Sparse Segment Identifications with Applications to DNA Copy Number Variation Analysis

Jessie Jeng, Tony Cai and Hongzhe Li
University of Pennsylvania

Abstract

Copy number variations (CNVs) are alterations of the DNA of a genome that results in the cell having an abnormal number of copies of one or more sections of the DNA. Germline CNVs have been shown to be associated with many complex diseases. Detecting and identifying all the CNVs in a given sample or in multiple population-based samples is an important first step in many CNV analyses. In this chapter, we review statistical methods for CNV identification, focusing on latest developed methods for sparse segment identifications in various settings. We review methods for optimal CNV identification for a single sample based on SNP allele intensity data, methods for robust CNV identification based on the next generation sequence (NGS) data. and methods for detection of recurrent CNVs in a population when a large set of samples are available. Our review focuses on problem formulations and optimal statistical properties of the procedures. We illustrate these methods using data from the 1000 Genomes Project and data from a large genome-wide association study of neuroblastoma. Areas that need further research are also presented.

1 Introduction

Structural variants in the human genome (Sebat et al., 2004; Feuk et al., 2006), including the copy number variants (CNVs) and balanced rearrangements such as inversions and translocations, play an important role in the genetics of complex diseases. Copy number variation refers to duplication or deletion of a segment of DNA sequences compared to a reference genome assembly. In normal genomic regions, there are two copies of DNAs, one from father, one from mother. 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. A high proportion of the genome, currently estimated at up to 12%, is subject to copy number variation (Hastings et al., 2009). Hastings et al. (2009) further provides possible molecular mechanisms of change in the gene copy number. Analysis of CNV in developmental and neuropsychiatric disorders (Feuk et al., 2006; Stefansson et al., 2008; Stone et al., 2008; Walsh et al., 2008) and in cancer (Diskin et al., 2009), has led to the identification of novel disease-causing copy number variant mutations, thus contributing important new insights into the genetics of these complex diseases.

This chapter reviews a few related statistical problems that arise from the CNV analysis for germline constitutional genomes, focusing on methods for sparse segment identifications. Current high-throughput genotyping technology is able to generate genome-wide observations in very high resolutions. In this type of ultrahigh dimensional data, the number of CNVs is relatively very small and the CNV segments are usually very short. These impose major difficulties in CNV identification (Zhang et al., 2009b). The emerging technologies of DNA sequencing have further enabled the identification of CNVs by the 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 structural variations of the target genome (for reviews, see Medvedev et al. (2009) and Alkan et al. (2011)). The general statistical problem is to identify sparse and subtle signal segments hidden in a long sequence of noisy data. Let $\mathbb{I}$ be the collection of all signal segments. The goal is two-fold: (1) to detect the existence of segments; and (2) to identify the locations of the segments if they exist. Precisely, we want to test

$$H_0 : \mathbb{I} = \emptyset \qquad \text{against} \qquad H_1 : \mathbb{I} \neq \emptyset, \tag{1.1}$$

and if H_1 is rejected, identify each $I_k \in \mathbb{I}$.

Although many methods have been developed for the CNV analysis, including methods based on hidden Markov models (Wang et al., 2007) and methods based on fused penalized regression (Zhang et al., 2010), these methods do not provide any theoretical results in term of optimality of the procedures. We review three approaches for testing the above null hypothesis and for identifying the CNVs at both individual sample and multiple-sample levels:

1. optimal CNV identification for a single sample based on single nucleotide polymorphism (SNPs) allele intensity data.

2. robust CNV identification based on the next generation sequence (NGS) data.

3. detection of recurrent CNVs in a population when a large set of samples are available.

These three problems can be formulated as general sparse segment identification problems in the high dimensional settings. Our review focuses on the problem formulations and some key steps of the procedure and some theoretical results. We also illustrate these methods with real data analyses. In addition, we discuss a few other interesting problems that require new statistical methodology, including methods for testing CNV associations and methods for CNV analysis based on mapping distances from the pair-end sequencing.

2 Optimal CNV identification for a single sample based on Log R ratio data from the SNP arrays

2.1 Statistical formulation and summary of theoretical results

The SNP data generated from SNP genotyping array platforms, such as the HumanHap550 array can be informative for CNV analysis. At a given SNP location, the observation is the Log R ratio (LRR) calculated as $\log_2(R_{obs}/R_{exp})$, where R_{obs} is the observed total intensity of both major and minor alleles for a given SNP, and R_{exp} is computed from linear interpolation of canonical genotype clusters (Peiffer et al., 2006). When there is no copy number change at a SNP location for the observed sample, R_{obs} should be the same as R_{exp}, and the LRR has a baseline level of zero. If there is a copy number deletion/duplication at a SNP location, R_{obs} can be smaller/larger than R_{exp}, therefore the LRR the deviates from zero to the negative/positive side, which implies deletion or duplication. See Figure 1 for an illustration on how different intensities can be used to infer the CNVs.

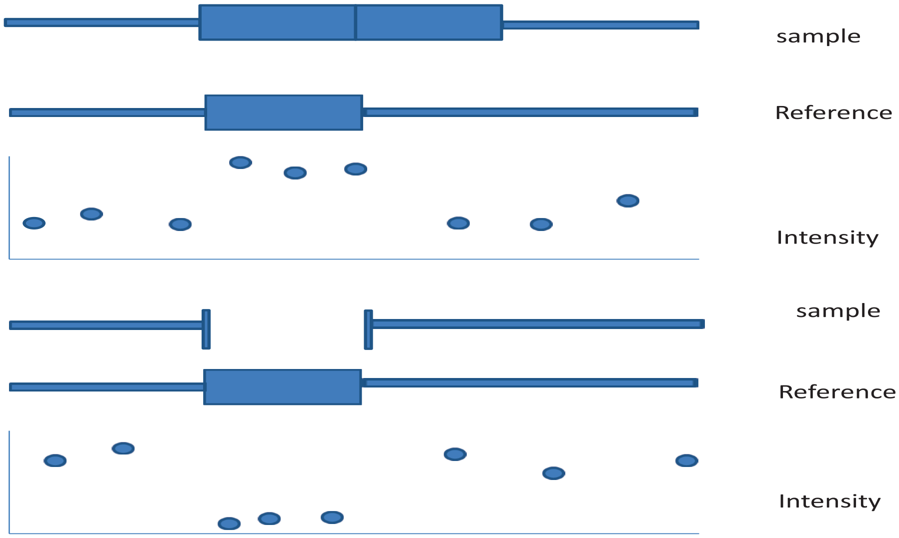

Figure 1: Illustration of detecting insertions and deletions based on the LRR values. Top panel: duplication; bottom panel: deletion.

