0)W_l\sqrt{m}$ under the null and that under the alternative tends to $1 - o(1/n)$, i.e. define

$$g(x) = \frac{f_{Z_l'+2h(0)\zeta_l'}[x - \sqrt{\{2(1-\varepsilon)\log(n)\}}]}{f_{Z_l'+2h(0)\zeta_l'}(x)},$$

where $f_{Z_l'+2h(0)\zeta_l'}$ represents the density function of $Z_l' + 2h(0)\zeta_l'$, and it is enough to show that

$$E[\sqrt{\{1-\nu+\nu\, g(X)\}}] = 1 - o(1/n),$$

where $\nu = |\mathbb{I}|/\lfloor T/\bar{s} \rfloor \leqslant |\mathbb{I}| \max_j |I_j|/n$ and $\mathcal{L}(X) = \mathcal{L}\{Z_l' + 2h(0)\zeta_l'\}$. Further, define $D_n = \{|X| \leqslant \sqrt{\{2\log(n)\}}\}$. Then $E[\sqrt{\{1-\nu+\nu\, g(X)\,\mathbf{1}_{D_n}\}}] \leqslant E[\sqrt{\{1-\nu+\nu\, g(X)\}}] \leqslant 1$. Applying Taylor's expansion gives

$$E[\sqrt{\{1-\nu+\nu\, g(X)\mathbf{1}_{D_n}\}}] = 1 - \frac{\nu}{2}E\{g(X)\,\mathbf{1}_{D_n^c}\} + \mathrm{err},$$

where, by the Cauchy–Schwarz inequality,

$$|\mathrm{err}| \leqslant C\nu^2\, E\{g(X)\,\mathbf{1}_{D_n} - 1\}^2 \leqslant C\nu^2[E\{g^2(X)\,\mathbf{1}_{D_n}\} + 1].$$

Now, by the range of ν and the conditions on $|\mathbb{I}|$ and $\max_j |I_j|$, it is sufficient to show that

$$E\{g(X)\,\mathbf{1}_{D_n^c}\} = o(1)$$

and

$$E\{g^2(X)\,\mathbf{1}_{D_n}\} = o(n).$$

The following lemma is implied by lemma 1 in Cai *et al.* (2011).

Lemma 3.

$$\int_{D_n^c} \phi[x - \sqrt{\{2(1-\varepsilon)\log(n)\}}]\,\mathrm{d}x = o(1)$$

and

$$\int_{D_n} \frac{\phi^2[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}]}{\phi(x)}\,\mathrm{d}x = o(n),$$

where ϕ is the density function of a standard normal random variable.

Then it is enough to show that

$$E\{g(X)\mathbf{1}_{D_n^c}\} = \int_{D_n^c} \phi[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}]\,\mathrm{d}x\{1+o(1)\}+o(1) \tag{36}$$

and

$$E\{g^2(X)\,\mathbf{1}_{D_n}\} = \int_{D_n} \frac{\phi^2[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}]}{\phi(x)}\,\mathrm{d}x\{1+o(1)\}. \tag{37}$$

Consider equation (36) first. By convolution,

$$\begin{aligned}
E\{g(X)\,\mathbf{1}_{D_n^c}\} &= \int_{D_n^c} f_{Z_l'+2h(0)\zeta_l'}[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}]\,\mathrm{d}x \\
&= \int_{D_n^c}\left(\int_{-\infty}^{\infty} f_{\zeta_l'}(w)\,\phi[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}-2h(0)w]\,\mathrm{d}w\right)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} f_{\zeta_l'}(w)\left(\int_{D_n^c}\phi[x-\sqrt{2(1-\varepsilon)\log(n)}-2h(0)w]\,\mathrm{d}x\right)\mathrm{d}w = \mathrm{I}+\mathrm{II},
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{I} &= \int_{|w|>4\sqrt{\log(n)/m}} f_{\zeta_l'}(w)\left(\int_{D_n^c}\phi[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}-2h(0)w]\,\mathrm{d}x\right)\mathrm{d}w \\
&\leqslant P\{|\zeta_l'|>4\sqrt{\log(n)/m}\}\to 0
\end{aligned}$$

