Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2012 November ; 74(5): 773–797. doi:10.1111/j.1467-9868.2012.01028.x.

Robust Detection and Identification of Sparse Segments in Ultra-High Dimensional Data Analysis

T. Tony Cai
Department of Statistics, University of Pennsylvania, Philadelphia, USA

X. Jessie Jeng[†]
Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, USA

Hongzhe Li
Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, USA

Summary

Copy number variants (CNVs) are alternations of DNA of a genome that results in the cell having a 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 different noise distributions. We present results from a CNV analysis of a HapMap Yoruban sample to further illustrate the theory and the methods.

Keywords

Robust segment detector; Robust segment identifier; optimality; DNA copy number variant; next generation sequencing data

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 DNA of a genome that results in the cell having a 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 tumor genomes; most are due to somatic

[†]*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 xjeng@upenn.edu.

mutations that occur during the clonal development of the tumor. The copy number changes in tumor genomes are often referred to as copy number aberrations (CNAs). In this paper, we focus on the CNVs from the germline 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 (SNP) 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 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/deletions (indels) 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 in order to identify gaps and segmental rearrangements in pair-wise alignments (Li et al., 2011). However, this approach requires data from two genomes.

Another important source of information useful for inferring CNVs from reads alignment is the read depth. 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, one expects to observe a smaller number of read counts or lower mapping density than the background read depth. In contrast, when the genomic location or bin is within an insertion or duplication, one expects 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 read-depth data provide more reliable information for large CNVs and CNVs flanked by repeats, where accurate mapping reads is difficult. The read depth 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 read depth data from the next generation sequencing. Several methods have been developed for such read depth data. Yoon et al. (2009) developed an algorithm for read depth data to detect CNVs, where they convert the read count of a window into a Z-score by subtracting the mean of all windows and dividing by the standard deviation and identify the CNVs by computing upper-tail and lower-tail probabilities by using a normality assumption on the RD data. The windows are then selected based on the extreme values of these probabilities controlling for the genome-wide false-positive rates. Abyzov et al. (2011) developed an approach to first partition the genome into a set of regions with different underlying copy numbers using mean-shift technique and then 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 read depth data when pairs of samples are available. The basic idea underlying these two methods is to convert the counts data into ratios and then apply existing copy number analysis methods developed for array CGH 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 due to the complex process of sequencing. Some recent literature assumes a constant read sampling rate across the genome and Poisson distribution or negative-binomial distribution for the read counts data (Xie and Tammi, 2009; Cheung et al., 2011). However, due to GC 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 the advances of sequencing technologies. To analyze 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. Misspecified data distribution can lead to a complete failure of these methods. Although some data distributions can be estimated by nonparametric methods, popular nonparametric methods such as permutation are often computationally expensive and not feasible for ultra-high 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 modeled as

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

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 *i.i.d.* random errors with median 0. Here positive μ_i implies duplication or insertion and 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 (GC) content-adjusted RD counts at genomic location or bin i, which can be regarded as continuous when coverage of the genome is sufficiently high, for example 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

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

and the second step is to identify the locations of the segments if there are any. A procedure called the 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, the 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, the 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 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 Cauchy (0, 1), and set the signal segment at [457 : 556] with a positive mean. Figure 1 compares the LRS with the RSI. In this example, the LRS fails to work at all by identifying too many false segments, while the RSI, on the other hand, 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 Cai and Zhou (2009) used local medians to turn the problem of nonparametric 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) shows 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 $\mu_j \sqrt{|I_j|} > \sqrt{\log n} / \left(\sqrt{2} h(0) \right)$ 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 robust segment identifier (RSI), which first transforms the data by binning and taking local medians, 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 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 ultra-high 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 robust segment detector (RSD). The robust segment identification procedure RSI is proposed and its theoretical properties are studied in Section 3. Numerical performance of RSI is investigated in Section 4 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 the Appendix.

2. Data Transformation and Robust Detection

In this section, we first introduce the local median transformation to tackle the problem of 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 reliably separate H_0 from H_1 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 there are q sparse and short signal segments in the observed data and the number of observations n is very large. The goal is to detect and identify these q segments. In order to do so, we first equally divide the n observations into $T = T_n$ groups with $m = m_n$ observations in each group. De ne the set of indices in the k-th group as $J_k = \{i : (k-1)m + 1 \le i \le km\}$, and generate the transformed dataset as

$$X_k = \text{median}\{Y_i : i \in J_k\}, \qquad 1 \le k \le T. \quad (2)$$

Set

$$\eta_k = \text{median}\{\xi_i : i \in J_k\}, \qquad 1 \le k \le T, \quad (3)$$

then the medians X_k can be written as

$$X_k = \theta_k + \eta_k, \qquad 1 \le k \le T, \quad (4)$$

where

$$\theta_k = \begin{cases} \mu_j, & J_k \subseteq i_j \text{ for some } I_j, \\ \mu_k^* \in \left[0, \mu_j\right], & J_k \cap I_j \neq \varnothing \text{ for some } I_j \text{ and } J_k \nsubseteq I_j, \\ 0, & \text{otherwise.} \end{cases}$$

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 variable for a wide range of noise distributions. Specifically, we assume that the distribution of ξ_i is symmetric about 0 with the density function h satisfying $h(0) > 0$ and

$$|h(y) - h(0)| \le C_{y^2} \quad (5)$$

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

Lemma 2.1. Assume (1), (5), and transformation (4), then η_k can be written as

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

where $Z_k \overset{iid}{\sim} N(0, 1)$ 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

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

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

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 $Z_k / \left(2h(0)\sqrt{m}\right)$, which follows $N(0, 1/(4h^2(0)m))$, so that after the data transformation in (4), existing methods for Gaussian noise can be applied to X_k, $1 \le k \le T$. It will be shown that by properly choosing the bin size m, a robust procedure can be constructed to reliably detect the signal segments. 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 $h(0)$ to be 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 $H_0: \mathbb{I} = \varnothing$ against $H_1: \mathbb{I} = \varnothing$. 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 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 (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 with their absolute values.