The problem of CNV identification using LRR data can be generalized as a problem of identifying sparse and short signal segments in a long sequence of noisy data. specifically, we observe $\{X_i, i = 1, ..., n\}$ with

$$X_i = \mu_1 1_{\{i \in I_1\}} + \ldots + \mu_q 1_{\{i \in I_q\}} + \sigma Z_i, \qquad 1 \leq i \leq n. \tag{2.2}$$

Here $q = q_n$ is the unknown number of the signal segments, possibly increasing with n, $I_1, \ldots I_q$ are disjoint intervals representing signal segments with unknown locations, $\mu_1, \ldots \mu_q$ are unknown positive means, σ is an unknown noise level, and $Z_i \overset{iid}{\sim} N(0,1)$.

The problem formulated in Section 1 for data described by model (2.2) pertains to statistical research in several areas. Without segment structure, it is closely related to large-scale multiple testing, which has motivated many novel procedures such as false discovery rate (FDR) (Benjamini and Hochberg, 1995) and higher criticism thresholding (HCT) (Donoho and Jin, 2008). Arias-Castro et al. (2005) considered the problem of detecting the existence of signals when there is only one signal segment. This is a special case of the testing part of our problem with $q = 1$. They showed that the detection boundary in this case is $\sqrt{2 \log n}/\sqrt{|I|}$, i.e., the signal mean should be at least $\sqrt{2 \log n}/\sqrt{|I|}$ in order for a signal with length $|I|$ to be reliably detected and that the generalized likelihood ratio test (GLRT) can be used for detecting the segment. A closely related result in Section 6 of Hall and Jin (2010) demonstrates the detection boundary under a wide range of signal sparsity when signals appear in several clusters. Further, Arias-Castro et al. (2005) and Walther (2010) studied detection of geometric objects and spatial clusters in 2-dimensional space, respectively, and Arias-Castro et al. (2008) provides detection threshold for the existence of an unknown path in a 2-dimensional regular lattice or a binary tree.

The problem considered here is also related to the problem of change-point detection, since it involves shifts in the characteristics of a sequence of data. Change-point detection in a single sequence has been extensively studied. See Zack (1983) and Bhattacharya (1994) for a review of the literature. Olshen et al. (2004) used the likelihood ratio based statistics for analysis of DNA copy number data, and Zhang and Siegmund (2007) proposed a BIC-based model selection criterion for estimating the number of change-points. Olshen et al. (2004) further developed an iterative circular binary segmentation procedure for segmentation of a single sequence and showed promising results in analysis of DNA copy number data, whereas Zhang et al. (2008) extended the problem of change-point detection from single sequence to multiple sequences in order to increase the power of detecting changes.

The problem is studied from another perspective in Jeng et al. (2010), which focuses on the recovery of sparse and subtle signals. For any signal segment, a statistical characterization of identifiable region is derived. When a signal segment is in the identifiable region, it is possible to reliably separate the segment from the noise; otherwise, it is impossible to do so. A likelihood ratio selection (LRS) procedure was proposed to identify the signal segments. The LRS involves scanning the linear data sequence of length (n) with all the segment of length less than a pre-specified interval length L and then calculate the likelihood ratio statistics for all these intervals. A threshold of $t_n = \sqrt{2\log(Ln)}$ is used to control for the genome-wide error rate. Specifically, the LRS has the following steps:

Step 1: Let $\mathbb{J}_n(L)$ be the collection of all possible subintervals in $\{1, \ldots, n\}$ with interval length less than or equal to L. Let $j = 1$. Define $\mathbb{I}^{(j)} = \{\tilde{I} \in \mathbb{J}_n(L) : X(\tilde{I}) > t_n\}$.
Step 2: Let $\hat{I}_j = \arg\max_{\tilde{I}\in\mathbb{I}^{(j)}} X(\tilde{I})$.
Step 3: Update $\mathbb{I}^{(j+1)} = \mathbb{I}^{(j)} \backslash \{\tilde{I} \in \mathbb{I}^{(j)} : \tilde{I} \cap \hat{I}_j \neq \emptyset\}$.
Step 4: Repeat Step 2-4 with $j = j + 1$ until $\mathbb{I}^{(j)}$ is empty.

Define 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 $\hat{\mathbb{I}}$.

Jeng et al. (2010) showed that the LRS provides consistent estimates for any signal segments in the identifiable region. In other words, the LRS procedure is an optimal procedure, which can reliably separate signal segments from noise as long as the signal segments can be identified. To elucidate the exact meaning of optimality, Jeng et al. (2010) introduced a quantity to measure the accuracy of an estimate of a signal segment. Recall that $\mathbb{I}$ is the collection of signal segments. Denote $\hat{\mathbb{I}}$ to be the collection of interval estimates. For any $\hat{I} \in \hat{\mathbb{I}}$ and $I \in \mathbb{I}$, define the dissimilarity between $\hat{I}$ and I as

$$D(\hat{I}, I) = 1 - |\hat{I} \cap I| / \sqrt{|\hat{I}||I|}, \tag{2.3}$$

where $|\cdot|$ represents the cardinality of a set. Note that $0 \leq D(\hat{I}, I) \leq 1$ with $D(\hat{I}, I) = 1$

indicating disjointness and $D(\hat{I}, I) = 0$ indicating complete identity. Similar quantity has been used in Arias-Castro et al. (2005) to measure the dissimilarity between intervals.
Definition 1 An identification procedure is *consistent* for a subset $\Omega \subseteq \mathbb{I}$ if its set of estimates $\hat{\mathbb{I}}$ satisfies

$$P_{H_0}(|\hat{\mathbb{I}}| > 0) + P_{H_1}(\max_{I_j \in \Omega} \min_{\hat{I}_j \in \hat{\mathbb{I}}} D(\hat{I}_j, I_j) > \delta_n) \to 0, \tag{2.4}$$

for some $\delta_n = o(1)$. Obviously, the first term on the left measures the type I error. The second term, which is the probability that some signal segments in Ω are not 'substantially matched' by any of the estimates, essentially measures the type II error.