by inequality (31), and

$$\begin{aligned}
\mathrm{II} &= \int_{|w|\leqslant 4\sqrt{\log(n)/m}} f_{\zeta_l'}(w)\left(\int_{D_n^c}\phi[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}-2h(0)w]\,\mathrm{d}x\right)\mathrm{d}w \\
&= \int_{|w|\leqslant 4\sqrt{\log(n)/m}} f_{\zeta_l'}(w)\,\mathrm{d}w\int_{D_n^c}\phi[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}]\,\mathrm{d}x+\mathrm{err} \\
&= \int_{D_n^c}\phi[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}]\,\mathrm{d}x\{1+o(1)\}+\mathrm{err}
\end{aligned}$$

where the last step is by inequality (31) again, and

$$\begin{aligned}
|\mathrm{err}| &\leqslant C\frac{\sqrt{\log(n)}}{m}\left|\int_{D_n^c}\phi'[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}]\,\mathrm{d}x\right|\int_{|w|\leqslant 4\sqrt{\log(n)/m}} f_{\zeta_l'}(w)\,\mathrm{d}w \\
&\leqslant C\frac{\sqrt{\log(n)}}{m}n^{-\{1-\sqrt{(1-\varepsilon)}\}^2}\to 0.
\end{aligned}$$

Summing up above gives equation (36).

Next, consider equation (37). By convolution again,

$$f_{Z_l'+2h(0)\zeta_l'}(x) = \int_{-\infty}^{\infty} f_{\zeta_l'}(w)\,\phi\{x-2h(0)w\}\,\mathrm{d}w = \mathrm{III}+\mathrm{IV}$$

where

$$\mathrm{III} = \int_{|w|>4\sqrt{\log(n)/m}} f_{\zeta_l'}(w)\,\phi\{x-2h(0)w\}\,\mathrm{d}w \leqslant C \int_{|w|>4\sqrt{\log(n)/m}} f_{\zeta_l'}(w)\,\mathrm{d}w \leqslant \frac{Cm}{\sqrt{\log(n)}} n^{-2}$$

by inequality (31). Note that, in D_n, $\phi(x) \geqslant Cn^{-1}$; then $\mathrm{III} = o\{\phi(x)\}$. Further, we have

$$\mathrm{IV} = \int_{|w|\leqslant 4\sqrt{\log(n)/m}} f_{\zeta_l'}(w)\,\phi\{x-2h(0)w\}\,\mathrm{d}w = \phi(x)\{1+o(1)\} + \mathrm{err}_1,$$

where

$$|\mathrm{err}_1| \leqslant C\frac{\sqrt{\log(n)}}{m}|\phi'(x)| \int_{|w|\leqslant 4\sqrt{\log(n)/m}} f_{\zeta_l'}(w)\,\mathrm{d}w = o\{\phi(x)\}$$

by the choice of m and the fact that $|\phi'(x)| \leqslant C\,\phi(x)$ for all x. Summing up above gives

$$f_{Z_l'+2h(0)\zeta_l'}(x) = \phi(x)\{1+o(1)\} \tag{38}$$

in D_n. Similarly,

$$f_{Z_l'+2h(0)\zeta_l'}[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}] = \phi[x-\sqrt{\{2(1-\varepsilon)\log(n)\}}]\{1+o(1)\} \tag{39}$$

in D_n. Substituting equations (38) and (39) into the definition of $E\{g^2(X)\mathbf{1}_{D_n}\}$ gives equation (37).