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

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

and threshold

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

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

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

$$X\left(\tilde{I}\right)=\frac{Z\left(\tilde{I}\right)}{2h\left(0\right)\sqrt{m}}+\frac{\zeta\left(\tilde{I}\right)}{\sqrt{m}} \text{ under } H_0, \quad (11)$$

were

$$Z\left(\tilde{I}\right)=\sum_{k\in\tilde{I}}Z_k/\sqrt{|\tilde{I}|} \qquad \text{and} \qquad \zeta\left(\tilde{I}\right)=\sum_{k\in\tilde{I}}\zeta_k/\sqrt{|\tilde{I}|}.$$

Since ζk are stochastically small random variables according to Lemma 2.1, $\max_{\tilde{I}\in\mathbb{J}_T}\zeta\left(\tilde{I}\right)$ should be much smaller than $\max_{\tilde{I}\in\mathbb{J}_T}Z\left(\tilde{I}\right)$ for large m. The following lemma provides the asymptotic bounds for both $\max_{\tilde{I}\in\mathbb{J}_T}\zeta\left(\tilde{I}\right)$ and $\max_{\tilde{I}\in\mathbb{J}_T}Z\left(\tilde{I}\right)$.

Lemma 2.2. Assume (1), (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, …, T}, we have

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

for some a > 4 and C > and

$$P\left(\max_{\tilde{I}\in\mathbb{J}_T}Z\left(\tilde{I}\right)>\sqrt{2\log T}\right)\leq\frac{C}{\sqrt{\log T}}. \quad (13)$$

Lemma 2.2 and (11) imply that the threshold on $\max_{\tilde{I}\in\mathbb{J}_T}X\left(\tilde{I}\right)$ should be approximately that of $\max_{\tilde{I}\in\mathbb{J}_T}Z\left(\tilde{I}\right)/2h\left(0\right)\sqrt{m}$, Which is log $\sqrt{2\log T}/\left(2h\left(0\right)\sqrt{m}\right)$ We set the threshold slightly more conservatively. The following theorem shows the control of family-wise type I error.

Theorem 2.1. Assume (1), (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, …, T},

$$P_{H_0}\left(\max_{\tilde{I}\in\mathbb{J}_T}X\left(\tilde{I}\right)>\lambda_n\right)\leq\frac{C}{\sqrt{\text{lot }T}}\to 0, \qquad T\to\infty.$$

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

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

Theorem 2.2. Assume (1), (5), and transformation (4) with m = log^{1+b} n for some b > 0. If there exists some segment $I_j \in \mathbb{I}$ that satisfies

$$|I_j|/m \to \infty. \quad (14)$$

and

$$\mu_j \sqrt{|I_j|} \geq \sqrt{2(1+\epsilon)\log n}/(2h(0)) \quad (15)$$

for some $\varepsilon > 0$, then 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 $|I_j|$. Note that 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 RSD cannot have asymptotically vanishing type I and type II errors. In fact, in this situation, no testing procedure works.

Theorem 2.3. Assume (1), (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 $I_j \in I_j \in \mathbb{I}$,

$$\log|I_j| = o(\log n), \quad (16)$$

$$\mu_j \sqrt{|I_j|} \leq \sqrt{2(1-\epsilon)\log n}/(2h(0)) \quad (17)$$

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 Theorem 2.2 and 2.3 imply that RSD is an near-optimal procedure to detect short signal segments based on $X_1, \ldots, X_T$. It can successfully separate H_0 and H_1 based on the transformed data whenever there exists some testing procedure that is able to do so.

The RSD is robust over a wide range of noise distributions when assumption (5) is satisfied. On the other hand, in the special case when the Gaussian assumption does hold for the noise distribution, the GLRT procedure specifically designed for this case is more efficient. The GLRT rejects H_0 if $\max_{\tilde{I} \in \mathbb{I}_n} Y(\tilde{I}) > \sqrt{2\log n}$, where $\mathbb{I}_n$ 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 2.1. Assume (1) and $\xi_i \sim N(0, 1)$. If there exists some segment $I_j \in \mathbb{I}$ that satisfies

$$\mu_j \sqrt{|I_j|} \geq \sqrt{2(1+\epsilon)\log n} \quad (18)$$

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. On the other hand, if $log\, q = o(log\, n)$, and for all segments $I_j \in \mathbb{I}$,

$$\log|I_j| = o(\log n), \quad (19)$$

$$\mu_j \sqrt{|I_j|} \leq \sqrt{2(1-\epsilon)\log n}, \quad (20)$$

for some $\epsilon > 0$, 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 2.1 with Theorem 2.2 and 2.3, we can see the exact power loss due to the local median transformation if noise distribution is known to be Gaussian. Note that 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 $\sqrt{2\log n}/(2h(0)) \approx 1.25 \times \sqrt{2\log n}$, 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, while 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 $\alpha_i = \mu_j 1_{\{i \in I_j\}}$ 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

$$\underline{s} \geq \log^2 n \quad \text{and} \quad \log \bar{s} = o(\log n), \quad (21)$$

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

$$\log q = o(\log n), \quad (22)$$

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

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

The ideal threshold for RSI should be the same as that of RSD, which is λ_n defined in (10). However, $h(0)$ is usually unknown in practice and needs to be estimated. By Lemma 2.1, $1/(2h(0)\sqrt{m})$ 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 (MAD) estimator:

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

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

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

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

Step 1: Transform data by (4). Let $\mathbb{J}_T(L)$ be the collection of all possible subintervals in $\{1,\ldots,T\}$ with interval length $\le L/m$.

Step 2: Let $j=1$. Define $\mathbb{I}^{(j)}=\left\{\tilde{I}\in\mathbb{J}_T(L):X\left(\tilde{I}\right)>\lambda_n^*\right\}$, where $X(\tilde{I})$ and λ_n^* are defined as in (9) and (24).

Step 3: Let $I_j^*=\arg\max_{\tilde{I}\in\mathbb{I}^{(j)}}X\left(\tilde{I}\right)$ and update $\mathbb{I}^{(j+1)}=\mathbb{I}^{(j)}\setminus\left\{\tilde{I}\in\mathbb{I}^{(j)}:\tilde{I}\cap I_j^*\neq\varnothing\right\}$.

Step 4: Repeat Step 3–4 with $j=j+1$ until $\mathbb{I}^{(j)}$ is empty.