2.2 Application to CNV analysis of LRR data from a trio

As an example, Jeng et al. (2010) presented an application of using the genotyping data for a father-mother-child trio from the Autism Genetics Resource Exchange (AGRC) collection (Bucan et al., 2009), genotyped on the Illumina HumanHap550 array. For each individual, the LRR data are observed at a total of 547,458 SNPs over 22 autosomes, and the numbers of SNPs on each chromosome range from 8,251 on chromosome 21 to 45,432 on chromosome 2. For each individual, the goal is to identify the CNVs by LRS. The purpose of using data from a trio is to partially validate the results since some CNVs are inherited from parents to child. Figure 2 shows the CNV segments with the likelihood ratio test scores (xstar) for the segments that the LRS algorithm selected for the child. The CNV segments identified in the parents are also plotted if they overlap with the CNV segments of the child. It is interesting to note that many of the CNV segments identified in the child were also observed in one of the parents, further indicating that some CNVs are inheritable and the LRS algorithm can effectively identify these CNVs. More details about the selected CNV segments and a comparison with the hidden Markov model (HMM)-based method implemented in PennCNV package (Wang et al., 2007) can be found in Jeng et al. (2010).

One of the key assumption made in Jeng et al. (2010) is that the noise follows a normal distribution throughout the genomes and the baseline mean values of the data are zero. In real data application, the LRR data can also be affected by genomic features such as GC contents and SNP densities. It is important to pre-process the data to ensure that they roughly follow normal distributions. Loess fitting provides method of normalization for the correction of wave like correlations in signal intensities across the genome Marioni et al. (2007).

3 Robust CNV identification for one sample based on NGS data

3.1 Statistical formulation and summary of theoretical results

Classical approaches for signal detection rely heavily on normality or other distributional assumptions on data observed. Examples include but not limited to the false discovery rate (FDR) (Benjamini and Hochberg, 1995), the higher criticism (HC) (Donoho and Jin, 2004) and the generalized likelihood ratio test (GLRT) (Arias-Castro et al., 2005). It is crucial

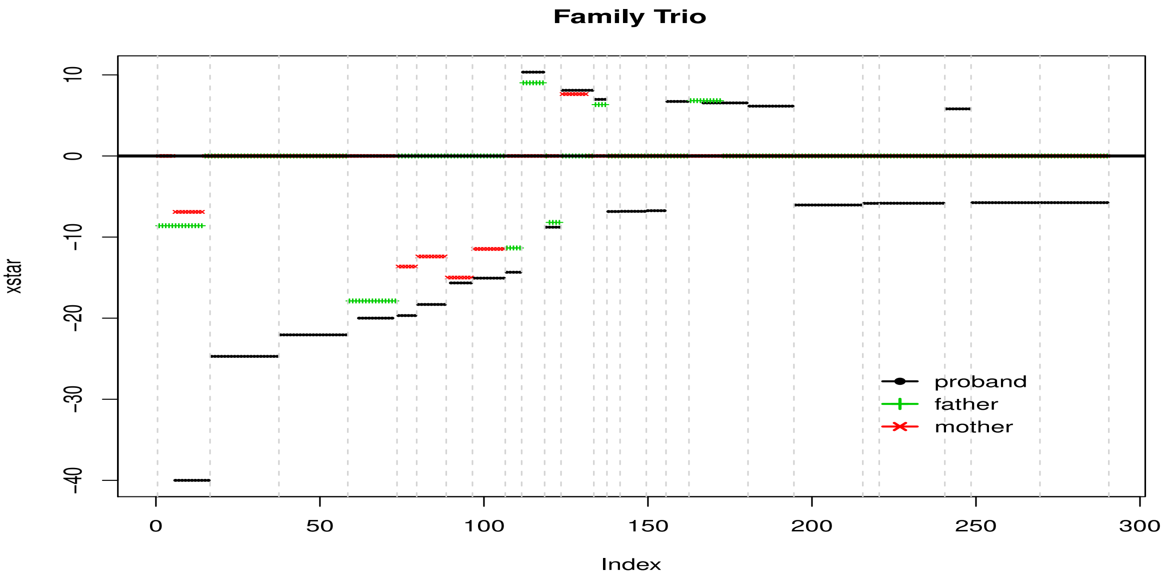

Figure 2: *Summary of results of LRS for CNV detection for a trio: the LR test statistics for the CNV segments identified by the proposed LRS procedure for the child, sorted by the absolute values of the likelihood ratio statistics. One segment with large statistics (-116.70 for the child) is truncated as -40 for better view.*

for these methods to specify the tail distribution of the test statistics under null hypothesis, so that false positive errors can be controlled at a desired level. However, the tail behavior critically depends on the noise distribution, which is usually unknown and hard to estimate in real applications. Although the FDR or the HC can be applied to p-values obtained from nonparametric methods, popular nonparametric methods such as permutation are often computationally expensive and not feasible for ultra-high dimensional data.

The emerging technologies of next generation sequencing enable CNV analysis at even higher resolutions. NGS can generate millions of short sequence reads along the whole human genome. When these short reads are mapped to the reference genome, read-depth data are generated to count the number of reads that cover a genomic location or a small bin along the genome. The read counts or read depth (RD) data provide important information about the CNVs (Shendure and Ji, 2008; Medvedev et al., 2009; Yoon et al., 2009; Mills et al., 2011; Sudmant et al., 2010) that a given individual carries. CNVs include large-scale deletions, duplications and insertions and form one type of genetic variation. 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. 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 examining the maximum p-value in a given interval. The p-values are obtained by a normality assumption on the RD data. 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) and Chiang et al. (2009) 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.

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, robust methods that are adaptive to unknown noise distribution and computationally efficient at the same time are greatly needed in order to minimize both false positive and false negative identification of CNVs and to estimate CNV break points more precisely.

The NGS data $\{Y_1, \cdots, Y_n\}$ is modeled as

$$Y_i = \mu_1 1_{\{i \in I_1\}} + \ldots + \mu_q 1_{\{i \in I_q\}} + \xi_i, \qquad 1 \leq i \leq n, \tag{3.5}$$

where 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). The above model is more general than (2.2) with the distribution of the noise ξ_i unspecified. 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. Under this more general model, parametric methods designed for Gaussian noise or any other tractable noise may fail completely and provide a large number of misidentifications.

To tackle this difficulty, a computationally efficient method called robust segment identifier (RSI) was proposed in Cai et al. (2011), which provides a robust and near-optimal solution for segment identification over a wide range of noise distributions. As an illustration, 1000 observations are generated based on Cauchy $(0, 1)$, and the signal segment is set at $[457 : 556]$ with a positive mean. Figure 3 compares the RSI with the LRS, which is an optimal procedure for Gaussian noise. 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, which was first introduced in Brown et al. (2008) and Cai and Zhou (2009) in the context of nonparametric regression. 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.