A.5. Proof of theorem 4

It is sufficient to show that the set $\hat{\mathbb{I}}$ of the RSI satisfies

$$P_{H_0}(|\hat{\mathbb{I}}|>0) \leqslant \frac{C}{\sqrt{\log(T)}} \tag{40}$$

and

$$P_{H_1}[\max_{I_j\in\mathbb{I}} \min_{\hat{I}_j\in\hat{\mathbb{I}}}\{D(\hat{I}_j, I_j)\} > \delta_n] \leqslant Cqn^{-C\varepsilon^2} + C\,q(\bar{s}/m)(L/m)n^{-C\delta_n^2} \tag{41}$$

for any δ_n such that $\sqrt{\{\log(q)+\log(\bar{s}/m)+\log(L/m)\}}/\sqrt{\log(n)} \ll \delta_n \ll 1$. Note that $\sqrt{\{\log(q)+\log(\bar{s}/m)+\log(L/m)\}}/\sqrt{\log(n)} = o(1)$ under conditions (21), (22) and (26), and the choice of m. Consider inequality (40) first. Since

$$P_{H_0}(|\hat{\mathbb{I}}|>0) \leqslant P_{H_0}[\max_{\tilde{I}\in\mathbb{J}_T(L)}\{X(\tilde{I})\} > \lambda_n^*] \leqslant P_{H_0}[\max_{\tilde{I}\in\mathbb{J}_T}\{X(\tilde{I})\} > \lambda_n^*]$$

and λ_n^* converges to λ_n at the order of $n^\gamma/\sqrt{\log(n)}$, then, by theorem 1 and some routine calculation, inequality (40) follows.

Next, we show inequality (41). Since all the elements in $\mathbb{J}_T(L)$ cannot reach more than one signal segment, the accuracy of estimating any $I_j \in \mathbb{I}$ by some element in $\hat{\mathbb{I}}$ is not influenced by the estimation of other segments in $\mathbb{I}$. This means that the accuracy of estimating any $I_j \in \mathbb{I}$ is equivalent to the case when only segment I_j exists. Define the following events:

$$A_j = \{\mathbb{I}^{(1)} \neq \emptyset \text{ when only } I_j \text{ exists}\}, \qquad B_j = \{D(\hat{I}_1, I_j) \leqslant \delta_n\}, \qquad j = 1, \ldots, q.$$

We have

$$\begin{aligned} P_{H_1}[\max_{I_j\in\mathbb{I}} \min_{\hat{I}_j\in\hat{\mathbb{I}}}\{D(\hat{I}_j, I_j)\} > \delta_n] &\leqslant P\{\exists I_j \in \mathbb{I} : A_j^{\mathrm{c}} \cup (A_j \cap B_j^{\mathrm{c}})\} \\ &\leqslant \sum_{j=1}^{q} P(A_j^{\mathrm{c}}) + \sum_{j=1}^{q} P(B_j^{\mathrm{c}}|A_j), \end{aligned}$$

and it is sufficient to show that, for any $j \in \{1, \ldots, q\}$,

$$P(A_j^{\mathrm{c}}) \leqslant Cn^{-C\varepsilon^2} \tag{42}$$

and

$$P(B_j^c|A_j) \leqslant Cn^{-C} + C(|I_j|/m)(L/m)n^{-C\delta_n^2}. \tag{43}$$

Consider inequality (42) first. By the construction of $\mathbb{I}^{(1)}$,

$$P(A_j^c) = P[\max_{\tilde{I}\in\mathbb{J}_T(L)}\{X(\tilde{I})\} \leqslant \lambda_n^*, \text{ only } I_j \text{ exists}] \leqslant P\{X(\tilde{I}_j) \leqslant \lambda_n^*, \text{ only } I_j \text{ exists}\},$$

where $\tilde{I}_j$ is defined in equation (33). By inequality (35), the fact that λ_n^* converges to λ_n at the order of $n^\gamma/\sqrt{\{2\log(n)\}}$, and some routine calculation, inequality (42) follows.