Step 5: For each I_j^* generated above, let $l_j=\left(\text{first element in } I_j^*-1\right)\times m+1$ and $r_j=\text{last element in } I_j^*\times m$, and denote $\widehat{I}_j=\left\{l_j,\ldots,r_j\right\}$.

Denote the collection of selected intervals as $\widehat{\mathbb{I}}=\left\{\widehat{I}_1,\widehat{I}_2,\ldots\right\}$. If $\widehat{\mathbb{I}}\neq\varnothing$, we reject the null hypothesis and identify the signal segments by all the elements in $\widehat{\mathbb{I}}$. 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 step 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 strong enough. Define the dissimilarity between any pair of $\widehat{I}\in\widehat{\mathbb{I}}$ and $I\in\mathbb{I}$ as

$$D\left(\widehat{I},I\right)=1-|\widehat{I}\cap I|/\sqrt{|\widehat{I}||I|}. \quad (25)$$

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

Theorem 3.1. Assume (1), (5), and transformation (4) with $m=\log^{1+b}n$ for $0<b<1$. Suppose that $\widehat{\sigma}$ is an n^{γ}-consistent estimator of the standard deviation of η_k for some $\gamma>0$, and L satisfies

$$\bar{s}\le L<\underline{d}, \quad \text{and} \quad \log L=o(\log n). \quad (26)$$

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

$$P_{H_0}\left(|\widehat{\mathbb{I}}|>0\right)+P_{H_1}\left(\max_{I_j\in\mathbb{I}}\min_{\widehat{I}_j\in\widehat{\mathbb{I}}}D\left(\widehat{I}_j,I_j\right)>\delta_n\right)\to 0 \quad (27)$$

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

The asymptotic result (27) essentially says that both the probability of having at least one false positive 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 $\left[\bar{s}, \underline{d}\right)$ is very large when signal segments are relatively short and rare as in the applications we are interested. The second part $\log L = o(\log T)$ is satisfied, if, for instance, $L = \log^a T$ for $a \geq 0$. 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 2.3, 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 $I_j \in \mathbb{I}$, $\mu_j\sqrt{|I_j|} \leq \sqrt{2(1-\epsilon)\log n}/(2h(0))$. This implies that the RSI is an 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 based on the transformed data.

4. Simulation Studies

We evaluate the finite-sample behavior of the RSI through simulation studies and compare its performance with the performances of LRS and another popular procedure, circular binary segmentation (CBS) (Olshen et al., 2004).

4.1. Performance of RSI under different 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 get thinner, and the t-distribution approaches to the standard normal. Nevertheless, the t-distributions satisfy the general assumption in (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 $|I_3| = 20$, respectively, and the signal mean for all segments at $\mu = 1.0$, 1.5, and 2.0, respectively. 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 $\lambda_n=\widehat{\sigma}\sqrt{2\log T}$, where $\widehat{\sigma}$ is calculated as in (23). Figure 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 is close to be normally distributed. In this case, $m = 20$ is large enough to stabilize the noise. More discussions on the choice of m are given in Section 4.3.

As used in Jeng et al. (2010), the identification accuracy of 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 over-selections. In detail, for each signal segment I_j, define

$$D_j = \min_{\widehat{I} \in \widehat{\mathbb{I}}} D\left(\widehat{I}, I_j\right),$$

where $D\left(\widehat{I}, I_j\right)$ is defined in (25). Obviously, smaller D_j represents better matching between I_J and some estimate $\widehat{I} \in \widehat{\mathbb{I}}$, and $D_j = 0$ iff $I_j = \widehat{\mathbb{I}}$. Define

$$\#O = \#\left\{\widehat{I} \in \widehat{\mathbb{I}} : \widehat{I} \cap I_j = \varnothing, \forall_j = 1, \ldots, q\right\}.$$

So $\#O$ is a non-negative integer, and $\#O = 0$ if there are no over-selected intervals. Note that according to Theorem 3.1, $\mu_j \sqrt{|I_j|}$ should be at least $\sqrt{\log n} / \left(\sqrt{2} h(0)\right) \approx 7.307$ for segment I_j to be consistently estimated by 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, 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 estimation for any of the signal segments when μ is large enough. In all the cases, the over-selection 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. A comparison with LRS and CBS

In the second set of simulations, we compare the performance of RSI with that of the original LRS and CBS under different 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 $t(3)$, the LRS has lower power and a large number of over-selections, while RSI remains very stable. In contrast, CBS fails to select any true intervals. When noise distribution is $t(30)$, which is very close to Gaussian, both LRS and CBS outperform RSI with better power and better identification of the signal segments. However, RSI still performs reasonably well and has no over-selection. This agrees with our theoretical results in Section 2.2.

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 of 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 and 1.5. Note that the noise variances are constant within an interval but different for different intervals. The bottom half of Table 2 shows the comparison of the three procedures. RSI still has the best overall performance. It results in smaller numbers of over-selections than LRS and better power than CBS. However, as the noise variance increases, RSI can result in more over-selections.

The 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. On the other hand, 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 evaluate 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 over-selection when m varies. Smaller m results in better power but more over-selections, especially when the noise distribution has heavy tail. The greater power is due to finer binning, which preserves the signal information better. On the other hand, 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 accurate enough to overcome the effect of the heavy-tailed noise on segment selection, which in term leads to more over-selections in the case of $t(1)$ and $m = 10$.

5. Application to Identification of CNVs Based on the NGS Data

To demonstrate our proposed methods, we analyze the short reads data on chromosome 19 of a HapMap Yoruban female sample (NA19240) from the 1000 Genomes Project. Let Y_i denote the GC 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 are mapping to the reference human DNA sequences, we obtain the RD data at $n = 54, 361, 060$ genomic locations. Figure 3 (a) shows the read depth for the first 10, 000 observations of the data set. The median of the count number over all n sites is 30.

The statistical challenges for CNV detection based on NGS data include both ultra-high dimensionality of the data that requires fast computation and unknown distribution of the read depths 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 modeling these RD data.