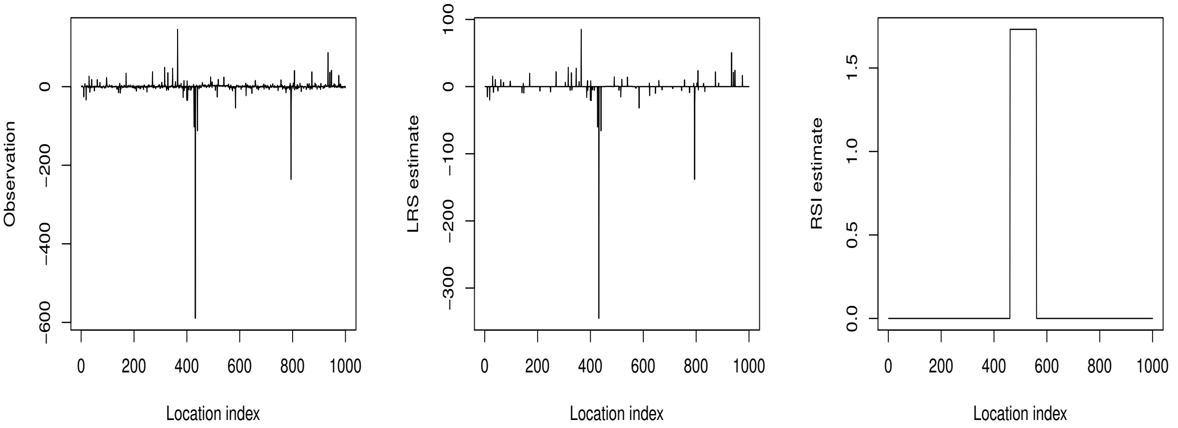

Figure 3: Effects of long-tailed error distribution on segment identification. Left plot: Data with Cauchy noise and a signal segment at [457 : 556]. Middle plot: Intervals identified and estimated interval means by LRS. Right plot: Interval identified and estimated means by RSI.

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, such as LRS, can then be applied to the new data set. Specifically, 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 k-th group as $J_k = \{i : (k-1)m + 1 \le i \le km\}$, and generate the transformed dataset as

$$X_k = \text{median}\{Y_i : i \in J_k\}, \qquad 1 \le k \le T. \tag{3.6}$$

Set

$$\eta_k = \text{median}\{\xi_i : i \in J_k\}, \qquad 1 \le k \le T, \tag{3.7}$$

then the medians X_k can be written as

$$X_k = \theta_k + \eta_k, \qquad 1 \le k \le T, \tag{3.8}$$

where

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

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

$$|h(y) - h(0)| \le Cy^2 \tag{3.9}$$

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 3.1 *Assume (4.14), (3.13), and transformation (3.8), then* η_k *can be written as*

$$\eta_k = \frac{1}{2h(0)\sqrt{m}} Z_k + \frac{1}{\sqrt{m}} \zeta_k, \tag{3.10}$$

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

$$E\zeta_{k1} = 0 \text{ and } E|\zeta_{k1}|^l \leq C_l m^{-l}, \tag{3.11}$$

$$P(\zeta_{k2} = 0) \geq 1 - C\exp(-am) \tag{3.12}$$

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 that after the data transformation in (3.8), existing methods for Gaussian noise can be applied to $X_k, 1 \leq k \leq T$. 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.

It is shown in Cai et al. (2011) that the RSI provides robust and near-optimal results as long as the distribution of ξ_i is symmetric about 0 with the density function h satisfying $h(0) > 0$ and

$$|h(y) - h(0)| \leq Cy^2 \tag{3.13}$$

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.

Like the LRR data that are subject to local genomic wave effects, the read depth data also depend on the local genomic features such as the GC contents of the genome. GC-content bias describes the dependence between fragment count (read coverage) and GC content found in high-throughput sequencing assays, particularly the Illumina Genome Analyzer technology. This bias can dominate the signal of interest for analyses that focus on measuring fragment abundance within a genome, such as copy number estimation. Benjamini and Speed (2011) proposed a new method to calculate predicted coverage and correct for the bias. This parsimonious model produces single bp prediction which suffices to predict the GC effect on fragment coverage at all scales, all chromosomes and for both strands. This model should be applied to estimate the GC-corrected read depths before our RSI procedure is used in order to reduce the effects of local feature on CNV identifications.

3.2 Application to CNV analysis of a trio sequencing data from the 1000 Genome Project

We applied this RSI procedure to a HapMap Yoruban trio and identified the CNVs independently for the parents and the child. After the short reads are mapping to the reference human DNA sequences, we obtain the RD data at $n = 54,361,060$ genomic locations. 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 read depth data.

We apply the RSI with $m = 400$ and $L = 150$ to each of the three individuals separately, 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. Figure 4 shows the concordant rates of the CNVs identified using the RSI when top 20, 50, 100, 200 and 300 CNVs identified for each of the three individuals on chromosome 19 are compared. We observe a very high concordant rates, which further validates the RSI for CNV detection based on the NGS data. On the same plot, we also present the CNV calling results assuming that the read depth data follow a negative binomial (NB) distribution. The concordant rates based on the NB distribution are slightly higher than the RSI procedure. As an example, Figure 5 shows plots of the read depth data for six different CNVs identified for the child, 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.

4 Detection of recurrent CNVs based on a population-based samples

4.1 Statistical formulation and summary of theoretical results

Germline CNVs, as a population level genetic variants, often occur recurrently in individuals from a population. Recurrent CNVs are important targets for association study and other down-stream population genetic analysis. When a large set of samples from a population are available, it is of great interest to pool information from multiple samples to identity recurrent CNVs in the population. This is especially relevant when the CNV signals from one single sample are not strong enough to be detect, however, pooling information across multiple samples can greatly increase the power of detecting such CNVs.

While efficient procedures have been developed for identifying CNVs in a long sequence of genome-wide observations, some type of post-processing is often used to select regions with highly recurrent CNVs. One problem with such an approach is that the power for identifying the recurrent CNVs does not improve with the increase of sample size. An important fact is that the locations of a recurrent CNV are mostly overlapping across