Now consider inequality (43). Define

$$\mathbb{K}_T(L) = \{\tilde{I} \in \mathbb{I}^{(1)} : D(\tilde{I}, \tilde{I}_j) > C\delta_n\}$$

for some $C > 0$. Since, for an interval I, $D(I, I_j) > \delta_n$ implies $D(\tilde{I}, \tilde{I}_j) > C\delta_n$, where $\tilde{I}$ is the collection of the index of J_k such that $J_k \subseteq I$, then B_j^c implies $I_1^* \in \mathbb{K}_T(L)$, where I_1^* is defined in step 3 of the RSI algorithm. This further implies the existence of some $\tilde{I} \in \mathbb{K}_T(L)$ such that $X(\tilde{I}) \geqslant X(\tilde{I}_j)$. Denote

$$\mathbb{K}_0 = \{\tilde{I} \in \mathbb{K}_T(L) : \tilde{I} \cap \tilde{I}_j = \emptyset\},$$
$$\mathbb{K}_1 = \{\tilde{I} \in \mathbb{K}_T(L) : \tilde{I} \cap \tilde{I}_j \neq \emptyset\}.$$

We have

$$\begin{aligned} P(B_j^c|A_j) &\leqslant P\{\exists \tilde{I} \in \mathbb{K}_0 : X(\tilde{I}) \geqslant X(\tilde{I}_j), |\mathbb{K}_0| \leqslant CL/m + \log(n)\} + P\{|\mathbb{K}_0| > CL/m + \log(n)\} \\ &\quad + P\{\exists \tilde{I} \in \mathbb{K}_1 : X(\tilde{I}) \geqslant X(\tilde{I}_j)\} \\ &\leqslant \{CL/m + \log(n)\}\, P\{X(\tilde{I}) - X(\tilde{I}_j) \geqslant 0, \tilde{I} \in \mathbb{K}_0\} + P\{|\mathbb{K}_0| > CL/m + \log(n)\} \\ &\quad + (|I_j|/m)\,(L/m)\, P\{X(\tilde{I}) - X(\tilde{I}_j) \geqslant 0, \tilde{I} \in \mathbb{K}_1\}. \end{aligned}$$

Since $|\mathbb{K}_0| = \Sigma_{\tilde{I}\in\mathbb{J}_T(L):\tilde{I}\cap\tilde{I}_j=\emptyset} \mathbf{1}\{X(\tilde{I}) > \lambda_n^*\}$, which converges to $\Sigma_{\tilde{I}\in\mathbb{J}_T(L):\tilde{I}\cap\tilde{I}_j=\emptyset} P\{X(\tilde{I}) > \lambda_n^*\}$ exponentially fast, and $\Sigma_{\tilde{I}\in\mathbb{J}_T(L):\tilde{I}\cap\tilde{I}_j=\emptyset} P\{X(\tilde{I}) > \lambda_n^*\} \leqslant CT(L/m)\, P[Z(\tilde{I}) > \sqrt{\{2\log(T)\}}] \leqslant CL/m$, then we have

$$P\{|\mathbb{K}_0| > CL/m + \log(n)\} \leqslant Cn^{-C}.$$

It is left to show that

$$P\{X(\tilde{I}) - X(\tilde{I}_j) \geqslant 0, \tilde{I} \in \mathbb{K}_0\} \leqslant Cn^{-C} \tag{44}$$

and

$$P\{X(\tilde{I}) - X(\tilde{I}_j) \geqslant 0, \tilde{I} \in \mathbb{K}_1\} \leqslant Cn^{-C\delta_n^2}. \tag{45}$$

For inequality (44), since $\tilde{I} \cap \tilde{I}_j = \emptyset$, then

$$X(\tilde{I}) \leqslant \mu_j + \frac{Z(\tilde{I})}{2h(0)\sqrt{m}} + \frac{\zeta(\tilde{I})}{\sqrt{m}},$$

where the first term on the right-hand side shows up because there is at most one position in $\tilde{I}$ that can possibly have mean μ_j, and other positions have mean 0. In contrast,

$$X(\tilde{I}_j) = \mu_j\sqrt{|\tilde{I}_j|} + \frac{Z(\tilde{I}_j)}{2h(0)\sqrt{m}} + \frac{\zeta(\tilde{I}_j)}{\sqrt{m}}.$$