We apply the proposed RSI with $m = 400$ and $L = 150$, which assumes that the maximum CNV based on our pre-processed data is $400 \times 150 = 60, 000$ base pairs (bps). This is sensible since typical CNVs include multi-kilobase deletions and duplications (McCarroll and Altshuler, 2007; Medvedev et al., 2009). We chose L=150 partially due to computational consideration. 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 75,991 bps in sizes, among these 24 are contiguous segments. After merging these contiguous segments, we obtained 101 CNVs, ranging from 400 to 125,440 bps in sizes with a median size of 38,860 bps. See Figure 3 (b) for the distributions of the sizes of CNVs identified. There are 8 CNVs of size of 400 bps, 4 CNVs of size of 800 bps and 4 CNVs of size of 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 identified by the RSI, Figure 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 than their neighboring regions. Examinations of the other CNV regions identified also show that these

regions contain more or fewer reads than their neighboring regions, further indicating the effectiveness of the RSI procedure in identifying the CNVs.

Mills et al. (2011) recently reported a map of copy number variants 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 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, respectively, based on 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 reported in the CNV map of (Mills et al., 2011), indicating high sensitivity of our method in detecting the CNVs based on the RD data since our CNVs calls are based on data from only a single sample (NA19240). We are not able to make a direct comparison on CNV calls for this particular sample since the sample-level CNV calls are not available from the publication of Mills et al. (2011).

6. Conclusion and Further Discussion

Motivated by CNV analysis based on the read depth data generated by the NGS technology, we have studied the problem of segment detection and identification from an ultra long linear sequence of data with an unknown noise distribution. We have developed a robust detection procedure RSD and a 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 read depths when compared to the neighboring 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 save a large amount of computational cost by greatly reducing the data dimension, which makes it particularly appealing for analyzing the ultra-high dimensional NGS data. As more and 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 noises are *i.i.d*, which can be violated for the RD data. The noise distribution of the RD data is directly related to the uncertainty and errors 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 *i.i.d* assumption for the RD data. While such errors can have great impact on analysis of single nucleotide polymorphisms and rare genetic variants, their impact on RD data distribution and CNV identification is not clear. Our simulations (Table 2) showed that when the noise has heterogeneous variances, RSI can still identify the correct CNVs unless the variance is very large.

The methods 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

$$\int_{-\infty}^{0} h(y) = 1/2, \quad h(0) > 0, \quad \text{and } h(y) \text{ is Lipschits at } y=0.$$

Obviously, this assumption is more general than (5) (Brown et al., 2008). To accommodate this more general assumption, a larger m is needed for the robust methods 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 among multiple individuals based on the NGS data and test for CNV associations. Our methods provide the basic tools for these extensions in order to consider structured signals with unknown and possibly heavy-tailed noise distributions.

Acknowledgments

Jeng and Li were supported by NIH grants ES009911 and CA127334. The research of Tony Cai was supported in part by NSF FRG Grant DMS-0854973. We thank Dr. Mingyao Li for providing the sequences data from the 1000 Genomes Project.

Appendix: Proofs

In this appendix we present the proofs for Lemma 2.2, Theorem 2.1, 2.2, 2.3, and 3.1.

Proof of Lemma 2.2

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

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

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

Let $A_k = m\zeta_{k1}$, then by (7) in Lemma 2.1, $EA_k = 0$ and $E|A_k|^l \le C_l$ for any $l > 0$. According to Lemma 2 in Zhou (2006), there exists some positive constant ϵ' such that for any $0 \le x \le \epsilon'$ and interval $\tilde{I}$,

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

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

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

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

$$P(|\zeta_2(\tilde{I})| > x/2) \le TP\left(|\zeta_{k2}| > x/2\sqrt{|\tilde{I}|}\right) \le CT \quad \exp(-am) \le Cn^{-C} \quad \log^b n \quad (30)$$

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

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

and consequently

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

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

Proof of Theorem 2.1

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

$$P_{H_0}\left(\max_{\tilde{I}\in\mathbb{J}_T}X\left(\tilde{I}\right) \quad >\lambda_n\right)=P_{H_0}\left(\max_{\tilde{I}\in\mathbb{J}_T}X\left(\tilde{I}\right) \quad >\lambda_n, \max_{\tilde{I}\in\mathbb{J}_T} \leq 5\sqrt{\log \mathrm{T}/\mathrm{m}}\right)+P_{H_0}\left(\max_{\tilde{I}\in\mathbb{J}_T}X\left(\tilde{I}\right) \quad >\lambda_n, \max_{\tilde{I}\in\mathbb{J}_T}>5\sqrt{\log \mathrm{T}/\mathrm{m}}\right)$$

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

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

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

Proof of Theorem 2.2

Since Theorem 2.1 implies that the type I error of RSD goes to 0, all we need to show is

$$P_{H1}\left(\max_{\tilde{I}\in\mathbb{J}_T}X\left(\tilde{I}\right) \leq \lambda_n\right) \to 0 \quad (32)$$

Suppose that under H_1 segment $I_j \in \mathbb{I}$ satisfies (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.

$$\tilde{I}_j=\left\{\kappa:J_\kappa \subseteq I_j\right\}, \quad (33)$$

then

$$|\tilde{I}_j| \leq \left\lfloor |I_j|/m \quad - \quad 1\right\rfloor > |I_j|/m \quad - \quad 2, \quad (34)$$

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

$$X_\kappa = \mu_j + \frac{Z_\kappa}{2h(0)\sqrt{m}} + \frac{\zeta_\kappa}{\sqrt{m}}, \quad \kappa \in \tilde{I}_j,$$

which further implies that

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

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

$$\mu\sqrt{|\tilde{I}_j|} \geq \frac{\mu_j\sqrt{|I_j|}}{\sqrt{m}}\sqrt{1-\frac{2m}{|I_j|}} \geq \frac{\mu_j\sqrt{|I_j|}}{\sqrt{m}}\sqrt{1-\frac{\epsilon}{2(1+\epsilon)}} \geq \frac{\sqrt{2(1+\epsilon/2)\log n}}{2h(0)\sqrt{m}}$$