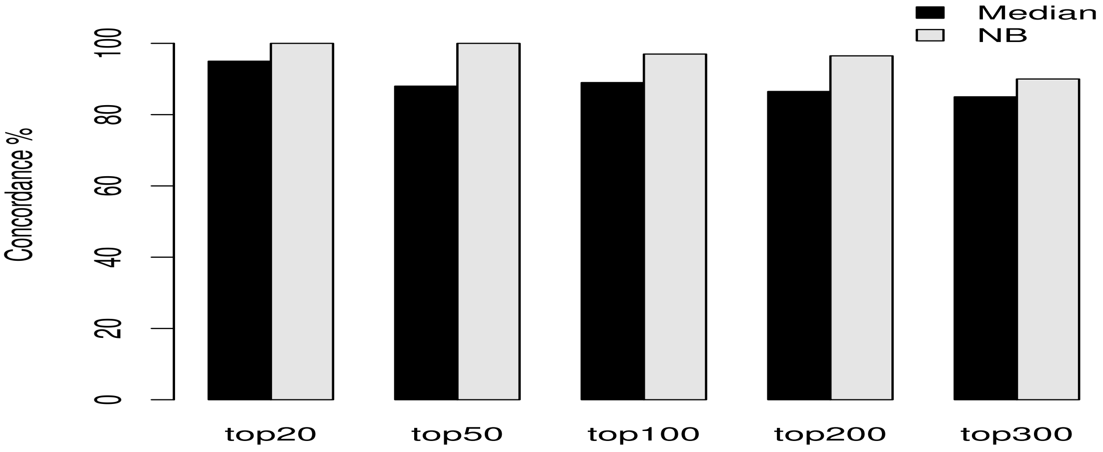

Figure 4: *Parents-child concordant rate of the CNV identified based on median (RSI) and negative binomial (NB) transformation of the RD data for the Yoruban trio.*

samples, so the improvement of identification power is possible if information from multiples samples can be efficiently pooled during the CNV identification step. In addition, most CNVs from the germline constitutional genome are very short and range mostly less than 20 single nucleotide polymorphism (SNPs) (Zhang et al. 2009) in typical Illumina 660K chip. Many of these short CNVs cannot be identified even by the optimal method based on data from a single sample (Jeng et al. 2010). Efficiently pooling information from multiple samples can greatly benefit the discovery of short CNVs that are missed in single-sample analysis. This has been nicely demonstrated by two recent publications. Zhang et al. (2010) introduced a method for detecting simultaneous change-points in multiple sequences that is only effective for detecting common CNVs. Siegmund et al (2010) further extended the method in Zhang et al. (2010) for identifying both the rare and common variants by introducing a prior probability of CNV frequencies that needs to be specified. No rigorous power studies were given in these two papers.

The model and the data for multiple sample CNV identification can be summarized as follows. Suppose there are N linear sequences (or samples) of noisy data and each sequence has T observations. Let X_{it} be the observed data for the ith sample at the tth location. If there are no signal variations, X_{it} scatters around 0 for any i and t. Suppose that at certain nonoverlapping segments (subintervals) $I_1, \ldots, I_q$ some samples have elevated or dropped means from the baseline (i.e., carriers) and others do not. Denote the collection of the segments as $\mathbb{I} = \{I_1, \ldots, I_q\}$, the carrier proportion at segment I_k in the population as π_k, and the magnitude of the segment for sample i as A_{ik}. Then an observation for sample $i \in \{1, \ldots, N\}$ at location $t \in \{1, \ldots, T\}$ can be modeled as

$$X_{it} = A_{ik} 1_{\{t \in I_k\}} + Z_{it} \quad \text{for some } I_k \in \mathbb{I}, \tag{4.14}$$

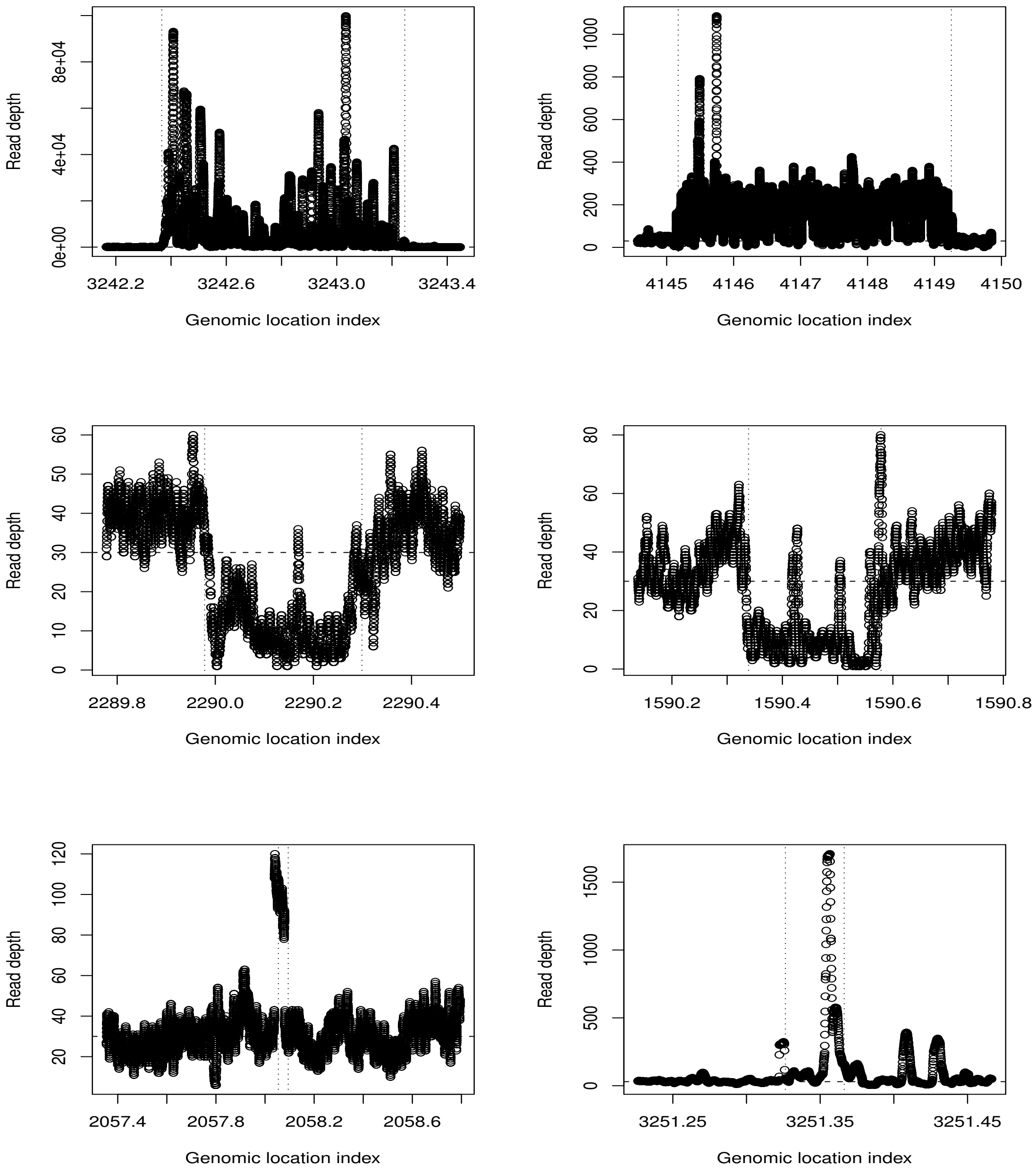

Figure 5: Examples of CNV identified by the RSI on chromosome 19 of NA19240 from the 1000 Genomes Project. Top two plots: duplications, regions with the highest scores; middle two plots: deletions, regions with the smallest scores; bottom two plots: the two shortest CNVs identified. For each plot, the 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.