So

$$\begin{aligned} P\{X(\tilde{I}) - X(\tilde{I}_j) \geqslant 0, \tilde{I} \in \mathbb{K}_0\} &\leqslant P\left\{\frac{Z(\tilde{I})}{2h(0)\sqrt{m}} - \frac{Z(\tilde{I}_j)}{2h(0)\sqrt{m}} \geqslant \mu_j\sqrt{|\tilde{I}_j|} - \mu_j + \frac{\zeta(\tilde{I}_j)}{\sqrt{m}} - \frac{\zeta(\tilde{I})}{\sqrt{m}}\right\} \\ &\leqslant P[Z(\tilde{I}) - Z(\tilde{I}_j) \geqslant \sqrt{\{2(1+\varepsilon/2)\log(n)\}} - \sqrt{\log(n)/m}] + Cn^{-C} \\ &\leqslant P[N(0,2) \geqslant \sqrt{\{2(1+C)\log(n)\}}] + Cn^{-C}, \end{aligned}$$

where the second inequality is because

$$\mu_j\sqrt{|\tilde{I}_j|}-\mu_j \geqslant \mu_j\sqrt{|\tilde{I}_j|}\sqrt{\left\{1-\frac{\varepsilon}{2(1+\varepsilon)}\right\}}$$

by conditions (21) and (15) and the choice of m, and

$$P\left\{\zeta(\tilde{I}_j)-\zeta(\tilde{I}) < -\frac{\sqrt{\log(n**)}}{m}\right\} \leqslant Cn^{-C} \tag{46}$$

by inequality (31). Therefore, inequality (44) follows.

For inequality (45), since $\tilde{I}\cap\tilde{I}_j \neq \emptyset$, we can write

$$X(\tilde{I})-X(I)=\mathrm{LR}_1+\mathrm{LR}_2+\mathrm{LR}_3,$$

$$\mathrm{LR}_1=\left(\frac{1}{\sqrt{|\tilde{I}|}}-\frac{1}{\sqrt{|\tilde{I}_j|}}\right)\sum_{k\in\tilde{I}\cap\tilde{I}_j}X_k=\left(\frac{1}{\sqrt{|\tilde{I}|}}-\frac{1}{\sqrt{|\tilde{I}_j|}}\right)\left\{\mu_j|\tilde{I}\cap\tilde{I}_j|+\frac{\sum_{k\in\tilde{I}\cap\tilde{I}_j}Z_k}{2h(0)\sqrt{m}}+\frac{\sum_{k\in\tilde{I}\cap\tilde{I}_j}\zeta_k}{\sqrt{m}}\right\},$$

$$\mathrm{LR}_2=\frac{1}{\sqrt{|\tilde{I}|}}\sum_{k\in\tilde{I}\setminus\tilde{I}\cap\tilde{I}_j}X_k\leqslant\frac{1}{\sqrt{|\tilde{I}|}}\left\{\mu_j+\frac{\sum_{k\in\tilde{I}\setminus\tilde{I}\cap\tilde{I}_j}Z_k}{2h(0)\sqrt{m}}+\frac{\sum_{k\in\tilde{I}\setminus\tilde{I}\cap\tilde{I}_j}\zeta_k}{\sqrt{m}}\right\}$$

$$\mathrm{LR}_3=-\frac{1}{\sqrt{|\tilde{I}_j|}}\sum_{k\in\tilde{I}_j\setminus\tilde{I}\cap\tilde{I}_j}X_k=\frac{1}{\sqrt{|\tilde{I}_j|}}\left\{-\mu_j|\tilde{I}_j\setminus\tilde{I}\cap\tilde{I}_j|-\frac{\sum_{k\in\tilde{I}_j\setminus\tilde{I}\cap\tilde{I}_j}Z_k}{2h(0)\sqrt{m}}-\frac{\sum_{k\in\tilde{I}_j\setminus\tilde{I}\cap\tilde{I}_j}\zeta_k}{\sqrt{m}}\right\}$$