Then

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

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

Proof of Theorem 2.3

Let $\bar{s} = max_{I_j \in \mathbb{I}} \lceil |I_j|/m \rceil$ and $\bar{I}_j = \{k : J_k \cap I_j \neq \theta\}$. Assume (A) 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 (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 assuming (A). Let

$$W_i = \left(X_{l\bar{s}+1} + \ldots + X_{(l+1)\bar{s}}\right)/\sqrt{\bar{s}}, \quad l = d, \ldots, \left\lfloor T/\bar{s} \right\rfloor - 1,$$

then W_l can be rewritten as

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

where $Z_l' \overset{iid}{\sim} N(0,1)$, $\zeta_l' = \left(\zeta_{l\bar{s}+1} + \ldots + \zeta_{(l+1)\bar{s}}\right)/\sqrt{\bar{s}}$, and $\theta_l' = 0$ at all but at most $|\mathbb{I}|$ positions where $\theta_l' \leq \sqrt{2(1-\epsilon)\log n}$ by (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 $2h(0)\sqrt{mW_l}$ under the null and that under the alternative tends to $1 - o(1/n)$, i.e, define

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

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

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

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

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

where, by Cauchy-Schwarz inequality,

$$|err| \le C\nu^2 E\left[g(X) \cdot 1_{D_n} - 1\right]^2 \le C\nu^2 \left(E\left[g^2(X) \cdot 1_{D_n}\right] + 1\right).$$

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

$$E\left[g(X)\, 1_{D_n^c}\right] = o(1) \quad \text{and} \quad E\left[g^2(X)\, 1_{D_n}\right] = o(n).$$

The following Lemma is implied by Lemma 3.1 in Cai et al. (2010).

Lemma 6.1.

$$\int_{D_n^c} \phi\left(x - \sqrt{2(1-\epsilon)\log n}\right) dx = o(1) \quad \textit{and} \quad \int_{D_n} \frac{\phi^2\left(x - \sqrt{2(1-\epsilon)\log n}\right)}{\phi(x)} dx = o(n),$$

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

Then it is enough to show

$$E\left[g(X)\, 1_{D_n^c}\right] = \int_{D_n^c} \phi\left(x - \sqrt{2(1-\epsilon)\log n}\right) dx\,(1 + o(1)) + o(1) \quad (36)$$

and

$$E\left[g^2(X)\,1_{D_n}\right]=\int_{D_n}\frac{\phi^2\left(x-\sqrt{2(1-\epsilon)\log n}\right)}{\phi(x)}dx\,(1+o(1)). \quad (37)$$

Consider (36) first. By convolution,

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

where

$$\mathrm{I}=\int_{|w|>4\sqrt{\log n}/m} f_{\zeta_l'}(w)\left(\int_{D_n^c}\phi\left(x-\sqrt{2(1-\epsilon)\log n}-2h(0)\,w\right)dx\right)dw\le P\left(|\zeta_l'|>4\sqrt{\log n}/m\right)\to 0$$

by (31), and

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

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

$$\begin{aligned}|err| &\le C\frac{\sqrt{\log n}}{m}\left|\int_{D_n^c}\phi'\left(x-\sqrt{2(1-\epsilon)\log n}\right)dx\right|\int_{|w|\le 4\sqrt{\log n}/m} f_{\zeta_l'}(w)\,dw\\ &\le C\frac{\sqrt{\log n}}{m}n^{-\left(1-\sqrt{1-\epsilon}\right)^2}\to 0.\end{aligned}$$

Summing up above gives (36).

Next, consider (37). By convolution again,

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

where

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

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

$$\mathrm{IV}=\int_{|w|\le 4\sqrt{\log n}/m} f_{\zeta_l'}(w)\,\phi(x-2h(0)\,w)\,dw=\phi(x)\,(1+o(1))+err_1,$$

where

$$|err_1| \le C\frac{\sqrt{\log n}}{m}|\phi'(x)|\int_{|w|\le 4\sqrt{\log n}/m} f_{\zeta_l'}(w)\,dw = o(\phi(x))$$

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

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

in D_n. Similarly,

$$f_{Z_l'+2h(0)\zeta_l'}\left(x-\sqrt{2(1-\epsilon)\log n}\right) = \phi\left(x-\sqrt{2(1-\epsilon)\log n}\right)(1+o(1)) \quad (39)$$

in D_n. Substitute (38) and (39) into the definition of $E[g^2(X)1_{Dn}]$ gives (37).

Proof of Theorem 3.1

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

$$P_{H_0}\left(|\widehat{\mathbb{I}}|>0\right) \le \frac{C}{\sqrt{\log T}} \quad (40)$$

and

$$P_{H_1}\left(\max_{I_j\in\mathbb{I}}\min_{\widehat{I}_j\in\widehat{\mathbb{I}}} D\left(\widehat{I}_j, I_j\right) > \delta_n\right) \le Cqn^{-C\epsilon^2} + Cq\left(\bar{s}/m\right)(L/m)\,n^{-C\delta_n^2} \quad (41)$$

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

Consider (40) first. Since

$$P_{H_0}\left(|\widehat{\mathbb{I}}|>0\right) \le P_{H_0}\left(\max_{\tilde{I}\in\mathbb{J}_T(L)} X\left(\tilde{I}\right) > \lambda_n^*\right) \le P_{H_0}\left(\max_{\tilde{I}\in\mathbb{J}_T} X\left(\tilde{I}\right) > \lambda_n^*\right)$$

and λ_n^* converges to λ_n at the order of $n^\gamma/\sqrt{\log n}$, then, by Theorem 2.1 and some routine calculation, (40) follows.

Next, we show (41). Since all the elements in $\mathbb{J}_T(L)$ can not reach more than one signal segments, the accuracy of estimating any $I_j \in \mathbb{I}$ by some element in $\widehat{\mathbb{I}}$ 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. De ne the following events:

$$A_j = \left\{\mathbb{I}^{(1)} \ne \varnothing \text{ when only } I_j \text{ exists}\right\}, \qquad B_j = \left\{D\left(\widehat{I}_1, I_j\right) \le \delta_n\right\}, \quad j = 1, \ldots, q.$$

We have

$$P_{H_1}\left(\max_{I_j\in\mathbb{I}}\min_{\widehat{I}_j\in\widehat{\mathbb{I}}} D\left(\widehat{I}_j, I_j\right) > \delta_n\right) \le P\left(\exists I_j \in \mathbb{I}: A_j^c \cup \left(A_j \cap B_j^c\right)\right) \le \sum_{j=1}^{q} P\left(A_j^c\right) + \sum_{j=1}^{q} P\left(B_j^c | A_j\right),$$