where

$$A_{ik} \sim (1-\pi_k)\delta_0 + \pi_k N(\mu_k, \tau_k^2), \qquad \mu_k \neq 0, \quad \tau_k^2 \geq 0, \tag{4.15}$$

δ_0 is a point mass at 0, and $Z_{it} \sim N(0, \sigma_i^2)$. The noise variance σ_i^2 for sample i can be easily estimated when T is large and the signal segments are sparse in the linear sequence of data for sample i. For example, the robust median absolute deviation (MAD) estimator can be applied. We assume $\sigma_i^2 = 1$ in theoretical analysis. All of the other parameters I_k, π_k, μ_k, τ_k, $1 \leq k \leq q$ are unknown. From this model, if t is not in any signal segment, X_{it} is Gaussian noise following $N(0, \sigma_i^2)$. If t is in a signal segment I_k, then

$$X_{it} \sim (1-\pi_k)N(0, \sigma_i^2) + \pi_k N(\mu_k, \sigma_i^2 + \tau_k^2). \tag{4.16}$$

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

Our goal is two-fold: (1) to detect the existence of recurrent segment variants across samples; and (2) to identify the locations of the segments. Precisely, we wish to first test

$$H_0 : \mathbb{I} = \emptyset \qquad \text{agains} \qquad H_1 : \mathbb{I} \neq \emptyset, \tag{4.17}$$

and if H_0 is rejected, detect each $I_k \in \mathbb{I}$.

A major challenge of pooling information from multiple samples to discover recurrent CNVs is that the CNV carrier's proportions vary a lot for different CNVs. Jeng et al. (2011) studied the problem from the perspective of sparse signal detection and proposed a proportion adaptive segment selection (PASS) procedure. Denote the set of the intervals with length less than or equal to L by $\mathbb{J}_{T,N}(L)$. For any interval $\tilde{I} \in \mathbb{J}_{T,N}(L)$, we calculate the standardized sum of observations in $\tilde{I}$ for each sample as

$$X_{\tilde{I},i} = \sum_{t \in \tilde{I}} X_{it} / \sqrt{|\tilde{I}|}, \qquad 1 \leq i \leq N. \tag{4.18}$$

By (4.14) and the assumption $\sigma_i^2 = 1$, $X_{\tilde{I},i}$ follows $N(0,1)$ under H_0. When $\tilde{I}$ overlaps with some signal segment, $X_{\tilde{I},i}$ follows a heterogeneous and heteroscedastic Gaussian mixtures according to (4.16). Specifically, when $\tilde{I} = I_k$ for some $I_k \in \mathbb{I}$,

$$X_{I_k,i} \sim (1-\pi_k)N(0, \sigma_i^2) + \pi_k N(\mu_k \sqrt{|I_k|}, \sigma_i^2 + \tau_k^2). \tag{4.19}$$

Note that the mean value of the second component includes the information of jump size and length of the segment variant at I_k.

The key of the PASS is to use the HC statistics (Donoho and Jin, 2004) to pool information across multiple samples based on the statistics $X_{\tilde{I},i}$ in order the identify the CNV regions. The PASS procedure can automatically adjusts to the unknown carrier's proportion and optimally detect both the rare and common CNVs. Jeng et al. (2011) showed that PASS has desirable theoretical and numerical properties. They further characterized the detection boundary that separates the region where a segment variant is detectable by some method from the region where it cannot be detected by any methods. Despite the fact that the detection boundaries are very different for the rare and common segment variants, Jeng et al. (2011) showed that PASS can reliably identify both the rare and common segment variants whenever they are detectable. Compared with methods for single sample analysis, PASS significantly gains power by pooling information from multiple samples.

Similar to the LRS procedure, PASS assumes that the noises of the LRR data follow a normal distribution. Great care must be taken to ensure that the data are approximately normal. An interesting problem for future research is the CNV identification by population-scale genome sequencing (Mills et al., 2011). The next generation sequencing technology can generate billions of counts data along the whole human genome and genotype much more DNA regions with rare variants. Due to complexity of the sequencing process, error distribution of the counts data is unknown and is difficult to characterize parametrically. An promising approach is to apply the median transformation to the read depths data and then to apply the PASS to the transformed data.

4.2 Application to analysis of neuroblastoma cases

We applied the PASS procedure to a sample of 674 neuroblamstoma cases that were collected as part of a large-scale genome-wide association study of neuroblastoma (Diskin et al., 2009). For each sample, about 600,000 SNPs were genotyped using the Illumina genotype platform and the log R-ratios data were obtained. In order to account for possible wave-effect or local effects, we performed similar processing as in Zhang et al. (2008) to obtain the normalized data, including subtracting the sample median, local adjustment by regressing on the first principal component. In our analysis, we considered only data from the chromosome 1, which includes $T = 40,929$ SNP log R-ratios.

PASS $L = 20$ resulted in selection of 335 CNVs with length of three or more SNPs, including 171 CNVs with three SNPs and 100 CNVs with 4 SNPs, and 11 CNVs with 10 or more SNPs. The median size of the CNVs identified is 4,165 bps with a range of 462 bps and 1,038,000 bps.

Since the identification of the short CNVs are more susceptible to local wave effects or other artifacts of the data, we should interpret the CNVs of three or four SNPs with caution and focus the following comparison on the identified CNVs of 5 or more SNPs. Among the CNVs identified, 64 have five or more SNPs. Among these 64 CNVs, 30 overlap with the CNVs in the database of genomic variants (http://projects.tcag.ca/variation/project.html). Note that this database only includes the CNVs identified in healthy human cases and are relatively common. To further demonstrate the power of PASS, we also performed single-sample CNV identification using the optimal CNV identification procedure LRS. Among the 64 CNVs with 5 or more SNPs identified by PASS, 20 of them did not reach the theoretical threshold of $\sqrt{2\log(TL)} = 5.22$ in any of the 674 samples, indicating a great loss of power of detecting the CNVs based on the single-sample analysis. Of these 20 CNVs missed by single sample analysis, 10 of them overlap with the CNVs in the genomic variants database.

5 Conclusion and Further Discussion

Our review focuses only on the germeline CNV detection problems, where the CNVs are short and sparse. There are a few other problems that require further methods development.