Note that LR_1, LR_2 and LR_3 are independent, and

$$\left(\frac{1}{\sqrt{|\tilde{I}|}}-\frac{1}{\sqrt{|\tilde{I}_j|}}\right)\frac{\sum_{k\in\tilde{I}\cap\tilde{I}_j}Z_k}{2h(0)\sqrt{m}}+\frac{1}{\sqrt{|\tilde{I}|}}\frac{\sum_{k\in\tilde{I}\setminus\tilde{I}\cap I}Z_k}{2h(0)\sqrt{m}}-\frac{1}{\sqrt{|\tilde{I}_j|}}\frac{\sum_{k\in\tilde{I}_j\setminus\tilde{I}\cap\tilde{I}_j}Z_k}{2h(0)\sqrt{m}}\sim\frac{1}{\sqrt{m}}N(0,\tau)$$

for some $\tau>0$. In contrast, $D(\tilde{I},\tilde{I}_j)\geqslant C\delta_n$ implies that

$$\left(\frac{|\tilde{I}\cap\tilde{I}_j|}{\sqrt{|\tilde{I}|}}-\sqrt{|\tilde{I}_j|}\right)\mu_j<-C\delta_n\sqrt{|\tilde{I}_j|}\mu_j.$$

Combining this with inequality (46) we have

$$P\{X(\tilde{I})-X(\tilde{I}_j)\geqslant 0,\tilde{I}\in\mathbb{K}_1\}\leqslant P\left\{N(0,\tau)\geqslant C\delta_n\mu_j\sqrt{|I_j|}-\frac{\sqrt{\log(n)}}{m}\right\}+Cn^{-C}.$$

Given expression (15) and the choice of m, inequality (45) follows for δ_n satisfying $\sqrt{\{\log(q)+\log(\bar{s}/m)+\log(L/m)\}}/\sqrt{\log(n)}\ll\delta_n\ll 1$.

References

Abyzov, A., Urban, A., Snyder, M. and Gerstein, M. (2011) CNVnator: an approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.

Alkan, C., Coe, B. and Eichler, E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–375.

Arias-Castro, E., Donoho, D. and Huo, X. (2005) Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theor.*, **51**, 2402–2425.

Bravo, H. and Irizarry, R. (2010) Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics*, **66**, 665–674.

Brown, L. D., Cai, T. T. and Zhou, H. H. (2008) Robust nonparametric estimation via wavelet median regression. *Ann. Statist.*, **36**, 2055–2084.

Cai, T. T., Jeng, X. J. and Jin, J. (2011) Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Statist. Soc.* B, **73**, 629–662.

Cai, T. T. and Zhou, H. H. (2009) Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann. Statist.*, **37**, 3204–3235.

Chen, K., Wallis, J., McLellan, M., Larson, D., Kalick, J., Pohl, C., McGrath, S., Wendl, M., Zhang, Q., Locke, D., Sho, X., Fulton, R., Ley, T., Ding, L. and Mardis, E. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Meth.*, **6**, 677–681.

Cheung, M., Down, T., Latorre, I. and Ahringer, J. (2011) Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.*, to be published, doi:10.1093/nar/gkr425.

Chiang, D., Getz, G., Jaffe, D., O'Kelly, M., Zhao, X., Carter, S., Russ, C., Nusbaum, C., Meyerson, M. and Lander, E. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Meth.*, **6**, 99–103.

Diskin, S. J., Hou, C., Glessner, J. T., Attiyeh, E. F., Laudenslager, M., Bosse, K., Cole, K., Moss, Y., Wood, A., Lynch, J. E., Pecor, K., Diamond, M., Winter, C., Wang, K., Kim, C., Geiger, E. A., McGrady, P. W., Blakemore, A. I. F., London, W. B., Shaikh, T. H., Bradfield, J., Grant, S. F. A., Li, H., Devoto, M., Rappaport, E. R., Hakonarson, H. and Maris, J. M. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, **459**, 987–991.