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

$$P\left(A_j^c\right) \le Cn^{-C\epsilon^2} \quad (42)$$

and

$$P\left(B_j^c | A_j\right) \le Cn^{-C} + C\left(|I_j|/m\right)(L/m)\, n^{-C\delta_n^2}. \quad (43)$$

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

$$P\left(A_j^c\right) = P\left(\max_{\tilde{I}\in\mathbb{I}_T(L)} X\left(\tilde{I}\right) \le \lambda_n^*, \text{ only } I_j \text{ exists}\right) \le P\left(X\left(\tilde{I}_j\right) \le \lambda_n^*, \text{ only } I_j \text{ exists}\right),$$

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

Now consider (43). Define

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

for some $C > 0$. Since for an interval I, $D(I, I_j) > \delta_n$ implies $D\left(\tilde{I}, \tilde{I}_j\right) > C\delta_n$, 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 the Step 3 of RSI algorithm. This further implies the existence of some $\tilde{I} \in \mathbb{K}_T(L)$ such that $X\left(\tilde{I}\right) \ge X\left(\tilde{I}_j\right)$. Denote

$$\mathbb{K}_0 = \left\{\tilde{I} \in \mathbb{K}_T(L): \tilde{I} \cap \tilde{I}_j = \varnothing\right\}, \qquad \mathbb{K}_1 = \left\{\tilde{I} \in \mathbb{K}_T(L): \tilde{I} \ne \tilde{I}_j = \varnothing\right\}.$$

We have

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

Since $|\mathbb{K}_0|=\sum_{\tilde{I}\in\mathbb{I}_T(L):\tilde{I}\cap\tilde{I}_j=\varnothing} 1\left\{X\left(\tilde{I}\right)>\lambda_n^*\right\}$, which converges to $\sum_{\tilde{I}\in\mathbb{I}_T(L):\tilde{I}\cap\tilde{I}_j=\varnothing} P\left(X\left(\tilde{I}\right)>\lambda_n^*\right)$ exponentially fast, and $\sum_{\tilde{I}\in\mathbb{I}_T(L):\tilde{I}\cap\tilde{I}_j=\varnothing} P\left(X\left(\tilde{I}\right)>\lambda_n^*\right)\le CT\,(L/m)\,P\left(Z\left(\tilde{I}>\sqrt{2\log T}\right)\right)\le CL/m$, then we have

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

It is left to show

$$P\left(X\left(\tilde{I}\right)-X\left(\tilde{I}_j\right)\ge 0,\tilde{I}\in\mathbb{K}_0\right)\le Cn^{-C} \quad (44)$$

and

$$P\left(X\left(\tilde{I}\right)-X\left(\tilde{I}_j\right)\ge 0,\tilde{I}\in\mathbb{K}_1\right)\le Cn^{-C\delta_n^2}. \quad (45)$$

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

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

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

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

So

$$P\left(X\left(\tilde{I}\right)-X\left(\tilde{I}_j\right)\ge 0,\tilde{I}\in\mathbb{K}_0\right)\le P\left(\frac{Z\left(\tilde{I}\right)}{2h\,(0)\,\sqrt{m}}-\frac{Z\left(\tilde{I}_j\right)}{2h\,(0)\,\sqrt{m}}\ge\mu_j\sqrt{|\tilde{I}_j|}-\mu_j+\frac{\zeta\left(\tilde{I}_j\right)}{\sqrt{m}}-\frac{\zeta\left(\tilde{I}\right)}{\sqrt{m}}\right)\le P\left(Z\left(\tilde{I}\right)-Z\left(\tilde{I}_j\right)\ge\sqrt{2\,(1+\epsilon/2)\log n}\right)+Cn^{-C}\le P\left(N\,(0,2)\ge\sqrt{2\,(1+C)\log n}\right)+Cn^{-C},$$

where the second inequality is because

$$\mu_j\sqrt{|\tilde{I}_j|}-\mu_j\ge\sqrt{|\tilde{I}_j|}\sqrt{1-\frac{\epsilon}{2\,(1+\epsilon)}}$$

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

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

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

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

$$X\left(\tilde{I}\right) - X\left(I\right) = LR_1 + LR_2 + LR_3,$$

$$LR_1 = \left(\frac{1}{\sqrt{|\tilde{I}|}} - \frac{1}{|\tilde{I}_j|}\right) \sum_{k \in \tilde{I} \cap \tilde{I}_j} X_k = \left(\frac{1}{\sqrt{|\tilde{I}|}} - \frac{1}{|\tilde{I}_j|}\right)\left(\mu_j |\tilde{I} \cap \tilde{I}_j| + \frac{\Sigma_{k \in \tilde{I} \cap \tilde{I}_j} Z_k}{2h\left(0\right)\sqrt{m}} + \frac{\Sigma_{k \in \tilde{I} \cap \tilde{I}_j} \zeta_k}{\sqrt{m}}\right),$$

$$LR_2 = \frac{1}{\sqrt{|\tilde{I}|}} \sum_{k \in \tilde{I} \setminus \tilde{I} \cap \tilde{I}_j} X_k \leq \frac{1}{\sqrt{|\tilde{I}|}}\left(\mu_j + \frac{\Sigma_{k \in \tilde{I} \setminus \tilde{I} \cap \tilde{I}_j} Z_k}{2h\left(0\right)\sqrt{m}} + \frac{\Sigma_{k \in \tilde{I} \setminus \tilde{I} \cap \tilde{I}_j} \zeta_k}{\sqrt{m}}\right)$$

$$LR_3 = \frac{1}{\sqrt{|\tilde{I}_j|}} \sum_{k \in \tilde{I}_j \setminus \tilde{I} \cap \tilde{I}_j} X_k = \sqrt{|\tilde{I}_j|}\left(-\mu_j |\tilde{I}_j \setminus \tilde{I} \cap \tilde{I}_j| - \frac{\Sigma_{k \in \tilde{I}_j \setminus \tilde{I} \cap \tilde{I}_j} Z_k}{2h\left(0\right)\sqrt{m}} - \frac{\Sigma_{k \in \tilde{I}_j \setminus \tilde{I} \cap \tilde{I}_j} \zeta_k}{\sqrt{m}}\right)$$

Note that LR_1, LR_2 and LR_3 are independent, and

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

for some $\tau > 0$. On the other hand, $D\left(\tilde{I}, \tilde{I}_j\right) \geq C\delta_n$ implies

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

Combing above with (46) we have

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

Given (15) and the choice of m, (45) follows for δn satisfying

$$\sqrt{\log q + \log\left(\bar{s}/m\right) + \log\left(L/m\right)} / \sqrt{\log} \ll \delta_n \ll 1$$

References

Abyzov A, Urban A, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research. 2011; 21:974–984. [PubMed: 21324876]

Alkan C, Coe B, Eichler E. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011; 12:363–375. [PubMed: 21358748]

Arias-Castro E, Donoho D, Huo X. Near-optimal detection of geometric objects by fast multiscale methods. IEEE Transactions on Information Theory. 2005; 51:2402–2425.