5.1 Statistical tests for CNV associations

One problem is to identify the CNVs that are associated with a clinical phenotype. This is often performed in a two-step approach. First all the CNVs are identified for all the samples by some CNV identification methods. These identified CNVs are then tested through some simple regression models. One limitation of this approach is the information of shared CNVs across multiple samples are not effectively utilized. In addition, the CNVs identified based on each sample separately often do not have exactly the same staring/ending boundaries. This makes the summaries of CNVs across multiple samples difficult. Finally, since the number of the tests is unknown prior to the CNV calls, it is not clear how one should adjust for multiple comparisons. An alternative to the two-stage procedure is to test for CNV association only for those known CNVs. Barnes et al. (2008) developed a robust statistical method for case-control association testing with CNVs using the EM algorithm treating the observed CNVs as latent variables.

Besides testing for association between relatively common CNVs and the clinical phenotypes, statistical methods for testing rare CNV (found in $< 1\%$ of the total sample) association are also needed. Global rare CNV burden are often compared between cases and controls regardless of where the rare CNVs are (Girirajan et al., 2011). Zhang et al. (2009a) presented a genome wide copy number variant (CNV) survey of 1001 Bipolar disorder cases and 1034 controls using the Affymetrix SNP 6.0 SNP and CNV platform. Singleton deletions (deletions that appear only once in the dataset) more than 100 kilobases in length are present in 16.2% of BD cases in contrast to 12.3% of controls (permutation p = 0.007), indicating potential importance of considering the cumulative effects of rare CNVs on disease risk. Methods that have been developed for testing rare genetic variants associations can also been applied for testing rare CNV association with the difficulty that the carriers of the rare variants have to be inferred from data, which can be challenging unless the signals are very strong.

5.2 CNV analysis based on mapping distances of pair-end sequencing

Another area that needs further statistical research is that CNV detection problem based on pair-end sequencing data, where the mapping distances between the mate pairs also provide important information about the structural changes of the genomes. Figure 6 gives an illustration of how mapping distances can be used for inferring the deletion and duplication base on the pari-end data. Beside the CNVs, the mapping distances or mapping orientations also provide information on translocations and inversions. Rigorous statistical formulations of both single-sample and multiple-sample CNV detection based on the mapping distances are required. Besides statistical formulation of the problems, efficient computational methods are also necessary since typical data sets often include several billions of end sequences of matepairs.

Methods of CNV analysis based on the matepair mapping distances relies on the computation of the expected distance between matepairs in the donor genome, which is referred to as insert size, d_i for the ith matepair. The distribution of the insert sizes d_i can be determined (Tuzun et al., 2005). An alignment of a paired-end read to reference genome is called concordant (Tuzun et al., 2005), if the distance between aligned ends of a pair

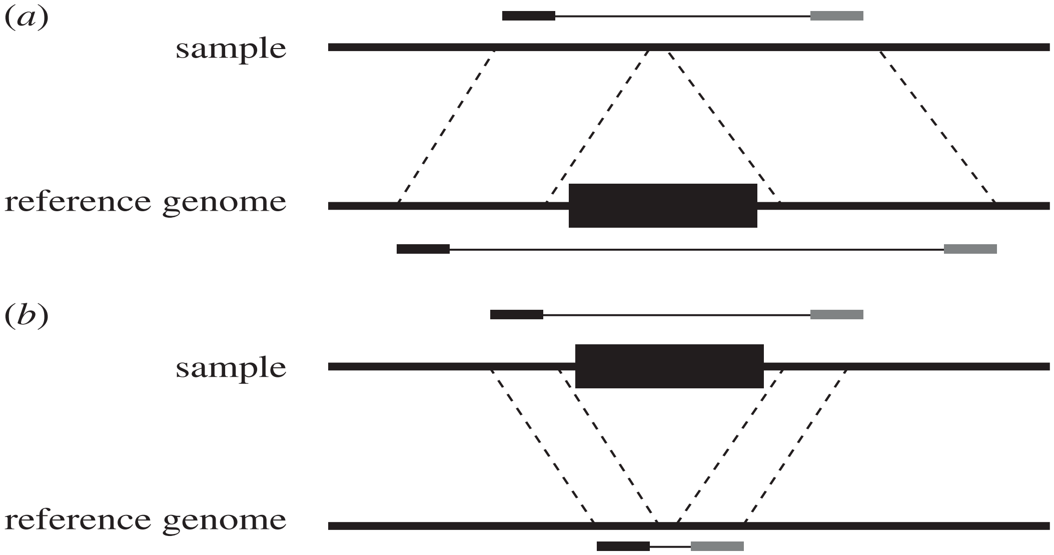

Figure 6: Illustration of detecting insertions and deletions based on the mapping distances from the paired-end mapping data. The ends of a DNA fragment from a sample individual are mapped to a reference genome (Schrider and Hahn, 2010). (a) If the portion of the reference genome spanned by the fragment ends is larger than expected, then the sample genome probably contains a deletion relative to the reference. (b) If the length of the region spanned by the locations of the end sequences in the reference genome is smaller than expected, then an insertion is inferred to be present in the sample genome.

in the reference genome is thought to come from the baseline insert size distribution, and both the orientation and the chromosome the paired-end read is aligned to are correct. For instance, in the Illumina platform, a paired-end read is considered to be aligned in the correct orientation if the left matepair is mapped to the "+" strand (which is represented by +), and the right mate pair is mapped to the "-" strand (which is represented by -). A paired-end read that has no concordant alignment in the reference genome (Tuzun et al., 2005; Lee et al., 2008; Hormozdiari et al., 2009) is called a discordant paired-end read, which indicates a possibility of a structure variant. Hormozdiari et al. (2009) proposed a combinatorial algorithms for structure variation detection and named their program *VariationHunter*. Our goal is to determine the discordant paired-end reads and use these reads to determine the CNV regions. A rigorous statistical formulation of the problem is needed in order to understand how true CNV lengths and depth of sequencing affect the power of detecting the CNVs.

5.3 CNV analysis by data integration

Finally, for the SNP chip data, B-allele frequencies also provide useful information for CNV detection and identifications. How to extend the LRS and PASS procedure to incorporate the B-allele frequencies data is an important topic in CNV research. For NGS data, an interesting research topic is to combine both read depth data and the mapping distances data in a unified framework for CNV detection.

Acknowledgments

This research is supported in part by NIH grant CA127334 and NSF FRG Grant DMS-0854973.

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 Research*, 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., Candes, E. J., Helgason, H., and Zeitouni, O. (2008), "Searching for a trail of evidence in a maze," *Ann. Statist.*, 36, 1726–1757.