Feuk, L., Carson, A. and Scherer, S. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Ivakhno, S., Royce, T., Cox, A., Evers, D., Cheetham, R. and Tavare, S. (2010) CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*, **26**, 3051–3058.

Jeng, J. J., Cai, T. T. and Li, H. (2010) Optimal sparse segment identification with application in copy number variation analysis. *J. Am. Statist. Ass.*, **105**, 1156–1166.

Kim, T., Luquette, L., Xi, R. and Park, J. (2010) rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinform.*, **11**, article 432.

Li, J., Jiang, H. and Wong, W. (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.*, **11**, article R50.

Li, Y., Zheng, H., Lou, R., Wu, H., Zhu, H., Li, R., Cao, H., Wu, B., Huang, S., Shao, H., Ma, H., Zhang, F., Feng, S., Zhang, W., Du, H., Tian, G., Li, J., Zhang, X., Li, S., Bolund, L., Kristiansen, K., de Smith, A., Blakemore, A., Coin, L., Yang, H. and Wang, J. (2011) Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat. Biotech.*, **29**, 723–730.

McCarroll, S. S. and Altshuler, D. M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, suppl., S37–S42.

Medvedev, P., Stanciu, M. and Brudno, M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Meth.*, **6**, S13–S20.

Miller, C. A., Hampton, O., Coarfa, C. and Milosavljevic, A. (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLOS ONE*, **6**, article e16327.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Sitz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A. and Korbel, J. O. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Nord, A., Lee, M., King, M. and Walsh, T. (2011) Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC Genom.*, **12**, article 184.

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

Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**, 496–501.

Redon, R., Ishikawa, S., Fitch, K., Feuk, L., Perry, G., Andrews, T., Fiegler, H., Shapero, M., Carson, A., Chen, W., Cho, E., Dallaire, S., Freeman, J., Gonzalez, J., Gratacs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J., Marshall, C., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D., Estivill, X., Tyler-Smith, C., Carter, N., Aburatani, H., Lee, C., Jones, K., Scherer, S. and Hurles, M. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T., Trask, B., Patterson, N., Zetterberg, A. and Wigler, M. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotech.*, **26**, 1135–1145.

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O., Ingason, A., Steinberg, A., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J. *et al.* (2008) Large recurrent microdeletions associated with schizophrenia. *Nature*, **455**, 178–179.

Stone, J., O'Donovan, M., Gurling, H., Kirov, G., Blackwood, D., Corvin, A., Craddock, N., Gill, M., Hultman, C., Lichtenstein, P. *et al.* (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, **455**, 237–241.

Urban, A., Korbel, J., Selzer, R., Richmond, T., Hacker, A., Popescu, G., Cubells, J., Green, R., Emanuel, B., Gerstein, M., Weissman, S. and Snyder, M. (2006) High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc. Natn. Acad. Sci. USA*, **103**, 4534–4539.

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, A. S., Kusenda, M., Malhotra, D., Bhandari, A., Stray, S. M., Rippey, C. F., Roccanova, P., Makarov, V., Lakshmi, B., Findley, R. L., Sikich, L., Stromberg, T., Merriman, B., Gogtay, N., Butler, P., Eckstrand, K., Noory, L., Gochman, P., Long, R., Chen, Z., Davis, S., Baker, C., Eichler, E. E., Meltzer, P. S., Nelson, S. F., Singleton, A. B., Lee, M. K., Rapoport, J. L., King, M.-C. and Sebat, J. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, **320**, 539–543.

Walther, G. (2010) Optimal and fast detection of spacial clusters with scan statistics. *Ann. Statist.*, **38**, 1010–1033.

Xie, C. and Tammi, M. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinform.*, **10**, article 80.

Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1568–1592.

Zhang, F., Gu, W., Hurles, M. and Lupski, J. (2009) Copy number variation in human health, disease and evolutions. *A. Rev. Genom. Hum. Genet.*, **10**, 451–481.

Zhou, H. (2006) A note on quantile coupling inequalities and their applications. *Technical Report*. Department of Statistics, Yale University, New Haven.