Bravo H, Irizarry R. Model-Based Quality Assessment and Base-Calling for Second- Generation Sequencing Data. Biometrics. 2010; 66:665–674. [PubMed: 19912177]

Brown LD, Cai TT, Zhou HH. Robust nonparametric estimation via wavelet median regression. Ann. Statist. 2008; 36:2055–2084.

Cai T, Jeng XJ, Jin J. Optimal detection of heterogeneous and heteroscedastic mixtures. 2010 Manuscript.

Cai TT, Zhou HH. Asymptotic equivalence and adaptive estimation for robust nonparametric regression. Ann. Statist. 2009; 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, Mardis E. BreakDancer: an algorithm for high-resolution mapping of genomic structrual variation. Nature Methods. 2009; 6:677–681. [PubMed: 19668202]

Cheung M, Down T, Latorre I, Ahringer J. Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic Acids Research. 2011 in press.

Chiang D, Getz G, Jaffe D, O'Kelly M, Zhao X, Carter S, Russ C, Nusbaum C, Meyerson M, Lander E. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nature Methods. 2009; 6:99–103. [PubMed: 19043412]

Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse1 K, Cole1 K, Moss Y, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K, Kim C, Geiger EA, McGrady PW, Blakemore AIF, London WB, Shaikh TH, Bradfield J, Grant SFA, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM. Copy number variation at 1q21.1 associated with neuroblastoma. Nature. 2009; 459:987–991. [PubMed: 19536264]

Feuk L, Carson A, Scherer S. Structural variation in the human genome. Nature Review Genetics. 2006; 7:85–97.

Ivakhno S, Royce T, Cox A, et al. CNAseg-a novel framework for identification of copy number changes in cancer from second-generation sequencing data. Bioinformatics. 2010; 26:3051–3058. [PubMed: 20966003]

Jeng JJ, Cai TT, Li H. Optimal sparse segment identification with application in copy number variation analysis. J. Am. Statist. Ass. 2010; 105:1156–1166.

Kim T, Luquette L, Xi R, et al. rSW-seq: Algorithm for detection of copy number alterations in deep sequencing data. BMC Bioinformatics. 2010; 11 DOI: 10.1186/1471-2105-11-432.

Li J, Jiang H, Wong W. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biology. 2010; 11:R50. [PubMed: 20459815]

Li Y, Zheng H, Lou R, et al. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. NATURE BIOTECHNOLOGY. 2011; 29:723–730.

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

Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nature Methods. 2009; 6:S13–S20. [PubMed: 19844226]

Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads. PLos ONE. 2011; 6(1):e16327. [PubMed: 21305028]

Mills RR, Walter K, Stewart C, Korbel JO. Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 470:59–65. [PubMed: 21293372]

Nord A, Lee M, King M, et al. Accurate and exact CNV identification from targeted high-throughput sequence data. BMC Genomics. 2011; 12 DOI: 10.1186/1471-2164-12-184.

Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004; 5(4)

Quackenbush J. Microarray data normalization and transformation. Nat Genet. 2002; 32:496–501. [PubMed: 12454644]

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, Gonzlez 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, Hurles M. Global variation in copy number in the human genome. Nature. 2006; 444:444–454. [PubMed: 17122850]

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al. Large-scale copy number polymorphism in the human genome. Science. 2004; 305:525–528-97. [PubMed: 15273396]

Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology. 2008; 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. Large recurrent microdeletions associated with schizophrenia. Nature. 2008; 455:178–179. [PubMed: 18784712]

Stone J, O'Donovan M, Gurling H, Kirov G, Blackwood D, Corvin A, Crad-dock N, Gill M, Hultman C, Lichtenstein P, et al. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 2008; 455:237–241. [PubMed: 18668038]

Urban A, Korbel J, Selzer R, Richmond T, Hacker A, Popescu G, Cubells J, Green R, Emanuel B, Gerstein M, Weissman S, Snyder M. High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proc. Natl. Acad. Sci. USA. 2006; 103:45344539.

Walsh T, McClellan J, McCarthy S, Addington A, Pierce S, Cooper G, Nord A, Kusenda M, Malhotra D, Bhandari A, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science. 2008; 320:539–543. [PubMed: 18369103]

Walther G. Optimal and fast detection of spacial clusters with scan statistics. Ann. of Stat. 2010

Xie C, Tammi M. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009; 10:80. [PubMed: 19267900]

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

Zhang F, Gu W, Hurles M, Lupski J. Copy number variation in human health, disease and evolutions. Annual Review of Genomics and Human Genetics. 2009; 10:451–481.

Zhou, H. Tech. rep. Department of Statistics, Yale University; 2006. A note on quantile coupling inequalities and their applications.

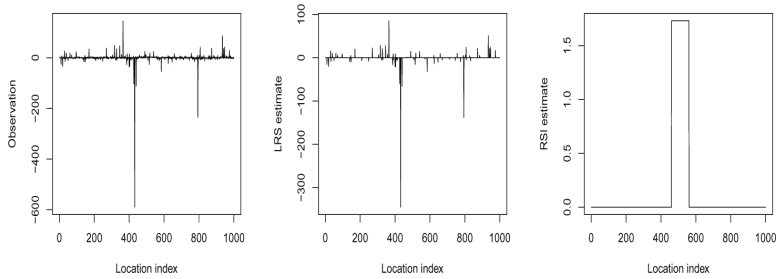


Fig. 1.

Effects of long-tailed error distribution on segment identification. Left plot: Data with Cauchy noise and a signal segment at [457 : 556]. Middle plot: Intervals identified and estimated interval means by LRS. Right plot: Interval identified and estimated means by RSI.