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

Barnes, C., Plagnal, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D., and Hurles, M. (2008), "A robust statistical method for case-control association testing with Copy Number Variation," *Nature Genetics*, 40, 1245–1252.

Benjamini, Y. and Hochberg, T. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. Royal Stat. Soc. B*, 85, 289–300.

Benjamini, Y. and Speed, T. P. (2011), "Estimation & correction for GC-content bias in high throughput sequencing," *Technical report, UC Berkeley.*

Bhattacharya, P. (1994), "Some aspects of change-point analysis," *In Change-point Problems, IMS Monograph 23, E. Carlstein, H. Muller and D. Siegmund, eds. Institute of Mathematical Statistics, Beachwood, OH*, 28–56.

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

Bucan, M., Abrahams, B., Wang, K., Glessner, J., Herman, E., Sonnenblick, L., Retuerto, A. A., Imielinski, M., Hadley, D., Bradfield, J., Kim, C., Gidaya, N., Lindquist, I., Hutman, T., , Sigman, M., Kustanovich, V., Lajonchere, C., Singleton, A., Kim, J., , Wassink, T., McMahon, W., Owley, T., Sweeney, J., Coon, H., Nurnberger, J., Li, M., Cantor, R., Minshew, N., Sutcliffe, J., Cook, E., Dawson, G., Buxbaum, J., Grant, S., Schellenberg, G., Geschwind, D., and Hakonarson, H. (2009), "Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes," *PLoS Genetics*, 5, e1000536.

Cai, T. T., Jeng, J. J., and Li, H. (2011), "Robust Detection and Identification of Sparse Segments in Ultra-High Dimensional Data Analysis," *JRSS-B, in press.*

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

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

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," *Nature Methods*, 6, 99–103.

Diskin, S. J., Hou, C., Glessner, J. T., Attiyeh, E. F., Laudenslager, M., Bosse1, K., Cole1, 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.

Donoho, D. and Jin, J. (2004), "Higher criticism for detecting sparse heterogeneous mixtures," *Ann. Statist.*, 32, 962–994.

— (2008), "Higher Criticism thresholding: optimal feature selection when useful features are rare and week," *Proc. Natl. Acad. Sci.*, 105, 14790–14795.

Feuk, L., Carson, A., and Scherer, S. (2006), "Structural variation in the human genome," *Nature Review Genetics*, 7, 85–97.

Girirajan, S., Brkanac, Z., Coe, B., Baker, C., L, L. V., and et al. (2011), "Relative Burden of Large CNVs on a Range of Neurodevelopmental Phenotypes," *PLoS Genetics*, 7(11), pages=.

Hall, P. and Jin, J. (2010), "Innovated Higher Criticism for detecting sparse signals in correlated noise," *Ann. Statist.*, 38(3), 1686–1732.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009), "Mechanisms of change in gene copy number," *Nature Review Genetics*, 10, 551–564.

Hormozdiari, F., Alkan, C., Eichler, E., and et al. (2009), "Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes," *Genome Research*, 19.

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.

— (2011), "Optimal discovery of rare and common segment variants," *Manuscript*.

Lee, S., Cheran, E., and Brudno, M. (2008), "A robust framework for detecting structural variations in a genome," *Bioinformatics*, 24, i59–i67.

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

Marioni, J., Thorne, N., Valsesia, A., Fitzgerald, T., Redon, R., Fiegler, H., Andrews, T., Stranger, B., Lynch, A., Dermitzakis, E., Carter, N., Tavare, S., and Hurles, M. (2007), "Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization," *Genome Biology*, 8(1), R228.

Medvedev, P., Stanciu, M., and Brudno, M. (2009), "Computational methods for discovering structural variation with next-generation sequencing," *Nature Methods*, 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(1), e16327.

Mills, R. R., Walter, K., Stewart, C., ..., and Korbel, J. O. (2011), "Mapping copy number variation by population-scale genome sequencing," *Nature*, 470, 59–65.

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 (4).

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J., Cheung, S. W., Shen, R. M., Barker, D. L., and Gunderson, K. L. (2006), "High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping," *Genome Res*, 16, 1136–1148.

Schrider, D. R. and Hahn, M. (2010), "Gene copy-number polymorphism in nature," *Proceedings of The Royal Society (B)*, 277, 3213–3221.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., and et al. (2004), "Large-scale copy number polymorphism in the human genome," *Science*, 305, 525–528–97.

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

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O., Ingason, A., Steinberg, A., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J., and 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., and et al. (2008), "Rare chromosomal deletions and duplications increase risk of schizophrenia," *Nature*, 455, 237–241.

Sudmant, P., Kitzman, J., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., Project, . G., and Eichler, E. (2010), "Diversity of human copy number variation and multicopy genes," *Science*, 330, 641–646.

Tuzun, E., Sharp, A. J., Bailey, J., Kaul, R., Morrison, V., Pertz, L., Haugen, E., Hayden, H., Albertson, D., and et al., D. P. (2005), "Fine-scale structural variation of the human genome," *Nature Genetics*, 37, 727732.

Walsh, T., McClellan, J., McCarthy, S., Addington, A., Pierce, S., Cooper, G., Nord, A., Kusenda, M., Malhotra, D., Bhandari, A., and et al. (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. of Stat.*

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S., Hakonarson, H., and Bucan, M. (2007), "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data," *Genome Research*, 17, 1665–1674.

Xie, C. and Tammi, M. (2009), "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BMC Bioinformatics*, 10, 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 Research*, 19, 1568–1592.

Zack, S. (1983), "Survey of classical and bayesian approaches to the change-point problem: Fixed sample and sequential procedures in testing and estimation," *In Recent Advances in Statistics, Academic Press, San Diego, CA*, 245–269.

Zhang, D., Cheng, L., Qian, Y., Alliey-Rodriguez, N., Kelsoe, J., Greenwood, T., Nievergelt, C., Barrett, T., McKinney, R., and et al (2009a), “Singleton deletions throughout the genome increase risk of bipolar disorder,” *Molecular Psychiatry*, 14, 376–380.

Zhang, F., Gu, W., Hurles, M., and Lupski, J. (2009b), “Copy number variation in human health, disease and evolutions,” *Annual Review of Genomics and Human Genetics*, 10, 451–481.

Zhang, N. R. and Siegmund, D. O. (2007), “A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomics Hybridization Data,” *Biometrics*, 63, 22–32.

Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. (2008), “Detecting Simultaneous Changepoints in Multiple Sequences,” *Biometrika*, 00(0), 1–18.

Zhang, Z., Lange, K., Ophoff, R., and Sabatti, C. (2010), “Reconstructing DNA copy number by penalized estimation and imputation,” *The Annals of Applied Statistics*, 4, 1749–1773.