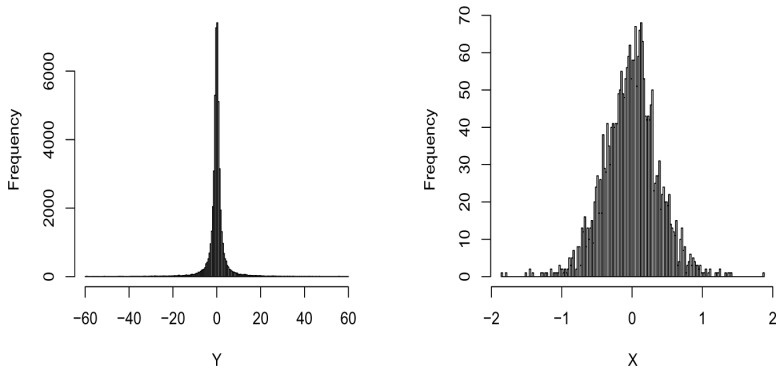
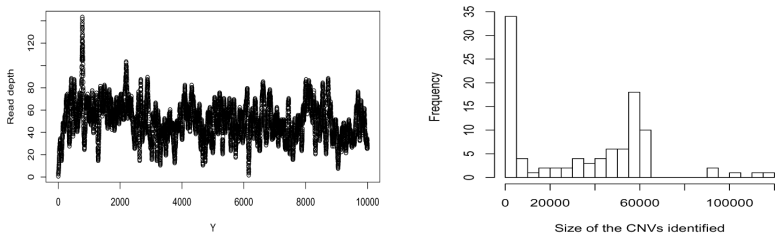


Fig. 2.
Simulated data. Left: histogram of the original data Y_i with $t(1)$ noise and $\mu = 1$. Right: histogram of the transformed data X_k.



(a) (b)

Fig. 3.

Analysis of the chromosome 19 data of individual NA19240 from the 1000 Genomes Project. (a): scatter plot of the first 10,000 observations; (b): histogram of the sizes of the CNVs identified in base pairs.

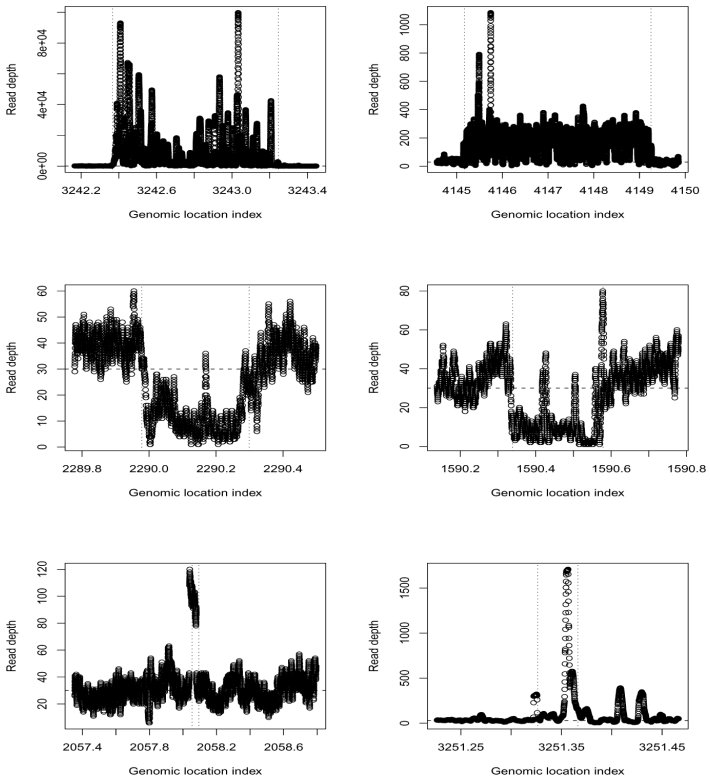


Fig. 4.

Examples of CNV identified on chromosome 19 of NA19240 from the 1000 Genomes Project. Top two plots: duplications, regions with the highest scores; middle two plots: deletions, regions with the smallest scores; bottom two plots: the two shortest CNVs identified. For each plot, the horizontal line presents the median count of 30 and the vertical dashed lines represent the estimated CNV boundaries. For each plot, x-axis is the genomic location in base pairs/10,000.

Table 1

Simulation results: medians of Dj and #O for RSI with $m = 20$ and $L = 6$. In Tables 1–3, the estimated standard errors based on the bootstrap appear in parentheses.

		$D_{1(\lvert I_1\rvert=100)}$	$D_{2(\lvert I_2\rvert=40)}$	$D_{3(\lvert I_3\rvert=20)}$	#O
t(1)	μ = 1.0	0.080(0.015)	1.000(0.026)	1.000(0.000)	2.000(0.330)
	μ = 1.5	0.087(0.003)	0.184(0.017)	1.00(0.000)	2.000(0.260)
	μ = 2.0	0.087(0.009)	0.150(0.020)	0.423(0.220)	2.000(0.140)
t(3)	μ = 1.0	0.087(0.005)	1.000(0.270)	1.000(0.000)	0.000(0.000)
	μ = 1.5	0.060(0.009)	0.175(0.029)	1.000(0.000)	0.000(0.000)
	μ = 2.0	0.050(0.008)	0.150(0.016)	0.293(0.019)	0.000(0.000)
t(30)	μ = 1.0	0.070(0.014)	1.000(0.32)	1.000(0.000)	0.000(0.000)
	μ = 1.5	0.065(0.012)	0.175(0.021)	1.000(0.245)	0.000(0.000)
	μ = 2.0	0.050(0.010)	0.175(0.019)	0.250(0.028)	0.000(0.000)

J R Stat Soc Series B Stat Methodol. Author manuscript; available in PMC 2013 November 01.

Table 2

Simulation comparisons of RSI, LRS, and CBS, where both homogeneous and heterogeneous noises are considered. Homogenous 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.

	RSI		LRS		CBS	
	$D_{2(\|I_2\|=40)}$	#O	$D_{2(\|I_2\|=40)}$	#O	$D_{2(\|I_2\|=40)}$	#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)

Table 3

Simulation results: effect of bin size m on the performance the RSI. μ is fixed at 2.

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